Drew University

College of Liberal Arts

What Makes a Good Quarterback?

Analysis of Longitudinal NFL Data Using Latent Variable Clustering Methods

A Thesis in Statistics

By

Lloyd Goldstein

Submitted in Partial Fulfillment

of the Requirements

for the Degree of

Bachelor of Science

With Specialized Honors in Statistics

May 2020

Table of Contents

Abstract

This thesis uses latent variable clustering methods to analyze longitudinal NFL quarterback data in a previously unexplored way. The main method used in this work is Latent Class Analysis (LCA) and its longitudinal extension Latent Transition Analysis (LTA). These methods use dichotomous, longitudinal performance data to create clusters of quarterbacks. Football performance data for 22 quarterbacks from 2012 to 2015 is used for analysis. The results of this clustering are then compared with results generated by other clustering methods. They are also compared with conventional football analysis from reputable websites such as ESPN. The results of the latent variable clustering methods are generally in line with those generated by other clustering methods. They also reflect conventional football wisdom quite accurately, but with a bit more specificity. For this football data set, latent variable clustering methods are effective and interpretable methods of quarterback classification.

1. Introduction

Statistical techniques and methods can be leveraged for many different purposes, from finding relationships between height and weight to predicting academic performance based on hours of studying. To do this, we use statistical techniques such as hypothesis tests, confidence intervals, and regression models. Another common type of statistical method is one used for clustering. Clustering is the process of finding groups in data (Kaufman and Rousseeuw 1990). For instance, say we had the following group of objects: orange, monkey, blue, green, kangaroo, elephant. Statistical clustering methods would sort these items into two groups; one containing ‘orange, blue, green’ and one containing ‘monkey, kangaroo, elephant’ (see Figure 1). We, the intelligent user, would then be able to deduce that one group contains colors and the other contains animals. These methods can be applied to almost any scenario imaginable, and in this case I will be using these methods to classify quarterbacks in professional football.

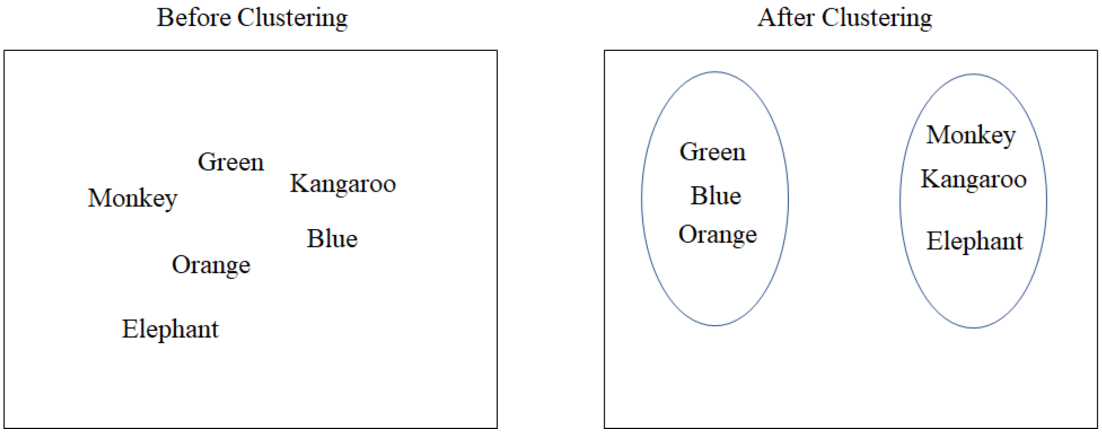

Figure 1. Visualization of Basic Application of Clustering Methods

In modern statistical practice, advanced computational power has increased the range of methods that can be utilized by statistical practitioners. Methods that previously were known but not practical for frequent or repeated use can now be repeated over and over with the click of a button. In addition to making known methods more practical and widespread, modern computational methods have allowed for the creation of entirely new statistical analysis techniques. One method that has only begun to come into mainstream usage since the turn of the millennium is a clustering method called Latent Class Analysis, abbreviated LCA, and its extensions. LCA is a latent variable method, just like the more well-known and established Factor Analysis. LCA differs from factor analysis in that it takes in dichotomous data instead of continuous data like factor analysis does. Its output also differs in that LCA outputs clusters that can be visualized instead of a mathematical model like factor analysis. One additional feature of LCA is that it can be easily extended to account for longitudinal data. This extension is referred to as Latent Transition Analysis (LTA). LCA and its extensions do have sound mathematical foundations, but make a strong conditional independence assumption and are much less established in the field of statistical analysis as a whole. One of the major goals of this thesis is to evaluate the performance of LCA and LTA compared to more established and well-known clustering methods.

The testing ground for this analysis will be longitudinal data on National Football League (NFL) quarterback performance. This thesis will attempt to model quarterback skill using latent variable methods. Since quarterback skill is a construct that most people agree exists but cannot be measured or observed directly in any way, it is a perfect test subject for modeling using LCA and LTA. This will be further expanded upon in the methods section, where I will discuss the difficulties with quantifying quarterback skill.

I will begin this thesis by providing background on how the sport of football works and the importance of the quarterback in the NFL. I will then give a brief overview of the different methods and statistics used to evaluate quarterbacks and other NFL players. Next, difficulties in quarterback evaluation and weaknesses in the current typical ways it is done will be discussed. The introduction will conclude by establishing the philosophy for NFL quarterback evaluation that will be used throughout the analysis.

The methods section of this thesis begins by establishing the data used for analysis and what variables will be examined. In this section, I will provide the justification for using the years 2012 through 2015 and the variables completion percentage, yards per attempt, yards per completion, touchdown percentage, and interception percentage. The section then introduces the LCA and LTA methodologies comprehensively. I then introduce two more popular clustering methods that will be used in this analysis: k-means clustering and hierarchical clustering. Lastly, the entire research and analysis process is walked through step by step.

The results section summarizes the findings of every part of the analysis, and compares the results of different statistical methodologies for NFL quarterback evaluation. The thesis concludes with discussion of the implications of these findings both for the future of the LCA and LTA methodologies and for the future of longitudinal NFL quarterback analysis.

2. Background

This chapter introduces the background information for my thesis. It introduces the NFL quarterback and its role and place in the league, then discusses how these players are typically evaluated. The chapter then introduces the statistical methods that will be used for the rest of the thesis, and concludes by explaining the overall goals and ideas of the analysis that will be conducted.

I. An Introduction to the NFL

The NFL is the most popular sports league in the United States. It draws tens of millions of viewers every week and is extremely profitable. This league consists of thirty-two teams who compete annually for a championship known as the Super Bowl in a season that spans from September to February.

In football, points are scored by moving the football down the field while the opposing team tries to prevent this from happening. This is accomplished by either running down the field with the ball or passing the ball to a teammate. The most important and valuable position in football is the quarterback, as he is usually the only offensive player on the field who is permitted to pass the ball down the field to his teammates. A good quarterback can throw the ball to his teammates consistently to move the ball down the field very quickly. However, quarterback is a very risky position to play; less-skilled quarterbacks may see their passes be caught, or ‘intercepted’, by members of the opposing defense. This transfers possession of the ball to the other team, causing a huge swing in the momentum and flow of the game. This means that the quarterback of an NFL team has more potential to both help and hurt his team than any other player on the roster.

II. How NFL Teams Acquire Quarterbacks

Since the quarterback is so integral to the success of an NFL team, both teams and fans of the NFL are strongly motivated to have the best possible quarterback on their team. NFL teams have two sources of new players. The first is by drafting one out of college. Although it is the easiest way to obtain promising new quarterbacks, this is a notoriously unreliable process. Teams have a generally poor track record at identifying talent out of college, with many highly-touted college prospects failing to succeed at the professional level and plenty of successful professional quarterbacks not being selected until later positions in the draft (Kuzmits and Adams 2008). For instance, Tom Brady, widely regarded as the greatest quarterback of all time, was not selected until pick number 199 in the draft. In contrast, Jamarcus Russell was taken first overall in the 2006 NFL draft and never had any success in the pros. There are good reasons why this is a difficult evaluation problem for teams to solve, however. College football is a meaningfully different game from NFL football, which makes projecting success in the NFL based on college production an exceptionally difficult task.

The other way to find a quarterback is to get one from another NFL team who has already had experience at the professional level, either through trade or by signing him in free agency. This can also be difficult because teams who are aware they have a good quarterback are generally unwilling to part with him because of how integral this player can be to a team's success. However, teams are still not reliable at evaluating and judging quarterbacks even when they have several seasons of NFL-level play under their belts. For instance, after being in the NFL for four years, Brock Osweiler was paid $72 million over four years by the Houston Texans. Osweiler was a total disappointment, however, and was traded to the Cleveland Browns after only one year of the contract had been paid out.

III. The Franchise Quarterback: Perception of NFL Quarterbacks

Fans and organizations have a convenient term for a quarterback that will enable their team to have long-term success: a 'franchise quarterback'. A franchise quarterback is generally seen as one who could win a Super Bowl, given a strong team and coaching staff surrounding him. Although it might seem like any quarterback could win a Super Bowl with a strong enough supporting cast, the reality is that the quarterback is such an important position that lackluster performance in this position can put a damper on what a team is able to accomplish. A good example of this is the 2018 Chicago Bears, who featured an incredibly talented defense and strong weapons at positions such as wide receiver and running back, but were held back by the mediocre play of quarterback Mitchell Trubisky. If a team really believes that its quarterback is holding back the rest of the team, it can even spark a wholesale rebuild of the roster. In contrast, having a good franchise quarterback means a team can never be completely out of contention. The ultimate example of this is the New England Patriots, who were one of the very top teams in the NFL from 2001 to 2019 largely thanks to their quarterback Tom Brady. Brady has received numerous accolades in this time frame, including four Super Bowl MVP awards and three NFL MVP awards (*Tom Brady*). Another factor to keep in mind when considering the importance of a franchise quarterback to a team's roster is that quarterbacks tend to have longer careers than most other positions. While a truly dominant running back or defensive lineman might be able to drag a team into temporary contention even without the support of a true franchise quarterback, these players tend to peak in their late twenties or early thirties and decline shortly thereafter. In contrast, quarterbacks can continue to play at a high level well into their mid and even late thirties. Quarterbacks are also viewed as the most important players in football by outside organizations; of the last twenty winners of the Associated Press NFL Most Valuable Player award, eighteen have been quarterbacks (*AP NFL MVP Winners)*. These reasons emphasize that what type of quarterback a team has is the biggest determining factor for every other organizational decision that a team will make for the foreseeable future. One of the main goals of this research will be to determine what exactly constitutes a franchise quarterback.

IV. Statistical Evaluation of NFL Quarterbacks

When teams are trying to obtain a quarterback from another organization or attempting to decide whether an upgrade is necessary in the first place, they evaluate the player to determine how helpful the player is to the success of the team. This concept of 'evaluating' quarterbacks has taken many forms over time, some of which involve the use of statistics.

In recent decades, the use of statistics in sports has been increasing at a rapid pace. This phenomenon is generally acknowledged to have first taken root in baseball in the early 2000s. Public awareness of this trend was critically boosted by Michael Lewis' book Moneyball and its movie adaptation starring Brad Pitt. Moneyball introduces a fundamental component of how statistics have come to be used in sports: to find and exploit market inefficiencies (Lewis 2004). This phase refers to a team's ability to find players or strategies that have been undervalued in some way by other teams and exploiting them to become more successful without having the same amount of resources as other teams. In football, teams can exploit market inefficiencies by finding players that other teams do not believe are strong contributors to winning and using them in more intelligent or more efficient ways. A main goal of this research is to investigate inefficiencies of this type in how NFL teams view their quarterbacks.

Naturally, for this type of inefficiency to exist in football, NFL teams must not have perfect understanding of how every player contributes to their chances of winning a game. There are a few reasons why statistical understanding of how the game operates has been slower to evolve in the NFL than in other popular sports leagues. The first reason for this is that there are so many different interactions happening on a football field at the same time. Each team has eleven players on the field at the same time, and each player can interact with every other player. This is a far cry from a sport such as baseball, where much of the time only two players (a pitcher and a hitter) are interacting. Another reason is that the definition of 'success' for an NFL player on a given play varies between plays, and is often completely unknown to third-party observers. In a given NFL play, each person on the field has a specific assignment he is tasked to fulfill by his coaches. Ideally, the ‘skill’ or ‘goodness’ of an NFL player could be found by evaluating how often he fulfills his assignment on a play properly. Unfortunately, there is often no way for spectators to be able to identify exactly what a player is supposed to be doing on a given play. For instance, offensive linemen are often instructed to ‘block’ a defender and prevent him from reaching the quarterback, but sometimes they are instructed to miss their blocks on purpose and allow a defensive player to run past them. This makes evaluating the skill of a football player just by looking at what they do individually on the field effectively impossible.

Quarterbacks are just as difficult a position to evaluate as any other in the NFL. This is largely because the team and coaches surrounding a quarterback are just as responsible for his production on the field as the player himself. For instance, it takes two players to make a complete pass: the quarterback and a receiver. This means that a quarterback’s receivers can be a strong determining factor in how effective the quarterback is at passing the football. This extends to plenty of other positions on the field, such as the offensive line whose job it is to block for the quarterback and give them time to throw and the coaches who try to produce good plays for the quarterback to run. Therefore, to evaluate the skill of a quarterback properly steps need to be taken to ensure that the statistics being used to judge the quarterback are as representative as possible of the skill of the quarterback and isolate him as much as possible from the effects of his team. Strategies for doing this in this research will be discussed in the Methods section.

V. Current Methods for Evaluating NFL Quarterbacks

There are many methods currently used to evaluate NFL quarterbacks. The first, and most obvious, is just to watch the player perform on the field and subjectively decide how skilled he is. This is generally referred to as ‘watching film’. Although it is possible for intelligent viewers to gain valuable insight from this, just like most subjective evaluation methods there is massive room for error and a general lack of consensus.

Because of the potential unreliability of subjective evaluation methods, quantitative methods are frequently used to evaluate quarterbacks. For instance, when the NFL’s awards are announced, they generally cite quantitative values and not anything related to film study (*NFL First Team All-Pro*). The first, and most basic, quantitative method for evaluating quarterbacks, is to look at raw, objective statistics. These can be separated into two major categories: ‘volume’ statistics and ‘efficiency’ statistics. One example of a volume statistic is the total number of touchdowns a quarterback passes for over a season. A touchdown is objectively a good thing; it is the best possible outcome when the offense is on the field. Therefore, it is reasonable to conclude that a quarterback who passes for a lot of touchdowns is a ‘good’ quarterback, or one that helps his team score points and win football games. However, volume statistics have some flaws. As with any measure with potentially variable sample sizes, volume statistics can make comparing quarterbacks with differing amounts of pass attempts difficult. For instance, a quarterback who throws 20 touchdowns on 200 pass attempts is clearly much more effective than one who passes for 20 touchdowns on 400 attempts. This is further exacerbated by the fact that sample sizes vary wildly between quarterbacks. Different teams tend to ask their quarterbacks to throw different numbers of passes; teams that run the ball a lot will therefore pass less, leading their quarterbacks to necessarily accumulate less impressive volume statistics even if they are not lacking in skill compared to a quarterback whose coach asks him to pass the ball more often. Furthermore, many quarterbacks often do not play full sixteen-game seasons. Football is a physical sport and injuries are common, meaning that many quarterbacks will miss games at some points because of injury. Teams that have secured playoff positions will often rest their quarterbacks for the last game of the season to avoid injury, further confusing comparisons based on sample size. Although looking at volume statistics is important to a basic understanding of quarterback performance, they have too many flaws to be used exclusively for quarterback evaluation.

The most common alternative to volume statistics is efficiency statistics, which look into the rates at which various events take place (*How is Total QBR Calculated?*). One example of an efficiency statistic is yards per completion, which is calculated by dividing total passing yards by number of completions. This statistic evaluates how efficiently a quarterback can pass the ball and gain yards to move the ball down the field, making it more useful in many situations than volume statistics like total passing yards. The drawback to efficiency statistics is that if not viewed in context they can be skewed by smaller sample sizes. For instance, a quarterback who throws one pass for 50 yards in a game is much less effective than one who throws 40 passes for 500 yards in a game, which a quick look at volume statistics would show clearly. However, efficiency statistics would paint entirely the wrong picture in a comparison between these two hypothetical quarterbacks. Overall, volume and efficiency statistics both have strengths and weaknesses. They can both be useful when viewed in context, but can easily become misleading without the proper context.

One way to provide some context to these statistics in an attempt to better evaluate quarterbacks is with artificially created statistics, which I will refer to as ‘adjusted statistics’. One example of an adjusted statistic is adjusted net yards per attempt (abbreviated as AY/A), which is calculated using a formula that incorporates sack yardage, passing yardage, interceptions, and touchdown (*ITP Glossary*). This is an adjusted statistic because although the actual statistic is objective and mathematical if one accepts that it is a valid way to evaluate quarterback performance, the way the measure is constructed is fairly subjective. Because of how the statistic is designed, it has the potential to favor ‘safe’ quarterbacks that throw few touchdowns and few interceptions over ‘risky’ quarterbacks that throw a lot of both; note that more touchdowns is better, while less interceptions is better. This could paint a picture of one style of playing quarterback being objectively better than another, whereas in reality this might be a matter for subjective debate.

Adjusted statistics tend to be how most people evaluate quarterbacks holistically. For instance, the NFL itself uses an adjusted statistic called ‘passer rating’ to determine its passing leader for each NFL season. Passer rating is calculated with a combination of efficiency statistics weighted in various different ways. This measure does have the advantage of combining four aspects of a quarterback’s performance into one convenient package: completions per attempts, yards per attempt, touchdown passes per attempt, and interceptions per attempt. Just like any adjusted statistic, though, passer rating has its flaws. These flaws are mostly related to how different parts of the passer rating calculation have hard caps. The formula used to calculate passer rating truncates any quarterback who passes for at least 77.5% completion, 12.5 yards per attempt, and an 11.875 touchdown percentage. For instance, passer rating will fail to differentiate between a quarterback who passes for 13 yards per attempt and one who passes for 18 yards per attempt, even though clearly the second quarterback is objectively more effective and efficient than the first. A more extreme example would be a comparison between one quarterback who throws for a 95% completion percentage, 20 yards per attempt, one touchdown per 5 attempts, and one interception per 20 attempts, and a second quarterback who throws for an 80% completion percentage, 13 yards per attempt, one touchdown per 7 attempts, and no interceptions. In this example, although the first quarterback described is more effective, the second player will actually have a higher passer rating. Passer rating also fails to incorporate rushing performance in any way, which for some quarterbacks is a large part of their production. This is not to say that passer rating is a bad statistic on an objective level, of course. The goal of these examples is to establish that adjusted statistics can be misinterpreted and misused the same way as volume and efficiency statistics; context is always important when looking at these measures, regardless of how all-encompassing the measure in question might seem.

With passer rating, the importance of context is particularly noteworthy with regards to the ‘perfect passer rating game’. Although the passer rating truncation does not tend to affect entire quarterback seasons because it is extremely difficult to maintain production of that caliber for an entire sixteen-game season, there have been plenty of instances of quarterbacks achieving the maximum possible passer rating of 158.3 for a single game. This would imply that these games where the maximum passer rating is achieved are perfect in every way and could not be improved on at all, but this is not necessarily true. For instance, Lamar Jackson had a perfect game on November 10, 2019 with 223 yards and 3 touchdown passes, while Jared Goff had one on September 27, 2018 with 465 yards and 5 touchdown passes. Although Goff's game clearly looks more impressive, passer rating fails to differentiate his game from Jackson's because of how it truncates very high-performing quarterbacks.

One mitigating factor in considering these current methods is that the general public is not aware of how NFL organizations internally evaluate quarterbacks, both those currently on their own team and prospective players they wish to bring into the organization. It is reasonable to assume that NFL coaches and executives spend more time actually watching footage of quarterbacks than fans and mass media, as they have more time to do so and are more generally invested in their teams' success than the general public. Nonetheless, it can still be assumed that they do make use of these statistics; if nothing else, the NFL itself gives out awards based on passer rating (*Football Encyclopedia of Players*).

Adjusted statistics are also one of the primary ways that the greater statistics community engages with statistics both in football and in the world of sports in general. Many different new adjusted statistics have been created to help paint a better and more holistic picture of player performance using statistical techniques (White and Berry 2002, Berry and Burke 2012). This can also be seen in other sports; basketball statisticians have created specialized statistics such as Career-Arc Regression Model Estimator with Local Optimization (CARMELO), a statistic used to create intuitive 'profiles' of players (Silver 2018). As with other sports, baseball is also ahead of the curve here, having several different complex ways to measure the amount of wins a player actually provides for his team; this is called Wins Above Replacement (WAR). Although baseball is much simpler than other sports to quantify, in an ideal world we would want to have WAR-like statistics for other sports as well; after all, the primary goal for any athlete in any sport is to win.

VI. Statistical Learning Methods

Now that we have a decent understanding of the basic numerical measures used to evaluate quarterbacks, we can begin to look into more complex statistical methods used to analyze their performances. Statistical learning methods can be sorted very generally into supervised and unsupervised learning methods (James et al. 2017, 26). Supervised learning methods tend to output a predicted value or number, such as a predicted income or attitude. They are often evaluated with training and test sets, allowing for an evaluation of how effective the created model is. These methods are commonly used by football statisticians because a crucial goal of evaluating quarterbacks as a whole is to predict how they will perform in the future. This is particularly important for NFL organizations. When an organization is looking to bring in a new quarterback, they are not doing it just because of what the player has already done but what they think the player can provide in the future. This means that they will be interested in projecting how many yards, touchdowns, and other types of production a quarterback will accumulate in coming years, not just looking back at the past.

In contrast, unsupervised learning methods do not have a predicted output and as such cannot be ‘correct’ or ‘incorrect’ (James et al. 2017, 26). One type of unsupervised learning methods is clustering methods. These methods have the goal of sorting observations into different groups, or clusters. This allows for observations to be compared to each other on a relative instead of absolute basis. There are many different ways to perform clustering, but every method has the ultimate goal of grouping together observations that are similar in some way. For instance, a clustering method for quarterbacks might group together quarterbacks who passed for similar numbers of yards in a given season. Unsupervised methods do not output a single number or equation in the way that supervised learning methods do, but can be just as useful with proper interpretation and context. For instance, with clustering methods it is important to look subjectively at which observations fall into which clusters and evaluate what constructs they might have in common.

VII. Challenges With Current Statistical Methodologies

Although the current statistical methods used in NFL quarterback evaluation certainly are useful and have plenty of merit, they have some global flaws. These flaws can be summed up as a lack of context. One way that these statistics lack context is that the NFL is a league that shifts from year to year. As we will see when we perform univariate analysis of our data, what constitutes an ‘average’ or ‘good’ quarterback performance has shifted over time in the NFL. In general, passing volume statistics have gone up over time as the years have passed in the NFL basically since its inception. One important aim when creating statistics to evaluate quarterbacks is to be able to compare the performance of quarterbacks in different years, and this is another notable area where football statistics lag behind sports such as baseball. In baseball era-adjusted statistics such as WRC+ have been created that allow for reasonably accurate comparison between players from different eras, but an equivalent to this for quarterback evaluation has yet to be created (*wRC and wRC+*). Clearly it should be possible to create era-adjusted passing statistics for football taking into account how well a given quarterback performs relative to his counterparts in the same year, but as of now this type of statistic has yet to appear in football analysis discourse to the best of my knowledge.

As discussed in section 1.V, another way that these evaluation methods tend to be subjective is that creating and selecting a statistic is in and of itself an inherently subjective process, since the evaluator is making a judgement of what traits they value in a quarterback. A basic example of this would be that in many methods of quarterback evaluation discussed above, quarterback rushing ability is not factored in in any way; one example of a measure with this flaw is ESPN’s Total QBR, which claims to be a holistic measure of quarterback performance but does not seem to take into account rushing (ESPN 2016). Quarterbacks are uniquely powerful in football because they have the power to either pass the ball or keep it and run with the football to advance it down the field. However, total QBR does not take this into account in any way. Although one might argue that a quarterback’s rushing ability has nothing to do with his passing ability, this is not the case. Quarterbacks who are a threat to run with the football force the opposing defense to play in a different way, making it easier for them to pass the ball more effectively. This means that Total QBR, although decent for capturing passing effectiveness, fails to capture the full picture of a quarterback's productivity. Although this might seem like a small nitpick, the point is that there are issues with every statistic that could possibly be devised to measure quarterback productivity, no matter how statistically sound. For this reason, statistics that do not claim to take into account every potential factor to create a measure of performance can actually be more useful because they tend to do a better job at accomplishing a smaller set of goals. For instance, passer rating does not take into account rushing performance, but nor does it claim to; it tries to be a full picture of quarterback passing performance, and does a good job of accomplishing that (although it is not flawless).

VIII. Goals for Longitudinal NFL Quarterback Evaluation

As I have discussed, any 'objective' statistic used to evaluate a quarterback, whether it is a raw statistic or something adjusted by statisticians, will inevitably have flaws. My goal is not to try to create a better method than everyone else who has tried. Many highly competent sports statisticians, analysts, and mathematicians have worked to create measurements to evaluate quarterbacks and other athletes, and none of them are exempt from criticism or nitpicks (Kuzmits and Adams 2008, Franks et al. 2008). In this research I will take a different angle by comparing quarterbacks to each other using clustering methods. Instead of trying to create a basis or underlying structure for quarterback evaluation, I will instead use the quarterbacks themselves as the baseline of performance for each other. In this way, I minimize the amount of my own personal bias present in the analysis. Another important component of this analysis is that it will be longitudinal; the data analyzed will come from several years. It should not be forgotten that the main goal of quarterback evaluation is to determine the skill of a given quarterback, and this simply cannot be done reliably with only one season of data. One season of football is not a large sample size, and there have been plenty of instances throughout NFL history of quarterbacks having one strong season and failing to reproduce it. Players only play sixteen regular season games a season maximum, which does not yield a large sample size. This makes rigorous statistical analysis more of a challenge than in a sport such as American major league baseball, where players go through 162 games and accumulate much more data. Using longitudinal analysis will allow us to take into account more information. This analysis will leverage past advances in sports statistics, particularly in NFL quarterback evaluation, to cluster quarterbacks across multiple seasons.

One noteworthy limitation of this work will be that it will not take into account quarterback rushing statistics. The first reason for this is that part of this analysis will include a comparison of the results for passer rating, which also does not include rushing statistics. The second is that comparing rushing production between quarterbacks is substantially more difficult than comparing passing production. This is because sample size for rushing attempts varies drastically between quarterbacks. For instance, in 2015 Cam Newton had 132 rushing attempts, while Joe Flacco only had 13 (*Football Encyclopedia of Players*). This would make any sort of attempt to compare the rushing production of these two quarterbacks entirely worthless.

3. Methods

This chapter will establish how the analysis will take place. I will discuss the data being used and variables being examined. After that, I will go through how the statistical analysis procedures will be run and how the results will be interpreted.

I. Data

a. Sources

The data used for this analysis comes from pro-football-reference.com. This website is a comprehensive database of football statistics that contains almost every player who has ever played in the NFL for almost every year of the league's existence. Since football is such a popular and well-documented sport there are plenty of options available for finding data, so I could have used any of several websites to source the data from. However, profootballreference has useful features such as options to export data to Excel and CSV formats and robust search functionality, so I used it to obtain my data.

b. Years

The first decision that had to be made in selecting data to use for this research was what years to examine. I decided to use a span four years for the analysis. This number of years means that the analysis will run across a long enough span of time to be able to identify longitudinal trends, but it will not span so many years that the results become messy and cluttered. I selected the years 2012-2015 to use for analysis. The goal of selecting this particular span of time was to have a set of years that are somewhat removed from the present so they are well documented, while also being recent enough where readers and audiences could engage meaningfully with the quarterbacks being analyzed.

c. Quarterbacks

After determining what years the data would be drawn from, the next step in the data selection process was to select which quarterbacks would be examined and clustered. The first requirement for a player to be included in this analysis was that he needed to play most of every season from 2012 to 2015. The bounds set were that the player needed to start at least 10 games for three of the four years, and at least 4 in the fourth; the allowance for one shorter year was necessary because long-term injuries are very common in the NFL. Since the goal of the analysis is to compare quarterback passing skill throughout the entire time span we are examining while including as much data as possible, we obviously want to maximize the data we have available for the quarterbacks we are analyzing. Some quarterbacks were technically in the NFL for these four seasons, but were his team's second-string quarterback and so did not play enough to accumulate meaningful data.

These criteria yielded the following list of quarterbacks for analysis.

Table 1: List of Quarterbacks Used for Analysis

Name	Team(s)
Tom Brady	New England Patriots
Drew Brees	New Orleans Saints

Jay Cutler	Chicago Bears
Andy Dalton	Cincinnati Bengals
Eli Manning	New York Giants
Ryan Fitzpatrick	Buffalo Bills, Tennessee Titans, Houston Texans, New York Jets
Joe Flacco	Baltimore Ravens
Nick Foles	Philadelphia Eagles, St. Louis Rams
Colin Kaepernick	San Francisco 49ers
Andrew Luck	Indianapolis Colts
Cam Newton	Carolina Panthers
Peyton Manning	Denver Broncos
Carson Palmer	Oakland Raiders, Arizona Cardinals
Philip Rivers	San Diego Chargers
Aaron Rodgers	Green Bay Packers
Ben Roethlisberger	Pittsburgh Steelers
Tony Romo	Dallas Cowboys
Matt Ryan	Atlanta Falcons

Alex Smith	San Francisco 49ers, Kansas City Chiefs
Matthew Stafford	Detroit Lions
Ryan Tannehill	Miami Dolphins
Russell Wilson	Seattle Seahawks

This is a total of 22 quarterbacks, which means that nearly three-quarters of the 32 NFL teams are represented in the analysis in any given year. The other ten teams cycled through multiple quarterbacks in this span of four years and as such did not have a representative in the data. For instance, the Jacksonville Jaguars used Chad Henne as their starting quarterback in 2012 and 2013 but changed to Blake Bortles for 2014 and 2015.

d. Variables Examined: Completion Percentage, Yards/Attempt, Yards/Completion, Touchdown Percentage, Interception Percentage

The last part in the data selection process was to determine what variables to examine for the quarterbacks chosen for analysis. There were multiple considerations when deciding what variables to use for this research. The first decision to be made was clearly what type of variable to use in the first place. I decided it would be better to use efficiency statistics rather than volume statistics, since although the sample sizes for all the quarterbacks in the analysis are relatively similar there are still fairly notable differences. These differences in games played most frequently happened because of injuries. Quarterbacks such as Cam Newton suffered injuries in some seasons that rendered them only able to play ten or eleven games. Although this might not sound like a big deal, it turns out to give quarterbacks who never miss games such as Philip Rivers a fairly substantial advantage. Although it could be argued that being healthy and available to play at all times is a skill, the goal of this analysis is to compare the effectiveness of these quarterbacks while they are on the field. Full season statistics were used instead of single-game statistics because individual football games are variable enough that looking at a single game is a poor measure of skill, and because the sample size for individual games would differ greatly between quarterbacks. Five efficiency statistics were chosen to use as variables for all methods. Although each of these statistics has their own individual weaknesses, by examining all of them together we can attempt to get a clear and holistic picture of overall quarterback passing performance.

The first statistic chosen for the analysis is Yards per Attempt (Y/A). This statistic is calculated by dividing the number of total passing yards a quarterback has in a season by his number of attempted passes. This is one of the most basic and universal efficiency statistics used for evaluating quarterbacks. The goal of this statistic is to see how effectively a quarterback moves the ball down the field, which is clearly one of the main goals of any football team. One notable weakness of this statistic is that it does not take into account touchdowns or interceptions, which matter much more to the outcome of football games than raw yardage. Another weakness of this statistic is that it is strongly affected by incomplete passes. Average completion percentage in the NFL is around 60%, which means that a large portion of any pass attempts a quarterback tries will fall incomplete.

The second statistic used is Yards per Completion (Y/C). Similar to Y/A, this statistic is calculated by dividing the number of total passing yards a quarterback has in a season by his number of completed passes. Although this may seem like it could cause issues by being too similar to Y/A, this is not the case. While Y/A is strongly affected by incomplete passes, Y/C is not. This means Y/C is more of a measure of how effective a quarterback's passes are, rather than how many passes he can complete. One weakness of this statistic is that it is disproportionately affected by extremely long completions. For example, a quarterback with five completions of ten yards will have a much lower Y/C than a quarterback with four incompletions and one completion of 90 yards, even though he might be more effective overall. Another weakness of Y/C is one that it shares with Y/A, which is that it does not take into account touchdowns or interceptions.

The third statistic used is Completion Percentage (CMP%). This statistic is simply the percentage of a quarterback's attempted passes are caught by a receiver from his team. This statistic is simple but very useful for evaluating certain qualities of a quarterback's performance, since whether or not a quarterback can pass the ball down the field successfully is a hugely important part of his overall skill. CMP% should never be used in isolation to evaluate quarterbacks, since it completely fails to take into account how effective these completed passes are. As part of a larger analysis, however, CMP% is a useful statistic that provides a good base measure of how effectively a quarterback can perform his primary task.

The fourth statistic used is Touchdown Percentage (TD%). This statistic is the percentage of a quarterback's attempted passes that are caught for a touchdown. This measure adds greatly to our analysis because it allows us to evaluate how useful a quarterback actually is at scoring points for his team, which is the overall goal of any football team's offense. One weakness of this statistic is that quarterbacks that play for teams that score more rushing touchdowns will necessarily have a lower TD%. This statistic also does not take into account how difficult it is for a team to drive down the field and score a touchdown in general. A quarterback that drives his team 60 yards down the field for a touchdown has a much more difficult job than a quarterback who only has to travel 20 yards, but if both drives end in a touchdown they will yield no meaningful difference in terms of TD%.

The fifth statistic used in this analysis is Interception Percentage (INT%). Similarly to TD%, INT% is the percentage of a quarterback's attempted passes that are intercepted, or caught by a defensive player from the other team. This is an important statistic because interceptions are one of the most impactful things that a quarterback can control in the game of football. However, INT% differs from the other statistics used in this analysis because it examines a negative impact a quarterback can have on his team. It should be noted that while most of the variables were dichotomized to be "1" if they were below median and "2" if they were above median, Int% was dichotomized in the opposite way. This was because the intent of the dichotomization was to have "1" represent low performance and "2" represent high performance. While for Cmp%, Y/A, Y/C, and TD% higher numbers indicate better performance, the opposite is true for Int% since throwing more interceptions is a bad thing. Although INT% is useful information to have about a quarterback, it clearly does not form a clear picture of quarterback performance in any way and thus must be supplemented with other statistics.

In combination, these five statistics give a well-rounded view of quarterback performance. These factors incorporate how effectively a quarterback can pass the ball to his teammates, how effectively he moves the ball down the field, how effectively he scores points for his own team, and how effectively he prevents the other team from obtaining the ball. These statistics do not account for quarterback rushing ability, but this analysis is only intended to address quarterback passing ability for the reasons outlined in section 1.VIII.

II. Univariate and Bivariate Analysis

Once the data are collected, I examine its distributions before conducting more in-depth statistical analysis. Next, I find basic summary statistics for each of the five measures used in the analysis. This is important because it gives us an idea of what sort of numbers we should be expecting and allows us to make better judgements of what constitutes a quarterback who is ‘good’ or ‘bad’ in these categories. We are trying to get a feel for the data before we start trying to make inferences about it; this is particularly relevant for those who do not have a background in football, as these people might have very minimal understanding of the typical (for example) yards per completion for an NFL quarterback. We also use histograms to visualize the shape and spread of the data. This is particularly useful in identifying outliers, which may correspond to quarterbacks who had unusually good or bad seasons.

After completing this initial univariate scan, we also create bivariate scatter plots of each pair of variables. This is important for identifying potential collinearity issues.

This is especially important to examine in regards to Y/C and Y/A, which logically could be correlated with each other. If there are observed collinearity issues, it may make sense to run the clustering procedures with different combinations of variables to see how much each individual variable adds to the results, as we do not want to use variables in our methods that do not add any information.

Univariate and bivariate analysis was all done with basic functions in R. Summary statistics were found with mean(), median(), and sd(). Histograms were made with hist(). Bivariate scatter plots were made with plot().

III. Statistical Learning Techniques

We will make use of a special type of statistical learning method called a latent variable method. A latent variable method is a technique that is designed to model an unobservable variable with zero error, here called a ‘latent’ variable (Collins and Lanza 2009, 4). This latent variable is believed to be associated with patterns of variables that can be observed. In this case, the latent variable that will be estimated is quarterback skill.

Clearly, quarterback skill cannot be measured on an absolute and objective basis. As we have established, however, there is no shortage of statistics that reflect the raw performance of a quarterback. This makes analyzing quarterback skill a great subject for latent variable methods. It is clear that quarterback skill must be somehow related to the statistical production of the quarterback on the football field, which is something that we have plentiful data about. If we accept that the five measures chosen form a holistic picture of skill, we should be able to use them in a latent variable method to accurately view quarterback skill.

Another factor that makes latent variable methods well-suited for this analysis is that they are generally used with a ‘person-first approach’ in mind (Collins and Lanza 2009, 8). This means that when using these methods, each observation should be viewed as a unique person whose traits should be viewed as parts of an entire person or being, rather than being the sum total of everything that they are. This also helps to make latent variable methods well-suited to our analysis.

This analysis will use a type of latent variable method called latent class analysis (LCA). Latent class analysis is a categorical clustering method that uses the latent variable method approach to grouping individuals. The method takes in categorical data; when scale data is used, as in our analysis, it is usually dichotomized as being above or below the median. It should be noted that this creates substantial risk of information loss by transforming widely ranging continuous data into binary data. This data is used to cluster each observation based on patterns of observed variables. In latent class analysis, each cluster is referred to as a ‘latent class’, as they are supposed to reflect a different value for the latent value being examined. Clustering is done by creating a likelihood function for each cluster with the expectation-maximization approach (Lanza and Collins 2009). In simpler terms, the likelihood function aggregates together the probability of each pattern of responses for each individual, weighted by the probability of being in each latent class. In this way, we attempt to cluster quarterbacks into different latent classes based on the latent variable of skill.

We explain the derivation of this likelihood function here for $i = 1,2,...,N$ individuals, $m = 1,2,...,M$ quarterback performance measures, and $k = 1,2,...,K$ latent classes. The LCA likelihood function begins with a binomial probability distribution of a specific response within a specific latent class, here being the probability of a specific performance measure being above or below the median. Here p_{km} represents the probability of the statistic being above the median in a specific latent class, and y_{ik} represents the response value. Note that y_{ik} can only be 1 or 0 here, since the data is dichotomized as above or below the median.

$$p_{km}^{y_{ik}}(1-p_{km})^{1-y_{ik}} \tag{1}$$

Next, the likelihood function assesses the probability of a given pattern of responses by taking the product of (1) across every response variable given *K* response variables, or in this case every quarterback performance measure.

$$\prod_{k=1}^{K} p_{km}^{y_{ik}}(1-p_{km})^{1-y_{ik}} \tag{2}$$

To account for each latent class, we take the summation of Equation 2 across every latent class given *M* latent classes. We weight the term of each latent class by the probability of a given observation falling into latent class π_m.

$$\sum_{i=1}^{M} \pi_m \prod_{k=1}^{K} p_{km}^{y_{ik}}(1-p_{km})^{1-y_{ik}} \tag{3}$$

To complete the likelihood function, we take the product of Equation 3 across the N individuals in the data set. This yields a likelihood function for latent class *π* and response probability *p* given observed value *Y*.

$$L(\pi, p|Y) = \prod_{i=1}^{N} \left[\sum_{i=1}^{M} \pi_m \prod_{k=1}^{K} p_{km}^{y_{ik}} (1 - p_{km})^{1-y_{ik}} \right] \tag{4}$$

In order to produce an optimal latent class model, this likelihood function must be maximized. This is accomplished here using the expectation maximization (EM) algorithm, which estimates a random likelihood function and tweaks it until it reaches a local maximum;. for more information, see e.g. Categorical Data Analysis by Alan Agresti (2003, 455-490).

Although LCA does not require many assumptions to hold to function properly, it does require one strong and important assumption to be met: the conditional independence assumption. This assumption states that within a given latent class, observed variables are independent. Although clearly observed variables will vary between latent classes, they must be independent within a given class. For instance, within a given latent class quarterback Y/C must be independent of Y/A. This assumption is necessary because the fundamental expression for LCA uses multiplication of two events to produce the probability of the intersection of the two events, which is only a mathematically valid calculation if the two events are independent. In practice, this assumption is usually checked by creating contingency tables and running statistical tests of independence such as chi-squared tests (Collins and Lanza 2009).

We have two different ways we can extend LCA to take into account our time series data. The first is with Repeated Measures LCA (RMLCA). RMLCA refers to the procedure of performing a separate LCA analysis at each time point. In our analysis, this means we would perform a separate LCA clustering procedure for each of the years we are examining. RMLCA is fairly simple and straightforward, as it is the most basic way to use LCA for longitudinal data. However, it does not enable us to use data from one year to inform our clustering and analysis decisions from other years. To do this, we would use an extension of LCA called Latent Transition Analysis (LTA). LTA differs from RMLCA because it is capable of taking into account data from several different time points with one analysis. It creates one likelihood function for each latent class that spans the entire time period of the analysis, with each likelihood function being the product of what the likelihood function of that latent class would be for each different time being considered in the analysis. For instance, in our analysis the LTA likelihood function for Latent Class 1 would be the product of the likelihood function of Latent Class 1 in 2012, the likelihood function of Latent Class 1 in 2013, the likelihood function of Latent Class 1 in 2014, and the likelihood function of Latent Class 1 in 2015 (Collins and Lanza 2009, 198). LTA is a very powerful analysis tool because it enables us to easily evaluate how quarterbacks move between latent classes over time. This is important for our analysis because one of our goals is to see if quarterback skill is a construct that can hold up consistently over time, or if it is less consistent than we might expect from normal quarterback evaluation.

It should also be noted that LTA uses the same conditional, or local, independence assumption as LCA. It also adds another assumption, however; that of measurement invariance across time. This assumption states that each variable used in the analysis must consistently measure the same construct at each time point. For instance, this could be violated in an analysis of standardized test results when the makeup of the standardized test changes dramatically between years. Our analysis will not have any issues with this assumption, since clearly the statistical categories we are using measure the same thing every year. If we were measuring in absolute amounts of yardage and other categories this might pose a problem because the overall average amount of passing yards, touchdowns, and interceptions thrown in the NFL changes every year, but since we are categorizing the data as being above or below the median value for that particular year this will not pose an issue.

IV. Analysis Procedure

a. Data initialization and transformation

To begin the process of analysis, the data first must be downloaded from Pro Football Reference. This is done using the website's "Convert to Excel" functionality for each quarterback, then saving the Excel sheet as a .csv file and reading that file in using R's read.csv() function. Data then needs to be dichotomized into two categories, since LCA and LTA work with binary data. This is done by dichotomizing the data by whether it is above or below the median of all quarterbacks for the particular statistic for the current year. This is done using an if() clause in R. Data values exactly equal to the median were categorized as being below the median.

b. Number of Clusters

The next decision to make here is how many clusters will be created. In the case of the LCA and LTA analysis, we will evaluate the optimal number of clusters to use in multiple different ways. The first way will be using goodness-of-fit statistics. In particular, the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) will be examined to see what number of clusters fit the data best. The second way that the optimal number of clusters will be found is by considering interpretability. Since this analysis is intended to have relevant and understandable conclusions and implications in the sports world, we cannot use a procedure that will render the process confusing or unintelligible to the vast majority of potential readers. For instance, even if AIC and BIC indicate that nine clusters is the optimal number to break the quarterbacks into for the analysis, we would not use this number because that is far too many clusters to draw any intuitive conclusions from. It should also be noted that we will use the same number of clusters for each time point. Even if the goodness of fit statistics indicate that different times might potentially have different optimal numbers of clusters, this will reduce the interpretability of the clusters produced to too great an extent for it to be worth it. This means that there is often a tradeoff between goodness of fit and interpretability in these procedures, so a number of clusters needs to be selected that both preserves the statistical integrity of the procedure and can be reasonably understood.

c. Running the RMLCA procedure

The RMLCA procedure was done in R using the poLCA package. As mentioned above, one complete run of the RMLCA analysis involves an individual LCA analysis for each time examined. In our case, this means we need to run an LCA procedure for 2012, 2013, 2014, and 2015. We create the LCA model for each time with the poLCA() function using our dichotomized data. The first step after this model is created is to examine the distribution of posterior probabilities. Posterior probabilities in an LCA model are a measure of how certain the model is that observations are being assigned to the correct clusters. Ideally these probabilities should be extremely high. If they are low enough, generally below 0.7 or 0.6, this means that the model is basically just guessing which cluster the observation falls into, which means that it is probably not very reliable.

If the posterior probabilities look viable, we move on to actually examining the distributions of observations across latent classes. To find which latent class a given quarterback falls into at a given time, we look at its posterior probabilities; the model is assigning the quarterback to the latent class that it has the highest probability of falling into. We can then construct a basic frequency table to show how many quarterbacks fall into each latent class at each time.

The next important step in LCA analysis is to examine the characteristics of each latent class. This is done graphically by creating bar graphs that show, for each latent class at a particular time, what proportion of the quarterbacks in that latent class are above median in each of the five observed variables. These graphs are vital for being able to characterize latent classes in terms of quarterback performance.

The final step in RMLCA analysis is to evaluate how quarterbacks move between the latent classes over time. Since each LCA model is independent of each other, for this transition analysis to take place we must subjectively decide which latent classes represent the same constructs at each time point. For instance, in this analysis we need to identify a ‘high performance’ latent class from the LCA results in each year; by identifying this particular cluster, we can evaluate how consistently different quarterbacks are seen as being high-performing by the LCA method. Once these clusters are labeled, we can use transition matrices to evaluate how quarterbacks move between latent classes over time using RMLCA.

d. Running the LTA Procedure

Contrary to the rest of this work, LTA must be run in SAS using the PROC LTA procedure, as there has not yet been an R package created to execute this method. The PROC LTA procedure was created by Dr. Stephanie T. Lanza and collaborators and makes running this analysis fairly simple (Collins and Lanza 2009). The data for this part of the process is first cleaned and dichotomized in R as described above, and then exported as a .csv file that can be imported into SAS and used in the PROC LTA analysis. When the PROC LTA procedure is run, it outputs two main pieces of information. The first is the results of the goodness-of-fit tests described above, tested for every possible number of clusters between 2 and 9. This is convenient since it means we do not have to perform the goodness-of-fit tests manually ourselves. The second the PROC LTA procedure outputs is the posterior probabilities for each output at each time; in other words, the predicted probability that each quarterback will fall into each latent class in each year. This data will be the main target for the rest of our analysis. We save these posterior probabilities as a .csv file that we can import into R for the remainder of our analysis.

e. Other Clustering Methods

As one of the goals of this analysis is to investigate the performance of LCA and LTA relative to more well-known clustering methods, we must also perform other clustering methods to compare to our main results. The first one of these is k-means clustering. K-means clustering is a clustering method that works by creating *k* centroids and then assigning each observation to the cluster with the most similar or 'closest' centroid in an iterative process. One of the dangers of k-means clustering is that it may converge to a local endpoint instead of a global endpoint, similar to the EM algorithm. This means that the starting points of k-means clustering can play a large role in how observations end up being classified, and that it should generally be run multiple times with several different starting points; in this analysis, this was handled by running the analysis several times with different random starting points and verifying that similar cluster sizes with similar distributions of quarterbacks were produced. For more detail see e.g. Finding Groups in Data (Kaufmann and Rousseeuw 1990). This is done in R using the kmeans package; the analysis is run with the same number of clusters that we used for our LCA and LTA procedures for easier comparison. Before we were able to run the k-means clustering procedure, we had to scale the data because the different variables have very different profiles in terms of absolute range (e.g. completion percentage tends to range between 55 to 65, while yards per attempt tends to range from 5 to 8). K-means clustering is run with non-dichotomized data in this analysis. This is done because k-means is usually run with continuous data, and part of the goal of this work is to compare results from LCA and LTA using dichotomous data with more popular clustering methods.

The second clustering method used for comparison is hierarchical clustering. Hierarchical clustering starts with each observation in its own cluster. Clusters that are the most similar (e.g. have the smallest distance between them) are then joined together over and over until there are only a couple large clusters remaining. Here, Euclidean distance is used as the distance measure and average linkage is used as the linkage function. For more information on hierarchical clustering see e.g. Hierarchical Clustering by Nielsen from Undergraduate Topics in Computer Science. One advantage of hierarchical clustering is the dendrogram, or tree graph, it produces. Unlike other clustering methods that only show the end product clusters, this dendrogram shows the user exactly what process the method went through to sort its observations into the best-fitting clusters. This is very useful for our analysis because it allows us to gauge exactly how similar different pairs or groups of quarterbacks are to each other. Hierarchical clustering was performed with the dichotomized data to maintain similarity with data used for the main analysis and draw additional comparisons to k-means clustering. Since k-means clustering and hierarchical clustering are both single-time methods that are not inherently longitudinal, they must be performed once for each year in a similar fashion to standard LCA. These results will be primarily compared with the LCA results for this reason.

V. Data and Output Interpretation

Now that we have performed the RMLCA and PROC LTA analyses, along with our other clustering methods, we can begin to interpret our results. The first steps here are to look at the output of each individual analysis procedure by itself before we begin comparing outputs to each other. We will begin this process by looking at each individual

LCA model generated in each year. First, as mentioned above, we look at how many quarterbacks fell into each latent class. This is a vital step in the process. A latent class with only one or two quarterbacks in it likely would be a sign of some strong outliers in performance for that year in some way, while if the latent classes were all of a decent size it would indicate that it is more likely that the quarterbacks fall into more interpretable groups. The next thing to evaluate in each model are the item response probabilities for each latent class. In LCA, item response probabilities represent the probability that a randomly chosen quarterback in the given latent class will be above median in a given observed variable. For instance, if the item response probability of above median for the variable Y/C in latent class 2 is 0.452, this means that 45.2% of the quarterbacks falling into latent class 2 in the LCA analysis are above median in yards per completion. We evaluate item response probabilities both graphically and through numerical tables.

Once we have examined each LCA analysis individually, we evaluate the RMLCA analysis as a whole by investigating how quarterbacks move between latent classes each year. In order to do this, as mentioned above we must decide which latent classes in each year correspond to each other latent classes in other years. For instance, we might decide that the 'high achieving' latent class of quarterbacks is latent class 1 in 2012, but latent class 2 in 2013 and latent class 1 again in 2014. These decisions are made by looking at the item response probabilities and grouping together latent classes that are the most similar in terms of their item response probabilities overall. Once this is done, we can construct transition matrices to evaluate how quarterbacks move between latent classes over time and how consistent the created classes are. We construct a transition matrix for each pair of adjacent years (e.g. 2012-2013, 2013-2014, etc.) and from the beginning of the analysis period to the end to gauge movement across the entire observed time span.

Next, we look at the results of the LTA analysis done in SAS. Once we export the data from SAS into R, we can conduct a fairly similar analysis to what we did with the RMLCA analysis. We start by evaluating the size of each latent class at each time, then move on to examining the item response probabilities. We do not have to create transition matrices ourselves, however, as the LTA analysis does this for us. One thing to keep in mind is that we need to be very careful with what each LTA-created latent class actually represents while examining these transition matrices. Although one might assume that since LTA creates all the latent classes at once they would represent the same constructs consistently over time, but careful examination of item response probabilities may reveal that this is not actually the case.

Our next step is to interpret the results of the k-means and hierarchical clustering. Like with the results of our LTA and LCA analysis, one thing we need to do here is subjectively determining what each cluster in each analysis represents. For instance, for each clustering result we try to identify a ‘high performance’ cluster for quarterbacks who had good statistics in most of the observed variables we are measuring, and a ‘low performance’ cluster for quarterbacks who did poorly in all the areas we examined. Once we have identified these clusters, we can compare how each clustering method places each quarterback. For instance, it is important to investigate if there are, for instance, certain quarterbacks who are classified as ‘high performance’ by LCA and ‘low performance’ by k-means clustering. We can also use the dendrogram produced by hierarchical clustering to investigate how similar certain pairs or groupings of quarterbacks are to each other.

4. Results

This chapter will lay out the full results of the analysis. I will discuss findings from the data and the statistical analysis methods. Then I will interpret the results of the different clustering methods used.

I. Univariate and Bivariate Analyses

Data before dichotomization can be found in Tables A1-A4. Univariate analysis of this data generally indicates that most of the variables examined in the analysis are roughly normally distributed. There is some variation and some of the variables have some slight skewness in some years, but there are no cases where the variables have severely skewed distributions. CMP% tends to be fairly normally distributed, although it has a slightly more uniform distribution in 2015. CMP% also increases over time, with its median increasing from 62.35 in 2012 to 64.40 in 2015. Y/A stands out in that it is the most likely variable examined to have positive skew, as it has more positive outliers than any of the other examined variables. Unlike CMP%, Y/A does not show any appreciable increase from 2012 to 2015 among the quarterbacks examined. Y/C has a fairly similar distribution to Y/A in that it is slightly positively skewed for the years 2012 to 2015, although it is less skewed than Y/A. Another similarity between Y/C and Y/A is that neither increase appreciably over time, with both the mean and median Y/C among the examined quarterbacks actually being lower in 2015 than in 2012. TD% is quite normally distributed overall, certainly as much as we could reasonably expect with our relatively small sample size. TD% increases steadily from 2012 to 2014, but then drops back down to 2012 levels in 2015. Int% is also normally distributed overall for most years, although there are some positive outliers in 2015. There is also no notable change over time in Int%, with the median barely varying over time and the mean increasing a bit only because of the aforementioned outliers. The means of each variable over time are shown in Table 2 here.

Table 2: Variable means, 2012-2015

Variable	2012	2013	2014	2015
Cmp%	62.42	62.85	64.17	63.45
Y/C	7.36	7.42	7.44	7.36
Y/A	11.82	11.81	11.59	11.61
TD%	4.77	5.11	5.09	4.77
Int%	2.38	2.50	2.23	2.53

Overall, the findings for our univariate analysis are largely unsurprising but encouraging for our analysis. At first glance, one might think that football statistics like those examined here might be positively skewed because there should be a small number of elite players who are much more successful than any others. Another interesting feature of the univariate analysis results is that none of the variables changed appreciably over time. The NFL has become a more pass-focused league over time, even over as short a time period as the one we are examining; for instance, the average NFL team completed 35.7 passes per game in 2015 as compared to 34.7 in 2012, which is not a large difference but is clearly an increase. Because NFL teams are passing more, we might expect the profiles of passing statistics to change as time passes. However, our univariate analysis indicates that this is generally not the case.

Although the methods we are using do not require the normality assumption to be met, there are still reasons to perform univariate analysis. First, we need to keep in mind that we are losing a lot of potential information by dichotomizing this data. This is particularly noteworthy if there are outliers present, since dichotomizing the data means that a huge positive outlier will receive the same value as a data point that is just above the median. By finding that there are few outliers, we are showing that we are losing relatively less information by dichotomizing the data. The second reason to perform this univariate analysis is to examine the measurement invariance assumption of LTA. If the structures of the variables change drastically over time, it could indicate that the constructs being measured in 2012 may not be exactly the same as what is being measured in 2015. However, since the overall profiles of the variables do not vary much over time (see Table 2), we can proceed confidently.

Bivariate analysis of potential correlations between pairs of variables indicates that there are not strong correlations between any pairs of observed variables. The scatter plots examining potential correlations between pairs of variables did not yield any strong correlations, with almost none of the correlation coefficient magnitudes being above 0.7, which is one commonly accepted threshold for a strong linear correlation (Mukaka 2012). In particular, Y/A and Y/C did not have tremendously strong correlations as one might expect, with their correlation only rising above 0.7 in one of the four years. Although this might seem alarmingly high, and probably would be if we had this correlation every year with a large sample size, we must keep in mind that our sample size is fairly small and so if there is not a clear and consistently obvious correlation between two variables we should be safe to proceed. Overall the bivariate analysis yielded very positive signs for our research, as it indicated that we should not have any major collinearity issues. We look for collinearity issues because they might indicate we are using unnecessary variables in the analysis. One important goal of this work is to use the minimum possible amount of variables for the best possible result, and strong collinearity between a pair of variables could mean that we would be better off removing one of them altogether; fortunately, bivariate analysis shows that issues of this type are not present.

II. Number of Clusters Used

Taking into account goodness-of-fit statistics and interpretability, we decided to use three clusters for our analysis. The primary source for our goodness-of-fit statistics was the LTA analysis performed in SAS, since this is the only single procedure used in our analysis that allows us to take into account all of the longitudinal data in one process. AIC and BIC increased as the number of clusters used for the LTA analysis increased (see Table 3). Since lower values of AIC and BIC indicate a better model fit, this means that the smaller the number of clusters used, the better the goodness-of-fit statistics. Taking this into account, the truly optimal number of clusters would be two solely considering the goodness-of-fit statistics. However, for interpretability purposes using only two clusters would not give us our desired amount of insight into how these quarterbacks performed, since it would strictly be separating quarterbacks with more above-median statistics overall from those with less. Adding one more cluster to bring the total number to three allows for much deeper subjective interpretation of what the clusters signify and creates more interesting patterns of quarterbacks moving between clusters over time. For these reasons, we decide to use three clusters for our LTA and RMLCA analysis, along with our other clustering methods.

Table 3: Goodness-of-Fit Statistics for Different Cluster Numbers

# of Clusters	Log Likelihood	AIC	BIC
2	-263.18	484.35	535.63
3	-233.92	491.83	579.12
4	-212.28	526.55	656.39
5	-190.92	573.84	752.77
6	-172.89	639.77	874.34
7	-161.84	731.67	1028.44

III. Interpretation of RMLCA Analysis

To perform the RMLCA and LTA analyses, we first dichotomize our data. This dichotomized data can be found in Tables A5-A8.

Using three clusters, we perform our RMLCA analysis by performing a separate LCA procedure on the data for every year from 2012 to 2015. After running the procedure, our first step is to check the posterior probabilities. As mentioned above, the posterior probabilities of an LCA or LTA model indicate how confident the model is that it is classifying the observation correctly; low posterior probabilities would indicate that the model is basically guessing how to cluster those corresponding observations, which would be a very troubling sign for the validity of our model. Fortunately, the posterior probabilities in this instance are very high. In all four years, every quarterback examined has a posterior probability basically equal to 1. This indicates that the model believes there is clear separation between latent classes. Since we are trying to search for significant differences in skill between quarterbacks, the knowledge that the model believes there are clear-cut differences in an underlying latent variable is very encouraging. We also briefly looked at cluster sizes (see Table 4). If there were any latent classes containing only one quarterback, for instance, that might be cause for concern since it would indicate that the clustering method did not really know what to do with that player. Fortunately, our smallest latent class size at any time was 5, which considering we grouped 22 quarterbacks into 3 clusters is not a cause for concern.

Table 4: RMLCA Raw Class Sizes

Year	LC 1	LC 2	LC 3
2012	6	6	10
2013	5	10	7
2014	6	7	9
2015	6	10	6

Our next step is to look at the item response probabilities for each latent class in each year, shown in Tables 13-16. By doing this, we will be able to characterize each latent class; this will give us the ability to more accurately evaluate how quarterbacks may change in skill over time.

Table 5: 2012 RMLCA Item Response Probabilities

	Cmp% Low	Cmp% High	Y/A Low	Y/A High	Y/C Low	Y/C High	TD% Low	TD% High	Int% High	Int% Low
class 1:	1.000	0.000	0.844	0.156	0.000	1.000	1.000	0.000	0.831	0.169
class 2:	0.667	0.333	1.000	0.000	1.000	0.000	0.500	0.500	0.500	0.500
class 3:	0.107	0.893	0.000	1.000	0.298	0.702	0.206	0.794	0.405	0.595

Table 6: 2013 RMLCA Item Response Probabilities

	Cmp% Low	Cmp% High	Y/A Low	Y/A High	Y/C Low	Y/C High	TD% Low	TD% High	Int% High	Int% Low
class 1:	0.400	0.600	0.000	1.000	0.200	0.800	0.600	0.400	1.000	0.000
class 2:	0.800	0.200	1.000	0.000	0.900	0.100	0.800	0.200	0.600	0.400
class 3:	0.143	0.857	0.000	1.000	0.143	0.857	0.000	1.000	0.000	1.000

Table 7: 2014 RMLCA Item Response Probabilities

	Cmp% Low	Cmp% High	Y/A Low	Y/A High	Y/C Low	Y/C High	TD% Low	TD% High	Int% High	Int% Low
class 1:	1.000	0.000	0.833	0.167	0.000	1.000	1.000	0.000	0.500	0.500

class 2:	0.429	0.571	0.000	1.000	0.000	1.000	0.000	1.000	0.571	0.429
class 3:	0.222	0.778	0.556	0.444	1.000	0.000	0.556	0.444	0.556	0.444

Table 8: 2015 RMLCA Item Response Probabilities

	Cmp%	Cmp%	Y/A	Y/A	Y/C	Y/C	TD%	TD%	Int%	Int%
	Low	High	Low	High	Low	High	Low	High	High	Low
class 1:	0.000	1.000	0.157	0.843	0.663	0.337	1.000	0.000	0.765	0.235
class 2:	0.795	0.205	1.000	0.000	0.603	0.397	0.503	0.497	0.642	0.358
class 3:	0.333	0.667	0.000	1.000	0.167	0.833	0.000	1.000	0.000	1.000

In 2012, latent class (LC) 1 tended to have above-median performance in Y/C and below-median performance in Cmp%, Y/A, TD%, and Int% (see Table 5). Since this LC had below-median performance in four out of the five observed variables, it can be classified as a “low performance” latent class. LC 2 was evenly split between above- and below-median in TD% and Int%, and majority below median in Cmp%, Y/A, and Y/C. Since this latent class had below median performance in a majority of the observed variables and did not excel in any observed variable, it can also be classified as a “low performance” latent class. The majority of the quarterbacks comprising LC 3 were above median in every single observed variable, so we can clearly classify it as a “high performance” LC. Overall, in 2012 LC 3 was the clear high performance cluster, with LCs 1 and 2 generally having profiles of low overall performance (see Figure 2).

These latent classes are visualized below in Figures 2 through 5 in graphics I will refer to as ‘latent class graphs’. These graphs show what proportion of each latent class is above median in each variable. The higher each orange bar is, the more quarterbacks in that latent class are above median. For instance, a column that is completely orange in TD% where every quarterback in that latent class was above median in touchdown percentage.

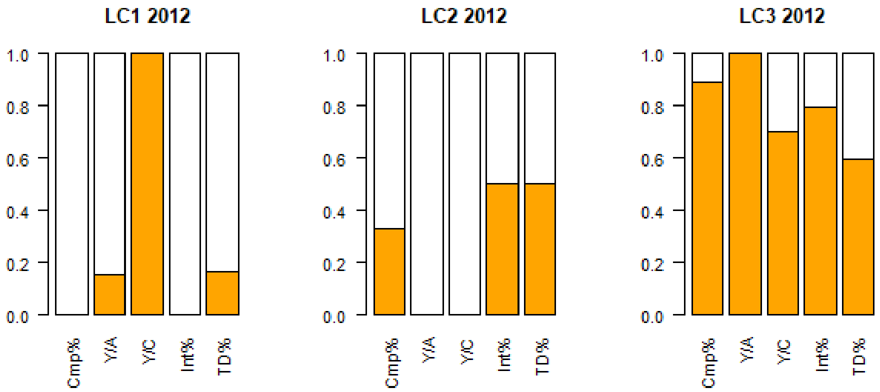

Figure 2. Latent Class Characteristics, 2012

In 2013, LC 1 was typically above median in Y/A and Y/C, worse than median in Int%, and varied with regards to Cmp% and TD%, so it could be classified as a ‘mixed performance’ LC. LC 2 was below median in every statistic besides Int%, which it was mixed in, so it could be classified as a ‘low performance’ LC. LC 3 was above median in every statistic, so clearly it could be classified as ‘high performance’. Overall, in 2013 LC 3 was the high performance cluster, with LC 1 being mixed and LC 2 generally showing low performance (see Figure 3).

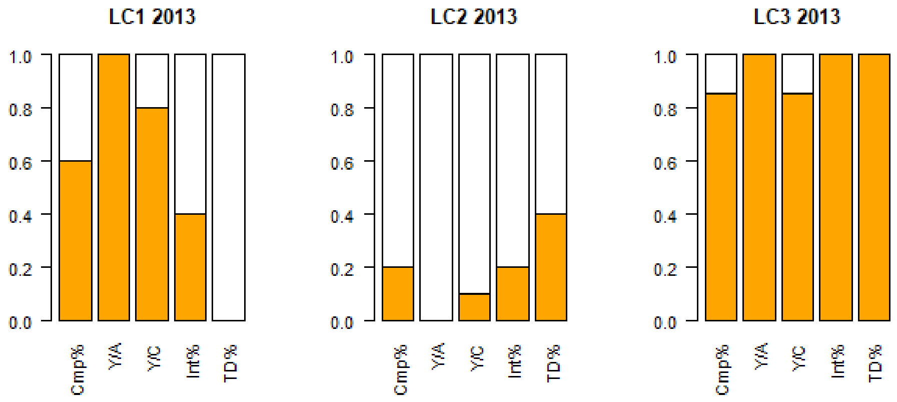

Figure 3. Latent Class Characteristics, 2013

In 2014, LC 1 was above median in Y/C, mixed in Int%, and below median in Cmp%, Y/A, and TD%, so it could be classified as a ‘low performance’ LC. LC 2 was above median in Y/A, Y/C, and TD% while being mixed in Cmp% and Int%, so it could be classified as a ‘high performance’ LC (although not overwhelmingly so). LC 3 was above median in Cmp%, below median in Y/C, and mixed in Y/A, TD%, and Int%, making it very much a ‘mixed performance’ LC. Overall, in 2014 LC 2 was the high performance cluster with LC 3 being mixed and LC 1 having overall low performance. However, it should be noted that 2014’s LC 2 was not as overwhelmingly high-performance as the designated high performance LCs from other years (see Figure 4).

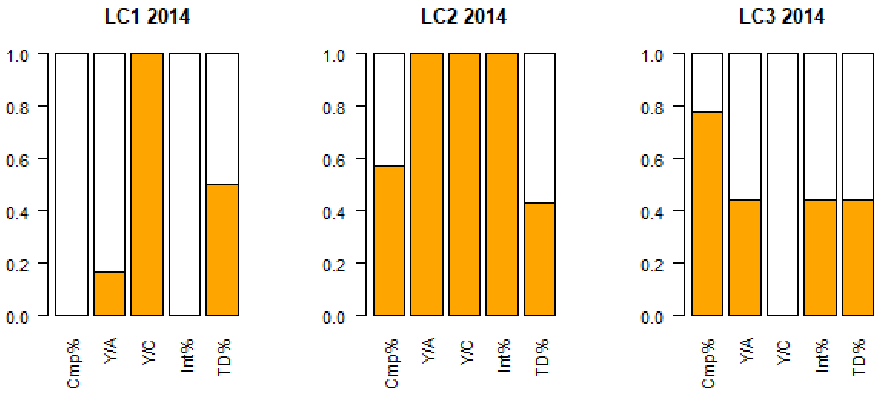

Figure 4. Latent Class Characteristics, 2014

In 2015, LC 1 was above median in Cmp% and Y/A while being below median in Y/C, TD%, and Int%, classifying it as a 'mixed performance' LC. LC 2 was mixed in TD% and below median in Cmp%, Y/A, Y/C, and Int%, making it clearly a 'low performance' LC. LC 3 was above median in every measured statistic, so it clearly falls into the 'high performance' category. Overall, in 2015 it is very clear that LC 3 corresponded with high performance, LC 1 with mixed performance, and LC 2 with low performance (see Figure 5).

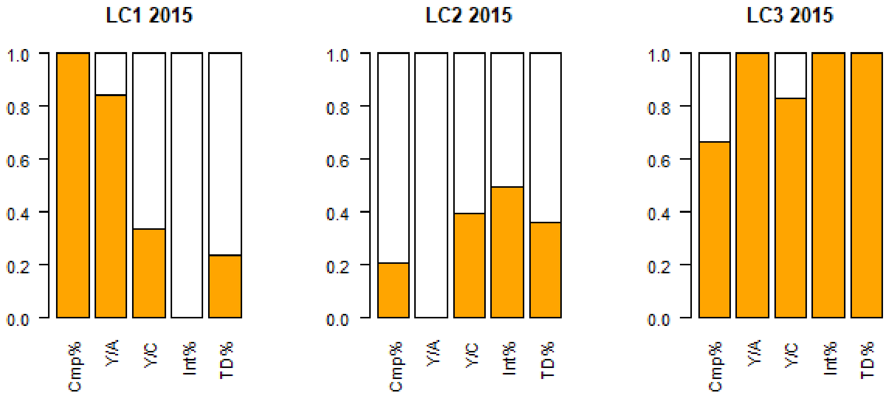

Figure 5. Latent Class Characteristics, 2015

We have now characterized every latent class produced by the RMLCA analysis for the years 2012 to 2015 (see Table 9). In general, LC 3 indicates high performance for every year besides 2014, where high performance is indicated by LC 2. The rest of the LCs are split between low and mixed performance, with 2012 being a notable exception in that it has two low performance clusters and no mixed performance cluster.

Table 9: RMLCA Latent Class Profiles

	2012	2013	2014	2015
LC 1	Low	Mixed	Low	Mixed
LC 2	Low	Low	High	Low
LC 3	High	High	Mixed	High

Table 10: RMLCA Clustering Results

Name	2012	2013	2014	2015
Brady	3	2	3	3
Brees	3	3	3	3
Cutler	1	1	3	1
Dalton	2	1	3	3
E. Manning	3	2	2	2
Fitzpatrick	2	2	2	2
Flacco	1	2	1	1
Foles	2	3	1	2

Kaepernick	3	3	1	2
Luck	1	2	2	2
Newton	1	2	1	3
P. Manning	3	3	2	2
Palmer	1	1	3	3
Rivers	2	3	3	2
Rodgers	3	3	2	2
Roethlisberger	2	1	2	1
Romo	3	2	2	1
Ryan	3	2	3	1
Smith	3	2	3	1
Stafford	2	1	1	2
Tannehill	1	2	3	2
Wilson	3	3	1	3

Now that we have characterized every latent class, we can begin to evaluate how the quarterbacks in our sample performed over time (see Table 10). Our first goal is to investigate and identify the quarterbacks who performed exceptionally well over time. Interestingly, none of the 22 quarterbacks in the sample fell into the high performance latent class all four years. This indicates that while some quarterbacks are more skilled than others, it is incredibly difficult for a quarterback to play at an elite level for several years without ever dropping in performance. Four quarterbacks fell into the high performance cluster three out of the four years: Drew Brees, Peyton Manning, Aaron Rodgers, and Russell Wilson. These quarterbacks are all recognized as some of the best in the NFL. Peyton Manning won the NFL's Most Valuable Player award in 2013, while Aaron Rodgers won it in 2014 (*AP NFL MVP Winners*). Some quarterbacks recognized as being among the best in football did not make it onto this list, such as Tom Brady and Matt Ryan (Clayton 2013, Sando 2014). However, it does appear that the quarterbacks recognized as being highly skilled by the RMLCA procedure are in fact generally characterized as being very good quarterbacks, which is strong evidence in favor of the usefulness and validity of our research.

After identifying the quarterbacks who performed exceptionally well, our next step is to attempt to identify quarterbacks who performed exceptionally poorly. Of all the quarterbacks in the sample, only four never fell into a high performance cluster: Jay Cutler, Joe Flacco, Matthew Stafford, and Ryan Tannehill. However, all four of these quarterbacks did fall into a mixed performance cluster at least once from 2012 to 2015. This indicates that although there are some quarterbacks who never perform exceptionally well, there are no quarterbacks in this sample who always perform poorly. This is likely partially a result of the fact that we are only examining quarterbacks who played most of a full season from 2012 to 2015; any quarterback who actually performed poorly all the time likely would no longer be a starting quarterback after multiple consecutive years of very poor performance. Of these four quarterbacks, Jay Cutler, Joe Flacco, and Ryan Tannehill were generally considered below-average quarterbacks for at least a large part of the time span of this analysis. Cutler was a below-average starting quarterback on the Bears, Flacco actually won the Super Bowl in the 2011 season but then dropped in performance precipitously, and Tannehill was largely considered a disappointment until his career revival with the Tennessee Titans in 2019 (*NFL QB Rankings*). The notable outlier here is Matthew Stafford, who was generally considered a very good quarterback during this time. However, he played for a consistently bad Detroit Lions team and a large part of his success came in volume statistics, so this analysis' focus on efficiency statistics will naturally work against him (*Football Encyclopedia of Players*). Just like with the quarterbacks that performed especially well, the quarterbacks that the RMLCA analysis characterized as performing especially poorly were quite reasonable and logically consistent with mainstream football opinion.

The other 14 quarterbacks in the sample all fell into a high-achieving cluster either once or twice. This indicates that these quarterbacks did not perform above average particularly consistently, but did not perform especially poorly either. Since the majority of the quarterbacks in the sample fell into this 'inconsistent' classification, this indicates that most quarterbacks do not consistently perform especially well or poorly when compared to their peers. It should again be noted that this is the result when comparing consistently starter-worthy quarterbacks to each other, however. If we were to do this analysis with all the starting quarterbacks in the NFL for a given season, it is likely that the vast majority of the quarterbacks used in this sample would fall into a high-achieving cluster.

IV. Interpretation of LTA Analysis

We performed our LTA analysis in SAS using the PROC LTA procedure. This analysis yielded a similar results structure to the RMLCA analysis procedure, and will be analyzed in a similar way. Our first step was to broadly investigate the posterior probabilities. Although the mean posterior probabilities were slightly lower than for the RMLCA procedure, they were still extremely high (mean was above 0.95) and as such were no cause for concern. There were two instances across the whole analysis of noticeably low posterior probabilities, which will be discussed later in the analysis when investigating profiles of quarterbacks and clusters.

Table 11: 2012 LTA Item Response Probabilities

	Cmp% Low	Cmp% High	Y/A Low	Y/A High	Y/C Low	Y/C High	TD% Low	TD% High	Int% High	Int% Low
class 1:	1.000	0.000	0.000	1.000	0.000	1.000	0.500	0.500	1.000	0.000
class 2:	0.000	1.000	0.182	0.818	0.455	0.545	0.182	0.818	0.364	0.636
class 3:	1.000	0.000	1.000	0.000	0.444	0.556	0.889	0.111	0.667	0.333

Table 12: 2013 LTA Item Response Probabilities

	Cmp% Low	Cmp% High	Y/A Low	Y/A High	Y/C Low	Y/C High	TD% Low	TD% High	Int% High	Int% Low
class 1:	0.400	0.600	0.000	1.000	0.200	0.800	0.600	0.400	1.000	0.000
class 2:	0.125	0.875	0.125	0.875	0.250	0.750	0.000	1.000	0.000	1.000

class 3:	0.889	0.111	1.000	0.000	0.889	0.111	0.889	0.111	0.667	0.333

Table 13: 2014 LTA Item Response Probabilities

	Cmp% Low	Cmp% High	Y/A Low	Y/A High	Y/C Low	Y/C High	TD% Low	TD% High	Int% High	Int% Low
class 1:	1.000	0.000	0.556	0.444	0.000	1.000	0.667	0.333	0.667	0.333
class 2:	0.000	1.000	0.000	1.000	0.333	0.667	0.000	1.000	0.500	0.500
class 3:	0.286	0.714	0.714	0.286	1.000	0.000	0.714	0.286	0.429	0.571

Table 14: 2015 LTA Item Response Probabilities

	Cmp% Low	Cmp% High	Y/A Low	Y/A High	Y/C Low	Y/C High	TD% Low	TD% High	Int% High	Int% Low
class 1:	0.000	1.000	0.000	1.000	0.569	0.431	0.708	0.292	0.566	0.434
class 2:	0.752	0.248	0.502	0.498	0.000	1.000	0.129	0.871	0.379	0.621
class 3:	0.571	0.429	1.000	0.000	1.000	0.000	0.714	0.286	0.571	0.429

Just as with RMLCA, our next step is to investigate the item response probabilities of each latent status at each time (see Tables 11-14) and characterize them to enable us to make meaningful statements about the quarterbacks in each cluster. These will also be visualized with latent class graphs, just as with the RMLCA results (see Figures 6-9).

In 2012, latent status (LS) 1 was above median in Y/A, mixed in TD%, and below median in Cmp%, Y/C, and Int%, which would make it a mixed performance LS. LS 2 was mixed in Y/C and above median in Cmp%, Y/A, TD%, and Int%, making it clearly a high performance LS. LS 3 was mixed in Y/C and below median in Cmp%, Y/A, TD%, and Int%, making it a low performance LS. In 2012, LS 2 was designated as high performance, LS 1 as mixed performance, and LS 3 as low performance.

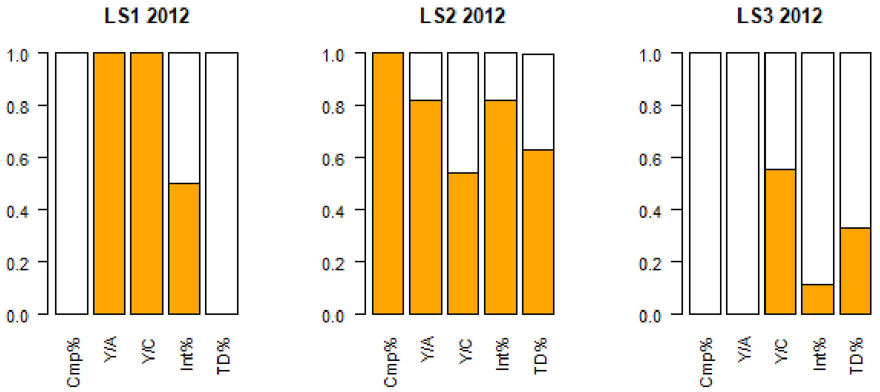

Figure 6. 2012 LTA Latent Status Characteristics

In 2013, LS 1 was above median in Cmp%, Y/A, and Y/C while being below median in TD% and Int%, meaning it would most likely be classified as a mixed performance class. Interestingly, this classification is the case despite LS 1 not actually being mixed in any of the observed variables. LS 2 was above median in all five variables, making it obviously a high performance LS. LS 3 was below median in all five variables, so clearly it can be classified as a low performance LS. Just like in 2012, in 2013 LS 2 was designated as high performance, LS 1 as mixed performance, and LS 3 as low performance.

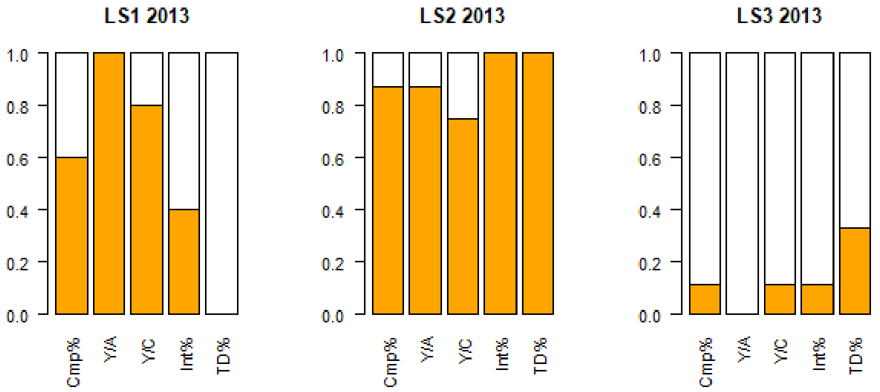

Figure 7. 2013 LTA Latent Status Characteristics

In 2014, LS 1 was above median in Y/C, mixed in Y/A, and below median in Cmp%, TD%, and Int%, meaning it would most likely be classified as a mixed performance class. LS 2 was mixed in Int% and above median in Cmp%, Y/A, Y/C, and TD%, meaning it could be classified as a high performance LS. LS 3 was above median in Cmp%, mixed in Int%, and below median in Y/A, Y/C, and TD%. 2014 stands out from the other years in the analysis in that it does not have a clear low performance cluster. Instead, LS 2 is once again the high performance cluster, with LS 1 and 3 both falling into the mixed classification.

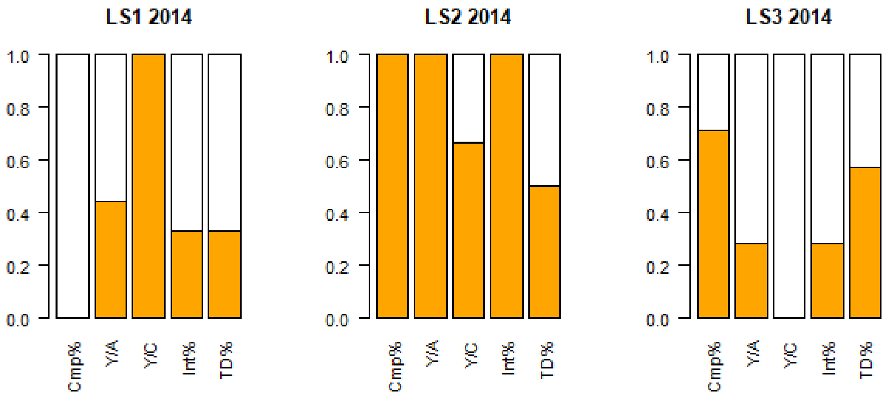

Figure 8. 2014 LTA Latent Status Characteristics

In 2015, LS 1 was above median in Cmp% and Y/A, mixed in Y/C and Int%, and below median in TD%, making it a mixed performance LS. LS 2 was above median in Y/C, TD%, and Int%, mixed in Y/A, and below median in Cmp%. LS 2 in 2015 is probably the most difficult cluster to characterize in the whole analysis. It does not quite fit the typical profile of a high performance cluster, which typically have four or five variables above median. On the other hand, having three above-median variables and one mixed makes it more high performing than the typical profile of a mixed performance cluster, which generally have multiple mixed variables and one or two variables both above and below median. Overall, we classify this LS as a high performance latent status. LS 3 is mixed in Cmp% and Int and below median in Y/A, Y/C, and TD%, making it a clear low performance cluster. Overall, in 2015 LS 2 was again classified as high performance, LS 1 as mixed performance, and LS 3 low performance.

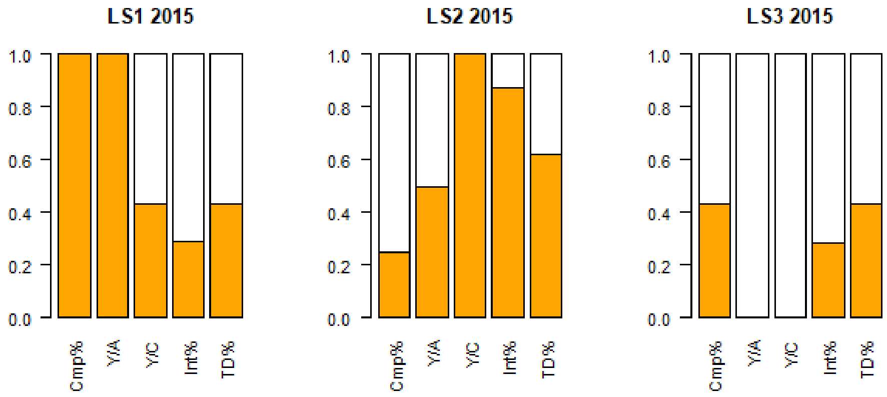

Figure 9. 2015 LTA Latent Status Characteristics

Now that we have characterized all of the clusters in our LTA analysis (see Table 15), we can again begin to profile different quarterbacks longitudinally (see Table 16). Just like with RMLCA, our first step is to try to find quarterbacks who performed exceptionally well. Again, we saw no quarterbacks fall into the high performance cluster (always LS 2) all four years of the analysis. However, it became more common for quarterbacks to fall into the high performance LS three of the four observed years. The following six quarterbacks accomplished this: Drew Brees, Peyton Manning, Philip Rivers, Aaron Rodgers, Tony Romo, and Russell Wilson. Interestingly, Brees, Manning, Rodgers, and Wilson were also the four quarterbacks that fell into high achieving clusters three of the four observed years in the RMLCA analysis. Philip Rivers and Tony Romo are also well-respected quarterbacks who played many years for the Chargers and Cowboys respectively, making them reasonably expected members of this list (Athlon 2014). The most important takeaway from this result is that RMLCA and LTA very closely agreed on which quarterbacks were the absolute best performing of the ones in the sample.

Next, we attempt to identify the quarterbacks that performed exceptionally poorly according to the LTA analysis. In this analysis, there were four quarterbacks who never fell into the high performance latent status: Jay Cutler, Andy Dalton, Joe Flacco, and Matthew Stafford. Three of these four quarterbacks are identical to the quarterbacks that were identified as being consistently low-performing in the RMLCA analysis, with the only change being that LTA placed Ryan Tannehill in the high performance LS in 2015 and replaced him with Andy Dalton in the low-performance group. Again, a major takeaway here is that the results in RMLCA are very similar to those given by LTA.

Table 15: LTA Latent Status Profiles

	2012	2013	2014	2015
LS 1	Mixed	Mixed	Mixed	Mixed
LS 2	High	High	High	High
LS 3	Low	Low	Mixed	Low

Table 16: LTA Clustering Results

Name	2012	2013	2014	2015
Brady	2	3	3	1
Brees	2	2	2	1
Cutler	3	1	3	1
Dalton	3	1	3	1
E. Manning	1	3	1	2

Fitzpatrick	3	3	1	2
Flacco	3	3	1	3
Foles	3	2	1	3
Kaepernick	2	2	1	3
Luck	3	3	1	2
Newton	1	3	1	2
P. Manning	2	2	2	3
Palmer	3	1	3	2
Rivers	2	2	2	3
Rodgers	2	2	2	3
Roethlisberger	2	1	2	1
Romo	2	2	2	1
Ryan	2	3	3	1
Smith	2	3	3	1
Stafford	3	1	1	3
Tannehill	3	3	3	2
Wilson	2	2	1	2

A useful feature of LTA compared to RMLCA is that because it is one singular longitudinal analysis procedure, it produces a series of full transition matrices that we can analyze to track patterns in how quarterbacks move between clusters over time. We will look at each of the three transitions (2012 to 2013, 2013 to 2014, and 2014 to 2015) to see how quarterbacks move between clusters over time. We will be most interested in seeing how stable the high performance latent statuses are, since that would signify a quarterback who is truly and exceptionally skilled if he continually falls into the high-performance cluster. However, we will also point out trends in transitions between the mixed and low performance clusters. These transition matrices are visualized in Tables 17-19.

Table 17: LTA Transition Matrix, 2012-2013

			Ending Status	
		1	2	3
	1	0	0	1
Starting Status	2	0.090909	0.636364	0.272727
	3	0.444444	0.111111	0.444444

Table 18: LTA Transition Matrix, 2013-2014

			Ending Status	
		1	2	3
	1	0.2	0.2	0.6
Starting Status	2	0.375	0.625	0
	3	0.555556	0	0.444444

Table 19: LTA Transition Matrix, 2014-2015

			Ending Status	
		1	2	3
	1	0	0.555556	0.444444
Starting Status	2	0.5	0	0.5
	3	0.57515	0.42485	0

From 2012 to 2013, the high performance latent class was quite stable. 63.64% of high-performing quarterbacks in 2012 stayed in that cluster in 2013. This is very strong evidence that quarterback skill is a stable, legitimate construct; if quarterbacks did not really have skill and were just moving between clusters randomly we would expect ⅓ of the quarterbacks in the high-performance latent status in 2012 to stay that way in 2013, but instead nearly double that amount remained in the high-performance cluster. Another interesting note about the first transition matrix is that every quarterback in the mixed performance cluster in 2012 moves to the low performance cluster in 2013, but with relatively small sample size this is not necessarily unexpected because the boundaries between the mixed and low performance latent statuses tend to be fairly narrow.

From 2013 to 2014, the high performance latent class was again stable. 62.5% of high-performing quarterbacks in 2013 stayed in that cluster in 2014. This is yet more evidence that quarterback skill is a real, stable construct. One interesting thing to note about this matrix is that none of the quarterbacks in the low performance cluster in 2013 jumped to the high performance cluster in 2014. This could be seen as evidence that quarterbacks who are lacking in skill in some way cannot just improve his skill to a large extent in a short period of time, which lends further credibility to the persistence and ‘stickiness’ of quarterback skill.

The transition matrix from 2014 to 2015 is not at all intuitive and runs somewhat counter to the previous two transition matrices. There is no stability in the high performance latent status from 2014 to 2015; in fact, not a single quarterback in LS 2 in 2014 stays there in 2015. Instead, the high-performing quarterbacks in 2015 are made up of a mix of quarterbacks from the mixed-performance clusters from 2014. This could be partially explained by a combination of three factors. First, as mentioned above, in 2015 the high performance latent status is not as ‘purely’ high-performance as the high performance clusters from the other three years, being much more mixed in profile. Thus, it is not as surprising that the quarterbacks we might expect to remain in a high performance cluster would not end up in this one. The second factor is that 2014 also had a moderately unusual composition of clusters, with two mixed performance latent statuses and no low performance latent status. The third, and most important, factor comes from looking at the posterior probabilities the model assigned each quarterback for each latent status in 2015. For two quarterbacks, Tom Brady and Andy Dalton, the model gave them only a slightly higher chance of falling into LS 1 than LS 2 (0.53 for LS 1 and 0.47 for LS 2). This means that the model easily could have placed them into the high performing LS 2, and then the transition matrix would look much more similar to the ones produced for the other two time transitions. If this type of poor posterior probability was common in the analysis that would be strong evidence against the usefulness of our model, but fortunately this only occurs twice out of the 88 clustering assignments made by the LTA model. Although the transition matrix from 2014 to 2015 shows slightly less stability than the other two, it is still reasonable and interpretable.

V. Interpretation of K-Means Clustering Results

For 2012, the k-means clustering method returns one cluster of size 2, one of size 8, and one of size 12 (see Tables 20-21). This indicates that the two quarterbacks in the smallest cluster, Colin Kaepernick and Cam Newton, likely were different in some noteworthy way from the other 20 quarterbacks in the sample. The cluster centers of the results are shown in Table 21. These centers are similar to the item response probabilities from RMLCA and LTA in that they can be used to characterize the different clusters, although they are means of scaled numerical data instead of probabilities that a number will take on a certain dichotomized value (see section 2.IV.f). Immediately, we can see that cluster 3 is definitely a low performance cluster, as it is below average in every observed variable. This leaves clusters 1 and 2 as two different profiles of higher performing classes. Cluster 2 is a more traditional high-performance cluster, as it is between 0.585 and 1.013 standard deviations above the mean in every observed variable besides Y/C, which it is slightly below average in. Cluster 1, by contrast, was very high above the mean in A/C (1.495 standard deviations above) and Y/C (2.210 standard deviations above the mean); this cluster was also slightly above the mean in Int% and slightly below the mean in Cmp% and TD%. This indicates that Colin Kaepernick and Cam Newton passed for yardage incredibly effectively, but were unspectacular with regards to the other examined variables. This is likely because Kaepernick and Newton both are quarterbacks that run with the football a lot, especially to score touchdowns. Since rushing touchdowns are not factored into this analysis, quarterbacks like Kaepernick and Newton would naturally be viewed as slight outliers by the clustering method. These two players are still clearly high performing quarterbacks, but in a different way.

In general, quarterbacks that are found to be high performing by k-means clustering are also found to be high performing by LCA in 2012. Of the 8 quarterbacks in the high-performing k-means cluster, all but one were also in the high-performing latent class in 2012. This indicates that there is a strong degree of agreement between the two methods; two entirely different methods using differently modified data (k-means clustering uses scale data, while LCA uses dichotomized data) are agreeing almost completely on which quarterbacks are performing better than their peers.

Table 20: K-Means Clustering Cluster Centers, 2012

	Cmp%	Y/A	Y/C	TD%	Int%
1	-0.600	1.495	2.210	-0.398	0.736
2	0.908	0.734	-0.087	1.013	0.585
3	-0.505	-0.739	-0.310	-0.609	-0.513

Table 21: K-Means Clustering Results, 2012

name	cluster
Brady	2 (High)

Brees	2 (High)
Cutler	3 (Low)
Dalton	3 (Low)
E. Manning	3 (Low)
Fitzpatrick	3 (Low)
Flacco	3 (Low)
Foles	3 (Low)
Kaepernick	1 (High/Unique)
Luck	3 (Low)
Newton	1 (High/Unique)
P. Manning	2 (High)
Palmer	3 (Low)
Rivers	3 (Low)
Rodgers	2 (High)
Roethlisberger	2 (High)
Romo	3 (Low)
Ryan	2 (High)
Smith	2 (High)
Stafford	3 (Low)

Tannehill	3 (Low)
Wilson	2 (High)

K-means clustering was also performed for the years 2013-2015 (see tables 22-27). Based on the center profiles, we can classify cluster 2 as high performing in 2013, cluster 2 as high performing in 2014, and cluster 3 as high performing in 2015. Once again, the quarterbacks we would expect tend to fall into the high performing clusters. For example, Russell Wilson, Aaron Rodgers, and Drew Brees all fall into the high achieving cluster in two of the three years between 2013 and 2015. This shows that k-means clustering using scale data identifies high performing quarterbacks with a similar accuracy to LCA and LTA, which tend to do a good job of classifying quarterbacks in a way consistent with conventional wisdom about quarterback skill.

Table 22: K-means clustering centers, 2013

	Cmp%	Y/A	Y/C	TD%	Int%
1	-0.645617427	-0.334104375	0.045045876	-0.455037266	0.861605373
2	1.099940801	1.364493498	0.91665645	1.230660754	-0.807274516
3	-0.112726852	-0.740003087	-0.843621656	-0.469804162	-0.415828751

Table 23: K-means clustering results, 2013

name	cluster
Brady	3

Brees	2
Cutler	1
Dalton	1
E. Manning	1
Fitzpatrick	1
Flacco	1
Foles	2
Kaepernick	1
Luck	3
Newton	1
P. Manning	2
Palmer	1
Rivers	2
Rodgers	2
Roethlisberger	3
Romo	3
Ryan	3
Smith	3
Stafford	1
Tannehill	3

Wilson	2

Table 24: K-means clustering centers, 2014

	Cmp%	Y/A	Y/C	TD%	Int%
1	0.605861258	-0.531877843	-1.053948132	-0.296990586	0.296175787
2	0.480116469	1.350507219	1.136124099	1.244140494	-0.26181285
3	-0.96594861	-0.481002571	0.201855058	-0.636114784	-0.099816149

Table 25: K-means clustering results, 2014

name	cluster
Brady	1
Brees	1
Cutler	1
Dalton	1
E. Manning	3
Fitzpatrick	2
Flacco	3
Foles	3
Kaepernick	3
Luck	2
Newton	3

P. Manning	2
Palmer	3
Rivers	1
Rodgers	2
Roethlisberger	2
Romo	2
Ryan	1
Smith	1
Stafford	3
Tannehill	1
Wilson	3

Table 26: K-means clustering centers, 2015

	Cmp%	Y/A	Y/C	TD%	Int%
1	0.498112731	0.610610805	0.387239889	0.437304112	-0.409086829
2	0.190113917	-0.451698704	-0.706845413	-0.970601809	2.497614244
3	-0.979384762	-1.004934722	-0.517203961	-0.534821405	0.046128612

Table 27: K-means clustering results, 2015

name	cluster
Brady	1

Brees	1
Cutler	1
Dalton	1
E. Manning	1
Fitzpatrick	3
Flacco	3
Foles	3
Kaepernick	3
Luck	3
Newton	1
P. Manning	2
Palmer	1
Rivers	1
Rodgers	3
Roethlisberger	1
Romo	2
Ryan	1
Smith	1
Stafford	1
Tannehill	3

Wilson	1

VI. Interpretation of Hierarchical Clustering Results

Hierarchical clustering in 2012 with dichotomous data, Euclidean distance, and average linkage produces the following dendrogram:

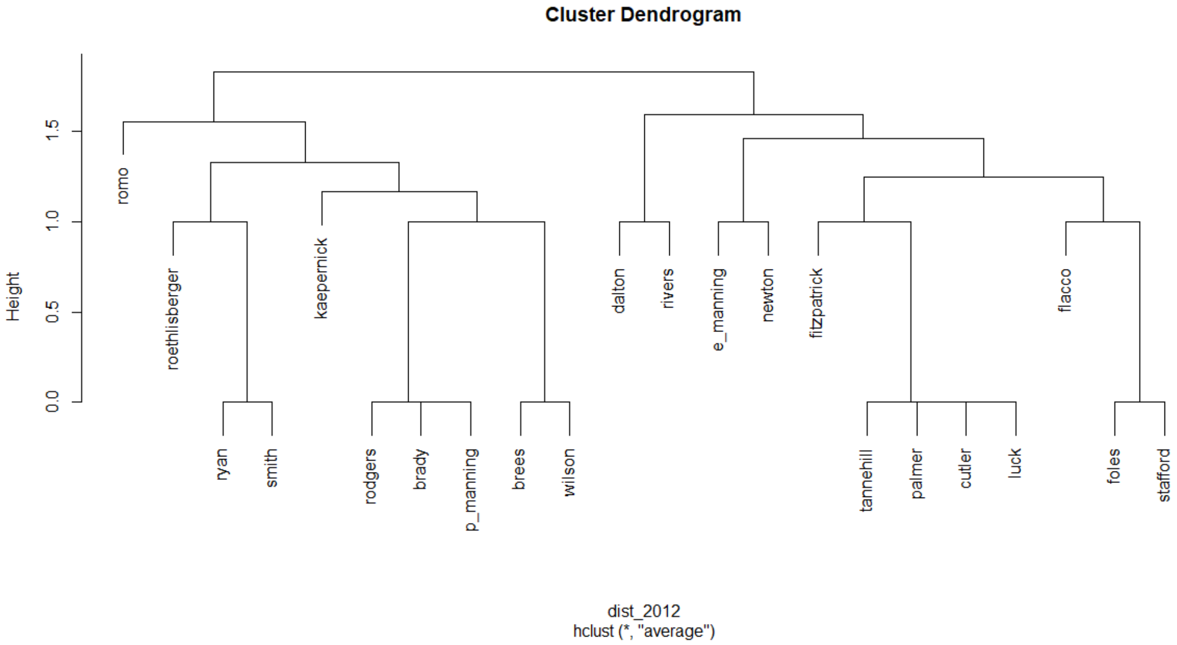

Figure 10. Hierarchical Clustering Dendrogram

When cut to produce three clusters, this dendrogram produces results that are not too far away from the results of our other clustering methods. The first and most noteworthy aspect of these results is that Tony Romo ends up in his own cluster. Looking at Tony Romo's statistics in 2012, he was above median in Cmp% and Y/A while being below median in everything else. No other quarterback shared this profile in 2012, so

Romo was different enough from the other 21 quarterbacks in the sample that he ended up in his own cluster even when separating 22 quarterbacks into 3 groups. However, the next level up on the dendrogram shows that there is relatively little distance between Romo and the rest of the large cluster on the left, showing that he is not too different from the quarterbacks in that group. With insight from our previous clustering methods, it appears that the large cluster on the left is likely a high performance cluster with the cluster to the right being a low performance cluster. The cluster on the left contains quarterbacks that have consistently been placed into high performance clusters by our other clustering methods, such as Drew Brees, Aaron Rodgers, Peyton Manning, and Russell Wilson. In contrast, the other large cluster contains quarterbacks who have generally been characterized as low performing, such as Jay Cutler and Joe Flacco. This means that hierarchical clustering also agrees with LCA, LTA, and k-means clustering on the general profile of how these quarterbacks should be grouped in 2012.

Of particular interest is the second-level cluster just to the left of the center of the dendrogram containing Rodgers, Brady, Peyton Manning, Brees, and Wilson. Four of these quarterbacks, all besides Brady, comprised the extremely high-performance group that fell into a high performance cluster three of the four years in both RMLCA and LTA, along with falling into the high-performance cluster in k-means clustering. This shows that hierarchical clustering agrees with our latent variable methods on who the absolute best quarterbacks in the sample are. It should be noted that although Tom Brady has not featured prominently in this analysis, he is widely regarded as the greatest NFL quarterback of all time and as such seeing him on an exclusive list of highly skilled quarterbacks like this one comes as no surprise (Harrison 2019).

5. Discussion

This chapter reflects on how the results of my thesis contribute to the larger literature. I also discuss both how the findings reflect on both the larger world of statistical analysis and how they contribute to the larger football discourse.

I. Comparing RMLCA and LTA Results

In general, for longitudinal clustering LTA is much more user-friendly and has more interpretable information. The fact that it runs as all one analysis offers many benefits to the user, such as only having to investigate one set of goodness-of-fit statistics. In some circumstances a user may want to use the same number of clusters for each time point in his or her analysis because he believes that there are consistent constructs that exist across time in his data set. Using RMLCA a user may encounter a situation where goodness-of-fit statistics indicate that different numbers of clusters are optimal for different times. Clearly, this problem can never occur with LTA. The way that LTA generates all of its clusters together also generally makes them easier to interpret. As seen in the Results section when I was attempting to characterize the different latent classes and statuses produced by both methods, LTA tends to yield clusters that are easier to characterize in a consistent pattern. For instance, with LTA the high-performing latent status was always cluster 2, while with RMLCA the high-performing latent status kept moving around. LTA producing its own transition matrices also helps greatly with its ease of use, since a major goal of any longitudinal cluster analysis is to examine how observations move between clusters over time. Particularly combined with the relatively consistent cluster profiles it produces, this makes LTA much easier to evaluate with regards to longitudinal transition than RMLCA.

However, RMLCA and LTA do appear to yield roughly equivalent results when broken down carefully. Our LTA results were slightly more consistent than the RMLCA results overall, as seen with the higher number of quarterbacks who were consistently clustered as either being very high performing or very low performing. However, all of the quarterbacks found to be consistently high performing by RMLCA were also found to be so by LTA, and all but one of the consistently low performing quarterbacks were the same between the two methods. In addition, generally each year had low-, mixed-, and high-performance clusters for each year; they were just more difficult and less intuitive to identify in the RMLCA results. Overall, RMLCA and LTA do yield similar results (as might be expected from two methods with the same mathematical basis), but LTA is much more intuitive and interpretable for longitudinal clustering work with this data set.

II. Comparing Latent Class Clustering Methods to Other Clustering Methods

One major goal of this research is to investigate advantages and disadvantages of LCA and LTA when compared to more traditional clustering methods. As discussed in the final part of the Results section, for the year 2012 the overall clustering results for LCA were very similar to those given by hierarchical clustering. K-means clustering also found similar results to LCA for all four years examined. All these methods agreed that quarterbacks such as Aaron Rodgers and Russell Wilson performed better than the average quarterback, while players such as Jay Cutler and Andy Dalton performed worse overall. It is important to keep in mind that this is just one analysis on a limited data set over a limited time span, and as such should not be taken as authoritative evidence on the usefulness of LCA and LTA. However, the evidence we do have strongly suggests that LCA and its siblings are just alternate methods to construct clusters.

In terms of results, there seems to be nothing extraordinary about LCA, which could be seen as both an advantage and disadvantage. This might be a disappointing finding to those who are expecting latent variable clustering methods to revolutionize cluster analysis. Although the likelihood function they use does add some diverse mathematical background to these clustering methods, it seems that they do not actually produce noticeably different results from more common distance-based procedures. On the other hand, this also provides evidence that LCA and its siblings are very valid clustering methods that provide another useful and interesting way of looking at data, and that they are not any less valid than more traditional clustering methods just because they are less well-known and established than the procedures that might traditionally come to mind when considering clustering methods.

While LCA and LTA may not be very different from more popular clustering methods in terms of results, they do have clear advantages and disadvantages in terms of the interpretability of their output and presentation. One advantage of latent variable clustering methods is that they explicitly give the user posterior probabilities. Knowing if the model is confident in a clustering assignment or not can be extremely useful if the user believes that something is somehow wrong with the cluster results, as seen in this analysis with the LTA results in 2015. Another useful feature of these clustering methods is that they are very explicit about the item response probabilities for clusters, which is very useful since it makes it much more intuitive to characterize the clusters produced by the procedure. One particular advantage of LTA is that it provides very clean and uniform transition matrices. With many longitudinal clustering methods (such as RMLCA itself), the user has to piece together transition matrices themselves by evaluating different clustering assignments. In contrast, LTA produces its own compact transition matrices, which is a highly useful and convenient feature. The major downside of using these models is that they give very little information in their output about how the likelihood functions and clusters based on them are generated; at times this can make them seem almost like a 'black box' of sorts. This runs in contrast to methods such as hierarchical clustering, which shows exactly which quarterbacks it believes to be most and least similar. Another important inherent disadvantage of these models is that they only work with categorical data, meaning scale data will have to be dichotomized to fit into these procedures. This means that there will necessarily be a large amount of information loss when using these techniques. These methods are incapable of picking up when some data is very different from average rather than just a bit different. One example of this from our analysis was how Colin Kaepernick and Cam Newton's massively high Y/A and Y/C were picked up by k-means clustering but not by LCA or LTA. Although in terms of technical results latent variable clustering methods do not appear to differ too much from more well-known clustering methods, in terms of interpretability and user-friendliness they have both clear advantages and notable disadvantages.

Overall, LCA and its sibling LTA tend to produce fairly similar results to more traditional clustering methods for this data set, but do have notable differences with regards to interpretability and usability. In conclusion, LCA and LTA are neither better nor worse than their more well-known counterparts; they are just different. I obviously do not believe that all statisticians should start exclusively using these methods. However, I do think this analysis shows that these methods are perfectly usable alternatives to more commonly accepted procedures, and are worthy of being used alongside methods such as k-means and hierarchical clustering for a different way of looking at data. LCA and LTA should not take over the proverbial world, but I strongly believe they should be brought more into the mainstream statistical dialogue and viewed as common and accepted clustering methods. This is particularly true in the case of longitudinal clustering, where LTA is the equal or superior of any other noteworthy longitudinal clustering procedure in terms of its usability and user-friendliness of its output. It should be noted, however, that these conclusions are only true for this data set. Although we can gain insight into these methods with this work, clearly this is not a global conclusion and we would have to perform this analysis on a multitude of data sets to truly understand how well these methods perform overall.

III. Insights Into NFL Quarterbacks

Another goal of this analysis was to try to provide some mathematical background to commonly accepted wisdom regarding what constitutes a ‘skilled’ NFL quarterback. In particular, one goal was to evaluate quarterbacks by comparing them with each other, rather than trying to create a new ‘objective’ statistic that inevitably would have had flaws. Overall, the results of this analysis mirrored conventional thinking about NFL quarterbacks to an almost startling degree. When evaluating the absolute best quarterbacks in the sample using latent variable and other clustering techniques, the quarterbacks that were found closely mirrored quarterbacks that are viewed to be at the top of the game in common football dialogue. This is perfectly illustrated by looking at the group of four quarterbacks who were found to be high performance in every facet by every clustering method used: Drew Brees, Peyton Manning, Aaron Rodgers, and Russell Wilson. All four of these quarterbacks are Super Bowl champions, all four have been selected to the Pro Bowl at least seven times, and all four have been named first- or second-team All-Pros by the Associated Press on at least one occasion (*Football Encyclopedia of Players*). If asked to name the very best quarterbacks in the NFL from 2012 to 2015, these four quarterbacks would appear on anyone’s list. The quarterbacks who were the least successful also mirror common opinion on which quarterbacks are better or worse than others, although for obvious reasons there are less awards for the very worst quarterbacks than the very best.

One way we can see how our results stack up against conventional football knowledge is by comparing the LTA clustering results to the NFL’s passer rating for the analyzed years. In general, the quarterbacks that fall into high performance clusters tend to have high passer ratings. In 2012, of the 11 quarterbacks with the highest passer rating who were included in this analysis, all 11 fell into the LTA high performance cluster. In 2013, the 7 quarterbacks in the high performance cluster contained 7 of the top 8 quarterbacks in passer rating included in the analysis. In 2014, 6 of the top 7 quarterbacks in passer rating fell into the high performance cluster created by LTA. 2015 was slightly less consistent with passer rating with its high performance class, but had a very welldefined low performance cluster; all four of the worst quarterbacks in the NFL in 2015 fell into LTA’s low performance latent status. This shows that latent variable clustering methods can reflect the most popular conventional football statistic quite accurately. This accuracy is in spite of how these methods only use dichotomous data, which one might expect to cause huge information loss.

The main differing point that sets this analysis apart from common football sense is that it acknowledges that even the best quarterbacks are not great all the time. This is illustrated by how there were no quarterbacks that fell into a high performance cluster all four years in the RMLCA and LTA analysis. When these elite quarterbacks have a down year, people generally do not point out this dip in performance because of their belief that the quarterback is truly skilled. In effect, oftentimes historically successful quarterbacks are protected from criticism by this very history of success. Clustering methods ignore this bias, however, and objectively identify when a generally successful quarterback has a season that is not up to his normal standards. As this analysis shows, no quarterback, no matter how skilled, is immune to having a ‘down season’. One major overall takeaway from this analysis is that even the most skilled quarterbacks are capable of having disappointing performances. Even if a quarterback is viewed as exceptionally talented, he is still capable of underperforming. Football analysis could potentially benefit from keeping this possibility in mind, rather than somewhat blindly defending generally ‘skilled’ quarterbacks as is often the case. The opposite is also the case; there were some quarterbacks who fell into low-performance clusters two or three of the four years but fell into a high performance cluster in one or two years. This shows that even quarterbacks who are generally viewed as not that skilled are capable of performing well, and that they should not be written off completely just because of a historical lack of success. Mainstream football thought does seem to do a fairly good job identifying the quarterbacks who are truly better or worse than the rest, but football analysts could stand to learn that performances do vary and are not always strictly correlated with skill.

6. Conclusion and Avenues for Future Research

This thesis provided useful insight into both NFL quarterback data and into novel latent variable clustering methods. In this way, it adds to the literature in providing a potential jumping-off point for further work in this area. However, this particular analysis is only the beginning of what would have to be done to firmly establish any of the conclusions I have drawn here. It would be irresponsible to pretend that analysis on one small data set should shift attitudes in the field of statistics drastically, but I do believe that I have laid groundwork for future progress in this area.

There are many different statistical analysis techniques and approaches that could be used to further this analysis. The most obvious next step would be to try to use factor analysis. Factor analysis is classified as a latent variable method like LCA and LTA. However, it differs in that it is not a clustering method, so it functions very differently in practice. There are also plenty of other clustering methods that could be tried to compare with the LCA and LTA results, such as mean-shift clustering and density-based clustering (Agresti 2003). It could also strengthen our evidence to repeat facets of this analysis in slightly different scenarios, such as using quarterbacks from a different span of years from 2012 to 2015. We could also strengthen the analysis by extending the data that is used. For example, we could pick a different span of years to examine or even do similar analysis with different sports. Another way to extend this analysis would be to incorporate rushing statistics in some way. Although this work did not use rushing statistics both because they would have been very difficult to include and to enable comparison with passer rating, being able to use these statistics would expand the analysis by taking into account more information.

There are plenty of ways that this work could be extended to learn more about both the statistical procedures used here and the athletic context they are examining. For instance, simulation studies could be used to learn more about the behavior of these latent variable clustering methods in different scenarios.

7. References

“Analyzing Repeated Categorical Response Data.” In *Categorical Data Analysis*, 455–90. John Wiley & Sons, Ltd, 2003. https://doi.org/10.1002/0471249688.ch11.

Pro-Football-Reference.com. “AP NFL Most Valuable Player Winners.” Accessed March 15, 2020. https://www.pro-football-reference.com/awards/ap-nfl-mvp-award.htm.

Berri, David J., and Brian Burke. “Measuring Productivity of NFL Players.” In *The Economics of the National Football League: The State of the Art*, edited by Kevin G. Quinn, 137–58. Sports Economics, Management and Policy. New York, NY: Springer, 2012. https://doi.org/10.1007/978-1-4419-6290-4_8.

Berri, David J., and Rob Simmons. “Catching a Draft: On the Process of Selecting Quarterbacks in the National Football League Amateur Draft.” *Journal of Productivity Analysis* 35, no. 1 (February 1, 2011): 37–49. https://doi.org/10.1007/s11123-009-0154-6.

ESPN.com. “Clayton: NFL Starting QB Rankings, 1-10,” August 29, 2013. https://www.espn.com/nfl/story/_/id/9592180/nfl-quarterback-rankings-john-clayton-reveals-2013-hierarchy-part-1.

Collins, Linda M., and Stephanie T. Lanza. *Latent Class and Latent Transition Analysis: With Applications in the Social, Behavioral, and Health Sciences*. 1 edition. Hoboken, N.J: Wiley, 2009.

Editors, I. T. P. "ITP Glossary: Adjusted Net Yards Per Attempt." *Inside The Pylon* (blog), January 8, 2016. http://insidethepylon.com/football-101/glossary-football-101/2016/01/08/itp-glossary-adjusted-net-yards-per-attempt/.

Elliot Harrison. "Top 25 Quarterbacks of All Time: Patriots' Tom Brady Leads List - NFL.Com." NFL.com, July 2, 2019. http://www.nfl.com/news/story/0ap3000001035041/article/top-25-quarterbacks-of-all-time-patriots-tom-Brady-leads-list.

Pro-Football-Reference.com. "Football Encyclopedia of Players." Accessed March 12, 2020. https://www.pro-football-reference.com/players/.

Franks, Alexander, Andrew Miller, Luke Bornn, and Kirk Goldsberry. "Counterpoints : Advanced Defensive Metrics for NBA," n.d.

Goldsberry, Kirk, and Eric Weiss. "The Dwight Effect : A New Ensemble of Interior Defense Analytics for the NBA," 2013.

ESPN.com. "How Is Total QBR Calculated? We Explain Our Quarterback Rating," September 8, 2016. https://www.espn.com/blog/statsinfo/post/_/id/123701/how-is-total-qbr-calculated-we-explain-our-quarterback-rating.

James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An Introduction to Statistical Learning: With Applications in R*. 1st ed. 2013, Corr. 7th printing 2017 edition. New York: Springer, 2017.

Kaufman, Leonard, and Peter J. Rousseeuw. *Finding Groups in Data: An Introduction to Cluster Analysis*. 99 edition. New York: Wiley-Interscience, 1990.

Kuzmits, Frank E., and Arthur J. Adams. "The NFL Combine: Does It Predict Performance in the National Football League?" *The Journal of Strength & Conditioning Research* 22, no. 6 (November 2008): 1721–1727. https://doi.org/10.1519/JSC.0b013e318185f09d.

MacQueen, J. "Some Methods for Classification and Analysis of Multivariate Observations." The Regents of the University of California, 1967. https://projecteuclid.org/euclid.bsmsp/1200512992.

Mukaka, MM. "A Guide to Appropriate Use of Correlation Coefficient in Medical Research." *Malawi Medical Journal : The Journal of Medical Association of Malawi* 24, no. 3 (September 2012): 69–71.

Pro-Football-Reference.com. "NFL First-Team All-Pro Selections Career Leaders." Accessed March 16, 2020. https://www.pro-football-reference.com/leaders/all_pros_first_team_career.htm.

Pro-Football-Reference.com. "NFL Pro Bowl History." Accessed March 16, 2020. https://www.pro-football-reference.com/probowl/index.htm.

NFLEmailEmailBioBioFollowFollowReporter, Mark Maske closeMark MaskeSports reporter covering the. "NFL QB Rankings from 1 to 32: Aaron Rodgers Tops Tom Brady, Peyton Manning." Washington Post. Accessed

March 15, 2020. https://www.washingtonpost.com/news/sports/wp/2014/11/13/nfl-qb-rankings-from-1-to-32-aaron-Rodgers-tops-tom-Brady-peyton-manning/.

Nielsen, Frank. “Hierarchical Clustering,” 195–211, 2016. https://doi.org/10.1007/978-3-319-21903-5_8.

“Other Mixture Models for Categorical Data.” In *Categorical Data Analysis*, 538–75. John Wiley & Sons, Ltd, 2003. https://doi.org/10.1002/0471249688.ch13.

ESPN.com. “Sando: Execs, Coaches Rank 32 NFL QBs,” June 30, 2014. https://www.espn.com/nfl/story/_/id/11156302/nfl-anonymous-league-insiders-rank-32-starting-quarterbacks-tiers.

Silver, Nate. “CARMELO NBA Player Projections.” FiveThirtyEight, July 1, 2018. https://projects.fivethirtyeight.com/carmelo/.

Sports, Athlon, 7/7/14, and 10:00 Am Edt. “2014 NFL Player Rankings: Quarterbacks.” AthlonSports.com. Accessed March 15, 2020. https://athlonsports.com/nfl/2014-nfl-player-rankings-quarterbacks.

Stimel, Derek. “A Statistical Analysis of NFL Quarterback Rating Variables.” *Journal of Quantitative Analysis in Sports* 5, no. 2 (2009). https://doi.org/10.2202/1559-0410.1166.

Wikipedia Contributors. “List of NFL Quarterbacks Who Have Posted a Perfect Passer Rating.” In *Wikipedia*, February 18, 2020.

https://en.wikipedia.org/w/index.php?title=List_of_NFL_quarterbacks_who_have_posted_a_perfect_passer_rating&oldid=941436001.

———. “Passer Rating.” In *Wikipedia*, March 4, 2020. https://en.wikipedia.org/w/index.php?title=Passer_rating&oldid=943950134.

“WRC and WRC+ | Sabermetrics Library.” Accessed March 16, 2020. https://library.fangraphs.com/offense/wrc/.

8. Appendix

Table A1: Non-Dichotomized Data, 2012

qb_names	Cmp%	Y/A	Y/C	TD%	Int%
Brady	63	7.6	12	5.3	1.3
Brees	63	7.7	12.3	6.4	2.8
Cutler	58.8	7	11.9	4.4	3.2
Dalton	62.3	6.9	11.2	5.1	3
E. Manning	59.9	7.4	12.3	4.9	2.8
Fitzpatrick	60.6	6.7	11.1	4.8	3.2
Flacco	59.7	7.2	12	4.1	1.9
Foles	60.8	6.4	10.6	2.3	1.9
Kaepernick	62.4	8.3	13.3	4.6	1.4
Luck	54.1	7	12.9	3.7	2.9
Newton	57.7	8	13.8	3.9	2.5
P. Manning	68.6	8	11.6	6.3	1.9
Palmer	61.1	7.1	11.6	3.9	2.5
Rivers	64.1	6.8	10.7	4.9	2.8
Rodgers	67.2	7.8	11.6	7.1	1.4
Roethlisberger	63.3	7.3	11.5	5.8	1.8
Romo	65.6	7.6	11.5	4.3	2.9

Ryan	68.6	7.7	11.2	5.2	2.3
Smith	70.2	8	11.4	6	2.3
Stafford	59.8	6.8	11.4	2.8	2.3
Tannehill	58.3	6.8	11.7	2.5	2.7
Wilson	64.1	7.9	12.4	6.6	2.5

Table A2: Non-Dichotomized Data, 2013

qb_names	Cmp%	Y/A	Y/C	TD%	Int%
Brady	60.5	6.9	11.4	4	1.8
Brees	68.6	7.9	11.6	6	1.8
Cutler	63.1	7.4	11.7	5.4	3.4
Dalton	61.9	7.3	11.8	5.6	3.4
E. Manning	57.5	6.9	12	3.3	4.9
Fitzpatrick	62	7	11.3	4	3.4
Flacco	59	6.4	10.8	3.1	3.6
Foles	64	9.1	14.2	8.5	0.6
Kaepernick	58.4	7.7	13.2	5	1.9
Luck	60.2	6.7	11.1	4	1.6
Newton	61.7	7.1	11.6	5.1	2.7

P. Manning	68.3	8.3	12.2	8.3	1.5
Palmer	63.3	7.5	11.8	4.2	3.8
Rivers	69.5	8.2	11.8	5.9	2
Rodgers	66.6	8.7	13.1	5.9	2.1
Roethlisberger	64.2	7.3	11.4	4.8	2.4
Romo	63.9	7.2	11.2	5.8	1.9
Ryan	67.4	6.9	10.3	4	2.6
Smith	60.6	6.5	10.8	4.5	1.4
Stafford	58.5	7.3	12.5	4.6	3
Tannehill	60.4	6.7	11	4.1	2.9
Wilson	63.1	8.2	13.1	6.4	2.2

Table A3: Non-Dichotomized Data, 2014

qb_names	Cmp%	Y/A	Y/C	TD%	Int%
Brady	64.1	7.1	11	5.7	1.5
Brees	69.2	7.5	10.9	5	2.6
Cutler	66	6.8	10.3	5	3.2
Dalton	64.2	7.1	11	4	3.5
E. Manning	63.1	7.3	11.6	5	2.3

Fitzpatrick	63.1	8	12.6	5.4	2.6
Flacco	62.1	7.2	11.6	4.9	2.2
Foles	59.8	7	11.6	4.2	3.2
Kaepernick	60.5	7	11.7	4	2.1
Luck	61.7	7.7	12.5	6.5	2.6
Newton	58.5	7	11.9	4	2.7
P. Manning	66.2	7.9	12	6.5	2.5
Palmer	62.9	7.3	11.5	4.9	1.3
Rivers	66.5	7.5	11.3	5.4	3.2
Rodgers	65.6	8.4	12.8	7.3	1
Roethlisberger	67.1	8.1	12.1	5.3	1.5
Romo	69.9	8.5	12.2	7.8	2.1
Ryan	66.1	7.5	11.3	4.5	2.2
Smith	65.3	7	10.8	3.9	1.3
Stafford	60.3	7.1	11.7	3.7	2
Tannehill	66.4	6.9	10.3	4.6	2
Wilson	63.1	7.7	12.2	4.4	1.5

Table A4: Non-Dichotomized Data, 2015

qb_names	Cmp%	Y/A	Y/C	TD%	Int%
Brady	64.4	7.6	11.9	5.8	1.1
Brees	68.3	7.8	11.4	5.1	1.8
Cutler	64.4	7.6	11.8	4.3	2.3
Dalton	66.1	8.4	12.7	6.5	1.8
E. Manning	62.6	7.2	11.5	5.7	2.3
Fitzpatrick	59.6	6.9	11.7	5.5	2.7
Flacco	64.4	6.8	10.5	3.4	2.9
Foles	56.4	6.1	10.8	2.1	3
Kaepernick	59	6.6	11.2	2.5	2
Luck	55.3	6.4	11.6	5.1	4.1
Newton	59.8	7.8	13	7.1	2
P. Manning	59.8	6.8	11.4	2.7	5.1
Palmer	63.7	8.7	13.7	6.5	2
Rivers	66.1	7.2	11	4.4	2
Rodgers	60.7	6.7	11	5.4	1.4
Roethlisberger	68	8.4	12.3	4.5	3.4
Romo	68.6	7.3	10.7	4.1	5.8
Ryan	66.3	7.5	11.3	3.4	2.6

Smith	65.3	7.4	11.4	4.3	1.5
Stafford	67.2	7.2	10.7	5.4	2.2
Tannehill	61.9	7.2	11.6	4.1	2
Wilson	68.1	8.3	12.2	7	1.7

Note: for all dichotomized data, “1” indicates below median and “2” indicates above median, except in the case of Int%, where “1” indicates above median and “2” indicates below median

Table A5: Dichotomized Quarterback Performance Statistics: 2012

qb_names	Cmp%	Y/A	Y/C	TD%	Int%
Brady	2	2	2	2	2
Brees	2	2	2	2	1
Cutler	1	1	2	1	1
Dalton	1	1	1	2	1
E. Manning	1	2	2	2	1
Fitzpatrick	1	1	1	1	1
Flacco	1	1	2	1	2
Foles	1	1	1	1	2
Kaepernick	2	2	2	1	2

Luck	1	1	2	1	1
Newton	1	2	2	1	1
P. Manning	2	2	2	2	2
Palmer	1	1	2	1	1
Rivers	2	1	1	2	1
Rodgers	2	2	2	2	2
Roethlisberger	2	1	1	2	2
Romo	2	2	1	1	1
Ryan	2	2	1	2	2
Smith	2	2	1	2	2
Stafford	1	1	1	1	2
Tannehill	1	1	2	1	1
Wilson	2	2	2	2	1

Table A6: Dichotomized Quarterback Performance Statistics: 2013

qb_names	Cmp%	Y/A	Y/C	TD%	Int%
Brady	1	1	1	1	2
Brees	2	2	1	2	2
Cutler	2	2	2	2	1
Dalton	1	2	2	2	1

E. Manning	1	1	2	1	1
Fitzpatrick	1	1	1	1	1
Flacco	1	1	1	1	1
Foles	2	2	2	2	2
Kaepernick	1	2	2	2	2
Luck	1	1	1	1	2
Newton	1	1	1	2	1
P. Manning	2	2	2	2	2
Palmer	2	2	2	1	1
Rivers	2	2	2	2	2
Rodgers	2	2	2	2	2
Roethlisberger	2	2	1	1	1
Romo	2	1	1	2	2
Ryan	2	1	1	1	1
Smith	1	1	1	1	2
Stafford	1	2	2	1	1
Tannehill	1	1	1	1	1
Wilson	2	2	2	2	2

Table A7: Dichotomized Quarterback Performance Statistics: 2014

qb_names	Cmp%	Y/A	Y/C	TD%	Int%
Brady	1	1	1	2	2
Brees	2	2	1	2	1
Cutler	2	1	1	2	1
Dalton	2	1	1	1	1
E. Manning	1	2	2	2	1
Fitzpatrick	1	2	2	2	1
Flacco	1	1	2	1	1
Foles	1	1	2	1	1
Kaepernick	1	1	2	1	2
Luck	1	2	2	2	1
Newton	1	1	2	1	1
P. Manning	2	2	2	2	1
Palmer	1	2	1	1	2
Rivers	2	2	1	2	1
Rodgers	2	2	2	2	2
Roethlisberger	2	2	2	2	2
Romo	2	2	2	2	2
Ryan	2	2	1	1	1
Smith	2	1	1	1	2

Stafford	1	1	2	1	2
Tannehill	2	1	1	1	2
Wilson	1	2	2	1	2

Table A8: Dichotomized Quarterback Performance Statistics: 2015

qb_names	Cmp%	Y/A	Y/C	TD%	Int%
Brady	2	2	2	2	2
Brees	2	2	1	2	2
Cutler	2	2	2	1	1
Dalton	2	2	2	2	2
E. Manning	1	1	2	2	1
Fitzpatrick	1	1	2	2	1
Flacco	2	1	1	1	1
Foles	1	1	1	1	1
Kaepernick	1	1	1	1	2
Luck	1	1	2	2	1
Newton	1	2	2	2	2
P. Manning	1	1	1	1	1
Palmer	1	2	2	2	2
Rivers	2	1	1	1	2

Rodgers	1	1	1	2	2
Roethlisberger	2	2	2	1	1
Romo	2	2	1	1	1
Ryan	2	2	1	1	1
Smith	2	2	1	1	2
Stafford	2	1	1	2	1
Tannehill	1	1	2	1	2
Wilson	2	2	2	2	2