Data Visualization EDA

Rahul Sagi

University of North Texas

Initial Questions:

1. Which QBs have completed the most passes from 2009-2018?

    - Transformed a 2009-2018 NFL Play-By-Play Dataset from a csv file to a Tableau workbook. Then I created a bar graph compare each QB's total completed passes from 2009-2018. Used descending sort and color to help visualize the differences in a clearer manner

2. Have QBs have the highest percentage of Total TDs from 2009-2018?

    - Created a tree map to find and compare the passers who have the highest percentage of Total Touchdowns from 2009-2018. Filtered Completed Passes to a value of at least 200 to have cleaner data.

3. Which QBs are throwing more first down passes and converting the important third downs?

    - Created a bar graph that shows the cumulative sum of all first down passes completed from 2009-2018 by each player. Filtered on Player Passer Name to only have 32 members and Completed Passes to a value of at least 200 to have cleaner data. Used descending sort and sort to help visualize the differences in a clearer manner.

- Created a Scatter Plot to see the correlation between 3rd Down Success and 3rd Down Failures. Used Size to show players who have completed the most passes. Filtered Completed Passes to a value of at least 200 to have cleaner data and used color to help visualize the differences in a clearer manner.

4. Which players have added the most Expected Points(EPA) from 2009-2018?

- Created a bar graph that shows the cumulative sum of all first down passes completed from 2009-2018 by each player. Filtered on Player Passer Name to only have 32 members and Completed Passes to a value of at least 200 to have cleaner data. Used desc. sort and color to help visualize the differences in a clearer manner.

5. Does a higher EPA average correlate to a higher Win Probability?

- Created a Scatter Plot to see the correlation between Estimated Points Added(EPA) per play and Win Probability Added(WPA) per play. Used Size to show players who have completed the most passes. Filtered Completed Passes to a value of at least 200 to have cleaner data and used color to help visualize the differences in a clearer manner.

6. Which players have the highest average EPA from 2009-2014?

- Created a bar graph that shows the average EPA for each player from completed from 2009-2014. Only used values from 2009-2014. Filtered on Player Passer Name to only have 40 members and Completed Passes to a value of at least 200 to have cleaner data. Used desc. sort and color to help see the trends better.

7. Which players have the highest average EPA from 2015-2018?

- Created a bar graph that shows the average EPA for each player from completed from 2009-2014. Only used values from 2014-2018. Filtered on Player Passer Name to only have 40 members and Completed Passes to a value of at least 200 to have cleaner data. Used descending sort and color to help visualize the differences in a clearer manner.

Dataset Information:

Where I got the Dataset:

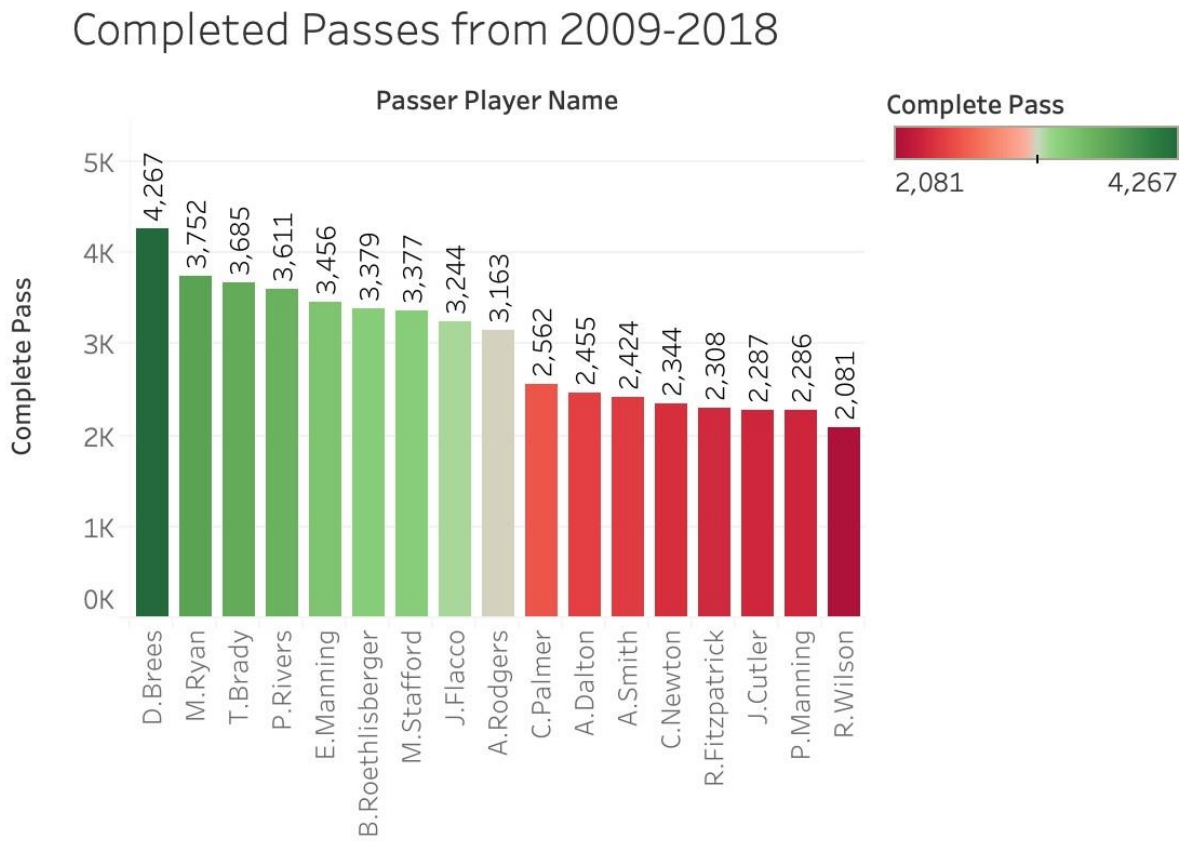
- I got my data set from Kaggle.com. The dataset was Detailed NFL Play-By-Play Data 2009-2018.

Dataset Format:

- Datasets was a csv file that I transformed into various visualizations in a Tableau Workbook

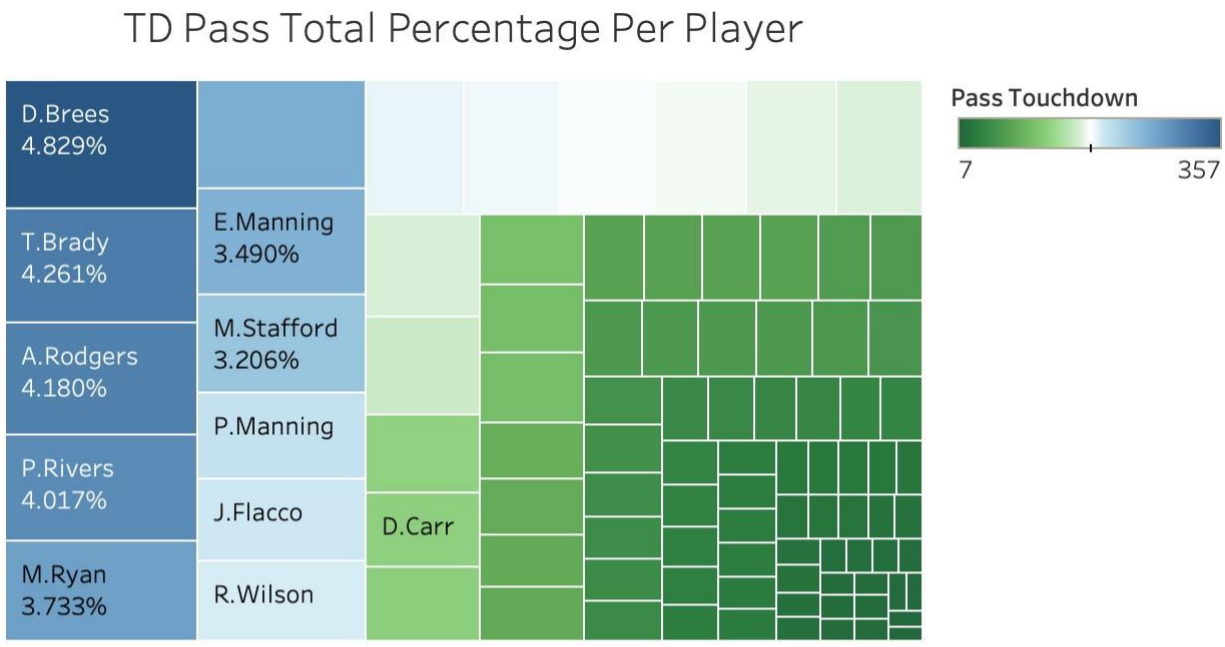
CSV File: https://www.kaggle.com/maxhorowitz/nflplaybyplay2009to2016

The NFL has slowly shifted into a passing league ever since the beginning of the 21st century. QBs have always been the one field position in the sport that influences a game's result. However, as rule changes favor the offense more year by year, passing numbers by QBs have become increasingly prolific. However, traditional statistics can't determine who the best QB is. Is there a metric or stat out there that combines all these factors to help determine the great from the mediocre? The following graphs below will try to determine who the best QB/best QBs of the past decade are as there have not been as many significant rule changes in the NFL throughout the past 10 years as they were in early 2000s.



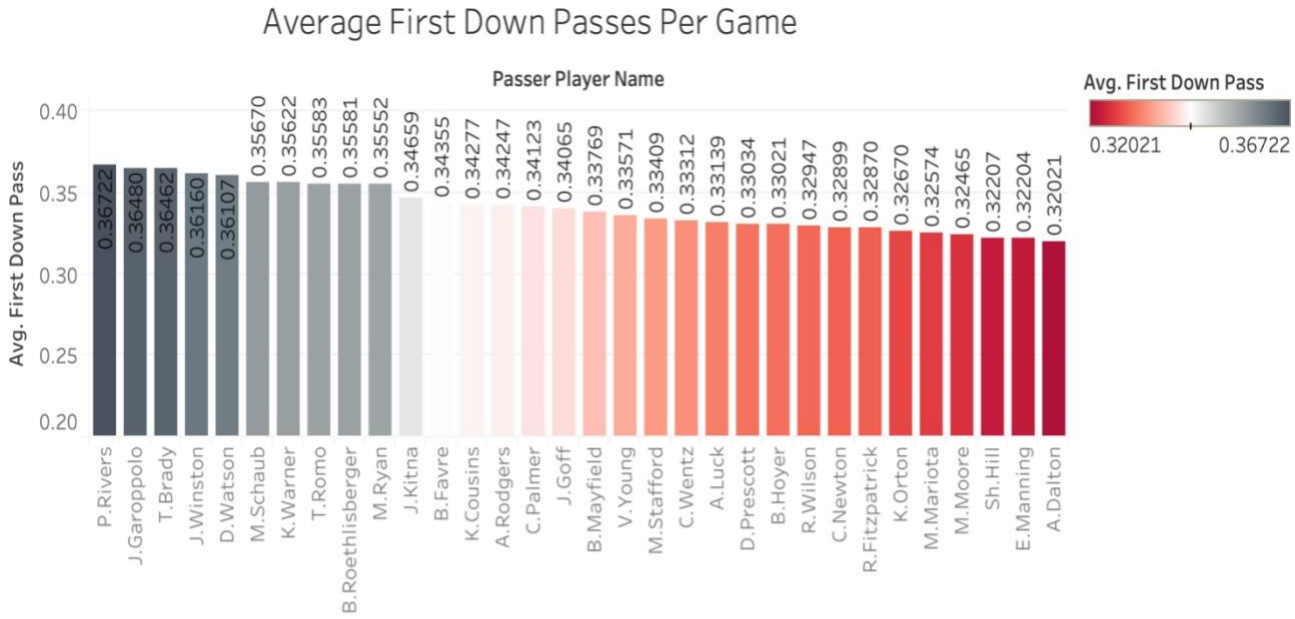
Sum of Complete Pass for each Passer Player Name. Color shows sum of Complete Pass. The marks are labeled by sum of Complete Pass. The data is filtered on sum of Comp Air Epa, which keeps non-Null values only. The view is filtered on Passer Player Name, which keeps 17 members.

Completed Passes tells us the number of completed passes a player has made from 2009-2019. I took this graph as my first measure to see who the most prolific passers from the past decade are. However, the problem with this is that there are many other metrics in judging player performance (such as TD passes) and this type of measurement is inherently biased towards players who have played throughout most of the timeframe. Due to the dataset being too large, I only took the top 17 values in this graph, and from it Drew Brees, Matt Ryan, and Tom Brady have completed the most passes. I used a bar graph here because it does a great job of comparing the number of passes completed between each QB. I used color to help clearly differentiate the max from the min. Furthermore, I filtered Passer Player Name so that only the top 17 values could be kept as there were too many players in the dataset.

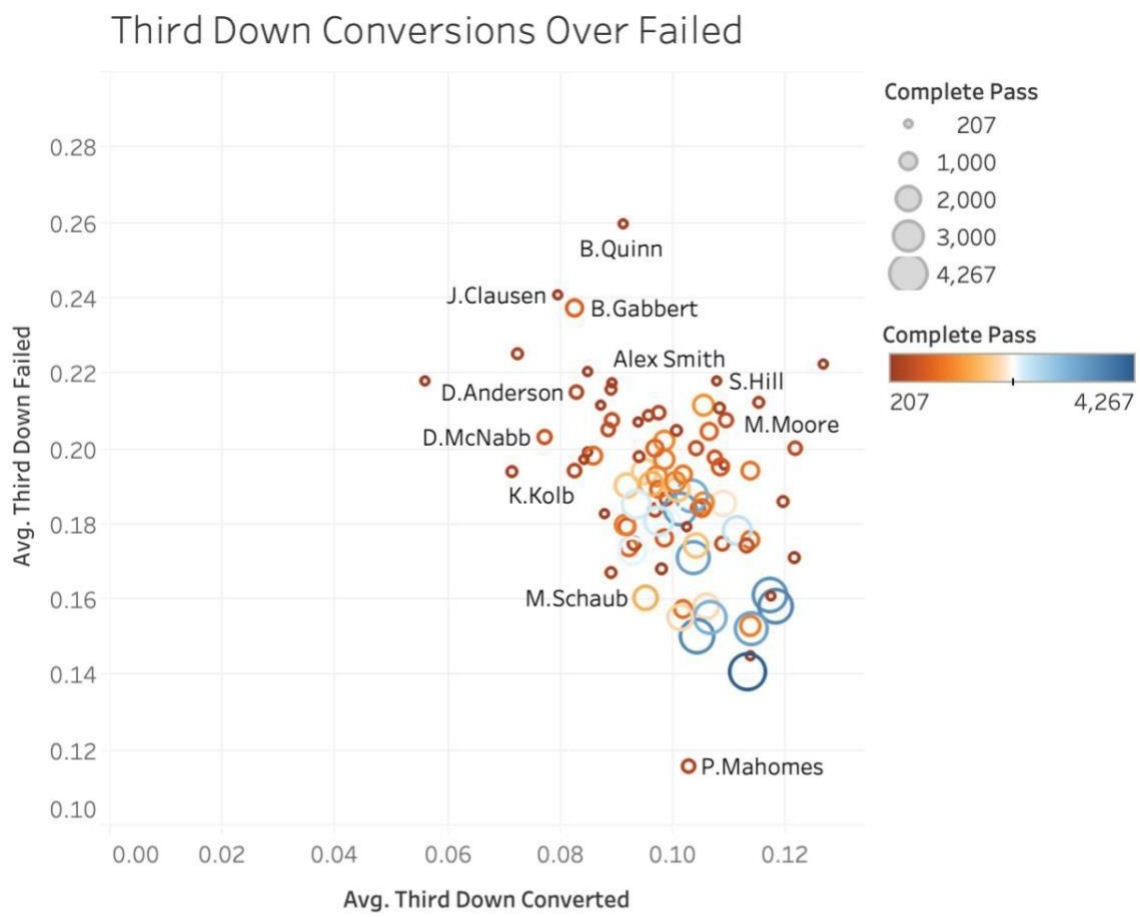


Passer Player Name and % of Total Pass Touchdown. Color shows sum of Pass Touchdown. Size shows sum of Pass Touchdown. The marks are labeled by Passer Player Name and % of Total Pass Touchdown. The data is filtered on sum of Complete Pass, which includes values greater than or equal to 200.

Next, I measured the total percentage of TD passes a player threw over these past 10 years. This is a better way of measuring how vital a QB was to his team's success because throwing TD passes puts up points on the scoreboard. The players with the highest percentage are the ones who are most likely the elite passers of the ball. From this graph, you can see that Drew Brees, Tom Brady, and Aaron Rodgers with percentages of 4.829, 4.261, and 4.180% respectively. I used a tree map here because it allows me to clearly differentiate the players who have the highest percentage of TD passes, and I used color to accentuate the differences. However, like the Completed Passes graph above, this is still a cumulative graph, and is biased towards longevity. Furthermore, this does not measure the efficiency of a player, and how much value that player adds over his contemporaries.



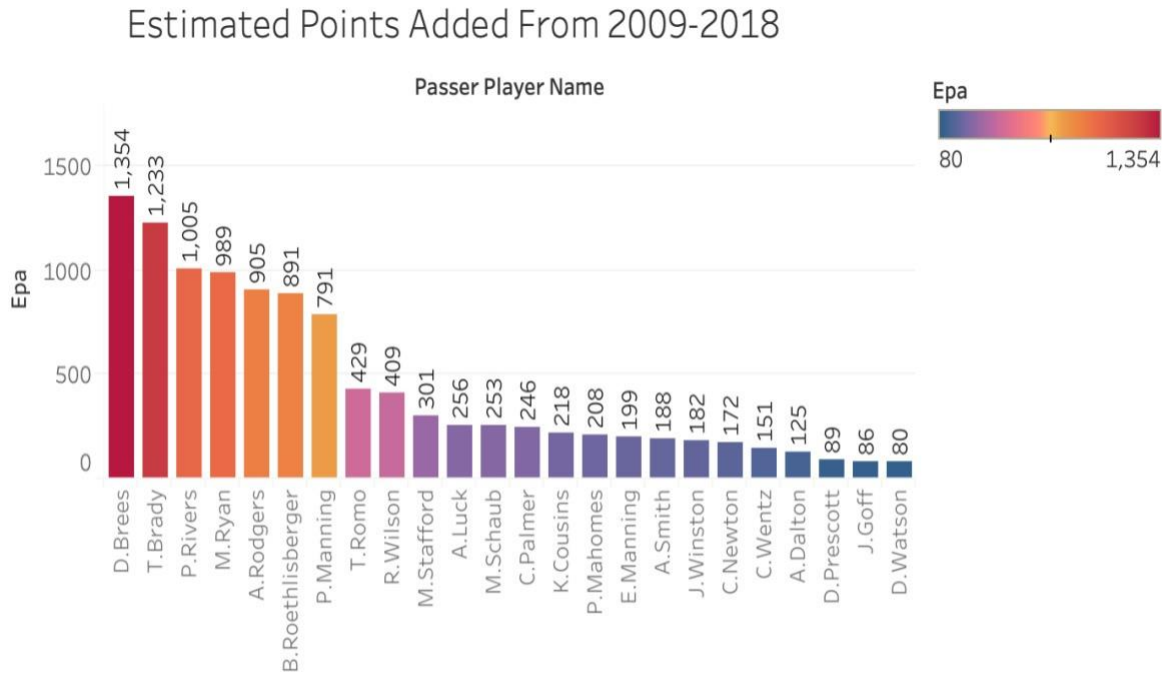
Average of First Down Pass for each Passer Player Name. Color shows average of First Down Pass. The marks are labeled by average of First Down Pass. The data is filtered on sum of Complete Pass, which ranges from 200 to 4,267. The view is filtered on Passer Player Name, which keeps 32 members.



Average of Third Down Converted vs. average of Third Down Failed. Color shows sum of Complete Pass. Size shows sum of Complete Pass. The marks are labeled by Passer Player Name. The view is filtered on sum of Complete Pass, which includes values greater than or equal to 200.

From the graphs above, there is a correlation between the number of passes completed and third down success rate. Furthermore, Tom Brady, Philip Rivers, and Jimmy Garoppolo average the most first down passes per play. The reason I picked these metrics is due to the NFL having 4 downs to get ten yards. A successful first down completion ensures that the QB's team stays on the field until the team gets a TD. Furthermore, by having a higher number of third downs completed when team defenses are fully defending against the pass(thus theoretically making it harder to pass the ball. From the scatterplot above, QBs that pass the ball more generally convert more third downs than ones who don't. However, this falls under the same

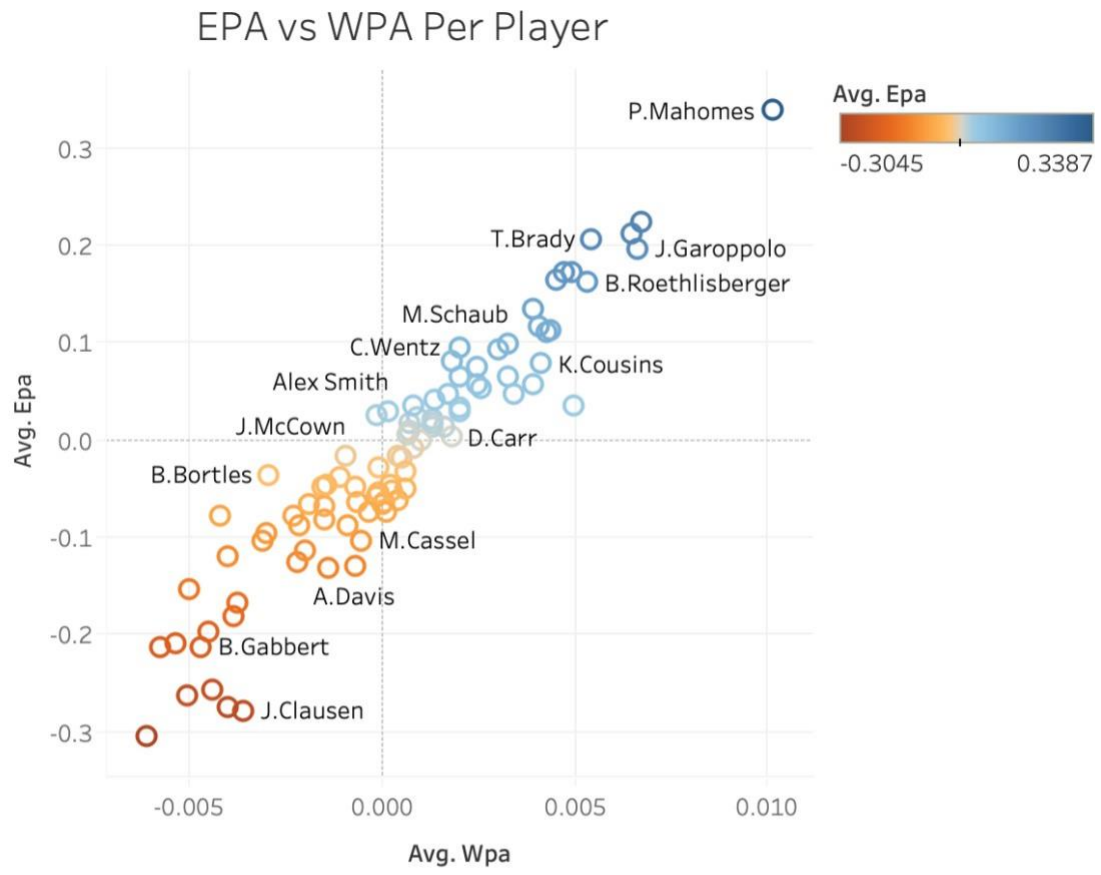
bias as the previous graphs in a sense due to players who have played the majority of the decade are favored in a sense over players who have not. I used a bar graph for Average First Down Per Pass as that helps show the players who stand out amongst their peers. Then I used a scatterplot to help show the correlation between the number of third downs converted versus failed. I used a filter that only showed players who have an average of 200 passes per season as I didn't want outliers or NULL values. I used Size and Color to help accentuate the differences in both graphs. However, is there a statistic out there that truly measures a QB's efficiency and the number of points he added?



Sum of Epa for each Passer Player Name. Color shows sum of Epa. The marks are labeled by sum of Epa. The view is filtered on Passer Player Name, which keeps 24 members.

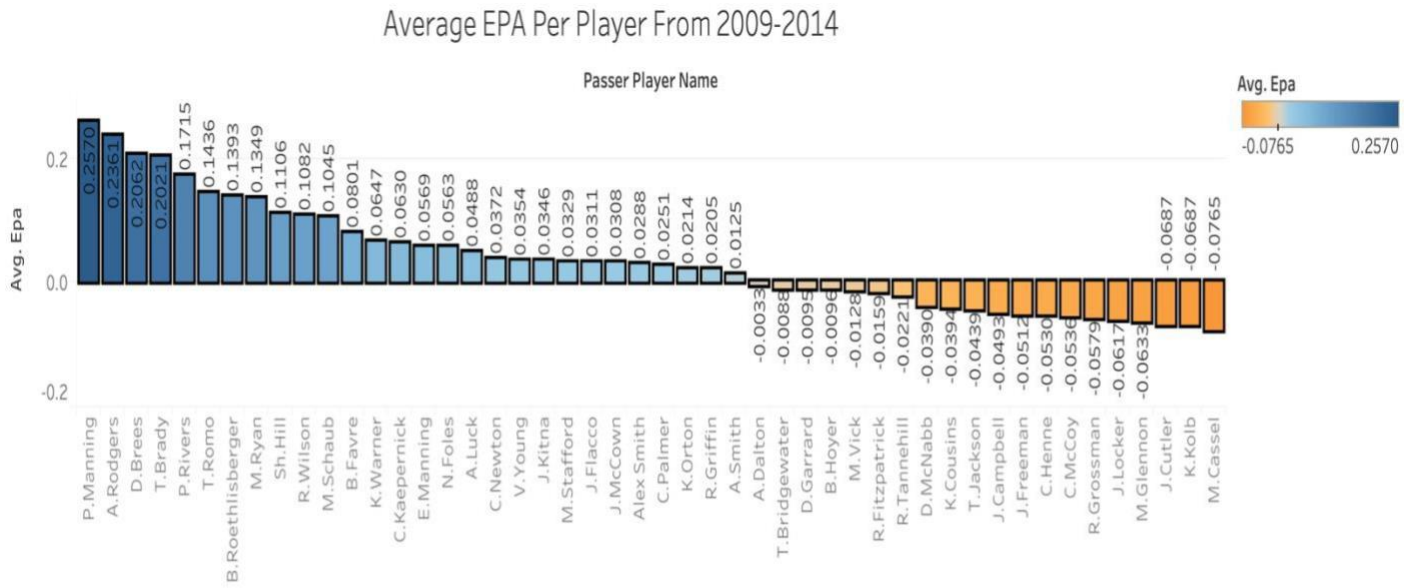
Estimated Points is a statistic that is cumulated by considering the down, the field position, and the remaining distance towards a TD. Based on the play that happened, the difference between the Estimated Points for that play and the Estimated Points before the play

is called Estimated Points Added(EPA). This is more or less measuring the efficiency of a player per given play. By taking the sun of EPA per player(who have completed 200 or more passes), the bar graph above displays that Drew Brees, Tom Brady, and Philipp Rivers have added been the most efficient QBs over the past 10 years. I used a bar graph here as it was the best form of measure to compare who have added the most value these past 10 years amongst their peers, and then I used color to help accentuate the differences in value amongst the best of the best. Furthermore, I only used the top 24 values as there are too many players in the dataset. However, we still don't know if EPA is a valid measure in this study as it has not been correlated to winning yet. The question now is if higher EPA lead to a higher Win Probability.



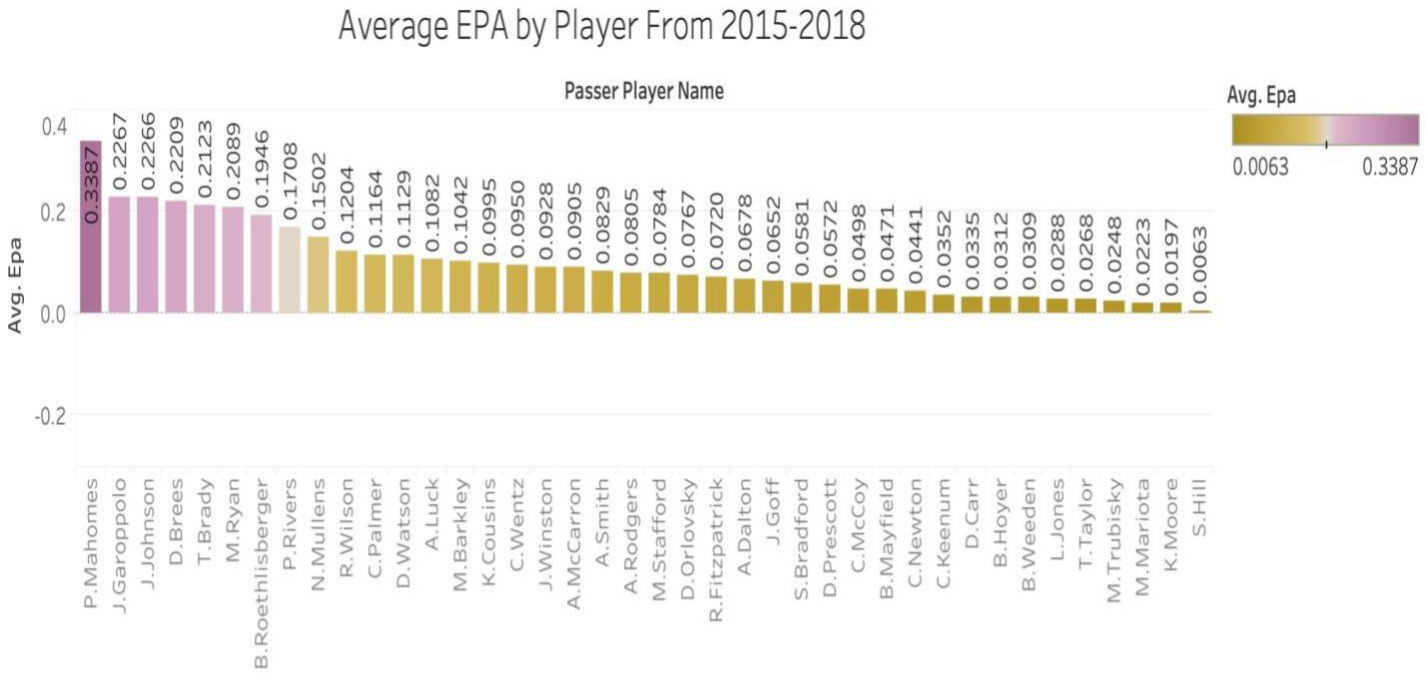
Average of Wpa vs. average of Epa. Color shows average of Epa. The marks are labeled by Passer Player Name. The data is filtered on sum of Complete Pass, which includes values greater than or equal to 200.

From the graph above, Win Probability Added Per Play(WPA Average) is highly correlated to EPA Average. What this means is that a higher EPA by a player will most likely lead to a team's Win Probability for that game to increase. Essentially, the more efficient you are, the more likely the team is going to win. From the scatterplot above, Patrick Mahomes, Jimmy Garoppolo, Ben Roethlisberger, and Tom Brady are the most efficient QBs per given play. I used a scatterplot here as I needed to see the correlation between two measures(EPA and WPA) and then labeled the player name for each point to help give me a better understanding of who the most and least efficient QBs are and how much they hurt their team's chances of winning per given play. I filtered the graph in such a way that only players who have completed 200 passes could be plotted on the graph(to make sure that there are no outliers). Then I used color to help distinguish the players who have the highest and lowest EPA Average.



Average of Epa for each Passer Player Name. Color shows average of Epa. The marks are labeled by average of Epa. The data is filtered on Game Date (MY) and sum of Complete Pass. The Game Date (MY) filter keeps 27 of 45 members. The sum of Complete Pass filter includes values greater than or equal to 200. The view is filtered on Passer Player Name, which keeps 48 of 377 members.

An inherent problem with the graphs above was that the timeframe was too large. It didn't account for the year to year differences, and some players were penalized for not playing in some of these years due to factors such as retirement. Thus, I filtered this particular graph to only include data from the 2009 to 2014 seasons. Then I wanted to measure how many points each QB added per play as one of my previous graphs was cumulative. This is a better measure of efficiency as it takes each QB's efficiency per any given play. Peyton Manning had the highest EPA Average from 2009 to 2014 with a value of .2570 while Matt Cassel had the lowest with a value of -.7605(Matt Cassel essentially lost his team -.7605 points per play). I used a bar graph here as this was the ideal form of measurement between the different players' EPA and the color helps me fully realize this difference.



Average of Epa for each Passer Player Name. Color shows average of Epa. The marks are labeled by average of Epa. The data is filtered on sum of Complete Pass and Game Date (MY). The sum of Complete Pass filter includes values greater than or equal to 20. The Game Date (MY) filter keeps 18 of 45 members. The view is filtered on Passer Player Name, which keeps 40 of 377 members.

Extending from my previous visualization, this next dataset explores Expected Points Added Per Play from 2015-2018. Furthermore, I filtered Passer Player Name and Completed passed to include the top 40 values and 200 or more passes completed as my dataset was too large in both instances. From the visualization above, Patrick Mahomes added .3387 EPA per play and has been the most efficient QB from 2015-2018. Other efficient passers after Mahomes are Brees, Brady, Ryan, and Garoppolo. Some of the names from the top of the bar graph from 2009-2014 either declined or dropped off severely like Peyton Manning. However, a few names have remained consistent and have maintained elite QB play through both decades such as Tom Brady, Drew Brees, Ben Roethlisberger, and Matt Ryan. While this timeframe isn't perfect in terms of compilation bias (for example Mahomes came into the NFL in 2018), it's undoubtedly better than the 10-year time frame in this regard and is not as subject to variance as a year to year timeframe. By implementing a bar graph, I am able to clearly display the most efficient passers from 2015-2018; moreover, I can use color to help differentiate the players' average EPA in a less cluttered manner.

In conclusion, players that complete more passes and throw TD passes are generally the ones that are most efficient as well and thus, help their teams win more often than less efficient QBs. The most consistent QBs throughout this EDA have been Tom Brady, Aaron Rodgers, Drew Brees, Matt Ryan, and Philip Rivers. While QBs who haven't played many years during his time frame have put up impressive efficiency numbers such as Patrick Mahomes, the aforementioned QBs' elite efficiency and volume simply make them the best QBs of this past decade. It isn't surprising then that these QBs were the QBs of teams who have been to 8 of the past 10 Super Bowls in this timeframe.

Why I Used Tableau to Visualize My Data

The reason I used Tableau is for its Automation Functionality. Tableau is a little more intuitive with creating processes, visualizations, and calculations. For example, when creating calculations in a tabular format, the formula can be typed once, stored as a field and applied to all rows referencing that source. Tableau's flexibility also allows users to create custom visualizations, and filters that aren't available in most of the tools. Compared to other tools such as Excel, it is a lot easier and more efficient tool.