Connor Saari

Problem Set 6

* Initial DataFrames
  + DataFrame 1
    - This is the initial dataframe that I pulled from Kaggle, listing the career stats for quarterbacks in the NFL. This dataframe was used because it included the two players I was primarily looking for. The bust was Johnny Manziel and the success was Ryan Fitzpatrick.
  + DataFrame 2
    - This second dataframe was included for the use of the visualizations. Altair uses dataframes with under 5000 data points, this secondary dataframe was used because it included the first 100 players as well as the rows associated with Manziel and Fitzpatrick.
      * For the use of this dataframe, particular columns had to be formatted and sorted to be displayed better in the Altair visualizations.
* Visualizations
  + Visualization 1
    - This visualization Is primarily concerned with the general context of the players. It is a circle plot showing the names of the players on the y axis and the number of games each player participated in per season on the x. This shows how Fitzpatrick has participated in many, many more games over several seasons for different teams, whereas Manziel only participated in two seasons and was bust after that point to never play again for another NFL team even though he was prospected to have great potential.
  + Visualization 2
    - This second visualization represents the total number of passes each player completed. As you can see the primary bar in this chart is Fitzpatrick at a high of 335 completed passes in a season, whereas Manziel only completed 129. When hovering over each you can compare their similar completion percentages, which is skewed greatly because of the limited number of passes Manziel threw making the chances of him having a higher completion percentage greater. Fitzpatrick had to maintain such a high percentage while also throwing double the amount of passes.
  + Visualization 3
    - This visualization is to emphasize the mistakes that Manziel had. This chart shows the names of the players against the total TD Passes by that player. Fitzpatrick as you can see had 31 touchdown passes and 19 sacks with the Jets. Manziel had the same amount of sacks as Fitzpatrick but only 7 touchdown passes with the Browns. This shows that Manziel had the same amount of sacks with a significantly smaller amount of completed touchdown passes.
  + Visualization 4
    - This area chart is to visually show the different careers of the two players. This chart represents each player’s total touchdown passes each year they were in the NFL. The pink and red areas on the right side represent Fitzpatrick’s and Manziel’s careers. As you can tell Fitzpatricks stayed and accomplished much greater results than Manziel, which directly corresponds to their accompanied characteristics.
* KNN
  + This was difficult to find. Primarily because clustering on a group of similar features was not attainable. The Position column used in your data, was not filled with any other data other than QB, because each of these players were quarterbacks. When aggregating on other features, the training data and query data did not match because there were players who had different sums than all the others. This made it impossible to group them into similar groups for comparison. However, some of the similar statistics I saw from other aggregation methods was that Fran Tarkenton was the most similar to Fitzpatrick, and Jack Scarbath was the most similar to Manziel. Based on similar features such as TD passes and Pass completion. Both of the successful players achieved career highs of 29 and 31 total touchdown passes in a year, with similar averages of about 61% completion. Whereas the busts also had very similar statistics, from career highs of 9 and 7 in one season, Manziel did somewhat worse in terms of completion percentage because of the amount of sacks he took that season.
    - After comparing several players based on the features I thought most appropriate, the players aboves are displayed in terms of their average completion percentage, passing yards per game, and touch down passes per season. These players are the most similar in the data frame I created, not the overall dataframe.
* KMeans
  + Based on the graph at the bottom of the code, the k that would be the most optimized for gain would be around 8. This is where the gains are noticeably stagnant and do not increase beyond this point.
* Synthesis
  + Overall there is a clear difference between the players that were successful and the players that were bust. The primary characteristic of bust players is that they generally spike in specific features like passes completed or games played, but when analyzing all the players statistics over time, it is clear the bust players were cut early and all that remains is a season or two of data. Contradictory to this assumption, successful players excel in areas like completion percentage and touchdown passes. As well their careers over time look much different from the bust, in that they have much more data over more seasons, generally starting with high average numbers in passing and completion which slowly increase as the seasons progress. For example, in both visualization 3 and 4, both successes and busts represent very similar patterns. As well, the averages below represent very similar statistics, from above 2000 passing yards a season for the success and sub 1000 passing yards per season for the busts. As well, the average touchdown passes per season show similar trends because of playing time similarities, and overall experience aggregated over the seasons. Overall, there are very similar trends for successful players and bust players that could help to find prospective players in the future.