This project will require you to the skills that you have learned throughout the semester. You will be performing some simple statistical analysis (data summaries and visualization) on data that you scrape from the web. The data will need to be cleaned or pre-processed before you can use it.

*You will turn in a copy of your annotated R script along with this document. Answer each of the questions below as you work.*

**Data Collection Requirements:** You must use at least two webpages, and you must scrape data for at least 5 variables total (between the different webpages). At least two should be categorical variables, and at least 3 should be quantitative variables. This means you must collect both numeric and character data. Keep this in mind as you answer the first few questions!

Choosing a Topic and Collecting Data:

1. Come up with a topic to research and explain why it is relevant to you.

Topic: NFL 2021 Quarterback Analysis

Relevance: Football is a sport where analytics are becoming more and more popular as the game evolves. My dream job (and my current hobby) is one where I can combine my passion for analytics and football. Quarterback is the most important position in football, so it is important to be able to evaluate their performance and find the best one. However, the cost of said quarterback is also very important. Why pay extra when you could get more value for a slight hit in production? Ultimately, the topic is relevant to me because it I enjoy both and I want to further my understanding of the subject.

1. Brainstorm at least seven possible research questions related to your topic that you would like to explore by scraping data from the web.

1. Who scored the most points for their team? (passing td + rushing td)

2. Who was the most efficient?

3. Who had the lowest pct of their total yards come from passing?

4. Who had the most first downs? (passing + rushing firstdowns)

5. Who outperforms their salary? Best value player? (Wins?,TDS?,Yards?)

6. Do higher paid QBs win more games? (Have a higher win percentage?)

7. Which conference spends the most on QB? Division?

8. Which conference scores the most? Division?

9. Does a larger investment at the position relate to number of wins?

10. Are higher paid Quarterbacks better? score more? More first downs? Less turnovers? Etc.

.

.

.

1. Find multiple web pages containing data that will help you explore as many of your questions from #2 as possible. Once you find appropriate webpages, read them into R and perform your web scraping. Provide a brief description here of each web page and what information you scraped from each. [See data collection requirements above!]

Webpage #1 Description and Scraped Data:

https://overthecap.com/position/quarterback/2021/

OverTheCap – The table that I scraped from the web page includes columns for player names, team names, cap number and cash spent. (2 categorical data, 2 numerical data). Essentially, it contains data for nfl player contracts including cap number and cash spent. Cap number is basically the players base salary that goes against the teams total cap. Cash spent is money that comes in the form of bonuses (signing bonus, workout bonus, etc.) and does not impact the team salary cap. The cash spent often varies year to year because players with new deals often have larger signing bonuses than players on their second or third year of their contract. Because of this, these players usually have a higher total compensation for the year even if their salary is lower than another player.

Webpage #2 Description and Scraped Data:

<https://www.pro-football-reference.com/years/2021/passing.htm>

PFR – the website stores football analytical data dating back to the 1920s before the NFL was even established. I scraped data from their 2021 passing statistics table. The data set includes many stats like games, passing yards, completions, attempts, touchdowns, interceptions, etc. There are also advanced statistics that rate performance. The three categorical data pieces are the players name, position, age, and team. The rest of the data is numeric. I will not use all the numerical columns, but the majority of them will come into play.

Webpage #3 Description and Scraped Data:

<https://www.pro-football-reference.com/years/2021/rushing.htm>

PFR – the website stores football analytical data dating back to the 1920s before the NFL was even established. I scraped data from their 2021 rushing statistics table. Since some quarterbacks offer more than just passing talents, it is important to incorporate rushing statistics too. This table contains quantitative variables like rushing yards, touchdowns, attempts, fumbles etc. It also includes categorical data like name, position, age, and team.

Preparing your data for use:

1. Clean up your data so that it is ready for analysis and visualization. As you work describe what you have done at each step. Along with any other work that needs to be done, make sure to cover the following:
   1. Convert variables/data into the correct format.
   2. Check for outliers and errors. If you find “problems”, decide how to handle them.
   3. Check for missing values. If you find missing values, decide how to handle them.
   4. Perform any necessary data transformations.

This is similar to the data pre-processing homework. Also see the R script for the web scraping examples.

Detailed list of steps taken to perform data-preprocessing/ data cleaning:

OTC preprocessing:

1. Scrape table from OTC and make dataframe called qb.2021compensation.
2. Change column names for from (Player, Team, Cap Number, Cash Spent) to (Player, Team.Name, Cap.Number, Cash.Spent). Essentially just removing the spaces and changing Team to Team.Name so I can add a column named Team that matches the other datasets’ Team columns.
3. Add in Taysom Hill contract details. Since he plays both Tight end and QB he was excluded from the dataframe. However, since he started multiple games for the Saints, I felt that he should be included so I found his contract details online at Spotrac (which is a notable professional sport contract website) and added him in.
4. Removed symbols from Cap.Number and Cash.Spent columns and then converted to numeric. Ex – ($32,000,000 to 32000000)
5. Adding Total.Comp column (Cap.Number + Cash.Spent). This column shows the total amount of money the player was given for the year (salary, bonuses, etc.)
6. Renaming player “Gardner Minshew” to “Gardner Minshew II” so that the data sets can merge correctly.
7. Create new data frame called teamdata with team information to allow merging for later. The Team column is compatible with other two datasets while the Mascot column matches the Team.Name column in the qb.2021compensation dataframe. Contains Team column with team abbreviations, Mascot column with team mascots, Conf column with team conference, and Div column with team division information.
8. Merge the teamdata datafram to the qb.2021compensation dataframe

PFR Passing preprocessing:

1. scrape passing table from PFR and make df called qb.2021passing
2. Renaming and removing useless columns. Renamed columns to Team from Tm, P.Att from Att, Cmp.Pct from `Cmp%`, TD.Pct from `TD%`, Int.Pct from `Int%`, P.Yds from Yds, P.TD from TD, P.First.Downs from `1D`, P.YdsPerAtt from `Y/A`, P.YdsPerCmp from `Y/C`, P.YdsPerGame from `Y/G`, Q4.Comebacks from `4QC`. Removed columns 1,18,20,26,27,28,29 that contained statistics that do not relate to questions and are somewhat useless for evaluation.
3. Separated QBrec into Wins, Losses, and Ties
4. Changed all columns that should be numeric to numeric from character. (the dataset came as only character data) This involved the columns Age, G, GS, Wins, Losses, Ties, Cmp, P.Att, Cmp.Pct, P.Yds, P.TD, TD.Pct, , Int, Int.Pct, P.First.Downs, P.YdsPerAtt, P.YdsPerCmp, P.YdsPerGame, Rate, QBR, Sk, Q4.Comebacks, and GWD.
5. Remove rows of passers that aren’t quarterback. Done by filtering if the position was ‘QB’, ‘qb’, or ‘’. Have to include these columns because the websites characterization of position involved playing time to distinguish between ‘QB’ and ‘qb’. Also had to take into account that some players (ex- Ben Roethlisberger) have empty position columns for an unknown reason. After filtering down to these remaining players, some other positions were still left because of the empty position column for them. So I had to remove the rows of these players (82,81,80,79,75,73,71,65) manually. Now all that is left are quarterbacks.
6. Since the only players remaining are QBs, we can fix the Pos categorization of the players and change them all to ‘QB’
7. Remove symbols by player names. Players who won awards have symbols by their names, so we have to remove them so that way the data sets can merge.
8. Removed quarterbacks with less than 1 GS (game started). These players have a very limited sample size (often less than 5 attempts) and many of their columns were empty or filled with NAs. It would be simpler to remove them to avoid outliers than to try to fill them with data that is not representative of their efforts.
9. Fix the NA’s for columns of GWD (game winning drives) and Q4.Comebacks (fourth quarter comebacks). These rows had NAs because no stat was recorded for GWD or Q4.Comebacks for these players. The website says that they were left empty because these players were not put into a situation to achieve these things. For the sake of this data frame, these NA’s will be changed to 0 because at the end of the day, they did not achieve these statistics so it would be unfair to assign any other value to its place.
10. There is still an error with player Josh Johnson’s team “2TM”, but we will fix that after merging the passing and rushing tables. This will later change to BAL as it was the most recent team he was on. Though since the rushing table has the same issue, it is easier to change it once after the merge than twice before hand.

PFR Rushing preprocessing:

1. scrape rushing table from PFR and make df called qb.2021rushing
2. column names were in first row, so I stored them in a variable and then rename the columns with them.
3. Remove and rename columns. Remove column 1 and 12. Rename columns to Team from Tm, R.Att from Att,R.Yds from Yds, R.TD from TD, R.First.Downs from `1D`, R.YdsPerAtt from `Y/A`, R.YdsPerGame from `Y/G`
4. Change columns that are character but should be numeric to numeric. Age, G, GS, R.Att, R.Yds, R.TD, R.First.Downs, R.YdsPerAtt, R.YdsPerGame, Fmb
5. Remove non QBs. Done by filtering if the position was ‘QB’, ‘qb’, or ‘’. Have to include these columns because the websites characterization of position involved playing time to distinguish between ‘QB’ and ‘qb’. Also had to take into account that some players (ex- Ben Roethlisberger) have empty position columns for an unknown reason. After filtering by position, we will filter by GS >=1. This will give us the players that have started a game (like our other data set) and further narrow the data set. The remaining players are mostly quarterbacks, however, some other positions made it through so we will have to manually remove the rows (1,2,6,10,17,19,32,60,61,66,67,68,69,73,74,75,77). Now only Quarterbacks are left.
6. Since the only players remaining are QBs, we can fix the Pos categorization of the players and change them all to ‘QB’
7. Removed symbols by player names. Players who won awards have symbols by their names, so we have to remove them so that way the data sets can merge.
8. There were no NA’s or empty cells to fix so this data set is done for now.
9. There is still an error with player Josh Johnson’s team, but we will fix that after merging the passing and rushing tables.
10. Now combine ALL of your data into a single data frame. Your data should be in a *tidy* format. Explain what that means for your data.

Steps done for merging

1. Merge passing and rushing dataframes
2. Fix the two quarterbacks that recorded zero rushing stats. (Nick Mullens and Garrett Gilbert). I looked on ESPN and NFL.com to confirm they had no rushing statistics and then filled in their rushing stats with 0s.
3. Fix the Josh Johnson Team issue. Change Team “2TM” to “BAL” because that was the team he played for most recently.
4. Finally, merge the compensation dataframe with the new dataframe containing both rushing and passing stats. This new dataframe is called qb.2021master

Tidy data explanation:

For this data frame, each player is an observation from the 2021 season. The data frame is tidy because each row is an observation and every column is a variable whether it be categorical like Team or quantitative like Wins. Each value is also in its own cell. Since it meets these conditions, the data frame is in tidy format (“long” format).

Data Analysis and Visualization:

1. Choose at least three of your questions from #2 to explore. State them here. (Feel free to alter your original questions to reflect the data you were able to find and scrape.)
2. Do higher paid QBs win more games? (Have a higher win percentage?)
3. Are higher paid quarterbacks better? score more? Less turnovers? More “clutch” (GWD, 4Q.Comebacks)?

3. Which conference spends the most on QB? Division? Do they win more?

Overall Question: Was a larger investment at the quarterback position worth it in 2021?

1. Now that your data is clean and tidy and ready for use, use dplyr and ggplot2 commands to extract and explore the data (numerically and graphically) in order to answer your question(s) from #6.

Explain your process as you work. What are you doing and why? In other words, what approach are you taking to answer each question (looking at what data, statistics and graphs)? Be specific. Approach your question(s) from *multiple* angles.

Use as many of the dplyr commands as possible, but only what makes sense for your question(s) and data.

Process to answer first question: Do higher paid QBs win more games? (Have a higher win percentage?)

First, created a subset with Player, Wins, Losses, Ties, Cap.Number, and Total.Comp. I compared Cap.Number and Total.Comp to Wins by creating a scatterplot for both comparisons with a line of best fit. By doing so, I noticed that higher paid quarterbacks tend to have more wins for both categories of Cap.Number and Total.Comp. However, wins aren’t the only thing to consider. The percentage of wins also matters. After making an adjustment to my dataset with the mutate function, I added a win.pct column. This allowed me to look at a players win.pct versus their Cap.Number and Total.Comp.

Chart, scatter chart

Description automatically generated Chart, scatter chart

Description automatically generated Chart, scatter chart

Description automatically generated Chart, scatter chart

Description automatically generated

However, after looking at the subset several times, I noticed the lack of amount of games started for some quarterbacks is throwing the win.pct off (limited sample size). So, I created a new subset with the same columns as before (including win.pct) but filtered out players with less than 4 games started. Then I plotted the same graph with the new subset.

Chart, scatter chart

Description automatically generated Chart, scatter chart

Description automatically generated Chart, scatter chart

Description automatically generated Chart, scatter chart

Description automatically generated

After looking at the four graphs (cap.number, vs wins, total.comp vs wins, cap.number, vs win.pct, total.comp vs win.pct), it seems clear that quarterbacks who get paid more (salary and in total) do win more games and win them at a higher percentage. However, these correlations do not mean that paying someone more money will win you more games. In fact, the graphs tell me that quarterbacks who get paid more get the opportunity to win more games (more games started), but that does not necessarily equate to a higher win percentage. Win percentage does trends that way, but it’s not as strong of a correlation as just wins and compensation. Instead, this feels more like a symptom of other things to consider.

Process to answer second question: Are higher paid quarterbacks better? (score more? Less turnovers? More efficient (passer rating)?

When you think of good/great quarterbacks, many have similar qualities. Most of the time you look at the number of touchdowns they score, how often they turn over the ball, and how do they react when the game is on the line. So, to decide if higher paid quarterbacks are better, we will look at the relationship between these things and how much they are paid.

1). First, lets look at if they score more. To do so, I made a subset with the columns Player, P.TD, R.TD, Cap.Number, Total.Comp. I include Cap Number as a comparison because it is representative to the salary that counts directly against the team while total compensation more accurately portrays what the quarterback is actually being paid. I also mutated the subset to include a total touchdown (Total.TD) column which was the sum of P.TD and R.TD. Then I compared Cap.Number and Total.Comp to Total.TD with a scatterplot. The graphs show us that quarterbacks who are paid more tend to score more touchdowns.

Chart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generated

Now lets limit the sample size (to quarterbacks with 4+ games started) and see if it remains true. So to do so we first have to create a new subset with a filter on games started. After doing so, we will make the same scatterplot as before with our new subset.

Chart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generated

These graphs tend to show the same story that quarterbacks who are paid more score more points, but do they score at a higher percentage? By changing our subset to include rushing and passing attempts and mutating it to include a Total.TD.Pct column ((P.TD+R.TD)/(P.Att+R.Att)), we are able to see how often these players score.

Chart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generated

These graphs tell us that the percentage of touchdowns per attempt is not that closely related to a quarterback’s compensation. However, looking at the graph, it is clear to see that players who are paid more have a smaller range of percentages that are in the upper half of the graph. This suggests that they do score at a higher percent than lesser paid quarterbacks. However, you also see several players with smaller contracts scoring very frequently in all graphs. These players are stuck on their rookie contracts and their earnings are limited until they are off of them. It is likely that these quarterbacks will join the higher paid quarterbacks once their contracts are up and it is time for a payday.

2). Now lets see if they turn the ball over less often. To explore this question, I thought about how players who play more will turn over the ball more naturally. So instead of looking at a volume stat like total turnovers, I decided to look at turnovers per attempt. Since all of my other graphs tended to be more clear with a filter of players who started more than four games, I went ahead and included that piece when making my subset for this question. I then selected the columns Player, P.Att,R.Att,Int,Fmb, Cap.Number, and Total.Comp. After that I added a column called TO.Pct ((Int+Fmb)/(P.Att+R.Att)). This would allow me to see the percentage of a players turnovers per attempt.

Chart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generated

Both graphs show a similar trend that players that are higher paid turn the ball over less. However, this also comes with experience. Any player in the right half of the graph has already finished their rookie contract and is more experienced than others. But experience is a valuable thing which is why they get paid the way they do. In either case, these quarterbacks that are paid more, do turn the ball over less.

3). Next lets look at efficiency. For quarterbacks, this is often measured in passer rating (Rate in the data frame). The max score is 158.3 and it takes things like completion percentage, yards per attempt, touchdowns per attempt, and interceptions per attempt, into account. Essentially, it calculates how effective a quarterback is at their job and is ultimately the best measure of a quarterbacks performance. To evaluate this, lets create a subset. Our subset will contain the columns Player, Rate, Cap.Number, and Total.Comp. Since we have done it in other parts of answering this question, lets also filter out players with less than 4 games started. Next we will create a scatter plot to compare passer rating (Rate) to Cap.Number and Cash.Spent.

Chart, scatter chart

Description automatically generated Chart, scatter chart

Description automatically generated

These graphs paint a similar picture to our other findings. They tell us that higher paid quarterbacks do perform better. However, there is a grouping on the right half that performs at the highest level while their total compensation is not very high. Who are these players? These are players like Justin Herbert, Lamar Jackson, and Joe Burrow that are performing incredibly well for how early they are in their careers. Their compensation is low because they haven’t yet “earned” their second contract. They are still on their four year rookie cost controlled deals. Essentially meaning that they are among the most valuable in terms of cost vs production. However even with their superior performance compared to others in their price range it is not enough to disprove that higher paid quarterbacks tend to be better.

Chart, scatter chart

Description automatically generated

Here is a better representation of what I mean. To make the graph above, I created a list of players still on their rookie deal (with a 4 game starting sample size). I used this list and the mutate function to add a column to the subset we just finished using that displays if the player is still on their rookie deal. Then I mapped color to this new column and created the same graph. As you can see, the grouping on the left that outperforms their contracts are mostly on their rookie deals.

At the end of the day, even accounting for rookie deals, players that are paid more, tend to perform better.

Process to answer third question: Which conference spends the most on QB? Division? Do they win more?

To answer the first question, the NFC spends the most. I got this answer by first selecting the columns Conf,Div,Total.Comp,Wins,Losses,Ties,Age. Then I used group\_by with the Conf variable. Then I computed the following summary statistics: (TotalSpent=sum(Total.Comp),TotalW=sum(Wins),TotalL=sum(Losses),TotalT=sum(Ties),WinPct=TotalW/(TotalW+TotalL),AvgAge=mean(Age))

This gave me the following table:

Conf TotalSpent TotalW TotalL TotalT WinPct AvgAge

1 NFC 680209527 136 135 1 0.502 29.3

2 AFC 403233075 135 136 1 0.498 27.5

This tells us that the NFC spent more and won more games. However, the difference is very small for spending ~18 million more.

Now, lets look at it by division. We will do the same steps as before but substitute in Div where Conf was in the group\_by(). This gives us the following table:

Div TotalSpent TotalW TotalL TotalT WinPct AvgAge

1 NFC West 190354741 40 28 0 0.588 29.6

2 AFC West 91746427 38 30 0 0.559 26.6

3 AFC North 94467638 35 32 1 0.522 28.6

4 AFC East 94247727 34 34 0 0.5 26.3

5 NFC South 159783021 34 34 0 0.5 30.3

6 NFC East 136824502 32 36 0 0.471 28

7 NFC North 193247263 30 37 1 0.448 29.4

8 AFC South 122771283 28 40 0 0.412 27.8

This tells us a different story. While the NFC West did spend the second most and win the most. The following three divisions spent the least and were 2nd 3rd and 4th in wins. This seems to tell us that the heavy spenders do not always win more. However, If you take into account the average age of the players in these divisions, you have to think about the many excellent players that are on rookie deals.

1. Now present your results! Pretend you are writing a short article on your topic/specific question(s) that will be published online. For example, see the articles from the Economist and FiveThirtyEight under in-class activities.

Your article should be 1-2 paragraphs in length, and it should include:

* An informative title that conveys the topic of your research.
* The answers to your questions with explanation.
* Supporting graphs, statistics, and/or other summaries. Note: You do not need to include everything from #7. Only present the most relevant information.
* A statement about where your data came from.

Bonus points will be given if your article contains a graph that displays multiple pieces of information that are clearly labeled within the graph itself. See the Joseph graph for example.

Article on next page

Your Article:

**Was a larger investment at the quarterback position worth it in 2021?**

With NFL Free Agency coming to an end and the draft quickly approaching, many fans have questions surrounding their team. Some wonder what the draft will bring to their team while others are left reeling about what their team paid or lost in free agency. However, one burning question sits at the heart of every NFL franchise: Is our quarterback worth what we are paying him? This particular question usually comes up several years into a deal when their cap number prevents the team from signing big name players or when they begin campaigning for a new contract. This often leaves teams wondering if the cost of their star players are worth what one player can bring to the field.

To evaluate the question of if investing heavily at quarterback is worth it, we need to consider a few things. Questions like “Do they win more?” and “Do they play better?” can help us better understand whether a quarterback is worth it. By using data from Pro Football Reference’s Rushing and Passing tables and Over The Cap’s 2021 Quarterback Compensation tables, I was able to answer such questions.

First, let’s compare compensation with the most important thing in any sport, winning. By looking at the graphs below, it seems clear that quarterbacks who get paid more tend to win more games at a higher percentage.

Chart, scatter chart

Description automatically generated Chart, scatter chart

Description automatically generated

However, just because they win more doesn’t necessarily mean they are better. By looking at a stat like passer rating compared to compensation in the graph below we begin to understand more of what is going on. If we take it a step further and distinguish between players who are on their rookie deals and not, we begin to see that many of the players that outperform their payments are on their rookie deals. This essentially mean that most players that are outperforming their contracts haven’t gotten the opportunity to become one of the higher paid players because they are on cost controlled deals. But how does that play out league wide?

Chart, scatter chart

Description automatically generated

Looking at splits by division and conference we begin to see things play out. While investing large sums of money works out for the NFC as a whole leading to more wins, when you look at it by division you begin to see something different. With the tables below, you see that the second highest spending division, the NFC West, won the most games. However, the next three divisions in wins were the ones that spent the least. While this might tell us that spending does not correlate to wins, it more clearly tells us the value of rookie deals. The divisions that spent the least all have low average ages which tells us that many of their quarterbacks are likely on rookies deals preventing them from having to pay large sums of money to the position. However, these valuable players on rookie deals are time bombs waiting to blow up into massive contracts.

Text

Description automatically generated with low confidenceTable

Description automatically generated with medium confidence

At the end of the day, unless you have a superstar quarterback on a rookie deal it is worth investing at the position. That doesn’t mean shell out your cash to a subpar player, but it means chase the good ones when they become available or prevent your player from coming available by paying them. So yes, in 2021 it was definitely worth investing in the position. Take the Rams for example, by shelling out the resources to bring in Matt Stafford, they were able to win a Super Bowl. However, the data also tells you that rookie deals offer the most for what you are paying for. So, if you don’t have a stud or aren’t able to lure one in free agency, draft one and hope for the best. With that being said, that is another form of investment, but draft capital is a topic for a different day. Ultimately, it was absolutely worth investing in the quarterback position in 2021.