Capstone Project: Milestone Report

1. Write a capstone project 1 milestone report (Google Doc, 5-6 pages) and include the following:
    a. Problem statement: Why it's a useful question to answer and for whom (source this from your proposal)
1. Can we utilize historical NFL statistics to predict future performance? This is a useful question to answer for anyone playing Fantasy Football. There are hundreds of thousands of fantasy football players across the world with an $18.6 billion market size. With these predictions players can maximize their chances of winning against the competition. It can also be used for those creating spreads and projections on the gambling side, such as DraftKings and FanDuel.

    b. Description of the dataset, how you obtained, cleaned, and wrangled it (source this from your data wrangling report)
1. Description of the dataset

The dataset can be described as a culmination of NFL offensive statistics from 2014 up to the 2018-2019 season. These stats include touchdowns (passing, rushing, receiving), yards (passing, rushing, receiving), attempts (passing, rushing, receiving), yards per attempt, completions, interceptions, red zone targets, to name a few. The dataset also includes Fantasy Football scoring data, such as Fantasy Points, Points Per Reception, DraftKings and FanDuel scoring, and rankings based off of fantasy points scored per season.

2. How I obtained the dataset

My dataset comes from several different sources in order to gain all of the statistical information needed. One set of data I downloaded as a CSV file, cleaned the data in Excel by removing unnecessary columns and rows before uploading it to my project notebook. This dataset came from Pro-Football-Reference.com.
One source of my data set was only available through SQL. To wrangle and incorporate this data called for me to use SQL to read the data into my Jupyter notebook. To integrate this data I had to use SQLite, which meant creating a cursor to use .execute on where I was then able to write SQL to sort and clean the data.

```
conn = sqlite3.connect('Data/nfl.db')
```

```
cursor = conn.execute('SELECT * FROM stats_offense')

# Query Database

cursor.execute('''
SELECT * FROM stats_offense
WHERE season > 2013
''')
stats_df = cursor.fetchmany(10)
stats_df = pd.DataFrame(stats_df)

cols = [i[0] for i in cursor.description]
print(cols)

stats_df.columns = cols
stats_df
```

3. Description of the cleaning and wrangling steps

Setting header and global index column when importing data.

```
pd.read_csv('Data/Fantasy_Stats_2018-19.csv', header=1, index_col='Rk')
```

Sorting the data by year by creating a ‘Year’ column to the newly imported data file.

```
fantasy_stats_2018['Year'] = '2018'
```

Cleaning column names by using the .rename method.

```
fantasy_stats_2018 = fantasy_stats_2018.rename(columns={'VBD▼': 'VBD'})
```

Cleaning up player names by stripping additional characters attached to the end of the name by using .str.strip() .str.replace(,) and .str.partition() methods.

```
qb['Player'] = qb['Player'].str.strip('*+')
qb_week_leaders_2018['Player'].str.replace('\\', "..")
qb_week_leaders_2018['Player'].str.partition("..")
```

I used these same cleaning steps for the running backs, wide receivers, and tight end data sources.

When running into null or missing values I used .fillna(0) method. This method was able to solve all missing value issues when dealing with numeric data.

```
qb = qb.fillna(0)
```

Again, I used this to deal with missing values across all of my different data sources.

c. Initial findings from exploratory analysis (source this from your data story and inferential statistics reports)

i. Summary of your findings

1. I have found several trends within the data. For example, there is a strong trend for running backs who have a high number of rushing attempts to produce higher fantasy points. This trend is consistent across all positions, with a few anomalies. This leads me to question what causes these outliers and if there is a way to identify them. If you can identify the outliers this will give you an advantage over your peers.
2. I gathered insight into what specific stats are most important to a player's success (total fantasy points scored). As well, I am able to determine higher correlations for a high scoring player and the team/coach/scheme they play for so that I am not relying only on a single player in vacuum but am taking into account the environment or system they play within.
3. My original hypothesis; The higher a players usage rate is (determined by attempts, percentage of snaps, percentage of team targets, or percentage of team attempts for the associated position) results in higher fantasy point production.

ii. Visuals and statistics to support those findings

1. I created a plot displaying the correlations of the positional stats that relate to the top performers for each position. For example, rushing yards and attempts have a high correlation to the top performing running backs.
2. I created scatterplots for multiple variables to visually display what stats consistently return higher fantasy point production. As well, these scatterplots point out any outliers in the data/stats.