Final Presentation
A Statistical Analysis of Fantasy Football Data

CSCI 4502 Final Project
Mark Haley
Varunjit Srinivas

Intro to Our Project

- We are Mark Haley and Varunjit Srinivas

- Our Project involves doing a statistical analysis of fantasy football data, specifically in the interval between 1999 and 2019.

- What we hope to do with our research is to give those who play Fantasy Football the evidence they need to pick the best players and score the most points possible by the end of the season.
  - We aimed to examine statistics that don't often factor in to pre-season predictions, specifically a proneness to in-season fatigue, career-spanning fatigue, and finally via a cluster analysis of total points amd strength of schedule

Main Tasks

- Preprocessing the Data

- Exploratory Analysis of Data

- Analysis of Average Player Performance

- Strength of Schedule K-means Clustering Analysis

Preprocessing

- We pulled all of our data from the a Github repository that automatically update with the most recent statistics from every NFL game.
- Then we put the data into pandas dataframes

```python
weekYearDf = pd.DataFrame()
tempYearDf = pd.DataFrame()
for years in range(1999, 2020):
    for weeks in range(1, 16):
        readFile = ("https://raw.githubusercontent.com/fantasydatapros/data/master/weekly/" + str(years) + "/week" + str(weeks)
        weeklyDf = pd.read_csv(readFile)
        weeklyDf.drop(['Int', 'PassingAtt', 'Cmp', 'RushingAtt', 'PassingTD',
                       'RushingTD', 'Rec', 'Tgt', 'ReceivingTD', 'FL',
                       'PPRFantasyPoints', 'HalfPPRFantasyPoints'], axis = 1, inplace=True)

        weeklyDf['Week'] = weeks
        tempYearDf = pd.concat([tempYearDf,weeklyDf])

    tempYearDf['Year'] = years
    weekYearDf = pd.concat([weekYearDf,tempYearDf])

#print(weekYearDf.tail())

QB_weekYearDf = weekYearDf.loc[weekYearDf['Pos'] == 'QB']
#print(weekYearDf['Year'])

RB_weekYearDf = weekYearDf.loc[weekYearDf['Pos'] == 'RB']
print(RB_weekYearDf.head())

WR_weekYearDf = weekYearDf.loc[weekYearDf['Pos'] == 'WR']
```
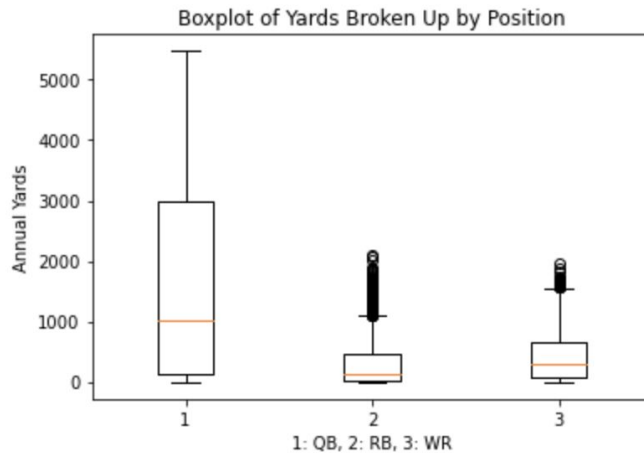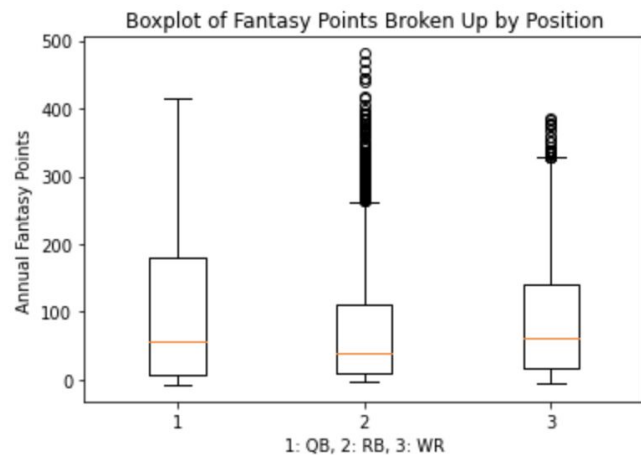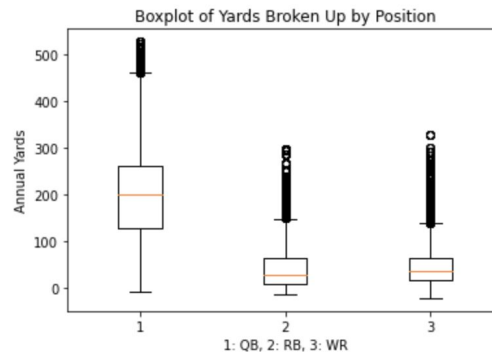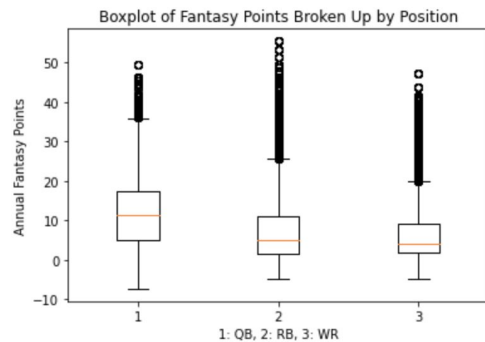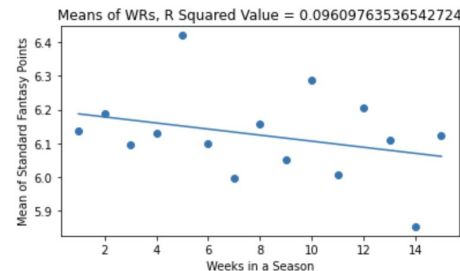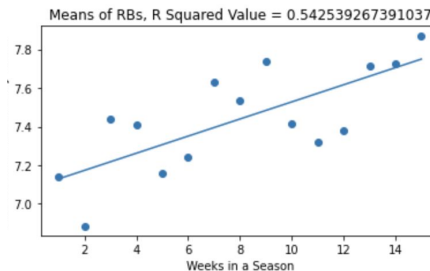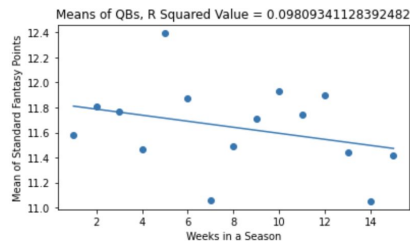
Exploratory Analysis of Annual Data
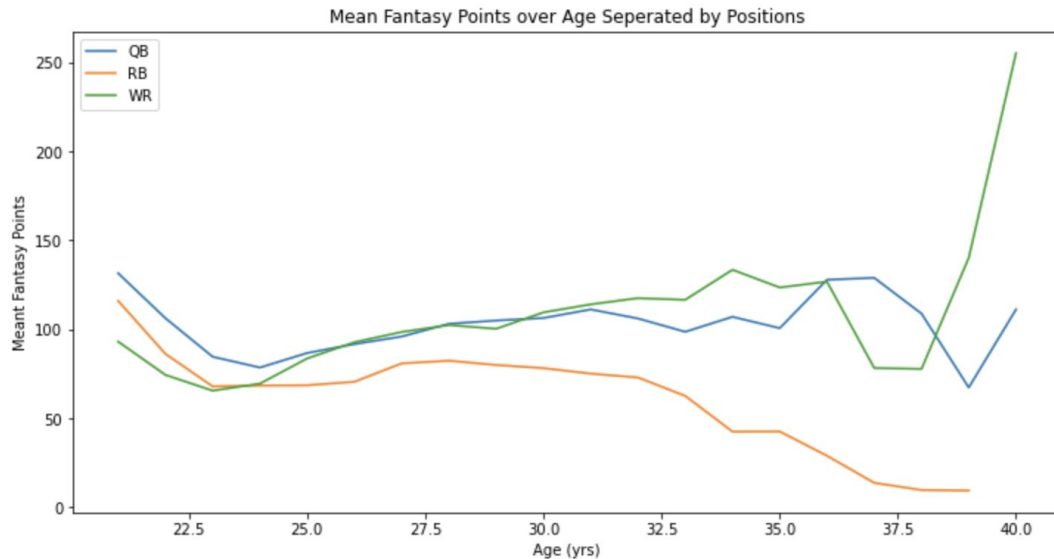
Exploratory Analysis of Weekly Data

Boxplot of Fantasy Points Broken Up by Position
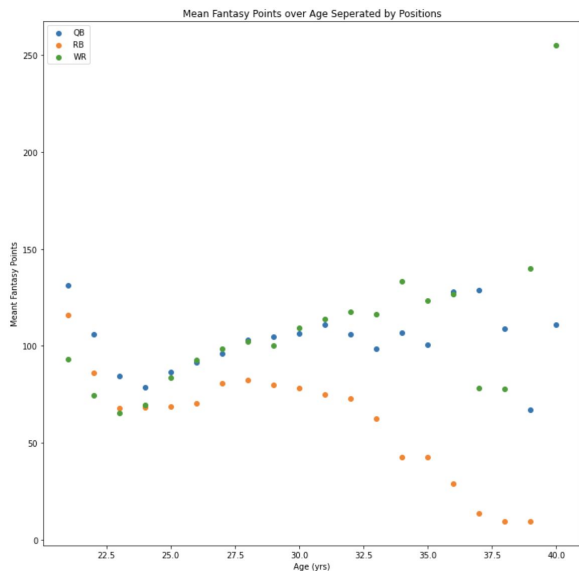

Boxplot of Yards Broken Up by Position

Analysis of Player Performance - In Season

- As evidenced by the graphs below, in particular their respective R^2 values for their linear regression lines, we determined that for QB's and WR's, there is no correlation between in season fatigue and performance. RB's however, actually see a noticeable positive correlation. We attribute this to opposing defenders fatiguing faster over the course of a season than a RB does.

Mean of Standard Points on a Weekly Basis

Means of QBs, R Squared Value = 0.09809341128392482

Means of RBs, R Squared Value = 0.542539267391037

Means of WRs, R Squared Value = 0.09609763536542724

Analysis of Player Performance by Age

Mean Fantasy Points over Age Seperated by Positions

Mean Fantasy Points over Age Seperated by Positions

Strength of Schedule K-Means Clustering Analysis

Scatter Plot of Data to be used for K Means Clustering
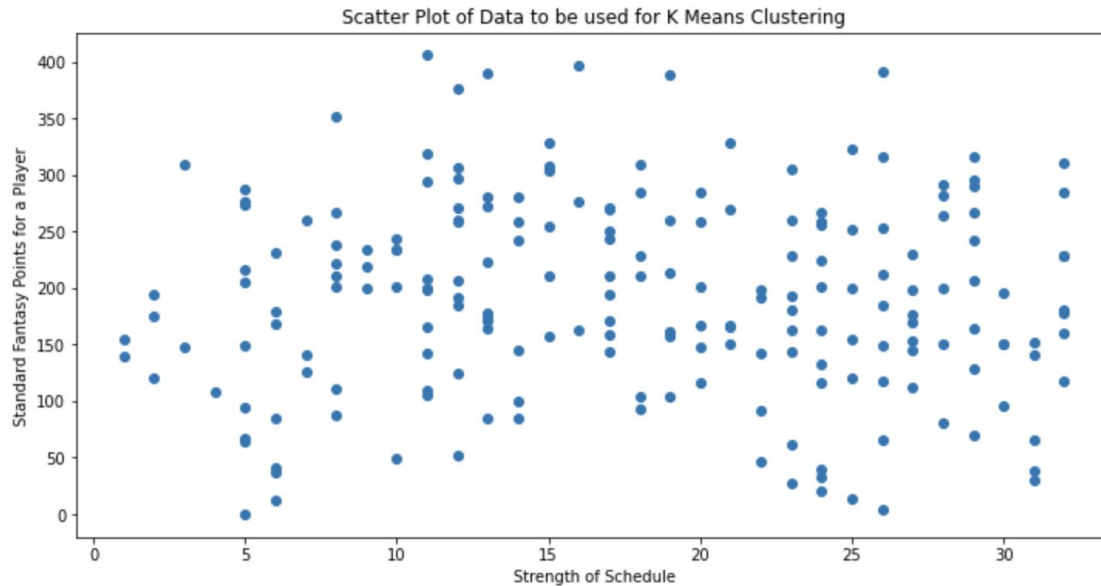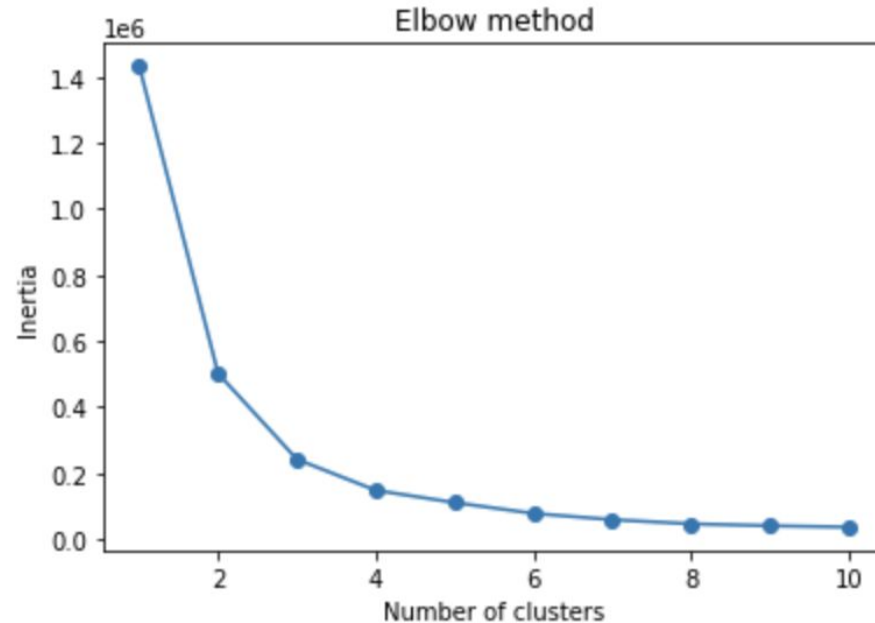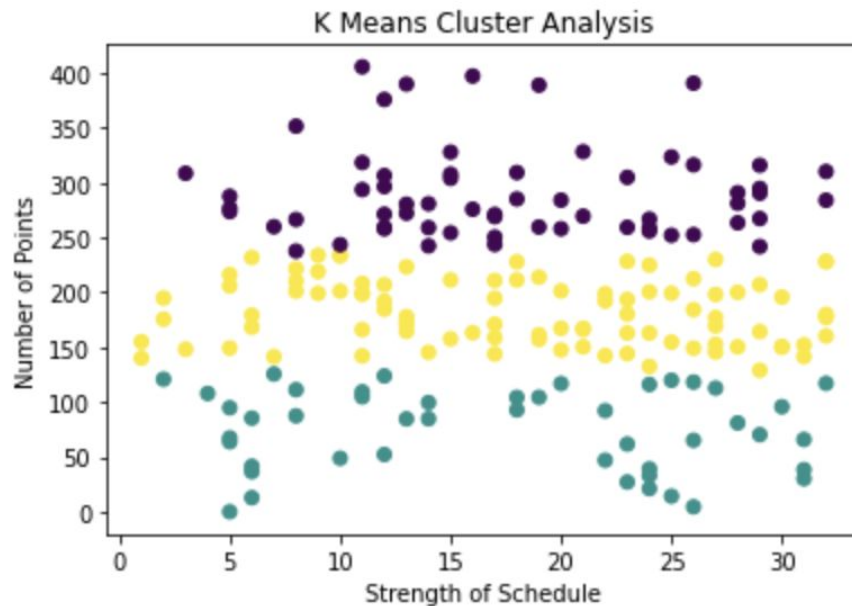
Strength of Schedule K-Means Clustering Analysis

Strength of Schedule K-Means Clustering Analysis

Points for Discussion

- Overall, our findings led us to define a series of heuristics that fantasy owners can use to help make predictions about players before a draft. They are as follows
  - From our exploratory analysis, players who start more than 12 games a season fall within a normal distribution of point values. Therefore, players who have shown a tendency to be injured, often have outlier low-point seasons. Fantasy owners should consider injury-proneness when drafting players.
  - From our analysis of weekly data, it was shown that running backs see increased performance as the season goes on. Fantasy owners can bank on better performances out of the running back position during later weeks.
  - From our annual analysis, we determined that players see a marked drop off in scoring their sophomore and sometimes 3rd season. Fantasy owners should expect 2nd and 3rd year players to underperform.
  - Finally, from our k means we observed that quality of a season is independent of opposing strength of schedule. Fantasy Owners should avoid letting strength of schedule play into their personnel decisions

Moving Forward from This Project

- The next step in using data mining to help fantasy owners, would involve using advanced stats to make predictions
    - Advanced stats are aggregate summary statistics of players that combine several individual stats.
    - Advanced stats are already widely used in basketball and in football they are becoming more accepted as a valuable metric for quantifying player performance.
    - Some examples of advanced stats are IAY (Intended Air Yards), IAYP (Intended Air Yards/Play), YAC (Yards After Contact), etc.
    - Similar analysis of advanced stats relative to age, week, and total points have a tremendous opportunity to learn more about fantasy football.

Related Work

- Goode, Hugh. "Data Driven Fantasy Football Lineup Optimization - Part 1." *Nycdatascience.com*, NYC Data Science Academy, https://nycdatascience.com/blog/student-works/data-driven-fantasy-football-lineup-optimization-part-1/.
  - Here is a completed project that used data science principles to optimize a daily-fantasy football lineup within standard salary cap constraints.

Questions?

Thank you for your time!