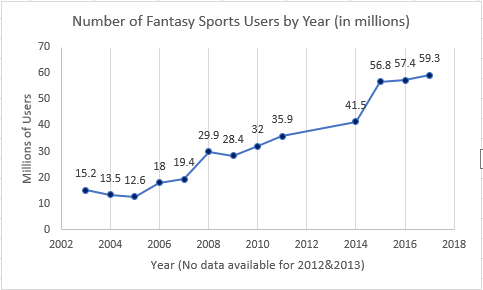
Statistics 2: Project 1

Sachin Chavan Sean Kennedy Kevin Thompson

## Introduction:

Fantasy sports are a big business. Generating nearly 7 billion dollars annually (cite) and with approximately 60 million players in the US/Canada (cite), fantasy sports – particularly football – have become as big as the sports that they mimic. Players can compete head to head in leagues across a wide range of providers (Yahoo, CBS, ESPN, DraftKings, DraftStreet etc.) – each with their own rulesets and stakes. Some are just friendly leagues set up between friends or coworkers, no real stakes other than bragging rights, others have significant monetary rewards for those that can get to the top of the leaderboard.



The explosion of weekly leagues over the course of the last few years has seen an already huge business get even larger. In the weekly cash leagues – each player is given a budget and drafts a completely new team every week. Each draft site uses a predictive model to set player salaries based on the number of points the model predicts that player to score. Budget constraints make it impossible to simply select the players that are predicted to score the most points – hence having a predictive model for which players will generate the best return on investment would be a huge advantage. In the following analysis we will attempt to build a position specific model that can accurately predict the number of points a given NFL player is likely to score in each week. In this analysis we are limiting our scope to quarterbacks.

## Data Description:

**\*All data for this project has been scraped from** [**https://fantasydata.com/**](https://fantasydata.com/) **unless otherwise noted.**

The training data that will be used for this model will consist of all QB data from the 2017 season with a minimum number of appearances by the QB equal to **8**. 8 was selected to capture as many of the consistent starting quarterbacks as possible without combing through the data by hand. The NFL season consists of 17 Weeks (16 playing weeks per team, 1 bye week) so any player that started a minimum of 8 games is likely to have more than 8 observations for the season. Since neither player nor team will be considered as factors, we have essentially created panel data for the population of quarterbacks that started played in a minimum of 8 games in the 2017 season.

The goal will be to select a set of features that analyzes the matchup between the quarterback and the defense matched up in the upcoming week. There are a variety of predictors that can be used to quantify the efficacy of an NFL quarterback. We will use season averages for various metrics (Completions, Touchdowns, Interceptions etc) to construct a set of features to represent a generic “quarterback”. We will assess normality for the lag variable across each week (since week/matchup is significant though we are not treating this as a time series problem for part 1).

The universe of features that we will select from are:

### Offense (QB):

* CumulativeAveragePassingYards
* CumulativeAveragePassingTouchdowns
* CumulativeAveragePassingInterceptions
* CumulativeAveragePassingRating
* CumulativeAverageCompletions
* CumulativeAverageCompletionPercentage
* CumulativeMaxPassingTouchdowns
* CumulativeMaxPassingYards
* CumulativeMaxPassingAttempts
* CumulativeMaxPassingRating
* CumulativeMaxCompletions
* CumulativeMaxPassYardsPerAttempt
* CumulativeMinPassingYards
* CumulativeMinPassingAttempts
* CumulativeMinPassingRating
* CumulativeMinCompletions
* CumulativeMinPassYardsPerAttempt
* CumulativeAverageFantasyPoints
* CumulativeMaxFantasyPoints
* CumulativeMinFantasyPoints
* **NextWeekFantasyPoints** (Target)
* HasThrownFor4TDsOrMore[[1]](#footnote-1)
* IsHomeNextWeek
* Team
* PlayerID
* Week
* ShortName

### Defense:

* AvgPassDefense [[2]](#footnote-2)
* AvgPointsAllowed
* AvgQBPointsAllowed
* OffensiveMatchup

We hope that the average columns will be a good indicator to the model for how the QB can expect to perform in each of the categories related to scoring fantasy points and that the min/max quantities will teach the model to assess risk level for each metric. It is important to note that all lag data is being used to predict **into the future** (i.e we are avoiding information leakage by using only trailing averages to predict future performance).

## Exploratory Analysis:

Football is a game of matchups and none is more important than the QB/Pass Defense matchup. In lieu of having a passing specific metric for the defense – we created a custom feature based on sacks, passes defended, quarterback hits and interceptions. Our goal will be to create a model that can learn some of the inner structures of the QB/PD matchup. Clearly, this is an ambitious goal. Sports in general are subject to a variety of hard to capture factors (weather, location, injuries, individual matchups of players with few statistics such as lineman etc.) and football - in particular - is notoriously hard to predict given the number of players on the field at any given time (22) and the complexity of the game itself given that all players can interact during a play.

### General EDA[[3]](#footnote-3):

###### Training Set:

* + - 2017 Season – filtering out data in weeks 1 and 2 (rolling averages were defined on a 3 period window) and week 17 (no target variables to predict) then filtering to ignore any QB that did not play in at least 8 games – there were 326 observations in the training set.

###### Test Set:

* + - 2018 Season – the test set contains all observations from the 2018 season subject to the same constraints as the 2017 season. There are 355 observations in the test set.

###### ANOVA:

* The ANOVA test set consists of 3 years of Fantasy Scoring data for quarterbacks that made at least 10 home/away appearances. There are 1060 observations in this set.

## Model Selection:

###### Type of Selection:

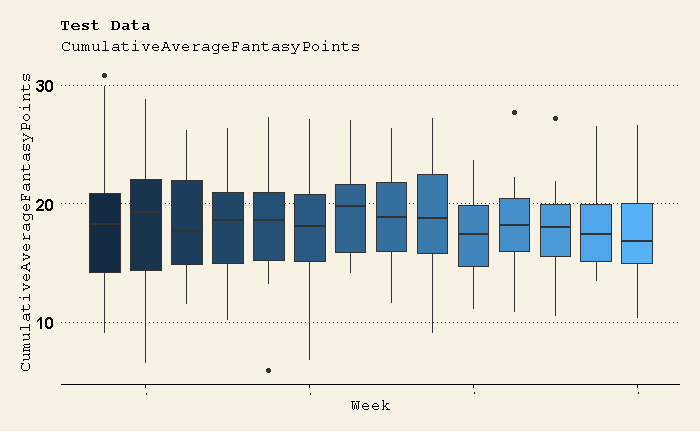
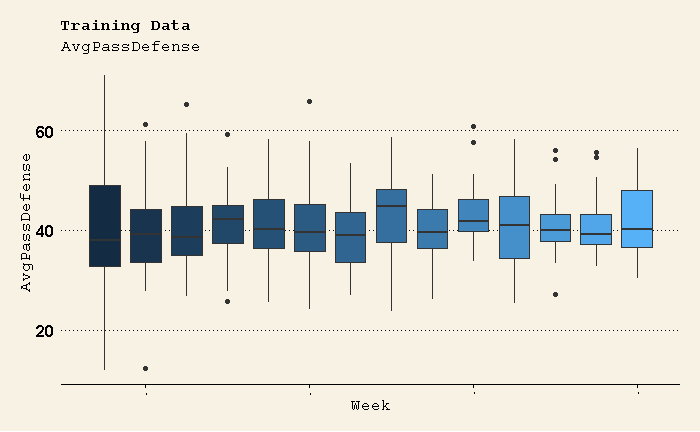
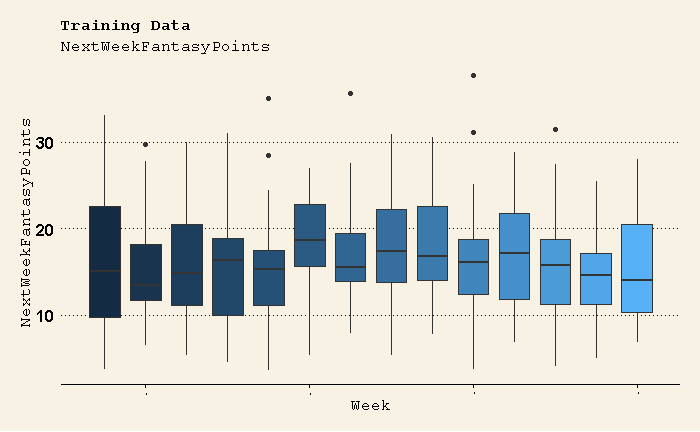
Model selection will be performed via a multitude of different approaches. Standard linear regression techniques will serve as the basis.Ultimately – we will select from the most predictive model generated by the following:

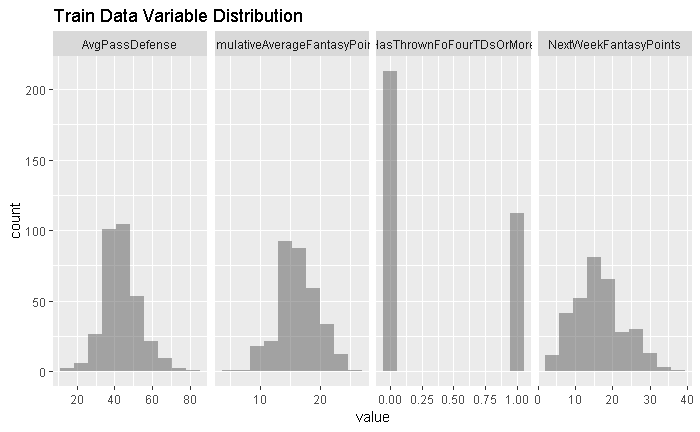
* General OLS Regression
* Ridge Regression
* LASSO Regression

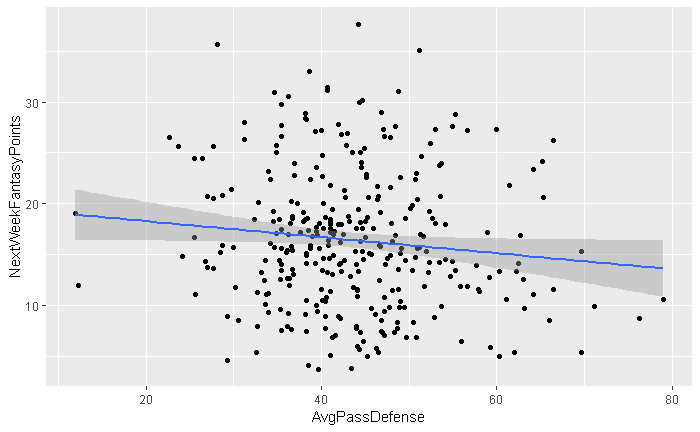
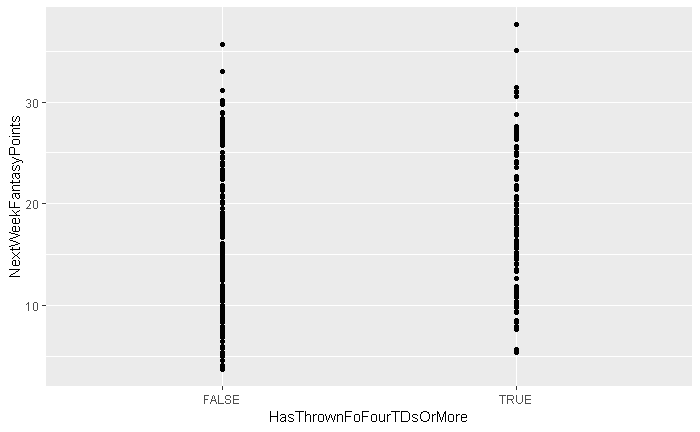
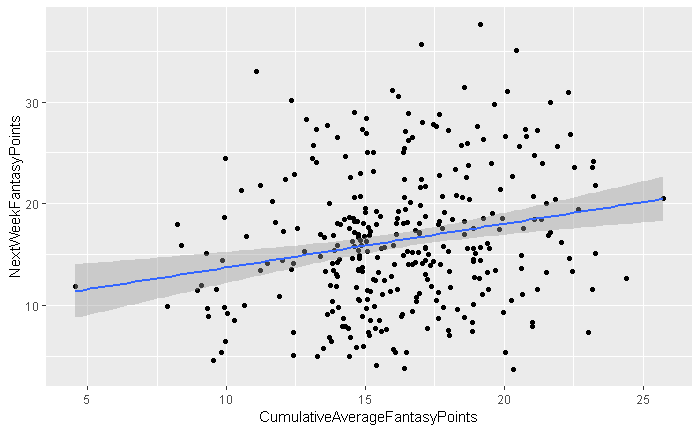
###### Checking Assumptions

* Normality/Linear Trend/ heteroscedasticity/Independence

Train Set:[[4]](#footnote-4)



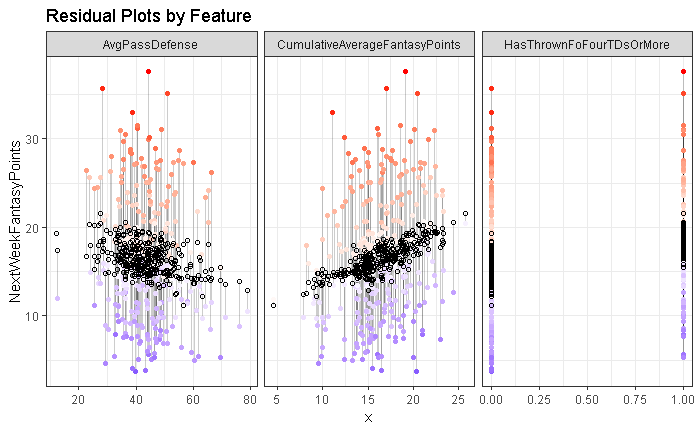


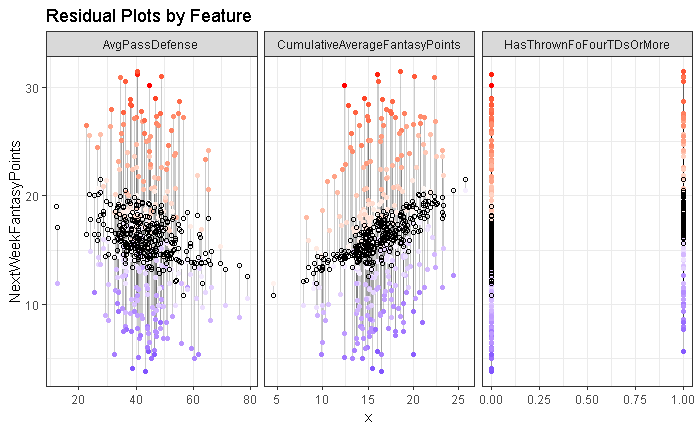
  

While there are clearly some outliers in any given week – there doesn’t seem to be any glaring violations of normality or heteroscedasticity across the dependent/independent variables.

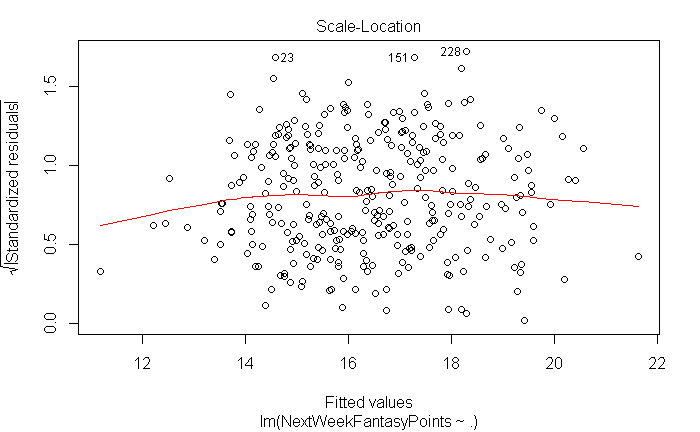
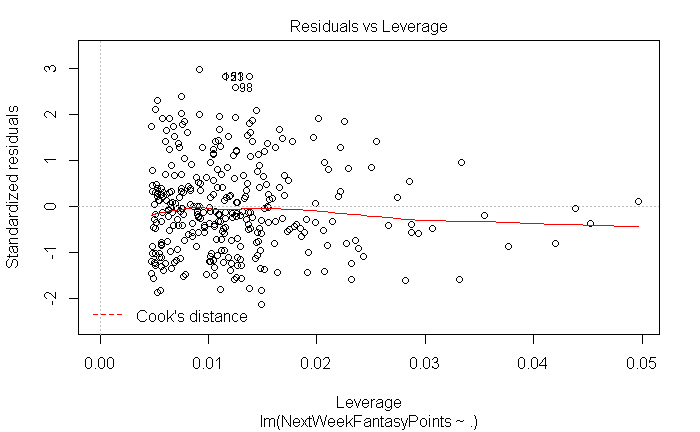
Independence is tricky in this situation. Theoretically, this is time series data – but due to the nature of the sport and the varying matchups week over week, we will proceed as if each game is an independent observation since the participants and interactions in each contest are very much unique.

###### Residual Plots:[[5]](#footnote-5)

[[6]](#footnote-6)

[[7]](#footnote-7)

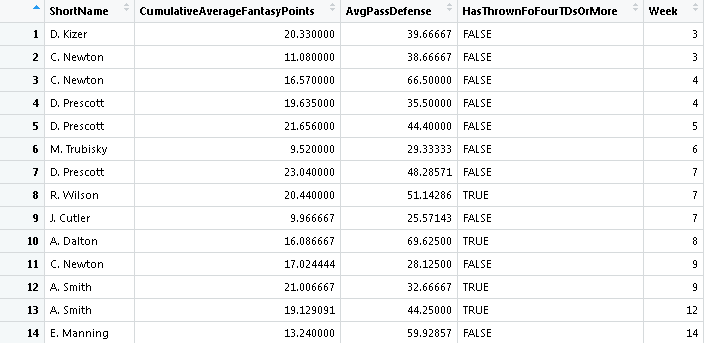
###### Influential point analysis (Cook’s D and Leverage)

Defining an influential observation as the following:

* RStudent > 3
* CooksD > 4/number of obs

Filters out the following 14 observations



Removing these observations improved R-squared and didn’t significantly impact our parameter estimates. We will proceed with the scrubbed data going forward.

###### Compare Competing Models:

Parameter Interpretation

Interpretation **Required**

Confidence Intervals **Required**

Final conclusions from the analyses of Objective 1 **Required**

In addition to overall conclusions, feel free to include additional insights or concerns gleaned from the analysis. What needs to be done next or how could we do it better next time?

Addressing Objective 2

State what route you are going to take 2way ANOVA or Time series and summarize the goal. **Required**

**-TS**

Main Analysis Content **Required**

This will depend on the route you take. I’m leaving it open here to see what you do.

**--ARIMA / lag analysis ACF/PACF**

Conclusion/Discussion **Required**

The conclusion should reprise the questions and conclusions of objective 2.

Appendix **Required**

Well commented SAS/R Code **Required**

Graphics and summary tables (Can be placed in the appendix or in the written report itself.)

1. Custom metric created in EDA process - see appendix [↑](#footnote-ref-1)
2. Custom metric created from domain understanding – see appendix for detailed calculation and comparison to known pass defense rankings [↑](#footnote-ref-2)
3. Graphical analysis and further summary data can be found in the appendix [↑](#footnote-ref-3)
4. Test Set analysis can be found in the appendix [↑](#footnote-ref-4)
5. Further analysis including Q/Q plots and pair plots can be found in the appendix [↑](#footnote-ref-5)
6. Residuals **PRE** outlier removal [↑](#footnote-ref-6)
7. Residuals **POST** outlier removal [↑](#footnote-ref-7)