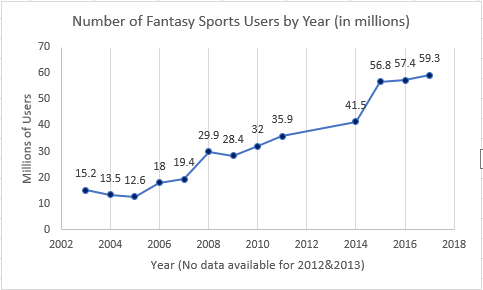
Statistics 2: Project 1

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:

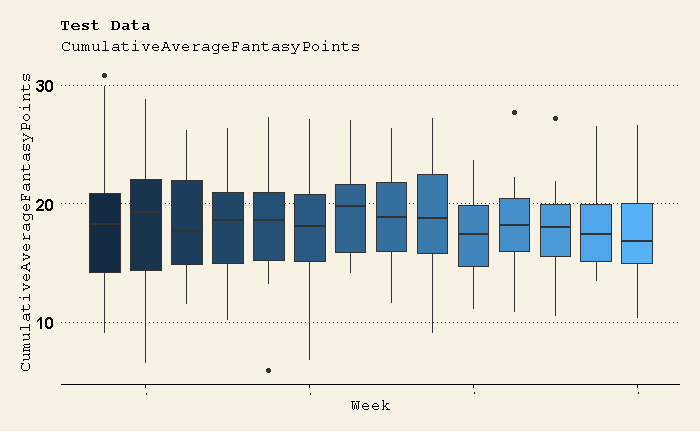
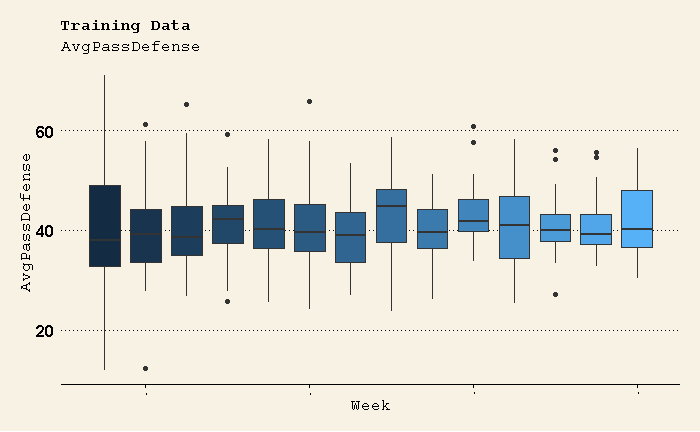
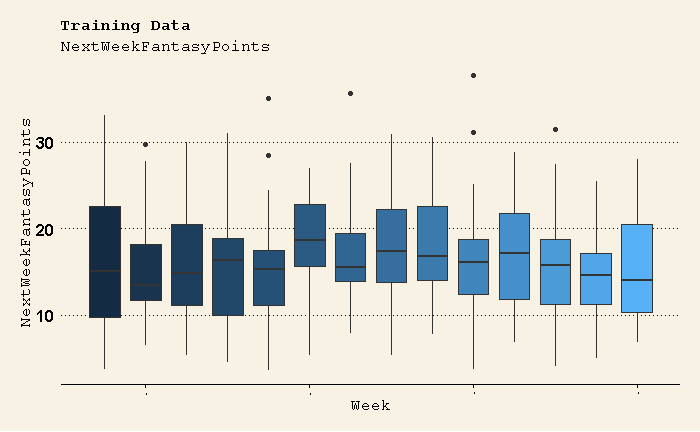
Model selection will be performed via a multitude of different approaches. Standard linear regression techniques will serve as the basis.Ultimately – we will select from the most predictive model generated by the following:

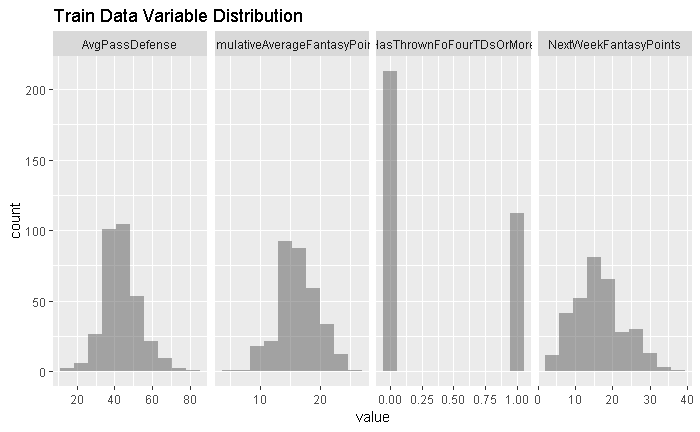
* General OLS Regression
* Ridge Regression
* LASSO Regression

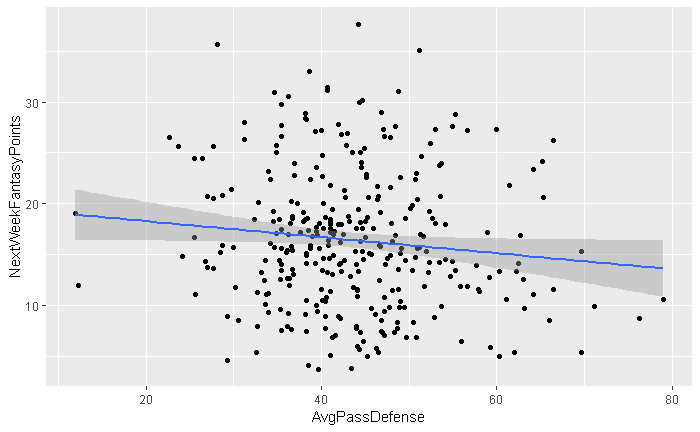
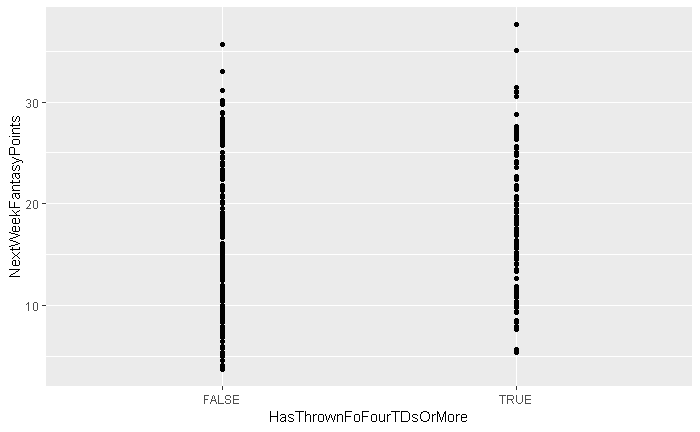
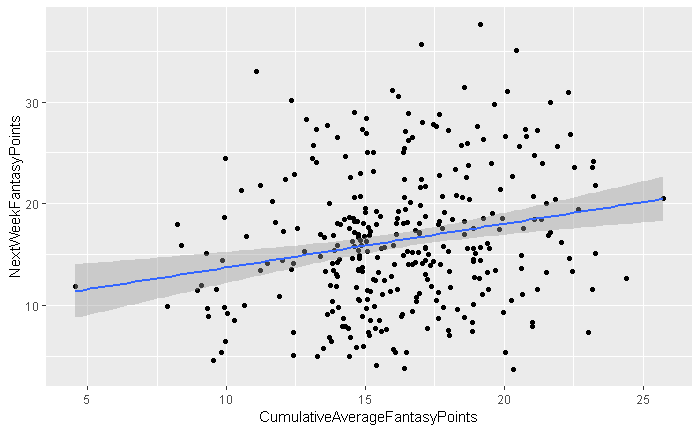
###### Checking Assumptions

* Normality/Linear Trend/ heteroscedasticity/Independence

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



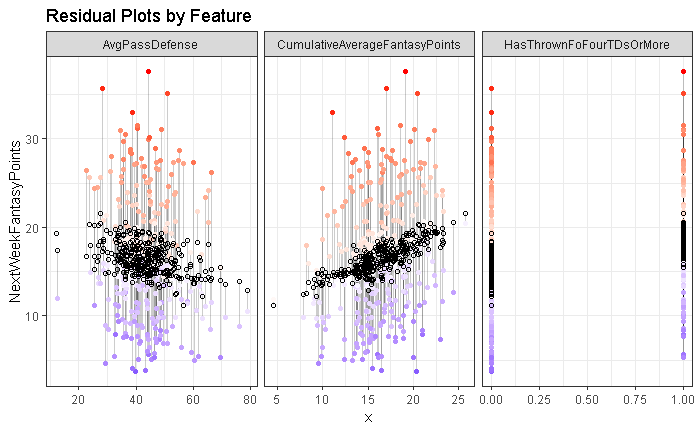
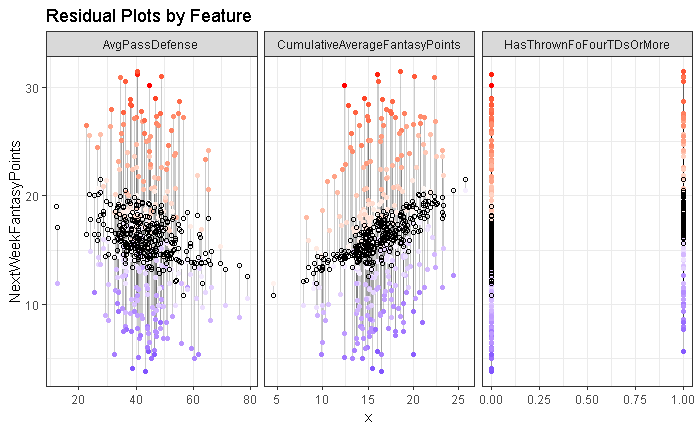


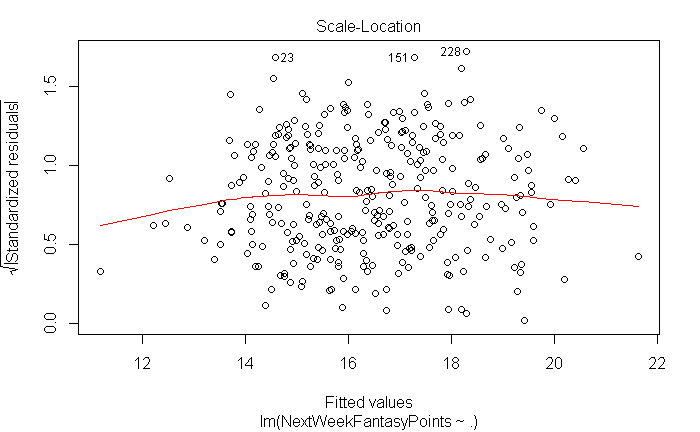
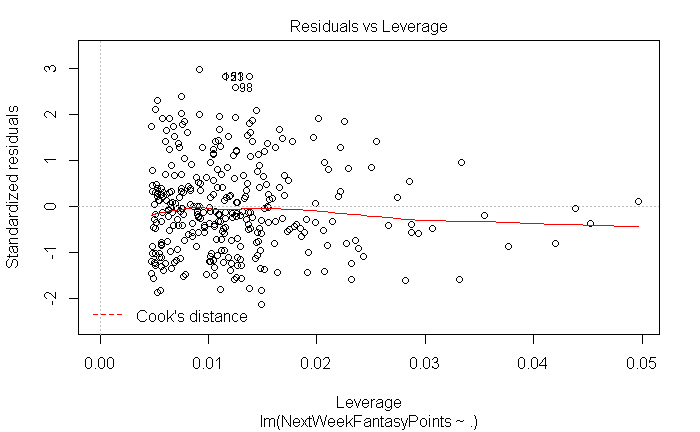
While there are clearly some outliers in any given week – there doesn’t seem to be any glaring violations of normality or heteroscedasticity across the dependent/independent variables. Linear trends seem strong and correct directionally.

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)

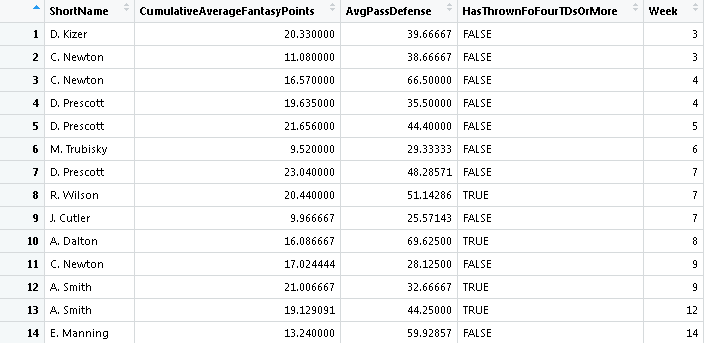
###### Influential point analysis (Cook’s D and Leverage)

Defining an influential observation as the following:

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

Filters out the following 14 observations

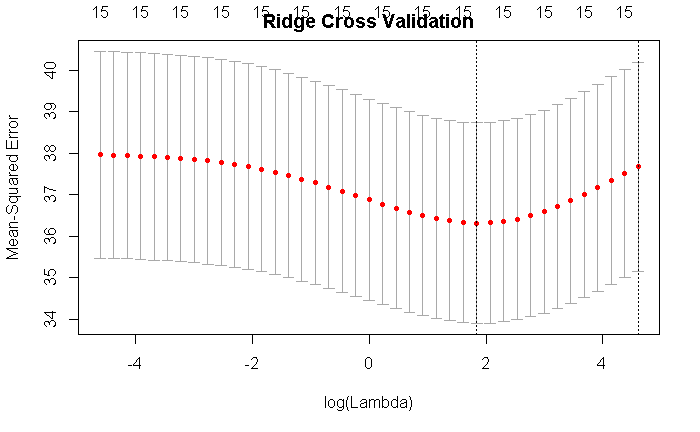
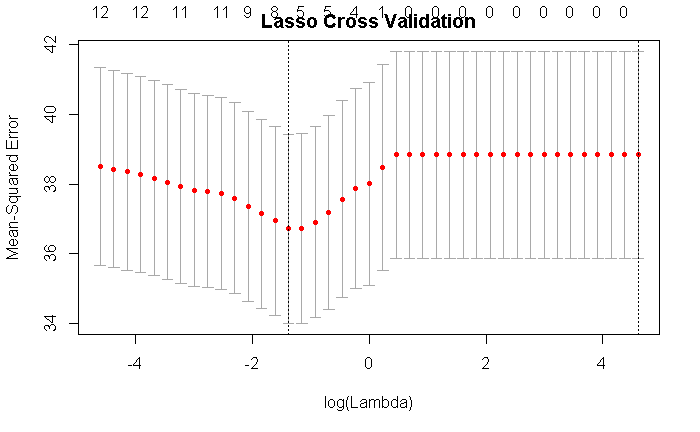


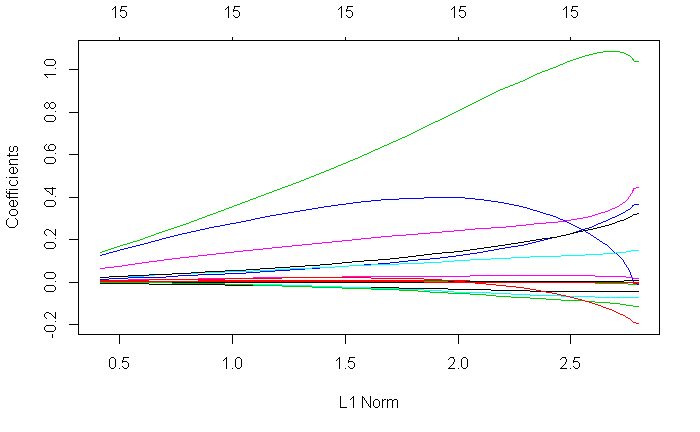
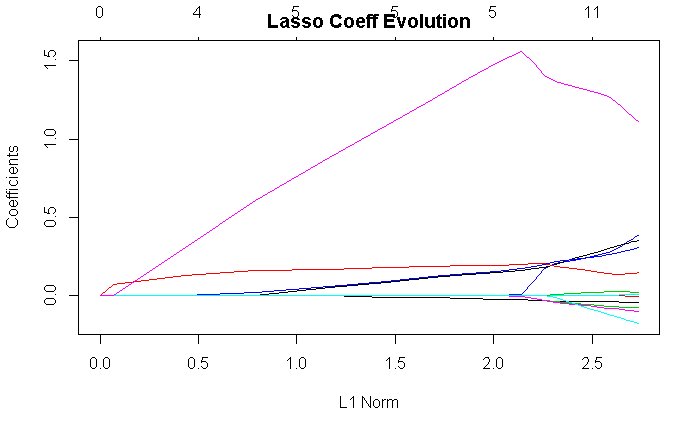
Removing these observations improved R-squared and didn’t significantly impact our parameter estimates.

###### Compare Competing Models:

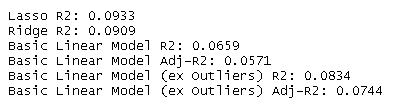
Performing a LASSO and Ridge regression to select variables of importance. We will proceed with the scrubbed data going forward. A 40-fold cross validation model with a lambda sequence of logarithmic steps (base 10) from 100 to 0.01. And will be scored on best MSE.

#### CV Results:

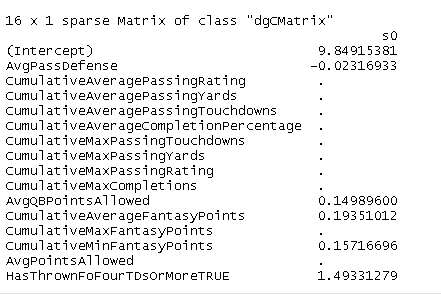




As expected, the LASSO procedure set 10 of the model coefficients = 0 after step 15 of the cross -validation process – whereas the Ridge procedure selected the model after iteration 30 of the CV process. Ultimately – the LASSO procedure kept 5 of the 15 potential variables - three of which were custom metrics that we created (AvgQBFantasyPointsAlllowed, AvgPassDefense and HasThrownForFourTDsOrMore) and the other two were CumulativeMinFantasyPoints (a good proxy for the expected player floor) and CumulativeAverageFantasyPoints (indicative of a players most likely output given recent history). The ridge regression never shrinks coefficients to zero, but it did select similar values to the parameters selected in the standard linear regression and LASSSO CV procedure.



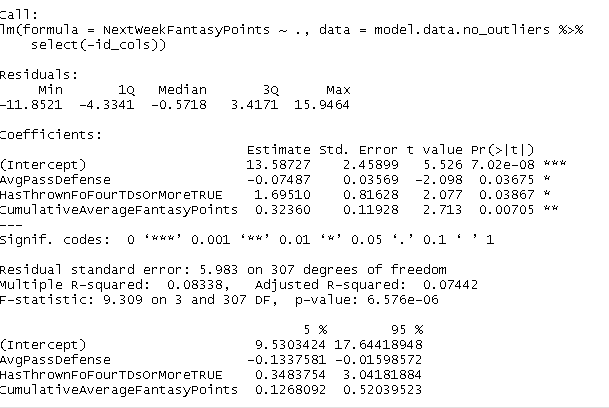
#### Lasso



#### Ridge

#### 

#### OLS



All things being equal – simpler is typically better. By using the suggestions from the LASSO selection process, the OLS model ultimately selected for predictions was limited to three predictors:

* AvgPassDefense
* HasThrownForFourTDsOrMore
* CumulativeAverageFantasyPoints

###### Parameter Interpretation

At an alpha level of .05, all parameters showed statistical significance. Given the intercept of 13.58, we can interpret this as an expectation for a QB facing an AvgPassDefense = 0, having a CumulativeAverageFantasyPoints = 0 and having never thrown for 4 or more TDs in a game to score roughly 13.6 points (9.5-17.6). This isn’t a realistic situation since there are no observations in the set where AvgPassDefense or CumulativeAverageFantasyPoints = 0.

The 1 parameter (AvgPassDefense) can be interpreted as a reduction in expected fantasy points of -.075 pts for each unit increase of AvgPassDefense

The 2 parameter (HasThrownForFourTDsOrMore) can be interpreted as an adjustment to the intercept (since it’s a Boolean param) of approximately 1.7 pts for QBs that have thrown for 4 TDs or more at least once in the season (prior to the week being analyzed)

The 3 parameter (CumulativeAverageFantasyPoints) can be interpreted as an increase in expected fantasy points of 0.323 pts for each unit increase of CumulativeAverageFantasyPoints

Hence, a quarterback that faces a defense with an AvgPassDefense = 20, having thrown for at least 4 TDs once and averaging 20 fantasy points per game, would be expected to score 20.26 points in our OLS model.[[8]](#footnote-8)

###### Confidence Intervals:

Confidence intervals for each of our parameters showed strong statistical significance. None of the parameters contained 0 as an estimate at the 95% CL threshold (see table at end of previous graphic for U/L limits for each).

###### Final conclusions:

As stated at the outset – predicting anything in the world of sports is difficult and fraught with potential pitfalls. By leveraging the power of LASSO model selection, we were able to limit our feature set to those that were most relevant to out target problem. While there are infinitely many new variables that we could add to our model (more matchup/positional specific metrics), we were happy to capture a portion of the variance. We were also encouraged by the fact that our custom metrics performed very well and showed a good understanding of the domain. Continuing this type of analysis and adding more predictors seems like a fruitful endeavor. As is always the case when adding features, multicollinearity should be avoided if possible – see the appendix for an analysis of the covariance of our custom predictors.

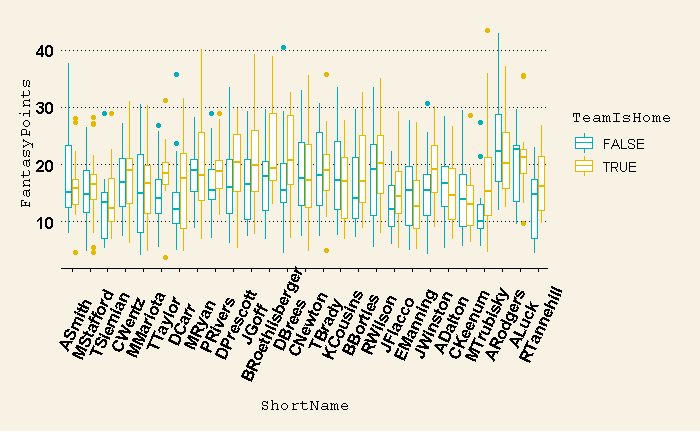
By far, the biggest issue we encountered in our analysis, was the underestimation of the amount of noise compared to signal in our dataset. While we believe that we can capture a large portion of that noise by adding new predictors, the question of how many predictors that would require is challenging.

## ANOVA: Who likes home cooking?

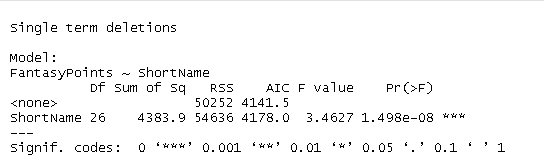
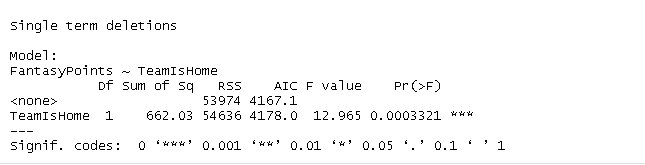
Home/away splits are often an important predictor for quarterback performance. Some QBs excel in their home stadium – riding the energy of their home crowd to better results or preferring the climate/conditions of their home stadium to those they are not familiar with. Others do not perform as well at home – perhaps feeling increased pressure to perform in front of home fans or not feeling comfortable in the climate in their home city (there are a variety of teams that play games in predominantly hot/cold weather climates and football is a sport where many games are played in deep winter). To distinguish if there is a significant difference in home/away splits and which players are helped/hurt the most by playing on the road, we will carry out a sequential ANOVA analysis on TeamIsHome and ShortName (i.e player).

###### Main Analysis Content**:**

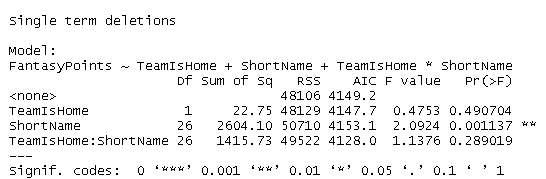
#### EDA:



Examining box plot of the QBs, shows that there is reason to believe that at least a few of them are significantly impacted by home/away splits. A RSS F-Test (type-3) shows the following:



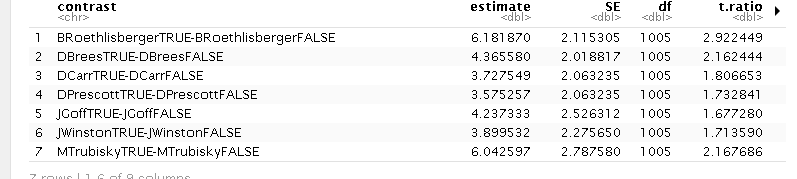
Running the test with interactions yields:

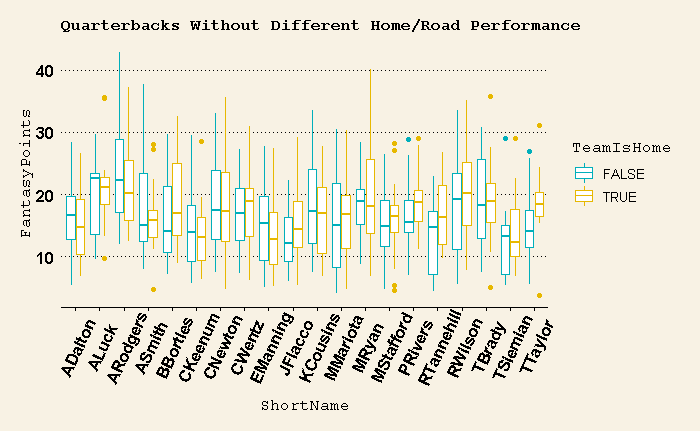
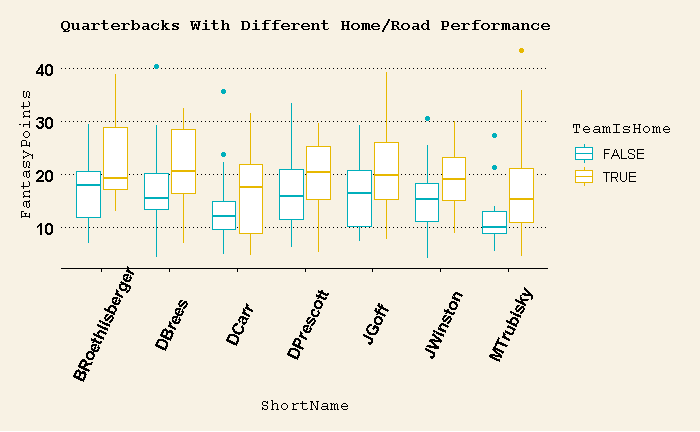


There seems to be an underlying relationship here, though the full model seems to be slightly confused as to which QBs are affected the most. To address this question – we will run a series of custom contrasts for each QB on their home/away splits to search for significant differences in their average fantasy points.

###### Conclusion/Discussion:

After running a contrast[[9]](#footnote-9) test to see which QBs had significant differences in home/road splits, there were 7 quarterbacks that were identified as being outperformers at home:





A comparison of two quarterbacks - B Roethlisberger and M Trubisky – is contained in the appendix. In Roethlisberger’s case, his 2016 season had a drastic impact on his home/away splits whereas Trubisky was poor more consistently on the road - his high QBR on the road in 2018 is a result of a high completion percentage – a stat that is not rewarded in fantasy leagues. Overall, the model did a fair job of segregating quarterbacks that perform well at home and not quite as well on the road.

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)
8. Prediction/Confidence intervals charts can be found in appendix [↑](#footnote-ref-8)
9. See appendix/R-code for contrast details [↑](#footnote-ref-9)