Springboard Capstone 2 Milestone Report 2: Analysis of NBA Dataset

Kaggle Dataset Repository

Final Report by: Michael Thabane

June 25, 2020

Contents

1 Introduction

1.1 The Business Problem

In the last 30 years the NBA has been one of the fastest growing sports in the world. With the advancement of sports analytics and the prevalence of gambling there is a high demand for player statistic predictions. The business problem we would like to solve is creating a prediction model for player stats, in particular points; using previous game data. We will also explore specific players to find potential variables that have a direct effect on the points a player scores. The plan is to model the data using time-series forecasting techniques to predict the number of points a player will score in an upcoming game.

1.2 Executive Summary

2 Approach

2.1 Data Acquisition and Wrangling

The dataset used for this project was acquired from Kaggle but was originally scraped from the NBA statistics website using the NBA API. The dataset consists of four different spreadsheets. The first spreadsheet (games.csv) contains team boxscore data which consists of total team statistics. The second spreadsheet (game_detail.csv) contains the player boxscore data which consists of individual player statistics for each game. The third spreadsheet (players.csv) contains player information for which team they play for. The fourth spreadsheet (ranking.csv) contains the nba standings for every day in the each season. In order to have all the relevant information in the same spreadsheet some data wrangling was need to get column into the game detail spreadsheet. First, we want to include the team stats for any given game to be in the player boxscore data. To achieve this we merged the team data from the games spreadsheet into the game detail spreadsheet but only included the team stats of the team the player is playing for. Then, we will manipulate data to create any variables that may have influence on the outcome of a players points. I have created the running average to include the players average over the previous three and five games.

2.2 Storytelling and Inferential Statistics

2.2.1 Descriptive Analysis

With the dataset containing information since 2003 it makes more sense to look into players that have had more of an impact scoring for their teams recently. Below is a graph which shows the top ten scorers in the last two seasons (2018 and 2019).

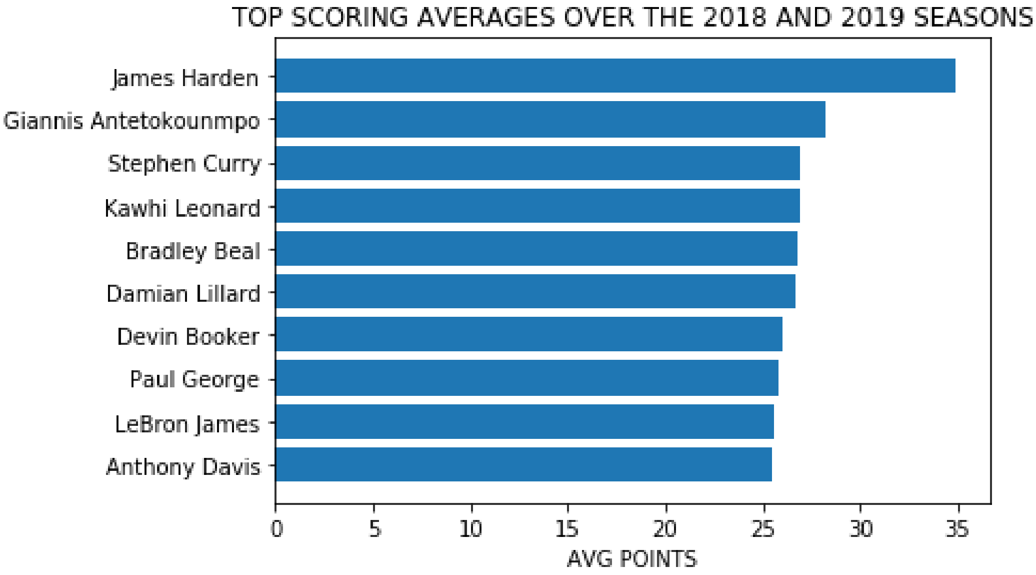

Figure 1: Top 10 Scorers Over 2018 and 2019 Seasons

From above you will notice that there are no centers in the top 10 list of scorers. Based on the recent NBA rule changes the game has strayed from making centers the focal point of an offense. Now the center position has been watered down with a lot less dominant big man in the league. Then new style of play has more spread out offenses and less focus on dominant interior play or in and out basketball. To examine this we can at look distribution of points scored by starters from the positions forward, guard and center and look to see if we can find this trend.

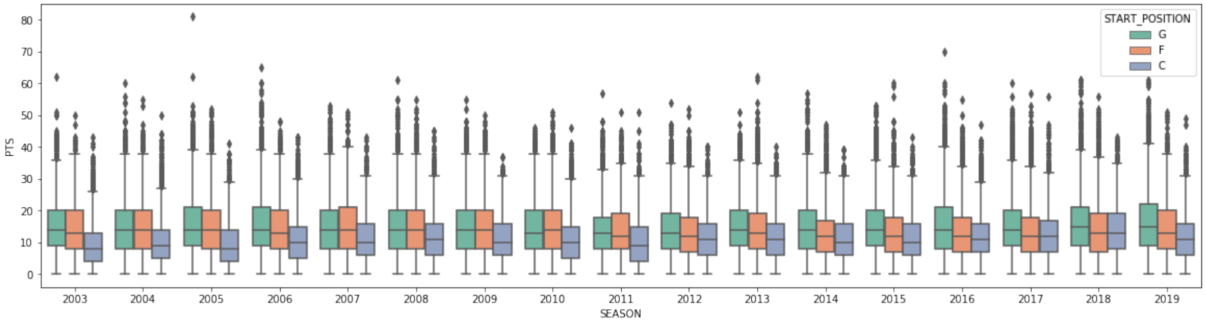

Figure 2: Boxplot of Starters Points

From the plot above it does not seem to show this trend as centers are scoring at a similar rate over the last 20 years. However if we plot the number of three-point field goals attempted it might show that interior players have gone from never shooting threes to spreading the floor.

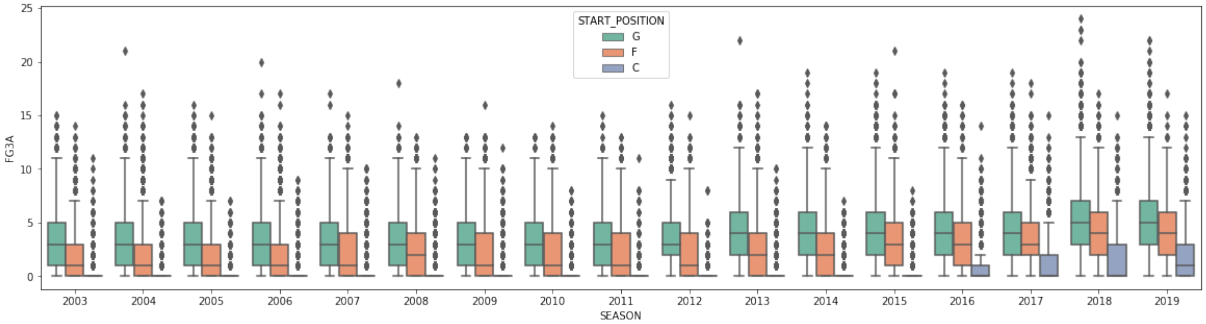

Figure 3: Boxplot of Starters 3PT Field Goals Attempted

From the plot above we can see that in the last four years starting centers have begun to spread the floor by shooting more threes. If we had the data of points scored in the paint that should show a trend of less points being scored in the paint by centers as the game has evolved.

From Figure 1 we can see that James Harden has had the highest scoring average. However, this has not always been the case throughout his career. His role for the OKC Thunder which is the team that drafted him was to provide scoring off the bench. Now his role with his current team is to be the focal point of the offense and provide scoring and leadership. To further investigate this we will use a violin plot to examine his point distribution throughout his career.

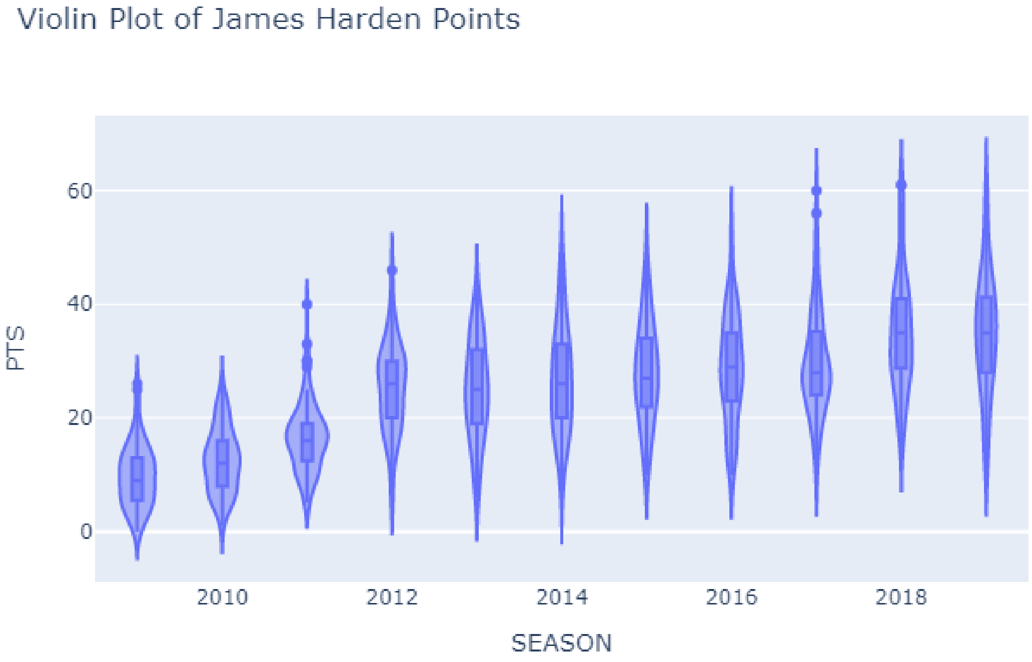

Figure 4: Violin Plot of James Harden

The plot above shows that when James Harden was traded from OKC to Houston at the start of the 2012 season his point average and distribution made a significant increase. This shows how player movement whether it is through free-agency or trades can make a significant impact on the volume of scoring a player will provide for his team. However, it is possible for a player's role to change with the team over time without a trade occurring. This can be attributed to both player development as well as team needs change from season to season. This can be perfectly illustrated by Pascal Siakam who has both developed as a player and has had an increased role due to the departure of the teams star player. A violin plot of Siakam's points help illustrate this.

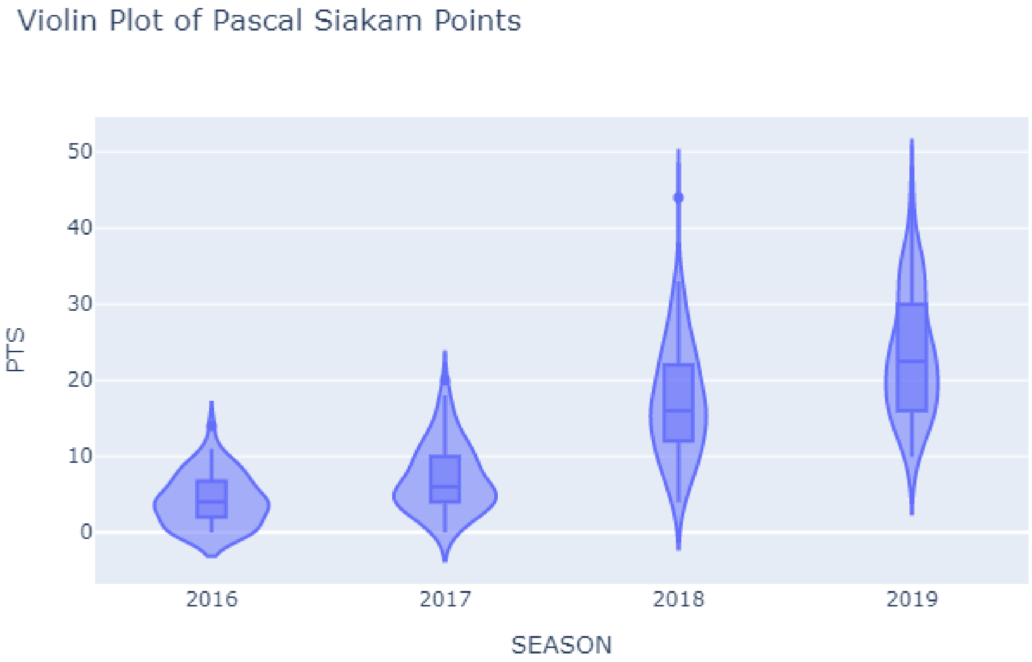

Figure 5: Violin Plot of Pascal Siakam Points

The above plot shows a big jump in point production from the 2017 to 2018 and slight increase from 2018 to 2019. However, the violin plot alone will not give you an idea of what has caused the increase. I will use a violin plot of Siakam's minutes to illustrate how his starting role change in the 2018 season impacted his points compared to how Kawhi's departure in the 2019 season has affected his point distribution.

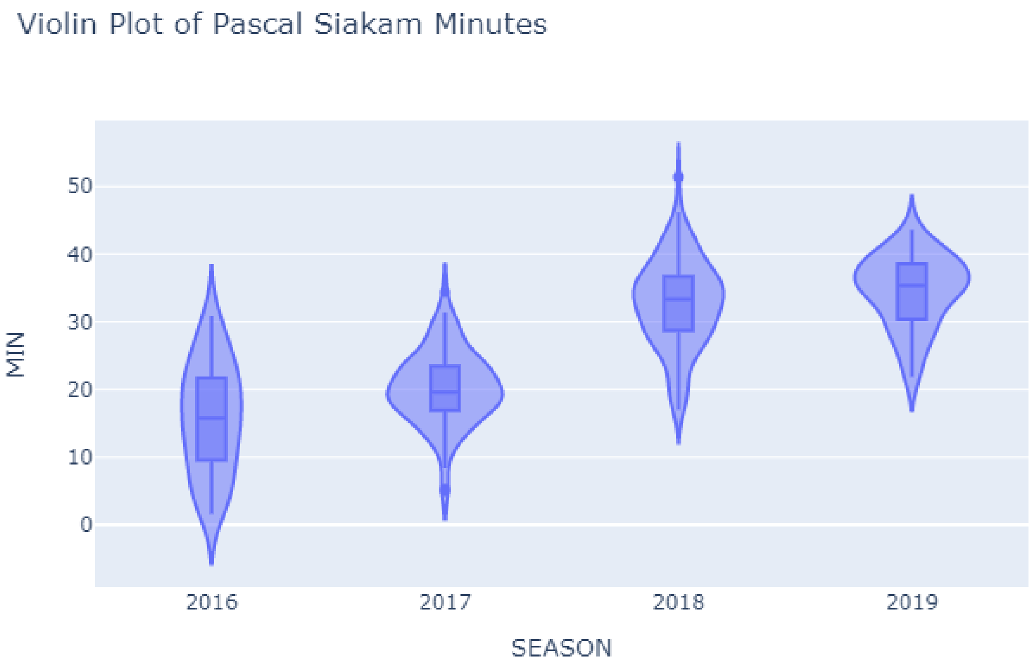

Figure 6: Violin Plot of Pascal Siakam Mins

We can see from the plot above that Siakam's minutes changes significantly between the 2017 and 2018 seasons. However, from 2018 to 2019 his minutes did not increase very much but his point distribution did. This was due to his impact in the offense changing with the departure of Kawhi Leonard.

2.2.2 Hypothesis Testing

The main objective of the hypothesis tests we will run in this section is to test whether the means of two groups are equal. We would like to find out if there is a significant difference in the average points scored between positions as well as testing the scoring averages of playoff performing teams vs. non playoff teams. Below is a boxplot showing the distribution of points scored at each position.

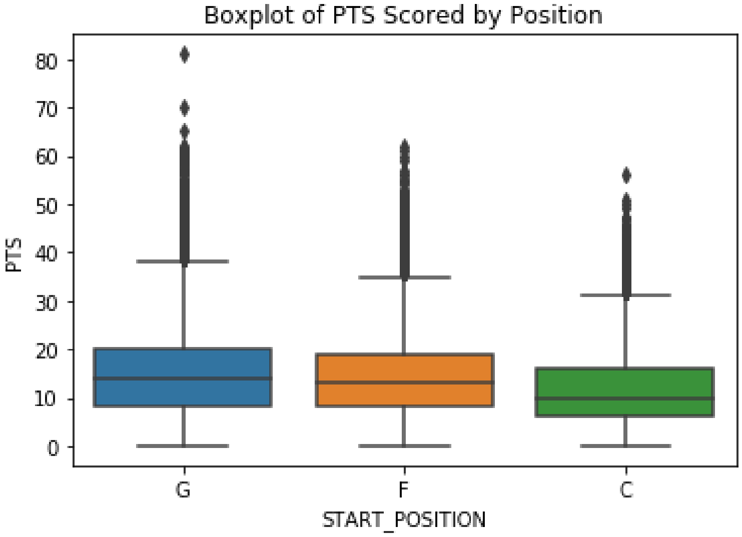

Figure 7: Boxplot of PTS Scored by Position

The boxplot only shows that centers clearly score less point than both guards and center. However, we will test the following hypotheses: 1) H0 = Are the average points scored by centers equal to the average points scored by guards. 2) H0 = Are the average points scored by centers equal to the average points scored by forwards. 3) H0 = Are the average points scored by guards equal to the average points scored by forwards. The results of the hypotheses tests are as follows: 1) Degrees of Freedom = 132821 , p-value = 0, test statistic = -76, critical value = 1. From this result we can conclude that there is a significant difference between average points scored by centers and the average points scored by guards; H0 is rejected since p-value less than 0.05. 2) Degrees of Freedom = 132823 , p-value = 0, test statistic = -58, critical value = 1. From this result we can conclude that there is a significant difference between average points scored by centers and the average points scored by forwards; H0 is rejected since p-value less than 0.05. 3) Degrees of Freedom = 177098

, p-value = 0, test statistic = 21, critical value = 1. From this result we can conclude that there is a significant difference between average points scored by guards and the average points scored by forwards; H0 is rejected since p-value less than 0.05.

The difference between points scored by players who play for playoff teams vs players who play for non-playoff will now be explored. First, we will look at a boxplot of points scored by players on play-off and non-playoff teams.

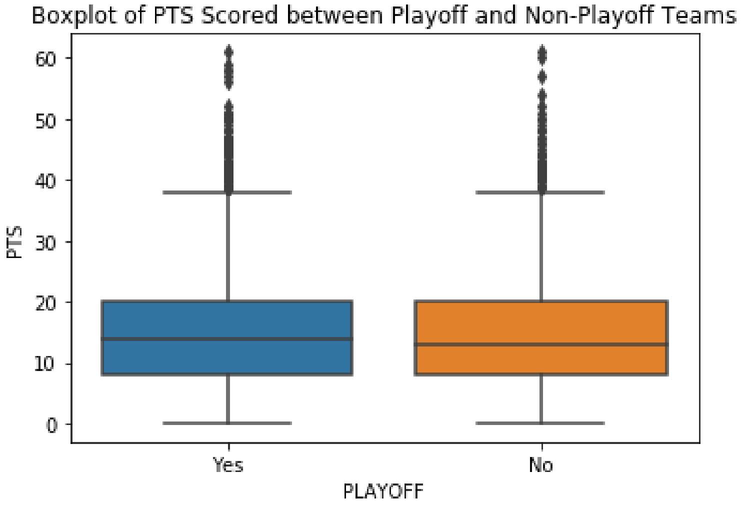

Figure 8: Boxplot of PTS Scored between Playoff and Non-Playoff Teams

From the boxplot above it is hard to tell if there is a difference between points scored by players who play for playoff teams vs players who play for non-playoff. Running a hypothesis test on H0: the average points scored for players who play for playoff teams vs the average points scored for players who play for non-playoff gives the following result: Degrees of Freedom = 13778 , p-value = 0, test statistic = 2, critical value = 1. From this result we can conclude that there is a significant difference between average points scored by players playing for playoff teams and the average points scored by players playing for non-playoff teams. However, this does not tell us which a playoff teams effect on specific players. To further examine this we will look at the points scored by each position between non-playoff and playoff team players. First by plotting the boxplots of we can see if the distribution of points scored at each position is different based on if that player plays for a playoff team.

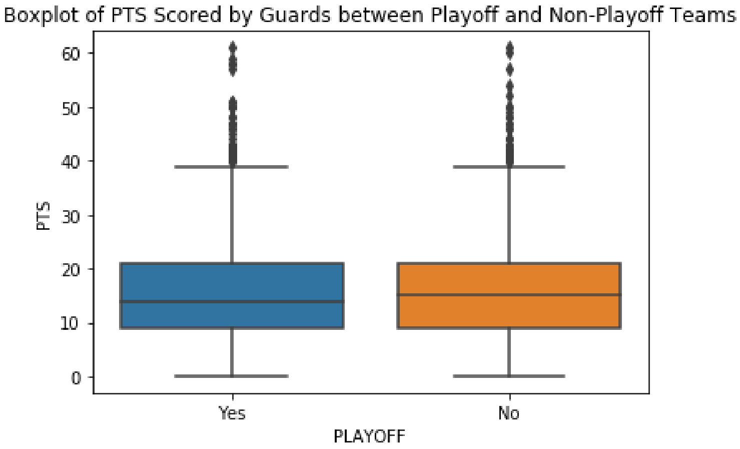

Figure 9: Boxplot of PTS Scored by Guards between Playoff and Non-Playoff Teams

It is hard to tell from this plot if there is a difference but it does seem that guards on not playoff teams score slightly more than guards playing for playoff teams. Below is the boxplot for points scored by playoff and non-playoff forwards.

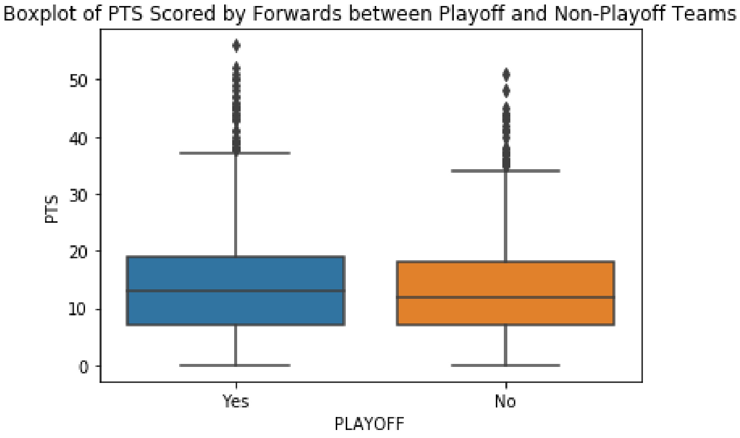

Figure 10: Boxplot of PTS Scored by Forwards between Playoff and Non-Playoff Teams

From this plot it seems that forwards playing for playoff teams score more than forwards playing for non-playoff teams. Below is the boxplot for points scored by playoff and non-playoff centers. the boxplot for points scored by playoff and non-playoff forwards.

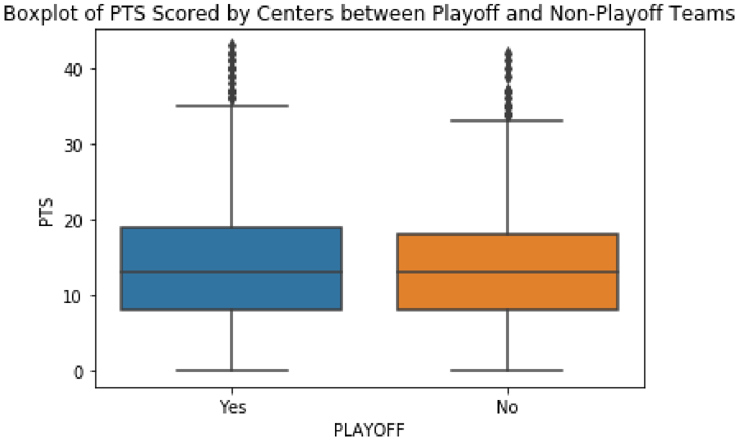

Figure 11: Boxplot of PTS Scored by Centers between Playoff and Non-Playoff Teams

It is hard to tell from this plot if there is a difference but in order to verify the answer to these questions we will test the following hypotheses: 1) H0: The average points scored by playoff guards is equal to the average points scored by non-playoff guards. 2) H0: The average points scored by playoff forwards is equal to the average points scored by non-playoff forwards. 3) H0: The average points scored by playoff centers is equal to the average points scored by non-playoff centers. The results of the hypotheses tests are as follows: 1) Degrees of Freedom = 5510, p-value = 0, test statistic = -1, critical value = 1. From this result we can conclude that there is a significant difference between average points scored by playoff guards and the average points scored by non-playoff guards; H0 is rejected since p-value less than 0.05. 2) Degrees of Freedom = 5510 , p-value = 0, test statistic = 3, critical value = 1. From this result we can conclude that there is a significant difference between the average points scored by playoff forwards and to the average points scored by non-playoff forwards; H0 is rejected since p-value less than 0.05. 3) Degrees of Freedom = 2754 , p-value = 0, test statistic = 1, critical value = 1. From this result we can conclude that there is a significant difference between average points scored by guards and the average points scored by forwards; H0 is rejected since p-value less than 0.05.

2.2.3 Baseline Modeling

In the baseline modeling section we would like to predict the number of points a player will score in a game. We will first start by examine how linear regression model performs when predicting the number of points Lebron James or Giannis Antetokounmpo will score. To do this we will use variables that were created in the Data Wrangling Notebook. These variables include the averages during the last five games of the following statistics: Points, Minutes, Field Goals Made, Field Goals Attempted,Field Goals Percentage, 3PT Field Goals Made, 3PT Field Goals Attempted, 3PT Field Goal Percentage, Free-Throws Made, Free-Throws Attempted, Free-Throw Percentage, Steals, Player Plus-Minus, Total Team Points. The dataset also includes the average points scoired in the last 3 games, the number of days off between the last game, and the number of years the player has been in the league. First we will create a subset of the boxscore data and only focus on predicting Lebron James' points. Since the 2019 season has not been completed due to the corona virus pandemic we will split the data in which the 2003-2017 seasons will the training set and the 2018 season will be used as the test set. The results of our initial model are below:

Dep. Variable:	PTS	**R-squared:**	0.044
Model:	OLS	**Adj. R-squared:**	0.026
Method:	Least Squares	**F-statistic:**	2.409
Date:	Wed, 24 Jun 2020	**Prob (F-statistic):**	0.00152
Time:	22:14:37	**Log-Likelihood:**	-2901.4
No. Observations:	849	**AIC:**	5837.
Df Residuals:	832	**BIC:**	5918.
Df Model:	16		

	coef	**std err**	**t**	**P> |t|**	**[0.025**	**0.975]**
const	23.4439	11.235	2.087	0.037	1.392	45.496
PTSL5	0.1329	0.342	0.388	0.698	-0.539	0.805
PTSL3	0.0641	0.117	0.547	0.585	-0.166	0.294
MINL5	0.2082	0.116	1.788	0.074	-0.020	0.437
FGML5	0.9391	0.651	1.442	0.150	-0.339	2.217
FGAL5	-0.5390	0.497	-1.084	0.279	-1.515	0.437
FG_PCTL5	-17.3894	17.022	-1.022	0.307	-50.801	16.022
FG3ML5	-1.9574	1.117	-1.752	0.080	-4.150	0.236
FG3AL5	0.6860	0.531	1.291	0.197	-0.357	1.729
FG3_PCTL5	-2.0303	4.100	-0.495	0.621	-10.077	6.016
FTML5	0.2122	0.753	0.282	0.778	-1.266	1.690
FTAL5	-0.0565	0.635	-0.089	0.929	-1.303	1.190
FT_PCTL5	-6.5451	4.751	-1.378	0.169	-15.871	2.780
STLL5	-0.6383	0.479	-1.332	0.183	-1.579	0.302
PLUS_MINUSL5	0.0509	0.051	0.995	0.320	-0.050	0.151
TOT_PTSL5	0.0621	0.056	1.110	0.267	-0.048	0.172
DAYSOFF	-0.1186	0.186	-0.638	0.523	-0.483	0.246
YEAR	-0.0644	0.149	-0.432	0.666	-0.357	0.228

Omnibus:	21.030	**Durbin-Watson:**	2.051
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	25.739
Skew:	0.294	**Prob(JB):**	2.58e-06
Kurtosis:	3.617	**Cond. No.**	2.18e+16

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The smallest eigenvalue is 2.55e-26. This might indicate that there are
strong multicollinearity problems or that the design matrix is singular.

Our r^2 value is very low which means that the variables being used for the model are having a hard time accurately predicting our outcome which is Points scored. Now we will apply the model to the test set to evaluate the accuracy of the model. Applying the model to the test set gives the following values for RMSE (root mean square error), r^2, and MAPE(mean absolute percentage error):

Model 1 - Lebron	Test - 2018 season
RMSE	7.753315379958686
r^2	-0.021090252249140296
MAPE	25.842116632470884

We notice that r^2 is negative which means the model is performing worse than using the avg points as a predictor. However we are using data from previous season to predict the following season. In practice we know that the year prior may have little to no impact on how the player will score in the season. To have a more realistic test and training split we add half of the games from the 2018 season to the training set and use a smaller test set which contains that later half of games from the 2018 season. Since the model will always be run using all the previous game information you will never run the model as previous. With the new training and test split we will rerun the model and examine how effective it is at predicting points. Our second model gives:

Dep. Variable:	PTS	**R-squared:**	0.042
Model:	OLS	**Adj. R-squared:**	0.024
Method:	Least Squares	**F-statistic:**	2.372
Date:	Wed, 24 Jun 2020	**Prob (F-statistic):**	0.00182
Time:	22:07:25	**Log-Likelihood:**	-3001.1
No. Observations:	876	**AIC:**	6036.
Df Residuals:	859	**BIC:**	6117.
Df Model:	16		

	coef	**std err**	**t**	**P> |t|**	**[0.025**	**0.975]**
const	22.8005	10.967	2.079	0.038	1.275	44.326
PTSL5	0.0064	0.330	0.019	0.984	-0.642	0.655
PTSL3	0.1037	0.115	0.905	0.366	-0.121	0.328
MINL5	0.1489	0.111	1.341	0.180	-0.069	0.367
FGML5	0.8834	0.646	1.367	0.172	-0.385	2.152
FGAL5	-0.4422	0.490	-0.902	0.367	-1.404	0.520
FG_PCTL5	-13.8495	16.755	-0.827	0.409	-46.735	19.036
FG3ML5	-1.9930	1.097	-1.817	0.070	-4.145	0.159
FG3AL5	0.8405	0.523	1.608	0.108	-0.185	1.866
FG3_PCTL5	-1.5997	4.031	-0.397	0.692	-9.512	6.313
FTML5	0.2327	0.746	0.312	0.755	-1.232	1.698
FTAL5	-0.0590	0.635	-0.093	0.926	-1.305	1.187
FT_PCTL5	-7.0632	4.758	-1.485	0.138	-16.401	2.275
STLL5	-0.6173	0.477	-1.294	0.196	-1.554	0.319
PLUS_MINUSL5	0.0502	0.050	1.014	0.311	-0.047	0.147
TOT_PTSL5	0.0807	0.053	1.516	0.130	-0.024	0.185
DAYSOFF	-0.0824	0.186	-0.442	0.658	-0.448	0.283
YEAR	-0.1427	0.147	-0.973	0.331	-0.430	0.145

Omnibus:	26.116	**Durbin-Watson:**	2.042
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	32.648
Skew:	0.331	**Prob(JB):**	8.14e-08
Kurtosis:	3.675	**Cond. No.**	6.08e+16

Warnings:

[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

[2] The smallest eigenvalue is 3.39e-27. This might indicate that there are
strong multicollinearity problems or that the design matrix is singular.

Our r^2 value is very similar to the first model. Below we will apply the model to the smaller test set to evaluate performance.

Model 2 - Lebron	Test - Half 2018 season
RMSE	5.962926546520658
r^2	0.015025279347450016
MAPE	19.398868355111855

We notice our MAPE improved and went from about 26 percent to 19 percent. Also our r^2 is no longer negative but it is still very low. Since we are only testing the models performance on Lebron it is better we examine another player to see how effect the model really is. Below we will run the same thing as the second model on Giannis' games dataset but use half of the 2019 season as test and include all seasons from 2013 to half of the 2019 season.

Dep. Variable:	PTS	**R-squared:**	0.461
Model:	OLS	**Adj. R-squared:**	0.442
Method:	Least Squares	**F-statistic:**	24.46
Date:	Wed, 24 Jun 2020	**Prob (F-statistic):**	1.38e-51
Time:	22:14:37	**Log-Likelihood:**	-1596.4
No. Observations:	474	**AIC:**	3227.
Df Residuals:	457	**BIC:**	3297.
Df Model:	16		

	coef	**std err**	**t**	**P> |t|**	**[0.025**	**0.975]**
const	5.3140	8.507	0.625	0.533	-11.404	22.032
PTSL5	-0.5187	0.469	-1.107	0.269	-1.440	0.402
PTSL3	-0.1119	0.147	-0.759	0.449	-0.402	0.178
MINL5	0.0122	0.141	0.086	0.932	-0.266	0.290
FGML5	0.4552	0.905	0.503	0.615	-1.323	2.233
FGAL5	0.6012	0.565	1.063	0.288	-0.510	1.712
FG_PCTL5	-2.3889	13.554	-0.176	0.860	-29.025	24.248
FG3ML5	-2.3471	2.115	-1.110	0.268	-6.503	1.809
FG3AL5	0.1062	0.776	0.137	0.891	-1.418	1.631
FG3_PCTL5	4.7041	5.061	0.930	0.353	-5.241	14.649
FTML5	0.9181	0.868	1.058	0.291	-0.787	2.623
FTAL5	0.0022	0.597	0.004	0.997	-1.171	1.175
FT_PCTL5	3.0808	4.464	0.690	0.490	-5.691	11.852
STLL5	0.1746	0.731	0.239	0.811	-1.262	1.612
PLUS_MINUSL5	0.0077	0.072	0.107	0.915	-0.134	0.149
TOT_PTSL5	-0.0441	0.065	-0.676	0.500	-0.172	0.084
DAYSOFF	-0.2625	0.273	-0.961	0.337	-0.799	0.274
YEAR	3.9341	0.604	6.515	0.000	2.747	5.121

Omnibus:	2.488	**Durbin-Watson:**	1.976
Prob(Omnibus):	0.288	**Jarque-Bera (JB):**	2.267
Skew:	0.133	**Prob(JB):**	0.322
Kurtosis:	3.209	**Cond. No.**	7.38e+16

Warnings:

[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

[2] The smallest eigenvalue is 1.18e-27. This might indicate that there are
strong multicollinearity problems or that the design matrix is singular.

r^2 for the model run on Giannis' data is much higher than the r^2 we observed from the models built on Lebron's data. However to examine the performance of the model we will apply it to the testing set.

Model 3 - Giannis	Test - Half 2019 season
RMSE	5.962926546520658
r^2	-0.04454537567408745
MAPE	23.254694325171055

From above we notice that r^2 is again negative and the MAPE is also 23%. From this testing we can conclude that a linear regression model may not be the best method or technique to predict a players points given historical boxscore data. In the next section we will examine other Machine Learning techiques which may result in a better proforming prediction model.