**Finance Project Report**

**NBA Champions Prediction**

Javier Cao Cao - 173034450,

Tiangfang Geng- 215825640,

Abdur Rafeh Shah - 215835790

Lazaridis School of Business and Economics: Wilfrid Laurier University

MGTA-632

Dr. Salar Ghamat

July 27th, 2022

**Table of Contents**

[**Overview**](#_hj5otfly17y9) **3**

[**Objectives**](#_nob8nhzfx4tq) **3**

[**Data**](#_vjovtxus8n4d) **3**

[**Data Collection**](#_6b1wjmmhv7rg) **3**

[**Data pre-processing**](#_cxtu51fb11r) **4**

[**Data Exploration**](#_6428txcc30sx) **5**

[**Predictive Analysis**](#_qfb8qxkwyqxn) **5**

[**Linear Regression**](#_57zmeywsc3dj) **5**

[**Random Forest**](#_viz7sr57xodz) **6**

[**Evaluation of results**](#_5j904dy7gqwg) **7**

[**Risks**](#_sj3ze3gawuop) **7**

[**Conclusion and Recommendation**](#_5962yxozwe2) **8**

[**Appendix**](#_eutopmj95o8o) **8**

# Overview and Business Problem

Basketball is one of the biggest global sports with approximately 2.2 million fans worldwide. From a business point of view, the global sport is also very profitable as The National Basketball Association (NBA) had a revenue of approximately 8.8 billion US dollars per year for the 2018/2019 season. While the Covid-19 pandemic did affect the sport’s annual revenue and viewership in a negative manner for the year 2020, the number of tickets sold, and viewership improved with projected revenue expected to surpass 10 billion USD in the next few years.

Sports betting is one of the most famous and profitable businesses that uses prediction to predict possible results and place bets on an outcome. Betting is ubiquitous as it is found in most major sports around the world where gamblers place bets on teams with higher chances of winning to make money. It is of no surprise that basketball being one of the biggest and most profitable global sports serves as a betting haven for gamblers. Many major betting companies and houses, such as Proline, Bet365, and Bet99, have become popular as they provide opportunities for new and seasoned gamblers to place their bets on their favorite teams. From a betting company’s point of view, providing a betting platform is very profitable as they utilize various strategies and predictions such as odds ratio to predict the outcome of matches. Like casinos, betting houses employ various prediction models and strategies that use data to predict teams that have the highest likelihood of winning. The companies then use this information for prediction outcomes. Since betting companies technically always have the upper hand, the likelihood of them losing money from bets is relatively slim.

At RJC Analytical Solutions, our job as consultants is to utilize past NBA season data to predict the team that has the highest likelihood of winning the playoff championships. A typical 8-month season starts in September and ends in April after which qualifying teams compete in a 2-month playoffs season to win the final championship. Theoretically, teams that perform well during the 8 months season are also expected to perform well during the playoffs. Our prediction model aims to utilize seasonal data to accurately predict the team that has the highest likelihood of winning the final championship. We plan to provide this information to major betting companies to help them become profitable by implementing the appropriate betting strategy. From a betting company’s point of view, it is crucial to accurately predict the likelihood of the outcome being profitable. At RJC Analytical Solutions, our team strongly believes that any betting house that utilizes our prediction model will gain the upper hand in terms of having the technical advantage, accurate outcome prediction, and increasing their likelihood of being profitable.

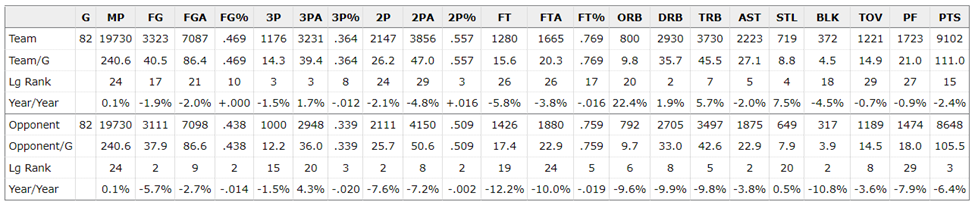
**Objectives**

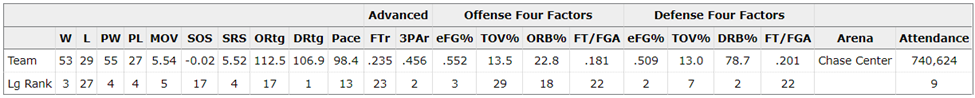
Our team’s main objective is to provide betting houses with analytical solutions that can substantially increase their likelihood of being profitable by accurately predicting the NBA champions for a specific season. To do this, our team will be using linear regression and random forest prediction models to accurately predict the outcome. At RJC Analytical Solutions, our goal is to become a trusted solutions and consulting firm and become a market leader in the betting industry.

# Data

## **Data Collection**

Basketball Reference is a basketball-related website that collects and provides every season, team, and player statistics in the history of the NBA. The website is the firm’s primary source of data and only regular season data will be used to predict the champions. A sample of the data is provided below, showing statistics for a single team during a specific season.





The data is divided into three types of statistics: "Team", "Opponent", and "Miscellaneous". As their name suggests, team and opponent stats represent each team and their average opponent per game numbers, respectively. Miscellaneous is a diverse and advanced type of statistics, some of which involve calculations with game numbers. The definition of each variable can be found in the data dictionary in appendix 1.The information from the dataset can be helpful in developing a profile of a championship team with its more notable characteristics. Additionally, these variables can provide an overall picture of each team's strengths and weaknesses which will be critical when calculating their likelihood of becoming a champion.

The dependent variable is the number of post-season (or playoff) wins. To become a champion, a team must reach 16 wins, i.e., win four 7-game series. Hence, the reasoning behind the analysis is that the team with the highest predicted number of wins will have the highest chance of winning the championship.

## **Data pre-processing**

The first step was to consider data after the 2002-2003 season. Starting that season, every playoff matchup, from the first round to the finals, was a best-of-seven series or the first to win four games. Including previous seasons would have skewed the results as the number of games required for winning the title has changed over time. Additionally, only teams who qualified for each year's playoff were included in the dataset (i.e. 16 teams, 8 for each conference).

One of the main problems when tracking statistics in sports is the constant change of game numbers due to improvements in coaching, playstyle, strategies, etc. The pace is defined as the total number of possessions per game and is an excellent metric to determine how the game is played during an era. **Appendix 2** shows an increase in pace from the 2003 to 2022 season, which led to higher points per game during the same stretch since more possessions lead to more opportunities to score. The following appendix (3) shows the number of 3-point attempts in the same time span. There is a noticeable increase in the number of 3-point attempts from the 2013 season, a year in which Stephen Curry, the greatest shooter of all time, broke for the first time the 3-point record. Since that year, the league noticed the value of the 3-pointer and started to apply it in their strategies. Those examples showed the constant change of the game, even for a small period of time. Therefore, to rectify this issue, ranks were used instead of game numbers. Finally, highly correlated variables were removed as they do not provide valuable insights, do not improve the model, and will help reduce the complexity of the model.

## **Data Exploration**

Before performing the predictive analysis, a brief exploratory analysis was performed to acquire insights into which features or patterns previous championship teams have in a visual manner. The ranks of some statistics (e.g., points per game) of every championship team since 2003 were plotted to better understand what leads a team to win the title.

Regarding offense, two statistics were plotted: offensive rating and effective field goal percentage (EFGP). The offensive rating measures offensive performance and refers to the number of points scored per 100 possessions. EFGP is a measurement of shooting effectiveness and is a better indicator than field goal percentage as it adds more value to the three-point. In appendix 4, it seems that most championship teams had a top 10 offensive rating in the league (red line) and a top 5 EFGP. This may suggest that teams that take advantage of every procession and high shooting efficiency have higher chances of winning more playoff matches.

On the defensive end, defensive rating and opponent EFGP were plotted. Similar to the previous two variables, defensive rating refers to the number of points a team allows, while opponent EFGP shows how well the opposing team shoots against your team. In appendix 5, both variables seem correlated, and championship teams have to be part of their top 10. Those statistics show the importance of being disciplined on defense and back the phrase "Defense wins championships".

In addition, some miscellaneous stats were plotted. Appendix 6 depicts the number of regular season wins a champion had. Only in two out of the 19 years, the champion positioned themselves outside of the top 5. Simple system rating (SRS) is "a [team evaluation rating](https://www.nbastuffer.com/team-evaluation-metrics) that takes into account average point differential and [strength of schedule](https://www.nbastuffer.com/analytics101/strength-of-schedule-sos/)" (O'Neill, 2021). The rating is fairly accurate when evaluating a team's performance; on the graph, it seems that champions fell in the top 5. Finally, the last chart (Appendix 7) illustrates the margin of victory, and the teams, on average, were ranked third among all teams in the league. This might indicate that winning games are insufficient; instead, it is essential to consider how a team wins (e.g., a score blowout) as it shows their dominance and superiority against their opponents.

# Predictive Analysis

## **Linear Regression**

This analysis used all team’s regular season data to predict the number of wins during playoffs. The regular season data from 2003 to 2021 was used as a training set to build the models, and each team's performance for this year (2021-2022) was used as a testing/predicting dataset.

The feature selection method was applied for the linear regression predicting models to find the best combination of predictor variables while avoiding potential overfitting problems. And the repeated k-fold Cross Validation method was used to help select the best model with the lowest error.

The first model included all the independent variables. The cross-validation result shows that predictor variables used in the model accounted for approximately 38.73% of the variance in ‘Playoff Wins’ (*RMSE* = 3.80, *MAE* = 3.03), *F*(48, 255) = 6.27, *p* < .001. The model is significant in predicting the dependent variable. Five predictors have significantly predicted Number of playoffs wins: SOS (*B* = 0.12, *p* < .01) and TOV\_perc (*B* = 0.33, *p* < .01) are showing a positive relationship with dependent variables; and TOV (*B* = -0.23, *p* < .05), PTS (*B* = -0.20, *p* < .05), and Number of Win (*B* = -0.51, *p* < .001) shows a negative relationship with Playoff Wins. The Top three Wins team forecast results for this model are: Phoenix Suns (11.29), Memphis Grizzlies (10.58), and Golden State Warriors (8.23). (Appendix 8-9)

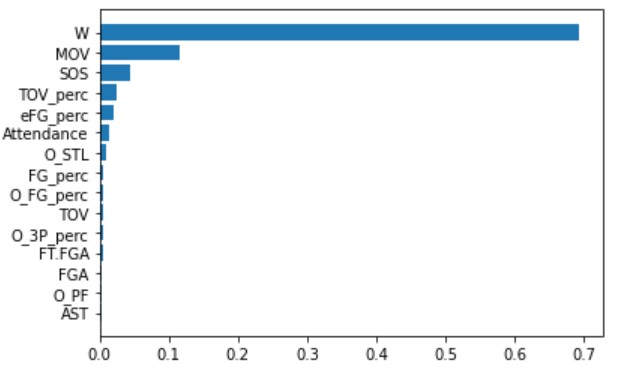
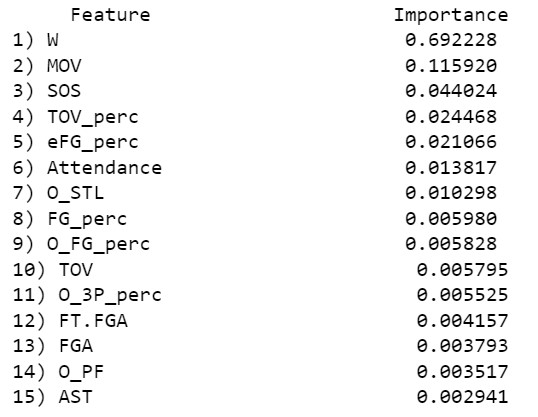
The second model applied forward feature selection and cross-validation methods. The model result shows that the predictor variables selected by forward feature selection accounted for approximately 48.89% of the variance in ‘Playoff Wins’, *F*(5, 298) = 57.26, *p* < .001. This model has a higher R^2 with a lower error level (*RMSE* = 3.41, *MAE* = 2.74), which means this is a less overfitting model. Totally five variables being selected and four have significantly predicted Number of ‘playoff wins’: SOS (*B* = 0.11, *p* < .001) and TOV\_perc (*B* = 0.09, *p* < .001) are showing a positive relationship with dependent variables; The O\_BLK (*B* = -0.05, *p* < .05), and Number of Win (*B* = -0.59, *p* < .001) shows a negative relationship with ‘Playoff Wins’. The Top three Wins team forecast results for this model are: Miami Heat (11.29), Golden State Warriors (10.29), and Phoenix Suns (10.17). (Appendix 10)

The third model applied backward feature selection and cross-validation methods (backward and stepwise return the same result). The model result shows that the predictor variables selected by forward feature selection accounted for approximately 49.62% of the variance in ‘Playoff Wins’, *F*(9,294) = 33.73, *p* <.001. Compared to the forward selection model, the backward/stepwise model didn’t improve much in terms of R^2 and error level (*RMSE* = 3.40, *MAE* = 2.71). Totally nine variables being selected and five have significantly predicted Number of ‘playoff wins’: SOS (*B* = 0.10, *p* < .001) and TOV\_perc (*B* = 0.31, *p* < .001) are showing a positive relationship with dependent variables; and TOV (*B* = -0.21, *p* < .01), PTS (*B* = -0.15, *p* < .01), and Number of Win (*B* = -0.58, *p* < .001) shows a negative relationship with ‘Playoff Wins’. The Top three Wins team forecast results for this model are: Phoenix Suns (10.39), Miami Heat (9.63), and Golden State Warriors (9.51). (Appendix 11)

By comparing the results from three models, both feature selection models are improved on the overfitting problem. However, they seem to return similar forecast results (same top three teams with different orders). Also, in the third model, both TOV and TOV\_perc are significant but have an opposite relationship with the target variable. For more accurate forecasting, a random forest model was used to explore the importance of the predictor variables and further validate the results.

## **Random Forest**

A Basic Random Forest model was applied to build a predictive model based on a training dataset (historical data). After turning the model with cross-validation and bootstrap methods, the random forest model contains the lowest error compared to other models (*RMSE* = 3.09, *MAE* = 1.55). The Top three Wins team forecast results for the random forest model are: Golden State Warriors (10.75), Miami Heat (9.85), and Phoenix Suns (9.60). (Appendix 12-13) And the importance of variables that selective by random forest model are shown below:



# From the figure above, the importance of the variables' results is linked back to the finding from the data exploration. The top three important variables are the Number of Wins, Margin of Victory, and Strength of the Schedule. Also, the TOV\_perc is more important than TOV.

# Risks

Although the results were satisfactory, the model has flaws and does not consider multiple common aspects in sports. It should be noted that we are only using statistics from the regular season, so any event that occurs during the playoffs will not be considered in the model. Injuries are the worst fear of every team, especially if it happens to the team's star players. They are unpredictable and can occur during the playoffs and completely change the course of a team's playoff run. Players' physical status is not the only scenario to care about; psychological status also matters. Being able to deal with the pressure of the game, which includes in-game distractions (e.g., fans), media, and the team's history on your back, and not suffer a mental breakdown comes with more exposure to such an event, which is also known as experience. There are more intangible factors that are impossible to collect, such as coaching, strategy, etc., which might considerably impact how the team will perform in the playoffs. Finally, all those variables plus the "unpredictability" nature of sports make the model not perfect but sufficient to fulfill the project's objective.

Another issue is the output of the models. When predicting the number of playoff wins, both regression and random forest championship team results had an expected number of playoff wins lower than 16. Also, regression obtained "negative wins" as output for teams. The size of the sample data does not allow a very accurate and robust model. However, the results can be interpreted as the team with the highest predicted wins will have the most chances of winning the title.

# Conclusion and Recommendation

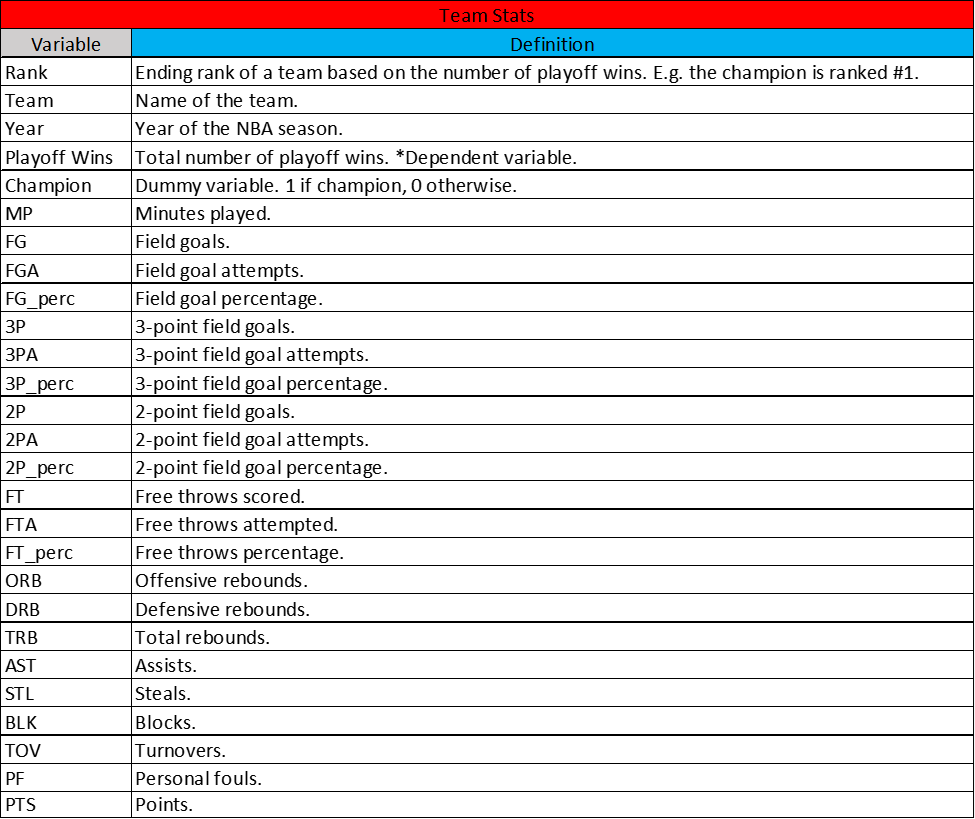
Basketball being one of the biggest global sports in the world is one of the most popular choices for betting in the betting industry. Every year, the NBA attracts thousands of bettors who place bets on teams that have the highest likelihood of winning the playoffs and becoming champions. From a betting house point of view, accurately predicting the outcome is crucial as it leads to higher profit. Failure to accurately predict the outcome can prove to be a huge financial burden as betting companies will have to pay out large sums of money if they lose. At RJC Analytical Solutions, we strive to become a market leader in the sports betting industry by providing accurate and reliable predictions that can be used by major betting houses to make a profit. The firm utilized regular 8-month season teams and players data from the years 2003 to 2020 from “Basketball Reference '' to develop various prediction models to determine the playoffs champions. The first 3 prediction models were regression models (base, forward-selection, and backward-selection) that yielded fairly similar results as the same teams were predicted as the top 3 teams (Phoenix Suns, Miami Heat, and Golden State Warriors) that had the highest likelihood of winning the championship. The fourth model was the basic Random Forest model which also yielded the same teams as the top 3 teams with the highest likelihood of winning the championship. Out of the 4 prediction models, the Random Forrest model was chosen as the model of choice as it had the lowest error compared to previous 3 models (*RMSE* = 3.09, *MAE* = 1.55). Based on the significant features and analysis, it was determined that teams favourable towards certain attributes, performance indicators and statistics have the highest likelihood towards winning the championship. The following significant attributes are as follows:

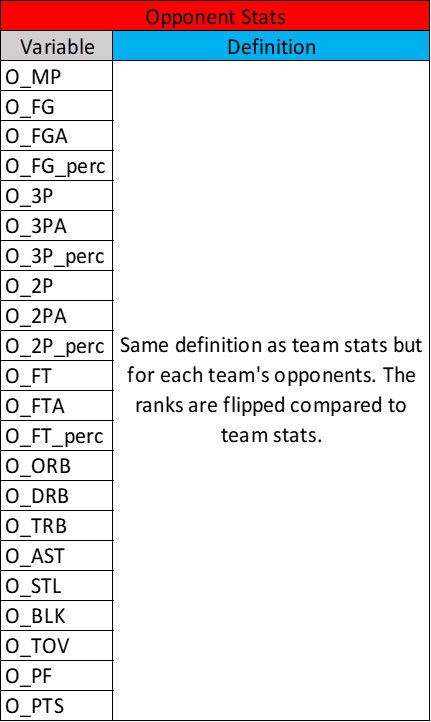
* Good overall performance in the regular season
* Having a high margin of victory (MOV) as it indicates team strength and dominance relative to other teams
* Be efficient when shooting
* Have high points per game
* Have low turnovers
* Have a relatively easier game schedule

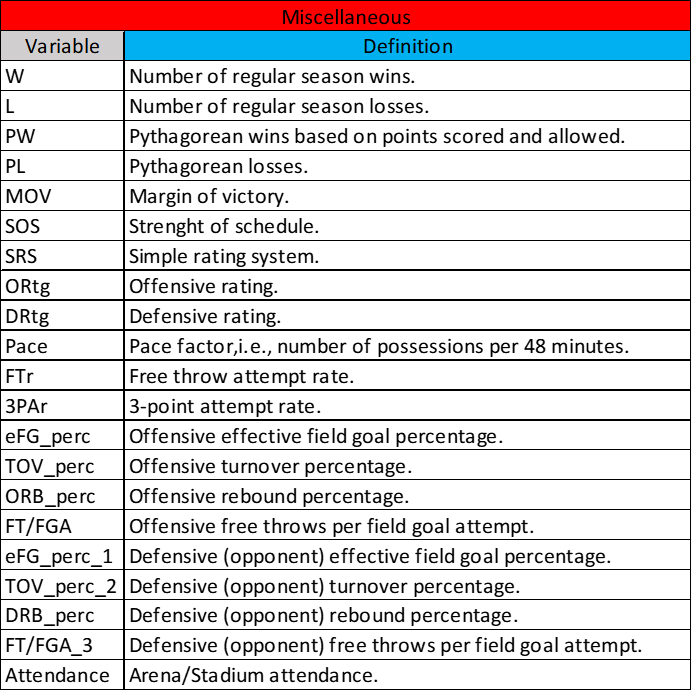
Based on our team’s analysis, it is recommended that the betting companies pay attention to the listed statistics and KPI’s to determine teams that have the highest likelihood of winning the championship.

# Appendices

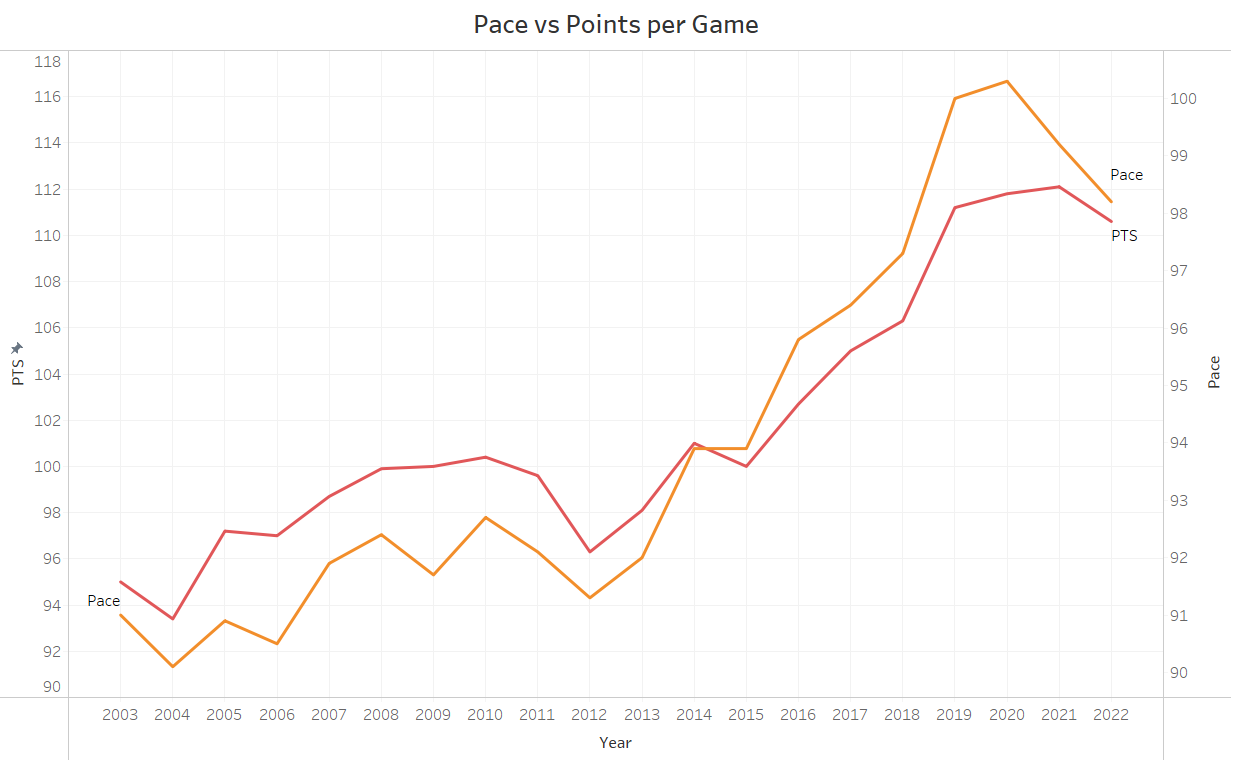
Appendix 1



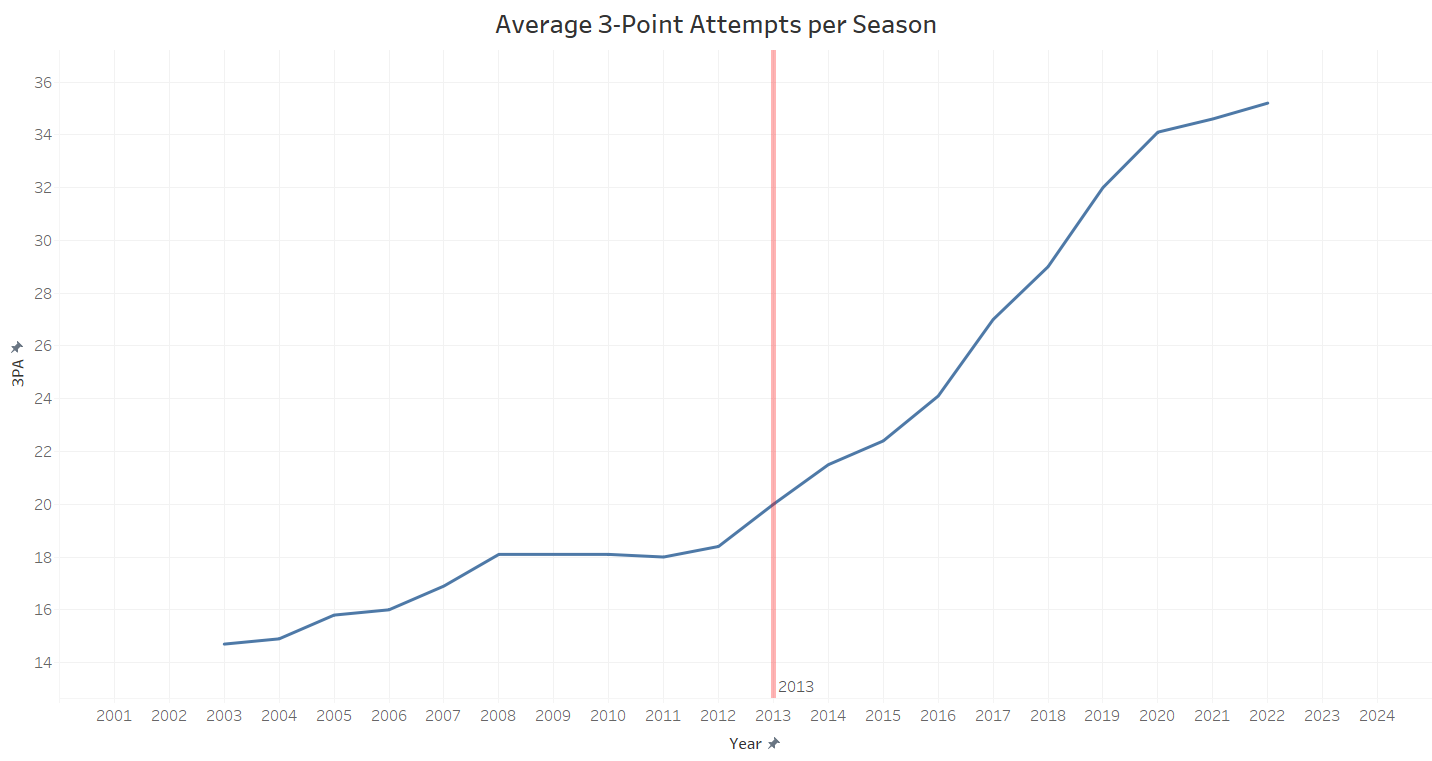




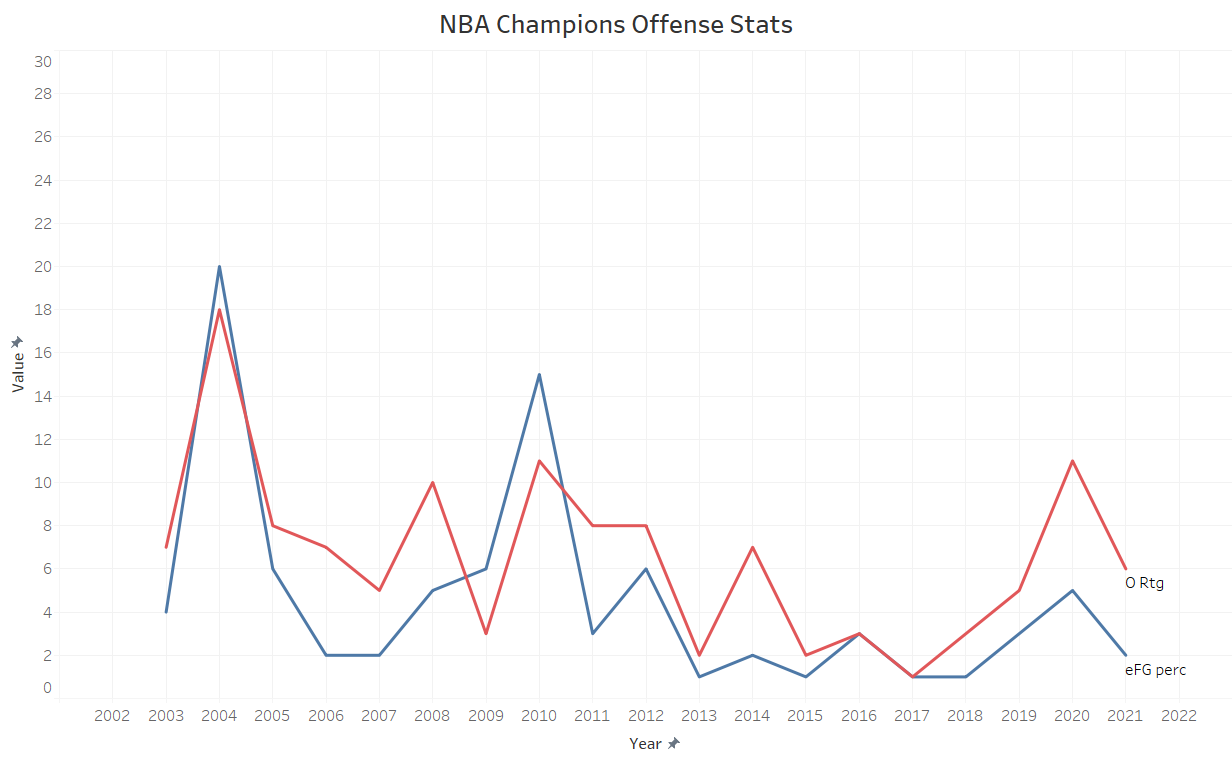
Appendix 2



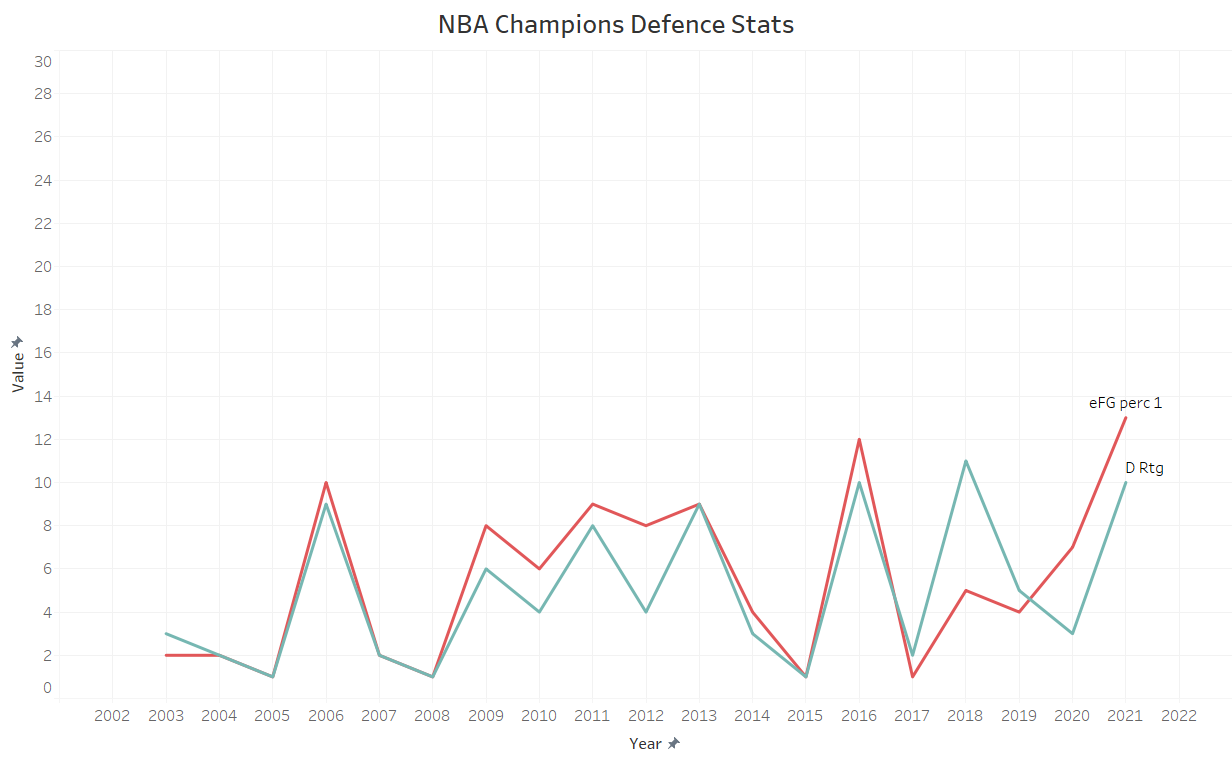
Appendix 3



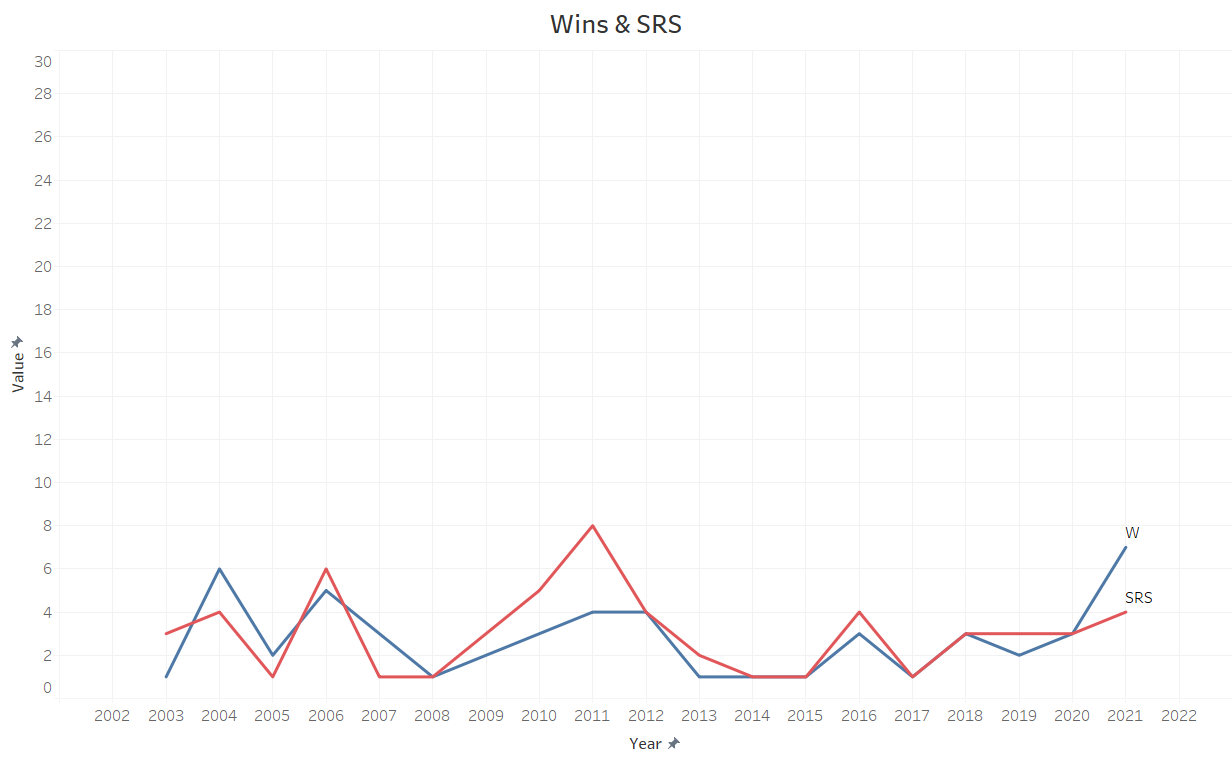
Appendix 4



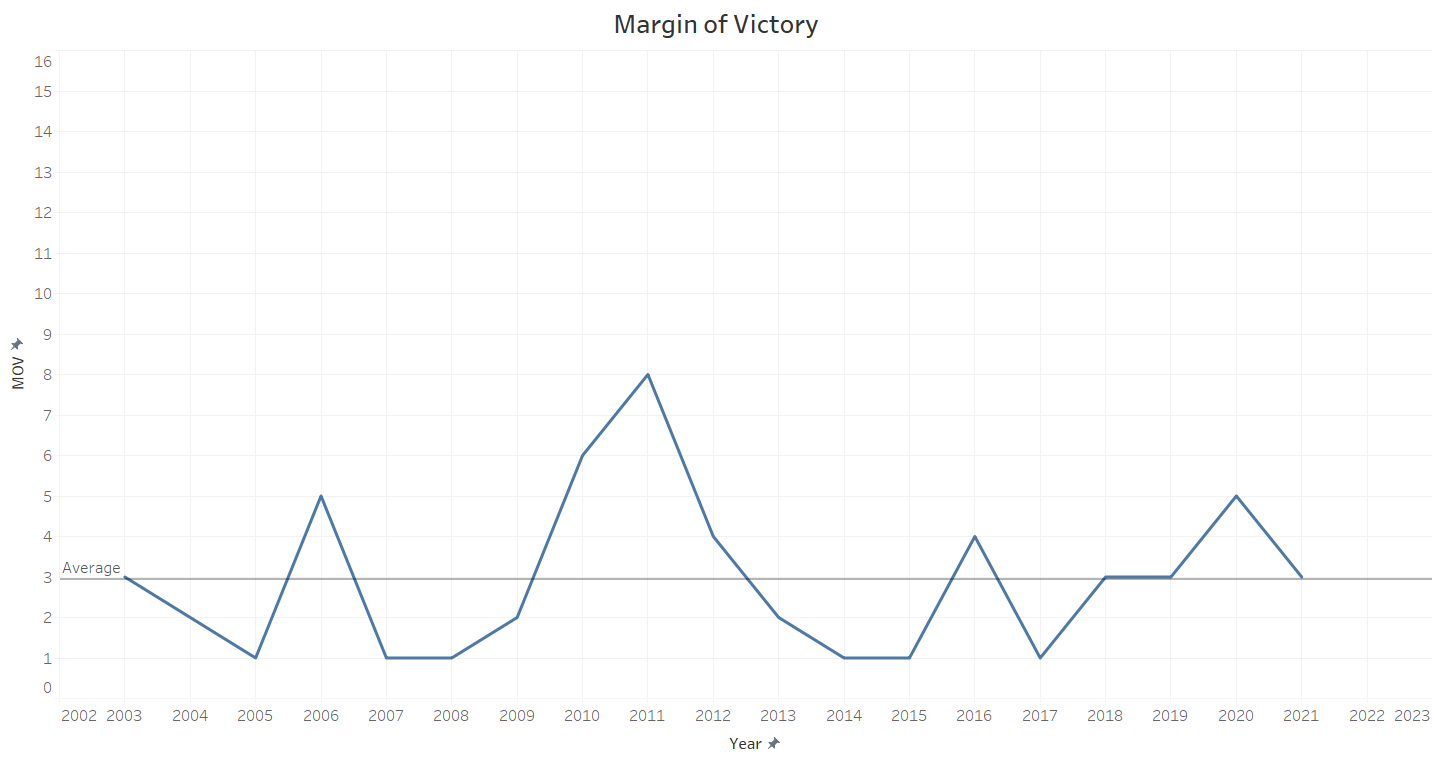
Appendix 5



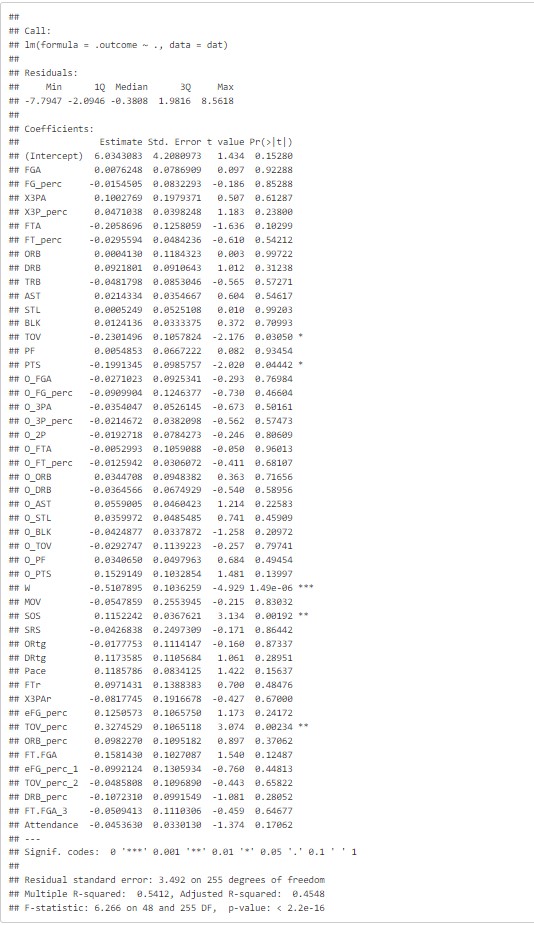
Appendix 6



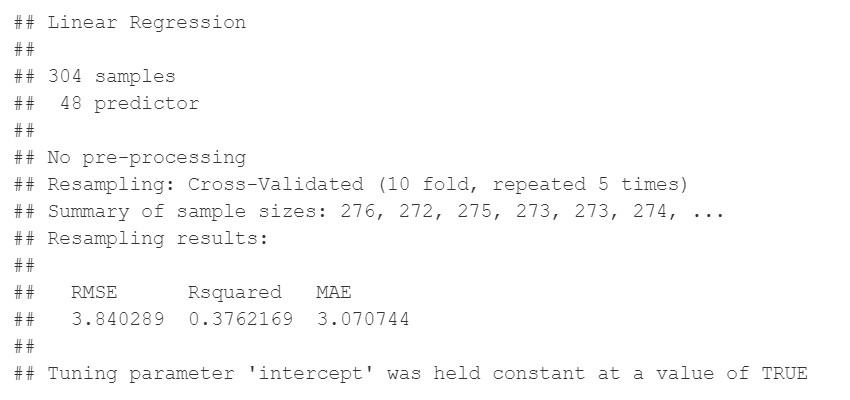
Appendix 7



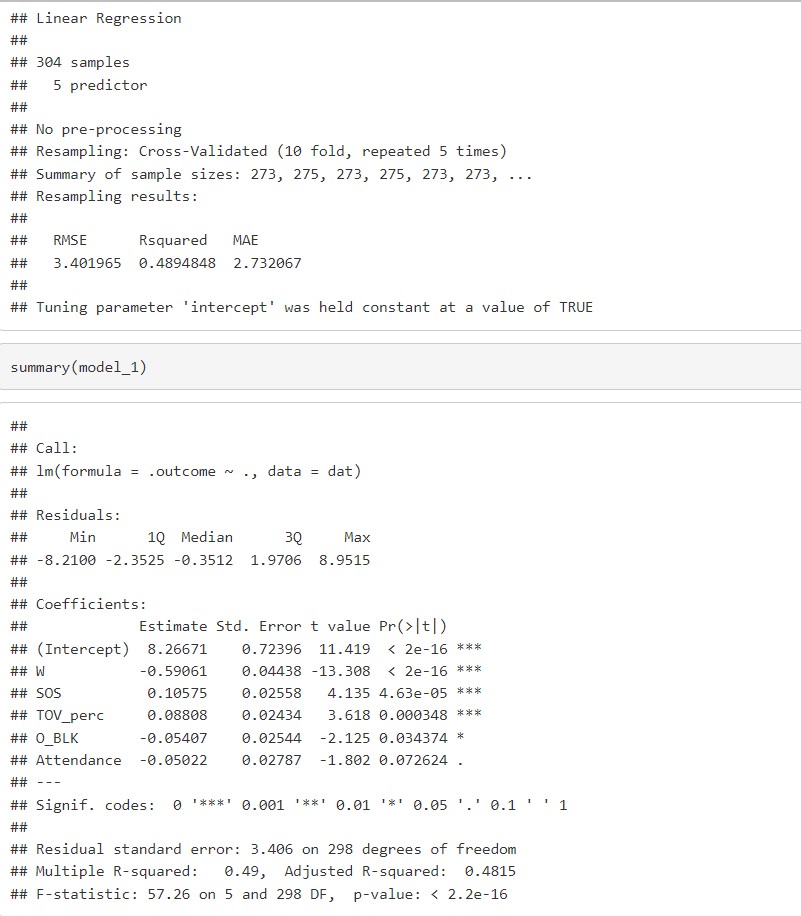
Appendix 8



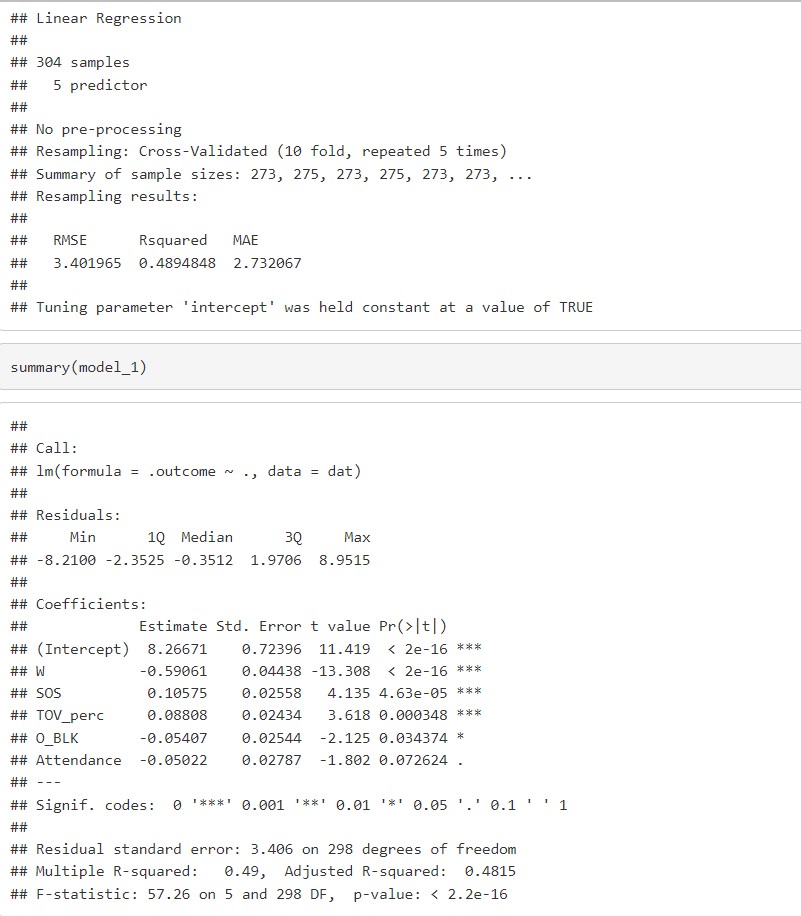
Appendix 9



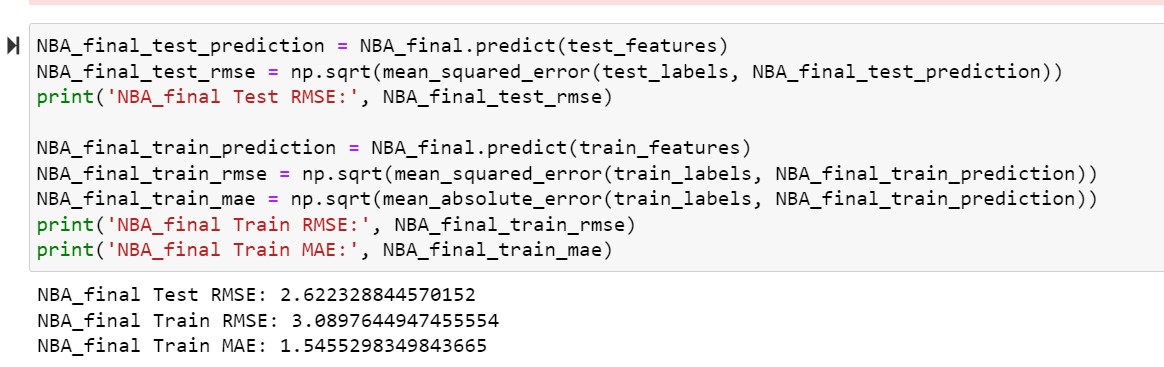
Appendix 10



Appendix 11



Appendix 12



Appendix 13

