Exploratory Data Analysis of the Toronto Raptors 2018-2019 Season

GDAA 1000 - Fundamentals of Geospatial Data Analytics

Aila Jalo

23/11/2021

Centre of Geographic Sciences

COGS | nscc

List of Figures

List of Tables

Introduction

The purpose of this assignment was to perform exploratory data analysis by looking at variance and co-variance between variables. The data had to come from an external data set that had a minimum of 100 observations across 5 variables where at least two of the variables were categorical.The data set used in this report is one of the team statistics for the Toronto Raptors (TR) of the National Basketball Association (NBA) from the 2018-2019 season.

One expected result would to see some correlation between certain stats and whether or not they contribute to a game being won. I expect to see that the TR won more games than they lost in most of the divisions. I also expect to see that the TR played better when they had the home court advantage.

Data Selection and Preparation

These data sets were chosen as I was watching a TR game and I realized that there would be both numerical and categorical variables in any sports data set as there are usually different divisions and subdivisions of teams, and that there are many different stats that are recorded for every game and player. Searching on Kaggle, I found a data set for the TR 2018-2019 season, which was when they won the NBA Championship. The data set contained four files, the player and team stats for the regular and playoff games. I chose to work with the team stats table only, as the player tables were proving difficult to work with due to the un-equal number of games played by each player due to injuries or being traded mid-season.

There were several things I did to tidy and re-arrange the data. I noticed that under the columns MATCHUP, the string format was either 'TOR vs MIA' or 'TOR @ MIA', where 'vs.' represented a game played at home, and '@' represented an away game. I then made a new column in team tables which would denote whether it was a home or away game base on whether the MATCHUP columns contained 'vs.' or '@'. Using the MATCHUP column again, I created a new column named 'teamAbbr' for the abbreviation of the opposing team name using the last three characters in the MATCHUP column. I added a new column named 'R_P' in the regular and playoff season tables where every row was either 'R' or 'P' to denote which part of the season the game was in. I then combined the two team tables together so that the playoff games would be in the same table as the regular season games.

The original data set did not include information about the opposing teams conference or division. I found another table on Kaggle that contained stats for every game from 2012 to 2018. I used Excel to delete all of the rows that were not in 2018, and then replaced a few abbreviations that only contained 2 characters with the proper 3 character abbreviations so that they would match the data sets for the TR data. In R, I then made a new data frame called 'nba_teams' which contained only the unique combinations of team abbreviations, conferences, and divisions from the Box Score table.

To create the final team tables, I used a left join to join the 'nba_teams' table to the team table using the 'teamAbbr' column, so that they would contain information about the conference and divisions of the opposing teams. To clean up the work space, I removed all of the tables except for the 'team' tables. I then converted the columns 'W.L' (win/loss), 'teamConf' (conference), 'teamAbbr' (team abbreviations) and 'teamDiv' (division) to factors.

Table 1: Structure of Team Table; 1136 obs. of 30 variables

Variable	Class	Observations
TEAM	character	TOR, TOR, TOR, TOR, TOR, TOR
DATE	character	01/13/2019, 11/29/2018, 02/13/2019, 10/30/2018, 03/18/2019, 11/10/2018
MATCHUP	character	TOR @ WAS, TOR vs. GSW, TOR vs. WAS, TOR vs. PHI, TOR vs. NYK, TOR vs. NYK
W.L	integer	W, W, W, W, W, W
MIN	integer	290, 265, 241, 240, 240, 239
PTS	integer	140, 131, 129, 129, 128, 128
FGM	integer	49, 47, 44, 47, 49, 42
FGA	integer	104, 90, 92, 92, 88, 79
FG.	double	47.1, 52.2, 47.8, 51.1, 55.7, 53.2
X3PM	integer	13, 15, 16, 11, 17, 14
X3PA	integer	36, 38, 38, 27, 40, 35
X3P.	double	36.1, 39.5, 42.1, 40.7, 42.5, 40
FTM	integer	29, 22, 25, 24, 13, 30
FTA	integer	36, 24, 30, 28, 22, 40
FT.	double	80.6, 91.7, 83.3, 85.7, 59.1, 75
OREB	integer	17, 7, 9, 9, 5, 12
DREB	integer	42, 30, 43, 34, 45, 34
REB	integer	59, 37, 52, 43, 50, 46
AST	integer	24, 25, 32, 29, 36, 24
STL	integer	16, 5, 8, 13, 5, 8
BLK	integer	8, 6, 8, 6, 6, 4
TOV	integer	21, 14, 14, 15, 14, 15
PF	integer	27, 23, 27, 26, 22, 20
X...	integer	2, 3, 9, 17, 36, 16
R_P	integer	R, R, R, R, R, R
H_A	integer	A, H, H, H, H, H
teamAbbr	integer	WAS, GSW, WAS, PHI, NYK, NYK
teamConf	integer	East, West, East, East, East, East
teamDiv	integer	Southeast, Pacific, Southeast, Atlantic, Atlantic, Atlantic

Data Summary

Data Structure

The structure of the team table can be viewed using the function str(); which prints the data as it is seen in the Environment tab in R. The output of the str(team) function can be seen in the appendix. It is also difficult to have this formatted into a nice table using kable.

One way to display the structure of the table using kable would be to first create a new data frame where the columns are the names of the variables, class, and observations from the player table. This way the kable styling and formatting can be applied to the data frame. One issue is that the number of observations and variables will not be printed, therefore they are added to the title of the table as seen in Table 1. Another issue is that the variables which are Factors shows the data types as being an Integer.

Data Summary

A summary of the data can be seen in Tale 2. This is a subset of the first 6 variables. A summary of the entire table can be seen in the appendix. A table with the names of the different stat codes can also be seen in the appendix.

Table 2: Summary of Team Table

TEAM	DATE	MATCHUP	W.L	MIN	PTS
Length:106	Length:106	Length:106	L:32	Min. :238.0	Min. : 86.0
Class :character	Class :character	Class :character	W:74	1st Qu.:239.0	1st Qu.:105.0
Mode :character	Mode :character	Mode :character	NA	Median :240.0	Median :114.0
NA	NA	NA	NA	Mean :242.3	Mean :112.7
NA	NA	NA	NA	3rd Qu.:241.0	3rd Qu.:121.8
NA	NA	NA	NA	Max. :290.0	Max. :140.0

Exploration of Variation

Variance is how near an observation is to the mean. A small variance means that the numbers in the data set do not vary much from the mean. A large variance means that the numbers fall far from the mean. An example of small variance would be Figure 9, as there is little variation in the length of the games. An example of large variance would be Figure 7, where the difference in points between the winning and losing team may fall further away from the mean.

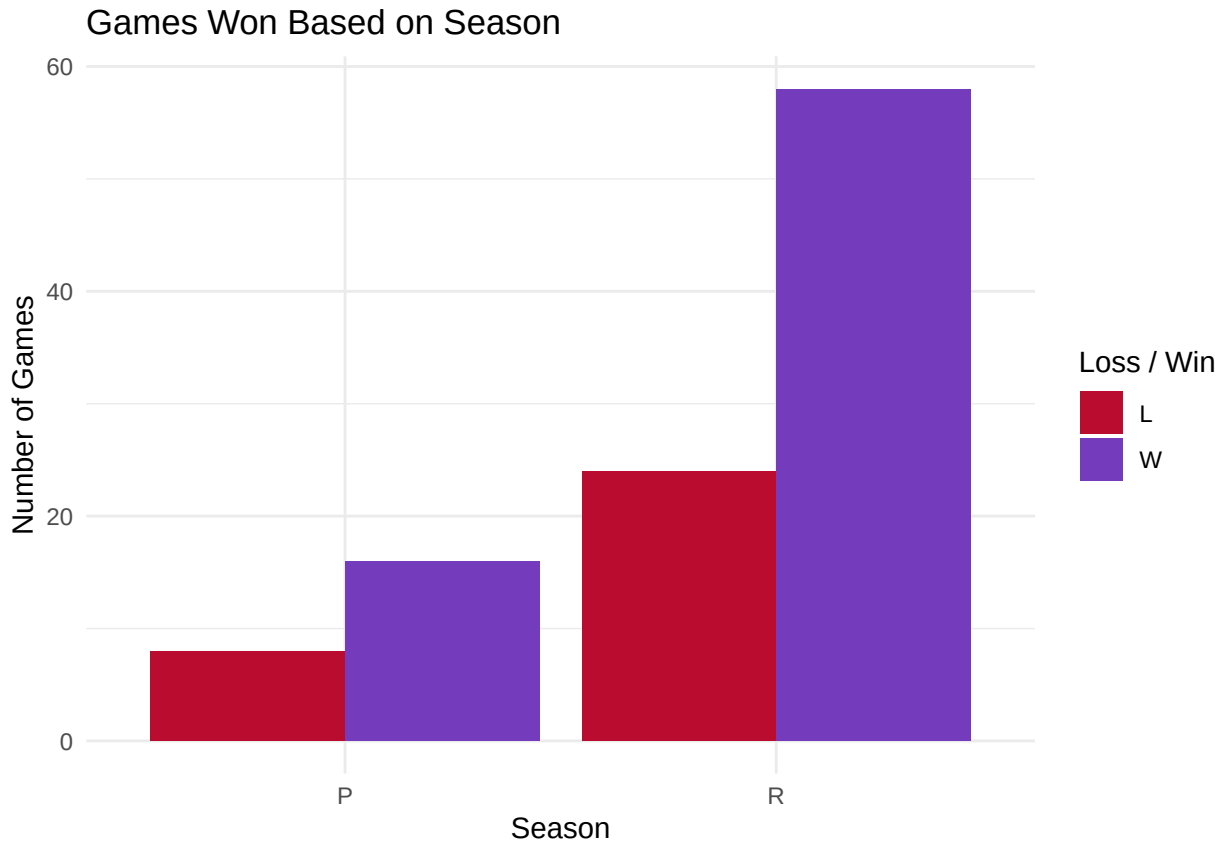

Figure 1: Games Won Based on Season

In Figure 1, it appears as though the TR won a larger percentage of their regular season games when compared to the playoff games. This could perhaps show that the TR were much more evenly matched by their opponents in the playoffs. This would make sense as all the teams making the playoffs are going to be the best in the league that year.

Figure 2 shows that the TR benefit greatly from playing in Toronto. The team still had a good chance at beating the opposing team if they were the away team.

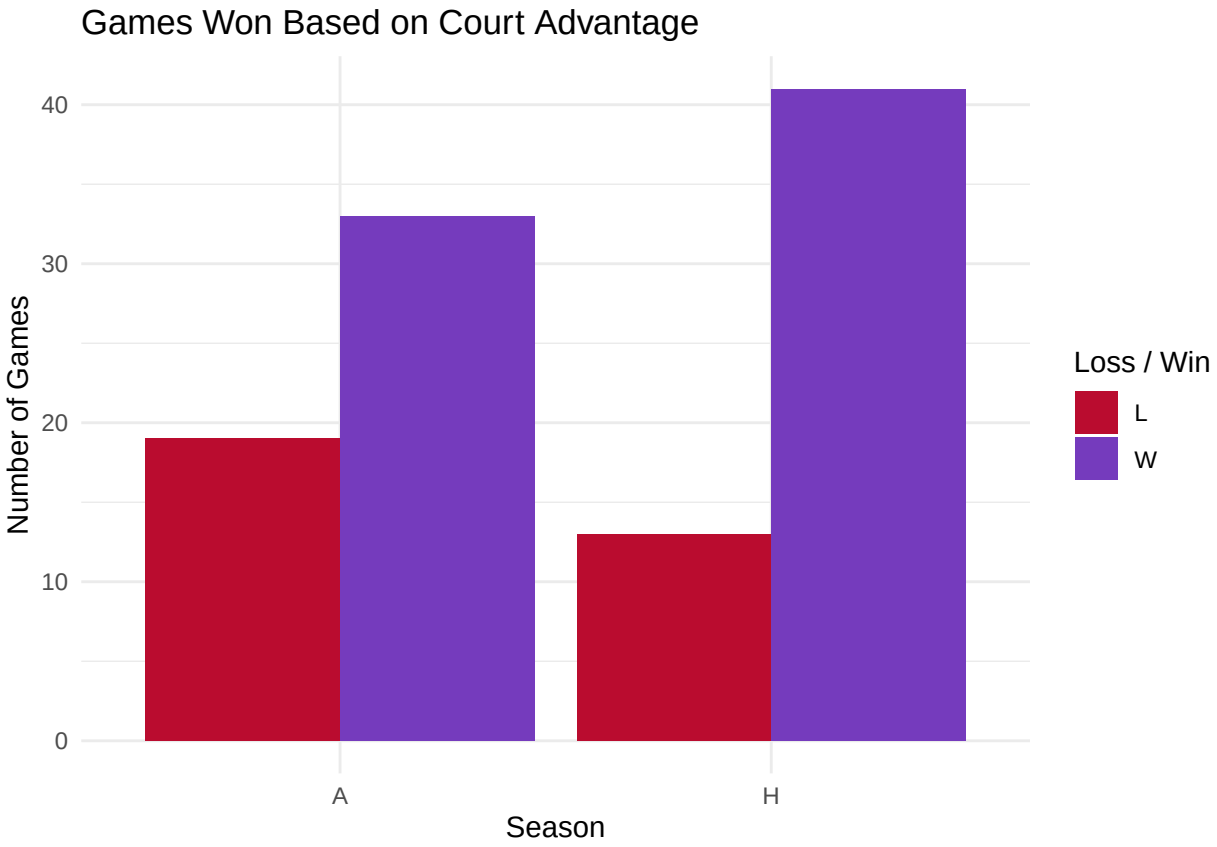

Figure 2: Games Won Based on Court Advantage

Figure 3 shows confirms the results from the first two figures. The TR were most likely going to defeat their opponents in both Home and Away games during the regular season, but they were much more evenly matched in the playoffs, and their odds of winning decreased slightly. Court advantage might not be as important of a factor in determining a win or loss during the playoff season.

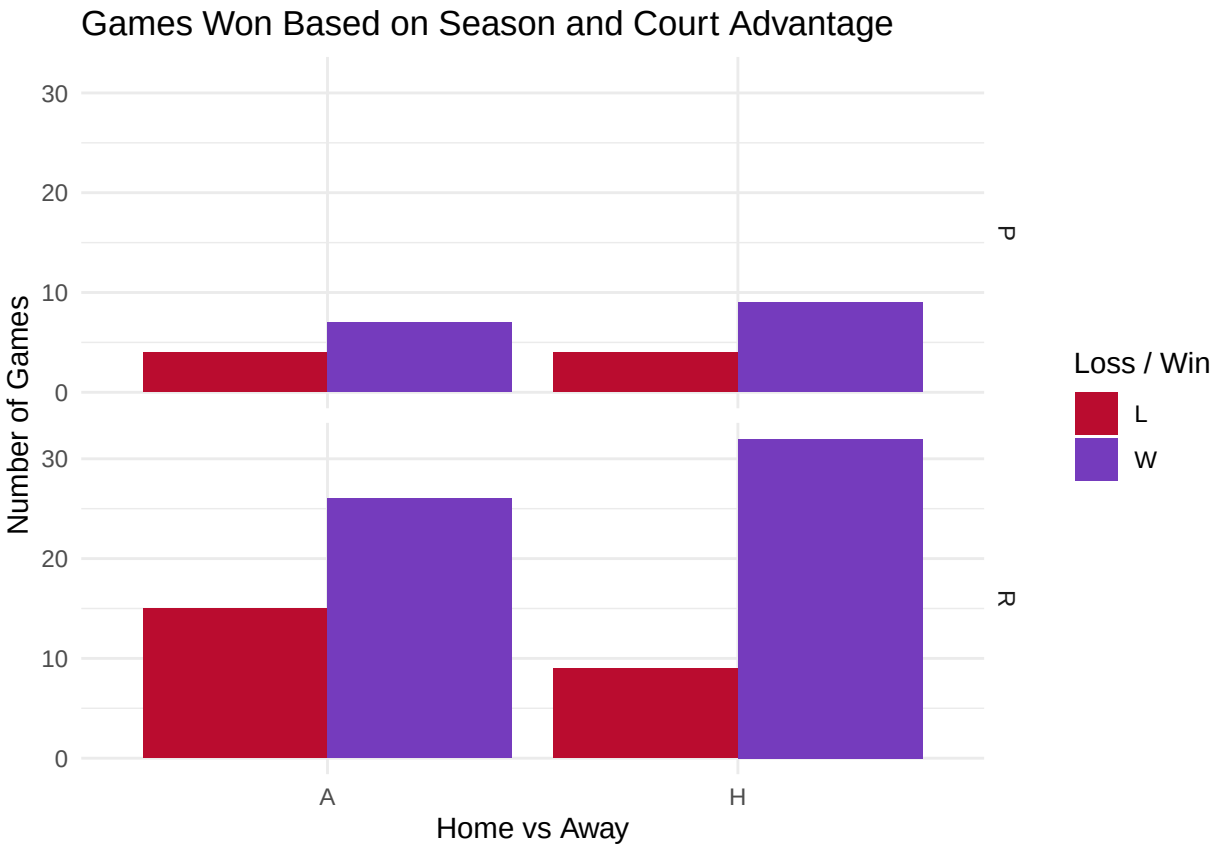

Figure 3: Games Won Based on Season and Court Advantage

Figure 4 shows the wins and losses based on the opposing teams conference and division. During the regular season will play, a team will play more games in their own conference than in the other. We can see that the TR played really well against teams in the Southeast and Atlantic divisions in the Eastern Conference. When the TR played a Western Conference team, they were relatively evenly matched in the Northwest and Southwest divisions, although played great against teams from the Pacific division.

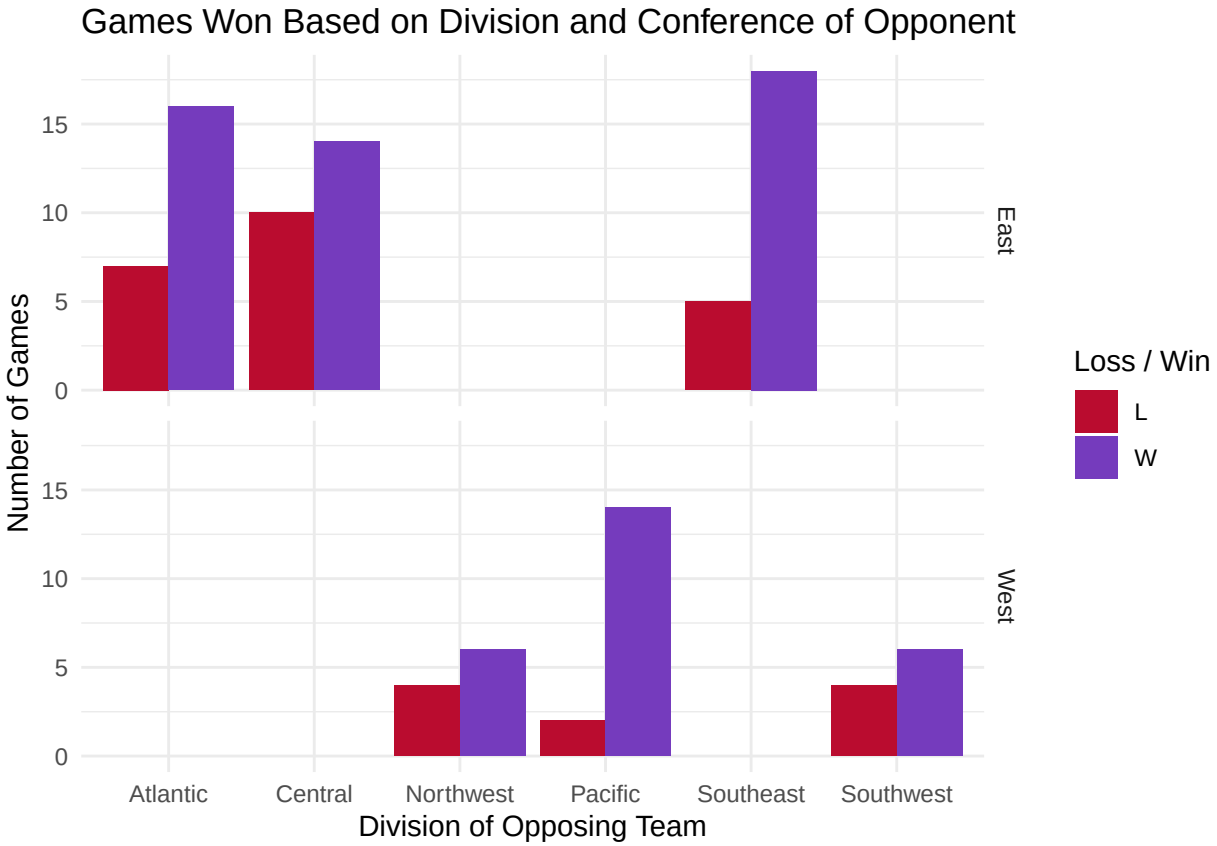

Figure 4: Games Won Based on Division and Conference of Opponent

In Figure 5, it is shown that the TR out-played every other team in the Eastern conference except for three teams. The TR swept 5 of the other teams in the Eastern conference. The TR were evenly matched against Boston and Milwaukee. The TR lost all of their games against Detroit, and lost more games to Charlotte than they won.

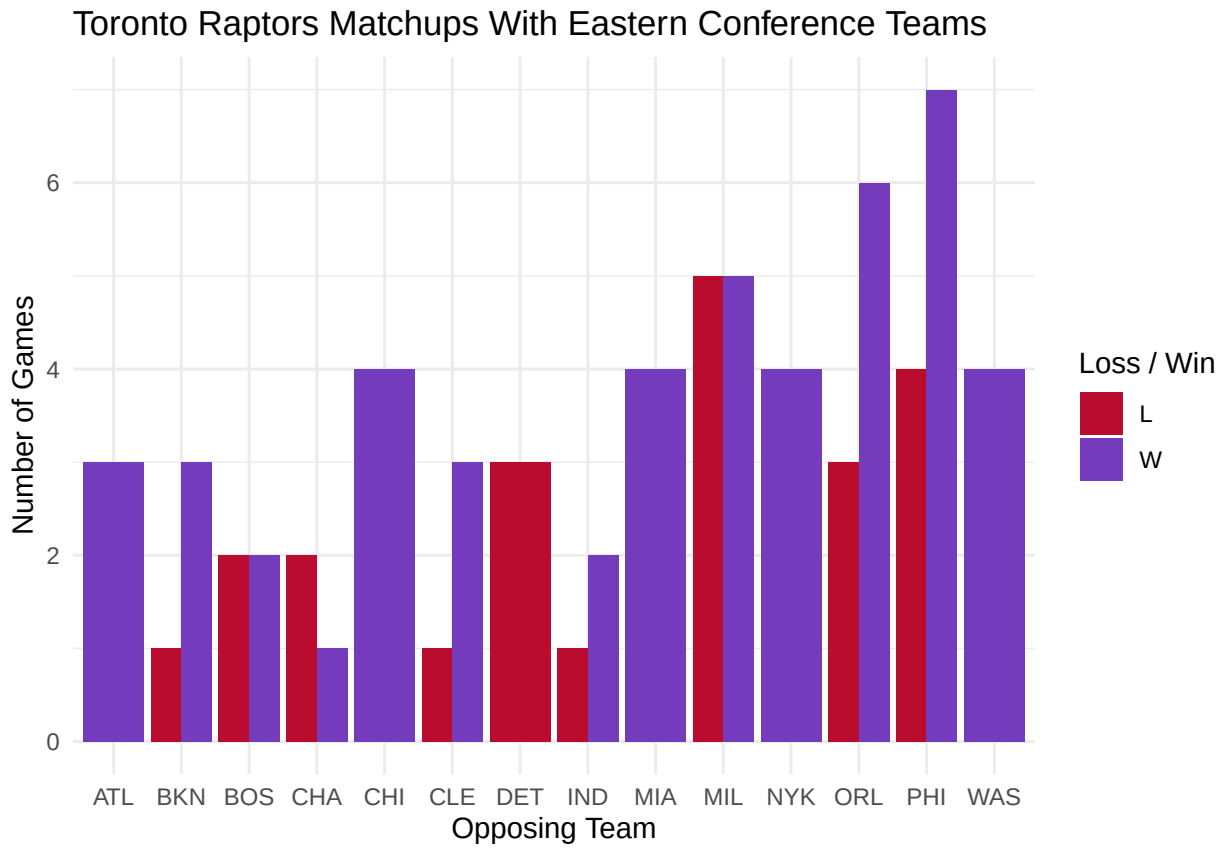

Figure 5: Toronto Raptors Matchups With Eastern Conference Teams

Figure 6 is interesting as it appears as though the TR swept more teams in the Western conference than they did in the East. It is interesting because the Western conference is usually said to be a much more difficult conference. The TR swept 8/15 of the teams in the West whereas they only swept 5/14 in the East. This could suggest that the teams in the West were not as strong as they had been previous years, or perhaps the teams in the West had more injured star players.

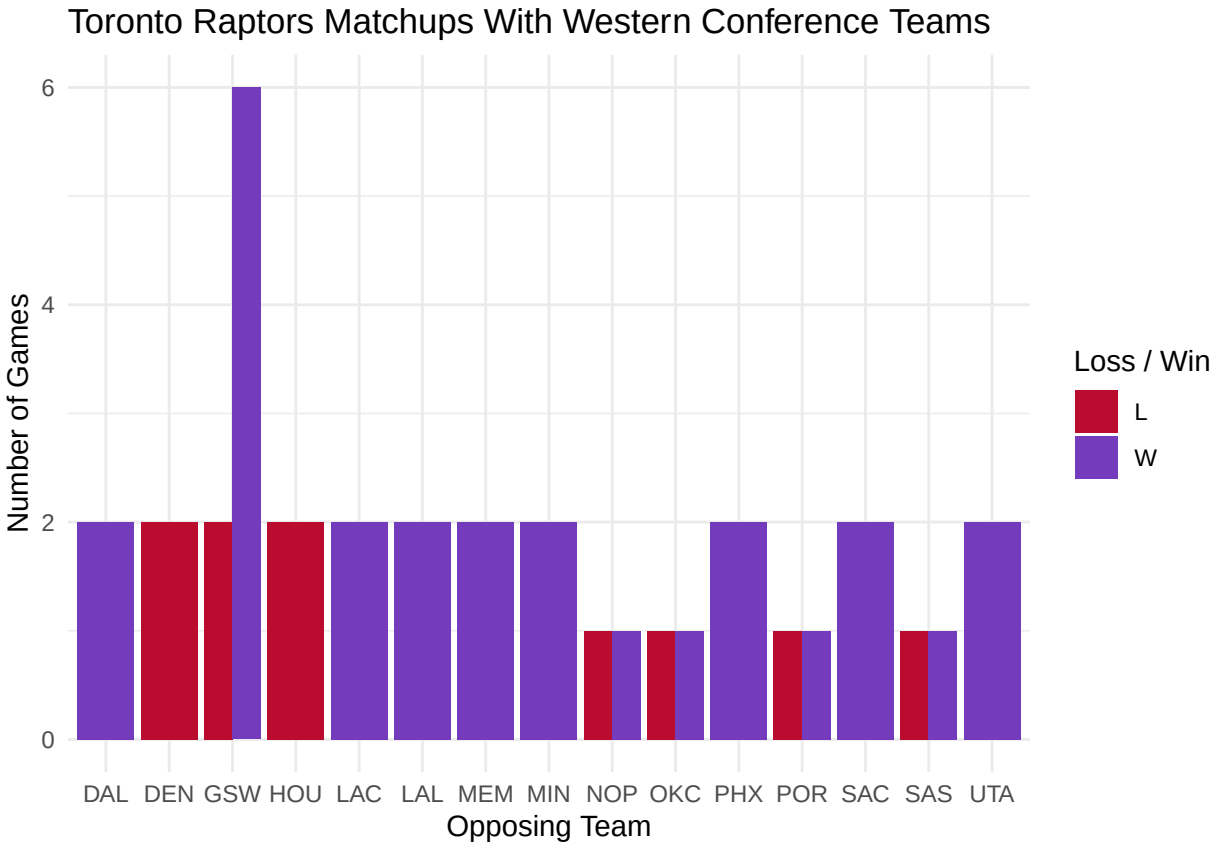

Figure 6: Toronto Raptors Matchups With Western Conference Teams

Figure 7 shows that the TR outscored their opponent most of the time. The graph also shows that whether they won or lost, the game was usually within 20 points.

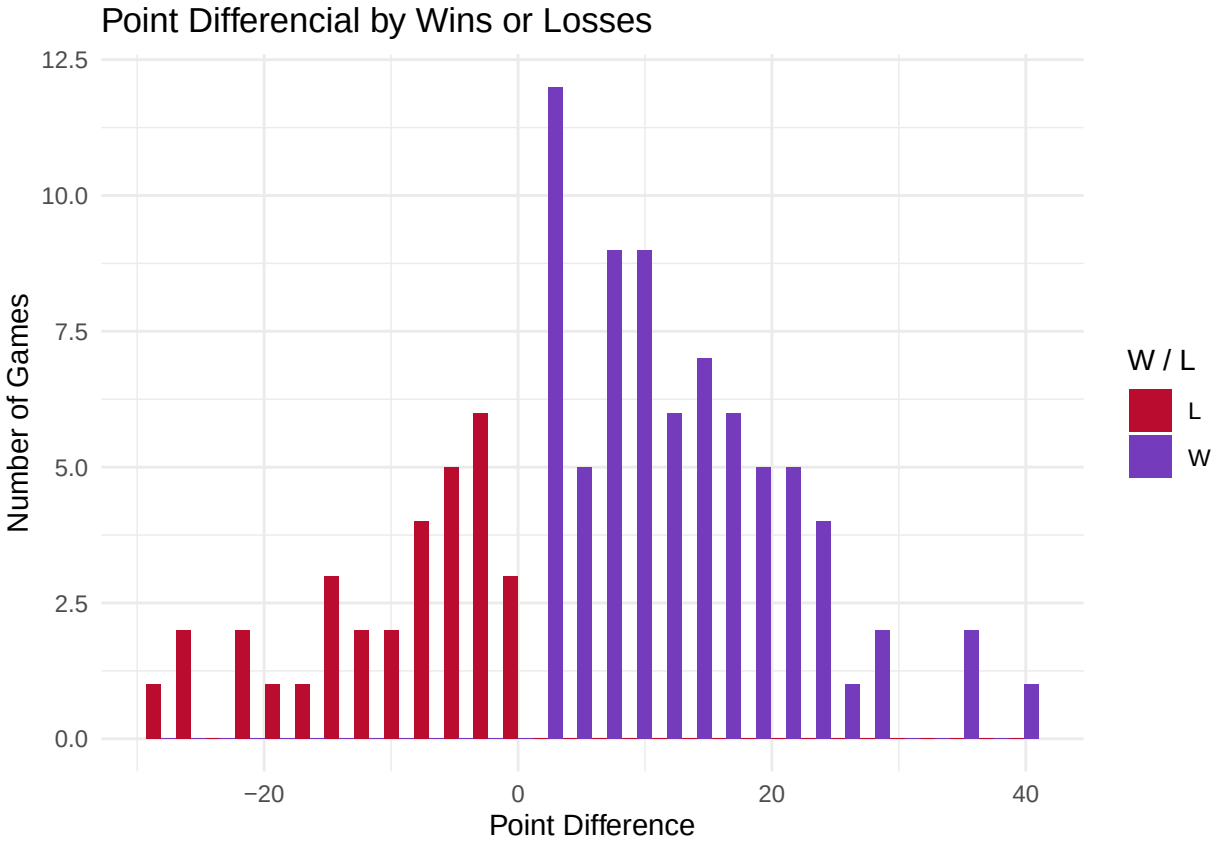

Figure 7: Point Differencial By Wins or Losses

Figure 8 shows a trend where most of the games won usually fell in the 115-125 point range. It also shows that most of the games lost were in the 105-115 range.

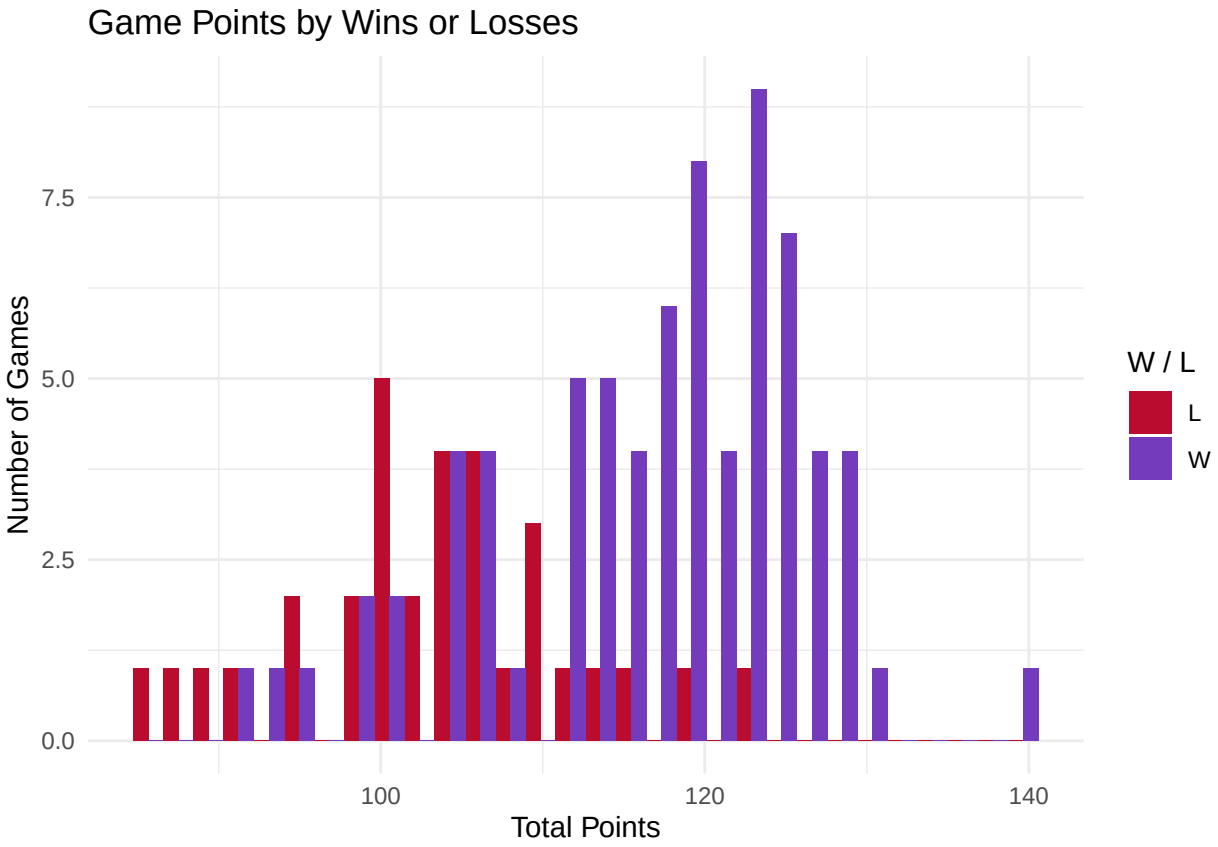

Figure 8: Points By Wins or Losses

In Figure 9, it is evident that there was very little variation in game time. This makes sense as every game is going to be the same length unless the game goes into overtime, or time is added to the clock by the officials. The games that were longer than 260 minutes were most likely overtime games, and it seems as though if a game went in to overtime, it was more likely that the TR would win.

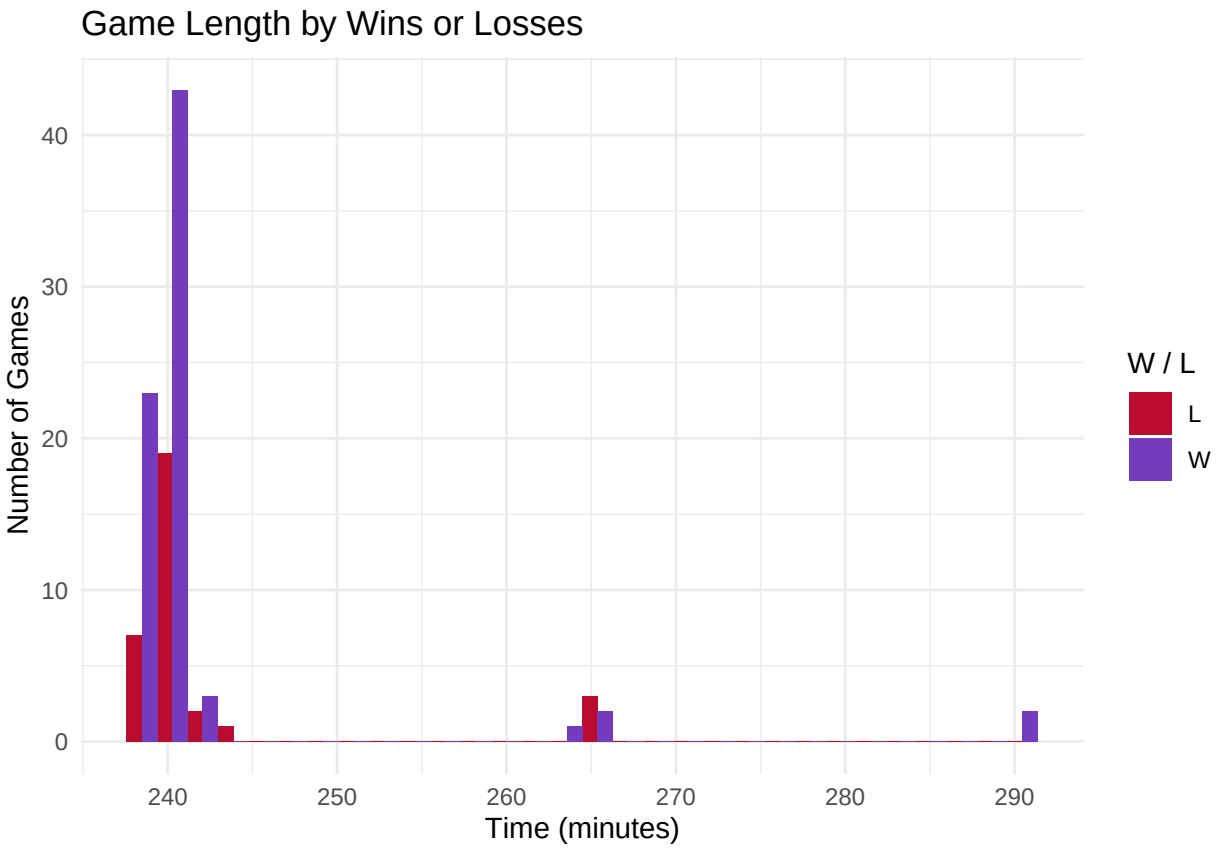

Figure 9: Game Length By Wins or Losses

Figure 10 shows that the TR were more likely to win if they made more than 40 field goals per game, whereas there is less correlation between the amount of field goals and losses.

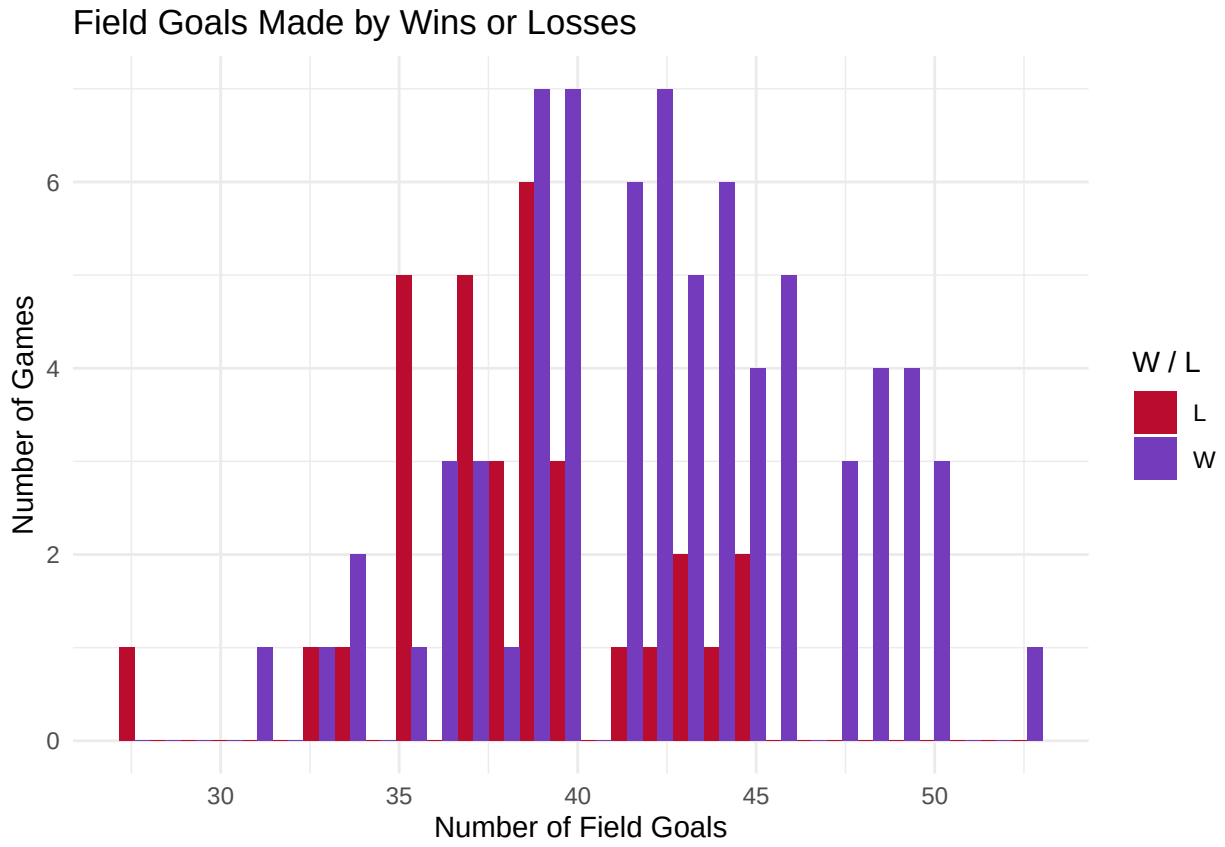

Figure 10: Field Goals Made By Wins or Losses

Figure 11 shows that there is not much correlation between the number of 3 point shots and whether the Raptors won or lost until the tails of the graph. If the TR made 18 3 point shots or more, they were probably going to win the game, therefore the number of 3 point shots made might not be a determining factor in the outcome of a game.

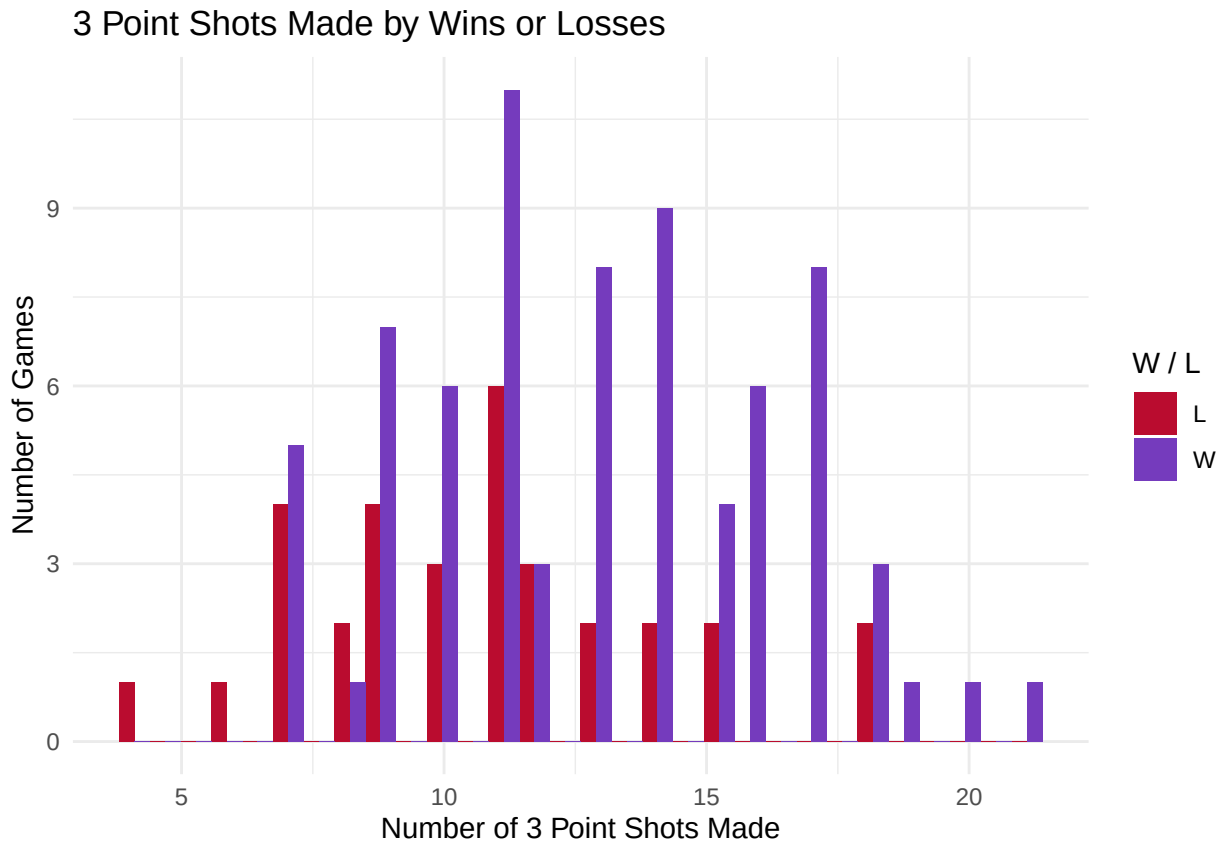

Figure 11: 3 Point Shots Made By Wins or Losses

Figure 12 also shows that the number of free throws might not be a good indicator of the outcome of a game, as it only becomes evident a game would be won if the TR made more than 25 free throws in a game.

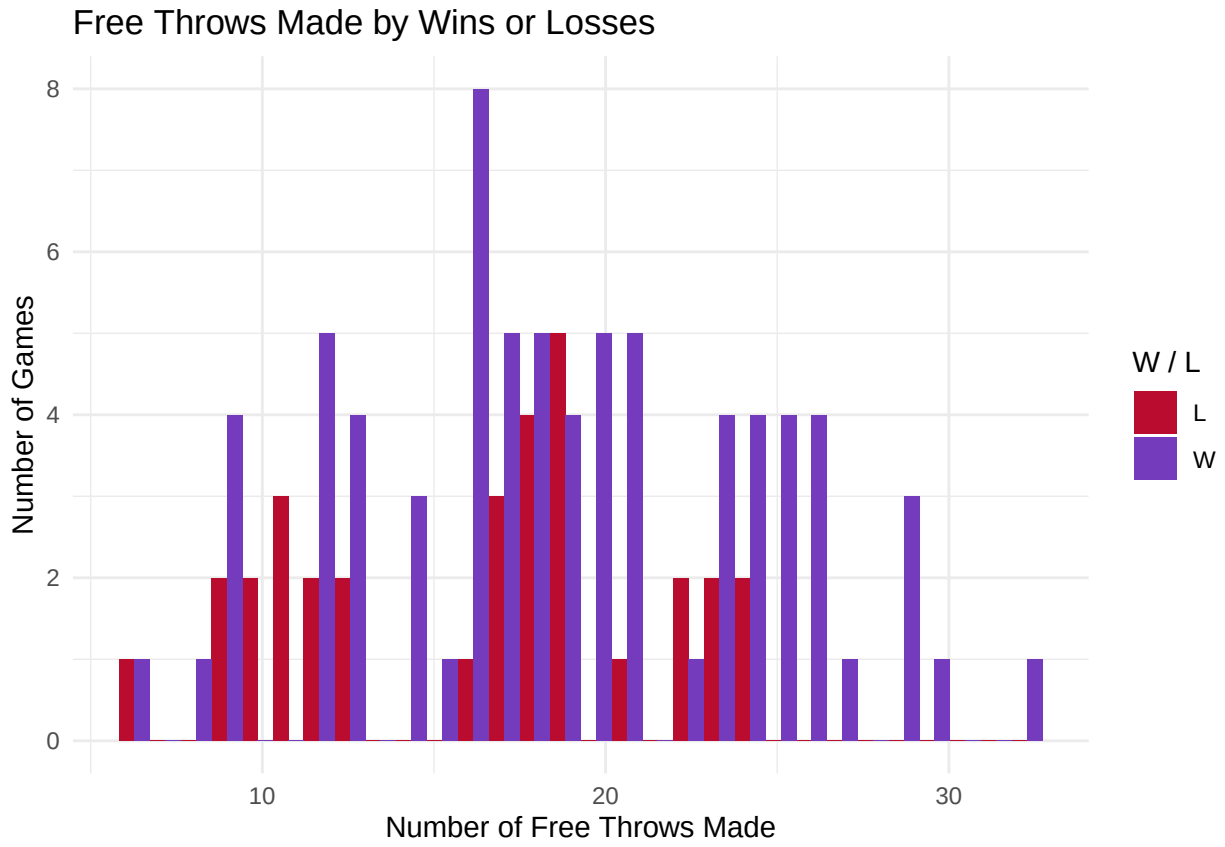

Figure 12: Free Throws Made By Wins or Losses

Figure 13 shows that a minimum of 20 rebounds were going to be required to win a game in the 2018-2019 season. It also shows that rebounds are not as important in determining the outcome of a game when playing against teams from the Northwest and Southwest divisions.

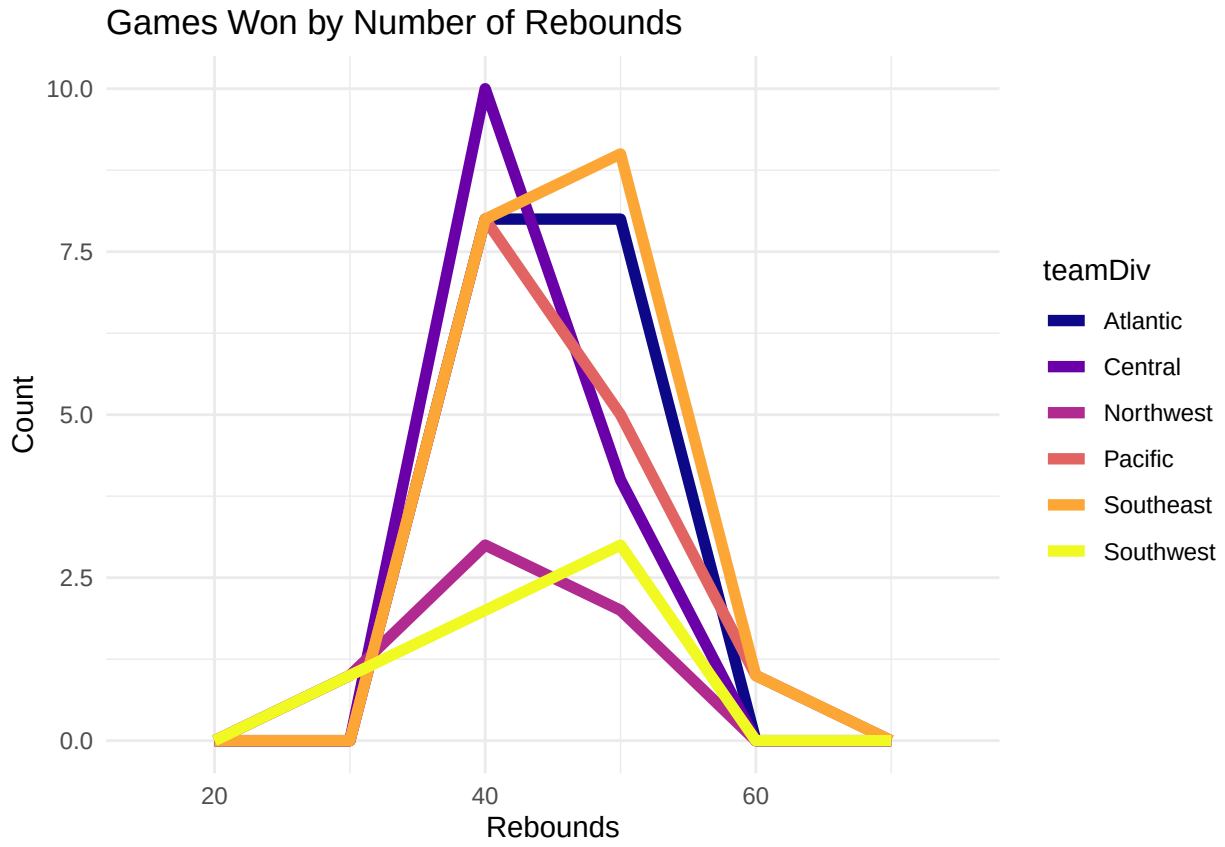

Figure 13: Games Won By Number of Rebounds and Division

Figure 14 shows that a player on the TR would have needed to make more passes and help create shots when playing teams from the Atlantic, Central, Pacific, and Southeast Divisions. The TR needed to have made more assist when playing teams from the Northwest and the Southwest.

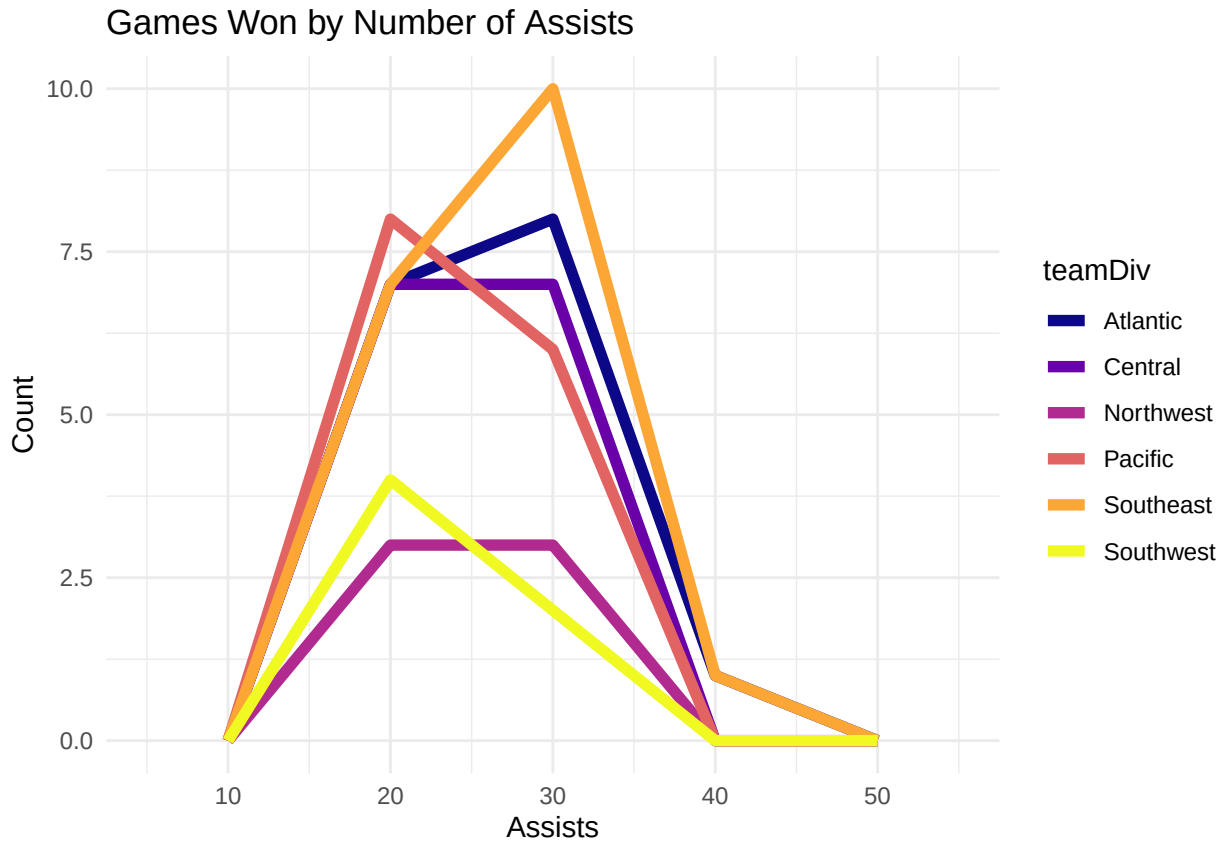

Figure 14: Games Won By Number of Assists and Division

Figure 15 shows that the teams from the Central division were good at drawing fouls from the TR players. It could also indicate referees calling fouls in favour of Central division teams more often than they do for other teams. The plays that the TR coaches plan out for Central division teams might have to be different than the ones they use for the rest of the league.

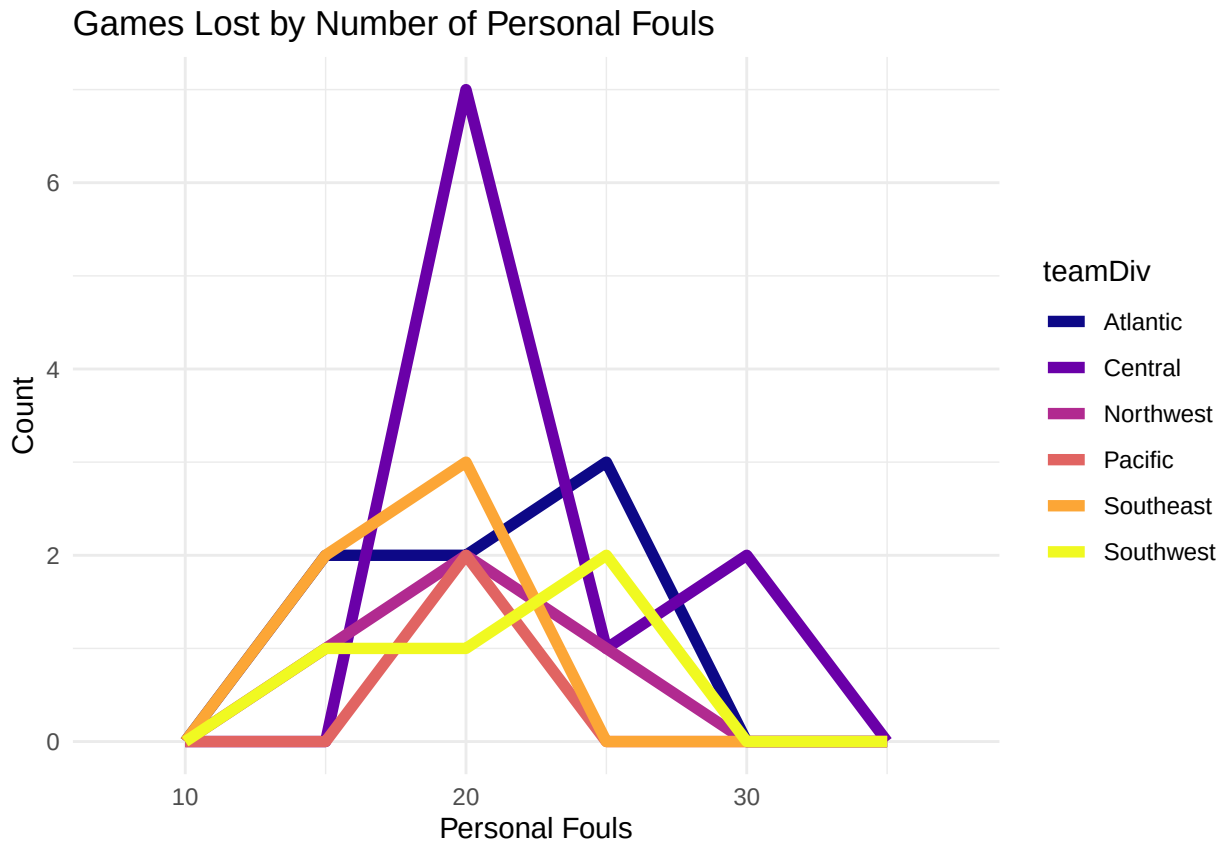

Figure 15: Games Lost By Number of Personal Fouls and Division

Exploration of Covariation

Covariance is how one variable corresponds to another. If one variable increase when another increases, then there is positive covariance between the two variables. If one variable decreases while another increases, then the covariance is negative.

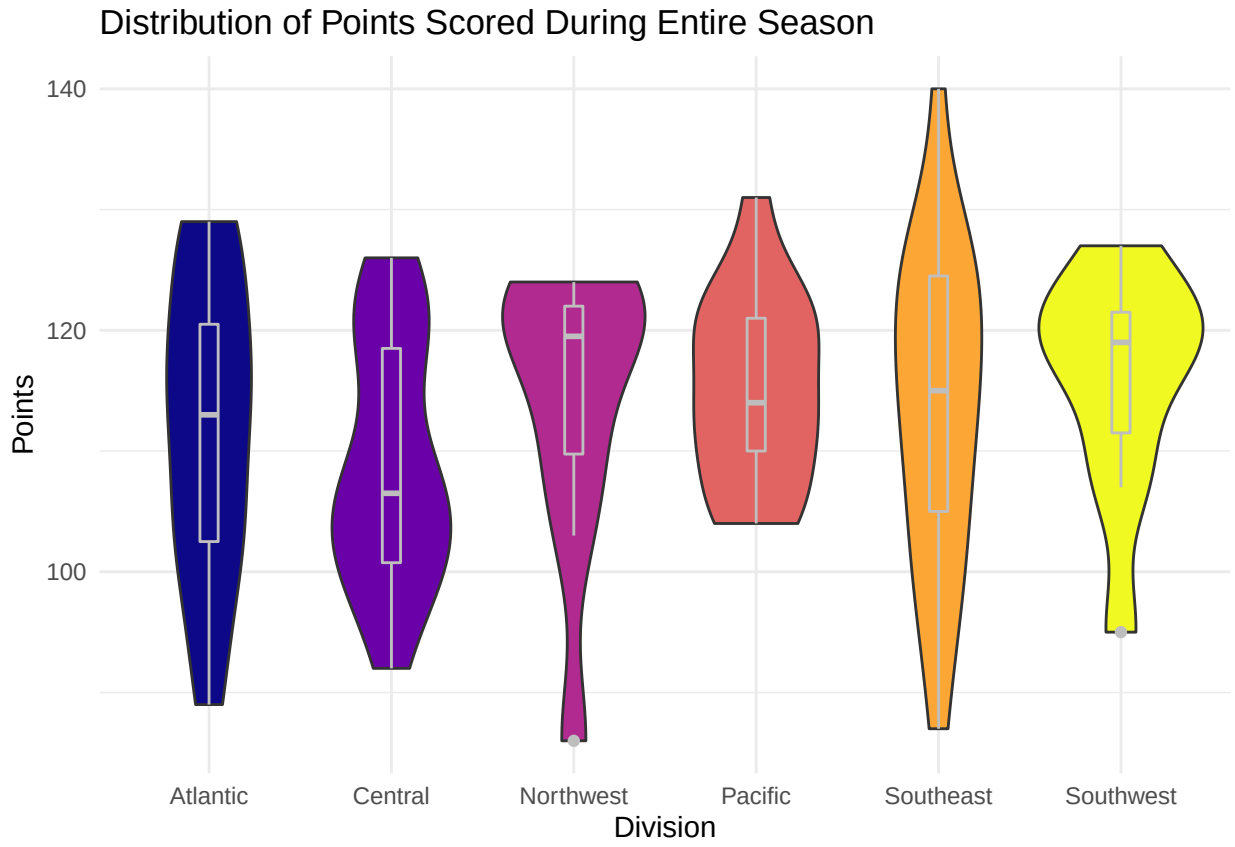

Figure 16: Distribution of Points Scored

Figure 16, 17 and 18 are violin plots with the box-plots over top. Figure 16 shows the distribution of points scored in a game. Figure 17 shows that when the Raptors played against teams from the Northwest and Southwest, the final scores in winning games were all going to be within a small range. Games played against teams from the other divisions had a much wider dispersion of final scores. Figure 18 shows that there was no correlation between the score of lost games and the division except for opponents from the Pacific division. Games lost to teams from the Pacific division were all going to be within a very small range of values.

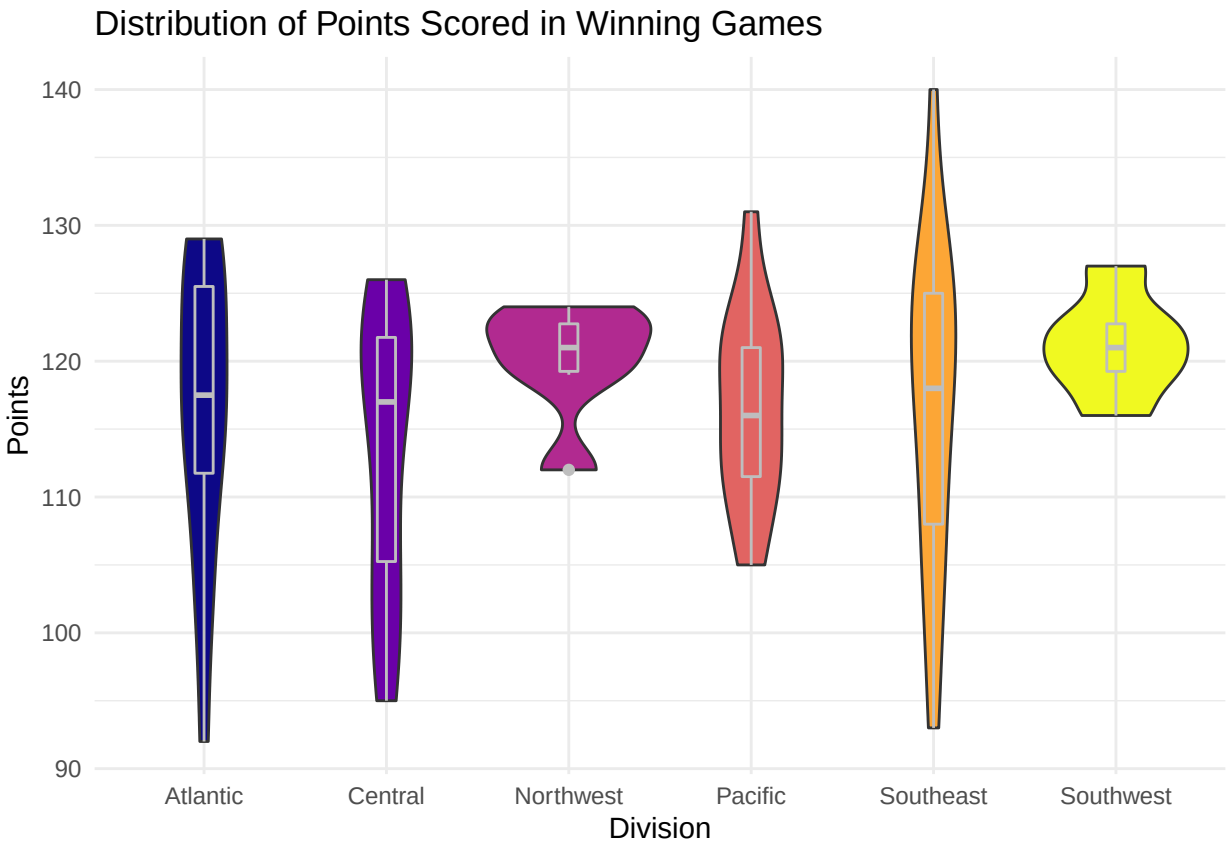

Figure 17: Distribution of Points Scored in Winning Games

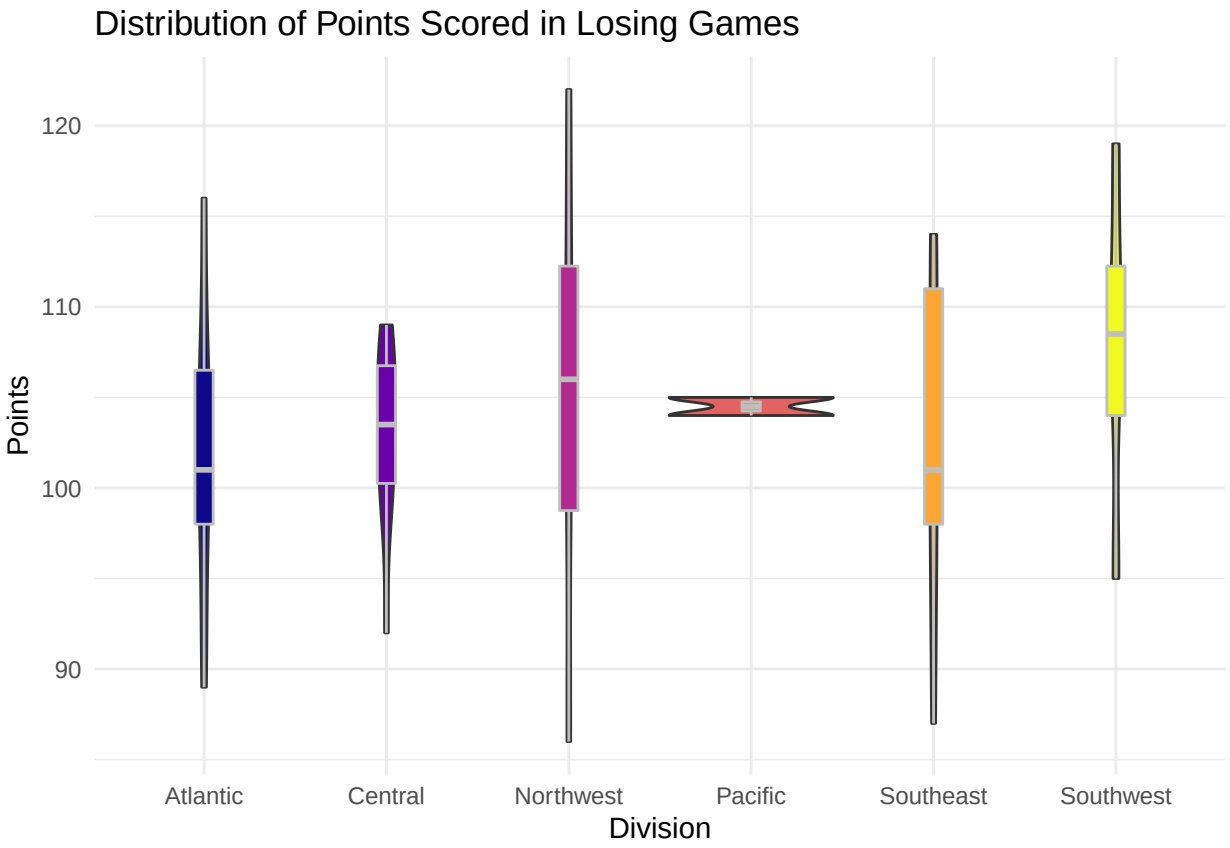

Figure 18: Distribution of Points Scored in Losing Games

Figure 19 shows a correlation matrix of the stats from winning games. Red represents a high correlation, and dark grey represents a low correlation. Looking at the plot, the more free throws that were attempted lead to more free throws made. The same can be said about the 3 point shots and the field goals. There is also a high correlation between rebounds and the attempted number of field goals. There is some correlation between assists and the various types of shots made away from the free throw line. There is a little correlation between personal fouls and free throws, which makes sense as getting personal fouls later in a game will lead to free throws. The more assists that are made correlate to the point differential in final game scores. There is also a high correlation between the number of rebounds and the attempted field goals, because as a player gets a rebound or a defensive rebound, they will most likely make another attempt at a basket.

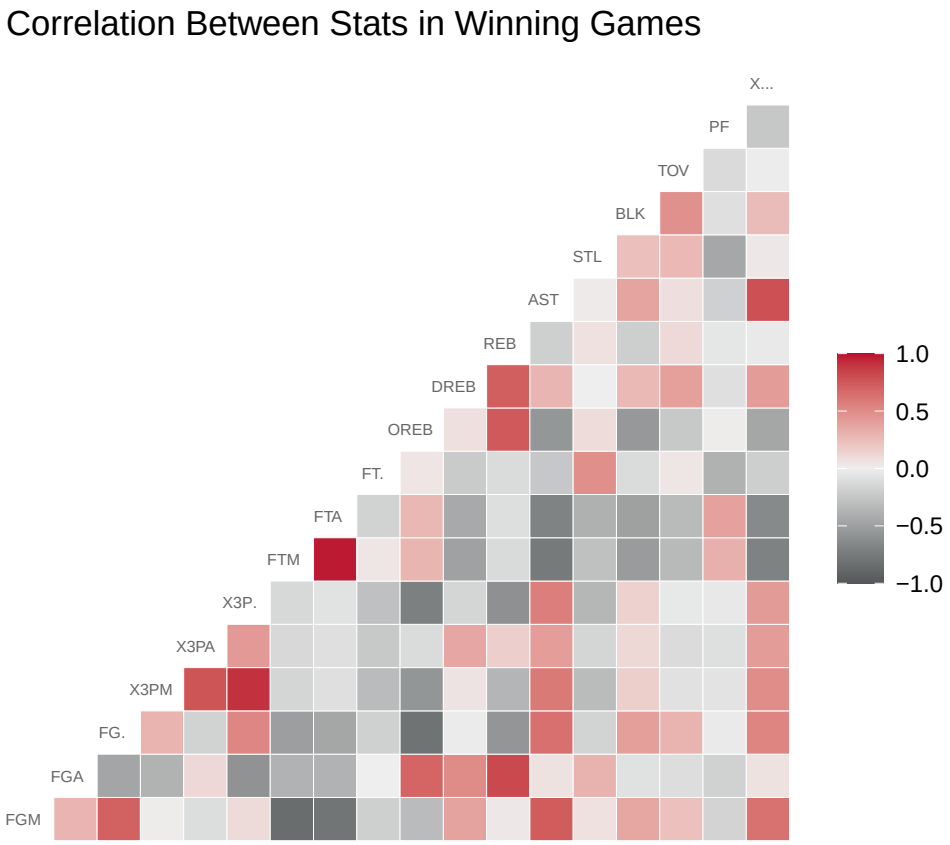

Figure 19: Correlation Between Stats From Winning Games

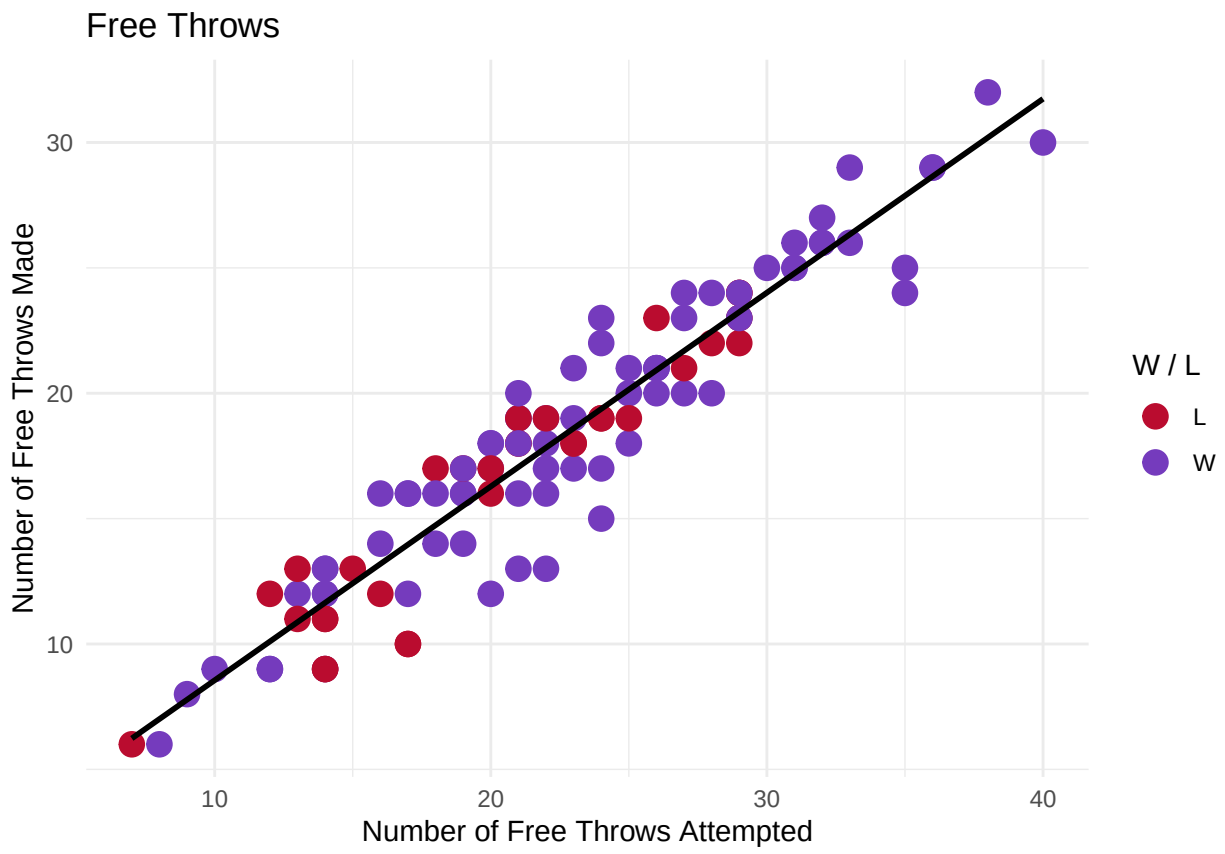

Figure 20: Free Throws

Figure 20 shows that there is a positive correlation between the number of attempted free throws and the number of free throws that are made, regardless of the outcome of the game.

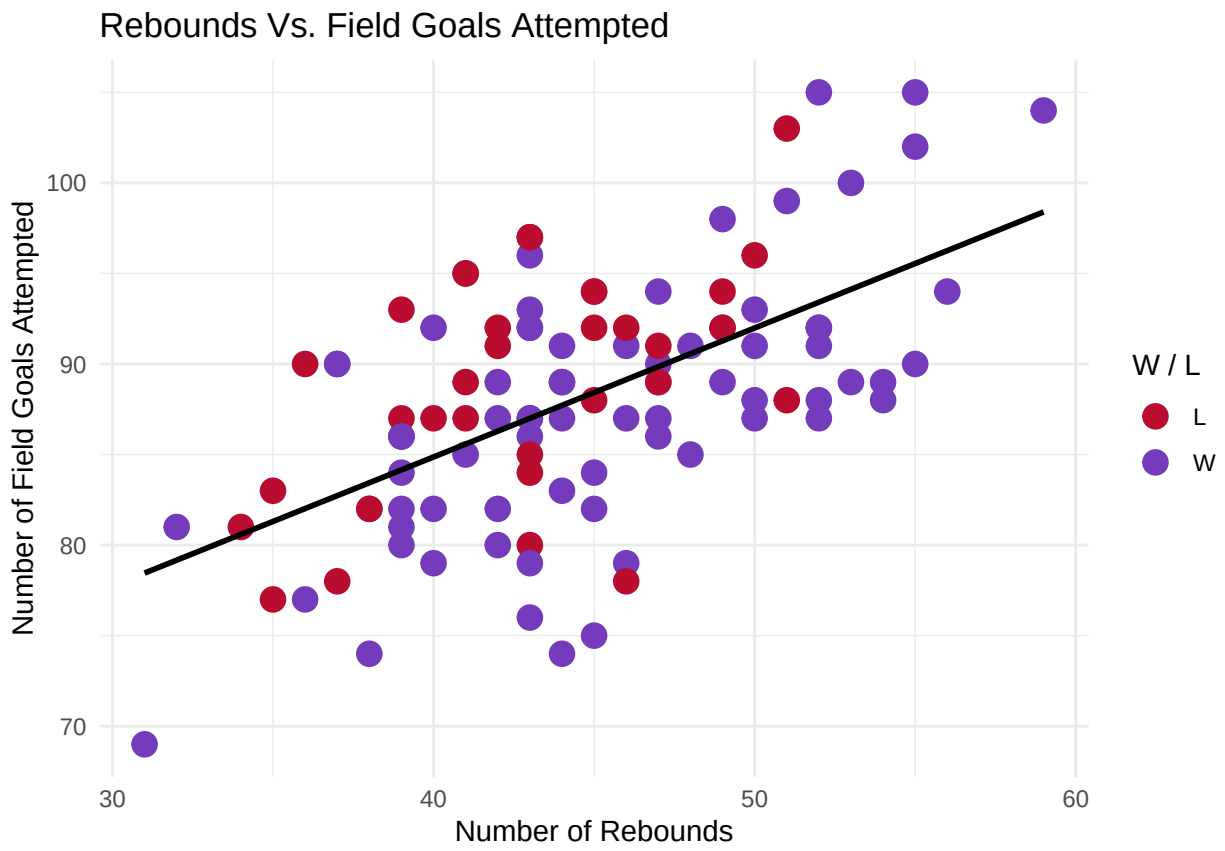

Figure 21: Rebounds Vs. Field Goals Attempted

Figure 21 shows a positive correlation between rebounds and attempted field goals, though the the correlation is not very strong as the slope is not as steep and the points are not grouped tightly to the line.

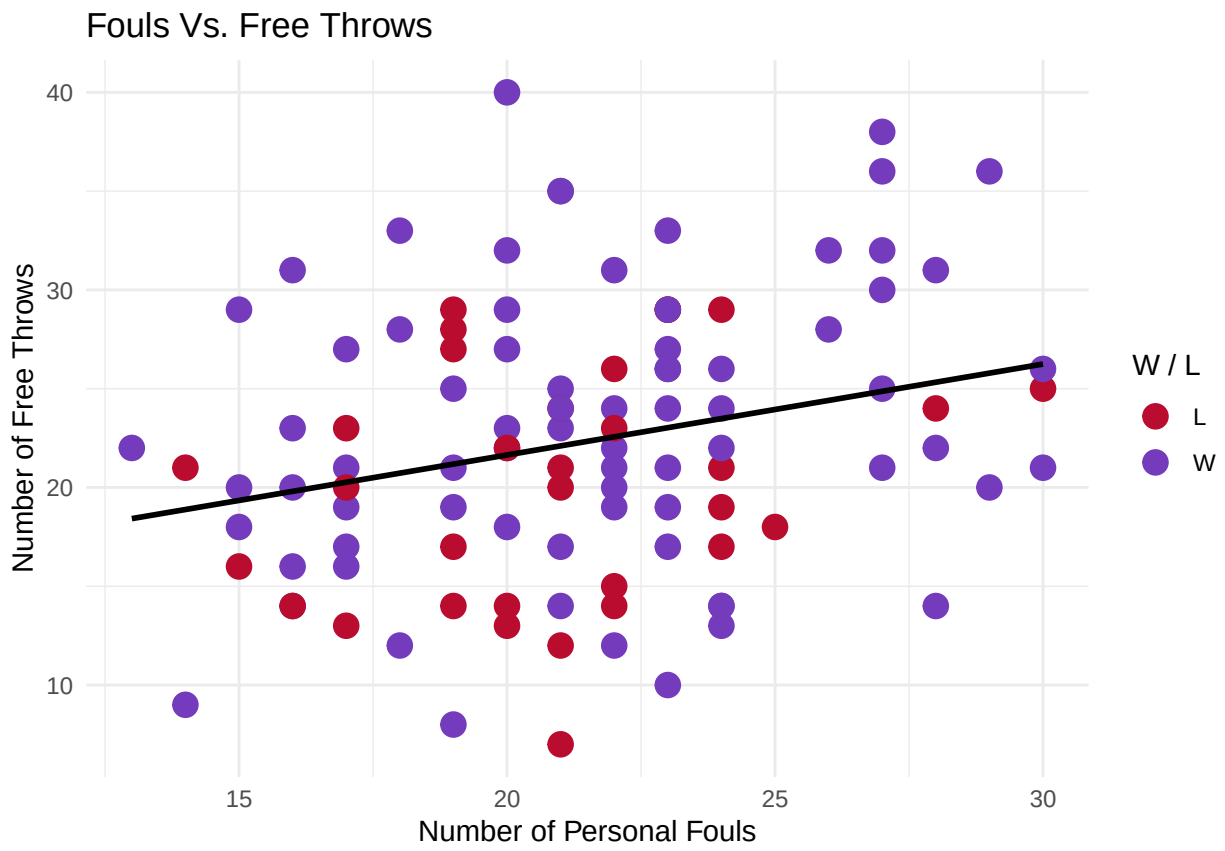

Figure 22: Fouls Vs. Free Throws

The co-variance is positive, but weak in Figure 22.

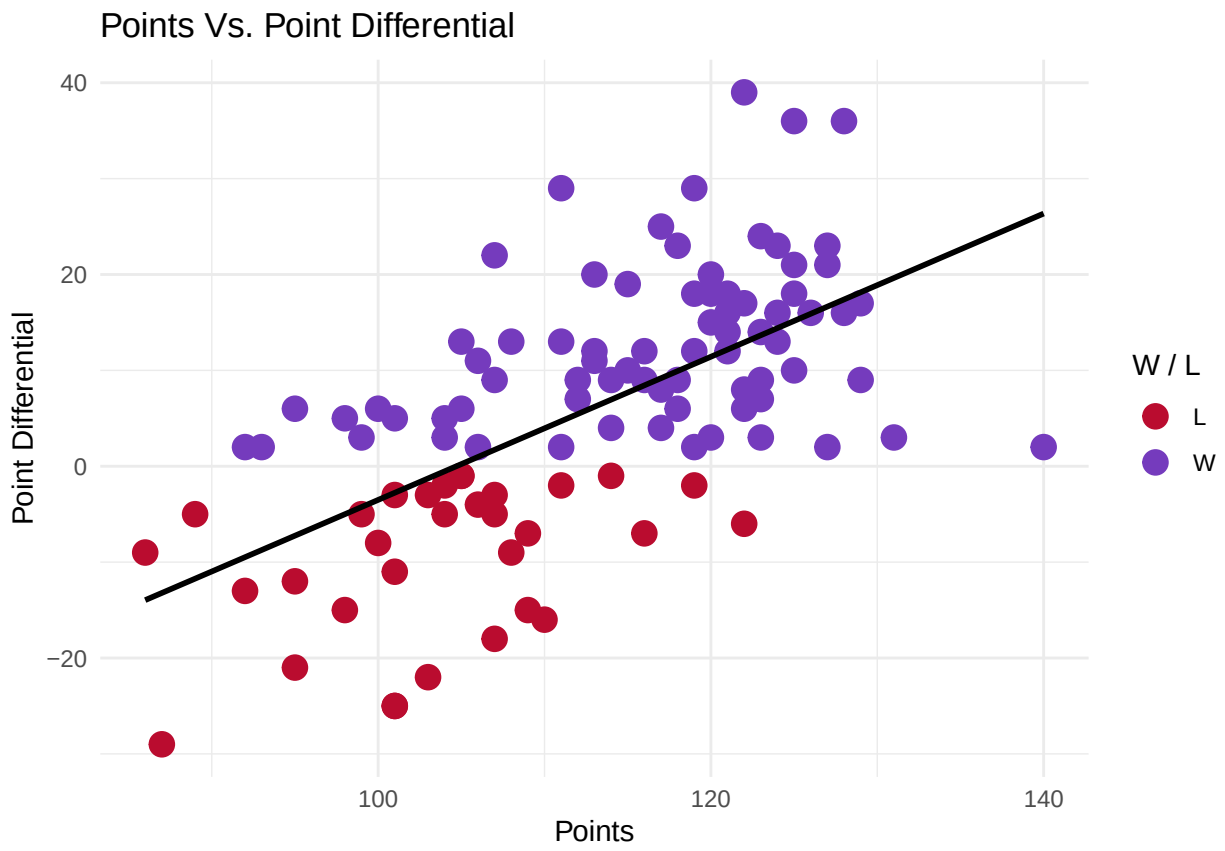

Figure 23: Points Vs. Point Differential

Figure 23 shows positive co-variance between the point differential and final score.

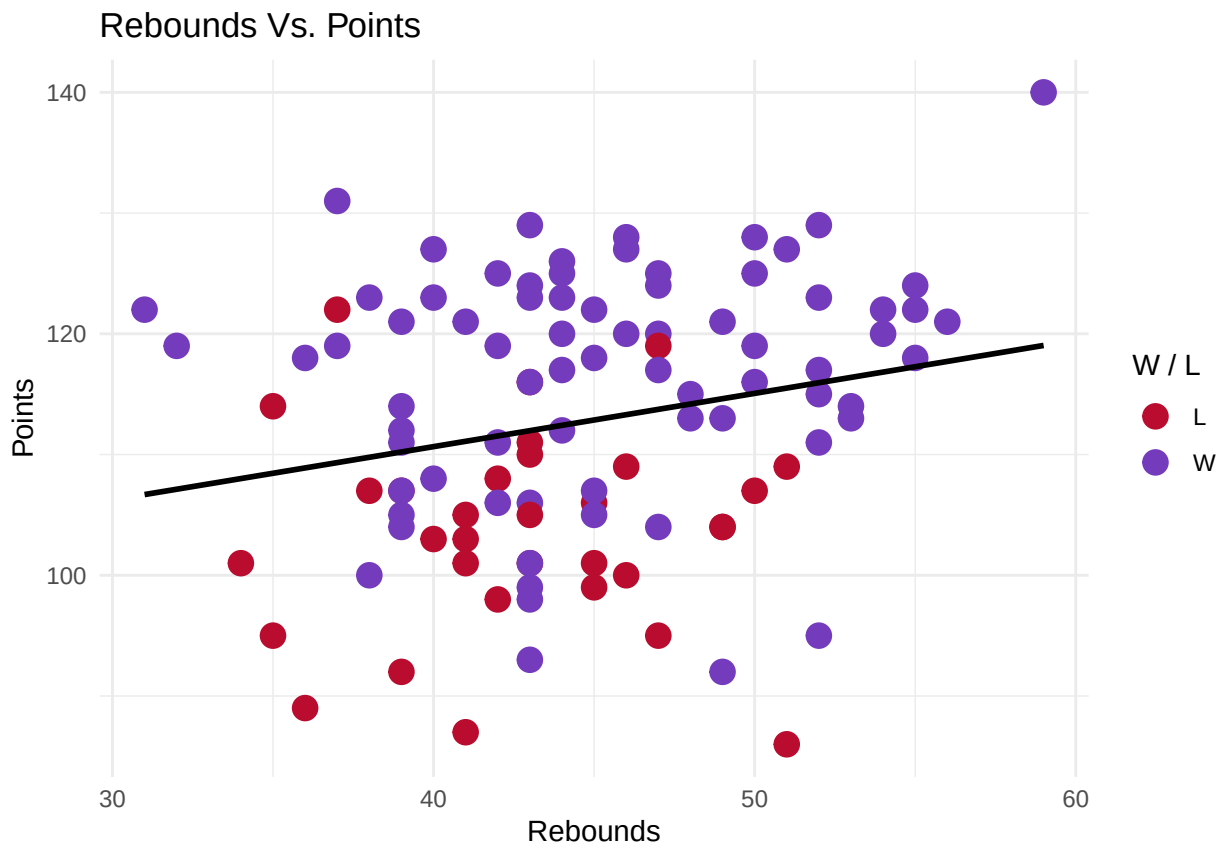

Figure 24: Rebounds Vs. Points

Figure 24 shows a positive covariance between rebounds and points, with most losses occuring when the number of rebounds is below the linear trendline.

Figure 25 shows a negative covariance between the number of turn-overs and the point differential, meaning that the more turn-overs the TR make, the more likely they will lose a game.

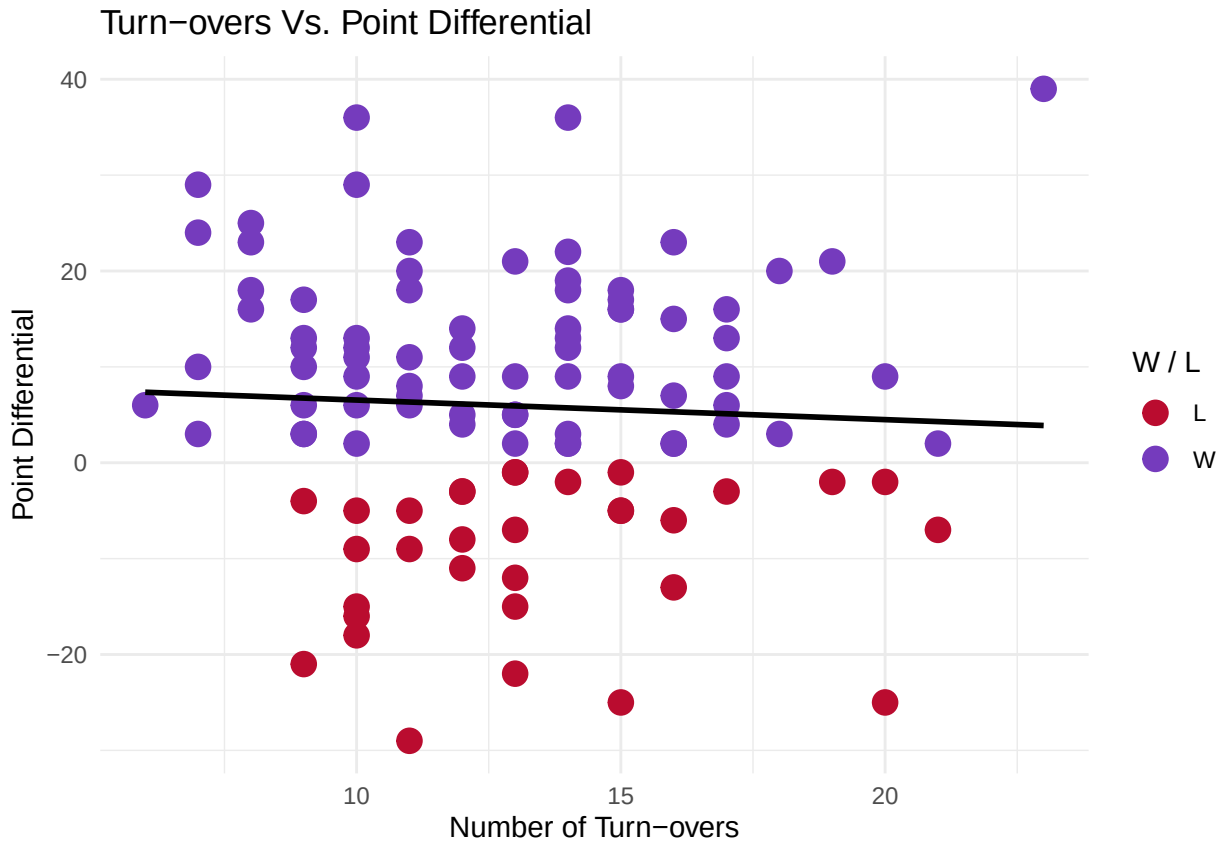

Figure 25: Turn-overs Vs. Point Differential

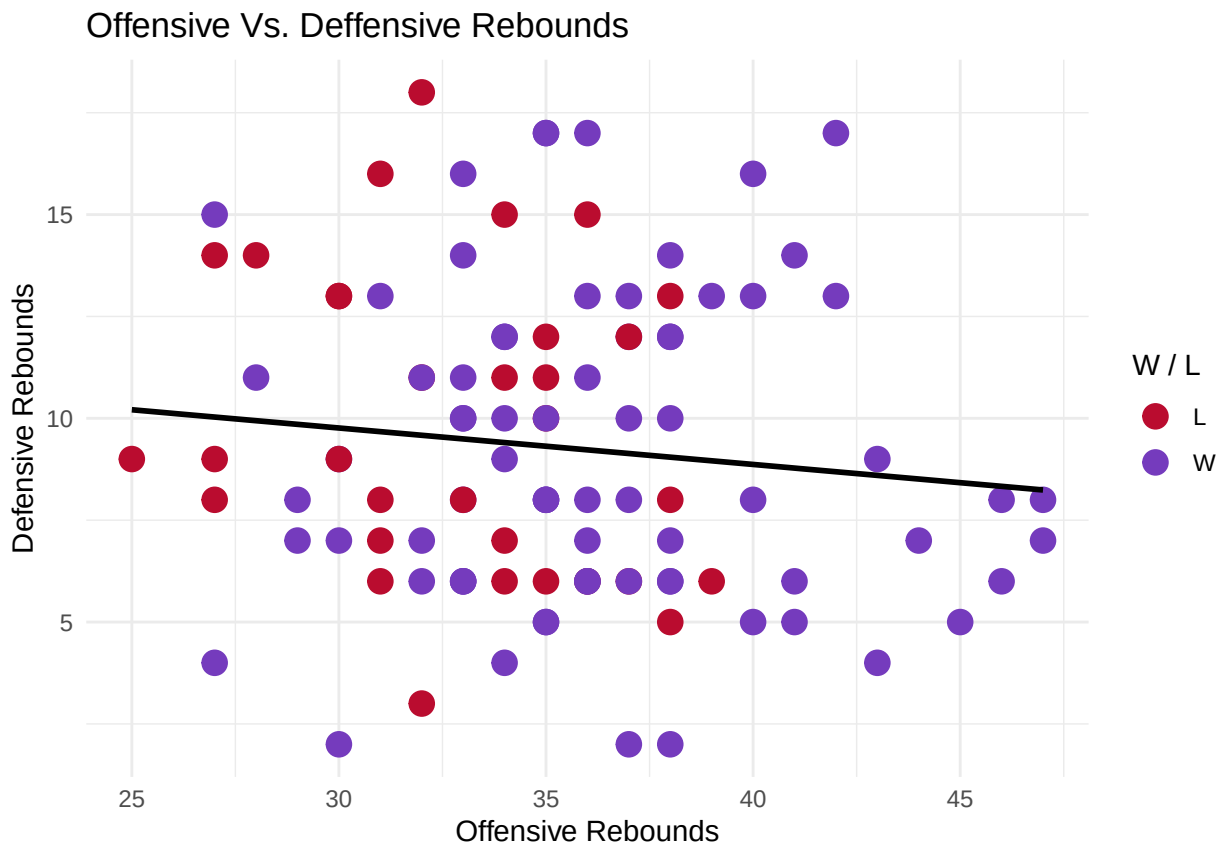

Figure 26: Offensive Vs. Defensive Rebounds

Figure 26 shows no correlation between the number of offensive rebounds and the number of defensive rebounds.

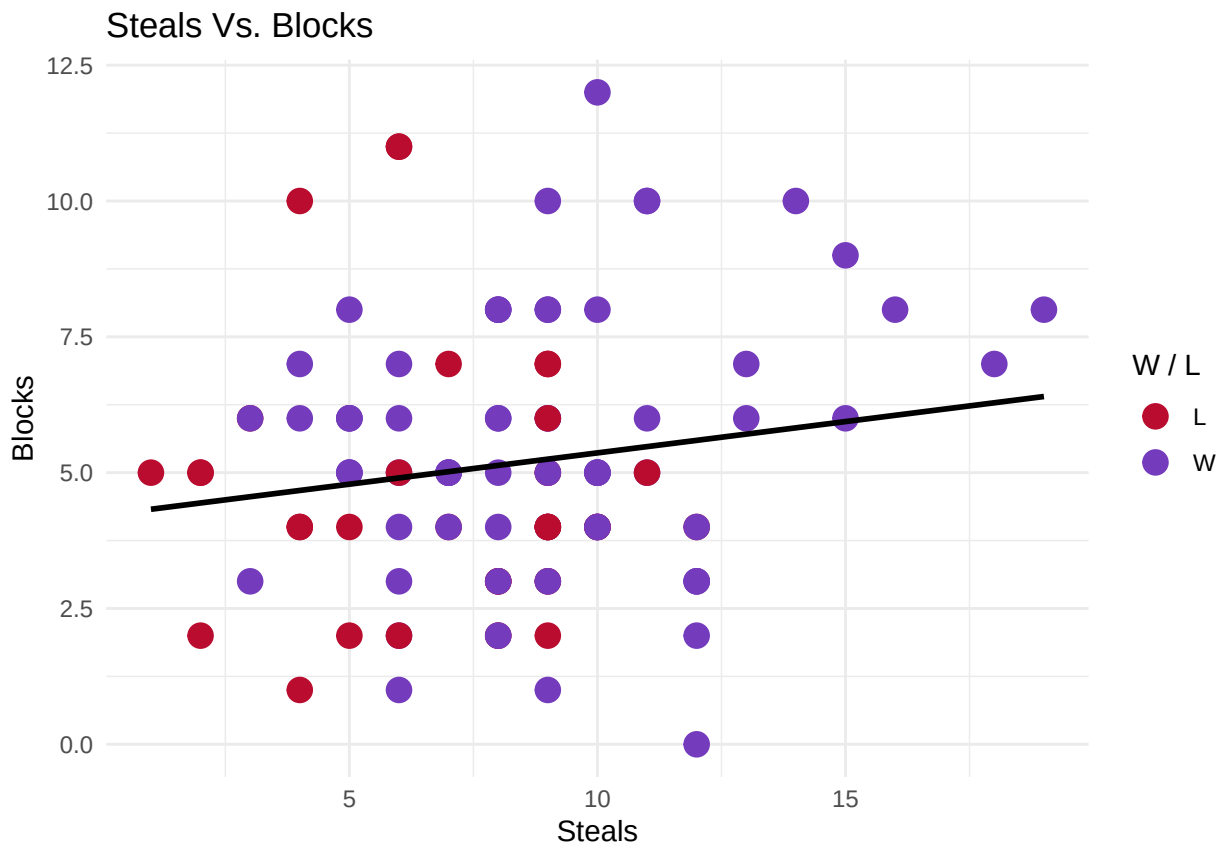

Figure 27: Steals Vs. Blocks

Figure 27 shows a positive correlation between the number of blocks and the number of steals, indicating that the more blocks the TR made, the more likely they would get a steal off of the play. The more steals and blocks the TR made, the more likely they would win a game.

Discussion

I had initially thought that there would be a stronger correlation between some stats and games being won. Through the Exploratory Data Analysis I think that the number of certain plays is not going to be a significant factor in whether or not a game is won. There are more positive or negative correlation between individual stats, such as in Figure 22 where the number of free throws attempted is strongly correlated to the number of free throws made. I think that there is not really a single type of play that is a strong contributor to the total number of points. I would think that this means that a NBA games over-all are not won by one team being really good at a certain play, therefore the teams can't really stack their rosters full of players who are all good at the same thing, they must have a variety of players who can contribute across a variety of plays.

I had also expected to see that the TR won more games than they lost in most of the divisions, which is true based on Figures 4, 5, and 6. I also guessed that having the home court-advantage would be a contributing factor to winning games. This was true as evident in Figure 3, where playing at home during the regular season lead to more games being won when compared to being the away team. Playing on the home court was still beneficial during the play-offs, although there was not as big of a difference between wins and losses during the playoffs.

Visualizing the data has helped in better understanding the data set. I think there could be much more done with this data set if a future analysis was done where it included another table with the stats of individual players. This would be interesting as it would enable you to see which players are contributing the most to which stats, and how that affects the outcome of games. It would also be interest to have a 2D or 3D map of the court to see where shots are being made from, and where each player has the best chance of making a shot. The analysis could probably get very complex if you included all the stats from all players in the NBA, to see which combination of players play well against certain combinations of players on opposing teams. I would assume that that is why NBA teams usually have several statisticians working for them and analysis things on the side of the court as games are being played. It would be interesting to know which stats the teams are most interested in, and how they use all the data for planning out their strategies for defeating other teams.

Appendix

Definitions of Stat Codes

Table 3: Stat Codes and Names

Code	**Definition**
TEAM	Toronto Raptors
DATE	Played On Date
MATCHUP	Match Up
W.L	Win or Loss
MIN	Minutes Played
PTS	Points
FGM	Field Goals Made
FGA	Field Goals Attempted
FG.	Field Goal Percentage
X3PM	3 Point Shots Made
X3PA	3 Point Shots Attempted
X3P.	3 Point Shot Percentage
FTM	Free Throws Made
FTA	Free Throws Attempted
FT.	Free Throw Percentage
OREB	Offensive Rebound
DREB	Deffensive Rebound
REB	Rebound Total
AST	Assists
STL	Steals
BLK	Blocks
TOV	Turn-overs
PF	Personal Fouls
X...	+/-
R_P	Regular or Playoff
H_A	Home or Away
teamAbbr	Opposing Team
teamConf	Opposing Team Conference
teamDiv	Opposing Team Division

Structure of Team Table

```
str(team)
```

```
## 'data.frame':    106 obs. of  29 variables:
##  $ TEAM    : chr  "TOR" "TOR" "TOR" "TOR" ...
##  $ DATE    : chr  "01/13/2019" "11/29/2018" "02/13/2019" "10/30/2018" ...
##  $ MATCHUP : chr  "TOR @ WAS" "TOR vs. GSW" "TOR vs. WAS" "TOR vs. PHI" ...
##  $ W.L     : Factor w/ 2 levels "L","W": 2 2 2 2 2 2 2 2 2 2 ...
##  $ MIN     : int  290 265 241 240 240 239 240 240 241 239 ...
##  $ PTS     : int  140 131 129 129 128 128 127 127 127 126 ...
##  $ FGM     : int  49 47 44 47 49 42 53 50 50 48 ...
##  $ FGA     : int  104 90 92 92 88 79 99 92 87 91 ...
##  $ FG.     : num  47.1 52.2 47.8 51.1 55.7 53.2 53.5 54.3 57.5 52.7 ...
##  $ X3PM    : int  13 15 16 11 17 14 13 9 15 9 ...
##  $ X3PA    : int  36 38 38 27 40 35 37 31 39 31 ...
##  $ X3P.    : num  36.1 39.5 42.1 40.7 42.5 40 35.1 29 38.5 29 ...
##  $ FTM     : int  29 22 25 24 13 30 8 18 12 21 ...
##  $ FTA     : int  36 24 30 28 22 40 9 20 14 26 ...
##  $ FT.     : num  80.6 91.7 83.3 85.7 59.1 75 88.9 90 85.7 80.8 ...
##  $ OREB    : int  17 7 9 9 5 12 7 5 5 13 ...
##  $ DREB    : int  42 30 43 34 45 34 44 35 41 31 ...
##  $ REB     : int  59 37 52 43 50 46 51 40 46 44 ...
##  $ AST     : int  24 25 32 29 36 24 31 32 36 21 ...
##  $ STL     : int  16 5 8 13 5 8 10 12 10 9 ...
##  $ BLK     : int  8 6 8 6 6 4 12 3 8 5 ...
##  $ TOV     : int  21 14 14 15 14 15 16 13 13 8 ...
##  $ PF      : int  27 23 27 26 22 20 14 16 21 23 ...
##  $ X...    : int  2 3 9 17 36 16 23 2 21 16 ...
##  $ R_P     : Factor w/ 2 levels "P","R": 2 2 2 2 2 2 2 2 2 2 ...
##  $ H_A     : Factor w/ 2 levels "A","H": 1 2 2 2 2 2 1 2 2 2 ...
##  $ teamAbbr: Factor w/ 29 levels "ATL","BKN","BOS",..: 29 10 29 23 20 20 19 2 4 6 ...
##  $ teamConf: Factor w/ 2 levels "East","West": 1 2 1 1 1 1 2 1 1 1 ...
##  $ teamDiv : Factor w/ 6 levels "Atlantic","Central",..: 5 4 5 1 1 1 6 1 5 2 ...
```

Summary of Team Table

```
##      TEAM               DATE              MATCHUP          W.L
##  Length:106         Length:106         Length:106         L:32
##  Class :character   Class :character   Class :character   W:74
##  Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##       MIN             PTS             FGM             FGA
##  Min.   :238.0   Min.   : 86.0   Min.   :28.00   Min.   : 69.0
##  1st Qu.:239.0   1st Qu.:105.0   1st Qu.:38.00   1st Qu.: 84.0
##  Median :240.0   Median :114.0   Median :41.00   Median : 88.5
##  Mean   :242.3   Mean   :112.7   Mean   :41.17   Mean   : 88.1
##  3rd Qu.:241.0   3rd Qu.:121.8   3rd Qu.:44.00   3rd Qu.: 92.0
##  Max.   :290.0   Max.   :140.0   Max.   :53.00   Max.   :105.0
##
##       FG.             X3PM            X3PA            X3P.
##  Min.   :29.50   Min.   : 4.00   Min.   :20.00   Min.   :20.00
##  1st Qu.:42.70   1st Qu.:10.00   1st Qu.:30.00   1st Qu.:29.00
##  Median :47.35   Median :12.00   Median :34.00   Median :35.80
##  Mean   :46.93   Mean   :12.27   Mean   :33.94   Mean   :36.04
##  3rd Qu.:51.58   3rd Qu.:15.00   3rd Qu.:38.00   3rd Qu.:41.45
##  Max.   :60.90   Max.   :21.00   Max.   :47.00   Max.   :58.10
##
##       FTM             FTA             FT.              OREB
##  Min.   : 6.00   Min.   : 7.00   Min.   : 58.80   Min.   : 2.000
##  1st Qu.:13.25   1st Qu.:17.00   1st Qu.: 76.38   1st Qu.: 6.000
##  Median :18.00   Median :22.00   Median : 81.55   Median : 8.000
##  Mean   :18.05   Mean   :22.28   Mean   : 81.39   Mean   : 9.292
##  3rd Qu.:22.00   3rd Qu.:27.00   3rd Qu.: 87.30   3rd Qu.:12.000
##  Max.   :32.00   Max.   :40.00   Max.   :100.00   Max.   :18.000
##
##       DREB            REB             AST             STL
##  Min.   :25.00   Min.   :31.00   Min.   :13.00   Min.   : 1.000
##  1st Qu.:32.25   1st Qu.:41.00   1st Qu.:21.00   1st Qu.: 6.000
##  Median :35.00   Median :44.00   Median :25.00   Median : 8.000
##  Mean   :35.25   Mean   :44.55   Mean   :24.77   Mean   : 8.236
##  3rd Qu.:38.00   3rd Qu.:49.00   3rd Qu.:29.00   3rd Qu.:10.000
##  Max.   :47.00   Max.   :59.00   Max.   :36.00   Max.   :19.000
##
##       BLK             TOV              PF             X...          R_P
##  Min.   : 0.00   Min.   : 6.00   Min.   :13.00   Min.   :-29.000   P:24
##  1st Qu.: 4.00   1st Qu.:10.00   1st Qu.:19.00   1st Qu.: -2.000   R:82
##  Median : 5.00   Median :13.00   Median :21.00   Median :  6.000
##  Mean   : 5.16   Mean   :12.92   Mean   :21.39   Mean   :  5.943
##  3rd Qu.: 6.00   3rd Qu.:15.00   3rd Qu.:24.00   3rd Qu.: 14.750
##  Max.   :12.00   Max.   :23.00   Max.   :30.00   Max.   : 39.000
##
##  H_A       teamAbbr  teamConf       teamDiv
##  A:52   PHI    :11   East:70   Atlantic :23
##  H:54   MIL    :10   West:36   Central  :24
##         ORL    : 9             Northwest:10
##         GSW    : 8             Pacific  :16
##         BKN    : 4             Southeast:23
##         BOS    : 4             Southwest:10
##         (Other):60
```