

PREDICTIVE MODELLING FOR NBA SPORTS LEAGUE

PROG8430 – Data Analysis, Modeling and Algorithms

Shubham Handa -8638369, Gurjyot Anand– 8578965,

Divya Chandwani-8636981, Hardeep Kaur-8638406

4/16/2019

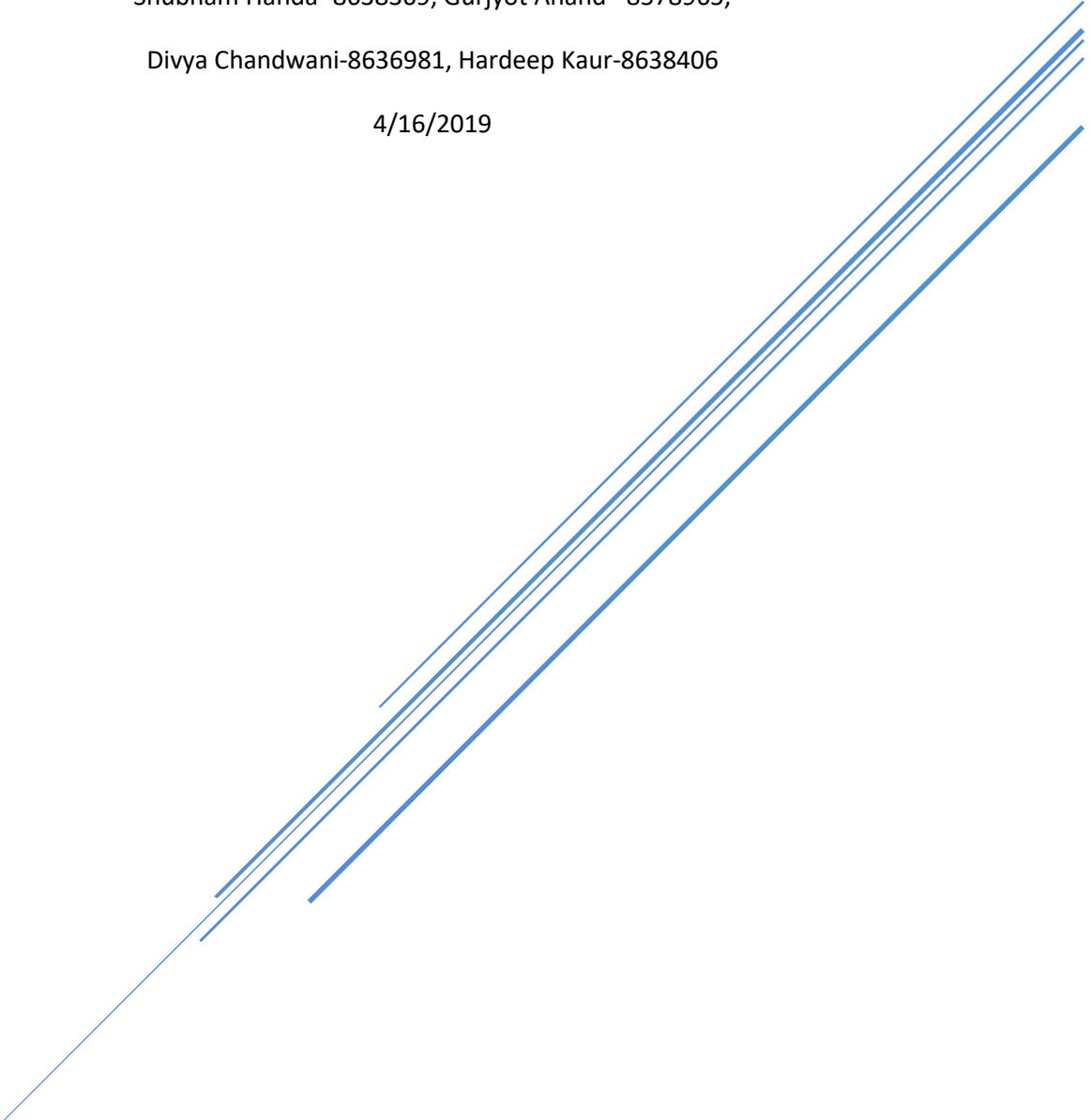
A decorative graphic consisting of several parallel blue lines of varying lengths, arranged diagonally from the bottom-left towards the top-right of the page.

Table of Contents

Introduction.....	2
Motivation.....	2
Dataset.....	3
Problem Statement, Explanations and Methodology.....	4
1) What factors define the income of a of player?	4
2) Do the officials favour Home team?	4
3) As the games goes on does physical fatigue play a factor?	4
4) Do back to back games affect the performance of a team?	4
5) What are the tendencies of a good player?	5
6) what factors are most impacting how the team wins?	5
7) How to solve the Conference mismatch Problem?	5
8) Does Being in a bigger Market help in getting better skilled players?	5
9) How the game has changed from 2012-2013 NBA season to 2017-2018 NBA Season?	6
10) METHODOLOGY (ALGORITHM OR ANALYSIS)	6
RESULTS , FINDINGS AND SOLUTIONS TO SOLVE PROBLEMS FOUND	8
1) What factors define the income of a of player?	8
2) Do the officials favour Home team?	9
3) As the games goes on does physical fatigue play a factor?	9
4) Do back to back games affect the performance of a team?	10
5) What are the tendencies of a good player?	11
6) What factors are most impacting how the team wins?	12
7) How to solve the Conference mismatch Problem?	13
8) Does Being in a bigger Market help in getting better skilled players?	13
9) How the game has changed from 2012-2013 NBA season to 2017-2018 NBA Season?	14
Collaboration	15
1) Group Members and Roles	15
2) Problems faced:	15
Reflection.....	16

Introduction

The most widely watched basketball league in the world. The NBA is any basketball fanatic (around 1 billion people watch basketball) dream to play in or cover. What makes this league so popular and great?

Motivation

The main motivation behind working on a sports league was the part to analyse the huge impact that this sport has that so many lives are dependent on it and people are so passionate about it, but what if we told you that this league is not perfect? What would happen to the people who watch this sport so passionately. What mistakes and what are imperfections of this league and how we can improve this. Statistics makes ever thing interesting and in depth and that is why sports and statistics are a perfect combination. Interested people looking at numbers to analyse their teams.

Dataset

The dataset that has been used in this Group project is “Enhanced NBA Statistics 2012-2018”. In this the dataset is described in three point of views.

- 1) the player’s point of view
- 2) the team’s point of view
- 3) Official’s point of view

Also, the salary dataset was used to calculate the salary of each player.

Data manipulation: All the following tables were merged together to form a master table through which all variables were calculated and analysed. Some of the values had to be transformed to make the data more interactive.

Data cleaning: The dataset had a few problems, while merging with the salary table a index had to be evaluated through which each player’s salary could be merged.

In some Cases, the games were duplicated (for eg: GSW game). Through thorough analysis it was referred that the dataset actually wanted to convey home/away fixtures. This was cleaned through introducing home/away variable.

Although, the data is highly accurate but still the data was checked for outliers and missing values and if founded they were treated.

The Dataset links are:

<https://www.kaggle.com/pablote/nba-enhanced-stats>

<https://www.kaggle.com/koki25ando/salary>

Problem Statement, Explanations and Methodology

1) What factors define the income of a of player?

Explanation: The correlation between income of players and all other factors was calculated, we had to merge two different tables for this the salary season 17-18 table and the player box score 17-18 table. But through this a lot of missing values were introduced because the salary for temporary contracts were not given in the dataset but the player statistics for these players were given. In order to make our calculations more accurate we replaced the missing values with the mean salary value. After merging and cleaning our data we calculated the summary of all statistics and plotted graphs for them. Then we identified outliers. For this dataset, We did not choose to treat outliers because the dataset was highly accurate and we wanted to keep the originality of the data. Before pursuing ahead we verified if the data was highly correct and we were satisfied. After that we calculated, correlations of all the variables. As the Dataset is quite extensive we only calculated correlations that are highly correlated (.05). Due to the extensiveness of the nature principal component analysis was used to reduce dimensionality of the variable. All the 9 principal component analysis were observed and it was observed that.

2) Do the officials favour Home team?

Explanation: The correlation between officials(referees) and whether the home team gets a advantage or not was calculated. The reason for this calculation was to identify whether the crowd plays a factor in the decision making of officials. In order to do this calculation we transformed our dataset into two tables one was for the home teams and one for the away teams. In this new table only values which can be affected by the officials(fouls, freethrows, team. turnonvers) were chosen. The observation that we observed were was during the home matches, the home team gets called for less fouls , indicating that indeed there is a difference and crowd tends to play a factor in the games. We used wilcox test between all variables to compare them. Suitable univariate graphs were also plotted along.

3) As the games goes on does physical fatigue play a factor?

Explanation: we wanted to calculate, whether professional players get tired as the length of the games proceeds towards end. For this we calculate and plotted the graphs between the pts scored in first quarter of the game and the points scored in the last two quarters. We found out that there was no difference and we conclude that players don't get physically tired and our used to playing in such high conditions.

4) Do back to back games affect the performance of a team?

Explanation: This problem has been under scrutiny for a long time, do back to back (teams playing 2 games within 2 days) affect how the team performs? The reason for this calculation to predict how unfair the schedule can be to the effective chances of performing well in the league. We observed that there was a difference between the performances i.e points and players were fouling more (could be a reason they are tired) but what we observed was there was no difference between the overall outcome i.e the

outcome wasn't getting affected by this situation. If we had a database that described injuries we would have been able to gather more insight into this.

5) What are the tendencies of a good player?

Explanation: In order to Solve this problem, we had to hypothesise our criteria for choosing a good player. We observed throughout the dataset and manually observing the best players are those who get more minutes per game i.e if you play more minutes it means you are a better player. To solve this problem we made a new table and calculated summary of each states and grouped it by each player. Then we calculated the correlations with minutes. Also, all the univariate graphs were plotted to get a more better idea. We were highly satisfied with the results we got , we got the factors and they seemed pretty good, but we wanted to increase the scope. We wanted to predict based on the factors whether he is a good player or not. In order to do this we made linear regression models and we plotted them. This can be useful in making future predictions and current predictions whether a player qualifies to get good amount of minutes or not

6) what factors are most impacting how the team wins?

Explanation: Similar to the question regarding qualities of a good player we wanted to find qualities of a good team. The difference in this was we had a good factor through which we can make our predictions. It was the result of each team. The teamboxscore table was summarized for each team and transformed then the result percentage of each team was used and correlations were plotted and evaluated. Based on the correlations and univariate graphs we made our models and through the models we can make predictions for other teams. A lot of models were made and a final model was proposed through which we can make predictions

7) How to solve the Conference mismatch Problem?

Explanation: The NBA has mainly two conferences eastern and western conference. The top 8 from each conference are selected and then they proceed to play in the playoffs for a chance to win a championship. It has been widely scrutinised that there is a huge misbalance amongst the conferences i.e one conference has better teams than the others. This was analysed to identify whether this is true or not. The reason for this calculation for this question was to analyse how unfair it is for some teams because the league has been divided into geographical conferences. For this we transformed our teamboxscore dataset on the bases of conferences (two tables one for each conference). After Data transformation we calculated summary statistics and univariate graphs and we ran comparison test for each table.

8) Does Being in a bigger Market help in getting better skilled players?

Explanation: There are big cities in The United states, the result was calculated to analyse do better players play for Big Markets (E.g.: Los Angeles). The reason for this calculation was as the media coverage is high

in these markets, the role of media in a sports league was analysed. Similar to our approach in conference mismatch problem we divided the datasets into two tables (big market and small market). teams based on their wealth and media coverage. Then all the summary graphs were plotted for each table and comparisons were run between two tables. The results were analysed it was observed that this does not make such a high difference (and it was observed New York has the worst players).

9) How the game has changed from 2012-2013 NBA season to 2017-2018 NBA Season?

Explanation: The main purpose for this was to calculate the change in the game of basketball over the years. For this, we transformed our table into two tables one for 2012-2013 season and one for 2017-2018 season. After transformation, all the summary graphs were plotted for each table and comparisons were run between two tables. It was observed the game has changed a lot, it has more scoring and more number of three pointers, also the pace of the league has increased in a huge way (almost 5 more possessions).

10) METHODOLOGY (ALGORITHM OR ANALYSIS)

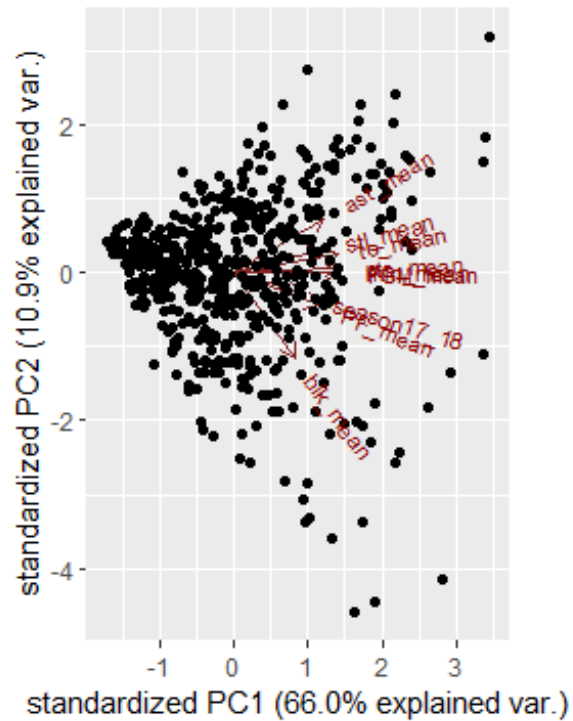
The methodology of this project was to import all the datasets. We could have merged all the tables and make a one master table but that would have made the whole analysis messy as the dataset for each table is quite extensive. So, we decided to transform the data as per each question that we asked. The string variables such as division and conference were converted into categorical variables. Then we transformed the player table and summarized the statistics for each player, this helped us in evaluating the statistics of each player. The result of this was joined with the salary table. This produced a lot of null values so they were replaced with mean values. After this we calculated outliers in the table, The reason for not removing outliers was because the data was highly accurate and we did not want to remove the originality of the data. We tried it with removing data and we found out that the results were not authentic enough to represent. After this the correlations of salary was calculated with each variable. As the dataset is quite extensive, we went with principal component analysis which is a dimension reduction technique to evaluate. Principal component analysis takes variables possibly correlated variables and reduces it into possible uncorrelated variables. After this, we calculated home court advantage and impact of officials in the game. For this we converted the dataset into home and away tables and then we ran comparison tests and plotted these tables graphically. The results were analysed. After this we calculated whether during the game the players get tired or not. Again, this was calculated using comparison tests such as wilcox test to analyse if they are statistically different and graphically the results were plotted. The next question we asked ourselves was how to tackle back to back games problem. We transformed the day by column no of days off between each game and then ran normalcy tests on the tables. We plotted the tables graphically and ran comparison tests to calculate. Then the next question we asked was how do we define what is a good player, for this we calculated correlations with mins played and then we calculated models based on. These model can help us in making future predictions. The same approach was followed for team success the only change being the choosing criteria. We chose teamRslt percentage as the criteria to evaluate and make our models. The next question was the misbalance in conferences, this was calculated by transforming the data into two conferences. Then comparison tests and graphical plots were used to compare the differences between the tables. The same approach was used for next question i.e

impact of big market and smaller market teams and does it help in getting better players. We defined by researching market value and media impact and we categorized 4 cities as big cities, then we calculated the result by again comparing the tables. The next part we asked was, is the game changing or has the game remained the same? To solve this problem, we took the table and split it into 2012-2013 season 2017-2018. Then we calculated results based on various comparison tests we ran and statistically and graphically analysing data. In the last, we calculated the average standings of each team to help us know which teams have had the best record in the past half decade.

RESULTS , FINDINGS AND SOLUTIONS TO SOLVE PROBLEMS FOUND

1) What factors define the income of a of player?

In this question we observed that there is no set criteria which drives a players income. One of the possible reasons for this could be the market in 2016 NBA season was random, This was due to the latest salary changes which allowed more money for the players. If the dataset was for some other duration we could have analysed a better outcome.



2) Do the officials favour Home team?

In this we observed that there was a lot of variance in the home fouls and away team fouls and also there was a difference in the free throw attempts. We concluded that , home crowd indeed plays a factor in officials decision. One of the techniques for solving this problem could be allowing a 4th official which corrects the decision of ground officials by watching the replay immediately.

The difference p values observed were

TeamPF :0.004

TeamFTA:0.005

3) As the games goes on does physical fatigue play a factor?

We found out that there was no difference and we conclude that players don't get physically tired and our used to playing in such high conditions.

The difference p values observed:

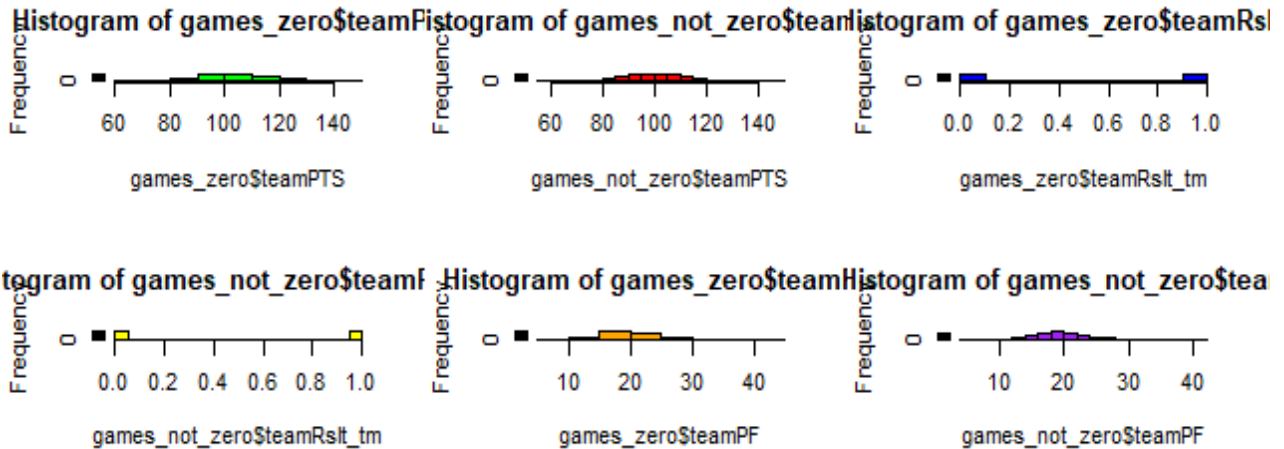
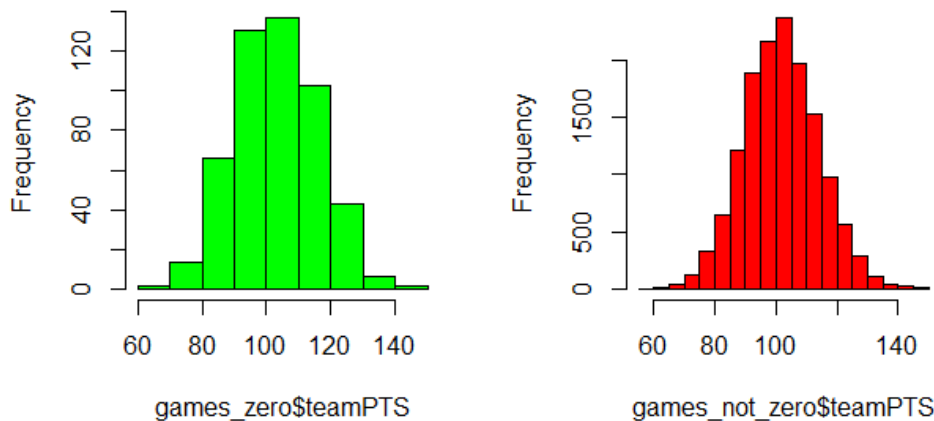
Quarter 1, Quarter 3: 0.2

Quarter 1, Quarter 4 :0.07

4) Do back to back games affect the performance of a team?

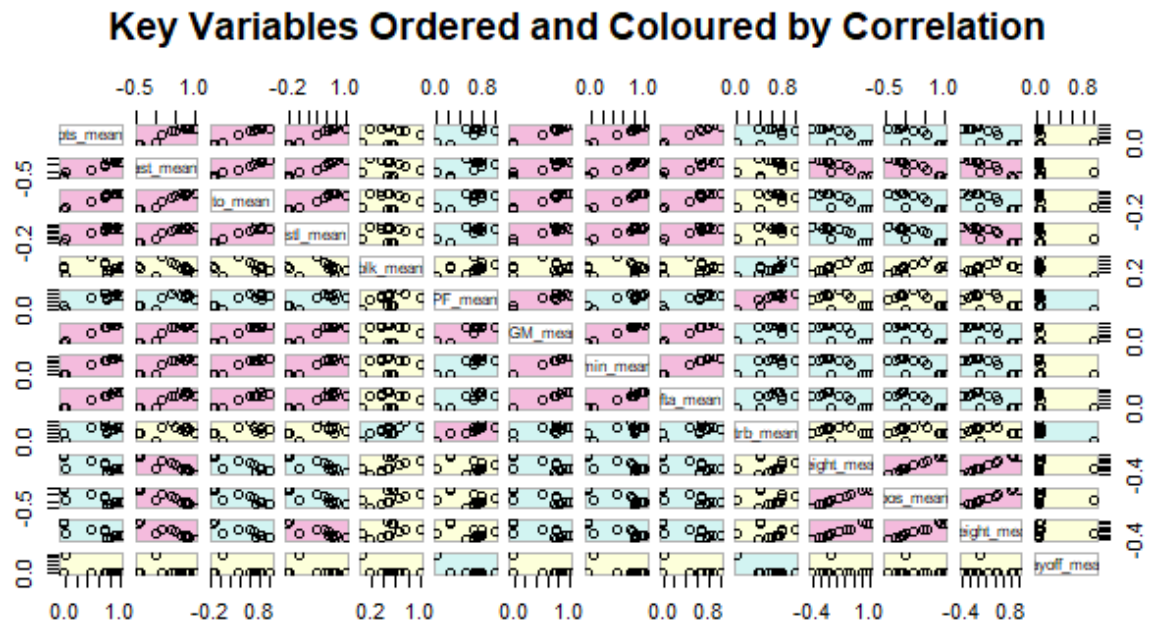
We observed that there was a difference between the performances i.e. points and players were fouling more (could be a reason they are tired) but what we observed was there was no difference between the overall outcome i.e. the outcome wasn't getting affected by this situation. A solution to solving this problem could be to start season early so that the dates are well spread and teams don't have to play back to back games.

Histogram of games_zero\$teamPTS Histogram of games_not_zero\$teamPTS



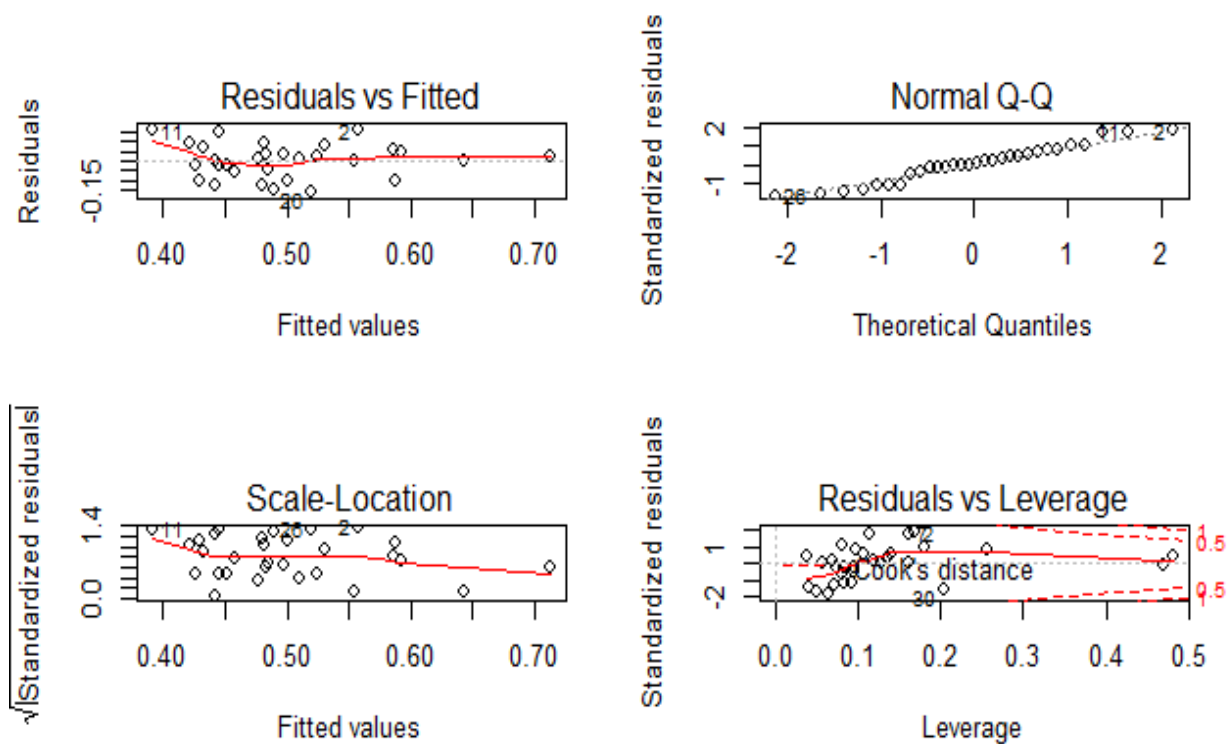
5) What are the tendencies of a good player?

The results we observed were that the factors that drives a persons minutes are, the no of points he scores the no. of assists he make, the number of free throw he attempts. We also observed position and height had no factor in deciding whether a player is good in NBA or not. Hence, breaking the wrong stigma that greater height you have the more successful you are in NBA.



6) What factors are most impacting how the team wins?

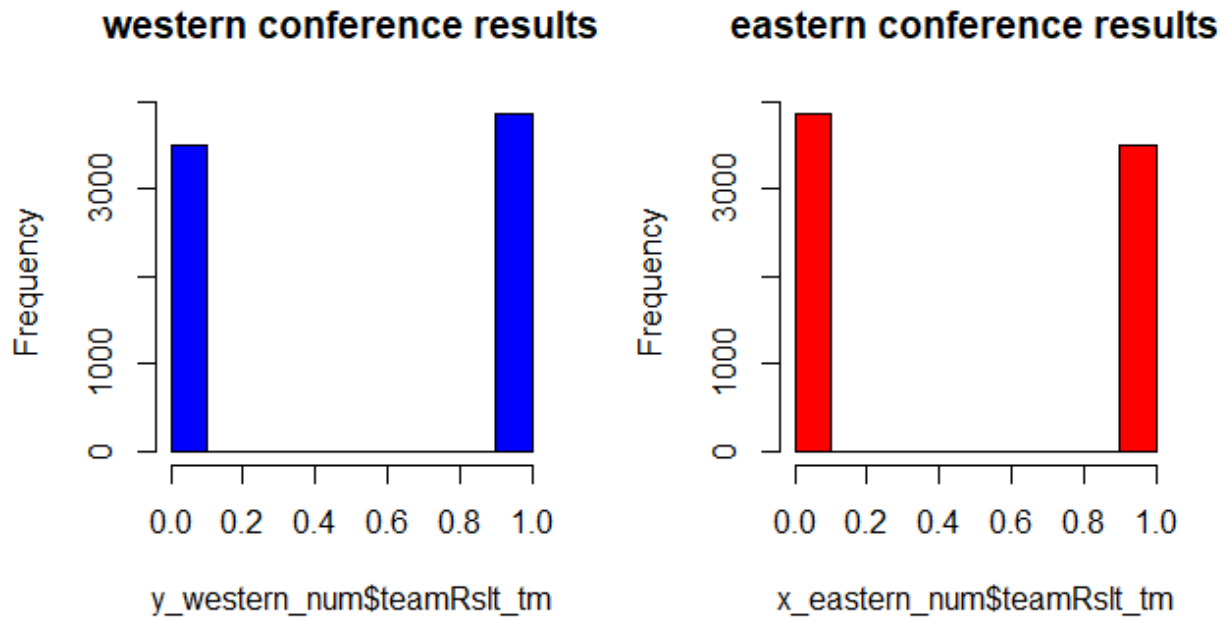
For this question a lot of models were made and we finally chose a model through the various factors such as p value and adjusted r squared value, also we had a look at the coefficients. Based on our observations. We observed that wins were highly correlated with points (obviously) , but two things that struck us was points in the third quarter had higher effect on wins than points at the fourth quarter i.e. if you win the third quarter your chances of winning are high. Also field goals attempted had higher correlation than field goals made. It is often said NBA is a Make or miss league. The statistics show otherwise.



7) How to solve the Conference mismatch Problem?

The results of this question were quite surprising, the best league in the NBA has a huge flaw in it, the league is skewed towards the western conference, i.e. teams in the western conference have better teams than teams in the eastern conference. This makes the teams in western conference a difficult task because only 8 teams from each conference is selected. A theory we proposed for solving this problem is to remove conferences and select top 16 teams from total. The results observed were

P Value: $4.872e-07$ Also, the result can be seen graphically,



8) Does Being in a bigger Market help in getting better skilled players?

The results for this comparison between bigger market teams and smaller market teams was not significant. In fact it was observed that New York had some of the worst squad players. The results for this question were calculated through various comparison tests such as Wilcoxon test. Also, graphical representation helped us in determining the value.

9) How the game has changed from 2012-2013 NBA season to 2017-2018 NBA Season?

To calculate the change in trend transformation of two tables were done. It was observed the pace of the game has increased, teams are scoring a lot more points, and the no of 3PM and attempted have increased meaning the game is moving away from the basket(free flowing). The result was calculated using Wilcox test

The results observed were:

P values for pace : 1.768×10^{-15}

95% confidence interval: -6.2 -4.3 indicating that the difference in mean is negative

P value for points : 4.369×10^{-11}

95% confidence interval: -10.2 -6.2

P value for 3pm: 2.4×10^{-12}

95% confidence interval -4.0 -2.5

Collaboration

1) Group Members and Roles

Group Members:	Role
Shubham Handa	Sports Guy (Researching, Analysing things, writing r scripts)
Gurjyot Singh Anand	statistician (Developing and verifying r scripts)
Divya Chandwani	Statistician (Developing and verifying r scripts)
Hardeep Kaur	Data Engineer (Finding Inconsistencies, verifying results bringing in new ideas)

The 3 ways we are functioning well as a team are:

- a) Everyone knows their role and are trying best to focus on that and perfect that skillset
- b) Everyone is honest to each other about their capabilities.
- c) Everyone is trying to push each other to improve.

2) Problems faced:

Time Issue: Due to lot of assignments and projects the team meetings have not been that long and hence not that productive as we wanted.

Knowledge issue: We have a lot of ideas to bring to the table, but being new to this subject we don't know whether we will be able to produce those ideas into a screen.

To function better we are planning to have long discussions and to remove that knowledge barrier in the earlier stages the sports guy will research the implementation of those complex ideas that we want to implement. He will then interact with the team members so they can proceed on to applying it on the dataset.

Reflection

The team learnt a lot from this assignment, the fact that how a team operates and how a team contributes can help you reduce work load, also learnt the power of data, never realised that simple data can transform the future and predict how the game will play out (referees and back to back question), I think if we had more knowledge about graphical representation of representing data we would have represented data in the form of hot spots i.e. from does each player have a better chance of scoring basket (for e.g.: LeBron James from right side under 24 feet has so and so probability of scoring). Also, some knowledge of implementation of machine learning would be nice as we would like the stats to generate automatically and based on that analyse them automatically. As basketball involves a large sample size in no. of shots, knowledge about clustering algorithms would be nice. For future students, the best advice would be to know what type of data you're getting into and research thoroughly as data might be inconsistent and the factors that you are trying to evaluate data might not be able to do that.