Comparing and Contrasting NBA Eras by Visualising Basketball Statistics Selected using Machine Learning Techniques

by

Noah Isaac Roberts

975804

SUPERVISOR: DR DANIEL ARCHAMBAULT

2020/21

This project dissertation is submitted to Swansea University in fulfilment for of the requirements for the Degree of Bachelor of Computer Science

Department of Computer Science Swansea University

Declaration

This work has not been previously accepted in substance for any degree and is not being con- currently submitted in candidature for any degree.

Signed

Date 26/04/2021

Statement 1

This thesis is the result of my own investigations, except where otherwise stated. Other sources are acknowledged by footnotes giving explicit references. A bibliography is appended.

Signed

Date 26/04/2021

Statement 2

I hereby give my consent for my thesis, if accepted, to be made available for photocopying and inter-library loan, and for the title and summary to be made available to outside organisations.

Signed

Date 26/04/2021

Abstract

Visualisation in sport is a relatively new field especially within basketball. Over the last decade teams have become very interested in analysing data to help become better and more efficient. This paper looks at what statistics have made teams successful during different eras and why they win game.

I have collected seasons worth of data from two very different NBA eras. I have performed machine learning techniques such permutation importance to find what are the most valuable statistics for an NBA team winning a game. I have created visualisations to compare and contrast the difference in NBA era and how teams have been successful and unsuccessful.

By reading this paper you will be able to see the importance of visualisation and why it has been adopted by this new era of sports and specifically NBA basketball.

Table of Contents

Table of Contents

Chapter 1

Introduction

1.1 Motivations

The use of statistics in sport has become extremely popular in recent years due to its vast uses in helping teams get better and prepare for games in much greater detail. Just a decade ago, if the Portland Trailblazers wanted to figure out how best to defend an elite player like Lebron James they would watch old game tape or send a scout to a live game. Using technology to track player movements and statistics to show how effective and productive they are being on the court; teams can use this in preparation for games and make sure they're being as efficient as possible. This leads to which players are picked to start games and how teams will set up to defend a rival team. Many teams who have bought into this have experienced much greater success since the rise of technology and statistical analysis of the game.

A popular example of this would be the book, turned film, Moneyball which glamorized sabermetrics – baseball statistics that measure in-game activity. The motive of the film was to find a "more objective method of analysing player performance and to find the most productive group of players within the team's salary cap (budget)" [0]. This was seen as a risky strategy and nearly led to Billy Beane, the general manager of the Oakland Athletics, being fired after a run of bad results. Sticking with him for the season led to the Athletics clinching the 2002 American League West title.

Statistics also play a large roll in sports betting. If you ask the average person how they decide on what bets to place they would say that they have picked the favourite or a team they support, but by looking into the statistics betters can have much greater success. For example, during my project I found that three-point percentage is one of the most important stats in determining if a team wins, with this information by placing a bet on a team that shoots high percentage threes a better would find much more success. Obviously, this is quite a simplified outlook but gives an idea of the use of basketball statistics in the wider world.

1.2 Objective

During my project I collected team data with statistics covering all aspects of an NBA game and used machine learning techniques to find the most important statistics for a win and also the most important statistics for a loss. I have done this for statistics from the 90's and the 2010's and have compared and contrasted between these different eras. I have created visualisations to prove that that they are in fact the most/least important statistics. I thoroughly enjoyed creating my project, as a fan of basketball for 15 years I found that I was very motivated throughout the year and was able to create useful visualisations that had a purpose in my project.

To conclude, the technical aims of my project were as follows:

- I have collected and trained NBA statistical data with random forest and used permutation importance to find the most important features for a win and loss
- I have created a visualisation to show the correlation between the statistics and wins to further back up the permutation importance result
- I have created visualisations to display the results of the permutation importance (top 5 and bottom 5 features for each NBA era)
- I have identified what makes teams successful, and compared the differences between eras using my visualisations

1.3 Background Information

Basketball

Basketball is a team sport consisting in two teams of five who are trying to score in a basket opposite ends of a rectangular court. The NBA is the National Basketball Association where 30 teams across America (1 in Canada) play against each other trying to win the NBA title. Base statistics were first recorded for the NBA in the 1946-47 season. Certain statistics were not always recorded which is why I used 90s and 2010s data in my work, by this time all advanced statistics were being recorded.

Machine Learning

Machine learning is an application of AI. Algorithms are used to build models using data given by the developer. The models are then used to make decisions without being “explicitly programmed to do so” [1]. During my project I used random forest to train data and permutation importance to find the most important team statistics for deciding a win in the NBA.

Visualisation

Data visualisation is way to graphically represent data. Using visualisations, such as graphs, it provides a way to find trends and patterns in selected data. It also helps to simply complicated data as you can display it in an intuitive way. During my project I have used a number of different visualisations techniques such as scatterplots, heatmaps and more to visualise the selected team statistics.

Chapter 2

Related Work

Here I will discuss previous pieces of work that influenced and related to my own project. I will discuss competing systems and how my project differs. Statistics being used in basketball is still relatively new, but I was able to find a lot of interesting academic papers that related to my own work.

2.1 Moneyball / Leicester F.C

As I talked about in my initial document the film Moneyball is “an American sports drama about a baseball team manager who adopts a risky new system for team selection” [2]. I watched this film multiple times before and during my project and found that it related very much. The assistant manager Billy Beane and the assistant manager Peter Brand used sabermetrics to analyse and select players for their team. In my own project I used a similar concept – permutation importance to analyse NBA teams in-game statistics. My own work differed to Beane’s analytics as I have analysed full team statistics instead of individual player statistics.

Also mentioned in my initial document the 2016 Premier League season where Leicester City defied the odds and won the title. The bookies had said “it was more likely that the Loch Ness monster would be discovered, or Elvis found alive” [3]. In the article, “How stats helped Leicester City F.C win the Premier League”, it was discussed how GPS data was tracked during training and games and presented to the players and coaches on a daily basis. They used this data to show how effective counter-attacking football was and based their training session around the data. This is related to my project as they found the most effective data and applied it to games which helped them score more goals, win more games and go onto win the title. Again, the data recorded was from each individual player but after collecting the results they were able to apply it to the whole team, meanwhile my project was based on whole team’s data.

2.2 Academic Papers

Information Visualisation in the NBA: The Short Chart is a paper written by Stephen Chu at the University of Berkeley [4]. In the paper he discusses how he created a shot chart to display the strengths and weaknesses in a team's offense. He colour-coded the locations on the court and teams with a below average percentage in a certain area would appear as grey and red for above average. My project differs to this as I take all offensive and defensive statistics into consideration which I believe gives me a more accurate representation of the most and least important statistics for a win. My project focused more on the team as a whole instead of individual players as this paper does. This paper also looked at individual games which can be good for learning about match ups with specific teams but not on how a team can improve as a whole, this is why I chose to use multiple seasons worth of data to make sure I was creating something that could be useful for a team.

NBA Game Result Prediction Using Feature Analysis and Machine Learning is a paper written by Thabtha, F, Zhang, L and Abdelhamid, N [5]. The paper talks about finding the influential features that affects NBA games and also forecasting the game outcomes. They use feature analysis to remove the least influential features which helps them predict outcomes more accurately. I have also used feature analysis in my own project, but I have not removed any features, I have simply taken the top 5 and worst 5 to analyse. This piece of work differs to my own as they have used the CFS method and the RIPPER algorithm for feature selection whereas I have used permutation importance using the eli5 library. My work also differs to this as I have made visualisations on the most and least important statistics instead of predicting game outcomes. Interestingly this paper did not have 3-point percentage as one of the most influential stats but did have field goal percentage which includes the 3-point percentage. However, this paper only used data from the 2014 season which is not a very accurate representation of modern NBA basketball. If they had used multiple seasons worth of data like I have done in my project I believe there would be more similarities in our work. This paper took into account the home variable which appeared as an influential attribute in two out of three of feature analysis algorithms. In my own work I did not use this attribute as I wanted to focus on the actual game statistics and the venue is an outside factor. In future work if I was to predict game outcomes, I would include this attribute.

Finding related work specifically for basketball was quite difficult, most pieces of work focused more on the machine learning side and predicting games. I was able to find an article about predictive analysis in the English Premier League that shared similarities to my own work. Predictive analysis and modelling football results using machine learning approach for EPL is a paper written by Rahul Baboota and Harleen Kaur at the Guru Gobind Singh Indraprastha University in New Delhi. Like my own work they used feature engineering and exploratory data analysis, but they used this to determine the most important factors for predicting the results of a football match. The way they found the most important factors was different to my work, in this paper they divided the features into two classes, class A and class B. They tested the feature set with the Gaussian naïve Bayes and gradient boosting models. I have used random forest with permutation importance. During their work they only used two years' worth of data whereas I am using a combined 10 years of historical and modern data. They found a much better performance with the second set of data as the features fit their model better, but this could be seen as a weakness in the analysis.

2.3 Moreyball

Daryl Morey was the former general manager of the Houston Rockets, a team in the NBA. He also co-founded the annual MIT Sloan Sports Analytics Conference, "an annual event that provides a forum for industry professionals and students to discuss the increasing role of analytics in the sports industry" [6]. Morey's basketball philosophy is massively reliant on data analytics, an example of this is that he favours three-point shots and lay-ups over mid-range shots as the percentage wise they are better shots to take. His style of play has been called "Moreyball" which is a nod towards the book Moneyball which I mentioned previously. As a former consultant and MIT graduate who had not played professional or college basketball "his previous work in statistical consultant led him to find a deeper understanding of the sport and how teams were operating inefficiently, costing them wins" [7].

As mentioned previously his main change to the Houston Rocket was to take an increased number of three-point shots. These shots are more difficult as they are further away from the basket, naturally. Morey recognised that the "50% uplift in points received for the three-point shot (compared to the two-point shot) made it more mathematically efficient than almost all two-point shots other than dunks and lay-ups" [8]. By doing this the team vastly increased their wins in a season becoming one of the tops teams in the NBA during the last 5 years. Another realisation was that "corner threes" had a higher percentage of going in, this is because of the shape of the three-point line which makes corner threes closer to the basket. This led to many set plays being designed to put good three-point shooting players on the corner meaning the shots are more likely to go in statistically. All these changes led to the Rockets breaking the record for most three-point shots made in a season.

During my analysis of the modern NBA era, I found that the 3-point percentage was the most important statistic in deciding if a team wins a game. I had assumed before my project that this would be one of the most important statistics after reading about "Moreyball" and it was very interesting to see this come true in the visualisations I created. My project differed to this as I focused on the top 5 statistics for a win instead of just focusing on the three-point statistics. I focused on the top 5 as I was comparing different eras. The three-point statistics were not as important during the 90's so during my project I didn't solely focus on independent statistics.

Chapter 3

Design

This section of the document discusses the design process of my project. I will discuss the structure of the software, alternatives that I considered and also what visualisations I selected and why.

3.1 Structure of Software

I will be diving a lot deeper into the software used, what I have created and the specifics during the implementation section of this paper, here I will be describing how I split my problem into parts, what was done in each part and how they fit together to meet my aims. I will also talk about the visualisations I selected and why.

The first task was to collect the data. Finding the data took some time at the start of my project as the official NBA statistics website does not allow you to use them for academic or personal use. This led me to look elsewhere. Basketball reference is a website that holds all statistics for a team, player or season that can be used for personal or academic use, this is where I collected all the data used for this project. I will go into further detail about how and the format it was collected in implementation. Next, I selected the program I would create the visualisations in which was Altair. I selected this because the documentation was very thorough and there were many YouTube tutorials on how to create different visualisations. I also used Altair during a previous university module, so I had some experience with it. I then used a library for permutation importance so that I could find the most important statistics for a team win (this was done for each era), this concluded my first aim.

My next aim was to create visualisations to show the correlation between chosen statistics and wins which would back up the permutation importance result. To do this I created a heatmap with labels which displayed the how much a win and statistic correlated. The label showed a decimal for how much they correlated, and it was also colour coded. The third aim was to create visualisations to graphically display the permutation importance results. I chose the top 5 and bottom 5 statistics from each era and created visualisations for each. By doing this I could see the trends in the data and visualise how they equated to more wins.

Lastly, after creating all the visualisations I was able to compare and contrast the different eras, this will be expanded on in the testing section of this paper. By following these steps, I was able to complete all my aims. I then found basketball reference which is a website containing all basketball statistics from when they were first recorded to present. I was able to download a season's worth of data for each team as csv files, which I could then change to my liking.

3.2 Alternatives

During this section I have discussed alternatives and also the chosen software – which will be expanded on during the implementation stage of this paper.

As previously mentioned, I first went to the official NBA website for statistics. Unfortunately, they did not allow the personal or academic use of their statistics, so I had to look elsewhere. I found an NBA API which would scrape the statistics from www.nba.com. The problem with this API was that it had not been updated for a few years and with the NBA website having changed it no longer worked. Basketball reference was free to use and allowed me to simply download seasons worth of data as a csv, this is why I chose to use it.

After collecting the data, I had to use permutation importance on it to find the most important statistics. I had a few different options in how I could do this, first I looked at Kaggle which used Python libraries NumPy, Pandas and Sklearn to train a model and show the importance's. I played around with Kaggle at the start of my project but didn't find the documentation very thorough or helpful. I looked for an alternative library and found ELI5, it allows visualisation of models using a unified API. The documentation for this library was very thorough which led me to use this library for my project.

When picking a software to use for visualisation I wanted something that could create complex graphics. I first came across matplotlib which is an open-source Python library. The library was "initially written by John D. Hunter, who was a neurobiologist. He authored Matplotlib at the time of his post-doctoral research in Neurobiology" [9]. The software was easy to use and learn but the problem I found with it was that the visualisations looked quite basic. The interactive features were also quite limited compared to alternatives. While speaking to my academic mentor he suggested taking a look at Altair which is a declarative statistical visualisation library for Python. The library is based off of Vega-Lite which I had used previously. After reading the Altair documentation and looking at the example gallery I decided to use it. The visualisations looked very sleek and modern and there were a lot of interactive features I was able to take advantage of.

3.3 Visualisations Selected

This will be a short section on why I have chosen to create the visualisations that I have in my project. (Some figures for this section will be shown in the appendix).

Scatter

A scatter diagram is a “mathematical diagram using Cartesian coordinates to display values for two variables of a dataset” [10]. As shown in figure 1 I have displayed the 3-point percentage and the wins with the teams as the label. I have also used tool tips to display the team’s name, wins, 3-point percentage, number of 3’s and the year the data is from, I have used tooltips to help the viewer of the visualisation understand what each plot represents. I have used scatter plots for each of the statistics chosen by permutation importance (for each era). I found scatter plots to be a good way to display many plots on a single graph. Figure 1 shows five seasons worth of data and by using the scaling feature on Altair I was able to display this data clearly. By using a scatter plot the range of data flow is shown clearly, the minimum and maximum values can be found.

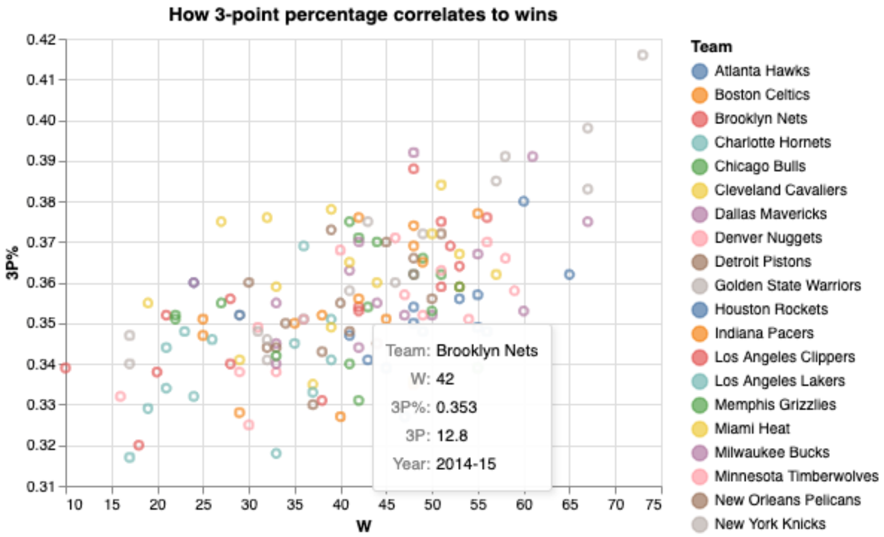


Figure 1: Simple scatter showing correlation between 3-point percentage and wins

Bar and Stacked Bar Charts

A bar chart displays categorical data using rectangular bars with heights proportional to the values that they represent. I chose to use bar graphs as an addition to the scatter plots, by interacting with plots of the scatter plots I created the bar graph would update only showing the selected plots. I did this as I wanted to see the number of three pointers alongside the percentages. This is shown in figure 2 below.

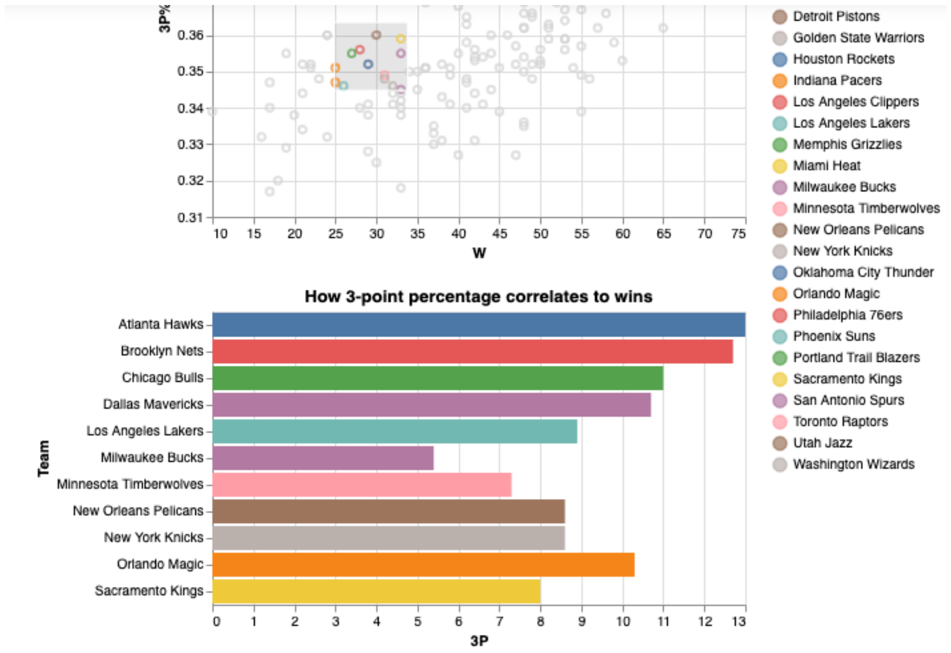


Figure 2: Showing the interactivity between the scatter and bar graph

A stacked bar chart "extends a standard bar chart from looking at numeric values across one categorical variable to two" [11]. As mentioned previously, 5 seasons worth of data was used for each time era. By using a stacked bar chart I showed the percentages of statistics for each season. This is shown in figure 3. I was able to easily compare the differences between teams and the statistics I had chosen for each season. For example, figure 3 shows each team's 3-point percentage for each of the 5 seasons, each bar is also colour coded to be darker the more wins the team has.

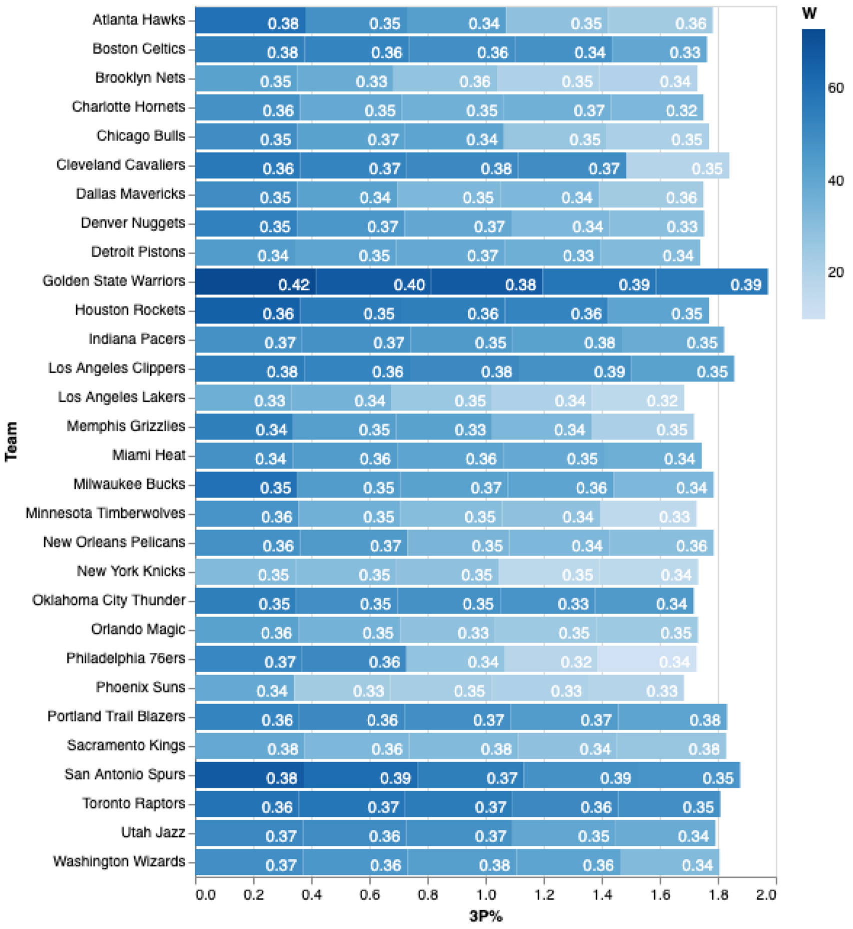


Figure 3: Stacked bar chart showing each team's 3-point percentage over 5 seasons

Heatmaps / Correlation Matrix

I used heatmaps for multiple things during this project. Firstly, I used a heatmap to show correlation between wins and the chosen statistics for the older data. I did this as it's a very nice way to view the data and shows the correlation very neatly and can easily be explained to someone, example of this is shown in the appendix. I also used a heatmap to prove that my permutation importance results were accurate, by putting the statistics against each other I had a visual representation of the correlation between each statistic and also a label given the numerical value of correlation. Figure 5 in the appendix shows my heatmap.

Line of Best Fit

Line of best fit is a line through a scatter plot of data that represents the relationship between the plots. This was a very important visualisation for me I used it for each statistic that I focused on from the permutation importance results. It was important as I could instantly find out if the data showed correlation and could see if it was a sensible line by looking at the plots. Figure 6 in the appendix shows an example of this.

Scatter Matrix

Scatter matrix is a grid of scatter plots that is used to "visualise bivariate relationships between combinations of variables" [12]. I created these as I was able to visualise and explore many relationships in one chart. This was very important for me during my project for when I was looking at comparisons between eras and patterns in the data. For example, I was surprised to see the correlation between the number of three pointers and the three-point percentage was not as strong as I first thought it would be. This will be expanded on during the testing section of my paper. An example of one scatter matrix I created is shown in the appendix (figure 7).

Chapter 4

Implementation

This section will be broken into three main parts: data collection, permutation importance and visualisations. I will be going into great detail on how I implemented my project, I will discuss what libraries I have used, the process I went through for each part and technical detail of all my aims.

4.1 Data Collection and Filtering

Basketball Reference is a website that was created by Sean Forman in August 2004. The website contains data from the inaugural 1946-47 season to current times. The website states that most of the data was donated by Sean Lahman but had modifications made to it. The website also holds data on each player and team (current and former).

I decided to use the “Team Stats for Season” from 1993-1998 and 2014-2019. The columns that this data featured were as follows: Rank, Team, Game, Minutes Played, Field Goal, Field Goal Attempts, Field Goal Percentage, 3 Pointer, 3-Point Attempt, 3-Point Percentage, 2 Pointer, 2-Point Attempt, 2-Point Percentage, Free Throw, Free Throw Attempt, Free Throw Percentage, Offensive Rebound, Defensive Rebound, Total Rebounds, Assists, Steals, Blocks, Turnovers, Personal Fouls and Points. I used these two different time periods as I wanted to compare the differences between eras especially with the rise of the 3-point shot in the last decade. The 90’s was known for “triangle offence” which became known when the Chicago Bulls used it under Phil Jackson and previous assistant coach Tex Winter. This offence is very different to today’s game so I thought that they would be good eras to compare.

Using the “Share & Export” button above the data I was able to view it as a csv and then copy and paste it into Microsoft Excel – where it would all be placed in one cell. By highlighting the cell and selecting text to columns I could split the data. I selected the “Delimited” option and then clicked next where I would select the comma option as this was the delimiter for the data, finally I clicked finished and would have the data correctly separated in their own columns. I followed this process 10 times as I could only copy a season’s worth of data at a time. I then combined the 5 seasons of 90’s data to one Excel sheet and did the same for the 5 seasons of modern data. The problem I found with the data was that it did not include the number of games, wins and or losses, my whole project revolved around the wins, so I had to do something about this. I found the extra columns I needed in a different section of basketball reference, so I modified my Excel sheets and added the columns that I needed. My sheet now included all the columns previously mentioned with the addition of the games, wins and losses. Finally, I saved my Excels files as CSVs ready to be used with pandas. I also deleted some unnecessary columns using the panda's library which will be expanded on during the permutation importance section.

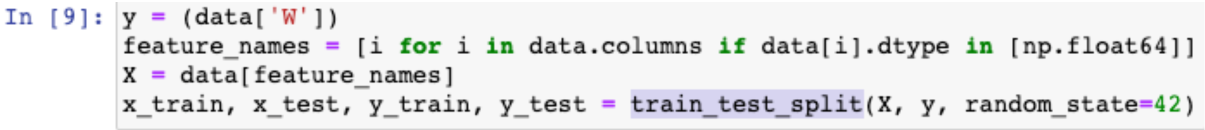
4.2 Permutation Importance

Permutation importance is defined to be the "decrease in a model score when a single feature value is randomly shuffled. This procedure breaks the relationship between the feature and the target, thus the drop in the model score is indictive of how much the model depends on the feature" [13].

I used Python to code permutation importance and my visualisations, and I used Jupyter Notebook as the environment. Jupyter Notebook is a "free, open-source, interactive web tool known as a computational notebook, which researchers can use to combine software code, computational output, explanatory text and multimedia resources in a single document" [14]. It was very useful as I could write some code in a cell and instantly see the output below. The web tool also uses auto save so my work was constantly being saved without worry of losing anything. I enjoyed using the tool as I could split code up between different cells and keep my work organised.

After collecting the data, I decided to find out which were the most important statistics towards a team winning a game. To do this I used permutation importance but first I had to clean up the data. I used pandas which is a library written for Python. Pandas is used for data manipulation and analysis and is particularly used for manipulating numerical tables which is what I needed to use it for. The columns I decided to delete were team (the name of each team) and game (the number of games played that season). First, I read in the CSV, I named the CSV data in Jupyter Notebook and then viewed first few lines of the file to see if it was correct with the head command. To delete the columns, I used the drop command. The axis part of the code was written because the column names were on the first line of the CSV. This was all done in a duplicate CSV called "perm" as I wanted to keep it separate to the original data that would be used for the visualisations.

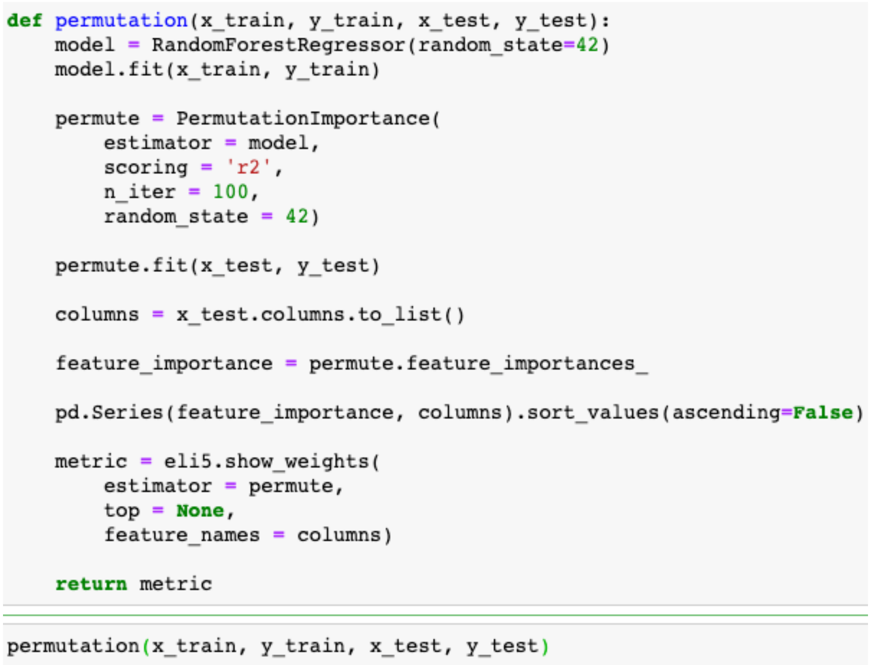
The next step was to load the data ready to be trained. Standard machine learning practise was followed so X was the features (data) and y was the feature we were taking out and seeing which features were most important for it. To extract the features for X all features that were of type "float64" were chosen and assigned to a variable called "X". To extract the "y" value I simply picked 'W' from data and assigned it to a variable called "y". Next, the "train_test_split" method was used to split the arrays (X and y) into random train and test subsets. A random state of 42 was chosen, Scikit-learn uses random permutations to generate splits. The random state that is provided is used as a seed for the random number generator – this ensures that the random numbers are generated in the same order [15]. This can be seen in the following figure.



```
In [9]: y = (data['W'])
        feature_names = [i for i in data.columns if data[i].dtype in [np.float64]]
        X = data[feature_names]
        x_train, x_test, y_train, y_test = train_test_split(X, y, random_state=42)
```

Figure 10: Loading data and setting test and train values

Next, a method for performing the permutation was created. To do this a model was made using a RandomForestRegressor with a random state of 42. A random forest is a "meta estimator that fits a number of classifying decision trees on various sub-samples of my dataset and uses averaging to improve the predictive accuracy and control over-fitting" [16]. Then the linear model was fit with the parameters X and y. A permute variable was then made which would perform the permutation importance. The estimator was set as the model, the scoring used was "r2", the iterator was 100 and the random state was 42. An estimator is an object that manages the estimation and decoding of a model – which is why it has been set to my model. R2 scoring gives a statistical description of how the samples fit along the linear model. I used this type of scoring as I found it was the most sensible scoring as everything will vary between 0 and 100%. A score of 100% would indicate that two variables are perfectly correlated, a low score would suggest a low level of correlation. The n_iter is defined in the Sklearn documentation as being the number of passes over the training data. By default, the iterator is set to 0 but I changed this to 100, this led to longer loading times, but I thought that more passes would be give a more accurate representation. The fit method was used again this time with the new permute variable with inputs x_test and y_test and a column variable was created for putting the features in a list. A new variable was made to store the actual feature importance value for each feature and the pandas' "Series" method was used to store the names and scores in a one-dimensional column – which was sorted by highest importance to least. Finally, the eli5 "show weights" feature was used to show how much weighting each score held – the higher the weight, the more important said feature was. Figure 11 shows the method that was made. I used this method on the 90s data and modern data so that I would be able to compare the differences. The top 5 and bottom 5 features were taken note of to be compared which will be talked about in the results section. This concluded the first aim of collecting and training NBA statistical data with random forest and using permutation importance to find the most important features for a win and loss.



```
def permutation(x_train, y_train, x_test, y_test):
    model = RandomForestRegressor(random_state=42)
    model.fit(x_train, y_train)

    permute = PermutationImportance(
        estimator = model,
        scoring = 'r2',
        n_iter = 100,
        random_state = 42)

    permute.fit(x_test, y_test)

    columns = x_test.columns.to_list()

    feature_importance = permute.feature_importances_

    pd.Series(feature_importance, columns).sort_values(ascending=False)

    metric = eli5.show_weights(
        estimator = permute,
        top = None,
        feature_names = columns)

    return metric
```

```
permutation(x_train, y_train, x_test, y_test)
```

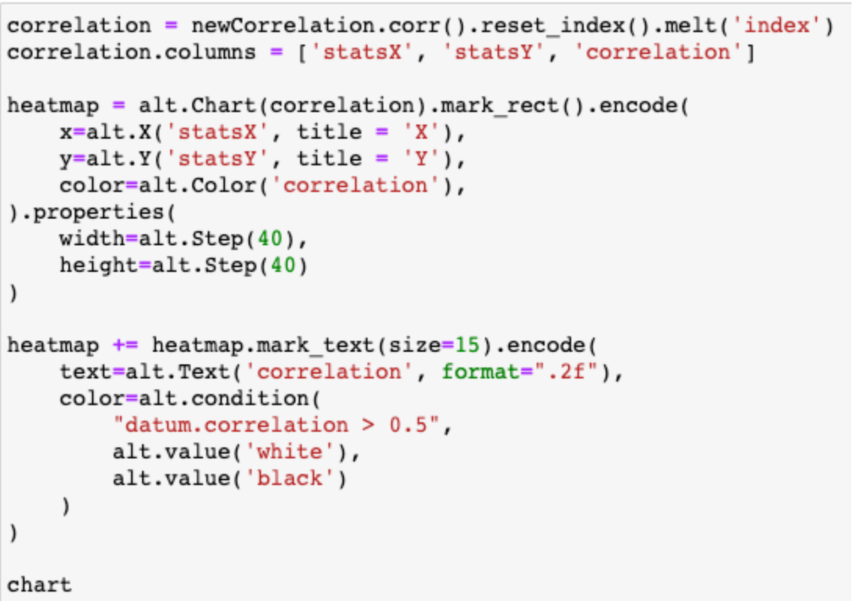
Figure 11: Permutation importance method

4.3 Visualisations

The process of creating visualisations was repeated multiple times for each statistic from the different sets of data, during this section the process for each unique visualisation will be explained e.g., correlation matrix, scatter graph, line of best fit. The most important statistic will be show in the figures for this section and the extra visualisations will be discussed in the results.

Correlation Matrix

The first visualisation made was a correlation matrix to confirm and further back up the permutation importance results. To create a matrix first the data was read and assigned to a variable. To find the correlation the “corr” method was used which finds the pairwise correlation of all columns in the data frame and the columns were split up. Next to create the matrix the named columns were assigned to the x and y values and the colour was set as the correlation. Finally, text was put on each square to show the numerical value as well as highlight all values with a correlation greater than 0.5 in white. This process was done twice for the old data and modern data and the process is shown in figure 12. This concluded the second aim of creating a visualisation to help back up and prove that the permutation importance results were accurate. Using the matrix, I was able to confirm if I was focusing on the correct features.



```
correlation = newCorrelation.corr().reset_index().melt('index')
correlation.columns = ['statsX', 'statsY', 'correlation']

heatmap = alt.Chart(correlation).mark_rect().encode(
    x=alt.X('statsX', title = 'X'),
    y=alt.Y('statsY', title = 'Y'),
    color=alt.Color('correlation'),
).properties(
    width=alt.Step(40),
    height=alt.Step(40)
)

heatmap += heatmap.mark_text(size=15).encode(
    text=alt.Text('correlation', format=".2f"),
    color=alt.condition(
        "datum.correlation > 0.5",
        alt.value('white'),
        alt.value('black')
    )
)

chart
```

Figure 12: Code to create and label correlation matrix

Interactive Scatter with Histogram

A scaled scatter graph was created to show how a team's 3-point percentage impacts their wins for the season. First, a brush variable was made using the selection_interval method. The method allows the user to interact with the graph by highlighting specific plots of the scatter. Next, the scatter was made by using Altair's "Chart" feature, the input was the dataset, and the graph was given a title. The teams' wins were set as the X axis and the 3-point percentage was set as the Y axis, and the colour was set as a team. The add_selection method was used, and the brush variable was called to give the graph interactivity. Next, the histogram was created, the X axis was set as the number of 3 pointers taken that season and the Y axis was the team. The brush was also called using the transform_filter method which allowed interactivity between the 2 graphs. Tooltips had also been used if a user was just going to look at the scatter plot, they would be able to click on a plot and see the team's name, number of wins, 3-point percentage, number of 3's and the year. By selecting a group of plots on the scatter graph I was able to see which teams plots I was looking at and the number of three pointers on the histogram simultaneously. This was helpful to spot patterns as I could see the type of numbers that the successful teams were putting up. Code for this is shown on figure 13 in the appendix.

Stacked Bar Chart

Creating the stacked bar chart was similar to the histogram. The chart tool was used to input the data set and mark bar was chosen as the graph chart. The encoded values were 3-point percentage as the X axis and the team as the Y axis. The bars were colour coded by the wins, by doing this we are able to see the multiple seasons of data on one graph with darker blue representing more wins and lighter blue showing that a team had less wins. These types of visualisations were very useful for viewing lots of data on one graph and we can spot the trends within the data set. The code for this is shown in figure 14 in the appendix.

Line of Best Fit

The line of best fit was an addition to the scatter graphs. It was used to express the relationship between the two selected features. It was a very useful in finding out if there was a trend or if the data was not at all related. To add the line of best fit to the scatter plot the transform regression tool was used with the two features as the input and to display the line on the graph the mark line method was used, and the colour was chosen as red to stand out from the plots.

Scatter Matrix

Scatter matrix were used as I was able to visualise multiple scatter graphs and show every relation between two features in one visualisation. Altair's repeat function was used to tie a channel to the rows and columns within a repeated graph. The graphs were also scaled to the lowest plot, this helped organise the plots and identify patterns easier.

This concluded my third aim as I had created visualisations of the top 5 and bottom 5 features that were selected from the results of the permutation importance. The next step was to analyse the patterns I found in the visualisations and to compare and contrast the differences between the two basketball eras.

Chapter 5

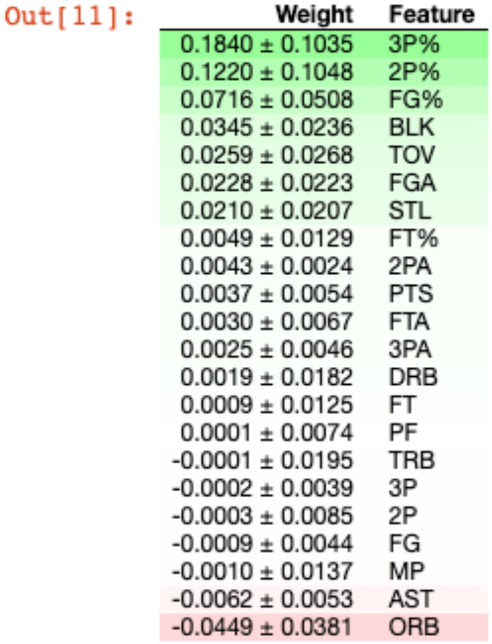
Testing and Results

5.1 Permutation Importance Results and Testing

2014-2019 NBA Data:

Permutation importance was first performed on the modern NBA data (2014-2019) to find the most important features that contribute to a team winning a game. I had previously assumed that the number of 3 pointers scored would be one of the most important features but as shown in figure 15 it was not very important to a team winning a game. The most important feature is the 3-point percentage which made sense as I have previously discussed that more teams are taking on the approach of settling for the benefit of shooting valuable 3 pointers instead of going to the basket and getting 2 points. The 3-point percentage has a larger score than the second most important feature by 0.062 so it is clearly the most important for this era. The top 5 features are the 3-point percentages, the 2-point percentage, the field goal percentage (which is the 3-point and 2-point percentage combined), the number of blocks per game and the number of turnovers per game. It made sense that the 2-point percentage was the second most important results due to the fact that although teams are taking less 2-point shots it is still important that they are making the 2-point shots that they do make. A team just relying on only scoring 3-pointer for a whole game will not be very successful, the 2-point shot is still an important shot for simple baskets and a team gaining momentum. The field goal percentage is then the third most important stat, but this is the 3-point percentage combined with the 2-point percentage, so this also made sense. Defence is a very important aspect of the game and blocks being a top 5 feature may not surprise some. By getting a block the opposing team will be a player down on defence so the offensive team takes control and can easily score by taking advantage of the one less defender. A big block will also swing the momentum of the game, the opposing team will be low on confidence and have to get back on defence. The most surprising feature in the top 5 was the turnovers per game. The turnover is a negative aspect of the game, it is when the team lose the ball to the opposition e.g., a player loses a dribble and gets the ball stolen. The reason this has appeared so high in the modern data is due to the fact that creative, attacking teams are more likely to lose the ball. Flashy players attempting skills or teams that know they are better than the opposition get overconfident and turn over the ball. So, although it's seen as a negative thing for a player to do it's usually the better teams having more turnovers which explains why it is so high.

The 5 least important features for an NBA team winning a game are as follows: the number of offensive rebounds, the numbers of assists, the number of field goals, the number of 2 pointers and 3 pointers. The MP – minutes played has been ignored as this is the same for all teams. The offensive rebounds being the least important was very surprising as it is an attacking statistic, however more offensive rebounds suggests that the team is missing more shots but gets the ball back to attempt another shot, this is why it may not have much impact on the overall game. Using figure 15 we showed that defensive rebounds are a lot higher in importance as the ball is being taken from the attacking team – which will have more of an impact on the game. The next least important was the number of assists in a game, this is an interesting feature to be not very important as you would think that more assists equate to more points which equates to more wins. However, points being scored doesn't always mean there's been an assist. If a player takes multiple dribbles on their own after being passed the ball this does not count as an assist. A team like the Golden State Warriors would have lower assists than a team worse than them due to the fact that Stephen Curry creates a lot of the teams points on his own. The next 3 features tie into each other, the number of field goals, 2-pointers and 3-pointers scored per game. You would think that these are important features but if a team is scoring a lot of 3 on a terrible percentage, they would not be impactful points. A team may score 10 3-pointers on 50 attempts whereas a team could get the same from 15 2-pointers – an easier shot.


Out[11]:

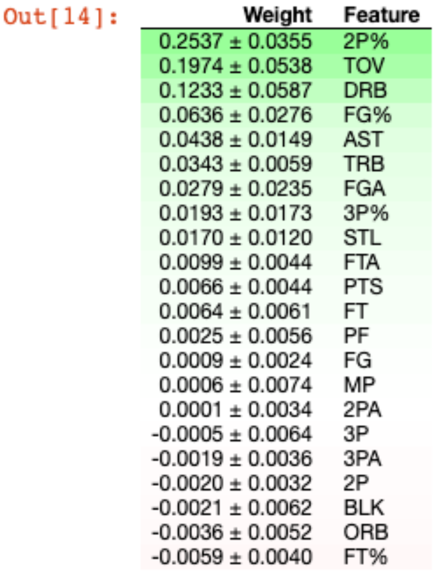
Weight	Feature
0.1840 ± 0.1035	3P%
0.1220 ± 0.1048	2P%
0.0716 ± 0.0508	FG%
0.0345 ± 0.0236	BLK
0.0259 ± 0.0268	TOV
0.0228 ± 0.0223	FGA
0.0210 ± 0.0207	STL
0.0049 ± 0.0129	FT%
0.0043 ± 0.0024	2PA
0.0037 ± 0.0054	PTS
0.0030 ± 0.0067	FTA
0.0025 ± 0.0046	3PA
0.0019 ± 0.0182	DRB
0.0009 ± 0.0125	FT
0.0001 ± 0.0074	PF
-0.0001 ± 0.0195	TRB
-0.0002 ± 0.0039	3P
-0.0003 ± 0.0085	2P
-0.0009 ± 0.0044	FG
-0.0010 ± 0.0137	MP
-0.0062 ± 0.0053	AST
-0.0449 ± 0.0381	ORB

Figure 15: Permutation Importance Results on Modern NBA Data

1993-98 NBA Data:

The top 5 most important stats for a win during the 90s era are as follows: 2-point percentage, number of turnovers per game, number of defensive rebounds per game, the field goal percentage and the numbers of assists per game. As mentioned in previous section the 90's era was known for triangle offense. This is where the offense tries to fill 5 set points, in doing so space is created between players and allows each player to pass to four teammates [17]. This offense explains why assists is such an important feature during this era as the ball is constantly being passed between teammates. The 2-point percentage is by far the most important with a score of 0.2537 which is 0.0563 higher than the second most important feature. During this era 3-pointers were shot a lot less as most players were told to "play their role", in today's basketball a 7ft centre can shoot threes but this was very uncommon during the 90s. Players relied on "safer" shots such a lay ups and midranges. Surprisingly the turnovers per game is also an important stat for this era, this was due to the top teams being freer with the ball, they would try riskier passes and skilful dribbles which meant they gave away the ball more. A notable difference to the modern game is that the defensive rebounds per game is the third most important features. The 90s was known as being tough and very defensive minded era. Games during this time were much lower in scoring due to the defence and type of shots that were being taken.

The five least important features for a win during the 90s were as follows: free-throw percentage, offensive rebounds, blocks, 2-points and 3-point attempts. Free throws were less common for this time as players could get away with harder fouls. Offensive rebounds were also less common as the defensive players e.g., the centres did not play as forward as they do now. Surprisingly blocks were not as important which we will look at in the visualisations to see if this is an anomaly. The two points being scored did not determine a win, as previously mentioned the games were a lot lower scoring, so if a team had averaged a higher number of points per game than the Chicago Bulls it does not mean they are better. Finally, the 3-point attempts were not seen as an important shot and were not as commonly taken which is why they did not have much impact on a team's wins.



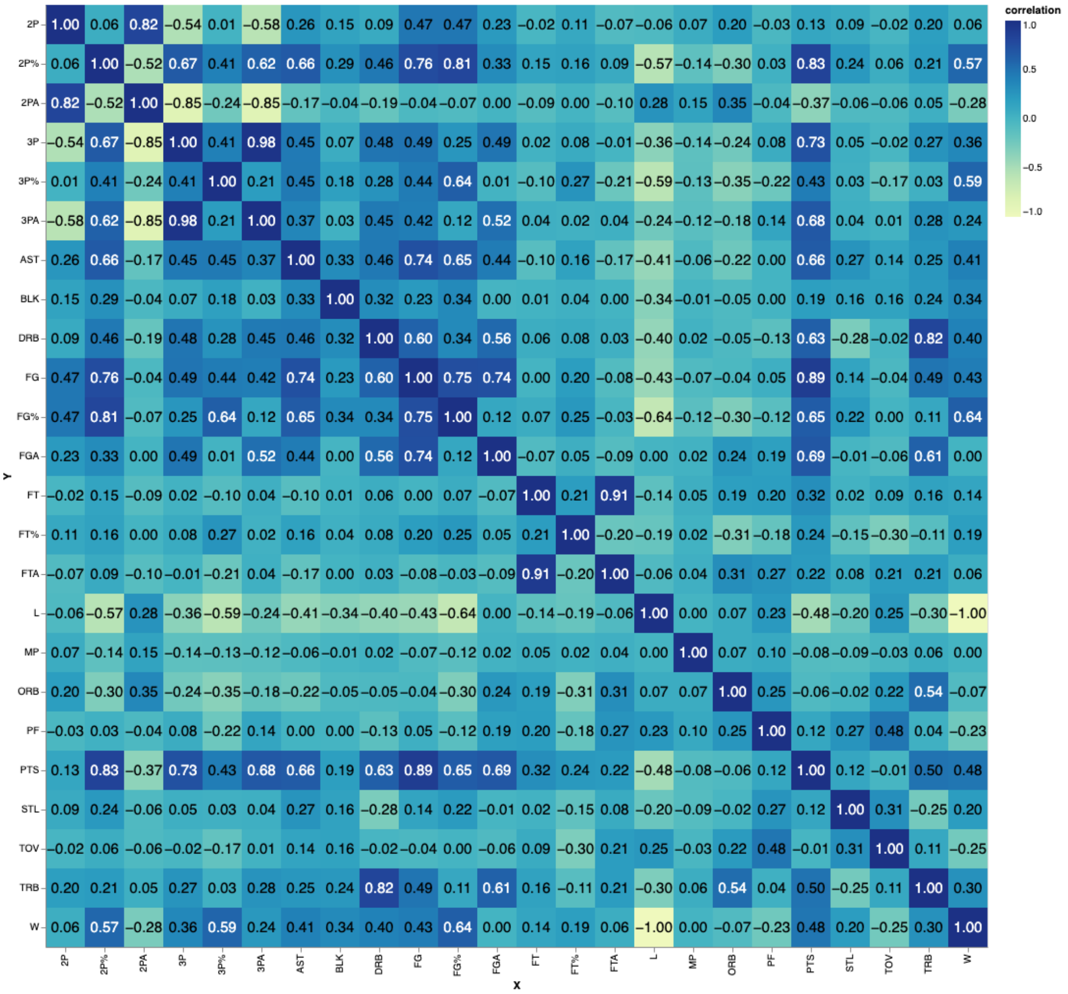
Out[14]:

Weight	Feature
0.2537 ± 0.0355	2P%
0.1974 ± 0.0538	TOV
0.1233 ± 0.0587	DRB
0.0636 ± 0.0276	FG%
0.0438 ± 0.0149	AST
0.0343 ± 0.0059	TRB
0.0279 ± 0.0235	FGA
0.0193 ± 0.0173	3P%
0.0170 ± 0.0120	STL
0.0099 ± 0.0044	FTA
0.0066 ± 0.0044	PTS
0.0064 ± 0.0061	FT
0.0025 ± 0.0056	PF
0.0009 ± 0.0024	FG
0.0006 ± 0.0074	MP
0.0001 ± 0.0034	2PA
-0.0005 ± 0.0064	3P
-0.0019 ± 0.0036	3PA
-0.0020 ± 0.0032	2P
-0.0021 ± 0.0062	BLK
-0.0036 ± 0.0052	ORB
-0.0059 ± 0.0040	FT%

Figure 16: Permutation Importance Results on 90s NBA Data

2014-2019 Data Testing:

As discussed in implementation a correlation matrix was created to see how accurate the permutation importance results were. We can see the matrix created in figure 17 below. By looking at the figure we showed the correlation between all of the features but more importantly the correlation between wins and our chosen statistics. Starting with the modern data's most important feature – the 3-point percentage. Using the figure, we showed that the correlation between the wins and 3-point percentage is 0.59, which is on the higher scale as shown by the correlation key. The correlation between wins and 2-point percentage is 0.57, we showed that the results are accurate as they followed permutation importance results. The field goal percentage and wins correlation is 0.64 which is higher than the previous two. This was not seen as an issue as the field goal percentage is the 2-point percentage and 3-point percentage combined, so the results are still indicating that the permutation importance is accurate. We showed that the blocks correlation with wins (0.34) is lower than features that were not selected as top features through permutation importance such as assists with 0.41. We can also see that the turnover correlation with wins is a surprising -0.25 which suggests no correlation contradicting the permutation importance result. In future work I would like to collect more data to identify if these are anomalies or follow a pattern.



	2P	2P%	2PA	3P	3P%	3PA	AST	BLK	DRB	FG	FG%	FGA	FT	FT%	FTA	L	MP	ORB	PF	PTS	STL	TOV	TRB	W
2P	1.00	0.06	0.82	−0.54	0.01	−0.58	0.26	0.15	0.09	0.47	0.47	0.23	−0.02	0.11	−0.07	−0.06	0.07	0.20	−0.03	0.13	0.09	−0.02	0.20	0.06
2P%	0.06	1.00	−0.52	0.67	0.41	0.62	0.66	0.29	0.46	0.76	0.81	0.33	0.15	0.16	0.09	−0.57	−0.14	−0.30	0.03	0.83	0.24	0.06	0.21	0.57
2PA	0.82	−0.52	1.00	−0.85	−0.24	−0.85	−0.17	−0.04	−0.19	−0.04	−0.07	0.00	−0.09	0.00	−0.10	0.28	0.15	0.35	−0.04	−0.37	−0.06	−0.06	0.05	−0.28
3P	−0.54	0.67	−0.85	1.00	0.41	0.98	0.45	0.07	0.48	0.49	0.25	0.49	0.02	0.08	−0.01	−0.36	−0.14	−0.24	0.08	0.73	0.05	−0.02	0.27	0.36
3P%	0.01	0.41	−0.24	0.41	1.00	0.21	0.45	0.18	0.28	0.44	0.64	0.01	−0.10	0.27	−0.21	−0.59	−0.13	−0.35	−0.22	0.43	0.03	−0.17	0.03	0.59
3PA	−0.58	0.62	−0.85	0.98	0.21	1.00	0.37	0.03	0.45	0.42	0.12	0.52	0.04	0.02	0.04	−0.24	−0.12	−0.18	0.14	0.68	0.04	0.01	0.28	0.24
AST	0.26	0.66	−0.17	0.45	0.45	0.37	1.00	0.33	0.46	0.74	0.65	0.44	−0.10	0.16	−0.17	−0.41	−0.06	−0.22	0.00	0.66	0.27	0.14	0.25	0.41
BLK	0.15	0.29	−0.04	0.07	0.18	0.03	0.33	1.00	0.32	0.23	0.34	0.00	0.01	0.04	0.00	−0.34	−0.01	−0.05	0.00	0.19	0.16	0.16	0.24	0.34
DRB	0.09	0.46	−0.19	0.48	0.28	0.45	0.46	0.32	1.00	0.60	0.34	0.56	0.06	0.08	0.03	−0.40	0.02	−0.05	−0.13	0.63	−0.28	−0.02	0.82	0.40
FG	0.47	0.76	−0.04	0.49	0.44	0.42	0.74	0.23	0.60	1.00	0.75	0.74	0.00	0.20	−0.08	−0.43	−0.07	−0.04	0.05	0.89	0.14	−0.04	0.49	0.43
FG%	0.47	0.81	−0.07	0.25	0.64	0.12	0.65	0.34	0.34	0.75	1.00	0.12	0.07	0.25	−0.03	−0.64	−0.12	−0.30	−0.12	0.65	0.22	0.00	0.11	0.64
FGA	0.23	0.33	0.00	0.49	0.01	0.52	0.44	0.00	0.56	0.74	0.12	1.00	−0.07	0.05	−0.09	0.00	0.02	0.24	0.19	0.69	−0.01	−0.06	0.61	0.00
FT	−0.02	0.15	−0.09	0.02	−0.10	0.04	−0.10	0.01	0.06	0.00	0.07	−0.07	1.00	0.21	0.91	−0.14	0.05	0.19	0.20	0.32	0.02	0.09	0.16	0.14
FT%	0.11	0.16	0.00	0.08	0.27	0.02	0.16	0.04	0.08	0.20	0.25	0.05	0.21	1.00	−0.20	−0.19	0.02	−0.31	−0.18	0.24	−0.15	−0.30	−0.11	0.19
FTA	−0.07	0.09	−0.10	−0.01	−0.21	0.04	−0.17	0.00	0.03	−0.08	−0.03	−0.09	0.91	−0.20	1.00	−0.06	0.04	0.31	0.27	0.22	0.08	0.21	0.21	0.06
L	−0.06	−0.57	0.28	−0.36	−0.59	−0.24	−0.41	−0.34	−0.40	−0.43	−0.64	0.00	−0.14	−0.19	−0.06	1.00	0.00	0.07	0.23	−0.48	−0.20	0.25	−0.30	−1.00
MP	0.07	−0.14	0.15	−0.14	−0.13	−0.12	−0.06	−0.01	0.02	−0.07	−0.12	0.02	0.05	0.02	0.04	0.00	1.00	0.07	0.10	−0.08	−0.09	−0.03	0.06	0.00
ORB	0.20	−0.30	0.35	−0.24	−0.35	−0.18	−0.22	−0.05	−0.05	−0.04	−0.30	0.24	0.19	−0.31	0.31	0.07	0.07	1.00	0.25	−0.06	−0.02	0.22	0.54	−0.07
PF	−0.03	0.03	−0.04	0.08	−0.22	0.14	0.00	0.00	−0.13	0.05	−0.12	0.19	0.20	−0.18	0.27	0.23	0.10	0.25	1.00	0.12	0.27	0.48	0.04	−0.23
PTS	0.13	0.83	−0.37	0.73	0.43	0.68	0.66	0.19	0.63	0.89	0.65	0.69	0.32	0.24	0.22	−0.48	−0.08	−0.06	0.12	1.00	0.12	−0.01	0.50	0.48
STL	0.09	0.24	−0.06	0.05	0.03	0.04	0.27	0.16	−0.28	0.14	0.22	−0.01	0.02	−0.15	0.08	−0.20	−0.09	−0.02	0.27	0.12	1.00	0.31	−0.25	0.20
TOV	−0.02	0.06	−0.06	−0.02	−0.17	0.01	0.14	0.16	−0.02	−0.04	0.00	−0.06	0.09	−0.30	0.21	0.25	−0.03	0.22	0.48	−0.01	0.31	1.00	0.11	−0.25
TRB	0.20	0.21	0.05	0.27	0.03	0.28	0.25	0.24	0.82	0.49	0.11	0.61	0.16	−0.11	0.21	−0.30	0.06	0.54	0.04	0.50	−0.25	0.11	1.00	0.30
W	0.06	0.57	−0.28	0.36	0.59	0.24	0.41	0.34	0.40	0.43	0.64	0.00	0.14	0.19	0.06	−1.00	0.00	−0.07	−0.23	0.48	0.20	−0.25	0.30	1.00

Figure 17: Correlation matrix for modern NBA data

1993-1998 Data Testing

Figure 18 in the appendix shows the correlation matrix for the 90s data. We showed that the most important feature is the 2-point percentage with 0.65 which matches the permutation importance result. The defensive rebound result shows a correlation of 0.51 making it the third most important feature after 2-point percentage and field goal percentage, both with 0.65. The assists also have a correlation of 0.51. Yet again the turnovers were shown as a negative in the correlation matrix with a score of -0.44.

Testing Conclusion

From the results it was concluded that the permutation importance results are accurate as 4 out of 5 of the features selected matched the correlation matrix. The anomaly found within the correlation matrix is that turnovers have a negative impact on a team's wins. Due to this the turnovers have not been used as a positive feature in the visualisations. In the future I would like to collect more seasons of data to analyse the turnover anomaly we found.

Comparing and Contrasting NBA Eras with Visualisations

The biggest differences found between this era is the offense. The data from 1993-98 shows a larger number of 2-point shots being taken. As previously mentioned, this is because of the triangle offence system. The triangle offense "combines perfect spacing with a series of actions based on player decisions resulting in a beautiful basketball offensive system" [18]. If we look at figure 19, we showed that the teams with a higher number of wins were scoring a higher percentage of the 2-point shots they were taking than teams with a low number of wins. Using the line of best fit we found a strong correlation between the 2-point percentage and a team winning.

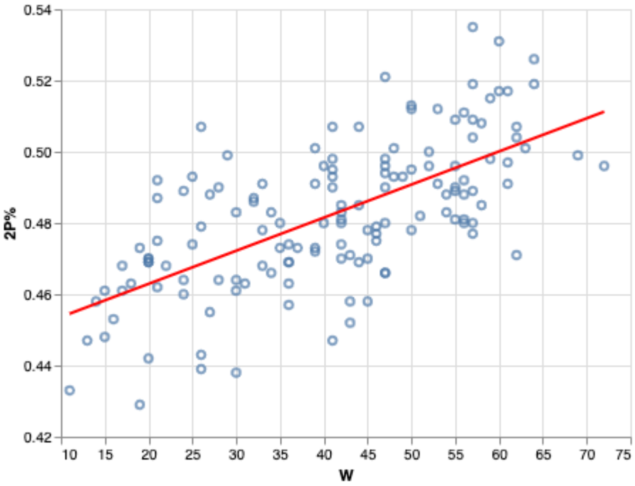


Figure 19: Line of best fit for 2P% and Wins

Using figures 20 and 21 we showed the decrease in 2-point shots scored between eras. An example of this is the Boston Celtics taking an average of 80.2 2-point shots per game during the 93-94 season, however in the 2014-15 season they are taking 56 2-point shots a game which is 24 less per game. This highlights the different type of offence that was played being played and shows that teams have been successful without taking many 2-point shots.

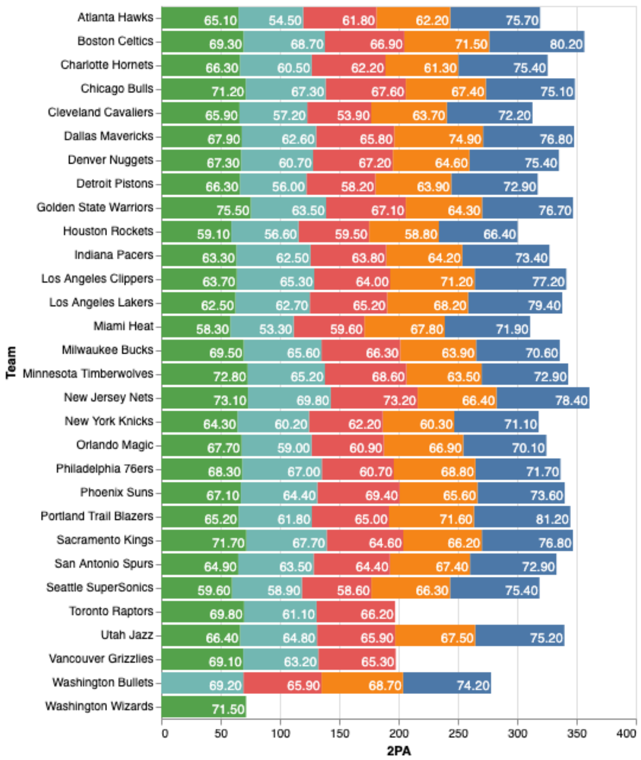


Figure 20: 1993-98 2-point attempts

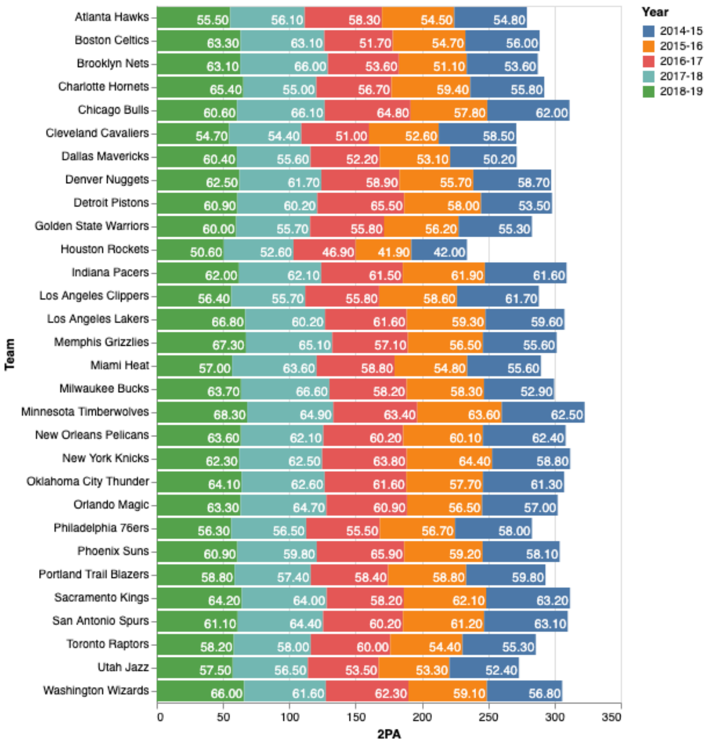


Figure 21: 2014-19 2-point attempts

Why are teams still thriving without taking and scoring as many 2-point shots?

The 3-point shot was introduced in 1961 to give "shorter" players a chance to make up for their lack of height [19]. It was seen as a way to make basketball more exciting and fast paced, however coaches saw it as a risky and unreliable shot and players during older eras simply did not have the skill set. This is why teams continued to trust the 2-point shot as they found more success with it. It did not have the impact on the NBA it was expected to… This was until the early 2010s, Steph Curry was drafted by the Golden State Warriors in 2009 and in the following years went on to solidify himself as the greatest three-point shooter of all time. He would test the limits and shoot from distances that had never been shot from before but most importantly he was doing this on great efficiency. The Golden State Warriors made the finals 5 times in a row from 2014 to 2019 and won 3 titles during this time span. Using the three-point shot as a weapon and being efficient in doing so led all teams around the league to change their game style. It is now expected by all players to be able to shoot the 3-point shot on good efficiency (30-40%). The triangle offense that was so successful during the 90s is no longer used in the NBA as 2-point shots simply aren't as valuable as a 3-point shot.

As talked about previously the Houston Rockets have been one of the most successful teams, winning 65 games during the 2017-18 season finishing top of their conference, using this new 3-point focused offence. Using figures 22 and 23 we showed that during the 1993-94 season the Rockets were shooting 15.7 3-pointers per game, this was higher than most of the other teams during this era, however looking at the 2014-15 season we can see that they attempted 45.40 3-pointers per game. This is an increase of 30 3-pointers per game.

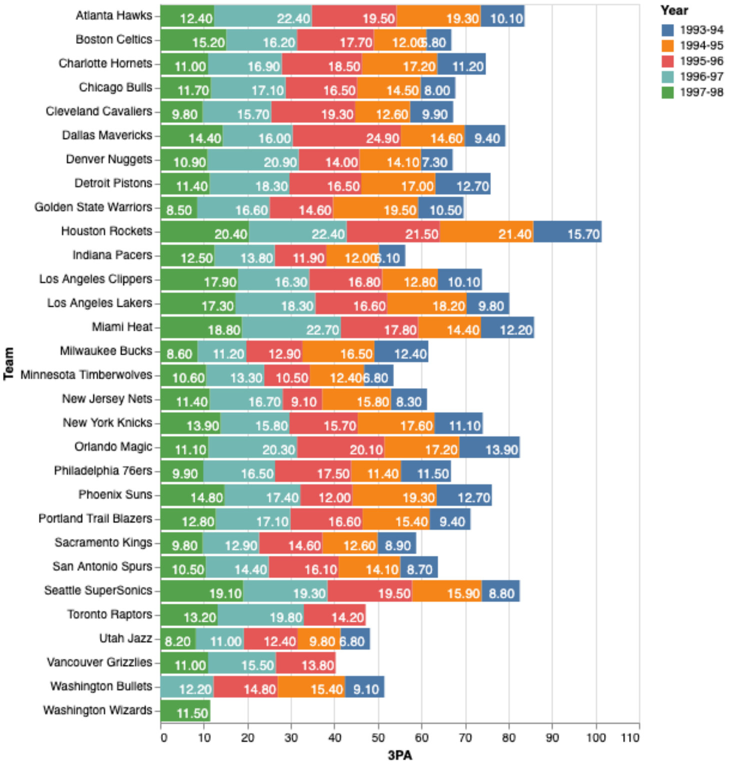


Figure 22: 3-point attempts during 1993-98 seasons

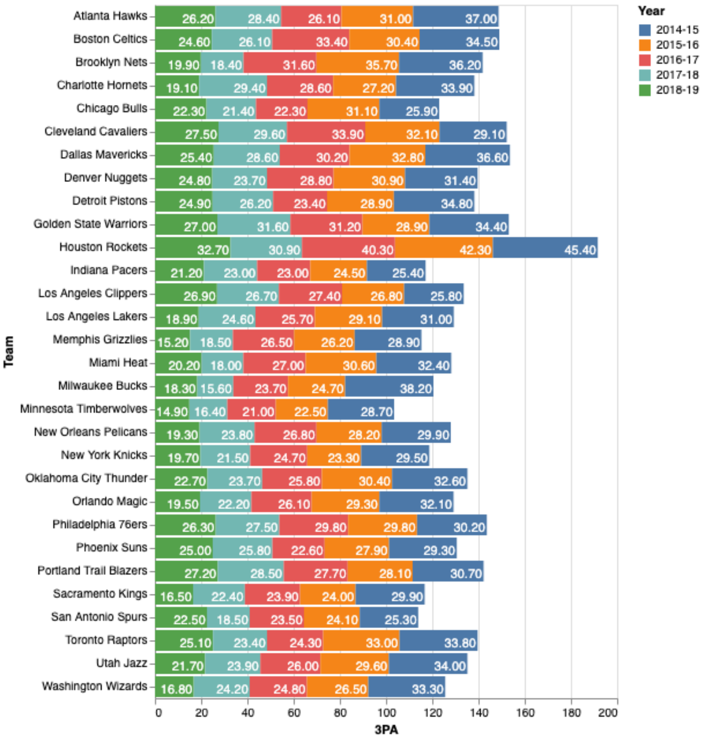


Figure 23: 3-point attempts during 2014-19 seasons

In figure 24 we showed that the teams scoring a higher percentage of the 3-pointers were able to win more games. The line of best fit showed a correlation between the percentage and wins. The top right plot showed that the Warriors during the 2015-16 season where they shot an average of 0.41% on threes for the season and achieved 73 wins which broke the NBA record for wins during a season.

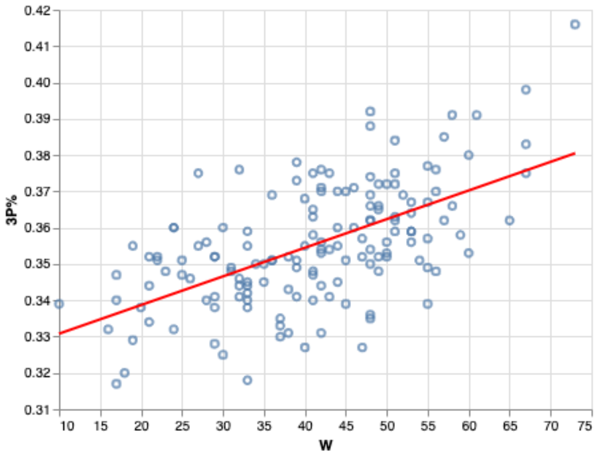


Figure 24: Correlation of wins and 3-point percentage during the 2014-19 season.

How has defence changed for successful teams?

During the permutation importance we found that defensive rebounds were the second most important feature for teams during the 90s era but was not a top 5 feature for the modern era. The term “defence wins championships” was coined by Bear Bryant a football coach but is this still the case? Using figures 25 and 26 in the appendix, we found that the number of defensive rebounds per game in the modern data were actually greater the number of defensive rebounds during the 90s era – yet the feature was a lot more important for teams during the 90s. The answer for this that was found is simple – more shots are being taken during the modern era, so more shots are missed which leads to more defensive rebounds. The influx in shots being taken has made the defensive rebound less important during the modern era. We can also see that shots being taken closer to the basket led to less misses and less defensive rebounds. Another important feature for the 90s era was the steals per game, using our visualisations we showed that teams were averaging more steals during this era compared to the modern data. This comes down to the triangle offense being played, a lot more ball movement between players which led to more steals by the opposition. During the 93-94 season we found that the Atlanta Hawks had 11.20 steals per game yet in the 2014-15 season they only averaged 8.20. We showed that defence plays less of a role in this era, players shooting from further away and giving up the ball with turnover less means there’s not as many chances to steal the ball for your opponent. Ironically, the assists were the fifth most important statistic during the 90s era, so although they gave up the ball more, the offense being pass first focused was a key factor in their own success. I was surprised after looking at the blocks, I found that although teams averaged more blocks per game during the 90s it was a more important statistic during the modern era. The majority of teams averaged between 4 and 5 blocks per game with a few anomalies such as Golden State averaging between 6 and 7 for all seasons of the modern data. I created a scatter matrix to analyse the defence for the modern era. Looking at figure 27 below we showed that there are many teams averaging over 40 wins per season without being on the high side end of defensive rebounds suggesting teams are winning without playing as much defence. This resulted in a low correlation between steals and wins as the plots are scattered all over the graph with no pattern. From this we concluded that there is a lower amount of defence being played in today’s NBA, yet this has not affected the top teams due to them covering this in other areas of the court e.g., offensively – scoring a higher percentage of threes than their opponents.

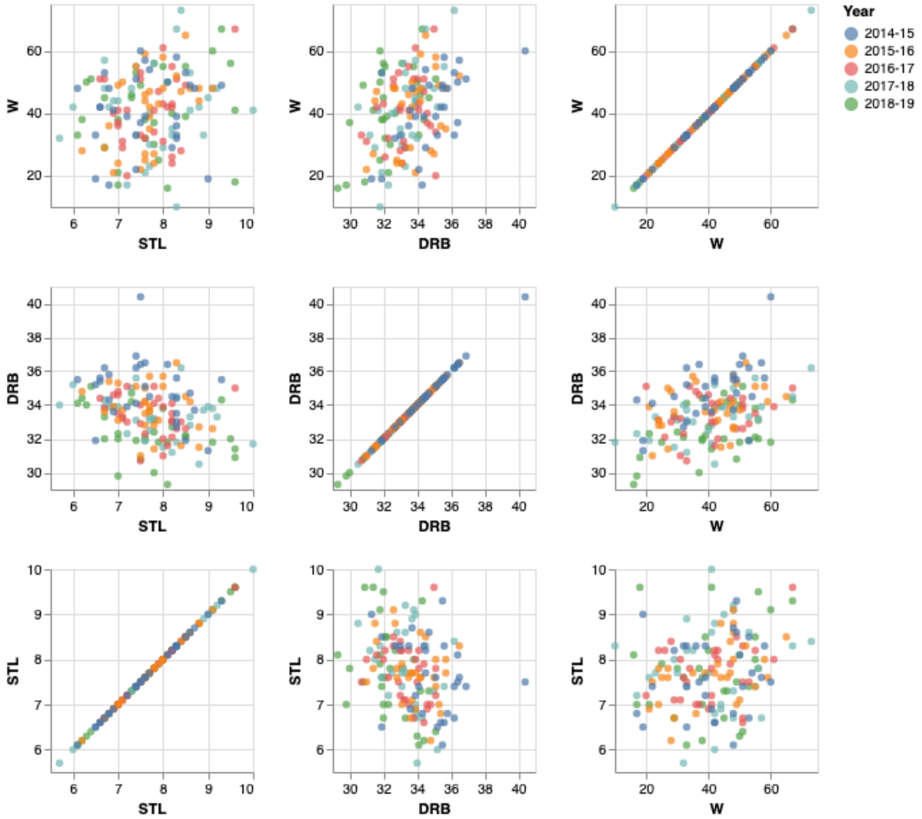


Figure 27: Scatter Matrix of defensive rebounds, wins and steals

Trends Between Eras

The offensive rebounds were one of the worst features for both eras coming last for the modern data and second from last for the 90s era. Using the visualisations, we found that teams were usually averaging more assists during the 90s than the modern era. This was due to the triangle offense which put players closer to the basket – making it easier to grab the rebound. The New Jersey Nets were the best offensive rebounding team from 1993-98 but found that this did not help their wins as they remained an average team for all those years, only just finishing over the 0.500 (balanced wins and losses) mark for wins.

It was surprising to find that the 2-pointers scored for the 90s era and 3-pointers scored for the modern era were low on the permutation importance results. Before this project I thought they would've been top features, however by creating visualisations I was able to show why this was not the case. We found that more shots do not correlate to more wins if the team is shooting at a low percentage – this was true for both eras and is shown in figure 19 and 24.

Chapter 6

Management

6.1 Management and Reflection

I believe I managed this project well for multiple reasons. Firstly, I started my implementation early, I did this to ensure I would meet all my aims and be able to change things if any risks were encountered. I set deadlines for myself such as collecting and filtering the data, training my data and using permutation importance on it and lastly creating visualisations. I allowed longer periods of time for tasks that would take the longest, I did this for the permutation importance as I had to learn and understand the machine learning concepts before implementation. Secondly, I was happy with the programs that I had chosen to use such as Altair, after gaining some experience with it during my university modules I was able to the visualisations I had hope to and more. I found that I was very motivated throughout my project, I have a big interest in sports analytics and visualisation so doing a project on that topic kept me on track. Following an agile approach was very important during this project, although I had set deadlines for myself, I wanted to be able to go back and improve parts of my project if need be. An example of this was collecting more seasons of data to improve the accuracy of my results, as I had initially started on 1 season of data from each era but increased this to 5 seasons (10 in total). I am happy with how my project has turned out, I was able to meet all of the aims that I set out and have made important analysations on modern and historic NBA data – showing the differences in eras and what statistics contribute most to a team winning.

6.2 Risks

I did not encounter many risks during my project, I attribute this to the way I managed my project, but I will discuss the few that I had. The first risk I encountered was during the data collection, I was struggling to find a free resource that allowed me to download and modify NBA data, the data I first found was subpar and did not show all of the game features. By going back and spending more time looking for a good source of data I was able to find Basketball Reference which had every statistic that I needed for each team since game data had first been recorded. This vastly improved my project as I was able to go into much deeper detail on what features attribute to a team's wins and how the NBA eras have changed. Another risk I prepared for was losing work that I had spent a lot of hours on. My mitigation strategy for this was control. I used Jupyter Notebooks autosave feature to ensure I did not lose any work and confirm that I was in control of my work. I also had backups stored online which I could access from different devices.

Chapter 7

Future Work and Conclusion

7.1 Future Work

As I have mentioned during this paper, if I had unlimited time for this project, I would've liked to collect more than 5 years of data for each era. The process of collecting data was longer than anticipated due to the fact that I had to combine multiple datasets as some did not contain all of the information, I needed e.g. I had to individually add a team's record (wins and losses) to their in-game features. By collecting more data, I would be able to obtain more accurate results. Another option would be to use live data so that my visualisations could update in real time, to do this I could use web scraping which I originally looked into before collecting the data from the website basketball reference.

To take my project further I would like to combine what I have done with an app or website. It would be very useful to see these visualisations at any moment and gave them update in real time with live data. I could then focus on the individual match outcomes using game simulations and make predictions. This could then be applied to gambling, punters could make "safer" bets using the visualisations that I have created. This is something I would find very useful when playing fantasy basketball against my friends and people around the world. I did not focus on the gambling side of things during this paper due to the negative connotations associated with it, but I think it could be an interesting route to go down.

7.2 Conclusion

This has been a challenging project that I had to learn a lot of new skills for. Prior to this project I had never used Python for anything other than a few simple programs. To be able to train a large volume of data with random forest and to use machine learning techniques such as permutation importance on that data to find valuable attributes was very satisfying to accomplish. Machine learning is viewed as a difficult concept, but I was happy with how I was able to apply it in my project. I was able to create a variety of different visualisations such as matrixes and stacked histograms to show trends and correlations between features in the data and learn why teams were successful in different eras.

During my related work I was surprised to see how little work had been done on basketball visualisation. I found that other academic papers focused on game predictions and the betting side of basketball statistics. This was a very worthwhile project due to the fact that not much work has been done like this. To be able to compare two different eras and why teams were successful in different decades is very important information. Modern teams can learn from previous decades and become more efficient and successful. I also found that other pieces of work would focus individual games and seasons whereas I have used a large range of data to make accurate statements.

The main take away from this project for a coach or team would be the change in offensive strategy. I have shown that the 3-point shot is more valuable in today's era than the 2-point shot, players during the last few years are becoming more skilful and the traditional positions in basketball are disappearing. All players coming into the league today would be expected to shoot a high percentage 3-pointer no matter their height. With this change in strategy basketball has become a much faster sport and has grown in popularity around the world. Smaller players being able to acquire a new, valuable skill has led to more and more people playing the sport. This paper has conveyed why teams have been successful in the past and present using machine learning techniques and visualisations. Using the visualisations created we have been able to find trends in the data shared across different eras of NBA basketball.

Bibliography

[0] A Guide to Sabermetric Research – Society for American Baseball Research, Sabr.org, https://sabr.org/sabermetrics.

[1] Koza, John R. 1996. How can computers learn to solve problems without being explicitly programmed? https://link.springer.com/chapter/10.1007%2F978-94-009-0279-4_9.

[2] British Board of Film Classification. 2011. Moneyball. United Kingdom. https://www.bbfc.co.uk/release/moneyball-film-qxnzzxq6vlgtotg3ntyy.

[3] Nick Giles, Arthur House. 2019. How stats helped Leicester City F.C win the Premier League. https://playr.catapultsports.com/blog/how-stats-helped-leicester-win-the-premier-league/.

[4] Stephen Chu. 2010. Information Visualization in the NBA: The Shot Chart. https://www.basketclubs.es/newyork/wp-content/uploads/sites/3/2015/09/NBA-statistics.pdf.

[5] Thabtah, F., Zhang, L. & Abdelhamid, N. 2019. NBA Game Result Prediction Using Feature Analysis and Machine Learning. https://doi.org/10.1007/s40745-018-00189-x.

[6] Daryl Morey, Jessica Gelman. 2006. MIT SLOAN SPORTS ANALYTIC CONFERENCE, http://www.sloansportsconference.com/about/.

[7] ZS. 2018. How the Houston Rockets have taken basketball analytics to the next level and risen to the top of the NBA. https://digital.hbs.edu/platform-digit/submission/moreyball-the-houston-rockets-and-analytics/.

[8] ZS. 2018. How the Houston Rockets have taken basketball analytics to the next level and risen to the top of the NBA. https://digital.hbs.edu/platform-digit/submission/moreyball-the-houston-rockets-and-analytics/.

[9] Priya Pedamkar. Matplotlib In Python. https://www.educba.com/matplotlib-in-python/.

[10] Jarrell, Stephen B. 1994. Basic Statistics (Special pre-publication ed).

[11] Mike Yi. 2019. A complete guide to stacked bar charts. https://chartio.com/learn/charts/stacked-bar-chart-complete-guide.

[12] ArcGIS. 2.7. Scatter plot matrix. https://pro.arcgis.com/en/pro-app/latest/help/analysis/geoprocessing/charts/scatter-plot-matrix.htm.

[13] L. Breiman. 2001. Random Forests, Machine Learning. https://link.springer.com/article/10.1023/A:1010933404324.

[14] Jeffrey M. Perkel. 2018. Why Jupyter is data scientists' computational notebook of choice. https://www.nature.com/articles/d41586-018-07196-1.

[15] Vumaasha. 2018. What is random state? https://stackoverflow.com/questions/49147774/what-is-random-state-in-sklearn-model-selection-train-test-split-example.

[16] RandomForestRegressor, Scikit Learn. https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html.

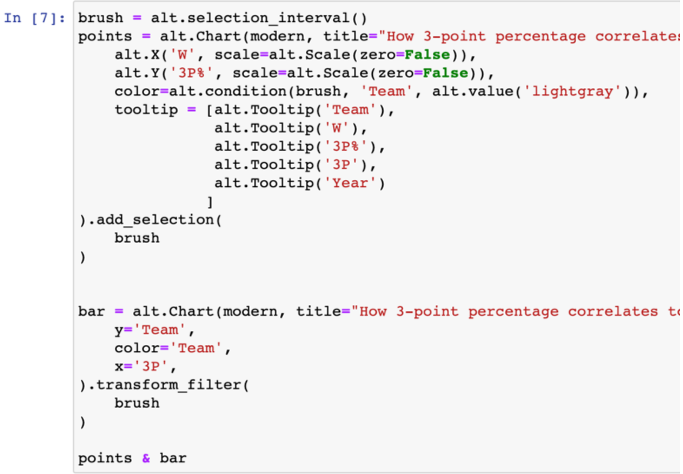
[17] 2016. Basketball – A game of Geometry, Relativity Digest. https://relativitydigest.com/2016/11/12/basketball-a-game-of-geometry/.

[18] Coach Mac. 2021. Triangle Offense, Basketball for Caoches. https://www.basketballforcoaches.com/triangle-offense/.

[19] Hoops Geek. 2021. The History and Evolution of the Three-Point Shot. https://www.thehoopsgeek.com/history-three-pointer/.

Appendix A

Implementation

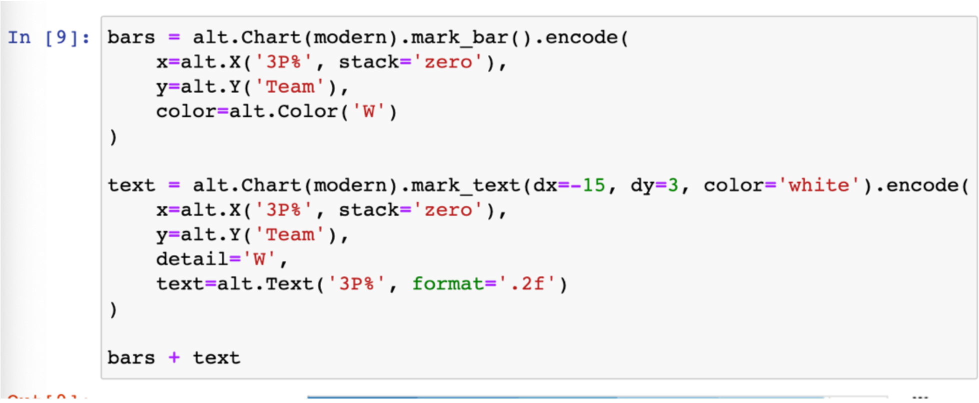


```
In [7]: brush = alt.selection_interval()
        points = alt.Chart(modern, title="How 3-point percentage correlates
            alt.X('W', scale=alt.Scale(zero=False)),
            alt.Y('3P%', scale=alt.Scale(zero=False)),
            color=alt.condition(brush, 'Team', alt.value('lightgray')),
            tooltip = [alt.Tooltip('Team'),
                       alt.Tooltip('W'),
                       alt.Tooltip('3P%'),
                       alt.Tooltip('3P'),
                       alt.Tooltip('Year')
                      ]
        ).add_selection(
            brush
        )


        bar = alt.Chart(modern, title="How 3-point percentage correlates t
            y='Team',
            color='Team',
            x='3P',
        ).transform_filter(
            brush
        )

        points & bar
```

Figure 13: Interactive scatter graph with histogram code



```
In [9]: bars = alt.Chart(modern).mark_bar().encode(
            x=alt.X('3P%', stack='zero'),
            y=alt.Y('Team'),
            color=alt.Color('W')
        )

        text = alt.Chart(modern).mark_text(dx=-15, dy=3, color='white').encode(
            x=alt.X('3P%', stack='zero'),
            y=alt.Y('Team'),
            detail='W',
            text=alt.Text('3P%', format='.2f')
        )

        bars + text
```

Figure 14: Stacked bar chart code

Appendix B

Visualisations

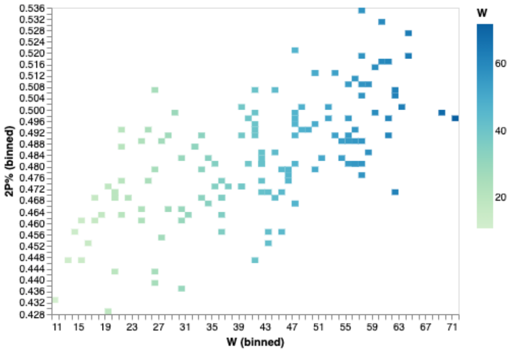
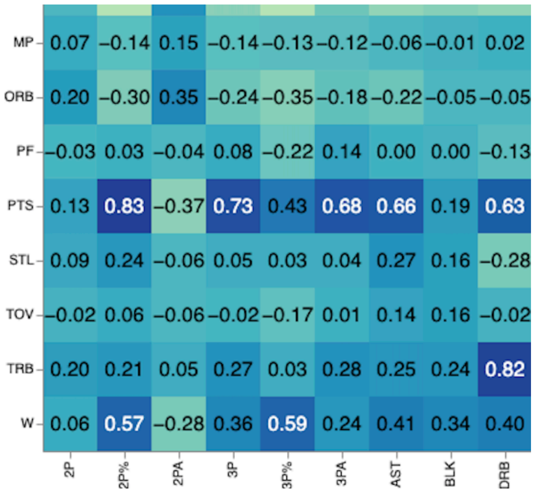


Figure 4: Binned heatmap using 90's era data to show correlation between wins and 2 point percentage



	2P	2P%	2PA	3P	3P%	3PA	AST	BLK	DRB
MP	0.07	−0.14	0.15	−0.14	−0.13	−0.12	−0.06	−0.01	0.02
ORB	0.20	−0.30	0.35	−0.24	−0.35	−0.18	−0.22	−0.05	−0.05
PF	−0.03	0.03	−0.04	0.08	−0.22	0.14	0.00	0.00	−0.13
PTS	0.13	0.83	−0.37	0.73	0.43	0.68	0.66	0.19	0.63
STL	0.09	0.24	−0.06	0.05	0.03	0.04	0.27	0.16	−0.28
TOV	−0.02	0.06	−0.06	−0.02	−0.17	0.01	0.14	0.16	−0.02
TRB	0.20	0.21	0.05	0.27	0.03	0.28	0.25	0.24	0.82
W	0.06	0.57	−0.28	0.36	0.59	0.24	0.41	0.34	0.40

Figure 5: Small snippet of correlation matrix

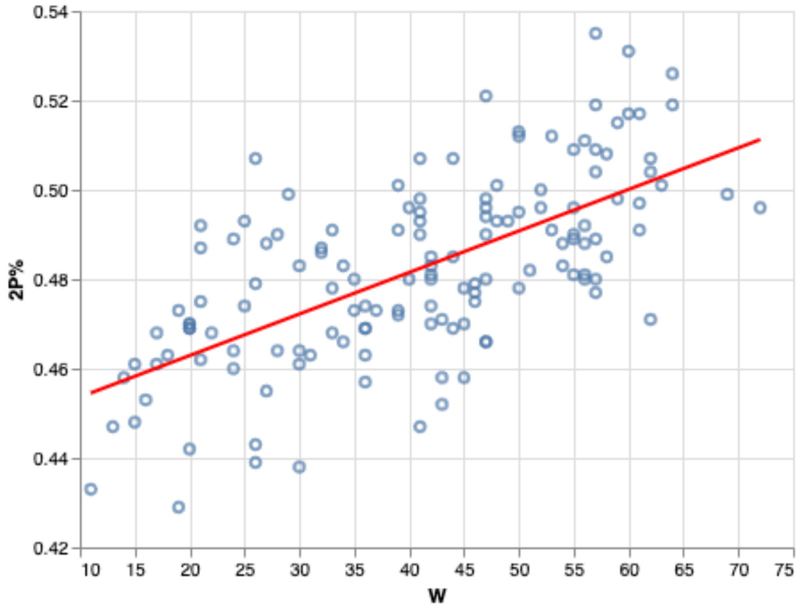


Figure 6: Line of best fit between the 2-point percentage and wins

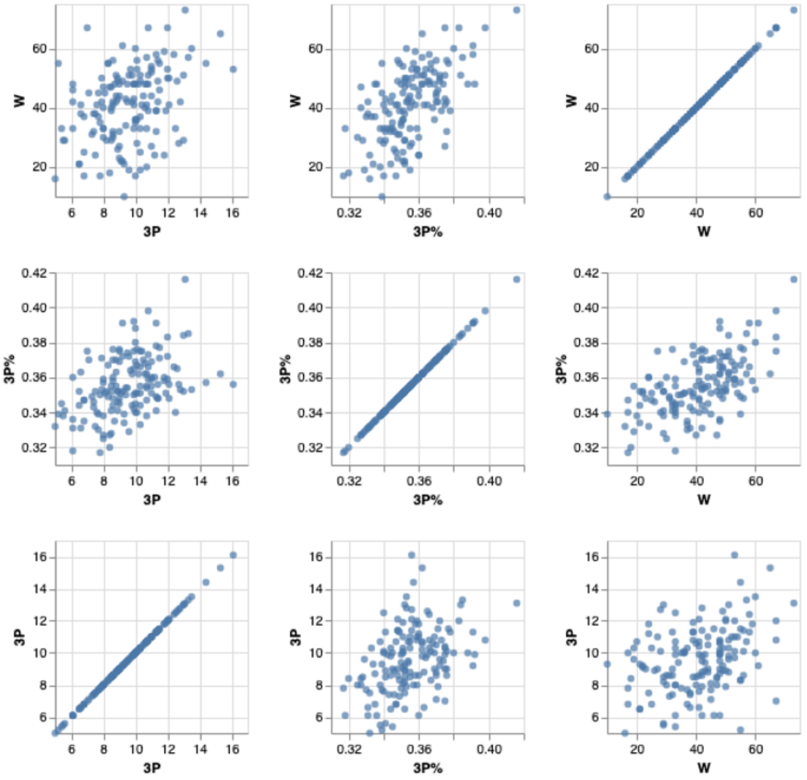


Figure 7: Scatter matrix between wins 3-pointers and 3-point percentages.

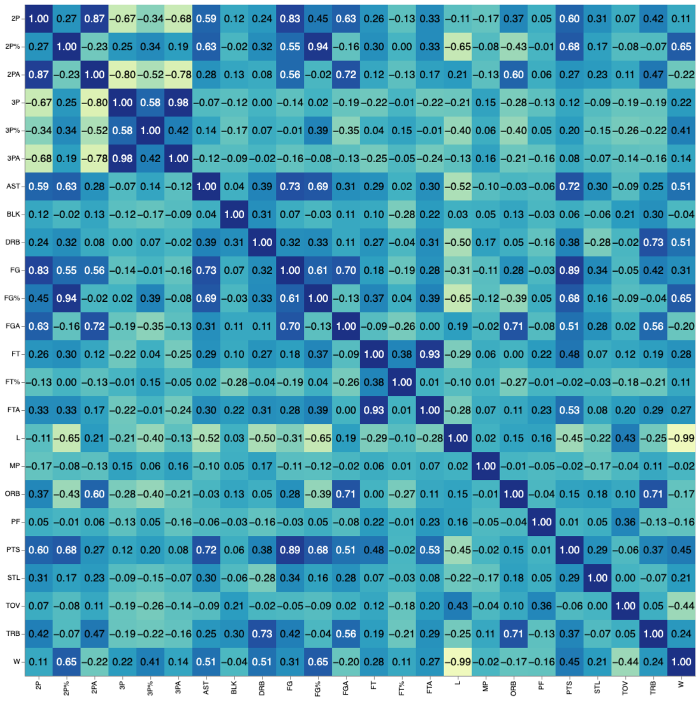


Figure 18: Correlation matrix for 90s NBA data

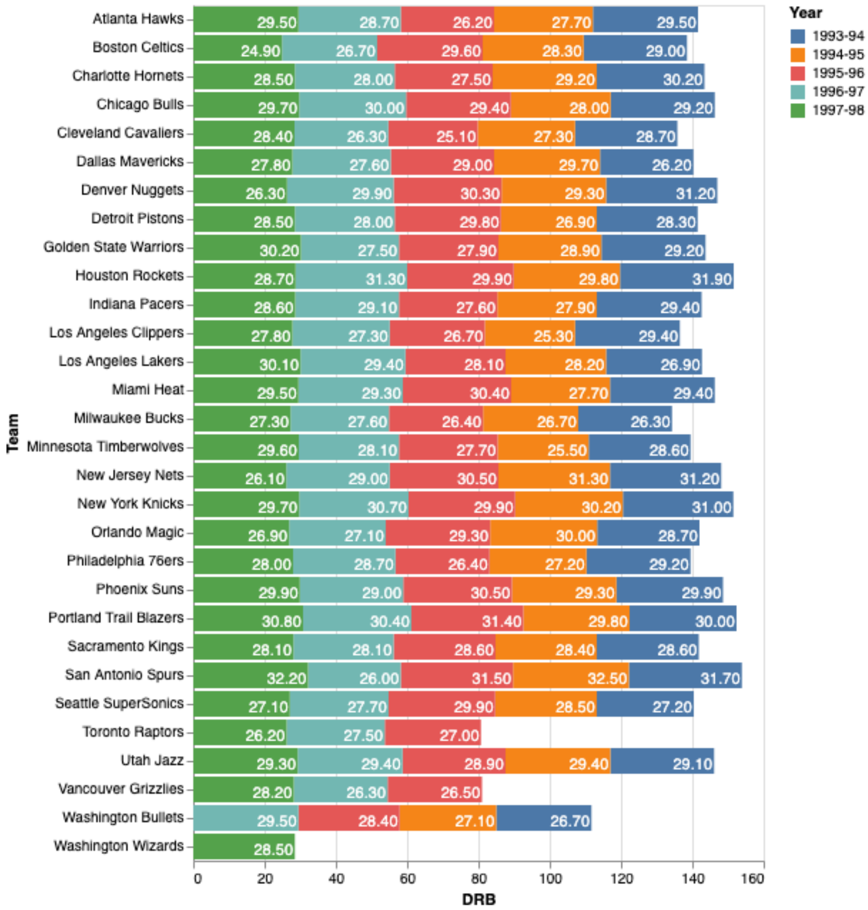


Figure 25: Defensive rebounds during 1993-98 seasons

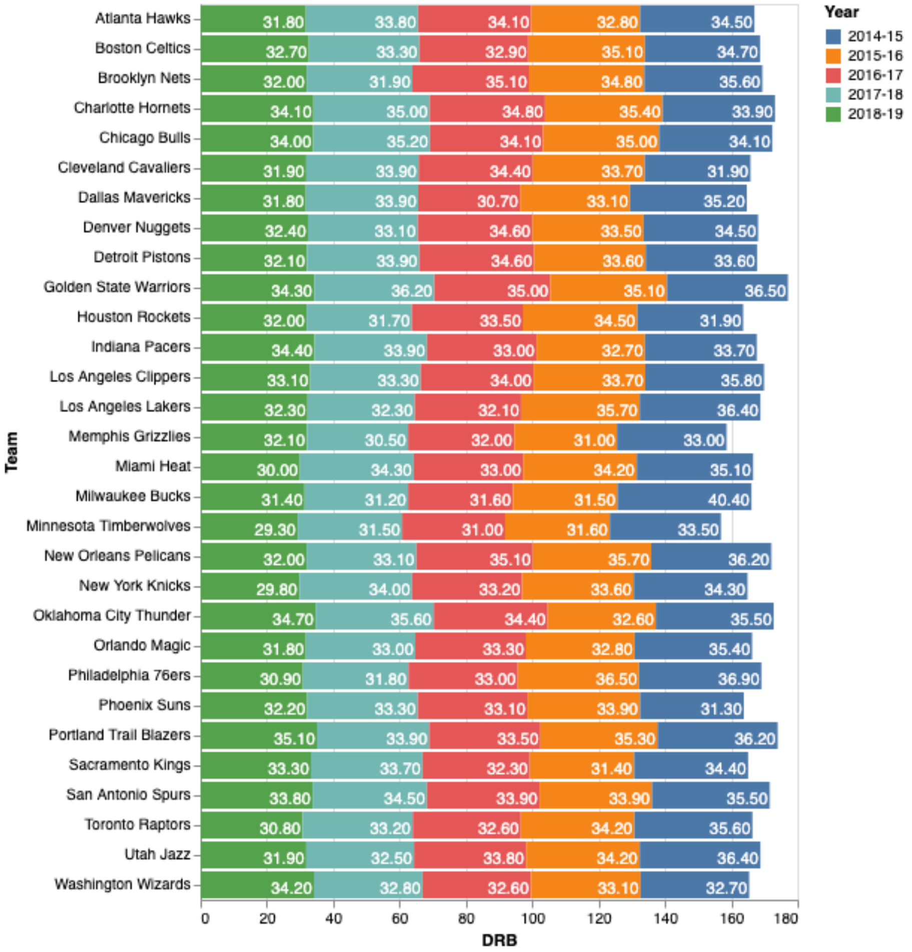


Figure 26: Defensive rebounds during 2014-19 seasons