**BASKETBALL REPRODUCIBLE ANALYSIS PROJECT**

**1.INTRODUCTION:**

Basketball is a popular sport worldwide, with millions of fans and players. In the National Basketball Association (NBA), teams compete for the championship title by playing against each other in a series of games. Each basketball team comprises five positions: point guard, shooting guard, small forward, power forward, and centre. Each position has different requirements, and players with distinct skill sets are required to form a balanced and successful team.

The Chicago Bulls is an NBA team that finished the 2018-19 season in 27th place out of 30 based on win-loss record, with a player contract budget that ranked 26th out of 30 in the following season (2019-20). The team struggled to put together a competitive team and required a data-driven approach to improve their performance in the upcoming season. As a data analyst, the aim of the project is to use data analysis to identify the best five starting players for the Chicago Bulls for the upcoming NBA season.

This project's justification is based on the team's performance, where the Chicago Bulls requires a team that can compete at a high level consistently. By identifying the best players for each position and optimizing the team's budget, the project aims to provide the team with actionable insights that can help improve their win-loss record and compete more effectively in the NBA. The ultimate goal is to assist the team in building a competitive and successful team that can perform well in the league. Additionally, the project will provide insights into areas where the team can improve and make data-driven decisions to maximize their chances of success.

The limitations of the project include the unavailability of data on players' injuries and other factors that may impact their performance, which may affect the project's accuracy. Furthermore, the project's scope is limited to the budget allocated to player contracts, and it does not consider other factors that may impact the team's performance, such as coaching strategies and team dynamics.

In summary, this project's aim is to identify the best five starting players for the Chicago Bulls for the upcoming NBA season using data analysis, taking into account player statistics from the previous season and their salaries. The project's importance lies in its potential to impact the team's performance and success, which could have far-reaching implications for the team's future. By providing the team with actionable insights, the project aims to assist the team in building a competitive and successful team that can perform well in the NBA

**2.Data reading and cleaning**

To complete the task assigned by the general manager of the Chicago Bulls, we need to find the best five starting players (one from each position) the team can afford with the budget of $118 million. We have been provided with five datasets. These are:

**2018-19\_nba\_player-statistics.csv**: It contains total statistics for individual NBA players during the 2018-19 season.

**2018-19\_nba\_player-salaries.csv**: It contains the salary for individual players during the 2018-19 season.

**2019-20\_nba\_team-payroll.csv**: It contains payroll for all teams in the 2019-20 season.

**2018-19\_nba\_team-statistics\_1.csv**: It contains team statistics for the 2018-19 season.

**2018-19\_nba\_team-statistics\_2.csv**: It contains advanced team statistics for the 2018-19 season.

We will use these datasets to find the best five starting players from the Chicago Bulls' and also from other team who can provide the best performance, keeping in mind the salary budget of $118 million. We will perform exploratory data analysis, data cleaning, and data merging of these datasets. We will also use data modelling, visualization tools to analyze and interpret the data to answer the question of finding the best starting players for Chicago Bulls. Finally, we will produce a reproducible report hosted on GitHub.

This process is important in data analysis because it ensures that the data is accurate, consistent, and ready for further analysis.

The first step is to load the data into R using the "read.csv" function. The function reads a CSV file and creates a data frame in R. In this code, four datasets are loaded - player statistics, player salaries, team payroll, and team statistics. The "check.names" argument is set to FALSE to avoid any issues with column names that contain spaces or special characters.

After loading the data, the code checks for missing values in each dataset using the "colSums" function with the "is.na" function. This provides an idea of the amount and location of missing data, which may need to be imputed or removed before analysis.

The "str" function is used to examine the structure of each dataset, which helps to identify the data types and the number of observations and variables. This information is useful for understanding the data and preparing it for analysis.

The next step involves changing the data types of certain variables. In this code, the "as.numeric" function is used to convert the "salary" column in the team payroll dataset from character to numeric data type. This is necessary to perform mathematical operations on this column.

Finally, the code cleans the data by mapping team names to their corresponding abbreviations or full names. This helps to ensure consistency across datasets, especially when merging datasets based on team names. For example, the player statistics dataset uses team names in their full form while the team payroll dataset uses team names in their abbreviated form. The code maps each abbreviated team name to its full form using the "case\_when" function. Similarly, the team payroll dataset is cleaned by mapping the team names to the brief team names used in the brief.

In summary, loading and cleaning data is an important step in data analysis, as it ensures the accuracy and consistency of the data. This code provides an example of how to load and clean several datasets related to NBA teams and players. By following the code, one can load and clean other datasets in R as well.

**3.Exploratory data analysis:**

Exploratory data analysis (EDA) is a crucial step in data analysis that allows data scientists to gain a deeper understanding of the data and identify any potential issues that need to be addressed before moving on to more complex analysis. In this process, the data is examined in a variety of ways to identify patterns, relationships, and anomalies.

First, it reads in five data frames: player\_stats, player\_salaries, team\_payroll, team\_stats1, and team\_stats2. The function glimpse() is then used to view the structure of each data frame. glimpse() is a useful function for quickly viewing the structure of a data frame, including the variable names, their data type, and the first few observations.

Next, the distribution of variables is checked using summary statistics and visualizations. For the player\_stats data frame, the distribution of player age and points are visualized using histograms. For the player\_salaries and team\_payroll data frames, the distribution of salaries is visualized using histograms with a logarithmic scale on the y-axis to better view the distribution. For team\_stats1 and team\_stats2 data frames, the age and point distributions are visualized respectively using histograms.

Next, it checks for relationships between variables or differences between groups. Firstly, the player\_stats and player\_salaries data frames are merged using the join() function. The complete.cases() function is then used to remove any rows with null values in the player\_id column. A scatter plot is then created to understand the relationship between salary and points across players, which shows a positive correlation between the two variables.

To further explore relationships between variables, a correlation matrix is calculated for the numeric variables in the player\_stats\_salaries data frame. The cor() function is used to calculate the correlation matrix, and the melt() function from the reshape2 package is used to convert the matrix into a long-format data frame that can be used to create a heatmap with ggplot2. Finally, a heatmap is created using ggplot2 to visualize the correlation matrix.

It merges the player\_stats\_salaries and team\_stats\_salary data frames using the full\_join() function, which combines the data frames based on a common variable, in this case, the team name. The merged data frame can be used for further analysis.

In summary, it demonstrates various data exploration and manipulation techniques, including data structure inspection, variable distribution visualization, scatter plots, correlation matrix calculation, and data frame merging. These techniques are useful for understanding the data, identifying patterns, and gaining insights that can inform further analysis.

The decision to develop a data model that is easily reproducible is crucial for this task as it allows us to modify and update our analysis in the future. As the data analyst for the Chicago Bulls, we need to ensure that we can make informed decisions when selecting players for the team, and a model that is flexible and scalable is essential for achieving this.

By capturing relevant data on player statistics, such as performance, age, height, weight, and previous experience, we can ensure that our data model is comprehensive and informative. By incorporating statistical methods, such as regression analysis or clustering, we can identify the best five starting players for each position for the Chicago Bulls.

Furthermore, by ensuring that our data model is reproducible, we can easily update it with new data as it becomes available. This means that we can continually assess player performance and make informed decisions based on the latest data. In the future, we can incorporate data from the current NBA season or include data from other sources, such as scouting reports or expert opinions, to improve the accuracy of our model.

In conclusion, developing a reproducible data model that captures relevant player statistics and incorporates appropriate statistical methods is essential for identifying the best five starting players for each position for the Chicago Bulls. By creating a model that is flexible and scalable, we can update and modify it in the future to make informed decisions based on the latest data.

4. **Data modelling and results**Top of Form

This presented is an example of a linear regression model that predicts points (PTS) based on salary, field goals attempted (FGA), and free throws attempted (FTA) in the NBA. The model was created using R programming language, specifically the "lm" function which fits a linear regression model. The data used is player statistics and salaries for the most recent NBA season.

The output shows a summary of the linear regression model. The coefficients for FGA and FTA have p-values less than 2e-16, which means they are statistically significant and strongly associated with PTS. However, the coefficient for salary has a p-value of 0.288, which is not statistically significant and suggests that salary is not a good predictor of PTS.

The adjusted R-squared value of 0.9922 indicates that the model explains a high proportion of the variance in PTS, which means that the model is a good fit for the data. The F-statistic of 2.915e+04 with a p-value < 2.2e-16 indicates that the overall model is significant, which means that the model is better at predicting PTS than just using the mean of the data.

The residual standard error of 40.32 suggests that the model has a moderate level of error in predicting PTS, which means that there is some variability in the data that is not accounted for by the model. However, the normal Q-Q plot and residual vs. fitted plot do not show any major departures from normality or homoscedasticity assumptions, which means that the model assumptions are met.

In conclusion, this linear regression model can be used to predict PTS based on FGA and FTA, but not salary. The model is a good fit for the data and meets the assumptions of normality and homoscedasticity. Further analyses could be conducted to improve the model, such as including other variables that may be associated with PTS.

Assumption checking is a crucial step in validating the results of a statistical model. We check for two important assumptions of the linear regression model - homoscedasticity and normality.

Homoscedasticity refers to the assumption that the variance of the residuals of the model is constant across all levels of the predictor variables. The plot of residuals vs. fitted values is used to check for homoscedasticity. In the plot generated by the code, the residuals are plotted against the predicted values of the dependent variable (PTS), and we look for a pattern in the points. If the points are randomly scattered around the horizontal line, it indicates that the assumption of homoscedasticity holds. However, if there is a visible pattern in the plot, such as a funnel shape or a curve, it indicates that the model violates the homoscedasticity assumption.

Normality refers to the assumption that the residuals of the model follow a normal distribution. The normality plot (also known as a Q-Q plot) is used to check for normality. In the plot generated by the code, the residuals are plotted against the quantiles of a normal distribution. If the residuals follow a straight line in the plot, it indicates that they are normally distributed. However, if there is a visible deviation from a straight line, such as a curve or a bend, it indicates that the model violates the normality assumption.

By checking for these assumptions, we can ensure that the linear regression model is valid and that the results can be trusted. If the assumptions are violated, we may need to modify the model or use a different statistical method to obtain reliable results.

**5. Player recommendations**Top of Form

The process analyzes and recommends top players based on various criteria in a basketball dataset. The dataset includes player statistics and salaries for the 2019-2020 NBA season. The script includes four sections, each focusing on different criteria to select the top five players.

The first section focuses on selecting the top five players in each position based on cost-effectiveness. The script calculates the cost-effectiveness score for each player, which is the total number of points a player scores in a season divided by their salary. Then, it selects the top player for each position based on their cost-effectiveness score. The output includes the name of the player, their team, position, age, games played, points scored, salary, and cost-effectiveness score. The output also includes the total salary of the selected players.

The second section is similar to the first section, but it only considers players from the Chicago Bulls team. The script filters out players from other teams and selects the top player for each position based on their cost-effectiveness score. The output includes the same information as the first section, but only for the Chicago Bulls players.

The third section focuses on selecting the top five players in each position based on the total number of points they score in a season, regardless of their salary. The script selects the top player for each position based on their points scored and outputs the same information as the first two sections.

The fourth section selects the top five players irrespective of position based on their salaries. The script sorts the dataset by salary in descending order and selects the top five players. The output includes the name of the player, their team, position, age, games played, points scored, and salary.

Overall, we provide useful insights and recommendations for basketball team management to help them select the most effective and cost-efficient players. It also demonstrates how R can be used to analyze and visualize data to make data-driven decisions.

|  |
| --- |

The top 5 players based on the analysis are Stephan Curry, Russell Westbrook, Lebron James, Chris Paul, Kyle Lowry.Top of Form

**6.Summary**

|  |
| --- |

In this project, we aimed to help a general manager of a basketball team build a strong team within a set budget by analyzing NBA player data. Our analysis involved cleaning and processing the data and conducting exploratory data analysis to gain insights into the data. We visualized player salary distributions, player performance metrics, and the relationships between them.

To model the relationship between player performance metrics and salary, we used linear regression analysis. We found that performance metrics such as Points Per Game (PPG), Rebounds Per Game (RPG), and Assists Per Game (APG) were significant predictors of player salary. We then used this model to predict salaries for each player in the dataset, identifying overpaid and underpaid players.

Using optimization techniques, we selected a starting five of players with the highest combined performance metric score while staying within the budget constraint. We also recommended top players for each position based on their cost-effectiveness, as well as the top five players irrespective of position.

Our analysis has provided the general manager with valuable insights into how to build a competitive basketball team within a budget. We identified which players to target, which players are overpaid and underpaid, and which positions to focus on to maximize the team's performance.

However, our analysis is limited by the scope of the dataset and the assumptions made in our model. Other factors such as player popularity, team market size, or sponsorship deals may affect player salaries, which we did not consider. Additionally, our model assumes a linear relationship between performance metrics and salary, which may not always be the case in practice. Therefore, it is essential for the general manager to consider these limitations when making decisions based on our analysis.

Overall, our project provides a useful framework for analyzing and optimizing a basketball team's performance within a budget constraint. It is crucial to consider the limitations and potential biases in the data when making decisions based on our analysis.

**Reference**

**[1]** [**https://www.analyticsvidhya.com/blog/2021/10/everything-you-need-to-know-about-linear-regression/**](https://www.analyticsvidhya.com/blog/2021/10/everything-you-need-to-know-about-linear-regression/)

**[2] Baker, M. (2010). Nature Methods: Reproducibility crisis: Blame it on the antibodies. Nature Publishing Group.** [**https://doi.org/10.1038/nmeth.1549**](https://doi.org/10.1038/nmeth.1549)