PSBDA6001RM Research Methodology and Statistical Methods

Analyzing IPL 2022 through Data summary and Data Visualization

MANDAR TARMALE
ROLL NO - 19

Dataset Description

This dataset contains match-wise data of IPL matches 2022 (March26 - May29 2022), The complete data of group stage matches.

Attributes	Description
match_id	Match Number
date	Date of the match
venue	Name of Venue
team1	Playing team 1
team2	Playing team 2
stage	Stage of the tournament
toss_winner	Toss winning team
toss_decision	Decision of toss winning team
first_ings_score	First innings score
first_ings_wkts	First innings wickets
second_ings_score	Second innings score
second_ings_wkts	Second innings wickets
match_winner	Winning Team

Attributes	Description
won_by	Specify whether won by runs or wickets
margin	Winning margin
player_of_the_ match	Player of the match (Best Performance)
top_scorer	Top scoring Batter in the match
highscore	Highscore in the match (Highest Individual score from both teams).
best_bowling	Bowler who has best bowling figure in the match.(if two or more bowlers has the same bowling figure ,bowler who takes more wickets from less number of overs is selected).
best_bowling_figure	Best bowling Figure in the match.(if two bowlers has the same bowling figure ,bowler who takes more wickets from less number of overs is selected).

Methodology

This study is based on secondary data which has been collected from the Kaggle website. This dataset contains a 74 observations of matches with 20 variables of the IPL 2022 season. The whole IPL analysis is done on both R programming language and MS-Excel. In this study, Descriptive Statistics is used here to check the summary of IPL data. The descriptive analysis has been done on R programming language. In this analysis, I have checked the data summary for numeric and for categorical attributes. I have also find the frequency of their performance and score.

Data visualization is done here to visualize the data. To visualize the relationship between the data attributes, I used grammar of graphics in R. Furthermore, MS-Excel has also used to do graphical analysis, providing an meaningful insights within data.

Overall, this study combined the power of statistical analysis in R programming language and the versatility of MS-Excel to thoroughly explore the IPL 2022 season data, giving us informative insights of the players performance with their scores and how the matches went.

DESCRIPTIVE ANALYSIS IN R

```
> data = read.csv("C:/Users/admin/Downloads//IPL2022.csv")
> View(data)
> # Cheking the dimensions of the dataset
> dim(data)
[1] 74 20
> # There were 74 observations with 20 variables.
> #Descriptive analysis of numeric data
> summary(data[sapply(data, is.numeric)])
    match_id     first_ings_score first_ings_wkts  second_ings_score second_ings_wkts     margin        highscore
 Min.   : 1.00   Min.   : 68.0    Min.   : 0.000   Min.   : 72.0     Min.   : 1.000   Min.   : 2.00   Min.   : 28.00
 1st Qu.:19.25   1st Qu.:154.2    1st Qu.: 5.000   1st Qu.:142.8     1st Qu.: 4.000   1st Qu.: 5.25   1st Qu.: 57.00
 Median :37.50   Median :169.5    Median : 6.000   Median :160.0     Median : 6.000   Median : 8.00   Median : 68.00
 Mean   :37.50   Mean   :171.1    Mean   : 6.135   Mean   :158.5     Mean   : 6.176   Mean   :16.97   Mean   : 71.72
 3rd Qu.:55.75   3rd Qu.:192.8    3rd Qu.: 8.000   3rd Qu.:176.0     3rd Qu.: 8.000   3rd Qu.:18.00   3rd Qu.: 87.75
 Max.   :74.00   Max.   :222.0    Max.   :10.000   Max.   :211.0     Max.   :10.000   Max.   :91.00   Max.   :140.00
```

1) (match_id): The data includes matches with IDs ranging from 1 to 74.
2) (first_ings_score): The scores during the first innings range from a minimum of 68 to a maximum of 222, with a mean(average) score of 171.1. The middle 50% of scores fall within the range of approximately 154.2 to 192.8.
3) (first_ings_wkts): The number of wickets taken during the first innings ranges from 0 to 10, with a mean of about 6.135. The interquartile range (IQR) for wickets is from 5 to 8 wickets.
4) (second_ings_score):Scores during the second innings vary from a minimum of 72 to a maximum of 211,with an average score of 158.5. The IQR for second innings scores falls between roughly 142.8 and 176.0.
5) (second_ings_wkts):The number of wickets taken during the second innings ranges from 1 to 10, with a mean of approximately 6.176. The IQR for second innings wickets is from 4 to 8.
6) (margin):The margin of victory or defeat in the matches varies from a minimum of 2 to a maximum of 91, with an average margin of approximately 16.97.
7) (highscore):The highest score achieved in the matches ranges from a minimum of 28 to a maximum of 140, with a mean highscore of approximately 71.72.

```
> #Descriptive analysis of categorical data
> library(dplyr)

Attaching package: 'dplyr'

The following objects are masked from 'package:stats':

    filter, lag

The following objects are masked from 'package:base':

    intersect, setdiff, setequal, union

Warning message:
package 'dplyr' was built under R version 4.2.3
> #Getting the frequency of top 5 player of the match
> data %>% group_by(player_of_the_match) %>% summarize(count = n()) %>%  arrange(desc(count)) %>%  head(5)
# A tibble: 5 × 2
  player_of_the_match count
  <chr>               <int>
1 Kuldeep Yadav           4
2 Jos Buttler             3
3 Avesh Khan              2
4 David Miller            2
5 Dinesh Karthik          2
```

To check the descriptive summary of categorical data, I have used 'dplyr' library . By using this library, I have find the frequency of top 5 player of the match in the season in which Kuldeep Yadav is being 4 times player of the match in the IPL 2022 season followed by Jos Buttler who is 3 times player of the match.

```
> #Geting the frequency of top 5 top_scorer of the season
> data %>% group_by(top_scorer) %>% summarize(count = n()) %>%  arrange(desc(count)) %>%  head(5)
# A tibble: 5 × 2
  top_scorer       count
  <chr>            <int>
1 Jos Buttler          7
2 Quinton de Kock      5
3 KL Rahul             4
4 Liam Livingstone     4
5 Shubman Gill         4
>
```

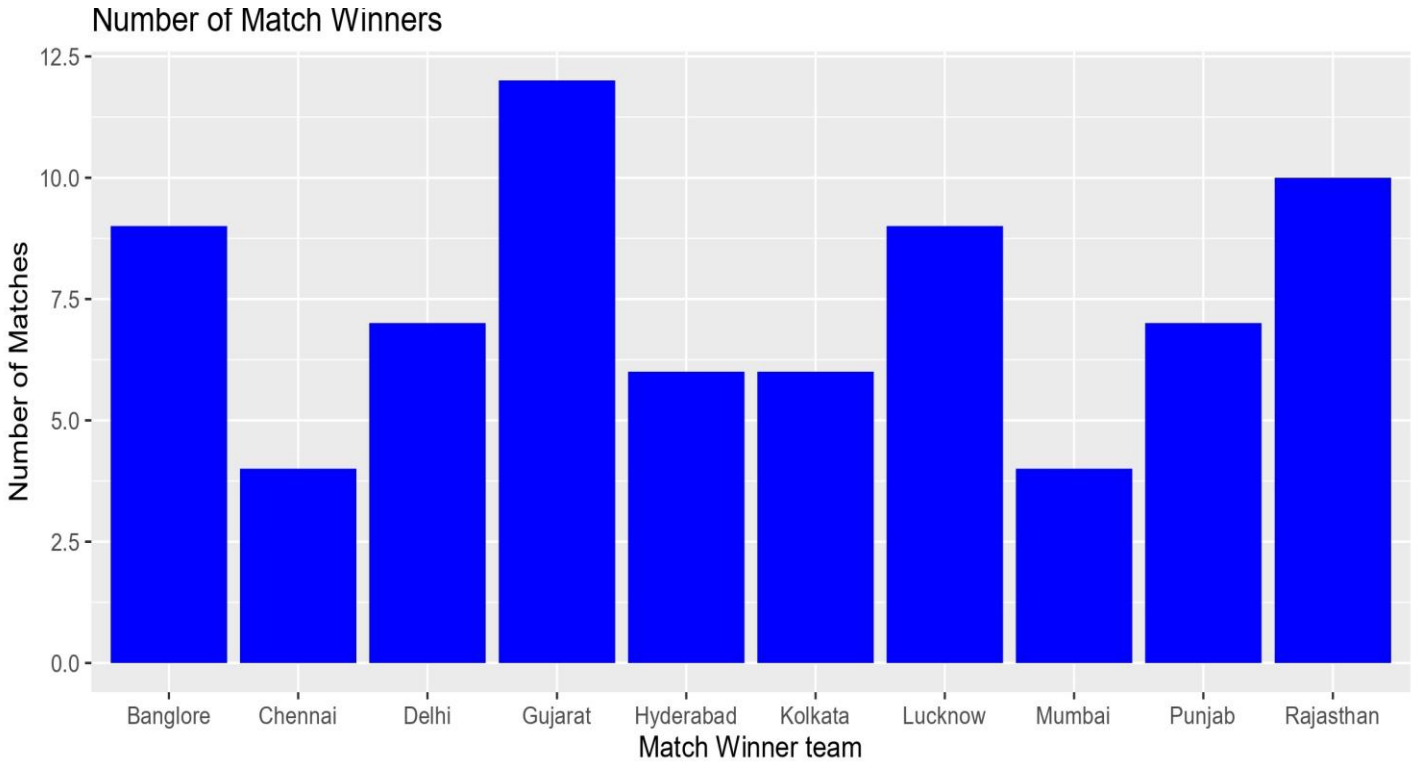
In the above, I have checked the frequency of top_scorer (The batsman who scored most runs) using dplyr library. So, I got Jos Buttler as top batter in the season who has scored highest runs for 7 times followed by Quinton de Kock 5 times. KL rahul, Liam Livingstone and Shubman Gill for 4 times.

GRAPHICAL ANALYSIS

BAR PLOT IN R

The command to draw bar plot in R is given as follows:

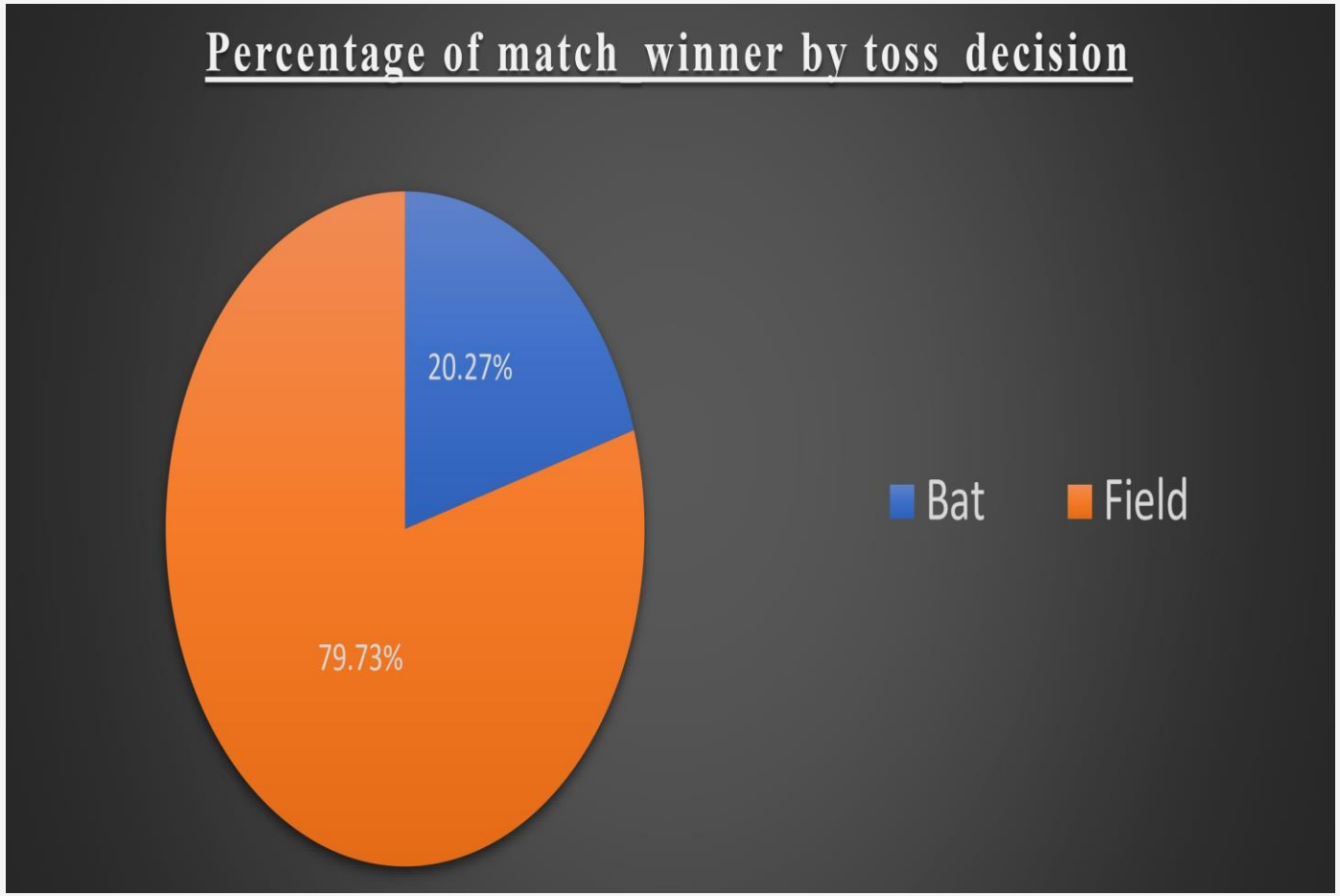
```
> #Creating Bar plot to see number of winner in IPL 2022
> library(ggplot2)
> ggplot(data,aes(match_winner)) + geom_bar(fill="blue") + labs(title="Number of Match Winner")
>
```



The above plot tells us that Gujrat team leads the IPL season 2022 followed by Rajasthan and Lucknow teams.So, It is an achievement for Gujrat as a new team in IPL.

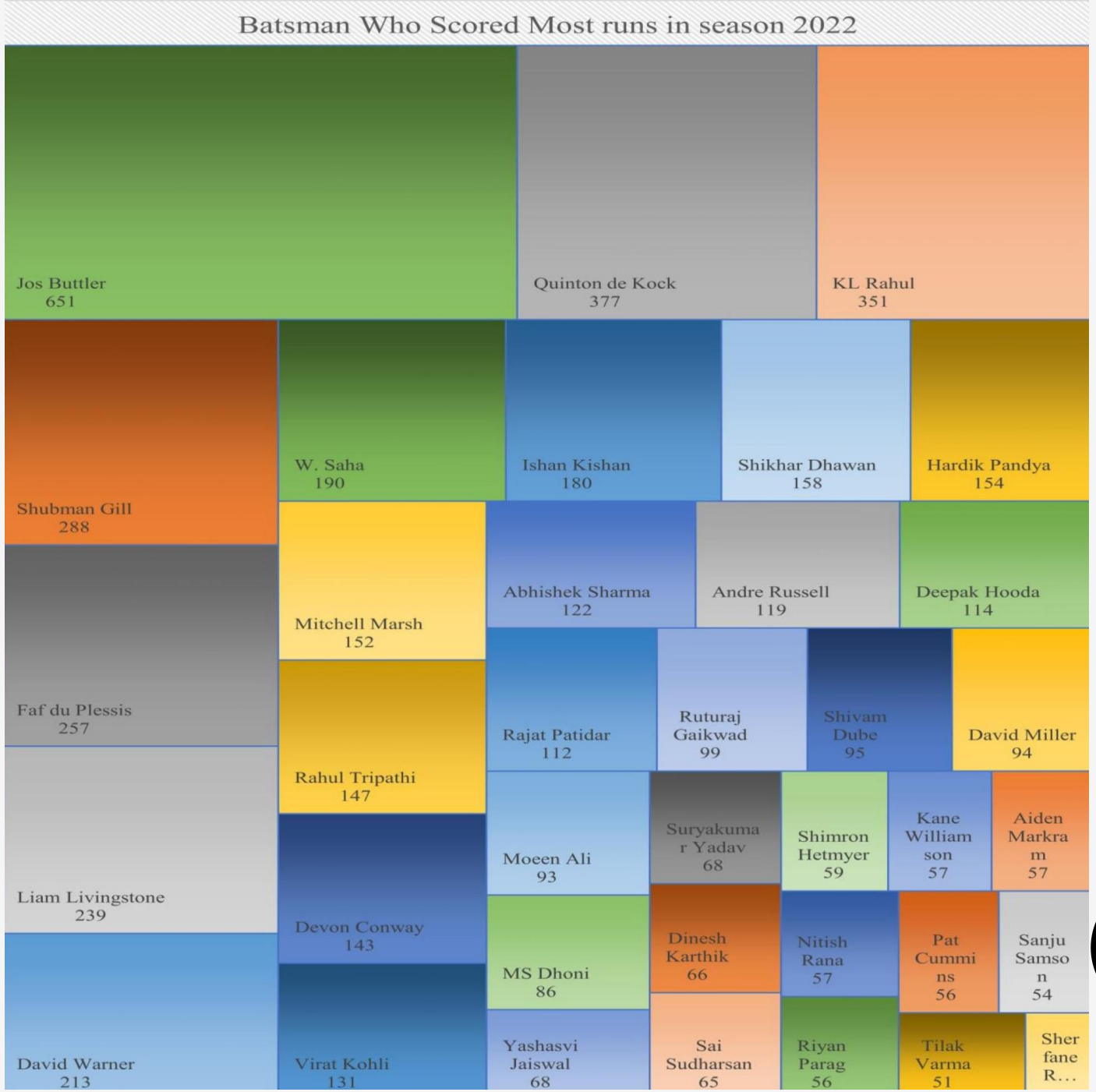
PIE CHART IN MS-EXCEL

To check percentage of match winner by toss decision, I have used pie chart in excel.



The pie chart states that the most teams prefer fielding over batting after winning toss. We can also say that most teams have won matches while chasing the target(batting second) and least teams have won while defending the target(batting first).

TREEMAP IN MS-EXCEL



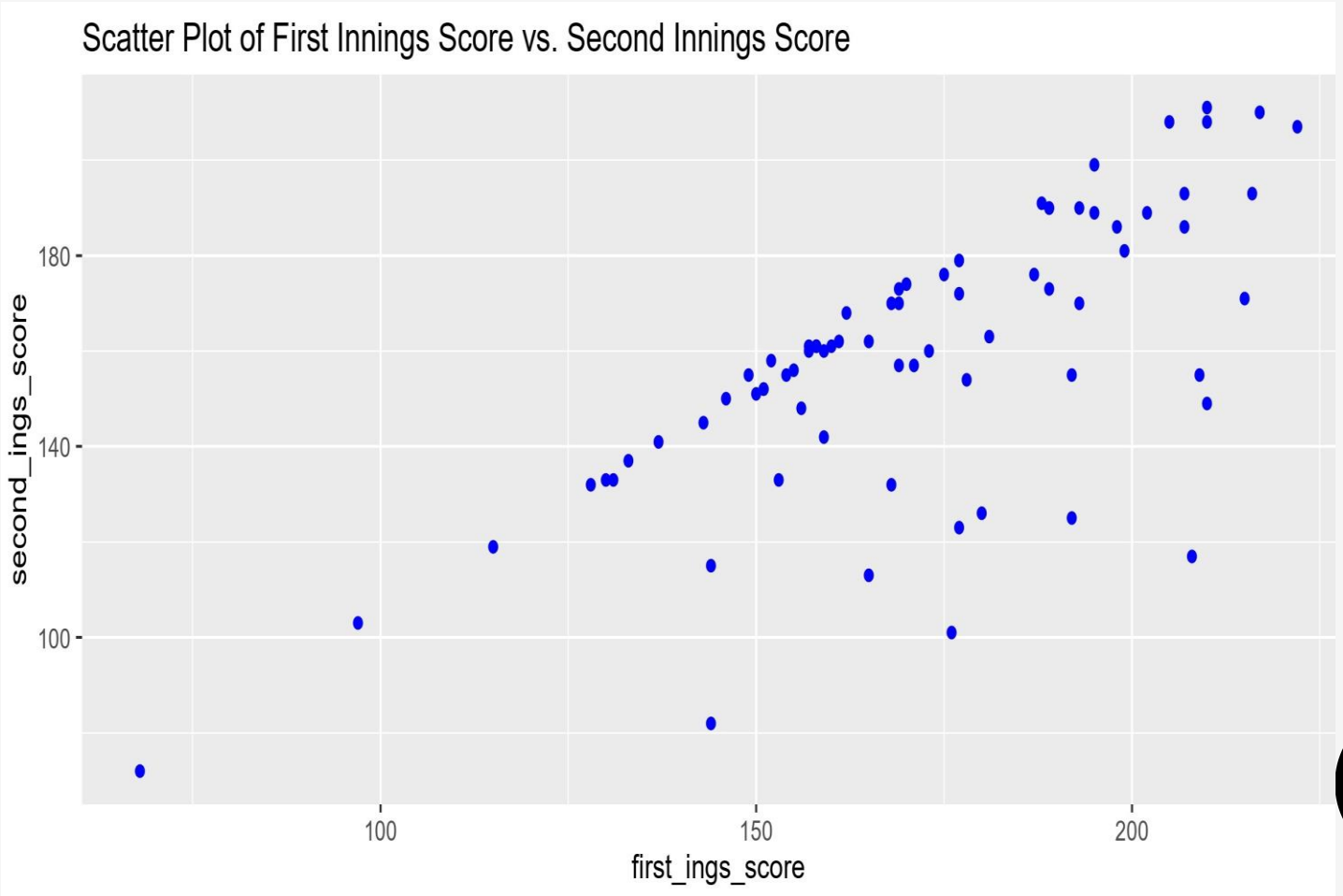
TREEMAP IN MS-EXCEL

I have make used of excel to draw above Treemap. The motive of plotting Treemap is to check who had scored most runs in the IPL season 2022.
The Treemap shows us that Jos Buttler being the top scorer in the season followed by Quinton de Kock and KL Rahul on the second and third position respectively.

SCATTER PLOT IN R

The Scatter plot had been plot by using below command in R;

```
> data = read.csv(file_path)
> # Create a scatter plot for both innings scores
> library(ggplot2)
> ggplot(data,aes(x=first_ings_score,y=second_ings_score)) +geom_point(color="blue")
>
```



The scatter plot displays points, each representing a cricket match. The x-axis represents the "First Innings Score," and the y-axis represents the "Second Innings Score“.

The majority of points are clustered in the central region of the plot. A diagonal line from the bottom-left corner to the top-right corner represents the line of equality, where the first innings score is equal to the second innings score. Points along the diagonal line indicate matches where both innings had similar scores.

Points spread above the line of equality indicate matches where the second innings score was higher than the first innings score and vice-versa. There are a few outliers, represented by points that are far from the main data. These outliers may represent extremely high-scoring or low-scoring second innings compared to the first innings.

Dataset Link

IPL DATASET LINK:

https://www.kaggle.com/datasets/aravindas01/ipl-2022dataset

IPL ANALYIS IN R :

https://drive.google.com/file/d/1JP61o8xTxWn7c071S3kBRn9d9rLJaE92/view?usp=drivesdk

THANK YOU!