WORKING WITH DATA- Assignment 1

This document is prepared for Working with Data Assignment. Below is the list of contents involved in this assignment.

Name: Deepti Balasubramaniam

Student ID: D20123887

CourseID: TU059(FT Data Science)

Table of Contents:

INTRODUCTION:

Indian Premier League (IPL) is a professional Twenty20 cricket league in India contested during March or April and May of every year by eight teams representing eight different cities in India. The league was founded by the Board of Control for Cricket in India (BCCI) in 2008.

For IPL, players are selected on auction basis. In that way a franchise wanted to chose the best players from the best performed team and therefore, they wanted a help of Data Scientist who can help them in exploring the performance of the team and players in the previous matches. With the questions like which team won highest number of times, which player bagged man of the match multiple times, how is the team in hitting boundaries, is Toss winning really helping to win, win percentage etc. franchise team can select the players based on analysis given.

Analysis of IPL dataset from Kaggle.

Link: https://www.kaggle.com/ramjidoolla/ipl-data-set

Have also uploaded the required datasets in github link. **And directly added the raw data link in the R script.**

https://github.com/DeeptiBSV/WWD-Dataset

matches.csv, **deliveries.csv**, **strikerate.csv**,**home_away_wins.csv**

Different sets of datasets available are:

Matches, deliveries, home, most_run_average_strike_rate, team_wise_home_away_wins, players.

WORKING WITH DATA- Assignment 1

I have merged the dataset Matches and home_away_win as one dataset, then deliveries, most_run_average_average_strike_rate as one.

With these merged datasets we will perform cleaning, exploration and plotting relevant graphs for easy visualization.

PACKAGE INSTALLATION:

Install the required packages to run the R script like tidyverse, dplyr, ggplot2, readxl, sqldf, data.table, treemap

LOADING AND MERGING:

Load the CSV and XLS datasets and merge the datasets using **left_join** with columns having same names and similar set of rows

CLEANING THE DATASET

Cleaning of dataset with start with identifying the NA values

- In this dataset we have few of the values as 0 which we can replace it as **na.omit**

PART II DATA EXPLORATION

Data Exploration is required to get an overview of the data.

1. *Find the total number of seasons available in the dataset and in which season highest number of matches were played.*
    a. Total No. of Season-12
    b. In IPL-2013 highest number of matches were played with the match count of 76
2. *Which team has won the match maximum number of times throughout the season and how many times?*
    a. **Kolkata Knight Riders** won 5 times throughout the season
3. *Which team scored maximum runs and minimum number of wickets*
    a. **Mumbai Indians** won by maximum run of 146
    b. **Kolkata Knight Riders, Chennai Super Kings** and **sunrise Hyderabad won** by minimum wicket of 1
4. *Which player has been awarded player of the match maximum number of times?*
    a. **Chris Gayle** has been awarded player of the match 21 times, which is the maximum number of times in the entire season
5. *Top 5 batsman in the entire IPL season*
    a. V.Kohli, SK.Raina ,RG.Sharma, DA.Warner, S.Dhawan are the top batsman in the entire season
6. Which team has won maximum number of times, plot the graph to show winning frequency greater than 10.
    a. **Mumbai Indians and Chennai Superkings** tops the list. Please refer Fig1 for graph
7. *Histogram of Win_by_Runs*

   a. Histogram gives the idea on how the values are distributed and plays a major role in data exploration and statistical analysis specially for inferential stats-Refer Appendix section Fig2

8. *Find total/maximum number of Matches played in different cities and create a bar plot for the same for easy visualization*
   a. Refer Fig 3 for graph
9. *How is win by runs is distributed for a team based on the decision to first batting or fielding. Explain it with a graph*
   a. Most of the teams win when they chose to field first. **Mumbai Indians** scored maximum runs when they chose to field. **Chennai Super Kings** scored maximum when they chose to bat.
      Refer Fig4 for graph

PART III -Exploring Dataset using R with subset of Dataset

1. *Which team got highest counts of 6s and 4s. Compare with the graphs for 6s and 4s in each team.*
   In the dataset we have a separate column for 6s and 4s from which we can count the total values of 6s and 4s and by using **melt** (reshape package) function we can derive a new variable for 6s and 4s in one column and its value in another column.
   a. Almost every team have high 4s than 6s
   b. **Mumbai Indians** have highest total count of 4s and **Rajasthan Royals** have highest total count of 6s-Please refer the Fig 5 for comparison with graph
2. *Is Toss a Decision Factor, does winning the toss actually increases the chances of winning the game.*

By winning the toss, team has provision to choose for first batting or fielding. Therefore, this gives a confidence and motivation to team to win when they win the toss.

In this dataset we have a column called toss_winner and a winner, from this we can derive a new variable called toss as win or loss by matching the teams in these 2 columns. With the win and loss count we can come to a conclusion below.

   a. Yes, winning the toss actually increases the chances of winning the game. Please refer Fig 6 for graph.

3. *How many times one team has won against the other teams. Consider Chennai Super kings (CSK) and Mumbai Indians (MI). Plot a treemap for easy visualization.*
   CSK and MI are most popular teams. Therefore, by knowing their winning strengths against other teams will help the team to evaluate their strengths and weakness. In This dataset we have a column called winner, team1and team2. By merging all these columns will be able to derive a new variable of win count against the team.
   a. **CSK** has won 15 times against Royal Challengers Bangalore and only 4 times
   b. **MI** has won 1 time against Delhi Capitals and Rising Pune Supergiants and won 19 times against kolkata knight Riders. Refer Fig 7 and 8 for graph

4. Derive a new variable called Win Percentage of each team. Identify the team having highest and lowest win percentage

As we have been exploring all the through win count but win percentage of each team will give an overall picture of which team have actually performed well throughout all the season.

We have columns like Team1, Team2 and Winner. By getting the count of each columns and merging it we can calculate Total Number of matches played and Total number of matches won, multiplying by 100 will give the win percentage

a. **Delhi Capitals** have highest win percentage and **Pune warriors** have lowest win percentage. Refer Fig 9 for graph

CONCLUSION:

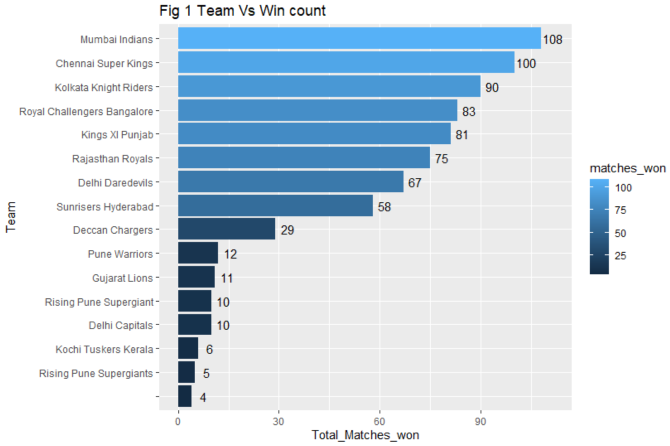
With the above explorations, graphs and hypothetical questions. We can give some conclusions like:

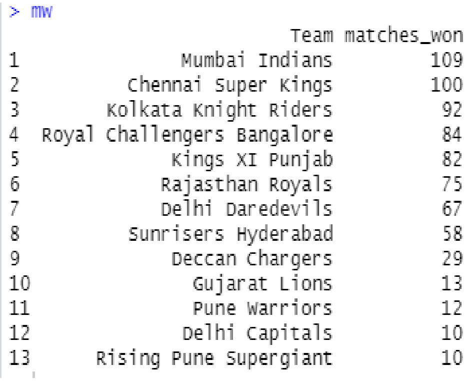
- Toss win is one of the decision factors in winning the match
- Team chances of winning are high when they chose to field first.
- Though Kolkata Knight Riders have won 5 times, win percentage of Delhi Capitals, Rising Pune Supergiants and Chennai Superkings are higher

APPENDIX:

Plots/GRAPHS:

Fig1.Team Vs Win





```
> mw
                          Team matches_won
1                Mumbai Indians         109
2           Chennai Super Kings         100
3         Kolkata Knight Riders          92
4   Royal Challengers Bangalore          84
5               Kings XI Punjab          82
6             Rajasthan Royals          75
7              Delhi Daredevils          67
8           Sunrisers Hyderabad          58
9               Deccan Chargers          29
10                Gujarat Lions          13
11                Pune Warriors          12
12               Delhi Capitals          10
13       Rising Pune Supergiant          10
```

Fig2 Histogram of Win_by_Runs

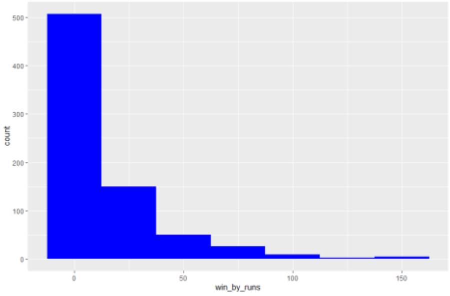


Fig 3 Total Matches in Each City

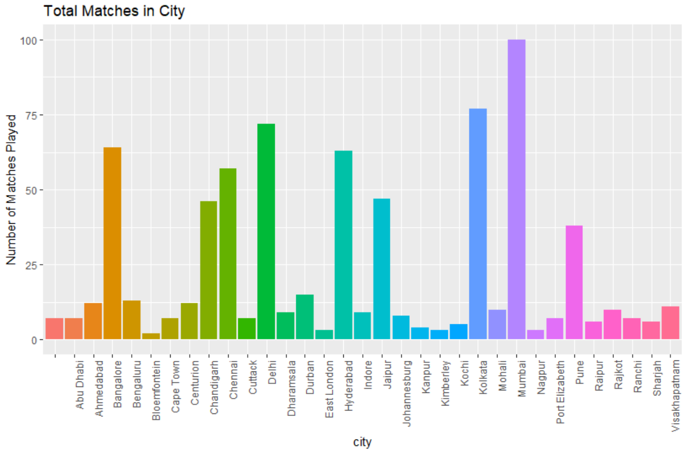


Fig 4 Total Runs by batting Vs Fielding

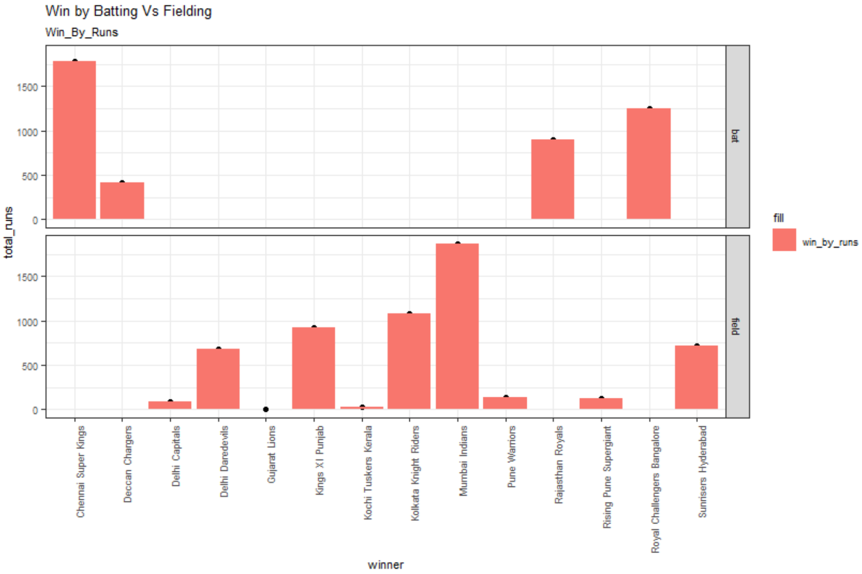


Fig5 Compare 6s and 4s of each team

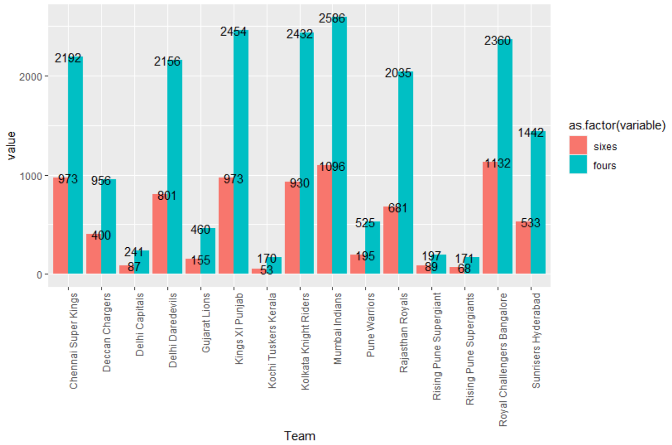


Fig6 Is Toss a decision factor.

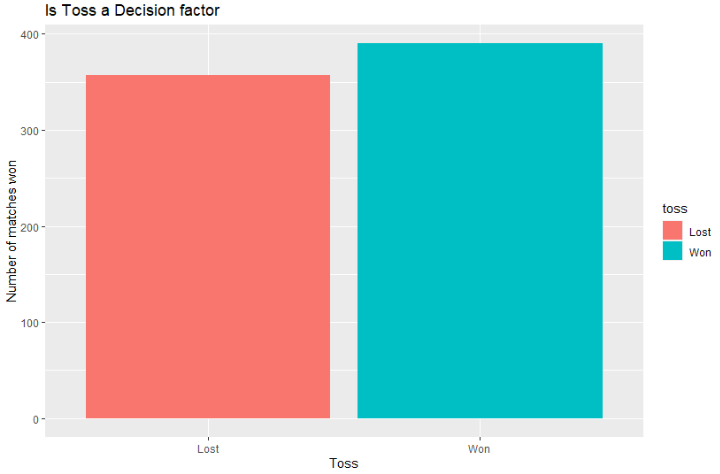


Fig 7 and Fig 8, Treemap of CSK and MI win Treemap against other teams.

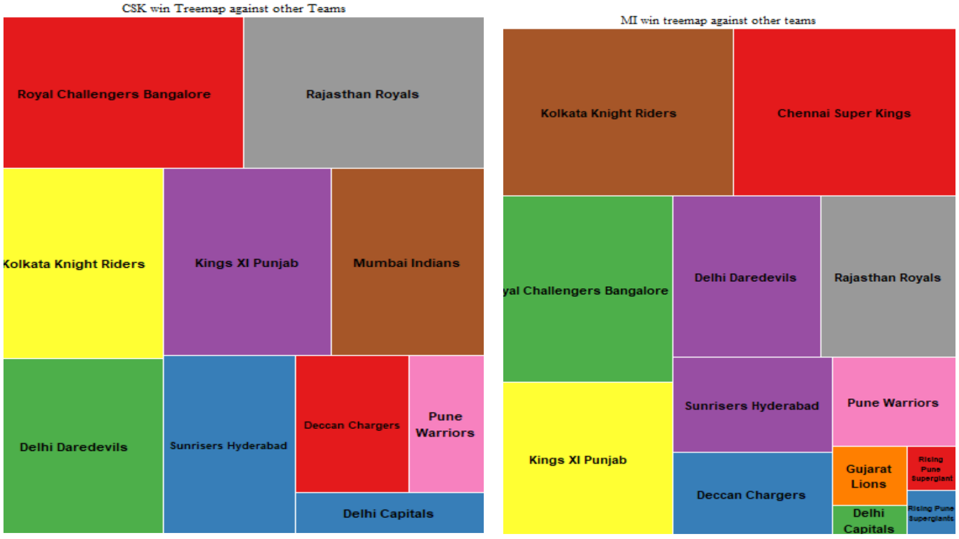
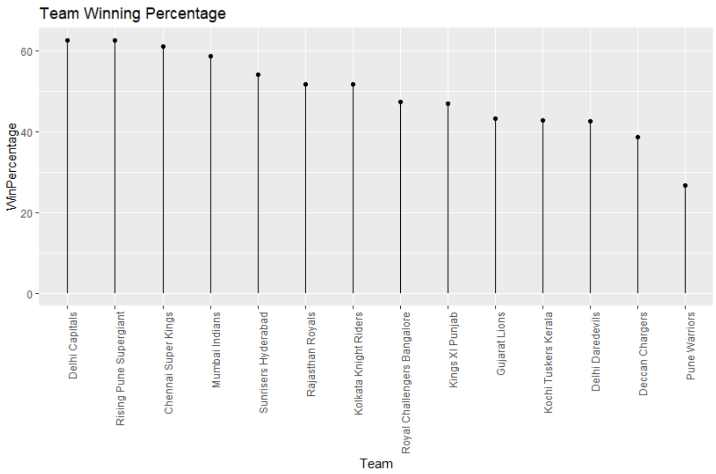
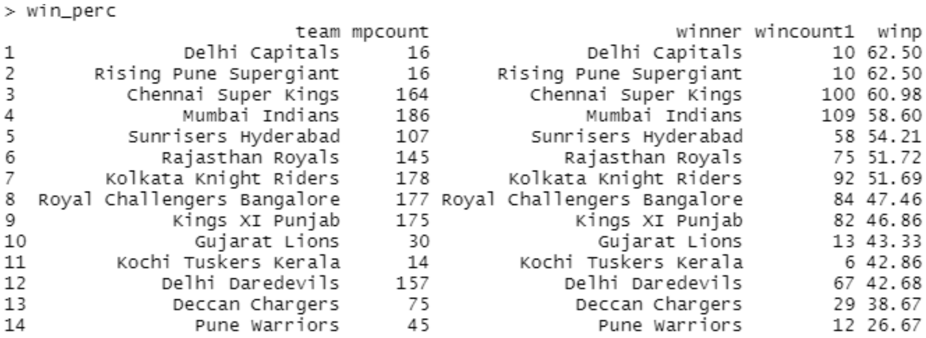


Fig 9 Team winning percentage.





```
> win_perc
                          team mpcount                      winner wincount1  winp
1               Delhi Capitals      16              Delhi Capitals        10 62.50
2       Rising Pune Supergiant      16      Rising Pune Supergiant        10 62.50
3          Chennai Super Kings     164         Chennai Super Kings       100 60.98
4               Mumbai Indians     186              Mumbai Indians       109 58.60
5          Sunrisers Hyderabad     107         Sunrisers Hyderabad        58 54.21
6             Rajasthan Royals     145            Rajasthan Royals        75 51.72
7        Kolkata Knight Riders     178       Kolkata Knight Riders        92 51.69
8  Royal Challengers Bangalore     177 Royal Challengers Bangalore        84 47.46
9              Kings XI Punjab     175             Kings XI Punjab        82 46.86
10               Gujarat Lions      30               Gujarat Lions        13 43.33
11        Kochi Tuskers Kerala      14        Kochi Tuskers Kerala         6 42.86
12            Delhi Daredevils     157            Delhi Daredevils        67 42.68
13             Deccan Chargers      75             Deccan Chargers        29 38.67
14               Pune Warriors      45               Pune Warriors        12 26.67
```

References:

Amittian/Exploratory-Data-Analysis-R. (n.d.). GitHub. Retrieved November 20, 2020, from https://github.com/amittian/Exploratory-Data-Analysis-R

IPL _Data_Set. (n.d.). Retrieved November 20, 2020, from https://kaggle.com/ramjidoolla/ipl-data-set

Holtz, Y. (n.d.). *Customize your R treemap*. Retrieved November 20, 2020, from https://www.r-graph-gallery.com/236-custom-your-treemap.html