

Data Visualization-Assignment 2

This document is prepared for Data Visualization Assignment 2. Below is the list of contents involved in this assignment.

Name: Deepti Balasubramaniam

Student ID: D20123887

CourseID: TU059(FT Data Science)

Data Visualization of IPL matches using R

Table of Contents:

- PART I
 - [Introduction](#)
 - [Installing packages using R](#)
 - [Loading and Dataset Using R](#)
 - [Cleaning the Dataset](#)
- PART II
 - [Exploring Dataset with Visualization](#)
- [CONCLUSION](#)
- [APPENDIX](#)

INTRODUCTION:

Indian Premier League (IPL) is a professional Twenty20 cricket league in India contested during March or April and May of every year by fifteen teams representing 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 can choose the best players from the best performed team. Franchise from **Mumbai** city in India are new to this field and they do not have much knowledge on what factors they need to choose the team therefore they wanted to analyze the overall performance of the teams in last 12 years, for which they wanted a help of Visual Designer who can help them in exploring the performance of the team previous matches. The major feature to focus on would be to analyze:

- Which team won highest number of times
- How is the team in hitting boundaries (6's and 4's)
- What is the win percentage overall for each team
- How Mumbai team has performed against other teams
- How is the performance of the team in each season overall etc.

As a Visual Designer, I will be able to benefit the Franchise company by providing useful insights.

Data Visualization-Assignment 2

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 code**, therefore just by giving run command will run the code overall, **no need to download any dataset**.

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

Different sets of datasets available are:

Matches and deliveries. We are going to analyze on these datasets.

- Matches dataset covers the information about season, city, Teams, winner team, toss_winner, win_by_runs, win_by_wicket, Umpire
- Deliveris dataset covers information about Players, boundaries.

PACKAGE INSTALLATION:

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

CLEANING THE DATASET

Cleaning of dataset with start with identifying the NA values

- In this dataset we have do not have much NA values and 0 values are indication of 0 runs. We do have few irrelevant columns, which we can drop it.
- In matches dataset dl_applied and umpire columns can be dropped
- In deliveries player_dismissed, dismissal_kind, fielder columns can be dropped

PART II DATA EXPLORATION USING VISUALIZATION

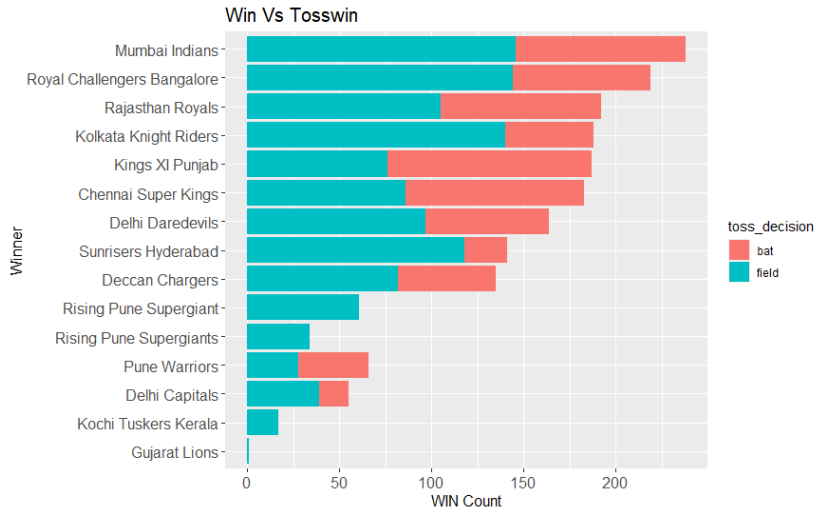
Data Exploration using Visualization is required to get an overview of the data.

Insight 1-Iteration 1:

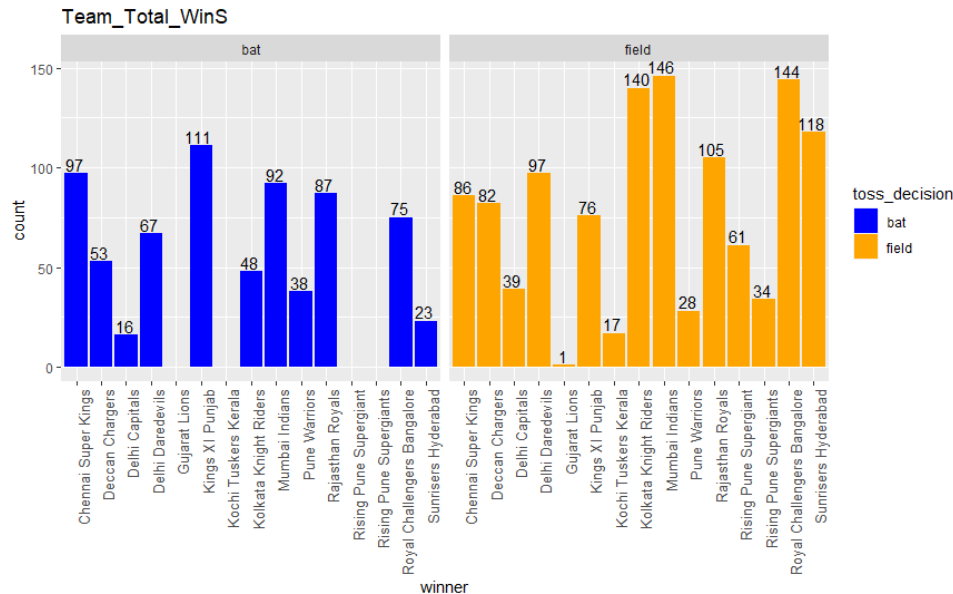
1. *What is maximum win by runs is for a team based on the decision to first batting or fielding. Explain it with a graph.*
 - a. Through toss, team can decide either to bat or field first, therefore when team wins the toss, they get to choose their decision.

Insight1 Iteration1: Plot

Data Visualization-Assignment 2



Insight1 Iteration 2: facet



Insight 1: Conclusion

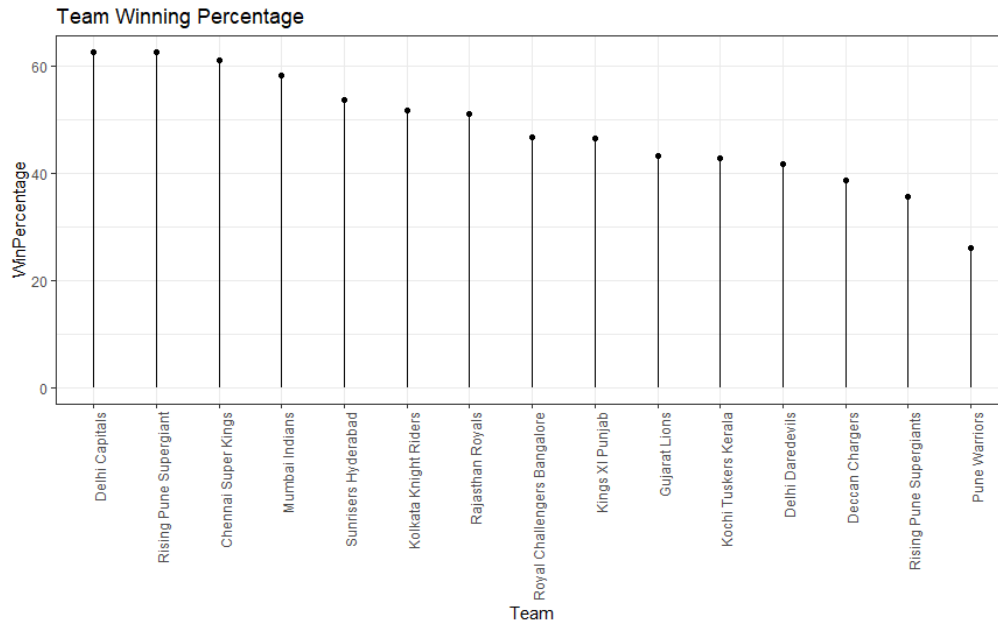
Most of the teams win when they chose to field first. **Mumbai Indians** scored maximum runs when they chose to field. **Chennai Super Kings** scored maximum when they chose to bat.

Insight 2 Iteration1:

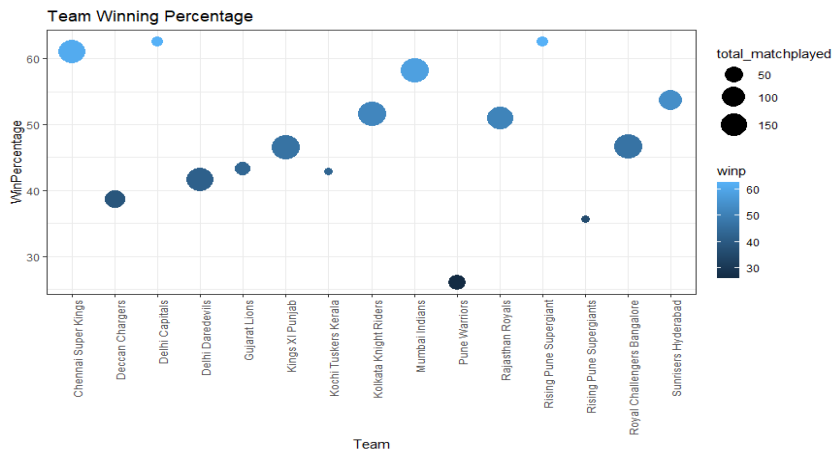
2. What is the win percentage of each team and which team have the highest win percentage?
 - a. In our matches dataset we have two column team 1 and team 2 we can **merge** these teams in to one team count the team values to get total match played, with another column called winner team we can add **new winp (variable win percentage)**.

Data Visualization-Assignment 2

Insight 2 Iteration 1: Plot



Insight2 Iteration 2: Plot



Insight 2 Iteration 2: Conclusion

From the plot we can see that **Delhi capitals** and **Rising Pune Supergiant** have high win percentage

Insight 3 Iteration 1:

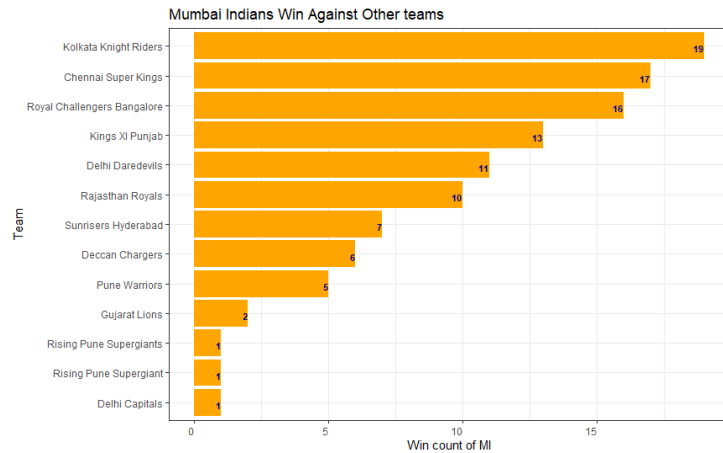
3. As the Franchise is from Mumbai city and so they are interested in Mumbai's performance compared to another team. Basically, they are more interested to know how many times Mumbai has won against the other team.

Data Visualization-Assignment 2

- a. We have In This dataset we have a column called winner, team1 and team2. By merging all these columns will be able to derive a new variable of win count against the team.

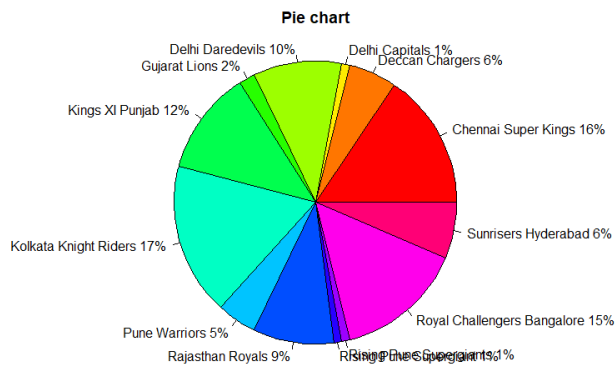
Insight 3 Iteration 1: Plot

Mumbai's jersey color is orange therefore it would be nice to add orange color to the bar plot.



Insight 3 Iteration 2: plot

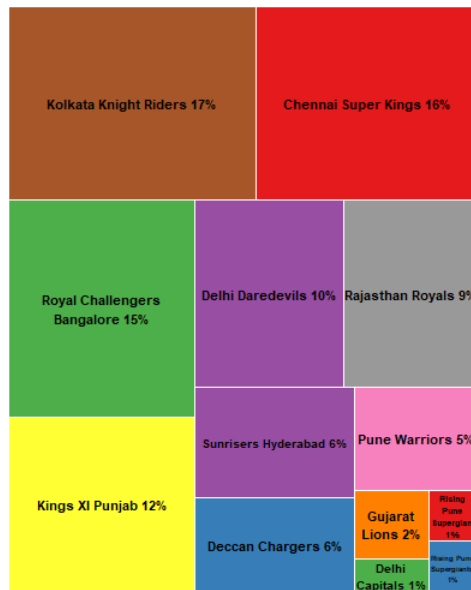
In pie chart we can see that Team names are overlapping and so we can try the same with treemap for much easier visualization



Insight 3 Iteration 3: plot

Data Visualization-Assignment 2

MI win treemap against other teams



Insight 3 Iteration 3: Conclusion

From the above graphs we can infer that Mumbai Indians has won against Kolkata Knight Riders and Chennai Super Kings the most.

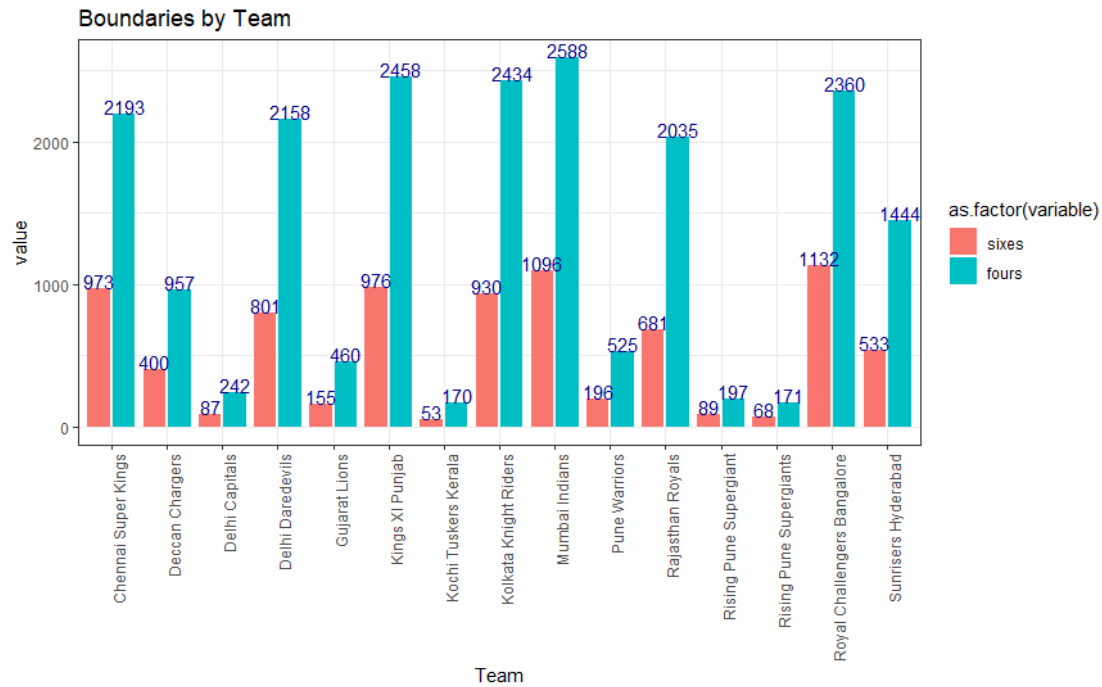
Insight 4 Iteration 1:

4. Which team got highest counts of 6s and 4s. Compare with the graphs for 6s and 4s in each team.

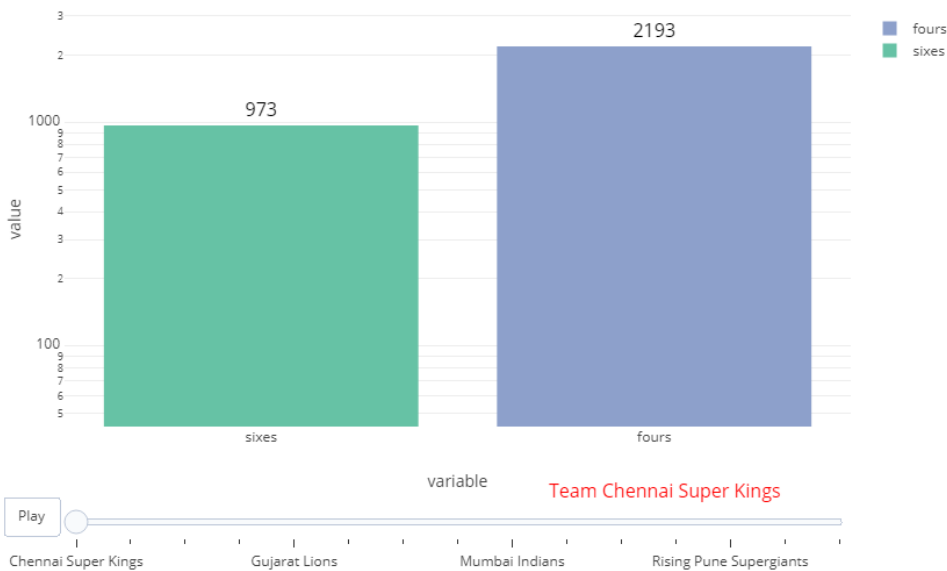
In the dataset we have a separate column for 6s and 4s from **deliveries** dataset from which we can count the total values of 6s and 4s and by using **melt** (reshape package) function we can derive a new variable for 6s and 4s in one column and its value in another column.

Data Visualization-Assignment 2

Insight 4 Iteration 1: Plot



Insight 4 Iteration 2: plot with Animation



Insight 4 Iteration 2: Conclusion

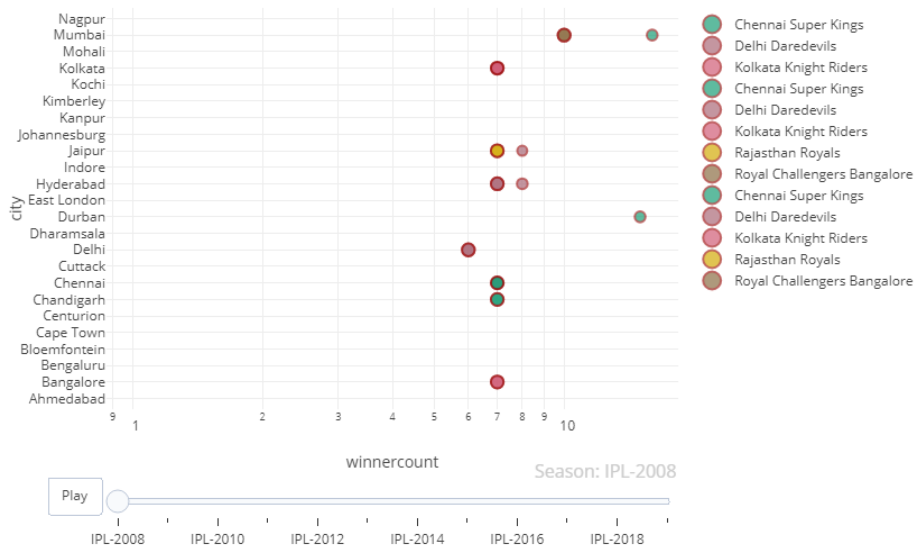
From the Graphs we can infer that **Mumbai Indians** has hit highest 4's boundaries and **Royal Challengers Bangalore** has hit highest 6's boundaries.

Data Visualization-Assignment 2

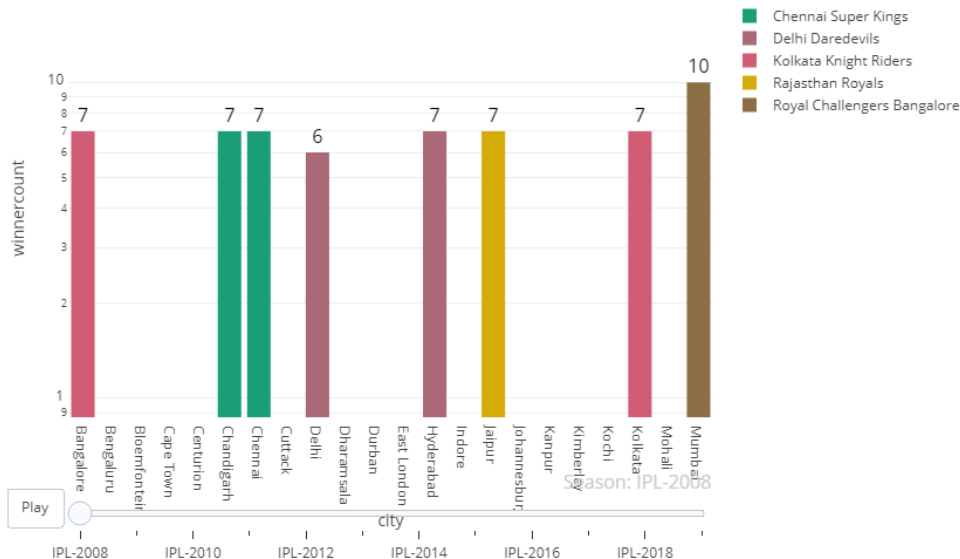
Insight 5 Iteration 1:

5. Franchise wants to view the overall performance of the teams throughout the seasons in different cities, therefore we can try animation on team's performance in each season.

Insight 5 Iteration 1: Plot



Insight 5 Iteration 2: Plot Animation with Bar



Data Visualization-Assignment 2

CONCLUSION:

With the above explorations and graphs, we can give some conclusions like:

- Team chances of winning are high when they chose to field first.
- Win percentage is high for Delhi Capitals and Rising Pune Sergiant
- Mumbai Indians have played well against Kolkata Knight Riders and Chennai Super kings
- Mumbai Indians performed well in hitting the boundaries

Therefore, from above few insights we infer that Mumbai Indians are one of the strong players.

APPENDIX:

R code for Iterations: Have attached the R code along with this document

References:

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>

Intro to Animations | R | Plotly. (n.d.). Retrieved December 20, 2020, from <https://plotly.com/r/animations/>