**Submitted by:**

*Samarthraj Balachandra Hegde(sh4636)*

*Shruthi Nemu Tandel(snt288)*

*Rakavee Anandan(ra2635)*

Analyzing Indian Premier League Cricket Dataset

Abstract:

Cricket is a very popular sport in Commonwealth countries. There are 11 players in each team which includes batsmen, bowlers and all-rounders. IPL (Indian Premier League) started in 2008 involving players from all the nations. So, we thought of analyzing the IPL cricket dataset, find some patterns and make some decisions which would be helpful for a team.

In this project, we have data about matches and deliveries in the form of csv files. Matches file has the information of all the cricket matches played from 2008 to 2017 in IPL. Deliveries file has information of each delivery (each ball bowled) for every cricket match. Each team plays two matches in a season. One at home ground and other at the opponent’s home ground. As the support from home team fans can affect the team’s win, we would like to analyze whether a team played at home or away has any effect over their wins. Also, as the toss plays an important role in game planning we will also analyze how crucial is to win a toss to win the game. Along with these we will analyze more features from the dataset to assist the team captain, coach to rebuild their strategy to increase chances of winning.

**CONTENTS**

**1 Introduction**  3

**2 Methodology** 4

2.1 Data Capture 4

2.2 Tech Stack 5

2.3 Data Cleaning 5

2.4 Data Analysis 6

**3 Results and Discussion** 7

3.1 Results obtained with tableau and matplotlib 7

3.1.1 Impact of toss on the outcome of the match 73.1.2 Analyzing which stadium is most suitable for batting first 8

3.1.3 Analyzing which stadium is most suitable for fielding first 9

3.1.4 Home Ground Wins Percentage 10

3.1.5 No. of wins by team and season in each city 10

3.1.6 Top 5 batsmen over the seasons 11

3.1.7 Top Bowlers over the seasons 12

3.1.8 Wicket Distribution in different overs 13

3.1.9 Runs Distribution in different overs 14

3.2 Other basic analysis 15

4 Code and References 15 4.1 Code 15 4.2 References 15

**Introduction**

The Indian Premier League (IPL), officially Vivo Indian Premier League for sponsorship reasons, is a professional Twenty20 cricket league in India contested during April and May of every year by teams representing Indian cities and some states. The league was founded by the Board of Control for Cricket in India (BCCI) in 2008, and is regarded as the brainchild of Lalit Modi, the founder and former commissioner of the league. IPL has an exclusive window in ICC Future Tours Program.

The IPL is the most-attended cricket league in the world and in 2014 ranked sixth by average attendance among all sports leagues. In 2010, the IPL became the first sporting event in the world to be broadcast live on YouTube. The brand value of IPL in 2017 was US$5.3 billion, according to Duff & Phelps. According to BCCI, the 2015 IPL season contributed ₹11.5 billion (US$182 million) to the GDP of the Indian economy.

There have been ten seasons of the IPL tournament. The current IPL title holders are the Mumbai Indians, who won the 2017 season.



**2 Methodology**

2.1 Data Capture

Data Set List:

1. Matches
2. Deliveries

IPL has published the open source data to the people to explore further and analyze their data. The data in Matches includes the below fields:

1. Season
2. Match date
3. Toss winner
4. Toss decision
5. Winner
6. Win by runs
7. Win by wickets
8. Player of the match
9. Venue

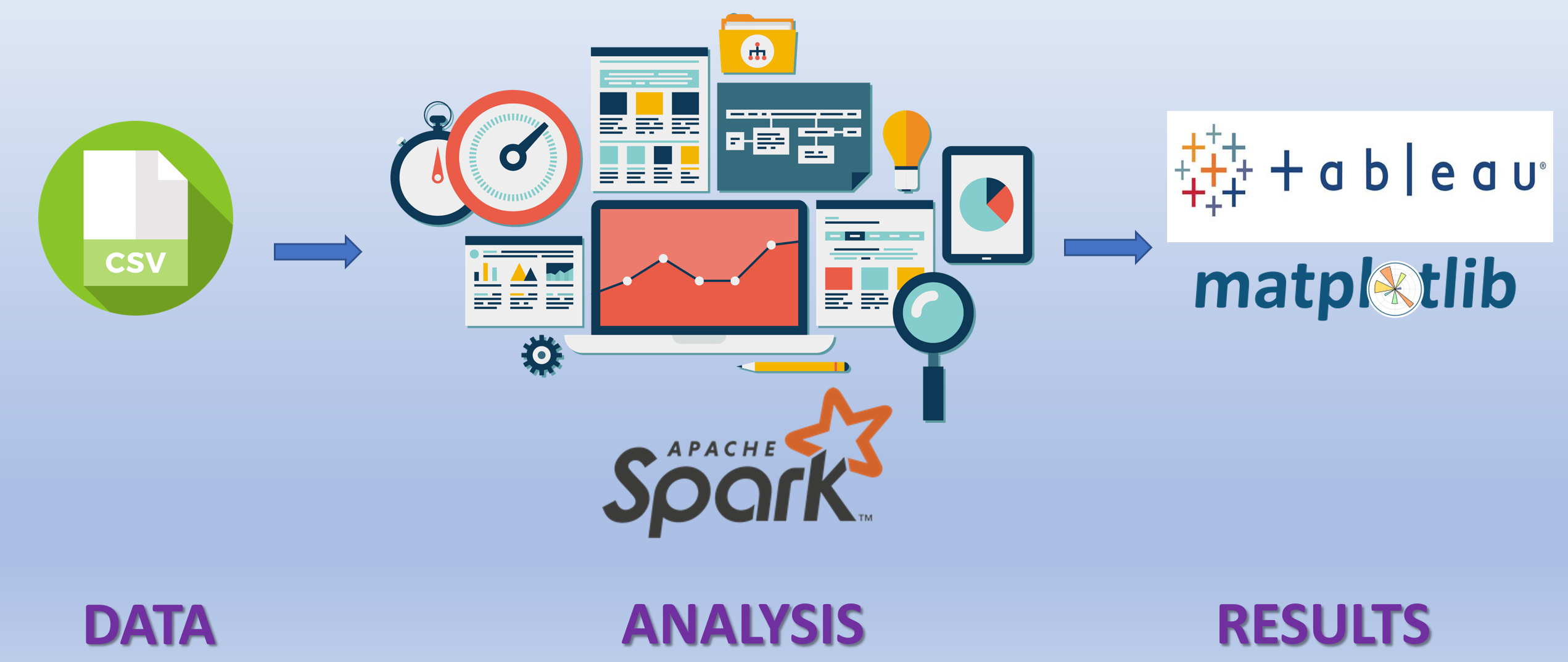
Deliveries has data about ever ball in every match, season-wise. It includes number of runs scored, batsman, bowler, wide runs, no ball details, wickets, etc., for every match played in the past 10 seasons.

2.2 Tech Stack

We implemented the project using the following technologies:

* PySpark
* Spark SQL
* Python Matplotlib
* Python pandas
* Tableau

The following figure shows the pipeline and architecture of the final system.



2.3 Data Cleaning

The IPL dataset was obtained from Kaggle and required data preprocessing.

The process of data cleaning involved several steps:

1. Collecting the data​: The data was collected was in a csv format and had to be transformed to a dataframe for the analysis.

2. Removing NA values​: NA values are critical since they can affect the analysis of our data, depending upon the number of NAs that are present in the data set. In our case, NAs were present in values and unit. So had to filter out those column to avoid manipulating the results.

**2.4 Data Analysis**

Since we have obtained the complete data in section 2.3, we can go ahead and start performing some analytics on the data. Some of the major analytics that we performed are:

1. Impact of toss on the outcome of the match

2. Analyzing which stadium is most suitable for batting first

3. Analyzing which stadium is most suitable for fielding first

4. Home Ground Wins Percentage

5. No. of wins by team and season in each city

6. Top 5 batsmen over the seasons

7. Top Bowlers over the seasons

8. Wicket Distribution in different overs

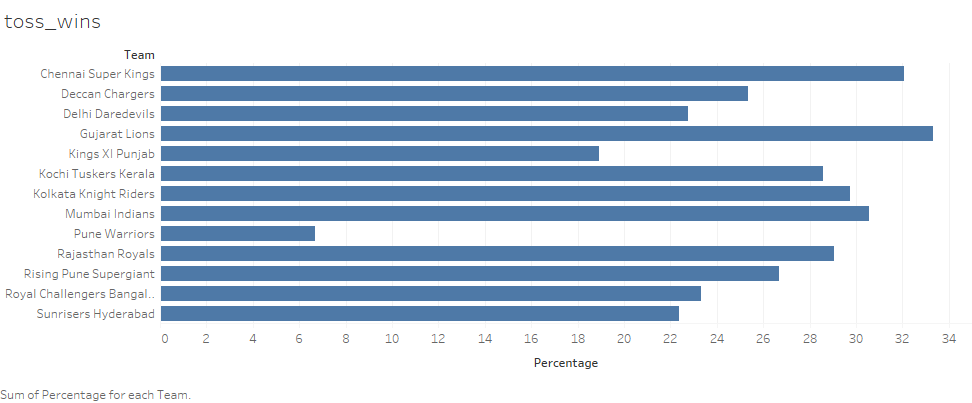
9. Runs Distribution in different overs

We also performed other basic analysis like Man of the Match, Centuries scored, Batsman and Bowler Aggregate, Total Number of Matches played, won and lost by a team, Top Batsmen and Bowlers over the seasons.

**3. Results and Discussion**

3.1 Results obtained with Tableau and matplotlib

3.1.1 Impact of toss on the outcome of the match



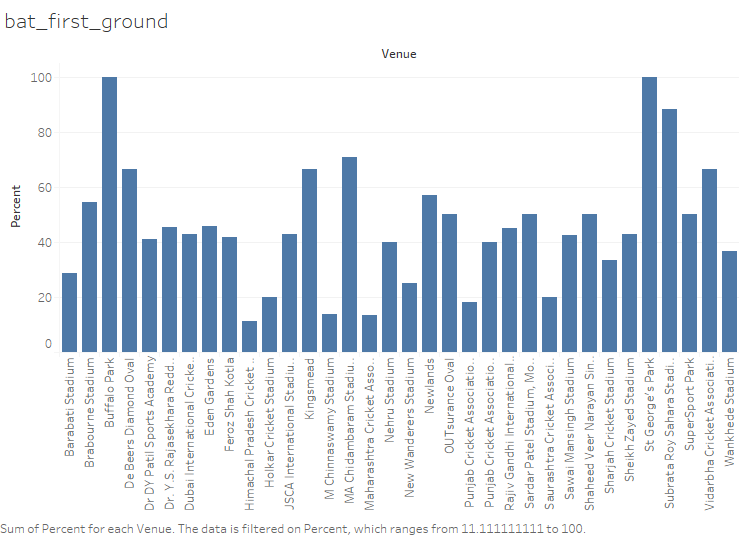
Winning the toss provides a small, but significant improvement to a team's chances of winning. The decision is of great tactical importance, and the captain will have considered many variables before arriving at his decision. The team winning toss can choose to bat or bowl first.

If the team is uncertain about the nature of the pitch or simply wants to play safe, also if the opposition bowling is strong, they often bat first. Sometimes the nature of the pitch deteriorates while the game progresses, making batting more difficult especially if facing spin bowling, batting first is then considered a good option.

The captain opts to bat second if he is confident that his team can successfully chase any total. Team’s bowling strength than batting skills helps win matches when chasing. The team just has to limit the opposition to a low score, and bat well to successfully chase the target.

From the graph above, we can see Chennai Super Kings has won the match 32% of the times after it won the toss.

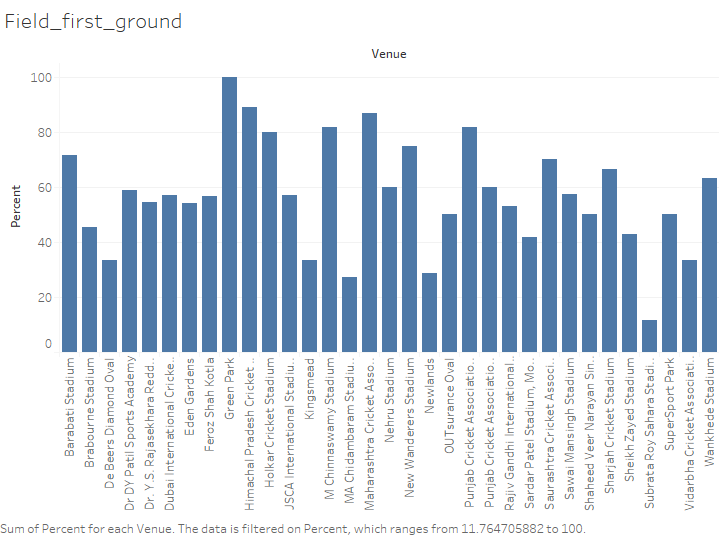
3.1.2 Analyzing which stadium is most suitable for batting first



There are various factors like pitch of the field, the ground size specific to the stadiums which can affect the decision whether to bat or bowl first at a stadium. If the pitch of the field deteriorates as the match progresses, then batting first is often considered a good option.

As we can see from the above graph, batting first at Buffalo Park stadium has led to winning the match 100% of the times and batting first at Himachal Pradesh Cricket stadium has winning percentage of only 10% approximately.

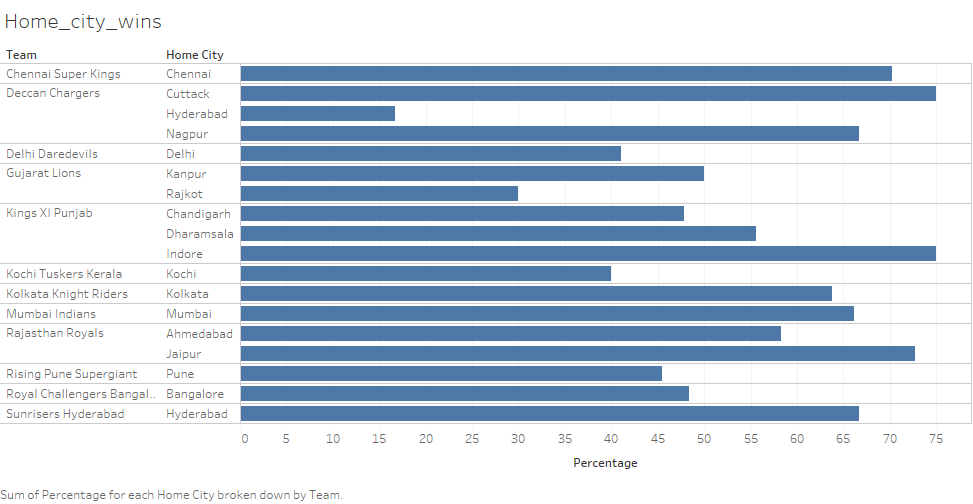
3.1.3 Analyzing which stadium is most suitable for fielding first



IPL matches are scheduled in the evenings which goes on till night. There are some grounds at Mumbai, Mohali and Dharamshala which accumulate meagre to a significant amount of dew in this period. Thus, ball becomes wet and it's difficult to grip, killing the ball variations. It becomes easier to play shots, as the ball reaches the bat straight, without much turn and swing. Wet outfield is slippery, making fielding unpleasant and tough for the fielders to hold on to catches. In such cases batting second is preferred.

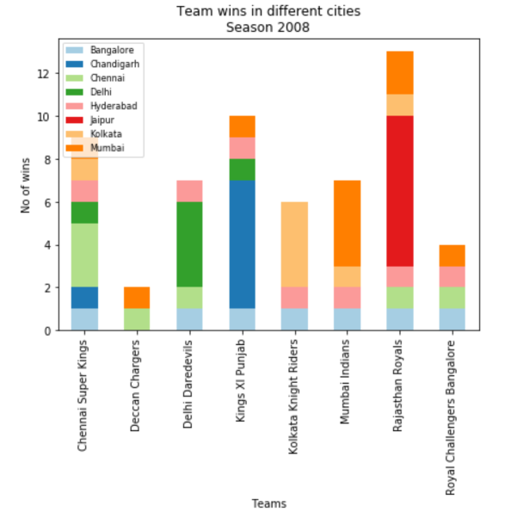
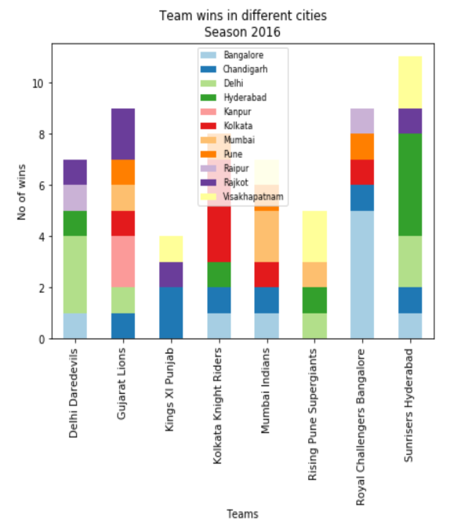
As we can see from the above graph, batting second at Himachal Pradesh Cricket stadium has led to winning the match roughly 90% of the times.

3.1.4 Home Ground Wins Percentage



From the above figure we can see the winning percentage of the teams when the match was held at their respective home cities. Deccan charges and Kings XI Punjab has the highest winning percentage of 75% when the match was held at their home cities Cuttack and Indore respectively.

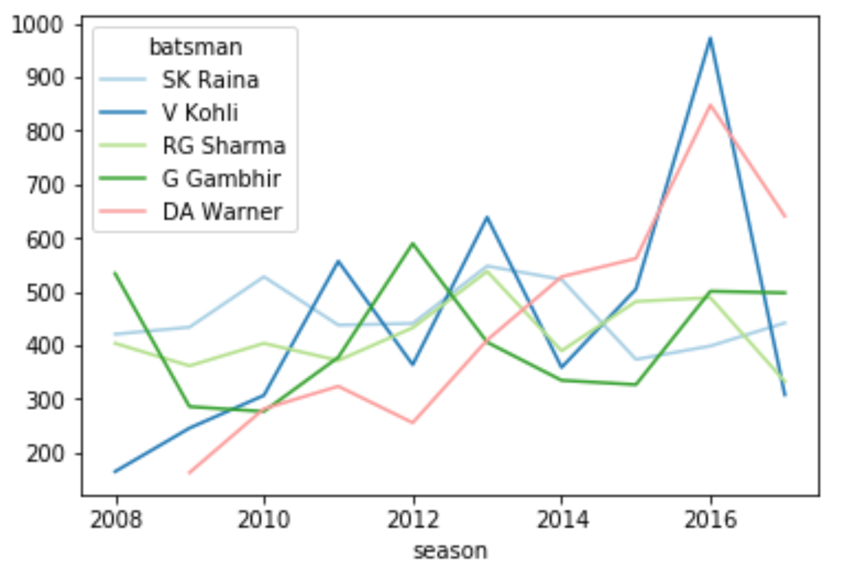
3.1.5 No. of wins by team and season in each city

Each team plays two matches against the other teams, one in its home city and other in the opponent’s home ground. Here we analyze whether a team playing at home or away has any impact their wins. The analysis is carried out for every season.

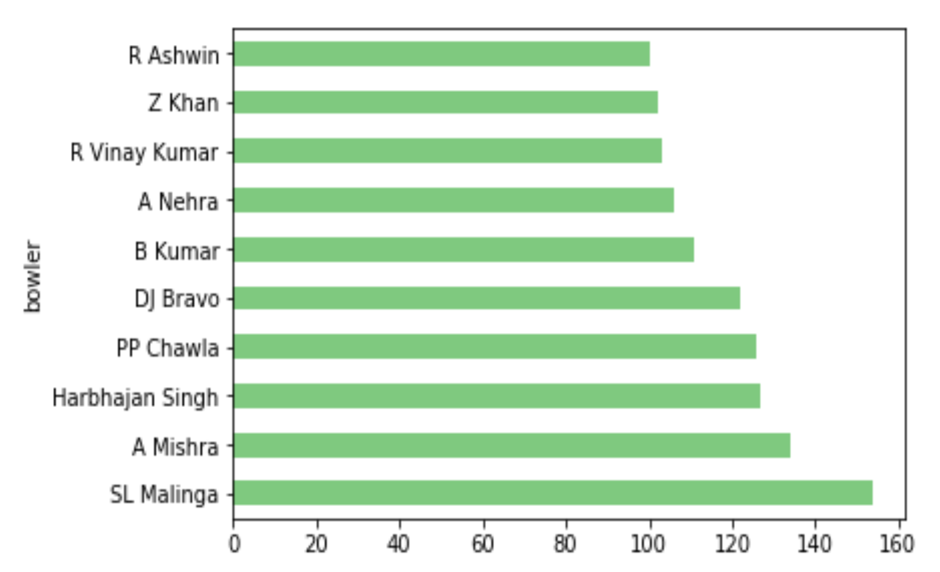
As we can see from the graph above, playing at home city has highest number of wins compared to other cities. For example, in season 2008, we can see that Rajasthan Royals has highest no of wins at its home city Jaipur. Similarly, in season 2016, Royal challengers Bangalore has highest no of wins at its home city Bangalore.

3.1.6 Top 5 batsmen over the seasons



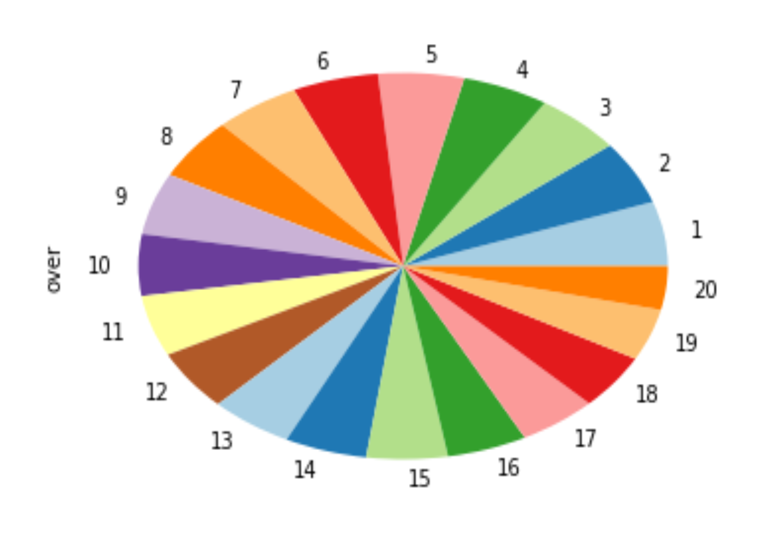
The figure above shows the top 5 batsmen in terms of their total runs over the seasons from 2008 to 2016. It shows that V Kohli has the highest number of runs scored over the seasons from 100 runs in 2008 to a cumulative 950 runs in 2016 approximately and DA Warner the second highest. We can infer that V Kohli shows a steady improvement over the seasons and SK Raina shows a slump.

3.1.7 Top Bowlers over the seasons



The figure above shows the top bowlers in terms of their total wickets over the seasons. It shows that SL Malinga has the highest total number of wickets earned over the seasons and A Mishra the second highest.

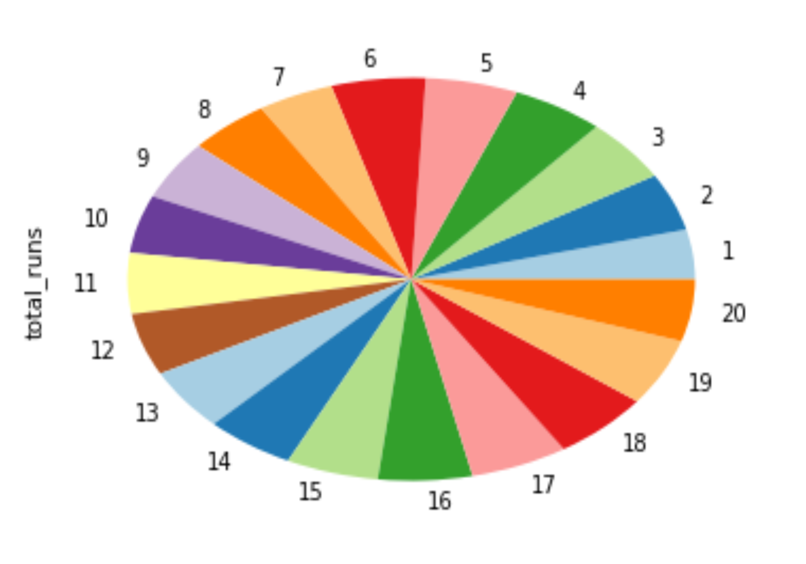
3.1.8 Wicket Distribution in different overs



IPL matches are played for 20 overs with 6 balls in each over excluding wide balls. Here we analyze the number of wickets in each over and in which over the maximum wickets were earned.

From the above figure we can see that wickets are more towards the first and last overs. During the first few overs, its possible that batsmen are not yet accustomed to the field, so the chances of getting out is high thus favoring the bowlers to score more wickets and during the last overs, the bowlers would try to score as much wickets as possible to win the match by limiting the opposition team to a low score.

3.1.9 Runs Distribution in different overs



Here we analyze the number of runs scored in each over and in which over the maximum runs were scored.

From the above figure, we can see that during the last few overs, there are more number of runs because the team would want to set a certain target score if it’s batting first and if it’s batting second it would try to score runs as much as possible to meet the target score set by the opponent’s team.

**3.2 Other analysis**

We also made other basic analysis as follows:

* Total Number of Matches Played by a team
* Total Number of Matches won by a team
* Total Number of Matches lost by a team
* Bowler Aggregate
* Batsmen Aggregate
* Man of the Match
* Centuries scored
* Top Batsmen over the seasons
* Top Bowlers in terms of Economy

**4 Code and References**

4.1 Code

The code is attached in the zip file.

4.2 References

●http://changhsinlee.com/install-pyspark-windows-jupyter/

● https://www.kaggle.com

● http://spark.apache.org/docs/latest/sparkr.html

● http://spark.apache.org/docs/2.1.0/api/python/pyspark.sql.html

● http://spark.apache.org/docs/2.1.0/api/python/pyspark.html

● <http://stackoverflow.com>

● https://www.tutorialspoint.com/tableau/index.htm