**ANALYSIS ON IPL MATCHES (2008-2017)**

|  |  |  |
| --- | --- | --- |
| Sarvesh Kedia | Yash Pradhan | Mukund Kadlabal |
| PES1201701256 | PES1201700262 | PES1201700648 |
| Department of CSE | Department of CSE | Department of CSE |
| PES University  Bengaluru, India | PES University  Bengaluru, India | PES University  Bengaluru, India |
| sarvesh512kedia@hotmail.com | yash.pradhan@gmail.com | mukund.kadlabal@gmail.com |

ABSTRACT

The dataset we have worked upon is about the matches in the IPL.

We have performed various operations on the dataset and have plotted different graphs to give us an insight into the data.

INTRODUCTION

The Indian Premier League or the IPL for short is one of the biggest cricketing extravaganzas in the world! Over the course of 10 years (2008-2017), there have been 13 different teams who have participated in the IPL.

In this project, we try and break down the statistics and trends observed in the IPL using Data Science. We bring forward some interesting facts & figures, which sure surprised us!

Our analysis is presented through graphs to better visualize the data & comparisons.

DATA SET

The data set is from kaggle.com.



Link:

https://www.kaggle.com/manasgarg/ipl

(matches.csv)

The various columns existing are the year, city played in, the teams that played, who won the toss and the subsequent decision, the winning team, the number of runs or wickets won by, man of the match, venue of the game and the umpires for the match.

The dataset we have worked upon is about the matches in the IPL.

We have performed various operations on the dataset and have plotted different graphs to give us an insight into the data.

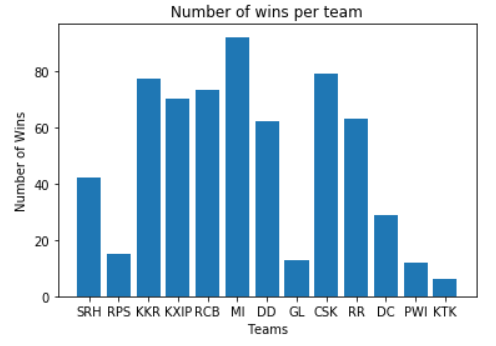
This research primarily tries to find the relations between the outcome of a match influenced by various parameters like venue of the game, outcome of the toss, decision to bat first and team played against using various statistical techniques and in turn predicting the outcomes of future games based off of our inferences.

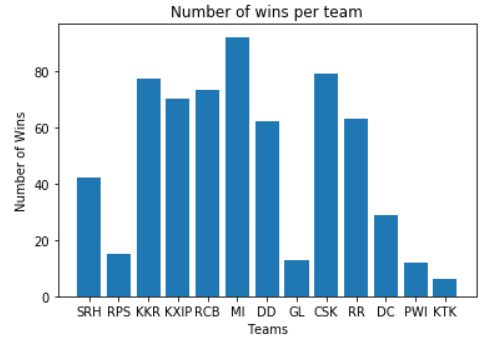
PRE-PROCESSING (DATA CLEANING)

We performed Data Cleaning on our data set to fill missing values & correct wrong values. E.g. The team Rising Pune Supergiant (RPS) was written as Rising Pune Supergiants in a few rows. This was likely to have been a typo, and we corrected the same using the Pandas module for python & replacing the wrong values in the CSV file with the correct ones**.**

DESCRIPTIVE ANALYSIS

1. Team Wins – Bar Graph

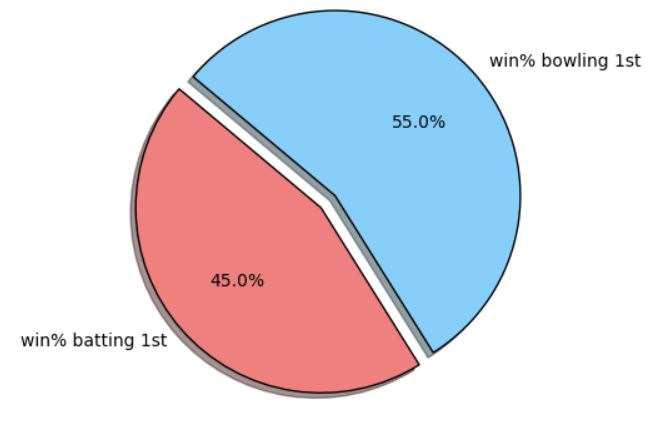




This bar graph represents the total number of matches won by a team since 2008. It clearly shows that. MI won most matches while KTK has won the least number of matches.



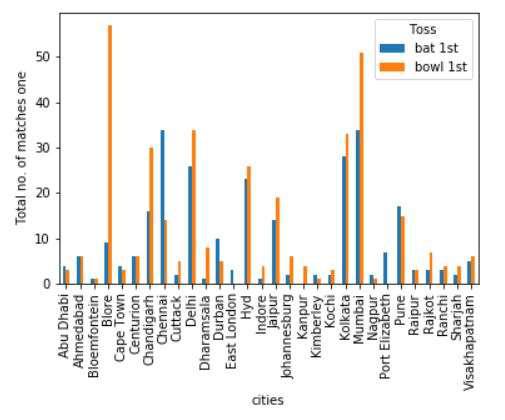
2. Wins: innings-wise – Pie Chart



This pie chart shows that teams batting first have won 45% of the games played while teams bowling first won 55% of the games played. Hence, we can infer that teams batting second have been more successful in IPL



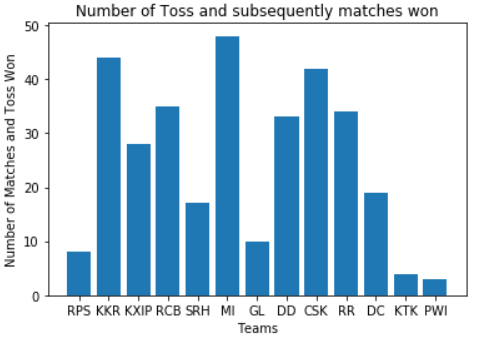
3. Toss outcomes at different venues



It shows that in venues like Bangalore and Mumbai, teams bowling first have been more successful while in venues like Chennai and Pune, teams batting first have been more successful**.**



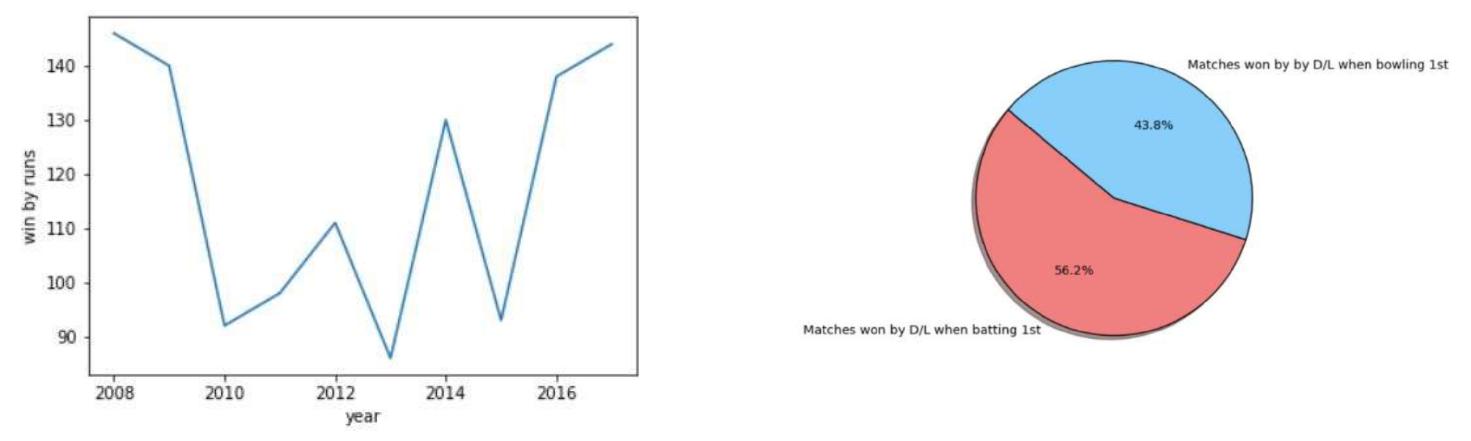
4. Teams that won the toss and subsequently won the match.



This bar graph represents the no. of matches where the team won, both the toss and the game. We can infer that teams like MI and KKR have won 48 games and tosses while PWI and KTK have won 3 games each.



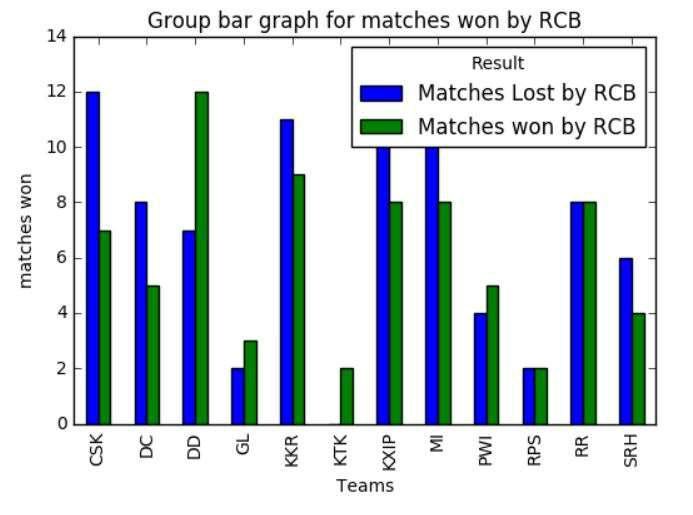
5. Margin of victory over the years 7. Results by D/L Method



This graph is simple representation of highest margin of victory by runs each year. WE can conclude that the margin of victory (in runs) in 2008 was the highest and least in 2013



6. RCB match results



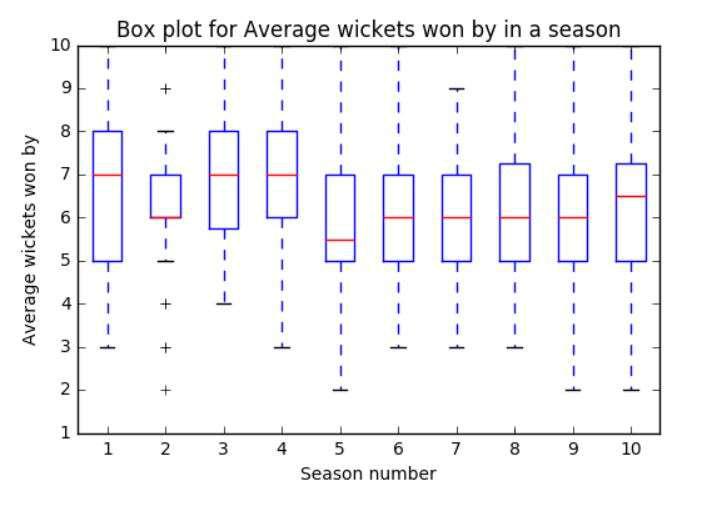
It shows number of wins and losses by RCB against each team. RCB have been more successful against DD and PWI and lost more matches against CSK and KKR.



This pie chart represents the success rate of winning by DL method while batting first or bowling first. It concludes that 56.2% of games have been won by teams bating first while 43.8% of the games have won by teams bowling first.



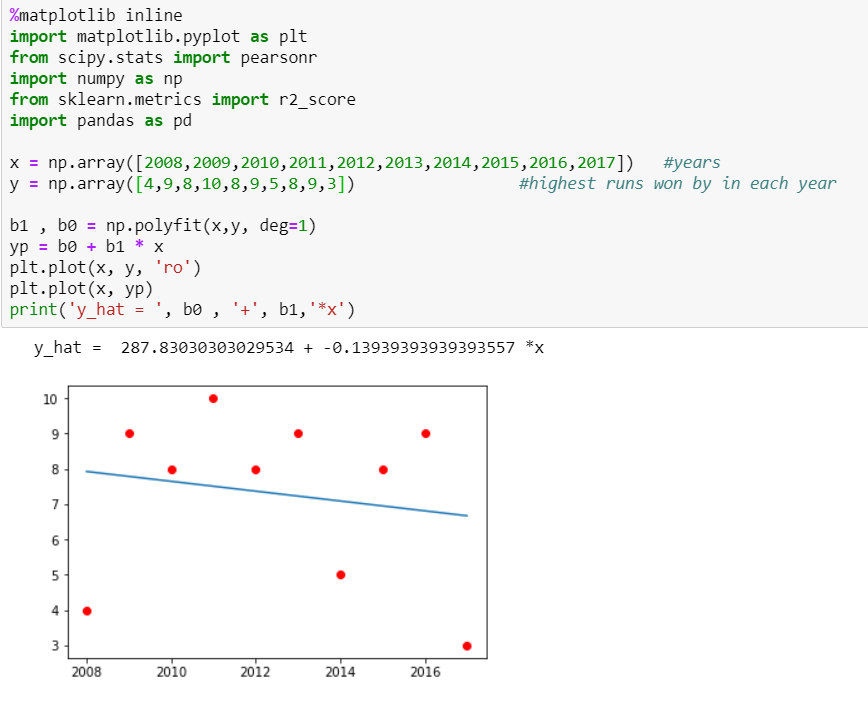
8. Average wickets won by – season wise

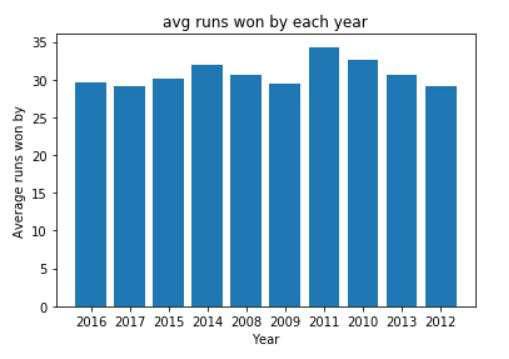


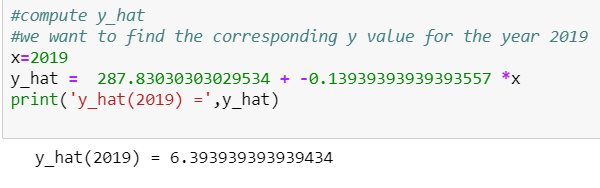
The Box Plot represents the average number of wickets won by in each season.

We observe that in first, third and tenth season the data is left skewed and in the second and fifth season, its right skewed.



9. Average runs won by – year wise





The graph represents the average number of

runs won by in each season.

The average runs won by was highest in 2011 and least in 2009



PARAMETRIC TESTING

*Simple Regression:*

Simple linear regression is a statistical method that allows us to summarize and study relationships between two continuous variables. One variable denoted *x*, is regarded as the **predictor** and the other variable denoted *y*, is regarded as the **response**.

For our data set we have performed simple Linear regression. Using this, we try to predict the number of matches Royal Challengers Bangalore will win in the 2019 edition of the Indian Premier League.

We get a y\_hat value of 6.39, so now we can predict that in the 2019 edition of the Indian Premier League RCB will win 6 Matches.



NON-PARAMETRIC TESTING

*Chi Square Test:*

Chi square is a non-parametric test that is used to asses the goodness of fit between a set of observed values and those expected theoretically.

For our dataset we use Chi Square test to check whether two variables are dependent on each other or not. In this case we are checking for the relationship between years and the highest margin of victory (in runs) in that year.

chisquare(x)

chisquare(y)

cont=pd.crosstab(x,y)

pvalue=scipy.stats.chi2\_contingency(cont)

[1]

After running the above code, we get a pvalue of 0.231. Since pvalue>0.05 we can infer that the two variables “highest margin of victory (in runs)” and “year” are independent of each other.

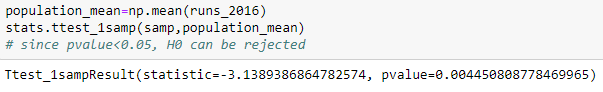


HYPOTHESIS TESTING

*Hypothesis 1:*

Null Hypothesis(H0): Average runs won by in the year 2016 is greater than population mean.

Alternate Hypothesis(H1): Average runs won by in the year 2016 is lesser than population mean.

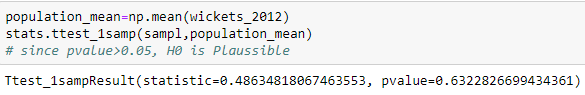


We get a pvalue of 0.017, since the pvalue is lesser than 0.05 we reject the Null Hypothesis.

*Hypothesis 2:*

Null Hypothesis (HO): Average wickets won by in the year 2012 is lesser than population mean

Alternate Hypothesis(H1): Average wickets won by in the year 2012 is greater than population mean



We get pvalue of 0.63, since it is greater than 0.05 we conclude that Null hypothesis is plausible.

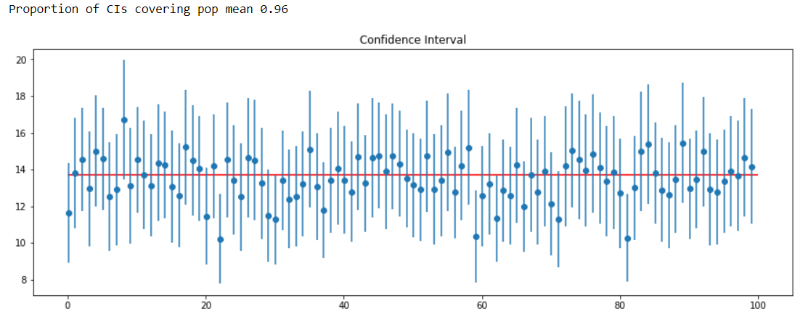


CONFIDENCE INTERVAL

It is a type of interval estimate computed from the statistics of the observed data, that might contain the true value of an unknown population parameter.

It quantifies the level of confidence that the parameter lies in the interval.

The proportion of confidence intervals that could cover the population mean is 0.96





CONCLUSION

The IPL has many statistical and predictive applications, in our project we have tried to implement one parametric and one non-parametric test on a sample population.

Using our descriptive analysis, we have made inferences on a variety of issues like outcomes of the toss at various venues, the number of matches won by D/L method when teams are batting 1st, number of matches won when batting 1st, matches won when the winning team has also won the toss and individual team performances against their counterparts.

We have used simple linear regression to predict the number of matches Royal Challengers Bangalore will win in the year 2019 and we have used the chi square test to test whether two variables are dependent on each other.

We have also performed simple independent tests to check certain Hypothesis and whether the Null hypothesis are Plausible or should be rejected.