INDIAN PREMIER LEAGUE

Analysis of IPL

CRICKET TOURNAMENT

Renganathan Lalgudi Venkatesan | Data Management | M12366827

Contents

Introduction

Indian Premier League (referred to as IPL) is an annual cricket tournament that is conducted every year in the Indian subcontinent. The objective of this project is to analyze the IPL tournament data set and to generate insights for teams based on their performances at different venues. The major part of the project involves analyzing the data collected over the past nine seasons of IPL from 2008 to 2016 and explore insights that are no commonly perceived. The performance of teams across venues and across time is analyzed using tools like SQL, R and Tableau. Finally, a summary of the analysis is also presented as per the findings from the analysis.

DESCRIPTION OF THE DATASET

The dataset used for this project has been obtained from Kaggle (as a part of a competition) (https://www.kaggle.com/nmaheshw/ipl-data-analysis/data). The data are divided into three CSV files each containing details about the Matches, Players and Teams.

OVERVIEW OF COLUMNS AND VALUES

The data sets considered has three different CSV files with a wide range of columns. Following is the description of the data fields considered in our analysis. The column description is spread across the different files as shown below:

Table -1: Match.csv:

Match_Id: The is a numeric value and serves as the primary key for this table. It is unique for each row of the table

Match_Date: The date on which the match was played. It is in DD-MM-YYYY format but is stored as a character string in the dataset. We will convert it to Date format during our analysis.

Team_Name_Id: There are 13 teams with Id's from one to thirteen. The description of each Id is present in the Team.csv file

Opponent_name_Id: There are 13 teams with Id's from one to thirteen. The description of each Id is present in the Team.csv file

Toss_Winner_Id: This field contains the Winning team Id. This again can be referenced from the Team table.

Toss_Decision: This is a categorical variable with two possible values: bat/field. These are in character string format in the table. We will convert them as factors in our analysis with R

Won_By: This is a numeric variable representing the margin by which the team won the match

Match_Winner_Id: This is a numeric Id representing which team won the match

City_Name: The venue the match was played at. It is a string variable

Table -2: Team.csv

Team_Id: A numeric variable with values from one to 13 representing the different teams

Team_name: A character string that has the names of the teams participating in the tournament

Team_Short_Code: Short forms for the team names

Table - 3: Player.csv

Player_Id: Each player has a unique numeric Id

Batting_hand: A player can be a right hand or left-hand batsman. This is a string variable

Country: The country the player represents at the international level

NORMALIZAION OF THE DATABASE

The dataset is normalized already. It has already been divided into independent tables with primary keys and foreign keys. The major concern here is though that the primary keys of one table are present as foreign keys in another table with a different name. We need be careful in selecting the ids when performing joins with the tables.

PROBLEMS IN THE DATASET

There were several challenges with the dataset as described below:

- The data set was in csv format. I tried for the first time using the Flat File Source. This way I could import the CSV files into SQL.
- Column "Venue_Name" had entries with special characters which could not be read from the CSV files into SQL using the Input Data Wizard. So, had to clean the data to have a proper data format before importing into SQL
- The data columns were also over lapping in certain cases after importing the file. That is, some of the columns when read through the input wizard were overlapping each other. So, had to get rid of the venue column as we are not using it for the analysis also.

ANALYSIS USING SQL

Given the database with three csv files, we first import the data into the database as tables. Then we performed the following analysis:

a. Looking at the data tables and the different variables present in each table

The first thing we want to do is to understand the distribution of various variables in the data. We can do this in SQL by looking at the different columns and their values. That is the first task to be performed. We do that by selecting all the columns in the given data set and looking at their values.

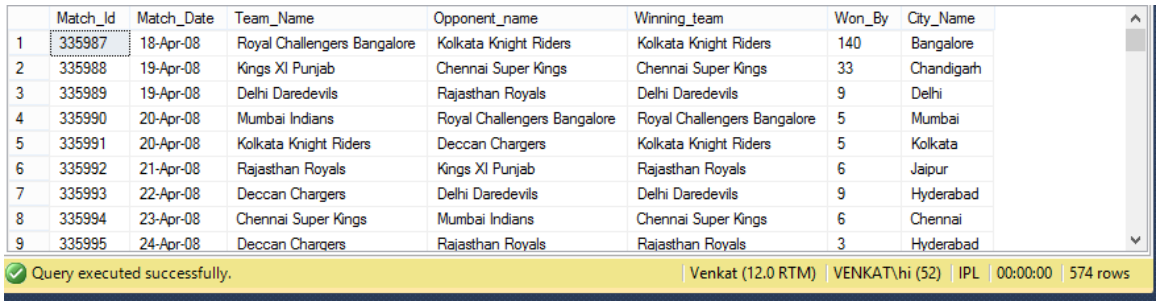
b. Data cleaning for NULL values in the Won_By Column

We find that there are NULL values in the Won_By column. We fill these values with a 0 as it is the runs by which the match was won and if this has a value of 0, it means there was no result and thus an equivalent way of saying it is by inserting a 0 in place of NULL.

c. Merging the different tables to obtain the required table for analysis:

In order to perform further analysis, many tables were needed to be merged. This was done and we had a final table with the columns of our interest. The final merged table that has been used for this analysis in SQL contains the following columns: Match_Id, Match_Date, Team_name, Opponent_name, Winning_Team, Won_By and City_Name.

We find that the new table has 574 rows of data with all the variables. This is a table with 574 unique matches that were played over the past nine seasons of IPL.



	Match_Id	Match_Date	Team_Name	Opponent_name	Winning_team	Won_By	City_Name
1	335987	18-Apr-08	Royal Challengers Bangalore	Kolkata Knight Riders	Kolkata Knight Riders	140	Bangalore
2	335988	19-Apr-08	Kings XI Punjab	Chennai Super Kings	Chennai Super Kings	33	Chandigarh
3	335989	19-Apr-08	Delhi Daredevils	Rajasthan Royals	Delhi Daredevils	9	Delhi
4	335990	20-Apr-08	Mumbai Indians	Royal Challengers Bangalore	Royal Challengers Bangalore	5	Mumbai
5	335991	20-Apr-08	Kolkata Knight Riders	Deccan Chargers	Kolkata Knight Riders	5	Kolkata
6	335992	21-Apr-08	Rajasthan Royals	Kings XI Punjab	Rajasthan Royals	6	Jaipur
7	335993	22-Apr-08	Deccan Chargers	Delhi Daredevils	Delhi Daredevils	9	Hyderabad
8	335994	23-Apr-08	Chennai Super Kings	Mumbai Indians	Chennai Super Kings	6	Chennai
9	335995	24-Apr-08	Deccan Chargers	Rajasthan Royals	Rajasthan Royals	3	Hyderabad

Query executed successfully. | Venkat (12.0 RTM) | VENKAT\hi (52) | IPL | 00:00:00 | 574 rows

d. Cities that have hosted the most number of matches in the tournament

We want to understand the distribution of matches across the venues. We do this by grouping the venues in and aggregating the number of matches played in a particular venue.

We find that most number of matches have been played at Mumbai with 77 matches being played in this city within the past nine seasons.

The venue that had hosted the least number of matches is Bloemfontein with just two matches being held at this Venue. When we dig deep we find that Bloemfontein is a city in South Africa and there was only one season of IPL that was played at SA. This explains why there were only two matches at this venue.

	City_Name	Num_of_matches
1	Mumbai	77
2	Bangalore	56
3	Kolkata	54
4	Delhi	52
5	Chennai	48
6	Chandigarh	42
7	Hyderabad	41
8	Jaipur	33

e. Teams that have won the most number of matches

We want to find how many teams have won how many matches over the years. This was done by finding out the matches won by each team by grouping the matches according to their teams.

We find that Mumbai has won the most number of matches (80) closely followed by Chennai (79). This also explains the number of times these teams have won the title in IPL in the past nine seasons. Mumbai had won it thrice and Chennai twice out of 9 seasons.

Pune has won the least number of matches. This again can be explained by their inception into the tournament. They joined the tournament only in 2014 and in two seasons they have won 5 matches in all. Still this is a low number of matches to win in three seasons.

	Winning_Team	Num_Times_Won
1	Mumbai Indians	80
2	Chennai Super Kings	79
3	Royal Challengers Bangalore	70
4	Kolkata Knight Riders	68
5	Rajasthan Royals	63
6	Kings XI Punjab	63
7	Delhi Daredevils	56
8	Sunrisers Hyderabad	34
9	Deccan Chargers	29

f. Matches won by different teams in different Venues (*)

We then calculate the number of matches won by a team at a particular venue. This will give us an indication as to if some team has some affinity to a certain venue or not.

We find an interesting fact here. Every team seems to have played and won most of their matches at their home venues. If we consider Mumbai and Chennai, they have almost won all the matches played at their home ground. This pattern is applicable to all other teams too. Every team tends to win more matches at their home ground than playing in a away stadium.

	Winning_team	City_name	Num_Won_Venue
1	Mumbai Indians	Mumbai	40
2	Kolkata Knight Riders	Kolkata	33
3	Chennai Super Kings	Chennai	33
4	Royal Challengers Bangalore	Bangalore	29
5	Rajasthan Royals	Jaipur	24
6	Kings XI Punjab	Chandigarh	20
7	Delhi Daredevils	Delhi	19
8	Sunrisers Hyderabad	Hyderabad	14

g. Distribution of number of matches played in each date

Usually in the IPL tournament, multiple matches may be played on a single day. We want to find out how many matches were played in general. We find that on a given date there can be either one or two matches. The distribution of the number of days with two matches being played and one match being played is found to be:

	Num_matches	days
1	1	238
2	2	168

ANALYSIS USING R

Having performed some preliminary analysis is SQL, I wanted to perform further analysis on the data tables created in SQL using R. I first connected the R on my PC to the SQL server and read all the required tables for the analysis as dataframes. Then I had performed the following analysis:

First of all, I wanted to understand the dimensions of the dataframes read into R. We find the following regarding the tables created:

Match table has 574 rows and 18 variables

Player table has 1046 rows with 8 variables

Team table has 26 rows with 3 variables

a. Understanding the number of players from each country

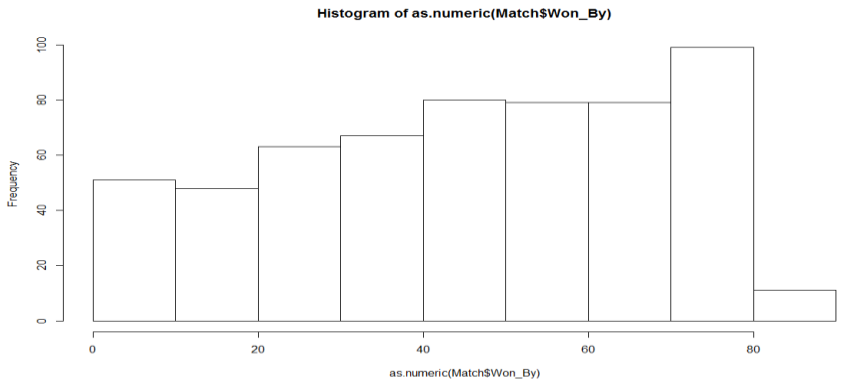
I wanted to identify how many players participated in the IPL tournament from different countries. Since it is an international tournament, there were several players participating in the league from various countries.

(*)As expected, the tournament was dominated with Indians with 582 players. It was amazing to see that countries like Australia had a really huge participation with 156 players involved in the tournament. The next big country that participated in the tournament was South Africa with 88 players. It was surprising to see that a country like Sri Lanka which was a part of the Indian subcontinent had a little participation in the tournament.

Australia	Bangladesh	England	India	Netherlands	New Zealand	Pakistan	South Africa	Sri Lanka	West Indies	Zimbabwea
156	10	36	582	2	52	30	88	44	40	6

b. Histogram of Scores in matches that were won by runs

We find that the margin of winning a match had almost a uniform distribution between one to 60 runs but there is a huge spike in the number of matches that were won by great margins(*)



c. Understanding Venues with their Won_By runs Statistics

We find that Bangalore as venue that has the maximum average winning margins. Then when we take a deep dive to understand why this was happening we find that of all the venues, Bangalore has the smallest ground size and this was affecting the margins by which teams won a match

City_Name	Num_matches	AvgWinning_Margin
Abu Dhabi	20	16
Pune	25	18
Hyderabad	41	13
Jaipur	33	18
Kolkata	54	12
Chandigarh	42	16
Chennai	48	17
Delhi	52	16
Mumbai	77	16
Bangalore	56	25

d. Choice at the Toss

Surprisingly, we find that most teams favored batting second.(*). Out of the 574 matches played, 315 times team that won the toss chose to field first and only 262 times the teams chose to bat first. This might be because of the nature of the game in which people have generally started to think that chasing a score is much easier compared to setting a target.

bat	field
262	315

e. Comparing the Toss Decisions Team wise

When we compare the decision at toss to bat/field, we find that most teams follow the general tendency to chase a target. But the only team that stands out of this is Chennai. This team has chosen to bat 50 times out of the 79 tosses it has won. This is a surprising number as it has only chosen to chase a score 29 times but the team has won the trophy twice in spite of this(*)

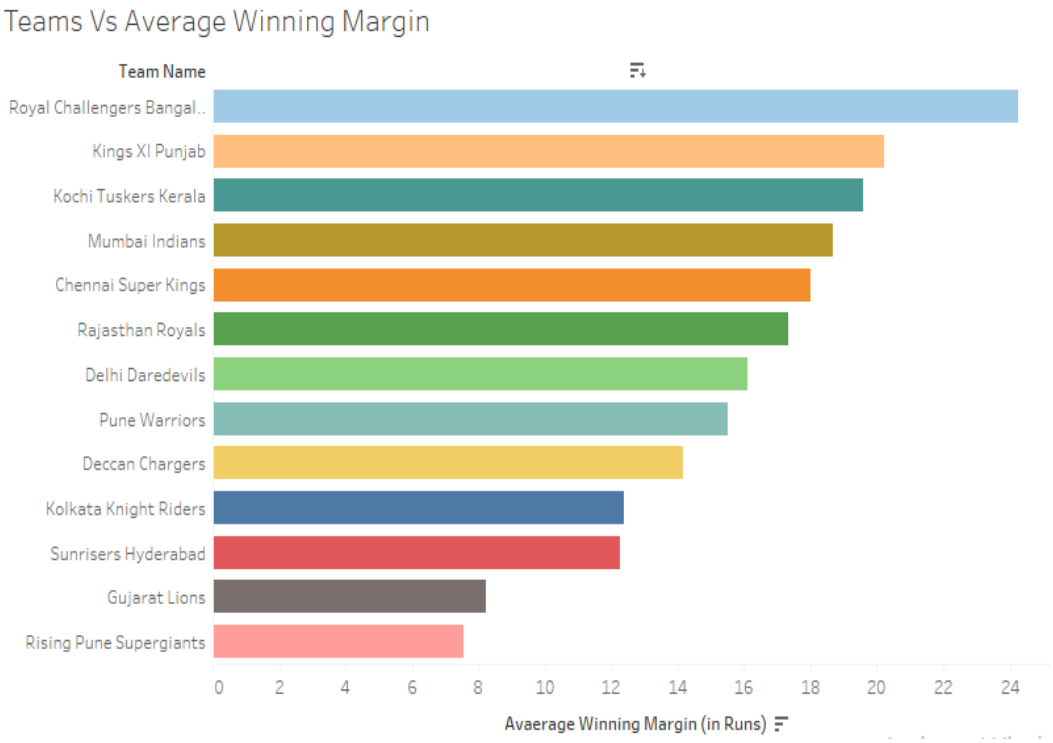
	bat	field
1	33	35
2	24	46
3	50	29
4	19	44
5	30	33
6	26	30
7	38	42
8	14	15
9	0	6
10	9	3
11	14	20
12	2	3
13	2	7

ANALYSIS USING TABLEAU

Then next I wanted to explore some features that were not explainable using the tabular form and thus wanted to make interactive visualizations in tableau to find more insights about the teams and venues.

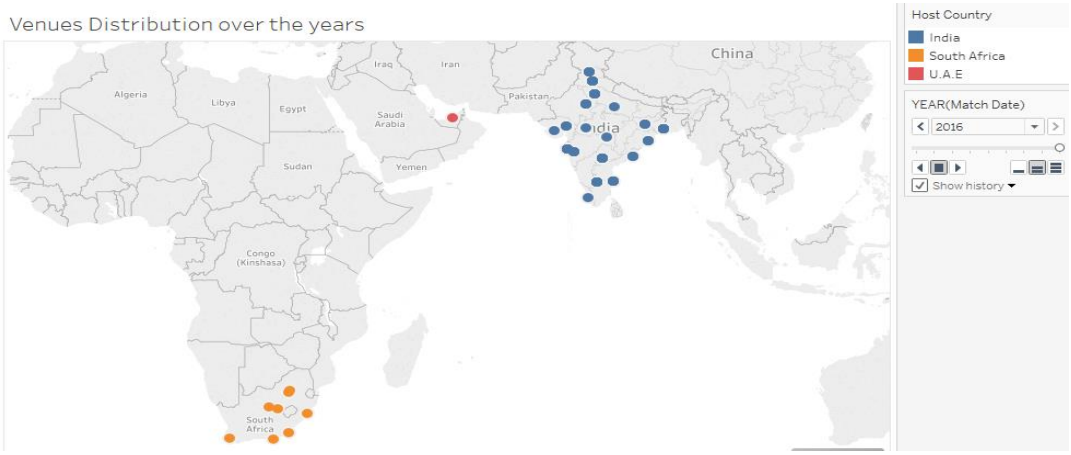
a. Average winning margins for the teams in the tournament

I wanted to compare the average winning margins for every team. We find that Bangalore team leads this. All the matches they have won have been by huge margins. So on an average they won a match by 24 runs which is quite phenomenal. Also if we look at other teams we find that Chennai and Mumbai are being consistent throughout. Both the teams have a big winning margin of 18 runs and 19 runs respectively which also explains them finishing top of the table most of the times and winning the tournament multiple times.



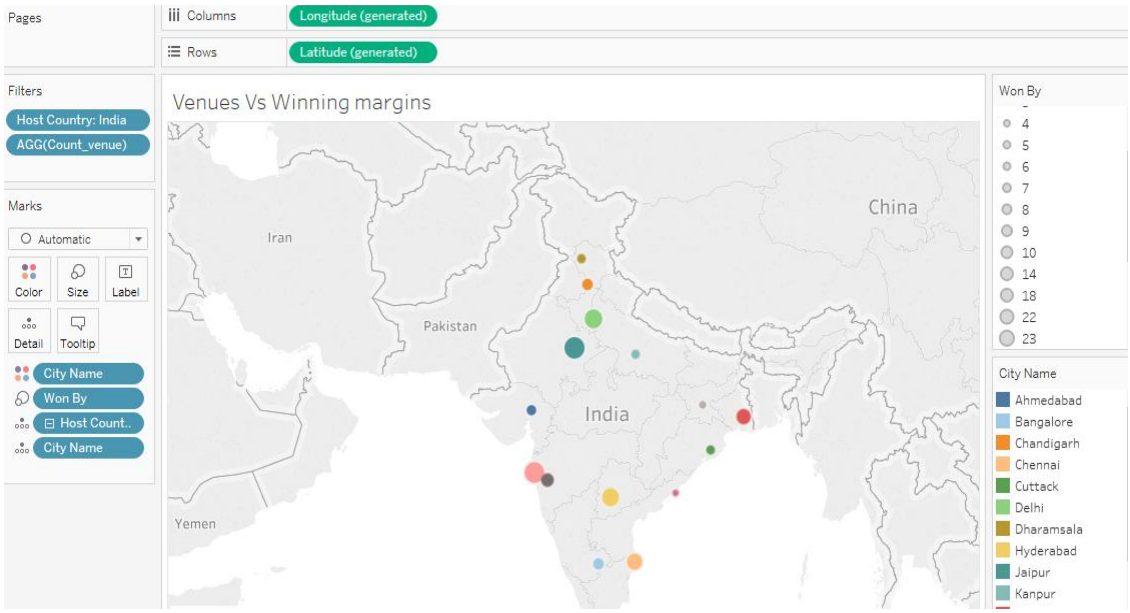
b. Venues Changed over the Years: Interactive Visualization

Using the power of tableau, I wanted to make a dynamic visualization of how the venues for IPL changed over the years through all the nine seasons. The results look amazing as we move over the years. With new teams created and some older ones split, the venues update according to the participating regions. We also find that in the year 2009, the tournament had been hosted at South Africa but still managed to maintain the same zeal and spot light in the cricketing circles(*)



c. Effect of Winning margins based on the venues

I also wanted to see the distribution of winning margins across the different venues where the games were played. We find that Mumbai, Delhi and Rajasthan were the major venues where the winnign margins were substantially high.



SUMMARY

Being an ardent cricket fan, I never knew there were so many things happening in IPL which I could not think of without performing an analysis. The following were the findings from the data set that I believe will be useful for any team manager:

- We find an interesting fact that every team seems to have played and won most of their matches at their home venues. If we consider Mumbai and Chennai, they have almost won all the matches played at their home ground. Thus, it is an important criterion to be looked if a team is expecting to win a match and bet for a location accordingly in times of critical knock out matches
- As expected, the tournament was dominated with Indians with 582 players. It was amazing to see that countries like Australia had a huge participation with 156 players involved in the tournament. The next big country that participated in the tournament was South Africa with 88 players. It was also surprising to see that a country like Sri Lanka which being a part of the Indian subcontinent had little participation in the tournament with just 40 players. So as a manager it would be great to look out for Australian talents and make the best of the opportunity.
- We find that the margin of winning a match had almost a uniform distribution between one to 60 runs but there is a huge spike in the number of matches that were won by great margins between 60 to 80 runs. This means that most teams panic while chasing big scores. It is good to develop a plan for such cases so that the margin of loss can be reduced, and it impacts the team's chances in the tournament
- Surprisingly, we find that most teams favored batting second. Out of the 574 matches played, 315 times team that won the toss chose to field first and only 262 times the teams chose to bat first. This might be because of the nature of the game in which people have generally started to think that chasing a score is much easier compared to setting a target. As a manager it is critical to look at the winning percentages at venue given a toss decision
- When we compare the decision at toss to bat/field, we find that most teams follow the general tendency to chase a target. But the only team that stands out of this is Chennai. This team has chosen to bat 50 times out of the 79 tosses it has won. This is a surprising number as it has only chosen to chase a score 29 times, but the team has won the trophy twice
- The venues for IPL changed over the years through all the nine seasons. The results look amazing as we move over the years. With new teams created and some older ones split, the venues update according to the participating regions. We also find that in the year 2009, the tournament had been hosted at South Africa but still managed to maintain the same zeal and spot light in the cricketing circles. As an IPL organizer this is a great insight for me if I am planning to conduct the tournament outside of India. The venue of the tournament doesn't seem to impact the revenue from the tournament

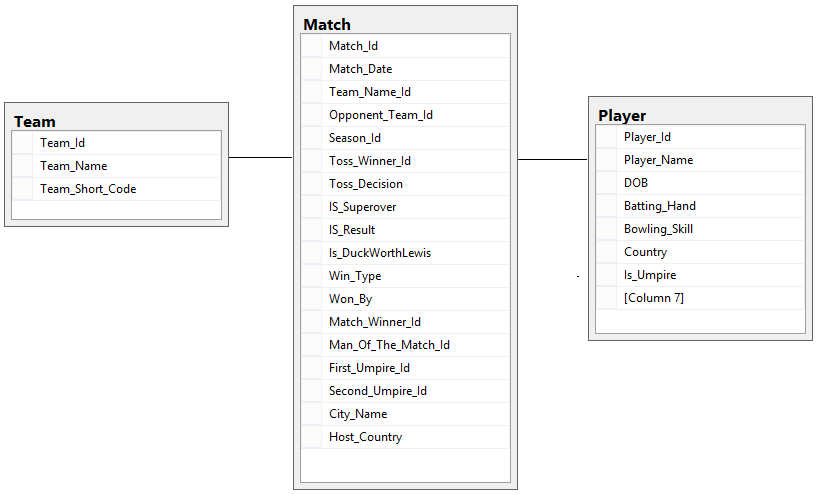
CHALLENGES

There were several challenges faced during the analysis:

- The data set was in csv format. I tried for the first time using the Flat File Source. This way I could import the CSV files into SQL.
- Column "Venue_Name" had entries with special characters which could not be read from the CSV files into SQL using the Input Data Wizard. So, had to clean the data to have a proper data format before importing into SQL
- The data columns were also over lapping in certain cases after importing the file. That is, some of the columns when read through the input wizard were overlapping each other. So, had to get rid of the venue column as we are not using it for the analysis also

APPENDIX

- ***Relation Schema***:



- Joins used in Tableau from the tables for performing the analysis