PES UNIVERSITY, Bangalore

(Established under Karnataka Act No. 16 of 2013)

UE18CS203

B.Tech, Sem III
Session : Aug-Dec, 2019

UE18CS203 – INTRODUCTION TO DATA SCIENCE

REPORT
ON
EXPLORATORY ANALYSIS ON
IPL DATASET

SECTION : A

#	SRN	Name	Contact No.	Email ID	Sign
1	PES1201800042	Revanth Babu P N	7411958095	putturevanth@gmail.com	
2	PES1201800230	Navyadhara G	9133132005	navyadhara79@gmail.com	

ABOUT THE DATA SET

Indian Premier League (Cricket)
This dataset contains two files: deliveries.csv and matches.csv. It contains the following:

All Indian Premier League Cricket matches between 2008 and 2016.
This is the ball by ball data of all the IPL cricket matches till season 9

The dataset contains 2 files: deliveries.csv and matches.csv.
deliveries.csv (150461rows*21columns)

1. match_id
2. Inning : Tells if the first set of batting was going on or second. 1: First Innings 2: Second Innings
3. Batting_team : The team name which is currently batting.
4. Bowling_team : The team name which is currently bowling.
5. Over : Describe the current over number.
6. Ball : Describe the current bowl no of the current over.
7. Batsman : Name of the batsman on striking end.
8. Non_striker : Name of the batsman on non-striking end.
9. bowler
10. is_super_over
11. wide_runs
12. bye_runs
13. legbye_runs
14. noball_runs
15. penalty_runs
16. batsman_runs
17. extra_runs
18. total_runs
19. player_dismissed
20. dismissal_kind
21. fielder

matches.csv(636rows*18columns)

1. id
2. season
3. city
4. date
5. team1
6. team2
7. toss_winner
8. toss_decision
9. result
10. Dl_applied : Duckworth Lewis method
11. winner

12. win_by_runs
13. win_by_wickets
14. player_of_match
15. venue
16. umpire1
17. umpire2
18. umpire3

ABSTRACT

The basic purpose of the assignment is to analyze and provide some useful insights about the dataset. The question we asked are how can this data be analyzed providing beautiful insights and also giving some facts. The analysis of the dataset gave us the answers. Our analysis can answer various questions like which batsman scored more runs, which team won a greater number of games, which bowlers' economy is better,which team has won the most seasons, which bowler has given the most runs/taken most wickets,prove or disprove-The winner of the toss is more likely to win the match, which batsman has scored most boundaries, singles, doubles, which bowler has given most extras, etc.

EXPLORATORY ANALYSIS

The dataset initially had some missing values and had a column with no data. It had some columns with duplicate names. The data is cleaned for all of the above cases and we arrived at a cleaner dataset. The dataset had two different datasets, matches.csv and deliveries.csv. Matches.csv has 636 rows and 17 columns. The other one has 150461 rows and 21 columns. Both of these datasets are used to analyze the data.

Python packages used: pandas,matplotlib, numpy, seaborn,scipy
Using these, we have plotted various graphs- bar chart, grouped bar chart, pie chart, grouped line graph,box plot and they have been used to arrive at various conclusions and insights.
Pie Charts- Is toss winner the match winner, Chances of chasing 200+ scores
Grouped line graph -Runs scored by top batsmen in different seasons, boundaries(4s and 6s) scored in Different seasons
Bar Graphs-Batsman with most runs, most player of the match awards, bowler with most wickets,most extras, bowler who has conceded most runs, the most common kind of dismissals.
Grouped bar chart - total 4s and 6s scored by different teams, toss decisions in different seasons

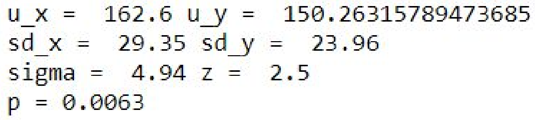
We have also framed hypothesis and tested using the data.
We have compared the average scores in 2009 vs that of in 2016. 2009 was hosted by South Africa.

Hypothesis Testing

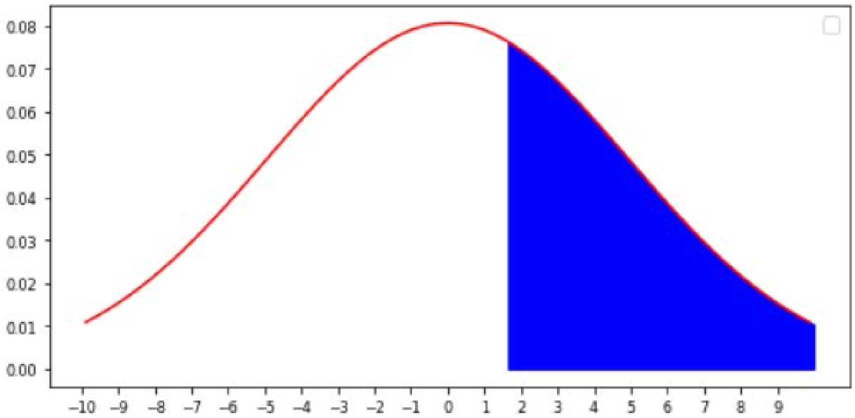
We have compared the average scores per match in 2016 vs that of in 2009.

x: denotes year 2016 , y: denotes year 2009

Ho : u_x - u_y <= 0 , Ha : u_x - u_y > 0



```
u_x =  162.6 u_y =  150.26315789473685
sd_x =  29.35 sd_y =  23.96
sigma =  4.94 z =  2.5
p = 0.0063
```



Since p_value = 0.0063 < 0.05

We reject the null hypothesis and conclude average scores in 2016 was greater than that in 2009.

IPL 2009 was hosted by South Africa. Thus we can conclude the pitches in South Africa are less batsmen supportive.This is also evident from the line graph, which indicated that the number of boundaries was least in 2009

CONCLUSIONS

1. It is not true that toss winners win the match
2. There is a high chance that the team scoring 200+ wins the match
3. David Warner's form looks to be improving season by season.
4. Virat Kohli holds the record of most runs in a season 973.There has been a sharp decline in Kohli's Runs from 2016 to 2017.
5. Raina has consistently scored 300+ runs in every season
6. Virat has most number of 1s and is among the top 5 batsman with most 1s,2s,4s and 6s.
7. Gayle has most number of 6s. Gambhir has most number of 4s. Dhoni has most number of 2s.
8. Suresh Raina has scored the most runs in IPL followed by Virat Kohli and Rohit sharma
9. RCB has scored most number of sixes and MI have scored most number of 4s
10. Average runs scored per match by RCB was highest in 2016.RCB has highest team total 263 runs.
11. The most common dismissal type in IPL is caught followed by bowled. There are very few instances of hit wicket as well.
12. Chris Gayle has won the most number of man of the match awards
13. In the initial seasons of IPL, there is no much difference in number of times toss winners chose to bat and field.In seasons 2009, 2010, 2013 batting first was the preferred choice.Whereas in the last few seasons it is clearly seen that the trend is to bowl first.

By hypothesis Testing,

We have come to a conclusion that the pitches in South Africa are less batsmen supportive than those of India. Alongside this we have arrived at multiple other conclusions about the performances of players.