SI 618 PROJECT 2
IPL DATA ANALYSIS

MOTIVATION:

Cricket is a bat-and-ball game played between two teams of eleven players on a field at the center of which is a 22-yard (20-metre) pitch with a wicket at each end, each comprising two bails balanced on three stumps [1]. Right from my childhood watching cricket and discussing about it has always been a part of my life. In India, cricket is not only considered as a game, but it is considered as an emotion. After watching this video on YouTube a few months back I realized how the insights from data can help teams to form strategies to win a cricket match. The main reason why I chose this dataset is because I wanted to apply the skills that I learnt to the data from the league that I enjoy the most. I have used the data from the Indian Premier League(IPL).IPL is a T20 cricket league contested by teams representing different cities in India.

DATA SOURCE:

I have used the IPL Complete Dataset (2008-2020) available on Kaggle. It contains ball by ball data and match wise summary statistic in two separates .csv files for all IPL matches played between the years 2008 and 2020. The ball-by-ball csv files contains about 193468 rows and 18 columns namely of which I will be using the columns of match id, the innings in which the ball was played, the runs scored in the ball and whether a wicket was taken in the ball. The match file contains about 816 rows and 17 columns of which I will be using the columns that contain the id of the match, the venue of the match, the two teams playing the match, the winner of the toss and the winner of the match. For some analysis joining of both the .csv files for would be required.

RESEARCH QUESTIONS:

Main Question:

What are the factors that influence the match outcome for a team in the IPL?

1. Does the team winning the toss have an advantage over the team losing the toss?
    a. Does the team winning the toss have a higher probability of winning the match?
    b. Does the decision made by the team(to field or to bat first) after winning the toss have an impact on the outcome of the match?
2. Does home advantage really play a role in the outcome of an IPL match?
3. How does powerplay impact the outcome of the match?
    a. The distribution of scoring elements in all the innings played by the winning teams vs the in all the innings played by the losing teams.
    b. Association between the outcome of the match for the chasing team associated and the percentage of the runs of the target they score during the powerplay.
    c. Association between the winning probability of a chasing team and the scoring elements during powerplay.

METHOD:

Both the datasets were loaded into the data frames and then some data exploration and preprocessing were done. During the data exploration I saw that multiple names pointed to the same stadium, so I changed them to a unique name. I saw that the name of Rising Pune Supergiant's was being repeated with a few minor changes, so I changed that into a single name. The Delhi Daredevils team was renamed into Delhi capitals after the year 2018 so for easy analysis purposes I'm changing the name to Delhi capitals for all the years. While looking for columns with missing values I found that some rows had missing values in the result and the winner column. Those matches were the matches that were interrupted, and no results were announced, and no winner was declared so I removed these rows. In the project for all the analysis

Submitted by,
Shivani Baskar

SI 618 PROJECT 2
IPL DATA ANALYSIS

I investigate the matches where at least one team has been declared the winner. So, I removed the rows containing the data about tied matches from the match dataset.

Analysis 1:

Manipulation of data: I created a column called "tossandgamewins" in the match dataset that indicates whether the winner of the toss was the winner of the game.

Handling missing data: As mentioned above the missing data was manipulated and the unwanted columns were removed in common for all three analyses.

Challenges encountered: I had some doubts initially on which plot would be the best to visualize the analysis. I then read about that online and then I was able to do it quickly.

Analysis 2:

Manipulation of data: First, I created a column called "home team" in the match dataset that indicates the home team of the venue in which the match was played. This was done by using a function that using the information provided in this webpage. Then I used the function label_homegame() to find whether a particular match was a home game for any of the two teams playing. For each team participating in the IPL I calculated the number of home games-the number of matches in which the team is either team1 or team2 and played in the team's home ground, the number of away games-the number of matches in which the team is either team1 or team2 and is not played in the team's home ground, the number of wins in home game, the number of wins in away games, the total number of wins and the total number of losses.

Handling missing data: As mentioned above the missing data was manipulated and the unwanted columns were removed in common for all three analyses.

Challenges encountered: I had some doubts initially on how to manipulate the data to perform the chi2 test. I then read about that online and then I was able to do the test quickly.

Analysis 3:

Manipulation of data: To calculate the total score of every innings I aggregated the ball-by-ball dataset based on the match id and the innings. According to the rules of IPL the first 6 overs of every inning in the match are powerplay overs. So, from the ball-by-ball dataset I filtered only the rows that had information about the first 6 overs. I then aggregated the data based on match id and innings to find the number of runs scored and the number of wickets lost by the batting team in the powerplay in each inning in each match. Then I selected the columns from match dataset that was needed for the analysis and then inner joined it with the powerplay data based on the match id.I then joined the death over data to the previous dataframe based on id and inning. I then used the functions batting() function to find the batting every inning in every match. I used the function batting_team_result() to find whether the batting team in the inning won the match or not. Now each row of the data frame contains data about the matchid,inning,name of the teams playing the match, winning team, the batting team whether the batting team won or lost the match, the total runs scored by the batting team in the powerplay ,the wickets lost by the batting team in the powerplay, the runs scored by the batting team.

Handling missing data: As mentioned above the missing data was manipulated and the unwanted columns were removed in common for all three analyses.

Challenges encountered: When I initially did the analysis, I found that the number of innings played by winning teams was not equal to the number of innings played by the losing teams. When I pondered into the dataset then I realized that there were some matches in which no winner or no result was declared due to cessation of play. Then I went back to preprocessing and removed these rows after which I got the correct analysis.

Submitted by,
Shivani Baskar

SI 618 PROJECT 2
IPL DATA ANALYSIS

ANALYSIS AND VISUALIZATION:

Analysis 1: Does the team winning the toss have an advantage over the team losing the toss?

Does the team winning the toss have a higher probability of winning the match?

Workflow:

To find the aggregate number of matches that were won and lost when the team won the toss, I aggregated the data from the match dataframe based on tossandgamewins then to visualize the analysis I used a pie plot. Then to analyze how the odds worked out for individual teams I created a crosstabulation in which each row gives the data about the percentage of matches won and lost by an individual team.

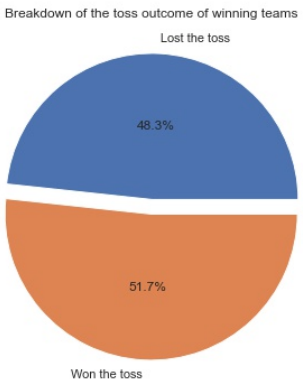


Figure 1:Breakdown of the toss outcome of winning teams

It can be seen from the plot that 51.7% of the times the team that won the toss won the match. This indicates that there is a slightly higher chance for the team winning the toss to win the match.

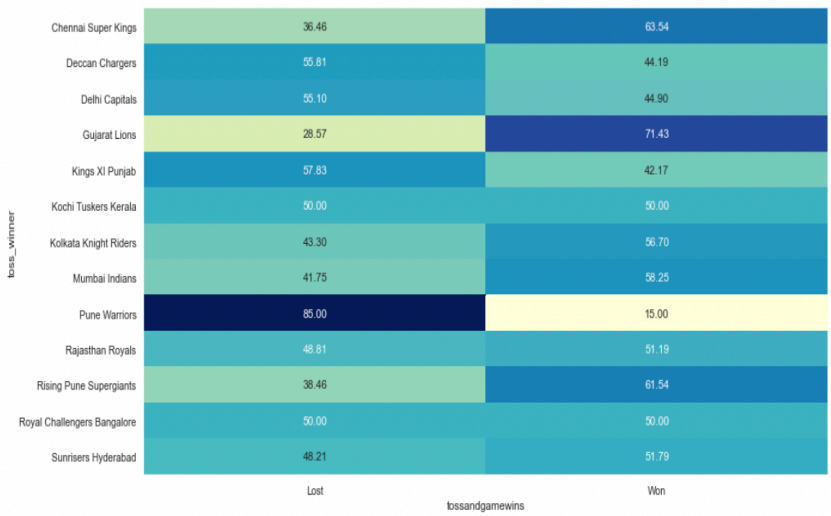


Figure 2:Relationship between the decision made by the individual teams when they won the toss related and their match result

It can be seen from the heatmap that for teams like Royal Challengers Bangalore, Kings Xl Punjab, Delhi Capitals, Deccan Chargers and Pune Warriors had lower odds of winning the match if they won the toss.

Submitted by,
Shivani Baskar

SI 618 PROJECT 2
IPL DATA ANALYSIS

Does the decision made by the team(to field or to bat first) after winning the toss have an impact on the outcome of the match?

Workflow:

To find this I created a mosaic plot to observe the relationship between the decision made by the toss winner and the match outcome for the toss winner. Then I wanted to see how the relationship between the decision made by the individual teams when they won the toss related to their match result. To do this I created a contingency table in which each row gives information about the team the number of matches they won while choosing to field after winning the toss, the number of matches they won while choosing to bat after winning the toss, the number of matches they lost while choosing to field after winning the toss and the number of matches they lost while choosing to field after winning the toss. To find if there is a statistically significant relationship between the match outcome of the team that won the toss and the team's decision to bat or bowl first, I did a chi2 test. The decision that the toss winner makes after the toss also depends on the venue the match is played. To analyze this, I created a similar cross tabulation as above but for each venue.

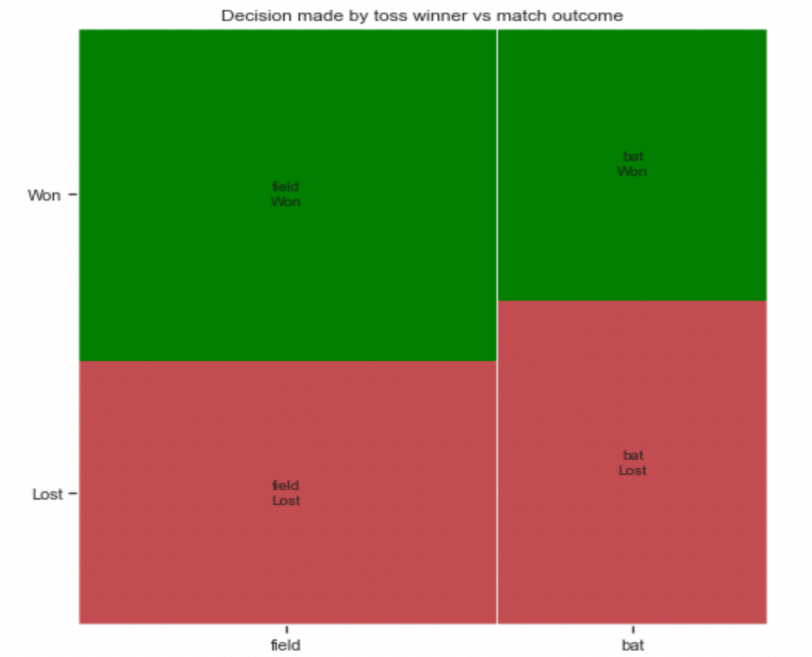


Figure 3:Mosaic plot to observe the relationship between toss and match wins

Most of the times the team that won the toss has decided to field first. This might because if they have a score to chase then they can plan their innings and play. But if they opt to bat first then they might want to set a good target which may or may not be chased by the opposite team. From the mosaic plot it can also be seen that the toss winner wins a lot of matches when they opt to field first.

Submitted by,
Shivani Baskar

SI 618 PROJECT 2
IPL DATA ANALYSIS

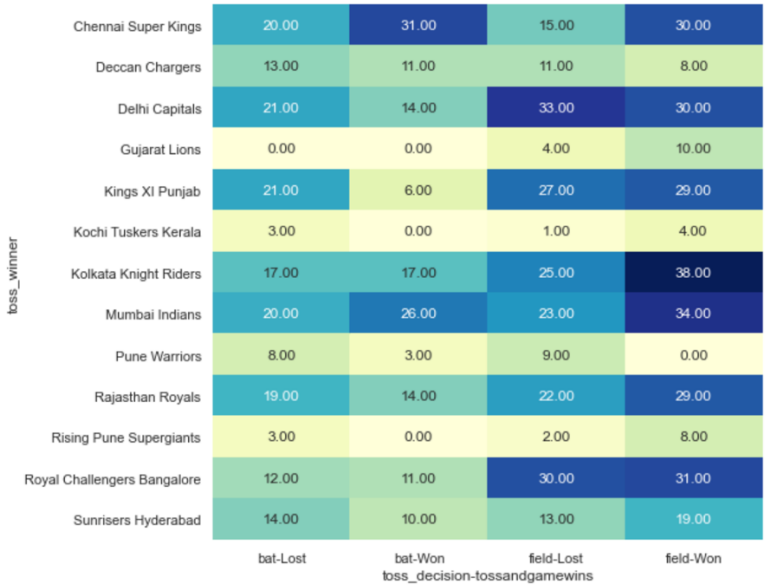


Figure 4:Relationship between toss decision and match wins for individual teams

From the figure except Delhi Capitals all the other teams follow the general result. Delhi capitals have lost a lot of matches when they decided to field first after winning the toss.

The pvalue(0.0064141) obtained as a part of the chi2 test, is lesser than alpha(0.05).This indicates that **there is a statistically significant relationship between the match outcome of the team that won the toss and the team's decision to bat or bowl first.**

toss_decision	bat		field	
tossandgamewins	Lost	Won	Lost	Won
venue				
Barabati Stadium	0	2	2	3
Brabourne Stadium	2	4	2	3
Buffalo Park	1	2	0	0
De Beers Diamond Oval	1	1	0	1
Dr DY Patil Sports Academy	4	3	4	6
Dr. Y.S. Rajasekhara Reddy ACA-VDCA Cricket Stadium	3	2	4	4
Dubai International Cricket Stadium	8	6	11	5
Eden Gardens	16	12	18	31
Feroz Shah Kotla	16	15	19	22
Green Park	0	0	0	4
Himachal Pradesh Cricket Association Stadium	0	1	4	4
Holkar Cricket Stadium	1	0	1	7
JSCA International Stadium Complex	2	1	1	3
Kingsmead	4	6	2	3
M Chinnaswamy Stadium	5	4	29	38
MA Chidambaram Stadium, Chepauk	14	22	12	8
Maharashtra Cricket Association Stadium	1	1	6	13
Nehru Stadium	1	1	2	1
New Wanderers Stadium	2	0	3	3
Newlands	1	3	1	1
OUTsurance Oval	0	1	0	1
Punjab Cricket Association Stadium, Mohali	11	6	18	21
Rajiv Gandhi International Stadium, Uppal	21	6	21	15
Sardar Patel Stadium, Motera	3	3	3	2
Saurashtra Cricket Association Stadium	2	0	3	4
Sawai Mansingh Stadium	13	6	9	19
Shaheed Veer Narayan Singh International Stadium	2	1	1	2
Sharjah Cricket Stadium	4	2	5	7
Sheikh Zayed Stadium	8	6	5	8
St George's Park	4	3	0	0
Subrata Roy Sahara Stadium	6	9	2	0
SuperSport Park	3	3	1	5
Vidarbha Cricket Association Stadium, Jamtha	1	1	1	0
Wankhede Stadium	11	10	25	26

One interesting observation that can be made from this table is that only in the venues of MA Chidambaram stadium and the Subatra Roy stadium most of the times the winner of the toss has decided to bat first. This might be because these grounds are good for hitting runs.

Submitted by,
Shivani Baskar

SI 618 PROJECT 2
IPL DATA ANALYSIS

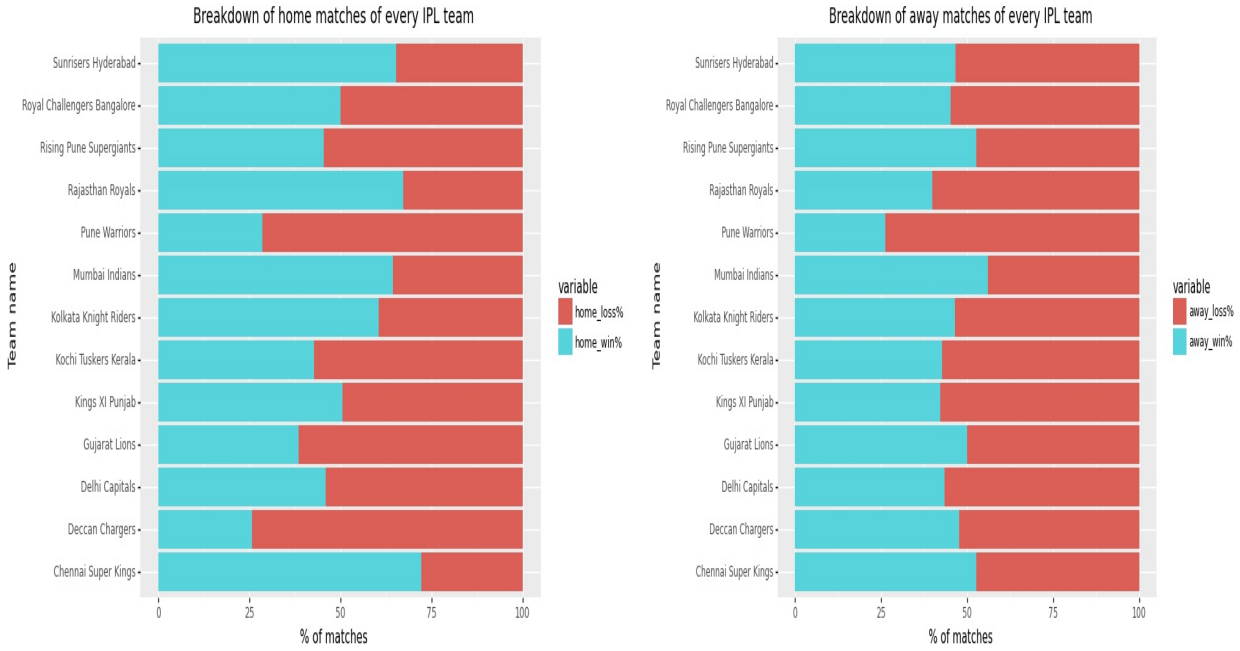
Analysis 2:

Does home advantage really play a role in the outcome of an IPL match?

There is a myth in the cricketing world that says that teams team play well in their home grounds than in away grounds.A sports team's home ground is their own playing field, as opposed to that of other teams. The main purpose of the question was to find out whether there is a statistically significant relationship between an individual team's match outcome(win or loss) and where the match was played(home or away ground).

Workflow:

To do the analysis I aggregated the information in the matches table as described in the method section above. To start with the analysis plotted the breakdown of home matches of all teams using stacked bar plots in plotnine. Then to make the visualization easier ,I converted the columns that had data about the wins to rows using melt() and converted the columns that had data about the home matches and away matches to rows using melt().I then plotted the breakdown of away matches of all teams using stacked bar plots in plotnine. Then to find whether the there is a statistically significant relationship between the team's match outcome(win or loss) and where the match was played(home or away ground for every team I extracted contingency tables for every individual team from the per_team_stats dataframe. Then I used the chi2_contingency() with an alpha value of 0.05 function from SciPy to find the statistical significance.



From the visualization it is evident that, Except for teams like Rising Pune Supergiants,Pune Warriors,Kochi Tuskers Kerala, Gujarat lions and Deccan Chargers who have played only a less number of games as they participated in the IPL only for less number of seasons for all the other teams it can be seen that the number of matches a team has won in the home ground is greater than the number of matches a team has lost in their home ground. This indicates the odds of winning a match while playing on the home ground is high.

Submitted by,
Shivani Baskar

SI 618 PROJECT 2
IPL DATA ANALYSIS

Except for teams like Chennai Super Kings and Mumbai Indians which are the most successful franchises in the IPL as they have won the greatest number of titles for all the other teams the number of matches a team has lost in an away ground is greater than the number of matches a team has lost in an away ground. This indicates the odds of a team losing a match while playing on away ground high.
From the output of the chi square tests it was observed that the teams for which there is a statistically significant relationship between the team's match outcome(win or loss) and where the match was played(home or away ground) are Chennai Super Kings and Rajasthan Royals and the teams for which there is no statistically significant relationship between the team's match outcome(win or loss) and where the match was played(home or away ground) are Deccan Chargers, Delhi Capitals, Gujarat Lions, Kings XI Punjab, Kochi Tuskers Kerala, Kolkata Knight Riders, Mumbai Indians, Pune Warriors, Rising Pune Supergiants, Royal Challengers Bangalore, and Sunrisers Hyderabad.

Analysis 3:
How does powerplay impact the outcome of the match?
The main purpose behind this analysis is to understand the importance of scoring runs and not losing wickets in the first six overs of the game.
The distribution of scoring elements in all the innings played by the winning teams vs the in all the innings played by the losing teams
To visualize the distribution ,I filtered out all the innings in which the batting team was the winner of the match and all the innings in which the batting team was the loser of the match then I plotted histograms with the kernel density estimators for the scoring elements in the powerplay for all the innings played by the winning teams and for all the innings played by the losing teams.

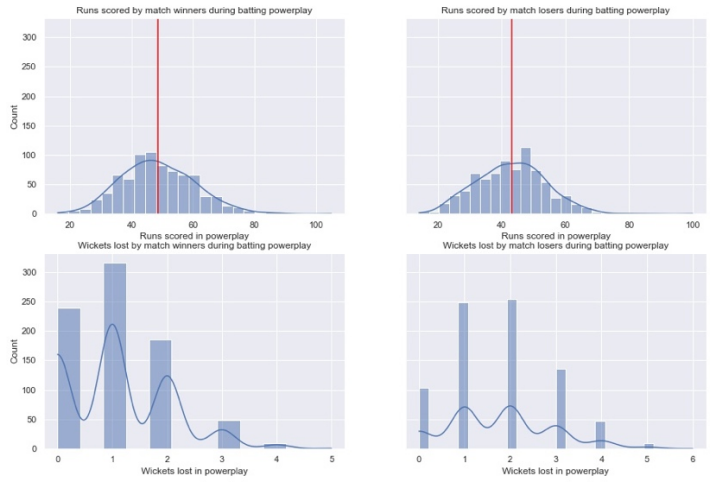


Figure 5:Distribution of the scoring elements(runs scored and wickets lost) in the powerplay in the innings played by the losing team vs in all the innings played by the winning teams in all matches

From the visual it can be seen that the distribution of runs scored during the powerplay in the innings played by the winning teams is very similar to the distribution of the runs scored in the innings played by the losing teams. The average runs scored in the powerplay overs in the innings played by the winning teams is 48 and the average runs scored in the powerplay overs in the innings played by the losing teams is 43.In the innings played by the winning teams it can be seen that the teams have mostly lost

Submitted by,
Shivani Baskar

zero or one wicket in the powerplay and in the innings played by the losing teams the teams have mostly lost 1 or two wickets during the powerplay.

Association between the outcome of the match for the chasing team associated and the percentage of the runs of the target they score during the powerplay

To do this analysis I first calculated the target score for all the chases(target score for chases is the total score of the first innings+1) and then merged it to a dataframe that contained the data only about the chasing innings(second innings).Then I calculated what percentage of the target runs did the team hit during the powerplay. I then categorized the percentage of target runs the team hit during powerplay as less than 30% of the target runs and greater than 30% of the target runs. I chose the threshold 30% because powerplay makes up only 30% of the overall innings(6 overs out of 20 overs). Then I plotted a mosaic plot to find out the association between the percentage of the runs of the target the chasing team scored during powerplay and the outcome of the match for the chasing team.

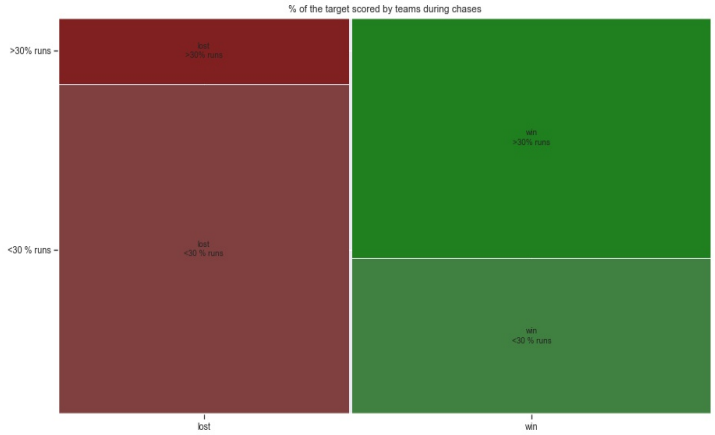


Figure 6:Association between the percentage of the runs of the target the chasing team scored during powerplay and the outcome of the match for the chasing team

It can be seen from the plot that the chasing team has lost most of the matches in which it had hit less than 30% of the target score during the powerplay and has won most of the matches in which it scored more than 30% of the runs of the target during powerplay. The teams that made more than 30% of the target have a high probability of winning the match. This indicates that there is a clear association between the percentage of the runs of the target scored during powerplay and the outcome of the match for the chasing team.

Association between the winning probability of a chasing team and the scoring elements during powerplay:

To do this I first created a cross tabulation that shows the percentage of outcome(win percentage, loss percentage) of the chasing team and the number of wickets lost they during powerplay. Then I drew a line plot to visualize the winning probability. Then categorized the chases based on the number of runs

Submitted by,
Shivani Baskar

they scored during the powerplay and then created a cross tabulation that shows the percentage of outcome(win percentage, loss percentage) of the chasing team and the number of runs they scored during the powerplay.

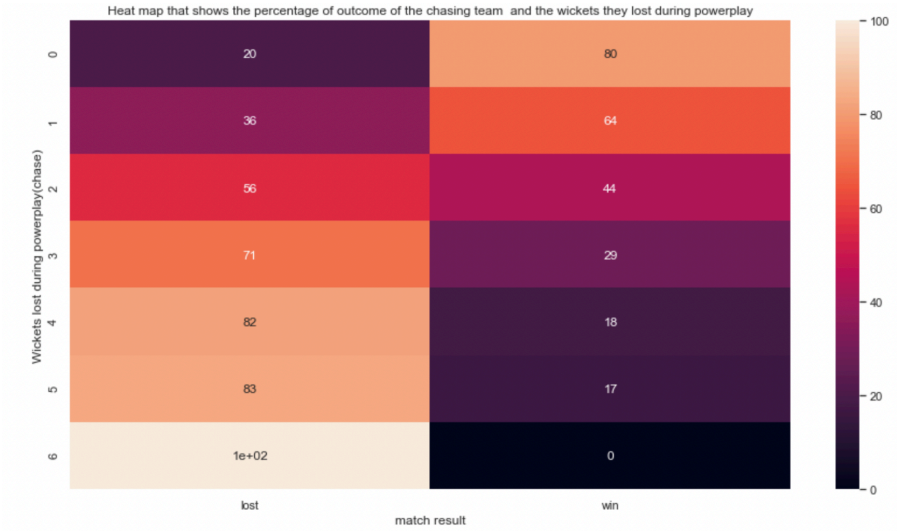


Figure 7:Cross tabulation that shows the percentage of outcome of the chasing team and the number of wickets lost they during powerplay

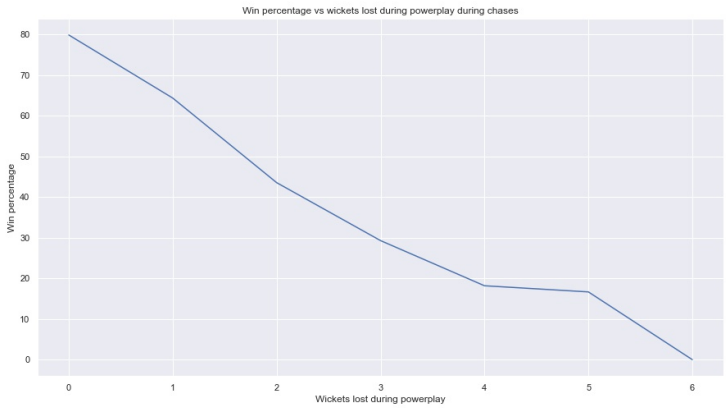


Figure 8:Win percentage vs Wickets lost in powerplay for chases

From the table we can see that the chance of winning a chase is correlated with the number of wickets that fall in the powerplay. The win probability reduces by approximately 15% to 20% with the loss of every wicket during powerplay.

Submitted by,
Shivani Baskar

SI 618 PROJECT 2
IPL DATA ANALYSIS

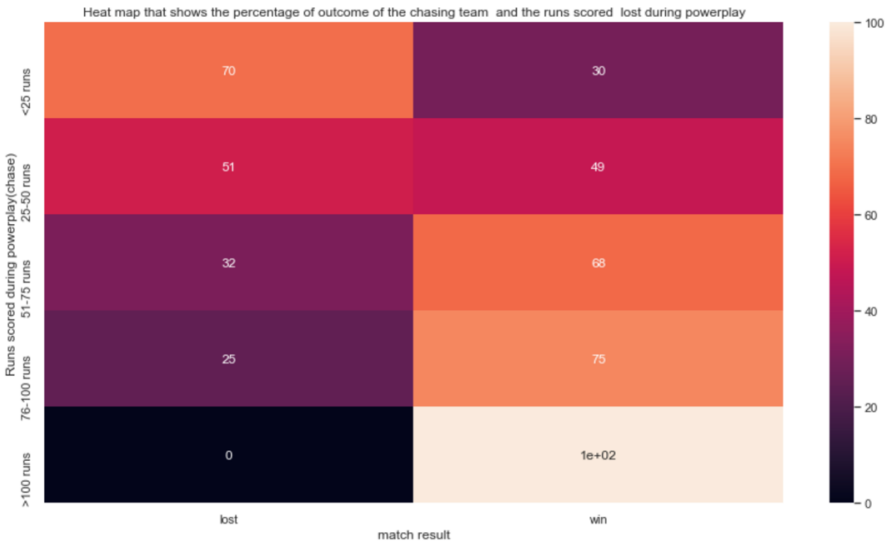


Figure 9:Cross tabulation that shows the outcome percentage for the chasing team and the number of runs they scored during powerplay

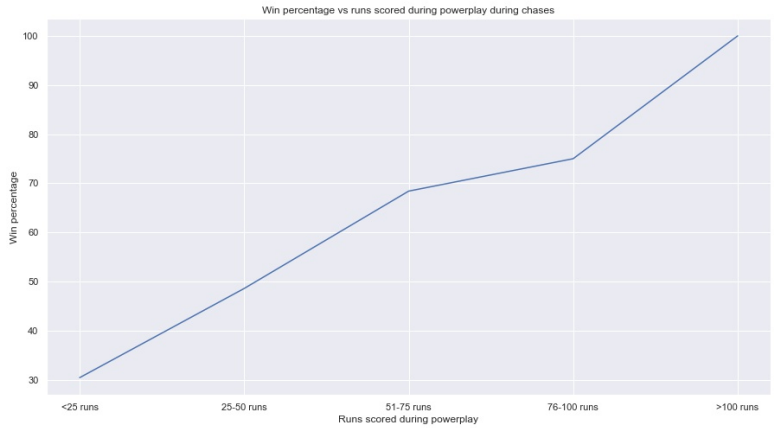


Figure 10:Win percentage vs Runs scored during powerplay while chasing

From the table we can see that the probability of winning a chase greatly increases with the number of runs scored in the powerplay. The win probability increases greatly with every 25 runs scored during the powerplay.

Conclusion:
From the analysis I conclude that the outcome of an IPL Match to an extent depends on the venue, the toss decision and the runs scored during the powerplay.

References:

1. IPL Complete Dataset (2008-2020)
2. https://wordpanda.net/definition/home-ground
3. https://en.wikipedia.org/wiki/List_of_Indian_Premier_League_venues

Submitted by,
Shivani Baskar