**622-FINAL**

**Ashley Dsilva**

Using CRISP-DM methodology for the given problem

Steps in CRISP-DM

1. Business understanding

2. Data understanding

3. Data preparation

4. Modelling/Evaluation

5. Deployment

1. Business understanding:

This analysis focuses on the various attributes of a Cricket tournament over the span of 2008-2020. The analysis can be useful in understanding various factor that come into play to win a match. Various analysis can be done to find out the performance of individual teams and its players. The analysis can also be used to find out which venue was most favorable for a specific team. All of this put together can help the team to perform better and help them fetch more winnings and win the tournament.

1. Data understanding:

Dataset used: Cricket (2008-2020)

The dataset has various sub datasets. Each of which has different attributes like the data about the team performance, individual player performance, individual innings and the details entailing that.

Attributes:

* Id
* City
* Date
* Player of the match
* Venue
* Team1
* Team2
* Toss decision
* Toss winner
* Winner
* Result
* Inning
* Over
* Ball
* Batsman
* Nonstriker
* Bowler
* Home wins
* Away wins
* Umpire1
* Umpire2

Unclean dataset preview

Table

Description automatically generated

Table

Description automatically generated

Table

Description automatically generated

1. Data preparation:

The dataset had a lot of NA values in specific columns, so I dropped those columns as they were of no use to me. Further, the names of some columns needed to be changed for better plotting of the graphs and to avoid error in the syntax. For example, home\_wins, away\_wins were changed to Home wins and Away wins. Next up, was to change the date column into separate day, month and year format for further analysis.

Clean dataset preview:

Table

Description automatically generated

Table

Description automatically generated

Table

Description automatically generated with medium confidence

1. Modelling and evaluation:

For the modelling part various graphs were used to plot and obtain insights to help answer desired questions. I used a mix of plots like bar, line, scatter, pie, and heat map to analyze the data and get valuable insights.

* For the first plot which is a bar plot I want to find out the number of matches played with respect to the season

Chart, bar chart

Description automatically generated

From the plot it can be seen that the greatest number of matches were played in the year 2011-2013, as during this period the tournament gained recognition and so the spike in the number of matches but then there was a sudden drop after 2013 as it did not have any new add-ons.

Using bar chart to answer the question felt the best as it could give specific insights just by looking at it.

* For the next I used stacked/grouped bar plot and I answered the question, which toss decision resulted in the most wins.

Chart, bar chart

Description automatically generated

From the above graph it can be deduced that in over 400 matches it was seen that the decision of opting to field first resulted in the most wins as compared to choosing to bat first. The reason behind this is that the teams that win the toss decide to field most of the times as it is easier to chase a total, because you know what run rate you need to follow in order to win instead of batting first. The stacked bar plot was the best choice as it could give a visual comparison of the toss decision and the results of the match.

* To find out the frequency of matches year-wise I used a line graph

Chart, line chart

Description automatically generated

It is seen that in the year 2013 total of 76 matches were played which is the highest followed by the year 2013 and 2011 with a frequency of 74 and 73 respectively.

* Next plot used is scatter plot to find out the number of runs scored and balls faced by my favorite batsmen

Chart, scatter chart

Description automatically generated

From the graph it can be seen that the most runs scored in the tournament was MS Dhoni followed by Virat Kohli. The fastest hundred was scored by MS Dhoni in just under 60 balls. Except RV Uthappa all the other batsmen scored a century of runs. Scatter plot best fitted to answer the question to visualize the runs scored given the number of balls faced of four different batsmen.

* Next up I wanted to find which team performed the best and had the most wins throughout the tournament. For this I used a pie chart.

Chart, pie chart

Description automatically generated

From the above graph it can be deduced that the team Mumbai Indians performed the best with a win percentage of 15 followed by teams Chennai Super Kings, Kolkata knight riders and Royal Challengers Bangalore with a percentage of 13.1, 12.3, and 11.1 respectively.

Pic chart visuals fit perfectly answering my question in a color coded manner.

* Next I used the heat map graph to find out various insights about all teams performance in the tournament.

Chart

Description automatically generated with medium confidence

From the above heat map, various insights about specific teams can be visualized, considering my favorite team Mumbai Indians, they had more wins playing in their home field compared to playing away. They had the same number of home and away matches. As for the team Kolkata knight riders, surprisingly they had more wins playing away than playing at a home field. Heat map was the fitting choice as it gives a shade wise visualization for different attributes for respective attribute chosen as columns.

* Another heat map was plotted to visualize the different attributes of the top 50 batsman in the tournament.

Chart, bar chart

Description automatically generated

In this heat map we can compare batsman and their performance, G Gambhir faced most number of balls. Most batsman played in the striker position than the non striker position which says they come under the tag of batsman and not a bowler who are sent to bat at the end of the wickets.

1. Results

From the various graphs we plotted and visualizations we carried out for the teams and the players, we can conclude with some insights, the year 2011-2013 was very successful for the tournament after which a bunch of viewers lost interest. In order to win a match it is better to opt to bowl first and then chase the goal as it shows more probability of winning as per our analysis. The top batsman of the tournament was MS Dhoni followed by Virat Kohli. And the team Mumbai Indians played well with highest wins in the tournament.