**Capstone Project Submission**

**Instructions:**

i) Please fill all the required information.

ii) Avoid grammatical errors.

|  |
| --- |
| **Please write a short summary of your Capstone project and its components. Describe the problem statement, your approaches and your conclusions. (200-400 words)** |
| **Problem Statement:**  I have given multiple datasets for analyzing the IPL T20 tournament. My primary task is to explore the datasets and grab meaningful insights from the data like what trend data follows, what useful information we can get. Find out the some specific facts and figures which tells the story of the combined and individual datasets by analyzing them.  **Approach:**  It contains 6 datasets:   1. matches.csv 2. Players.xlsx 3. Deliveries.csv 4. Teams.csv 5. most\_runs\_average\_strikerate.csv 6. teamwise\_home\_and\_away.csv   First of all I have take a look at all the datasets given, exploring all the features available in all the datasets. Then I moved on to data cleaning process. I have analyzed all the datasets individually and find out there are multiple variables/features in which the records are missing, so firstly I have started imputing those missing values and if the values missing are greater than 20% then I drop that particular feature, if not I have imputed those values with respect to their feature. For example, if a variable is categorical in nature then I have imputed the mode value for that particular feature in missing records of that feature.  Then there are some features which requires replacing the records as those values are identical in nature with little bit difference in their spelling. This could impact the results, therefore I have replaced those identical records.  After that I have visualize the data distribution to see if any major change before and after data cleaning with the help of histogram.  **Conclusion:**  I have found out some interesting facts and figures by analyzing and visualizing these datasets which are mentioned below:-   * Total number of matches played over the entire tournament **756**. * Maximum number of runs in **season 2013** . * Average runs scored per match is around **300**. * Team who won the toss maximum number of times is **Mumbai Indians**. * Most commonly decisions made after winning the toss across all seasons if **Field first**. * By visualizing the data distribution we can say that the teams **winning by runs** is in the range of **0-40**. * **Mumbai Indians** is the team who won the match after batting first by winning a toss. * Winning percentage of **Mumbai Indians** is the highest among all the teams across all seasons. * By visualizing the data distribution we can say that the teams **winning by wickets** is in the range of **4-8**. * **Kolkata Knight Riders** is the team with most number of wins after batting second. * Most number of times a team who won the toss wins the match also. * Most number of times a team who won the toss and chose to field first wins the match also. * The top run scorer batsman over the entire tournament is **Virat Kohli**. * The maximum number of times **Player of the Match** award wins by **Chris Gayle**. |
| **Please paste the drive link to your deliverables folder. Ensure that this folder consists of the project Colab notebook, project presentation and video.** |
| https://drive.google.com/drive/folders/1UwUig98FIvtjzz9VLsHIwDhMDh87Fzpb?usp=sharing |