**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)** |
| So what can we expect from the multiple IPL datasets? 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   **Data Cleaning:** As we can see our project had multiple datasets to be cleaned. The way I approached these datasets was, I did a detailed analysis of each dataset and looked up for the content of each column and figured out that there are both numerical and categorical missing values in all the datasets, here I used the mean and the mode method to treat those missing values. Plus I had to drop few columns too which were not a part of our analysis. Apart from this I checked for the unique values in particular column whenever need and replaced it with a specific unique names. Once I was done with all these cleaning and imputing, I plotted the heat maps before and after treating the missing values for all the datasets.  Number of Players from different countries:  Players dataset contains information about player’s country. We can use group players by country to count numbers of player from particular country.  Number of Matches played in each city:  Matches data set contains information about city where the match is hosted. We can count number of matches hosted in the city by grouping the cities data together.  Number of Matches per season:  Matches dataset contains details of season. We can group the matches by its season data and count the number of matches that particular season have.  Number of boundaries:  Delivery dataset contains information about each ball delivery, we can calculate number of 4 and 6 runs from each delivery and group them according to seasons to count number of boundaries per season.  Player of the match:  From matches dataset we can calculate number of time a particular player won the player of the match award.  Number of matches a Team won:  Matches dataset contains information about winner team. We can calculate number of wins for each team by counting the winner team names.  Toss winning:  Matches dataset contains information about toss winning, match winning and toss decision. So we can check if winning a toss affects result of winning the match and we can also calculate what decision is made most frequently when a team won the toss.  Centuries:  Delivery dataset contains information about match id and run for given ball delivery. We can group data by match and then we can calculate each players run individually, from runs we can calculate number of centuries and half centuries. |
| **Please paste the drive link to your deliverables folder. Ensure that this folder consists of the project Colab notebook, project presentation and video.** |
|  |