APPROACH

- **Data Prep:**
  - Quality checks performed / Errors found:
    - **Knockout_match2** column contains value other than 0 and 1. This column should be binary like this first one.
    - **Match_event_id** column has missing values which cannot be predicted.
  - Data preprocessing steps:
    - **Finding Missing values** of various columns. Majorly **Mean** and **Mode** are used to assign values to NA.
    - For Home/Away column, missing values were calculated using **regular expression** and using Lat/Lon column as well.
    - **Data type conversion** for **correlation**
- **EDA:**
  - Feature generation:
    - Converting factor variables to **dummy variables**.
    - Using Colum1 for shot_id_number because it had missing values.
    - Separate out Latitude and Longitude variables.
  - Exploratory data analysis:
    - **Mean** and **Mode** of various variables.
    - **Correlation Matrix**
    - **P-Value**
- **Model building**:
  - Model Choice:
    - Since we had to calculate probability of goal, we had to use regression and in that **Logistic Regression**. So, I have used **Generalized Linear Model (GLM).**
    - **Dimensionality reduction** is been done using **PCA**
- **Conclusion**:
  - Important Features*:*
    - Knockout_Match
    - Power_of_shot
    - Remaining_min
    - Home.away
    - Range_of_shot