Problem Statement

Cricket is a religion in our country. It has become a craze after the advent of IPL. Over the years IPL has gone from just a game to a matter of pride for fans as well as owners. Every penny which goes into auctions, planning, training, marketing and broadcasting the matches needs to be proved its worth. This calls for data driven analysis and strategies to come up with the best plans for teams (with the goal of lifting the IPL for every team)

Data Science offers great promise towards answering some of the pertinent questions teams and owners may have which could help them design the best teams possible with the limited budgets they have. Some of the questions could be :

- Which are the most explosive batsmen?
- Which are the most consistent batsmen?
- Which overs are best suited for charging the bowlers?
- Which batsmen need to be put against which bowlers for maximum returns (in terms of no. of runs)?
- Which bowlers have the best consistency?
- Which batsmen are more vulnerable to spin?
- What combination of bowlers should be used in the beginning spell?
- Which pitches are more batsmen/bowler friendly?
- Which batsmen are bunnies for a given bowler?

And the possibilites are just endless!!

You are a Data Analyst hired by Bangalore Royal Challengers which is struggling badly at IPL for past few seasons..

Your job would be to devise questions, metrics, dimensions concerning the given problem statement, collect, clean and process the data and in the end build a dashboard which would help RCB to gather actionable insights which would in turn help them come up with strategies to form the best team, win against teams and hence add value to the franchise

Please Note :

- Go through the questions below and solve using **Excel ONLY**
- Please ensure that you include all your worked out files in a folder, zip the same and upload the same as attachment while submitting
- In the absence of worked out files, your submission will stand **INVALID**

Question 1

- What is the **granularity** of the dataset **ipl_matches.xlsx**?
- What does **each row** represent in business terms?
- Repeat the above steps for :
    - **ipl_batting.xlsx**

Answer :

Granularity of the dataset **ipl_matches.xlsx:**

The level of detail is on matches level i,e. each granular record contains details about each individual match that were ever played in IPL.

Bussiness term of each row in **ipl_matches.xlsx:**

As it's granularity is match level, each individual row represents details of each individual match such as date, ground, both teams, toss result (who won and their respective choice) and win result(winning team, winning mode and margin)

Granularity of the dataset **ipl_batting.xlsx:**

The level of detail is on inning level i,e. each granular record contains details about each individual innings that were ever played in IPL.

Bussiness term of each row in **ipl_batting.xlsx:**

As it's granularity is innings level, each individual row represents characteristics of each individual innings such as match number, both team's name, batsman's name, wicket status(how he was out or out or not out etc), Runs scored, Length of a inning in terms of balls, maiden overs, No. os Fours and Sixes, his strike rate, Player's special role(normal player, captain, vice captain etc.) and match date.

Question 2

- Great, now since you have figured out the granularity go ahead and create the unique key for the below 2 datasets in the given format :

1. **ipl_matches.xlsx** : A combination of **date** (in **YYYYMMDD** format) + **team_code**
    - Example : if the **match_date** is **18-04-2018** and the **match_winner** is **Chennai Super Kings** then the **match_key** would be **20180418CSK**
2. **ipl_deliveries** : Same as above

Question 3

- Perform the mentioned data quality checks
    - For **ipl_batting.xlsx** :

        - Fix the **Ground** column for non-standard names
        - Check and fix duplication (if any)
    - For **ipl_matches.xlsx** :
        - Fix the match_date column for non-conformity
        - fix the R and SR column for missing values
- Write down an explanation of how you treated missing values

---

Answer :

Data checks in **ipl_matches.xlsx:**

Changed Ground: **Bengaluru, Banglore** to **Bengaluru**

Changed Ground: **Mohali, Chandigarh** to **Mohali**

3 duplicate records found:

2 records of **20080502CSKDC**, 1 deleted.

2 records of **20080517RRRCB**, 1 deleted.

3 records of **20090425CSKKKR**, 2 deleted.

Data checks in **ipl_batting.xlsx:**

Changed date format from **MDY** to **DMY** for non conforming dates to create proper heirachical data.

Handling missing **Runs**:

Created new column named **R_without_missing.**

If runs from column **R** are missing derive those runs details from column **SR** & column **B**(balls played). Else Derive runs from column **R**.

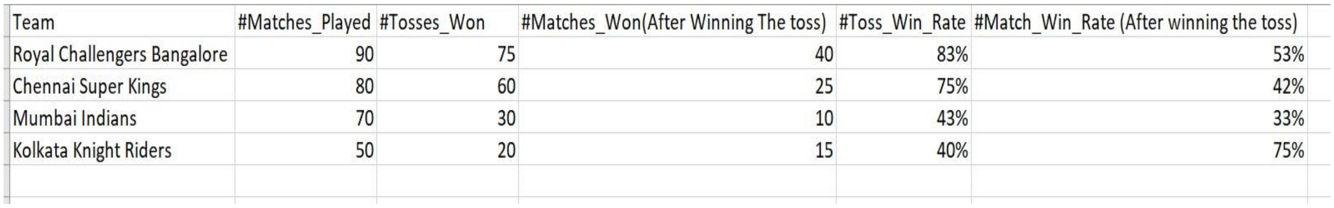
Handling missing **SR**:

Similar to missing runs. Derive from runs and balls played.

Question 4

- Use the dataset : **ipl_matches_long.xlsx**
- It is said : **The side winning the toss wins half the match**

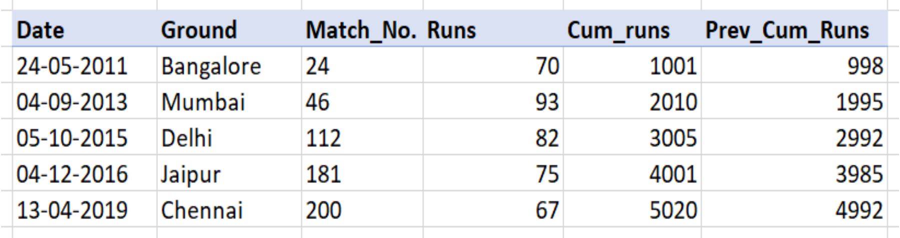
- This means is a side wins the toss and loses the match, we can say the team did either did not leverage the opportunity well or did not read the conditions properly and took a wrong decision to start with.
- Let's validate this with data
- Your goal would be to create team wise below output :



Team	#Matches_Played	#Tosses_Won	#Matches_Won(After Winning The toss)	#Toss_Win_Rate	#Match_Win_Rate (After winning the toss)
Royal Challengers Bangalore	90	75	40	83%	53%
Chennai Super Kings	80	60	25	75%	42%
Mumbai Indians	70	30	10	43%	33%
Kolkata Knight Riders	50	20	15	40%	75%

Question 5

- We are interested in knowing various **landmark** innings for **Virat Kohli** in IPL
- Perform the following activities :
  - **Merge** the datasets **ipl_matches.xlsx** and **ipl_batting.xlsx**
  - **Filter** the result for **Virat Kohli**
  - Create a new column called **Cum_Runs** which shows the cumulative runs as on date (**Hint** : Recall **window operations**)
  - Create a new column called **Prev_Cum_Runs** which shows the cumulative runs till the previous played match
  - Create a new column which sets a flag value of **1** when runs accumulated crosses **1000,2000,3000,4000 or 5000** in a match else **0**
  - **Filter** out all the matches for those innings where **Virat Kohli** has reached a landmark
  - **Expected Outcome : (sample output)**



Date	Ground	Match_No.	Runs	Cum_runs	Prev_Cum_Runs
24-05-2011	Bangalore	24	70	1001	998
04-09-2013	Mumbai	46	93	2010	1995
05-10-2015	Delhi	112	82	3005	2992
04-12-2016	Jaipur	181	75	4001	3985
13-04-2019	Chennai	200	67	5020	4992