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 **PowerQuery (Within 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

- Use the **Data Preview** feature of **PowerQuery** to find out the columns having missing values in the dataset **ipl_batting.xlsx**
- Write down the names of the columns
- Perform the **column profiling** on entire data (instead of default 1000 rows)
- Write down the **exact %** of missing values in an **excel sheet** as shown below :

Column_Name	Total_Row_Count	Missing_Row_Count	Missing_%
col1	11500	70	0.61%
col2	11500	120	1.04%
col3			

- **Treat** the columns for **missing values** (You have already done this in **part 1** of the **Capstone**. You may repeat them using **PowerQuery** this time)

Question 2

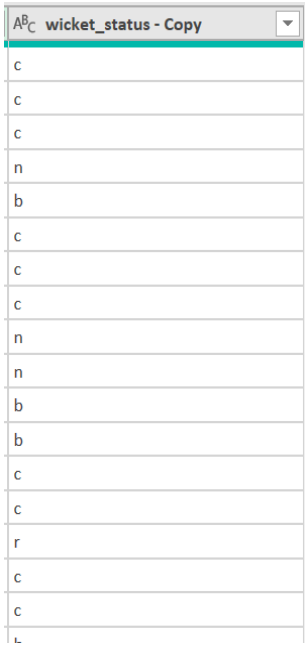
- Your Team's **batting coach** would like to do an assessment of **performance of openers**. He has following questions in mind :
    - Are the openers making **enough run**s to help the team set **strong targets** or help in the chase?
    - Are they maintaining **good run rate** (until the first wicket falls down)?
    - Are they making runs at a **fast pace** (Strike rate)
    - Where do our openers stand **compared** to other teams?

Let's do the following **activities** to make an attempt to answer some of the questions asked above :

1. **Ingest** the below datasets in **Power Query**
    - **ipl_batting.xlsx**
    - **batting_position.xlsx**
2. Find and treat **missing values**
3. Create a new column called **match_player_key** by concatenating match_key and player_key. Sample column output : **20080418KKR1190**
4. Repeat **steps 2 & 3** for the query **batting_position**
5. Merge the queries **ipl_batting** and **batting_position** on **match_player_key** column
6. **Note :** Keep only **Team**, **Innings** & **Position** columns from the query : **batting_position**

Question 3

- Good job with the previous Question. Let's proceed.
- We are concerned with the performance of the **opening pair**
- Let's prepare a new column called **is_Opening**. It should have values **Opening** if the batting position is **1 or 2** otherwise **Others**
- Also, it is expected that the opening pair (especially in IPL) would contribute as much as possible in boundaries
- Create another **new column** calculating the runs scored in **boundaries** (4s and 6s)
- The opening pair has a delicate job of making runs at fast pace as well as making good amount of runs on an average. (Look up the web for the definition of batting average)
    - Use the column **wicket status** to extract the very first character of the string. The outcome may look something like below :



wicket_status - Copy
c
c
c
n
b
c
c
c
n
n
b
b
c
c
r
c
c

    - Create a new column called **dismissal_type** to map the single characters to meaningful names like : **c** for **CAUGHT**, **h** for **HIT WICKET**, **b** for **BOWLED** etc.

Question 4

- Really **superb job** guys till now! Let's push ourselves a bit more...
- Create a summarized query at a **Team** & **Is_Opening** level to calculate the below metrics :
    - **Total runs scored**
    - **Batting Average**
    - **Strike Rate**
    - **Boundary Contribution**
- Create a **summarized query** as shown below where Run% is the contribution of runs scored by opening pair of total

Team	Opening	Others	Run%
Team1	737	720	50.58%
Team2	182	203	47.27%
Team3	577	675	46.09%
Team4	1496	1841	44.83%
Team5	1863	2323	44.51%
Team6	669	915	42.23%
Team7	340	480	41.46%
Team8	1348	2003	40.23%

- What rank does **RCB** hold in terms of run contribution by openers?
- Create another summarized query just like above (this time for **average**)
- Create a column to calculate **ratio of averages** of openers vs others and sort with this column
- What's the **rank** of RCB here?
- Do you think RCB has done well or not in their batting?
