# Important Criteria

## Step 1: Choose a Dataset

1. Atleast 1 GB data
2. Data should be structured or easily converted
3. Check column types and values
4. Atleast 3-4 tables out of dataset and should be connected via some Foreign Keys
5. Check data quality shouldn’t have a lot of missing data (Null)
6. Business Domain related dataset

## Step 2: Business Understanding

1. Ask yourself questions about the dataset that you have chosen, at the end you should achieve a couple of goals:

1. Why this data has been gathered?

2. What can be done with data? What can we achieve?

3. What are some of the goals/targets we have regarding the business that we can achieve by investigating this data?

4. What insightful information can this data provide us that can be used to improve the business?

5. Why are we studying this data? 6. Are there any problems in our business (based on the given data)?

7. Can we find any solutions to these problems by studying this data?

8. What are some of the things we can optimize/improve in our business by studying this data?

## Step 3: Data Understanding

* Understand the data by each column, data type of columns, range of values in each column
* If required change the name of columns for better understanding.
* If there are significant number of missing values, then what is the percentage of such columns.
* Is there any duplicate data?
* Do you believe there are outliers in the data?
* Give simple statistics of data in each column. For example: range, mode, mean, median, variance, counts (frequency)
* Describe what these values mean, especially if you found something interesting
* Try to understand the relationships between the columns of the data. What relationships can you find between the columns?
* Are there any functional dependencies in the data?