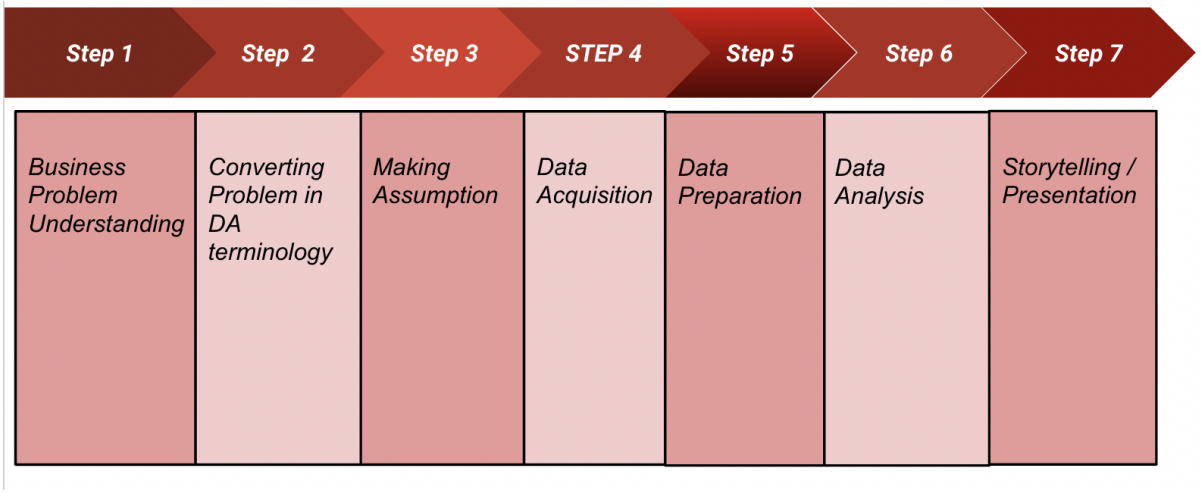
Data Analysis Framework

As we have seen, there is a way, and by following it step by step, we can convert our data into insights based on which we can make decisions. Those decisions help companies to grow and to make more profit.

There are broadly 7 steps to perform Data Analysis.



1. **Business Problem Understanding** :
    - Discussion with the SAMe, Stakeholders or Domain expert to understand the problem
    - Understanding of all the deliverables
    - Getting the idea of the kind of task

2. **The Problem in Data Analysis Terms**: Once a business problem is finalised and understood, then the problem needs to be converted into a data analysis problem.
    - Asking the right question about the data
    - Converting the problem in terms of what kind of data is required.
    - Getting a snapshot of the data so that the data acquisition can take place.

3. **Business Problems Assumptions :**
    - The assumption is created to decide any threshold to get the correct result in terms of business problems.
    - Assumptions are decided after the discussion with domain experts and SMEs.

4. **Data Acquisition :**
    - Once the problem is fully understood in data analysis terms and all the assumptions checked and validated, now comes the data acquisition phase.
    - Now, analysts need to decide which data sets are necessary to solve the problem, where these data sets would be residing and which format. Often, analysts might need help from the DBA's and other data engineering teams to understand the availability of data.

5. **Data Preparation :**
    - Once the data is finalised, now data needs to be prepared in the format on which analysis can be performed.
    - This particular step is the combination of multiple steps.
        a. Data Merging
            - If data is residing in multiple files, now to perform analysis if we require those data to be at one place Data analyst will perform merging.
            - To perform data merging, both these files should have a common column/attribute by which they can be merged.
        b. Data Type Validation
            - All the features should be in an expected format like all the date columns/features like birth date, transaction date etc., need to be in date format.
            - The correct format helps in the analysis as well as in feature engineering methods.

c. Handling the Missing Values
    - As per the data concern, there are chances some data is not available for any attribute. That data is known as missing values.
    - If that particular attribute which is having missing values is not required for our data analysis, we can ignore those values, but if that particular column is required for the analysis, we need to handle those missing values.
    - To handle these missing values, there are various method lines: deleting those rows(if missing value's rows are too less in number if we compare to the whole dataset), putting zero or replacing the value using the statistics method. (all these methods depend on the dataset and the need of the attribute in the analysis)
    - Handling this missing value is required as it impacts the final analysis, giving us wrong insight if data is not available.
d. Handling Outliers
    - Outliers are the data points or values which do not fall into all the other values. For example, if all values in an attribute are between 1 to 100 and there are two values which are 2000 and 3000, then these two will be outliers.
    - Outliers can be an error or exceptional cases in any particular data.
    - We need to handle these types of values as it impacts the result of the analysis. These kinds of outliers can impact mathematical computation.
    - We can handle these by removing those rows or by bucketing methods.

e. Feature Engineering
    - This is more of an art than science and heavily related to the domain, as there can be impactful new features that can be derived from already existing features.

- These new features can be helpful in analysis.

6. Data Analysis:

- There are various methods through which analysis can be done on the processed data.
- Statistical methods, regression or machine learning algorithms can be used.
- In this step, all four types of analysis occur, and at the end of the analysis, DA comes up with the solution for the business problem.

7. Storytelling / Presentation:

- It is the last step of the framework.
- After completing the analysis, the results need to be presented to the stakeholders.
- These results should be in such a way they should be understandable to the stakeholder, so it is always good to have them in pictorial form or less technical content.

Following are additional resources that can help you in the learning framework:

Resource 1

Resource 2