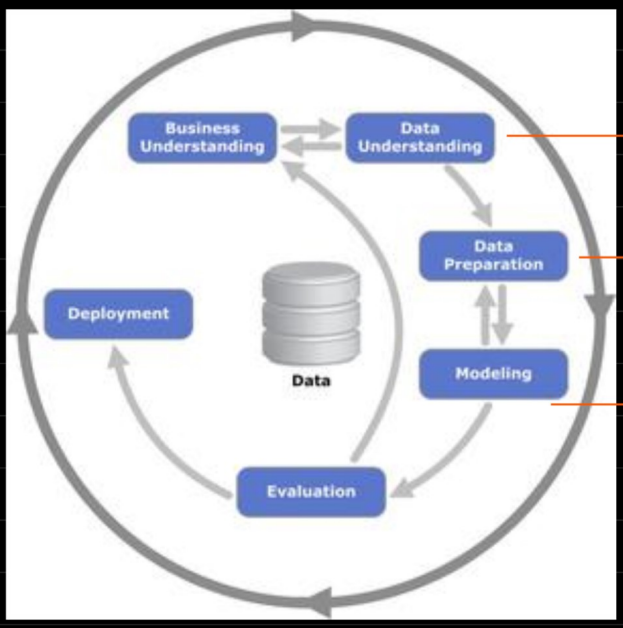
EDA - Exploratory Data Analysis

CRISP - DM

Cross Industry standard process for Data mining.

Data Preparation:

- missing value handling
- Class imbalance
- Data encoding
- Outlier treatment



→ Exploratory data Analysis

→ Missing value treatment
Data encoding
Outlier detection

→ ML Algorithms

Journey of DS project

① Data ingestion (.csv, .tsv, .xlsx)

② Exploratory data Analysis

③ Data Preparations / Feature Engineering

④ Model Building

⑤ Model Evalution.

EDA

- → Profile of data (No of rows, columns, data type, duplicate)
- → Statistical based Analysis. (five point summary, Outliers, correlation, covariance)
- → Graph based analysis (Scatter plot, Heatmap, pairplot, Box plot)

⇔ Once you understood the data well then
↓
→ Select relevant features
→ Convert to numerical.

How to do EDA ??

$(f_1\ f_2\ f_3\ \dots\ f_{100})$

To gain insights from the data — EDA

- Uni variate analysis
  - 1 feature
  - eg. distplot
- Bi-variate analysis
  - 2 features at a time
  - eg. Scatterplot
- Multivariate analysis
  - Multiple features at a time
  - eg. Heatmap, pairplot.

≠ datasets

- Google playstore
- HR analytics
- Cronic
- Travel
- Student performance
- US- Visa
- flight

19M → 19 × 1024

1MB = 1024 KB