**What is EDA? What are the steps usually taken to do this?**

Get an understanding of the data in the context of the domain/problem we are meaning to solve, questions we are meaning to answer. What are the various features that are available? What insights can be drawn? What questions can be answered? We use various statistical and visualizations techniques to achieve this.

Identify the features and the target variable.

**Data cleaning**

* Remove unnecessary/irrelevant data
* Is the data complete? If data is missing, then we may have to impute the data (wherever possible and valid)
* Is all the data in a common format? If not, adjustments (conversions, standardizations) must be done

**Data Analysis**

* Univariate analysis – how individual features
* Bi/multi-variate analysis – correlational coefficients – how multiple features behave wrt each other
* Any outliers
* Numerical, categorical

The result of this exercise would also to be provide an input to the subsequent modelling phase.

**Are missing values always missing due to chance alone?**

There could be human errors if the data collection is manual. For example, if feedback is being collected in a written form and this data is manually entered in digital format. There could be errors in data transfer. The handwriting is unclear, so the field was left empty.

There could be bugs in the software – rare but possible.

Example: Shared cab dataset. User chooses not to provide certain optional or sensitive data like age, gender, income etc.

Example: ICU patient dataset. It might not be possible to take some tests or patient might not be in a position to give the personal information (age, ethnicity, origin country etc.)

Example: Student marks dataset. Students might not have given the test and hence few marks are not available.

So, it is not always chance that certain values are missing. It could be out of users’ choice or it is just that data is not available due to which the field has to be left empty.

MCAR, MAR etc

**What is the approach to missing data?**

* If lot of data (50% or more) is missing for a column, then we may ignore the column for our analysis. Missing due to chance
* If many columns (50% or more) are missing for few rows, then we may ignore those rows. Missing due to chance
* If target value is missing, the rows can be ignored.
* In some cases, we may decide to impute the data based on the other values in the same column (mean/median for numerical data, mode for categorical data) or other columns (predictive models) whichever makes the most business sense. For example, XXX
* If imputing the values does not make sense or there is no reliable way to impute the data, we may leave it empty. For example, XXX

D**oes a correlation coefficient of 0 between 2 numeric variables mean no relationship between them?**

Correlation Coefficient represents the **linear** relationship between two numerical variables. Typically, it falls between the values -1 and 1. A +1 indicates that there is a strong positive correlation between the variables. A -1 indicates that there is a strong negative correlation between the variables. A correlation coefficient of 0 only means that there is no linear relationship. But it is possible that there is a non-linear relationship between the 2 numerical variables. We cannot rule that out

Example

**It is always observed that the more the ice creams sold in a city, higher are the number of murders in a city i.e., there seems to be a high correlation between the ice creams sold and the murders committed in a city. Can we say that because of higher ice cream sales there are higher number of murders in the city?**

Correlation always does not mean there is a causation. While correlation means there is a relationship or pattern between the values of two variables, Causation means that one event causes another event to occur.

Two variables can be strongly correlated but we cannot conclude merely based on the correlation that there one variable causes the other. To decide whether one occurrence of one variable results in the occurrence of the other, there should be stronger evidence.

Example: Ice cream sales and people killed due to shark attacks both have a strong correlation. But it does not mean that higher ice cream sales cause shark attacks or vice versa. In this case, it was a third common reason - summer – that resulted in both – higher ice cream sales and higher people going to beaches & hence more people getting killed by shark attacks.

**How do we confirm causation between the variables?**

Causation can only be determined from an appropriately designed experiment.

Experimental studies

Observation studies

**Given a pair of numerical variables with n observations, how will you visualize it? In what case would a line plot not be suitable?**

Box plots, histograms are suitable for individual numerical variables.

Scatter plots are best suited for a pair of numerical variables.

Line plots are best suited for sequential data i.e. one the variable is of type datetime

If we bin the numerical values (convert them to categorical values), then bar charts would be suitable

**While dealing with multi-dimensional data, how can we visualize more than 2 variables (3 or 4 variables) in 2 dimensions without using dimensionality reduction techniques?**

Heatmaps are suitable for 3 variables

Scatterplots (bubble plots) are also suitable for 3 variables where the size of the dot can represent the 3rd dimension

Color, size can be used to convey multiple dimensions

A single scatterplot can be divided into 2 scatterplots based on the facet.

**What are the different ways outliers can be handled?**

Outliers are data points that are far away from most of the data points

Various methods of visualizing/identifying outliers – boxplots for a univariate analysis, scatterplots for bivariate analysis

If the outliers are not normal and erroneous, then these can be either deleted or treated as missing values and imputed.

If the outliers are valid values, we could either bin them or cap them to a specific value

If we are looking for typical values (like a typical salary of a group of people), we should not include the outliers. We can use median for that.

If we are looking for average values (like average cost of an employee to a company), we should include the outliers. We should use mean in that case.

**What would you prefer as a measure of variability in the data – “standard deviation” or “Interquartile Range (IQR)”?**

Standard deviation indicates how the data points are spread wrt the mean. It considers all the values (including outliers if any) in the dataset.

IQR represents the spread of the central 50% of the datapoints. It does not include the outliers in the dataset.

If the data contains extreme values, then IQR is a better metric to consider.

If the data does not contain extreme values, then Standard deviation is a better metric to consider.