Exploratory Data Analysis

In data preprocessing, EDA gets most of the effort and unavoidable steps. Exploratory data analysis is the first and foremost step to analyze any kind of data. Rather than a specific set of procedures, EDA is an approach, or a philosophy, which seeks to explore the most important and often hidden patterns in a data set.

1. Data sourcing
2. Data cleaning
3. Univariate analysis
4. Bivariate analysis
5. Derived metrics

1. Data sourcing

To solve a business problem using analytics, you need to have historical data to come up with actionable insights. Data is the key — the better the data, the more insights you can get out of it.

2. Data cleaning

There are various types of quality issues when it comes to data, and that's why data cleaning is one of the most time-consuming steps of data analysis.

Though data cleaning is often done in a somewhat haphazard way and it is too difficult to define a 'single structured process', we will study data cleaning in the following steps:

1. Fix rows and columns

Checklist for Fixing Rows:

- **Delete summary rows:** Total, Subtotal rows
- **Delete incorrect rows:** Header rows, Footer rows
- **Delete extra rows:** Column number, indicators, Blank rows, Page No

Checklist for Fixing Columns:

- **Merge columns for creating unique identifiers if needed:** E.g. Merge State, City into Full address
- **Split columns for more data:** Split address to get State and City to analyze each separately
- **Add column names:** Add column names if missing
- **Rename columns consistently:** Abbreviations, encoded columns
- **Delete columns:** Delete unnecessary columns
- **Align misaligned columns:** Dataset may have shifted columns

2. Fix missing values

- **Set values as missing values:** Identify values that indicate missing data, and yet are not recognised by the software as such, e.g treat blank strings, "NA", "XX", "999", etc. as missing.
- **Adding is good, exaggerating is bad:** You should try to get information from reliable external sources as much as possible, but if you can't, then it is better to keep missing values as such rather than exaggerating the existing rows/columns.
- **Delete rows, columns:** Rows could be deleted if the number of missing values are insignificant in number, as this would not impact the analysis. Columns could be removed if the missing values are quite significant in number.
- **Fill partial missing values using business judgment:** Missing time zone, century, etc. These values are easily identifiable.

3. Standardize values

- **Standardize units:** Ensure all observations under a variable have a common and consistent unit, e.g. convert lbs to kgs, miles/hr to km/hr, etc.
- **Scale values if required:** Make sure the observations under a variable have a common scale
  (https://github.com/prajwala0/MachineLearning/blob/main/N2_1_Feature_Scaling.pdf)

- **Standardize precision for better presentation of data:** e.g. 4.5312341 kgs to 4.53 kgs.
- **Remove outliers:** Remove high and low values that would disproportionately affect the results of your analysis.

4. Fix invalid values

A data set can contain invalid values in various forms. Some invalid values can be corrected.

- **Encode unicode properly:** In case the data is being read as junk characters, try to change encoding. E.g. CP1252 instead of UTF-8.
- **Convert incorrect data types:** Correct the incorrect data types to the correct data types for ease of analysis. E.g. string to numeric values
- **Correct values that go beyond range:** Correct the values are beyond logical range, E.g. temperature less than -273° C (0° K),
- **Correct values not in the list:** Remove values that don't belong to a list. E.g. strings "E" or "F" in a data set containing blood groups of individuals,
- **Correct wrong structure:** Values that don't follow a defined structure can be removed. E.g.12 digit pin codes of Indian cities
- **Validate internal rules:** If there are internal rules such as a date of a product's delivery must definitely be after the date of the order, they should be correct and consistent.

5. Filter data

- **Deduplicate data:** Remove identical rows, remove rows where some columns are identical
- **Filter rows:** Filter by segment, filter by date period to get only the rows relevant to the analysis
- **Filter columns:** Pick columns relevant to the analysis Aggregate data: Group by required keys, aggregate the rest

3. Univariate analysis

It deals with analyzing variables one at a time. The agenda of univariate analysis is to understand:

- **Metadata description** : Information such as the size of the data set, how and when the data set was created, what the rows and variables represent, etc. are captured in metadata .
  Types of Variables :
  - Ordered ones have some kind of ordering. Example : Salary = High-Medium-low
  - Unordered ones do not have the notion of high-low, more-less etc.
- **Data distribution plots** reveal interesting insights about the data. You can observe various visible patterns in the plots and try to understand how they came to be.
- **Summary metrics** are used to obtain a quantitative summary of the data.

Segmented Univariate Analysis

The broad agenda of “Segmented Univariate Analysis” is as follows:

- **Basis of segmentation**
  The entire segmentation process can be divided into four parts:
    - Take raw data
    - Group by dimensions
    - Summarize using a relevant metric such as mean, median, etc.
    - Compare the aggregated metric across groups/categories
- **Comparison of averages**
- **Comparison of other metrics**

4. Bivariate analysis

- Bivariate analysis on continuous variables
  - Correlation is a metric to find the relationship between the variables .
- Bivariate analysis on categorical variables
  - To see the distribution of two categorical variables.
  - To see the distribution of two categorical variables with one continuous variable.

Multivariate Analysis

Multivariate analysis looks at more than two variables. A heat map is widely used for Multivariate Analysis. Heat Map gives the correlation between the variables, whether it has a positive or negative correlation.

5. Derived metrics

There are three different types of derived metrics::

- **Type-driven metrics**
    - **Nominal variables**: Categorical variables, where the categories differ only by their names; there is no order among categories, e.g. color (red, blue, green)
    - **Ordinal variables:** Categories follow a certain order, but the mathematical difference between categories is not meaningful, e.g. education level (primary school, high school, college)
    - **Interval variables:** Categories follow a certain order, and the mathematical difference between categories is meaningful, e.g. temperature in degrees celsius
    - **Ratio variables**: Apart from the mathematical difference, the ratio (division/multiplication) is possible, e.g. sales in dollars ($100 is twice $50)
- **Business-driven metrics :** It is derived from the existing variables but it requires domain expertise.
- **Data-driven metrics :** Data-driven metrics can be created based on the variables present in the existing data set. For example, BMI instead weight and height

Categorical Data Encoding

- Integer Encoding : Each unique label is mapped to an integer
- One Hot Encoding : Each label is mapped to a binary vector
- Learning Embedding : A distributed representation of the categories is learned