Gaurav Topre Follow
Feb 18 · 3 min read · Listen

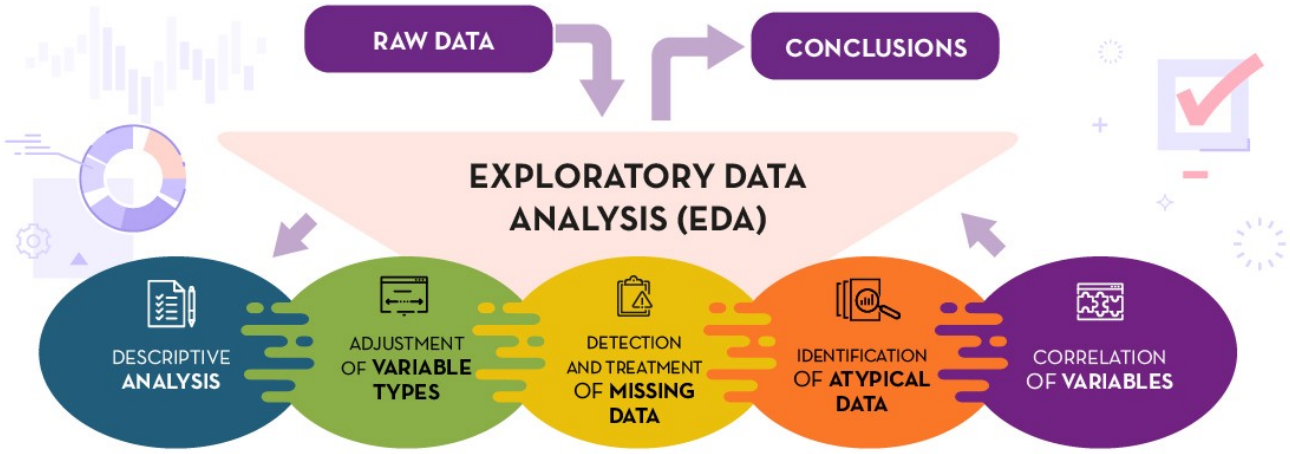
# Questions to ask while EDA

### What is Exploratory Data Analysis ?

In simple words we can say that EDA (Exploratory Data Analysis) introduces data with us. It is the process where we do critical investigation on data to draw few basic conclusions from the data such as the presence of null/duplicate values, Data types, Presence of outliers, etc.

While Performing EDA on any dataset available we need to ask few question to ourselves and our data will answer those questions.

We categorize these questions in 5 categories as follows.




### 1. Data Structure & Distributions

Questions to ask:

1. How many features do you have?
2. How many observations do you have?
3. What is the data type of each feature?
4. From what you know about the features of your dataset, do the data types make sense? Do you need to change any?
5. Do you have null values?
6. How much memory does this dataset use? Could this pose a problem for you later on?
7. What is the distribution of each variable?
8. Do there appear to be outliers?

11. What is the mean for each variable? What do the means tell you about your dataset as a whole?

### 2. Null Values & Duplicates

Questions to ask

1. Check the duplicated sum.
2. Is the null value a result of the way data was recorded? (Was it kept Null Intentionally — we can decide this by carefully understanding the data)
3. Can you drop the rows with null values without it significantly affecting your analysis?
4. Looking at the distributions of the variables, can you justify filling in the missing values with the mean or median for that variable? (We use mean for Numerical Data with no Outliers, Median for Numerical Data with Outliers and we use Mode for Categorical data)
5. If your data is time-series data, can you fill the missing values with interpolation?
6. Are there so many missing values for a variable that you should drop that variable from your dataset?

### 3. Outliers

Questions to Ask

1. Do you have outliers (represented as dark circles on the boxplots) in your variables?
2. Why do you think you have outliers?
3. Do the outliers represent real observations (i.e. not errors)?
4. Should you exclude these observations?

### 4. Correlations/Relationships

Questions to Ask

1. Which variables are most correlated with your target variable? (If applicable)
2. Is there multicollinearity? (Two features that have a correlation $> 0.8$) How will this affect your model?
3. Do you have variables that represent the same information? Can one be dropped?

### 5. Feature Engineering

Variable Transformation:

- One-Hot-Encoding
- Standardization
- np.log()
- np.sqrt()
- box-cox-transformation, etc.

Creating New Features

- You suspect that the relationship of an outcome and a feature depends on a second feature → Create an interaction variabl
- You want to create linear relationships → Create quadratic or higher level functions