Exploratory Data Analysis

Mhd. Arsya Fikri | 191402066

Apa yang akan kita bahas?

What is EDA?

Why do we need EDA?

Principles of EDA

Hands-On

.

.

While there are principles that guides us, EDA is not a one-size-fits-all process. Different projects, different use cases, different requirements, will create a different EDA routine.

01

Exploratory Data Analysis

What is it?

What is EDA?

- Exploratory data analysis (EDA) is used by data scientists to analyze and investigate data sets and summarize their main characteristics, often employing data visualization methods. It helps determine how best to manipulate data sources to get the answers you need, making it easier for data scientists to discover patterns, spot anomalies, test a hypothesis, or check assumptions.

02

Exploratory Data Analysis

Why EDA?

Why EDA?

- To understand our problem and our dataset.
- To have a better overview of our bigger picture, while also knowing the intricate details.
- To plan what to do next, and what can be done in the dataset.

03

Exploratory Data Analysis

Principles of EDA

Principles of EDA

- Ask Relevant Questions

Read or gain at least a background info on the dataset that you want to analyze, and have a few questions in mind first.

Don’t just ask general questions, ask more creative/critical questions!

Principles of EDA

- Understand Each Column

You should also understand each column. What does each column represent, and are they already in the correct data format?

Principles of EDA

- Check for Missing Values

Next, we should check for missing values. In a real-world situation, we should check with the stakeholder about the missing value (and not just impute it ourselves without a clear understanding)

Principles of EDA

- Check for Duplicate Values

The ‘opposite’ of missing values. Sometimes, there are instances where data should not be duplicated.

Examples:

- Student examination data. Each student should only have 1 score for each test.
- Worker email address. Each employee should only have 1 email address within the same company.

Principles of EDA

- Visualize!

There are various types of visualizations –

1. Univariate analysis: This type of data consists of only one variable. The analysis of univariate data is thus the simplest form of analysis since the information deals with only one quantity that changes. It does not deal with causes or relationships and the main purpose of the analysis is to describe the data and find patterns that exist within it.

Principles of EDA

- Visualize!

2. Bi-Variate analysis: This type of data involves two different variables. The analysis of this type of data deals with causes and relationships and the analysis is done to find out the relationship among the two variables.

3. Multi-Variate analysis: When the data involves three or more variables, it is categorized under multivariate.

Principles of EDA

- Visualize!

Don’t forget to visualize your data using visualization principles which we have discussed in previous days.

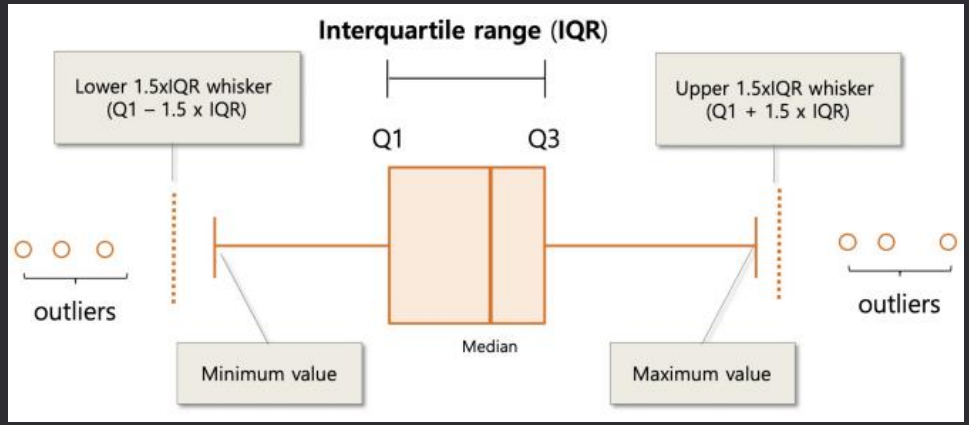
However, here’s an interesting study (yes, dinosaurs are somehow involved)

https://towardsdatascience.com/how-to-turn-a-dinosaur-dataset-into-a-circle-dataset-with-the-same-statistics-64136c2e2ca0

Principles of EDA

- Check for Outliers

Outliers (pencilan) are data which does not follow ‘the others’. They can be too small or too large, and if plotted using box plots, are beyond ±1.5*IQR



Principles of EDA

- Check for Outliers

Like missing values, outliers should not be immediately removed!

We should investigate them, and ask relevant stakeholders as it could shine on a previously undiscovered, bigger problem.

Anomaly usually refer to outliers which are unexplainable (or are wrong).

Principles of EDA

- Do some Feature Engineering

Feature engineering is a task to understand the useful features from the raw data and also, create new features from the existing features that have an impact on the results or, manipulating the features such that they are model ready or can enhance the results.

Example:
Given you have a sales data that has 3 columns.
1. Date
2. Product Name
3. Quantity Sold

Principles of EDA

- Do some Feature Engineering

- Separate date into day, month, year. Some products are seasonal, and knowing the day, month, and year, can be beneficial.
- Identify the product type from each product name. (For example, is it an electronic or food product?)
- Another good indicator of quantity of sales would be the price of the product

Sky is the limit!
You can do anything in the feature engineering step. However, usually this step is emphasized AFTER initial modelling.

04

Exploratory Data Analysis

Hands On

Thanks!

Do you have any questions?

CREDITS: This presentation template was created by **Slidesgo**, including icons by **Flaticon**, and infographics & images by **Freepik**