DS049-EDA & Data Quality Check 1

By: Eng. Esraa Madhi

What is EDA?

Gain an initial **understanding of the data**, its structure, and summarize its main characteristics, often with visual / statistical methods.

EDA is about 'listening' to what the data has to tell us, rather than imposing preconceived notions on it. It's a crucial step that can **lead to more robust, accurate, and insightful results in any data analysis or data science project.**

The output of EDA is typically a better understanding of the data's structure, a refined list of potential variables for modeling(extracting important variables and leaving behind useless variables) , a set of identified data quality issues, and a series of answers (maximize your insights of a dataset) and another bunch of questions or hypotheses that might warrant further investigation and testing.

To summarize, **Understanding the data better** would help us in the following**:**

1. **Discover Patterns:** Uncover any underlying patterns, trends, or relationships between variables that may exist in the dataset.

2. **Spot Outliers**: Identify any outliers or anomalies in the data that may warrant further investigation or could affect the results of later analyses.
3. **Prepare for Further Analysis**: Prepare the data for subsequent analysis steps, including cleaning, transforming, and feature engineering.
4. **Communicate Results**: Facilitate the communication of findings through visualizations and summary statistics that can be understood by stakeholders who may not have a technical background.

Key Aspects / tasks / steps of EDA

1. **Data Profiling**: This is the first step in EDA where you:
   a. Summarize the main characteristics of the dataset (Descriptive Analysis).
   b. Check data types, ranges of values, unique counts, and the presence of null values and the rest of data quality checks.
2. **Data Cleaning**: Preliminary findings from data profiling can lead to cleaning the data by:
   a. Handling missing values
   b. Correcting errors
   c. Dealing with outliers.
3. **Univariate Analysis**: This involves examining single variables to understand their distribution, central tendency, dispersion, and shape.
4. **Bivariate/Multivariate Analysis**: Here, you look at the relationships between two or more variables. This can involve looking for correlations, patterns, and trends that suggest a relationship or an association.

EDA is not a one-time process (Iterative Process) As you perform EDA, you may uncover more data quality issues, which then leads you back to additional cleaning and checks. Likewise, high-quality data enables more effective and accurate EDA, leading to better insights and decision-making.

To perform EDA in Python, you can use libraries like Pandas, NumPy, Matplotlib, and Seaborn. These libraries provide functions and tools for data manipulation, visualization, and statistical analysis, which facilitate the process of exploring and understanding the data.

Pandas library :

pandas is an open-source library built on top of numpy providing high performance, easy to use data structures and data analysis tools for python. It allows for fast analysis and data cleaning and preparation.

Numpy library:

NumPy is a library for Python that adds support for large, multi-dimensional arrays and matrices, along with a large collection of high-level mathematical functions to operate on these arrays.

Matplotlib library :

Matplotlib is a comprehensive library for creating static, animated, and interactive visualizations in Python.

Seaborn library :

Seaborn is a Python data visualization library based on matplotlib. It provides a high-level interface for drawing attractive and informative statistical graphics.

Resources:

- https://medium.com/@ethan.duong1120/hotel-booking-project-exploratory-data-analysis-48bcfb7ae7cd
- https://www.analyticsvidhya.com/blog/2022/07/step-by-step-exploratory-data-analysis-eda-using-python/
- https://medium.com/@lamsampathkumar0/eda-exploratory-data-analysis-project-using-python-de90cbf4e128
- https://www.analyticsvidhya.com/blog/2021/05/exploratory-data-analysis-eda-a-step-by-step-guide/
- https://jovian.com/aishwaryakeshari18/exploratory-data-analysis-project
- https://www.knowledgehut.com/blog/data-science/eda-data-science