MAIN FLOW SERVICES & TECHNOLOGIES

Data Science with Python Internship: Task 4

JUNE 24, 2024

SAKSHI RAJESH BHAVSAR
bsakshi2019@gmail.com

Main Flow Services & Technologies (batch: 25May-25July 2024)

Notes from task 4:

Exploratory Data Analysis

Description:

This task involves performing exploratory data analysis on a dataset.

Responsibility:

Create visualizations to understand the distribution of variables, identify outliers, and check for correlations between variables.

bsakshi2019@gmail.com

Concepts of Exploratory Data Analysis (EDA)

1. **Data Cleaning**:
    - **Handling Missing Values**: Identify and manage missing data to prevent skewed results.
    - **Removing Duplicates**: Eliminate redundant entries to ensure data accuracy.
    - **Correcting Errors**: Fix incorrect or inconsistent data points.
2. **Data Transformation**:
    - **Normalization and Scaling**: Adjust data to a common scale without distorting differences in values.
    - **Encoding Categorical Variables**: Convert categorical data into numerical format for analysis.
    - **Date and Time Parsing**: Extract meaningful components (year, month, day) from datetime fields.
3. **Data Visualization**:
    - **Histograms and Density Plots**: Understand the distribution of individual variables.
    - **Box Plots**: Identify outliers and understand the spread of data.
    - **Scatter Plots**: Examine relationships between two numerical variables.
    - **Correlation Heatmaps**: Visualize the strength of relationships between variables.
4. **Summary Statistics**:
    - **Descriptive Statistics**: Calculate measures like mean, median, variance, and standard deviation.
    - **Frequency Distribution**: Understand how often different values occur in a dataset.
5. **Feature Engineering**:
    - **Creating New Features**: Derive new variables from existing data to provide additional insights.
    - **Feature Selection**: Identify and select the most relevant features for analysis.

bsakshi2019@gmail.com

Interesting Points

- **Interdisciplinary Applications**: EDA is valuable across various fields such as marketing, finance, healthcare, and engineering, demonstrating its versatility.

- **Real-Time Analysis**: With advancements in technology, EDA can now be performed on real-time data, enabling businesses to react swiftly to changing conditions.

- **Empowers Non-Technical Stakeholders**: Through intuitive visualizations, EDA allows non-technical stakeholders to understand and leverage data insights, promoting data-driven cultures.

- **Foundation for Machine Learning**: EDA is often the first step in building robust machine learning models, providing a deep understanding of the data before model training.

bsakshi2019@gmail.com