

Course End Project: Data Analysis Module

Exploratory Data Analysis

Objective:

The objective of this project is to conduct an exploratory data analysis (EDA) on a dataset of your choice. You have the option to use the Airbnb Listings & Reviews dataset or another dataset of your preference, subject to confirmation with the instructor. The goal remains the same: to analyze the dataset, understand its characteristics, and draw meaningful insights from it.

Dataset Options:

- **Airbnb Listings & Reviews Dataset:** This dataset contains information about Airbnb listings, including property details, host information, pricing, availability, and guest reviews. It comprises data for 250,000+ listings in 10 major cities, along with over 5 million historical reviews.
- **Alternate Dataset:** You may propose an alternative dataset of your choice, but you must confirm its suitability with the instructor. The dataset should be sufficiently rich in attributes and suitable for exploratory data analysis.

Tasks:

1. Data Loading and Initial Exploration:

- Load the selected dataset into your preferred data analysis environment.
- Explore the structure of the dataset (number of rows and columns, data types, etc.).
- Examine the first few rows of the dataset to understand its contents.

2. Data Cleaning:

- Handle missing values appropriately (e.g., imputation, deletion, etc.).
- Check for and handle any duplicate entries.
- Convert categorical variables into the appropriate data type if necessary.

3. Exploratory Data Analysis:

- Conduct univariate analysis to understand the distribution of individual variables (e.g., listing prices, number of reviews, etc.).
- Perform bivariate analysis to explore relationships between variables (e.g., price vs. property type, price vs. neighborhood, etc.).
- Visualize the data using appropriate plots (e.g., histograms, box plots, scatter plots, etc.).
- Calculate summary statistics for relevant variables.

4. Feature Engineering:

- Create new features that might be useful for analysis (e.g., average rating, occupancy rate, etc.).
- Encode categorical variables if necessary (e.g., one-hot encoding, label encoding, etc.).

5. Hypothesis Testing (Optional):

- Formulate hypotheses related to the dataset based on different factors.
- Perform statistical tests (e.g., t-test, ANOVA, etc.) to test these hypotheses.

6. Conclusion:

- Summarize the findings from the exploratory data analysis.
- Discuss any insights or patterns observed in the data.
- Make recommendations based on the analysis conducted.

Deliverables:

1. Jupyter Notebook (or equivalent) containing all the code and analysis.
2. Visualizations (plots, graphs) to support the analysis.
3. A brief report summarizing the findings and conclusions.

Note:

- If choosing the alternate dataset option, you must confirm its suitability with the instructor before proceeding.
- Make sure to document your code and analysis steps clearly for better understanding.
- You can refer to online resources or books for guidance on data analysis techniques and best practices.
- Collaboration with peers or seeking guidance from the instructor is encouraged during the project.