Exploring the Data

OBJECTIVE

The purpose of this task is to perform a comprehensive data quality assessment and exploration of a given dataset. You will generate key visualizations and reports to understand the data distribution, detect patterns, and address any data quality issues, such as missing or incorrect data.

STEPS

1 Load the Dataset

- Load the dataset from the file listings_new_york_2024.csv[1].
- Do preliminary data exploration:
  1. Identify number of rows and columns.
  2. Display data types of each column and convert data types if necessary (e.g., dates, categorical variables).
  3. Generate summary statistics (mean, median, mode, etc.) for numerical columns.
  4. Count unique values, find the first and the second mode, the frequency of the first and the second mode for categorical columns.

2 Data Quality Report

- Missing Data Analysis:
  - Identify columns with missing values and the percentage of missing data in each column.
- Incorrect Data Detection:
  - Detect potential outliers or incorrect data entries.

3 Handling Incorrect Data

[1] This file is taken from the : Inside Airbnb dataset, New York City, United States, file date: July 5, 2024.

- Identify and handle outliers or incorrect data entries.
- Use domain knowledge to filter or replace incorrect values where necessary.
- Explain any assumptions made and the process for correcting these errors.

4 Dealing with Missing Data

- Apply strategies to handle missing data:
  - Remove rows or columns with a high percentage of missing data.
  - Impute missing values using mean/median (for numerical data) or mode (for categorical data).
- Document your approach and reasoning behind handling missing data.

5 Data Exploration

- Histograms:
  - Create histograms for numerical columns to understand the data distribution (normal, skewed, etc.).
- Bar Plots:
  - Generate bar plots for categorical columns to examine the distribution of categories.
- Scatter Plots:
  - Create scatter plots to explore relationships between pairs of numerical columns.
- Correlation Matrix:
  - Calculate the correlation matrix for numerical variables.
  - Visualize the correlation matrix using a heatmap to identify highly correlated variables.

6 Final Data Summary

- Provide a summary of the cleaned dataset, including the final number of rows and columns, and a comparison with the original dataset.
- Comment on any transformations or imputations applied.

DELIVERABLES

You are required to submit an IPython notebook containing:

- A detailed data quality report for the original dataset.
- Categorization of all variables (features) in the dataset.
- Visualizations including histograms, bar plots, scatter plots, and a correlation matrix heatmap.
- A data quality plan, outlining strategies for handling missing and incorrect data.
- Answers to the following questions:
    1. What is the distribution of property prices across different neighborhoods, and are there significant differences between them?
    2. How does the room type (Entire home/apt, Private room, etc.) affect the price? Are certain room types consistently more expensive?
    3. What is the correlation between the number of reviews and the availability of listings (availability_365)? Do listings with more reviews tend to be less available?
    4. Are there any outliers in the price or minimum night stays? How do they compare to typical listings?
    5. How do hosts with multiple listings compare to those with a single listing in terms of reviews, pricing, and availability?