IST 652 PROJECT PROPOSAL

The final project for IST652 involves locating an open data set or a group of data sets of interest, formulating an inquiry or set of inquiries that could be addressed with the data,processing the data set(s) in a Jupyter Notebook environment using Python, and conducting some analyses on the data to illuminate the inquiry. The project focuses on open data in order to ensure that your chain of transformations and analysis is reproducible.

This is the FIRST DELIVERABLE

Project Objective

Primary objectives for the project are ..

- Demonstrate your ability to write Python scripts to access and process data.
- Describe steps taken to prepare the data for analysis. For example how did you access and ingest the data, data wrangling, formatting, feature engineering and other steps.
- Develop a research questions you are hoping to answer from the data collected.
- Clearly articulate findings from analysis and summarizes impactful findings.
- Collaborate as a team.

Analysis Team

List team members below and their roles (note roles may be modified in the second deliverable)

```
--== Double-Click and Write Your Project Summary Below This Line ==--
```

Kunal Ahirrao:

Handling data wrangling, cleaning, analysis, and creating visualizations. Currently preparing and preprocessing the data, conducting analysis, and designing meaningful visualizations to address the research questions.

Praveen Bharathi:

Focusing on conducting in-depth analysis and crafting impactful visualizations to derive insights and support findings.

Phase 1: Ideation

The goal of this phase is to outline the specific goals and objectives of your project; include evidence of its feasibility by including citations of resources you will use to complete the code.

Step 1: Project Summary

Write a brief summary of your project ideas, In 250 - 500 words.

```
--== Double-Click and Write Below this Line  ==--
```

This project will analyze Airbnb listings data for New York City to develop insights into the following aspects: pricing, room type, availability, and guest interactions. The dataset contains detailed information on neighborhoods, room types, reviews, and availability that will enable an in-depth analysis of the short-term rental market. The project, by using Python in data analysis and visualization, addresses the key research questions: average price for each neighborhood group and room type; how the number of reviews relates to the price; price variation across room types in the top neighborhoods; how availability affects the pricing; and what distinguishes high-priced listings from low-priced listings. Through this analysis, the project provides actionable insights for both travelers and hosts in making informed decisions.

The analysis will include cleaning, transforming data, and statistical modeling that identifies the presence of patterns and relationships: for example, how minimum night requirements impact availability, how host experience correlates with revenue, and how proximity to landmarks affects pricing. The results of this study can be used by hosts to create an optimized listing and travelers to make more-informed booking decisions.

Given the fine open data provided by Airbnb and the team's skills in Python and data analysis, this project is feasible. Some main tools and libraries which will be used include Pandas, Matplotlib, Seaborn, NumPy, and Scikit-learn. Visualization such as heatmaps, treemaps, and radial plots will add visual value to the presentation of findings.

Step 2: Datasets Research

Select a dataset or a combination of datasets for your project. Many data sets are available at sites such as the World Bank (http://data.worldbank.org), the U.S. Federal Government (http://www.data.gov), - other potential sites for data sets will be provided by the instructor but it is recommended that you search for open data sets too on your own. However, do not use datasets from Kaggle.com.

Note: The number of records (rows) present in your dataset (or total combination of datasets) must exceed 4,000 with at least 8 different categories (columns) of data.

Clearly describe from where your data was located.Why is this resource an authority. Provide a shortlist of datasets your team is considering for your final project. Provide references to the dataset as applicable. Include any other components necessary.

```
--== Double-Click and Write Below this Line  ==--
```

Name: New York City Airbnb Open Data

Source: https://insideairbnb.com/get-the-data/

Authority:

This dataset is curated by Inside Airbnb, an independent, open-source platform providing detailed data for public research and analysis. The dataset includes over 37,541 rows and 18 columns covering various listing features.

Why This Dataset:

The dataset is extensive and well-documented, with clear categories such as price, room type, location, and reviews. It is reliable, frequently updated, and used widely for Airbnb-related research.

Dataset Components:

Listings: Includes prices, room types, and availability.

Hosts: Attributes such as the number of managed listings and reviews.

Locations: Geospatial data for mapping proximity to landmarks.

Reviews: Monthly and total review counts for popularity analysis.

Policies: Minimum nights and cancellation policies.

Step 2a: Objectives

What have you learned about your dataset(s) so far, and what are the questions you plan to answer with the data (a minimum of 5 questions is a good start).

```
--== Double-click and write below this line  ==--
```

Project Objective

The primary objectives of the project are as follows:

Demonstrate the ability to write Python scripts to access, clean, process, and analyze data.

Describe the steps taken to prepare the data for analysis, including data ingestion, wrangling, formatting, and feature engineering.

Formulate generalized research questions to guide the analysis and derive actionable insights.

Clearly articulate findings from the analysis in a way that is accessible and impactful.

Collaborate effectively as a team to ensure project success.

Data Preparation

To prepare the dataset for analysis, the following steps will be taken:

Data Cleaning: Handle missing values, remove redundant columns, and filter outliers to ensure consistency.

Feature Engineering: Create additional features, such as availability bins and price categories, to enhance the analysis.

Data Transformation: Format numerical and categorical columns for compatibility with Python libraries like Pandas, Seaborn, and Matplotlib.

Exploratory Data Analysis (EDA): Summarize the data using descriptive statistics and initial visualizations to understand trends.

Research Questions:

What is the average price by neighborhood group? This question explores pricing trends across different neighborhood groups to identify premium and budget-friendly locations.

What is the average price by room type? This analysis determines how room types (e.g., entire homes, private rooms, shared rooms) influence pricing.

What is the relationship between price and the number of reviews? This examines whether listings with more reviews are priced higher or lower and identifies patterns in guest engagement.

How does average price vary by room type in the top neighborhood? This question focuses on understanding room-type-specific pricing within the most popular neighborhoods.

How does availability affect the average price of listings? This investigates how the availability of listings (measured in days per year) impacts pricing trends.

What distinguishes high-priced listings from low-priced listings? This explores the characteristics of listings at different price points, such as room type, availability, and review patterns.

Analysis Techniques

The analysis will involve:

Descriptive statistics for summarizing data.

Data visualizations (e.g., heatmaps, bar charts, scatter plots) to reveal patterns.

Correlation analysis to explore relationships between features.

Comparative analysis of high-priced and low-priced listings to identify trends.

Tools:

Python Libraries: Pandas, NumPy, Matplotlib, Seaborn. Jupyter Notebook for interactive analysis.

References

```
--== Double-click and write below this line  ==--
```

References

Inside Airbnb: http://insideairbnb.com/get-the-data.html

Official Airbnb Data Dictionary: https://insideairbnb.com/how-we-use-data.html

Python Libraries: Pandas, Seaborn, Matplotlib, NumPy, Scikit-learn.

Supporting Articles: Historical Airbnb trends in NYC. Analytical methods for short-term rental data analysis.

In [ ]: