Homework Assignment 1 – Exploratory Data Analysis

Dr. Gail Gilboa-Freedman
Dr. Naveh Eskinazi
Mr. Asaf Lev

Submission: 13/4/2023

GENERAL INSTRUCTIONS

The purpose of this exercise is to align all students with importing Python libraries and modules and perform basic exploratory data analysis.

The work will be based on a CSV named **“abb_nyc_data.csv”** located in the course’s Moodle site.

SUBMISSION:

Through assignment box within the course Moodle, submit a **Jupyter Notebook file named HWA1_<student name>.ipynb** (e.g. HWA1.avia_malka.ipynb)
Should include all the relevant code needed to perform the assignment’s tasks along with code's output.
(Recommendation: Add headers and sub-headers using the Markdown option)

Good Luck!

PART1: PREQUISITES

TASK 1: SETTING THE FOLDER

1. Create a new and blank Jupyter Notebook named **HWA1_<student name>.ipynb.**
2. Download from the CSV file named **“abb_nyc_data.csv”** from Moodle.
3. Upload the CSV file to Jupyter (Note: make sure the file is placed in the same location as your Jupyter Notebook)

TASK 2: IMPORT LIBRARIES & MODULES

4. Import the following libraries and modules within your notebook: **scipy, numpy, pandas, and matplotlib.pyplot**

PART 2: EXPLORATORY DATA ANALYSIS

To complete the following tasks, use what you've learned in the lecture and tutorial and rely on the **abb_nyc_data** dataset.

TASK 3: DESCRIBE STATISTICS

Use Python commands (e.g., shape, describe, iloc) to plot and provide answers to the following questions:

5. How many rows and columns are in the data?
6. What are the apartments' neighborhoods in rows **4 till 9**?
7. How many types of **Rooms** there are?
8. How many apartments are priced above **$1000** per night?
9. What was the date of the **last review** given for an apartment?
10. What was the **standard deviation** of **price per night**?
11. What was the **lowest number of reviews**?
12. What is the **highest Latitude**? Explain the meaning of this result using your own words.
13. What is the **mean number of days over the year** in which the apartments are available?
14. What is the **total number of reviews** given?
15. What is the **number of apartments** located in **Queens** district?
16. What is the **most** popular district?
17. What was the **least** popular district?

TASK 4: VISUALIZE STATISTICS

18. Create a data frame of the apartments that their **price per night** ranges **between $1000 and $5000** (Including these values).
19. Using a Box & Whisker plot on data frame you've created in question 18 and answer the following questions:
    - What represents the horizontal line below the box?
    - What represents the horizontal line above the box?
    - What represents the horizontal line which goes through the box?
20. Based on the data frame you've created in question 18, show a histogram of prices.
21. Using a Bar chart, what is the **Room Type** in which the number of reviews was **the highest**? (Hint: before creating the chart, you need to group by the data using the aggregation function sum)

Good Luck!