MIE 223 Data Science
Lab and Assignment 02:
Data Cleaning with Python

In this lab and assignment, you will learn and gain hands-on experience on the use of Pandas for data cleaning.

- Programming language: Python (Google Colab Environment)
- Due Date: Posted in Syllabus

Marking scheme and requirements: You are required to provide appropriate answers to the questions in the Jupyter notebook named `Assignment_02.ipynb` and commit all changes to your assignment repository.

This assignment has *4 points* in total and the point allocation is shown below:

- Solution to Questions (1 point):
  - Each question is worth 0.1 point. That is Q1(a): 0.1, Q1(b) 0.1, and so on.
- Code review (1 point)
- Quiz (2 points)

What/how to submit your work:

- All your code should be included in the provided notebook named `Assignment_02.ipynb`.
- Commit and push your work to your github repository in order to submit it. Your last commit and push before the assignment deadline will be considered to be your submission. You can check your repository online to make sure that all required files have actually been committed and pushed to your repository.
- A link to create a personal repository for this assignment is posted on QUERCUS.

Notes that you should pay attention to

1. All your code should be written in the provided notebook named `Assignment_02.ipynb`.
2. All functions must *return* the specified return type.

3. Commit and push your work to your GitHub repository in order to submit it. Your last commit and push before the assignment deadline will be considered to be your submission. You can check your repository online to make sure that all required code has actually been committed and pushed to your repository.

4. Your code should show all required outputs asked in the assignment questions. If any of the assignment question asks you to explain your outputs, please use markdown cells to clearly explain them.

5. As a backup plan, also upload a PDF copy of your assignment notebook to GitHub. This is useful especially when GitHub cannot render some images or outputs.

6. During the code review, TAs will go over your answer submissions to assess your understanding of the concepts relating to the assignment and the logic behind your codes. Please, ensure your codes are readable and well commented using python comments and markdown cells when necessary. Not have readable code may impact your code review score.

7. Please note the plagiarism policy in the syllabus.

1 Main Assignment

You are provided with aggregated data of room and property rentals in Toronto listed on Airbnb's website on January 8th, 2024. Data was obtained from Inside Airbnb. You can see their data dictionary here (the sheet for the data used in this assignment is named "**listings.csv summary v2**").

Please answer the questions below in the provided notebook named `Assignment_02.ipynb` using Google Colab. **Pandas functions should be used for computing your answers to all the questions**.

Q1. Data Inconsistency

(a) Output a Pandas Series that shows the data type of each column.

(b) Convert the data type of the "**rating**" column to float and output its new data type

Q2. Summary Statistics

Output the following summary statistics for the "**rating**" column: count, mean, standard deviation, min, 25th percentile, Median, 75th percentile, and Max. These should be calculated using one line of code, the output should be a Pandas Series that contains all the summary statistics.

Q3. Missing Values

(a) Show the percentage (expressed as a fraction, no need to multiply by 100) of missing values per column in a Pandas Series. The keys of the resulting Series should be the column names, while the values should be the percentage of missing values. The resulting Series should be sorted in descending order of its values.

(b) Filter the data to only show the listings (rows) with missing value in the "**host_name**" column. The resulting DataFrame should show all the columns in the dataset.

(c) Drop the listings with missing values in the "**host_name**" column. Be sure that this operation modifies the original dataframe named **df**. Print the new number of rows in **df** and repeat **Q3(a)**.

(d) Impute missing values in "**price**" column conditioned on the average price in their "**neighbourhood**". Print the number of missing values in the "**price**" column before and after the conditional imputation.

Q4. Data Transformation and Outlier Detection

(a) Create a new column (name it "**z-score_of_log_price**") whose values are the z-scores of the log of the values of "**price**" column. That is:

$$z_r = \frac{\log(p_r) - \mu}{\sigma} \, \forall \, r \in \{1, 2, ..., R\}$$

where:

R is the number of observations (rows) in the dataset.
z_r is the "**z-score_of_log_price**" value in row r
p_r is the value of the "**price**" column in row r
μ is the sample mean value of $\log(p_r)$
σ is the sample standard deviation value of $\log(p_r)$

(b) Output the rows where the values of "**z-score_of_log_price**" is greater than 3. Your output should show only the following columns: "id" "name" "host_id" "host_name" "neighbourhood" "room_type" "price" "number_of_reviews" and "rating"; and should be sorted by "**price**" column in descending order. You should also print the shape of the resulting DataFrame.

(c) Repeat **Q4(b)** for rows where the values of "**z-score_of_log_price**" is less than -3. In one sentence, use a markdown (text) cell to explain the reason that we might look for listings with z-scores less than -3. (*Hint*: are these typical listings? If not, why not?)