Key Tasks

In this project, you will perform data profiling and cleansing for the Passengers’ Satisfaction Survey dataset Download Passengers’ Satisfaction Survey dataset. You’ll examine and assess the data to expose technical and data issues to plan and conduct data cleansing using Python. Here is the Passenger's Satisfaction Survey data dictionary Download Passenger's Satisfaction Survey data dictionarywhere you can find a description for each column in the dataset. All your work will be performed using a Jupyter Notebook with annotations that explain and document your work. Your submission will also include a CSV file for the cleaned dataset.

Set up one Jupyter notebook to complete the following key tasks and annotate your work in the notebook:

Conduct Data Access and Profiling:

To begin, read the survey dataset into a pandas DataFrame.

Then, create a subset of the DataFrame to include all passenger satisfaction data about only the Business Class. Name the created DataFrame subset “df\_Business”.

Collect information about the “df\_Business” dataset and its validity. Report the following:

The number of observations (or rows) and the number of variables (or columns) in the dataset.

The name and data type of each data column.

The unique values of each column.

The number of missing values of each column.

Summary statistics for each data column.

Frequency distribution (list the number of times each unique value appears) for each data column.

The number of fully duplicated data rows.

Based on the information that you collected, provide a list of the issues that the data contains as an annotation.

Supporting Materials

Conduct Data Cleansing:

Implement the following tasks:

Drop the Class column from the created Business Class passenger satisfaction survey (“df\_Business”).

Rename “df\_Business” columns by making all the columns’ names in lowercase, and if the column name has whitespace, replace it with an underscore. (For example: “Customer Type” will be renamed to “customer\_type”).

Modify improperly formatted data and handle bad data.

Handle the missing values (You should have zero missing values in your cleaned dataset).

Remove the fully duplicated observations.

Fix the columns datatypes.

Export the cleaned dataset to CSV file.

Use the following naming convention: Project\_2\_Group#.csv

Compare the Data Characteristics Before and After Data Cleansing.

At the end of the Jupyter Notebook, provide a brief write-up annotation to compare the characteristics of each column before and after data cleansing (this includes, the column name, datatype, unique values, and the number of missing values).