WIA1007
Individual Assignment

Data Wrangling with Pandas

Instructions

In this assignment, you will work with a dataset and perform data wrangling tasks using the Pandas library in Python using the Jupiter notebook. Follow the steps below to complete the assignment:

Step 1: Data Loading
Download the dataset from Kaggle onto your local drive. Load the dataset into a Pandas DataFrame. The dataset must be in CSV format.

Step 2: Initial Data Exploration
Display the first 5 rows of the dataset to get an initial sense of the data. You may use the **info()** and **describe()** functions to gather basic information about the dataset, such as data types, missing values, and summary statistics (e.g., min, max, mean, std deviation, and quartiles).

Step 3: Data Cleaning
Handle missing values using appropriate Panda codes: If there are missing values in the dataset, decide on a strategy to deal with them (e.g., fill with a specific value or drop rows row columns). Check for and handle duplicate entries if they exist.

Step 4: Data Selection and Filtering
Select a subset of the data based on a specific condition (e.g., select rows where a particular column meets a certain criteria). Filter the data to include only the columns that are relevant to your analysis. You may use **regex()**.

Step 5: Data Transformation
Create a new column that combines information from one existing column. Suppose you want to create a new column that combines a date (in datetime format) that combines one existing column, for e.g. total revenue of the day.

Step 6: Data Aggregation
Group the data by a particular column and calculate summary statistics (e.g., mean, sum) for other columns within each group. For example, by grouping using **groupby()** the data by month and calculating the total revenue for each month.

Step 7: Data Visualization
Create at least two data visualizations using Matplotlib or Seaborn to illustrate trends or insights in the data (e.g., bar chart, scatter plot, histogram).

Step 8: Data Export
Save the cleaned and transformed dataset to a CSV file for future analysis.

Step 9: Documentation
Provide comments and explanations in your code cells to describe the steps you've taken by using the markdown feature in Jupiter.

Step 10: Conclusion
Summarize your findings, insights, and trends from the data wrangling process. Explain how the dataset can be laveraged for decision-making making. What type of Machine learning problems, e.g. classification / clustering / regression is suitable for the dataset and decision-making.

Submission:
Save your Jupyter Notebook containing the Pandas code and analysis. Submit your Jupyter Notebook as well as the cleaned dataset (in CSV format).

Note:
Feel free to choose a dataset related to your area of interest.
Iris dataset is not allowed for this assignment.

Grading Criteria:
You will be evaluated on the completeness of each step, accuracy in data cleaning and transformation, clarity of code and comments, and the quality of data visualizations and insights provided in your analysis. This assignment covers common data wrangling tasks using Panda and strongly encourages students to explore more intermediate Panda code for data wragling for extra marks.

Good luck with your assignment!

[20 marks]