Assignment 4- Data Wrangling Process

GGE6505/GGE5405 Introduction to Big Data and Data Science

Data Wrangling is the process of gathering, collecting, and transforming Raw data into another format for better understanding, decision-making, accessing, and analysis in less time. Data Wrangling is also known as Data Munging.

In this assignment, you will learn how to explore and clean a dataset. Following the class notes each group needs to perform the following actions on their assigned dataset:

- **Data Exploration and visualization**:
    a) Explain the data. Find the features in the dataset. Print their names and dimensions.
    b) Plot the data distribution of a few features. Discuss about their mean and variance.
- **Data Pre-processing**:
    a) Data cleaning: Find missing data. Remove them and justify your choice.
    b) Data Cleaning: Identify noise in the data. How did you identify the noise? Justify and demonstrate the technique you would use to reduce noise.
    c) Data transformations: Perform standardizations and normalization. Justify your chosen normalization method.
    d) Any other techniques which are required for your dataset such as adding data head

Please create a python notebook and demonstrate your steps. You will present your codes in a notebook and not PowerPoint.

assigned datasets:

Group 1: Taxi Trajectory Data | Kaggle

Group 2: Get the Data - Inside Airbnb. Adding data to the debate.(New York City, New York, United States)

Group 3: UCI Machine Learning Repository: Automobile Data Set

Group 4: UCI Machine Learning Repository: Adult Data Set

Group 5: Titanic dataset | Kaggle

Group 6: UCI Machine Learning Repository: Movie Data Set

Group 7: The McGill Billboard Project · DDMAL

Group 8: Get the Data - Inside Airbnb. Adding data to the debate.(Vancouver, British Columbia, Canada)

Group 9: UCI Machine Learning Repository: Exasens Data Set

Submission:

1. Due Date for presentation (in class) and file submission (11 a.m.): Wednesday, March 2.
2. Group assignment (2 students per group)
3. 5 minutes presentation per group to answer all questions. Remember that you will be penalized if you go over time.
4. Upload your files in D2L and GitHub

Grading metrics:

1. Content and Organization

2. Communication and Engagement. Presentation style.

3. Comprehension. Answers to assignment questions.

4. Finishing presentation within Time limit.

5. Team engagement.