National University of Computer and Emerging Sciences, Lahore Campus

Course:	Data Science	Course Code:	DS-2001
Program:	BS(Computer Science)	Semester:	Fall 2023
Duration:	-	Total Marks:	40
Due Date:	19-Nov-23	Weight	4%
Section:	A	Page(s):	2
Exam:	Project-I	Roll No.	

Instruction/Notes:

- Read the assignment carefully. Make sure you understand the requirements and expectations of the assignment.
- Ensure that you have all the necessary files and documents ready for submission in the CORRECT format.
- Only group leader should submit the files.
- The assignment must be submitted before the announced DEADLINE. One mark will be deducted for each day of late submission.

PART 1: DATASET DESCRIPTION

By this time, you are expected to have your dataset ready for the next step in your project. In this part of the assignment you are required to submit a brief document explaining your dataset for this project. It should include:

- Problem statement
- Source of the dataset
- Brief description about the dataset
- Description of the attributes/variables/columns of the dataset

PART 2: DATA WRANGLING, PREPROCESSING AND TRANSFORMATION

Data wrangling is the process of cleaning, transforming, and preparing raw data into a format that is suitable for analysis. In this assignment, you will demonstrate your understanding of data wrangling techniques and tools.

Your task is to wrangle a dataset chosen for your term project.

Data cleaning: Identify and address missing data, duplicates, and outliers. Consider the Reading assignment 1 when choosing the method for filling the missing values.

Data transformation: Transform the data into a format that is suitable for analysis. This may include converting data types, creating new variables, encoding variables, and merging data from multiple sources.

Normalization and Standardization: Normalize your data if applicable. Normalization and standardization are both techniques used to transform numerical data into a standardized format that can be more easily analyzed and compared. Normalization scales the data to a range between 0 and 1, while

standardization transforms the data to have a mean of 0 and a standard deviation of 1. The choice of technique depends on the specific data and analysis requirements.

Dimensionality Reduction: Working with too many variables is expensive and may lead to unexpected result. It is referred as the Curse of Dimensionality. Look through your dataset and use an appropriate Dimensionality Reduction Algorithm to cut down the variables that are not useful for data modeling step.

Additional Instruction In Case Of Image Data:

The data wrangling process for an image dataset involves preparing the images for analysis by cleaning, transforming, and validating the data. Here are the steps that could be followed in a typical data wrangling process for an image dataset:

Cleaning: In the cleaning phase, the images should be inspected for any errors, inconsistencies, or duplicates. This may involve removing any corrupted or invalid images, checking for naming inconsistencies, and removing any duplicates.

Normalization: Normalization is a common pre-processing step in image data analysis that is used to bring the pixel values of an image into a specific range or scale. The purpose of normalization is to improve the performance of deep learning models by making the data more consistent and easier to process. In image data, pixel values are typically represented as integers ranging from 0 to 255, where 0 represents black and 255 represents white. Normalization rescales these pixel values to a range of values between 0 and 1 or -1 and 1.

Transformation: In the transformation phase, the images may be resized, cropped, or converted to a standard format. This may involve resizing the images to a uniform size, cropping them to remove any irrelevant background or foreground, or converting them to a standard format such as JPEG or PNG. It also involves.

Augmentation: Image augmentation is a technique used in computer vision and image processing to increase the size of the dataset by artificially creating new, transformed versions of the existing images. The process involves applying a set of image transformations such as rotation, scaling, flipping, and cropping to the original images, creating a larger and more diverse dataset.

PART 3: EXPLORATORY DATA ANALYSIS AND VISUALIZATION

You have collected, cleaned, and transformed the dataset and are now ready to explore and visualize it. In this assignment, you will use various statistical techniques and data visualization tools to analyze and understand the data better.

The purpose of this part is to:

- Identify patterns, trends, and relationships in the dataset.
- Gain insights into the data and its characteristics.
- Generate visualizations that can communicate your findings effectively.

Instructions:

Understanding the Dataset: Before diving into exploratory data analysis, you need to have a clear understanding of the dataset. Review the data dictionary and the metadata to understand the structure, variables, and their meanings.

Univariate Analysis: Begin by analyzing each variable individually. Use summary statistics and visualizations such as line charts, histograms, box plots, bar/pie charts, word clouds, and density plots to understand the distribution, central tendency, and variability of each variable.

Bivariate Analysis: Next, analyze the relationship between pairs of variables. Use scatter plots, correlation matrices, multiple line charts, and heat maps to identify correlations, dependencies, and interactions between variables.

Descriptive Analysis: Explore the data and provide descriptive analysis of data in form of reports and charts/graph.

Deliverables: Your deliverables for this assignment should include:

- A Data description document (pdf).
- A Jupyter Notebook that documents your data wrangling and exploration process. Your notebook should include code and explanations of your methods and findings.
- A final cleaned dataset in a CSV format that can be used for further analysis. In case of image dataset, share the Google drive folder containing the dataset.

Grading Criteria: Your assignment will be evaluated based on the following criteria:

Completeness of data description document. (10)

Completion of all required tasks: Cleaning, transformation, exploration and visualization. (25)

Clarity and organization of the Jupyter Notebook. (5)

Submission: Submit your Jupyter Notebook, final cleaned dataset, and data description document as a zip file on the classroom. Rename the file to your project title before submitting.