Assignment-I: Statistical Interpretation and Data Visualization.

Siman Giri, Anmol Adhikari

December 2022

1 Assignment Details

Due	Marks	Submission
December-20, 15:00 NST.	10	Rendered .ipynb file//see details below

2 Assignment Overview

In this assignment, you will implement what you have learned till now. The datasets have been provided to you in the google classroom to complete your (Statistical Data Analysis and Data Visualization) task and follow the instructions.

It is primarily an exercise in applying programming knowledge and skills to data analysis, demonstrating your skills for problem-solving and critical thinking.

Please note that plagiarism is a serious academic offense, for which penalties are severe.

All suspected cases of plagiarism will be reported.

3 Submission Guidelines

The final date for submission is 20-December-2022 and 15:00 PM-NST. .

3.1 Naming Conventions:

You are supposed to follow naming conventions strictly any file not following the naming conventions will be marked"0".

File Name: WLVIDFullName(firstname+last).ipynb

Example: 00000ABC Sharma.ipynb

3.2 How to submit:

You are expected to submit completely rendered .ipynb file named after following naming convention.

3.3 Where to submit:

Designated Portal opened at **Google Classroom**, where you are supposed to upload the **rendered.ipynb**, correctly named before the deadline.

No Late submission allowed.

3.4 After Submission

After the submission you are expected to give small viva based on the work you submitted.

Please Note: No marking without Viva.

Consult with your respected tutor for your viva schedule. .

4 Learning Outcomes

Learning outcomes can be following but not limited to:

1. Use Pandas as the primary tool to process structured data in Python with CSV files,
2. Use matplotlib and seaborn library to produce various plots for visualization,
3. Extract various information from a given dataset using statistical and visualizing techniques.

5 Dataset Description

The data given here in the assignment is of student achievements in secondary education of two Portuguese schools namely "Gabriel Pereira" and "Mousinho da Silveira". The dataset's attributes include student grades in demographic, social, and mathematics and many other school-related features, and all the information was collected by using the school reports and questionnaire. The data description of the data file is provided below:
Filename: "performance.csv"
available for download from google classroom.
The data set contains 395 student records. Each record consists of 33 variables, which include information about the students. Variable 33, G3 – final grade (numeric: 0 – 20), is the target variable)

6 Tasks and Marks Division

6.1 Data Understanding and Cleaning: [1]

1. Use pandas to upload your data.
2. Once you make your initial observation, please explain why do you think the data was collected, what kind of information you can extract from the dataset. Please write your answer in text cell of the Jupyter Notebook.
3. Check for null values and datatypes for all the columns present in the dataset.

6.2 Data Transformation: [4]

Write code to transform variables according to the following instructions:

1. "School", "sex", "address", "schoolsup", "famsup", "activites", "nursery", "internet", and "romantic" into binary: 0 or 1 (create new columns without overwriting the existing ones).
2. "Medu", "Fedu", "reason", "guardian", "studytime", "freetime", and "health" into ordinal numbers based on the number cases in the data set (create news columns without overwriting the existing ones).
3. Convert column "age" to interval datatype. i.e. Create a new column name category_age whose values should be based on the frequency in the column "age", You can create categorical data with following interval.
   (a) interval1: [15-17]
   (b) interval2: [18-20]
   (c) interval3: [21-all]
4. Create a new column name passed (yes or no) whose values should be based on the values present in the G3 column ($>= 8$–$yes, < -no$).

6.3 Descriptive Analysis: [1]

Write a code to show the summary statistic (sum, mean, median, standard deviation, max, and min) of the variables age, absences, G1, G2, and G3).

6.4 Data Exploration and Visualization: [4]

1. Write a code to show **histogram plots** and **boxplots** to visualize the distribution of the variables "age", "absence", and "G3". Interpret the results and comment on the distribution of each variable. Write your answer in text cell of jupyter notebook.
2. Write a code to show a **bar graph** of the total number of students who passed the final term grouped according to the school that they belong to. Use proper labels in the graph and interpret the results.
3. Write a code to show a **pie chart** of analysis of students' grades as per their school. Use proper labels in the graph and interpret the results.
4. Write a code to show a **bar graph** with the relation of the father's occupation with the grade of the student. Use proper labels in the graph and interpret the results.

7 The End

"Please follow Good and Clean Coding Practise."