Data Science

Working with Data

**Task III: Advanced Data Visualization and Analysis using Matplotlib and Seaborn**

**Part I:**

Load the dataset using Pandas.

Perform data preprocessing by:

a. Handling missing values, if any

b. Encoding categorical variables, if required

**b.** Encoding categorical variables is not required in this data frame, because all the variables have been encoded.

**Python code:**

import numpy as np

import pandas as pd

import matplotlib.pyplot as plt

import seaborn as sns

df1 = pd.read\_csv("D:\StudentsPerformance.csv")

#Display 1st 5 rows with all columns

df1.head()

**A screenshot of a computer

Description automatically generated**

**Python code:**

#Checking missing values is there or not, 0 = no missing values

df1.isna().sum()

**A screenshot of a computer

Description automatically generated**

**Python code:**

#missing values is there or not, 0 = no missing values

df1.isnull().sum()

**A screenshot of a computer

Description automatically generated**

**Python code:**

#how many rows and columns

df1.shape

**A screenshot of a computer

Description automatically generated**

**Python code:**

#creating a copy of the data frame before to a major change.

df1\_copy = df1.copy()

#how many rows and columns

df1\_copy.shape

**A screenshot of a computer

Description automatically generated**

**Python code:**

#in real world data is unclear or might be having duplicate rows, so need to remove them before to any operation

#entire rows that are duplicated will be removed

df1\_copy.drop\_duplicates(inplace=True)

#how many rows and columns

df1\_copy.shape

**A screenshot of a computer

Description automatically generated**

According to this result, we can confirm that there are no duplicate rows in the data source.

**Part II:**

Use Matplotlib to create a stacked bar plot displaying the average scores in Math, Reading, and Writing exams for each parental level of education. Customize the plot by adding a title, axes labels, and a legend.

Create a Seaborn heatmap visualizing the correlation between the numerical variables (Math, Reading, and Writing scores) for each gender. Customize the heatmap by adding a color map, annotations, and ensuring the tick labels are readable.

Based on the visualizations, provide a brief analysis (100-150 words) discussing the relationships between the exam scores, parental level of education, and gender, as well as any interesting patterns or trends you observe.

**Python code:**

**A screenshot of a computer program

Description automatically generated**

**A graph of a graph of a graph

Description automatically generated with medium confidence**

**Observations:**

1. Comparatively, nearly most of the students at the Associate's degree level and Some college level have average scores that are similar across all three academic areas.
2. The research indicates that students are more actively involved in math, reading, and writing in their master's degree programs than they are in high schools.
3. Overall, the total number of students engaging with Maths, Reading and Writing subjects with their Associate’s degrees and Bachelor’s degree are approximately equal.

Using Seaborn, create a pair plot showing the relationships between the numerical variables (Math, Reading, and Writing scores) along with their respective distributions. Customize the plot by selecting an appropriate color palette and setting the plot style.

**Python code:**

A computer code with text

Description automatically generated with medium confidence

A graph of a test results

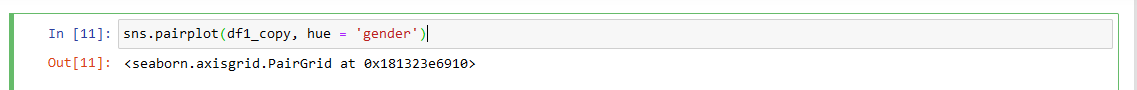
Description automatically generated

The basis for the pairs plot is the scatter plot and the histogram. While scatter plots show the relationship between two variables, the diagonal part of histogram lets us see how a single variable is spread.

For example, the scatter plot of the corelation between the two variables, the math score and the reading score is so higher than the corelation between the two variables of the scatter plot of reading score and the writing score.

Let’s look at the same by using a categorical variable, the gender.

**Python code:**



A graph of a number of red and orange dots

Description automatically generated with medium confidence

According to the aforementioned graphs, male students show significant correlations between the reading-math and writing-math scores while female students have large correlations between the math-writing and math-reading scores.

**Or for example, we can write the Python code as below by using the ‘coolwarm’ color palette:**

A close-up of a white background

Description automatically generated

A graph of a test

Description automatically generated with medium confidence

**Box Plot**

**Python code:**

A screen shot of a computer

Description automatically generated

A diagram of a graph

Description automatically generated

x axis = data to be plotted

y axis = the distribution of the frequency

From the above, we can see that the median math score remains between 60 and 70, while the medians for reading and writing are close to 70. Additionally, this shows the outliers quite clearly, and those are roughly less than 25.

|  |
| --- |
| (geeksforgeeks, 2022) References geeksforgeeks, 2022. *geeksforgeeks.* [Online]  Available at: https://www.geeksforgeeks.org/box-plot-in-python-using-matplotlib/ [Accessed 12 September 2023].  (built in, 2023) References built in, 2023. *built in.* [Online]  Available at: https://builtin.com/data-science/boxplot [Accessed 12 September 2023].  (ProjectPro, 2023) References Available at: https://www.projectpro.io/recipes/make-boxplot-and-interpret-it [Accessed 12 September 2023]. |

**Python code:**

# Heatmap using Seaborn

corr = df[['maths\_score', 'reading\_score', 'writing\_score']].corr()

plt.figure(figsize=(8, 6))

sns.heatmap(corr, annot=True,annot\_kws={"size": 15}, cmap='Blues', center=0, linewidths=0.5, fmt='.2f')

plt.title('Correlation Heatmap of Exam Scores by Gender', y=1.03)

plt.show()

**Results:**

A screenshot of a graph

Description automatically generated

A screenshot of a computer

Description automatically generated

**Or**

**Python code:**

A screenshot of a computer screen

Description automatically generated

For instance, the corelation between a writing score and a math score is 0.8, and in a similar vein, the corelation between a writing score and a reading score is 0.95. As opposed to the other two corelations, writing and reading score have a significant correlation.