

Principles of Data Science Assignment 1- question 2.

Hari chandra prasad Pasupuleti

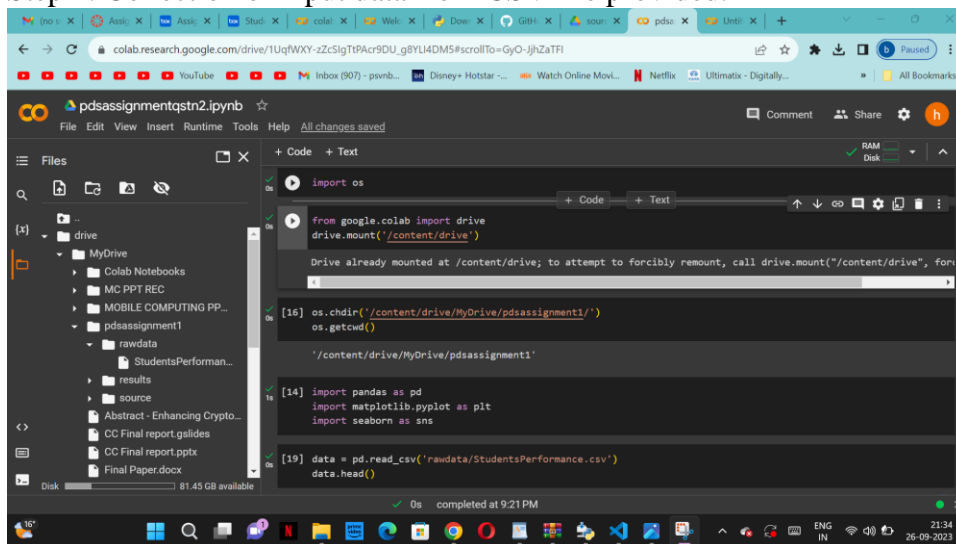
16341883

2) Perform 5 data visualization tasks on the student performance dataset given in the link below (create

5 different visualizations). Explain what kind analysis has become easier with each of the visualizations.

Create the folder structure for this question similar to question 1. (15 points)

Step 1: Collection of input data from CSV file provided.



The screenshot shows a Google Colab notebook titled 'pdsassignmentqstn2.ipynb'. The left sidebar displays the file explorer with a folder structure: 'drive' > 'MyDrive' > 'pdsassignment1' > 'rawdata'. The main code area contains the following Python code:

```
import os

from google.colab import drive
drive.mount('/content/drive')

Drive already mounted at /content/drive; to attempt to forcibly remount, call drive.mount("/content/drive", force_remount=True).

[16] os.chdir('/content/drive/MyDrive/pdsassignment1/')
os.getcwd()

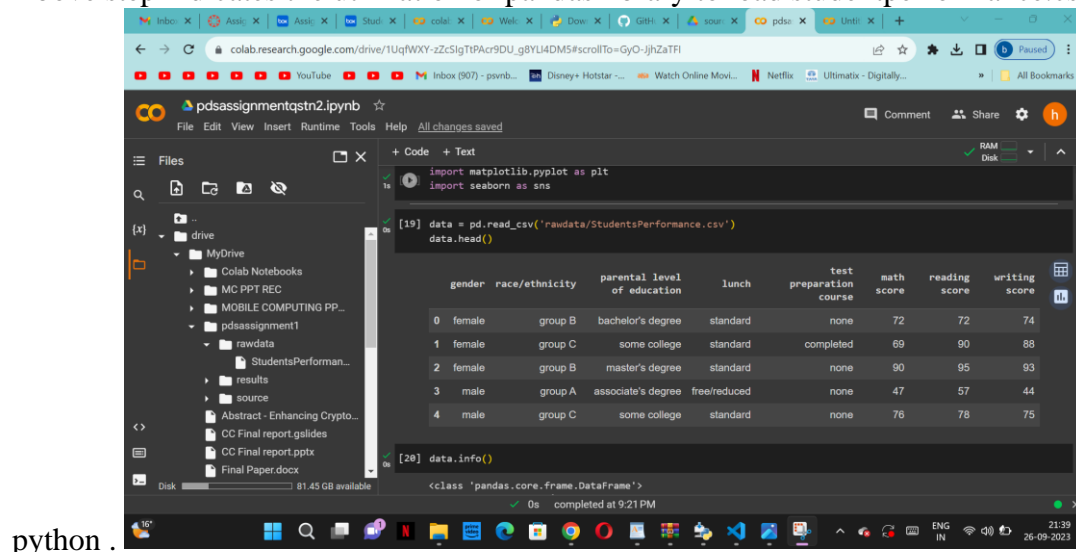
'/content/drive/MyDrive/pdsassignment1'

[14] import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

[19] data = pd.read_csv('rawdata/StudentsPerformance.csv')
data.head()
```

The notebook interface shows the code is executed successfully, with a status bar at the bottom indicating 'completed at 9:21 PM'.

Above step indicates the utilization of pandas library to load studentperformance.csv file into



The screenshot shows a Google Colab notebook titled 'pdsassignmentqstn2.ipynb'. The code cell contains the following Python code:

```
import matplotlib.pyplot as plt
import seaborn as sns

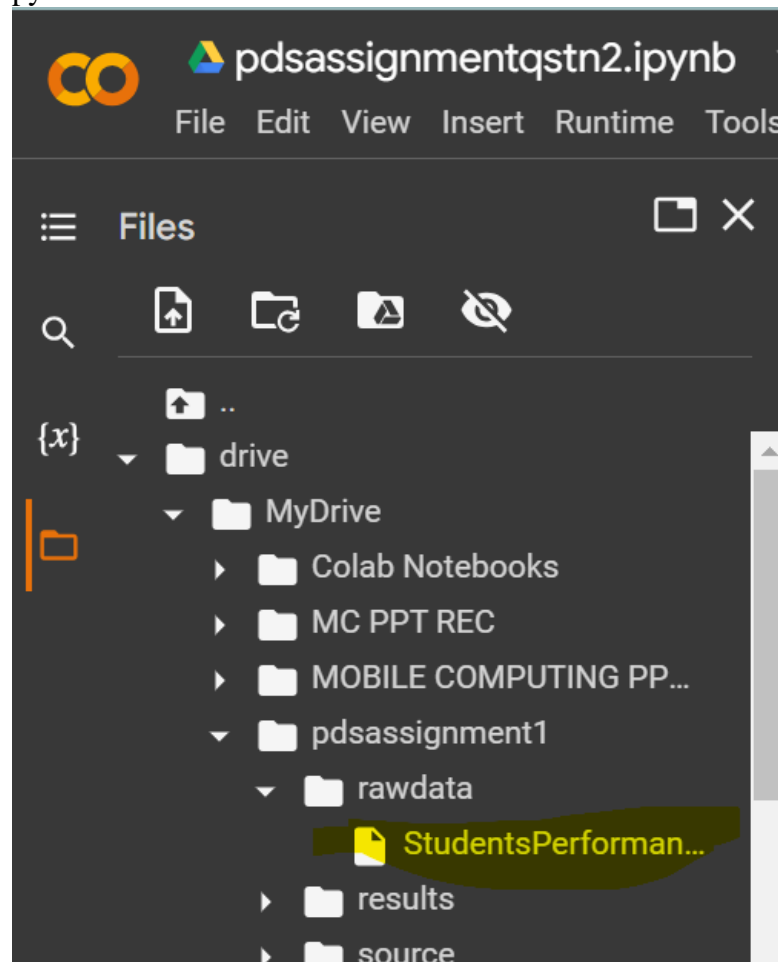
[19] data = pd.read_csv('rawdata/StudentsPerformance.csv')
data.head()
```

The output of the code is a preview of the first five rows of the 'StudentsPerformance.csv' file. The data is as follows:

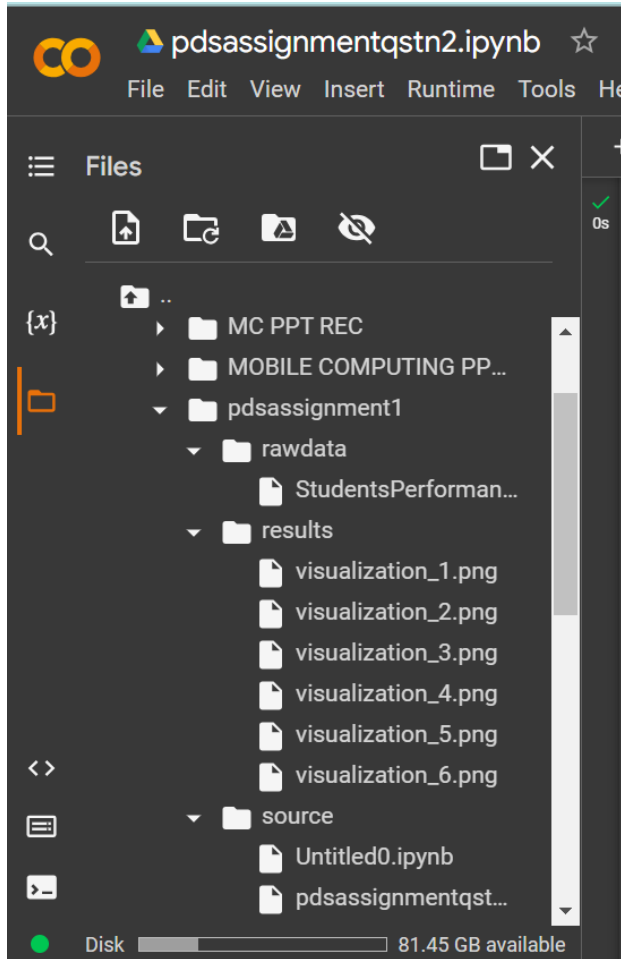
	gender	race/ethnicity	parental level of education	lunch	test preparation course	math score	reading score	writing score
0	female	group B	bachelor's degree	standard	none	72	72	74
1	female	group C	some college	standard	completed	69	90	88
2	female	group B	master's degree	standard	none	90	95	93
3	male	group A	associate's degree	free/reduced	none	47	57	44
4	male	group C	some college	standard	none	76	78	75

Below the table, the output of `data.info()` is shown, indicating the data is a `<class 'pandas.core.frame.DataFrame'>` with 0 rows and completed at 9:21 PM.

python .



The above picture indicates the file structure of the folder created.



This file structure indicates the final folder structure with all the data visualization images obtained.

Step 2: Processing of the data.

The screenshot shows a Google Colab notebook titled 'pdsassignmentqstn2.ipynb'. The left sidebar displays the file explorer with a tree view of the drive, including folders like 'Colab Notebooks', 'MC PPT REC', 'MOBILE COMPUTING PP...', 'pdsassignment1', 'rawdata', 'StudentsPerforman...', 'results', 'source', and files like 'Abstract - Enhancing Crypto...', 'CC Final report.galides', 'CC Final report.pptx', and 'Final Paper.docx'. The main area shows a code cell with the following code:

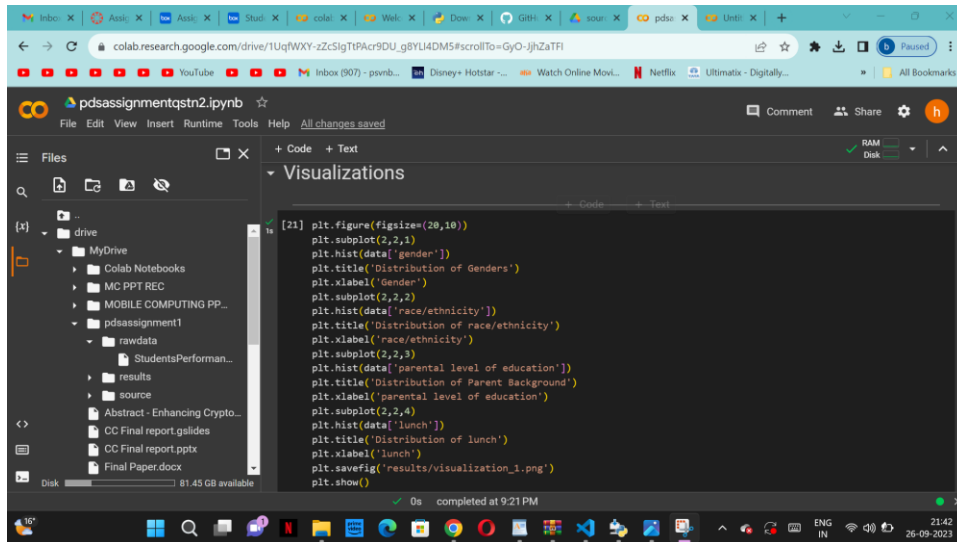
```
data.info()
```

The output of the code cell is displayed below the code:

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1000 entries, 0 to 999
Data columns (total 8 columns):
 #   Column              Non-Null Count  Dtype  
---  --
 0   gender              1000 non-null  object  
 1   race/ethnicity       1000 non-null  object  
 2   parental level of education 1000 non-null  object  
 3   lunch               1000 non-null  object  
 4   test preparation course 1000 non-null  object  
 5   math score          1000 non-null  int64   
 6   reading score       1000 non-null  int64   
 7   writing score        1000 non-null  int64   
dtypes: int64(3), object(5)
memory usage: 62.6+ KB
```

Below the code cell, there is a section titled 'Visualizations' which is currently collapsed. The bottom status bar indicates that the code was completed at 9:21 PM on 26-09-2023.

Step 3: Visualization of the data.

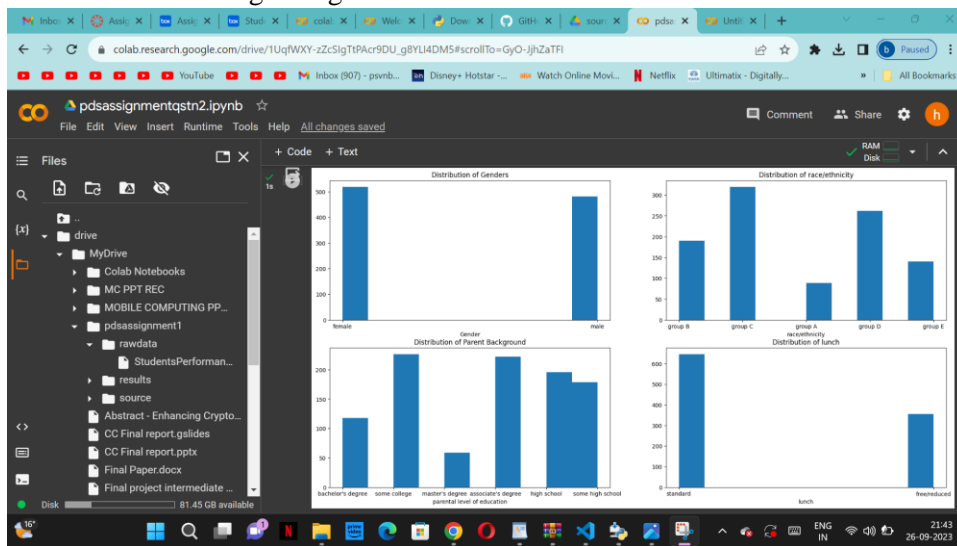


The screenshot shows a Google Colab notebook interface. The left sidebar displays a file explorer with a directory structure including 'drive', 'MyDrive', 'Colab Notebooks', 'MC PPT REC', 'MOBILE COMPUTING PP...', 'pdsassignment1', 'rawdata', 'StudentsPerforman...', 'results', 'source', 'Abstract - Enhancing Crypto...', 'CC Final report.gslides', 'CC Final report.pptx', and 'Final Paper.docx'. The main area is titled 'Visualizations' and contains the following Python code:

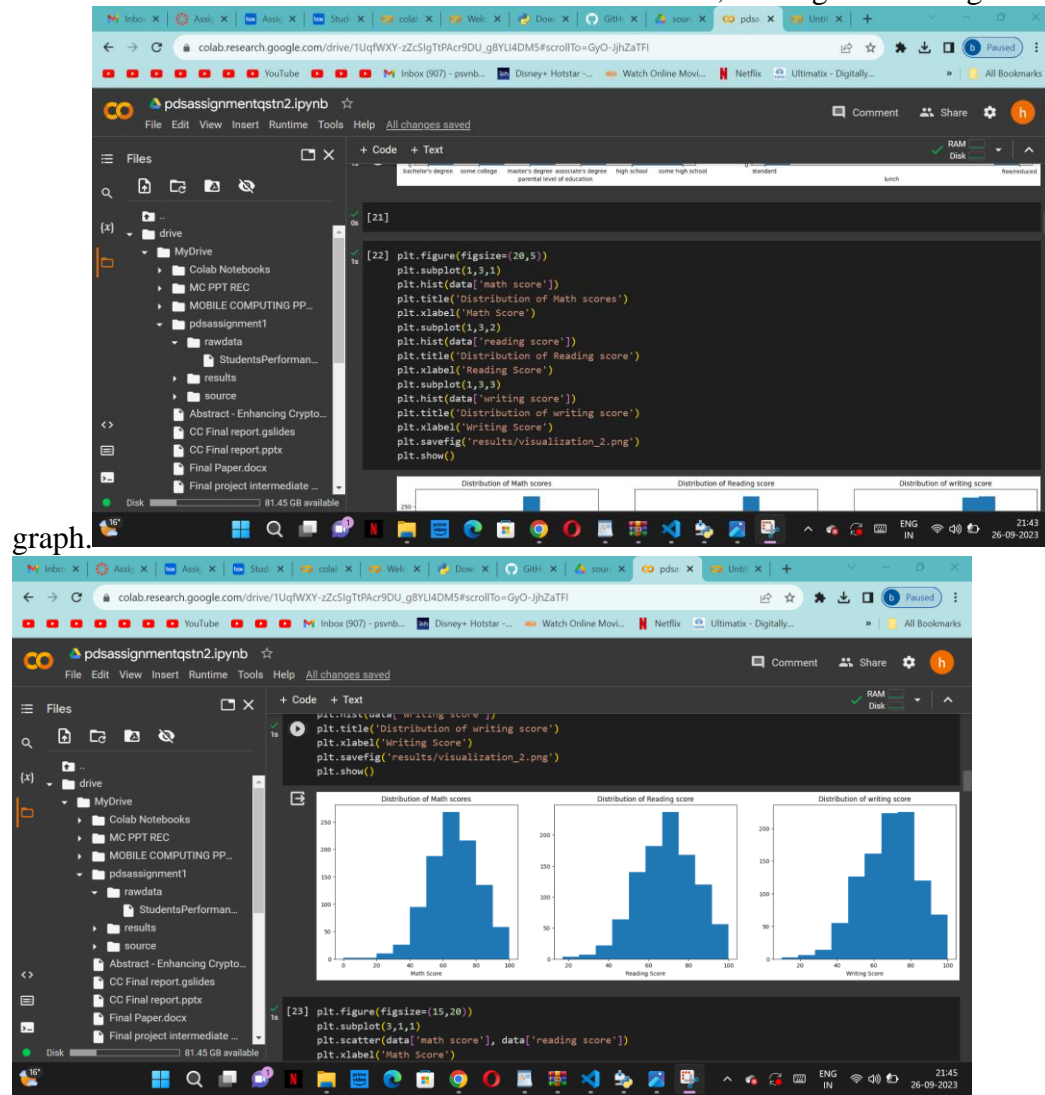
```
[21] plt.figure(figsize=(20,10))
plt.subplot(2,2,1)
plt.hist(data['gender'])
plt.title('Distribution of Genders')
plt.xlabel('Gender')
plt.subplot(2,2,2)
plt.hist(data['race/ethnicity'])
plt.title('Distribution of race/ethnicity')
plt.xlabel('race/ethnicity')
plt.subplot(2,2,3)
plt.hist(data['parental level of education'])
plt.title('Distribution of Parent Background')
plt.xlabel('parental level of education')
plt.subplot(2,2,4)
plt.hist(data['lunch'])
plt.title('Distribution of lunch')
plt.xlabel('lunch')
plt.savefig('results/visualization_1.png')
plt.show()
```

The status bar at the bottom indicates '0s completed at 9:21 PM'.

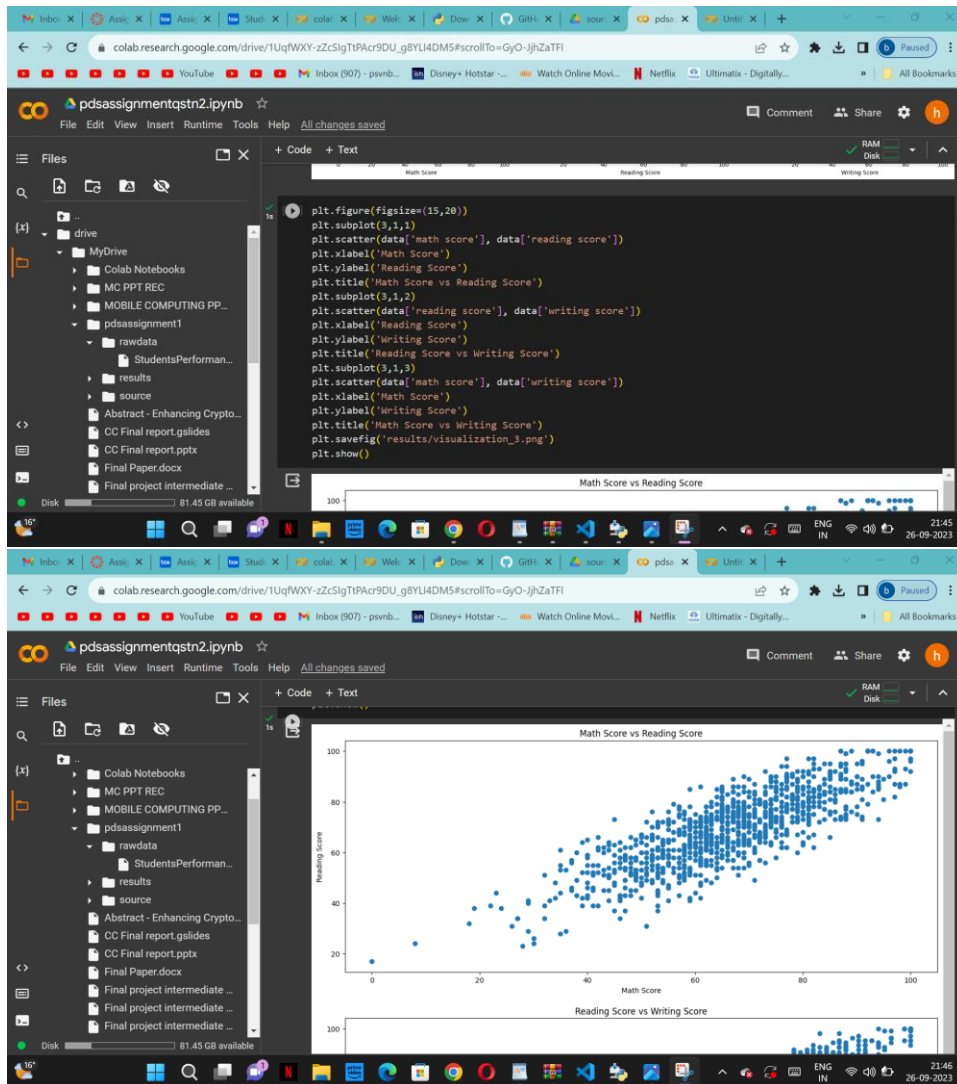
1: Visualization using histogram.

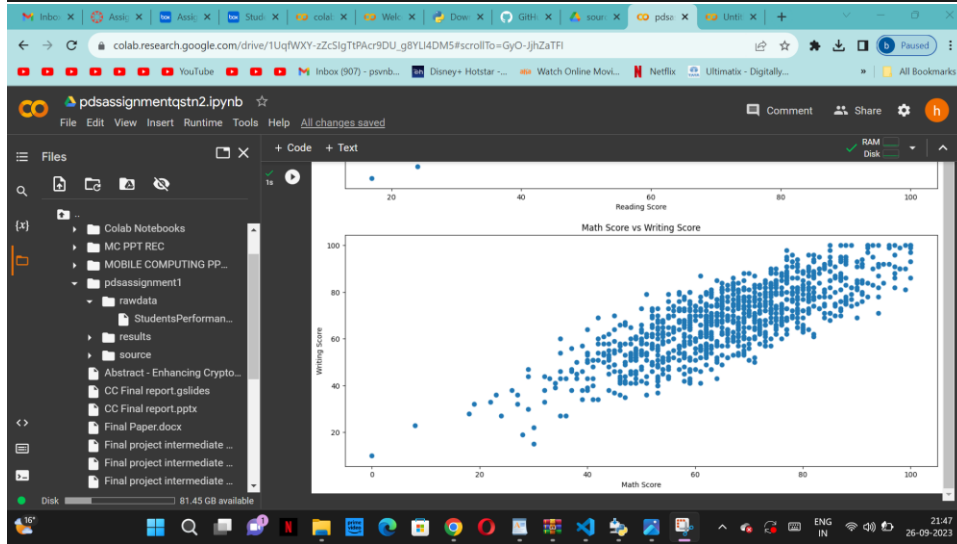
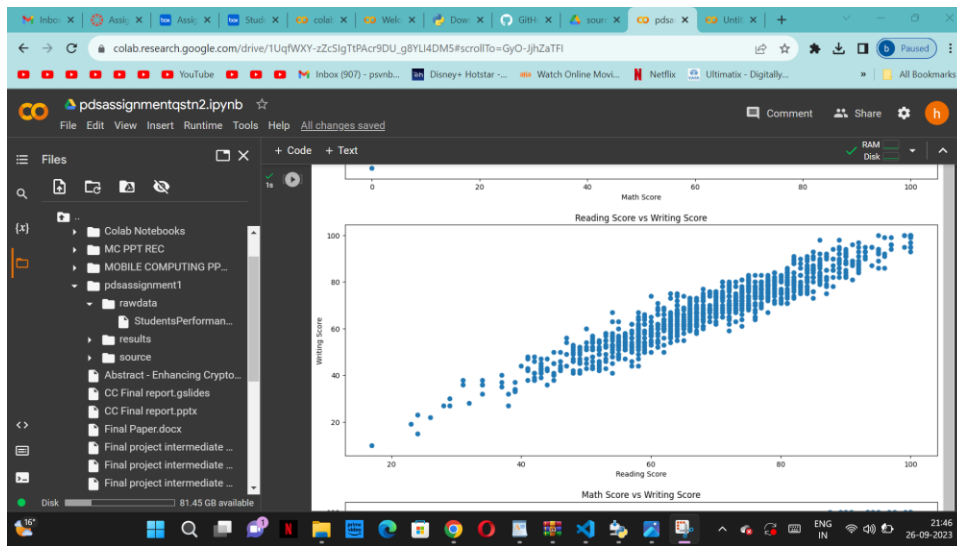


2. Visualization to show distribution between the maths, reading and writing scores on a bar

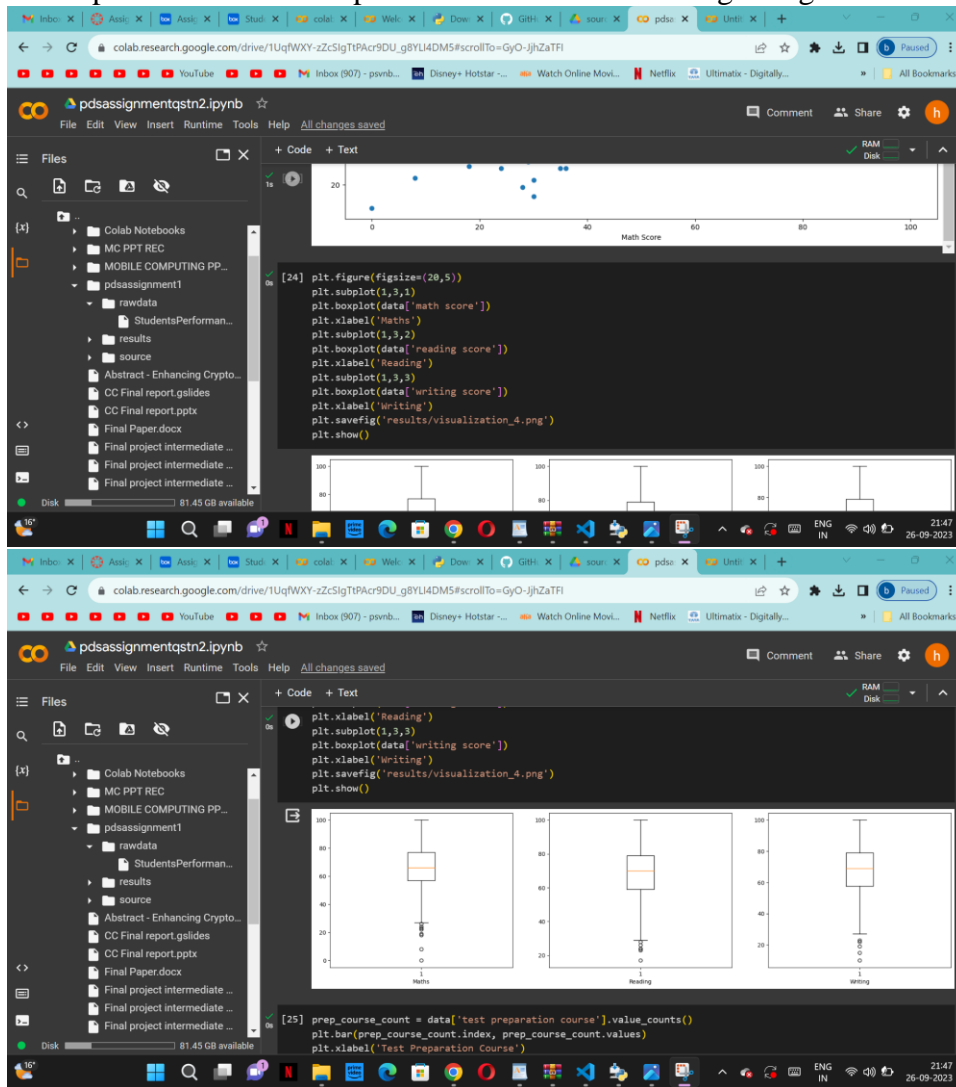


3. Visualization of data in form of scatterplot to show comparison between maths vs reading vs writing scores.

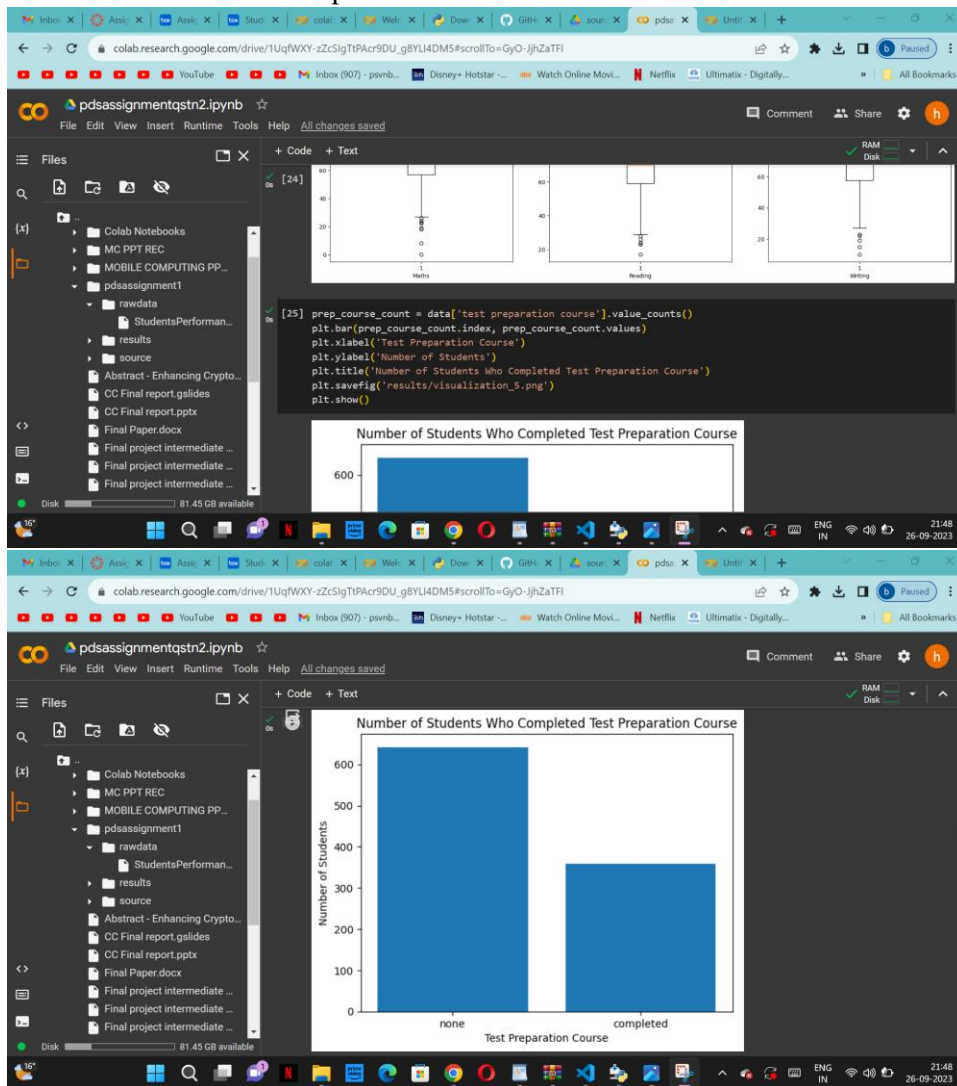




4.Box plot visualization to provide more information regarding the data.



5. Bar chart visualization to provide relation between number of students and test preparation course.



6. Box plot visualization, this visualization can help identify whether there is a difference in test scores between students whose parents have different levels of education.

