Assignment 1

1) Based on the following table, design the three stages of reproducible workflow, includes the work you can do and the folder structure in each stage (reference study case in chapter 3). (5 points)

Height (Inches)	Weight (Pounds)	Age	Grip strength	Frailty
65.8	112	30	30	N
71.5	136	19	31	N
69.4	153	45	29	N
68.2	142	22	28	Y
67.8	144	29	24	Y
68.7	123	50	26	N
69.8	141	51	22	Y
70.1	136	23	20	Y
67.9	112	17	19	N
66.8	120	39	31	N

Stage 1: Raw data

- PDSAssignment (main folder)
  - raw_data (sub-folder)
  - raw_people_data.csv (raw data file)

Stage 2: Data Cleaning

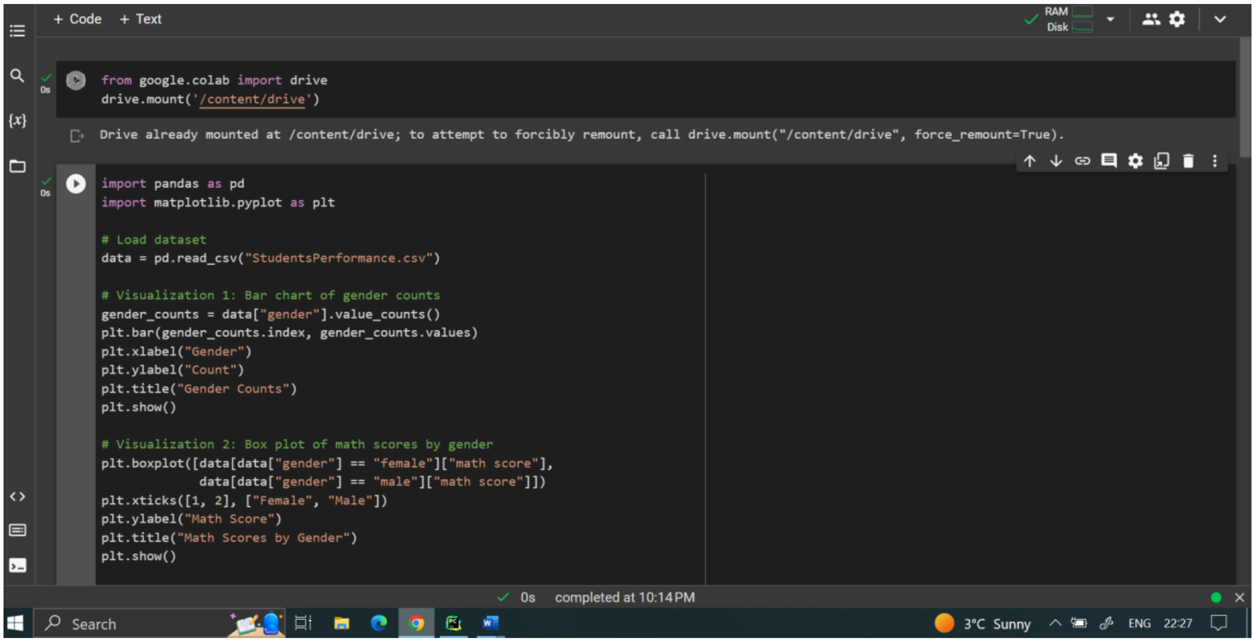
- PDSAssignment (main folder)
  - raw_data (sub-folder)
    - raw_people_data.csv (raw data file)
  - clean_data (sub-folder)
    - clean_people_data.csv (clean data file)
  - data_cleaning_script.py (Python script for cleaning the data)

Stage 3: Data Analysis and Visualization

- PDSAssignment (main folder)
  - raw_data (sub-folder)
  - raw_people_data.csv (raw data file)
  - clean_data (sub-folder)

- clean_people_data.csv (clean data file)
- data_cleaning_script.py (Python script for cleaning the data)
- data_analysis_script.py (Python script for analyzing and visualizing the data)
- age_distribution.png (image file showing the age distribution by frailty status)

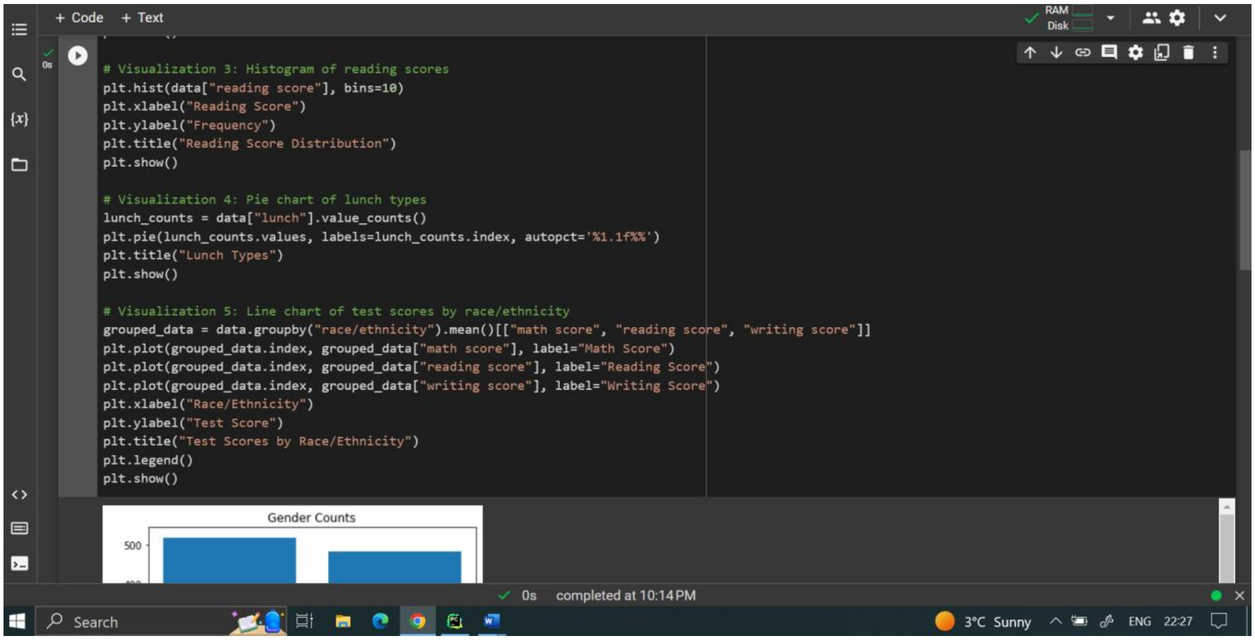
2) Perform 5 data visualization tasks on the student performance dataset given in the link below (create 5 different visualizations). Explain what kind analysis has become easier with each of the visualizations. Create the folder structure for this question similar to question 1. (15 points).



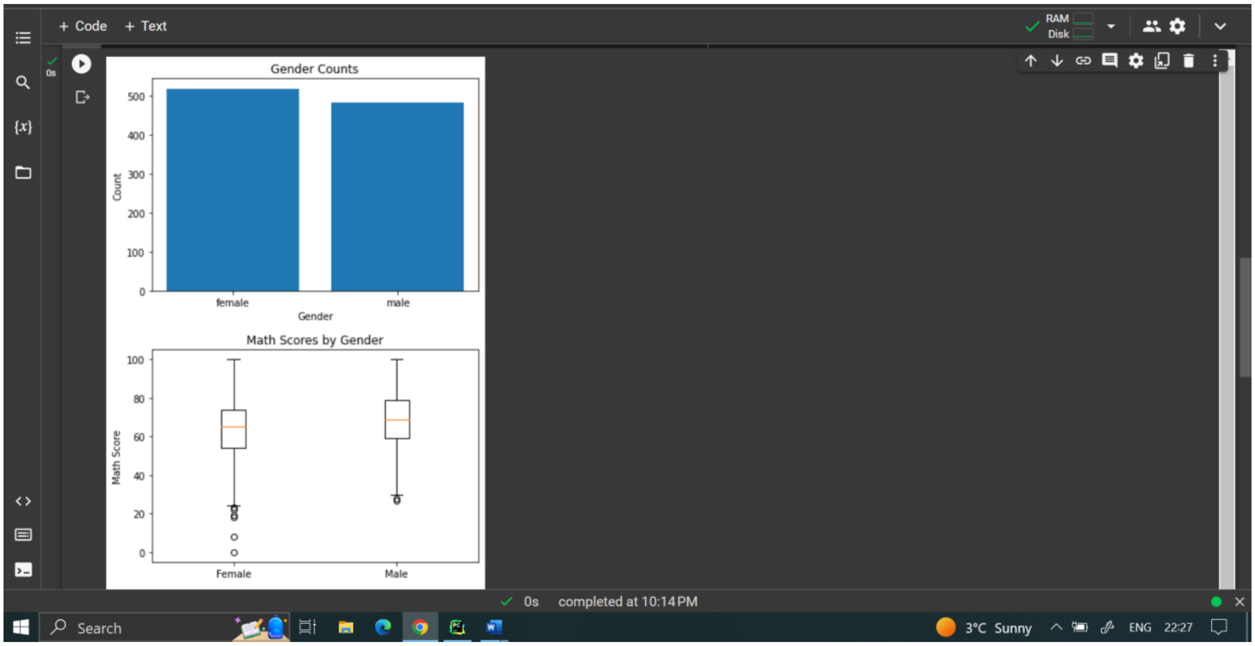
```
from google.colab import drive
drive.mount('/content/drive')
```

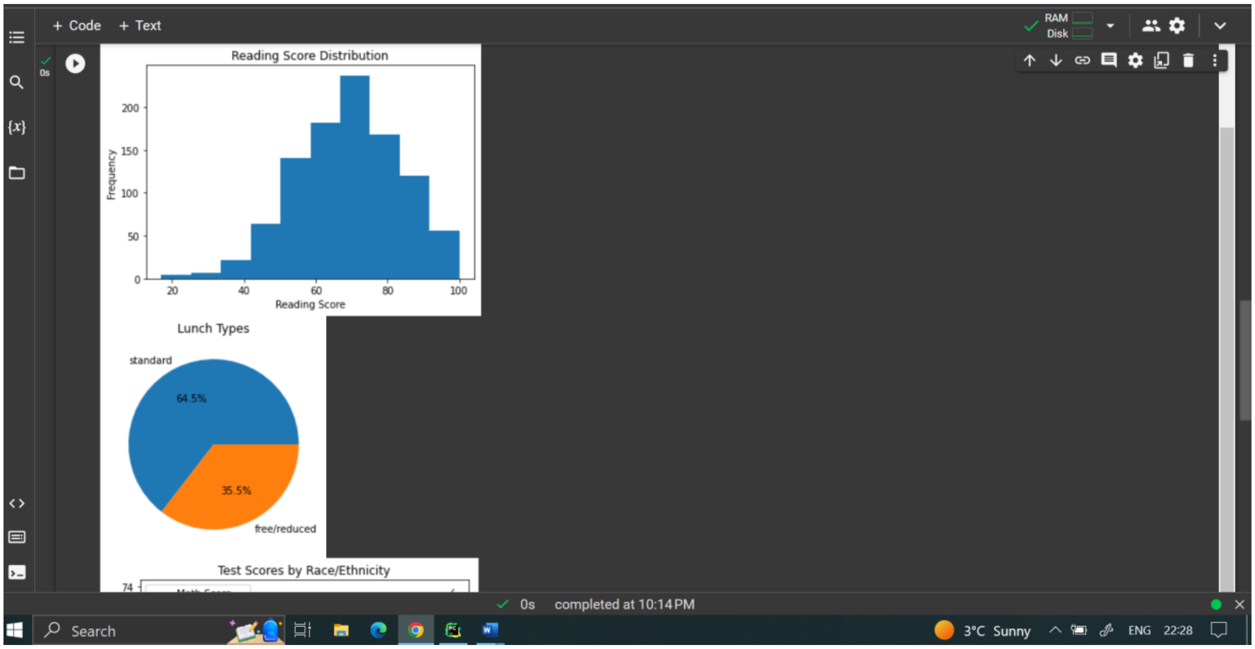
```
Drive already mounted at /content/drive; to attempt to forcibly remount, call drive.mount("/content/drive", force_remount=True).
```

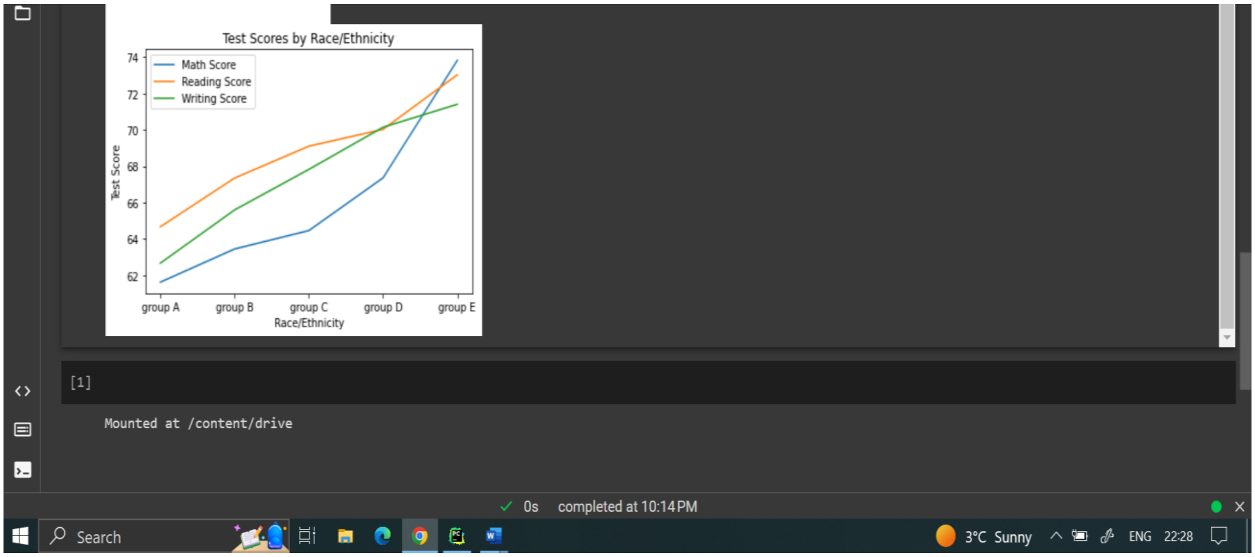
```
import pandas as pd
import matplotlib.pyplot as plt

# Load dataset
data = pd.read_csv("StudentsPerformance.csv")

# Visualization 1: Bar chart of gender counts
gender_counts = data["gender"].value_counts()
plt.bar(gender_counts.index, gender_counts.values)
plt.xlabel("Gender")
plt.ylabel("Count")
plt.title("Gender Counts")
plt.show()

# Visualization 2: Box plot of math scores by gender
plt.boxplot([data[data["gender"] == "female"]["math score"],
             data[data["gender"] == "male"]["math score"]])
plt.xticks([1, 2], ["Female", "Male"])
plt.ylabel("Math Score")
plt.title("Math Scores by Gender")
plt.show()
```



```
# Visualization 3: Histogram of reading scores
plt.hist(data["reading score"], bins=10)
plt.xlabel("Reading Score")
plt.ylabel("Frequency")
plt.title("Reading Score Distribution")
plt.show()

# Visualization 4: Pie chart of lunch types
lunch_counts = data["lunch"].value_counts()
plt.pie(lunch_counts.values, labels=lunch_counts.index, autopct='%1.1f%%')
plt.title("Lunch Types")
plt.show()

# Visualization 5: Line chart of test scores by race/ethnicity
grouped_data = data.groupby("race/ethnicity").mean()[["math score", "reading score", "writing score"]]
plt.plot(grouped_data.index, grouped_data["math score"], label="Math Score")
plt.plot(grouped_data.index, grouped_data["reading score"], label="Reading Score")
plt.plot(grouped_data.index, grouped_data["writing score"], label="Writing Score")
plt.xlabel("Race/Ethnicity")
plt.ylabel("Test Score")
plt.title("Test Scores by Race/Ethnicity")
plt.legend()
plt.show()
```







```
[1]

Mounted at /content/drive
```

- Visualization 1 (Gender Counts Bar Chart): This visualization makes it easier to compare the number of male and female students in the dataset. This can be useful for performing gender-based analyses, such as comparing test scores by gender or investigating gender-based performance gaps.

- Visualization 2 (Math Scores by Gender Box Plot): This visualization makes it easier to compare the distribution of math scores between male and female students. It shows the median, quartiles, and outliers for each gender, making it easier to identify differences in the distributions.

- Visualization 3 (Reading Score Distribution Histogram): This visualization makes it easier to see the distribution of reading scores in the dataset. It shows the frequency of scores in each range, making it easier to identify patterns such as clusters or outliers.

- Visualization 4 (Lunch Types Pie Chart): This visualization makes it easier to compare the distribution of lunch types among the students. This can be useful for investigating the effect of lunch type on test scores or comparing the socioeconomic status of students with different lunch types.

- Visualization 5 (Test Scores by Race/Ethnicity Line Chart): This visualization makes it easier to compare the average test scores for different racial/ethnic groups. It shows the trend of scores across the different groups, making it easier to identify which groups have higher or lower scores on average. This can be useful for investigating performance gaps based on race/ethnicity or identifying which groups may need additional support.

Folder Structure:

```
project_folder/
| ├── data_raw/
|       └── StudentsPerformance.csv
| ├── data_cleaned/
|        └── StudentsPerformance_cleaned.csv
|
├── results/
 |    ├── plots/
|        |   ├── math_scores_hist.png
|        |   ├── reading_scores_hist.png
|        |   ├── writing_scores_hist.png
|          |    ├── test_scores_boxplot.png
|          |    |— gender_scores_boxplot.png
|          ├── summary_stats.csv
|          |— hypothesis_tests.csv
|
|— student_performance_workflow.py
```

Submission:

Create a public GitHub repo and upload the folders for both the questions on the GitHub and submit the link to Canvas.