Assignment 1, data analysis

All assignments must be done in anaconda environment by the use of Jupyter notebook

Part 1 [30 points]

- Create a NumPy array named "my_array" with the following values: [10, 20, 30, 40, 50].
- Calculate and print the mean, median, and standard deviation of "my_array".
- Import the Pandas library and create a DataFrame named "my_dataframe" with the following data:

Name	Age	Gender
Alice	25	F
Bob	30	M
Charlie	35	M
David	40	M
Emma	28	Prefer not to say

- Display the first two rows of "my_dataframe".
- Calculate and print the maximum age from "my_dataframe".

Part 2 [10 points]

- Import the Matplotlib library and use it to plot a line graph with the following data:
  - X-axis values: [1, 2, 3, 4, 5]
  - Y-axis values: [10, 15, 7, 20, 12]
- Add labels to the X-axis and Y-axis of the plot as "X" and "Y" respectively.
- Set the title of the plot as "Line Graph".
- Save the plot as a PNG file named "line_graph.png".

-

Part 3 [60 points]

You will be working with a dataset named "sales_data.csv". The dataset contains sales data with the following columns:
Order ID: Unique identifier for each order.
Date: Date of the order.
Product: Product name.
Price: Price of the product.
Quantity: Quantity ordered.
Total: Total sales for each order.
Tasks:

- Load the "sales_data.csv" dataset into a Pandas DataFrame named "sales_df". Ensure that the necessary Pandas library is imported.
- Display the first 5 rows of the "sales_df" DataFrame.
- Check the data types of each column in the DataFrame and ensure they are appropriate. Convert the "Date" column to the datetime data type.
- Check for missing values in the DataFrame and handle them appropriately. Remove any rows with missing values.

- Clean the "Price" column by removing any leading or trailing whitespace and convert it to a numeric data type.
- Clean the "Quantity" column by removing any non-numeric characters and convert it to a numeric data type.
- Create a new column named "Revenue" in the DataFrame by multiplying the "Price" and "Quantity" columns
- Calculate and print the average revenue per order in the dataset.

здесь обратите внимание что в датасете order_id должен повторяться в некоторых местах тк в одном заказе может быть несколько товаров.
Также quantity должен содержать помимо числовых значений также и текстовые, далее вы их удалите (это один из подпунктов задания)