**ASSIGGNMENT -1**

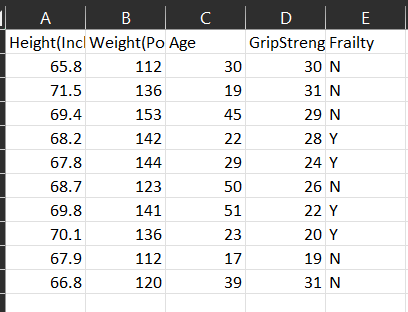
**Name: Sowmya Patlolla**

**Std Id: 16338256**

**Frailty Analysis**

**Step 1: Data Collection**

* The given data is manually inputted in the spreadsheet and saved as CSV file format and named as “raw\_yield\_data.csv”.
* The data represents 10 different persons with their age, weight, height, grip strength and the person with frailty or not(Y,N).



**Step 2: Data Preprocessing**

* In the given raw data, there is no need to remove or clean to proceed to further analysis, but just in case we have large data. So, I’m checking for null values in the dataset.
* I have used df.isnull().sum() to check for missing values in each column of the DataFrame. This is a crucial data processing step as handling missing data is essential for accurate analysis. In this case, I didn’t find any missing values in any of the columns.
* After that, I’m moving cleaned data file i.e.,” [clean\_raw\_yield\_data](http://localhost:8888/edit/OneDrive/Desktop/Sowmya_PDS_Assignment1/FrailtyAnalysis/Clean_data/clean_raw_yield_data)” into a new folder named “Clean\_data”.

A screenshot of a computer

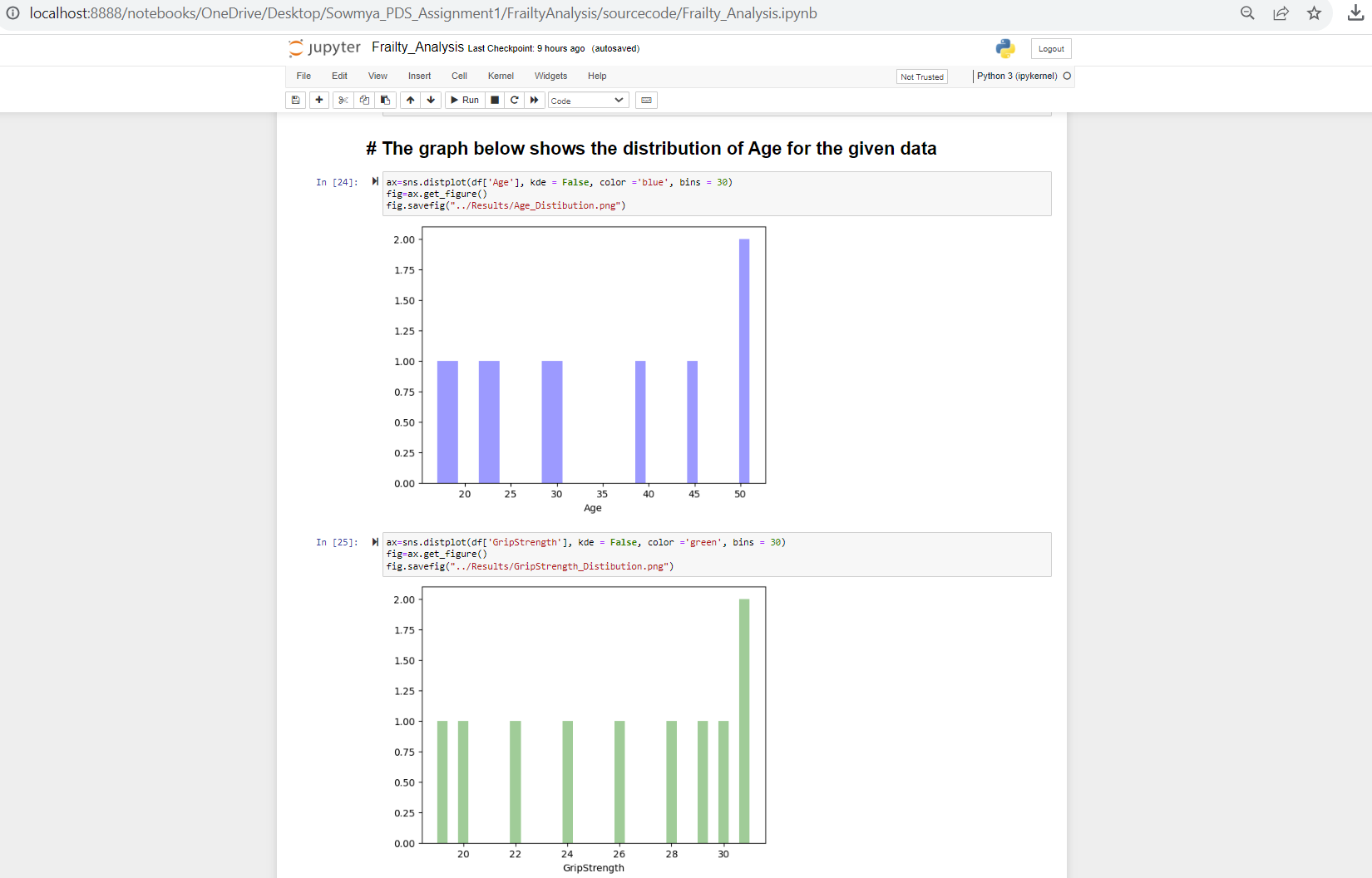
Description automatically generated

**Step 3: Data Analysis**

* First, I have imported the required panda’s package for analysis and then I have loaded the cleaned data file for further analysis.
* I have imported matplotlib, seaborn and NumPy libraries for plotting the analysis that made on the cleaned data.

A screenshot of a computer

Description automatically generated

****

**A screenshot of a computer

Description automatically generated**

**Student Performance Analysis**

**Step 1: Data Collection**

* The data is collected from the below link, that has been provided in the question and stored that data in the “data\_raw “folder.

**A screenshot of a computer

Description automatically generated**

**Step 2: Data Preprocessing**

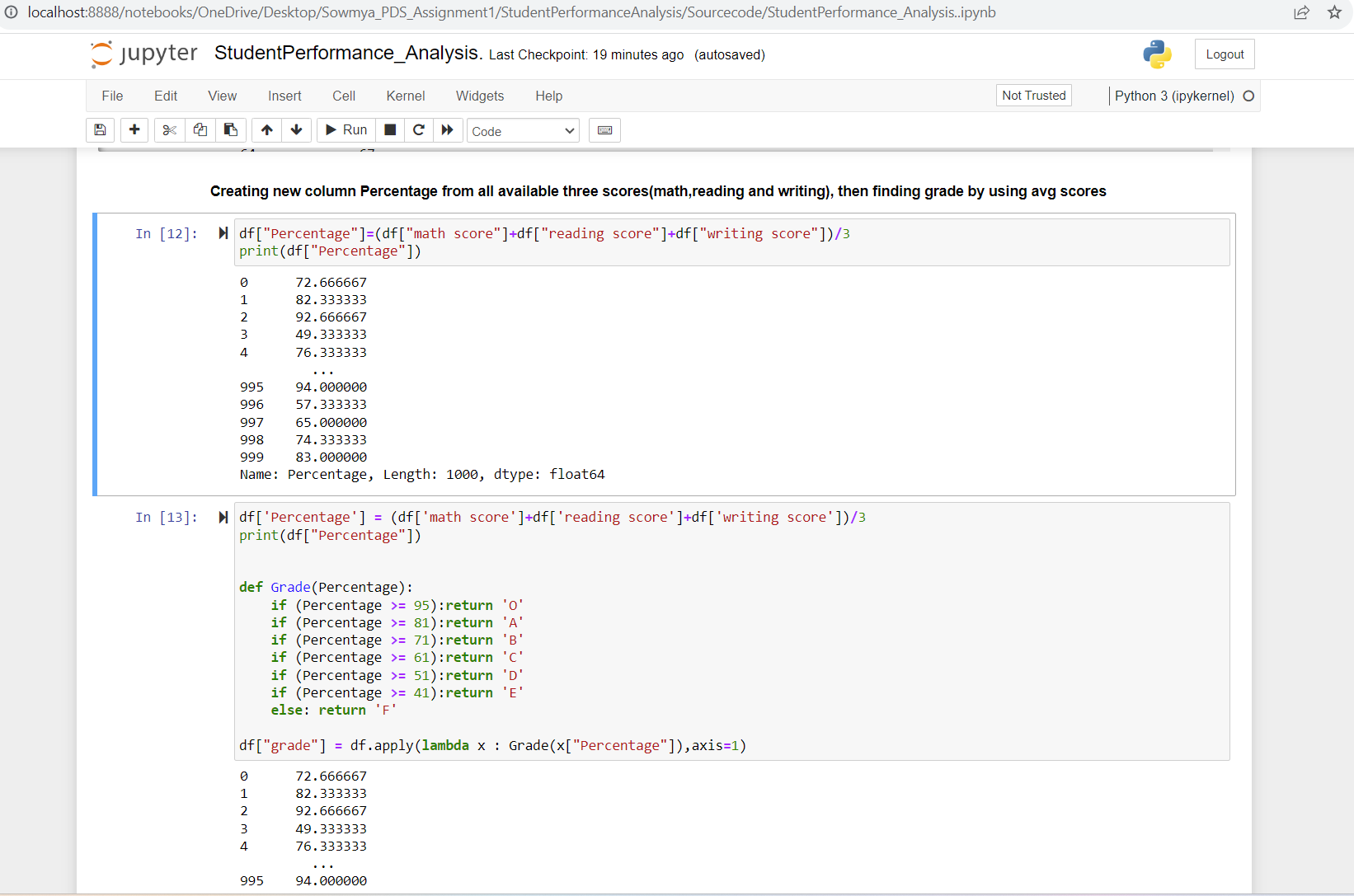
* I have used df.isnull().sum() to check for missing values in each column of the DataFrame. This is a crucial data processing step as handling missing data is essential for accurate analysis. In this case, I didn’t find any missing values in any of the columns.
* After that, I’m moving cleaned data file i.e.,” [Cleaned\_StudentsPerformance.csv](http://localhost:8888/edit/OneDrive/Desktop/Sowmya_PDS_Assignment1/StudentPerformanceAnalysis/data_clean/Cleaned_StudentsPerformance.csv)” into a new folder named “data\_clean”.

**A screenshot of a computer

Description automatically generated**

**Step 3: Data Analysis**

* First, I imported the required panda’s package for analysis and then I loaded the cleaned data file for further analysis.’
* Creating new column Percentage from all available three scores (math, reading and writing), then finding grade by using avg scores.
* I have defined a function called "Grade" that assigns a grade based on the percentage score. This value is then applied to each row in the DataFrame to create a new column named "grade”.



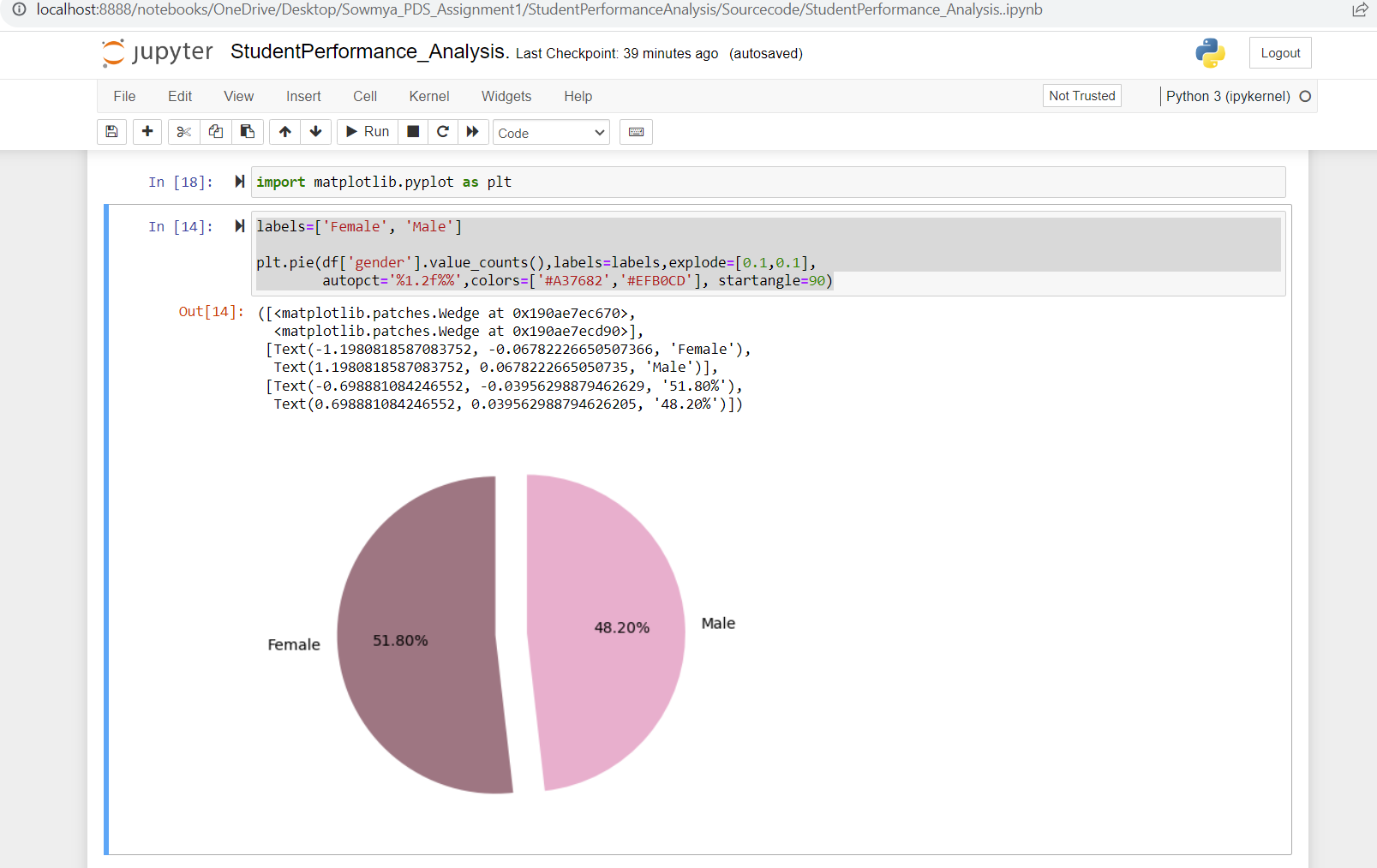
**Step 4: Data Visualization**

* First, I have imported the required python libraries for performing visualized analysis by considering the various features of the cleaned data.
* I have used barplot to analyze the relation between percentage and gender, it compares the standard deviation of the "Percentage" variable across different genders.
* Each bar represents a gender category.
* Larger standard deviations indicate greater variability in the "Percentage" variable for that gender category.
* Smaller standard deviations indicate less variability and greater consistency.

**A screenshot of a computer screen

Description automatically generated**

* Below, I’m creating a pie chart using Matplotlib to visualize the distribution of values in the "gender" column of your DataFrame (df).
* The chart shows two segments, one for "Female" and one for "Male."
* The segment sizes represent the proportion of each gender category in the dataset.
* The explosion effect creates a slight separation of the segments for better visibility.
* In this case, "Female" appears to be more prevalent than "Male."
* The percentages displayed inside the chart provide precise information about the proportion of each category.

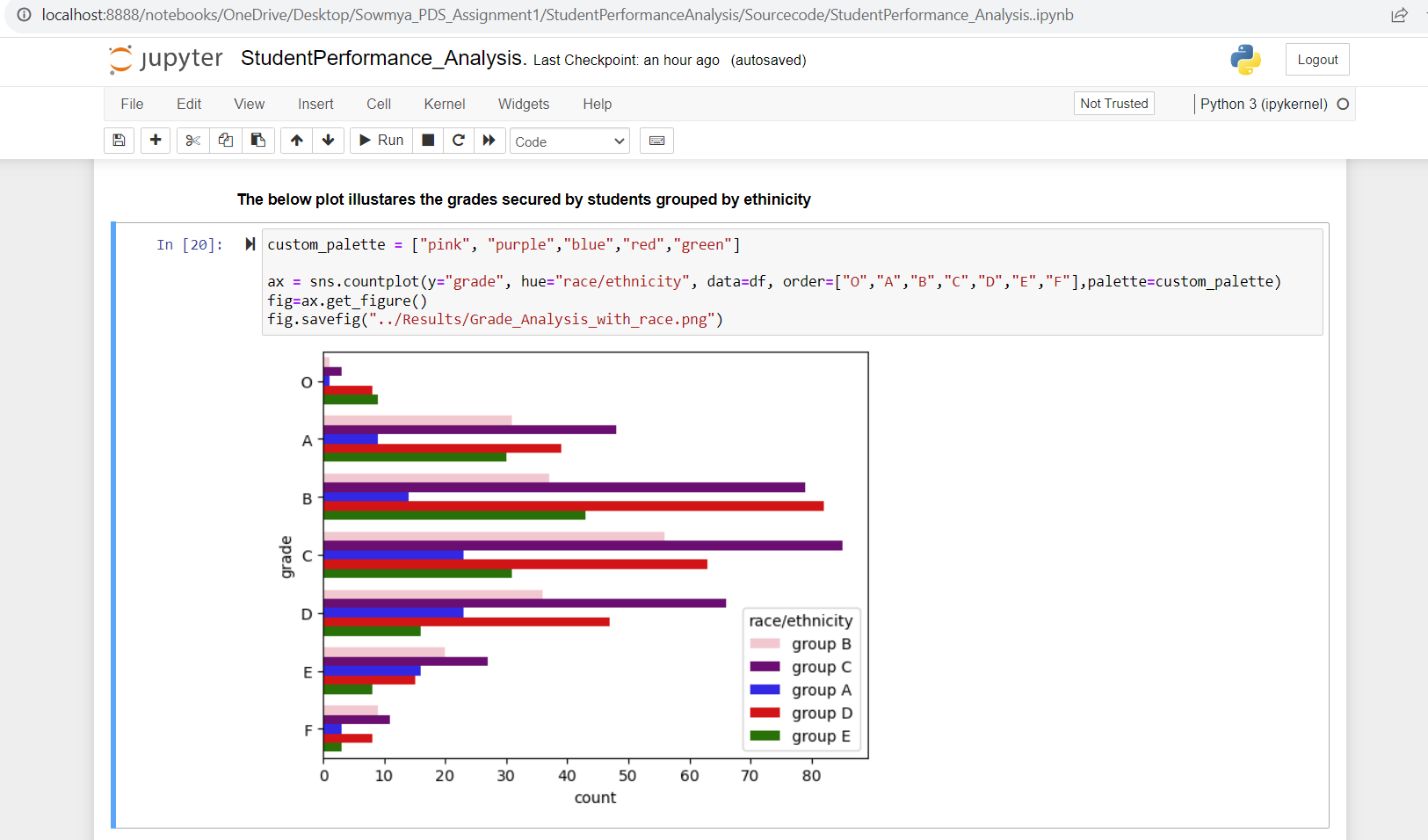


* Below, I’m creating a countplot using Seaborn to visualize the distribution of grades ("grade" variable) by gender ("gender" variable) and customizes the colors of the bars.
* The countplot visually represents the distribution of grades ("O," "A," "B," "C," "D," "E," "F") for each gender ("Female" and "Male").
* The bars represent the count of each grade category, and the bars are grouped by gender, so we can see how grades are distributed within each gender category.
* This plot allows us to compare the grade distributions between genders. So, we can see how many students from each gender received each grade category.

**A screenshot of a graph

Description automatically generated**

* Below, I’m creating a countplot using Seaborn to visualize the distribution of grades ("grade" variable) by race/ethnicity ("race/ethnicity" variable).
* The countplot visually represents the distribution of grades ("O," "A," "B," "C," "D," "E," "F") for each race/ethnicity category.
* The bars represent the count of each grade category, and the bars are grouped by race/ethnicity, so we can see how grades are distributed within each race/ethnicity category.
* This plot allows us to compare the grade distributions among different race/ethnicity categories.
* We can see how many students from each race/ethnicity category received each grade category.

****

* Below, I’m creating a countplot using Seaborn to visualize the distribution of grades ("grade" variable) by parental level of education ("parental level of education" variable).
* The countplot visually represents the distribution of grades ("O," "A," "B," "C," "D," "E," "F") for each parental level of education category.
* The bars represent the count of each grade category, and the bars are grouped by parental education level, allowing you to see how grades are distributed within each parental education category.
* This plot allows us to compare the grade distributions among different parental education levels.
* We can see how many students with parents in each education level received each grade category.

A screenshot of a graph

Description automatically generated

* Below, I’m creating countplot using Seaborn to visualize the distribution of grades ("grade" variable) by lunch type ("lunch" variable).
* The countplot visually represents the distribution of grades ("O," "A," "B," "C," "D," "E," "F") for each lunch type category.
* The bars represent the count of each grade category, and the bars are grouped by lunch type, allowing us to see how grades are distributed within each lunch type category.
* This plot allows you to compare the grade distributions between different lunch types.
* You can see how many students with different lunch types received each grade category.

A screenshot of a computer

Description automatically generated

* Below, I’m creating countplot using Seaborn to visualize the distribution of grades ("grade" variable) by test preparation ("test preparation" variable).
* The countplot visually represents the distribution of grades ("O," "A," "B," "C," "D," "E," "F") for each test preparation course category.
* The bars represent the count of each grade category, and the bars are grouped by test preparation course, allowing us to see how grades are distributed within each course category.
* This plot allows us to compare the grade distributions between students who took the test preparation course and those who did not.
* We can see how many students from each group received each grade category.

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer screen

Description automatically generated

**Conclusion:**

Based on the above all visualization analysis, we can that It's easier to compare and assess potential gender-based differences in academic performance using that visualization. It facilitates the identification of disparities or patterns in academic performance among various racial and ethnic groups. It assesses whether parental education has an impact on academic performance, making it easier to explore this relationship.It facilitates the investigation of whether the type of lunch (standard or free/reduced) affects academic performance. It assesses the impact of test preparation on academic performance and makes it easier to compare the grades of prepared and unprepared students.

In each case, the visualizations provide a clear and intuitive representation of the data, making it easier to draw initial insights and comparisons. However, it's important to note that these visualizations serve as a starting point for analysis.

The choice of analysis will ultimately depend on our research questions and the specific insights we want to gain from the data. These visualizations serve as valuable tools for data exploration and hypothesis generation.