**Big Data Solution Architecture, School of Applied Computer Science & Information Technology**

**PROG8420-23S-Sec1-Programming for Big Data**

**Assignment 10**

**Instructor: Jomis Varikayanickal John**

Tasks:

1. Task 1: Dataset Selection

Dataset url:

<https://github.com/awesomedata/awesome-public-datasets/blob/master/Datasets/titanic.csv.zip>

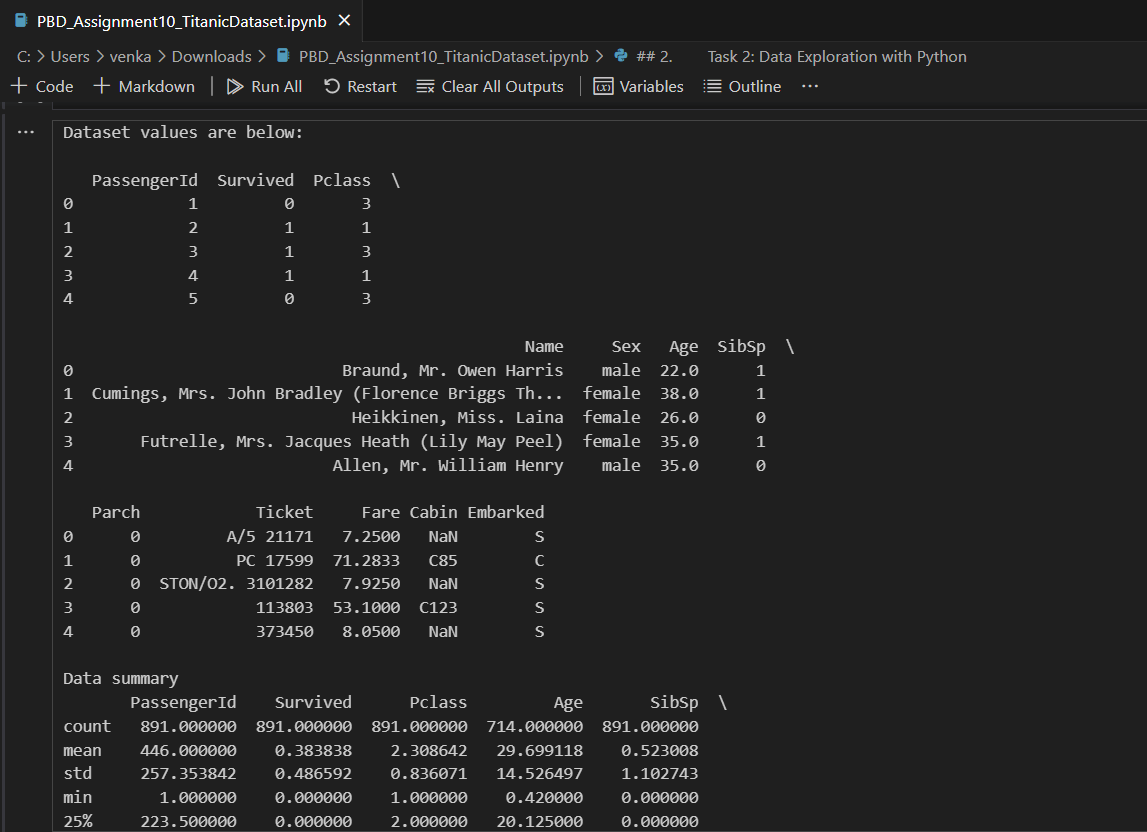
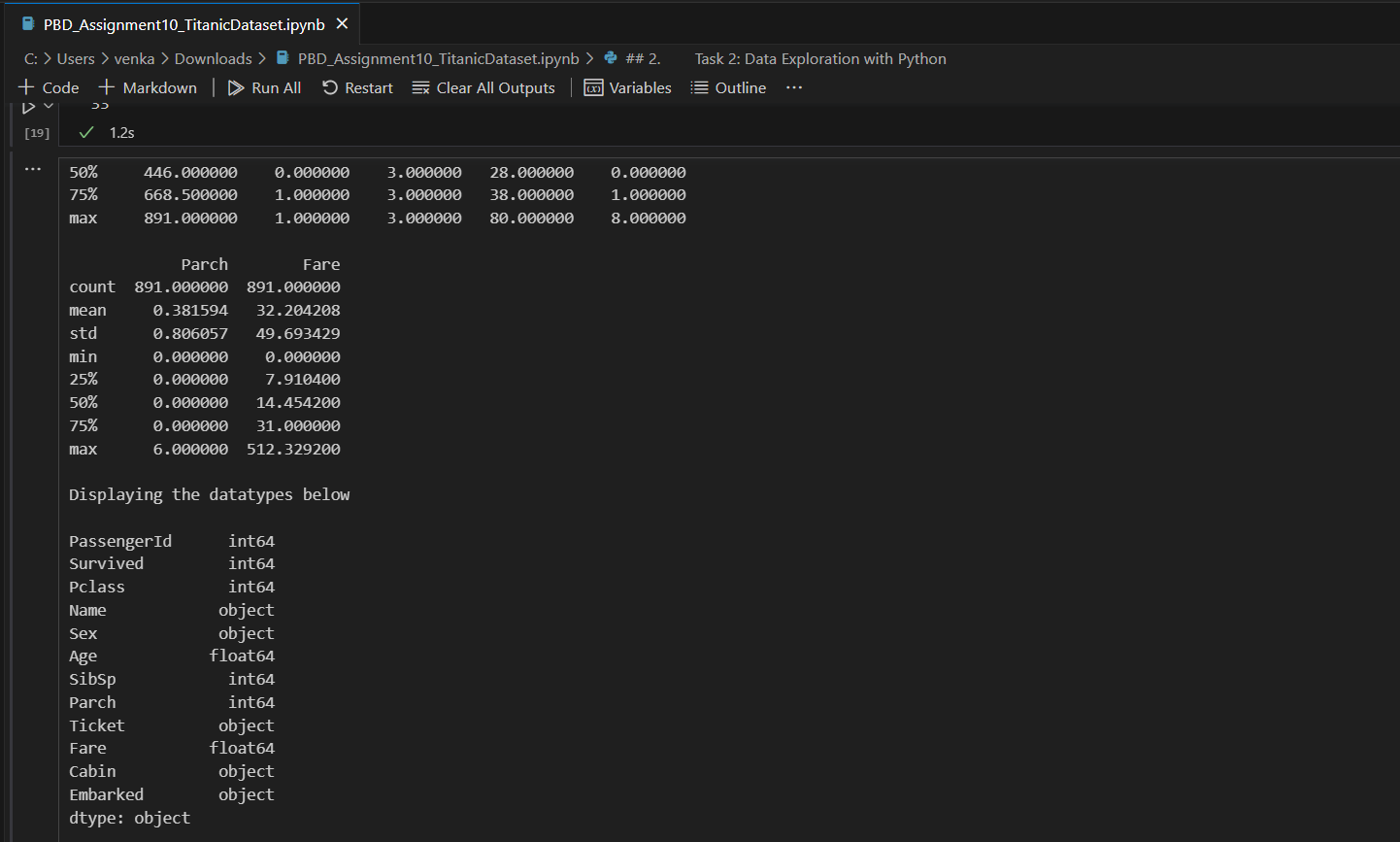
Git url: https://github.com/venkatapechetty-01/Lab-10

Justification:

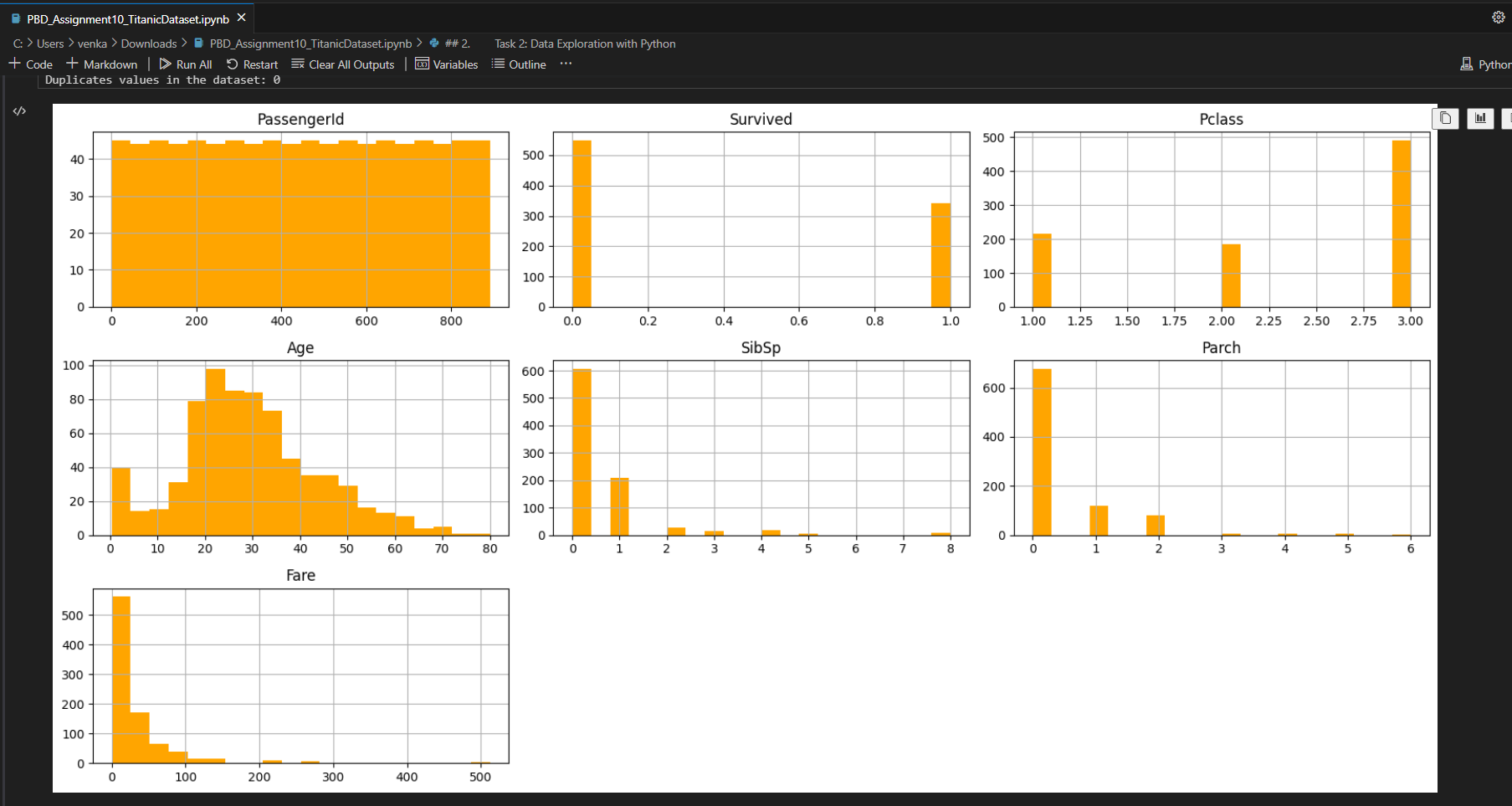
The contents of the Titanic dataset are a well-known the data collection that is crucial for predictive modeling. It contains information on the passengers of the Titanic, such as their age, gender, class, and whether or not they survived. This dataset can be used for binary classification tasks where the objective is to decide on whether or not a passenger survived using the available features. Managing missing values and categorical categories, which are common in datasets from the real world, is another challenge the dataset presents for data preprocessing. Through the examination of this dataset, can help to explore more complex preprocessing techniques and model evaluation.

1. Task 2: Data Exploration with Python
   1. Perform exploratory data analysis (EDA) using Python for the first dataset.
   2. Generate summary statistics, identify data types, and visualize the data distribution to gain insights into the dataset.

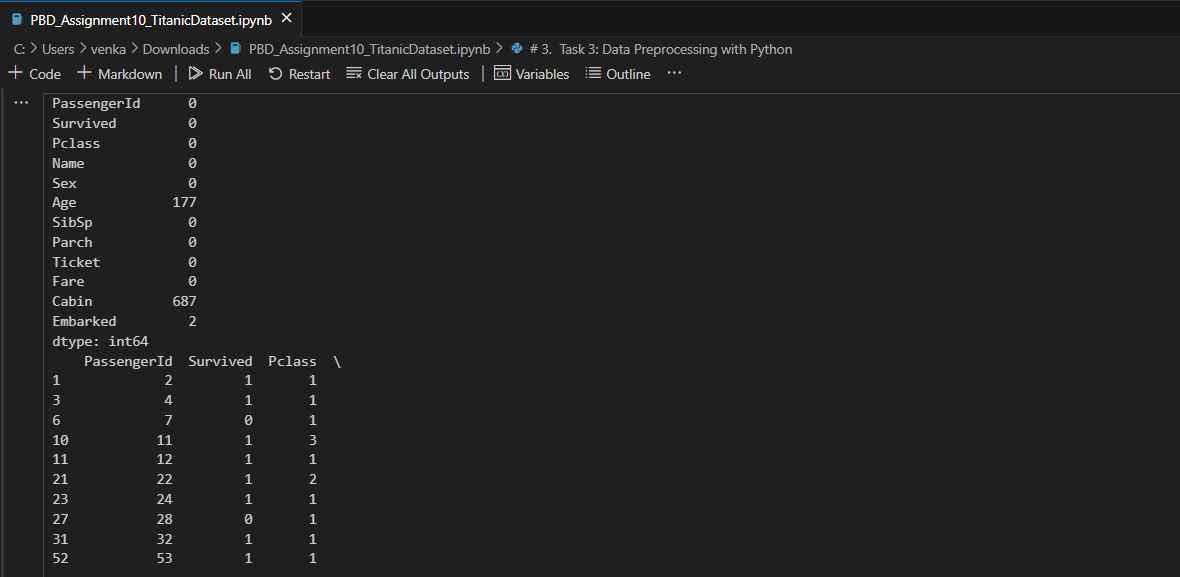
…continued

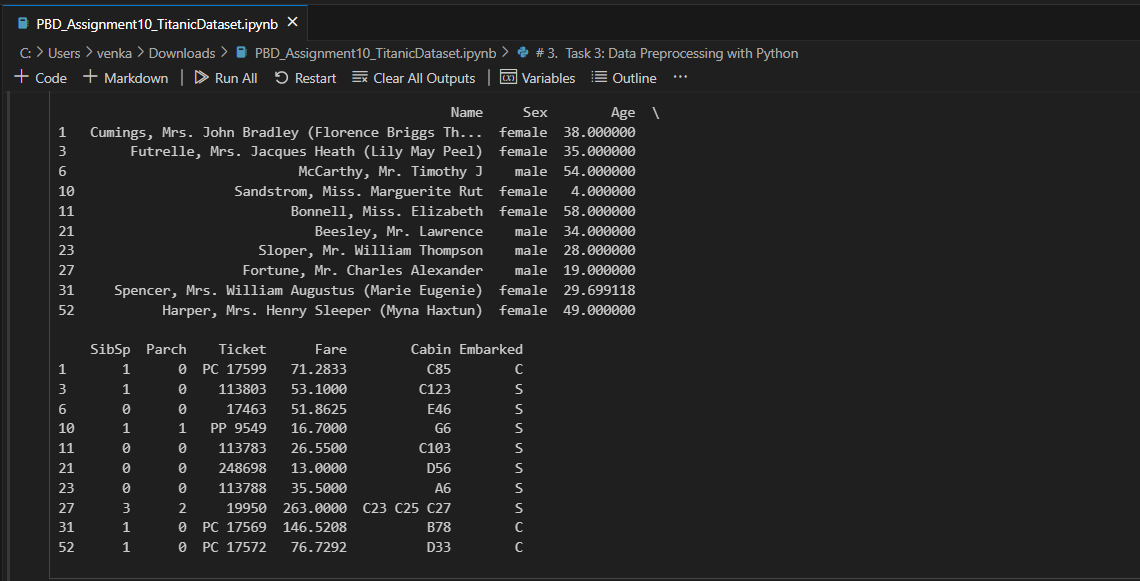
 

Histograms for all the numerical columns

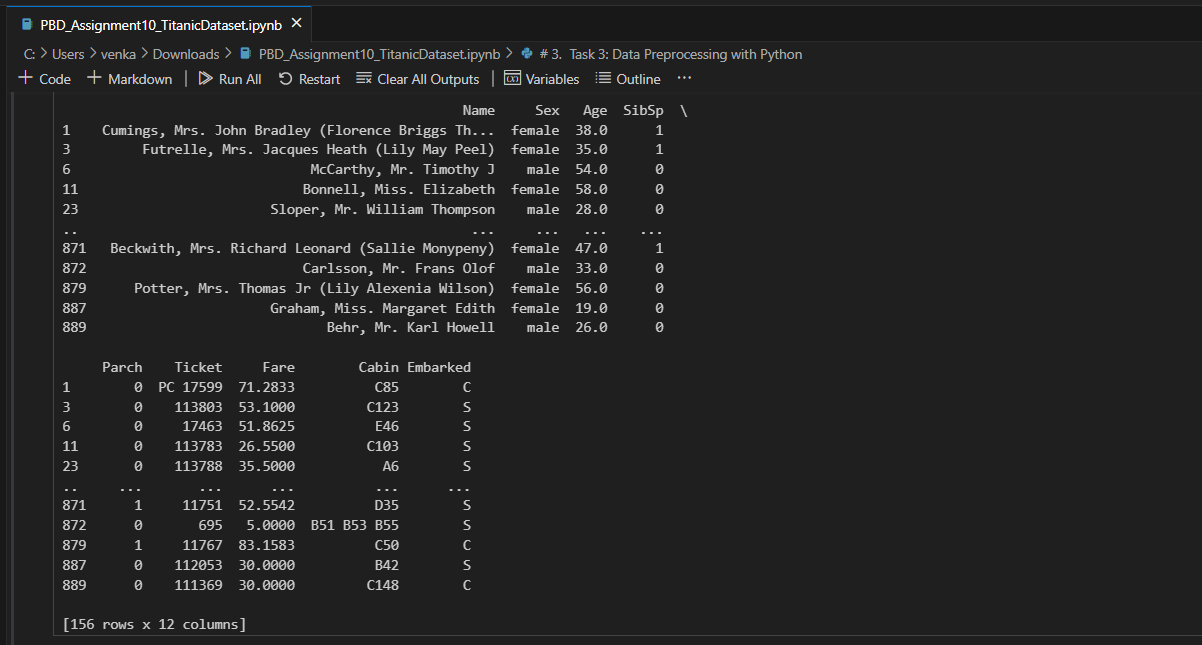


1. Task 3: Data Preprocessing with Python
   1. Preprocess the data from the first dataset using Python.
   2. Handle missing values, outliers, and perform feature engineering when necessary to prepare the data for machine learning models.



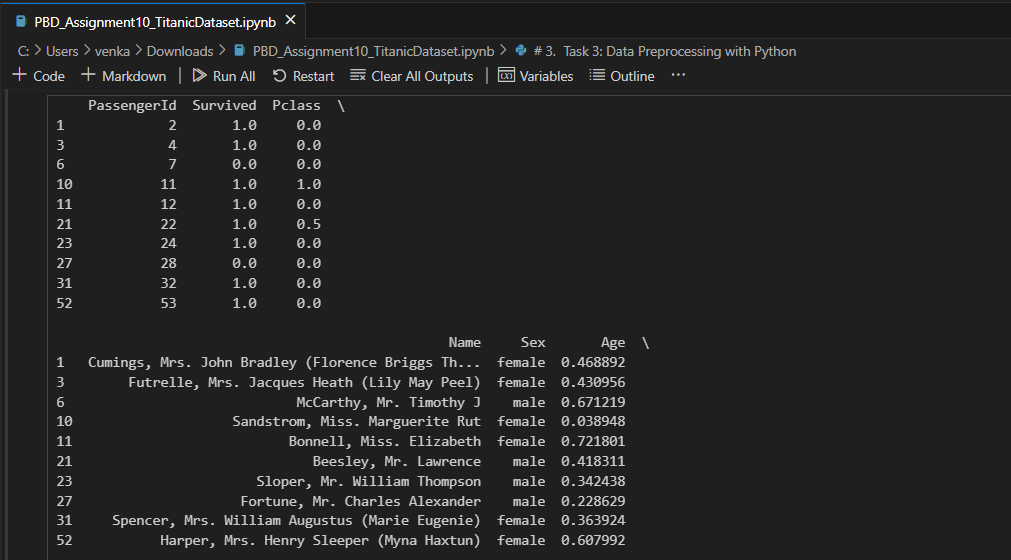




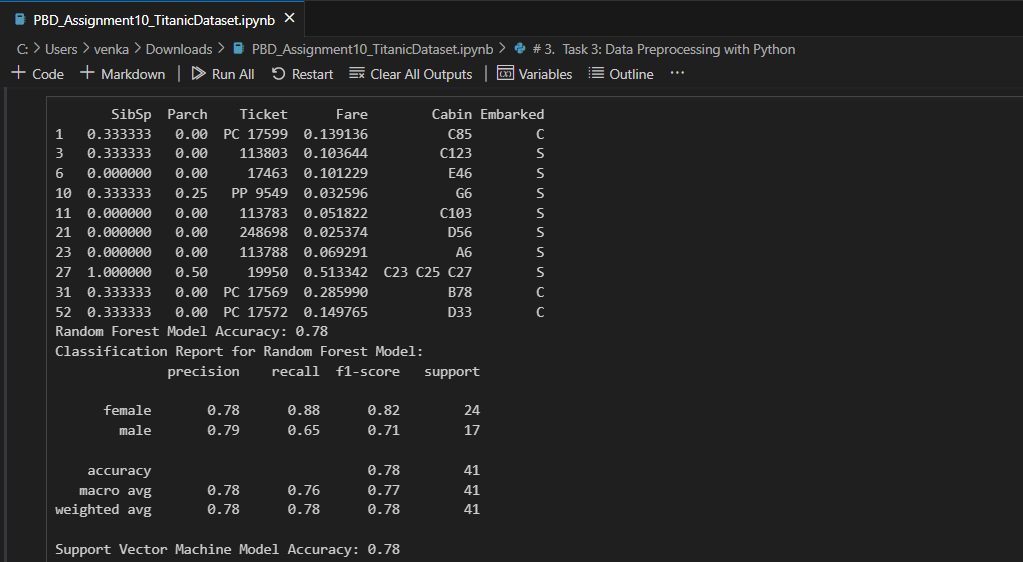


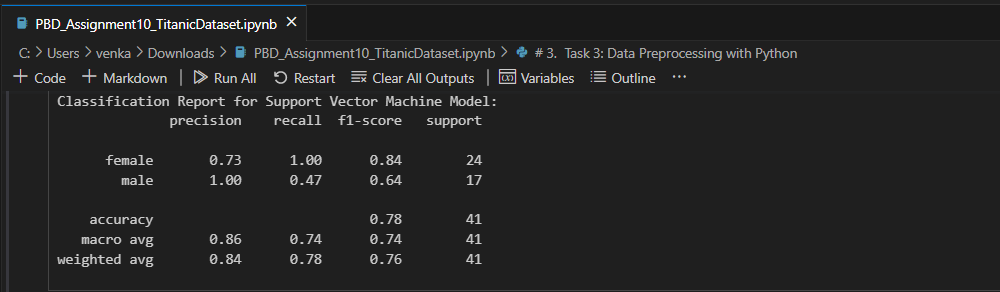
Displaying after removing outliers





1. Task 4: Implement Machine Learning Models with Python
   1. Implement at least two different machine learning models using Python.
   2. Evaluate and compare the performance of each model using appropriate metrics to determine the most suitable model for the dataset.





1. Task 5: Visualization with Python
   1. Create meaningful visualizations (e.g., scatter plots, heatmaps, bar charts) for the first dataset using Python.
   2. Use libraries like Matplotlib, Seaborn to create clear and insightful visual representations of the dataset.

