**Python Assignment 10**

Rajatkumar Bharatkumar Maisuriya(8870363)

Big-data Solution and Architecture, Conestoga

PROG8420 : Programing for Big Data

Jomis John

August 2, 2023

**Tasks:  
1.Task1:Dataset Selection**  
a. Choose two datasets from the provided repositories.

Answer: I have chosen the following two datasets from the provided repositories:

Dataset 1: Titanic Dataset

Dataset 2: Economic Impact of Tourism

**b. Justify your selection for each dataset based on its relevance to machine learning tasks. Include a brief paragraph explaining the dataset's potential for analysis and its suitability for machine learning applications.**

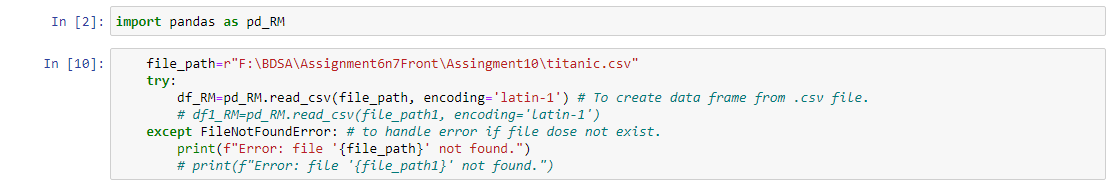
Answer: Dataset 1: Titanic Dataset

Justification: Titanic dataset has many merits for analysis, it's also a relatively small and simplified dataset. While it's a great starting point, real-world datasets can be more complex and nuanced. It contains information about the passengers aboard the RMS Titanic during its tragic maiden voyage, including whether they survived or not.

Answer: Dataset 2: Economic Impact of Tourism

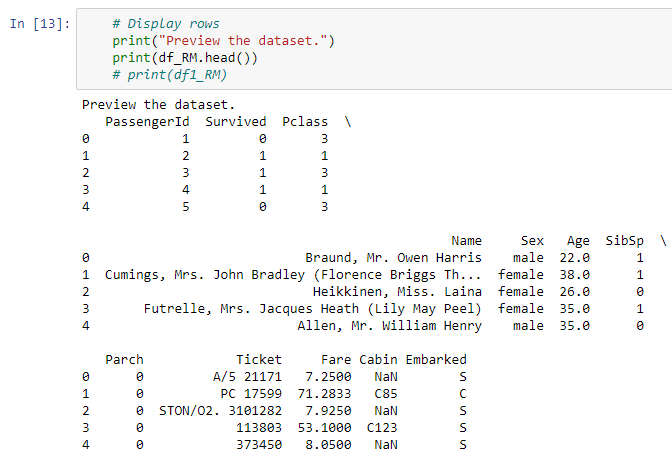
Justification: The economic impact of the tourism dataset is rich in potential for both data analysis and machine learning. It has financial data that are a good use for predicting and analyzing. Its structured format, relevance, and versatility make it a valuable resource for showcasing the various ways data can be explored, preprocessed, analyzed, and leveraged for predictive modeling.

Importing pandas and creating data frame by reading (.csv) file.

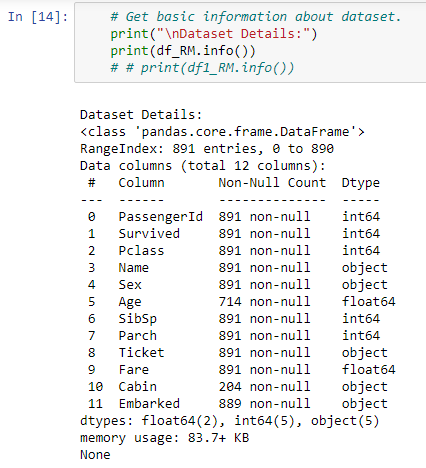


2.Task2: Data Exploration with Python  
a. Perform exploratory data analysis (EDA) using Python for the first dataset.  
b. Generate summary statistics, identify data types, and visualize the data distribution to gain insights into the dataset.

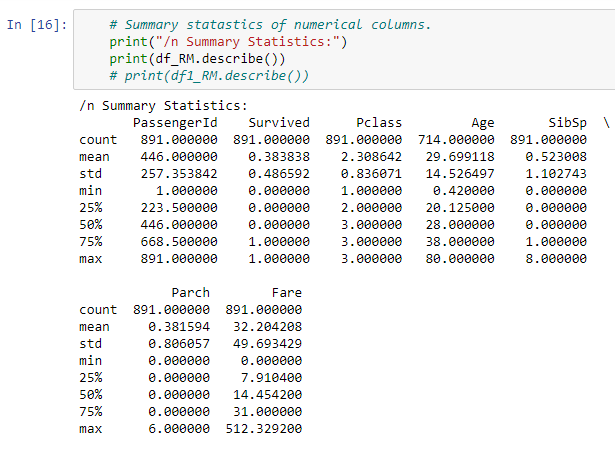
Display data of dataset.



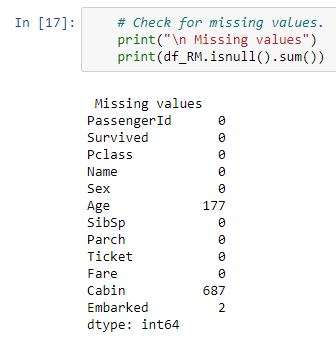
Get the information of the dataset.



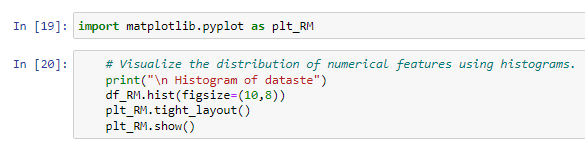
Summary the statistics of numerical coiumns.

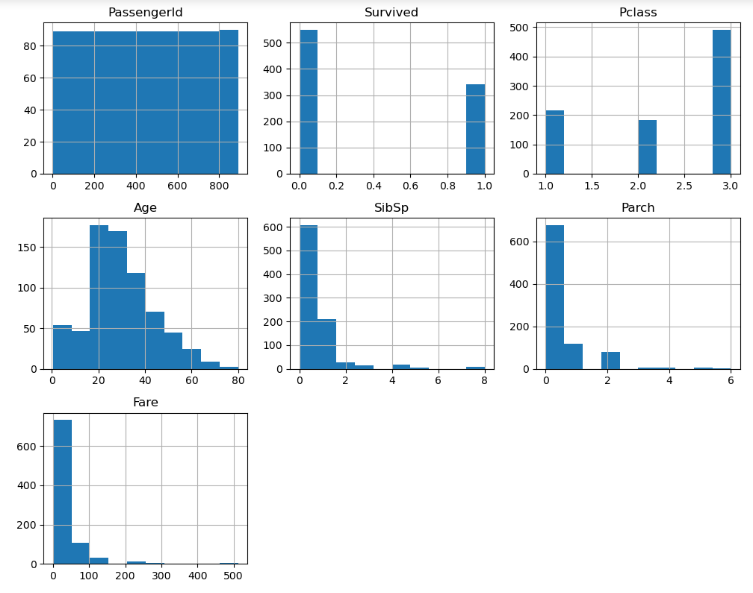


Checking for missing values.

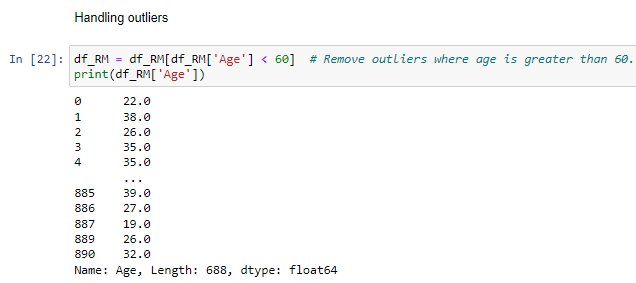


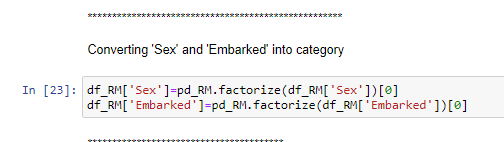
Visualisation of datasets

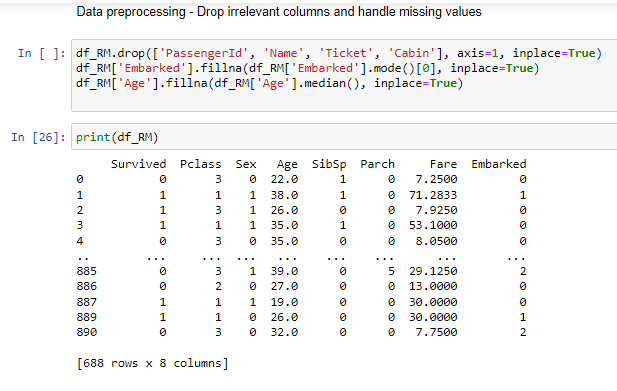




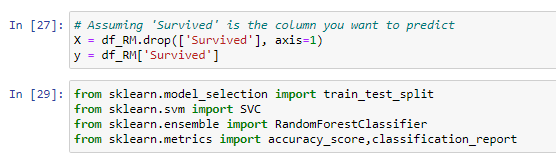
3. Task 3: Data Preprocessing with Python  
a. Preprocess the data from the first dataset using Python.  
b. Handle missing values, and outliers, and perform feature engineering when necessary to prepare the data for machine learning models.

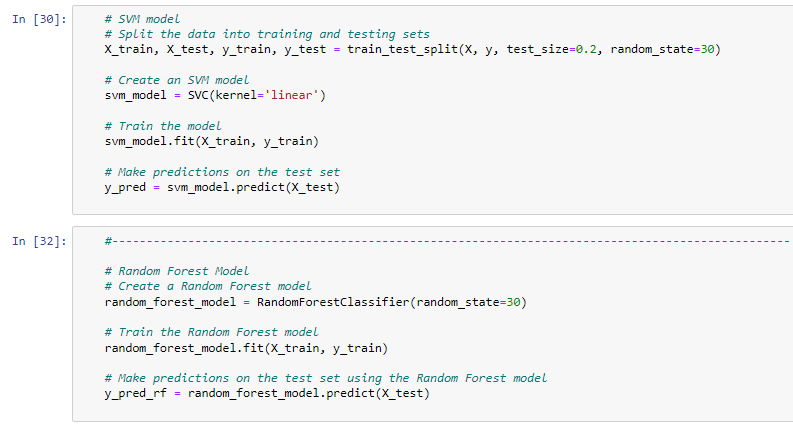




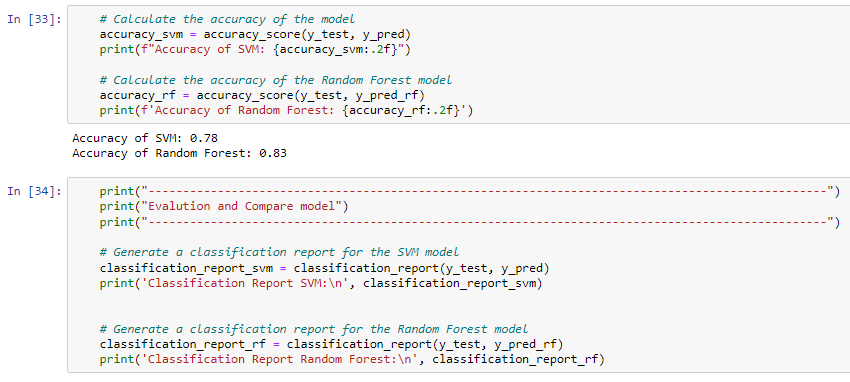


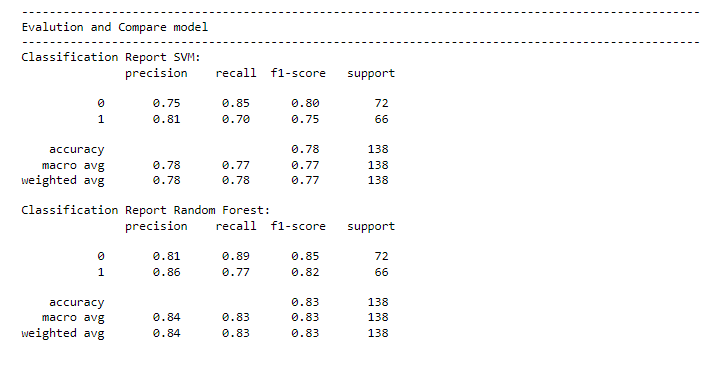
4. Task 4: Implement Machine Learning Models with Python  
a. Implement at least two different machine learning models (e.g., SVM, Random Forest, Neural Network) for the first dataset using Python.





b. Evaluate and compare the performance of each model using appropriate metrics to determine the most suitable model for the dataset.





5. Task 5: Visualization with Python  
a. Create meaningful visualizations (e.g., scatter plots, heatmaps, bar charts) for the first dataset using Python.

b. Use libraries like Matplotlib, Seaborn, or Plotly to create clear and insightful visual representations of the dataset.

