**Bansilal Ramnath Agarwal Charitable Trust’s**

**VISHWAKARMA INSTITUTE OF INFORMATION TECHNOLOGY,**

**PUNE-48**

**Department of Information Technology**

**ITUA40201: DATA SCIENCE AND ANALYTICS**

**Assignment-1**

**Shreyas Shripad Kulkarni**

**BTech A Division**

**Roll No.: 431048**

**PRN: 22010443**

**AIM:** Data Exploration and Visualization Exercise

**OBJECTIVE:** Perform exploratory data analysis and create visualizations to gain insights from a given dataset.

**TASKS:**

1. Load a dataset (e.g., customer sales data, stock market data).
2. Perform data cleaning and pre-processing.
3. Conduct descriptive statistical analysis.
4. Create visualizations (e.g., bar charts, scatter plots, box plots) to explore relationships and patterns in the data.
5. Interpret the findings and present the insights.

**THEORY:**

* **DATA PREPROCESSING**: Data preprocessing is a fundamental and critical step in the data preparation phase of any data analysis. It involves a series of operations and techniques to clean, transform, and organize raw data into a format that is suitable for analysis or modelling. Data preprocessing aims to improve the quality, consistency, and usability of data, making it ready for further exploration and application.
* Key Aspects of Data Preprocessing:
  + Data Cleaning
  + Data Transformation
  + Data Reduction
  + Data Integration
  + Feature Selection
  + Handling Imbalanced Data
* **DATA CLEANING**: Data cleaning is the process of fixing or removing incorrect, corrupted, incorrectly formatted, duplicate, or incomplete data within a dataset. Data cleaning is the process that removes data that does not belong in your dataset.

There are several other data cleaning techniques:

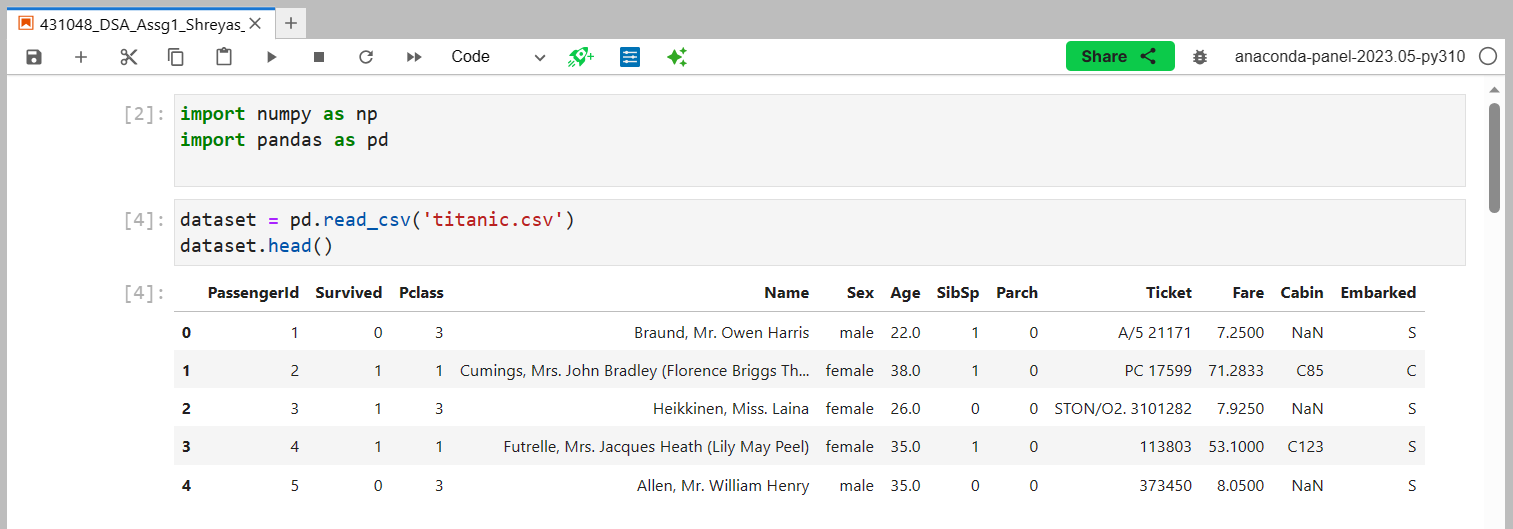
1. Handling Missing Values:
   * Drop rows or columns with missing values: You can use the `dropna()` function to remove rows or columns with missing values.
   * Fill missing values with a specific value: You can use the `fillna()` function to replace missing values with a specified value.
   * Fill missing values with mean, median, or mode: You can use the `fillna()` function with the mean, median, or mode of the column to replace missing values.
2. Removing Duplicates:
   * Use the `drop\_duplicates()` function to remove duplicate rows from the dataset.
3. Handling Outliers:
   * Identify outliers using statistical methods such as z-score or IQR (Interquartile Range).
   * Remove outliers by filtering the dataset based on the identified outliers.

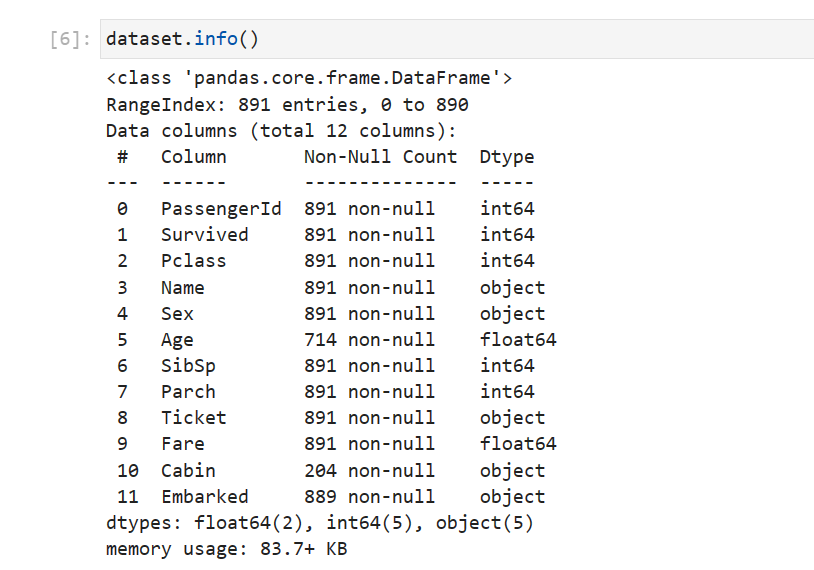
* **DESCRIPTIVE STATISTICAL ANALYSIS**: Descriptive statistical analysis is a branch of statistics that focuses on summarizing and presenting data in a meaningful and interpretable way. Descriptive statistics are fundamental in data analysis as they provide a clear and concise snapshot of the data, making it easier to understand and interpret.

**IMPLEMENTATION:**

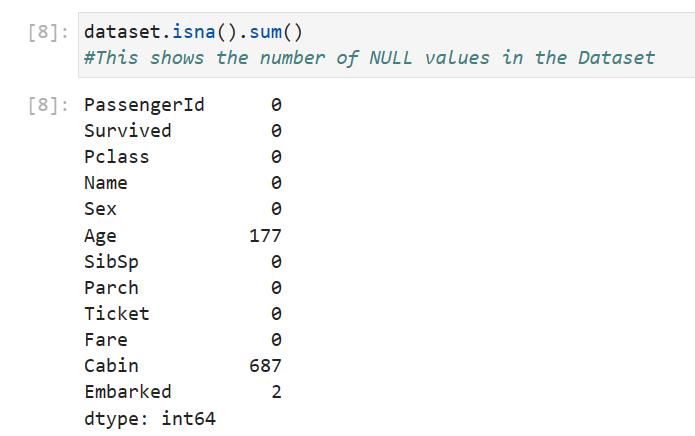
In this assignment we will be using the standard dataset (Titanic).

Step 1: Load Dataset.



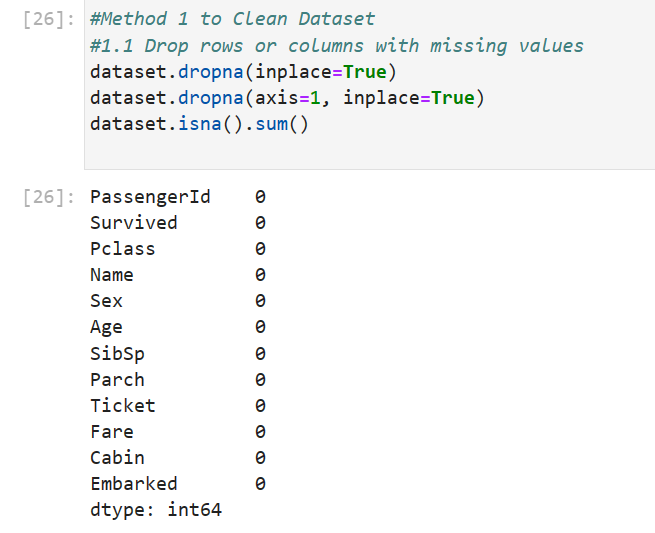


Step 2: Data Preprocessing & Data Cleaning

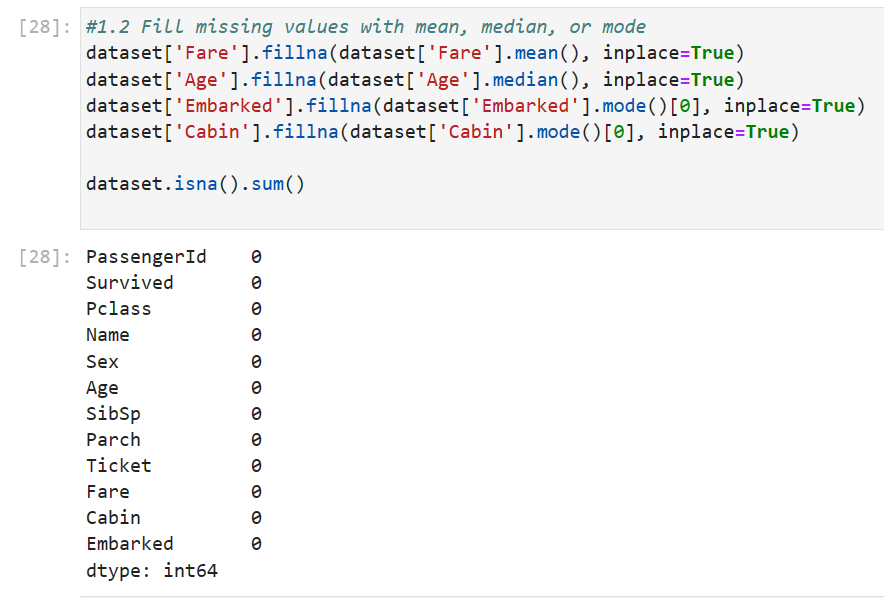


* + Drop rows or columns with missing values: You can use the `dropna()` function to remove rows or columns with missing

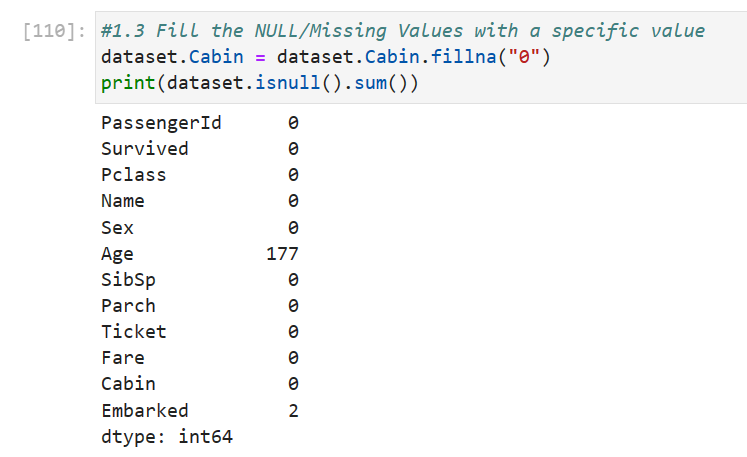
values.



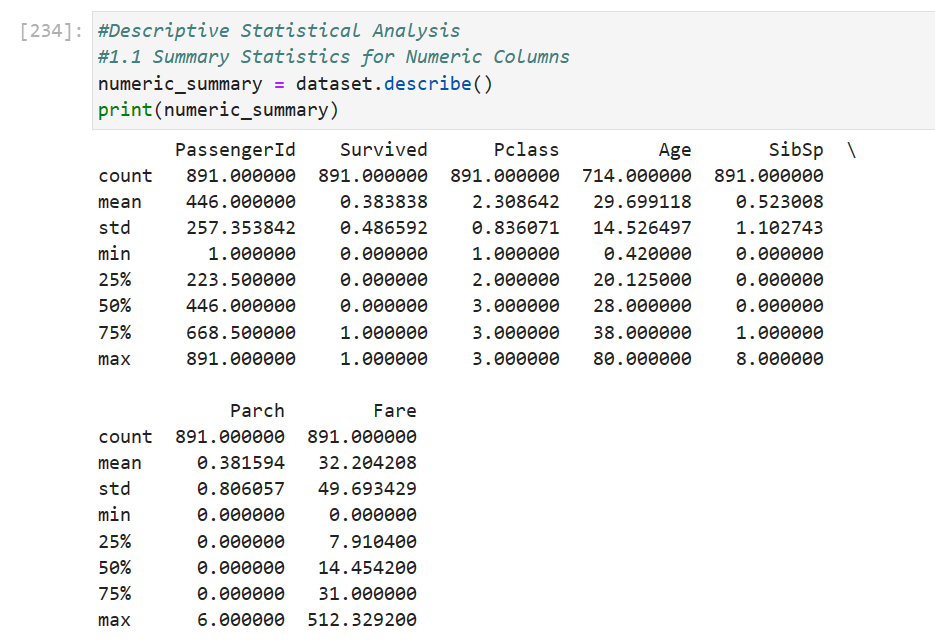
* + Fill missing values with a specific value: You can use the `fillna()` function to replace missing values with a specified value.

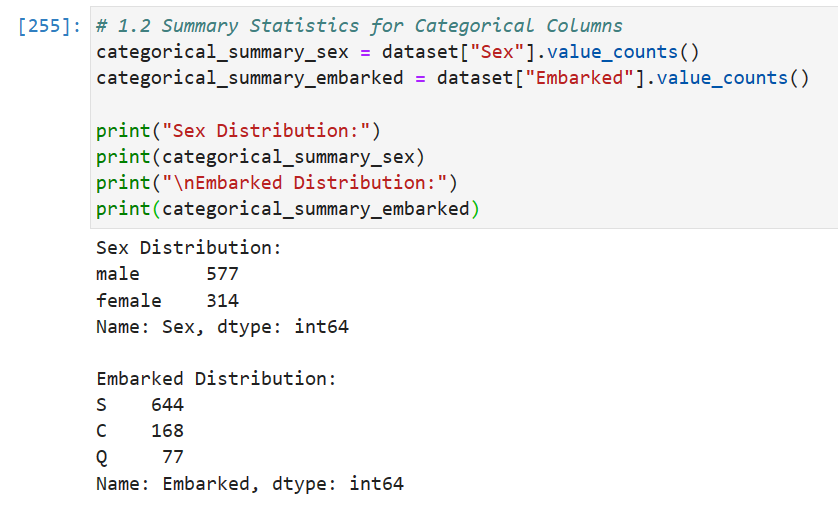


* + Fill missing values with mean, median, or mode: You can use the `fillna()` function with the mean, median, or mode of the column to replace missing values.



Step 3: Descriptive statistical analysis

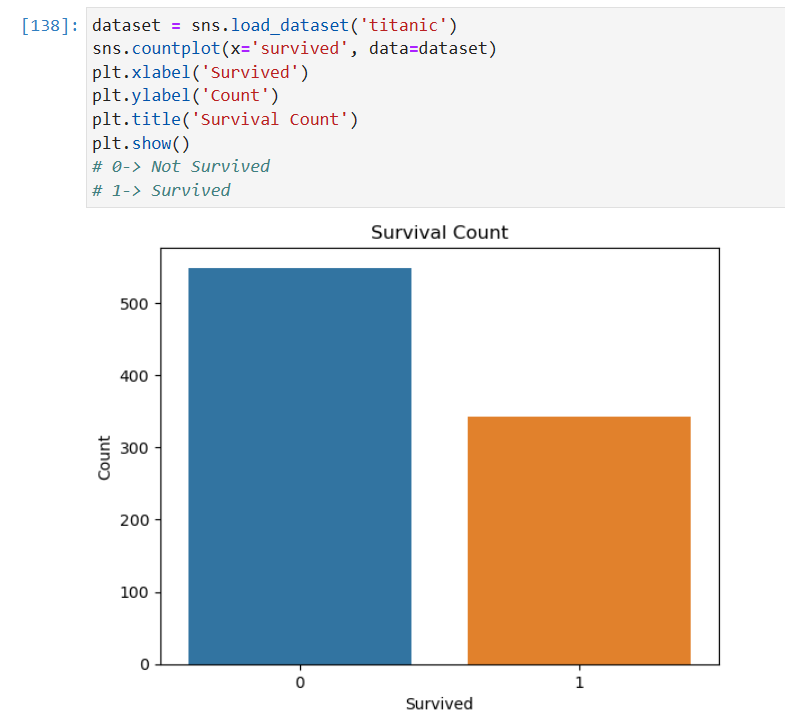




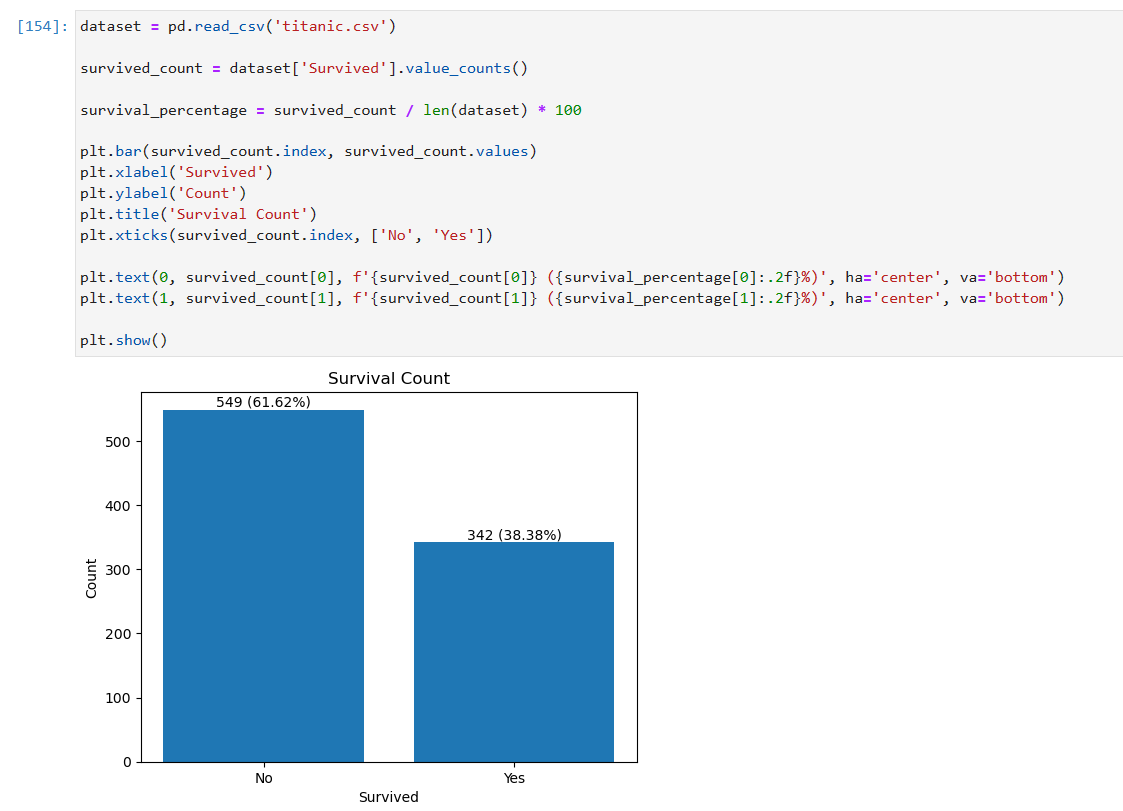


Step 4: Create visualizations

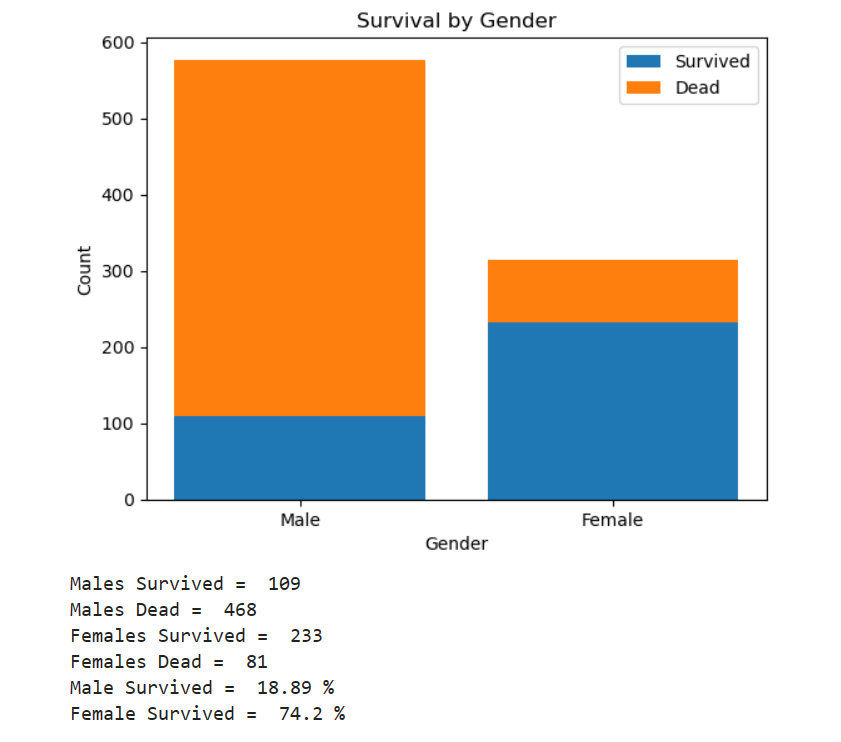




Step 5: Interpret Findings and present Insights







Findings & Insights from the Dataset:

* 109 Males & 233 Females Survived.
* 468 Males & 81 Females Dead
* 18.89 % Males & 74.2 % Females Survived
* 38.38 % People Survived

**CONCLUSION**: We have learnt, understood and performed exploratory data analysis and created visualizations to gain insights from titanic dataset.

[GitHub Repository](https://github.com/Shreyas100100/DATA-SCIENCE-AND-ANALYTICS)