Shri Vile Parle Kelavani Mandal's
DWARKADAS J. SANGHVI COLLEGE OF ENGINEERING
(Autonomous College Affiliated to the University of Mumbai)
NAAC Accredited with "A" Grade (CGPA : 3.18)

Department of Artificial Intelligence (AI) and Data Science
B.Tech. Sem: IV Subject: Data Mining and Analytics Laboratory
Experiment 3

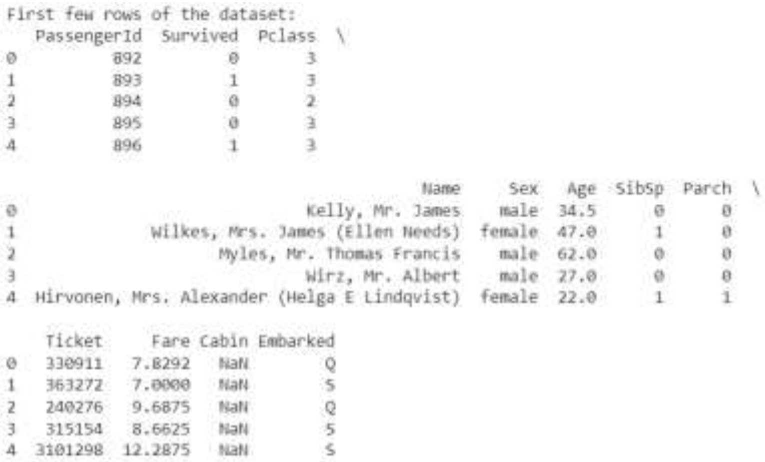
Name: Parth Karia **SAP ID: 60018230108**

Batch: A2 **Course Code: DJS22ADL403**

	Experiment Title: Data preparation using NumPy and Pandas **I. Collect data from a specific source (e.g., CSV file, API, database) and inspect its structure.**
Aim	To prepare Data using Numpy and Pandas in Python.
Software	Google Colab
Implementation	Code:

```
import numpy as np
import pandas as pd
from google.colab import files
uploaded = files.upload()
file_name = list(uploaded.keys())[0]
data = pd.read_csv(file_name)
```

Output:

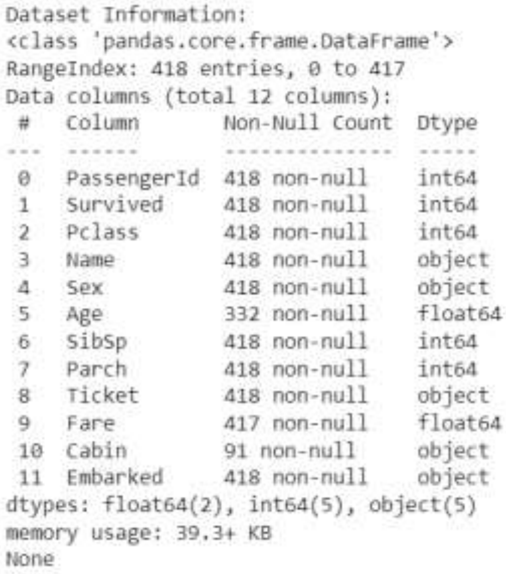


```
First few rows of the dataset:
   PassengerId  Survived  Pclass  \
0          892         0       3
1          893         1       3
2          894         0       2
3          895         0       3
4          896         1       3

                                           Name     Sex   Age  SibSp  Parch  \
0                              Kelly, Mr. James    male  34.5      0      0
1              Wilkes, Mrs. James (Ellen Needs)  female  47.0      1      0
2                     Myles, Mr. Thomas Francis    male  62.0      0      0
3                              Wirz, Mr. Albert    male  27.0      0      0
4  Hirvonen, Mrs. Alexander (Helga E Lindqvist)  female  22.0      1      1

    Ticket     Fare Cabin Embarked
0   330911   7.8292   NaN        Q
1   363272   7.0000   NaN        S
2   240276   9.6875   NaN        Q
3   315154   8.6625   NaN        S
4  3101298  12.2875   NaN        S
```

Code:

```
print("\nDataset Information:")
print(data.info())
```

Output:

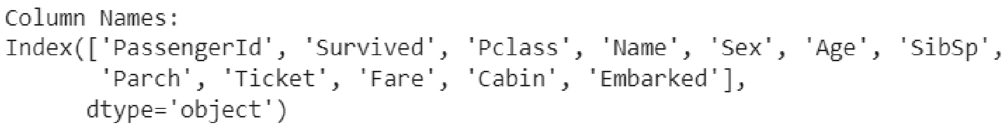


```
Dataset Information:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 418 entries, 0 to 417
Data columns (total 12 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   PassengerId  418 non-null    int64
 1   Survived     418 non-null    int64
 2   Pclass       418 non-null    int64
 3   Name         418 non-null    object
 4   Sex          418 non-null    object
 5   Age          332 non-null    float64
 6   SibSp        418 non-null    int64
 7   Parch        418 non-null    int64
 8   Ticket       418 non-null    object
 9   Fare         417 non-null    float64
 10  Cabin        91 non-null     object
 11  Embarked     418 non-null    object
dtypes: float64(2), int64(5), object(5)
memory usage: 39.3+ KB
None
```

Code:

```
# Display the column names
print("\nColumn Names:")
print(data.columns)
```

Output:

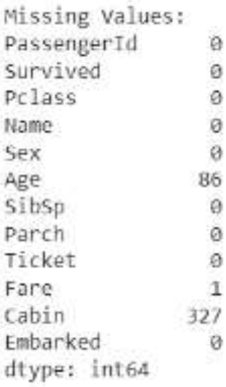


```
Column Names:
Index(['PassengerId', 'Survived', 'Pclass', 'Name', 'Sex', 'Age', 'SibSp',
       'Parch', 'Ticket', 'Fare', 'Cabin', 'Embarked'],
      dtype='object')
```

Code:

```
# Check for missing values in each column
print("\nMissing Values:")
print(data.isnull().sum())
```

Output:

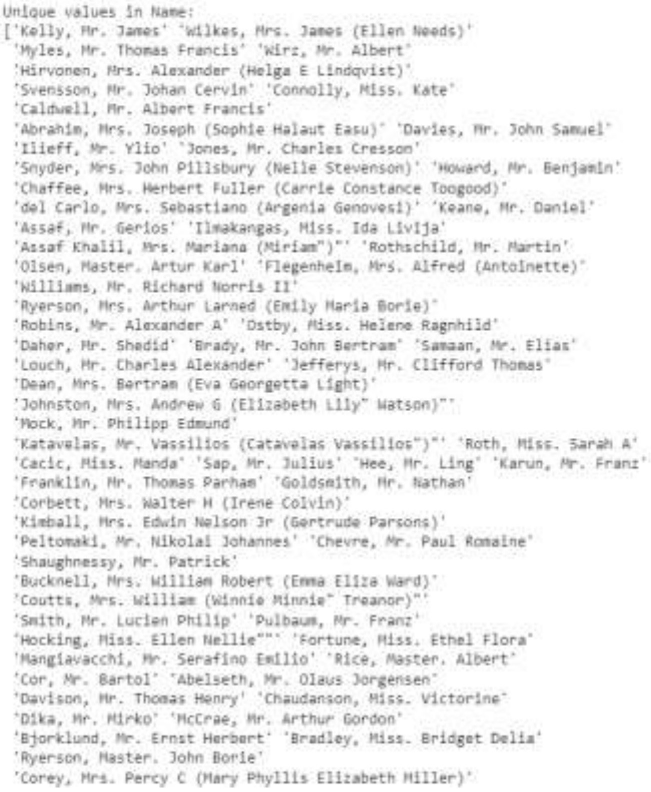


```
Missing Values:
PassengerId      0
Survived         0
Pclass           0
Name             0
Sex              0
Age             86
SibSp            0
Parch            0
Ticket           0
Fare             1
Cabin          327
Embarked         0
dtype: int64
```

Code:

```
# Display unique values in categorical columns
categorical_columns = data.select_dtypes(include=['object']).columns
for column in categorical_columns:
    print(f"\nUnique values in {column}:")
    print(data[column].unique())
```

Output:



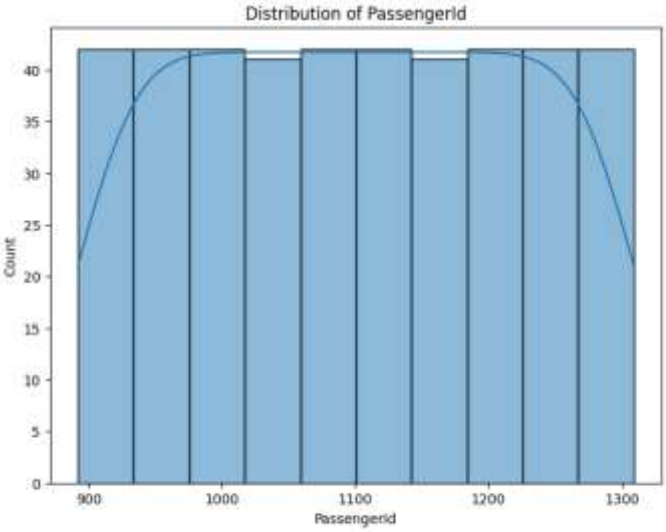
```
Unique values in Name:
['Kelly, Mr. James' 'Wilkes, Mrs. James (Ellen Needs)'
 'Myles, Mr. Thomas Francis' 'Wirz, Mr. Albert'
 'Hirvonen, Mrs. Alexander (Helga E Lindqvist)'
 'Svensson, Mr. Johan Cervin' 'Connolly, Miss. Kate'
 'Caldwell, Mr. Albert Francis'
 'Abrahim, Mrs. Joseph (Sophie Halaut Easu)' 'Davies, Mr. John Samuel'
 'Ilieff, Mr. Ylio' 'Jones, Mr. Charles Cresson'
 'Snyder, Mrs. John Pillsbury (Nelle Stevenson)' 'Howard, Mr. Benjamin'
 'Chaffee, Mrs. Herbert Fuller (Carrie Constance Toogood)'
 'del Carlo, Mrs. Sebastiano (Argenia Genovesi)' 'Keane, Mr. Daniel'
 'Assaf, Mr. Gerios' 'Ilmakangas, Miss. Ida Livija'
 'Assaf Khalil, Mrs. Mariana (Miriam")"' 'Rothschild, Mr. Martin'
 'Olsen, Master. Artur Karl' 'Flegenheim, Mrs. Alfred (Antoinette)'
 'Williams, Mr. Richard Norris II'
 'Ryerson, Mrs. Arthur Larned (Emily Maria Borie)'
 'Robins, Mr. Alexander A' 'Ostby, Miss. Helene Ragnhild'
 'Daher, Mr. Shedid' 'Brady, Mr. John Bertram' 'Samaan, Mr. Elias'
 'Louch, Mr. Charles Alexander' 'Jefferys, Mr. Clifford Thomas'
 'Dean, Mrs. Bertram (Eva Georgetta Light)'
 'Johnston, Mrs. Andrew G (Elizabeth Lily" Watson)"'
 'Mock, Mr. Philipp Edmund'
 'Katavelas, Mr. Vassilios (Catavelas Vassilios")"' 'Roth, Miss. Sarah A'
 'Cacic, Miss. Manda' 'Sap, Mr. Julius' 'Hee, Mr. Ling' 'Karun, Mr. Franz'
 'Franklin, Mr. Thomas Parham' 'Goldsmith, Mr. Nathan'
 'Corbett, Mrs. Walter H (Irene Colvin)'
 'Kimball, Mrs. Edwin Nelson Jr (Gertrude Parsons)'
 'Peltomaki, Mr. Nikolai Johannes' 'Chevre, Mr. Paul Romaine'
 'Shaughnessy, Mr. Patrick'
 'Bucknell, Mrs. William Robert (Emma Eliza Ward)'
 'Coutts, Mrs. William (Winnie Minnie" Treanor)"'
 'Smith, Mr. Lucien Philip' 'Pulbaum, Mr. Franz'
 'Hocking, Miss. Ellen Nellie""' 'Fortune, Miss. Ethel Flora'
 'Mangiavacchi, Mr. Serafino Emilio' 'Rice, Master. Albert'
 'Cor, Mr. Bartol' 'Abelseth, Mr. Olaus Jorgensen'
 'Davison, Mr. Thomas Henry' 'Chaudanson, Miss. Victorine'
 'Dika, Mr. Mirko' 'McCrae, Mr. Arthur Gordon'
 'Bjorklund, Mr. Ernst Herbert' 'Bradley, Miss. Bridget Delia'
 'Ryerson, Master. John Borie'
 'Corey, Mrs. Percy C (Mary Phyllis Elizabeth Miller)'
```

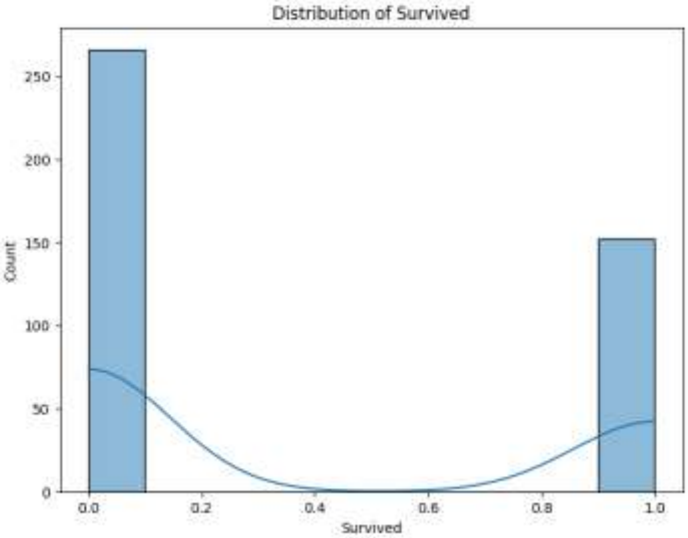
II. Generate summary statistics for a given dataset, including mean, median, standard deviation, and quartiles for numerical columns.

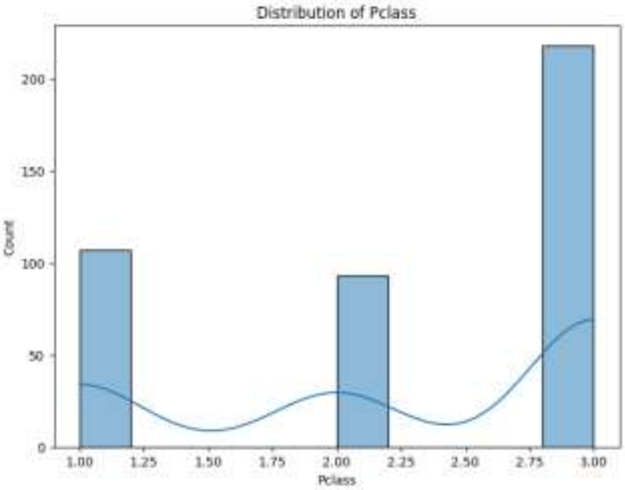
Code:

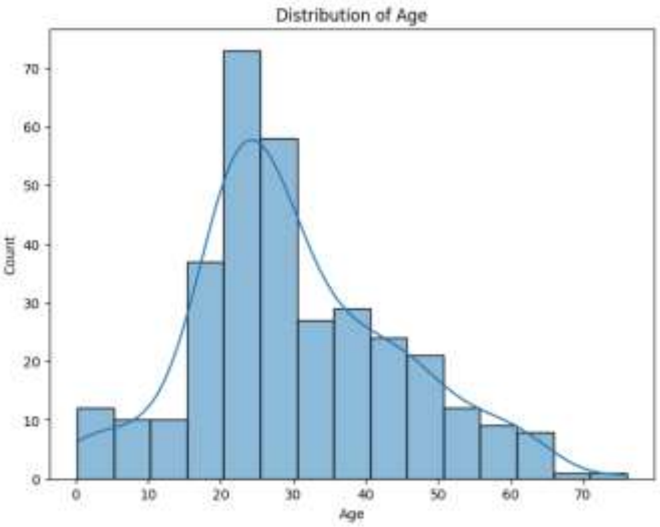
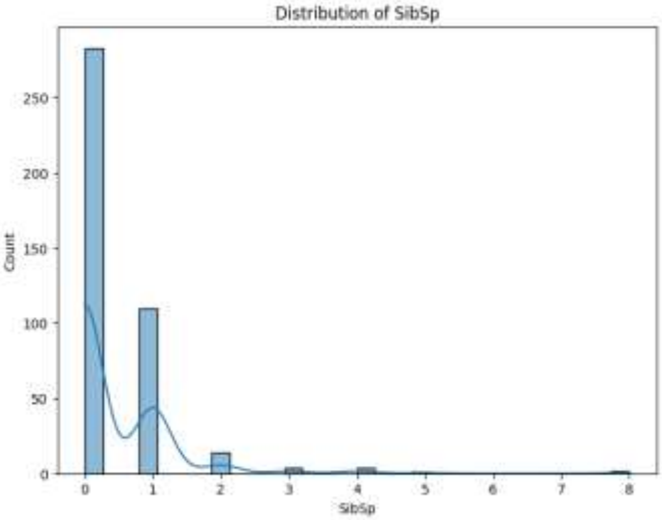
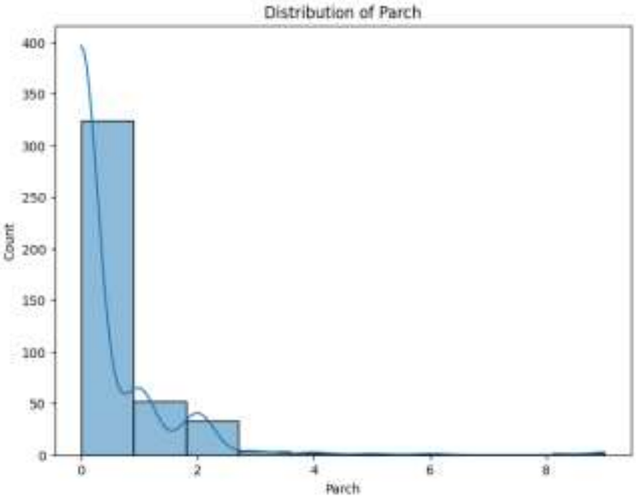
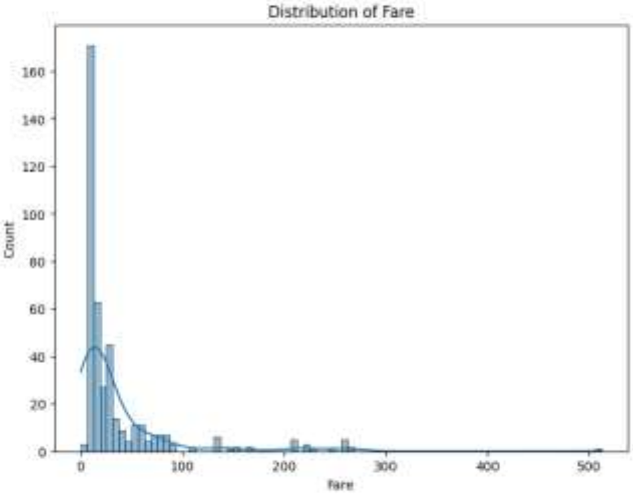
```
import matplotlib.pyplot as plt
import seaborn as sns
numerical_columns = data.select_dtypes(include=['float64',
'int64']).columns
for column in numerical_columns:
    plt.figure(figsize=(8, 6))
    sns.histplot(data[column], kde=True)
    plt.title(f'Distribution of {column}')
    plt.show()
```

Output:


Distribution of PassengerId


Distribution of Survived

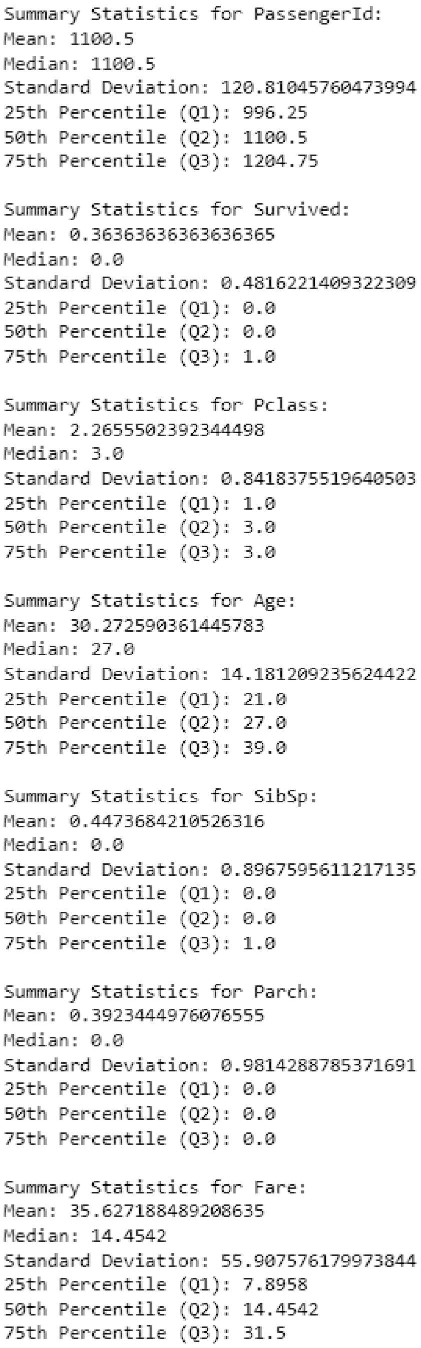

Distribution of Pclass

	Distribution of Age Count 0 10 20 30 40 50 60 70 Age Distribution of SibSp Count 0 50 100 150 200 250 1 2 3 4 5 6 7 8 SibSp Distribution of Parch Count 0 50 100 150 200 250 300 350 400 2 4 6 8 Parch Distribution of Fare Count 0 20 40 60 80 100 120 140 160 200 300 400 500 Fare

Code:

```
# Display mean, median, standard deviation, and quartiles for each
numerical column
for column in data.select_dtypes(include=['float64',
'int64']).columns:
    print(f"\nSummary Statistics for {column}:")
    print(f"Mean: {summary_statistics[column]['mean']}")
    print(f"Median: {data[column].median()}")
    print(f"Standard Deviation: {summary_statistics[column]['std']}")
    print(f"25th Percentile (Q1): {data[column].quantile(0.25)}")
    print(f"50th Percentile (Q2): {data[column].quantile(0.50)}")
    print(f"75th Percentile (Q3): {data[column].quantile(0.75)}")
```

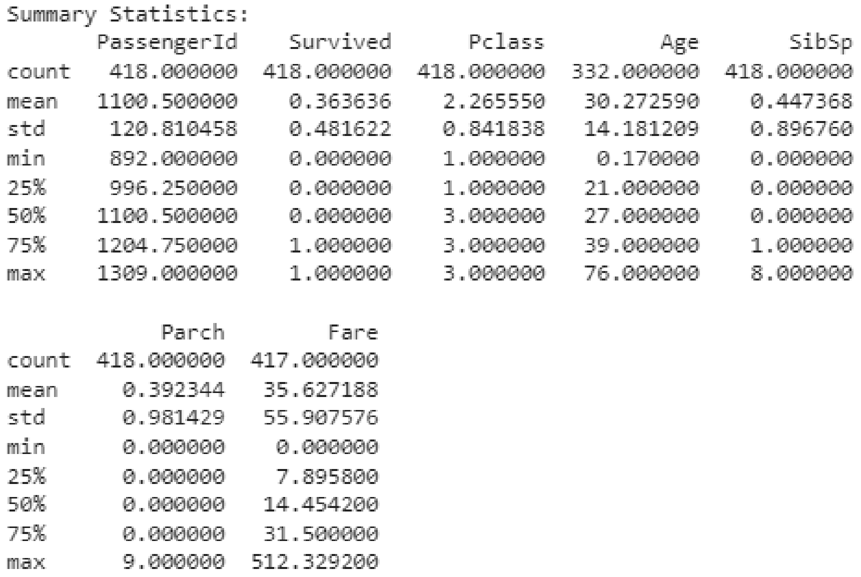
Output:



```
Summary Statistics for PassengerId:
Mean: 1100.5
Median: 1100.5
Standard Deviation: 120.81045760473994
25th Percentile (Q1): 996.25
50th Percentile (Q2): 1100.5
75th Percentile (Q3): 1204.75

Summary Statistics for Survived:
Mean: 0.36363636363636365
Median: 0.0
Standard Deviation: 0.4816221409322309
25th Percentile (Q1): 0.0
50th Percentile (Q2): 0.0
75th Percentile (Q3): 1.0

Summary Statistics for Pclass:
Mean: 2.2655502392344498
Median: 3.0
Standard Deviation: 0.8418375519640503
25th Percentile (Q1): 1.0
50th Percentile (Q2): 3.0
75th Percentile (Q3): 3.0

Summary Statistics for Age:
Mean: 30.272590361445783
Median: 27.0
Standard Deviation: 14.181209235624422
25th Percentile (Q1): 21.0
50th Percentile (Q2): 27.0
75th Percentile (Q3): 39.0

Summary Statistics for SibSp:
Mean: 0.4473684210526316
Median: 0.0
Standard Deviation: 0.8967595611217135
25th Percentile (Q1): 0.0
50th Percentile (Q2): 0.0
75th Percentile (Q3): 1.0

Summary Statistics for Parch:
Mean: 0.3923444976076555
Median: 0.0
Standard Deviation: 0.9814288785371691
25th Percentile (Q1): 0.0
50th Percentile (Q2): 0.0
75th Percentile (Q3): 0.0

Summary Statistics for Fare:
Mean: 35.627188489208635
Median: 14.4542
Standard Deviation: 55.907576179973844
25th Percentile (Q1): 7.8958
50th Percentile (Q2): 14.4542
75th Percentile (Q3): 31.5
```

Code:

```python
# Display summary statistics of the dataset
print("\nSummary Statistics:")
print(data.describe())
```



```
Summary Statistics:
       PassengerId    Survived      Pclass         Age       SibSp
count   418.000000  418.000000  418.000000  332.000000  418.000000
mean   1100.500000    0.363636    2.265550   30.272590    0.447368
std     120.810458    0.481622    0.841838   14.181209    0.896760
min     892.000000    0.000000    1.000000    0.170000    0.000000
25%     996.250000    0.000000    1.000000   21.000000    0.000000
50%    1100.500000    0.000000    3.000000   27.000000    0.000000
75%    1204.750000    1.000000    3.000000   39.000000    1.000000
max    1309.000000    1.000000    3.000000   76.000000    8.000000

            Parch        Fare
count  418.000000  417.000000
mean     0.392344   35.627188
std      0.981429   55.907576
min      0.000000    0.000000
25%      0.000000    7.895800
50%      0.000000   14.454200
75%      0.000000   31.500000
max      9.000000  512.329200
```

Conclusion	Hence, we have learned how to prepare data using Numpy and Pandas in Python