MIT | Academy of Engineering

EDS Project on:

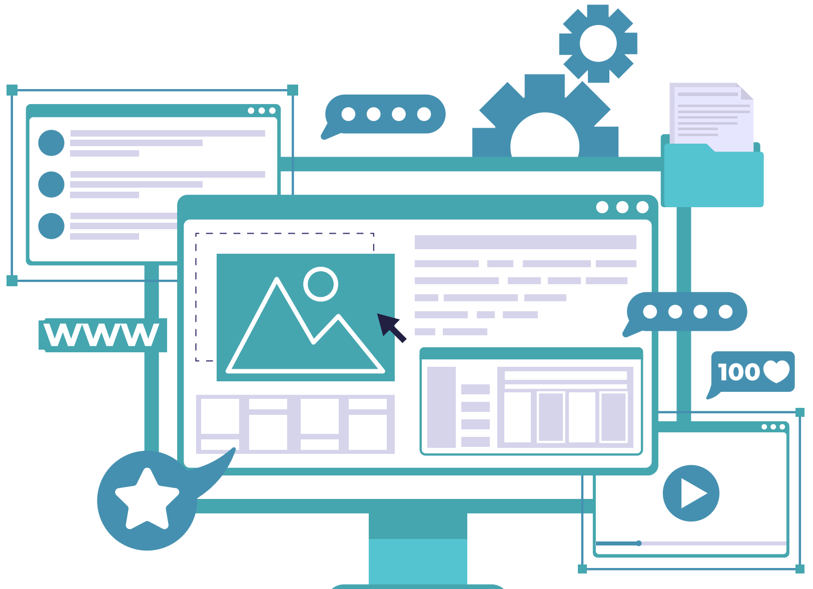
Exploring the Titanic data

Guided By: S.P Kale (MITAOE)

Presented By:

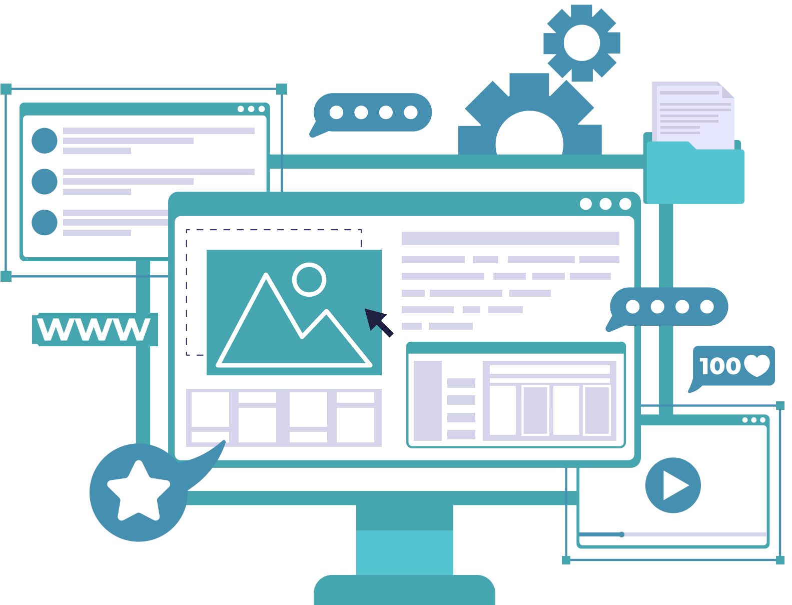
230 : Aditya Babar (202201070130)
231 : Janhavi Kawadkar (202201060008)
236 : Rohan Agrawal (202201040208)

FYBTech,MITAOE



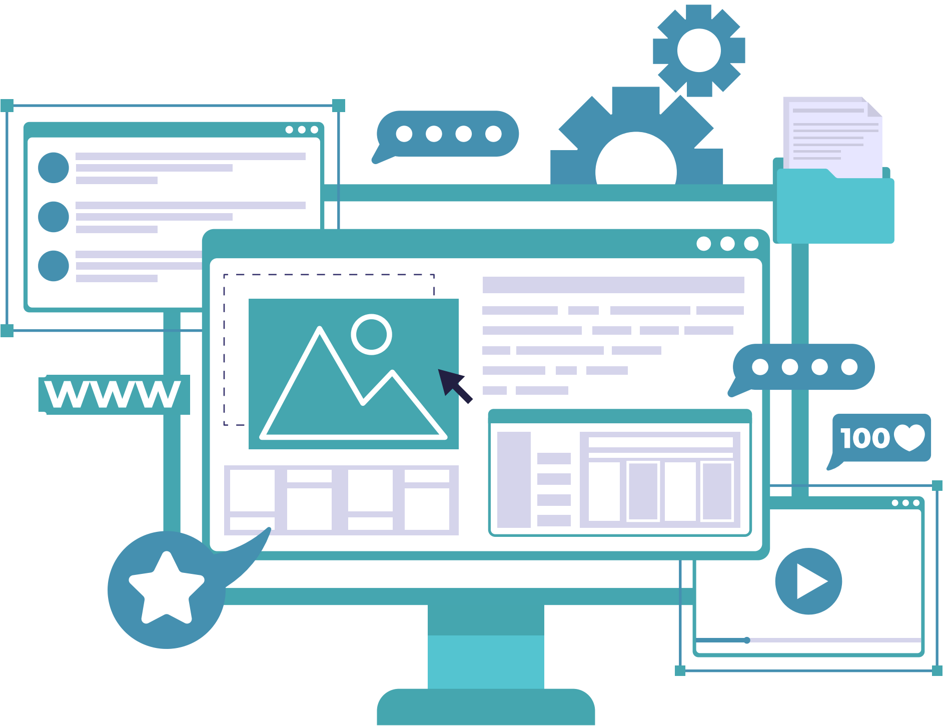
Introduction

- Data analytics is the process of examining vast volumes of data to extract meaningful patterns, trends, and correlations.
- In the context of the Titanic dataset, data analysis becomes a window through which we can do various analysis such as data manipulation, data visualization ,etc.
- By applying robust data analysis techniques, we aim to uncover the factors that influenced survival rates, understand the demographics of the passengers, and reveal intriguing correlations within the dataset.



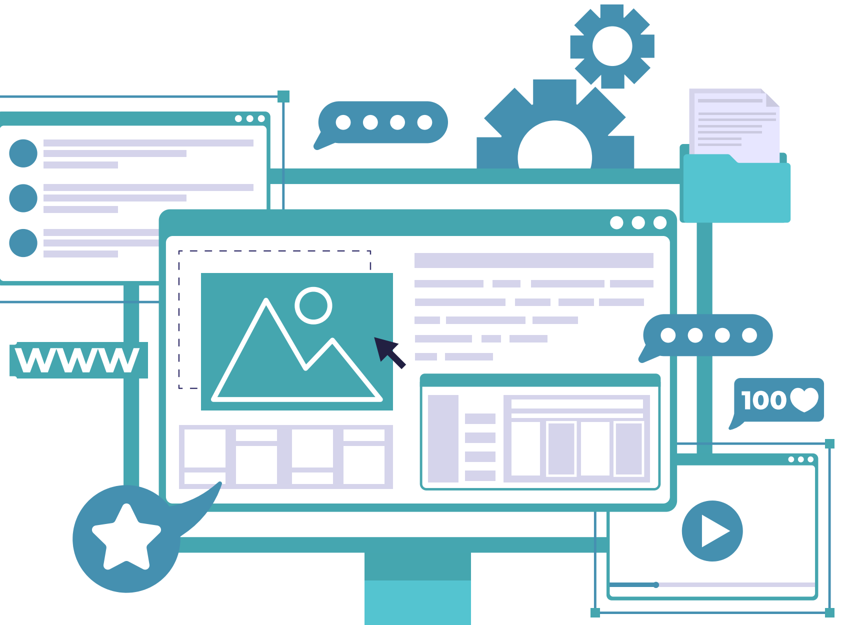
Motivation

The sinking of the Titanic is a well-known historical event, and the dataset provides information about the passengers on board, including their demographics, ticket class, cabin, fare, survival status, etc. Analyzing this dataset allows us to gain insights into various factors that might have influenced the survival of passengers during the tragedy.



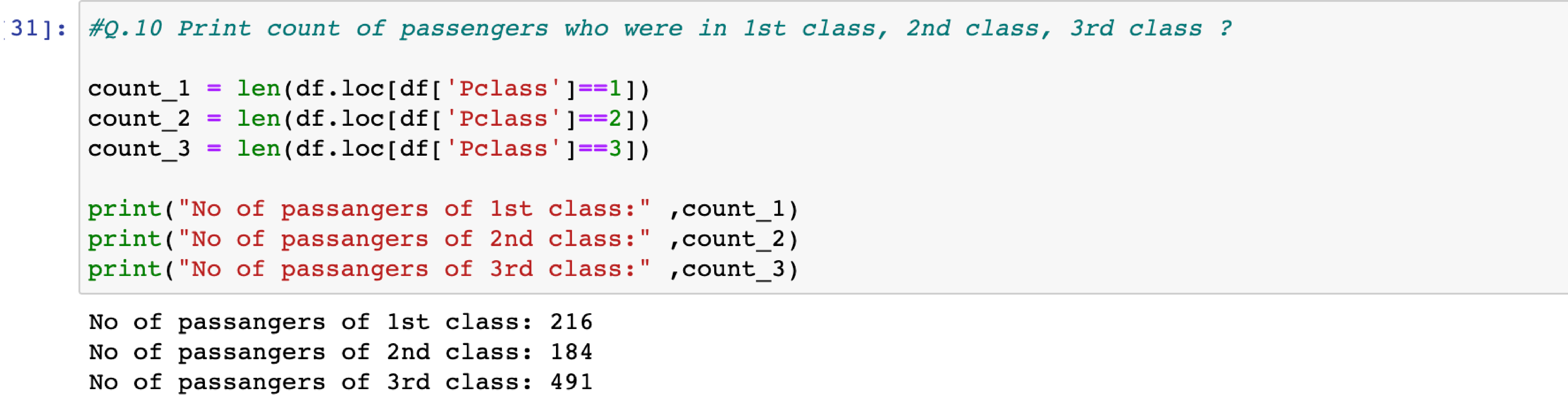
Details of Dataset

- Name: Titanic dataset
- Number of features: 14
- Number of records: 891



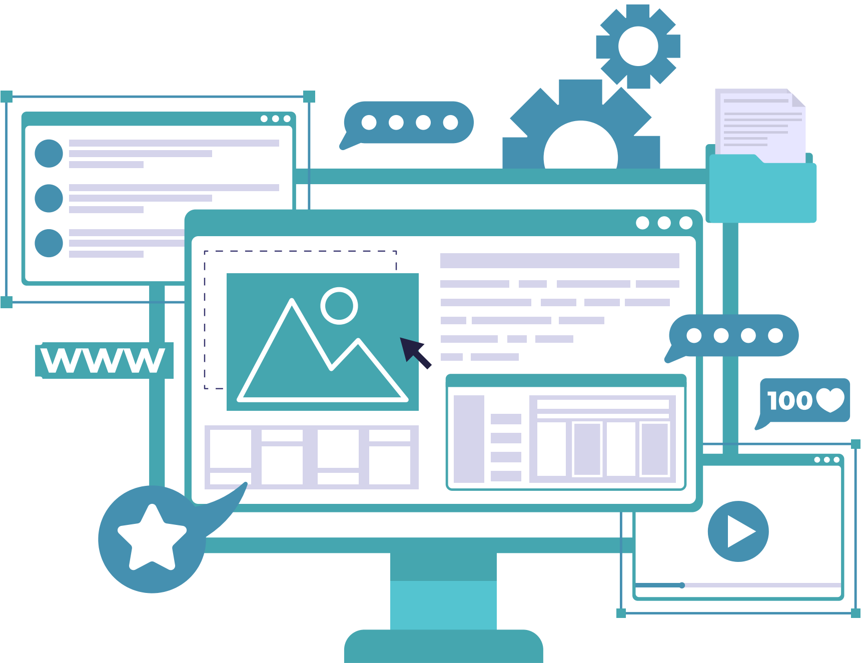
Data Manipulation

Data manipulation is a fundamental process in data analysis that involves transforming and preparing raw data to make it suitable for further exploration and analysis. It encompasses a range of operations aimed at ensuring data quality, consistency, and usability. Missing values can be imputed or removed, while outliers can be addressed through various methods such as transformation.



```
31]: #Q.10 Print count of passengers who were in 1st class, 2nd class, 3rd class ?

     count_1 = len(df.loc[df['Pclass']==1])
     count_2 = len(df.loc[df['Pclass']==2])
     count_3 = len(df.loc[df['Pclass']==3])

     print("No of passangers of 1st class:" ,count_1)
     print("No of passangers of 2nd class:" ,count_2)
     print("No of passangers of 3rd class:" ,count_3)
```

```
No of passangers of 1st class: 216
No of passangers of 2nd class: 184
No of passangers of 3rd class: 491
```



In [9]:
```
#Q.8 Print count of passengers who survived ?

count = len(df.loc[df['Survived']==1])
print("No of passangers who survived:" ,count)
```

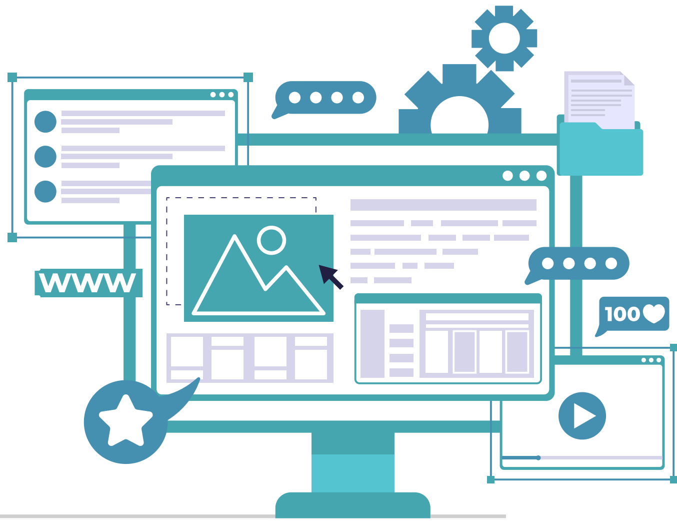
```
No of passangers who survived: 342
```

In [10]:
```
#Q.9 Print count of passengers who didnt survived ?

count = len(df.loc[df['Survived']==0])
print("No of passangers who didnt survived:" ,count)
```

```
No of passangers who didnt survived: 549
```

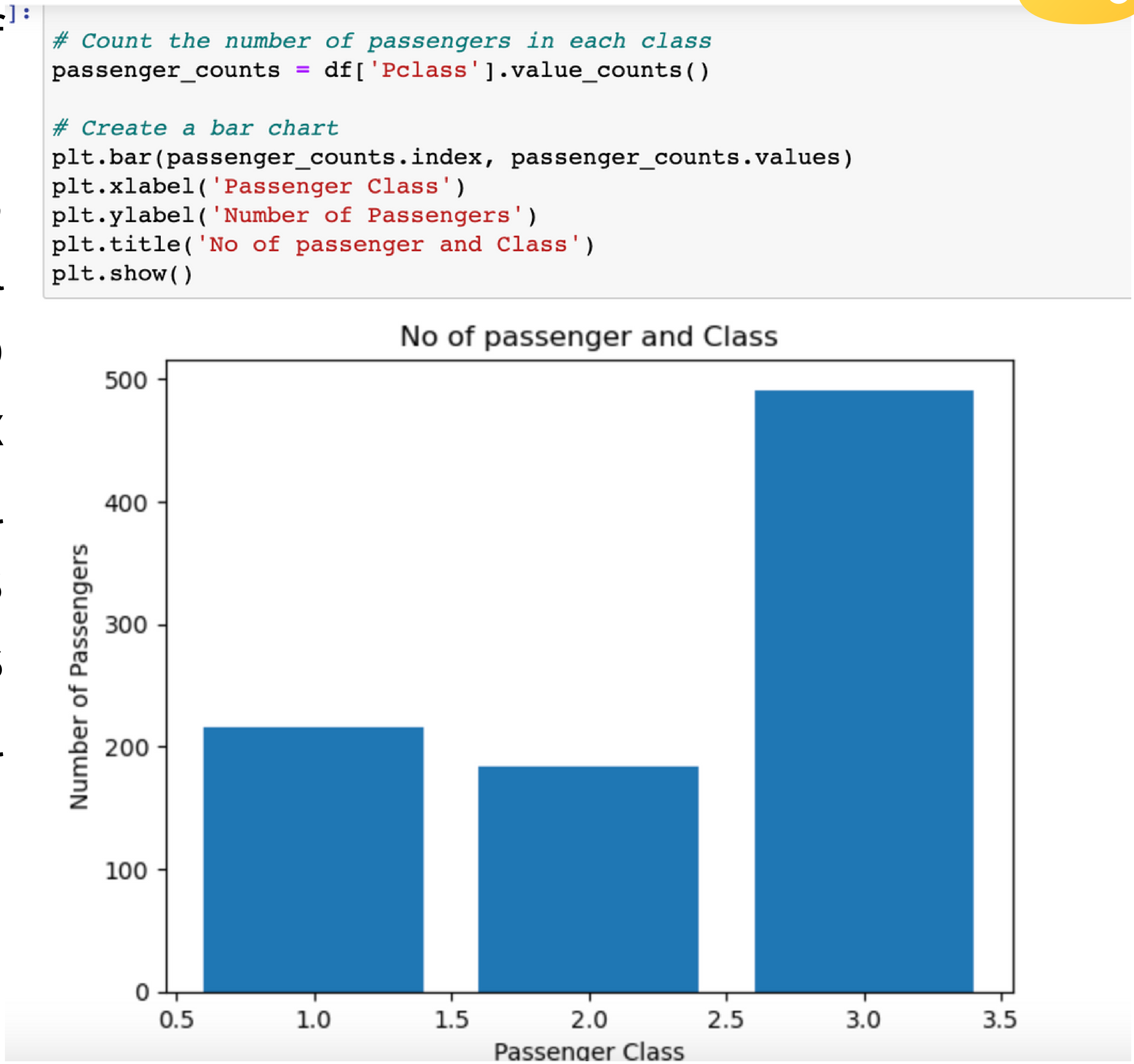
In [16]:
```
#Q.13 Use groupby function to count the number of passengers embarked from C,Q,S spot ?

a=df.groupby("Embarked").count()
print(a)
```

```
          Age  Cabin  Fare  Name  Parch  PassengerId  Pclass  Sex  SibSp  \
Embarked
C         169     70   169   169    169          169     169  169    169
Q          77      4    77    77     77           77      77   77     77
S         645    130   645   645    645          645     645  645    645

          Survived  Ticket  Title  Family_Size
Embarked
C              169     169    169          169
Q               77      77     77           77
S              645     645    645          645
```

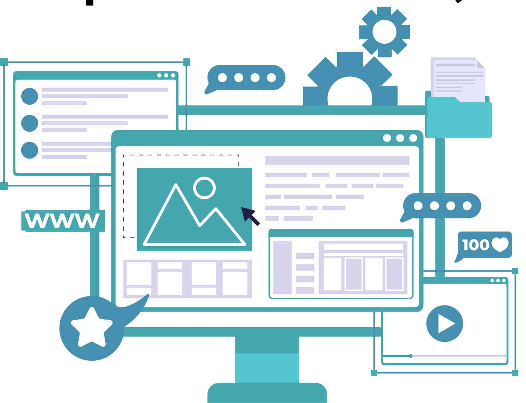


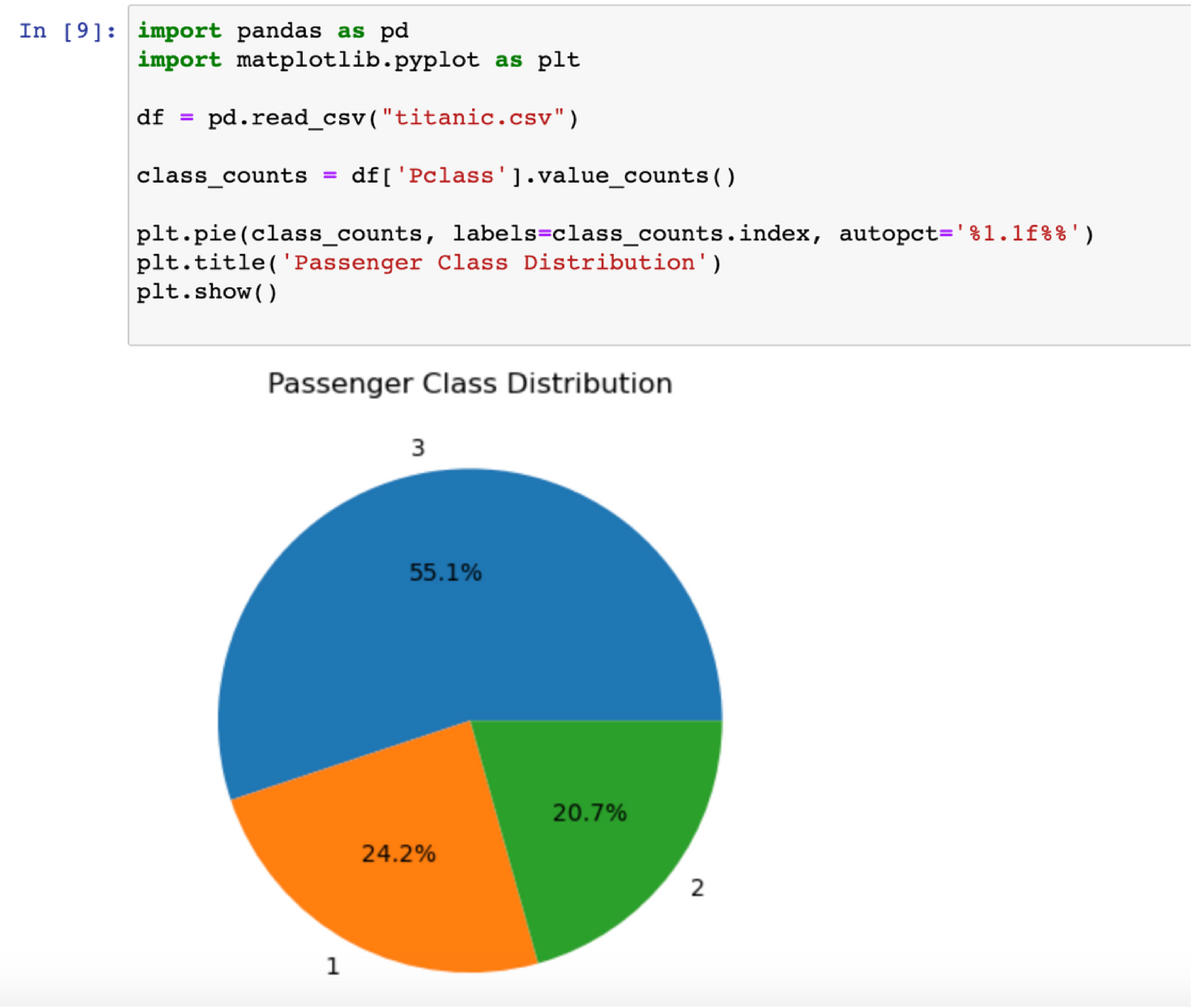
Data Visualization

Data visualization is the process of representing data and information visually through charts, graphs, maps, and other graphical elements. It is a powerful technique that allows us to effectively communicate complex concepts, patterns, and trends in a visual format. Data visualization transforms complex data into visual representations that enhance understanding, reveal patterns, and support decision-making.

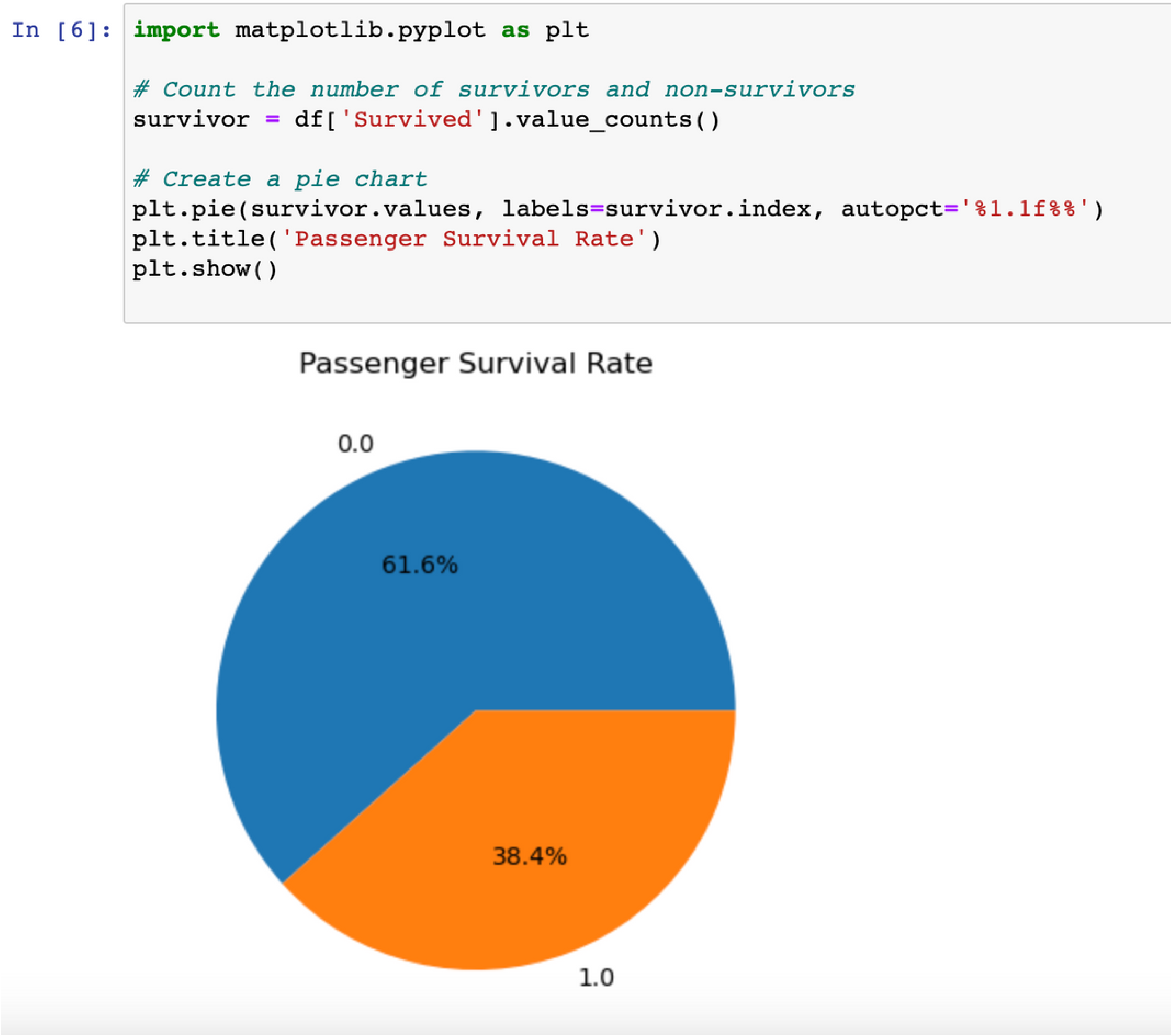


```python
# Count the number of passengers in each class
passenger_counts = df['Pclass'].value_counts()

# Create a bar chart
plt.bar(passenger_counts.index, passenger_counts.values)
plt.xlabel('Passenger Class')
plt.ylabel('Number of Passengers')
plt.title('No of passenger and Class')
plt.show()
```

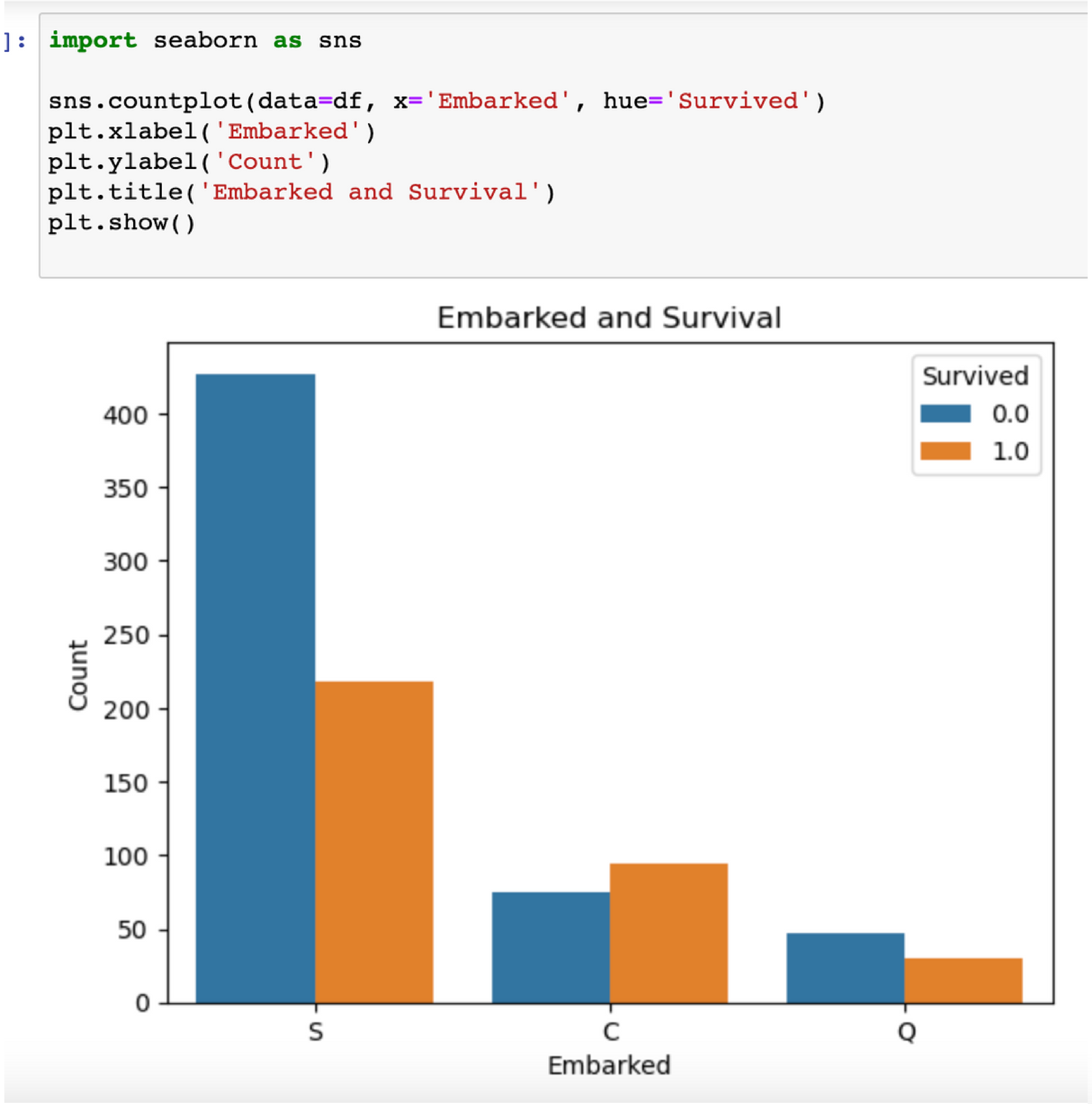




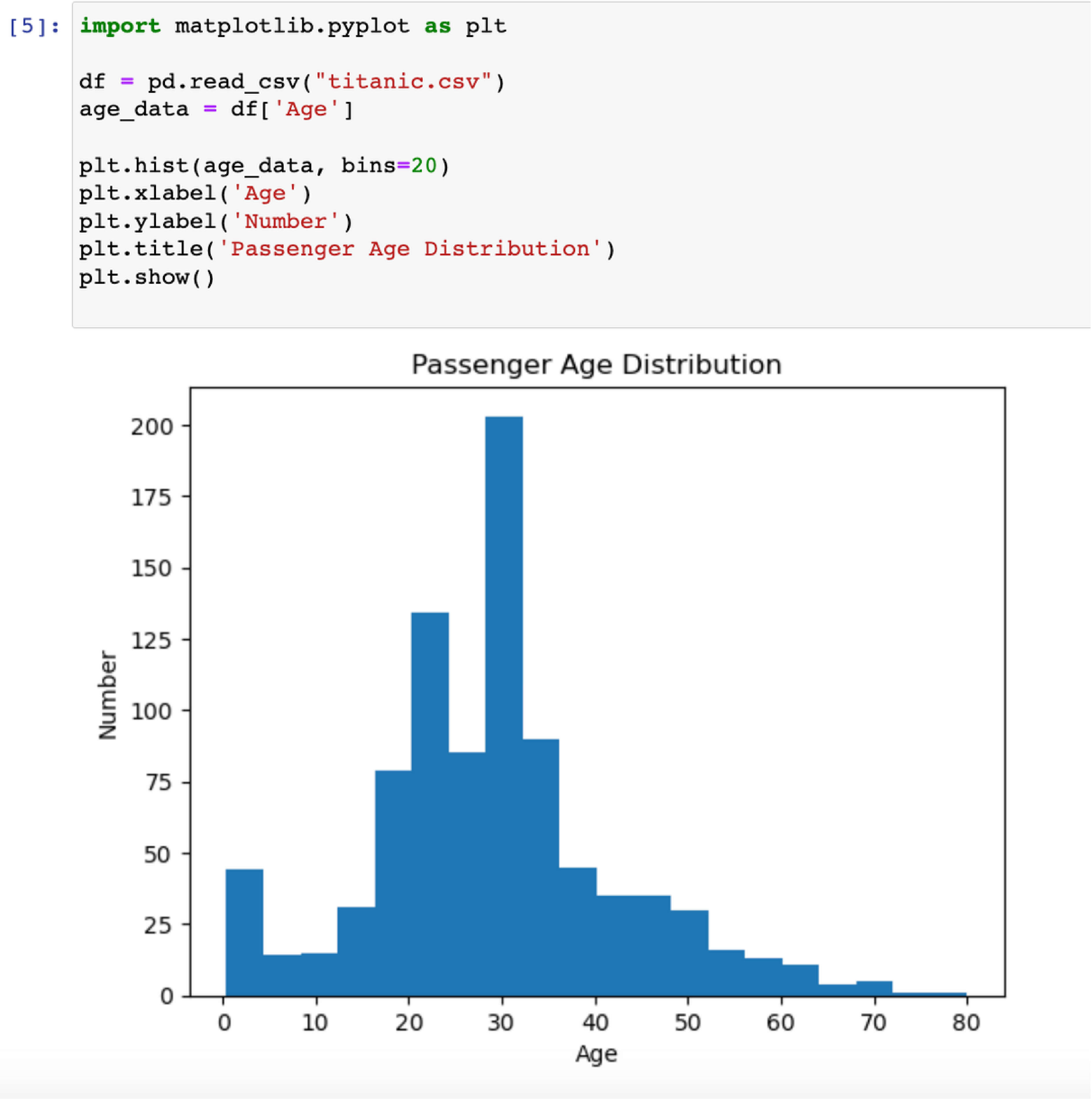
In [9]:
```python
import pandas as pd
import matplotlib.pyplot as plt

df = pd.read_csv("titanic.csv")

class_counts = df['Pclass'].value_counts()

plt.pie(class_counts, labels=class_counts.index, autopct='%1.1f%%')
plt.title('Passenger Class Distribution')
plt.show()
```



In [6]:
```python
import matplotlib.pyplot as plt

# Count the number of survivors and non-survivors
survivor = df['Survived'].value_counts()

# Create a pie chart
plt.pie(survivor.values, labels=survivor.index, autopct='%1.1f%%')
plt.title('Passenger Survival Rate')
plt.show()
```

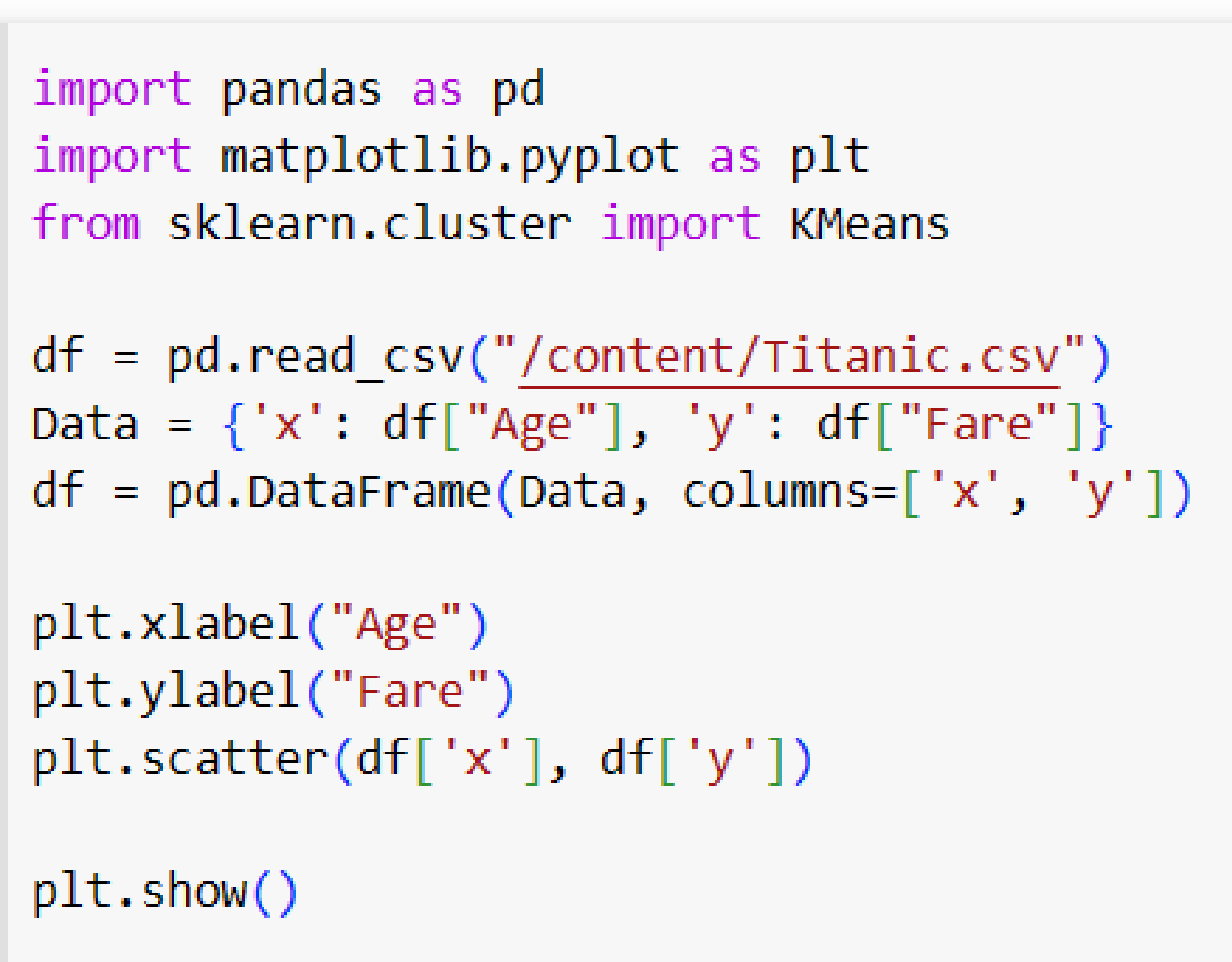


```
import seaborn as sns

sns.countplot(data=df, x='Embarked', hue='Survived')
plt.xlabel('Embarked')
plt.ylabel('Count')
plt.title('Embarked and Survival')
plt.show()
```

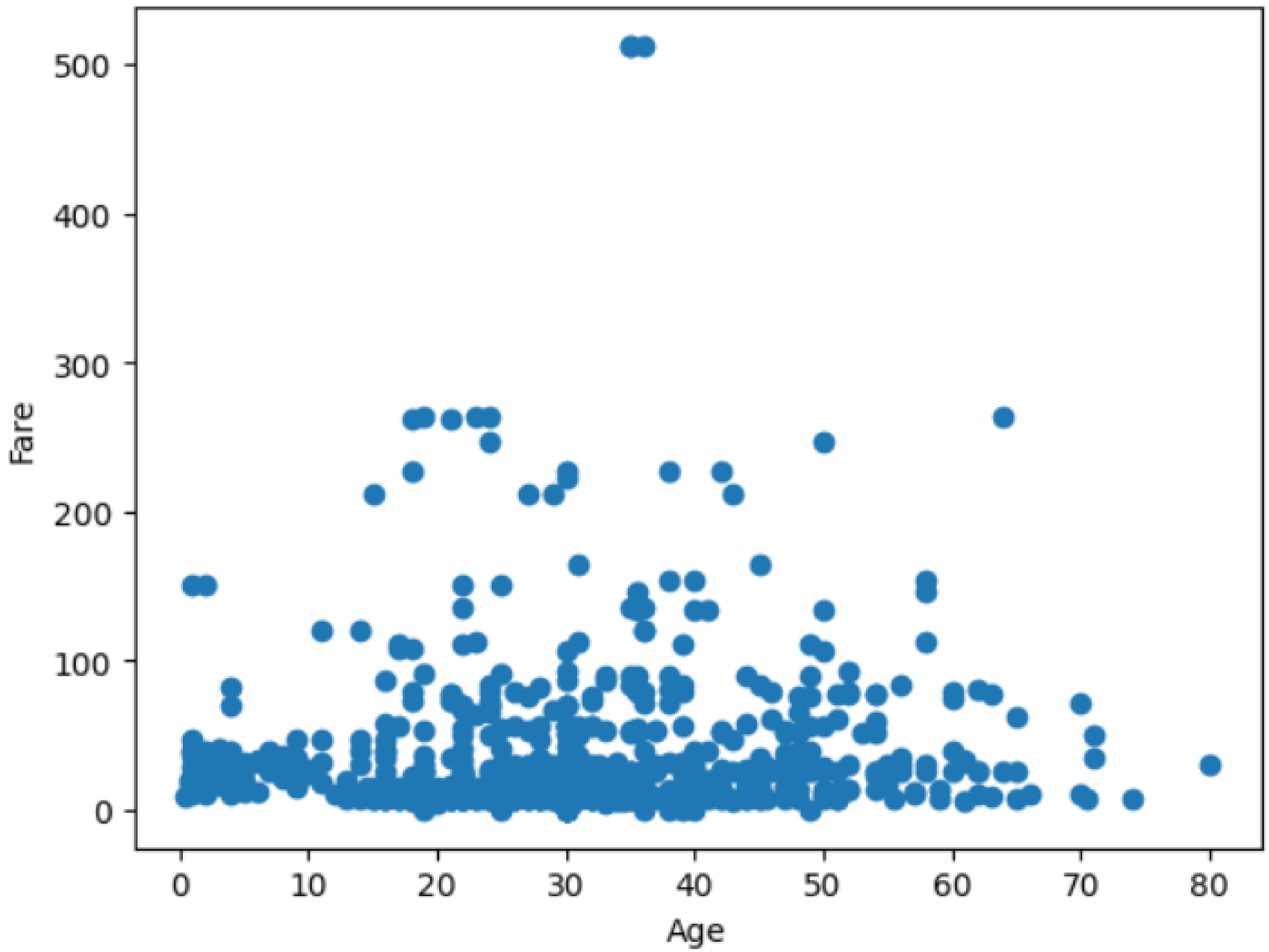


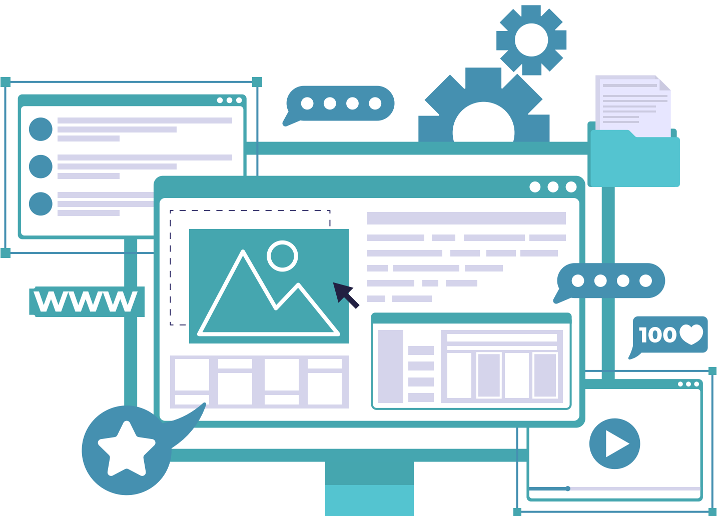
```
import matplotlib.pyplot as plt

df = pd.read_csv("titanic.csv")
age_data = df['Age']

plt.hist(age_data, bins=20)
plt.xlabel('Age')
plt.ylabel('Number')
plt.title('Passenger Age Distribution')
plt.show()
```

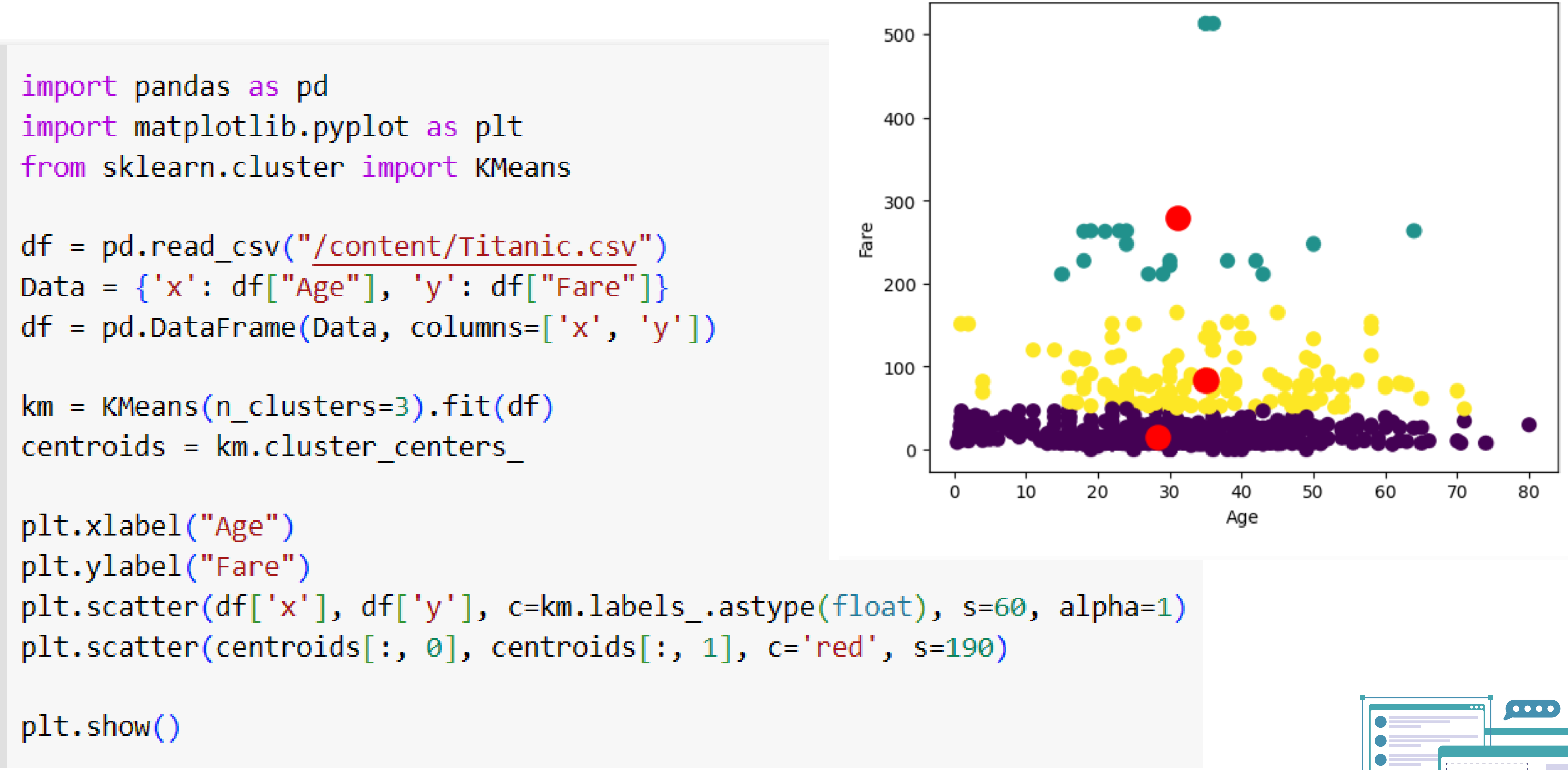
Predictive Technique (K Means)



```
import pandas as pd
import matplotlib.pyplot as plt
from sklearn.cluster import KMeans

df = pd.read_csv("/content/Titanic.csv")
Data = {'x': df["Age"], 'y': df["Fare"]}
df = pd.DataFrame(Data, columns=['x', 'y'])

plt.xlabel("Age")
plt.ylabel("Fare")
plt.scatter(df['x'], df['y'])

plt.show()
```







```
import pandas as pd
import matplotlib.pyplot as plt
from sklearn.cluster import KMeans

df = pd.read_csv("/content/Titanic.csv")
Data = {'x': df["Age"], 'y': df["Fare"]}
df = pd.DataFrame(Data, columns=['x', 'y'])

km = KMeans(n_clusters=3).fit(df)
centroids = km.cluster_centers_

plt.xlabel("Age")
plt.ylabel("Fare")
plt.scatter(df['x'], df['y'], c=km.labels_.astype(float), s=60, alpha=1)
plt.scatter(centroids[:, 0], centroids[:, 1], c='red', s=190)

plt.show()
```

Application

- By performing data manipulation techniques such as cleaning, filtering, and transforming the dataset, you can gain a deeper understanding of the data. Exploring summary statistics, distributions, and correlations between variables can provide insights into the characteristics and relationships within the dataset.

- Visualizing the Titanic dataset can help uncover patterns, trends, and relationships between variables. Plots such as histograms, scatter plots and bar charts can provide visual representations.

- After performing data manipulation, visualizing the data, and clustering using K-means, the resulting clusters can serve as new features for predictive modeling. The cluster labels can be used as input features to build a classification model to predict survival or any other relevant outcome.

Conclusion

- In conclusion, our analysis of the Titanic dataset has provided valuable insights into the passengers and the factors influencing their survival.
- We discovered significant correlations between survival and variables such as age, gender, passenger class, and family size.
- The analysis highlighted the importance of preparedness, class disparities, and gender biases during this tragic event.
- Through data cleaning, preprocessing, visualization, and modeling, we were able to extract meaningful information.