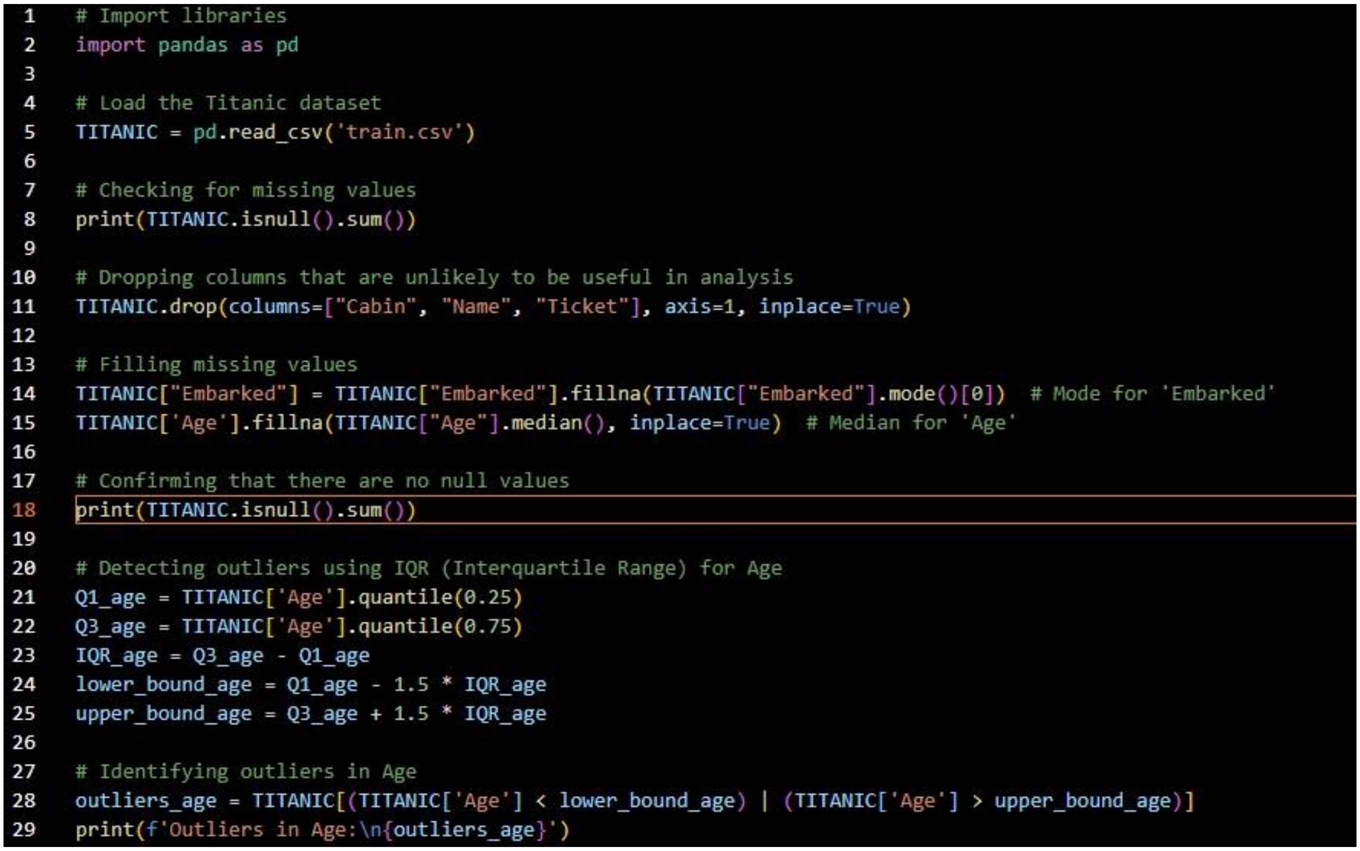
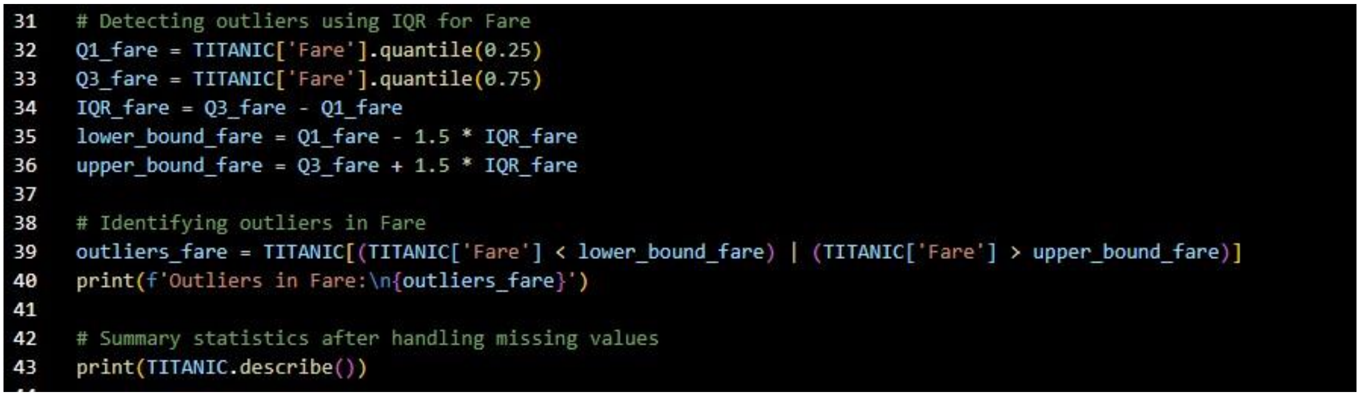
TASK 2: SUMMARY STATISTICS

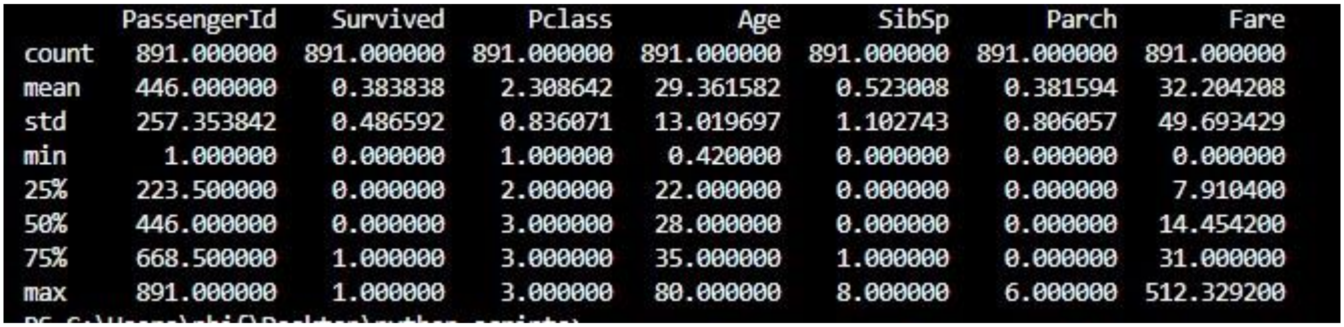
The describe () function is used to obtain summary statistics: mean, count, standard deviation, percentiles and the minimum and maximum values in each of the columns containing numerical values. Before calculating summary statistics, the data is cleaned.



```
# Import libraries
import pandas as pd

# Load the Titanic dataset
TITANIC = pd.read_csv('train.csv')

# Checking for missing values
print(TITANIC.isnull().sum())

# Dropping columns that are unlikely to be useful in analysis
TITANIC.drop(columns=["Cabin", "Name", "Ticket"], axis=1, inplace=True)

# Filling missing values
TITANIC["Embarked"] = TITANIC["Embarked"].fillna(TITANIC["Embarked"].mode()[0])  # Mode for 'Embarked'
TITANIC['Age'].fillna(TITANIC["Age"].median(), inplace=True)  # Median for 'Age'

# Confirming that there are no null values
print(TITANIC.isnull().sum())

# Detecting outliers using IQR (Interquartile Range) for Age
Q1_age = TITANIC['Age'].quantile(0.25)
Q3_age = TITANIC['Age'].quantile(0.75)
IQR_age = Q3_age - Q1_age
lower_bound_age = Q1_age - 1.5 * IQR_age
upper_bound_age = Q3_age + 1.5 * IQR_age

# Identifying outliers in Age
outliers_age = TITANIC[(TITANIC['Age'] < lower_bound_age) | (TITANIC['Age'] > upper_bound_age)]
print(f'Outliers in Age:\n{outliers_age}')
```



```
# Detecting outliers using IQR for Fare
Q1_fare = TITANIC['Fare'].quantile(0.25)
Q3_fare = TITANIC['Fare'].quantile(0.75)
IQR_fare = Q3_fare - Q1_fare
lower_bound_fare = Q1_fare - 1.5 * IQR_fare
upper_bound_fare = Q3_fare + 1.5 * IQR_fare

# Identifying outliers in Fare
outliers_fare = TITANIC[(TITANIC['Fare'] < lower_bound_fare) | (TITANIC['Fare'] > upper_bound_fare)]
print(f'Outliers in Fare:\n{outliers_fare}')

# Summary statistics after handling missing values
print(TITANIC.describe())
```



```
       PassengerId    Survived      Pclass         Age       SibSp       Parch        Fare
count   891.000000  891.000000  891.000000  891.000000  891.000000  891.000000  891.000000
mean    446.000000    0.383838    2.308642   29.361582    0.523008    0.381594   32.204208
std     257.353842    0.486592    0.836071   13.019697    1.102743    0.806057   49.693429
min       1.000000    0.000000    1.000000    0.420000    0.000000    0.000000    0.000000
25%     223.500000    0.000000    2.000000   22.000000    0.000000    0.000000    7.910400
50%     446.000000    0.000000    3.000000   28.000000    0.000000    0.000000   14.454200
75%     668.500000    1.000000    3.000000   35.000000    1.000000    0.000000   31.000000
max     891.000000    1.000000    3.000000   80.000000    8.000000    6.000000  512.329200
```

Insights from the output:

a) General Observation

   - The count value for each of the columns indicates that the data had a total of 891 entries.

b) Survival rate

   - The mean statistic (0.383838) indicates that only 38.38% of the passengers survived while the other 61.62% perished in the tragedy.

c) Passenger class

   - A mean of 2.308642 indicates that majority of the passengers were in second and third class since the mean is closer to 2.
   - The minimum (1) and maximum (3) values confirms that there were only three passenger classes.

d) Age

   - A mean of 29.36 suggests a young passenger demographic onboard i.e. around 29 years.
   - The oldest passenger was 80 years old and the youngest was about five months old (0.42 years)
   - A median value of 28 years may suggest normal distribution of the age data since the median is close to the mean.

e) No. of siblings and spouses on board (SibSp)

   - The mean (0.52) suggests that majority of the passengers either travelled with one or no sibling or spouse.

- The maximum value (8) reveals that there was a passenger who had 8 family members which is significantly higher compared to the average.

f) No. of parents and children (Parch)

- The mean being close to zero indicates that majority of the passengers had no parents or children on board.
- The maximum value (6) confirms that there were few passengers who travelled with larger families.

g) Fare

- On average, a ticket costed $32.20. However, this statistic is likely skewed due to the high cost of first class tickets.
- The minimum cost ($0) suggests that there were passengers on board who did not pay for tickets. These may have been crew members who do not require to purchase tickets to board the ship.