ANALYSIS OF THE TITANIC DATASET

OVERVIEW OF THE DATASET

The titanic dataset contains different data types:

PassengerId: Unique identifier for each of the passengers

Survived: Binary indicator of survival where (0= Did not survive, 1=Survived)

Pclass: Passenger class (1=First class, 2=Second class, 3=Third class)

Name: Full name of the passenger

Sex: Gender of the passenger (male or female)

SibSp: Number of siblings or spouses on board

Parch: Number of parents or children on board

Ticket: Ticket Number

Fare: Amount paid for the ticket by a passenger

Cabin: Cabin number

Embarked: Port where the passenger Embarked (C=Cherbourg, Q= Queenstown, S=Southampton)

LOADING THE DATASET

The first step involves loading the data, getting an overview of the data and understanding the structure.

```python
# Import libraries
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt

# Load CSV file
TITANIC = pd.read_csv('train.csv')

# Viewing the first rows of the dataset
print(TITANIC.head())

# Viewing the last rows of the dataset
print(TITANIC.tail())

# Viewing a random line
print(TITANIC.sample())

# Overview of the data
print(TITANIC.info())

```

```
   PassengerId  Survived  Pclass                                               Name     Sex   Age  SibSp  Parch            Ticket     Fare Cabin Embarked
0            1         0       3                            Braund, Mr. Owen Harris    male  22.0      1      0         A/5 21171   7.2500   NaN        S
1            2         1       1  Cumings, Mrs. John Bradley (Florence Briggs Th...  female  38.0      1      0          PC 17599  71.2833   C85        C
2            3         1       3                             Heikkinen, Miss. Laina  female  26.0      0      0  STON/O2. 3101282   7.9250   NaN        S
3            4         1       1       Futrelle, Mrs. Jacques Heath (Lily May Peel)  female  35.0      1      0            113803  53.1000  C123        S
4            5         0       3                           Allen, Mr. William Henry    male  35.0      0      0            373450   8.0500   NaN        S
     PassengerId  Survived  Pclass                                      Name     Sex   Age  SibSp  Parch      Ticket   Fare Cabin Embarked
886          887         0       2                     Montvila, Rev. Juozas    male  27.0      0      0      211536  13.00   NaN        S
887          888         1       1              Graham, Miss. Margaret Edith  female  19.0      0      0      112053  30.00   B42        S
888          889         0       3  Johnston, Miss. Catherine Helen "Carrie"  female   NaN      1      2  W./C. 6607  23.45   NaN        S
889          890         1       1                     Behr, Mr. Karl Howell    male  26.0      0      0      111369  30.00  C148        C
890          891         0       3                       Dooley, Mr. Patrick    male  32.0      0      0      370376   7.75   NaN        Q
     PassengerId  Survived  Pclass                 Name   Sex  Age  SibSp  Parch Ticket   Fare Cabin Embarked
522          523         0       3  Lahoud, Mr. Sarkis  male  NaN      0      0   2624  7.225   NaN        C
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 891 entries, 0 to 890
Data columns (total 12 columns):
 #   Column       Non-Null Count  Dtype  
---  ------       --------------  -----  
 0   PassengerId  891 non-null    int64  
 1   Survived     891 non-null    int64  
 2   Pclass       891 non-null    int64  
 3   Name         891 non-null    object 
 4   Sex          891 non-null    object 
 5   Age          714 non-null    float64
 6   SibSp        891 non-null    int64  
 7   Parch        891 non-null    int64  
 8   Ticket       891 non-null    object 
 9   Fare         891 non-null    float64
 10  Cabin        204 non-null    object 
 11  Embarked     889 non-null    object 
dtypes: float64(2), int64(5), object(5)
```

TASK 1: **DATA CLEANING**

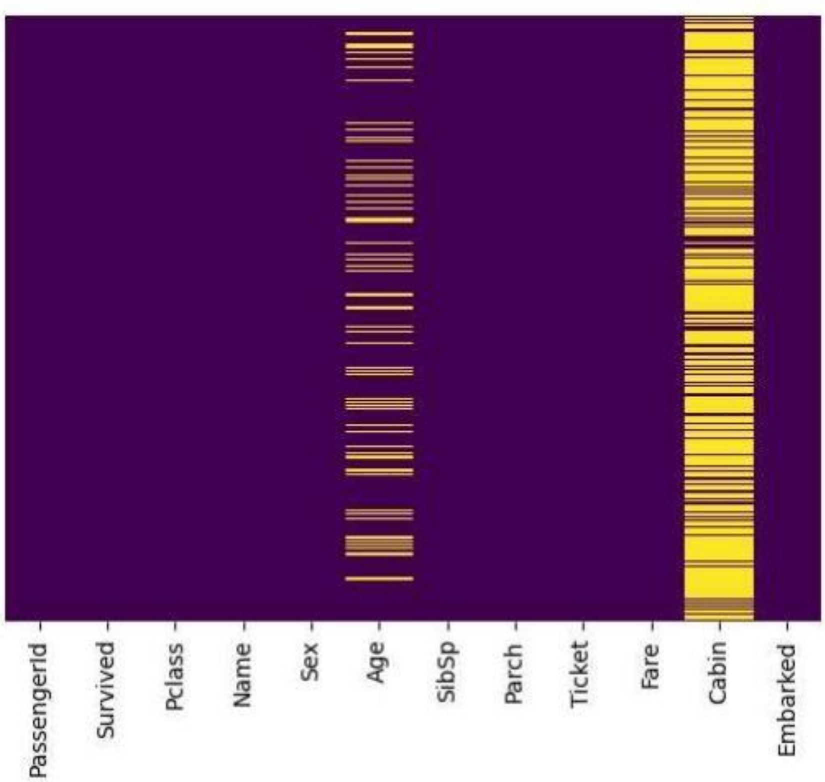
This step involves identifying missing values, outliers and duplicates.

```python
# Checking columns with missing values
print(TITANIC.isnull().sum())

# Heatmap to show the missing values in each column
sns.heatmap(TITANIC.isnull(), yticklabels=False, cmap="viridis", cbar=False)

#  show the plot
plt.show()
```

```
PassengerId      0
Survived         0
Pclass           0
Name             0
Sex              0
Age            177
SibSp            0
Parch            0
Ticket           0
Fare             0
Cabin          687
Embarked         2
dtype: int64
```



From the output there are no duplicates in the data but there are missing values in three columns, i.e. Age column, Cabin Column and Embarked Column.

- **Cabin**: Since the Cabin column contains a significant number of missing values it is dropped. The Name and Ticket Columns are also dropped alongside the cabin column since they are not used for further data analysis.
- **Embarked:** There are 2 missing entries in the embarked column. These two entries are replaced with the mode of the column.

- **Age:** To replace the missing values in the age column, we use the median of the age.

```
# Dropping columns
TITANIC.drop(columns=["Cabin", "Name", "Ticket"], axis=1, inplace=True)

# Filling null values in the 'Embarked' column with the mode (most frequent value)
TITANIC["Embarked"] = TITANIC["Embarked"].fillna(TITANIC["Embarked"].mode()[0])

# Filling missing age values with the median
TITANIC['Age'].fillna(TITANIC["Age"].median(), inplace=True)

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

```
PassengerId    0
Survived       0
Pclass         0
Sex            0
Age            0
SibSp          0
Parch          0
Fare           0
Embarked       0
dtype: int64
```

To remove the outliers, we use the interquartile range method. In this method, outliers are defined as values outside the range of:

Q1 -1.5*IQR and Q3+1.5*IQR

where Q1 is the first quartile, Q3 is the third quartile and IQR is the interquartile range.

```
# Remove outliers in Age
TITANIC = TITANIC[(TITANIC['Age'] >= lower_bound_age) & (TITANIC['Age'] <= upper_bound_age)]

# 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

# Remove outliers in Fare
TITANIC = TITANIC[(TITANIC['Fare'] >= lower_bound_fare) & (TITANIC['Fare'] <= upper_bound_fare)]
```

```
     PassengerId  Survived  Pclass     Sex   Age  SibSp  Parch     Fare Embarked
0              1         0       3    male  22.0      1      0   7.2500        S
2              3         1       3  female  26.0      0      0   7.9250        S
3              4         1       1  female  35.0      1      0  53.1000        S
4              5         0       3    male  35.0      0      0   8.0500        S
5              6         0       3    male  28.0      0      0   8.4583        Q
..           ...       ...     ...     ...   ...    ...    ...      ...      ...
886          887         0       2    male  27.0      0      0  13.0000        S
887          888         1       1  female  19.0      0      0  30.0000        S
888          889         0       3  female  28.0      1      2  23.4500        S
889          890         1       1    male  26.0      0      0  30.0000        C
890          891         0       3    male  32.0      0      0   7.7500        Q

[718 rows x 9 columns]
```