```python
import pandas as pd

# Load the training data
df = pd.read_csv(r'C:\Users\8noor\Downloads\Titanic.csv')

# Find the number of rows and columns
rows, columns = df.shape
print(f"Number of rows: {rows}")
print(f"Number of columns: {columns}")

# Find the list of columns
columns_list = df.columns.tolist()
print(f"List of columns: {columns_list}")

# Find the number of missing values in each column
missing_values = df.isnull().sum()
print("Number of missing values in each column:")
print(missing_values)

# Statistical analysis of the column values
statistics = df.describe(include='all')
print("Statistical analysis of the column values:")
print(statistics)

# Create a data quality report
data_quality_report = pd.DataFrame({
    'Data Type': df.dtypes,
    'Missing Values': df.isnull().sum(),
    'Unique Values': df.nunique(),
    'Mean': df.mean(numeric_only=True),
    'Median': df.median(numeric_only=True),
    'Standard Deviation': df.std(numeric_only=True)
})
print("Data Quality Report:")
print(data_quality_report)

# Identify important features
important_features = ['Pclass', 'Sex', 'Age', 'SibSp', 'Parch', 'Fare']
print(f"Important features: {important_features}")

# List columns to drop and the reason
columns_to_drop = ['PassengerId', 'Name', 'Ticket', 'Cabin']
print(f"Columns to drop: {columns_to_drop}")
```

```python
# Reasons for dropping columns:
# - PassengerId: Unique identifier, not useful for prediction.
# - Name: Mostly unique, but title extraction could be useful.
# - Ticket: Unique identifier, requires complex feature engineering.
# - Cabin: Too many missing values.

# Drop the specified columns
df = df.drop(columns=columns_to_drop)

# Design or create new relevant features
# Example: Creating a new feature 'FamilySize' from 'SibSp' and 'Parch'
df['FamilySize'] = df['SibSp'] + df['Parch'] + 1
print("New feature 'FamilySize' created.")

# Display the first few rows of the updated DataFrame
print(df.head())
```

```
Number of rows: 418
Number of columns: 16
List of columns: ['PassengerId', 'Survived', 'Pclass', 'Name', 'Sex', 'Age', 'SibSp', 'Parch', 'Ticket', 'Fare', 'Cabin', 'Embarked', 'Constant_Featur
e', 'Pclass_Duplicate', 'Quasi_Constant_Feature', 'PClass']
Number of missing values in each column:
PassengerId               0
Survived                  0
Pclass                    0
Name                      0
Sex                       0
Age                      86
SibSp                     0
Parch                     0
Ticket                    0
Fare                      1
Cabin                   327
Embarked                  0
Constant_Feature          0
Pclass_Duplicate          0
Quasi_Constant_Feature    0
PClass                    0
dtype: int64
Statistical analysis of the column values:
        PassengerId    Survived      Pclass              Name   Sex  \
count    418.000000  418.000000  418.000000               418   418
unique          NaN         NaN         NaN               418     2
top             NaN         NaN         NaN  Kelly, Mr. James  male
freq            NaN         NaN         NaN                 1   266
mean    1100.500000    0.363636    2.265550               NaN   NaN
std      120.810458    0.481622    0.841838               NaN   NaN
min      892.000000    0.000000    1.000000               NaN   NaN
25%      996.250000    0.000000    1.000000               NaN   NaN
50%     1100.500000    0.000000    3.000000               NaN   NaN
75%     1204.750000    1.000000    3.000000               NaN   NaN
max     1309.000000    1.000000    3.000000               NaN   NaN

               Age       SibSp       Parch    Ticket        Fare  \
count   332.000000  418.000000  418.000000       418  417.000000
unique         NaN         NaN         NaN       363         NaN
top            NaN         NaN         NaN  PC 17608         NaN
freq           NaN         NaN         NaN         5         NaN
mean     30.272590    0.447368    0.392344       NaN   35.627188
std      14.181209    0.896760    0.981429       NaN   55.907576
min       0.170000    0.000000    0.000000       NaN    0.000000
25%      21.000000    0.000000    0.000000       NaN    7.895800
50%      27.000000    0.000000    0.000000       NaN   14.454200
75%      39.000000    1.000000    0.000000       NaN   31.500000
max      76.000000    8.000000    9.000000       NaN  512.329200
```

```
              Age       SibSp       Parch    Ticket        Fare  \
count   332.000000  418.000000  418.000000       418  417.000000
unique         NaN         NaN         NaN       363         NaN
top            NaN         NaN         NaN  PC 17608         NaN
freq           NaN         NaN         NaN         5         NaN
mean     30.272590    0.447368    0.392344       NaN   35.627188
std      14.181209    0.896760    0.981429       NaN   55.907576
min       0.170000    0.000000    0.000000       NaN    0.000000
25%      21.000000    0.000000    0.000000       NaN    7.895800
50%      27.000000    0.000000    0.000000       NaN   14.454200
75%      39.000000    1.000000    0.000000       NaN   31.500000
max      76.000000    8.000000    9.000000       NaN  512.329200

                  Cabin Embarked Constant_Feature  Pclass_Duplicate  \
count                91      418              418        418.000000
unique               76        3                1               NaN
top     B57 B59 B63 B66        S         constant               NaN
freq                  3      270              418               NaN
mean                NaN      NaN              NaN          2.265550
std                 NaN      NaN              NaN          0.841838
min                 NaN      NaN              NaN          1.000000
25%                 NaN      NaN              NaN          1.000000
50%                 NaN      NaN              NaN          3.000000
75%                 NaN      NaN              NaN          3.000000
max                 NaN      NaN              NaN          3.000000

       Quasi_Constant_Feature      PClass
count                     418  418.000000
unique                      2         NaN
top                  constant         NaN
freq                      412         NaN
mean                      NaN   22.655502
std                       NaN    8.410681
min                       NaN   10.000000
25%                       NaN   10.000000
50%                       NaN   30.000000
75%                       NaN   30.000000
max                       NaN   30.000000
```

```
Data Quality Report:
                       Data Type  Missing Values  Unique Values         Mean  \
Age                      float64              86             79    30.272590
Cabin                     object             327             76          NaN
Constant_Feature          object               0              1          NaN
Embarked                  object               0              3          NaN
Fare                     float64               1            169    35.627188
Name                      object               0            418          NaN
PClass                     int64               0              6    22.655502
Parch                      int64               0              8     0.392344
PassengerId                int64               0            418  1100.500000
Pclass                     int64               0              3     2.265550
Pclass_Duplicate           int64               0              3     2.265550
Quasi_Constant_Feature    object               0              2          NaN
Sex                       object               0              2          NaN
SibSp                      int64               0              7     0.447368
Survived                   int64               0              2     0.363636
Ticket                    object               0            363          NaN

                           Median  Standard Deviation
Age                       27.0000           14.181209
Cabin                         NaN                 NaN
Constant_Feature              NaN                 NaN
Embarked                      NaN                 NaN
Fare                      14.4542           55.907576
Name                          NaN                 NaN
PClass                    30.0000            8.410681
Parch                      0.0000            0.981429
PassengerId             1100.5000          120.810458
Pclass                     3.0000            0.841838
Pclass_Duplicate           3.0000            0.841838
Quasi_Constant_Feature        NaN                 NaN
Sex                           NaN                 NaN
SibSp                      0.0000            0.896760
Survived                   0.0000            0.481622
Ticket                        NaN                 NaN
Important features: ['Pclass', 'Sex', 'Age', 'SibSp', 'Parch', 'Fare']
Columns to drop: ['PassengerId', 'Name', 'Ticket', 'Cabin']
```

```
New feature 'FamilySize' created.
   Survived  Pclass     Sex   Age  SibSp  Parch     Fare Embarked  \
0         0       3    male  34.5      0      0   7.8292        Q
1         1       3  female  47.0      1      0   7.0000        S
2         0       2    male  62.0      0      0   9.6875        Q
3         0       3    male  27.0      0      0   8.6625        S
4         1       3  female  22.0      1      1  12.2875        S

  Constant_Feature  Pclass_Duplicate Quasi_Constant_Feature  PClass  \
0         constant                 3               constant      30
1         constant                 3               constant      30
2         constant                 2               constant      20
3         constant                 3               constant      30
4         constant                 3               constant      30

   FamilySize
0           1
1           2
2           1
3           1
4           3
```

Based on data analysis, I identified several key features that are typically important for predicting survival on the Titanic. These include:

- **Pclass**: The passenger class serves as an indicator of socioeconomic status, which historically had a significant impact on survival rates during the Titanic disaster.
- **Sex**: Gender has been a critical factor in survival probabilities, with women and children being given priority during life-saving procedures.
- **Age**: Age is another important factor, as children and younger individuals might have been prioritized for evacuation.
- **SibSp**: The number of siblings and spouses aboard provides insight into family size, which could influence the chances of survival.
- **Parch**: Similarly, the number of parents and children aboard can indicate family size and dynamics, affecting survival strategies.
- **Fare**: This reflects the wealth and potentially the cabin location, which could correlate with quicker access to lifeboats.

  After analyzing the dataset and considering the utility of each column, I decided to drop the following columns for specific reasons:

- **PassengerId**: This is merely a unique identifier for each passenger and holds no predictive value regarding survival.
- **Name**: Although names themselves are not directly useful for prediction, extracting titles could be beneficial. However, for simplicity in initial models, I decided to drop this column.
- **Ticket**: The ticket number is largely a unique identifier with complex patterns that would require significant feature engineering to possibly extract any meaningful insight.
- **Cabin**: Due to the high percentage of missing values, it's challenging to use this feature effectively without substantial data imputation, which could introduce bias.
- **Embarked**: While the port of embarkation might provide some insights into demographics, it generally does not strongly correlate with survival outcomes compared to other available features. It could be explored further in more detailed analyses.

The meaningful feature to significantly enhance this model's predictive accuracy I implemented a new feature as Family size.

> I designed this feature by combining `SibSp` and `Parch`. This new feature represents the total number of family members on board. The rationale is that individuals with family might have different survival odds compared to those traveling alone, either through increased assistance or perhaps a higher motivation to secure spots on lifeboats.