```python
# importing libraries
import numpy as np
import matplotlib.pyplot as plt
import pandas as pd
import seaborn as sns

df = pd.read_csv('/content/titanic_dataset.csv')
print(df)
```

```
     PassengerId  Survived  Pclass  \
0              1         0       3
1              2         1       1
2              3         1       3
3              4         1       1
4              5         0       3
..           ...       ...     ...
886          887         0       2
887          888         1       1
888          889         0       3
889          890         1       1
890          891         0       3

                                                 Name     Sex   Age
SibSp  \
0                             Braund, Mr. Owen Harris    male  22.0
1
1    Cumings, Mrs. John Bradley (Florence Briggs Th...  female  38.0
1
2                              Heikkinen, Miss. Laina  female  26.0
0
3         Futrelle, Mrs. Jacques Heath (Lily May Peel)  female  35.0
1
4                            Allen, Mr. William Henry    male  35.0
0
..                                                 ...     ...   ...
...
886                            Montvila, Rev. Juozas    male  27.0
0
887                     Graham, Miss. Margaret Edith  female  19.0
0
888           Johnston, Miss. Catherine Helen "Carrie"  female   NaN
1
889                              Behr, Mr. Karl Howell    male  26.0
0
890                                Dooley, Mr. Patrick    male  32.0
0

     Parch            Ticket     Fare Cabin Embarked
0        0         A/5 21171   7.2500   NaN        S
1        0          PC 17599  71.2833   C85        C
```

```
2         0  STON/O2. 3101282   7.9250   NaN        S
3         0            113803  53.1000  C123        S
4         0            373450   8.0500   NaN        S
..      ...               ...      ...   ...      ...
886       0            211536  13.0000   NaN        S
887       0            112053  30.0000   B42        S
888       2        W./C. 6607  23.4500   NaN        S
889       0            111369  30.0000  C148        C
890       0            370376   7.7500   NaN        Q

[891 rows x 12 columns]
```

```
df.describe()
```


```
{"summary":"{\n  \"name\": \"df\",\n  \"rows\": 8,\n  \"fields\": [\n{\n      \"column\": \"PassengerId\",\n      \"properties\": {\n\"dtype\": \"number\",\n        \"std\": 320.8159711429855,\n\"min\": 1.0,\n        \"max\": 891.0,\n        \"num_unique_values\":6,\n        \"samples\": [\n          891.0,\n          446.0,\n668.5\n        ],\n        \"semantic_type\": \"\",\n\"description\": \"\"\n      }\n    },\n    {\n      \"column\":\"Survived\",\n      \"properties\": {\n        \"dtype\":\"number\",\n        \"std\": 314.8713661874558,\n        \"min\":0.0,\n        \"max\": 891.0,\n        \"num_unique_values\": 5,\n\"samples\": [\n          0.3838383838383838,\n          1.0,\n0.4865924542648575\n        ],\n        \"semantic_type\": \"\",\n\"description\": \"\"\n      }\n    },\n    {\n      \"column\":\"Pclass\",\n      \"properties\": {\n        \"dtype\": \"number\",\n\"std\": 314.2523437079694,\n        \"min\": 0.836071240977049,\n\"max\": 891.0,\n        \"num_unique_values\": 6,\n\"samples\": [\n          891.0,\n          2.308641975308642,\n3.0\n        ],\n        \"semantic_type\": \"\",\n\"description\": \"\"\n      }\n    },\n    {\n      \"column\":\"Age\",\n      \"properties\": {\n        \"dtype\": \"number\",\n\"std\": 242.9056731818781,\n        \"min\": 0.42,\n        \"max\":714.0,\n        \"num_unique_values\": 8,\n        \"samples\": [\n29.69911764705882,\n          28.0,\n          714.0\n        ],\n\"semantic_type\": \"\",\n        \"description\": \"\"\n      }\n    },\n    {\n      \"column\": \"SibSp\",\n      \"properties\": {\n        \"dtype\": \"number\",\n        \"std\": 314.4908277465442,\n\"min\": 0.0,\n        \"max\": 891.0,\n        \"num_unique_values\":6,\n        \"samples\": [\n          891.0,\n0.5230078563411896,\n          8.0\n        ],\n\"semantic_type\": \"\",\n        \"description\": \"\"\n      }\n    },\n    {\n      \"column\": \"Parch\",\n      \"properties\": {\n        \"dtype\": \"number\",\n        \"std\": 314.65971717879,\n\"min\": 0.0,\n        \"max\": 891.0,\n        \"num_unique_values\":5,\n        \"samples\": [\n          0.38159371492704824,\n6.0,\n          0.8060572211299483\n        ],\n\"semantic_type\": \"\",\n        \"description\": \"\"\n      }\
```

```
n    },\n    {\n      \"column\": \"Fare\",\n      \"properties\": {\n
\"dtype\": \"number\",\n        \"std\": 330.6256632228578,\n
\"min\": 0.0,\n        \"max\": 891.0,\n        \"num_unique_values\":
8,\n        \"samples\": [\n          32.204207968574636,\n
14.4542,\n          891.0\n        ],\n        \"semantic_type\":
\"\",\n        \"description\": \"\"\n      }\n    }\n  ]\
n}","type":"dataframe"}
```


```
df.info()
```

```
<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     882 non-null    object
dtypes: float64(2), int64(5), object(5)
memory usage: 83.7+ KB
```

```
df.head()
```


```
{"summary":"{\n  \"name\": \"df\",\n  \"rows\": 891,\n  \"fields\": [\
n    {\n      \"column\": \"PassengerId\",\n      \"properties\": {\n
\"dtype\": \"number\",\n        \"std\": 257,\n        \"min\": 1,\n
\"max\": 891,\n        \"num_unique_values\": 891,\n
\"samples\": [\n          710,\n          440,\n          841\n
],\n        \"semantic_type\": \"\",\n        \"description\": \"\"\n
}\n    },\n    {\n      \"column\": \"Survived\",\n
\"properties\": {\n        \"dtype\": \"number\",\n        \"std\":
0,\n        \"min\": 0,\n        \"max\": 1,\n
\"num_unique_values\": 2,\n        \"samples\": [\n          1,\n
0\n        ],\n        \"semantic_type\": \"\",\n
\"description\": \"\"\n      }\n    },\n    {\n      \"column\":
\"Pclass\",\n      \"properties\": {\n        \"dtype\": \"number\",\n
\"std\": 0,\n        \"min\": 1,\n        \"max\": 3,\n
\"num_unique_values\": 3,\n        \"samples\": [\n          3,\n
1\n        ],\n        \"semantic_type\": \"\",\n
\"description\": \"\"\n      }\n    },\n    {\n      \"column\":
\"Name\",\n      \"properties\": {\n        \"dtype\": \"string\",\n
```

\"num_unique_values\": 891,\n   \"samples\": [\n
\"Moubarek, Master. Halim Gonios (\\\"William George\\\")\",\n
\"Kvillner, Mr. Johan Henrik Johannesson\"\n   ],\n
\"semantic_type\": \"\",\n   \"description\": \"\"\n   }\
n   },\n   {\n   \"column\": \"Sex\",\n   \"properties\": {\n
\"dtype\": \"category\",\n   \"num_unique_values\": 2,\n
\"samples\": [\n   \"female\",\n   \"male\"\n   ],\
n   \"semantic_type\": \"\",\n   \"description\": \"\"\n
}\n   },\n   {\n   \"column\": \"Age\",\n   \"properties\": {\
n   \"dtype\": \"number\",\n   \"std\": 14.526497332334042,\
n   \"min\": 0.42,\n   \"max\": 80.0,\n
\"num_unique_values\": 88,\n   \"samples\": [\n   0.75,\n
22.0\n   ],\n   \"semantic_type\": \"\",\n
\"description\": \"\"\n   }\n   },\n   {\n   \"column\":
\"SibSp\",\n   \"properties\": {\n   \"dtype\": \"number\",\n
\"std\": 1,\n   \"min\": 0,\n   \"max\": 8,\n
\"num_unique_values\": 7,\n   \"samples\": [\n   1,\n
0\n   ],\n   \"semantic_type\": \"\",\n
\"description\": \"\"\n   }\n   },\n   {\n   \"column\":
\"Parch\",\n   \"properties\": {\n   \"dtype\": \"number\",\n
\"std\": 0,\n   \"min\": 0,\n   \"max\": 6,\n
\"num_unique_values\": 7,\n   \"samples\": [\n   0,\n
1\n   ],\n   \"semantic_type\": \"\",\n
\"description\": \"\"\n   }\n   },\n   {\n   \"column\":
\"Ticket\",\n   \"properties\": {\n   \"dtype\": \"string\",\n
\"num_unique_values\": 681,\n   \"samples\": [\n
\"11774\",\n   \"248740\"\n   ],\n
\"semantic_type\": \"\",\n   \"description\": \"\"\n   }\
n   },\n   {\n   \"column\": \"Fare\",\n   \"properties\": {\n
\"dtype\": \"number\",\n   \"std\": 49.6934285971809,\n
\"min\": 0.0,\n   \"max\": 512.3292,\n
\"num_unique_values\": 248,\n   \"samples\": [\n
11.2417,\n   51.8625\n   ],\n   \"semantic_type\":
\"\",\n   \"description\": \"\"\n   }\n   },\n   {\n
\"column\": \"Cabin\",\n   \"properties\": {\n   \"dtype\":
\"category\",\n   \"num_unique_values\": 147,\n
\"samples\": [\n   \"D45\",\n   \"B49\"\n   ],\n
\"semantic_type\": \"\",\n   \"description\": \"\"\n   }\
n   },\n   {\n   \"column\": \"Embarked\",\n   \"properties\":
{\n   \"dtype\": \"category\",\n   \"num_unique_values\":
3,\n   \"samples\": [\n   \"S\",\n   \"C\"\n
],\n   \"semantic_type\": \"\",\n   \"description\": \"\"\n
}\n   }\n]\n}","type":"dataframe","variable_name":"df"}

```
df.tail()
```

{"summary":"{\n \"name\": \"df\",\n \"rows\": 5,\n \"fields\": [\n
{\n   \"column\": \"PassengerId\",\n   \"properties\": {\n
\"dtype\": \"number\",\n   \"std\": 1,\n   \"min\": 887,\n
\"max\": 891,\n   \"num_unique_values\": 5,\n   \"samples\":

```
[\n          888,\n          891,\n          889\n        ],\n
\"semantic_type\": \"\",\n        \"description\": \"\"\n      }\
n    },\n    {\n      \"column\": \"Survived\",\n      \"properties\":
{\n        \"dtype\": \"number\",\n        \"std\": 0,\n
\"min\": 0,\n        \"max\": 1,\n        \"num_unique_values\": 2,\n
\"samples\": [\n          1,\n          0\n        ],\n
\"semantic_type\": \"\",\n        \"description\": \"\"\n      }\
n    },\n    {\n      \"column\": \"Pclass\",\n      \"properties\":
{\n        \"dtype\": \"number\",\n        \"std\": 1,\n
\"min\": 1,\n        \"max\": 3,\n        \"num_unique_values\": 3,\n
\"samples\": [\n          2,\n          1\n        ],\n
\"semantic_type\": \"\",\n        \"description\": \"\"\n      }\
n    },\n    {\n      \"column\": \"Name\",\n      \"properties\": {\n
\"dtype\": \"string\",\n        \"num_unique_values\": 5,\n
\"samples\": [\n          \"Graham, Miss. Margaret Edith\",\n
\"Dooley, Mr. Patrick\"\n        ],\n        \"semantic_type\": \"\",\
n        \"description\": \"\"\n      }\n    },\n    {\n
\"column\": \"Sex\",\n      \"properties\": {\n        \"dtype\":
\"category\",\n        \"num_unique_values\": 2,\n        \"samples\":
[\n          \"female\",\n          \"male\"\n        ],\n
\"semantic_type\": \"\",\n        \"description\": \"\"\n      }\
n    },\n    {\n      \"column\": \"Age\",\n      \"properties\": {\n
\"dtype\": \"number\",\n        \"std\": 5.354126134736337,\n
\"min\": 19.0,\n        \"max\": 32.0,\n        \"num_unique_values\":
4,\n        \"samples\": [\n          19.0,\n          32.0\
n        ],\n        \"semantic_type\": \"\",\n
\"description\": \"\"\n      }\n    },\n    {\n      \"column\":
\"SibSp\",\n      \"properties\": {\n        \"dtype\": \"number\",\n
\"std\": 0,\n        \"min\": 0,\n        \"max\": 1,\n
\"num_unique_values\": 2,\n        \"samples\": [\n          1,\n
0\n        ],\n        \"semantic_type\": \"\",\n
\"description\": \"\"\n      }\n    },\n    {\n      \"column\":
\"Parch\",\n      \"properties\": {\n        \"dtype\": \"number\",\n
\"std\": 0,\n        \"min\": 0,\n        \"max\": 2,\n
\"num_unique_values\": 2,\n        \"samples\": [\n          2,\n
0\n        ],\n        \"semantic_type\": \"\",\n
\"description\": \"\"\n      }\n    },\n    {\n      \"column\":
\"Ticket\",\n      \"properties\": {\n        \"dtype\": \"string\",\n
\"num_unique_values\": 5,\n        \"samples\": [\n
\"112053\",\n          \"370376\"\n        ],\n
\"semantic_type\": \"\",\n        \"description\": \"\"\n      }\
n    },\n    {\n      \"column\": \"Fare\",\n      \"properties\": {\n
\"dtype\": \"number\",\n        \"std\": 10.09253436952285,\n
\"min\": 7.75,\n        \"max\": 30.0,\n        \"num_unique_values\":
4,\n        \"samples\": [\n          30.0,\n          7.75\
n        ],\n        \"semantic_type\": \"\",\n
\"description\": \"\"\n      }\n    },\n    {\n      \"column\":
\"Cabin\",\n      \"properties\": {\n        \"dtype\": \"category\",\
n        \"num_unique_values\": 2,\n        \"samples\": [\n
```

\"C148\",\n   \"B42\"\n   ],\n   \"semantic_type\": \"\",\n   \"description\": \"\"\n   }\n   },\n   {\n   \"column\": \"Embarked\",\n   \"properties\": {\n   \"dtype\": \"string\",\n   \"num_unique_values\": 3,\n   \"samples\": [\n   \"S\",\n   \"C\"\n   ],\n   \"semantic_type\": \"\",\n   \"description\": \"\"\n   }\n   }\n]\n}","type":"dataframe"}

```
print(df.shape)
```

```
(891, 12)
```

```
df.isnull().sum()
```

```
PassengerId      0
Survived         0
Pclass           0
Name             0
Sex              0
Age            177
SibSp            0
Parch            0
Ticket           0
Fare             0
Cabin          687
Embarked         9
dtype: int64
```

```
df.isnull().count()
```

```
PassengerId    891
Survived       891
Pclass         891
Name           891
Sex            891
Age            891
SibSp          891
Parch          891
Ticket         891
Fare           891
Cabin          891
Embarked       891
dtype: int64
```

```
df['Age'].fillna(df['Age'].median(), inplace = True)
```

```
df.isnull().sum()
```

```
PassengerId      0
Survived         0
Pclass           0
```

```
Name             0
Sex              0
Age              0
SibSp            0
Parch            0
Ticket           0
Fare             0
Cabin          687
Embarked         9
dtype: int64
```

```
print(df.duplicated().sum())
```

```
0
```

```
# encoding categorical variables
df['Sex'] = df['Sex'].map({'male':1, 'female':2})
print(df['Sex'])
```

```
0      1
1      2
2      2
3      2
4      1
      ..
886    1
887    2
888    2
889    1
890    1
Name: Sex, Length: 891, dtype: int64
```

```
# one-hot encoding
df = pd.get_dummies(df, columns = ['Embarked'], drop_first = True)
print(df)
```

```
     PassengerId  Survived  Pclass  \
0              1         0       3
1              2         1       1
2              3         1       3
3              4         1       1
4              5         0       3
..           ...       ...     ...
886          887         0       2
887          888         1       1
888          889         0       3
889          890         1       1
890          891         0       3

                                                  Name  Sex   Age
SibSp  \
```

```
0                               Braund, Mr. Owen Harris    1  22.0
1
1    Cumings, Mrs. John Bradley (Florence Briggs Th...    2  38.0
1
2                                Heikkinen, Miss. Laina    2  26.0
0
3          Futrelle, Mrs. Jacques Heath (Lily May Peel)    2  35.0
1
4                              Allen, Mr. William Henry    1  35.0
0
..                                                   ...  ...   ...    .
..
886                              Montvila, Rev. Juozas    1  27.0
0
887                       Graham, Miss. Margaret Edith    2  19.0
0
888           Johnston, Miss. Catherine Helen "Carrie"    2  28.0
1
889                              Behr, Mr. Karl Howell    1  26.0
0
890                                Dooley, Mr. Patrick    1  32.0
0

     Parch            Ticket     Fare Cabin  Embarked_Q  Embarked_S
0        0         A/5 21171   7.2500   NaN       False        True
1        0          PC 17599  71.2833   C85       False       False
2        0  STON/O2. 3101282   7.9250   NaN       False        True
3        0            113803  53.1000  C123       False        True
4        0            373450   8.0500   NaN       False        True
..     ...               ...      ...   ...         ...         ...
886      0            211536  13.0000   NaN       False        True
887      0            112053  30.0000   B42       False        True
888      2        W./C. 6607  23.4500   NaN       False        True
889      0            111369  30.0000  C148       False       False
890      0            370376   7.7500   NaN        True       False

[891 rows x 13 columns]
```

```
df = df.drop(['Name', 'Ticket', 'Cabin'], axis =1)
df
```

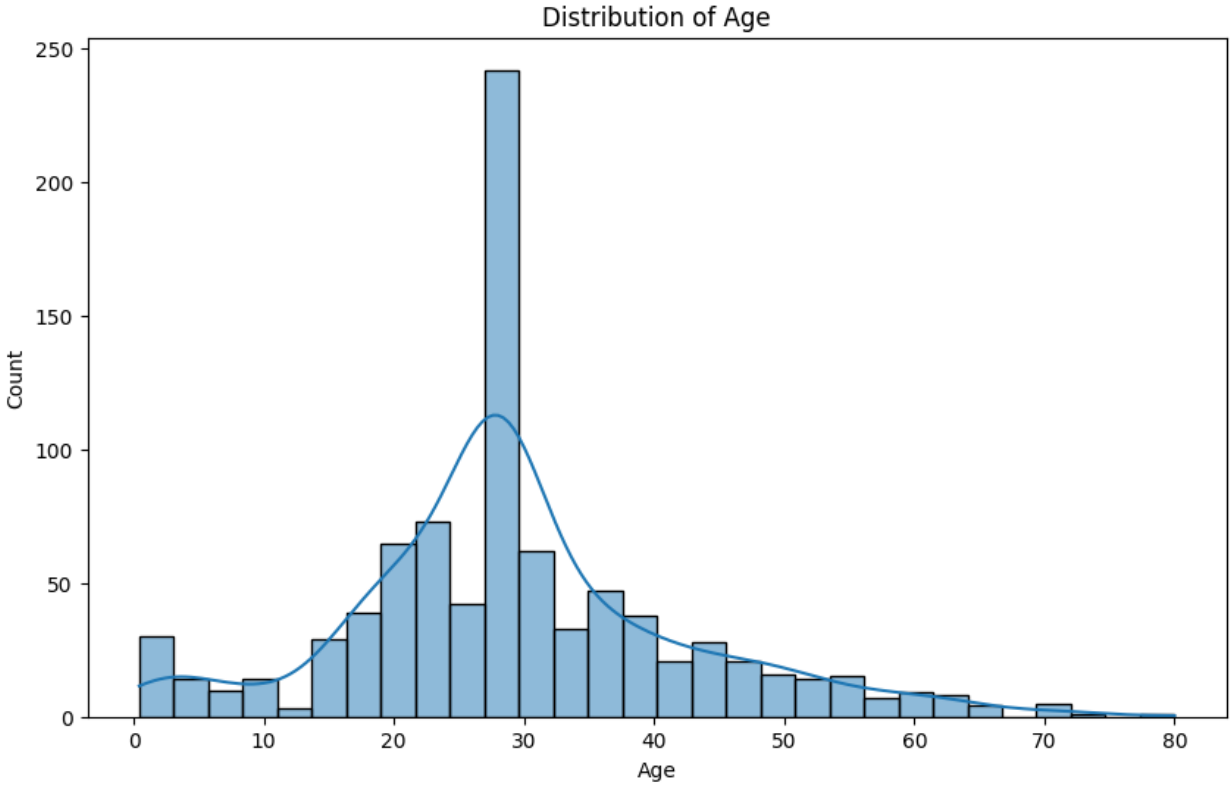

```
{"summary":"{\n  \"name\": \"df\",\n  \"rows\": 891,\n  \"fields\": [\
n    {\n      \"column\": \"PassengerId\",\n      \"properties\": {\n
\"dtype\": \"number\",\n        \"std\": 257,\n        \"min\": 1,\n
\"max\": 891,\n        \"num_unique_values\": 891,\n
\"samples\": [\n          710,\n          440,\n          841\n
],\n        \"semantic_type\": \"\",\n        \"description\": \"\"\n
}\n    },\n    {\n      \"column\": \"Survived\",\n
\"properties\": {\n        \"dtype\": \"number\",\n        \"std\":
0,\n        \"min\": 0,\n        \"max\": 1,\n
```

```
\"num_unique_values\": 2,\n        \"samples\": [\n          1,\n
0\n        ],\n        \"semantic_type\": \"\",\n
\"description\": \"\"\n      }\n    },\n    {\n      \"column\":
\"Pclass\",\n      \"properties\": {\n        \"dtype\": \"number\",\n
\"std\": 0,\n        \"min\": 1,\n        \"max\": 3,\n
\"num_unique_values\": 3,\n        \"samples\": [\n          3,\n
1\n        ],\n        \"semantic_type\": \"\",\n
\"description\": \"\"\n      }\n    },\n    {\n      \"column\":
\"Sex\",\n      \"properties\": {\n        \"dtype\": \"number\",\n
\"std\": 0,\n        \"min\": 1,\n        \"max\": 2,\n
\"num_unique_values\": 2,\n        \"samples\": [\n          2,\n
1\n        ],\n        \"semantic_type\": \"\",\n
\"description\": \"\"\n      }\n    },\n    {\n      \"column\":
\"Age\",\n      \"properties\": {\n        \"dtype\": \"number\",\n
\"std\": 13.019696550973201,\n        \"min\": 0.42,\n        \"max\":
80.0,\n        \"num_unique_values\": 88,\n        \"samples\": [\n
0.75,\n          22.0\n        ],\n        \"semantic_type\": \"\",\n
\"description\": \"\"\n      }\n    },\n    {\n      \"column\":
\"SibSp\",\n      \"properties\": {\n        \"dtype\": \"number\",\n
\"std\": 1,\n        \"min\": 0,\n        \"max\": 8,\n
\"num_unique_values\": 7,\n        \"samples\": [\n          1,\n
0\n        ],\n        \"semantic_type\": \"\",\n
\"description\": \"\"\n      }\n    },\n    {\n      \"column\":
\"Parch\",\n      \"properties\": {\n        \"dtype\": \"number\",\n
\"std\": 0,\n        \"min\": 0,\n        \"max\": 6,\n
\"num_unique_values\": 7,\n        \"samples\": [\n          0,\n
1\n        ],\n        \"semantic_type\": \"\",\n
\"description\": \"\"\n      }\n    },\n    {\n      \"column\":
\"Fare\",\n      \"properties\": {\n        \"dtype\": \"number\",\n
\"std\": 49.6934285971809,\n        \"min\": 0.0,\n        \"max\":
512.3292,\n        \"num_unique_values\": 248,\n        \"samples\":
[\n          11.2417,\n          51.8625\n        ],\n
\"semantic_type\": \"\",\n        \"description\": \"\"\n      }\
n    },\n    {\n      \"column\": \"Embarked_Q\",\n
\"properties\": {\n        \"dtype\": \"boolean\",\n
\"num_unique_values\": 2,\n        \"samples\": [\n          true,\n
false\n        ],\n        \"semantic_type\": \"\",\n
\"description\": \"\"\n      }\n    },\n    {\n      \"column\":
\"Embarked_S\",\n      \"properties\": {\n        \"dtype\":
\"boolean\",\n        \"num_unique_values\": 2,\n        \"samples\":
[\n          false,\n          true\n        ],\n
\"semantic_type\": \"\",\n        \"description\": \"\"\n      }\
n    }\n  ]\n}","type":"dataframe","variable_name":"df"}
```

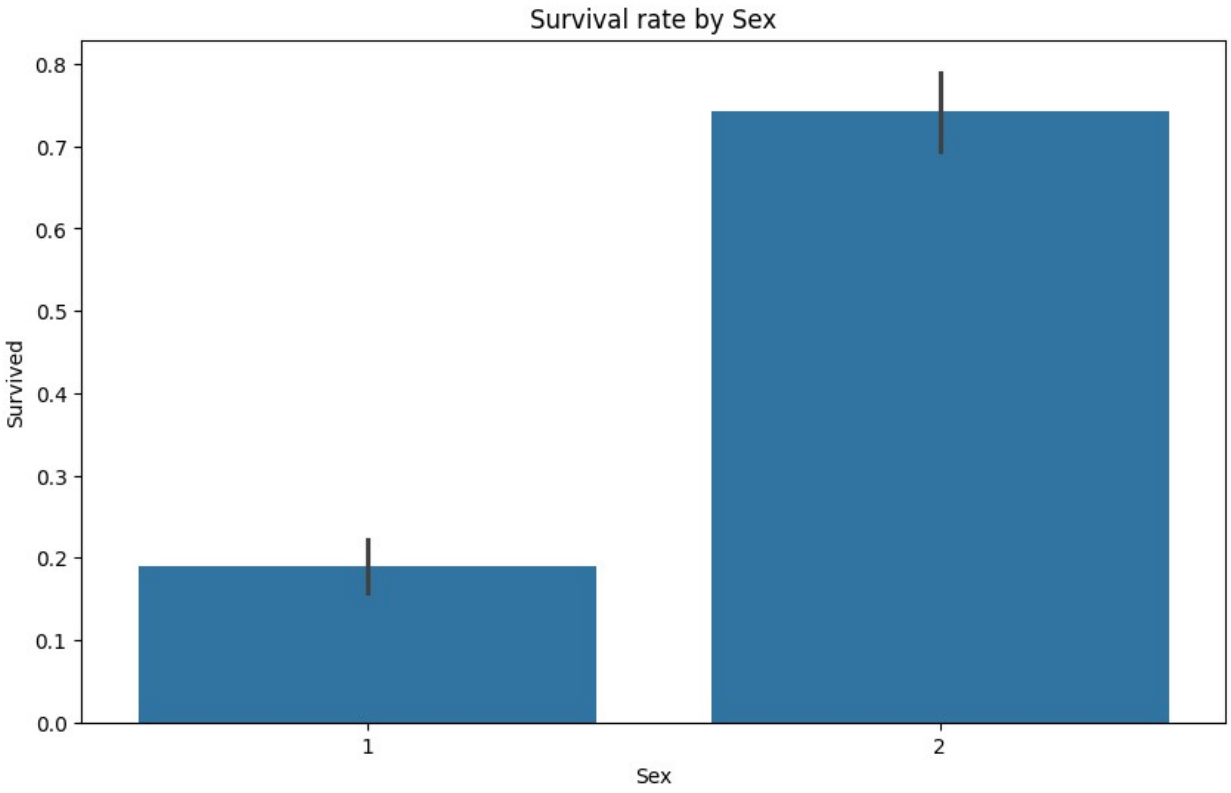

Apply univariate, bivariate, and multi-variate analysis

```
# univariate analysis
plt.figure(figsize = (10,6))
sns.histplot(df['Age'], bins = 30, kde = True)
```

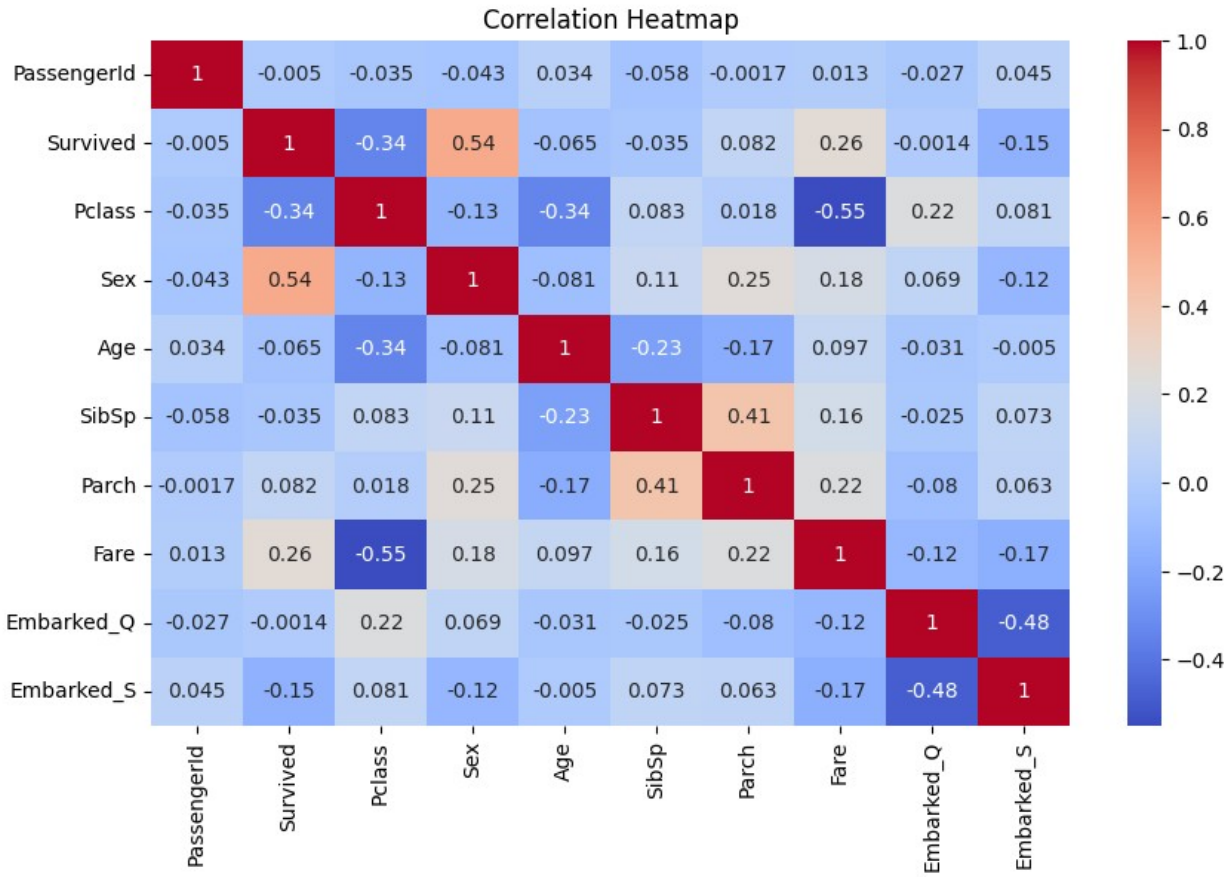
```
plt.title('Distribution of Age')
plt.show()
```



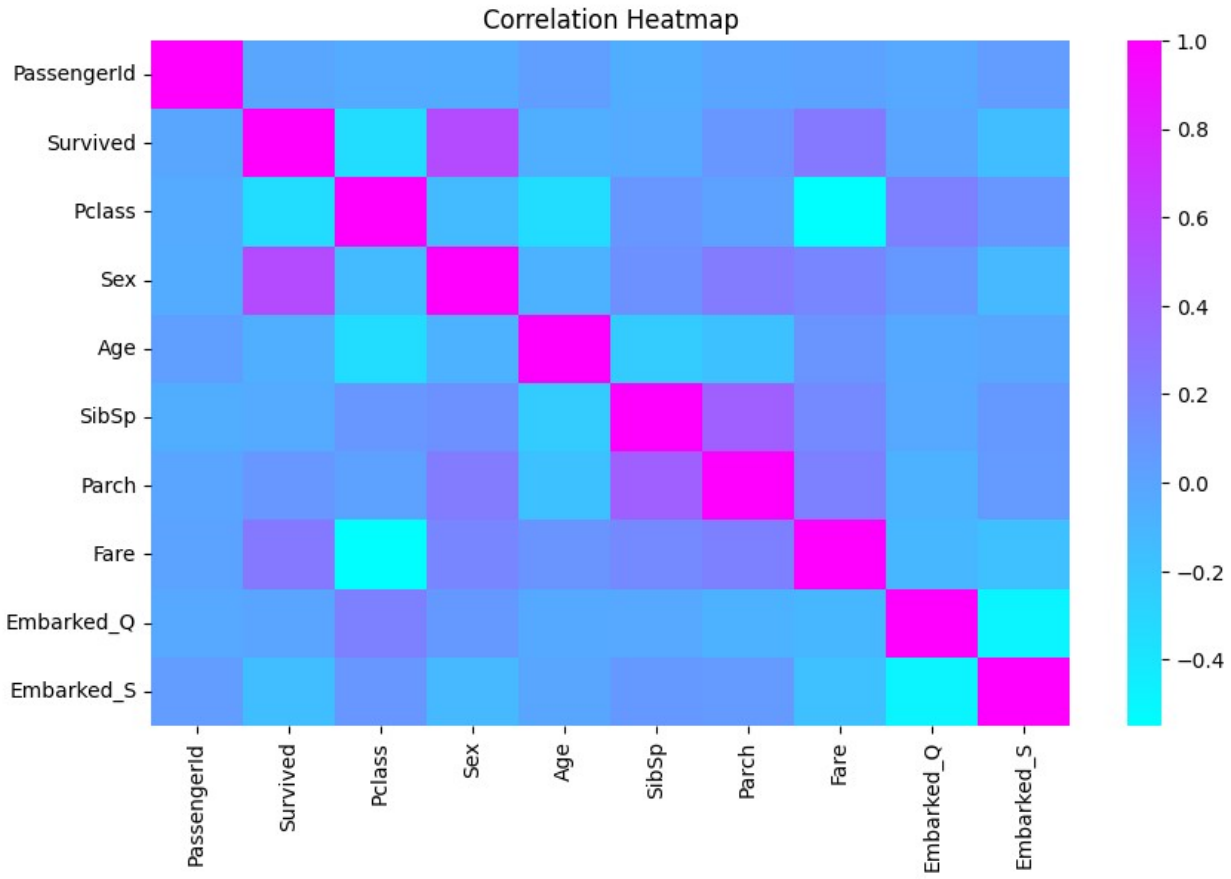

```
# bivariate analysis
plt.figure(figsize = (10,6))
sns.barplot(x = 'Sex', y = 'Survived', data = df)
plt.title('Survival rate by Sex')
plt.show()
```




```
# multivariate analysis
plt.figure(figsize = (10,6))
sns.heatmap(df.corr(), annot = True, cmap = 'coolwarm')
plt.title('Correlation Heatmap')
plt.show()
```




```
# multivariate analysis
plt.figure(figsize = (10,6))
sns.heatmap(df.corr(), cmap = 'cool')
plt.title('Correlation Heatmap')
plt.show()
```




```
df =
df.drop(['PassengerId','SibSp','Parch','Fare','Embarked_Q','Embarked_S
'], axis = 1)
print(df)
```

```
     Survived  Pclass  Sex   Age
0           0       3    1  22.0
1           1       1    2  38.0
2           1       3    2  26.0
3           1       1    2  35.0
4           0       3    1  35.0
..        ...     ...  ...   ...
886         0       2    1  27.0
887         1       1    2  19.0
888         0       3    2  28.0
889         1       1    1  26.0
890         0       3    1  32.0

[891 rows x 4 columns]
```

```
from google.colab import files
df.to_csv('titanic_cleaned.csv', index=False)
files.download('titanic_cleaned.csv')
```

```
<IPython.core.display.Javascript object>
<IPython.core.display.Javascript object>
```