

SmartInternz

AIML – Assignment 4

Rahul Palanivel

21BCE7828

## Assignment 15 sep

### Perform Data preprocessing on Titanic dataset

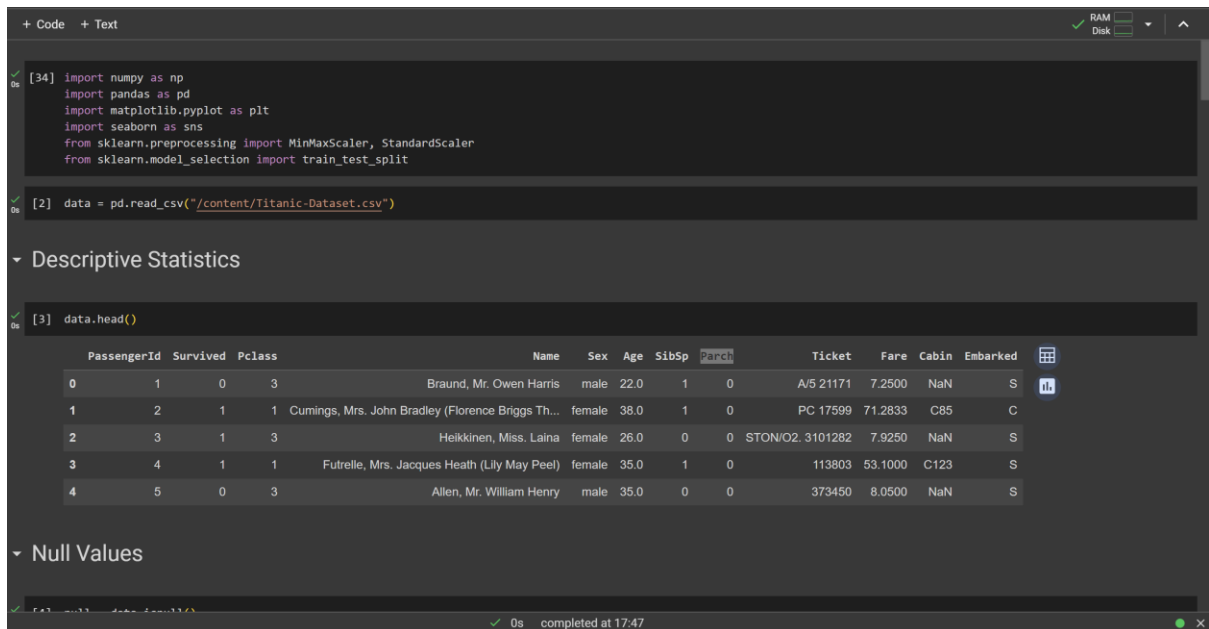
#### 1.Data Collection.

Please download the dataset from

<https://www.kaggle.com/datasets/yasserh/titanic-dataset>

#### 2.Data Preprocessing

- o Import the Libraries.
- o Importing the dataset.
- o Checking for Null Values.
- o Data Visualization.
- o Outlier Detection
- o Splitting Dependent and Independent variables
- o Perform Encoding
- o Feature Scaling.
- o Splitting Data into Train and Test



```
+ Code + Text
[34] import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.preprocessing import MinMaxScaler, StandardScaler
from sklearn.model_selection import train_test_split

[2] data = pd.read_csv("/content/Titanic-Dataset.csv")

Descriptive Statistics

[3] data.head()
```

	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

```
Null Values

[4] pd.isnull(data).sum()
```

completed at 17:47

+ Code + Text

RAM  
Disk

Null Values

```
0 null = data.isnull()
1 nullval = null.sum()
2 print(nullval)
```

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

Data visualisation

```
0 [16] sns.histplot(data['Age'], bins=20, kde=True)
1 plt.title('Age Distribution')
2 plt.xlabel('Age')
3 plt.ylabel('Count')
4 plt.show()
```

Age Distribution

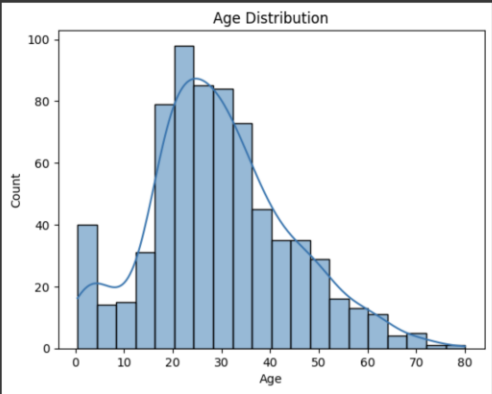
completed at 17:47

+ Code + Text

RAM  
Disk

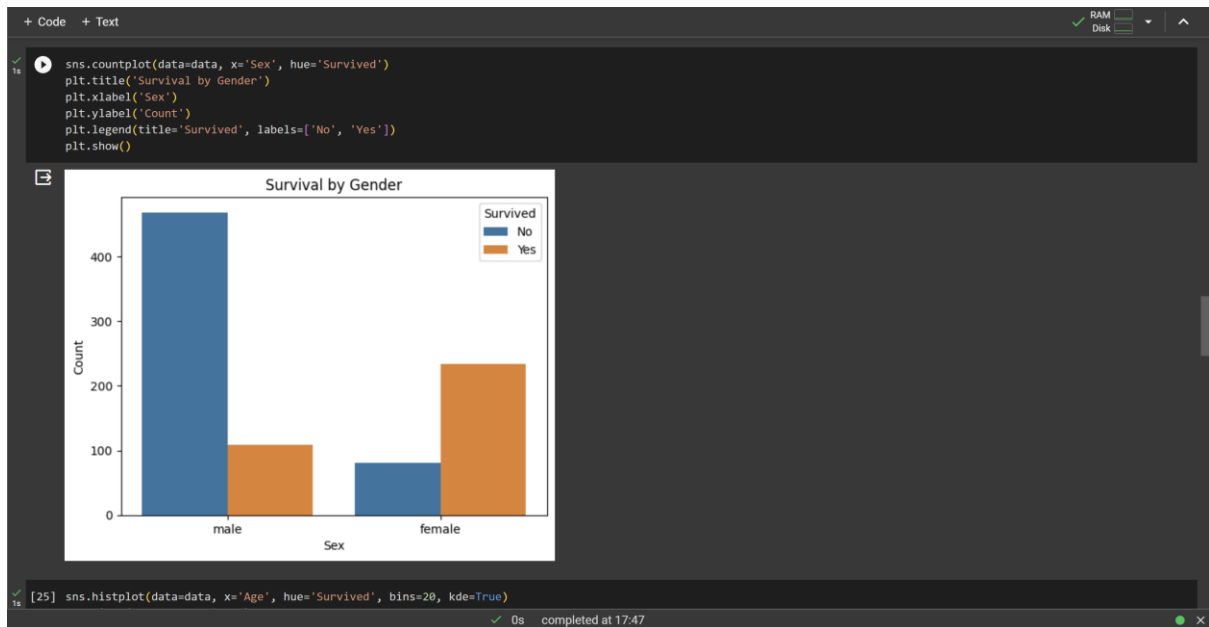
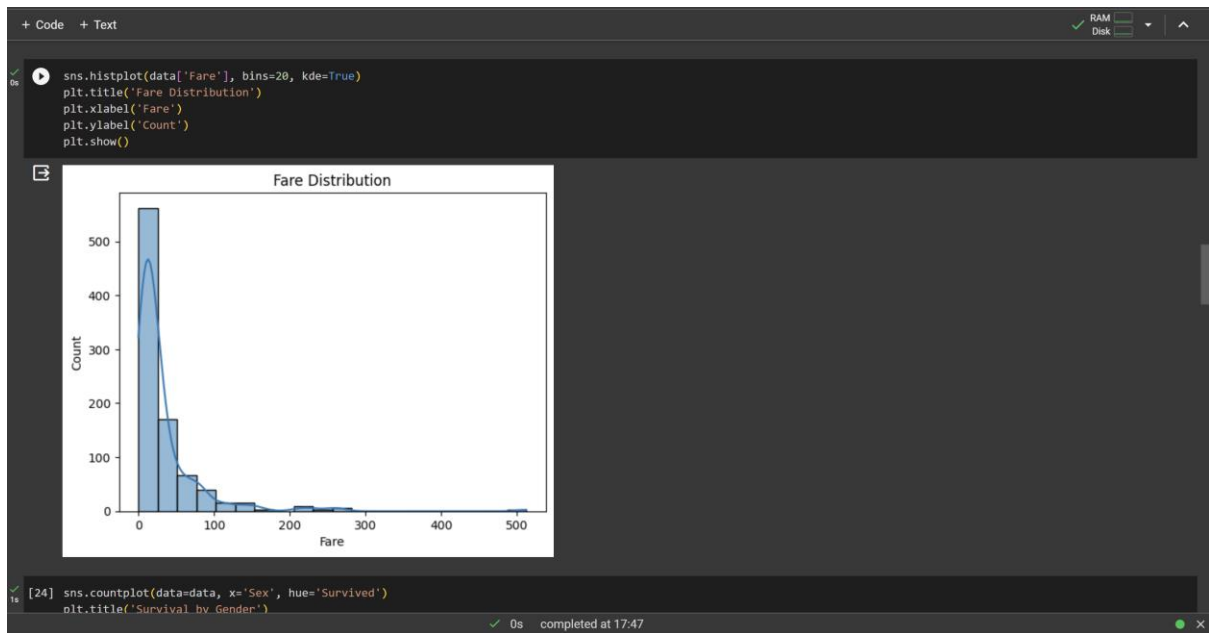
```
0 sns.histplot(data['Age'], bins=20, kde=True)
1 plt.title('Age Distribution')
2 plt.xlabel('Age')
3 plt.ylabel('Count')
4 plt.show()
```

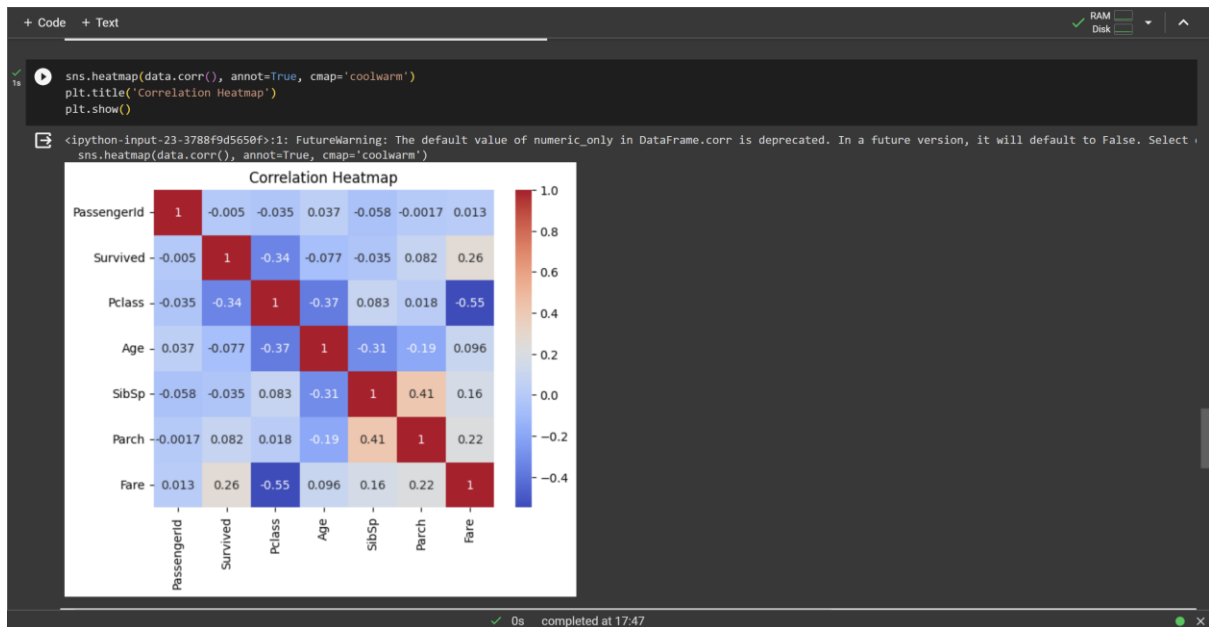
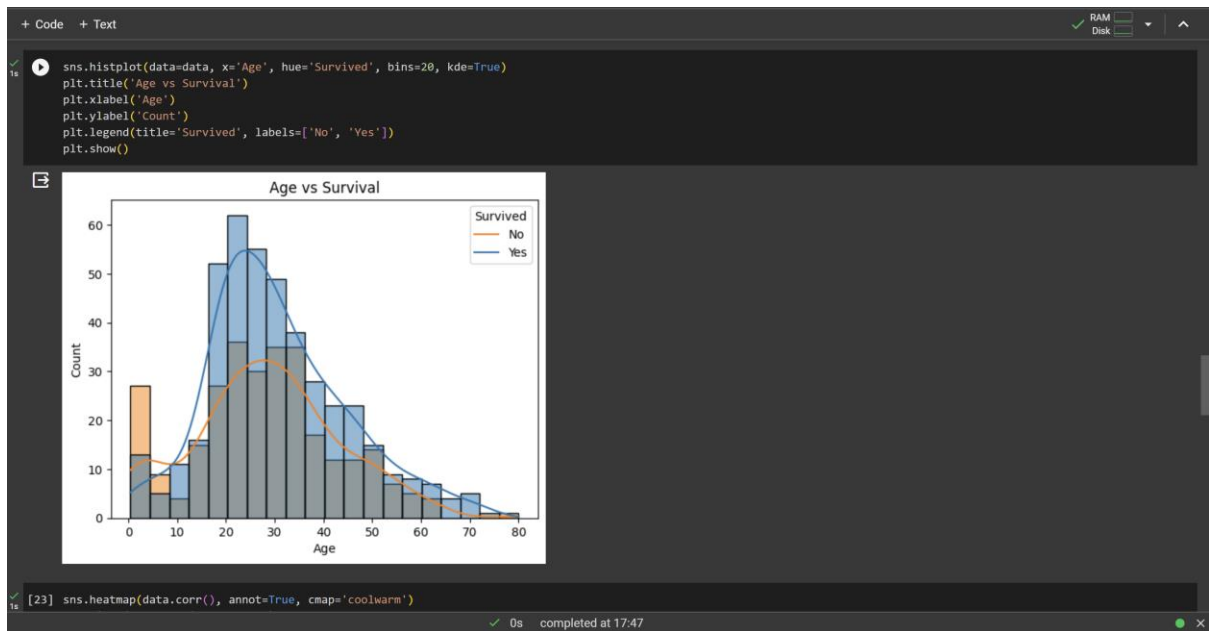
Age Distribution



```
0 [19] sns.histplot(data['Age'], bins=20, kde=True)
1 plt.title('Age Distribution')
```

completed at 17:47





+ Code+ Text

RAM  
Disk

Splitting Dependent and Independent variables

```
[26] dep = data['Survived']
      indep = data.drop(columns=['Survived'])
      X = dep.values
      y = indep.values
```

Feature Scaling

```
columns_to_scale = ['Pclass', 'Parch', 'SibSp', 'Fare']
min_max_scaler = MinMaxScaler()
data_min_max_scaled = data.copy()
data_min_max_scaled[columns_to_scale] = min_max_scaler.fit_transform(data[columns_to_scale])
print("Min-Max Scaled Data:")
print(data_min_max_scaled.head())
```

Min-Max Scaled Data:

PassengerId	Survived	Pclass	\
0	1	0	1.0
1	2	1	0.0
2	3	1	1.0
3	4	1	0.0
4	5	0	1.0

	Name	Sex	Age	SibSp	\
0	Braund, Mr. Owen Harris	male	22.0	0.125	
1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	0.125	
2	Heikkinen, Miss. Laina	female	26.0	0.000	
3	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	0.125	
4	Allen, Mr. William Henry	male	35.0	0.000	

0s completed at 17:47

+ Code+ Text

RAM  
Disk

Train Test Split

```
X = data.drop(columns=['Survived'])
y = data['Survived']
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

print("Training data shape:", X_train.shape, y_train.shape)
print("Testing data shape:", X_test.shape, y_test.shape)
```

Training data shape: (712, 11) (712,)
Testing data shape: (179, 11) (179,)

0s completed at 17:47