Prodigy InfoTech Internship:

Task-2

Perform data cleaning and exploratory data analysis (EDA) on a dataset of your choice, such as the Titanic dataset from Kaggle. Explore the relationships between variables and identify patterns and trends in the data.

Sample Dataset: Titanic

```
import warnings
warnings.filterwarnings('ignore')
import numpy as np
import pandas as pd
```

```
from google.colab import drive
drive.mount('/content/drive')
```

2 Understand the shape of the data

```
[ ] df= pd.read_csv('/content/drive/My Drive/TITANIC.csv' , index_col='PassengerId')
```

```
[ ] df.head()
```

PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 1309 entries, 1 to 1309
Data columns (total 11 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   Survived  1309 non-null   int64
 1   Pclass    1309 non-null   int64
 2   Name      1309 non-null   object
 3   Sex       1309 non-null   object
 4   Age       1046 non-null   float64
 5   SibSp     1309 non-null   int64
 6   Parch     1309 non-null   int64
 7   Ticket    1309 non-null   object
 8   Fare      1308 non-null   float64
 9   Cabin     295 non-null    object
 10  Embarked  1307 non-null   object
dtypes: float64(2), int64(4), object(5)
memory usage: 122.7+ KB
```

Aditya Rathore
Linkedin

```
df.describe()
```

	Survived	Pclass	Age	SibSp	Parch	Fare
count	1309.000000	1309.000000	1046.000000	1309.000000	1309.000000	1308.000000
mean	0.377387	2.294882	29.881138	0.498854	0.385027	33.295479
std	0.484918	0.837836	14.413493	1.041658	0.865560	51.758668
min	0.000000	1.000000	0.170000	0.000000	0.000000	0.000000
25%	0.000000	2.000000	21.000000	0.000000	0.000000	7.895800
50%	0.000000	3.000000	28.000000	0.000000	0.000000	14.454200
75%	1.000000	3.000000	39.000000	1.000000	0.000000	31.275000
max	1.000000	3.000000	80.000000	8.000000	9.000000	512.329200

```
df.drop(columns='Cabin').isna().mean()
```

```
Survived    0.000000
Pclass      0.000000
Name        0.000000
Sex         0.000000
Age         0.200917
SibSp       0.000000
Parch       0.000000
Ticket      0.000000
Fare        0.000764
Embarked    0.001528
dtype: float64
```

```
[ ] df = df.drop(columns='Cabin').dropna(subset=['Embarked'])
```

3 Data Cleaning

```
[ ] df['Age'] = df['Age'].round()
    df['Pclass'] = df['Pclass'].map({1: 'Upper', 2: 'Middle', 3: 'Lower'})
    df['Embarked'] = df['Embarked'].map({'C': 'Cherbourg', 'Q': 'Queenstown', 'S':'Southampton'})
    df['Survived'] = df['Survived'].map({0: 'Survived', 1: 'Not Survived'})
    df['Sex'] = df['Sex'].str.title()
    categorical_columns = ['Sex', 'Parch', 'SibSp', 'Pclass', 'Embarked','Survived']
    df[categorical_columns] = df[categorical_columns].astype('category')
```

```
cols = [
'Name', 'Sex', 'Age', 'Parch', 'SibSp',
'Ticket', 'Pclass', 'Embarked', 'Fare',
'Survived',]
df = df[cols]
```

```
df.head()
```

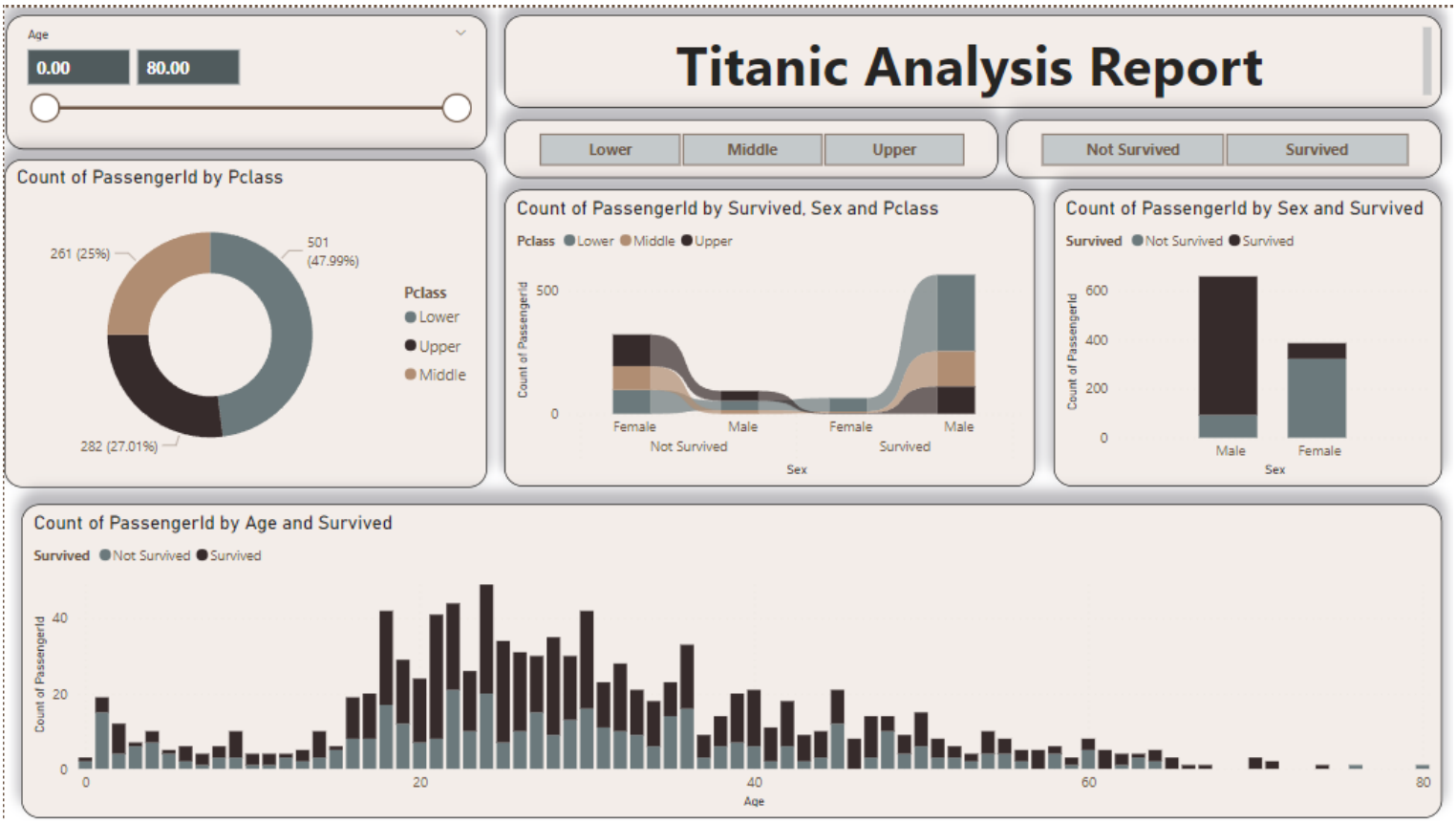
PassengerId	Name	Sex	Age	Parch	SibSp	Ticket	Pclass	Embarked	Fare	Survived
1	Braund, Mr. Owen Harris	Male	22.0	0	1	A/5 21171	Lower	Southampton	7.2500	Survived
2	Cumings, Mrs. John Bradley (Florence Briggs Th...	Female	38.0	0	1	PC 17599	Upper	Cherbourg	71.2833	Not Survived
3	Heikkinen, Miss. Laina	Female	26.0	0	0	STON/O2. 3101282	Lower	Southampton	7.9250	Not Survived
4	Futrelle, Mrs. Jacques Heath (Lily May Peel)	Female	35.0	0	1	113803	Upper	Southampton	53.1000	Not Survived
5	Allen, Mr. William Henry	Male	35.0	0	0	373450	Lower	Southampton	8.0500	Survived

Aditya Rathore
Linkedin

```
df.dropna(subset=['Age'], inplace=True)
```

```
df.to_csv('/content/drive/My Drive/titanic@2.csv')
```

4 Data Visualization using Power BI



Aditya Rathore
Linkedin