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 Dataset

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

➢ Understanding the shape of the Dataset:

```
data = pd.read_csv("/content/drive/MyDrive/Project_Datasets/Titanic_Dataset/train.csv")
data.head()  # View the top rows
```

	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

```
data.info()  # Data types and non-null counts
data.describe()  # Statistical summary for numerical columns
```

```
<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)
memory usage: 83.7+ KB
```

By: Aman Kumar Jha
LinkedIn

	PassengerId	Survived	Pclass	Age	SibSp	Parch	Fare
count	891.000000	891.000000	891.000000	714.000000	891.000000	891.000000	891.000000
mean	446.000000	0.383838	2.308642	29.699118	0.523008	0.381594	32.204208
std	257.353842	0.486592	0.836071	14.526497	1.102743	0.806057	49.693429
min	1.000000	0.000000	1.000000	0.420000	0.000000	0.000000	0.000000
25%	223.500000	0.000000	2.000000	20.125000	0.000000	0.000000	7.910400
50%	446.000000	0.000000	3.000000	28.000000	0.000000	0.000000	14.454200
75%	668.500000	1.000000	3.000000	38.000000	1.000000	0.000000	31.000000
max	891.000000	1.000000	3.000000	80.000000	8.000000	6.000000	512.329200

➢ Creating DataFrame:

```
1 df=pd.DataFrame(data)
```

➢ Data Cleaning:

```
1 df.drop(columns='Cabin').isna().mean()
2 df.drop(columns='Cabin').dropna(subset=['Embarked'])
```

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	S
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	S
...	...	...	...	...	...	...	...	...	...	...	...
886	887	0	2	Montvila, Rev. Juozas	male	27.0	0	0	211536	13.0000	S
887	888	1	1	Graham, Miss. Margaret Edith	female	19.0	0	0	112053	30.0000	S
888	889	0	3	Johnston, Miss. Catherine Helen "Carrie"	female	NaN	1	2	W./C. 6607	23.4500	S
889	890	1	1	Behr, Mr. Karl Howell	male	26.0	0	0	111369	30.0000	C
890	891	0	3	Dooley, Mr. Patrick	male	32.0	0	0	370376	7.7500	Q

```
1 df.dropna(subset=['Age'],inplace=True)
2 df.drop('Cabin', axis=1, inplace=True)
3 df
```

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	S
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	S
...	...	...	...	...	...	...	...	...	...	...	...
885	886	0	3	Rice, Mrs. William (Margaret Norton)	female	39.0	0	5	382652	29.1250	Q
886	887	0	2	Montvila, Rev. Juozas	male	27.0	0	0	211536	13.0000	S
887	888	1	1	Graham, Miss. Margaret Edith	female	19.0	0	0	112053	30.0000	S
889	890	1	1	Behr, Mr. Karl Howell	male	26.0	0	0	111369	30.0000	C
890	891	0	3	Dooley, Mr. Patrick	male	32.0	0	0	370376	7.7500	Q

By: Aman Kumar Jha
LinkedIn

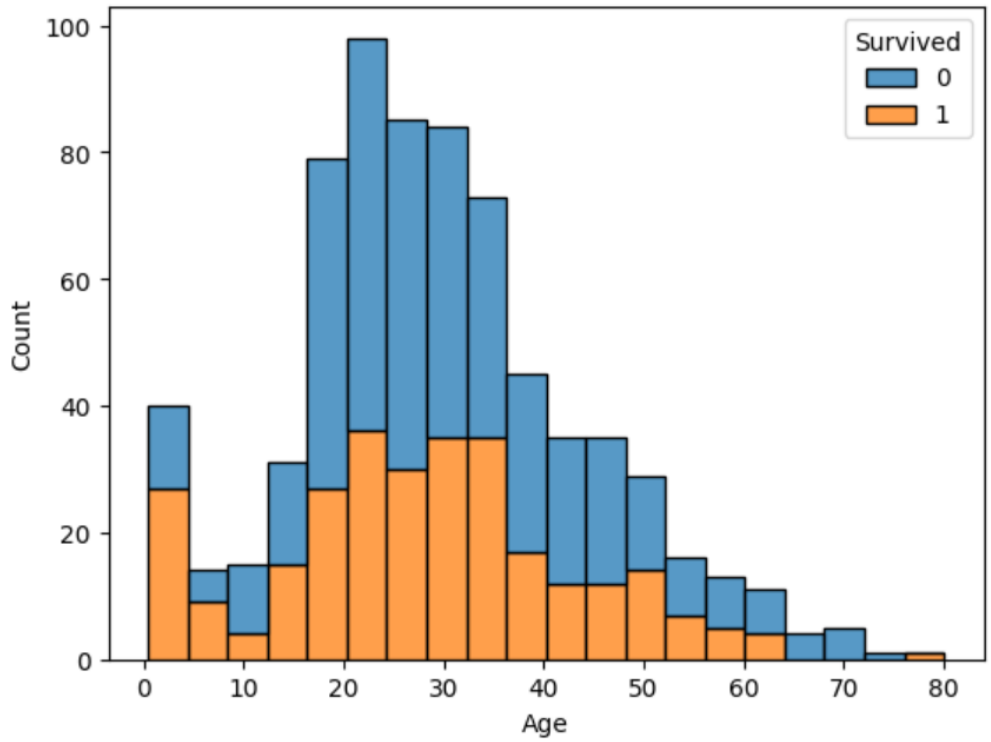
➢ Exporting to Csv for Power BI Visualisation:

```
[ ]  1 df.to_csv("/content/drive/MyDrive/Project_Datasets/Titanic_Dataset/train_Cleaned.csv")
```

➢ Data Exploration :-

```
1 sns.histplot(data=data, x="Age", hue="Survived", multiple="stack")
```

```
<Axes: xlabel='Age', ylabel='Count'>
```




```
[12]  1 survived_female_counts = data[data['Sex'] == 'female']['Survived'].value_counts()
      2 print(f"Survived: {survived_female_counts[1]}")
      3 print(f"Not Survived: {survived_female_counts[0]}")
```

```
Survived: 233
Not Survived: 81
```

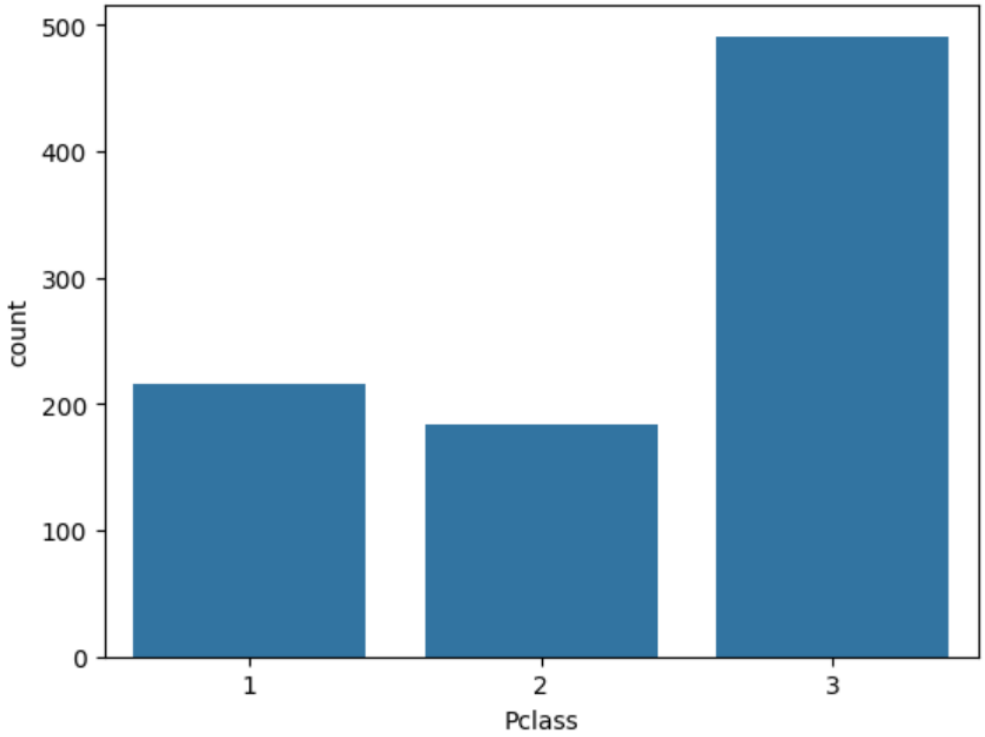
```
1 survived_male_counts = data[data['Sex'] == 'male']['Survived'].value_counts()
2 print(f"Survived: {survived_male_counts[1]}")
3 print(f"Not Survived: {survived_male_counts[0]}")
```

```
Survived: 109
Not Survived: 468
```

By: Aman Kumar Jha
LinkedIn

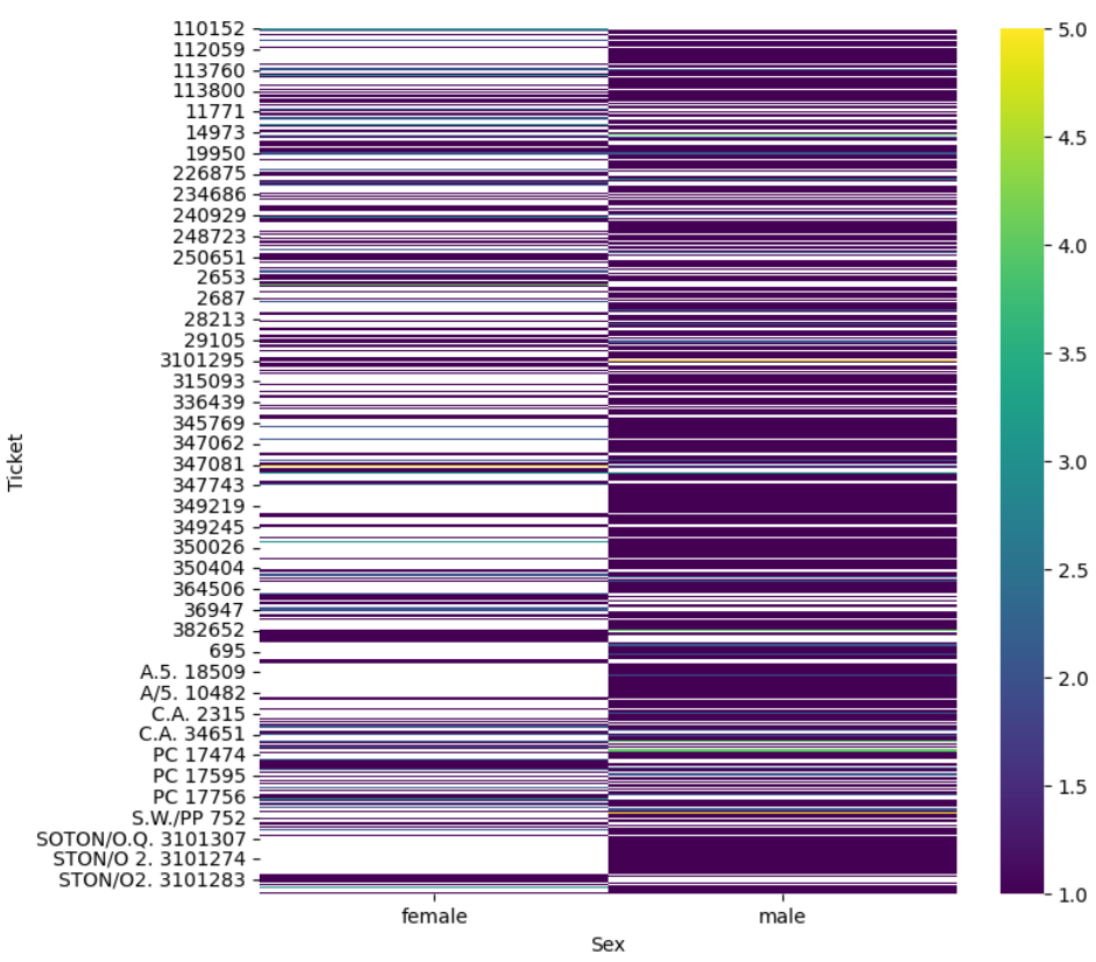
```
1 data['Survived'].value_counts()  # Counts of 'Survived' and 'Not Survived'
2 sns.countplot(x='Pclass', data=data)  # Bar plot of passenger class distribution
```

```
<Axes: xlabel='Pclass', ylabel='count'>
```

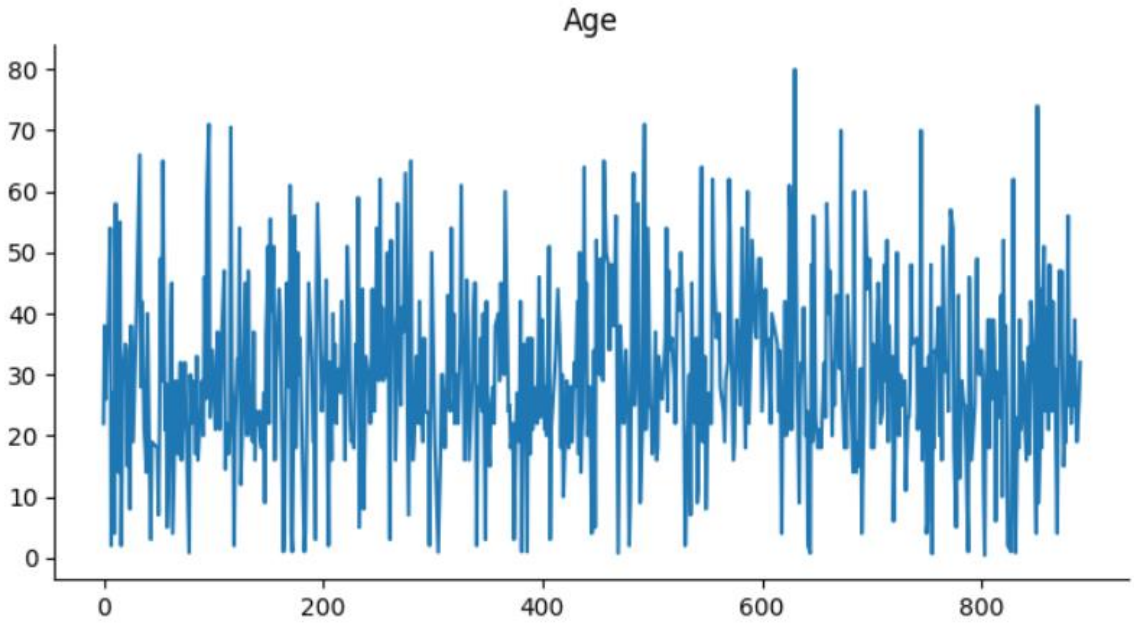



```
1 plt.subplots(figsize=(8, 8))
2 df_2dhist = pd.DataFrame({
3     x_label: grp['Ticket'].value_counts()
4     for x_label, grp in df.groupby('Sex')
5 })
6 sns.heatmap(df_2dhist, cmap='viridis')
7 plt.xlabel('Sex')
8 _ = plt.ylabel('Ticket')
```

By: Aman Kumar Jha
LinkedIn

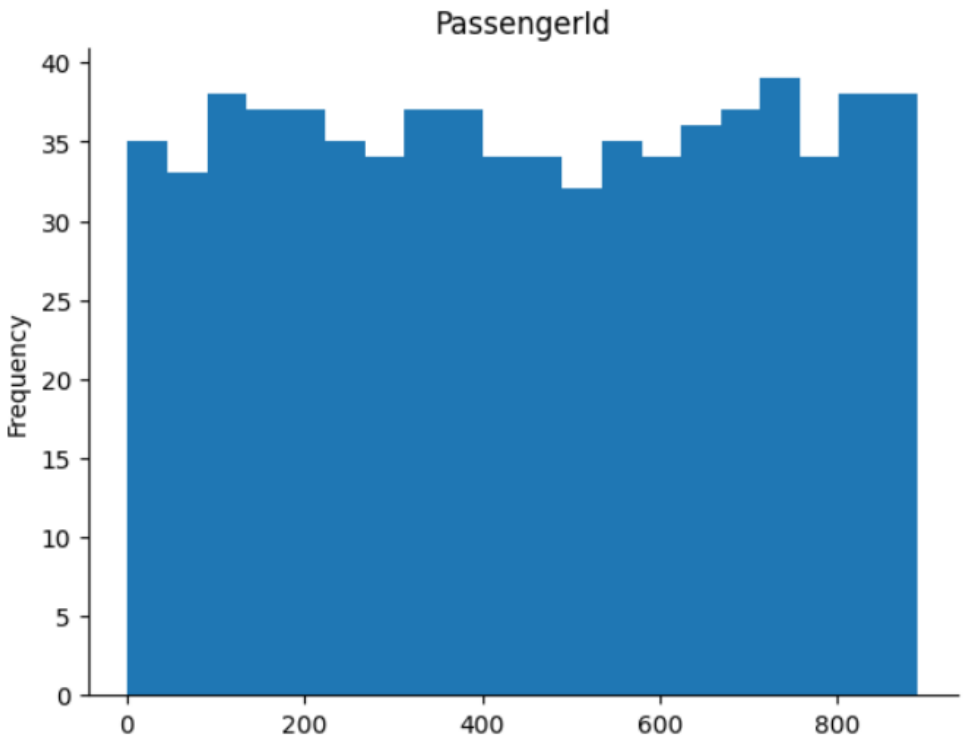



```
[16]  1 df['Age'].plot(kind='line', figsize=(8, 4), title='Age')
      2 plt.gca().spines[['top', 'right']].set_visible(False)
```



By: Aman Kumar Jha
LinkedIn

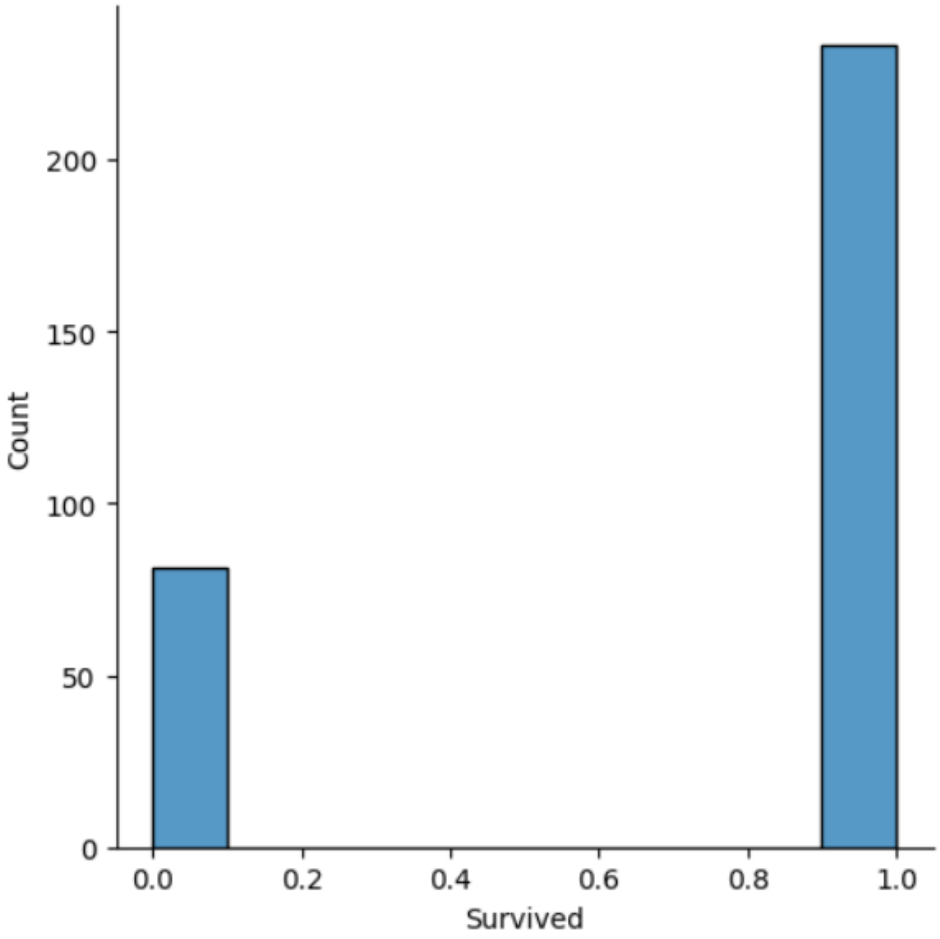
```
1 df['PassengerId'].plot(kind='hist', bins=20, title='PassengerId')
2 plt.gca().spines[['top', 'right',]].set_visible(False)
```



By: Aman Kumar Jha
LinkedIn

```
1 sns.displot(data[data['Sex'] == 'female']['Survived'])
```

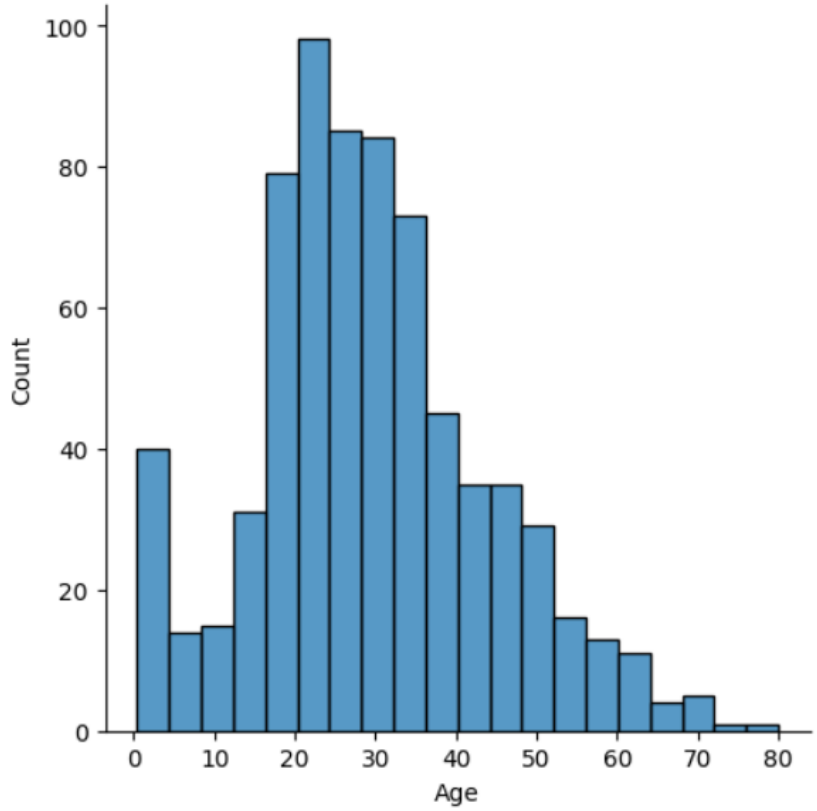
```
<seaborn.axisgrid.FacetGrid at 0x7d6a49f67940>
```



By: Aman Kumar Jha
LinkedIn

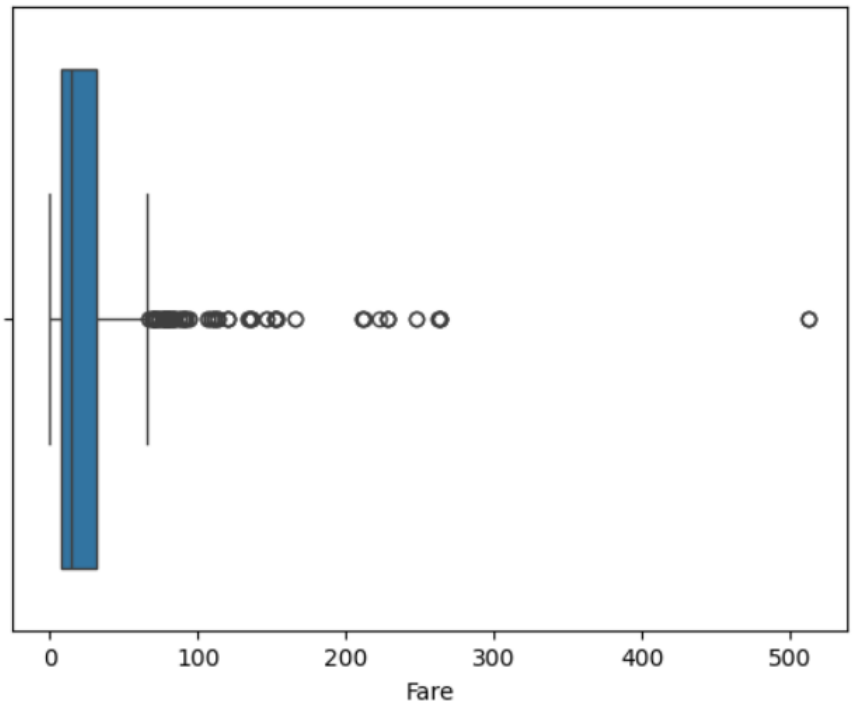
```
1 sns.displot(data['Age'])  # Histogram of age
```

```
<seaborn.axisgrid.FacetGrid at 0x7d6a49ee8a90>
```




```
1 sns.boxplot(x='Fare', data=data)  # Box plot of fares
```

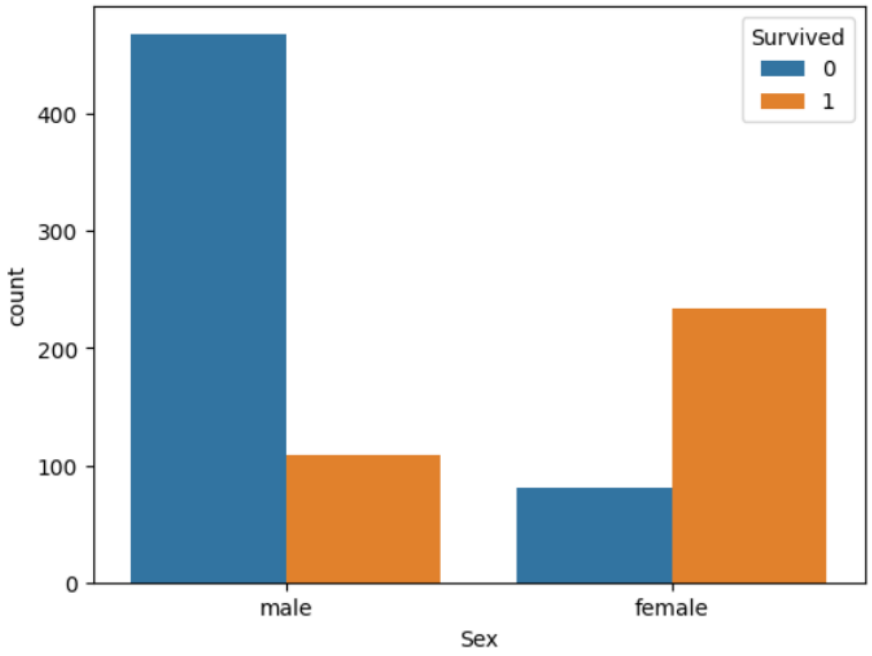
```
<Axes: xlabel='Fare'>
```



By: Aman Kumar Jha
LinkedIn

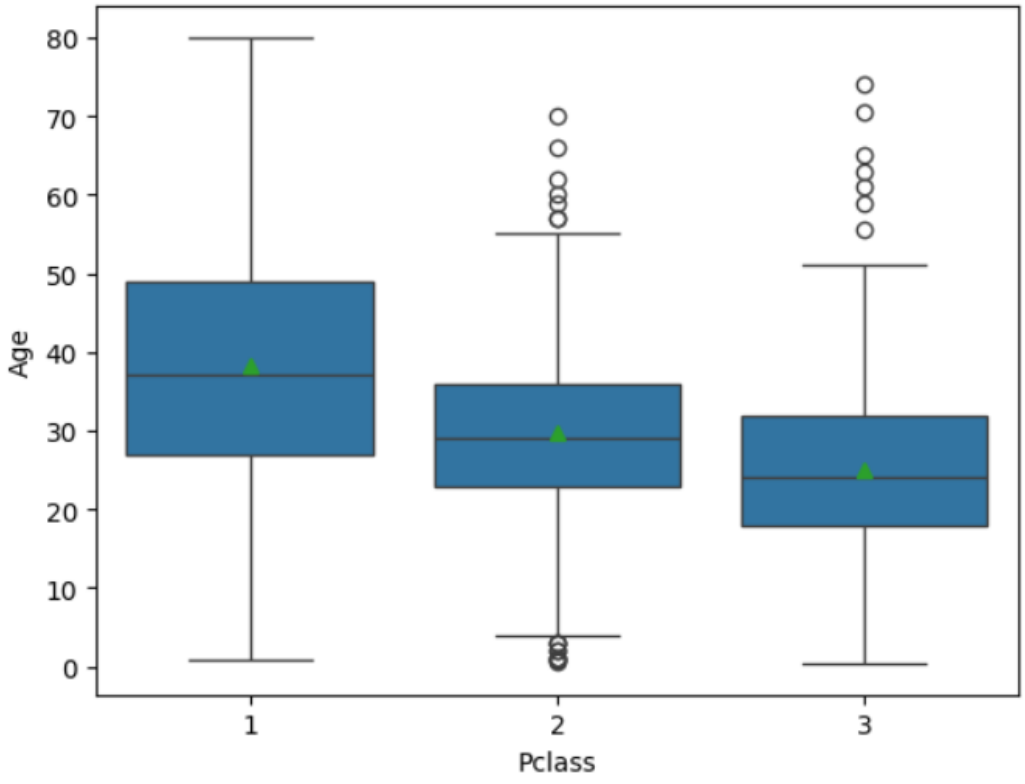
```
1 sns.countplot(x='Sex', hue='Survived', data=data)  # Survival by gender
```

```
<Axes: xlabel='Sex', ylabel='count'>
```




```
1 sns.boxplot(x='Pclass', y='Age', showmeans=True, data=data)
2
```

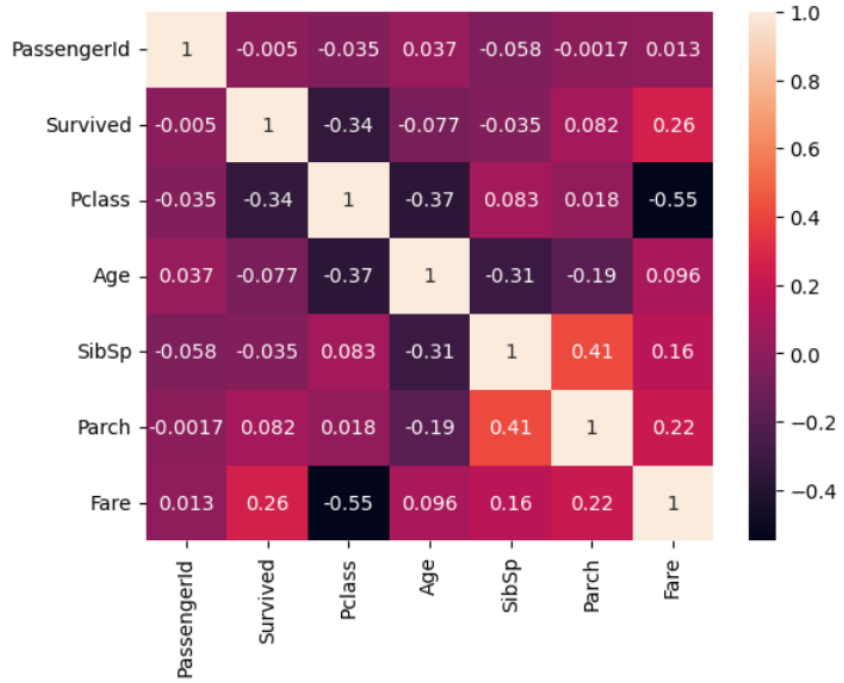
```
<Axes: xlabel='Pclass', ylabel='Age'>
```



By: Aman Kumar Jha
LinkedIn

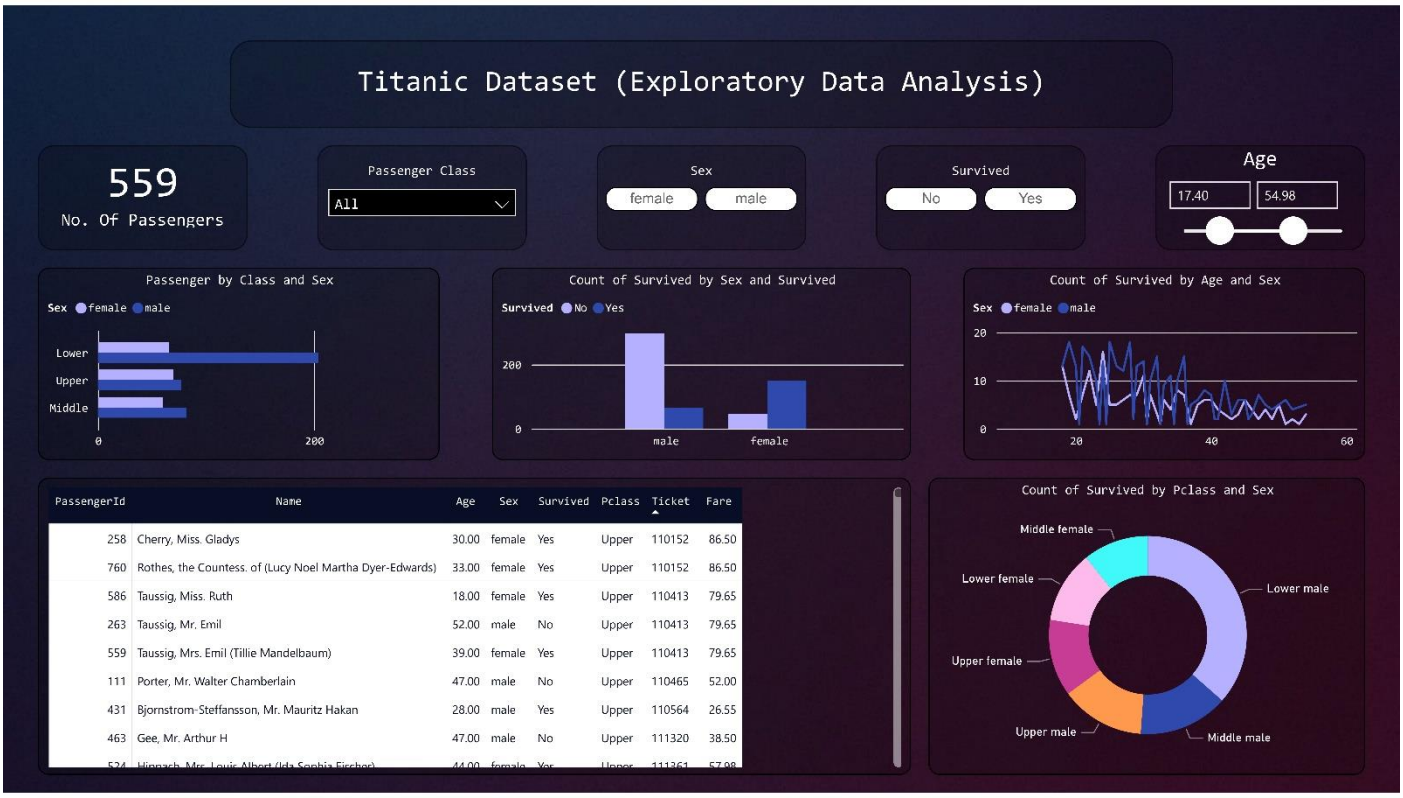
```
1 sns.scatterplot(x='Age', y='Fare', data=data)
2 numerical_data = data.select_dtypes(include=['number'])  # Select numerical columns
3 sns.heatmap(numerical_data.corr(), annot=True)
4
```

<Axes: >



➢ Data Visualization (Using Power BI):

Check out the PowerBI Visualisation Here



By: Aman Kumar Jha
LinkedIn