Data and Artificial Intelligence
Cyber Shujaa Program

Week 3 Assignment
Titanic Exploratory Data Analysis

Student Name: Cherotich Mercy

Student ID: CS-DA01_25091

Introduction

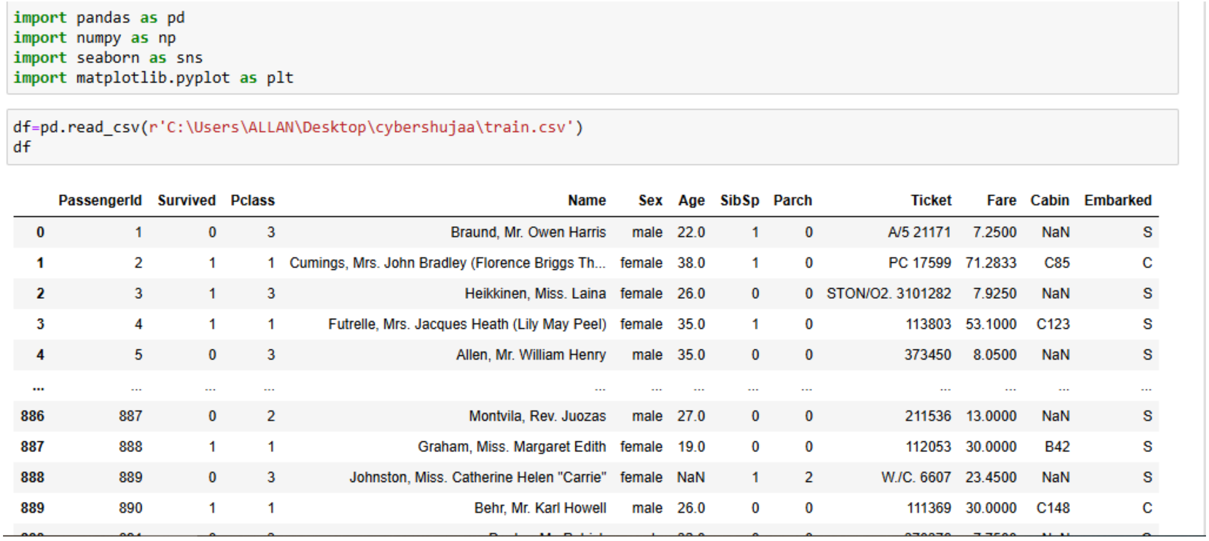
This week's assignment was to do Exploratory Data Analysis. I was not new to the tools we were introduced to. I used Jupiter notebook to write my code and finally uploaded the final work on my GitHub account.

The objectives of the assignment were:

1. Initial Data Exploration
2. Handling Missing Values and Outliers
3. Univariate Analysis
4. Bivariate Analysis
5. Multivariate Analysis
6. Target Variable Analysis

Tasks

Imported pandas,seaborn,numpy and matplotlib. Finally read the csv file.

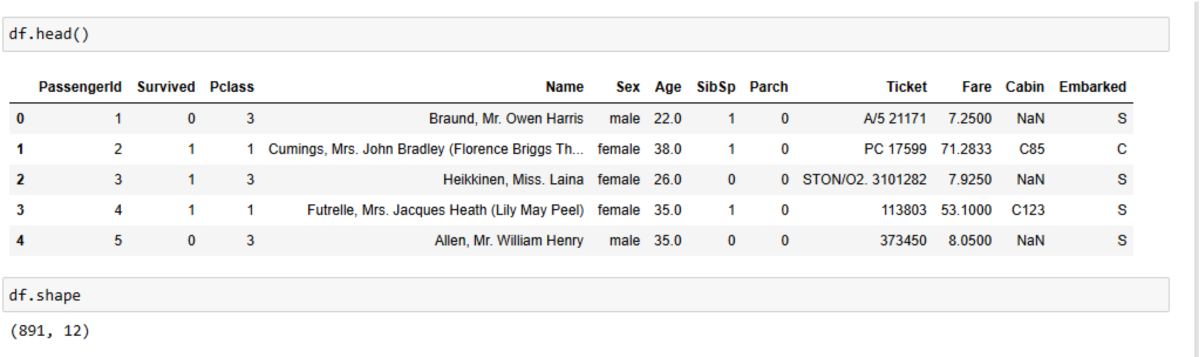


```
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
```

```
df=pd.read_csv(r'C:\Users\ALLAN\Desktop\cybershujaa\train.csv')
df
```

	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
...	...	...	...	...	...	...	...	...	...	...	...	...
886	887	0	2	Montvila, Rev. Juozas	male	27.0	0	0	211536	13.0000	NaN	S
887	888	1	1	Graham, Miss. Margaret Edith	female	19.0	0	0	112053	30.0000	B42	S
888	889	0	3	Johnston, Miss. Catherine Helen "Carrie"	female	NaN	1	2	W./C. 6607	23.4500	NaN	S
889	890	1	1	Behr, Mr. Karl Howell	male	26.0	0	0	111369	30.0000	C148	C

Found the head of the dataset and its shape.



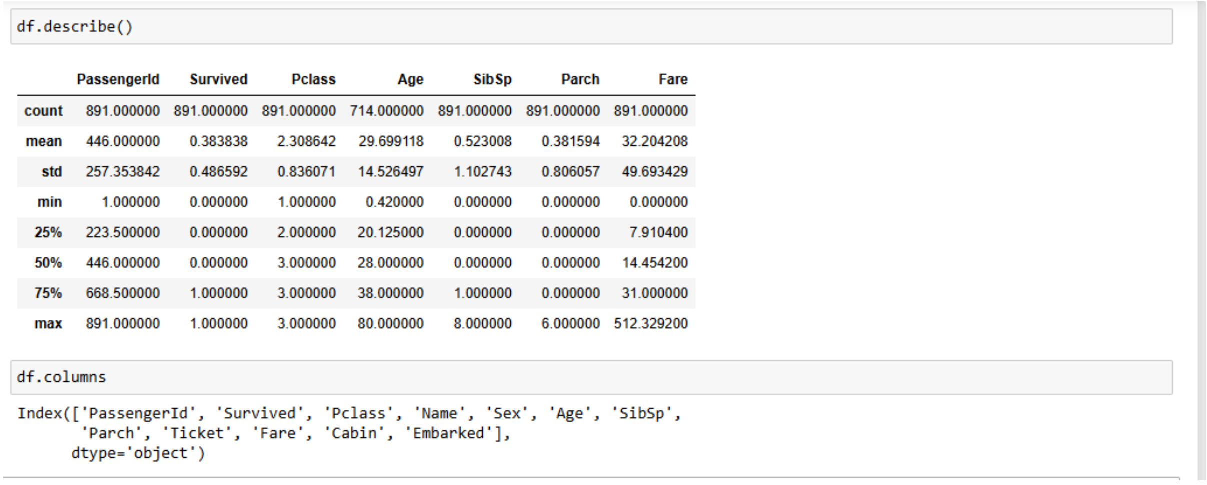
```
df.head()
```

	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

```
df.shape
```

```
(891, 12)
```

Df.describe generated aggregation of the column and finally generated the columns of the dataset.



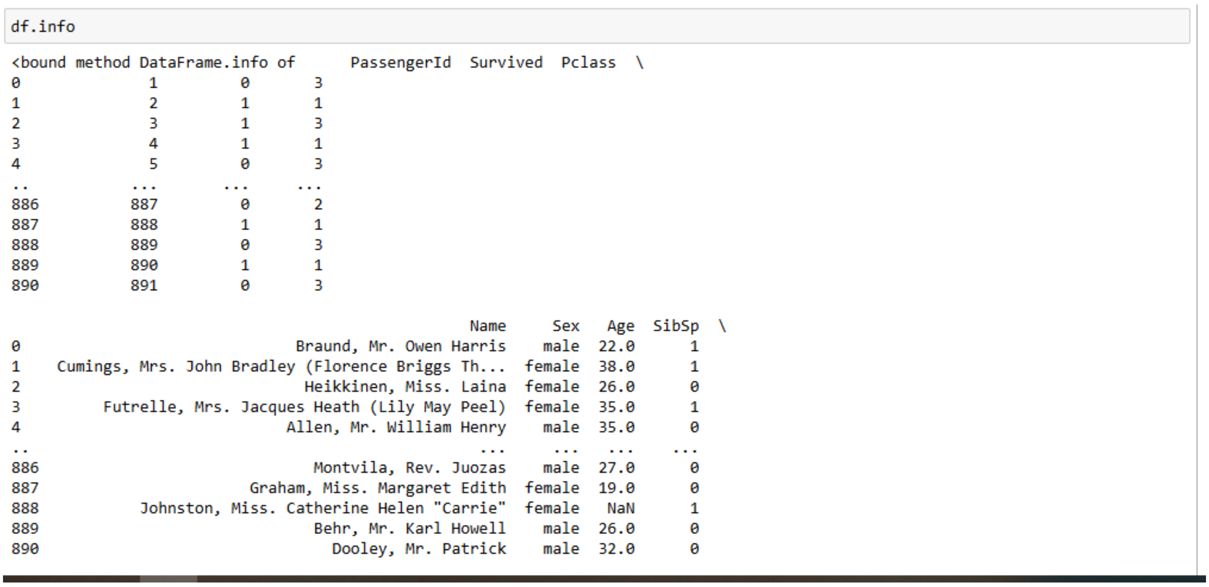
```
df.describe()
```

	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

```
df.columns
```

```
Index(['PassengerId', 'Survived', 'Pclass', 'Name', 'Sex', 'Age', 'SibSp',
       'Parch', 'Ticket', 'Fare', 'Cabin', 'Embarked'],
      dtype='object')
```

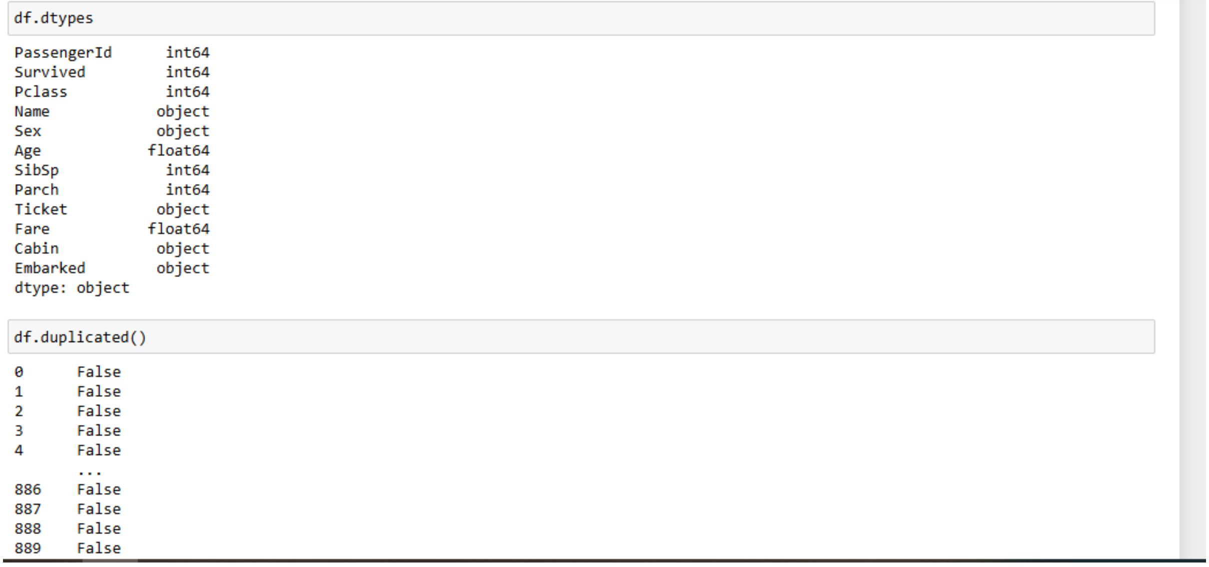
Df.info to find the information of the dataset.



```
df.info
```

```
<bound method DataFrame.info of      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
```

Found the data types of the data and checked if there were any duplicates



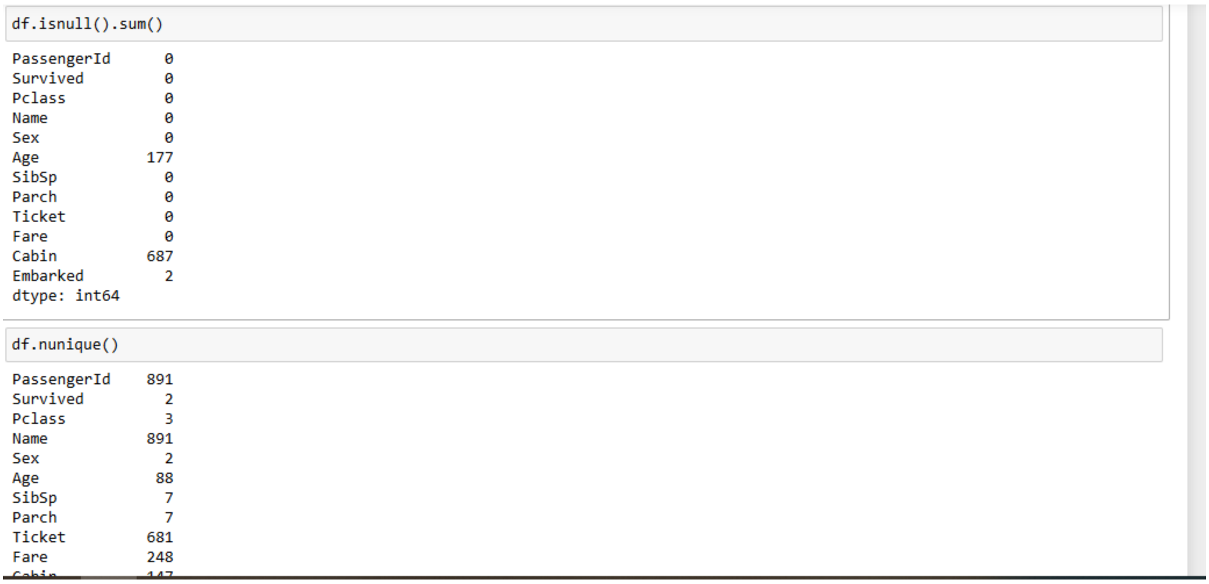
```
df.dtypes
```

```
PassengerId      int64
Survived         int64
Pclass           int64
Name            object
Sex             object
Age            float64
SibSp            int64
Parch            int64
Ticket          object
Fare           float64
Cabin           object
Embarked        object
dtype: object
```

```
df.duplicated()
```

```
0      False
1      False
2      False
3      False
4      False
       ...
886    False
887    False
888    False
889    False
```

Checked the total number of null values in each column and unique values



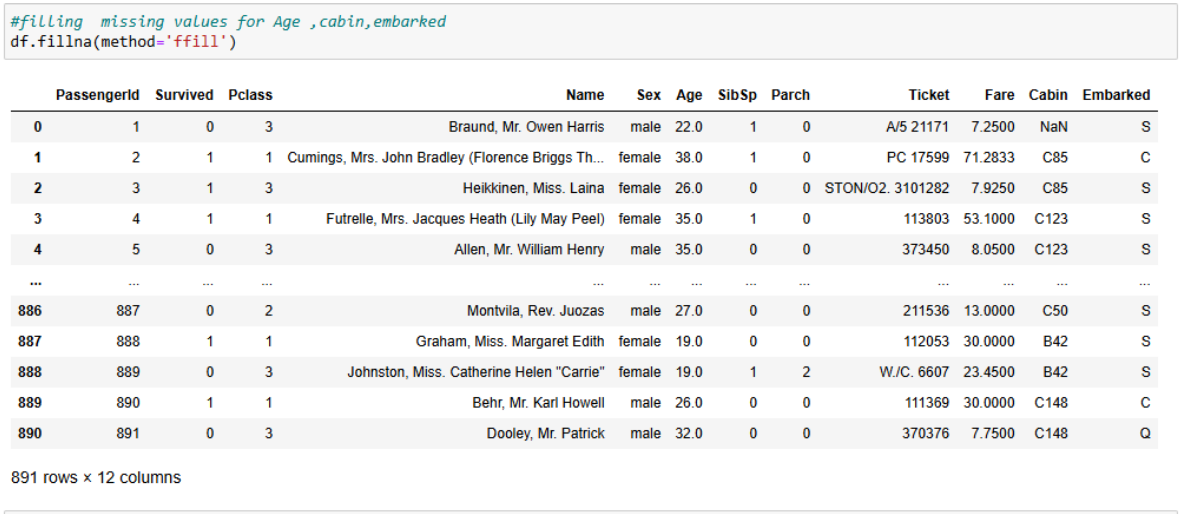
```
df.isnull().sum()
```

```
PassengerId      0
Survived         0
Pclass           0
Name             0
Sex              0
Age            177
SibSp            0
Parch            0
Ticket           0
Fare             0
Cabin          687
Embarked         2
dtype: int64
```

```
df.nunique()
```

```
PassengerId    891
Survived         2
Pclass           3
Name           891
Sex              2
Age             88
SibSp            7
Parch            7
Ticket         681
Fare           248
Cabin          147
```

Used ffill method to fill null values

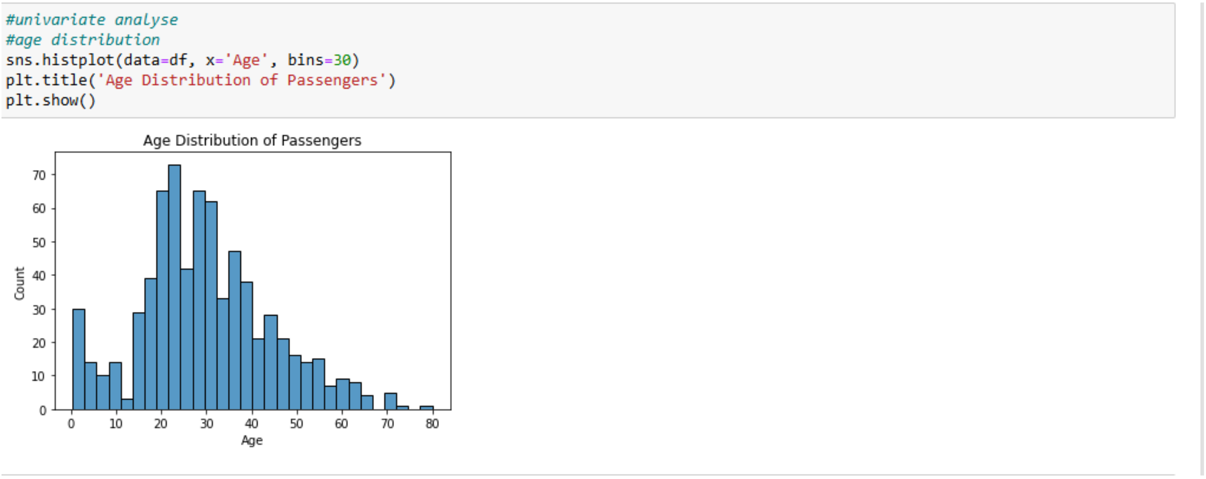


```
#filling  missing values for Age ,cabin,embarked
df.fillna(method='ffill')
```

	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	C85	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	C123	S
...	...	...	...	...	...	...	...	...	...	...	...	...
886	887	0	2	Montvila, Rev. Juozas	male	27.0	0	0	211536	13.0000	C50	S
887	888	1	1	Graham, Miss. Margaret Edith	female	19.0	0	0	112053	30.0000	B42	S
888	889	0	3	Johnston, Miss. Catherine Helen "Carrie"	female	19.0	1	2	W./C. 6607	23.4500	B42	S
889	890	1	1	Behr, Mr. Karl Howell	male	26.0	0	0	111369	30.0000	C148	C
890	891	0	3	Dooley, Mr. Patrick	male	32.0	0	0	370376	7.7500	C148	Q

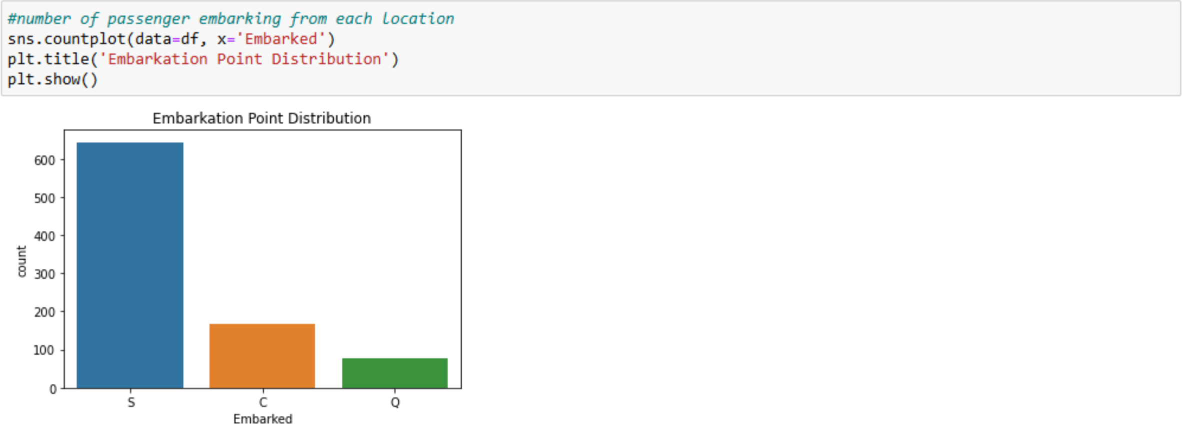
891 rows × 12 columns

Univariate analysis .Found the age distribution.



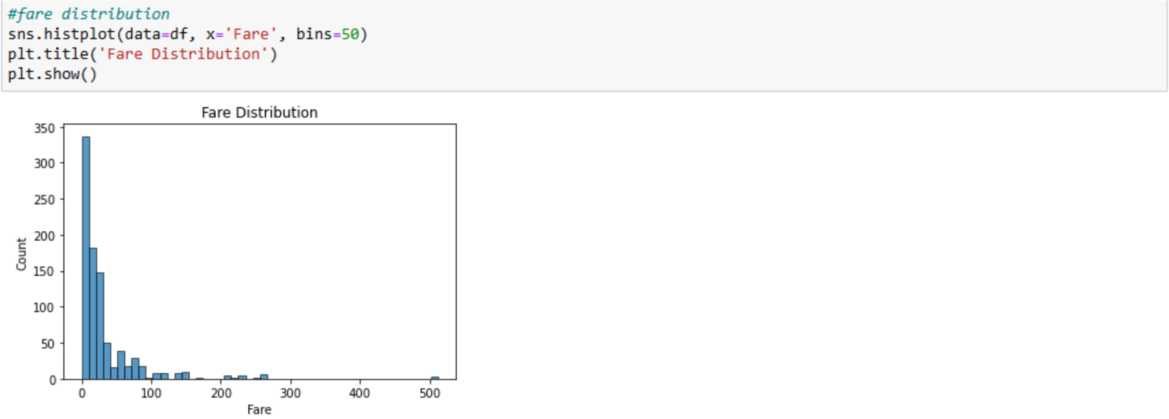
```
#univariate analyse
#age distribution
sns.histplot(data=df, x='Age', bins=30)
plt.title('Age Distribution of Passengers')
plt.show()
```

Number of passenger embarking from each location



```
#number of passenger embarking from each location
sns.countplot(data=df, x='Embarked')
plt.title('Embarkation Point Distribution')
plt.show()
```

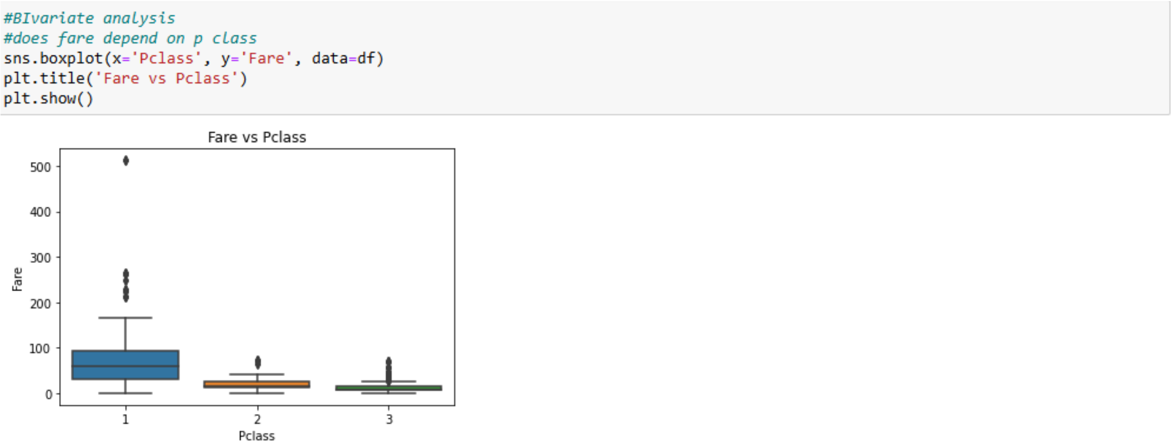
Fare distribution



```
#fare distribution
sns.histplot(data=df, x='Fare', bins=50)
plt.title('Fare Distribution')
plt.show()
```

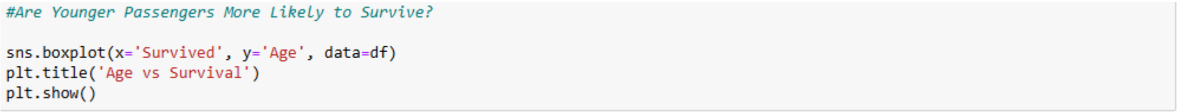
Bivariate analysis

Does fare depend on pclass

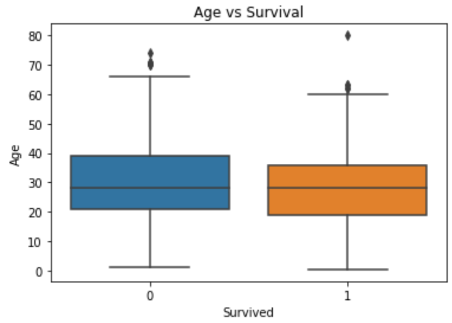


```
#BIvariate analysis
#does fare depend on p class
sns.boxplot(x='Pclass', y='Fare', data=df)
plt.title('Fare vs Pclass')
plt.show()
```

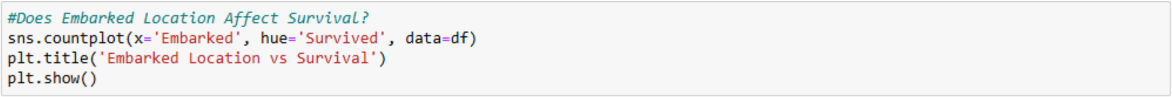
Are younger passengers more likely to survive



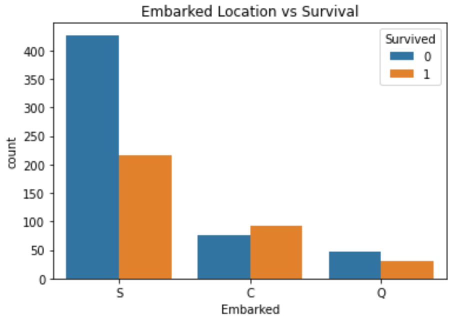
```
#Are Younger Passengers More Likely to Survive?

sns.boxplot(x='Survived', y='Age', data=df)
plt.title('Age vs Survival')
plt.show()
```



Does embarked location affect survive

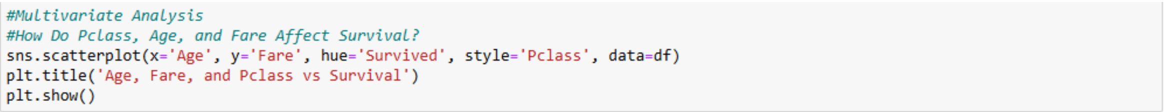


```
#Does Embarked Location Affect Survival?
sns.countplot(x='Embarked', hue='Survived', data=df)
plt.title('Embarked Location vs Survival')
plt.show()
```

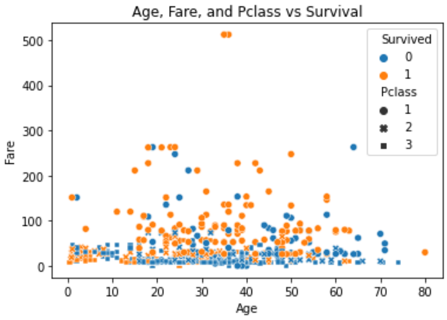


Multivariate analysis

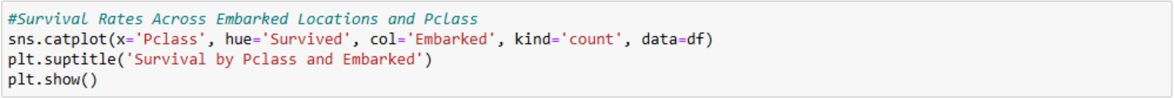
How do pclass,age,fare affect survival



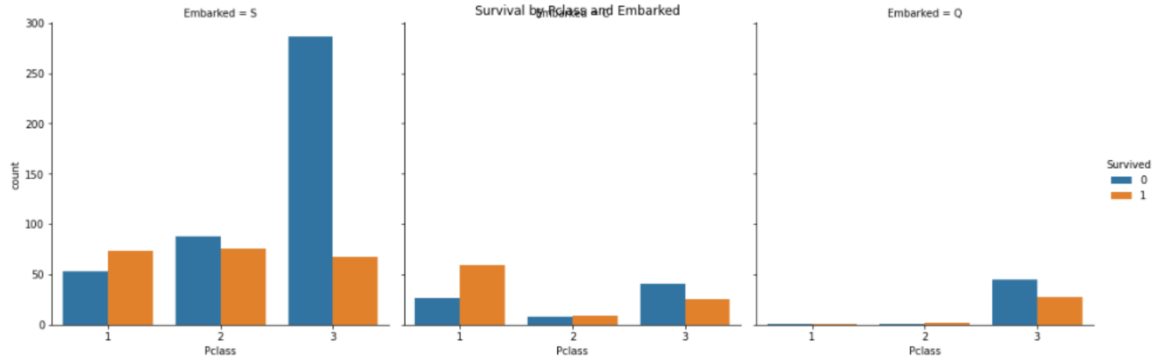
```
#Multivariate Analysis
#How Do Pclass, Age, and Fare Affect Survival?
sns.scatterplot(x='Age', y='Fare', hue='Survived', style='Pclass', data=df)
plt.title('Age, Fare, and Pclass vs Survival')
plt.show()
```



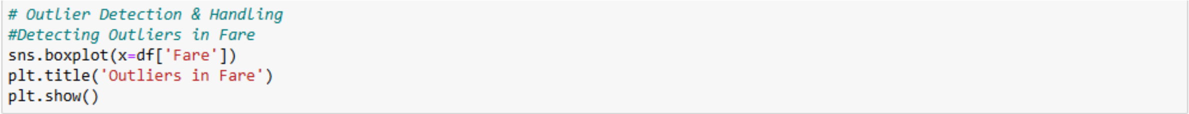
Survival rates across embarked location and pclass



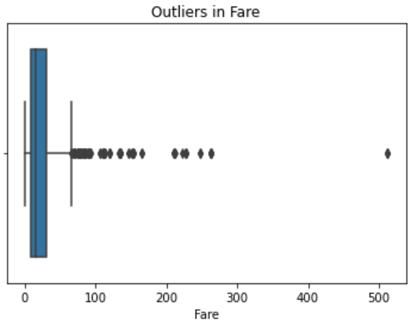
```python
#Survival Rates Across Embarked Locations and Pclass
sns.catplot(x='Pclass', hue='Survived', col='Embarked', kind='count', data=df)
plt.suptitle('Survival by Pclass and Embarked')
plt.show()
```



Detecting outliers in fare

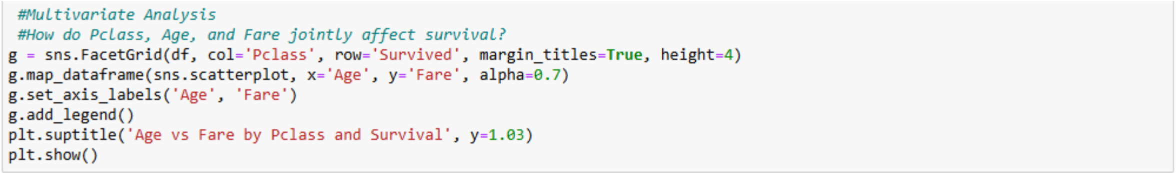


```python
# Outlier Detection & Handling
#Detecting Outliers in Fare
sns.boxplot(x=df['Fare'])
plt.title('Outliers in Fare')
plt.show()
```

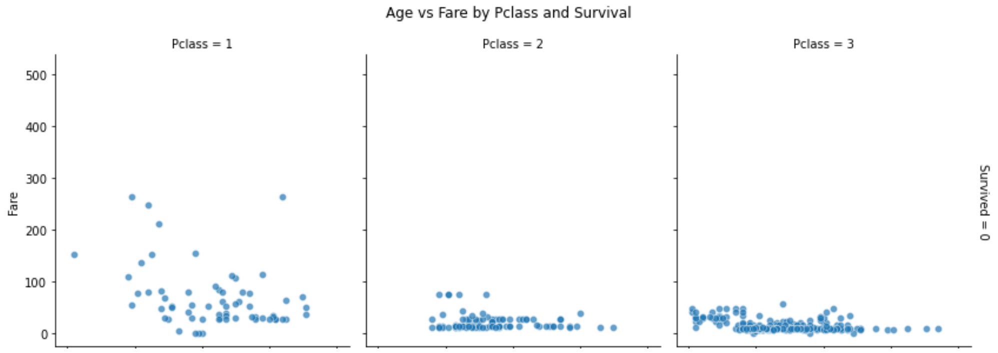


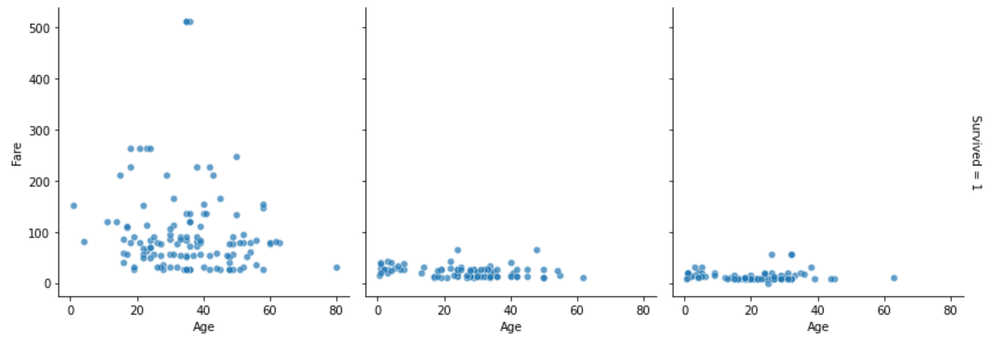
Multivariate analysis

How do pclass ,age and fare jointly affect survival

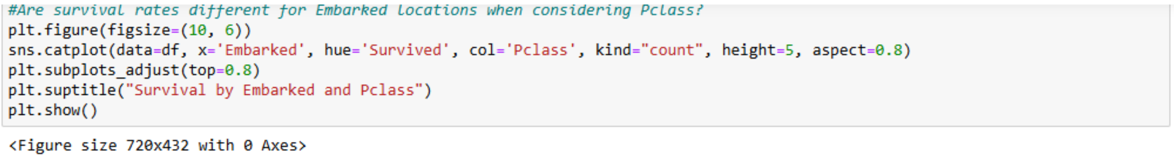


```python
#Multivariate Analysis
#How do Pclass, Age, and Fare jointly affect survival?
g = sns.FacetGrid(df, col='Pclass', row='Survived', margin_titles=True, height=4)
g.map_dataframe(sns.scatterplot, x='Age', y='Fare', alpha=0.7)
g.set_axis_labels('Age', 'Fare')
g.add_legend()
plt.suptitle('Age vs Fare by Pclass and Survival', y=1.03)
plt.show()
```



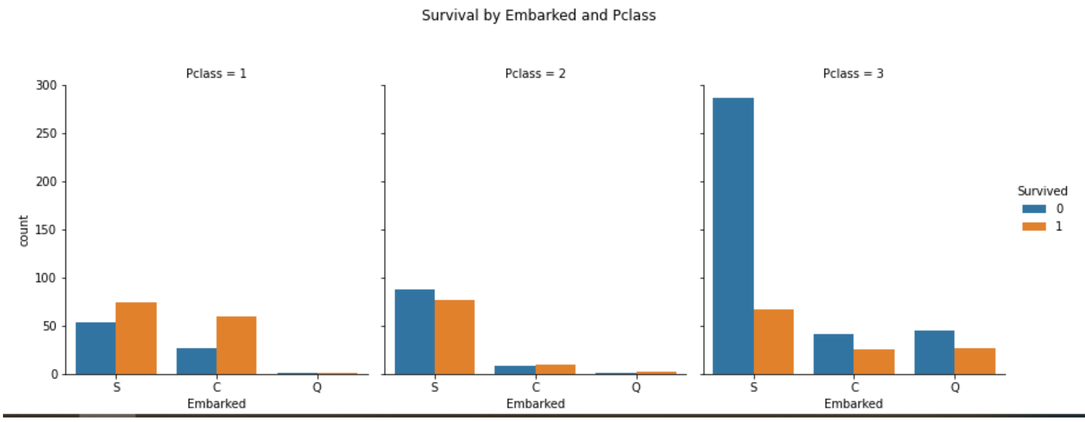


Are survival rates different for embarked when considering pclass.



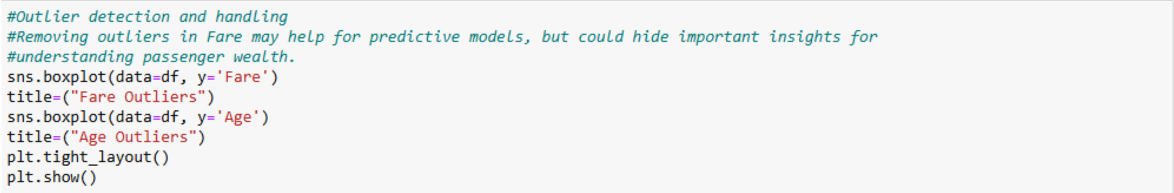
```
#Are survival rates different for Embarked locations when considering Pclass?
plt.figure(figsize=(10, 6))
sns.catplot(data=df, x='Embarked', hue='Survived', col='Pclass', kind="count", height=5, aspect=0.8)
plt.subplots_adjust(top=0.8)
plt.suptitle("Survival by Embarked and Pclass")
plt.show()
```

```
<Figure size 720x432 with 0 Axes>
```

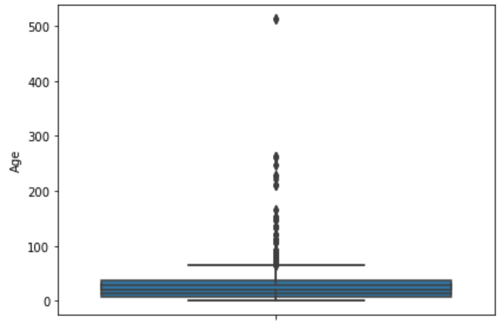


Outlier detection and handling

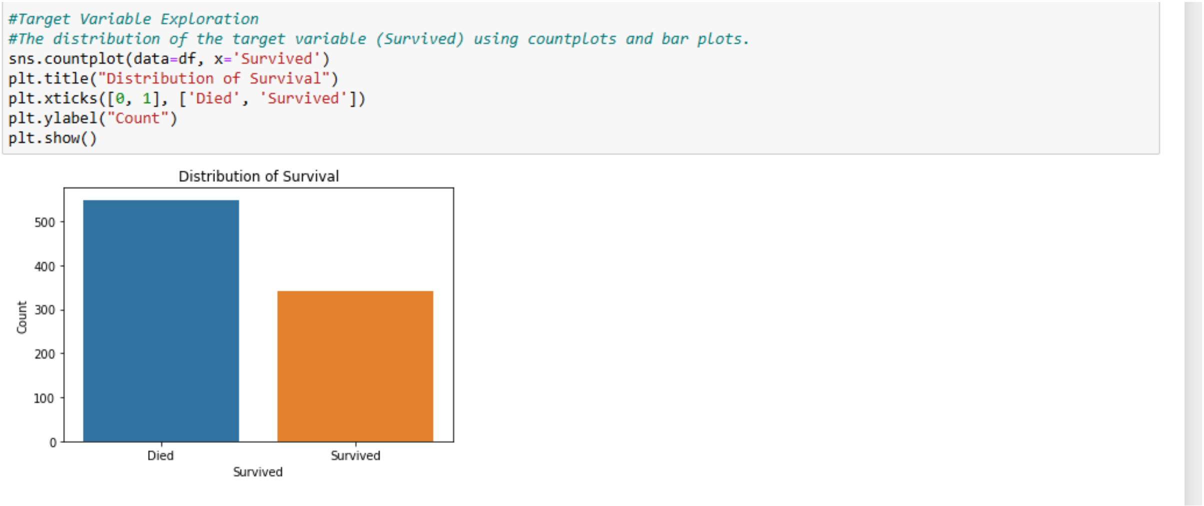
Removing outliers in fare may help for predictive model but could hide important insights for understanding passenger wealth



```
#Outlier detection and handling
#Removing outliers in Fare may help for predictive models, but could hide important insights for
#understanding passenger wealth.
sns.boxplot(data=df, y='Fare')
title=("Fare Outliers")
sns.boxplot(data=df, y='Age')
title=("Age Outliers")
plt.tight_layout()
plt.show()
```

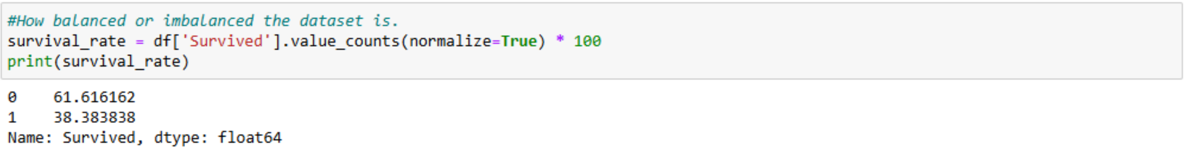


Target variable eploration



```
#Target Variable Exploration
#The distribution of the target variable (Survived) using countplots and bar plots.
sns.countplot(data=df, x='Survived')
plt.title("Distribution of Survival")
plt.xticks([0, 1], ['Died', 'Survived'])
plt.ylabel("Count")
plt.show()
```

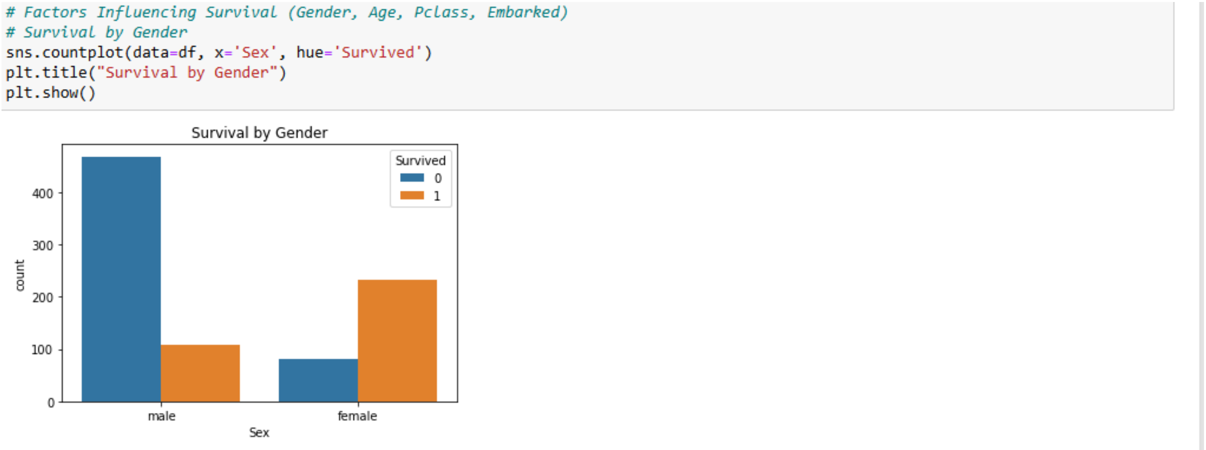
Determining if the data id balanced or not



```
#How balanced or imbalanced the dataset is.
survival_rate = df['Survived'].value_counts(normalize=True) * 100
print(survival_rate)
```

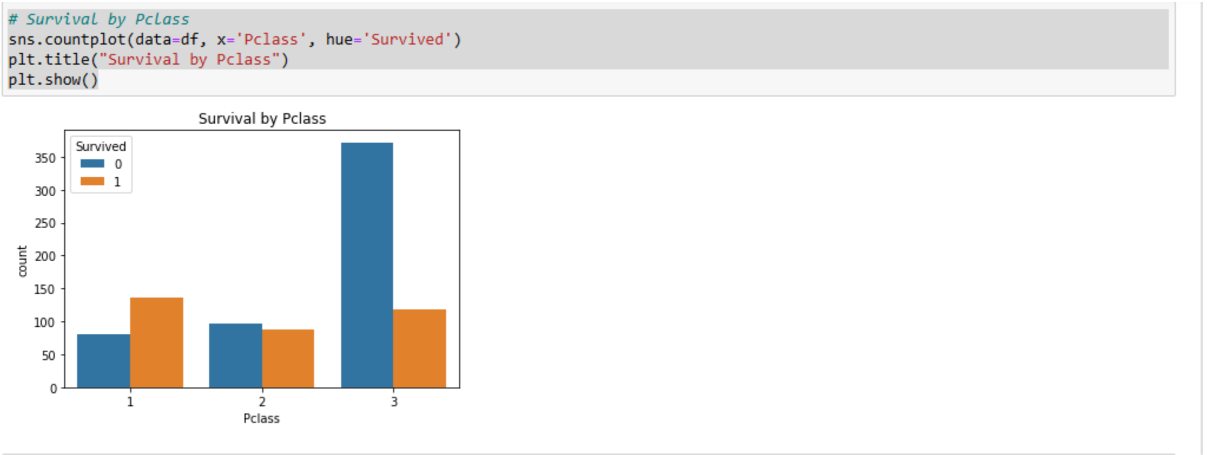
```
0    61.616162
1    38.383838
Name: Survived, dtype: float64
```

Survival by gender



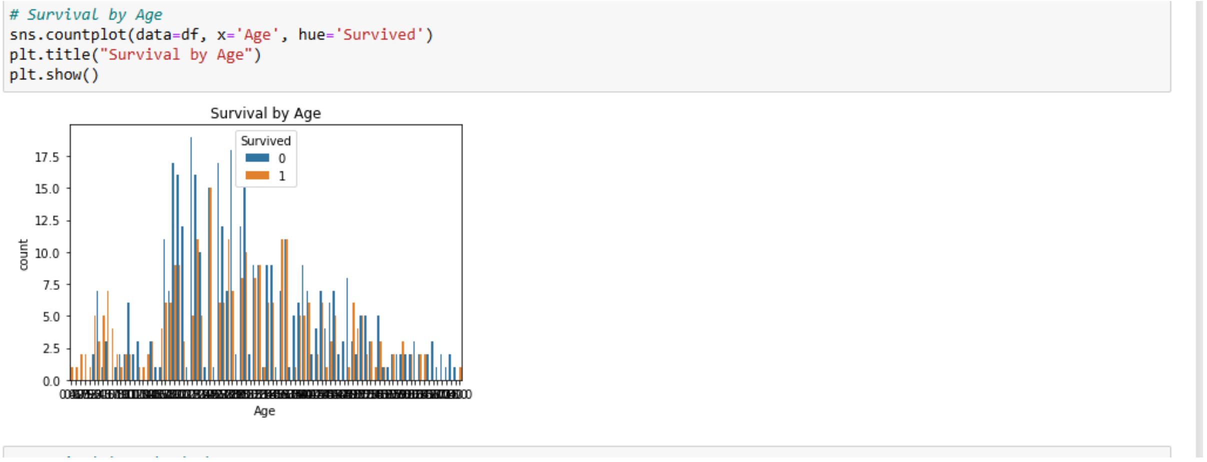
```
# Factors Influencing Survival (Gender, Age, Pclass, Embarked)
# Survival by Gender
sns.countplot(data=df, x='Sex', hue='Survived')
plt.title("Survival by Gender")
plt.show()
```

Survival factor by pclass



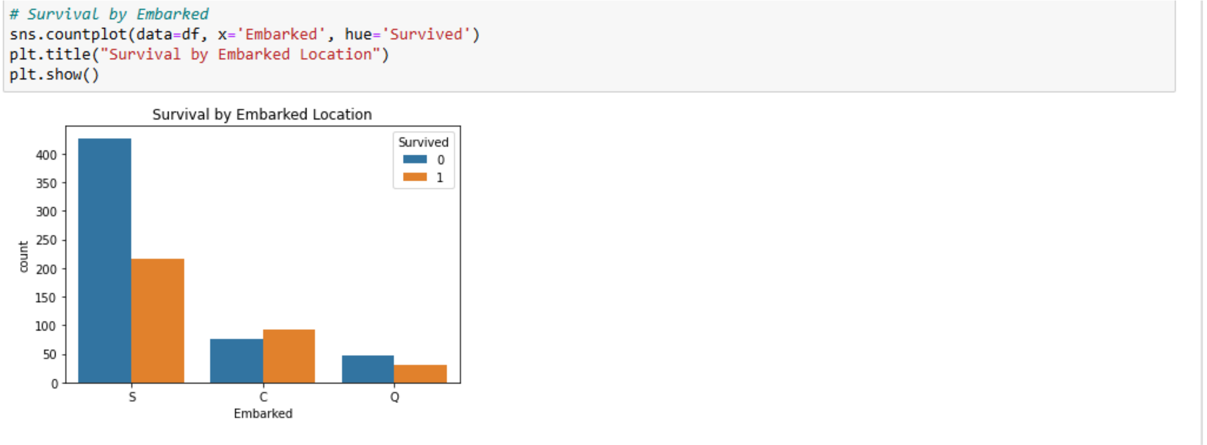
```
# Survival by Pclass
sns.countplot(data=df, x='Pclass', hue='Survived')
plt.title("Survival by Pclass")
plt.show()
```

Survival by age



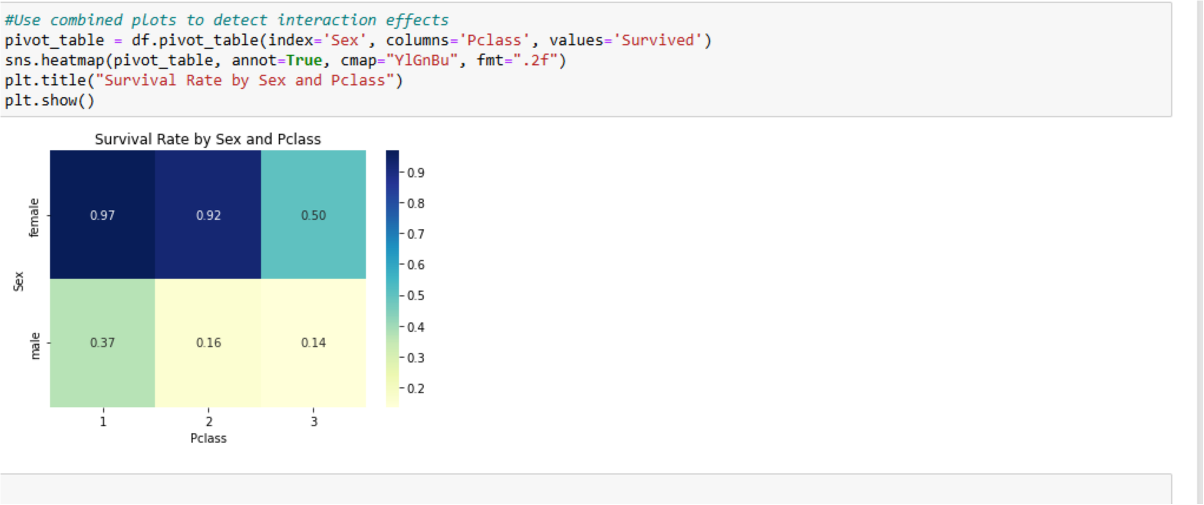
```
# Survival by Age
sns.countplot(data=df, x='Age', hue='Survived')
plt.title("Survival by Age")
plt.show()
```

Survival by embarked



```
# Survival by Embarked
sns.countplot(data=df, x='Embarked', hue='Survived')
plt.title("Survival by Embarked Location")
plt.show()
```

Using combine plots to detect interaction effects.



```
#Use combined plots to detect interaction effects
pivot_table = df.pivot_table(index='Sex', columns='Pclass', values='Survived')
sns.heatmap(pivot_table, annot=True, cmap="YlGnBu", fmt=".2f")
plt.title("Survival Rate by Sex and Pclass")
plt.show()
```

Link: **https://github.com/Chero-dev/Cyber-shujaa-EDA-week-3.git**

Conclusion

This week I gained a lot of insights and knowledge on data exploration build. I have uploaded my work on my GitHub and I look forward to building a portfolio that I can showcase on my CV as I look for jobs in Data and AI.