# EDA PEER TO PEER ASSIGNMENT

## Brief description of the data set and a summary of its attributes

I have data set consisting of 891 data points and 12 columns representing features.

1. **PassengerId:** This is the ID of ever passengers.
2. **Survived:** This feature have values 0 and 1. 0 is for not survived and 1 is for survived.
3. **Pclass:** These are 3 classes of passengers. Class1, Class2 and Class3.
4. **Name:** Name of each passengers.
5. **Sex:** Gender of passengers.
6. **Age:** Age of passengers.
7. **SibSp:** Indication that passenger have siblings and spouse.
8. **Parch:** Whether a passenger is alone or with family.
9. **Ticket:** Ticket no of passenger.
10. **Fare:** Indicating the fare.
11. **Cabin:** Cabin of passengers.
12. **Embarked:** Embarked category.

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

```python
df = pd.read_csv('Titanic.csv')
df.head()
```

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

```python
df.shape
```

```
(891, 12)
```

```python
df.columns
```

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

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

```python
df.describe()
```

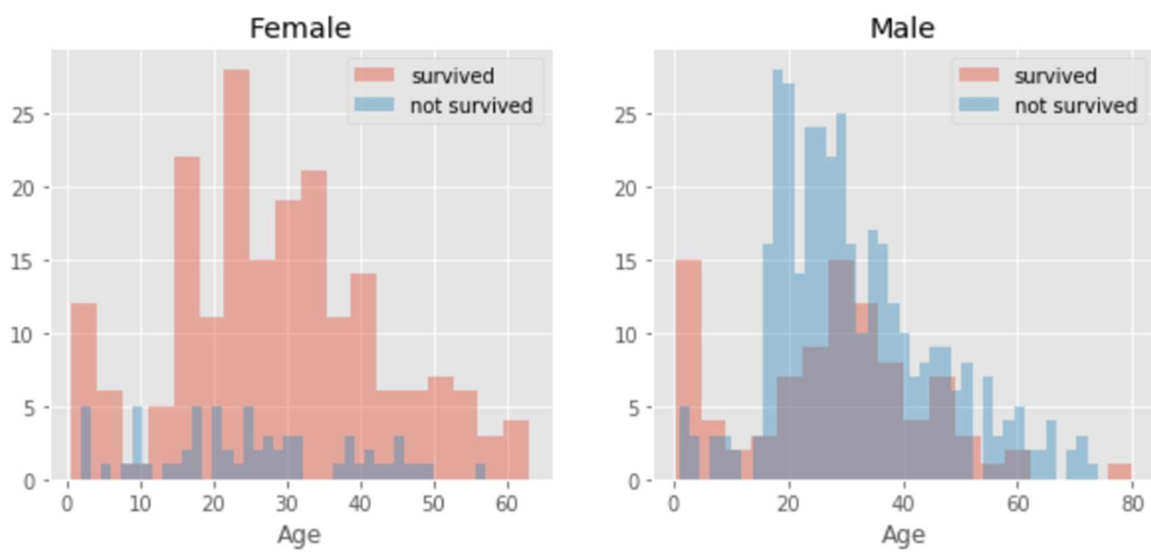
| | PassengerId | Survived | Pclass | Age | SibSp | Parch | Fare |
|---|---|---|---|---|---|---|---|
| **count** | 891.000000 | 891.000000 | 891.000000 | 714.000000 | 891.000000 | 891.000000 | 891.000000 |
| **mean** | 446.000000 | 0.383838 | 2.308642 | 29.699118 | 0.523008 | 0.381594 | 32.204208 |
| **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 |

**Initial plan for data exploration**

Exploratory data analysis (EDA) is important in the sense that by gaining proper insight in our data we can ensure that the feature that we are using for our machine learning model are relevant and will give us correct and interpreted results.

- By using different plot for data visualization i.e (pairplot, heatmap etc) , I found there are some features which are not important for our target variable like cabin, Ticket No etc, So we cam simply remove them from our dataset.
- By using seaborn distplot graph for Sex column , I observed that women has better chance of survival than males.

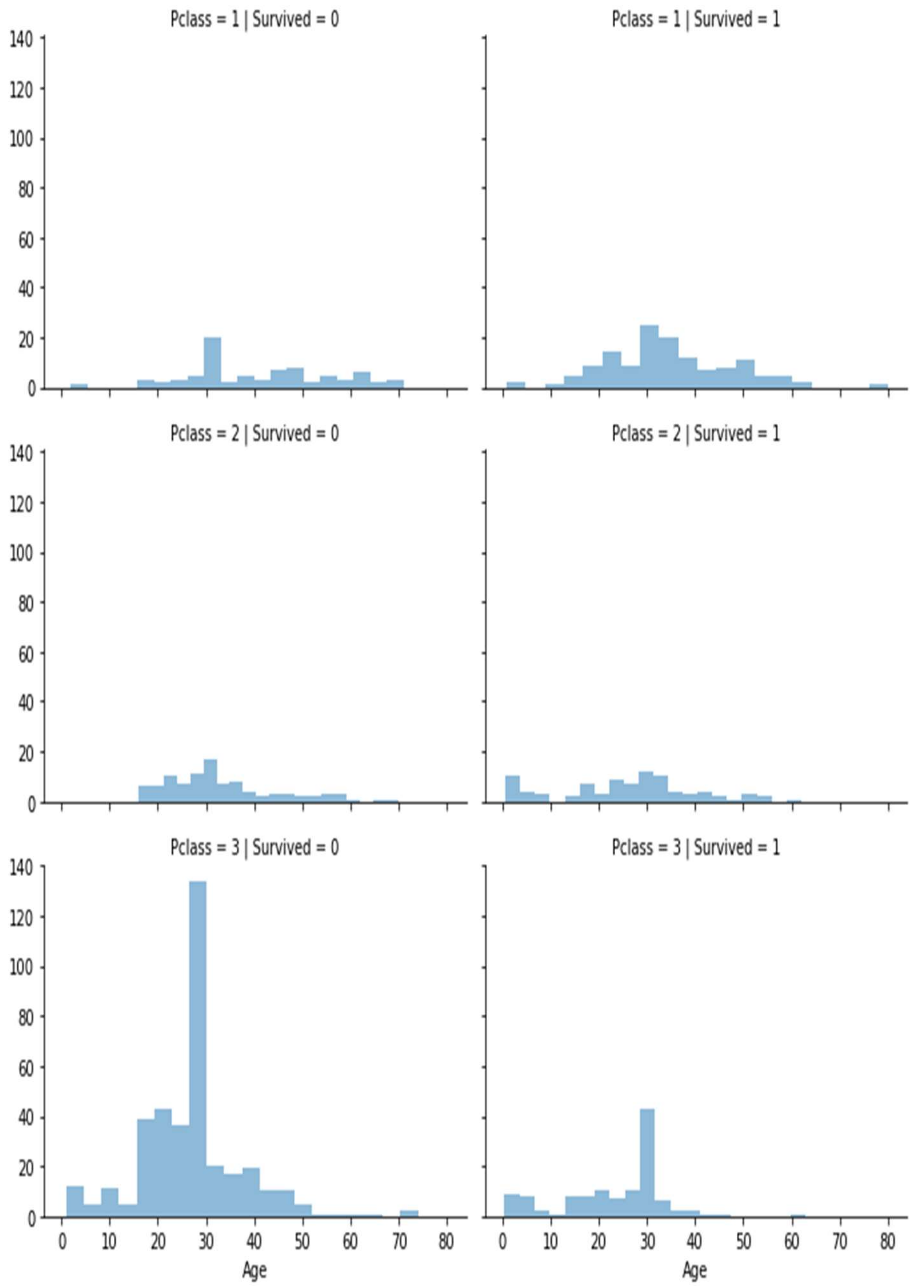
```
women = df[df['Sex']=='female']
men = df[df['Sex']=='male']
fig, axes = plt.subplots(nrows=1,ncols=2,figsize=(12,6))
ax = sns.distplot(women[women['Survived']==1].Age,bins=18,label='survived',ax=axes[0],kde=False)
ax = sns.distplot(women[women['Survived']==0].Age,bins=40,label='not survived',ax=axes[0],kde=False)
ax.legend()
ax.set_title('Female')
ax = sns.distplot(men[men['Survived']==1].Age,bins=18,label='survived',ax=axes[1],kde=False)
ax = sns.distplot(men[men['Survived']==0].Age,bins=40,label='not survived',ax=axes[1],kde=False)
ax.legend()
ax.set_title('Male')
```



- Similarly I found Pclass column in out dataset have relation with target variable which can be observed through graph that Higher class person specially women have greater chance of survival.

```
]: grid = sns.FacetGrid(df, col='Survived', row='Pclass', size=3, aspect=1.6)
   grid.map(plt.hist, 'Age', alpha=.5, bins=20)
   grid.add_legend();
```

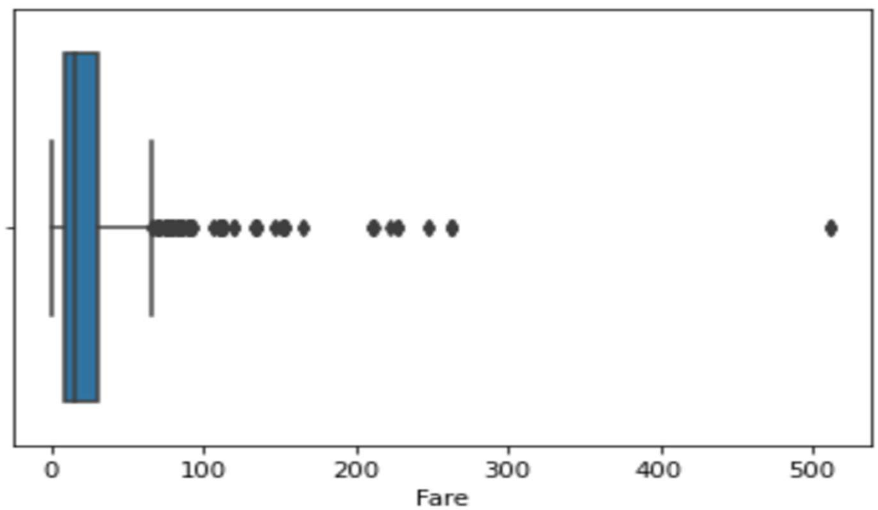
```
C:\Users\Hp\anaconda3\lib\site-packages\seaborn\axisgrid.py:316: UserWarning: The `size` parameter has been
`; please update your code.
  warnings.warn(msg, UserWarning)
```



- More ever there are some column i.e(Age, Fare etc) which are having outliers that can be visualize using boxplot.

```
: sns.boxplot(x='Fare',data=df)
```

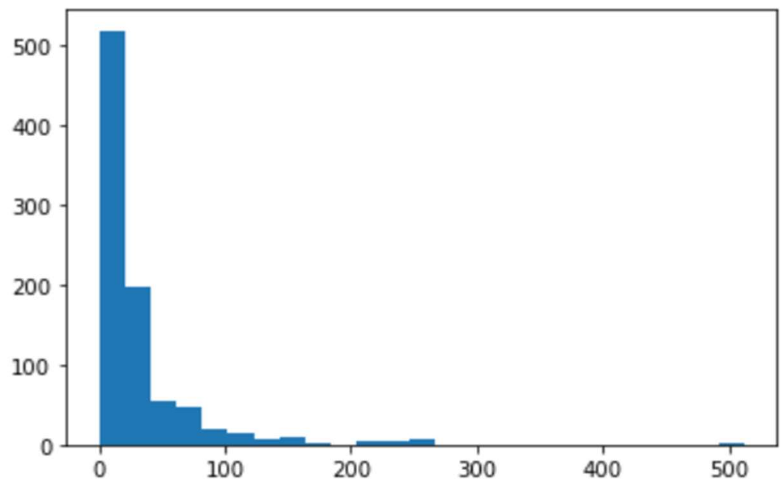
```
: <AxesSubplot:xlabel='Fare'>
```




- Also there is Fare column in out dataset which is right skewed and can be visualize using hist plot.

```
plt.hist(df['Fare'],bins=25)
```

```
(array([519., 197.,  55.,  47.,  20.,  15.,   7.,   9.,   2.,   0.,   5.,
          4.,   8.,   0.,   0.,   0.,   0.,   0.,   0.,   0.,   0.,   0.,
          0.,   0.,   3.]),
 array([  0.       ,  20.493168,  40.986336,  61.479504,  81.972672,
        102.46584 , 122.959008, 143.452176, 163.945344, 184.438512,
        204.93168 , 225.424848, 245.918016, 266.411184, 286.904352,
        307.39752 , 327.890688, 348.383856, 368.877024, 389.370192,
        409.86336 , 430.356528, 450.849696, 471.342864, 491.836032,
        512.3292  ]),
 <BarContainer object of 25 artists>)
```




**Actions taken for data cleaning and feature engineering.**

- First of all I used **.isnull().sum()** pandas builtin functions to check how many entries in every column is having null value.

```
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
```

- In result I found three column i.e Age, Cabin, Embarked having missing values and I used **fillna().mean()** function for Age column to fill missing values with mean of column.

```
df['Age'].fillna(df['Age'].mean(),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         2
dtype: int64
```

- Before checking for null value, I also checked all column data type by using **Info()** function. In result I found there are some columns i.e Cabin and Embarked are having object datatype(**Categorical data**) and we can also Judge these columns are not having any significant impact on our target

```
df.drop(['Cabin','Embarked','SibSp','PassengerId','Ticket','Name'],axis=1,inplace=True)
```

```
df.head()
```

| | Survived | Pclass | Sex | Age | Parch | Fare |
|---|---|---|---|---|---|---|
| **0** | 0 | 3 | male | 22.0 | 0 | 7.2500 |
| **1** | 1 | 1 | female | 38.0 | 0 | 71.2833 |
| **2** | 1 | 3 | female | 26.0 | 0 | 7.9250 |
| **3** | 1 | 1 | female | 35.0 | 0 | 53.1000 |
| **4** | 0 | 3 | male | 35.0 | 0 | 8.0500 |

- Similarly I found Sex column is **categorical** but it have impact on target column like greater chance for survival if gender is female, So converted categorical data into numeric form by using pandas get dummies function.

```
df['Sex'] = pd.get_dummies(df['Sex'],drop_first=True)
```

```
df.head()
```

| | Survived | Pclass | Sex | Age | Parch | Fare |
|---|---|---|---|---|---|---|
| **0** | 0 | 3 | 1 | 22.0 | 0 | 7.2500 |
| **1** | 1 | 1 | 0 | 38.0 | 0 | 71.2833 |
| **2** | 1 | 3 | 0 | 26.0 | 0 | 7.9250 |
| **3** | 1 | 1 | 0 | 35.0 | 0 | 53.1000 |
| **4** | 0 | 3 | 1 | 35.0 | 0 | 8.0500 |

- As we know from EDA that Fare column of our dataset is right skewed, So I converted it into normal distribution using numpy np.log1p function.

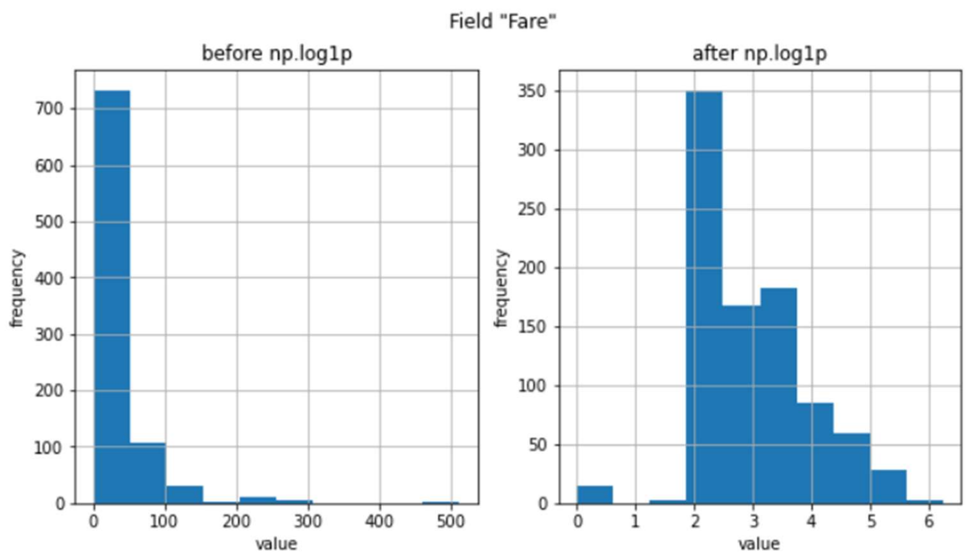
```python
field = "Fare"

# Create two "subplots" and a "figure" using matplotlib
fig, (ax_before, ax_after) = plt.subplots(1, 2, figsize=(10, 5))

# Create a histogram on the "ax_before" subplot
df[field].hist(ax=ax_before)

# Apply a log transformation (numpy syntax) to this column
df[field].apply(np.log1p).hist(ax=ax_after)

# Formatting of titles etc. for each subplot
ax_before.set(title='before np.log1p', ylabel='frequency', xlabel='value')
ax_after.set(title='after np.log1p', ylabel='frequency', xlabel='value')
fig.suptitle('Field "{}"'.format(field));
```




## Key Findings and Insights, which synthesizes the results of Exploratory Data Analysis in an insightful and actionable manner

- My key finding from EDA is that there are many columns that are not important for our target column also known as redundant columns, So we can simply from them from our datasets.
- There is also some outliers and missing values in different columns.
- There are some features that are very important for our target variable like Fare, Sex and Pclass etc.
- But there was also some skewed data in our dataset that, I normalized successfully.

## Formulating at least 3 hypotheses about this data

**Hypothesis One:**

**Null hypothesis:**

If Pclass is high, then person have 50 % chance of survival.

**Alternative hypothesis:**

Pclass is high but person do not have 50% chance of survival.

**Hypothesis Two:**

**Null hypothesis:**

If Sex is female, then there is more than 70% chance of survival.

**Alternative hypothesis:**

If Sex is female, then chance is not more than 70% of survival.

**Hypothesis Three:**

**Null hypothesis:**

If Fair is high, then their more chance of survival.

**Alternative hypothesis:**

High Fair does not affect chance of survival.

**Conducting a formal significance test for one of the hypotheses and discuss the results**

```
from scipy.stats import binom
prob = 1 - binom.cdf(75, 100, 0.70)

print(str(round(prob*100, 1))+"%")
```

```
11.4%
```

As probability value is 11.4% which is greater than 5%, So from significance test I can conclude that my null hypotheses is correct.

**Suggestions for next steps in analyzing this data**

Furthermore we can apply different visualization technique to find which algorithm best matches this data for training our model.

**A paragraph that summarizes the quality of this data set and a request for additional data if needed**

Although my dataset was not much ambiguous, and it was small in size. But if I have larger and clean dataset my model will have more data for training and testing and hence it will be more accurate.