Titanic Dataset EDA and Hypothesis Testing

Brief description of Dataset and summary of attributes:

The titanic dataset describes the survival status of individual passengers on board. I have taken dataset from Kaggle.

The attributes, description and data types are as follows:

Attribute	Description	Data Type
passenger_id	ID of passenger on board	integer
pclass	pclass refers to passenger's socio-economic class: 1 – Upper Class 2 – Middle Class 3 – Lower Class	integer
name	name of passenger	object
sex	Male or Female	object
age	Age of passenger in years. Some fractional values for infants.	float
Sibsp	Number of siblings/Spouses	integer
parch	Number of Parents/Children	integer
ticket	Ticket number	object
fare	Ticket fare	float
cabin	Cabin	Object
embarked	Port of embarkation -> C = Cherbourg Q = Queenstown S = Southampton	object
boat	Lifeboat	object
body	Body identification number	float
home.dest	Home/Destination	object
survived	Survival status 0 – No 1 - Yes	integer

Plan for Data Exploration:

1. First, we will do some Basic Data Exploration and check for presence of null values.
2. After basic data exploration, we will do some exploratory data analysis for checking relationship of some variables with survival status. The aim here is to check if there are any variables have effect on survival status.
3. Then we will do feature engineering and null value treatment.
4. Then we will convert some categorical variables into numeric type.
5. We will do one Hot encoding for some categorical variables.
6. Finally, we will recheck data for cleanliness
7. Then we will go for Hypothesis formulation and testing

Basic Data Exploration:

Shape of data: Data frame has 850 rows with 15 columns.

Data Types: The columns are of data types integer, float and object. Some of the object type variables like name, sex, embarked are converted to categorical variables and dummy encoding is done to convert them to numeric type attributes.

Outliers: By looking at description statistic of data, it is seen that there are no outliers in data.

Null values: Null values are present in some columns which are treated using appropriate methods.

Exploratory Data Analysis:

By plotting distribution of survival status (left plot in below figure), it is clear less than half of passengers survived accident. The number is approximately 40% of all passengers.

In right plot in below figure, the survival status is plotted with respect to gender. It is seen that out of all survived passengers, around twice number of female passengers survived than male passengers.

```
Out[7]: <matplotlib.axes._subplots.AxesSubplot at 0x1af472c44a8>
```

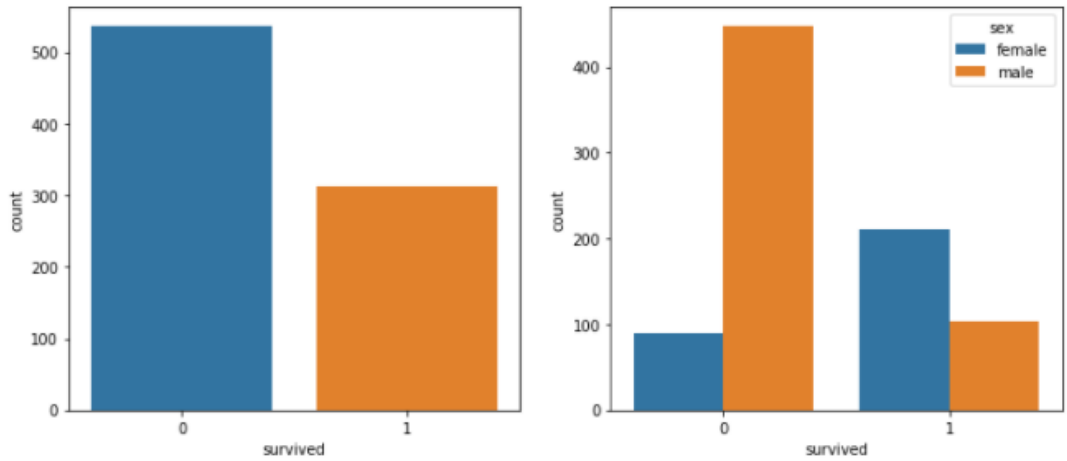
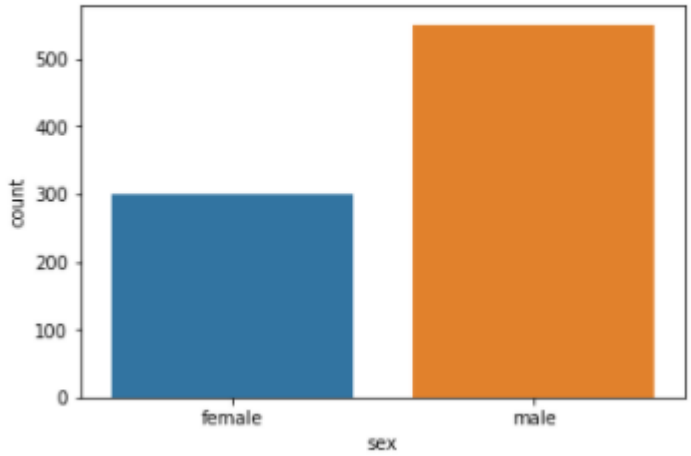


Figure below shows distribution of passengers on boat. Male passengers were more than female passengers. Combining the analysis with above plot, it seems men have given preference to save women.

```
Out[8]: <matplotlib.axes._subplots.AxesSubplot at 0x1af474840f0>
```



We can check for survival rate with respect to age and sex.

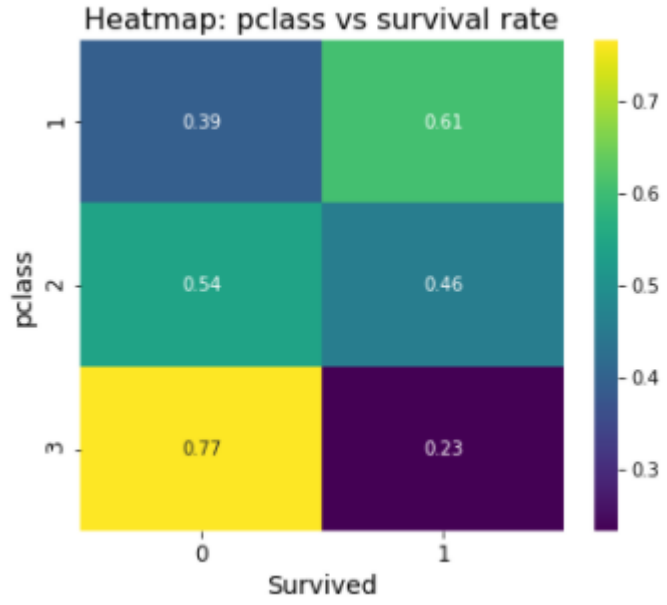
Below violin plot shows that Survival rate is:

Good for children

High for women and

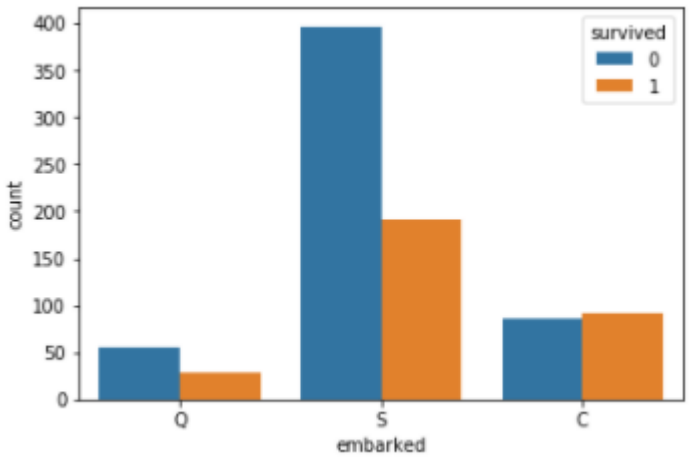
Low for men

Now we will plot a heatmap to show survival status with respect to class. The annotations are out of 1. It is clearly seen that 61 % passengers in class 1 survived and the percentage drops as class decreases.

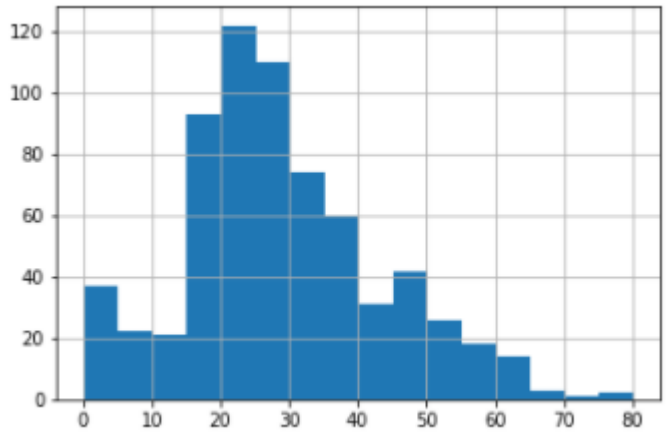


Let us examine relation between embarkation and survival status. Most of the passengers embarked from Southmpton. More than half passengers who boarded from Cherbourg survived.

C = Cherbourg, Q = Queenstown, S = Southampton

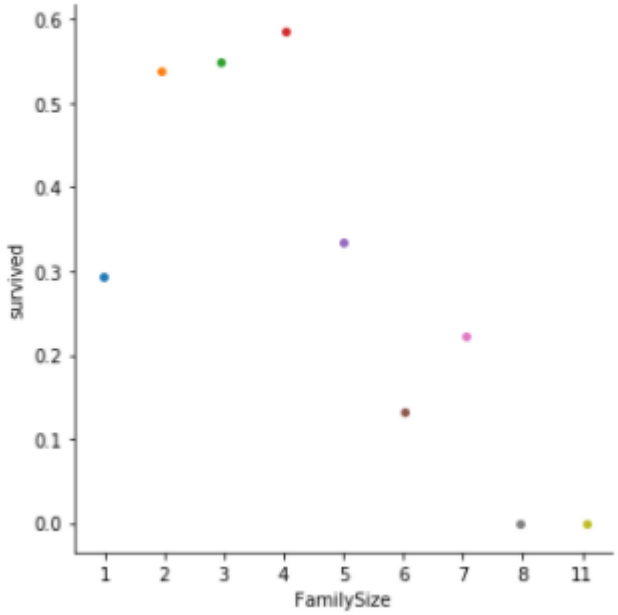


Histogram of age shows that A good number of passengers were from age group 15 to 30. Age feature is right skewed. The skewness will be taken care while grouping age by band.



The size of family and being alone also has effect on survival status. Being alone or having family size greater than 4 had low chances of survival.

```
Out[28]: <seaborn.axisgrid.FacetGrid at 0x1af473a0c18>
```



Summary of exploratory data analysis:

From the exploratory data analysis, it is clear that some features are impacting survival status. Number of male passengers was high than female passengers. Being female had more chances of survival. Violin plot shows that Survival rate is Good for children, High for women and Low for men. Heatmap of class versus survival status shows higher socio-economic class had higher chances of survival. More number of passengers embarked from Southmpton port. The reason for this is not clear from data. Histogram of age feature shows more passengers are from age interval 15 to 30. Family size also impacted survival status. Being alone or having family size greater than 4 had low chances of survival.

Missing value treatment:

The data frame contains some missing values which are filled by taking appropriate actions. Embarked feature was missing a single value which is filled by median of data. The median was Southampton (S) which is also clear from plot of embarked location with respect to survival status.

Age feature had around 174 missing values. As age feature is somewhat normally distributed the missing values are filled from random numbers from normal distribution.

Feature Engineering:

A new feature tile in created from attribute name. Title feature is further converted to 5 categories and then dummy encoded. Embarked feature is converted to numeric using dummy encoding. Age and fare features are first converted in bands and then to numeric by manual ordinal encoding. A new feature FamilySize is created by combining sibsp and parch features.

Sample for Fare feature conversion given below. Features AgeBand and FareBand are dropped later in feature selection step.

```
In [24]: # Fare feature

         # As fare is categorical variable replace missing Fare values with the median of Fare.

         titanic['fare'] = titanic['fare'].fillna(titanic['fare'].median())
```

```
In [25]: # Create new column FareBand. We divide the Fare into 4 category range.

         titanic['FareBand'] = pd.qcut(titanic['fare'], 4)
         print (titanic[['FareBand', 'survived']].groupby(['FareBand'], as_index=False).mean())
```

```
          FareBand  survived
0    (-0.001, 7.896]  0.207048
1    (7.896, 14.108]  0.285000
2   (14.108, 30.924]  0.433333
3  (30.924, 512.329]  0.553991
```

```
In [26]: # Map Fare according to FareBand

         titanic.loc[ titanic['fare'] <= 7.896, 'fare'] = 0
         titanic.loc[(titanic['fare'] > 7.896) & (titanic['fare'] <= 14.108), 'fare'] = 1
         titanic.loc[(titanic['fare'] > 14.108) & (titanic['fare'] <= 30.924), 'fare']   = 2
         titanic.loc[ titanic['fare'] > 30.924, 'fare'] = 3
         titanic['fare'] = titanic['fare'].astype(int)

         titanic.head(3)
```

Out[26]:

	passenger_id	pclass	name	sex	age	sibsp	parch	ticket	fare	cabin	embarked	boat	body	home.dest	survived	title	AgeBand	FareBand
0	1216	3	Smyth, Miss. Julia	female	1	0	0	335432	0	NaN	Q	13	NaN	NaN	1	Miss	(16.0, 32.0]	(-0.001, 7.896]
1	699	3	Cacic, Mr. Luka	male	2	0	0	315089	1	NaN	S	NaN	NaN	Croatia	0	Mr	(32.0, 48.0]	(7.896, 14.108]

Finally feature selection is done by dropping unimportant features.

Hypothesis Formulation:

1. Null Hypothesis: The gender of passenger does not have effect on survival rate

   Alternative Hypothesis: The gender of passenger does hhave effect on survival rate

2. Null Hypothesis: The port of embarkation does not affect survival rate

   Alternative Hypothesis: The port of embarkation does have an effect on survival rate.

3. Null Hypothesis: High class of people do not have effect on survival rate

Alternative Hypothesis: High class of people do have effect on the survival rate

We will conduct significance test for 3rd hypothesis:

* Null Hypothesis: High class of people do not have effect on survival rate

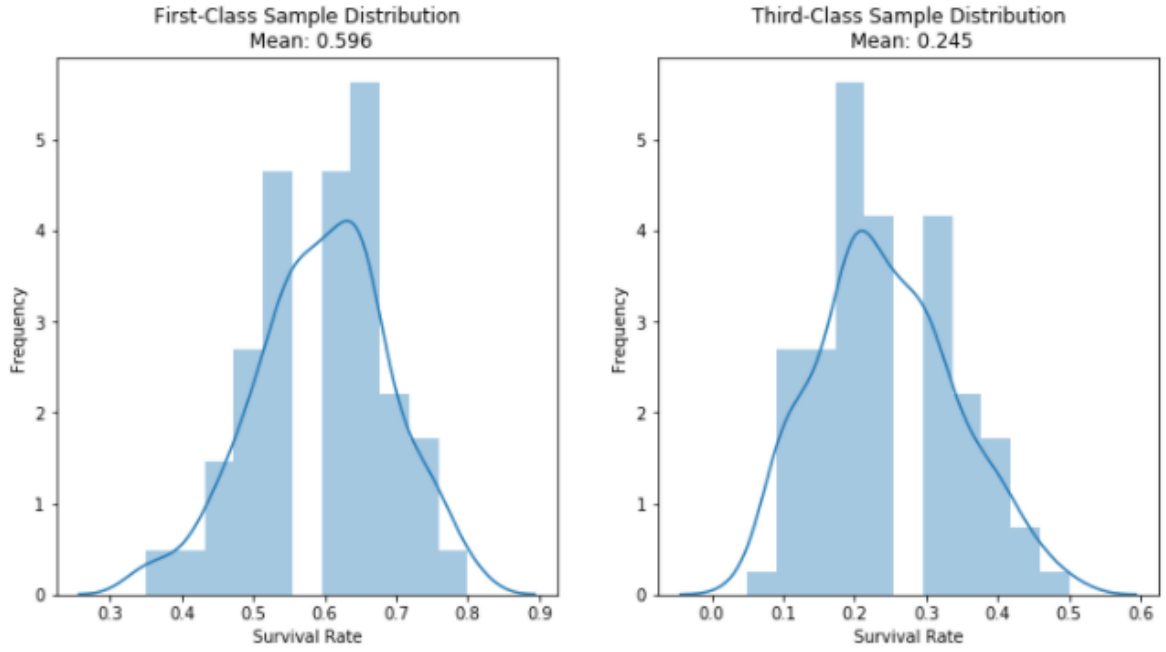
* Alternative Hypothesis: High class of people do have effect on the survival rate

Hypothesis Testing - Significance test for 3rd hypothesis:

We will do Z test for testing hypothesis. For z test we need normally distributed samples hence taken 100 means from the data.

```
In [38]: First_Class_Sample = np.array([np.mean(titanic[titanic["pclass"]==1].sample(20)["survived"].values) for i in range(100)])
         third_Class_Sample = np.array([np.mean(titanic[titanic["pclass"]==3].sample(20)["survived"].values) for i in range(100)])
```

The sample distribution is plotted below with mean:



Then we found the p value which is very much less than standard alpha value of 0.05. Hence, we can reject null hypothesis that High class of people do not have effect on survival rate. This contradicts visual analysis however we dot accept the alternative hypothesis.

In [40]:
```
# According to the sample distributions, the effect of the class is 0.608-0.239 = 0.369

effect = np.mean(First_Class_Sample) - np.mean(third_Class_Sample)
sigma_first = np.std(First_Class_Sample)
sigma_third = np.std(third_Class_Sample)
sigma_difference = np.sqrt((sigma_first**2)/len(First_Class_Sample)  +  (sigma_third**2)/len(third_Class_Sample))
z_score = effect / sigma_difference
```

In [41]:
```
from scipy import stats as st
pvalue = st.norm.sf(abs(z_score))*2
pvalue
```

Out[41]:
```
2.5899130773153576e-150
```

In [42]:
```
print("The p value ({}) is smaller than, alpha = 0.05, hence we can reject the null hypothesis!".format(pvalue))
```

```
The p value (2.5899130773153576e-150) is smaller than, alpha = 0.05, hence we can reject the null hypothesis!
```

Suggestions for next steps in analyzing this data:

Further analysis can be done to check relation of Fare feature with respect to survival rate. There can be other relationships also which can be checked among the features. I have checked only one hypothesis. Different hypothesis can be formulated and checked using either z score or t value.

Summary and Conclusion:

The titanic dataset is formulated by humans hence there is some chance or errors while recording data. There are many null values in some columns. These columns can be checked if those are important for prediction of survival rate. If not, these columns can be dropped.

If there is a different source of data than Kaggle request, you to provide links.

From exploratory data analysis, it is clear that many features had impact on survival rate. That might be due to coincidence however different hypothesis can be formulated and checked. I have done formulation and checking for hypothesis which checks if socio economic class have impact on survival rate and we got sufficient evidence to reject Null hypothesis. This does not mean we can accept alternative hypothesis.