**Data Description**

The size and type of variables in the dataset were as follows.

<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

Dataset had 891 rows with 12 columns.

Survived is a categorical variable with two values 0 & 1.

0 being did not survive and 1 being survived.

Pclass is for the ticket class. 1 = 1st class , 2 = 2nd class , 3 = 3rd class. It’s a proxy for socio-economic status (SES)  
1st = Upper  
2nd = Middle  
3rd = Lower

Embarked variable is for the port of Embarkation .

Its values are defined as

C = Cherbourg

Q = Queenstown

S = Southampton

age: Age is fractional if less than 1. If the age is estimated, is it in the form of xx.5  
  
sibsp: The dataset defines family relations in this way...  
Sibling = brother, sister, stepbrother, stepsister  
Spouse = husband, wife (mistresses and fiancés were ignored)  
  
parch: The dataset defines family relations in this way...  
Parent = mother, father  
Child = daughter, son, stepdaughter, stepson  
Some children travelled only with a nanny, therefore parch=0 for them.

**Missing Data**

Before making any conclusions from the data, it needs to verified if data is infact complete for the total number of observations we see.

With the help of df.isna().any() checked for any missing values.

PassengerId False

Survived False

Pclass False

Name False

Sex False

Age True

SibSp False

Parch False

Ticket False

Fare False

Cabin True

Embarked True

dtype: bool

Dataset had missing values for the variables Age , Cabin & Embarked.

And the count of missing values were as follows.

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

**If survival rate is associated with the class of passengers.**

Chart, bar chart

Description automatically generated

Chart, bar chart

Description automatically generated

Chart, bar chart

Description automatically generated

Performed chi-square test on survived and P class variables .

| **Survived** | **0** | **1** |
| --- | --- | --- |
| **Pclass** |  |  |
| **1** | 80 | 136 |
| **2** | 97 | 87 |
| **3** | 372 | 119 |

P value of the chi-square test: 4.549251711298793e-23

The p value was less then 0.05, we can reject null hypothesis. The null hypothesis was that survived and p class are independent of each other.

Statistically, there is a dependency between survived and p class variable.

**If survival rate is associated with gender :**

**Chart, bar chart

Description automatically generated**

Chart, bar chart

Description automatically generated

Chart, bar chart

Description automatically generated

**If Survival rate is associated with age:**

We have 177 missing values in age out of 891 . Around 20 % of observations do not have age data.

Chart, bar chart

Description automatically generated

Chart, bar chart

Description automatically generated

