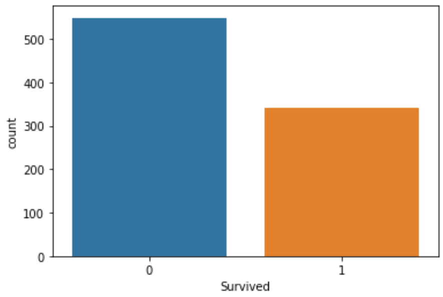
Summary of my Findings (Written by Cyrus Look)

Data Dictionary of Titanic Data Set	
Variable	**Definition, Key / Notes**
Survived (Categorical Variable)	Survival; 0 = No, 1 = Yes
Pclass (Categorical Variable)	Ticket class; 1 = 1st = Upper, 2 = 2nd = Middle, 3 = 3rd = Lower
Sex (Categorical Variable)	Sex
SibSp (Categorical Variable)	Number of siblings / spouses aboard the Titanic
Parch (Categorical Variable)	Number of parents / children aboard the Titanic
Embarked (Categorical Variable)	Port of Embarkation; C = Cherbourg, Q = Queenstown, S = Southampton
Age (Numerical Variable)	Age in years; Age is fractional if less than 1 & in the form of xx.5 if estimated.
Fare (Numerical Variable)	Passenger fare

2-categorical-variable analyses are used for the first 5 hypotheses. To learn about the relationship between two categorical variables, **Chi-square analysis**, involving Chi-square statistic and p-value, would be used. In other words, I would test the **independence** between the two categorical variables. As two-tailed tests are used in hypothesis testing, I would adopt 5% (0.05) significance level. If p-value is smaller than 0.05, null hypothesis (H0) is rejected; If p-value is larger than 0.05, null hypothesis (H0) is accepted.

2-numerical-variable analysis is used for the last hypothesis. To learn about the relationship between two numerical variables, **correlation analysis**, involving scatter plot, would be used. In other words, I would test the **correlation** between the two numerical variables.

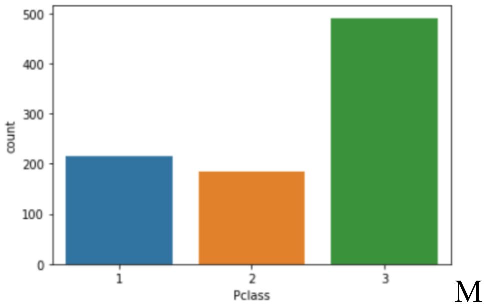
Since the survival rate is involved in the first 5 hypotheses, I would make a countplot for 'Survived'.



There were more victims than survivors in the titanic accident.

1) Determine if the survival rate is associated to the class of passenger.

First, make a countplot for 'Pclass'.



Most passengers were in the 3rd (Lower) class.

Then, set up two opposing statements: null hypothesis (*H*0) and alternative hypothesis (*H*a).

H0: 'Survived' and 'Pclass' are **independent** to each other among all subjects in the population.

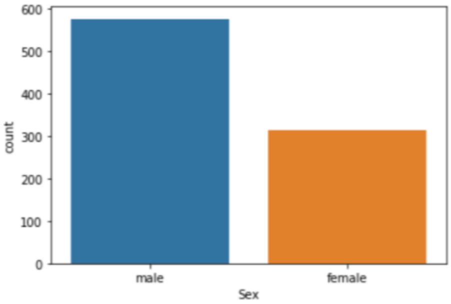
Ha: 'Survived' and 'Pclass' are **not independent** to each other among all subjects in the population.

After running of codes, Chi-square statistic = 102.88898875696056 , p-value = 4.549251711298793e^{-23}

Since p-value is smaller than 0.05, null hypothesis (*H*0) is rejected. I conclude that 'Survived' and 'Pclass' are **not independent** to each other among all subjects in the population. In layman's term, **different classes of passenger have significantly different survival rates**.

2) Determine if the survival rate is associated to the gender.

First, make a countplot for ‘Sex’.



There were more male passengers in the titanic.

Then, set up two opposing statements: null hypothesis (*H*0) and alternative hypothesis (*H*a).

H0: ‘Survived’ and ‘Sex’ are **independent** to each other among all subjects in the population.

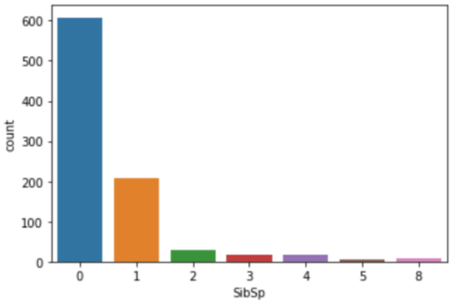
Ha: ‘Survived’ and ‘Sex’ are **not independent** to each other among all subjects in the population.

After running of codes, Chi-square statistic = 260.71702016732104 , p-value = 1.1973570627755645e^{-58}

Since p-value is smaller than 0.05, null hypothesis (*H*0) is rejected. I conclude that ‘Survived’ and ‘Sex’ are **not independent** to each other among all subjects in the population. In layman’s term, **different genders have significantly different survival rates**.

3) Determine if the survival rate is associated to the number of siblings / spouses aboard the Titanic.

First, make a countplot for ‘SibSp’.



Most passengers did not have siblings / spouses aboard the Titanic.

Then, set up two opposing statements: null hypothesis (*H*0) and alternative hypothesis (*H*a).

H0: ‘Survived’ and ‘SibSp’ are **independent** to each other among all subjects in the population.

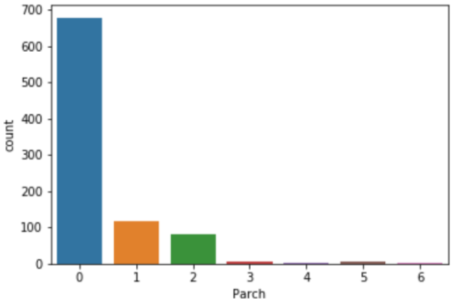
Ha: ‘Survived’ and ‘SibSp’ are **not independent** to each other among all subjects in the population.

After running of codes, Chi-square statistic = 37.2717929152043 , p-value = 1.5585810465902147e^{-06}

Since p-value is smaller than 0.05, null hypothesis (*H*0) is rejected. I conclude that ‘Survived’ and ‘SibSp’ are **not independent** to each other among all subjects in the population. In layman’s term, **different number of siblings / spouses aboard the Titanic have significantly different survival rates**.

4) Determine if the survival rate is associated to the number of parents / children aboard the Titanic.

First, make a countplot for ‘Parch’.



Most passengers did not have parents / children aboard the Titanic.

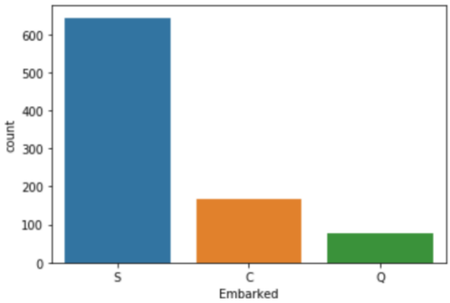
Then, set up two opposing statements: null hypothesis (*H*0) and alternative hypothesis (*H*a).

H0: ‘Survived’ and ‘Parch’ are **independent** to each other among all subjects in the population.

Ha: 'Survived' and 'Parch' are **not independent** to each other among all subjects in the population.
After running of codes, Chi-square statistic = 27.925784060236168 , p-value = 9.703526421039997e^{-05}
Since p-value is smaller than 0.05, null hypothesis ($H0$) is rejected. I conclude that 'Survived' and 'Parch' are **not independent** to each other among all subjects in the population. In layman's term, **different number of parents / children aboard the Titanic have significantly different survival rates**.

5) Determine if the survival rate is associated to the port of embarkation.

First, make a countplot for 'Embarked'.

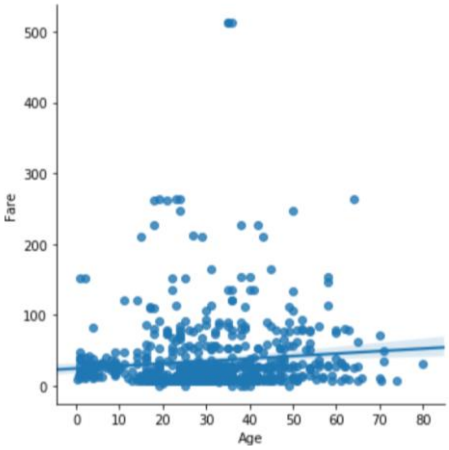


Most passengers embarked at Southampton (S).

Then, set up two opposing statements: null hypothesis ($H0$) and alternative hypothesis (Ha).
H0: 'Survived' and 'Embarked' are **independent** to each other among all subjects in the population.
Ha: 'Survived' and 'Embarked' are **not independent** to each other among all subjects in the population.
After running of codes, Chi-square statistic = 26.48914983923762 , p-value = 1.769922284120912e^{-06}
Since p-value is smaller than 0.05, null hypothesis ($H0$) is rejected. I conclude that 'Survived' and 'Embarked' are **not independent** to each other among all subjects in the population. In layman's term, **different ports of embarkation have significantly different survival rates**.

6) Determine if the age is associated to the passenger fare.

Make a scatter plot where x = 'Age' and y = 'Fare'.



From the scatter plot and the regression line, there is a weak positive linear association with many outliers. We can say that **the age and the passenger fare is positively correlated**. The older the passenger was, the higher passenger fare he / she had.