TITANIC DATASET

1. Introduction:

In this project, we will analyze the Titanic Dataset with the objective evaluate the 3 given hypothesis and conclude which are accepted.

2. Summary:

The process conducted in this dataset was:

2.1 Created two-way frequency tables with the purpose of calculate the percentage of the given variables related to each hypothesis (tables 1, 2, 4, 5 – Graphs and Tables file).
2.2 Plotted the frequency tables to represent the relation between the variables.
2.3 For the Age variable, we had to clean the data by calculating the median (due to the number of outliers in the variable (graph 7 – Graphs and Tables file)) and converting floats to integers.
2.4 Apply the chi-square analysis to verify the independence between the variables of each hypothesis.

3. Results:

3.1 Hypothesis 1: Determine if the survival rate is associated to the class of passenger

3.1.1 In the graph (graph 3 – Graphs and Tables file) and in the proportion table we could see that the people in the 3rd class had more probability of dying (Survived 0) than those in the other passenger class.

3.1.2 Apply the chi-square test to verify the independence of the two variables:

- Null hypothesis: Pclass variable and Survived variable are independence to each other.
- Alternative hypothesis: Pclass variable and Survived variable are not independence to each other.

3.1.3 Through the Chi-square Tesis we could identify that since the chi-squared result is greater than the decision point, we have enough evidence to reject the null hypothesis, which means that the Class of Passenger is associated with the Survival rate.

3.2 Hypothesis 2: Determine if the survival rate is associated to the gender

3.2.1 We could verify in the heatmap (graph 6 – Graphs and Tables file) that females had more than double of probability surviving than males.

3.2.2 Apply the Chi-Square test to verify the independence of the two categorical variables:

- Null hypothesis: Gender variable and Survived variable are independence to each other.
- Alternative hypothesis: Gender variable and Survived variable are not independence to each other

3.2.3 Through the Chi-square Tesis we could identify that since the chi-squared result is greater than the decision point, we have enough evidence to reject the Null Hypothesis, which means that the Gender (Sex) is strongly associated with the Survival rate.

3.3 Hypothesis 3: Determine the survival rate is associated to the age

3.3.1 We could verify in the graph (graph 8 – Graphs and Tables file) that people in their 20s had more possibility to survive than people in other ranges.

3.3.2 Apply the Chi-Square test to verify the independence of the two categorical variables:

- Null hypothesis: Age variable and Survived variable are independence to each other.
- Alternative hypothesis: Age variable and Survived variable are not independence to each other.

3.3.3 Through the Chi-square Tesis we could identify that since the chi-squared result is greater than the decision point, we have enough evidence to reject the Null Hypothesis, which means that Age is strongly associated with the Survival rate.

4. Conclusions:

4.1 At the end of this analysis, we could conclude that all hypotheses were accepted which means, these results were supported not just by the proportion of each variable but also by the chi-square test performed.

4.2 It is always important to validate the quality of the data before performing any analysis with the objective to have the most accurate results.