**CPSC4800 – W01**

**Assignment 3 Part 2**

This report summarizes the result of an exploratory data analysis on the Titanic dataset from Kaggle. The dataset contains 891 observations and 12 variables. Below is the information on the variables of the dataset.

<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

This data analysis will examine and decide on three hypothesis:

1. Determine if the survival rate is associated to the class of passenger
2. Determine if the survival rate is associated to the gender of passenger
3. Determine if the survival rate is associated to the age of passenger
4. **Survival rate vs the Class of Passenger**

Chi-Square test was performed :

H0 : Survival rate and the class of passenger are independent of each other

Ha : Survival rate and the class of passenger are not independent of each other

|  |  |
| --- | --- |
|  | Pclass 1 2 3 Total  Survived  0 80 97 372 549  1 136 87 119 342  Total 216 184 491 891  **chi-square value is 102.88898875696056**  df is 2 and dp is 5.991  **Dependent (reject H0)**  p value is 4.549251711298793e-23  **Dependent (reject H0)** |
| **Survival rate is associated to the Class of Passenger** | |

1. **Survival rate vs the Gender of Passenger**

Chi-Square test was performed :

H0 : Survival rate and the gender of passenger are independent of each other

Ha : Survival rate and the gender of passenger are not independent of each other

|  |  |
| --- | --- |
|  | Sex female male Total  Survived  0 81 468 549  1 233 109 342  Total 314 577 891  chi-square value is **260.71702016732104**  df is 1 and dp is 3.841  **Dependent (reject H0)**  pvalue is 1.1973570627755645e-58  **Dependent (reject H0)** |
| **Survival rate is associated to the Gender of Passenger** | |

1. **Survival rate vs the Age of Passenger**

T-test was performed :

H0 – the mean of age of survivors = the mean of age of the dead

Ha – the mean of age of survivors not = the mean of age of the dead

Before performing the test, the observations with null age value were dropped, resulting in 714 observations only.

|  |  |
| --- | --- |
| Chart, histogram  Description automatically generated |  |
|  | |
| Ttest\_indResult(statistic=2.06668694625381, pvalue=0.03912465401348249)  **p-value < 0.05, reject H0** | |
| **Survival rate is associated to the Age of Passenger** | |