# Data Analysis Report

*Author: Geraldi Gunawan*

## Findings from DataVisual.py:

1) It was observed that young passengers whose age relatively between 0 to 10 years old has higher survival rate.

2) It was observed that female has higher survival rate and lesser casualties than male. Hence, proofing this statement by doing Pearson chi square test.

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Survived** | **Casualties** |  |
| **Female** | 233 | 81 | 314 |
| **Male** | 109 | 468 | 577 |
|  | 342 | 549 | 891 |

Degree of freedom = 1

This is also supported by chi squared testing which is 264.16 with p-value of 4.355x10^-58. The chi2 value is significantly bigger than the critical value in X^2 table for degree of freedom of 1. The p-value is much smaller than 0.05 which indicates that we have to reject null hypothesis. As a result, we can conclude that there is significant difference of survival/casualties rate between gender.

3) It was observed that that passenger with ticket class no 1 and 2 has higher survival rate compared to ticket class no 3. Hence, proofing this statement by doing Pearson chi square test.

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Survived** | **Casualties** |  |
| **Class 1** | 136 | 80 | 216 |
| **Class 2** | 87 | 97 | 184 |
| **Class 3** | 119 | 372 | 491 |
|  | 342 | 549 | 891 |

Degree of freedom = 2

This is supported by chi squared testing which is 101.8 with p-value of 6.15x10^-22. The chi2 value is significantly bigger than the critical value in X^2 table for degree of freedom of 2. The p-value is much smaller than 0.05, which means we have to reject null hypothesis. As a result, we can conclude that there is significant difference of survival/casualties rate between ticket classes.

4) Cabin information are mostly retrieved for passenger with Pclass = 1 (176 out of 204), rest of the people’s cabin info are not recorded.

5) It was observed that there are more survivors compared to casualties for passengers who embarked at Cherbourg.

Observed

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Survived** | **Casualties** |  |
| **S (Souththampton)** | 219 | 427 | 646 |
| **C (Cherbourg)** | 93 | 75 | 168 |
| **Q (Queenstown)** | 30 | 47 | 77 |
|  | 342 | 549 | 891 |

Degree of freedom: 2

Expected

Souththampton: 72.5%

Cherbourg: 18.85%

Queenstown: 8.64%

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Survived** | **Casualties** |  |
| **S (Souththampton)** | 248 | 398 | 646 |
| **C (Cherbourg)** | 65 | 103 | 168 |
| **Q (Queenstown)** | 29 | 48 | 77 |
|  | 342 | 549 | 891 |

This is supported by chi squared testing which is 25.2 with p-value of 1.38^-5. The chi2 value is significantly bigger than the critical value in X^2 table for degree of freedom of 2. The p-value is much smaller than 0.05, which means we have to reject null hypothesis. As a result, we can conclude that there is significant difference of survival/casualties rate between embarkation port.

6) It was visibly enough observed that passengers with siblings has higher survival rate compared to passengers without siblings. Hence, proofing this statement by doing Pearson chi square test.

Observed

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Survived** | **Casualties** |  |
| **With Siblings** | 132 | 151 | 283 |
| **Without Siblings** | 210 | 398 | 608 |
|  | 342 | 549 | 891 |

Expected

31.7% with siblings

68.3% without siblings

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Survived** | **Casualties** |  |
| **With Siblings** | 109 | 174 | 283 |
| **Without Siblings** | 233 | 375 | 608 |
|  | 342 | 549 | 891 |

df= 1

This is supported by chi squared testing which is 11.57 with p-value of 0.00306. The chi2 value is significantly bigger than the critical value in X^2 table for degree of freedom of 1. The p-value is much smaller than 0.05, which means we have to reject null hypothesis. As a result, we can conclude that there is indeed a difference in terms of survival rate based on whether passenger has siblings or not.

## Prediction criteria:

Based on the info retrieved, the following information is what was believed to be good predictors:

1. It is statistically proven that female has higher survival rate and lower casualties rate compared to male.
2. It is statistically proven that Class 1 and Class 2 passengers have higher survival rate than passengers in class 3.
3. It is visually observed that passengers whose ages are in between 0 - 10 have higher survival rate.
4. It is statistically proven that passengers that have siblings have higher chance of survival rate compared to passengers who do not have any siblings.

5) it is statistically proven that passengers embarkation port plays a factor in survival rate.