Bivariate Analysis of Titanic Dataset

Rajput, Kaustubh Rajendra

University of the Cumberlands

MSDS-530-A01: Fundamentals of Data Science

Dr. Afshin Zarenejad

January 19, 2025

Bivariate Analysis of the Titanic Dataset

The Titanic dataset is widely used to analyze factors influencing survival outcomes during the Titanic disaster. This study examines the relationships between qualitative and quantitative variables, specifically *Sex* and *Survived* (qualitative) and *Age* and *Fare* (quantitative). Understanding these relationships helps to reveal insights into demographic and socioeconomic factors that may have impacted survival rates.

Qualitative variables and the association between them

For the qualitative variables, we selected *Sex* and *Survived* for analysis. Both are categorical variables, where *Sex* refers to the gender of the passengers, and *Survived* indicates whether a passenger survived the disaster. The study of these variables can shed light on how gender may have influenced survival, reflecting societal norms and the prioritization of women and children during rescue operations.

A Chi-square test was performed to assess the relationship between *Sex* and *Survived*. The test results are as follows:

Chi-square Statistic: 260.717

p-value: 1.197×10^{-58}

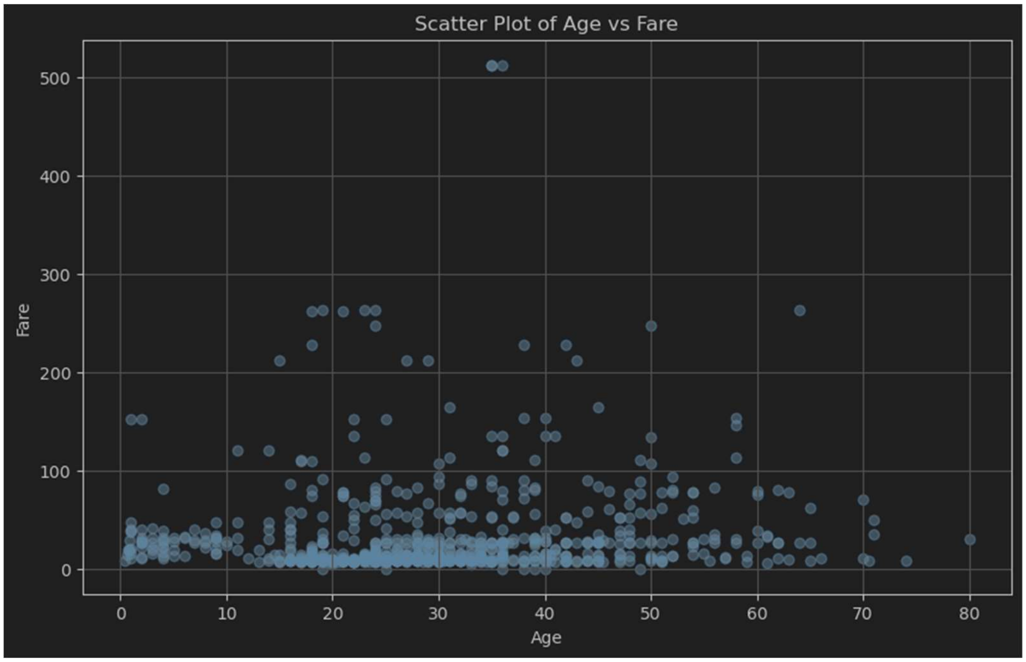
Explanation: The high Chi-square statistic indicates a significant deviation between observed and expected frequencies, suggesting that *Sex* and *Survived* are associated. The extraordinarily small p-value, much lower than conventional significance levels (e.g., 0.05), leads to rejecting the null hypothesis. This provides compelling evidence that the two variables are not independent and have a significant association between them.

Analysis: The degrees of freedom align with the 2×2 contingency table used for the Chi-square test. The significant difference between observed and expected frequencies indicates that gender was crucial to survival outcomes. Historical accounts support the idea that women were prioritized in the evacuation, as the statistical findings reflect.

Quantitative variables and the association between them

For the quantitative variables, we selected *Age* and *Fare* for analysis. *Age* represents the distribution of passengers across different life stages, while *Fare* reflects their economic status, with higher fares typically associated with first-class tickets. Analyzing these variables together provides insights into the socioeconomic factors that may have influenced survival outcomes.

Scatter plot



A scatter plot of *Age* vs. *Fare* reveals several key trends:

Fare Distribution: Most passengers paid fares below 100, indicating that most were likely traveling in second or third class. However, there were several outliers, with fares exceeding 200, and some passengers paid over 500, suggesting they traveled in first class or had special accommodations.

Age Distribution: Most passengers were between 0 and 50 years old, with fewer passengers in the older age range (above 60).

Patterns in Age vs. Fare

Younger Passengers (Age < 20): Most younger passengers paid lower fares, which suggests they were more likely to be in third class or benefited from family group discounts.

Middle-Aged Passengers (20–50): This group exhibited a broader fare spread, with some passengers paying significantly higher amounts, likely representing first- and second-class accommodations.

Older Passengers (50–80): Although fewer, some older passengers paid higher fares, suggesting they were also represented in first-class accommodations.

Statistical Analysis

Covariance Matrix: The covariance between *Age* and *Fare* is 73.85, indicating a slight positive relationship. *Fare* also tends to increase as *Age* increases, although the relationship is weak. However, covariance alone does not standardize the relationship or clarify the association's strength due to unit dependence.

Pearson’s Correlation Coefficient: The Pearson correlation coefficient is 0.096, suggesting a weak positive linear relationship between *Age* and *Fare*. The p-value of 0.0102 indicates that this relationship is statistically significant, despite the minimal correlation.

Conclusion

In conclusion, analyzing both qualitative and quantitative variables reveals significant insights into the Titanic dataset. The Chi-square test highlights a strong association between *Sex* and *Survived*, confirming that gender played a crucial role in survival outcomes. Women were more likely to survive, reflecting societal norms and prioritization during evacuation. On the other hand, the relationship between *Age* and *Fare* is weak, with a slight positive correlation. While age may have influenced the class of travel, other factors, such as passenger class or group travel, likely had a more significant impact on the fare paid. The weak correlation between these two variables emphasizes the need for further investigation, incorporating additional factors such as passenger class and family size to understand survival patterns better.