David Koyrakh

Final Project

College of Science and Technology, Bellevue University

DSC-530: Data Exploration and Analysis

Professor Metzger

11/18/2024

The goal of this study was to investigate the famous Titanic disaster and the spectrum of variables at play. In particular, I was interested to learn whether passengers' ticket fares, a known socioeconomic variable, influenced passengers' likelihood of surviving. My hypothesis was that passengers who paid higher fares had a better chance of survival. This led to the prediction that the observed average fare amongst survivors must be higher than non-survivors.

My Exploratory Data Analysis (EDA) for this study examined five variables in Stanford's Titanic dataset: survival outcome, class, age, fare price, and family size aboard. Various patterns emerged from the data throughout the EDA: Age demonstrated a strong relationship with survival, with younger passengers surviving at higher rates. Age was fit decently-well with a normal distribution, with the exception of a spike at the beginning of the graph around infants and very young children. Potentially, this spike may indicate that saving younger children was prioritized during the Titanic disaster. Socioeconomic variables, including ticket fare and passenger class, were also shown to correlate with survival outcomes.

To assess the correlation (and possible effect) of socioeconomic variables on survival, I calculated the point-biserial correlation and analyzed box plots, rather than Pearson's correlation and scatter plots. This decision was made on the basis that passenger class is categorical and survival is binary (either the passenger survived or not). While it is possible to make scatter plots with a binary variable, it is less standard to do so, as they are less visually insightful. Therefore, I went ahead and created box plots which highlight the main regions where the variables overlapped as well as providing a broad sense of their overall distributions. The box plots and correlations supported my hypothesis that higher socioeconomic status correlates with better survival outcomes. My hypothesis was also supported by the permutation test, which refuted the

null hypothesis by confirming that there was significant difference in mean fares between survivors and non-survivors.

This study was not without its limitations. Further analysis would benefit from the inclusion of more variables, such as cabin location and passenger embarkation point. Besides potentially uncovering previously-unknown survival patterns, the effect of these variables on survival could be compared against the effect from socioeconomic variables. Another limitation is around the nature of correlation, which is not causation; it could be that external factors were at play, causing the variables to correlate with survival coincidentally. The main challenge in this study and analysis was in dealing with non-continuous variables. This necessitated the use of getting creative with alternative statistical tools, visuals, and calculations.

Overall, this EDA study on the Titanic dataset provided me with an insightful and fascinating exercise for putting new learned tools and methods into action.