Titanic Dataset Interpretations & Analysis

A graph of a number of passengers

Description automatically generated

The histogram of passengers’ ages shows a high concentration of younger individuals, with a peak around ages 20-30. This shows a skew, with much fewer older passengers. This is expected, as the traveling demographic around the time was mostly young adults.

A graph of a family size

Description automatically generated with medium confidenceA pie chart with numbers and a number on it

Description automatically generatedHowever, the distribution of the classes (as seen in the pie chart) is a little different than expected. Surprisingly, there were noticeably less second-class passengers than first class. I was surprised by this, as I was expecting more of a 60/30/10 split between the three. This highlights that there is little socioeconomic diversity aboard the titanic, with a majority being third and second class with only the wealthiest aboard being able to afford first class.

This scatter plot shows the relationship between Ticket Price and Total Family Size (SibSp + Parch). This shows that although there is some tie, Fare price is not significantly dependent on family size. More likely they were impacted by class and cabin location, rather than the accompanying family.

A chart of different colored squares

Description automatically generatedA screenshot of a computer screen

Description automatically generatedI found the correlation scores to be most interesting, and even used them to accurately predict columns of data to compare against one another. This figure shows that among all else, fare had the highest correlation to Survived, in which I expected Pclass to play more of a role. It also demonstrates some relationships my new column of data (FamilySize) has towards fare. Fare has a solid correlation to class, as expected. This graph is useful to visually see what is related and what is unrelated data.

A screenshot of a computer

Description automatically generatedThe chi-squared test I used to determine survivability via class. As seen, first class is the ONLY class with >50% survivability. This shows not only the difference between classes, but the unfair survivability of the titanic, where class is one of the biggest factors (along with age/sex) in whether a person survived or not.

A screenshot of a computer

Description automatically generated

Comparing the breakdown of classes departing from each of the Titanic’s three ports, we can see a clear class distribution. It can be interpreted that Queenstown is a poorer area than the rest, with only 6.5% of passengers being able to afford a ticket above third class. This also shows that Cherbourg is likely a very wealthy area, with over 50% of passengers from there being first class. This data was interesting to compare to the socioeconomic demographics from each of the three ports of the titanic.