Titanic Dataset Report

BUS674

Chelsea Schuebel

**Overview/Description of Data:**

The sinking of the Titanic is one of the most infamous shipwrecks in history. In 1912, during her maiden voyage, the Titanic was labeled unsinkable. This dataset is a compilation of passengers that were aboard the Titanic.

**Methods Used:**

Minimal Data scrubbing/cleaning was required to work with the dataset. There were only some missing values in the age variable, so those have been excluded from the analysis.

Methods used in this data analysis include the following:

* Descriptive Statistics
* Graphs
* Correlation Matrix
* Linear Regression (Simple and Complex)

**Results From Analysis:**

Descriptive Statistics from this analysis tell us that the dataset has a count of 891. The dataset frequencies also tell us that 342 or 38.4% of these passengers survived the sinking of the titanic. There were 3 ticket classes that passengers were able to purchase for boarding.

The range for each variable are as follows:

* Age: .42-80 years old
* Fare: $0-$512.33
* Siblings/Spouse: 0-8
* Children/Parent: 0-6

The ranges for these values vary, so it is nice to see the means/average of each variable to get a better sense of the data.

The means in the data:

* Age: 30 years old
* Fare: $32.20
* Siblings/Spouse onboard: .52
* Children/Parent onboard: .38

A histogram was created based on age, and it is skewed to the right.

The graph in the appendix shows the relationship between Age and Fare, which is positive. The older people on the ship spent more on their tickets.

Correlation: As you can see from the correlation index in the appendix, there is a significant correlation between ticket class and fare, so that is how I chose to create my simple linear regression.

This simple linear regression model of dependent variable being the ticket class and the independent variable being the fare. This gave an R value of .549 which is not bad, but it is worth looking to see if there are more key factors that determine the ticket class.

A complex linear regression was created with the dependent variable being the ticket class, and this time having independent variables as follows: Age, Fare, Sibling/Spouse onboard. This regression gave us a better R value of .640, but it shows that the sibling/spouse onboard is not significant to the model.

Another regression was made with the independent variables changing to just age and fare. This gave us an R value of .639, which is not much different from when we included the sibling/spouse variable.

**Executive Summary:**

There are a number of things that we can learn from this analysis. It seems that the age range on the Titanic was quite broad (age .42 up to 80), but most people aboard the ship were younger than 40. There is also a positive relationship between age and the price of fare. We can guess that the older the individuals were, the more they were able to spend and in turn were in a higher ticket class. Our regression model shows us that ticket class is dependent upon these two variables. There really is no significance between ticket class and the number of siblings/spouses or parent/children aboard.

I also looked at the relationship between whether the passengers survived (independent) and what ticket class (dependent) they were in. The linear regression model came up with an R value of .338. This tells me that there is a relationship between the two and it would be interesting to dig into further analysis to see how ticket class and survival are linked.

To better understand the data, we could also do some external learning about the Titanic and its passengers, and in the future, this may be helpful in creating better and more accurate models.

**Appendix:**



