Data Visualization on the Titanic Dataset

Jawadul Chowdhury

February 10, 2025

1 Abstract

To be worked on.

2 Introduction

The Titanic was a British passenger ship that sank in the Atlantic Ocean on April 15, 1912. The ship had struck an iceberg on its maiden voyage from Southampton, England to New York City.

We use exploratory data analysis on the titanic dataset, using different visualizatoin tehcniques as well as determining which features are should be included in a machine learning model.

3 Dataset Description

When examining the dataset, there are a total of 16 features. Such variables are listed as follows:

- `PassengerId`: The ID of the passenger. This is a discrete numerical data type as the difference between units is constant.
- `Survived`: Whether the passenger survived or not.This is a nominal binary categorical data type as there is no order information. 0 means passenger didn't survive and and 1 means passenger did survive.
- `Pclass`: This is the passenger class. It is a ordinal categorical data type, as 1 means 1st ticket class and so on for 2 and 3, to identify passenger class.
- `Name`: The name of the passenger. This is a nominal categorical data type, as the names of the passenger have no order information.
- `Sex`: This is the sex of the passenger. This is a nominal categorical data type,as genders can't be ordered and is rather binary.
- `Age`: Age of the Passenger. This is a continuous numerical data type, as the difference between units is constant and can be counted.
- `SibSp`: Number of Siblings / Spouses of the passenger. It is a discrete numerical data type as it can be counted and has a constant difference between units.
- `Parch`: Number of Parents / Children abroad the Titanic. It is a discrete numerical data type, as it can be counted and can only take certain values.
- `Ticket`: This is the ticket number. This is a nominal categorical data type as there is no ordering information.
- `Fare`: This is the passenger fare. This is a ratio numerical data type as there is a true zero, where zero means that the passenger has not paid any fare.
- `Cabin`: This is the cabin number of the passenger. It is a nominal categorical data type as there is no order information bur rather a quantitvate classifcation.
- `Embarked`: This is the port where the passenger embarked. C is Cherbourg, Q is Queenstown and S is Southampton. This is a categorical nominal data type as there is no order information and has quantitative classification.
- `Age_fill_mean`: This is a copy of the Age column but the blanks have been filled in with the mean. This is a ratio numerical data type, because there are fractional values and has a true zero point.
- `Age_fill_median`: This is a copy of the Age column but the blanks have been filled in with the median. This is a discrete numerical data type because the age is defined to be continuous numerical.
- `Age_fill_mode`: This is a copy of the Age column but the blanks have been filled in with the mode. This is a discrete numerical data type, because the age is defined to be continuous numerical.
- `Age_fill_knn`: This is a copy of the Age column but the blanks have been filled in with the mean. This is a ratio numerical data type, because there are fractional values and has a true zero point.

4 Data Visualization on Survived

Now, we move on to visualizting some of the variables we have in the Titanic dataset against the `Survived` variable. We also look for any predictive relationships using the plots.

4.1 Fare vs Survived

In order to determine a appropriate plot type, we need to identify the type of data. We know that the `Survived` variable is binary and categorical, and we know that `Fare` is a ratio numerical data type. For this application, it would be best to use a box plot since we're dealing with categorical / numerical data, as shown in Figure 1.

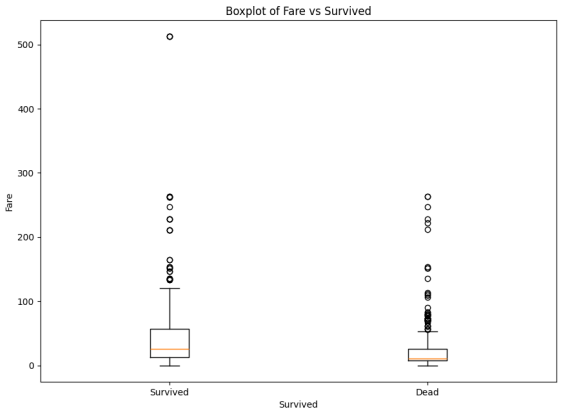

Figure 1: Box Plot of Fare vs Survived

4.2 Sex vs Survived

Next up, we now plot the `Sex` variable against `Survived` variable. We know that the `Sex` variable is a categorical variable, which uses male and female. We also have the survived variable which is also categorical as discussed previously. For this application, we can use a Heat Map since we're dealing with categorical / categorical data as shown in Figure 2.

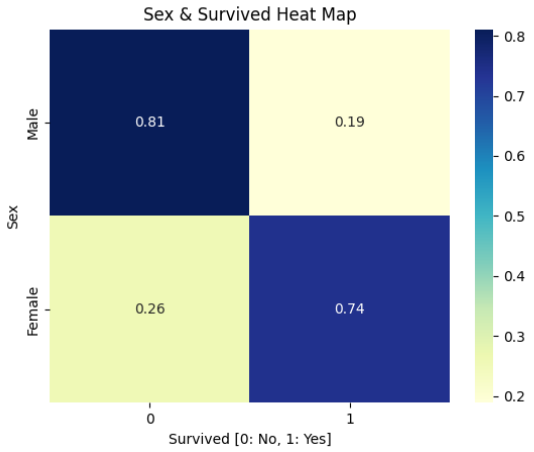

Figure 2: Heat Map of Sex vs Survived

4.3 Sibling / Spouse vs Survived

Next up, we now plot the `SibSp` variable against the `Survived` variable. We know that the `SibSp` variable is a discrete numerical data type, as it can be counted. As a result, since we're plotting a numerical data type against a categorical data type, we can use a bot plot to best represent this numerical / categorical data as shown in Figure 3.

4.4 Parent / Children vs Survived

Moreover, we now plot the `Parch` variable against the `Survived` variable. We know that the `Parch` variable is a discrete numerical data type, as it can be counted. As a result, since we're plotting a numerical data type against a categorical data type, we can use a bot plot to best represent this as shown in Figure 4.

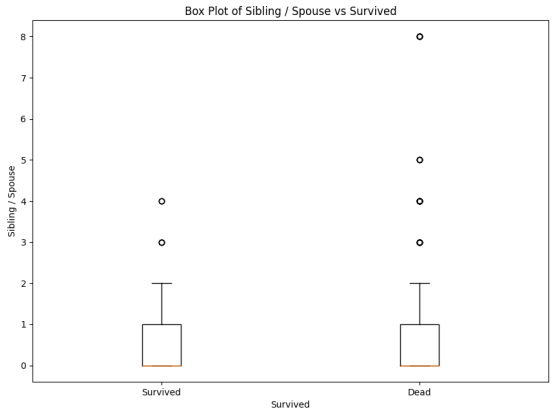

Figure 3: Heat Map of Sibling / Spouse vs Survived

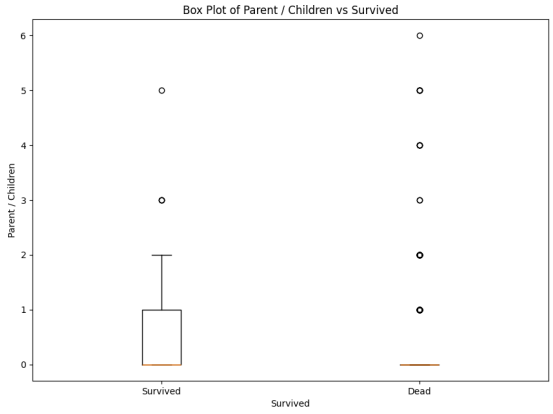

Figure 4: Heat Map of Parent / Children vs Survived

4.5 Embarked vs Survived

Moving on, we now plot the `Embarked` variable against the `Survived` variable. We know that the `Embarked` variable is a categorical data type, and since we're plotting the a categorical data type against a categorical data type, we can then use a heat map to best represent this as shown in Figure 5.

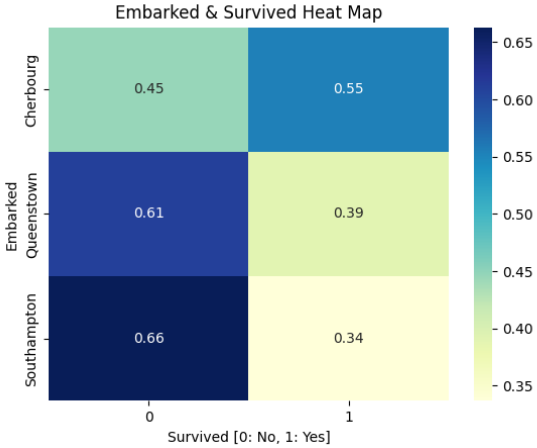

Figure 5: Heat Map of Parent / Children vs Survived

4.6 Passenger Class vs Survived

Next, we now plot the `Pclass` variable against the `Survived` variable. We know that the `Pclass` variable is a categorical data type, and since we're plotting the a categorical data type against a categorical data type, we can then use a heat map to best represent this as shown in Figure 6.

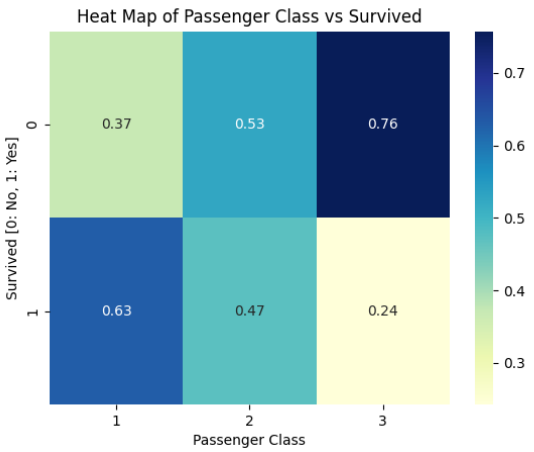

Figure 6: Heat Map of Parent / Children vs Survived

4.7 Age Fill Mean vs Survived

Now, we plot the `Age_fill_mean` variable against the `Survived` variable. We know that the `Age_fill_mean` is a ratio numerical data type and `Survived` is a categorical data type. To create a plot where we plot a categorical data type against a numerical data type, we can use a violin plot, as shown in Figure 7.

Figure 7: Violin Plot of Age (Mean) vs Survived

4.8 Age Fill KNN vs Survived

Now, we plot the `Age_fill_KNN` variable against the `Survived` variable. We know that the `Age_fill_KNN` is a ratio numerical data type and `Survived` is a categorical data type. To create a plot where we plot a categorical data type against a numerical data type, we can use a violin plot, as shown in Figure 8.

Figure 8: Violin Plot of Age (KNN) vs Survived

4.9 Age Fill Median vs Survived

Next, we plot the `Age_fill_Median` variable against the `Survived` variable. We know that the `Age_fill_Median` is a discrete numerical data type and `Survived` is a categorical data type. To create a plot where we plot a categorical data type against a numerical data type, we can use a violin plot, as shown in Figure 9.

Figure 9: Violin Plot of Age (Median) vs Survived

4.10 Age Fill Mode vs Survived

Lastly, we plot the `Age_fill_Mode` variable against the `Survived` variable. We know that the `Age_fill_Mode` is a discrete numerical data type and `Survived` is a categorical data type. To create a plot where we plot a categorical data type against a numerical data type, we can use a violin plot, as shown in Figure 10.

5 Relationships

Now that we have produced the graphs, we need to understand the predictive relationships that can be analyzed from them. In order to do this, we look at each visualization that has been produced.

Figure 10: Violin Plot of Age (Mode) vs Survived

5.1

6 Sources