Assignment 4: R programming

Members: zhp6 tow6

In this assignment, we used the dataset from Kaggle.com about the Titanic: Machine learning from Disaster. There are in total 12 attributes and 891 instances, (PassengerId, Survived, Pclass, Name, Sex, Age, SlibSp, Parch, Ticket, Fare, Cabin, Embarked) and we found that some data has already been lost. However, we could omit it. And we set the dataset for the variable 'mydata4' in R language. Among these data, age, sex, fare and family size have great impact on the passengers' survival. In the following, we will give multiple diagrams related to the dataset in order to analyze those attributes. We will use the package "ggplot2" in R, which is really powerful.

a. Whisker-plot

Whisker-plot is from the package "ggplot2", which can also be called as "box-plot", Here we will give two diagrams of different attributes, one is (Age, Survived), another is (Fare, Survived).

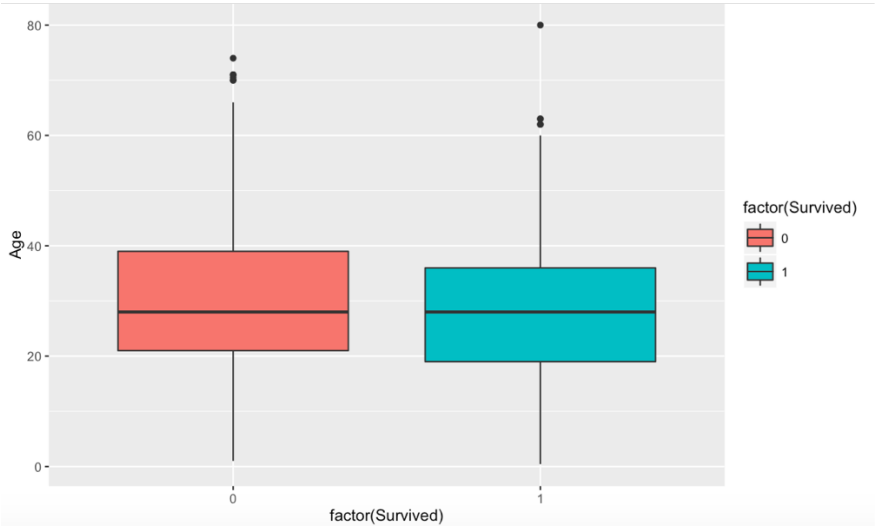

Figure I Whisker-plot of Survived on Age

1) From Figure I, we can see that the range of age 20-40 is the majority of people that are both survived and un-survived, which indicates that the passengers are mostly the young people.

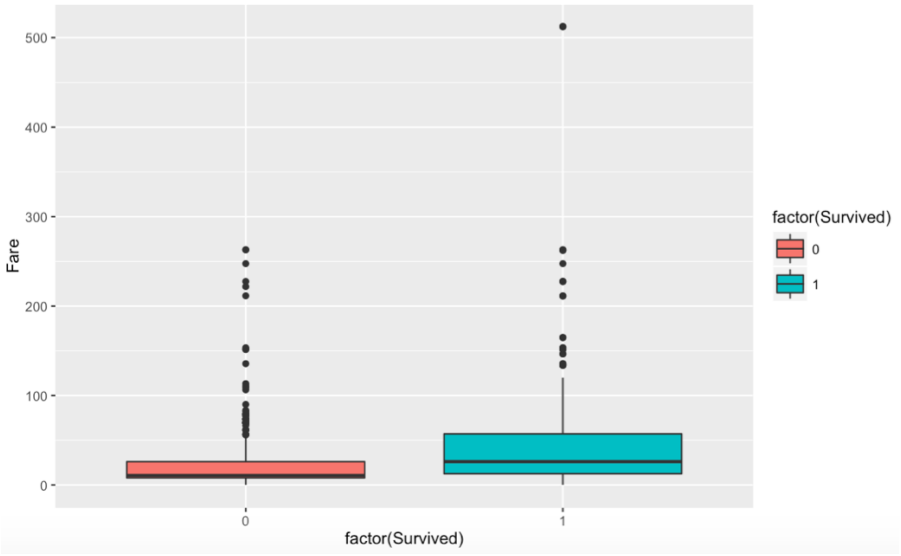

Figure 2 Whisker-plot of Survived on Fare

2)We can see from Figure 2 that passengers who bought a cheaper ticket have a high probability not to survive, although there are some exceptions existing. While the visualization of this diagram is very nice.

b. Histogram

A **histogram** is a graphical representation of the distribution of numerical data. It is an estimate of the probability distribution of a continuous variable (quantitative variable) and was first introduced by Karl Pearson.

1)

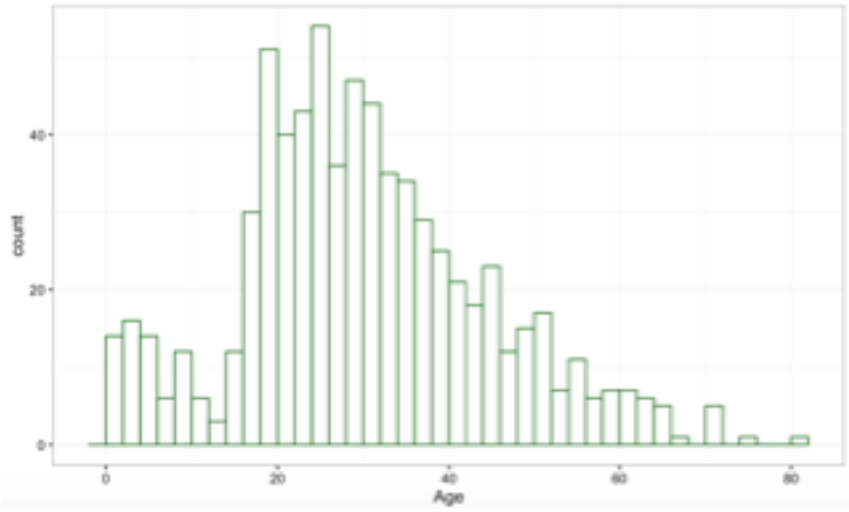

Figure 3 Histogram of Age distribution

We can see from Figure 3 that the distribution of the age of passengers are nearly the normal distribution, the range of the age 20-40 are the most comparing with others. The oldest passengers are over 80 years old and the youngest passengers were just born.

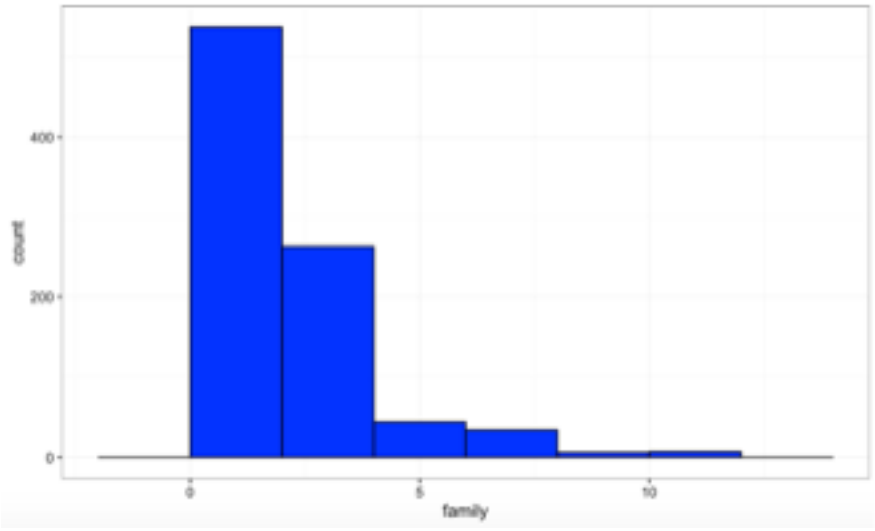

Figure 4 Histogram of family members' distribution

While the family attribute consists of one's siblings and parent on board, we can see from figure 4 that the family members less than 3 are dominant among all the families. And family members more than 3 are very little on board.

c. Facet grid

Facet grid plot can show grouped grids to better generate comparisons.

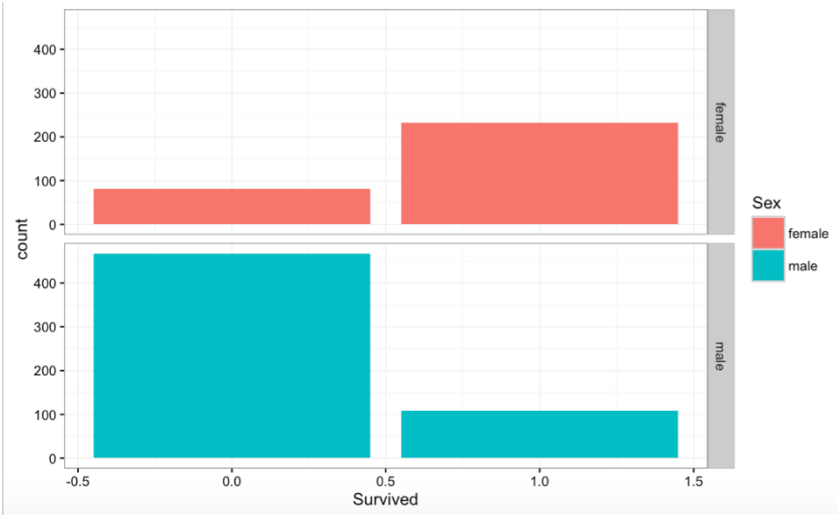

Figure 5 Facet grid divided by sex in terms of survive

While we can see from the diagram that the male passengers have great probabilities not to be survived, while for the female passengers, they have great probabilities to be survived. We can learn from this result that 'lady first' when met the disaster.

Figure 6 Facet grid divided by Embarked in terms of survive

While, from the Figure 6 we can see that the port S has the highest probability of not being survived, while for the port C and port Q, there is no obvious difference between survived and not survived. We can come to the conclusion that passengers got on board from port S was more perilous than other ports.

d. Violin plot

Violin plot is similar to box plot, which can show the probability density of the data at different values.

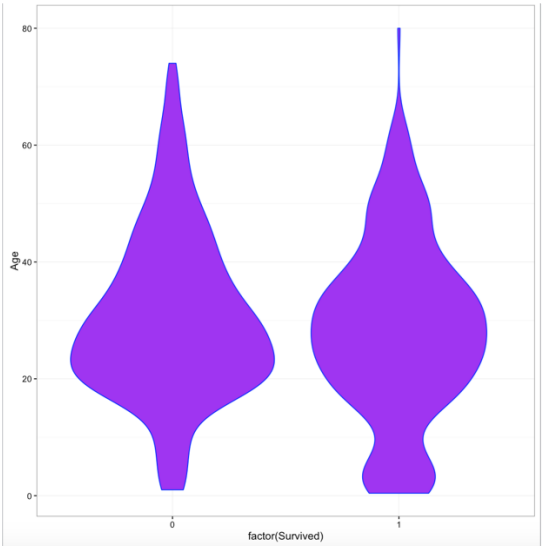

Figure 7 violin plot of survived or unsurvived for the Age of passengers

While we can see from Figure 7 that for the unsurvived passengers, the range of the age was mostly during 20-40 years old, and for the survived paassengers, the range of the age was also mostly during 20-40 years old. However, for the kids and the elders, there were also many survivors.

2)

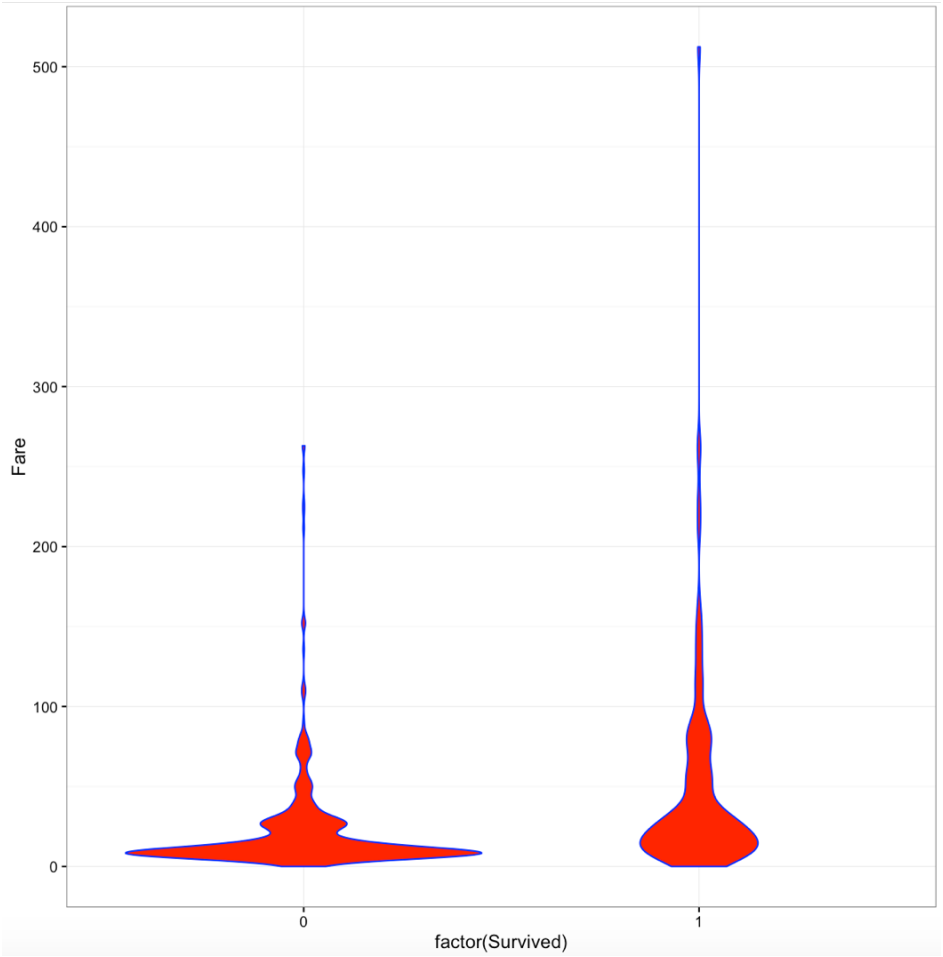

Figure 8 Violin plot of survived or un-survived for the fare of the tickets

While we can see from the diagram that when passengers bought a ticket that the price was lower than 50 dollars, they were more likely not to be survived. However, if the passenger bought the ticket that the price was more than 50 dollars, then they were more likely to be survived.

e. Heatmap

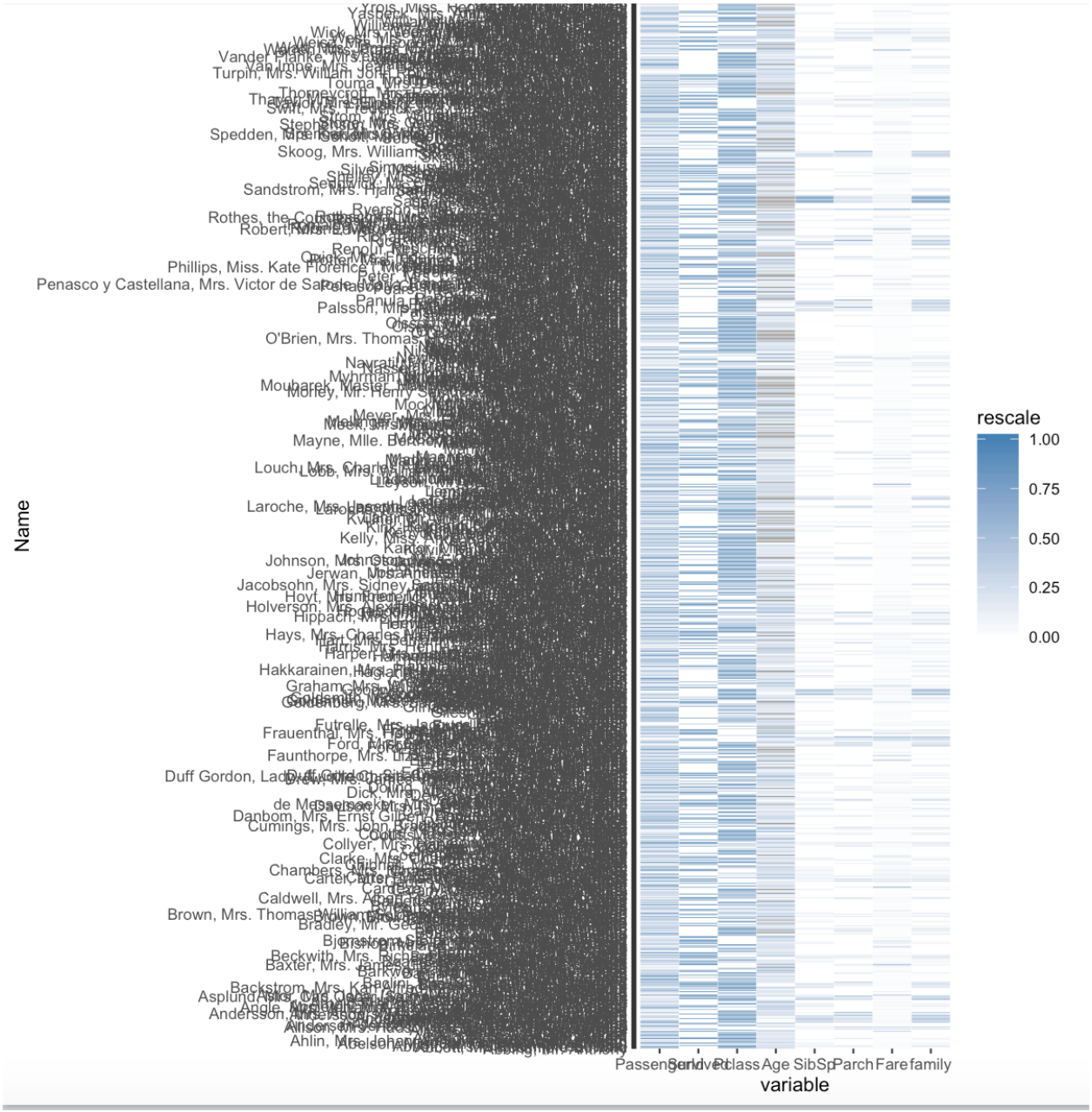

Figure 9 the heat map of attributes

While we can see from the diagram that the deeper the color is, the larger the value is. For example, for the age columns, we could see multiple of degrees for one color, which indicate different ages among the passengers.