INFSCI 2725 Data Analytics
Assignment 4 - R Programming

Date: 10/21/2015

Student name: Ting LI, Hongyuan Cui

pittUserLogin: til42, hoc27

Tutor name: Marek Druzdzel

Executive Summary

This report offers a brief graph summary for Titanic training dataset using whisker-plot, histogram, facet-grid, violin plot and heat map. Referring to those plots, we find that age, sex, pclass, family size calculated from variables sibsp(Number of Siblings/Spouses Aboard) and parch (Number of Parents/Children Aboard), and fare all influence the survival rate of passengers. The age of passengers survived basically centers on 30s. Females, people who paid higher fare, and passengers with higher social-economic class have higher survival rate. Family size can also impact survival rate in that family size less than three have more survival.

1. Whisker-plot

Figure 1, Figure 2 and Figure 3 below are the whister plot for variable age, fare and family size respectively. Figure 1 shows that the age of passengers survived is more centered on 30s. Almost all people older than 60 died except for few exceptions. We can observe from Figure 2 that passengers who paid higher fare have high likelihood of survival. The passenger who paid more than 500 survived. Figure 3 illustrates that although passengers with family size less than three have higher survival rate, there are passenger with family size between three to six survived. Moreover, it it obvious that barely no one survived with a family size greater than six.

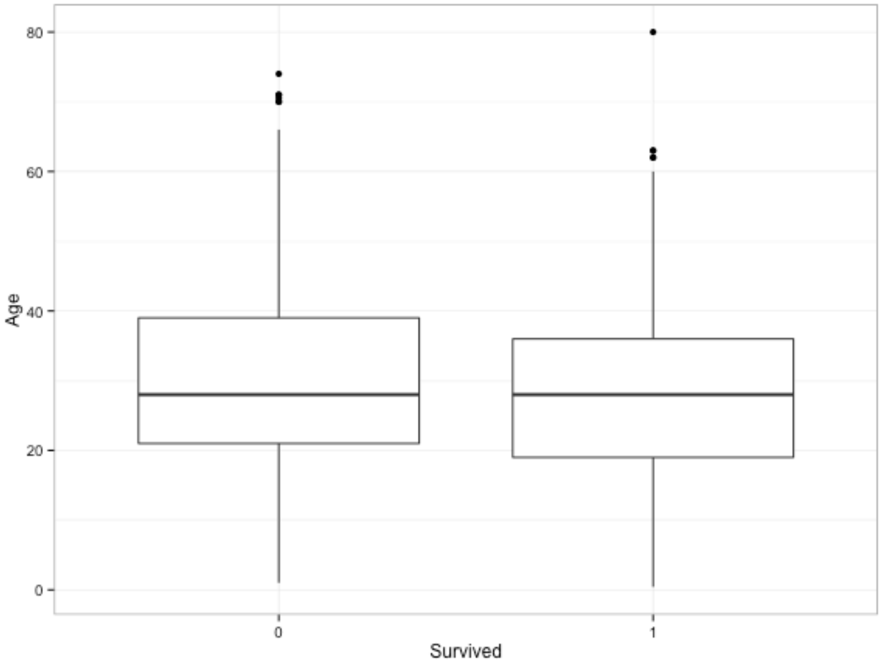

Figure 1. Whisker-plot of Survived on Age

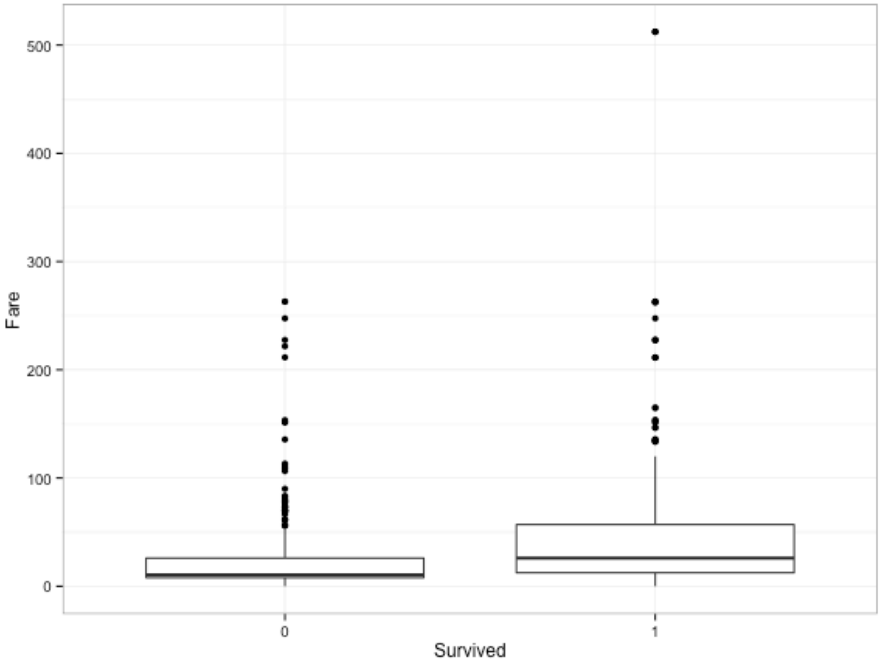

Figure 2. Whisker-plot of Survived on Fare

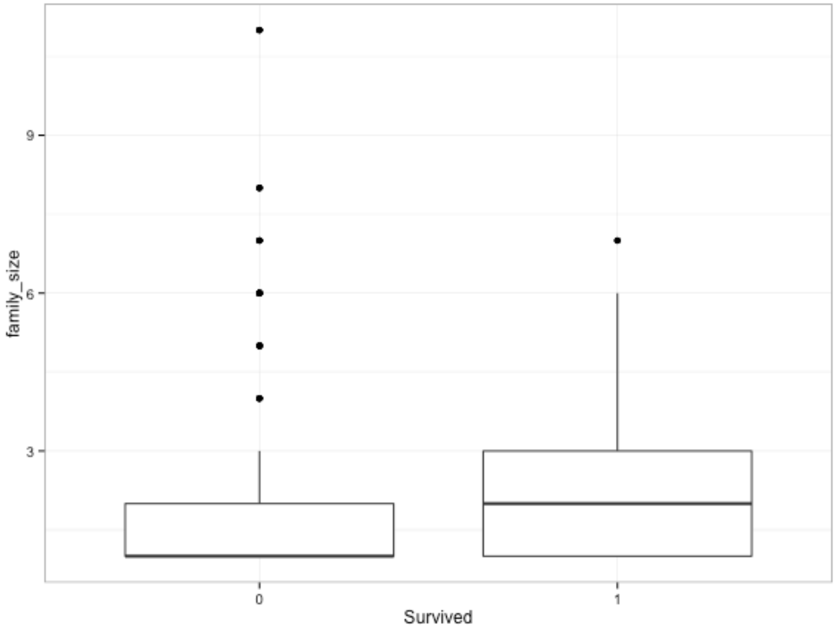

Figure 3. Whisker-plot of Survived on FamilySize

2. Histogram

Two histograms below demonstrate the distribution of age and family size correspondingly. It is noticeable that the age of most passengers on Titanic are from 20 to 40. Only a few passengers are older than 70. The number of children is larger than that of the elder. Figure 5 tell us that most people are single in the ship, followed by a family size of two or three. Family with size greater than six is few.

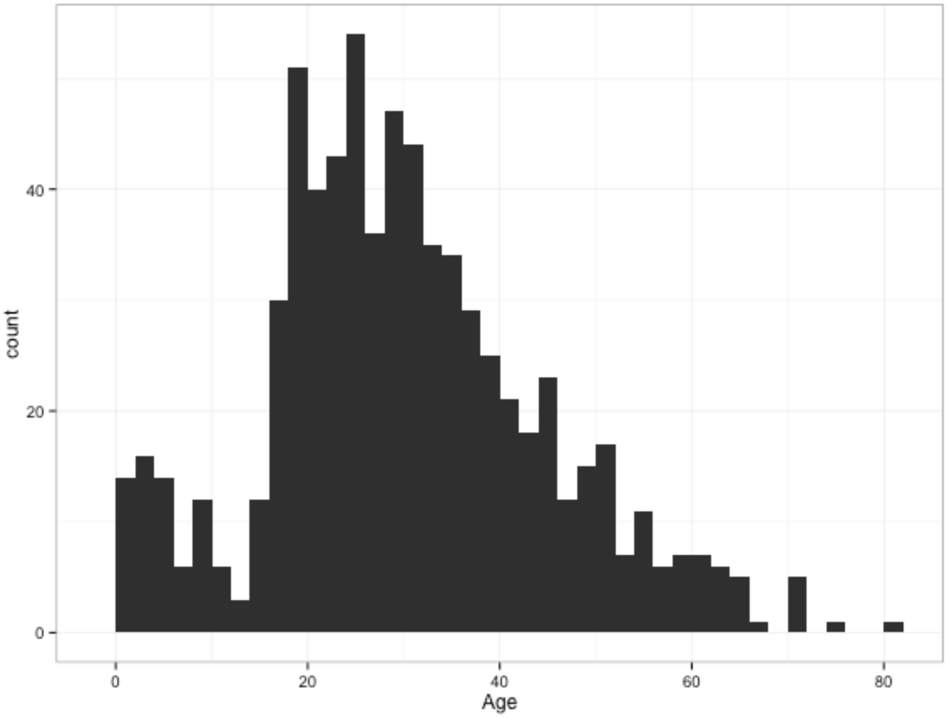

Figure 4. Histogram of Age

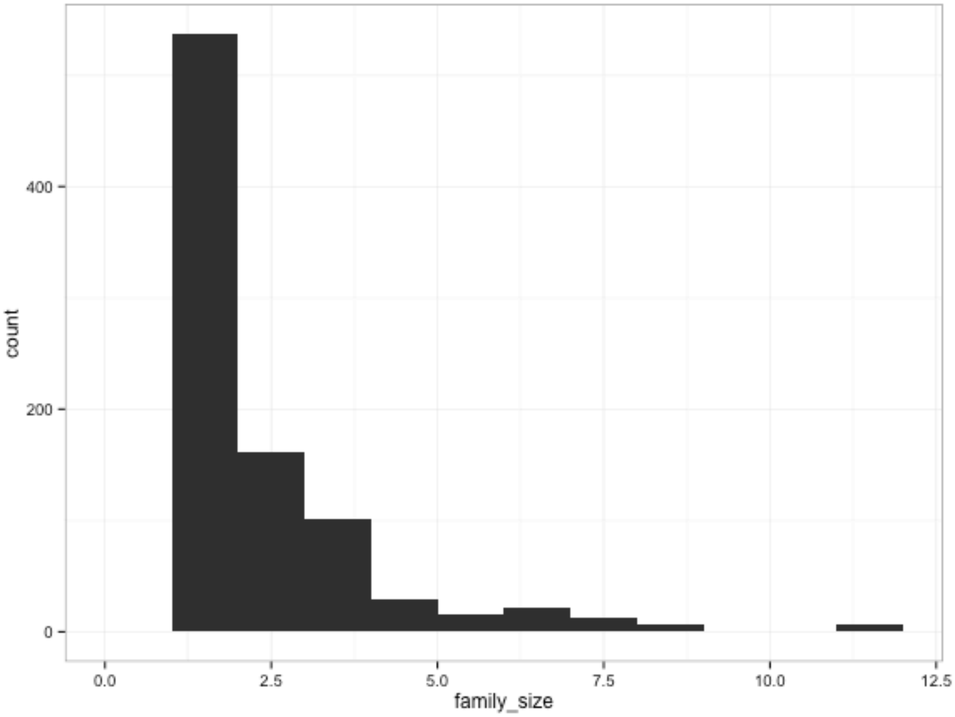

Figure 5. Histogram of FamilySize

3. Facet_grid

Figure 6 illustrates the sex distribution of passengers conditioning on whether they have survived. There are about 600 males and 300 females in the training dataset. However, only around 100 males survived suggesting a survival rate of only 17%. In contrast with the lower survival rate of male, the survival rate of female is over 60%. This significant difference demonstrate that female enjoys the privilege when being rescued. Figure 7 presents the comparison of number of survival among different classes. Pclass is a proxy for socio-economic status, and 1st, 2nd and 3rd refer to upper, middle and lower correspondingly. It is obvious that the highest death rate appears in passengers belonging to 3rd class, and the highest survival rate is in 1st class. The amount of passengers in 2nd class basically has equal amount of death and survival. Therefore, we can believe that the social-economic status has great influence on the sequence of rescue.

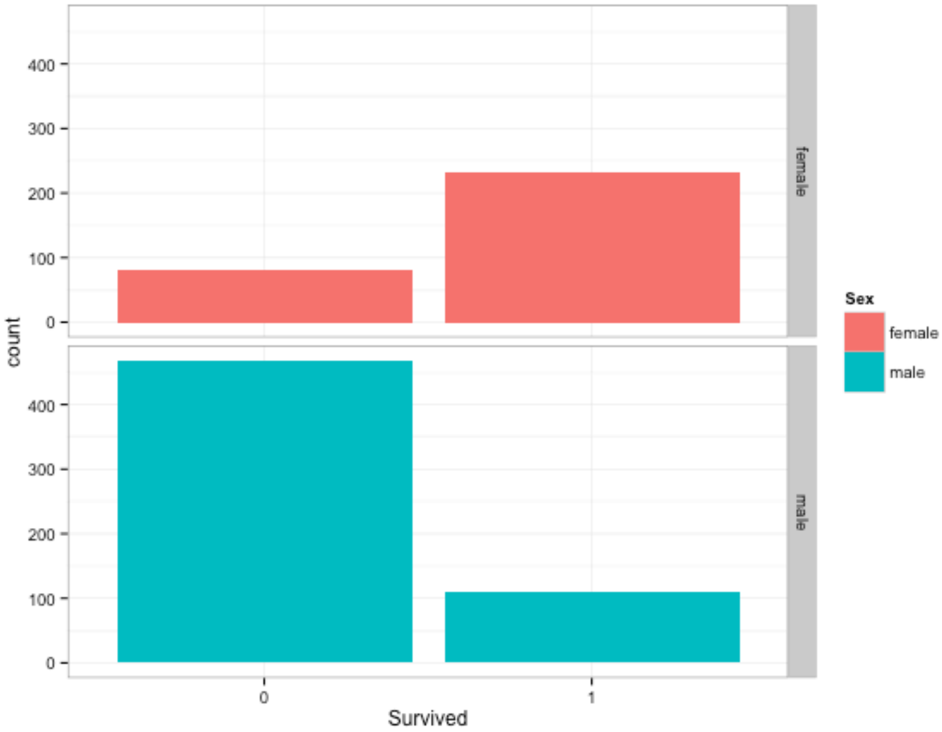

Figure 6. Conditional Histogram of Survived on Sex

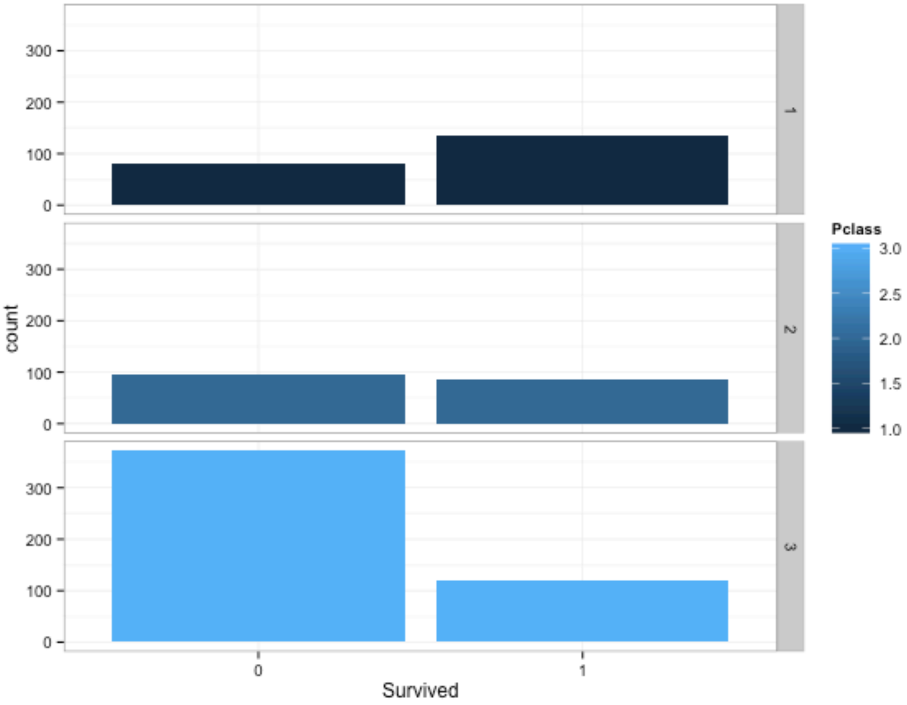

Figure 7. Conditional Histogram of Survived on Pclass

4. Violin plot

Figure 8 below shows two age distributions separately for whether passengers have survived. According to this figure, we can see children (age below 10) are likely to survive and people around 20 years old have a little bit more probability of death. For people close to 80 years old, there is no exception for survival. Besides, people between 40 and 70 has the similar possibility to survive. The figure 9 demonstrates fare distribution conditioning for whether people survived or not. We can see the fare vary from 0 to over 500. For people whose fare over 300, they all survived, comparing to those who pay the ticket less than 50, most of them did not survive.

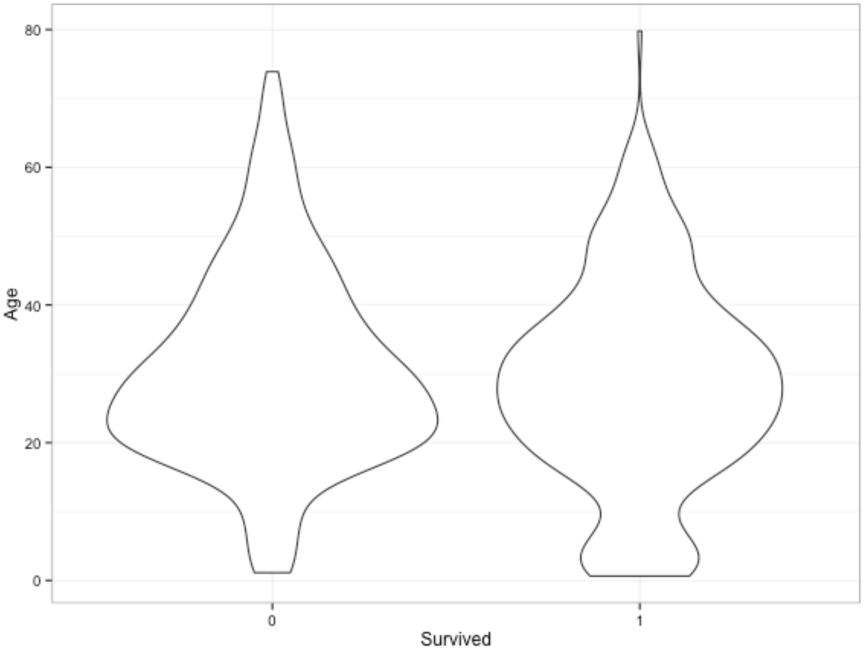

Figure 8. Violin Plot of Survived on Age

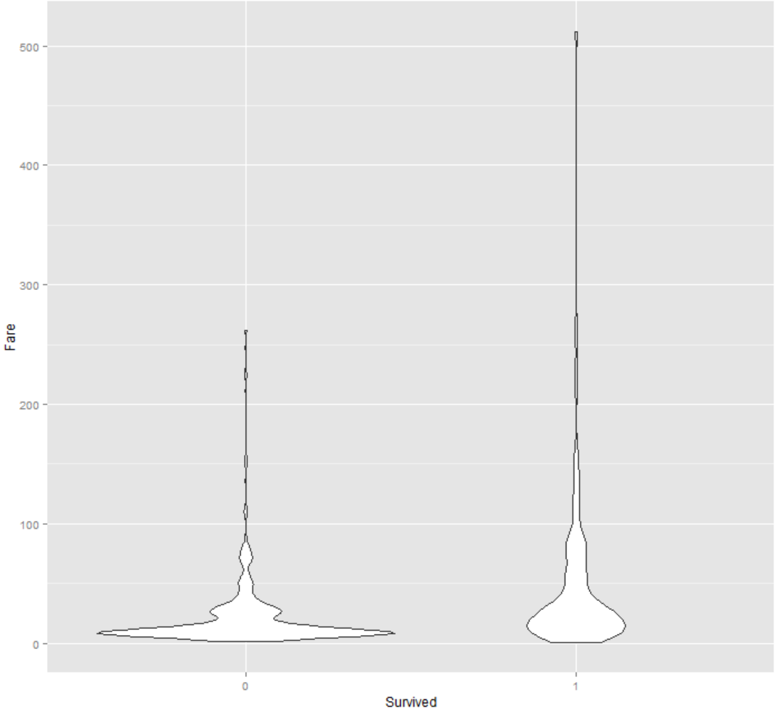

Figure 9. Violin plot of Survived on Fare

5. Heat map

Figure 10 is a heat map on Pclass, Sex, Age, Fare and Family_size of 100 survived people who are randomly sampled. This figure illustrates that survived people contains lots of upper Pclass(deeper color) and female(deeper color). We can also see that the color in age column is almost deep orange, which means that survived people have many young people. In the fare column, the color is shallower, showing that most survived people bought expensive tickets. In the family_size column, the color is almost deep so among the survived people, there are more people with few family members.

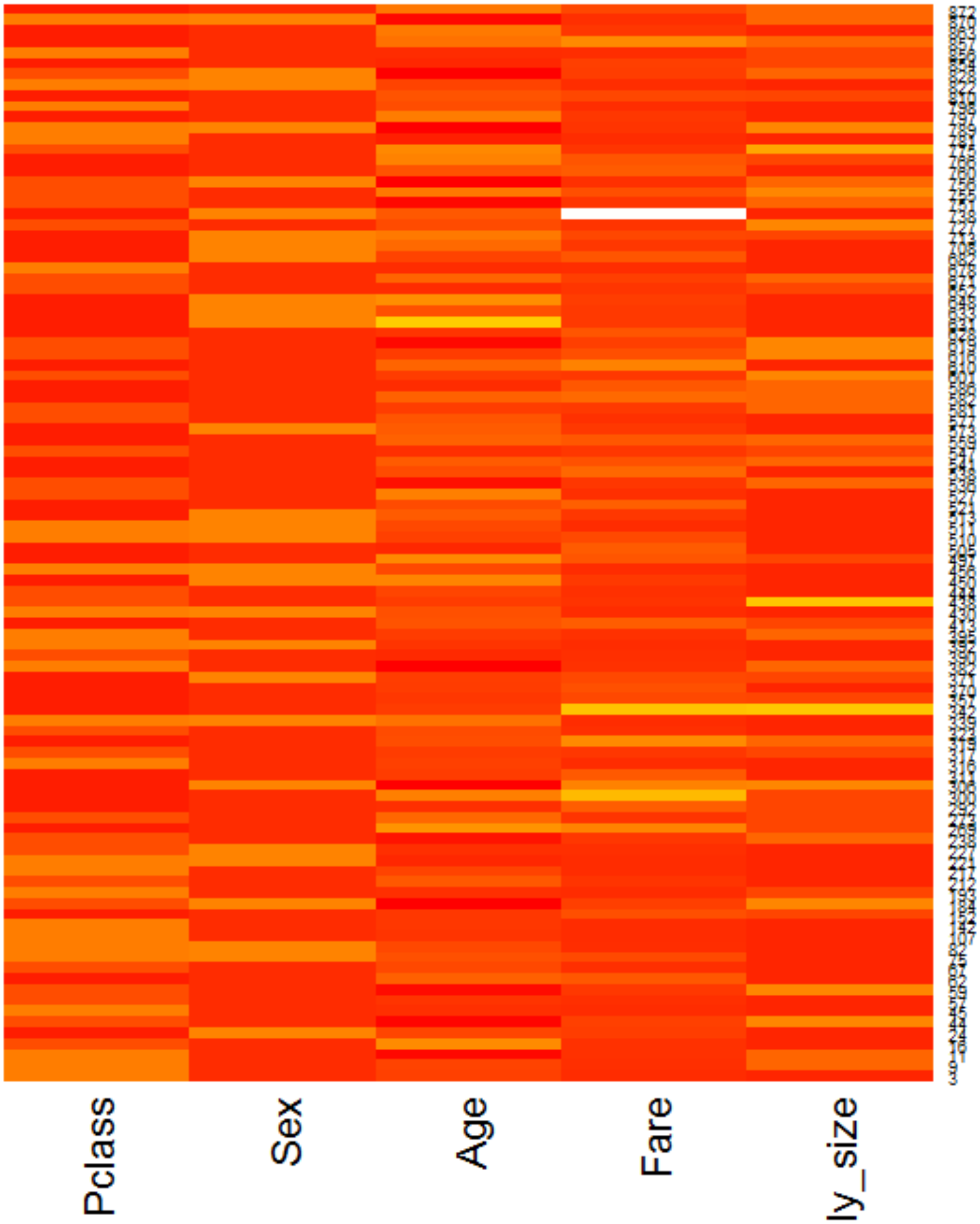

Figure 10. Heat Map of 100 random samples of Survived Persons on Pclass, Sex, Age, Fare and Family_Size