Data Analytics for Titanic Disaster

Lab 4 - October 13, 2015

# 1. Overview

## 1.1 Group Members

* Siying Zhang ([siying.zhang@pitt.edu](mailto:siying.zhang@pitt.edu));
* Chukun Xia([CHX26@pitt.edu](mailto:CHX26@pitt.edu));
* Abdulaziz Abdulrahman Almuzaini ([aaa169@pitt.edu](mailto:aaa169@pitt.edu));
* Yue Li ([YUL134@pitt.edu](mailto:YUL134@pitt.edu));
* Abhishek Mukherjee (ABM84@pitt.edu)

## 1.2 Data Source

* **Titanic: Machine Learning from Disaster** <https://www.kaggle.com/c/titanic/data> -> train.csv

# 2. Introduction

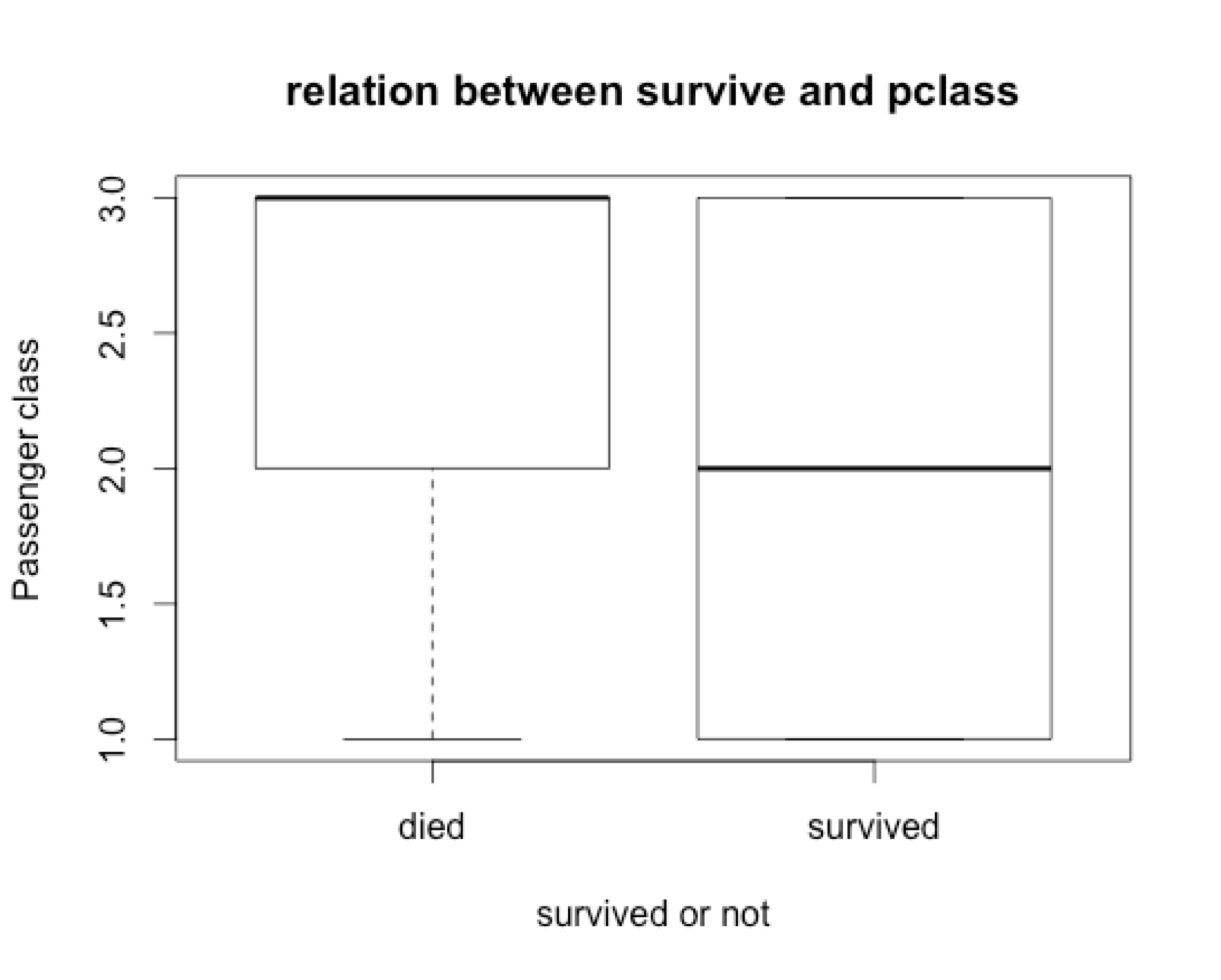
Lab 4 is truly easy, but it makes us feel truly heavy when we analyse these disaster data.

Firstly, we read data from the downloaded csv file, while could be done by the method read.csv. And then select data that we want to analyse to make a further data visualisation with ggplot.

According to the requirements, we did data visualisation with five different types of diagram, whisker plot, histogram, facet grid, violin plot and heat map.

# 3. Detailed Explanation

## 3.1 Whisker plot

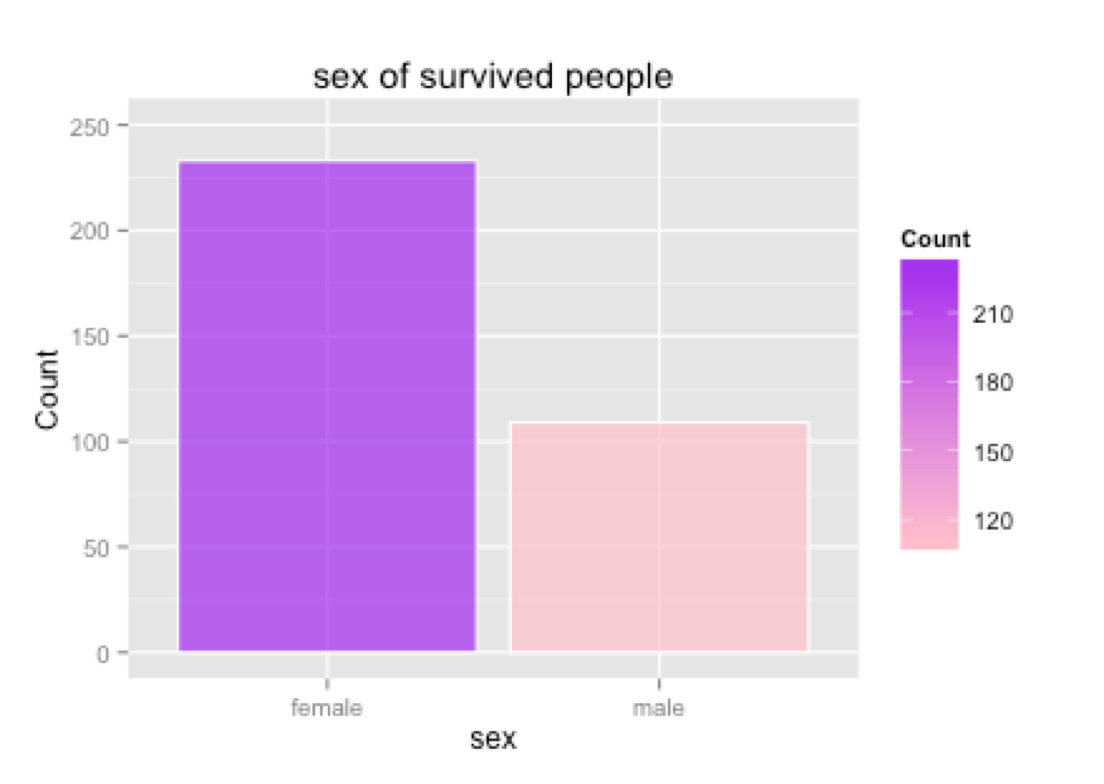


This graph shows the relation between passenger class and whether they are survived. It is obvious that passengers in 2nd and 3rd class were less likely to die, as first class cabins were closer to the deck of the ship. It’s hard to see something if you only concentrate on the survived box because its average line is in the middle. The reason may be the data we selected is separate, the passengers’ classes have three types rather than a continuous type.

## 3.2 Histogram

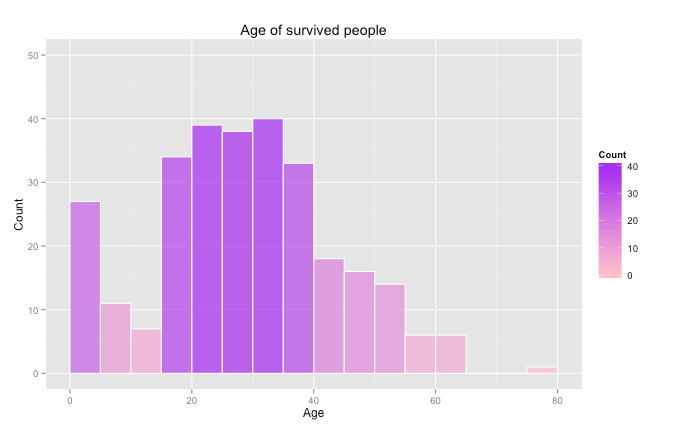
Histogram looks like the bar diagram, which is mostly used in the data analysis. While the histogram can not only represent the the value merely but also could represent it through the changes of colour. We did two histograms.

# 3.2.1 Sex of survived people



This graph shows the number of survived people group by gender. It is obvious that the number of female is two times more than male. The reason may be that women are more likely to be taken care of and have more opportunities to survive their lives. So “sex” is a good predictor when we are doing analysis.

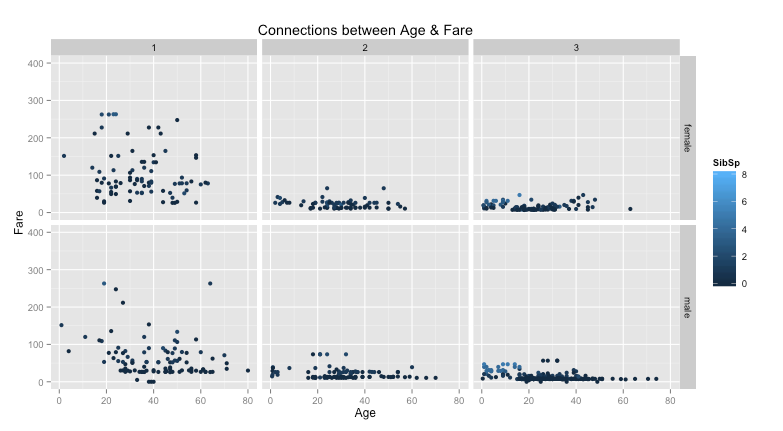
# 3.2.2 Age of survived people



This histogram represents that the amount of people in different ages distribution in survived group. As you can see, most survived people are between 15 to 40, they are strong and full of energy. Also, children between 0 to 5 years old also survived a lot, which indicates that they were protected by their parents carefully.

## 3.3 Facet grid

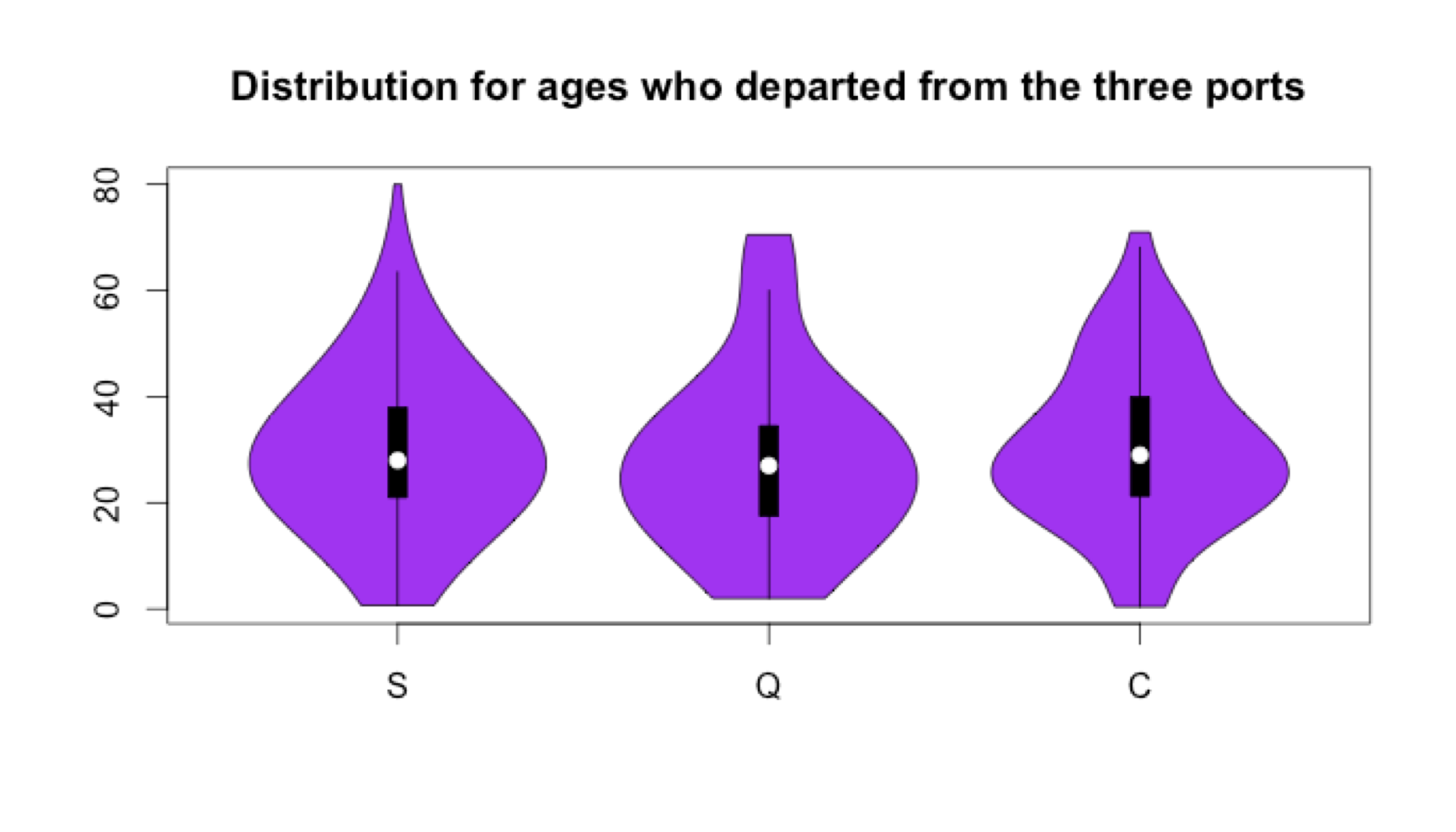
Facet grid is really a “big” diagram. It could represent plenty of data in one diagram.



In the graph above, it represents the relationship between fare and age in different sexes and classes. High fare appears mostly in the first class, and especially for female. No matter female or male, the fare in class 2 & 3 is fairly low and almost in the same range.

## 3.4 Violin plot

Violin plot is good to represent the density about a range of data. In this lab, we use it to represent the distribution for ages who departed from the three ports.

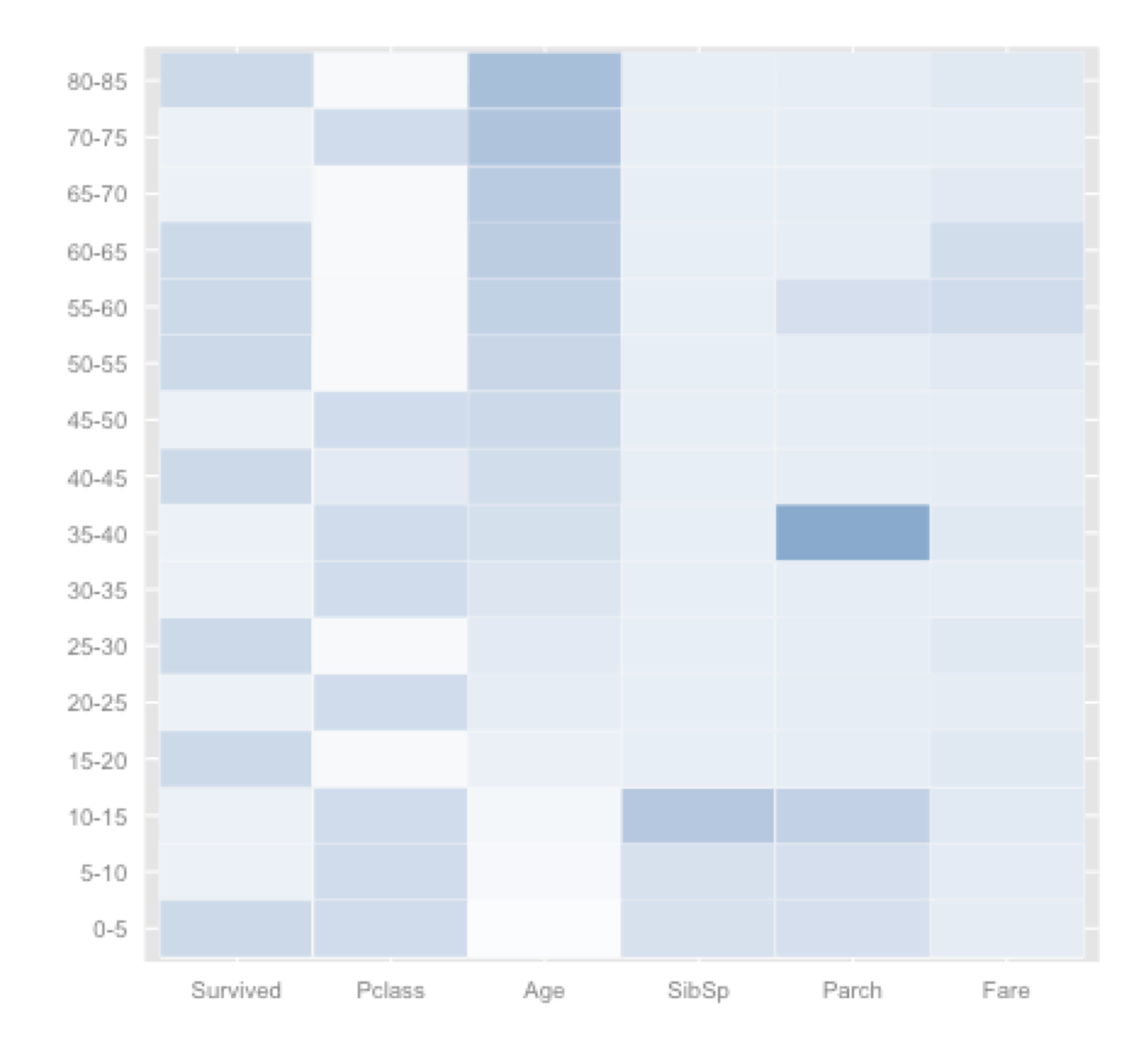


This graph makes it easier to determine the distribution and the density among disparate travelers’ ages that departed from the three cities (“S”,”C”,”Q). For those different cities, it is obvious to determine the median is between 25-30. Also, this graph shows us the normal distribution for city S is between age 10 -50, for city Q is from infant – age 45, and finally for city C is quite between from 20 to 40.

## 3.5 Heat Map

Heat map is suitable to represent the density with the differences of colour.

We adopt heat map to describe the density of almost every parameters on different age levels.



The graph is a Heatmap on different age level, it shows that:

- Kids and elders are more likely to survive (deeper blue);

- People in lower class cabins are more likely to survive (lighter blue);

- When comparing “Survived” with “Pclass”, some correlation appears: people in higher level of cabin are less likely to survived (Cells have lighter blue in “Pclass” have darker blue in “Survived”);

- People under age 15 are more likely to have Siblings/Spouses/Parents on aboard, meanwhile people age 35-40 and 55-60 are more likely to have Parents/Children on aboard;

- Fare paid by people age between 55-65 are higher than fare paid by people in other age.

# 4. Summary

R and especially ggplot are really powerful, which could represent data in a more visualized way. Different kinds of diagrams have different concentrates. It’s worthy for us to explore and study deeplier.