Lab 5

Enter Your Name and UNI Here

February 15, 2018

Instructions

Make sure that you upload an RMarkdown file to the canvas page (this should have a .Rmd extension) as well as the PDF output after you have knitted the file (this will have a .pdf extension). Note that since you have already knitted this file, you should see both a **Lab5__UNI.pdf** and a **Lab5__UNI.Rmd** file in your UN2102 folder. Click on the **Files** tab to the right to see this. The files you upload to the Canvas page should be updated with commands you provide to answer each of the questions below. You can edit this file directly to produce your final solutions. The lab is due 11:59 pm on Tuesday, February 20th.

Part 1 (Iris)

Background

The R data description follows:

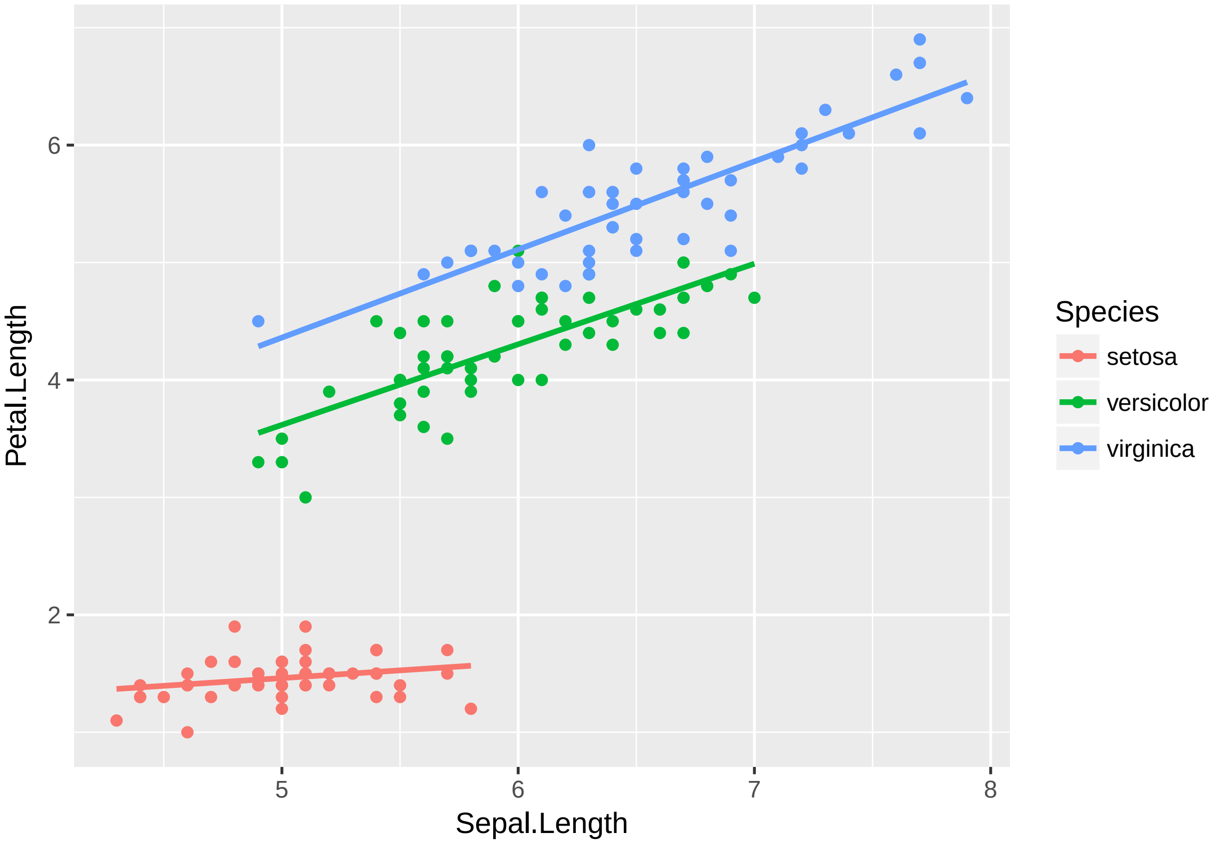
This famous (Fisher's or Anderson's) iris data set gives the measurements in centimeters of the variables sepal length and width and petal length and width, respectively, for 50 flowers from each of 3 species of iris. The species are Iris setosa, versicolor, and virginica.

Task

Produce the exact same plot from Lab 3 using `ggplot` as opposed to Base **R** graphics. That is, plot **Petal Length** versus **Sepal Length** split by **Species**. The colors of the points should be split according to **Species**. Also overlay three regression lines on the plot, one for each **Species** level. Make sure to include an appropriate legend and labels to the plot. Note: The function **coef()** extracts the intercept and the slope of an estimated line.

```r
library(ggplot2)
## Plot.
set.seed(1)
iris$Species <- factor(iris$Species)

p <- ggplot(data = iris) + geom_point(mapping = aes(x = Sepal.Length, y = Petal.Length, color = Species
p
```



Part 2 (World's Richest)

Background

We consider a data set containing information about the world's richest people. The data set us taken form the World Top Incomes Database (WTID) hosted by the Paris School of Economics [http://topincomes.g-mond.parisschoolofeconomics.eu]. This is derived from income tax reports, and compiles information about the very highest incomes in various countries over time, trying as hard as possible to produce numbers that are comparable across time and space.

Tasks

1) Open the file and make a new variable (dataframe) containing only the year, "P99", "P99.5" and "P99.9" variables; these are the income levels which put someone at the 99th, 99.5th, and 99.9th, percentile of income. What was P99 in 1993? P99.5 in 1942? You must identify these using your code rather than looking up the values manually. The code for this part is given below.

```r
setwd("/Users/salimmjahad/Desktop/STAT_COMP/lab5")
wtid <- read.csv("wtid-report.csv", as.is = TRUE)
wtid <- wtid[, c("Year", "P99.income.threshold","P99.5.income.threshold", "P99.9.income.threshold")]
names(wtid) <- c("Year", "P99", "P99.5", "P99.9")
subset(wtid, Year %in% c(1993, 1942))
```

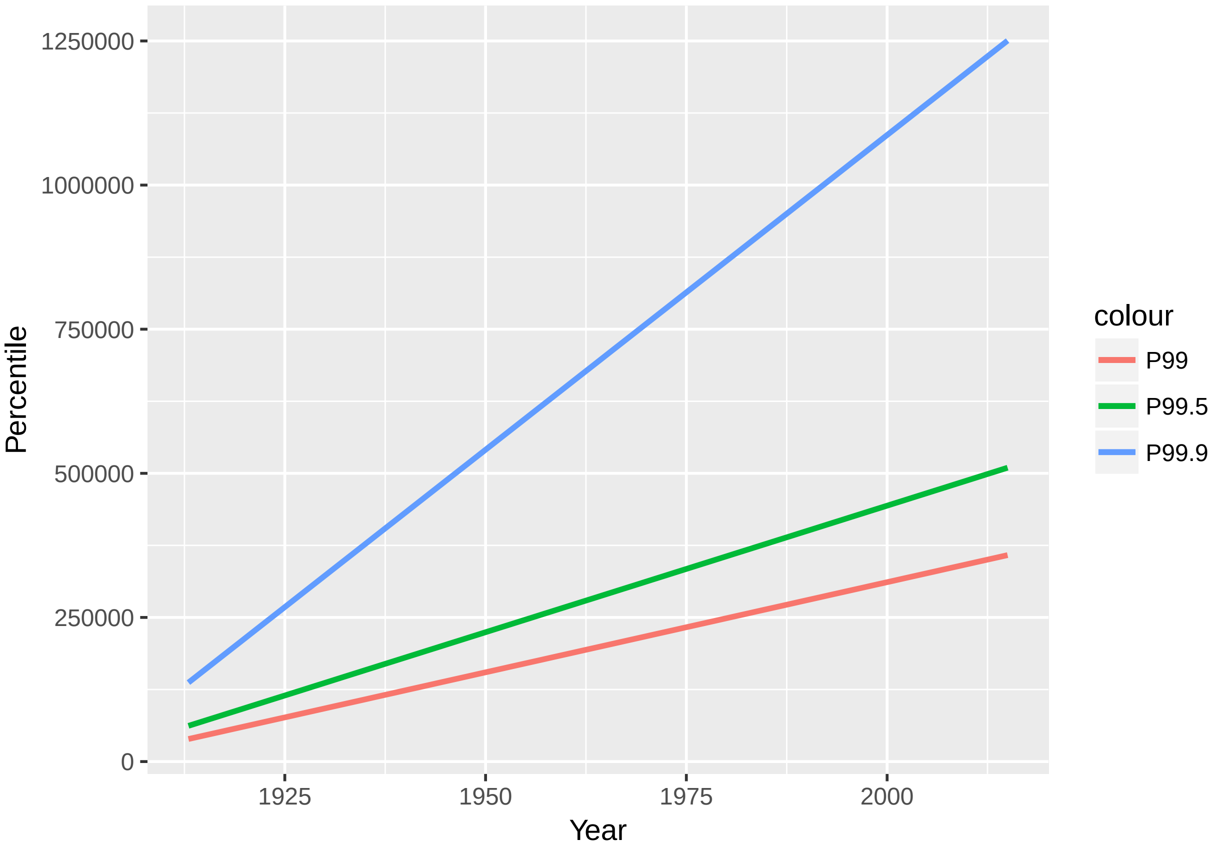
```
##    Year      P99    P99.5    P99.9
## 30 1942 120306.1 189140.6 499711.5
## 81 1993 273534.9 387241.2 928224.4
```

P99 in 1993 273534.9 and P99.5 in 1942 189140.6

2) Using `ggplot`, display three line plots on the same graph showing the income threshold amount against time for each group, P99, P99.5 and P99.9. Make sure the axes are labeled appropriately, and in particular that the horizontal axis is labeled with years between 1913 and 2012, not just numbers from 1 to 100. Also make sure a legend is displayed that describes the multiple time series plot. Write one or two sentences describing how income inequality has changed throughout time.

```
## Plot
p <- ggplot(data = wtid) + geom_smooth(aes(x = Year, y = P99, color = "P99"), method = "lm", se = FALSE

p
```



Income inequality has increased throughout time as the gaps between even the richest people have increased.

Part 3 (Titanic)

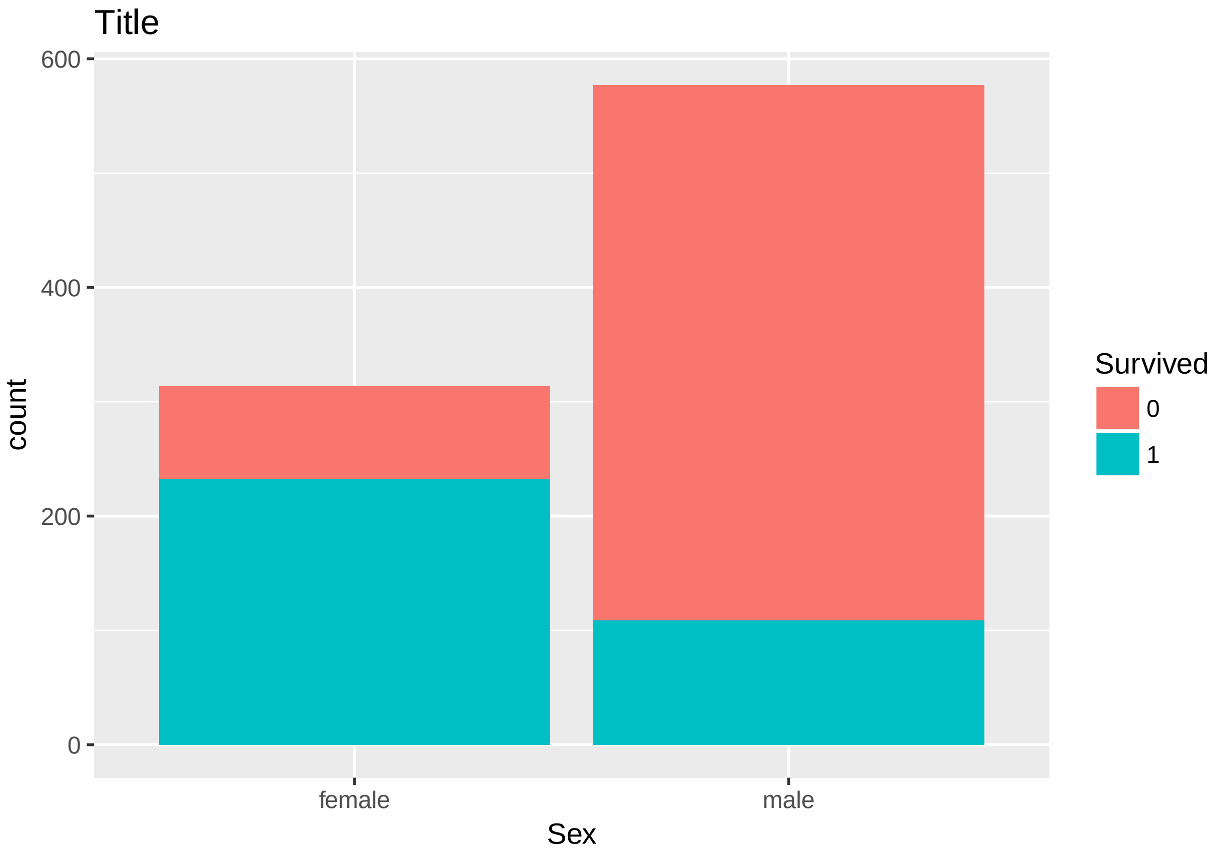
Background

In this part we'll be studying a data set which provides information on the survival rates of passengers on the fatal voyage of the ocean liner *Titanic*. The dataset provides information on each passenger including, for example, economic status, sex, age, cabin, name, and survival status. This is a training dataset taken from the Kaggle competition website; for more information on Kaggle competitions, please refer to https://www.kaggle.com. Students should download the data set on Canvas.

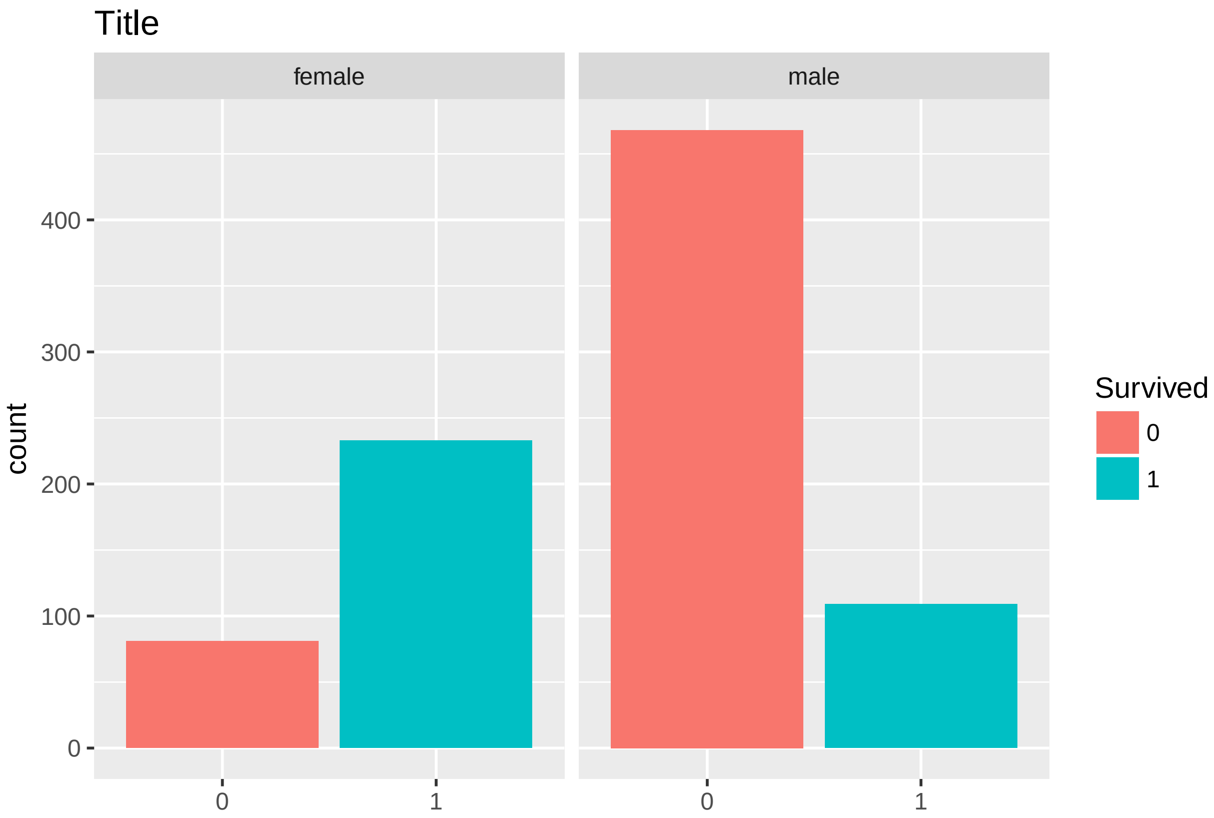
Tasks

1) Run the following code and describe what the two plots are producing

```r
# Read in data
titanic <- read.table("Titanic.txt", header = TRUE, as.is = TRUE)
# Plot 1
ggplot(data=titanic) +
  geom_bar(aes(x=Sex,fill=factor(Survived)))+
  labs(title = "Title",fill="Survived")
```



```r
# plot 2
ggplot(data=titanic) +
  geom_bar(aes(x=factor(Survived),fill=factor(Survived)))+
  facet_grid(~Sex)+
  labs(title = "Title",fill="Survived",x="")
```

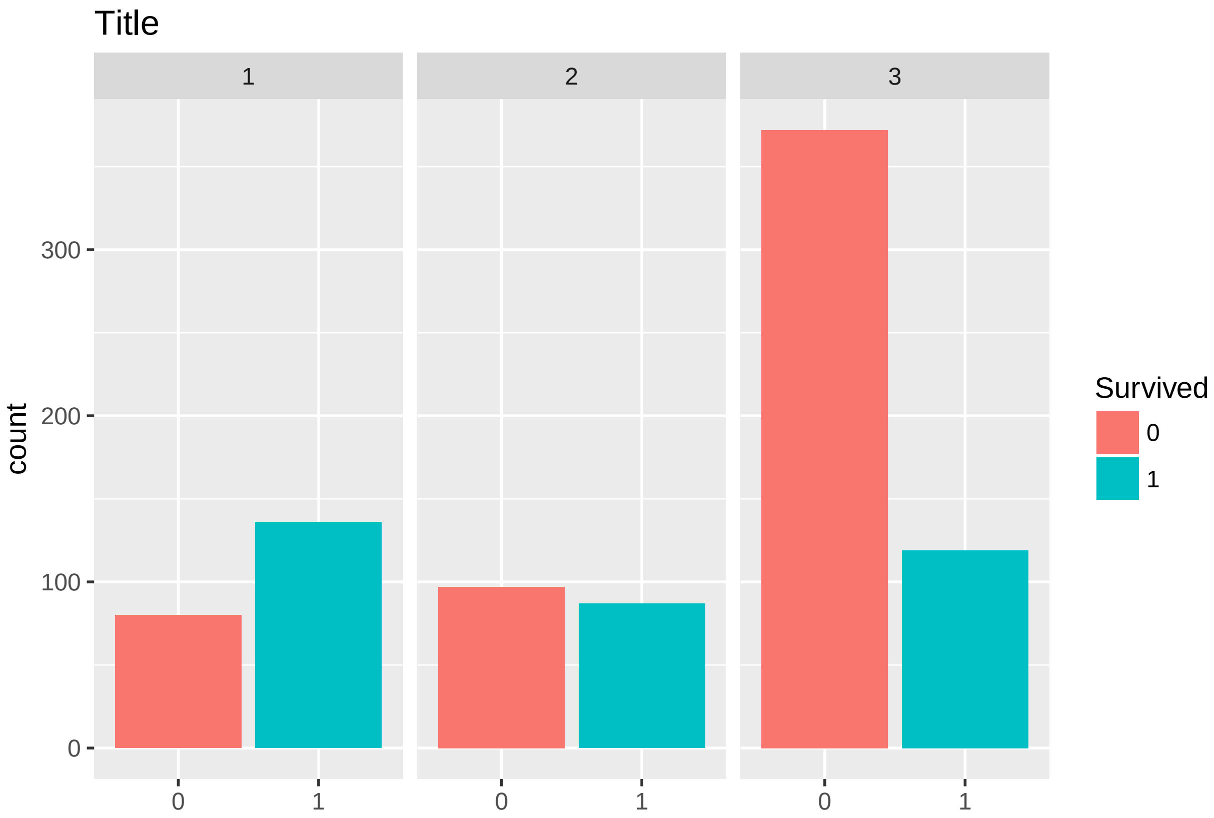


The first plot is a bar plot where the x-axis is categorical variable Gender that is male or female and the y axis is the number of passengers of that gender. There is a third variable which is represented using colors that separate each bar to a blue color for those who survived and a red one for those who did not.

The second plot is basically two separate bar plots. One is describing the count of males who were on the titanic (y-axis) depending on whether they survived or not (x-axis). The other one is similar but for female.

2) Create a similar plot with the variable **Pclass**. The easiest way to produce this plot is to **facet** by **Pclass**. Make sure to include appropriate labels and titles. Describe your

```
# Plots
ggplot(data=titanic) +
  geom_bar(aes(x=factor(Survived),fill=factor(Survived)))+
  facet_grid(~Pclass)+
  labs(title = "Title",fill="Survived",x="")
```



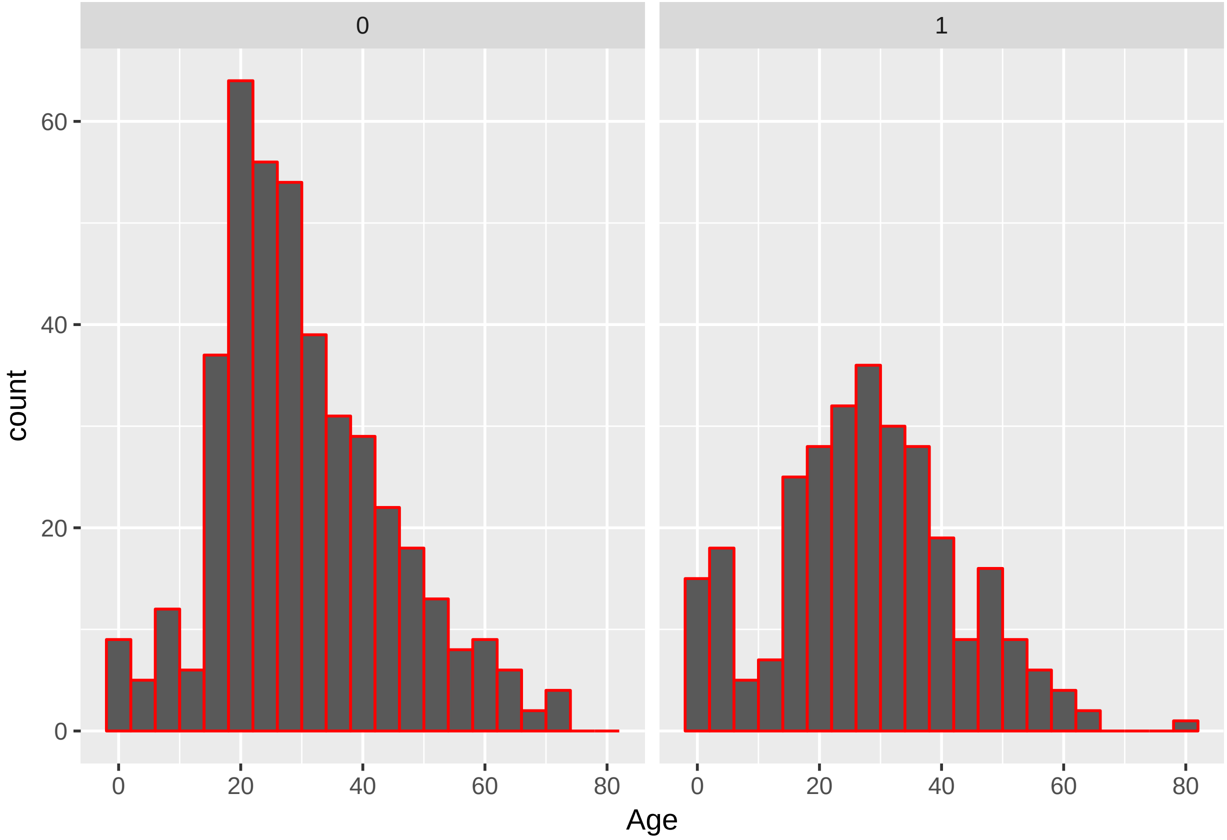
People from the first class have more survivors than not, the second class has almost a 1:1 ratio of survivors to non survivors, and the lower class definitely has many more deaths than survivals. The lower the class (better) the higher the ratio of survivors to deaths.

3) Create one more plot of your choice related to the **titanic** data set. Describe what information your plot is conveying.

```
# Plots
ggplot(data=titanic) + geom_bar(aes(x=Age), stat="bin", color = "red", binwidth = 4) + facet_grid(~Survi
```

```
## Warning: `geom_bar()` no longer has a `binwidth` parameter. Please use
## `geom_histogram()` instead.

## Warning: Removed 177 rows containing non-finite values (stat_bin).
```



We can see the distribution of non-survivors is sharper in the 18-30 age than the survivors one. This can show that someone who is between 18 and 30 was more likely to die in titanic.