# Visualisation with ggplot2

*Kevin Chesis*

*7/17/2019*

Figure 1: Chesis

# Contents



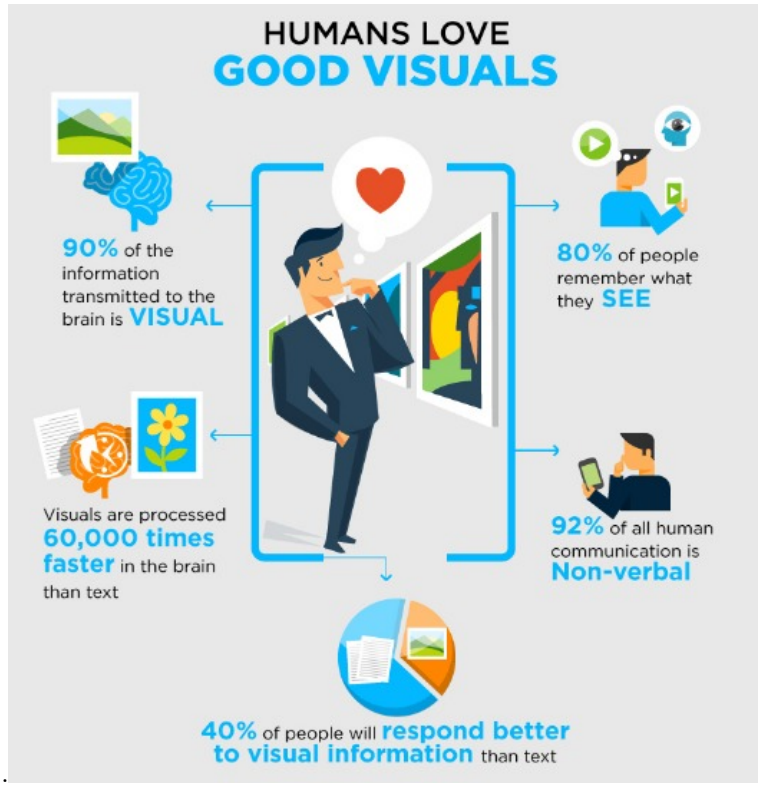
```
setwd("C:/Users/fibon/Desktop/Principal-Component-Analysis/Visualization")
```

# Data Visualisation

Data visualisation is the graphical representation of information and data. Visual elements aids to see and understand trends, outliers, and patterns in data. Some of the common visual elements include charts, graphs and maps. In the world of big data, data visualisation tools and technologies are essential to annalyze massive amounts of information and make data-driven decisions.

# Advantages of good data visualisation

Our eyes are drawn to colors and patterns. Our culture as human beings are visual including everything from art and advertisements in TV and movies.




The figure above clearly suggests that 80% of people remember what they see and visuals are processed 60,000 times faster in the brain than text. Data visualisation is a form of visual art that grabs interest and keeps our eyes on the message. If you stare on a spreadsheet of data you may not observe any trend but with a visualization, this could be attained effectively. Data visualization is storytelling with a purpose.

Most professional industry benefit from making data more understandable. Every STEM field benefits form data visualization. Powerful softwares such as Tableau have been developed just for the sole purpose of data storytelling through visualization. Data visualization is a skill that most if not all aspiring Data Scientist **MUST** have. In this article I am going to attempt to address the data visualization in R using *ggplot2*. This is a data visualization package for the statistical programming language R.

## Common General Types of Data Visualization

- Charts
- Tables

- Graphs
- Maps
- Infographics
- Dashboards

## Effective Visualization

1. Conveys the right information without distorting facts.
2. Is simple but elegant. It should not force you to think much in order to get it.
3. Aesthetics supports information rather than overshadow it.
4. Is not overloaded with information.

Load the required package and set a theme for clear visualizations.

## Scatter Plot

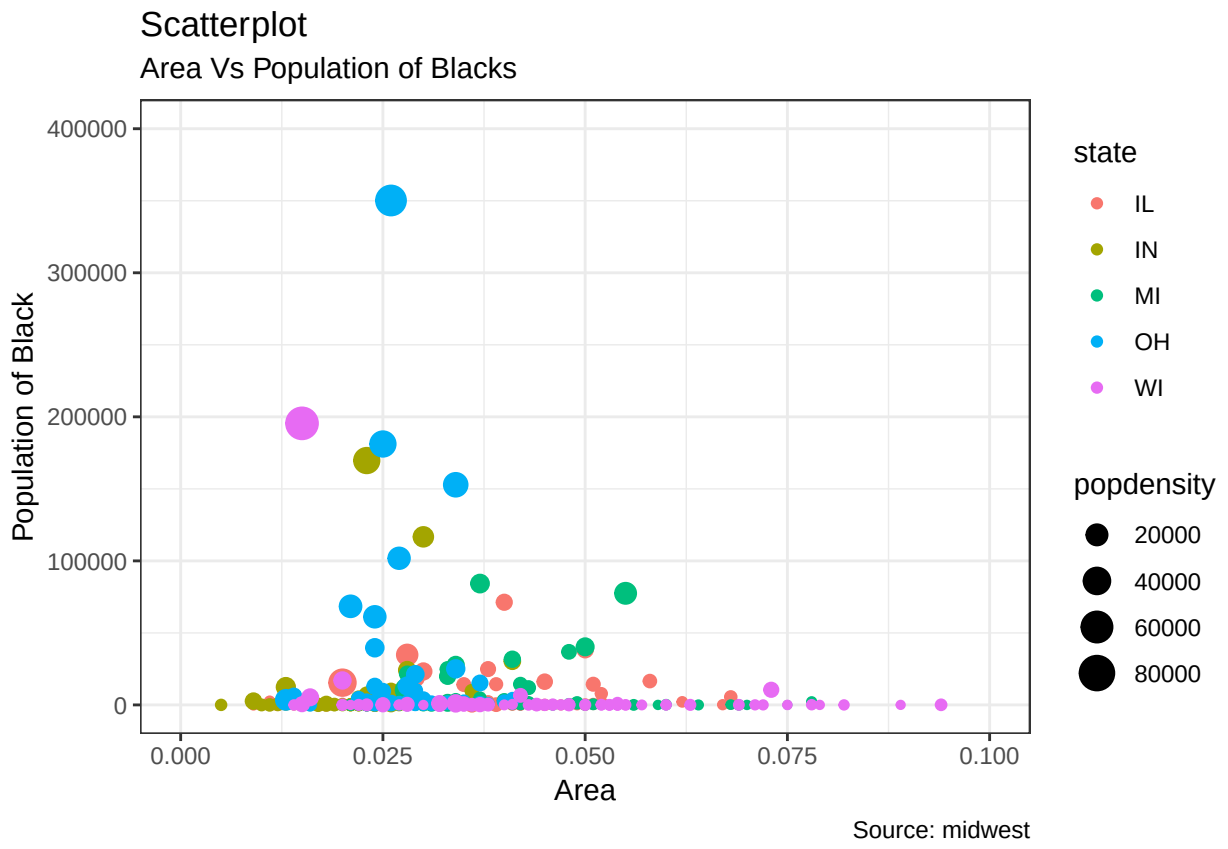
This is the most frequently used plot for data analysis. Whenever you want to visualize the relationship between two variables, the first choice would be the scatterplot.

```
#scatter plot of the inbuilt midwest data


options(scipen = 999) # Turns-off scientific notation
kable(head(midwest, 5))
```

| PID | county | state | area | poptotal | popdensity | popwhite | popblack | popamerindian | popasian |
|---|---|---|---|---|---|---|---|---|---|
| 561 | ADAMS | IL | 0.052 | 66090 | 1270.9615 | 63917 | 1702 | 98 | 249 |
| 562 | ALEXANDER | IL | 0.014 | 10626 | 759.0000 | 7054 | 3496 | 19 | 48 |
| 563 | BOND | IL | 0.022 | 14991 | 681.4091 | 14477 | 429 | 35 | 16 |
| 564 | BOONE | IL | 0.017 | 30806 | 1812.1176 | 29344 | 127 | 46 | 150 |
| 565 | BROWN | IL | 0.018 | 5836 | 324.2222 | 5264 | 547 | 14 | 5 |

```
ggplot(midwest,aes(x = area, y = popblack)) +
            geom_point(aes(col = state, size = popdensity, na.rm=TRUE)) +
            geom_smooth(method = " lm ",na.rm=TRUE) +
            xlim(c(0, 0.1)) +
            ylim(c(0, 400000)) +
            labs(title = "Scatterplot",
                 subtitle = "Area Vs Population of Blacks",
                 y = "Population of Black",
                 x = "Area",
                 caption = "Source: midwest")
```




According to the scatter plot, there is a large population of black in the state Ohio. In all the statet the area occupied by the Black population in the midwest is approximately 0.025. Population density of black population is higher in Ohio and Wichita States. I will get back to this data in a bit. For now, let me jump over to the titanic data.

Most people are familiar with the titanic story.

## Titanic Data

The goal of the project in the now famous kaggle competition is to determine the survival rate of the passengers. The response variable is Survived with 1 if the passenger survived or 0 if they perished. There are 8 different independent variable. In this article, I am not going to build a machine learning model based on the data. I am going down to the very beginning and check for any patterns and trends in the data.

```
titanic.data <- read.csv("train.csv")
datatable(titanic.data, options = list(pageLength = 5))
```

```
str(titanic.data)
```

```
## 'data.frame':    891 obs. of  12 variables:
##  $ PassengerId: int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Survived   : int  0 1 1 1 0 0 0 0 1 1 ...
##  $ Pclass     : int  3 1 3 1 3 3 1 3 3 2 ...
##  $ Name       : Factor w/ 891 levels "Abbing, Mr. Anthony",..: 109 191 358 277 16 559 520 629
##  $ Sex        : Factor w/ 2 levels "female","male": 2 1 1 1 2 2 2 2 1 1 ...
##  $ Age        : num  22 38 26 35 35 NA 54 2 27 14 ...
##  $ SibSp      : int  1 1 0 1 0 0 0 3 0 1 ...
##  $ Parch      : int  0 0 0 0 0 0 0 1 2 0 ...
##  $ Ticket     : Factor w/ 681 levels "110152","110413",..: 524 597 670 50 473 276 86 396 345
##  $ Fare       : num  7.25 71.28 7.92 53.1 8.05 ...
##  $ Cabin      : Factor w/ 148 levels "","A10","A14",..: 1 83 1 57 1 1 131 1 1 1 ...
##  $ Embarked   : Factor w/ 4 levels "","C","Q","S": 4 2 4 4 4 3 4 4 4 2 ...
```

The independent variables include

- Passenger Id

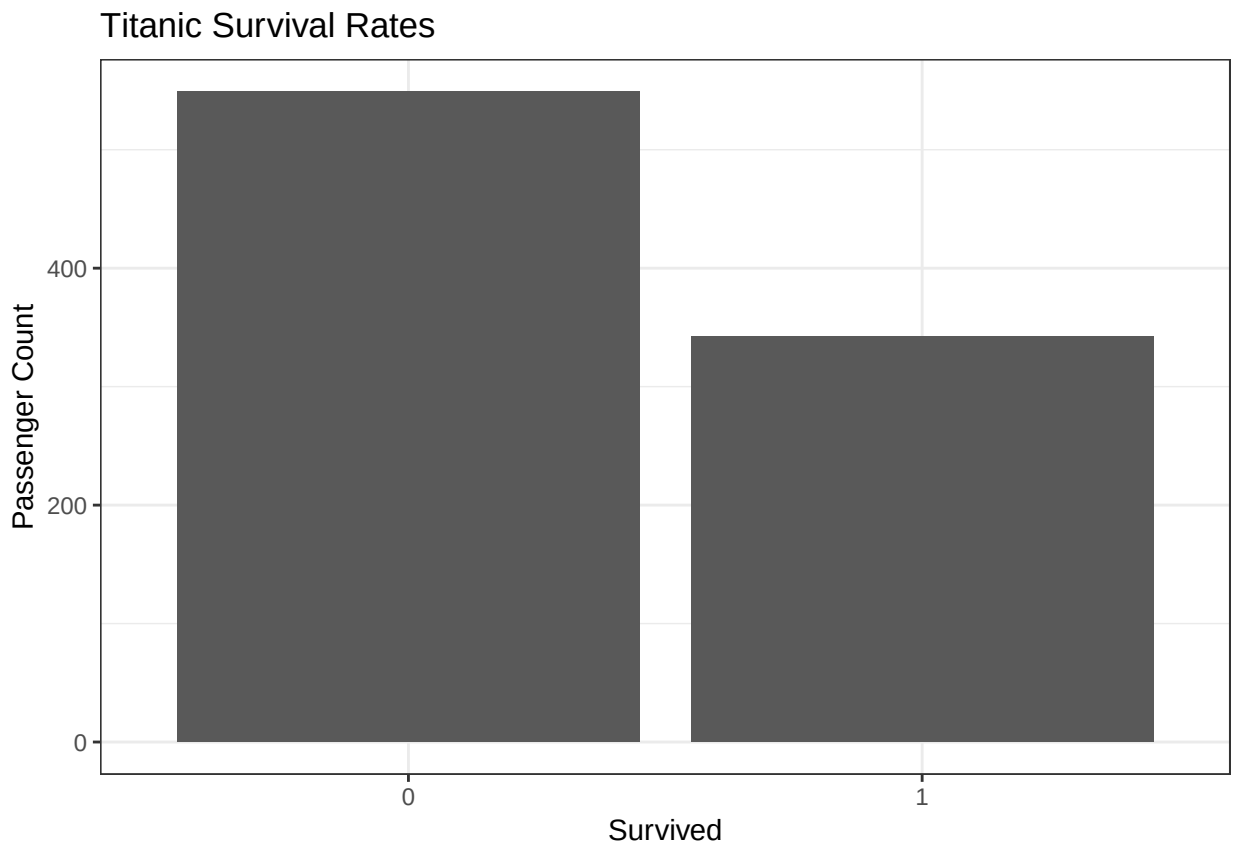
- Pclass : categorical variable 1st, 2nd & 3rd
- Name
- Sex : Categorical variable; Male & Female
- Age : Numerical variable
- SibSp : Number of siblingd and/or spouse on board
- Parch
- Ticket
- Fare
- Carbin
- Embarked: Which port they boarded the ship.

It is important to know the different variable types so that those that are factors cannot be confused for numeric data type.

## Survival Rate

Lets look at the bar graph that shows survival patterns.

```
p <- ggplot(titanic.data, aes(x = Survived)) +
  theme_bw() +
  geom_bar() +
  labs(y = "Passenger Count",
       title = "Titanic Survival Rates")
p
```




```
ggplotly(p)
```

```
# from the bar graph we can tell that more people perished than survived.
prop.table(table(titanic.data$Survived))
```

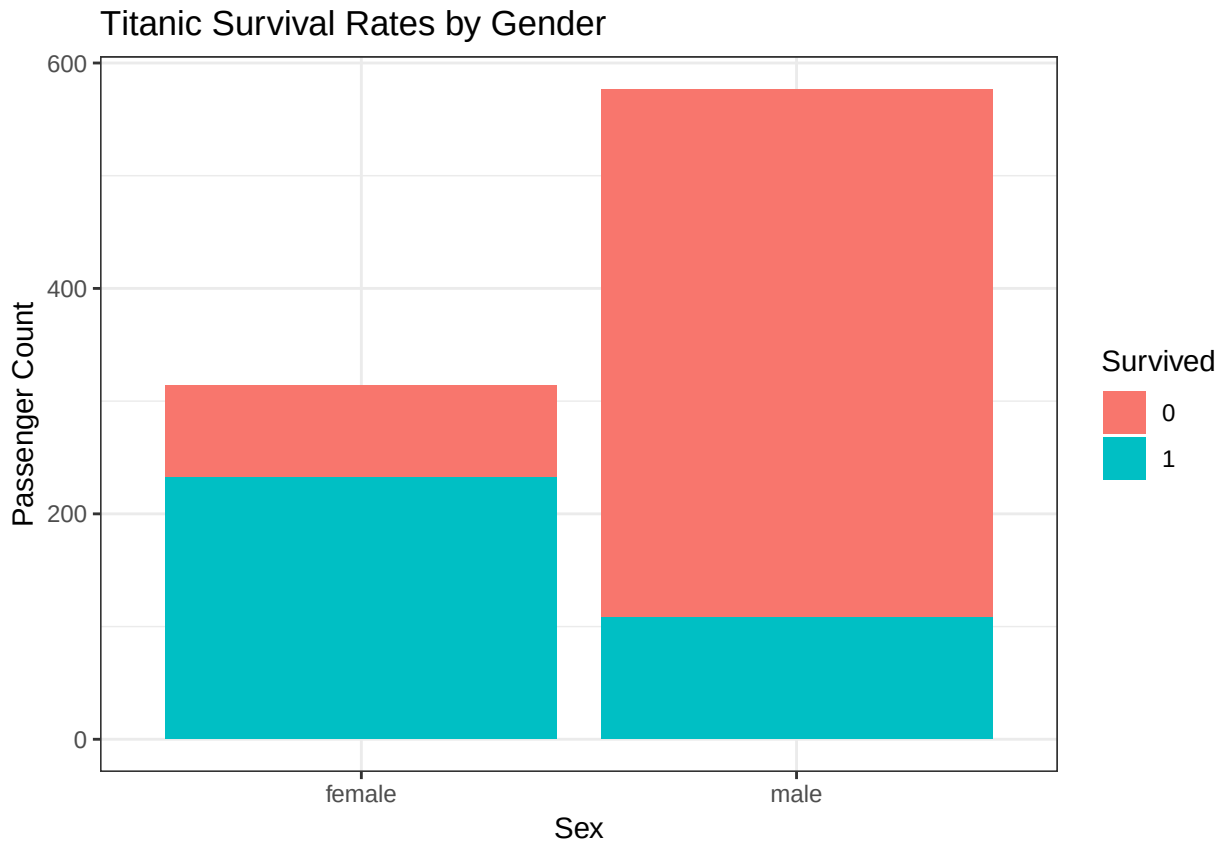
```
##
##         0         1
## 0.6161616 0.3838384
```

```
#Close to 61 % of the passengers perished while about 38% survived
```

## Survival Rate by Gender

Did Gender play a role? The adage "Women and Children First" is greatly used whenever we talk about the Titanic tragedy. It is reasonable to look at survival rate based on gender.

```
p <- ggplot(titanic.data, aes(x = Sex, fill = Survived)) +
  theme_bw() +
  geom_bar() +
  labs(y = "Passenger Count",
       title = "Titanic Survival Rates by Gender")
p
```



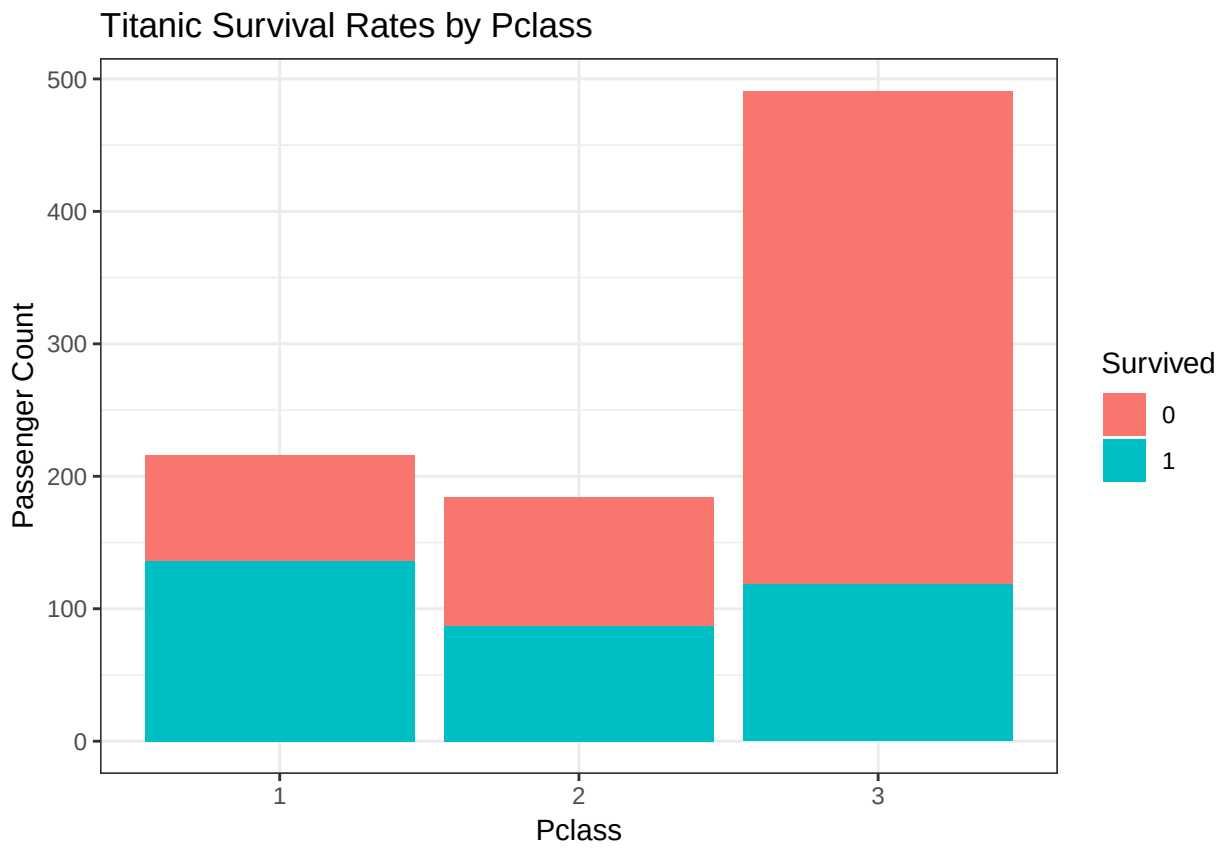

```
ggplotly(p)
```

From the output we can see that most men were in the Titanic. However, most men perished did not survive. Furthermore, there were fewer Females in the Titanic but most of them survived. Clearly, Gender played a role in the survival of the Titanic.

## Survival Rate by Class of Ticket

Did class of ticket play a role in Survival? 1st class were located higher in the titanic than 3rd class. As a data storyteller, we may be interested to know if class of ticket played a role in the survival patterns.

```
p <- ggplot(titanic.data, aes(x = Pclass, fill = Survived)) +
  theme_bw() +
  geom_bar() +
  labs(y = "Passenger Count",
       title = "Titanic Survival Rates by Pclass")
p
```

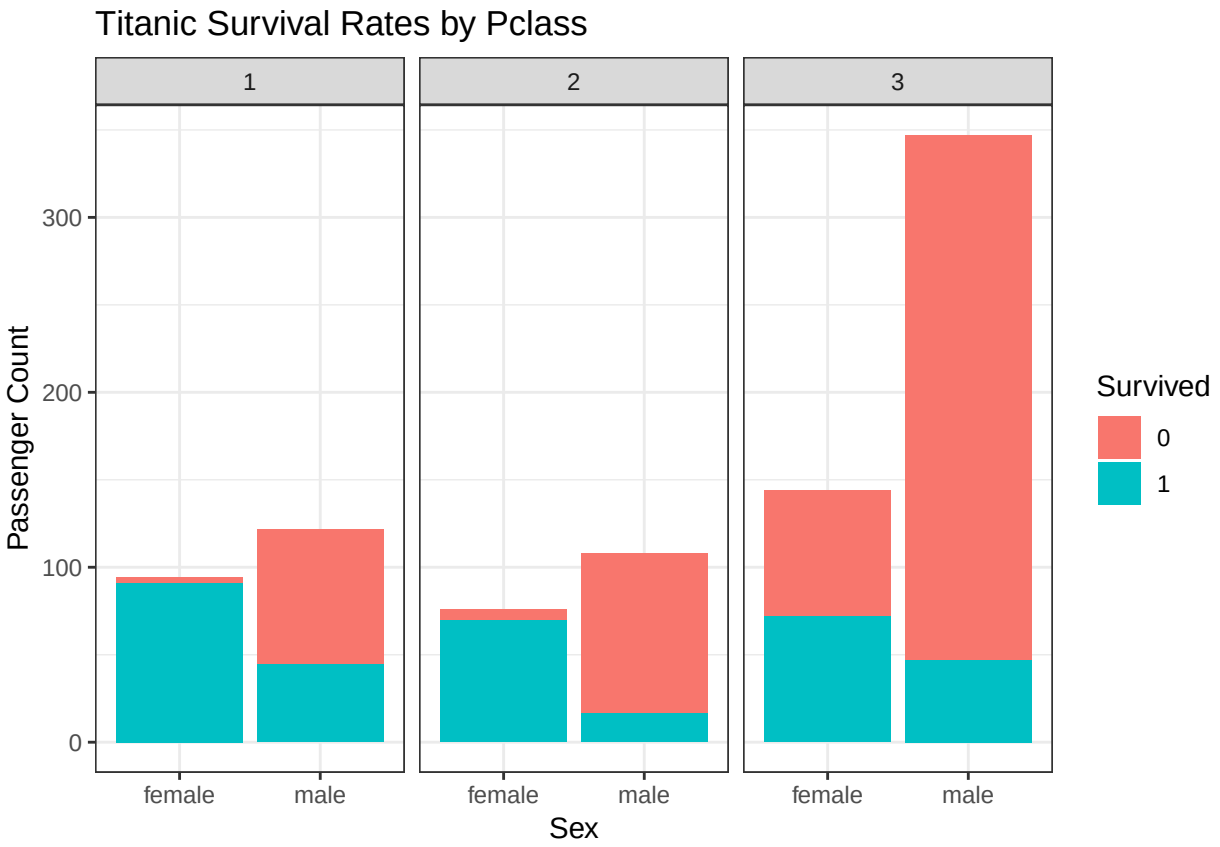


```
ggplotly(p)
```

From the plot, there were more passengers in 3rd class. Unfortunatley, most passengers in 3rd class perished. There more than 50% survival in 1st class. It is not suprising that more people who survived were in the 1st class due to their close proximity to life boats and help as compared to 3rd class passengers.

So far it seems that both Gender and Ticket class played a role in the survival chances of the passengers. How about we combine and see how the two independent variables jointly determined the survival patterns.

```
p <- ggplot(titanic.data, aes(x = Sex, fill = Survived)) +
  theme_bw() +
  facet_wrap(~Pclass) +
  geom_bar() +
  labs(y = "Passenger Count",
       title = "Titanic Survival Rates by Pclass")
p
```




```
ggplotly(p)
```

Females in the 1st class overwhelmingly survived with a percentage of survival at approximately 97%. The same would be said about the in the 2nd class. In the 3rd class, there was a 50% chance of female survival. Males had much lower survival rates across the different classes. Most men in 3rd class had lower chances of survival.

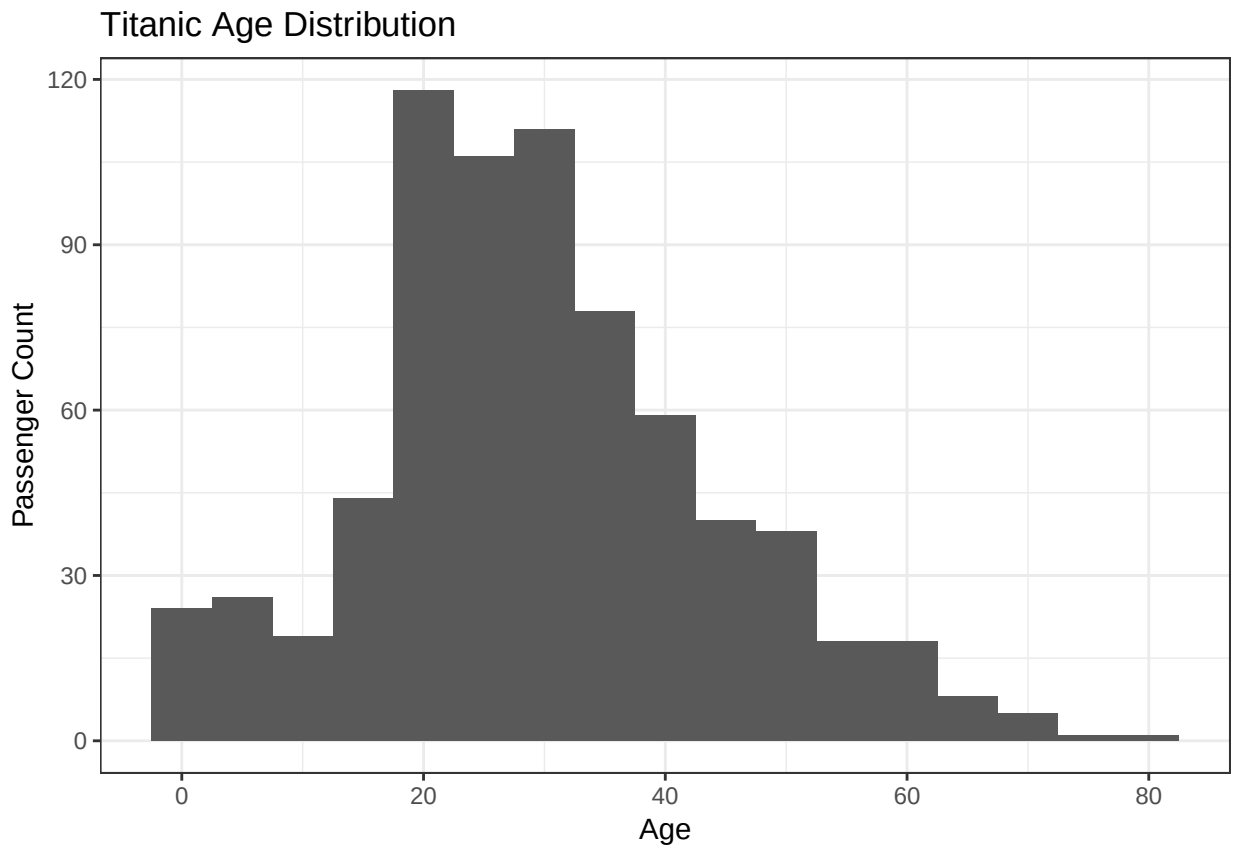
## Survival Rate by Age

We may be interested in knowing the range of age of the passengers in the titanic. In particular we want to know if age played a role on survival chances.

```
p <- ggplot(titanic.data, aes(x = Age)) +
  theme_bw() +
  geom_histogram(binwidth = 5) +
```

```
  labs(y = "Passenger Count",
       x = "Age",
       title = "Titanic Age Distribution")
p
```

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



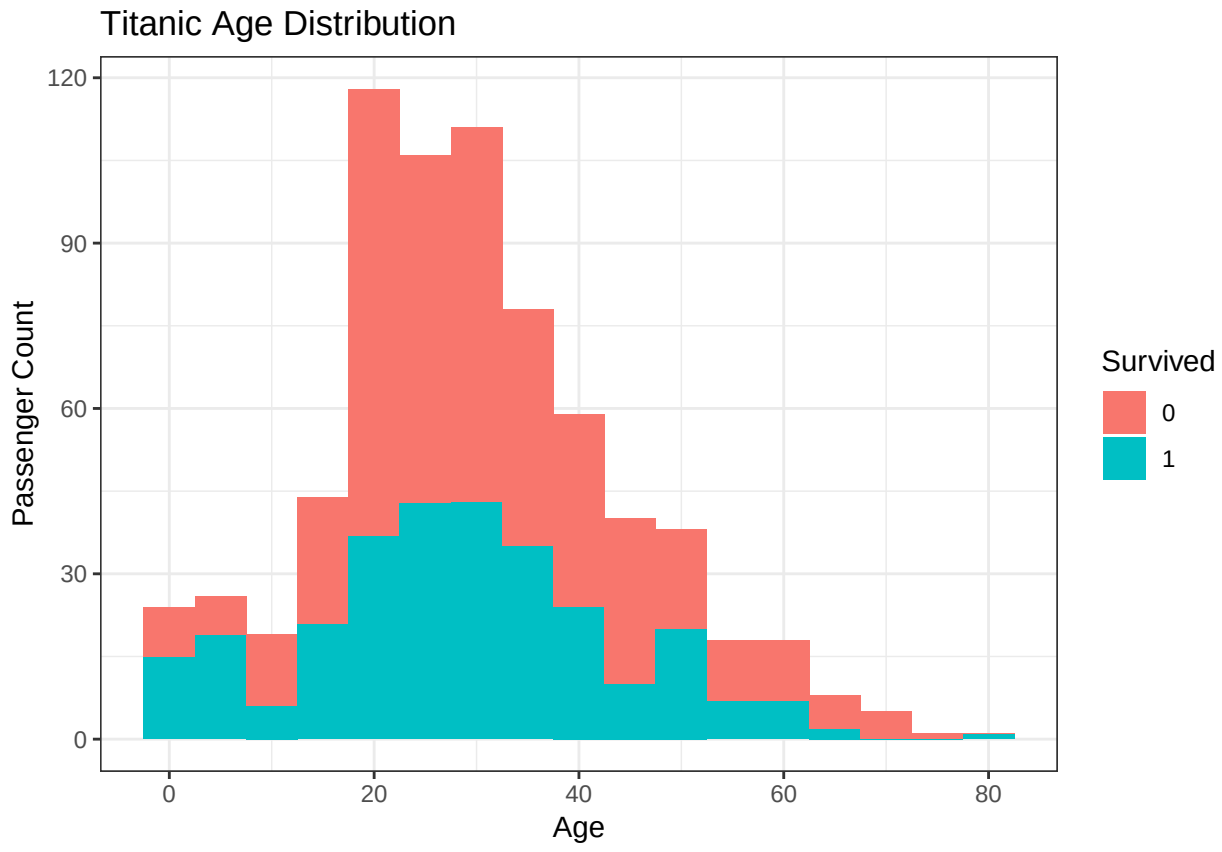

```
ggplotly(p)
```

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

Most passengers are betweeen the ages of 20 - 50 years. There was longer tail to the right for old people up to 80 years. We can also see there were young children from ages between 0 - 4 years as well. The next logical question to ask is of these different age brackets, what's the proportion of those who survived. Especially I am (just me) interested to of the young children old people, how many of them survived.

```
p <- ggplot(titanic.data, aes(x = Age, fill = Survived)) +
  theme_bw() +
  geom_histogram(binwidth = 5) +
  labs(y = "Passenger Count",
       x = "Age",
       title = "Titanic Age Distribution")
p
```

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




```
ggplotly(p)
```

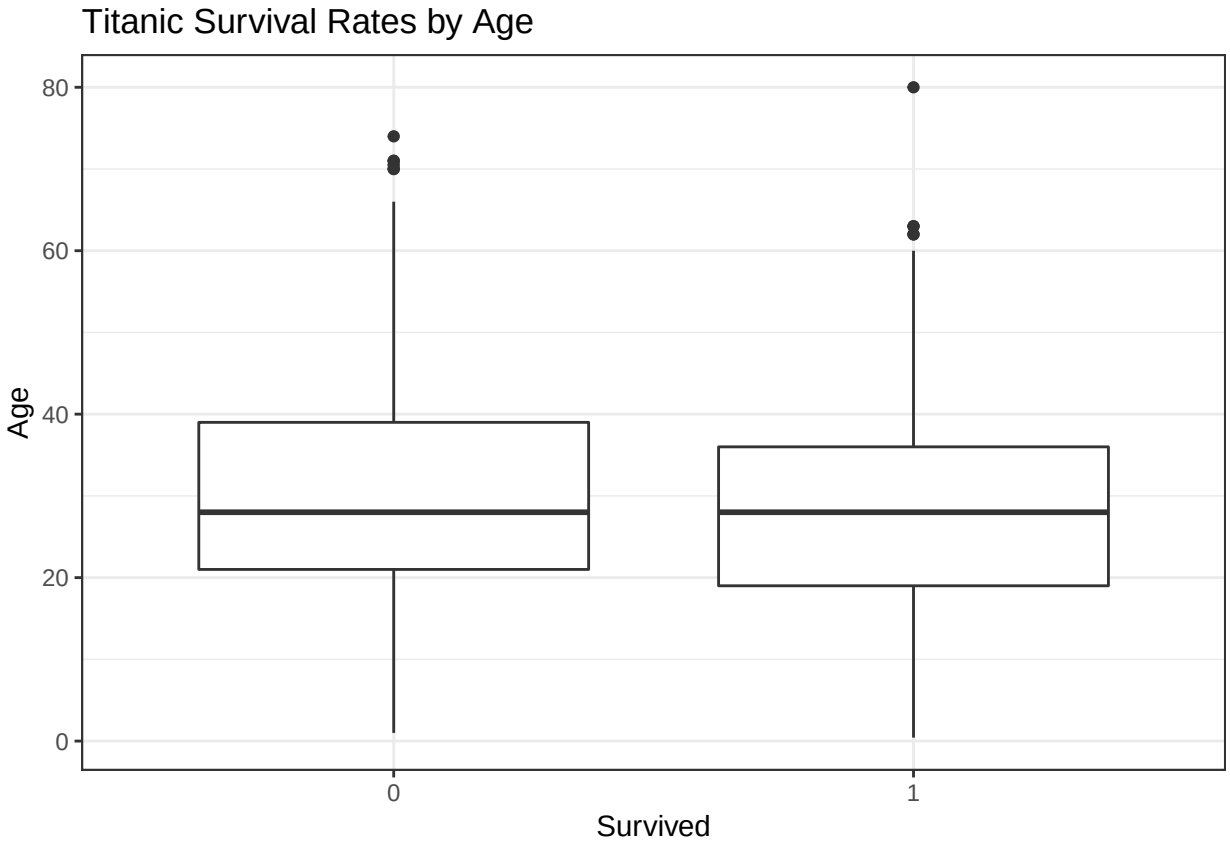
```
## Warning: Removed 177 rows containing non-finite values (stat_bin).
```

In this particular visualization, we can see that more than 50 % of children survived. For the old folks greater than 50 years had a low chances of survival. In general those who survived tended to

be between the ages of 20 - 50 years. We can further observe this pattern by looking a box and whisker plot.

```
p <- ggplot(titanic.data, aes(x = Survived, y = Age)) +
  theme_bw() +
  geom_boxplot() +
  labs(y = "Age",
       x = "Survived",
       title = "Titanic Survival Rates by Age")
p
```

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



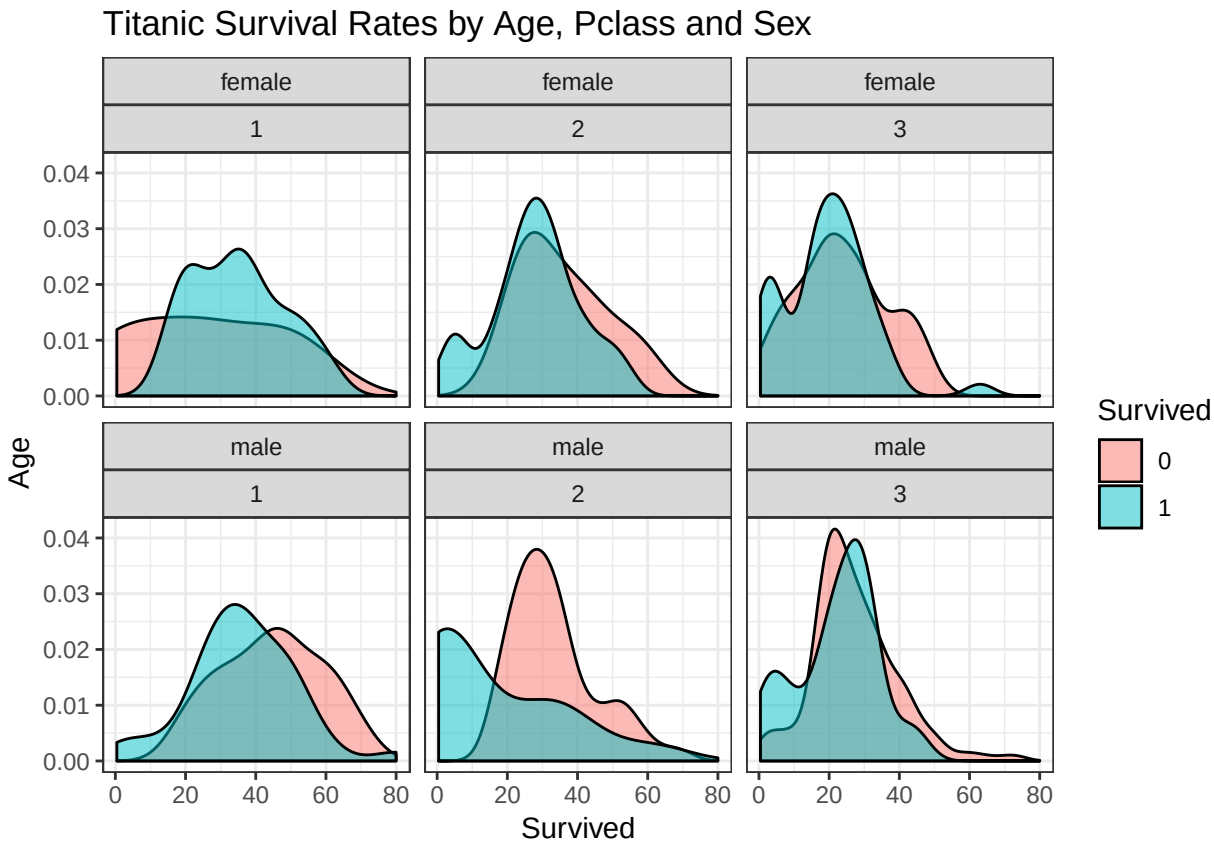
```
ggplotly(p)
```

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

From the Box and Whisker plot, we can see that the oldest survivor was 80 year old and youngest was 0.42 year old. Median age for those who perished and those who survived was 28 year old. The plot further confirms that majority of those who survived and perished are between 20 - 50 years.

```
p <- ggplot(titanic.data, aes(x = Age, fill = Survived)) +
  theme_bw() +
  facet_wrap(Sex ~ Pclass) +
  geom_density(alpha = 0.5) +
  labs(y = "Age",
       x = "Survived",
       title = "Titanic Survival Rates by Age, Pclass and Sex")
p
```

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




```
ggplotly(p)
```

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

From the density plot, we can see that a combination of Age, Gender and ticket class.

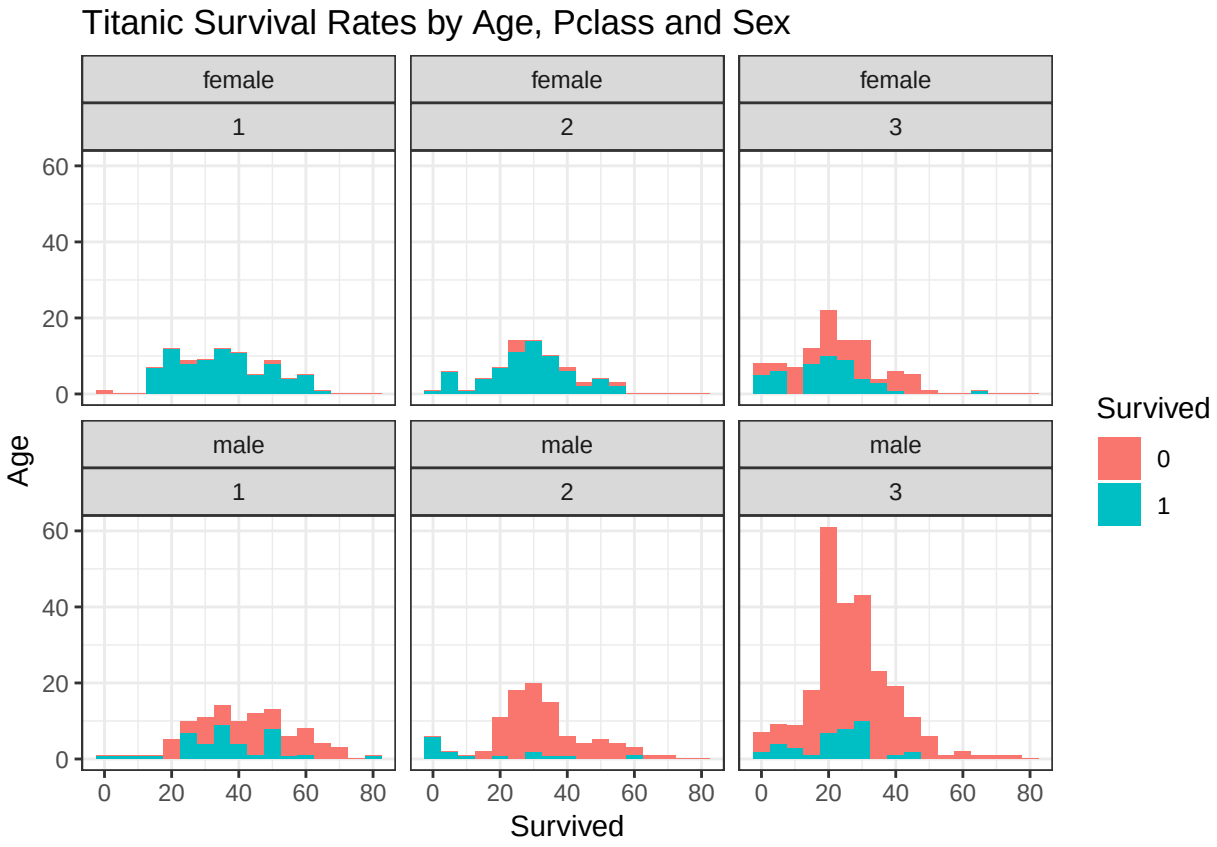
```
p <- ggplot(titanic.data, aes(x = Age, fill = Survived)) +
  theme_bw() +
  facet_wrap(Sex ~ Pclass) +
  geom_bar(binwidth = 5) +
  labs(y = "Age",
       x = "Survived",
       title = "Titanic Survival Rates by Age, Pclass and Sex")
```

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

```
## `geom_histogram()` instead.
```

```
p
```

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




```
ggplotly(p)
```

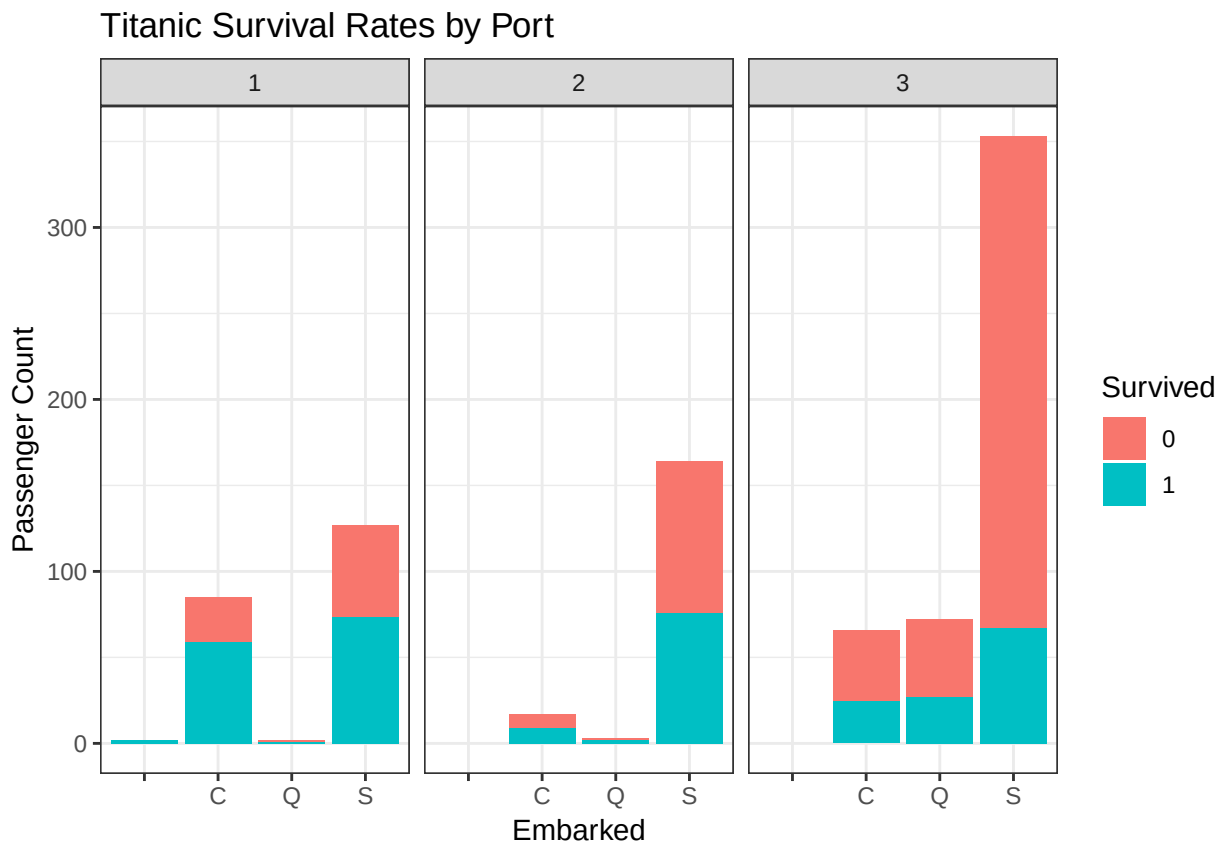
```
## Warning: Removed 177 rows containing non-finite values (stat_bin).
```

The bar graphs tell a similar story and much better story than the density plot. Most females in 1st and 2nd class survived. There was disproportionate survival of the females in the 3rd class. Older males in 2nd and 3rd class had lower chances of survival. All young males in first class survived.

## Survival by Port of Embarking

There were three ports of embarking and the goal as data storyteller is whether these may tell us anything about the survival chances.

```
p <- ggplot(titanic.data, aes(x = Embarked, fill = Survived)) +
  theme_bw() +
  facet_wrap(~Pclass) +
  geom_bar() +
  labs(y = "Passenger Count",
       title = "Titanic Survival Rates by Port")
p
```



```
ggplotly(p)
```

Most passenger boarded from Port of Southampton. However, there is no clear patterns in terms of the survival patterns based on Port of entry

## Conclusion

Data visualization in ggplot2 is a skill that every data storyteller or data scientist should possess. I am glad I was able to write about this visualization article after researching and looking out for content in the world of R data storytellers.