Analytics of Business Intelligence

Chapter # 3 - Exploratory Data Analysis

Dr. Wajahat Gilani

Rutgers Business School

April 20, 2020

Exploratory Data Analysis

Exploratory data analysis is a term attributyed to the statistician John Tukey in his book *Exploratory Data Analysis*. Exploratory data analysis means examining a dataset to discover its underlying characteristics with an emphasis on visualization. The point is that it helps you during analysis design to determine if you should gather more data, suggest hypotheses to test, and identify models to develop. In this chapter, we will cover the following four topics related to exploratory data analysis:

- Understanding exploratory data analysis
- Analyzing a single data variable
- Analyzing two variables together
- Exploring multiple variables simultaneously

We will be using the text file Ch3_marketing.csv.

Understanding Exploratory Data Analysis

> *"An approximate answer to the right question is worth a great deal more than a precise answer to the wrong question."*

- John Tukey

Questions matter, but how do you know what questions to answer. We saw when doing basic summary analysis some things stood out, generating natural questions.

Analyzing/Adjusting The Data

```
Classes 'data.table' and 'data.frame':  172 obs. of  7 variables:
 $ google_adwords : num  65.7 39.1 174.8 34.4 78.2 ...
 $ facebook       : num  47.9 55.2 52 62 40.9 ...
 $ twitter        : num  52.5 77.4 68 86.9 30.4 ...
 $ marketing_total: num  166 172 295 183 150 ...
 $ revenues       : num  39.3 38.9 49.5 40.6 40.2 ...
 $ employees      : int  5 7 11 7 9 3 10 6 6 4 ...
 $ pop_density    : chr  "High" "Medium" "Medium" "High" ...
 - attr(*, ".internal.selfref")=<externalptr>
```

```
1 mark=copy(Ch3_marketing)
2 library(data.table)
3 setDT(mark)
4
5 mark[,pop_density:=factor(pop_density,levels=c('Low','
      Medium','High'),ordered = T)]
```

Analyzing/Adjusting The Data

```
Classes ‘data.table’ and 'data.frame':  172 obs. of  7 variables:
 $ google_adwords : num  65.7 39.1 174.8 34.4 78.2 ...
 $ facebook       : num  47.9 55.2 52 62 40.9 ...
 $ twitter        : num  52.5 77.4 68 86.9 30.4 ...
 $ marketing_total: num  166 172 295 183 150 ...
 $ revenues       : num  39.3 38.9 49.5 40.6 40.2 ...
 $ employees      : int  5 7 11 7 9 3 10 6 6 4 ...
 $ pop_density    : Ord.factor w/ 3 levels "Low"<"Medium"<..: 3 2 2 3 1 3 1 3 2 1 ...
 - attr(*, ".internal.selfref")=<externalptr>
```

- Expenditures for Google AdWords advertising
- Expenditures for Facebook advertising
- Expenditures for Twitter advertising
- Total marketing budget
- Revenues associated with that specific facility
- Number of employees
- Market as defined by population density (Low, Medium, High)

5-Number Summary (Including Mean)

```
1 summary(mark$google_adwords)
```

or

```
1 mark[,summary(google_adwords)]
```

Min. 1st Qu. Median Mean 3rd Qu. Max.
23.65 97.25 169.47 169.87 243.10 321.00

```
1 mark[,fivenum(google_adwords)]
```

23.650 97.225 169.475 243.160 321.000 5-number summary provides information on central tendency (median), spread (quartiles) and range (minimum and maximum)

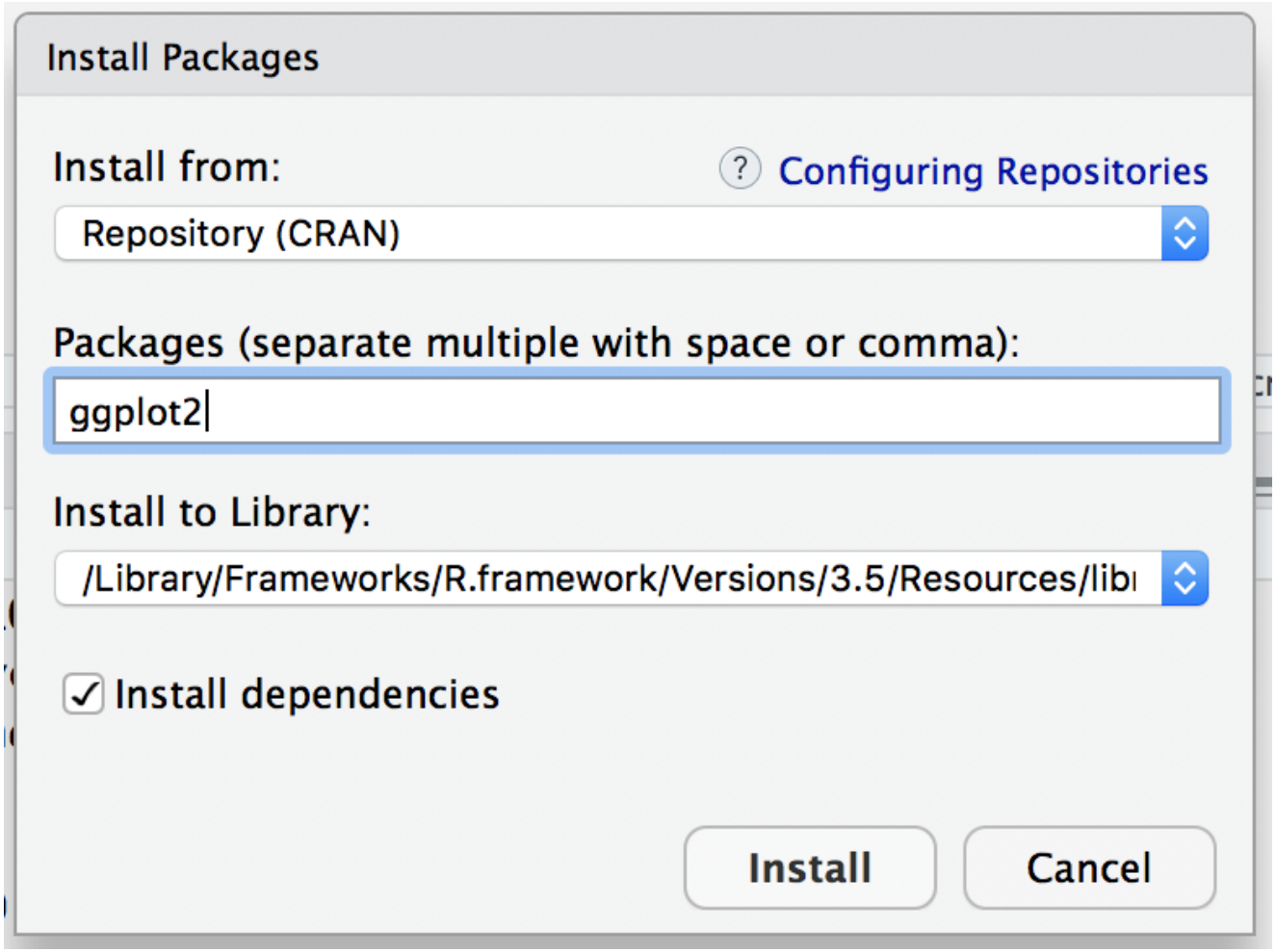
Summary of categorical data

```
1  mark[,summary(pop_density)]
```

Low Medium High
68 52 52

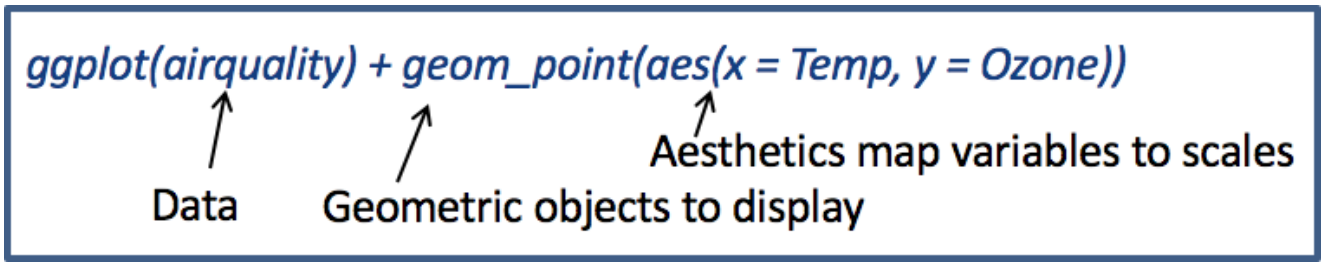
Visualizing Data - ggplot

Even though R has natively built in plot() functions, data analysts tend to use the package ggplot() because its designed to be used for the purpose of visual analytics. Import the package ggplot2:



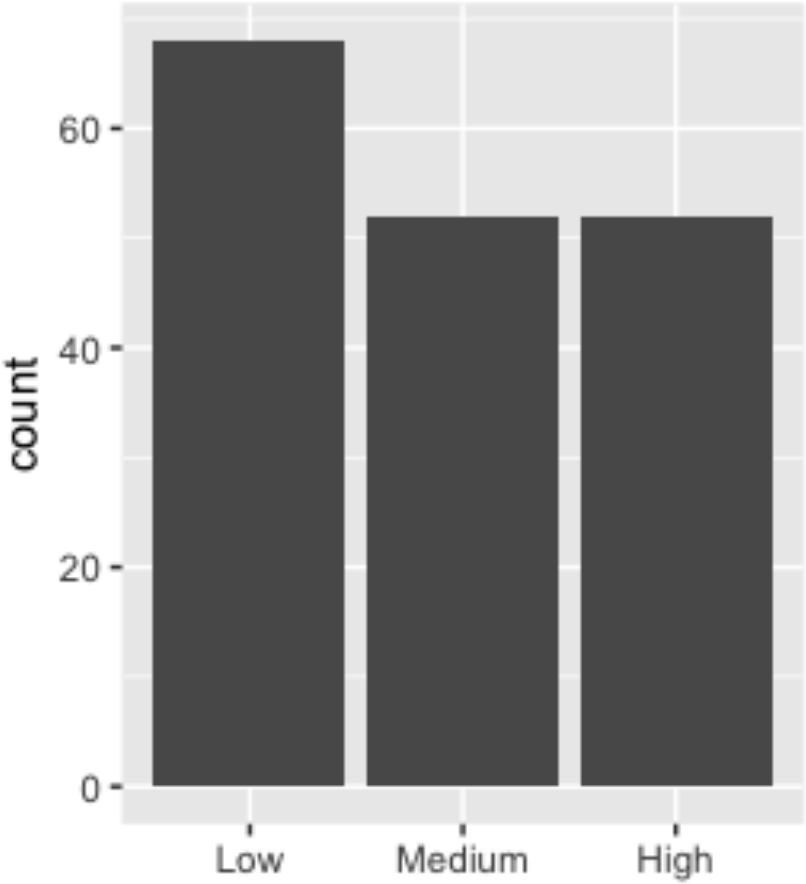
Building a plot in ggplot2

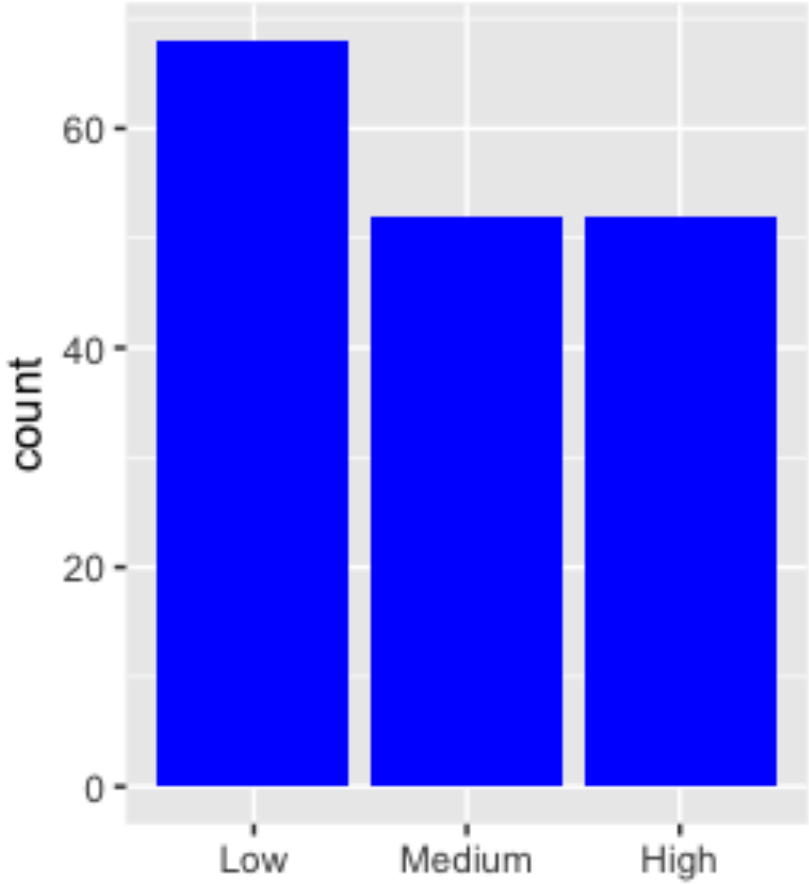
- data to visualize (a data.frame or data.table)
- map variables to aesthetic attributes
- geometric objects – what you see (points, bars, etc)
- statistical transformations – summarize data
- scales map values from data to aesthetic space
- faceting subsets the data to show multiple plots
- coordinate systems put data on plane of graphic



Bar Graph pop_density

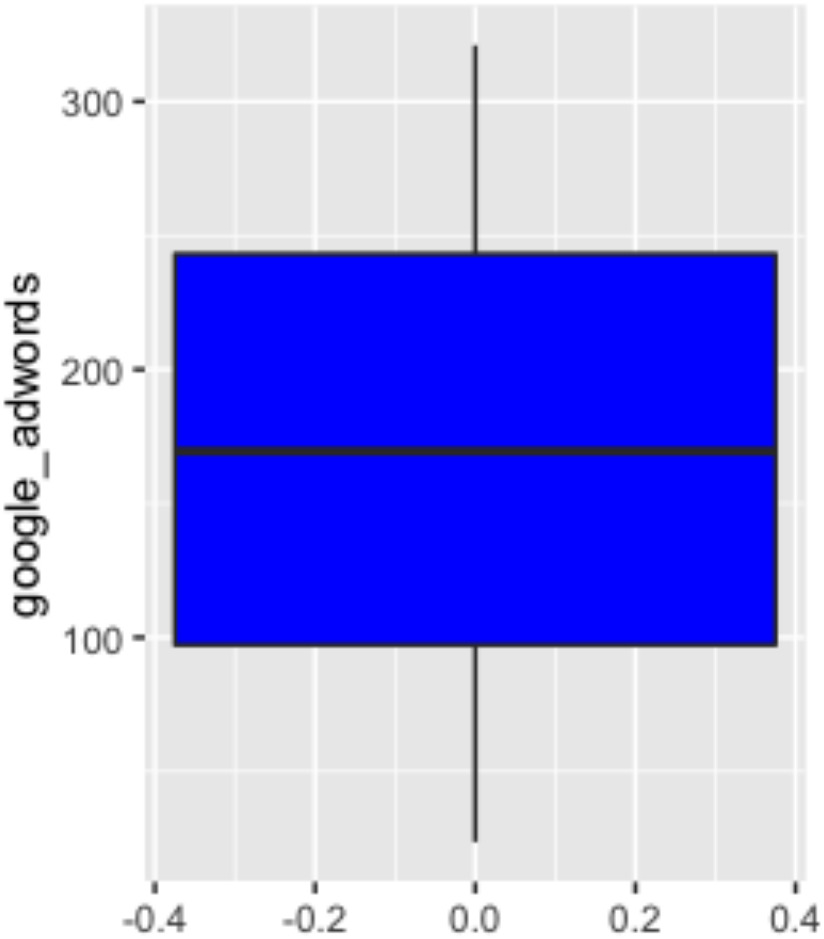
```
1 library(ggplot2)
2
3 ggplot(mark,aes(x=pop_density)) + geom_bar()
4 ggplot(mark,aes(x=pop_density)) + geom_bar(fill='blue'
    )
```

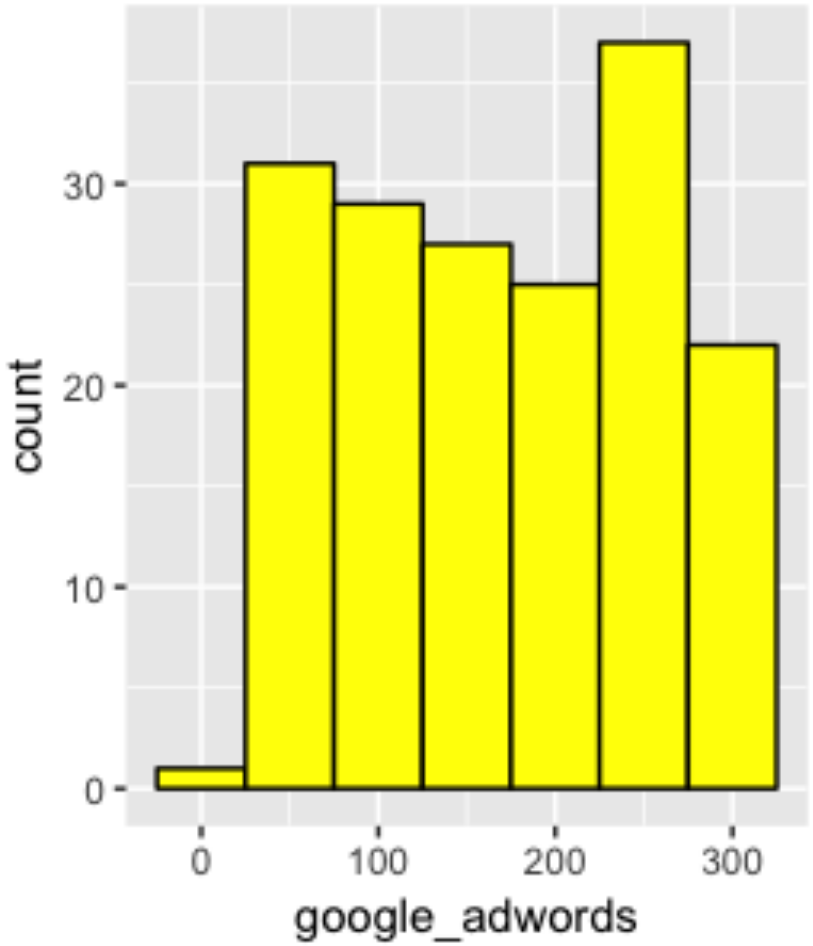




Box Plot - Histogram google_adwords

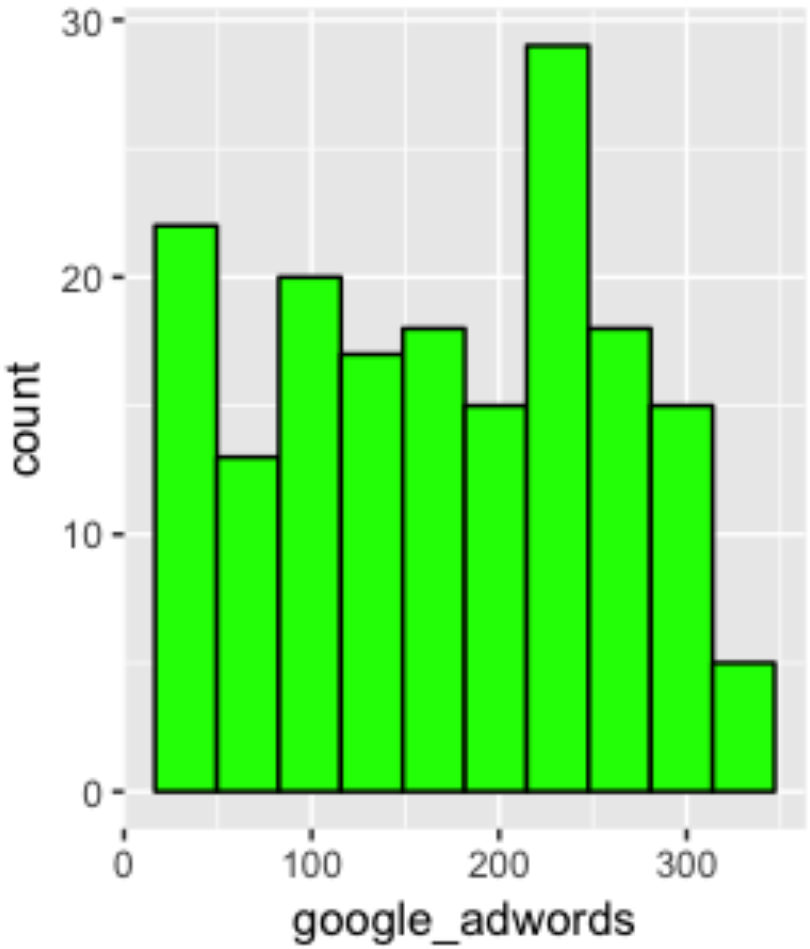
```
1 ggplot(mark,aes(y=google_adwords)) + geom_boxplot(fill
    ='blue')
2 ggplot(mark,aes(x=google_adwords)) + geom_histogram(
    fill='yellow',color='black', binwidth=50)
```

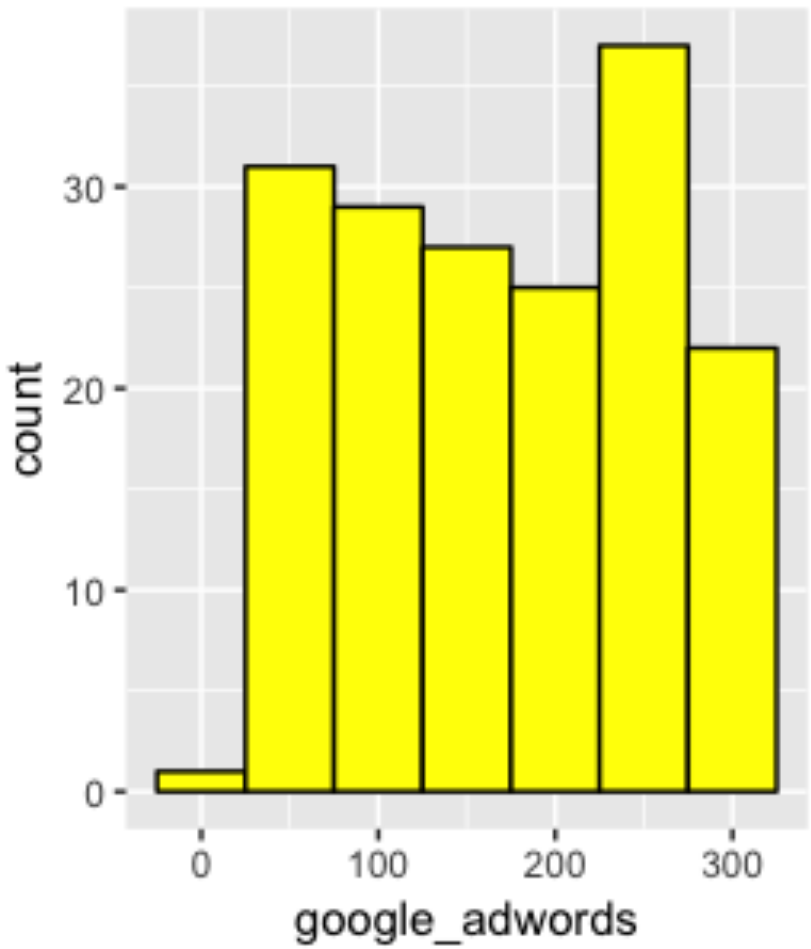




Box Plot - Histogram google_adwords

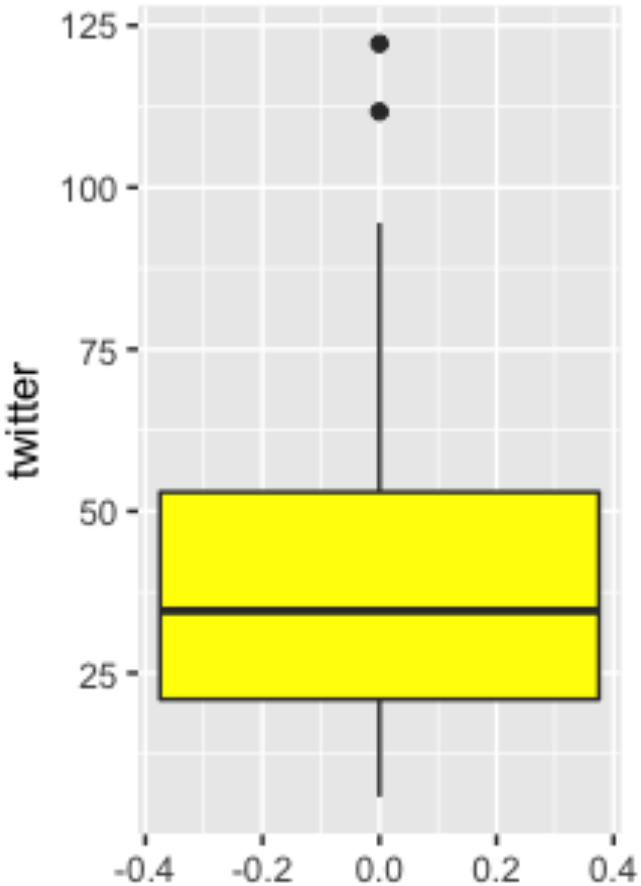
```
1 ggplot(mark,aes(x=google_adwords)) + geom_histogram(
    fill='green',color='black', bins =10)
2 ggplot(mark,aes(x=google_adwords)) + geom_histogram(
    fill='yellow',color='black', binwidth=50)
```

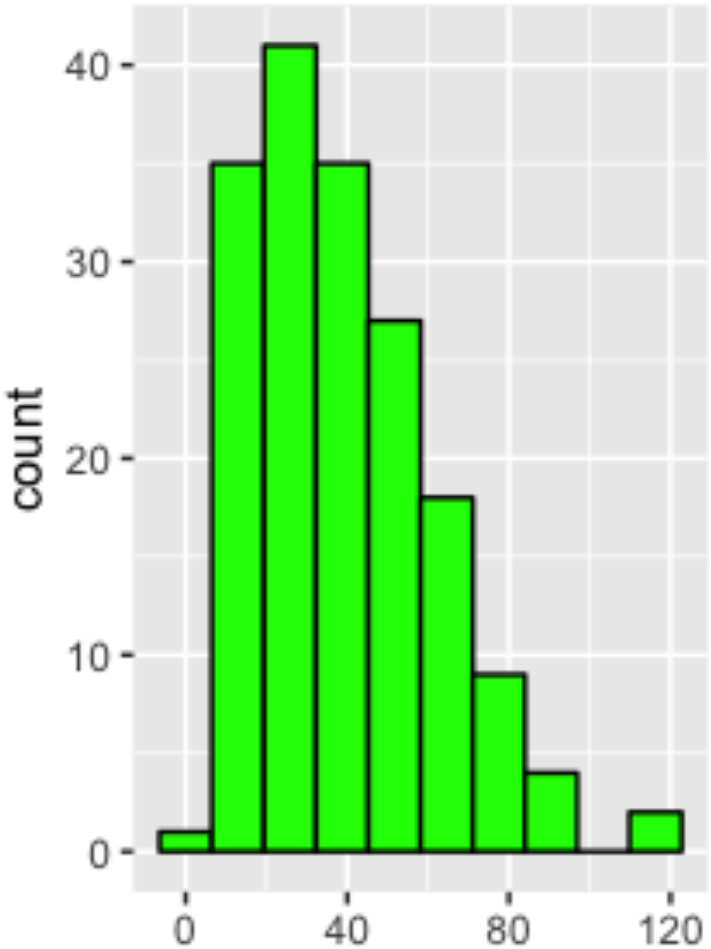




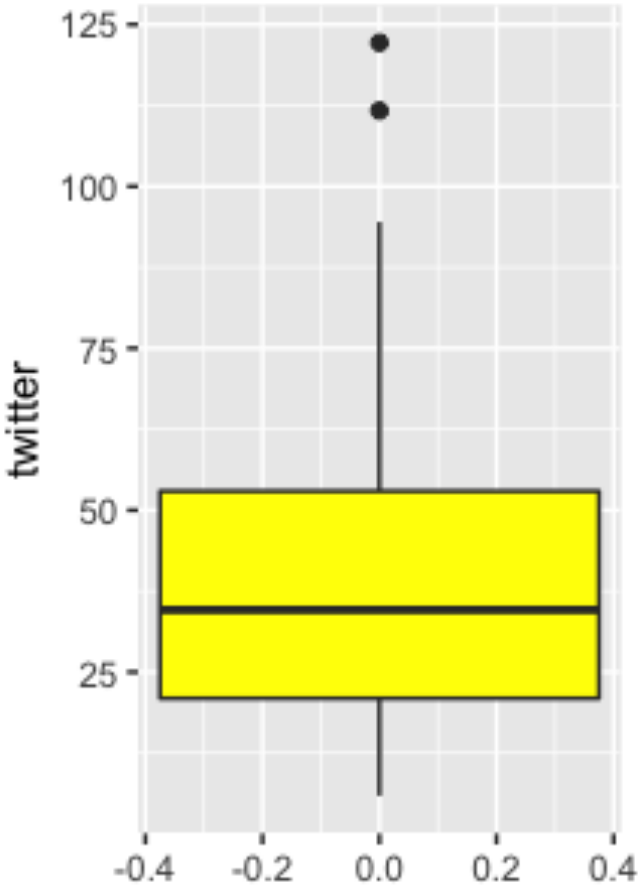
Box Plot - Histogram twitter

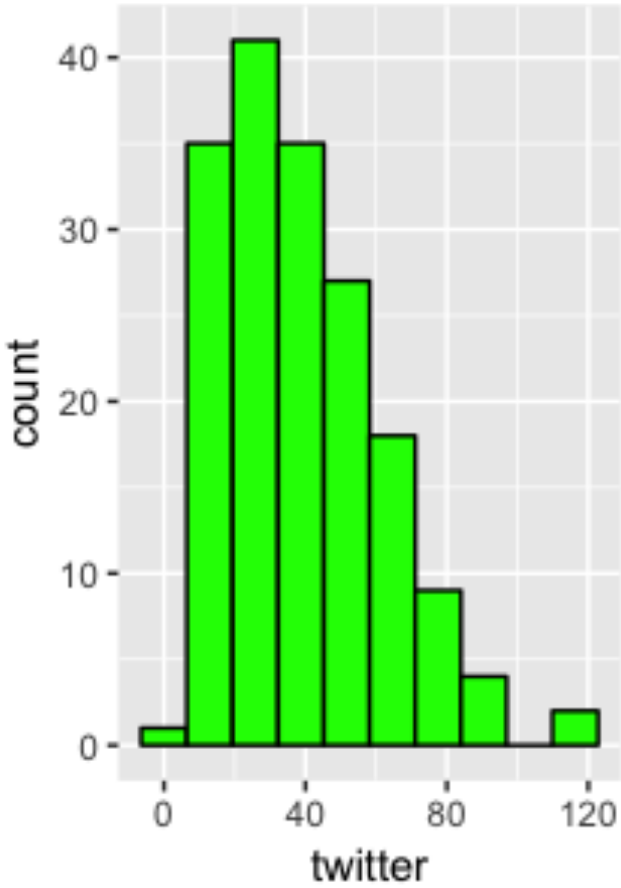
```
1 ggplot(mark,aes(y=twitter)) + geom_boxplot(fill='
    yellow')
2 ggplot(mark,aes(x=twitter)) + geom_histogram(fill='
    green',color='black', bins =10)
```





Box Plot - Histogram twitter





The twitter boxplot has dots above it, that means that the top line is top quartile, and the outlier dot points are the max point. The bar graph shows the outliers as well. You can clearly see that the data is skewed to the right.

Analyzing 2 Variables Together

Sometimes you want to test possible relationships between data, these 4 questions serve as a guide:

- What does the data look like?
- Is there any relationship between two variables?
- Is there any correlation between the two?
- Is the correlation significant?

What if we wanted to test for a relationship between employees and pop_density. One problem is that employees are numeric values while pop_density are categorical values.

To test for this relationship, pop_density has to be converted to categorical data.

Transforming a Numerical Data To Categorical

Create a new column empFactor, that will be the employees column subdivided into groups. We will assume 2 groups. To do this we will use the cut() function. The cut() function subdivides values into groups, decided by the parameter:

```
1 mark[,empFactor:=cut(employees,2)]
```

Since we typed 2, empFactor will have 2 groups (factors):

```
1 table(mark$empFactor)
```

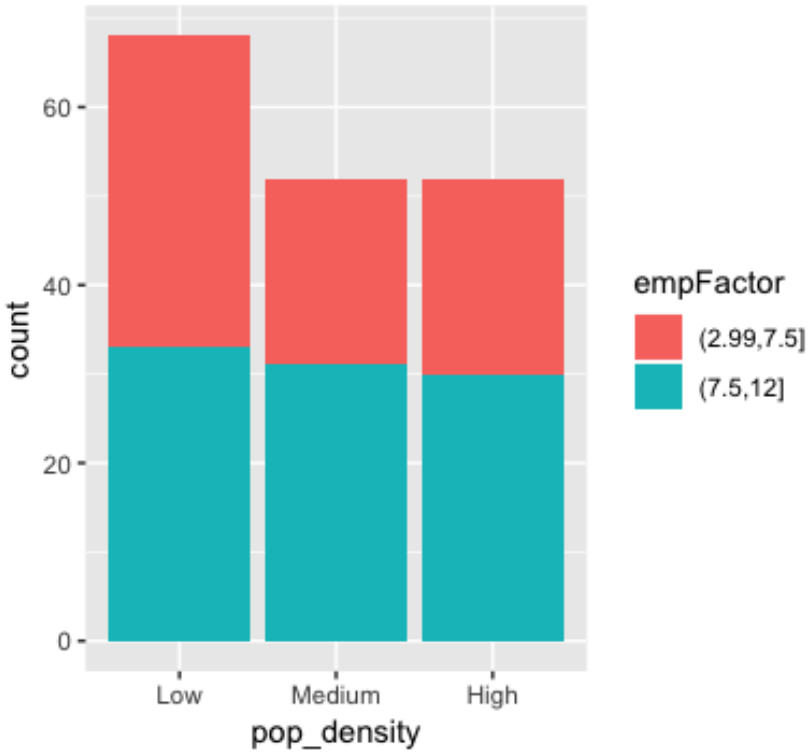
```
(2.99,7.5] (7.5,12]
78         94
```

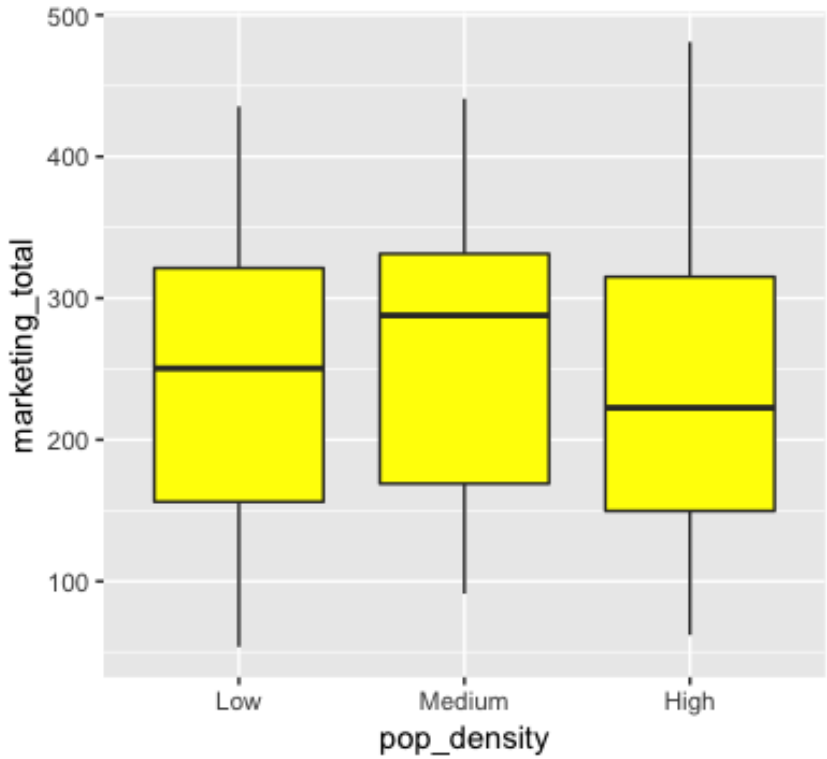
78 rows fall into the 2.9 (exclusive) to 7.5 category and 94 rows fall into the 7.5 (exclusive) to 12 category.

```
1 table(mark$empFactor, mark$pop_density)
```

	Low	Medium	High
(2.99,7.5]	35	21	22
(7.5,12]	33	31	30

Measuring Combinations Visually






```
1 ggplot(mark,aes(x=pop_density,fill=empFactor)) + geom_
    bar()
2
3 ggplot(mark,aes(y=marketing_total,x=pop_density)) +
    geom_boxplot(fill='yellow')
```

Notice the ordering of the box-plot x-axis.

Measuring Combinations Visually - Numerical/Numerical


```
1 ggplot(mark,aes(x=revenues,y=google_adwords)) + geom_
    point(color='purple')
```

What kind of correlation are we seeing? Is it strong?

Correlation

```
1 cor(mark$google_adwords,mark$revenues)
2 mark[,cor(google_adwords,revenues)]
```

0.7662461

- **Sign** will either be positive or negative
- **Value** will be between 0 and 1

But is the correlation significant? In statistical sense, significant means that your correlation is due to circumstances other than random chance. R has the null test for correlation built in:

```
1 cor.test(mark$google_adwords,mark$revenues)
```

```
        Pearson's product-moment correlation

data:  mark$google_adwords and mark$revenues
t = 15.548, df = 170, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.6964662 0.8216704
sample estimates:
      cor
0.7662461
```

Correlation

```
        Pearson's product-moment correlation

data:  mark$google_adwords and mark$revenues
t = 15.548, df = 170, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.6964662 0.8216704
sample estimates:
      cor
0.7662461
```

We will talk about the t-test when we do regression, but essentially the idea here is to test the null and alternative hypothesis's which are:

$\beta = 0$

$\beta \neq 0$

The null hypothesis basically says that nothing is going on and there is no correlation, where as the alternative is stating there is something going on and there seems to be some sort of correlation. The **t-test** is basically asking how surprising is it to see this degree of correlation if the 2 variables truly are not correlated. This requires us to use the t-statistic formula to arrive at a t-statistic score using coefficients and standard errors (linear regression), the answer yielded a score of 15.548.

Correlation

```
        Pearson's product-moment correlation

data:  mark$google_adwords and mark$revenues
t = 15.548, df = 170, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.6964662 0.8216704
sample estimates:
      cor
0.7662461
```

The p-value then tells in a normal probability distribution, how surprising would this result be, and the answer is .0000000000000022
, that is probabilistically a pretty low chance of happening if the null hypothesis were true.

Thats why the results are saying, true correlation is not equal to 0.

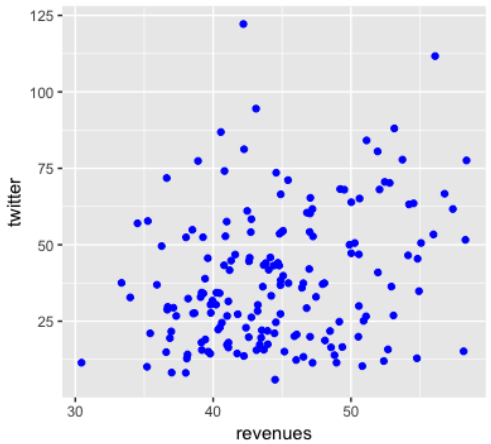
Twitter and Facebook vs Revenues

```
1 cor.test(mark$twitter,mark$revenues)
2 cor.test(mark$facebook,mark$revenues)
3 mark[,.(cor.test(twitter,revenues),cor.test(facebook,
    revenues))]
```

	V1	V2
1:	3.651564	9.230834
2:	170	170
3:	0.000346737	1.05071e-16
4:	0.2696854	0.5778213
5:	0	0
6:	two.sided	two.sided
7:	Pearson's product-moment correlation	Pearson's product-moment correlation
8:	twitter and revenues	facebook and revenues
9:	0.1250993,0.4030549	0.4687127,0.6695639

Visualize ScatterPlots

```
1 ggplot(mark,aes(x=revenues,y=google_adwords)) + geom_
    point(color='purple')
2 ggplot(mark,aes(x=revenues,y=twitter)) + geom_point(
    color='blue')
3 ggplot(mark,aes(x=revenues,y=facebook)) + geom_point(
    color='red')
```



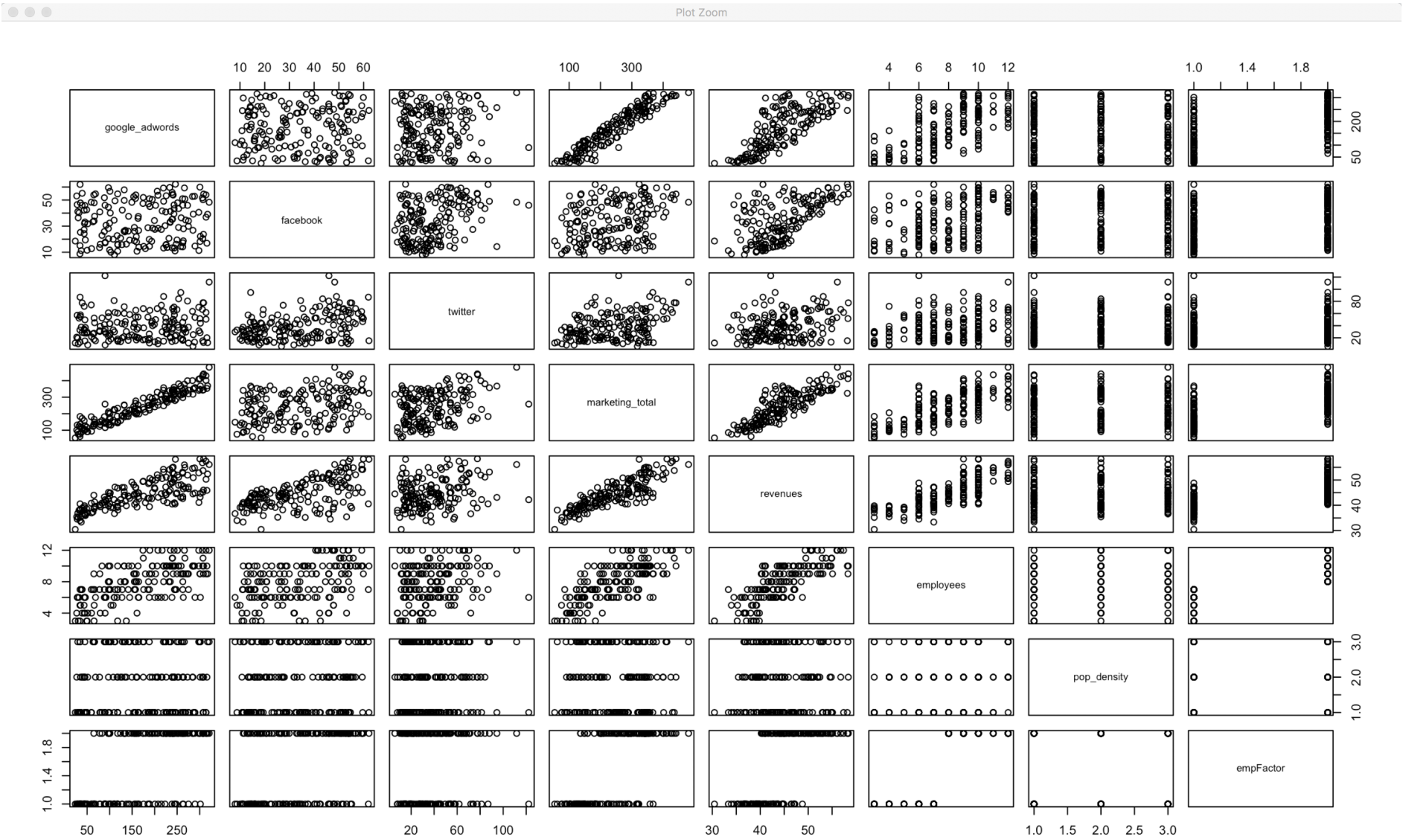
0.7662461

0.2696854

0.5778213

Pairs Function

```
1 pairs(mark)
```



Cor Function

```
1 cor(mark[,1:6])
```

	google_adwords	facebook	twitter	marketing_total	revenues	employees
google_adwords	1.00000000	0.07643216	0.0989750	0.9473566	0.7662461	0.6610312
facebook	0.07643216	1.00000000	0.3543410	0.3102232	0.5778213	0.4101966
twitter	0.09897500	0.35434096	1.0000000	0.3758691	0.2696854	0.2290618
marketing_total	0.94735659	0.31022316	0.3758691	1.0000000	0.8530354	0.7210171
revenues	0.76624608	0.57782131	0.2696854	0.8530354	1.0000000	0.7656857
employees	0.66103123	0.41019661	0.2290618	0.7210171	0.7656857	1.0000000

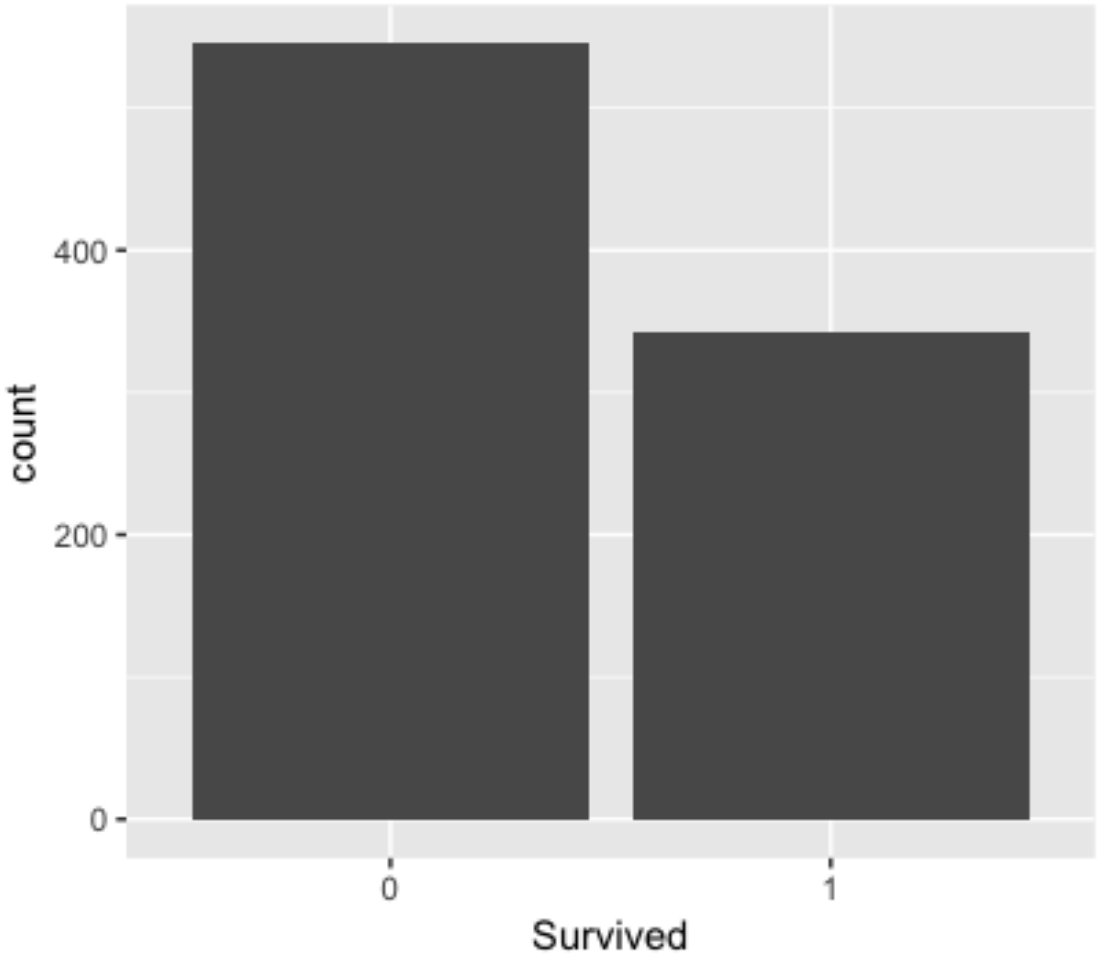
Titanic - Women and Children First?

We want to now do a Visual Analytics example using a famous data-set. Import the titanic dataset into R. We want to go ahead and answer the question if indeed it was women and children first, when the titanic was sinking. We start with viewing the data. As we can see there are 3 columns that have numbers as values, but are clearly categorical in nature: Pclass, Survived, and Sex. Lets convert them into categorical data.

```
1 titanic$Pclass=as.factor(titanic$Pclass)
2 titanic$Survived=as.factor(titanic$Survived)
3 titanic$Sex=as.factor(titanic$Sex)
```

Titanic - View Survived

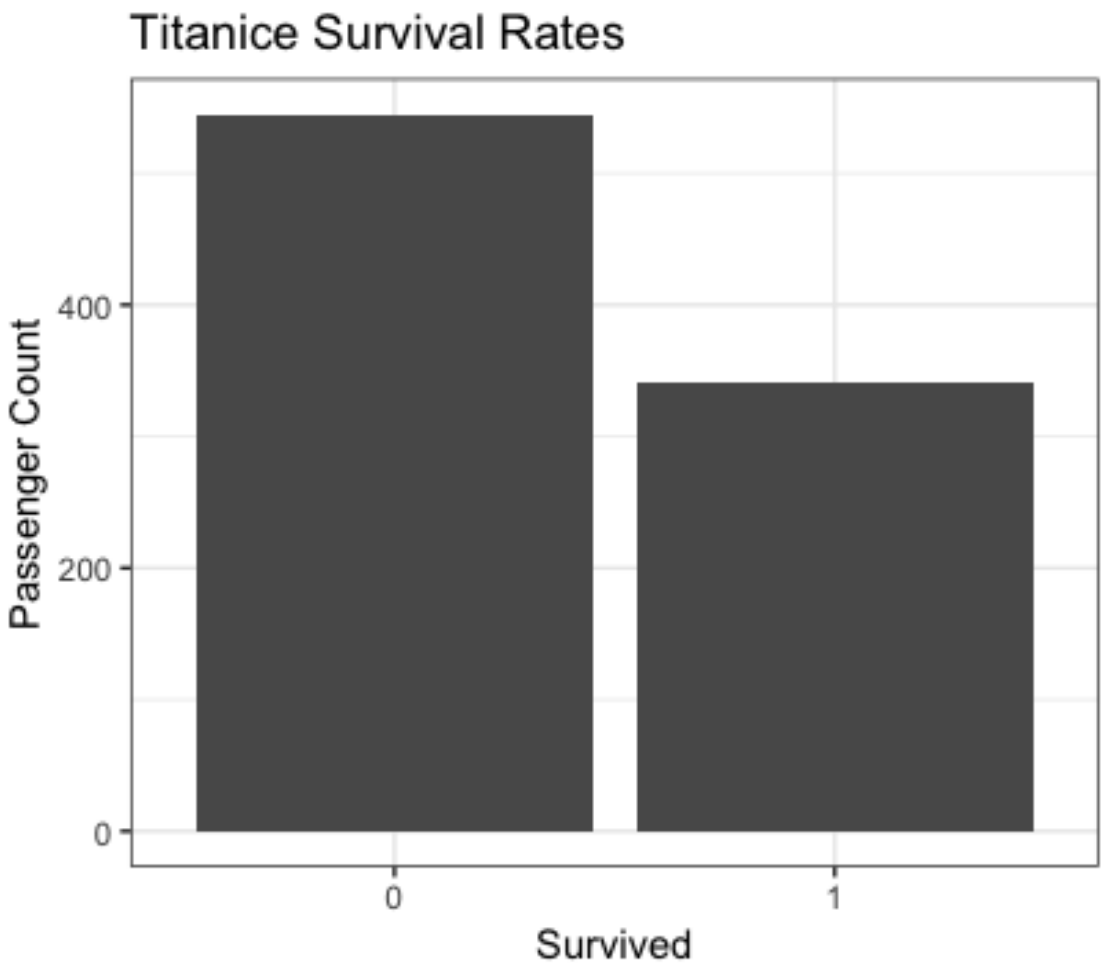
```
1 ggplot(titanic, aes(x=Survived)) + geom_bar()
```



More people died than survived.

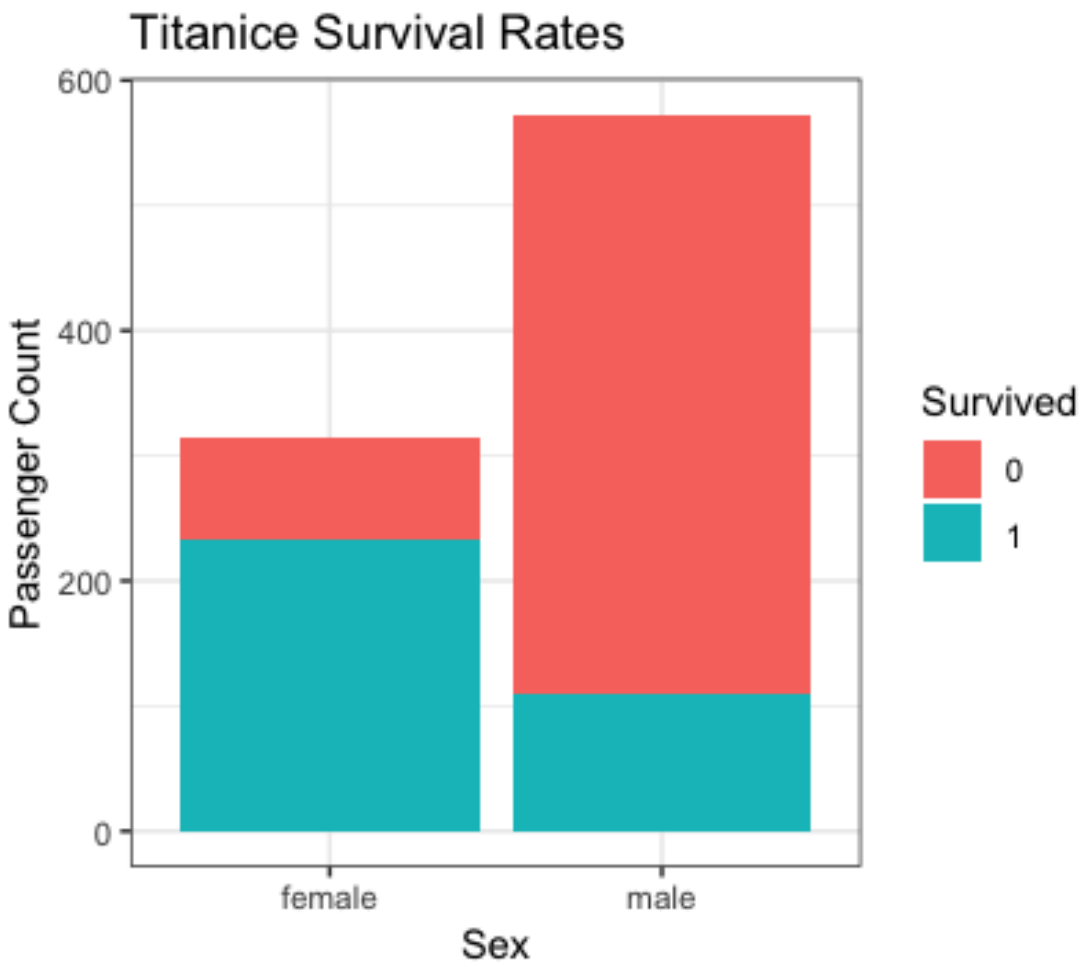
Titanic - View Survived (Nicer)

```
1 ggplot(titanic, aes(x=Survived)) + geom_bar() + theme_
    bw() + labs(y= "Passenger Count", title = "Titanice
     Survival Rates")
```



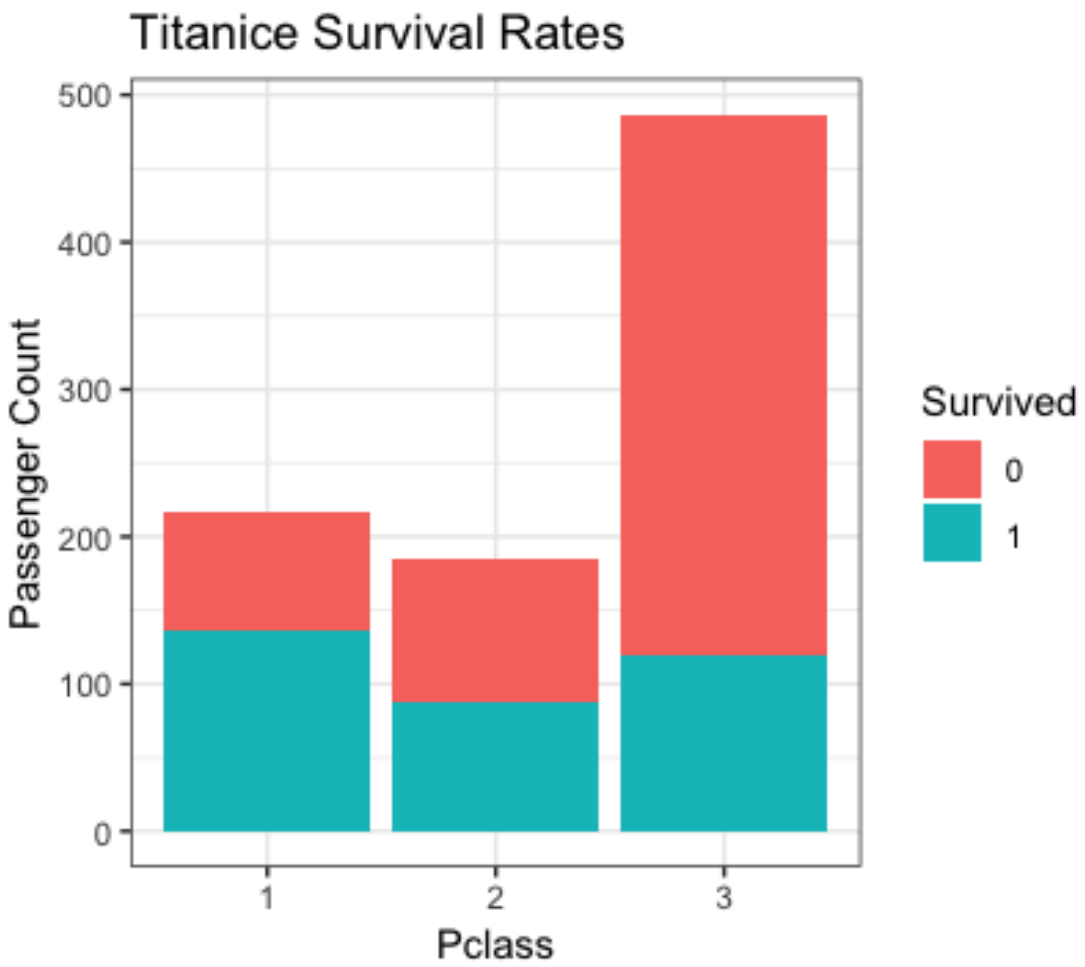
Titanic - Survival Rate By Gender

```
1 ggplot(titanic, aes(x=Sex, fill=Survived)) + geom_bar
    () + theme_bw() + labs(y= "Passenger Count", title
    = "Titanice Survival Rates")
```



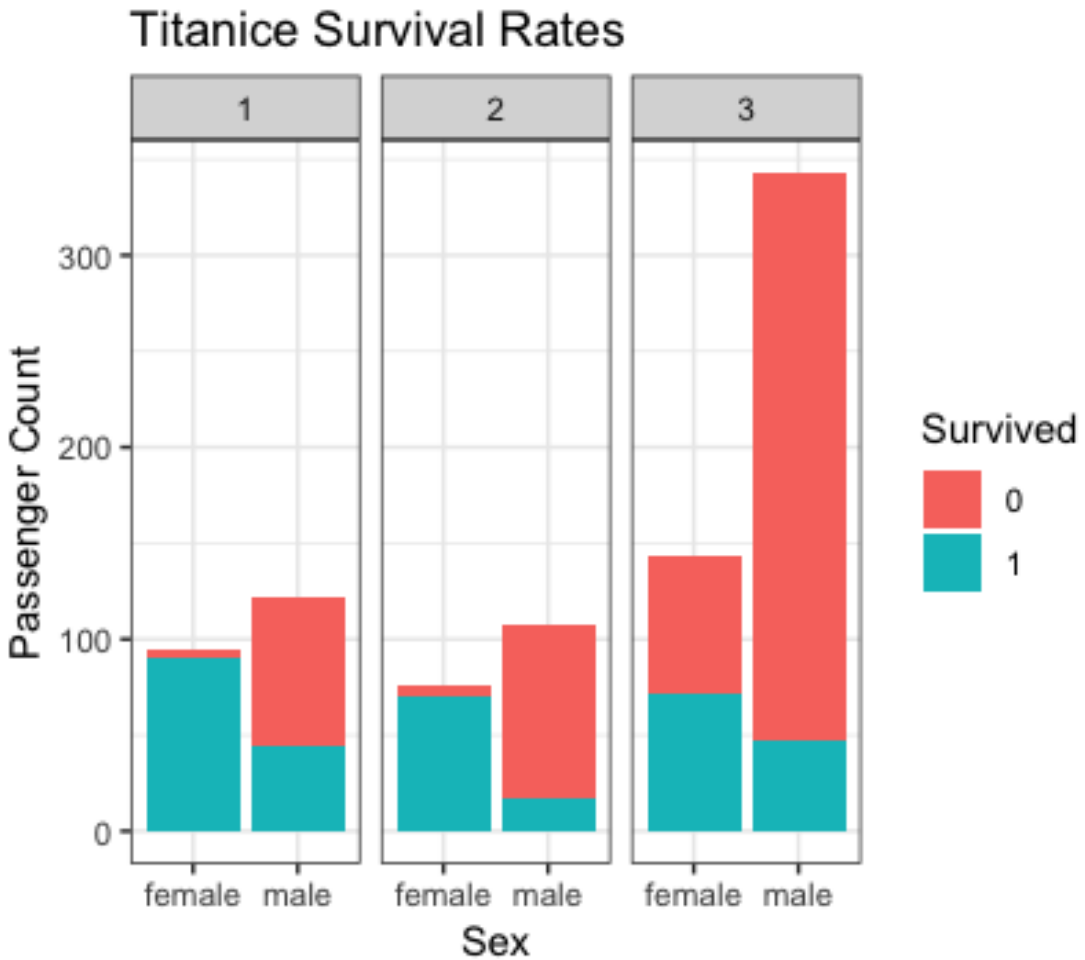
Titanic - Survival Rate By Class

```
1 ggplot(titanic, aes(x=Pclass, fill=Survived)) + geom_
    bar() + theme_bw() + labs(y= "Passenger Count",
    title = "Titanice Survival Rates")
```



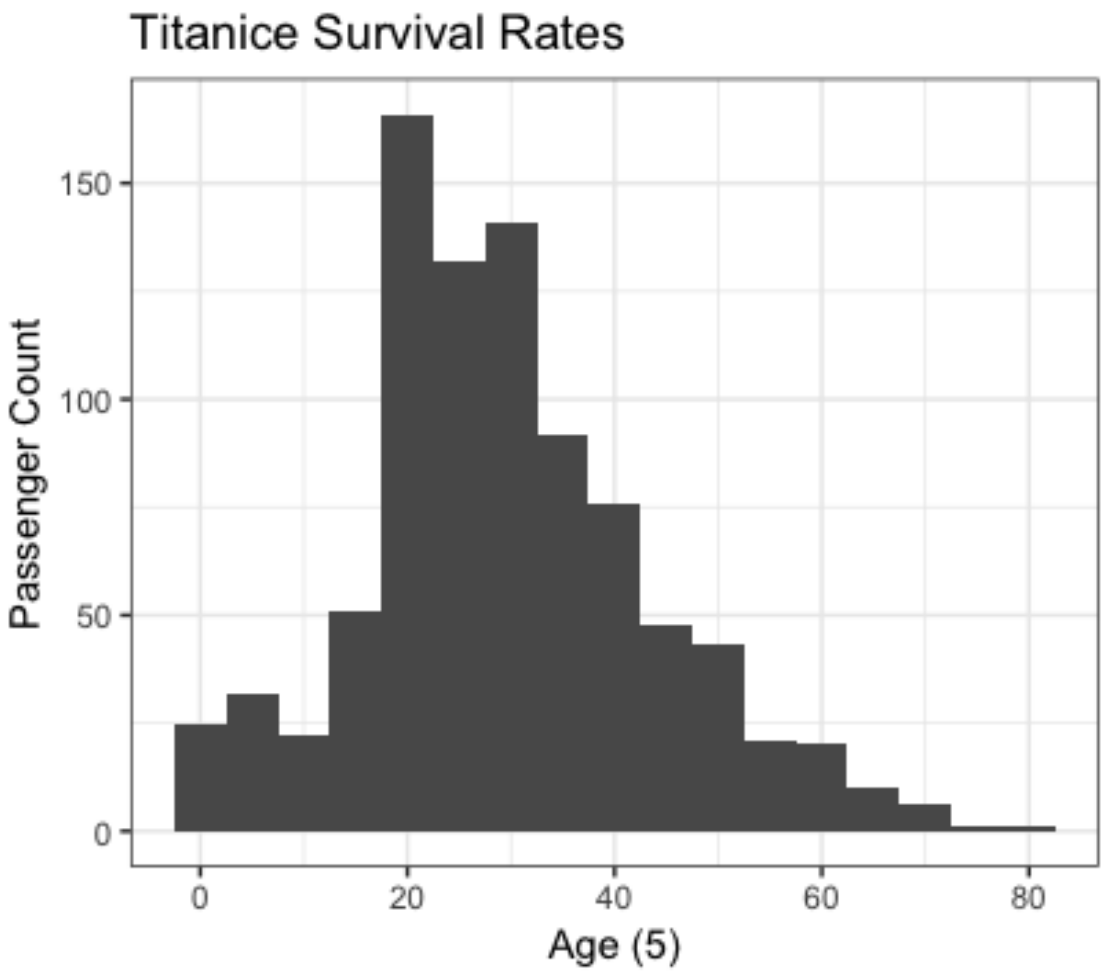
Titanic - Survival Rate By Gender AND Class (facet_wrap)

```
1 ggplot(titanic, aes(x=Sex, fill=Survived)) + geom_bar
    () + theme_bw() + facet_wrap(~ Pclass) + labs(y= "
    Passenger Count", title = "Titanice Survival Rates"
    )
```



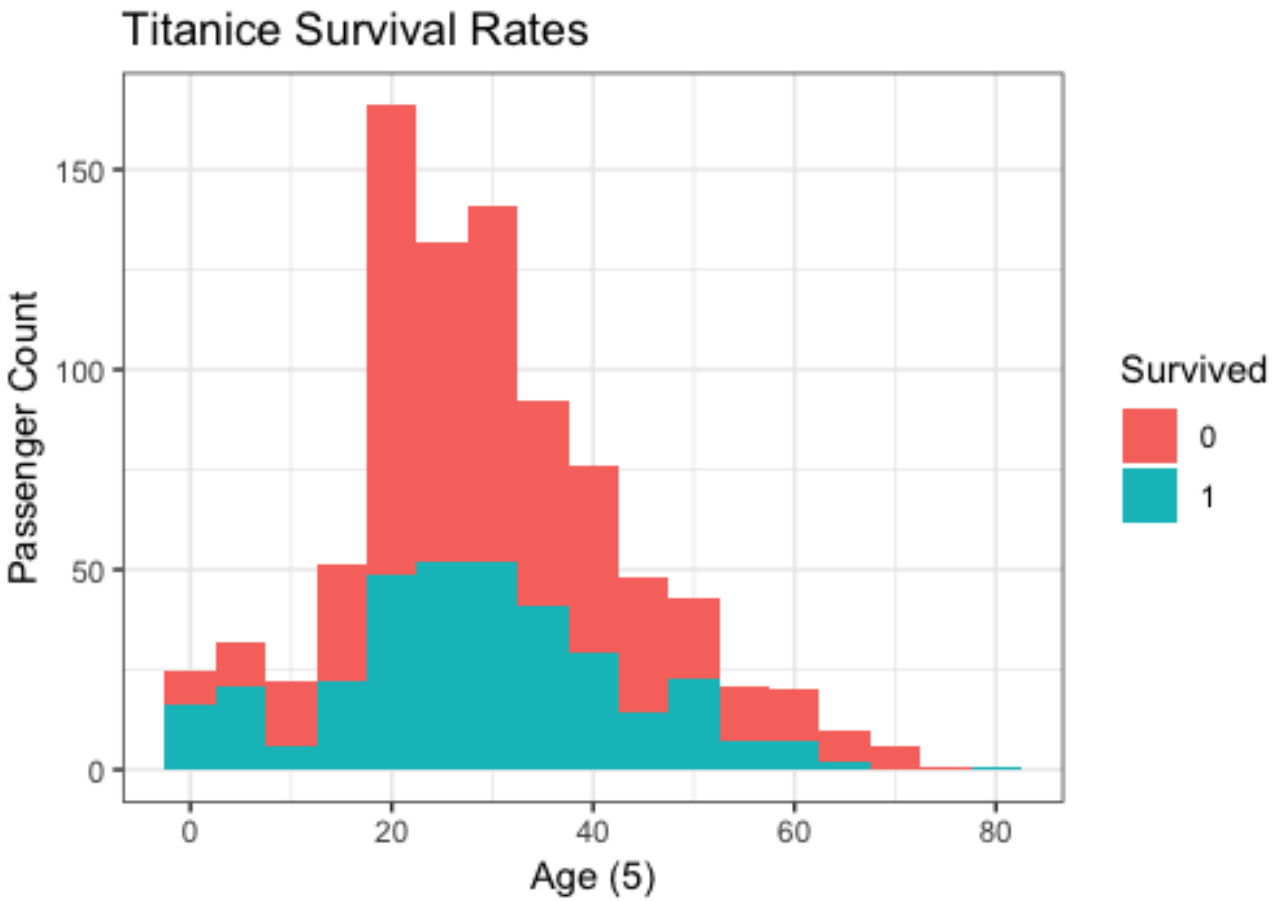
Titanic - Age Distribution

```
1 ggplot(titanic, aes(x=Age)) + geom_histogram(binwidth
    = 5) + theme_bw() + labs(y= "Passenger Count",x="
    Age (5)", title = "Titanice Survival Rates")
```



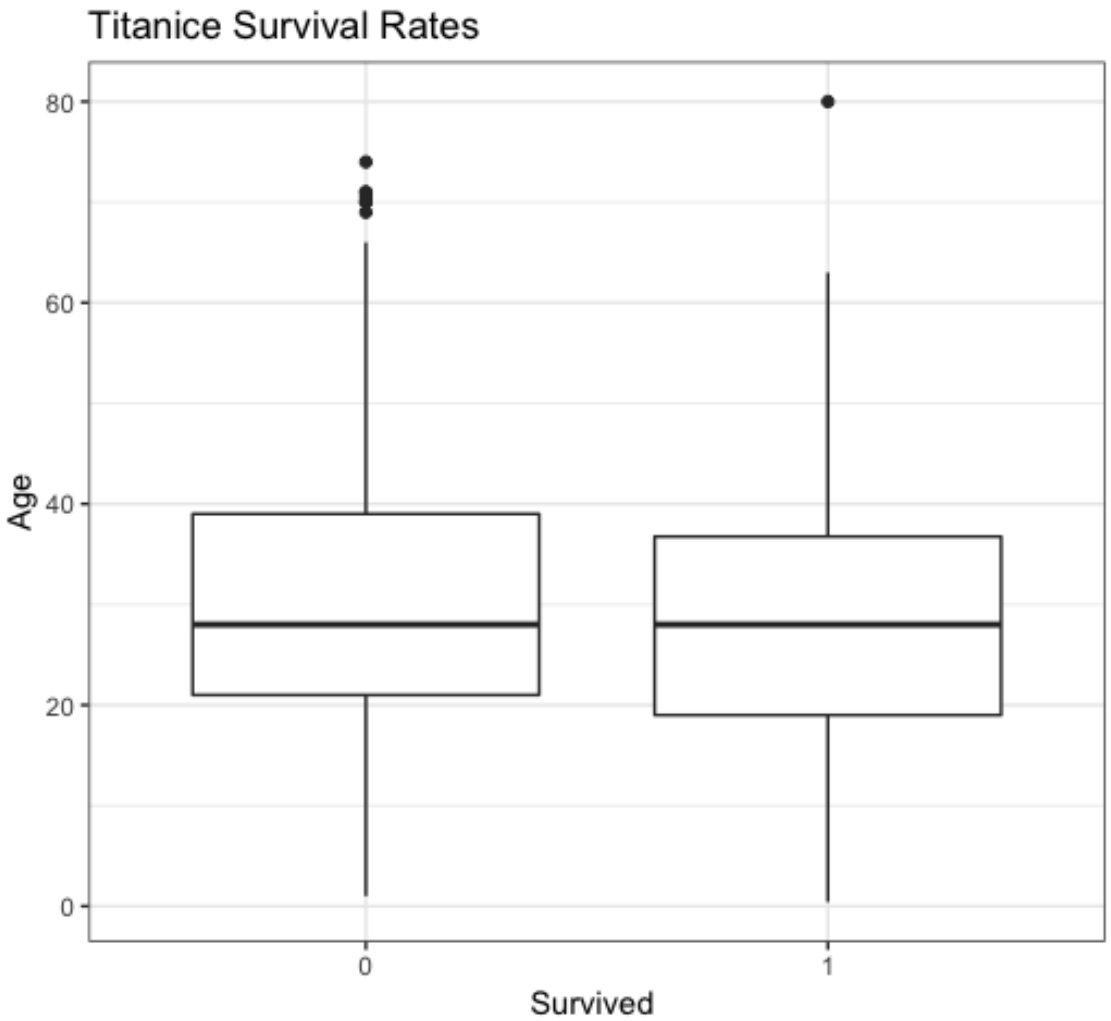
Titanic - Age Distribution (By Survival)

```
1 ggplot(titanic, aes(x=Age, fill=Survived)) + geom_
    histogram(binwidth = 5) + theme_bw() + labs(y= "
    Passenger Count",x="Age (5)", title = "Titanice
    Survival Rates")
```



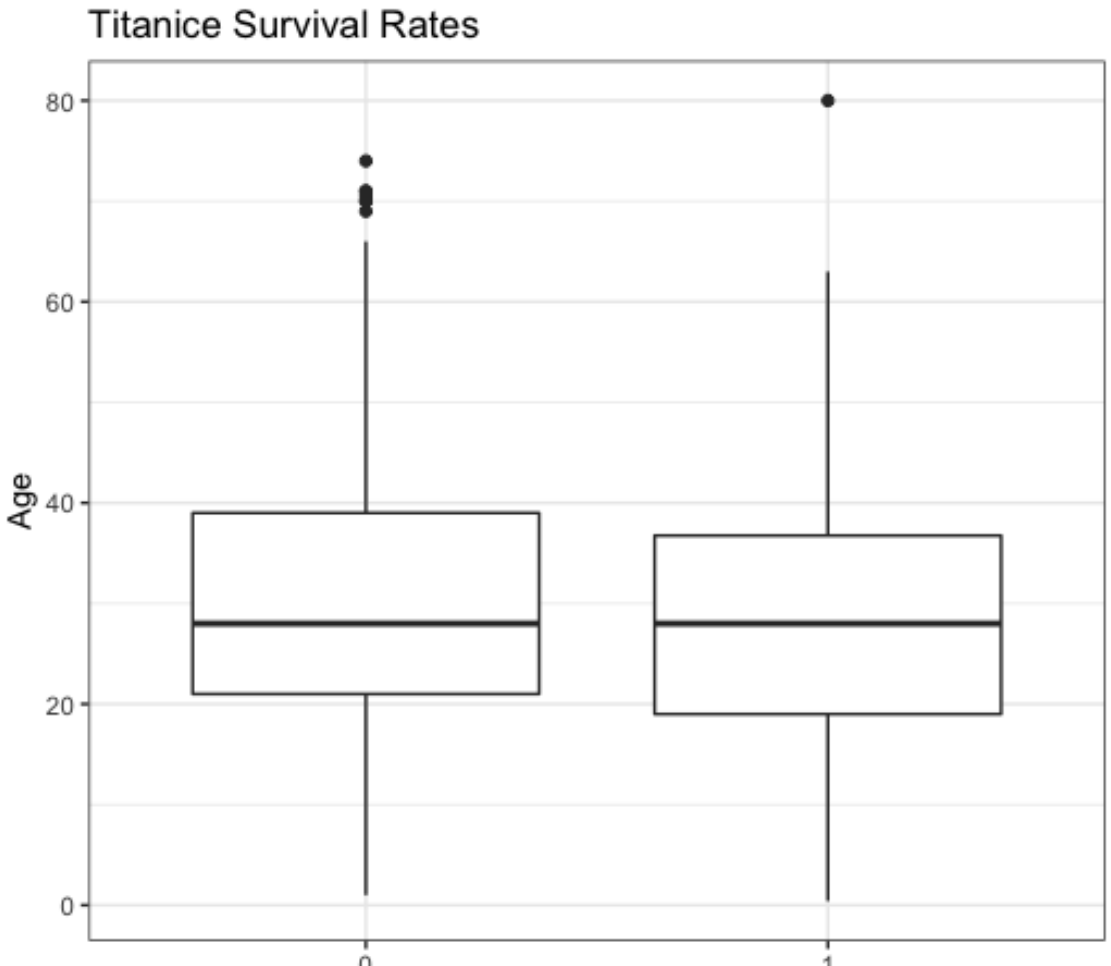
Titanic - Age Distribution Boxplot

```
1 ggplot(titanic, aes(x=Survived, y=Age)) + geom_boxplot
    () + theme_bw() + labs(y= "Age",x="Survived ",
    title = "Titanice Survival Rates")
```



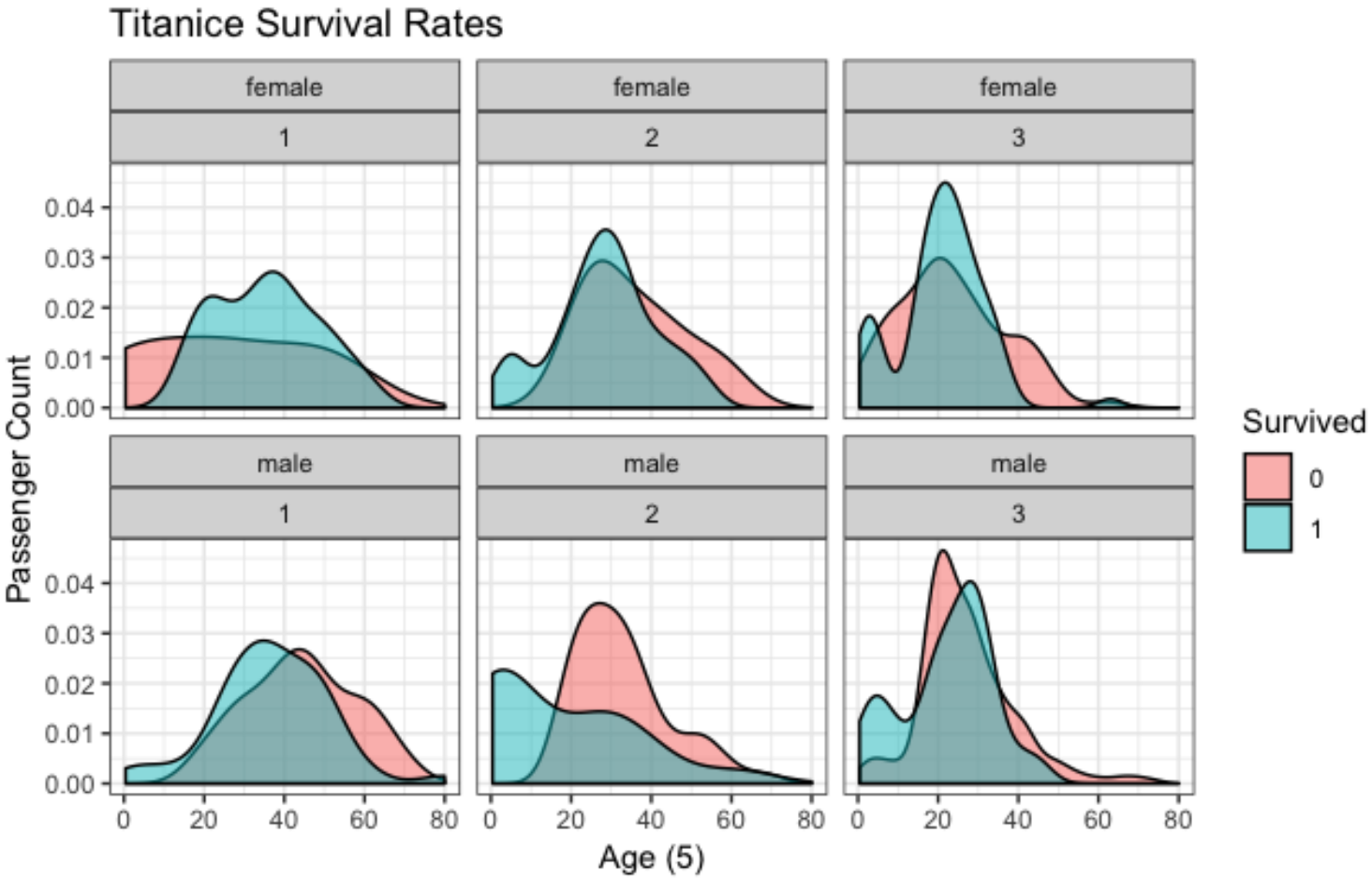
Titanic - Age/Gender/Class Density Curve

```
1 ggplot(titanic, aes(x=Age, fill=Survived)) + geom_
    density(alpha = 0.5) + theme_bw() + facet_wrap(Sex
    ~ Pclass) + labs(y= "Passenger Count",x="Age (5)",
    title = "Titanice Survival Rates")
```



Titanic - Age/Gender/Class Density Curve

```
1 ggplot(titanic, aes(x=Age, fill=Survived)) + geom_
    density(alpha = 0.5) + theme_bw() + facet_wrap(Sex
    ~ Pclass) + labs(y= "Passenger Count",x="Age (5)",
    title = "Titanice Survival Rates")
```



Titanic - Age/Gender/Class Density Histogram

```
1 ggplot(titanic, aes(x=Age, fill=Survived)) + geom_
    histogram(binwidth = 5) + theme_bw() + facet_wrap(
    Sex ~ Pclass) + labs(y= "Passenger Count",x="Age
    (5)", title = "Titanic Survival Rates")
```