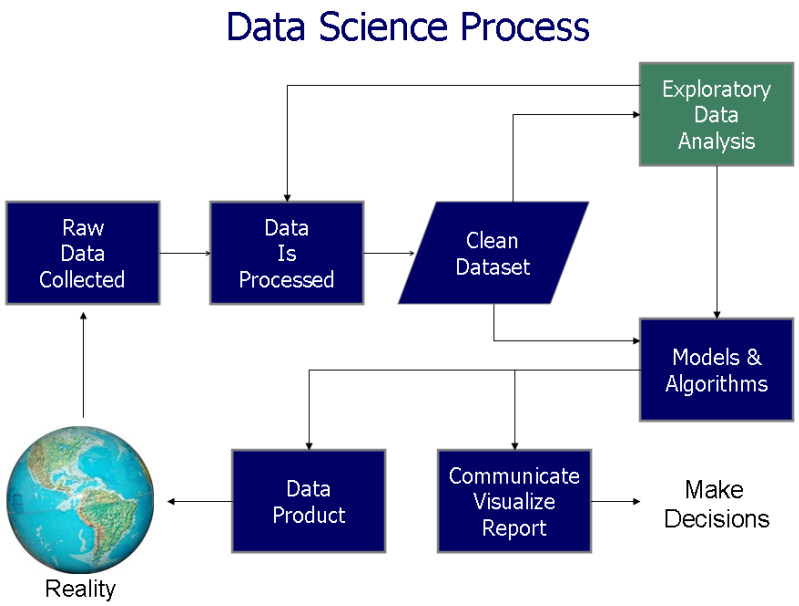
Notes 3 - Exploratory Data Analysis

Jillian Morrison

September 24, 2019



By Farcaster at English Wikipedia, CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=40129394

Exploratory Data Analysis

"Procedures for analyzing data, techniques for interpreting the results of such procedures, ways of planning the gathering of data to make its analysis easier, more precise or more accurate, and all the machinery and results of (mathematical) statistics which apply to analyzing data." ~Tukey (1961)

In other words, you want to perform initial investigations on data so as to:

- discover patterns
- to spot anomalies
- to test hypothesis
- to check assumptions with the help of summary statistics and graphical representations

By https://towardsdatascience.com/exploratory-data-analysis-8fc1cb20fd15

What will we learn to do?

Plots

- Histogram
- Bar Chart
- Pie Chart
- Box Plot
- Scatter Plot
- Correlation Plots

Cluster Analysis

- K-means Clustering
- Hierarchical Clustering

Main packages we will use

`{base}`

`{ggplot2}`

`{plotly}`

Plots

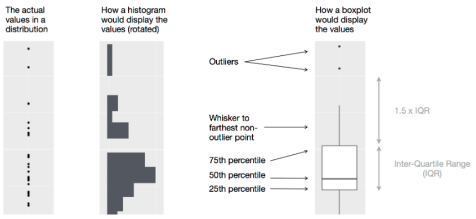
Here are other resources you can use besides texts listed in syllabus

http://www.sthda.com/english/wiki/be-awesome-in-ggplot2-a-practical-guide-to-be-highly-effective-r-software-and-data-vis

http://www.sthda.com/english/wiki/r-base-graphs

http://r-statistics.co/Top50-Ggplot2-Visualizations-MasterList-R-Code.html

Boxplots



Spread of distribution and symmetric vs skewed

- the box - from the 25th percentile to the 75th percentile (or IQR)
- line in middle of the box - median (50th percentile) These

Outliers

- points that fall more than 1.5 times the IQR
- line (or whisker) that extends from the box shows the farthest non-outlier point

Box plot using{base} i.e. boxplot()

using ToothGrowth dataset in {datasets}. This library should be loaded automatically when R starts.

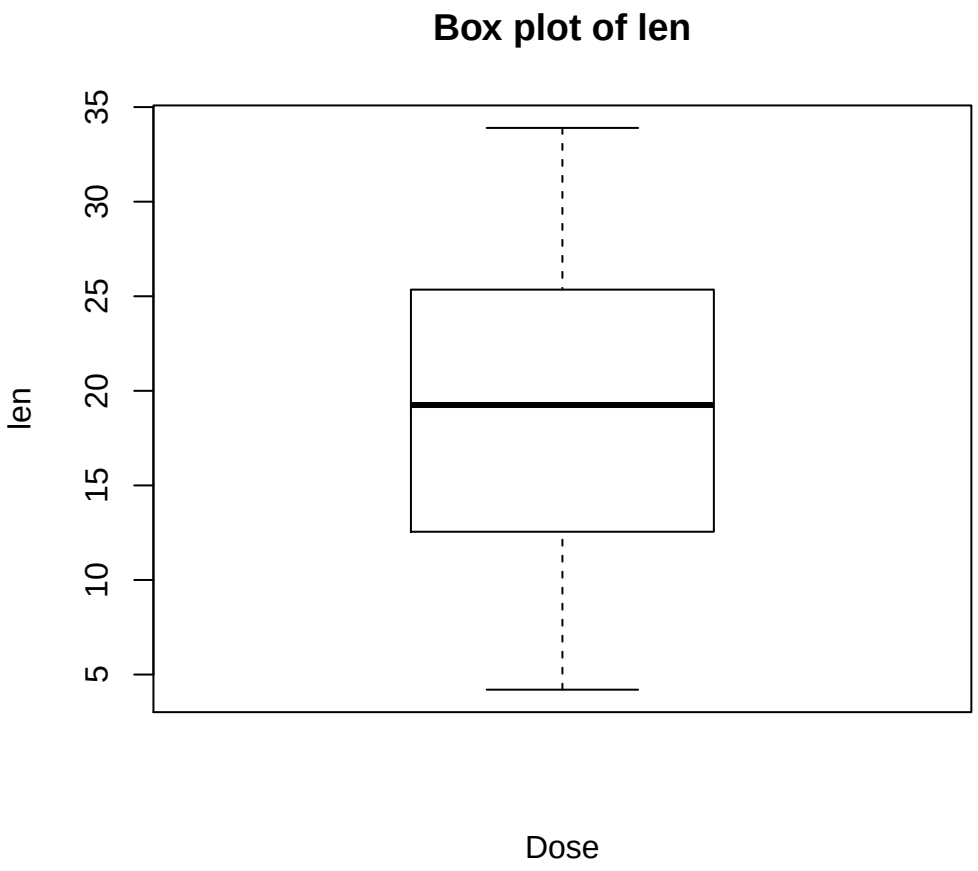
```
> library(dplyr)
> Tooth=ToothGrowth  #Saving dataset to Tooth
> Tooth2=Tooth%>%mutate(dose=factor(dose)) # changes dose to FACTOR (was numeric)
```

```
> #To see what is inside the dataset
> head(Tooth, n=3)
   len supp dose
1  4.2   VC  0.5
2 11.5   VC  0.5
3  7.3   VC  0.5
> #We see first 6 rows of dataset
```

```
> Tooth%>%group_by(supp)%>%slice(1)
# A tibble: 2 x 3
# Groups:   supp [2]
    len supp   dose
  <dbl> <fct> <dbl>
1  15.2 OJ      0.5
2   4.2 VC      0.5
> #We see 2 types of supplements, OJ and VC
```

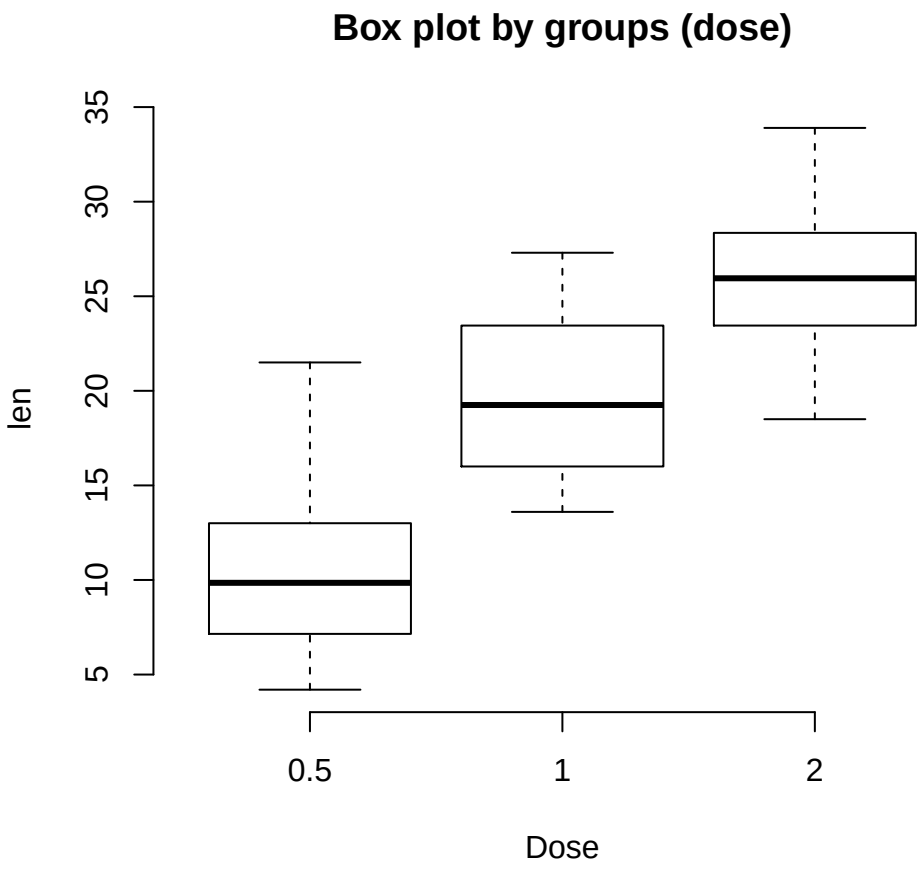
Boxplot of one Variable

```
> boxplot(Tooth2$len, main="Box plot of len",  xlab="Dose", ylab="len")
```



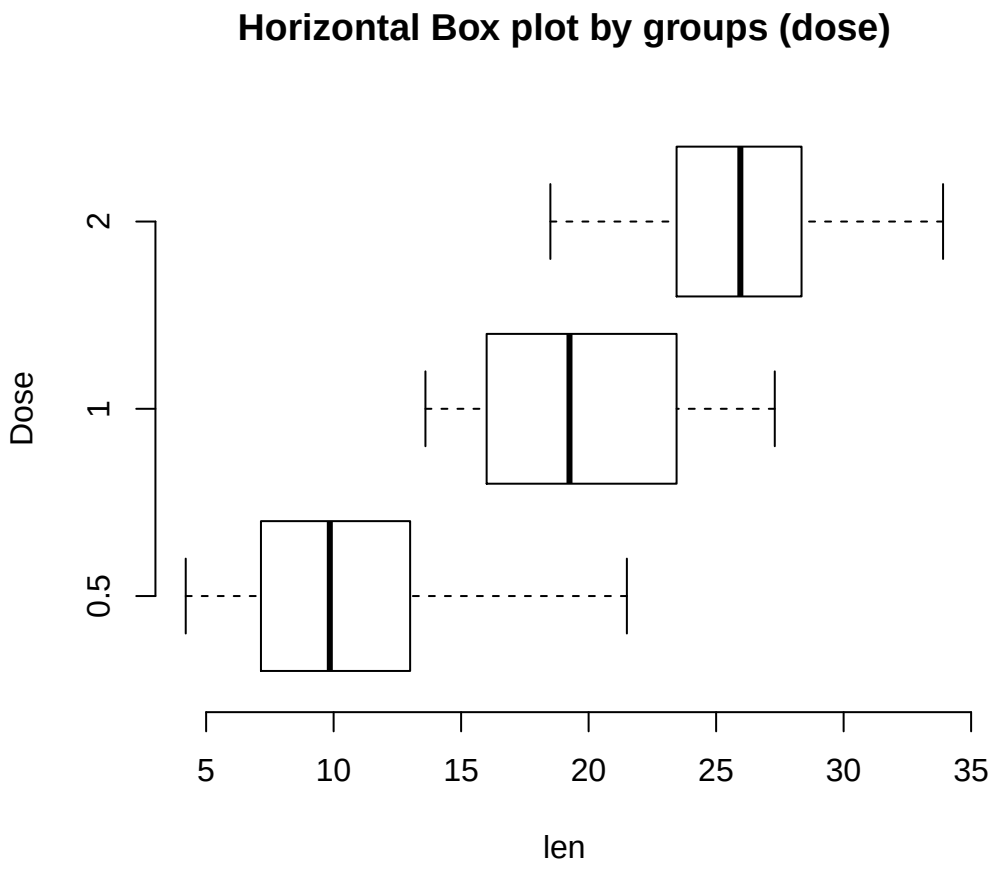
Box plots by groups (dose) with frame removed

```
> boxplot(len ~ dose, data = Tooth2, frame = FALSE, main="Box plot by groups (dose)",
+         xlab="Dose", ylab="len")
```



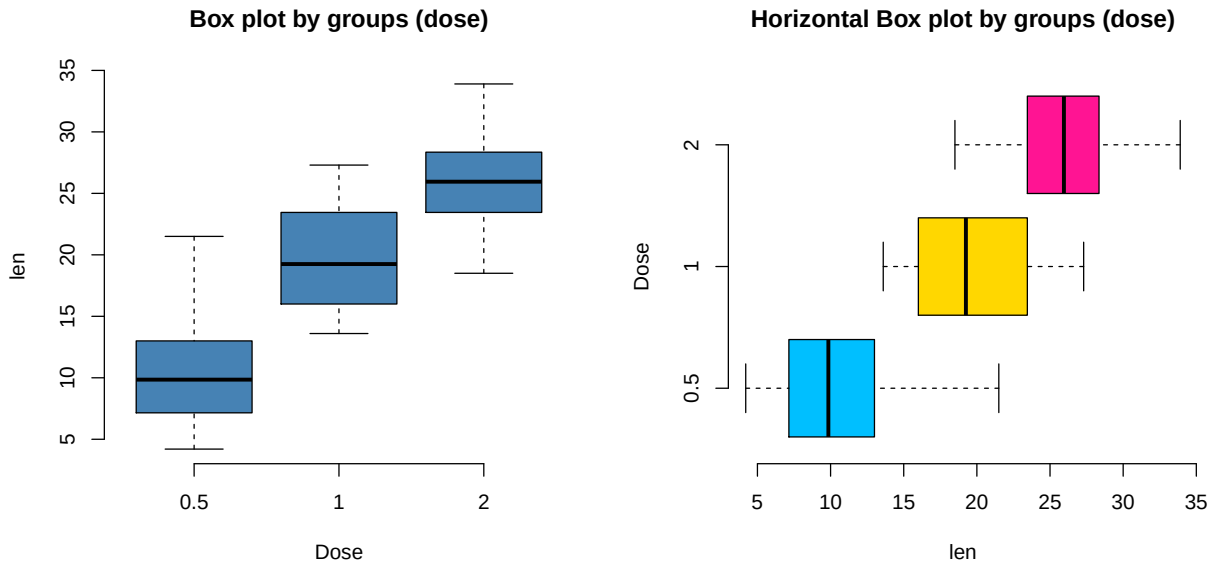
Horizontal box plots

```
> boxplot(len ~ dose, data = Tooth2, frame = FALSE,
+         horizontal = TRUE, main="Horizontal Box plot by groups (dose)", ylab="Dose",
+         xlab="len")
```



Adding color and multiple plots on page

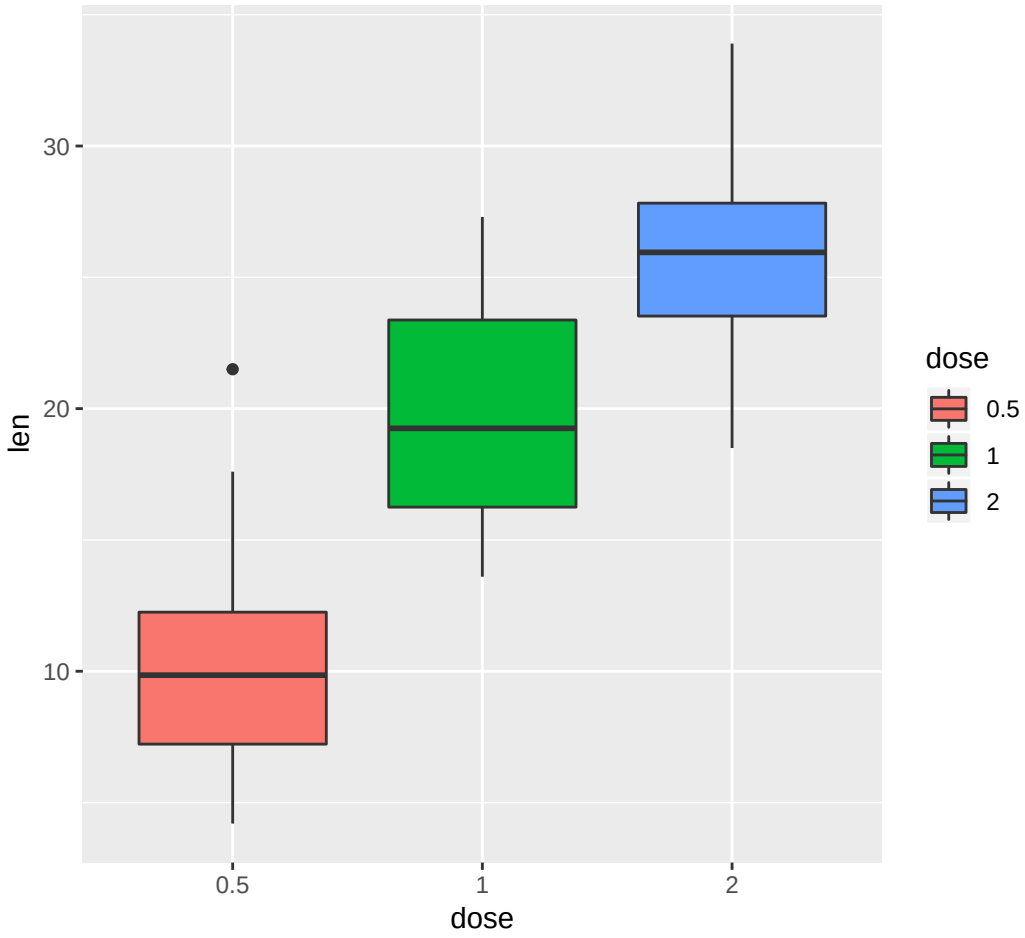
```
> par(mfrow=c(1,2))  #Puts number of plots by c(row, column)
>
> # All plots the same color
> boxplot(len ~ dose, data = Tooth2, frame = FALSE, main="Box plot by groups (dose)",
+         xlab="Dose", ylab="len", col = "steelblue")
>
> # Specific color per plot
> boxplot(len ~ dose, data = Tooth2, frame = FALSE,
+         horizontal = TRUE, main="Horizontal Box plot by groups (dose)", ylab="Dose",
+         xlab="len",col = c("deepskyblue", "gold", "deeppink"))
```



For more colors see: http://www.stat.columbia.edu/~tzheng/files/Rcolor.pdf

Using `ggplot()` in {ggplot2}

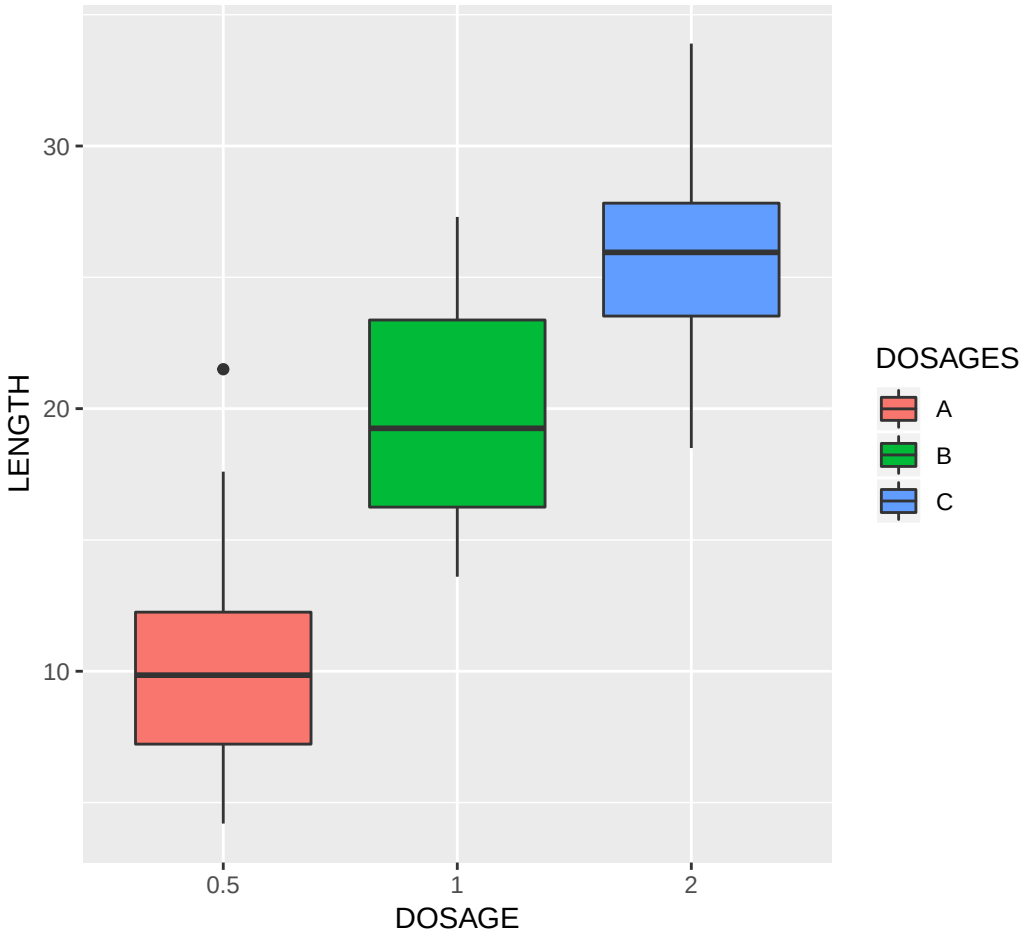
```
> library(ggplot2)
> ggplot(data = Tooth2, mapping = aes(x = dose, y = len, fill=dose)) + geom_boxplot()
```



- mapping (i.e. aes()): variables that map to the aestethics of the plot
  - x,y: coresponding x and y variables
  - fill: grouping variable - fills with color
- `geom_boxplot()`: specifies type of plot - boxplot in this case
- +: allows you to layer plot with more options

ADDING axis labels

```
> ggplot(data = Tooth2, mapping = aes(x = dose, y = len, fill=dose)) + geom_boxplot()+
+   xlab("DOSAGE")+ylab("LENGTH")+scale_fill_discrete(name="DOSAGES", labels=c("A","B","C"))
```



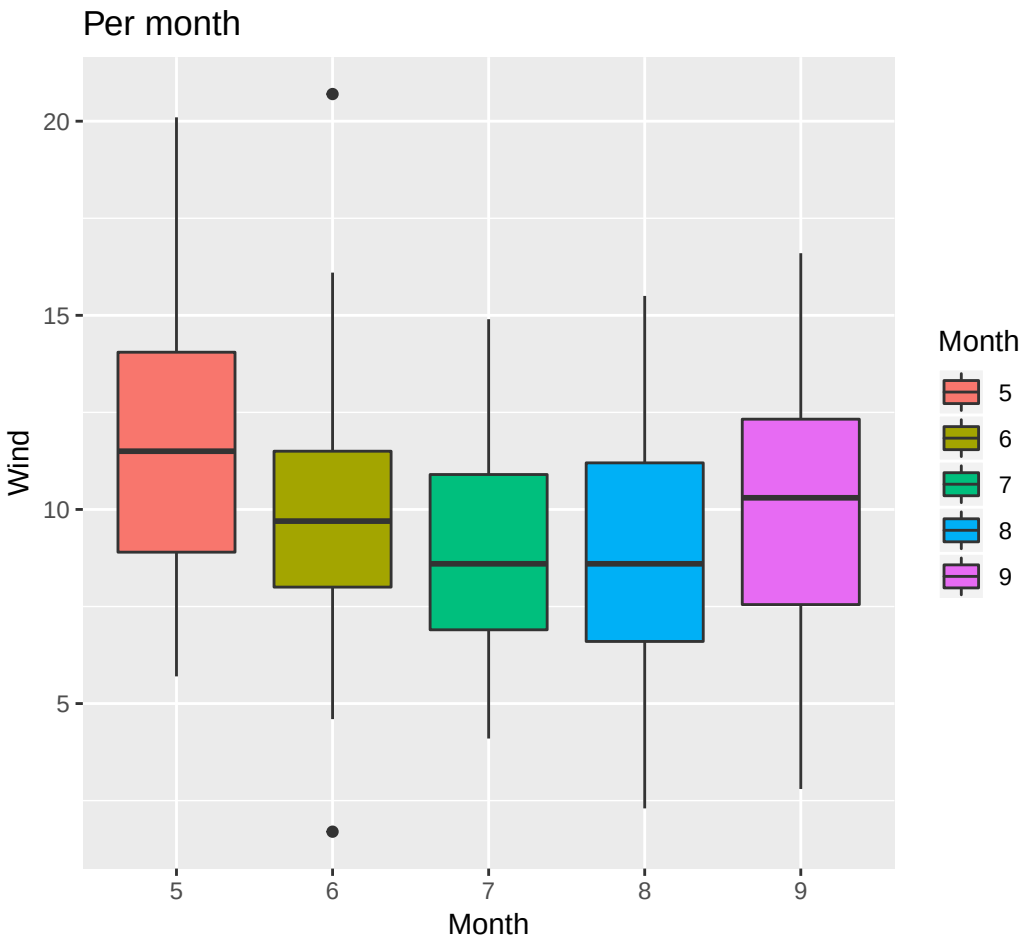
Exercise

Use the `airquality` dataset in `{datasets}` to draw boxplots of wind by month. Solutions will be provided with `ggplot()`

```
> AIR=airquality
> head(AIR)
  Ozone Solar.R Wind Temp Month Day
1    41     190  7.4   67     5   1
2    36     118  8.0   72     5   2
3    12     149 12.6   74     5   3
4    18     313 11.5   62     5   4
5    NA      NA 14.3   56     5   5
6    28      NA 14.9   66     5   6
```

Solution

```
> AIR2=AIR%>%mutate(Month=factor(Month)) #Month needs to be a factor
> ggplot(data=AIR2,
+        mapping=aes(x=Month, y=Wind,fill=Month))+geom_boxplot()+ggtitle("Per month")
```



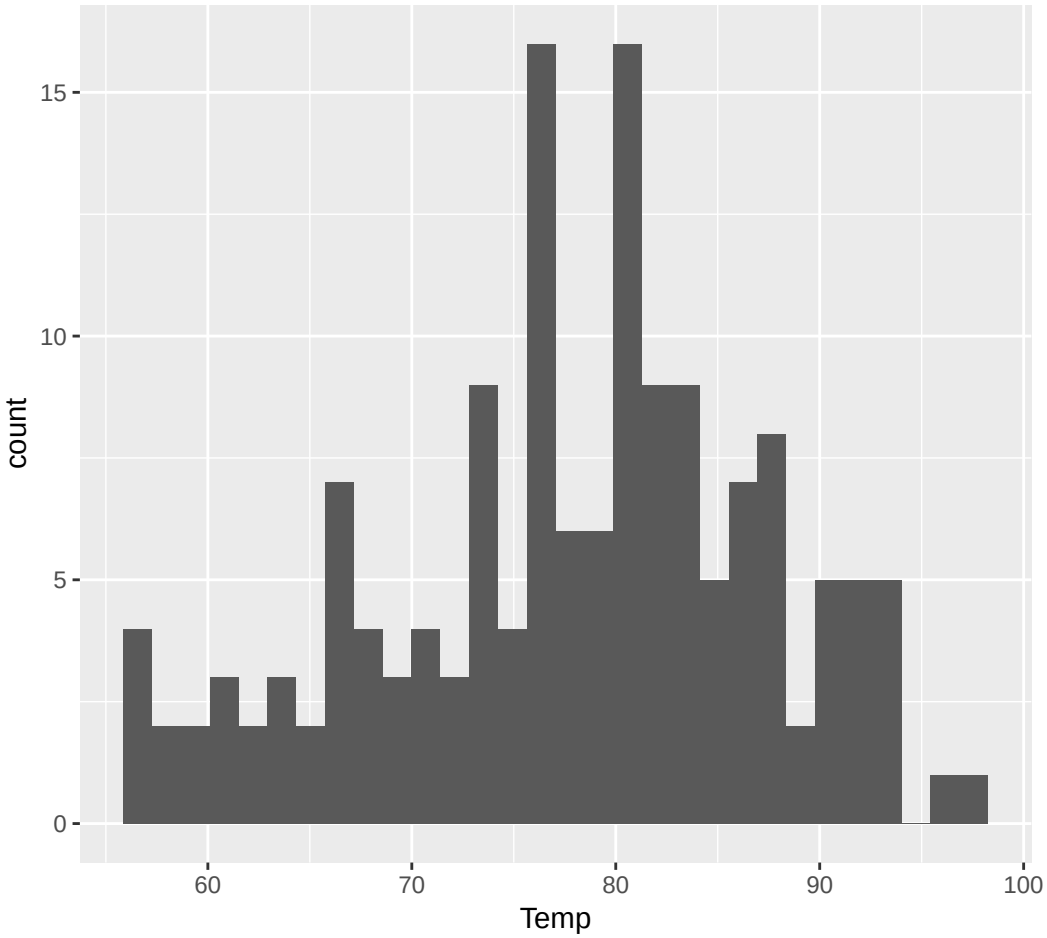
Notice I layered with `+ ggtitle("TITLE of plot")` where I added a title to the plot. TRY IT OUT!

Histogram

Histogram is usually used to see the distribution of the data. Data must be quantative (in R, numerical)

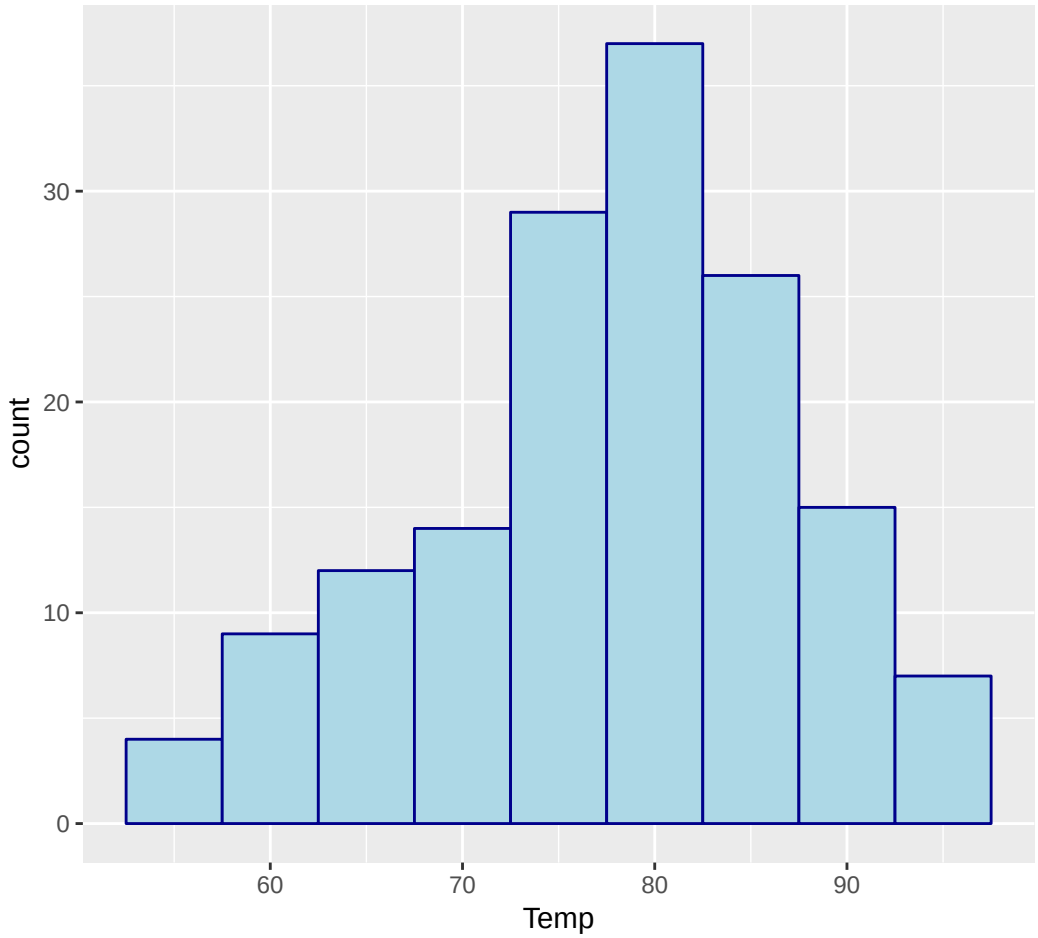
For Example: plotting histogram of `Temp` in `airquality` dataset

```
> ggplot(data=AIR2, aes(x=Temp))+geom_histogram()
```



Histogram- changing the bin width and adding color

```
> ggplot(data=AIR2, aes(x=Temp))+geom_histogram(binwidth = 5,
+                                               color="darkblue", fill="lightblue")
```

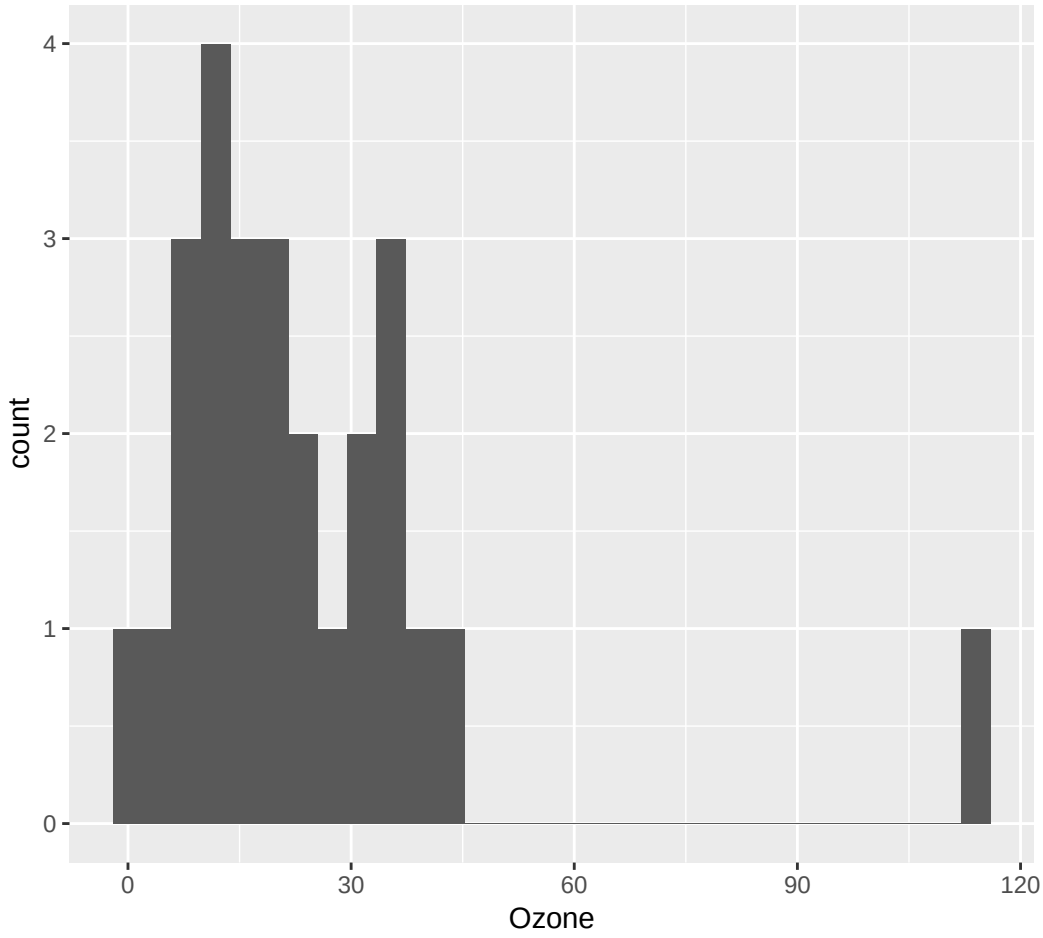


Exercise

Filter the `airquality` dataset for Month 5 (Hint: Save this as a new object). Draw a histogram of `Ozone`

Solution

```
> AIR3=AIR2%>%filter(Month =="5")
> ggplot(AIR3, aes(x=Ozone))+geom_histogram()
```

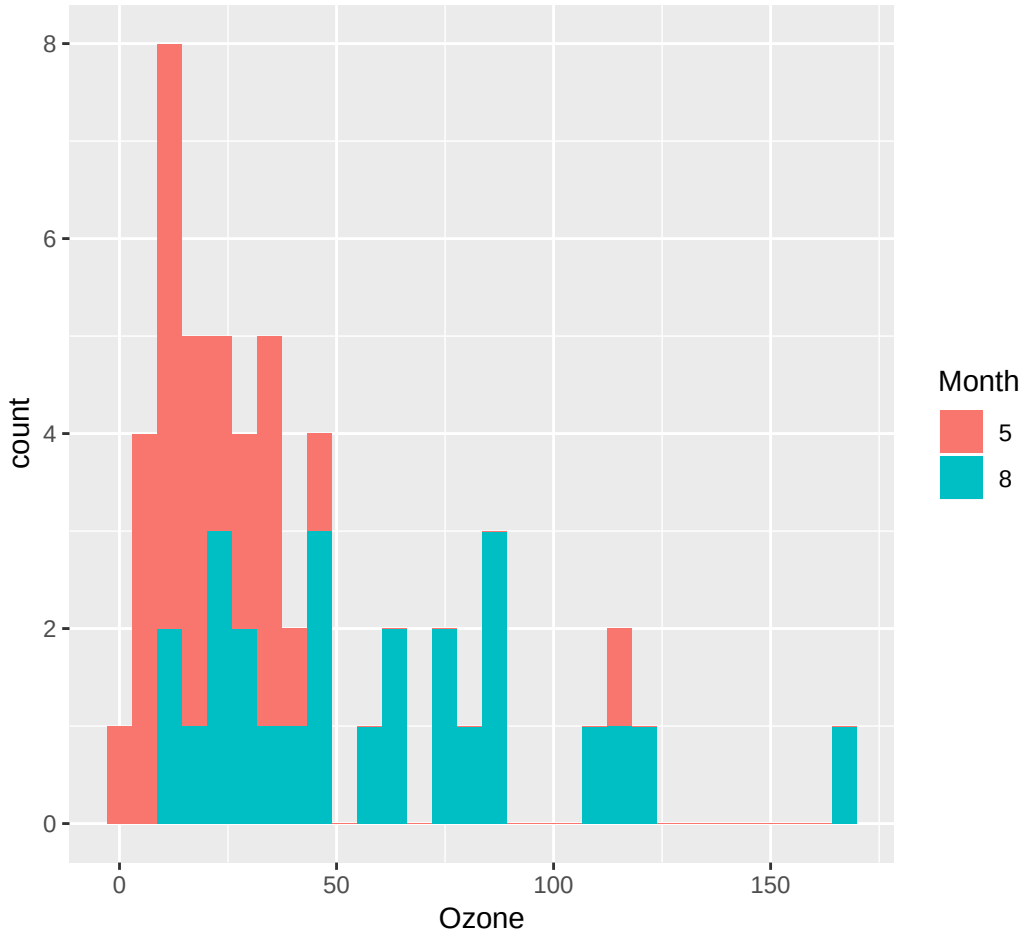


Exercise

Filter for Months 5 and 8 Draw histograms (on the same plot) of Ozone and group by Months 5 and 8

Solution

```
> AIR4=AIR2%>%filter(Month=="5" | Month =="8")
> ggplot(AIR4, aes(x=Ozone, fill=Month))+geom_histogram()
```



Barplots

Barplots are usually used to visualize categorical data. In R, these are usually factors or characters.

Using the `titanic_train` dataset in `{titanic}`

```
> #install.packages("titanic")
> library(titanic)
> Tit=titanic_train%>%mutate(Survived=factor(Survived), Sex=factor(Sex), Pclass=factor(Pclass))  #Mutat
>
> glimpse(Tit)
Observations: 891
Variables: 12
$ PassengerId <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12,...
$ Survived    <fct> 0, 1, 1, 1, 0, 0, 0, 0, 1, 1, 1, 1, 0,...
$ Pclass      <fct> 3, 1, 3, 1, 3, 3, 1, 3, 3, 2, 3, 1, 3,...
$ Name        <chr> "Braund, Mr. Owen Harris", "Cumings, M...
$ Sex         <fct> male, female, female, female, male, ma...
$ Age         <dbl> 22, 38, 26, 35, 35, NA, 54, 2, 27, 14,...
$ SibSp       <int> 1, 1, 0, 1, 0, 0, 0, 3, 0, 1, 1, 0, 0,...
$ Parch       <int> 0, 0, 0, 0, 0, 0, 0, 1, 2, 0, 1, 0, 0,...
$ Ticket      <chr> "A/5 21171", "PC 17599", "STON/O2. 310...
```

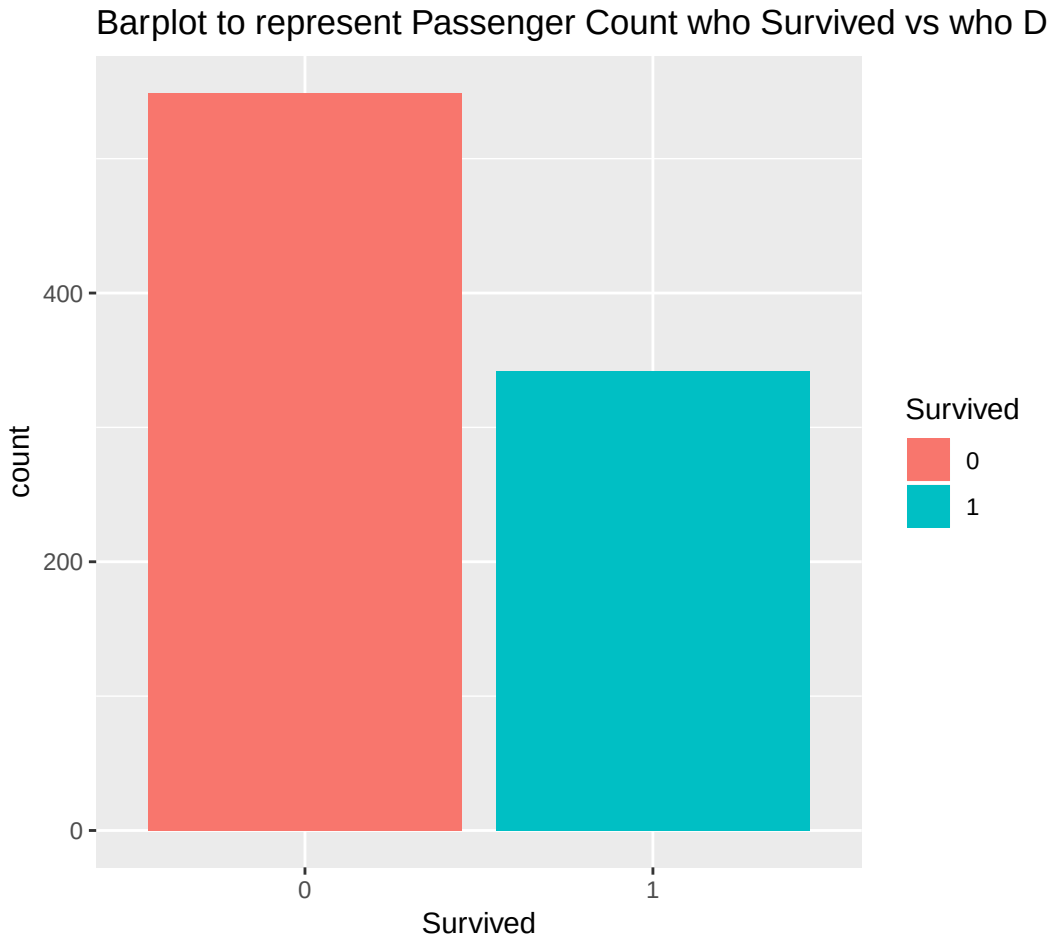
```
$ Fare        <dbl> 7.2500, 71.2833, 7.9250, 53.1000, 8.05...
$ Cabin       <chr> "", "C85", "", "C123", "", "", "E46", ...
$ Embarked    <chr> "S", "C", "S", "S", "S", "Q", "S", "S"...
```

provides information on the fate of passengers on the fatal maiden voyage of the ocean liner "Titanic", summarized according to economic status (class), sex, age and survival.

Titanic barplot - Using counts

Want to plot to see how many people survived or died.

```
> ggplot(Tit, aes(x = Survived, fill = Survived))+
+   geom_bar()+
+   ggtitle("Barplot to represent Passenger Count who Survived vs who Died")
```



Note: This just counts number of the levels of the Survived variable (i.e. 0 if Survived, 1 if Died)

Titanic barplot - Using variable from dataset

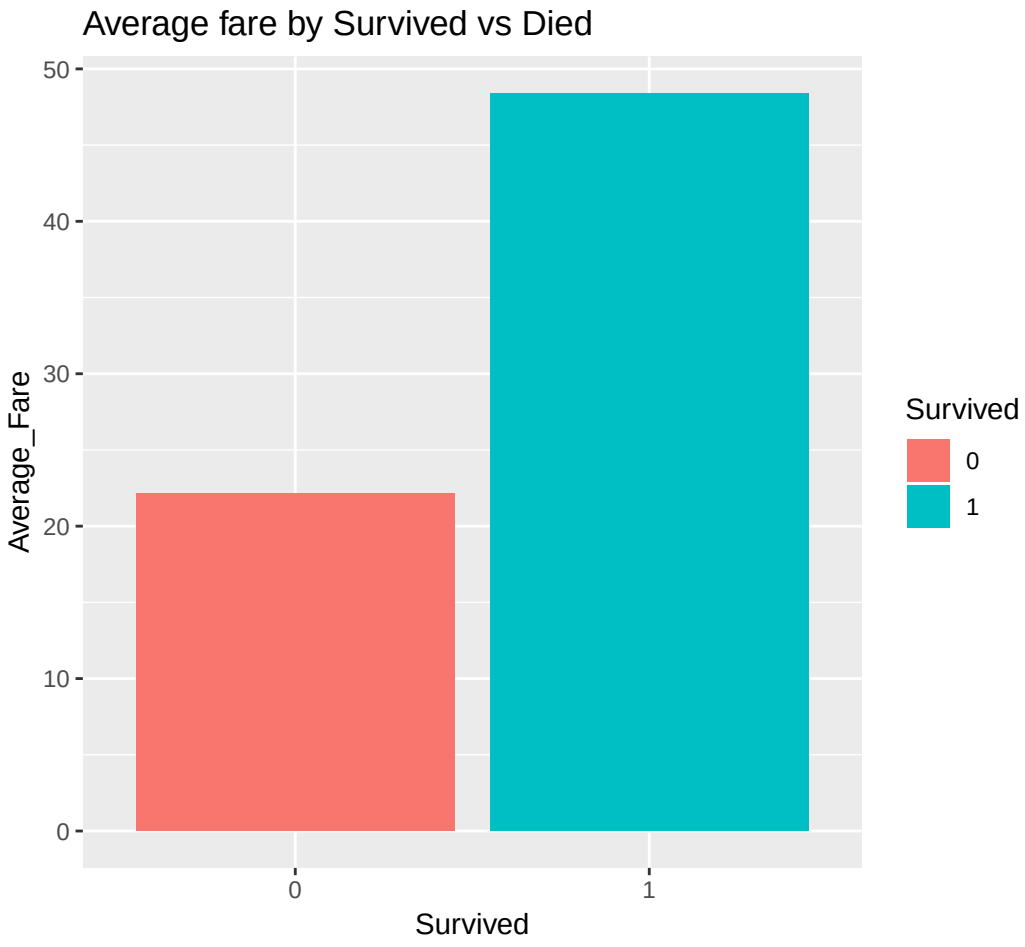
Want a barplot of average Fare by Survival status

- Must summarize data to have ONLY Average fare and survival status (i.e. use `summarize()`)
- Plot barplot, however, stat = "identity" MUST be added when you want to use an existing value from the dataset instead of count.

```
> Tit3=Tit%>%group_by(Survived)%>%summarize(Average_Fare=mean(Fare))
> Tit3
# A tibble: 2 x 2
  Survived Average_Fare
  <fct>           <dbl>
1 0                22.1
2 1                48.4
```

Titanic barplot - Using variable from dataset (ctd)

```
> ggplot(Tit3, aes(x=Survived,y=Average_Fare, fill = Survived))+
+   geom_bar(stat = "identity")+
+   ggtitle("Average fare by Survived vs Died")
```



Titanic Plot - Using variable in dataset and grouping

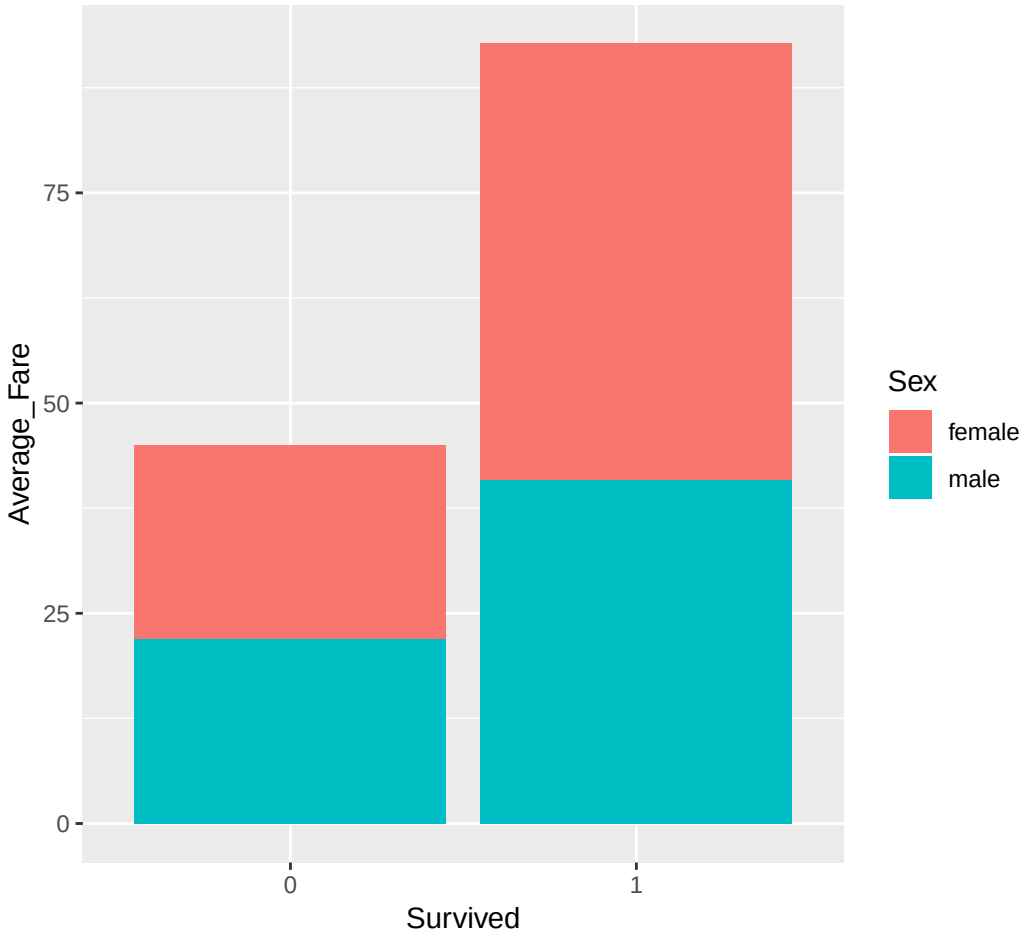
Want to Plot Average fare by Survival status, but group by Sex...

AGAIN: Need to Summarize to have table with ONLY the NEEDED variables

```
> Tit4=Tit%>%group_by(Survived, Sex)%>%summarize(Average_Fare=mean(Fare))
> Tit4
# A tibble: 4 x 3
# Groups:   Survived [?]
  Survived Sex    Average_Fare
  <fct>    <fct>         <dbl>
1 0        female         23.0
2 0        male           22.0
3 1        female         51.9
4 1        male           40.8
```

Titanic Plot - Using variable in dataset and grouping (ctd.)

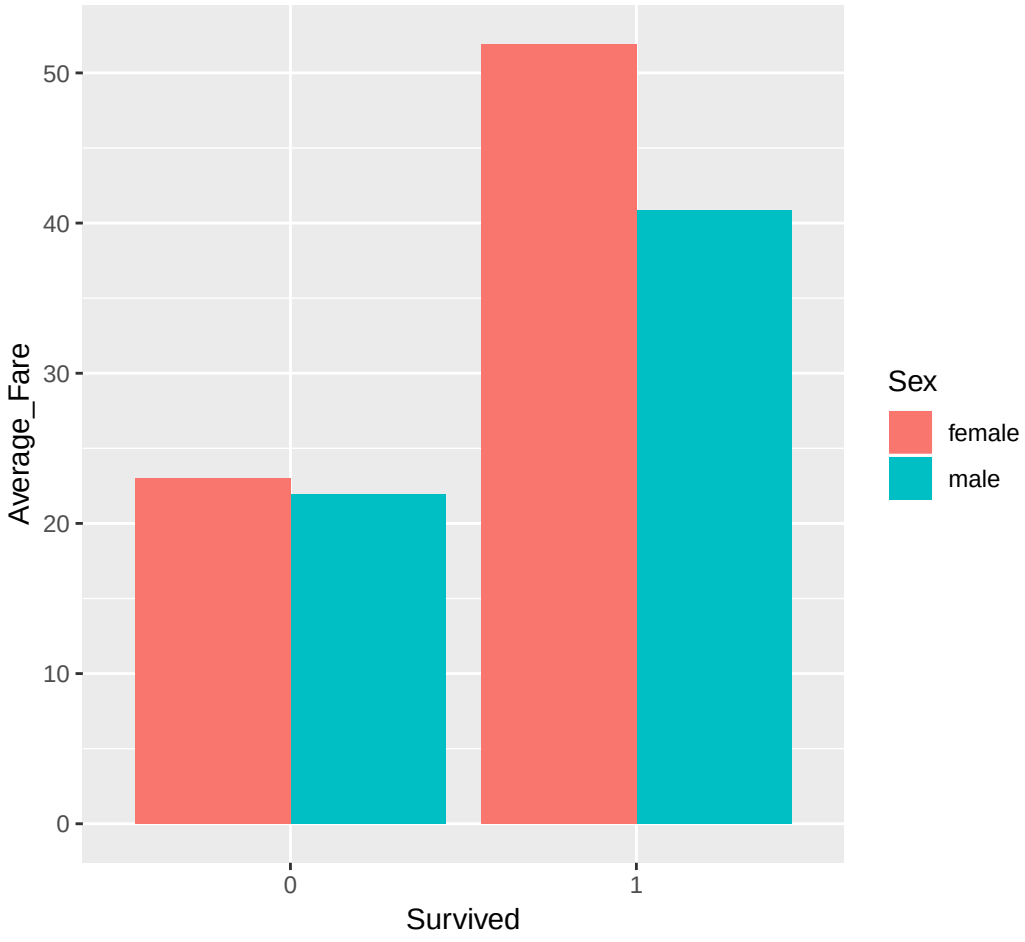
```
> ggplot(Tit4, aes(x=Survived, y=Average_Fare, fill=Sex))+geom_bar(stat = "identity")
```



Titanic Plot - Using variable in dataset and grouping (ctd.)

Changing the Style of the barplot we just created

```
> ggplot(Tit4, aes(x=Survived, y=Average_Fare, fill=Sex))+geom_bar(stat = "identity", position='dodge')
```



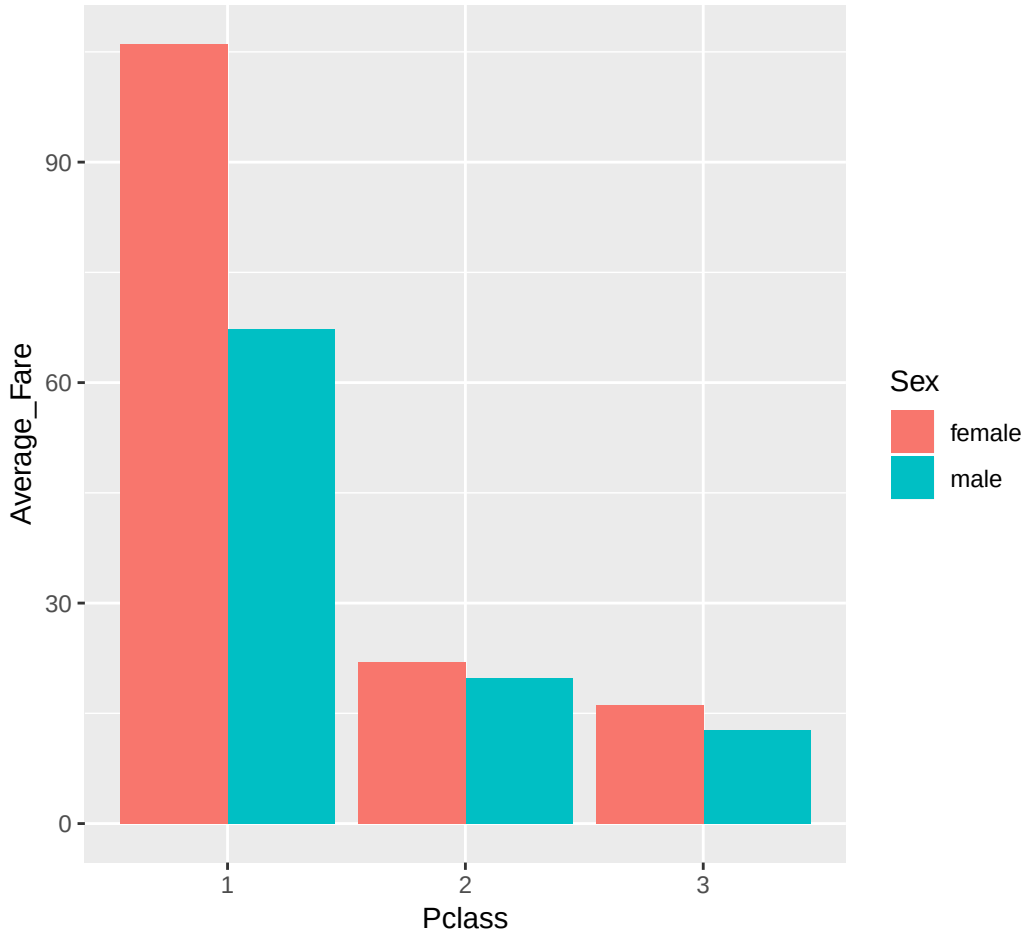
Exercise

Plot Average fare by Pclass (class of ticket), but group by Sex...

REMEMBER: You have to summarize data by Average fare, Pclass and Sex before plotting

Solution

```
> Tit5=Tit%>%group_by(Pclass, Sex)%>%summarize(Average_Fare=mean(Fare))
> ggplot(Tit5, aes(x=Pclass, y=Average_Fare, fill=Sex))+geom_bar(stat = "identity", position='dodge')
```



Want to change color palette?

USE

`+ scale_color_brewer(palette="Dark2")` for example.

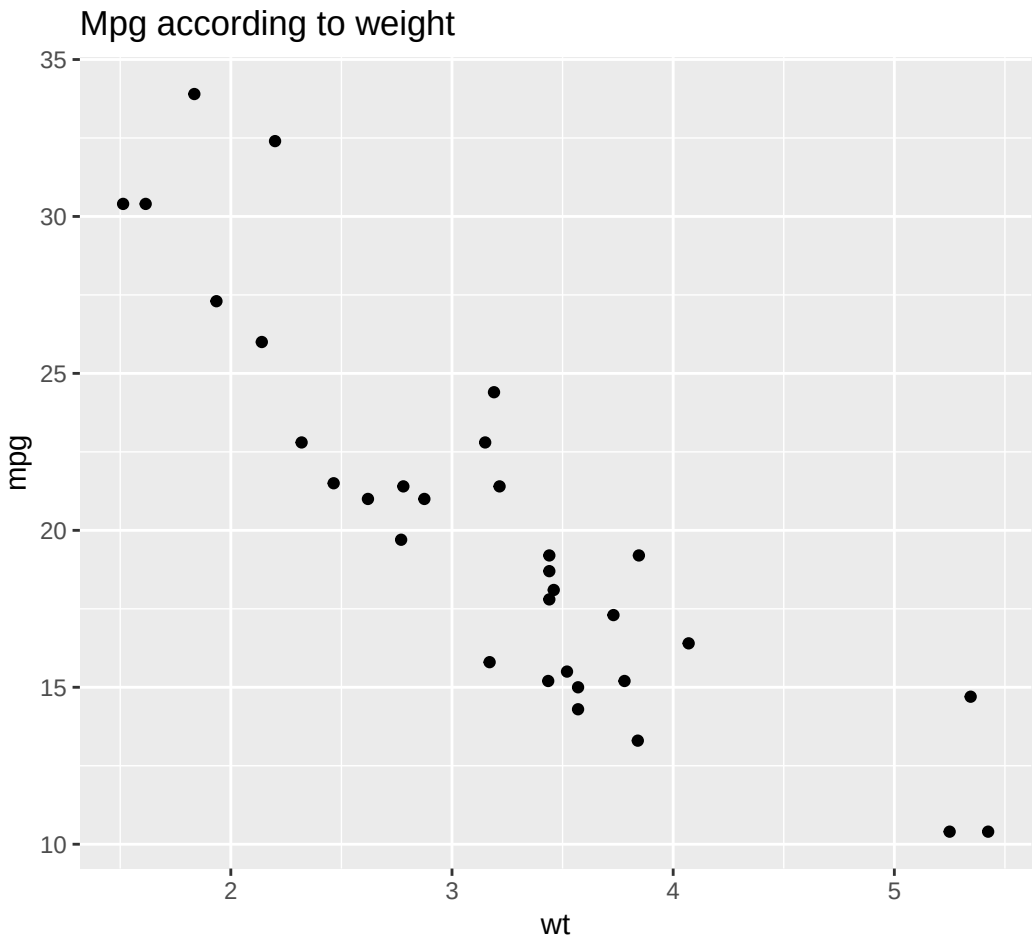
Scatter Plot

Using `mtcars` dataset from `{datasets}`

```
> Car=mtcars%>%mutate(cyl=factor(cyl))
> glimpse(Car)
Observations: 32
Variables: 11
$ mpg  <dbl> 21.0, 21.0, 22.8, 21.4, 18.7, 18.1, 14.3, 24....
$ cyl  <fct> 6, 6, 4, 6, 8, 6, 8, 4, 4, 6, 6, 8, 8, 8, 8, ...
$ disp <dbl> 160.0, 160.0, 108.0, 258.0, 360.0, 225.0, 360...
$ hp   <dbl> 110, 110, 93, 110, 175, 105, 245, 62, 95, 123...
$ drat <dbl> 3.90, 3.90, 3.85, 3.08, 3.15, 2.76, 3.21, 3.6...
$ wt   <dbl> 2.620, 2.875, 2.320, 3.215, 3.440, 3.460, 3.5...
$ qsec <dbl> 16.46, 17.02, 18.61, 19.44, 17.02, 20.22, 15....
$ vs   <dbl> 0, 0, 1, 1, 0, 1, 0, 1, 1, 1, 1, 0, 0, 0, 0, ...
$ am   <dbl> 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...
$ gear <dbl> 4, 4, 4, 3, 3, 3, 3, 4, 4, 4, 4, 3, 3, 3, 3, ...
$ carb <dbl> 4, 4, 1, 1, 2, 1, 4, 2, 2, 4, 4, 3, 3, 3, 4, ...
```

Scatterplot of mpg versus wt

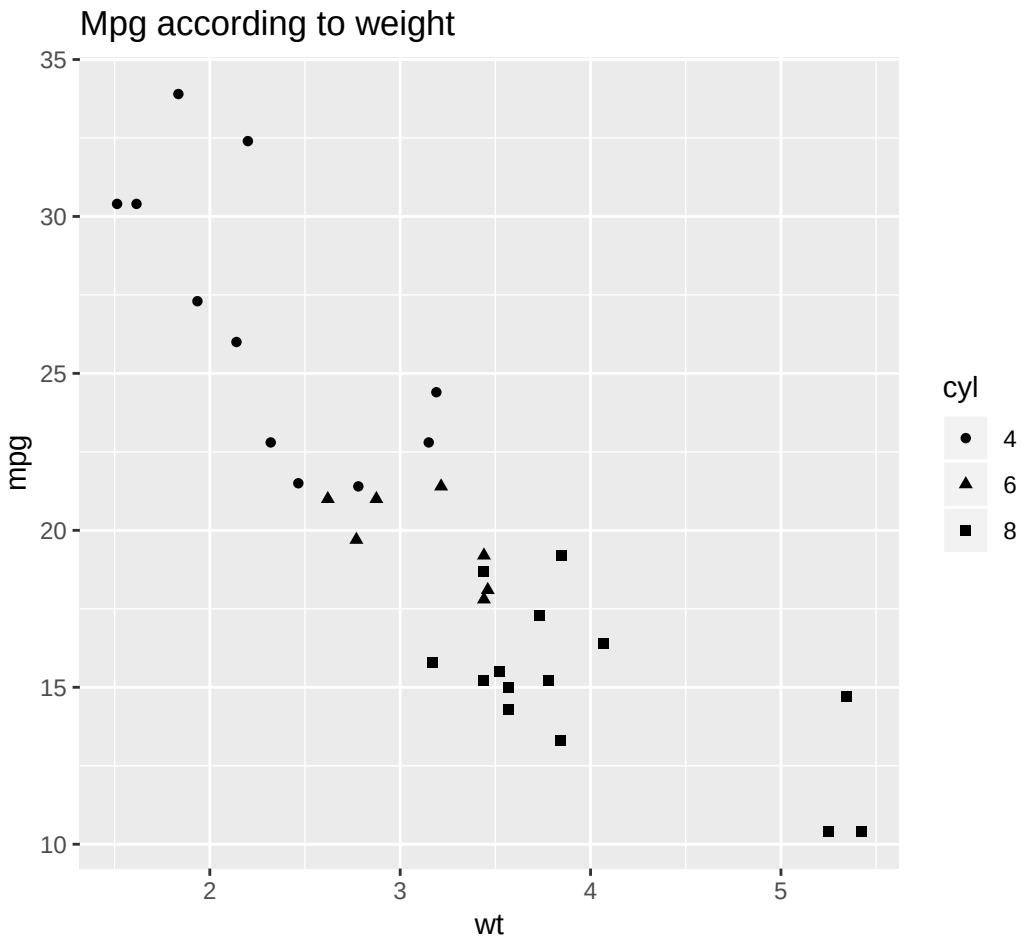
```
> ggplot(Car, aes(x=wt, y=mpg))+geom_point()+ggtitle("Mpg according to weight")
```



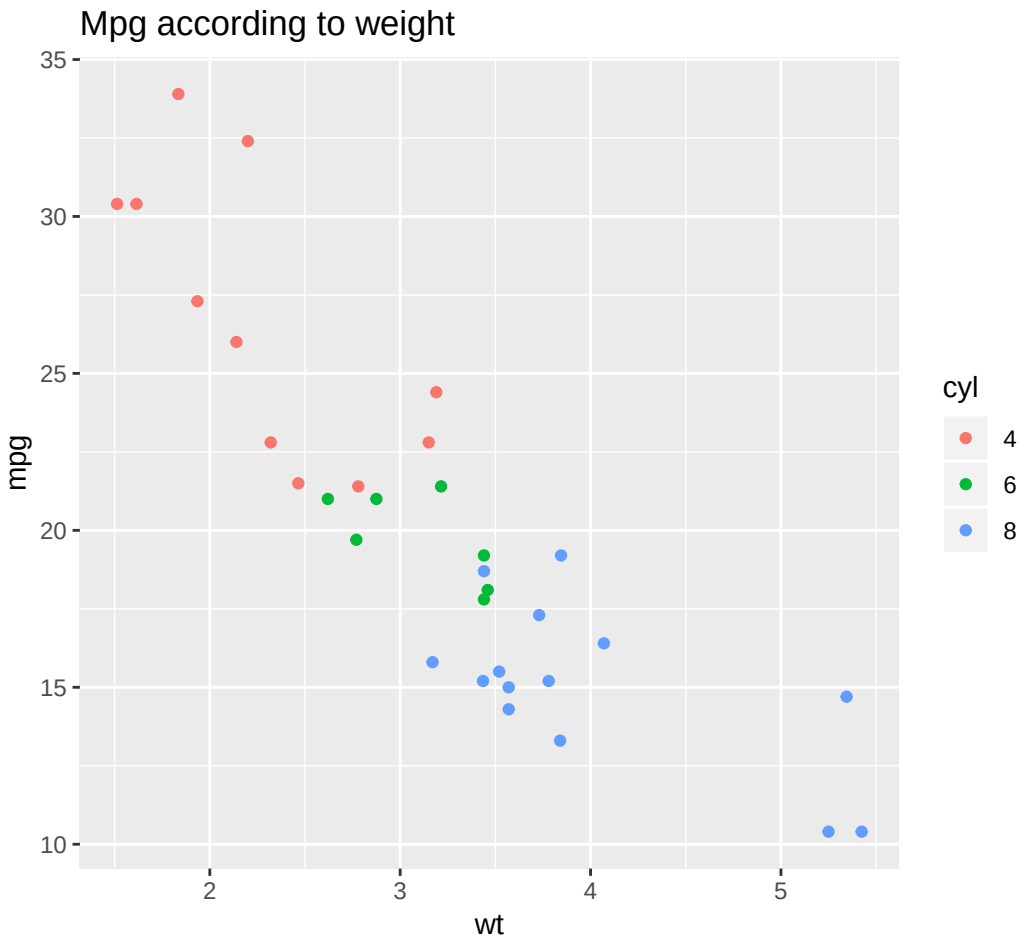
Scatterplot of mpg versus wt

Also want to group by cyl

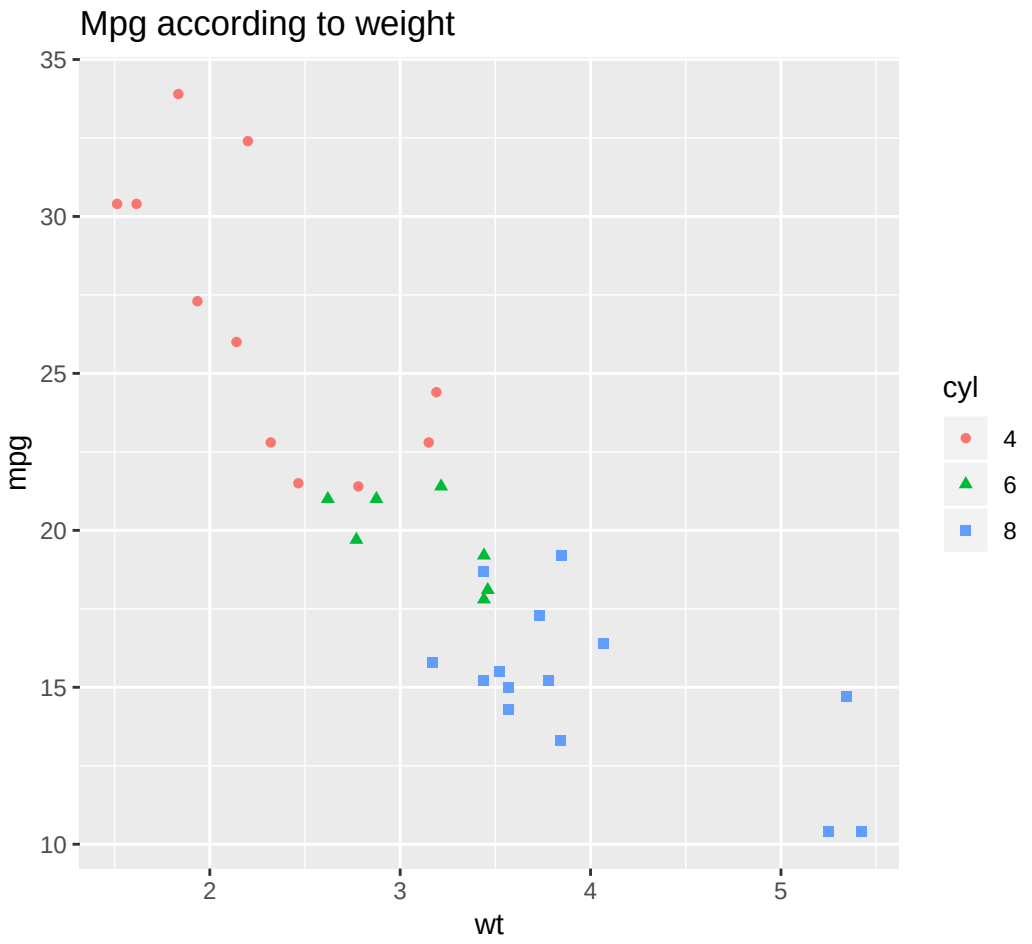
```
> #Shape
> ggplot(Car, aes(x=wt, y=mpg, shape=cyl ))+geom_point()+ggtitle("Mpg according to weight")
```



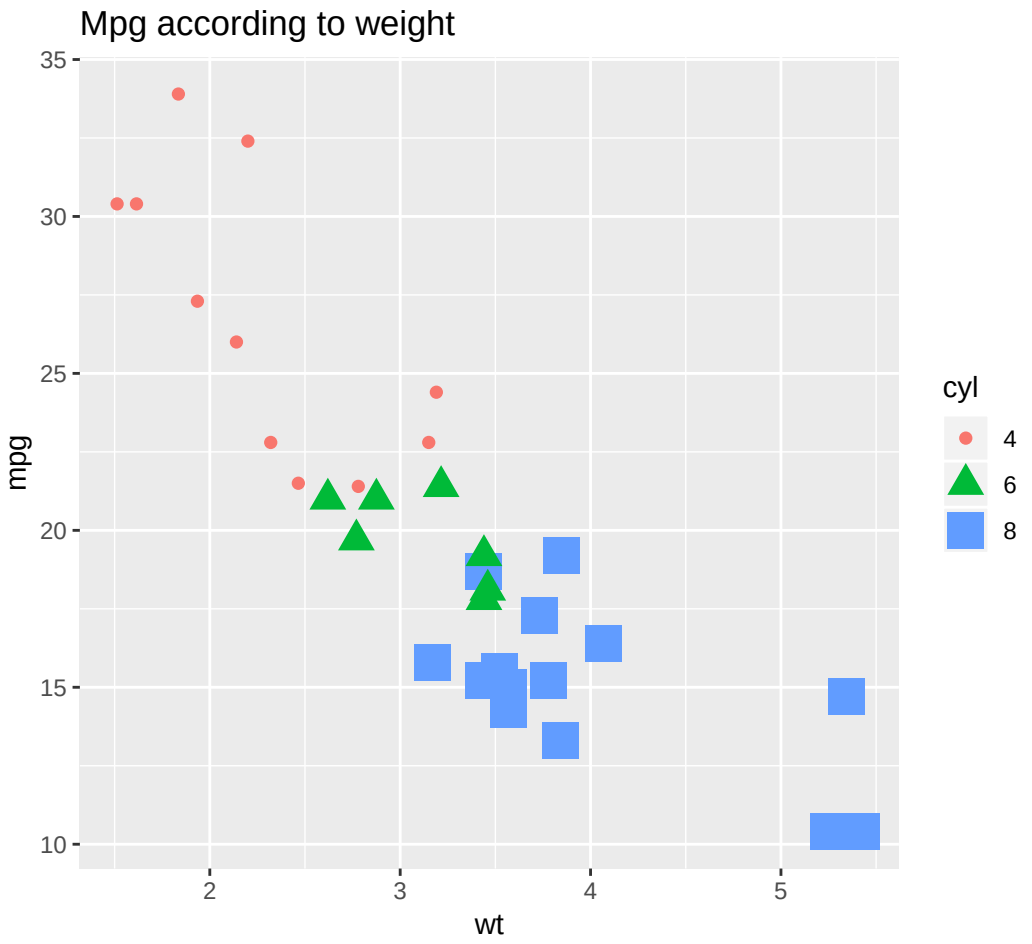

```
> #Color
> ggplot(Car, aes(x=wt, y=mpg,color=cyl ))+geom_point()+ggtitle("Mpg according to weight")
```




```
> #Color and Shape
> ggplot(Car, aes(x=wt, y=mpg, color=cyl, shape=cyl))+geom_point()+ggtitle("Mpg according to weight")
```

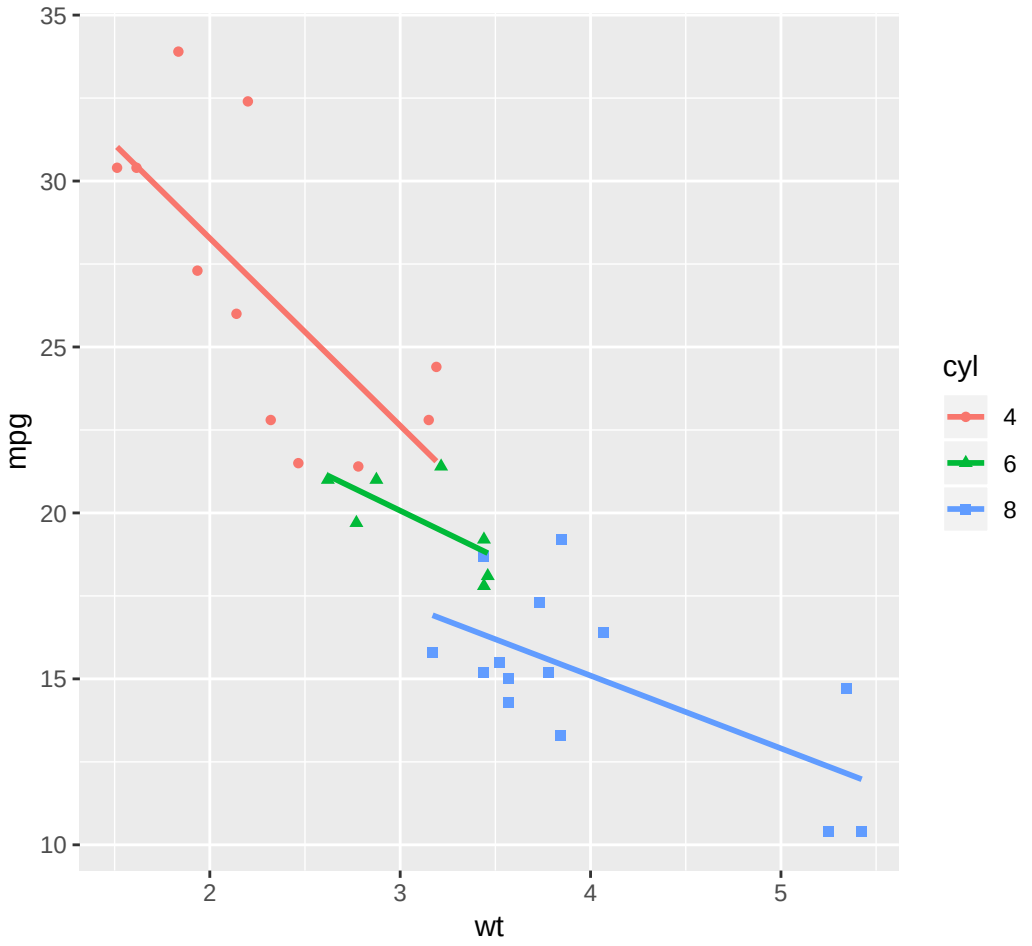



```
> #Color Shape and Size
> ggplot(Car, aes(x=wt, y=mpg, color=cyl, shape=cyl, size=cyl ))+geom_point()+ggtitle("Mpg according to
```

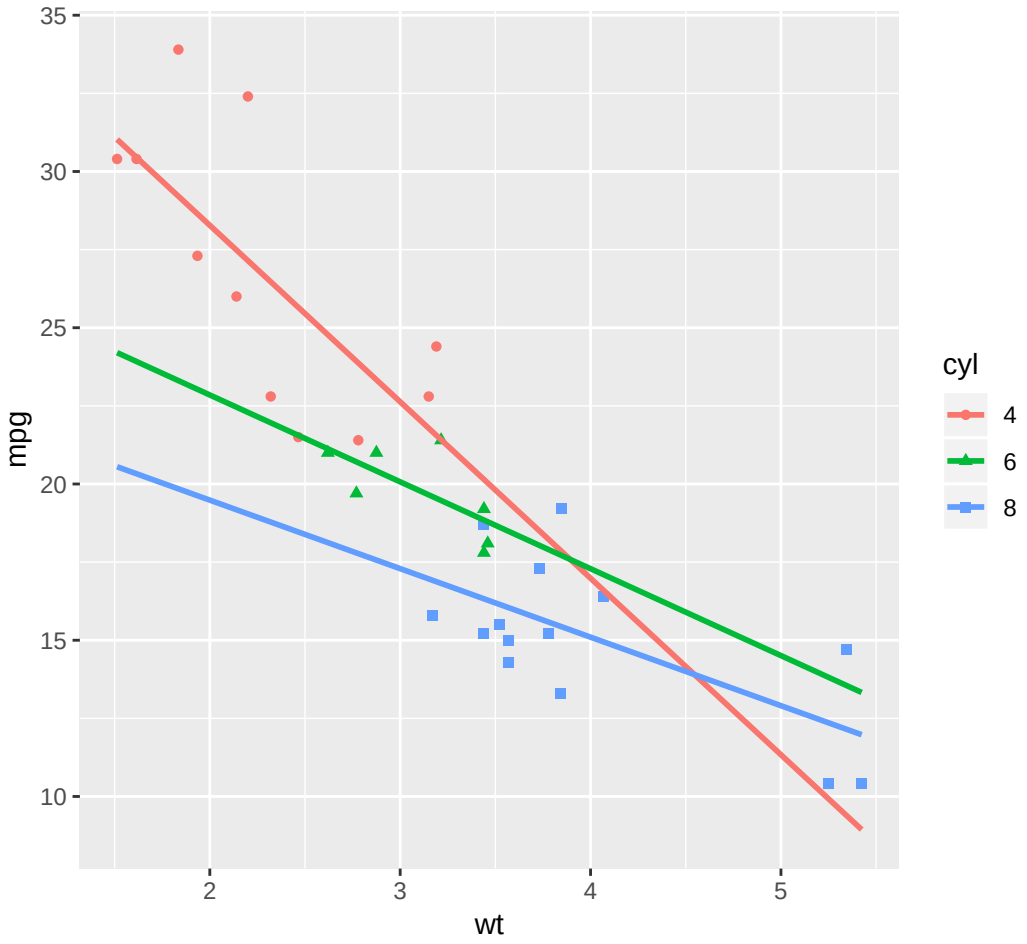


Scatterplot - adding linear regression line

```
> #Adding line - se is for standard error/confidence intervals, method = lm is for linear model
> ggplot(Car, aes(x=wt, y=mpg, color=cyl, shape=cyl)) +
+   geom_point() +   geom_smooth(method=lm, se=FALSE)
```




```
> #Adding full range line
> ggplot(Car, aes(x=wt, y=mpg, color=cyl, shape=cyl)) +
+   geom_point() +   geom_smooth(method=lm, se=FALSE, fullrange=TRUE)
```



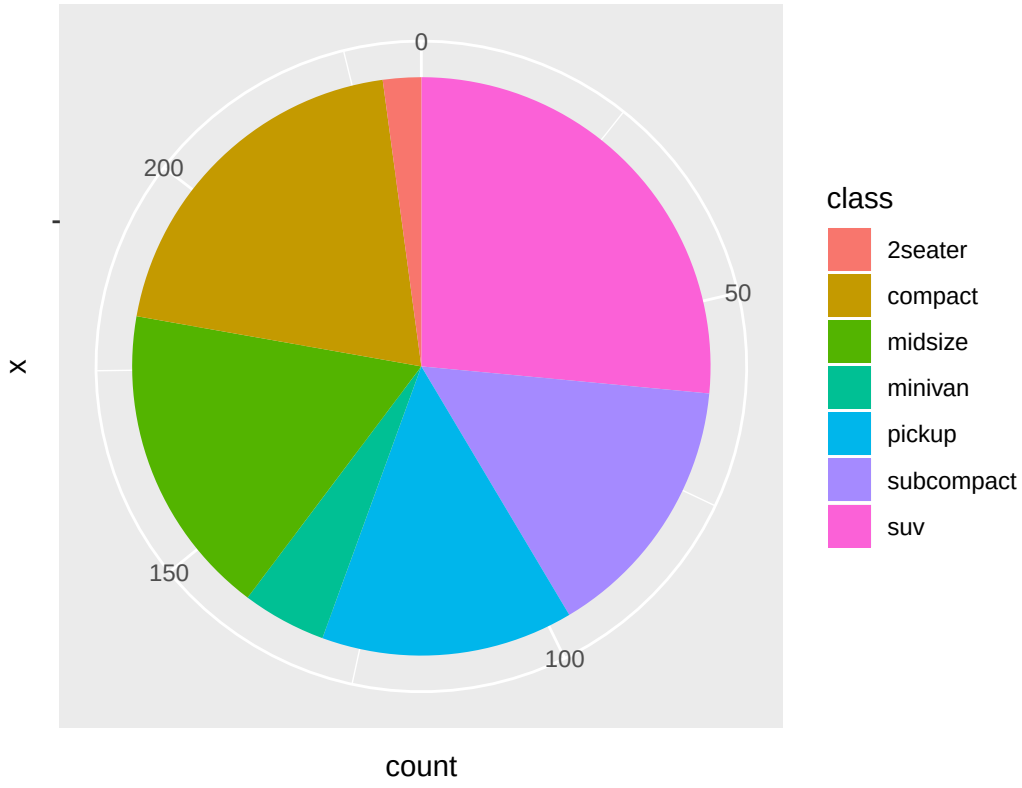
Pie Chart

Using `mpg` dataset in `{ggplot2}`

```
> Efficiency=mpg %>% mutate(class=factor(class))
> glimpse(Efficiency)
Observations: 234
Variables: 11
$ manufacturer <chr> "audi", "audi", "audi", "audi", "audi...
$ model        <chr> "a4", "a4", "a4", "a4", "a4", "a4", "...
$ displ        <dbl> 1.8, 1.8, 2.0, 2.0, 2.8, 2.8, 3.1, 1....
$ year         <int> 1999, 1999, 2008, 2008, 1999, 1999, 2...
$ cyl          <int> 4, 4, 4, 4, 6, 6, 6, 4, 4, 4, 4, 6, 6...
$ trans        <chr> "auto(l5)", "manual(m5)", "manual(m6)...
$ drv          <chr> "f", "f", "f", "f", "f", "f", "f", "4...
$ cty          <int> 18, 21, 20, 21, 16, 18, 18, 18, 16, 2...
$ hwy          <int> 29, 29, 31, 30, 26, 26, 27, 26, 25, 2...
$ fl           <chr> "p", "p", "p", "p", "p", "p", "p", "p...
$ class        <fct> compact, compact, compact, compact, c...
```

Pie Chart

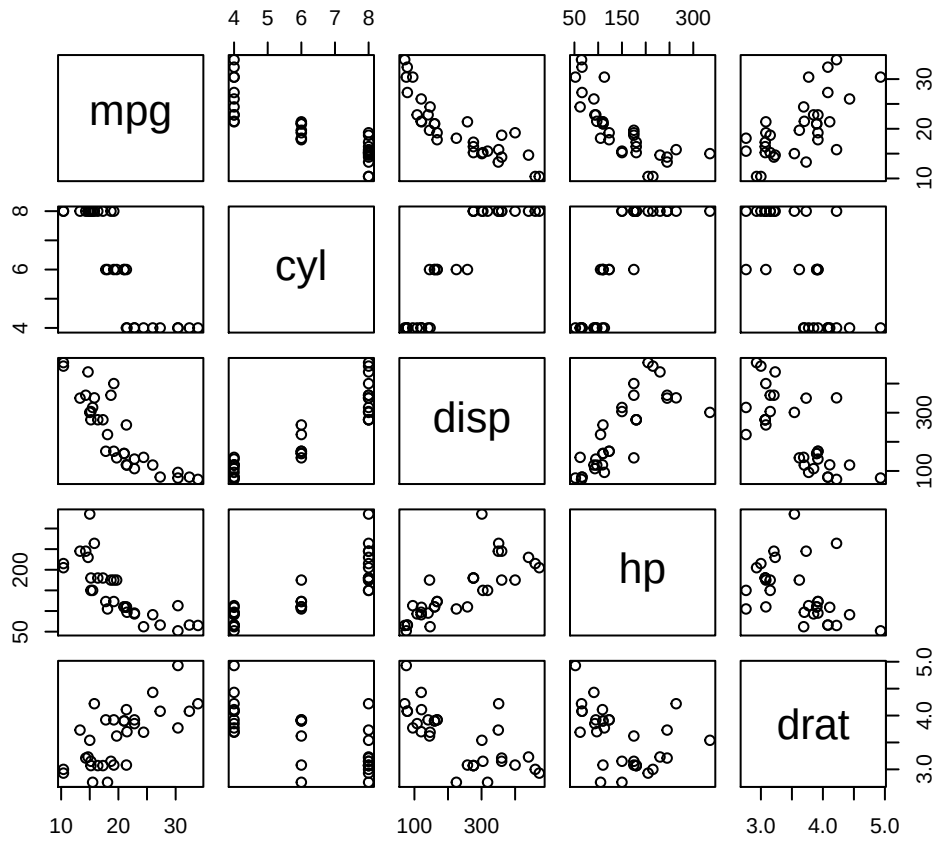
```
> Eff2=Efficiency%>%group_by(class)%>% summarize(count=n())
> ggplot(Eff2, aes(x=" ",y=count,fill=class))+geom_bar(width=1, stat='identity') +coord_polar("y")
```



Correlation plot

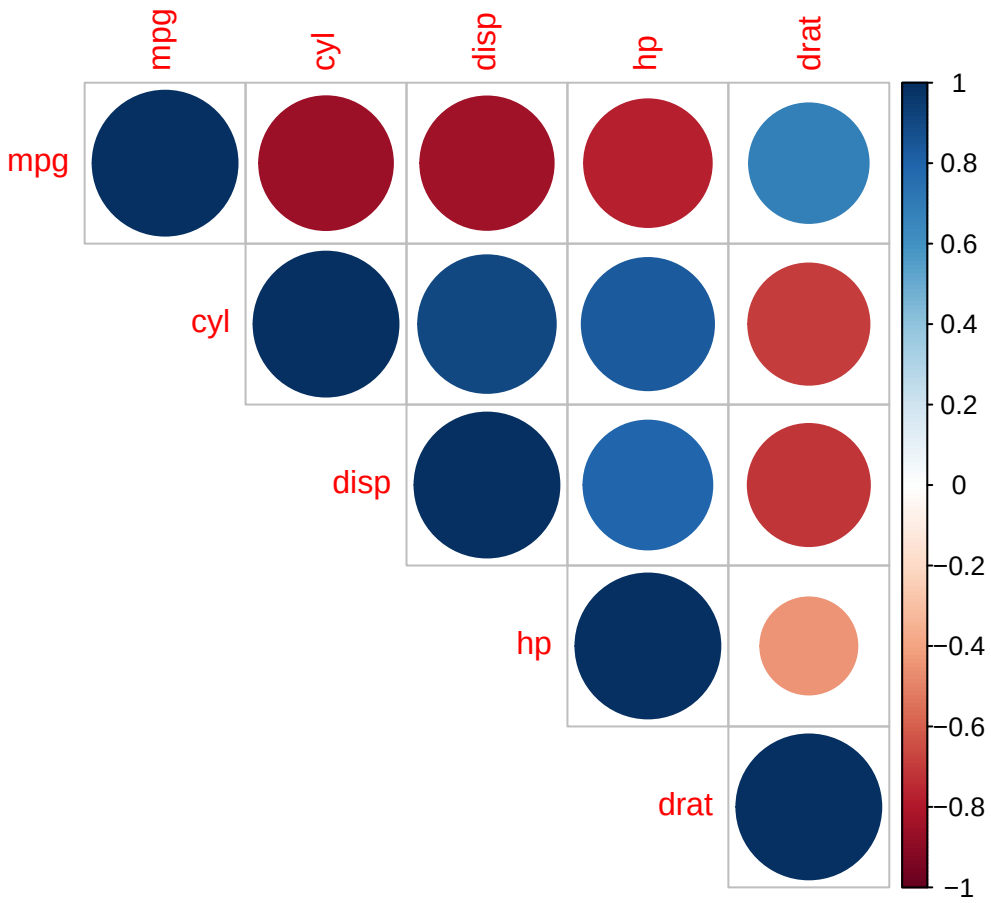
Using `mtcars` dataset in `{datasets}`

```
> #Selecting variables to find correlation of
> Data=mtcars
> Data2=Data%>%select(mpg:drat)
> plot(Data2)
```



Correlation Plot 2

```
> #install.packages("corrplot")
> library(corrplot)
>
> corr_matrix=cor(Data2)
> corrplot(corr_matrix, type="upper")
```


mpg
cyl
disp
hp
drat
1
0.8
0.6
0.4
0.2
0
−0.2
−0.4
−0.6
−0.8
−1