ggplot from the scratch

shyamjith c

2023-10-15

Introduction

Data is huge and it is everywhere but along with that comes the need to understand data and base our decisions after drawing inferences from data.

One of the major steps that we have in the field in data science is first exploring the data thereby presenting it in the form of informative plots and it is referred to as Data Visualization.

We are very visual creatures as a large portion of brain dedicates itself to visual processing. Images are able to grab our attention easily, we are immediately drawn to them.

John Tukey was an American mathematician best known for the development of the FFT algorithm and boxplot.Here are some of his notable quotes on the importance of visualization through graphs.

> "The simple graph has brought more information to the data analyst's mind than any other device." — John Tukey

Step 1: installing ggplot2 library

```r
#install.packages("ggplot2") # if not installed
require("ggplot2") # check whether the package is available if not install
```

```
## Loading required package: ggplot2
```

```r
library(ggplot2)
```

Loading the dataset

```r
mydata=mtcars
str(mydata)
```

```
## 'data.frame':    32 obs. of  11 variables:
##  $ mpg : num  21 21 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 ...
##  $ cyl : num  6 6 4 6 8 6 8 4 4 6 ...
##  $ disp: num  160 160 108 258 360 ...
##  $ hp  : num  110 110 93 110 175 105 245 62 95 123 ...
##  $ drat: num  3.9 3.9 3.85 3.08 3.15 2.76 3.21 3.69 3.92 3.92 ...
##  $ wt  : num  2.62 2.88 2.32 3.21 3.44 ...
##  $ qsec: num  16.5 17 18.6 19.4 17 ...
##  $ vs  : num  0 0 1 1 0 1 0 1 1 1 ...
##  $ am  : num  1 1 1 0 0 0 0 0 0 0 ...
##  $ gear: num  4 4 4 3 3 3 3 4 4 4 ...
##  $ carb: num  4 4 1 1 2 1 4 2 2 4 ...
```

Working with `ggplot`

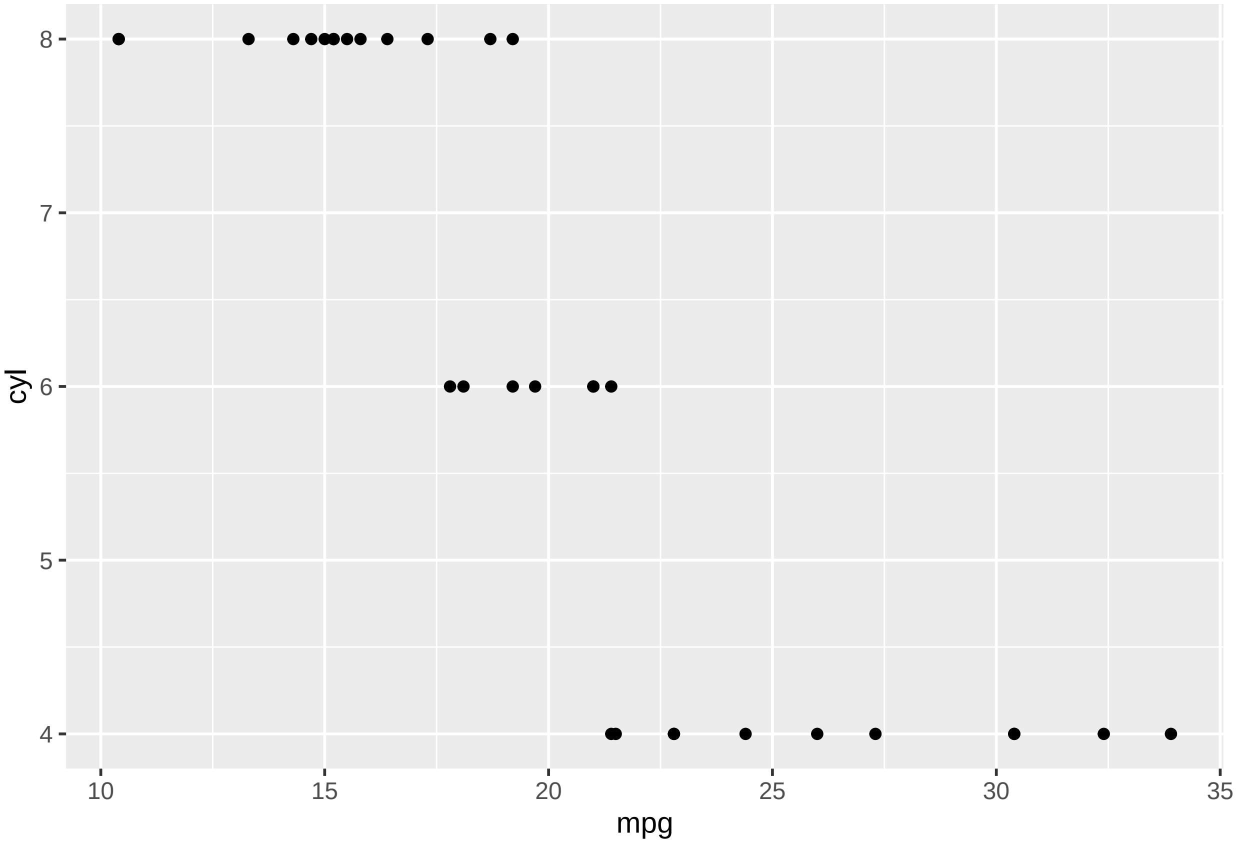
Scatterplots:

They are used to represent distribution between different variables in form of points scattered all over. They are are most basic and simple to make and every data point gets a chance to be represented however it becomes somewhat less distinct to see the image unlike the case with a line chart.

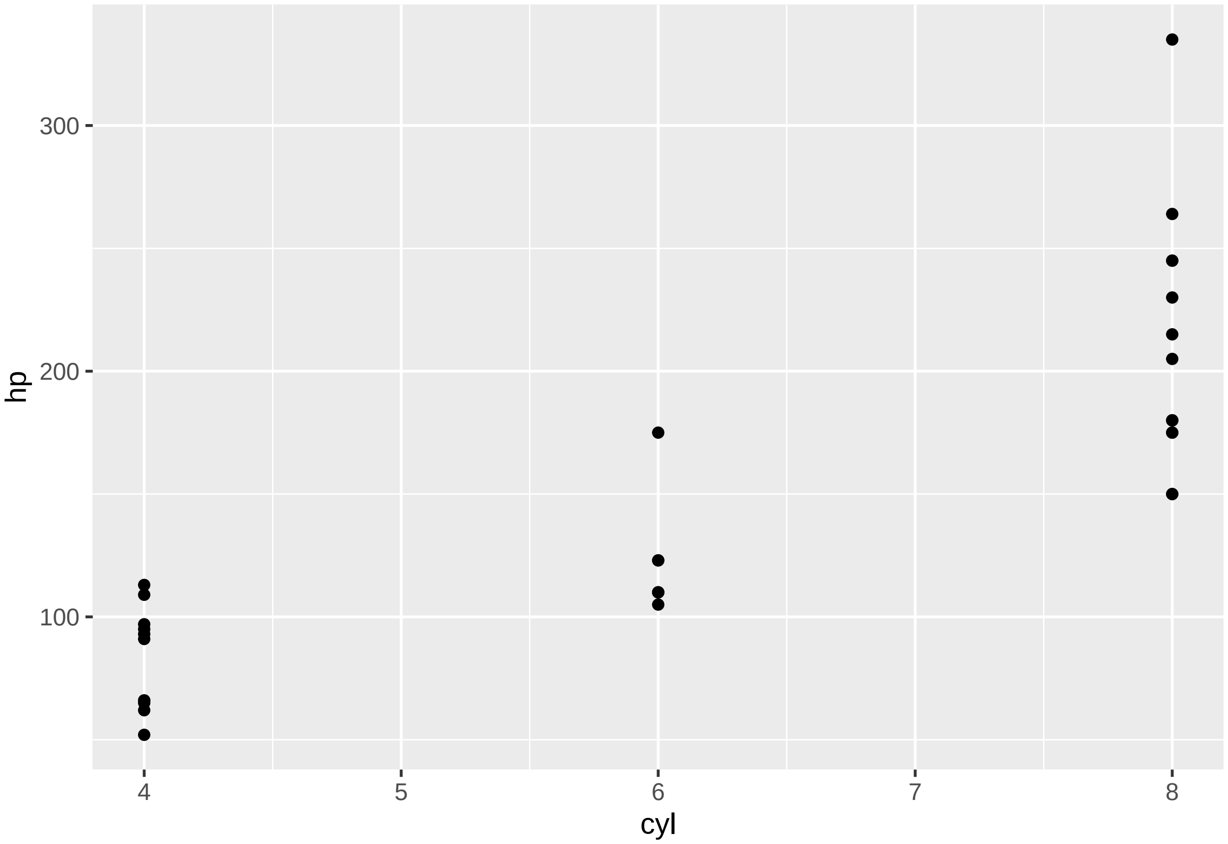
Language of Grammer of Graphics: A `geom` is a geometrical object that a plot uses to represent data and different plots use different `geoms` like bar chart uses bar `geoms` , line chart uses line `geoms`, boxplot uses boxplot `geoms`.

mappings are used to map different aesthetics (written in the brackets under aes) assigning different axis x and y to variables and other aesthetics like color, fill also associated with other variables to be represented in the plot.

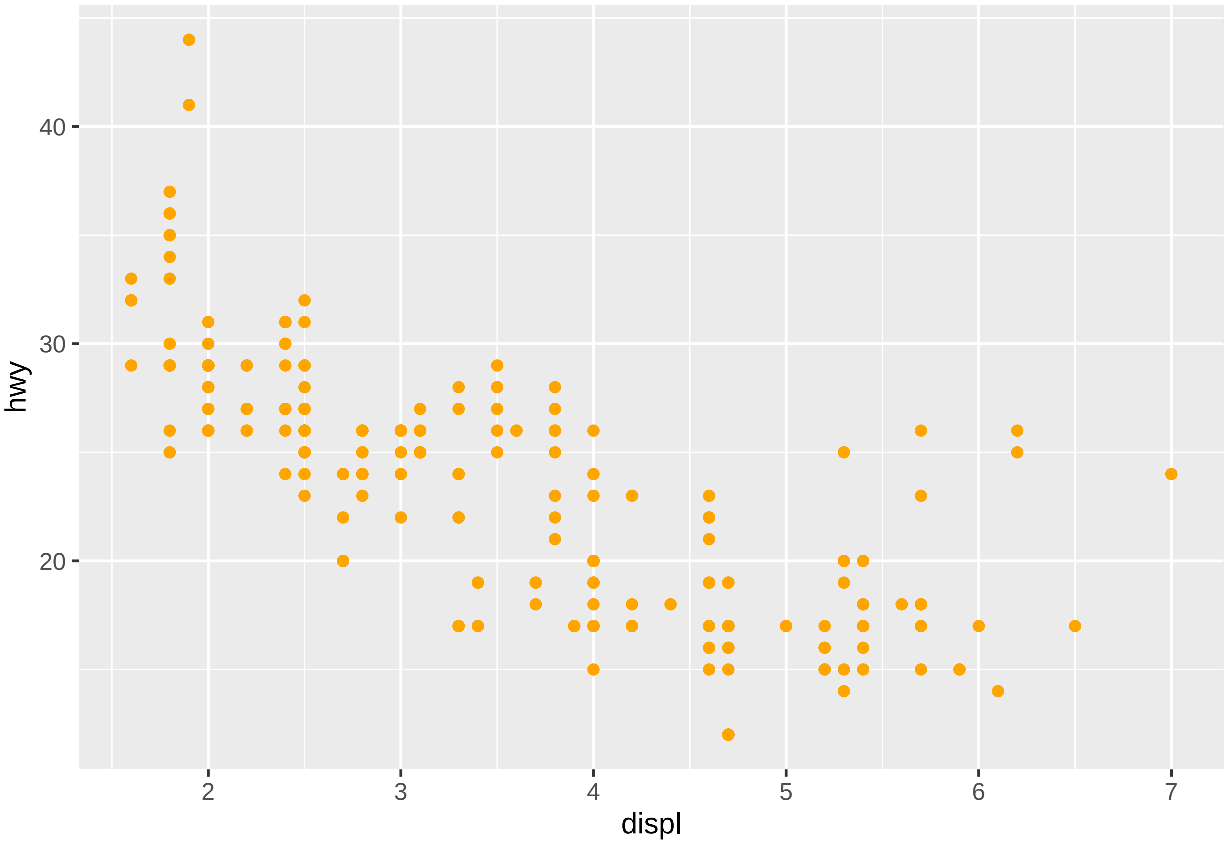
```r
ggplot(data = mydata)+
geom_point(mapping = aes (x = mpg, y= cyl))# giving  different axis to different variables
```



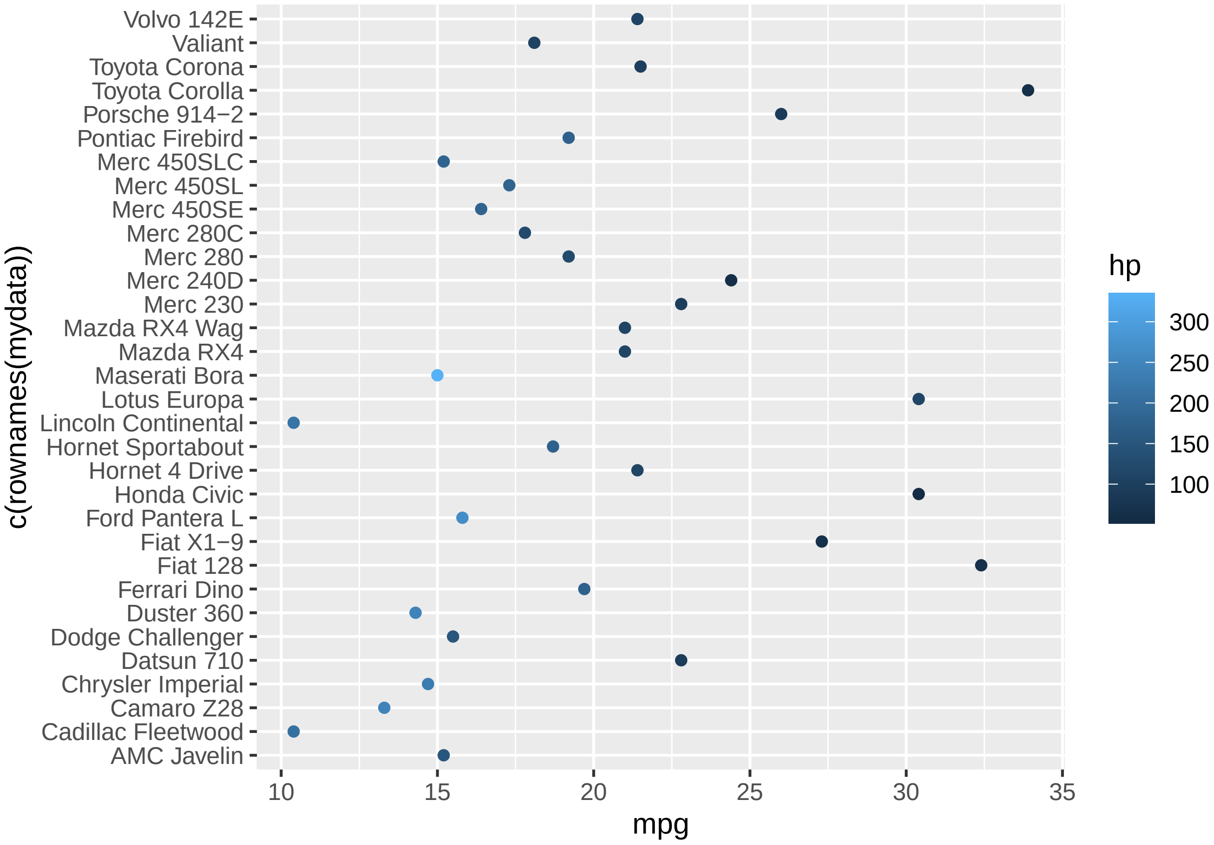
```r
ggplot(data = mtcars)+
geom_point(mapping = aes(x = cyl, y = hp))
```



```r
ggplot(data = mpg) +
geom_point(mapping = aes(x = displ, y = hwy), color = "orange")
```



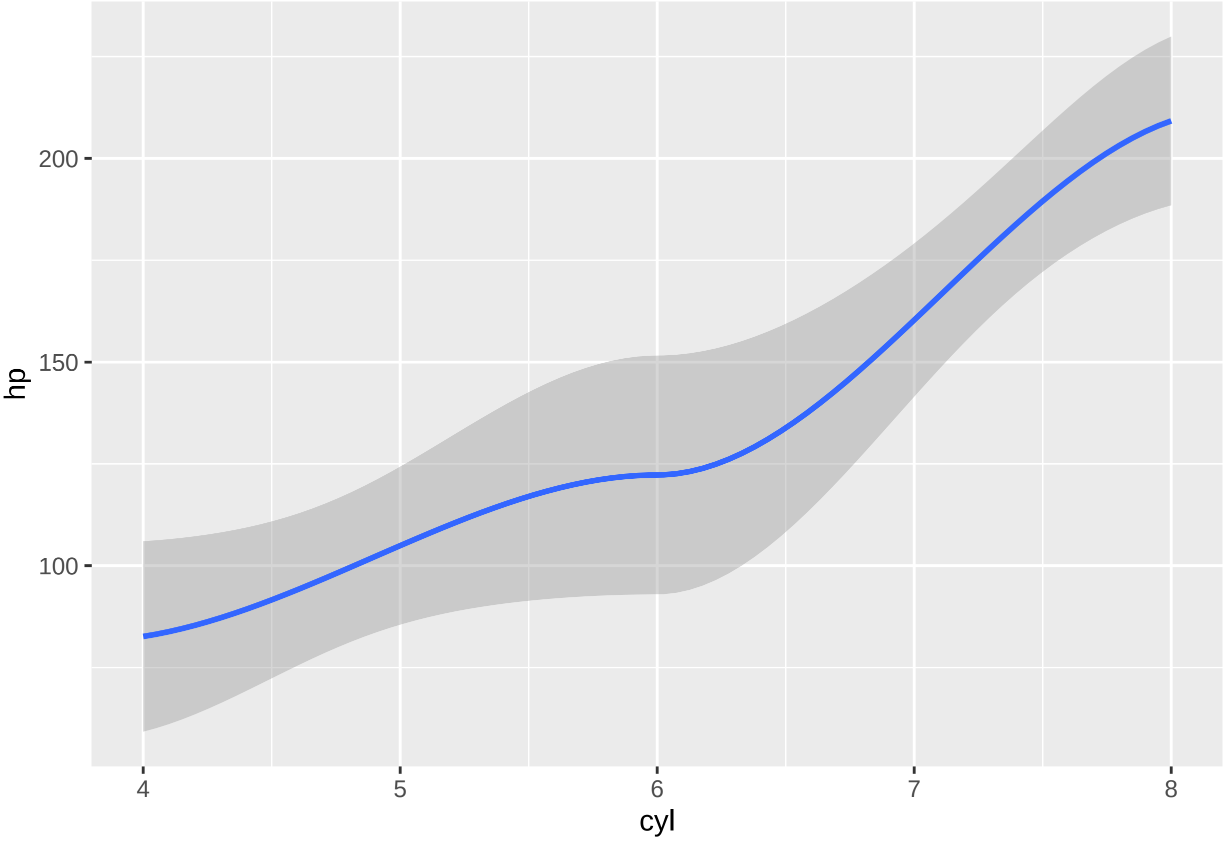
```
ggplot(data = mydata)+
geom_point(mapping = aes(x = mpg , y = c(rownames(mydata)),color =hp))
```



Line chart:

```
ggplot(data = mydata) +
geom_smooth(mapping = aes(x = cyl, y = hp))
```

```
## `geom_smooth()` using method = 'loess' and formula = 'y ~ x'

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric = parametric,
## : pseudoinverse used at 3.98

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric = parametric,
## : neighborhood radius 4.02

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric = parametric,
## : reciprocal condition number 1.2905e-16

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric = parametric,
## : There are other near singularities as well. 16.16

## Warning in predLoess(object$y, object$x, newx = if (is.null(newdata)) object$x
## else if (is.data.frame(newdata))
## as.matrix(model.frame(delete.response(terms(object)), : pseudoinverse used at
## 3.98

## Warning in predLoess(object$y, object$x, newx = if (is.null(newdata)) object$x
## else if (is.data.frame(newdata))
## as.matrix(model.frame(delete.response(terms(object)), : neighborhood radius
## 4.02

## Warning in predLoess(object$y, object$x, newx = if (is.null(newdata)) object$x
## else if (is.data.frame(newdata))
## as.matrix(model.frame(delete.response(terms(object)), : reciprocal condition
## number 1.2905e-16

## Warning in predLoess(object$y, object$x, newx = if (is.null(newdata)) object$x
## else if (is.data.frame(newdata))
## as.matrix(model.frame(delete.response(terms(object)), : There are other near
## singularities as well. 16.16
```

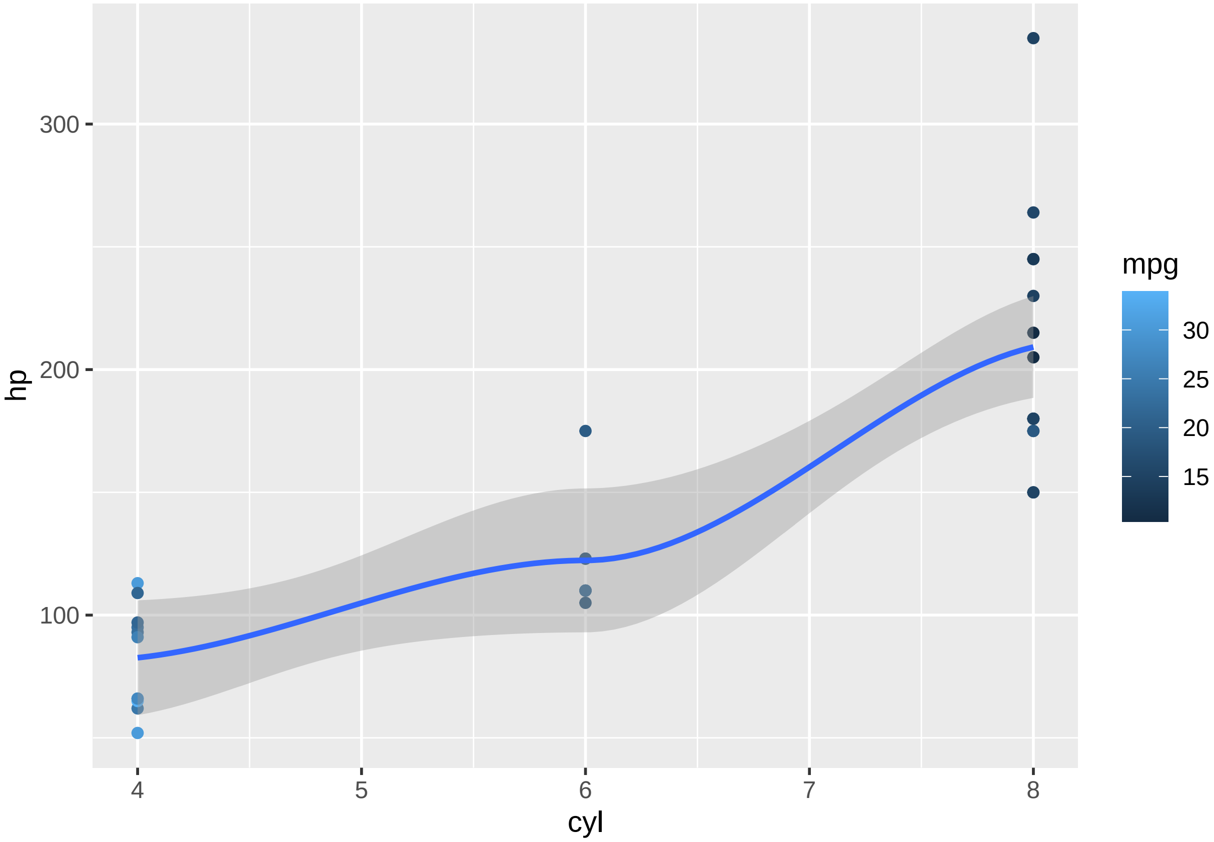


```r
ggplot(data = mydata, mapping = aes(x = cyl, y = hp),method=NULL) +
 geom_point(mapping = aes(color = mpg)) +
 geom_smooth()
```

```
## `geom_smooth()` using method = 'loess' and formula = 'y ~ x'

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric = parametric,
## : pseudoinverse used at 3.98

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric = parametric,
## : neighborhood radius 4.02

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric = parametric,
## : reciprocal condition number 1.2905e-16

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric = parametric,
## : There are other near singularities as well. 16.16

## Warning in predLoess(object$y, object$x, newx = if (is.null(newdata)) object$x
## else if (is.data.frame(newdata))
## as.matrix(model.frame(delete.response(terms(object)), : pseudoinverse used at
## 3.98
## Warning in predLoess(object$y, object$x, newx = if (is.null(newdata)) object$x
## else if (is.data.frame(newdata))
## as.matrix(model.frame(delete.response(terms(object)), : neighborhood radius
## 4.02
```

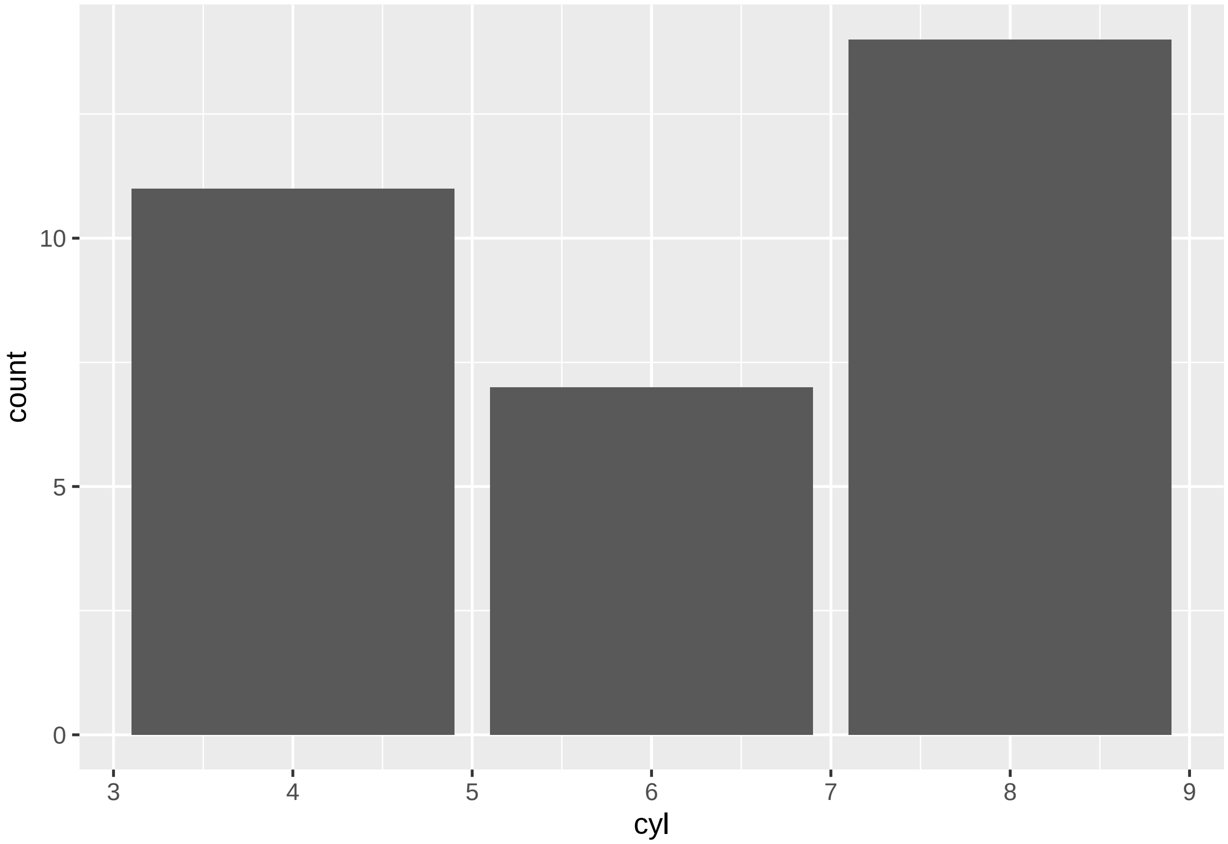
```
## Warning in predLoess(object$y, object$x, newx = if (is.null(newdata)) object$x
## else if (is.data.frame(newdata))
## as.matrix(model.frame(delete.response(terms(object)), : reciprocal condition
## number 1.2905e-16
```

```
## Warning in predLoess(object$y, object$x, newx = if (is.null(newdata)) object$x
## else if (is.data.frame(newdata))
## as.matrix(model.frame(delete.response(terms(object)), : There are other near
## singularities as well. 16.16
```

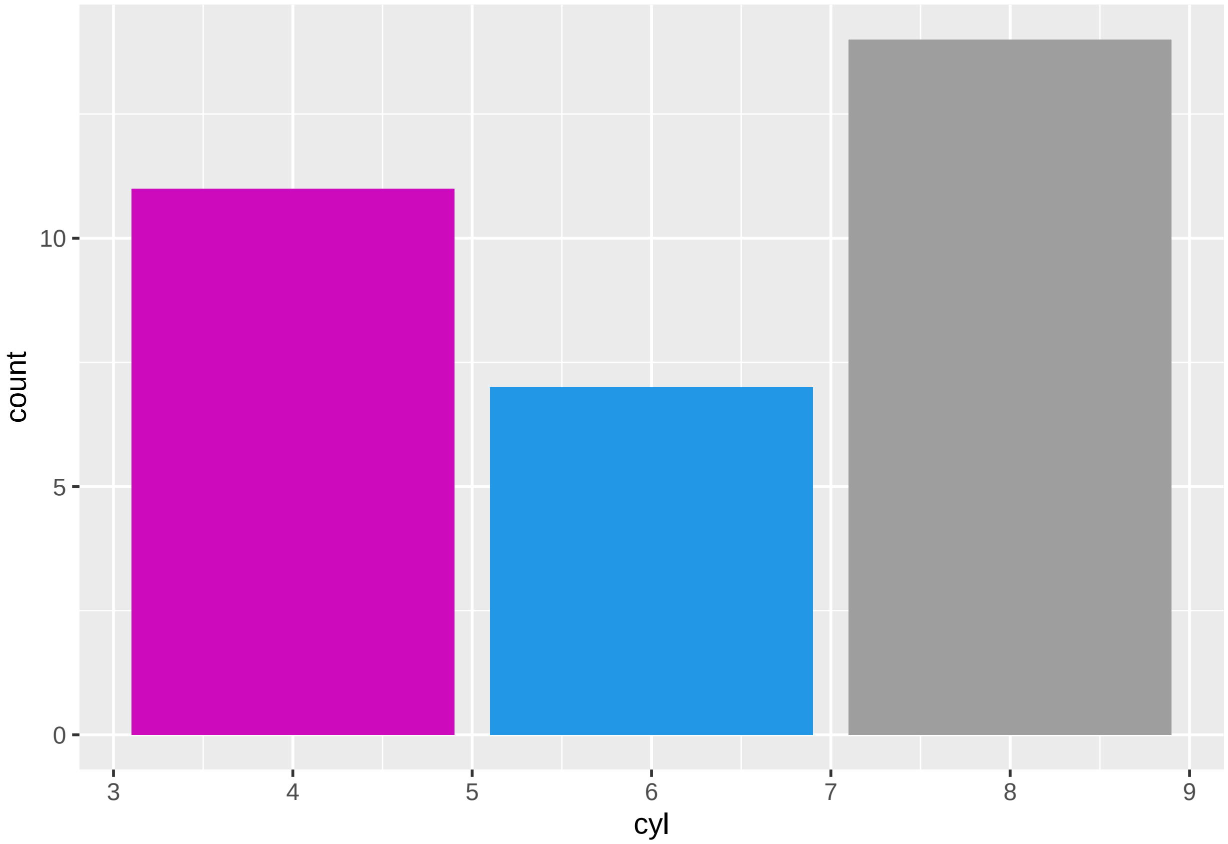


Bar charts:

```
ggplot(data = mydata) +
 geom_bar(mapping = aes(x = cyl))
```

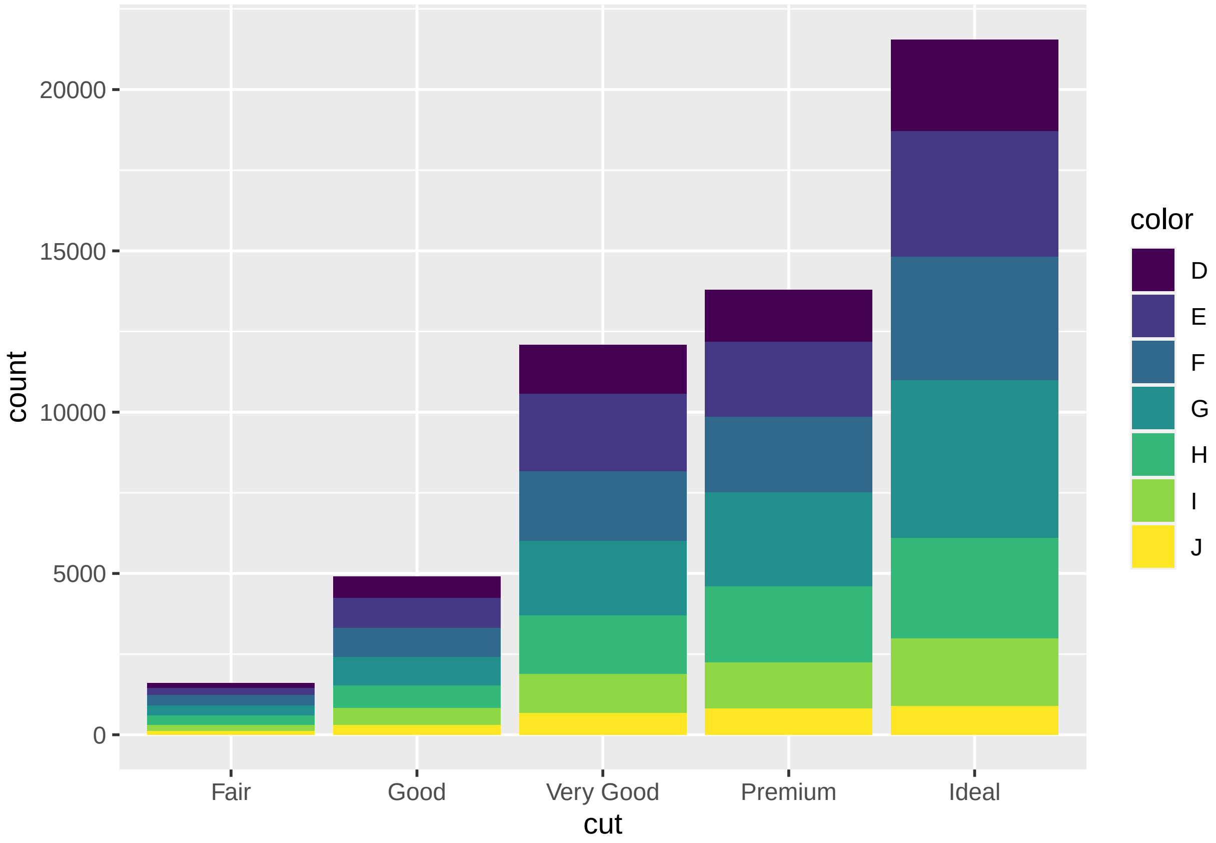


```
ggplot(data = mydata) +
 geom_bar(mapping = aes(x = cyl),fill=unique(mydata$cyl))
```

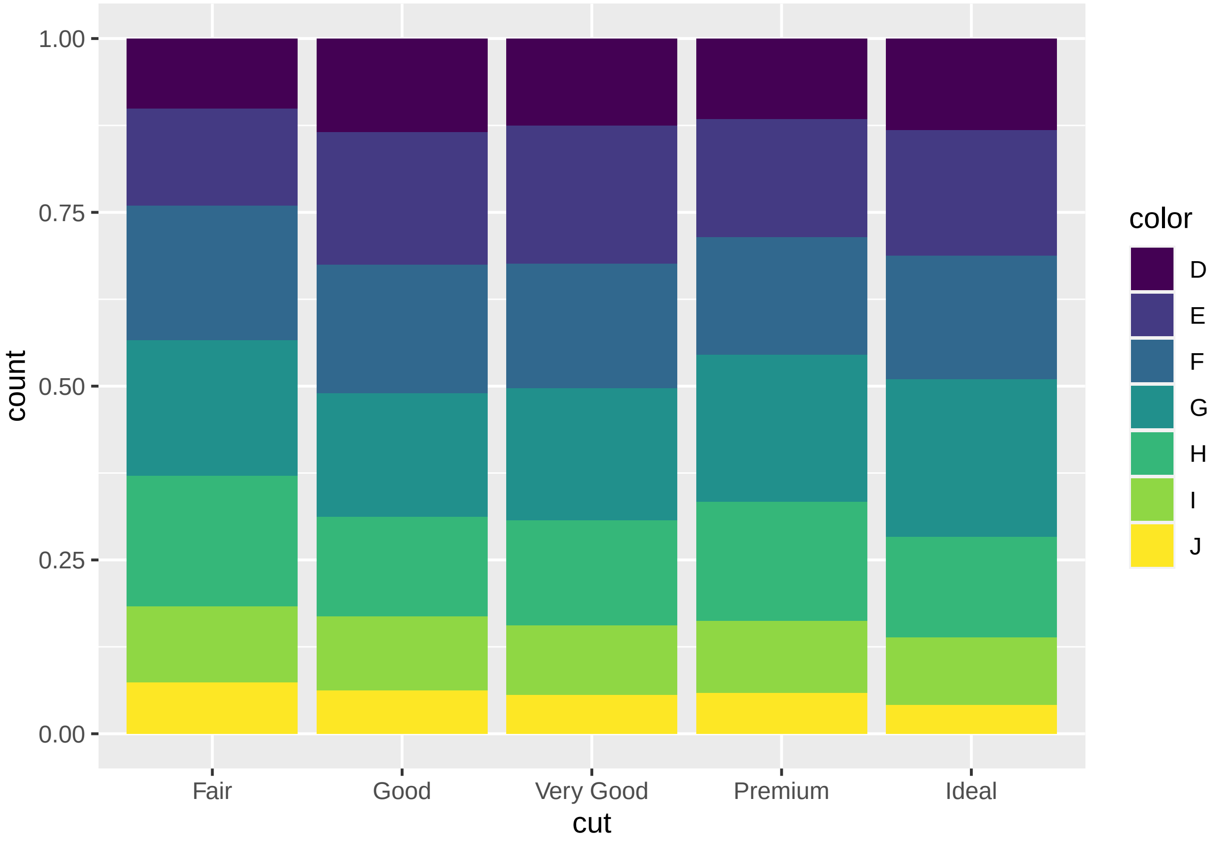


Another example:

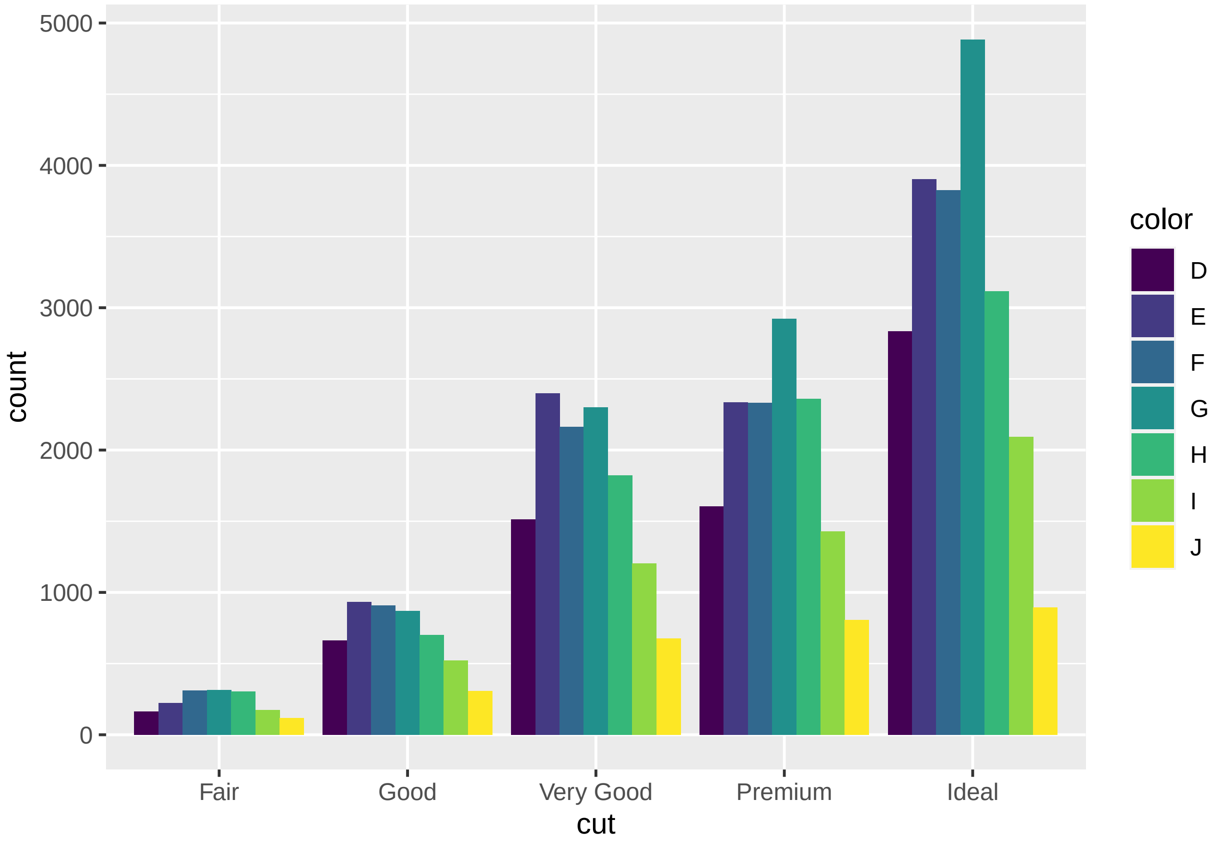
```
ggplot(data = diamonds) +
  geom_bar(mapping = aes(x = cut, fill = color))
```



```r
ggplot(data = diamonds) +
  geom_bar(mapping = aes(x = cut, fill = color), position = "fill")
```



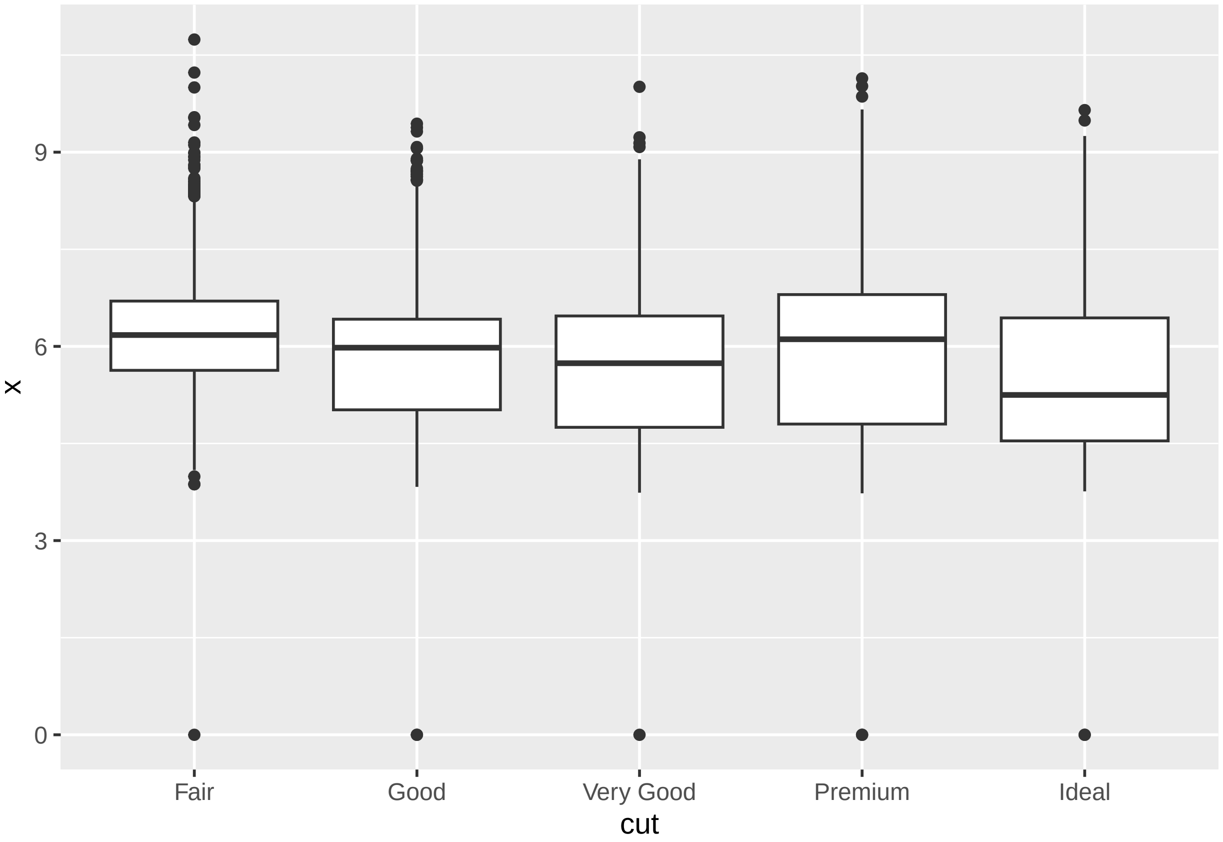
```r
ggplot(data = diamonds) +
    geom_bar(mapping = aes(x = cut, fill = color),position = "dodge")
```



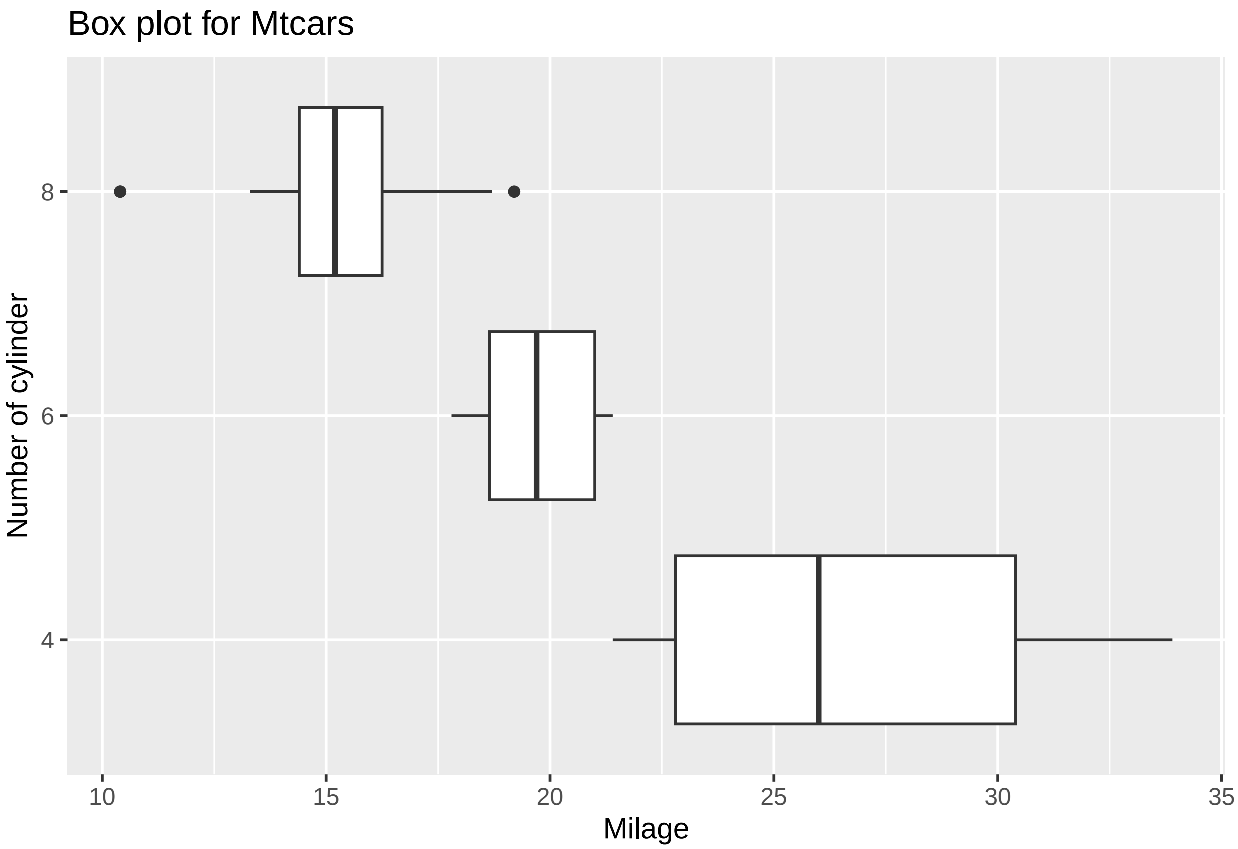
Boxplot:

They are mainly used to represent outliers and it is their ability to represent the difference between distributions and showing outliers for different categories of a variable.

```r
ggplot(data = diamonds, mapping = aes(x = cut, y = x)) +
  geom_boxplot()
```



```
mydata$cyl=as.factor(mydata$cyl)
p=ggplot(data = mydata, mapping = aes(x = cyl, y = mpg))+
  geom_boxplot()+
  coord_flip()
p+labs(title="Box plot for Mtcars",
        x ="Number of cylinder", y = "Milage")
```


Box plot for Mtcars
Number of cylinder
8
6
4
10
15
20
25
30
35
Milage