Visualization in R

Data Visualization

- Data visualization is a technique used for the graphical representation of data.
  - Eg. scatter plots, histograms, maps, etc.,
- Make our data more understandable
- Makes it easy to recognize patterns, trends, and exceptions in our data.
- Enables us to convey information and results in a quick and visual way.

Data Visualization in R

- Base Graphics
- Grid Graphics
- Lattice Graphics
- ggplot2

Basic plots

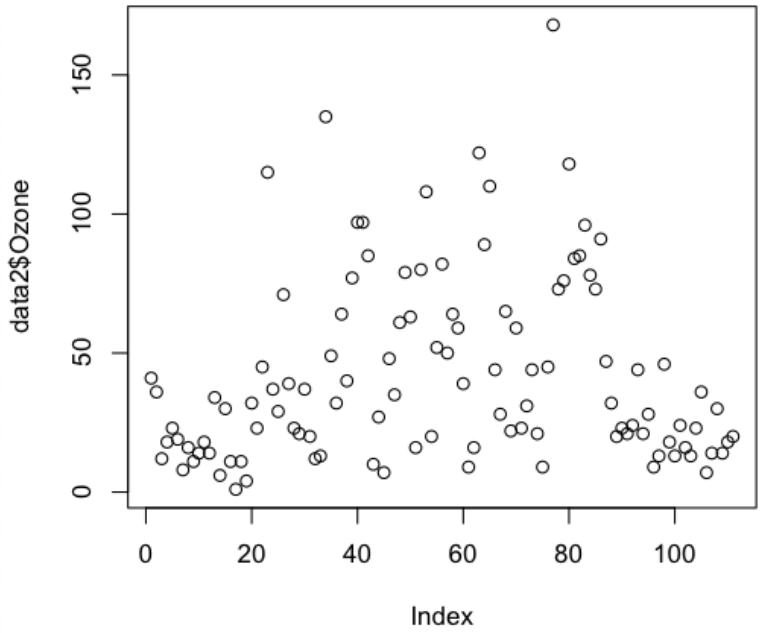
- The **graphics** package is used for plotting **base** graphs like scatter plot, box plot etc.
- A complete list of functions with help pages can be obtained by typing :

```
library(help = "graphics")
```

plot()

- The plot() function is a kind of a generic function for plotting of **R** objects.

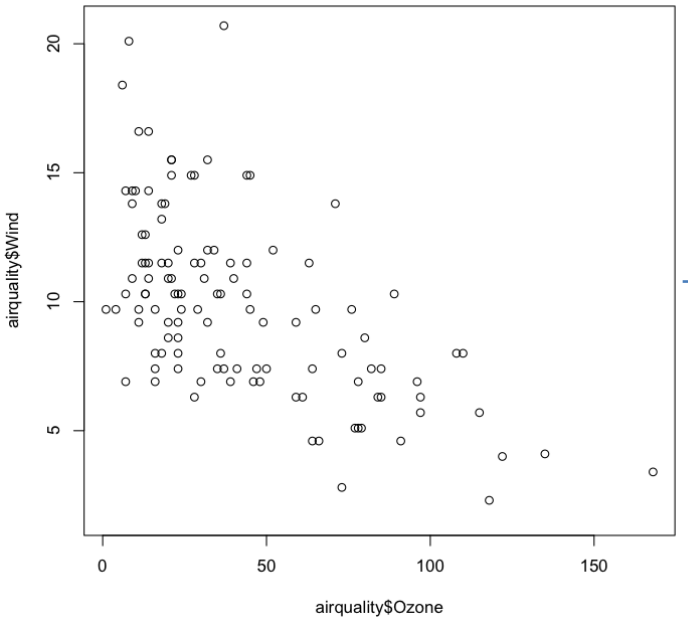
  plot(dat$Ozone)



→ 1D scatter plot

Scatter plot

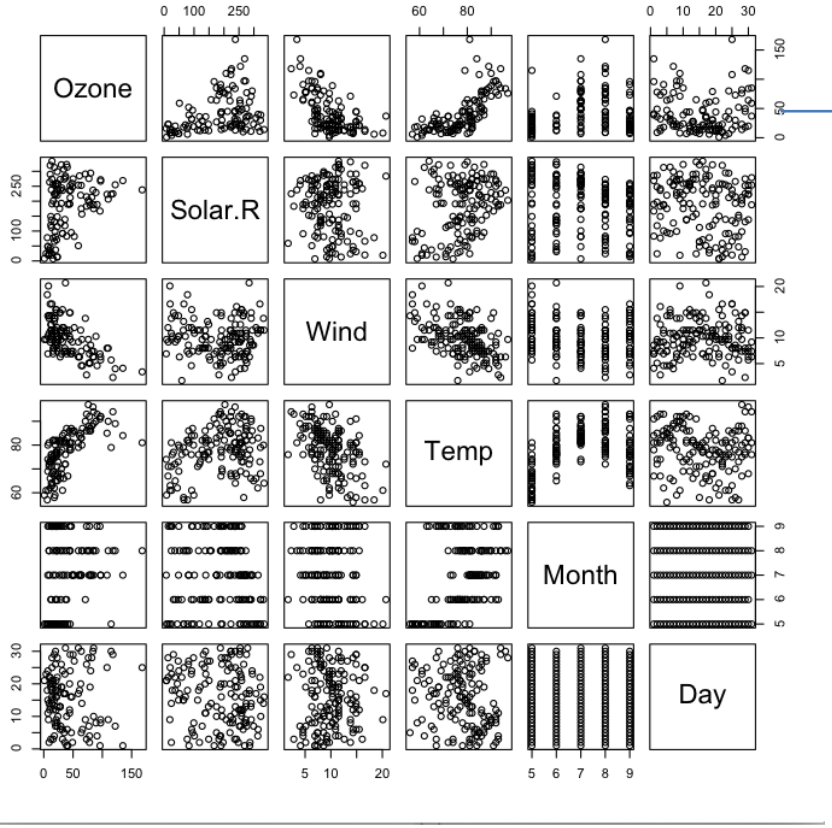
- Used to get relationship between two variables
  - To study the relationship between the Ozone and Wind values
- plot(dat$Ozone, dat$Wind)



→ Negative Correlation

Scatter plot (contd.)

- When plot command is used with the entire dataset, a matrix of scatterplots is obtained which is a correlation matrix of all the columns.



Ozone and Wind - Negative Correlation

Ozone and Temperature – Positive Correlation

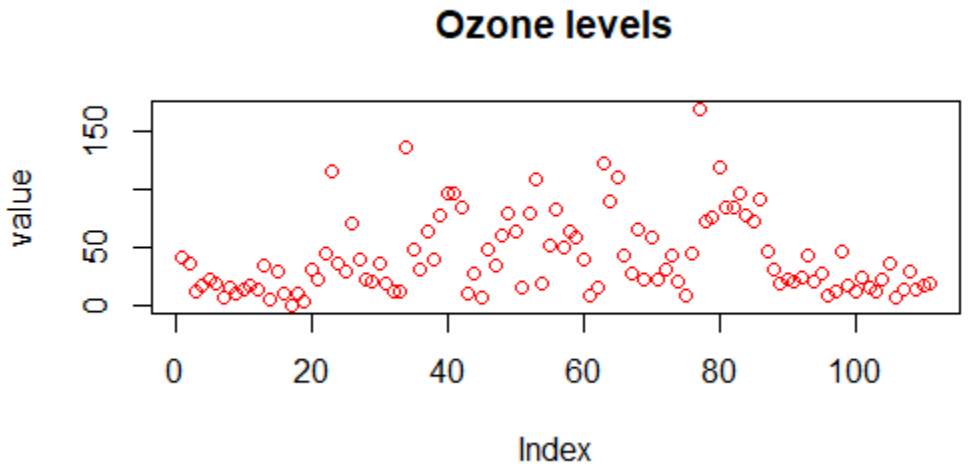
Wind and Temperature – Negative Correlation

Argument in plot()

- type argument
  - Take in values like **p: points**, **l: lines**,**b: both** etc. This decides the shape of the output graph.
  - **h:high density lines**

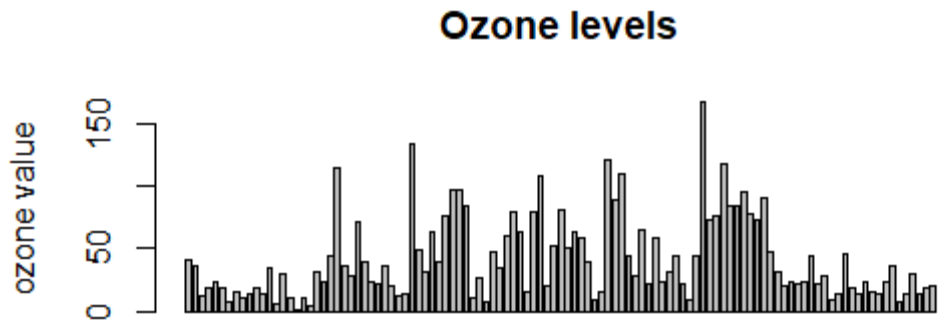
Argument in plot() – (contd.)

- Titles & Labels
  - main argument – Title
  - xlab, ylab argurments– x-axis & y-axis label respectively



Bar plot

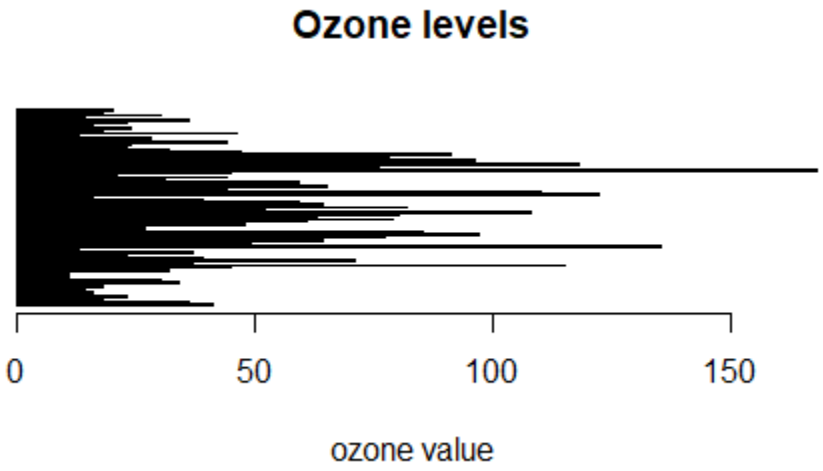
- Data is represented in the form of rectangular bars
- Length of the bar is proportional to the value of the variable
  - barplot(dat$Ozone, main = 'Ozone levels', ylab = 'ozone value')



Bar plot (contd.)

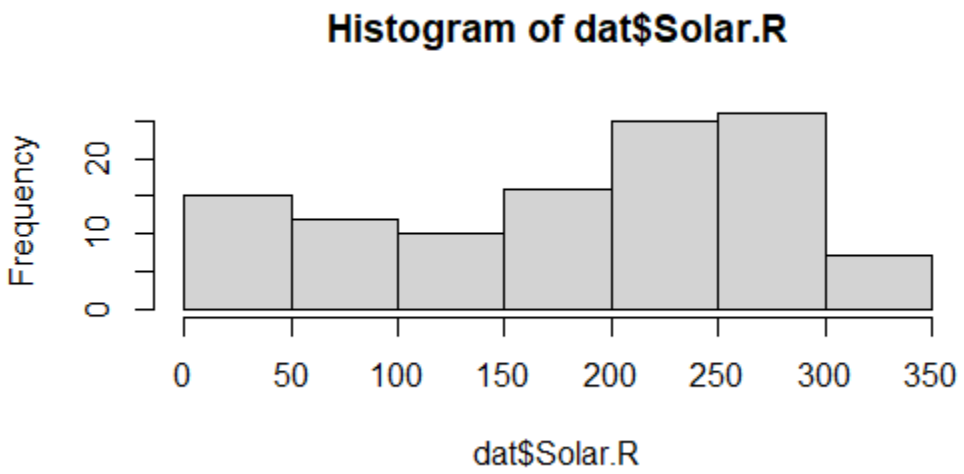
- Both horizontal, as well as a vertical bar chart, can be generated by tweaking the **horiz** parameter.

```
barplot(dat$Ozone, main = 'Ozone levels', xlab = 'ozone value',horiz = TRUE)
```



Histogram

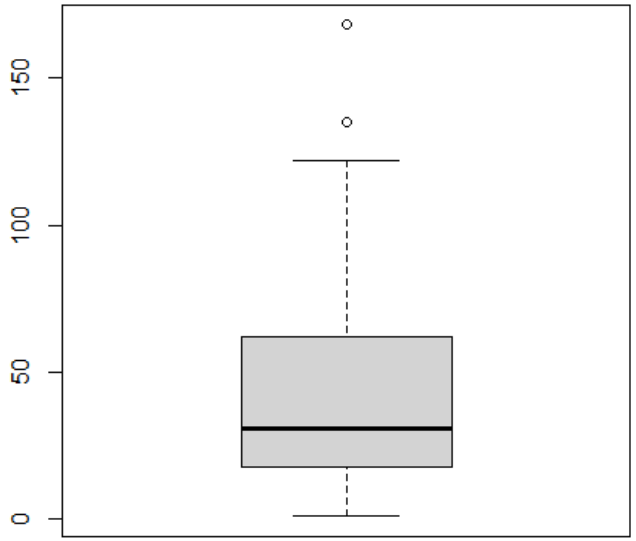
- Represents the frequencies of values of a variable bucketed into ranges
- Similar to a bar chart except that it groups values into continuous ranges
- hist(dat$Solar.R)



Box plot

- Displays the descriptive statistics graphically in the form of quartiles

```
boxplot(dat$Ozone)
```

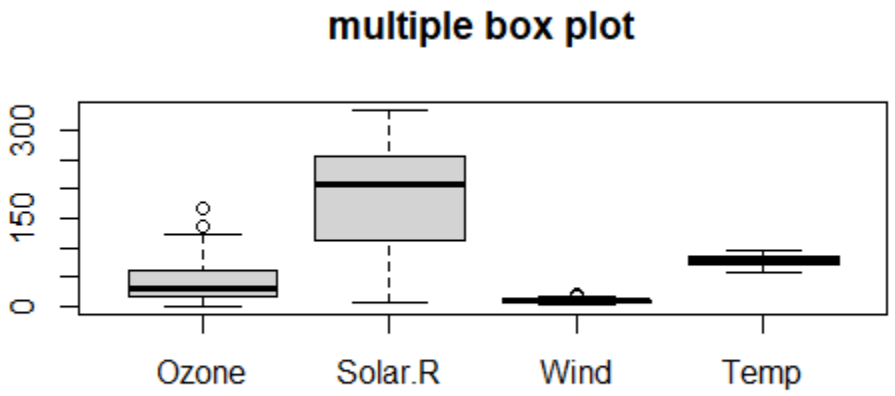



```
summary(dat$Ozone)
Min. 1st Qu. Median Mean 3rd Qu. Max.
1.0   18.0    31.0  42.1  62.0   168.0
```

Box plot (Contd.)

- Multiple box plot

```
boxplot(dat[,1:4],main='multiple box plot')
```

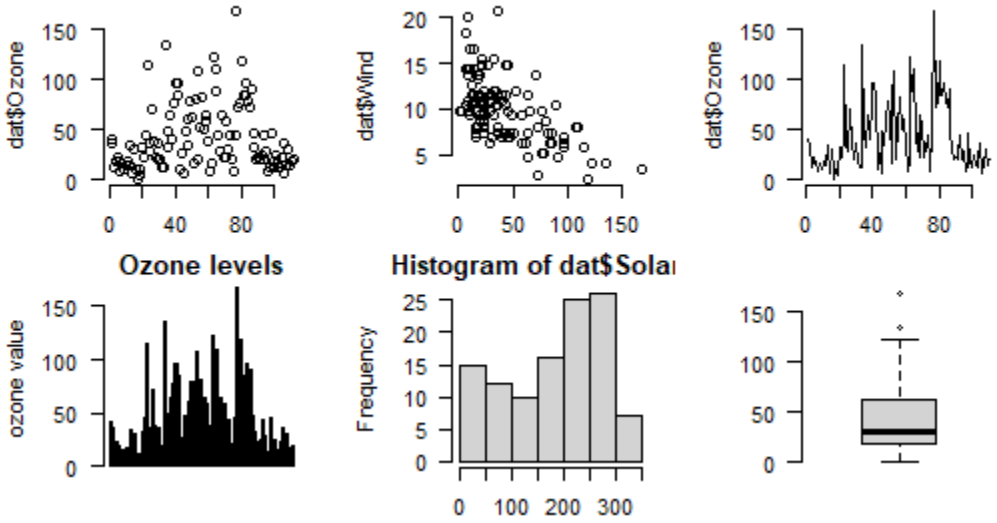


Grid of charts

- Enables plotting multiple charts at once
- For drawing a grid, the first argument should specify certain attributes like
  - the margin of the grid(mar)
  - no of rows and columns(mfrow)
  - whether a border is to be included(bty) and
  - position of the labels(las: 1 for horizontal, las: 0 for vertical).

Grid of charts (contd.)

```
par(mfrow=c(2,3),mar=c(2,5,2,1),las=1, bty='n')
plot(dat$Ozone)
plot(dat$Ozone,dat$Wind)
plot(dat$Ozone,type ='l')
barplot(dat$Ozone, main = 'Ozone levels', ylab = 'ozone value')
hist(dat$Solar.R)
boxplot(dat$Ozone)
```



Lattice Graphs

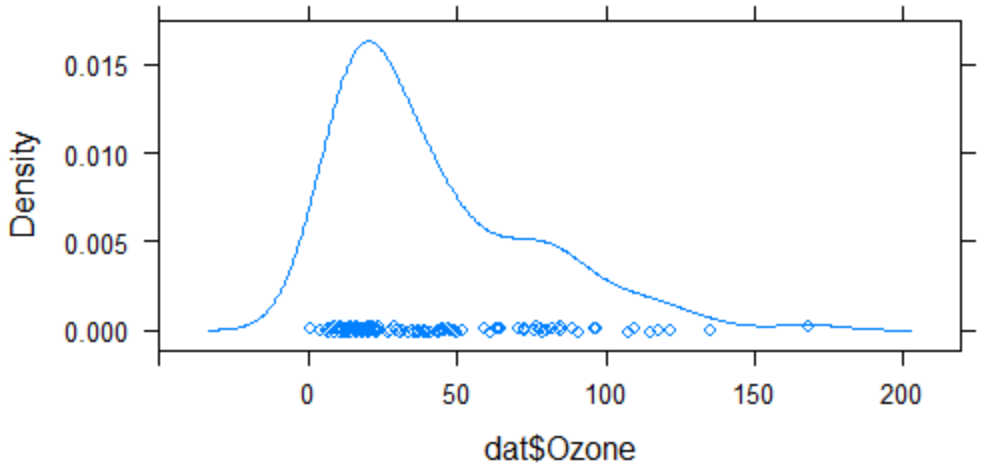
- Lattice package is used to visualize multivariate data.
- Lattice enables the use of ***trellis graphs***.
- Trellis graphs exhibit the relationship between variables which are dependent on one or more variables.

```
library(lattice)
```

Lattice Graphs (contd.)

- Kernel density plot

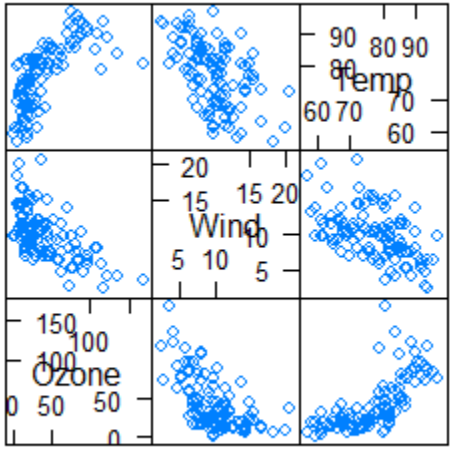
densityplot(dat$Ozone)



Lattice Graphs (contd.)

- scatter plot matrix

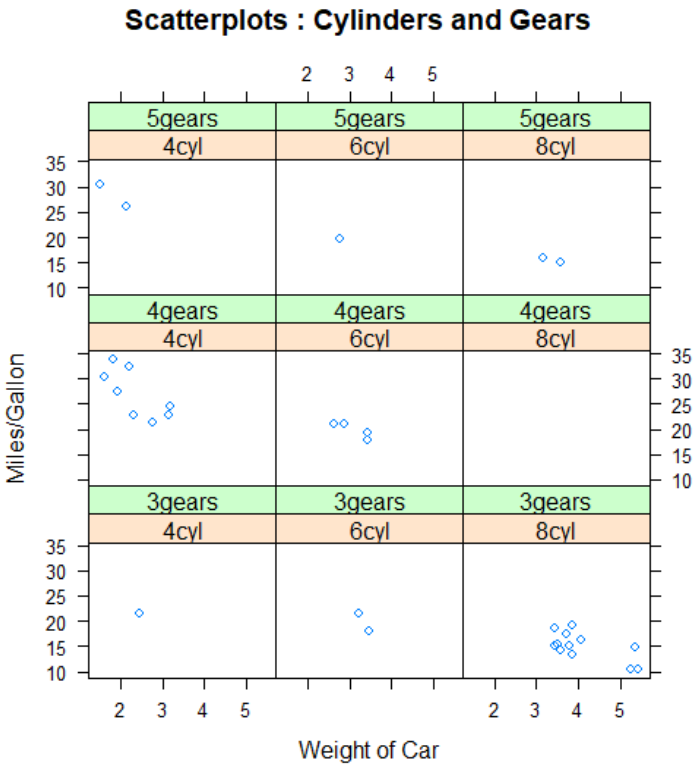
splom(dat[c(1,3,4)])



Lattice Graphs (contd.)

- scatter plot depicting the combination of 2 factors

```
xyplot(mpg~wt|cyl_factor*gear_factor,
      main="Scatterplots : Cylinders and Gears",
      ylab="Miles/Gallon", xlab="Weight of Car")
```



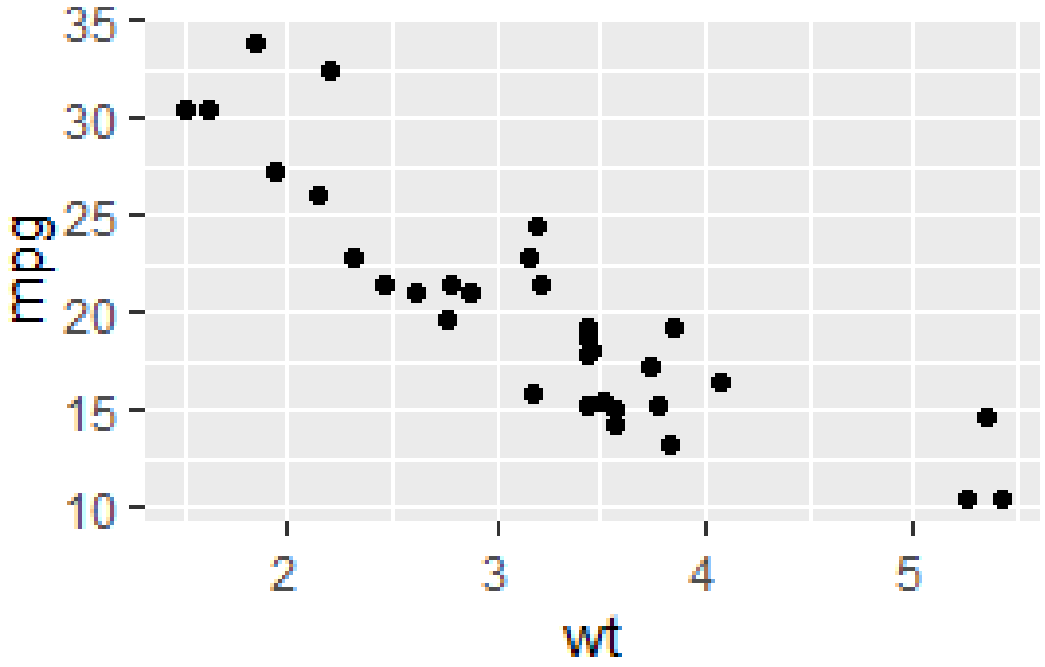

```
#preprocessing
unique(gear)
gear_factor<-factor(gear,levels=c(3,4,5),
            labels=c("3gears","4gears","5gears"))

unique(cyl)
cyl_factor<-factor(cyl,levels=c(4,6,8),
            labels=c("4cyl","6cyl","8cyl"))
```

ggplot()

- Stands for **grammar of graphics**
- Introduced by **Hadley Wickham, Winston Chang** in the year 2007.
- Used for creating elegant and more sophisticated visualization with little code
- Builds graph in layers
  - build a a complex graph by starting with a simple graph and adding additional elements, one at a time

ggplot() –scatter plot

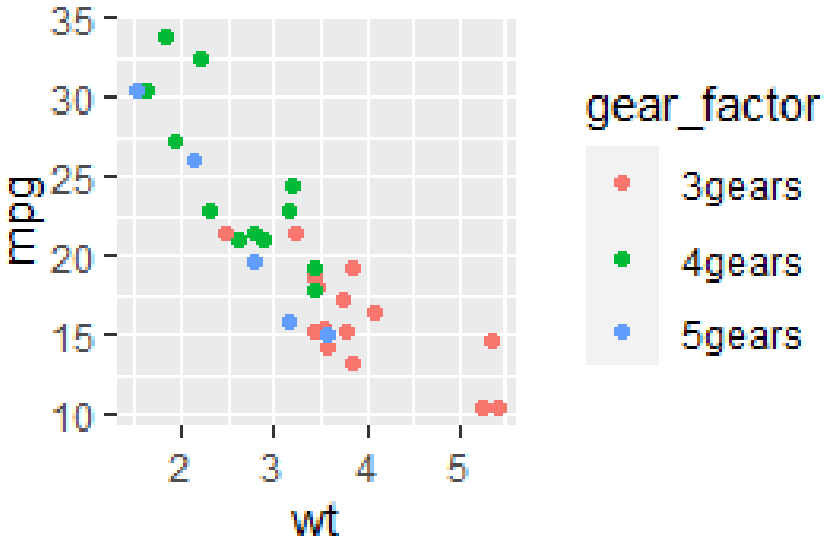
```
library(ggplot2)
ggplot(data = mtcars,
       mapping = aes(x=wt,y=mpg))+geom_point()
```



ggplot() –scatter plot (contd.)

- Grouping – Allows plotting multiple variables in a single graph
- Split the plot with factor variable using color paramerter

```
ggplot(mtcars,aes(x=wt,y=mpg,color=gear_factor))+geom_point()
```

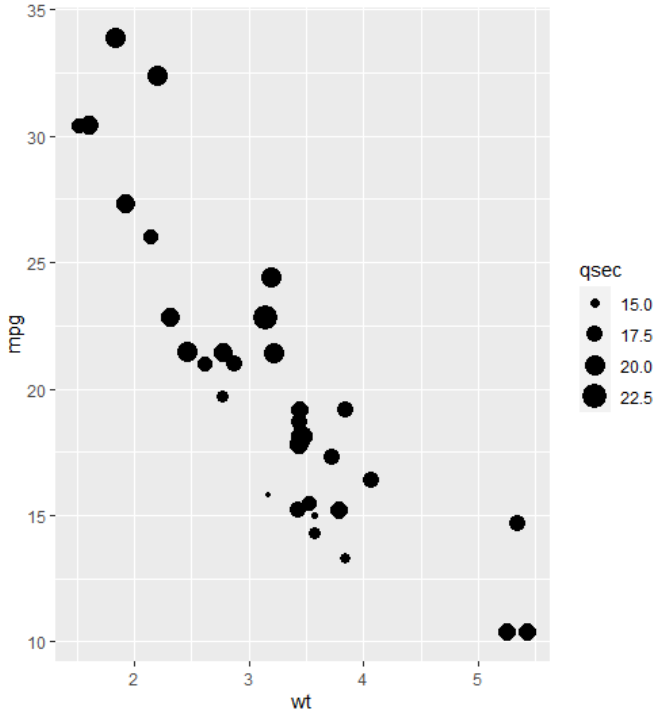


ggplot() –scatter plot (contd.)

- Split the plot with factor variable using size paramerter

```
ggplot(mtcars,aes(wt,mpg,size=qsec))+geom_point()
```

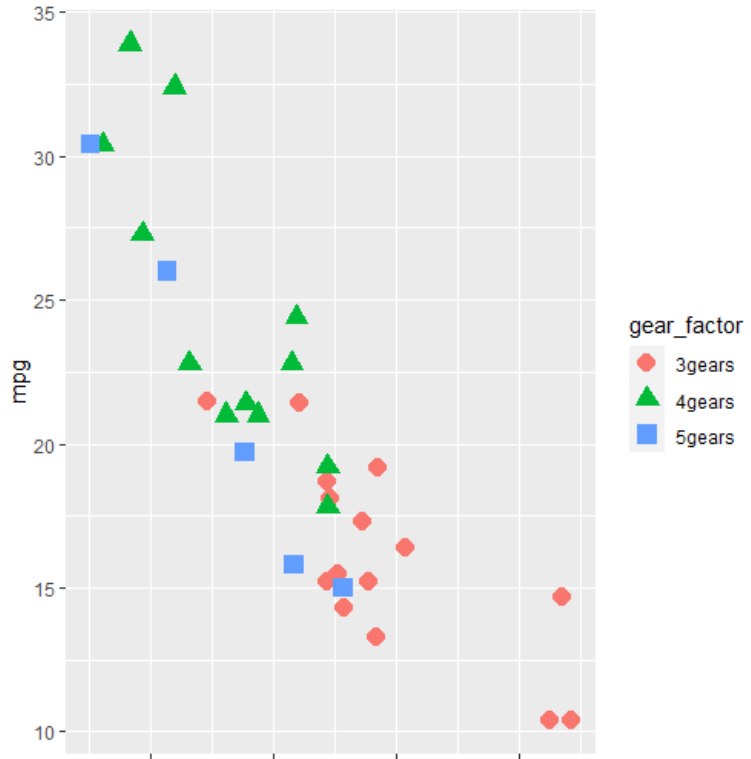
The value of qsec indicates the acceleration which decides the size of the points



ggplot() –scatter plot (contd.)

- Differentiating the data with both shape and color

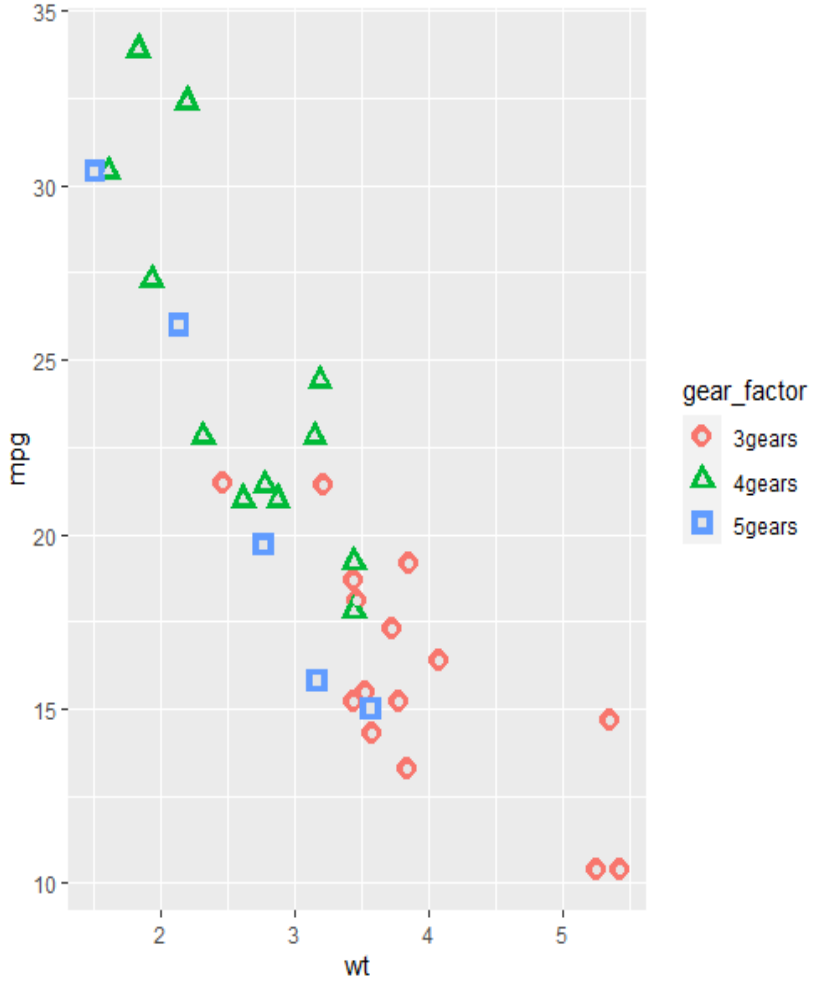
```
ggplot(mtcars,aes(wt,mpg,shape=gear_factor))+geom_point(aes(color=gear_factor),size=4)
```



ggplot() –scatter plot (contd.)

- Adding layers

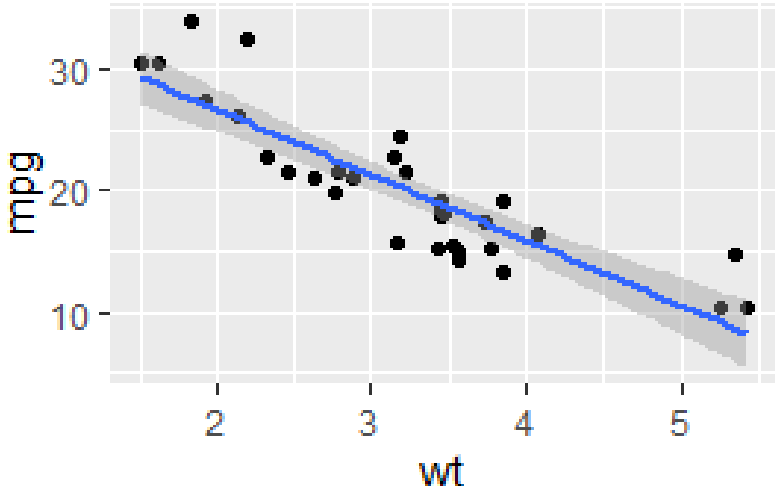
```
ggplot(mtcars,aes(wt,mpg,shape=gear_factor))
+geom_point(aes(color=gear_factor),size=4)
+geom_point(color='grey90',size=1.5)
```



ggplot() –scatter plot

- Adding best fit line

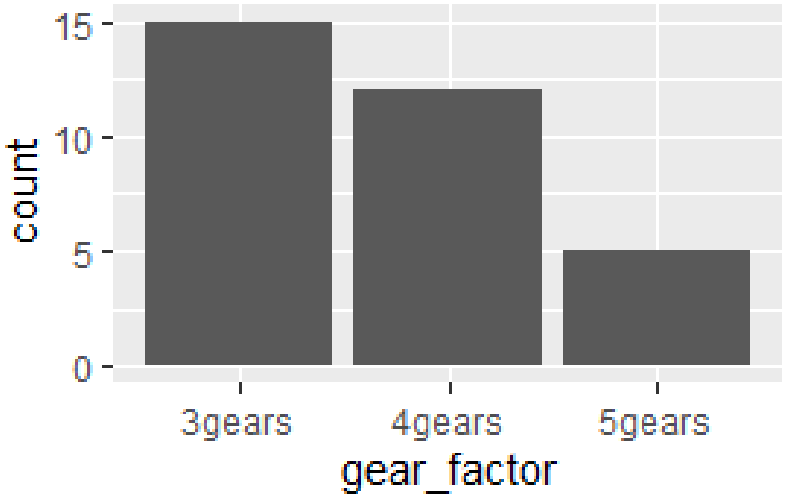
```
ggplot(data = mtcars,mapping =
aes(x=wt,y=mpg))+ geom_point()+
geom_smooth(method = 'lm')
```



ggplot –bar plot

- plotting the distribution of cylinder

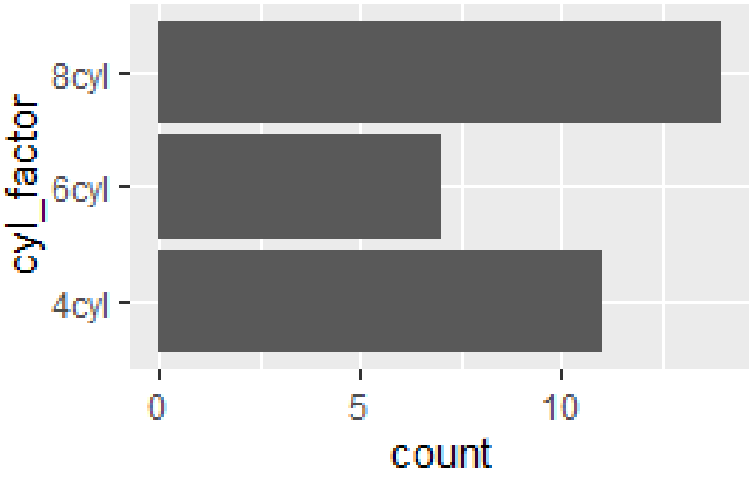
```
ggplot(mtcars,aes(x=gear_factor))+geom_bar()
```



ggplot –bar plot

- plotting the distribution of cylinder – flipping the bar direction

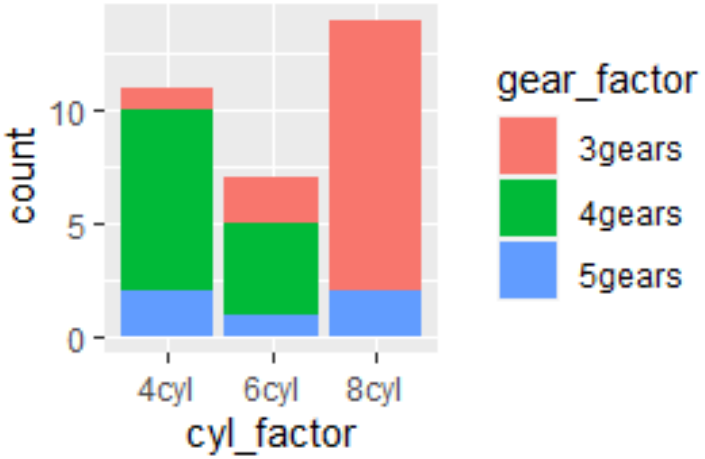
```
ggplot(mtcars,aes(x=gear_factor))+geom_bar()+
coord_flip()
```



ggplot –bar plot – 2 variables

- plotting the distribution of cylinder and gears as stacked bar

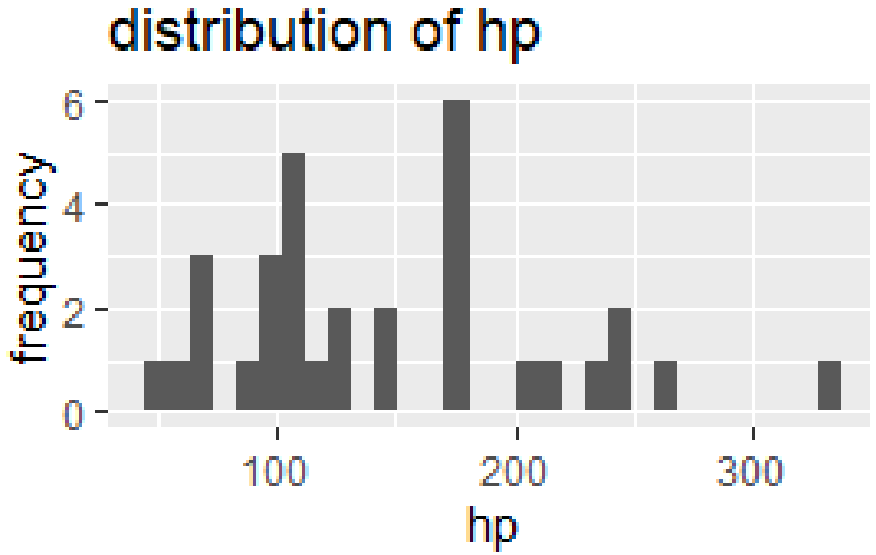
```
ggplot(mtcars,aes(x=cyl_factor,fill=gear_factor))
+geom_bar(position = "stack")
```



ggplot –histogram

- plotting the distribution of hp (horse power)

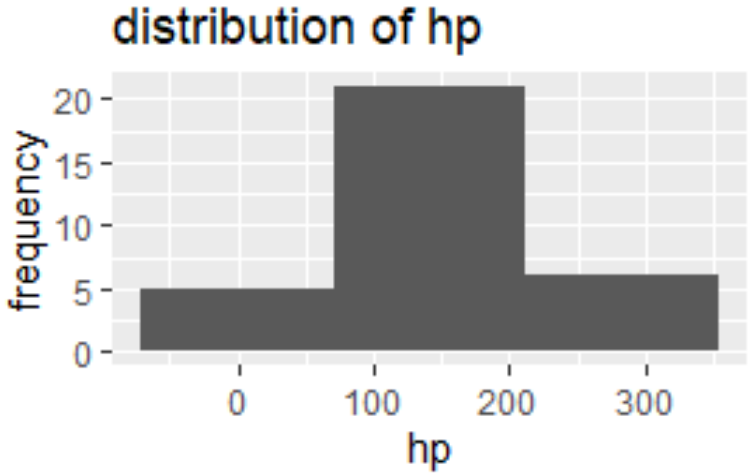
```
ggplot(mtcars,aes(x=hp))+geom_histogram()+la
bs(title='distribution of hp',y='frequency')
```



ggplot –histogram

- plotting the distribution of hp (horse power)- change the no.of bins

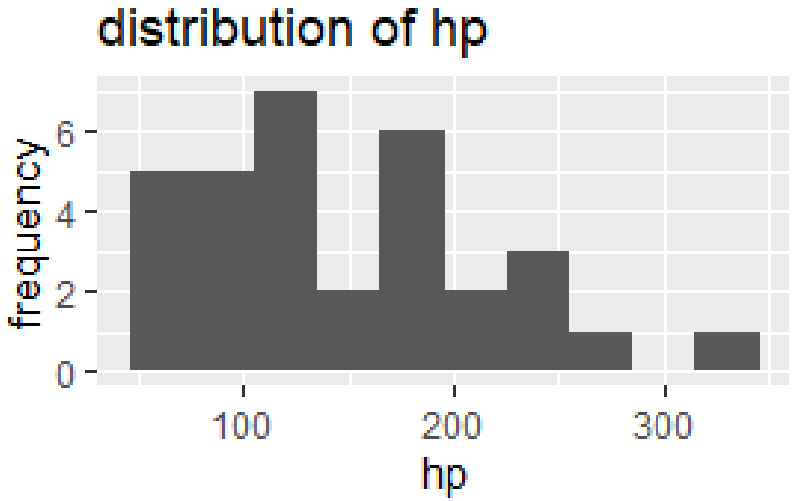
```
ggplot(mtcars,aes(x=hp))+geom_histogram(bins = 3)+labs(title='distribution of hp',y='frequency')
```



ggplot –histogram

- plotting the distribution of hp (horse power) – can change the binwidth

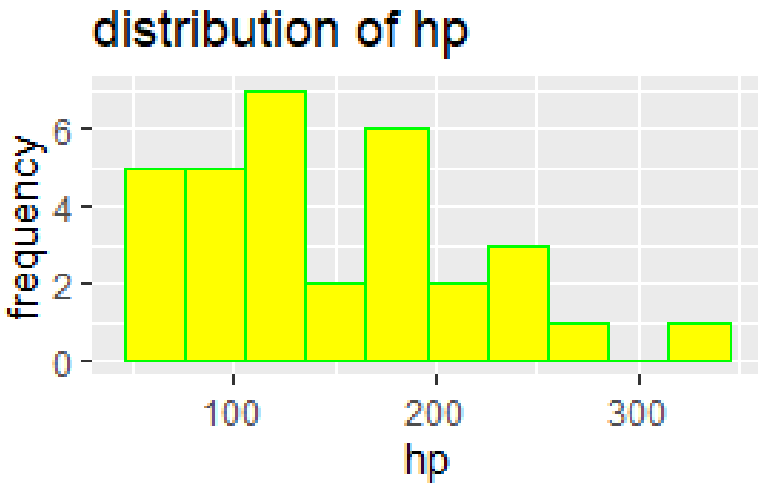
```
ggplot(mtcars,aes(x=hp))+geom_histogram(binwidth = 30)+labs(title='distribution of hp',y='frequency')
```



ggplot –histogram

- plotting the distribution of hp (horse power) – can specify border and fill color

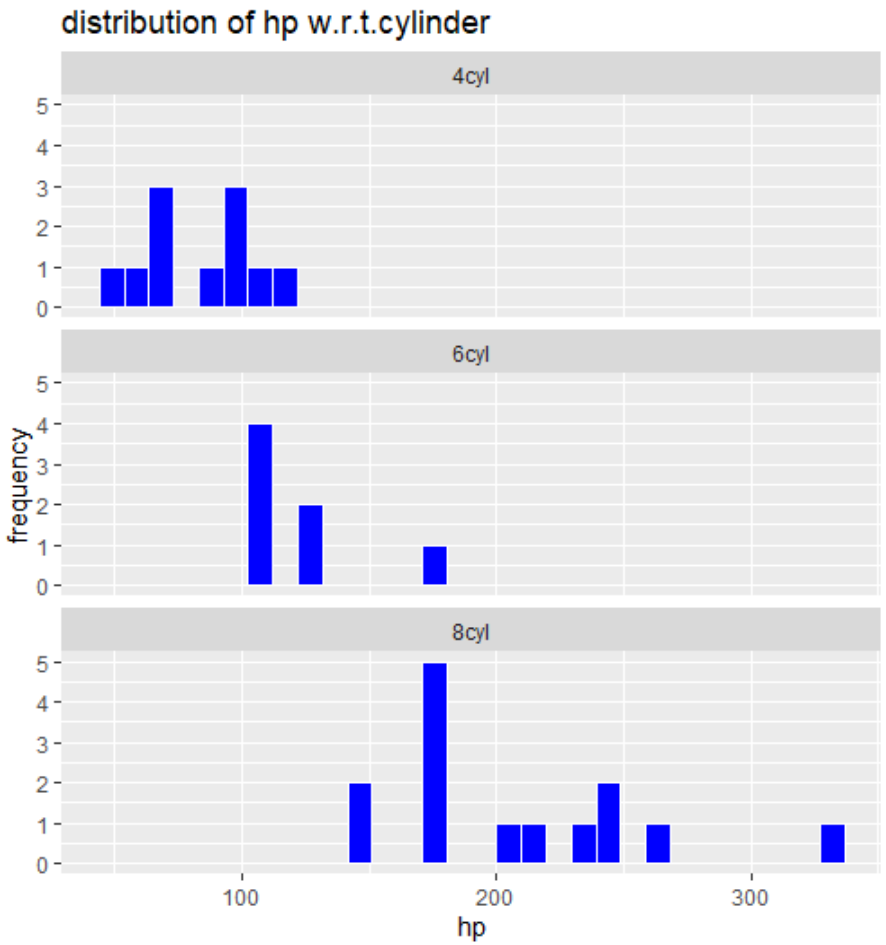
```
ggplot(mtcars,aes(x=hp))+geom_histogram(binwidth = 30,color='green',fill='yellow')+labs(title='distribution of hp',y='frequency')
```



ggplot –histogram

- Faceting – A graph consisting of several plots
- plotting the distribution of hp (horse power) based on cylinder values (cyl)

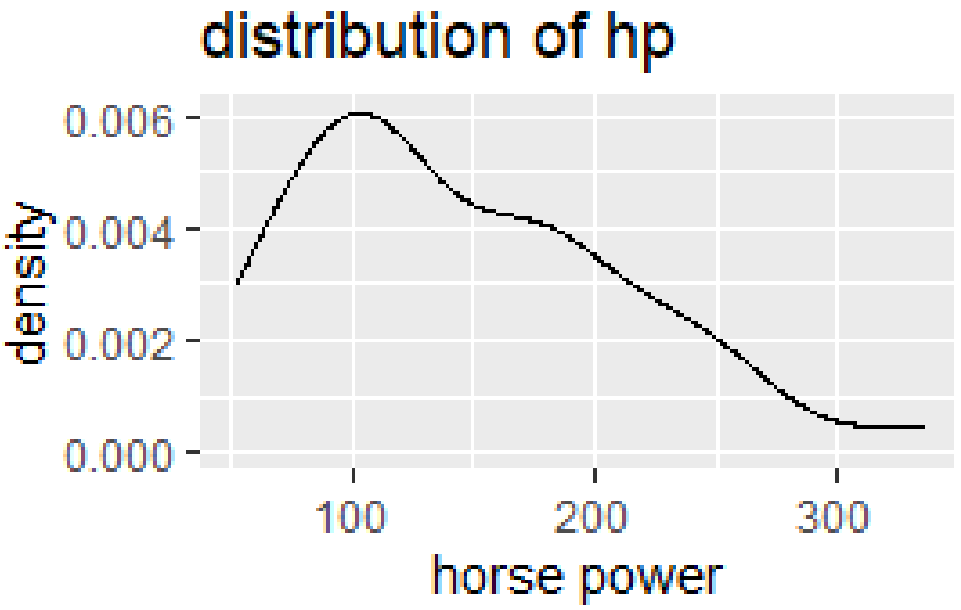
```
ggplot(mtcars,aes(x=hp))+
geom_histogram(fill='blue',color='white')+
facet_wrap(cyl_factor,ncol=1)+
labs(title='distribution of hp w.r.t.cylinder',y='frequency')
```



ggplot –Kernel density curve

- Plotting the kernel density curve of hp (horse power)

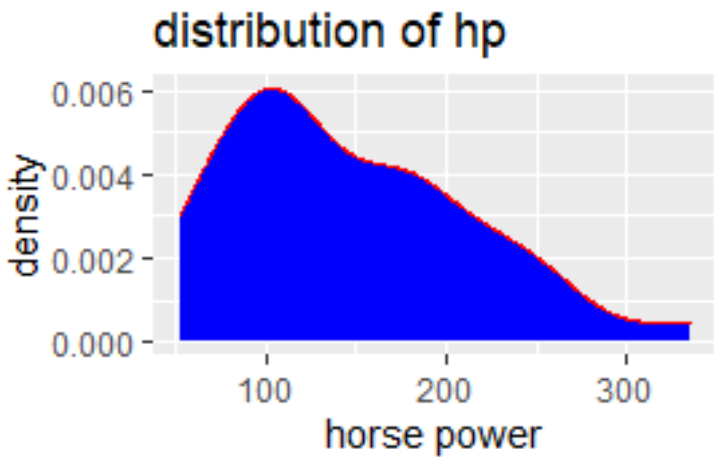
```
ggplot(mtcars,aes(x=hp))+geom_density()+
labs(x="horse power",y="density",title="distribution of hp")
```



ggplot –Kernel density curve

- Plotting the kernel density curve of hp (horse power) – fill with yellow color

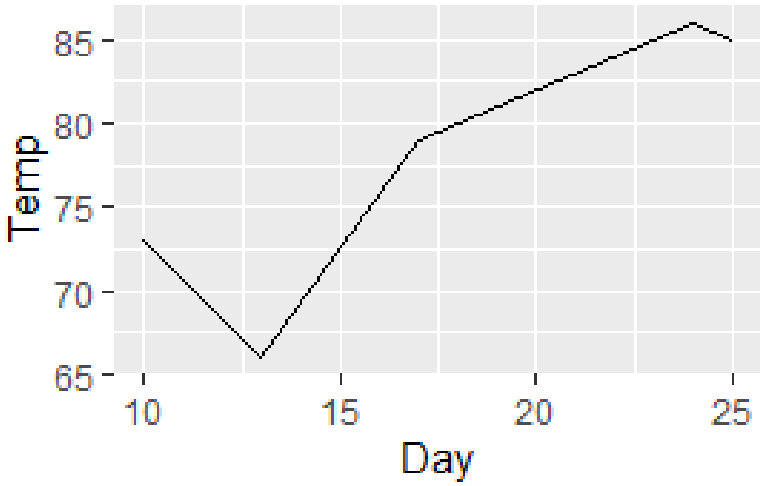
```
ggplot(mtcars,aes(x=hp))+geom_density(fill='blue',color='red')+
labs(x="horse power",y="density",title="distribution of hp")
```



ggplot – line plot

- Line plot of Days vs. Temp in airquality dataset which is read in dat dataframe

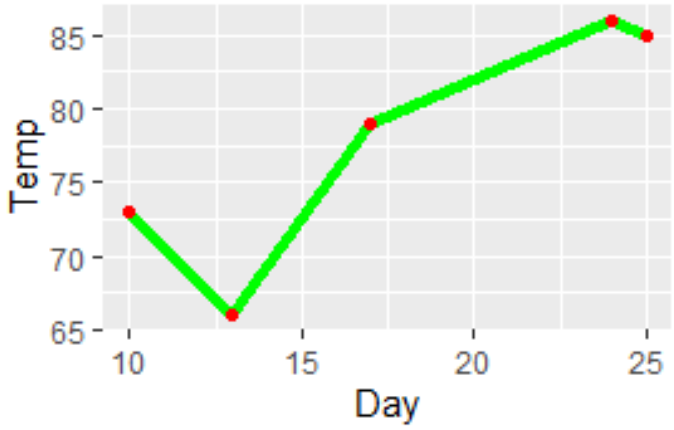
```
dat1 <- dat[sample(nrow(dat),5),] # sampling random 5 rows
ggplot(dat1,aes(x=Day,y=Temp))+geom_line()
```



ggplot – line plot

- Line plot of Days vs. Temp in airquality dataset which is read in dat dataframe
  - With varied thickness and color with points

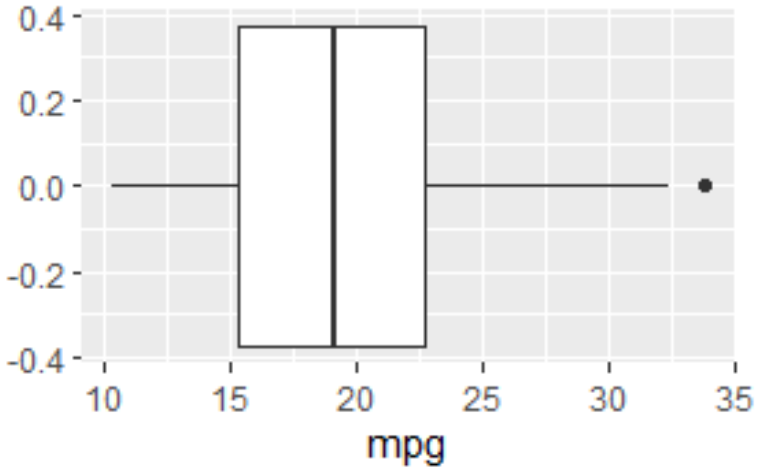
```
ggplot(dat1,aes(x=Day,y=Temp))+geom_line(size=1.5,color='green')+geom_point(size=1.5,color='red')
```



ggplot – box plot

- Box plot showing the summary statistics of miles per gallon (mpg) variable

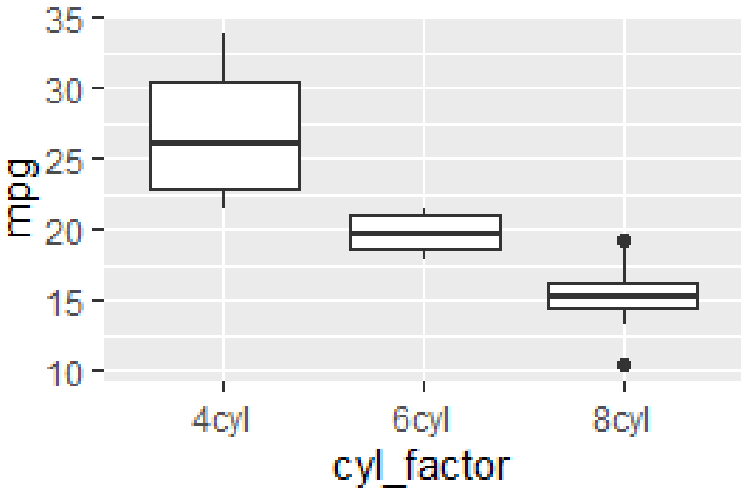
```
ggplot(mtcars,aes(x=mpg))+geom_boxplot()
```



ggplot – box plot

- Box plot showing the summary statistics of miles per gallon (mpg) variable for varied cylinder (cyl) values

```
ggplot(mtcars,aes(x=cyl_factor,y=mpg))+geom_boxplot()
```



Reference

- https://towardsdatascience.com/a-guide-to-data-visualisation-in-r-for-beginners-ef6d41a34174#c517
- https://rkabacoff.github.io/datavis/Univariate.html#categorical
- https://www.analyticsvidhya.com/blog/2015/07/guide-data-visualization-r/
- https://intellipaat.com/blog/tutorial/r-programming/data-visualization-in-r/