LECTURE 2 Linear Modelling

Objectives:

- To decipher graphical displays used for quantitative and categorical variables
- Graph relationships between variables using scatterplot.
- To visualize how data is right skewed or left skewed or symmetrical.

Graphical Summaries Displays: Is a methodology used in the initial phase of data exploration and statistical modelling. Graphical summaries enables us to visually ascertain the mean, median, mode (measures of central tendencies), standard deviation (spread/dispersion/deviation) and the shape (skewed or bell shaped) of the distribution.

Univariate Graphical Displays

Histogram

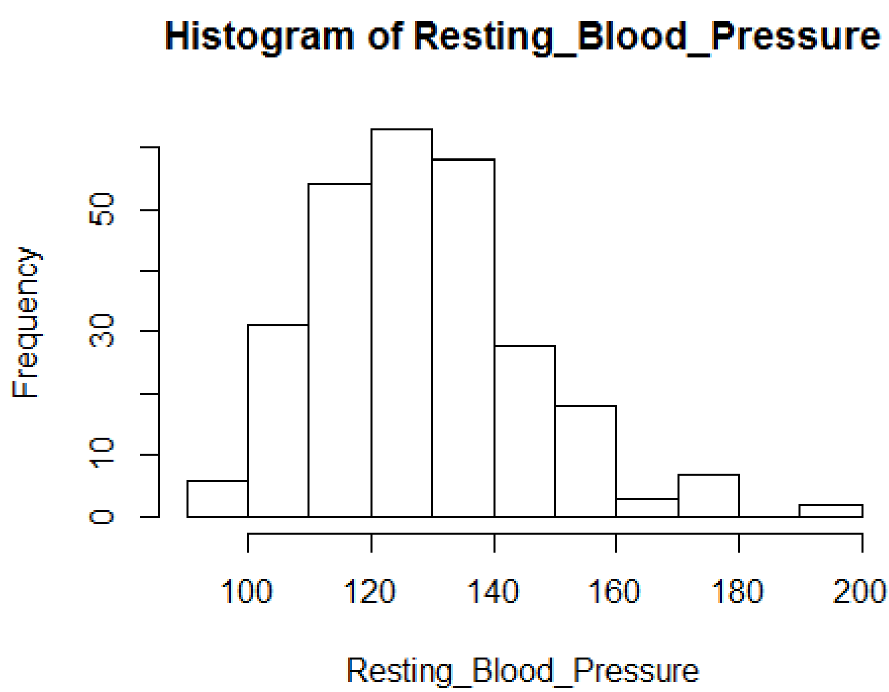
- Consists of parallel vertical bars that depict the frequency distribution of the quantitative data set. The height of the bar is the frequency of that data interval. Each bin has the same width This kind of histogram is defined as Frequency Histogram.
- If each bin height corresponds to the area within that range then the histogram is called the density histogram.
- The choice of number of bins is important. Very less number of bins eliminates the important details of the data that need to be perceived wheras too many bin shows extraneous details and does not display the general pattern of the data.
- The number of bins can be ascertained by the formula given by Freedman and Diaconis:
  $$\frac{n^{1/3}(\max - \min)}{2(Q_3 - Q_1)}$$
- Stem and leaf graphical representation is used for a small data set.

Using R for generating Frequency Histograms:

```
HeartAttack<-read_excel("HeartAttack.xlsx")
head(HeartAttack)

##   Age Sex Chest_Pain_Type Resting_Blood_Pressure Serum_Cholesterol
## 1  70   1               4                    130               322
## 2  67   0               3                    115               564
## 3  57   1               2                    124               261
## 4  64   1               4                    128               263
## 5  74   0               2                    120               269
## 6  65   1               4                    120               177
##   Fasting_Blood_MoreThan_120 Resting_Electrocardiographic_Reading
## 1                          0                                    2
## 2                          0                                    2
## 3                          0                                    0
## 4                          0                                    0
## 5                          0                                    2
## 6                          0                                    0
##   Maximum_Heart_Rate  Exercise_Induced_Angina Old_Peak Slope
```

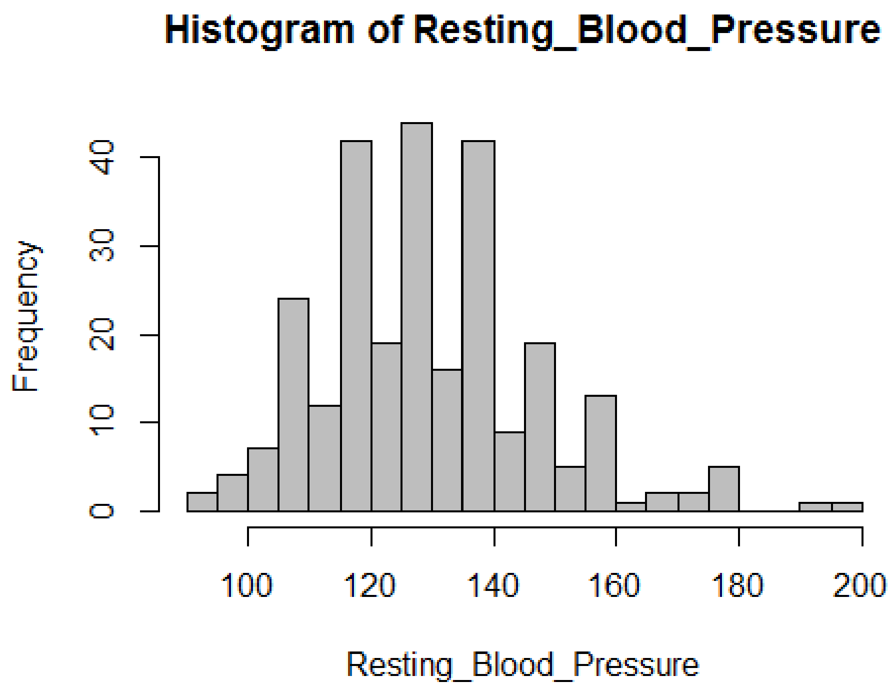
```
## 1                  109                          0       2.4      2
## 2                  160                          0       1.6      2
## 3                  141                          0       0.3      1
## 4                  105                          1       0.2      2
## 5                  121                          1       0.2      1
## 6                  140                          0       0.4      1
##   Number_Blood_Vessels_Calcified thal Heart_Attack_Diagnosis Residual
## 1                              3    3                      2  52.9060
## 2                              0    7                      1 298.6518
## 3                              0    7                      2   8.1378
## 4                              1    7                      1   1.3976
## 5                              1    3                      1  -5.0884
## 6                              0    7                      1 -85.8510
```

```
with(HeartAttack,hist(Resting_Blood_Pressure))
```




```
with(HeartAttack,hist(Resting_Blood_Pressure,breaks="FD",col="gray"))
```

```
args(hist.default)
function (x, breaks = "Sturges", freq = NULL, probability = !freq,
    include.lowest = TRUE, right = TRUE, density = NULL, angle = 45,
    col = NULL, border = NULL, main = paste("Histogram of", xname),
    xlim = range(breaks), ylim = NULL, xlab = xname, ylab, axes = TRUE,
    plot = TRUE, labels = FALSE, nclass = NULL, warn.unused = TRUE,
    ...)
NULL
```



Using R for generating Density Histograms:

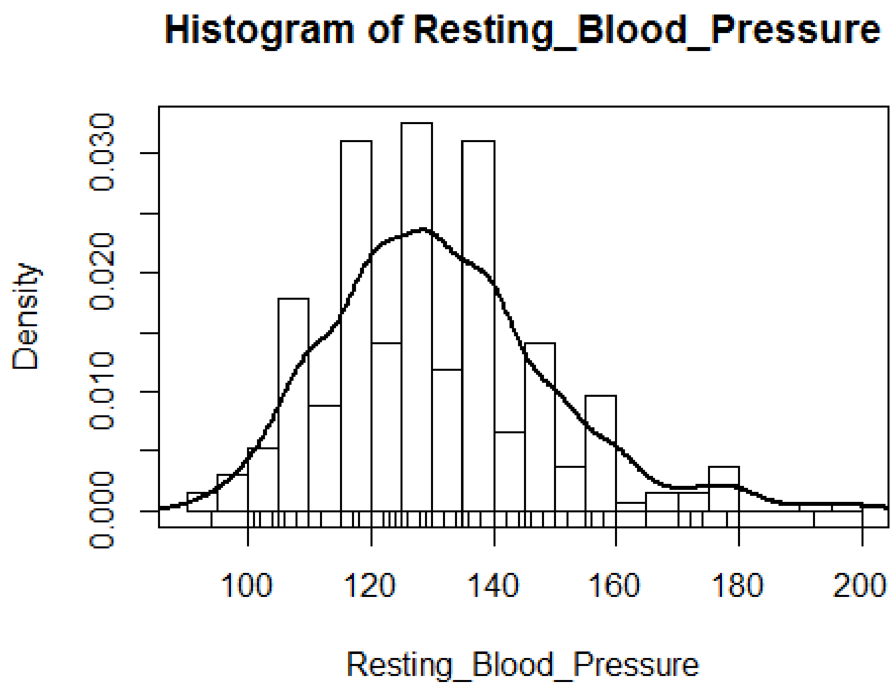
- Density estimation constructs the probability density distribution of a variable given a sample.
- Kernel Density estimate encloses an area of 1.
- Kernel estimation smooths the roughness of a naïve estimator

The width of the kernel estimation is obtained by trial and error in order to display the requisite detail but still eliminating unnecessary noise.

http://www.mvstat.net/tduong/research/seminars/seminar-2001-05/2

#ylab command defines the y axis. Freq=FALSE defines Density graph instead of Frequency. Lines #command draws the density. lwd=2 draws double thick line. box creates a one dimensional scatterplot.

```
with(HeartAttack, {
  hist(Resting_Blood_Pressure,breaks="FD",freq=FALSE,ylab="Density")
  lines(density(Resting_Blood_Pressure),lwd=2)
  rug(Resting_Blood_Pressure)
  box()
})
```

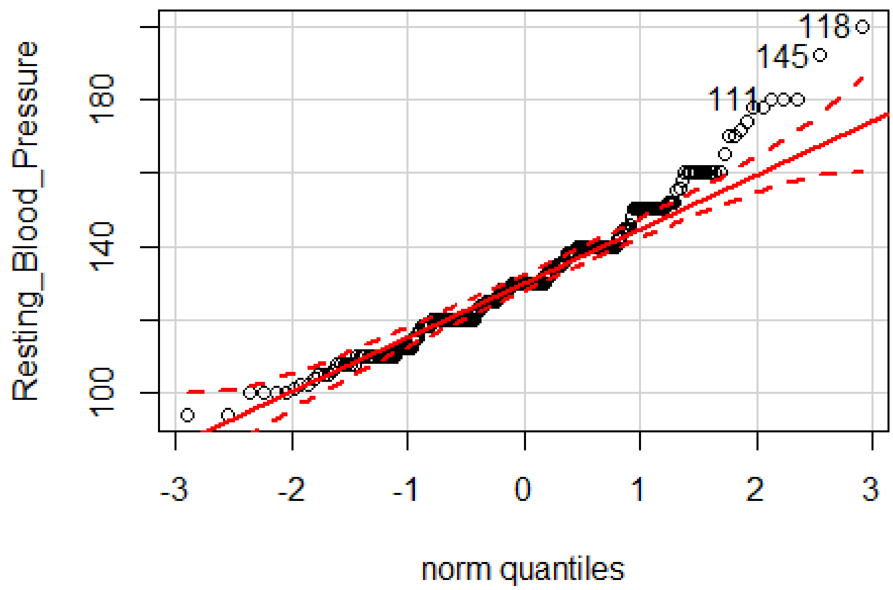


Quantile Comparison Plots

- QQ Plots are used to compare the empirical sample distribution with theoretical distributions like Normal Distribution
- Quantile function which is an inverse CDF is used to make these plots.

#The qqplot() function in the car package provides the 95% CI around the line fit to the plot. The #row name provides the Resting_Blood_Pressure being evaluated and the id.n=3 provides #marking the three most extreme points.

```
with(HeartAttack,qqPlot(Resting_Blood_Pressure,labels = row.names(HeartAttack
),id.n=3))
```




```
## 118 145 111
## 270 269 266
```

- R provides the qq plot for any know distributions like Chi Square, binomial, Poisson etc Where the root is the distribution like chisq the density and quantile function is dchisq and qchisq. To generate a random variable from that distribution rchisq function is used.

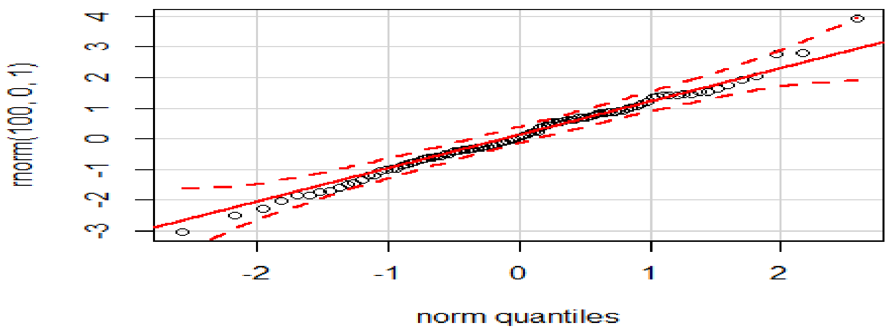
Demonstrating symmetrical, right skew, thick tails distribution.

1) A symmetrical distribution will more or less follow the comparison line and will be within the confidence interval bounds.

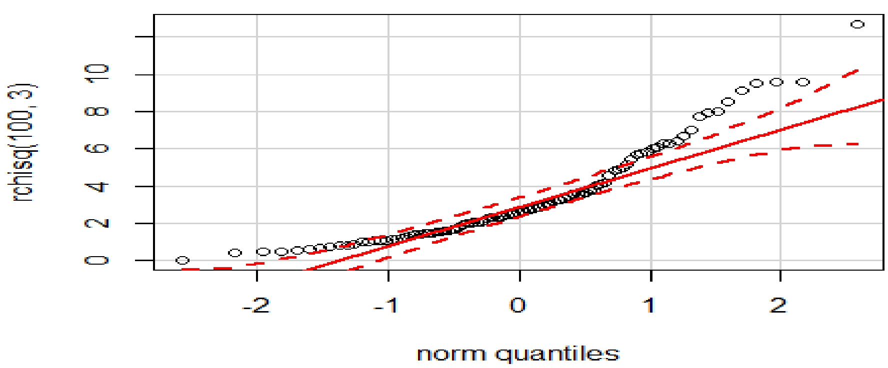
2) A right skewed distribution will have points that lie above the comparison line in both tails. For negative skewed distribution the points will lie below the comparison line in both tails .

3) For a heavy tailed distribution like the t distribution the upper tail lie above the normal quantiles and the lower tail lie below the normal quantiles.

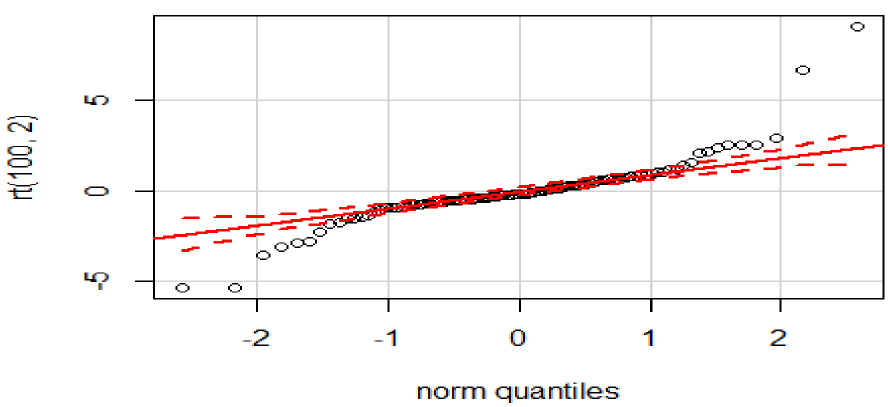
```
qqPlot(rnorm(100,0,1))
```




```
qqPlot(rchisq(100,3))
```



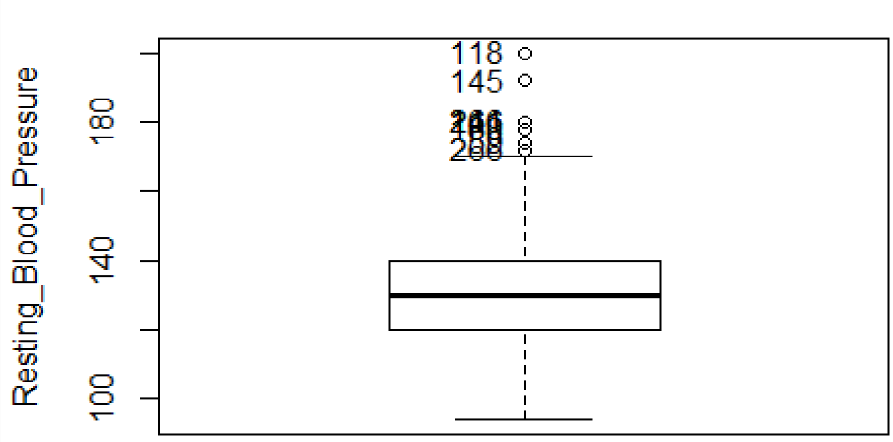

```
qqPlot(rt(100,2))
```



Boxplots

- The boxplots is a graphical representation for quantitative variables that provides the five number summary: Least, Q1, Q2 (Median), Q3, Maximum.
- Boxplot unlike histogram does not display all the observations so it does not preserve the data. Boxplot is used for focusing on main characteristics of the distribution, comparison of distributions and for deciding the transformation to make the distribution symmetrical.
- To identify outliers ,they can be identified by the following rule:
  Outliers are any values that are to the right and left of the boundaries $(L-1.5*IQR, H+1.5*L)$ .These are inner fences. Beyond the inner fences the values are outlies.
- To identify extreme values, they can be identified by the following rule:
  Extreme values are any values (outliers) that are to the right and left of the boundaries $(L-3*IQR, H+3*L)$.
- Bimodality is not identified by boxplots.
- If the explanatory variables are discrete then the parallel boxplots can be used. If the explanatory variables is qualitative/categorical then also parallel boxplots can be used for comparisons.

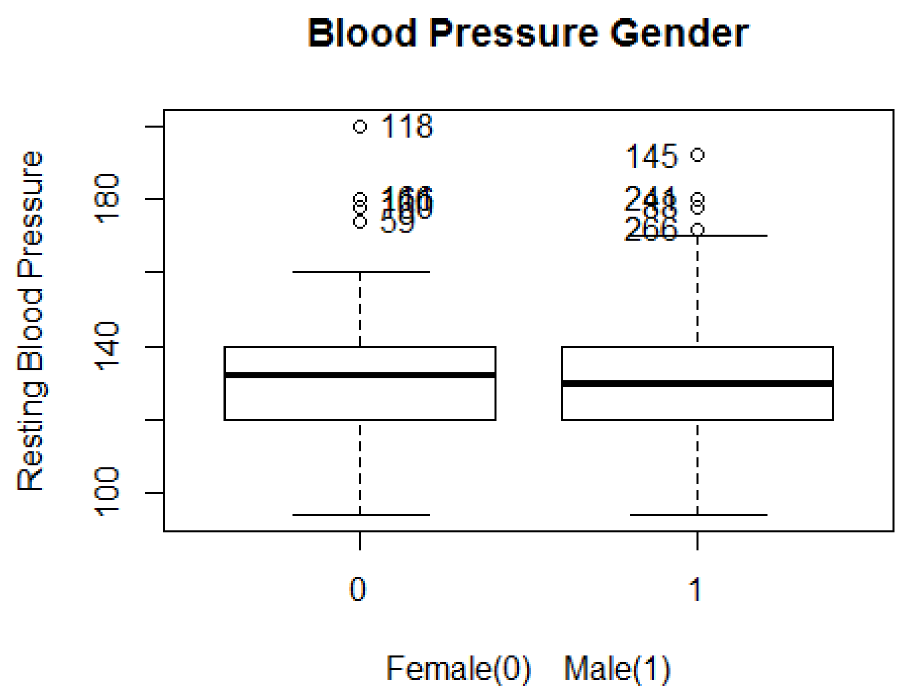
```
Boxplot(~  Resting_Blood_Pressure,data=HeartAttack)
```




```
## [1] "59"  "88"  "111" "118" "145" "160" "166" "241" "266
```

Parallel Boxplot

```r
Boxplot(Resting_Blood_Pressure~Sex,data=HeartAttack, main="Blood Pressure Gender", xlab="Female(0)     Male(1)", ylab="Resting Blood Pressure")
```

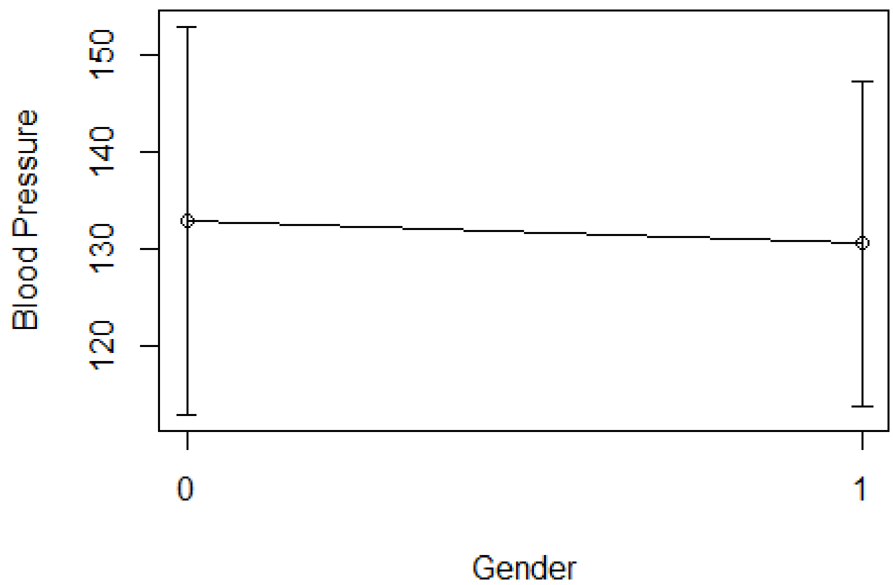


The median and IQR Resting Blood Pressure for Female is marginally higher than Men. The range of Resting Blood Pressure is higher for men as compared to female.

Boxplots of Means

This plot is used for scientific literature. This depicts the boxplot representation of means of the groups. It graphs the means and the error bars with $\pm 1SD$ around the mean.

```r
#Boxplots of Means
mean_values<-with(HeartAttack,tapply(Resting_Blood_Pressure,Sex,mean))
standard_deviation<-with(HeartAttack,tapply(Resting_Blood_Pressure,Sex,sd))
# tapply() computes means ans sds for each gender
plotCI(1:2,mean_values,standard_deviation,xaxt="n",xlab="Gender",ylab="Blood Pressure")
# plotCI() function helps draw the graph.The parameters are coordinates on horizontal axis ,means
# on vertical axis,standard deviations,xaxt="n" suppresses the x ticks.
lines(1:2,mean_values)
#lines joins the means with the lines
axis(1,at=1:2,labels=names(mean_values))
```




```
# axis specifies the groups on the x axis.
```

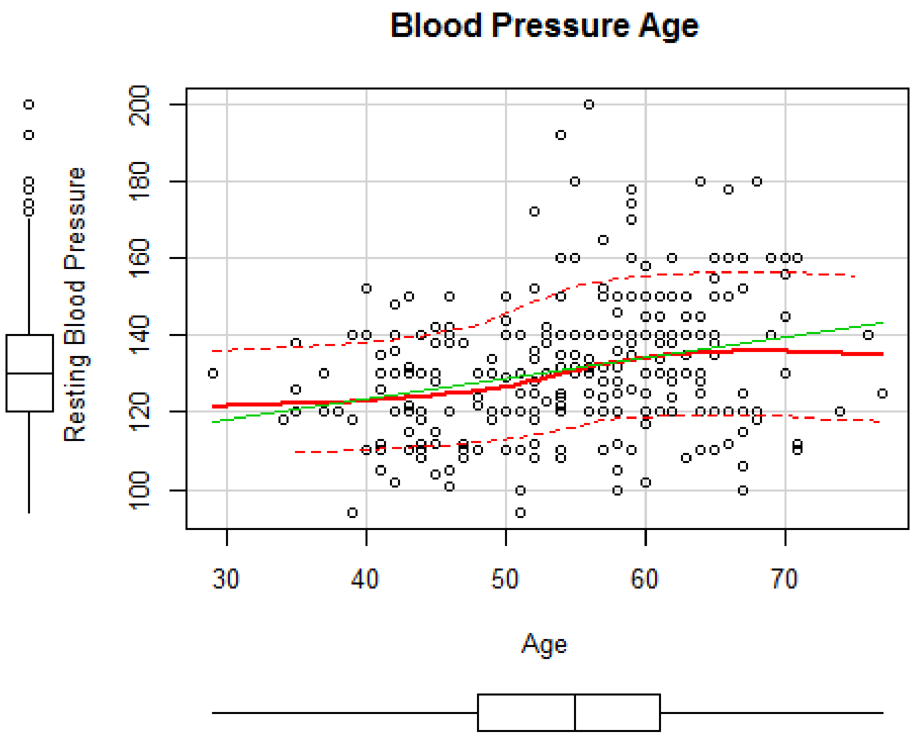
Bivariate Graphical Distributions

Scatter plot Geometrical distribution of two quantitative variables.

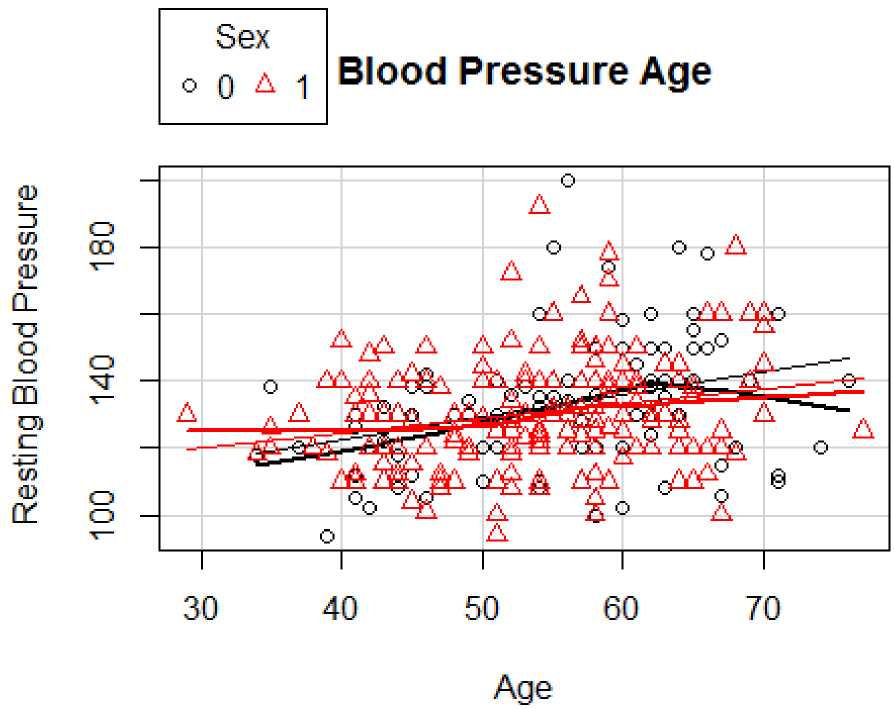
- The scatterplots are affected by outliers ie skewed data.
- The scatterplots are sometimes not easy to examine especially when variables are discrete. To solve this issue and remove the overlap a random quantity is added to each of the variables. This random quantity could be a quantity between [-.5,+.5] to each of the
- Non parametric regression curved line can also be plotted as a smoother. The lowess (locally weighted scatterplot smoother) uses the local linear regression to fit the data with the maximum weights assigned to the values that are closed to the focal x values. The weightage decreases as the x value gets farther away from the focal x value.
- Degree of smoothness is controlled by the span parameter which is determined by the fraction of data included in each local regression fit.
- Larger spans create smoother lines.
- The objective is to obtain a balance between variance and bias. Smaller span captures more detail and will showcase more variance as opposed to bias. Therefore a larger span is better for a smoother curve.
- Variance Bias tradeoff: http://rstudio-pubs-static.s3.amazonaws.com/1690_b6906d2174654e339c33a07d020e6cc3.html

Scatterplots

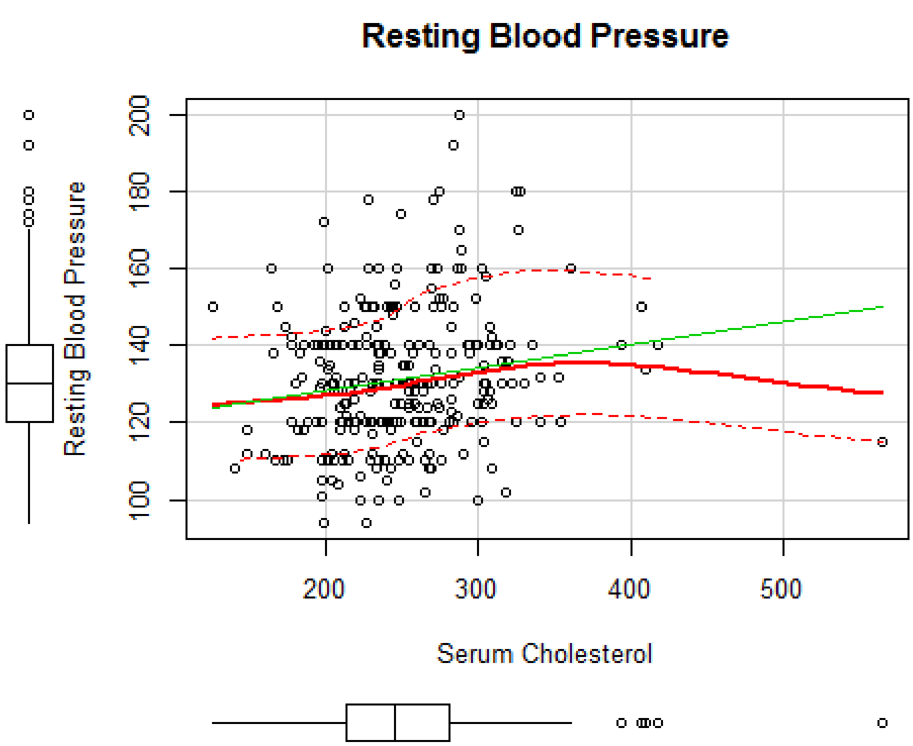
```
scatterplot(Resting_Blood_Pressure~Age,data=HeartAttack, main="Blood Pressure
Age", xlab="Age", ylab=" Resting Blood Pressure")
```



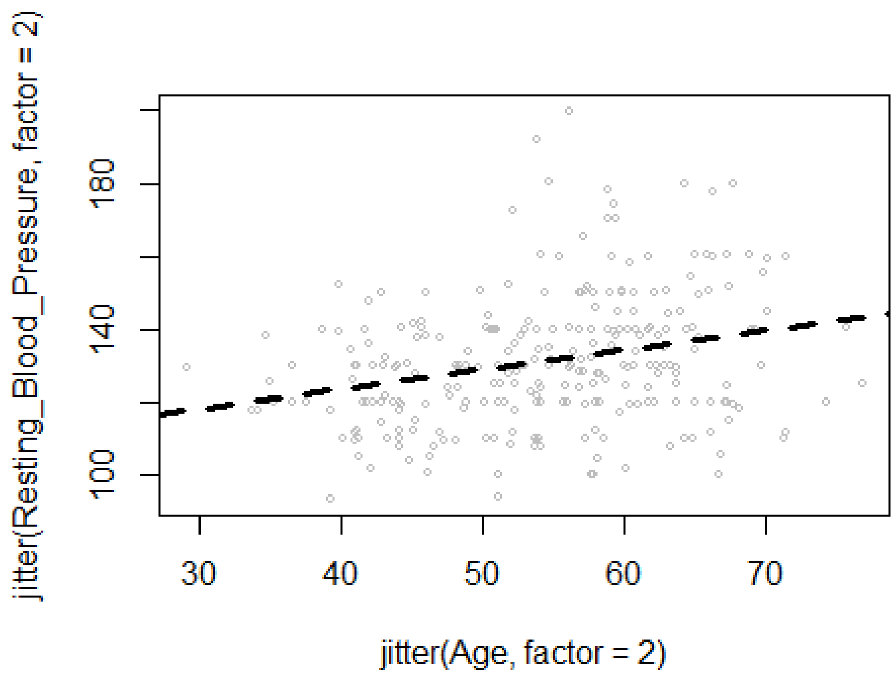

```
# scatterplot is a function in the car package that provides two smoothers on
the scatterplot.
# The first smoother is the Ordinary Least Squares(OLS) which can be suppress
ed by reg.line=FALSE
# The second is a solid curved line which is the non parameteric regression s
moother produced by # # the lowess smmoother(locally weighted scatterplot smo
other) and this can be supressed by lowess # function in R.The variance is se
en by the dotted line around the lowess curve.The conditional # # distributio
n of Resting Blood Pressure with Age can be evaluated by the vertical line be
tween
# dotted variance curve line.This also gives us a boxplot for both variables.

# Coded Scatter plot
scatterplot(Resting_Blood_Pressure~Age|Sex,data=HeartAttack, main="Blood Pres
sure Age", xlab="Age", ylab=" Resting Blood Pressure")
```




```
scatterplot(Resting_Blood_Pressure~Serum_Cholesterol,data=HeartAttack, main="
Resting Blood Pressure", xlab="Serum Cholesterol", ylab="Resting Blood Pressu
re")
```



```
#Jittered Scatterplot
plot(jitter(Resting_Blood_Pressure,factor=2)~jitter(Age,factor=2),col="gray",
cex=.5,data=HeartAttack)
with(HeartAttack, {
  abline(lm(Resting_Blood_Pressure~Age),lwd=3,lty="dashed")
  lines(lowess(Resting_Blood_Pressure,Age,f=0.2),lwd=3)
  # least square line is created by abline using the lm function
})
```




```
# Three dimensional scatterplot
scatter3d(Resting_Blood_Pressure~Age+Serum_Cholesterol,id.n=3,data=HeartAttac
k)
# The 3d graph shows Resting blood pressure on y axis and Age and Serum
Chaolesterol on the
# x axis and z axis.The three furthest points were identified by the
Mahalanobis distances
# (Point of means)
```

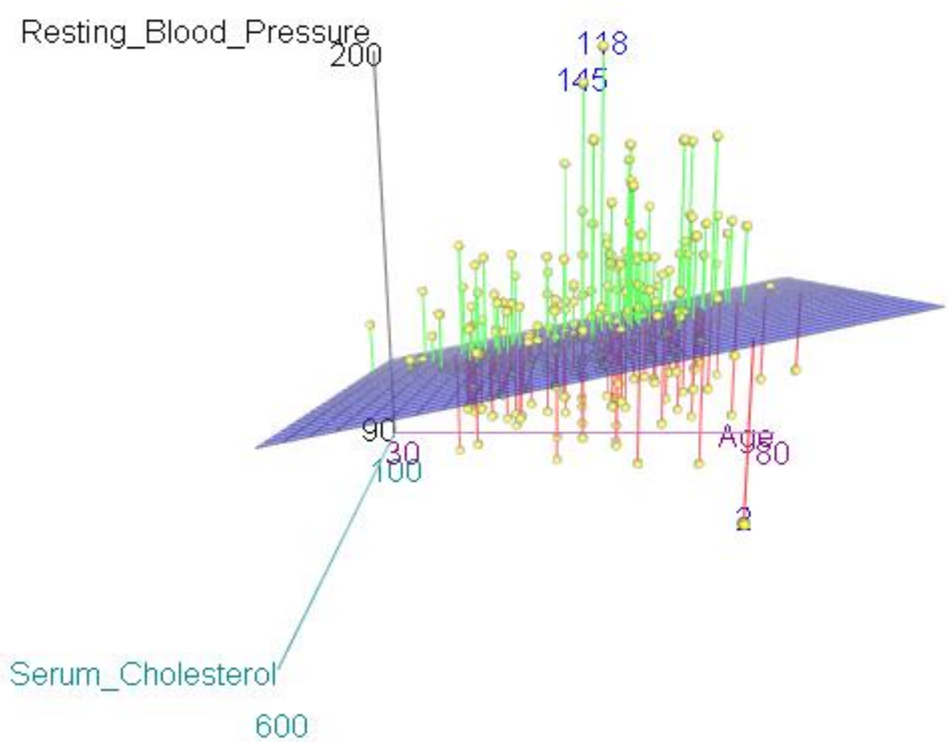
Multivariate Data Distributions

Three Dimensional Plots

- For data in which there is one numeric response variable but multiple predictor variables (quantitative or categorical) the graph can be represented in higher dimensions. Two predictor variable's relationship will be depicted on a three dimensional graph.

- There are various packages like lattice, wireframe, rggobi that have additional functionality for multivariate graphing capabilities. Non parametric regression can also be achieved by these techniques.

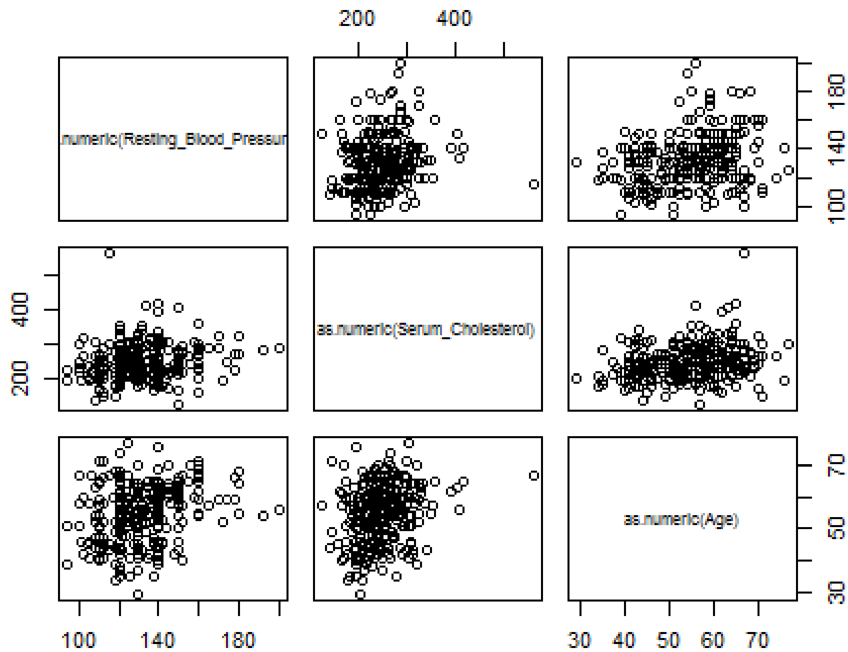
```
#Three dimensional scatterplot
scatter3d(Resting_Blood_Pressure~Age+Serum_Cholesterol,id.n=3,data=HeartAttack)
# The 3d graph shows Resting blood pressure on y axis and Age and Serum Chaolesterol on the
# x axis and z axis.The three furthest points were identified by the
#mahalanobis distances (Point of means)
```



Scatterplot Matrices

- Scatterplot matrices show case the correlation between a combination of variables taken a pair at a time.
- The pairs() method in the package car implements this functionality.

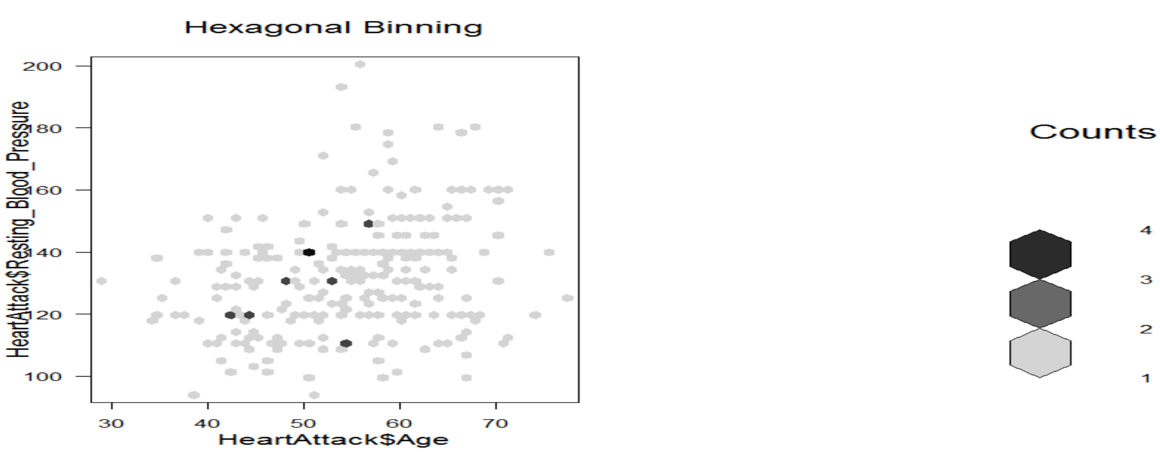
```r
# Scatterplot Matrices
pairs(~ as.numeric(Resting_Blood_Pressure) + as.numeric(Serum_Cholesterol) + as.numeric(Age) ,data=HeartAttack)
```



 # High Density Scatterplot with Binning

bin<-hexbin(HeartAttack$Resting_Blood_Pressure,HeartAttack$Age,xbins=50)

plot(bin,main="Hexagonal Binning")



```
boxplot(Serum_Cholesterol~Sex, data = HeartAttack, lwd = 2, ylab = 'Cholesterol')
stripchart(Serum_Cholesterol~Sex, vertical = TRUE, data = HeartAttack,
   method = "jitter", add = TRUE, pch = 20, col = 'blue')
```