Week 3 Data Visualization Assignment

Nitin

September 20, 2020

Read and Analyze Data

Leat us analisis diffrence between when we read file as csv or xlsx

```
library(readxl)
XlData <- read_excel("data.xlsx")
str(XlData)
```

```
## Classes 'tbl_df', 'tbl' and 'data.frame':    361 obs. of  25 variables:
##  $ Month                        : POSIXct, format: "1990-04-01" "1990-05-01" ...
##  $ Coarse wool Price            : num  482 447 441 418 418 ...
##  $ Coarse wool price % Change   : chr  "-" "-7.2700000000000001E-2" "-1.4E-2" "-5.11E-2" ...
##  $ Copra Price                  : num  236 234 216 205 198 196 198 236 237 233 ...
##  $ Copra price % Change         : chr  "-" "-8.5000000000000006E-3" "-7.6899999999999996E-2" "-5.0900
##  $ Cotton Price                 : num  1.83 1.89 1.99 2.01 1.79 1.79 1.79 1.82 1.85 1.85 ...
##  $ Cotton price % Change        : chr  "-" "3.2800000000000003E-2" "5.2900000000000003E-2" "1.01E-2"
##  $ Fine wool Price              : num  1072 1057 898 896 951 ...
##  $ Fine wool price % Change     : chr  "-" "-1.35E-2" "-0.15029999999999999" "-2.7000000000000001E-3"
##  $ Hard log Price               : num  161 173 182 188 186 ...
##  $ Hard log price % Change      : chr  "-" "7.2300000000000003E-2" "5.0999999999999997E-2" "3.4599999
##  $ Hard sawnwood Price          : num  550 492 495 486 488 ...
##  $ Hard sawnwood price % Change: chr  "-" "-0.1055" "7.1000000000000004E-3" "-1.9199999999999998E-2"
##  $ Hide Price                   : num  100 99.5 97.9 96.8 91.9 ...
##  $ Hide price % change          : chr  "-" "-5.4000000000000003E-3" "-1.5699999999999999E-2" "-1.17E-
##  $ Plywood Price                : num  312 350 374 378 365 ...
##  $ Plywood price % Change       : chr  "-" "0.12089999999999999" "6.8000000000000005E-2" "1.21E-2" ..
##  $ Rubber Price                 : num  0.84 0.85 0.85 0.86 0.88 0.9 0.9 0.9 0.88 0.87 ...
##  $ Rubber price % Change        : chr  "-" "1.1900000000000001E-2" "0" "1.18E-2" ...
##  $ Softlog Price                : num  121 124 129 124 130 ...
##  $ Softlog price % Change       : chr  "-" "0.03" "4.1599999999999998E-2" "-4.0300000000000002E-2" ..
##  $ Soft sawnwood Price          : num  219 213 200 210 208 ...
##  $ Soft sawnwood price % Change: chr  "-" "-2.63E-2" "-6.0999999999999999E-2" "5.0299999999999997E-2
##  $ Wood pulp Price              : num  829 843 831 799 819 ...
##  $ Wood pulp price % Change     : chr  "-" "1.5900000000000001E-2" "-1.32E-2" "-3.9100000000000003E-2
```

Create new data frame with columns of our Intrest.

```
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 3.5.3
```

```
## -- Attaching packages --------------------------------------------------------------------- t
```

```
## v ggplot2 3.2.0       v purrr   0.2.5
## v tibble  2.1.3       v dplyr   0.8.0.1
## v tidyr   0.8.1       v stringr 1.3.1
## v readr   1.1.1       v forcats 0.3.0
```

```
## Warning: package 'ggplot2' was built under R version 3.5.3
## Warning: package 'tibble' was built under R version 3.5.3

## Warning: package 'dplyr' was built under R version 3.5.3

## -- Conflicts ------------------------------------------------------------------------------- tidyvers
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
df2 <- select(XlData, -c(3,5,7,9,11,13,15,17,19,21,23,25))
str(df2)
```

```
## Classes 'tbl_df', 'tbl' and 'data.frame':    361 obs. of  13 variables:
##  $ Month              : POSIXct, format: "1990-04-01" "1990-05-01" ...
##  $ Coarse wool Price  : num  482 447 441 418 418 ...
##  $ Copra Price        : num  236 234 216 205 198 196 198 236 237 233 ...
##  $ Cotton Price       : num  1.83 1.89 1.99 2.01 1.79 1.79 1.79 1.82 1.85 1.85 ...
##  $ Fine wool Price    : num  1072 1057 898 896 951 ...
##  $ Hard log Price     : num  161 173 182 188 186 ...
##  $ Hard sawnwood Price: num  550 492 495 486 488 ...
##  $ Hide Price         : num  100 99.5 97.9 96.8 91.9 ...
##  $ Plywood Price      : num  312 350 374 378 365 ...
##  $ Rubber Price       : num  0.84 0.85 0.85 0.86 0.88 0.9 0.9 0.9 0.88 0.87 ...
##  $ Softlog Price      : num  121 124 129 124 130 ...
##  $ Soft sawnwood Price: num  219 213 200 210 208 ...
##  $ Wood pulp Price    : num  829 843 831 799 819 ...
```

```
names(df2)<-str_replace_all(names(df2), c(" " = "." , "," = "" ))
str(df2)
```

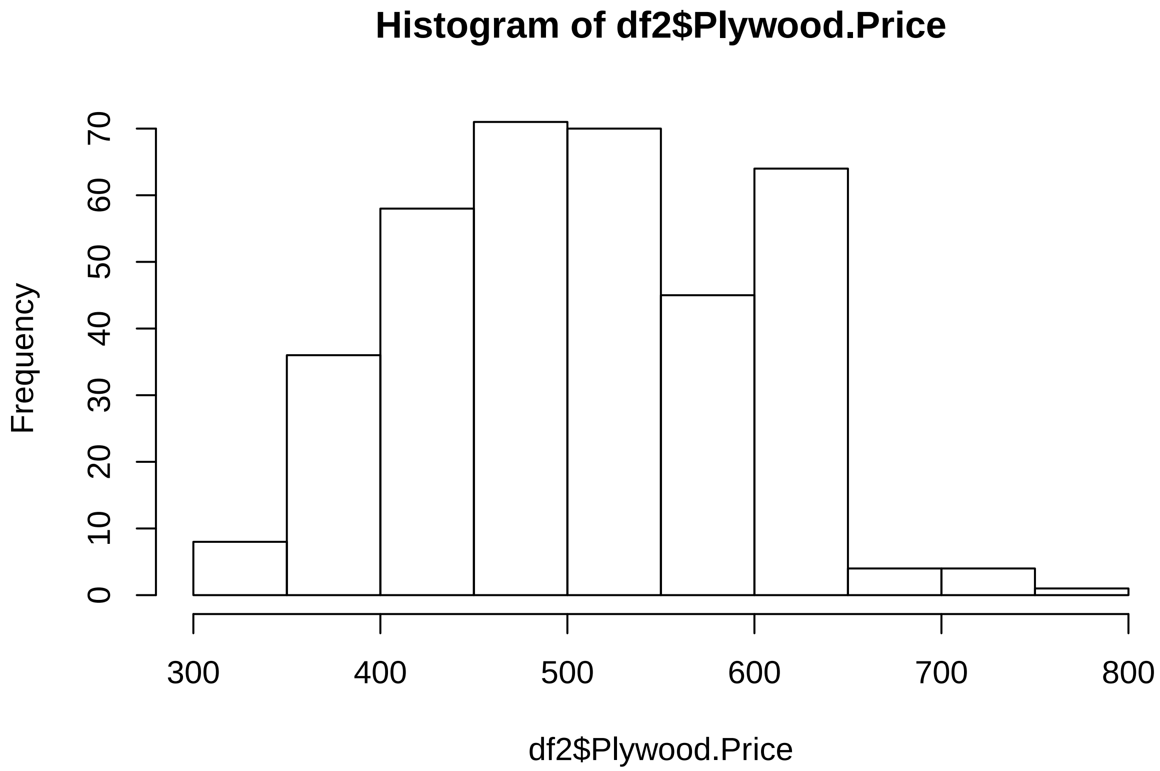
```
## Classes 'tbl_df', 'tbl' and 'data.frame':    361 obs. of  13 variables:
##  $ Month              : POSIXct, format: "1990-04-01" "1990-05-01" ...
##  $ Coarse.wool.Price  : num  482 447 441 418 418 ...
##  $ Copra.Price        : num  236 234 216 205 198 196 198 236 237 233 ...
##  $ Cotton.Price       : num  1.83 1.89 1.99 2.01 1.79 1.79 1.79 1.82 1.85 1.85 ...
##  $ Fine.wool.Price    : num  1072 1057 898 896 951 ...
##  $ Hard.log.Price     : num  161 173 182 188 186 ...
##  $ Hard.sawnwood.Price: num  550 492 495 486 488 ...
##  $ Hide.Price         : num  100 99.5 97.9 96.8 91.9 ...
##  $ Plywood.Price      : num  312 350 374 378 365 ...
##  $ Rubber.Price       : num  0.84 0.85 0.85 0.86 0.88 0.9 0.9 0.9 0.88 0.87 ...
##  $ Softlog.Price      : num  121 124 129 124 130 ...
##  $ Soft.sawnwood.Price: num  219 213 200 210 208 ...
##  $ Wood.pulp.Price    : num  829 843 831 799 819 ...
```

Step 1. Create Univariate plot for the variable of your interest (your Y variable).

Calculate skewness, kurtosis and describe the results.

Plot Histogram of target variable Plywood_Price

```
hist(df2$Plywood.Price)
```



Histogram above shows frequency of occurrence plywood price in the price range along 'X' axis.

Plot density distribution of target variable Plywood_Price

```
plot(density(df2$Plywood.Price))
```

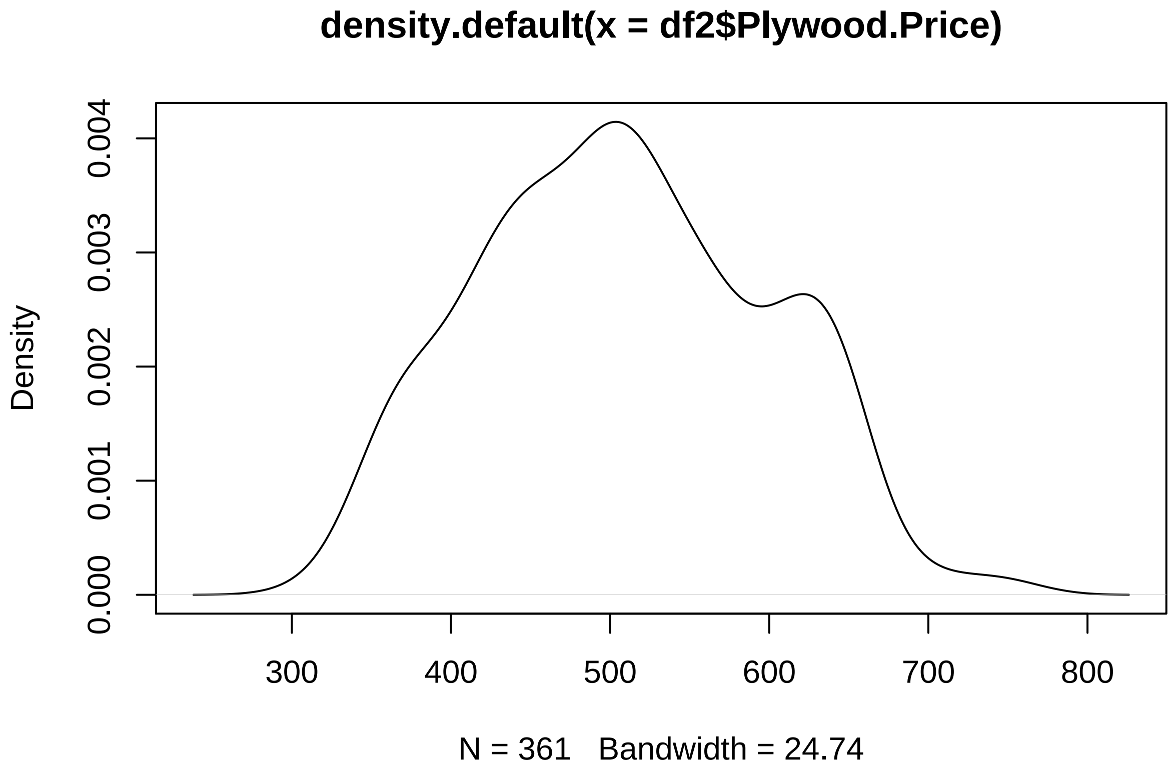


Figure above shows density distribution of plywood price. From density distribution shows distribution is skewed to right

Calculate distribution skewness of target variable Plywood.Price.

```
library(moments)
```

```
## Warning: package 'moments' was built under R version 3.5.2
```

Calculate Skewness value

Skewness is used to measure of symmetry of distribution.

```
skewness(df2$Plywood.Price)
```

```
## [1] 0.1435799
```

As skewness of distribution of plywood price is between -0.5 and 0.5, it is fairly symmetrical.

Calculate kurtosis value

Kurtosis value is a measure of outliers present in the distribution.

```
kurtosis(df2$Plywood.Price)
```

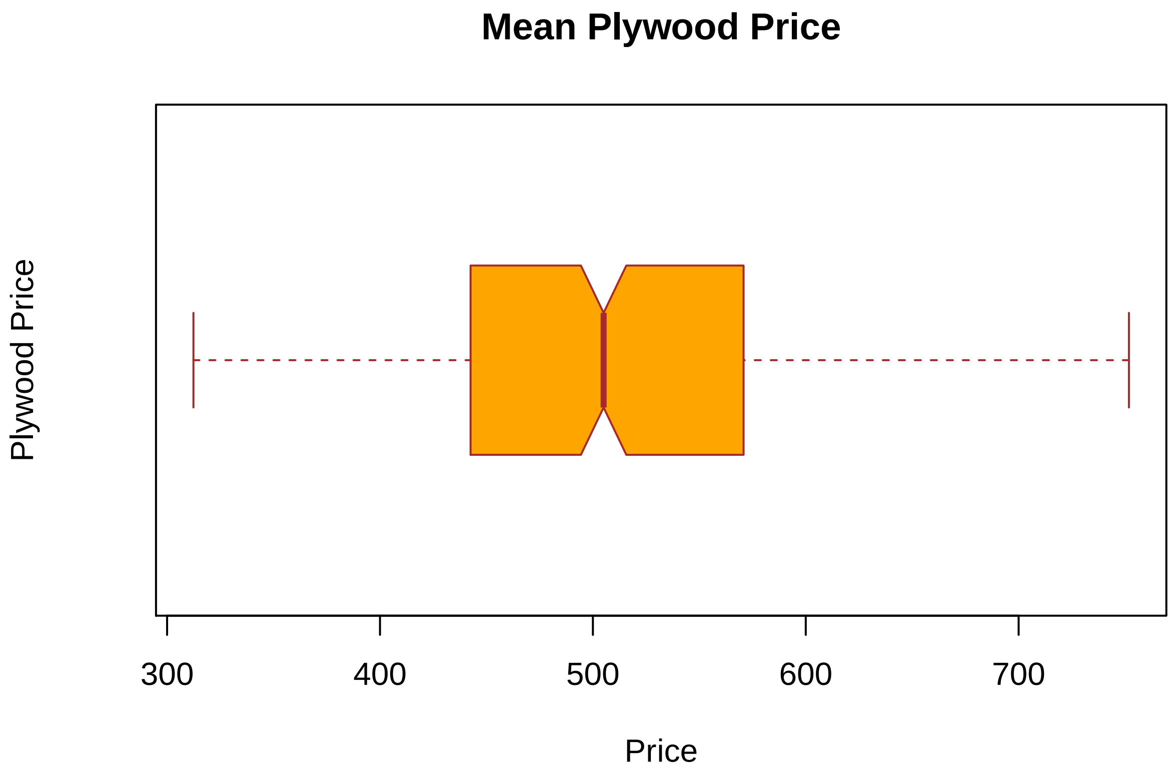
```
## [1] 2.325423
```

Since kurtosis value is less than 3 it could be observed that distribution is shorter, tails are thinner.

Step 2. Create Bivariate plot Box Plot for your Y variable and one of other important metrics (your X).

Box Plot

```
boxplot(df2$Plywood.Price,
main = "Mean Plywood Price",
xlab = "Price",
ylab = "Plywood Price",
col = "orange",
border = "brown",
horizontal = TRUE,
notch = TRUE
)
```

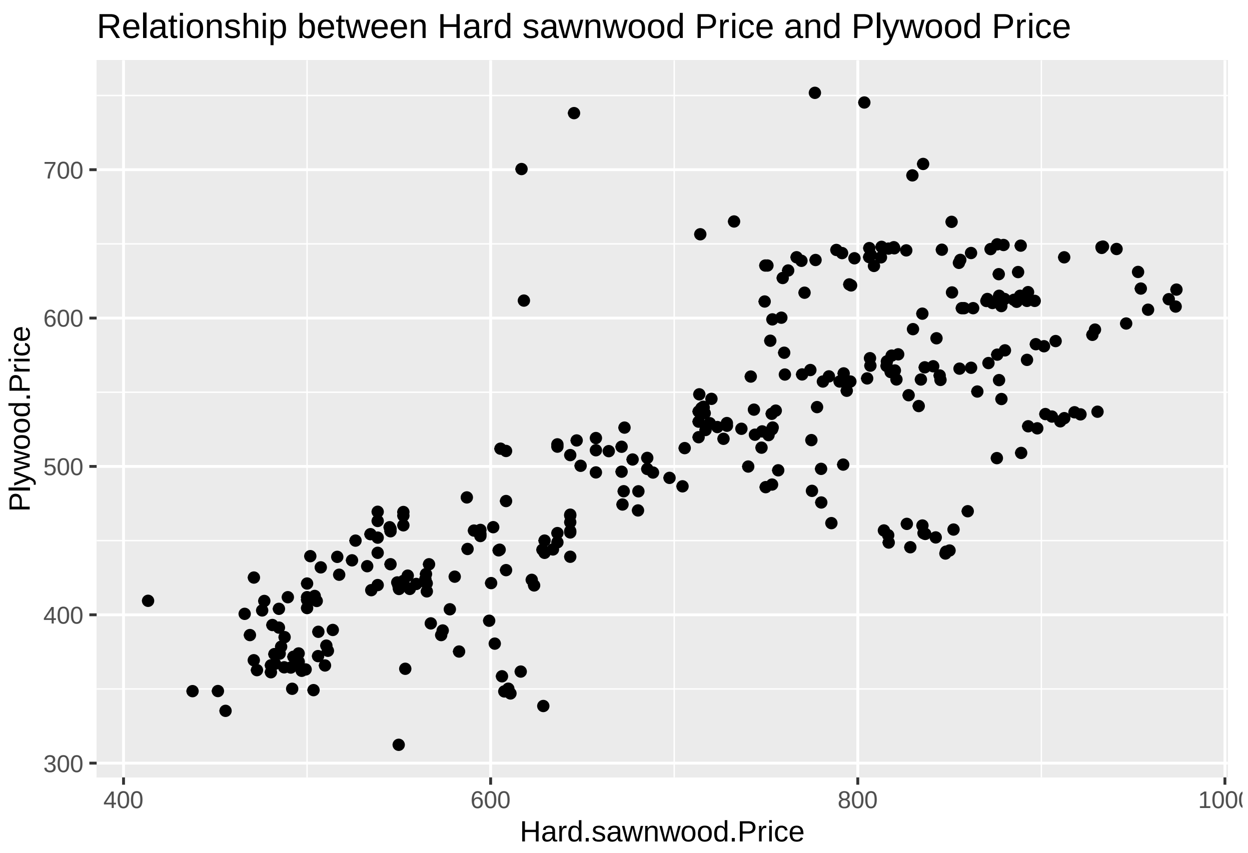


Above is box plot for plywood price. From above box plot it could be observed that Median price is 505.

Bivariate plot

```
df2 %>%
    ggplot(aes(Hard.sawnwood.Price, Plywood.Price)) +
    geom_point() +
    labs(title = "Relationship between Hard sawnwood Price and Plywood Price")
```

```
## Warning: Removed 34 rows containing missing values (geom_point).
```



Above is bivariate plot between plywood price and hard sandwood price. From this bivariate plot it could be observed that Hard sawnwood price is linearly proportional to plywood price.

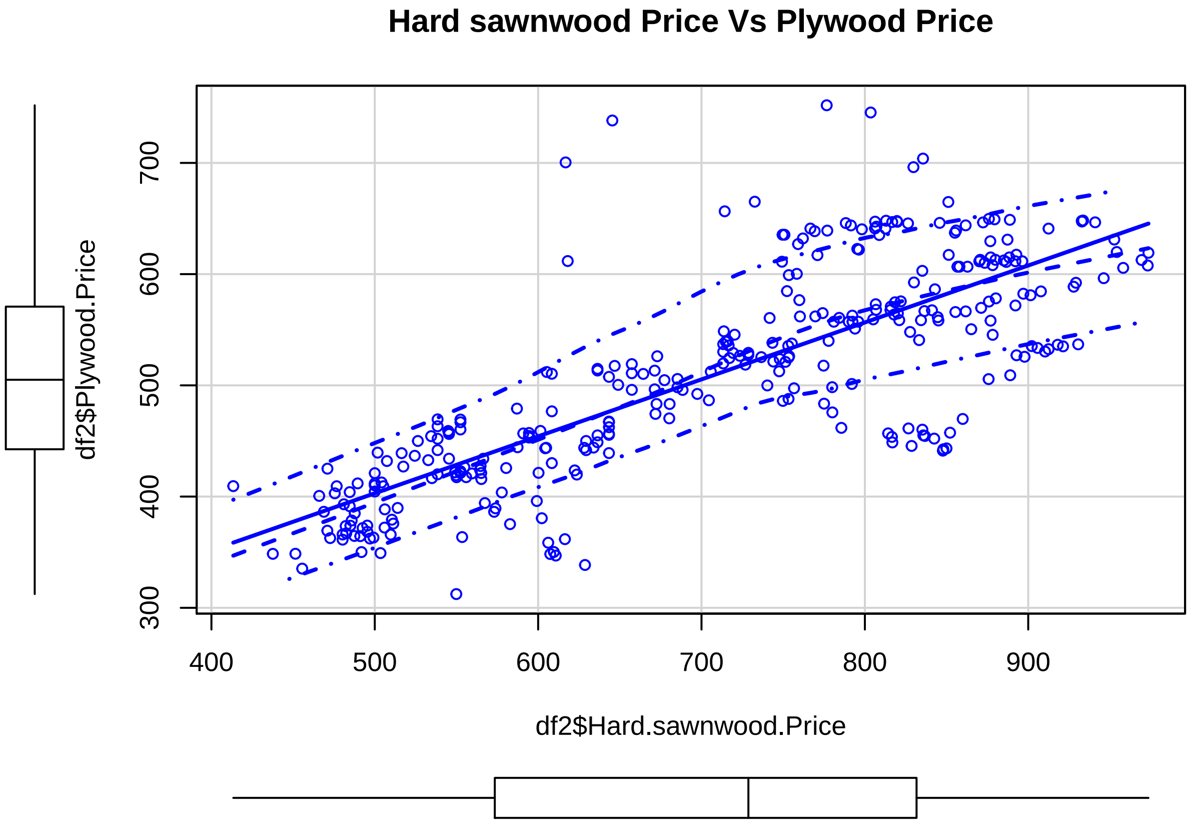
Step 3. Create scatter plots between between your Y and your X variables.

Describe the relationship between features (x) and Outcome (y) - linear, non-linear, no relation, positive, negative?

```
library(car)
```

```
## Warning: package 'car' was built under R version 3.5.3

## Loading required package: carData

##
## Attaching package: 'car'

## The following object is masked from 'package:dplyr':
##
##     recode

## The following object is masked from 'package:purrr':
##
##     some
```

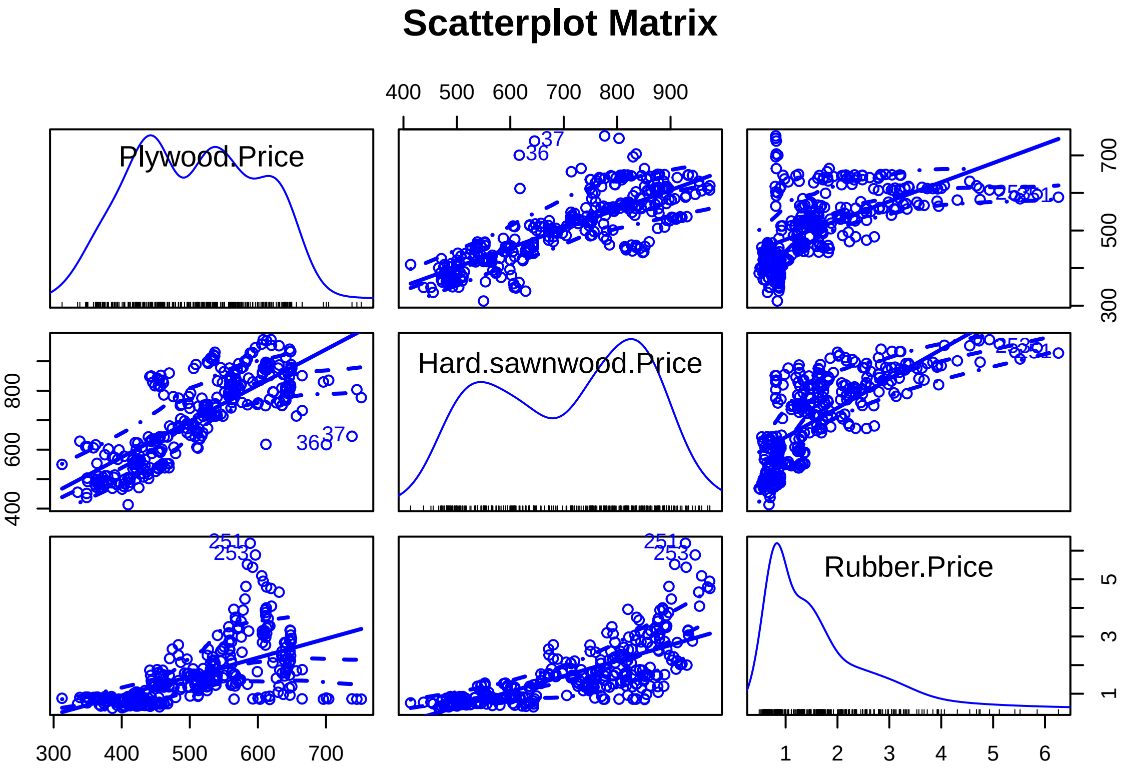
```
scatterplot(df2$Hard.sawnwood.Price, df2$Plywood.Price, main = "Hard sawnwood Price Vs Plywood Price")
```



Above is scatter plots for plywood price Vs Hard sawnwood price. Hard sawnwood price and plywood price are positively related.

Step 4. Create a multivariate plot - Use one scatter plot between X and Y from step 3 but add another X variable using colored symbols.

```
scatterplotMatrix(~ Plywood.Price + Hard.sawnwood.Price + Rubber.Price, data=df2, id=TRUE,  legend = TRU
```



Above are multivariate plots for plywood price, hard sawnwood price and rubber price. From above multivariate plots it could be observed that Plywood Price, Hard Sawnwood Price and Rubber Price are positively linearly related to each other.