Week 5 - Regression Assignment

Nitin

October 3, 2020

Read and Analyze Data

```r
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
```

Remove space in the variable names

```r
library(tidyverse)
```

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

## -- Attaching packages ---------------------------------------------------------------------------

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

```
names(XlData)<-str_replace_all(names(XlData), c(" " = "." , "," = "" ))
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" "3.459999
##  $ 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
```

Select columns of our interest from dataset.

```
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 ...
```

Step 1. Scale or normalize your data. Make sure to apply imputation if needed. 5pts [train/test split or K-fold CV if needed)

```
library(naniar)
miss_var_summary(df2)
```

```
## # A tibble: 13 x 3
##    variable            n_miss pct_miss
##    <chr>                <int>    <dbl>
##  1 Coarse.wool.Price       34    9.42
##  2 Fine.wool.Price         34    9.42
##  3 Hard.sawnwood.Price     34    9.42
##  4 Hide.Price              34    9.42
##  5 Softlog.Price           34    9.42
##  6 Soft.sawnwood.Price     34    9.42
##  7 Copra.Price             22    6.09
##  8 Wood.pulp.Price          1    0.277
##  9 Month                    0    0
## 10 Cotton.Price             0    0
## 11 Hard.log.Price           0    0
## 12 Plywood.Price            0    0
## 13 Rubber.Price             0    0
```

It could be observed from result display above that miss data is less than 10 % it could be ignored but we will delete rows in data set where data is missing.

```
df2 = na.omit(df2)
str(df2)
```

```
## Classes 'tbl_df', 'tbl' and 'data.frame':    327 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 ...
##  - attr(*, "na.action")= 'omit' Named int  328 329 330 331 332 333 334 335 336 337 ...
##   ..- attr(*, "names")= chr  "328" "329" "330" "331" ...
```

Scale Data

Scale continious verible Month

```
df2$Months <- scale(df2$Month)
```

```
data_scaled <- as.data.frame(scale(df2[,c(2:14)]))
summary(data_scaled)
```

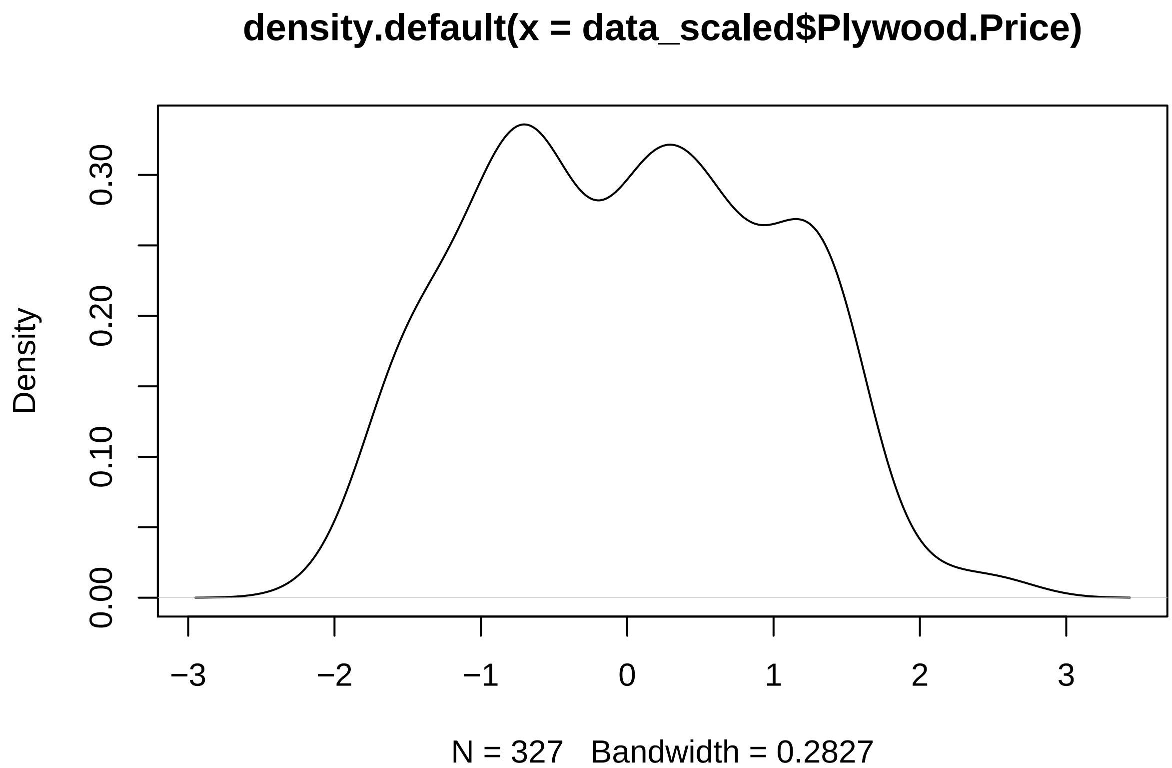
```
##  Coarse.wool.Price  Copra.Price       Cotton.Price     Fine.wool.Price
##  Min.   :-1.2657   Min.   :-1.3145   Min.   :-1.5042   Min.   :-1.5177
```

```
##  1st Qu.:-0.8567   1st Qu.:-0.6102   1st Qu.:-0.6501   1st Qu.:-0.7148  
##  Median :-0.3380   Median :-0.3035   Median :-0.1527   Median :-0.3576  
##  Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.0000  
##  3rd Qu.: 0.7368   3rd Qu.: 0.4832   3rd Qu.: 0.3916   3rd Qu.: 0.5955  
##  Max.   : 2.5535   Max.   : 3.6875   Max.   : 6.4543   Max.   : 3.5616  
##  Hard.log.Price     Hard.sawnwood.Price   Hide.Price      
##  Min.   :-1.68612   Min.   :-2.0377     Min.   :-3.65043  
##  1st Qu.:-0.79114   1st Qu.:-0.9303     1st Qu.:-0.66262  
##  Median :-0.02425   Median : 0.1436     Median :-0.09617  
##  Mean   : 0.00000   Mean   : 0.0000     Mean   : 0.00000  
##  3rd Qu.: 0.55355   3rd Qu.: 0.8556     3rd Qu.: 0.54295  
##  Max.   : 3.96122   Max.   : 1.8376     Max.   : 2.63416  
##  Plywood.Price       Rubber.Price     Softlog.Price    
##  Min.   :-2.10345   Min.   :-1.0966   Min.   :-1.7650  
##  1st Qu.:-0.80448   1st Qu.:-0.7687   1st Qu.:-0.7250  
##  Median : 0.03114   Median :-0.3098   Median :-0.1624  
##  Mean   : 0.00000   Mean   : 0.0000   Mean   : 0.0000  
##  3rd Qu.: 0.77138   3rd Qu.: 0.4629   3rd Qu.: 0.6127  
##  Max.   : 2.58725   Max.   : 4.3077   Max.   : 3.7288  
##  Soft.sawnwood.Price Wood.pulp.Price       Months          
##  Min.   :-3.1498     Min.   :-1.8616   Min.   :-1.7229354  
##  1st Qu.:-0.3949     1st Qu.:-0.8463   1st Qu.:-0.8617266  
##  Median : 0.1143     Median :-0.1019   Median :-0.0008646  
##  Mean   : 0.0000     Mean   : 0.0000   Mean   : 0.0000000  
##  3rd Qu.: 0.5805     3rd Qu.: 0.9697   3rd Qu.: 0.8620788  
##  Max.   : 2.3902     Max.   : 1.8182   Max.   : 1.7250222
```

Step 2. Build a multiple linear regression model or logistic regression (based on your Y) 10pts [or random forest, time series regression, stepwise, ridge, lasso]

Plot density distribution of target variable Plywood_Price.

```r
plot(density(data_scaled$Plywood.Price))
```



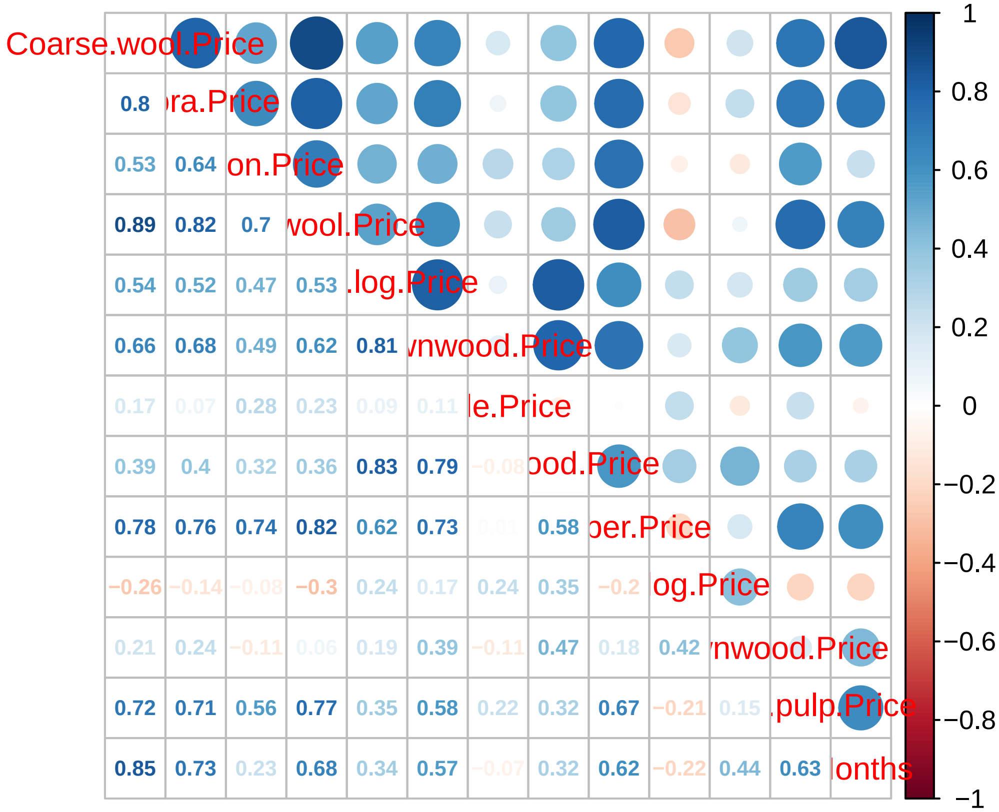
It could be observed that distribution is normal.

```r
library(corrplot)
```

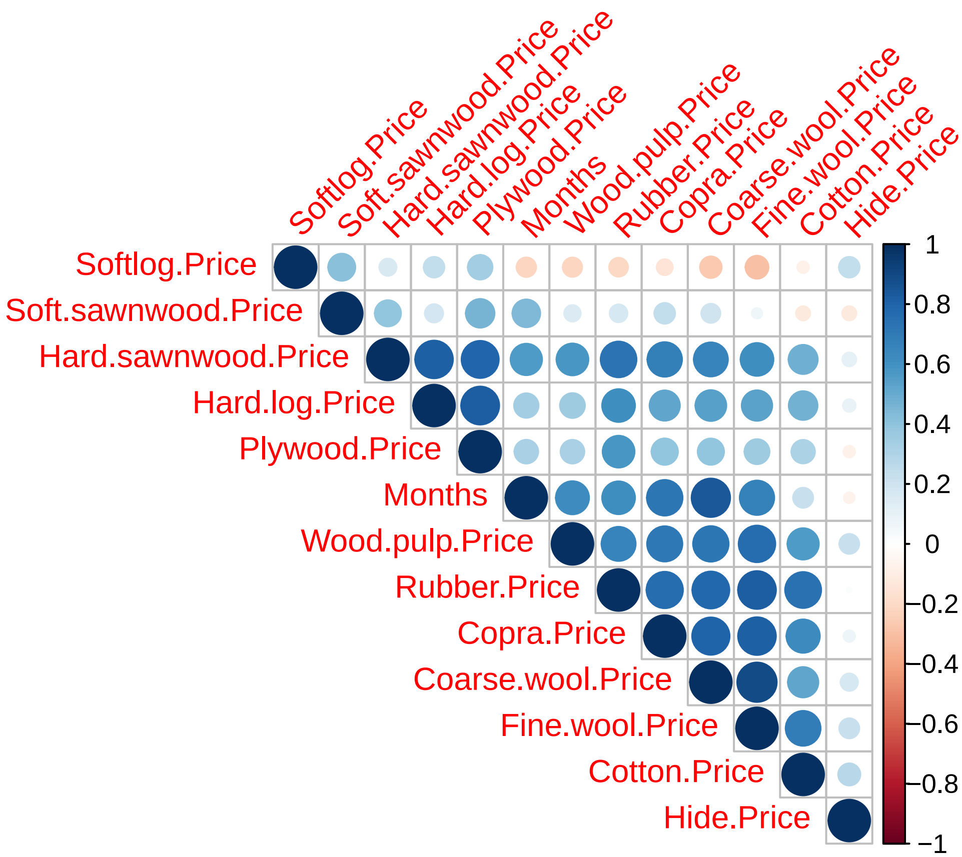
```
## Warning: package 'corrplot' was built under R version 3.5.3
```

```
## corrplot 0.84 loaded
```

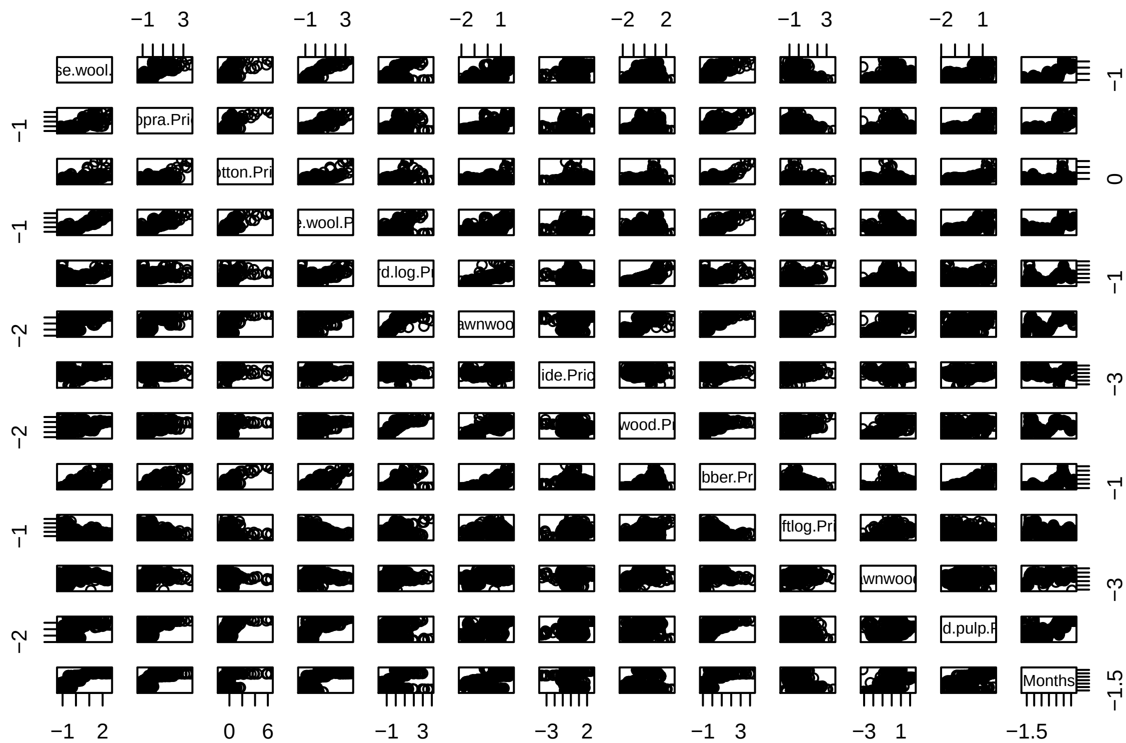
```r
cor1 = cor(data_scaled)
corrplot.mixed(cor1, number.cex = .7)
```




```
corrplot(cor1, order = "hclust", type='upper', tl.srt=45)
```




```
plot(data_scaled)
```



Build a multiple linear regression model using stepwise forward selection approuch.

```
#model1 <- lm(Plywood.Price~(Hard.sawnwood.Price + Hide.Price + Months + Coarse.wool.Price + Copra.Pric
model1 <- lm(Plywood.Price~ ., data=data_scaled)
#head(data_scaled)
formula(model1)
```

```
## Plywood.Price ~ Coarse.wool.Price + Copra.Price + Cotton.Price +
##     Fine.wool.Price + Hard.log.Price + Hard.sawnwood.Price +
##     Hide.Price + Rubber.Price + Softlog.Price + Soft.sawnwood.Price +
##     Wood.pulp.Price + Months
```

Step 3. Print summary and interpret table (see lecture slides). Describe the summary or the output of your regression. 15 pts

```
summary(model1)
```

```
##
## Call:
## lm(formula = Plywood.Price ~ ., data = data_scaled)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.16676 -0.19425  0.01522  0.17979  1.29131
##
## Coefficients:
```

```
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)         -4.365e-16  2.056e-02   0.000 1.000000
## Coarse.wool.Price   -9.449e-02  7.702e-02  -1.227 0.220805
## Copra.Price         -1.754e-01  4.649e-02  -3.774 0.000192 ***
## Cotton.Price        -1.248e-01  4.605e-02  -2.709 0.007116 **
## Fine.wool.Price     -6.042e-02  6.504e-02  -0.929 0.353625
## Hard.log.Price       6.136e-01  4.529e-02  13.548  < 2e-16 ***
## Hard.sawnwood.Price  1.908e-01  5.197e-02   3.672 0.000283 ***
## Hide.Price          -1.124e-01  2.915e-02  -3.856 0.000140 ***
## Rubber.Price         3.774e-01  5.620e-02   6.715 8.82e-11 ***
## Softlog.Price        7.983e-02  3.497e-02   2.283 0.023119 *
## Soft.sawnwood.Price  2.608e-01  3.238e-02   8.054 1.68e-14 ***
## Wood.pulp.Price      1.475e-01  3.669e-02   4.019 7.32e-05 ***
## Months              -1.503e-01  6.133e-02  -2.450 0.014818 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3718 on 314 degrees of freedom
## Multiple R-squared:  0.8668, Adjusted R-squared:  0.8618
## F-statistic: 170.3 on 12 and 314 DF,  p-value: < 2.2e-16
```

From summary for model1 it could be observed that for Fine.wool.Price Pr vlaue is not significant, second insignificant Pr value is for Coarse.wool.Price for second model (model2) let us eliminated one variable Fine.wool.Price from the model1. It could be observed that Adjusted R-squared for model1 is 0.8618. Also use Akaike Information Criterion (AIC) approach to conform our decision.

Akaike Information Criterion (AIC) from model1

```r
step(model1, direction = "backward")
```

```
## Start:  AIC=-634.31
## Plywood.Price ~ Coarse.wool.Price + Copra.Price + Cotton.Price +
##     Fine.wool.Price + Hard.log.Price + Hard.sawnwood.Price +
##     Hide.Price + Rubber.Price + Softlog.Price + Soft.sawnwood.Price +
##     Wood.pulp.Price + Months
##
##                       Df Sum of Sq    RSS     AIC
## - Fine.wool.Price      1    0.1193 43.528 -635.41
## - Coarse.wool.Price    1    0.2081 43.617 -634.74
## <none>                             43.409 -634.31
## - Softlog.Price        1    0.7203 44.130 -630.93
## - Months               1    0.8300 44.239 -630.11
## - Cotton.Price         1    1.0146 44.424 -628.75
## - Hard.sawnwood.Price  1    1.8642 45.273 -622.56
## - Copra.Price          1    1.9688 45.378 -621.80
## - Hide.Price           1    2.0551 45.464 -621.18
## - Wood.pulp.Price      1    2.2332 45.642 -619.90
## - Rubber.Price         1    6.2337 49.643 -592.43
## - Soft.sawnwood.Price  1    8.9683 52.377 -574.90
## - Hard.log.Price       1   25.3735 68.783 -485.80
##
## Step:  AIC=-635.41
## Plywood.Price ~ Coarse.wool.Price + Copra.Price + Cotton.Price +
##     Hard.log.Price + Hard.sawnwood.Price + Hide.Price + Rubber.Price +
```

```
##     Softlog.Price + Soft.sawnwood.Price + Wood.pulp.Price + Months
##
##                      Df Sum of Sq    RSS     AIC
## <none>                            43.528 -635.41
## - Coarse.wool.Price    1    0.4517 43.980 -634.03
## - Months               1    0.7775 44.306 -631.62
## - Softlog.Price        1    0.8672 44.396 -630.96
## - Cotton.Price         1    1.1271 44.656 -629.05
## - Wood.pulp.Price      1    2.1200 45.649 -621.86
## - Hard.sawnwood.Price  1    2.1331 45.662 -621.77
## - Hide.Price           1    2.3353 45.864 -620.32
## - Copra.Price          1    2.4113 45.940 -619.78
## - Rubber.Price         1    6.1406 49.669 -594.26
## - Soft.sawnwood.Price  1    8.9928 52.521 -576.00
## - Hard.log.Price       1   25.5334 69.062 -486.47
```

```
##
## Call:
## lm(formula = Plywood.Price ~ Coarse.wool.Price + Copra.Price +
##     Cotton.Price + Hard.log.Price + Hard.sawnwood.Price + Hide.Price +
##     Rubber.Price + Softlog.Price + Soft.sawnwood.Price + Wood.pulp.Price +
##     Months, data = data_scaled)
##
## Coefficients:
##         (Intercept)    Coarse.wool.Price          Copra.Price
##          -4.434e-16           -1.255e-01           -1.870e-01
##        Cotton.Price       Hard.log.Price  Hard.sawnwood.Price
##          -1.304e-01            6.063e-01            2.002e-01
##          Hide.Price         Rubber.Price        Softlog.Price
##          -1.176e-01            3.668e-01            8.600e-02
## Soft.sawnwood.Price      Wood.pulp.Price               Months
##           2.612e-01            1.379e-01           -1.448e-01
```

From Akaike Information Criterion (AIC) results above it could be observed that when we do not eliminate any variable AIC will be -634.31 but if we remove variable Fine.wool.Price from the model AIC improves to -635.41. Let us remove variable Fine.wool.Price from model and perform analysis again.

Step 4. Perform another model and evaluate which model performs better. 10pts [if you have stepwise regression - you do not need to create another model - just explain which model is the best]

```r
model2 <- lm(Plywood.Price~Coarse.wool.Price + Copra.Price + Cotton.Price +
    Hard.log.Price + Hard.sawnwood.Price + Hide.Price + Rubber.Price +
    Softlog.Price + Soft.sawnwood.Price + Wood.pulp.Price + Months, data=data_scaled)
```

```r
summary(model2)
```

```
##
## Call:
## lm(formula = Plywood.Price ~ Coarse.wool.Price + Copra.Price +
##     Cotton.Price + Hard.log.Price + Hard.sawnwood.Price + Hide.Price +
##     Rubber.Price + Softlog.Price + Soft.sawnwood.Price + Wood.pulp.Price +
##     Months, data = data_scaled)
##
## Residuals:
```

```
##      Min       1Q   Median       3Q      Max
## -1.1931 -0.1979  0.0159  0.1781  1.3166
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)         -4.434e-16  2.056e-02   0.000 1.000000
## Coarse.wool.Price   -1.255e-01  6.941e-02  -1.808 0.071557 .
## Copra.Price         -1.870e-01  4.477e-02  -4.177 3.82e-05 ***
## Cotton.Price        -1.304e-01  4.565e-02  -2.856 0.004575 **
## Hard.log.Price       6.063e-01  4.461e-02  13.593  < 2e-16 ***
## Hard.sawnwood.Price  2.002e-01  5.096e-02   3.929 0.000105 ***
## Hide.Price          -1.176e-01  2.860e-02  -4.111 5.03e-05 ***
## Rubber.Price         3.668e-01  5.503e-02   6.666 1.18e-10 ***
## Softlog.Price        8.600e-02  3.433e-02   2.505 0.012744 *
## Soft.sawnwood.Price  2.612e-01  3.237e-02   8.067 1.53e-14 ***
## Wood.pulp.Price      1.379e-01  3.520e-02   3.917 0.000110 ***
## Months              -1.448e-01  6.103e-02  -2.372 0.018289 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3717 on 315 degrees of freedom
## Multiple R-squared:  0.8665, Adjusted R-squared:  0.8618
## F-statistic: 185.8 on 11 and 315 DF,  p-value: < 2.2e-16
```

It could be observed from summary of model2 that Adjusted R-squared for model2 is 0.8618, which is same as model1.