MIS40970 - Data Mining for Business Analytics

Deepak Kumar Gupta (MScBA - 16200660)

Due Date: Feb 15, 2017

Contents

1 Load the oj.csv data into R. 2

2 How many records and how many attributes are in the orange juice dataset? 3

3 What is the mean, standard deviation and range of the price of orange juice? 3

4 What is the median of the 'log of number of units sold' (logmove)? 3

5 What are the names of the 3 orange juice brands? 3

6 Create a histogram of prices for each brand of orange juice. 4
6.1 Method 1: using hist() . . . 4
6.2 Method 2: Using ggplot2() . . . 4
6.3 Method 3: Using histogram() . . . 5

7 Generate a boxplot, which includes a separate plot for the prices of each brand 6

8 What does the boxplot tell us about the relative prices of each brand? 6

9 Generate a scatterplot of the logmove compared to price, and color the points according to their brand 7
9.1 Method 1: Using pairs() . . . 7
9.2 Method 2: Using plot() . . . 7
9.3 Method 3: Using plotly() . . . 8

10 Based on what you observe in the scatterplot, what can we say about the price of each brand of chilled orange juice and the volume of sales? 8

11 Calculate the mean price of orange juice sold each week, and create aline plot of this timeseries. 10
11.1 Method 1: using tapply() . . . 10

12 Extract the mean weekly price of orange juice sold each week according to each brand. 11
12.1 Method 1: using aggregate() . . . 11
12.2 Method :2 using tapply() . . . 11

13 Create a plot which compares the mean weekly price of orange juice for all brands versus each individual brand. 11

14 When there is an advertising campaign for orange juice does it impact on the number of units sold? 12

15 Can you create a line plot of the mean weekly units sold without a promotion overlayed with the mean weekly units sold with a promotion? What is interesting about this plot 12

16 Descriptive Analytics **13**
16.1 Answer: Yes, there are patterns among demographics variables, which can be use for marketing of orange juice at given stores. . . . 13
16.2 Overview . . . 13
16.3 Descriptive Analytics Summary for demographic data . . . 14
16.4 Scatter Plot Matrices in R . . . 15
16.4.1 Analysis of Correlation among demographics variables. . . . 15
16.4.1.1 1. Positive Correlations: . . . 15
16.4.1.2 1. Negative Correlations: . . . 15

---

1 Load the oj.csv data into R.

```r
options(warn = 1)
oj <- read.csv(file="oj.csv",head=TRUE,sep=",")
summary(oj)
```

```
##      store                brand           week          logmove
##  Min.   :  2.00   dominicks  :9649   Min.   : 40.0   Min.   : 4.159
##  1st Qu.: 53.00   minute.maid:9649   1st Qu.: 70.0   1st Qu.: 8.490
##  Median : 86.00   tropicana  :9649   Median :101.0   Median : 9.034
##  Mean   : 80.88                      Mean   :100.5   Mean   : 9.168
##  3rd Qu.:111.00                      3rd Qu.:130.0   3rd Qu.: 9.765
##  Max.   :137.00                      Max.   :160.0   Max.   :13.482
##       feat            price           AGE60              EDUC
##  Min.   :0.0000   Min.   :0.520   Min.   :0.05805   Min.   :0.04955
##  1st Qu.:0.0000   1st Qu.:1.790   1st Qu.:0.12210   1st Qu.:0.14598
##  Median :0.0000   Median :2.170   Median :0.17065   Median :0.22939
##  Mean   :0.2373   Mean   :2.282   Mean   :0.17313   Mean   :0.22522
##  3rd Qu.:0.0000   3rd Qu.:2.730   3rd Qu.:0.21395   3rd Qu.:0.28439
##  Max.   :1.0000   Max.   :3.870   Max.   :0.30740   Max.   :0.52836
##      ETHNIC            INCOME          HHLARGE           WORKWOM
##  Min.   :0.02425   Min.   : 9.867   Min.   :0.01351   Min.   :0.2445
##  1st Qu.:0.04191   1st Qu.:10.456   1st Qu.:0.09794   1st Qu.:0.3126
##  Median :0.07466   Median :10.635   Median :0.11122   Median :0.3556
##  Mean   :0.15556   Mean   :10.617   Mean   :0.11560   Mean   :0.3592
##  3rd Qu.:0.18776   3rd Qu.:10.797   3rd Qu.:0.13517   3rd Qu.:0.4023
##  Max.   :0.99569   Max.   :11.236   Max.   :0.21635   Max.   :0.4723
##     HVAL150            SSTRDIST          SSTRVOL          CPDIST5
##  Min.   :0.002509   Min.   : 0.1321   Min.   :0.4000   Min.   :0.7725
##  1st Qu.:0.123486   1st Qu.: 2.7670   1st Qu.:0.7273   1st Qu.:1.6262
##  Median :0.346154   Median : 4.6507   Median :1.1154   Median :1.9634
##  Mean   :0.343766   Mean   : 5.0973   Mean   :1.2073   Mean   :2.1204
##  3rd Qu.:0.528313   3rd Qu.: 6.6506   3rd Qu.:1.5385   3rd Qu.:2.5337
##  Max.   :0.916700   Max.   :17.8560   Max.   :2.5714   Max.   :4.1079
##     CPWVOL5
##  Min.   :0.09456
##  1st Qu.:0.27167
##  Median :0.38323
##  Mean   :0.43891
##  3rd Qu.:0.56024
##  Max.   :1.14337
```

2 How many records and how many attributes are in the orange juice dataset?

There are 28947 records and 17 attributes.

```
#Method 1
dim(oj)
```

```
## [1] 28947    17
```

```
#Method 2
nrow(oj)
```

```
## [1] 28947
```

```
ncol(oj)
```

```
## [1] 17
```

```
names(oj)
```

```
##  [1] "store"    "brand"    "week"     "logmove"  "feat"     "price"
##  [7] "AGE60"    "EDUC"     "ETHNIC"   "INCOME"   "HHLARGE"  "WORKWOM"
## [13] "HVAL150"  "SSTRDIST" "SSTRVOL"  "CPDIST5"  "CPWVOL5"
```

3 What is the mean, standard deviation and range of the price of orange juice?

```
mean(oj$price)
```

```
## [1] 2.282488
```

```
sd(oj$price)
```

```
## [1] 0.6480007
```

```
range(oj$price)
```

```
## [1] 0.52 3.87
```

4 What is the median of the 'log of number of units sold' (logmove)?

Median is 9.03408

```
median(oj$logmove)
```

```
## [1] 9.03408
```

5 What are the names of the 3 orange juice brands?

- Juice Brands are:

1. Dominicks
2. Minute.maid
3. Tropicana

```
#Method 1
unique(oj$brand)
```

```
## [1] tropicana   minute.maid dominicks
## Levels: dominicks minute.maid tropicana
```

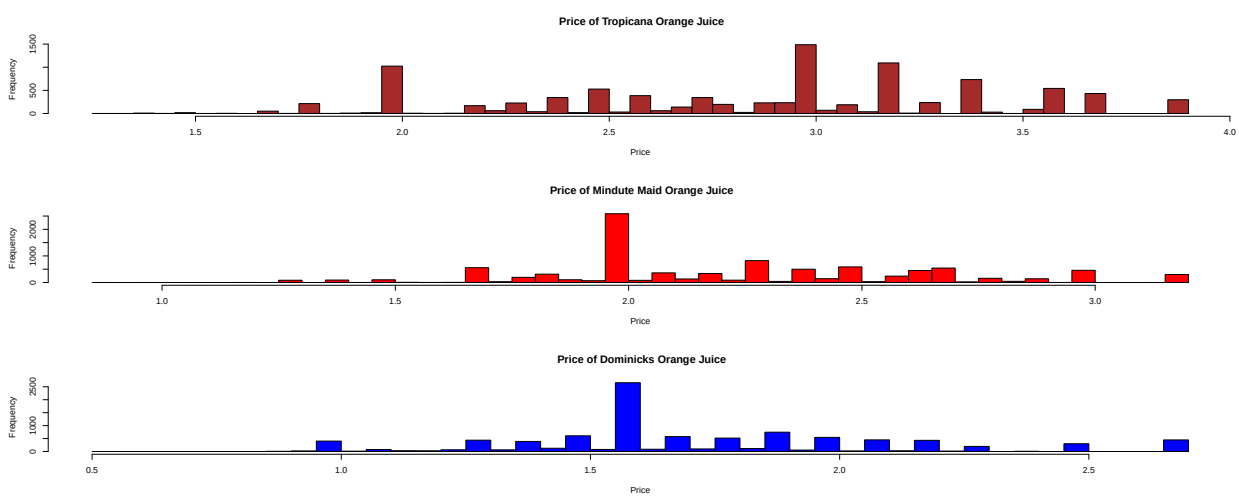
```
#Method 2
juicebrand<- factor(oj$brand)
table(juicebrand)
```

```
## juicebrand
##   dominicks minute.maid   tropicana
##        9649        9649        9649
```

6 Create a histogram of prices for each brand of orange juice.

6.1 Method 1: using hist()

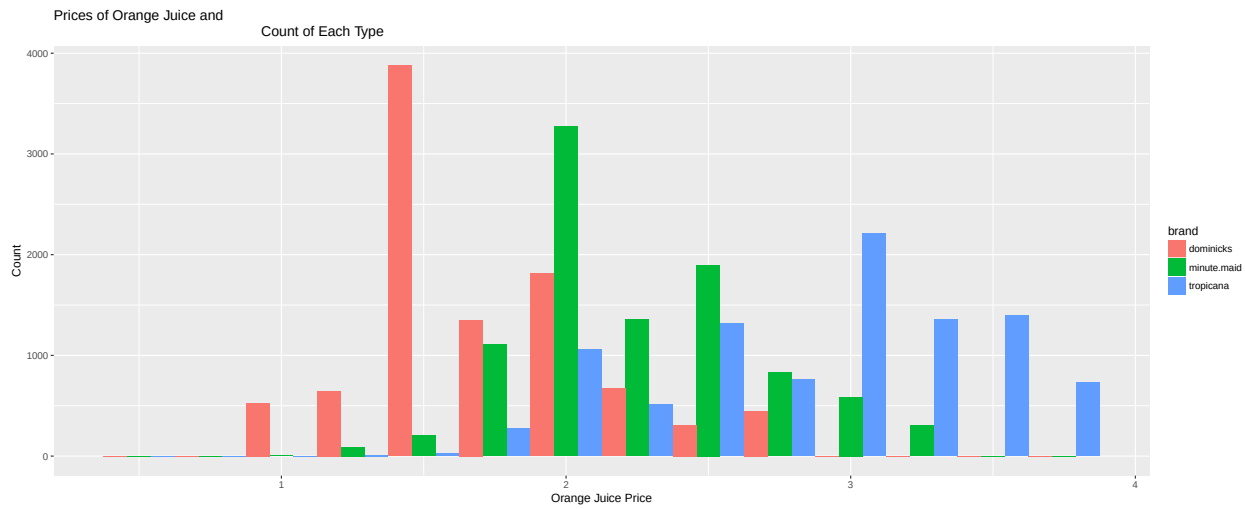
```
library(lattice)
par(mfrow = c(3,1))
hist(oj$price[oj$brand=='tropicana'],main=" Price of Tropicana Orange Juice"
       ,xlab="Price",breaks =50, col = 'brown')
hist(oj$price[oj$brand=='minute.maid'],main=" Price of Mindute Maid Orange Juice"
       ,xlab="Price",breaks =50,col='Red')
hist(oj$price[oj$brand=='dominicks'],main=" Price of Dominicks Orange Juice"
     ,xlab="Price",breaks =50, col = 'blue')
```



6.2 Method 2: Using ggplot2()

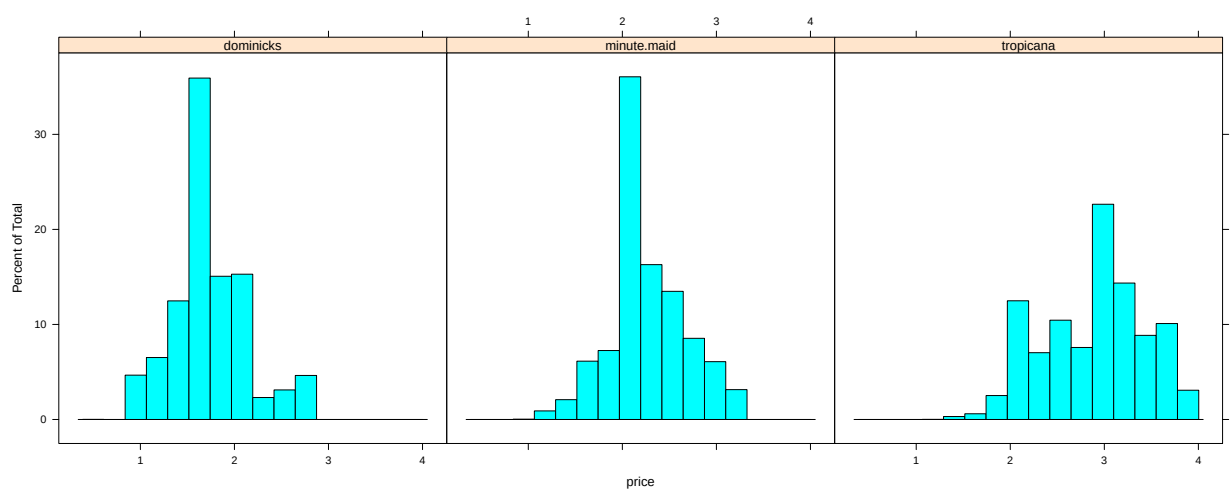
```
library(ggplot2)
par(mfrow = c(3,3))
```

```
ggplot(oj, aes(x= oj$price, fill = brand)) + geom_histogram(binwidth=0.25,position = "dodge")+
ylab("Count") + xlab("Orange Juice Price") + ggtitle("Prices of Orange Juice and
                                                      Count of Each Type")
```

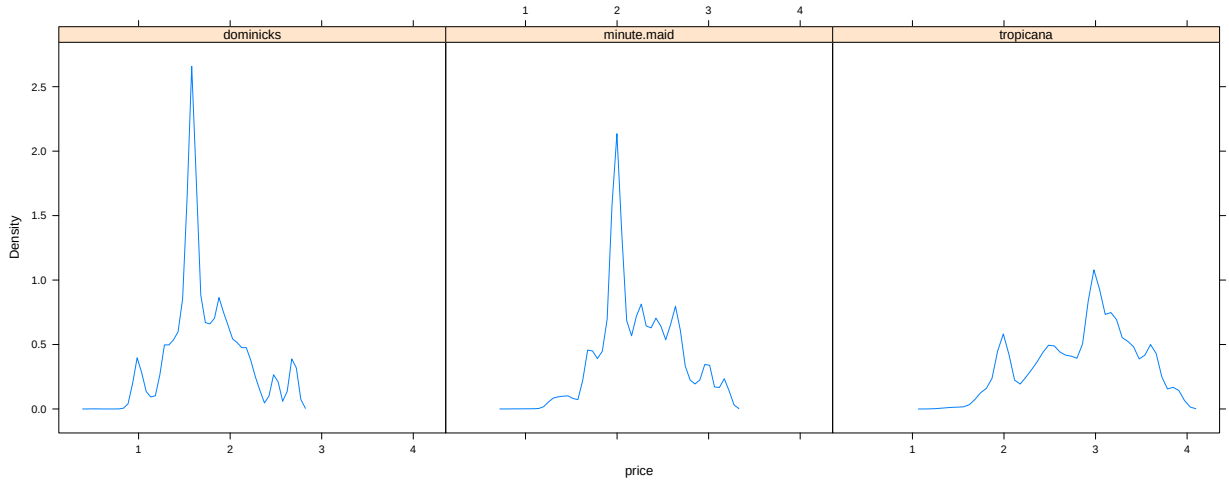


6.3 Method 3: Using histogram()

```
histogram(~price|brand,data=oj,layout=c(3,1))
```

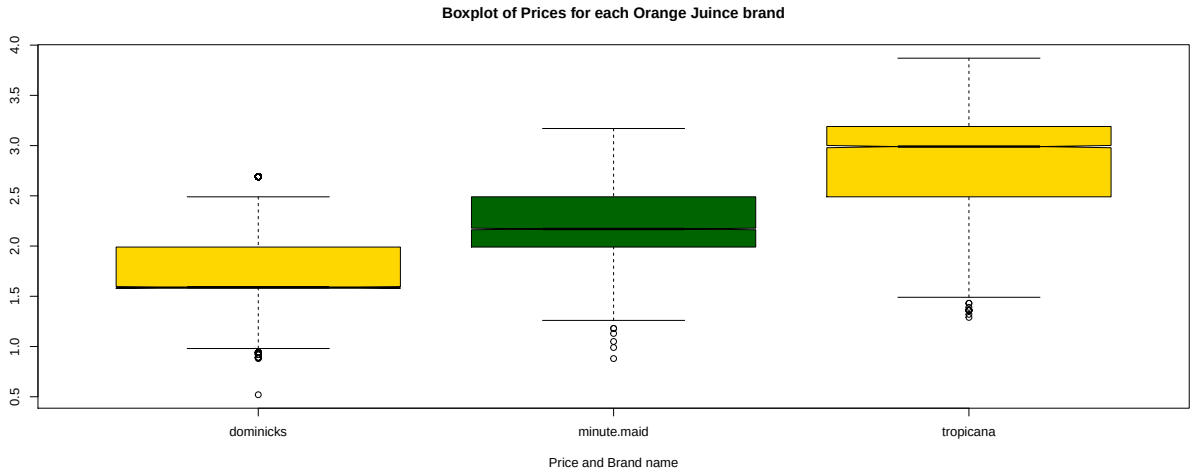



```
densityplot(~price|brand,data=oj,layout=c(3,1),plot.points=FALSE)
```



7 Generate a boxplot, which includes a separate plot for the prices of each brand

```
boxplot(oj$price~oj$brand, data=oj, notch=TRUE,
  col=(c("gold","darkgreen")),main="Boxplot of Prices for each Orange Juince brand",
  xlab="Price and Brand name")
```



8 What does the boxplot tell us about the relative prices of each brand?

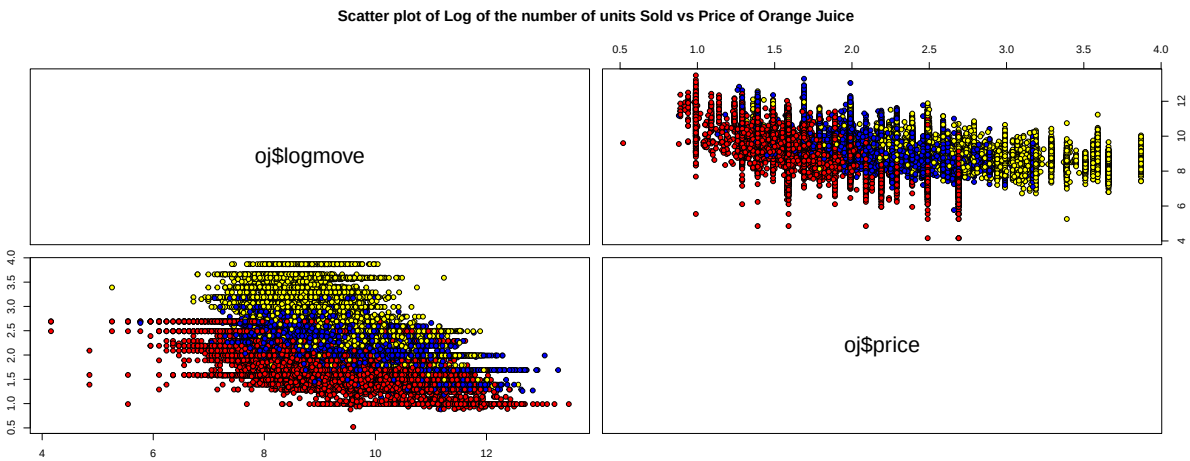
- Dominick's has outlier less than 3/2 times of lower quartile. one outlier is more than 3/2 time of upper quartile. Also, majority of price range of Dominick's is more than the median value and it has range of 1.6 to 2 (approx)
- Minute.maid has about 5 outliers less than 3/2 times of lower quartile.

- Tropicana has few outliers closer to the minimum of lower quartile whiskers. Overall, with the help of boxplot, one can interpret the distribution of figure around quartile of a median value. That will help a user to identify outlier values and then perform detailed data analysis to find trends and pattern among datasets.

9 Generate a scatterplot of the logmove compared to price, and color the points according to their brand

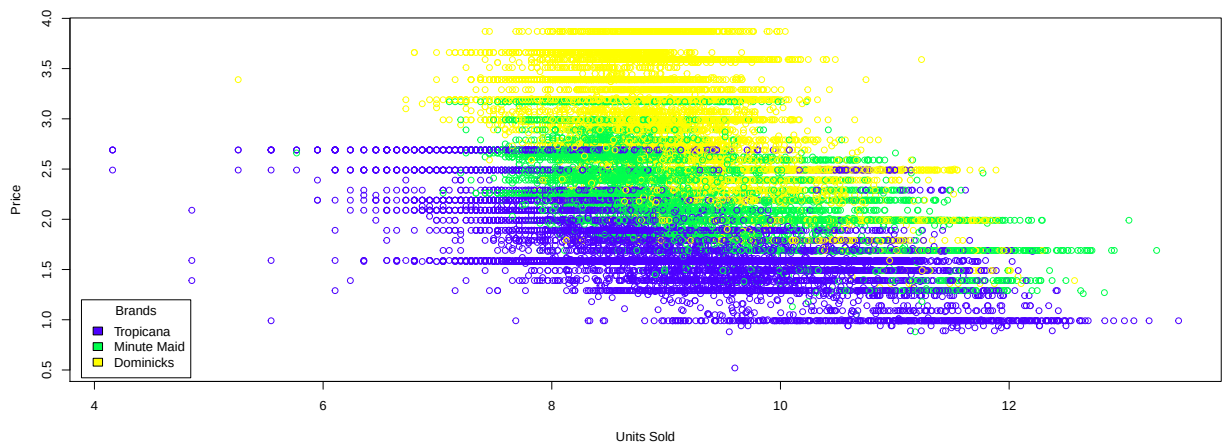
9.1 Method 1: Using pairs()

```
library(scatterplot3d)
attach(oj)
pairs(~oj$logmove+oj$price, pch=21,
      main="Scatter plot of Log of the number of units Sold vs Price of Orange Juice",
      bg= c("red", "blue", "yellow")[unclass(oj$brand)])
```



9.2 Method 2: Using plot()

```
plot(oj$logmove,oj$price, xlab="Units Sold",ylab="Price",pch=21,col=topo.colors(3)
     [unclass(oj$brand)])
legend("bottomleft", inset=0.01, title="Brands",c("Tropicana","Minute Maid","Dominicks"),
       fill=topo.colors(3), horiz=FALSE)
```



9.3 Method 3: Using plotly()

```
library(plotly)
Sys.setenv("plotly_username"="deepak91g")
Sys.setenv("plotly_api_key"="Gb3tMDu2UzMuGdXsMB9n")

#p <- plot_ly(data = oj, x = ~logmove, y = ~price, color = ~brand, colors = "Set1")

# Create a shareable link to your chart
# Set up API credentials: https://plot.ly/r/getting-started
#chart_link = plotly_POST(p, filename="colorbrewer")
#chart_link
```

10 Based on what you observe in the scatterplot, what can we say about the price of each brand of chilled orange juice and the volume of sales?

- Scatterplot helps one to identify the relation between two variables if one changes then how the change in other is reflected. It could be no change, increase or decrease in the value. Similarly, for orange juice data, the scatterplot of Units Sold vs Price shows that the Price of Tropicana juice is lowered compared to other two, and sold the number of units compared to Minute Maid and Dominicks.
- Dominicks has sold units in range of 7-11, but its price is highest among all three brands. On the other hand, minute maid unit sold is between 8-12, and average price per unit is 2.5.
- Overall, Maximum price of a unit juice is 4.0 and minimum price is approx 1.0. Max unit sold is 14 and minmum number of unit sold is 4.

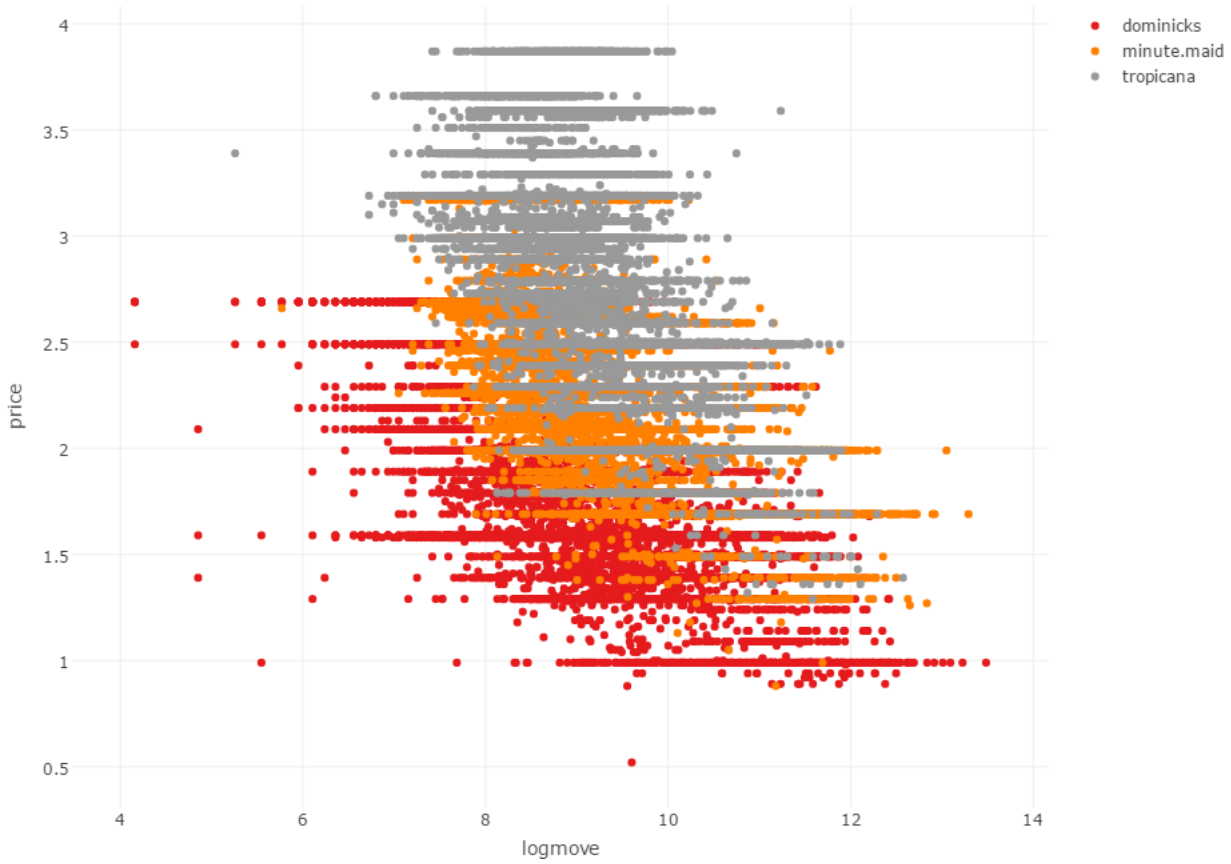
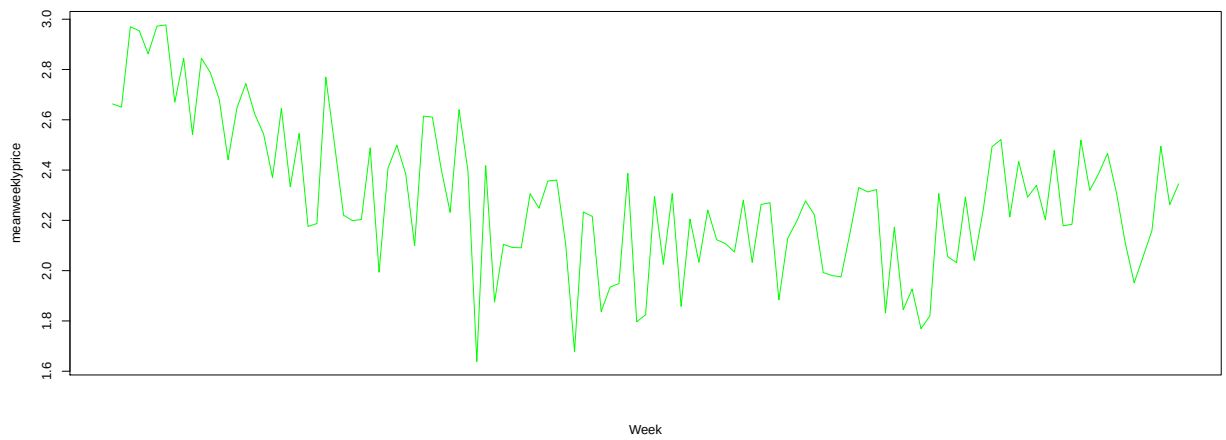


Figure 1: Scatter Plot using Ploty library

11 Calculate the mean price of orange juice sold each week, and create aline plot of this timeseries.

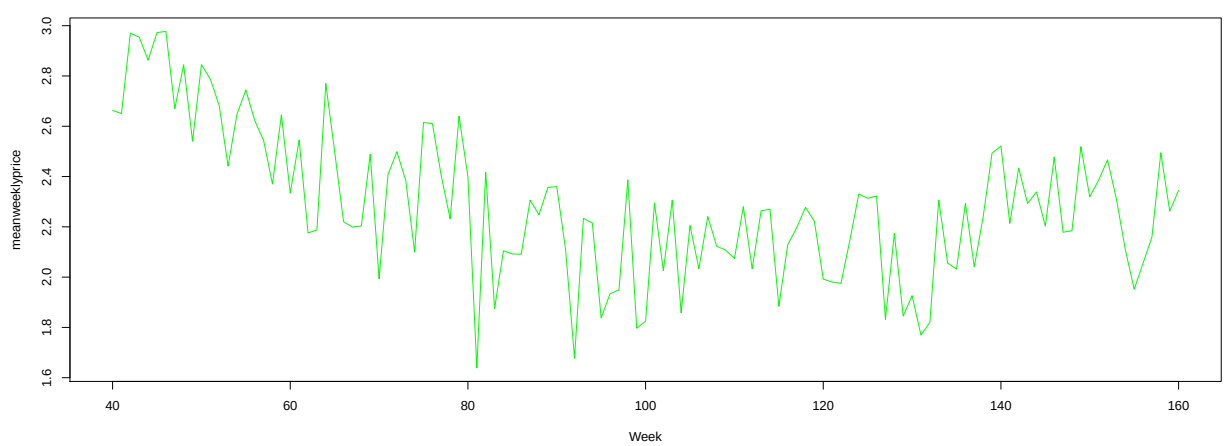
11.1 Method 1: using tapply()

```
xnames <- tapply(oj$price,oj$week,FUN=mean,na.rm=TRUE)
plot(tapply(oj$price,oj$week,FUN=mean,na.rm=TRUE),xaxt="n",pch=16,type="l",
     col="green",xlab = "Week",ylab="meanweeklyprice")
```



##Method 2: using aggreagate()

```
meanweeklyprice <- aggregate(oj[ ,6], list(oj$week), mean)
plot(meanweeklyprice$Group.1,meanweeklyprice$x,type="l",col="green",
     xlab = "Week",ylab="meanweeklyprice")
```



12 Extract the mean weekly price of orange juice sold each week according to each brand.

12.1 Method 1: using aggregate()

```
mean_weekly_brand_price <- aggregate(oj[ ,6], list(Brand=oj$brand,Week=oj$week), mean)
#mean_weekly_brand_price  #Uncomment this to view full data frame
head(mean_weekly_brand_price)
```

```
##         Brand Week        x
## 1   dominicks   40 1.590000
## 2 minute.maid   40 2.890000
## 3   tropicana   40 3.509315
## 4   dominicks   41 2.447015
## 5 minute.maid   41 1.990000
## 6   tropicana   41 3.514627
```

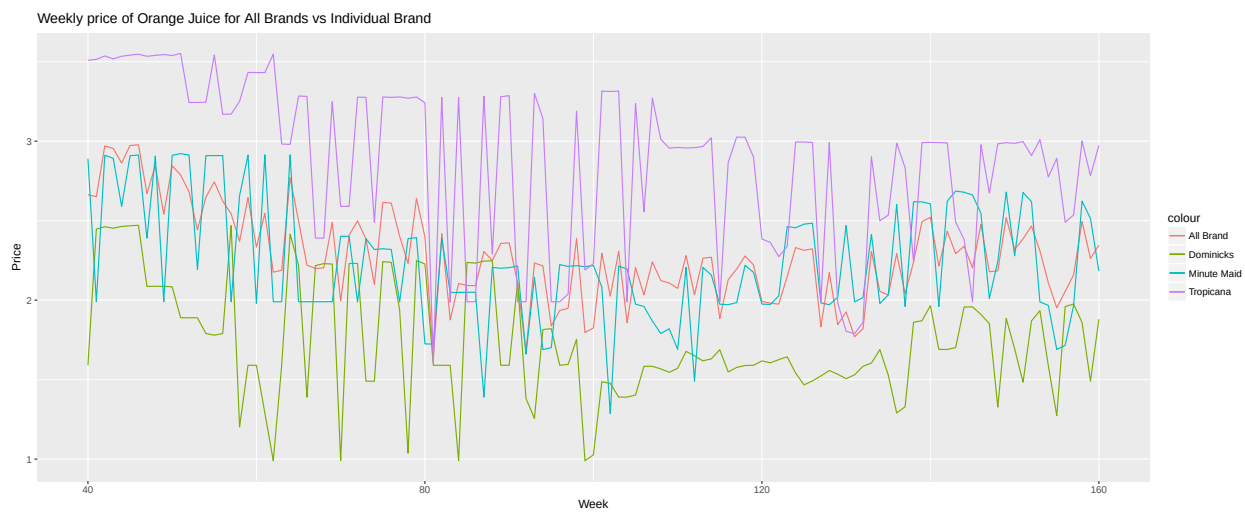
12.2 Method :2 using tapply()

```
weekly_brand_price<-tapply(oj$price,list(Week=oj$week,oj$brand),FUN=mean,na.rm=TRUE)
#weekly_brand_price  #Uncomment this to view full data frame
head(weekly_brand_price)
```

```
##
## Week dominicks minute.maid tropicana
##   40  1.590000    2.890000  3.509315
##   41  2.447015    1.990000  3.514627
##   42  2.461940    2.911343  3.536567
##   43  2.452597    2.892597  3.517792
##   44  2.463506    2.590000  3.534156
##   45  2.467838    2.908649  3.541486
```

13 Create a plot which compares the mean weekly price of orange juice for all brands versus each individual brand.

```
weekly_brand_price<-tapply(oj$price,list(Week=oj$week,oj$brand),FUN=mean,na.rm=TRUE)
allbrand <- tapply(oj$price,list(Week=oj$week),FUN=mean,na.rm=TRUE)
df <- cbind(allbrand,weekly_brand_price)
write.table(df, file = 'q13.csv',sep = ",",row.names = T)
library(readr)
q13 <- read_csv("E:/Smurfit/Semester 2/DM/R/Data-Mining/q13.csv",col_names = FALSE, skip = 1)
ggplot(q13, aes(q13$X1),pch=21) +
  geom_line(aes(y = q13$X2, colour = "All Brand")) +
  geom_line(aes(y = q13$X3, colour = "Dominicks")) +
  geom_line(aes(y = q13$X4, colour = "Minute Maid")) +
  geom_line(aes(y = q13$X5, colour = "Tropicana")) + xlab("Week") +ylab("Price")+
  ggtitle("Weekly price of Orange Juice for All Brands vs Individual Brand")
```



14 When there is an advertising campaign for orange juice does it impact on the number of units sold?

Yes, more number of units sold when there is an advertising campaign

```
adcap <- factor(oj$feat)
tapply(oj$logmove,adcap,FUN=mean,na.rm=TRUE)
```

```
##         0         1
##  8.862963 10.148049
```

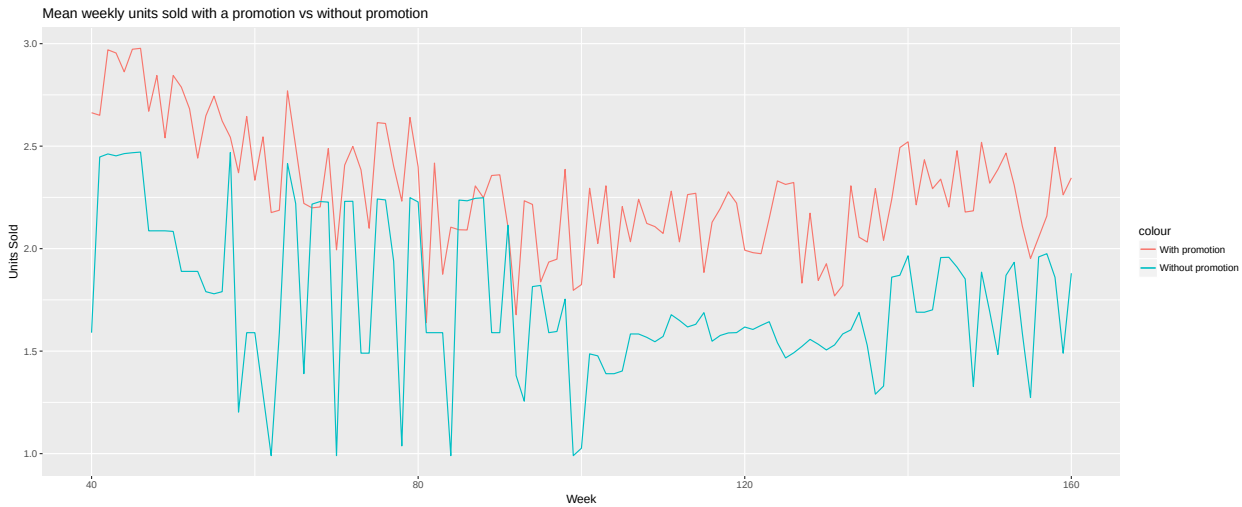
15 Can you create a line plot of the mean weekly units sold without a promotion overlayed with the mean weekly units sold with a promotion? What is interesting about this plot

With a promotion more number of units sold as compared to campaign without a promotion

```
unitsold<-tapply(oj$logmove,list(oj$week),FUN=mean,na.rm=TRUE)
head(unitsold)
```

```
##       40       41       42       43       44       45
## 8.598766 8.942384 8.298204 8.429358 8.741467 8.140365
```

```
sold <- tapply(oj$logmove,list(Week=oj$week,oj$feat),FUN=mean,na.rm=TRUE)
df <- cbind(sold)
write.table(df, file = 'q15.csv',sep = ",",row.names = T)
library(readr)
q15 <- read_csv("E:/Smurfit/Semester 2/DM/R/Data-Mining/q15.csv",col_names = FALSE, skip = 1)
ggplot(q15, aes(q13$X1,xlab="Week")) +
  geom_line(aes(y = q13$X2, colour = "With promotion")) +
  geom_line(aes(y = q13$X3, colour = "Without promotion"))+ xlab("Week") +ylab("Units Sold")+
  ggtitle("Mean weekly units sold with a promotion vs without promotion")
```



16 Descriptive Analytics

Consider the demographic and competitive variables1. Using descriptive analytics are there patterns you can observe that might suggest the potential for profiling individual stores or customers, which might then be used for marketing purposes?

16.1 Answer: Yes, there are patterns among demographics variables, which can be use for marketing of orange juice at given stores.

1. Data Cleaning: This help in identifying and removing duplicate records along with outlier, which might diverge business decision.
2. Descriptive Analytics: Using this one understand the behaviour, size, and range etc about data points.
3. Correlations: One can use highly correlated variables into account while forming a new marketing strategy. Such as, in given data, EDUC and HVAL150 are very much related to each other, this give us an insight that potential customer who are well educated also lives in an expensive house and vice verse. Similarly, HVAL150, INCOME and EDUC also correlated with each other.
4. Distribution: By plotting probability distribution of chart, one can figure out the probability of future business decision outcome.
5. Store Rating: Using statistical analysis, one can identify the best and worst store, also can rate each store depending on units sold by them. It is found that store 75 is poorest and store 62 is the wealthiest store among 83 stores.
6. Linear Models: By generating linear models of each variable, one can analysis the trend of data, which will help in forecasting future trend as well.

16.2 Overview

Actual data contains 28947 observations and 12 variables for demographic data.

```
d1 <-(oj[c(1,7:17)])
dim(d1)
```

```
## [1] 28947    12
```

On analysing the data, it is found that, there were only 83 unique records out of 28947 obsservations. Therefore, data should be cleaned before performing any further descriptive analytics.

```
d2 <- unique(oj[c(1,7:17)])
dim(d2)
```

```
## [1] 83 12
```

16.3 Descriptive Analytics Summary for demographic data

This analysis will help one to understand the data, how it is distrbuted and how many observations are there etc.

```
oj_subset <- unique(oj[c(1,7:17)])
oj_stats <- do.call(data.frame,
           list(Mean = apply(oj_subset, 2, mean),
                Std_Dev = apply(oj_subset, 2, sd),
                Median = apply(oj_subset, 2, median),
                Min = apply(oj_subset, 2, min),
                Max = apply(oj_subset, 2, max),
                N = apply(oj_subset, 2, length)))
oj_stats
```

```
##                Mean     Std_Dev      Median        Min         Max  N
## store    80.9277108 35.93448752 86.00000000 2.00000000 137.0000000 83
## AGE60     0.1729724  0.06221013  0.17065481 0.05805397   0.3073979 83
## EDUC      0.2257762  0.11114318  0.22939040 0.04955029   0.5283620 83
## ETHNIC    0.1546351  0.18782770  0.07465643 0.02424657   0.9956908 83
## INCOME   10.6176755  0.28344624 10.63532646 9.86708287  11.2361965 83
## HHLARGE   0.1156640  0.03035444  0.11122120 0.01350636   0.2163543 83
## WORKWOM   0.3591536  0.05280191  0.35563513 0.24446267   0.4723083 83
## HVAL150   0.3446528  0.24092599  0.34615385 0.00250941   0.9166995 83
## SSTRDIST  5.0969450  3.48872894  4.65068687 0.13209682  17.8559508 83
## SSTRVOL   1.2104260  0.53090387  1.11538461 0.40000000   2.5714286 83
## CPDIST5   2.1184928  0.73761928  1.96341236 0.77252973   4.1079018 83
## CPWVOL5   0.4393132  0.22067903  0.38322679 0.09456175   1.1433666 83
```

```
stats_data =apply(as.matrix(oj_subset),2,quantile)
cat("Statisical Summary of Orange Juice Data",fill=TRUE)
```

```
## Statisical Summary of Orange Juice Data
```

```
print(stats_data)
```

```
##      store      AGE60       EDUC     ETHNIC    INCOME    HHLARGE   WORKWOM
## 0%     2.0 0.05805397 0.04955029 0.02424657  9.867083 0.01350636 0.2444627
## 25%   53.5 0.12394915 0.14878867 0.04294112 10.468048 0.09804559 0.3129127
## 50%   86.0 0.17065481 0.22939040 0.07465643 10.635326 0.11122120 0.3556351
## 75%  110.5 0.21362903 0.28227241 0.17995974 10.795248 0.13345868 0.4021066
## 100% 137.0 0.30739786 0.52836201 0.99569076 11.236197 0.21635434 0.4723083
##         HVAL150   SSTRDIST   SSTRVOL   CPDIST5    CPWVOL5
## 0%   0.00250941  0.1320968 0.4000000 0.7725297 0.09456175
## 25%  0.12895177  2.8399843 0.7272727 1.6346475 0.28043820
## 50%  0.34615385  4.6506869 1.1153846 1.9634124 0.38322679
## 75%  0.51703001  6.5742868 1.5504808 2.5336557 0.55988695
## 100% 0.91669949 17.8559508 2.5714286 4.1079018 1.14336661
```

16.4 Scatter Plot Matrices in R

16.4.1 Analysis of Correlation among demographics variables.

The closer correlation coeff(r) is to +1 or -1, the more closely the two variables are related. If r is close to 0, it means there is no relationship between the variables. Below are few sample correlated variables.

16.4.1.1 1. Positive Correlations:

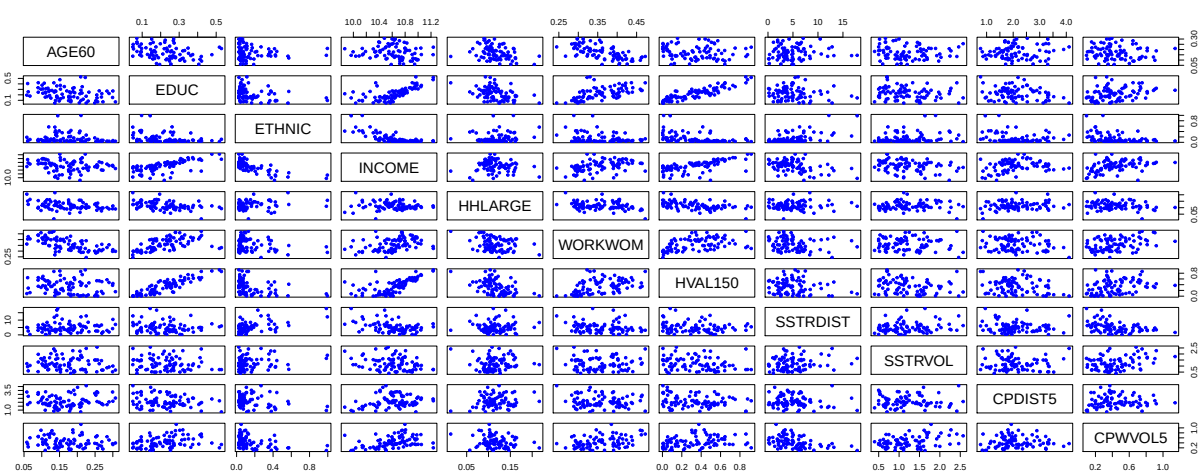
```
1. EDUC ~ CPWVOL5
2. WORKWOM ~ CPWVOL5
3. EDUC ~ WORKWOM
4. EDUC ~ HVAL150
5. WORKWOM ~ HVAL150
6. SSTRDIST ~ SSTRVOL
```

16.4.1.2 1. Negative Correlations:

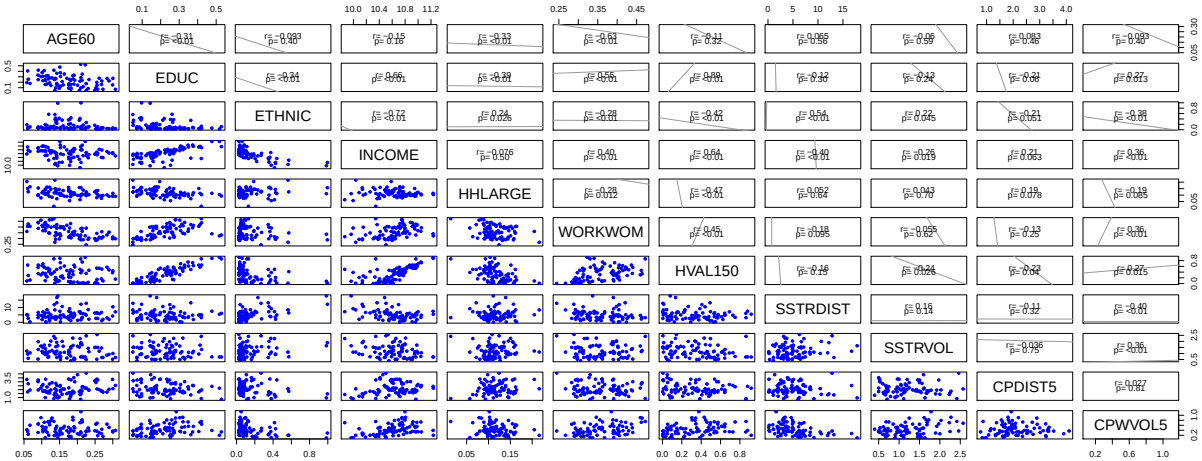
```
1. EDUC ~ CPWVOL5
2. WORKWOM ~ CPWVOL5
3. EDUC ~ WORKWOM
4. EDUC ~ HVAL150
5. WORKWOM ~ HVAL150
6. EUDC ~ AGE60
7. EDUC ~ ETHNIC
8. INCOME ~ ETHNIC
9. INCOME ~ HHLARGE
10.WORKWOM ~ HHLARGE
11. HVAL150 ~ SSTRDIST
12. SSTRVOL ~ CPDIST5
```

EDUC and HVAL150 are highly correlated with each other, having r = 0.89 and p <= 0.01. This signifies that educated people having house which has value more than $150,000.

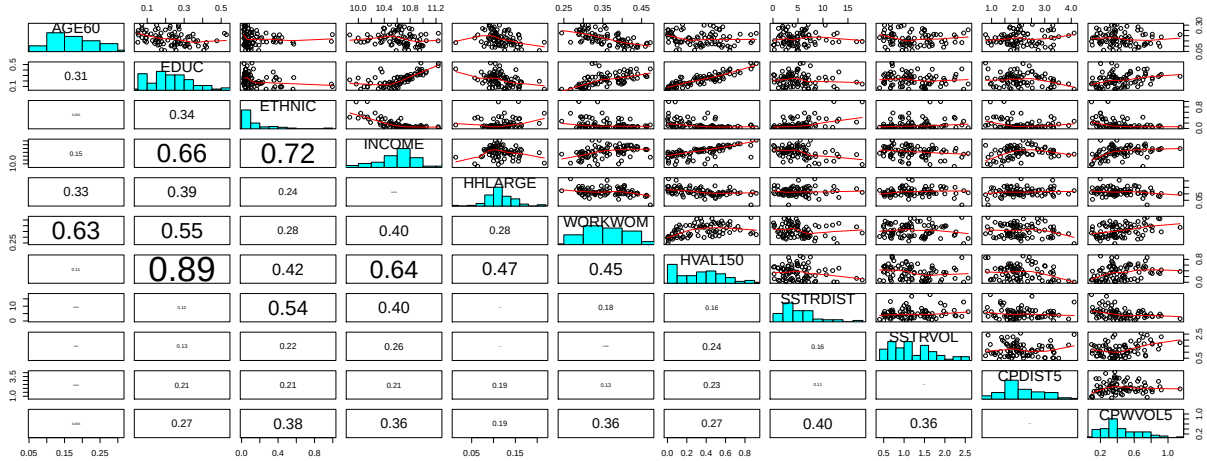
```r
#Scatter Plot to find correlation among demographics variables
pairs(oj_subset[,-1],col="blue",pch=20)
```



```r
#Reference: https://www.r-bloggers.com/scatter-plot-matrices-in-r/


#Scatter Plot with Correlation and P Value
panel.cor <- function(x, y, digits = 2, cex.cor=.8, ...)
{

  #Plotting Regression Line
  reg1 <- lm(x~y)
  abline(reg1,col = "gray60")

  usr <- par("usr"); on.exit(par(usr))
  par(usr = c(0, 1, 0, 1))
  # correlation coefficient
  r <- cor(x, y)
  txt <- format(c(r, 0.123456789), digits = digits)[1]
  txt <- paste("r= ", txt, sep = "")
  text(0.5, 0.6, txt)

  # p-value calculation
  p <- cor.test(x, y)$p.value
  txt2 <- format(c(p, 0.123456789), digits = digits)[1]
  txt2 <- paste("p= ", txt2, sep = "")
  if(p<0.01) txt2 <- paste("p= ", "<0.01", sep = "")
  text(0.5, 0.4, txt2)
}

pairs(oj_subset[,-1], upper.panel = panel.cor,col="blue",pch=20)
```




```r
library(s20x)
pairs20x(oj_subset[,-1])
```



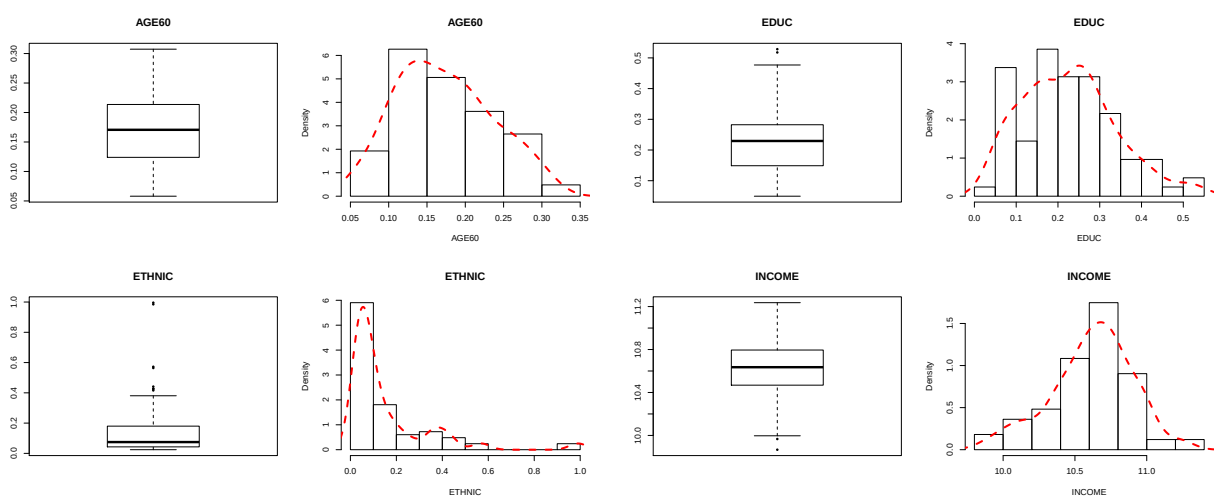
Boxplot and Histogram for Data Analysis

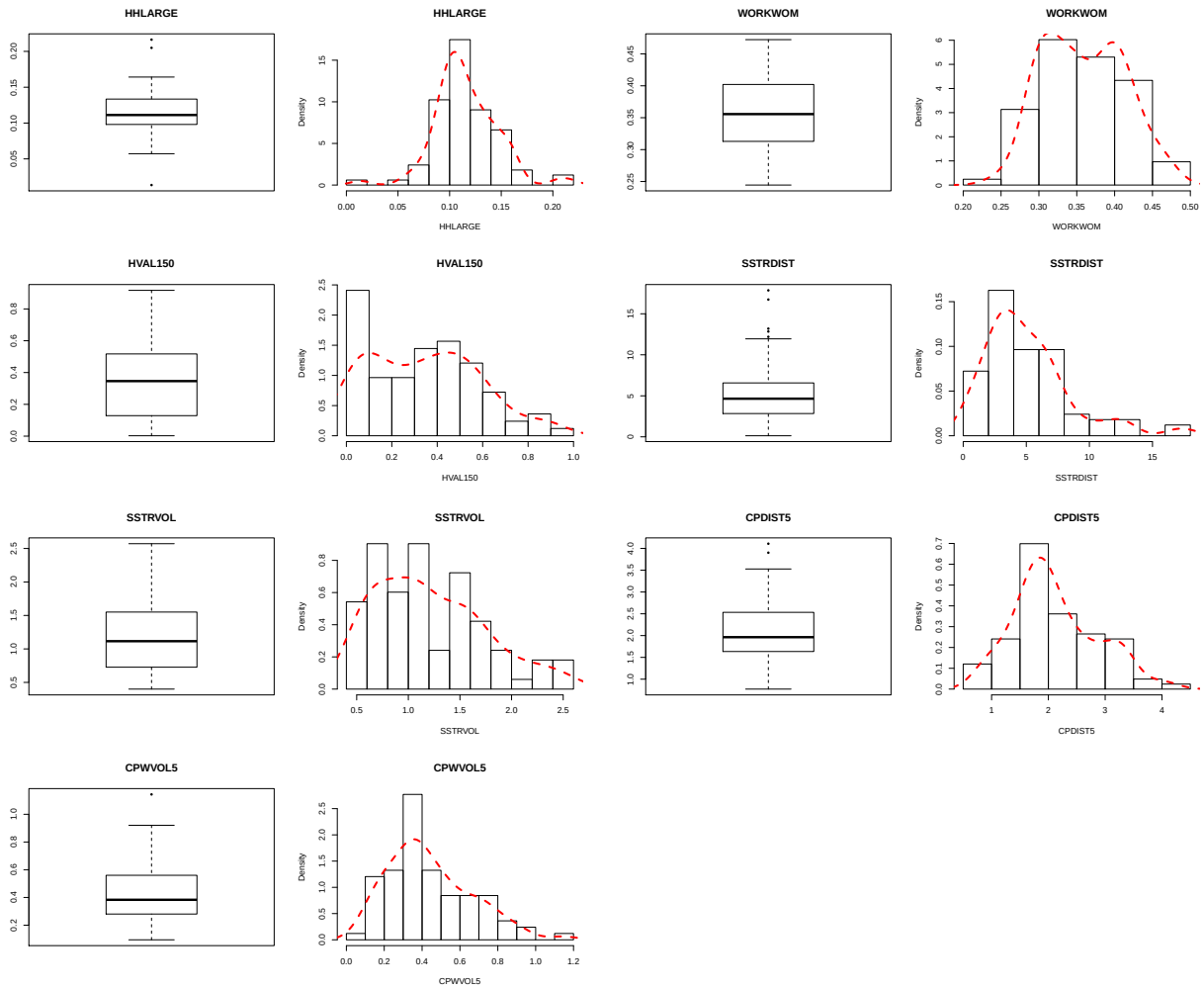
With the help of Boxplot, one is able to identify any outliers in the data, thereafter histogram plots help in checking the distribution of data, i.e whether data is normally distributed or not.

Also, the p-values of the variable greater than 0.05, suggests that there is no strong evidence to reject the null hypothesis that is orange juice variables data are normally distributed.

1. EDUC, ETHNNIC, INCOME, HHLARGE, SSTRDIST, CPDIST5 and CPWVOL5 has outliers

```
par(mfrow = c(2, 4))
varlist <- names(oj_subset)[2:12]
box_hist_plot<-function(index) {
  boxplot(oj_subset[,index], main=names(oj_subset[index]),pch=16)
  hist(oj_subset[,index],main=names(oj_subset[index]),xlab=names(oj_subset[index]),prob=T)
  lines(density(oj_subset[,index]),lty="dashed",lwd=2.5,col="red")
  }
lapply(2:12,FUN=box_hist_plot)
```





Wealthiest Store and Poorest Store Analysis

```
##Reference: Data Mining and Business Analytics with R Book by Johannes Ledolter
library(lattice)
par(mfrow = c(1, 2))
t21=tapply(oj$INCOME,oj$store,FUN=mean,na.rm=TRUE)
t21[t21==max(t21)]
```

```
##      62
## 11.2362
```

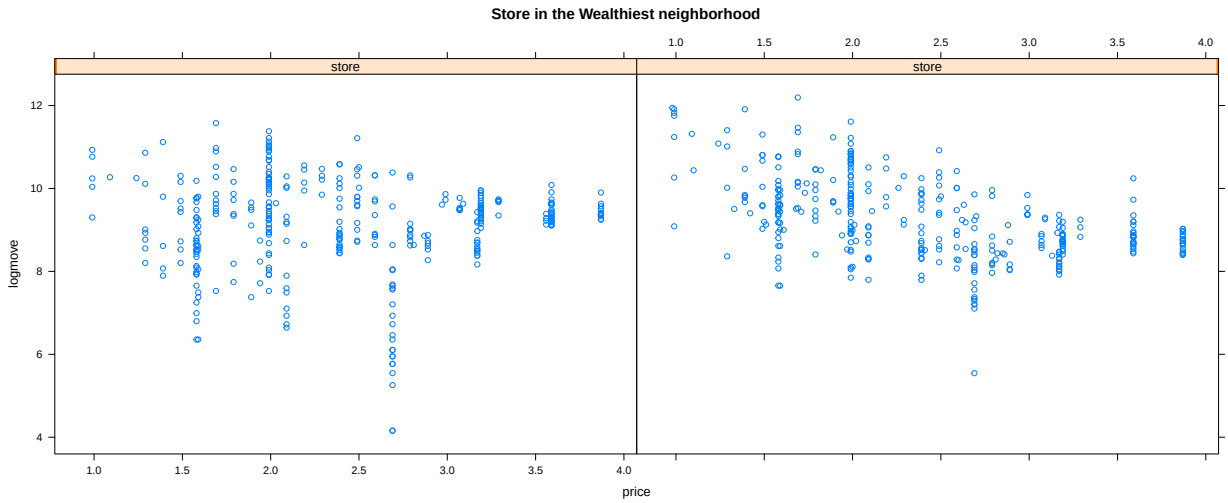
```
t21[t21==min(t21)]
```

```
##       75
## 9.867083
```

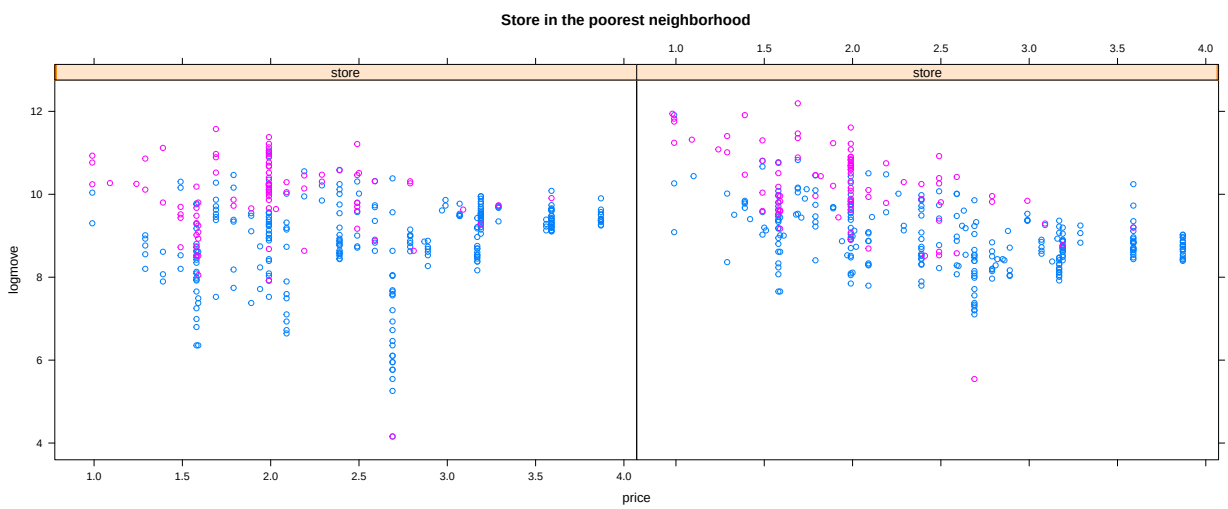
```
print(t21[t21==min(t21)])
```

```
##       75
## 9.867083
```

```
oj1=oj[oj$store == 62,]
oj2=oj[oj$store == 75,]
oj3=rbind(oj1,oj2)
xyplot(logmove~price|store,data=oj3,main="Store in the Wealthiest neighborhood")
```




```r
xyplot(logmove~price|store,groups=feat,data=oj3,main="Store in the poorest neighborhood")
```




```r
## store in the wealthiest neighborhood
mhigh=lm(logmove~price,data=oj1)
summary(mhigh)
```

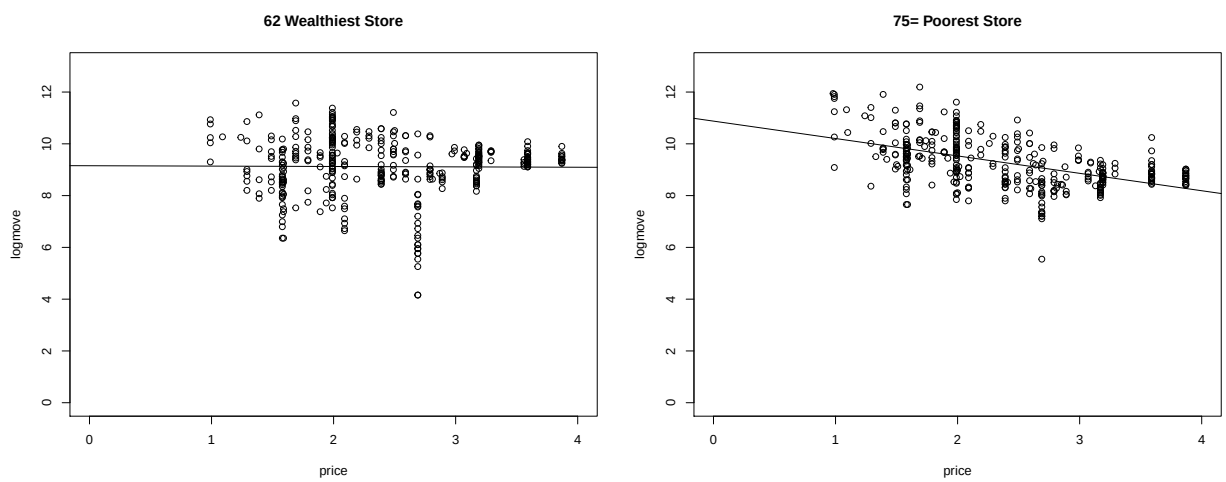
```
##
## Call:
## lm(formula = logmove ~ price, data = oj1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.9557 -0.4934  0.1815  0.6557  2.4454
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  9.15394    0.21112  43.359   <2e-16 ***
## price       -0.01461    0.08381  -0.174    0.862
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 1.142 on 349 degrees of freedom
## Multiple R-squared:  8.712e-05,  Adjusted R-squared:  -0.002778
## F-statistic: 0.03041 on 1 and 349 DF,  p-value: 0.8617
```

```r
plot(logmove~price,data=oj1,xlim=c(0,4),ylim=c(0,13),main="62 Wealthiest Store")
abline(mhigh)
## store in the poorest neighborhood
mlow=lm(logmove~price,data=oj2)
summary(mlow)
```

```
##
## Call:
## lm(formula = logmove ~ price, data = oj2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.5235 -0.5606  0.0392  0.5090  2.4523
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 10.87695    0.15184   71.63   <2e-16 ***
## price       -0.67222    0.06071  -11.07   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8383 on 352 degrees of freedom
## Multiple R-squared:  0.2584, Adjusted R-squared:  0.2563
## F-statistic: 122.6 on 1 and 352 DF,  p-value: < 2.2e-16
```

```r
plot(logmove~price,data=oj2,xlim=c(0,4),ylim=c(0,13),main="75= Poorest Store")
abline(mlow)
```

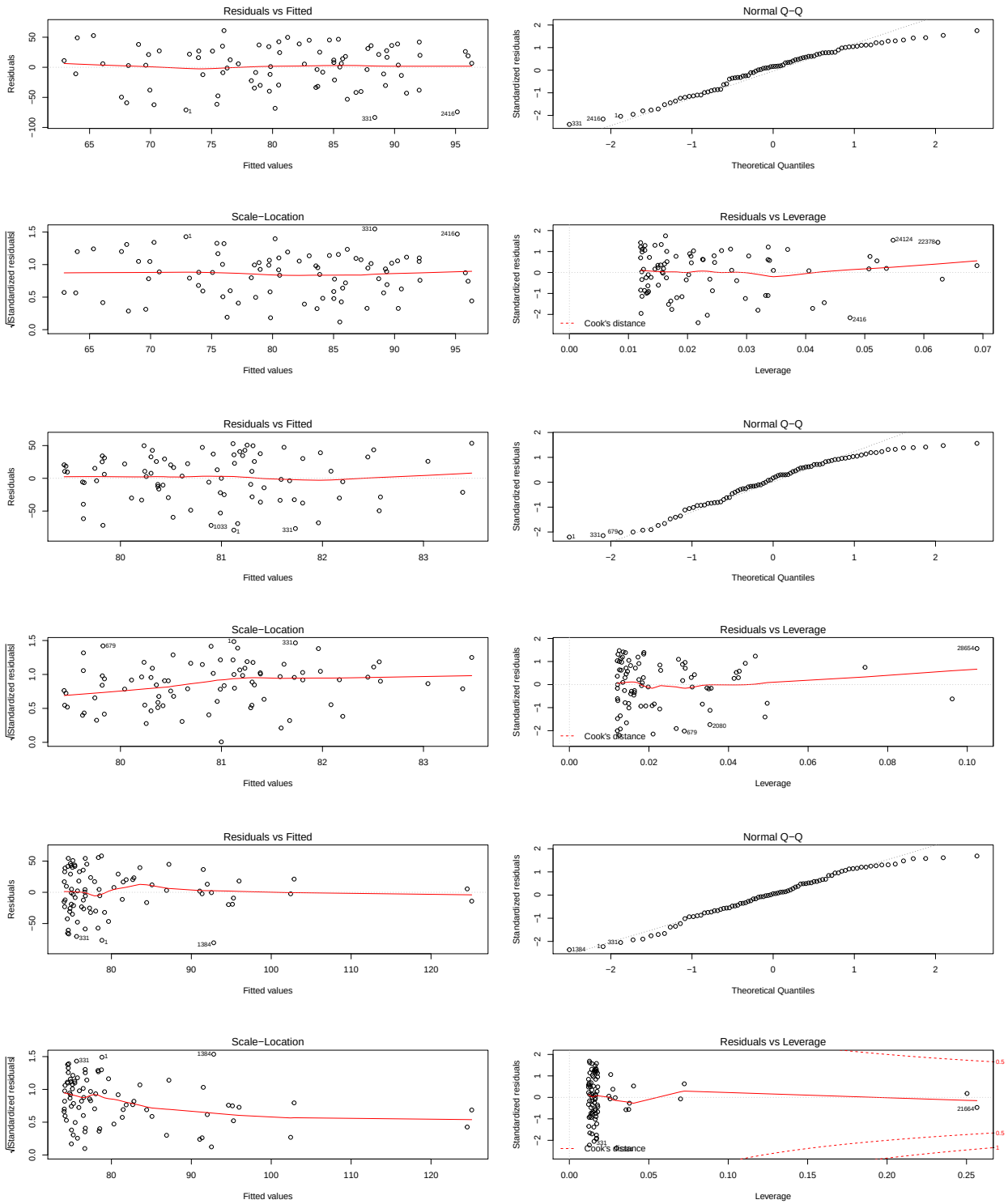


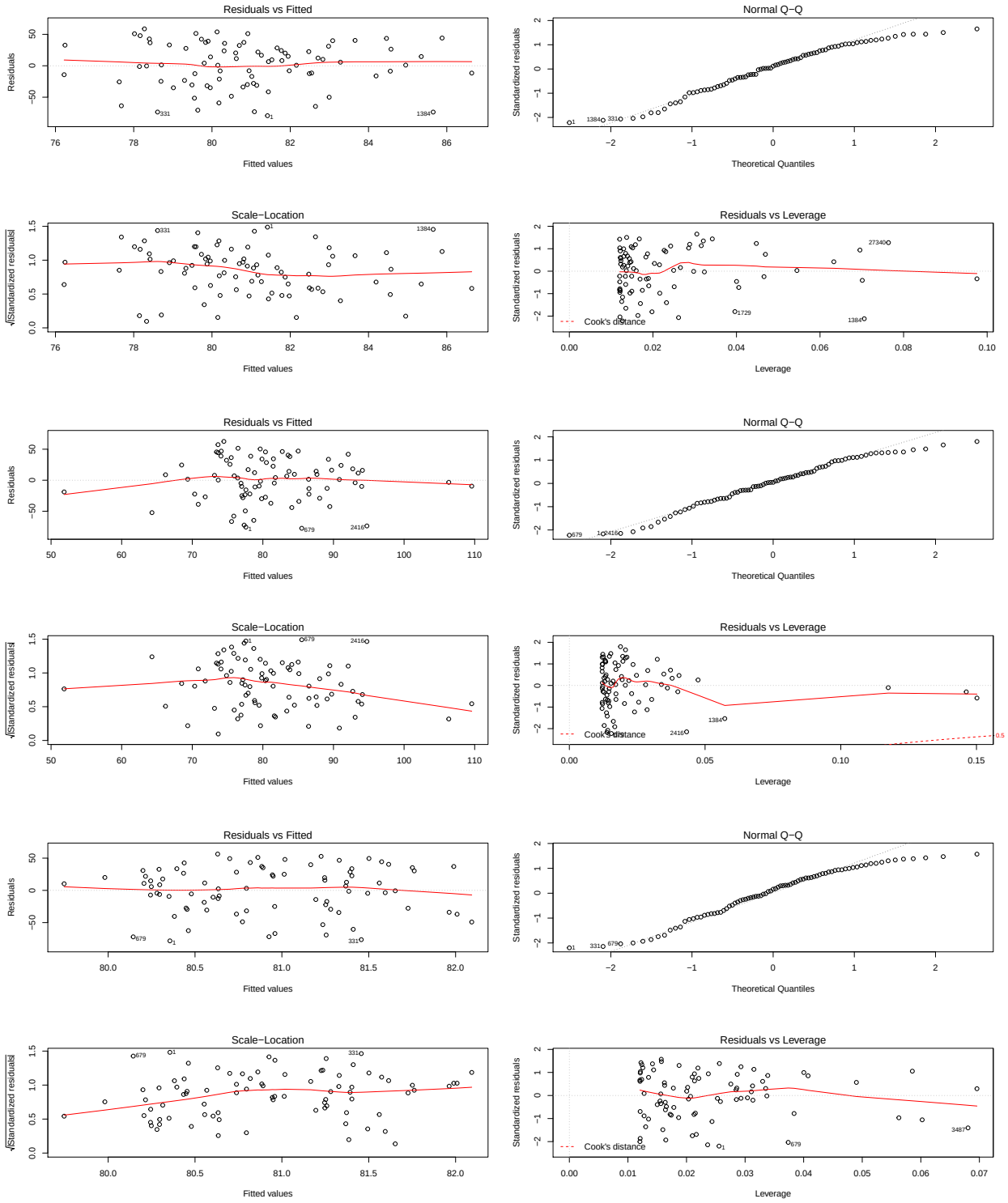
Linear Regression Models

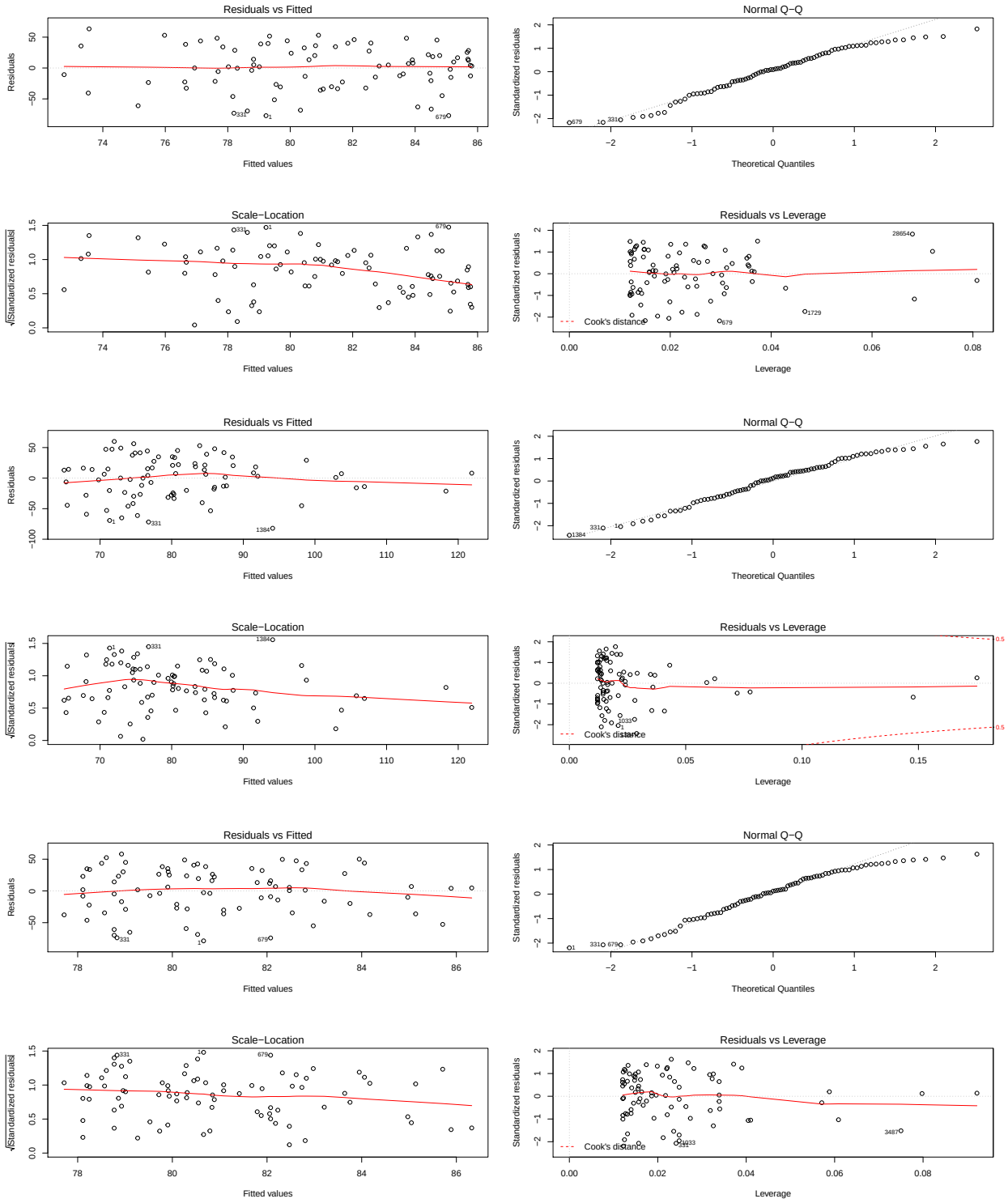
Reference UCLA.EDU

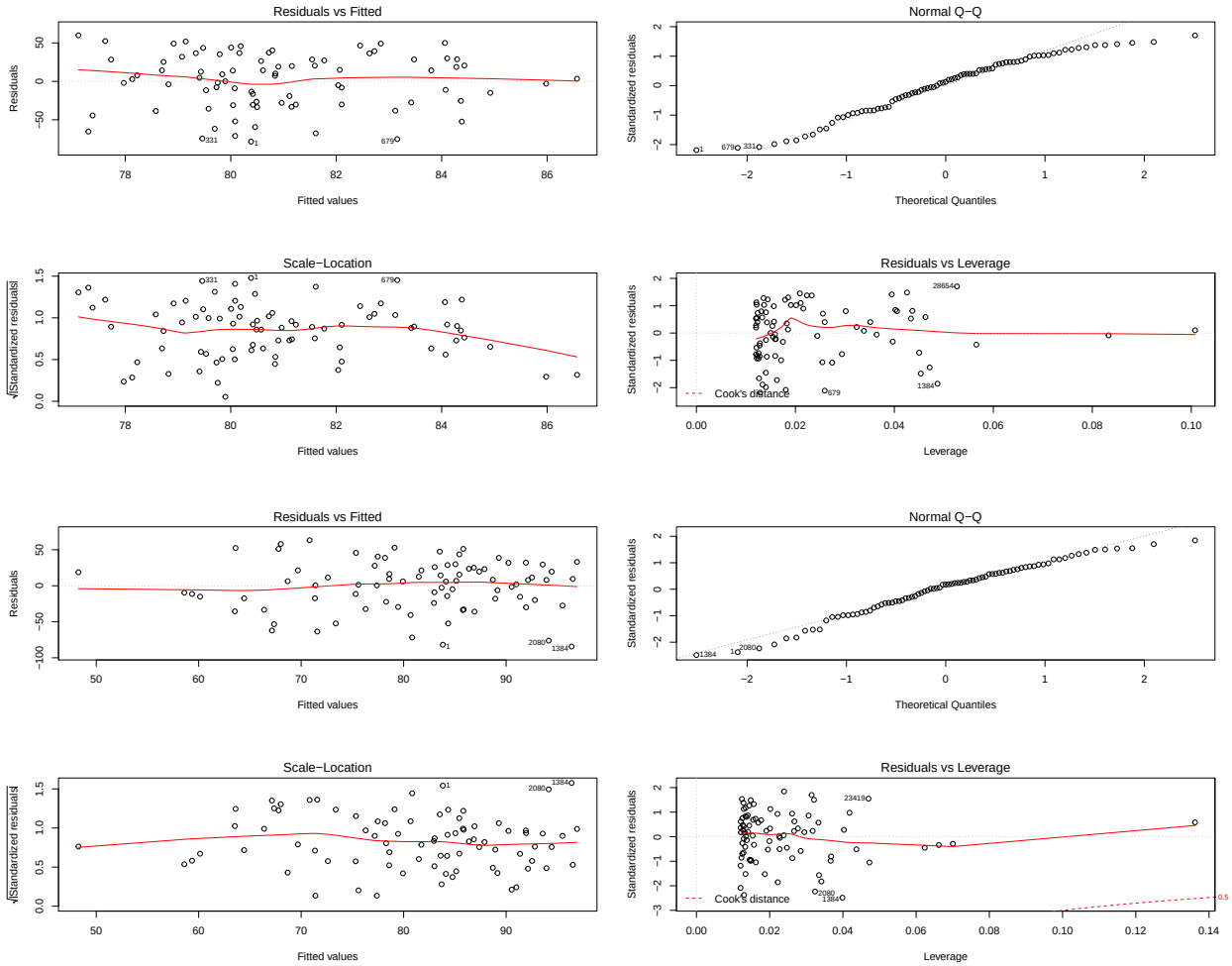
```r
#Reference: http://www.ats.ucla.edu/stat/r/pages/looping_strings.htm

varlist <- names(oj_subset)[2:12]
models <- lapply(varlist, function(x) {
```

```
    lm(substitute(store ~ i, list(i = as.name(x))), data = oj_subset)
})
par(mfrow = c(2, 2))
invisible(lapply(models, plot))
```




Residuals vs Fitted
Residuals
Fitted values
Normal Q−Q
Standardized residuals
Theoretical Quantiles
Scale−Location
√|Standardized residuals|
Fitted values
Residuals vs Leverage
Standardized residuals
Cook's distance
Leverage
Residuals vs Fitted
Residuals
Fitted values
Normal Q−Q
Standardized residuals
Theoretical Quantiles
Scale−Location
√|Standardized residuals|
Fitted values
Residuals vs Leverage
Standardized residuals
Cook's distance
Leverage
Residuals vs Fitted
Residuals
Fitted values
Normal Q−Q
Standardized residuals
Theoretical Quantiles
Scale−Location
√|Standardized residuals|
Fitted values
Residuals vs Leverage
Standardized residuals
Cook's distance
Leverage


Residuals vs Fitted
Residuals
Fitted values
Normal Q−Q
Standardized residuals
Theoretical Quantiles
Scale−Location
√|Standardized residuals|
Fitted values
Residuals vs Leverage
Standardized residuals
Cook's distance
Leverage
Residuals vs Fitted
Residuals
Fitted values
Normal Q−Q
Standardized residuals
Theoretical Quantiles
Scale−Location
√|Standardized residuals|
Fitted values
Residuals vs Leverage
Standardized residuals
Cook's distance
Leverage
Residuals vs Fitted
Residuals
Fitted values
Normal Q−Q
Standardized residuals
Theoretical Quantiles
Scale−Location
√|Standardized residuals|
Fitted values
Residuals vs Leverage
Standardized residuals
Cook's distance
Leverage


Residuals vs Fitted
Residuals
Fitted values
Normal Q−Q
Standardized residuals
Theoretical Quantiles
Scale−Location
√|Standardized residuals|
Fitted values
Residuals vs Leverage
Standardized residuals
Cook's distance
Leverage
Residuals vs Fitted
Residuals
Fitted values
Normal Q−Q
Standardized residuals
Theoretical Quantiles
Scale−Location
√|Standardized residuals|
Fitted values
Residuals vs Leverage
Standardized residuals
Cook's distance
Leverage