# Data Mining Assignment 1

Louis Carnec
15204934
April 5, 2016

## 1.

CODE:

```
oj <- read.csv("/Users/Carnec/Desktop/Business_Analytics
   /Data Mining/Assignment1/oj.csv")
attach(oj)
```

## 2.

There are 28947 records and 17 attributes.

CODE:

```
dim(oj)
```

## 3.

The mean of orange juice prices is 2.2825. The standard deviation is 0.6480. The range is from 0.52 to 3.87.

CODE:

```
summary(oj$price)
mean(oj$price)
# mean = 2.2825
sd(oj$price)
# sd = 0.6480
min(oj$price)
# min = 0.52
```

```
max(oj$price)
# max = 3.87
```

OUTPUT:

```
> summary(oj$price)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.520   1.790   2.170   2.282   2.730   3.870
> mean(oj$price)
[1] 2.282488
> sd(oj$price)
[1] 0.6480007
> min(oj$price)
[1] 0.52
> max(oj$price)
[1] 3.87
```

## 4.

The median of the log of number of units sold is 9.034.

CODE:

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

OUTPUT:

```
> median(oj$logmove)
[1] 9.03408
```

## 5.

The names of the three orange juice brands are Dominicks, Minute Maid and Tropicana.

CODE:

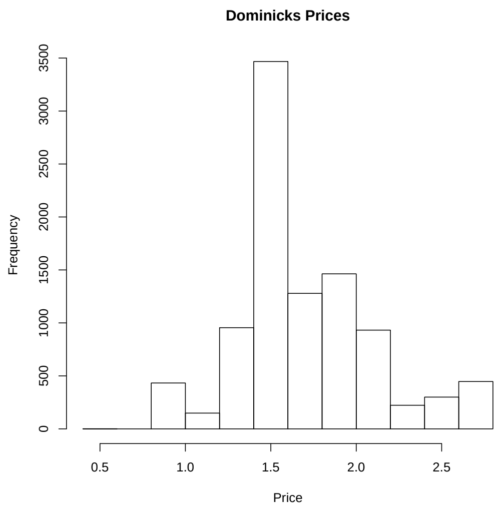
```
brands <- factor(oj\$brand)
table(brands)
```

OUTPUT:

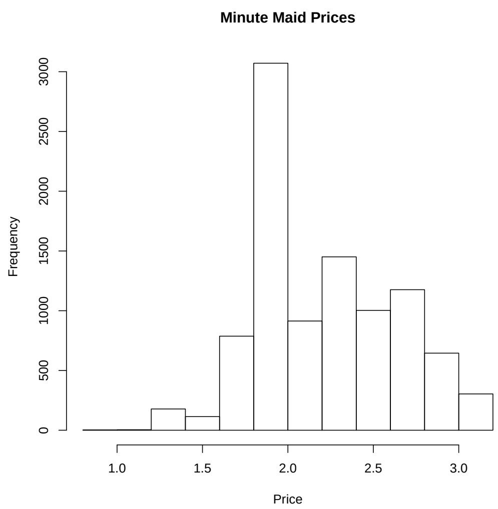
```
> brands <- factor(oj\$brand)
> table(brands)
```

```
brands
  dominicks minute.maid   tropicana
       9649        9649        9649
```

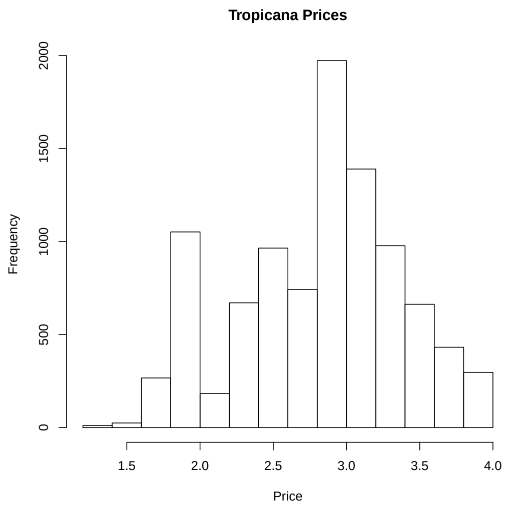
# 6.




*Plot:* Histogram of Dominicks prices

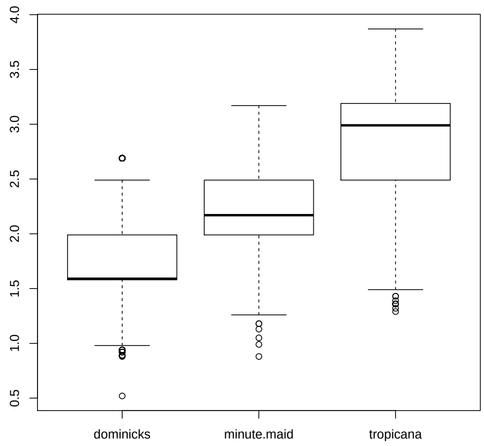



*Plot:* Histogram of Minute Maid prices




*Plot:* Histogram of Tropicana prices

# 7.



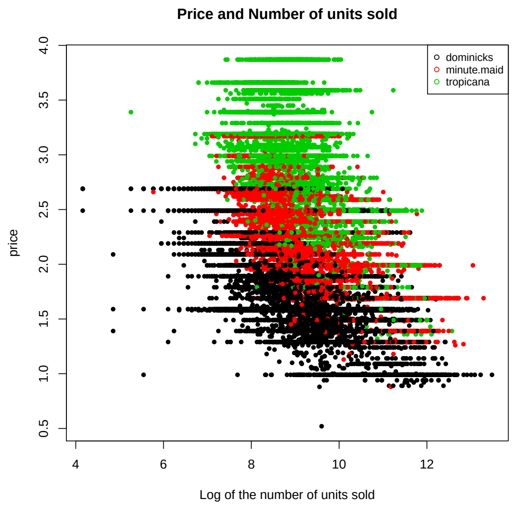

*Plot:* Boxplot of prices per brand

# 8.

The boxplot tells us that the most expensive orange juice is Tropicana. The second most expensive is Minute Maid. And the least expensive is Dominicks.

Tropicana is the most expensive orange juice. It's median price is almost as high as the maximum price of the second most expensive orange juice (Minute Maid). Quite a large proportion of Tropicana juices can be purchased for considerably less than the median price of Tropicana. Tropicana has the largest range (difference between max and min prices) of the three orange juice brands.
Interestingly there are several outliers below the minimum value for all three brands but only one on the upper end in the case of Dominicks. Thus, there is one store which is selling Dominicks at a much higher price than other stores.

# 9.




*Plot:* Scatterplot of logmove compared to price

# 10.

There is a negative relationship between price and log number of units sold for each of the brands.This relationship is more obvious in the case of Dominicks and Minute Maid, however this relationship is not as clear-cut in the case of Tropicana where there seems to be a non-linear relationship. Beyond a particular price (around 3.0) the log of units sold do not seem to decrease anymore as price increases compared to the relationship we can observe for lower Tropicana prices.

# 11.

CODE:

```
week.price <- tapply(oj$price,oj$week,FUN=mean,na.rm=
   TRUE)
meanweekprice <- as.data.frame(week.price)
print(weekprice)
pdf("timeseries.pdf")
plot.ts(meanweekprice,main="Mean price of orange juice
   sold each week",ylab="Mean Price",xlab = "Weeks")
graphics.off()
```


*Plot:* Time series of the mean price of orange juice

# 12.

CODE:

```
meanpriceperbrand <- t(as.data.frame(tapply(oj$price,
   INDEX=list(oj$brand,oj$week),FUN=mean,na.rm=TRUE)))
```

# 13.

CODE:

```
pdf("pricestimeseries.pdf")
plot.ts(meanpriceperbrand,plot.type=c("single"),ylab="
   Weekly Price",main="Weekly Prices",col=c("
   darkgoldenrod1","aquamarine2","deeppink4"))
lines(weekprice,col=c("coral2"))
legend(90,3.6,cex=0.6,legend=c("Mean","dominicks","
   tropicana","minute maid"),lty=c(1,1),col=c("coral2","
   darkgoldenrod1","deeppink4","aquamarine2"))
graphics.off()
```


*Plot:* Time series comparing mean orange juice price with mean individual prices

# 14.

Through ANOVA analysis, we find the impact of advertisement on units sold is statistically significant at the 0.001 significance level, thus we reject the null hypothesis that advertisement has no impact on the number of units sold.
CODE:

```
advertisement <- factor(oj$feat)
tapply(oj$logmove,advertisement,FUN=mean)
aov.out = aov(oj$logmove~advertisement)
summary(aov.out)
```

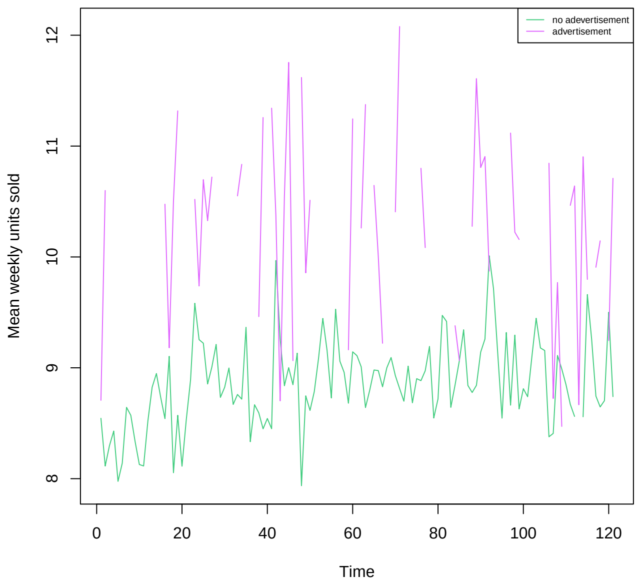
OUTPUT:

```
> summary(aov.out)
                 Df Sum Sq Mean Sq F value Pr(>F)
advertisement     1   8651    8651   11686 <2e-16 ***
Residuals     28945  21428       1
---
Signif. codes:  0 ?***? 0.001 ?**? 0.01 ?*? 0.05 ?.? 0.1
    ? ?
```

## 15.

The plot of mean weekly units sold with and without promotion is interesting because it shows that the number of units sold with promotion is highly significant (the pink line is consistently high than the green line for no advertisement). Units sold with promotion follow the general trend of the units sold without promotion, this tells us that advertisement does not counter the up and down movements in demand felt throughout the market.
Where the advertisement trend meets with the no advertisement trend, the promotional campaigns were failures (in terms of increasing units sold). This phenomenon becomes more regular towards the end of the time period.
CODE:

```
pdf("advetisementlineplot.pdf")
logmove_feat <- tapply(oj$logmove,INDEX=list(oj$week,
   oj$feat),FUN=mean,na.rm=TRUE)
plot.ts(logmove_feat,plot.type="single",ylab="Mean
   weekly units sold", col=c("seagreen3","mediumorchid1
   "))
legend("topright",cex=0.6,legend=c("no adevertisement","
   advertisement"),lty=c(1,1),col=c("seagreen3","
   mediumorchid1"))
graphics.off()
```




*Plot:* Mean weekly units sold with and without promotion

# 16.

## (a)

In order to uncover potential patterns which may suggest potential for profiling of individual stores and customers, I will produce a range of descriptive statistics and will create plots if and when appropriate.
A systematic output is produced for each variable. First, a summary is produced allowing us to quickly realise the shape of that variable's distribution. Then we use the tapply function to access the mean, median and standard deviation for the log number of units sold with respect to the variable we are interested in and the brand purchased. Lastly, we calculate correlation values for the mean of the variable we are interested in and the log number of units sold and repeat for each individual orange juice brand. Calculating correlation will allow us to infer the direction and strength of the relationship between the log number of units sold and the variable we are interested in.
Some of the variables in the data set were left out as they did not produce patterns which, at first sight, may be useful to marketers.

**CODE**:

```
attributes(oj)

#Summary of variables
summary(oj)
ojdescrive <-describe(oj)
stargazer(ojdescrive,summary=FALSE)

###Interested in number of unit sold!
summary(logmove)
tapply(logmove,brand,summary)
tapply(logmove,brand,mean)
tapply(logmove,brand,median)
tapply(logmove,brand,sd)
```

**OUTPUT**:

```
> summary(logmove)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  4.159   8.490   9.034   9.168   9.765  13.480
> tapply(logmove,brand,summary)
$dominicks
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  4.159   8.393   9.122   9.175   9.955  13.480

$minute.maid
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  5.768   8.476   9.026   9.217   9.829  13.290

$tropicana
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  5.257   8.566   8.987   9.111   9.534  12.570

> tapply(logmove,brand,median)
  dominicks minute.maid   tropicana
   9.121728    9.026418    8.987197
> tapply(logmove,brand,sd)
  dominicks minute.maid   tropicana
  1.1929370   0.9852867   0.8473800
```

**CODE**:

```
###AGE###
summary(AGE60)
oj$percAGE <- cut(AGE60,breaks=c
   (0,0.10,0.15,0.20,0.25,0.30,0.35,0.40,0.45,0.50))
summary(oj$percAGE)

#mean units sold to fraction of customers in age bracket
    over 60
tapply(logmove,list(oj$percAGE,brand),mean)

#mean units sold to customers in age bracket over 60 per
    brand
tapply(logmove,list(oj$AGE60,brand),mean)
pdf("boxplotlogmoveover60.pdf")
boxplot(tapply(logmove,list(oj$AGE60,brand),mean),ylab="
   logmove",main="Boxplot of logmove to >60 per brand")
graphics.off()

#median units sold to fraction of customers in age
   bracket over 60
tapply(logmove,list(percAGE,brand),median)

#standard deviation of units sold to fraction of
   customers in age bracket over 60
tapply(logmove,list(percAGE,brand),sd)

#Correlation between mean logmove per store and mean
   AGE60 per store
cor(tapply(logmove,store,FUN=mean,na.rm=TRUE),tapply(
   AGE60,store,FUN=mean,na.rm=TRUE))

#Correlation between mean logmove per store and mean
   AGE60 per store for dominicks
cor(tapply(logmove[brand=="dominicks"],store[brand=="
   dominicks"],FUN=mean,na.rm=TRUE),tapply(AGE60[brand
   =="dominicks"],store[brand=="dominicks"],FUN=mean,na.
   rm=TRUE))

#Correlation between mean logmove per store and mean
   AGE60 per store for minute.maid
```

```
cor(tapply(logmove[brand=="minute.maid"],store[brand=="
   minute.maid"],FUN=mean,na.rm=TRUE),tapply(AGE60[brand
   =="minute.maid"],store[brand=="minute.maid"],FUN=mean
   ,na.rm=TRUE))

#Correlation between mean logmove per store and mean
   AGE60 per store for tropicana
cor(tapply(logmove[brand=="tropicana"],store[brand=="
   tropicana"],FUN=mean,na.rm=TRUE),tapply(AGE60[brand
   =="tropicana"],store[brand=="tropicana"],FUN=mean,na.
   rm=TRUE))
```

**OUTPUT**:

```
> summary(AGE60)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.05805 0.12210 0.17070 0.17310 0.21390 0.30740
> summary(oj$percAGE)
   (0,0.1] (0.1,0.15] (0.15,0.2] (0.2,0.25] (0.25,0.3]
      (0.3,0.35] (0.35,0.4] (0.4,0.45] (0.45,0.5]
      2733       9105       7347       5193       3864
                 705          0          0          0
> #Correlation between mean logmove per store and mean
   AGE60 per store
> cor(tapply(logmove,store,FUN=mean,na.rm=TRUE),tapply(
   AGE60,store,FUN=mean,na.rm=TRUE))
[1] 0.3074491
>
> #Correlation between mean logmove per store and mean
   AGE60 per store for dominicks
> cor(tapply(logmove[brand=="dominicks"],store[brand=="
   dominicks"],FUN=mean,na.rm=TRUE),tapply(AGE60[brand
   =="dominicks"],store[brand=="dominicks"],FUN=mean,na.
   rm=TRUE))
[1] 0.2074369
>
> #Correlation between mean logmove per store and mean
   AGE60 per store for minute.maid
> cor(tapply(logmove[brand=="minute.maid"],store[brand
   =="minute.maid"],FUN=mean,na.rm=TRUE),tapply(AGE60[
   brand=="minute.maid"],store[brand=="minute.maid"],FUN
   =mean,na.rm=TRUE))
```

```
[1] 0.2711299
>
> #Correlation between mean logmove per store and mean
   AGE60 per store for tropicana
> cor(tapply(logmove[brand=="tropicana"],store[brand=="
   tropicana"],FUN=mean,na.rm=TRUE),tapply(AGE60[brand
   =="tropicana"],store[brand=="tropicana"],FUN=mean,na.
   rm=TRUE))
[1] 0.2866523
```

**CODE**:

```
###EDUC###
summary(EDUC)
oj$percEDUC <- cut(EDUC,breaks=c
   (0,0.10,0.15,0.20,0.25,0.30,0.35,0.40,0.45,0.50,0.55,0.60)
   )
summary(percEDUC)

#mean units sold to fraction of customers with a college
    degree
tapply(logmove, list(percEDUC,brand),mean)

#median units sold to fraction of customers with a
   college degree
tapply(logmove,list(percEDUC,brand),median)

#standard deviation of units sold to fraction of
   customers with a college degree
tapply(logmove,list(percEDUC,brand),sd)

#Correlation between mean logmove per store and mean
   EDUC per store
cor(tapply(logmove,store,FUN=mean,na.rm=TRUE),tapply(
   EDUC,store,FUN=mean,na.rm=TRUE))

#Correlation between mean logmove per store and mean
   EDUC per store for dominicks
cor(tapply(logmove[brand=="dominicks"],store[brand=="
   dominicks"],FUN=mean,na.rm=TRUE),tapply(EDUC[brand=="
   dominicks"],store[brand=="dominicks"],FUN=mean,na.rm=
```

```
   TRUE))

#Correlation between mean logmove per store and mean
   EDUC per store for Minute MAid
cor(tapply(logmove[brand=="minute.maid"],store[brand=="
   minute.maid"],FUN=mean,na.rm=TRUE),tapply(EDUC[brand
   =="minute.maid"],store[brand=="minute.maid"],FUN=mean
   ,na.rm=TRUE))

#Correlation between mean logmove per store and mean
   EDUC per store for tropicana
cor(tapply(logmove[brand=="tropicana"],store[brand=="
   tropicana"],FUN=mean,na.rm=TRUE),tapply(EDUC[brand=="
   tropicana"],store[brand=="tropicana"],FUN=mean,na.rm=
   TRUE))
```

**OUTPUT**:

```
> summary(EDUC)
    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.04955 0.14600 0.22940 0.22520 0.28440 0.52840

> #Correlation between mean logmove per store and mean
   EDUC per store
> cor(tapply(logmove,store,FUN=mean,na.rm=TRUE),tapply(
   EDUC,store,FUN=mean,na.rm=TRUE))
[1] 0.02565469
>
> #Correlation between mean logmove per store and mean
   EDUC per store for dominicks
> cor(tapply(logmove[brand=="dominicks"],store[brand=="
   dominicks"],FUN=mean,na.rm=TRUE),tapply(EDUC[brand=="
   dominicks"],store[brand=="dominicks"],FUN=mean,na.rm=
   TRUE))
[1] -0.4966506
>
> #Correlation between mean logmove per store and mean
   EDUC per store for Minute MAid
> cor(tapply(logmove[brand=="minute.maid"],store[brand
   =="minute.maid"],FUN=mean,na.rm=TRUE),tapply(EDUC[
   brand=="minute.maid"],store[brand=="minute.maid"],FUN
```

```
   =mean,na.rm=TRUE))
[1] 0.1108335
>
> #Correlation between mean logmove per store and mean
   EDUC per store for tropicana
> cor(tapply(logmove[brand=="tropicana"],store[brand=="
   tropicana"],FUN=mean,na.rm=TRUE),tapply(EDUC[brand=="
   tropicana"],store[brand=="tropicana"],FUN=mean,na.rm=
   TRUE))
[1] 0.4350975
```

**CODE**:

```
###ETHNIC###

summary(ETHNIC)
oj$percETHNIC <- cut(ETHNIC,breaks=c
   (0,0.10,0.15,0.20,0.25,0.30,0.35,0.40,0.45,0.50,0.55,0.60,0.65,0.7
   )
summary(percETHNIC) #OUTLIER

pdf("ETHNIChist.pdf")
hist(tapply(ETHNIC,store,mean),xlab="percent of
   population that is black or hispanic")
graphics.off()

#mean units sold to ethnic fraction of customers
tapply(logmove, list(percETHNIC,brand),mean)

#median units sold to ethnic fraction of customers
tapply(logmove, list(percETHNIC,brand),mean)

#standard deviation of units sold to ethnic fraction of
   customers
tapply(logmove,list(percETHNIC,brand),sd)

#Correlation between mean logmove per store and mean
   ETHNIC per store
cor(tapply(logmove,store,FUN=mean,na.rm=TRUE),tapply(
   ETHNIC,store,FUN=mean,na.rm=TRUE))

#Correlation between mean logmove per store and mean
```

```
   ETHNIC per store for dominicks
cor(tapply(logmove[brand=="dominicks"],store[brand=="
   dominicks"],FUN=mean,na.rm=TRUE),tapply(ETHNIC[brand
   =="dominicks"],store[brand=="dominicks"],FUN=mean,na.
   rm=TRUE))

#Correlation between mean logmove per store and mean
   ETHNIC per store for Minute MAid
cor(tapply(logmove[brand=="minute.maid"],store[brand=="
   minute.maid"],FUN=mean,na.rm=TRUE),tapply(ETHNIC[
   brand=="minute.maid"],store[brand=="minute.maid"],FUN
   =mean,na.rm=TRUE))

#Correlation between mean logmove per store and mean
   ETHNIC per store for tropicana
cor(tapply(logmove[brand=="tropicana"],store[brand=="
   tropicana"],FUN=mean,na.rm=TRUE),tapply(ETHNIC[brand
   =="tropicana"],store[brand=="tropicana"],FUN=mean,na.
   rm=TRUE))
```

**OUTPUT**:

```
> summary(ETHNIC)
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
0.02425 0.04191 0.07466 0.15560 0.18780 0.99570
> #Correlation between mean logmove per store and mean
   ETHNIC per store
> cor(tapply(logmove,store,FUN=mean,na.rm=TRUE),tapply(
   ETHNIC,store,FUN=mean,na.rm=TRUE))
[1] 0.1967332
>
> #Correlation between mean logmove per store and mean
   ETHNIC per store for dominicks
> cor(tapply(logmove[brand=="dominicks"],store[brand=="
   dominicks"],FUN=mean,na.rm=TRUE),tapply(ETHNIC[brand
   =="dominicks"],store[brand=="dominicks"],FUN=mean,na.
   rm=TRUE))
[1] 0.4310104
>
> #Correlation between mean logmove per store and mean
   ETHNIC per store for Minute MAid
> cor(tapply(logmove[brand=="minute.maid"],store[brand
```

```
   =="minute.maid"],FUN=mean,na.rm=TRUE),tapply(ETHNIC[
   brand=="minute.maid"],store[brand=="minute.maid"],FUN
   =mean,na.rm=TRUE))
[1] 0.1918715
>
> #Correlation between mean logmove per store and mean
   ETHNIC per store for tropicana
> cor(tapply(logmove[brand=="tropicana"],store[brand=="
   tropicana"],FUN=mean,na.rm=TRUE),tapply(ETHNIC[brand
   =="tropicana"],store[brand=="tropicana"],FUN=mean,na.
   rm=TRUE))
[1] -0.1029473
```

**CODE**:

```
###INCOME###
summary(INCOME)
hist(INCOME)
hist(tapply(INCOME,store,mean))

#mean log units sold for different customer incomes
tapply(logmove, list(INCOME,brand),mean)

#median log units sold for different customer incomes
tapply(logmove, list(INCOME,brand),median)

#standard deviation of log units sold for different
   customer incomes
tapply(logmove, list(INCOME,brand),sd)

#Correlation between mean logmove per store and mean
   INCOME per store
cor(tapply(logmove,store,FUN=mean,na.rm=TRUE),tapply(
   INCOME,store,FUN=mean,na.rm=TRUE))

#Correlation between mean logmove per store and mean
   INCOME per store for dominicks
cor(tapply(logmove[brand=="dominicks"],store[brand=="
   dominicks"],FUN=mean,na.rm=TRUE),tapply(INCOME[brand
   =="dominicks"],store[brand=="dominicks"],FUN=mean,na.
   rm=TRUE))
```

```
#Correlation between mean logmove per store and mean
   INCOME per store for Minute MAid
cor(tapply(logmove[brand=="minute.maid"],store[brand=="
   minute.maid"],FUN=mean,na.rm=TRUE),tapply(INCOME[
   brand=="minute.maid"],store[brand=="minute.maid"],FUN
   =mean,na.rm=TRUE))

#Correlation between mean logmove per store and mean
   INCOME per store for tropicana
cor(tapply(logmove[brand=="tropicana"],store[brand=="
   tropicana"],FUN=mean,na.rm=TRUE),tapply(INCOME[brand
   =="tropicana"],store[brand=="tropicana"],FUN=mean,na.
   rm=TRUE))
```

**OUTPUT**:

```
> summary(INCOME)
    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  9.867  10.460  10.640  10.620  10.800  11.240
  > #Correlation between mean logmove per store and mean
      INCOME per store
> cor(tapply(logmove,store,FUN=mean,na.rm=TRUE),tapply(
   INCOME,store,FUN=mean,na.rm=TRUE))
[1] -0.1362701
>
> #Correlation between mean logmove per store and mean
   INCOME per store for dominicks
> cor(tapply(logmove[brand=="dominicks"],store[brand=="
   dominicks"],FUN=mean,na.rm=TRUE),tapply(INCOME[brand
   =="dominicks"],store[brand=="dominicks"],FUN=mean,na.
   rm=TRUE))
[1] -0.4788677
>
> #Correlation between mean logmove per store and mean
   INCOME per store for Minute MAid
> cor(tapply(logmove[brand=="minute.maid"],store[brand
   =="minute.maid"],FUN=mean,na.rm=TRUE),tapply(INCOME[
   brand=="minute.maid"],store[brand=="minute.maid"],FUN
   =mean,na.rm=TRUE))
[1] -0.06028755
>
> #Correlation between mean logmove per store and mean
```

```
   INCOME per store for tropicana
> cor(tapply(logmove[brand=="tropicana"],store[brand=="
   tropicana"],FUN=mean,na.rm=TRUE),tapply(INCOME[brand
   =="tropicana"],store[brand=="tropicana"],FUN=mean,na.
   rm=TRUE))
[1] 0.1869724
```

**CODE**:

```
###HHLARGE###
summary(HHLARGE)
oj$percHHLARGE <- cut(HHLARGE,breaks=c
   (0.05,0.1,0.15,0.20,0.25))
summary(percHHLARGE)
hist(tapply(HHLARGE,store,mean))

#mean log units sold to percentage of households with
   more than 5 people
tapply(logmove,list(percHHLARGE,brand),mean)

#median log units sold to percentage of households with
   more than 5 people
tapply(logmove,list(percHHLARGE,brand),median)

#sd of log units sold to percentage of households with
   more than 5 people
tapply(logmove,list(percHHLARGE,brand),sd)

#Correlation between mean logmove per store and mean
   HHLARGE per store
cor(tapply(logmove,store,FUN=mean,na.rm=TRUE),tapply(
   HHLARGE,store,FUN=mean,na.rm=TRUE))

#Correlation between mean logmove per store and mean
   HHLARGE per store for dominicks
cor(tapply(logmove[brand=="dominicks"],store[brand=="
   dominicks"],FUN=mean,na.rm=TRUE),tapply(HHLARGE[brand
   =="dominicks"],store[brand=="dominicks"],FUN=mean,na.
   rm=TRUE))

#Correlation between mean logmove per store and mean
   HHLARGE per store for Minute MAid
```

```
cor(tapply(logmove[brand=="minute.maid"],store[brand=="
   minute.maid"],FUN=mean,na.rm=TRUE),tapply(HHLARGE[
   brand=="minute.maid"],store[brand=="minute.maid"],FUN
   =mean,na.rm=TRUE))

#Correlation between mean logmove per store and mean
   HHLARGE per store for tropicana
cor(tapply(logmove[brand=="tropicana"],store[brand=="
   tropicana"],FUN=mean,na.rm=TRUE),tapply(HHLARGE[brand
   =="tropicana"],store[brand=="tropicana"],FUN=mean,na.
   rm=TRUE))
```

**OUTPUT**:

```
> summary(HHLARGE)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.01351 0.09794 0.11120 0.11560 0.13520 0.21640
> summary(percHHLARGE)
(0.05,0.1] (0.1,0.15] (0.15,0.2] (0.2,0.25]       NA's
      7662      17157       3075        702        351
#Correlation between mean logmove per store and mean
   HHLARGE per store
cor(tapply(logmove,store,FUN=mean,na.rm=TRUE),tapply(
   HHLARGE,store,FUN=mean,na.rm=TRUE))

#Correlation between mean logmove per store and mean
   HHLARGE per store for dominicks
cor(tapply(logmove[brand=="dominicks"],store[brand=="
   dominicks"],FUN=mean,na.rm=TRUE),tapply(HHLARGE[brand
   =="dominicks"],store[brand=="dominicks"],FUN=mean,na.
   rm=TRUE))

#Correlation between mean logmove per store and mean
   HHLARGE per store for Minute MAid
cor(tapply(logmove[brand=="minute.maid"],store[brand=="
   minute.maid"],FUN=mean,na.rm=TRUE),tapply(HHLARGE[
   brand=="minute.maid"],store[brand=="minute.maid"],FUN
   =mean,na.rm=TRUE))

#Correlation between mean logmove per store and mean
   HHLARGE per store for tropicana
cor(tapply(logmove[brand=="tropicana"],store[brand=="
```

```
   tropicana"],FUN=mean,na.rm=TRUE),tapply(HHLARGE[brand
   =="tropicana"],store[brand=="tropicana"],FUN=mean,na.
   rm=TRUE))
```

**CODE**:

```
###WORKWOM###

summary(WORKWOM)
oj$percWORKWOM <- cut(WORKWOM,breaks=c
   (0,0.05,0.1,0.15,0.20,0.25,0.30,0.35,0.40,0.45,0.50))
summary(percWORKWOM)
hist(tapply(WORKWOM,store,mean))

#mean log units sold to percentage of women in full-time
    jobs
tapply(logmove,list(percWORKWOM,brand),mean)

#median log units sold to percentage of women in full-
   time jobs
tapply(logmove,list(percWORKWOM,brand),median)

#sd of log units sold to percentage of women in full-
   time jobs
tapply(logmove,list(percWORKWOM,brand),sd)

#Correlation between mean logmove per store and mean
   WORKWOM per store
cor(tapply(logmove,store,FUN=mean,na.rm=TRUE),tapply(
   WORKWOM,store,FUN=mean,na.rm=TRUE))

#Correlation between mean logmove per store and mean
   WORKWOM per store for dominicks
cor(tapply(logmove[brand=="dominicks"],store[brand=="
   dominicks"],FUN=mean,na.rm=TRUE),tapply(WORKWOM[brand
   =="dominicks"],store[brand=="dominicks"],FUN=mean,na.
   rm=TRUE))

#Correlation between mean logmove per store and mean
   WORKWOM per store for Minute MAid
cor(tapply(logmove[brand=="minute.maid"],store[brand=="
```

```
   minute.maid"],FUN=mean,na.rm=TRUE),tapply(WORKWOM[
   brand=="minute.maid"],store[brand=="minute.maid"],FUN
   =mean,na.rm=TRUE))

#Correlation between mean logmove per store and mean
   WORKWOM per store for tropicana
cor(tapply(logmove[brand=="tropicana"],store[brand=="
   tropicana"],FUN=mean,na.rm=TRUE),tapply(WORKWOM[brand
   =="tropicana"],store[brand=="tropicana"],FUN=mean,na.
   rm=TRUE))
```

**OUTPUT**:

```
> summary(WORKWOM)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.2445  0.3126  0.3556  0.3592  0.4023  0.4723
> summary(percWORKWOM)
  (0,0.05] (0.05,0.1] (0.1,0.15] (0.15,0.2] (0.2,0.25]
     (0.25,0.3] (0.3,0.35] (0.35,0.4] (0.4,0.45]
     (0.45,0.5]
          0          0          0          0        354
               4578       8661       7629       6306
               1419
> #Correlation between mean logmove per store and mean
   WORKWOM per store
> cor(tapply(logmove,store,FUN=mean,na.rm=TRUE),tapply(
   WORKWOM,store,FUN=mean,na.rm=TRUE))
[1] -0.2659654
>
> #Correlation between mean logmove per store and mean
   WORKWOM per store for dominicks
> cor(tapply(logmove[brand=="dominicks"],store[brand=="
   dominicks"],FUN=mean,na.rm=TRUE),tapply(WORKWOM[brand
   =="dominicks"],store[brand=="dominicks"],FUN=mean,na.
   rm=TRUE))
[1] -0.4141641
>
> #Correlation between mean logmove per store and mean
   WORKWOM per store for Minute MAid
> cor(tapply(logmove[brand=="minute.maid"],store[brand
   =="minute.maid"],FUN=mean,na.rm=TRUE),tapply(WORKWOM[
   brand=="minute.maid"],store[brand=="minute.maid"],FUN
```

```
   =mean,na.rm=TRUE))
[1] -0.2156699
>
> #Correlation between mean logmove per store and mean
   WORKWOM per store for tropicana
> cor(tapply(logmove[brand=="tropicana"],store[brand=="
   tropicana"],FUN=mean,na.rm=TRUE),tapply(WORKWOM[brand
   =="tropicana"],store[brand=="tropicana"],FUN=mean,na.
   rm=TRUE))
[1] -0.0455048
```

**CODE**:

```
###HVAL150###
summary(HVAL150)
oj$percHVAL150 <- cut(HVAL150,breaks=c
   (0,0.10,0.20,0.30,0.40,0.50,0.60,0.70,0.80,0.90,1.00)
   )
summary(percHVAL150)
hist(tapply(HVAL150,store,mean))

#mean log units sold to percentage HH worth > 150,000
tapply(logmove,list(percHVAL150,brand),mean)

#median log units sold to percentage of HH worth >
   150,000
tapply(logmove,list(percHVAL150,brand),median)

#sd of log units sold to percentage of HH worth >
   150,000
tapply(logmove,list(percHVAL150,brand),sd)

#Correlation between mean logmove per store and mean
   HVAL150 per store
cor(tapply(logmove,store,FUN=mean,na.rm=TRUE),tapply(
   HVAL150,store,FUN=mean,na.rm=TRUE))

#Correlation between mean logmove per store and mean
   HVAL150 per store for dominicks
cor(tapply(logmove[brand=="dominicks"],store[brand=="
   dominicks"],FUN=mean,na.rm=TRUE),tapply(HVAL150[brand
   =="dominicks"],store[brand=="dominicks"],FUN=mean,na.
```

```
   rm=TRUE))

#Correlation between mean logmove per store and mean
   HVAL150 per store for Minute MAid
cor(tapply(logmove[brand=="minute.maid"],store[brand=="
   minute.maid"],FUN=mean,na.rm=TRUE),tapply(HVAL150[
   brand=="minute.maid"],store[brand=="minute.maid"],FUN
   =mean,na.rm=TRUE))

#Correlation between mean logmove per store and mean
   HVAL150 per store for tropicana
cor(tapply(logmove[brand=="tropicana"],store[brand=="
   tropicana"],FUN=mean,na.rm=TRUE),tapply(HVAL150[brand
   =="tropicana"],store[brand=="tropicana"],FUN=mean,na.
   rm=TRUE))
```

**OUTPUT**:

```
> summary(HVAL150)
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
0.002509 0.123500 0.346200 0.343800 0.528300 0.916700
> summary(percHVAL150)
  (0,0.1] (0.1,0.2] (0.2,0.3] (0.3,0.4] (0.4,0.5]
     (0.5,0.6] (0.6,0.7] (0.7,0.8] (0.8,0.9]   (0.9,1]
     6966      2826      2835      4113      4521
        3549      2088       696      1002       351
> #Correlation between mean logmove per store and mean
   HVAL150 per store
> cor(tapply(logmove,store,FUN=mean,na.rm=TRUE),tapply(
   HVAL150,store,FUN=mean,na.rm=TRUE))
[1] 0.06603706
>
> #Correlation between mean logmove per store and mean
   HVAL150 per store for dominicks
> cor(tapply(logmove[brand=="dominicks"],store[brand=="
   dominicks"],FUN=mean,na.rm=TRUE),tapply(HVAL150[brand
   =="dominicks"],store[brand=="dominicks"],FUN=mean,na.
   rm=TRUE))
[1] -0.540369
>
> #Correlation between mean logmove per store and mean
   HVAL150 per store for Minute MAid
```

```
> cor(tapply(logmove[brand=="minute.maid"],store[brand
   =="minute.maid"],FUN=mean,na.rm=TRUE),tapply(HVAL150[
   brand=="minute.maid"],store[brand=="minute.maid"],FUN
   =mean,na.rm=TRUE))
[1] 0.1466044
>
> #Correlation between mean logmove per store and mean
   HVAL150 per store for tropicana
> cor(tapply(logmove[brand=="tropicana"],store[brand=="
   tropicana"],FUN=mean,na.rm=TRUE),tapply(HVAL150[brand
   =="tropicana"],store[brand=="tropicana"],FUN=mean,na.
   rm=TRUE))
[1] 0.5377712
```

**CODE**:

```
###SSTRDIST###

summary(SSTRDIST)
hist(tapply(SSTRDIST,store,mean))

#Correlation between mean logmove per store and mean
   distance to the nearest warehouse store per store
cor(tapply(logmove,store,FUN=mean,na.rm=TRUE),tapply(
   SSTRDIST,store,FUN=mean,na.rm=TRUE))

##Correlation between mean logmove per store and mean
   distance to the nearest warehouse store per store for
    dominicks
cor(tapply(logmove[brand=="dominicks"],store[brand=="
   dominicks"],FUN=mean,na.rm=TRUE),tapply(SSTRDIST[
   brand=="dominicks"],store[brand=="dominicks"],FUN=
   mean,na.rm=TRUE))

##Correlation between mean logmove per store and mean
   distance to the nearest warehouse store per store for
    Minute MAid
cor(tapply(logmove[brand=="minute.maid"],store[brand=="
   minute.maid"],FUN=mean,na.rm=TRUE),tapply(SSTRDIST[
   brand=="minute.maid"],store[brand=="minute.maid"],FUN
   =mean,na.rm=TRUE))
```

```
##Correlation between mean logmove per store and mean
   distance to the nearest warehouse store per store for
    tropicana
cor(tapply(logmove[brand=="tropicana"],store[brand=="
   tropicana"],FUN=mean,na.rm=TRUE),tapply(SSTRDIST[
   brand=="tropicana"],store[brand=="tropicana"],FUN=
   mean,na.rm=TRUE))
```

**OUTPUT**:

```
> summary(SSTRDIST)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.1321  2.7670  4.6510  5.0970  6.6510 17.8600
```

**CODE**:

```
###CPDIST5###
summary(CPDIST5)
hist(tapply(CPDIST5,store,mean))

#Correlation between mean logmove per store and average
   distance in miles to the nearest 5 supermarkets
cor(tapply(logmove,store,FUN=mean,na.rm=TRUE),tapply(
   SSTRDIST,store,FUN=mean,na.rm=TRUE))


pdf("lovemovebrandsts.pdf")
plot.ts(tapply(logmove,INDEX=list(week,brand),FUN=mean),
   plot.type=("single"),ylab="log of the number of units
    sold",col=c("red","blue","green"))
legend("topright",cex=0.6,legend=c("Dominicks","Minute
   Maid","Tropicana"),lty=c(1,1),col=c("red","blue","
   green"))
graphics.off()
```

**CODE**:

```
###REGRESSION###

hist(tapply(logmove,INDEX=list(AGE60,brand),FUN=mean))

lmout = lm(formula = logmove ~ EDUC + INCOME, data = oj)
```

```
summary(lmout)
stargazer(lmout)
```

**OUTPUT**:

Table 1

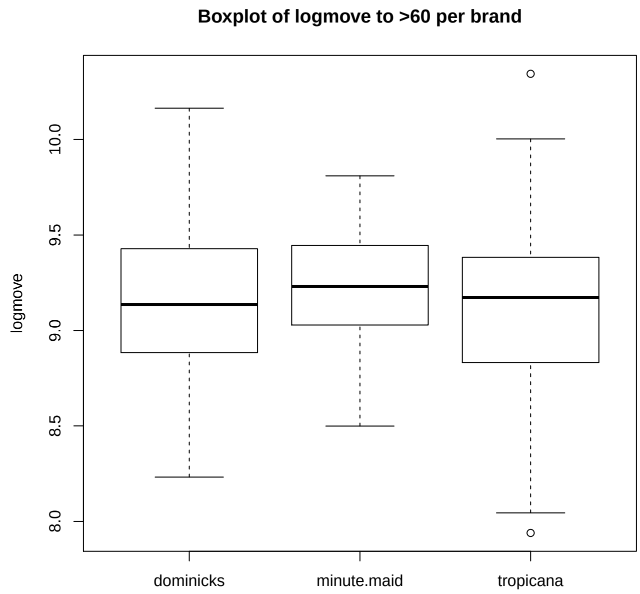
| | *Dependent variable:* |
|---|---|
| | logmove |
| EDUC | $-0.639^{***}$ |
| | (0.132) |
| INCOME | $0.102^{**}$ |
| | (0.046) |
| ETHNIC | $0.569^{***}$ |
| | (0.054) |
| HHLARGE | $-2.549^{***}$ |
| | (0.262) |
| HVAL150 | $0.293^{***}$ |
| | (0.059) |
| Constant | $8.336^{***}$ |
| | (0.466) |
| Observations | 28,947 |
| $R^2$ | 0.010 |
| Adjusted $R^2$ | 0.010 |
| Residual Std. Error | 1.014 (df = 28941) |
| F Statistic | $60.145^{***}$ (df = 5; 28941) |
| *Note:* | $^{*}$p<0.1; $^{**}$p<0.05; $^{***}$p<0.01 |

## (b) and (c)

We have found several findings which may be significant.

### AGE60

There is a 0.30 correlation between logmove and AGE60. This correlation is positive but not very strong. It seems that this correlation is strongest for the Tropicana brand. Tropicana could therefore be marketed to individuals over the age of 60 in stores where there is a high percentage of those individuals.



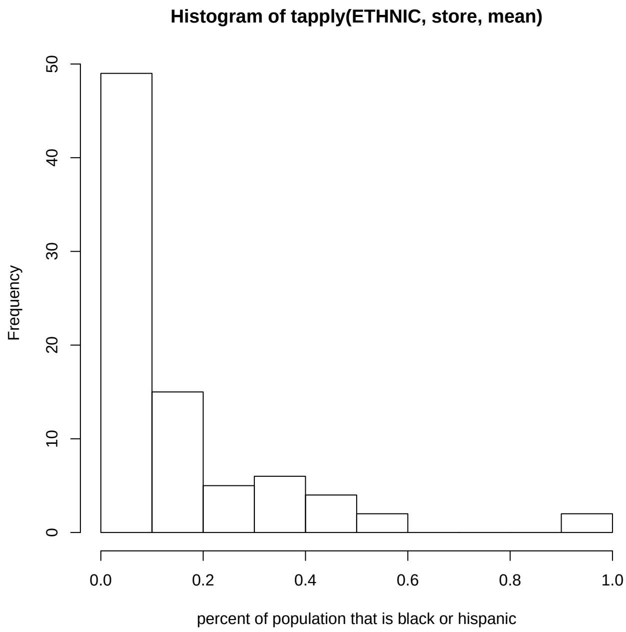

*Plot:* Boxplot logmove over 60

### EDUC

There is a strong negative relationship between percentage of the population with a college degree and the sale of Dominicks (cor = -0.49). Reversely, there is a strong positive relationship for the Tropicana brand. Educated individuals shun Dominicks and buy Tropicana.

### ETHNIC

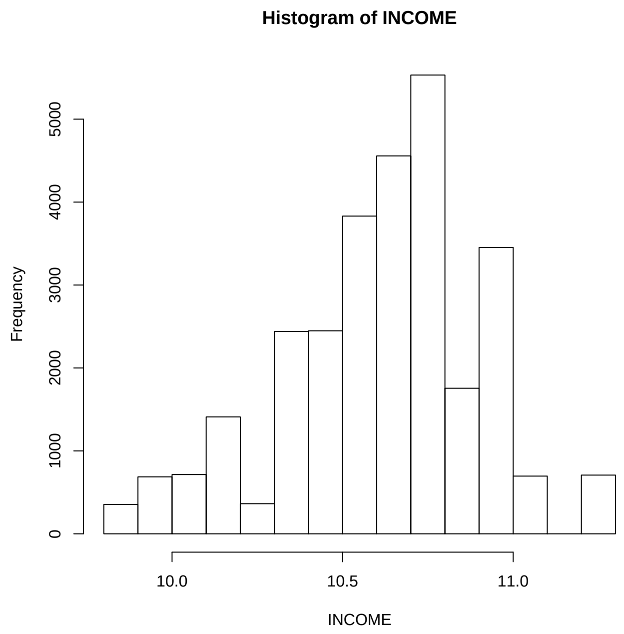
The histogram reveals to us that there is a small set of stores with a very high percentage of ethnic individuals (close to 100%). This group would therefore be easily marketed to. The statistics show that the percentage of ethnic individuals has a strong positive correlation (cor = 0.43) with the log number of units sold for the Dominicks brand. We do not know however whether this may be related to the fact that some ethnic individuals may be less educated. We may need to run a multivariate regression to uncover this question. Marketers may want to market Dominicks to ethnic areas.




*Plot:* Histogram of percentage ethnic per store

**INCOME**

Income is strongly negatively correlated with how many units of Dominicks are sold. Again the effect of income on logmove for Dominicks may be confounded with other variables. Reversely there is a weak positive relationship with the number of Tropicana units sold as we have come to expect.



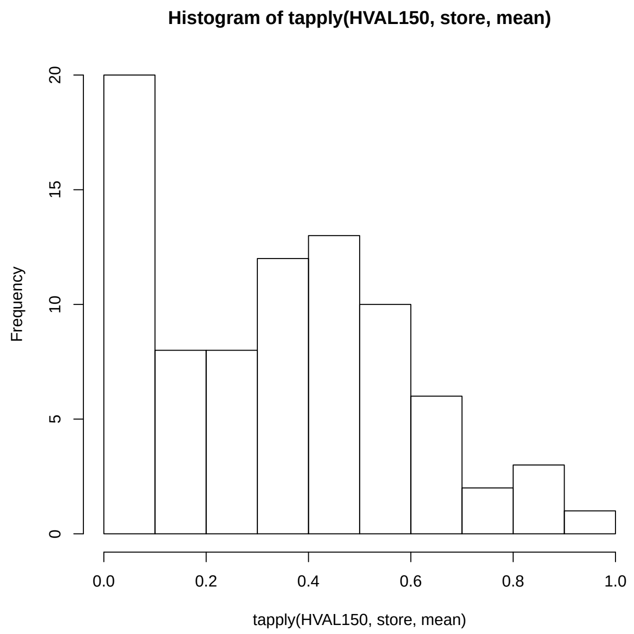

*Plot:* Histogram of median income

**HHLARGE**

The relationship between the percentage of large households and the number of units of Tropicana sold is strong and negative. This is expected, large households may be on a tighter budget and thus avoid to buy the most expensive orange juice brand. Marketers should avoid marketing that product to large households.

**HVAL150**

The percentage of households worth more than $150,000 have a very strong positive relationship with the number of units of Tropicana sold. Marketers may want to market Tropicana to wealthy neighbourhoods.




*Plot:* Histogram of households with income over 150,000

## (c)

The marketing team may want to customise marketing mix variables to the store level. Marketers can take advantage of the variation in buying behaviour across stores while combining the advantage of a large scale operation such as a store chain.
This report's descriptive statistics have uncovered several purchasing patterns which may be useful to marketers. Using these uncovered patterns marketers may wish to conduct other types of analysis (exploratory, inferential, predictive) in making decisions about marketing strategy.
These further analyses could study potential pricing or micro-pricing (different prices at the store level to account for demographic and competitive differences across stores) strategies by modelling changes in store-level demand as prices for the three different brands are altered [1]. Montgomery uses a hierarchical Bayesian model in modelling the store-level demand for a similar dataset.
We have uncovered that older, more educated and wealthier individuals are less likely to purchase dominicks and more likely to purchase Tropicana. With predictive analysis, marketers could calculate product-level price response for each store to take advantage of price and non-price sensitive customers.
For example, in affluent areas, the price of Tropicana could be increased relative

to the other two brands to take advantage of the low-price sensitivity in those areas. This low price sensitivity was uncovered to high prices for Tropicana was uncovered in the answer two question **9.**, where the negative relationship between price and logmove seems to disappear over a price threshold of around 3.

# APPENDIX

Table 2

| | vars | n | mean | sd | median | trimmed | mad | min | max | range | skew | kurtosis | s |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| store | 1 | 28,947 | 80.884 | 35.577 | 86 | 83.354 | 40.030 | 2 | 137 | 135 | -0.478 | -0.695 | 0.2 |
| brand* | 2 | 28,947 | 2 | 0.817 | 2 | 2 | 1.483 | 1 | 3 | 2 | 0 | -1.500 | 0.0 |
| week | 3 | 28,947 | 100.460 | 34.692 | 101 | 100.516 | 44.478 | 40 | 160 | 120 | -0.013 | -1.196 | 0.2 |
| logmove | 4 | 28,947 | 9.168 | 1.019 | 9.034 | 9.119 | 0.922 | 4.159 | 13.482 | 9.323 | 0.407 | 0.392 | 0.0 |
| feat | 5 | 28,947 | 0.237 | 0.425 | 0 | 0.172 | 0 | 0 | 1 | 1 | 1.235 | -0.474 | 0.0 |
| price | 6 | 28,947 | 2.282 | 0.648 | 2.170 | 2.252 | 0.712 | 0.520 | 3.870 | 3.350 | 0.382 | -0.620 | 0.0 |
| AGE60 | 7 | 28,947 | 0.173 | 0.062 | 0.171 | 0.171 | 0.067 | 0.058 | 0.307 | 0.249 | 0.282 | -0.699 | 0.0 |
| EDUC | 8 | 28,947 | 0.225 | 0.110 | 0.229 | 0.219 | 0.113 | 0.050 | 0.528 | 0.479 | 0.487 | -0.112 | 0.0 |
| ETHNIC | 9 | 28,947 | 0.156 | 0.188 | 0.075 | 0.116 | 0.058 | 0.024 | 0.996 | 0.971 | 2.547 | 7.501 | 0.0 |
| INCOME | 10 | 28,947 | 10.617 | 0.282 | 10.635 | 10.632 | 0.240 | 9.867 | 11.236 | 1.369 | -0.414 | 0.039 | 0.0 |
| HHLARGE | 11 | 28,947 | 0.116 | 0.030 | 0.111 | 0.115 | 0.025 | 0.014 | 0.216 | 0.203 | 0.327 | 2.041 | 0.0 |
| WORKWOM | 12 | 28,947 | 0.359 | 0.053 | 0.356 | 0.358 | 0.065 | 0.244 | 0.472 | 0.228 | 0.155 | -0.893 | 0.0 |
| HVAL150 | 13 | 28,947 | 0.344 | 0.239 | 0.346 | 0.330 | 0.292 | 0.003 | 0.917 | 0.914 | 0.351 | -0.698 | 0.0 |
| SSTRDIST | 14 | 28,947 | 5.097 | 3.472 | 4.651 | 4.645 | 2.793 | 0.132 | 17.856 | 17.724 | 1.382 | 2.349 | 0.0 |
| SSTRVOL | 15 | 28,947 | 1.207 | 0.527 | 1.115 | 1.163 | 0.607 | 0.400 | 2.571 | 2.171 | 0.624 | -0.318 | 0.0 |
| CPDIST5 | 16 | 28,947 | 2.120 | 0.730 | 1.963 | 2.087 | 0.588 | 0.773 | 4.108 | 3.335 | 0.527 | -0.219 | 0.0 |
| CPWVOL5 | 17 | 28,947 | 0.439 | 0.219 | 0.383 | 0.423 | 0.203 | 0.095 | 1.143 | 1.049 | 0.715 | 0.153 | 0.0 |



Louis Carnec
15204934

# References

[1] Montgomery, A. L. Creating micro-marketing pricing strategies using supermarket scanner data. *Marketing Science 16*, 4 (1997), 315–337.