Orange Juice data

Emanuele Taufer

Orange Juice Data

The data contain weekly sales of refrigerated orange juice in 64-ounce containers for 83 stores in a chain (Dominick’s Finer Foods (DFF)), in the Chicago area.

The data refer to sales for 121 weeks and three different brands (Dominicks, MinuteMaid and Tropicana).

Each line of the data set provides the store sales (in logarithms: logmove), the brand, the price, the presence / absence of advertising and the demographic characteristics of the neighborhood in which the point of sale operates.

There are 28,947 lines (units).

The data are taken from the *bayesm* (or *ISLR*) package by P. Rossi for R and they have been used by Montgomery (Marketing Science, Vol. 16, No. 4, (1997), pp. 315-337).

The data

```
OJ<-read.table("http://www.cs.unitn.it/~taufer/Data/oj.csv",header=T,sep=",")
str(OJ)
```

```
## 'data.frame':    28947 obs. of  17 variables:
##  $ store   : int  2 2 2 2 2 2 2 2 2 2 ...
##  $ brand   : Factor w/ 3 levels "dominicks","minute.maid",..: 3 3 3 3 3 3 3 3 3 3 ...
##  $ week    : int  40 46 47 48 50 51 52 53 54 57 ...
##  $ logmove : num  9.02 8.72 8.25 8.99 9.09 ...
##  $ feat    : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ price   : num  3.87 3.87 3.87 3.87 3.87 3.87 3.29 3.29 3.29 3.29 ...
##  $ AGE60   : num  0.233 0.233 0.233 0.233 0.233 ...
##  $ EDUC    : num  0.249 0.249 0.249 0.249 0.249 ...
##  $ ETHNIC  : num  0.114 0.114 0.114 0.114 0.114 ...
##  $ INCOME  : num  10.6 10.6 10.6 10.6 10.6 ...
##  $ HHLARGE : num  0.104 0.104 0.104 0.104 0.104 ...
##  $ WORKWOM : num  0.304 0.304 0.304 0.304 0.304 ...
##  $ HVAL150 : num  0.464 0.464 0.464 0.464 0.464 ...
##  $ SSTRDIST: num  2.11 2.11 2.11 2.11 2.11 ...
##  $ SSTRVOL : num  1.14 1.14 1.14 1.14 1.14 ...
##  $ CPDIST5 : num  1.93 1.93 1.93 1.93 1.93 ...
##  $ CPWVOL5 : num  0.377 0.377 0.377 0.377 0.377 ...
```

Variables

1. STORE - store number
2. BRAND - brand indicator
3. WEEK - week number
4. LOGMOVE -log of the number of units sold
5. PRICE
6. FEAT - feature advertisement
7. AGE60 - percentage of the population that is aged 60 or older
8. EDUC - percentage of the population that has a college degree

9. ETHNIC - percent of the population that is black or Hispanic

10. INCOME - median income

11. HHLARGE - percentage of households with 5 or more persons

12. WORKWOM - percentage of women with full-time jobs

13. HVAL150 - percentage of households worth more than $150,000

14. SSTRDIST - distance to the nearest warehouse store

15. SSTRVOL - ratio of sales of this store to the nearest warehouse store

16. CPDIST5 - average distance in miles to the nearest 5 supermarkets

17. CPWVOL5 -ratio of sales of this store to the average of the nearest five stores

```r
head(OJ)
```

```
##   store     brand week  logmove feat price     AGE60      EDUC    ETHNIC
## 1     2 tropicana   40 9.018695    0  3.87 0.2328647 0.2489349 0.1142799
## 2     2 tropicana   46 8.723231    0  3.87 0.2328647 0.2489349 0.1142799
## 3     2 tropicana   47 8.253228    0  3.87 0.2328647 0.2489349 0.1142799
## 4     2 tropicana   48 8.987197    0  3.87 0.2328647 0.2489349 0.1142799
## 5     2 tropicana   50 9.093357    0  3.87 0.2328647 0.2489349 0.1142799
## 6     2 tropicana   51 8.877382    0  3.87 0.2328647 0.2489349 0.1142799
##     INCOME   HHLARGE   WORKWOM   HVAL150 SSTRDIST  SSTRVOL CPDIST5
## 1 10.55321 0.1039534 0.3035853 0.4638871 2.110122 1.142857 1.92728
## 2 10.55321 0.1039534 0.3035853 0.4638871 2.110122 1.142857 1.92728
## 3 10.55321 0.1039534 0.3035853 0.4638871 2.110122 1.142857 1.92728
## 4 10.55321 0.1039534 0.3035853 0.4638871 2.110122 1.142857 1.92728
## 5 10.55321 0.1039534 0.3035853 0.4638871 2.110122 1.142857 1.92728
## 6 10.55321 0.1039534 0.3035853 0.4638871 2.110122 1.142857 1.92728
##     CPWVOL5
## 1 0.3769266
## 2 0.3769266
## 3 0.3769266
## 4 0.3769266
## 5 0.3769266
## 6 0.3769266
```

Objectives of the analysis

Build a predictive and inferential model for sales of orange juice

Understand if it is appropriate to differentiate prices at the level of areas / stores rather than use a uniform pricing strategy for all stores.

Strategy of analysis

Since the main objective is to analyze sales' trends against predictors and compare sales between different stores and / or zones, the linear regression model is probably the most suitable model.

In fact, the RL model allows, through the specification of appropriate equations, to construct an analysis structure adequate to the problem and to the questions of analysis.

As there are 83 stores in the data set, it is advisable to check whether there are any homogeneous groupings.

Structure

In the example, three main steps will be used in the analysis

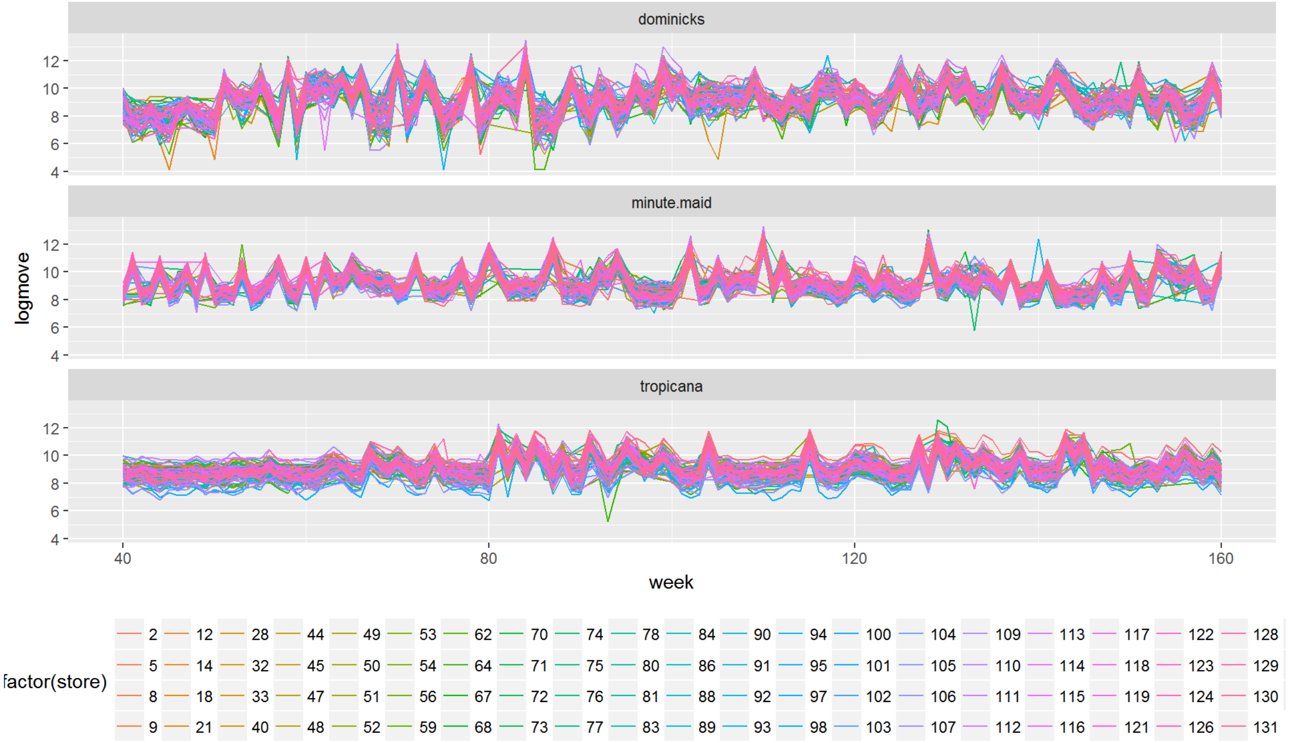
- Use of the PCA as an exploratory analysis on demographic variables.

- On the basis of the indications obtained from the PCA, we will proceed to identify groups of sales points (k-means), homogeneous by area of ♂♂operation.

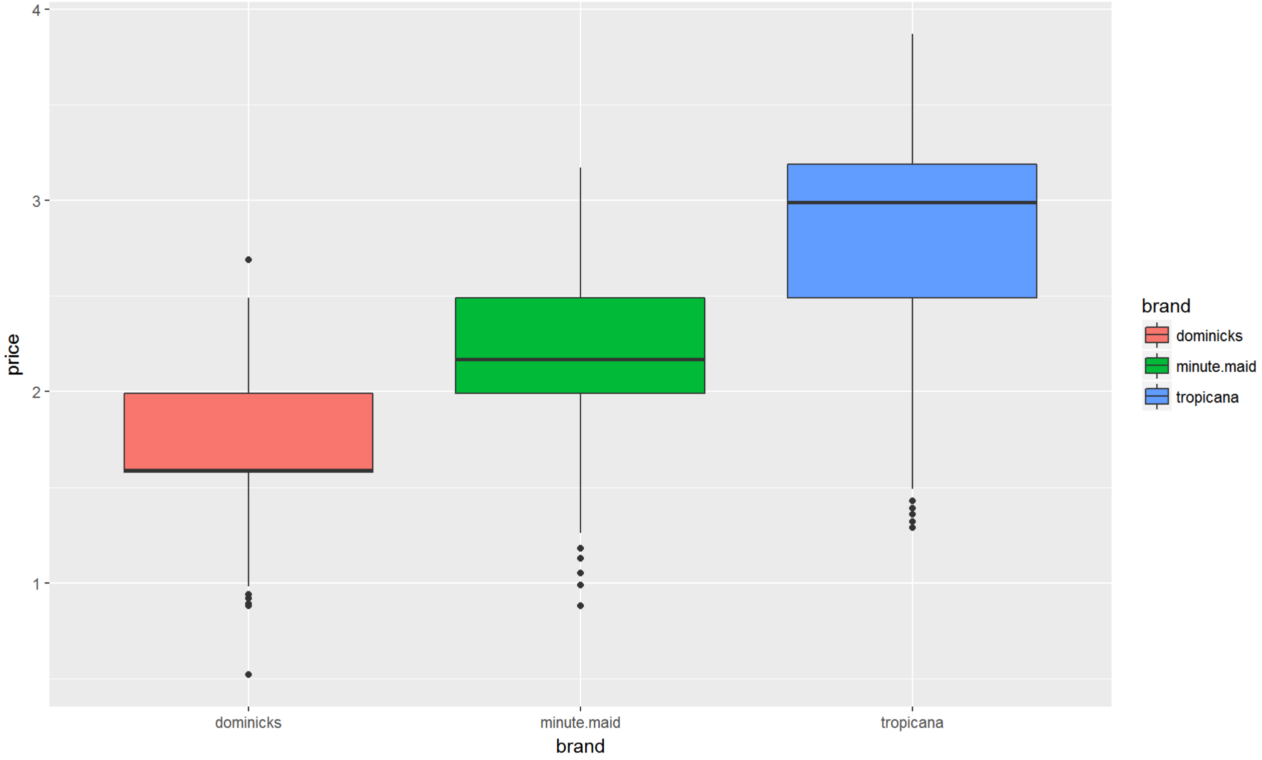
- Estimation and analysis of some regression models using the groups (and other predictors) identified with the cluster analysis

Some preliminary analysis

```
g1<-ggplot(OJ,aes(x=week,y=logmove,color=factor(store)))+geom_line()
g1+facet_wrap(~brand,nrow=3,ncol=1)+theme(legend.position="bottom",
          legend.direction = 'horizontal') + guides(col = guide_legend(ncol = 21))
```



```
g0<-ggplot(OJ,aes(x=brand,y=price))+geom_boxplot(aes(fill=brand))
g0
```



Why LOGMOVE?

The LOGMOVE variable represents the logarithm of the quantity sold of orange juice (per week, of a certain brand, in a certain store)

Why is the logarithm used?

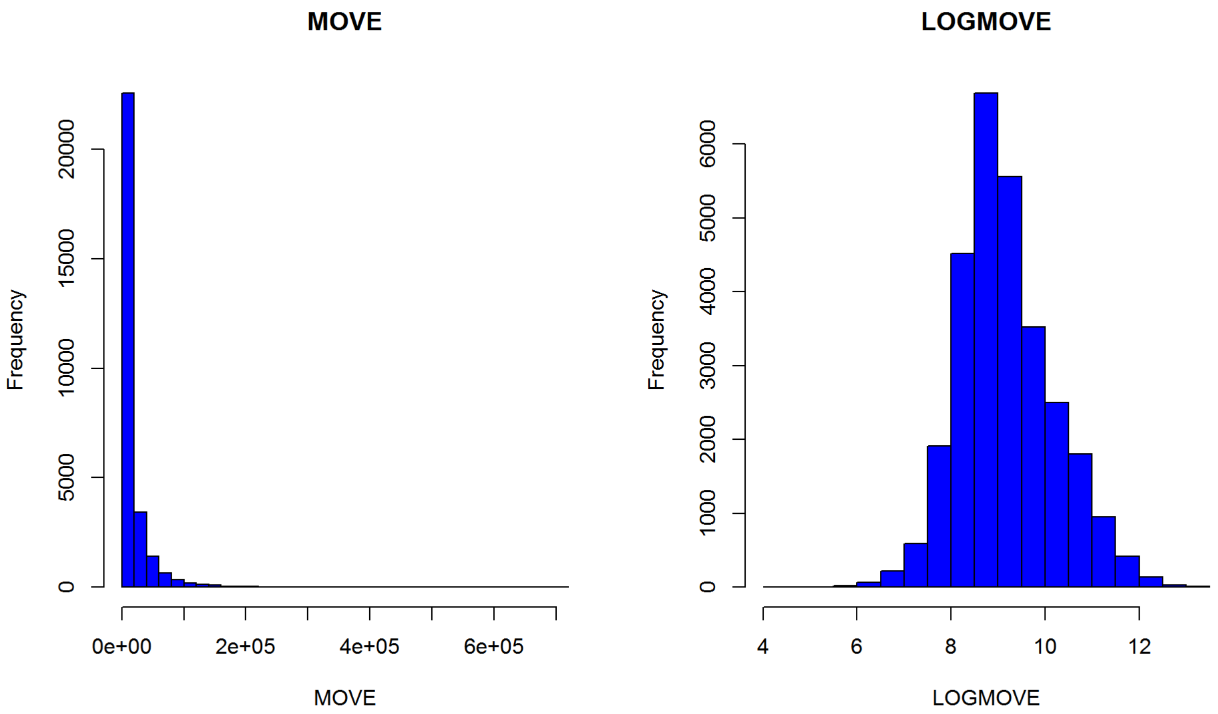
If you look at the sales histogram (not the logarithmic terms), which we calculate as

$$\mathrm{MOVE} = \exp\{\mathrm{LOGMOVE}\},$$

we note that it is strongly asymmetric.

The use of a dependent variable with strongly asymmetric distribution can have consequences on the reliability of the models.

Therefore, it is customary to use a logarithmic transformation that reduces the data asymmetry.



MOVE

LOGMOVE

PCA

Objective: summarize the characteristics of the area in which each sales point operates

Since the data on the demographic characteristics of the area in which the *store* operates are constant, if we want to try to synthesise them through a PCA, it is necessary to construct a reduced data set that contains:

- the *stores* in the rows (the units)
- the demographic variables in the column

Use *duplicated()* to select only one row (the first) of the OJ dataset for each *store*

```
OJ.S=OJ[!duplicated(OJ$store),]
rownames(OJ.S)<-OJ.S$store
OJ.S=OJ.S[,7:13]
```

The OJ. data.frame contains for each of the 83 stores the demographic variables

```
head(OJ.S)
```

```
##        AGE60       EDUC    ETHNIC    INCOME     HHLARGE   WORKWOM
## 2  0.2328647 0.24893493 0.11427995 10.553205 0.10395341 0.3035853
## 5  0.1173680 0.32122573 0.05387528 10.922371 0.10309158 0.4105680
## 8  0.2523940 0.09517327 0.03524333 10.597010 0.13174970 0.2830747
## 9  0.2691190 0.22217232 0.03261883 10.787152 0.09683047 0.3589945
## 12 0.1783414 0.25341297 0.38069799  9.996659 0.05721242 0.3909416
## 14 0.2139493 0.34829302 0.03417874 11.043929 0.10789429 0.3623057
##       HVAL150
## 2  0.46388706
## 5  0.53588335
## 8  0.05422716
## 9  0.50574713
## 12 0.38662791
## 14 0.75076991
```

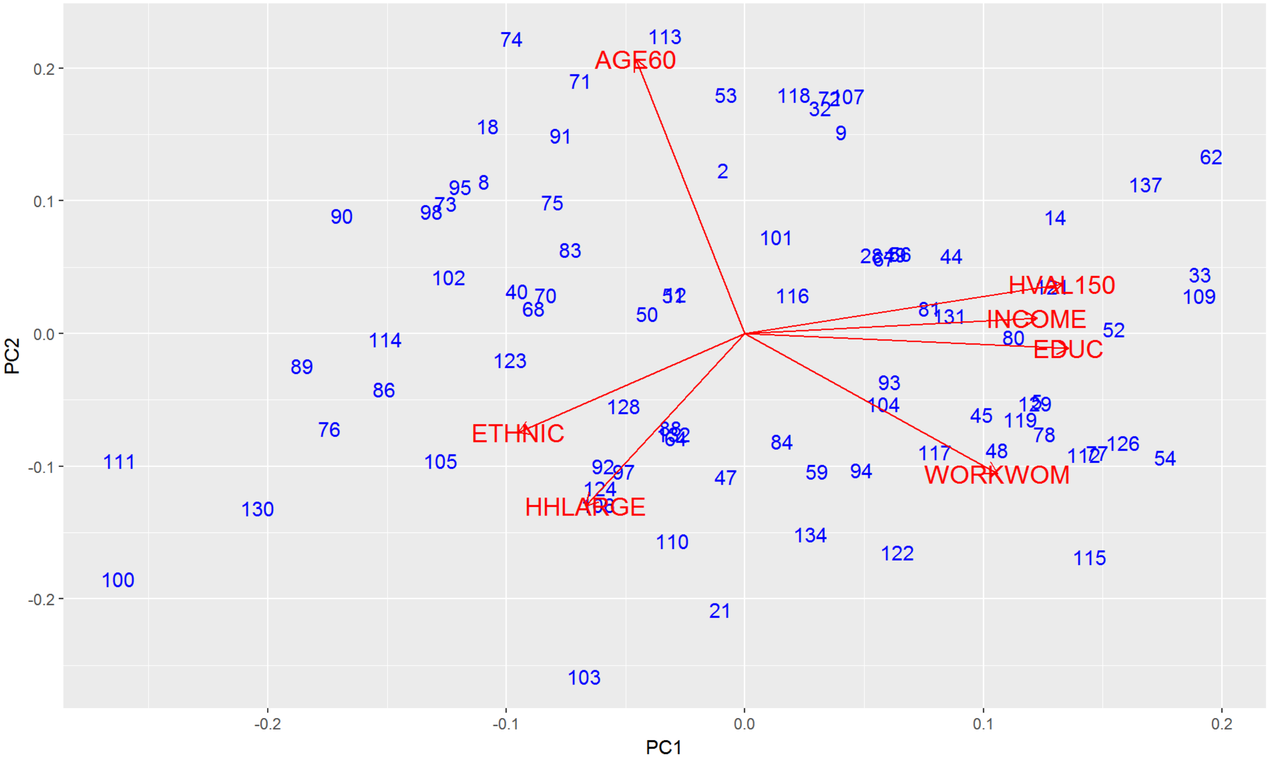
PCA on OJ.S

```
pca.OJ=prcomp(OJ.S,scale=T)
summary(pca.OJ)
```

```
## Importance of components:
##                           PC1    PC2    PC3    PC4     PC5     PC6     PC7
## Standard deviation     1.8515 1.2453 1.0215 0.7946 0.39740 0.34143 0.26811
## Proportion of Variance 0.4897 0.2215 0.1490 0.0902 0.02256 0.01665 0.01027
## Cumulative Proportion  0.4897 0.7113 0.8603 0.9505 0.97308 0.98973 1.00000
```

Biplot

```r
autoplot(pca.OJ,shape = FALSE, colour="blue",label.size = 4,loadings = TRUE,
  loadings.colour = 'red',loadings.label = TRUE, loadings.label.size = 5)
```



The biplot analysis clearly shows three distinct zones:

- on the right, an area characterized by the level of education, income and high valued houses,

- at the top, an area with high percentage of population over 60

- in the lower left, an area with high percentages of large families, Hispanic or African-American

This information can be used to try to perform a cluster analysis with 3 or 4 groups

Cluster analysis

Let’s try to build a cluster analysis (k-means) with $k = 4$.

We standardize the variables before applying the k-means algorithm

```
SOJ.S=scale(OJ.S)
```

kmeans with $k = 4$

```
set.seed(1)
km=kmeans(SOJ.S,centers=4,nstart=20)
km
```

```
## K-means clustering with 4 clusters of sizes 29, 25, 14, 15
## 
## Cluster means:
##         AGE60       EDUC      ETHNIC      INCOME     HHLARGE    WORKWOM
## 1 -0.41709719  1.0232181 -0.45627721  0.86471328 -0.5137673  0.8047049
## 2  1.10468515 -0.6277896 -0.40225578 -0.14662237 -0.2579298 -0.8412513
## 3 -0.06533376 -0.7614684  1.72778884 -1.61464490  0.3721148 -0.6549377
## 4 -0.97377584 -0.2212019 -0.06004068  0.07959352  1.0758593  0.4575979
##      HVAL150
## 1  0.9634415
## 2 -0.4331624
## 3 -0.6863681
## 4 -0.5001059
## 
## Clustering vector:
##   2   5   8   9  12  14  18  21  28  32  33  40  44  45  47  48  49  50 
##   2   1   2   2   3   1   2   4   1   2   1   2   1   1   4   1   1   2 
##  51  52  53  54  56  59  62  64  67  68  70  71  72  73  74  75  76  77 
##   2   1   2   1   1   4   1   4   1   3   2   2   2   2   2   3   3   1 
##  78  80  81  83  84  86  88  89  90  91  92  93  94  95  97  98 100 101 
##   1   1   1   2   4   3   4   3   2   2   4   1   4   2   4   2   3   2 
## 102 103 104 105 106 107 109 110 111 112 113 114 115 116 117 118 119 121 
##   2   4   1   3   4   2   1   4   3   1   2   3   1   2   1   2   1   1 
## 122 123 124 126 128 129 130 131 132 134 137 
##   4   3   3   1   3   1   3   1   4   4   1 
## 
## Within cluster sum of squares by cluster:
## [1] 92.16334 60.89011 70.41467 24.68977
##  (between_SS / total_SS =  56.8 %)
## 
## Available components:
## 
## [1] "cluster"      "centers"      "totss"        "withinss"    
## [5] "tot.withinss" "betweenss"    "size"         "iter"        
## [9] "ifault"
```

Analysis of groups

Observing the centroids you notice some peculiarities of the groups:

- Group 1: high values for EDUC, INCOME, WORKWOM, HVAL150 - we call the group **AFFLUENT**

- Group 2: above average value only for AGE60 - **ELDER**

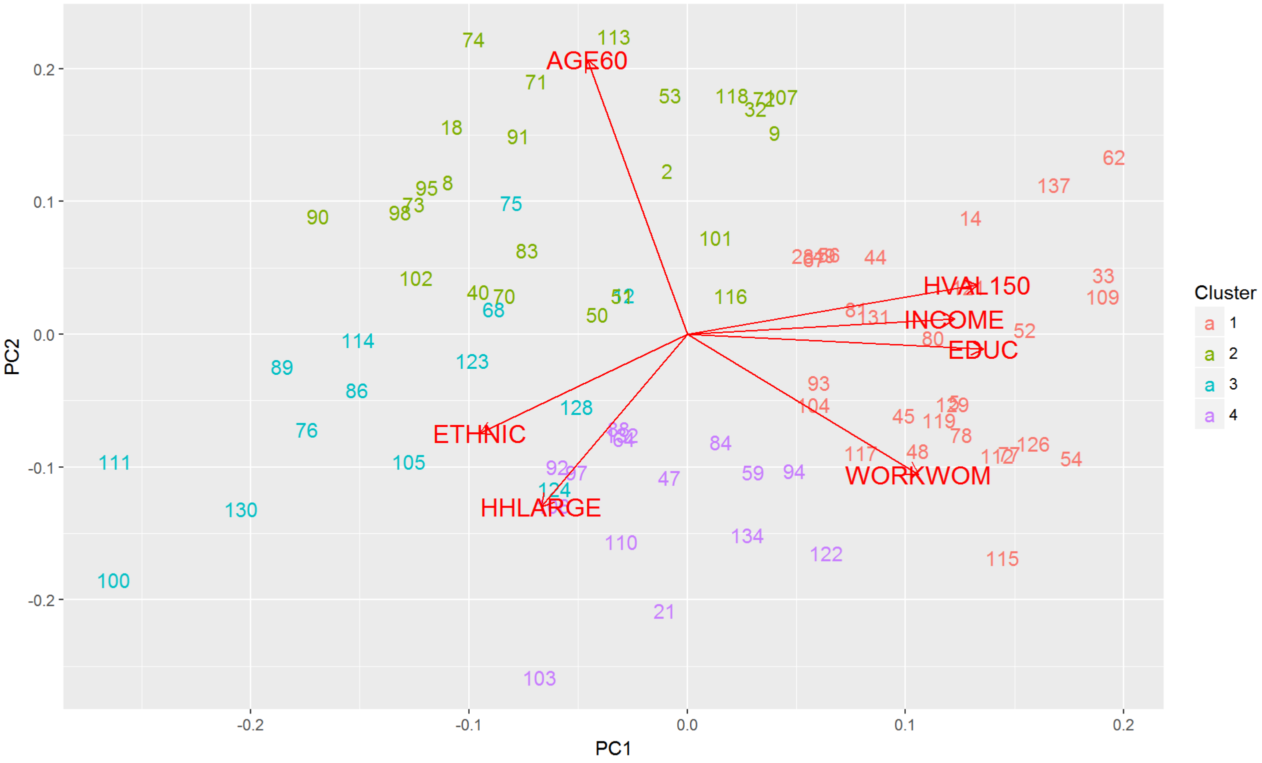
- Group 3: above average value only for ETHNIC - **ETHNIC**

- Group 4: above average values ♂♂for HHLARGE, WORKWOM - **YOUNG**

We insert clusters in the `data.frame` and plot using the PCA

The clustering identified by the PCA seems appropriate

```
OJ.S$Cluster<-factor(km$cluster)
pca.OJ=prcomp(OJ.S[,-8],scale=T)
autoplot(pca.OJ,shape = FALSE, data=OJ.S,colour="Cluster",label.size = 4,
         loadings = TRUE, loadings.colour = 'red',loadings.label = TRUE,
         loadings.label.size = 5)
```



Preparation of the OJ dataset for predictive analysis

Insert the group variable in the original dataset OJ

```
dfc=data.frame(km$cluster)
dfc$shop=rownames(dfc)
dfc$shop=as.numeric(dfc$shop)

g1<-dfc[dfc$km.cluster==1,2]
g2<-dfc[dfc$km.cluster==2,2]
g3<-dfc[dfc$km.cluster==3,2]
g4<-dfc[dfc$km.cluster==4,2]

OJ$group=NA

OJ$group[is.element(OJ$store,g1)]<-"AFFLUENT"
OJ$group[is.element(OJ$store,g2)]<-"ELDER"
OJ$group[is.element(OJ$store,g3)]<-"ETHNIC"
OJ$group[is.element(OJ$store,g4)]<-"YOUNG"
OJ$group=factor(OJ$group)
str(OJ)
```

```
## 'data.frame':    28947 obs. of  18 variables:
##  $ store   : int  2 2 2 2 2 2 2 2 2 2 ...
##  $ brand   : Factor w/ 3 levels "dominicks","minute.maid",..: 3 3 3 3 3 3 3 3 3 3 ...
##  $ week    : int  40 46 47 48 50 51 52 53 54 57 ...
##  $ logmove : num  9.02 8.72 8.25 8.99 9.09 ...
##  $ feat    : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ price   : num  3.87 3.87 3.87 3.87 3.87 3.87 3.29 3.29 3.29 3.29 ...
##  $ AGE60   : num  0.233 0.233 0.233 0.233 0.233 ...
##  $ EDUC    : num  0.249 0.249 0.249 0.249 0.249 ...
##  $ ETHNIC  : num  0.114 0.114 0.114 0.114 0.114 ...
##  $ INCOME  : num  10.6 10.6 10.6 10.6 10.6 ...
##  $ HHLARGE : num  0.104 0.104 0.104 0.104 0.104 ...
##  $ WORKWOM : num  0.304 0.304 0.304 0.304 0.304 ...
##  $ HVAL150 : num  0.464 0.464 0.464 0.464 0.464 ...
##  $ SSTRDIST: num  2.11 2.11 2.11 2.11 2.11 ...
##  $ SSTRVOL : num  1.14 1.14 1.14 1.14 1.14 ...
##  $ CPDIST5 : num  1.93 1.93 1.93 1.93 1.93 ...
##  $ CPWVOL5 : num  0.377 0.377 0.377 0.377 0.377 ...
##  $ group   : Factor w/ 4 levels "AFFLUENT","ELDER",..: 2 2 2 2 2 2 2 2 2 2 ...
```

Regression models

We adapt two regression models:

- An additive model

- A model with interactions

We use the variables PRICE, WEEK, BRAND, FEAT, GROUP and SSTRDIST, SSTRVOL, CPDIST5, CPWVOL5 as predictors

We do NOT use the STORE, AGE60, .. variables. Their effect should be summarized in GROUP

The simplification made through the use of GROUP will be particularly effective when we are going to build a model with interactions

Estimation of the additive model

```
M1=lm(logmove~price+week+brand+feat+group+SSTRDIST+SSTRVOL+CPDIST5+CPWVOL5,data=OJ)
```

The output of the **M1** estimate indicates that two variables: SSTRDIST and CPDIST5 do not seem to be important.

To decide whether to exclude them from the analysis, we build a reduced model and use a partial F test

```
M2=lm(logmove~price+week+brand+feat+group+SSTRVOL+CPWVOL5,data=OJ)
anova(M2,M1)
```

```
## Analysis of Variance Table
##
## Model 1: logmove ~ price + week + brand + feat + group + SSTRVOL + CPWVOL5
## Model 2: logmove ~ price + week + brand + feat + group + SSTRDIST + SSTRVOL +
##     CPDIST5 + CPWVOL5
##   Res.Df   RSS Df Sum of Sq      F Pr(>F)
## 1  28936 14345
## 2  28934 14343  2    1.4211 1.4333 0.2385
```

… that tells us that **M2**, the reduced model, is preferred. We therefore exclude SSTRDIST and CPDIST5 from the analysis

Output M2 (additive)

```
summary(M2)
```

```
##
## Call:
## lm(formula = logmove ~ price + week + brand + feat + group +
##     SSTRVOL + CPWVOL5, data = OJ)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.7799 -0.4366  0.0018  0.4270  3.0637
##
## Coefficients:
##                      Estimate Std. Error  t value Pr(>|t|)
## (Intercept)      11.3250414  0.0298087  379.924  < 2e-16 ***
## price            -1.1214819  0.0102198 -109.736  < 2e-16 ***
## week             -0.0018525  0.0001255  -14.758  < 2e-16 ***
## brandminute.maid  0.5806321  0.0114305   50.797  < 2e-16 ***
## brandtropicana    1.2914759  0.0151946   84.996  < 2e-16 ***
## feat              0.9064761  0.0103944   87.208  < 2e-16 ***
## groupELDER        0.0571186  0.0109153    5.233 1.68e-07 ***
## groupETHNIC       0.1718437  0.0148254   11.591  < 2e-16 ***
## groupYOUNG       -0.3024589  0.0129785  -23.305  < 2e-16 ***
## SSTRVOL          -0.0446117  0.0093345   -4.779 1.77e-06 ***
## CPWVOL5          -0.4312476  0.0242979  -17.748  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7041 on 28936 degrees of freedom
## Multiple R-squared:  0.5231, Adjusted R-squared:  0.5229
## F-statistic:  3174 on 10 and 28936 DF,  p-value: < 2.2e-16
```

Some comments

All variables are significant and inconsistencies are not noticed

Effects:

- PRICE: the increase in price reduces the quantity sold
- WEEK: there is a slight negative trend in sales
- BRAND: Tropicana is the brand that sells, on average, more
- GROUP: ELDER and ETHNIC buy, on average, more than YOUNG and AFFLUENT; YOUNG purchases less than AFFLUENT
- SSTRVOL and CPWVOL5: competition effect

Model with interactions (M3)

```
M3=lm(logmove~price+week+brand*feat+feat*group+price*group+price*brand+SSTRVOL+CPWVOL5,data=OJ)
summary(M3)
```

```
## 
## Call:
## lm(formula = logmove ~ price + week + brand * feat + feat * group + 
##     price * group + price * brand + SSTRVOL + CPWVOL5, data = OJ)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -5.2901 -0.4051 -0.0046  0.4053  2.9175 
## 
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)    
## (Intercept)            11.8755929  0.0463189 256.388  < 2e-16 ***
## price                  -1.5047160  0.0206443 -72.888  < 2e-16 ***
## week                   -0.0018475  0.0001208 -15.300  < 2e-16 ***
## brandminute.maid       -0.6919671  0.0542700 -12.750  < 2e-16 ***
## brandtropicana         -0.4331857  0.0528947  -8.190 2.73e-16 ***
## feat                    0.8405005  0.0209748  40.072  < 2e-16 ***
## groupELDER              0.6746253  0.0395414  17.061  < 2e-16 ***
## groupETHNIC             1.0857573  0.0466555  23.272  < 2e-16 ***
## groupYOUNG              0.7860460  0.0468388  16.782  < 2e-16 ***
## SSTRVOL                -0.0399478  0.0088562  -4.511 6.49e-06 ***
## CPWVOL5                -0.4395636  0.0230527 -19.068  < 2e-16 ***
## brandminute.maid:feat   0.2755377  0.0228468  12.060  < 2e-16 ***
## brandtropicana:feat    -0.2401109  0.0258596  -9.285  < 2e-16 ***
## feat:groupELDER         0.0477572  0.0239771   1.992  0.04640 *  
## feat:groupETHNIC        0.0680050  0.0285575   2.381  0.01726 *  
## feat:groupYOUNG        -0.0729360  0.0280161  -2.603  0.00924 ** 
## price:groupELDER       -0.2759218  0.0158137 -17.448  < 2e-16 ***
## price:groupETHNIC      -0.4028006  0.0180837 -22.274  < 2e-16 ***
## price:groupYOUNG       -0.4754149  0.0190866 -24.908  < 2e-16 ***
## price:brandminute.maid  0.6724678  0.0256285  26.239  < 2e-16 ***
## price:brandtropicana    0.8578157  0.0227432  37.718  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.6678 on 28926 degrees of freedom
## Multiple R-squared:  0.5711, Adjusted R-squared:  0.5708 
## F-statistic:  1926 on 20 and 28926 DF,  p-value: < 2.2e-16
```

Compare M2-M3

```
anova(M2,M3)
```

```
## Analysis of Variance Table
##
## Model 1: logmove ~ price + week + brand + feat + group + SSTRVOL + CPWVOL5
## Model 2: logmove ~ price + week + brand * feat + feat * group + price *
##     group + price * brand + SSTRVOL + CPWVOL5
##   Res.Df   RSS Df Sum of Sq      F    Pr(>F)
## 1  28936 14345
## 2  28926 12900 10    1444.5 323.89 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

M3 is preferred to M2

Comments

Note that all interactions are significant; they indicate a rather elaborate structure in the phenomenon that is read more appropriately than the additive model.

Note, for example, that the positive value of the interaction between PRICE and TROPICANA does not indicate that sales, for this BRAND, increase as the price increases.

In fact, the main term must also be taken into account.

The price effect is therefore

- -1.5047 per unit of LOGMOVE for DOMINICS juice

- $-1.5047 + 0.6724 = -0.8323$ for MINUTE MAID

- $-1.5047 + 0.8578 = -0.6469$ for TROPICANA

The detailed analysis for all the effects, taking into account costs and revenues, will suggest the most appropriate price strategies for the different points of sale.

Final comments

The final M3 model seems a good compromise between interpretive simplicity and reading of the phenomenon.

You could try to build a more precise model using directly the main components (how many to be determined with cross-validation) on the demographic variables and those relating to the competitors.

If we need to cross validate this model, since we have data that also have a temporal dimension, the most appropriate procedure is that of the validation set approach where the last 4 weeks (for example) of data are used for validation.