

# R: Basics

## Session 7.3: Installation of R

For these exercises, download the files:

“Business Analytics – Week 7 Instructions.doc”

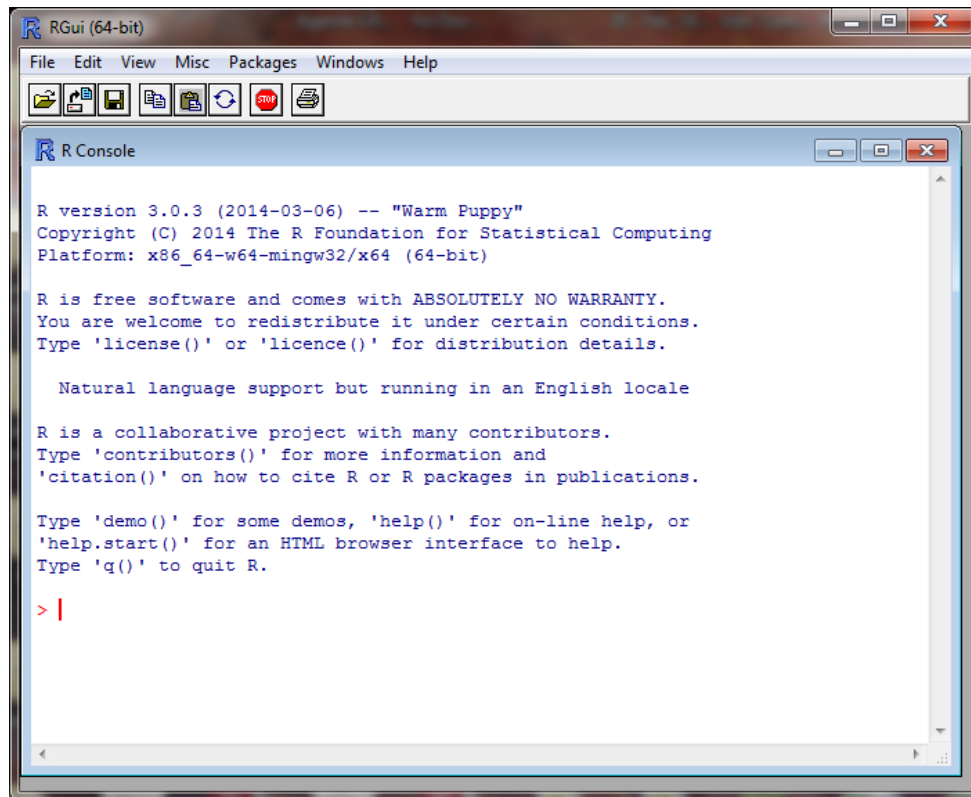
“Business Analytics – Week 7 oj.xls”

R is a free downloadable package capable of performing sophisticated statistical analysis and data mining. The software is already installed on the classroom laptops. To install on your own personal computer:

1. go to the website: <https://cran.r-project.org/>
  - a. Windows: <http://cran.r-project.org/bin/windows/base/>
  - b. Mac: <https://cran.r-project.org/bin/macosx/>
2. Click on Download R 3.2.2 (note: the version changes frequently)
3. Click on Run, and follow the install instructions

## Starting R

1. Click on the Start button in the lower left corner of Windows
2. Click on All Programs, then click on the R folder, then R

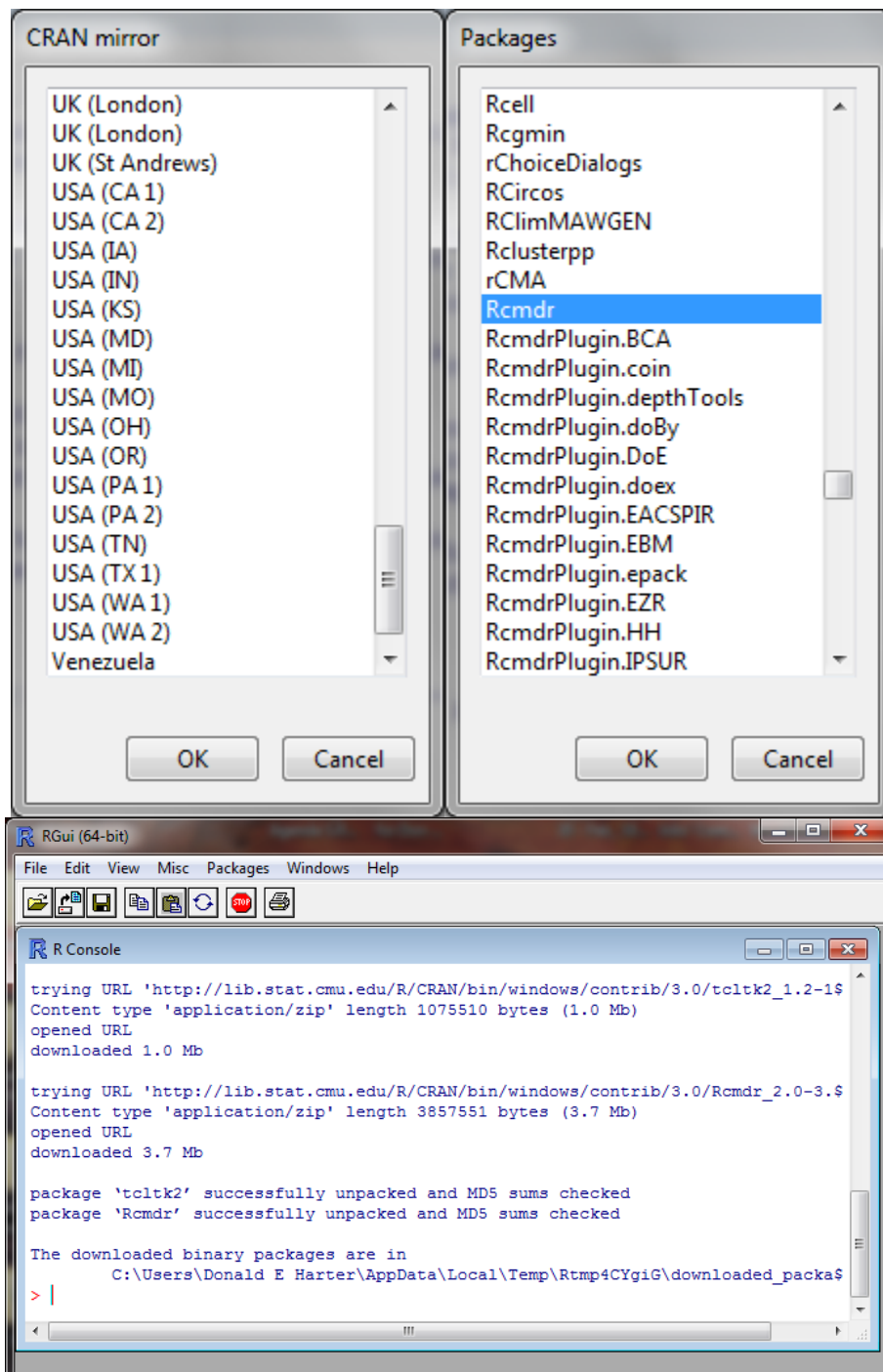


This is the command line screen. You can enter commands, but need to know the syntax. There is a simpler approach to running R, called Rcmdr (R Commander). If you are running a Whitman computer, Rcmdr is already installed. If not, you need to install it.

## Installing R Commander

Follow these steps only if you don't already have Rcmdr installed.

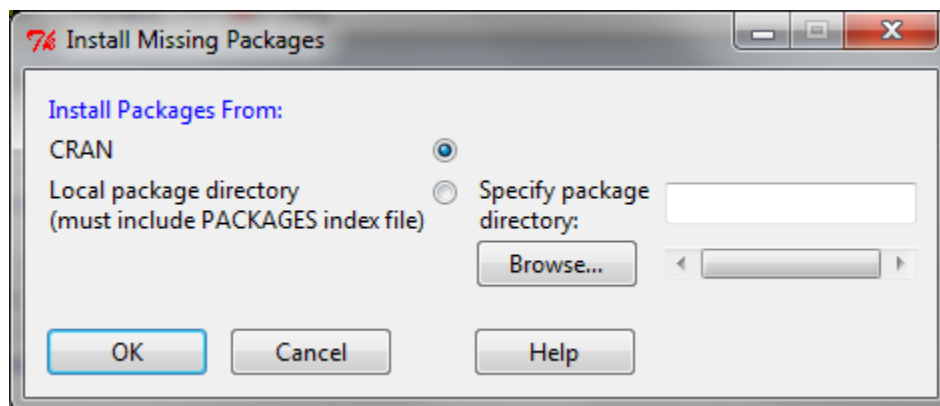
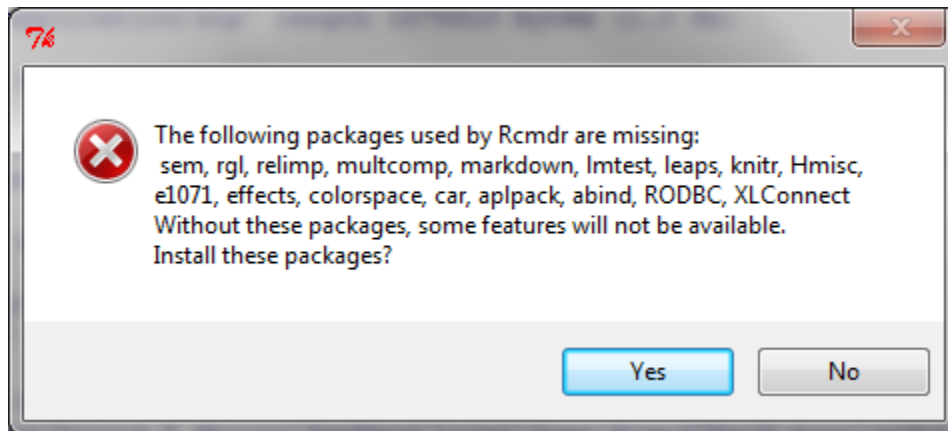
1. At the top of the screen, click on Packages
2. In the drop down menu, click on Install Package(s)
3. In the CRAN mirror, select the location closest to you; then click OK
4. In the Packages screen, click on Rcmdr, then OK
5. When prompted to create a personal library, click Yes

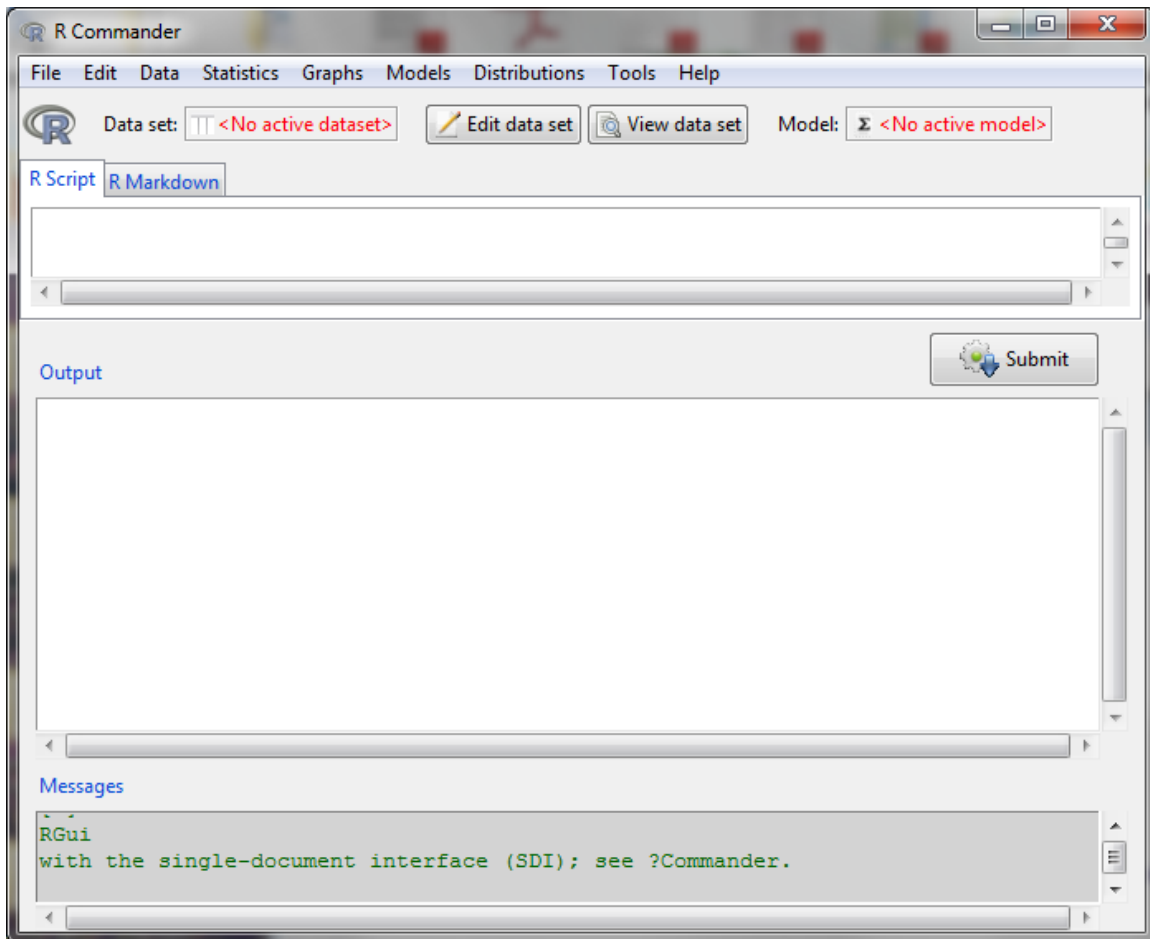


## Launch Rcmdr (R Commander)

Rcmdr is a graphical user interface (GUI) that is easier to use than the command line. To launch Rcmdr:

1. Type library(Rcmdr)
2. If you receive a warning message that some packages are missing, it will ask if you want them installed. Click Yes.
3. On the Install Missing Packages screen, click OK
4. R will install the necessary software
5. The R Commander screen will appear





## Session 7.4: Download Datasets

To access some excellent data sets used in the book “Data Mining and Business Analytics with R,” by Johannes Ledolter:

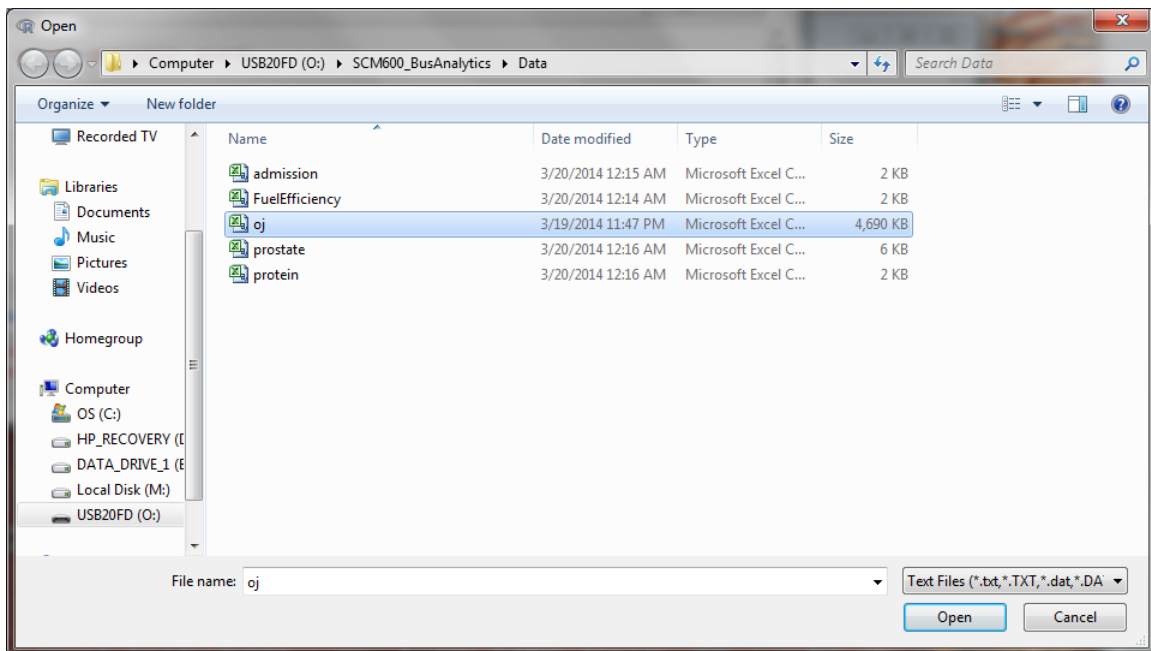
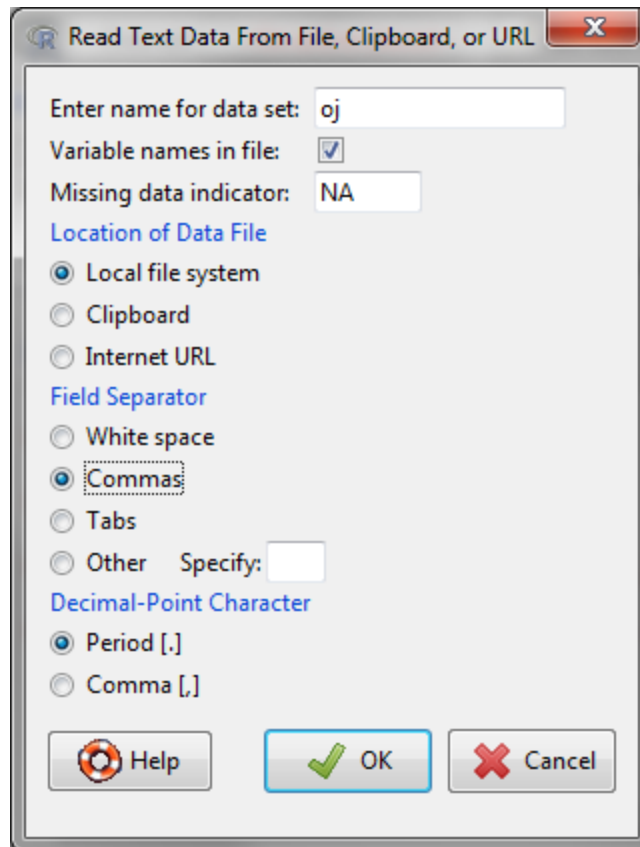
1. Go to the website:  
<http://www.biz.uiowa.edu/faculty/jledolter/DataMining>
2. Click on Data Text
3. Right click on oj.csv, then save on your computer
4. Remember where you saved the file

The Business Analytics - Week 7 oj.csv (orange juice) file can be downloaded from the course website.

## Loading Data

To load data into R:

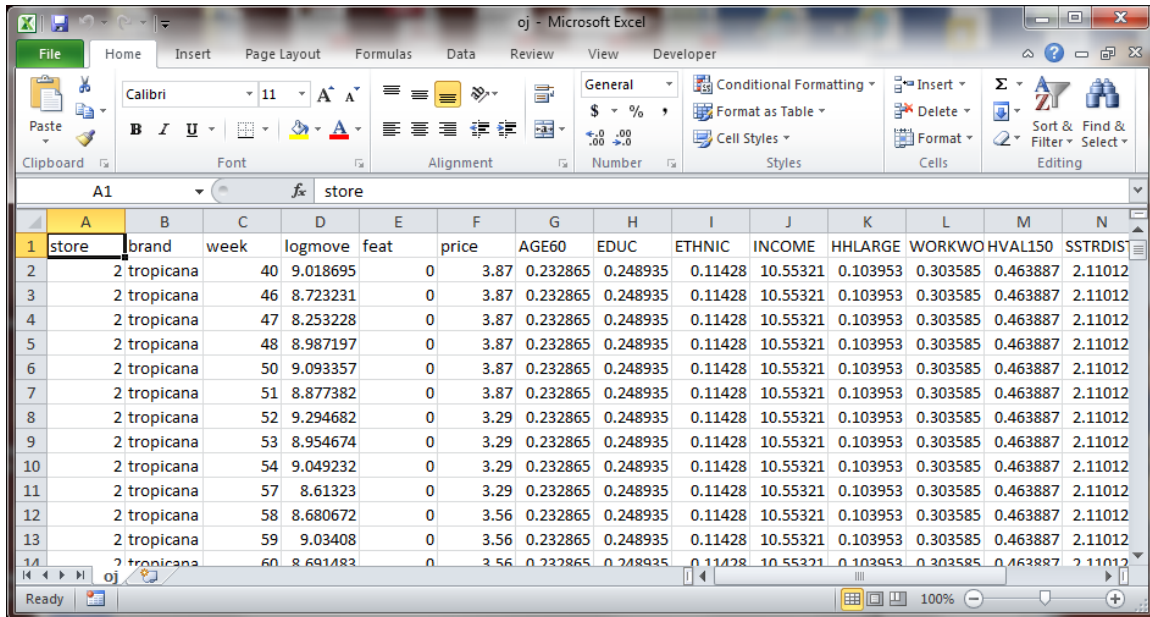
1. Click on Data at the top of the screen
2. Click on Import Data > From text file ...
3. Enter the name that you would like to use for this data set; type in oj
4. Change Field Separator to Commas, then OK
5. Click on the oj file, then Open



Note that the dataset oj has 28,947 rows and 17 columns.

## Viewing data fields

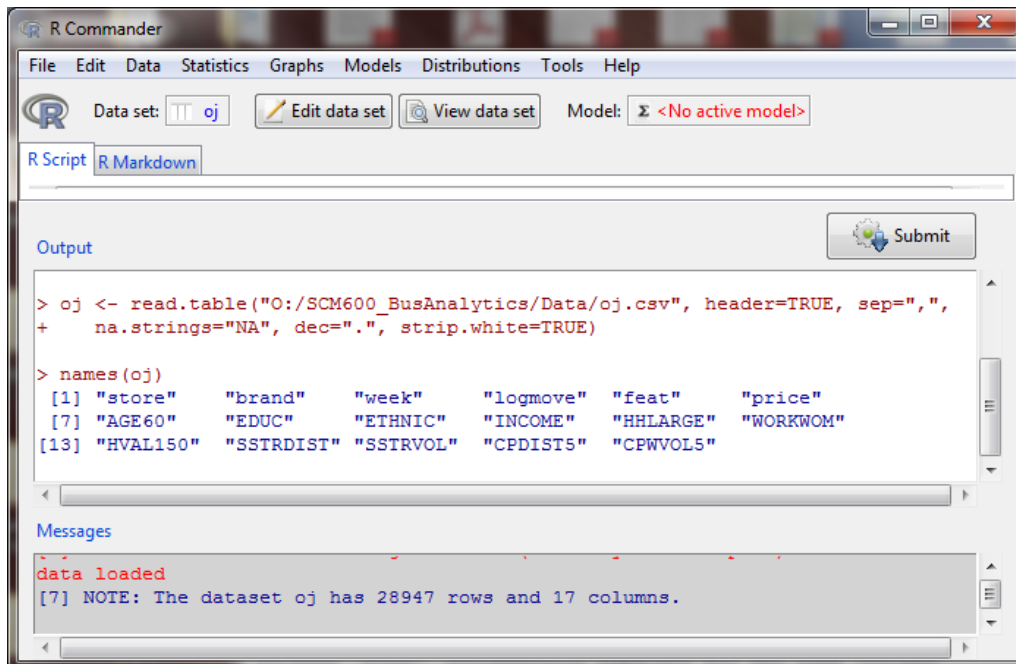
This data set lists weekly sales over 83 stores for three brands of products. Let's view the data. The easiest way to view is simply by opening the original Excel spreadsheet. Find the spreadsheet oj.csv that you downloaded and double click on it. The variable logmove is the logarithm of sales (how much product moved in a week).



store	brand	week	logmove	feat	price	AGE60	EDUC	ETHNIC	INCOME	HHLARGE	WORKWOM	HVAL150	SSTRDIST
2	2 tropicana	40	9.018695	0	3.87	0.232865	0.248935	0.11428	10.55321	0.103953	0.303585	0.463887	2.11012
3	2 tropicana	46	8.723231	0	3.87	0.232865	0.248935	0.11428	10.55321	0.103953	0.303585	0.463887	2.11012
4	2 tropicana	47	8.253228	0	3.87	0.232865	0.248935	0.11428	10.55321	0.103953	0.303585	0.463887	2.11012
5	2 tropicana	48	8.987197	0	3.87	0.232865	0.248935	0.11428	10.55321	0.103953	0.303585	0.463887	2.11012
6	2 tropicana	50	9.093357	0	3.87	0.232865	0.248935	0.11428	10.55321	0.103953	0.303585	0.463887	2.11012
7	2 tropicana	51	8.877382	0	3.87	0.232865	0.248935	0.11428	10.55321	0.103953	0.303585	0.463887	2.11012
8	2 tropicana	52	9.294682	0	3.29	0.232865	0.248935	0.11428	10.55321	0.103953	0.303585	0.463887	2.11012
9	2 tropicana	53	8.954674	0	3.29	0.232865	0.248935	0.11428	10.55321	0.103953	0.303585	0.463887	2.11012
10	2 tropicana	54	9.049232	0	3.29	0.232865	0.248935	0.11428	10.55321	0.103953	0.303585	0.463887	2.11012
11	2 tropicana	57	8.61323	0	3.29	0.232865	0.248935	0.11428	10.55321	0.103953	0.303585	0.463887	2.11012
12	2 tropicana	58	8.680672	0	3.56	0.232865	0.248935	0.11428	10.55321	0.103953	0.303585	0.463887	2.11012
13	2 tropicana	59	9.03408	0	3.56	0.232865	0.248935	0.11428	10.55321	0.103953	0.303585	0.463887	2.11012
14	2 tropicana	60	8.691483	0	3.56	0.232865	0.248935	0.11428	10.55321	0.103953	0.303585	0.463887	2.11012

Now return to R. To view the variables in R,

1. Click on Data, Active Data Set, Variables in Active Data Set



```
> oj <- read.table("O:/SCM600_BusAnalytics/Data/oj.csv", header=TRUE, sep="," ,
+ na.strings="NA", dec=".", strip.white=TRUE)

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

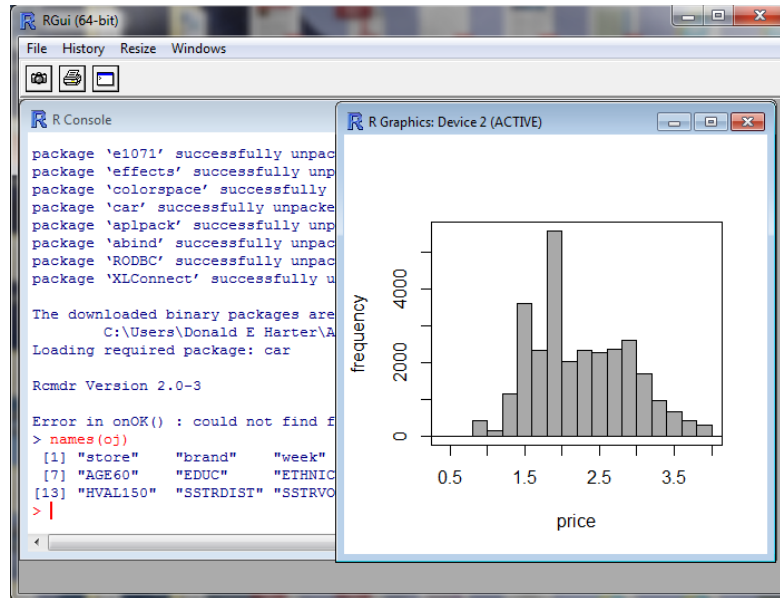
data loaded  
[7] NOTE: The dataset oj has 28947 rows and 17 columns.

Notice that R generates the command names(oj). This is the command line version.

## Session 7.5: Histograms

To create a histogram,

1. Click on Graphs, Histogram, price, OK

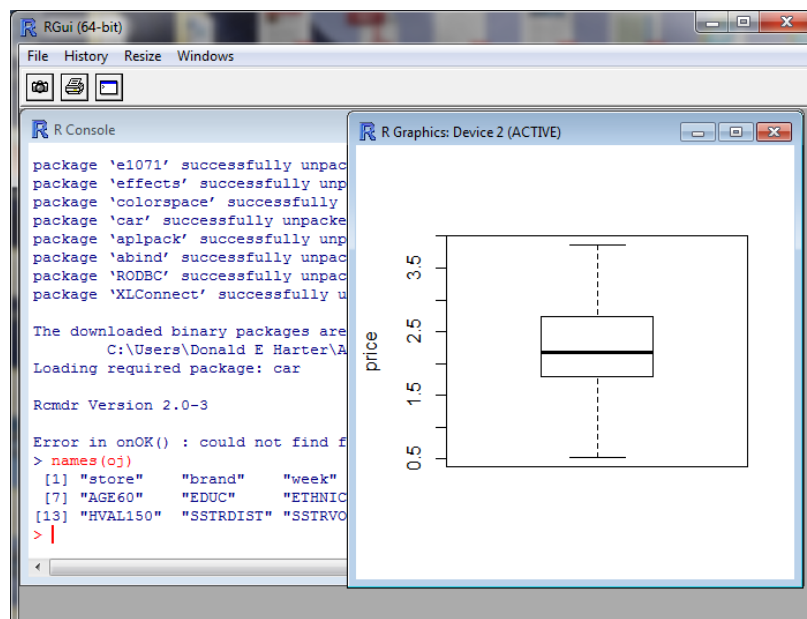


Try it again, but in the Options tab, change to Percentages (click Apply), Densities, and change number of bins to 5.

## Boxplots

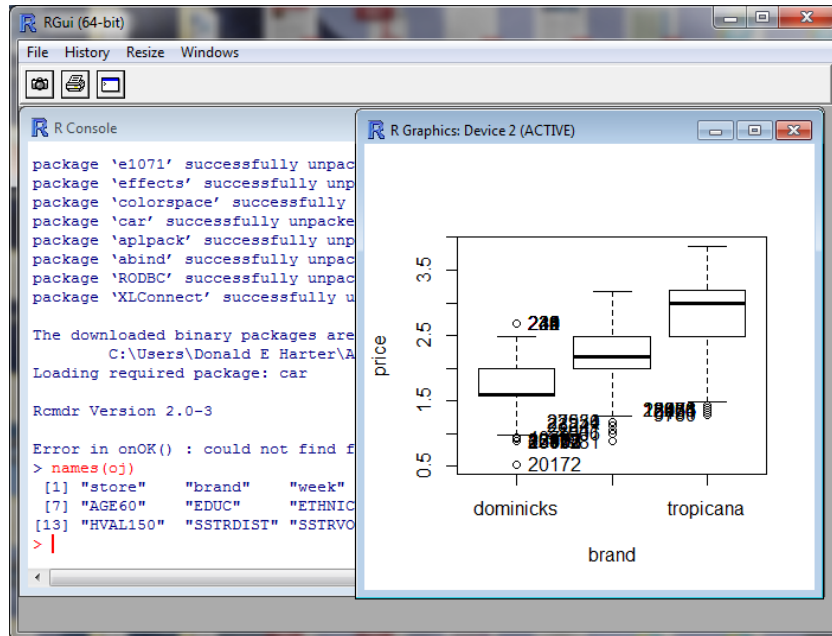
To create a boxplot,

1. Click on Graphs, Boxplot, price, OK



To create a boxplot by brand,

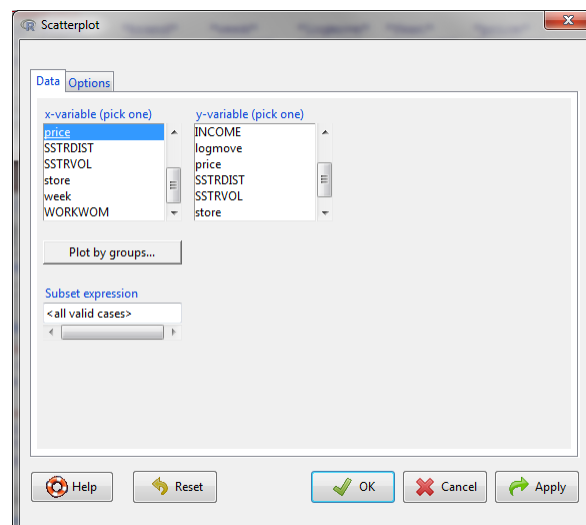
1. Click on Graphs, Boxplot, price
2. Click on Plot by groups, select brand, OK
3. Click OK



## Scatterplots

To generate a scatter plot,

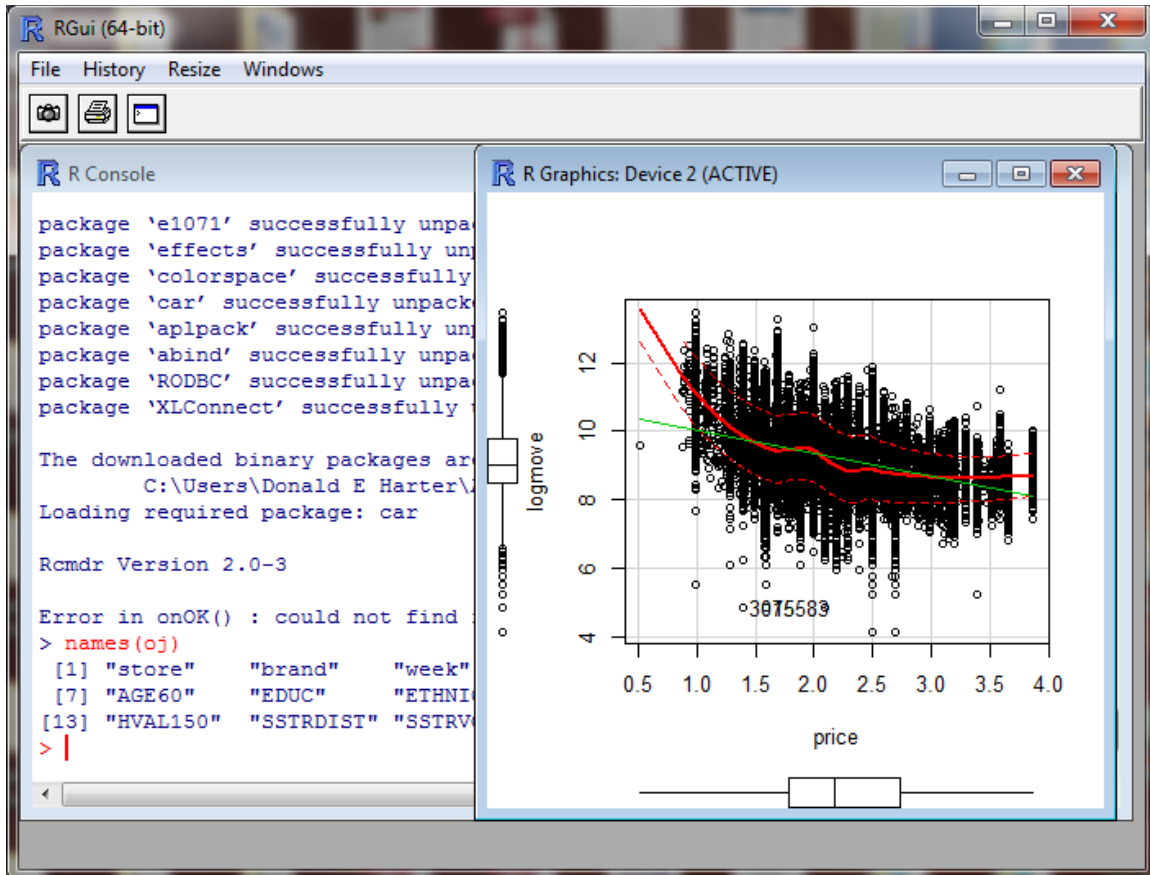
1. Click on Graphs, Scatterplot
2. Select price as the x-variable
3. Select logmove as the y-variable
4. Click on OK





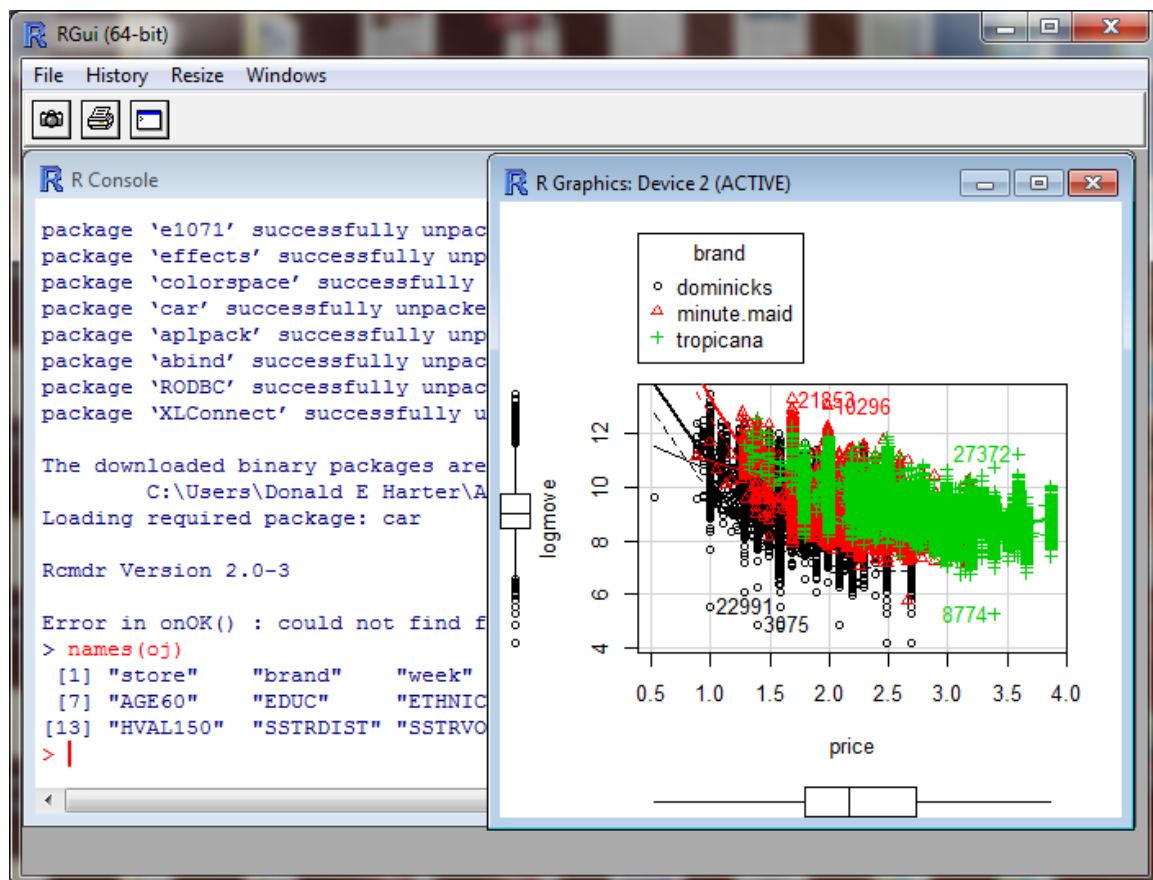
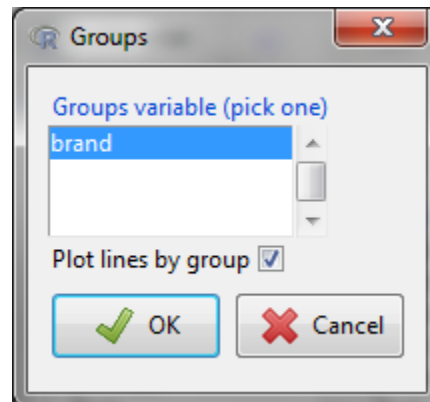
To interpret the chart:

1. The black dots are the price versus log(sales) for each time period, store and brand.
2. The green line is the linear regression line through the data
3. The red lines are the averages and plus or minus one standard deviation
4. Below and to the left of the chart are box and whisker diagrams
  - a. The center line is the average
  - b. The box is the 25%-ile to 75%-ile range
  - c. The whiskers show the range



Now generate a scatter plot by brand,

1. Click on Graphs, Scatterplot
2. Select price as the x-variable
3. Select logmove as the y-variable
4. Click on Plot by Groups, select brand, then OK.
5. Click on OK



### Interpretation

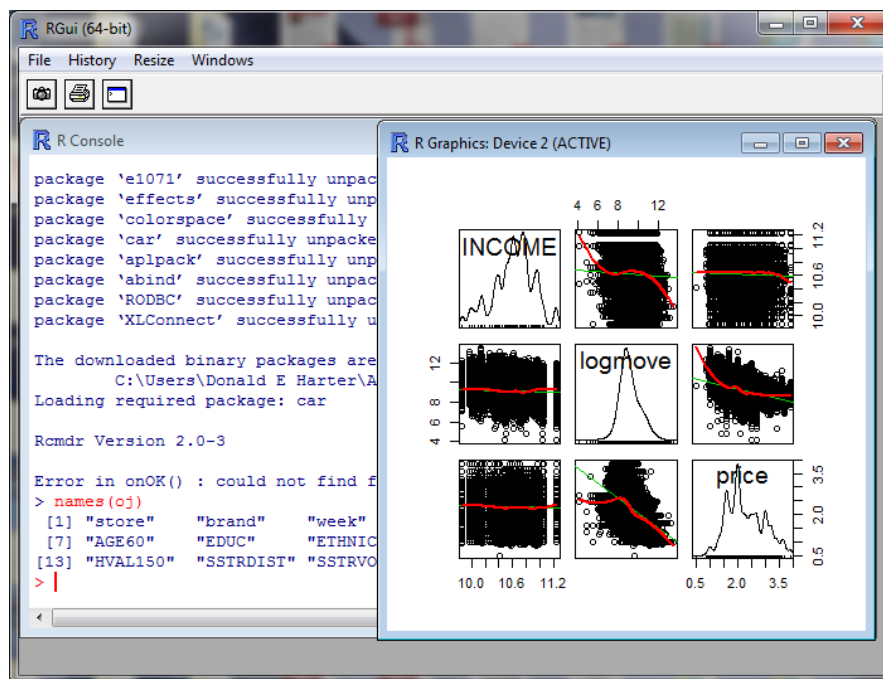
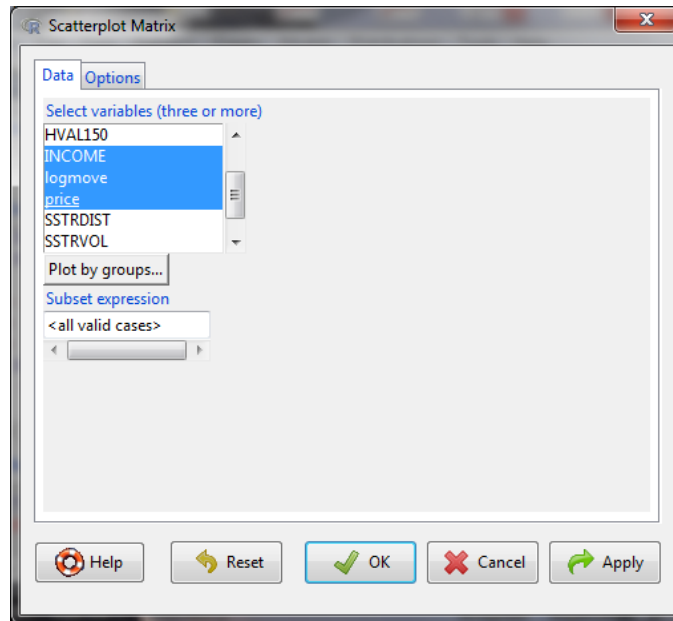
1. Brand dominicks is in black
2. Brand minute maid is in red
3. Brand Tropicana is in green

Which is the premium brand?

## Plotting pairwise scatterplots with more than two variables

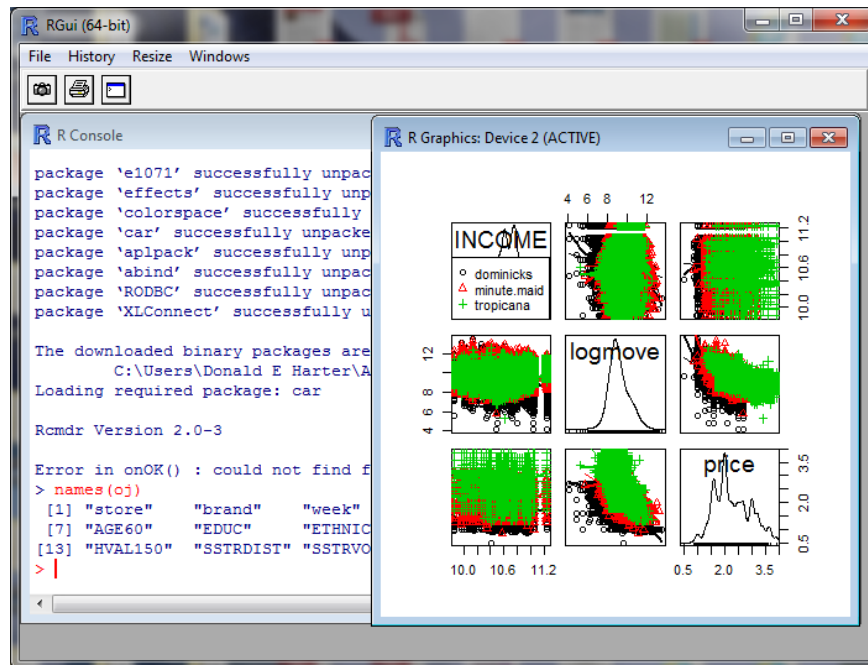
To create a matrix of scatterplots with more than two variables,

1. Click on Graphs
2. Click on Scatterplot Matrix
3. To select multiple variables, hold down the control key, then select INCOME, logmove, and price
4. Click OK



The diagonal is the distribution of data points (density function). Off-diagonal are the scatterplots for the pair of variables listed to the side and above/below the scatterplot.

Now perform a Scatterplot Matrix by Groups (brand)

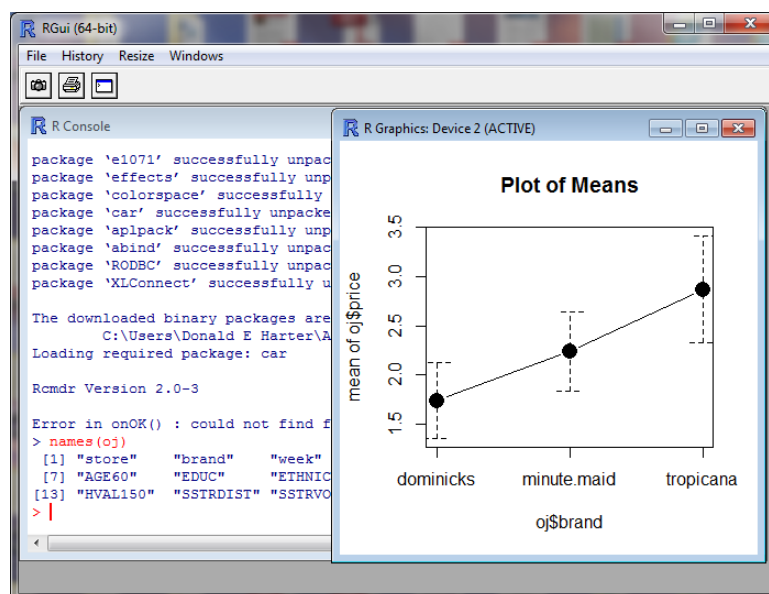


Once again, the brands are color coded.

## Plot of Means

To determine if the different brands have different prices, on average, plot the means:

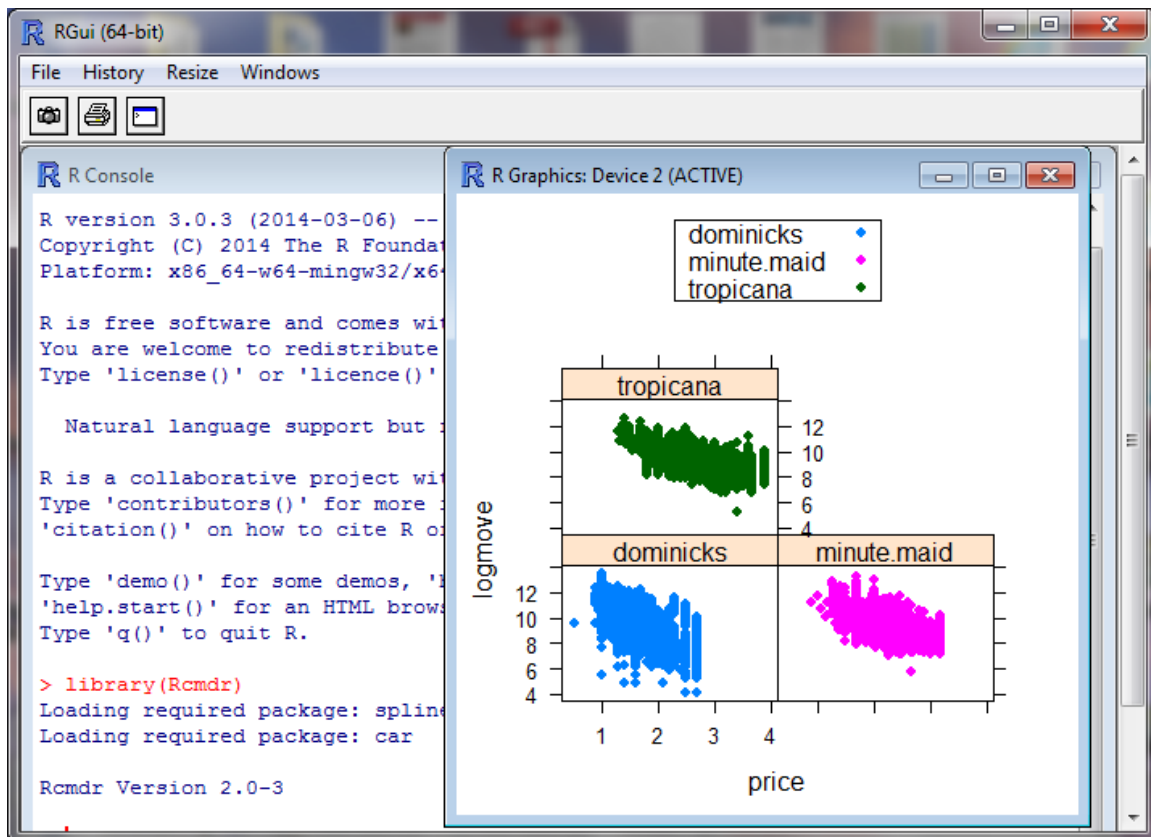
1. Click on Graphs, Plot of Means
2. Select price
3. In the Options tab, click on standard deviations, then OK



## XY Plots

Now let's generate XY plots by brand.

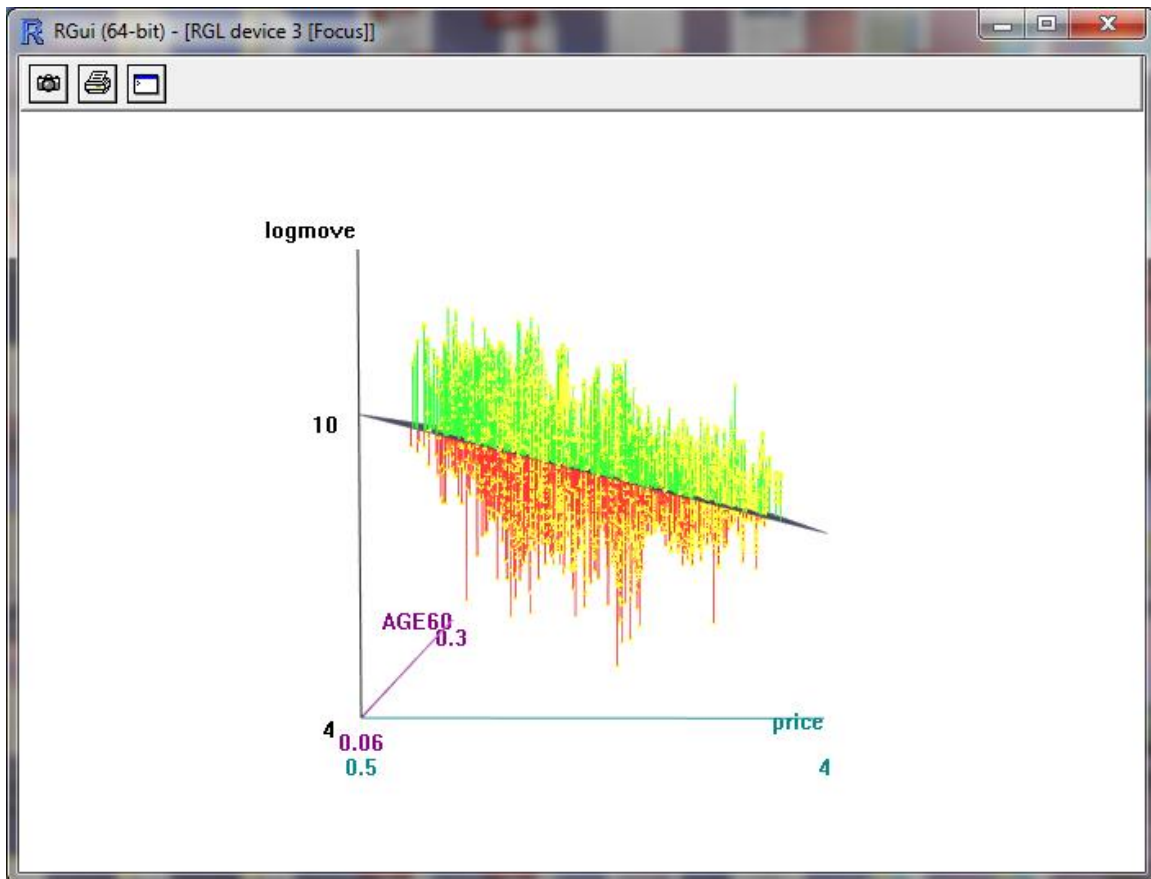
1. Click on Graphs, XY Conditioning Plot
2. Select price for the explanatory variable
3. Select logmove for the response variable
4. Click on brand for each
5. Click OK



## Session 7.6: 3D Graphs

To generate 3D graphs,

1. Graphs, 3D Graph, 3D Scatterplot
2. Select AGE60 and price as explanatory variables by holding down the control key, then clicking on AGE60 and price
3. Select logmove (log of sales) as the response variable
4. Click on the Options tab and check the box "Linear least-squares"
5. Click OK
6. Expand the window by clicking on the box in the upper right corner of your graph
7. Rotate the graph by clicking on the graph with your mouse, hold the mouse button down, and move

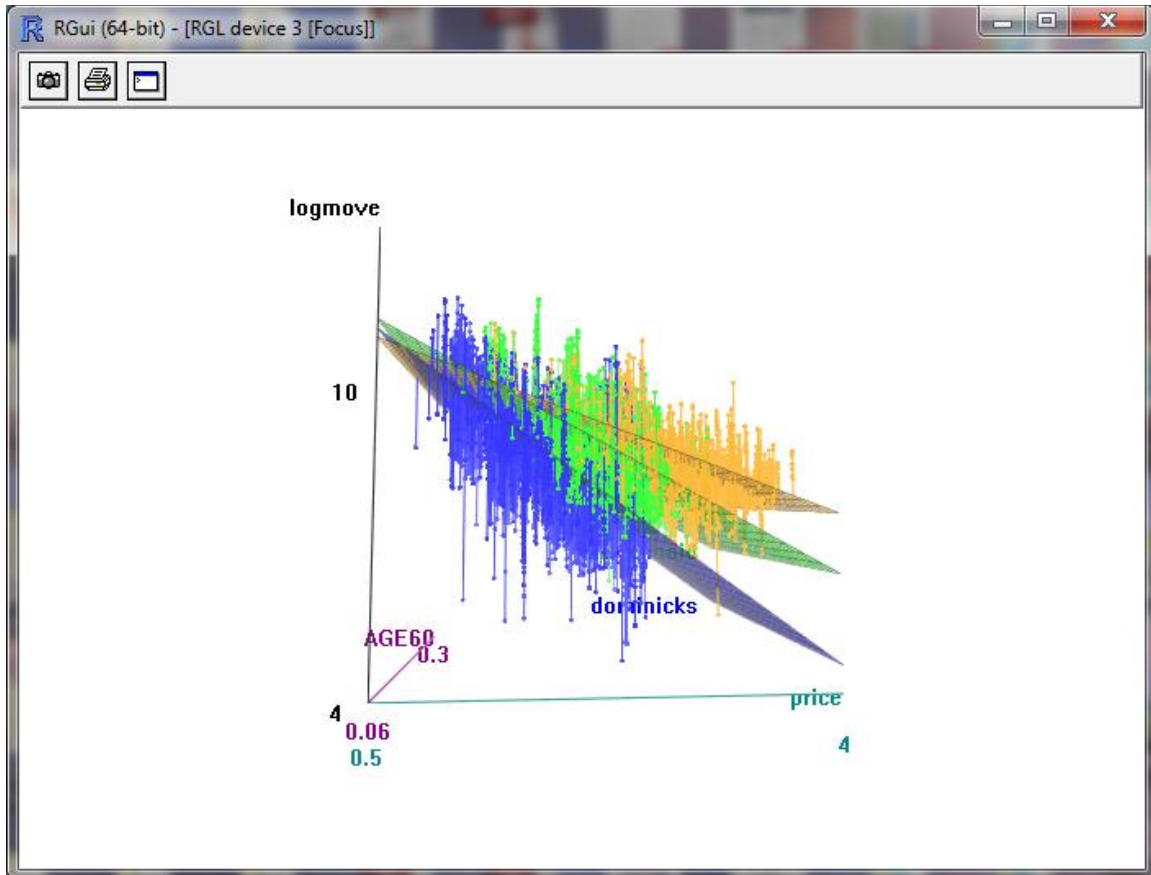


Does price affect sales?

Does age affect sales?

To generate 3D graphs by brand,

1. Graphs, 3D Graph, 3D Scatterplot
2. Select AGE60 and price as explanatory variables by holding down the control key, then clicking on AGE60 and price
3. Select logmove (log of sales) as the response variable
4. Click Plot by groups, select brand, then OK
5. Click OK
6. Expand the window by clicking on the box in the upper right corner of your graph
7. Rotate the graph by clicking on the graph with your mouse, hold the mouse button down, and move



Which brand is more sensitive to price?

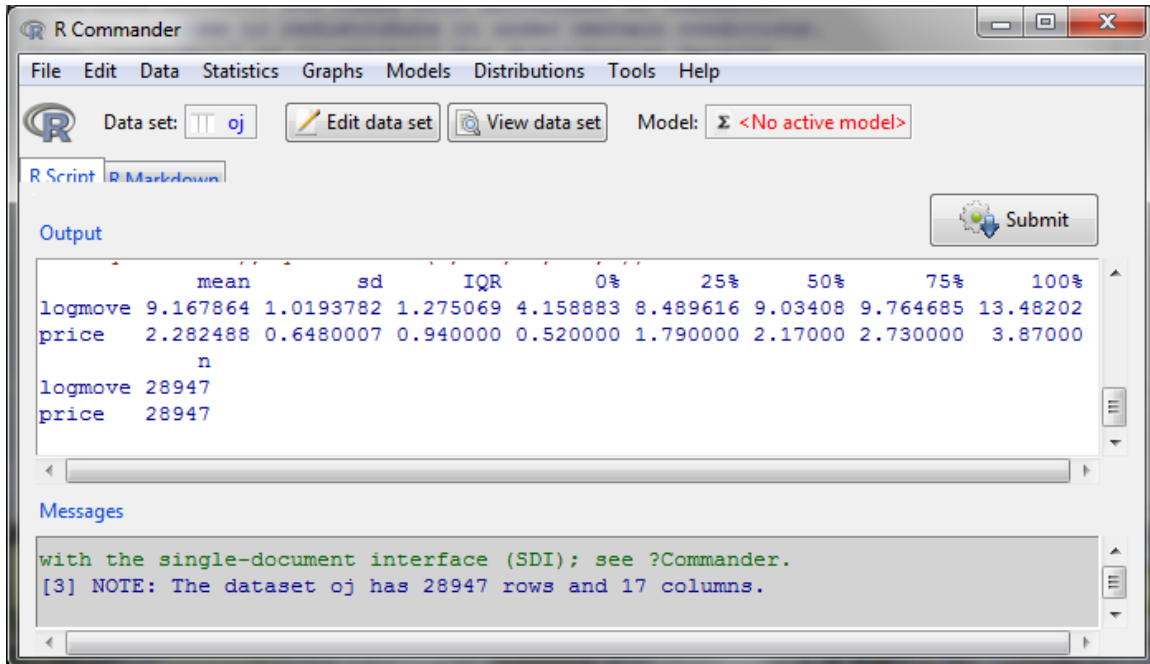
### Saving Graphs

You can save graphs by clicking Graphs, Save Graph to File, then select the type of file.

## Session 7.7: Statistical Summaries

The mean, standard deviation and quartiles can be found by:

1. Click on Statistics, Summaries, Numerical Summaries
2. Select logmove, price by holding down the control key
3. Click OK



To categorize by brand, the mean, standard deviation and quartiles can be found by:

1. Click on Statistics, Summaries, Numerical Summaries
2. Select logmove, price by holding down the control key
3. Click Summarize by Groups, then OK
4. Click OK



R Commander

File Edit Data Statistics Graphs Models Distributions Tools Help

Data set: **oj** Edit data set View data set Model: **<No active model>**

R Script R Markdown

```

numSummary(oj[,c("logmove", "price")], statistics=c("mean", "sd", "IQR",
  "quantiles"), quantiles=c(0,.25,.5,.75,1))
numSummary(oj[,c("logmove", "price")], groups=oj$brand, statistics=c("mean",
  "sd", "IQR", "quantiles"), quantiles=c(0,.25,.5,.75,1))
numSummary(oj[,c("logmove", "price")], groups=oj$brand, statistics=c("mean",
  "sd", "IQR", "quantiles"), quantiles=c(0,.25,.5,.75,1))

```

Output

Variable: logmove

	mean	sd	IQR	0%	25%	50%	75%
dominicks	9.174831	1.1929370	1.5619512	4.158883	8.392990	9.121728	9.954941
minute.maid	9.217278	0.9852867	1.3523928	5.768321	8.476371	9.026418	9.828764
tropicana	9.111483	0.8473800	0.9685592	5.257495	8.565602	8.987197	9.534161

100% n

dominicks	13.48202	9649
minute.maid	13.29018	9649
tropicana	12.57205	9649

Variable: price

	mean	sd	IQR	0%	25%	50%	75%	100%	n
dominicks	1.735809	0.3858380	0.41	0.52	1.58	1.59	1.99	2.69	9649
minute.maid	2.241162	0.4045146	0.50	0.88	1.99	2.17	2.49	3.17	9649
tropicana	2.870493	0.5485578	0.70	1.29	2.49	2.99	3.19	3.87	9649

Messages

```

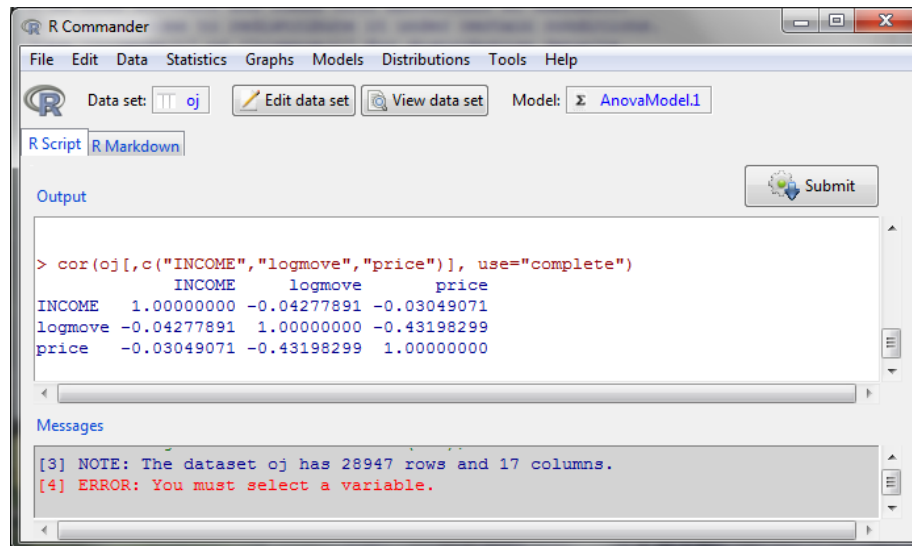
[3] NOTE: The dataset oj has 28947 rows and 17 columns.
[4] ERROR: You must select a variable.

```

## Session 7.8: Correlation

To generate a correlation matrix,

1. Click on Statistics, Summaries, Correlation Matrix
2. Hold down the control key and select INCOME, logmove, price
3. Click OK



The screenshot shows the R Commander interface. The 'Data set' is 'oj'. The 'Model' is 'AnovaModel.1'. The 'Output' pane displays the following R code and its result:

```
> cor(oj[,c("INCOME", "logmove", "price")], use="complete")
```

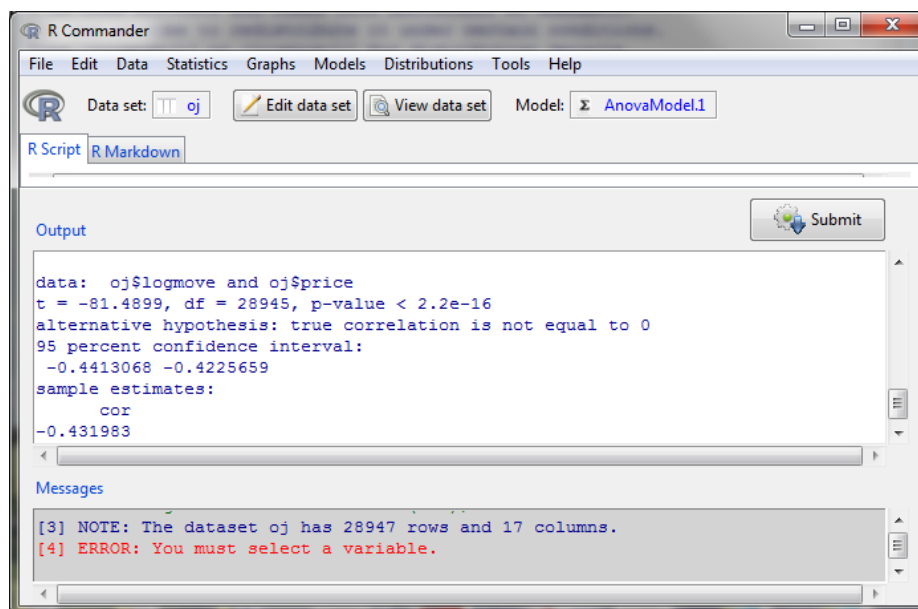
	INCOME	logmove	price
INCOME	1.00000000	-0.04277891	-0.03049071
logmove	-0.04277891	1.00000000	-0.43198299
price	-0.03049071	-0.43198299	1.00000000

The 'Messages' pane shows the following messages:

```
[3] NOTE: The dataset oj has 28947 rows and 17 columns.  
[4] ERROR: You must select a variable.
```

The matrix shows the correlation, but not the statistical significance. To calculate significance,

1. Click on Statistics, Summaries, Correlation Test
2. Select both logmove and price
3. Click OK



The screenshot shows the R Commander interface. The 'Data set' is 'oj'. The 'Model' is 'AnovaModel.1'. The 'Output' pane displays the following R code and its result:

```
data: oj$logmove and oj$price  
t = -81.4899, df = 28945, p-value < 2.2e-16  
alternative hypothesis: true correlation is not equal to 0  
95 percent confidence interval:  
-0.4413068 -0.4225659  
sample estimates:  
cor  
-0.431983
```

The 'Messages' pane shows the following messages:

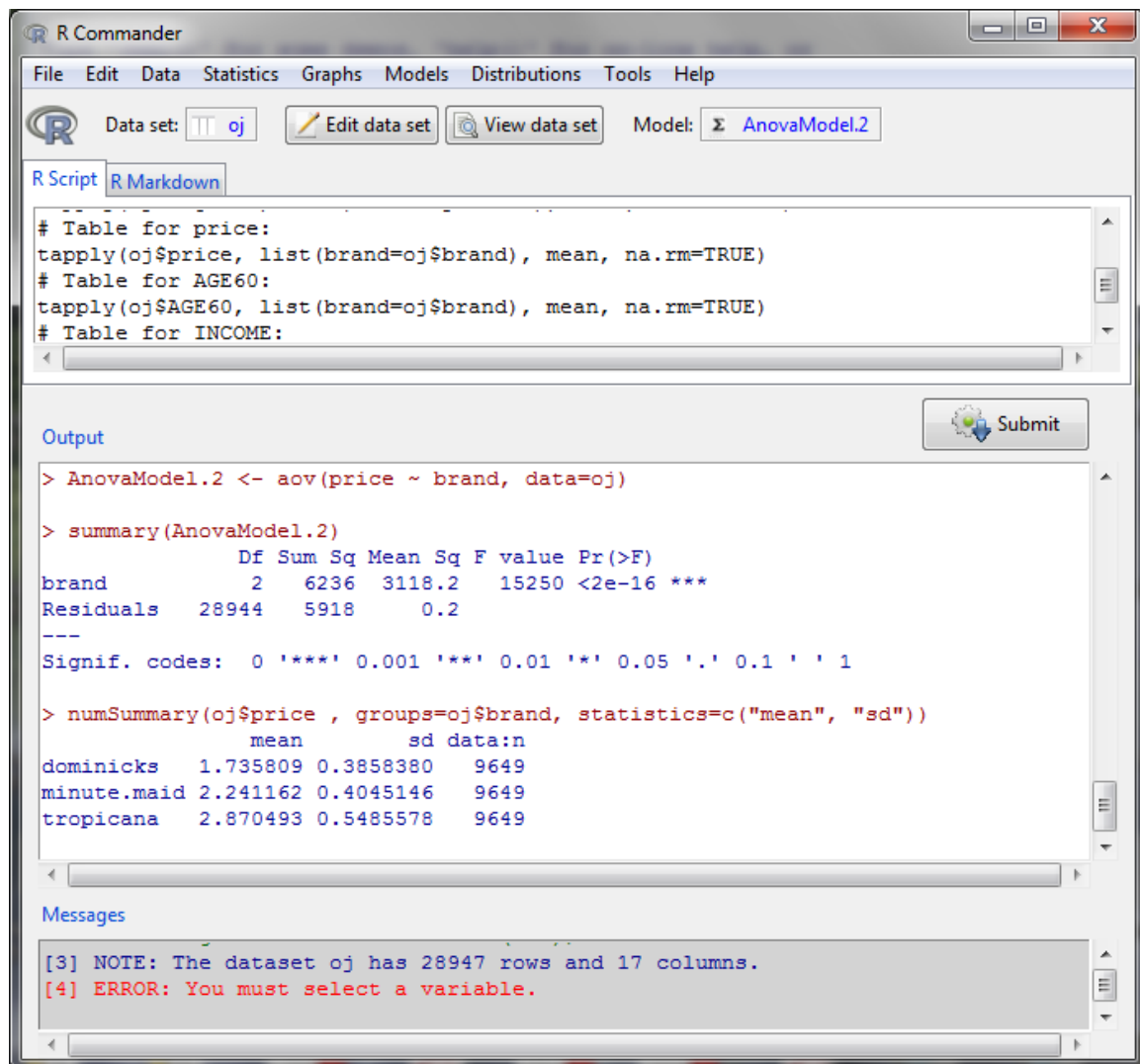
```
[3] NOTE: The dataset oj has 28947 rows and 17 columns.  
[4] ERROR: You must select a variable.
```

The p-value is less than 0.05, so the correlation is statistically significant.

## Session 7.9: ANOVA

ANOVA stands for Analysis of Variance. It compares the means of several groups to determine if the groups are different. Let's see if prices are different across brands.

1. Click on Statistics, Means, One-way ANOVA
2. For the response variable, click on price
3. Click OK



The screenshot shows the R Commander window with the 'oj' dataset selected. The R Script pane contains the following code:

```
# Table for price:
tapply(oj$price, list(brand=oj$brand), mean, na.rm=TRUE)
# Table for AGE60:
tapply(oj$AGE60, list(brand=oj$brand), mean, na.rm=TRUE)
# Table for INCOME:
```

The Output pane shows the results of the ANOVA test:

```
> AnovaModel.2 <- aov(price ~ brand, data=oj)
> summary(AnovaModel.2)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
brand	2	6236	3118.2	15250	<2e-16 ***
Residuals	28944	5918	0.2		

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

```
> numSummary(oj$price, groups=oj$brand, statistics=c("mean", "sd"))
```

	mean	sd	data:n
dominicks	1.735809	0.3858380	9649
minute.maid	2.241162	0.4045146	9649
tropicana	2.870493	0.5485578	9649

The Messages pane shows the following messages:

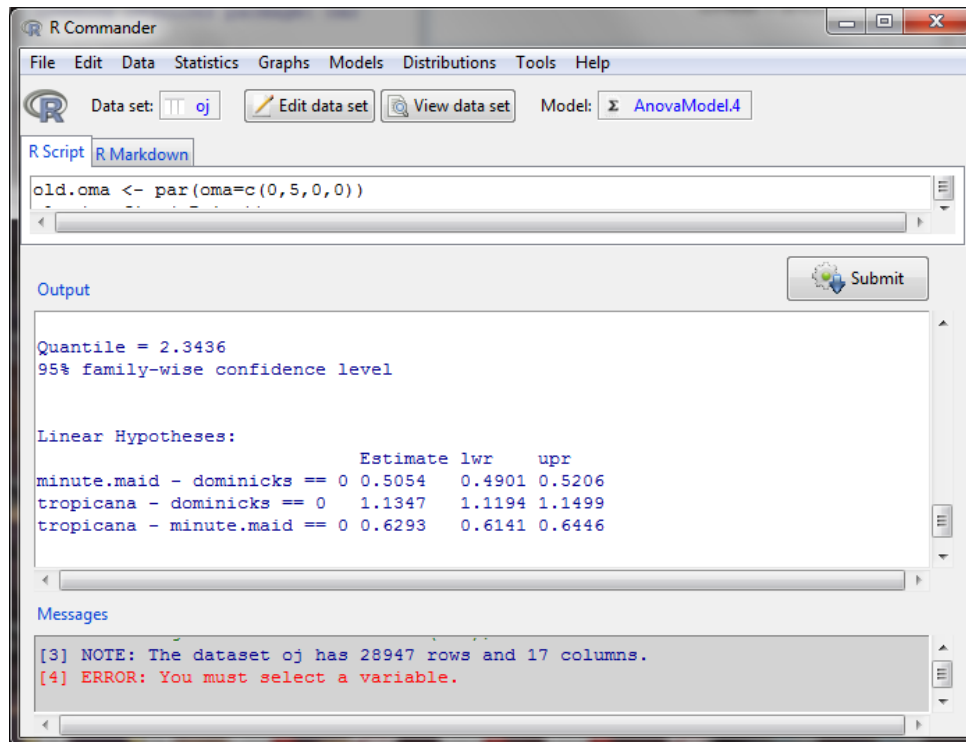
```
[3] NOTE: The dataset oj has 28947 rows and 17 columns.
[4] ERROR: You must select a variable.
```

The F-statistic p-value [Pr(>F)] is less than 0.05. That means that one of the brands has a price that is statistically different from the others.

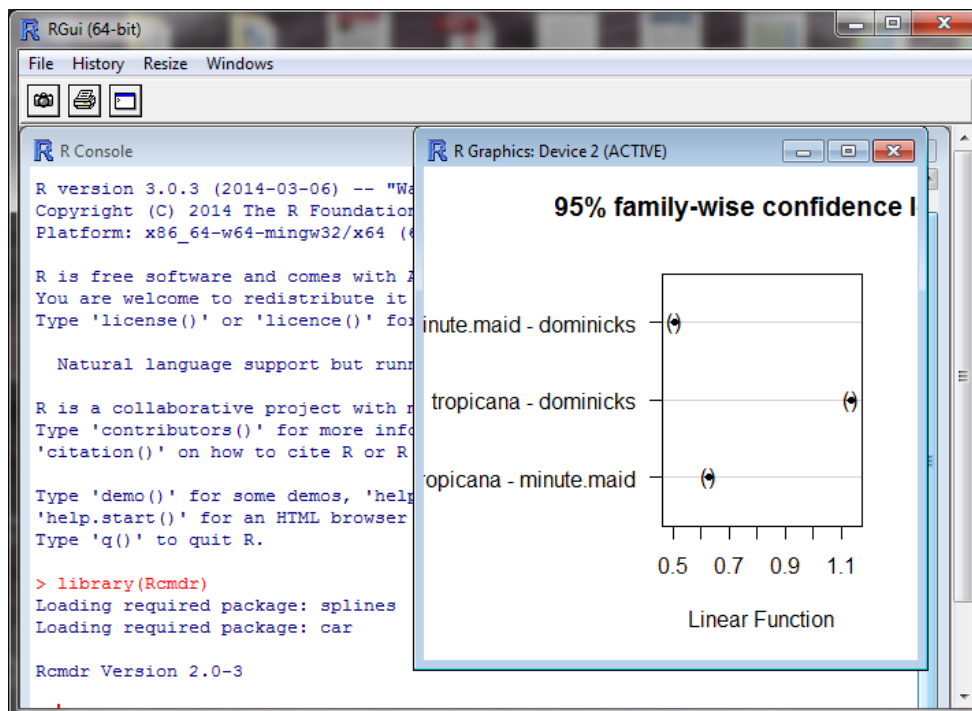
Let's now perform a pairwise comparison.

1. Click on Statistics, Means, One-way ANOVA
2. For the response variable, click on price
3. Check the box Pairwise comparison of means
4. Click OK

The pairwise comparison estimates the price and confidence interval for each brand (lower and upper interval values). Do they overlap?



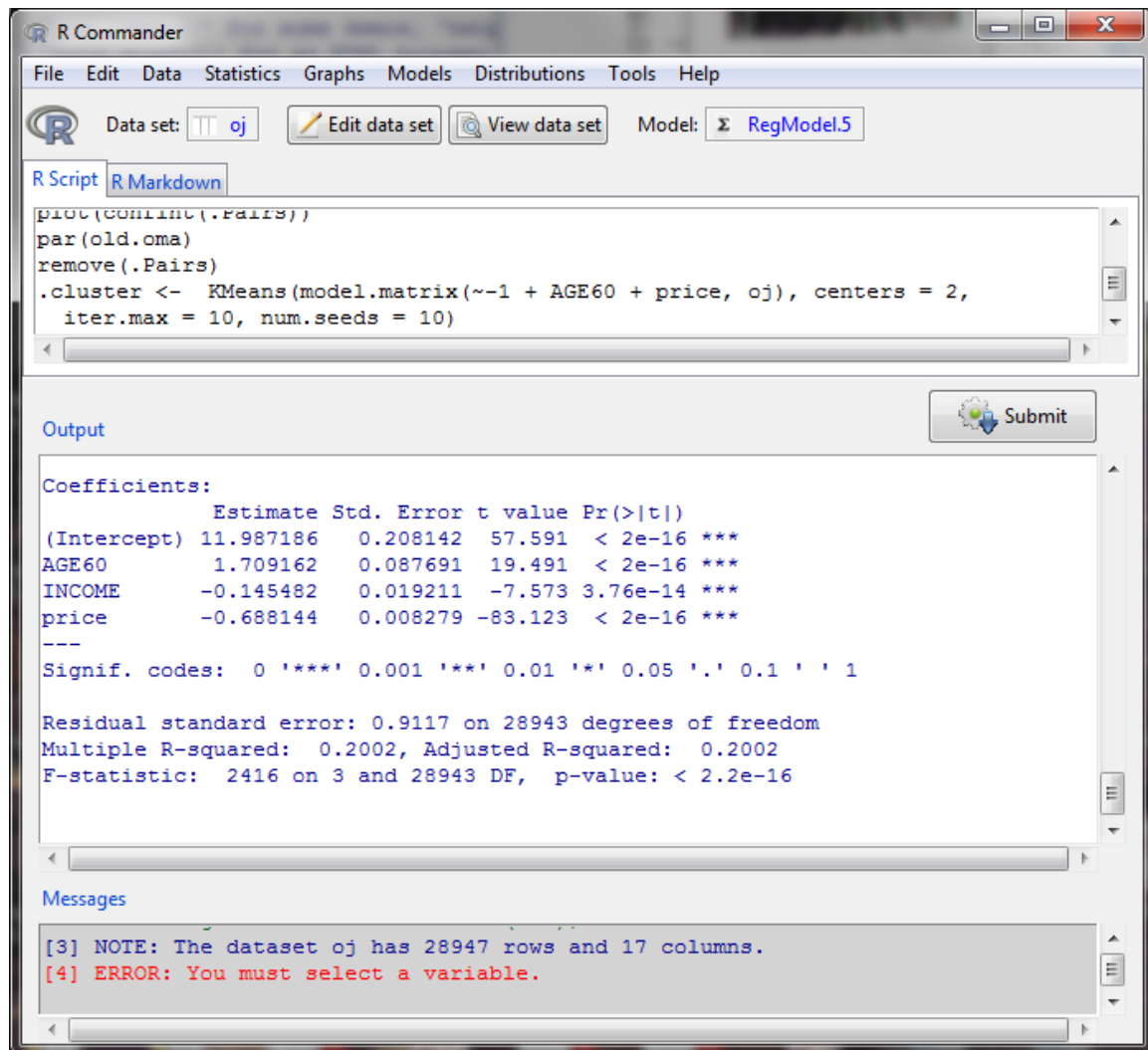
The graph portrays the estimate of price for each brand and the confidence interval.



## Session 7.10: Regression

Linear regression of the log of sales against age, income and price can be performed by:

1. Click on Statistics, Fit Models, Linear Regression
2. For response variable, click on logmove
3. For explanatory variables, hold down the control key and click on AGE60, INCOME, price
4. Click OK



The screenshot shows the R Commander window with the following components:

- Menu Bar:** File, Edit, Data, Statistics, Graphs, Models, Distributions, Tools, Help.
- Toolbar:** Data set: oj, Edit data set, View data set, Model: RegModel.5.
- R Script:**

```
plot(logmove(.Pairs))
par(old.oma)
remove(.Pairs)
.cluster <- KMeans(model.matrix(~-1 + AGE60 + price, oj), centers = 2,
  iter.max = 10, num.seeds = 10)
```
- Output:**

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	11.987186	0.208142	57.591	< 2e-16 ***
AGE60	1.709162	0.087691	19.491	< 2e-16 ***
INCOME	-0.145482	0.019211	-7.573	3.76e-14 ***
price	-0.688144	0.008279	-83.123	< 2e-16 ***

---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9117 on 28943 degrees of freedom  
Multiple R-squared: 0.2002, Adjusted R-squared: 0.2002  
F-statistic: 2416 on 3 and 28943 DF, p-value: < 2.2e-16
- Messages:**

```
[3] NOTE: The dataset oj has 28947 rows and 17 columns.
[4] ERROR: You must select a variable.
```

Is the equation statistically significant?

How much of the variability in the log of sales is explained by the explanatory variables?

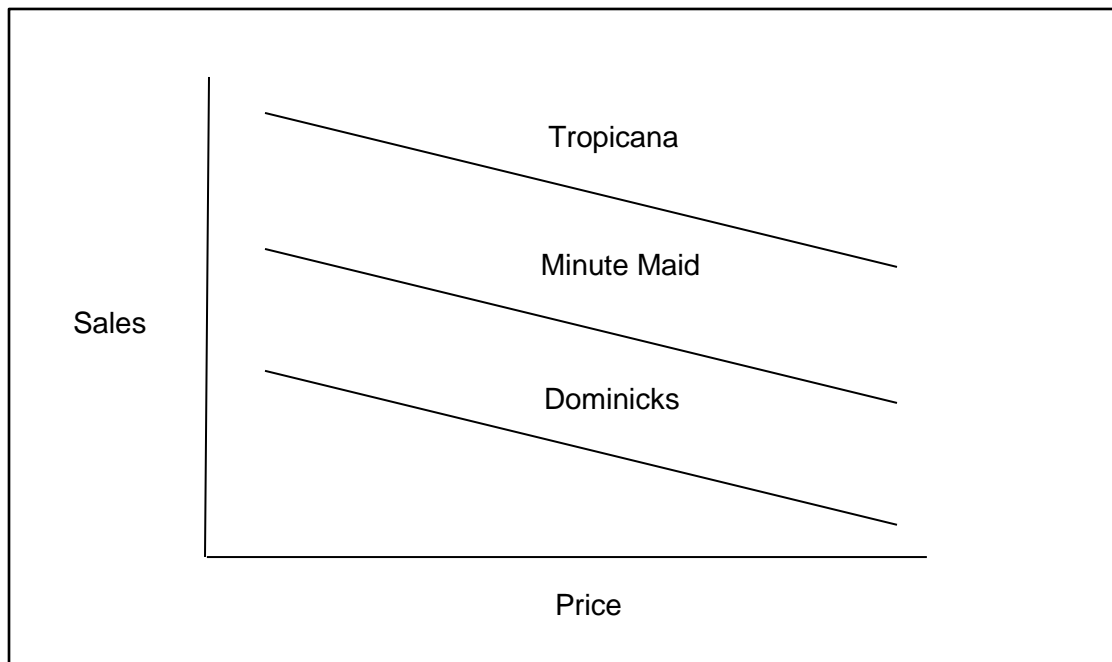
Which explanatory variables are statistically significant?

How does each explanatory variable affect sales? Which affect it positively and which negatively?  
How do you interpret this (what does it really mean)?

## Session 7.11: Regression with Dummy Variables

A dummy variable in a regression can assist in determining if the constant changes when the brand changes.

In the example below, the slopes of the lines are the same, which means the coefficients in the regression are the same. However, the intercept terms are different. This difference in intercept term is represented by a dummy variable (zero or one) where the coefficient in the dummy variable is the change in the intercept. One brand is identified as the base case, e.g., in this case it's Dominicks. The coefficient of the dummy variable for Tropicana is the shift in the intercept for Tropicana; similarly the coefficient of the dummy variable for Minute Maid is the shift in the intercept for Minute Maid.



To perform this more sophisticated regression,

1. Click on Statistics, Fit Models, Linear Model
2. Click Reset
3. Double click on logmove
4. Double click on AGE60
5. Double click on INCOME
6. Double click on price
7. Double click on brand (notice it says it's a factor)
8. Click on OK

The screenshot shows the R Commander window with the following content:

**R Script:**

```
LinearModel.6 <- lm(logmove ~ AGE60 + INCOME + price +AGE60*price, data=oj)
summary(LinearModel.6)
LinearModel.7 <- lm(logmove ~ AGE60 + INCOME + price +EDUC, data=oj)
summary(LinearModel.7)
LinearModel.8 <- lm(logmove ~ AGE60 + INCOME + price + EDUC*price, data=oj)
summary(LinearModel.8)
LinearModel.9 <- lm(logmove ~ AGE60 + INCOME + price +brand, data=oj)
summary(LinearModel.9)
LinearModel.10 <- lm(logmove ~ AGE60 +INCOME + price + brand, data=oj)
summary(LinearModel.10)
```

**Output:**

```
lm(formula = logmove ~ AGE60 + INCOME + price + brand, data = oj)

Residuals:
    Min       1Q   Median       3Q      Max
-4.8363 -0.5416 -0.0591  0.5025  3.2653

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   13.12952    0.18533   70.85  <2e-16 ***
AGE60          1.96658    0.07792   25.24  <2e-16 ***
INCOME        -0.18339    0.01706  -10.75  <2e-16 ***
price         -1.35277    0.01055 -128.26  <2e-16 ***
brand[T.minute.maid] 0.72607    0.01282   56.66  <2e-16 ***
brand[T.tropicana]  1.47162    0.01670   88.10  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.8095 on 28941 degrees of freedom
Multiple R-squared:  0.3695, Adjusted R-squared:  0.3694
F-statistic: 3393 on 5 and 28941 DF, p-value: < 2.2e-16
```

**Messages:**

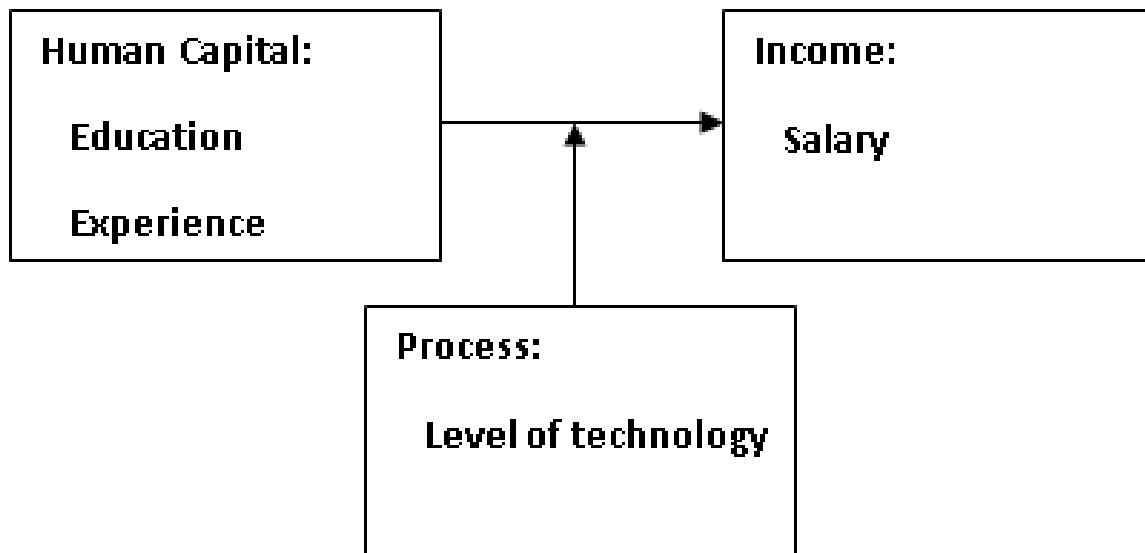
```
[3] NOTE: The dataset oj has 28947 rows and 17 columns.
[4] ERROR: You must select a variable.
```

In the Output section, the intercept is 13.12952. This is the intercept for dominicks. The intercept for minute maid is  $13.12952 + 0.72607$ . The intercept for Tropicana is  $13.12952 + 1.47162$ .

This means, all else being equal, the log of sales is highest for Tropicana.

### Session 7.12: Moderating effects (interactions of price and brand)

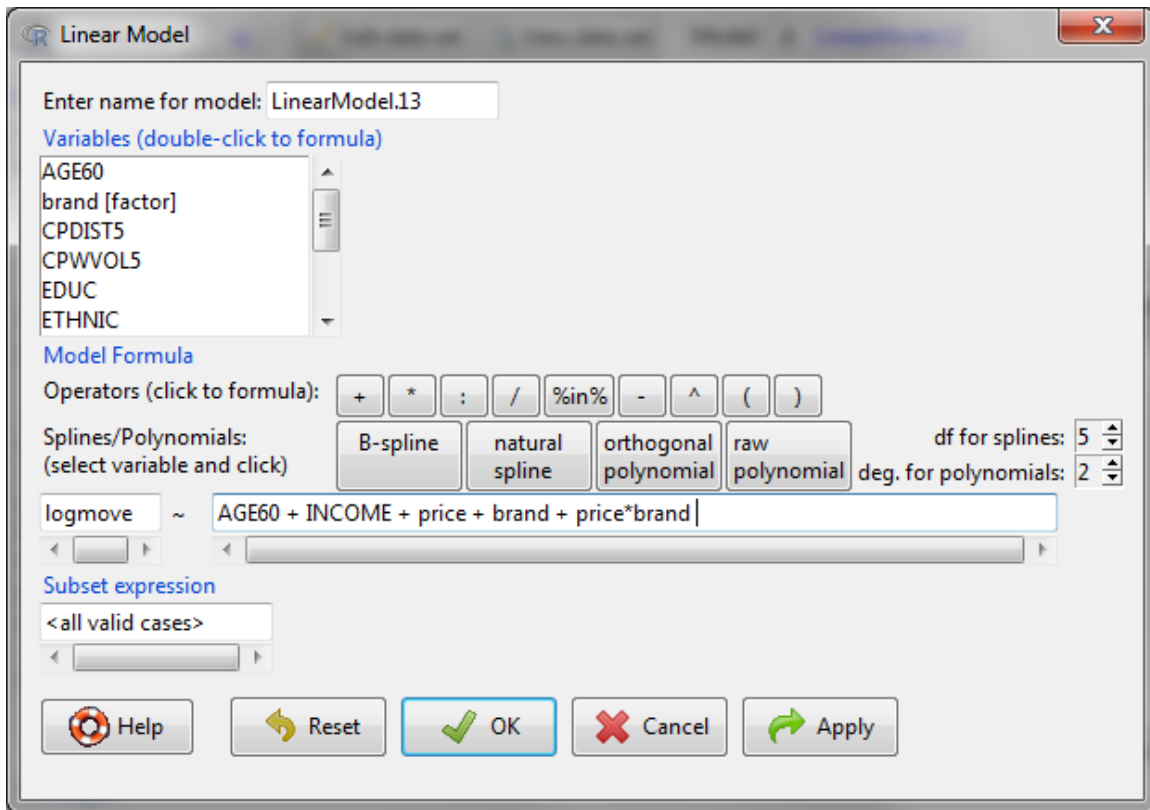
The interaction of two variables can be modeled and is called a moderating effect. A moderating effect magnifies the effect of one variable by changing another variable. In the following example, salary is dependent on education and experience of an individual. The technology that a person has available can enhance the effect of education and experience on their salary. Think of this as a catalyst which can increase or decrease the effect of a variable. This moderation is represented by an interaction, or multiplying two variables together.



In the previous example, we examined if the intercept is different for each brand. It's possible that the slope of the relationship between price and sales varies by brand. To test this, we create what is called an interaction term. An interaction is two variables multiplied together.

1. Click on Statistics, Fit Models, Linear Model
2. Click Reset
3. Double click on logmove
4. Double click on AGE60
5. Double click on INCOME
6. Double click on price
7. Double click on brand (notice it says it's a factor)
8. Double click on price (again)
9. Click on the multiplication sign (\*)
10. Click on brand
11. Click on OK





On the next page, the coefficient on price is -1.94480. That means as price increases, the log of sales declines. But since we included an interaction term, this only applies to dominicks. We need to include the price\*brand effect for minute maid and Tropicana. For minute maid, the coefficient is  $-1.94480 + 0.47545 = -1.46935$ . For Tropicana, the coefficient is  $-1.94480 + 0.94817 = -0.99663$ .

Which brand is more sensitive to price?

R Commander

File Edit Data Statistics Graphs Models Distributions Tools Help

Data set: **oj** Edit data set View data set Model: **LinearModel.13**

R Script R Markdown

```
summary(LinearModel.7)
LinearModel.8 <- lm(logmove ~ AGE60 + INCOME + price + EDUC*price, data=oj)
summary(LinearModel.8)
LinearModel.9 <- lm(logmove ~ AGE60 + INCOME + price +brand, data=oj)
summary(LinearModel.9)
LinearModel.10 <- lm(logmove ~ AGE60 +INCOME + price + brand, data=oj)
summary(LinearModel.10)
LinearModel.12 <- lm(logmove ~ AGE60 + INCOME + price + brand + price*brand,
  data=oj)
summary(LinearModel.12)
LinearModel.13 <- lm(logmove ~ AGE60 + INCOME + price + brand + price*brand,
  data=oj)
summary(LinearModel.13)
```

Output

Call:

```
lm(formula = logmove ~ AGE60 + INCOME + price + brand + price *
    brand, data = oj)
```

Residuals:

Min	1Q	Median	3Q	Max
-5.1482	-0.5318	-0.0519	0.4937	3.5030

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	14.20227	0.18385	77.248	<2e-16 ***
AGE60	1.96066	0.07606	25.776	<2e-16 ***
INCOME	-0.18754	0.01666	-11.259	<2e-16 ***
price	-1.94480	0.02086	-93.218	<2e-16 ***
brand[T.minute.maid]	-0.04030	0.05854	-0.689	0.491
brand[T.tropicana]	-0.57834	0.05668	-10.204	<2e-16 ***
price:brand[T.minute.maid]	0.47545	0.02882	16.500	<2e-16 ***
price:brand[T.tropicana]	0.94817	0.02549	37.194	<2e-16 ***

---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7902 on 28939 degrees of freedom  
Multiple R-squared: 0.3992, Adjusted R-squared: 0.3991  
F-statistic: 2747 on 7 and 28939 DF, p-value: < 2.2e-16

Messages

```
[3] NOTE: The dataset oj has 28947 rows and 17 columns.
[4] ERROR: You must select a variable.
```