**R: Basics**

**Session 7.3: Installation of R**

For these exercises, download the files:

“Business Analytics – Week 7 Instructions.doc”

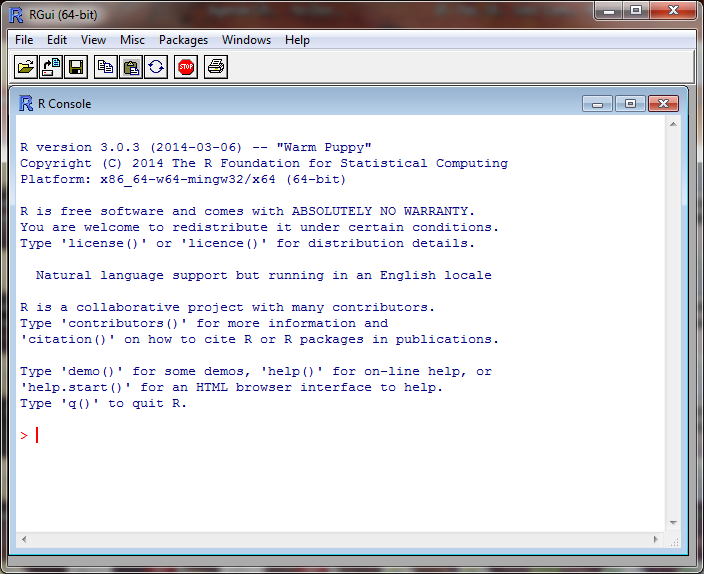
“Business Analytics – Week 7 oj.xls”

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/>
   1. Windows: <http://cran.r-project.org/bin/windows/base/>
   2. 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

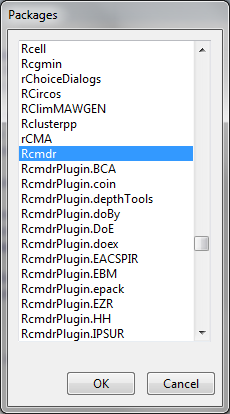
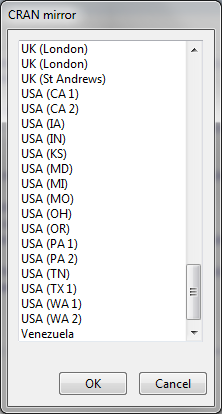


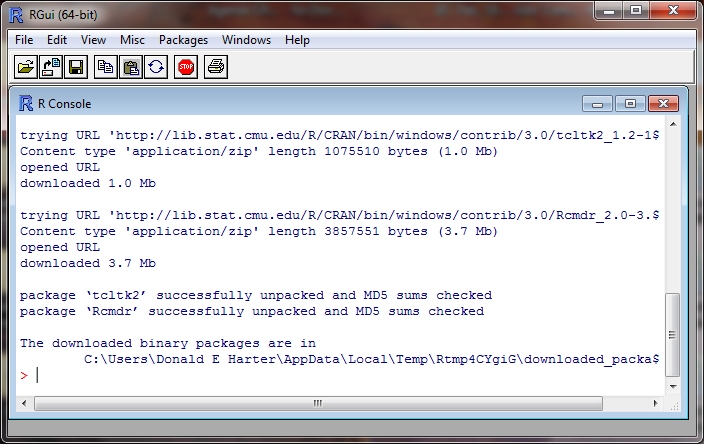
This is the command line screen. You can enter commands, but need to know the syntax. There is a simpler approach to running R, called Rcmdr (R Commander). If you are running a Whitman computer, Rcmdr is already installed. If not, you need to install it.

**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

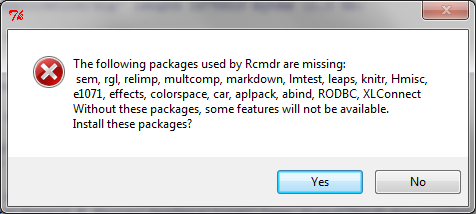


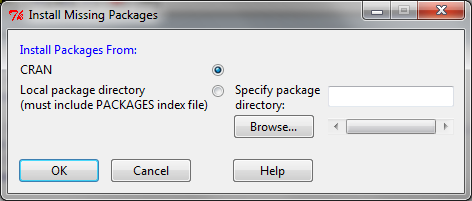


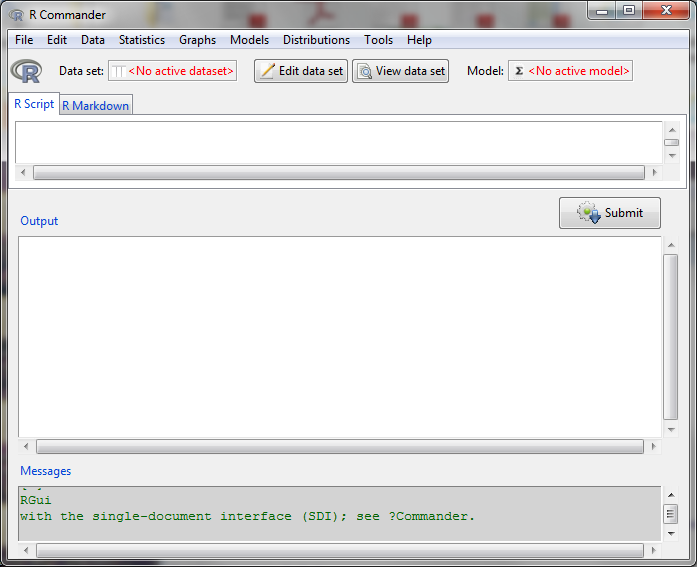
**Launch Rcmdr (R Commander)**

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

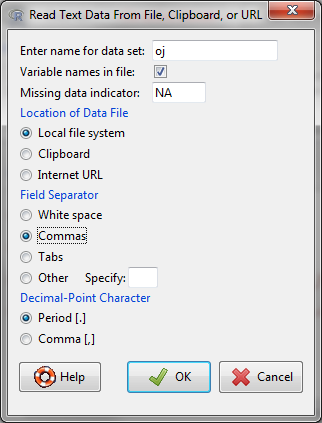
1. Click on Data Text
2. Right click on oj.csv, then save on your computer
3. Remember where you saved the file

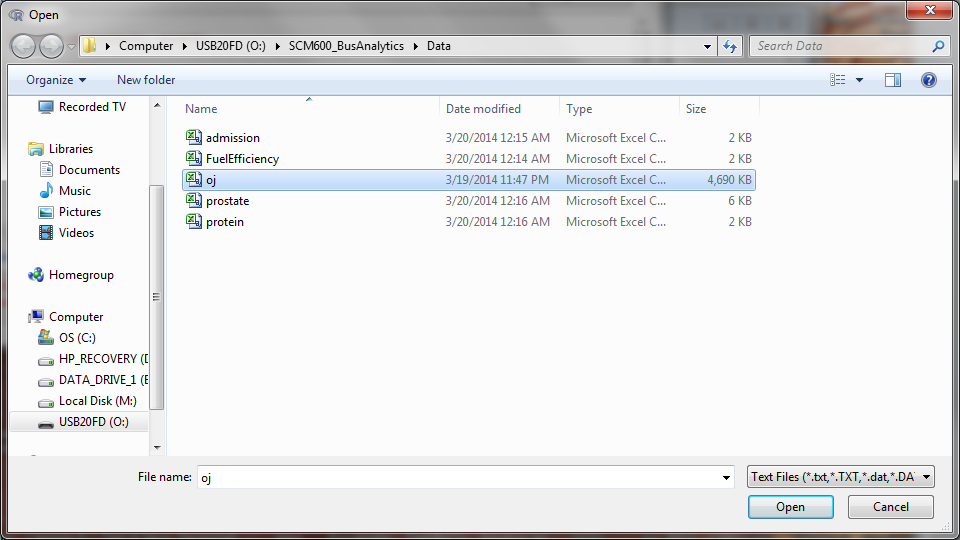
The Business Analytics - Week 7 oj.csv (orange juice) file can be downloaded from the course website.

**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

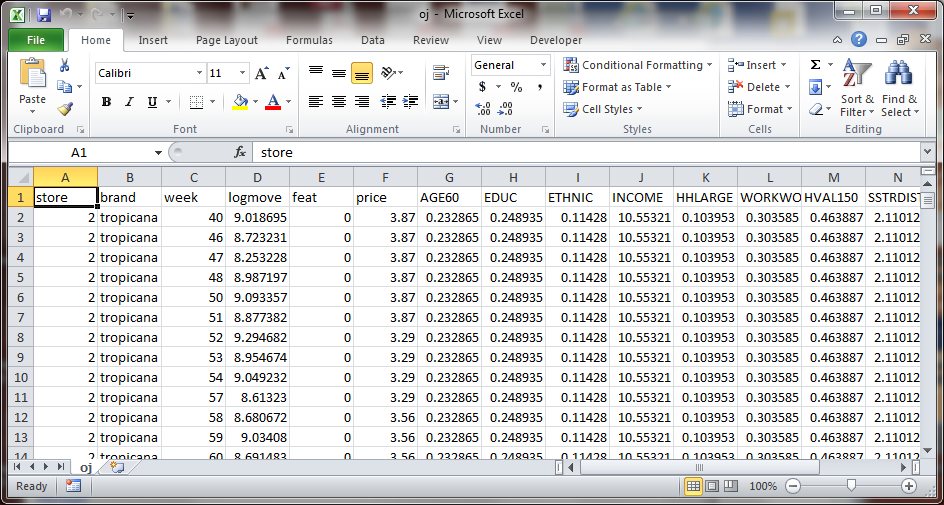




Note that the dataset oj has 28,947 rows and 17 columns.

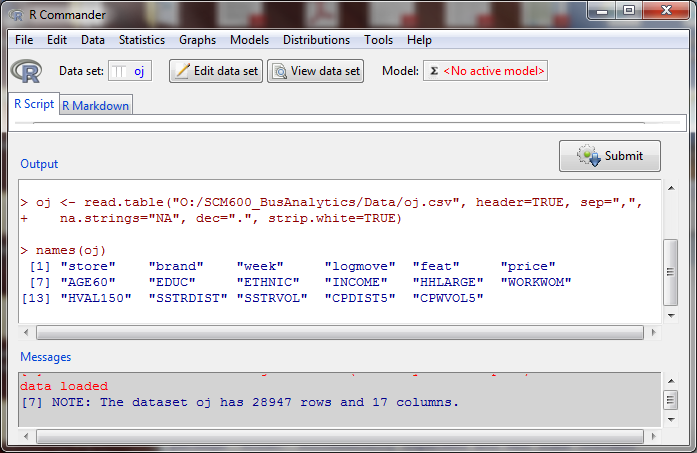
**Viewing data fields**

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).



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

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

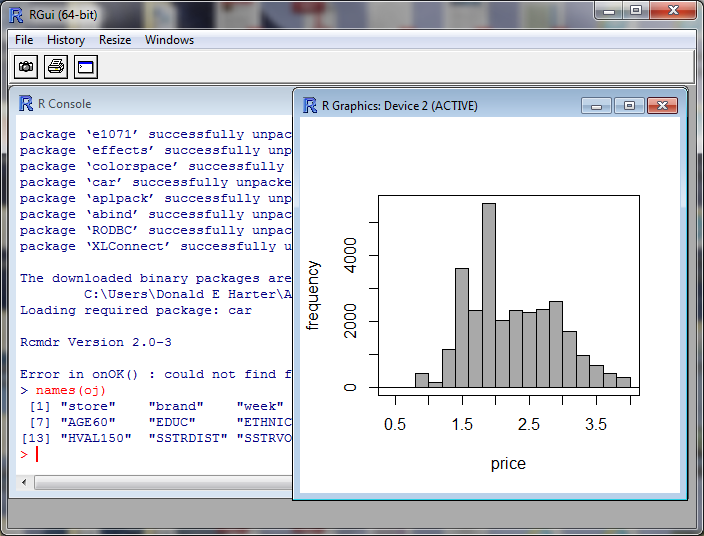


Notice that R generates the command names(oj). This is the command line version.

**Session 7.5: Histograms**

To create a histogram,

1. Click on Graphs, Histogram, price, OK

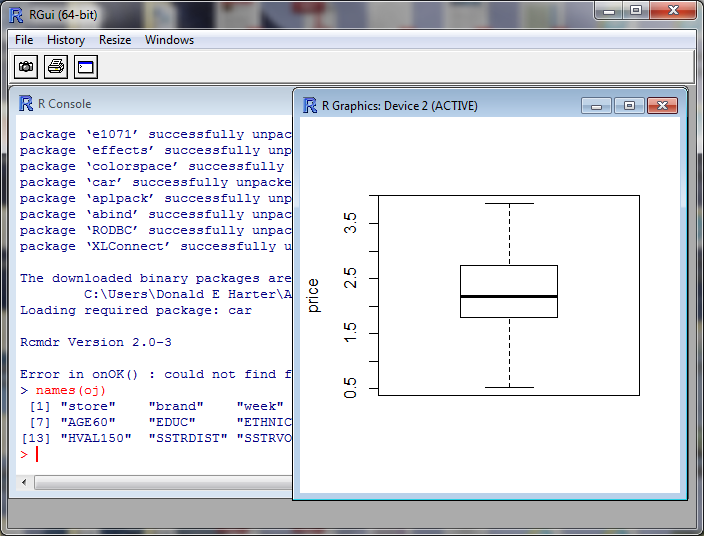


Try it again, but in the Options tab, change to Percentages (click Apply), Densities, and change number of bins to 5.

**Boxplots**

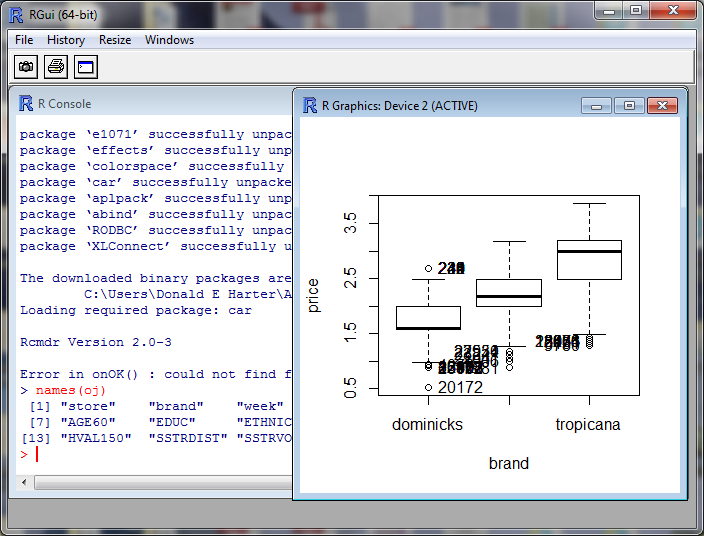
To create a boxplot,

1. Click on Graphs, Boxplot, price, OK



To create a boxplot by brand,

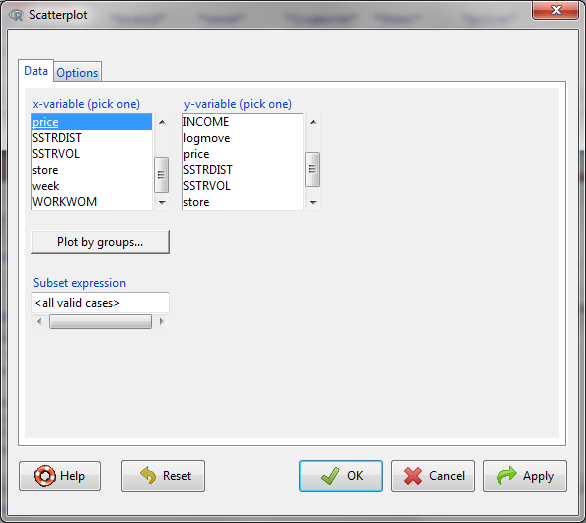
1. Click on Graphs, Boxplot, price
2. Click on Plot by groups, select brand, OK
3. Click OK



**Scatterplots**

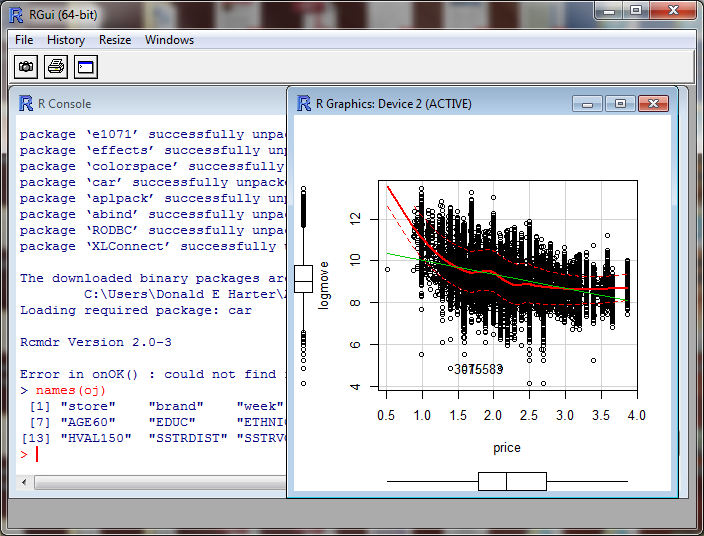
To generate a scatter plot,

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



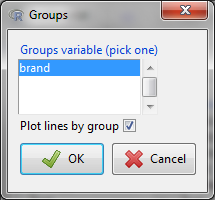
To interpret the chart:

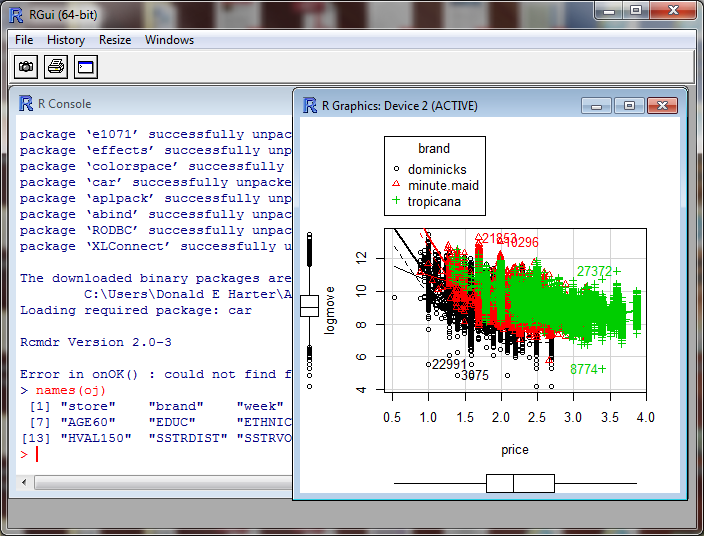
1. The black dots are the price versus log(sales) for each time period, store and brand.
2. The green line is the linear regression line through the data
3. The red lines are the averages and plus or minus one standard deviation
4. Below and to the left of the chart are box and whisker diagrams
   1.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

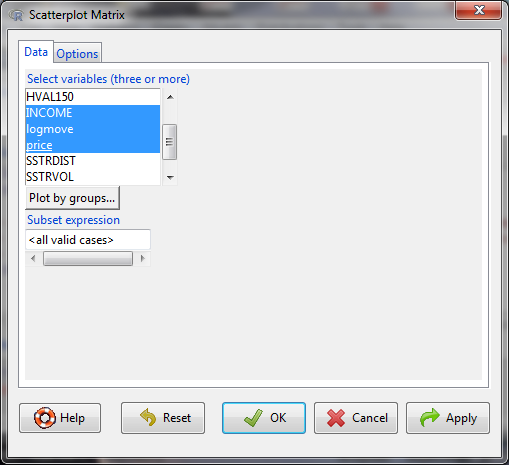
1. Brand dominicks is in black
2. Brand minute maid is in red
3. Brand Tropicana is in green

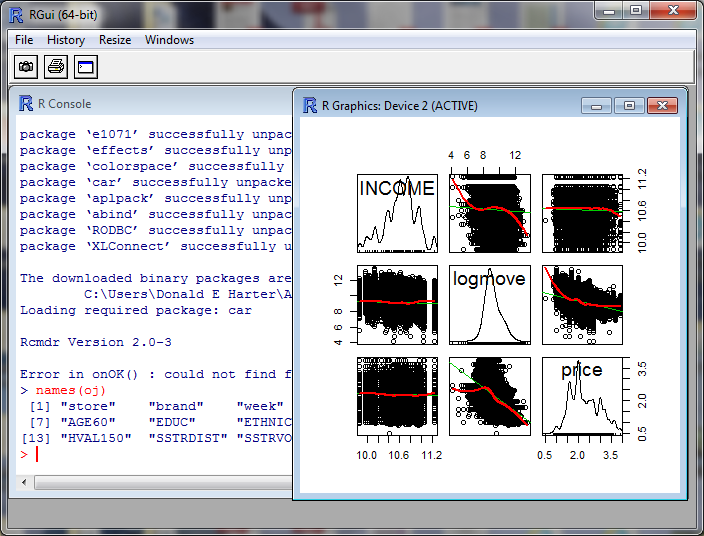
Which is the premium brand?

**Plotting pairwise scatterplots with more than two variables**

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

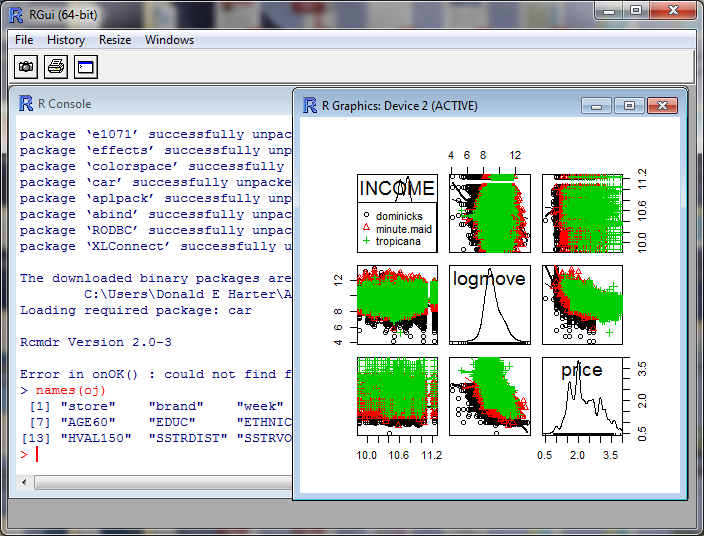
1. Click on Graphs
2. Click on Scatterplot Matrix
3. To select multiple variables, hold down the control key, then select INCOME, logmove, and price
4. Click OK





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)

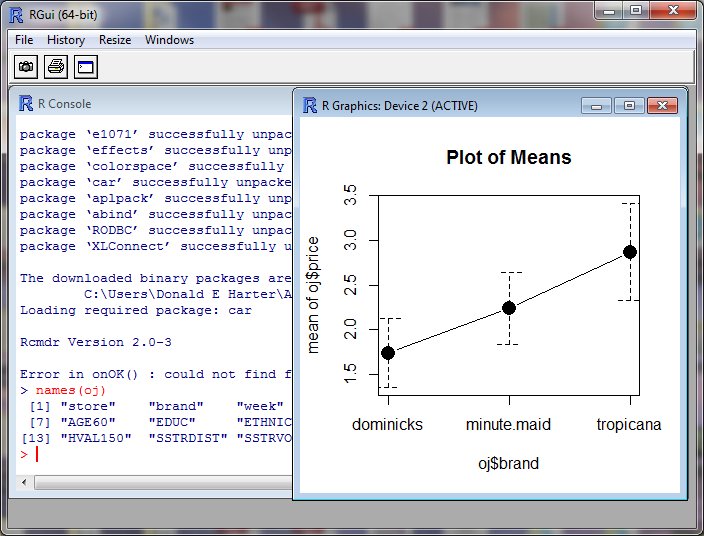


Once again, the brands are color coded.

**Plot of Means**

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

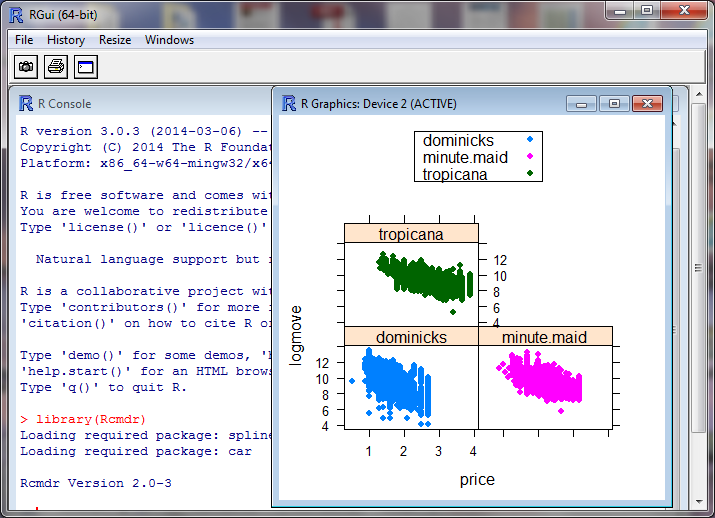
1. Click on Graphs, Plot of Means
2. Select price
3. In the Options tab, click on standard deviations, then OK



**XY Plots**

Now let’s generate XY plots by brand.

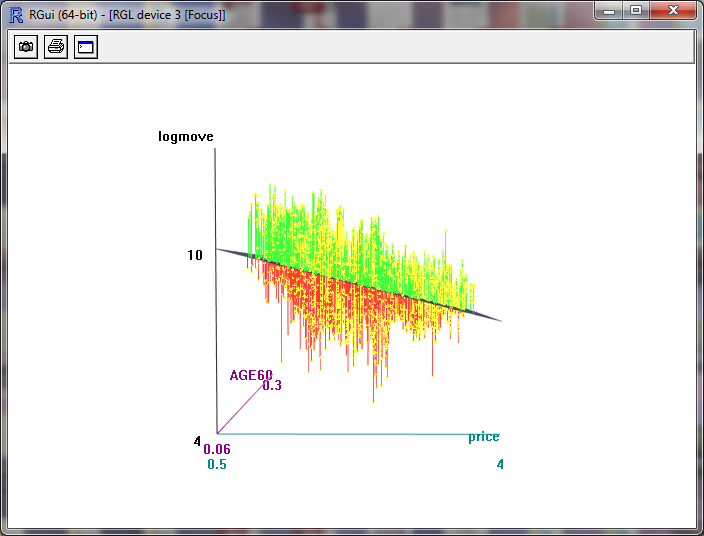
1. Click on Graphs, XY Conditioning Plot
2. Select price for the explanatory variable
3. Select logmove for the response variable
4. Click on brand for each
5. Click OK



**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 OK
5. Expand the window by clicking on the box in the upper right corner of your graph
6. Rotate the graph by clicking on the graph with your mouse, hold the mouse button down, and move

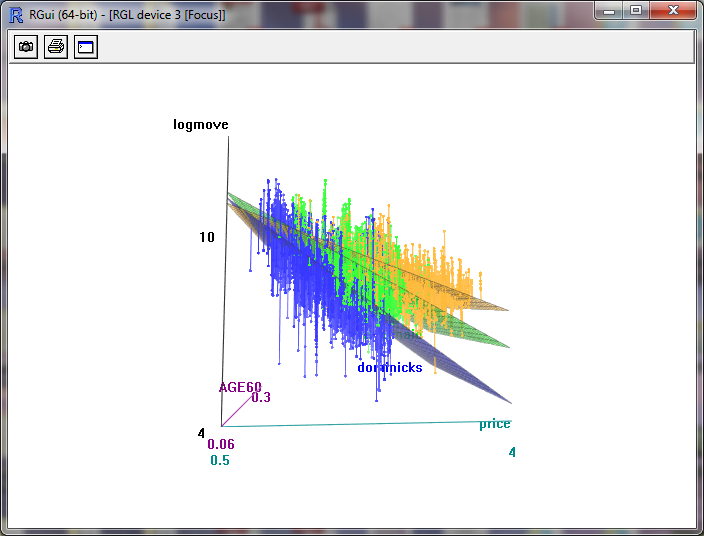


Does price affect sales?

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?

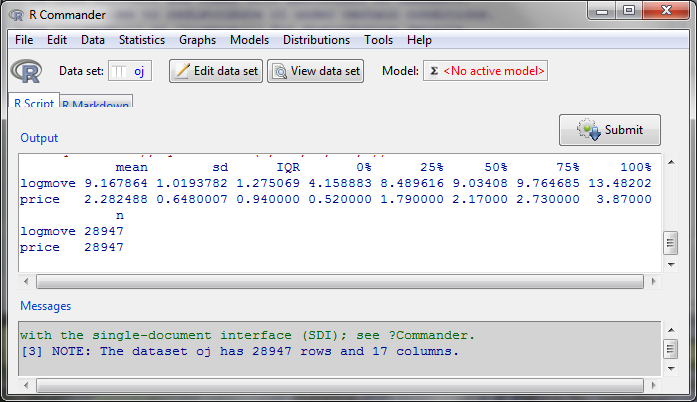
**Saving Graphs**

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

**Session 7.7: Statistical Summaries**

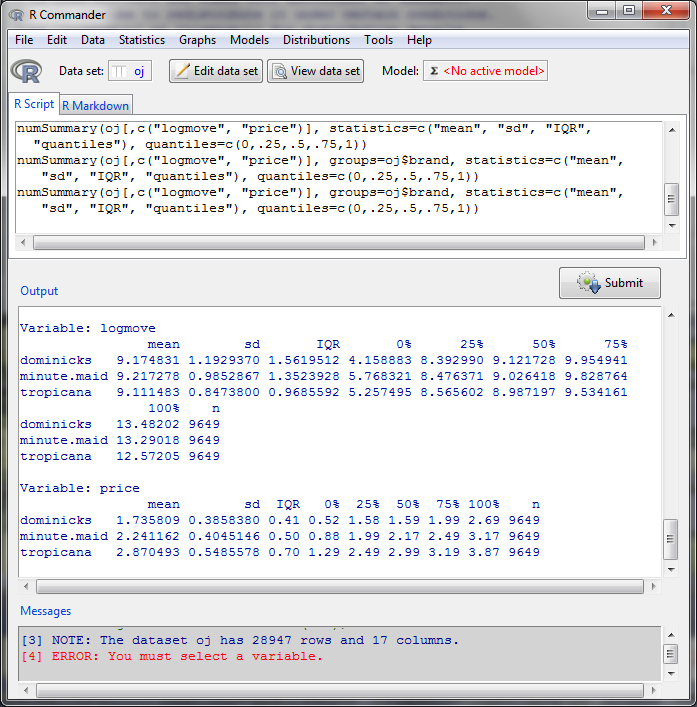
The mean, standard deviation and quartiles can be found by:

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:

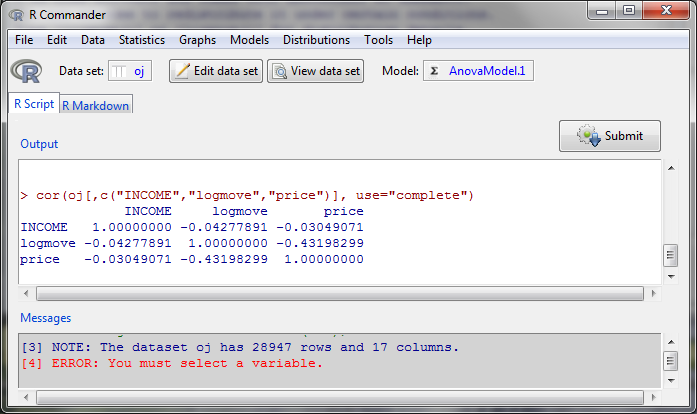
1. Click on Statistics, Summaries, Numerical Summaries
2. Select logmove, price by holding down the control key
3. Click Summarize by Groups, then OK
4. Click OK



**Session 7.8: Correlation**

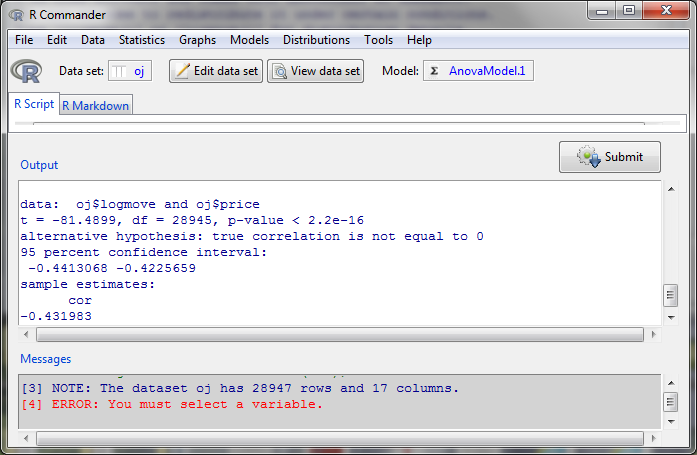
To generate a correlation matrix,

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



The 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

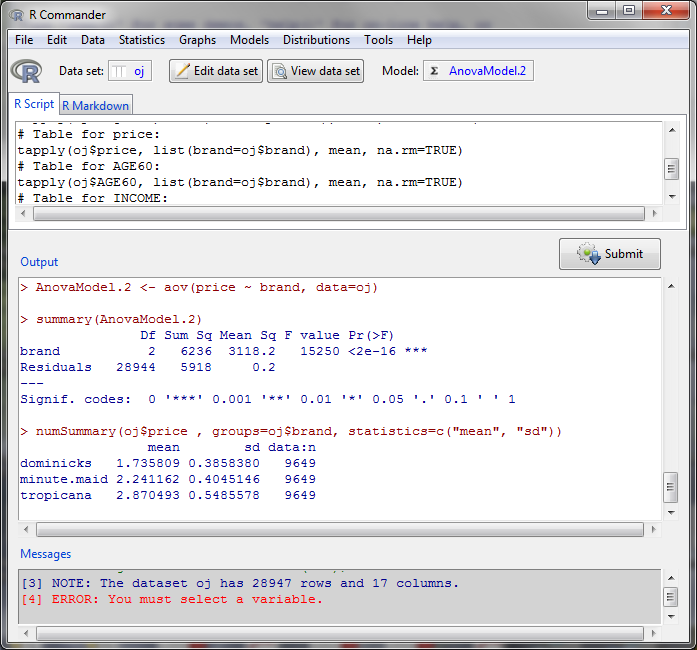


The p-value is less than 0.05, so the correlation is statistically significant.

**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

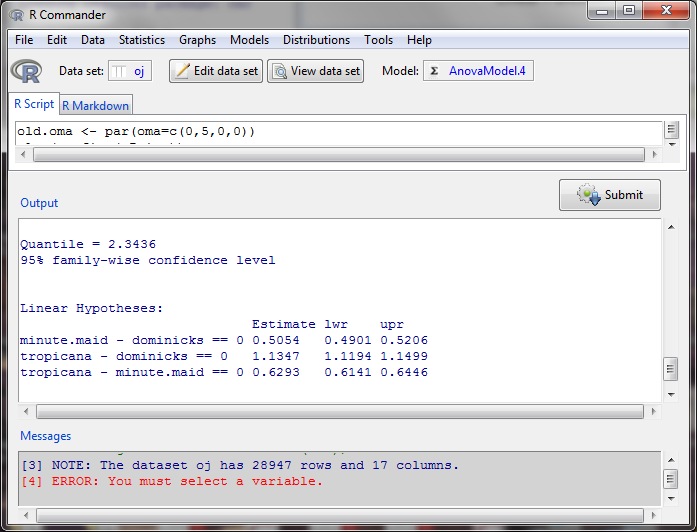


The F-statistic p-value [Pr(>F)] is less than 0.05. That means that one of the brands has a price that is statistically different from the others.

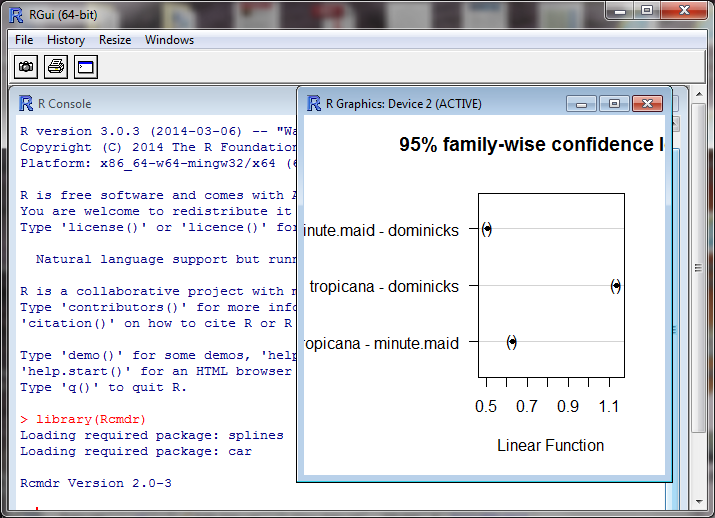
Let’s now perform a pairwise comparison.

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?



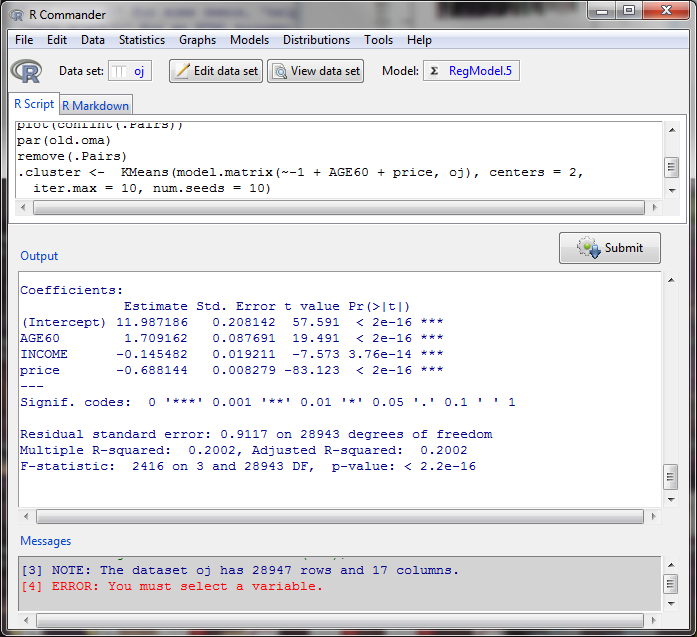
The graph portrays the estimate of price for each brand and the confidence interval.



**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



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.

Tropicana

Minute Maid

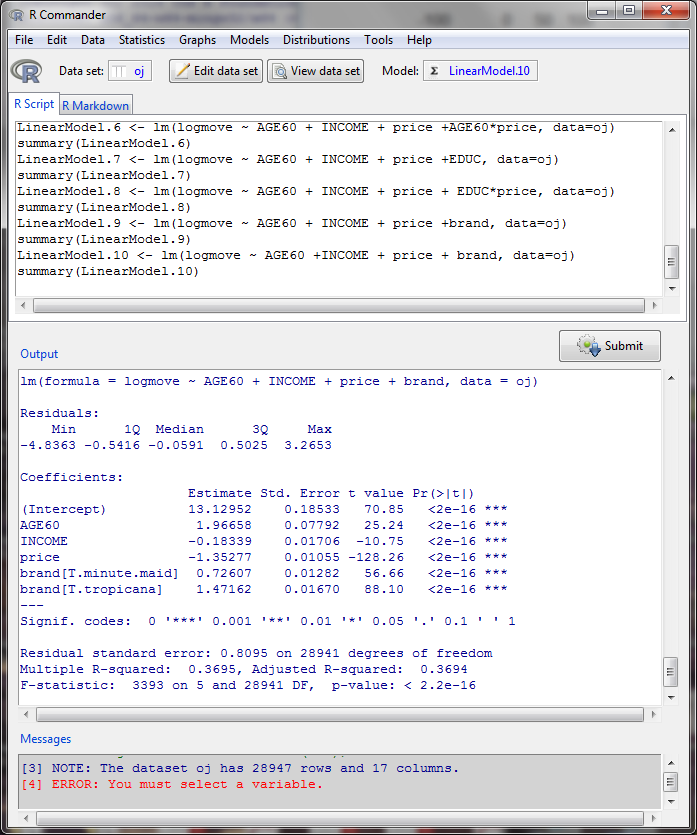
Sales

Dominicks

Price

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

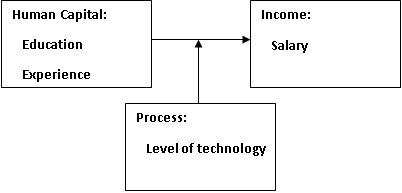


In the Output section, the intercept is 13.12952. This is the intercept for dominicks. The intercept for minute maid is 13.12952 + 0.72607. The intercept for Tropicana is 13.12952 + 1.47162.

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

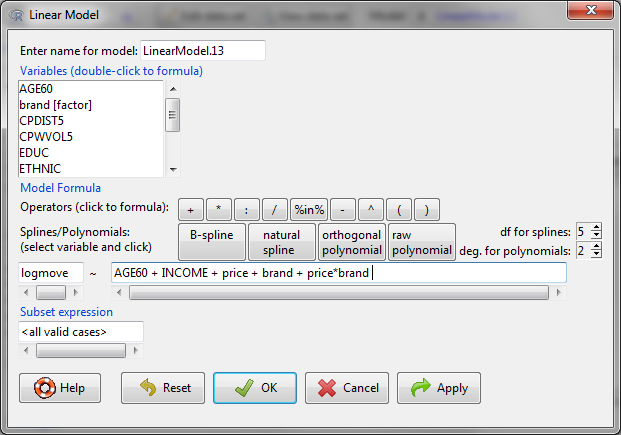
**Session 7.12: Moderating effects (interactions of price and brand)**

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?

