

R: Basics

Overview

The following instructions for installing R are for your own computer. You can use rds.syr.edu to access the server version of R.

Installation of R

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/>

The Comprehensive R Archive Network

Download and Install R

Precompiled binary distributions of the base system and contributed packages, **Windows and Mac** users most likely want one of these versions of R:

- [Download R for Linux](#) (Debian, Fedora/Redhat, Ubuntu)
- [Download R for macOS](#)
- [Download R for Windows](#)

R is part of many Linux distributions, you should check with your Linux package management system in addition to the link above.

Source Code for all Platforms

Windows and Mac users most likely want to download the precompiled binaries listed in the upper box, not the source code. The sources have to be compiled before you can use them. If you do not know what this means, you probably do not want to do it!

- The latest release (2022-03-10, One Push-Up) [R-4.1.3.tar.gz](#), read [what's new](#) in the latest version.
- Sources of [R alpha and beta releases](#) (daily snapshots, created only in time periods before a planned release).
- Daily snapshots of current patched and development versions are [available here](#). Please read about [new features and bug fixes](#) before filing corresponding feature requests or bug reports.
- Source code of older versions of R is [available here](#).
- Contributed extension [packages](#)

2. Click on the Download link for your operating system
 - a. For Mac Users, click on R for (Mac) OS X. Note that if you have X11, you must install XQuartz, since it is no longer part of OS X. See the directions for R 4.1.3 on the page displayed after clicking on (Mac) OS X. Click on R-4.1.3.pkg
 - b. For Windows users, click on R for Windows, then click on "install R for the first time"
3. Click on Run, and follow the install instructions

Remote Access

To run R and RStudio on the server:

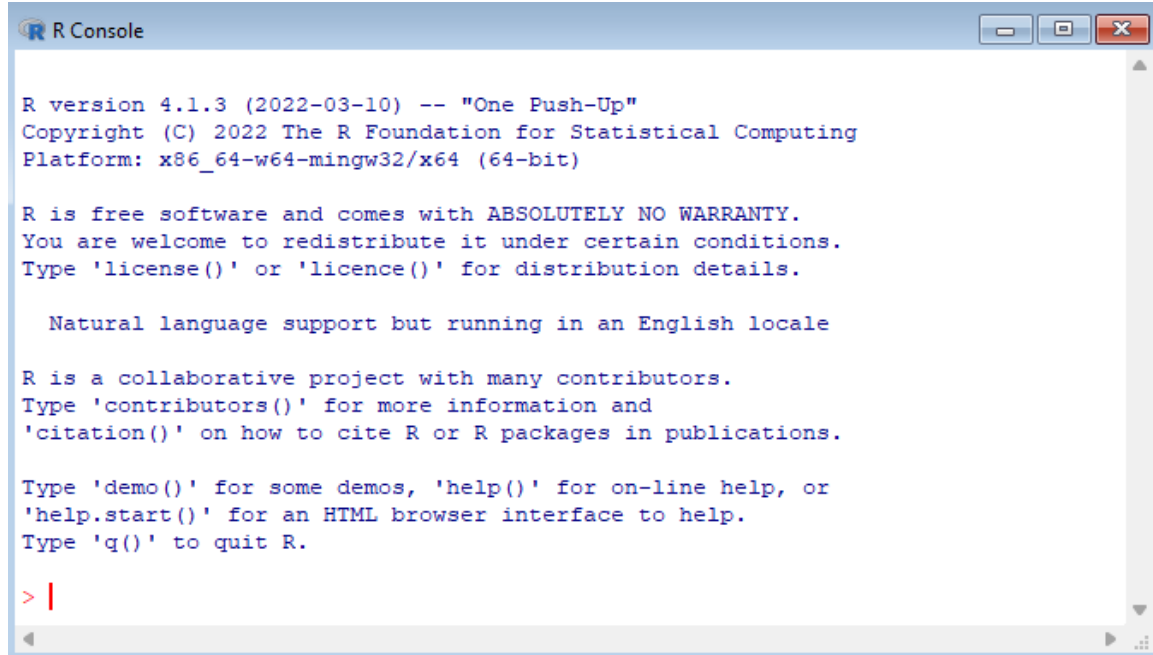
1. Go to rds.syr.edu
2. In the upper right corner click on Settings (picture of gear), then Download the rdp file.
3. Click on Whitman Remote Desktop in the upper left corner; this will download the Whitman Remote Desktop.rdp file
4. Click on the Whitman Remote Desktop.rdp file in the lower left corner
5. When the Whitman Remote Desktop Connection window opens, click Connect
6. The login screen will appear; enter your Syracuse University email address and Syracuse University password
7. After several seconds to a minute, your rds desktop will appear

R versus RStudio

R is a command line system where you can enter R commands. RStudio is a programmer development environment for the R language. RStudio has some bugs in it which will cause it to crash in this course. Please use R and RCommander for this course.

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 x64 or R x386
3. This is the command line screen.

A screenshot of the R Console window. The window has a title bar that says "R Console" and standard Windows window controls (minimize, maximize, close). The main area is a white text box with blue text. It displays the R version (4.1.3), copyright information (© 2022 The R Foundation for Statistical Computing), and platform (x86_64-w64-mingw32/x64 (64-bit)). It also includes a disclaimer about warranty and a list of useful commands like 'license()', 'contributors()', 'citation()', 'demo()', 'help()', 'help.start()', and 'q()'. The prompt ">|" is visible at the bottom left of the text area.

```
R Console

R version 4.1.3 (2022-03-10) -- "One Push-Up"
Copyright (C) 2022 The R Foundation for Statistical Computing
Platform: x86_64-w64-mingw32/x64 (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

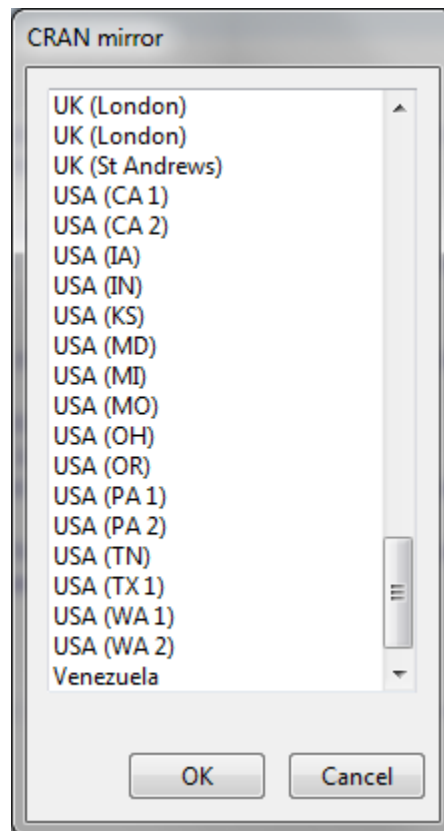
> |
```

Installing R Commander

Follow these steps only if you do not already have Rcmdr installed.

1. In R, type the command:

```
install.packages("Rcmdr", dependencies = TRUE)
```
2. In the CRAN mirror, select the location closest to you; use a USA location near you, then click OK
3. If prompted to create a personal library, click Yes
4. If prompted to add missing packages, click Yes

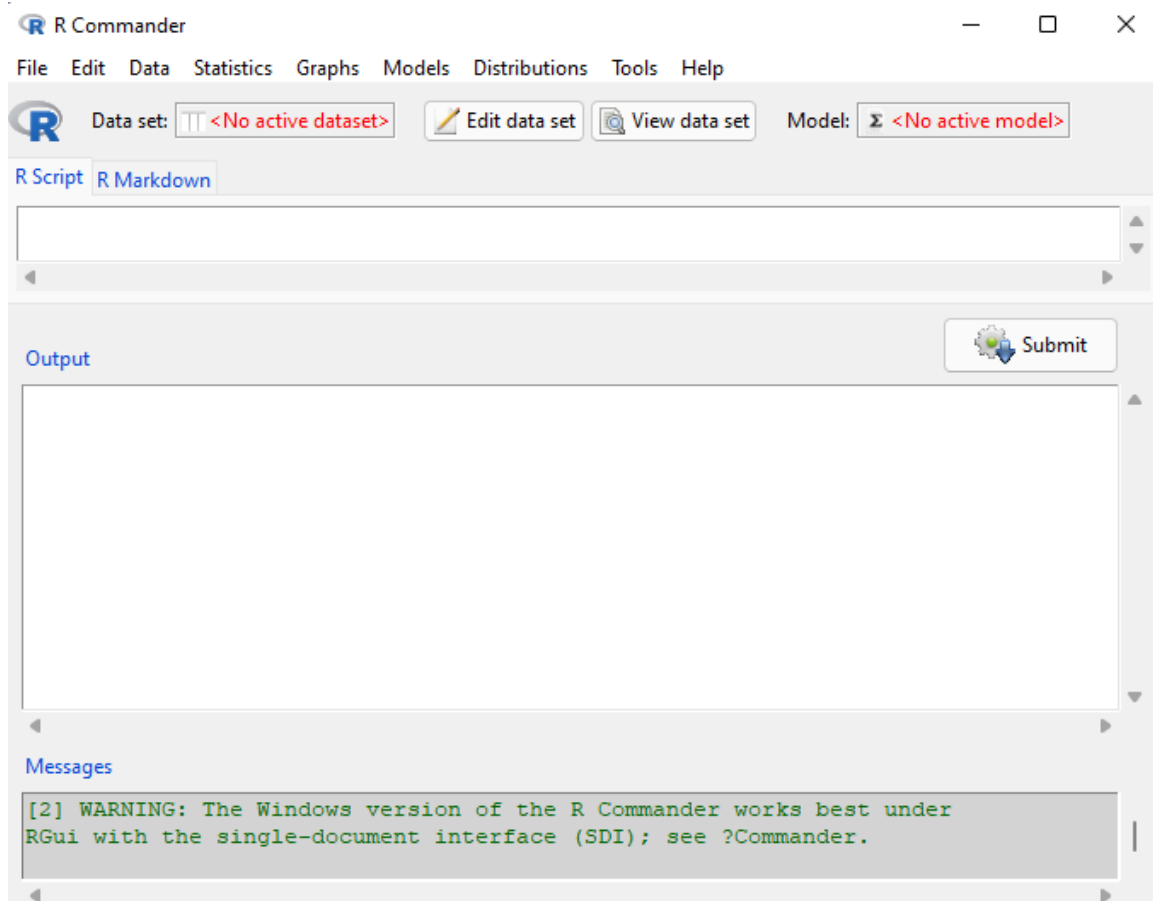


Launch Rcmdr (R Commander)

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

1. Type in the command:

```
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



Download Datasets

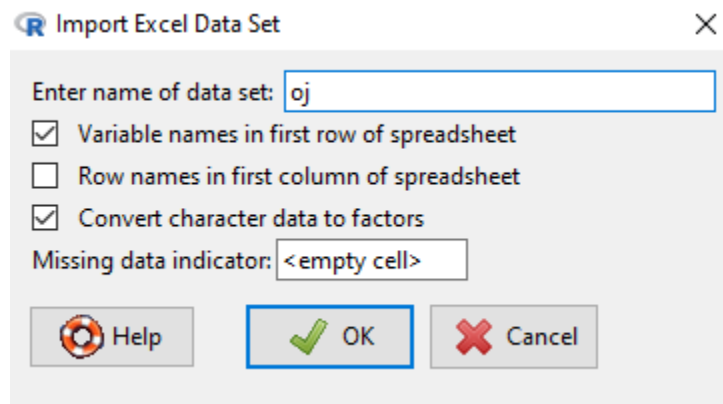
To access this and other excellent data sets used in the book “Data Mining and Business Analytics with R,” by Johannes Ledolter:

1. go to BlackBoard and right click on the file oj.xls and save the file to your desktop. This file is also on the G: drive.

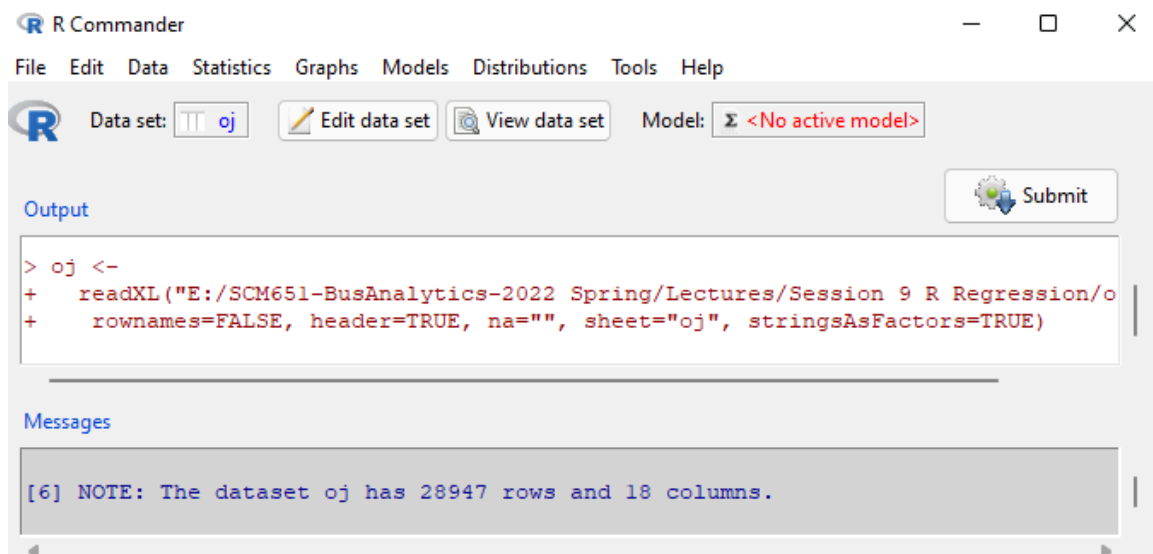
Loading Data

To load text data or .csv files (comma separated values files) into R:

1. Click on Data at the top of the Rcmdr screen
2. Click on Import Data > From Excel file ...
3. Enter the name that you would like to use for this data set; type in oj (stands for orange juice), then click OK



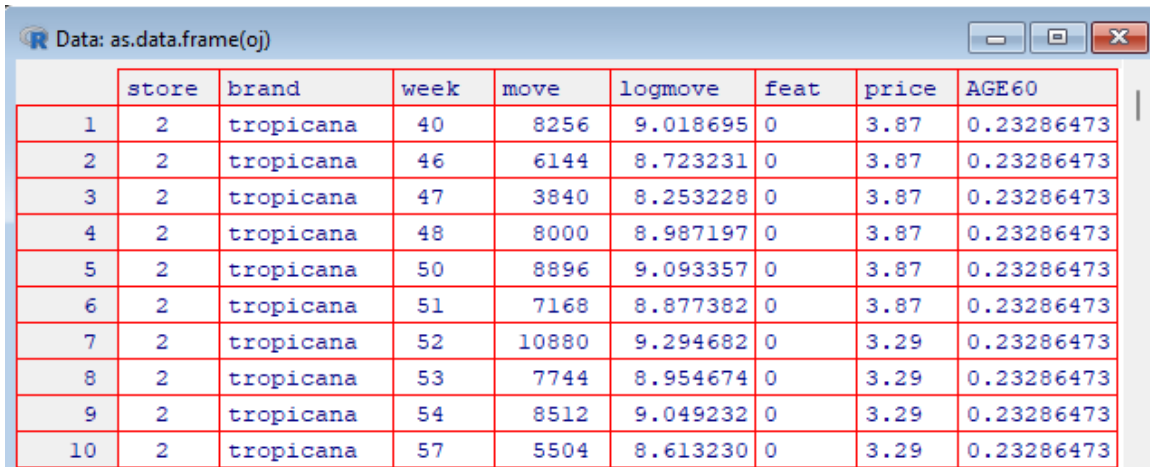
4. Click on the oj file, then Open
5. Note that the dataset oj has 28,947 rows and 18 columns.



Viewing data fields

This data set lists weekly sales over 83 stores for three brands of products.

1. Let's view the data. The variable logmove is the logarithm of sales (how much product moved in a week). Click on the button View data set in Rcmdr.



The screenshot shows the Rcmdr interface with the title bar 'Data: as.data.frame(oj)'. The window displays a table with 9 columns: an index column (1-10), 'store', 'brand', 'week', 'move', 'logmove', 'feat', 'price', and 'AGE60'. All data is for the 'tropicana' brand. The 'move' values range from 5504 to 10880, and 'logmove' values range from 8.613230 to 9.018695. The 'price' is 3.87 for weeks 40-51 and 3.29 for weeks 52-57. The 'AGE60' column is constant at 0.23286473.

	store	brand	week	move	logmove	feat	price	AGE60
1	2	tropicana	40	8256	9.018695	0	3.87	0.23286473
2	2	tropicana	46	6144	8.723231	0	3.87	0.23286473
3	2	tropicana	47	3840	8.253228	0	3.87	0.23286473
4	2	tropicana	48	8000	8.987197	0	3.87	0.23286473
5	2	tropicana	50	8896	9.093357	0	3.87	0.23286473
6	2	tropicana	51	7168	8.877382	0	3.87	0.23286473
7	2	tropicana	52	10880	9.294682	0	3.29	0.23286473
8	2	tropicana	53	7744	8.954674	0	3.29	0.23286473
9	2	tropicana	54	8512	9.049232	0	3.29	0.23286473
10	2	tropicana	57	5504	8.613230	0	3.29	0.23286473

2. To view a list of the variables in R, click on Data, Active Data Set, Variables in Active Data Set

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: <No active model>.
- Script Editor:** Contains the command `names(oj)`.
- Output Console:** Displays the results of the commands:

```
+ readXL("E:/SCM651-BusAnalytics-2022 Spring/Lectures/Session 9 R Regression/o
+ rownames=FALSE, header=TRUE, na="", sheet="oj", stringsAsFactors=TRUE)

> oj <-
+ readXL("E:/SCM651-BusAnalytics-2022 Spring/Lectures/Session 9 R Regression/o
+ rownames=FALSE, header=TRUE, na="", sheet="oj", stringsAsFactors=TRUE)

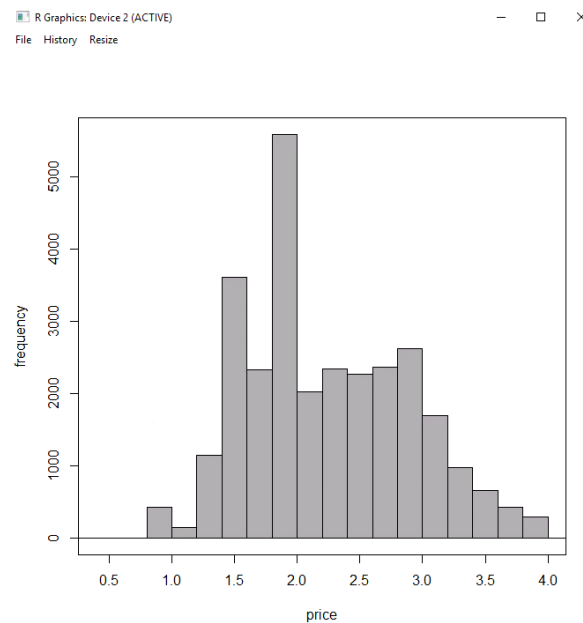
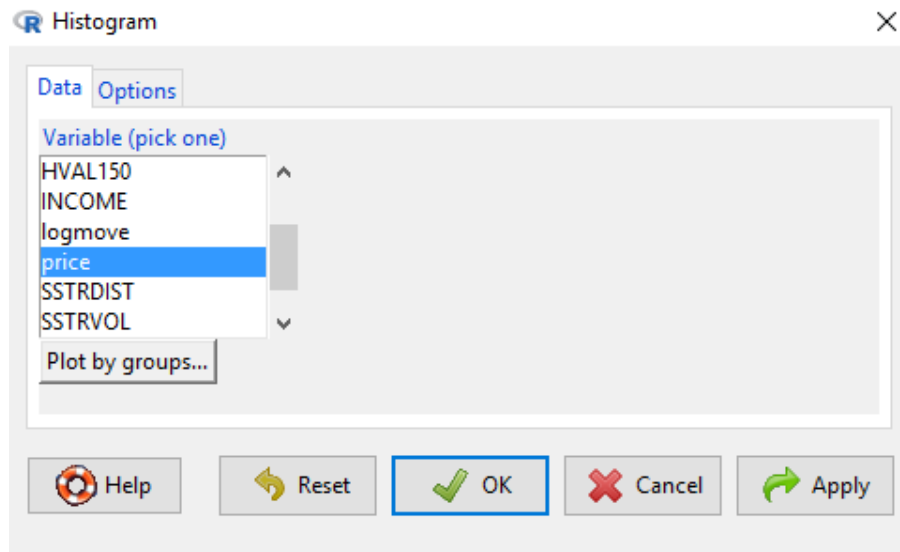
> names(oj)
[1] "store"    "brand"    "week"     "move"     "logmove"  "feat"
[7] "price"    "AGE60"    "EDUC"     "ETHNIC"   "INCOME"   "HHLARGE"
[13] "WORKWOM"  "HVAL150"  "SSTRDIST" "SSTRVOL"  "CPDIST5"  "CPWVOL5"
```
- Messages Console:** Displays the message: `[6] NOTE: The dataset oj has 28947 rows and 18 columns.`

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

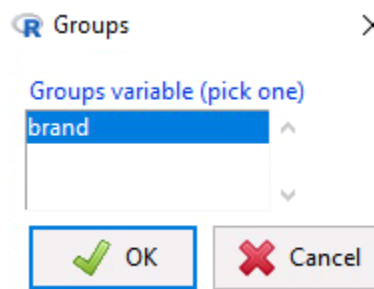
Histograms

To create a histogram,

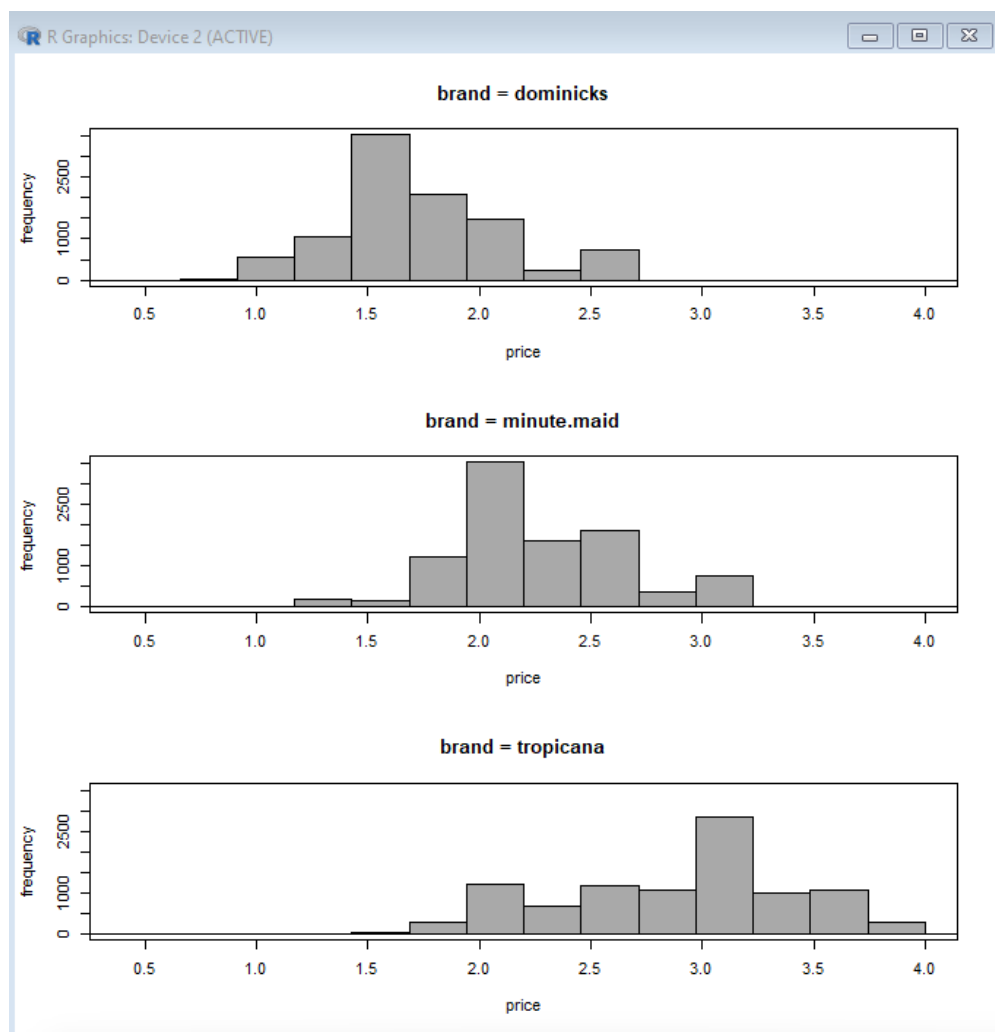
1. Click on Graphs, Histogram
2. Click on the variable price, OK



- Next, plot by groups. Click on Graphs, Histogram, Plot by groups



- Click on brand, then OK
- Click on OK to generate the histograms

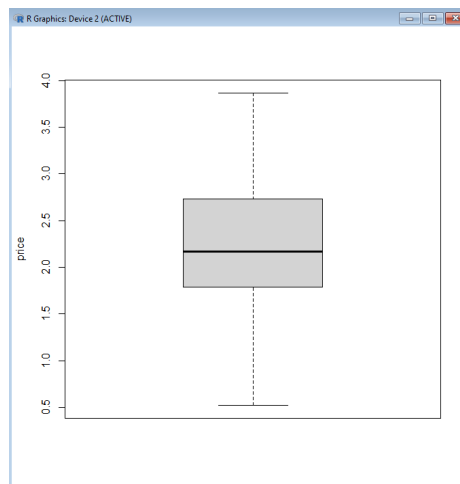


- Which brand is the premium brand?

Boxplots

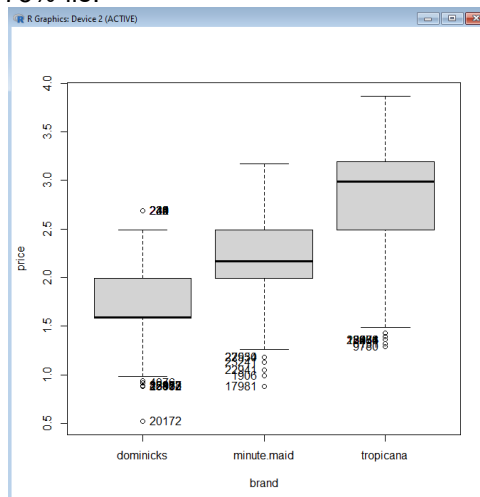
To create a boxplot,

1. Click on Graphs, Boxplot, price, OK
 - a. The upper line is the maximum, up to 1.5 times the interquartile range (box size: 75%-ile minus 25%-ile)
 - b. The lower line is the minimum, up to 1.5 times the interquartile range (box size: 75%-ile minus 25%-ile)
 - c. The middle line is the median (50%-ile)
 - d. The top of the box is the 75%-ile
 - e. The bottom of the box is the 25%-ile
 - f. The inter-quartile distance is the distance between the 25%-ile and 75%-ile



To create a boxplot by brand,

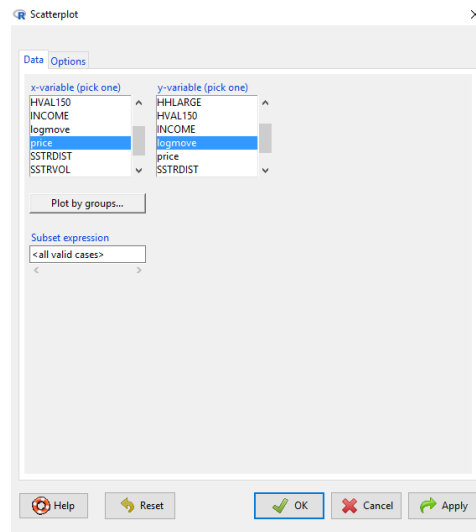
1. Click on Graphs, Boxplot, price
2. Click on Plot by groups, select brand, OK, then click OK again
3. Data points beyond the whiskers are outliers
4. Outliers for boxplots are points that are more than 1.5 times the inter-quartile distance from the 25%-ile or 75%-ile.



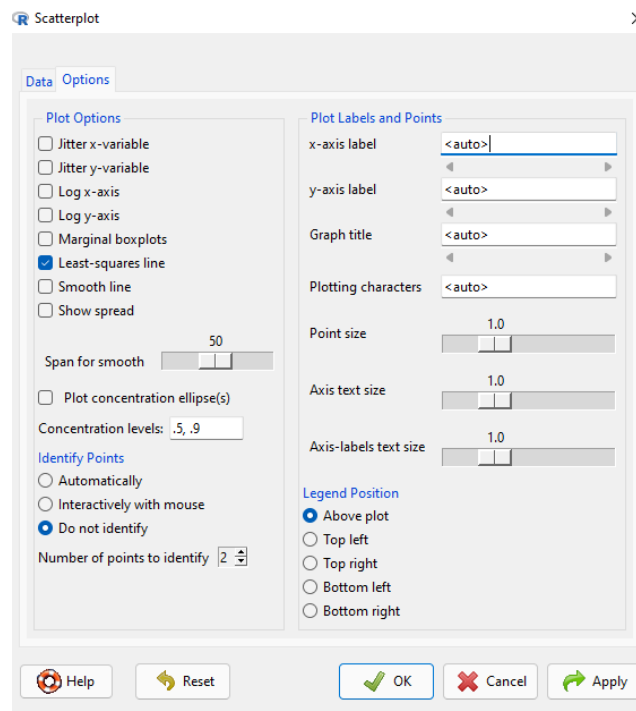
Scatterplots

To generate a scatter plot,

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

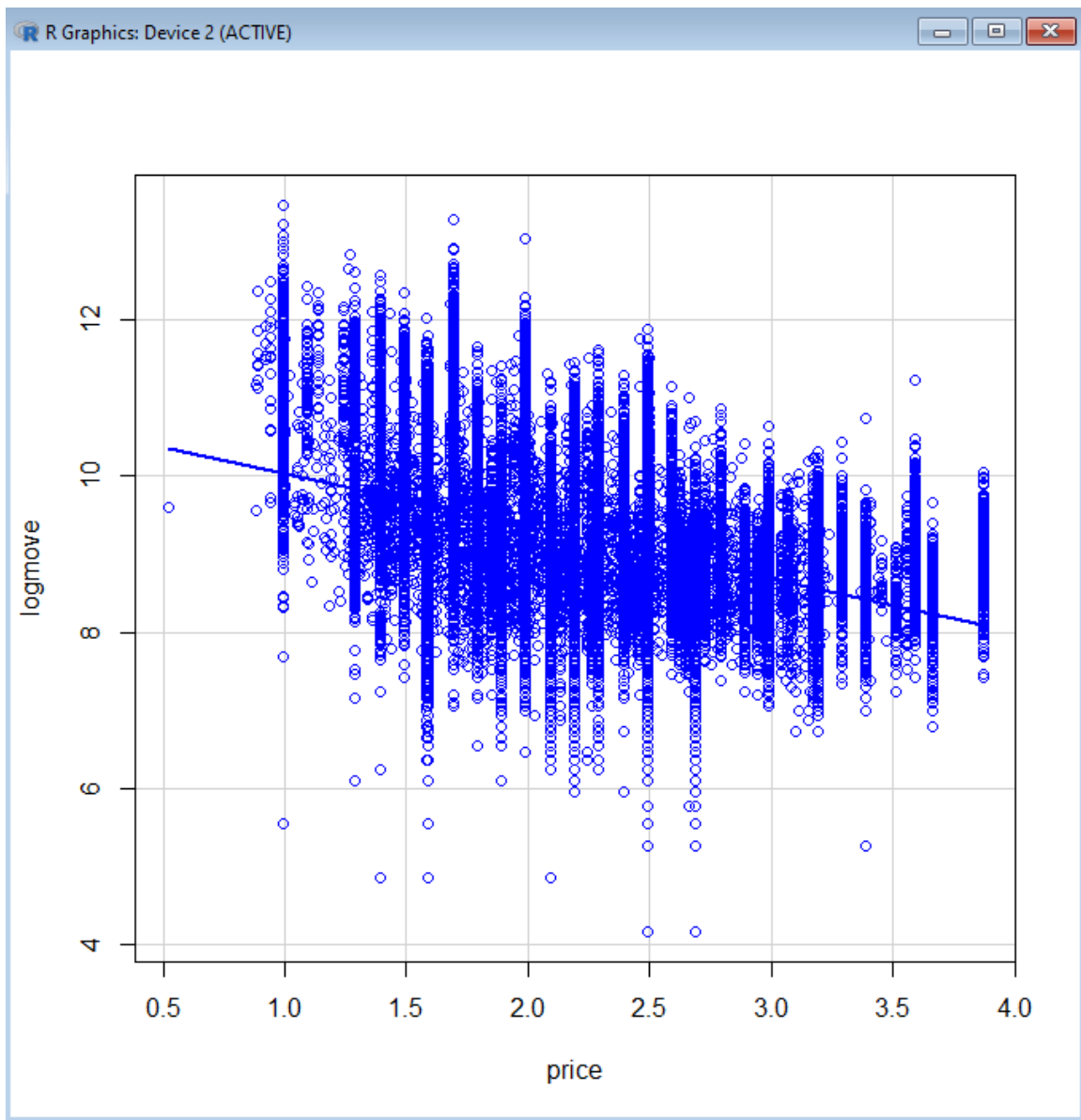


4. Click on the Options tab, select Least-squares line
5. Click on OK



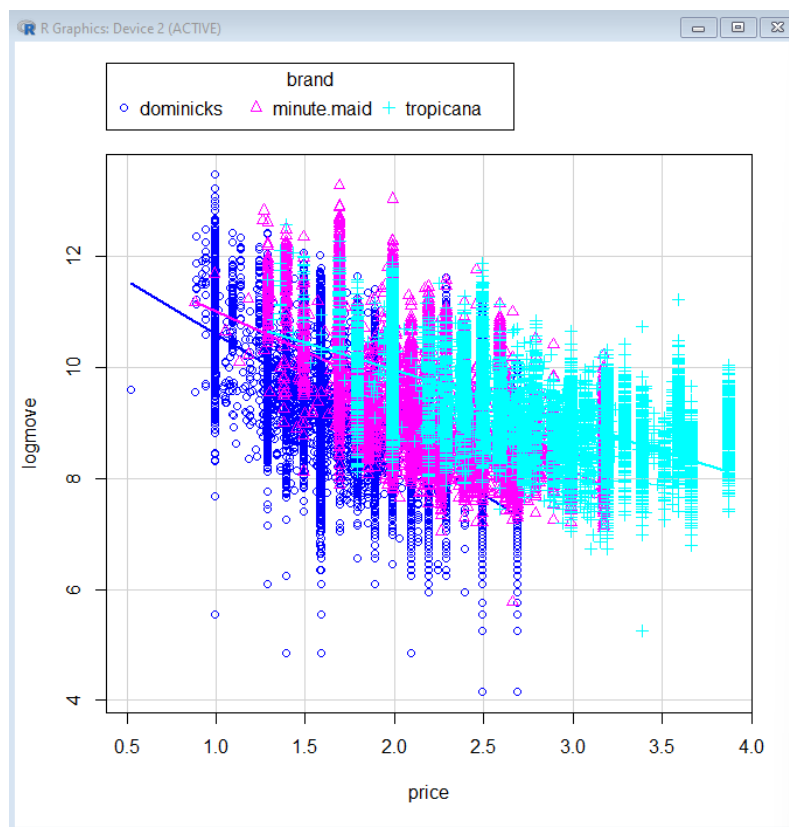
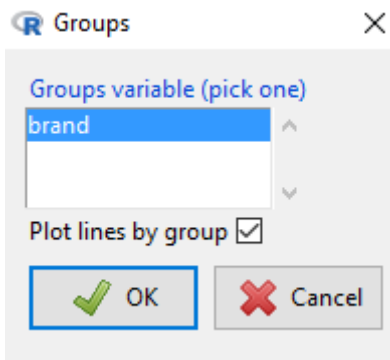
To interpret the chart:

6. The blue dots are the price versus $\log(\text{sales})$ for each time period, store and brand.
7. The blue solid line is the linear regression line through the data



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

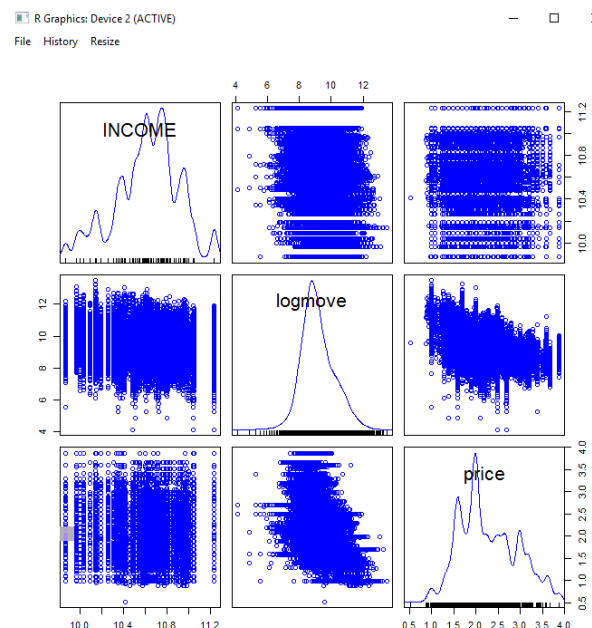
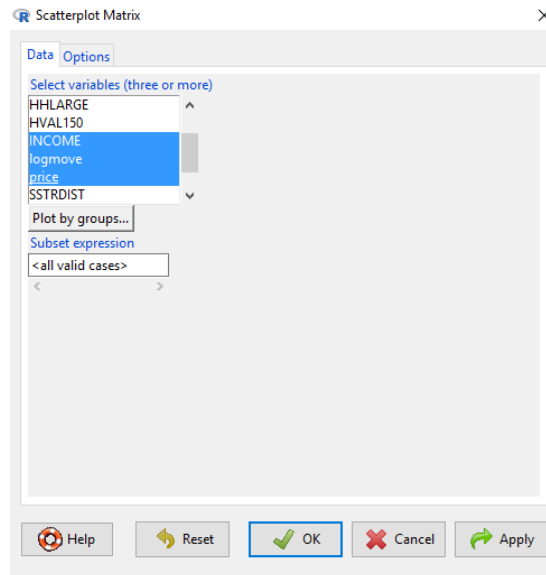


6. Interpretation
 - a. Brand Dominicks is in dark blue
 - b. Brand Minute Maid is in purple/pink
 - c. Brand Tropicana is in light blue
7. Which is the premium brand? Why?

Plotting pairwise scatterplots with more than two variables

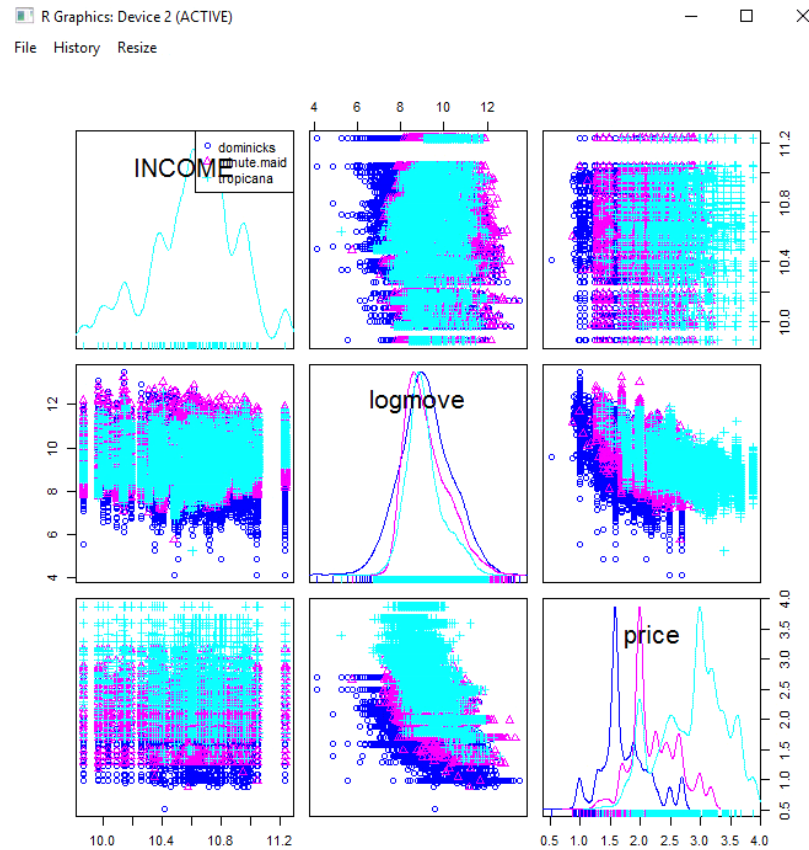
To create a matrix of scatterplots with more than two variables,

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.

5. Now perform a Scatterplot Matrix by Groups (brand)
6. Click on Graphs, Scatterplot Matrix
7. Check that INCOME, logmove and price are still highlighted
8. Click Plot by Groups
9. Click on brand, then OK
10. Click OK again

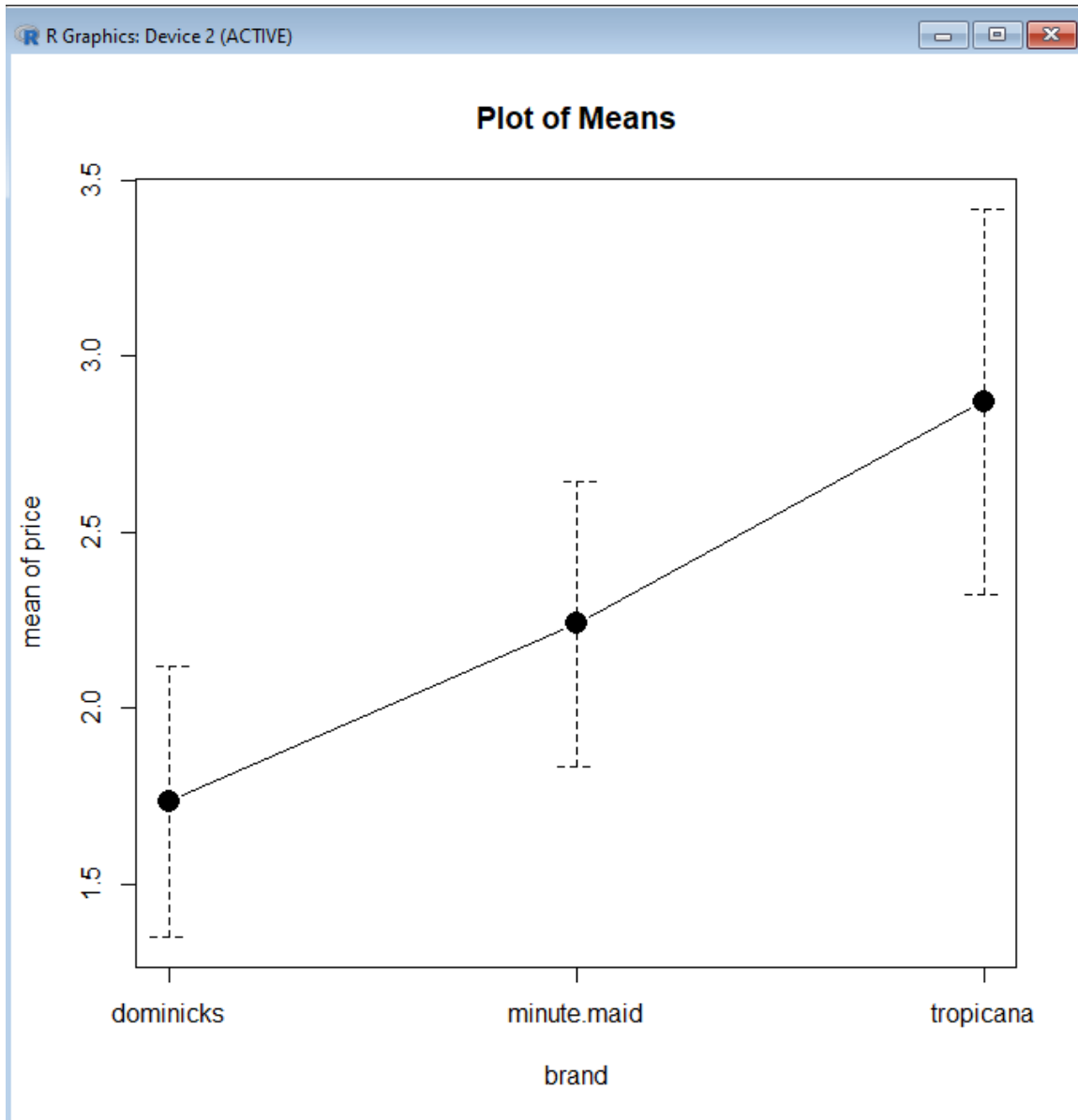


11. Once again, the brands are color coded.
 - a. Brand Dominicks is in dark blue
 - b. Brand Minute Maid is in purple/pink
 - c. Brand Tropicana is in light blue

Plot of Means

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

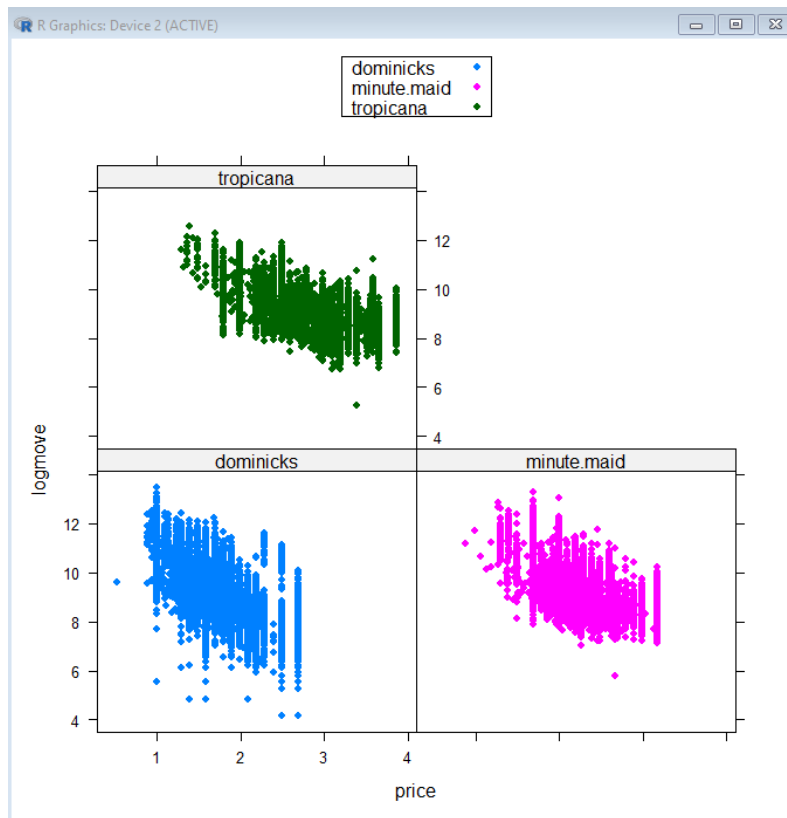
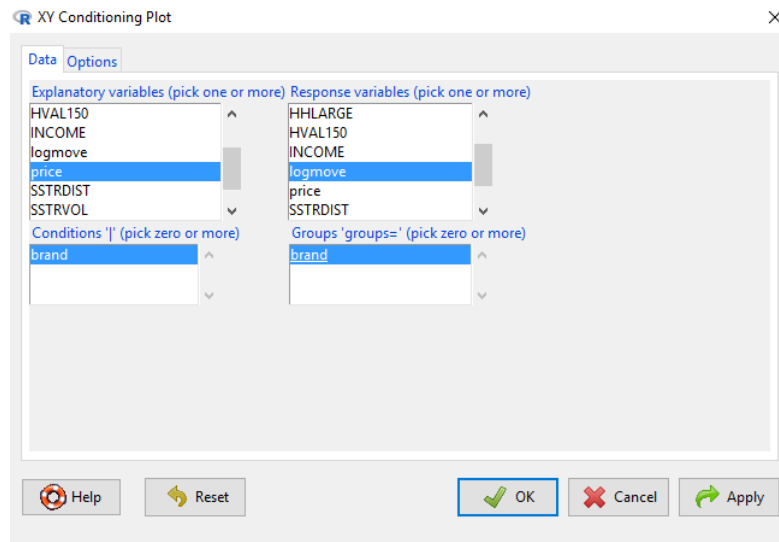
1. Click on Graphs, Plot of Means
2. Select price
3. In the Options tab, click on standard deviations, then OK



XY Plots

Now let's generate XY plots by brand.

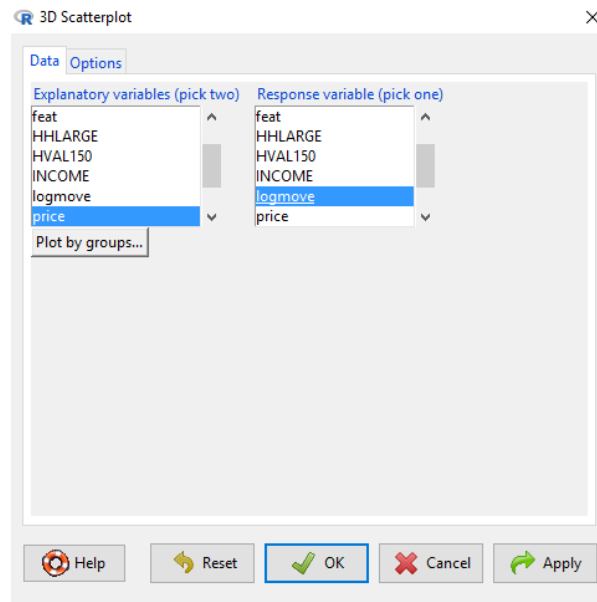
1. Click on Graphs, XY Conditioning Plot
2. Select price for the explanatory variable
3. Select logmove for the response variable
4. Click on brand for each
5. Click OK



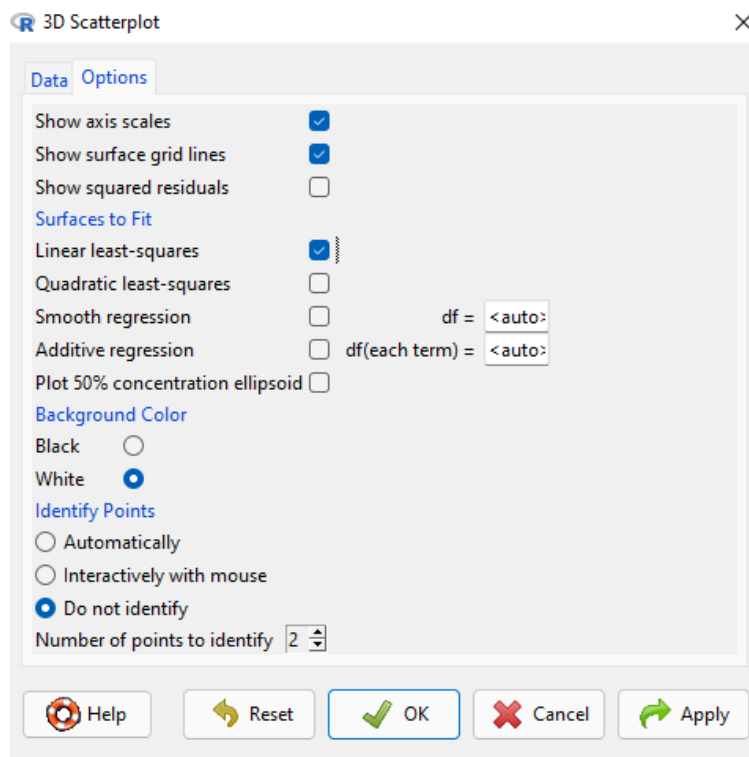
3D Graphs

To generate 3D graphs,

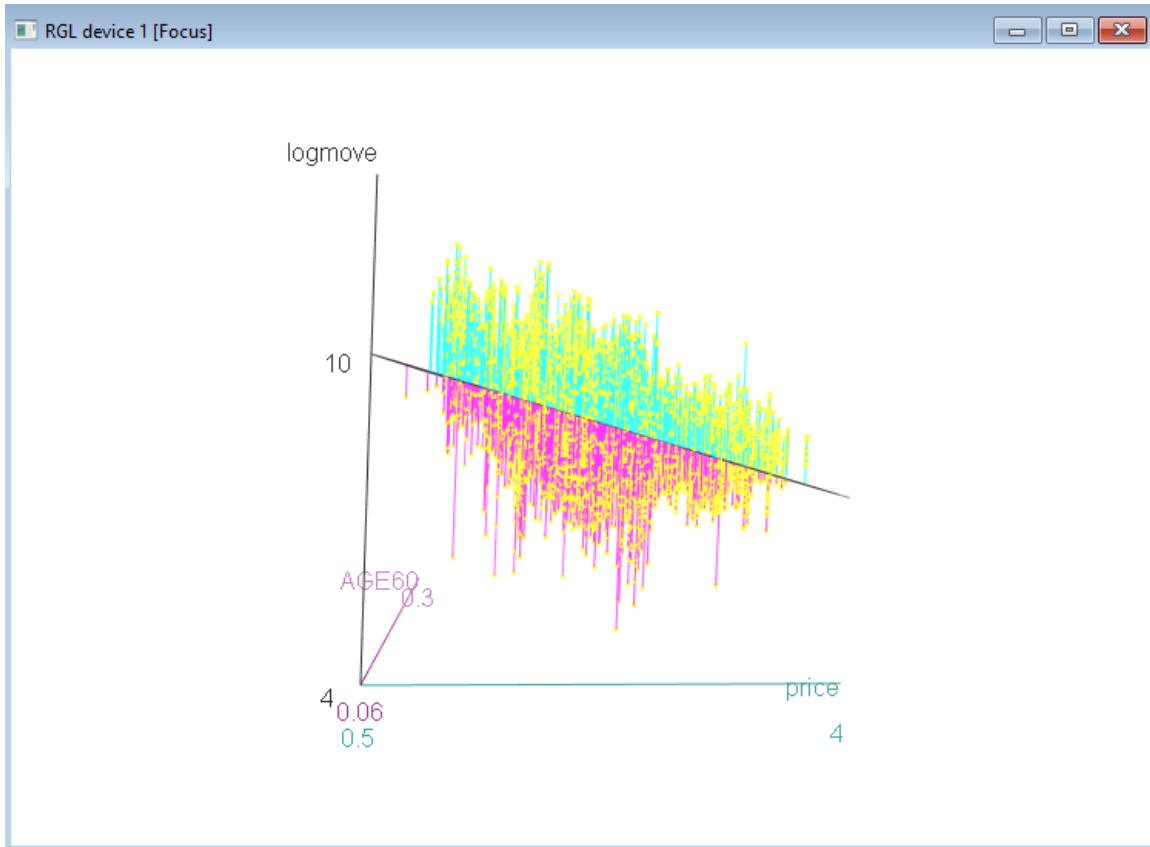
1. Graphs, 3D Graph, 3D Scatterplot
2. Select AGE60 and price as explanatory variables by holding down the control key, then clicking on AGE60 and price
3. Select logmove (log of sales) as the response variable



4. Click on the Options tab, select Linear least-squares



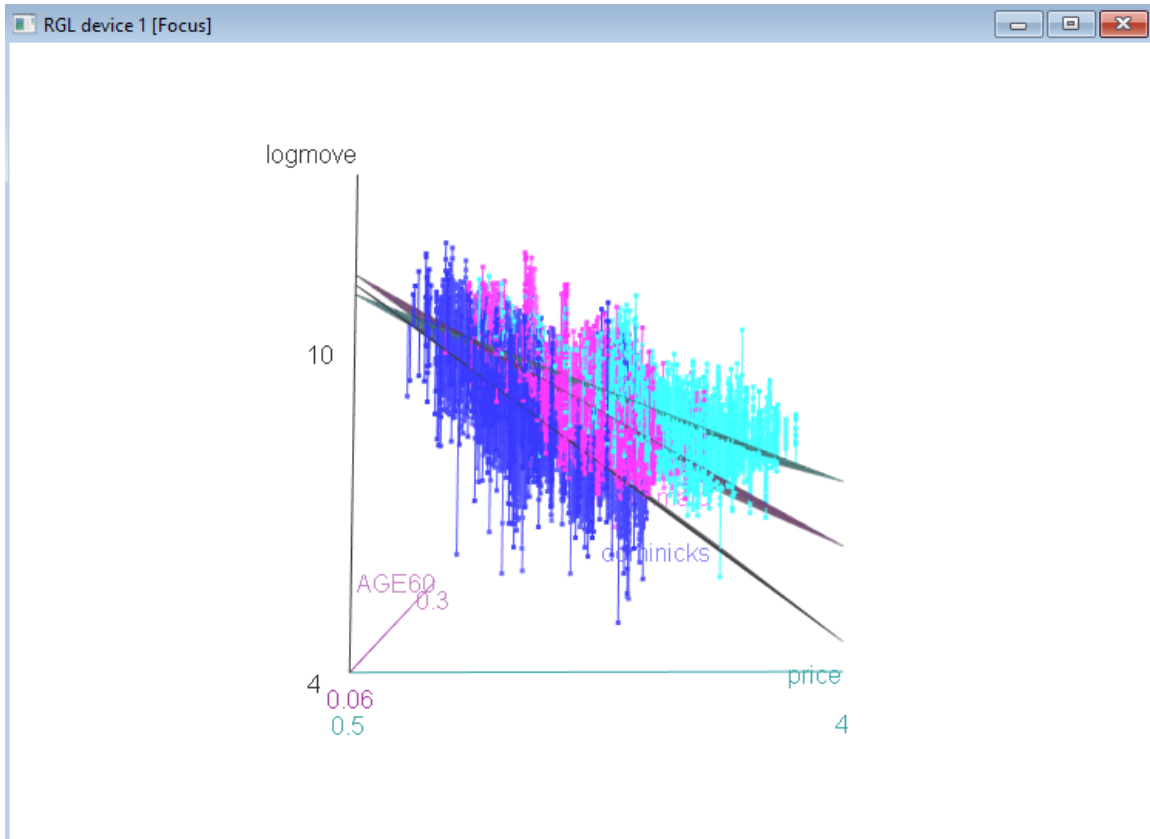
5. Click OK
6. Note: the graph might be behind one of your screens. Expand the window by clicking on the box in the upper right corner of your graph
7. Rotate the graph by clicking on the graph with your mouse, hold the mouse button down, and move



8. The plane is the regression plane that shows how price and age affect the logarithm of sales (logmove)
9. The dots are the data points
10. The lines from the plane to the dots are error terms, called residuals
11. Does price affect sales?
12. Does age affect sales?

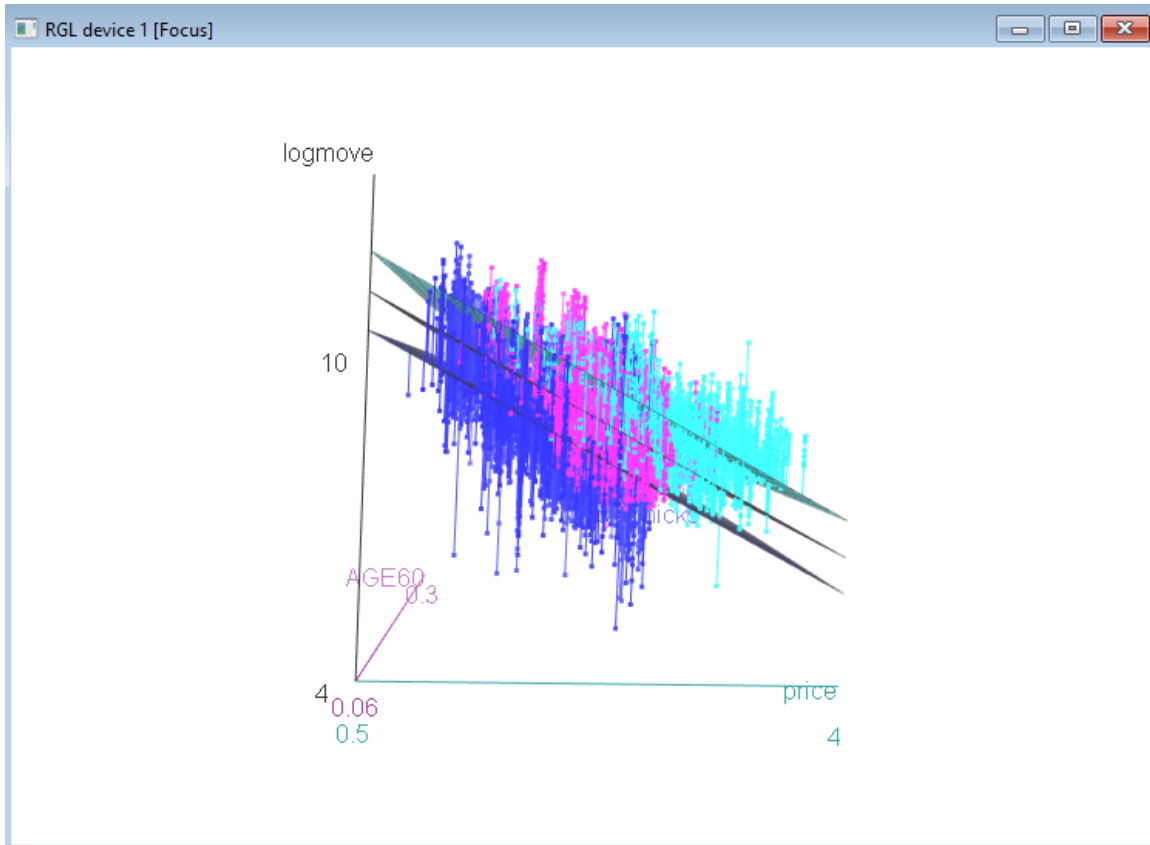
To generate 3D graphs by brand,

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



20. Which brand is more sensitive to price?
21. This model has a different slope for each brand
22. Different slopes reflect different elasticities of demand

23. Now rerun the 3D graph by clicking on Graphd, 3D Graphs, 3D Scatterplot.
24. Click on Plot by: brand, then check the box Parallel regression surfaces, then click OK, and OK again
25. This model has a different intercept for each brand, but the same slope
26. Different intercepts for each brand represent brand premium.



27. The technical term for different slopes and different intercepts are:
 - a. Dummy variables produce different intercepts
 - b. Moderating effects (interactions) produce different slopes

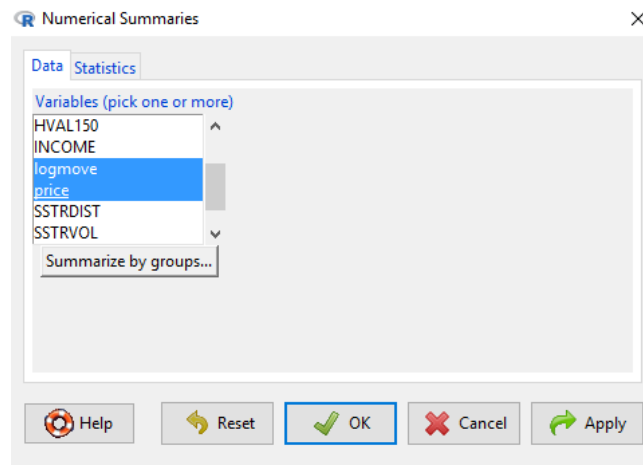
Saving Graphs

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

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

```
Rcmdr> numSummary(oj[, c("logmove", "price"), drop = FALSE], statistics = c("mean", "sd",  
Rcmdr+ "IQR", "quantiles"), quantiles = c(0, 0.25, 0.5, 0.75, 1))  
      mean      sd      IQR      0%      25%      50%      75%      100%  
logmove 9.167864 1.0193782 1.275069 4.158883 8.489616 9.03408 9.764685 13.48202  
price    2.282488 0.6480007 0.940000 0.520000 1.790000 2.17000 2.730000 3.87000  
      n  
logmove 28947  
price    28947
```

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

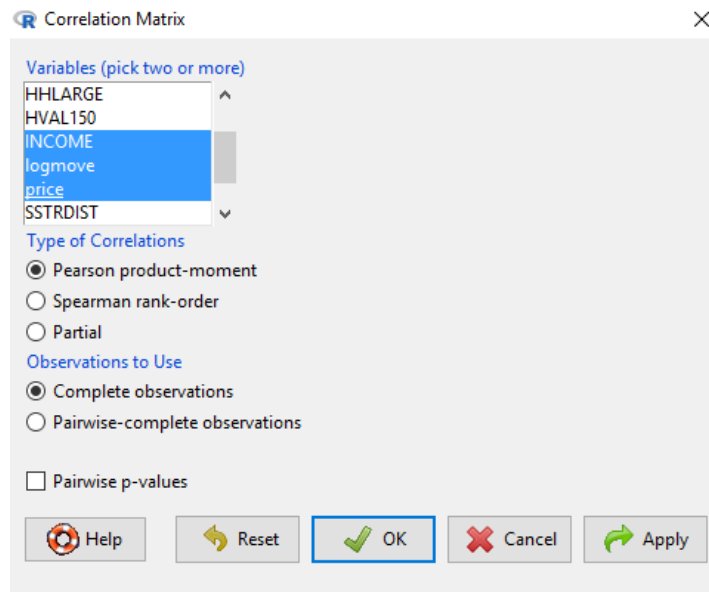
4. Click on Statistics, Summaries, Numerical Summaries
5. Select logmove, price by holding down the control key
6. Click Summarize by Groups, click brand, then OK
7. Click OK

```
variable: logmove  
      mean      sd      IQR      0%      25%      50%      75%      100%      n  
dominicks 9.174831 1.1929370 1.5619512 4.158883 8.392990 9.121728 9.954941 13.48202 9649  
minute.maid 9.217278 0.9852867 1.3523928 5.768321 8.476371 9.026418 9.828764 13.29018 9649  
tropicana 9.111483 0.8473800 0.9685592 5.257495 8.565602 8.987197 9.534161 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
```

Correlation

To generate a correlation matrix,

1. Click on Statistics, Summaries, Correlation Matrix, Pearson (for linear relationships)
2. Hold down the control key and select INCOME, logmove, price



3. Click OK

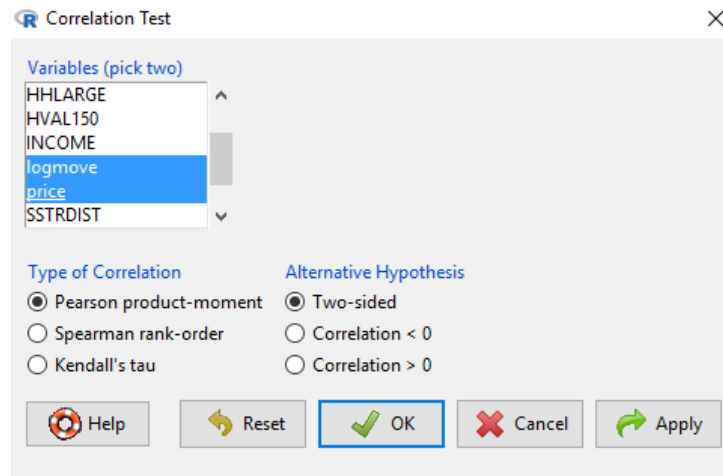
```
              INCOME      logmove      price
INCOME  1.00000000 -0.04277891 -0.03049071
logmove -0.04277891  1.00000000 -0.43198299
price   -0.03049071 -0.43198299  1.00000000
```

4. Click on Statistics, Summaries, Correlation Matrix, Spearman (for non-linear relationships)
5. Hold down the control key and select INCOME, logmove, price
6. Click OK

```
              INCOME      logmove      price
INCOME  1.00000000 -0.009233931 -0.02342516
logmove -0.009233931  1.00000000 -0.44228695
price   -0.023425164 -0.442286955  1.00000000
```

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

7. Click on Statistics, Summaries, Correlation Test, Pearson (for linear relationships)
8. Select both logmove and price



9. Click OK

Pearson's product-moment correlation

```
data: logmove and price
t = -81.49, 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
```

10. The p-value is less than 0.05, so the correlation is statistically significant.
11. Click on Statistics, Summaries, Correlation Test, Spearman (for non-linear relationships)
12. Select both logmove and price
13. Click OK

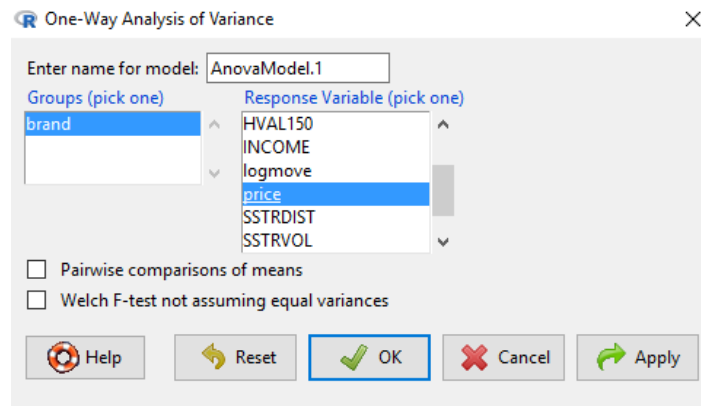
Spearman's rank correlation rho

```
data: logmove and price
S = 5830571263731, p-value < 2.2e-16
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
-0.442287
```


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

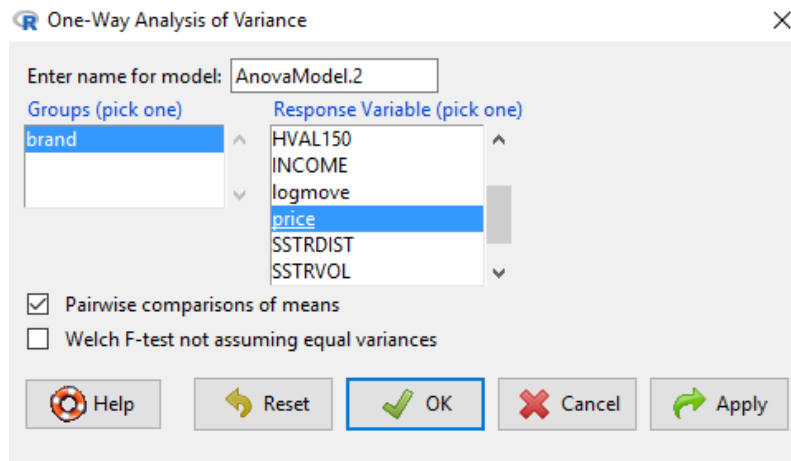
```
> summary(AnovaModel.1)
              Df Sum Sq Mean Sq F value Pr(>F)
brand           2   6236   3118.2   15250 <2e-16 ***
Residuals    28944    5918     0.2
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> with(oj, numSummary(price, groups=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
```

4. 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.

To determine which products are different, we need to perform a pairwise comparison.

5. Click on Statistics, Means, One-way ANOVA
6. For the response variable, click on price
7. Check the box Pairwise comparison of means



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

Simultaneous Tests for General Linear Hypotheses

Multiple Comparisons of Means: Tukey Contrasts

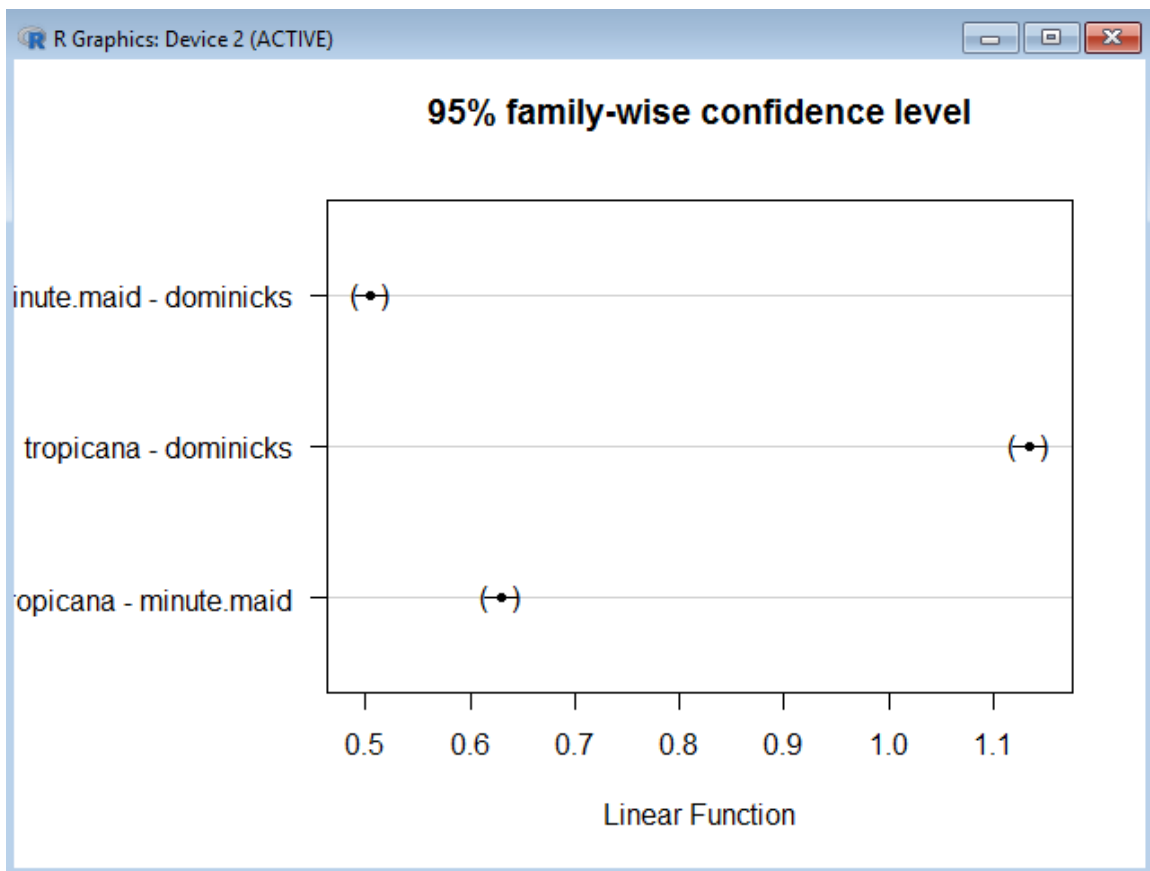
```
Fit: aov(formula = price ~ brand, data = oj)
```

Linear Hypotheses:

	Estimate	Std. Error	t value	Pr(> t)
minute.maid - dominicks == 0	0.50535	0.00651	77.62	<2e-16 ***
tropicana - dominicks == 0	1.13468	0.00651	174.29	<2e-16 ***
tropicana - minute.maid == 0	0.62933	0.00651	96.67	<2e-16 ***

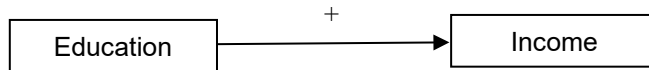
 Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
 (Adjusted p values reported -- single-step method)

10. The graph portrays the estimate of price difference between brands and the confidence interval. If the difference is not zero, then we can conclude that the product prices compared are different.



Modeling – Regression

So far, we have been performing regressions on a dependent variable Y against an independent variable X. For example, we can examine how education (X) affects income (Y). Pictorially, this would appear as:



The line and arrow identify a relationship between education and income. The plus sign above the line indicates that the relationship is positive, i.e., if education increases, then income increases.

This relationship can be written as:

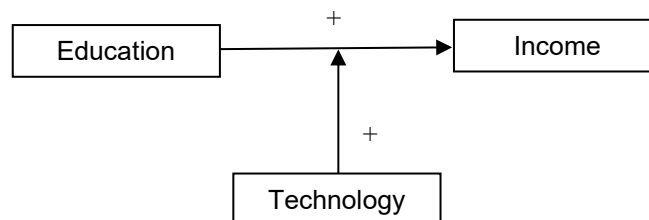
$$\text{Income} = f(\text{Education})$$

Which means that income is a function of education. One formulation of this could be the linear relationship:

$$\text{Income} = \beta_0 + \beta_1 * \text{Education}$$

Where β_0 is the intercept and β_1 is the coefficient for education.

Now consider a third variable: technology. Technology has the potential for increasing the value of educated employees. Technology itself does not generate income for an employee but affects the value of education. This is called a moderating variable and is shown as:



This new model means that as education increases, income increases. The moderating effect of technology on education implies that technology further increases the value of education. This is modeled as an interaction term:

$$\text{Income} = \beta_0 + \beta_1 * \text{Education} + \beta_2 * \text{Education} * \text{Technology}$$

Therefore, the effect of education on income is influenced by the level of technology that an employee has.

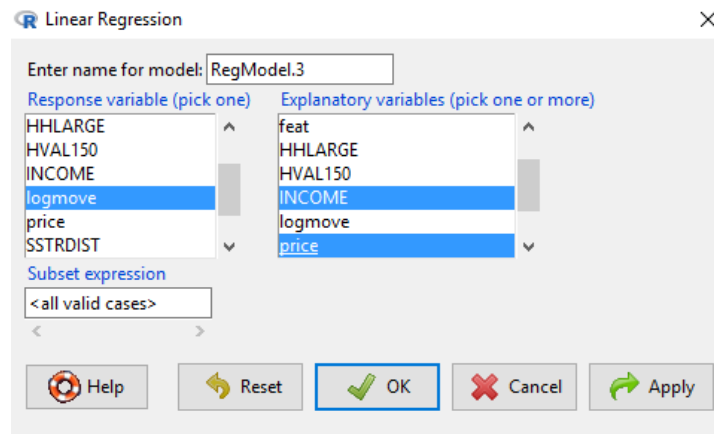
Summary

A dummy variable changes the intercept. A moderating effect (interaction) changes the slope.

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

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
```

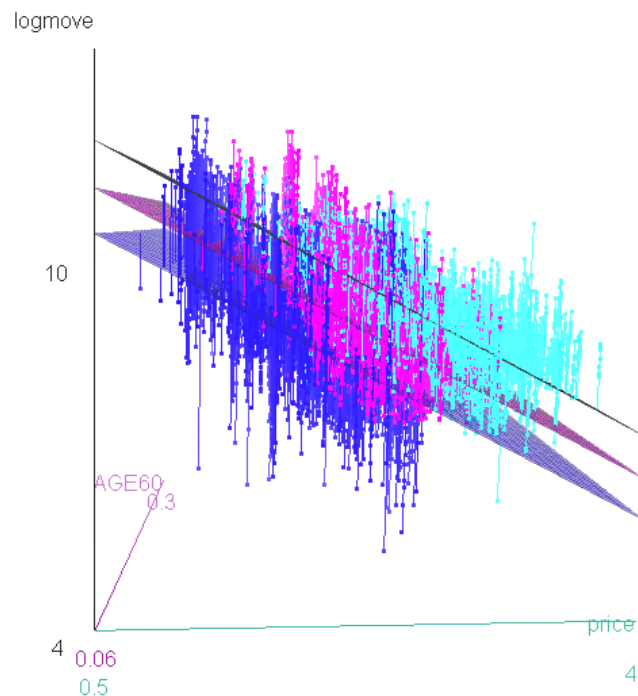
5. Is the equation statistically significant? If the p-value of the F-statistic is < 0.05 , then the equation is statistically significant.
6. How much of the variability in the log of sales is explained by the explanatory variables? The R^2 measures explanatory power.
7. Which explanatory variables are statistically significant?
8. How does each explanatory variable affect sales? Which affect it positively and which negatively? How do you interpret this (what does it really mean)?

Dummy Variables

1. Dummy variables are variables that take on the value of zero or one. For example, a dummy variable for homeowner would be 1 when the person is a homeowner and zero when the person is not a homeowner.
2. Dummy variables change the intercept in a regression equation.
3. Any categorical variable can be coded as a dummy variable. For example, education status can be coded as 1 for student, 0 for non-student. Similarly, employment status can be coded as 1 for employed, 0 for unemployed
4. When a category has two possibilities (student, non-student), you only need one variable to represent the two categories
5. When a category has three or more possibilities, (Tropicana, Minute Maid, Dominicks), then you need $n-1$ dummy variables, where n is the number of categories. In the orange juice example, we would have a dummy variable for Minute Maid and Tropicana. The intercept for Dominicks would be the regular intercept (called base case). The intercept for Minute Maid would be the base intercept plus the Minute Maid dummy coefficient. The intercept for Tropicana would be the base intercept plus the Tropicana dummy coefficient.
6. The following picture is an example of the use of a dummy variable

RGL device 1 [Focus]

— □ ×

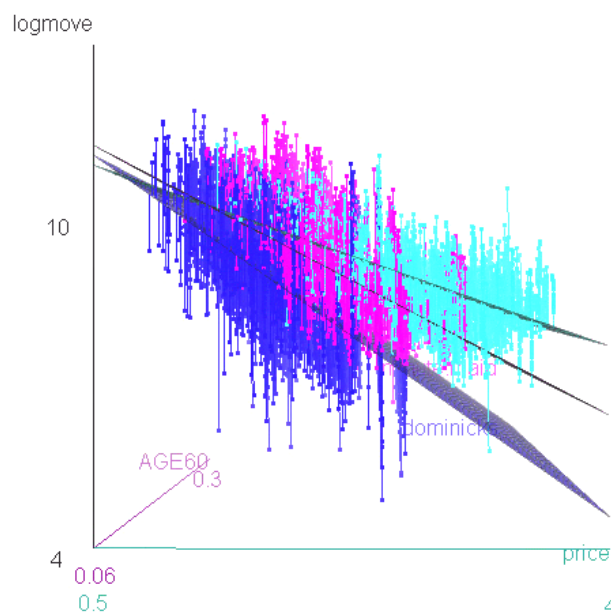


Moderating Effect (also called Interaction)

1. A moderating effect occurs when one variable magnifies the effect of another variable.
2. A moderating effect changes the slope in a regression equation.
3. A moderating effect is modeled by multiplying two variables together.
4. The following picture is an example of a moderating effect (interaction) where the brand interacts with price.
5. The effect of price on logmove for each brand is reflected in the different slopes.

RGL device 2 [Focus]

— □ ×



Regression with Dummy Variables

A dummy variable in a regression can assist in determining if the intercept changes when the brand changes. To perform this more sophisticated regression,

1. Click on Statistics, Fit Models, Linear Model
2. Click Reset
3. Double click on logmove to make it the Y or dependent variable
4. Double click on AGE60 to make it an X or explanatory variable
5. Double click on INCOME to make it an X or explanatory variable
6. Double click on price to make it an X or explanatory variable
7. Double click on brand (notice it says it's a factor) to create a dummy variable for the brands. A dummy variable takes on the value of zero or one; zero if it is not that brand, one if it is that brand.
8. Click on OK

```
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
```

9. The equation becomes:

$$\begin{aligned}\text{Logmove} = & 13.13 + 1.97 * \text{AGE60} - 0.18 * \text{INCOME} - 1.35 * \text{price} \\ & + 0.73 * \text{brand}(\text{minute.maid}) \text{ dummy} \\ & + 1.47 * \text{brand}(\text{tropicana}) \text{ dummy}\end{aligned}$$

10. Notice that there is a dummy for minute maid and Tropicana, but not dominicks. Dominicks is the base, so the intercept represents dominicks intercept.
11. 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 $13.12952 + 1.47162$.
12. This means, all else being equal, the log of sales is highest for Tropicana.

Moderating effects (interactions of price and brand)

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

The screenshot shows the 'Linear Model' dialog box in R. The 'Enter name for model:' field contains 'LinearModel.5'. The 'Variables (double-click to formula)' list includes AGE60, brand [factor], CPDIST5, CPWVOL5, EDUC, and ETHNIC. The 'Model Formula' section shows the formula 'logmove ~ AGE60 + INCOME + price + brand + price*brand'. The 'Operators (click to formula):' section includes +, *, :, /, %in%, -, ^, (, and). The 'Splines/Polynomials: (select variable and click)' section includes B-spline, natural spline, orthogonal polynomial, and raw polynomial. The 'df for splines:' is set to 5 and 'deg. for polynomials:' is set to 2. The 'Subset expression' is set to '<all valid cases>' and the 'Weights' are set to '<no variable selected>'. The 'OK' button is highlighted.

12. 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.
13. We need to include the price*brand effect for minute maid and Tropicana. For minute maid, the coefficient on price is $-1.94480 + 0.47545 = -1.46935$.
14. For Tropicana, the coefficient is $-1.94480 + 0.94817 = -0.99663$.
15. Which brand is more sensitive to price?

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