Overview of R

Overview of R

- R is an open-source programming language for statistical analysis, data mining, and machine learning.
  - The S language was created at Bell Labs in 1976.
  - S was augmented with documentation and called S-Plus (S plus documentation) in 1988.
  - Ross Ihaka and Robert Gentleman at the University of Auckland, New Zealand, launched R in 1995 as a free and open-source version of S (they named it after their first initials).
- R is hosted worldwide on mirrored servers; the intent is to have multiple identical copies available to increase availability.
- R can be run via an R Console or the RStudio development platform.
- R also has available the Rcmdr (R Commander) graphical user interface and Rattle data mining packages, in addition to many others.

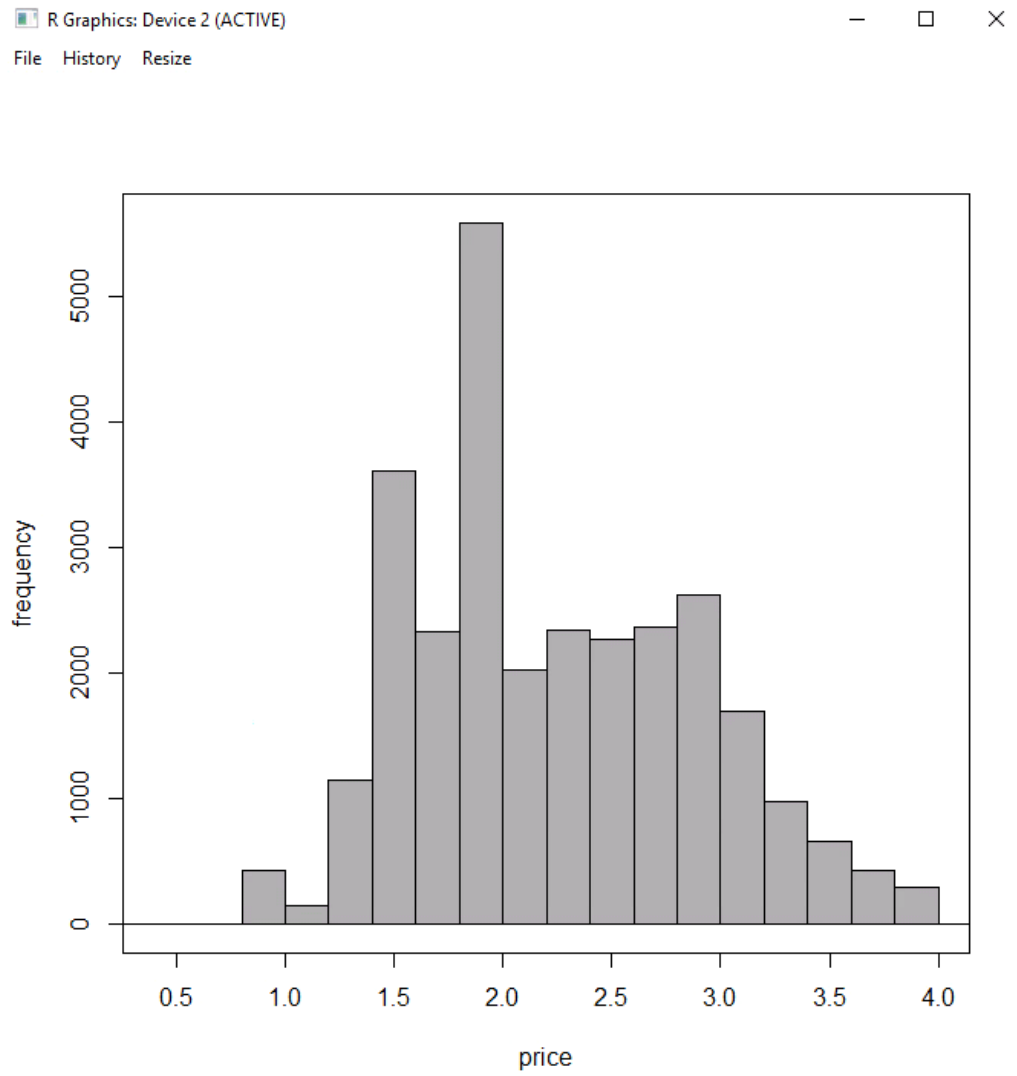
The End

R

Histograms, Box Plots, Scatter Plots, Mean Plots, XY Plots

R: Histograms, Box Plots, Scatter Plots, Mean Plots, XY Plots, Part I

Histograms display the frequency of data within an interval, often called a bin



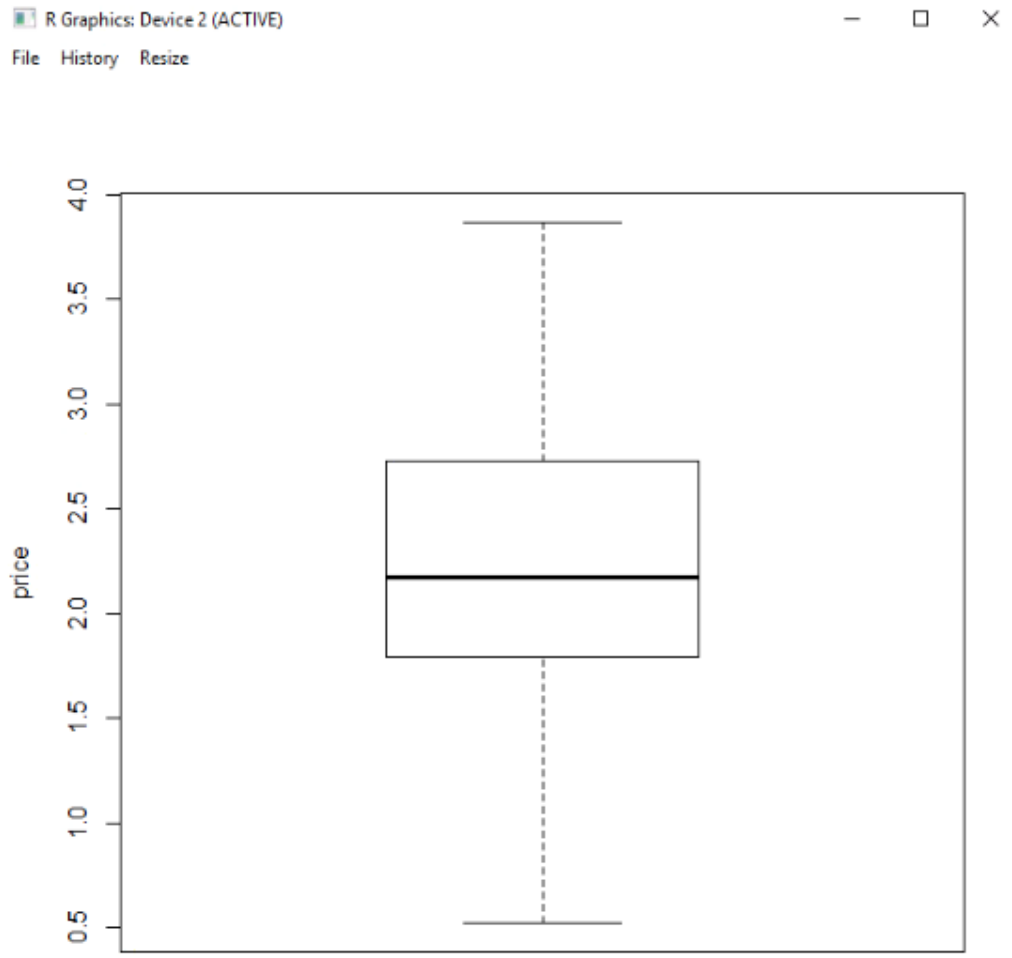
R: Histograms, Box Plots, Scatter Plots, Mean Plots, XY Plots, Part II

Box plots show the quartiles of data

- Minimum
- First quartile or 25%-ile
- Second quartile or 50%-ile or median
- Third quartile or 75%-ile
- Maximum
- Interquartile range (IQR) (third quartile minus first quartile)
- Outliers: data points more than 1.5 times IQR above the third quartile or below the first quartile

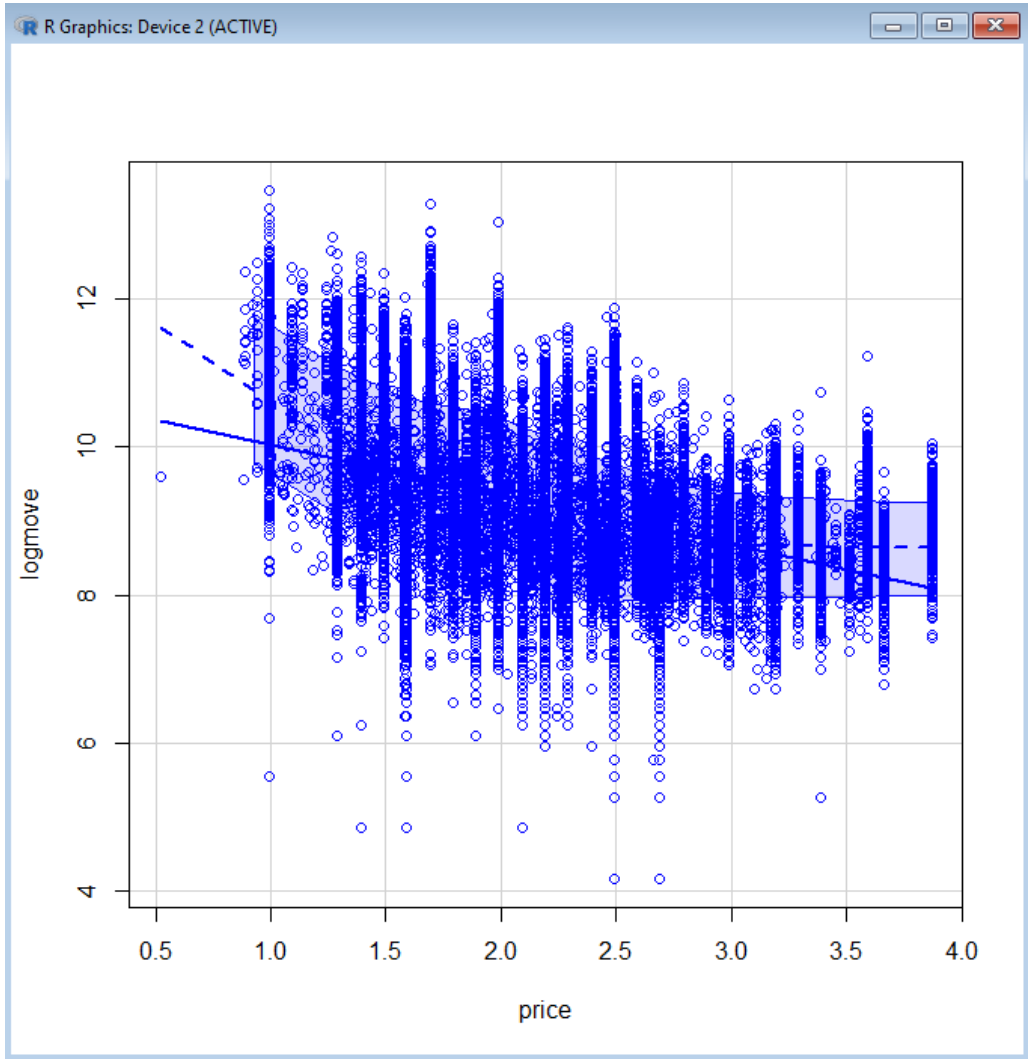
R: Histograms, Box Plots, Scatter Plots, Mean Plots, XY Plots, Part III

Box plot example



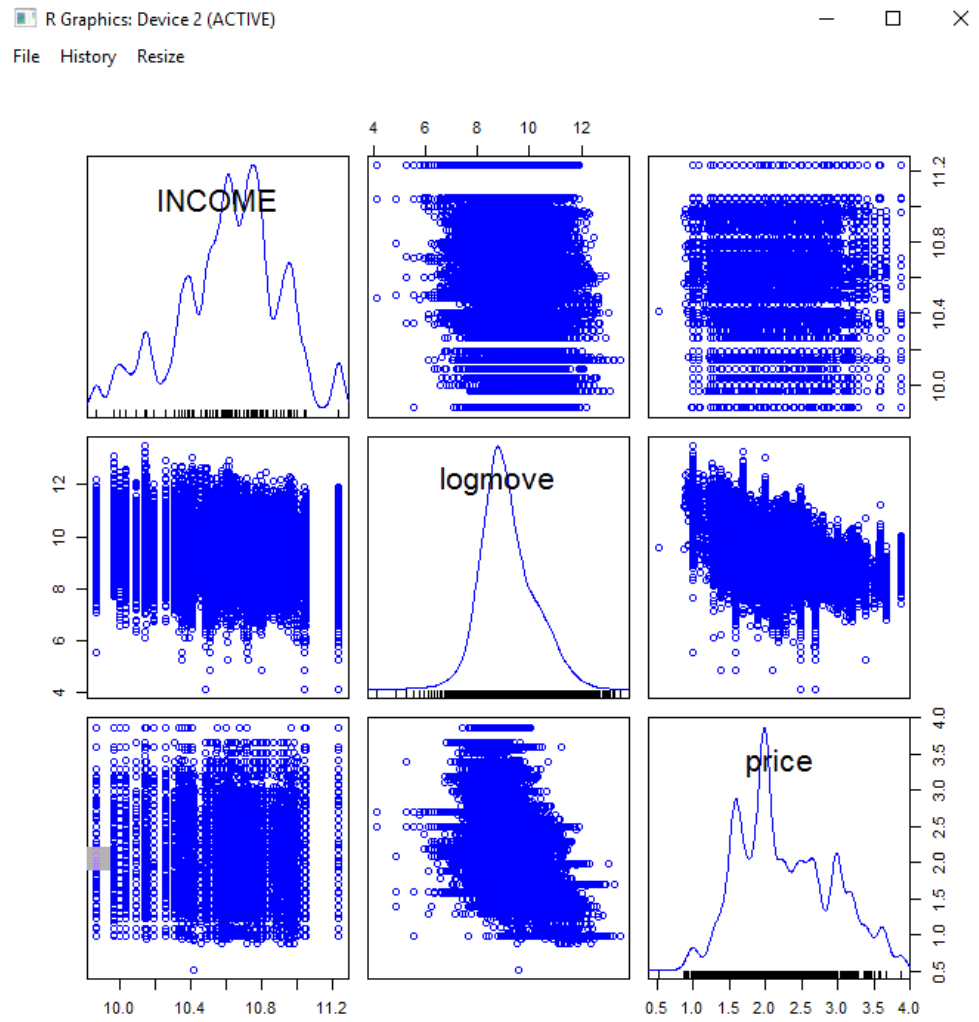
R: Histograms, Box Plots, Scatter Plots, Mean Plots, XY Plots, Part IV

Scatter plots display the data points on an X and Y axis, with only one X variable and one Y variable



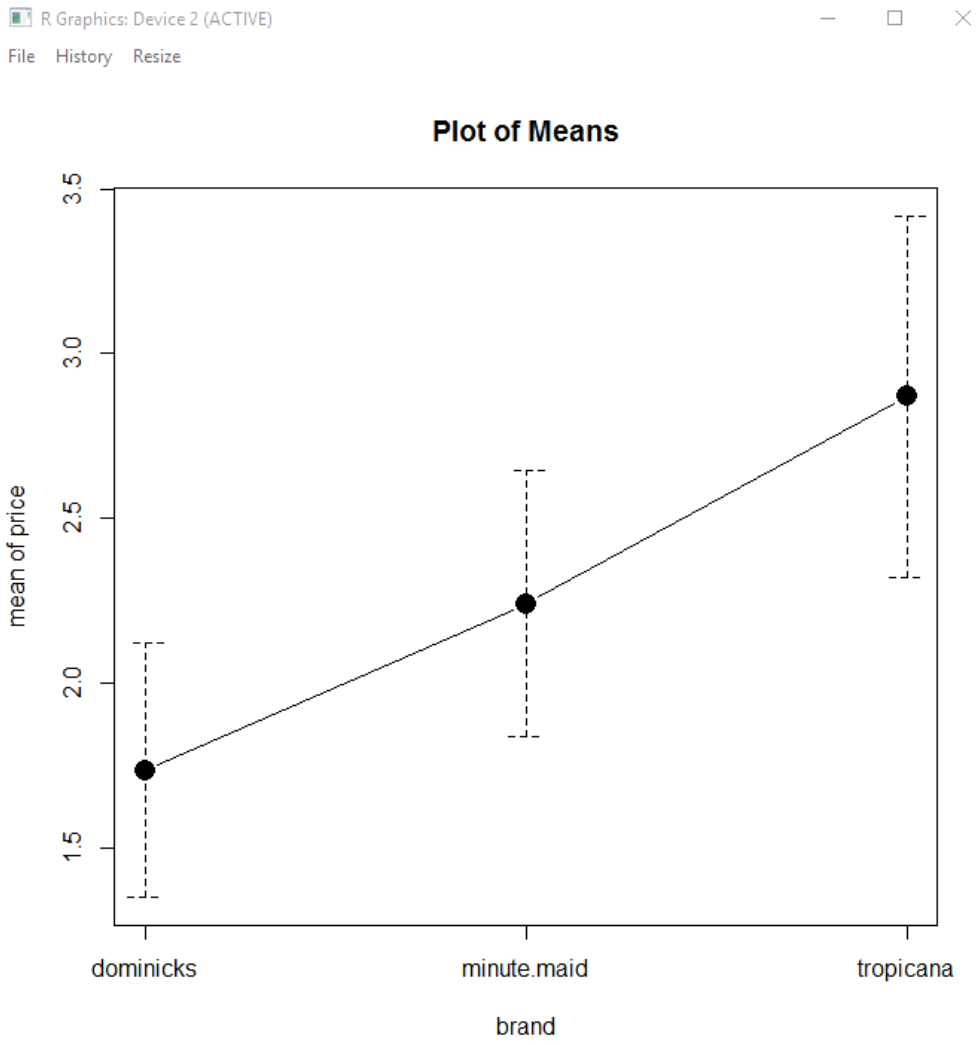
R: Histograms, Box Plots, Scatter Plots, Mean Plots, XY Plots, Part V

Scatter plot matrix shows data points on multiple scatter plots simultaneously, with multiple X and Y variables



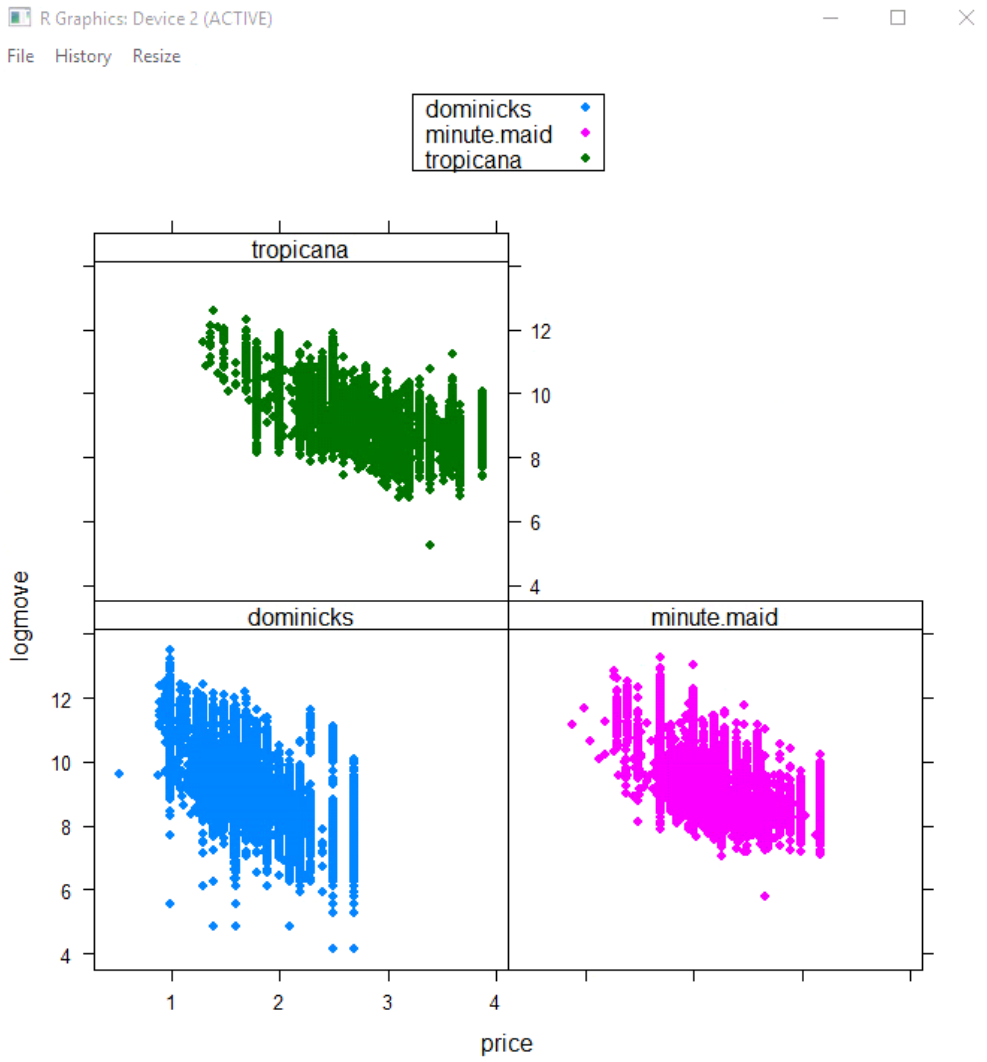
R: Histograms, Box Plots, Scatter Plots, Mean Plots, XY Plots, Part VI

Mean plots display the averages of multiple groups



R: Histograms, Box Plots, Scatter Plots, Mean Plots, XY Plots, Part VII

XY plots show scatter plots with different groups in each scatter plot



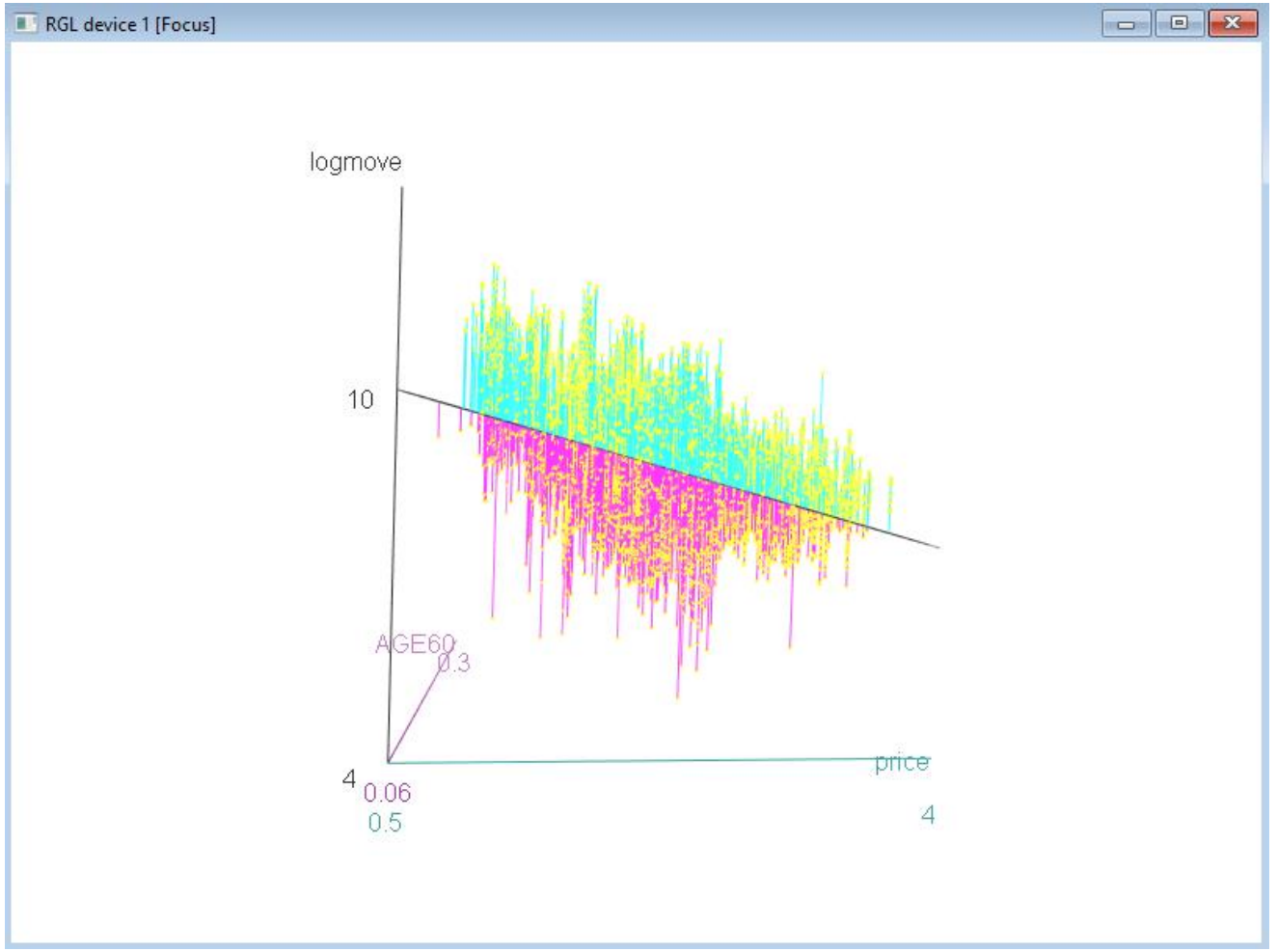
The End

R

3-D Graphs

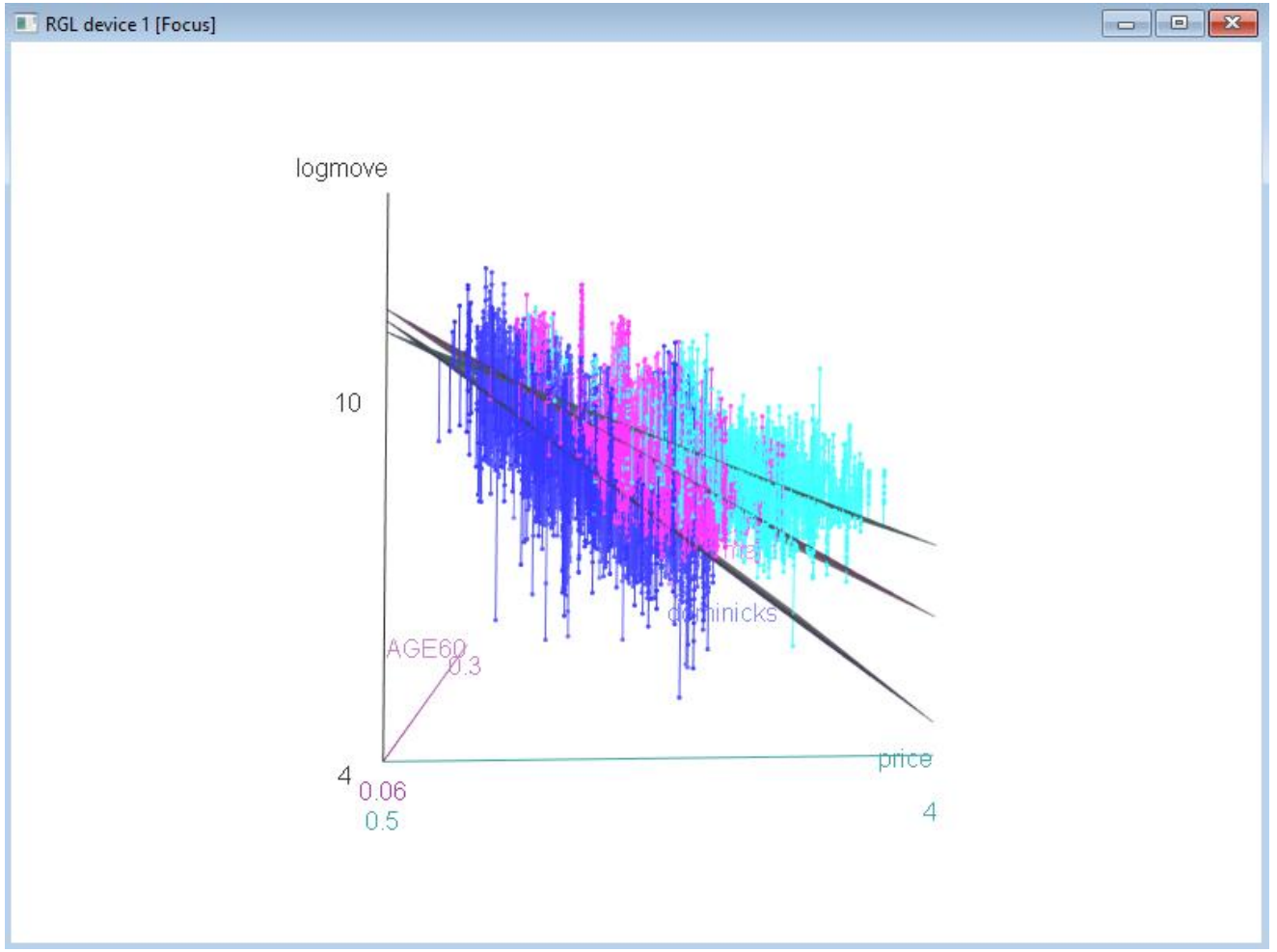
R: 3-D Graphs

3-D graphs display data in three dimensions with the ability to rotate the graph.



R: 3-D Graphs (cont.)

3-D graphs can plot data by groups to identify differences among groups.



The End

R

Statistical Summaries

R: Statistical Summaries

- R generates descriptive statistics including:
  - Mean
  - Standard deviation
  - Interquartile range
  - Quartiles
- Descriptive statistics can be calculated for an entire dataset or by groups

The End

R

Correlations

R: Correlations

Correlations measure the direction of relationship between two variables and the strength of the relationship.

- Correlation ranges from –1 to +1.
- Positive correlation means the variables move in the same direction.
- Negative correlation means the variables move in different directions.
- Correlations close to 0 are weak; far from 0 are strong.

The End

R

ANOVA

R: ANOVA

- Analysis of variance (ANOVA) examines groups and determines if the mean (average) of one group is different from the others.
- Pairwise ANOVA identifies which group average is different from the others.

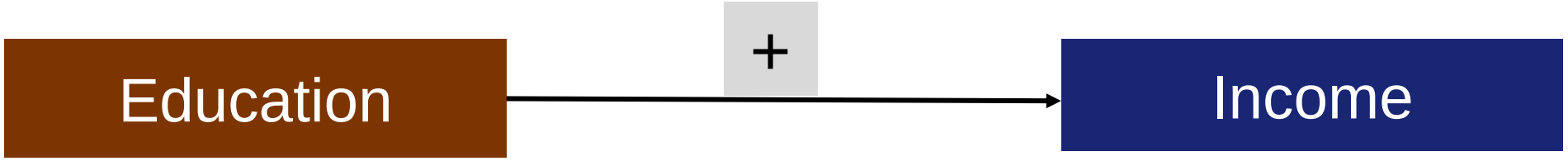
The End

R

Regression and Modeling

R: Regression and Modeling

- Regression determines the relationship between one or more X-variables and one Y-variable, measuring the direction and magnitude of change of Y when X changes

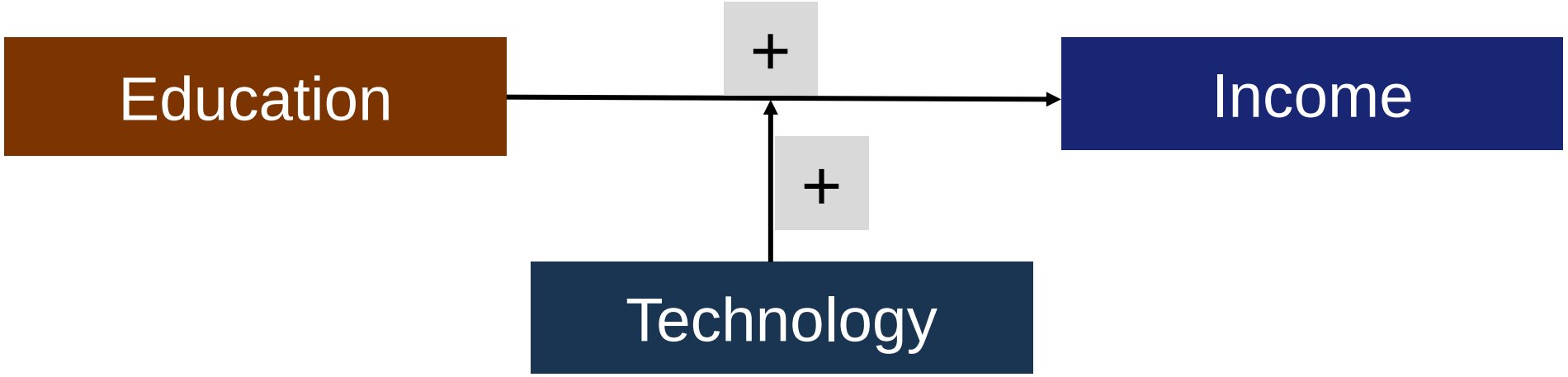


- In the example of income as the Y (dependent variable) and X (independent variable), the equation would be:

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

R: Regression and Modeling (cont.)

- Consider a third variable, technology, that magnifies or reduces the effect of education on income (interaction)



- In this example, the equation would be:

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

The End

R

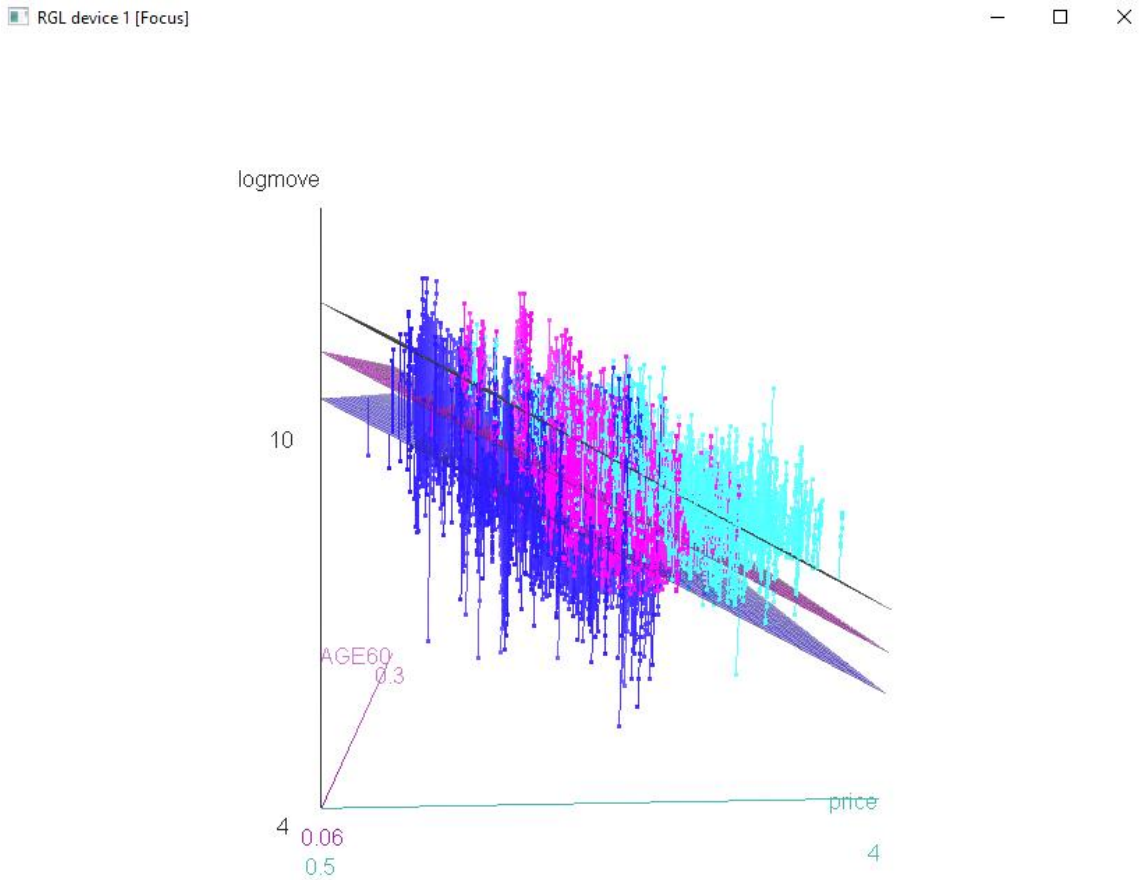
Regression with Dummy Variables

R: Regression with Dummy Variables

- Dummy variables allow one to change the intercepts in a regression for different groups.
- The number of dummy variables required is one less than the number of groups, e.g., if examining the demand curves for three types of orange juice, then two dummy variables are required.

R: Regression with Dummy Variables (cont.)

- Demand curves with three types of orange juice.
- Dummy variables create different intercepts.
- The difference between the lines represent brand premium.



The End

R

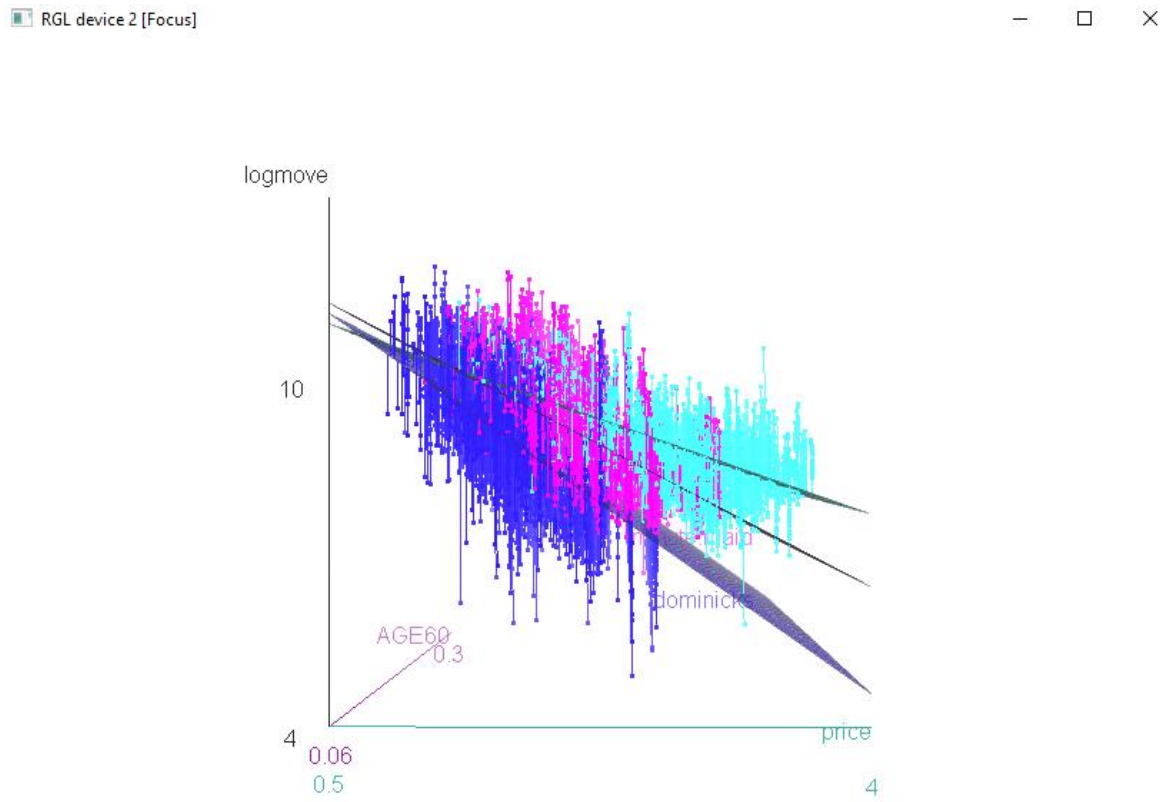
Regression with Moderating Effects

R: Regression with Moderating Effects

- Moderating effects change the slopes in a regression for different groups.
- Moderating effects, also called interaction terms, are created by multiplying together two variables.

R: Regression with Moderating Effects (cont.)

- Demand curves with three types of orange juice.
- Moderating effects change the slope for each brand of orange juice.
- The different slopes represent different elasticities of demand.



The End