DS3: Visualizing Data and Regression

Instructors: Jacob LaRiviere

Emails: jlariv@microsoft.com

1) Download the orange juice data from the course website and create an Rmd script for this assignment.
2) Change the working directory so that R knows where to look for the data (tip: create a Econ404 folder and save datasets there). See `setwd()`. [You can type `?setwd` to see the help file.]
3) Read in the data, see `read.csv`. `oj` is a data frame with many variables. You can click on the dataframe in the top right corner of Rstudio to explore. You can refer to any variable with `oj$var` where "`var_name`" is the variable of interest. We will also refer to `df` as a generic term for a "dataframe"
4) Visualizing price.
    a. Make a box plot of price.
        i. Use the ggplot2 package to do this. ggplot2 is kind of quirky but powerful package. You'll need to start by calling the package once you've installed it:
        ```
        library(ggplot2)
        ggplot(df, aes(factor(var_name1), var_name2)) +
        geom_boxplot(aes(fill = factor(brand)))
        ```
        The first line above calls the ggplot and tells it to use the dataframe `df`.
        `aes` is short for "aesthetics"
        the term `factor(var_name1)` tells it to create a unique plot by each unique value in `var_name1`.
        the second variable listed `var_name2` tells it to use that variable in creating the boxplot.
        The second part of the line `+ geom_boxplot(aes(fill = factor(var_name1)))` tells it to make a boxplot and color each one by `var_name1`.

    b. Make a box plot of log price.
    c. Make a box plot of price, but separate out each brand.
    d. Do the same for log price.
    e. **What do these graphs tell you about the variation in price? Why do the log plots look different? Do you find them more/less informative?**

5) Visualizing the quantity/price relationship
    a. Plot `logmove` (log quantity) vs. `log(price)` for each brand. For this one the appropriate second part of the ggplot command will be: `+ geom_point(aes(color = factor(var_name)))`
        i. **What do insights can you derive that were not apparent before?**
6) Estimating the relationship.

a. Do a regression of log quantity on *log price*. **How well does the model fit? What is the elasticity, does it make sense?**
b. Now add in an intercept term for each brand (add brand to the regression), **how do the results change?**
c. Now figure out a way to allow the elasticities to differ by brand. Search "interaction terms" and "dummy variables" if you don't remember this from econometrics. Note the estimate coefficients will "offset" the base estimates. **What is the insights we get from this regression? What is the elasticity for each firm? Do the elasticities make sense?**
d. Super Star Status: Hold out 20% of your sample randomly. Estimate the model on the remaining 80%. Use the predict command to see how well the model fits on the rest of the data (e.g., `y_hat <- predict(my_model, newx = test_matrix))`

7) Impact of "featuring in store". The "`feat`" variable is an indicator variable which takes the value of one when a product is featured (e.g., like on an endcap display)
   a. Which brand is featured the most? **Make a ggplot to show this**. Hint: using `position = "jitter",` within the `aes(color = factor(var_name))` of ggplot is one way to do this.
      i. What is the average price and featured rate of each brand? Hint: `aggregate(df[, x:y], list(df$var_name), mean)` where `x` and `y` are the column numbers of the two variables you care about. See if you can do this with the `dplyr` package.
   b. How should incorporate the feature variable into our regression? Start with an additive formulation (e.g. feature impacts sales, but not through price).
   c. Now run a model where features can impact sales and price sensitivity.
   d. Now add what you think are the most relevant sociodemographic controls and **produce the regression results from that regression as well.**
8) Overall analysis
   a. **Based on your work, which brand has the most elastic demand, which as the least elastic?**