DS3: Visualizing Data and Regression

Instructor: Jacob LaRiviere

Email: [jlariv@microsoft.com](mailto:jlariv@microsoft.com)

* Download the orange juice data from the course website and create an Rmd script for this assignment.
* Change the working directory so that R knows where to look for the data (tip: create a folder and save datasets there). See setwd(). [You can type ?setwd to see the help file.]
* 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”
* Visualizing price.
* Make a box plot of price.
* 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.

* Make a box plot of log price.
* Make a box plot of price, but separate out each brand.
* Do the same for log price.
* **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?**log plot makes the outliers more visible.

there are alot more outlier on the lower side of price for all brands.

* Visualizing the quantity/price relationship
* 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)))
* **What do insights can you derive that were not apparent before?**

there is an inverse relationship between log quantity and log price.

* Estimating the relationship.
* Do a regression of log quantity on *log price*. **How well does the model fit? What is the elasticity, does it make sense?**
* Now add in an intercept term for each brand (add brand to the regression), **how do the results change?**
* 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?**
* 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))
* 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](https://easyshiftapp.zendesk.com/hc/en-us/articles/206348625-Endcap-Display))
* 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.
* 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.

Now do this with the dplyr package: ddply(oj, .(brand), summarize, feat=mean(feat),price=mean(price)). That package is really nifty. You can also break this out by feat and not-featured: ddply(oj, .(brand,feat), summarize, mean\_price=mean(price),sd\_price=sd(price), obs=length(price))

* How should incorporate the feature variable into our regression? Start with an additive formulation (e.g. feature impacts sales, but not through price).
* Now run a model where features can impact sales and price sensitivity.
* Now add what you think are the most relevant sociodemographic controls and **produce the regression results from that regression as well.**
* Overall analysis
* **Based on your work, which brand has the most elastic demand, which as the least elastic?**