**Data Wrangling and Analysis Process - Amazon versus In-Store Prices**

Ming Waters

An open source data set produced by Alberto Cavallo, co-founder of MIT’s Billion Price Project was used for this project. The original data set was used to determine if prices in store was significantly different than prices on-line. Cavallo and team gathered prices from multiple products, in multiple countries on-line and in-store. Multiple teams gathered prices manually and through a price scrapping app developed by Cavallo.

For this project, a subset of the full data set was used. The data set contained 5 categories of products with prices on multiple items in-store and online.

**Data Set:**

My project will wrangle and analyze a small subset of Cavallo’s multi-country data. I will focus on the US price comparisons only comparing the in store prices with the Amazon price.

Raw Data file was downloaded from: <https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi%3A10.7910%2FDVN%2FXXOUHF>

Goal1– To determine if there is a cost benefit of purchasing various product categories at Amazon or in the physical store

Goal2 – To determine if there is a cost benefit of purchasing at Amazon or in the store when the items are regular priced or on sale regardless of the categories.

**Data Wrangling Approach:**

* Import data original\_amazon\_compare.csv into R
* Used dplyr library to wrangle data set
* Used ggplot2 library to visualize data
* Remove outliers
* Determine if the percent difference in price by category has a normal distribution and an ANOVA analysis can be performed to determine significant difference in pricing by category.
* Determine next steps if normality is not achieved

The data contains many columns (variables) that are not needed for my analysis so I will first clean the data table and only select the columns I need...

price – price in store

id – item number

price\_amazon – price on amazon

sale\_online – the item was on sale online

PRICETYPE – if the item was on sale in store

datediff – difference between time purchase online – in store

category – groupings of product type (electronics, office supply, etc.)

The data contained prices across multiple price points. To offset weighing deltas from higher priced items, percent difference was used to compare difference.

p\_difference was calculated based on the delta of price from in-store versus on-line

p\_difference = (amazon – in-store)/ in-store

**Data Errors –**

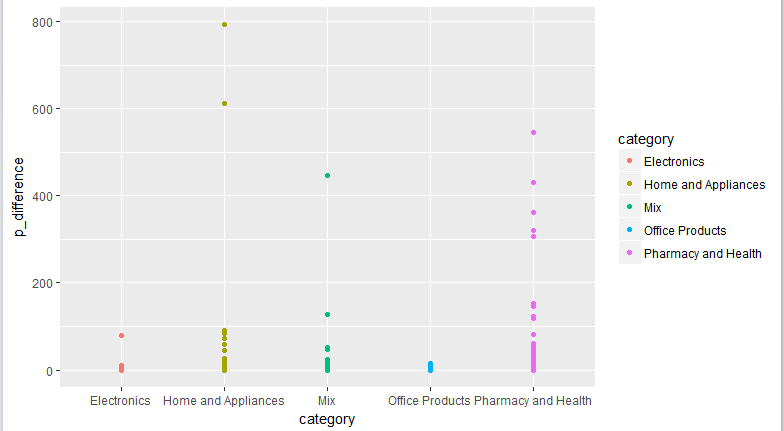
The next step in analyzing data was to visualize the various categories to develop an approach to analysis. Upon inspection, it was evident that there were several large errors embedded in the data. Data errors include –

* Product ID transferring into In-Store Prices
* Errors in decimal placement - ex – 6.99 (in-store) but typed as 69.9 (amazon)
* Errors in data entry – ex – Door has online price 2,130 and in-store price of 3.97
* There were subtle errors where some items did appear to be erroneous but differences and product type were unclear since amazon may charge higher prices for small items to compensate for shipping and handling

The data set contained over 3900 rows so manual inspection of each item was not feasible.

**Handling Outliers –**

Using a scattered plot, I could identify the p\_difference that was the threshold for extreme outliers.



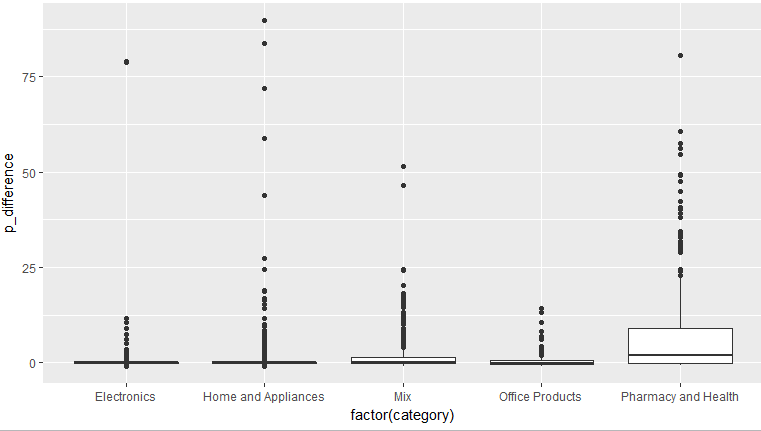
**Figure 1 – Scatter plot of p\_difference by category**

From the plot I found that after a p\_difference of >100, the price gap was too extreme to be a reasonable difference.

I used that difference to remove the obvious difference. To remove values, I created a vector containing the values that exceeded my threshold, then removed that vector from the data set.

I then attempted to used statistical tests to identify outliers, using the Outlier packaged, I applied the Grubbs test to identified the extreme outliers from each end of the distribution. This worked but needed multiple iterations as it only identified the one point. The other issue was that this test was designed for normal distribution data sets. This data was non-normal in its distribution.

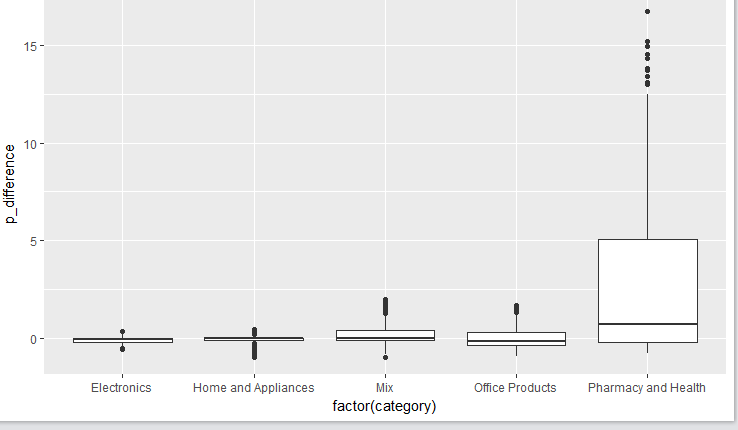
Since the data was non-normal, I applied the median 1.5 IQR method to determine thresholds for upper and lower point removal. Calculated each category’s upper and lower IQR, then examined the data at those thresholds to determine if the removal of points beyond the upper and lower thresholds were reasonable to remove.



**Figure 2 -Box Plot prior to removal of outliers**

Once I determined the reasonable thresholds, I removed the data points exceeding those thresholds on the upper and lower ends.

The new data set continues to show excessive spread but the difference seems more reasonable and more a function of actual price differences versus differences due to entry errors.

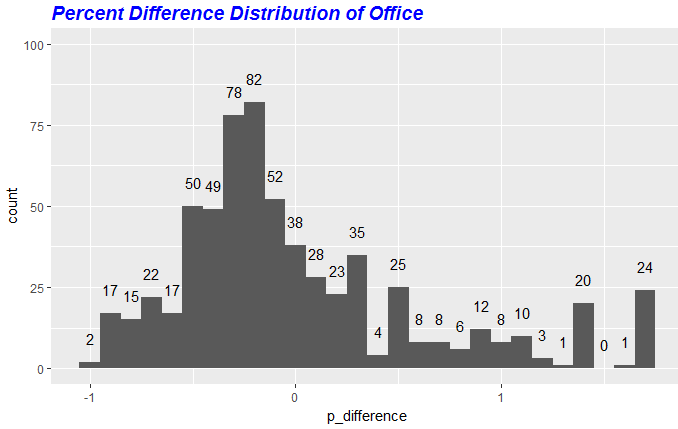


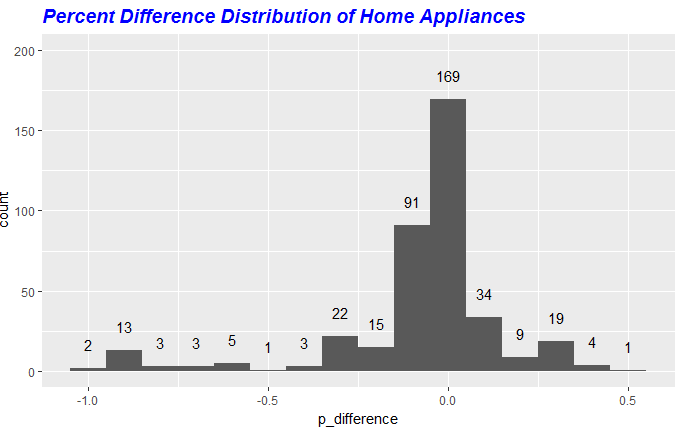
**Figure 3 – Box Plot after outlier removal**

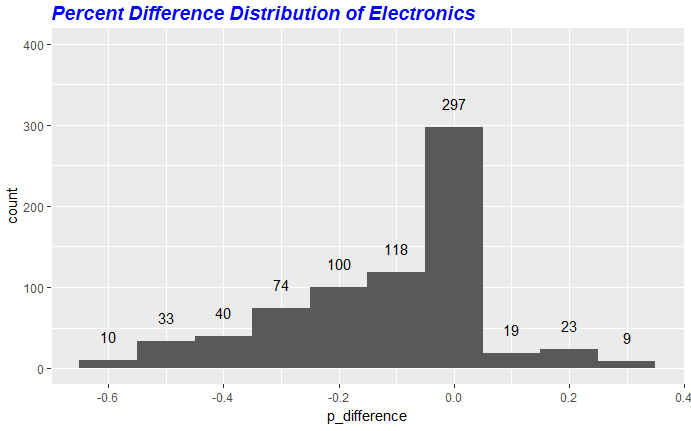
**Testing Hypothesis –**

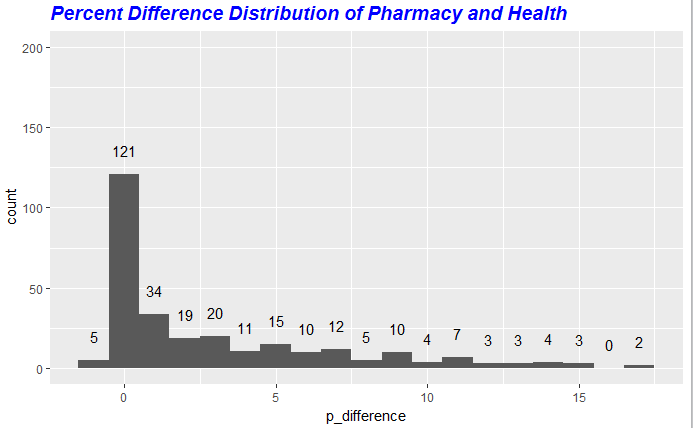
Since the data is non-normal, even after the removal of outliers, I cannot test my hypothesis or predict all price differences with my current sample set.

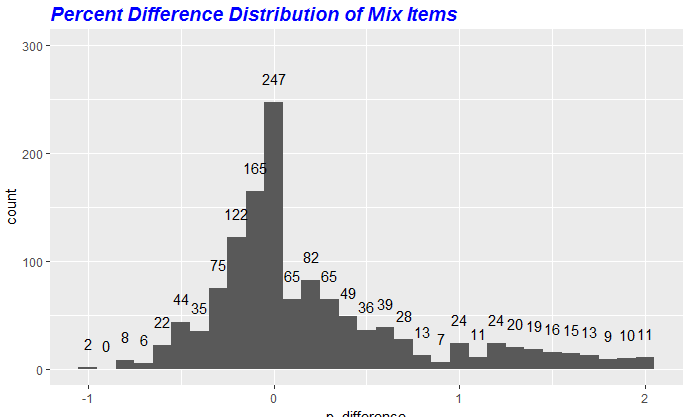
My first hypothesis is to determine if one product has a better price distribution or skew towards Amazon versus in Store. I can test this visually with a simple distribution plot











To determine the percentage of time it was better or neutral to purchase in store, I created new column assigning 0 if the price was better on-line and 1 if the price was better in store or same price.

The averages of the product categories showed –

mean(pharm\_health$bias)

[1] 0.625

> mean(electronics\_set$bias)

[1] 0.2365145

> mean(home\_app$bias)

[1] 0.5101523

> mean(mix$bias)

[1] 0.5304212

> mean(office$bias)

[1] 0.369906

> mean(pharm\_health$bias)

[1] 0.625

From the averages, you can see a price advantage if you purchase office and electronic products online.

**Determining if Sale Items Correlates to Better Price On-line or in store:**

My next analysis was to determine if the price of the item was on sale in the store and on-line, which mode of shopping would be best to purchase the item, if the items were regular priced, which mode would be better. If the price was regular priced in the store but on sale on-line, would the pricing still be better on-line.

To perform this analysis, I used the columns, “sale\_online” and “PRICETYPE”. The columns identified if the items were on sale or regular priced. I determine that there were four price conditions that could be met:

* Category 1 – Reg. Price Online, Reg. Price in Store
* Category 2 – Sale Online, Reg. Price in Store
* Category 3 – Reg. Online, Sale in Store
* Category 4 – Sale Online, Sale in Store

I researched multinomial logistic regression to determine if I could use the categories above to predict if the price was better or neutral in store = 1 or if the price was better on-line = 0

The article - <http://stats.idre.ucla.edu/r/dae/multinomial-logistic-regression/> provide a good library and analysis of this test but I could not determine if I could apply categorical conditions as my predictor values to determine my dependent value bias. I will continue to investigate.

Until I find a test to predict best price, I went forward with using straight analysis to determine what the data set provided in terms of best mode of purchase.

I created a new column in the data set, “cat\_price” that coded each value Cat 1 – 4, depending on the conditions above. There were blank values in the PRICETYPE category and I could not assume that the blank field was a sale price or regular, therefore, I coded the cat\_price value to “NA”. There were 577 missing values out of 3,325 observations so I was ok with excluding these values. Once I created the category field, I extracted the rows by each category type and calculated the mean of the bias variable to determine the percentage of times it was better to purchase on-line when the items were regular priced on sale.

The results are below:

mean(cat1$bias)

[1] 0.4309091

> mean(cat2$bias)

[1] 0.2614379

> mean(cat3$bias)

[1] 0.6818182

> mean(cat4$bias)

[1] 0.3818182

> View(cat2)

From the mean values I found that it was better to purchase the items in the store only when the price was on sale in the store. Every other condition, showed a slightly better condition to purchase on-line.

**Wrap-up and Value:**

This project was a good start to exploring open source data and using R to wrangle and explore unknown data sets. Prior to starting this project, I had no exposure to the vast amounts of interesting open source data sets that exist. I was challenged in trying to develop a project or test based on unknown large data sets. As a scientist, I am used to developing a hypothesis or a problem, then structuring my experiments and data collection based on testing my problem statement. It was difficult for me to look at large data sets then develop the hypothesis.

The data used in this project was not ideal for statistical predictions but was very good at helping me learn and practice the critical skill of data wrangling. This was the most important goal for me so I am happy that I was able to practice visualizing, exploring and cleaning data sets.

While am not confident that my statistical analysis of this set was worth making a broad suggestion or prediction to a potential consumer searching for the best mode of shopping, I do feel like the analysis did provide some general awareness which mode of shopping would provide better pricing. If I have time, I may go on a shopping trip and explore how my test data compares to some small sampling of store and online prices for the categories in this project.