Online v. Offline Prices:
Being a More Informed Market Participant

MIDS-INFO-W18-Section 1
Project 2 Final
Colby Carter, Chase Smiegiel, and Yisang Yoon

Topic

With the development of internet and distribution of various tools that access it, online sales have had a growing presence in retail sales. The attraction of online purchases attributes to various aspects that appeal to both sellers and consumers.

Perhaps the greatest appeal of the online markets is that they connect the world of sellers with the world of buyers without confinement of location that limit the traditional offline markets. This is especially appealing because, at least theoretically, the competition amongst a greater number of sellers should lower the price of a product.

In our study, we test the widespread notion that online markets indeed offer lower prices for consumers than do offline markets. Moreover, we identify which product categories account for the greatest differences and by what degree.

This information should be useful for both consumers and sellers in building a market strategy to attain the best price for various product classes. The information also contributes to a more efficient market as it helps identify price spreads in various products.

Focus

We have used historical prices from Amazon and online multi-channel retailers to represent online prices and prices at major retailers by zip-code (US) or country to represent offline prices. We have focused on data from 2015.

Before diving into the dataset, our group has asked the following questions to guide the direction of the study:

1. Generally, what is the price difference between online and offline markets? Which one is more inexpensive and by how much?
2. What are some of the most popular items sold? Does this differ for online vs. offline?
3. Are there certain price ranges where the price difference increases or decreases vs the average?
4. Which products are more expensive offline? How about online?
5. How do offline/online prices compare around the world?

Data

The primary datasets are sources from the Billion Prices Project (BPP) at Massachusetts Institute of Technology (MIT), which is a project that collected information on online prices since 2008. Our analysis combined multiple datasets, including a comparison between Amazon pricing with alternative online and offline retailers.

The dataset offers several useful variables besides the obvious comparison between online and offline prices. Noteworthy ones include country of transaction, type of product (e.g. electronic), and date the price was valid.

While the currency of the prices is not labeled, we have concluded from observation that the prices are in the reporting country's local currency. Therefore, we focused our analysis to degree of price differences rather than on unit of price differences.

Limitations

For our purposes, we have excluded the cost of shipping and handling, as well as taxes, for online prices. Likewise, we have excluded the greater length of time and effort (e.g. transportation cost) needed to buy a product at a physical retail store.

In addition, while there are countless number of product classes, we have focused on a select few that are publicly popular. We are also limited in the number of countries surveyed.

Also, we took caution regarding seasonality differences in countries. Most of these limitations derive from the dataset that we have worked with.

Data Validation

We have observed several issues worth addressing.

First, there were products that showed a difference in price of 10x or more, which did not seem reasonable. We uncovered that many of these unusually great differences were due to differences in quantity. While one market was selling a single item, another market was selling a package of several items. Yet there were others that showed a difference of two decimal points for reasons we could not uncover. These outliers were less than 0.1% of the dataset, and we have removed them from our study.

Second, we have worked with a dataset that had disproportionate regional representation. Of the ten countries observed, data from the United States represented the largest piece, accounting for roughly half of the observations. In addition, a significantly large amount (approximately 75%) of the information pertaining to the United States comes from Massachusetts. It is important to note that most of the towns in the United States are urban. In addition, information from China seemed insufficient as it accounted for the lowest observation count amongst the ten countries observed.

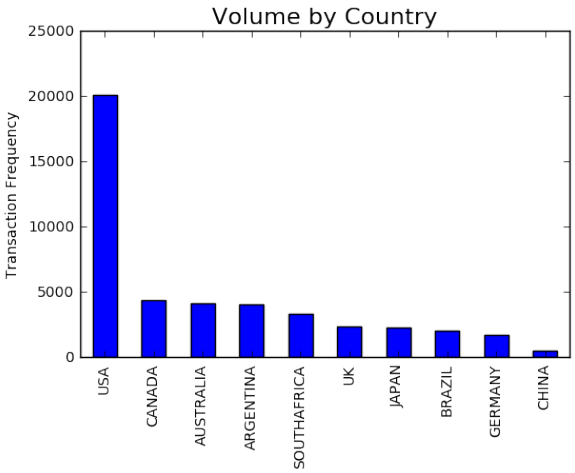


Figure 1 – Most of the data originated from the United States while information from China seemed insufficient.

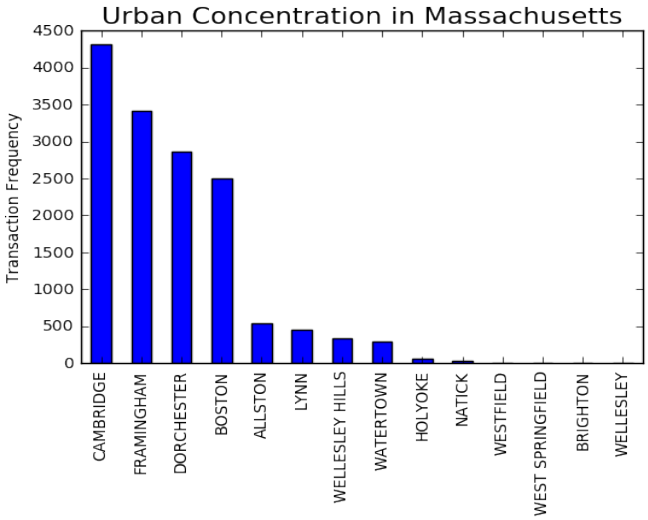


Figure 2 – Most of the information came from urban areas.

Our group had initially planned to test comparisons between prices originating in different regions such as developed and developing countries or urban and rural areas. However, we have decided that the dataset we were analyzing were not diverse enough to lead to meaningful conclusions.

Therefore, we have decided to exclude such regional comparisons. We have, however, proceeded with conducting analysis of each country, focusing on the United States which was abundant in data.

Third, we have added Amazon.com prices to the main dataset from MIT. Amazon is the largest online retailer in the world by total sales and market cap. By adding prices from Amazon, we were able to compare them to online prices in general in addition to the comparison to offline retailers.

However, general online price information was missing from the file that contained Amazon prices in roughly 30% of the dataset. Our observations have indicated that Amazon prices in the dataset were 10x higher than offline prices about 5% of the time, and 10x lower than offline prices about 1% of the time.

We have used the pandas package to execute the cleanup, and the matplotlib library to organize the findings. Many of our exhibits are shown in "box plots" using matplotlib, since we found it useful to show several useful information (max, min, mean) easily.

Analysis

United States

Price data originating from the United States makes up the largest component amongst data collected from 10 different countries. Our analysis on MIT's BPP

dataset has revealed that, by and large, there is no difference between online and offline retailers when it comes to *average* price. The *average* price difference for the entire US dataset was near zero ($0.44). This finding contradicts the popular notion that goods online can be found at a cheaper price than its offline counterparts.

Figure 3 – A clear positive correlation exists between online and offline prices.

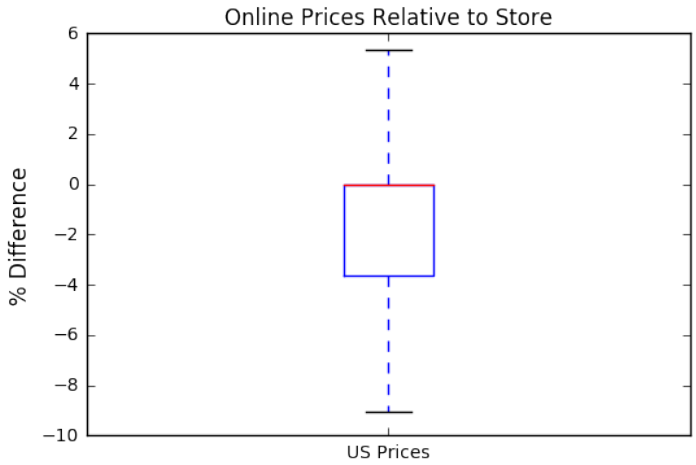


Figure 4 – Overall price difference between US online and offline retailers was near zero.

However, the opportunity to save by shopping online does appear with the highest percentiles. The highest prices (items priced in the 99th percentile) showed an online savings of $38.01. The other end of prices (items priced in the 1st percentile) on the other hand showed an *offline* savings of $35.24

Then, do the savings increase as price of items diverge away from the median? According to our data, that doesn't seem to be the

case. While there were close to 0 savings observed for the 50th percentile category, the savings did not necessarily show an increasing (or decreasing) savings in either directions. Only in the 95th percentile and above did we observe notable online savings. Likewise, offline savings were observed in 5th percentile and below.

	Store Price ($)	Online Price ($)	Difference ($)	Difference (%)
Mean	35.02	34.59	0.44	0.80
St. Deviation	66.33	66.73	13.15	34.24
Minimum	0.44	0.40	-449.02	-88.04
1%	1.39	1.29	-35.24	-53.30
5%	2.54	2.50	-6.00	-33.34
10%	3.48	3.29	-0.75	-20.01
25%	6.48	6.29	0.00	-3.62
40%	10.49	10.00	0.00	0.00
50%	14.99	14.99	0.00	0.00
60%	19.99	19.99	0.00	0.00
75%	36.94	34.99	0.30	0.00
90%	79.99	79.99	4.00	8.73
95%	129.99	128.00	11.98	33.30
99%	323.38	327.68	38.01	126.74
Maximum	1,199.99	999.99	300.00	898.40

Table 1 – From 20,133 observations, 95th percentile+ showed significant online savings.

Perhaps we would be able to further break down the analysis by dividing up the dataset into three subsets: (1) a subset of "low" prices which is comprised of offline retail products that are less than $10, (2) a subset of "moderate" prices which is comprised of offline retail products that are $10 and above up to $50, and finally (3) a subset of "high" prices which is comprised of offline retail products that are $50 and up.

Interestingly, the average prices of all three subsets showed very little differences between online and offline retailers.

Figure 5 – A comparison of online and offline prices for products that are under $10: price difference was not significant.

It can be observed, however, that the "moderate" and "high" price subsets included much lower prices when comparing lowest prices of the subsets. The lowest online price for the "moderate" subset was $2.49 compared to the $10.00 offline price. The lowest online price for the "high" subset was $11.99 compared to the $50.01 offline price. In contrast, the highest prices were similar for both online and offline (for "moderate", $50.00 offline vs. $63.72 online at the 99th percentile; for "high", $1,199.99 offline vs. $999.99 online).

Figure 6 – A comparison of online and offline prices for products that are between $10 and $50: Again, price difference was not significant.

Figure 7 – A comparison of online and offline prices for products that are over $50: Same story as the previous two.

For the "low" subset, the lowest price was comparable ($0.44 offline vs. $0.40 online), but showed a 50% difference in the highest price ($9.99 for offline vs. $14.99 for online at the 99th percentile).

We can infer from these findings that consumers were more likely to find a cheaper online price with the expensive items, though overall the vast majority of products will be similarly priced.

Amazon.com

The second dataset we looked at was one that compared offline retail prices to two online prices: Amazon.com and online ex-Amazon. For this exercise, we divided up the dataset to five different product classes (electronics, home and appliances, office

products, pharmacy and health, and others labeled as “mix”) as labeled to see if there were any product class that showed noteworthy online or offline savings, and how Amazon prices compare.

With electronics, we did observe slight savings on Amazon. Interestingly, for office products, general online prices were *higher* than offline retailers while Amazon prices were lower. For the other three categories (Home and Appliance, Pharmacy and Health, and “Mix”), Amazon prices were higher than both general online prices and offline prices, which were about the same.

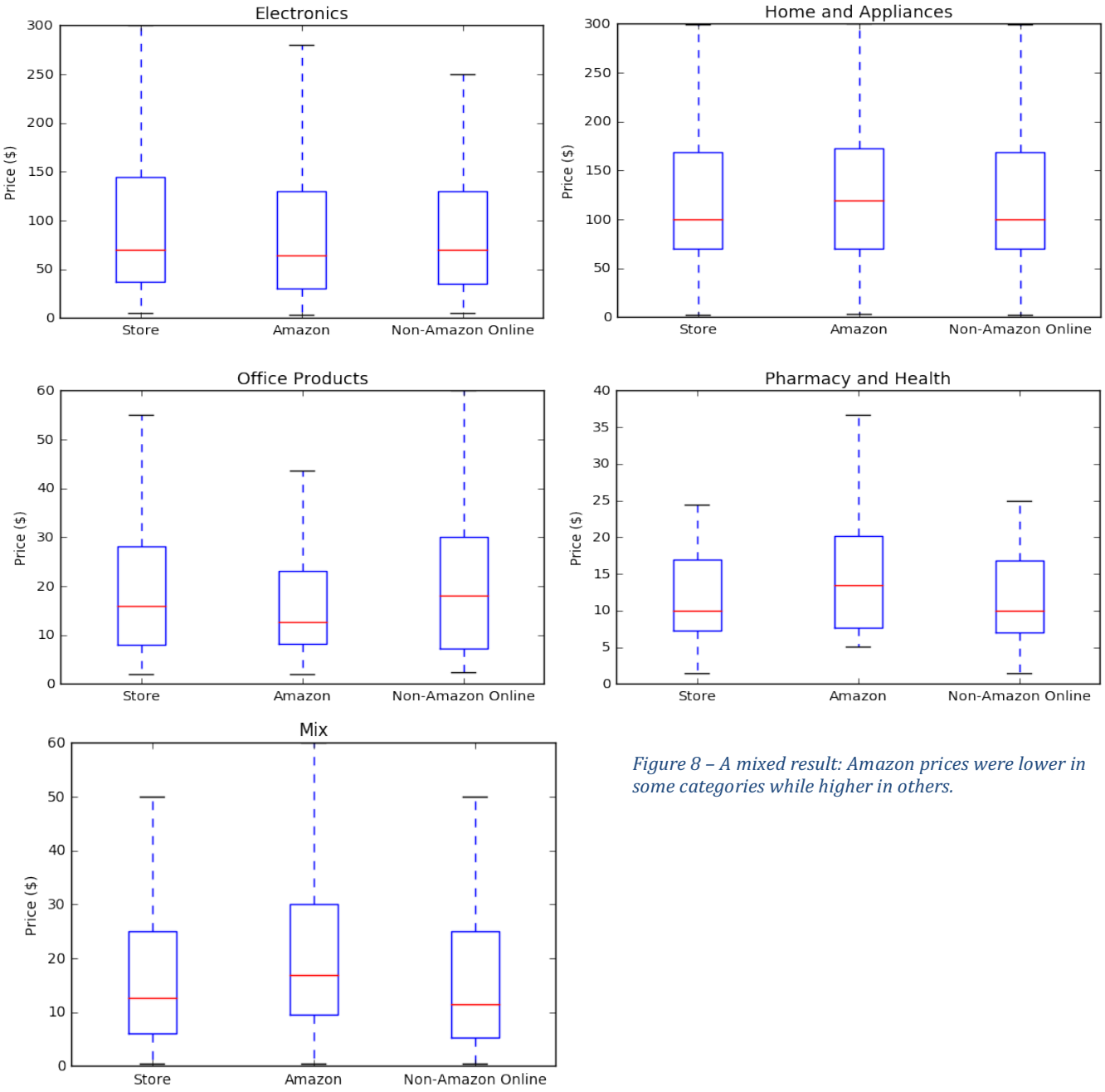
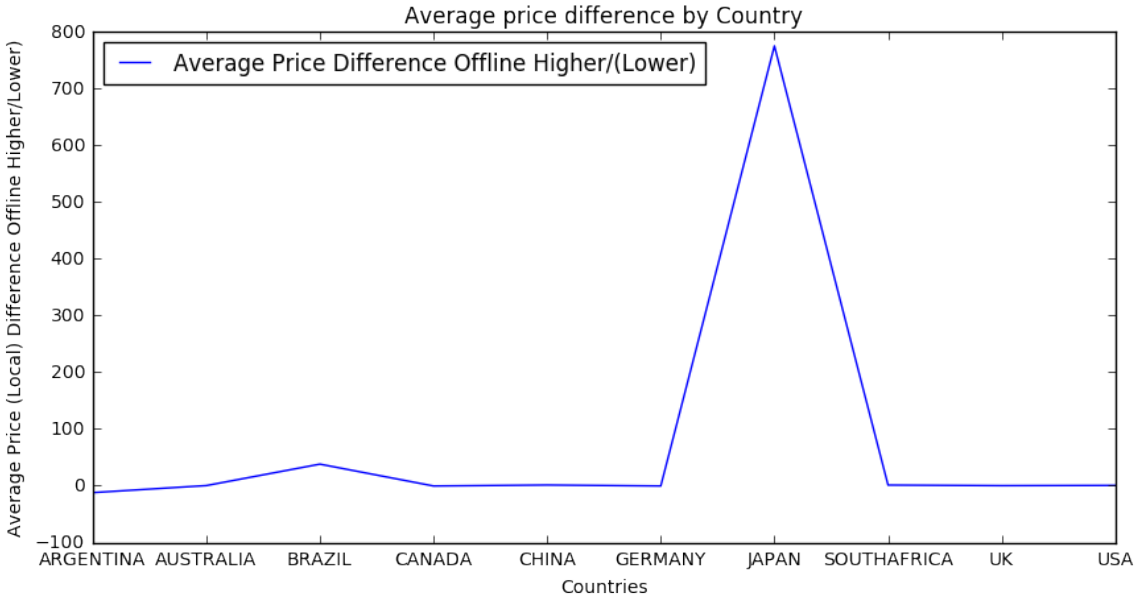


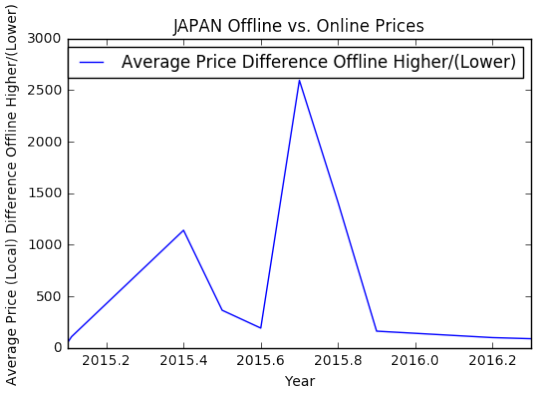
Figure 8 – A mixed result: Amazon prices were lower in some categories while higher in others.

Other Countries

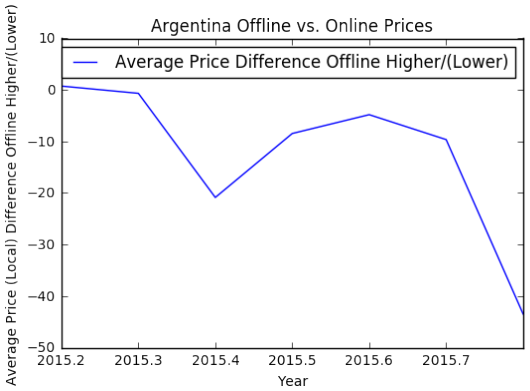
Our dataset was primarily comprised of price comparisons in the United States, but it did include some price point data from other countries including: Argentina, Australia, Brazil, Canada, China, Germany, Japan, South Africa and the United Kingdom. The data wasn't consistent in terms of currency (local), time frames, or products we set up a script to analyze the available country price data that was available.

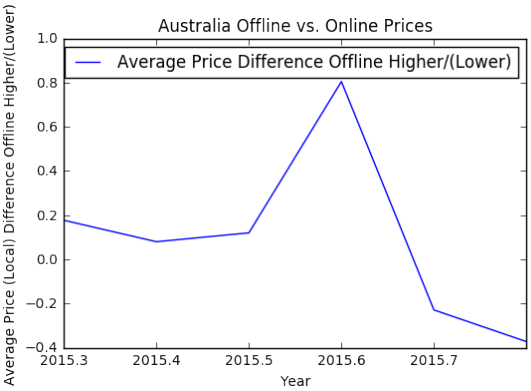


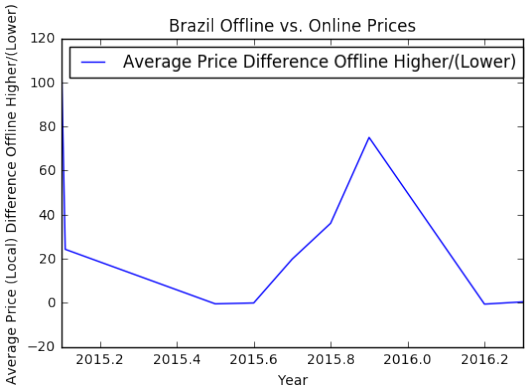
While the end result was very noisy and didn't really allow us to draw any concrete conclusions about the countries themselves, or comparisons between the countries, we were able to set up a framework that looked at the average price difference within each country over-time (separate plots) and each countries average price difference (on one plot).

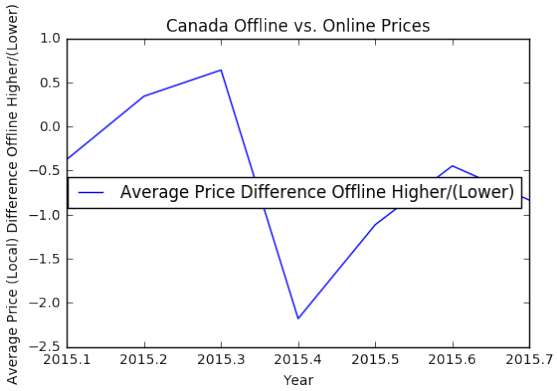


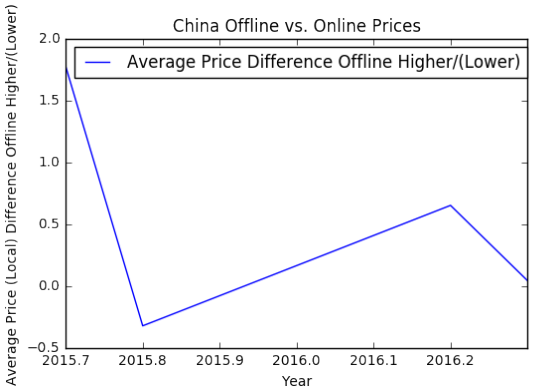
Had there been more data available it would have been really interesting to see any sort of seasonality associated with price differences or if any countries or types of countries had cheaper online or offline prices. If a more extensive dataset had been available we could have focused more on this topic and weighted the price differences by the prices of the products available, converted all prices into one common currency and/or segmented the price difference by product category.

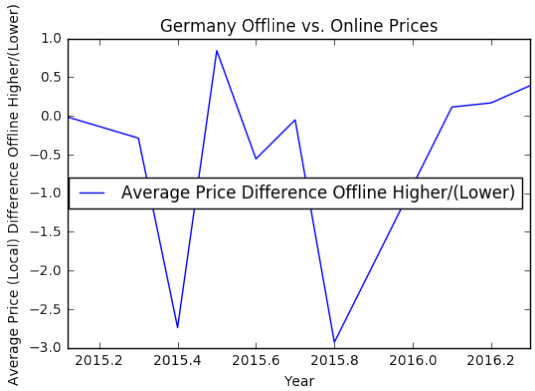

Argentina Offline vs. Online Prices
Average Price Difference Offline Higher/(Lower)
Average Price (Local) Difference Offline Higher/(Lower)
Year

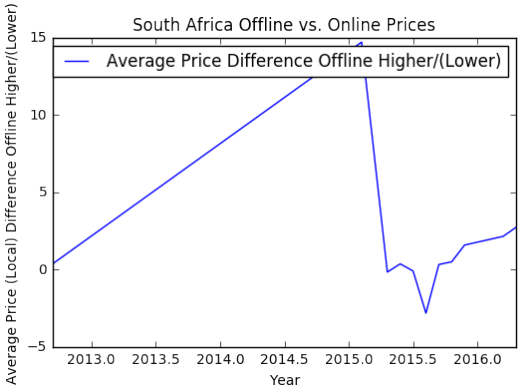

Australia Offline vs. Online Prices
Average Price Difference Offline Higher/(Lower)
Average Price (Local) Difference Offline Higher/(Lower)
Year

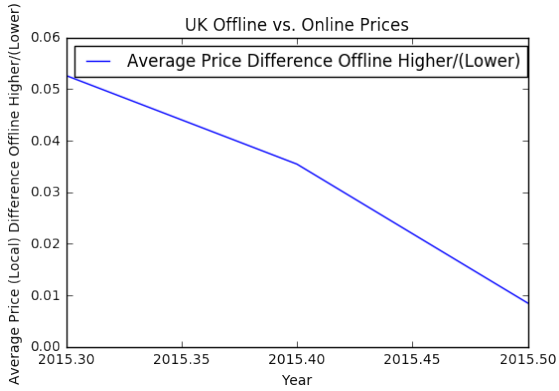

Brazil Offline vs. Online Prices
Average Price Difference Offline Higher/(Lower)
Average Price (Local) Difference Offline Higher/(Lower)
Year


Canada Offline vs. Online Prices
Average Price Difference Offline Higher/(Lower)
Average Price (Local) Difference Offline Higher/(Lower)
Year


China Offline vs. Online Prices
Average Price Difference Offline Higher/(Lower)
Average Price (Local) Difference Offline Higher/(Lower)
Year


Germany Offline vs. Online Prices
Average Price Difference Offline Higher/(Lower)
Average Price (Local) Difference Offline Higher/(Lower)
Year


South Africa Offline vs. Online Prices
Average Price Difference Offline Higher/(Lower)
Average Price (Local) Difference Offline Higher/(Lower)
Year


UK Offline vs. Online Prices
Average Price Difference Offline Higher/(Lower)
Average Price (Local) Difference Offline Higher/(Lower)
Year

Conclusion

Findings

Our analysis showed that, by and large, there were not noteworthy differences between online and offline prices. We did identify significant online savings in expensive items. There were various subsets that showed different results. But on average, the prices did not differ much.

While the result may be somewhat uninteresting, to disprove a popular notion that online prices will almost always be lower than retail prices is an important conclusion.

Challenges

The biggest difficulty we had was with the original dataset we began with. We were constantly questioning the validity of the data and were looking for outliers that would throw off the results. In addition, we had to cut down on some of the questions we wanted to analyze due to the limitations in the scope of the dataset.

A more abundant and diverse dataset that is more complete and more uniform would have been much easier to work with. This experience gave us a taste of what it would be like working with data in the real world and how frustrating it can be.