CS838 DATA SCIENCE PROJECT

Data Analysis

Tuan Dinh
dinh5@wisc.edu

Sam Gelman
sgelman2@wisc.edu

Varun Sah
varun.sah@wisc.edu

May 7, 2017

CONTENTS

1 DATA

The dataset we worked with was created by cleaning, matching and merging data instances crawled from the websites of Amazon and Newegg. The dataset consists of electronics products belonging to several categories and their associated attributes like price, brand and number of reviews. The following subsections present some more details regarding the aforementioned dataset which was obtained as the result of the previous project stage (Data Merging) and serves as the starting point of the Data Analysis stage.

1.1 SCHEMA

The schema for the combined table includes the following attributes:

- ASIN
- NID
- RID
- NAME
- PRICE
- CATEGORY
- BRAND
- AMAZON_INFO
- NEWEGG_INFO
- AMAZON_NUM_REVIEWS
- NEWEGG_NUM_REVIEWS

1.2 SIZE

There are 971 tuples in the combined table.

ASIN	*NID*	*RID*	*NAME*	*PRICE*	*CATEGORY*	*BRAND*	*AMAZON_INFO*	*NEWEGG INFO*	*AMAZON NUM REVIEWS*	*NEWEGG NUM REVIEWS*
B00005ARK3	9SIA0AJ0NA7913	33-124-002	linksys befw11s4 wireless router ieee 802.3/3u ieee 802.11b | cisco-linksys wireless-b cable/dsl	49.98	Wireless Routers	linksys	8.2 x 7.2 x 2.1 inches 2.5 pounds 2.5 pounds b00005ark3 befw11s4 3.4 out of 5 stars may 2 2006 none none	befw11s4	1108	69
B00080DSEM	9SIA1N82853897	26-105-166	microsoft compact optical mouse 500 - black	11.59	Mice	microsoft	3.4 x 1.9 x 1 inches 4 ounces 4 ounces b00080dsem u81-00009 4.1 out of 5 stars june 6 2005 none none	u81-00009	83	26

ASIN	*NID*	*RID*	*NAME*	*PRICE*	*CATE-GORY*	*BRAND*	*AMAZON_INFO*	*NEWEGG INFO*	*AMA-ZON NUM REVI-EWS*	*NEW-EGG NUM REV-IEWS*
B004RI5EHA	9SIA1UH50Y2009	17-182-023	rosewill rv350-2 350w atx 12v v2.2 power supply | 350-watt atx12v	27.27	Power Sup-plies	rosewill	b004ri5eha 3.8 out of 5 stars 7.3 pounds item can be shipped within u.s. this item is not eligible for international shipping. december 17 2010 rosewill rv350-2 3.5 pounds 5.5 x 5.9 x 3.4 inches 5.51 x 5.91 x 3.35 inches silver 220 volts none	rv350-2	74	142
B00603QXPM	9SIAAEE5B48720	19-116-492	intel core i7-3930k sandy bridge-e 6-core 3.2ghz 3.8ghz turbo lga 2011 130w bx80619i73930k desktop processor | hexa-core 3.2 ghz 12 mb cache -	507.82	CPUs	intel	4.9 x 2 x 1.8 inches 3.2 ounces 4.8 ounces item can be shipped within u.s. this item is not eligible for international shipping. b00603qxpm bx80619i73930k 4.5 out of 5 stars september 10 2011 none none	bx80619-i73930k	135	356

1.4 Enrichment

In order to better analyze customer satisfaction metrics, we enriched the merged dataset (combined_products.csv) with additional fields representing customer response to products. We added average customer ratings as well as the distribution of these ratings obtained from Amazon raw data (jtotal.json) for each of the 971 products. This resulted in an enriched version of the combined dataset with additional attributes pertaining to customer ratings included (products.csv).

1.4.1 Enriched Schema

The enriched schema for the includes the following attributes:

- ASIN
- NID
- RID
- NAME
- PRICE
- CATEGORY
- BRAND
- AMAZON_INFO
- NEWEGG_INFO
- AMAZON_NUM_REVIEWS
- NEWEGG_NUM_REVIEWS
- AMAZON_RATING_1
- AMAZON_RATING_2
- AMAZON_RATING_3
- AMAZON_RATING_4
- AMAZON_RATING_5
- AMAZON_RATING_AVG

1.4.2 Enriched example tuples

ASIN	*NID*	*RID*	*NAME*	*PRICE*	*CAT-EGO-RY*	*BR-AND*	*AMA-ZON-_IN-FO*	*NEW-EGG-_IN-FO*	*AMA-ZON NUM REV-IEWS*	*NEW-EGG NUM REV-IEWS*	*AMA-ZON-_RAT-ING-_1*	*AMA-ZON-_RAT-ING-_2*	*AMA-ZON-_RAT-ING-_3*	*AMA-ZON-_RAT-ING-_4*	*AMA-ZON-_RAT-ING-_5*	*AMA-ZON-_RAT-ING-_AVG*
B0000-5ARK3	9SIA-0AJ0N-A7913	33-124-002	linksys befw-11s4 wire-less router ...	49.98	Wire-less Rout-ers	link-sys	8.2 x 7.2 x 2.1 inch-es 2.5 pou-nds ...	befw-11s4	1108	69	20	11	8	20	41	3.51
B0000-5LLY4	9SIA-5AD4Y-Z4354	22-136-292	re-fur-bished west-ern digi-tal blue wd80-0bb 80gb ...	37.74	In-ter-nal Hard-drives	west-ern dig-ital	5.8 x 4 x 1 inch-es 1.3 pou-nds ...	wd80-0bb	159	97	16	8	5	19	52	3.83

ASIN	*NID*	*RID*	*NAME*	*PRICE*	*CATEGORY*	*BRAND*	*AMAZON_INFO*	*NEWEGG_INFO*	*AMAZON NUM REVIEWS*	*NEWEGG NUM REVIEWS*	*AMAZON_RATING_1*	*AMAZON_RATING_2*	*AMAZON_RATING_3*	*AMAZON_RATING_4*	*AMAZON_RATING_5*	*AMAZON_RATING_AVG*
B0000-5LLY4	9SIA-AEE3X-R1053	22-144-102	western digital caviar wd80-0bb 80gb ...	32.25	Internal Hard-drives	western digital	5.8 x 4 x 1 inches 1.3 pounds ...	wd800bb	159	506	16	8	5	19	52	3.83
B0000-5QBUU	9SIA-0AJ0Z-B3820	26-159-210	sony mdr-e818lp font-opia ear-bud head-phones ...	49.85	Head-phones	sony	3.9 x 10 x 3.9 inches 4 ounces ...	mdr-e818lp	284	16	10	8	8	23	51	3.97

2 Data Analysis and Results

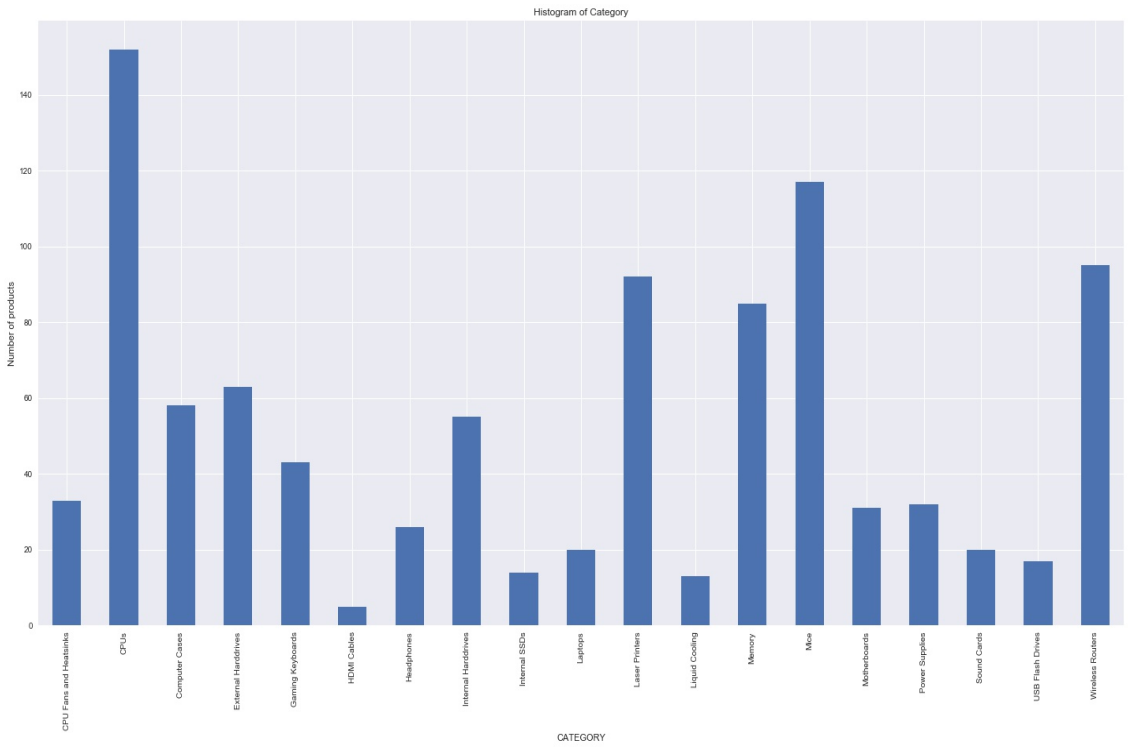
We started the data analysis with simple data exploration with respect to a single attribute to get a sense of what the distribution looked like. We then inspected the data with respect to a pair of attributes to gauge if there was any obvious correlation or association. The pair-wise analysis also served the purpose of assisting us in locating anomalies. The ultimate goal was to discover patterns and similarities between groups of products that elicit similar response from customers, which was achieved by means of clustering on the basis of customer ratings.

2.1 Data Exploration - OLAP

In order to get a deeper understanding of the nature of the dataset, we looked at the frequency distribution of products with respect to category, brand and price.

2.1.1 Category wise product distribution

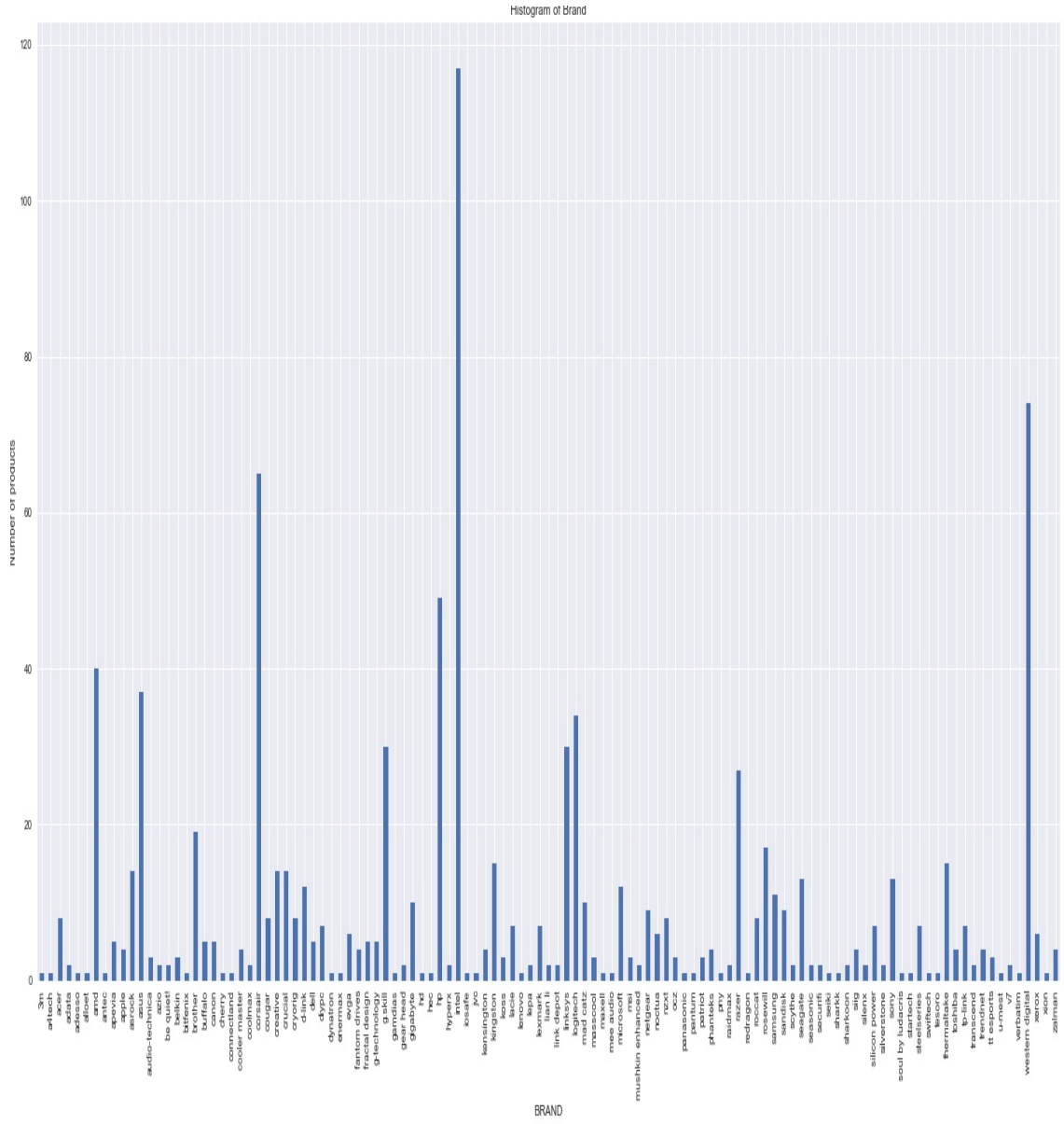
The histogram below illustrates the product frequency distribution with respect to categories. There were 19 distinct categories under which products were catalogued in the dataset.



"CPUs" was found to be the most frequent category followed by "Mice" and "Wireless Routers". The least common category was "HDMI Cables" with "Liquid Cooling" and "Internal SSDs" being other categories with less number of products.

2.1.2 BRAND WISE PRODUCT DISTRIBUTION

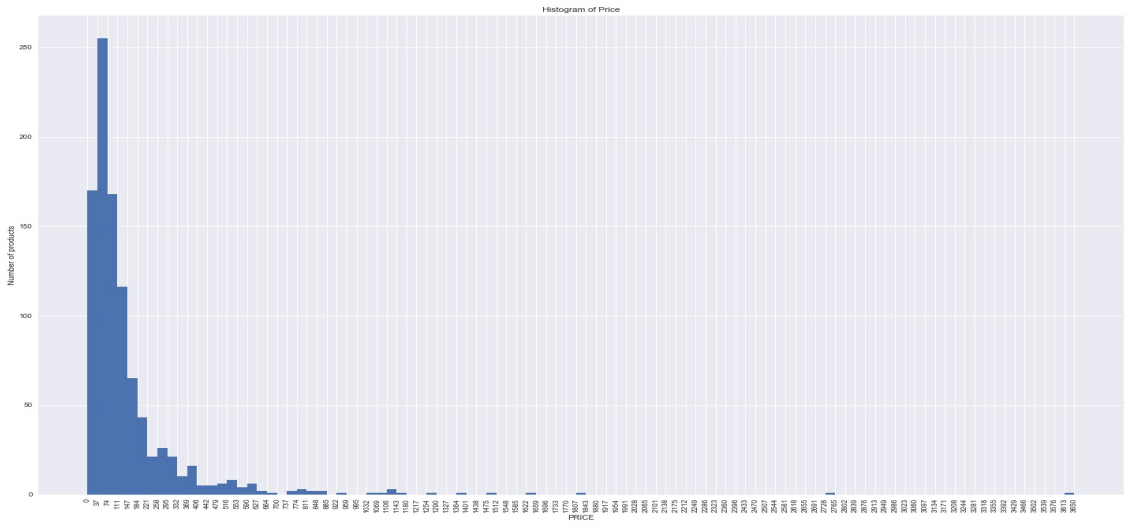
The histogram below illustrates the product frequency distribution with respect to brands. There were 114 distinct brands to which the products in the dataset belong.



"Intel" was the brand with the most number of products followed by "Western Digital" and "Corsair".

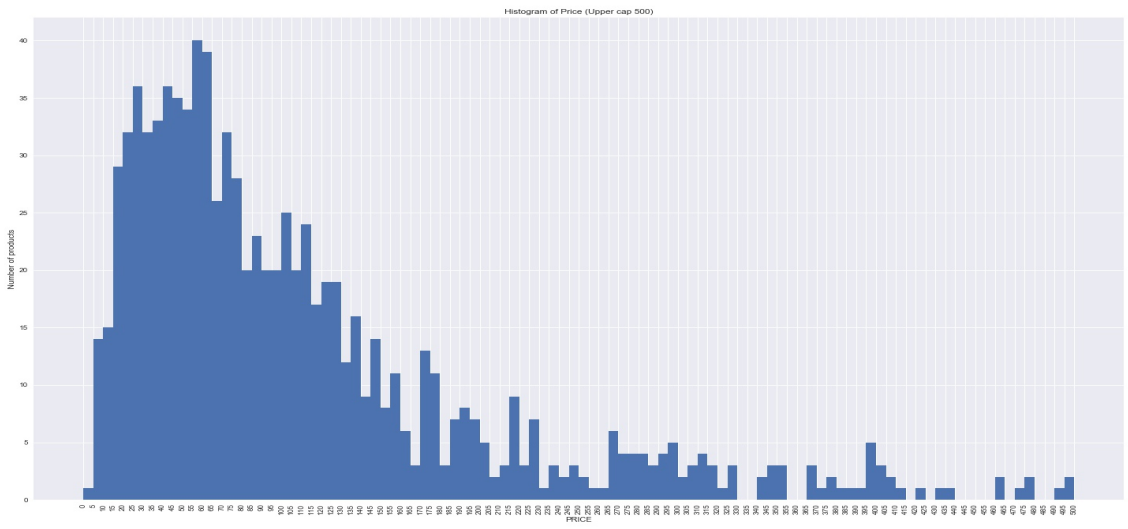
2.1.3 PRICE WISE PRODUCT DISTRIBUTION

The histogram below illustrates the product frequency distribution with respect to price.



The above graph shows a great skew in terms of prices. The dataset is dominated by cheap or moderately expensive products. Expensive products are few in number and make it difficult to observe the distribution of the majority of the data.

Below we slice on price to select only products with a price less than $500

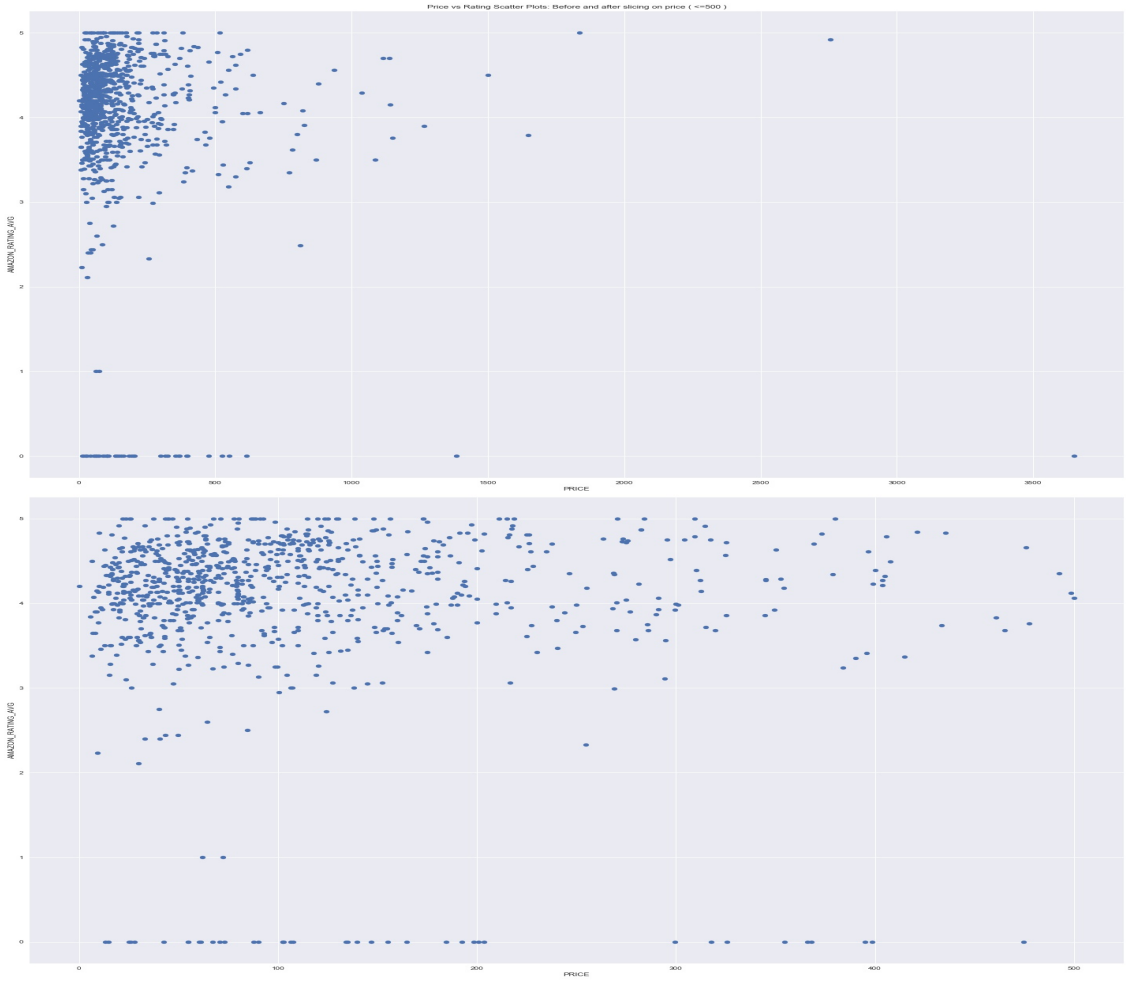


2.2 Graphical Attribute Correlation and Anomaly Exploration

We looked at scatter plots of pairs of attributes to determine if there was any obvious, observable relationship between them.

2.2.1 Price and Average Rating

The following scatter plots were drawn to observe the relationship between a product's price and its average rating.

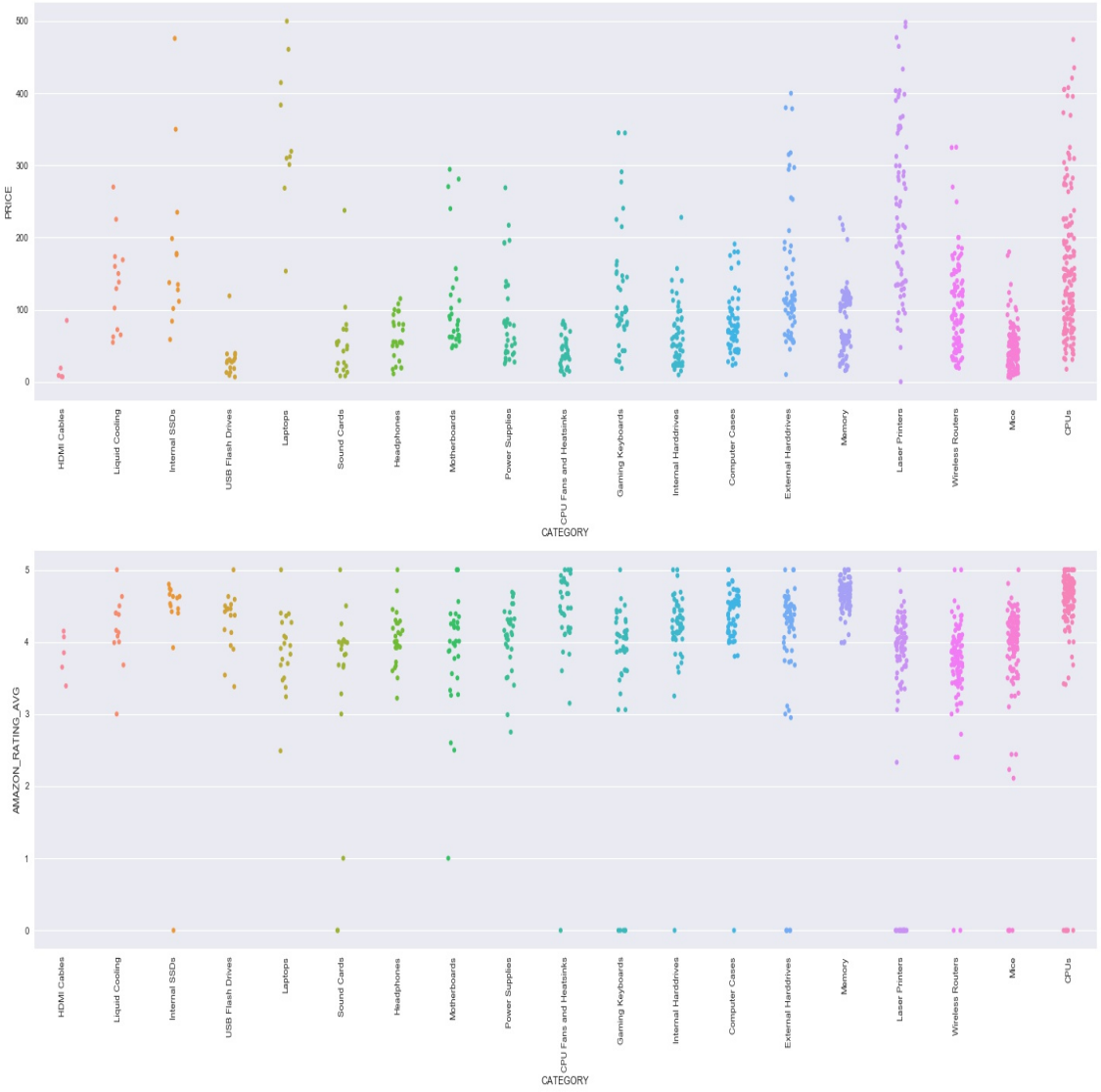


The first plot helped us locate outliers that were significantly more expensive than other products. Based on the frequency distribution of section 2.1.3, we decided that the products more expensive than $500 are extremely few in number and do not gainfully contribute to the distribution as a whole. Their presence, however, makes the already difficult task of vizualization even more formidable. To get a better view, we sliced the data to remove products costing more than $500 and created the second scatter plot.

2.2.2 Category wise Price, Rating

We used strip plots to observe the density distribution of products belonging to each category against the price of the product.

A similar strip plot was drawn to observe the density distribution of products belonging to each category against the rating of the product.

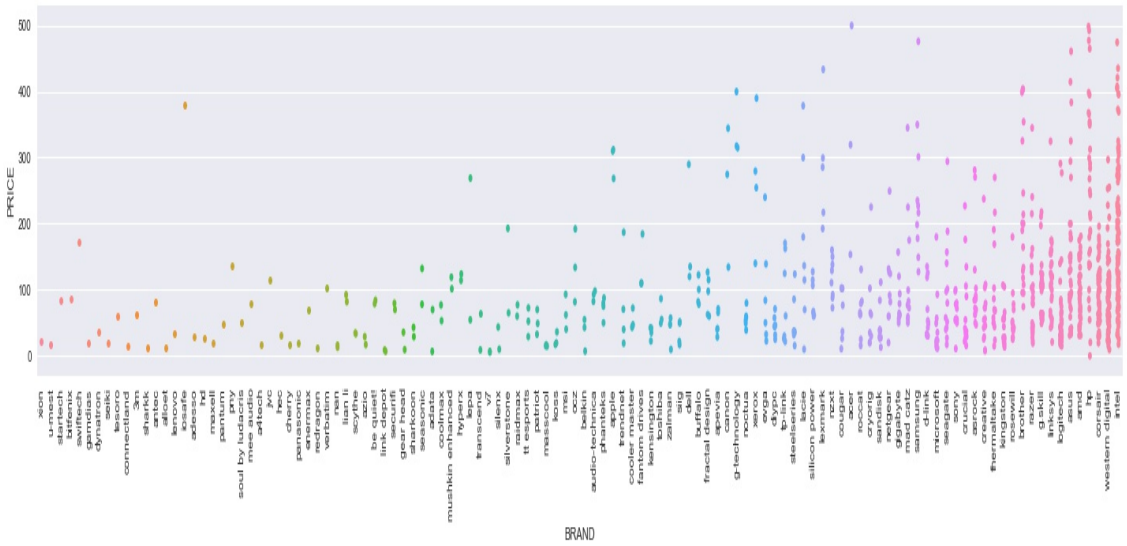


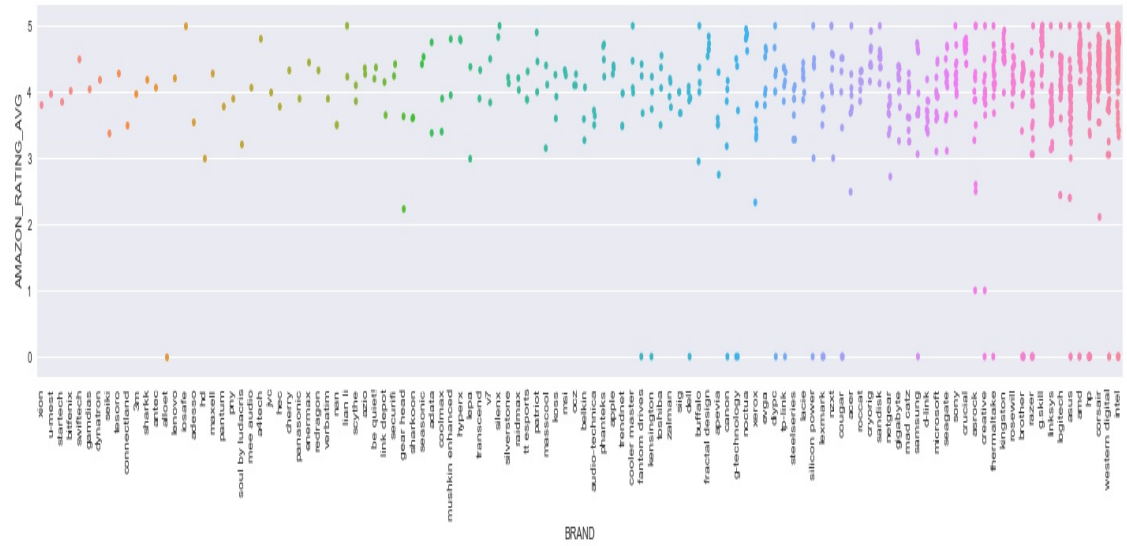
It can be seen from the above plots that "Laptops" is the most expensive category on average. On the other end of the spectrum, "HDMI Cables", "USB Flash Drives" and "Mice" are most affordable product categories.

2.2.3 Brand wise Price, Rating

We used strip plots to observe the density distribution of products belonging to each brand against the price of the product.

A similar strip plot was drawn to observe the density distribution of products belonging to each brand against the rating of the product.



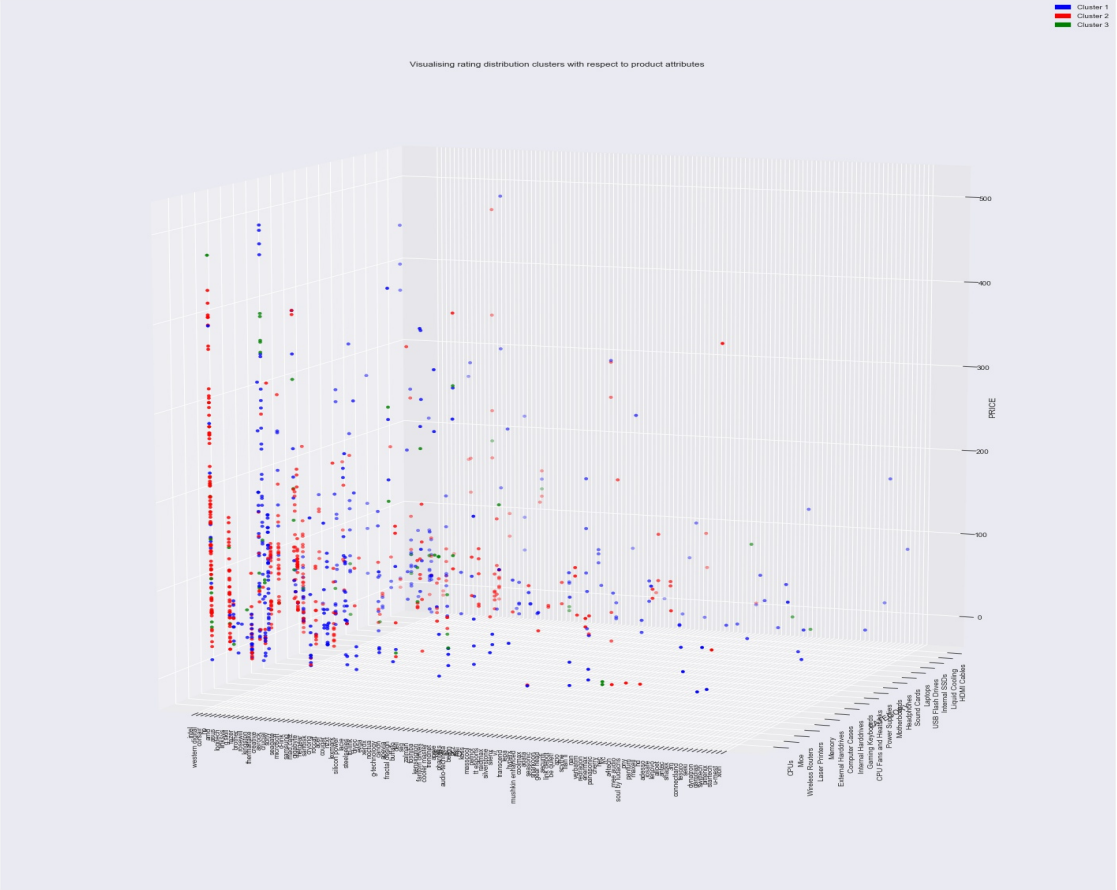


2.3 Clustering and Analysis of Clusters

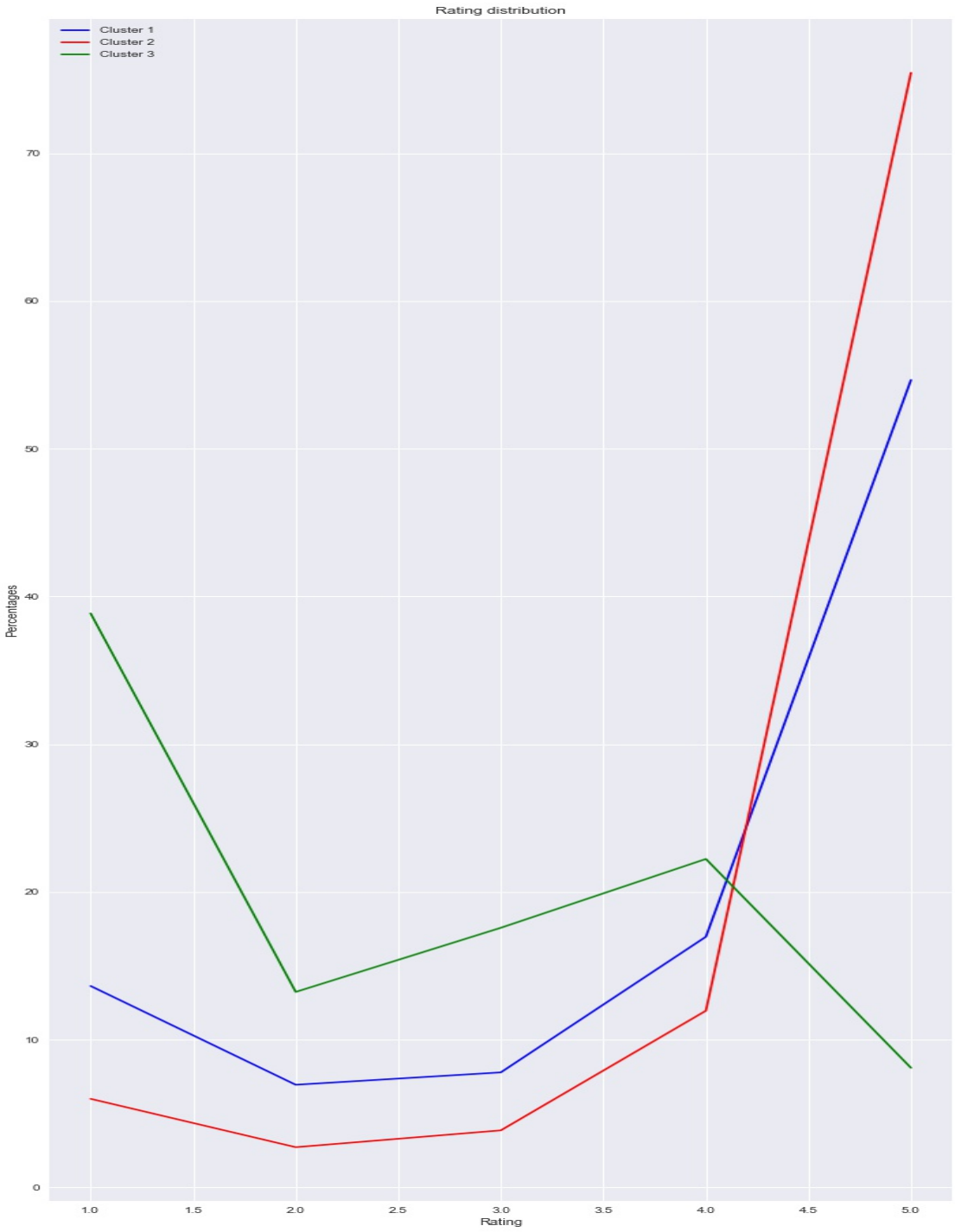
We then moved to the task of clustering products based on the distribution of customer ratings in order to discover patterns and similarities between groups with similar rating distributions.

We used both Agglomerative(heirarchical) and KMeans clustering as part of the analysis. We only present the plots pertaining to KMeans clustering due to the similar nature of the results.

Here we present a visualization of the clusters using brand, category and price as product attributes (attributes that were not involved in the clustering to begin with).



To better understand and visualize the set of various rating distributions characteristics existing in the data, we created a canonical representative of each cluster. Interestingly, the three canonical distributions appeared to be representative of three product segments.



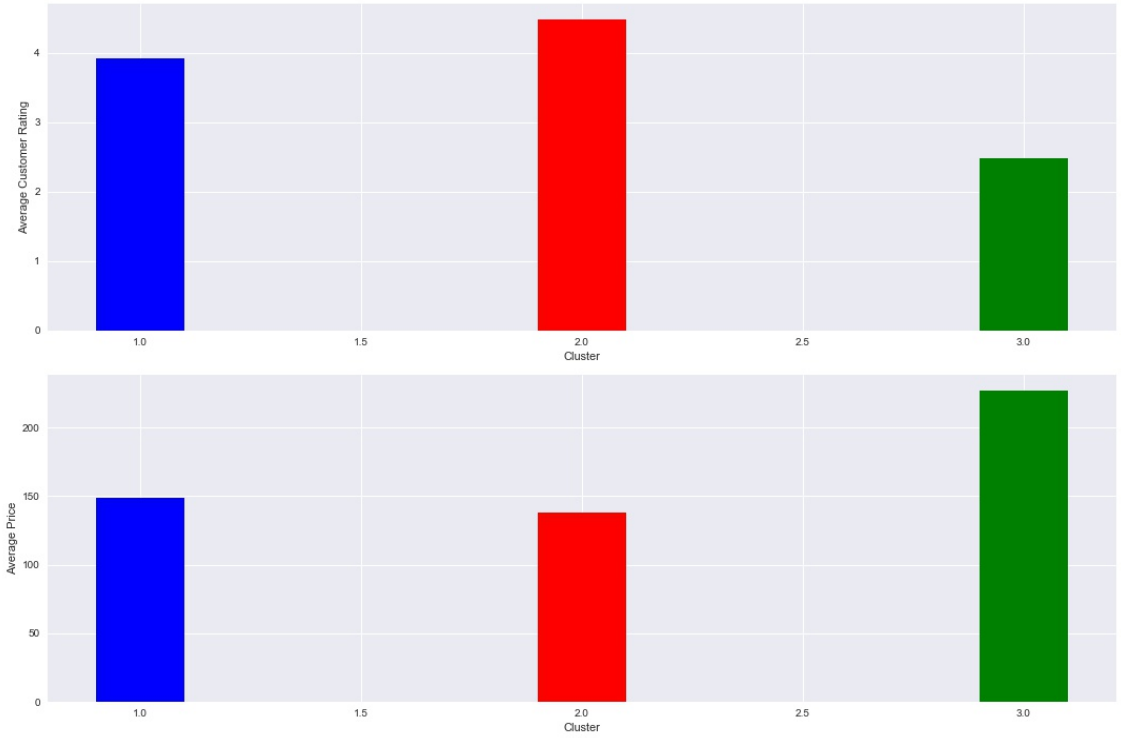
We interpreted the above graph as follows:

- The first cluster seemed to represent moderately rated products with a seemingly bal-

anced rating distribution. Although a majority of the ratings were high (>50 %), it is much more balanced than the second cluster. We can attribute this slight positive trend in the rating distribution to the inherent bias in all ecommerce data where even flawed products have good ratings by virtue of paid or fake reviews.

- The second cluster seemed to represent extremely highly rated products which possess a heavily positively-skewed rating distribution. An overwhelming majority (>75%) of the ratings of this product segment were 5-star ratings.
- The third cluster seemed to represent unfavorably rated products which have an overall negative trend in their rating distribution. A majority of the reviews on such products had ratings 1 and 2 (>50%).

We continued further analysis of each cluster. For each cluster, we computed the average customer rating as well as the average price.



The above plots allude to an inverse relationship between customer rating and price. The most highly rated segment of products was also the least expensive. The least favorably rated segment of products contained the most expensive products. And the third segment was intermediate both in terms of ratings and price.

3 HYPOTHESIS TESTING (ACCURACY METRICS)

To determine if there is infact an association between product price and rating, we performed statistical tests of association.

The null hypothesis for these tests was that there is no association between a product being expensive and having a poor rating. As per statistical conventions, in order to be able to reject the null hypothesis, we need a p-value of less than 0.05.

We considered a product to be poorly rated if its average rating was less than 2.5. We considered a product to be expensive if it costs more than $150. The thresholds of 2.5 for rating and $150 for price were NOT chosen arbitrarily. They were observed as the boundary values while inspecting the clusters from the previous section. We ran both Pearson's Chi-Squared Test of Association and Fisher's Exact Test. The results were as follows:

Actual Contingency Table

		Price	
		<=150	*>150*
Rating	*<=2.5*	35	23
	>2.5	685	228

Expected Contingency Table

		Price	
		<=150	*>150*
Rating	*<=2.5*	43.00721	14.99279
	>2.5	676.99279	236.00721

Residuals

		Price	
		<=150	*>150*
Rating	*<=2.5*	-1.2209856	2.0679495
	>2.5	0.3077437	-0.5212170

Chi-Squared Test:

- $X - squared = 5.3915$
- $p - value = 0.02024$

Fisher's Exact Test:

- $p - value = 0.01939$
- 95 percent confidence interval: 0.2843279 0.9188525

The results of both Pearson's Chi-Squared Test and Fisher's Exact Test were in agreement. The p-values (0.02024 and 0.01939) were both less than 0.05 which enabled us to reject the null hypothesis that there was no association between the price and rating. Hence, our intuition of the inverse relationship between price and rating was confirmed and found to be statistically significant.

4 Conclusion and Problems

During the course of the analysis, we faced certain difficulties in observing patterns due to the inherent bias and skew in e-commerce data where a majority of the products have high ratings (some of which may be due to fake or paid reviews while some can be attributed to customers' bias after reading other positive reviews). Due to the skew, it was not possible to directly observe any correlation between customer ratings and average price.

We were able to make interesting observations (statistically significant p-value < 0.05) based on the clustering procedure we described in earlier in the report:

- Extremely highly rated products (with a heavily positively-skewed rating distribution) are, on average, also the cheapest products (lowest average price). They have extremely high average ratings (greater than 4.3)
- Moderately rated products (with a balanced rating distribution) are moderate in the price department as well. They have intermiedaite average ratings.
- The least favorably rated products (with a negative trend in rating distribution), are, on average, the most expensive products. They have subpar average ratings (lesser than 2.5).
- The data illustrates that the 2-star rating is the least used rating by users irrespective of the ratings of product.

5 Future Prospects

We feel that the proces of adhering to the course timeline as well as the specified requirements for each project stage, restricted the scope of analysis to a certain extent. In other words, given more time, we would have tried to utilize the data that we had crawled to its full potential. With the large amount of data that we had collected for Amazon (17518 products and the millions of reviews associated with those products) before the matching and merging with Newegg data, we could have explored problems like:

- Classification of customer reviews as fake or authentic based on the characteristics of the review.
- Classification of customer reviews as helpful or not based on the characteristics of the review.

The aforementioned classification tasks, although extremely interesting, are particularly difficult are rather time intensive as they would involve advanced natural language processing techniques for feature engineering itself.