Trends and Analysis of eCommerce Data

Francesca Barrios
CSPB 4502
University of Colorado Boulder
frba7936@colorado.edu

Laura Brown
CSPB 4502
University of Colorado Boulder
labr1398@colorado.edu

Seth Ely
CSPB 4502
University of Colorado Boulder
seel6470@colorado.edu

Darby Hansen
CSPB 4502
University of Colorado Boulder
daha2377@colorado.edu

Abstract

In the dynamic landscape of online eCommerce retail, understanding consumer behavior and market trends is pivotal for success. Our group's objective is to leverage insights obtained from a dataset sourced from a multi-category online store, covering the period from October 2019 to April 2020. We aim to uncover valuable trends and patterns across various brands and categories, reflecting real-world marketing scenarios.

Here are some key questions our group seeks to answer with this analysis:

Which brands, or item categories exhibit the highest volume of sales during the specified period?

Among the features of brands and item categories, which ones demonstrate the highest purchases per view ratio?

Can seasonal patterns be identified within the 7-month window of the dataset?

Which products are similar, and can we generate a suggested list of items for users based on this data?

Answering these questions can help us to understand consumer needs and behavior. For example, uncovering which brands or items consistently lead in sales can shed light on consumer preferences and brand loyalty within the online retail space.

Identifying brands or items with high purchases per view ratios can reveal effective marketing strategies or product attributes that drive purchases, providing actionable insights for optimizing sales and marketing.

Discovering seasonal trends can unveil fluctuations in consumer demand influenced by factors like holidays, weather changes, or cultural events, enabling retailers to plan strategies and offerings accordingly.

By looking into these factors, we were able to discover which types of products users are most likely to purchase from the applicable eCommerce vendor with the hopes of targeting commonly purchased product categories and brands for further market analysis to drive revenue growth.

Introduction

The first question we are seeking to answer is to determine which brands and item categories exhibit the highest revenue sales during the timeframe we have access to through our dataset. This is an extremely important question to answer, as it will demonstrate user purchase preferences and help the eCommerce vendor to better understand their target audience. The answer to this question may also give insight into market trends and where to effectively spend resources into marketing for the target audience.

The second question will be to determine which brands and categories have the highest purchases per view. This metric may be insightful as it speaks to the intentionality of users. Some products may have high traffic and visibility, but that traffic may not translate directly to revenue, while other products may have appeal to users in a way that drives them to purchase more often. Items with high purchase per view values could represent products that users are finding uniquely with this vendor, products that are more optimized for common search engines, or possibly products that are in demand and may the availability of which may cause the user to purchase before the product runs out.

The third question will be to determine what trends over time we can gather from the data. The time range of data we specifically have spans the dates of October 2019 to April 2020, which entails the COVID 19 shutdown that began in March. We expect to see some trends that could be explained by this global epidemic, along with insights that will also further explain the behavior of the eCommerce vendor's target audience.

Related Work

The world of eCommerce is a dynamic landscape, constantly evolving and presenting new opportunities and challenges. It is also a foundational wealth of data which can present information to anticipate customer needs and drive profits and sales.

Most major eCommerce corporations are very protective of their data, considering it proprietary. The analysis of this data is what helps them to create recommendation algorithms, identify seasonal trends and sales patterns, or forecast consumer behavior with precision.

By harnessing the power of big data and advanced analytics, businesses can tailor their strategies to meet the evolving needs and preferences of their customers. Whether it is predicting the next big trend, optimizing pricing strategies, or personalizing the shopping experience, data-driven insights are the cornerstone of success in the digital marketplace.

One such comprehensive analysis is presented in "Global eCommerce Market Analysis & Trends" by Grand View Research. This study found that the global e-commerce market was worth around USD 9.09 trillion in 2019 and was expected to grow significantly, reaching USD 27.15 trillion by 2027, driven by factors like increasing small and medium enterprises and smartphone usage. The analysis indicated that Asia Pacific dominates this market, with key players including Alibaba, Amazon, and Walmart shaping its trajectory. (Grand View Research, 2020).

Since most companies keep their data private, the dataset we found does not include the name or organizational information where the data originates. However, the public availability of this dataset through the OpenCDP Project posted on Kaggle has allowed widespread analysis and insight collection by many individuals seeking to gain similar insights as our group.

One such project, by user Tshephisho Sefara, analyzed this data to create a predictive model using XGBoost to predict whether a user would add an item to their cart given the date/time as well as other features present in the data. (Sefara, 2020)

With his detailed analysis, it is no wonder his project has the most upvotes on the hosted site, Kaggle, and while we seek to find similar trends and insights as this user did, we believe we can find more compelling trends and prediction points than the 68% accuracy his final model was able to produce. We are interested in answering questions similar to this study, and while many other users have gathered similar insights, we do not seek to replicate their work but gather our own information and create more compelling insights with greater correlation.

Data Set

The project dataset is hosted on Kaggle, found here: [https://www.kaggle.com/datasets/mkechinov/ecommerce-behavior-data-from-multi-category-store.](https://www.kaggle.com/datasets/mkechinov/ecommerce-behavior-data-from-multi-category-store)

Title: "ECommerce behavior data from multicategory store"

This dataset contains 7 months of behavior data from a multi category online store, ranging from October 2019 to April 2020. The data was originally collected by Open CDP project, and due to the size of the total dataset, is embedded as seven datasets, one for each month.

It contains 67,501,979 total values and a storage size of 16.45 Gb. Because of the size of the dataset, it is split into seven individual datasets, one for each month which we will have to join as part of our data cleaning. All of us plan to download the entirety of the dataset and collaborate on the project together.

The dataset contains nine features:

event_time: Continuous data type representing the date and time of the event.

event_type: Nominal data type indicating different types of events (view, purchase, remove from cart).

product_id: Nominal data type identifying unique products.

category_id: Nominal data type denoting product category identifiers.

category_code: Nominal data type describing the category for the product.

brand: Nominal data type specifying brand name of the product.

price: Continuous data type representing product prices.

user_id: Nominal data type identifying unique users.

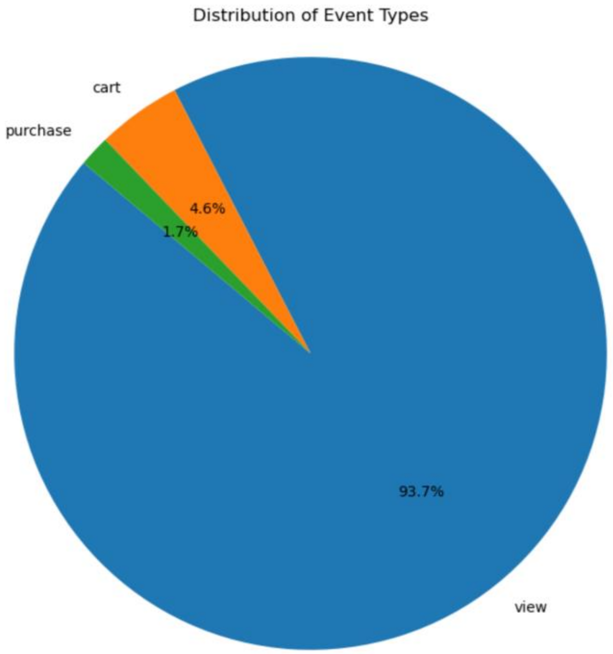
user_session: Nominal type data representing user sessions.

We will use these features accordingly to gather the insights needed to answer our questions.

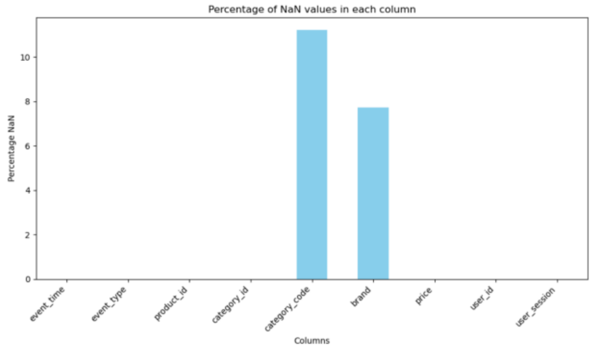
Main Techniques Applied

The biggest hurdle to overcome has been the mere size of the data we were dealing with. Since our dataset contains over 67 million data entries and compressed is over 20 Gb in size, creating a single dataframe to load into memory was not a feasible option.

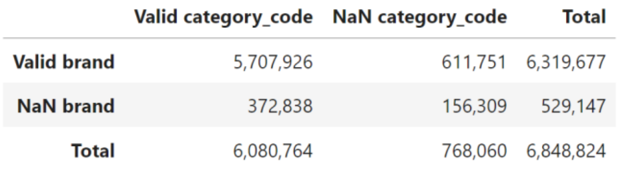
We performed some initial exploratory data analysis by creating a sample dataframe made up of 50,000 entries randomly selected from each month. Turning this into a Pandas dataframe, we were able to determine that a vast majority of our data represented views as opposed to purchases. Since our first two questions involved specifically purchases, we could filter these view entries out.



This diagram shows the results found from this sample dataset. As you can see, over 90% of these datapoints were views, while only 1.7% of all events represented actual purchases. After filtering for only purchased entries, we did some further EDA to gain insight into missing values. The following visualizations represents the breakdown of NaN values and where they were listed:



We determined that the only fields with missing data were the category_code and brand features, with only four missing values in user_session for some unknown reason. To understand the missing values further, we visualized the breakdown using the following table:



	Valid category_code	NaN category_code	Total
Valid brand	5,707,926	611,751	6,319,677
NaN brand	372,838	156,309	529,147
Total	6,080,764	768,060	6,848,824

With this information, we could see that most missing values came from the category_code feature. Of the entries that had a missing category_code field, around 20% were also missing the brand feature. Given these factors, and the size of the data, we determined that dropping any entries with missing values would be preferred over imputing these missing values.

Furthermore, we determined that the user_id and user_session features were not useful for the purposes of the specific questions we were seeking to answer, and so filtered these out as well. The product id values ended up being less informative than our initial thought, given that the eCommerce vendor is unknown and a quick search for many of these id's ended up at many dead ends, so these were filtered out as well.

Finally, we converted the event_time feature to a date_time format in Pandas to be able to be used to answer our second question referring to purchases over time.

With these changes, we were able to have a dataset that had around 5.7 million data entries and took the size of only around 50 Mb compressed using gzip.

Using this cleaned data, we then were able to find the insights we were seeking for in this massive dataset without losing much of the variability.

We also utilized clustering by implementing the K Means algorithm over the brand and category_code features to create fifty clusters to be used for creating suggested brands/item_categories for users on any given product page.

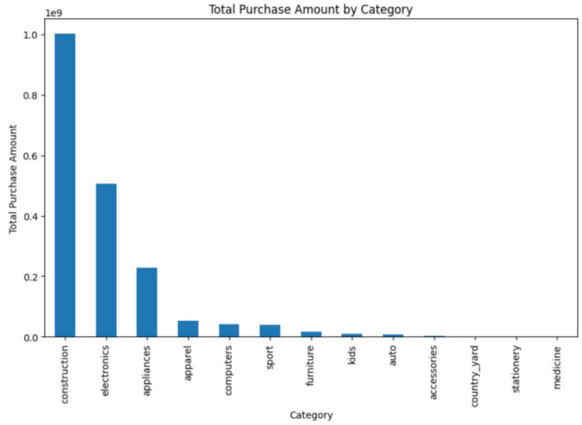
Key Results

Some initial changes in the direction of our product occurred when we were able to look at the data more closely. Initially, we believed we would be able to determine insight on specific products, however since these products are represented by a product ID number that is specific to an unknown vendor, we determined that this feature would not prove to be as insightful.

The initial findings while doing exploratory data analysis and the data cleaning have been explained previously, so we won't reiterate them here. However, there were many interesting findings we were able to uncover by answering the project questions in our project proposal.

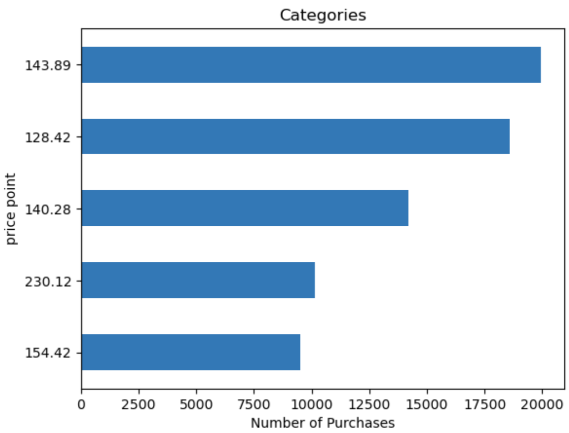
1 Which brands, or item categories, exhibit the highest volume of sales and profits during the specified period?

Since the category_code was parsed with parent and sub-categories (e.g. electronics.audio.headphone), we initially parsed these categories to see the biggest revenue streams coming from the highest parent category. Here is a graphical representation of these findings:



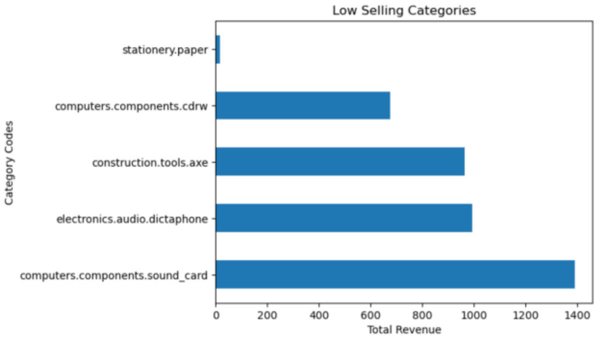
We were surprised to see that the leading contributor of profits was the construction category, followed by eletronics and appliances. To gain additional insight. We sought to gain additional insight to see what might be contributing to the high sales for this category.

Initially, it was thought that construction products might have a high cost, possibly contributing to less sales of items but higher revenue overall. To test this hypothesis, we combed through the data of purchases to find the number of purchases per unique price values in the construction category. Here is a breakdown of these findings:



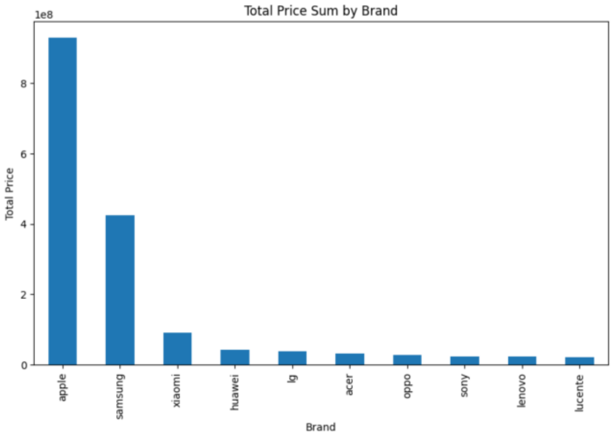
We can see that between $100 and $200 items tend to be the most purchased items in this category, leading us to believe that the volume of sales in this category, rather than cost of items, may be driving the total revenue.

We also took a look at the lowest revenue producers to see the following insights:



Paper seems to be a very low selling product, with some specific computer and electronic components present as low revenue producing products as well. This is insightful in that the marketplace vendor may seek to not spend as much marketing in these areas since users may not be as interested in these niche products.

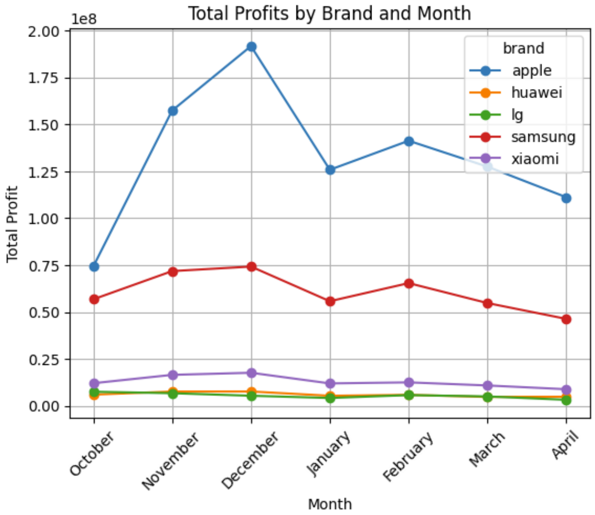
For brands, the findings were much more of what we would expect:



Apple leads the way followed by Samsung as leading cost drivers. This may be consistent with the electronics/smartphones being a major contender for revenue. It is possible that the construction category might have multiple brands that make up its revenue, which may be why it is not represented as much in this graph.

2 Among brands and categories, how do profits fluctuate over the 7 month time frame?

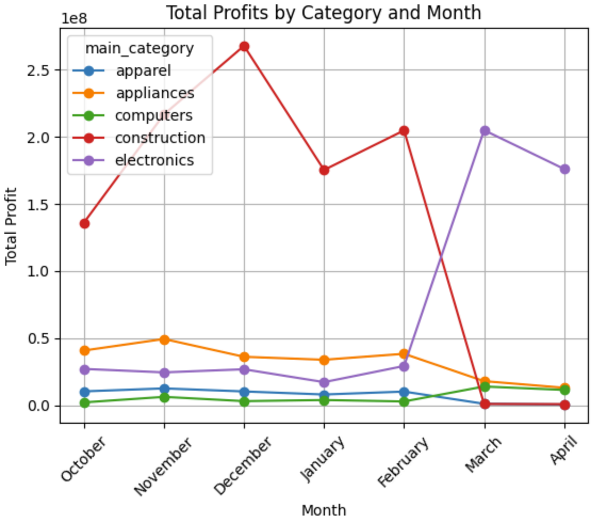
Since we were able to convert the event_time category to a date_time format in our Pandas dataframe, we were able to create visualizations for the breakdown of revenue over time for the top leading brands and categories.



Here we can see Apple continues to show its hold on purchases with the other contenders trailing close behind. The months of November and December show strong peaks for Apple and milder peaks for the other leading brands. This can easily be explained by the Holiday shopping season, as this is commonly a high market trend for purchases during this timeframe.

Interestingly, the COVID 19 shutdown in March and April is not as apparent as one would think, but I would expect popular brands such as these to be fairly insulated from downward market trends caused by this event.

The data for the category code values, however, does show some powerful correlations with the COVID 19 shutdown that started in March of 2020.

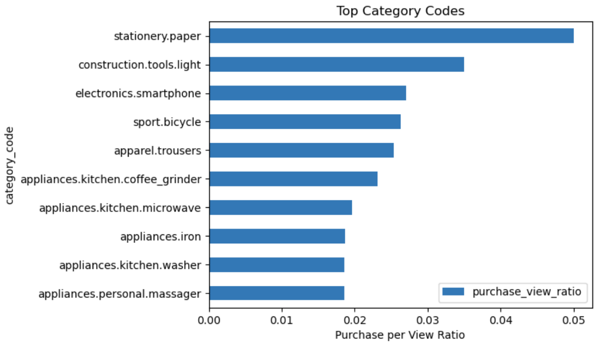


Here we can see sales for the construction category plummet in March, which is easy to correlate with the COVID 19 shutdown. Conversely, electronics sees an uptick at this time. It is easy to explain this trend as people during the shutdown were probably increasing spending on electronics for home office setups, as well as entertainment and connection for friends and family during lockdown.

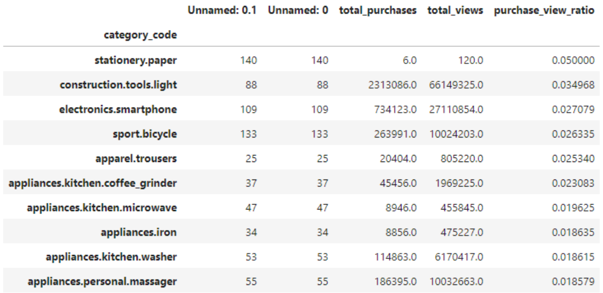
3 Among brands and categories, how do profits fluctuate over the 7 month time frame?

This question obviously involves looking at the data in its totality, as opposed to just purchases. As before, the chunk attribute of the Pandas read_csv method was crucial to crawling through the data to gain insight.

Doing so gave some rather illuminating insight into the types of products that had the most purchases per view.



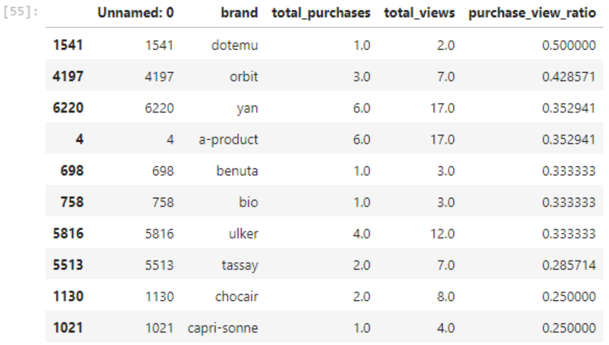
Here we can see that our lowest revenue category has the highest purchases per view, with our leading revenue categories trailing close behind. Looking at the actual views and purchases for these categories can give us further explanation as to why.



	Unnamed: 0.1	Unnamed: 0	total_purchases	total_views	purchase_view_ratio
category_code					
stationery.paper	140	140	6.0	120.0	0.050000
construction.tools.light	88	88	2313086.0	66149325.0	0.034968
electronics.smartphone	109	109	734123.0	27110854.0	0.027079
sport.bicycle	133	133	263991.0	10024203.0	0.026335
apparel.trousers	25	25	20404.0	805220.0	0.025340
appliances.kitchen.coffee_grinder	37	37	45456.0	1969225.0	0.023083
appliances.kitchen.microwave	47	47	8946.0	455845.0	0.019625
appliances.iron	34	34	8856.0	475227.0	0.018635
appliances.kitchen.washer	53	53	114863.0	6170417.0	0.018615
appliances.personal.massager	55	55	186395.0	10032663.0	0.018579

We can see right away that stationary has a very low number of views and an even lower number of purchases, which would likely skew this category as a significant result. The other categories give us further evidence that they are, in fact, leading revenue contenders and users are likely more intentional about purchasing these items as opposed to simply browsing for them.

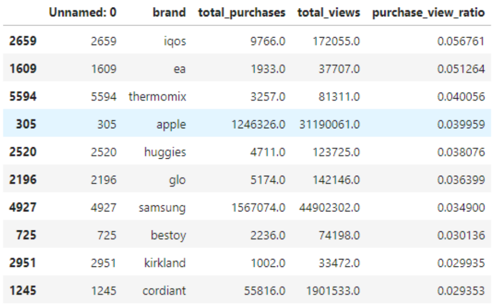
Brands seemed to give us less information initially:



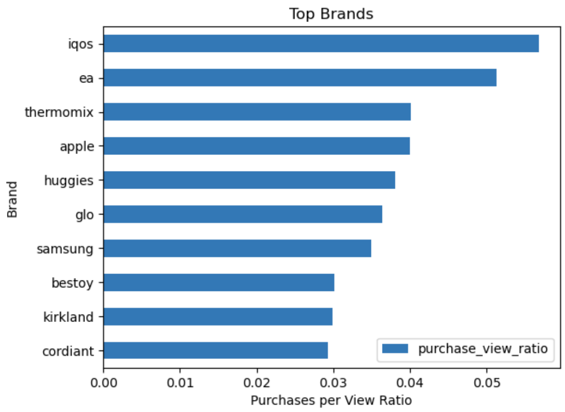
[55]:

	Unnamed: 0	brand	total_purchases	total_views	purchase_view_ratio
1541	1541	dotemu	1.0	2.0	0.500000
4197	4197	orbit	3.0	7.0	0.428571
6220	6220	yan	6.0	17.0	0.352941
4	4	a-product	6.0	17.0	0.352941
698	698	benuta	1.0	3.0	0.333333
758	758	bio	1.0	3.0	0.333333
5816	5816	ulker	4.0	12.0	0.333333
5513	5513	tassay	2.0	7.0	0.285714
1130	1130	chocair	2.0	8.0	0.250000
1021	1021	capri-sonne	1.0	4.0	0.250000

You can see that, much like the stationery category, the top brands with the highest purchase per view ratio have miniscule view and purchase numbers. This could be due to an oversaturation of brands with obscure or niche off-brand products. To gain more information, we decided to only look at brands that have at least 1000 purchases as well as views.



	Unnamed: 0	brand	total_purchases	total_views	purchase_view_ratio
2659	2659	iqos	9766.0	172055.0	0.056761
1609	1609	ea	1933.0	37707.0	0.051264
5594	5594	thermomix	3257.0	81311.0	0.040056
305	305	apple	1246326.0	31190061.0	0.039959
2520	2520	huggies	4711.0	123725.0	0.038076
2196	2196	glo	5174.0	142146.0	0.036399
4927	4927	samsung	1567074.0	44902302.0	0.034900
725	725	bestoy	2236.0	74198.0	0.030136
2951	2951	kirkland	1002.0	33472.0	0.029935
1245	1245	cordiant	55816.0	1901533.0	0.029353



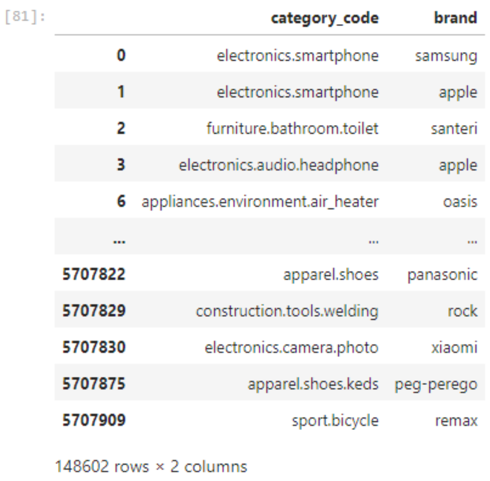
Here we can see Apple and Samsung lower in our list, but the highest brands are IQOS, EA, and Thermomix, with Huggies closely following Apple. These could be brands the eCommerce vendor may want to understand further as to why users may be more likely to be intentional about visiting the product page to purchase.

The figures could indicate the popularity of the brand/category from above, or possibly that the search engine optimization causes more users to be directed to these pages. Another interpretation would be whether the products in these categories/brands are unique to this eCommerce site, contributing to more people coming to them specifically to purchase rather than view.

3 Which products are similar, and can we generate a suggested list of items for users based on this data?

To answer this question, we decided to implement a clustering model using the K-Means algorithm using unique brand and category pairs. These two features should be closely related and the clustering would be helpful in creating a related items list for each product page taken from products that are located nearby in the same cluster.

We discovered that there were only 148,602 unique category/brand pairs:

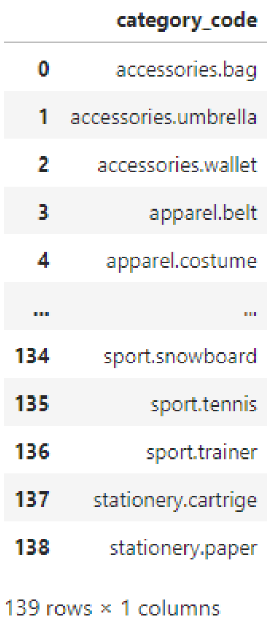


[81]:

	category_code	brand
0	electronics.smartphone	samsung
1	electronics.smartphone	apple
2	furniture.bathroom.toilet	santeri
3	electronics.audio.headphone	apple
6	appliances.environment.air_heater	oasis
...	...	...
5707822	apparel.shoes	panasonic
5707829	construction.tools.welding	rock
5707830	electronics.camera.photo	xiaomi
5707875	apparel.shoes.keds	peg-perego
5707909	sport.bicycle	remax

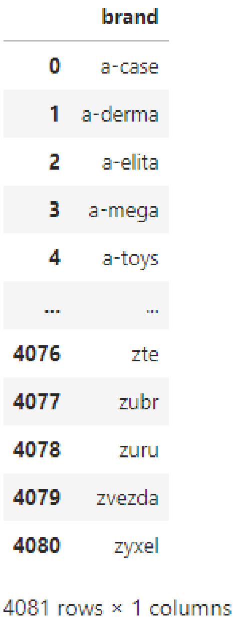
148602 rows × 2 columns

Additionally, we were able to see that only 139 unique categories were present in the purchased data with 4,081 different brands.



	category_code
0	accessories.bag
1	accessories.umbrella
2	accessories.wallet
3	apparel.belt
4	apparel.costume
...	...
134	sport.snowboard
135	sport.tennis
136	sport.trainer
137	stationery.cartrige
138	stationery.paper

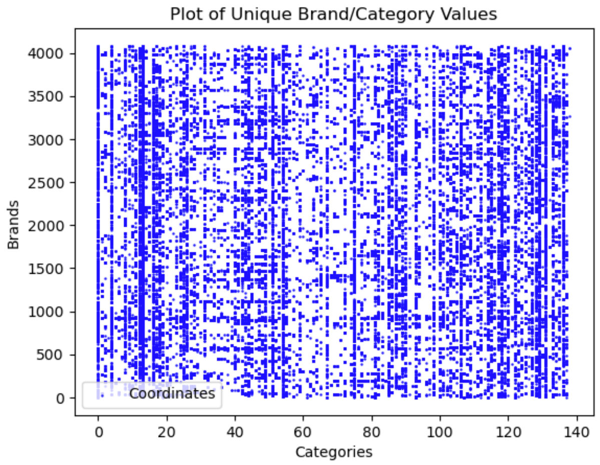
139 rows × 1 columns



	brand
0	a-case
1	a-derma
2	a-elita
3	a-mega
4	a-toys
...	...
4076	zte
4077	zubr
4078	zuru
4079	zvezda
4080	zyxel

4081 rows × 1 columns

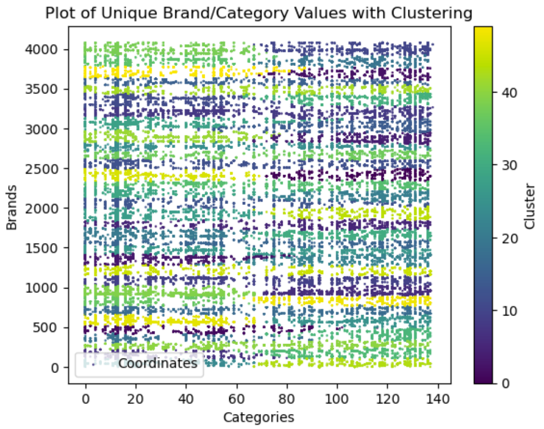
We then converted the unique brand/category_code pairs into numerical x and y values and graphed the corresponding results.



As you can see, there is a very dense representation with a few sparse categories towards the middle of the graph. Because of the density and the quantity of data points, it may be best to create many clusters as opposed to only a few.

Running the K-Means algorithm creates centroids for each group, assigns each data point to the nearest centroid and recalculates each centroid as the mean of the newly assigned members. This process continues until convergence is reached, or the members of each cluster and the centroids are no longer updated.

Here is a graphical representation of the K-Means algorithm performed with a K value of 50 to accommodate the density and quantity of data points:



As you can see, the clustering algorithm seems to prioritize horizontal relationships, representing the different categories among single brands. This may be because brands might sell products in neighboring categories, whereas clustering by categories might not present as much similarity. Since K-Means uses Euclidean distance, we can assume there is greater density between brands than between categories as a result.

Given the output of the K-Means algorithm, we would easily be able to create a new characteristic for each product representing it's associated cluster, and a random generator could be used to retrieve several items belonging to the same cluster as the product being viewed.

APPLICATIONS

As stated previously, these insights give great insight into target user behavior, marketing trends caused by the typical Holiday shopping season and the atypical COVID 19 pandemic and resulting lockdown. Additionally, the insights we gained would create increased marketability, even providing our hypothetical client with a way to create targeted product recommendations from our clustering model. In summation, these insights have real-world value that could result in a greater understanding of the market and the users who interface with the eCommerce marketplace, which could potentially drive profits and revenue.

REFERENCES

[1] Kechinov, Michael, 2020. "eCommerce behavior data from multi category store." Retrieved from [https://www.kaggle.com/datasets/mkechinov/ecommerce-behavior-data-from-multi-category-store].

[2] Grand View Research, 2020. "Global eCommerce Market Analysis & Trends." Retrieved from [https://www.grandviewresearch.com/industry-analysis/e-commerce-market].

[3] Tshephisho, Sefara, 2020. "eCommerce behaviour using XGBoost." Retrieved from [https://www.kaggle.com/code/tshephisho/ecommerce-behaviour-using-xgboost/notebook#Know-your-Customers].