Readme

https://github.com/shubhdhebar/Consumer-Behavior-Analysis-on-E-Commerce-Platform

Problem Statement

In today's world, shopping has largely become an errand that is run online. Hence, an enormous amount of data is generated on a daily basis on various e-commerce platforms. This data contains potential insights that can allow e-commerce businesses to implement strategies that will help them optimize their product placements, and eventually the revenue. This project intends to harness the power of Big Data Analytics in order to generate these insights.

Solution

The data collected from the eCommerce platform can be processed using the concepts of parallelization by Apache Spark. The DataPrep and MatPlotLib libraries are then used to generate visualizations of data. It is divided on the basis Event-Type to allow the e-commerce platform to gain insights such as:

1. What are the most popular products viewed/added to cart/purchased?
2. How likely is a viewed product added to the cart?
3. How likely is a product in the cart purchased?
4. What products are frequently bought together?

K-Means clustering can then be used to perform Market Segmentation for the platform.

Dataset

The dataset contains consumer behavior data from November 2019 for a major e-commerce multi-category store. It is a large dataset (5GB) taken from Kaggle: https://www.kaggle.com/datasets/mkechinov/ecommerce-behavior-data-from-multi-category-store

Columns

event-time: timestamp of the collected datapoint
event-type: view/cart/remove from cart/purchase depending on the customer action
product_id: the product that is undergoing the event
category_id: id corresponding to the category of the product
category_code: code corresponding to the category of the product
brand: name of the brand
price

user_id
session_id

Tools Used

1. PySpark
2. DataPrep
3. MatPlotLib
4. SparkML
5. Pandas
6. Kaggle Notebook

Data Pre-Processing

1. Missing Values: Rows carrying missing values were dropped.

```
df_market = df_market.na.drop(how="any")
```

2. Duplicate values: Rows with duplicate values were also dropped.

```
df_market= df_market.dropDuplicates()
```

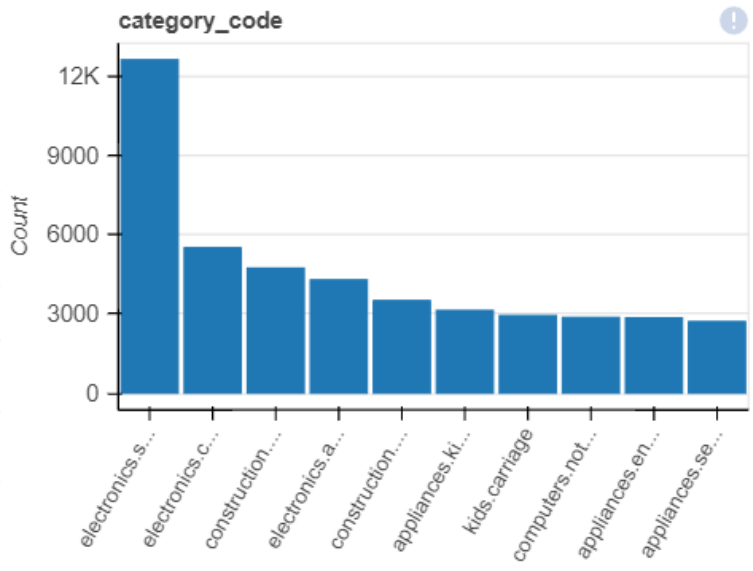
Data Analysis and Visualization

For Event-Type: View

category_code
categorical

Show Details

Approximate Distinct Count	126
Approximate Unique (%)	0.1%
Missing	0
Missing (%)	0.0%
Memory Size	10.4 MB

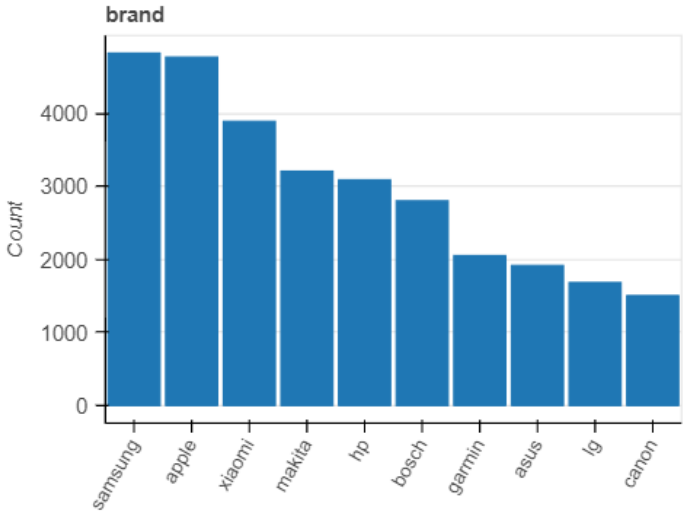


Top 10 of 126 category_code

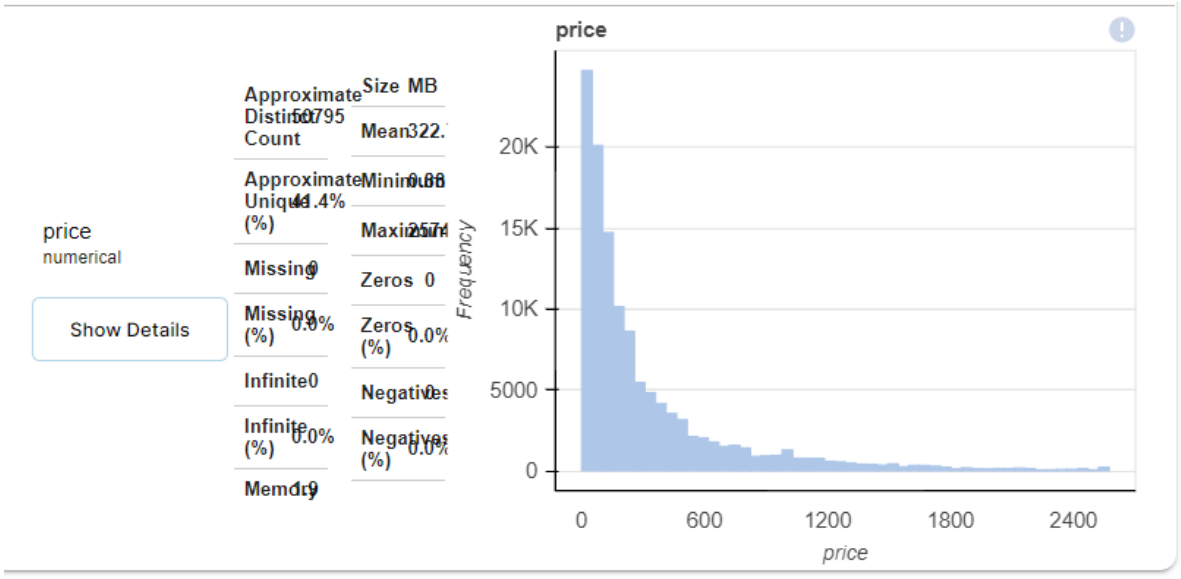
brand
categorical

Show Details

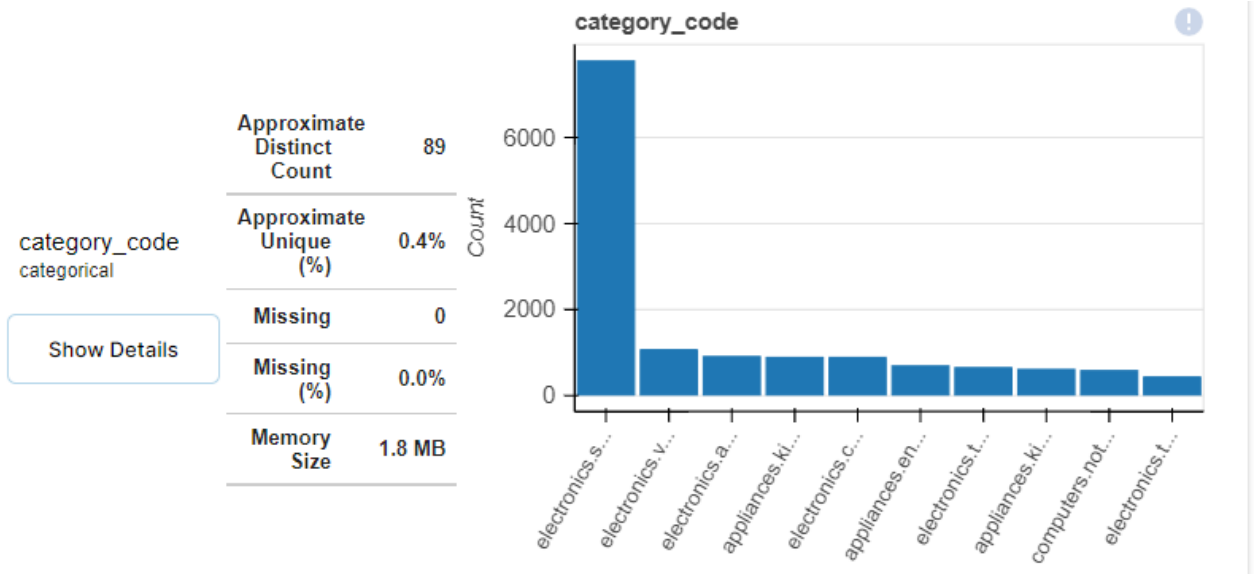
Approximate Distinct Count	1731
Approximate Unique (%)	1.4%
Missing	0
Missing (%)	0.0%
Memory Size	8.3 MB



Top 10 of 1731 brand



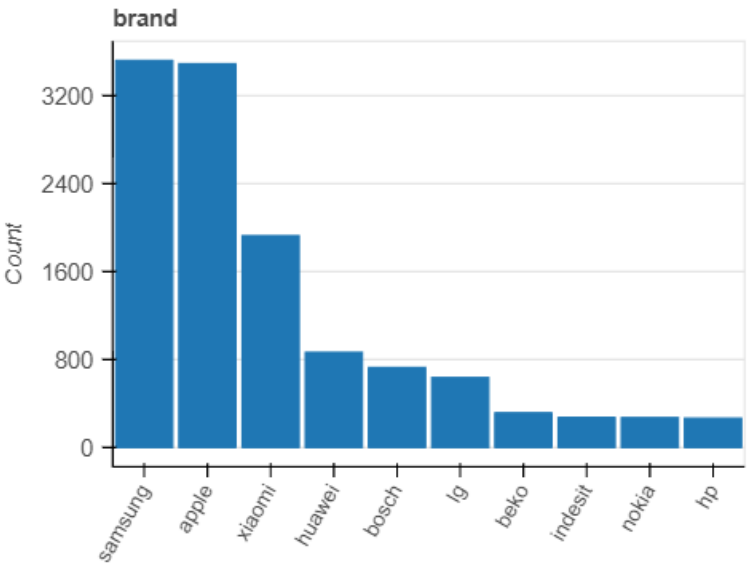
For Event-type: Cart



brand
categorical

Show Details

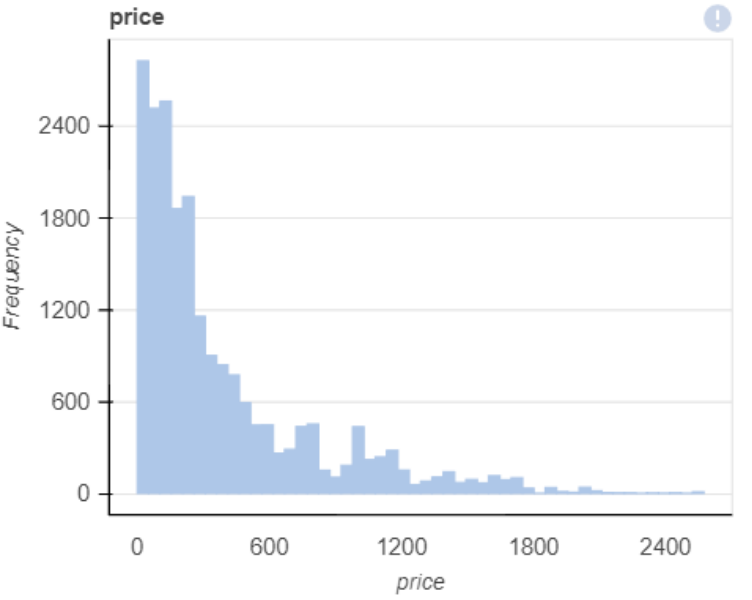
Approximate Distinct Count	369
Approximate Unique (%)	1.7%
Missing	0
Missing (%)	0.0%
Memory Size	1.4 MB



price
numerical

Show Details

Approximate Distinct Count	1700	Size KB	
Approximate Unique (%)	69.7%	Mean	383.
Missing	0	Minimum	1
Missing (%)	0.0%	Maximum	257
Infinite	0	Zeros	0
Infinite (%)	0.0%	Zeros (%)	0.0%
Memory	29.8	Negatives	0
		Negatives (%)	0.0%

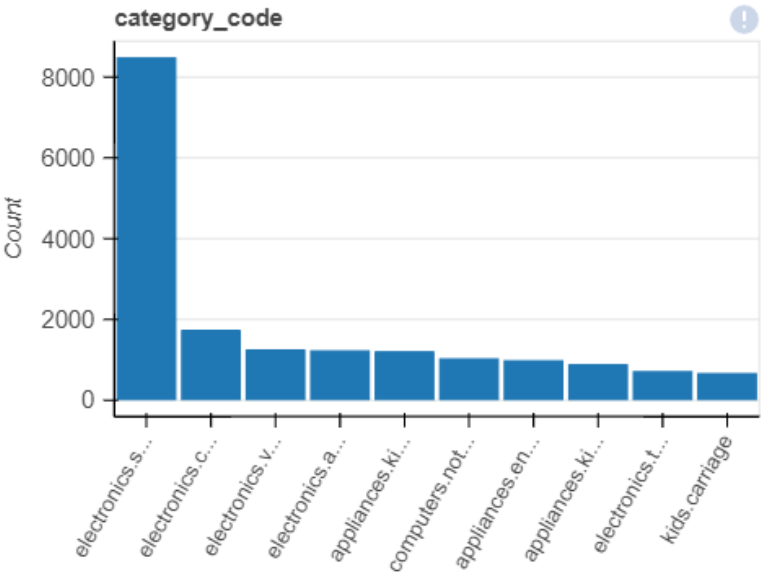


For Event-type: Purchase

category_code
categorical

Show Details

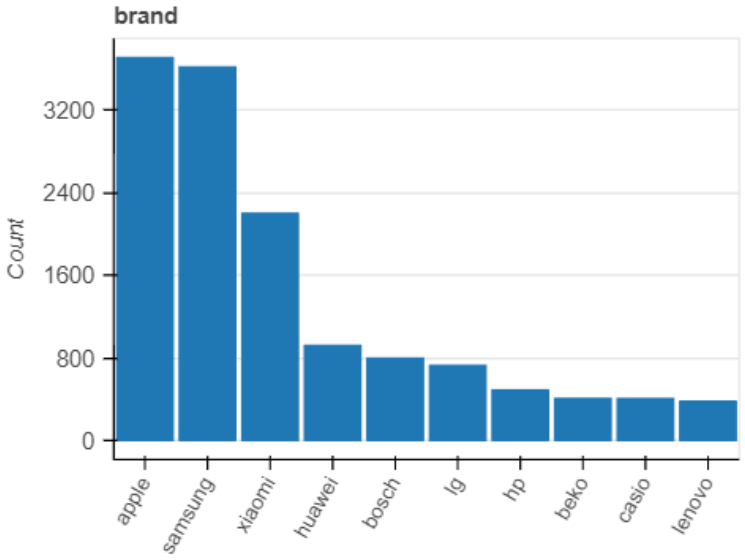
Approximate Distinct Count	120
Approximate Unique (%)	0.4%
Missing	0
Missing (%)	0.0%
Memory Size	2.7 MB



brand
categorical

Show Details

Approximate Distinct Count	1052
Approximate Unique (%)	3.2%
Missing	0
Missing (%)	0.0%
Memory Size	2.2 MB



price
numerical

Show Details

Approximate Distinct Count	1839	Size KB	
Approximate Unique (%)	56.6%	Mean	324.
Missing	0	Minimum	0.88
Missing (%)	0.0%	Maximum	2577
Infinite	0	Zeros	0
Infinite (%)	0.0%	Zeros (%)	0.0%
Memory	507.9	Negatives	0
		Negatives (%)	0.0%