E-commerce Analytics

DEEPA KUMARI

Introduction

In the rapidly evolving digital marketplace, an electronics store aims to stay ahead by optimizing its e-commerce platform. The primary goal of this project is to enhance marketing strategies and improve customer experience, thereby driving increased sales and fostering customer loyalty. The store seeks to gain a deeper understanding of customer behaviors, preferences, and purchasing patterns through advanced data analysis techniques.

Aim

The project aims to enhance marketing strategies, improve customer experience, and ultimately increase sales and customer loyalty.

The Process

- Data Collection
- Data cleaning
- Data Analysis
- Insights

Data Collection

- The data has been collected from Kaggle in the form zip file, then extracted in the form of a csv file

- The data has 26,33,521 rows and 8 columns

Data Dictionary

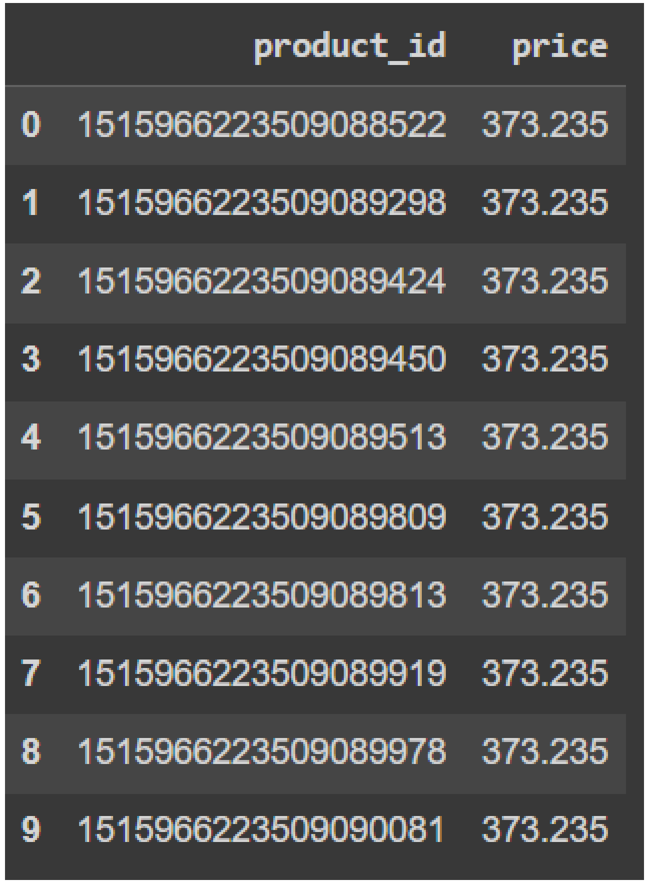
- Event_time: Represents the timestamp indicating the occurrence of a purchase or related event (e.g. adding to cart, viewing). Vital for analyzing purchase patterns across time.
- Order_Id: A unique identifier assigned to each other, facilitating individual transaction tracking and crucial for distinguishing between different orders during analysis.
- Product_Id: Unique identification for each product purchased, pivotal for product-level analysis and identification of specific items
- Category_id: An exclusive identifier for the category of each product. Aids in categorizing products for comprehensive analysis
- Category_code: Possibly a textual or descriptive representation of the product category. It offers a more intuitive understanding than category IDs regarding product types
- Brand: Signifies the brand of the product, important for brand-level analysis and understanding of customer brand preferences
- Price: The selling price of the product. Essential for revenue analysis and comprehensive purchasing patterns
- User_Id: A distinctive identifier assigned to each customer

Data Cleaning

- In category_id, category_code, brand, price, user_id has around 16%, 23%, 19%, 16% and 78% missing values
- Since user_id has more than 50% of data as missing values, so decided to drop that column. For other columns, used the imputation method for handling missing values
- Changed the data type of event_time column into datetime
- There are 675 duplicate values, so dropped the duplicate values

Data Analysis

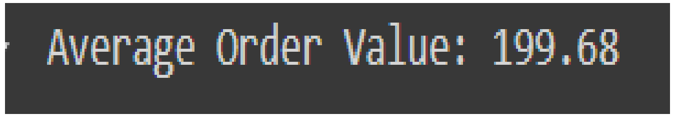
1. Identify the Top 10 most expensive products



	product_id	price
0	1515966223509088522	373.235
1	1515966223509089298	373.235
2	1515966223509089424	373.235
3	1515966223509089450	373.235
4	1515966223509089513	373.235
5	1515966223509089809	373.235
6	1515966223509089813	373.235
7	1515966223509089919	373.235
8	1515966223509089978	373.235
9	1515966223509090081	373.235

These are the Top 10 most expensive products with the same price value

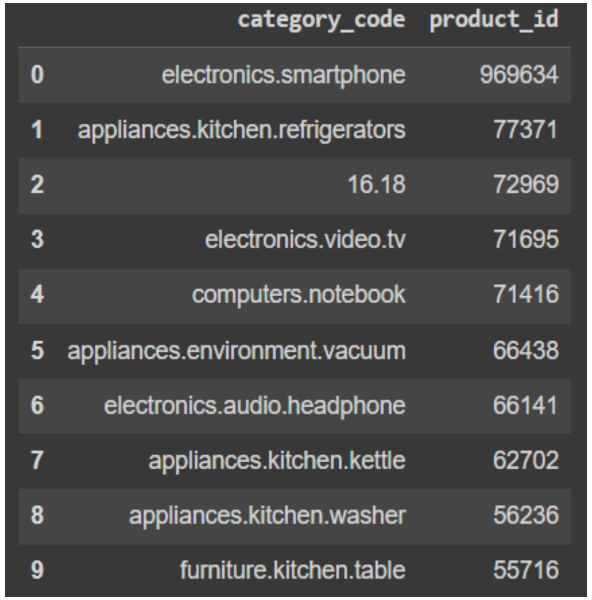
2. Calculate the average order value



```
Average Order Value: 199.68
```

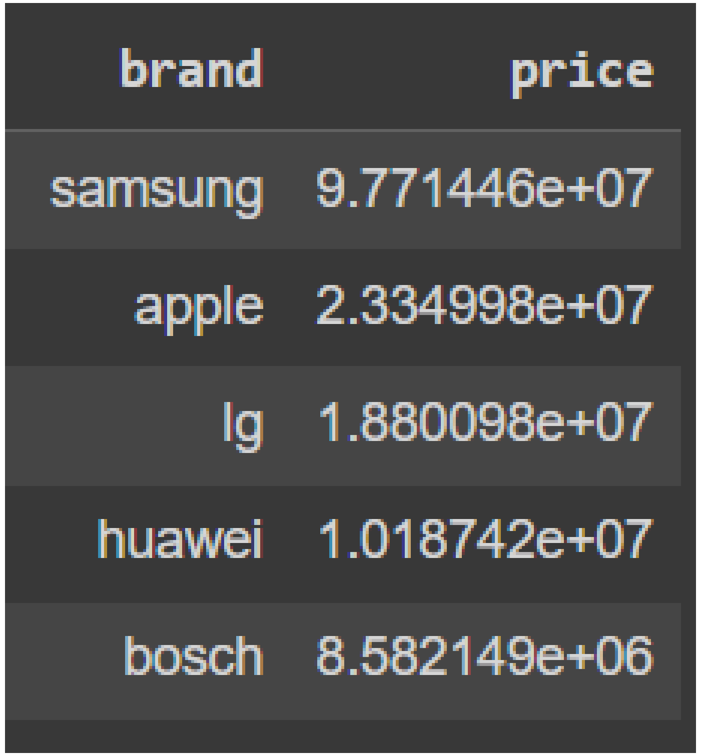
The average customer spends 199.68 rupees per order, indicating a moderate spending pattern. This can help in understanding customer purchasing power and habits.

3. Determine the most popular product categories



	category_code	product_id
0	electronics.smartphone	969634
1	appliances.kitchen.refrigerators	77371
2	16.18	72969
3	electronics.video.tv	71695
4	computers.notebook	71416
5	appliances.environment.vacuum	66438
6	electronics.audio.headphone	66141
7	appliances.kitchen.kettle	62702
8	appliances.kitchen.washer	56236
9	furniture.kitchen.table	55716

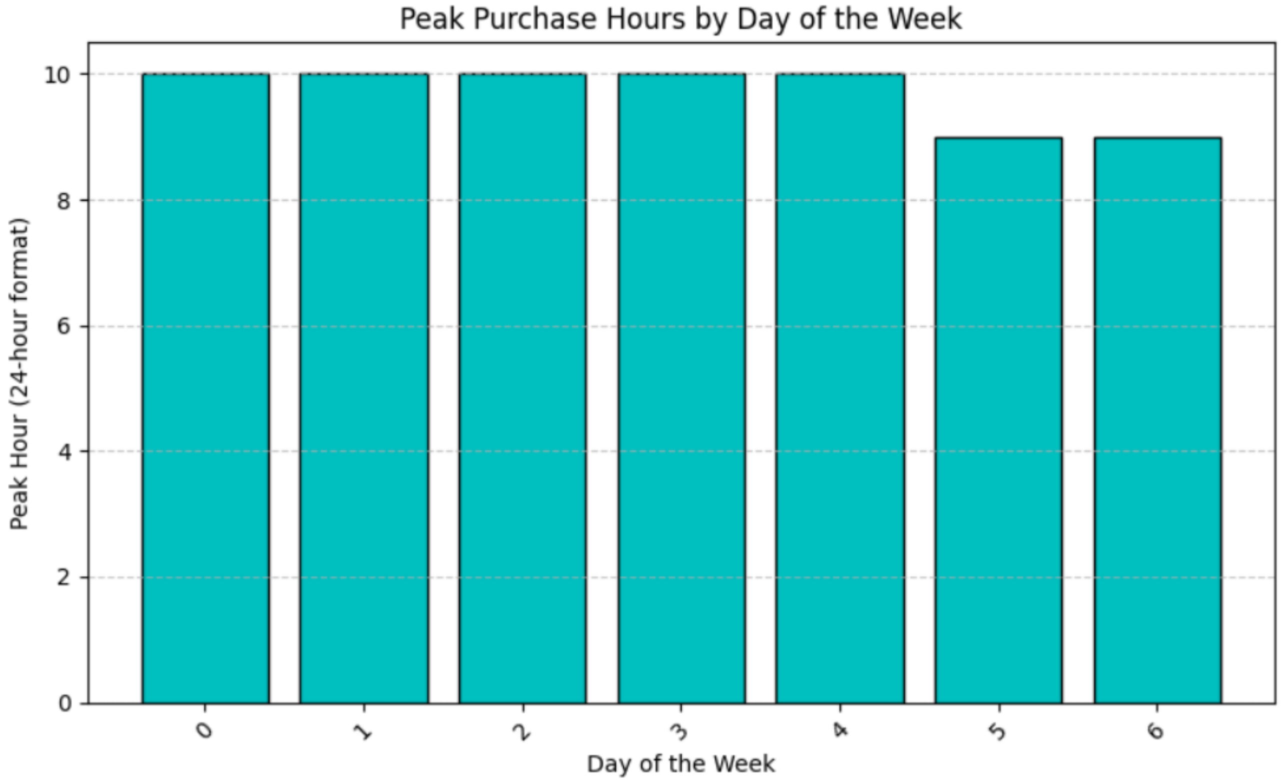
4. Determine which brand has the highest sales



brand	price
samsung	9.771446e+07
apple	2.334998e+07
lg	1.880098e+07
huawei	1.018742e+07
bosch	8.582149e+06

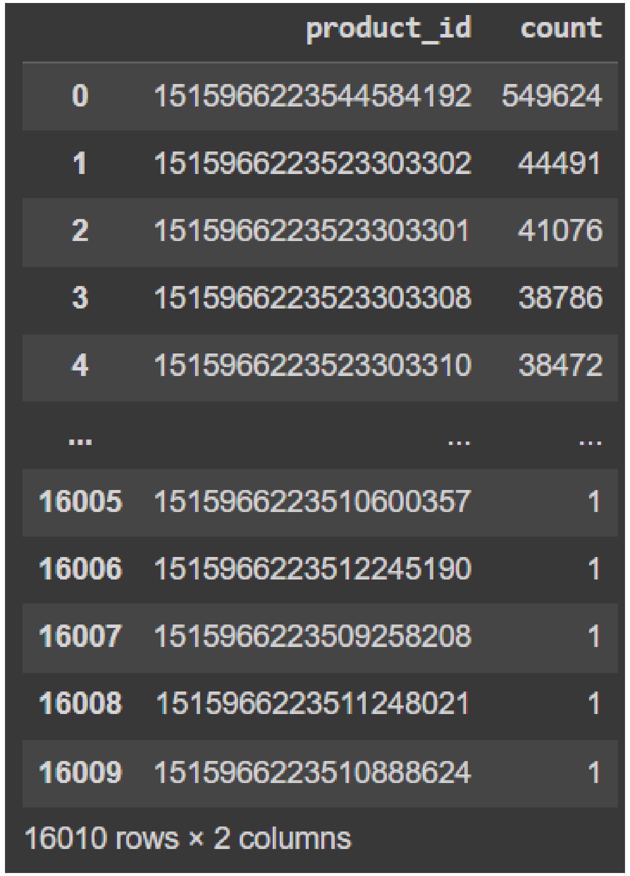
Samsung, Apple, LG, Huawei, and Bosch are the top 5 brands which have the highest sales

5. Finding peak purchase hours of each day of the week



The peak purchase hours are 9 and 10 on most days of the week

6. Identify the frequency of purchases for different products

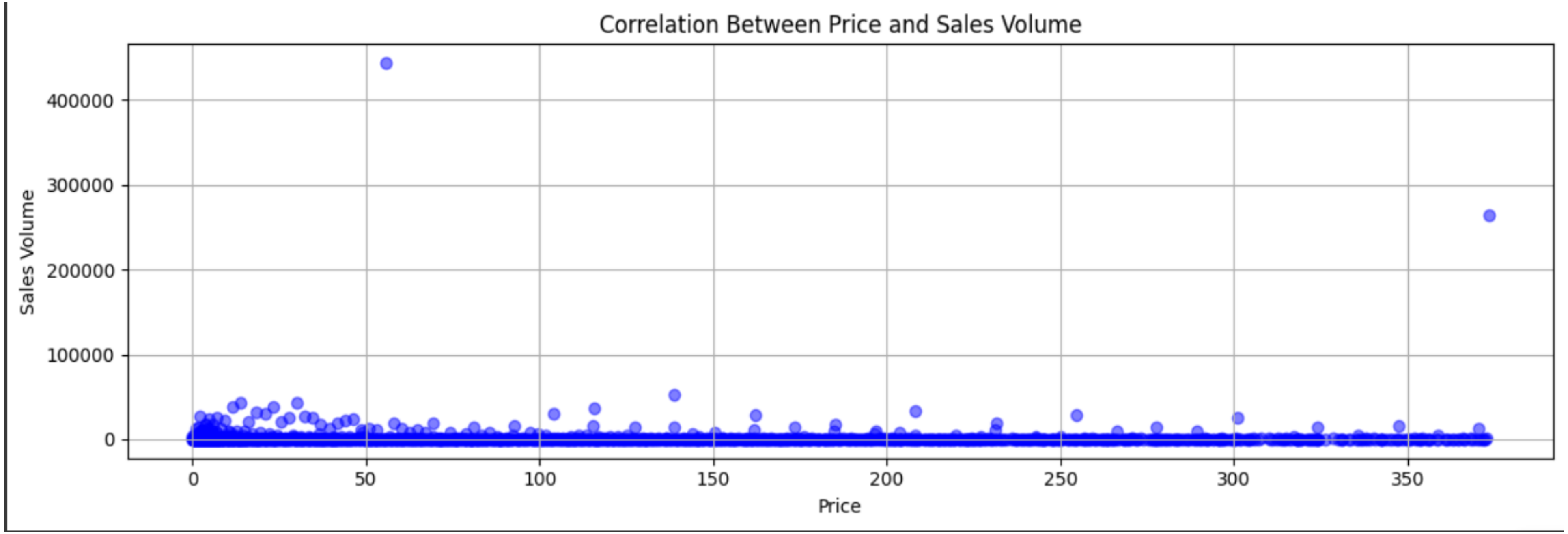


	product_id	count
0	1515966223544584192	549624
1	1515966223523303302	44491
2	1515966223523303301	41076
3	1515966223523303308	38786
4	1515966223523303310	38472
...	...	...
16005	1515966223510600357	1
16006	1515966223512245190	1
16007	1515966223509258208	1
16008	1515966223511248021	1
16009	1515966223510888624	1

16010 rows × 2 columns

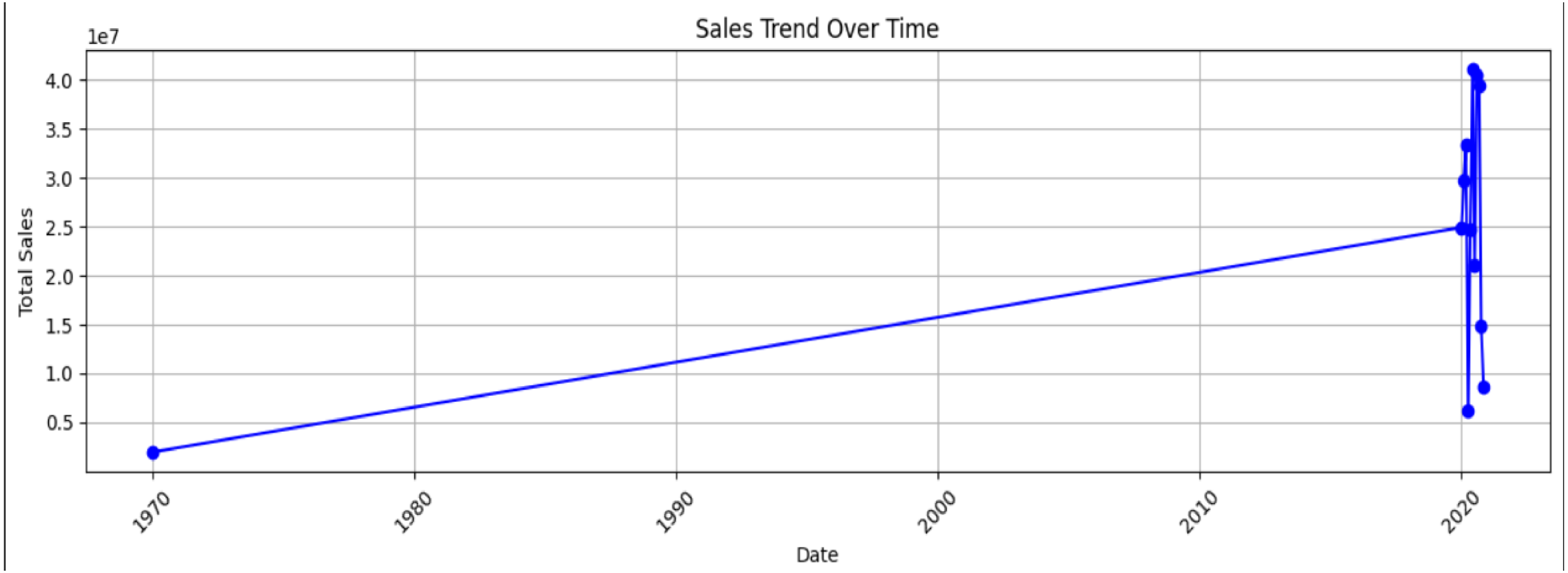
The product ID 1515966223544584192 has the highest frequency of purchases i.e. 5,49,624

7. Analyze the impact of price change on sales volume



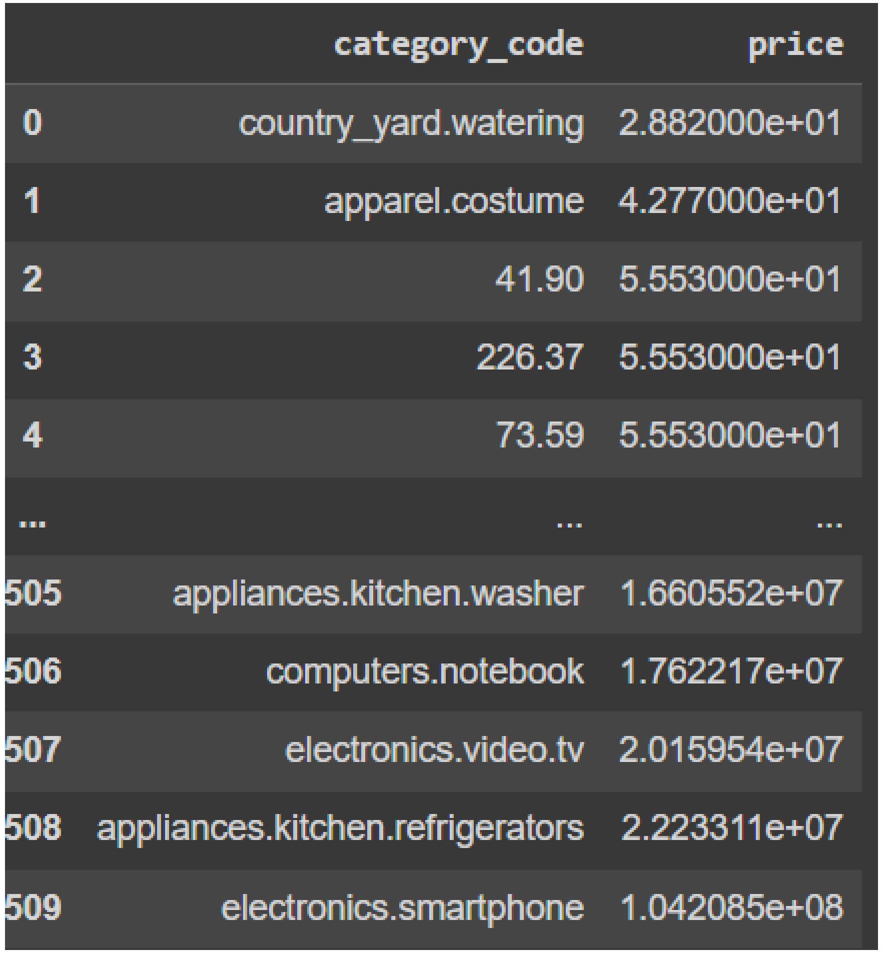
The graph shows that the price change has a significant impact on sales volume. As the price is less, the sales volume is high, and on increasing the price, sales volume also decreases.

8. Identify trends in sales over time



- The graph shows a steady increase in total sales from around 1970 until 2020.
- This suggests consistent growth over the decades.

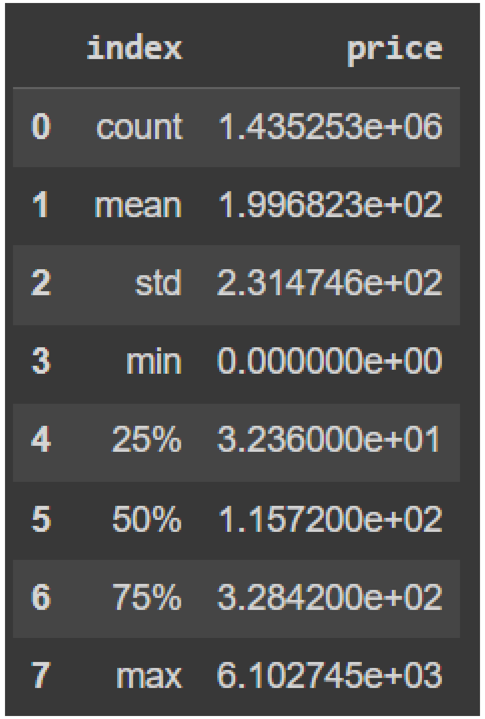
9. Analyze how spending is distributed across different categories



	category_code	price
0	country_yard.watering	2.882000e+01
1	apparel.costume	4.277000e+01
2	41.90	5.553000e+01
3	226.37	5.553000e+01
4	73.59	5.553000e+01
...	...	...
505	appliances.kitchen.washer	1.660552e+07
506	computers.notebook	1.762217e+07
507	electronics.video.tv	2.015954e+07
508	appliances.kitchen.refrigerators	2.223311e+07
509	electronics.smartphone	1.042085e+08

The categories like "appliances.kitchen.washer," "computers.notebook," "electronics.video.tv," "appliances.kitchen.refrigerators," and "electronics.smartphone" have notably high prices (in the range of millions). This suggests that these categories are associated with high-end or expensive products.

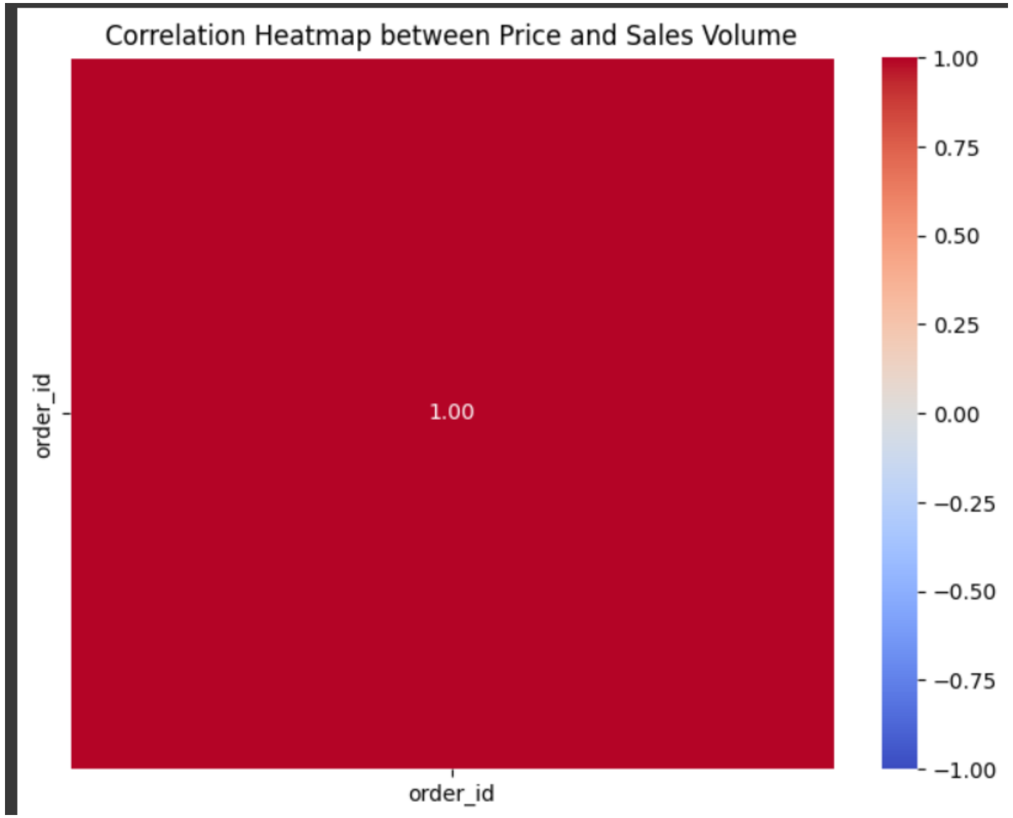
10. Determine the distribution of Order value



	index	price
0	count	1.435253e+06
1	mean	1.996823e+02
2	std	2.314746e+02
3	min	0.000000e+00
4	25%	3.236000e+01
5	50%	1.157200e+02
6	75%	3.284200e+02
7	max	6.102745e+03

- The 25th percentile of the price data is 32.36 units, meaning 25% of the prices are below this value.
- The 75th percentile of the price data is 328.42 units, meaning 75% of the prices are below this value.
- The maximum price recorded is 6,102.75 units, indicating the highest price in the dataset.
- The minimum price recorded is 0. This might represent free items or possibly erroneous entries.

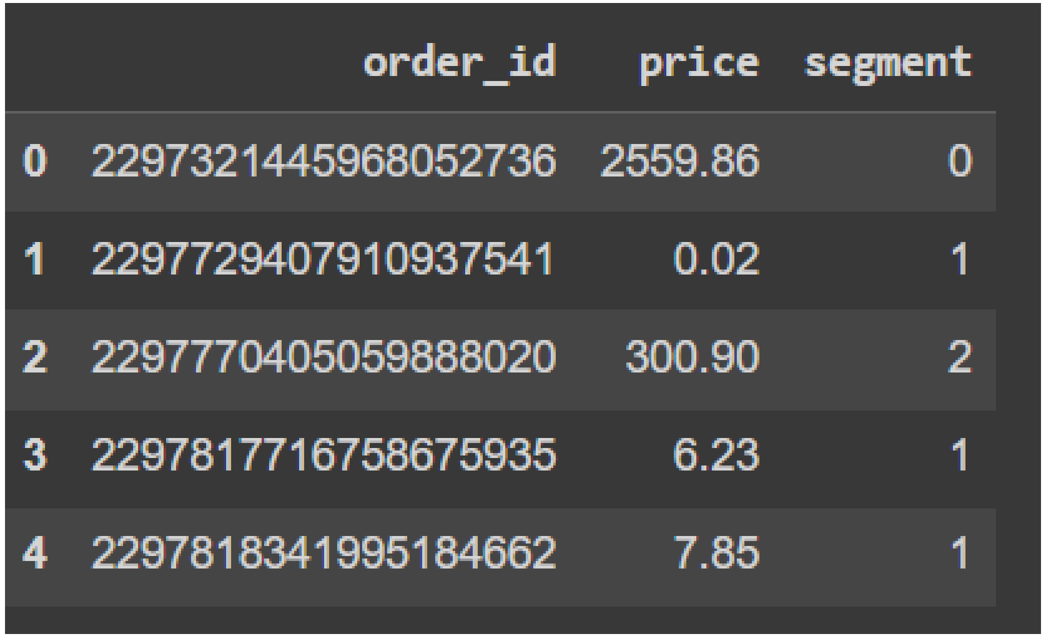
11. Examine the correlation between price and sales volume



The correlation heatmap shows a perfect positive correlation between price and sales volume. This means that as price increases, sales volume also increases proportionally. However, this is likely due to a confounding variable or measurement error. It's unlikely that a direct relationship exists between price and sales volume, as higher prices generally lead to lower demand

A third factor might influence price and sales volume, causing them to appear correlated. For example, a popular product might have a higher price and higher sales volume due to its popularity.

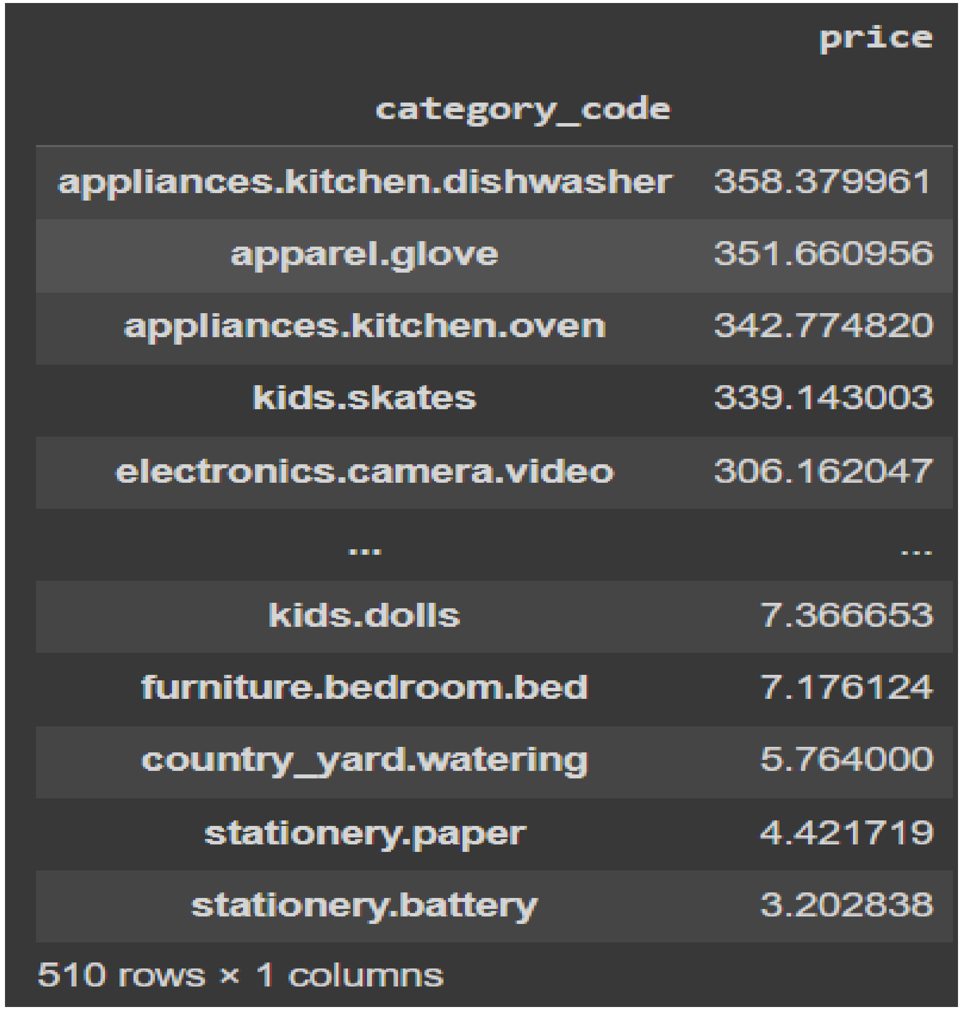
12. Segment customers based on their purchase patterns and behaviors .



	order_id	price	segment
0	2297321445968052736	2559.86	0
1	2297729407910937541	0.02	1
2	2297770405059888020	300.90	2
3	2297817716758675935	6.23	1
4	2297818341995184662	7.85	1

Customers in Segment 0 are high-value customers who spend large amounts on each order, segment 1 is low-value customers and Segment 2 are average-level customers who spend an average amount on each order

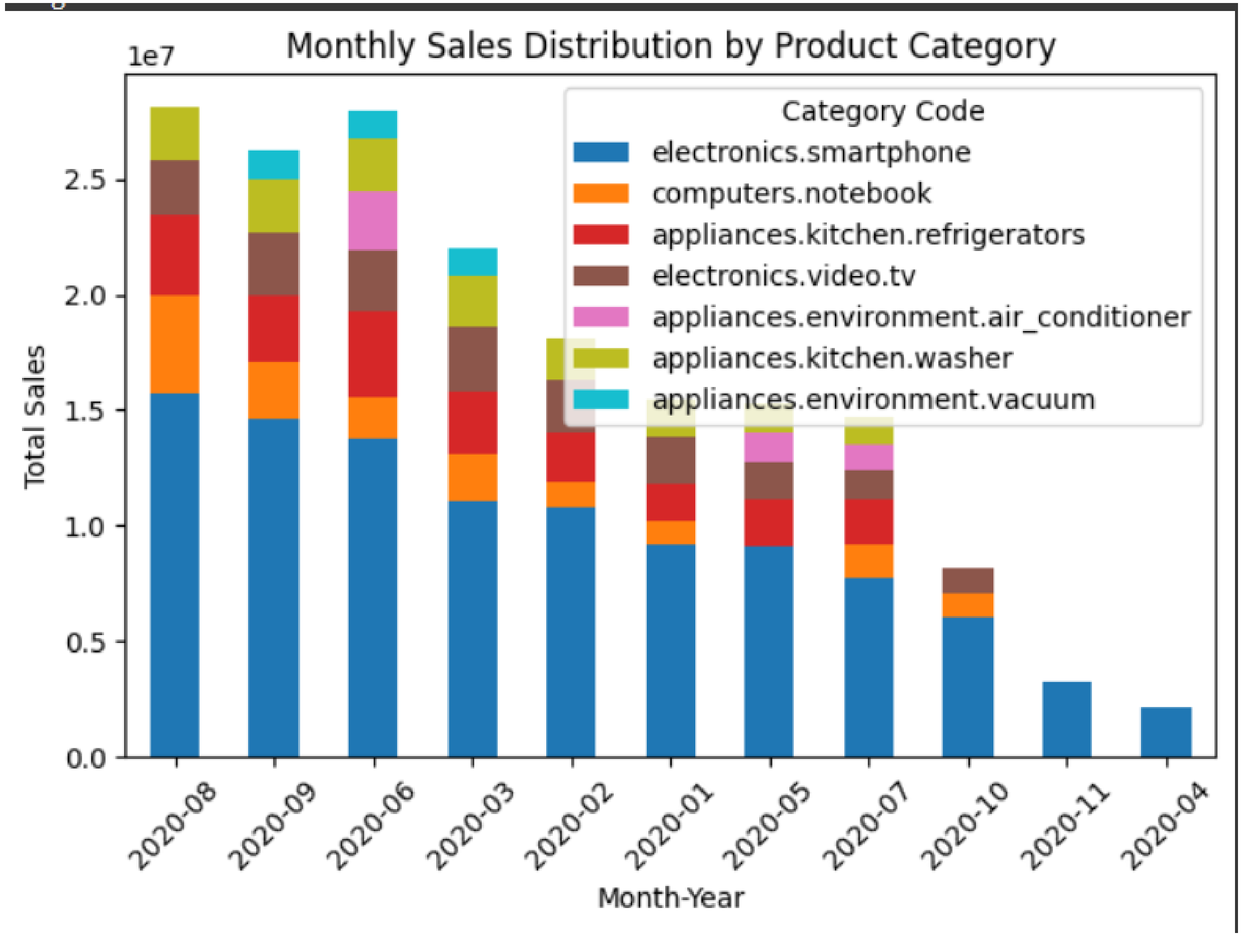
13. Average price of products within each category



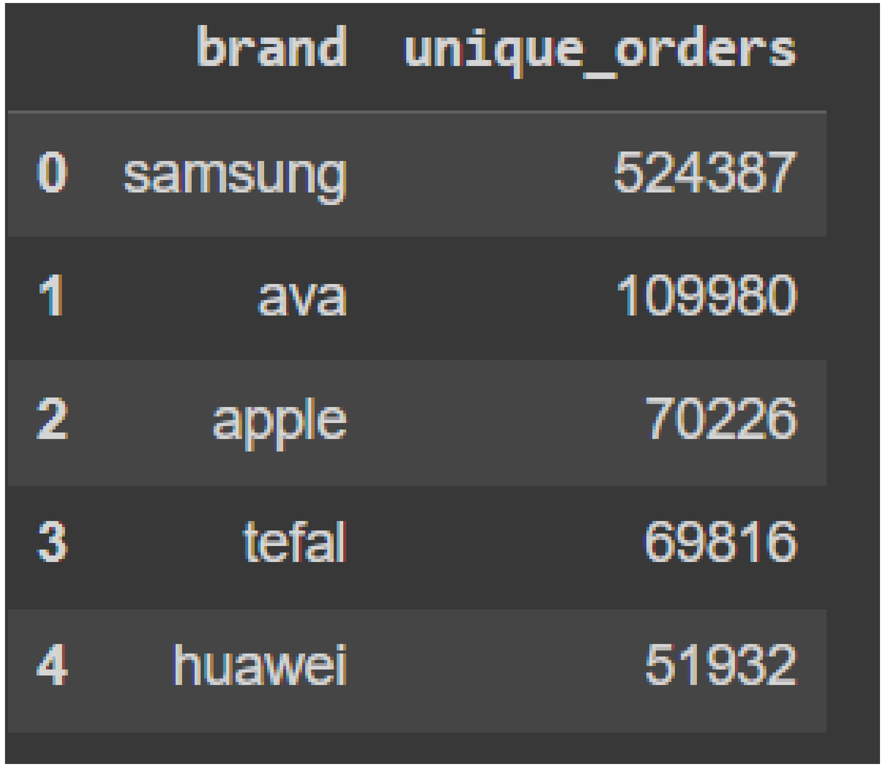
	price
category_code	
appliances.kitchen.dishwasher	358.379961
apparel.glove	351.660956
appliances.kitchen.oven	342.774820
kids.skates	339.143003
electronics.camera.video	306.162047
...	...
kids.dolls	7.366653
furniture.bedroom.bed	7.176124
country_yard.watering	5.764000
stationery.paper	4.421719
stationery.battery	3.202838

510 rows × 1 columns

14. Identify monthly sales distribution by top 10 product category



15. Top 5 brands have the highest customer loyalty, measured by the number of repeat purchases within the same brand?



	brand	unique_orders
0	samsung	524387
1	ava	109980
2	apple	70226
3	tefal	69816
4	huawei	51932

Samsung, Ava, Apple, Tefal, and Huawei are the top 5 brands with the highest repeat purchases

Thank You