SYSTEMS ANALYSIS AND DESIGN
CIS 5200-01

Ecommerce Behavior Data from Multi Category Store

Project – 2

Lekha AjitKumar
Sushmitha Dandu
Dauren Omarov
Navyasree Sriramoju

DATA SPECIFICATION

Data Size : 15.83 GB

Data Source
URL: https://www.kaggle.com/datasets/mkechinov/ecommerce-behavior-data-from-multi-category-store?select=2019-Oct.csv

Dataset Column: 10

November Dataset: 6.7 million rows

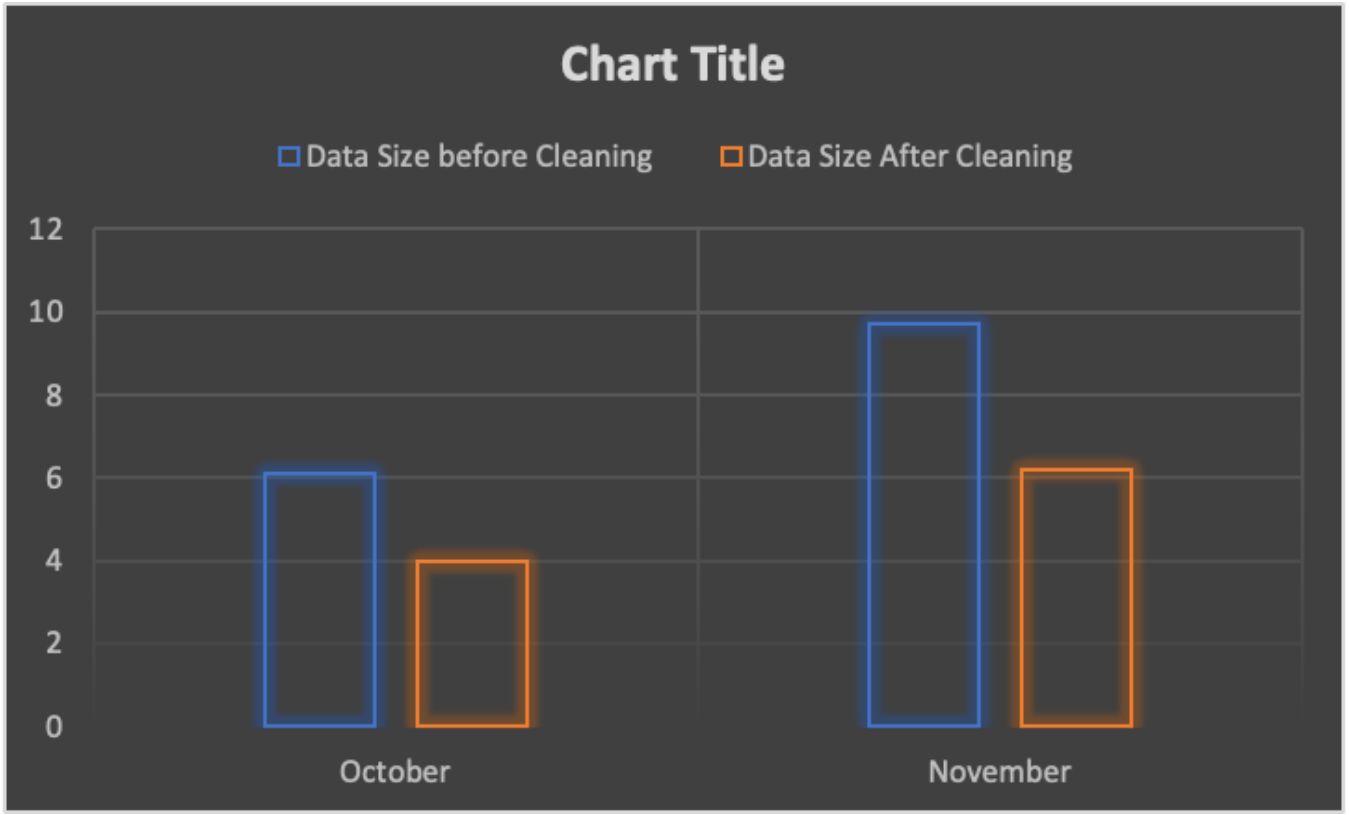
October Dataset: 4.2 million rows

Number of Files: 2

GitHub Link: https://github.com/Lekha19202/E-commerce-customer-behaviour-uding-Hadoop.git

DATA CLEANING

File Name	Data Size before Cleaning	Data Size After Cleaning
October	6.11 GB	3.98 GB
November	9.72 GB	6.2 GB



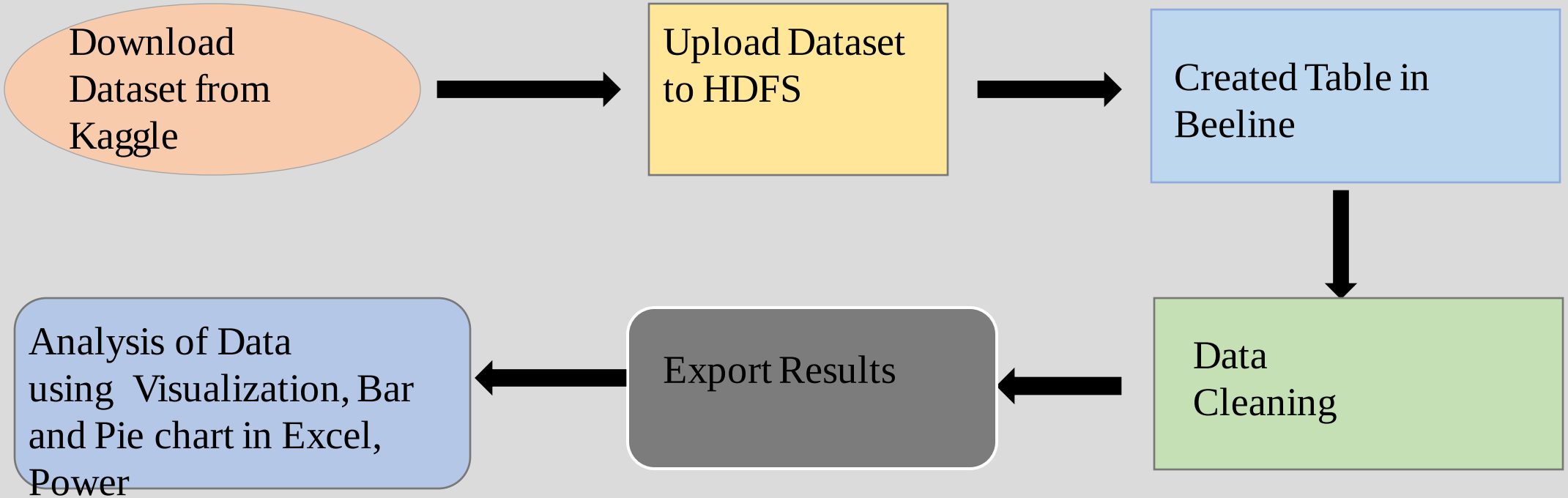
H/W EXPERIMENTAL SPECIFICATIONS

CLUSTER VERSION	HADOOP 3.1.2
Number of Nodes	5 Nodes (2 Master and 3 Worker nodes)
Memory Size	390.7 GB
CPU Speed	1995.309 MHz
Number of Core CPU	4

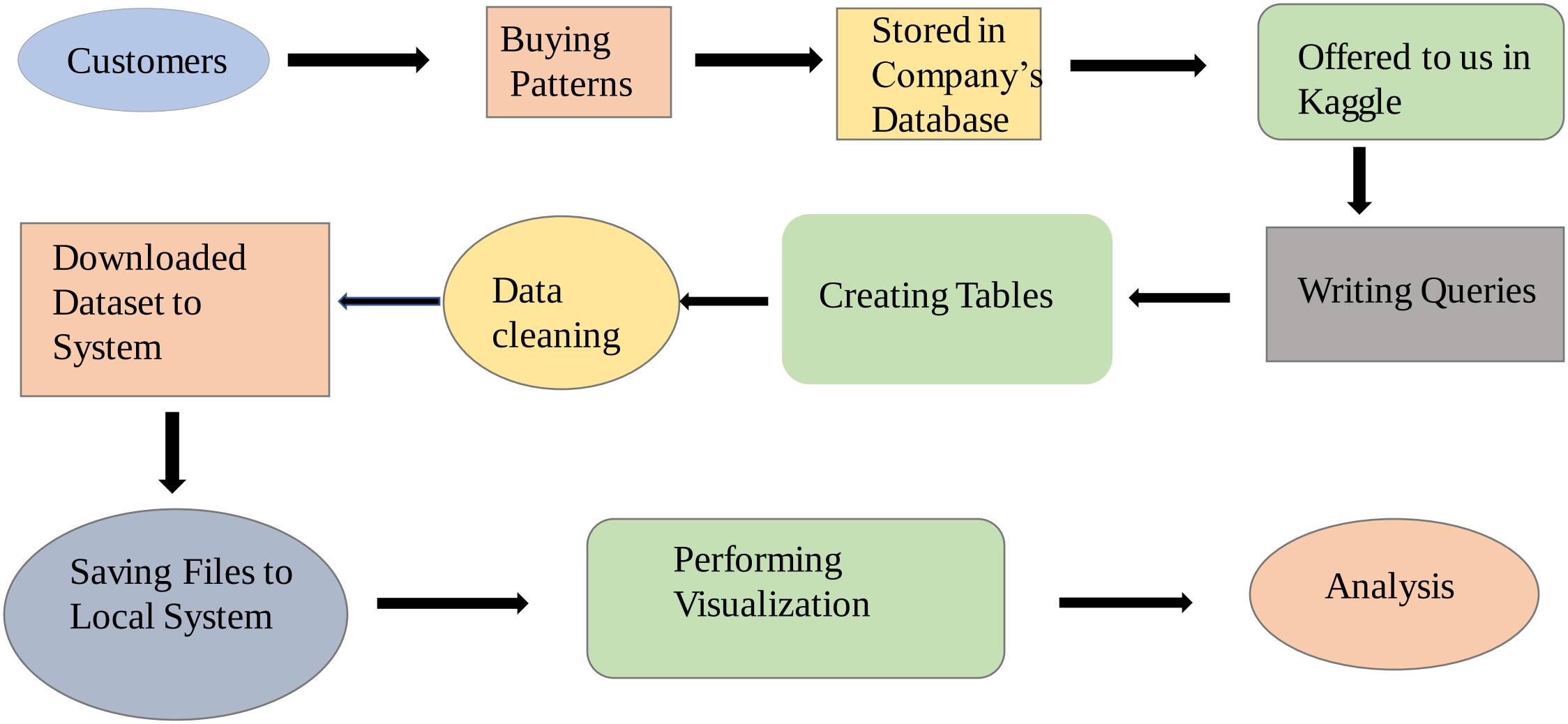
INTRODUCTION

- This dataset contains customer behavior data for October and November 2019. The size of the dataset is 5GB. There are 10 columns in this set, which are, as follows: time, type, product id, category id, category code, brand, price, user id, and user session.
- In this data set our goal is to analyze customer buying behavior. We have Data from 2 different months, hence we also compare the buying and viewing patterns of customers. We have chosen this dataset because it will help companies understand requirements of the users and help maximize their sales and production of demanding goods.

Workflow Chart



Architecture of Implementation Chart



AGENDAS

- Top 10 popular categories in October and November
- Top 10 Least popular categories in October and November
- Top 10 purchased categories and their sales count and average price in October and November.
- Top 10 popular brands October and November
- Top 10 Purchased Brands of October and November
- Top 10 Least Purchased Brands of October and November
- Views, Purchases, In-Carts in October and November
- Sum of Sales in both October and November
- Exit rate
- Top 5 hours with most purchases in November
- Top 5 days with most purchases in October
- Top 10 Users who made the most purchases in November

AGENDA-1

Top 10 Popular categories in October and November

- October

select category_code, count(category_code) as count from cleanedoctober group by category_code order by count(category_code) desc limit 10;

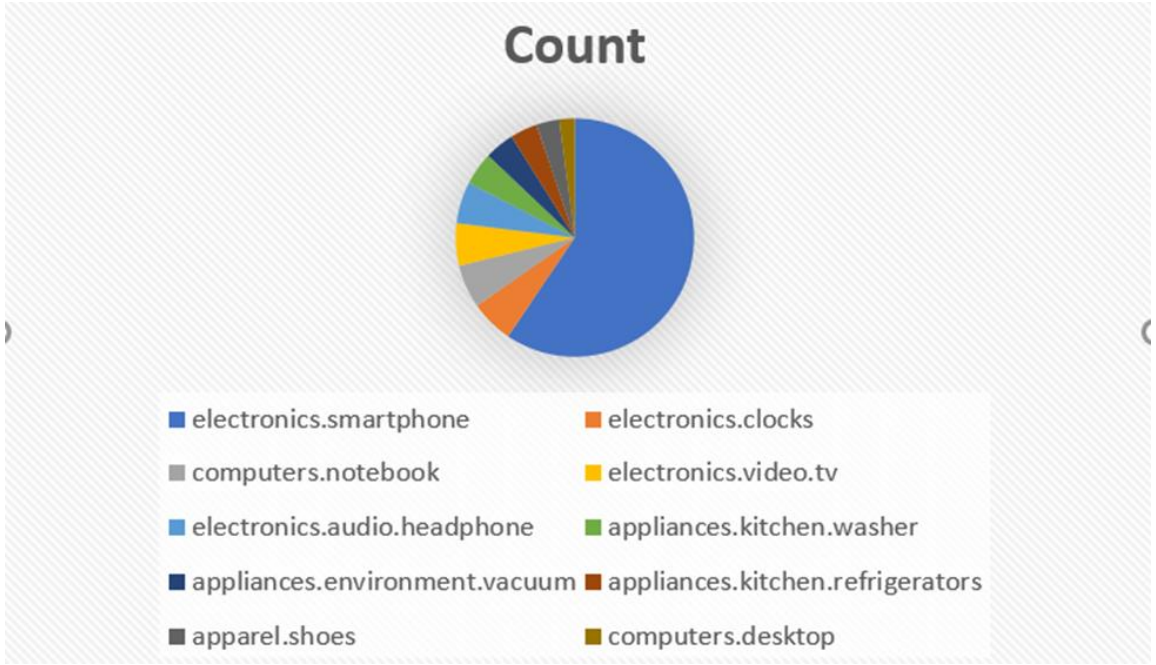
category_code	count
electronics.smartphone	11485320
electronics.clocks	1132207
computers.notebook	1131269
electronics.video.tv	1112047
electronics.audio.headphone	1092952
appliances.kitchen.washer	860417
appliances.environment.vacuum	778587
appliances.kitchen.refrigerators	712119
apparel.shoes	604625
computers.desktop	403070

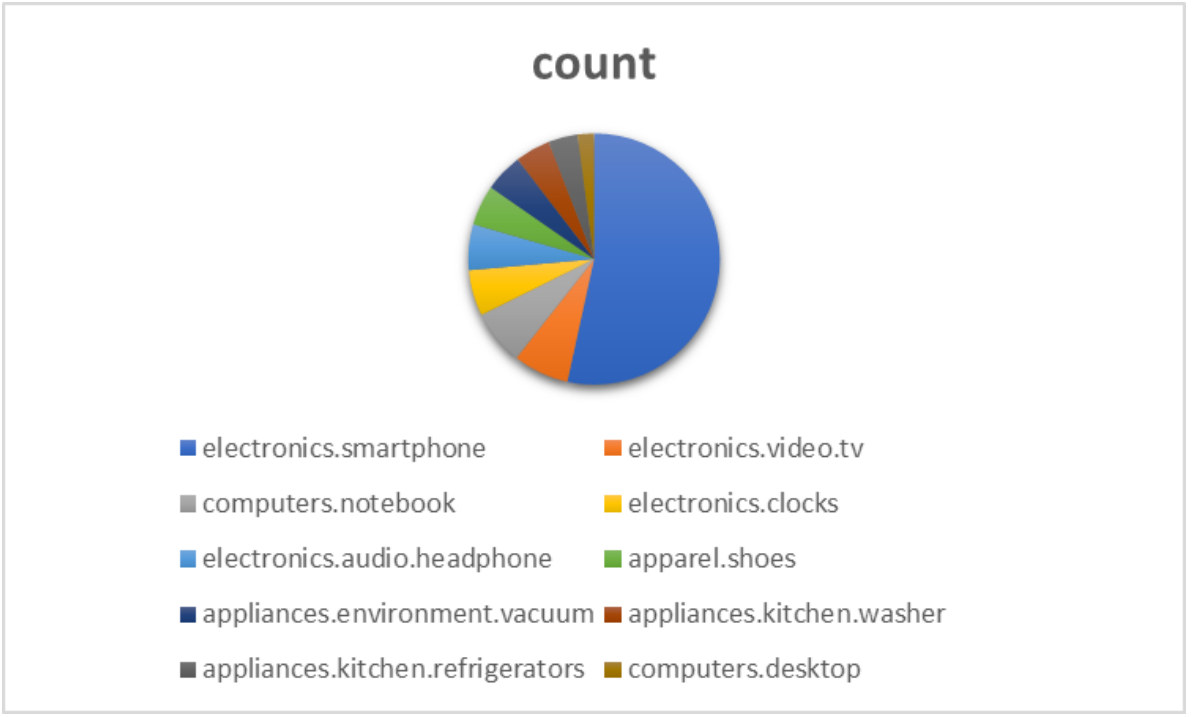
- November

select category_code, count(category_code) as count from cleanednovember group by category_code order by count(category_code) desc limit 10;

category_code	count
electronics.smartphone	16353579
electronics.video.tv	2195118
computers.notebook	2164657
electronics.clocks	1811325
electronics.audio.headphone	1803893
apparel.shoes	1587667
appliances.environment.vacuum	1510004
appliances.kitchen.washer	1389808
appliances.kitchen.refrigerators	1149533
computers.desktop	647867

Top 10 Popular categories in October and November





AGENDA-2

Top 10 Least popular categories in October and November

- October

select category_code, count(category_code) as count from cleanedoctober group by category_code order by count(category_code) limit 10;

category_code	count
country_yard.furniture.bench	190
construction.tools.soldering	201
auto.accessories.anti_freeze	296
apparel.belt	370
apparel.shorts	423
apparel.jacket	436
apparel.skirt	685
country_yard.furniture.hammok	1214
apparel.shoes.step_ins	1326
apparel.shoes.espadrilles	1398

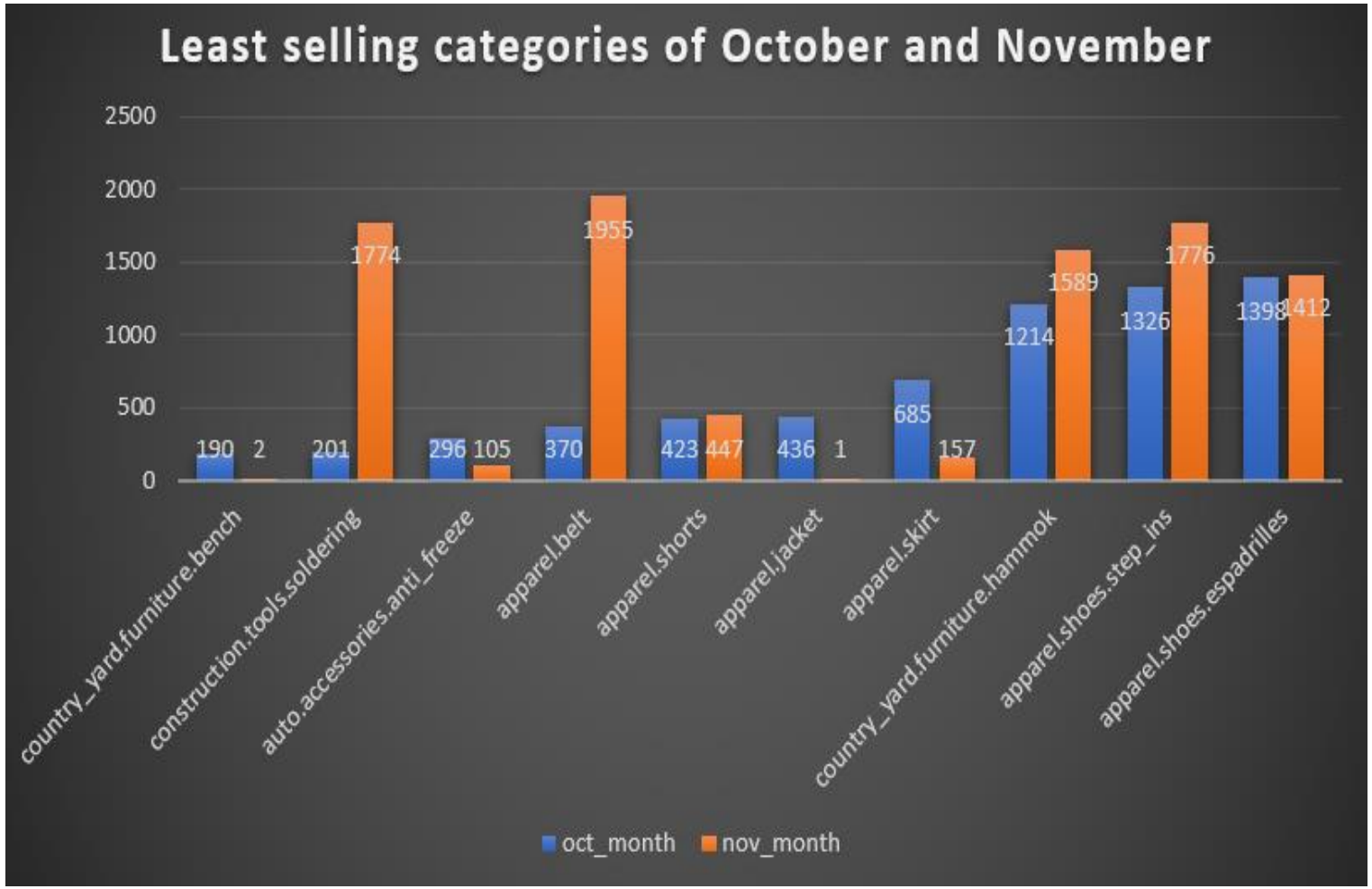
- November

select category_code, count(category_code) as count from cleanednovember group by category_code order by count(category_code) limit 10;

category_code	count
apparel.jacket	1
country_yard.furniture.bench	2
appliances.kitchen.fryer	105
construction.tools.screw	157
apparel.shorts	447
apparel.shoes.espadrilles	1412
country_yard.furniture.hammok	1589
construction.tools.soldering	1774
apparel.shoes.step_ins	1776
apparel.belt	1955

Bar Graph

Top 10 Least popular categories in October and November



Agenda -3
Top 10 purchased categories, sales count and average price in October and November.

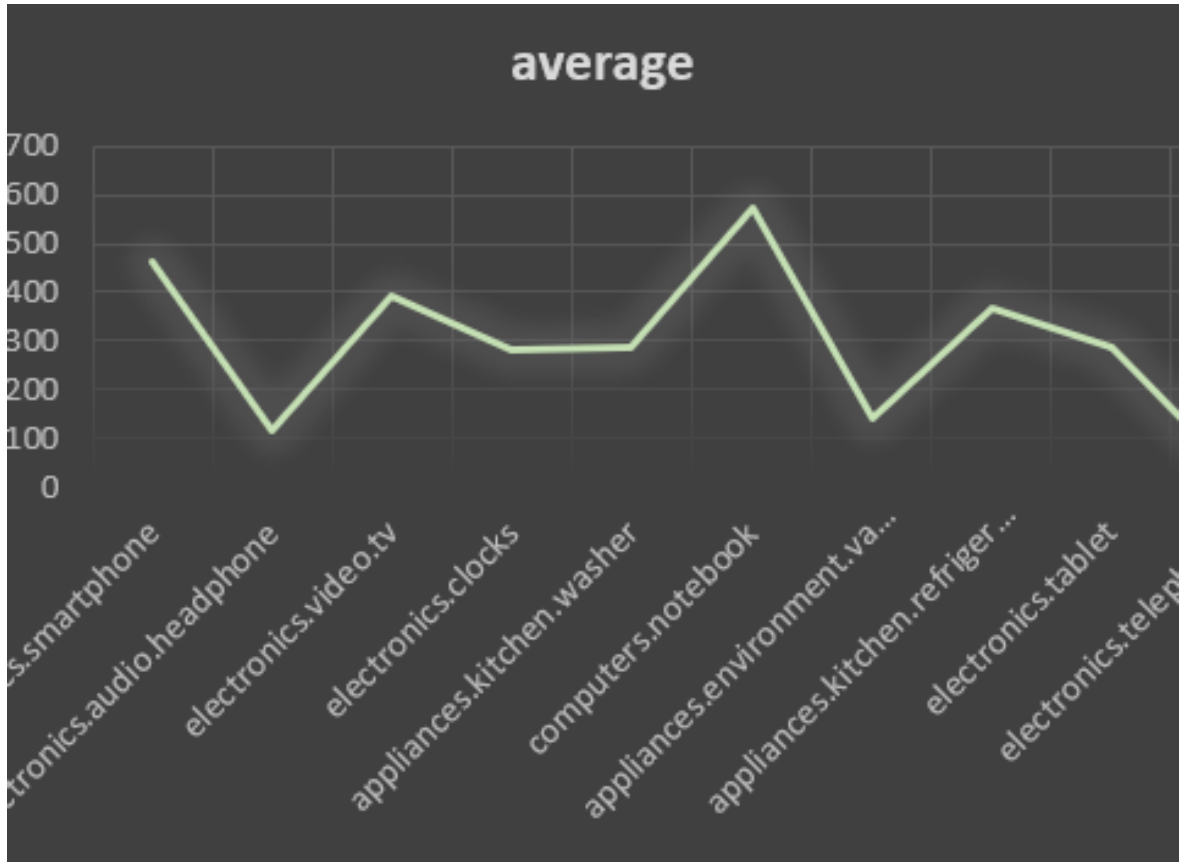
- select category_code as category_name, count(category_code) as count, cast(sum(price) as bigint) as sales, avg(price) as average_price from cleanedoctober where event_type like 'purchase' group by category_code order by count(category_code) desc limit 10;

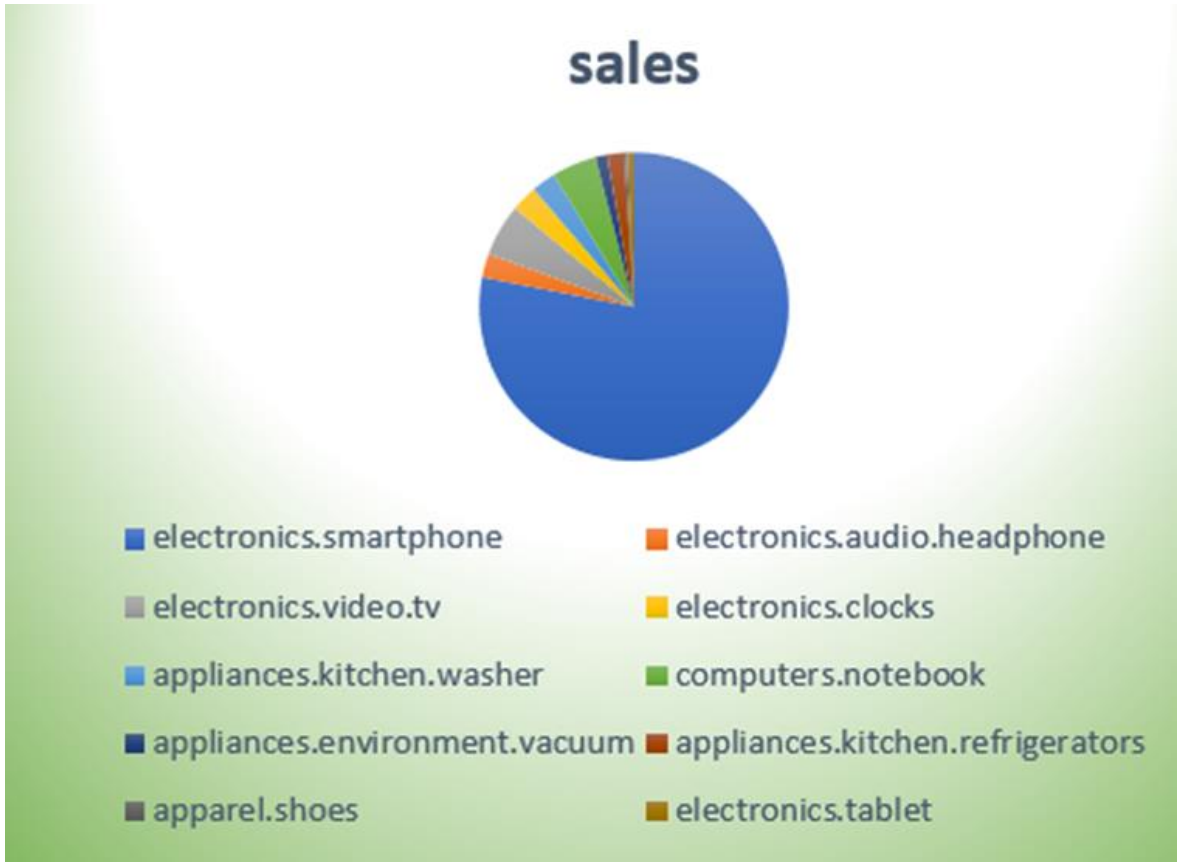
category_name	count	sales	average_price
electronics.smartphone	337575	156745645	464.32835944604443
electronics.audio.headphone	30439	3537007	116.19986727554131
electronics.video.tv	21548	8416411	390.5889845925363
electronics.clocks	16647	4648698	279.25141887427515
appliances.kitchen.washer	16059	4638860	288.86357120617663
computers.notebook	15547	8948500	575.5773165240855
appliances.environment.vacuum	12218	1708631	139.84539286298966
appliances.kitchen.refrigerators	8871	3268251	368.41970014654663
electronics.tablet	5599	1609957	287.5436881585982
electronics.telephone	3733	126609	33.91627645325482

- select category_code as category_name, count(category_code) as count, cast(sum(price) as bigint) as sales, avg(price) as average_price from cleanednovember where event_type like 'purchase' group by category_code order by count(category_code) desc limit 10;

category_name	count	sales	average_price
electronics.smartphone	382492	177747817	464.7098962070141
electronics.audio.headphone	40742	5664176	139.02548647588023
electronics.video.tv	30178	12430585	411.90886109085903
electronics.clocks	21426	6261585	292.24238168580564
appliances.kitchen.washer	19680	5786011	294.0046702235795
computers.notebook	18323	10614351	579.2911220869877
appliances.environment.vacuum	18122	2757834	152.18159143582253
appliances.kitchen.refrigerators	10420	4088907	392.4095969289827
apparel.shoes	8768	767080	87.4864016879559
electronics.tablet	6123	1519396	248.14576351461776

Top 10 purchased categories, sales count and average price in October and November





Agenda-4
Top 10 popular brands October and November

- October

select brand, count(brand) as count from cleanedoctober group by brand order by count(brand) desc limit 10;

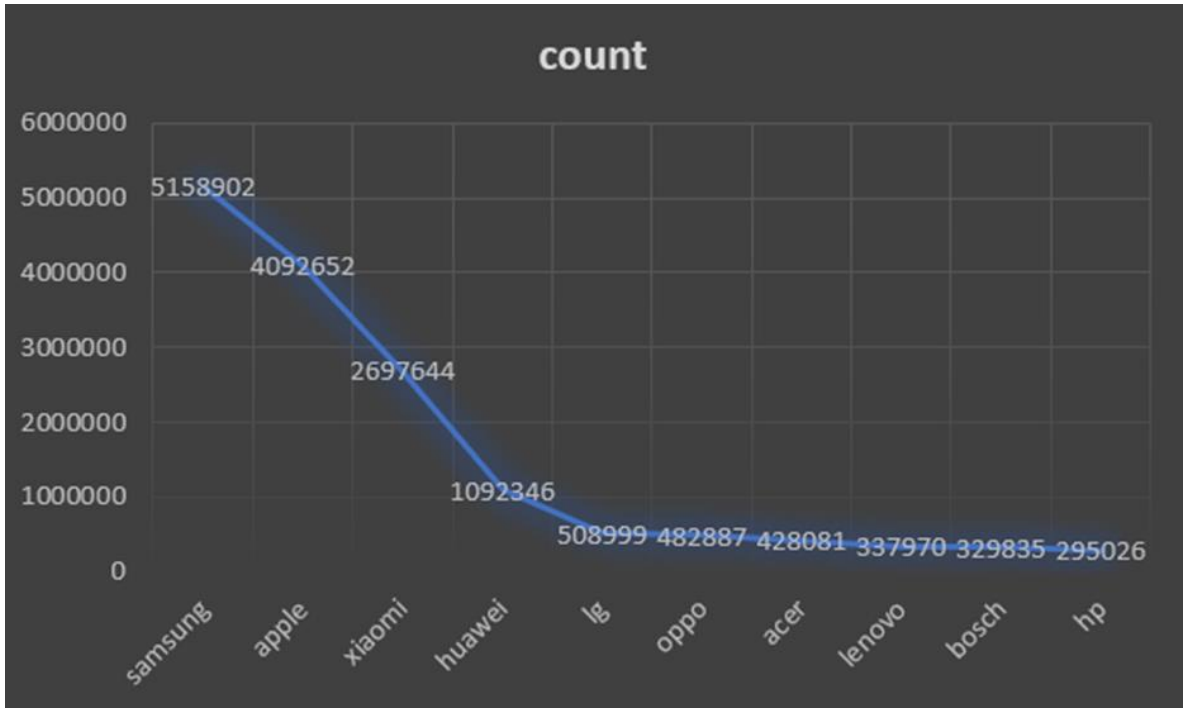
- November

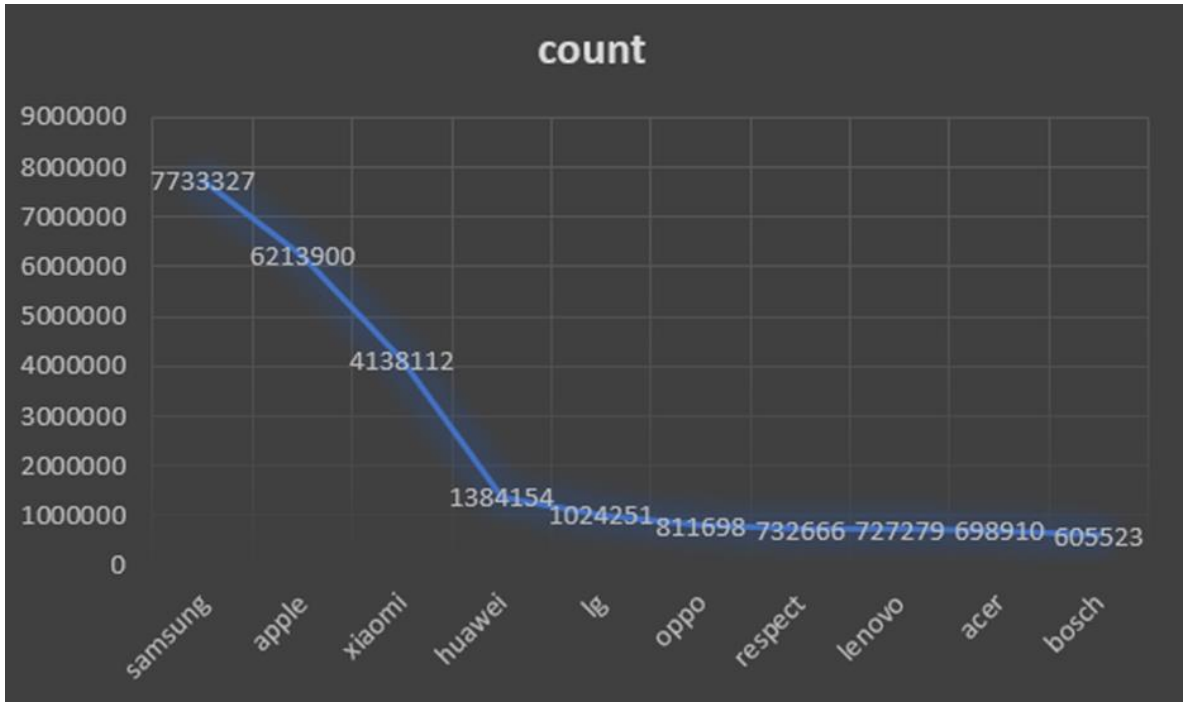
select brand, count(brand) as count from cleanednovember group by brand order by count(brand) desc limit 10;

brand	count
samsung	5158902
apple	4092652
xiaomi	2697644
huawei	1092346
lg	508999
oppo	482887
acer	428081
lenovo	337970
bosch	329835
hp	295026

brand	count
samsung	7733327
apple	6213900
xiaomi	4138112
huawei	1384154
lg	1024251
oppo	811698
respect	732666
lenovo	727279
acer	698910
bosch	605523

Top 10 popular brands October and November





Agenda-5
Top 10 Purchased Brands of October and November

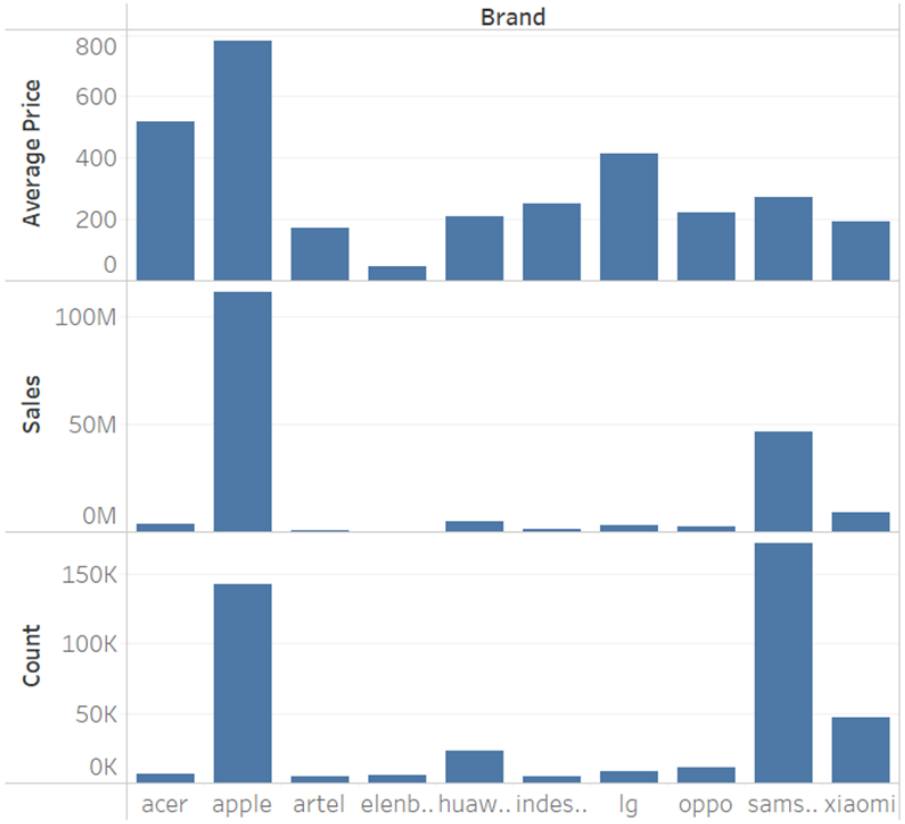
- select brand, count(brand) as count, cast(sum(price) as bigint) as sales, avg(price) as average_price from cleanedoctober where event_type like 'purchase' group by brand order by count(brand) desc limit 10;
- select brand, count(brand) as count, cast(sum(price) as bigint) as sales, avg(price) as average_price from cleanednovember where event_type like 'purchase' group by brand order by count(brand) desc limit 10;

brand	count	sales	average_price
samsung	171706	46350825	269.9429601761183
apple	142577	111189822	779.8580576811813
xiaomi	46595	8869391	190.35071702971942
huawei	23294	4872029	209.15384219112144
oppo	10891	2412959	221.55539068956136
lg	7831	3225784	411.92498276081864
acer	6882	3576719	519.720941586754
elenberg	5435	244570	44.99914075437048
indesit	5023	1249809	248.81727652797156
artel	4717	807799	171.25283230866924

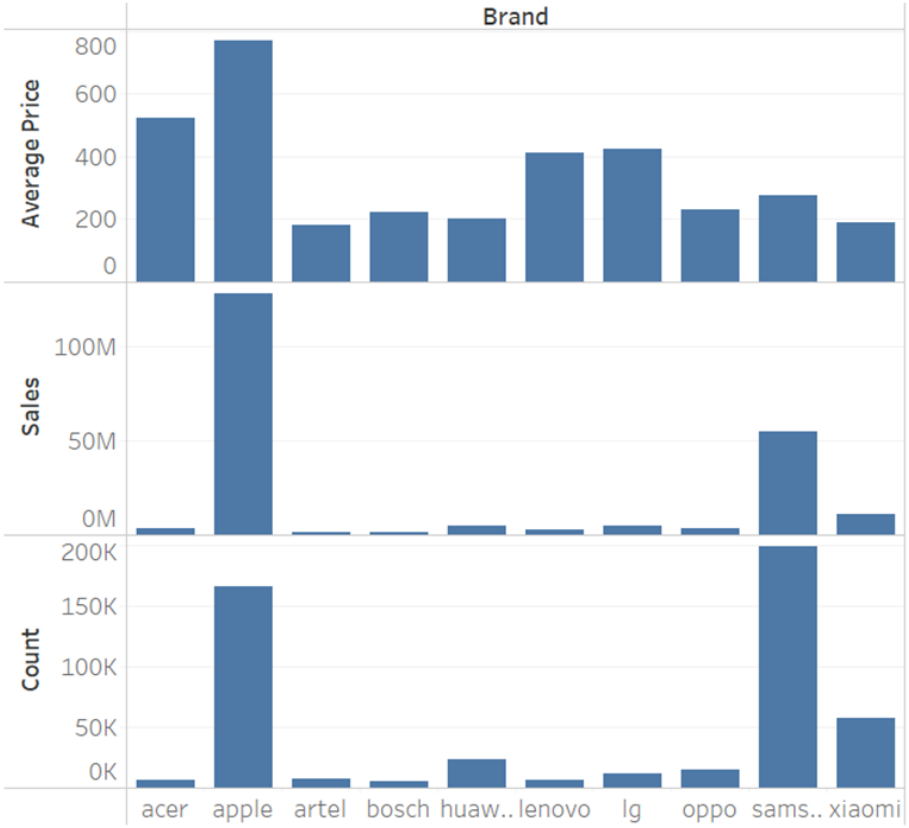
brand	count	sales	average_price
samsung	198670	54790697	275.78747470683527
apple	165681	127490496	769.4937659116308
xiaomi	57909	10874049	187.7782249736615
huawei	23466	4768995	203.23002769965083
oppo	15080	3488540	231.3355941644597
lg	11828	5029641	425.2317923571167
artel	7269	1329815	182.94340074288164
lenovo	6546	2698104	412.17599450045907
acer	6402	3347306	522.8532536707261
bosch	5718	1276557	223.25236271423637

Top 10 Purchased Brands of October and November

Top Selling Brands, Total Sales and Average Price of October

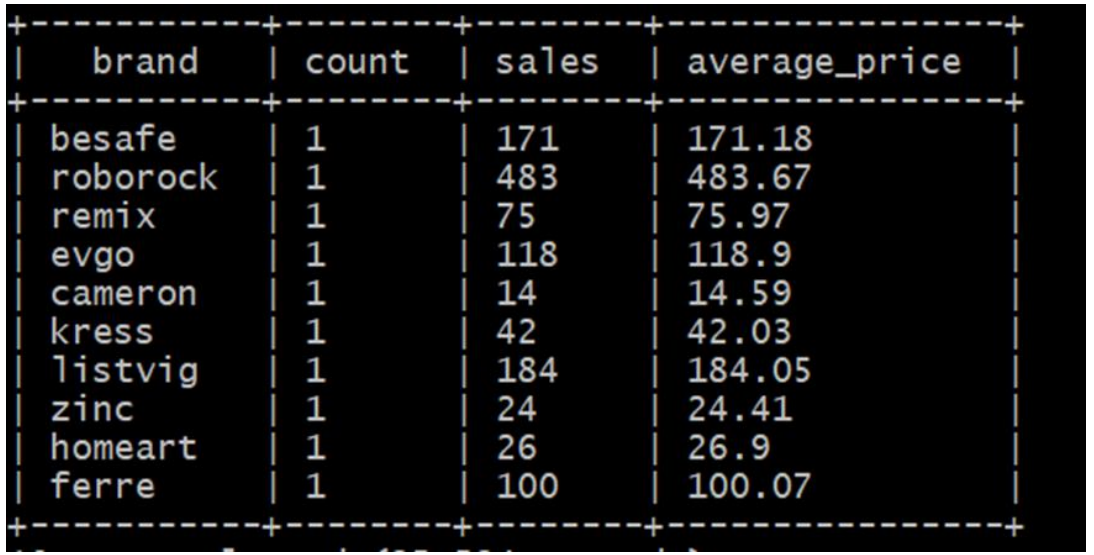


Top Selling Brands, Total sales and Average Price of November

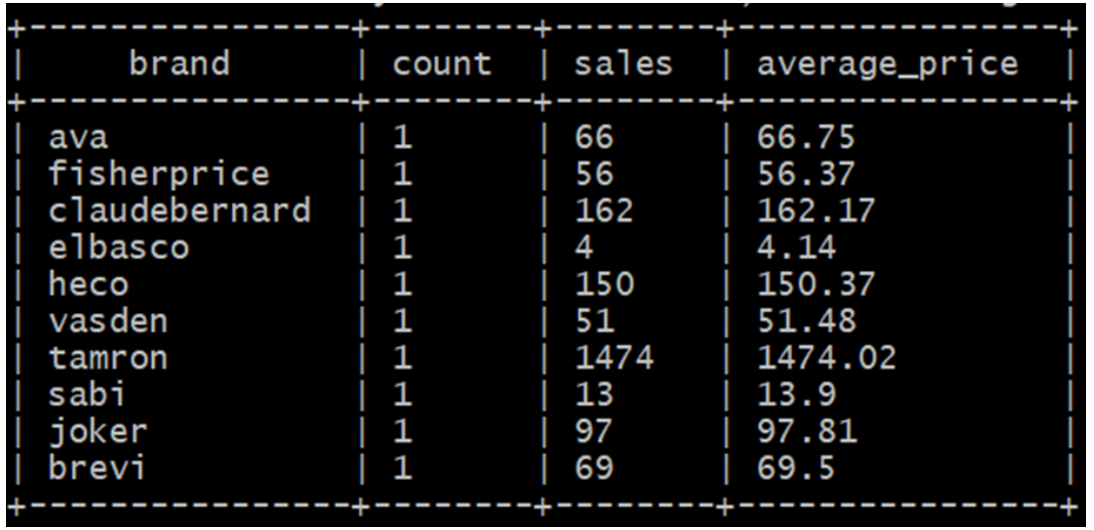


Agenda-6
Top 10 Least Purchased Brands of October and November

- select brand, count(brand) as count, cast(sum(price) as bigint) as sales, avg(price) as average_price from cleanedoctober where event_type like 'purchase' group by brand order by count(brand) limit 10;
- select brand, count(brand) as count, cast(sum(price) as bigint) as sales, avg(price) as average_price from cleanednovember where event_type like 'purchase' group by brand order by count(brand) limit 10;

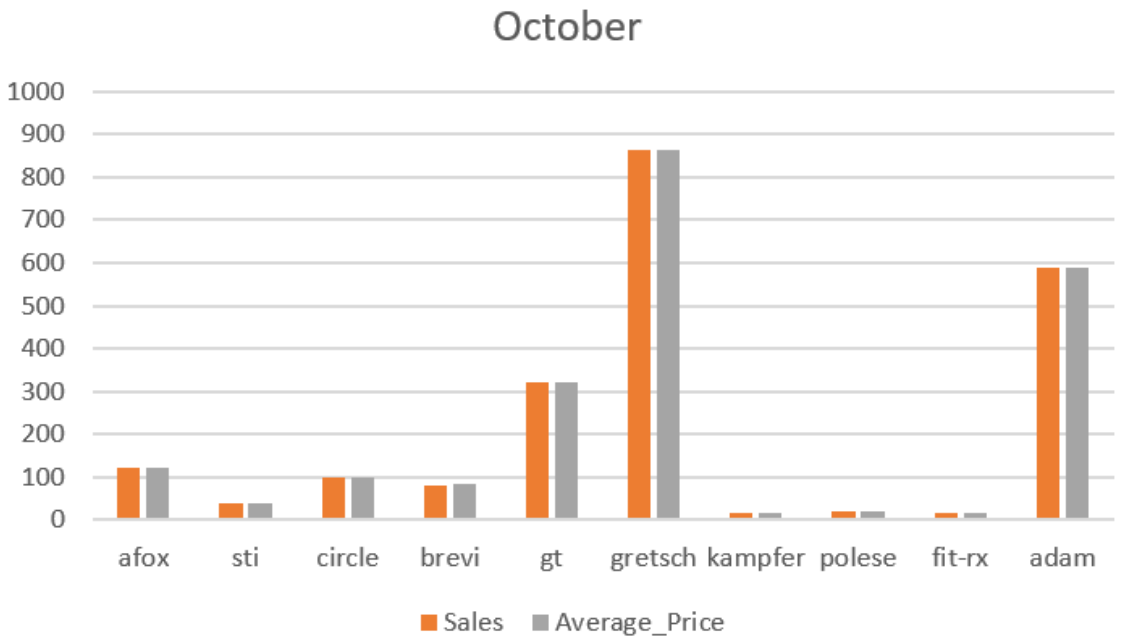


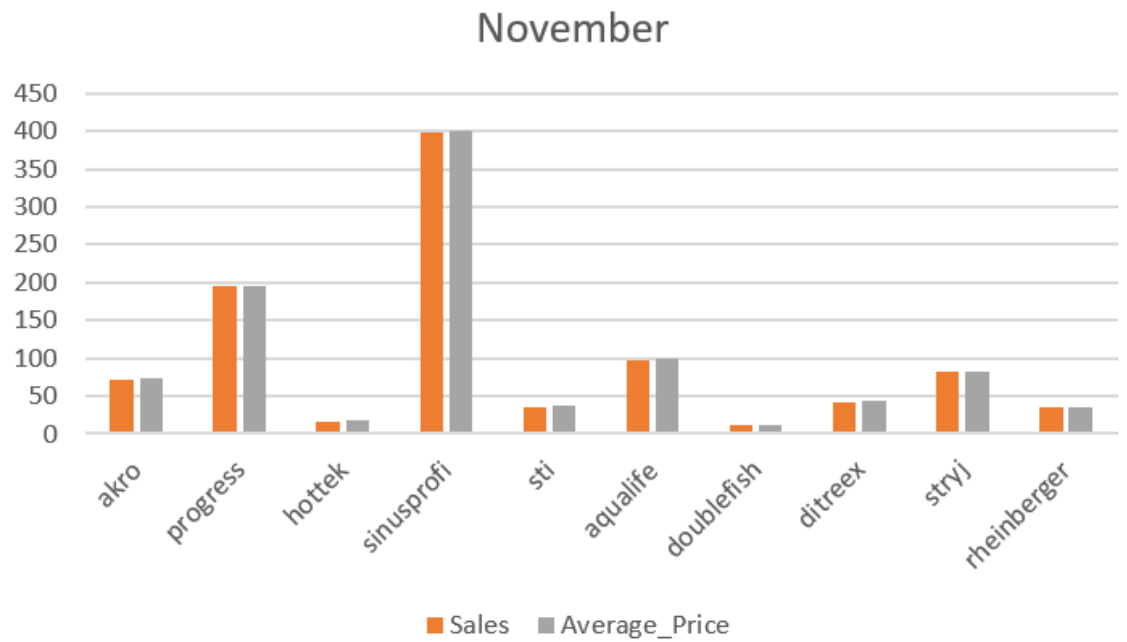
brand	count	sales	average_price
besafe	1	171	171.18
roborock	1	483	483.67
remix	1	75	75.97
evgo	1	118	118.9
cameron	1	14	14.59
kress	1	42	42.03
listvig	1	184	184.05
zinc	1	24	24.41
homeart	1	26	26.9
ferre	1	100	100.07



brand	count	sales	average_price
ava	1	66	66.75
fisherprice	1	56	56.37
claudebernard	1	162	162.17
elbasco	1	4	4.14
heco	1	150	150.37
vasden	1	51	51.48
tamron	1	1474	1474.02
sabi	1	13	13.9
joker	1	97	97.81
brevi	1	69	69.5

Top 10 Least Purchased Brands of October and November





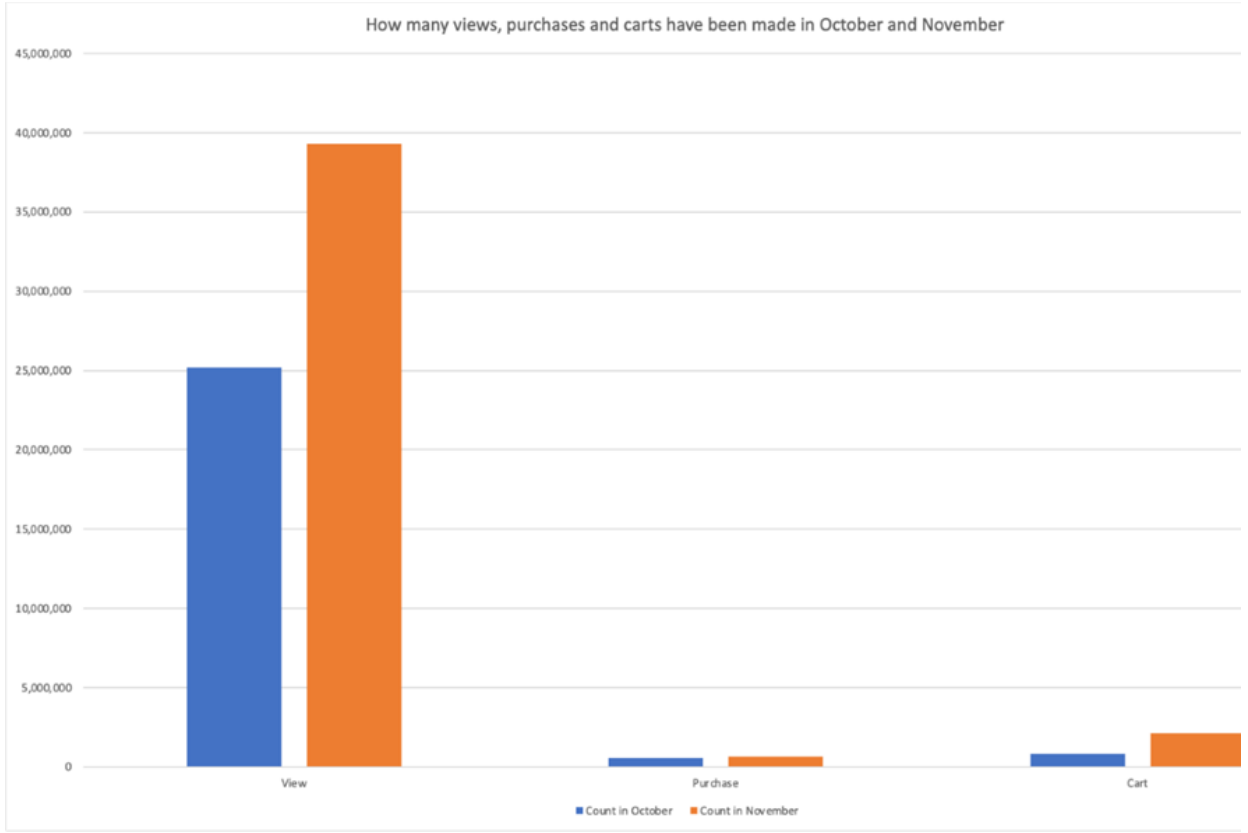
Agenda-7
Views, Purchases, In-Carts in October and November

- select event_type, count(event_type) as count from cleanedoctober group by event_type;
- select event_type, count(event_type) as count from cleanednovember group by event_type;

event_type	count
view	25201706
purchase	549507
cart	809407

event_type	count
view	39315226
cart	2115082
purchase	659256

Views, Purchases, In-Carts in October and November



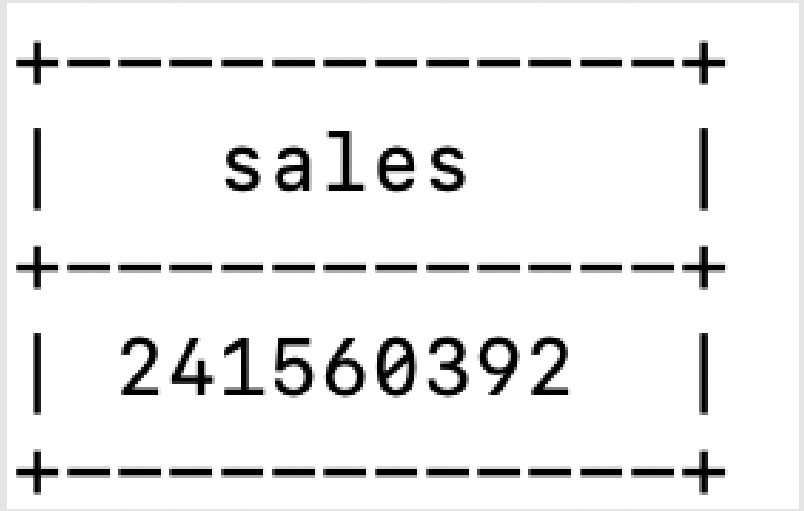
AGENDA-8
Sum of Sales in both October and November

October

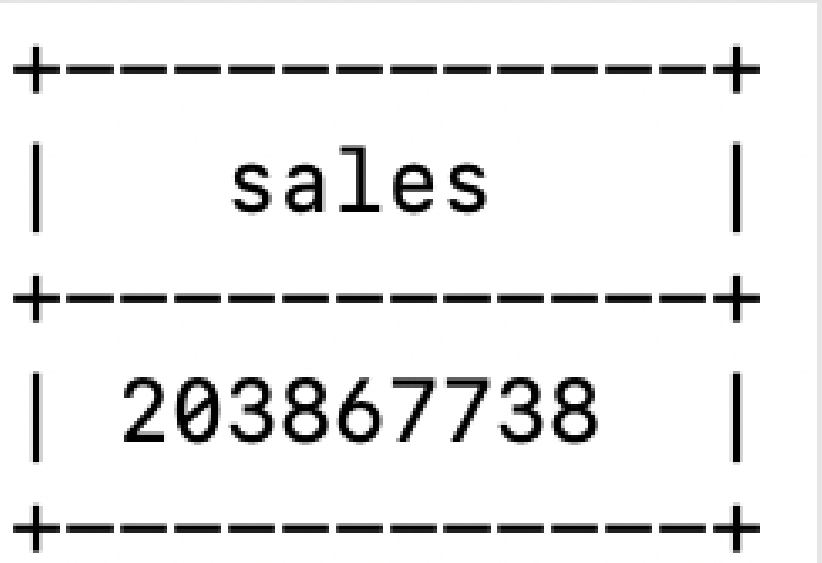
```
select cast(sum(price) as bigint) as sales from
cleanedoctober where event_type like 'purchase';
```

November

```
select cast(sum(price) as bigint) as sales from
cleanednovember where event_type like 'purchase';
```

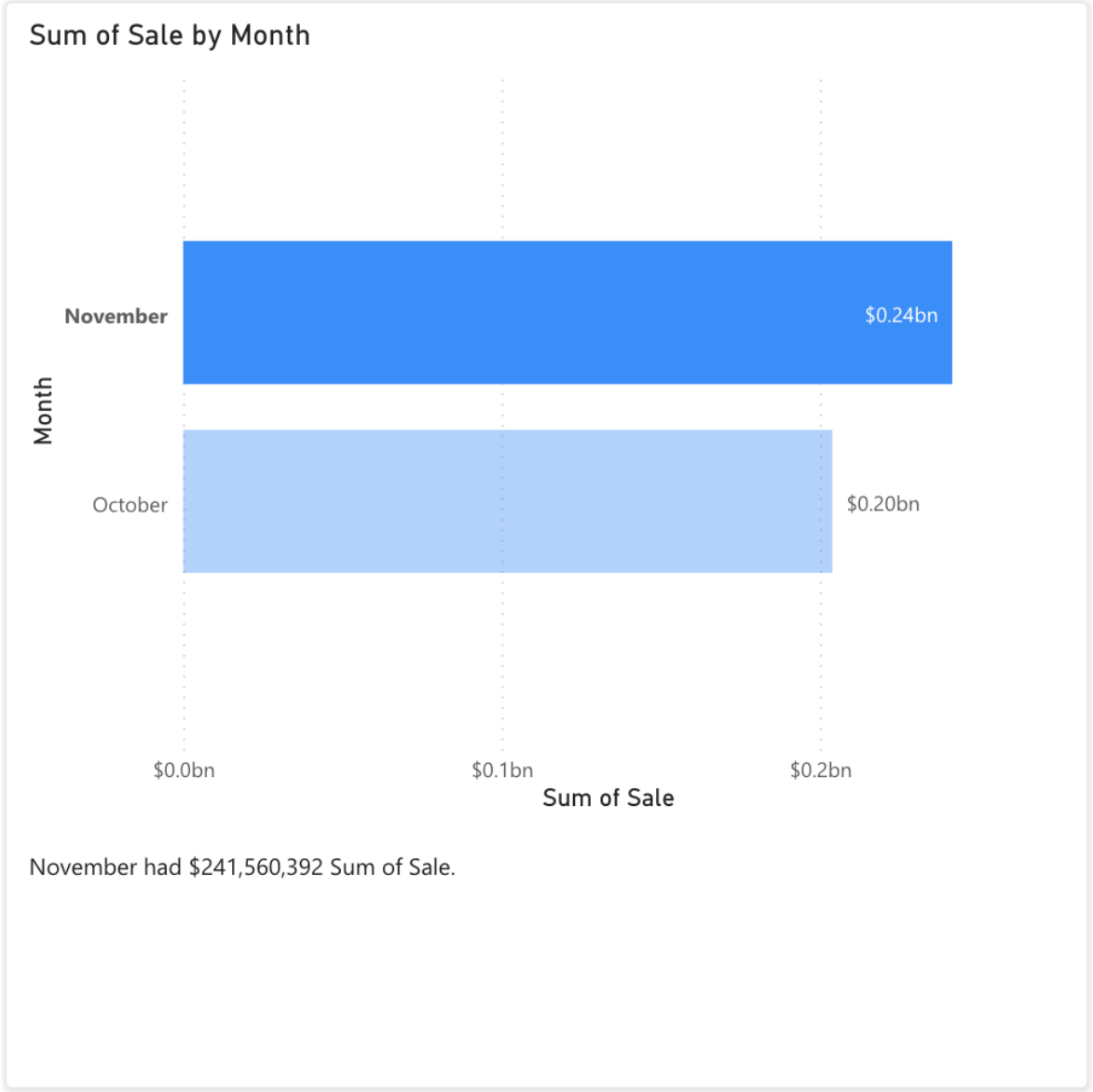


```
+-------------+
|    sales    |
+-------------+
| 241560392   |
+-------------+
```



```
+-------------+
|    sales    |
+-------------+
| 203867738   |
+-------------+
```

Sum of Sales in October and November

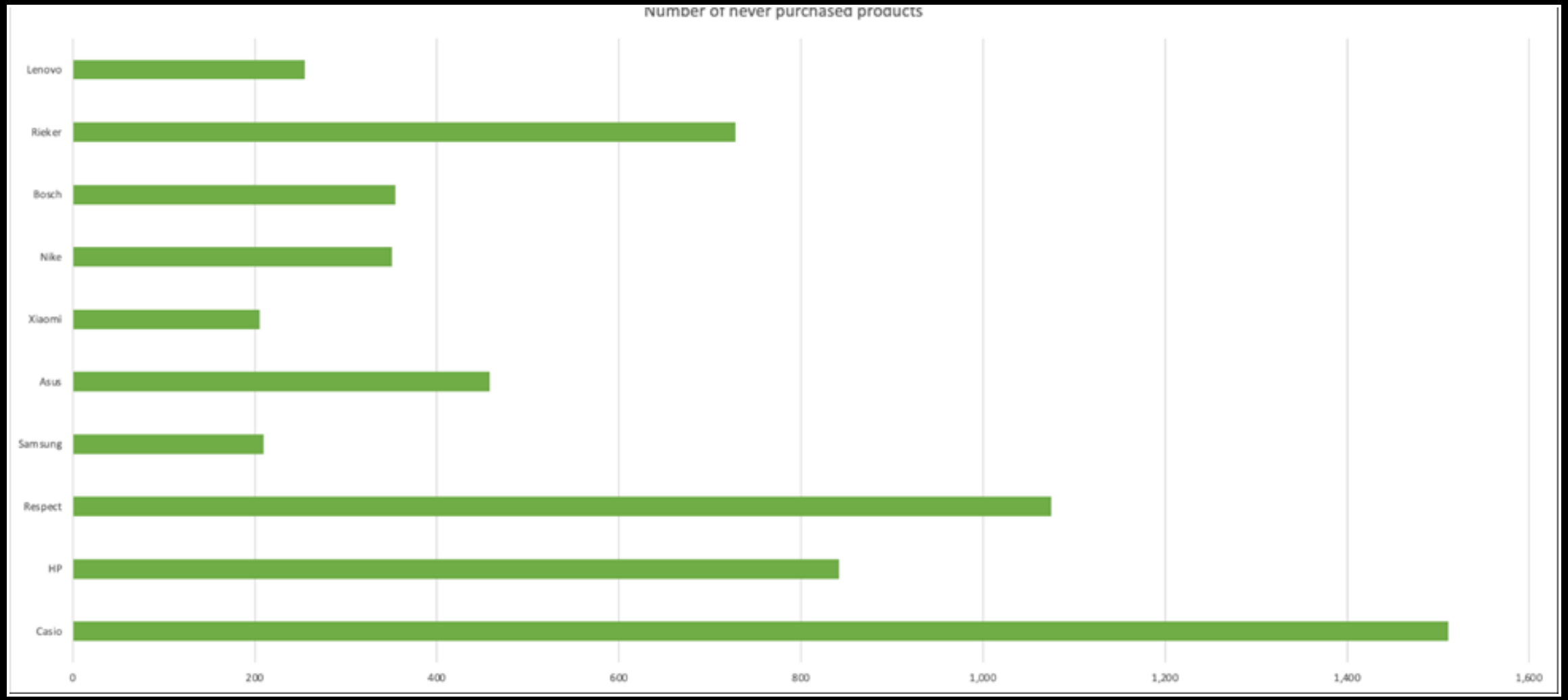


November had $241,560,392 Sum of Sale.

Agenda-9
Exit Rate - Most viewed brand but not purchased

```
select brand, count(distinct
product_id) as count from
cleanedoctober where event_type
= 'view'  and product_id NOT IN
(select product_id from
cleanedoctober where event_type =
'purchase') group by brand order
by count(product_id) desc limit 10;
```

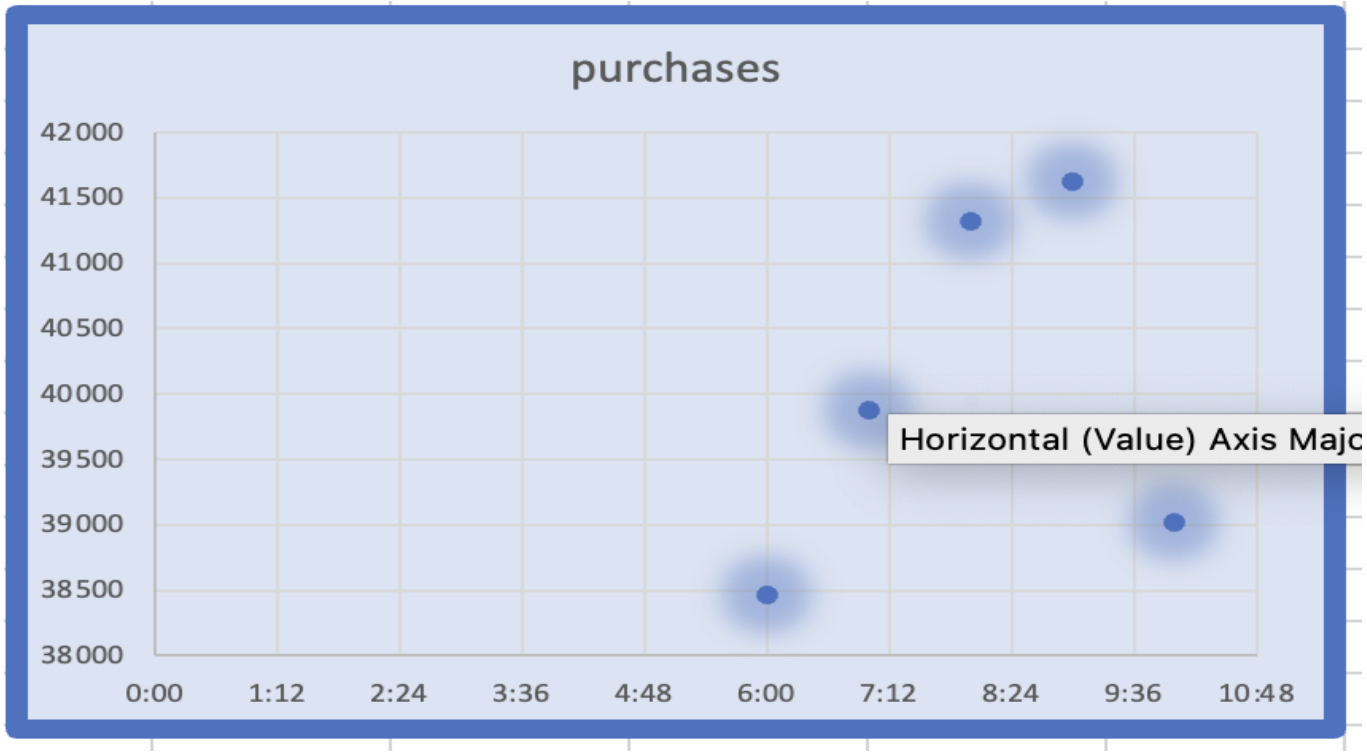
brand	count
casio	1511
hp	842
respect	1075
samsung	210
asus	458
xiaomi	205
nike	351
bosch	354
rieker	728
lenovo	255



Exit Rate - most viewed brand but not purchased

Agenda -10
Top 5 hours with most purchases in November

```
Select substr(event_time, 12, 2) as hour,
count(substr(event_time, 12, 2)) as count from
cleanednovember where event_type like 'purchase' group by
substr(event_time, 12, 2) order by count(substr(event_time, 12,
2)) desc limit 5;
```



hour	count
09	41622
08	41325
07	39874
10	39015
06	38467

Agenda-11
Top 5 days with most purchases in October

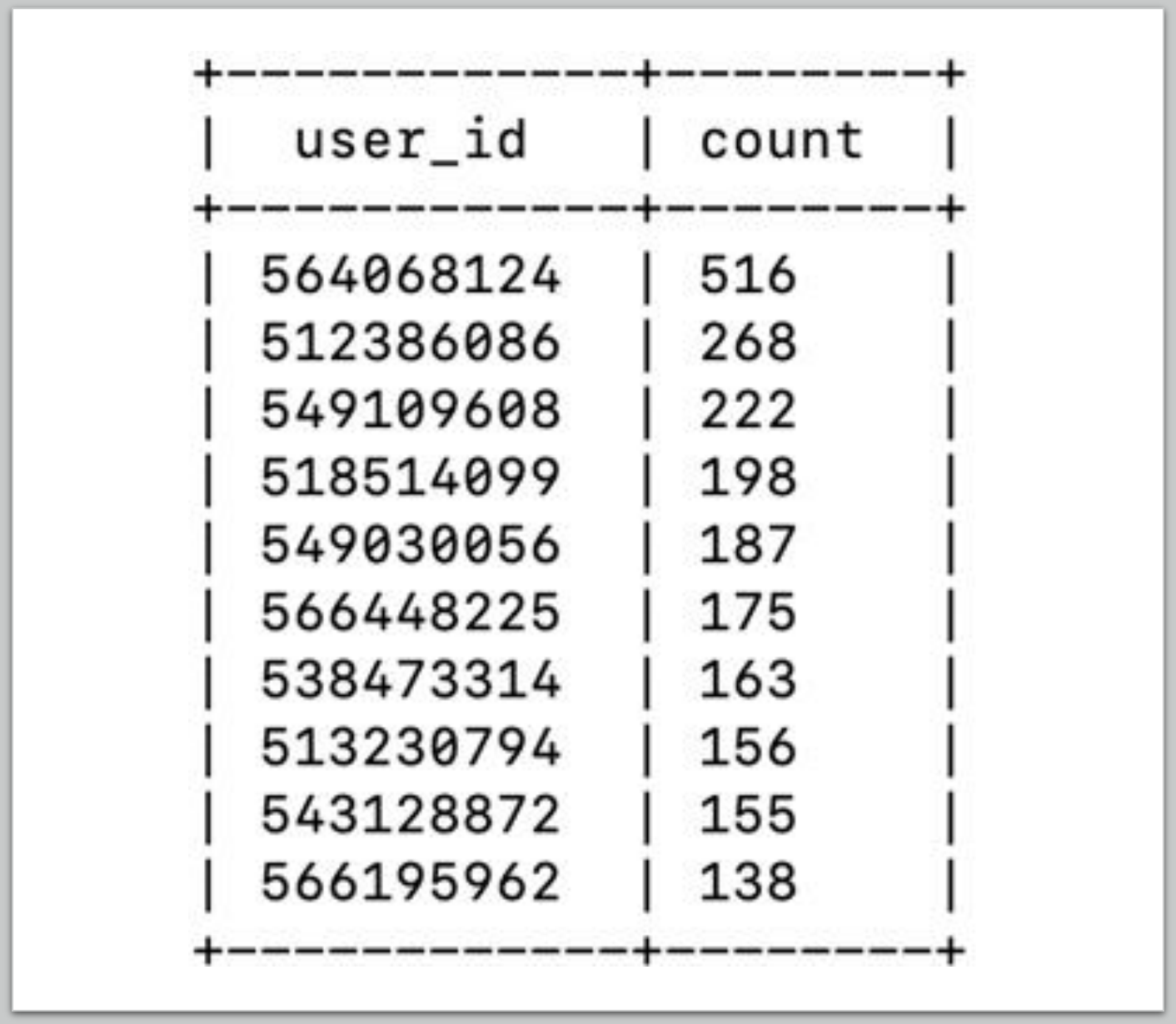
```
Select substr(event_time, 9, 2) as
day, count(substr(event_time, 9,
2)) as count from cleanedoctober
where event_type =
'purchase' group
by substr(event_time, 9, 2) order
by count(substr(event_time, 9,
2)) desc limit 5;
```

day	count
16	23976
14	22044
17	21324
13	20468
04	20455

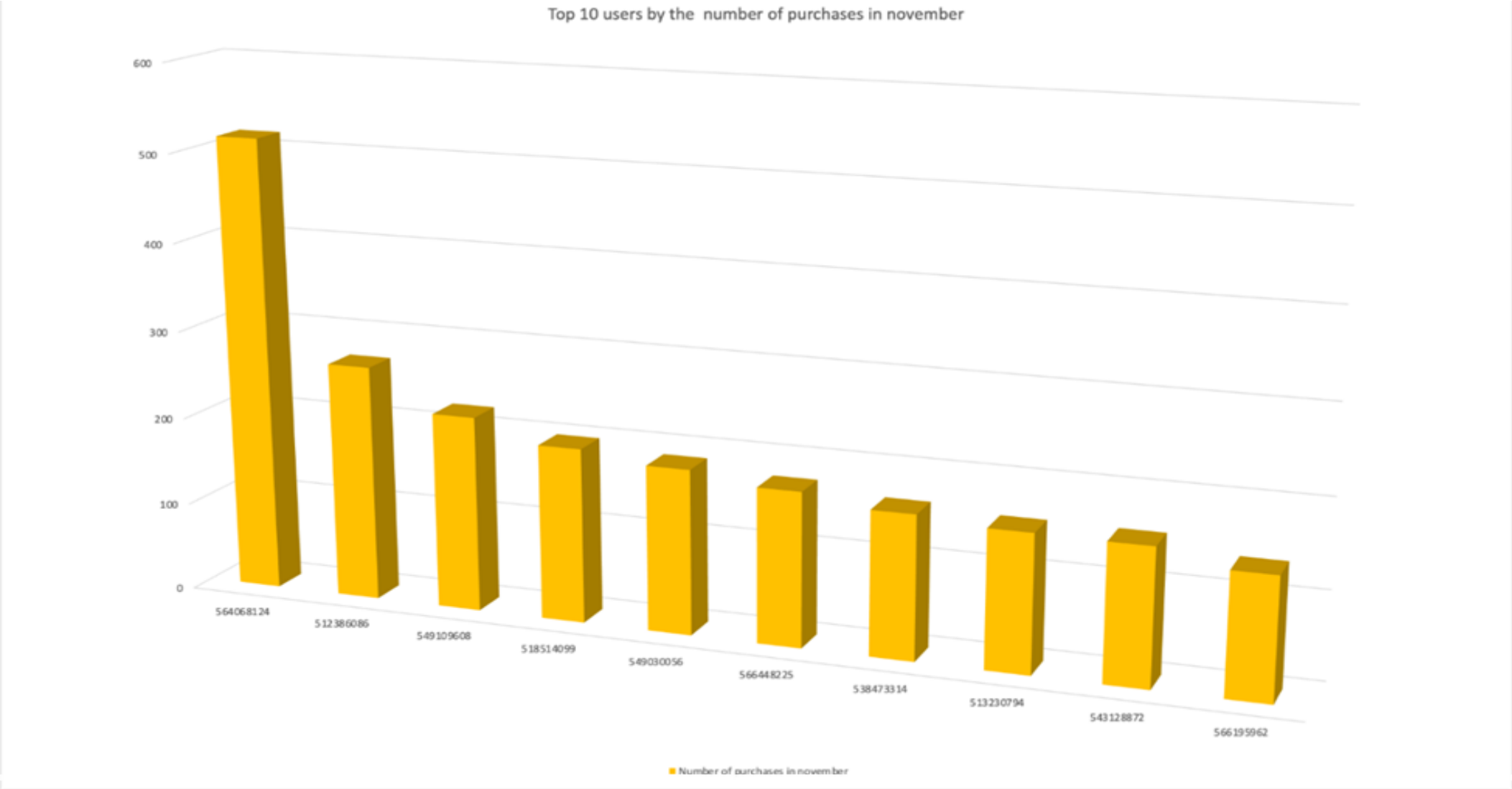
Top 5 days with most purchases in October5 days with most purchases in October

Agenda-12
Top 10 Users who made the most purchases in November

select user_id, count(user_id) as count from cleanednovember where event_type = 'purchase' group by user_id order by count(user_id) limit 10;



user_id	count
564068124	516
512386086	268
549109608	222
518514099	198
549030056	187
566448225	175
538473314	163
513230794	156
543128872	155
566195962	138



Top 10 Users who made the most purchases in November

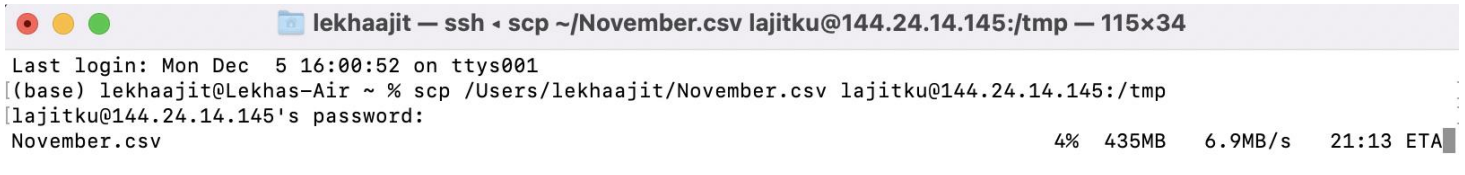
CHALENGES FACED

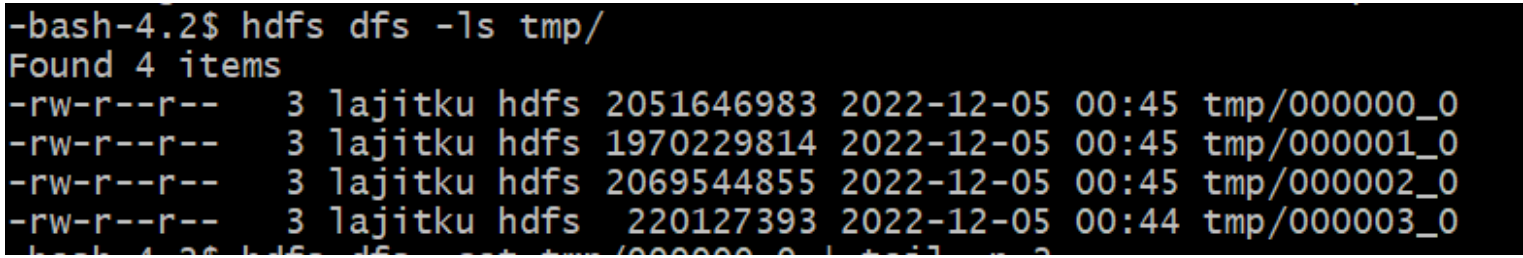
Data Uploading and Downloading

- Time consuming cause data set was large
- Data got downloaded in multiple files for our data set due to its large file size

Running a query from HDFS

- Execution time for a query to fetch data from HDFS was more than 2-8 hrs.





Conclusion

From all the above work we can conclude the following:

- Smartphones are the most popular & purchased category in both the Months.
- Furniture Bench and Jackets are the least purchased categories in October and November, respectively.
- Samsung is the most popular & purchased brand of October and November.
- Besafe and Ava are the least purchased brands in October and November respectively.
- Users viewed, added the Products in Cart and Purchased mostly in November than in October.
- Sales in November are more than Sales in October.
- Most Viewed but not purchased brand is Casio.
- Most of the purchases happened around 9’o clock.
- Most of the purchases happened around 16th of October.
- User id – 564068124 has made most of the Purchases.

THANK YOU