CIS5200 Term Project Tutorial

Authors: Lekha Ajitkumar, Sushmitha Dandu, Dauren Omarov, Navyasree Sriramoju

Instructor: Jongwook Woo

Date: 07/12/2022

Lab Tutorial

Lekha Ajitkumar (lajitku@calstatela.edu)

Sushmitha Dandu (sdandu@calstatela.edu)

Dauren Omarov (domarov@calstatela.edu)

Navyasree Sriramoju (nsriram@calstatela.edu)

07/12/2022

Ecommerce Behavior Data from Multi Category Store

Objectives

In this hands-on lab, you will learn how to:

- Download dataset from the Kaggle website
- Using SCP upload the data to the Hadoop cluster
- Create Hive tables in HDFS using HiveQL
- Create HiveQL queries to manipulate and analyze the data
- Visualize the result in Excel, Power BI and Tableau

Platform Spec

- Cluster Version: Hadoop 3.1.2
- CPU Speed: 1995.309 MHz
- # of CPU cores: 4
- # of nodes: 3
- Total Memory Size: 390.7 GB

Dataset Details

- DATASET NAME: Ecommerce Behavior Data from Multi Category Store

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

- TOTAL SIZE: 15.83 GB

- MONTHS CONSIDERED: October and November

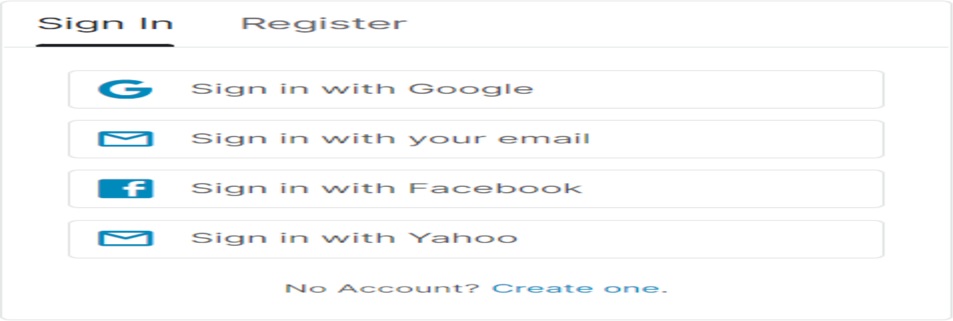
- NUMBER OF FILES: 2

- FILE FORMAT: CSV

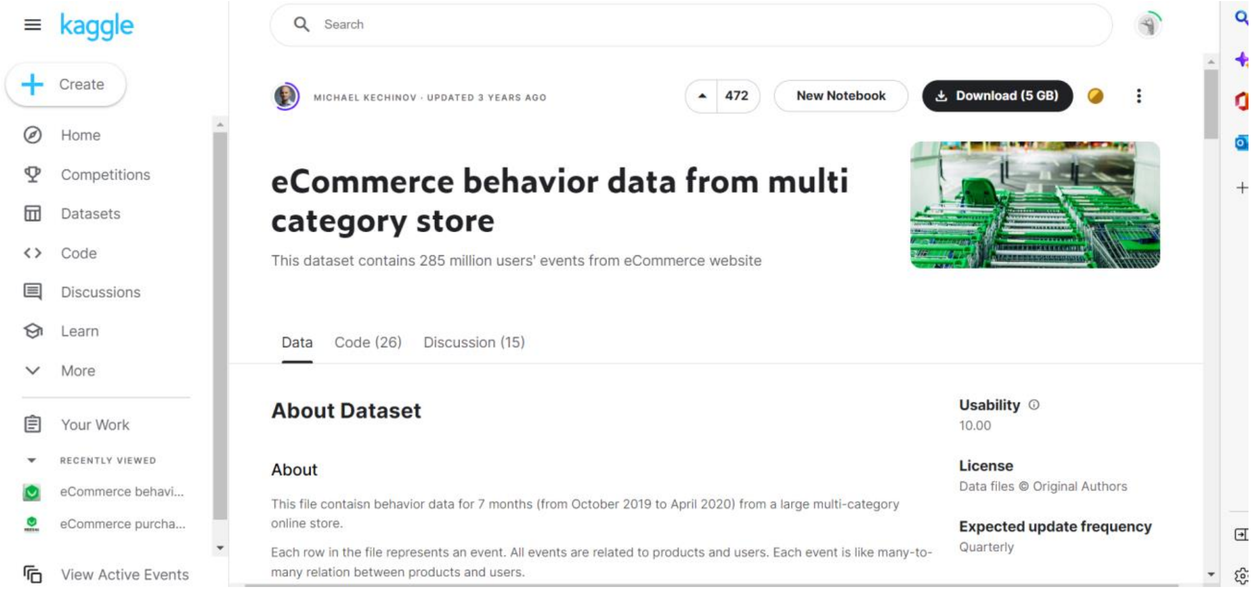
Step 1: Download the Dataset

This step is to get data manually. You need to remotely access your Oracle Cloud Big Data Compute Editions that you executed in your Oracle Cloud account using ssh using the information - ip address and connect command in beeline CLI

Ecommerce Behavior Data from Multi Category Store Dataset - Download Dataset to local machine from Kaggle Website, Sign in to Kaggle with any of the following Options.



Scroll down until you find the 2 csv files on right side.



eCommerce behavior data from multi category store

This dataset contains 285 million users' events from eCommerce website

About Dataset

About

This file contaisn behavior data for 7 months (from October 2019 to April 2020) from a large multi-category online store.

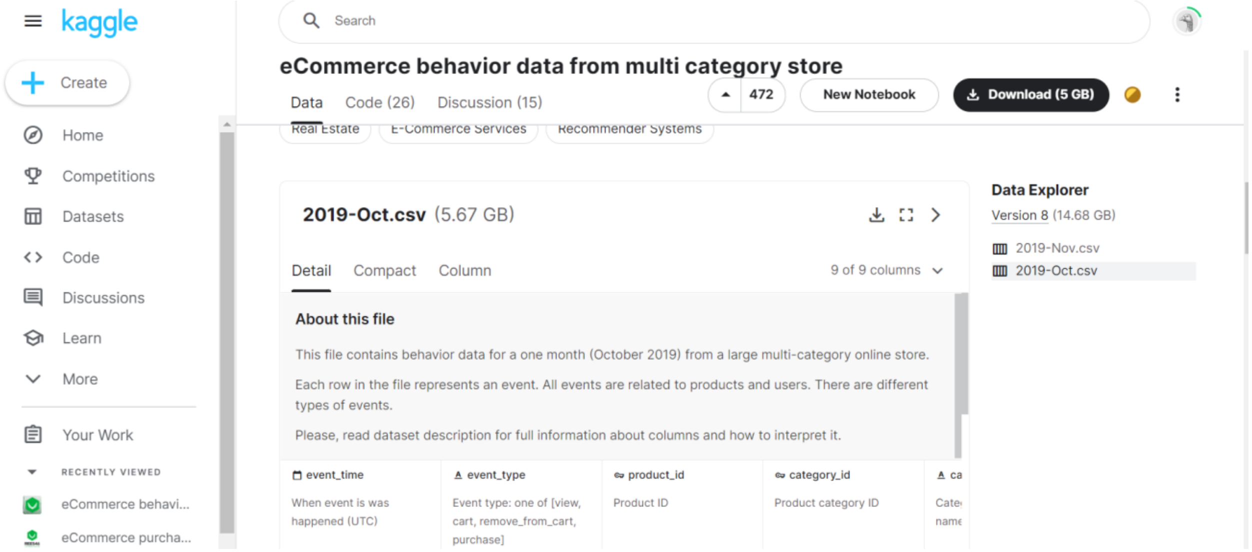
Each row in the file represents an event. All events are related to products and users. Each event is like many-to-many relation between products and users.

Usability
10.00

License
Data files © Original Authors

Expected update frequency
Quarterly

Download 2019-Nov.csv and 2019-Oct.csv, You will see Two zip files in downloads of your Personal Computer



eCommerce behavior data from multi category store

2019-Oct.csv (5.67 GB)

About this file

This file contains behavior data for a one month (October 2019) from a large multi-category online store.

Each row in the file represents an event. All events are related to products and users. There are different types of events.

Please, read dataset description for full information about columns and how to interpret it.

event_time	event_type	product_id	category_id	ca
When event is was happened (UTC)	Event type: one of [view, cart, remove_from_cart, purchase]	Product ID	Product category ID	Cate name

Extract the Zip files then you can find 2 csv files of October & November which should be uploaded in HDFS.

Step 2: Upload Files to Hadoop File System (HDFS)

Using SCP:

Open a command prompt session and from the directory of the extracted files in the previous step and perform the following commands:

scp /Users/lekhaajit/November.csv lajitku@144.24.14.145:/tmp
scp /Users/lekhaajit//October.csv lajitku@144.24.14.145:/tmp

Note: Use your own userid and server ip address.

Connect to server provided by the instructor.

You need to remotely access your server provided by the instructor using ssh. Your CalStateLA username(lajitku) should be a username/password to connect to the Hadoop cluster as follows:

Note: Do not forget to change lajitku with your username.

ssh lajitku@144.24.14.145

Create Directories and transfer the October and November files from tmp to ecommerce1 and ecommerce2 respectively.

Hdfs dfs -mkdir ecommerce1

Hdfs dfs -mkdir ecommerce2

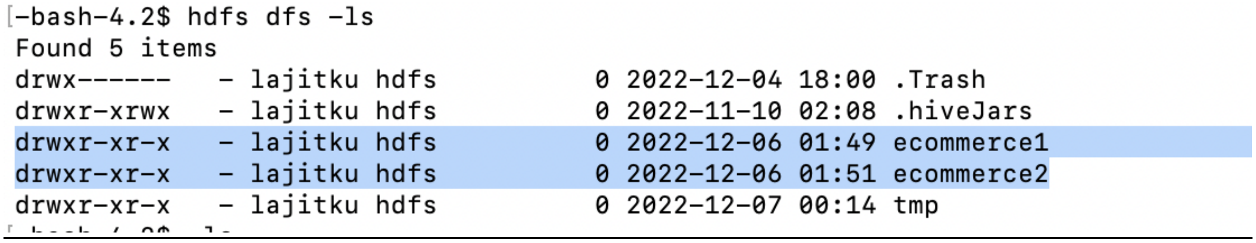
Cd tmp/

hdfs dfs -put 2019-Oct.csv ecommerce_behavior1/

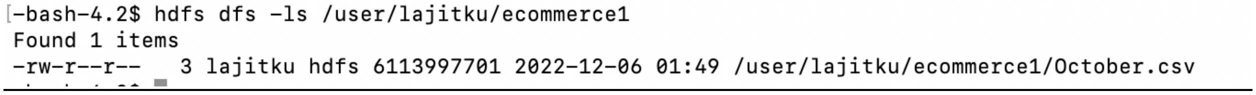
hdfs dfs -put 2019-Nov.csv ecommerce_behavior2/

Confirm files transferred using ls command.

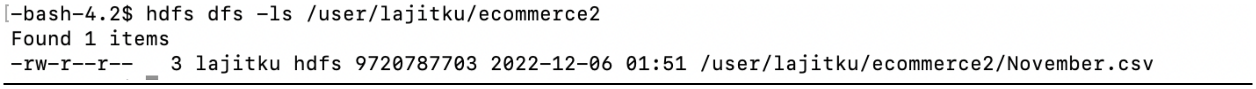
Hdfs dfs -ls



```
[-bash-4.2$ hdfs dfs -ls
 Found 5 items
 drwx------   - lajitku hdfs          0 2022-12-04 18:00 .Trash
 drwxr-xrwx   - lajitku hdfs          0 2022-11-10 02:08 .hiveJars
 drwxr-xr-x   - lajitku hdfs          0 2022-12-06 01:49 ecommerce1
 drwxr-xr-x   - lajitku hdfs          0 2022-12-06 01:51 ecommerce2
 drwxr-xr-x   - lajitku hdfs          0 2022-12-07 00:14 tmp
```



```
[-bash-4.2$ hdfs dfs -ls /user/lajitku/ecommerce1
 Found 1 items
 -rw-r--r--   3 lajitku hdfs 6113997701 2022-12-06 01:49 /user/lajitku/ecommerce1/October.csv
```



```
[-bash-4.2$ hdfs dfs -ls /user/lajitku/ecommerce2
 Found 1 items
 -rw-r--r--   3 lajitku hdfs 9720787703 2022-12-06 01:51 /user/lajitku/ecommerce2/November.csv
```

Step 3: Create Hive Tables

The following Hive statement creates an external table that allows Hive to query data stored in HDFS. External tables preserve the data in the original file format while allowing the Hive to perform queries against the data within the file.

The Hive statements below creates a new table, by describing the fields and the delimiter (Comma) between fields from the file.

Now you have to open another terminal window and login into your account using ssh command.

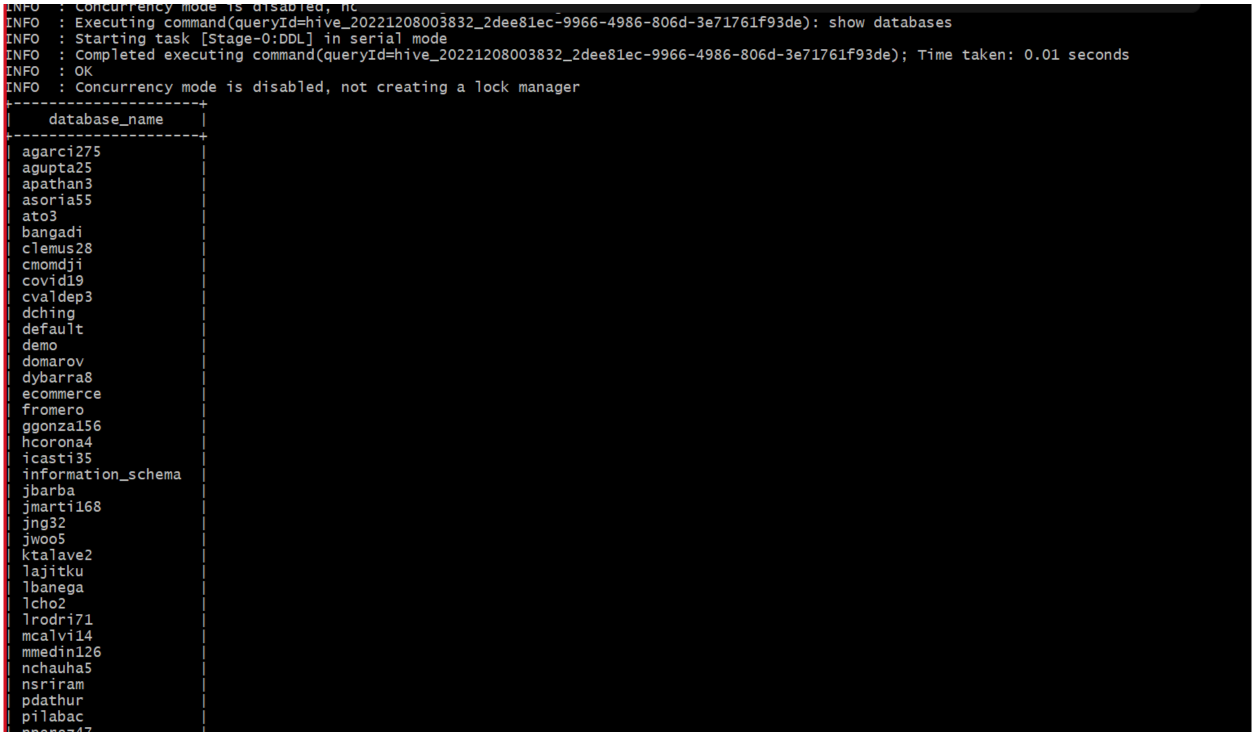
Open beeline Command Line Interface using the following command to run hive queries. Beeline is for multiple users access to Hive Server 2 of a Hadoop cluster.

```
-bash-4.2$ beeline
```

Now you must create your database with your username to separate your tables from other users. For example, the user (lajitku) should run the following:

```
0: jdbc:hive2://bigdaiwn0.sub02180640120.trai> CREATE DATABASE IF NOT EXISTS lajitku;

0: jdbc:hive2://bigdaiwn0.sub02180640120.trai> show databases;
```



```
INFO  : Concurrency mode is disabled, nc
INFO  : Executing command(queryId=hive_20221208003832_2dee81ec-9966-4986-806d-3e71761f93de): show databases
INFO  : Starting task [Stage-0:DDL] in serial mode
INFO  : Completed executing command(queryId=hive_20221208003832_2dee81ec-9966-4986-806d-3e71761f93de); Time taken: 0.01 seconds
INFO  : OK
INFO  : Concurrency mode is disabled, not creating a lock manager
+---------------------+
|    database_name    |
+---------------------+
| agarci275           |
| agupta25            |
| apathan3            |
| asoria55            |
| ato3                |
| bangadi             |
| clemus28            |
| cmomdji             |
| covid19             |
| cvaldep3            |
| dching              |
| default             |
| demo                |
| domarov             |
| dybarra8            |
| ecommerce           |
| fromero             |
| ggonza156           |
| hcorona4            |
| icasti35            |
| information_schema  |
| jbarba              |
| jmarti168           |
| jng32               |
| jwoo5               |
| ktalave2            |
| lajitku             |
| lbanega             |
| lcho2               |
| lrodri71            |
| mcalvi14            |
| mmedin126           |
| nchauha5            |
| nsriram             |
| pdathur             |
| pilabac             |
```

```
0: jdbc:hive2://bigdaiwn0.sub02180640120.trai> use lajitku;
```

Note: use your database name instead of lajitku

October month:

```
CREATE EXTERNAL TABLE IF NOT EXISTS Octuncleaned (
sno INT,
event_time STRING,
event_type STRING,
product_id INT,
category_id BIGINT,
category_code STRING,
brand STRING,
price DOUBLE,
user_id INT,
user_session STRING)
ROW FORMAT DELIMITED FIELDS TERMINATED BY ','
STORED AS TEXTFILE LOCATION '/user/lajitku/ecommerce1/'
TBLPROPERTIES ('skip.header.line.count'='1');
```

November month:
CREATE EXTERNAL TABLE IF NOT EXISTS Novuncleaned (
sno INT,
event_time STRING,
event_type STRING,
product_id INT,
category_id BIGINT,
category_code STRING,
brand STRING,
price DOUBLE,
user_id INT,
user_session STRING)
ROW FORMAT DELIMITED FIELDS TERMINATED BY ','
STORED AS TEXTFILE LOCATION '/user/lajitku/ecommerce2/'
TBLPROPERTIES ('skip.header.line.count'='1');

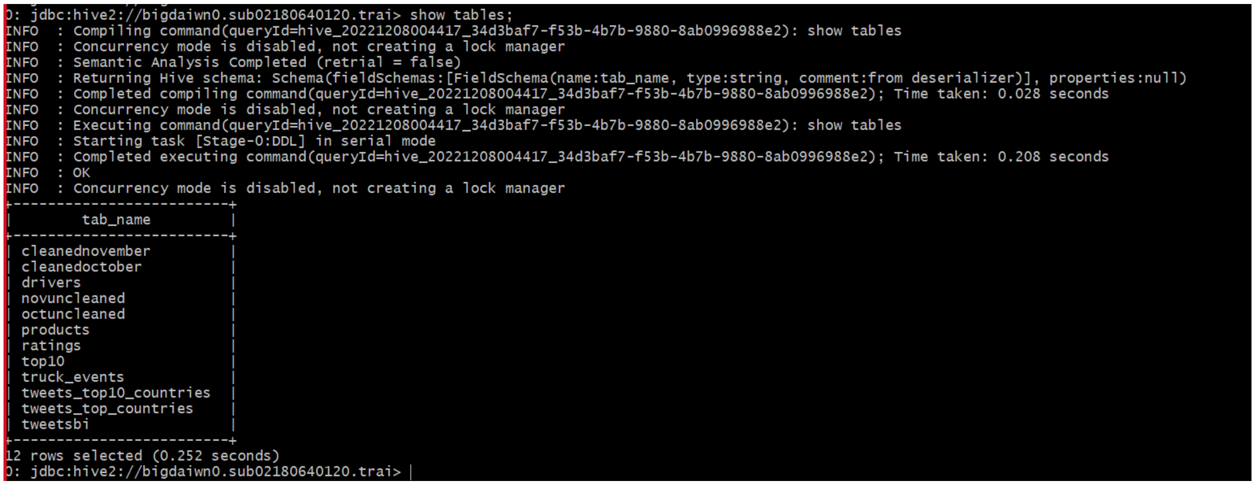
Data Cleaning and Creation of New Tables:

October month:
CREATE TABLE IF NOT EXISTS cleanedoctober
AS SELECT * from octuncleaned
where category_code not like "NULL" AND brand not like "NULL" AND user_session not like "NULL";

November month:
CREATE TABLE IF NOT EXISTS cleanednovember
AS SELECT * from novuncleaned
where category_code not like "NULL" AND brand not like "NULL" AND user_session not like "NULL";

Confirm the Tables creation using Show Tables;

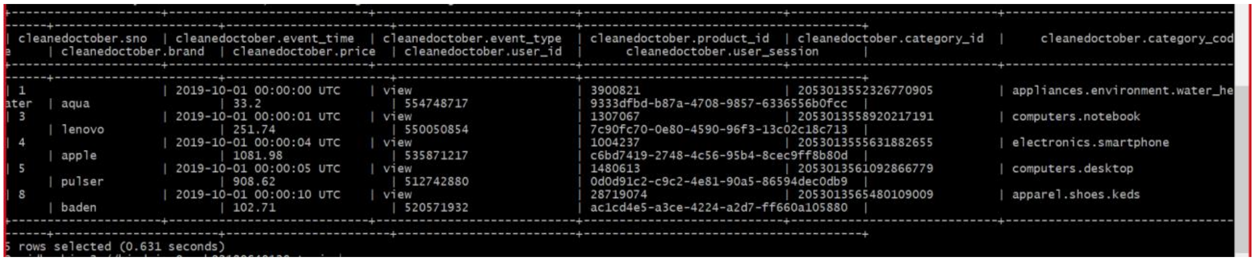
```
INFO  : Compiling command(queryId=hive_20221208004417_34d3baf7-f53b-4b7b-9880-8ab0996988e2): show tables
```



```
0: jdbc:hive2://bigdaiwn0.sub02180640120.trai> show tables;
INFO  : Compiling command(queryId=hive_20221208004417_34d3baf7-f53b-4b7b-9880-8ab0996988e2): show tables
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Semantic Analysis Completed (retrial = false)
INFO  : Returning Hive schema: Schema(fieldSchemas:[FieldSchema(name:tab_name, type:string, comment:from deserializer)], properties:null)
INFO  : Completed compiling command(queryId=hive_20221208004417_34d3baf7-f53b-4b7b-9880-8ab0996988e2); Time taken: 0.028 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20221208004417_34d3baf7-f53b-4b7b-9880-8ab0996988e2): show tables
INFO  : Starting task [Stage-0:DDL] in serial mode
INFO  : Completed executing command(queryId=hive_20221208004417_34d3baf7-f53b-4b7b-9880-8ab0996988e2); Time taken: 0.208 seconds
INFO  : OK
INFO  : Concurrency mode is disabled, not creating a lock manager
+-------------------------+
|        tab_name         |
+-------------------------+
| cleanednovember         |
| cleanedoctober          |
| drivers                 |
| novuncleaned            |
| octuncleaned            |
| products                |
| ratings                 |
| top10                   |
| truck_events            |
| tweets_top10_countries  |
| tweets_top_countries    |
| tweetsbi                |
+-------------------------+
12 rows selected (0.252 seconds)
0: jdbc:hive2://bigdaiwn0.sub02180640120.trai> |
```

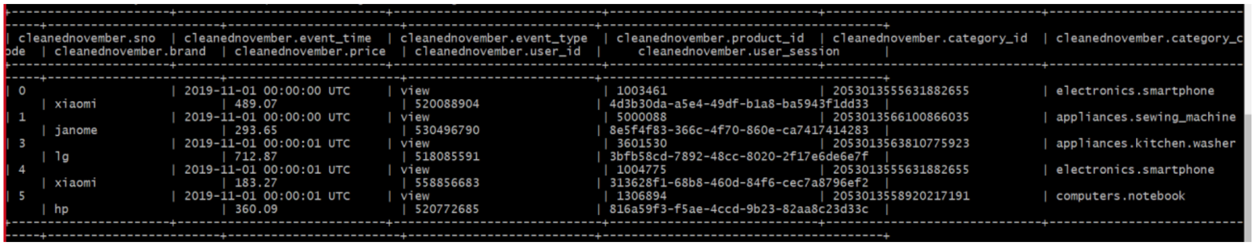
Confirm contents in table with the SELECT statement.

```
0: jdbc:hive2://bigdaiwn0.sub02180640120.trai> SELECT * from cleanedoctober limit 5;
```



```
+--------------------+--------------------------+---------------------------+---------------------------+-----------------------------+------------------------------
-----+--------------------+--------------------------+----------------------------------+
| cleanedoctober.sno  | cleanedoctober.event_time  | cleanedoctober.event_type  | cleanedoctober.product_id  | cleanedoctober.category_id  |     cleanedoctober.category_cod
e     | cleanedoctober.brand  | cleanedoctober.price  | cleanedoctober.user_id  |       cleanedoctober.user_session       |
+--------------------+--------------------------+---------------------------+---------------------------+-----------------------------+------------------------------
-----+--------------------+--------------------------+----------------------------------+
| 1                   | 2019-10-01 00:00:00 UTC    | view                       | 3900821                    | 2053013552326770905         | appliances.environment.water_he
ater  | aqua                  | 33.2                  | 554748717               | 9333dfbd-b87a-4708-9857-6336556b0fcc  |
| 3                   | 2019-10-01 00:00:01 UTC    | view                       | 1307067                    | 2053013558920217191         | computers.notebook
      | lenovo                | 251.74                | 550050854               | 7c90fc70-0e80-4590-96f3-13c02c18c713  |
| 4                   | 2019-10-01 00:00:04 UTC    | view                       | 1004237                    | 2053013555631882655         | electronics.smartphone
      | apple                 | 1081.98               | 535871217               | c6bd7419-2748-4c56-95b4-8cec9ff8b80d  |
| 5                   | 2019-10-01 00:00:05 UTC    | view                       | 1480613                    | 2053013561092866779         | computers.desktop
      | pulser                | 908.62                | 512742880               | 0d0d91c2-c9c2-4e81-90a5-86594dec0db9  |
| 8                   | 2019-10-01 00:00:10 UTC    | view                       | 28719074                   | 2053013565480109009         | apparel.shoes.keds
      | baden                 | 102.71                | 520571932               | ac1cd4e5-a3ce-4224-a2d7-ff660a105880  |
+--------------------+--------------------------+---------------------------+---------------------------+-----------------------------+------------------------------
-----+--------------------+--------------------------+----------------------------------+
5 rows selected (0.631 seconds)
```

```
0: jdbc:hive2://bigdaiwn0.sub02180640120.trai> SELECT * from cleanednovember limit 5;
```



```
+---------------------+----------------------------+-----------------------------+-----------------------------+-------------------------------+-----------------------------
-----+---------------------+----------------------------+-------------------------------------+
| cleanednovember.sno  | cleanednovember.event_time  | cleanednovember.event_type  | cleanednovember.product_id  | cleanednovember.category_id  | cleanednovember.category_c
ode  | cleanednovember.brand  | cleanednovember.price  | cleanednovember.user_id  |       cleanednovember.user_session       |
+---------------------+----------------------------+-----------------------------+-----------------------------+-------------------------------+-----------------------------
-----+---------------------+----------------------------+-------------------------------------+
| 0                    | 2019-11-01 00:00:00 UTC     | view                        | 1003461                     | 2053013555631882655          | electronics.smartphone
     | xiaomi                 | 489.07                 | 520088904                | 4d3b30da-a5e4-49df-b1a8-ba5943f1dd33  |
| 1                    | 2019-11-01 00:00:00 UTC     | view                        | 5000088                     | 2053013566100866035          | appliances.sewing_machine
     | janome                 | 293.65                 | 530496790                | 8e5f4f83-366c-4f70-860e-ca7417414283  |
| 3                    | 2019-11-01 00:00:01 UTC     | view                        | 3601530                     | 2053013563810775923          | appliances.kitchen.washer
     | lg                     | 712.87                 | 518085591                | 3bfb58cd-7892-48cc-8020-2f17e6de6e7f  |
| 4                    | 2019-11-01 00:00:01 UTC     | view                        | 1004775                     | 2053013555631882655          | electronics.smartphone
     | xiaomi                 | 183.27                 | 558856683                | 313628f1-68b8-460d-84f6-cec7a8796ef2  |
| 5                    | 2019-11-01 00:00:01 UTC     | view                        | 1306894                     | 2053013558920217191          | computers.notebook
     | hp                     | 360.09                 | 520772685                | 816a59f3-f5ae-4ccd-9b23-82aa8c23d33c  |
+---------------------+----------------------------+-----------------------------+-----------------------------+-------------------------------+-----------------------------
-----+---------------------+----------------------------+-------------------------------------+
```

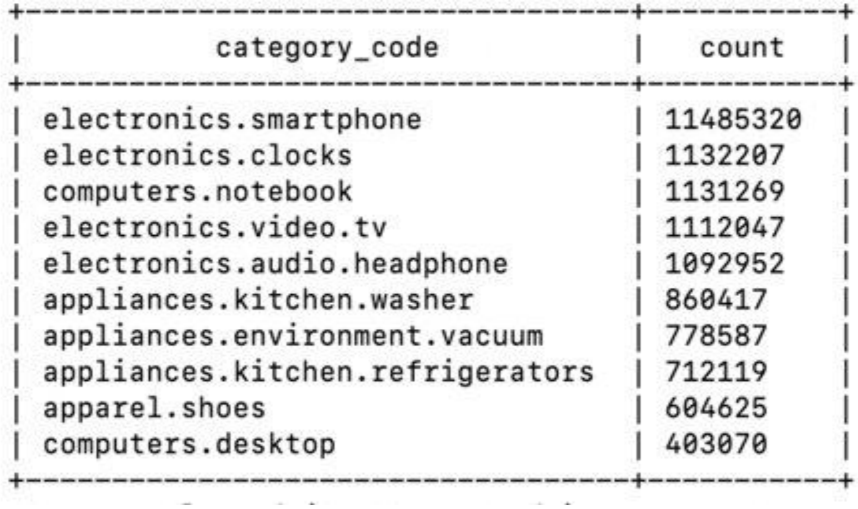
Step 4: Create Hive Table Queries

The following Queries will help us to figure out the Visualization and analyze the Customer Behavior

Top 10 popular categories in October and November

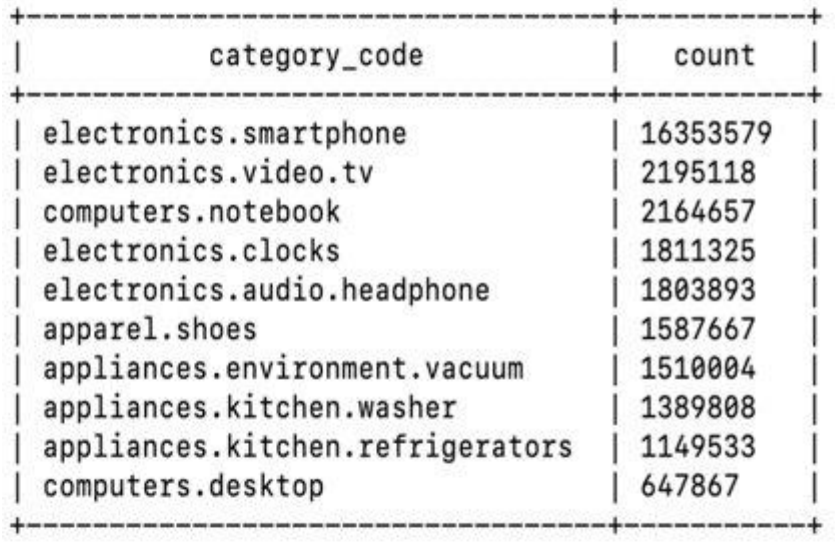
October

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    |
+-----------------------------------+-----------+
```

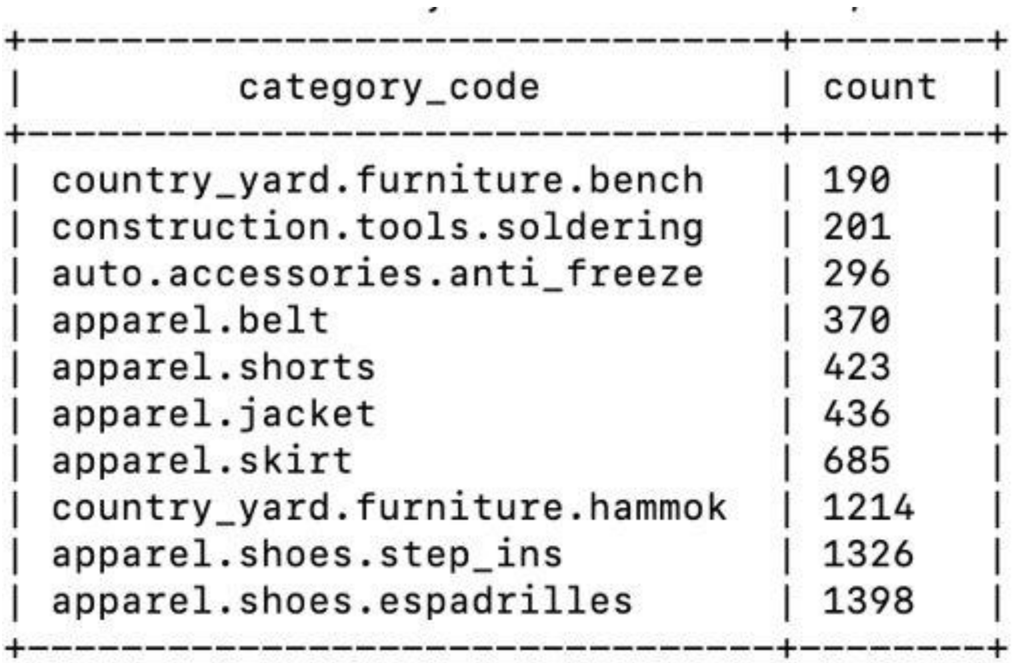
November
select category_code, count(category_code) as count from cleanednovember group by category_code order by count(category_code) desc limit 10;



```
+----------------------------------+----------+
|          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   |
+----------------------------------+----------+
```

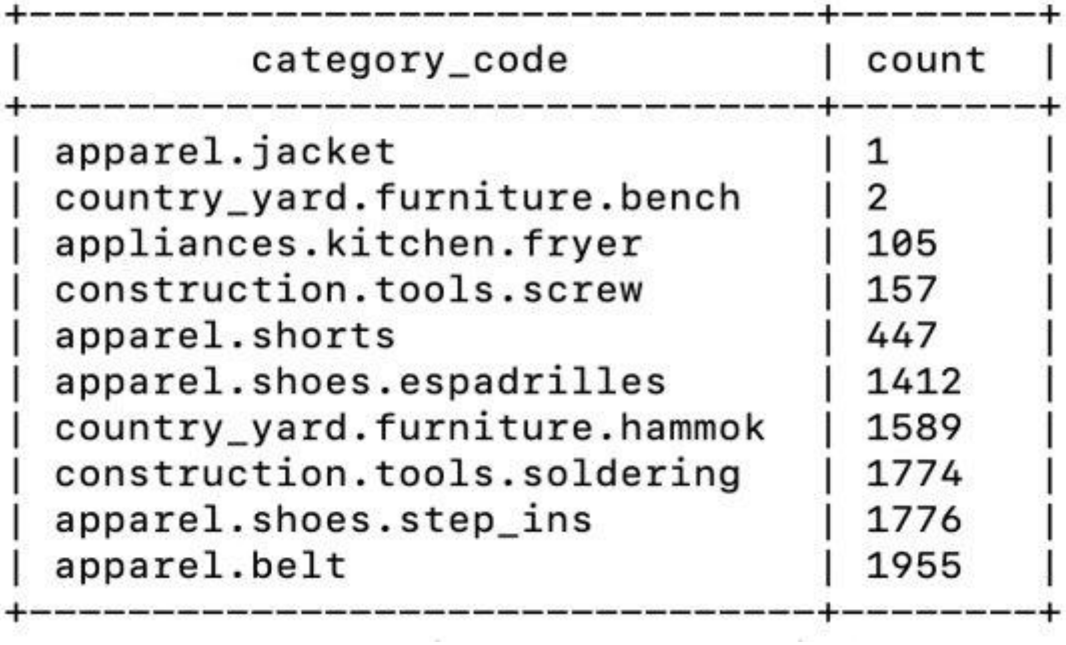
Top 10 Least popular categories in October and November

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   |
+----------------------------------+--------+
```

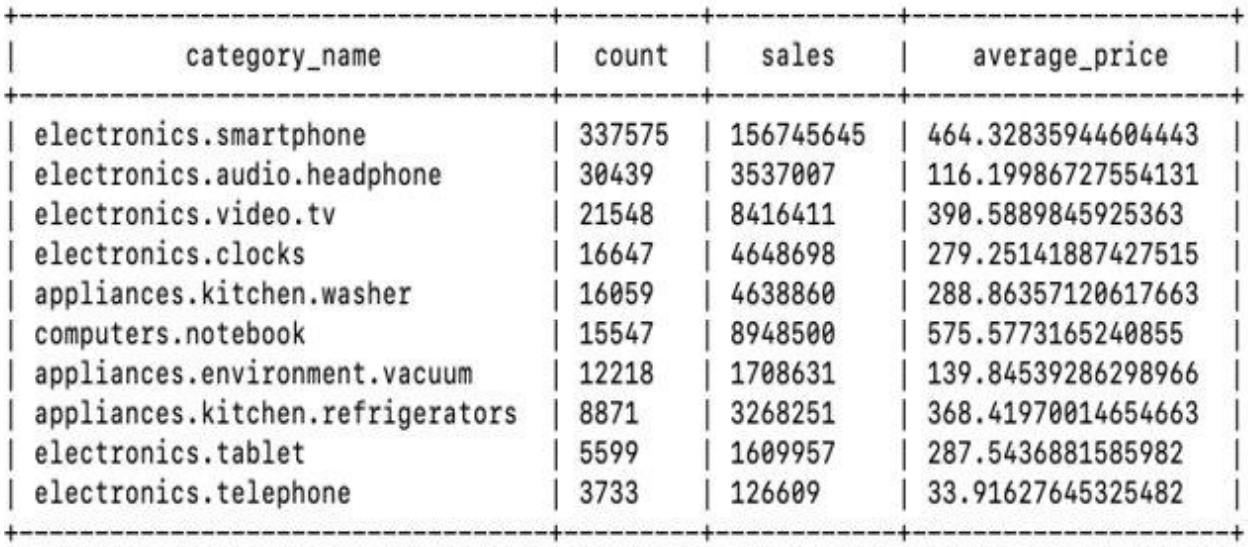
November
select category_code, count(category_code) as count from cleanednovember group by category_code order by count(category_code) limit 10;



```
+---------------------------------+--------+
|          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   |
+---------------------------------+--------+
```

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

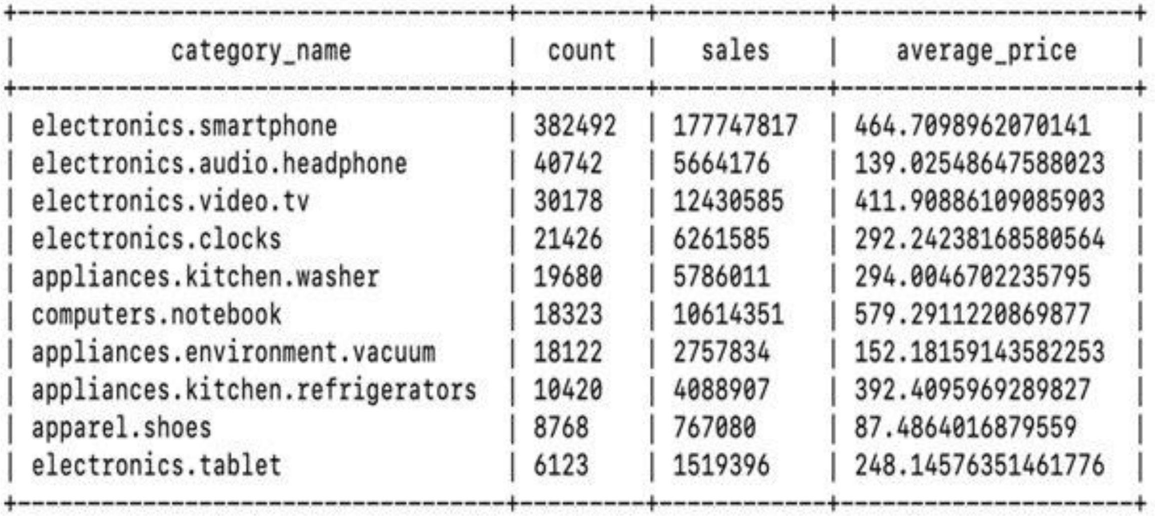
October
select category_name as category_name, count(category_code) as count, cast(sum(price) as bigint) as sales, avg(price) as average_price from cleanedoctober where event_type like 'purchase' group by category_code order by count(category_code) desc limit 10;



```
+-----------------------------------+---------+-----------+--------------------+
|           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  |
+-----------------------------------+---------+-----------+--------------------+
```

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 cleanednovember where event_type like 'purchase' group by category_code order by count(category_code) desc limit 10;

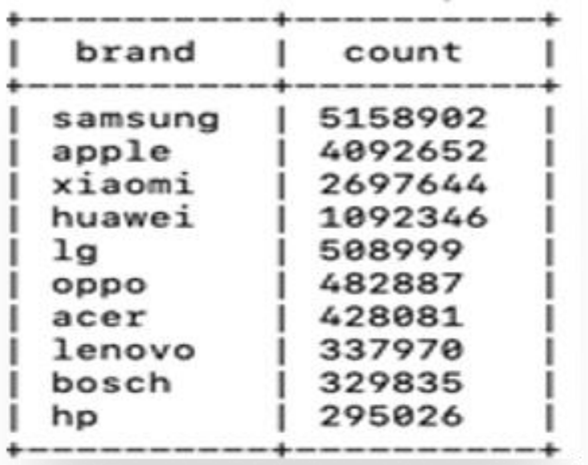


```
+----------------------------------+---------+------------+--------------------+
|           category_name          |  count  |   sales    |   average_price    |
+----------------------------------+---------+------------+--------------------+
| electronics.smartphone           | 382492  | 177747817  | 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 popular brands October and November

October

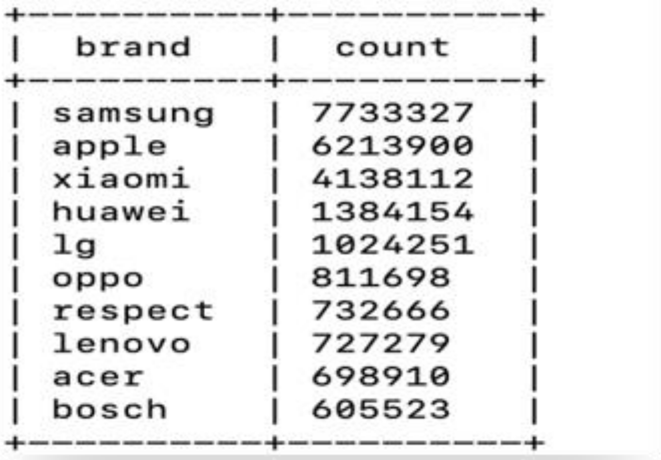
select brand, count(brand) as count from cleanedoctober group by brand order by count(brand) desc limit 10;



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

November

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

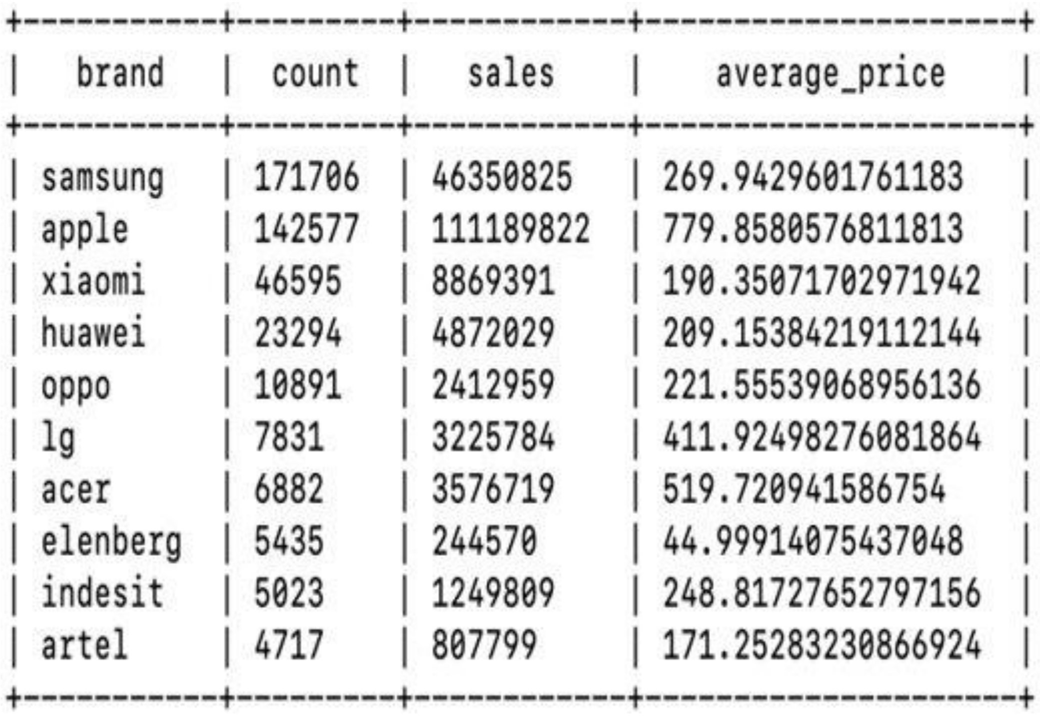


```
+-----------+-----------+
|  brand    |  count    |
+-----------+-----------+
| samsung   | 7733327   |
| apple     | 6213900   |
| xiaomi    | 4138112   |
| huawei    | 1384154   |
| lg        | 1024251   |
| oppo      | 811698    |
| respect   | 732666    |
| lenovo    | 727279    |
| acer    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

Top 10 Purchased Brands of October and November

October

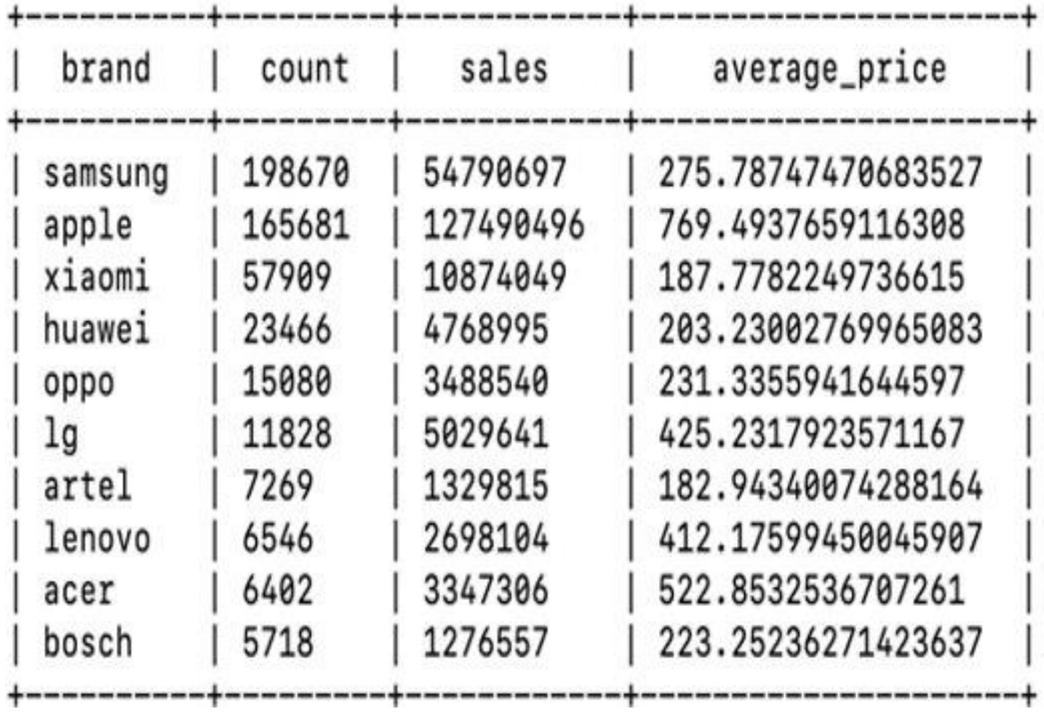
select brand, count(brand) as count, cast(sum(price) as bigint) as sales, avg(price) as average_price from cleanedoctober where event_type like 'purchase' group by brand order by count(brand) desc limit 10;



```
+----------+---------+-----------+--------------------+
|  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 |
+----------+---------+-----------+--------------------+
```

November

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;

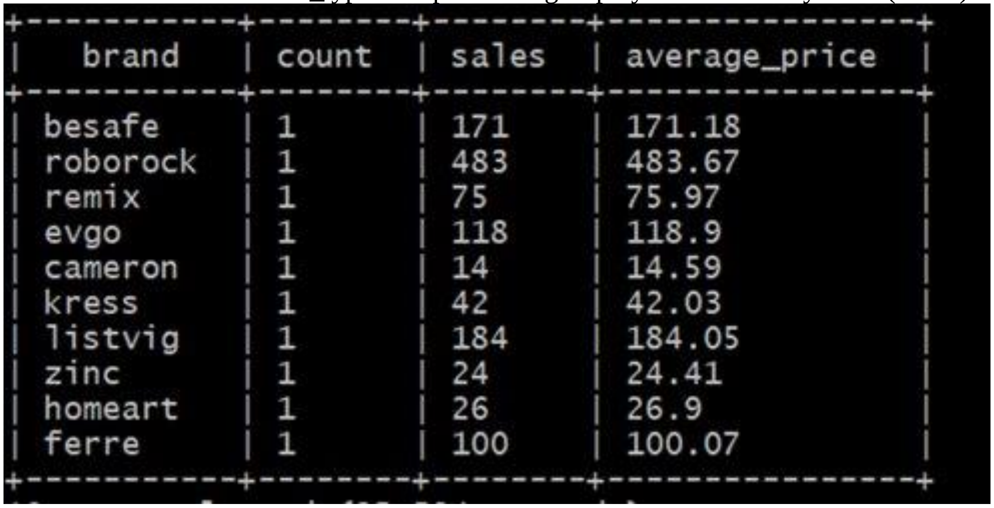


```
+----------+---------+-----------+--------------------+
|  brand   |  count  |   sales   |    average_price   |
+----------+---------+-----------+--------------------+
| samsung  | 198670  | 54790697  | 275.78747470683527 |
| apple    | 165681  | 127490496 | 769.4937659116308  |
| xiaomi   | 57909   | 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   | 223.2523627142363 7|
+----------+---------+-----------+--------------------+
```

Top 10 Least Purchased Brands of October and November

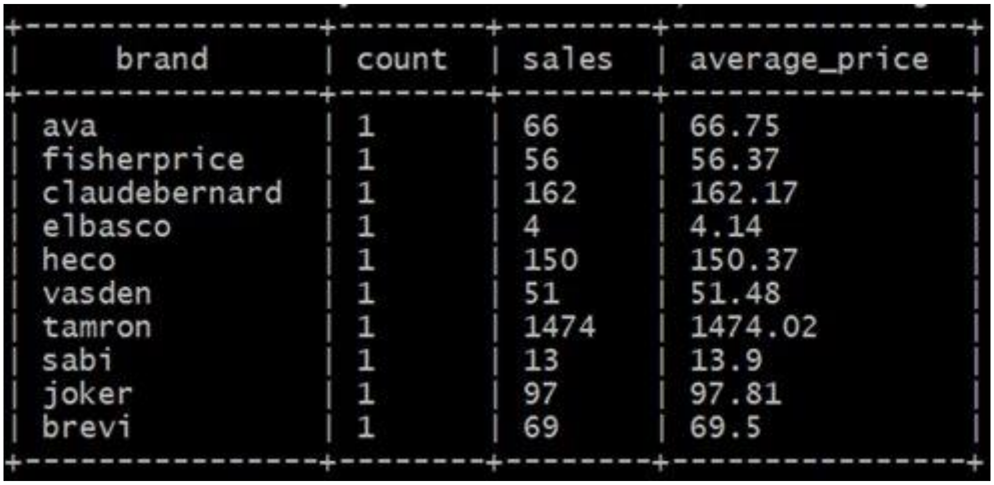
October

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;



```
+-----------+--------+--------+---------------+
|   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        |
+-----------+--------+--------+---------------+
```

November
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;

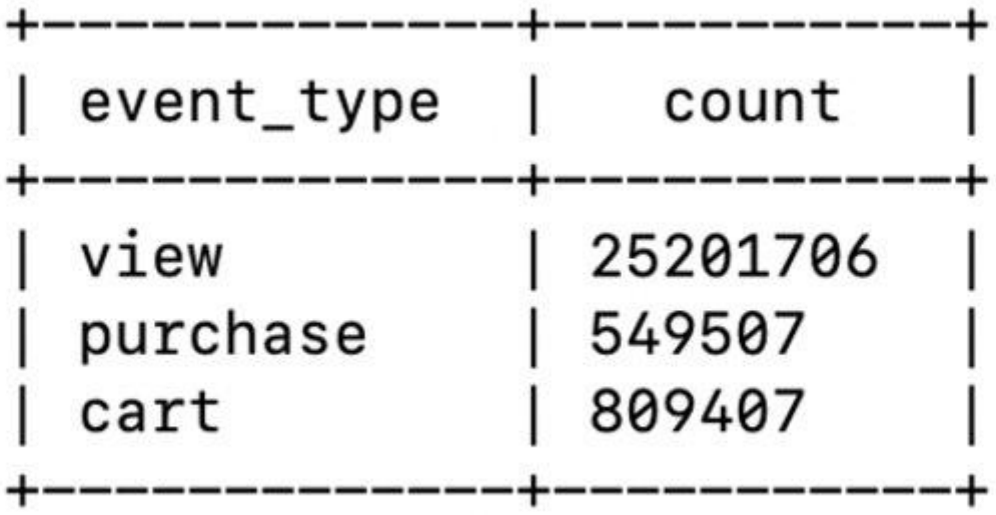


```
+---------------+--------+--------+---------------+
|     brand     | count  | sales  | average_price |
+---------------+--------+--------+---------------+
| ava           | 1      | 66     | 66.75         |
| fisherprice   | 1      | 56     | 56.37         |
| claudebernard | 1      | 162    | 162.17        |
| 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          |
+---------------+--------+--------+---------------+
```

Views, Purchases, In-Carts in October and November

October

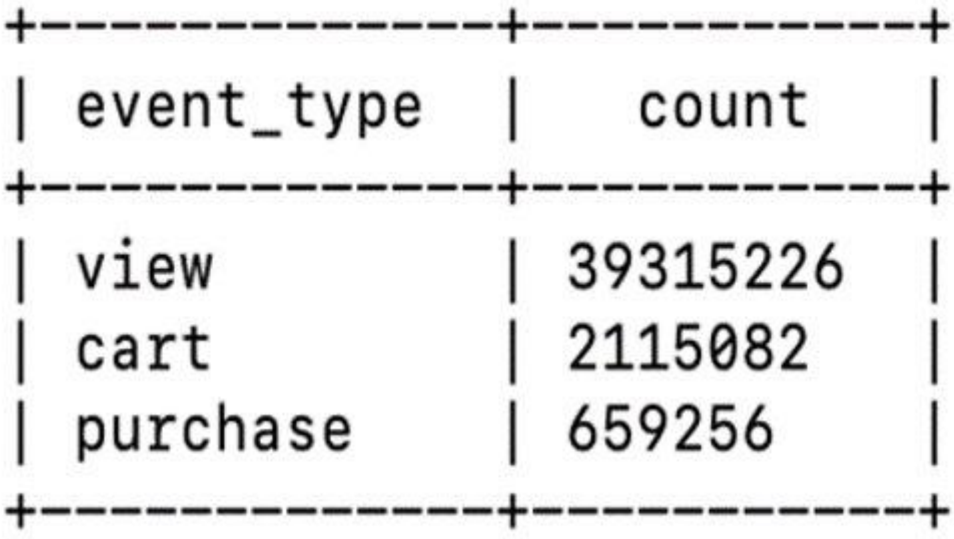
select event_type, count(event_type) as count from cleanedoctober group by event_type;



```
+-------------+-----------+
| event_type  |   count   |
+-------------+-----------+
| view        | 25201706  |
| purchase    | 549507    |
| cart        | 809407    |
+-------------+-----------+
```

November

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

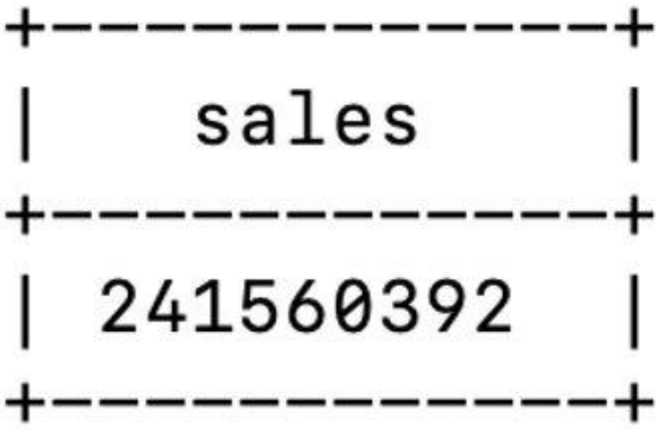


```
+-------------+-----------+
| event_type  |   count   |
+-------------+-----------+
| view        | 39315226  |
| cart        | 2115082   |
| purchase    | 659256    |
+-------------+-----------+
```

Sum of Sales in both October and November

October

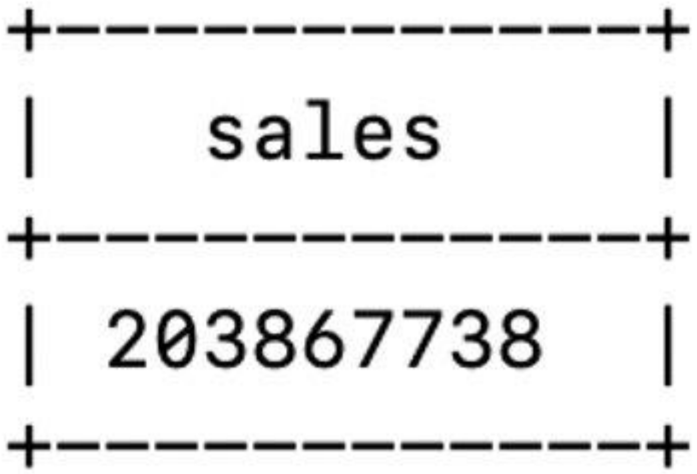
select cast(sum(price) as bigint) as sales from cleanedoctober where event_type like 'purchase’;



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

November

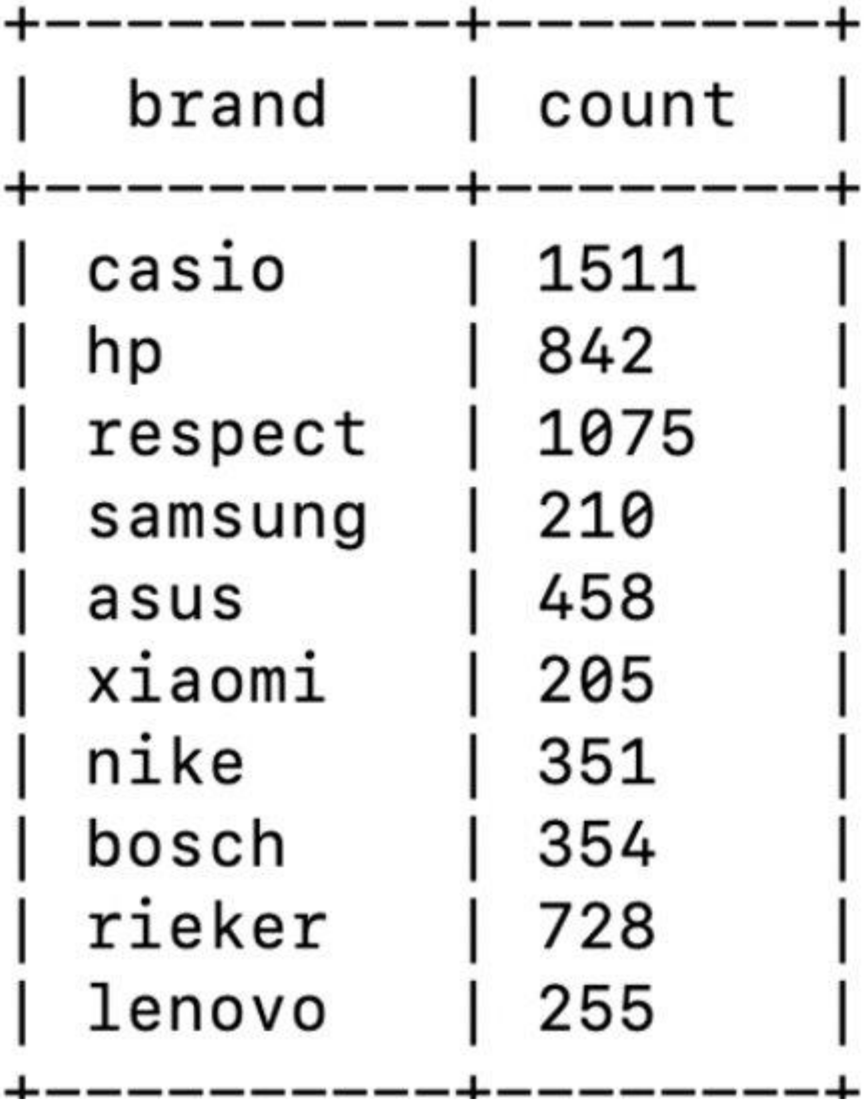
select cast(sum(price) as bigint) as sales from cleanednovember where event_type like 'purchase';



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

Exit rate- Most viewed brand but not purchased

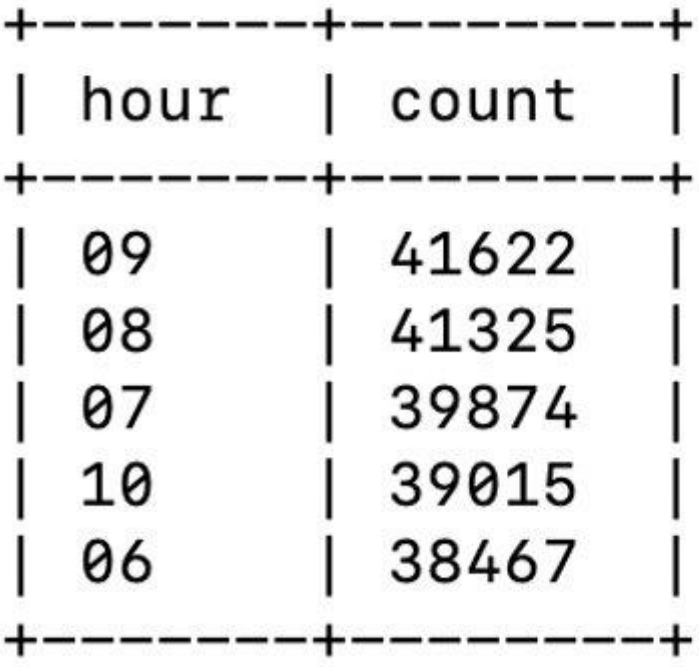
```
select brand, count(distinct product_id) as count from cleanedoctober where event_type = 'view' and
product_id NOT IN (select product_id from cleanedoctober where event_type = 'purchase') group by
brand order by count(product_id) desc limit 10;
```



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

Top 5 hours with most purchases in November

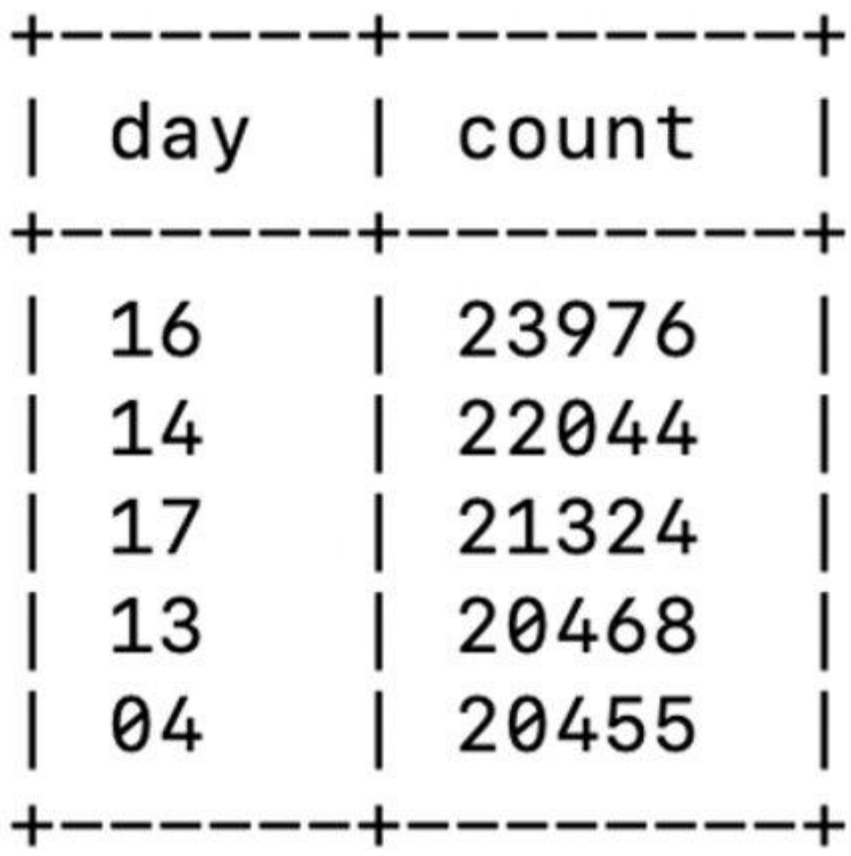
```
Select substr(event_time, 12, 2) as hour, count(substr(event_time, 12, 2)) as count from cleanednovember where event_type like 'purchase' group by substr(event_time, 12, 2) order by count(substr(event_time, 12, 2)) desc limit 5;
```



```
+-------+--------+
| hour  | count  |
+-------+--------+
| 09    | 41622  |
| 08    | 41325  |
| 07    | 39874  |
| 10    | 39015  |
| 06    | 38467  |
+-------+--------+
```

Top 5 days with most purchases in October

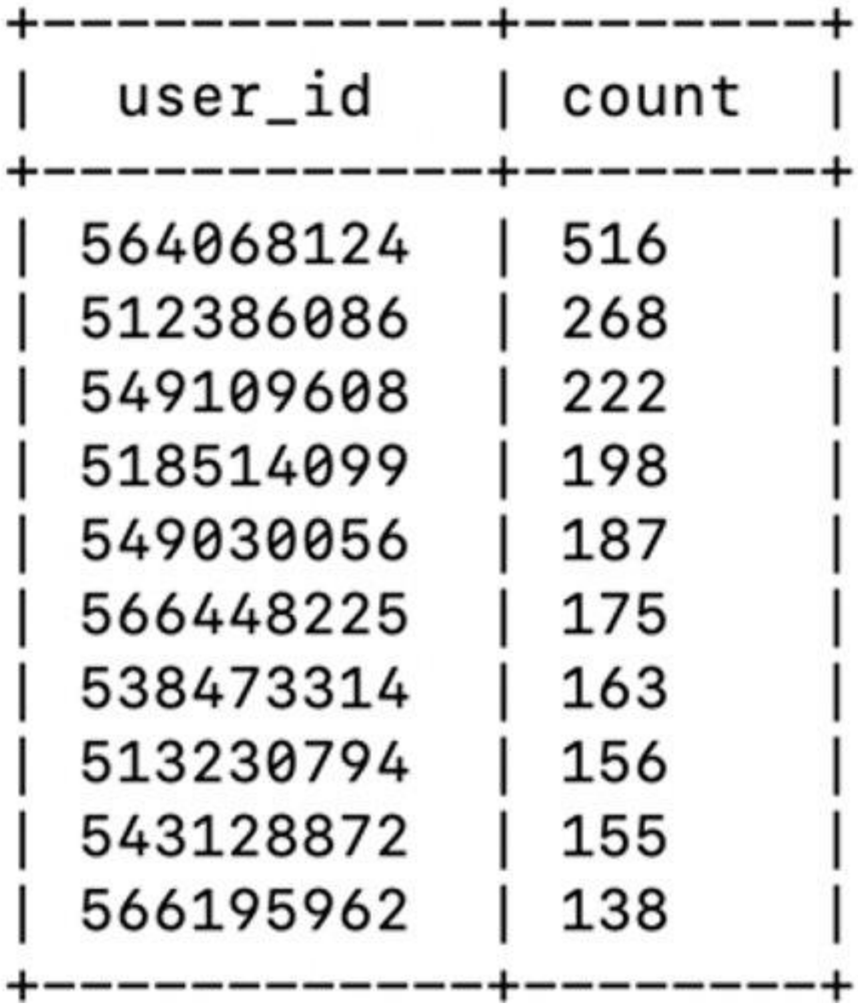
Select substr(event_time, 9, 2) as day, count(substr(event_time, 9, 2)) as count from cleanedoctober where event_type = 'purchase' group by substr(event_time, 9, 2) order by count(substr(event_time, 9, 2)) desc limit 5;



```
+-------+---------+
| day   | count   |
+-------+---------+
| 16    | 23976   |
| 14    | 22044   |
| 17    | 21324   |
| 13    | 20468   |
| 04    | 20455   |
+-------+---------+
```

Top 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;



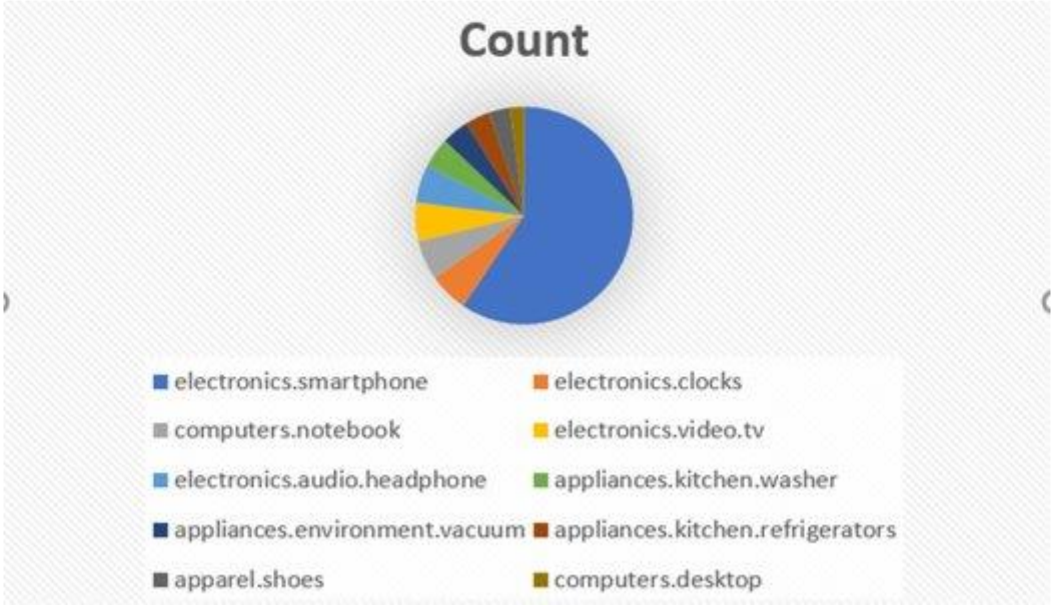
```
+------------+---------+
| user_id    | count   |
+------------+---------+
| 564068124  | 516     |
| 512386086  | 268     |
| 549109608  | 222     |
| 518514099  | 198     |
| 549030056  | 187     |
| 566448225  | 175     |
| 538473314  | 163     |
| 513230794  | 156     |
| 543128872  | 155   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

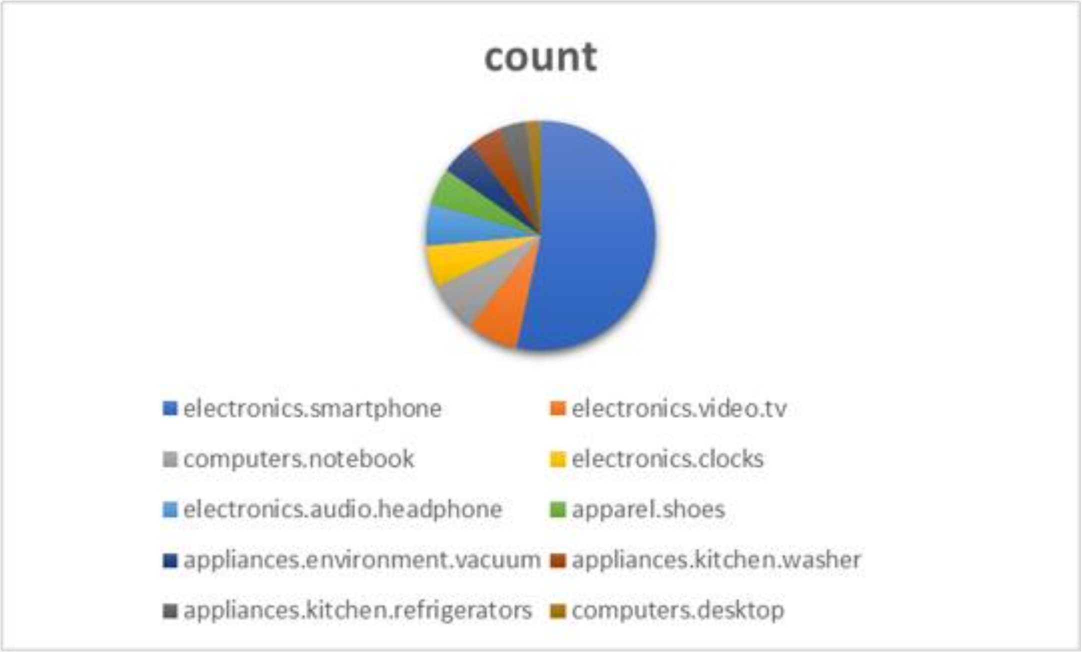
Step 5: Visualization

This step is to show the Visualization for the above Queries.

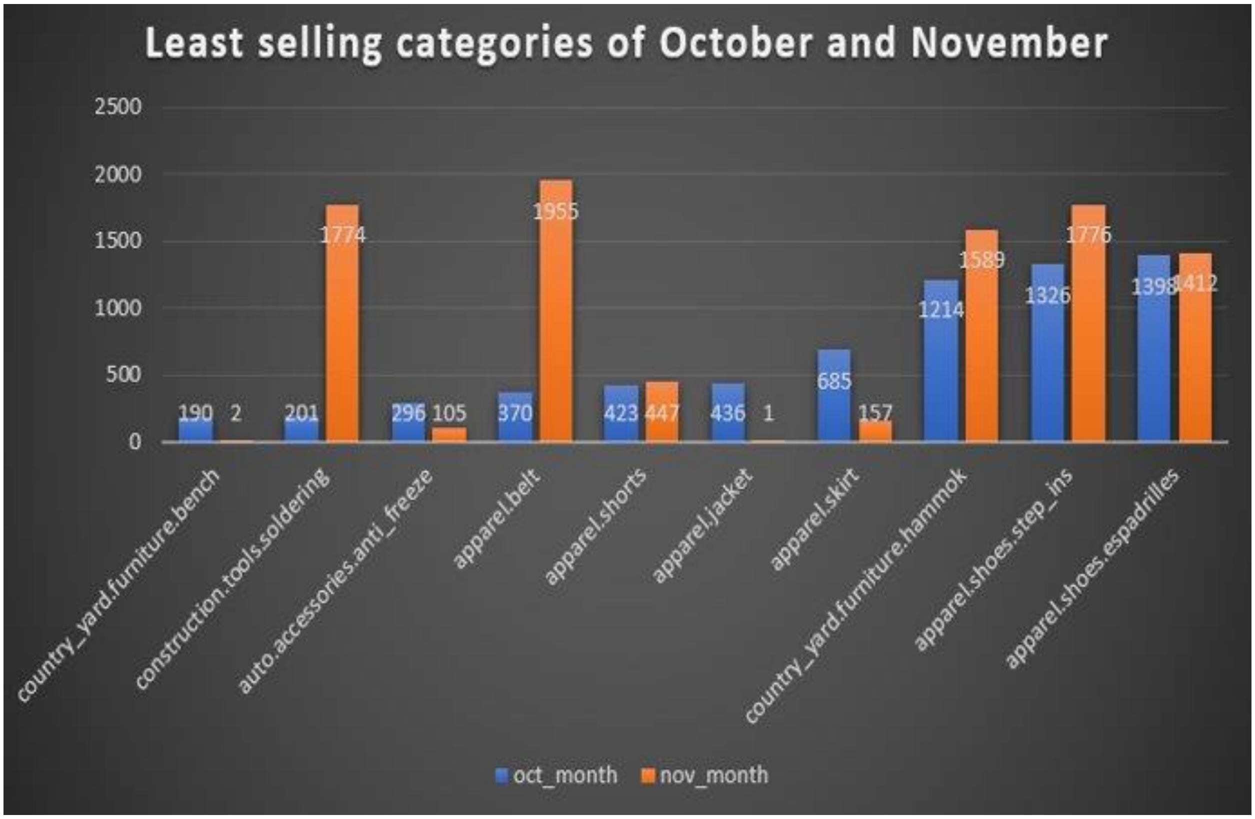
To visualize results on Graphs, convert csv file to excel and click on Graphs button under insert tab.

Top 10 Popular categories in October and November

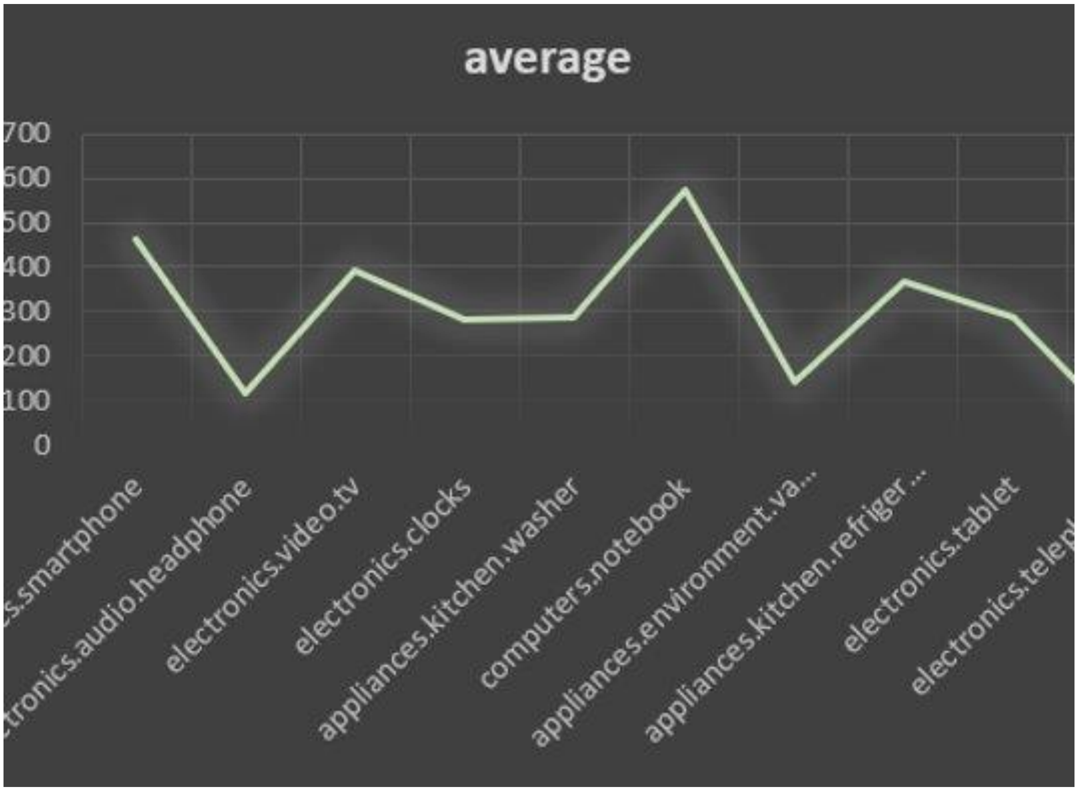


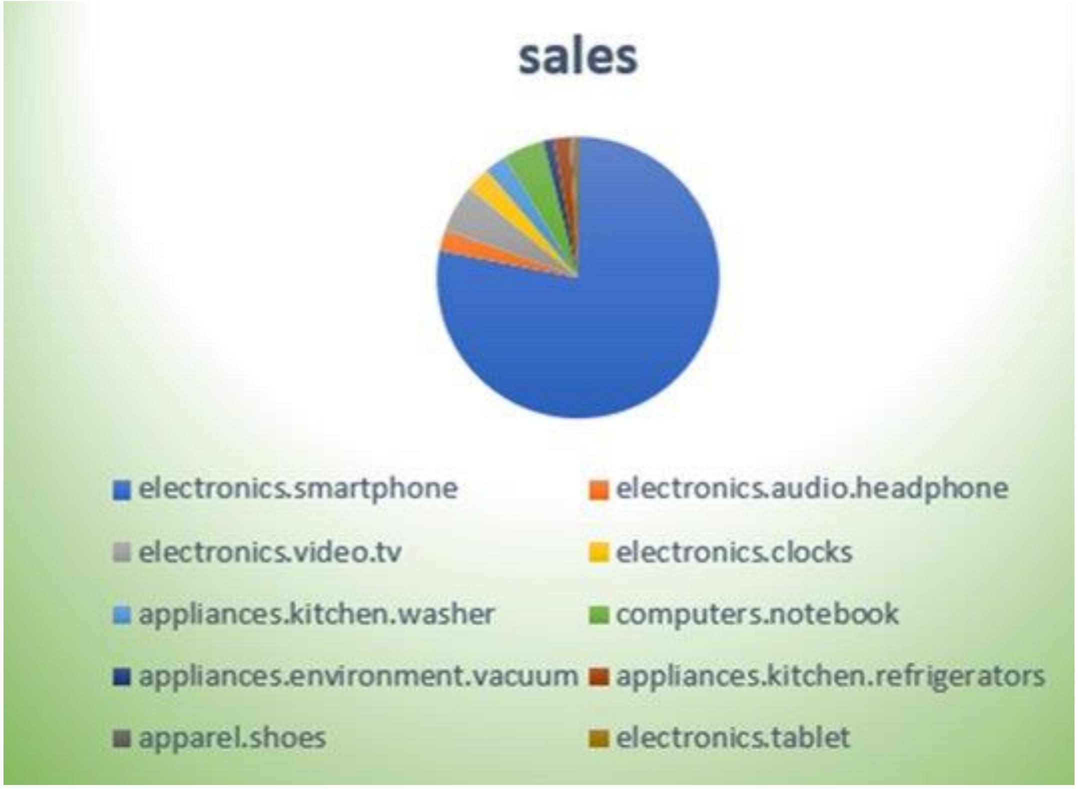


Top 10 Least popular categories in October and November

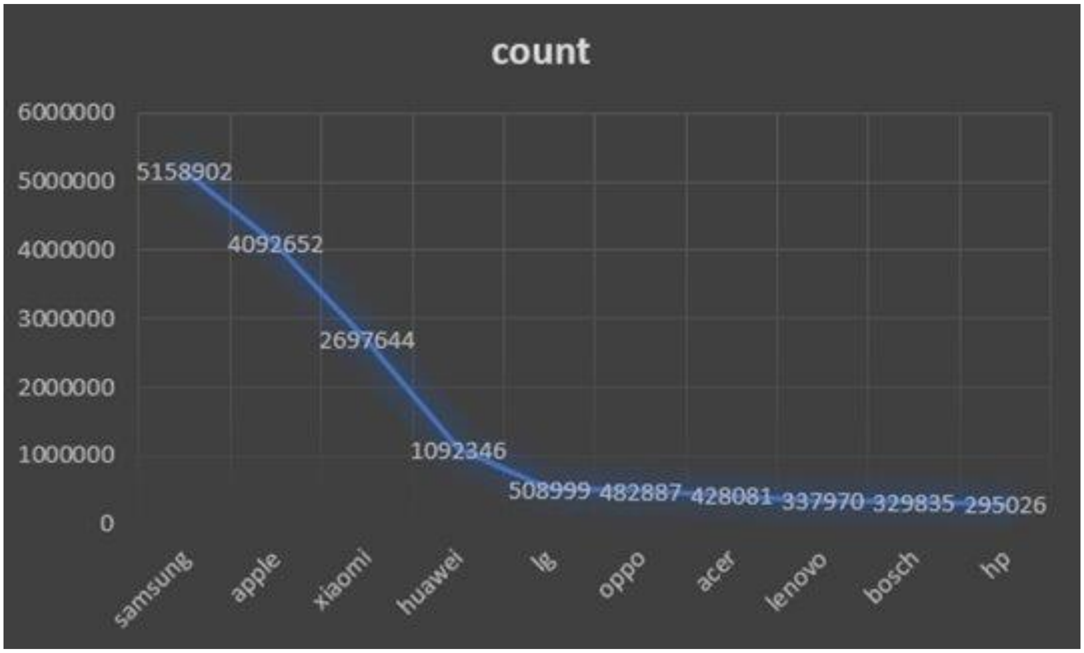


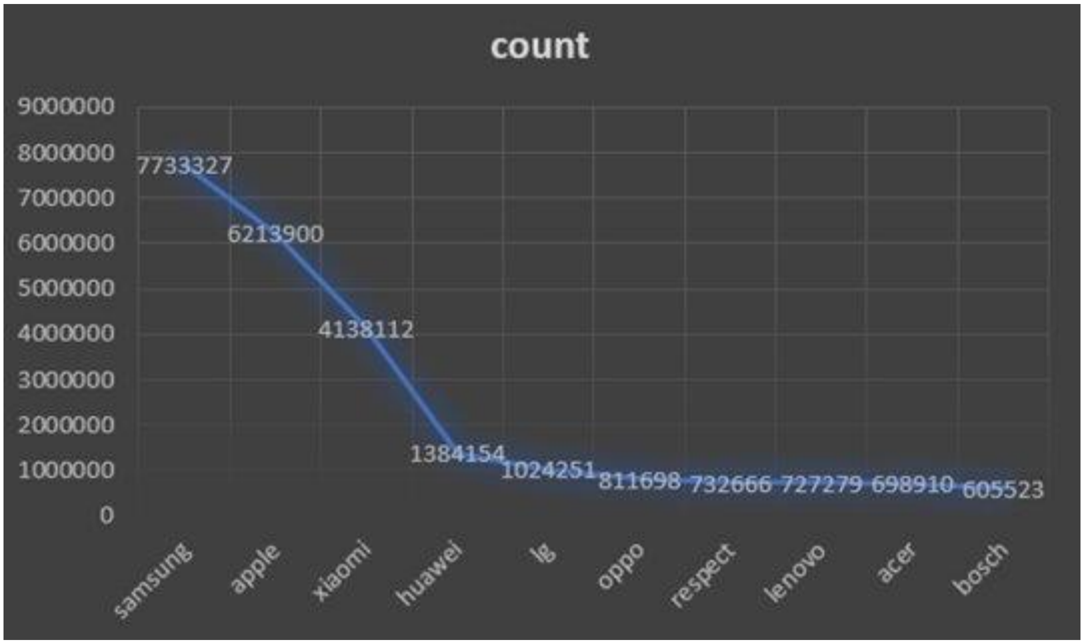
Top 10 purchased categories, sales count and average price in October and November.



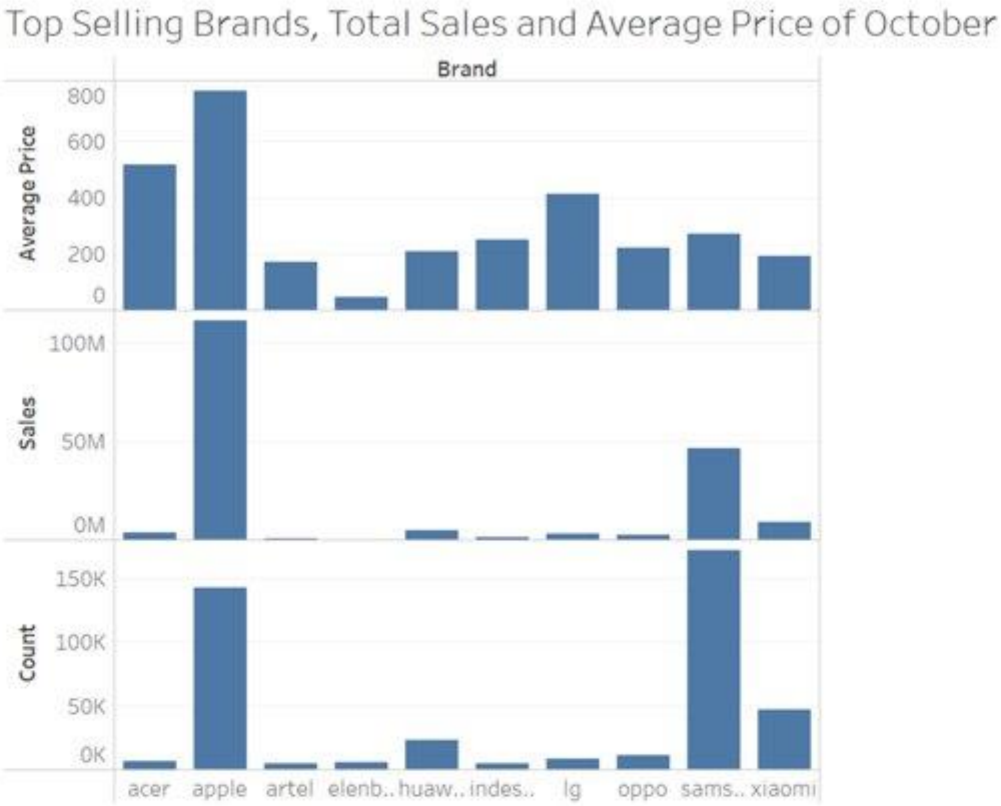


Top 10 popular brands October and November

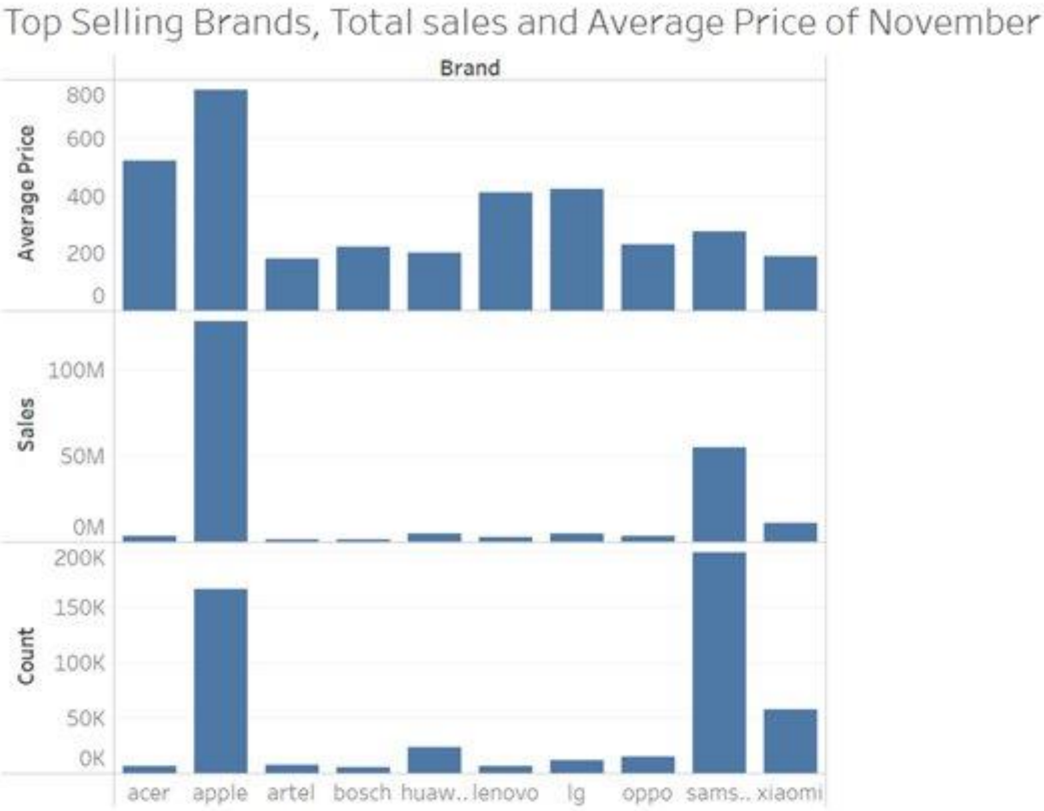




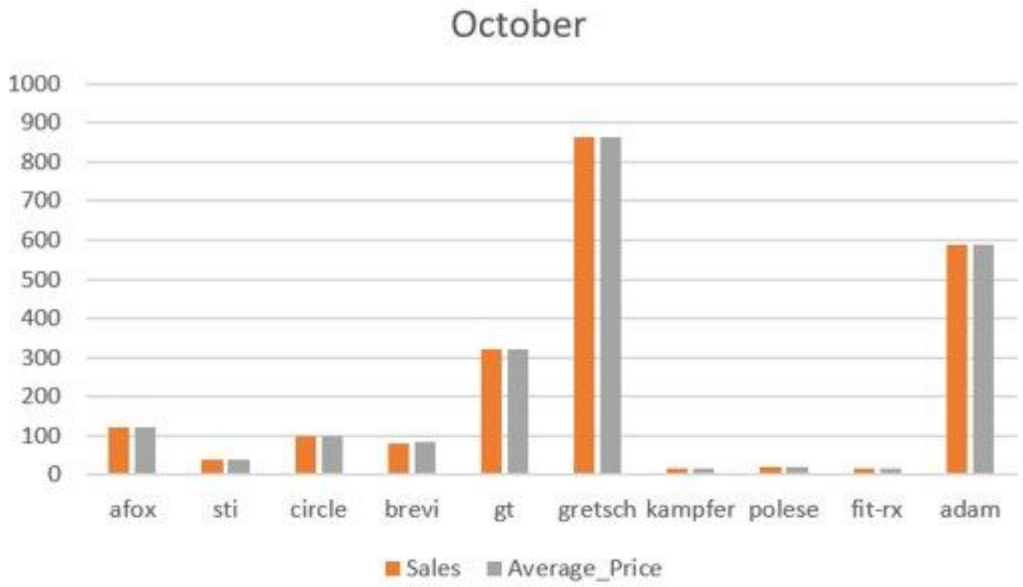
Top 10 Purchased Brands of October and November

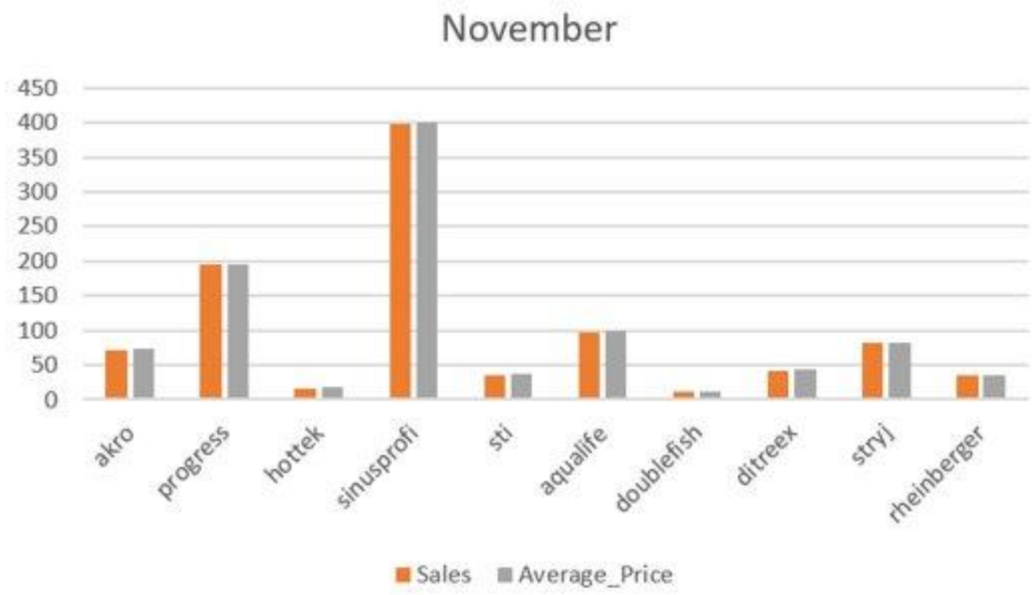


Top Selling Brands, Total Sales and Average Price of October

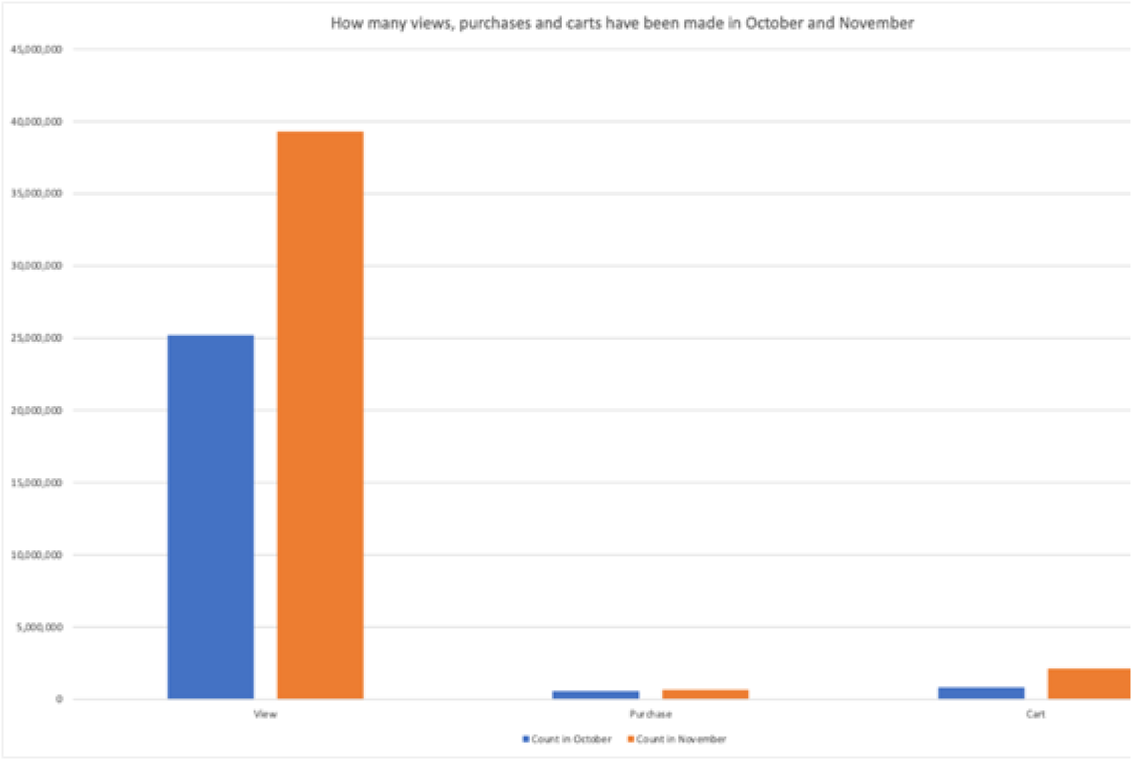


Top 10 Least Purchased Brands of October and November

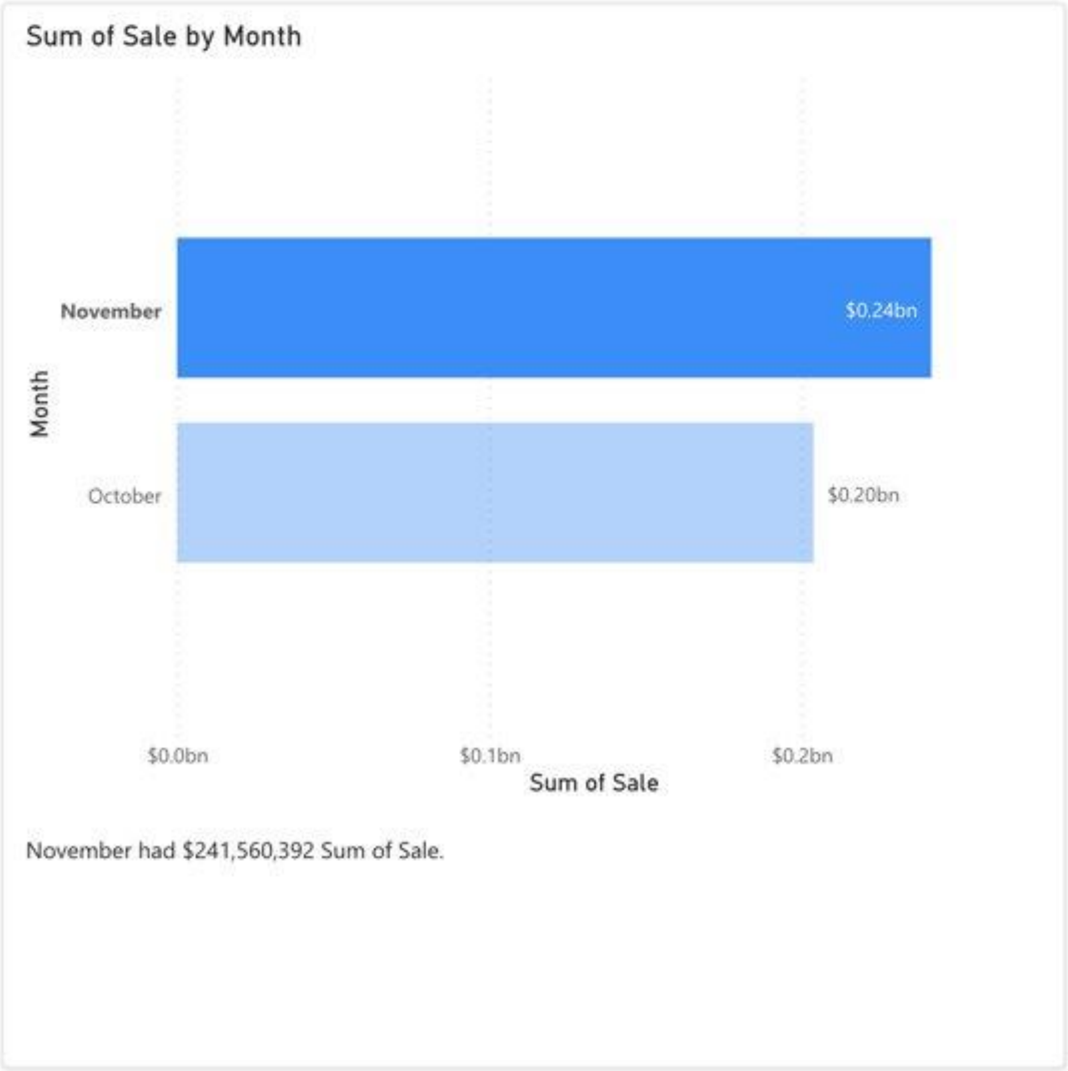




Views, Purchases, In-Carts in October and November

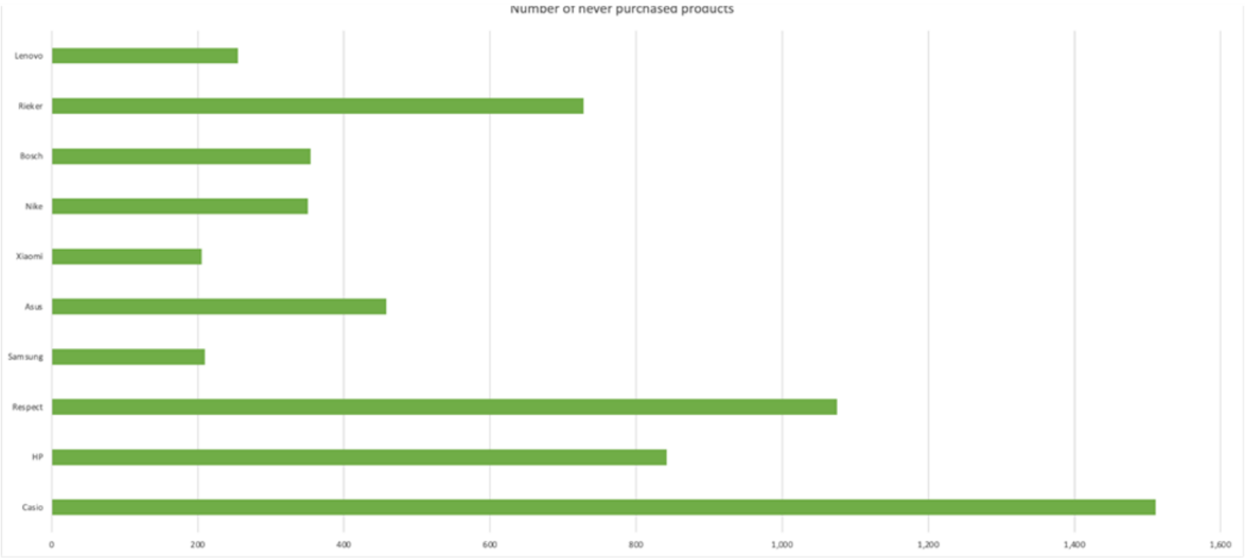


Sum of Sales in both October and November

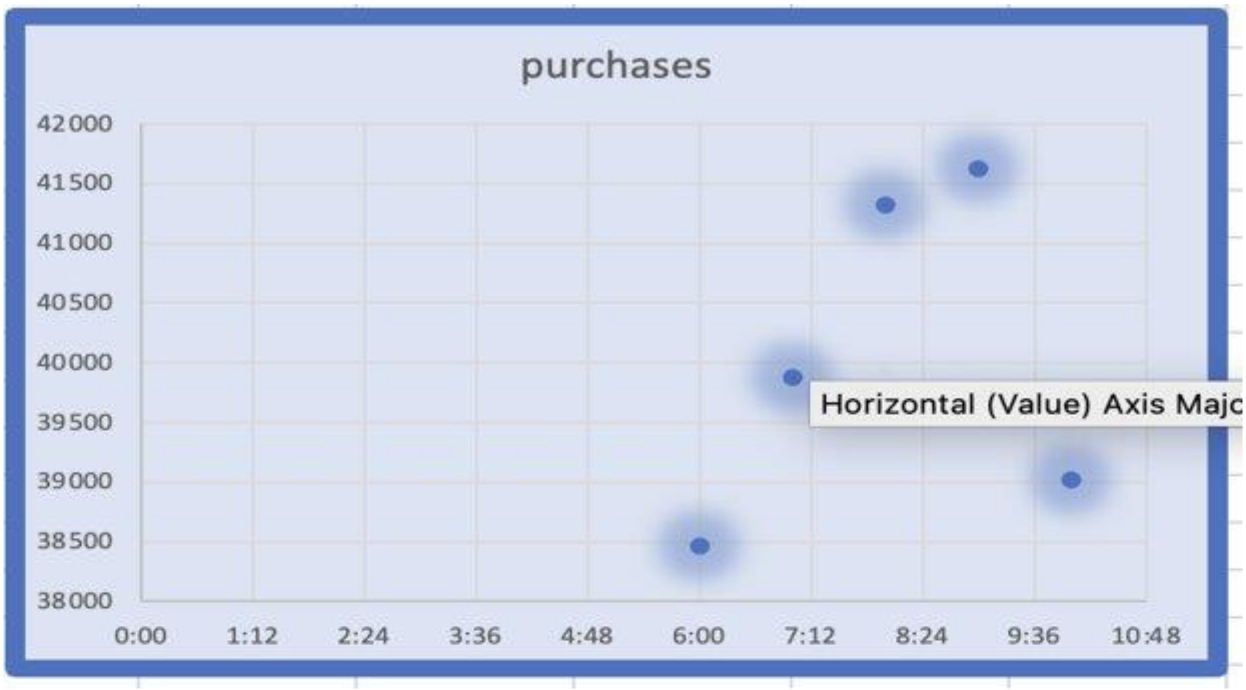


November had $241,560,392 Sum of Sale.

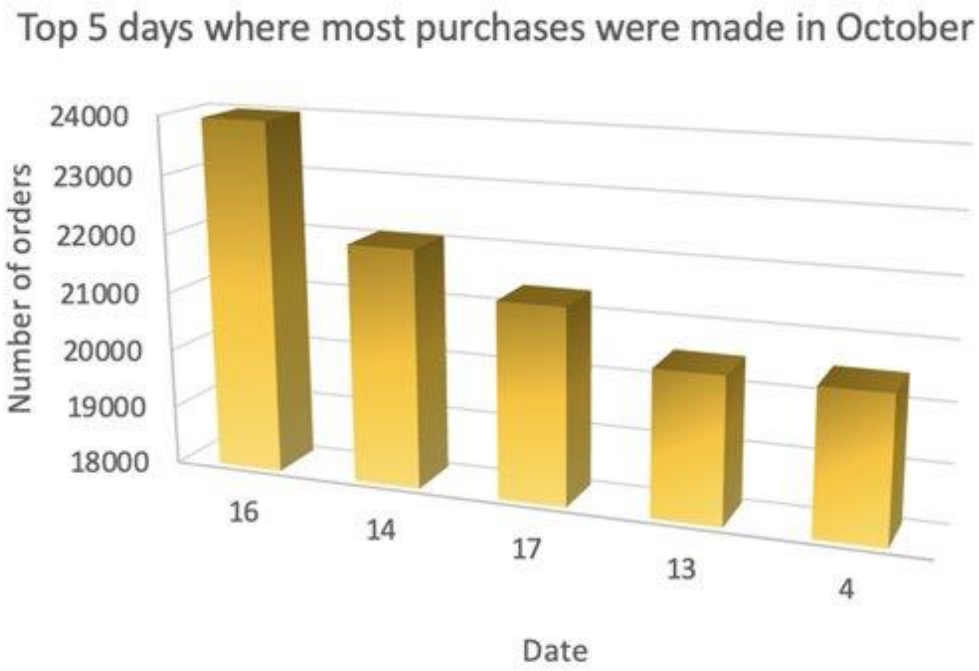
Exit Rate - Most viewed brand but not purchased



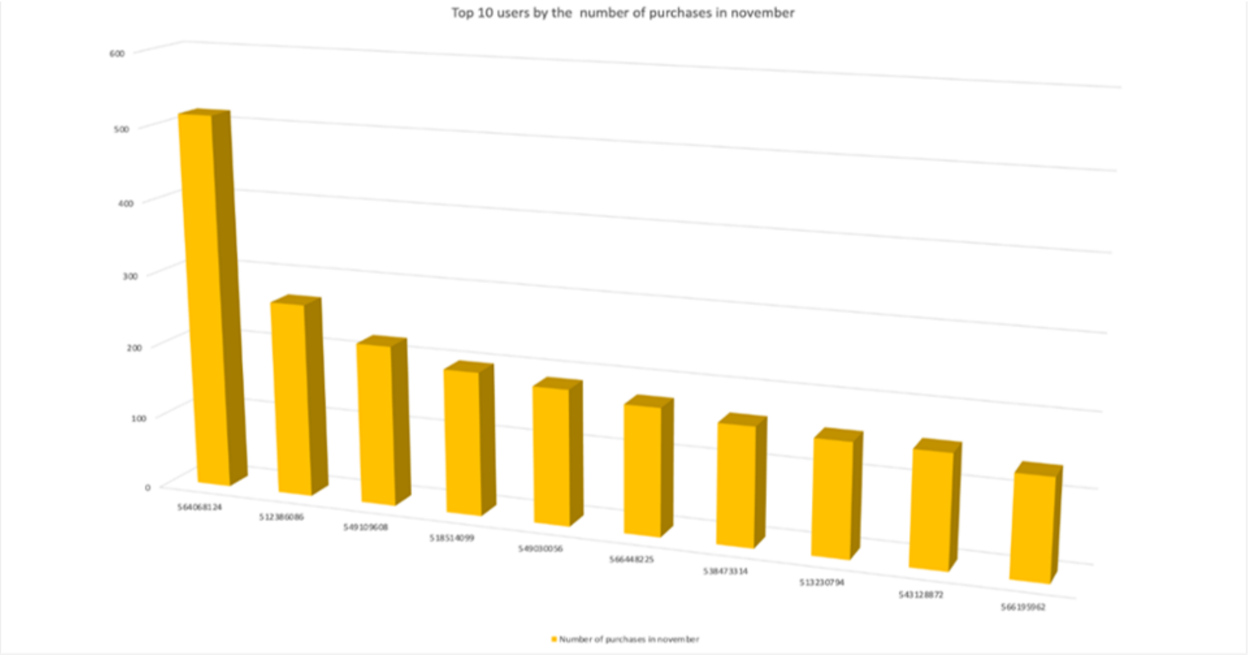
Top 5 hours with most purchases in November



Top 5 days with most purchases in October



Top 10 Users who made the most purchases in November



References

1. URL of Data Source: eCommerce behavior data from multi category store | Kaggle
2. URL of your Github: https://github.com/Lekha19202/E-commerce-customer-behaviour-uding-Hadoop.git
3. URL of References: https://sanyasachdeva1.github.io/Portfolio/files/Analysis%20of%20e-commerce%20behavior%20in%20Multi-Category%20Store.pdf

   https://stackoverflow.com/questions/51097895/hive-sql-find-most-popular-value-across-multiple-columns