CAL STATE LA

CIS 5200

SYSTEM ANALYSIS & DESIGN

ECOMMERCE BEHAVIOR DATA

Project Team 5

Prathima Sarvani

Abhiram Natakarani

Vamsi Reddy

Jeremy Esquivias

AGENDA

Introduction

Dataset Specification

Data Sample

Data Cleaning

H/W Specification

Implementation – Workflow

Data Visualizations

Summary

References

INTRODUCTION

The ecommerce dataset is a valuable resource for data analysts seeking to understand consumer behavior and trends in online shopping.

This dataset contains a vast array of information, including product information, purchase history, and website engagement metrics.

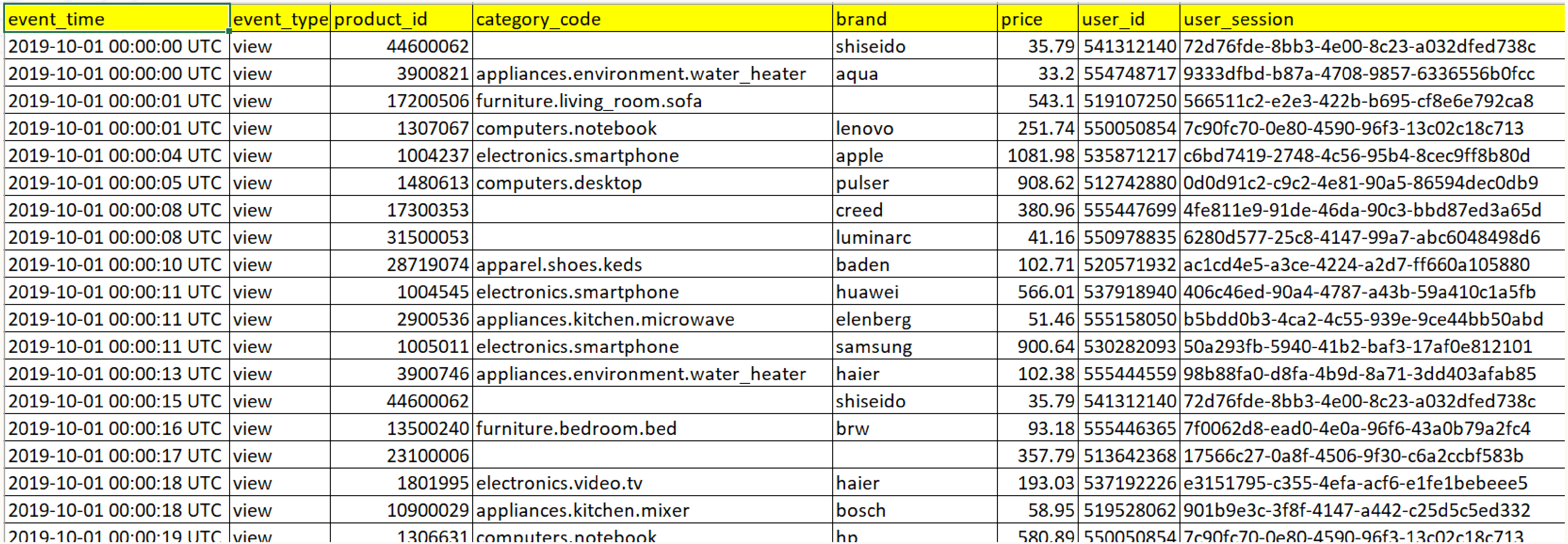
By analyzing this data, businesses can gain insights into customer preferences, identify opportunities for growth, and optimize their online sales strategies.

DATASET SPECIFICATION

- **Data Set** : Ecommerce Behavior Data
- **About** : This file contains behavior data for the month of 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. Each event is like many-to-many relation between products and users.
- **Size** : 5 GB
- **URL** : eCommerce behavior data from multi category store | Kaggle

  Filename: 2019-Oct.csv
- **GitHub** : https://github.com/prathimasarvani/5200-System-Analysis-and-Design

DATASET SAMPLE



event_time	event_type	product_id	category_code	brand	price	user_id	user_session
2019-10-01 00:00:00 UTC	view	44600062		shiseido	35.79	541312140	72d76fde-8bb3-4e00-8c23-a032dfed738c
2019-10-01 00:00:00 UTC	view	3900821	appliances.environment.water_heater	aqua	33.2	554748717	9333dfbd-b87a-4708-9857-6336556b0fcc
2019-10-01 00:00:01 UTC	view	17200506	furniture.living_room.sofa		543.1	519107250	566511c2-e2e3-422b-b695-cf8e6e792ca8
2019-10-01 00:00:01 UTC	view	1307067	computers.notebook	lenovo	251.74	550050854	7c90fc70-0e80-4590-96f3-13c02c18c713
2019-10-01 00:00:04 UTC	view	1004237	electronics.smartphone	apple	1081.98	535871217	c6bd7419-2748-4c56-95b4-8cec9ff8b80d
2019-10-01 00:00:05 UTC	view	1480613	computers.desktop	pulser	908.62	512742880	0d0d91c2-c9c2-4e81-90a5-86594dec0db9
2019-10-01 00:00:08 UTC	view	17300353		creed	380.96	555447699	4fe811e9-91de-46da-90c3-bbd87ed3a65d
2019-10-01 00:00:08 UTC	view	31500053		luminarc	41.16	550978835	6280d577-25c8-4147-99a7-abc6048498d6
2019-10-01 00:00:10 UTC	view	28719074	apparel.shoes.keds	baden	102.71	520571932	ac1cd4e5-a3ce-4224-a2d7-ff660a105880
2019-10-01 00:00:11 UTC	view	1004545	electronics.smartphone	huawei	566.01	537918940	406c46ed-90a4-4787-a43b-59a410c1a5fb
2019-10-01 00:00:11 UTC	view	2900536	appliances.kitchen.microwave	elenberg	51.46	555158050	b5bdd0b3-4ca2-4c55-939e-9ce44bb50abd
2019-10-01 00:00:11 UTC	view	1005011	electronics.smartphone	samsung	900.64	530282093	50a293fb-5940-41b2-baf3-17af0e812101
2019-10-01 00:00:13 UTC	view	3900746	appliances.environment.water_heater	haier	102.38	555444559	98b88fa0-d8fa-4b9d-8a71-3dd403afab85
2019-10-01 00:00:15 UTC	view	44600062		shiseido	35.79	541312140	72d76fde-8bb3-4e00-8c23-a032dfed738c
2019-10-01 00:00:16 UTC	view	13500240	furniture.bedroom.bed	brw	93.18	555446365	7f0062d8-ead0-4e0a-96f6-43a0b79a2fc4
2019-10-01 00:00:17 UTC	view	23100006			357.79	513642368	17566c27-0a8f-4506-9f30-c6a2ccbf583b
2019-10-01 00:00:18 UTC	view	1801995	electronics.video.tv	haier	193.03	537192226	e3151795-c355-4efa-acf6-e1fe1bebeee5
2019-10-01 00:00:18 UTC	view	10900029	appliances.kitchen.mixer	bosch	58.95	519528062	901b9e3c-3f8f-4147-a442-c25d5c5ed332
2019-10-01 00:00:19 UTC	view	1306631	computers.notebook	hp	580.89	550050854	7c90fc70-0e80-4590-96f3-13c02c18c713

DATA CLEANING

- Have segregated the Category Code into three different Category, Subcategory and Product
- **Example:**
- Category Code : appliances.environment.water_heater
- Category: Appliances
- Subcategory: Environment
- Product: Water Heater

- Loaded the complete file into the hive table. Created a new table by eliminating any records having null value in Category and brand.

- CREATE TABLE ECOMMERCE AS SELECT * FROM ECOMMERCE_ORIGINAL WHERE (CATEGORY != '' OR BRAND != '') AND (CATEGORY != '' AND BRAND != '') ;

DATA CLEANING

- Modified Event Time column, added random Date and Time using the below HiveQL

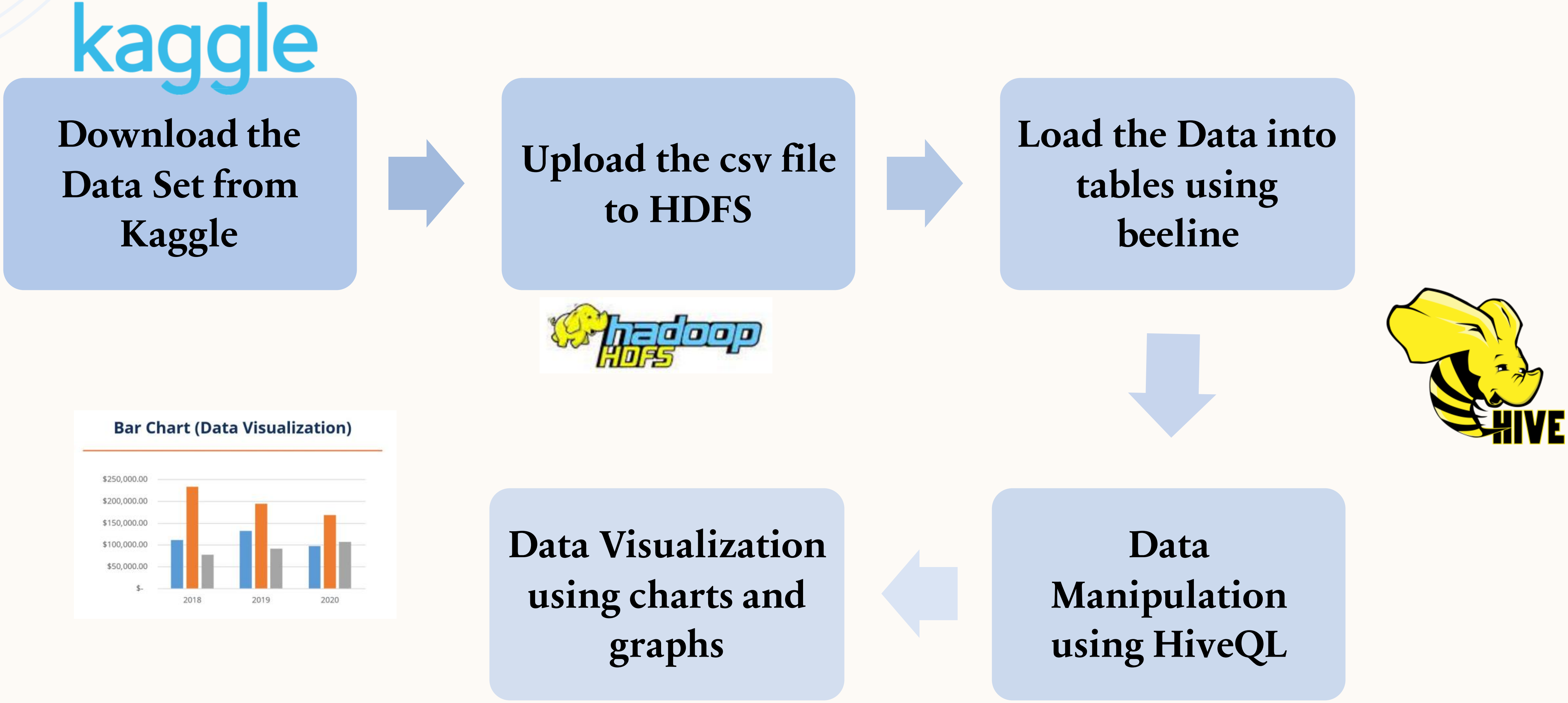
```
FROM_UNIXTIME(UNIX_TIMESTAMP(
 CAST(DATE_FORMAT('2019-10-01','yyyy-MM-dd') AS STRING) || ' ' ||
LPAD(CAST(FLOOR(RAND() * 24) AS STRING), 2, '0') || ':' ||
 LPAD(CAST(FLOOR(RAND() * 60) AS STRING), 2, '0') || ':' ||
 LPAD(CAST(FLOOR(RAND() * 60) AS STRING), 2, '0')) + (CAST(RAND() * 30 AS INT)  * 24 *
60 * 60),'yyyy-MM-dd HH:mm:ss' )
```

- Added location column (Random Top Cities in the world) in the excel using formulas

H/W SPECIFICATION

Cluster Version	Hadoop 3.1.2
Cluster Number of Nodes	2 Master + 3 slave nodes = 5 total
Memory Size	~390GB
CPU Speed	1995.312 MHz
Number of CPU	8

IMPLEMENTATION - WORKFLOW



Download the Data Set from Kaggle

Upload the csv file to HDFS

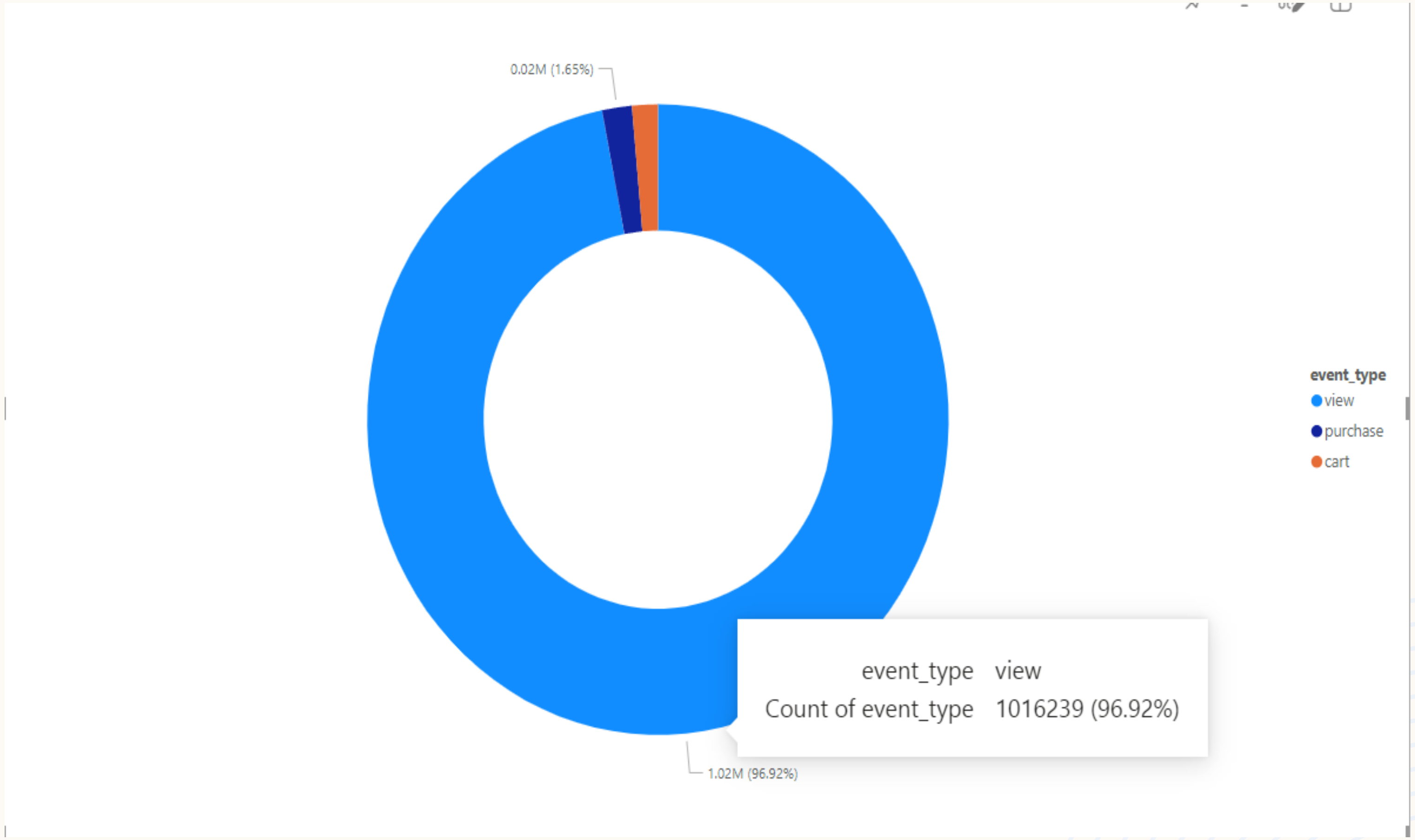
Load the Data into tables using beeline

Data Manipulation using HiveQL

Data Visualization using charts and graphs

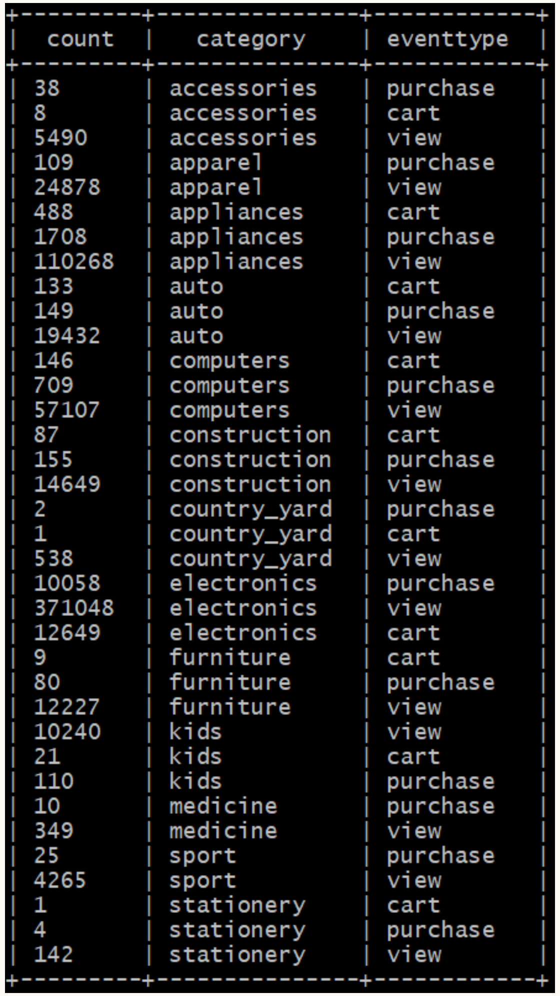
STATISTICS OF VIEWS, CART AND PURCHASES

- This pie chart depicts the distribution of views, cart additions and purchases.
- Here we can say there is a significant difference between the number of views, purchases and cart additions
- There is a low conversion rate from views to purchases which may indicate that there is potential issues in the purchasing process or business needs to improve the user experience.



VIEWS, CART & PURCHASES OF EACH CATEGORY

```
CREATE TABLE STATS_CATEGORY
ROW FORMAT DELIMITED FIELDS TERMINATED BY ','
STORED AS TEXTFILE LOCATION
'ecommerce/tmp/data/stats_category'
AS
SELECT COUNT(*) COUNT, CATEGORY, EVENTTYPE
FROM ECOMMERCE
GROUP BY CATEGORY,EVENTTYPE
ORDER BY CATEGORY;
```



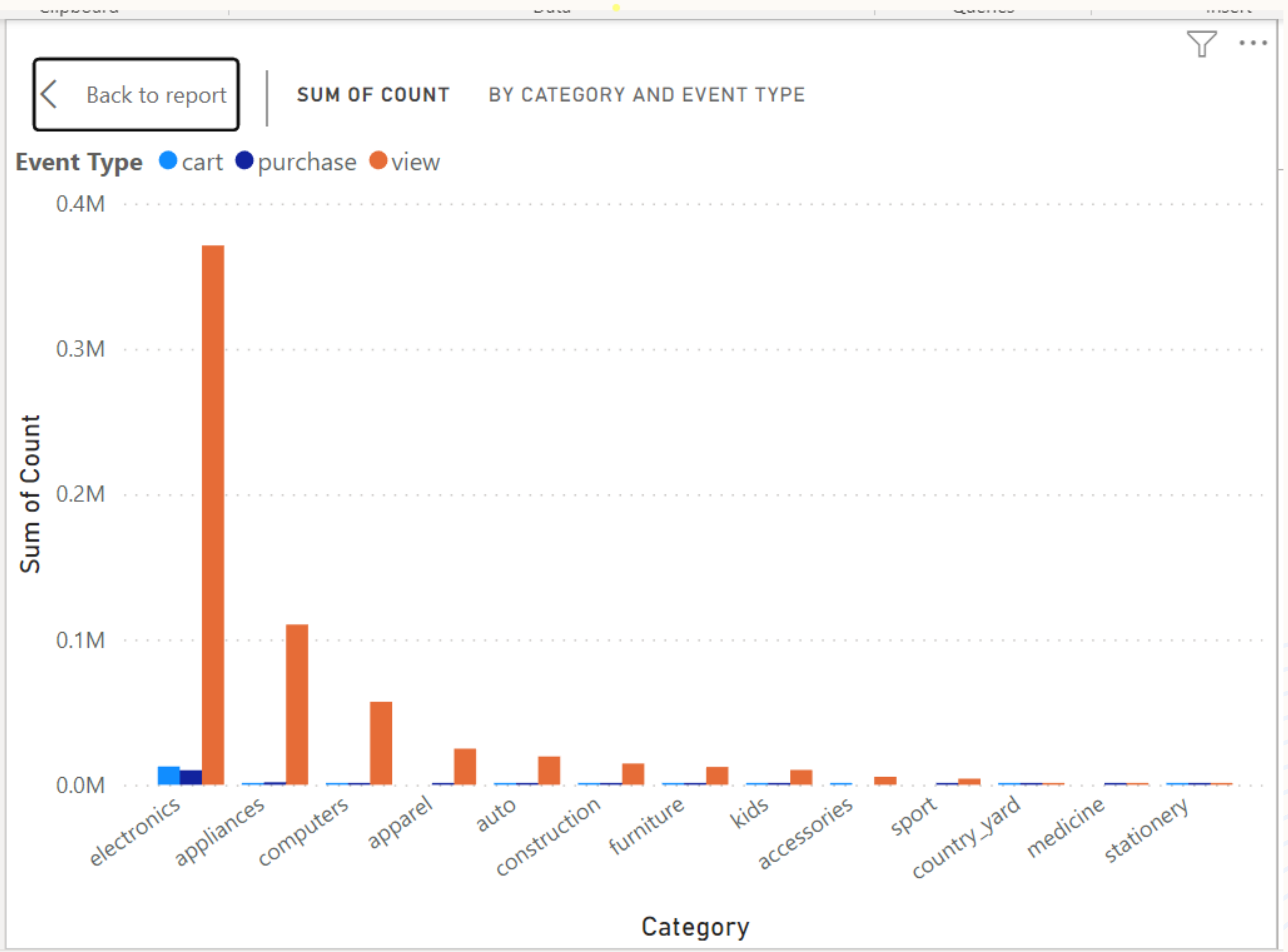
```
+----------+----------------+-------------+
|  count   |    category    |  eventtype  |
+----------+----------------+-------------+
| 38       | accessories    | purchase    |
| 8        | accessories    | cart        |
| 5490     | accessories    | view        |
| 109      | apparel        | purchase    |
| 24878    | apparel        | view        |
| 488      | appliances     | cart        |
| 1708     | appliances     | purchase    |
| 110268   | appliances     | view        |
| 133      | auto           | cart        |
| 149      | auto           | purchase    |
| 19432    | auto           | view        |
| 146      | computers      | cart        |
| 709      | computers      | purchase    |
| 57107    | computers      | view        |
| 87       | construction   | cart        |
| 155      | construction   | purchase    |
| 14649    | construction   | view        |
| 2        | country_yard   | purchase    |
| 1        | country_yard   | cart        |
| 538      | country_yard   | view        |
| 10058    | electronics    | purchase    |
| 371048   | electronics    | view        |
| 12649    | electronics    | cart        |
| 9        | furniture      | cart        |
| 80       | furniture      | purchase    |
| 12227    | furniture      | view        |
| 10240    | kids           | view        |
| 21       | kids           | cart        |
| 110      | kids           | purchase    |
| 10       | medicine       | purchase    |
| 349      | medicine       | view        |
| 25       | sport          | purchase    |
| 4265     | sport          | view        |
| 1        | stationery     | cart        |
| 4        | stationery     | purchase    |
| 142      | stationery     | view        |
+----------+----------------+-------------+
```

VIEWS, CART & PURCHASES OF EACH CATEGORY

- In the electronics category, only 2.5% of users made a purchase.

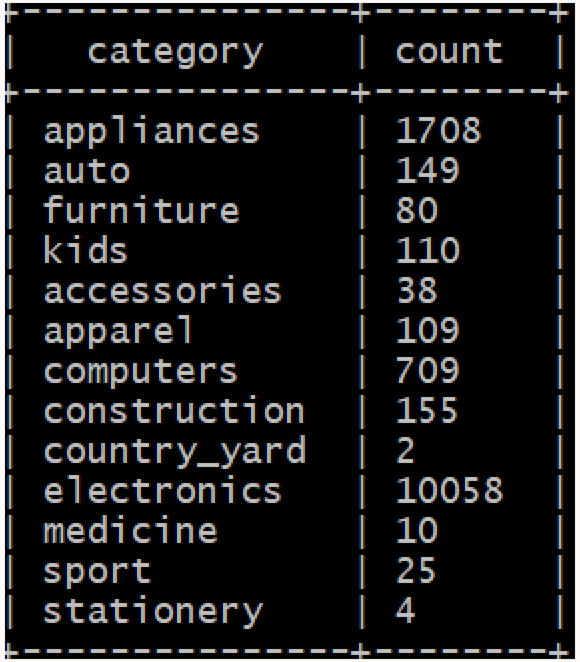
- 3.2% of users in the electronics category added items to their cart.

- Majority of the users in this electronics category, engaged in product views without making a purchase or adding items to their cart.



HIGHLY PURCHASED CATEGORIES

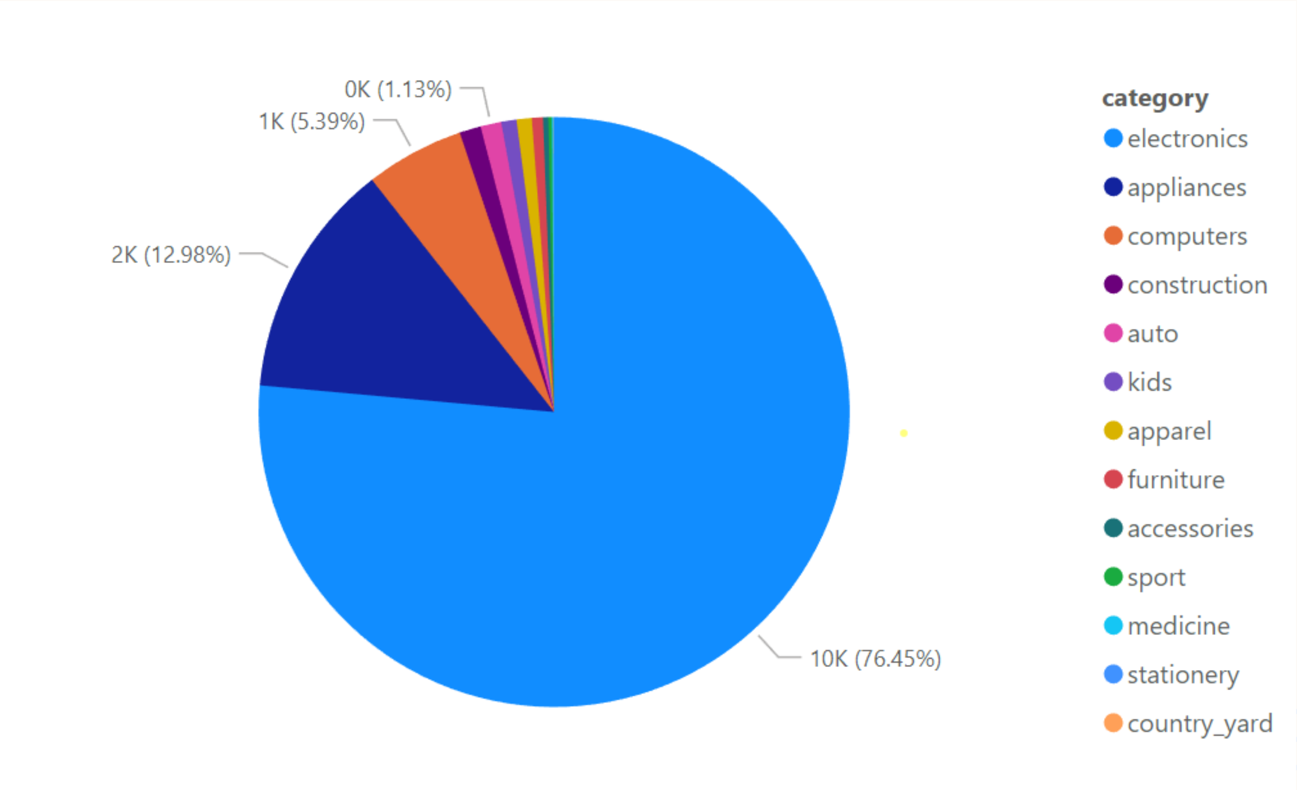
```
CREATE TABLE HIGHEST_SELLING_CATEGORY
ROW FORMAT DELIMITED FIELDS TERMINATED BY ','
STORED AS TEXTFILE
LOCATION  'ecommerce/tmp/data/highest_selling_category'
 AS
SELECT CATEGORY ,COUNT(*) COUNT
FROM ECOMMERCE
WHERE EVENTTYPE = 'purchase'
GROUP BY CATEGORY;
```



```
+----------------+---------+
|    category    |  count  |
+----------------+---------+
| appliances     | 1708    |
| auto           | 149     |
| furniture      | 80      |
| kids           | 110     |
| accessories    | 38      |
| apparel        | 109     |
| computers      | 709     |
| construction   | 155     |
| country_yard   | 2       |
| electronics    | 10058   |
| medicine       | 10      |
| sport          | 25      |
| stationery     | 4       |
+----------------+---------+
```

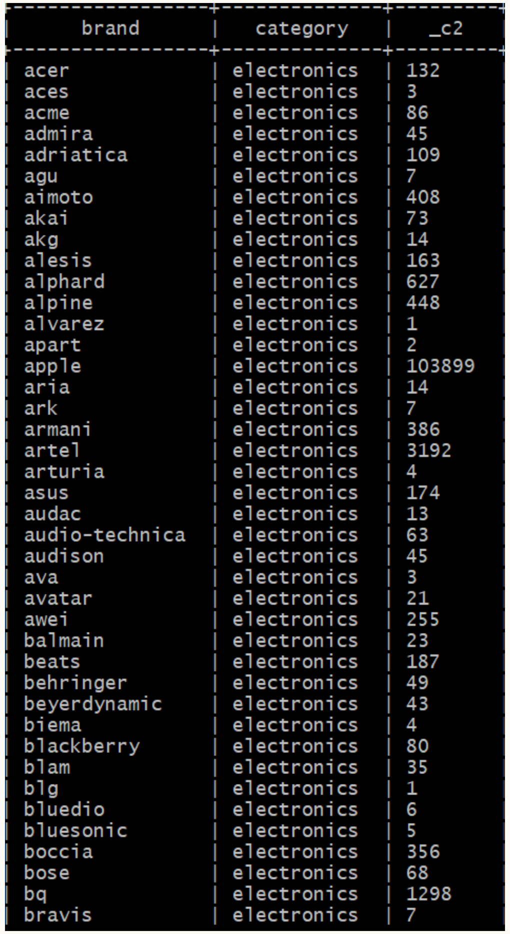
HIGHLY PURCHASED CATEGORIES

- The highly purchased category is electronics and then appliances. These are the categories which are popular among consumers
- These categories generated highest revenue.
- The least purchased category is stationery and country yard.
- This can guide in managing inventory and marketing strategies.
- Also, the Business can understand the market trends.



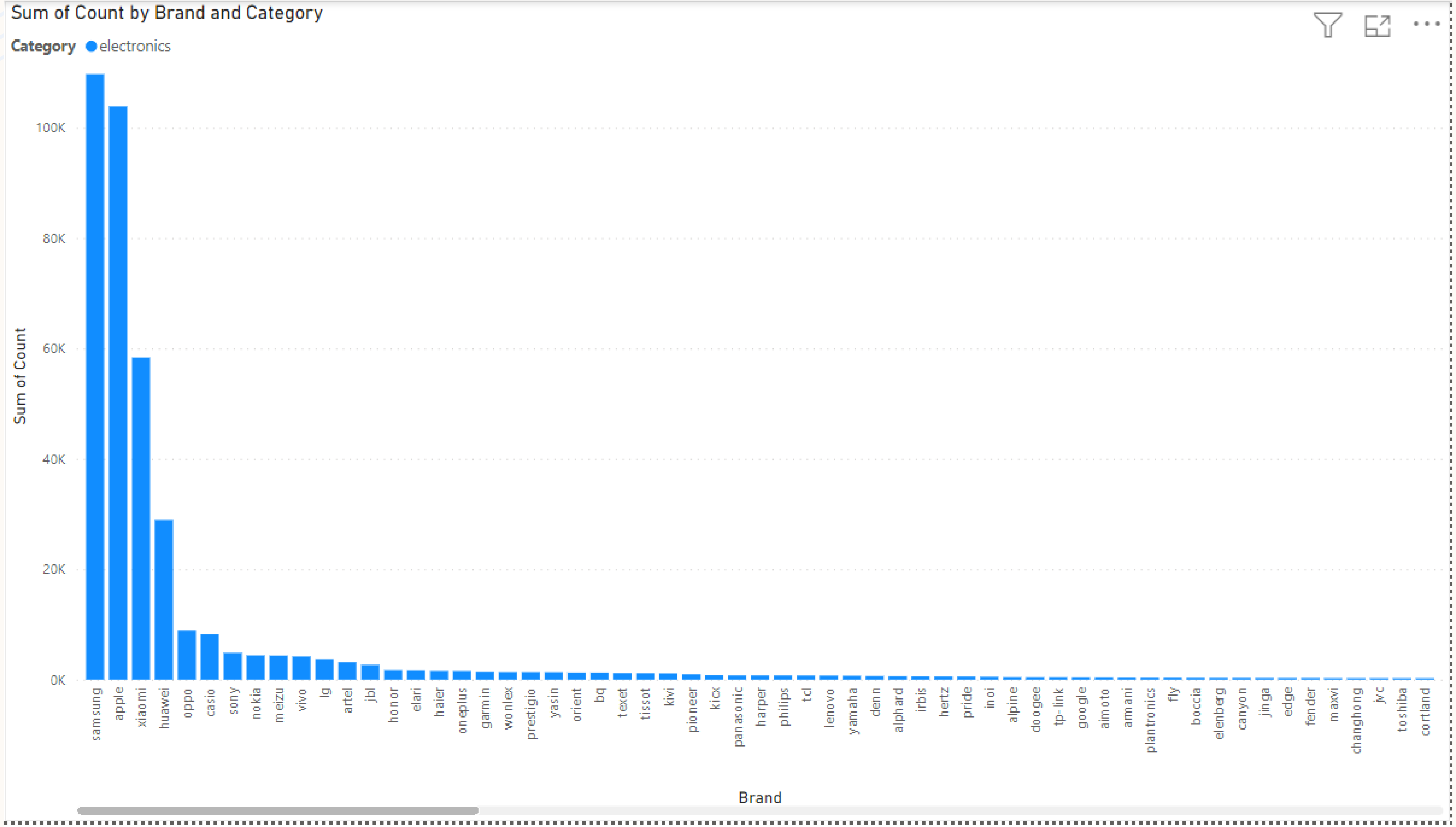
POPULAR BRANDS IN HIGHLY PURCHASED CATEGORY

```
CREATE TABLE BRAND_STATISTICS_HPC
ROW FORMAT DELIMITED FIELDS TERMINATED BY ','
STORED AS TEXTFILE  LOCATION
'ecommerce/tmp/data/brand_stats_hpc'
AS SELECT BRAND, CATEGORY, COUNT(*)  FROM
ECOMMERCE
WHERE CATEGORY IN (
SELECT CATEGORY FROM ECOMMERCE
GROUP BY CATEGORY
ORDER BY COUNT(*)  DESC LIMIT 1)
GROUP BY BRAND , CATEGORY ;
```



```
+------------------+---------------+----------+
|      brand       |   category    |   _c2    |
+------------------+---------------+----------+
| acer             | electronics   | 132      |
| aces             | electronics   | 3        |
| acme             | electronics   | 86       |
| admira           | electronics   | 45       |
| adriatica        | electronics   | 109      |
| agu              | electronics   | 7        |
| aimoto           | electronics   | 408      |
| akai             | electronics   | 73       |
| akg              | electronics   | 14       |
| alesis           | electronics   | 163      |
| alphard          | electronics   | 627      |
| alpine           | electronics   | 448      |
| alvarez          | electronics   | 1        |
| apart            | electronics   | 2        |
| apple            | electronics   | 103899   |
| aria             | electronics   | 14       |
| ark              | electronics   | 7        |
| armani           | electronics   | 386      |
| artel            | electronics   | 3192     |
| arturia          | electronics   | 4        |
| asus             | electronics   | 174      |
| audac            | electronics   | 13       |
| audio-technica   | electronics   | 63       |
| audison          | electronics   | 45       |
| ava              | electronics   | 3        |
| avatar           | electronics   | 21       |
| awei             | electronics   | 255      |
| balmain          | electronics   | 23       |
| beats            | electronics   | 187      |
| behringer        | electronics   | 49       |
| beyerdynamic     | electronics   | 43       |
| biema            | electronics   | 4        |
| blackberry       | electronics   | 80       |
| blam             | electronics   | 35       |
| blg              | electronics   | 1        |
| bluedio          | electronics   | 6        |
| bluesonic        | electronics   | 5        |
| boccia           | electronics   | 356      |
| bose             | electronics   | 68       |
| bq               | electronics   | 1298     |
| bravis           | electronics   | 7        |
```

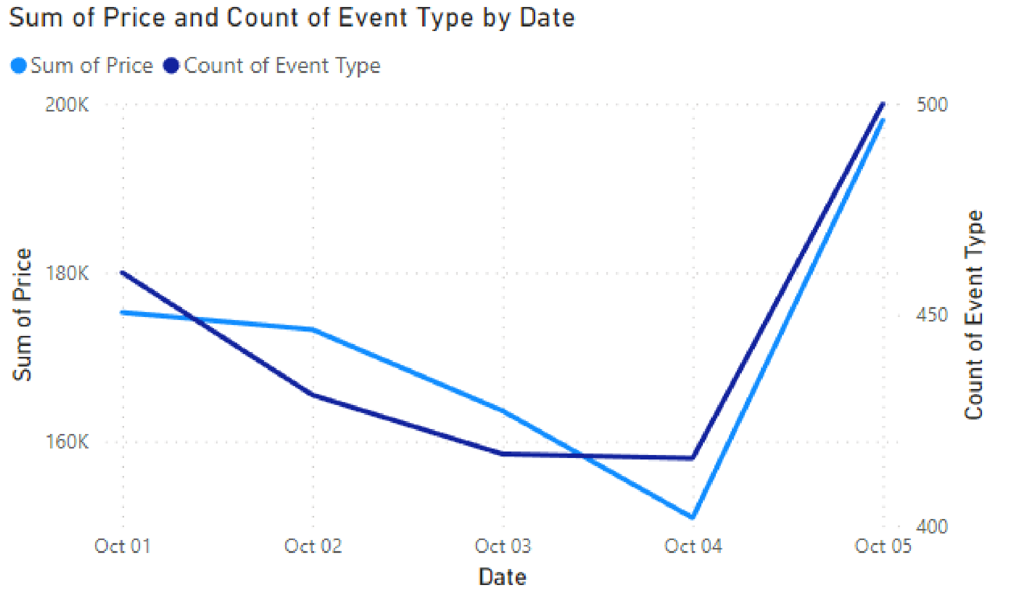
POPULAR BRANDS IN HIGHLY PURCHASED CATEGORY

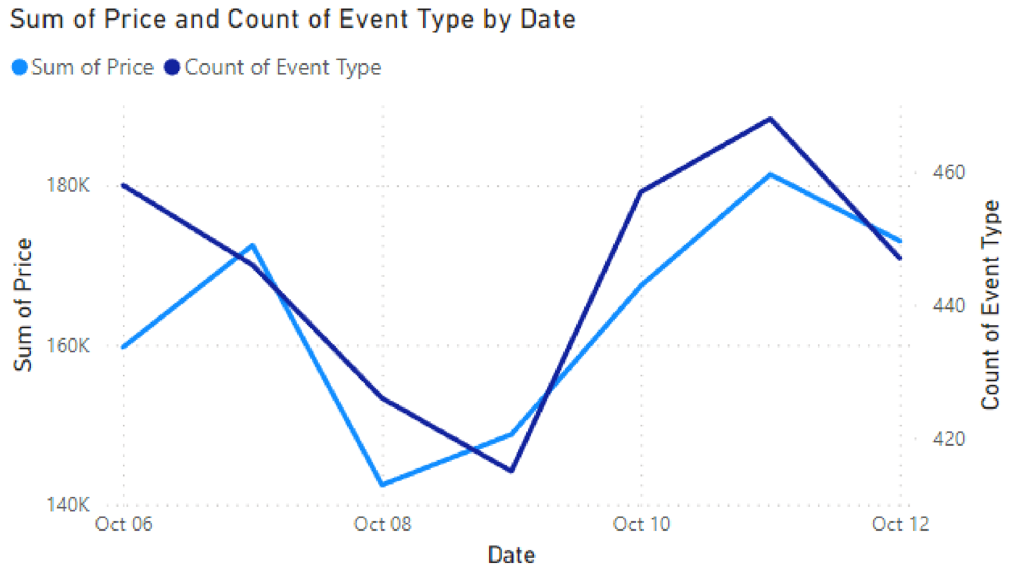


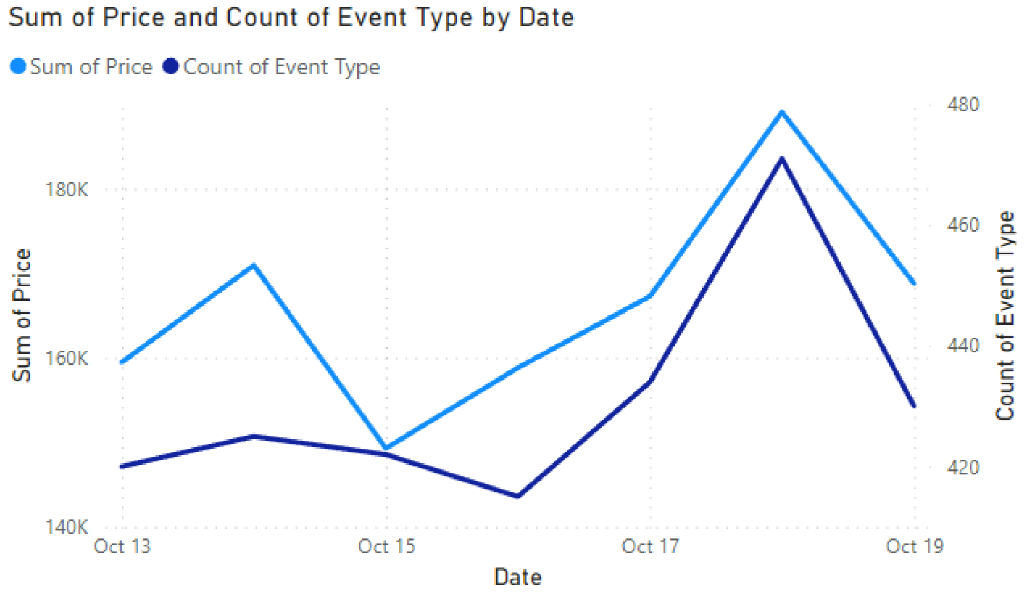
POPULAR BRANDS IN HIGHLY PURCHASED CATEGORY

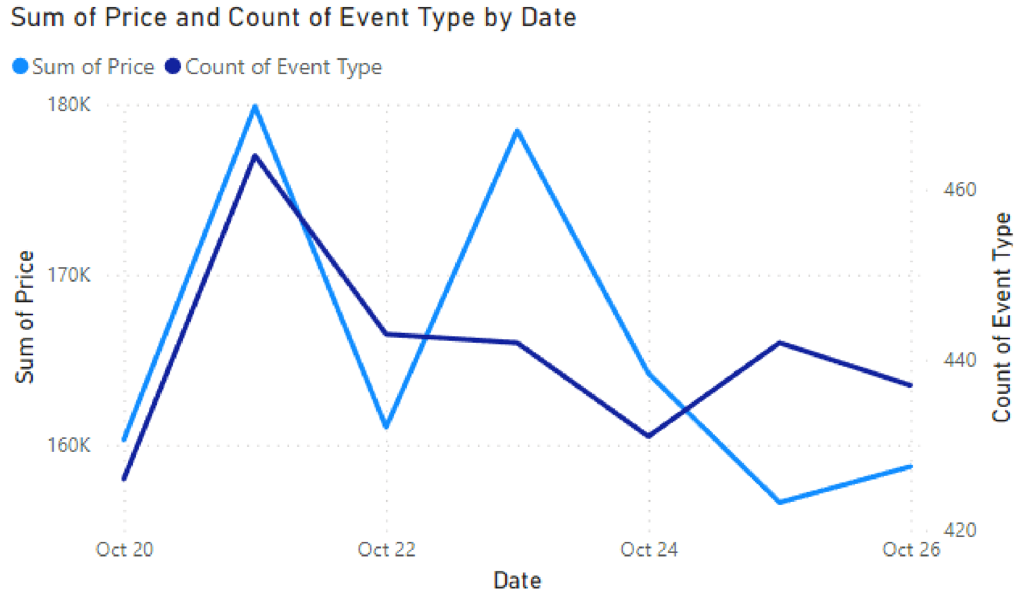
- Samsung and Apple are the highly purchased brands in the highly purchased category electronics. We can say that the customers are likely attracted to the features, quality and reputation associated with these brands.
- This bar chart can help business to get an idea on the customer preferences
- Also, it can guide the business to know about the unpopular brands and introduce gift bundles and marketing strategies to improves the sales.

WEEKLY REPORT – REVENUE (WEEK 1 - 4)





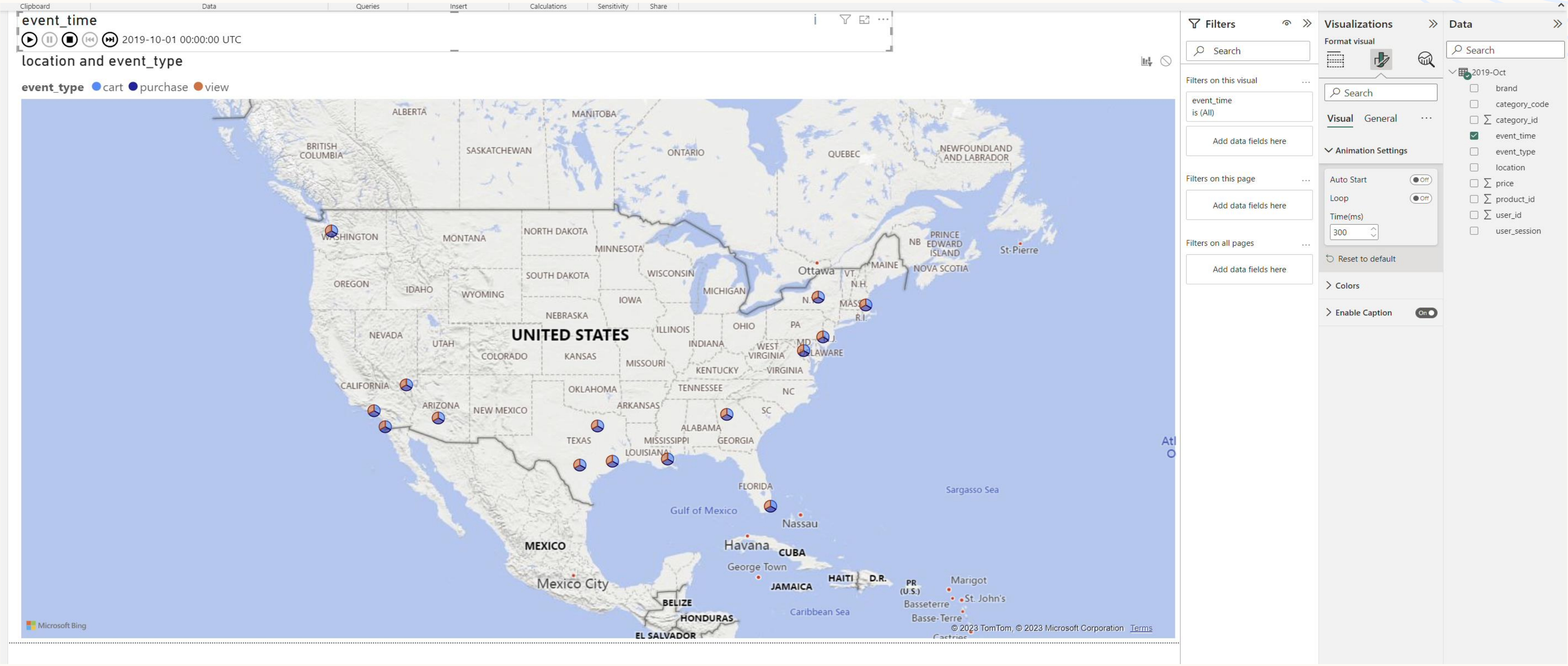




WEEKLY REPORT - REVENUE

- The 4-week line graphs in the Weekly Report's Oct 2019 revenue breakdown illustrate the data from the Purchases.
- The first week reveals that sales decreased on October 4th and increased during the weekend.
- The following week witnessed maybe balanced sales at the beginning of the week, a modest decline, and a recovery by the end.
- Purchase trend in the third week followed the 2nd week .
- The last week of the month had a volatile trend in purchases, with an upswing at first and a sharp decline in sales that persisted through the weekend and concluded with the fewest purchases overall.
- Overall, the month of October in 2019 has seen a purchase trend being volatile over the course making the sales revenue fluctuate significantly.
- In the end, businesses looking for steadiness in their sales revenue faced hurdles due to the unpredictability of the purchasing trend in October 2019. It served as a reminder of the fluidity of consumer behavior and the necessity of taking preventative action to lessen the effects of unstable market conditions.

VISUAL ANALYSIS OF EVENT TYPE BY LOCATION AND TIME



event_time

2019-10-01 00:00:00 UTC

location and event_type

event_type ● cart ● purchase ● view

Filters

Search

Filters on this visual

event_time
is (All)

Add data fields here

Filters on this page

Add data fields here

Filters on all pages

Add data fields here

Visualizations

Format visual

Search

Visual General

Animation Settings

Auto Start Off

Loop Off

Time(ms) 300

Reset to default

Colors

Enable Caption On

Data

Search

2019-Oct

- □ brand
- □ category_code
- □ ∑ category_id
- ☑ event_time
- □ event_type
- □ location
- □ ∑ price
- □ ∑ product_id
- □ ∑ user_id
- □ user_session

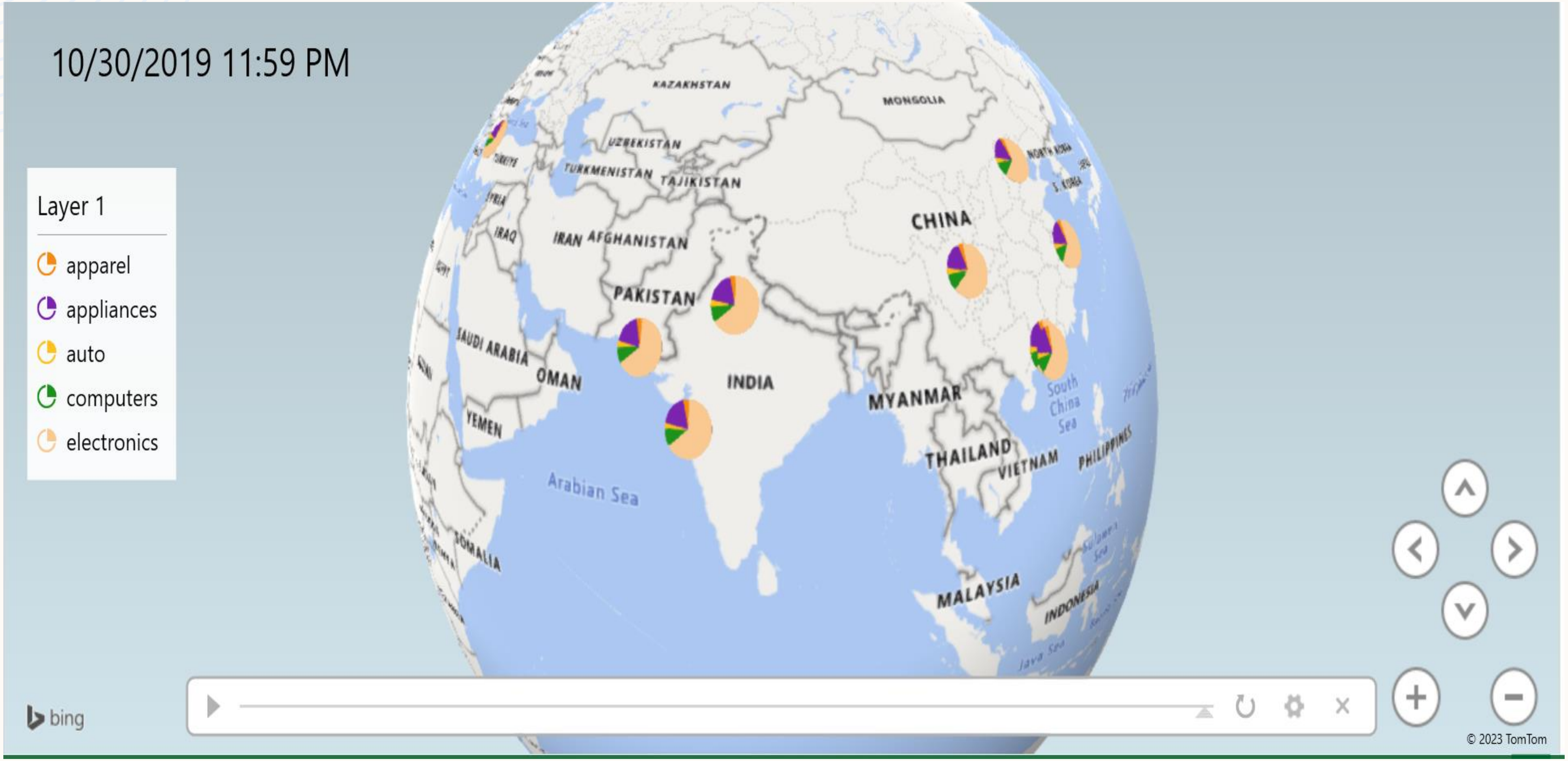
VISUAL ANALYSIS OF EVENT TYPE BY LOCATION AND TIME CONTINUED

- Power BI used for the visualization analysis
- Can get data from the following query:

```
SELECT eventtype, location, eventtime
FROM mcs_org
ORDER BY location, eventtime;
```

- Visual analysis confirms that most people viewed products throughout the day

MAP CHART – EVENT TYPE, LOCATION, TIME, TOP CATEGORIES (EXCEL 3D MAP)



SUMMARY

- This presentation provides an overview of ecommerce data, highlighting the categories with high purchase activity, popular brands within those categories, and the tempo spatial distribution of customer engagement.
- These visualizations can guide decision-making processes, drive revenue growth, and improve customer targeting and satisfaction

REFERENCES

- Lab Tutorial Documents
- https://chat.openai.com/
- Google
- YouTube

THANK YOU