**Analysis of customer shopping behavior on multi category e-commerce platform**

|  |
| --- |
| A Project Report Presented to  DATA-228-11  Fall, 2021 |
|  |

|  |
| --- |
| By |
| Lakshmi Naga Meghana Polisetty [015918708]  Sakshi Jain [015918708]  Sakshi Tongia [015443389]  Snehal Dashrath Karad [015935114] |
|  |

|  |
| --- |
| Copyright © 2021 |
|  |
| ALL RIGHTS RESERVED |

**ABSTRACT**

**Analysis of customer shopping behavior on multi category e-commerce platform**

In today’s world with the advancement in technology, e-commerce has become a vital part of the economy and retail industry. The continued expansion of e-commerce retail has driven consumers of all age groups to use e-retail as a part of their everyday life. The shift in retail sales from brick-and-mortar stores to digital was evident more than ever during the pandemic in 2020. With consumers having a myriad of options to choose from, it has become increasingly difficult for merchants and online shopping portals to engage customers. In such a scenario, it is important for them to deliver a good customer experience. This is where we can leverage the benefit of data that online shopping websites gather by analyzing and drawing meaningful insights. It helps to understand the customers better, recommend products accordingly and increase the conversion rate from a shopper to a customer.  In this project we have analyzed the dataset of a multicategory e-commerce store and made visualization for the key factors affecting the sales like month over month analysis, market basket analysis and trend analysis.

**Acknowledgements**

The authors are deeply indebted to Professor Andrew H. Bond for his invaluable comments and assistance in the development of this project.

Table of Contents

**Chapter 1 Introduction**

* 1. Project goals and objectives
  2. Problem and motivation
  3. Data Description
  4. Project results and deliverables

**Chapter 2 Architecture and High-Level Design**

* 1. ELT Pipeline

2.2 Description

* 1. Tools Used

**Chapter 3 Development Process**

3.1 Data Loading and Transformation

3.2 Data Analysis

3.3 Web Integration

**Chapter 4 Visualization**

4.1. Data Visualization

4.2. Insight 1: Major Categories Sold

4.3. Insight 2: Top subcategories sold

4.4. Insight 3: Sales Ratio per category

4.5 Insight 4: Month over month analysis

4.6 Insight 5: Category RFM

4.7 Insight 6: Activity trend analysis

4.8 Insight 7: User Session Funnel

4.9 Insight 8: Event occurred distribution

**Chapter 5 Collaboration**

5.1. GitHub

5.2. Team Member Id’s

**Chapter 6 Conclusion and Future Work**

6.1 Conclusion

6.2 Challenges Faced

6.3 Future work

List of Figures

Figure 1. Deliverables 1

Figure 2. ELT Pipeline 2

Figure 3. Data Transformation 4

Figure 4. Data Transformation 2

Figure 5. Athena 5

Figure 6. AWS Redshift 6

Figure 7. Tableau Public Extraction 6

Figure 8. Web Integration 7

Figure 9. Website 8

Figure 10. Major Categories Sold 9

Figure 11. Top Categories Sold 10

Figure 12. Sales Ratio 10

Figure 13. MoM 11

Figure 14. Category RFM 12

Figure 15. Activity Trend Analysis 12

Figure 16. User Session Funnel 13

Figure 17. Event Occurrence 13

# Introduction

* 1. **Project goals and objectives**

With our in-depth analysis of the dataset, we aim to gain insights that will be useful for business to make strategic decisions. In this data driven world, we can leverage the benefit of data by analyzing and drawing meaningful insights. These insights will help to understand the customers better, target the right audience, better product recommendations, increase in the conversion rate and get appropriate merchandise. All the analysis will make one understand the overview of the transactions, movement of customers on the website and take appropriate actions to increase the revenue and provide better services.

**1.2** **Problem and motivation**

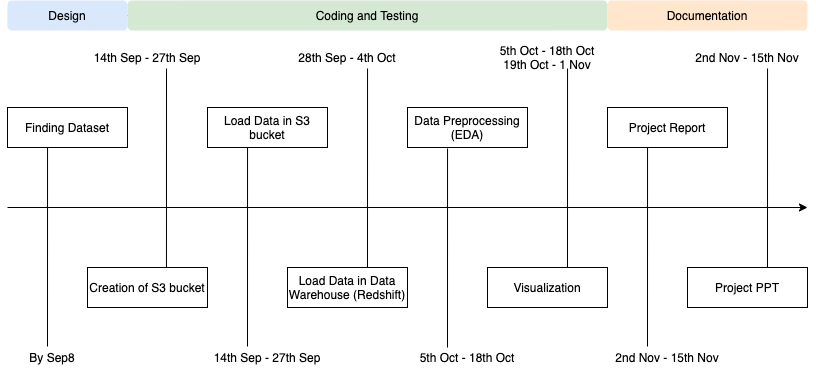
These days e-commerce plays a key role in retail. E-commerce has amplified especially during the time of pandemic. For any business, it is crucial to understand the sales and to know their customers to come up with appropriate business strategies to increase revenue. To leverage the data to understand various aspects of the business, in depth analysis of the customer buying behavior, merchandise, depth and breadth of the category plays a significant role. This has led us to choose this topic of analyzing the customer behavior for e-commerce website and come up with insights on the key aspects of the business.

**1.3** **Data Description**

In this project, we chose a dataset from a Multicategory e-commerce store (Source: Kaggle). The dataset comprised of 285 million rows of users shopping activity from Oct 2019 – April 2020. The size of the available data is 30GB, but we used 15GB of it. The data consists of columns describing product, product category, brand, price, user details, user session details, type of event- view, add to cart and purchase.

**1.4.** **Project results and expected deliverables**

Fig. 1. Deliverables

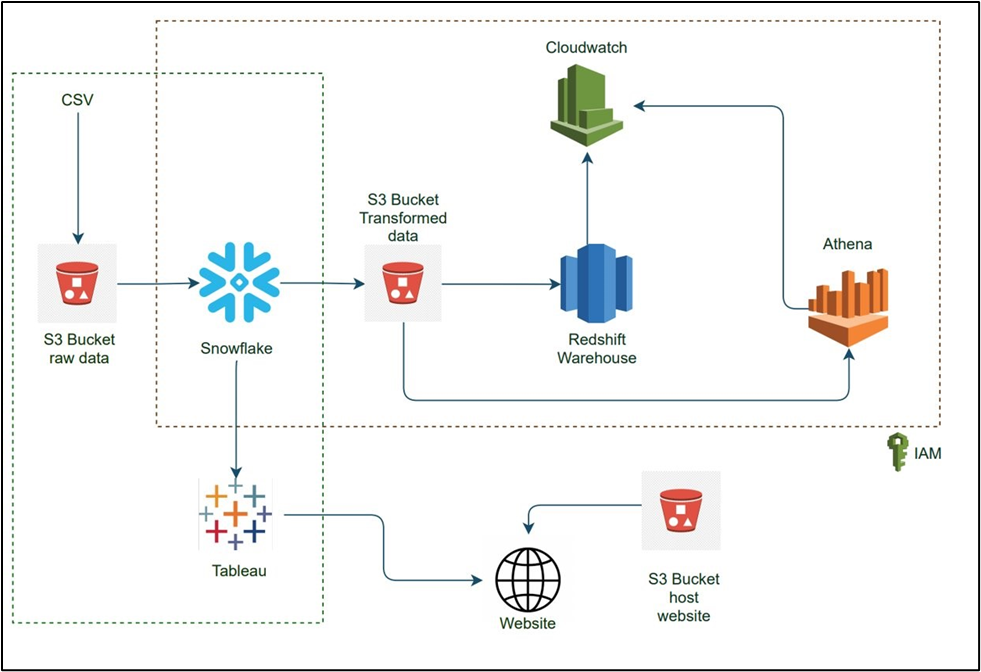


# Chapter 2 Architecture and High-Level Design

* 1. **ELT Pipeline**

The following was the ETL pipeline architecture followed to carry out analysis on the selected dataset.

.



**Fig. 2. ELT Pipeline**

* 1. **Description**

We used IAM access for the project. We used one of the public data from Kaggle. The data was available in csv format. This data was uploaded to S3 bucket. The data from S3 was then fetched into Snowflake for ELT. Data was transformed into Snowflake. The transformed data was uploaded back into S3 in CSV format.

Data from S3 was uploaded into Redshift, where Redshift is our data warehouse. Chunks of data(categorical) was then unloaded from Redshift into S3. We connected Athena to S3 for querying purposes and used CloudWatch to monitor logs from Redshift and Athena.

The other part of the architecture is where we are performing visualization using the data from Snowflake, here Snowflake acts as our data warehouse. We connected Snowflake to Tableau to use this transformed data for visualization. We used S3 bucket to host a static website to upload the visualizations from Tableau to a web server.

* 1. **Tools Used**

We used Snowflake for data transformation, Snowflake uses SQL and has a user-friendly interface. Also, there was no cost associated with using Snowflake as compared to Amazon Glue.

The visualization generated through Tableau is more appealing than compared to other tools available in the market. We could connect Tableau easily to Snowflake and there was a free desktop version available for use.

We used AWS Athena to perform querying operations to extract meaningful data from the dataset. Performing SQL querying on a huge data warehouse containing humongous volume of data is both not cost effective as well as performance effective. Athena provides agility and flexibility in querying the data and at the same time is cost effective.

|  |  |
| --- | --- |
| **Name** | **Usage** |
| Snowflake | Data Loading and Transformation |
| Tableau | Data Visualization |
| AWS Athena | Analytical Queries |
| AWS Redshift | Data Loading |

# Chapter 3 Development Process

**3.1.** **Data Loading and Transformation**

We performed Exploratory Data Analysis and ELT on the dataset which consisted of removing/replacing missing values, removing duplicates, handling outliers, handling correlated columns.

After carefully analyzing the dataset the following cleaning and transformation were made to the dataset:

* The event\_time column had a timestamp for the events but had ‘UTC’ with the timestamp so that needed to be dropped for it to be changed to the timestamp datatype.
* All the columns were changed to the relevant datatypes.
* Column Category was split into three new columns as the column category had main category, subcategory and the special category listed in one column with a delimiter ‘.’

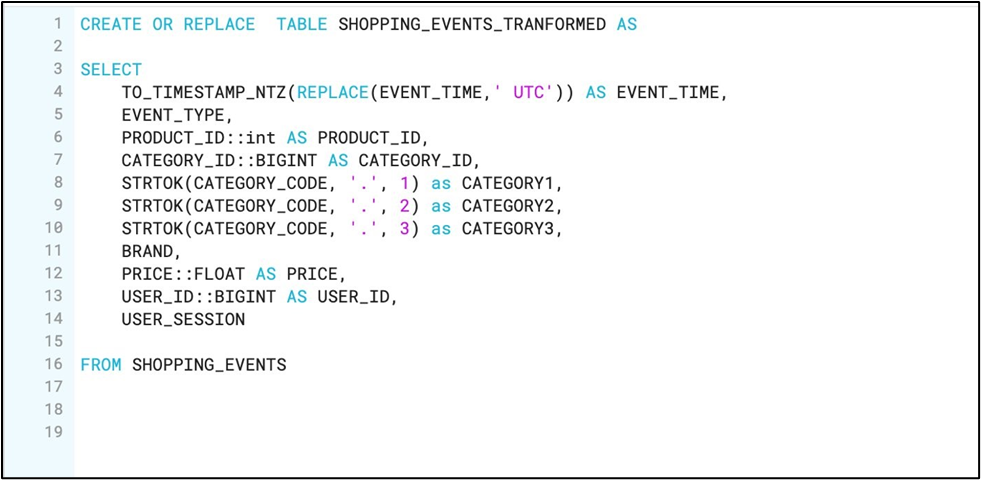


Fig. 3. Data Transformation

Once the transformations were done to the data, the new table shopping\_events\_transformed was created. The figure below shows the code for shifting the transformed data back into S3 bucket so that required AWS Athena functions and queries can be performed on it.

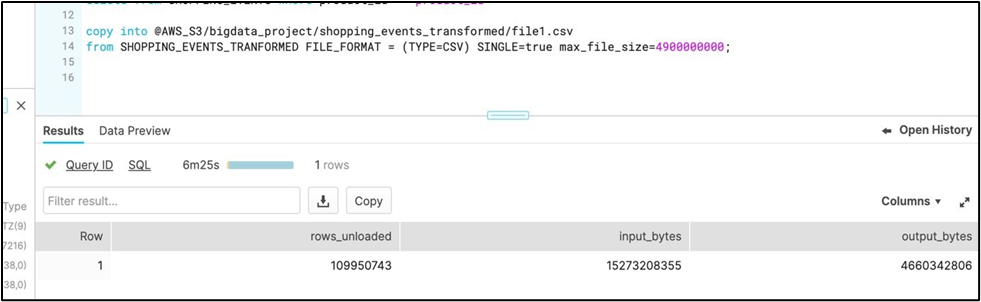


Fig. 4. Data Transformation

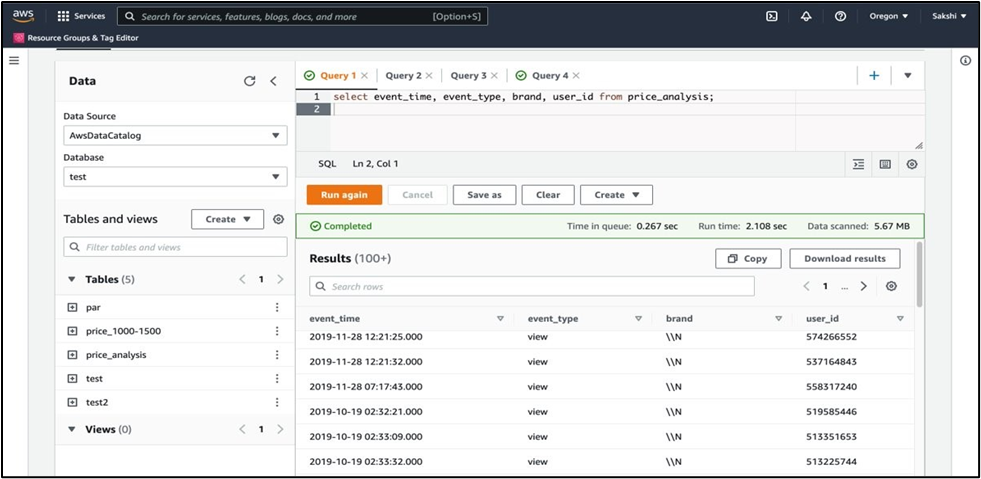
**3.2.** **Data Analysis**

When we use Amazon Simple Storage Service (Amazon S3) data lakes and Amazon Redshift as our data warehouse, the combination of using these two is termed as a data lake house approach. In the real-world scenario, companies deal with huge volumes of data and as we know, Amazon Redshift Spectrum does an excellent job of unifying our data lake and data warehouse.

However, performing some SQL querying on a huge data warehouse containing humongous volume of data is both not cost effective as well as performance effective.

Athena an interactive query service, is integrated as it is easy to use which simply points to our data in Amazon S3, defines the schema, and start querying using standard SQL. Most results are delivered within seconds, and we pay only for the queries that we run.

Fig. 5. Athena



We are dealing with a huge data of 15GB. Using Amazon Redshift commands, we have unloaded chunks of data from Amazon Redshift based on a pricing category (0-500,500-1000 etc.) in Parquet files. The output files are placed in S3. We used Parquet file format here since they offer reduced storage costs, and improvement in reading data when compared to other file formats.

Using Athena query service, we have performed a few meaningful queries on the categorical pricing data. Athena is integrated, as it provides agility and flexibility in querying our data.

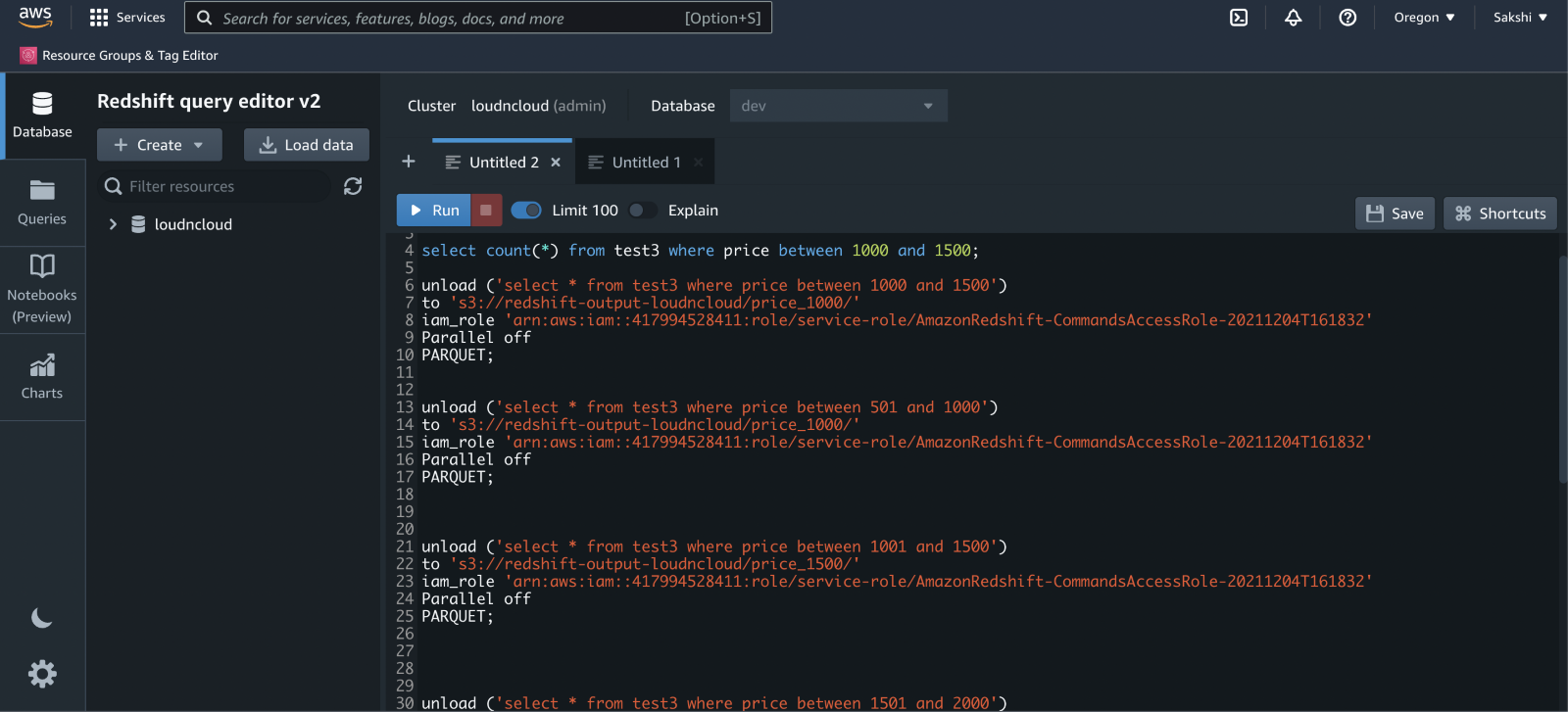


Fig. 6. AWS Redshift

**3.3.** **Web Integration**

The web integration was done using a public tableau server and hosting the website on S3 bucket.

Data was extracted and was uploaded along with visual models to tableau public.

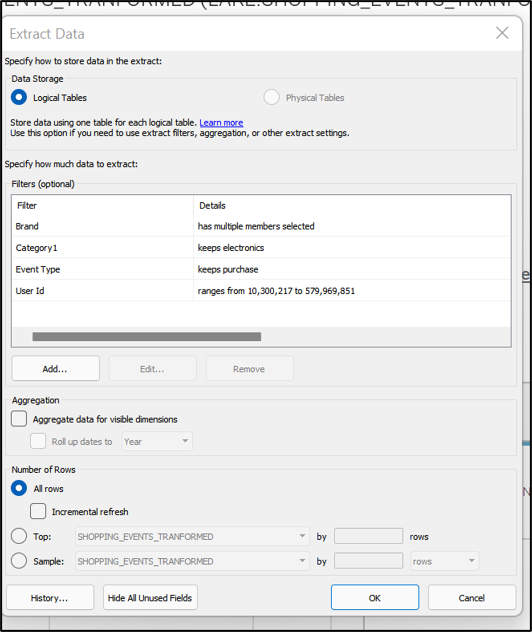


Fig. 7. Tableau Public Extraction

The server URL was generated from this public server and was used in the java script file of the website code as a part of linking.

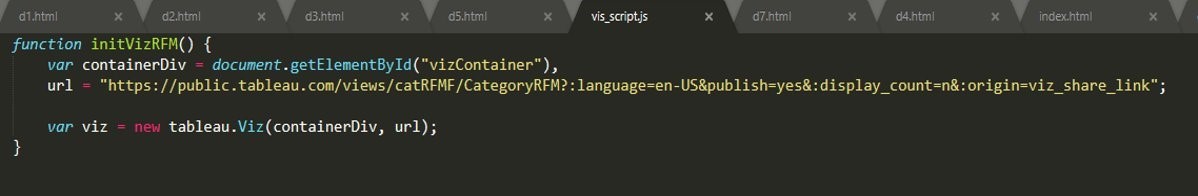


Fig. 8. Web Integration

This website code was then hosted on S3 bucket.

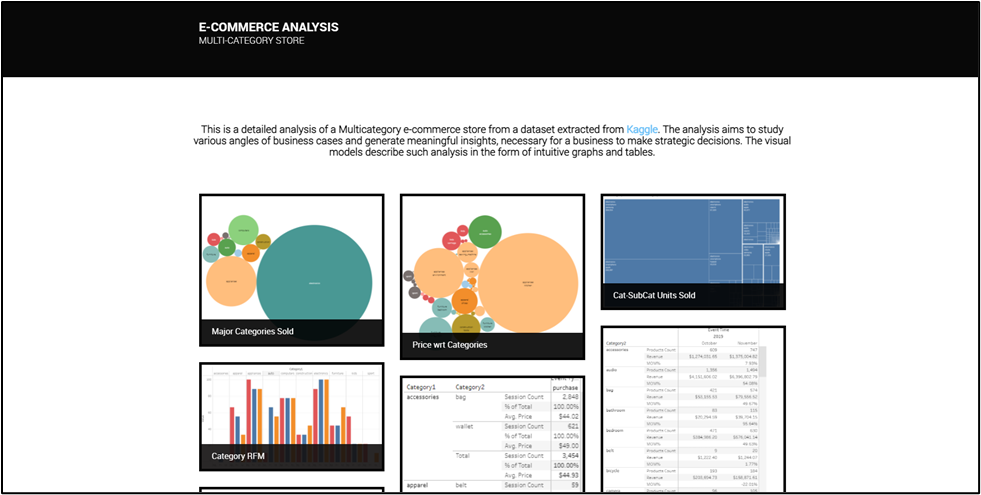
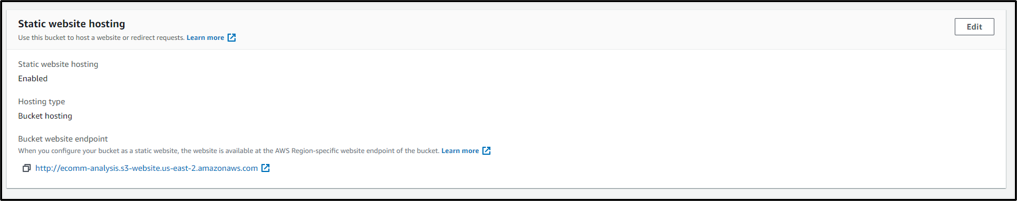
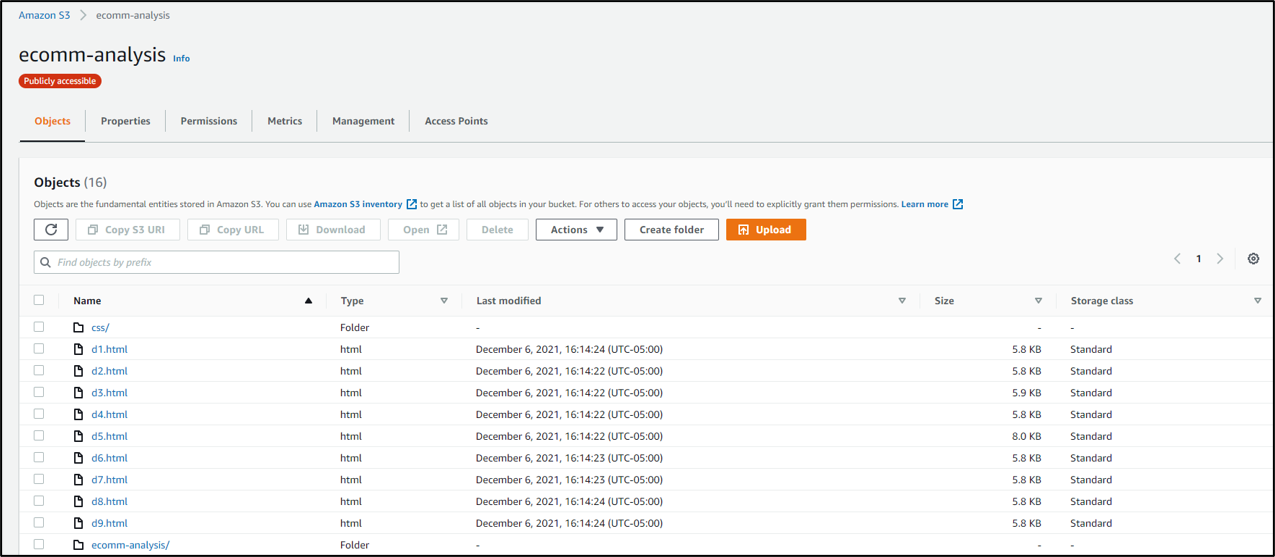


Fig. 9. Website

**Static Website**: https://us-west-1.console.aws.amazon.com/console/home?region=us-west-1

# Chapter 4 Visualization

**4.1**  **Data Visualization**

We performed data visualization to identify patterns in the data and generate insights related to segment customers based on amount spent, category of product purchased, time spent on the website, comparison in customer shopping behavior. Our project aimed to do an analysis for gaining insights into customer shopping behavior, not limited to customer segmentation. The visualizations were created by connecting transformed data in Snowflake warehouse to Tableau. The following key insights were generated after prior analysis.

* 1. **Insight 1: Major Categories Sold**

This visualization shows that electronics is the most sold category on the website followed by appliances and computers.

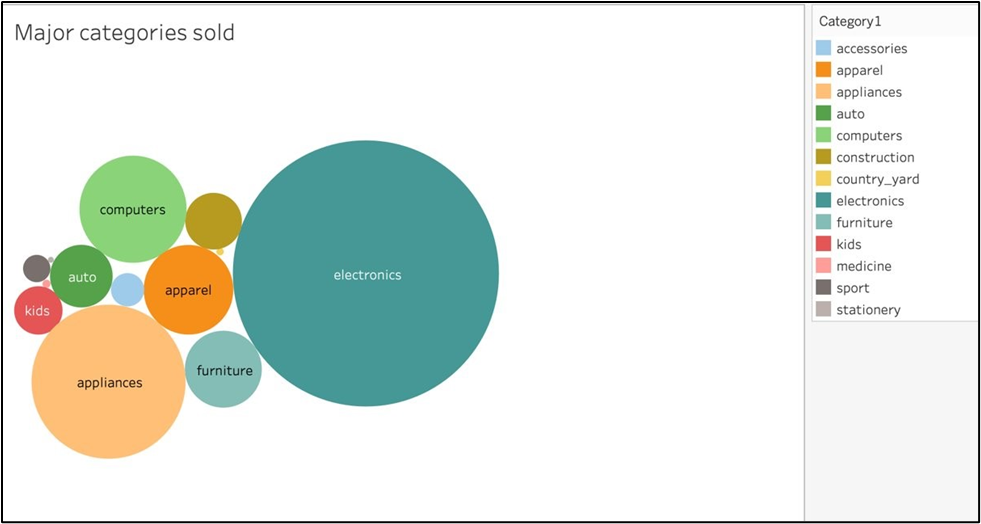


Fig. 10. Major Categories Sold

* 1. **Insight 2: Top subcategories sold**

This Visualization shows the Subcategories and brands in the Electronics Category with brand name and units sold. It clearly shows that in electronics Smartphones are doing the best for the brand Samsung followed by Apple.

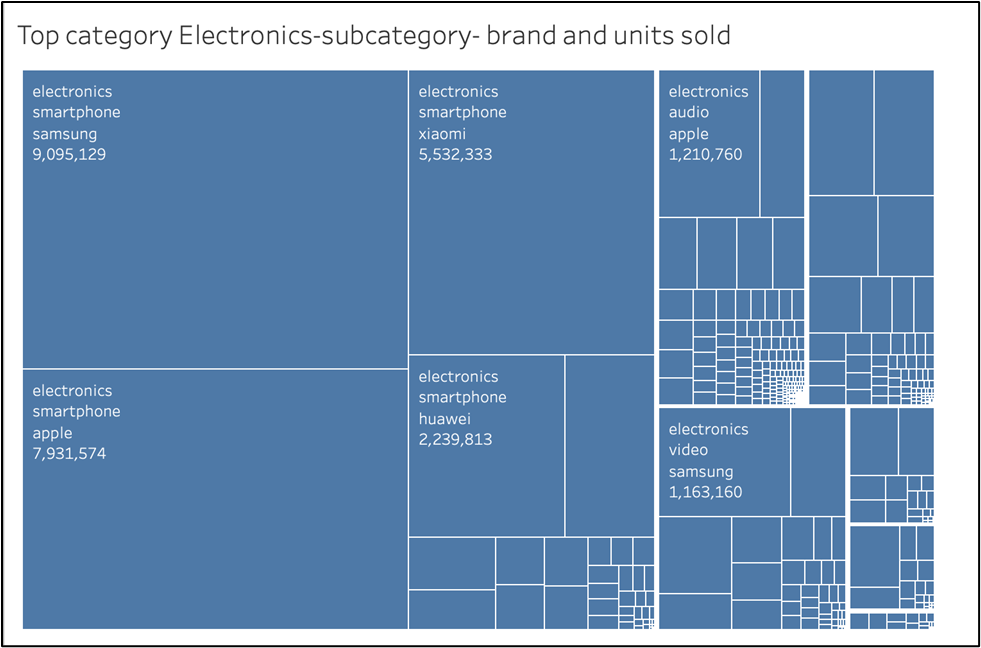


Fig 11. Top Categories Sold

* 1. **Insight 3: Sales Ratio per category**

The visualization below shows the ratio of sales based on monetary value for categories other than electronics (cell phones, laptops, headphones, camera, etc.)

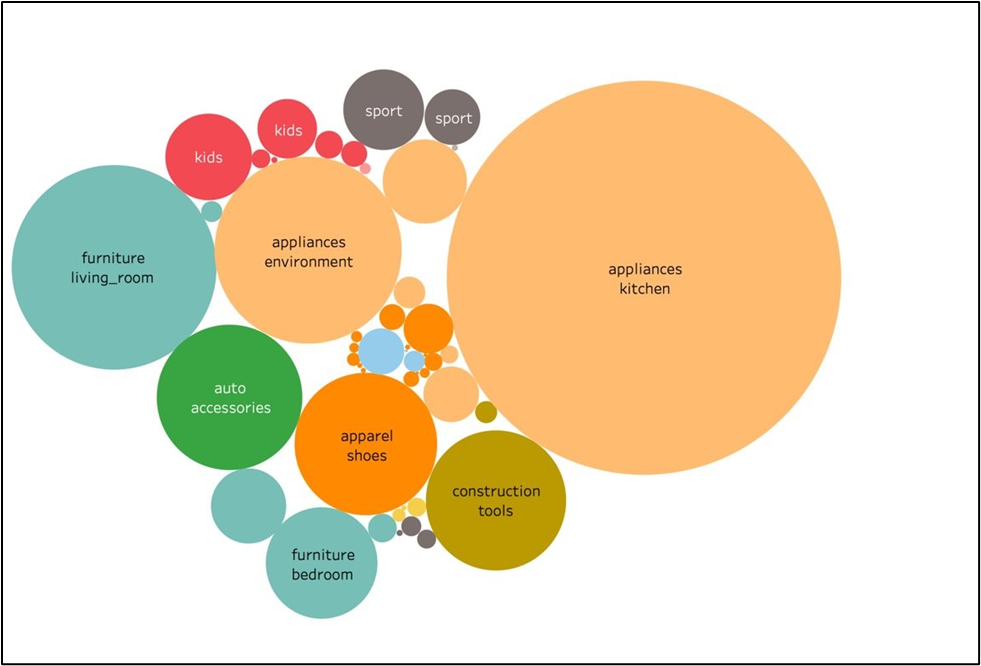


Fig. 12. Sales Ratio

* 1. **Insight 4: Month over month analysis**

This visualization shows Month over month analysis of all the major categories for the two months of October and November of 2019. For every month it shows the number of products sold and revenue generated from that category per month. MoM% shows the change in percentage of sales as compared to the previous month.

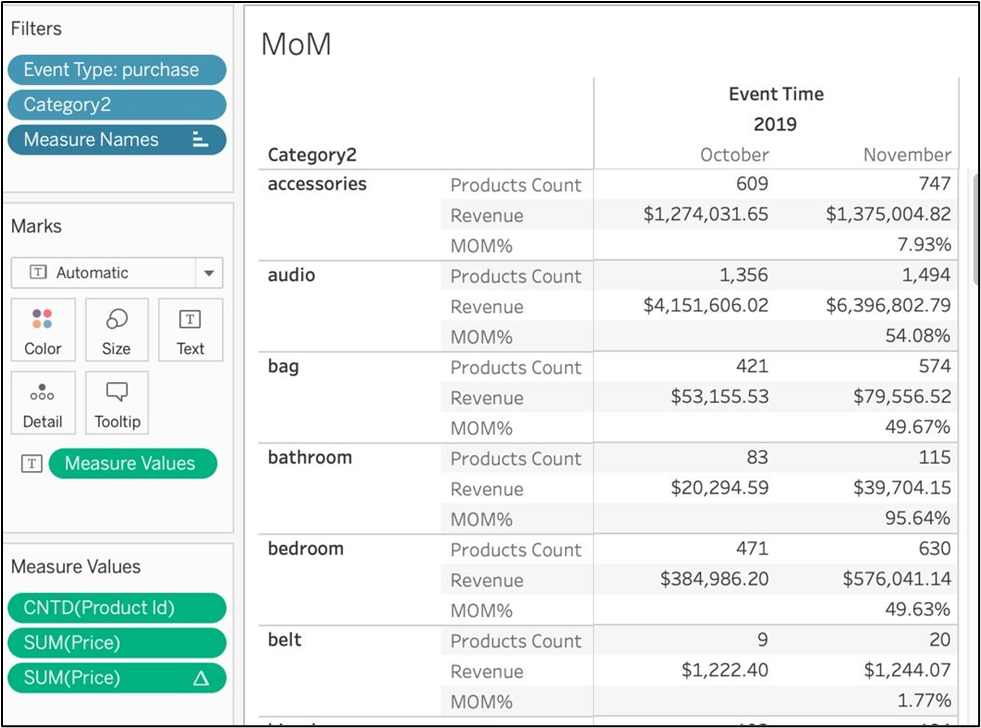


Fig. 13. MoM

* 1. **Insight 5: Category RFM**

The visualization shows the Recency, Frequency, and monetary analysis for the various categories. This visualization indicates that the recency is highest in category appliances whereas frequency and monetary is highest in category Electronics. It can also be deduced that categories accessories are not performing good on the website as its recency and monetary percentile as compared to other categories is nil.

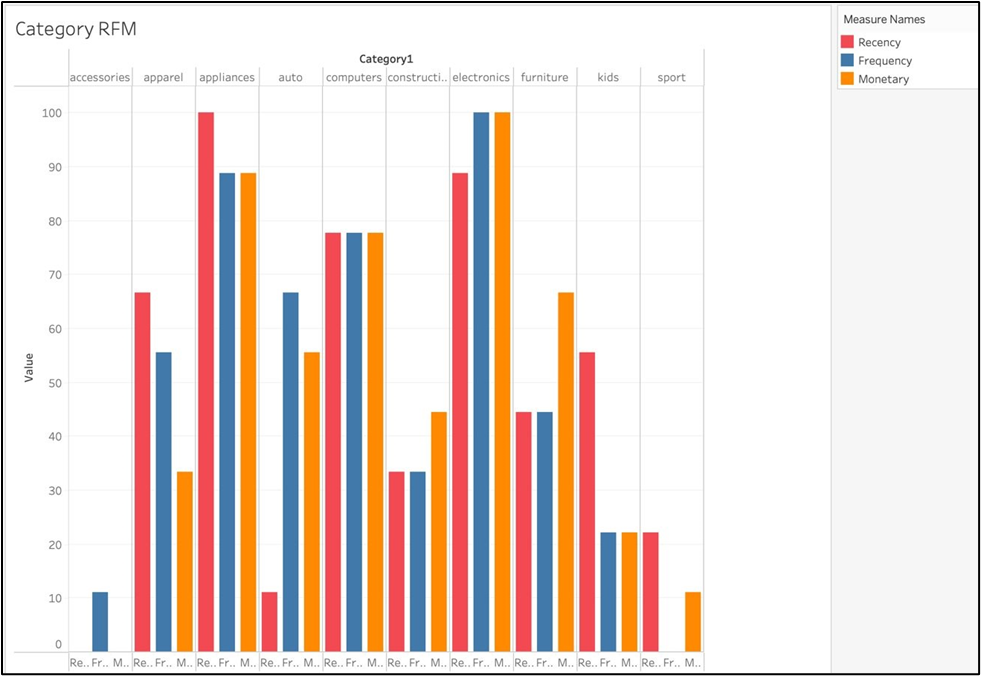


Fig. 14. Category RFM

* 1. **Insight 6: Activity trend analysis**

The visualization above shows the digital footprint on website at various times of the day. We can see there is a clear variation in the traffic on website as the day progresses. Activity on the website is the heaviest during the evening hours. We can use this insight to promote the deals/products during this time of the day and increase the sales.

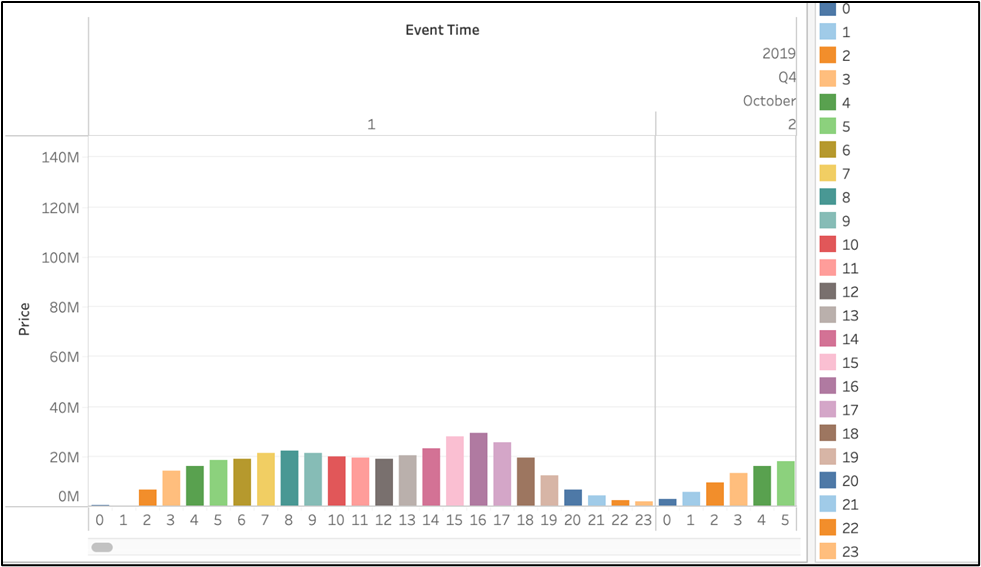


Fig. 15. Activity Trend Analysis

* 1. **Insight 7: User Session Funnel**

This visualization shows the flow of user activity on the website in accordance with the event types associated. This gives a clear view on how many products in each category were viewed, added to the cart, and then purchased. It also provides the average price of the products being added to each event. The % of total field compares the cart and purchases with respect to the views that have been made in the category and subsequently added to the cart or purchased.

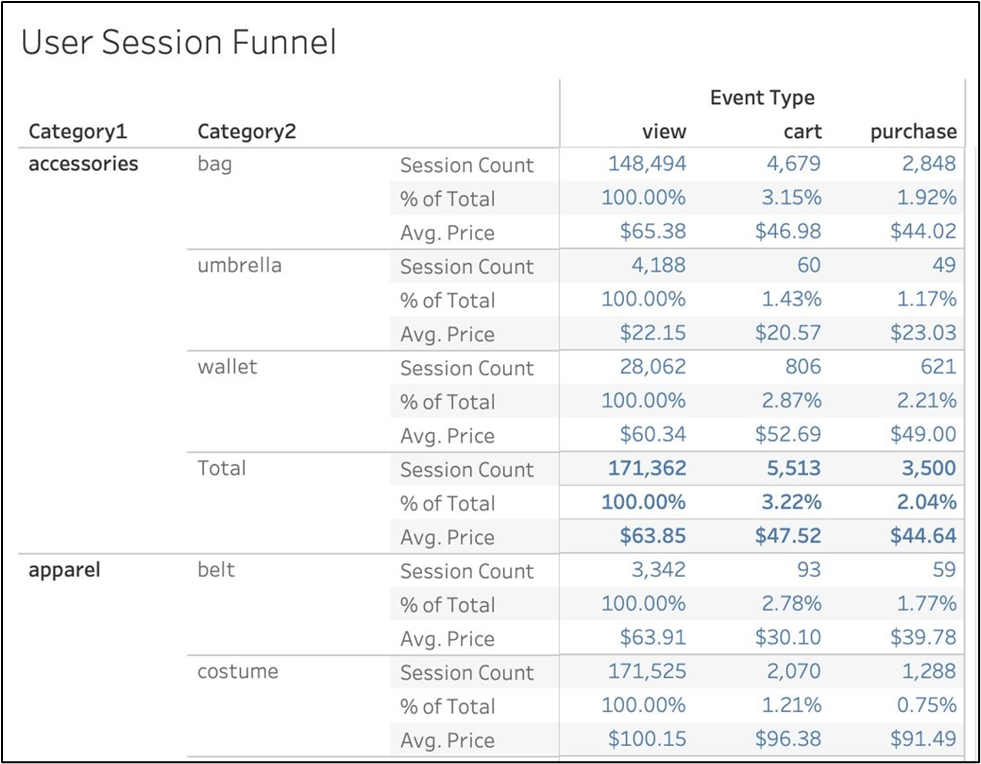


Fig. 16. User Session Funnel

* 1. **Insight 8: Event occurred distribution**

The data and visualization above show that an overall 94.89% of user visiting the website end up only browsing the data, 3.6% add the products to cart and only 1.51% make the purchase.

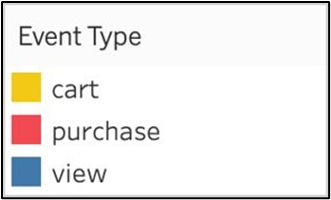
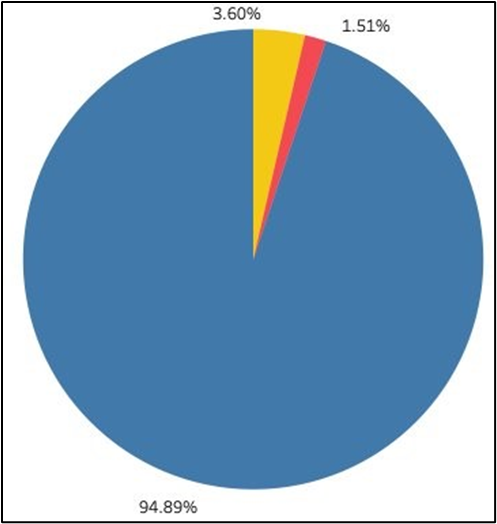


Fig. 17. Event Occurrence

# Chapter 5 Collaboration

**5.1.** **GitHub**

All source code and finished documentation will be uploaded to GitHub repository.

Repository URL: <https://github.com/Sakshijain3/Bigdata_SNOWFLAKE_PROJECT>

**5.2.** **Team members and Id’s**

|  |  |  |
| --- | --- | --- |
| **Name** | **Student ID** | **GitHub ID** |
| Snehal Dashrath Karad | 015935114 | 77168964 |
| Sakshi Jain | 015691936 | 89425256 |
| Lakshmi Naga Meghana Polisetty | 015918708 | 90289879 |
| Sakshi Tongia | 015443389 | 90562631 |

**Chapter 6** **Conclusion and Future Work**

**6.1 Conclusion**

This analysis helped us to understand the transactions that happened over the given dataset for two months. We were able to gain insights into the categories that sold the most, brand performance, user conversions, peak traffic times on the website. By doing analysis by creating User Funnel, Month over month analysis and RFM for the data, we were able to understand in depth the user buying trends, category, and brand performances.

**6.2**  **Challenges Faced**

We faced issues while loading the data in the snowflake using appropriate datatypes, so we loaded all the columns as string. Later, we were able to change the data types to the appropriate ones for all the columns using custom SQL for transformation.

**6.3** **Future work**

Dive deep into understanding why users are browsing content, but a small percent of that end up making a purchase, this will help in making changes to the website, availability of products etc., and increase the conversion ratio.

The data can be used to provide recommendations to the customers using their purchase history data and building a machine learning model to predict their next purchase product/time.

Cluster the customers based on their buying behavior we can optimize their shopping experience by showing them relevant products (what other customers who brought this item also brought).

References

* <https://www.kaggle.com/mkechinov/ecommerce-behavior-data-from-multi> category-store?select=2019-Nov.csv
* https://towardsdatascience.com/scalable-efficient-big-data-analytics-machine-learning-pipeline-architecture-on-cloud-4d59efc092b5
* https://help.tableau.com/current/pro/desktop/en-us/examples\_snowflake.htm
* https://docs.aws.amazon.com/IAM/latest/UserGuide/tutorials.html
* https://docs.aws.amazon.com/athena/latest/ug/work-with-data-stores.html