DATA VISUALIZATION PROJECT REPORT

(Project Semester January-May 2024)

PROJECT REPORT

ON

Data Visualization on Supermarket Sales

Submitted by

JAGADEESH

KUMAR

12111531

Program: Bachelor of Technology

Section: K21SH

Course Code: INTB233

Under the Guidance of

BALJINDER KAUR:27952

Discipline of CSE/IT

Lovely School of Computer Science & Engineering

LovelyProfessional University, Phagwara

DECLARATION

I, JAGADEESH KUMAR, student of Computer Science & Engineering under CSE/IT Discipline at, Lovely Professional University, Punjab, hereby declare that all the information furnished in this project report is based on my own intensive work and is genuine.

Date: Signature: **JAGADEESH**

Registration.no.**12111531** Name of the student:**JAGADEESH**

CERTIFICATE

This is to certify that **I, JAGADEESH KUMAR** bearing Registration no: **12106526** has completed INTB233 project titled, **"Data Visualization on Supermarketsales"** under my guidance and supervision. To the best of my knowledge, the present work is the result of his/her original development, effort and study.

Signature and Name of the Supervisor
Designation of the Supervisor
School of Computer Science & Engineering
Lovely Professional University
Phagwara, Punjab.

Date:10-03-2024

ACKNOWLEDGEMENT

Primarily I'd thank God for being able to complete my project with success. Then I'd like to thank my mentor **Ms.BALJINDER KAUR**, whose valuable guidance has been the ones that helped me patch this project and make it full proof success in contribution towards the completion of this project.

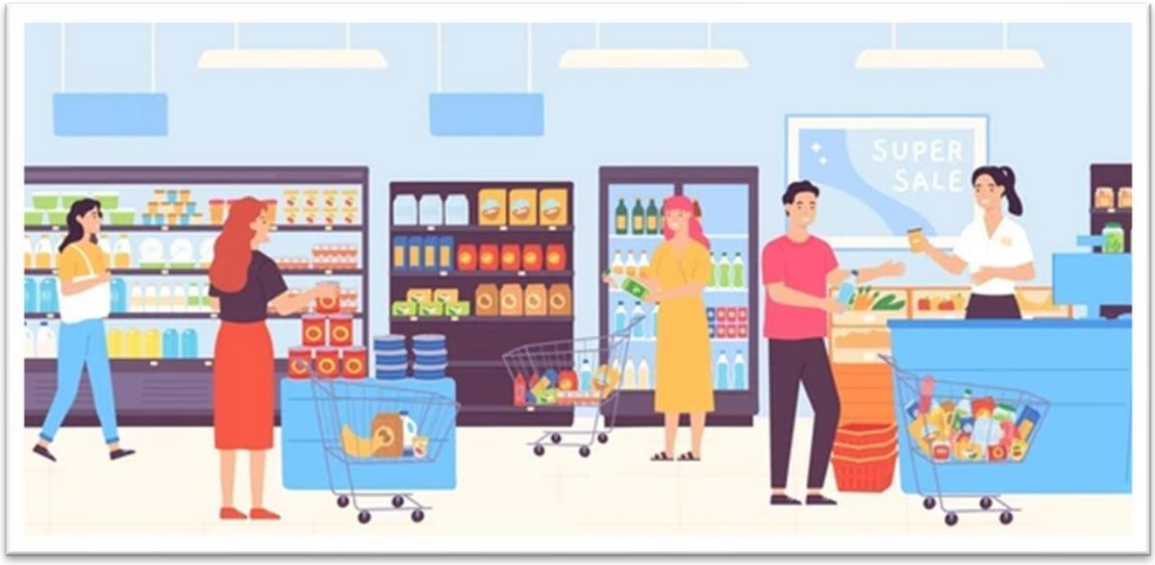
Finally, I'd rather thanks to **Lovely Professional University,** and my parent's inspiration, who gave me this golden opportunity to learn many new things, to learnanother aspect of life.

-JAGADEESH KUMAR

CONTENTS:

INTRODUCTION:

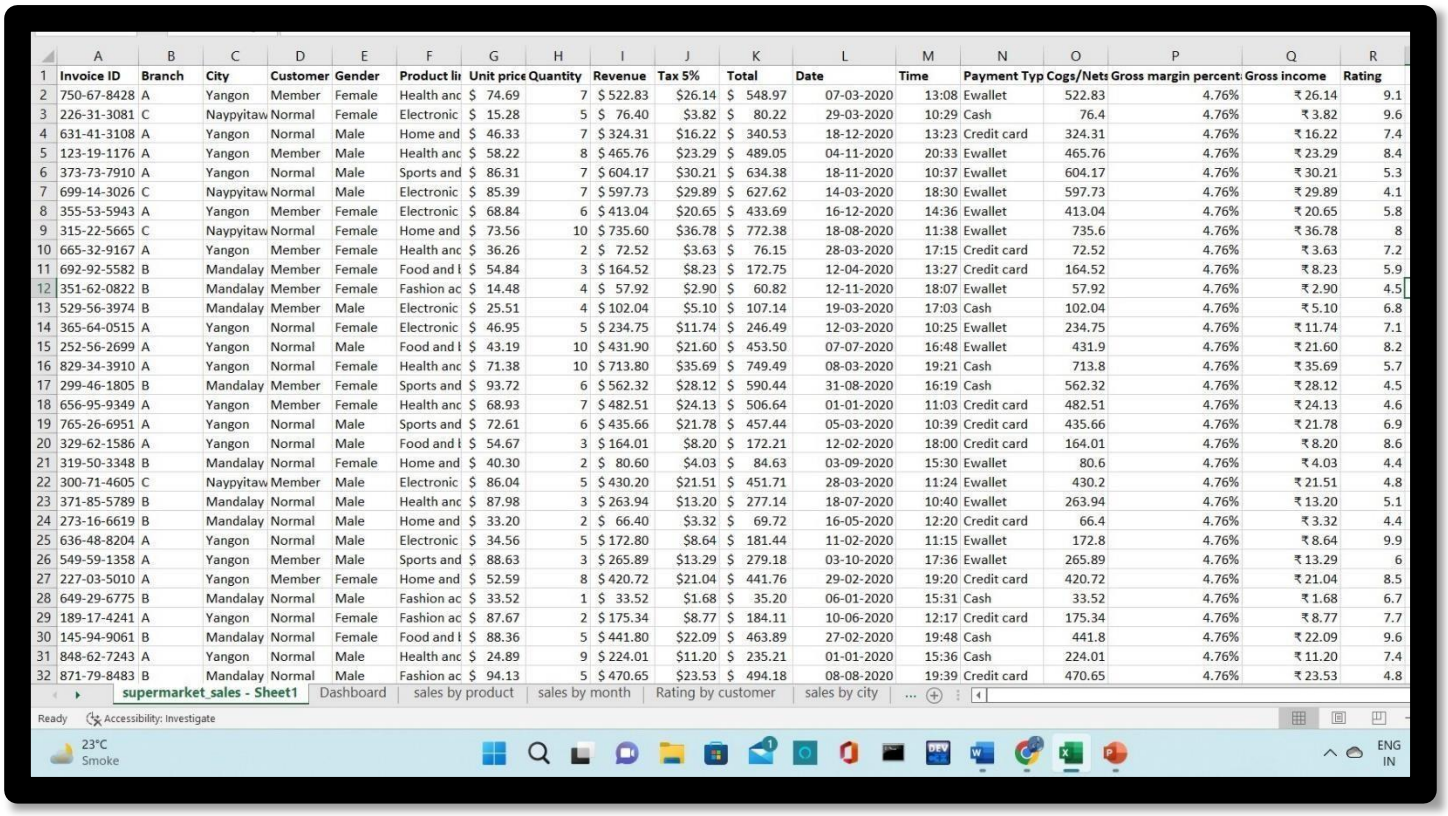
This project report analyzes the Supermarket Sales dataset consisting of sales data from three major cities in Myanmar: Yangon, Mandalay, and Naypyitaw. The dataset contains over 1,000 transactions recorded in 2020. It provides a comprehensive overview of sales patterns in the supermarkets located in these cities, offering valuable insights into consumer behavior and purchasing trends.



By analyzing the sales data from these three cities, we aim to gain a deeper understanding of consumer behavior in Myanmar and identify opportunities for improving supermarket sales and marketing strategies.The insights obtained from this analysis will be useful for supermarket owners, marketers, and other stakeholders in the retail industry who are looking to enhance their understanding of the Myanmar market and maximize their sales potential. The growth of supermarkets in most populated cities are increasing and market competitions are also high.

Overall, the Supermarket Sales dataset is a valuable resource for data analysts, researchers, and businesses interested in gaining insights into sales performance and customer behavior in the retail industry. The growth of supermarkets in most populated cities are increasing and market competitions are also high. The dataset is one of the historical sales of supermarket company which has recorded in 3 different branches for 1 year data (2020).

The supermarketsales raw data contains many parameters of customer data and transaction details like invoice_id, Payment type and rating by customer and many more.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1	Invoice ID	Branch	City	Customer	Gender	Product li	Unit price	Quantity	Revenue	Tax 5%	Total	Date	Time	Payment Typ	Cogs/Nets	Gross margin percent	Gross income	Rating
2	750-67-8428	A	Yangon	Member	Female	Health anc	$ 74.69	7	$ 522.83	$26.14	$ 548.97	07-03-2020	13:08	Ewallet	522.83	4.76%	₹ 26.14	9.1
3	226-31-3081	C	Naypyitaw	Normal	Female	Electronic	$ 15.28	5	$ 76.40	$3.82	$ 80.22	29-03-2020	10:29	Cash	76.4	4.76%	₹ 3.82	9.6
4	631-41-3108	A	Yangon	Normal	Male	Home and	$ 46.33	7	$ 324.31	$16.22	$ 340.53	18-12-2020	13:23	Credit card	324.31	4.76%	₹ 16.22	7.4
5	123-19-1176	A	Yangon	Member	Male	Health anc	$ 58.22	8	$ 465.76	$23.29	$ 489.05	04-11-2020	20:33	Ewallet	465.76	4.76%	₹ 23.29	8.4
6	373-73-7910	A	Yangon	Normal	Male	Sports and	$ 86.31	7	$ 604.17	$30.21	$ 634.38	18-11-2020	10:37	Ewallet	604.17	4.76%	₹ 30.21	5.3
7	699-14-3026	C	Naypyitaw	Normal	Male	Electronic	$ 85.39	7	$ 597.73	$29.89	$ 627.62	14-03-2020	18:30	Ewallet	597.73	4.76%	₹ 29.89	4.1
8	355-53-5943	A	Yangon	Member	Female	Electronic	$ 68.84	6	$ 413.04	$20.65	$ 433.69	16-12-2020	14:36	Ewallet	413.04	4.76%	₹ 20.65	5.8
9	315-22-5665	C	Naypyitaw	Normal	Female	Home and	$ 73.56	10	$ 735.60	$36.78	$ 772.38	18-08-2020	11:38	Ewallet	735.6	4.76%	₹ 36.78	8
10	665-32-9167	A	Yangon	Member	Female	Health anc	$ 36.26	2	$ 72.52	$3.63	$ 76.15	28-03-2020	17:15	Credit card	72.52	4.76%	₹ 3.63	7.2
11	692-92-5582	B	Mandalay	Member	Female	Food and b	$ 54.84	3	$ 164.52	$8.23	$ 172.75	12-04-2020	13:27	Credit card	164.52	4.76%	₹ 8.23	5.9
12	351-62-0822	B	Mandalay	Member	Female	Fashion ac	$ 14.48	4	$ 57.92	$2.90	$ 60.82	12-11-2020	18:07	Ewallet	57.92	4.76%	₹ 2.90	4.5
13	529-56-3974	B	Mandalay	Member	Male	Electronic	$ 25.51	4	$ 102.04	$5.10	$ 107.14	19-03-2020	17:03	Cash	102.04	4.76%	₹ 5.10	6.8
14	365-64-0515	A	Yangon	Normal	Female	Electronic	$ 46.95	5	$ 234.75	$11.74	$ 246.49	12-03-2020	10:25	Ewallet	234.75	4.76%	₹ 11.74	7.1
15	252-56-2699	A	Yangon	Normal	Male	Food and b	$ 43.19	10	$ 431.90	$21.60	$ 453.50	07-07-2020	16:48	Ewallet	431.9	4.76%	₹ 21.60	8.2
16	829-34-3910	A	Yangon	Normal	Female	Health anc	$ 71.38	10	$ 713.80	$35.69	$ 749.49	08-03-2020	19:21	Cash	713.8	4.76%	₹ 35.69	5.7
17	299-46-1805	B	Mandalay	Member	Female	Sports and	$ 93.72	6	$ 562.32	$28.12	$ 590.44	31-08-2020	16:19	Cash	562.32	4.76%	₹ 28.12	4.5
18	656-95-9349	A	Yangon	Member	Female	Health anc	$ 68.93	7	$ 482.51	$24.13	$ 506.64	01-01-2020	11:03	Credit card	482.51	4.76%	₹ 24.13	4.6
19	765-26-6951	A	Yangon	Normal	Male	Sports and	$ 72.61	6	$ 435.66	$21.78	$ 457.44	05-03-2020	10:39	Credit card	435.66	4.76%	₹ 21.78	6.9
20	329-62-1586	A	Yangon	Normal	Male	Food and b	$ 54.67	3	$ 164.01	$8.20	$ 172.21	12-02-2020	18:00	Credit card	164.01	4.76%	₹ 8.20	8.6
21	319-50-3348	B	Mandalay	Normal	Female	Home and	$ 40.30	2	$ 80.60	$4.03	$ 84.63	03-09-2020	15:30	Ewallet	80.6	4.76%	₹ 4.03	4.4
22	300-71-4605	C	Naypyitaw	Member	Male	Electronic	$ 86.04	5	$ 430.20	$21.51	$ 451.71	28-03-2020	11:24	Ewallet	430.2	4.76%	₹ 21.51	4.8
23	371-85-5789	B	Mandalay	Normal	Male	Health anc	$ 87.98	3	$ 263.94	$13.20	$ 277.14	18-07-2020	10:40	Ewallet	263.94	4.76%	₹ 13.20	5.1
24	273-16-6619	B	Mandalay	Normal	Male	Home and	$ 33.20	2	$ 66.40	$3.32	$ 69.72	16-05-2020	12:20	Credit card	66.4	4.76%	₹ 3.32	4.4
25	636-48-8204	A	Yangon	Normal	Male	Electronic	$ 34.56	5	$ 172.80	$8.64	$ 181.44	11-02-2020	11:15	Ewallet	172.8	4.76%	₹ 8.64	9.9
26	549-59-1358	A	Yangon	Member	Male	Sports and	$ 88.63	3	$ 265.89	$13.29	$ 279.18	03-10-2020	17:36	Ewallet	265.89	4.76%	₹ 13.29	6
27	227-03-5010	A	Yangon	Member	Female	Home and	$ 52.59	8	$ 420.72	$21.04	$ 441.76	29-02-2020	19:20	Credit card	420.72	4.76%	₹ 21.04	8.5
28	649-29-6775	B	Mandalay	Normal	Male	Fashion ac	$ 33.52	1	$ 33.52	$1.68	$ 35.20	06-01-2020	15:31	Cash	33.52	4.76%	₹ 1.68	6.7
29	189-17-4241	A	Yangon	Normal	Female	Fashion ac	$ 87.67	2	$ 175.34	$8.77	$ 184.11	10-06-2020	12:17	Credit card	175.34	4.76%	₹ 8.77	7.7
30	145-94-9061	B	Mandalay	Normal	Female	Food and b	$ 88.36	5	$ 441.80	$22.09	$ 463.89	27-02-2020	19:48	Cash	441.8	4.76%	₹ 22.09	9.6
31	848-62-7243	A	Yangon	Normal	Male	Health anc	$ 24.89	9	$ 224.01	$11.20	$ 235.21	01-01-2020	15:36	Cash	224.01	4.76%	₹ 11.20	7.4
32	871-79-8483	B	Mandalay	Normal	Male	Fashion ac	$ 94.13	5	$ 470.65	$23.53	$ 494.18	08-08-2020	19:39	Credit card	470.65	4.76%	₹ 23.53	4.8

It consists of different columns containing different categories like

- Invoice id: Computer generated sales slip invoice identification number.
- Branch: Branches of supercenter in city (3 branches are available identified by A, B and C).
- City: Location of supercenters (Yangon, Mandalay and Naypyitaw)
- Customer type: Type of customers, recorded by Members for customers using member and Normal for without member card.
- Gender: Gender type of customer (Male/Female).
- Product line: General item categorization groups - Electronic accessories, Fashion accessories, Food and beverages, Health and beauty, Home and lifestyle, Sports and travel.
- Unit price: Price of each product (in $).
- Quantity: Number of products purchased by customer.

- Tax: 5% tax fee for customer buying (in $).
- Total: Total price including tax (in $).
- Date: Date of purchase (Record available from January 2020 to December 2020).
- Time: Purchase time (10am to 9pm).
- Payment Type: Payment used by customer for purchase (3 methods are available – Cash,Credit card and E-wallet)
- Rating: Customer stratification rating on their overall shopping experience (On ascale of 1 to 10).

Scope of the Analysis/Objectives:

The aim of this project is to gain insights into the supermarket sales patterns in Myanmar, focusing on the differences and similarities between the three cities. We will examine various aspects of the data, including product categories, sales volume, revenue, and customer demographics, to identify key trends and patterns. Our analysis will involve a range of data visualization techniques, including charts, graphs, and tables, to provide a clear and concise overview of the sales data.

After analysis of the dataset, the aim of this project is to give answer of given objectives in easy way:

- Total Sales by Product type (**Top product)**
- Total Sales in each month
- Average rating on product (**Top Product**).
- Total sales with respect to Payment Type
- Most Sales for Each hour (**Hourly sales**)
- Quantity with respect to City
- Forecast of sales for 2021

Sales by product category: Analyzing sales by product category can help identify the most popular products in the supermarket and which categories are performing well or poorly. This can help the supermarket to make informed decisions on product placement, promotions, and pricing strategies.

Sales by city: Analyzing sales by store location can help identify which stores are performing well or poorly, and which regions have the highest demand for specific products. This can help the supermarket to make informed decisions on store expansion or closure.

Sales by Hours: Analyzing sales by time of day can help identify peak shopping times and optimize staffing and inventory levels accordingly**.** It also helps the management that when to add the staff to increase the sales and to reduce the staff to decrease the expenses.

Monthly Sales: The analysis can provide insights into the sales trends for each month, such as identifying which months have higher or lower sales, and any patterns in monthly sales fluctuations. The analysis can identify customer behavior patterns, such as identifying if customers tend to purchase certain products more frequently during specific months.

Average rating on product: The analysis can identify which product lines have the highest ratings, and assess the relationship between product line and sales performance. It can assess how customer ratings impact purchasing behavior, such as identifying whether products with higher ratings tend to sell more frequently.

Forecast of sales for 2021: The scope of analysis for a data visualization of the forecast of sales for the next 4 months based on 1 year of data might involve displaying the monthly sales data for the past 12 months in a line chart or combination chart format. The forecasted sales for the next 4 months can be displayed as a separate line or series on the chart, with clear labels and annotations to indicate the forecast method and any relevant external factors.

Existing System/ Drawbacks or limitations of existing system

There are some limitations for the existing system

- Limited data sources: The existing systems may only have access to a limited set of data sources, which could result in incomplete or inaccurate information. This can affect the accuracy and usefulness of the data visualizations.

- Lack of customization: Existing systems may not allow for a high degree of customization in terms of data visualization options. This means that the data visualizations may not be tailored to the specific needs of the project report.

- Limited interactivity: Many existing systems may have limited interactivity, which can make it difficult to explore the data in detail or to identify trends and patterns.

- Poor data quality: Data quality is essential for accurate data visualization, but existing systems may not have effective measures in place to ensure data quality. This can result in errors or inconsistencies in the data, which can negatively impact the usefulness of the data visualizations.

- Technical limitations: Existing systems may have technical limitations that prevent them from processing and visualizing large amounts of data effectively. This can limit the usefulness of the data visualizations and make it difficult to identify trends and patterns in the data.

- Lack of real-time data: Many existing systems may not provide real-time data, which can make it difficult to track changes and respond quickly to changes in the market or customer behavior.

- Security concerns: With the increasing concern of data breaches and cyber threats, existing systems may not have adequate security measures in place to protect sensitive data. This can compromise the privacy and security of customer and business data.

- Cost: Some existing systems can be expensive to purchase or maintain, particularly for smaller supermarkets with limited budgets. This can be a barrier to adoption for some users.

SOURCE OF DATASET:

https://www.kaggle.com/datasets/aungpyaeap/supermarket-sales

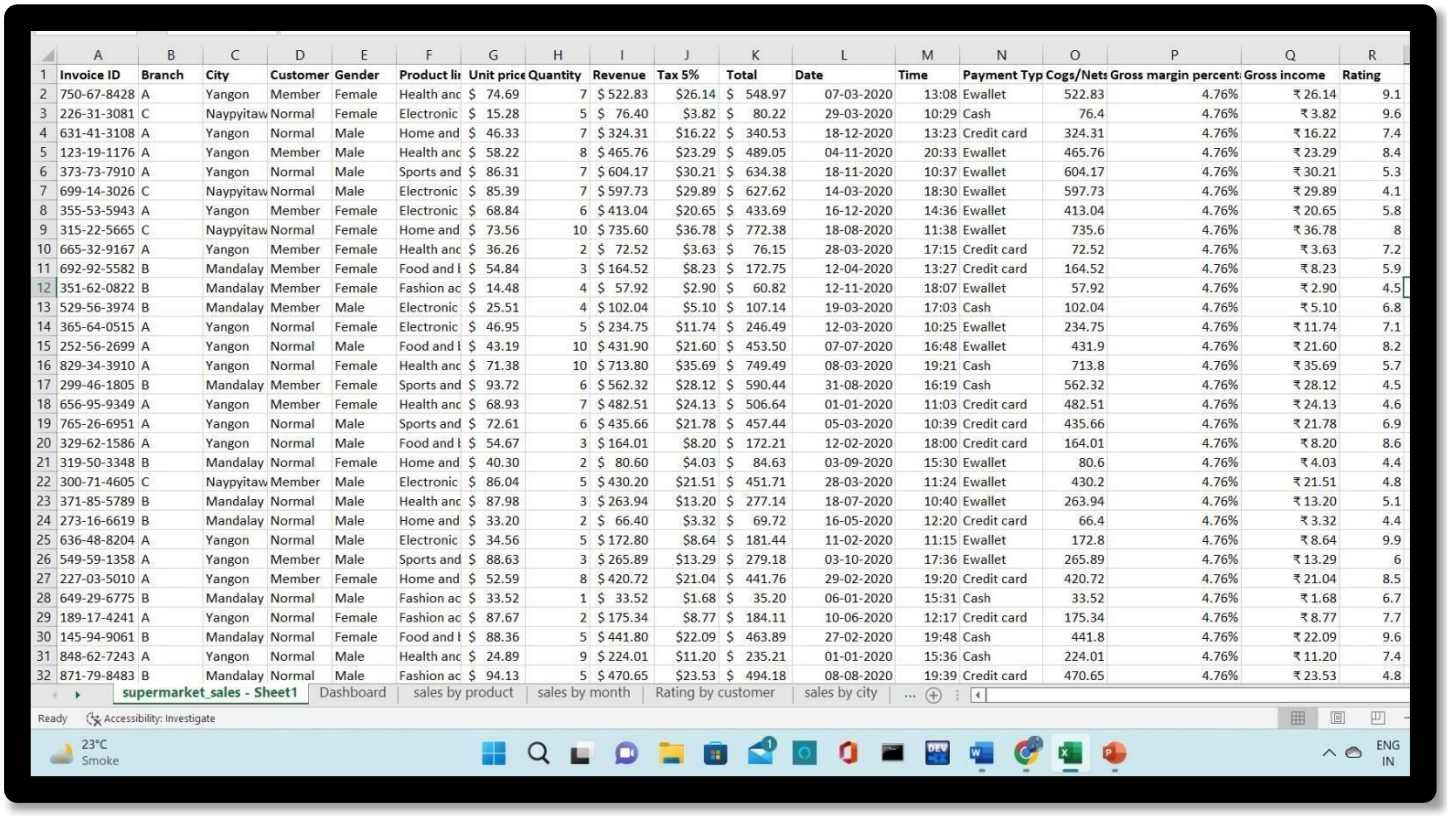
KAGGLE:

Kaggle is an online community for data scientists and machine learners, developedby Google. Kaggle allows users to find and publish data sets, explore, and build models in a web-based data-science environment, work with other data scientists and machine learning engineers, and enter competitions to solve data science challenges. Kaggle got its start by offering machine learning competitions and nowalso offers a public data platform, a cloud - based workbench for data science, and short form Al education. On 8 March 2017, Google announced that they were acquiring Kaggle.

This data science project analyzes the Supermarket sales dataset. It was created forthe (B. Tech CSE sixth semester Data Visualization course) project.

Sample of Dataset with DataFields:

The dataset contains the following columns like invoice id, city, product and so on.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1	Invoice ID	Branch	City	Customer	Gender	Product li	Unit price	Quantity	Revenue	Tax 5%	Total	Date	Time	Payment Typ	Cogs/Nets	Gross margin percent	Gross income	Rating
2	750-67-8428	A	Yangon	Member	Female	Health anc	$ 74.69	7	$ 522.83	$26.14	$ 548.97	07-03-2020	13:08	Ewallet	522.83	4.76%	₹ 26.14	9.1
3	226-31-3081	C	Naypyitaw	Normal	Female	Electronic	$ 15.28	5	$ 76.40	$3.82	$ 80.22	29-03-2020	10:29	Cash	76.4	4.76%	₹ 3.82	9.6
4	631-41-3108	A	Yangon	Normal	Male	Home and	$ 46.33	7	$ 324.31	$16.22	$ 340.53	18-12-2020	13:23	Credit card	324.31	4.76%	₹ 16.22	7.4
5	123-19-1176	A	Yangon	Member	Male	Health anc	$ 58.22	8	$ 465.76	$23.29	$ 489.05	04-11-2020	20:33	Ewallet	465.76	4.76%	₹ 23.29	8.4
6	373-73-7910	A	Yangon	Normal	Male	Sports and	$ 86.31	7	$ 604.17	$30.21	$ 634.38	18-11-2020	10:37	Ewallet	604.17	4.76%	₹ 30.21	5.3
7	699-14-3026	C	Naypyitaw	Normal	Male	Electronic	$ 85.39	7	$ 597.73	$29.89	$ 627.62	14-03-2020	18:30	Ewallet	597.73	4.76%	₹ 29.89	4.1
8	355-53-5943	A	Yangon	Member	Female	Electronic	$ 68.84	6	$ 413.04	$20.65	$ 433.69	16-12-2020	14:36	Ewallet	413.04	4.76%	₹ 20.65	5.8
9	315-22-5665	C	Naypyitaw	Normal	Female	Home and	$ 73.56	10	$ 735.60	$36.78	$ 772.38	18-08-2020	11:38	Ewallet	735.6	4.76%	₹ 36.78	8
10	665-32-9167	A	Yangon	Member	Female	Health anc	$ 36.26	2	$ 72.52	$3.63	$ 76.15	28-03-2020	17:15	Credit card	72.52	4.76%	₹ 3.63	7.2
11	692-92-5582	B	Mandalay	Member	Female	Food and I	$ 54.84	3	$ 164.52	$8.23	$ 172.75	12-04-2020	13:27	Credit card	164.52	4.76%	₹ 8.23	5.9
12	351-62-0822	B	Mandalay	Member	Female	Fashion ac	$ 14.48	4	$ 57.92	$2.90	$ 60.82	12-11-2020	18:07	Ewallet	57.92	4.76%	₹ 2.90	4.5
13	529-56-3974	B	Mandalay	Member	Male	Electronic	$ 25.51	4	$ 102.04	$5.10	$ 107.14	19-03-2020	17:03	Cash	102.04	4.76%	₹ 5.10	6.8
14	365-64-0515	A	Yangon	Normal	Female	Electronic	$ 46.95	5	$ 234.75	$11.74	$ 246.49	12-03-2020	10:25	Ewallet	234.75	4.76%	₹ 11.74	7.1
15	252-56-2699	A	Yangon	Normal	Male	Food and I	$ 43.19	10	$ 431.90	$21.60	$ 453.50	07-07-2020	16:48	Ewallet	431.9	4.76%	₹ 21.60	8.2
16	829-34-3910	A	Yangon	Normal	Female	Health anc	$ 71.38	10	$ 713.80	$35.69	$ 749.49	08-03-2020	19:21	Cash	713.8	4.76%	₹ 35.69	5.7
17	299-46-1805	B	Mandalay	Member	Female	Sports and	$ 93.72	6	$ 562.32	$28.12	$ 590.44	31-08-2020	16:19	Cash	562.32	4.76%	₹ 28.12	4.5
18	656-95-9349	A	Yangon	Member	Female	Health anc	$ 68.93	7	$ 482.51	$24.13	$ 506.64	01-01-2020	11:03	Credit card	482.51	4.76%	₹ 24.13	4.6
19	765-26-6951	A	Yangon	Normal	Male	Sports and	$ 72.61	6	$ 435.66	$21.78	$ 457.44	05-03-2020	10:39	Credit card	435.66	4.76%	₹ 21.78	6.9
20	329-62-1586	A	Yangon	Normal	Male	Food and I	$ 54.67	3	$ 164.01	$8.20	$ 172.21	12-02-2020	18:00	Credit card	164.01	4.76%	₹ 8.20	8.6
21	319-50-3348	B	Mandalay	Normal	Female	Home and	$ 40.30	2	$ 80.60	$4.03	$ 84.63	03-09-2020	15:30	Ewallet	80.6	4.76%	₹ 4.03	4.4
22	300-71-4605	C	Naypyitaw	Member	Male	Electronic	$ 86.04	5	$ 430.20	$21.51	$ 451.71	28-03-2020	11:24	Ewallet	430.2	4.76%	₹ 21.51	4.8
23	371-85-5789	B	Mandalay	Normal	Male	Health anc	$ 87.98	3	$ 263.94	$13.20	$ 277.14	18-07-2020	10:40	Ewallet	263.94	4.76%	₹ 13.20	5.1
24	273-16-6619	B	Mandalay	Normal	Male	Home and	$ 33.20	2	$ 66.40	$3.32	$ 69.72	16-05-2020	12:20	Credit card	66.4	4.76%	₹ 3.32	4.4
25	636-48-8204	A	Yangon	Normal	Male	Electronic	$ 34.56	5	$ 172.80	$8.64	$ 181.44	11-02-2020	11:15	Ewallet	172.8	4.76%	₹ 8.64	9.9
26	549-59-1358	A	Yangon	Member	Male	Sports and	$ 88.63	3	$ 265.89	$13.29	$ 279.18	03-10-2020	17:36	Ewallet	265.89	4.76%	₹ 13.29	6
27	227-03-5010	A	Yangon	Member	Female	Home and	$ 52.59	8	$ 420.72	$21.04	$ 441.76	29-02-2020	19:20	Credit card	420.72	4.76%	₹ 21.04	8.5
28	649-29-6775	B	Mandalay	Normal	Male	Fashion ac	$ 33.52	1	$ 33.52	$1.68	$ 35.20	06-01-2020	15:31	Cash	33.52	4.76%	₹ 1.68	6.7
29	189-17-4241	A	Yangon	Normal	Female	Fashion ac	$ 87.67	2	$ 175.34	$8.77	$ 184.11	10-06-2020	12:17	Credit card	175.34	4.76%	₹ 8.77	7.7
30	145-94-9061	B	Mandalay	Normal	Female	Food and I	$ 88.36	5	$ 441.80	$22.09	$ 463.89	27-02-2020	19:48	Cash	441.8	4.76%	₹ 22.09	9.6
31	848-62-7243	A	Yangon	Normal	Male	Health anc	$ 24.89	9	$ 224.01	$11.20	$ 235.21	01-01-2020	15:36	Cash	224.01	4.76%	₹ 11.20	7.4
32	871-79-8483	B	Mandalay	Normal	Male	Fashion ac	$ 94.13	5	$ 470.65	$23.53	$ 494.18	08-08-2020	19:39	Credit card	470.65	4.76%	₹ 23.53	4.8

ETL PROCESS:

ETL is defined as a process that extracts the data from different RDBMS source systems, then transforms the data (like applying calculations, concatenations, etc.) and finally loads the data into the Data Warehouse system. ETL full form is Extract,Transform and Load.

It's tempting to think a creating a Data warehouse is simply extracting data from multiple sources and loading into database of a Data warehouse. This is far from the truth and requires a complex ETL process. The ETL process requires active inputs from various stakeholders including developers, analysts, testers, top executives and is technically challenging.

Need of ETL Process

- ETL process allows sample data comparison between the source and the target system.
- ETL is a predefined process for accessing and manipulating source data intothe target database.
- Allow verification of data transformation, aggregation, and calculations rules.

When it comes to the implementation of the ETL process, the itinerary of tasks can bedivided up into the full form of its acronym.

1. **E – Extraction**
2. **T – Transformation**
3. **L – Loading**

Extract is the process of fetching (reading) the information from the database. Atthis stage, data is collected from multiple or different types of sources.A staging area is required during ETL load. There are various reasons why staging area is required.The source systems are only available for specific period of time to extract data. This period of time is less than the total data-load time. Therefore, staging area allows you to extract the data from the source system and keeps it in the staging area before the time slot ends.

In data transformation, you apply a set of functions on extracted data to load it intothe target system. Data, which does not require any transformation is known as direct move or pass-through data,

Transform is the transformation step is where the raw data is cleaned, prepared, and transformed into a format that is suitable for visualization. Here are some possible transformations that you may need to perform:

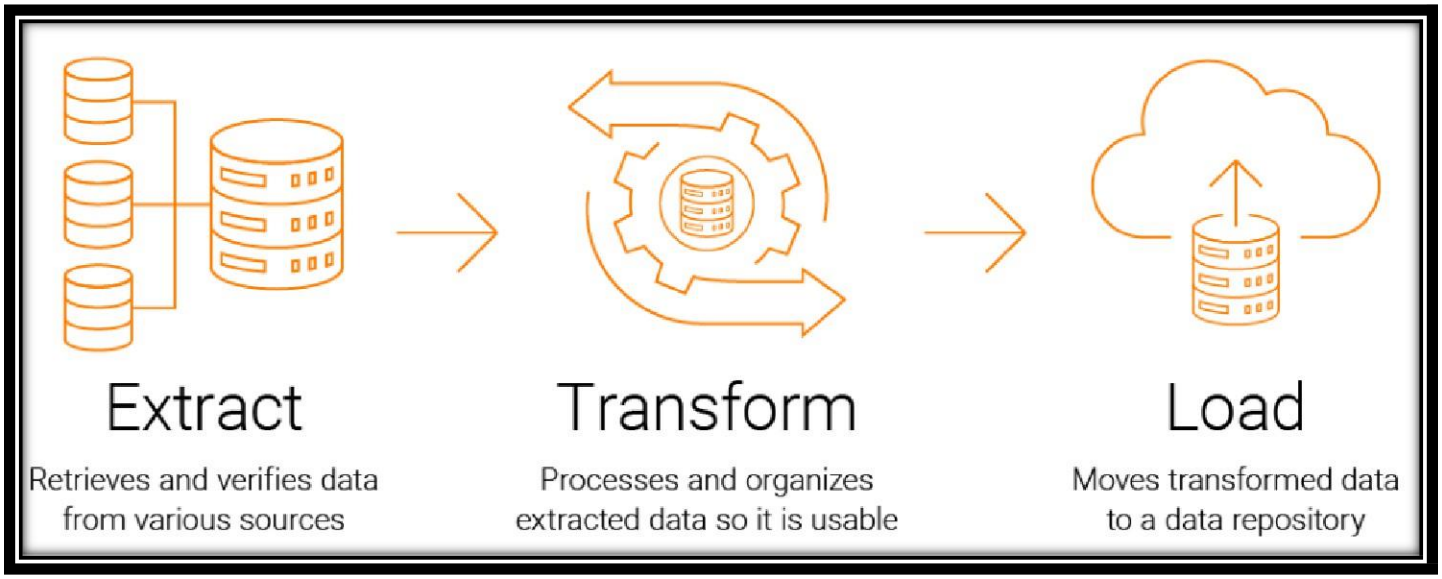
Cleaning the data: This involves removing any duplicate rows or inconsistent values that can interfere with the accuracy of the analysis.

Removing unnecessary columns: If there are columns that are not relevant to the project report, they can be removed to reduce clutter and improve readability.

Renaming columns: If column names are not descriptive or easy to understand, they can be renamed to make them more meaningful.

Converting data types: Dates, for example, may need to be converted to a more standardized format to ensure accurate analysis.

Aggregating data: Summarizing or aggregating the data by creating calculated fields or summarizing data can be useful for analysis.



Load is the process of writing the data into the target database. During Load phase,data is loaded into the end-target system and it can be a flat file or a Data Warehouse system. During the load step, it is necessary to ensure that the load is performed correctly and with as little resources as possible. The target of the Load process is often a database. In order to make the load process efficient, it is helpful to disable any constraints and indexes before the load and enable them back only after the load completes. The referential integrity needs to be maintained by ETL tool to ensure consistency.

ETL Process used in Project

Extract:

The extraction step involves obtaining the raw data from its source, which in this case is Kaggle. I downloaded the dataset from Kaggle in CSV format, which I have saved it to a folder on your computer.

Transform:

In the Supermarket Sales dataset, there are several columns with null values, such as the Product Weight column. One approach to handling null values is to replace them with the mean or median value of the column. Another approach is to remove the rows with null values entirely. I cleaned the data having null values and duplicate values using Tableau Prep with ETL tools.

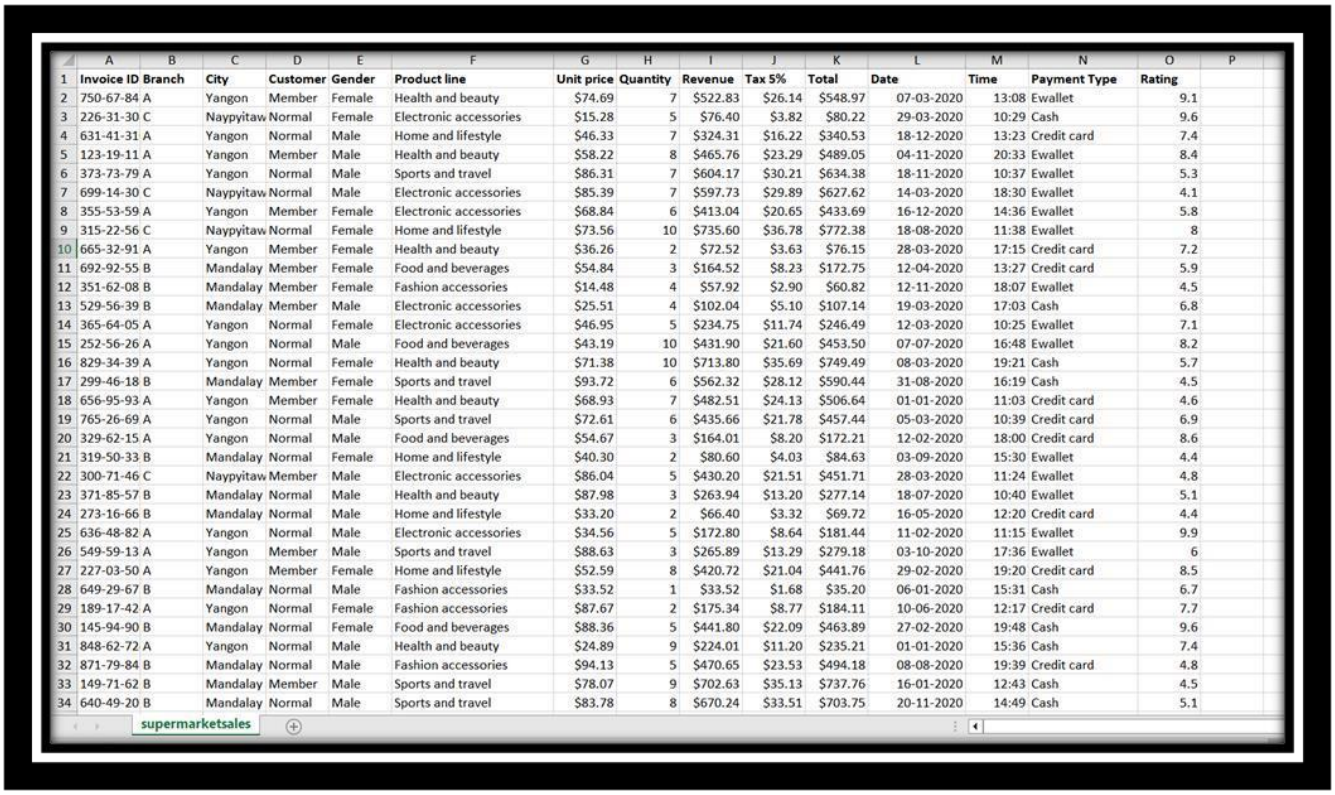
I also removed the Unnecessary columns like Gross income and Gross margin percentage, which is not useful for my data visualization and data analysis. It is also a reducing the data set complexity.I also gave the currency value to the total sales and unit price(in $).

Adding a new column: We can add a new column to the dataset that calculates the Revenue for each transaction by multiplying the Unit Price by the Quantity. We can do this by using the following formula:

I added the new column with name Revenue using formula which is multiplying the Quantity and Unit price without including the tax.

Formula: Revenue = Quantity * Unit price of Product Quantity = No. of items sold

After writing the formula in formula bar and dragging the Dragger to the end of the table, we can get the new column with Revenue.



Invoice ID	Branch	City	Customer	Gender	Product line	Unit price	Quantity	Revenue	Tax 5%	Total	Date	Time	Payment Type	Rating
750-67-84	A	Yangon	Member	Female	Health and beauty	$74.69	7	$522.83	$26.14	$548.97	07-03-2020	13:08	Ewallet	9.1
226-31-30	C	Naypyitaw	Normal	Female	Electronic accessories	$15.28	5	$76.40	$3.82	$80.22	29-03-2020	10:29	Cash	9.6
631-41-31	A	Yangon	Normal	Male	Home and lifestyle	$46.33	7	$324.31	$16.22	$340.53	18-12-2020	13:23	Credit card	7.4
123-19-11	A	Yangon	Member	Male	Health and beauty	$58.22	8	$465.76	$23.29	$489.05	04-11-2020	20:33	Ewallet	8.4
373-73-79	A	Yangon	Normal	Male	Sports and travel	$86.31	7	$604.17	$30.21	$634.38	18-11-2020	10:37	Ewallet	5.3
699-14-30	C	Naypyitaw	Normal	Male	Electronic accessories	$85.39	7	$597.73	$29.89	$627.62	14-03-2020	18:30	Ewallet	4.1
355-53-59	A	Yangon	Member	Female	Electronic accessories	$68.84	6	$413.04	$20.65	$433.69	16-12-2020	14:36	Ewallet	5.8
315-22-56	C	Naypyitaw	Normal	Female	Home and lifestyle	$73.56	10	$735.60	$36.78	$772.38	18-08-2020	11:38	Ewallet	8
665-32-91	A	Yangon	Member	Female	Health and beauty	$36.26	2	$72.52	$3.63	$76.15	28-03-2020	17:15	Credit card	7.2
692-92-55	B	Mandalay	Member	Female	Food and beverages	$54.84	3	$164.52	$8.23	$172.75	12-04-2020	13:27	Credit card	5.9
351-62-08	B	Mandalay	Member	Female	Fashion accessories	$14.48	4	$57.92	$2.90	$60.82	12-11-2020	18:07	Ewallet	4.5
529-56-39	B	Mandalay	Member	Male	Electronic accessories	$25.51	4	$102.04	$5.10	$107.14	19-03-2020	17:03	Cash	6.8
365-64-05	A	Yangon	Normal	Female	Electronic accessories	$46.95	5	$234.75	$11.74	$246.49	12-03-2020	10:25	Ewallet	7.1
252-56-26	A	Yangon	Normal	Male	Food and beverages	$43.19	10	$431.90	$21.60	$453.50	07-07-2020	16:48	Ewallet	8.2
829-34-39	A	Yangon	Normal	Female	Health and beauty	$71.38	10	$713.80	$35.69	$749.49	08-03-2020	19:21	Cash	5.7
299-46-18	B	Mandalay	Member	Female	Sports and travel	$93.72	6	$562.32	$28.12	$590.44	31-08-2020	16:19	Cash	4.5
656-95-93	A	Yangon	Member	Female	Health and beauty	$68.93	7	$482.51	$24.13	$506.64	01-01-2020	11:03	Credit card	4.6
765-26-69	A	Yangon	Normal	Male	Sports and travel	$72.61	6	$435.66	$21.78	$457.44	05-03-2020	10:39	Credit card	6.9
329-62-15	A	Yangon	Normal	Male	Food and beverages	$54.67	3	$164.01	$8.20	$172.21	12-02-2020	18:00	Credit card	8.6
319-50-33	B	Mandalay	Normal	Female	Home and lifestyle	$40.30	2	$80.60	$4.03	$84.63	03-09-2020	15:30	Ewallet	4.4
300-71-46	C	Naypyitaw	Member	Male	Electronic accessories	$86.04	5	$430.20	$21.51	$451.71	28-03-2020	11:24	Ewallet	4.8
371-85-57	B	Mandalay	Normal	Male	Health and beauty	$87.98	3	$263.94	$13.20	$277.14	18-07-2020	10:40	Ewallet	5.1
273-16-66	B	Mandalay	Normal	Male	Home and lifestyle	$33.20	2	$66.40	$3.32	$69.72	16-05-2020	12:20	Credit card	4.4
636-48-82	A	Yangon	Normal	Male	Electronic accessories	$34.56	5	$172.80	$8.64	$181.44	11-02-2020	11:15	Ewallet	9.9
549-59-13	A	Yangon	Member	Male	Sports and travel	$88.63	3	$265.89	$13.29	$279.18	03-10-2020	17:36	Ewallet	6
227-03-50	A	Yangon	Member	Female	Home and lifestyle	$52.59	8	$420.72	$21.04	$441.76	29-02-2020	19:20	Credit card	8.5
649-29-67	B	Mandalay	Normal	Male	Fashion accessories	$33.52	1	$33.52	$1.68	$35.20	06-01-2020	15:31	Cash	6.7
189-17-42	A	Yangon	Normal	Female	Fashion accessories	$87.67	2	$175.34	$8.77	$184.11	10-06-2020	12:17	Credit card	7.7
145-94-90	B	Mandalay	Normal	Female	Food and beverages	$88.36	5	$441.80	$22.09	$463.89	27-02-2020	19:48	Cash	9.6
848-62-72	A	Yangon	Normal	Male	Health and beauty	$24.89	9	$224.01	$11.20	$235.21	01-01-2020	15:36	Cash	7.4
871-79-84	B	Mandalay	Normal	Male	Fashion accessories	$94.13	5	$470.65	$23.53	$494.18	08-08-2020	19:39	Credit card	4.8
149-71-62	B	Mandalay	Member	Male	Sports and travel	$78.07	9	$702.63	$35.13	$737.76	16-01-2020	12:43	Cash	4.5
640-49-20	B	Mandalay	Normal	Male	Sports and travel	$83.78	8	$670.24	$33.51	$703.75	20-11-2020	14:49	Cash	5.1

supermarketsales

Analysis of DataSet:

1) Total Sales for Product type

Introduction

By performing this analysis, we will get number of Sales with respect to Product type

Description: It is to see the number of Sales for each product.I used Pie chart for this analysis which is sui suitable for easy analysis and visualization.

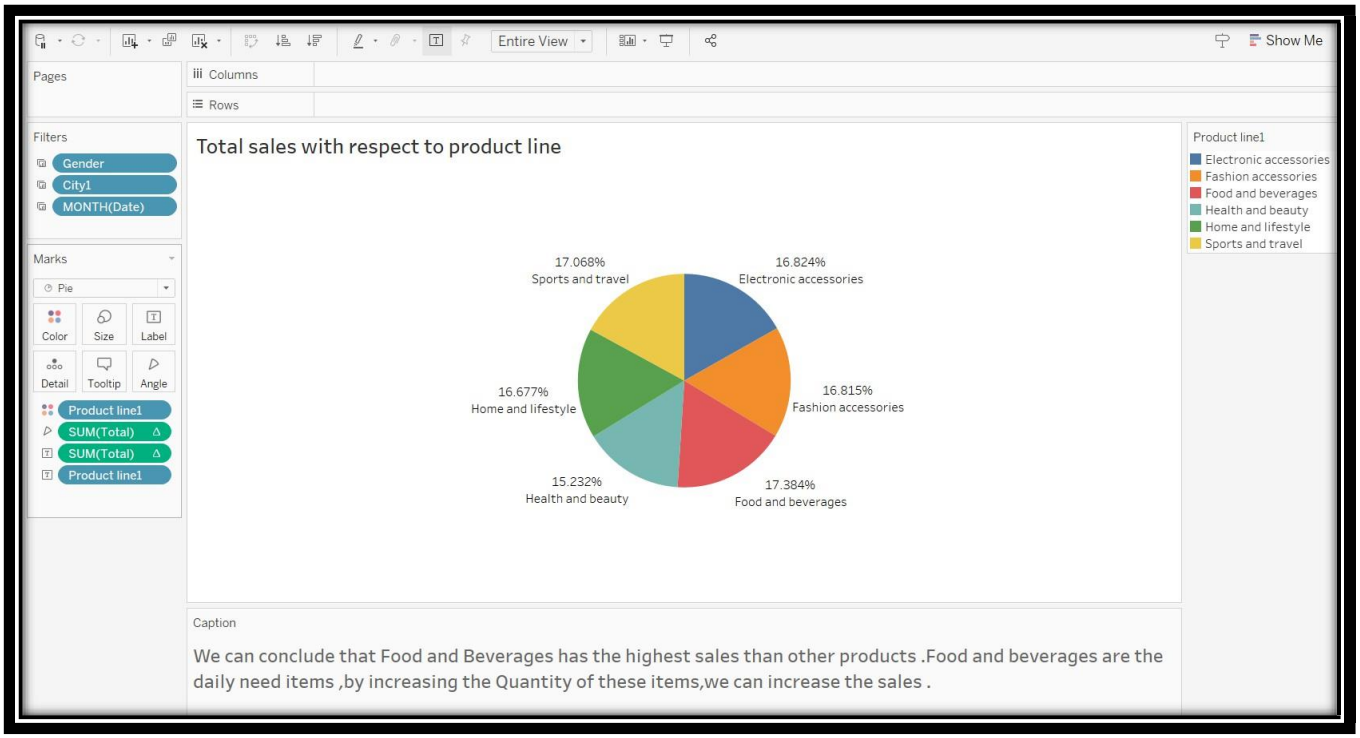
Specific requirements, functions, and formulas:

For counting number of sales, we can use count function: =Sum()

And Used the filters for Specific analysis and easy visualization

Analysis results: In the year 2020, most sales are with respect toproduct :

The Food and Beverages has highest sales i.e;(17.384% of total sales) and Home and lifestyle has lowest sales with (15.232% oftotal sales) .



2) Total Sales by Month

Introduction

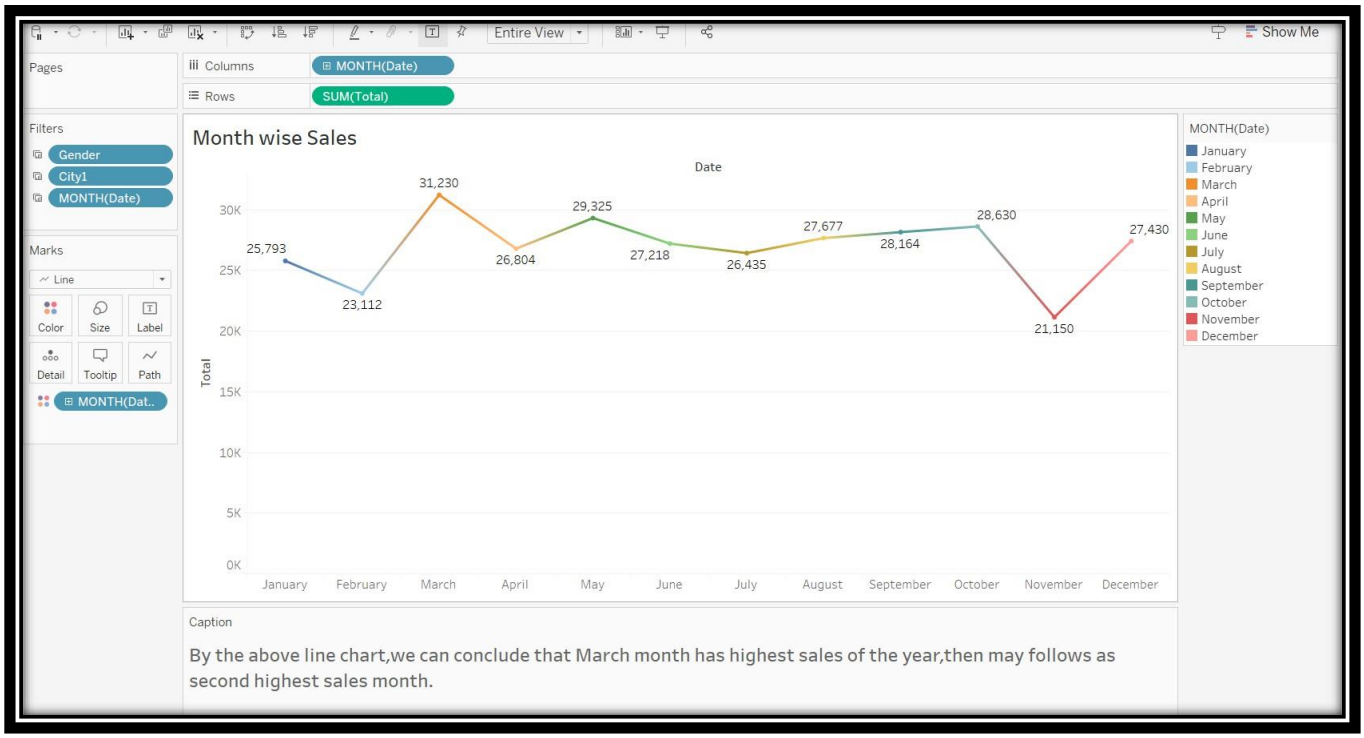
By performing this analysis, we will get number of Sales for eachmonth of year 2020

Description*:* It is to see the most sales in a month of year 2020.I used Line chart for this analysis which show the increase and decrease in sales in months wise.

Function Used : I have used sum function to add up the sales =SUM()

Analysis results: In the year 2020, most Sales in each month ofyear 2020 :

In March,2020 has highest sales with value $31230 and In November has lowest sales with value $21150.



3) Average Rating on Product :

Introduction

By performing this analysis, we will get Average rating for each Product type.

Description: It is to see the most top-rated product type in 2020.I used Bar chart for this analysis which is suitable for this analysis.

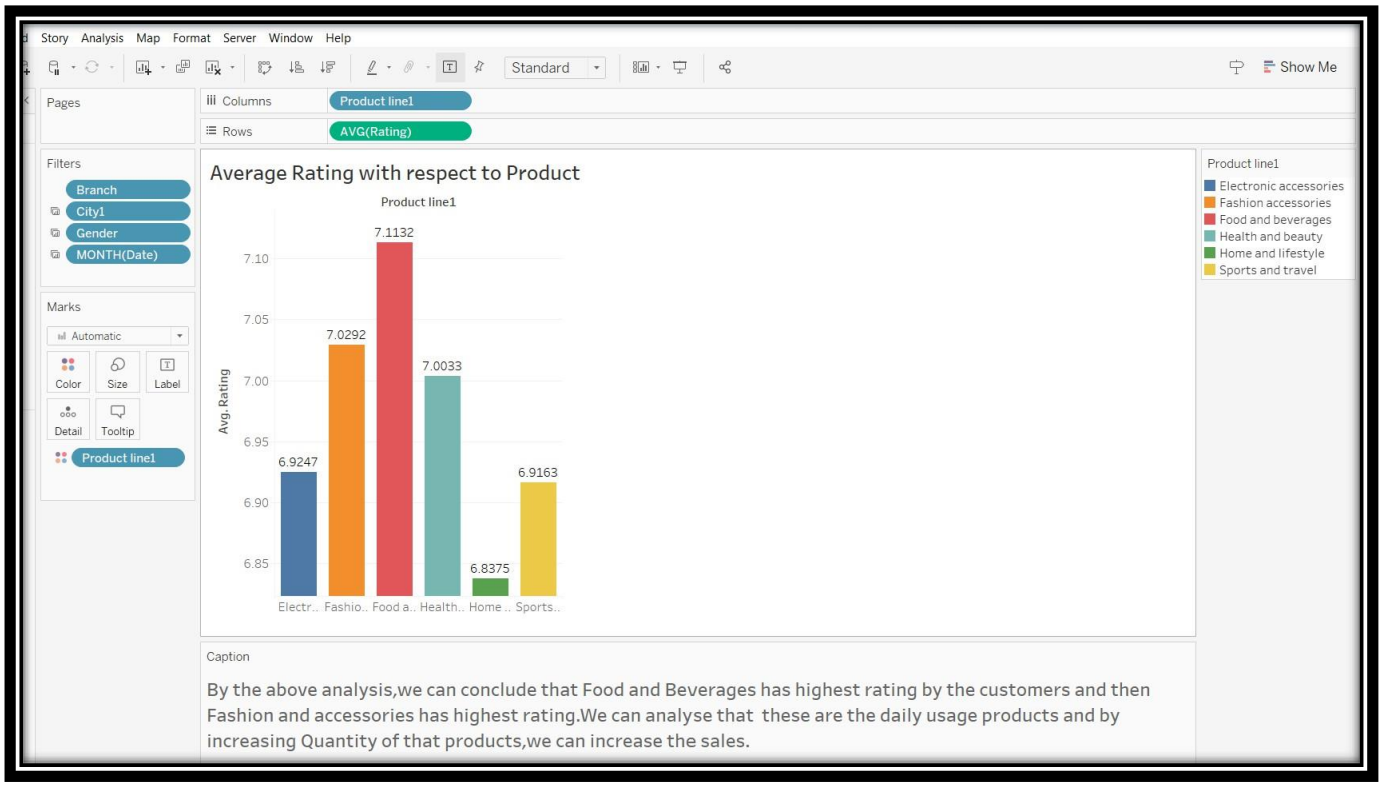
Formula/Function Used:

We calculated the average rating by using function=Average ()

Average = (sum of all ratings of customers)/no.of total customers

Analysis results: In the year 2020, most top-rated product:I used the Bar chart for this analysis which is used for easy analysis of data.

In 2020, the most rated product type is Food and beverages with average rating of 7.11(out of 10) and Home and lifestyle is lowestrated product with 6.84(out of 10)



4) Most Sales by Time (Hours)

Introduction

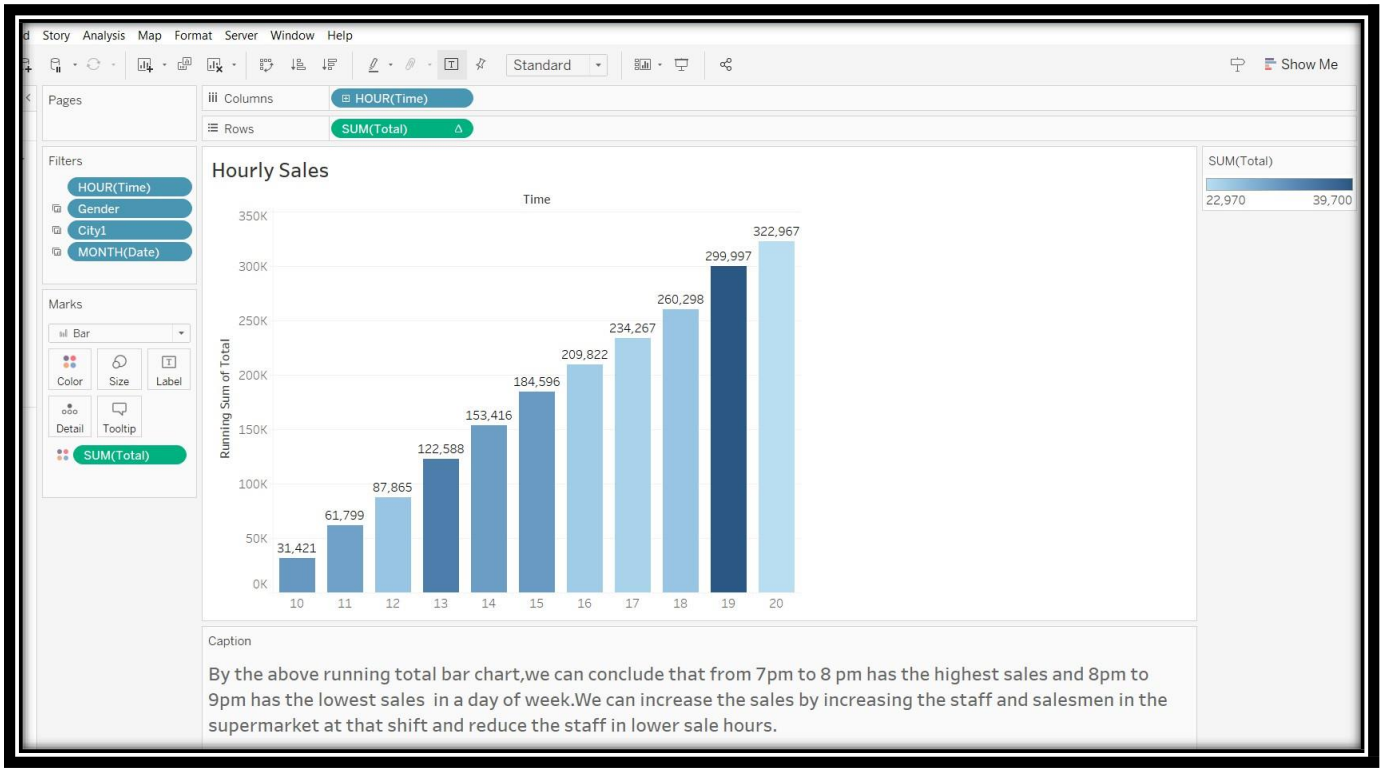
By performing this analysis, we will get the Hourlysales in day.By which we can conclude which hour has highest sales and lowest sales.

Description*:* It is to see the most hourly sales in 2020.I used Bar chart for this analysis and it is best for this objective.

Formulas and Functions Used : I used the **Running Total** where we can see the variation in increase in sales from one hour to another.

Analysis results: In the year 2020, most hourly sales in 2020 :

In 2020, the most hourly sales are between 7pm – 8pm and least sales is in between 8pm – 9pm



5) Total sales with respect to Payment Type

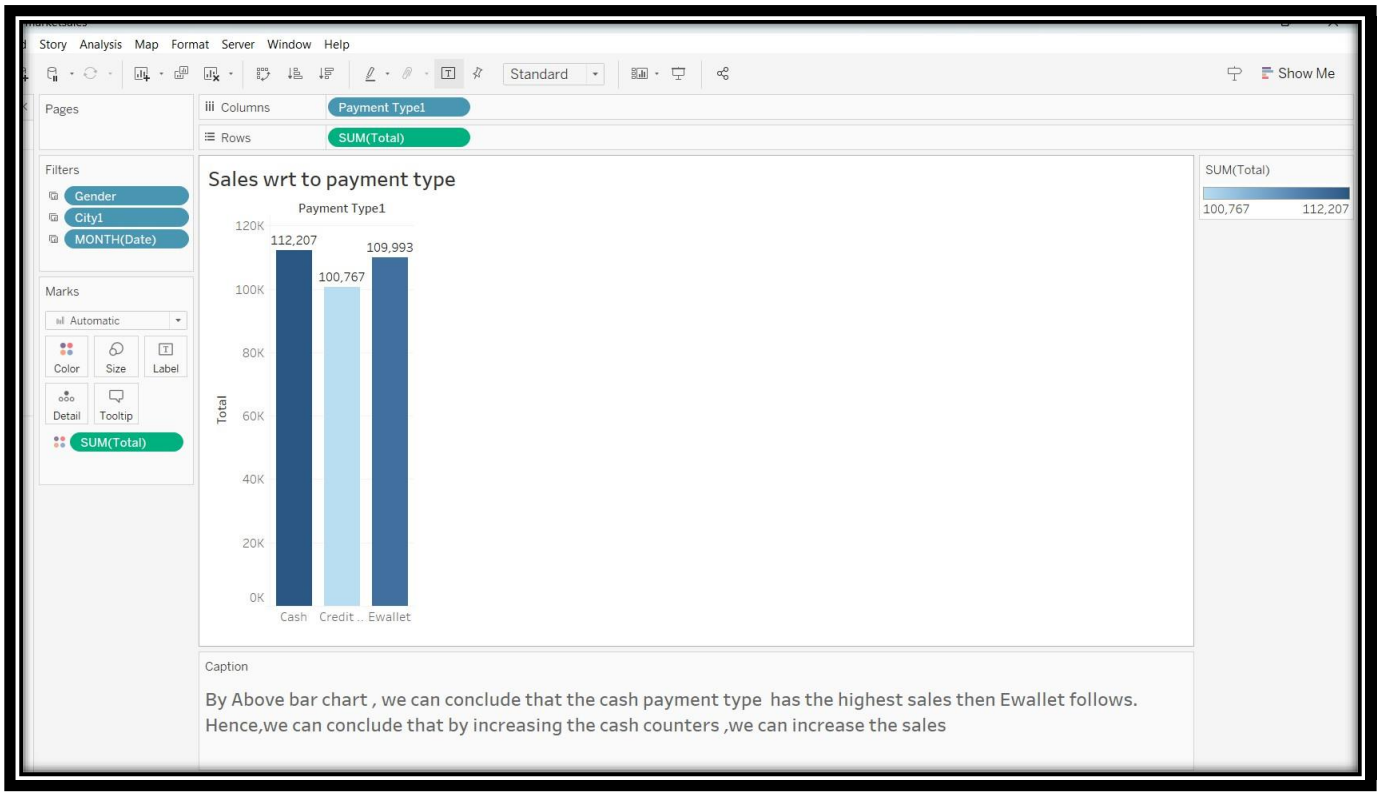
Introduction

By performing this analysis, we will get the conclusion of what payment type is used by the customers for payments.

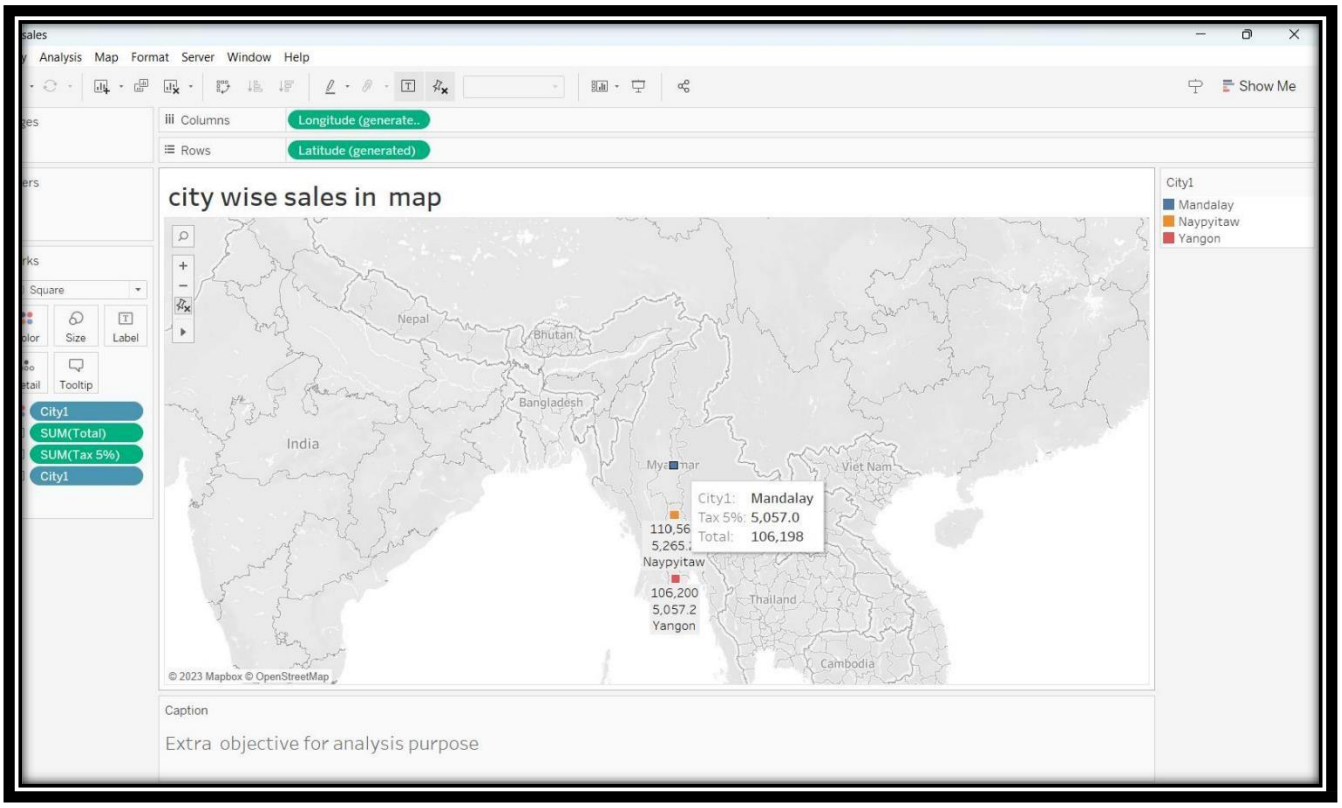
Description*:* It is to see the most sales with respect to the payment type done by customers. I used the bar chart for this analysis and it is suitable for easy analysis for this objective.

Functions /Formulas Used: I have used to sum function to calculate totalratings by customers=SUM() and I used the filters to analyze the effectively.

Analysis results: In the year 2020, we can conclude that Cash payment type has the highest sales compared to other payment types .



6) Sales, Tax of cities in geographical Location(map):

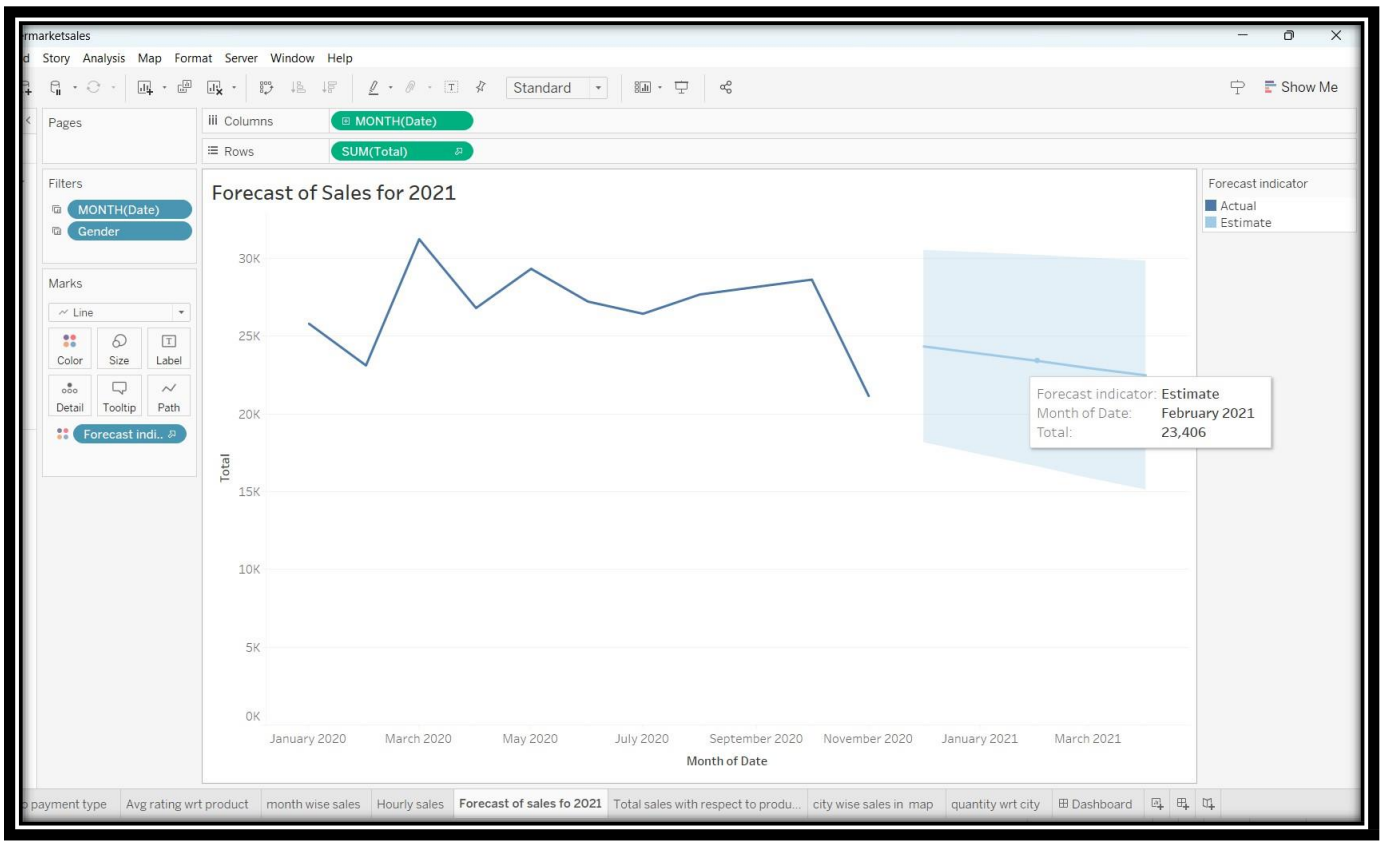


7) Forecast of sales for 2021:

Introduction: In this analysis, we are predicting the future sales of four months of year 2021 using the data of 1-year sales of 2020.

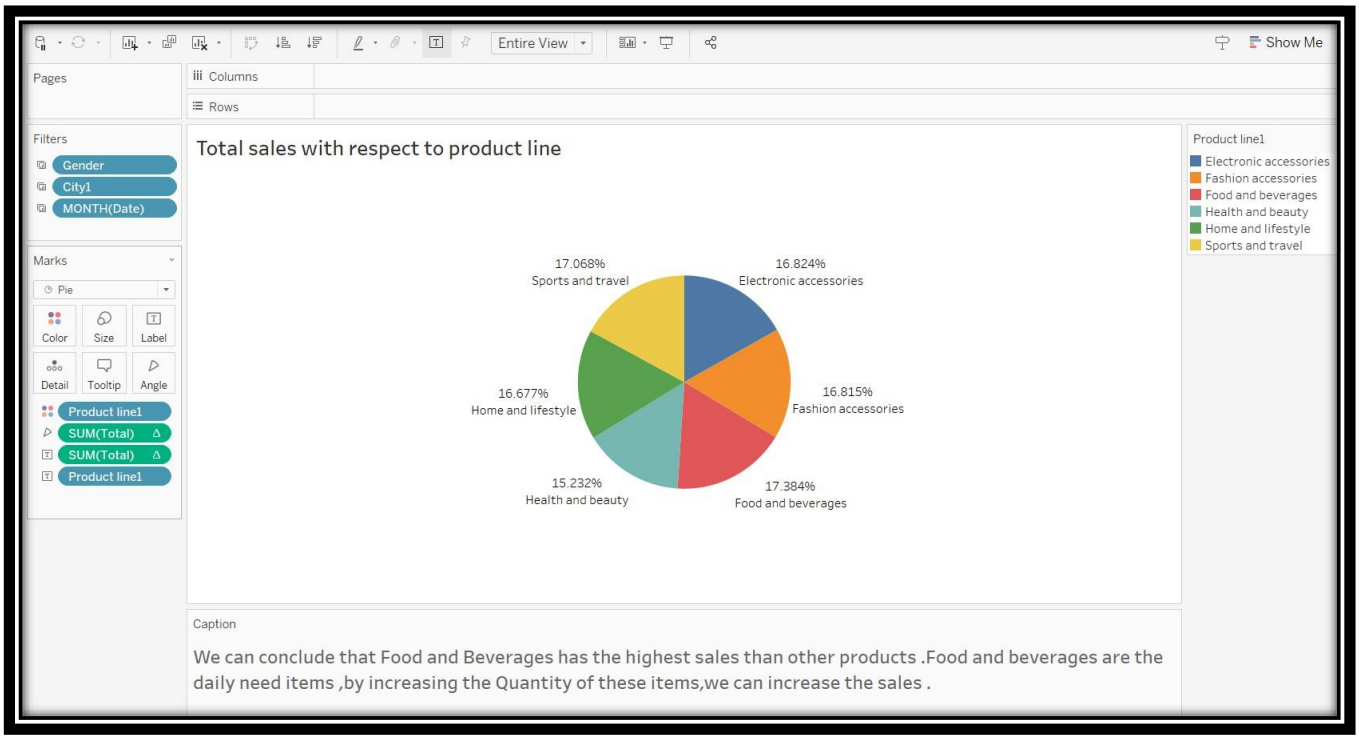
Description: In this analysis, I had used Forecasting for future prediction of sales and It is only for estimation purposes of sales of supermarket sales.

Analysis and Results: By the Forecasting we can conclude that there is an estimation of loss in sales for next four months in year 2021.

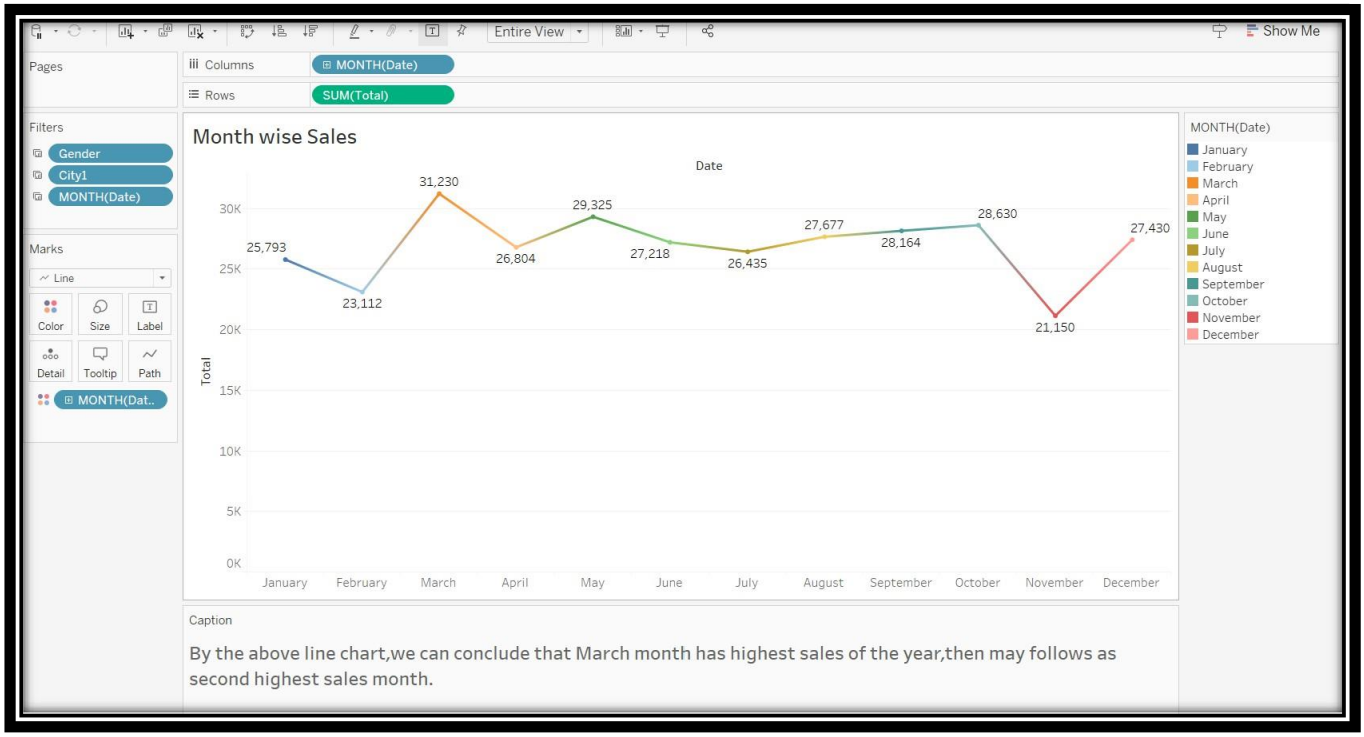


List of Analysis of Results :

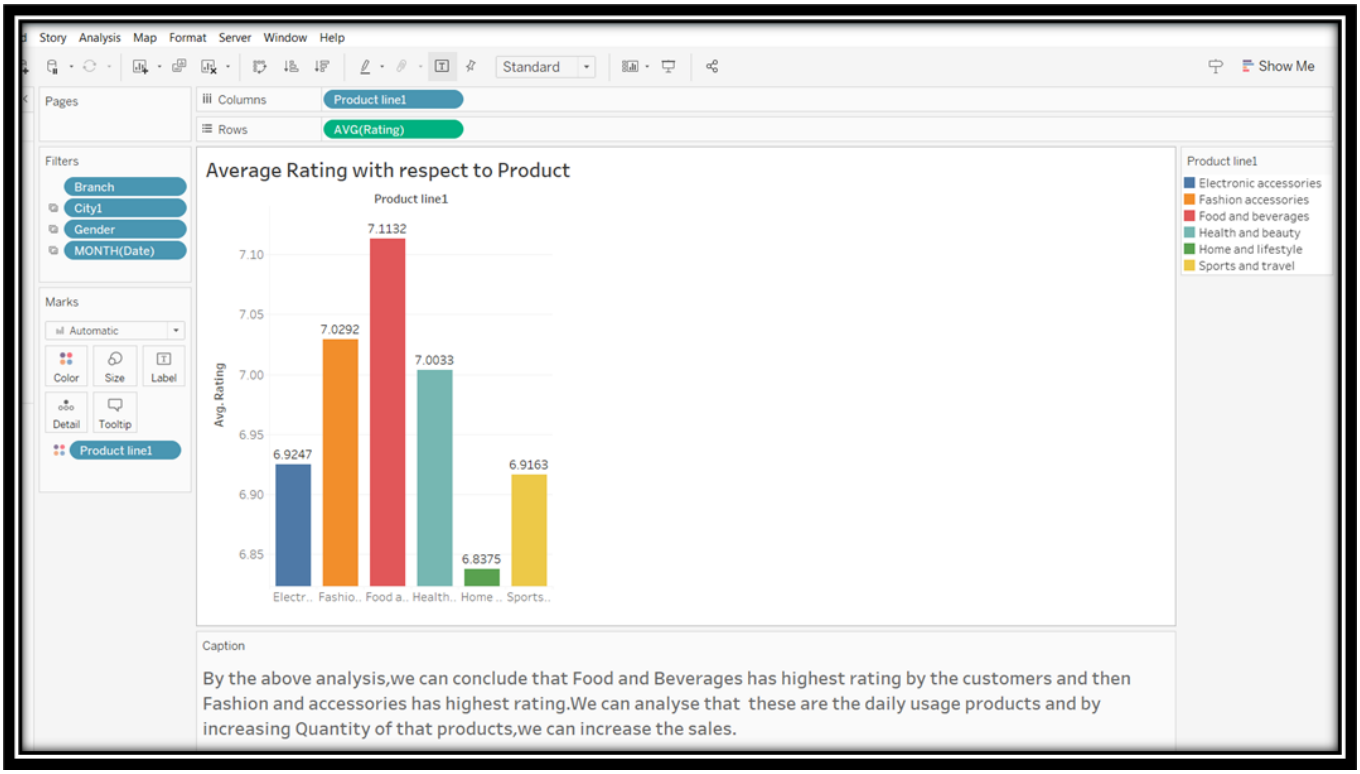
1) Total Sales by product Type:



2) Total Sales By Month

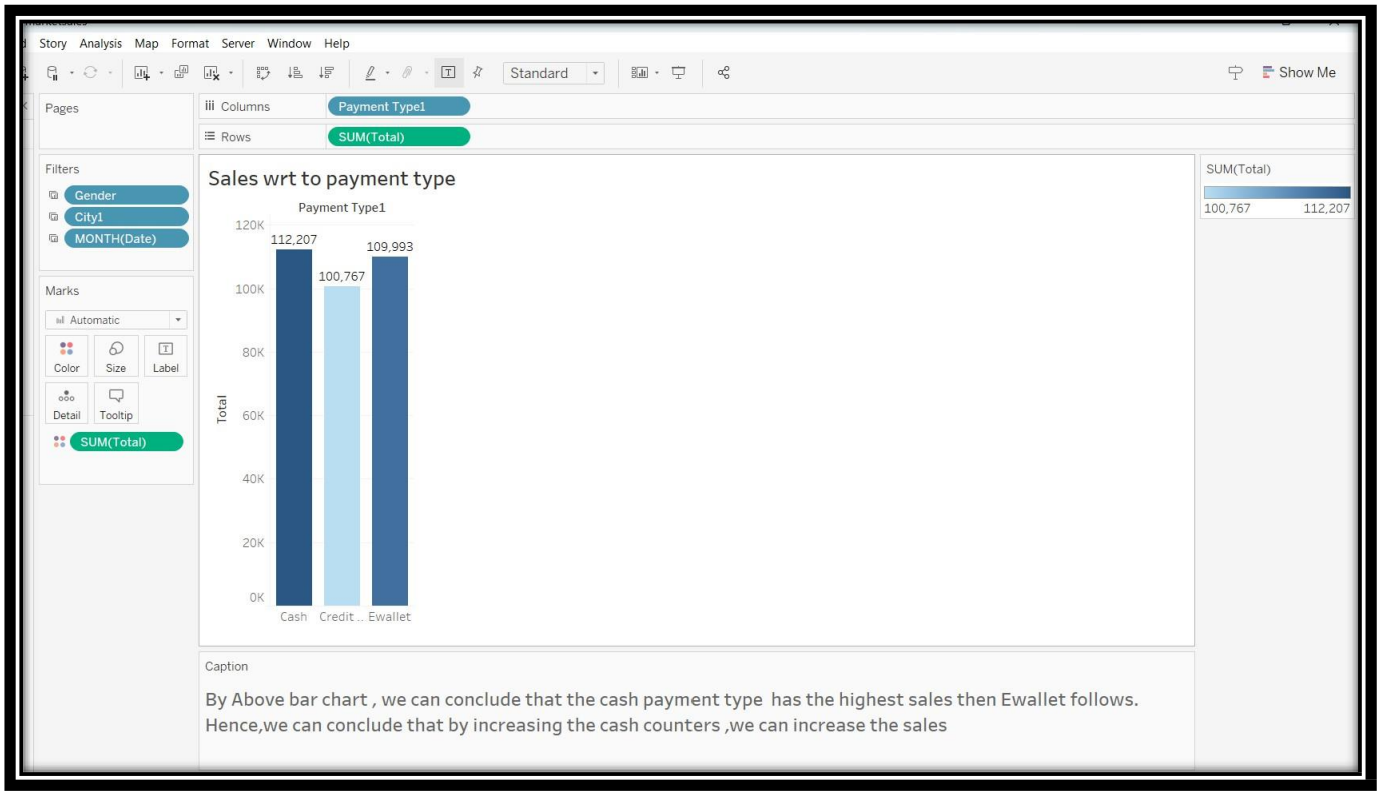


3) Average rating on product :

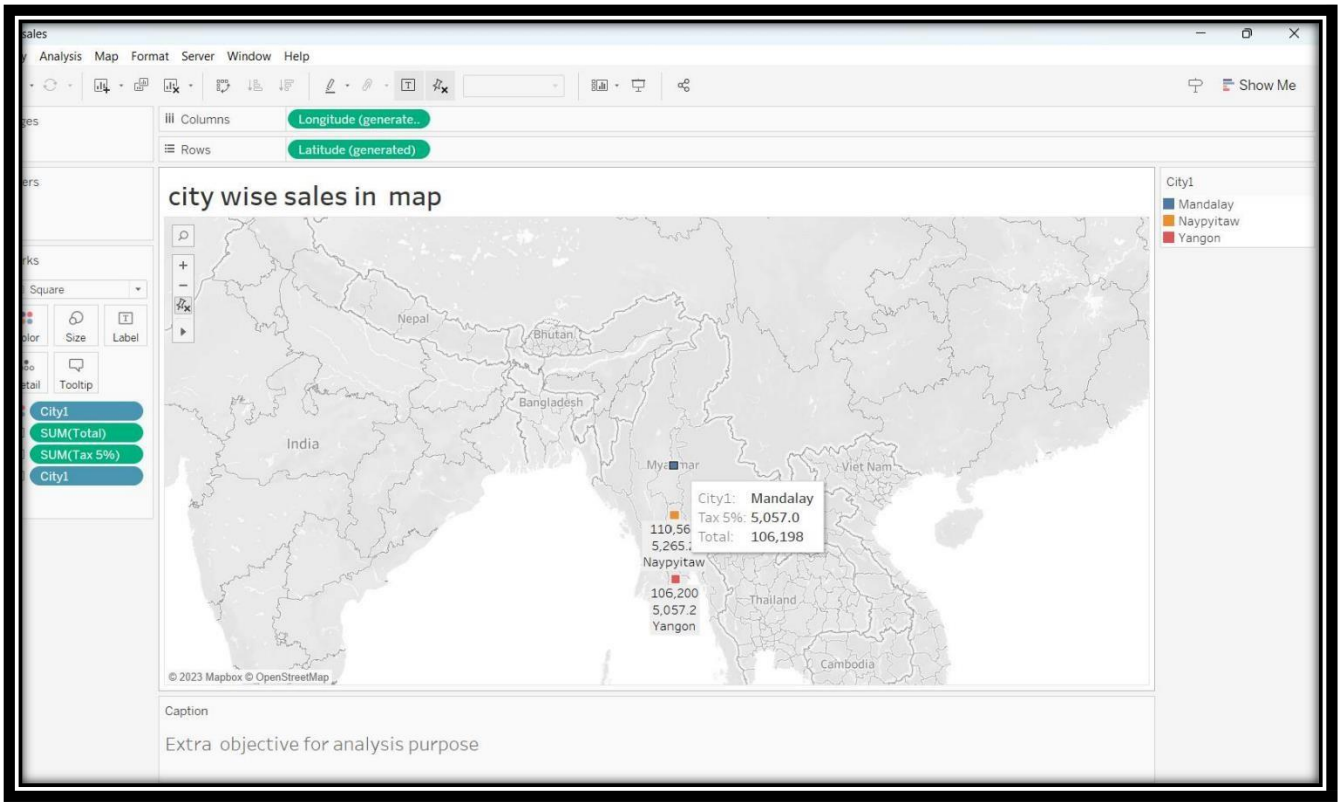


4) Most Sales by Time :

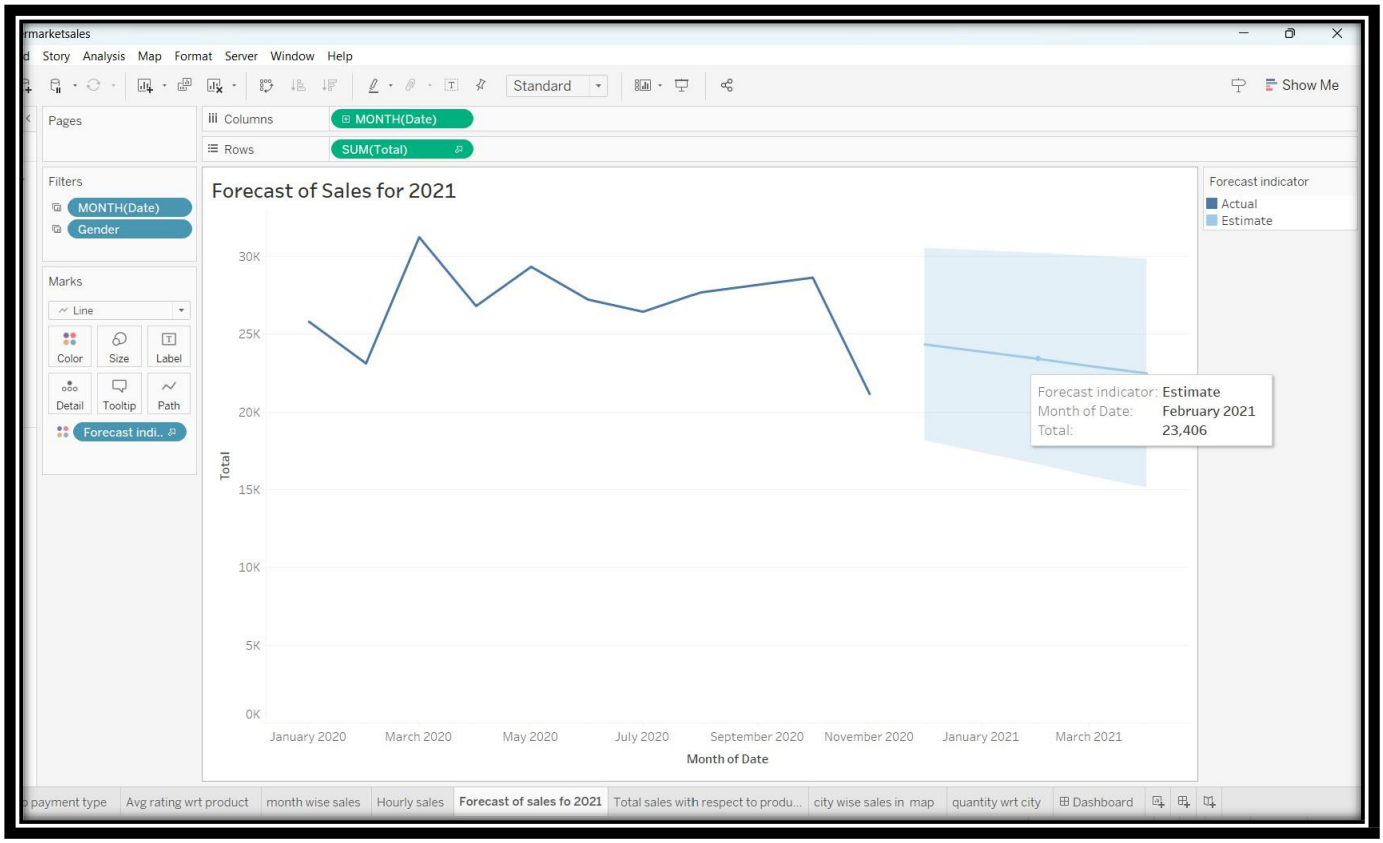
5) Total Sales with respect to Payment type



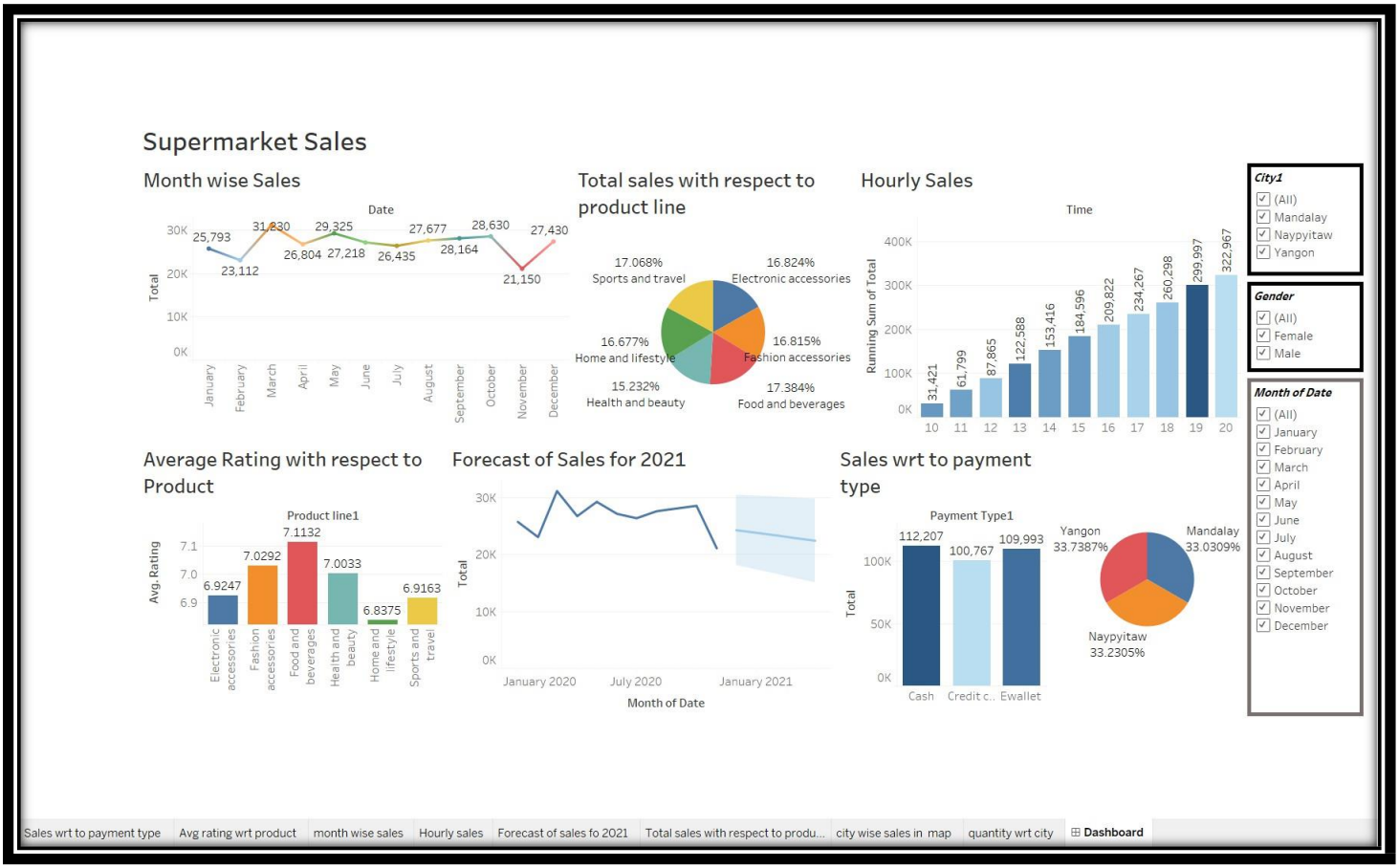
6) Sales, Tax of cities in geographical Location(map):



7) Forecast of sales for 2021



Final Dashboard :



Future scope :

The dataset provides information on the sales of products across different categories, stores, and regions, along with other relevant features such as product price, promotion, and customer demographics. Some of the potential future scopes for using this dataset in a project report are:

Predictive modeling: The supermarket sales dataset can be used for building predictive models to forecast future sales trends. Machine learning algorithms such as regression, time series analysis, and clustering can be used to identify patterns in the data and make accurate sales predictions.

Customer segmentation: The dataset can be used for customer segmentation analysis to identify different customer segments based on their buying behavior, demographics, and product preferences.

This information can help retailers to tailor their marketing and promotional strategies to specific customer segments.

Inventory management: The dataset can be used for inventory management to optimize the product stock levels across different stores and regions. This can help retailers to avoid stock-outs and overstocking, which can result in loss of sales and profit.

Product recommendations: The dataset can be used for building product recommendation systems based on Customer purchase history and preferences. This can help retailers to offer personalized product recommendations to customers and improve customer satisfaction.

Store performance analysis: The dataset can be used for analyzing store performance metrics such as sales growth, foot traffic, and customer satisfaction. This can help retailers to identify areas of improvement and optimize their operations to increase sales and profitability. Pricing optimization: The dataset can be used for pricing optimization to determine the optimal price for different products based on customer demand, competition, and other factors. This can help retailers to maximize their revenue and profit margins.

Promotions analysis: The dataset can be used for analyzing the effectiveness of different promotional strategies such as discounts, coupons, and loyalty programs. This can help retailers to identify the most effective promotions and optimize their promotional spending.

Market basket analysis: The dataset can be used for market basket analysis to identify the products that are frequently purchased together. This can help retailers to improve their cross-selling and upselling strategies and increase their average transaction value.

Geospatial analysis: The dataset can be used for geospatial analysis to identify the sales trends and patterns across different regions and stores. This can help retailers to identify the high-performing stores and regions and make informed expansion decisions.

Customer lifetime value analysis: The dataset can be used for customer lifetime value analysis to determine the profitability of different customer segments over time. This can help retailers to identify the most valuable customers and allocate their marketing and promotional resources accordingly.

Overall, the supermarket sales dataset on Kaggle offers a wide range of possibilities for data analysis and can be used to address various business challenges in the retail industry.

References:

[1] S. Garg, "Data Visualization through Tableau," Coursera, 2024. [Online]. Available: https://www.coursera.org/specializations/data-visualization. [Accessed: Apr. 07, 2023].

[2] Tableau Software, "Tableau Tutorial for Beginners," YouTube, 2021. [Online]. Available: https://www.youtube.com/playlist?list=PLWPirh4EWFpGXTBu8ldLZGJCUeTMBpJFK. [Accessed: Apr. 07, 2023].

[3] N. Sharma, N. Jain and N. K. Singh, "Data Visualization through Tableau," in 2018 3rd International Conference on Computing Methodologies and Communication (ICCMC), Erode, India, 2018, pp. 129-131, Doi: 10.1109/ICCMC.2018.8472975.

[4] Tableau Software, "Charts," Tableau, 2022. [Online]. Available: https://www.tableau.com/data-insights/reference-library/visual-analytics/charts. [Accessed: Apr. 07, 2023].

Yang, H., Zhang, Y., & Gao, M. (2020). Visual Analysis of Supermarket Sales Data Using Tableau. In Proceedings of the 2nd International Conference on Electronics, Communications and Control Engineering (pp. 177-181).

Baek, H., & Kim, Y. (2020). Analysis of supermarket sales data using Tableau: a case study. Journal of the Korean Society of Marine Environment and Safety, 26(1), 78-84.

Kim, D., & Kim, K. (2019). An Analysis of Supermarket Sales Data using Tableau. Journal of the Korea Society of Computer and Information, 24(9), 79-8

Bibliography

Dataset from Kaggle

https://www.kaggle.com/datasets/aungpyaeap/supermarket-sales

Youtube videos for Etl process

https://www.youtube.com/watch?v=_jmiEGZ6PIY