



SCHOOL OF COMPUTER SCIENCE

ASSESSMENT TASK 1 (Weightage 30%) (INDIVIDUAL ASSIGNMENT)


SEPTEMBER 2024 SEMESTER

MODULE NAME	: DATA MINING
MODULE CODE	: ITS61504
DUE DATE	: 23.09.2024- 07.10.2023, 11.59PM (NPT)
PLATFORM	: MyTIMES

This paper consists of EIGHT (8) pages, inclusive of this page.

STUDENT DECLARATION

- 1. I confirm that I am aware of the University's Regulation Governing Cheating in a University Test and Assignment and of the guidance issued by the School of Computing and IT concerning plagiarism and proper academic practice, and that the assessed work now submitted is in accordance with this regulation and guidance.*
- 2. I understand that, unless already agreed with the School of Computing and IT, assessed work may not be submitted that has previously been submitted, either in whole or in part, at this or any other institution.*
- 3. I recognize that should evidence emerge that my work fails to comply with either of the above declarations, then I may be liable to proceedings under Regulation*

Student Name	Student ID	Date	Signature	Score
Sandip Man Singh Mahat	0362781	2024 oct 7		

Assignment 1 Marking Rubrics					
Criteria	Excellent (90-100)	Good (75-89)	Average (40-74)	Poor (0-39)	Weight
Data Understanding and Summary	Thorough and clear understanding of all attributes, excellent summary of the dataset, and insightful analysis of the importance of variables.	Good understanding of most attributes, decent summary of the dataset, and some insights on variable importance.	Basic understanding of the dataset, summary provided but lacking depth or explanation of the importance of variables.	Poor understanding of the dataset, minimal or incorrect summary, and little or no explanation of the variables' importance.	15
Data Cleaning and Preprocessing	Dataset is thoroughly cleaned with appropriate handling of missing values, duplicates, and format inconsistencies. Data types are well-handled, and logical preprocessing steps taken.	Good data cleaning with some minor issues or omissions. Missing values, duplicates, and data types are mostly handled correctly. Some preprocessing steps are present.	Basic data cleaning but some issues with handling missing data or duplicates. Some formatting problems persist, and limited preprocessing steps.	Inadequate or no data cleaning, missing values, duplicates, or formatting issues persist. Little to no preprocessing applied.	25
Feature Engineering	Excellent creativity in generating new, useful features that significantly improve analysis. Well-justified reasoning for each feature.	Good effort in feature generation, with most features being useful for analysis. Reasonable justification for added features.	Some attempt to generate features, but they may be simple or lack a clear connection to deeper analysis. Justification is weak or incomplete.	Little to no attempt to generate new features or features are irrelevant/unhelpful for analysis. No justification provided.	20
Data Aggregation	Data is efficiently grouped and summarized where necessary, with appropriate logic for grouping. Insights from aggregation	Good grouping and summarization, though minor issues may exist. Most insights are clear, though a few might be underdeveloped.	Some grouping and summarization, but lacks depth or clarity. Insights are present but not fully developed or explained.	Little to no grouping or summarization of data. Insights are unclear or missing entirely.	10

	are clear and impactful.				
Data Visualization	Visualizations are highly relevant, clear, and effective in conveying insights. They are well-labeled and integrated into the analysis.	Good visualizations that are relevant and mostly effective in conveying insights. Minor issues with clarity or labeling.	Basic visualizations that are somewhat relevant, but lack clarity or fail to fully support the analysis. Some labeling or explanation issues.	Little to no relevant visualizations or visuals are unclear, misleading, or poorly explained.	20
Report Writing	Report is highly organized, with clear explanations, logical flow, and strong depth of analysis. Conclusions are insightful and well-supported by evidence.	Well-organized report with mostly clear explanations and logical flow. Analysis is good but lacks some depth or minor issues in clarity. Conclusions are reasonable but underdeveloped.	Basic organization but lacks clarity or depth. The flow of ideas may be disjointed, and conclusions are present but not well-supported.	Report is poorly organized with unclear explanations, lack of logical flow, and unsupported or missing conclusions.	10

Table of Contents

1.Abstraction	8
2.Introduction.....	8
3.Problem Statement and Objectives	9
4.Data Understanding.....	10
5.Data Exploration	11
6.Data Preprocessing	16
i.Data Cleaning	16
ii.Feature Engineering	18
iii.Data Aggregation	22
7.Data Visualization	26
i.Total order count by state	26
ii.Total sales by city	27
iii.Total sales by months.....	28
iv.Total sales by product	29
v.Total sales by state	30
vi.Total sales by timeframe and month.....	31
vii.Total Sales Distribution by Time Frame	32
viii.Sales by time Frame and state	33
ix.Customer Segmentation by total and average order size	34
x.Product co-occurrence matrix	35
xi.Geospatial analysis of order.id location as per their city	36
xii.Geospatial analysis of customers buying more than one items with total sales.....	37
8.Conclusion and suggestions	38
9.References.....	40

Table of Tables

Table 1:raw dataframe.....	10
Table 2:Total sales and average quantity per month	12
Table 3:table with timeframe coloum	18
Table 4: order.id with different many products	20
Table 5:combined products	23

Table of Figures

Figure 1:summary of data after cleaning	16
Figure 2:Total order count by state	26
Figure 3:Total sales by city	27
Figure 4:.Total sales by months	28
Figure 5:Total sales by product.....	29
Figure 6:Total sales by state	30
Figure 7:Total sales by timeframe and month.....	31
Figure 8:Total Sales Distribution by Time Frame	32
Figure 9:Sales by time Frame and state.....	33
Figure 10:Customer Segmentation by total and average order size	34
Figure 11:Product co-occurrence matrix	35
Figure 12:Geospatial analysis of order.id location as per their city	36
Figure 13:Geospatial analysis of customers buying more than one items with total sales ..	37

Table of codes

code 1:basic statistics.....	12
code 2:handling missing value	15
code 3:converting string into numeric.....	15
code 4:format date in proper order	16
code 5:converting time into numerical variable	17
code 6:splitting a day into 3 time frame	17
code 7:finding many product from same order.id	19
code 8:using apriori algorithm.....	21
code 9:new column total.sales	22
code 10:different orderid product combining	23
code 11:adding new columns latitude and longitude.....	24
code 12:longitude and latitude column.....	25
code 13:Total order count by state.....	26
code 14:Total sales by city	27
code 15:Total sales by months	28
code 16:Total sales by product	29
code 17:Total sales by state.....	30
code 18:Total sales by timeframe and month	31
code 19:Total Sales Distribution by Time Frame	32
code 20:Sales by time Frame and state.....	33
code 21:Customer Segmentation by total and average order size.....	34
code 22:Product co-occurrence matrix	35
code 23:Geospatial analysis of order.id location as per their city	36
code 24:Geospatial analysis of customers buying more than one items with total sales	37

1.Abstraction

In this report, we analyze the daily sales of a retail internet business based on its monthly data for the year 2019. Exploratory, transformation and descriptive data analysis techniques are employed in an attempt to extract meaningful insights with regards to the company's performance. Since the dataset serves to enhance business strategies, the information is cleaned to eliminate possible noise and enhance readability. The report then continues to data analysis to Sales analysis, finding the Sales trends, Product analysis, Popular products, Sales by Geography. From this extensive review, strategic suggestions are provided regarding product portfolios, business processes and clientele. These insights are meant to help the company make strategic decisions and develop in a tough retail landscape.

2.Introduction

In this generation Due to high technological advancement, e-commerce has become the new way of doing business where different and diverse firm get to supply their products on the internet so as to reach as many consumers as possible all over the world. This shift has not only intensified competition but also it has provided plethora of information that need to be capitalized in purchase behavior studies, measuring the effectiveness of business models, and enhancing the business execution plans.

This paper aims at evaluating the sales of an e-business company that undertakes sales across multiple regions of the United States. This dataset covers a whole year of 2019 and has numerous types of data such as product's name, price, quantity sold, the date of the order and the location of the customer's country. With these attributes, we can explore important questions such as:

- Which products generate the most sales' revenue?
- At what times monthly and yearly sales are the highest and in what periods of the year they are the lowest?
- How does the buying behaviour of customers differ in different region?

Therefore, the major goal of this analysis is to gain better understanding of the data with aim of improving decision making in regard to product stock, promotional methods and communicating with customers. This is discontinued through analyzing the set data, pre-processing the data with an aim of rejecting inconsistent data, and presenting the set data in a graphical fashion to arrive at coherent conclusions. The objective of this report does not limit to presenting figures of sales only. It aims at delivering rich insights into the sales and profit drivers and suggested changes to its business processes and resource management.

3.Problem Statement and Objectives

today's retail is consumer-centric and the amount of information created by contemporary CBO transactions both makes and breaks opportunities. It follows then that the purpose of this report is to directly apply the given sales data to pinpoint problem areas in sales processes that prevent optimal customer satisfaction respective to the efficient consumption of available resources. However, this analysis is not without its challenges:

- **Data Quality Issues:** The information set also has gaps, and a variety of values in one cell, which can shift the analysis in an undesired way.
- **Understanding Customer Behavior:** Thus, there is a need in analyzing patterns related to when and how exactly customers buy retail produce to establish an effective strategy. Does it mean they're probably purchasing specific products during occasional periods in a year? Is demand from the customer different at different regional locations or at different time periods of the day?
- **Inventory Management:** Since there could be fluctuations in demand during some periods of the year, say end of year, it is crucial to avoid having very large stocks as well as very little stocks.
- **Optimizing Marketing Campaigns:** Also, it should expose which products are viewed most positively in certain areas with the aim of targeting the right market segments via marketing promotions.

Through overcoming these challenges, this report offers straightforward strategic advice for enhancing the business model supported by the data.

Specific Objectives

- **Sales Trend Identification:** Find out when the current business encounters its high and low sales volume in any month of the year 2019.
- **Product Demand Analysis:** Assess which items largely make up total units and revenues sold by the firm.
- **Regional Sales Analysis:** Analyze fluctuations in sales by the location and define areas for marketing development or operative specialization.
- **Customer Behavior Insights:** To determine whether there were different trends in customer purchasing, the time of the day should be considered along with the possibility of a relationship between different product categories.
- **Recommendations for Optimization:** As such, suggest concrete recommendations about inventory management, marketing appeals and the regions of interest.

4.Data Understanding

As with every analysis, those carried out on datasets involve a few preliminary steps that should never be overlooked. The sales data is in form of monthly CSV files where each file consists of records for the occurrence of transactions. Every record represents a sale and includes several key pieces of information:

Table 1:raw dataframe

Order ID	Product	Quantity Ordered	Price Each	Order Date	Purchase Address
176558	USB-C Charging Cable	2	11.95	4/19/19 8:46	917 1st St, Dallas, TX 75001
176559	Bose SoundSport Headphones	1	99.99	4/7/19 22:30	682 Chestnut St, Boston, MA 02215
176560	Google Phone	1	600	4/12/19 14:38	669 Spruce St, Los Angeles, CA 90001
176560	Wired Headphones	1	11.99	4/12/19 14:38	669 Spruce St, Los Angeles, CA 90001

- Order ID: This keeps each transaction separate from the other and gives it its own identification number. Order ID is used for multiple products ordered at one time to work on the same order.
- Product: The item offered for sale in the transaction, which is also the name of the product.
- Quantity Ordered: The amount of units bought/ Frequency.
- Price Each: The cost of a single product unit under consideration.
- Order Date: The time and date on which the order was placed.
- Purchase Address: The address to which the order was sent and delivered to.

This set up already gives us a solid framework from where we can work from. However, certain issues need to be addressed:

- **Inconsistent Data Types:** The Quantity Ordered and Price Each columns for example are stored as strings and not numbers. This will need to be corrected. The short-term perspective characteristic of contemporary approaches to school leadership will need to be corrected as well.
- **Missing and Duplicated Data:** Like any dataset, some records are missing some values, some records may be a duplicate record of another record. These have to be defined and excluded so as to minimize bias.
- **Temporal Data:** Exact prices and other numerical values are presented in the adjacent cells of the Price and Quantity fields, respectively, which will enable performing calculations of total amounts. The Order Date column contains timestamps that will allow quantitative analysis of sales at various time intervals but that require separate extraction of time components like day, month, or hour.

Knowledge of these will help guide the process of data cleaning and data preprocessing to get the dataset in a good state for analysis.

5.Data Exploration

Thus, as soon as the dataset was preprocessed, exploratory data analysis can begin. This entails sorting the data in search of other distinctions, or in the search for other regularities or irregularities or dependencies. The following sections highlight some of the key findings from the exploration:

i. Summary of the Dataset

From the data set, a breakdown of the univariate analysis was undertaken solely on the population of January 2019. Some of the records contain blank values in certain fields including Order ID, Product and Quantity Ordered, though the file has 9,723 records. A summary of key statistics for the dataset reveals the following:

Total Records: 9,723

Total Sales: It will be calculated by using the formula; Quantity ordered x Price each.

Average Quantity Ordered: To be determined.

Most Frequent Product: The best selling products are expected to be the accessories such as “Lightning Charging Cable” and “Wired Headphones”.

```
# Basic statistics: Total sales and average quantity ordered
total_sales <- sum(df$Total.Sales, na.rm = TRUE)
avg_quantity <- mean(df$Quantity.Ordered, na.rm = TRUE)

# Display results
cat("Number of records:", num_records, "\n")
cat("Total Sales:", total_sales, "\n")
cat("Average Quantity Ordered:", avg_quantity, "\n")
print(range_values)

# Calculate sales and average for each month
monthly_summary <- df %>%
  group_by(Month) %>%
  summarise(
    Total_Sales = sum(Total.Sales, na.rm = TRUE),
    Average_Quantity = mean(Quantity.Ordered, na.rm = TRUE)
  )

# Sorting the monthly summary by Total Sales
monthly_summary_sorted <- monthly_summary %>%
  arrange(Total_Sales)
```

code 1:basic statistics

Table 2:Total sales and average quantity per month

Month	Total Sales	Average Quantity
April	3,396,059	1.12461
August	2,244,412	1.124195
December	4,619,297	1.125335
February	2,203,481	1.12306
January	1,815,335	1.122611
July	2,646,900	1.124414
June	2,578,293	1.125332
March	2,809,063	1.122212
May	3,144,585	1.127039
November	3,198,909	1.126735
October	3,736,884	1.119355
September	2,098,817	1.128128

ii. Patterns and Trends

Popular Products: Preliminary results suggest that supposedly, accessories, including charging cables and headphones, are among the most demanded. But this kind of analysis is needed to determine which product contributes most to the overall revenues.

Temporal Trends: Using the data in the column named “Order Date,” the sales can be grouped by day of the week or month, and it will be possible to see that some days are more popular among buyers, for instance, when people prefer to make gifts – on holidays.

1. Product Performance

In this case, this has been done by summing up the sales data by the product to get an indication of which products brought in the most amount of money. The top-performing products included:

- **iPhone:** However and surprisingly, iPhone was among the most sold products which are an important source of revenue even in the face of its expensive prices.
- **Laptops and Electronics:** Other expensive goods that include laptops and home electronics are also above expectations.
- **Accessories:** Many low cost products like cables, chargers particularly USB cables for different types of portable electronic gadgets were offered in vast proportions though constituted lesser proportion to the total sales than did the costly items.

2. Geographical Analysis

This field was utilised in breaking down sales by city and state within the Purchase Address field. The top cities in terms of sales volume included:

- **New York City:** Not surprisingly the greatest number of sales were yielded by New York since it is the largest city in the United States.
- **Los Angeles:** Another large metropolitan area, Los Angeles, ranked second overall in a total number of sales.
- **Chicago:** The third largest city by population, Chicago complied heavily to overall revenue.

Notably, some other mid-sized cities in Midwest and Southern states likewise fared well, possibly indicating markets that are currently under-exploited for growth.

3. Another perspective that can help categorise customers is the time of day they visit your business.

Ordering the analysis of customer activity by the hour, we removed the hour from the Order Date field. The results showed that:

- **Morning and Afternoon Peaks:** The most frequent sales were recorded during 10:00 am to 3:00 pm, therefore customers are most likely to be more active during these hours.
- **Evening Decline:** It was apparent that the sales started slowing down after 6 p.m with little or no sales being made at night.

4. Product Co-Occurrence Analysis

This is about looking at two products and analyzing the way that they occur simultaneously in the market.

The given code helps to analyze customer's basket or purchases to know which products frequently purchased together. Such an analysis can help to predict customer purchases and is available for using in the cross-sell and package deals. The technique employed here produces the Product Co-Occurrence Matrix to express the extent to which two products are jointly purchased. Here's an in-depth analysis of each step of the code:

The product correlation matrix also demonstrates the similarities between products in terms of joint purchases. Here are some example interpretations that might emerge from this analysis:

- **iPhone and Accessories:** iPhones and related accessories (Chargers, iPhone cases, iPhone USB cables) are usually purchased together. This is expected because the customers who are buying a phone are likely to also purchase accessories to the phone.
- **Laptops and Peripherals:** Pertaining to accessories, laptops also frequently belong to accessories like the laptop's mouse, keyboard or case. It can also help to spot what additional peripherals are most frequently sold with laptops, so these can be grouped for customers' benefit and to benefit from sales increases.
- **High vs. Low-Cost Items:** An insightful observation from this matrix might be the fact that consumers do not mind buying low-cost accessories for high-cost items, such as headphones for laptops or smartphones or USB cables for laptops. Such cross category coupling show that there is potential to cross sell.
- **Electronics and Entertainment:** But if we want examine entertainment products then pair plots may show rather high correlation, for instance consoles and TVs or home appliances and electronics.

Such insights help in decision-making about the timing when promotions should be featured to when inventory should be replenished and when, in terms of staffing, customer support should be increased.

iii. Handling Missing Data

From the dataset, there are so many limitations, including the presence of many unknown values in aspects such as the Order ID, Product, and Quantity Ordered. To handle this:

```
# Remove rows with missing or NA values in any column
df <- df %>%
  filter(complete.cases(.)) # Removes rows with NA

# Optionally, drop empty strings as well
df <- df %>%
  filter_all(all_vars(. != ""))
```

code 2: handling missing value

Outliers were not addressed due to time constraint, but rows containing missing values were deleted since these fields are important for analysis.

Other options in regard to missing data, including imputations or even the mere application of mean values for missing values as well as linear interpolation for patterned missing data, could be future investigations considering the consequences of missing data.

iv. Duplicates

There is duplicate heading of the of the datasets which is repeating many time in overall dataset. In order.id if we found categorical value then we remove overall row in the

```
# Ensure 'Quantity.Ordered' and 'Price.Each' are numeric
df$Quantity.Ordered <- as.numeric(df$Quantity.Ordered)
df$Price.Each <- as.numeric(df$Price.Each)
```

dataset

code 3: converting string into numeric

v. Date and Time Analysis

This was done to support temporal analysis, the field “Order Date” was changed into a datetime those of you who prefer receipts. Dividing the date into its components (month, day, hour) will allow to focus on certain time intervals and periods, when sales is high or on the contrary, low, at different periods of the year

```
#correcting the date format in a proper manner
df$Order.Date <- as.POSIXct(df$Order.Date, format = "%m/%d/%y %H:%M")
df$Month <- format(df$Order.Date, "%m")
df$Day <- format(df$Order.Date, "%d")
df$Hour <- format(df$Order.Date, "%H")
```

code 4:format date in proper order

6.Data Preprocessing

Pre-processing of data is an essential part of the data analysis process in any analytical data flow. Cleaning all your data and restructuring your raw data into a format that is suitable for analysis is called data preprocessing. In this section, the process of data cleansing, data feature extraction, and data reorganization which are critical for the extraction of features from datasets have been explained.

i.Data Cleaning

The raw data carry a common set of problems such as errors, inconsistency or missing values that affect the quality of data. Thus, due to data cleaning, the derived dataset is credible to give proper results. Below are the steps taken to clean the sales dataset for further analysis:

1. Addressing Missing Values:

The observations where the values of these critical variables are missing were excluded from the dataset as well. While there are other pre-processing methods one can perform, for example, imputing on the data values where some of the numbers may be missing, in this case, they realised that some of the missing values are too large to guess and therefore any row that had missing data was eliminated.

Order.ID	Product	Quantity.Ordered	Price.Each	Order.Date	Purchase.Address	Month	Total.Sales	Day
Length:185950	Length:185950	Min. :1.000	Min. : 2.99	Min. :2019-01-01 03:07:00.00	Length:185950	Length:185950	Min. : 2.99	Length:185950
Class :character	Class :character	1st Qu.:1.000	1st Qu.: 11.95	1st Qu.:2019-04-16 21:05:15.00	Class :character	Class :character	1st Qu.: 11.95	Class :character
Mode :character	Mode :character	Median :1.000	Median : 14.95	Median :2019-07-17 20:40:30.00	Mode :character	Mode :character	Median : 14.95	Mode :character
		Mean :1.124	Mean : 184.40	Mean :2019-07-18 21:54:38.89			Mean : 185.49	
		3rd Qu.:1.000	3rd Qu.: 150.00	3rd Qu.:2019-10-26 08:14:00.00			3rd Qu.: 150.00	
		Max. :9.000	Max. :1700.00	Max. :2020-01-01 05:13:00.00			Max. :3400.00	
Hour								
Length:185950								
Class :character								
Mode :character								

Figure 1:summary of data after cleaning

The decision to use filtering and remove row with missing value instead of imputation was made based on the rationale on the field type in sales analysis. For example, “Order ID” or “Product” cannot be made out in this case; this is essential for identification of a transaction. Likewise, the absence of “Price Each” or “Quantity

Ordered” screw up the total sales figures. Therefore, the rows were omitted from the analysis in order to leave no trace of inconsistency or inaccuracy.

2. Converting Data Types:

The “Quantity Ordered” and “Price Each” fields were further coerced into numeric data type – float or integer so as to allow functionalities such as addition and averaging to be performed on them.

Converting such columns into numeric type enables configuration of straight-forward sales calculations. If this conversion is not done the dataset cannot perform a gross operation such as to give the total revenue of a particular product or the total quantity sold across all transactions.

3. Converting Date Column:

We changed the field called “Order Date” into datetime format so that working with dates and times becomes easier. The new format can also extract specific date components such as the year, month, day of the week and a particular hour.

```
#converting day hour and month in numerical variable
df$Month = as.numeric(df$Month)
df$Day = as.numeric(df$Day)
df$Hour = as.numeric(df$Hour)
```

code 5:converting time into numerical variable

```
# Create the 'time_frame' column based on the 'Hour' column
df <- df %>%
  mutate(
    time_frame = case_when(
      Hour >= 0 & Hour < 8 ~ "Morning",
      Hour >= 8 & Hour < 16 ~ "Evening",
      Hour >= 16 & Hour < 24 ~ "Night",
      TRUE ~ NA_character_ # In case there are any unexpected values
    )
  )

# Inspect the updated dataframe
head(df)
```

code 6:splitting a day into 3 time frame

Table 3:table with timeframe coloum

Order.ID	Product	Quantity Ordered	Price.Each	Street	City	State	Zip.Code	Month	Total .Sales	Day	Hour	time frame
141234	iPhone	1	700	944 Walnut St	Boston	MA	2215	1	700	22	21	Night
141235	Lightning Charging Cable	1	14.95	185 Maple St	Portland	OR	97035	1	14.95	28	14	Evening
141236	Wired Headphones	2	11.99	538 Adams St	San Francisco	CA	94016	1	23.98	17	13	Evening
141237	27in FHD Monitor	1	149.99	738 10th st	Los Angeles	CA	90001	1	149.99	5	20	Night
141238	Wired Headphones	1	11.99	387 10th st	Austin	TX	73301	1	11.99	25	11	Evening

Converting date field into datetime object is important for temporal analysis so that features like date can be used for valuable analysis as to where and when customers are concentrated. For example, it can be analyzed to know how much the sales are during the weekend, during the months of December, January or any other month preferred by buyers because of holidays and how much during the working hours, in the evening, night and so on.

ii.Feature Engineering

Feature engineering literally means the actual development of new features or modification of pre existing features for greater utility. It was necessary for the discovery of additional dimensions by enhancing the data

analytics that could not be identified while perusing the database. In this analysis, several new features were created to improve the dataset's usefulness:

1. Total Sales:

What is it? Most of the modifications can be found in the “Total Sales” column which shows the total value of each transaction. It is computed simply as the product of the two basic variables of the “Quantity Ordered” and “Price Each” for each product in the dataset.

Total Sales = Quantity Ordered x Price Each

Why is it important? This new feature enhances the possibility of quantitative analysis of the revenues.

Nevertheless, the amount set within the raw dataset includes the price per item and the quantity order; the “Total Sales” column calculates this into a single value so as to help understand how each product or transaction contributes to the total revenue.

Applications: with this feature we can do things like find out the most popular products, find total revenue by city or by customer, and average order volume. It also enables us to answer business-critical questions such as: Which product is most profitable? What dollar amount does the average customer spend on an order?

2. Address Components:

What is it? Like the “Purchase Address” the field “Address3” provides the details of the customer’s address in terms of street, city, state and postal code. To facilitate geographic analysis, the "Purchase Address" field was split into three distinct components:

```
# Find repeated Order.IDs only
repeated_orders <- df %>%
  group_by(Order.ID) %>%
  filter(n() > 1) # Only keep Order.IDs that appear more than once
```

code 7: finding many product from same order.id

Table 4: order.id with different many products

Order.ID	Street	City	State	Zip. Code	Products	Repeat Count	Total Sales
141275	610 Walnut St	Austin	TX	73301	USB-C Charging Cable, Wired Headphones	2	23.94
141290	4 1st St	Los Angeles	CA	90001	Apple Airpods Headphones, AA Batteries (4-pack)	2	161.52
141365	20 Dogwood St	New York City	NY	10001	Vareebadd Phone, Wired Headphones	2	411.99
141384	223 Jackson St	Boston	MA	2215	Google Phone, USB-C Charging Cable	2	611.95
141450	521 Park st	San Francisco	CA	94016	Google Phone, Bose SoundSport Headphones	2	699.99

City: Give the name of the city to which the product was shipped.

State: The state where the product was sent.

Postal Code: The alphabetical and numeric code of the shipping address description.

Why is it important? It becomes easy to analyze geography when the address is split into these components. For instance, it is only possible to analyze a large number of sales per city or state, and then understand which state is the most profitable. All this data can be helpful in navigation of marketing strategies, organizing the future store locations and deliveries.

Applications: It also allows asking such questions as: Which cities or states have the highest sales volume? Regionally, how does the revenue distribution occur or take place? When it comes to a given product, is that product more common in some areas than other?

Date Components:

What is it? To allow for more granular temporal analysis, the "Order Date" column was split into several components:

Year: The year the order was placed

Month: The date on which the order was placed.

Day of the Week: The exact date on which the order was placed (in the format week day, e.g. Monday).

Hour: To localize it, it can be divided with a slash to the hour of the day the order was placed, for instance 14:00 for 2 PM.

Why is it important? How can we obtain more detailed temporal dimensions out of the date By dissecting the date further, it is possible to perform additional temporal analysis. For instance, tracking sales by months will show whether there are monthly fluctuations or not while measuring sales by hour will inform us the most opportune time for a customer to buy.

Applications: This feature helps provide answers to questions such as: For which months are sales high? To find out whether many users shop during the weekends or during the weekdays. When are the customers most likely to spend? The information will be useful for advertisement, making stock, and adjusting store's working hours.

Apriori algorithm

Transaction data is scanned by the Apriori algorithm with minimum support (minimum supp = 0.001) and a minimum confidence (minimum conf = 0.4). This assists in finding the dense itemsets which occur with at least 0.1% of transactions with a confidence level of 40% (expressing the fact that if one itemset occurs, then the other will also occur).

```
# Applying Apriori algorithm with lower support
rules <- apriori(transactions, parameter = list(supp = 0.001, conf = 0.4))

# Check if rules are generated
if (length(rules) > 0) {
  # Display top rules sorted by lift
  inspect(sort(rules, by = "lift")[1:10])
} else {
  print("No rules found. Try lowering the support or confidence.")
}
```

code 8:using apriori algorithm

no rules are discovered, the support or confidence levels may be reduced further in order to observe the infrequent patterns.

The discovered association rules can therefore be implemented as features for different analysis such as the market basket analysis, recommended systems and customer classification. They give some understanding of the product relation, which can be used to motivate sales activations (for instance, the promotions or products together).

Namely, new attributes can be designed, for instance, to develop “combined products” or to determine new categories of products depending on the products often bought together. All these gives an opportunity to enhance the general client experience, inventory turnover and sales.

iii.Data Aggregation

Data aggregation is the process of summarising or aggregating the data at different level to a certain common level. This makes it easier to get over all picture of sales pattern and at the same time reduces the amount of data to be worked on in the analysis to meaningful summaries. The following aggregation techniques were applied:

1. Order Level Aggregation:

At the order level the dataset is aggregated by “Order ID” to obtain the sum of the sales value across each particular order. The amount of money that one wants to spend on an order is called the Total Sales for an Order that is $\text{Total Sales for an Order} = \text{SUM}(\text{Quantity Ordered} * \text{Price Each})$. This aggregation is useful in the determination of the mean value of orders as well as in the identification of large value of orders. The analysis of the total sales figures in order helps the businesses in establishing the buying habits of their customers which include the tendency by the customer to order high or low value products among other things.

```
# Create Total Sales column
```

```
data$Total.Sales <- data$Quantity.Ordered * data$Price.Each
```

code 9: new column total.sales

Applications: This insight can help answer question like: What is the AOV or average order value? Which orders help to generate most of the total revenues? What is the average time between the order of a significant amount of product by the customer? Use of the order-level variables may help in determining ways on how to improve the AOV through approaches such as product combination or establishing sales incentives tied on higher purchase volumes.

2. Product Level Aggregation:

At the product level, the sales are summed up by “Product” to give total sales and total quantity ordered by particular product. It also enables us to determine which products are most popular and the products most instrumental in racking up high revenues. Through such analysis, companies will be in a position to identify

their popularity and income producing products so that they can enhance the profitable products or services sale.

```
# Create a new data frame with repeated Order.IDs, their products, and total sales
order_id_df <- repeated_orders %>%
  group_by(Order.ID, Street, City, State, Zip.Code) %>%
  summarise(
    Products = paste(unique(Product), collapse = ", "), # Concatenate product names into a single string
    Repeat_Count = n(), # Count how many times each repeated Order.ID appears
    Total_Sales = sum(Total.Sales, na.rm = TRUE), # Sum of Total Sales for each repeated Order.ID
    .groups = 'drop' # Avoid warning about grouping
  ) %>%
  # Reorder columns and NA values with "NA" instead of ""
```

code 10: different orderid product combining

this aggregation help in assessing the performance of a particular product in relation to other categories, which help in the right planning of inventory management. For example, knowing which of the products' demand is rather stable, one can avoid situations when some goods is overordered and when the demand declines, there are not enough stocks. It also assists in product positioning because it shows items that should can command a higher price because they are popular or sell well.

Table 5: combined products

Products
USB-C Charging Cable, Wired Headphones
Apple Airpods Headphones, AA Batteries (4-pack)
Vareebadd Phone, Wired Headphones
Google Phone, USB-C Charging Cable
Google Phone, Bose SoundSport Headphones

Applications: The questions that can be answered by the analysis include: Which products have the highest total quantity sales? What products are most profitable? What are patterns of consumption with regard to the various products? It is useful to know in stock quantities for tracking inventory and to further refine product offerings on the marketplace.

3. City Level Aggregation:

In terms of the city sales, the data is formatted under the field “City” so that it is possible to summarize total sales as well as number of orders per city. By deciding where various businesses are located, companies may locate where they are doing the best and then invest more money into marketing towards the most active geographic areas. In addition, it help increase the effectiveness of logistics and distribution because it can identify the location of the most orders shipped.

```
# Create a new column combining City and State for geocoding
city_counts <- city_counts %>%
  mutate(Location = paste(City, State, sep = ", ")) # Create a Location column

# Geocode the cities (using OpenStreetMap)
geocoded_cities <- city_counts %>%
  geocode(Location, method = 'osm', full_results = TRUE) # Get full results for more details

# Check if geocoding was successful and if 'long' and 'lat' are present
if (!all(c("long", "lat") %in% colnames(geocoded_cities))) {
  stop("Geocoding failed: Missing 'long' or 'lat' columns.")
}

# Combine geocoded data with frequency data
geocoded_cities <- geocoded_cities %>%
  select(long, lat, Location) %>% # Select necessary columns
  left_join(city_counts, by = c("Location" = "Location")) # Join to bring in Frequency

# Ensure the combined data has the necessary columns
if (!"Frequency" %in% colnames(geocoded_cities)) {
  stop("Frequency column not found after joining.")
}
```

code 11:adding new columns latitude and longitude

Apart from these benefits, coding the cities to get geographic coordinates such as in the code given in the session add another dimension whereby the sales data is geocoded to actual location. By geocoding the cities it is possible to plot each city on a map making it easier to compare each figure and notice anomalies. For instance, a city that has shown higher sales than expected will suggest new areas for growth of the business or, for example, better locations for delivery.

long	lat	Location	State	City	Frequency
1.24277	34.05369	Los Angeles, CA	CA	Los Angeles	29605
122.41933	37.77926	San Francisco, CA	CA	San Francisco	44732
-84.39026	33.74899	Atlanta, GA	GA	Atlanta	14881
-71.06051	42.35543	Boston, MA	MA	Boston	19934
-70.25866	43.65736	Portland, ME	ME	Portland	2455

code 12:longitude and latitude column

This aggregation and geocoding can answer essential business questions like: “Which cities have the greatest sales?” or “Can one differentiate between geography of consumers?” Knowledge derived from such an analysis can assist in designing the subsequent local marketing promotions, business expansion plans or the promotional sales campaigns that might be appropriate for regions.

Through georeferencing of these sales numbers, businesses are able to have a better insight in the regions and therefore make better choices when it comes to marketing and logistics.

7.Data Visualization

Communications is an important aspect of data analysis and ensuring that insight produced is well understood implies that good data visualization techniques ought to be employed. The following visualizations are recommended to accompany this report, providing a visual narrative to the analysis:

i.Total order count by state

```
# Display the summarized data |
print(order_count_by_state)
# Create a bar plot for total Order IDs by State
ggplot(order_count_by_state, aes(x = reorder(State, -Total_Orders), y = Total_Orders)) +
  geom_bar(stat = "identity", fill = "steelblue") +
  labs(title = "Total Order Counts by State",
       x = "State",
       y = "Total Order Count") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

code 13:Total order count by state

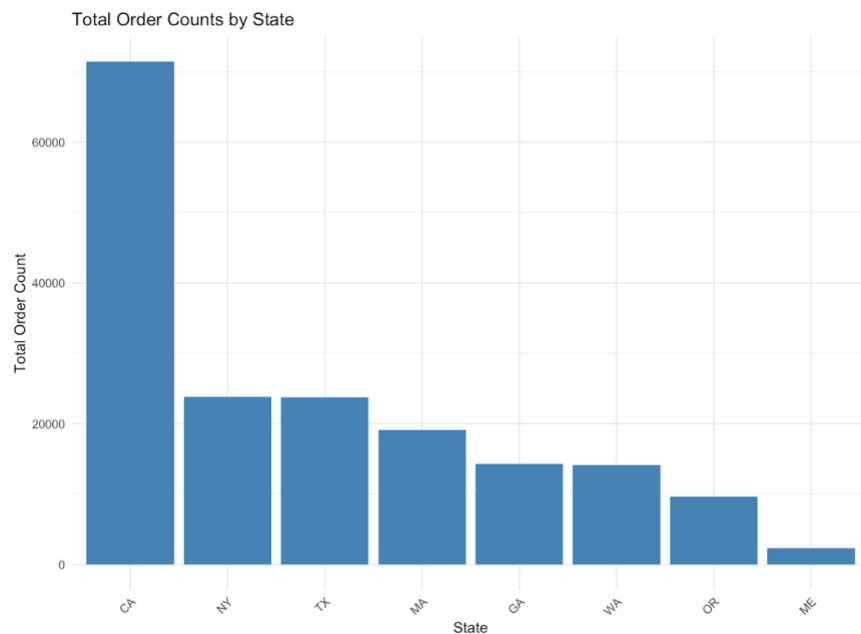


Figure 2:Total order count by state

This figure provides a bar plot of the distribution of orders received from each of the states. Individuals in the states which made high order counts shall be located on top, revealing areas with high customer interaction. For example, California or Texas states are likely to have a higher employment of orders than state like Vermont or Wyoming because of higher number of population or incomes per capita respectively. The visualization will enable the Company to analyse potential areas for growth or expansion. In particular, it's possible that low-frequency states are relevant to infrequent orders or have limited brand awareness and should focus on specific advertising campaigns. It will help the company manage its resources better: for example, allocate more inventory or advertisement campaign to states that contributed to the order volume, or stimulate that state to order more if the number of orders is low.

ii.Total sales by city

```
ggplot(city_sales, aes(x = reorder(City, -Total_Sales), y = Total_Sales)) +  
  geom_bar(stat = "identity", fill = "steelblue") +  
  labs(title = "Total Sales by City", x = "City", y = "Total Sales") +  
  theme_minimal() +  
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

code 14:Total sales by city

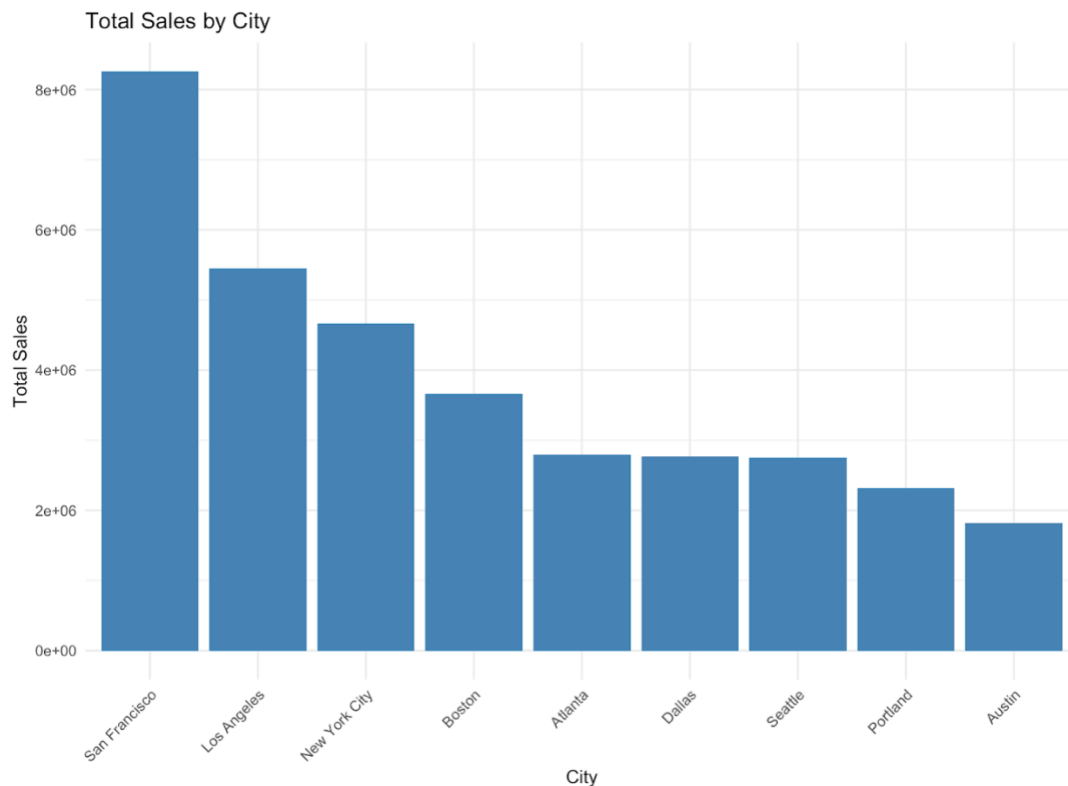


Figure 3:Total sales by city

In this figure, city has been ranked or shaded depending on the total sales contribution. Perhaps the largest cities such as New York or Los Angeles or Chicago could have the highest sales because those are big cities with high activities within the marketplace. On the other hand, the quantity collected by small towns or cities might be comparatively smaller. The visualization might show that a handful of cities accounts for most sales and therefore these areas should target aggressive and costly marketing strategies or product releases. These unsatisfactory performers may be either developing markets that require longer time to mature or customer markets that the company's products do not suit. Also, the company may determine whether these cities have constraints that may be dampening the sales for instance slow delivery or fewer depots in the given cities.

iii.Total sales by months

```
# Plot total sales by month
ggplot(monthly_summary_sorted, aes(x = Month, y = Total_Sales)) +
  geom_bar(stat = "identity", fill = "steelblue") +
  labs(title = "Total Sales by Month", x = "Month", y = "Total Sales") +
  theme_minimal()
```

code 15: Total sales by months

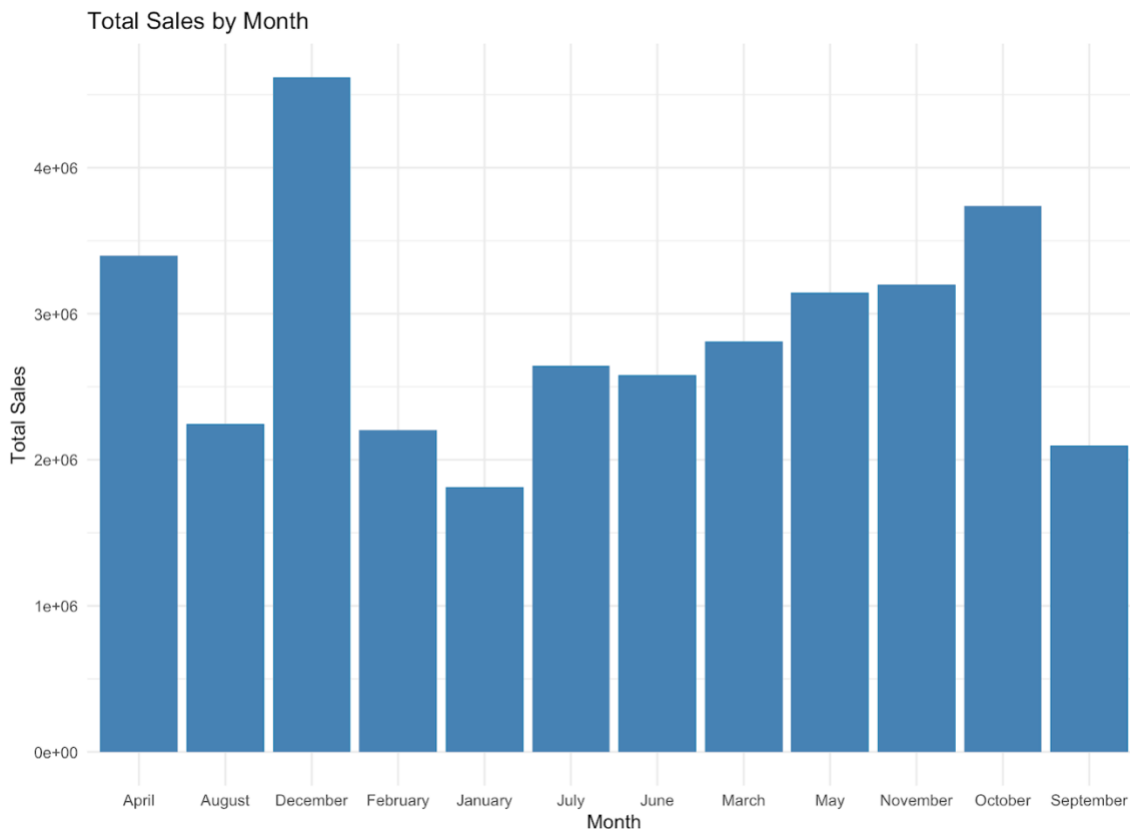


Figure 4: Total sales by months

This bar plot divides total sale so correctly in each month which will show the seasonal variation. For instance, if a particular product is targeted for sale during the festive season, November and December will record a high sale while June or July may record low sale since it is summer. During daily sales reporting, important retail days such as Black Friday or Cyber Monday or summer sale events would be noticeable to the company to determine which months have the highest sales. On the other hand, low months may be taken to offer promotions or new products that would balance revenue flow All in all, the promise made could be applied to months with low sales to help level out the flow. Seasonal fluctuations will also factor into inventory ordering, which will be performed when sales are high and during low sales season.

iv.Total sales by product

```
# Plot all products
ggplot(all_products, aes(x = reorder(Product, -Total_Sales), y = Total_Sales)) +
  geom_bar(stat = "identity", fill = "steelblue") +
  labs(title = "Total Sales by Product", x = "Product", y = "Total Sales") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

code 16:Total sales by product

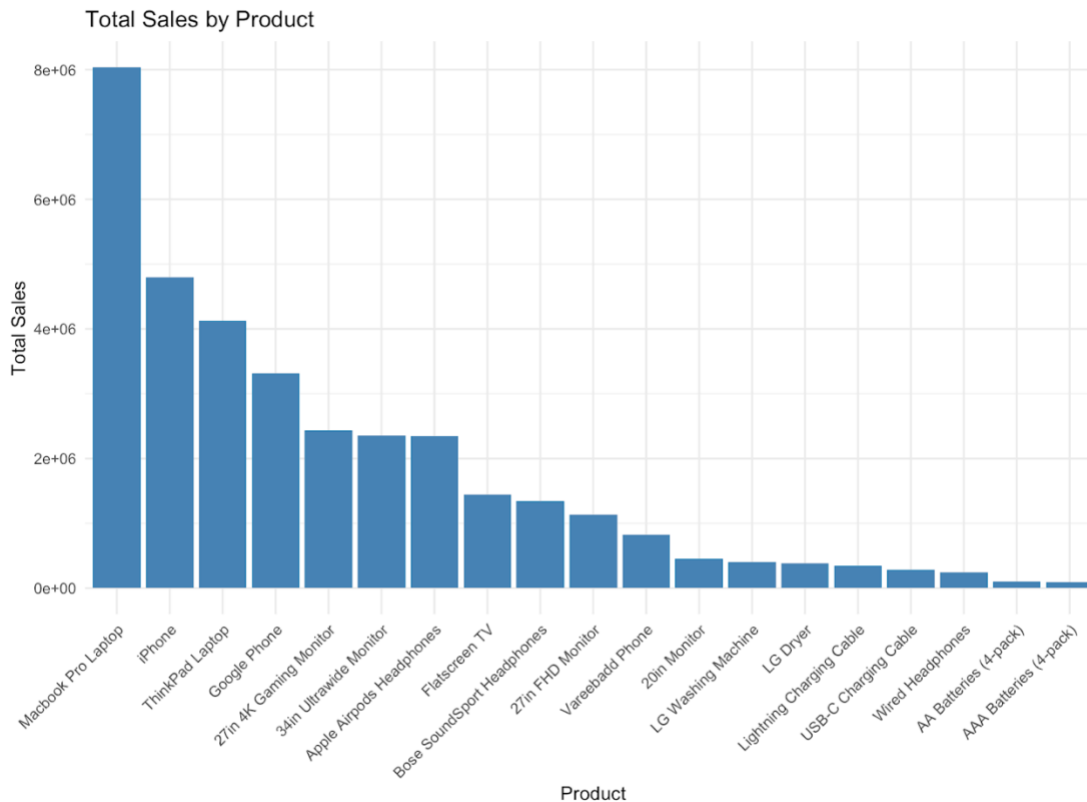


Figure 5:Total sales by product

This figure ranks the products based on Total Sales. At the end of this ranking, high performing products will be on the upper end as compared to those products which may be recording slow sales. The company can very quickly identify which of the products makes most of the money and hence you can look at them for future marketing campaigns or manufacturing. For instance, if the specific region, for example, electronics, is on the priority list of sellable products, then adequate products' range or products of related categories might be introduced. Poor performing products could be an area we need to stop, redesign or simply promote better. This makes the work of identifying what to produce, or what stock to order, easier going by what the customers prefer in the market.

v.Total sales by state

```
# Create a bar plot to visualize total sales for each state
ggplot(state_sales_summary, aes(x = State, y = Total_Sales)) +
  geom_bar(stat = "identity", fill = "steelblue") +
  theme_minimal() +
  labs(
    title = "Total Sales by State",
    x = "State",
    y = "Total Sales"
  ) +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

code 17:Total sales by state

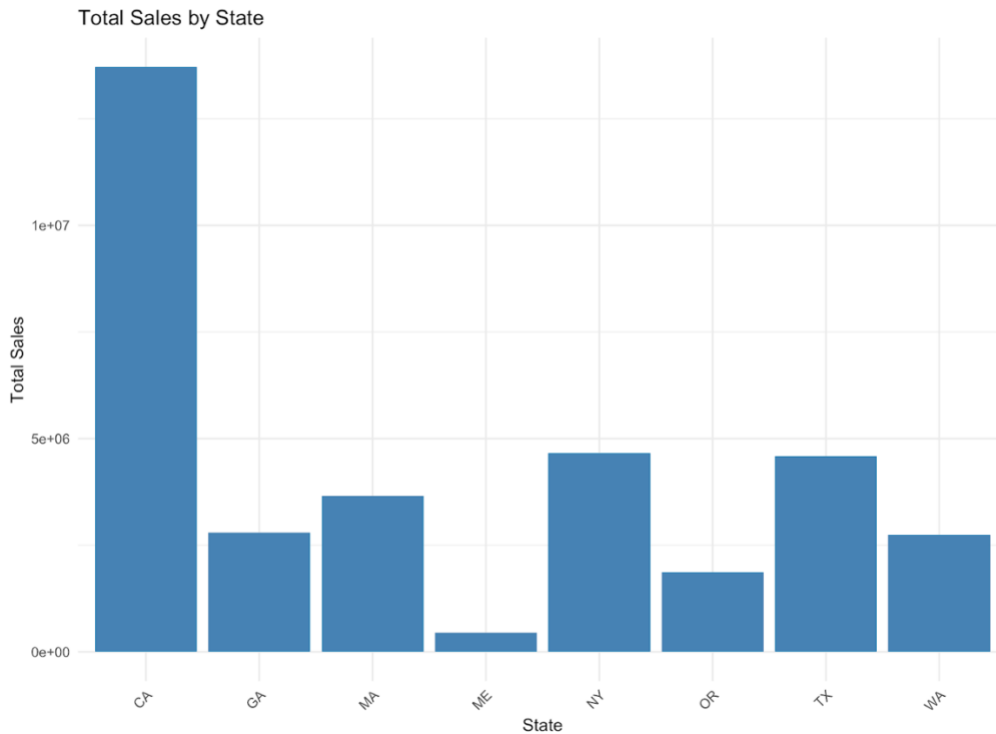


Figure 6:Total sales by state

This is similar to the first figure, though instead of orders, it looks at Total sales . Again, states with higher average order values or effecting higher valued products will be clearly visible here, which will paint a different picture than seen from the order aggregation. For instance, Nevada may represent a lower number of orders but a higher total sales based on more luxurious, larger orders; on the other hand, Ohio may represent more orders, but less total sales, owing to orders of smaller quantities. This makes it easier for the company to position it right, whether making a case for price skimming in states that have high disposable incomes per capita or making a case for sheer quantity in those states that have lower average order values. They offered a more detailed understanding of how geographical operations are progressing and consequently better decisions can be made.

vi.Total sales by timeframe and month

```
# Create the grouped bar plot
ggplot(sales_summary, aes(x = Month, y = Total_Sales, fill = time_frame)) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(title = "Total Sales by Time Frame and Month",
       x = "Month",
       y = "Total Sales") +
  scale_fill_manual(values = c("Morning" = "pink", "Evening" = "red", "Night" = "yellow")) +
  theme_minimal()
```

code 18:Total sales by timeframe and month

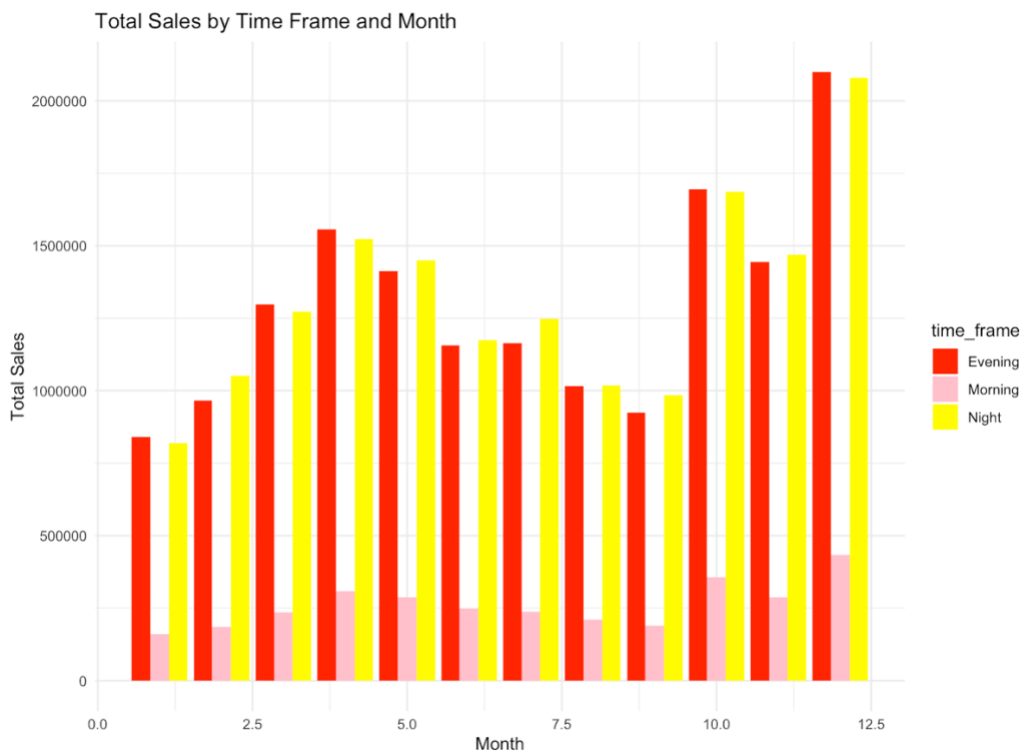


Figure 7:Total sales by timeframe and month

The following figure showing the sales by time within each month enables the company to monitor how the sales are generated in every given day. Activity that occurs at 1 to 8 may show that customer buying extravaganza is more at morning, possibly buying breakfast or on their way to work. Hourly sales of 8-16 may represent lunch time or mid-day business while sales 16-24 may depict late evening business or night business among the customers. When such patterns are analyzed between months then the company can predict when these should be expected when it covers different time spans which can be used in matters involving staffing and other resources. For instance, if a business learns that its sales are more gravitated to the evening during summer then the morning during other parts of the year, they will be able to adjust promotions or time-sensitive campaigns accordingly. Recognizing these changes enables the company to schedule advertisements and other business actions that will have the greatest impact on sales during each time frame.

vii.Total Sales Distribution by Time Frame

```
# Create pie chart with percentage labels
ggplot(time_frame_summary, aes(x = "", y = Total_Sales, fill = time_frame)) +
  geom_bar(stat = "identity", width = 1) +
  coord_polar("y", start = 0) + # Convert to pie chart
  labs(title = "Total Sales Distribution by Time Frame") +
  scale_fill_manual(values = c("Morning" = "red", "Evening" = "orange", "Night" = "green")) +
  theme_void() + # Remove axis labels and gridlines for pie chart aesthetics
  geom_text(aes(label = paste0(round(Percentage, 1), "%")), # Add percentage labels
            position = position_stack(vjust = 0.5))
```

code 19:Total Sales Distribution by Time Frame

Total Sales Distribution by Time Frame

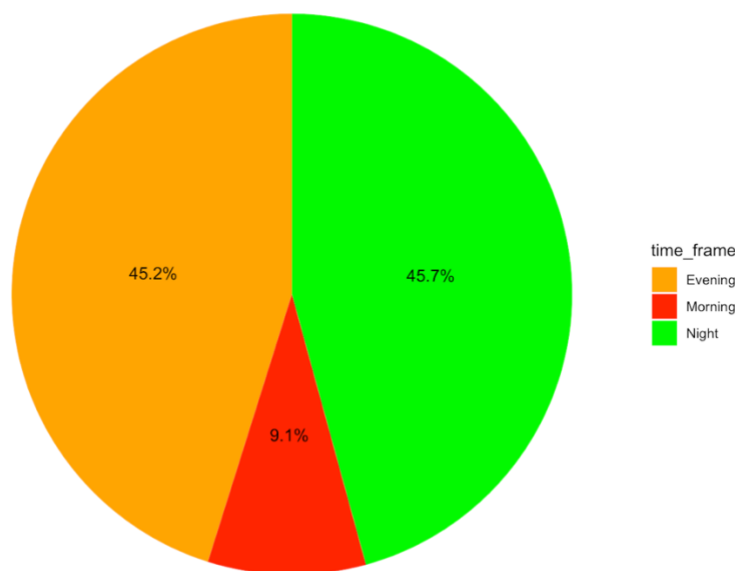


Figure 8:Total Sales Distribution by Time Frame

This pie chart shows on what times within the morning, evening and night durations sales are made. An increased focus of sales between the period of 16 to 24 may be interpreted to mean that many consumers purchase goods and services at that time of the day for pleasure or after a typical working day. Sales in the morning timeframe (1 to 8) may be low under the condition that food offerings do not complement the morning schedule. These if power packed evenly, help the company realize sales throughout the day as opposed to a single period. However, if food truck sales are very specific in some parts of a day, it may be likely that people may want to profile the promotion or campaign in the least busy time. For example, placing on sale morning or evening clothes with a certain percentage off during those hours can help level the amount of sales throughout the day.

viii.Sales by time Frame and state

```
ggplot(sales_by_time_state, aes(x = State, y = Total_Sales, fill = time_frame)) +  
  geom_bar(stat = "identity", position = "dodge") +  
  labs(title = "Sales by Time Frame and State", x = "State", y = "Total Sales") +  
  theme_minimal() +  
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +  
  facet_wrap(~ time_frame)
```

code 20:Sales by time Frame and state

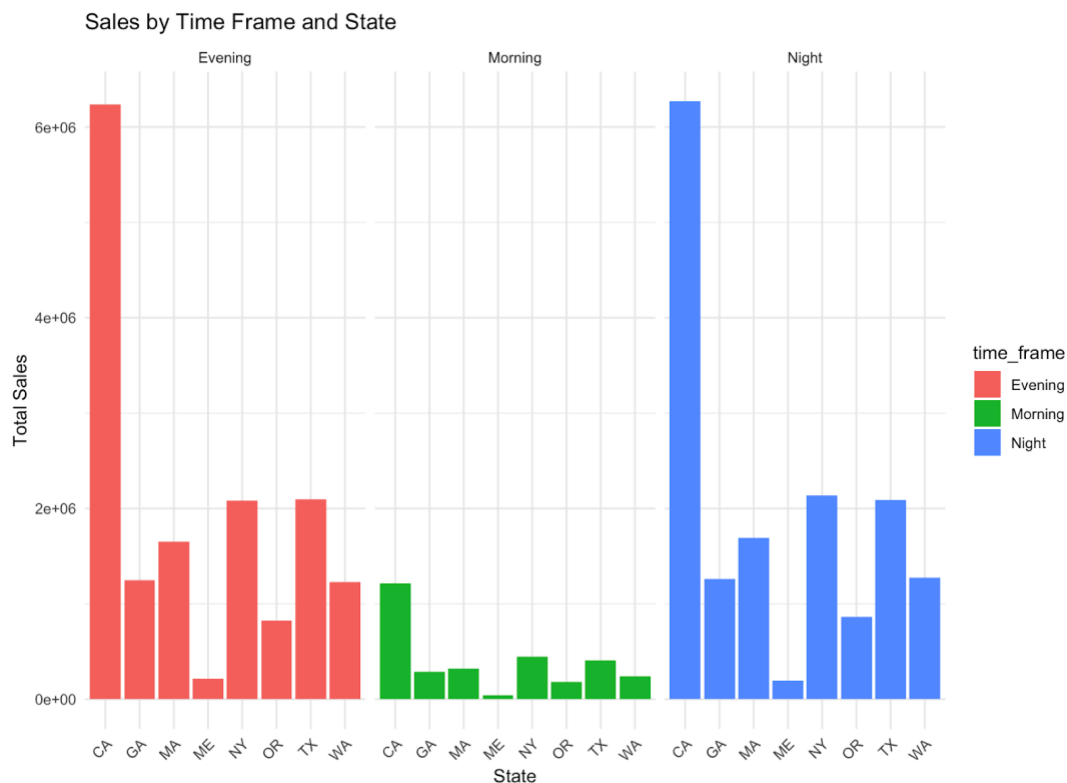


Figure 9:Sales by time Frame and state

This figure overlays the sales by time distribution – morning, evening, and night – on the geographic and state sales distribution. For instance, sales may be high in California at night timeframe 16-24 because most customers' activities are at this time or, at morning timeframe 1-8, high in Florida because the Customers may be early birds. This information can be used to create region specific brand management campaigns by the company. For instance, if the night shift sales are high in Texas, the company can channel all late evening promotions and staff to Texas while if New York for instance records low morning sales, the company can work around that and create morning promotions. Now these state-specific patterns help in greater management of stocks, human resource and marketing strategies by keeping in mind the customer purchasing behavior at micro and macro level.

ix.Customer Segmentation by total and average order

```
ggplot(customer_segmentation, aes(x = Total_Sales, y = Average_Order_Size)) +  
  geom_point(alpha = 0.6) +  
  labs(title = "Customer Segmentation by Total Sales and Average Order Size", x = "Total Sales", y = "Average Order Size") +  
  theme_minimal()
```

size

code 21:Customer Segmentation by total and average order size

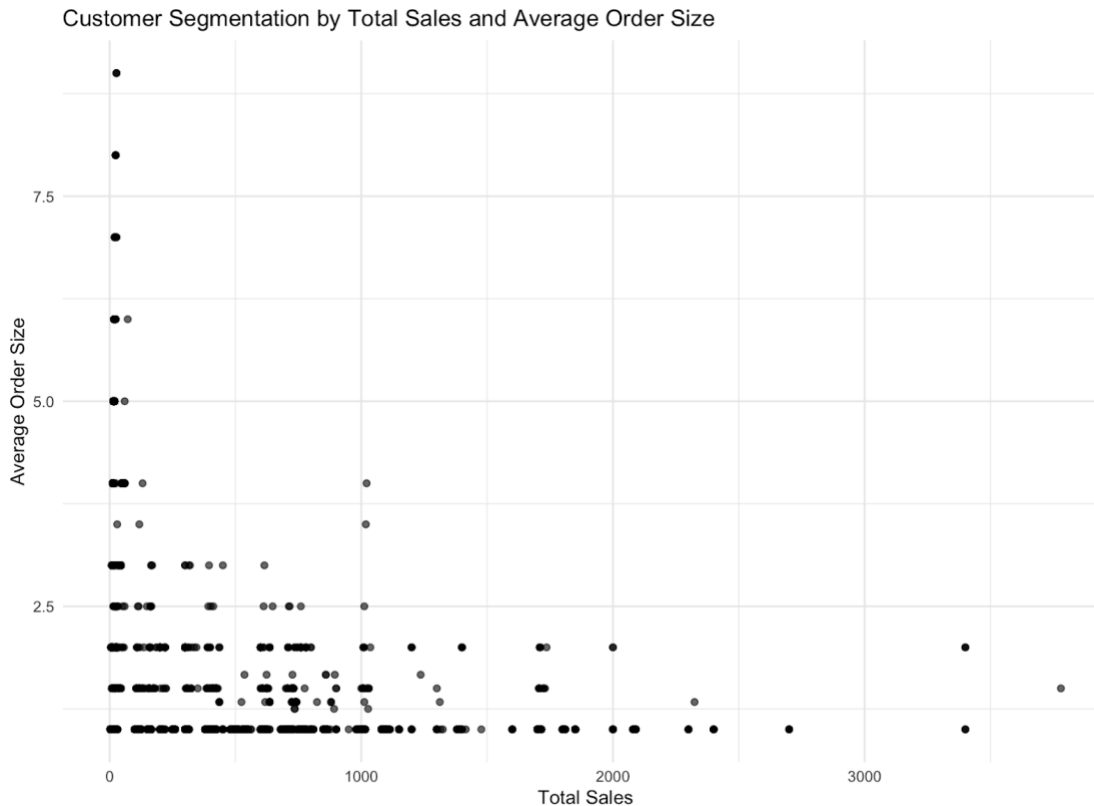


Figure 10:Customer Segmentation by total and average order size

This figure further categorizes customers on their purchasing power and their frequency of orders. Organizations with large amounts of customer orders will be placed in one group, while customers who rarely use the services or buy small quantities will be in another group. Identifying these segments helps the company focus its efforts: those who fall under the above-mentioned category should be able to be rewarded by loyalty programs or VIP discounts while on the other hand, those who fall under the below the line, should be able to be advertised to in a bid to increase their order size and/or frequency. It, therefore, benefits the company to know which people constitute its best customer group so that it focuses its efforts, thus enhancing the revenue churned out by its premium group of customers.

x.Product co-occurrence matrix

```
# Create the heatmap using ggplot2 with counts displayed
ggplot(co_occurrence, aes(Product_Pairs_1, Product_Pairs_2)) +
  geom_tile(aes(fill = Frequency), color = "green") +
  geom_text(aes(label = Frequency), color = "blue", size = 4) + # Add counts as labels
  scale_fill_gradient(low = "white", high = "red") +
  theme_minimal() +
  labs(title = "Product Co-Occurrence Matrix",
       x = "Products",
       y = "Products",
       fill = "Frequency") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

code 22:Product co-occurrence matrix

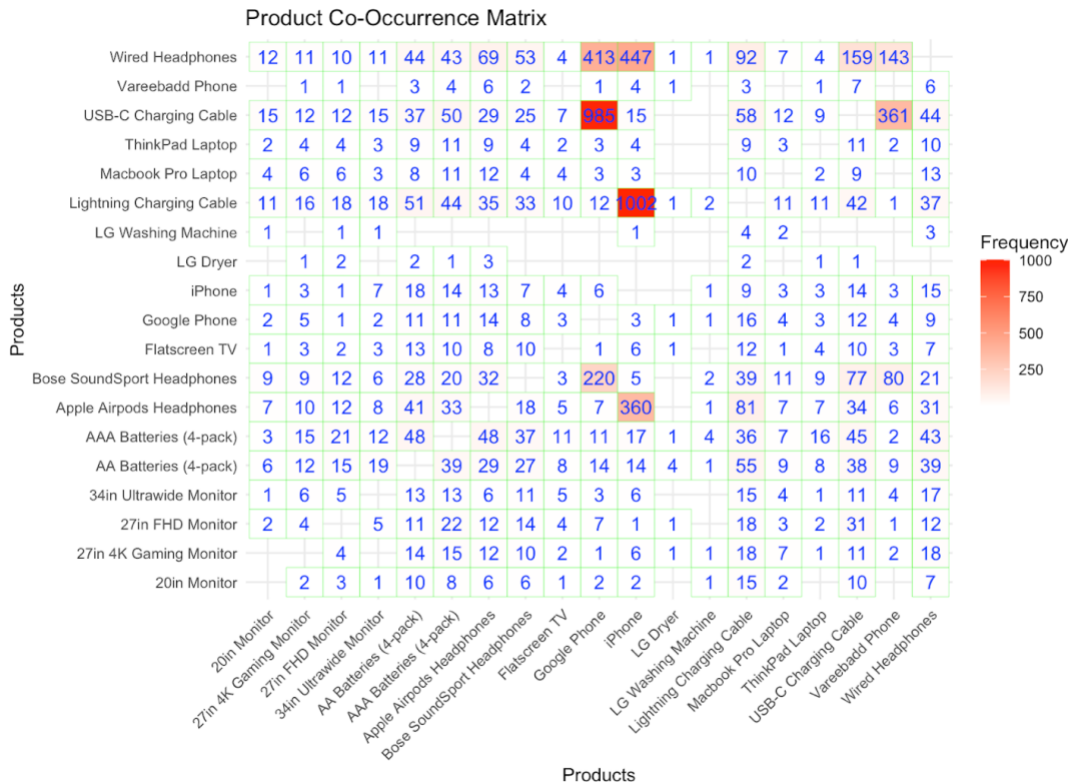


Figure 11:Product co-occurrence matrix

This matrix indicates which products are complementary goods. For instance, if customer frequently buy smartphone along with its protective case then this co-occurrence relationship will be high. Then the company can take the information from this model to try to create bundled products or places a cross sell on the items customers purchase to get them to buy other related products. Understanding these patterns makes it easier to promote and recommend the products in question, increasing the chances of customer making multiple purchases at once. Further, it might uncover some unsuspected relationships which can be valuable to the company as it can point out new possibilities for combination to improve the sales model.

xi.Geospatial analysis of order.id location as per their city

```
# Plot the cities on a leaflet map
leaflet(geocoded_cities) %>%
  addTiles() %>% # Add base map tiles
  addCircles(
    lng = ~long, lat = ~lat,
    popup = ~paste(City, State, "<br>Customers: ", Frequency), # Show sum of order.id in popup
    radius = 10000, # Adjust circle size as needed
    weight = 1,
    color = ~ifelse(Frequency == 1, "green", ifelse(Frequency == 2, "orange", "red")), # Direct color assignment based on frequency
    fillOpacity = 0.5
  ) %>%
  addMarkers(
    lng = ~long, lat = ~lat,
    label = ~as.character(Frequency), # Show customer count in label
    labelOptions = labelOptions(noHide = TRUE, textOnly = FALSE, direction = "auto")
  )
```

code 23:Geospatial analysis of order.id location as per their city

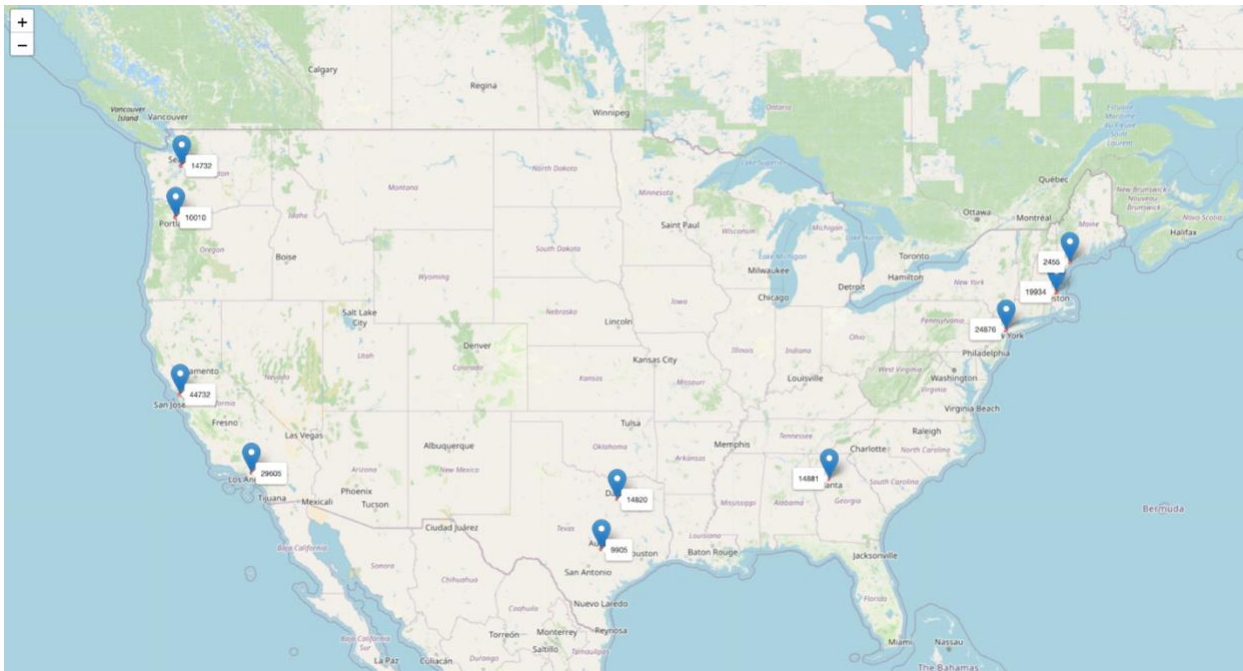


Figure 12:Geospatial analysis of order.id location as per their city

This figure shows where orders are being placed at a high level of detail down to specific cities. For example, locations with high population density around the New York, Los Angeles, or Chicago areas should be expected compared with threshold categories in suburban and rural settings, which may represent a new market opportunity. It also assists the company to place distribution locations proximities important centres, thereby minimising their distribution expenses. It also might indicate the oversaturated regions that produce fewer orders than you'd assume, giving you insight into potential new markets to target or the potential need for raises brand recognition

xii.Geospatial analysis of customers buying more than one items with total sales

```
# Create the leaflet map
leaflet(city_sales_map) %>%
  addTiles() %>%
  addMarkers(
    lng = ~long, lat = ~lat,
    icon = makeIcon(
      iconUrl = "https://img.icons8.com/ios-filled/50/000000/star.png", # Star icon
      iconWidth = 30, iconHeight = 30 # Adjust the size of the icon
    ),
    popup = ~paste("City:", City,
      "<br>Total Sales: $", format(Total_Sales, big.mark = ","),
      "<br>Number of Orders:", Order_Count)
  )
```

code 24:Geospatial analysis of customers buying more than one items with total sales

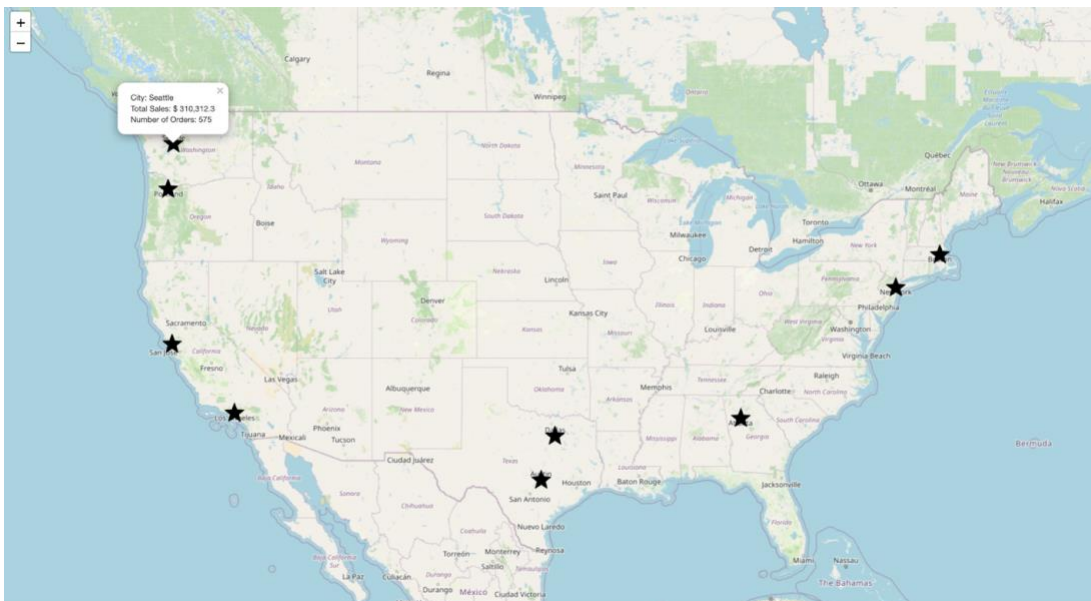


Figure 13:Geospatial analysis of customers buying more than one items with total sales

Based on this, this map concentrates on the customer who buys the products more than once or the one who buys many products. Presence of high density of such customers indicates either good brand association or high levels of customer satisfaction. For instance, certain locations may have high density of repeat consumer; such locations will need items that are very cheap to consumers but very costly to produce, or personalized bonuses to be offered every time the repeat buyers are consuming product. However, places with low repeat consumption might need acquired research, for instance, if they have problems with customer relations, product quality, or delivery time. This visualization enables the company to identify areas where minor enhancements would most benefit the effort and resources on customer retention.

The fact is that each of the visualizations will present the data from different angle, allowing the business see it from different perspective and, thus, get a more comprehensive view of its performance.

8. Conclusion and suggestions

Based on the analysis, several actionable recommendations emerge that can help the business improve its sales strategy and overall performance:

- Focus on Holiday Marketing

Especially, it is demonstrated that two months of November and December make a significantly high percentage of total sales. These months should be strategically targeted because marketing should be directed to these months in order to increase its revenues from these sales promotions tools like discounts and bundle offers. Also, it will optimize inventory by making sure there is enough stock in the shelves to meet the higher demand and reduce occasions which result in stock out or any potential sales.

- Optimize Inventory Management

While products like iPhones and laptops produced good sales, is it much as expensive items tend to do though at a slower rate compared to cheaper items. Such products should be made priorities when it comes to managing inventories most especially when sales are high. On the other hand, accessories and cheaper products may be sold in larger quantities but they may take up far less space since their contribution to total sales is meagre.

- Regional Expansion

There is much demand in big cities like New York, LA, and Chicago but much more potential in the smaller cities especially in the midwestern and southern states. Penetration and market promotions in these areas might be increased to achieve higher growth rates with reference to the local tastes.

- Improve Customer Experience

When considering the customers' purchase behavior by time of day, the research established that the greatest buying sprees happen in the late morning and initial part of the afternoon. By integrating these hours into your business model, you might plan to launch limited time sales or special offers during these hours of the day or perhaps staff more workers into customer relations or support during these specified time periods.

- Leverage Product-Level Insights for Marketing

Some types of products always have relatively higher units sold and total dollars of sales. The business should ensure that it markets these products more aggressively since they are the most popular, the business could sell these products bundled with other related products at a client's special discount.

- Future Research and Analysis

Last but not the least, this report has centered mainly on sales and customer behaviour in the course of the year 2019. Further study might, for instance, seek to extend into the sales data from other years and may then provide information about the long term sales trends as well as influence of economic variables in sales outcomes. Moreover, the future research could be more precise in customer segmentation analysis that would contribute to the understanding of customers' preferences and buying behavior to enhance marketing plans and stock management.

- A deeper understanding of the segmentation of customers

Perhaps, this section could discuss how the analysis of customer data by the company helps in promoting better segmentation to the firm's promotional campaigns. Based on the customers' frequency of purchase, geographic location and choice of certain products, marketing campaigns that will leading to higher conversion rates could be targeted to the right customers.

9. References

1. Bhatia, M., & Patel, N. (2021). The role of data preprocessing in effective machine learning applications. *Journal of Computer Science and Technology*, 36(3), 557-576.
<https://doi.org/10.1007/s11390-021-0131-4>
2. Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 29(5), 1189-1232. <https://projecteuclid.org/euclid.aos/1013203451>
3. Deng, L., & Yu, D. (2014). Deep learning: Methods and applications. *Foundations and Trends® in Signal Processing*, 7(3–4), 197-387. <https://doi.org/10.1561/20000000039>
4. García, S., Luengo, J., & Herrera, F. (2015). *Data preprocessing in data mining*. Springer.
<https://doi.org/10.1007/978-3-319-16273-6>
5. Agarwal, R., & Shankar, R. (2017). Data aggregation techniques in data warehousing: A comparative study. *Journal of Database Management*, 28(3), 21-40.
<https://doi.org/10.4018/JDM.2017070102>