HANOI SCHOOL OF BUSINESS AND MANAGEMENT
VietNam National University

----o0o----

REPORT

REVENUE MANAGEMENT AND PRICING

Subject: Introduction to Data Science

Lecture: Dr. Emmanuel Plan
PhD. Nguyen Huy Anh

Class: MET4

Group: 08

HaNoi - 2024

ACKNOWLEDGEMENTS

Group 8 would like to express our sincere thanks to Dr. Emmanuel Plan and Master Nguyen Huy Anh (DMS researcher), who have supported us a lot in our Introduction to Data Science assignment. We would like to express our sincere thanks to the lecturers, whose passion, dedication and extensive knowledge have helped us complete this report as thoroughly as possible.

The dedication and passion of the lecturers in teaching and scientific research have aroused in us a love and interest in this field. The comments from the lecturers have helped us gain a deeper insight into the problem of our group's assignment, and also helped us improve our teamwork skills as well as our ability to analyze and synthesize information.

The group's success is thanks to the support and guidance of the teachers, we really appreciate that. I hope the teachers always stay healthy and be more successful in their research and teaching work!

GROUP 8

TABLE OF CONTENTS

1. INTRODUCTION

1.1. Background

In the context of the growing digital economy, e-commerce has become a key sales channel, playing an important role in the global economy. With the rise of platforms such as Amazon, Alibaba, Shopee, and Lazada, the e-commerce industry not only creates opportunities to reach a wide range of customers but also comes with the challenge of optimizing revenue and pricing products strategically.

Revenue Management and Pricing strategies have proven to play an important role in helping e-commerce businesses achieve maximum profits. Pricing decisions in the e-commerce environment are affected by many factors, including:

- **Demand fluctuations:** Customers on e-commerce platforms tend to change their shopping behavior over time, according to promotional events, or socio-economic factors.
- **Fierce Competition:** Online retailers face pressure from competitors to offer fair prices to retain customers.
- **Data and Technology:** With a huge amount of data from customers, businesses can apply artificial intelligence and machine learning (AI/ML) to predict demand and optimize prices.

An effective revenue management and pricing strategy in e-commerce not only helps businesses increase revenue but also builds trust and maintains relationships with customers. However, implementing these strategies faces challenges such as: managing complexity in the supply chain, price transparency, and legal issues related to consumer protection.

1.2. Objectives of the Report

The objective of this report is to provide an in-depth look at how to manage revenue and build an effective pricing strategy in the e-commerce industry.

- **Optimizing pricing strategies:** How to adjust prices to each customer segment while maximizing profits
- **Revenue forecasting:** Using historical data, how to predict future revenue trends
- **Customer behavior analysis:** Gain a better understanding of shopping habits, product preferences, and factors that drive purchase decisions.
- **Optimizing inventory management:** Predicting product demand helps reduce inventory costs and avoid stockouts.

1.3. Potential stakeholders

- CEO of the company: The leadership needs to understand the results to build the company's strategy and business plan to make appropriate moves.
- Sales or account executives: Improve inventory management and distribution of goods based on customer demand analysis
- Marketing team: Use information about consumption trends and customer models to deploy appropriate campaigns.
- Product development department: Focus on investing and developing products in the potential product category.

1.4. Question

- What factors affect total revenue?
- Which time of the year or season tends to have the most sales? Any particular product that stands out?
- How does revenue change based on Quantity Ordered and Price each?
- How can we determine if a customer is high value(potential)?

2. DATA

2.1. Data Source

The "E-commerce Sales and Order Details Dataset" provides a comprehensive record of sales transactions within an online retail platform during the month of December 2019. It offers insights into consumer preferences, purchasing patterns, and sales trends across different cities and times of day. [1]

2.2. Features

Order ID: Unique identifier for each order.
Product Category: The category to which the product belongs (e.g., Laptops and Computers, Home Appliances, Charging Cables,
Product: Name of the product.
Quantity Ordered: Number of units ordered for the product
Price Each: Unit price of the product.
Order Date: Date and time of the order placement.
Purchase Address: Address where the purchase was made
Month: Numeric representation of the month (e.g., 12 for December)
Sales: Numeric representation of the month (e.g., 12 for December)

3. METHODS

3.1. Data Preprocessing

This dataset was already clean, requiring minimal preprocessing before analysis

3.2. Data Visualization

Data is visualized using charts to better understand trends, relationships, and key features.

3.3. Machine learning

- K-Means Clustering

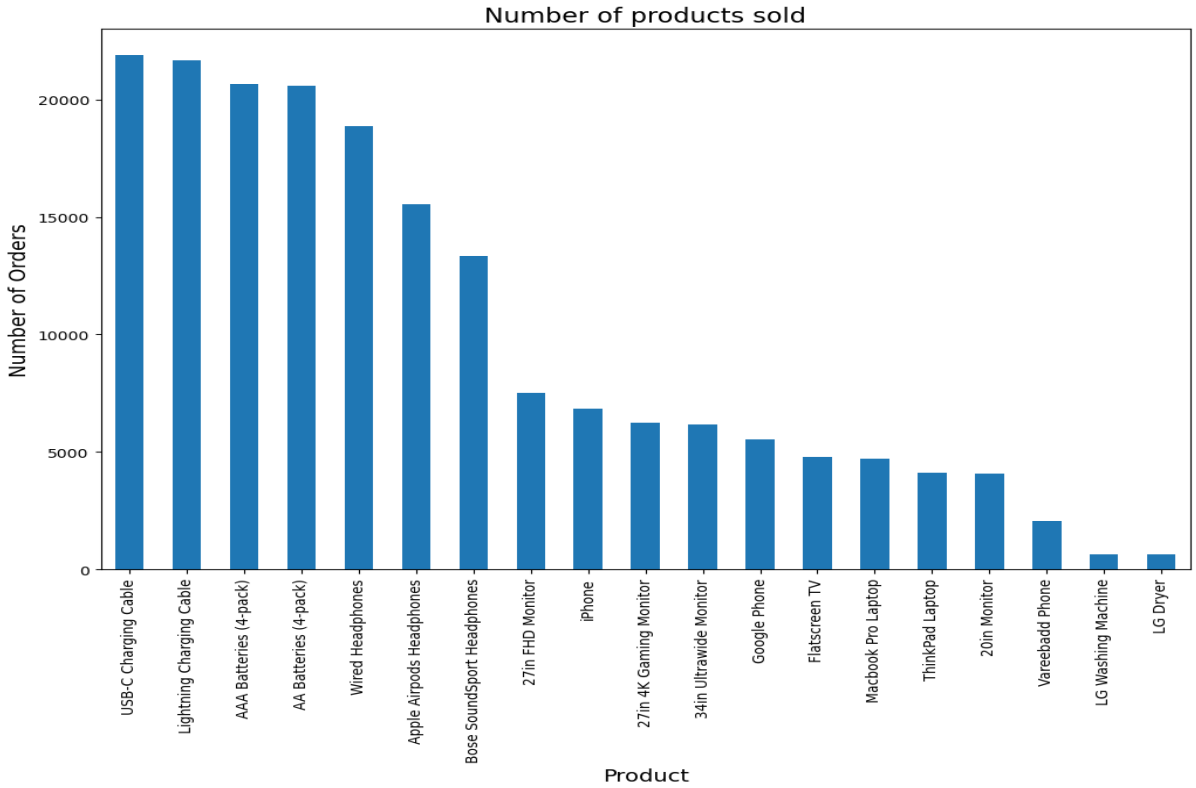
K-Means is applied to cluster customers based on the RFM (Recency, Frequency, Monetary) model. The optimal number of clusters is determined through the Elbow Method, then the data is grouped into clusters of customers with similar characteristics.[2]

- Linear Regression

Linear Regression is a technique that predicts the value of unknown variables by using another related and know data value

4. DESCRIPTIVE ANALYSIS

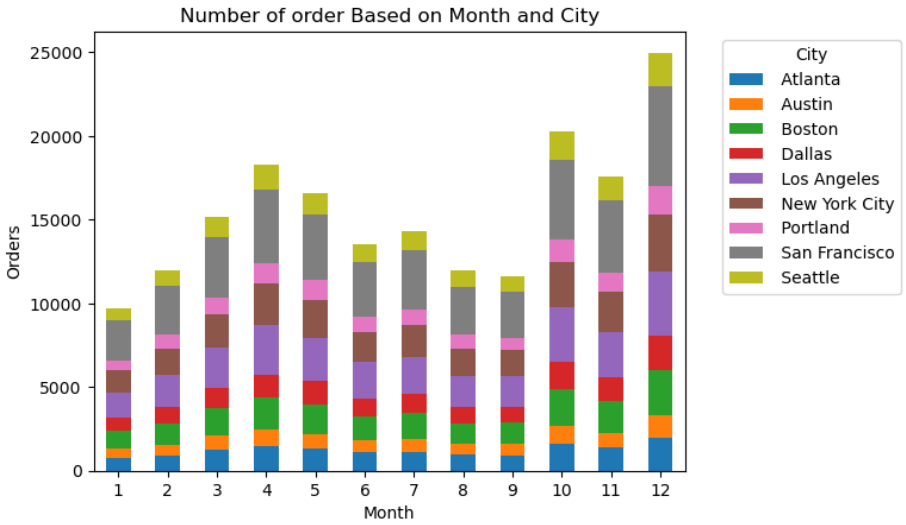
4.1. Number of Products sold



The chart shows that USB-C and Lightning Charging Cables are the top-selling products, each exceeding 20,000 orders, followed by AAA and AA Batteries, Wired Headphones, and Apple Airpods, all being compact, high-demand items. In contrast, products like the Vareebadd Phone, LG Washing Machine, and LG Dryer have the fewest orders, below 2,000, likely due to high prices or low market demand for large household appliances. The significant disparity between top-selling and low-selling products may be related to high pricing or low market demand for large household appliances.

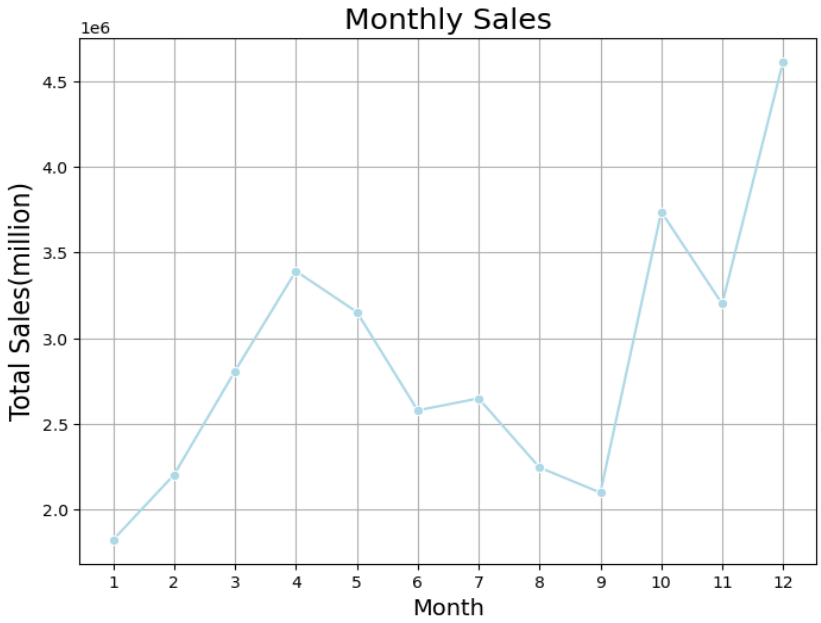
From this chart, it is clear that electronic accessories and small consumer products have strong revenue potential and should remain a focus for the business. For large household products like washing machines and dryers, the company needs to develop suitable marketing strategies and promotions to boost sales

4.2. Number of orders Based on Month and City



The stacked bar chart displays the number of orders distributed by month and city. The trend shows a steady increase in orders throughout the year, peaking in December, which is the busiest shopping period. Seattle has the highest number of orders year-round, followed by New York and Los Angeles, highlighting these as key markets. In contrast, cities like Austin and Dallas have lower order volumes. These results suggest that the business should allocate resources to the year-end shopping season and focus on major cities to maximize revenue.

4.3. Sales by Month



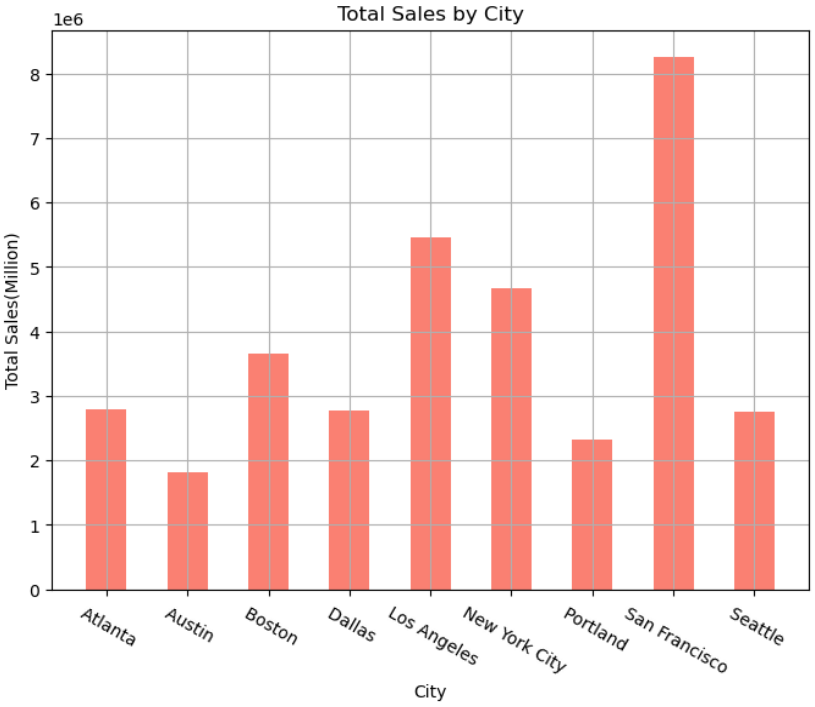
In general, we can see that revenue tends to increase gradually at the beginning of the year, from January to April, with April having the highest figure of approximately 3,390 million USD. From May to September, revenue tends to decrease gradually with September having the lowest figure (2,097 million USD). At the end of the year from October, revenue begins to increase again, peaking in December at approximately 4,613 million USD.

The beginning of the year has an increase due to early year promotions, holidays or shopping and spending needs of each family. Revenue tends to decrease gradually from May to September because this is probably the summer time, people tend to spend more on travel and vacation. The end of the year saw a return to growth due to October and November being the back-to-school season and major sales events such as Black Friday. More significantly,

December sales reached their highest figures of the year due to factors such as the holiday season and major year-end sales.

4.4. Sales by City

The chart shows a significant disparity in revenue between cities, with San Francisco and some other large cities such as Los Angeles and New York having much higher revenue than smaller cities. This may reflect differences in market size, economic development, and local business presence.

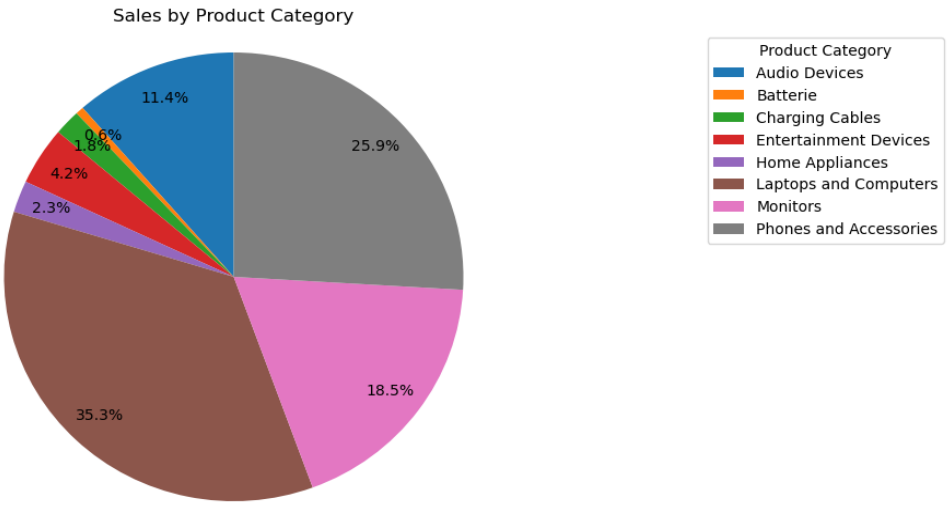


San Francisco leads the way with revenue of approximately $8.6 million. This is the highest-earning city among the cities listed. This is followed by Los Angeles with approximately $5.8 million, New York with $5 million, Boston with $3.6 million, and Atlanta with $2.9 million. Lower income cities include Dallas ($2.8 million), Seattle ($2.7 million), Portland ($2.1 million), and Austin ($1.9 million).

San Francisco is one of the world's largest tech hubs, home to many of the world's leading tech companies. This creates a high demand for tech products like laptops, phones, and electronics.

San Francisco is one of the wealthiest and most populous cities, and most importantly, has the highest per capita income in the US. This means that people here have more purchasing power than

4.5. Sales by Product

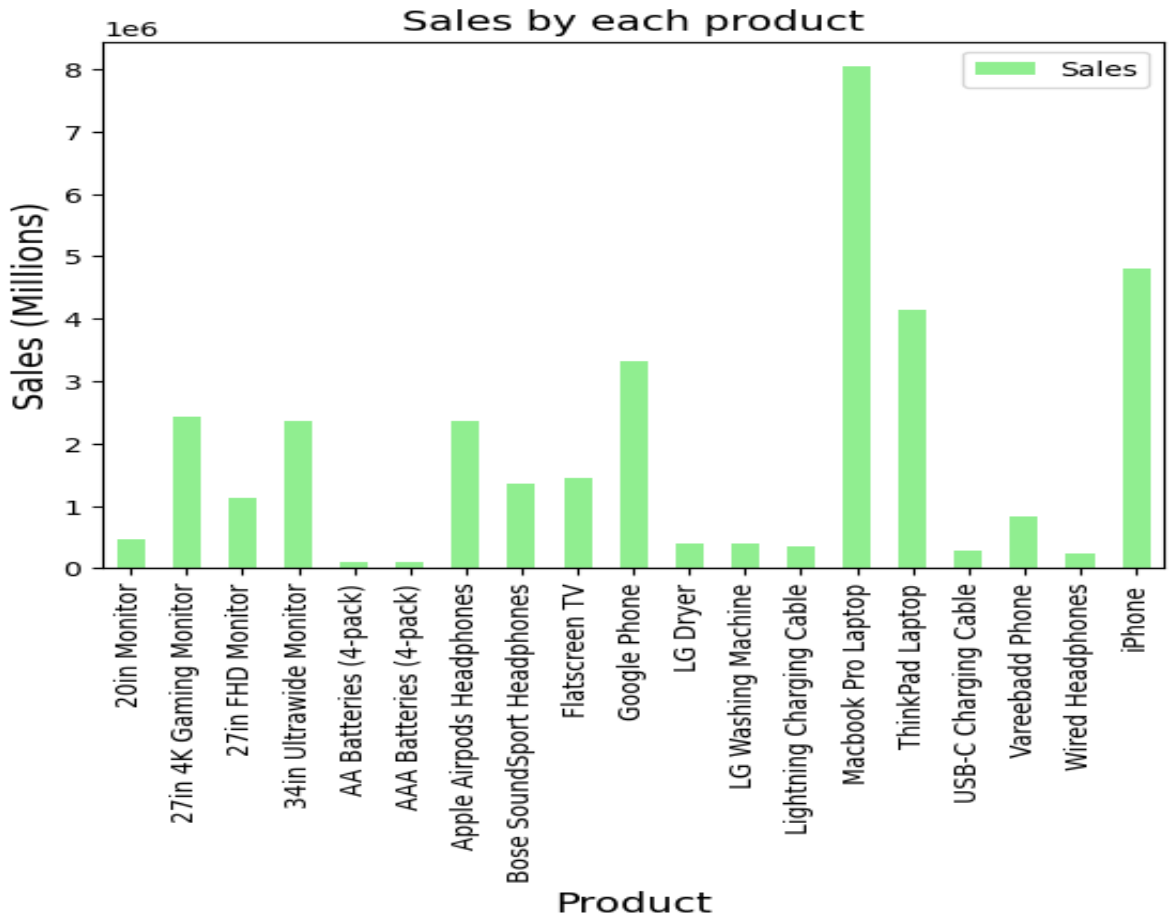


The pie chart highlights the revenue distribution across eight product categories, revealing that Laptops and Computers dominate, contributing 75% of total revenue

and solidifying their role as the company's core product group. Phones and Accessories rank second with 25.9%, underscoring their significant contribution. Monitors account for 18.5%, presenting a promising area for growth. Meanwhile, Audio and Entertainment devices contribute 11.4% and 9.2% respectively, suggesting opportunities for further investment and promotion. Household Appliances and Printing Supplies make up only 0.5% and 0.2% of revenue, indicating untapped potential for future development. Overall, while the company's revenue structure is diverse, it heavily relies on Laptops and Computers.

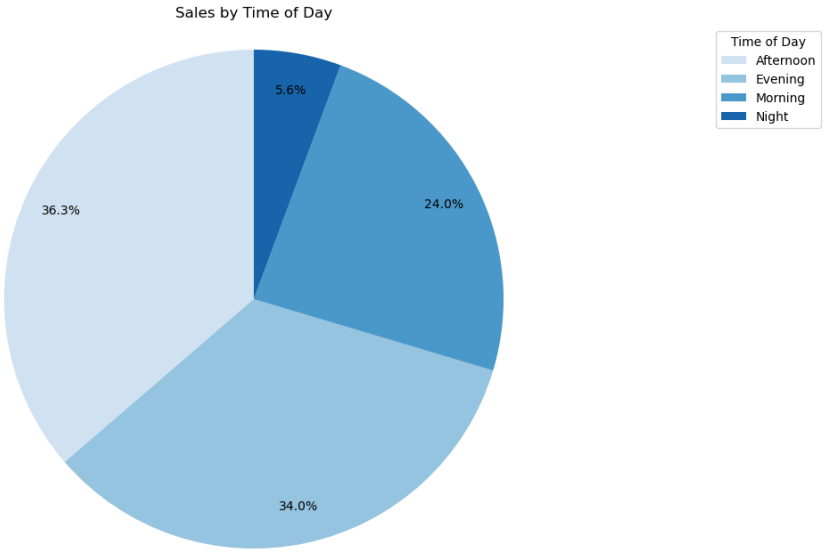
In short, the chart shows that the company's revenue structure is quite diverse but still mainly focuses on the Laptop and Computer product groups. In the coming time, the company can consider diversifying its product portfolio, investing and developing other product groups to balance the revenue structure.

4.6. Sales by each product



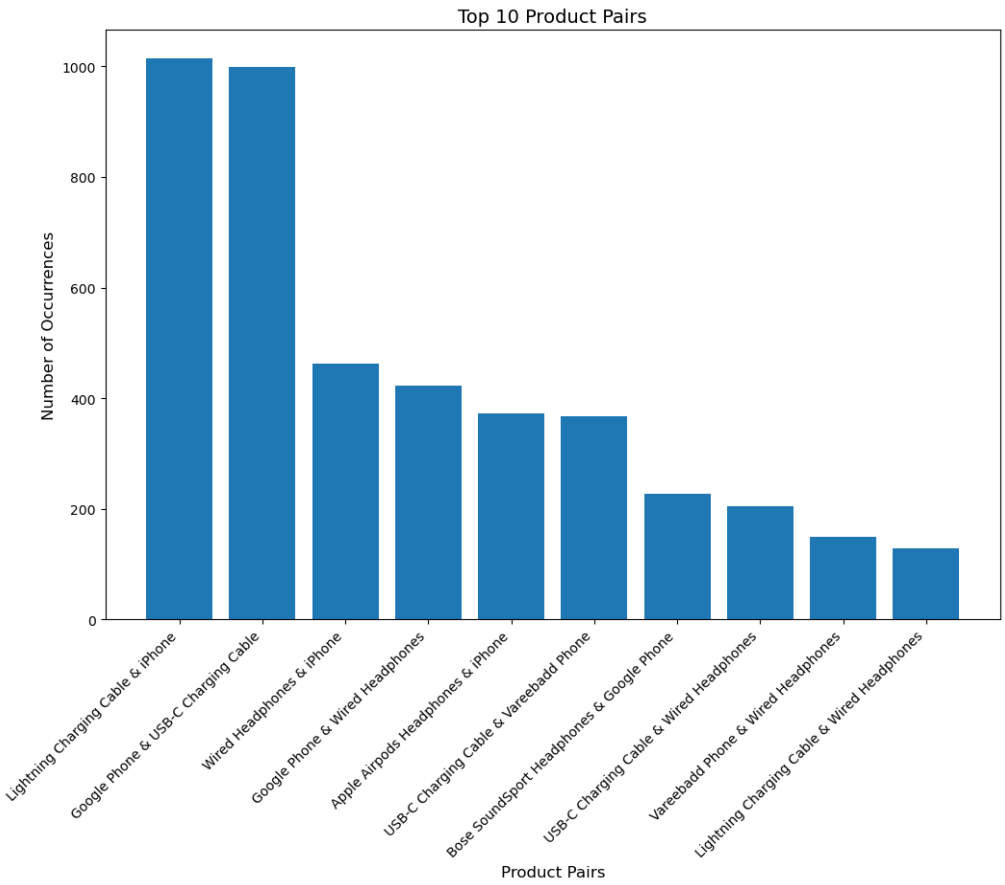
The product that generated the highest revenue was the MacBook Pro Laptop, worth about $7.6 million. This shows that the MacBook Pro Laptop is the main and popular product in the company's product portfolio. Next, phones accounted for about $5 million in revenue, equivalent to 65% of the MacBook Pro laptop. This is also an important product that attracts consumers. On the other hand, products like "USB-C Charging Cable" and "20in Monitor" have relatively low sales, suggesting that specific strategies are needed to boost their performance. Overall, high-value products and tech accessories contribute significantly to total revenue.

4.7. Sales Trends by Time of Day



The pie chart shows the proportion of sales according to the time of day. Morning sales (36.3%) and afternoon sales (34.0%) account for over 70% of the total, indicating that these are the most active shopping hours. In contrast, evening and night sales are lower, with night sales contributing only 5.6%. Customers tend to shop more during the day, when they have more time and convenience to research, compare and decide whether to buy a product. They are less likely to shop in the evening and at night. The company can focus its marketing, product promotion and sales activities during the day and early evening after work, when customers are most likely to be exposed to and interested in the product.

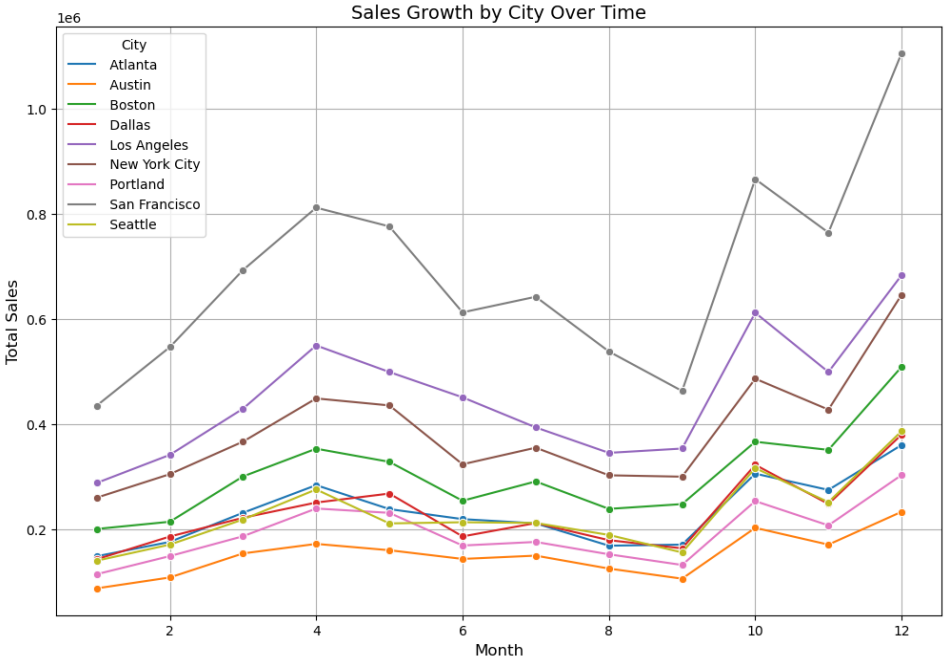
4.8. Top 10 Product Pairs



The chart shows the overall purchase frequency for different product combinations. The most popular pair is "Lightning Charging Cable & iPhone," followed by "Google Phone & USB-C Charging Cable." This suggests that customers

often purchase these products together, suggesting potential complementarities or combinations. Analyzing this data can help identify potential cross-sell and up-sell opportunities. For example, if a customer purchases an iPhone, a company might recommend a Lightning Charging Cable. Similarly, if a customer purchases a Google Phone, recommending a USB-C Charging Cable could increase sales.
By understanding these product relationships, a company can optimize product recommendations, inventory management, and marketing strategies to drive sales and improve customer satisfaction.

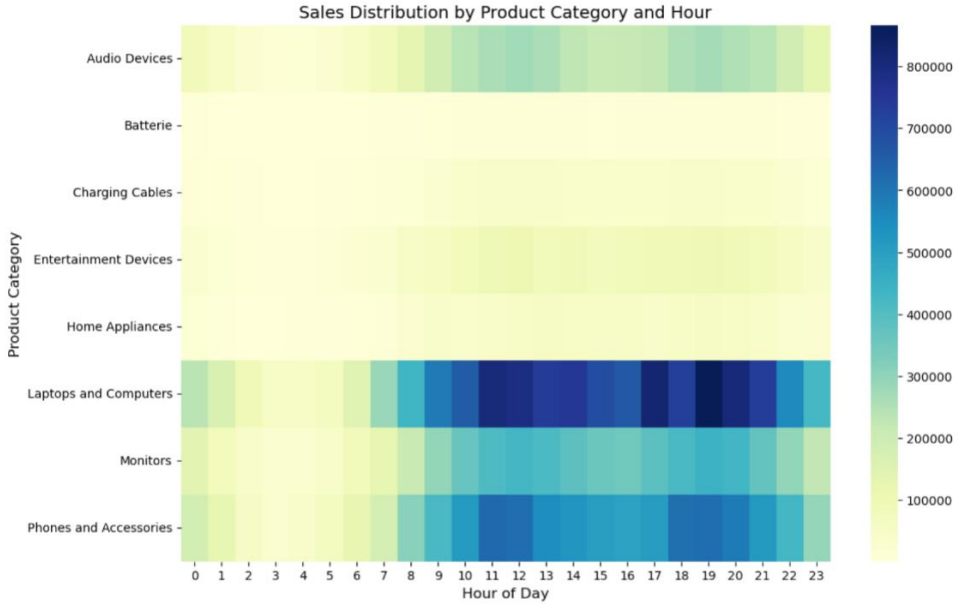
4.9. Sales Growth by City Over



The line chart shows sales growth by city over time for 9 different cities. Most cities show an upward trend in income over time, especially from October to December, which indicates a recovery or growth. Strong growth in the last quarter of the year. Income in many cities tends to increase towards the end of the year, possibly due to seasonal factors such as holidays and year-end shopping.
The two cities New York and Los Angeles show the most consistent and strong growth, especially in the last quarter of the year. San Francisco also sees significant growth, especially starting in October. Atlanta, Boston, and Dallas have relatively stable growth rates compared to other cities. For Portland and Seattle, incomes fluctuate quite a bit from month to month, but the overall trend is still growth.

Each city has different revenue per product, depending on many factors such as geographic location, income, population, and factors such as user demand. In the final months of the year, the growth rate of many cities tends to increase faster than previous months. In addition, seasonal events such as holidays and tourist seasons can greatly affect the city's revenue.

4.10. Sales Distribution by Product Category and Hour



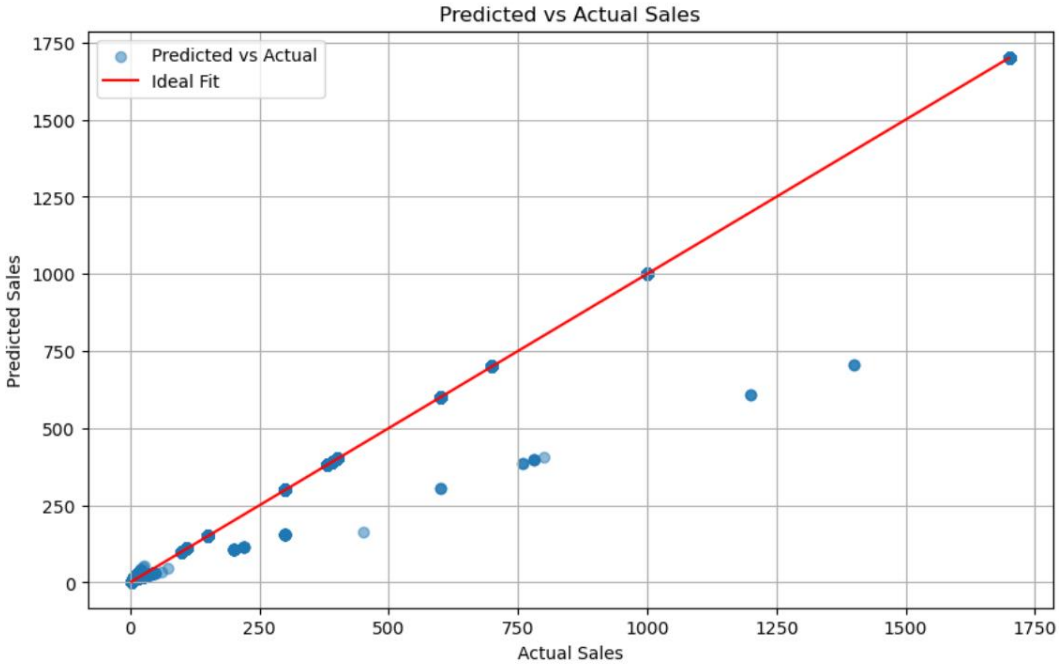
The heatmap reveals distinct sales patterns across different product categories and hours of the day. Laptops and computers exhibit peak demand during office hours, likely driven by business and productivity needs. Phones and accessories demonstrate consistent demand throughout the day, reflecting their ubiquitous use. Monitors follow a similar pattern to laptops and computers, with higher sales during work hours. Home appliances show steady demand, indicating a diverse range of products catering to various household needs. Audio devices, batteries, charging cables, and entertainment devices have relatively lower demand compared to other categories, suggesting potential opportunities for targeted marketing and promotions.

5. ADVANCED ANALYTICAL MODEL

5.1. Linear Regression

```
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
Z = df[['Quantity Ordered', 'Price Each']]
t = df[['Sales']]
Z_train, Z_test, t_train, t_test = train_test_split(Z, t, test_size=0.2, random_state=0)
```

```
The cofficient(s) b is(are) [[6.63241127 1.0011644 ]]
The intercept a is [-6.60019477]
The linear relationship: 0.9986937223894214
```



Firstly, linear regression is a statistical method used to model the relationship between a dependent variable and one or more independent variables. The main purpose in this case is to predict ‘sales’, so the dependent variable will be ‘Sales’ and the two independent variables are 'Quantity Ordered' and 'Price Each'. We name the independent variables Z and the dependent one y, then extract them from the initial dataframe. After that we start to train the data so the machine can do the regression, in our case, there is 80% of the data will be used in the training set and the remaining used in test set and the data will be choose randomly using ‘random state=0’, it can randomly choose the data for each set and ensure the data division will always be the same every time we run the code. We also calculate the intercept, coefficients and model score.

As a result the intercept aa is -6.60019477, which is the expected value of sales when the predictors are zero and the coefficients are 6.63241127 and 1.0011644. This can show us how much the sales can change with each unit change in prediction, to be more specific, we can use the formula:

$$y = -6.60019477 + 6.63241127x1 + 1.0011644x2$$

The linear relationship is very high, approximately 0.9986, which means this model has a very strong linear relationship. Therefore, the blue dots which represent the actual data points mostly lie on the red line which represents the ideal fit, where the predicted values would exactly equal the actual values. However, there are still some cases when the predictions are not accurate and they are the blue dots under the red line. These are the values that are predicted but lower than the actual one, or this can be understood as the model's prediction being lower than the actual value. Overall, our linear regression model is performing very well, with a strong ability to predict sales based on the features provided.

5.2. OLS Regression Results

OLS Regression Results

Dep. Variable:	Sales	**R-squared:**	1.000
Model:	OLS	**Adj. R-squared:**	1.000
Method:	Least Squares	**F-statistic:**	4.244e+31
Date:	Wed, 18 Dec 2024	**Prob (F-statistic):**	0.00
Time:	10:39:36	**Log-Likelihood:**	3.1225e+06
No. Observations:	130165	**AIC:**	-6.245e+06
Df Residuals:	130160	**BIC:**	-6.245e+06
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-5.171e-12	1.01e-13	-50.976	0.000	-5.37e-12	-4.97e-12
Quantity Ordered	7.999e-12	5.98e-14	133.667	0.000	7.88e-12	8.12e-12
Price Each	-3.469e-15	1.92e-15	-1.807	0.071	-7.23e-15	2.94e-16
Sales	1.0000	1.92e-15	5.22e+14	0.000	1.000	1.000
Hour	2.193e-14	4.72e-15	4.650	0.000	1.27e-14	3.12e-14

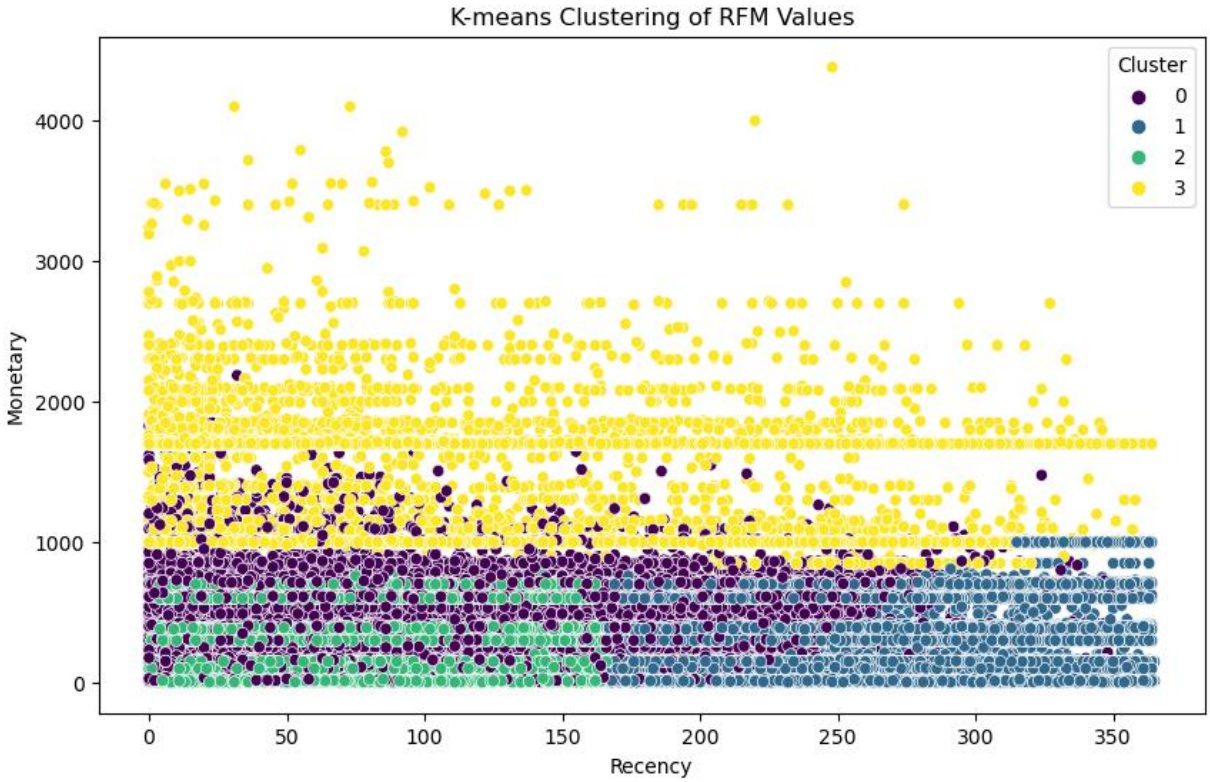
Omnibus:	62780.814	**Durbin-Watson:**	0.826
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	351553.777
Skew:	-2.332	**Prob(JB):**	0.00
Kurtosis:	9.563	**Cond. No.**	2.33e+03

The result below show good things:

- **The p-value of the F-statistic is very small**: The linear regression model is statistically significant, proving that not all coefficients are equal to 0.
- **Adjusted R^2**: In multiple regression, Adjusted R^2 is more important as it adjusts for the number of predictors. Here, Adjusted R^2 is very strong and close to the original R^2, indicating a strong linear relationship between the variables.
- **Coefficients**:
    - The coefficients for "Quantity Ordered" and "Hour" are statistically significant since their p-values are very small (0.000).
    - The coefficient for "Price Each" is not statistically significant (p-value = 0.071).

- **Intercept (const)**: The p-value for the intercept (0.000) indicates that it is statistically significant, so there is no need to remove the intercept from the model.

5.3. K-means Clustering



We used Kmeans combined with the RFM model for customer segmentation. RFM is a way to segment customers by using three criteria: R - recency, the number of days since the last time the customer made a purchase, F - frequency, the number of times when the customer returned to make a purchase, M - monetary, the total amount of money the customer spent on shopping.The analysis chart shows individual RFM data points, with each point representing a customer. The points are based on the cluster they were assigned to by the K-means algorithm. The x-axis represents the Recency value, which measures how recently a customer has made a purchase. The y-axis represents the Monetary value, which measures the total amount of money a customer has spent. In addition, we also determined the nearest date to calculate recency as 2020-01-01, we converted the data in the Orderdate column into year - month - day format and used the function " df['Order Date'].max() " to determine the nearest date to calculate recency, which is 2020-01-01. With the three criteria above, we decide to divide the customers into 4 group:

- Group 1 - blue, these are customers who do not buy frequently and spend little. These customers are likely to have low engagement with the brand and are at risk of churning.
- Group 2 - purple, representing customers with spending value from under $1000 to around $3500, with the last purchase ranging from 0 to 200 days, this group can be called the mid-range customer group. These customers are important as they have potential to either upgrade to higher-value groups or downgrade to less frequent buyers.

- Group 3 - green are customers with high purchase frequency, significant spending, and with the last purchase ranging from 0 to around 100 days. This group represents the core loyal customer base and contributes significantly to revenue.
- Group 4 - yellow, which represents customers who use a lot of money and have high purchase frequency but long time no purchase. This group represents significant revenue potential if re-engaged effectively.

With this segmentation, we will have a more accurate insight into customer segmentation based on RFM metrics. This can help identify different customer groups and make marketing strategies to each group: for example prioritize with promotions and gifts to group 3 in order to create a loyal customer base.

6. CONCLUTION & RECOMMENDATIONS

6.1. Conclusions

Monthly revenue shows a steady increase from the beginning of the year to April, a gradual decline until September, and a sharp rise toward December, consistent with holiday spending trends in retail. Revenue by product category reveals that laptops and computers dominate with 75% of total revenue, while phones and accessories contribute 25.9%; lower-performing categories like audio, home appliances, and printing may require targeted sales strategies. Revenue by city highlights San Francisco, Los Angeles, and New York as top earners due to their economic strength, while smaller cities like Dallas, Seattle, Portland, and Austin generate less revenue, possibly due to geographic and economic factors. Revenue by hour peaks midday and declines toward evening, reflecting consumer habits and providing opportunities to optimize advertising and promotions. By product, the MacBook Pro ($7.6 million) and iPhone ($5 million) lead in sales, while items like monitors and headphones show weaker performance, suggesting the need for specialized strategies. Lastly, analyzing sales growth and discrepancies between forecasted and actual revenue offers valuable insights to adjust business and marketing strategies effectively.

6.2. Business and Marketing Recommendations

- ***Targeted Marketing Strategy***

Tailoring marketing campaigns to different customer segments based on customer shopping behavior

For example: Offer personalized promotions, discounts, and messages based on segment characteristics and shopping behavior to improve customer engagement and loyalty.

Introduce seasonal sales initiatives to compensate for slower or shorter months like February. Introduce new products or upgrades to capture customer attention early in the year.

For example: Offer special discounts or curated product sets for Valentine's Day, catering to the holiday gifting market.

Focus on high-revenue cities: San Francisco, Los Angeles, and New York: These cities are leading in revenue

For example: Investing in advanced technology and innovative products, will help meet the needs of high-tech consumers. Emphasize advanced product features and benefits to attract this group of customers.

- ***Product and Quality Optimization***

Use sales data and trends to optimize inventory management activities, ensuring adequate inventory levels for high-demand/high-margin products such as mobile phones, while minimizing excess inventory for lower-volume products such as home appliances.

Implement packaging strategies to encourage larger purchases and increase revenue by offering additional products at a discount.
For example: Offer bundles for multiple purchases

The success of the MacBook Pro and phones clearly shows the high demand of the market for these products. To maintain and grow revenue, the company should continue to invest in its flagship products and improve its competitiveness through marketing activities, product innovation, and customer loyalty programs.
For low-performance products, it is necessary to conduct a comprehensive assessment of market potential and competitiveness, conduct market research and collect customer feedback to better understand consumer preferences.

- ***Customer retention and loyalty programs***

Implement targeted customer retention initiatives such as loyalty programs or exclusive perks, to reduce churn and encourage repeat purchases.
For example: Offer VVIP or member-specific benefits and rewards.

Focus on improving overall customer satisfaction by addressing issues related to product quality, customer service, and delivery experience, which can contribute to increased customer retention.

- ***Performance monitoring and analysis***

Set up regular performance monitoring measurements to track key metrics (KPIs) such as sales revenue, churn, customer retention, and customer satisfaction.

Conduct regular quality audits and customer feedback surveys to address any issues promptly and continually improve the overall customer experience.

Conduct ongoing analysis and market research to identify emerging trends, patterns, and opportunities for product improvement or innovation.

- ***Develop sustainably***

Continue to be maintained and developed. Invest in research and development (R&D) to improve the quality and features of these products. Strengthen marketing and promotion campaigns to maintain market share.

To maximize growth, the company should diversify its portfolio into areas such as audio equipment, entertainment, and home appliances, supported by profit margin analysis for effective investment decisions. Expanding product lines, improving after-sales service, and adopting advanced technologies can attract more customers. Additionally, potential growth lies in household appliances, printing supplies, and entertainment products, which require market research, R&D, and targeted marketing to meet consumer needs and boost sales.

7. REFERENCE

[1] 'Sales Dataset (E-Commerce Sales)'. Accessed: Dec. 20, 2024. [Online]. Available: https://www.kaggle.com/datasets/naofilahmad/sales-datset-product-sample

[2] 'RFM Analysis: An Effective Customer Segmentation technique using Python | by Anand Singh | Capillary Data Science | Medium'. Accessed: Dec. 20, 2024. [Online]. Available: https://medium.com/capillary-data-science/rfm-analysis-an-effective-customer-segmentation-technique-using-python-58804480d232

TEAM MEMBERS

Student ID	Student Name	Task	Contributions
22080022	Nguyễn Thị Thùy Dung (Leader)	Code Report	100%
22080003	Đoàn Hoàng Anh	Chart Analysis Business Recomendations	100%
22080005	Lê Thị Vân Anh	Report	100%
2208006	Nguyễn Bùi Châu Anh	Business Recomendations Powerpoint	100%
22080020	Trương Hoàng Diệp	Report	100%
22080039	Nguyễn Vũ Huy	Code Chart Analysis	100%
22080068	Dương Yến Nhi	Chart Analysis Business Recomendations	100%
22080080	Vũ Bảo Sơn	Data Visualization	100%
22080081	Đoàn Trọng Tấn	Data Visualization Powerpoint	100%