Hanoi University of Science and Technology
School of Informations and Comunication of Technology

Sales Data Analysis

Nguyen The An	20210006
Le Trung Kien	20214907
Nguyen Trung Truc	20214936
Do Dinh Kien	20214906

December, 2023

Contents

1 Introduction

Business analytics is a discipline that leverages data-driven methods and quantitative analysis to gain insights, make informed decisions, and optimize business processes. It involves the use of statistical analysis, predictive modeling, data mining, and other analytical techniques to extract meaningful patterns and trends from large sets of data.

The primary goal of business analytics is to support data-driven decision-making within organizations, helping them solve complex problems, identify opportunities, and enhance overall performance. By examining historical data and applying advanced analytical methods, businesses can uncover valuable insights that contribute to strategic planning, operational improvements, and better understanding of customer behavior.

Business analytics is widely applicable across various industries and functional areas, including finance, marketing, supply chain management, human resources, and more. Organizations that embrace business analytics gain a competitive edge by making data-driven decisions that contribute to efficiency, innovation, and overall business success. As technology continues to advance, the role of business analytics in shaping strategic initiatives and fostering a data-driven culture becomes increasingly crucial for modern businesses.

2 Problem Description

Sales data analysis involves the process of inspecting, cleaning, transforming, and modeling sales-related data with the goal of discovering useful information, drawing conclusions, and supporting decision-making within a business. This analytical approach is crucial for businesses to gain insights into their sales performance, customer behavior, and overall market trends. It contains some basic processes:

1. Data Collection

   Sales data analysis begins with the collection of relevant data. This data may include information about individual sales transactions, such as the date of sale, products sold, quantities, prices, customer details, and more.

2. Data Cleaning and Preprocessing

   Raw data collected may contain errors, inconsistencies, or missing values. Data cleaning involves the process of rectifying these issues to ensure accurate and reliable analysis. This may also include transforming data into a usable format.

3. Exploratory Data Analysis (EDA)

   EDA involves the initial examination of the data to summarize its main characteristics, often using statistical graphics and other data visualization methods. Analysts explore patterns, trends, and relationships within the data to form hypotheses.

4. Descriptive Analytics

   Descriptive analytics involves summarizing and presenting key features of the data, such as total sales, average transaction value, and product/category performance. It provides a snapshot of historical data to understand what has happened in the past.

Sales data analysis might contains more processes such as: trend analysis, customer segmentation,... based on our objective and business logic.

3 Dataset

In this project we will work with Retail Sales Data in Istanbul. Our dataset contains shopping information from 10 different shopping malls between 2021 and 2023. We have gathered data from various age groups and genders to provide a comprehensive view of shopping habits in Istanbul. The dataset includes essential information such as invoice numbers, customer IDs, age, gender, payment methods, product categories, quantity, price, order dates, and shopping mall locations. Information about each attribute:

1. **invoice_no:** Invoice number. Nominal. A combination of the letter 'I' and a 6-digit integer uniquely assigned to each operation.

2. **customer_id:** Customer number. Nominal. A combination of the letter 'C' and a 6-digit integer uniquely assigned to each operation.
3. **gender:** String variable of the customer's gender.
4. **age:** Positive Integer variable of the customer's age.
5. **category:** String variable of the category of the purchased product.
6. **quantity:** The quantities of each product (item) per transaction. Numeric.
7. **price:** Unit price. Numeric. Product price per unit in Turkish Liras (TL).
8. **payment_method:** String variable of the payment method (cash, credit card, or debit card) used for the transaction.
9. **invoice_date:** Invoice date. The day when a transaction was generated.
10. **shopping_mall:** String variable of the name of the shopping mall where the transaction was made.

The dataset contains 99457 records.

4 Designing System

4.1 Data Preprocessing

4.1.1 Check for missing and duplicated data

Before conducting data analysis, it is essential to check for missing and duplicated data and handle them. Below is our code for checking missing and duplicated data

```
def show_missing(df):
    feature_names = []
    data_types = []
    total_count = []
    unique_count = []
    missing_count = []
    missing_percentage = []

    for feature in df.columns:
        feature_names.append(feature)
        data_types.append(df[feature].dtype)
        total_count.append(len(df[feature]))
        unique_count.append(len(df[feature].unique()))
        missing_count.append(df[feature].isna().sum())
        missing_percentage.append(round((df[feature].isna().sum() / len(df[feature])) * 100, 2))

    output_df = pd.DataFrame({
        'feature': feature_names,
        'dtype': data_types,
        'total_count': total_count,
        'unique_count': unique_count,
        'missing_count': missing_count,
        'missing_percentage': missing_percentage
    })

    return output_df.sort_values("missing_count", ascending=False).reset_index(drop=True)

# Show any duplicated value
df[df.duplicated() == True]

# Show missing Value
show_missing(df)
```

There is no value returns for duplicated value and missing value:

4.1.2 Change data type

After taking a look at our data, we seem to have some remarks:

1. The data type of ”invoice_date” is incorrect.

2. There is no missing data.

3. There are no unusual expressions such as ? - */ . null, minus price, age, quantity, etc.

4. There are no different classes representing the same class (e.g., abcxxx → xxxabc).

5. Each customer is unique, meaning that customers have visited only once.

6. Each invoice is unique, meaning that only one product has been purchased in each invoice.

7. ”invoice_no” and ”customer_id” serve as indexes and represent each row.

We need to change the data type of ”invoice_date”. Below is our code:

```
# change type
df['invoice_date'] = df['invoice_date'].astype('datetime64')
```

4.2 Exploratory Data Analysis

Exploratory data analysis (EDA) is the process of analyzing data using simple concepts from statistics and probability theory and presenting the results in an easy-to-understand format that supports visuals.

4.2.1 Gender Distribution

First, we analyzed the data to find the answers to the following questions:

1. How is the shopping distribution according to gender?

2. Which gender did we sell more products to?

3. Which gender generated more revenue?

4. Distribution of purchase categories relative to other columns?

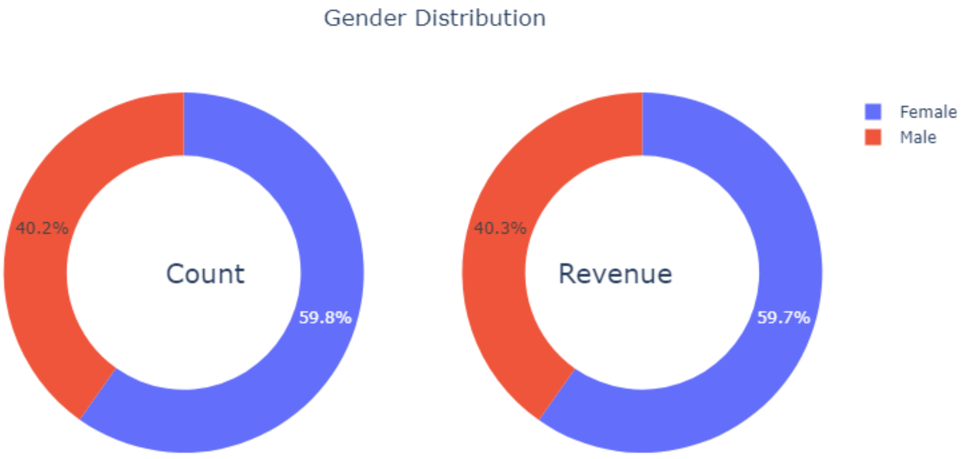

Figure 1: Gender Distribution

From 2 above graphs, we have some remarks about gender distribution:

1. We have generated more revenue from women.
2. We have sold more products to women.
3. The average revenue per product shows men generate slightly more revenue than women.

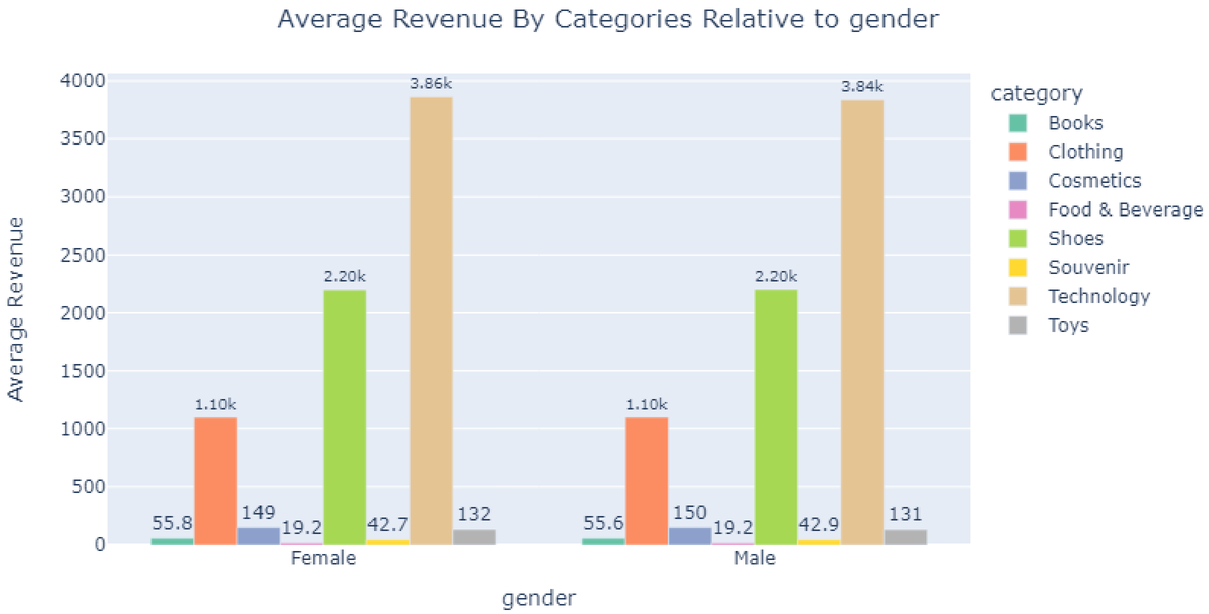

Figure 2: Gender Distribution over Average Revenue by each Category

As we can see on the graph:

1. Average spending is the same and has not changed according to any category.
2. We expect something different, such as women spending more on clothing and men spending more on technology.
3. The highest revenue has been obtained in the categories of technology, clothing, and shoes.

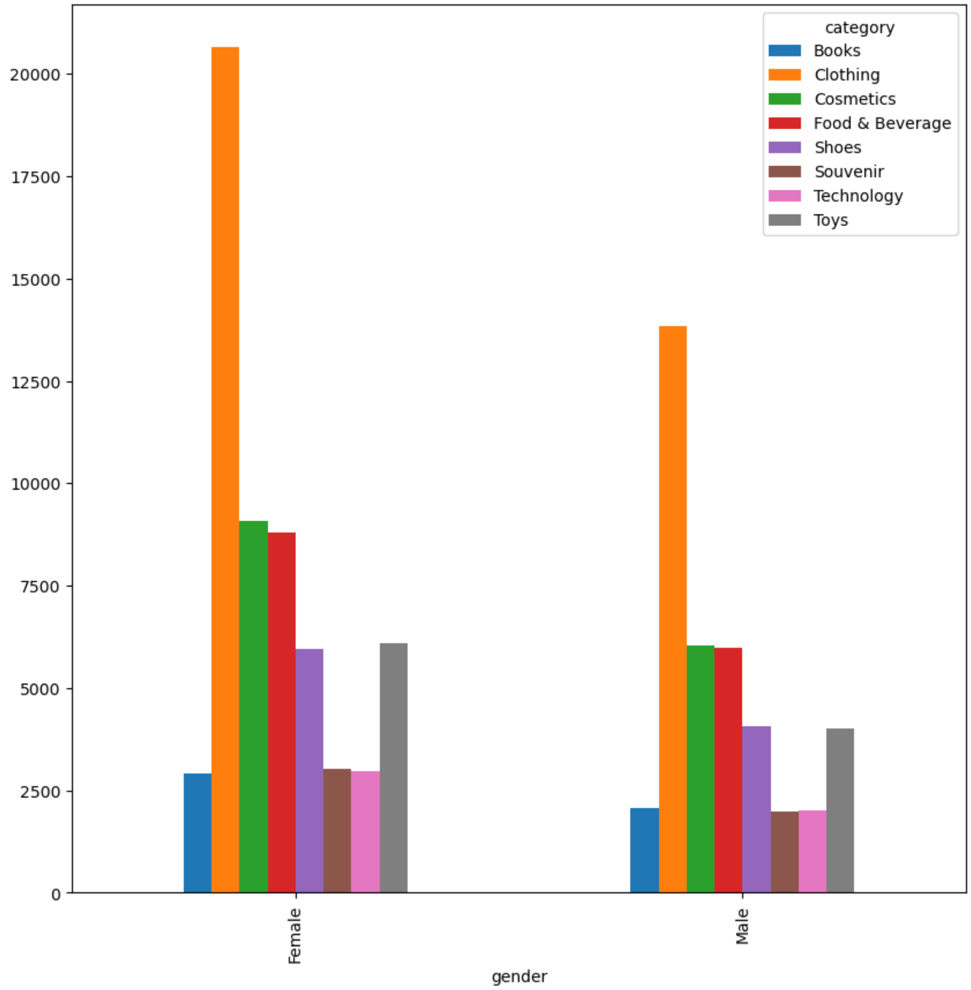

Figure 3: Gender Distribution over Number of Sold Products by each Category

Two important remarks about number of sold products

1. The ratio of products sold to women compared to products sold to men appears to be the same when broken down by category and gender.

2. Women have purchased 1.5 times more products than men in all categories.

4.2.2 Age Distribution

1. How is the shopping distribution according to age?

2. Which age category did we sell more products to?

3. Which age category generated more revenue?

4. Distribution of purchase categories relative to other columns?

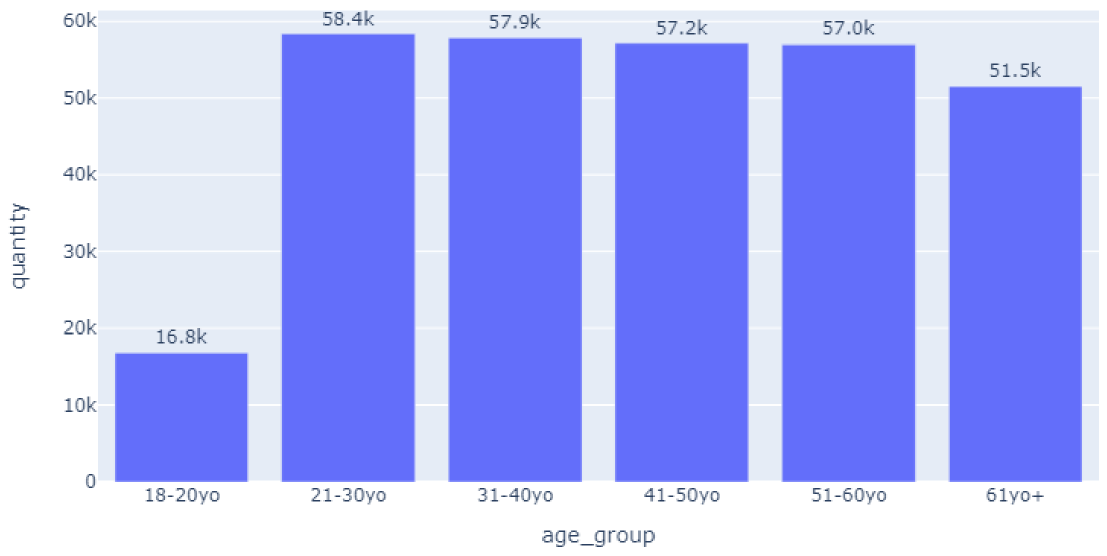

Figure 4: Age Distribution over quantity of products

According to the data, the highest quantity of products sold occurred in the middle age group, while young people had the lowest quantity of products sold.

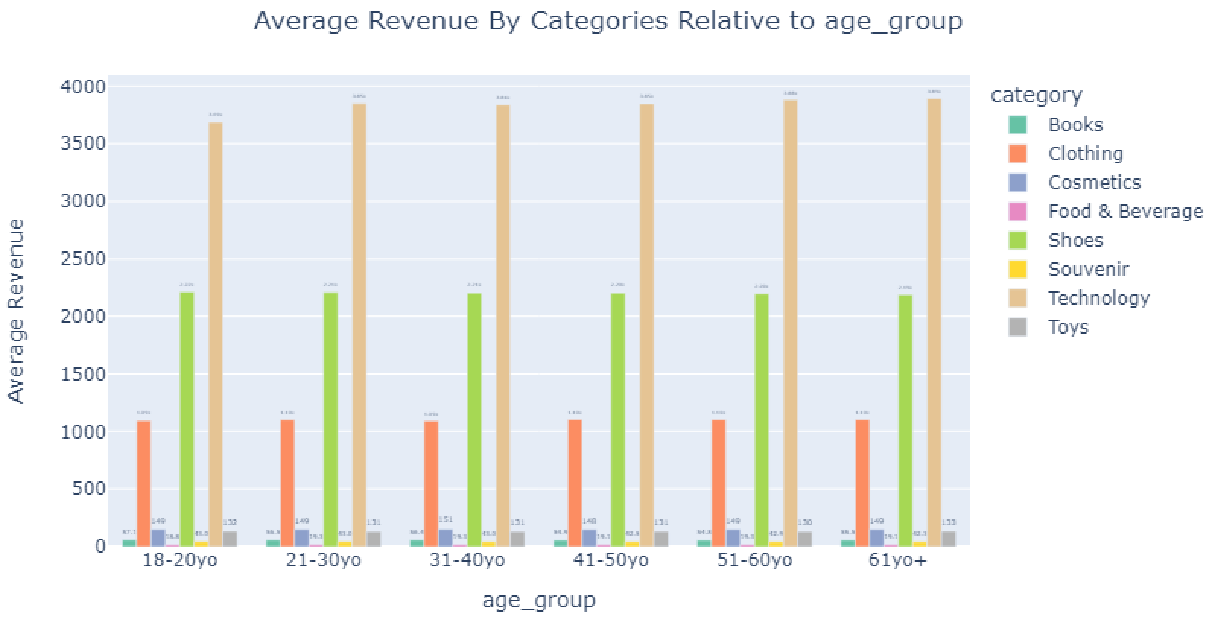

Figure 5: Age Distribution over Average Revenue by each Category

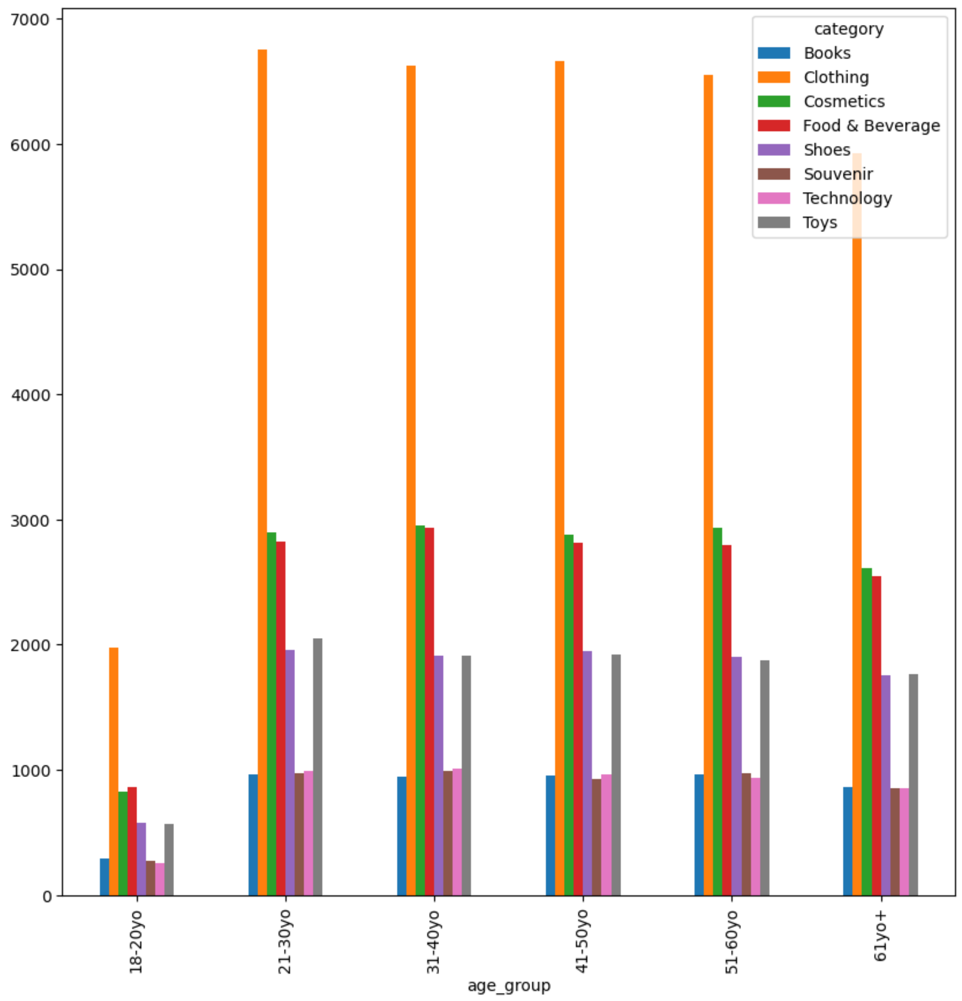

Figure 6: Age Distribution over Quantity of Sold Products by each Category

We can conclude that:

1. As seen, average spending is the same and has not changed according to any age category.
2. We expected different spending for different age category.
3. Middle-aged people have bought the same amount of products as old people.
4. Middle-aged people bought 1.5 times more products in all categories than young people.

4.2.3 Payment Method Distribution

We going to find out about:

1. Does the payment method have a relation with other columns?
2. How is the distribution of the payment method?

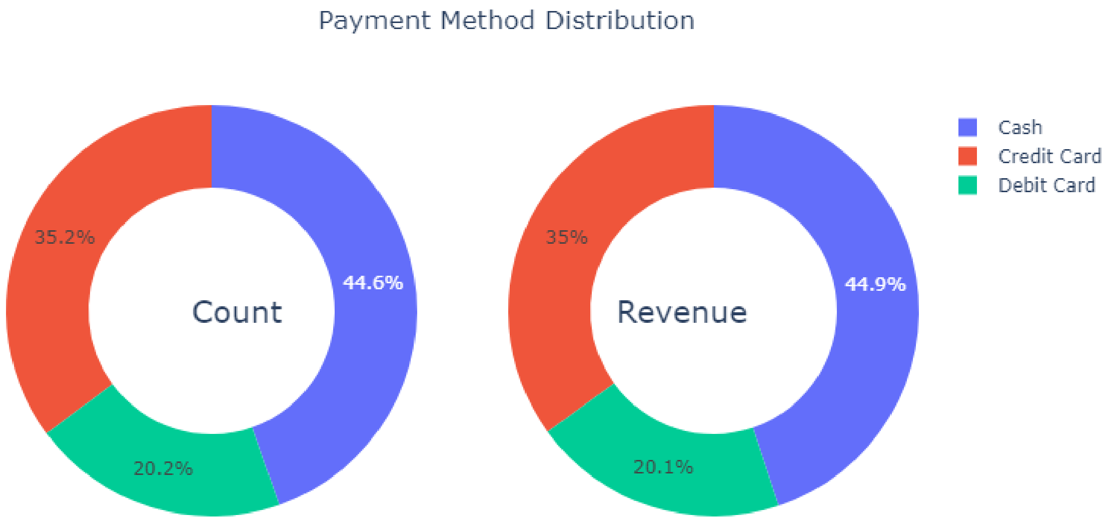

Figure 7: Payment Distribution

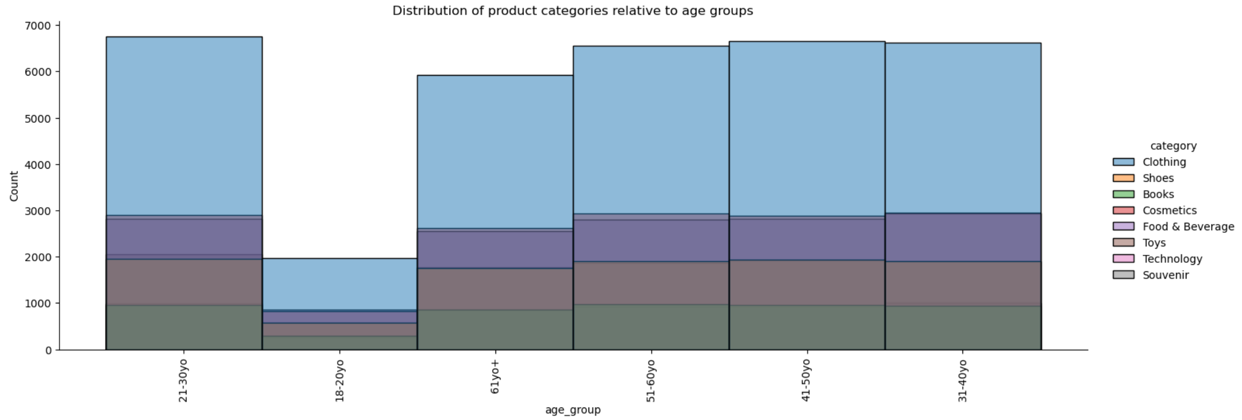

Figure 8: Payment Distribution over Age Group by Category

1. According to the payment distribution, the highest revenue and sales volume come from cash payments, followed by credit card payments and lastly debit card payments.
2. There seems to be no relationship between the age group variable and payment method.
3. In all product categories, the amount of cash payments is 2.2 times higher than payments made with a bank card.

4.2.4 Date Distribution

We are going to find out about if the date affects the number of purchased products or not.

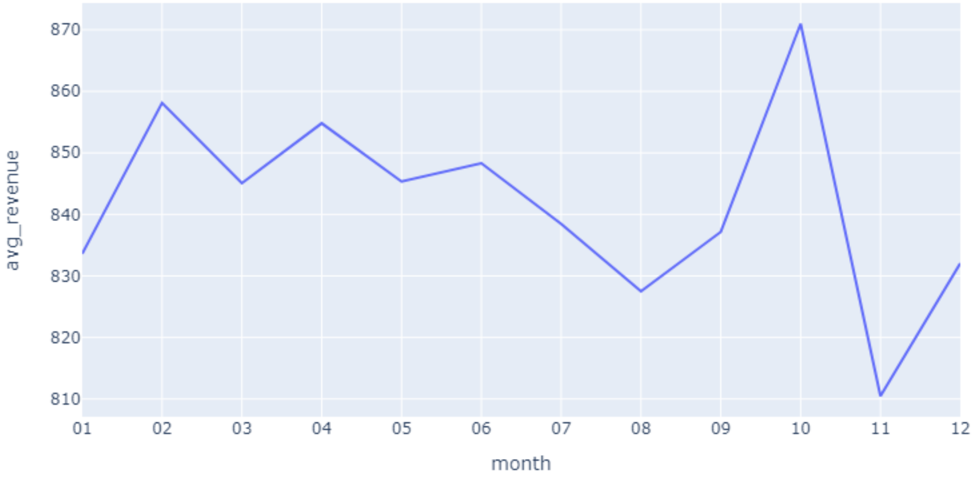

Figure 9: Revenue over Month

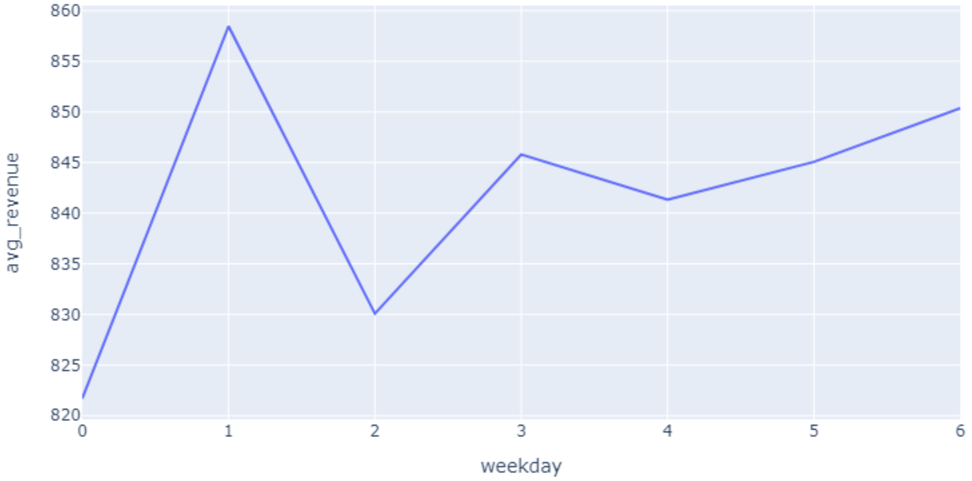

Figure 10: Revenue over Weekday

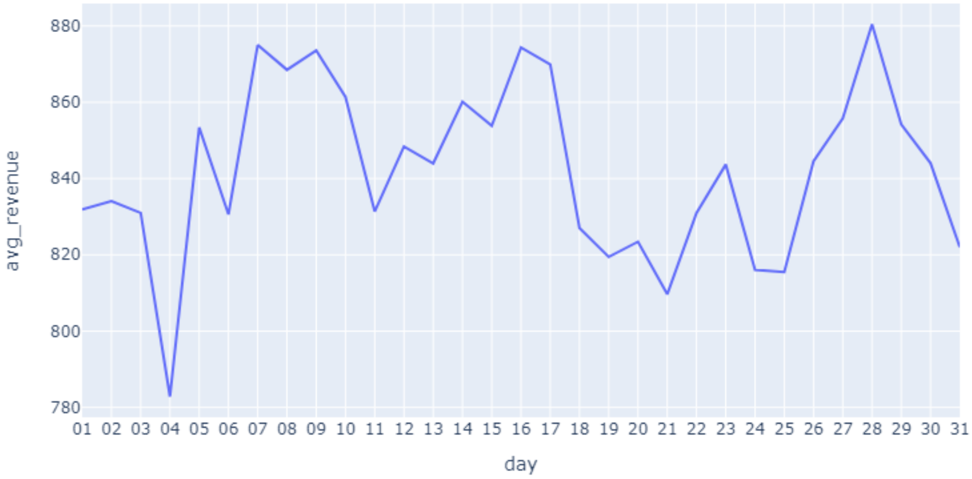

Figure 11: Revenue over Day in Month

The annual difference is the same, there is a 100 TL difference daily, a 60 TL difference monthly, and a 40 TL difference weekly. The date does not seem to have any effect on the shopping.

4.3 Model

For retail data, we conducted monthly sales forecasting for the next year (2024) to see how sales would recover and build strategies accordingly. We approach this problem as a time series forecasting problem, with the input being our historical monthly sales for each mall and the output is the forecast.

We built a Long Short Term Memory (LSTM) neural network to predict our sales data. We will not go deep into how LSTMs actually learn from historical data. An example forecast is shown below:

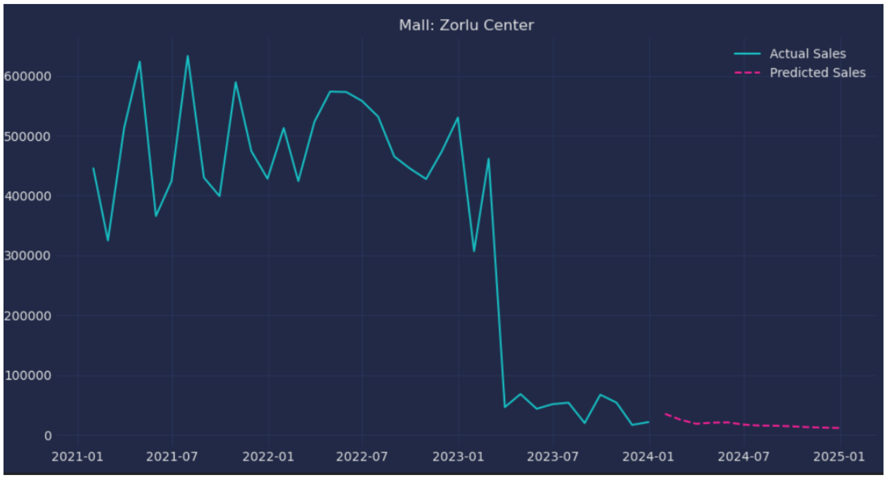

Figure 12: Sales forecasting

We can see that the sales took a huge plummet at the start of 2023, most likely due to COVID and still cannot recover from the pandemic. Our forecast shows that the sales will not increase in the next year. This is the common trend for all malls. Therefore we would need appropriate marketing and advertising strategies to promote customers shopping.

5 Achievements

5.1 Insights

After analyze the data and visualize key features, some key insights can be derived from the data.

- **Key demographics**
  - **Gender Distribution:** We have observed that we generate more revenue from women, and they also purchase more products than men. However, the average revenue per product is slightly higher for men.
  - **Age Distribution:** The highest quantity of products sold is within the middle age group, while young people contribute the least to the overall sales volume. Interestingly, middle-aged individuals consistently purchase 1.5 times more products across all categories compared to young people.
- **Purchasing patterns**
  - **Monthly Revenue Peaks:** The periods generating the most average revenue are mid-month (around days 15 and 16) and end of the month (around day 28).
  - **Category Revenue:** The highest revenue is observed in the categories of technology, clothing, and shoes.
- **Product Sales Ratios**
  - **Gender-Based Product Sales:** Although the ratio of products sold to women compared to men seems consistent across categories, women have purchased 1.5 times more products than men in all categories.
  - **Age-Based Product Sales:** Middle-aged individuals consistently buy 1.5 times more products in all categories compared to young people.
- **Payment Methods**
  - **Payment Method Distribution:** Cash payments contribute the highest revenue and sales volume, followed by credit card payments and debit card payments.
  - **Age Group and Payment Method:** No significant relationship is observed between the age group variable and the chosen payment method.
- **Customer Value Segmentation**
  - **Most Valuable Customers:** Middle-aged women emerge as the most valuable customers, consistently contributing higher revenue.
  - **Least Valuable Customers:** Conversely, young men appear to be the least valuable customers, contributing less to overall sales.

5.2 Strategic Recommendations

From the insights and the data, we propose some marketing and price optimization strategies for malls to follow:

- **Targeted Campaigns for Middle-Aged Women:** Develop marketing campaigns specifically tailored to middle-aged women, our most valuable customer segment. Highlight promotions, discounts, and exclusive offers on technology, clothing, and shoes to attract and retain this demographic.
- **Young Men Engagement Campaigns:** Create targeted campaigns to engage young men, our least valuable customer segment. Consider promotions, discounts, or product bundles that may appeal to this demographic and address any barriers to their participation.
- **Loyalty Program Enhancements:** Evaluate and potentially enhance the existing loyalty program to further incentives repeat business. Consider offering exclusive discounts, early access to sales, or bonus points for purchases in high-revenue categories.

- **Payment Method Promotions:** Encourage the use of specific payment methods by offering promotions or discounts. For example, create cash-back offers for customers using debit or credit cards to diversify payment methods.
- **Communication of Consistent Pricing:** Emphasize the consistency in pricing, highlighting that the date does not seem to influence shopping behavior. This can be used to build trust with customers and emphasize transparency.
- **Referral Program for Women:** Implement a referral program targeted at women, given that they have purchased 1.5 times more products than men across all categories. Offer incentives for women customers to refer friends or family

All of these strategies should be applied since sales are still recovering after COVID. According to our sales forecasting, sales will take years to recover to its peak but with adequate strategies, this will not take as long.

6 Discussion and Conclusion

With the data we retrieved from Kaggle, we have conducted data analysis and visualize to retrieve insights, build Machine Learning model to forecast sales in upcoming year and propose business strategies for shopping malls. There are much more that can be done such as customer segmentation to define segments customer and adjust our strategies accordingly. We originally planned to conduct RFM (Recency, Frequency, Monetary) Analysis, you can refer to here for more detail about that. RFM helps create segments of 'important customers'. Those customers are the ones most likely to have the highest return on marketing investments. Specifically, the analysis can help improve targeting, reduce costs, and increase return on advertising investments. However, this could not be done due to our data having only 1 customer per invoice, making the frequency of each customer being 1. This makes our RFM misleading since we cannot correctly define customers that are 'important'.

In the course of doing the projects, we have learnt many knowledge, including how data can tell us about business and how do we use those data to gain insights, to help our business. We have worked our teamwork, gain a much deeper and more concise understanding of the subject. For that we would like to thank Dr. Nguyen Binh Minh for giving us this opportunity.

References

[1] Customer shopping dataset - retail sales data. https://www.kaggle.com/datasets/mehmettahiraslan/customer-shopping/dataset/data.