CITY of ISTANBUL



Retail sale report



Ha Luong

September 2023

Analysis of Retail Sales in Istanbul 2021-2023

Introduction

This report is tailored for the executive leadership team of a prominent retail company situated in Istanbul. It encompasses an extensive analysis of sales patterns within ten distinct shopping malls across the city, spanning the period from 2021 to March 2023. Additionally, this report proposes the development of a predictive model aimed at estimating both the total daily retail sales and the daily sales figures for each individual shopping mall.

The analysis conducted in this report leverages sophisticated data analysis software, specifically Tableau and Python, to extract meaningful insights from the comprehensive sales data available. Through the utilization of these tools, we have unearthed invaluable information that can play a pivotal role in guiding informed strategic decisions and improving the overall performance of your retail operations.

Objectives

- Identify trends and seasonality in Istanbul shopping mall sales.
- Segment and analyze customer groups.
- Analyze product sales across various categories.
- Determine the best model to predict daily sales in Istanbul shopping malls among: ARIMA, SARIMA and XGBoost models.

Data extract

The report utilizes shopping information from 10 different shopping malls between 2021 and 2023. Access the data at:

https://www.kaggle.com/datasets/mehmettahiraslan/customer-shopping-dataset

Attribute Information:

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

From the name of shopping mall, geographical coordinates of shopping malls is generated using the Geopy package in Python in order to create map in analysis.

Data wrangling

Data cleaning:

- Convert **invoice_date** to date type
- Check for null and duplicate: No missing and duplicate data found in the dataset
- Create **sales** = **quantity** * **price** (as float type)

To generate daily sales forecasts, we have organized the daily sales data by aggregating sales based on the invoice_date. Furthermore, we have computed the daily sales figures for each shopping mall by applying filters specific to each mall's data.

Data cleaning for daily sale data:

- Check for complete of time series dataset:

Shopping mall	**Number of days**
Emaar Square Mall	797
Istinye Park	797
Kanyon	797
Mall of Istanbul	797
Metrocity	797
Metropol AVM	797
Cevahir AVM	**796**
Viaport Outlet	**796**
Forum Istanbul	**795**
Zorlu Center	**794**

There are some missing days in the time series of 4 shopping malls: Cevahir AVM, Viaport Outlet, Forum Istanbul and Zorlu Center.

- Identify the missing dates, introduce new rows for these missing dates, and substitute the missing sales values with the average sales figure, as there has been relatively little fluctuation in sales throughout the period.

Our output comprises:

- The cleaned original data (in transaction-level detail)
- Cleaned data for total daily sales
- Daily sales data for the 10 individual shopping malls.

General analysis

Observing the map of the 10 shopping malls across Istanbul, we notice that the three largest shopping malls, represented by the size of the circles, are consistently Mall of Istanbul, Kanyon, and Metrocity (Figure 1). These top 3 shopping malls have maintained their positions over the past three years.

The sales line graph displays a horizontal trend line, indicating that there has been no significant increase or decrease in sales from 2021 to March 2023.

The dataset also exhibits no distinct patterns or seasonality whether analyzed by month of the year or by day of a week. When examining each shopping mall, there is also little to no pattern and seasonality.

In Istanbul, cash payment still remains the prevailing method, accounting for 44.7% of all transactions.

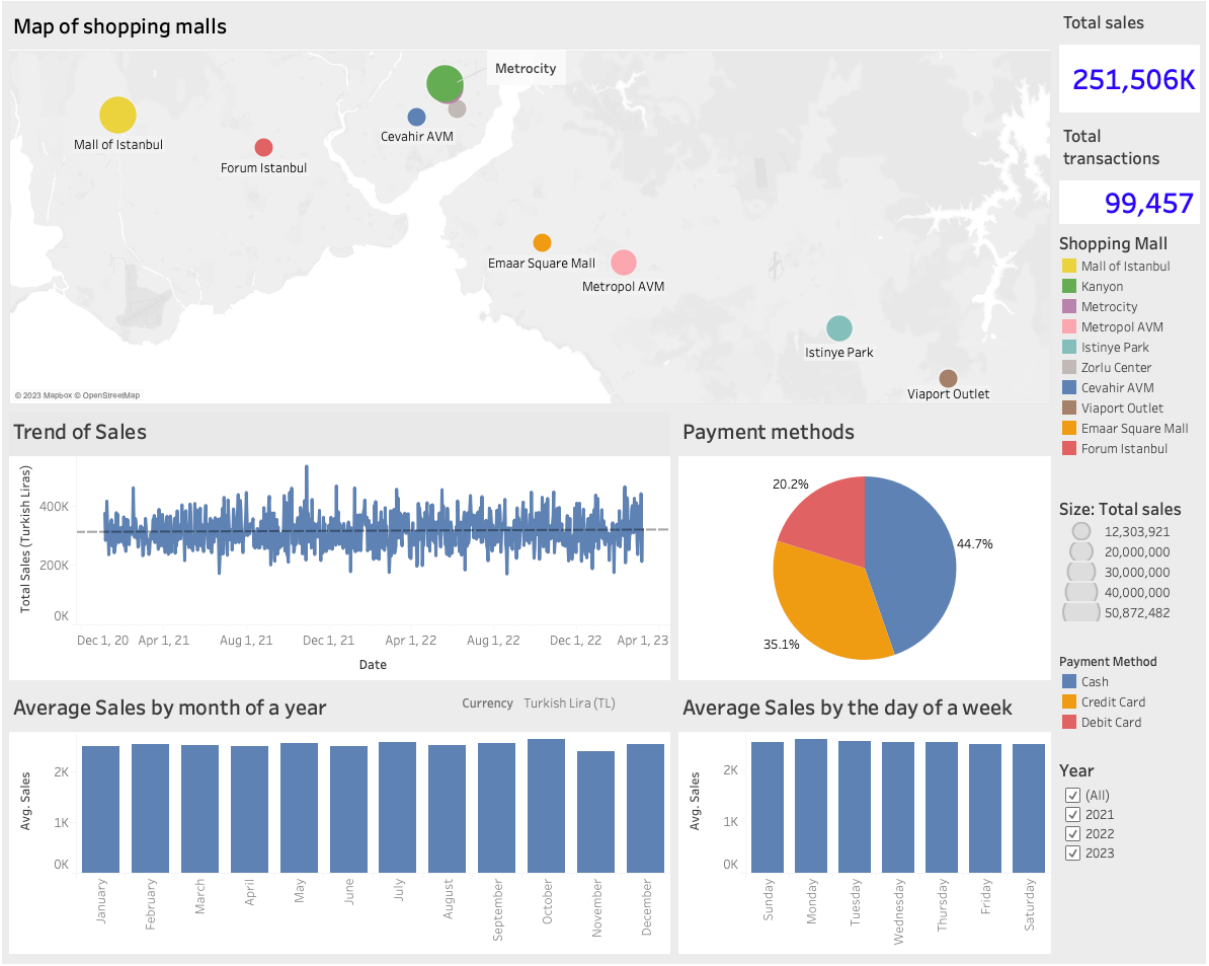


Figure 1: General analysis of total sales

Customer segmentation analysis

The customer segmentation analysis reveals that the majority of customers fall within the 25 to 64-year-old age range, with fewer customers younger than 25 or older than 64. Females represent nearly 60% of both the total customer base and total sales (Figure 2).

The bar chart below break down total sales by gender and age groups. What is interesting here is the different in spending habits among age groups and between genders. The color in the sales distribution chart indicate the average sales per person.

The highest-spending group consists of male customers aged 35-44 and over 65 (highlighted in dark red on the sales chart). Conversely, the lowest-spending customers are male groups under 25 and between 54-64 years old.

In contrast, there is less variation among female groups. The highest spending group comprises those under 25 years old, while the lowest-spending group falls within the 25-34 age range. This trend contrasts with the spending patterns of male groups.

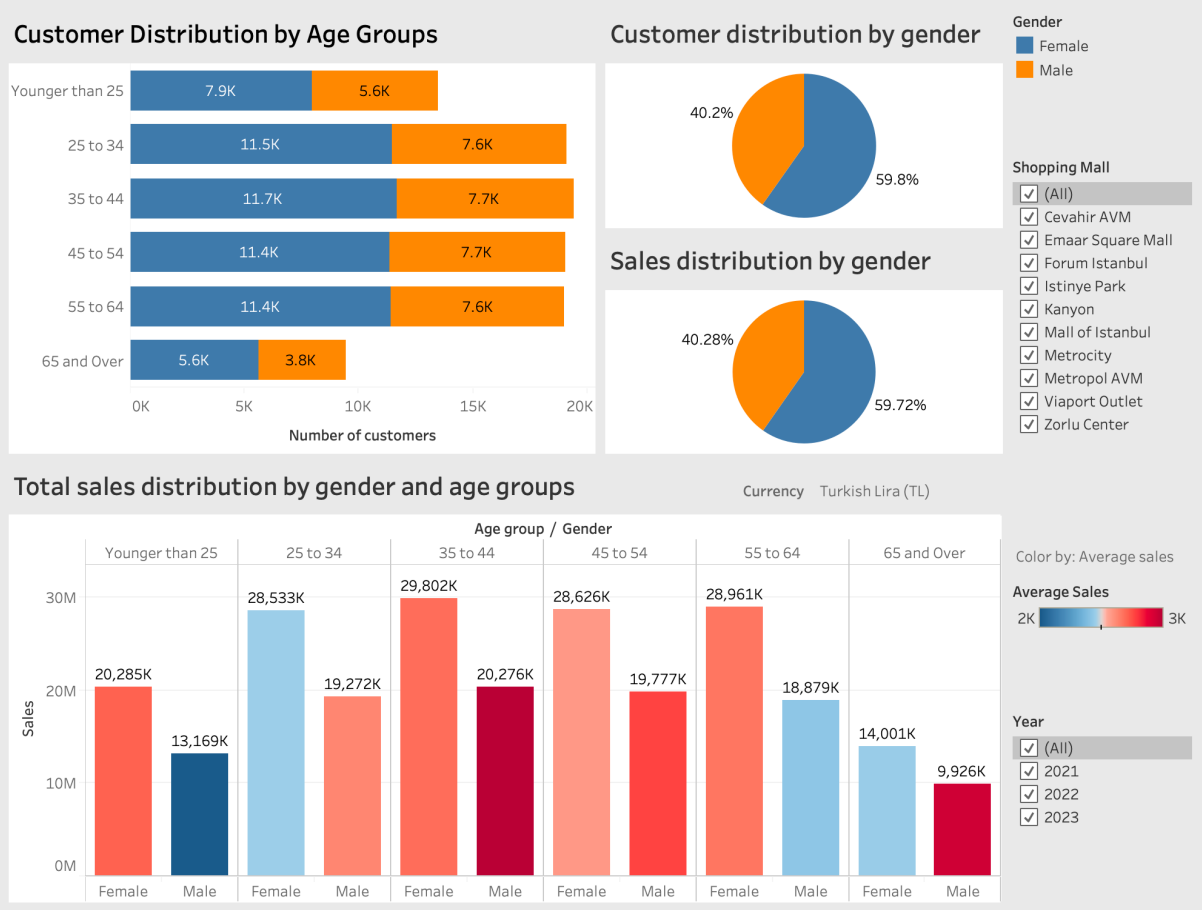


Figure 2: Customer segmentation analysis

Product category analysis

Clothing stands out as the most popular and revenue-generating category, leading in both the number of transactions and total sales. It represents nearly 35% of the total number of transactions and contributes to 45% of the total sales (Figure 3).

Cosmetics and Food and Beverage secure the 2nd and 3rd positions in terms of the number of transactions, each accounting for approximately 15%. However, these categories do not significantly contribute to total sales.

Shoes and Technology claim the 2nd and 3rd positions in total sales.

Technology stands out with the highest unit price, resulting in a substantial contribution to total sales despite its lower sales volume.

Finally, the sales lines show little variation in total sales for each product category.

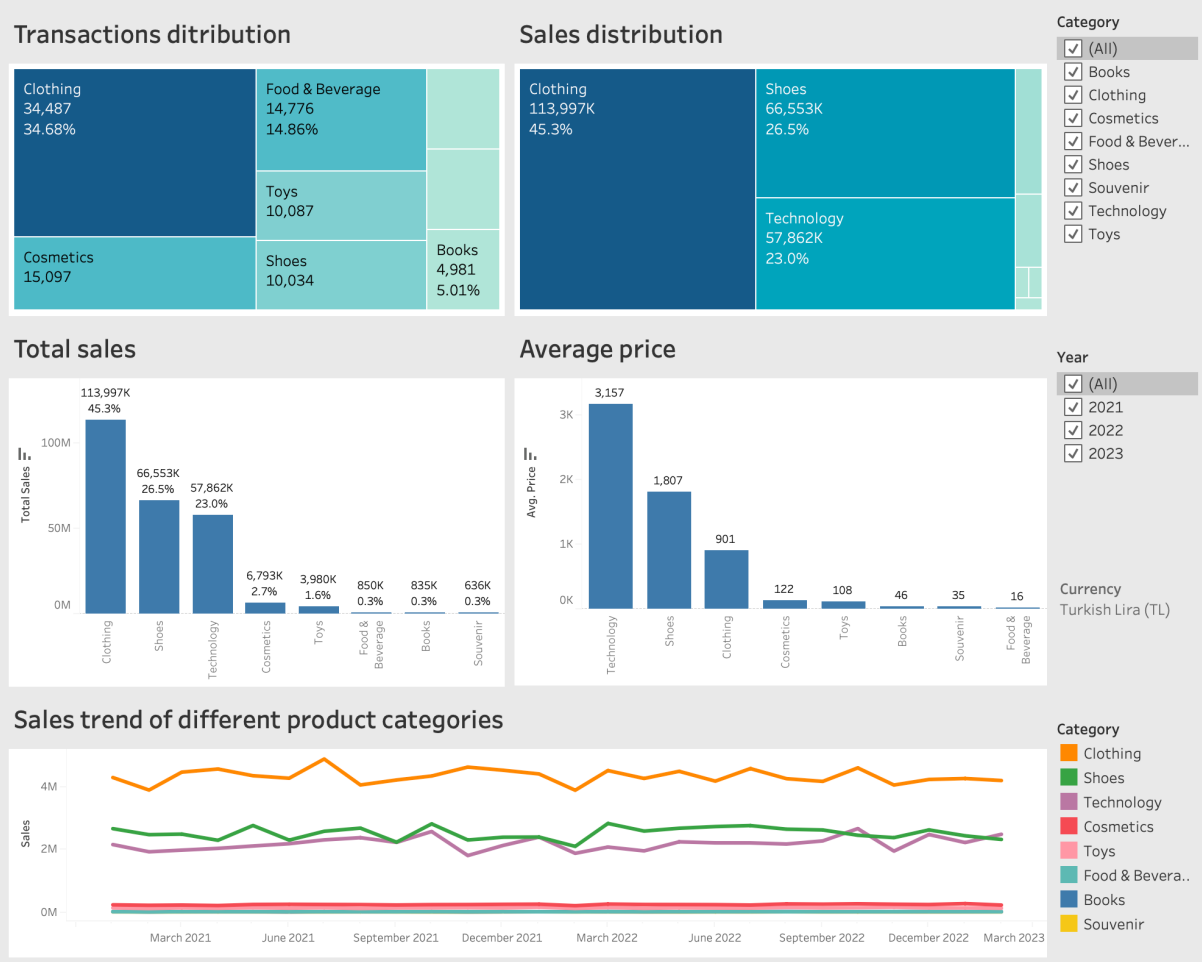


Figure 3: Product category analysis

Total Daily Sales Forecast

In this section, the report will use ARIMA, SARIMA and XGBoost model to predict to total daily sale of shopping malls in Istanbul. The dataset will be split into training and test set:

- Training set: data from Jan 2021 to September 2022 (21 months)
- Test set: data from October 2022 to March 2023 (5+ months)

Exploring time series

Even after aggregating the dataset on a daily basis, there are no discernible patterns or seasonality observed, whether analyzed on a monthly or weekly basis.

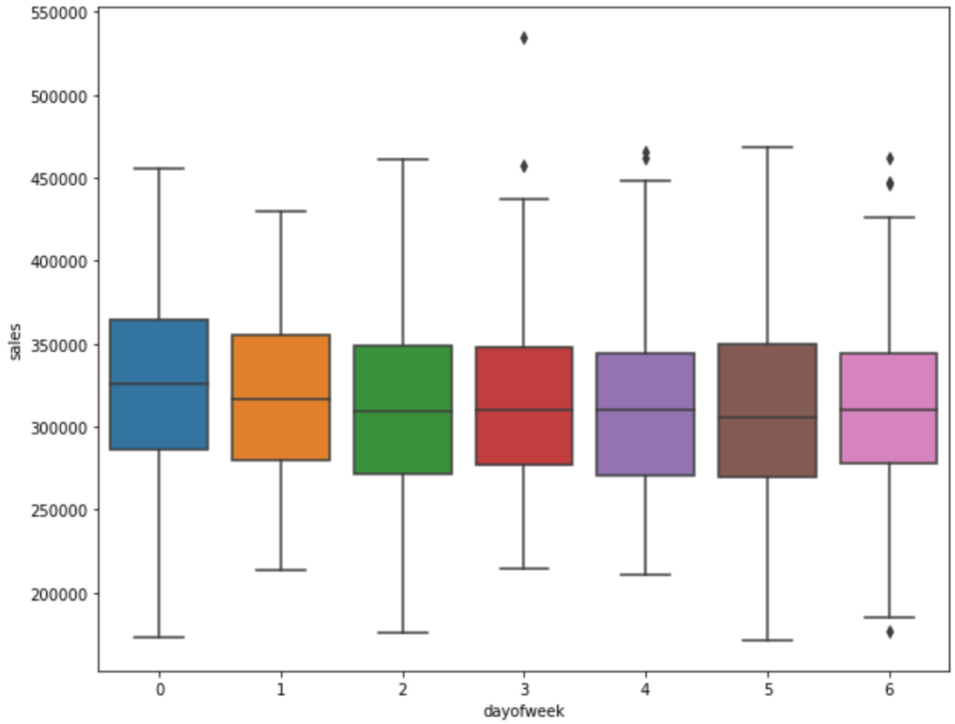


Figure 4: Distribution of sales per day of the week

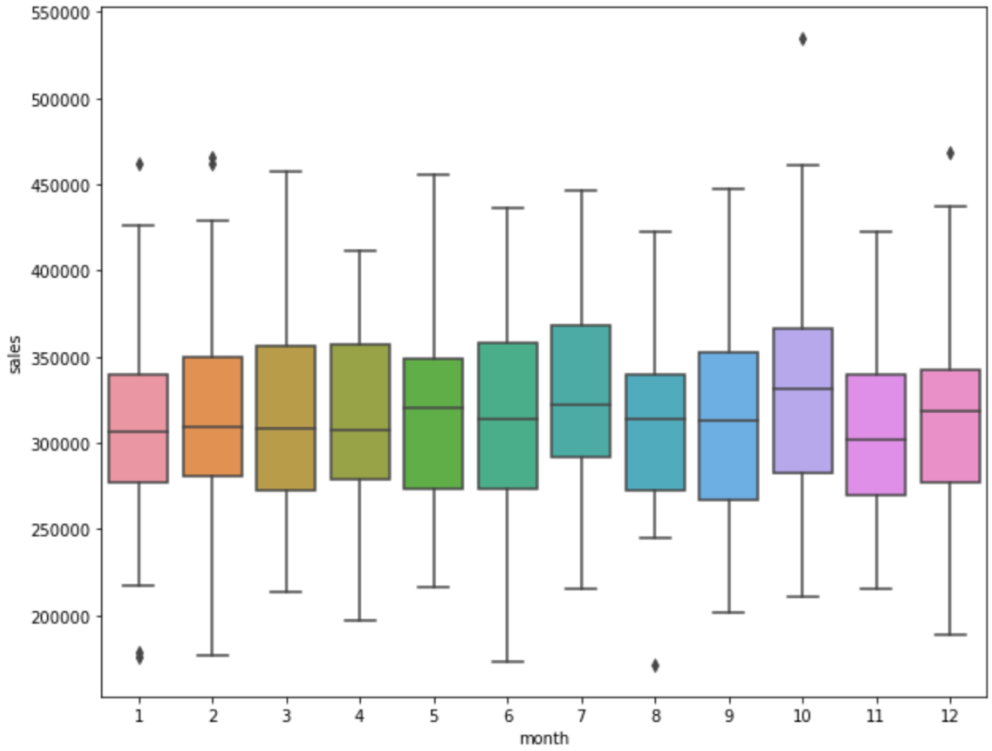


Figure 5: Distribution of sales per month of year

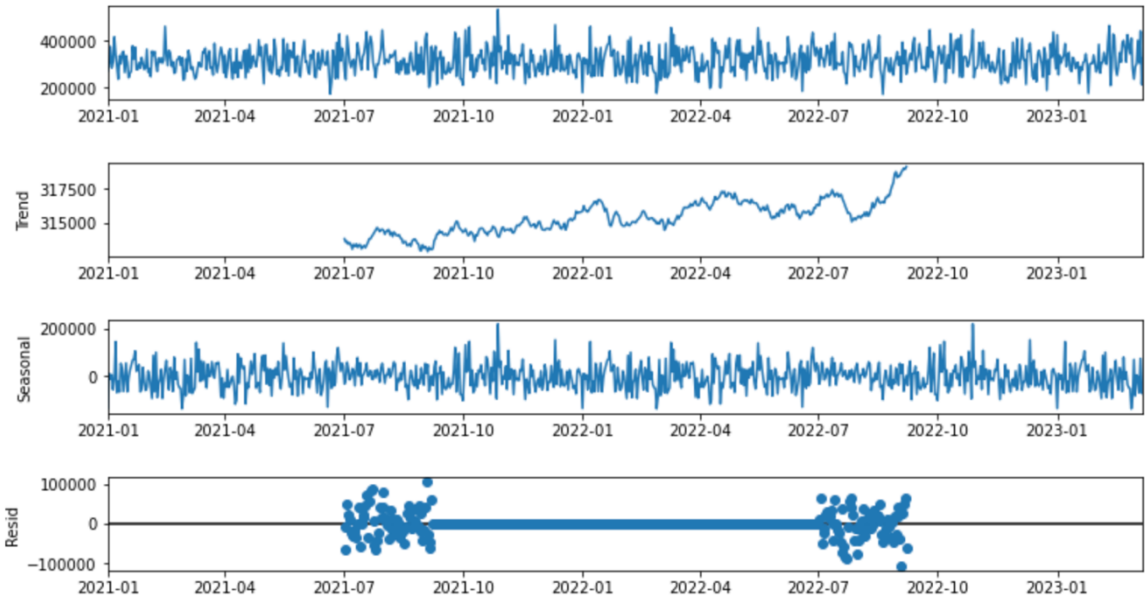


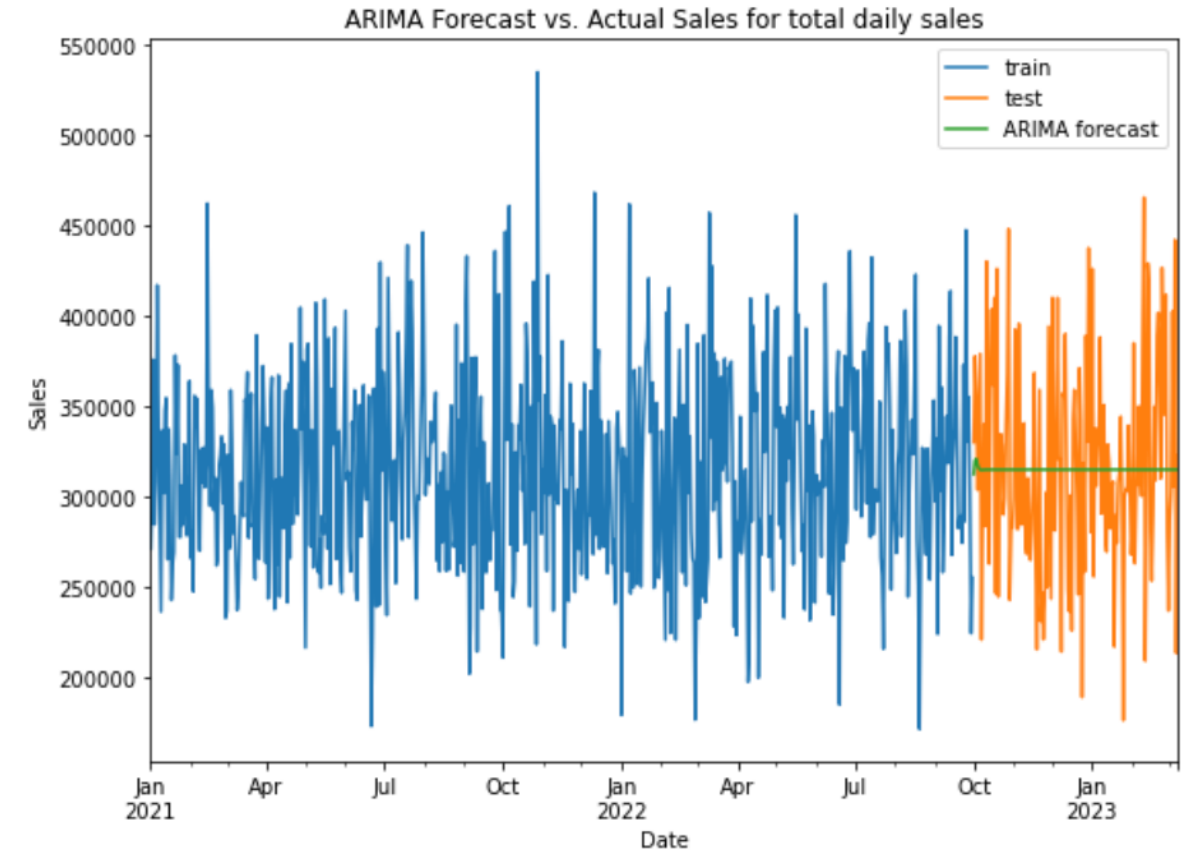
Figure 6: Decomposition plot for the daily sales data

ARIMA and SARIMA Prediction of Daily Sales

First of all, both ARIMA and SARIMA model does not provide good prediction. The forecasts generated by both models closely resemble the average line, indicating their limited effectiveness (Lool at Figure 7).

Additionally, SARIMA fails to outperform ARIMA with a little higher RMSE, 58473,3 compared to 58446.

This aligns with the dataset's absence of clear trend and seasonality, as identified during the exploratory data analysis (EDA).



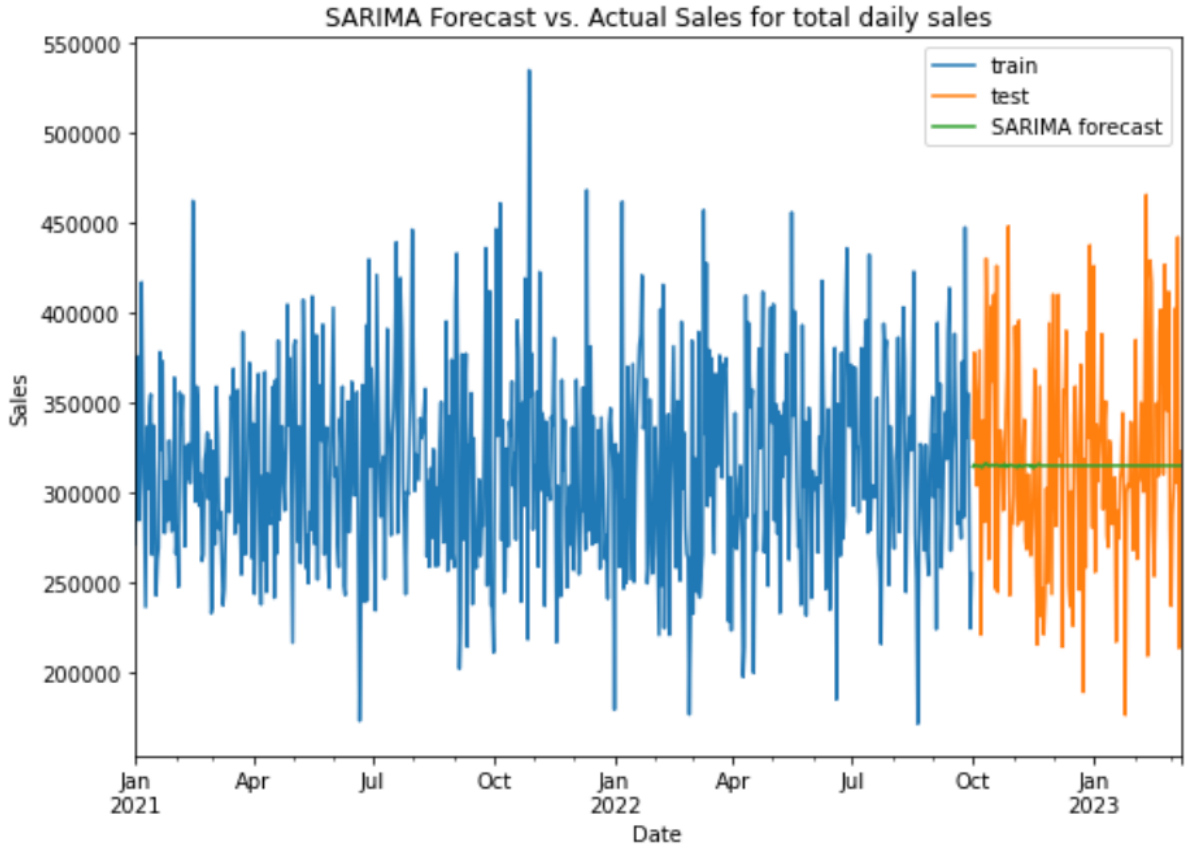


Figure 7: Forecast and actual total daily sales by ARIMA and SARIMA model

Basic XGBoost and Tuned XGBoost Prediction of Daily Sales

XGBoost (Extreme Gradient Boosting) is a machine learning algorithm that belongs to the gradient boosting family. It's designed for supervised learning tasks, both regression and classification. It works by combining the predictions from multiple weak learners (typically decision trees) into a strong ensemble model.

The 1st XGBoost model use random parameter and called basic XGBoost model. After conducting a parameter optimization process by running a loop with various parameters to identify the best configuration, we got the tuned XGBoost model. The prediction of total daily sales is visualized in Figure 8.

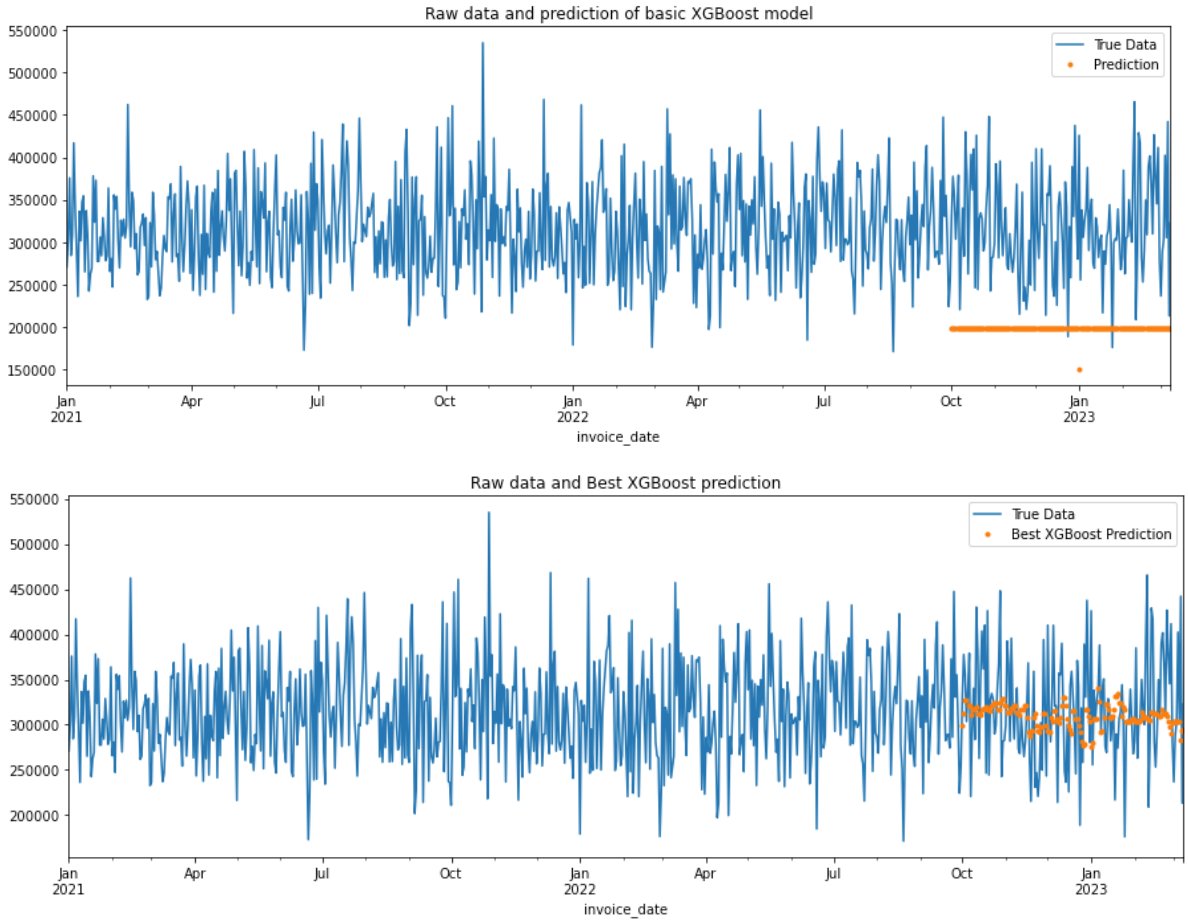


Figure 8: Forecast and actual total daily sales by XGBoost models

When visually inspecting the charts, the tuned XGBoost model appears to provide better fit predictions than ARIMA and SARIMA. However, not as I expected, in terms of RMSE, both the basic and tuned XGBoost models perform worse than ARIMA and SARIMA models.

Table 1: Root Mean Squared Error (RMSE) for each model used to predict the total daily sales.

ARIMA	SARIMA	XGBoost	Tuned XGBoost
58446.57	58473.28	132860.28	59913.35

Prediction of Daily Sales for each Shopping Mall

The figures belowed show the result of daily sale prediction for Mall of Istanbul, biggest mall, the biggest shopping mall in Istanbul.

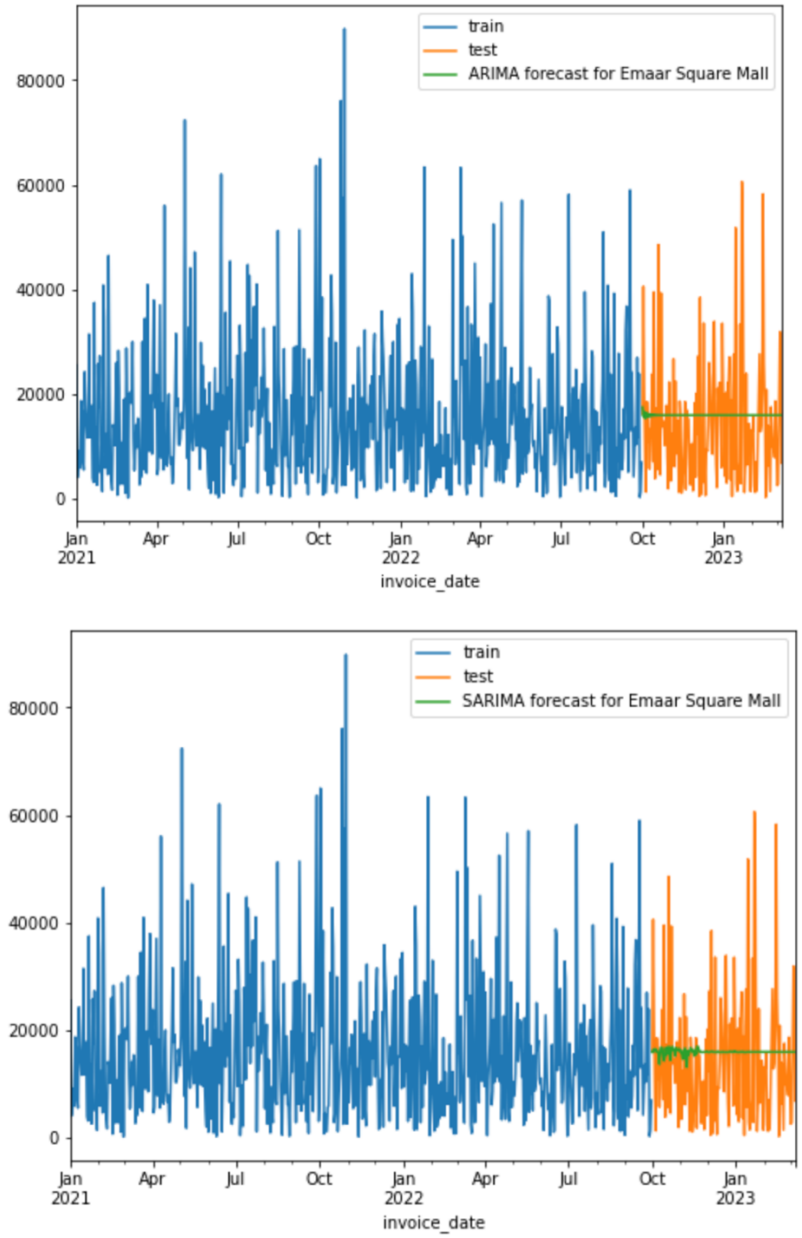


Figure 9: Forecast and actual daily sales for Mall of Istanbul by ARIMA and SARIMA model

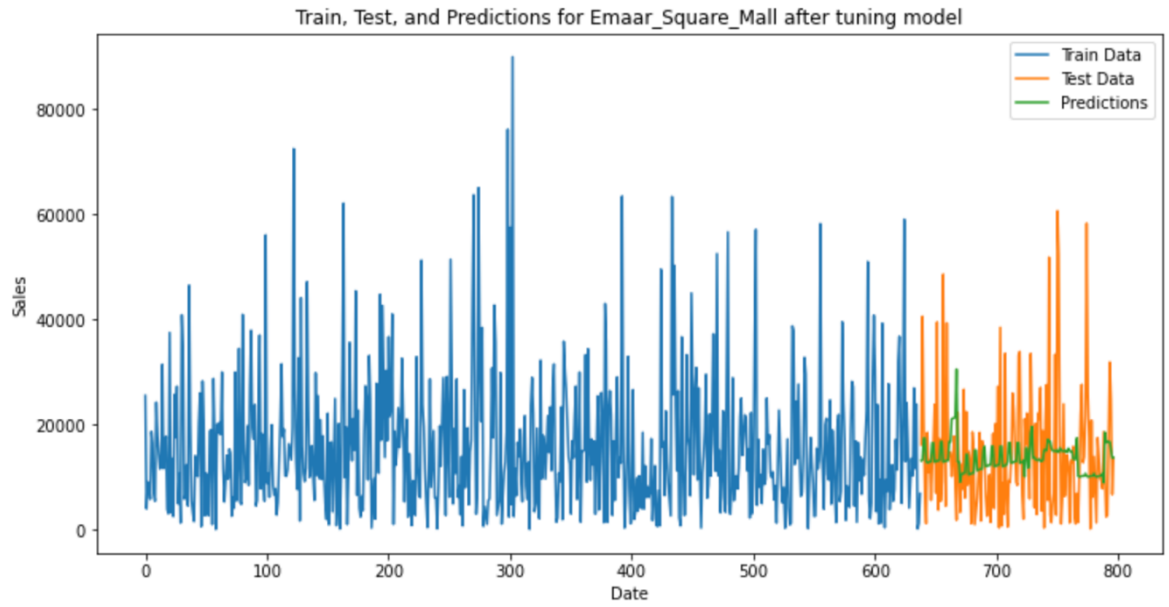


Figure 10: Forecast and actual daily sales for Mall of Istanbul by XGBoost models

When using ARIMA and SARIMA, the predictions of daily sale for Mall of Istanbul do not look better than the prediction of total dataset (compare Figure 7 and Figure 9). In general, the predictions for each shopping mall by ARIMA and SARIMA do not look better than the prediction of total dataset.
With XGBoost model, the prediction seem to be overfitted than the prediction of total dataset (compare Figure 8 and Figure 10).
However, when looking at the RMSE, XGBoost model prediction also does not perform better than ARIMA and SARIMA.

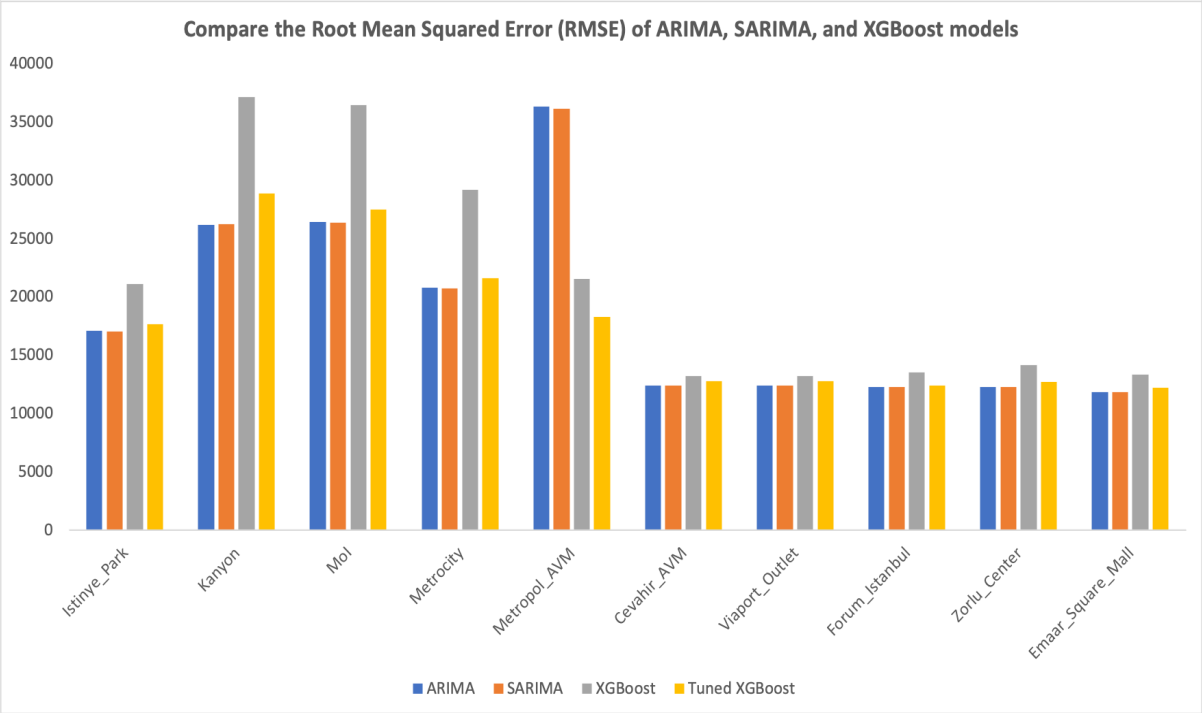


Figure 11: Comparison of RMSE among ARIMA, SARIMA, basic and Tuned XGBoost models for each shopping mall

After comparing the prediction of different models, I suggest using ARIMA for this current dataset. Several factors support this recommendation:

- Lowest RMSE: ARIMA achieves the lowest Root Mean Square Error (RMSE) among the models considered, indicating its superior predictive accuracy.
- Interpretability: ARIMA and SARIMA models are more interpretable as they rely on time series components such as trend and seasonality. In contrast, XGBoost is a more complex model, making it less straightforward to explain.
- Simple model: Since there are currently no external data sources available, ARIMA remains a practical and effective choice.

It's crucial to emphasize that XGBoost has the potential to deliver remarkable performance when combined with external data sources. Therefore, we strongly advise considering the collection of a more extensive time series dataset and the inclusion of exogenous features such as public holiday and lag features when implementing the XGBoost model in future applications.

Conclusions and Recommendations

The implementation of a targeted marketing strategy is essential to boost sales, particularly given the lack of significant growth from 2021 to 2023. Here are key recommendations:

- **Target Peak Times**: Focus on promoting sales during potential peak periods within the week or year, such as weekends or the year-end holiday season.
- **Gender Diversification**: While women constitute the primary customer base, it's crucial to attract male customers as well. Notably, specific male customer segments, including those aged 35-44 and over 65, exhibit notably higher average spending per transaction.
- **Product Emphasis**: Pay special attention to the top three products in terms of total sales: clothing, shoes, and technology. Additionally, consider the top three products based on the number of transactions: clothing, cosmetics, and food and beverage.
- **Improved Forecasting**: Enhance the accuracy of daily sales forecasts by exploring the possibility of gathering a more extensive time series dataset. Additionally, consider incorporating exogenous features like public holiday data and lag features when implementing the XGBoost model in future applications.

By adhering to these recommendations, the retail company can strategically revitalize its sales performance, ultimately fostering growth and success.