Customer-Centric Retail Analysis: Trends, Correlations, and Time Series Forecasts

Submitted by

Name	Reg No:
MUHAMMED AJMAL T	**223036**
MUHAMMED FARIS MUKTHAR M V	**223037**
NASREEN OT	**223038**
NIDIN V NANDAN	**223039**

In partial fulfillment of the requirements for the award of Master of Sciencein Computer Science with Specialization in Data Analytics

Of

School of Digital Sciences

Kerala University of Digital Sciences, Innovation, and Technology

(Digital University Kerala)

Technocity Campus, Thiruvananthapuram, Kerala – 695317

July 2023

BONAFIDE CERTIFICATE

This is to certify that the project report entitled **Customer-Centric Retail Analysis: Trends, Correlations, and Time Series Forecasts**
submitted by

Name	Reg No:
MUHAMMED AJMAL T	**223036**
MUHAMMED FARIS MUKTHAR M V	**223037**
NASREEN OT	**223038**
NIDIN V NANDAN	**223039**

In partial fulfilment of the requirements for the award of Master of Science in Computer Science with Specialization in Data Analytics is a Bonafide record of thework carried out at KERALA UNIVERSITY OF DIGITAL SCIENCES, INNOVATION AND TECHNOLOGY under our supervision.

Supervisor

Prof. MANOJ KUMAR TK

School Of Digital Sciences

DUK

Course Coordinator

Prof. MANOJ KUMAR TK

School of Digital Sciences

DUK

Head of Institution Prof.

SAJI GOPINATH

Vice Chancellor

DUK

DECLARATION

We,**MUHAMMED AJMAL T,MUHAMMED FARIS MUKTHAR M V,NASREEN OT, NIDIN V NANDAN**,students of Master of Science in Computer Science with Specialization in Data Analytics, hereby declare that this report is substantially the result of our own work, and has been carried out during the period March 2022-July 2022

Place: TRIVANDRUM

Date:08/09/2023

MUHAMMED AJMAL T

MUHAMMED FARIS MUKTHAR M V

NASREEN OT

NIDIN V NANDAN

ACKNOWLEDGEMENT

I would like to express my sincere and deepest gratitude to my guide Dr. T.K. Manoj Kumar, Associate Professor, Digital University Kerala, Trivandrum, for his valuable guidance, advice, and support, which enabled me to complete this project successfully.

I also like to express a deep sense of gratitude to Prof. Saji Gopinath for providing me with a good environment, valuable guidance, and educational facilities that enhanced my ability to undertake a project of this scale.

I would also like to utilise this opportunity to thank my friends, and my family for their valuable assistance, encouragement, and support during the execution of this project work.

ABSTRACT

In today's market, satisfying customer needs plays a crucial role in keeping customers happy and attracting them back to the shop. So there we need a proper analysis of the customer data in order to identify the various factors of the customer such as age group of the customer, for which category of product the demand is more, the method of payment made by the customer. All these things play a crucial role to understand the trends in the customer and if we can find out the trends in these we will be able to meet most of the needs by customers and hence the business will be having a high demand as well as growth.

So apart from these if we have a transaction history and if we are able to forecast the transaction possibility in the future dates it will play a crucial role in meeting the supply needs of those products. This is possible through the time series analysisof the data that we are having

Another key strategy to boost market performance is identifying the correlation that exists among the product categories in the market. By placing related products nearby, businesses can increase the likelihood of customers purchasing multiple items.

So in this project we are analysing the customer shopping dataset of Istanbul city inorder to identify these customer trends,finding correlation among different categories to find which all product can we sold together and forecasting the futuretransactions price using time series analysis to identify on which all days there is a greater chance of selling more products to meet the supply demands.

So through the analysis of the dataset if we are able to find out the above mentioned trends,correlation and forecast, then it can bring out a huge impact on the sales of products in the market as well satisfying the customer needs

CONTENTS

	Page No:
INTRODUCTION	07
DATASET DESCRIPTION	08
METHODOLOGY	10
RESULTS AND INSIGHTS	14
CONCLUSION	57
REFERENCES	58

INTRODUCTION

In today's highly competitive market, understanding and satisfying customer needs is paramount for the success of any business. A deep dive into customer data can unveil crucial insights that keep customers happy and attract them back to the shop.This report embarks on a journey through the bustling retail landscape of Malls in Istanbul city, delving into the intricacies of customer behaviour, product demand, payment methods, and the art of forecasting future transactions.

In an era where data is the new currency, harnessing the power of information is not just an advantage but a necessity. To this end, we explore various facets of customer data, including the age group of customers, the categories of products with the highest demand,and the preferred payment methods. By dissecting these components, we aim to unravel the underlying trends in customer behaviour. Armed with this knowledge, businesses can tailor their strategies better to meet their clientele's diverse and evolving needs, fostering customer satisfaction and business growth.

Moreover, we delve into the realm of predictive analytics, particularly time series analysis, to forecast transaction possibilities in the future. The ability to anticipate demand on specific days empowers businesses to optimise their supply chain management, ensuring that products are readily available when and where they are most likely to be sold.

Another strategic focus of this report is the identification of product category correlationswithin the market. By understanding which products are frequently purchased together, businesses can strategically position related items nearby, increasing the likelihood of customers making multiple purchases. This not only enhances the shopping experience but also maximises revenue.

In general, this report serves as a comprehensive exploration of the Istanbul city customershopping dataset. Through rigorous analysis, we aim to uncover essential customer trends, discover valuable product category correlations, and harness the power of time series forecasting to align supply with demand. The ultimate goal is to revolutionise market performance by meeting and exceeding customer expectations, resulting in increased sales and customer satisfaction. Join us in this insightful journey as we unveil the keys to unlocking success in the bustling retail landscape of Istanbul.

DATASET DESCRIPTION

Data Collection:

The Istanbul shopping dataset presents a comprehensive shopping dataset collected from 10 different shopping malls from 2021 to 2023. This dataset is a valuable resource for researchers, data analysts, and machine learning enthusiasts interested in understanding shopping trends in Istanbul. It includes essential information such as unique invoice numbers, customer IDs, age, gender, payment methods, product categories, quantity, price, order dates, and shopping mall locations. The dataset offers a diverse perspective on shopping habits, sourced from various age groups and genders, and covers transactions from different malls. Data collection methods involved point-of-sale systems, surveys, and demographic profiling to ensure completeness and accuracy.

The malls in the istanbul city used for collecting these data are:

Kanyon Mall,Forum Istanbul Mall,Metrocity Mall,Metropol AVM Mall,Istinye Park Mall,Mall of Istanbul,Emaar Square Mall,Cevahir AVM Mall,Viaport Outlet,Zorlu CenterMall

Attribute Information:

- invoice_no - Invoice number of the customer,combination of 'I' and a 6-digitinteger specific to each operation.
- customer_id - An id number which is a combination of ‘C’ and a 6 digit integerthat is specific to each operation
- gender - Male and Female string variables
- age - Integer variable representing customers age
- category - A string variable representing the category of the product purchased
- quantity - numeric value representing the quantity of each product purchased

- price - price of product per unit in Turkish Liras(TL)
- payment_method - A string variable representing the method of payment used for the transaction
- invoice_date – Represents date on which the transaction was done
- shopping_mall- A string variable representing the name of the mall in the Istanbul city

The dataset consist of a total of 99457 rows

The sample of the dataset that is used for the analysis is given below:

	invoice_no	customer_id	gender	age	category	quantity	price	payment_method	invoice_date	shopping_mall
0	I138884	C241288	Female	28	Clothing	5	1500.40	Credit Card	2022-08-05	Kanyon
1	I317333	C111565	Male	21	Shoes	3	1800.51	Debit Card	2021-12-12	Forum Istanbul
2	I127801	C266599	Male	20	Clothing	1	300.08	Cash	2021-11-09	Metrocity
3	I173702	C988172	Female	66	Shoes	5	3000.85	Credit Card	2021-05-16	Metropol AVM
4	I337046	C189076	Female	53	Books	4	60.60	Cash	2021-10-24	Kanyon
...	...	...	...	...	...	...	...	...	...	...
99452	I219422	C441542	Female	45	Souvenir	5	58.65	Credit Card	2022-09-21	Kanyon
99453	I325143	C569580	Male	27	Food & Beverage	2	10.46	Cash	2021-09-22	Forum Istanbul
99454	I824010	C103292	Male	63	Food & Beverage	2	10.46	Debit Card	2021-03-28	Metrocity
99455	I702964	C800631	Male	56	Technology	4	4200.00	Cash	2021-03-16	Istinye Park
99456	I232867	C273973	Female	36	Souvenir	3	35.19	Credit Card	2022-10-15	Mall of Istanbul

99457 rows × 14 columns

METHODOLOGY

Feature Engineering:

A critical phase in the data preprocessing pipeline is feature engineering, which aims to improve the accuracy and usefulness of the features utilised in our study. To more accurately depict the underlying patterns in the data, this procedure includes generating new features, transforming existing ones, and extracting relevant information from the dataset.

For example, here in the dataset, we are creating a new column named “Total_Payment” by multiplying the columns “quantity” and the column “price”

Outlier Detection:

We used the well-known box plotting method to find potential outliers in the dataset. Box plots, often referred to as box-and-whisker plots, provide a visual depiction of the central tendency and spread of the data, making them a useful tool for identifying data points thatsignificantly vary from the distribution as a whole. Data points outside the box plot's "whiskers" were categorised as outliers. Following the standard definition of outliers in box plots, we specifically regarded data points located below the lower bound (Q1 - 1.5 * IQR) or above the upper bound (Q3 +1.5 * IQR) as potential outliers. To make it easier to recognise detected outliers, theywere graphically highlighted on the box plots. We were able to evaluate the size and effect of outliers on the data distribution because of this visual inspection.

Outlier Removal:

Capping Method :

Capping outlier removal, also referred to as "trimming," is a data preprocessing technique employed to mitigate the impact of extreme data points (outliers) on statistical analyses, visualisations, and machine learning models.With this technique, a predetermined threshold is established, usually based on the Interquartile Range (IQR), and outlier values that fall outside of the threshold are either capped or replaced with the threshold values themselves. Identifying possible outliers in the dataset was the first step in the outlier reduction procedure.The difference between the data's third quartile (Q3) and first quartile (Q1) was used to generate the interquartile range (IQR). This metric represents the middle 50% of the data's distribution.

The upper threshold for identifying potential outliers was established as Q3 plus 1.5 times the IQR. The lower threshold for identifying potential outliers was established as Q1 minus 1.5 times the IQR. Any data point exceeding both these thresholds was considered as a potential outlier. Outliers identified were treated to capping, which replaced their values with the upper threshold value (up_lim). This procedure ensures that the remaining data points are preserved while bringing extreme values inside a certain boundary.

Grouped Bar Chart:

In our study, we used a grouped bar chart as a visual tool to investigate and highlight interactions between important variables in our dataset. Using this charting technique, we could compare and contrast results across many categories while preserving a strong visual distinction.

Pearson Correlation Coefficient:

Our study employed the Pearson correlation coefficient as a fundamental statistical measure to assess the strength and direction of linear relationships between pairs of continuous variables within our dataset. This method allowed us to quantify the degree towhich variables move together or in opposite directions linearly.

Using Python's data analysis libraries, we calculated the Pearson correlation coefficient for all pairs of selected variables. This coefficient measures the linear association between two variables and ranges from -1 to 1.

- A coefficient of 1 indicates a perfect positive linear relationship, where an increasein one variable corresponds to an exact linear increase in the other.
- A coefficient of 0 suggests no linear relationship; the variables are not correlated.
- A coefficient of -1 indicates a perfect negative linear relationship, where an increase in one variable corresponds to an exact linear decrease in the other.

We interpreted the resulting correlation coefficients to discern the nature of the relationships between variables. Positive coefficients indicated positive linear correlations, while negative coefficients indicated negative linear correlations. Coefficients close to 0 suggested weak or no linear association.

Pie-Chart:

A pie chart is a circular graphic which demonstrate data to show how various parts or groups within a whole are distributed in terms of relative proportions. It is a popular form of chart that displays the division of a single data set into smaller portions and the appropriate percentages with regard to the entire.

Line Plot:

A line plot, represents data points as a series of individual points connected by straight lines. It is a fundamental and widely used method for visualising and analysing data trends and changes over a continuous interval or time period.

Correlation Matrix Heatmap:

In our analysis, we employed a correlation heatmap as a powerful visualisation tool to explore and visualise the relationships between multiple variables within our dataset. This heat map provided an intuitive and comprehensive overview of the pairwise correlations between variables.We utilised Python's seaborn library to create the correlation heatmap. The *sns.heatmap* function was applied to the correlation matrix, where each cell represented the correlation coefficient between a pair of variables. We incorporated colour mapping to convey the strength and direction of the correlations.

ARIMA model:

A type of linear model called an ARIMA model uses past data to predict future data.The term "Auto Regressive Integrated Moving Average" describes a family of models that predict future values by using an equation that explains a particular time series based on its own historical values, or, more specifically, on its own lags and lagged prediction errors.

Auto Regression: A model that incorporates the dependence between being an observation and a predetermined amount of lag observations is known as auto regression. Integrated: The practice of subtracting a raw measurement from an observation taken at a previous time step in order to stabilise the time series.

Moving Average: A model which utilises the correlation between such a lagged observable and a moving average model's residual error.

Where, p is the order of the AR term here we are using p=4

q is the order of the MA term here we are using q=1

d is the number of differencing required here we are using d=3

Augmented Dickey-Fuller(ADF) Test:

A time series is deemed stationary if it displays zero trend, constant variance, and constant autocorrelation throughout time. To evaluate whether a time series is stationary or otherwise, one may do an enhanced Dickey-Fuller test that uses the null and alternative hypotheses provided below.

Null Hypothesis, H0: The time series is non-stationary.

Alternate Hypothesis, HA: The time series is stationary.

If the test's p-value is less than a predetermined significance threshold (for example, α 0.01), we may dismiss the null hypothesis and conclude that the time series is stationary.

RESULTS AND INSIGHTS

Exploratory Data Analysis (EDA)

In this analysis, we conducted Exploratory Data Analysis (EDA) to uncover the dataset's structure, shape, patterns, and the relationships between its features.

- Analysing the dataset, we can see that the dataset is free of null values
- Using the info function from the pandas library, we can see the informationabout the data frame.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 99457 entries, 0 to 99456
Data columns (total 10 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   invoice_no      99457 non-null  object
 1   customer_id     99457 non-null  object
 2   gender          99457 non-null  object
 3   age             99457 non-null  int64
 4   category        99457 non-null  object
 5   quantity        99457 non-null  int64
 6   price           99457 non-null  float64
 7   payment_method  99457 non-null  object
 8   invoice_date    99457 non-null  object
 9   shopping_mall   99457 non-null  object
dtypes: float64(1), int64(2), object(7)
```

Here, we can see that other than columns 'age',' quantity', and' price' all other columns belong to the object data type

Here, the invoice date is in object data type, which won't be suitable for our analysis as it is a date we have to convert to datetime format

Code snippet to convert to datetime format:

" df['invoice_date'] = pd.to_datetime(df['invoice_date'],dayfirst=True) "

Now, checking the column's data types, we can see that the invoive_date column is in datetime format

```
invoice_no                 object
customer_id                object
gender                     object
age                         int64
category                   object
quantity                    int64
price                     float64
payment_method             object
invoice_date       datetime64[ns]
shopping_mall              object
```

Feature Engineering in the dataset

- For extracting trends out of the dataset, we are creating three new columns called **"Day" , "Month ", " year"** Out of the invoice_date column
- Creating a new column called "**Age_Range**" from the "age" column to have the age ranges between 10-20,20-30, etc., to 50-70 to understand the trends for each age group
- Analysing the dataset, we can see that theprice is given for a single item, and there is no total payment column, so using the feature engineering, we are going to create a new column named **"Total_Payment"** by multiplying the column quantity with the column price.

So the new dataset with the added new columns is:

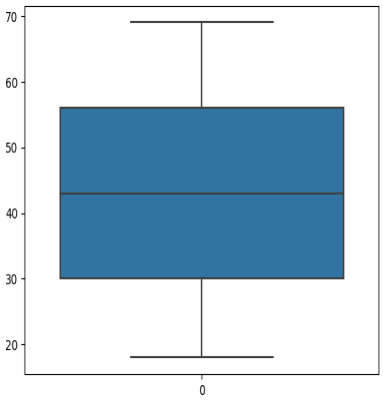
	invoice_no	customer_id	gender	age	category	quantity	price	payment_method	invoice_date	shopping_mall	Day	Month	Year	Age_Range	Total_Payment
0	I138884	C241288	Female	28	Clothing	5	1500.40	Credit Card	2022-08-05	Kanyon	Friday	August	2022	21-30	7502.00
1	I317333	C111565	Male	21	Shoes	3	1800.51	Debit Card	2021-12-12	Forum Istanbul	Sunday	December	2021	21-30	5401.53
2	I127801	C266599	Male	20	Clothing	1	300.08	Cash	2021-11-09	Metrocity	Tuesday	November	2021	10-20	300.08
3	I173702	C988172	Female	66	Shoes	5	3000.85	Credit Card	2021-05-16	Metropol AVM	Sunday	May	2021	51-70	15004.25
4	I337046	C189076	Female	53	Books	4	60.60	Cash	2021-10-24	Kanyon	Sunday	October	2021	51-70	242.40
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
99452	I219422	C441542	Female	45	Souvenir	5	58.65	Credit Card	2022-09-21	Kanyon	Wednesday	September	2022	31-50	293.25
99453	I325143	C569580	Male	27	Food & Beverage	2	10.46	Cash	2021-09-22	Forum Istanbul	Wednesday	September	2021	21-30	20.92
99454	I824010	C103292	Male	63	Food & Beverage	2	10.46	Debit Card	2021-03-28	Metrocity	Sunday	March	2021	51-70	20.92
99455	I702964	C800631	Male	56	Technology	4	4200.00	Cash	2021-03-16	Istinye Park	Tuesday	March	2021	51-70	16800.00
99456	I232867	C273973	Female	36	Souvenir	3	35.19	Credit Card	2022-10-15	Mall of Istanbul	Saturday	October	2022	31-50	105.57

99457 rows × 15 columns

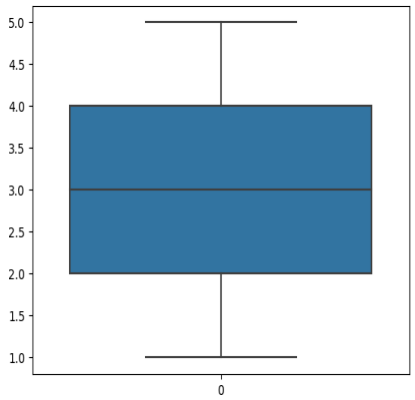
Checking for Outliers:

The method used to check outliers is visualising the columns using *sns.barplot*.

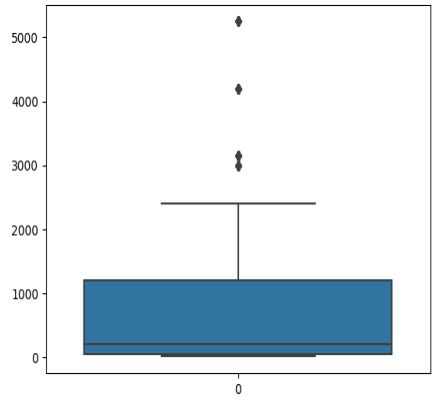
The results obtained are:



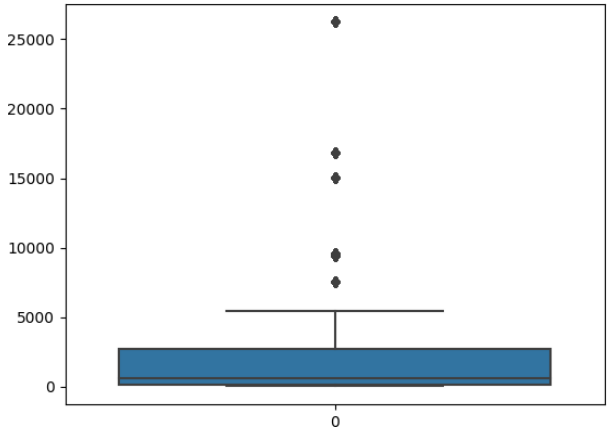
Column "age"



Column "quantity"



Column "price"



Column "Total_Payment"

The above bar plot shows that outliers were present in the 'price' and 'Total_Payment' columns. We also can understand that the outliers are above the upper limit, so we must replace the values with the upper threshold value. So, to remove the outliers present in it, we have to apply the capping or outlierhandling method using the Interquartile Range (IQR) method.

The code snippet for doing that is:

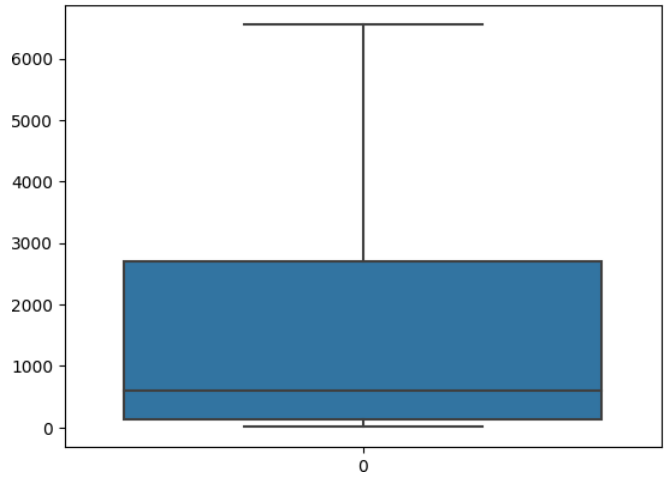
For 'Total_Payment' column

```
Q1=np.percentile(df["price"],25,interpolation="midpoint")
Q2=np.percentile(df["price"],50,interpolation="midpoint")
Q3=np.percentile(df["price"],75,interpolation="midpoint")
IQR=Q3-Q1
low_lim=Q1-1.5*IQR
up_lim=Q3+1.5*IQR


index1 = (df['price']>up_lim)
index1 = df.loc[index1].index
print(index1)
df.loc[list(index1),'price']=up_lim

sns.boxplot(df['price'])
```

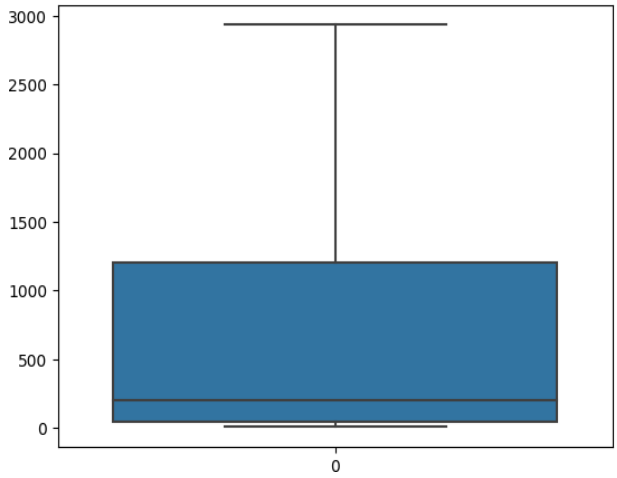
For 'price' column

```
Q1=np.percentile(df["price"],25,interpolation="midpoint")
Q2=np.percentile(df["price"],50,interpolation="midpoint")
Q3=np.percentile(df["price"],75,interpolation="midpoint")
IQR=Q3-Q1
low_lim=Q1-1.5*IQR
up_lim=Q3+1.5*IQR


index1 = (df['price']>up_lim)
index1 = df.loc[index1].index
print(index1)
df.loc[list(index1),'price']=up_lim

sns.boxplot(df['price'])
```

Result of running the above code snippet:



Bar plot for column "Total_Payment"



Bar plot for column "price"

So after running the outlier removal method and plotting the new bar plot it is clearly understood that the outliers has been removed.

Lets dive into the dataset to understand the unique things present in the columns

- In the “category” column the unique strings present are:

  'Clothing', 'Shoes', 'Books', 'Cosmetics', 'Food & Beverage','Toys', 'Technology', 'Souvenir'

  (A total of 8 items)

  So these are the categories of products that are being selling in the malls in Istanbul city

- In the “payment_method” column the unique string present are:

  'Credit Card', 'Debit Card', 'Cash'

  (A total of 3 items)

  These are the payment methods that is being used for the mode of transaction in the malls

- In the “shopping_mall” column the unique strings present are:

  'Kanyon', 'Forum Istanbul', 'Metrocity', 'Metropol AVM', 'Istinye Park','Mall of Istanbul', 'Emaar Square Mall', 'Cevahir AVM', 'Viaport Outlet', 'Zorlu Center'

  (A total of 10 items)

  These 10 malls are considered for the study

Value count for each item in category column of the dataset:

Clothing	34487
Cosmetics	15097
Food & Beverage	14776
Toys	10087
Shoes	10034
Souvenir	4999
Technology	4996
Books	4981

The value counts for each item category in the dataset provide valuable insights into the distribution and popularity of products across different categories.

- The "Clothing" category stands out as the most prevalent category, with a substantial count of 34,487. This suggests that clothing items are frequently purchased and are likely to be a major focus for retailers.
- "Cosmetics" and "Food & Beverage" are also popular categories, with 15,097 and14,776 counts, respectively. This suggests that consumers place importance on personal care and food-related products.
- "Souvenir," "Technology," and "Books" have relatively lower counts compared tothe top categories. This could indicate that these categories are less frequently purchased or are more specialised.

So from the above insights we can conclude that there is a more chance for success if a retailer is investing in the clothing category as it is the most popular product that is having more sail across the different malls in istanbul

The presence of diverse categories suggests that shoppers have a range of interests and preferences. Retailers can use this information to tailor their product offerings and marketing efforts to specific customer segments.

Values Count for each item in shopping_mall column of the dataset :

Mall of Istanbul	19943
Kanyon	19823
Metrocity	15011
Metropol AVM	10161
Istinye Park	9781
Zorlu Center	5075
Cevahir AVM	4991
Forum Istanbul	4947
Viaport Outlet	4914
Emaar Square Mall	4811

The value counts for each shopping mall in the dataset provide insights into the popularity and foot traffic of different shopping destinations in Istanbul. Here are some insights that can be drawn from the result:

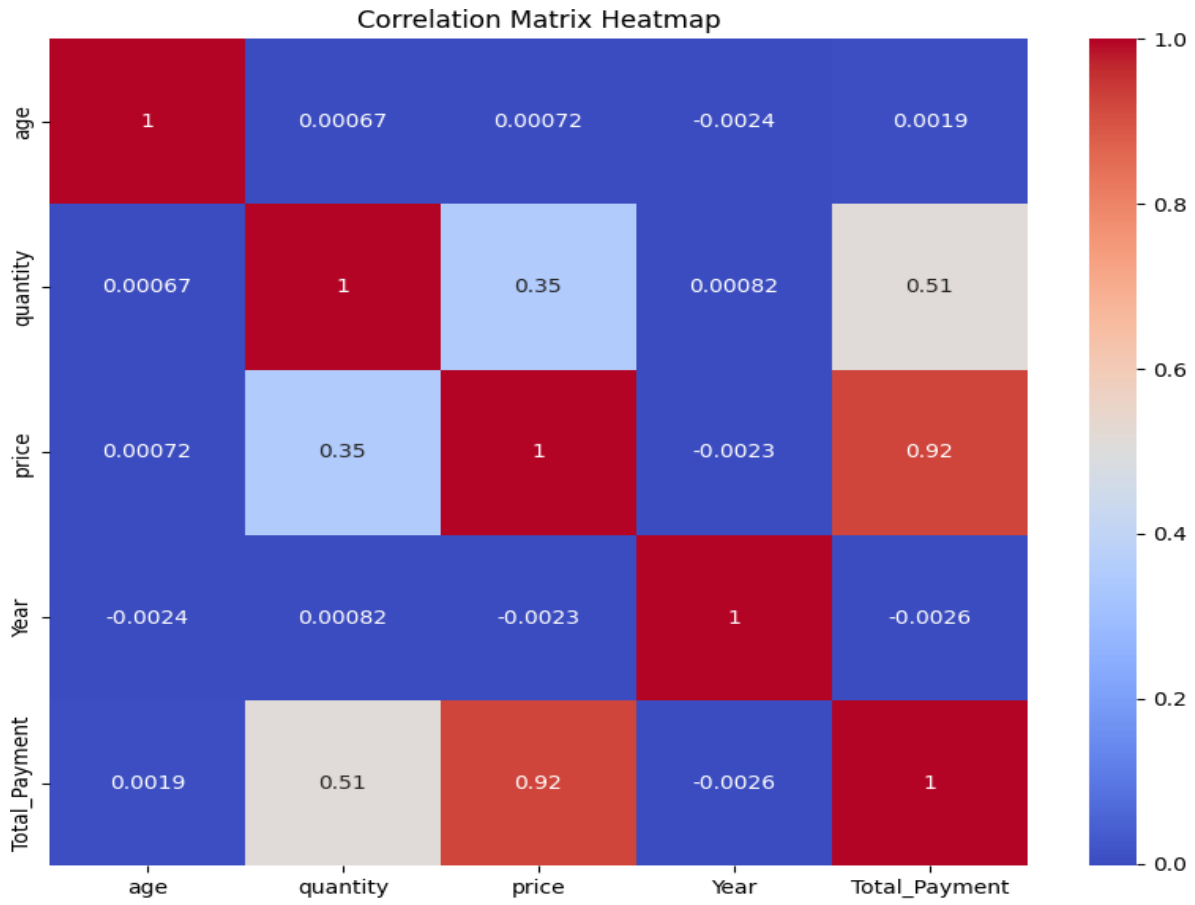
- "Mall of Istanbul" and "Kanyon" stand out as the two most popular shopping malls, with similar high visitation counts of approximately 19,943 and 19,823,respectively. These malls are likely major shopping hubs in Istanbul.
- "Metrocity" and "Metropol AVM" also have substantial visitation counts, with15,011 and 10,161, respectively. These malls attract a significant number of shoppers and are important retail destinations.
- "Istinye Park" follows closely behind with 9,781 visits. It's another growing mall, indicating its popularity among shoppers.
- Zorlu Center, Cevahir AVM, Forum Istanbul, Viaport Outlet, Emaar Square MallThese malls have lower visitation counts compared to the top performers butstill attract a notable number of shoppers. The visitation counts for these malls range from 4,811 to 5,075.

So, from the above insights, we can conclude that the presence of multiple malls with varying visitation counts suggests that Istanbul offers diverse shopping choices to its residents and visitors. Shoppers have access to a range of retail experiences and brand offerings.
Retailers may consider the popularity of malls when deciding on the location of their stores and the malls where they wish to have a presence."Mall of Istanbul" and "Kanyon" will be a better choice as these malls have the most visits by the customer

VISUALISATIONS

Correlation among different variables:

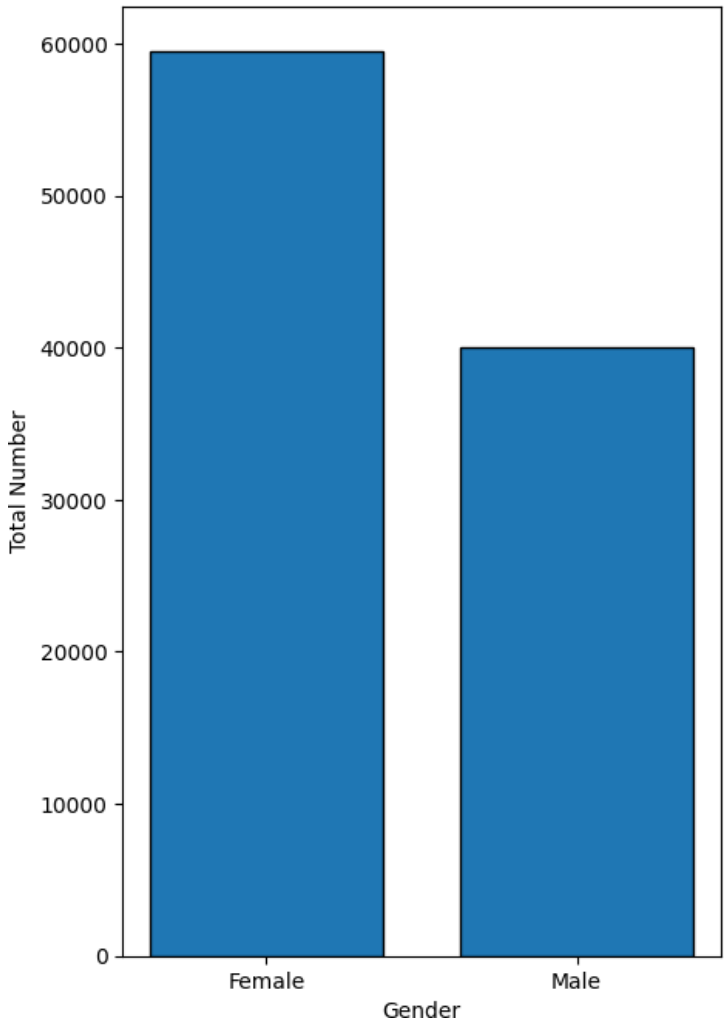
The correlation among existing different variables can be found by plotting the heatmap:



In the above figure the correlation between price and total payment is 0.92 ,which indicates the fact that as the price increases as it is multiplied by the quantity the Total_Payment will increase. Also the correlation between Total_Payment and quantity is 0.51 even though it doesn't show a much correlation there is a slight correlation indicating that as the quantity increases the total payment will also increase. Most of the other variables are independent of each other as there is less correlation between them.

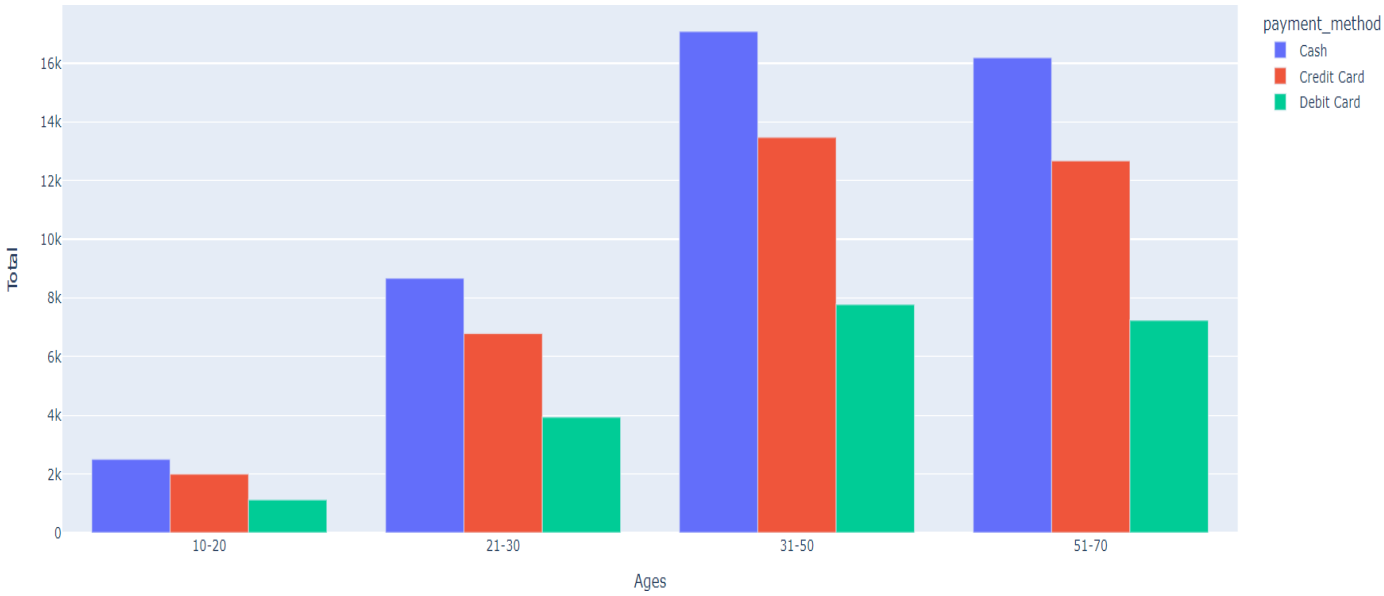
Counting the occurrences of genders based on their purchasing behaviour

Based on the purchase information, a barplot was generated to represent the "gender" column. The barplot provides a visual representation of the total number of purchases made by males and females. In a dataset containing a total of 99,457 purchases, it's evident that the majority of purchases are attributed to females, accounting for 59.80% ofthe total. In contrast, the number of purchases made by males is slightly lower at 40.10%, approximately 20% less than that of females.



This information highlights the insight that there is a greater chance for the sale of the products for females, and Designing marketing strategies that specifically target female consumers can be effective. Tailored advertising, promotions, and messaging that resonate with female preferences and interests can help attract and retain this significant customer segment.

Bar graph to visualise the usage of payment methods among different age groups:

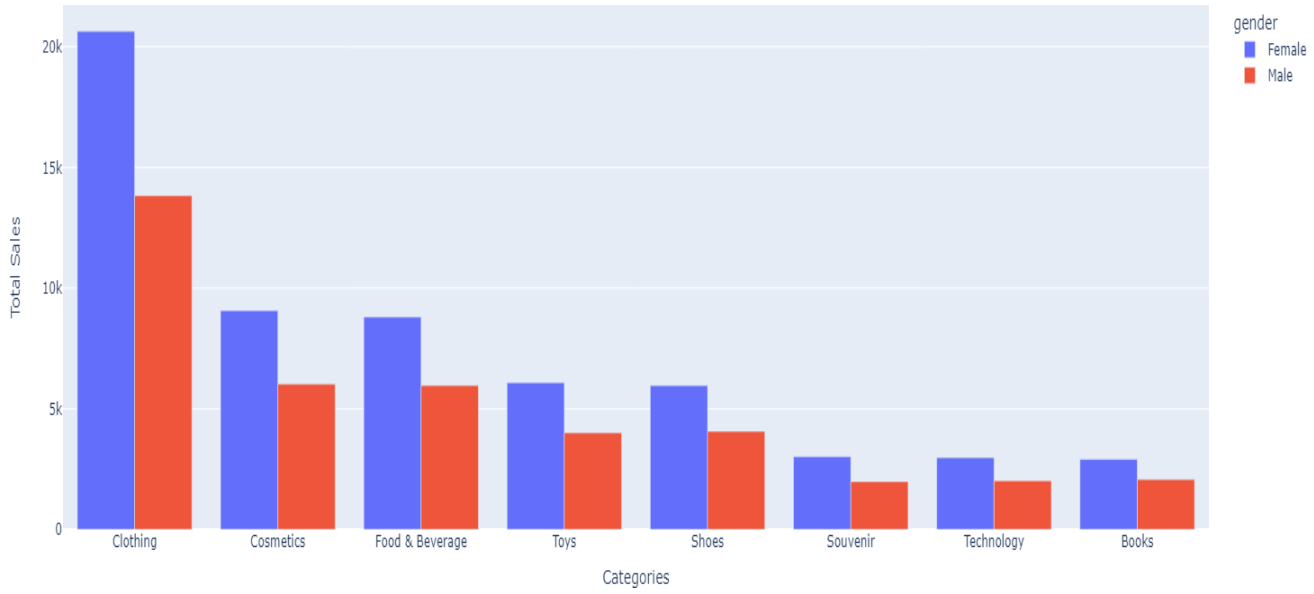


Looking at the provided bar graph, it's evident that we can easily determine the most effective and commonly used payment method within specific age groups. **Debit Card usage is relatively low across all age ranges**. The most frequently employed methods are Cash and Credit Cards. **Age groups between 31 and 70 tend to make more purchases**.

So, we can conclude that

- Cash and Credit Cards are the most frequently employed payment methods across all age groups. This suggests that these two payment methods are the preferred choices for a wide range of consumers. Businesses may need to ensure that they can accommodate both cash and credit card payments effectively.
- The bar graph reveals age-related patterns in payment method usage. Specifically, age groups between 31 and 70 tend to make more purchases. This insight is crucial for marketers and businesses, as it suggests that targeting and tailoring marketing efforts toward these age groups may yield higher sales and engagement.
- Businesses can use these insights to inform their marketing strategies and payment processing options. For instance, offering multiple payment methods, promoting credit card rewards, or implementing contactless payment solutions may alignwith customer preferences and enhance the shopping experience.

Category-wise Total sales for gender :



The above bar chart, which displays total sales by gender and category, offers valuable insights into purchasing behaviour based on gender and product category.

Observing this graph, as previously noted, females account for the majority of purchases, with the "clothing" category seeing the highest purchase activity. On the other hand, the categories of "Books" and "Technology" show the least or minimal purchase levels.

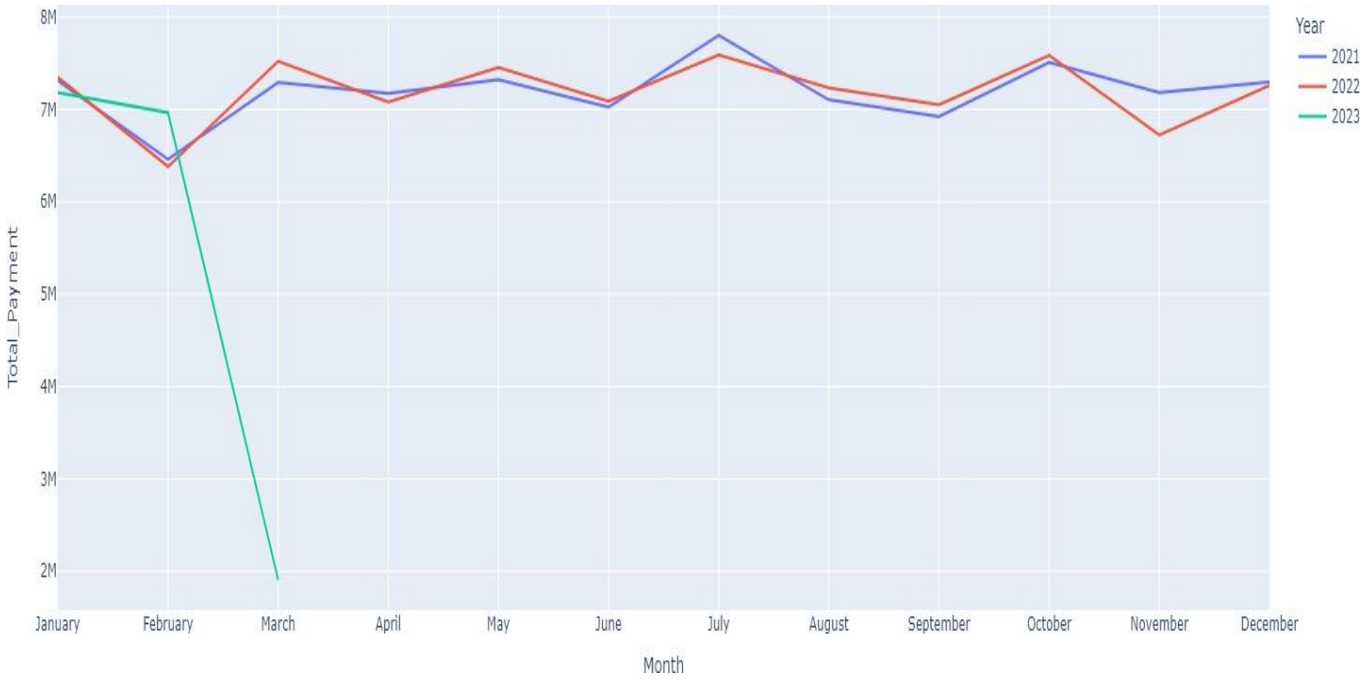
The general insights that can be derived from the above figure are:

- The bar chart vividly illustrates the product category preferences of different genders. For example, within the "Clothing" category, females have a notably higher total sales count (20,652) compared to males (13,835). This suggests that clothing is a more popular choice among female shoppers in the dataset.
- "Cosmetics" and "Food & Beverage" also show higher total sales among females compared to males. This implies that these categories tend to attract more female consumers.
- In contrast, categories like "Toys" and "Shoes" show a more balanced distribution between genders. While females still contribute to a significant portion of sales in these categories, males also engage in substantial purchases.
- Categories like "Technology" and "Books" exhibit a more even distribution, with relatively comparable total sales between genders.

- Interestingly, the "Souvenir" category also demonstrates a relatively even distribution, suggesting that both genders show interest in souvenir purchases.
- This data highlights the importance of customer segmentation based on gender. Understanding the differing preferences of male and female customers can enable businesses to personalise their offerings and enhance customer satisfaction.
- Businesses may consider offering promotions or discounts on products that align with the preferences of their target gender demographic. This can lead to increased sales and customer loyalty.

So, in general, clothing is the category with the most demand in the malls being run in Istanbul city. And the female category is making most of the purchases in the malls. So businesses can spend more money on the products that are more likely to be purchased by the female category to make more profit.

Line plot displaying total payment amount over months for differentyears

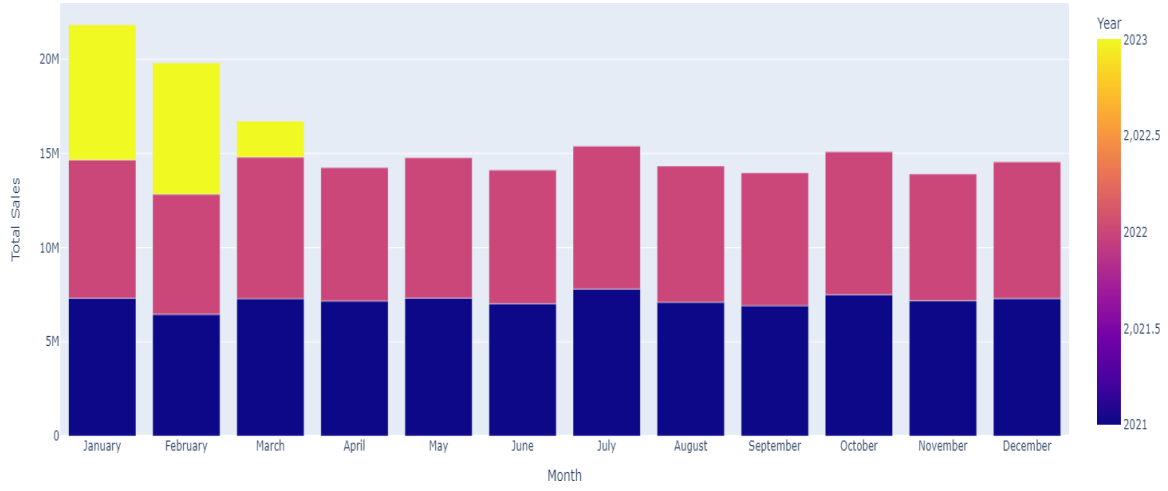


The provided line chart displays the total payment amount over months for different years. By analysing this, we can understand in which all months the sales are gonna be high and in which all month there will be a decrease in the business

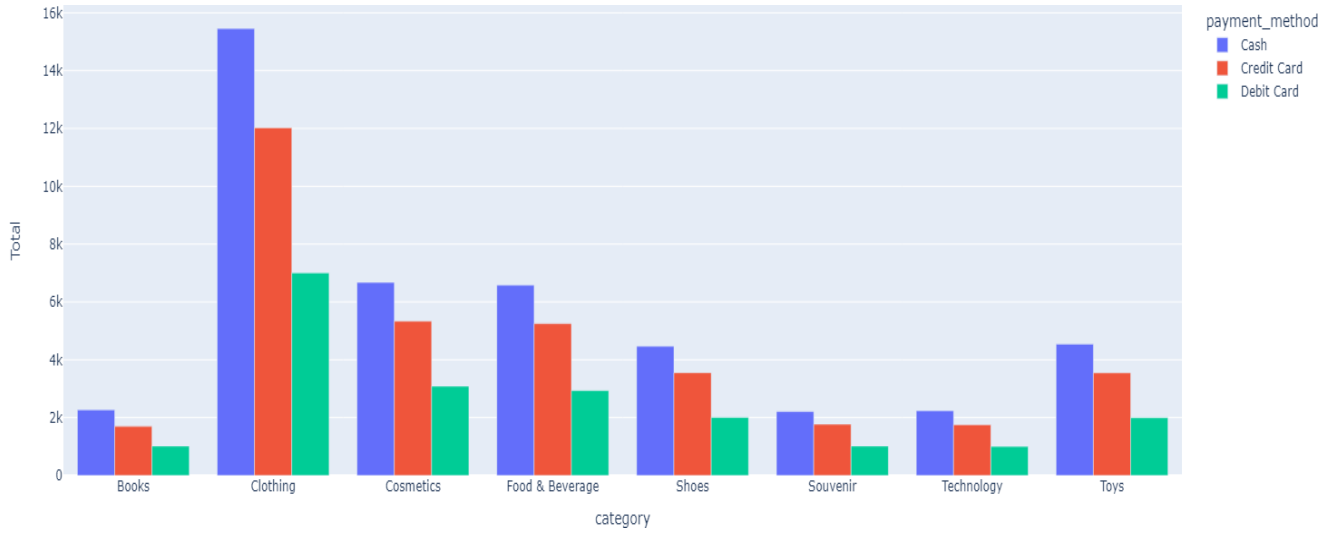
Here, the dataset is available only up to 8/03/2023, that is why there is a dip in sales in March
So the insights that we get from this are:

- A recurring pattern in total payments follows a yearly cycle. Total payments tend to peak and dip around the same months each year. This suggests the presence of seasonal factors influencing payment behaviour.
- The highest total payments occur in the months of January and July. This may be because of the new year and some other special festival that is happening in Istanbul
- Despite fluctuations, there appears to be a general trend of consistent or growing total payments over the years. This suggests that the business or market is experiencing overall growth.
- In the month of february there is a dip in the sale and this pattern is following inall the 3 years 21,22, and 23 so we can conclude that this month is a off season for the mall.
- But after the dip in the february month there comes a peak in the march month so this month is very important for the retailers as they have to arrange the stock and all those things i order to meet the needs of the customer
- After march most of the months have a constant scale of sale only a peak is shownin the month of july
- Also we can see that there is a sale dip in the month of november, it may be because of the coming christmas in the next month

The above visualisation can also be represented in grouped bar chart to convey the idea more clearly , This also convey the same insights from above



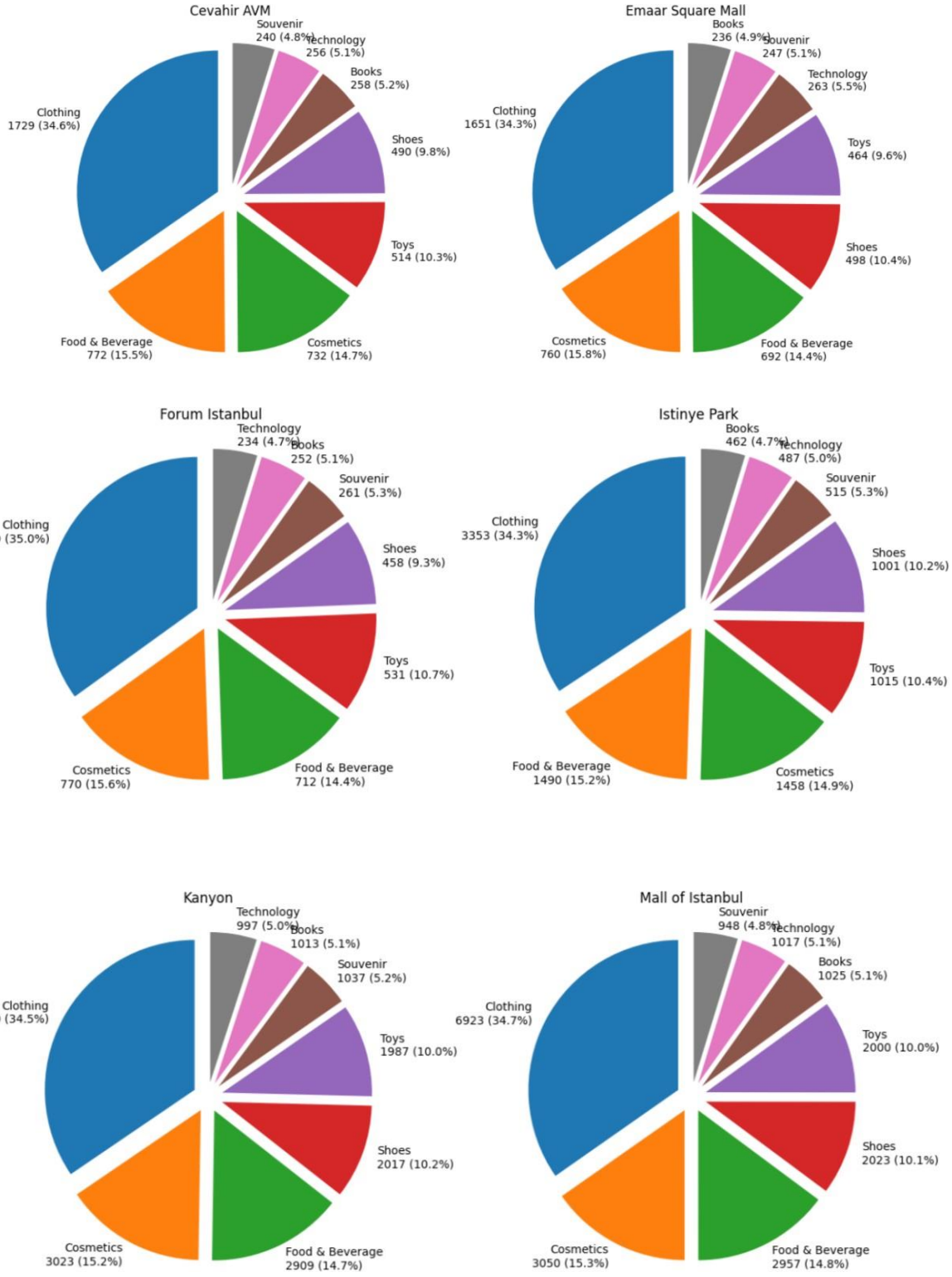
Grouped bar chart representing the distribution of payment methods within different product categories

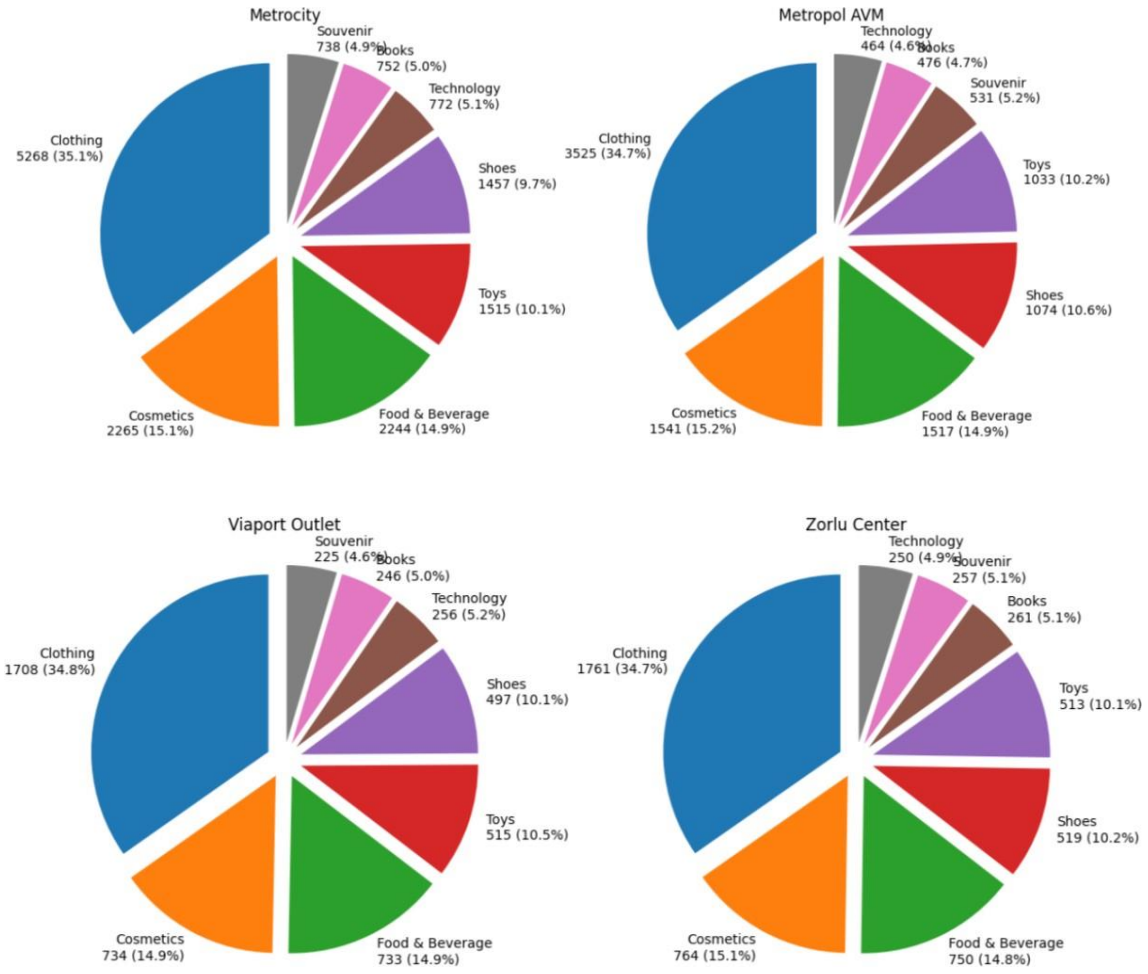


The insights that we can derive from these are:

- "Cash" is the most frequently used payment method. This suggests that many customers prefer to pay in cash when purchasing items from these categories.
- "Credit Card" is a commonly used payment method across various product categories, including "Books," "Clothing," "Cosmetics," "Food and beverage," "Shoes," "Souvenirs," and "Technology",” Toys” rather than debit cards. It demonstrates that credit cards are a popular choice among customers for a wide range of products.
- So, as we can see, credit cards are the most used card, Retailers can collaborate with banks to give cashback on buying products using the credit cards of those banks. This will generate more business for the mall and encourage the bank's business.
- Businesses can use these insights to optimise their payment processing strategies. For instance, if cash is the dominant payment method in a specific category, retailers may need to ensure they have the necessary cash-handling infrastructure in place.
- The choice of payment method can also impact the efficiency of inventory management and the checkout process. Businesses should consider the implications for their operations. This is necessary in order to keep the customers happy as the customers don’t want to waste their time in the queue. So the retailersshould have necessary stuff to make the checkout process as fast as possible

Pie chart representing sales for different categories in each mall:



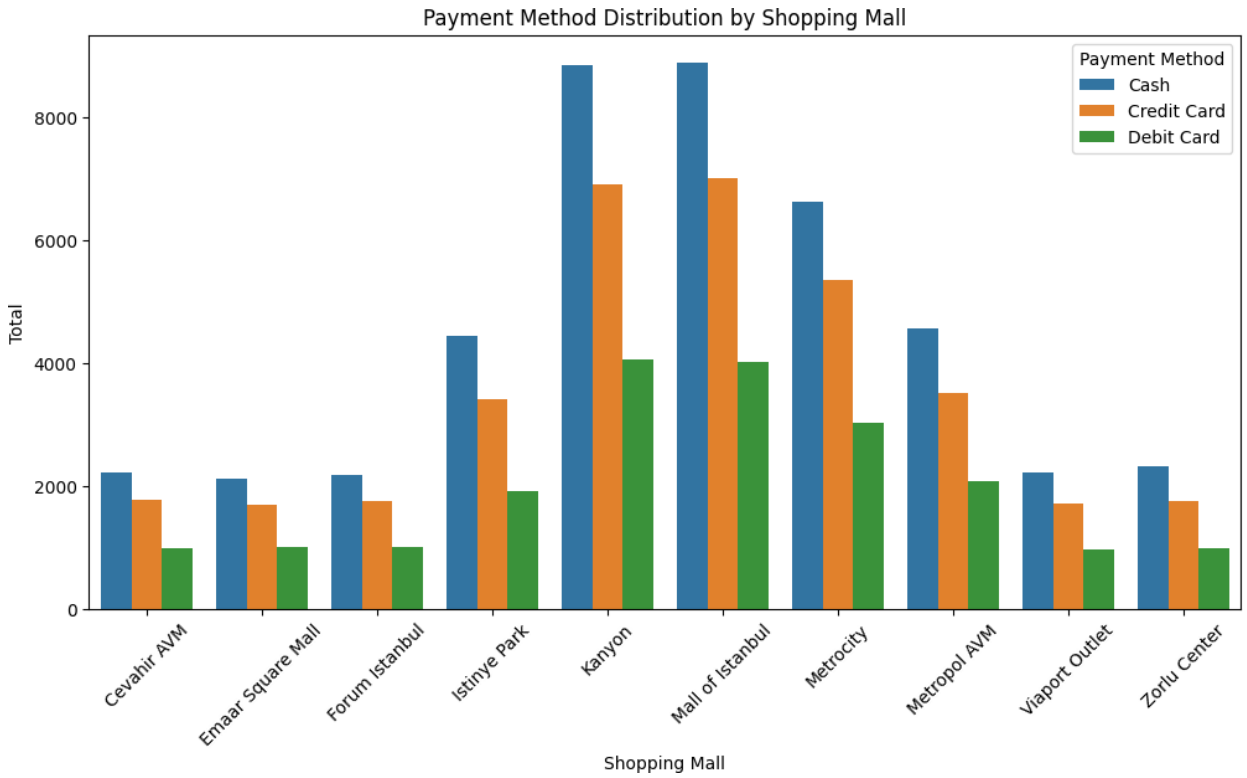


Analysing the pie chart we can arrive at the conclusion that:

- The sales of clothing is more in the **Mall of Istanbul**(6923) followed by **Kanyon Mall**(6840).
- Analysing the pie charts it is evident that the sales of all categories are high in the **Mall of Istanbul** and **Kanyon Mall** so this gives as clear evidence that these are the central and popular malls in the istanbul city.
- **Emaar Square Mall** is the mall with the smallest sales among all the other malls and the mall authority should invest more in advertising, giving offers etc. to the customers
- Analysing we can conclude that **Books** is the category that is having less sales in the malls followed by souvenir
- The category **Food & Beverages** and **Cosmetic** have almost same count of salesin almost every mall
- **Clothing** category is the most dominant category in terms of sales across all the mall.

From these insights the mall operators can take necessary steps like giving more offers andall to the customers to increase the sale in the mall

Distribution of payment methods across different shopping malls :



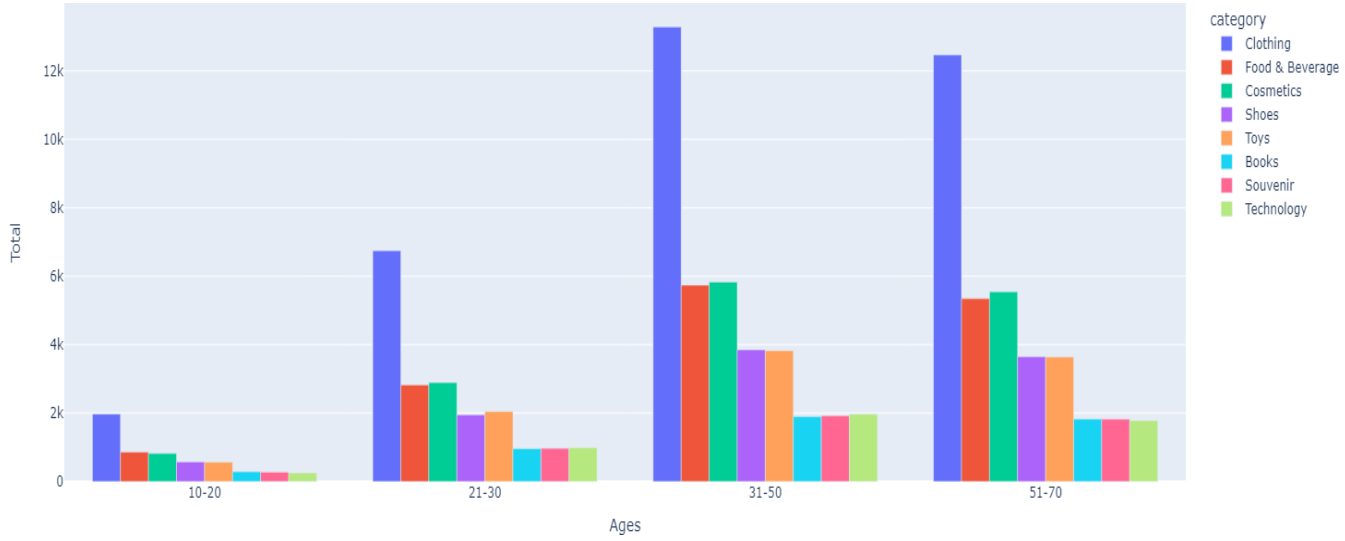
Here the bar plot visualises the distribution of payment methods (Cash, Credit Card, Debit Card) across different shopping malls.

Here are the insights from the plot:

- In all the malls most of the payments are done through cash mode. This leads tothe insights that still digital payments have not found much popularity in the istanbul city.
- As we have already identified the **Mall of Istanbul** and **Kanyon Mall**,as a result most of the card used payments are also found in this mall. As a result it opens up business opportunities for the bank sector for the distribution of their credit/debit cards,as they can collaborate with the shops in the mall and provide cashbacks for card purchases.

These insights can help shopping mall managers and retailers tailor their payment processing systems and marketing strategies to accommodate the preferred paymentmethods of their customers.

Distribution of product categories across different age ranges



The bar plot depicts the distribution of product categories across different age ranges. Here are the insights from the plot:

- The age group "31-50" tends to make the highest number of purchases across all product categories. This group has the highest total purchases in "Clothing," "Cosmetics," "Food & Beverage," "Shoes," "Toys," "Books," "Souvenir," and "Technology."
- The age group "51-70" is the second-largest consumer group across most categories, with substantial purchases in "Clothing," "Cosmetics," "Food & Beverage," "Shoes," "Toys," "Books," "Souvenir," and "Technology."
- The age group "21-30" follows closely behind the "51-70" group regarding total purchases for most categories.
- "Clothing" and "Cosmetics" are the most popular categories among all age groups.
- "Technology" and "Books" appear less popular regarding total purchases across all age ranges.

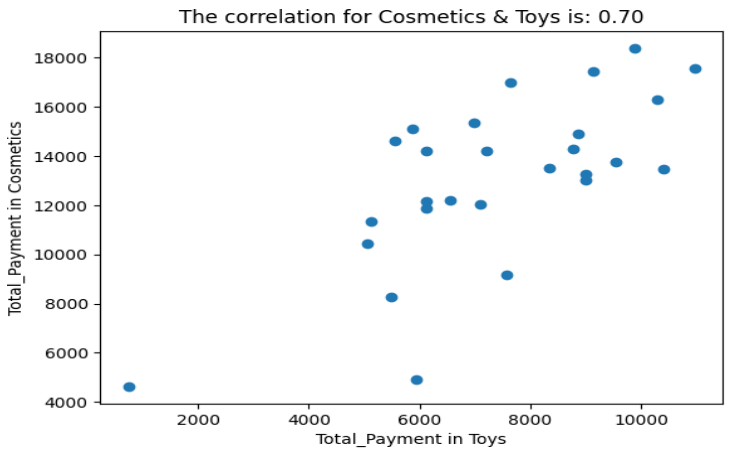
These insights can inform marketing strategies, product placement, and inventory management decisions, helping businesses tailor their offerings to the preferences of specific age groups for optimised sales and customer satisfaction.

Correlation between Different Categories for Forum istanbul Mall:


```
{'The correlation for Clothing & Shoes is': 0.3074573873915598,
 'The correlation for Clothing & Books is': 0.5549155759757178,
 'The correlation for Clothing & Cosmetics is': 0.5509662503846081,
 'The correlation for Clothing & Food & Beverage is': 0.4141645589369376,
 'The correlation for Clothing & Toys is': 0.5728478228813645,
 'The correlation for Clothing & Technology is': 0.3764138897866189,
 'The correlation for Clothing & Souvenir is': 0.27791530624343413,
 'The correlation for Shoes & Books is': -0.014252223684123353,
 'The correlation for Shoes & Cosmetics is': -0.07639124524638165,
 'The correlation for Shoes & Food & Beverage is': 0.48680907536870405,
 'The correlation for Shoes & Toys is': 0.13191983351127032,
 'The correlation for Shoes & Technology is': -0.14511456782568313,
 'The correlation for Shoes & Souvenir is': 0.1720938673951914,
 'The correlation for Books & Cosmetics is': 0.19695701372235533,
 'The correlation for Books & Food & Beverage is': 0.14817530875048868,
 'The correlation for Books & Toys is': 0.12984225455578807,
 'The correlation for Books & Technology is': 0.03212139668770539,
 'The correlation for Books & Souvenir is': 0.1776640674720151,
 'The correlation for Cosmetics & Food & Beverage is': 0.2872881489898245,
 'The correlation for Cosmetics & Toys is': 0.6977429478693891,
 'The correlation for Cosmetics & Technology is': 0.25507545382155616,
 'The correlation for Cosmetics & Souvenir is': 0.24628678201033924,
 'The correlation for Food & Beverage & Toys is': 0.4150369283547908,
 'The correlation for Food & Beverage & Technology is': 0.18343489506650357,
 'The correlation for Food & Beverage & Souvenir is': 0.40480348436600294,
 'The correlation for Toys & Technology is': 0.1620563080094655,
 'The correlation for Toys & Souvenir is': 0.322583438926367,
 'The correlation for Technology & Souvenir is': 0.22112264208548799}
```


From this filtering out categories with correlation greater than 0.65We get

'The correlation for Cosmetics & Toys is': 0.6977429478693891



This suggest that there is strong relationship exist between cosmetic and toys in Forum Istanbul mall and there is a chance for selling both item if they are placed nearby

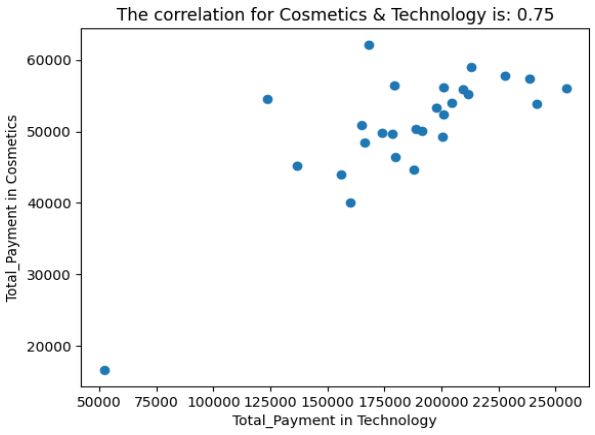
Correlation between Different Categories for Kanyon Mall:

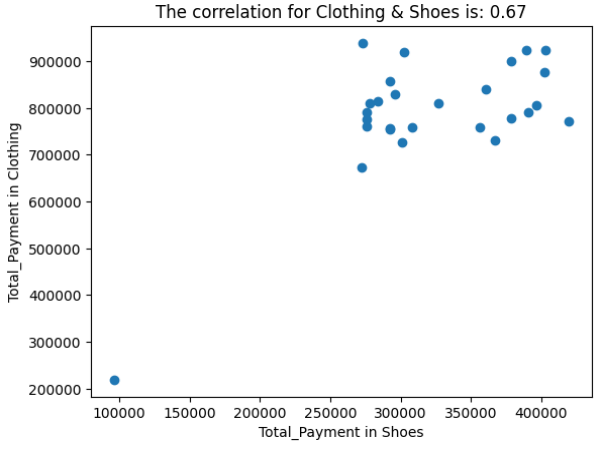
```
{'The correlation for Clothing & Shoes is': 0.6685875249956222,
 'The correlation for Clothing & Books is': 0.6347660031233875,
 'The correlation for Clothing & Cosmetics is': 0.6330615361296567,
 'The correlation for Clothing & Food & Beverage is': 0.5383813992252717,
 'The correlation for Clothing & Toys is': 0.6182222363586319,
 'The correlation for Clothing & Technology is': 0.5195285234094996,
 'The correlation for Clothing & Souvenir is': 0.5479684143243225,
 'The correlation for Shoes & Books is': 0.6205166168035346,
 'The correlation for Shoes & Cosmetics is': 0.490314604830705,
 'The correlation for Shoes & Food & Beverage is': 0.6128407779515969,
 'The correlation for Shoes & Toys is': 0.44728310234246293,
 'The correlation for Shoes & Technology is': 0.3559354920268592,
 'The correlation for Shoes & Souvenir is': 0.4074231114899687,
 'The correlation for Books & Cosmetics is': 0.6058691194649276,
 'The correlation for Books & Food & Beverage is': 0.5056233731605047,
 'The correlation for Books & Toys is': 0.5283538136697531,
 'The correlation for Books & Technology is': 0.5836642390277581,
 'The correlation for Books & Souvenir is': 0.30985042591861883,
 'The correlation for Cosmetics & Food & Beverage is': 0.6110745008873885,
 'The correlation for Cosmetics & Toys is': 0.6191706501330513,
 'The correlation for Cosmetics & Technology is': 0.7477605501526325,
 'The correlation for Cosmetics & Souvenir is': 0.4221680765163051,
 'The correlation for Food & Beverage & Toys is': 0.47977616431856446,
 'The correlation for Food & Beverage & Technology is': 0.4416279217044051,
 'The correlation for Food & Beverage & Souvenir is': 0.30534099639735096,
 'The correlation for Toys & Technology is': 0.3440905698045374,
 'The correlation for Toys & Souvenir is': 0.5697739073700453,
 'The correlation for Technology & Souvenir is': 0.35676179742741554}
```

From this filtering out categories with a correlation greater than 0.65We get

'The correlation for Clothing & Shoes is': 0.6685875249956222

'The correlation for Cosmetics & Technology is': 0.7477605501526325





The insights that we get from these are:

- The strong positive correlation between Clothing and Shoes suggests that thesetwo categories often have a complementary relationship
- Customers who purchase clothing may be more likely to buy shoes during thesame shopping trip, and vice versa.
- Customers who buy cosmetics may also be more inclined to purchasetechnology-related products, and vice versa.
- This correlation could be a result of fashion trends, seasonal shopping patterns, or store layouts that encourage customers to explore both the sections.
-

Correlation between Different Categories for Metrocity_Mall

```
{'The correlation for Clothing & Shoes is': 0.4612198696547712,
 'The correlation for Clothing & Books is': 0.36700573592328306,
 'The correlation for Clothing & Cosmetics is': 0.5970300188142166,
 'The correlation for Clothing & Food & Beverage is': 0.6075016153383657,
 'The correlation for Clothing & Toys is': 0.7440721460174076,
 'The correlation for Clothing & Technology is': 0.433183471609875,
 'The correlation for Clothing & Souvenir is': 0.49634433345665285,
 'The correlation for Shoes & Books is': 0.30159749406908537,
 'The correlation for Shoes & Cosmetics is': 0.6761210888921592,
 'The correlation for Shoes & Food & Beverage is': 0.38738328270485134,
 'The correlation for Shoes & Toys is': 0.5883952504352989,
 'The correlation for Shoes & Technology is': 0.5826100308607862,
 'The correlation for Shoes & Souvenir is': 0.3664219072803385,
 'The correlation for Books & Cosmetics is': 0.37506882997743485,
 'The correlation for Books & Food & Beverage is': 0.4178680454366621,
 'The correlation for Books & Toys is': 0.612392783857441,
 'The correlation for Books & Technology is': 0.43623982760346547,
 'The correlation for Books & Souvenir is': 0.27795476035968336,
 'The correlation for Cosmetics & Food & Beverage is': 0.5418987252094232,
 'The correlation for Cosmetics & Toys is': 0.7080544351956704,
 'The correlation for Cosmetics & Technology is': 0.6684093115884789,
 'The correlation for Cosmetics & Souvenir is': 0.5599614168160386,
 'The correlation for Food & Beverage & Toys is': 0.4670704172731319,
 'The correlation for Food & Beverage & Technology is': 0.40669255509139485,
 'The correlation for Food & Beverage & Souvenir is': 0.47957186565010174,
 'The correlation for Toys & Technology is': 0.6272721104729579,
 'The correlation for Toys & Souvenir is': 0.5425099492390306,
 'The correlation for Technology & Souvenir is': 0.4944532964373654}
```

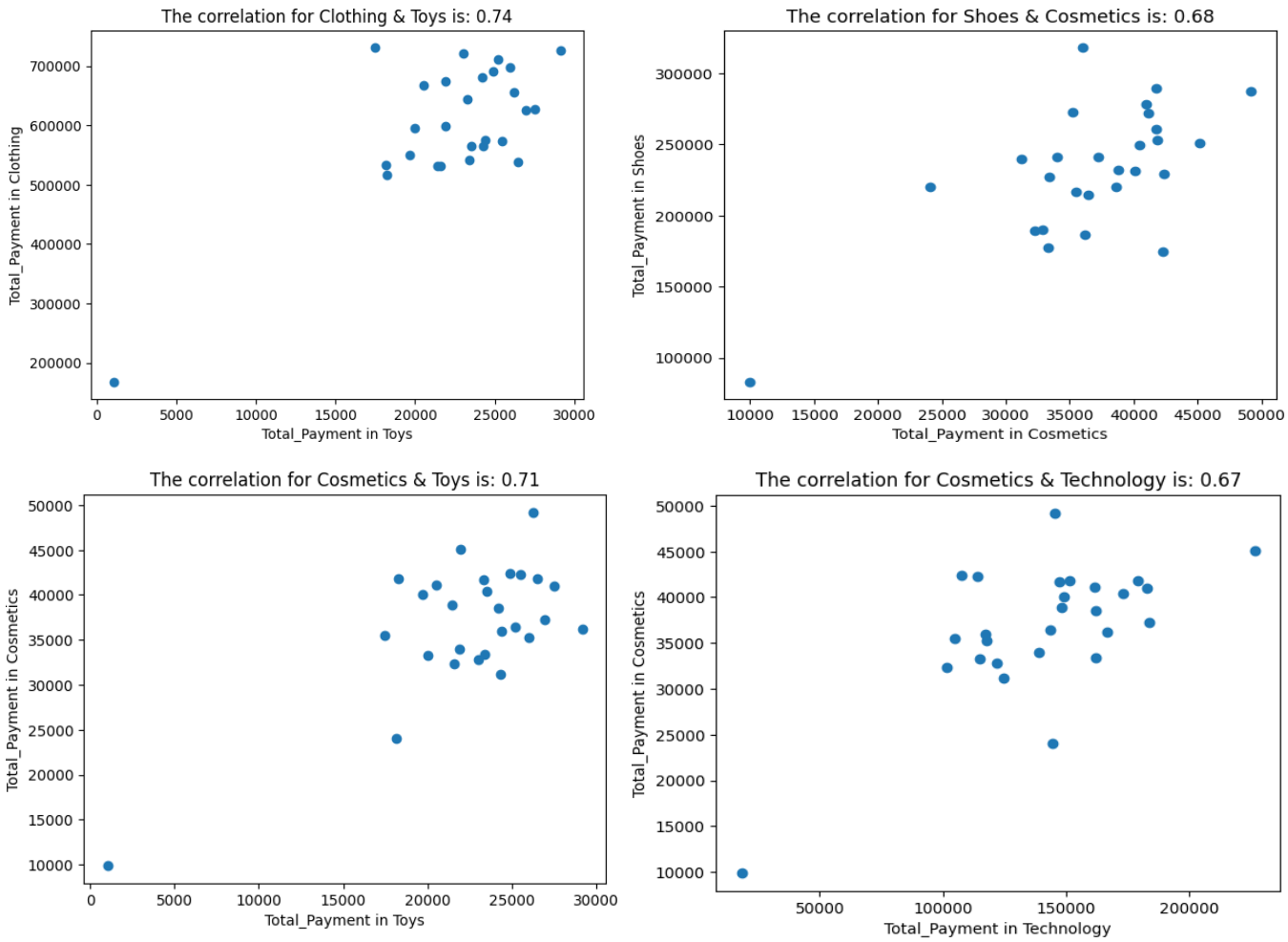
From this filtering out categories with correlation greater than 0.65We get

'The correlation for Clothing & Toys is': 0.7440721460174076

'The correlation for Shoes & Cosmetics is': 0.6761210888921592

'The correlation for Cosmetics & Toys is': 0.7080544351956704

'The correlation for Cosmetics & Technology is': 0.6684093115884789



The insights that we get from this are:

- In metrocity mall, there exists a strong positive correlation between **Clothing & Toys, Shoes & Cosmetics, Cosmetics & Toys and Cosmetics & Technology**

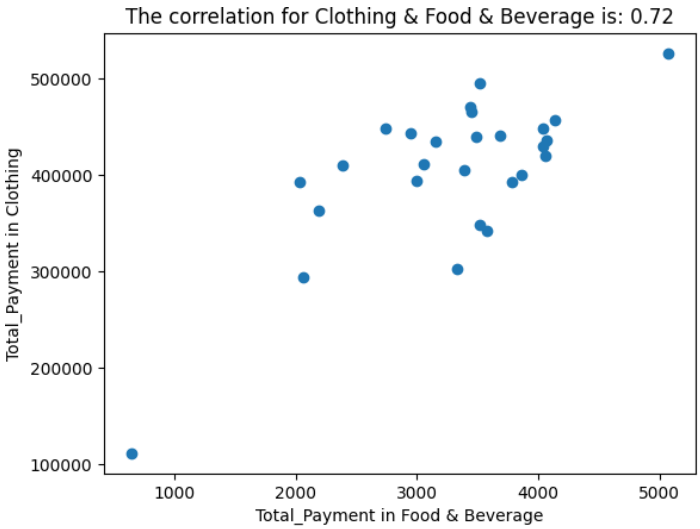
  So there is a great chance for the sale of both products if these items are kept nearby, which increases the chance for picking up both products by the customers whoare visiting the Metrocity Mall.

Correlation between Different Categories for Metropol_AVM

```
{'The correlation for Clothing & Shoes is': 0.49042040654830826,
 'The correlation for Clothing & Books is': 0.35980739640333037,
 'The correlation for Clothing & Cosmetics is': 0.5350979781967146,
 'The correlation for Clothing & Food & Beverage is': 0.7197522148842125,
 'The correlation for Clothing & Toys is': 0.5153292857923887,
 'The correlation for Clothing & Technology is': 0.3722710085890232,
 'The correlation for Clothing & Souvenir is': 0.46202257118188905,
 'The correlation for Shoes & Books is': 0.43981732701424664,
 'The correlation for Shoes & Cosmetics is': 0.33659187621556397,
 'The correlation for Shoes & Food & Beverage is': 0.6435208707961968,
 'The correlation for Shoes & Toys is': 0.44268177508299994,
 'The correlation for Shoes & Technology is': -0.049766788679319815,
 'The correlation for Shoes & Souvenir is': 0.2443033707934291,
 'The correlation for Books & Cosmetics is': 0.2971665316798064,
 'The correlation for Books & Food & Beverage is': 0.23937857938953094,
 'The correlation for Books & Toys is': 0.4295208410908584,
 'The correlation for Books & Technology is': 0.2674331063386054,
 'The correlation for Books & Souvenir is': 0.19373112545975893,
 'The correlation for Cosmetics & Food & Beverage is': 0.2829896935567778,
 'The correlation for Cosmetics & Toys is': 0.5616277710740806,
 'The correlation for Cosmetics & Technology is': 0.2332337777054933,
 'The correlation for Cosmetics & Souvenir is': 0.2781128550425158,
 'The correlation for Food & Beverage & Toys is': 0.22726323603100776,
 'The correlation for Food & Beverage & Technology is': 0.12453832535998932,
 'The correlation for Food & Beverage & Souvenir is': 0.18822852288942893,
 'The correlation for Toys & Technology is': 0.19576823277267227,
 'The correlation for Toys & Souvenir is': 0.40930851245976557,
 'The correlation for Technology & Souvenir is': 0.16008616350254098}
```

From this filtering out categories with correlation greater than 0.65We get

'The correlation for Clothing & Food & Beverage is': 0.7197522148842125



So in metropol AVM mall there is a strong correlation between Clothing and Food & Beverage category. So there is greater chanceof selling this product together.

Correlation between Different Categories for Istinye_Park_Mall

```
{'The correlation for Clothing & Shoes is': 0.5324769373620396,
 'The correlation for Clothing & Books is': 0.21112768100150572,
 'The correlation for Clothing & Cosmetics is': 0.6073300385530757,
 'The correlation for Clothing & Food & Beverage is': 0.5800042623096683,
 'The correlation for Clothing & Toys is': 0.5121053962949841,
 'The correlation for Clothing & Technology is': 0.4182651000596105,
 'The correlation for Clothing & Souvenir is': 0.4969047322191963,
 'The correlation for Shoes & Books is': 0.20617348181907952,
 'The correlation for Shoes & Cosmetics is': 0.2942964724274402,
 'The correlation for Shoes & Food & Beverage is': 0.5939684635412801,
 'The correlation for Shoes & Toys is': 0.38497584788136596,
 'The correlation for Shoes & Technology is': 0.21496429580387616,
 'The correlation for Shoes & Souvenir is': 0.624918045536435,
 'The correlation for Books & Cosmetics is': 0.2534639638530651,
 'The correlation for Books & Food & Beverage is': 0.2883754241848799,
 'The correlation for Books & Toys is': 0.15694736843770568,
 'The correlation for Books & Technology is': 0.2238363638077626,
 'The correlation for Books & Souvenir is': 0.02523615371283216,
 'The correlation for Cosmetics & Food & Beverage is': 0.40684853299863255,
 'The correlation for Cosmetics & Toys is': 0.6258954873459099,
 'The correlation for Cosmetics & Technology is': 0.4926292615134862,
 'The correlation for Cosmetics & Souvenir is': 0.18983683431792606,
 'The correlation for Food & Beverage & Toys is': 0.5302986124371574,
 'The correlation for Food & Beverage & Technology is': 0.4716696961646053,
 'The correlation for Food & Beverage & Souvenir is': 0.3665743735759392,
 'The correlation for Toys & Technology is': 0.4784222341077253,
 'The correlation for Toys & Souvenir is': 0.4389135035883296,
 'The correlation for Technology & Souvenir is': 0.26313522955237845}
```

Analysing this we can confirm there is no evidence of strong positive correlation betweendifferent categories as like the other malls. Here only the **Clothing and Cosmetic** and **Cosmetic and Toys** have at least a correlation value above 0.6 even though it cannot be considered as a significant for the Istinye_Park_Mall this items have a significant importance in the part of sales. So this mall can considering giving offers on purchasing both items together etc to increase their sales

Correlation between Different Categories for Mall_of_Istanbul

```
{'The correlation for Clothing & Shoes is': 0.7350082742464816,
 'The correlation for Clothing & Books is': 0.52252708537011,
 'The correlation for Clothing & Cosmetics is': 0.5792746043858747,
 'The correlation for Clothing & Food & Beverage is': 0.7092508418873701,
 'The correlation for Clothing & Toys is': 0.7451204248666052,
 'The correlation for Clothing & Technology is': 0.64736822953744,
 'The correlation for Clothing & Souvenir is': 0.4370160886370177,
 'The correlation for Shoes & Books is': 0.48992732715401904,
 'The correlation for Shoes & Cosmetics is': 0.6119008140356245,
 'The correlation for Shoes & Food & Beverage is': 0.7472160716138211,
 'The correlation for Shoes & Toys is': 0.5860085355769747,
 'The correlation for Shoes & Technology is': 0.5667310493792752,
 'The correlation for Shoes & Souvenir is': 0.39103400754889445,
 'The correlation for Books & Cosmetics is': 0.377444637206964,
 'The correlation for Books & Food & Beverage is': 0.40379365917358134,
 'The correlation for Books & Toys is': 0.46958250840667465,
 'The correlation for Books & Technology is': 0.49618846824358315,
 'The correlation for Books & Souvenir is': 0.6728630486620977,
 'The correlation for Cosmetics & Food & Beverage is': 0.6523000470603415,
 'The correlation for Cosmetics & Toys is': 0.6193117054797499,
 'The correlation for Cosmetics & Technology is': 0.48502132003230264,
 'The correlation for Cosmetics & Souvenir is': 0.540712299677522,
 'The correlation for Food & Beverage & Toys is': 0.5316728620657278,
 'The correlation for Food & Beverage & Technology is': 0.6405044508633844,
 'The correlation for Food & Beverage & Souvenir is': 0.448861368648651,
 'The correlation for Toys & Technology is': 0.5156052466993425,
 'The correlation for Toys & Souvenir is': 0.5586628775862555,
 'The correlation for Technology & Souvenir is': 0.5473229772329091}
```

From this filtering out categories with correlation greater than 0.65We get

'The correlation for Clothing & Shoes is': 0.7350082742464816

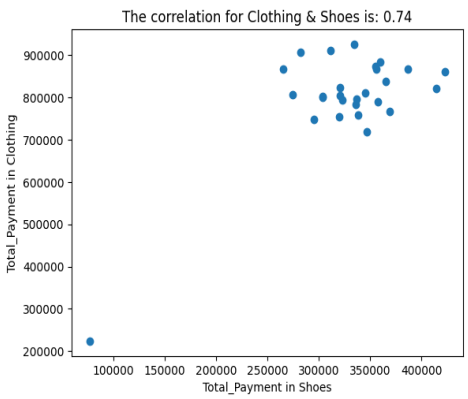
'The correlation for Clothing & Food & Beverage is': 0.709250841887370

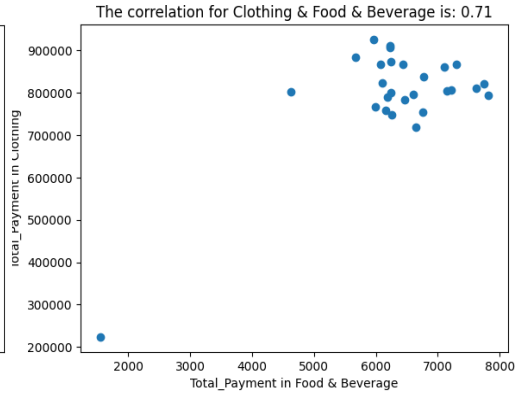
'The correlation for Clothing & Toys is': 0.7451204248666052

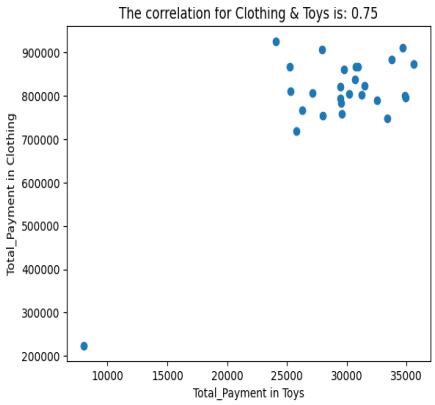
'The correlation for Shoes & Food & Beverage is': 0.7472160716138211

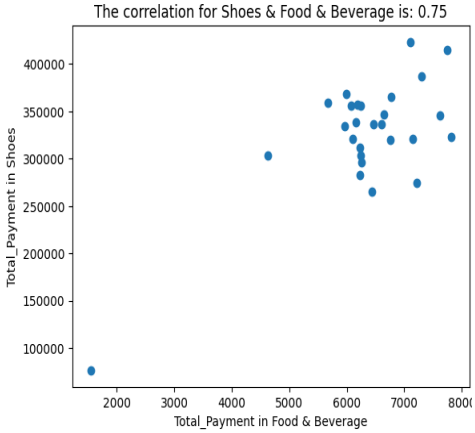
'The correlation for Books & Souvenir is': 0.6728630486620977

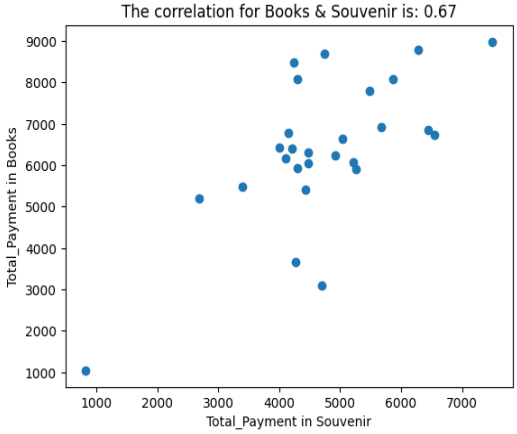
'The correlation for Cosmetics & Food & Beverage is': 0.6523000470603415

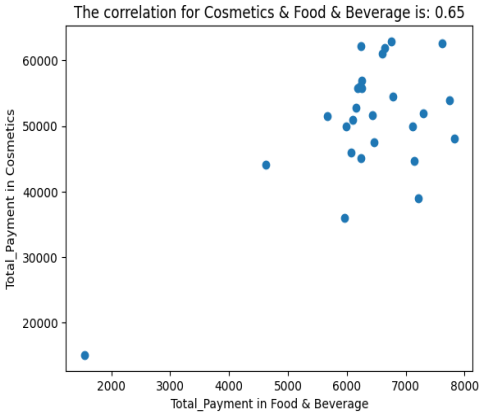












We have already identified that the Mall of Istanbul is a popular mall with bigger sales allover the year so as a result it is common to find a more correlation between different categories

Here we got 6 strong correlation between different categoriesThe

insights that we get from this are:

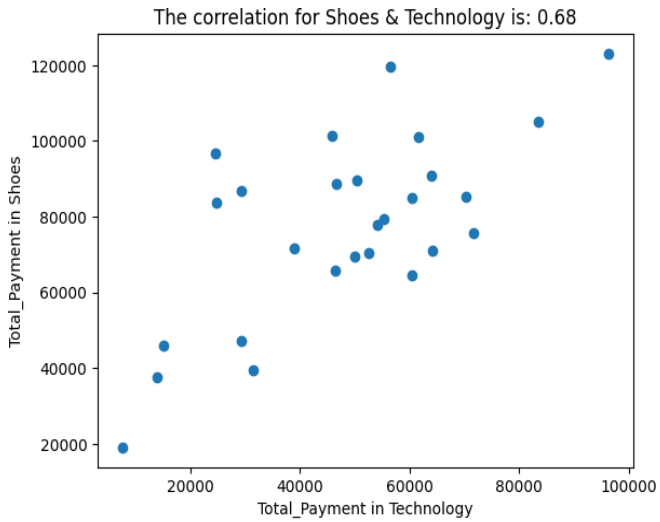
- There is greater chance for the sale of both the product clothing and shoes if we can keep it nearby as it shows a strong relation in sales
- There is a greater chance for purchasing Food & beverages by a customer who purchases Clothing or shoes or cosmetic so if we keep this items nearby then thereis a greater chance for customers buying this together
- We can also see that there is a strong correlation between books and souvenir category in most of the malls. The sales of books is less compared to other category products so this insight is a valuable insight that there's a chance of sales of both these items together and which results in a greater sales of books.

Correlation between Different Categories for Emaar_Square_Mall

```
{'The correlation for Clothing & Shoes is': 0.3278020618306169,
 'The correlation for Clothing & Books is': 0.35452576682005277,
 'The correlation for Clothing & Cosmetics is': 0.16554668619560742,
 'The correlation for Clothing & Food & Beverage is': 0.4248284129780588,
 'The correlation for Clothing & Toys is': 0.5359235393531103,
 'The correlation for Clothing & Technology is': 0.28222748203393827,
 'The correlation for Clothing & Souvenir is': -0.0765584086361198,
 'The correlation for Shoes & Books is': -0.04364038304063776,
 'The correlation for Shoes & Cosmetics is': 0.30143071294100615,
 'The correlation for Shoes & Food & Beverage is': 0.03378515907438762,
 'The correlation for Shoes & Toys is': 0.5054942024009358,
 'The correlation for Shoes & Technology is': 0.676238794007046,
 'The correlation for Shoes & Souvenir is': -0.009542952950467979,
 'The correlation for Books & Cosmetics is': 0.06052652817195238,
 'The correlation for Books & Food & Beverage is': 0.16626915711427206,
 'The correlation for Books & Toys is': 0.11602881469388247,
 'The correlation for Books & Technology is': 0.022985092684262163,
 'The correlation for Books & Souvenir is': -0.14894381174598906,
 'The correlation for Cosmetics & Food & Beverage is': 0.18990045839403122,
 'The correlation for Cosmetics & Toys is': 0.04685034422770366,
 'The correlation for Cosmetics & Technology is': 0.05488175385793237,
 'The correlation for Cosmetics & Souvenir is': 0.2523149584072286,
 'The correlation for Food & Beverage & Toys is': 0.02825178355476964,
 'The correlation for Food & Beverage & Technology is': 0.13665745807580026,
 'The correlation for Food & Beverage & Souvenir is': 0.1067920814037407,
 'The correlation for Toys & Technology is': 0.25492231783631114,
 'The correlation for Toys & Souvenir is': -0.008367793838269673,
 'The correlation for Technology & Souvenir is': 0.009873909500682029}
```

From this filtering out categories with correlation greater than 0.65We get

'The correlation for Shoes & Technology is': 0.676238794007046



Here we can witness a correlation between category shoes and technology in the Emaar square mall. So there's a chance for purchasingboth these product together by a customer whois coming to this mall

Correlation between Different Categories for Cevahir_AVM_Mall

```
{'The correlation for Clothing & Shoes is': 0.5112703119970333,
 'The correlation for Clothing & Books is': 0.30365478210234964,
 'The correlation for Clothing & Cosmetics is': 0.5137851963096712,
 'The correlation for Clothing & Food & Beverage is': 0.307247292277524,
 'The correlation for Clothing & Toys is': 0.49612663942474294,
 'The correlation for Clothing & Technology is': 0.08763994421135347,
 'The correlation for Clothing & Souvenir is': -0.2209860256167467,
 'The correlation for Shoes & Books is': 0.4730686151271357,
 'The correlation for Shoes & Cosmetics is': 0.21647095942459998,
 'The correlation for Shoes & Food & Beverage is': 0.17570000855787063,
 'The correlation for Shoes & Toys is': 0.37727195299646815,
 'The correlation for Shoes & Technology is': -0.10879274005722196,
 'The correlation for Shoes & Souvenir is': -0.1796778447487324,
 'The correlation for Books & Cosmetics is': 0.2261588818065977,
 'The correlation for Books & Food & Beverage is': -0.049692170374093445,
 'The correlation for Books & Toys is': 0.3715170444206959,
 'The correlation for Books & Technology is': 0.02410337156929968,
 'The correlation for Books & Souvenir is': -0.03420642533365947,
 'The correlation for Cosmetics & Food & Beverage is': 0.16359052700467788,
 'The correlation for Cosmetics & Toys is': 0.2071967393443757,
 'The correlation for Cosmetics & Technology is': 0.15710705067983396,
 'The correlation for Cosmetics & Souvenir is': 0.12092019813492172,
 'The correlation for Food & Beverage & Toys is': 0.41466114719990715,
 'The correlation for Food & Beverage & Technology is': -0.05050362191056839,
 'The correlation for Food & Beverage & Souvenir is': 0.31681667109177464,
 'The correlation for Toys & Technology is': 0.1343577170695028,
 'The correlation for Toys & Souvenir is': 0.07918161108904163,
 'The correlation for Technology & Souvenir is': 0.1488553233485559}
```

Analysing the above correlations we can say that there is no strong correlation between different categories of products that are being sold in the mall. The reason is because the Cevahir AVM mall is having less sales compared to all other malls in the Istanbul city.
This may be because of the less popularity of the mall.

So the insight that we can give from this to the Cevahir AVM mall is that

- They should focus on more advertising to the products that is being selling in themall
- They should find insights from the other malls for organising strategies to increase the sales
- They should consider selling multiple categories of items with offers so that more customers will be attracted to the mall and both the categories will be sold simultaneously

Correlation between Different Categories for Viaport_Outlet_Mall

```
{'The correlation for Clothing & Shoes is': 0.3533871332870977,
 'The correlation for Clothing & Books is': 0.3133594820193293,
 'The correlation for Clothing & Cosmetics is': 0.3377102101745824,
 'The correlation for Clothing & Food & Beverage is': 0.6389534143891689,
 'The correlation for Clothing & Toys is': 0.3766618997053186,
 'The correlation for Clothing & Technology is': 0.22317422918505916,
 'The correlation for Clothing & Souvenir is': 0.45211184686772726,
 'The correlation for Shoes & Books is': 0.41071269330733695,
 'The correlation for Shoes & Cosmetics is': 0.10309799221607702,
 'The correlation for Shoes & Food & Beverage is': 0.4268580027572636,
 'The correlation for Shoes & Toys is': 0.2164812598592762,
 'The correlation for Shoes & Technology is': 0.19427536194142447,
 'The correlation for Shoes & Souvenir is': 0.31394720128777087,
 'The correlation for Books & Cosmetics is': 0.21416786912410724,
 'The correlation for Books & Food & Beverage is': 0.015102769203562525,
 'The correlation for Books & Toys is': 0.35759386792983094,
 'The correlation for Books & Technology is': 0.006032303096295922,
 'The correlation for Books & Souvenir is': 0.3165095246720418,
 'The correlation for Cosmetics & Food & Beverage is': 0.01334006389427057,
 'The correlation for Cosmetics & Toys is': 0.059647435915931045,
 'The correlation for Cosmetics & Technology is': 0.015458646170937023,
 'The correlation for Cosmetics & Souvenir is': 0.20433913657818928,
 'The correlation for Food & Beverage & Toys is': 0.09337446817040289,
 'The correlation for Food & Beverage & Technology is': 0.4756420726340407,
 'The correlation for Food & Beverage & Souvenir is': 0.15076050299443905,
 'The correlation for Toys & Technology is': 0.12078080818277317,
 'The correlation for Toys & Souvenir is': 0.5192226808136319,
 'The correlation for Technology & Souvenir is': 0.12026378987184383}
```

Analysing this, there is no strong correlation between different categories of items that are being sold in the mall. Viaport Outlet Mall also has less product sales compared to other malls. This may be the reason for the lack of correlation between the products in the malls.

So, insights that we can derive from the correlation are:

- Here, the customers are used to purchasing only single products rather than multiple ones.
- So the mall authority should take necessary steps to make the customers buymultiple products together by availing offers, discounts etc..

Correlation between Different Categories for Zorlu_Center_Mall

```
{'The correlation for Clothing & Shoes is': 0.5391709479729078,
 'The correlation for Clothing & Books is': 0.47990841781644056,
 'The correlation for Clothing & Cosmetics is': 0.481246074103121,
 'The correlation for Clothing & Food & Beverage is': 0.44759036649010986,
 'The correlation for Clothing & Toys is': 0.3848555174568406,
 'The correlation for Clothing & Technology is': 0.4509171017913562,
 'The correlation for Clothing & Souvenir is': 0.13143431016809162,
 'The correlation for Shoes & Books is': 0.28864721780363434,
 'The correlation for Shoes & Cosmetics is': 0.4066500936789613,
 'The correlation for Shoes & Food & Beverage is': 0.08719195074814938,
 'The correlation for Shoes & Toys is': 0.43992350740226527,
 'The correlation for Shoes & Technology is': 0.31872514760801623,
 'The correlation for Shoes & Souvenir is': 0.0918841970499685,
 'The correlation for Books & Cosmetics is': 0.09129207902389713,
 'The correlation for Books & Food & Beverage is': 0.4643334414121627,
 'The correlation for Books & Toys is': 0.07507988607391886,
 'The correlation for Books & Technology is': 0.5617859853467883,
 'The correlation for Books & Souvenir is': 0.10135511681980168,
 'The correlation for Cosmetics & Food & Beverage is': -0.03173177990435617,
 'The correlation for Cosmetics & Toys is': 0.354181370942329,
 'The correlation for Cosmetics & Technology is': 0.13473853896380203,
 'The correlation for Cosmetics & Souvenir is': 0.14278394116237395,
 'The correlation for Food & Beverage & Toys is': 0.11406055178689034,
 'The correlation for Food & Beverage & Technology is': 0.4637580995305792,
 'The correlation for Food & Beverage & Souvenir is': -0.25822053814340223,
 'The correlation for Toys & Technology is': 0.041083415104211714,
 'The correlation for Toys & Souvenir is': -0.023203680747774406,
 'The correlation for Technology & Souvenir is': 0.03044558314999141}
```

Analysing the correlation between categories for Zorlu Center mall also shows that there is no significant correlation between different categories of products

So in this dataset 4 malls are showing no correlation between categories they are:

Cevahir AVM, Istinye Park ,Viaport Outlet ,Zorlu Center

As analysing the sales of the malls itself we can notice that these malls are lacking sales compared to other malls. So the insights that we can generate common for these malls are:

- They should gain insights from the others malls to build strategies for sales.
- They should invest more in advertising to make the malls popular to attractcustomers.
- They should start providing offers and discounts to attract more customers
- They can collaborate with banks to offer cashbacks on card payments this willattract more card holders to the mall.

Dataframe showing the categories in malls having strong correlation

	city	mall	correlation	item1	item2
0	Istanbull	Forum istanbul Mall	0.697743	Cosmetics	Toys
1	Istanbull	Kanyon Mall	0.668588	Clothing	Shoes
2	Istanbull	Kanyon Mall	0.747761	Cosmetics	Technology
3	Istanbull	Metrocity Mall	0.744072	Clothing	Toys
4	Istanbull	Metrocity Mall	0.676121	Shoes	Cosmetics
5	Istanbull	Metrocity Mall	0.708054	Cosmetics	Toys
6	Istanbull	Metrocity Mall	0.668409	Cosmetics	Technology
7	Istanbull	Metropol AVM	0.719752	Clothing	Food & Beverage
8	Istanbull	Mall of Istanbul	0.735008	Clothing	Shoes
9	Istanbull	Mall of Istanbul	0.709251	Clothing	Food & Beverage
10	Istanbull	Mall of Istanbul	0.745120	Clothing	Toys
11	Istanbull	Mall of Istanbul	0.747216	Shoes	Food & Beverage
12	Istanbull	Mall of Istanbul	0.672863	Books	Souvenir
13	Istanbull	Mall of Istanbul	0.652300	Cosmetics	Food & Beverage
14	Istanbull	Emaar Square Mall	0.676239	Shoes	Technology

This is an important data frame because it shows the categories in malls having strong correlation and this is the correlation data frame of the entire Istanbul city. So by analysing this, we can find out the general correlation trends that is being shown in the Istanbul City
The insights that we can gain from these are:

- The category **Cosmetic and Toys** have a strong correlation in 2 malls in the city, which gives information that their is a chance that the customers in the city buying cosmetics also buy toys too, so we can keep this product nearby so that customers will show a tendency to buy the products together and thereby increasethe sales

- The category **Clothing and Shoes** have a strong correlation in 2 malls in the city, which gives as information that there is the chance that the customers in the city buying clothing also buy shoes along with it so we can keep this product nearbyso that customers will show a tendency to buy the products together and thereby increase the sales
- The category **Cosmetic and Technology** have a strong correlation in 2 malls in thecity, which gives information that there is a chance that the customers in the citybuying cosmetics also buy technology products, so we can keep this product nearby so that customers will show a tendency to buy the products together and thereby increase the sales
- The category **Clothing and Toys** have a strong correlation in 2 malls in the city, which gives us information that there is a chance that the customers in the city buying clothing also buy toys, so we can keep these products nearby so that customers will show a tendency to buy the products together and thereby increasethe sales
- The category **Clothing and Food and Beverage** have a strong correlation in 2 malls in the city, which gives information that there is a chance that the customers in the city buying clothing also buy food and beverage items, so we can keep these products nearby so that customers will show a tendency to buy the products together and thereby increase the sales

So these are the categories that show a strong correlation trends across the city they are:

- **Cosmetic and Toys**
- **Clothing and Shoes**
- **Cosmetic and Technology**
- **Clothing and Toys**
- **Clothing and Food & Beverage**

We have already identified clothing and cosmetics as the top-selling products across all the malls, and these two products are making a correlation with others also. So we can conclude that most of the customers are attracted to the malls for the products, clothing and cosmetic
So if the malls want to increase the sale of other products, they can make combo offerswith any of the clothing or cosmetic products, the chance of more sales is very high.

Time series Analysis

Time series for Food & Beverage in Metrocity mall

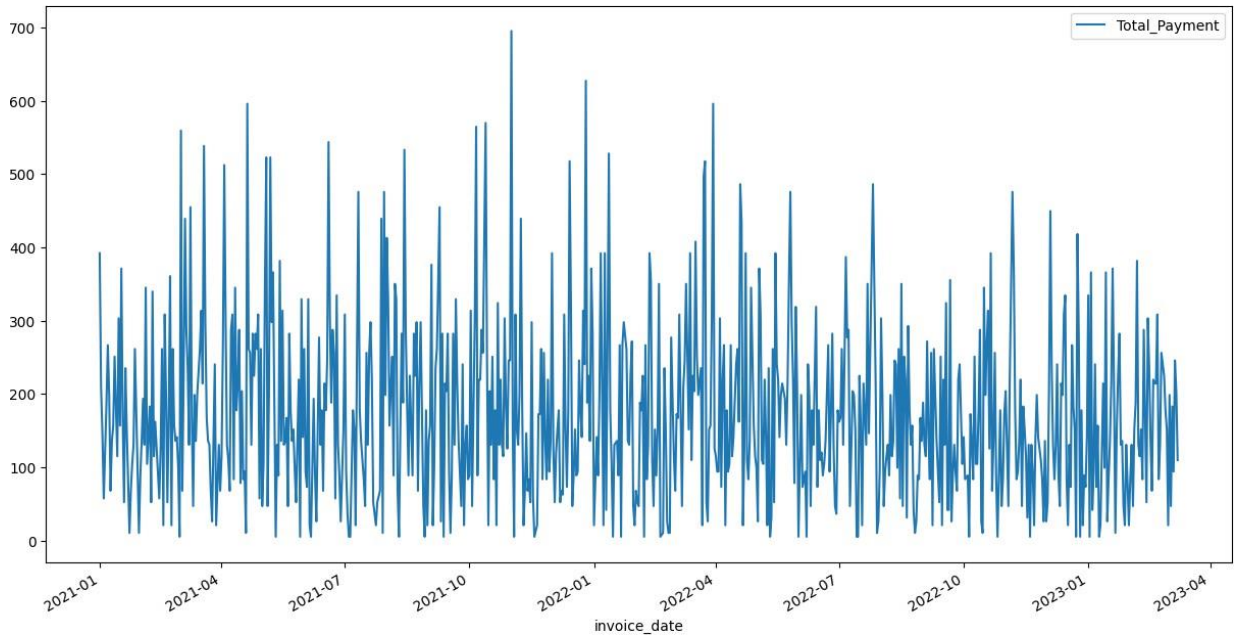
Dataframe created by grouping sum of the invoice date for the category "Food & Beverage" for "Metrocity mall"

invoice_date	Total_Payment
2021-01-01	392.25
2021-01-02	209.20
2021-01-03	146.44
2021-01-04	57.53
2021-01-06	188.28
...	...
2023-03-04	183.05
2023-03-05	94.14
2023-03-06	245.81
2023-03-07	203.97
2023-03-08	109.83

747 rows × 1 columns

Here the invoice date is set to the index position in order to facilitate the Time series analysis.

Plotting the above data frame using the plot function



From the above figure we can see the variation in the values based on each month.

Checking whether the data frame is stationary or not using The Augmented Dickey-Fuller (ADF) test,a statistical test used to determine the stationarity of a time series data.

Code snippet:

```
from statsmodels.tsa.stattools import adfuller
```

```
def adf_test(series):
    result=adfuller(series)
    print('ADF Statistics: {}'.format(result[0]))
    print('p- value: {}'.format(result[1]))
    if result[1] <= 0.05:
        print("strong evidence against the null hypothesis, reject the null hypothesis. Data has no unit root and is stationary")
    else:
        print("weak evidence against null hypothesis, time series has a unit root, indicating it is non-stationary ")
```

For the given dataset, we performed the ADF test and obtained the following results:

- ADF Statistics: -26.172871041530314
- p-value: 0.0

The ADF Statistics value is highly negative, indicating a strong deviation from the null hypothesis. The extremely low p-value suggests we can confidently reject the null hypothesis. Therefore, we have strong evidence against the null hypothesis: "The data has no unit root and is stationary."

This result indicates a stationary time series, which is crucial for many time series analysis techniques that assume stationarity. The rejection of the null hypothesis implies that the data's statistical properties do not change over time, making it more amenable forvarious analytical purposes.

It is important to note that achieving stationarity is often desirable fortime series data, as it simplifies the modelling process and allows for more accurate predictions and insights.

In conclusion, based on the ADF test results, we can confidently state that the provided dataset exhibits strong evidence of stationarity and lacks a unit root.

Now the dataset has split into 80% of train data and 20% of test data

invoice_date	Total_Payment
2022-10-04	88.91
2022-10-05	5.23
2022-10-06	172.59
2022-10-07	130.75
2022-10-08	83.68
...	...
2023-03-04	183.05
2023-03-05	94.14
2023-03-06	245.81
2023-03-07	203.97
2023-03-08	109.83

150 rows × 1 columns

Test Data

invoice_date	Total_Payment
2021-01-01	392.25
2021-01-02	209.20
2021-01-03	146.44
2021-01-04	57.53
2021-01-06	188.28
...	...
2022-09-27	219.66
2022-09-28	240.58
2022-09-30	104.60
2022-10-01	141.21
2022-10-02	83.68

597 rows × 1 columns

Train Data

Building ARIMA Model for Time Series

Code snippet:

```
## create a ARIMA model
from statsmodels.tsa.arima.model import ARIMA
```

```
model_ARIMA=ARIMA(k1['Total_Payment'],order=(4,3,1))
```

```
model_Arima_fit=model_ARIMA.fit()
```

ARIMA model summary:

SARIMAX Results

Dep. Variable:	Total_Payment	**No. Observations:**	747
Model:	ARIMA(4, 3, 1)	**Log Likelihood**	-4859.803
Date:	Sat, 02 Sep 2023	**AIC**	9731.606
Time:	10:26:56	**BIC**	9759.278
Sample:	0	**HQIC**	9742.272
	- 747		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	-1.2924	0.032	-40.835	0.000	-1.354	-1.230
ar.L2	-1.1735	0.048	-24.321	0.000	-1.268	-1.079
ar.L3	-0.7857	0.051	-15.399	0.000	-0.886	-0.686
ar.L4	-0.3484	0.033	-10.552	0.000	-0.413	-0.284
ma.L1	-1.0000	0.050	-19.879	0.000	-1.099	-0.901
sigma2	2.723e+04	1.85e-06	1.47e+10	0.000	2.72e+04	2.72e+04

Ljung-Box (L1) (Q):	8.10	**Jarque-Bera (JB):**	3.38
Prob(Q):	0.00	**Prob(JB):**	0.18
Heteroskedasticity (H):	0.60	**Skew:**	0.15
Prob(H) (two-sided):	0.00	**Kurtosis:**	3.15

Insights from the above summary:

Model Specification

- Dependent Variable: Total_Payment
- No. of Observations: 747
- Model: ARIMA(4, 3, 1)
- Log Likelihood: -4859.803
- AIC (Akaike Information Criterion): 9731.606

- BIC (Bayesian Information Criterion): 9759.278
- HQIC (Hannan-Quinn Information Criterion): 9742.272

Coefficients

The estimated coefficients of the model are as follows:

- ar.L1: -1.2924
- ar.L2: -1.1735
- ar.L3: -0.7857
- ar.L4: -0.3484
- ma.L1: -1.0000

These coefficients provide insights into the relationships and impacts of past observationsand past forecast errors on the current value of the dependent variable.

Here, the ar.L1,arL2 etc. represent the autoregressive coefficient at a specific lag.

For example, "ar.L1" indicates the autoregressive coefficient for the first lag , "ar.L2" for the second lag, and so on. In comparison, the "MA" term with a number (e.g., "ma.L1") represents the moving average coefficient at a specific lag. For example, "ma.L1" indicates the moving average coefficient for the first lag.

Model Diagnostic Measures

- Ljung-Box (L1) (Q): 8.10 (Significant)
- Jarque-Bera (JB): 3.38 (Not Significant)
- Heteroskedasticity (H): 0.60 (Significant)
- Skewness: 0.15
- Kurtosis: 3.15

The Ljung-Box test measures the presence of autocorrelation in the residuals, and the Jarque-Bera test assesses the normality assumption of residuals. Heteroskedasticity measures the variance of residuals across different levels of the independent variables.

Testing the model using the test data:

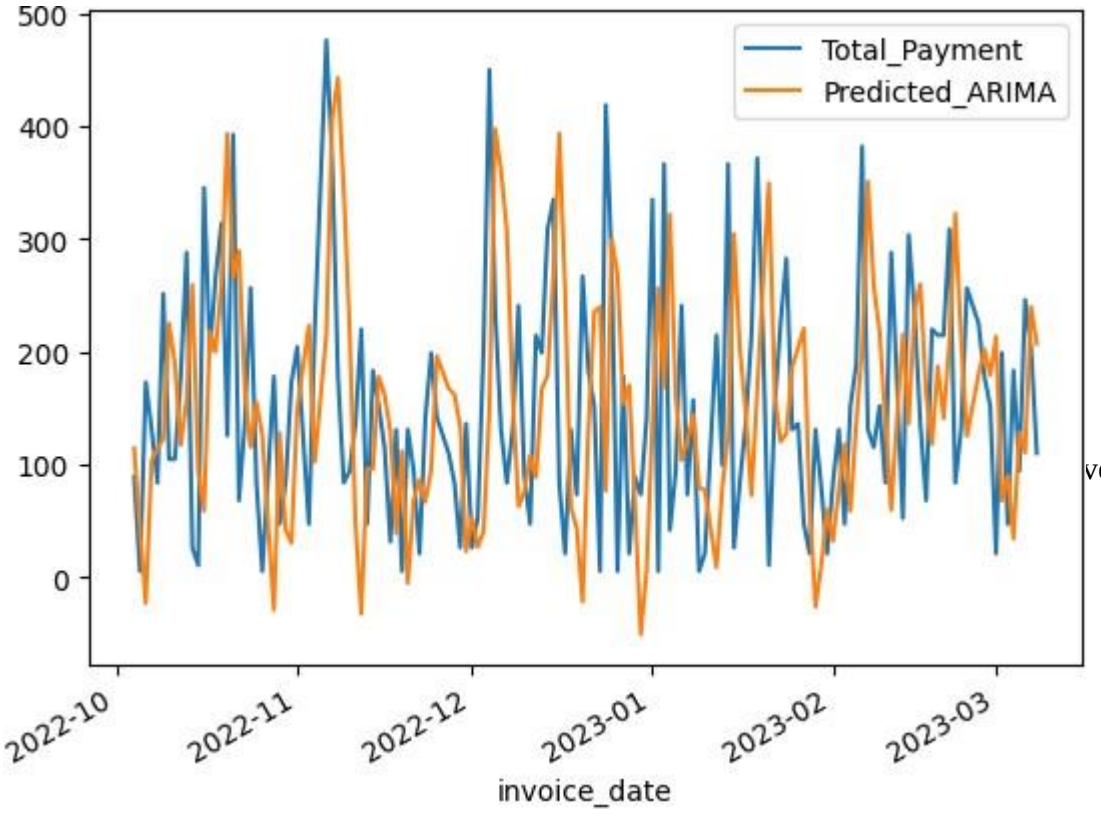
Predicting the values for the test data

Starting date: 2022-10-04

End date:2023-03-08

```
invoice_date
2022-10-04    114.321431
2022-10-05     42.727517
2022-10-06    -23.235531
2022-10-07    103.763896
2022-10-08    113.158844
                  ...
2023-03-04     34.175579
2023-03-05    128.283521
2023-03-06    110.023858
2023-03-07    239.730797
2023-03-08    206.842076
Name: predicted_mean, Length: 150, dtype: float64
```

Plotting the Predicted price with the actual price:



ver.

Here we can see that the predicted values are moving close to the actual values which means that the model is performing well.

Now, forecasting the new values for the next 10 days starting from 2023-03-09 using the forecast function in the model.

Date	Forecasted_Values
2023-03-09	200.005222
2023-03-10	181.903259
2023-03-11	196.260461
2023-03-12	169.450666
2023-03-13	178.941826
2023-03-14	202.361714
2023-03-15	186.536592
2023-03-16	191.222942
2023-03-17	192.184607
2023-03-18	200.191493

The above are the predicted values for the next 10 days for the category food and beverages of Forum Istanbul mall by having this information the shops in the mall sellingthe food and beverages can arrange the necessary things in order to satisfy the customer needs by understanding the fact that on which all days the shopping is going to be high

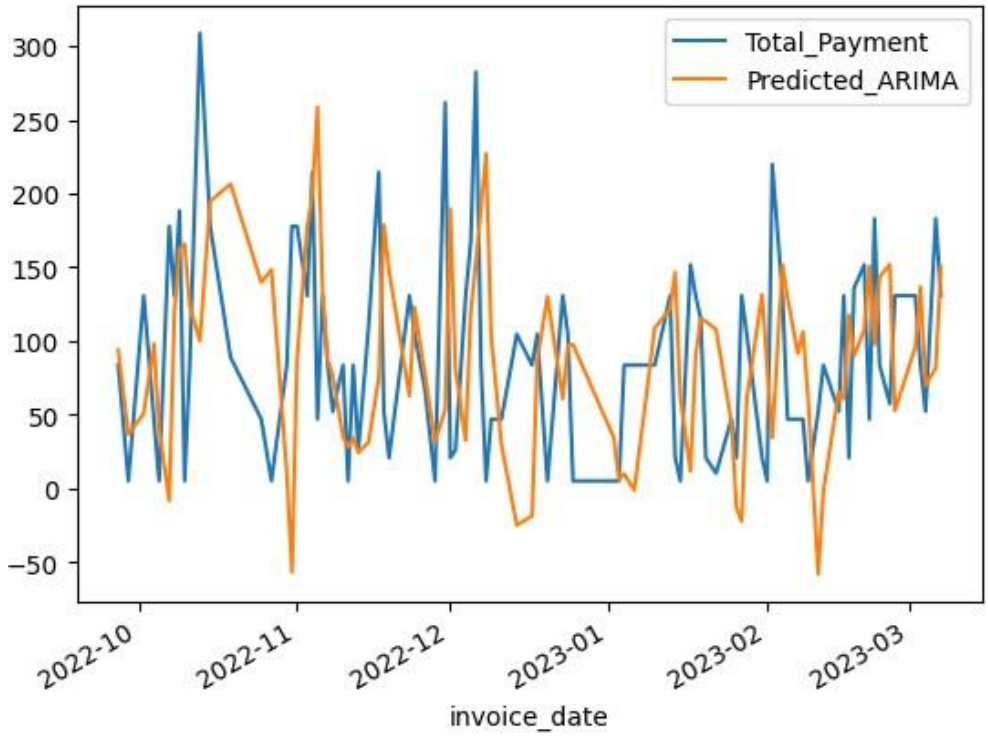
This is a crucial insights that can lead to advanced knowledge about the amount of sales that may occur in the future by knowing it already the retailers and mall organisers can arrange necessary steps to satisfy the needs of the customers
If a future day price can we known in advance then the retailers can give more offers, discounts etc to attract the customers

Similarly we can run this ARIMA Model for forecasting the future prices for any other category belonging to any other mall
In the same way discussed above we are forecasting the future price forFood &
Beverage in Forum Istanbul mall.

Time series for Food & Beverage in Forum Istanbul mall

Here also we are following the same steps as discussed above

After building the ARIMA model and fitting it we are testing the model performancePlotting the Predicted price with the actual price:



Here also we can see the close movement of predicted price with the actual price. Here also the model is trained on the training data and tested on the testing data

So we can say that the model is accurately predicting the values neer to its test data Now let's forecast the future price

Forecasted future price for Food & Beverage in Forum Istanbul mall for the next 10 days

Date	Forecasted_Values
2023-03-09	130.224768
2023-03-10	132.037882
2023-03-11	165.563121
2023-03-12	181.124706
2023-03-13	159.342146
2023-03-14	179.615145
2023-03-15	192.287913
2023-03-16	204.324409
2023-03-17	203.848967
2023-03-18	210.542187

By using this model, we can predict future purchase amounts and make proactive arrangements for necessary stock, cash balance, and other resources based on that foresight.This is not limited to 10 days we can also run it for future dates and adding the future data to the database will increase the accuracy of the model for predicting further

CONCLUSION

In this analysis, we started with analysing the customer shopping dataset of various malls in Istanbul city to gain insights about the customer shopping trends correlations existing among the products in the dataset. We found much information from the dataset, which can lead to better business for the malls in the city in the future.
We generally focused on answering questions like “Which age group is making most of the purchases?” “Which product has the most sales?” etc., to help the retailers make necessary strategies to make their business more profitable. And we were successfully done with it, we extracted all the possible trends out of the dataset to help them.

We also generated an ARIMA model which uses time series analysis to forecast the future purchase rate. This is a very useful tool that will give insights to purchasing behaviour in the future. Using this model, the retailers can understand the range of purchase amounts they will get on a particular day in the future. Keeping this knowledge, they can make the necessary arrangements and build a strategy to maximise their profit.

This study uses the customer shopping dataset of Istanbul City. Similarly, we can use any shopping dataset and can extract information like this to improve their business and using ARIMA model a forecasting model can be built on that dataset too for predicting the future purchase rates.

REFERENCES

https://machinelearningmastery.com/arima-for-time-series-forecasting-with-python/

https://towardsdatascience.com/an-introduction-to-time-series-analysis-with-arima-a8b9c9a961fb

https://neptune.ai/blog/arima-sarima-real-world-time-series-forecasting-guide

https://www.analyticsvidhya.com/blog/2021/06/statistical-tests-to-check-stationarity-in-time-series-part-1/#:~:text=The%20ADF%20test%20is%20used,time%20series%20is%20non%2Dstationary.

https://gilberttanner.com/blog/introduction-to-data-visualization-inpython/

https://www.machinelearningplus.com/time-series/augmented-dickey-fuller-test/

https://www.analyticsvidhya.com/blog/2021/08/effective-data-visualization-techniques-in-data-science-using-python/