CHRIST
(DEEMED TO BE UNIVERSITY)
BENGALURU · INDIA

CASE STUDY

Exploring Consumer Spending Patterns in Istanbul's Shopping Malls

by

Champaka R (2348021)

Jarin JV (2348031)

Shirley Sharon C (2348059)

Lithish R (2348073)

A project report submitted in partial fulfillment of the requirements for the award of degree of MSc. Data Science of the CHRIST (Deemed to be University)

August - 2023

Tables and Figures

Table of Content

Exploring Consumer Spending Patterns in Istanbul's Shopping Malls: A Data-Driven Analysis

Introduction

The retail landscape in Istanbul has been evolving with changing consumer preferences and the emergence of new shopping destinations. Understanding customer shopping behavior has become crucial for businesses and shopping malls to stay competitive and meet consumer demands effectively. This case study aims to conduct a comprehensive analysis of shopping data collected from 10 different shopping malls in Istanbul over the period from 2021 to 2023. The dataset covers diverse age groups and genders, providing valuable insights into shopping habits in the city.

Description of data

This report illustrates an analysis of shopping data in Istanbul. The data is gathered from 10 different shopping malls between 2021 and 2023. Data is collected from various age groups and genders in Istanbul to provide a comprehensive view of shopping habits and gain valuable insights into shopping patterns. The dataset comprises invoice numbers, customer IDs, age, gender, payment methods, product categories, quantity, price, order dates, and shopping mall locations. The dataset has a broad scope; it offers numerous opportunities for a detailed analysis, for example identifying trends in customer demographics, understanding, predicting customer purchase behavior, and exploring relationships between data attributes.

Problem Statement

The primary objective of this case study is to analyze customer shopping behavior in Istanbul and gain valuable insights into trends related to customer demographics, popular products, payment methods, and shopping mall preferences. By understanding and predicting customer purchase behavior, businesses can enhance the shopping experience, optimize their offerings, and make informed decisions to stay ahead in the competitive retail market.

The main aim of this data analysis is to understand and identify trends in customer purchase behavior in Istanbul based on shopping data collected from 10 different shopping malls between 2021 and 2023. The study aims to gain valuable insights into shopping patterns, customer demographics, popular products, payment methods, and shopping mall preferences in the city.

Objective

The intended purpose of data analysis is to understand and identify the trends in customer purchase behavior, analyze which products and payment methods are most frequently used by the customers, which malls customers have visited the most, and how different customer demographics come into play while they shop. Extracting such information can help make data-driven decisions, enhance the customer's experience, evaluate potential product offerings, and improve opportunity areas. To be precise, this data analysis helps determine the factors that impact customer behavior across various demographic groups in Istanbul.

Methodology

Customer Shopping Data Analysis

1. Collect shopping data from 10 shopping malls in Istanbul (2021-2023).
2. Preprocess the data by handling missing values and converting categorical variables.
3. Use descriptive statistics and visualizations to explore shopping patterns.
4. Analyze customer demographics, popular products, payment methods, and shopping mall preferences.
5. Examine correlations between variables.
6. Present findings through visual plots and concise interpretations.
7. Provide actionable recommendations for businesses based on the analysis.

In this data analysis, Python was used to perform various exploratory data analysis (EDA) tasks, including descriptive statistics, visualizations, and correlation analysis. Additionally, an Ordinary Least Squares (OLS) regression model was employed to predict and understand the impact of specific factors on customer purchase behavior.

Analysis

Examine the data source

Count columns and rows:

The original data has total 99,457 customers and 10 columns.

Preprocess data: Reformat date and group customers' age in 10's (<20 years old), 20's (20 to 30 years old), 30's (30 to 40 years old), 40's (40 to 50 years old), the rest is 60's

Preliminary descriptive statistics analysis

Table representing the summary of the data set.

Table 1: Data describe

	age	quantity	price
count	13808.000000	13808.000000	13808.000000
mean	43.527955	2.996524	688.020031
std	14.934368	1.418913	951.230454
min	18.000000	1.000000	5.230000
25%	31.000000	2.000000	40.660000
50%	43.000000	3.000000	203.300000
75%	56.000000	4.000000	1200.320000
max	69.000000	5.000000	5250.000000

Count: There are 99457 records in the data.

Age

Mean: The mean age of the customers is approximately 43.43.

Standard Deviation: The standard deviation of age is approximately 14.99 meaning the data points vary 14.99 approximately from the mean points.

Min: The minimum age of the customers in the data is 18.

25%: 25% of the data points have an age value below 30.

50% (Median): The median age is 43, which means half of the data points have an age below 43 and the other half have an age above 43.

75%: 75% of the data points have an age value below 56.

Max: The maximum age in the data is 69 years.

Quantity - Number of items purchased in the shopping category.

Mean: The mean quantity is approximately 3. On average customers have purchased 3 goods.

Standard Deviation (std): The standard deviation of quantity is approximately 1.41.

Min: The minimum quantity in the data is 1.

25%: 25% of the data points have a quantity value below 2.

50% (Median): The median quantity is 3, which means half of the data points have a quantity below 3 and the other half have a quantity above 3.

75%: 75% of the data points have a quantity value below 4.

Max: The maximum quantity in the data is 5. There exists no record of purchases made above 5 quantities.

Price

Mean: The mean price is approximately 689.26 (in rupees).

Standard Deviation: The standard deviation of price is approximately 941.18 (in rupees).

Min: The minimum price in the data is Rs 5.23.

25%: 25% of the data points have a price value below 45.45.

50% (Median): The median price is Rs 203.30, which means half of the data points have a price below Rs 203.30 and the other half have a price above Rs 203.30.

75%: 75% of the data points have a price value below Rs 1200.32.

Max: The maximum price in the data is Rs 5250.00.

Malls from which the data is collected:

array(['Kanyon', 'Forum Istanbul', 'Metrocity', 'Metropol AVM', 'Istinye Park', 'Mall of Istanbul', 'Emaar Square Mall', 'Cevahir AVM', 'Viaport Outlet', 'Zorlu Center', nan], dtype=object)

Genders of the customers visiting the malls:

array(['Female', 'Male'], dtype=object)

Various categories from which the customers have purchased:

array(['Clothing', 'Shoes', 'Books', 'Cosmetics', 'Food & Beverage', 'Toys', 'Technology', 'Souvenir'], dtype=object)

Mode of payment used for purchasing and the number of times these modes are used:

array(['Credit Card', 'Debit Card', 'Cash', nan], dtype=object)

'Credit Card': 34931, 'Debit Card: 20079', 'Cash': 44447

44% of purchases are made through cash payment. While 35% through Credit card and the rest have used their debit cards for purchases.

Table represents the number of purchases made in different shopping malls.

Table 2: Malls and their purchases

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

Mall of Istanbul and Kanyon foresees maximum customers between the years 2021 and 2023. An equal spread of shoppers visited the Cevahir AVM, Istanbul, Viaport Outlet, Emaar Square Mall.

<u>Represents the count of female and male shoppers:</u>

Female : 59482

Male : 39975

The dataset has more female customers (59,482) compared to male customers (39,975). Thus, these malls attract more female customers.

<u>Represents the count of shoppers in each category:</u>

Table 3: Count of the Shoppers in each Category

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

Clothing seems to be the most popular product category, followed by Cosmetics and Food & Beverage.

Toys, Shoes, Souvenir, Technology, and Books are also significant product categories, though with relatively fewer purchases.

Range of prices:

range: 5250.0 - 5.23

The prices are quite spread out across this range. The prices column in the data set are affected by extreme outliers.

Standard Deviation of prices:

941.184567215467

The standard deviation of 941.18 means that the prices in the dataset tend to deviate from the mean price by around 941.18 units, on average. Thus, there is a wider range of prices in the data set.

Variance:

885828.389564566

Variance is another statistical measure that quantifies the spread or dispersion of a dataset.

The variance of 885828.39 means that, on average, the squared differences between each price and the mean price in the dataset sum up to 885828.39.

The standard deviation we calculated (approximately 941.18) is the square root of the variance (885828.39).

Graphical Representaion of Data

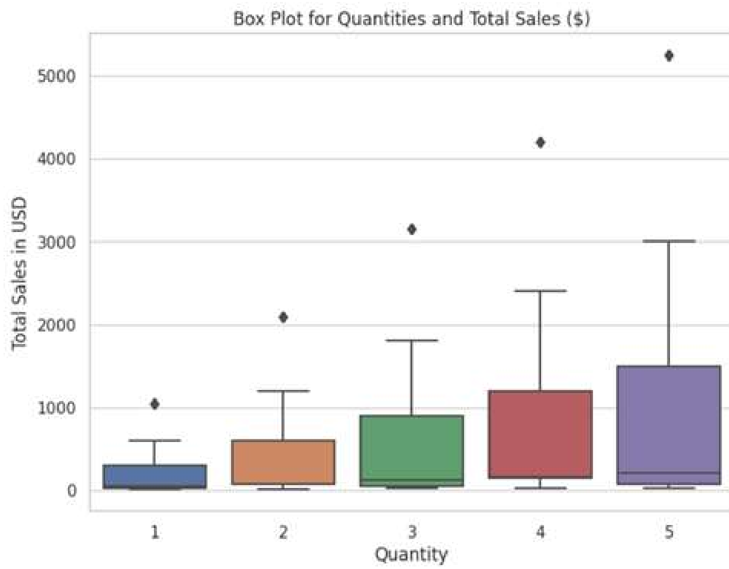

Fig 1: Boxplot for quantities and price

The boxes in the graph show positive skewness for the considered data and presence of outliers. ‘Quantity 5’ has more dispersed data than others along with the highest maximum value. Hence, as quantity increases, price increases too.

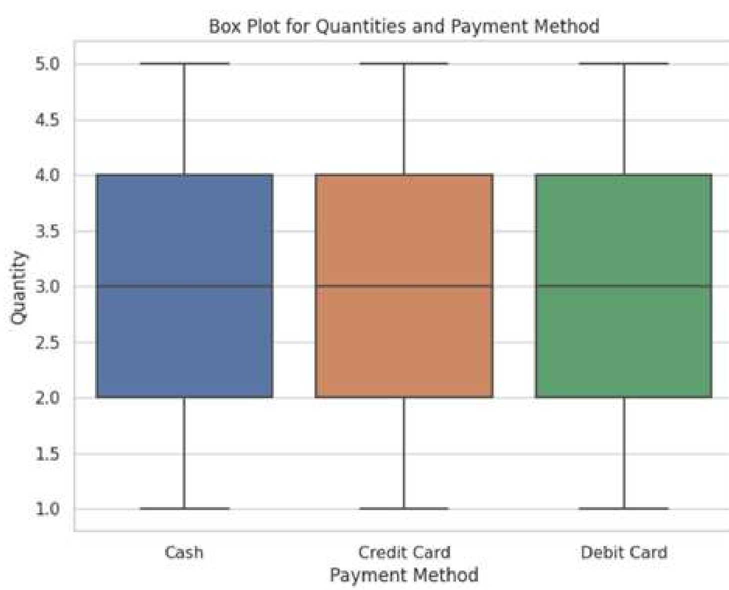

Fig 2: Boxplot for Quantities and Payment method

All three methods of payment used in the considered data has equal dispersion and levels of data, i.e, payment through cash, credit card or debit card were preferred equally.

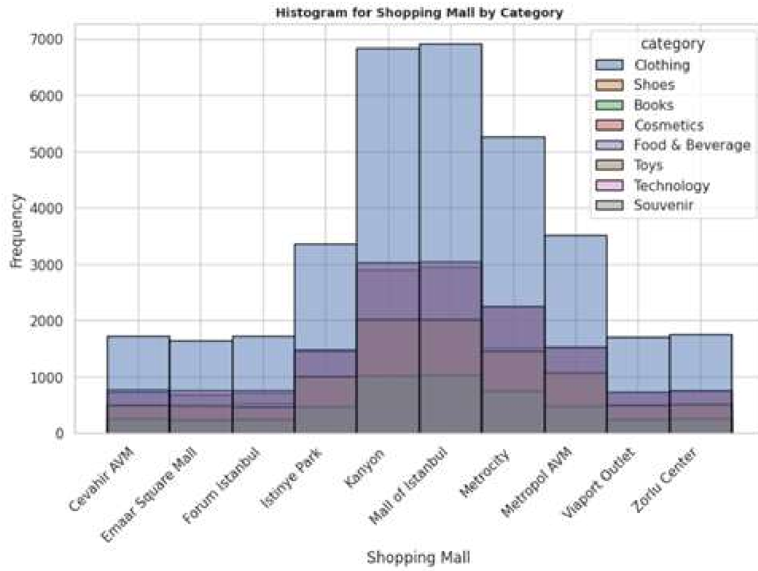

Fig 3: Histogram for shopping mall by category

Clearly, 'clothing' is the most frequently shopped category in shopping malls, followed by 'technology' and 'cosmetics'.

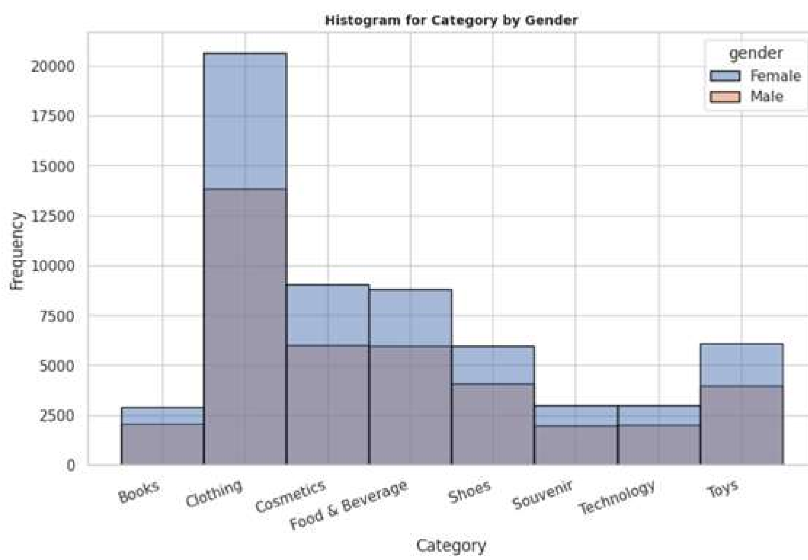

Fig 4: Histogram for Gender by category

All the categories in the shopping mall have been visited by females more frequent number of times. Especially 'clothing', where females have visited upto 20,000 times.

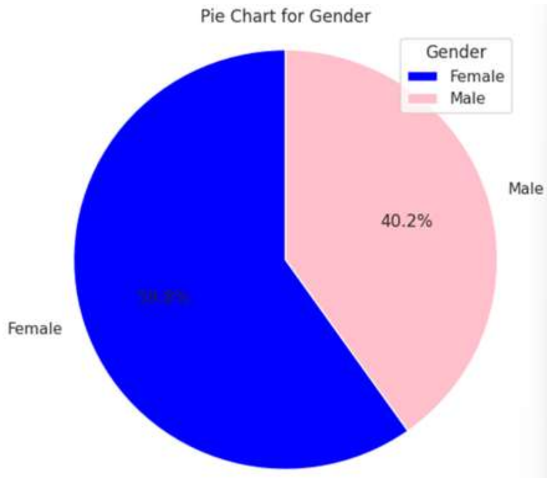

Fig 5: Pie chart for Gender

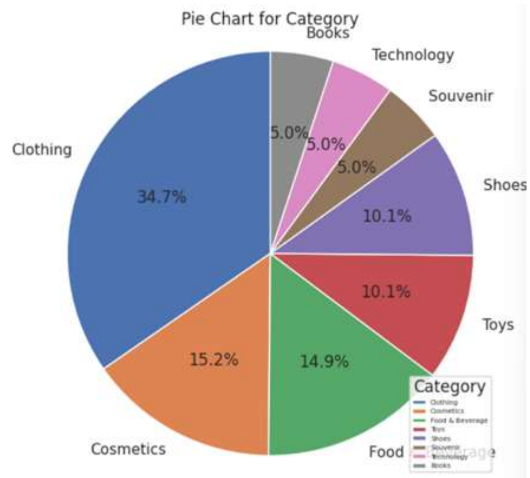

Fig 6: Pie Chart for Categories

About 60% of the customers are female. Clothing is the most frequently visited category with about 35% of the customers shopping for clothes.

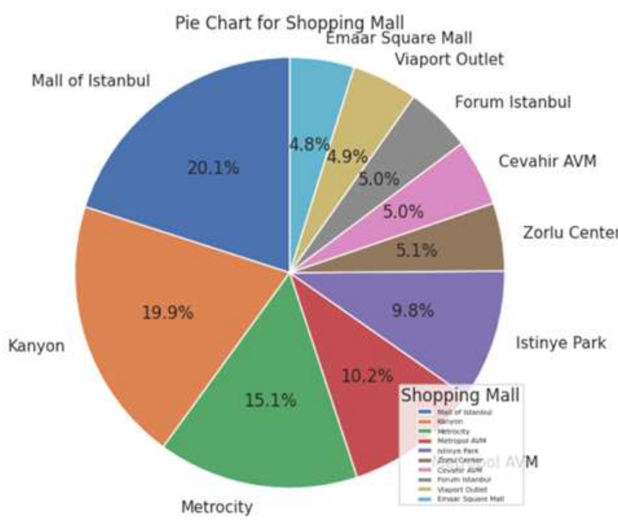

Fig 7: Pie Chart for Shopping Mall

Mall of Istanbul is the most frequently visited shopping malls, followed by Kanyon and Metrocity.

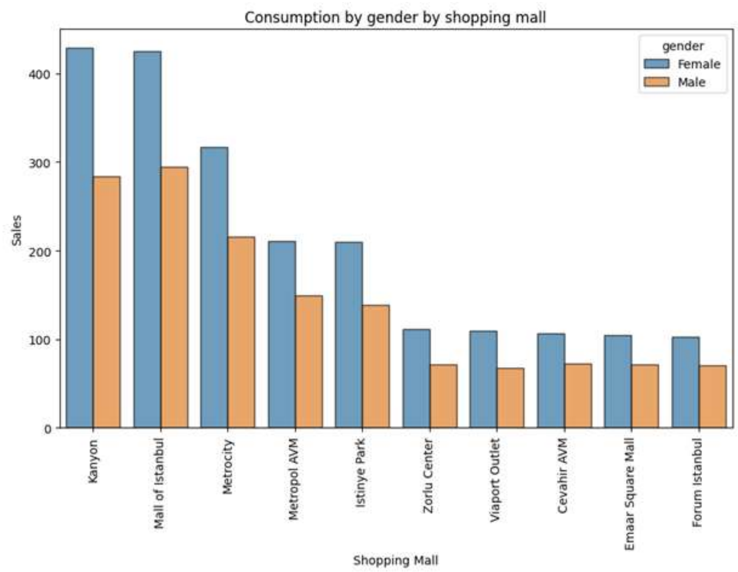

Fig 8: Bar Chart for Gender by Shopping Mall

In all the shopping malls females have visited a greater number of times than males by considerable amounts.

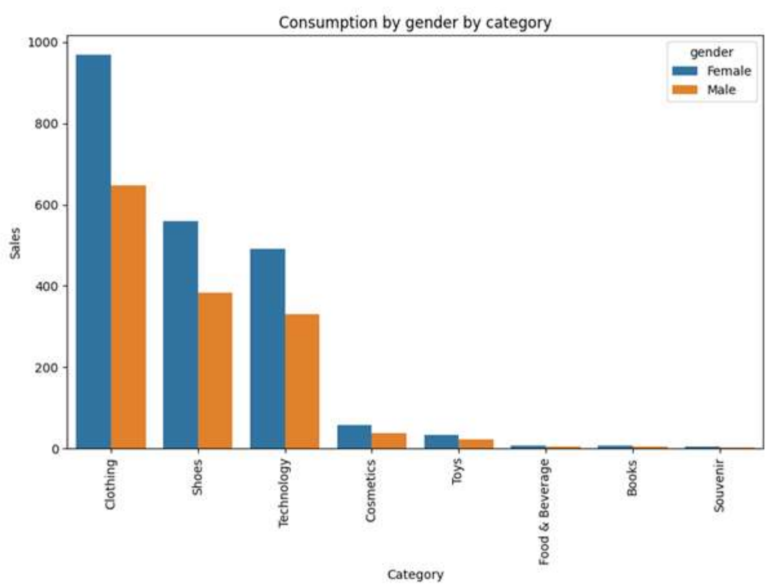

Fig 9: Bar Chart for Gender by Category

Clothing is the most frequently visited category and again females have shopped more than males in each category.

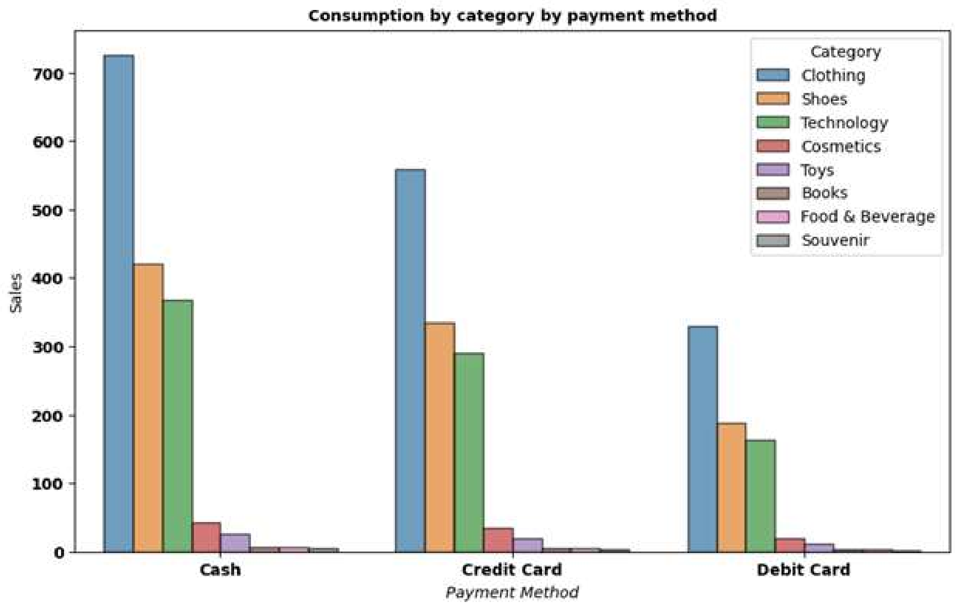

Fig 10: Bar Chart for Payment Methods by Category.

Although all categories are equally distributed for each payment methods, we can see that payment through cash for clothing is more frequent than other payment methods.

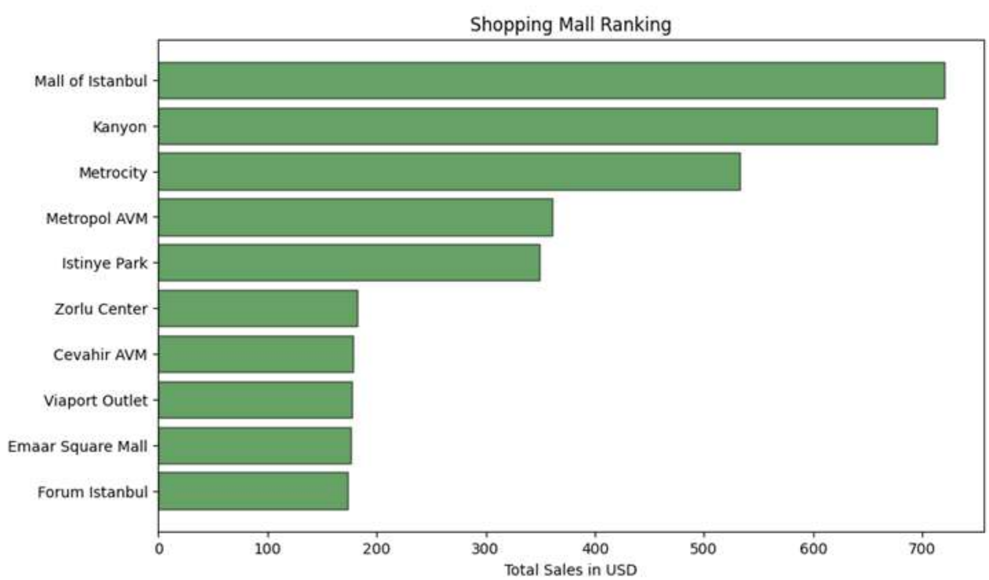

Fig 11: Bar Chart for ranking Shopping Mall based on Price.

Mall of Istanbul and Kanyon are the most frequently visited shopping malls with a total price of about 700 USD, thereby being the most popular shopping malls of the locality.

Correlations between various combinations of variables

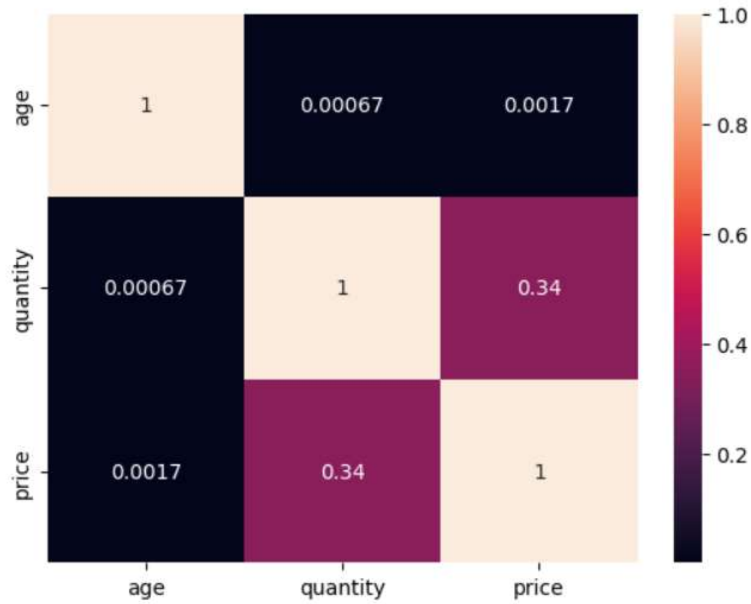

Fig 12: Correlation Heatmap

The graph shows that there is a positive correlation between the age and price of the product. This means that as the age of the product increases, the price of the product also tends to increase. However, the correlation is not very strong, so there are some outliers in the data. For example, there are a few points that show that the price of a product decreased as the age of the product increased.

There is also a positive correlation between the quantity and price of the product. This means that as the quantity of the product increases, the price of the product also tends to increase. However, the correlation is not as strong as the correlation between the age and price of the product.

Overall, the graph shows that there is a positive correlation between the age, quantity, and price of a product. However, the correlations are not very strong, so there is a lot of variation in the data.

Table 4: Correlation Coefficients

Variable 1	Variable 2	Correlation Coefficient
Age	Price	0.34
Quantity	Price	0.0017
Age	Quantity	0.00067

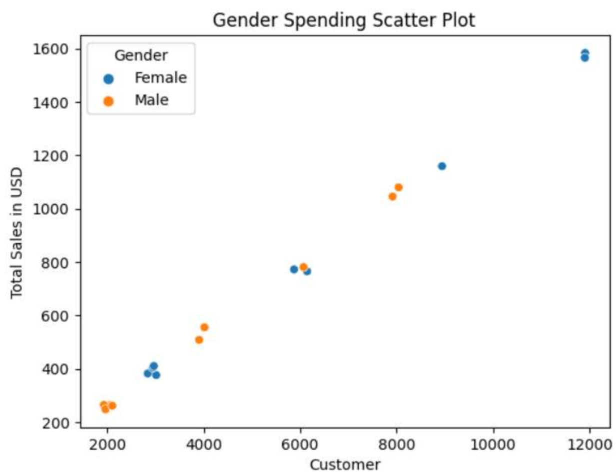

Fig 13: Gender Spending Scatter Plot

As expected, females in Istanbul have a stronger purchasing tendency than males. This can come from Young Single ones and Married Females who go shopping for the whole household.

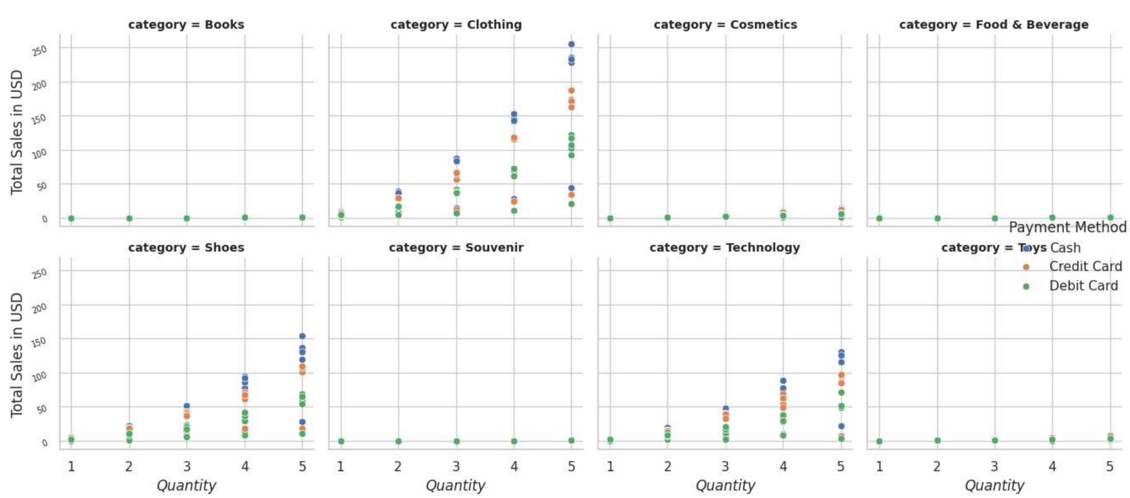

Fig 14: Purchasing Behavior by Category Scatter Plot

Cash is still the primary payment method, with the highest usage in both numbers of customers and sales amount.

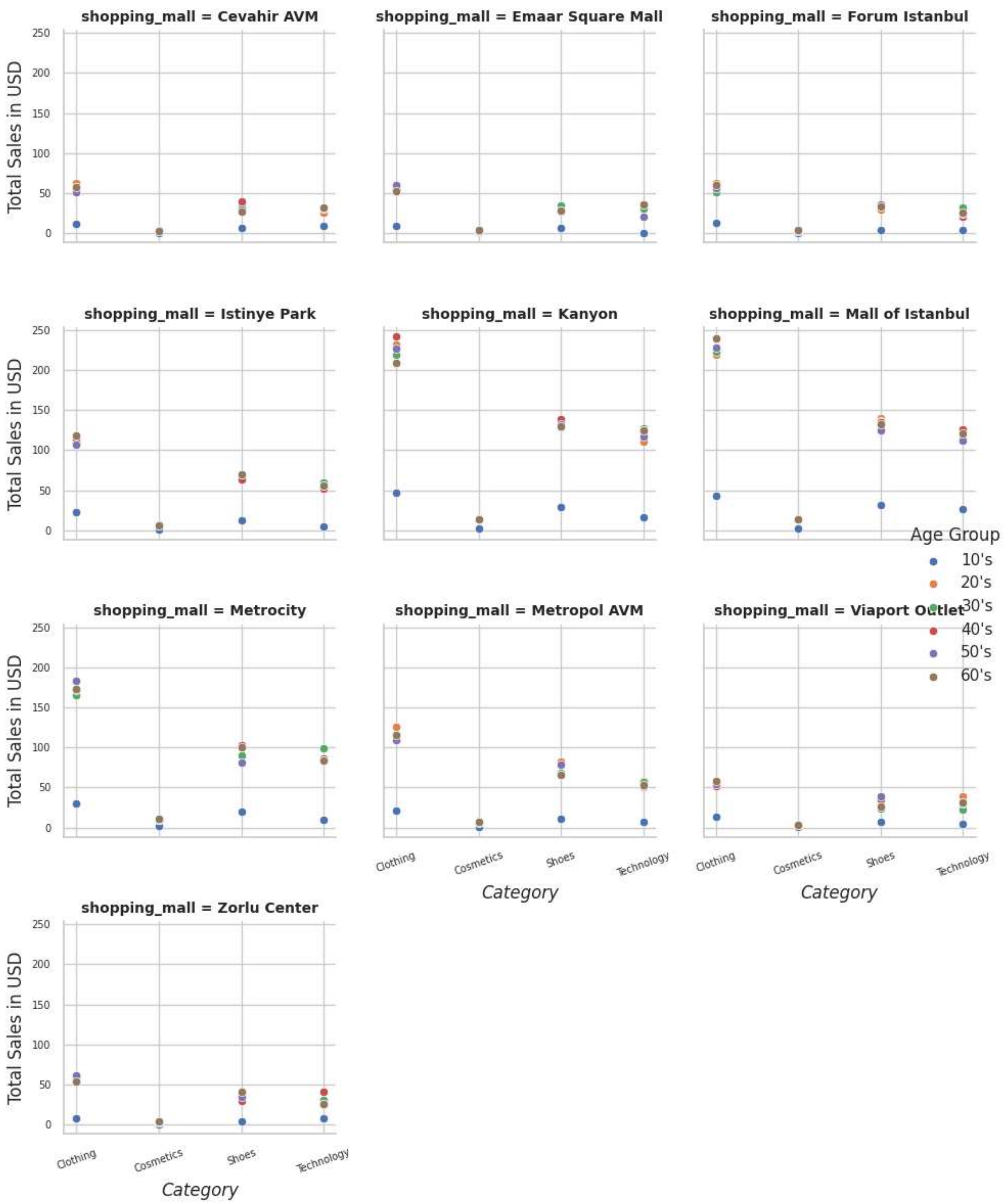

Fig 15: Shopping malls preferred by customers by age group filtering top 4 categories 1 = 1,000$

All 4 categories have a similar trend in every mall. In detail, the ‘40s (between 40 and 50 years old) customers and the ‘60s (more than 60 years old) purchase aggressively, followed by the ‘20s (between 20 and 30 years old) and the ‘50s (between 50 and 60 years old); yet the pattern varies between stores - for example: in Mall of Istanbul, the ‘40s and the ‘60s share the same power, while in Kanyon, the ‘40s has the more substantial purchasing power, which may be influenced by the demographic surrounding each mall.

Regression

Hypothesis 1: Age and Gender Impact Shopping Behavior

Plot analysis: The quantity bought increases with age for Males, whereas it decreases for Females.Inference: From the output of the regression analysis for Hypothesis 1 the regression model results suggest that the 'age' and 'gender' variables do not have a statistically significant impact on the 'quantity'. The R-squared is close to zero, indicating that the model explains very little of the variability in the 'quantity'. Additionally, the p-values for both 'age' and 'gender' are high, suggesting that these variables are not statistically significant predictors of 'quantity' in this model. The residuals are not normally distributed, but there is no significant autocorrelation or multicollinearity. Overall, the model does not seem to be a good fit for the data,

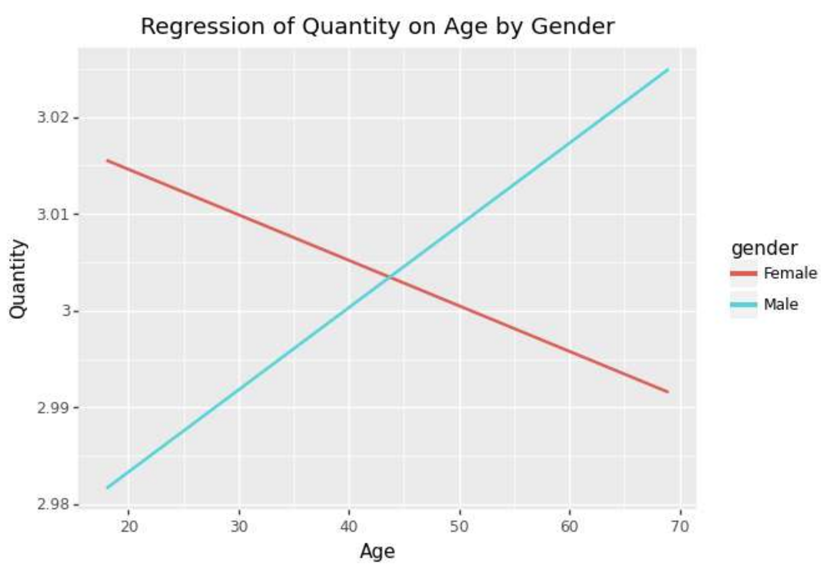

Fig 16: Regression of Quantity on Age by Gender

Hypothesis 2: Payment Method Affects Purchase Amount

Plot Analysis: Payment Method (Credit Card or Debit Card) does not significantly impact Purchase Amount in this dataset, as it is not a statistically significant predictor.the regression model results suggest that the 'method_Cash', 'method_Credit Card', and 'method_Debit Card' variables do not have a statistically significant impact on the 'price'. The R-squared and adjusted R-squared are close to zero, indicating that the model does not explain much of the variability in the 'price'. The F-statistic and p-values also suggest that the model is not statistically significant. Additionally, there might be potential issues with the data, such as large coefficients and multicollinearity. Further investigation and possibly different independent variables are needed to build a meaningful and statistically significant model for predicting 'price'.

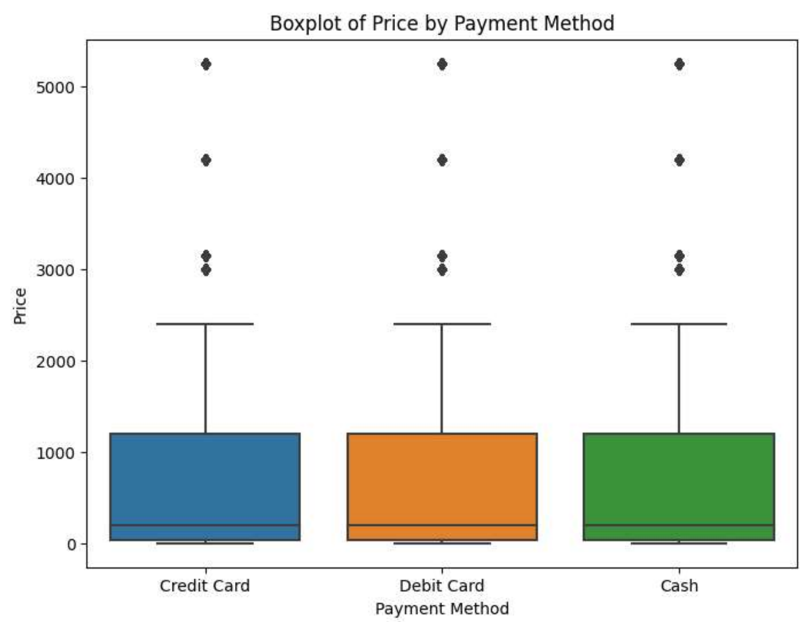

Fig 17: Boxplot of price by the payment method

Hypothesis 3: Shopping Mall Location Influences Purchase Frequency

The results of the analysis show that the location of shopping malls does not seem to have a significant impact on the purchase frequency. The R-squared value, which measures how well the model explains the variation in the data, is very close to zero (0.000). This indicates that the location of the shopping malls does not explain much of the variation in the purchase frequency.

By looking at the coefficients for each shopping mall location, none of them are statistically significant. This means that the differences in purchase frequency between different malls are likely due to random chance rather than being driven by the specific location of the malls.
We cannot conclude that the location of shopping malls has a meaningful influence on the frequency of purchases made by customers. Other factors not included in this analysis may be more relevant in explaining variations in purchase frequency.

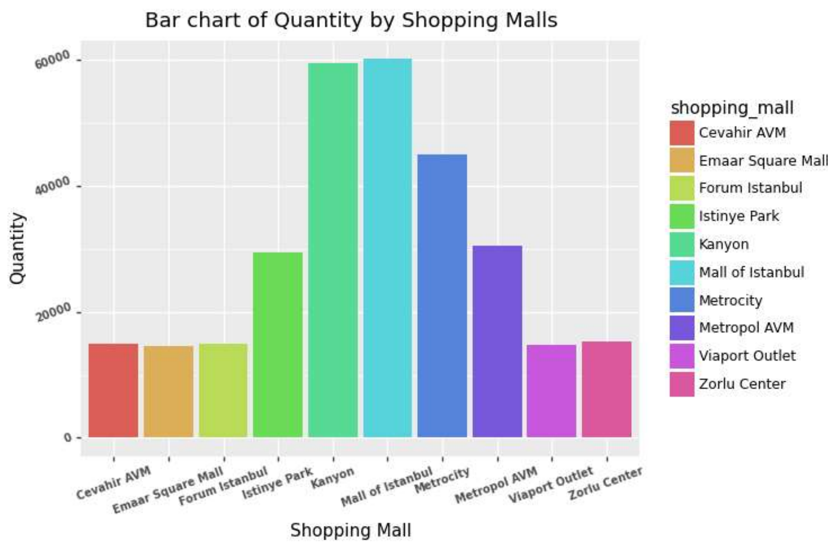

Fig 18: Bar Chart of Quantity by Shopping Malls

<u>Hypothesis 4: Product Category Affects Purchase Amount</u>

The results of the analysis show that the product category does have a significant impact on the purchase amount. The R-squared value, which measures how well the model explains the variation in the data, is 0.723. This indicates that the product category explains approximately 72.3% of the variation in purchase amounts.
The coefficients for each product category represent the average difference in purchase amount compared to a reference category (which is not explicitly mentioned in the output). For example:

- Clothing: On average, customers spend $901.08 more on clothing purchases compared to the reference category.
- Technology: On average, customers spend $3156.94 more on technology-related purchases compared to the reference category.
- Books: On average, customers spend $45.57 more on book purchases compared to the reference category.

By the p-values associated with each coefficient are very close to zero (0.000). This indicates that the differences in purchase amounts between different product categories are statistically significant, meaning they are unlikely to have occurred by chance.

We can conclude that the product category significantly affects the amount of money spent on purchases. Customers tend to spend varying amounts depending on the type of product they are buying, and the product category is a strong predictor of the purchase amount.

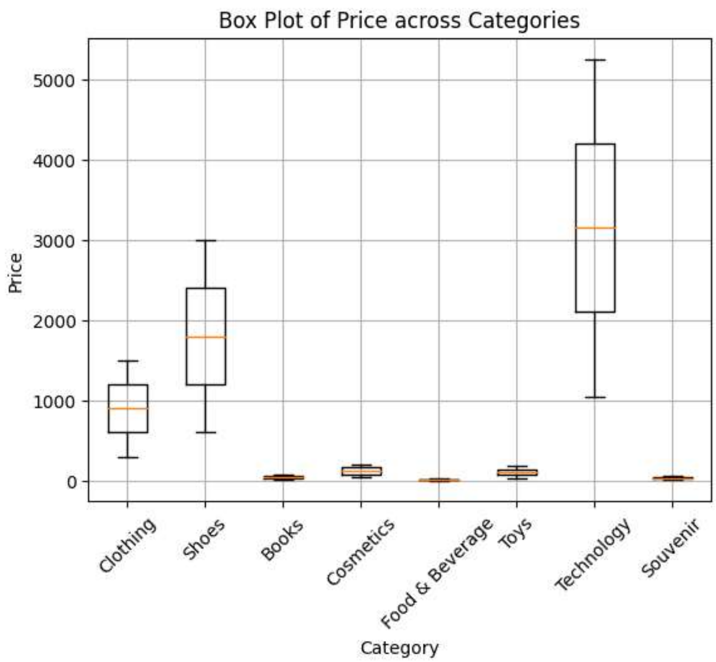

Fig 19: Box plot of Price across Categories

Solutions

In this data-driven analysis, we studied how people shop in Istanbul's malls from 2021 to 2023. We looked at different age groups and genders to understand their shopping habits. The data included information like age, gender, what they bought, how much they bought, and how they paid.

We found that clothing, cosmetics, and food & beverage are popular categories. Mall of Istanbul and Kanyon are the most visited malls.

Women shop more often than men in all categories and malls. Most people pay with cash, and fewer use credit or debit cards.

As people get older or buy more things, they tend to spend more money. The type of product they buy also affects how much they spend.

Based on our findings, we suggest businesses focus on marketing to specific age groups and genders, offer popular products, promote digital payments, improve less-visited malls, and use data to make better decisions.

By following these suggestions, businesses can improve customer satisfaction, increase sales, and stay competitive in the retail market. Analyzing data is essential for understanding customers and making smart choices for success.

Implementation

<u>Examining the dataset:</u>

We used a special set of tools called "numpy libraries" to look closely at the data and make it easier to understand and analyze. These libraries helped us turn the data into useful information so we could study and make sense of it more effectively.

```python
import numpy as np

customerDF['invoice_date'] = pd.to_datetime(customerDF['invoice_date'], format='%d/%m/%Y')

customerDF['total_sale'] = customerDF['price'] * customerDF['quantity']

customerDF['age_group'] = np.select(
    [
        customerDF['age'] < 20,
        customerDF['age'] < 30,
        customerDF['age'] < 40,
        customerDF['age'] < 50,
        customerDF['age'] < 60,
    ],
    ["10's", "20's", "30's", "40's", "50's"],
    default="60's"
)

selected_columns = [
    'shopping_mall', 'invoice_date', 'invoice_no', 'customer_id', 'gender', 'age',
    'age_group', 'category', 'quantity', 'price', 'total_sale', 'payment_method'
]

customerDF = customerDF[selected_columns]

customerDF.head()
```

	shopping_mall	invoice_date	invoice_no	customer_id	gender	age	age_group	category	quantity	price	total_sale	payment_method
0	Kanyon	2022-08-05	I138884	C241288	Female	28	20's	Clothing	5	1500.40	7502.00	Credit Card
1	Forum Istanbul	2021-12-12	I317333	C111565	Male	21	20's	Shoes	3	1800.51	5401.53	Debit Card
2	Metrocity	2021-11-09	I127801	C266599	Male	20	20's	Clothing	1	300.08	300.08	Cash
3	Metropol AVM	2021-05-16	I173702	C988172	Female	66	60's	Shoes	5	3000.85	15004.25	Credit Card
4	Kanyon	2021-10-24	I337046	C189076	Female	53	50's	Books	4	60.60	242.40	Cash

Fig 20: Data Examination

Descriptive analysis:

We looked at the dataset in detail to understand its different parts. We used a handy tool called "dataframe" from the pandas library to help us with this. With the help of this dataframe, we were able to describe and summarize the data, which gave us valuable information about its contents.

```
1 customerDF.describe()
```

	age	quantity	price
count	13808.000000	13808.000000	13808.000000
mean	43.527955	2.996524	688.020031
std	14.934368	1.418913	951.230454
min	18.000000	1.000000	5.230000
25%	31.000000	2.000000	40.660000
50%	43.000000	3.000000	203.300000
75%	56.000000	4.000000	1200.320000
max	69.000000	5.000000	5250.000000

```
1 df.shopping_mall.value_counts()
```

```
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
Name: shopping_mall, dtype: int64
```

```
1 g=df.gender.value_counts()
2 c=df.category.value_counts()
3
4 print(g)
5 print(c)
```

```
Female    59482
Male      39975
Name: gender, dtype: int64
Clothing           34487
Cosmetics          15097
Food & Beverage    14776
Toys               10087
Shoes              10034
Souvenir            4999
Technology          4996
Books               4981
Name: category, dtype: int64
```

Fig 21: Descriptive analysis

Graphical Representation:

We used libraries such as matplotlib and seaborn, to create visual representations of the data. With matplotlib, we made different types of plots, like bar graphs or line charts, to show patterns and relationships in the data. Seaborn was used specifically for creating a heatmap, which helped us see how different elements in the data were related to each other, visually highlighting any correlations. These graphical representations made it easier for us to understand the data and find important insights.

Regression:

We performed a type of analysis called regression, which helps us understand the relationships between variables. We did this on four different cases. To build our models, we used a tool called "sklearn" that provides helpful functions. Specifically, we used the OLS (Ordinary Least Squares) method for modeling.

In our dataset, we had some qualitative data, which means data with categories like gender and payment method. To work with this data in our models, we needed to convert it into a numerical representation. For this purpose, we used a technique called "label encoder" to assign unique numbers to each category. This allowed us to use all the data in our regression models effectively.

```python
le = LabelEncoder()
customerDF['gender'] = le.fit_transform(customerDF['gender'])

X = customerDF[['age', 'gender']]
X = sm.add_constant(X)
y = customerDF['quantity']

shopping_model1 = sm.OLS(y, X).fit()


print(shopping_model1.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:               quantity   R-squared:                       0.000
Model:                            OLS   Adj. R-squared:                 -0.000
Method:                 Least Squares   F-statistic:                   0.02294
Date:                Sat, 29 Jul 2023   Prob (F-statistic):              0.977
Time:                        08:38:50   Log-Likelihood:            -1.7551e+05
No. Observations:               99457   AIC:                         3.510e+05
Df Residuals:                   99454   BIC:                         3.511e+05
Df Model:                           2
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          3.0009      0.014    210.998      0.000       2.973       3.029
age         6.281e-05      0.000      0.210      0.834      -0.001       0.001
gender        -0.0004      0.009     -0.041      0.967      -0.018       0.018
==============================================================================
Omnibus:                   344979.444   Durbin-Watson:                   1.997
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             6959.159
Skew:                          -0.001   Prob(JB):                         0.00
Kurtosis:                       1.704   Cond. No.                         149.
==============================================================================
```

Hypothesis 1

```python
X = pd.get_dummies(cdf['payment_method'], prefix='method')
X = sm.add_constant(X)
y = customerDF["price"]

shopping_model1 = sm.OLS(y, X).fit()

print(shopping_model1.summary())

plt.figure(figsize=(8, 6))
sns.boxplot(x='payment_method', y='price', data=cdf)
plt.xlabel("Payment Method")
plt.ylabel("Price")
plt.title("Boxplot of Price by Payment Method")
plt.show()
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  price   R-squared:                      -0.000
Model:                            OLS   Adj. R-squared:                 -0.000
Method:                 Least Squares   F-statistic:                   -0.4947
Date:                Sat, 29 Jul 2023   Prob (F-statistic):               1.00
Time:                        08:44:02   Log-Likelihood:            -8.2212e+05
No. Observations:               99457   AIC:                         1.644e+06
Df Residuals:                   99453   BIC:                         1.644e+06
Df Model:                           3
Covariance Type:            nonrobust
=======================================================================================
                          coef    std err          t      P>|t|      [0.025      0.975]
---------------------------------------------------------------------------------------
const               -6.423e+14   6.12e+14     -1.050      0.294   -1.84e+15    5.57e+14
method_Cash          6.423e+14   6.12e+14      1.050      0.294   -5.57e+14    1.84e+15
method_Credit Card   6.423e+14   6.12e+14      1.050      0.294   -5.57e+14    1.84e+15
method_Debit Card    6.423e+14   6.12e+14      1.050      0.294   -5.57e+14    1.84e+15
==============================================================================
Omnibus:                    46070.151   Durbin-Watson:                   1.995
Prob(Omnibus):                  0.000   Jarque-Bera (JB):           241377.666
Skew:                           2.247   Prob(JB):                         0.00
Kurtosis:                       9.168   Cond. No.                     4.80e+14
==============================================================================
```

Hypothesis 2

```python
shopping_model3 = ols(formula='quantity ~ shopping_mall', data=cdf).fit()
print(shopping_model3.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:               quantity   R-squared:                       0.000
Model:                            OLS   Adj. R-squared:                 -0.000
Method:                 Least Squares   F-statistic:                    0.4002
Date:                Sat, 29 Jul 2023   Prob (F-statistic):              0.936
Time:                        08:39:36   Log-Likelihood:            -1.7551e+05
No. Observations:               99457   AIC:                         3.510e+05
Df Residuals:                   99447   BIC:                         3.511e+05
Df Model:                           9
Covariance Type:            nonrobust
===================================================================================================
                                      coef    std err          t      P>|t|      [0.025      0.975]
---------------------------------------------------------------------------------------------------
Intercept                           2.9952      0.020    149.747      0.000       2.956       3.034
shopping_mall[T.Emaar Square Mall]  0.0189      0.029      0.663      0.507      -0.037       0.075
shopping_mall[T.Forum Istanbul]     0.0070      0.028      0.248      0.804      -0.049       0.063
shopping_mall[T.Istinye Park]       0.0173      0.025      0.703      0.482      -0.031       0.065
shopping_mall[T.Kanyon]             0.0042      0.022      0.188      0.851      -0.040       0.048
shopping_mall[T.Mall of Istanbul]   0.0191      0.022      0.854      0.393      -0.025       0.063
shopping_mall[T.Metrocity]         -0.0045      0.023     -0.193      0.847      -0.050       0.041
shopping_mall[T.Metropol AVM]       0.0094      0.024      0.386      0.699      -0.038       0.057
shopping_mall[T.Viaport Outlet]    -0.0005      0.028     -0.017      0.986      -0.056       0.055
shopping_mall[T.Zorlu Center]       0.0066      0.028      0.234      0.815      -0.049       0.062
==============================================================================
Omnibus:                   344605.612   Durbin-Watson:                   1.997
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             6958.699
Skew:                          -0.001   Prob(JB):                         0.00
Kurtosis:                       1.704   Cond. No.                         15.5
==============================================================================
```

Hypothesis 3

```python
shopping_model4 = sm.OLS(cdf['price'], pd.get_dummies(cdf['category']))

results = shopping_model4.fit()
print(results.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  price   R-squared:                       0.723
Model:                            OLS   Adj. R-squared:                  0.723
Method:                 Least Squares   F-statistic:                 3.706e+04
Date:                Sat, 29 Jul 2023   Prob (F-statistic):               0.00
Time:                        08:39:55   Log-Likelihood:            -7.5831e+05
No. Observations:               99457   AIC:                         1.517e+06
Df Residuals:                   99449   BIC:                         1.517e+06
Df Model:                           7
Covariance Type:            nonrobust
===================================================================================
                      coef    std err          t      P>|t|      [0.025      0.975]
-----------------------------------------------------------------------------------
Books              45.5686      7.021      6.491      0.000      31.808      59.329
Clothing          901.0840      2.668    337.718      0.000     895.854     906.314
Cosmetics         122.4486      4.033     30.364      0.000     114.545     130.353
Food & Beverage    15.6719      4.076      3.845      0.000       7.683      23.661
Shoes            1807.3886      4.947    365.384      0.000    1797.693    1817.084
Souvenir           34.8943      7.008      4.979      0.000      21.159      48.630
Technology       3156.9355      7.010    450.338      0.000    3143.196    3170.675
Toys              107.7332      4.934     21.837      0.000      98.064     117.403
==============================================================================
Omnibus:                    11900.643   Durbin-Watson:                   2.002
Prob(Omnibus):                  0.000   Jarque-Bera (JB):           131104.730
Skew:                          -0.009   Prob(JB):                         0.00
Kurtosis:                       8.625   Cond. No.                         2.63
==============================================================================
```

Hypothesis 4

Fig 22: Hypothesis (Regression)

Conclusion

In conclusion, our data-driven analysis of consumer spending patterns in Istanbul's shopping malls from 2021 to 2023 has provided valuable insights. We found that clothing, cosmetics, and food & beverage are the most popular product categories. Mall of Istanbul and Kanyon are the most visited shopping malls.

We observed that women tend to shop more frequently than men, and most customers prefer to pay with cash. However, the method of payment does not significantly impact the purchase amount.

Age has a mild influence on shopping behavior, with older customers tending to spend more. Product category, on the other hand, has a significant impact on the purchase amount, with customers spending more on technology-related products compared to other categories.

Based on these findings, businesses can tailor their marketing strategies to target specific age groups and genders. They can also focus on offering popular product categories to attract more customers. Encouraging digital payment options and improving the appeal of less-visited malls can enhance the overall shopping experience.

By implementing data-driven decisions, businesses can increase customer satisfaction, optimize their offerings, and stay competitive in the ever-evolving retail market. Analyzing customer behavior and preferences through data analysis will be instrumental in driving success in the retail industry in Istanbul.