SHOPPING MALL TRENDS IN ISTANBUL

4/20/2023

SPRING 2023

Submitted by: Alejandra Mejia, Sarah Alikhan, Sri Ramya Simhadri, Swetha Chukka, Valeria Latorraca

Submitted to: Kutsal Dogan

ISM 6562 - Data Visualization

Table of Contents

ABSTRACT

This report intends to analyze the different consumer behavior trends that can serve to identify the main customer segments and their main interests in 10 big shopping malls in Istanbul. Throughout the report we will answer different segmentation and consumer behavior questions through data visualization of the data. We will explain the design process, our key findings and at last provide conclusions on all of our analysis.

INTRODUCTION

Project Purpose:

We want to understand and visualize different customer segments per mall, the amount of money spent on different categories by different genders and age groups, and product analysis of shopping malls in Istanbul. We expect that after finishing the project we will be able to fully understand the sales dynamics of each mall and how each one is different from each other. We also expect to be able to understand how the different customer segments change by mall and if the consumption behavior changes depending on the mall they go to.

Project Description:

With the aid of Tableau and potentially Excel for some data cleaning, we want to build visualizations that show top product categories per mall, and customer segments. We will also be able to build time series charts that help understand how the sales vary during the year and which events can contribute to these fluctuations. Finally, we would like users to be able to conclude the dynamics of each mall and what marketing, sales, and strategy efforts can be done to improve the profitability of their businesses.

A description of the intended users and tasks:

1. **Brands opening up new outlets.**
   Companies wanting to enter the Turkish market to open up an outlet and want to determine which mall(s) they would want to go to.
2. **Shopping malls management.**
   To determine how they are performing against other competitors, their strengths, weak season, reasons for weaknesses etc.

3. **Advertisers/Marketing companies.**
   The marketing companies or brands wanting to put up Kiosks or run any promotional activity would like to determine which mall and which day and seasons would be the ideal place and timing for their brands/promotion based on shopper profiles, peak traffic days.
4. **Credit and store card companies.**
   Financial institutions may use the data to design or offer promotions to a specific mall, geography, and/or outlets link discounts to promote usage of credit cards versus debit cards.

PREVIOUS WORK

From the information found on Kaggle, there have been at least 33 previous different projects undertaken using this dataset conducting EDA (exploratory data analysis), trend analysis, regressions analysis and models. But all the work previously done using this dataset has been using python.

DESIGN PROCESS

Data Set Selection

The original dataset was obtained from www.kaggle.com. Kaggle is the world's largest data science community with powerful tools and resources. The following elements were taken into consideration while selecting the dataset.

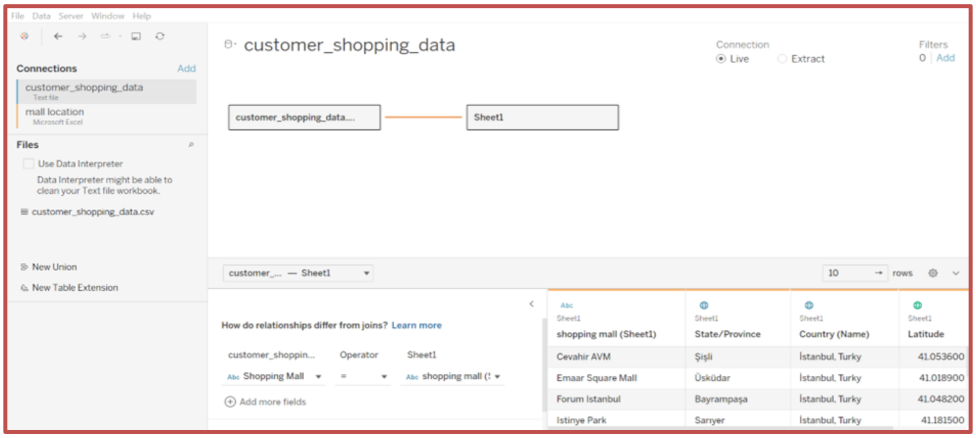
- High volume: There were a total of 99,458 observations in the dataset.
- Historical: At least 2 full years of complete data were available for analysis.
- Consistent: The data across the board was in a consistent format with similar data formats.
- Multivariate & dimensionally structured: The data contained multiple variables such as the amount spent, quantities purchased, payment methods used, etc. Similar data were available for each mall over a period of 2.5 years.
- Atomic, Clean & Clear: It was a clean dataset without any missing value, multi-variate cell, format, or data-related issues. So, most of our time and focus went into creating visualizations.
- Richly segmented: It was decently segmented data. Some additional segmentation we would have liked to have been missing but derived from the purpose of our analysis.

Data Cleaning

The initial dataset downloaded contained information about 10 Istanbul malls. It had 10 variables that included: Invoice number, customer ID, invoice date, Mall name, customer age, customer gender, average unit price, product quantity, product category, and payment method. In Excel, we created two additional variables that were useful for our analysis. First, we calculated the total order value by multiplying the average price by the quantity. Then we created the variable *Weekday* to be able to explore sales by day of the week.

Data Exploration

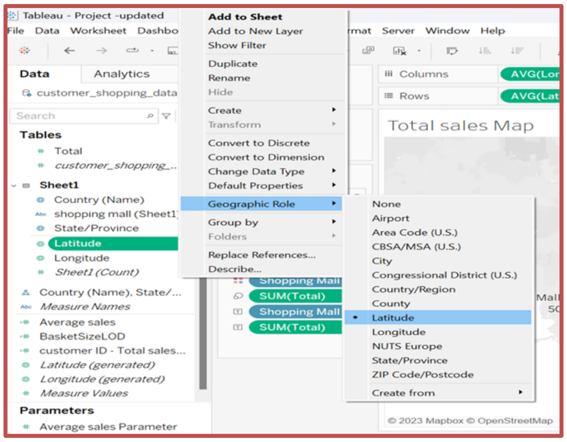
1. The dataset obtained did not contain geographical information. So, the first step for us is to identify the location of those shopping malls for our visualizations. To create a map with mall locations, we created a new Excel with mall names and longitude and latitude values. We added a relationship between the "customer_shopping_data" and "mall location" files using the common values in "mall name".



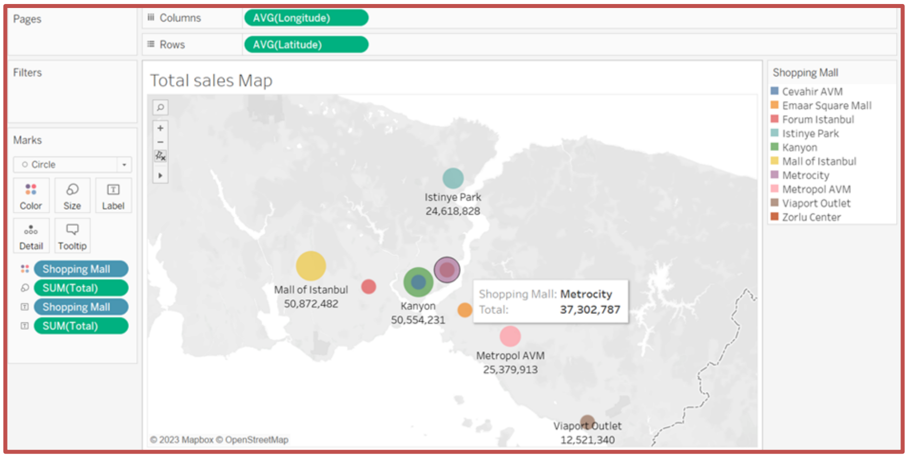
customer_shoppin...	Operator	Sheet1
Shopping Mall	=	shopping mall (!

shopping mall (Sheet1)	State/Province	Country (Name)	Latitude
Cevahir AVM	Şişli	İstanbul, Turky	41.053600
Emaar Square Mall	Üsküdar	İstanbul, Turky	41.018900
Forum Istanbul	Bayrampaşa	İstanbul, Turky	41.048200
Istinye Park	Sarıyer	İstanbul, Turky	41.181500

Then, we needed to ensure that longitude and latitudes are of data type "Number". We also converted their Geographic Role to longitudes and latitudes respectively.



We created a symbol map with these geographical locations, then assigned a color to each mall, and displayed the size of circle marks based on the SUM of total sales in that mall.



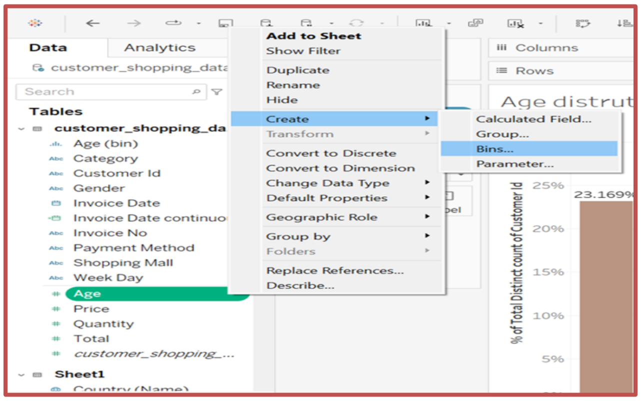
Columns: AVG(Longitude)

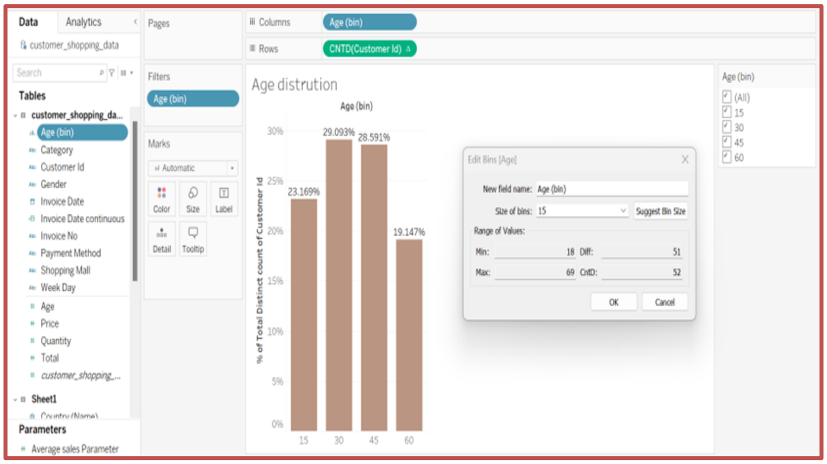
Rows: AVG(Latitude)

Marks: Circle

We noticed that the Mall of Istanbul had the highest sales with ₺50,872,482, followed by Kanyon with total sales of ₺50,554,231, and the third highest sales came from Metrocity Mall.

2. We had customer data between ages 18 - 69, we created an Age(bin) field with bin size 15. We visualized the age distribution in percentages and created a bar chart with Age(bin) column and CNTD(Customer Id).





Columns: Age (bin)

Rows: CNTD(Customer Id) Δ

Filters: Age (bin)

Marks: Automatic

Edit Bins [Age]

New field name: Age (bin)

Size of bins: 15

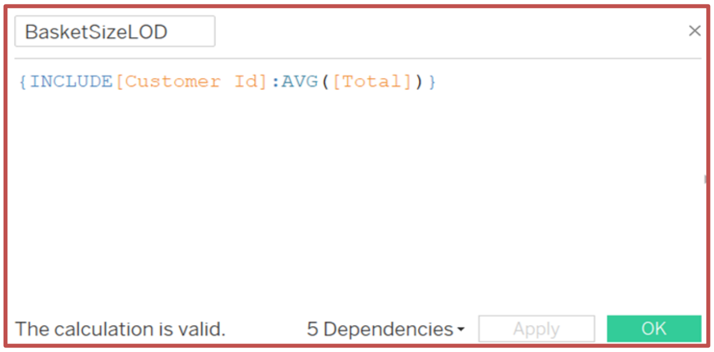
Range of Values:

Min: 18 Diff: 51

Max: 69 CntD: 52

LODs Used:

1. **BasketSizeLOD:**

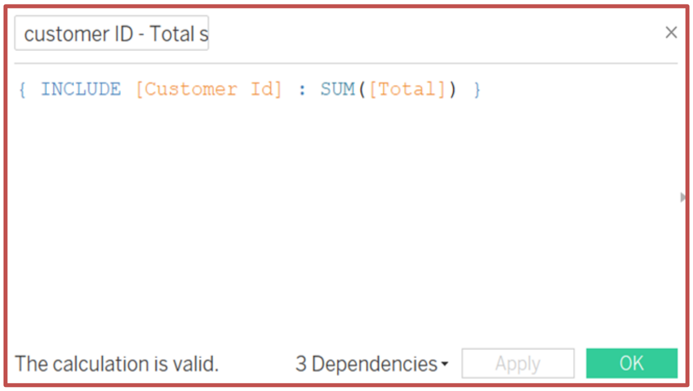


BasketSizeLOD

```
{INCLUDE[Customer Id]:AVG([Total])}
```

The calculation is valid.

2. **customer ID - Total sales LOD:**



customer ID - Total s

```
{ INCLUDE [Customer Id] : SUM([Total]) }
```

The calculation is valid.

VISUALIZATIONS:

We intended to answer all the proposed questions through our visualizations. They were the baseline for our storyboard/visualizations, and we will discuss them at each step along with their respective visualizations. With each visualization, we will also discuss the data design and integrity techniques we considered while creating them. Also, given below is a summary of the number of visualizations, dashboards, and stories our project was split into while planning its design.

- Worksheets: 16
- Dashboards with filters: 5
- Story: 1

<u>Interaction Techniques and Other System Features Used</u>

1. **Information Overview:**

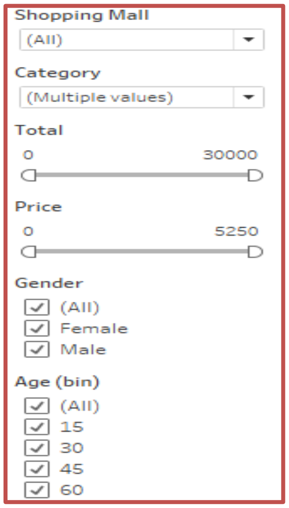
We have given an overview for the total population to address each question via dashboards and the story created.

2. **Zoom, Filter and Detail on Demand:**

Moreover, we have added the following filters on each visualization page to enable the user to filter for the required information and to obtain details on demand based on their requirement:

- Total Spent
- Item Price
- Categories
- Shopping Malls
- Age (bins)
- Gender

One or more dashboards are created to address each question and to allow the user to see the relationship between various elements impacting a variable from different angles. In addition, a filter for time is provided on relevant visualizations to look at details by different periods.



Shopping Mall
(All)

Category
(Multiple values)

Total
0 30000

Price
0 5250

Gender
☑ (All)
☑ Female
☑ Male

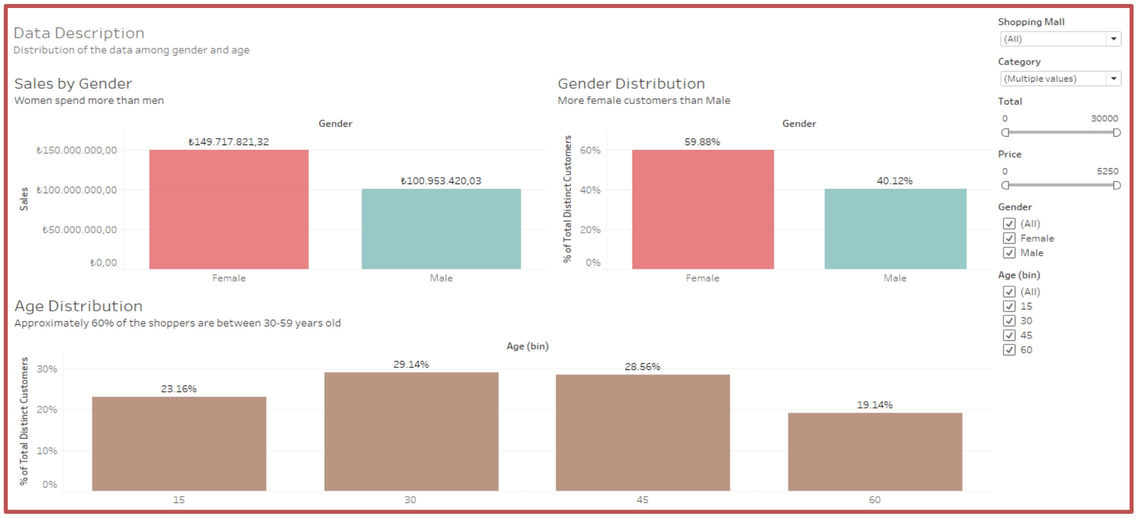
Age (bin)
☑ (All)
☑ 15
☑ 30
☑ 45
☑ 60

STEPS OF VISUALIZATION CREATION:

1. **Customer Profiling:**

To explore the gender distribution, we added gender in Column and CNTD(CustomerID) in the Rows shell and created a bar chart. We also created a bar chart with gender in column and SUM(Total) to identify which gender group was spending more. We could see that women are spending more when compared to men.

For further analysis we created an Age distribution histogram by placing Age(bin) in columns and CNTD(Customer Id) in Rows and formatted the Y-axis in percentages, so that we can understand the percentage of customers present in different age groups. The bin size selected was 15 years. Meaning that for the first bin there are customers between 15-29 years old, for the second bin there are customers aged 30-44 years old, for the third bin there customer aged 45-59 years old and for the last bin is all customers 60-69 years old which is the oldest customer recorded in the dataset.



Shopping Mall: (All)

Category: (Multiple values)

Total: 0 – 30000

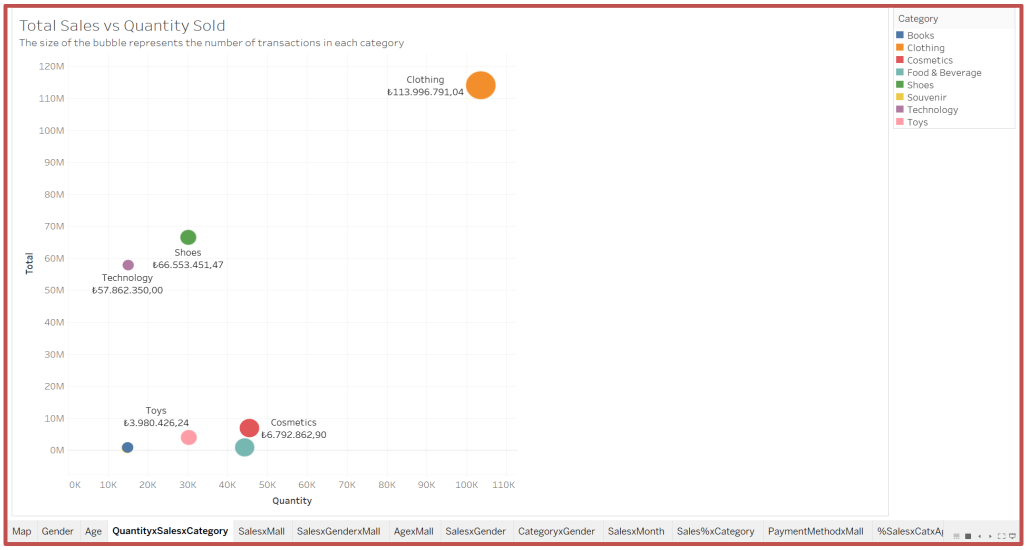
Price: 0 – 5250

Gender: ☑ (All) ☑ Female ☑ Male

Age (bin): ☑ (All) ☑ 15 ☑ 30 ☑ 45 ☑ 60

2. Data Overview

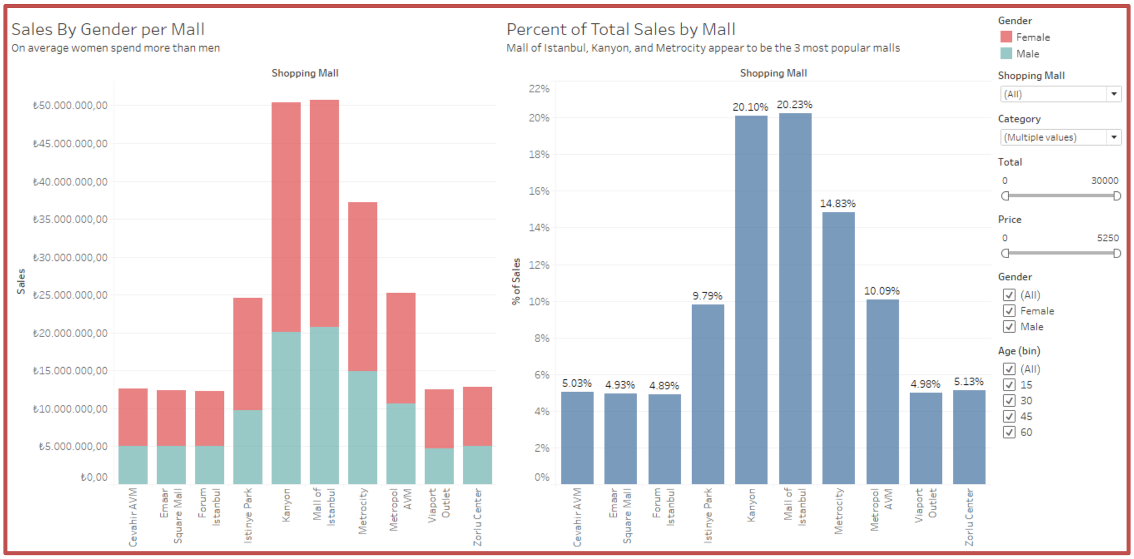
We established a color coding for various variables that we used in the different visualizations and maintained that consistency throughout the report. In the visualization below, a certain color palette is used to depict various categories. In addition, the position is shown to highlight the relative spent in each category, and the size of the circle depicts the number of transactions against the others.



3. Types of Customers by Mall

To analyze the sales by gender in different males, we placed Shopping Mall field in columns shelf and SUM(Total) in Rows shelf, and we added gender in color so that we can view each mall sales for both Females and males. To compare the sales with the percentage of total sales in each mall, we created a new sheet by adding the Shopping Mall field in columns shelf and SUM(Total) in Rows shelf and changed the Y-axis to percentages to get the Percentage of total sales by Mall.

Consistency is maintained throughout the report ***via color coding*** for gender. Females are represented with color red while males are represented with a light blue color. This makes the interpretation of visualization very intuitive.



Shopping Mall
(All) ▾

Category
(Multiple values) ▾

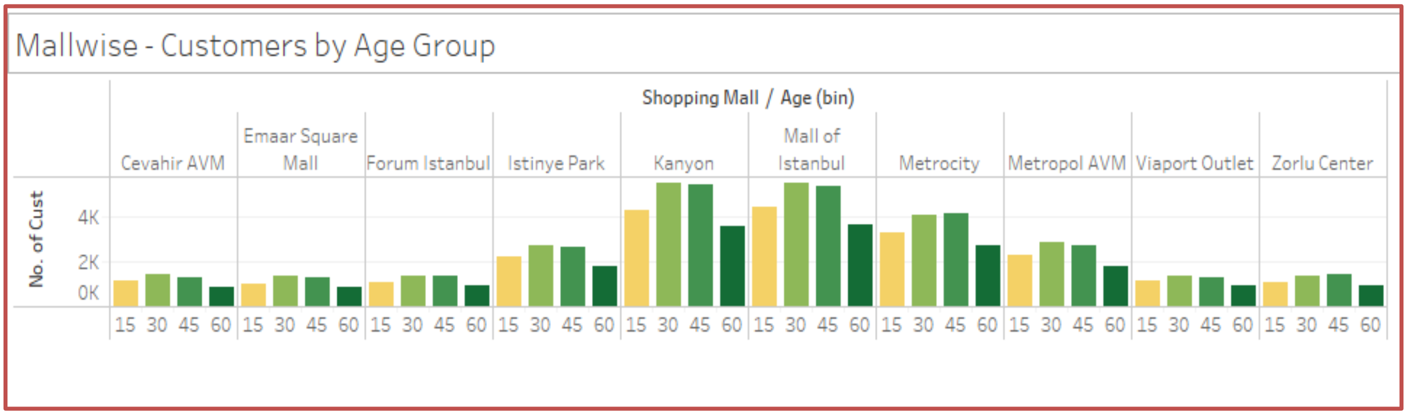
Total
0 30000

Price
0 5250

Gender
☑ (All)
☑ Female
☑ Male

Age (bin)
☑ (All)
☑ 15
☑ 30
☑ 45
☑ 60

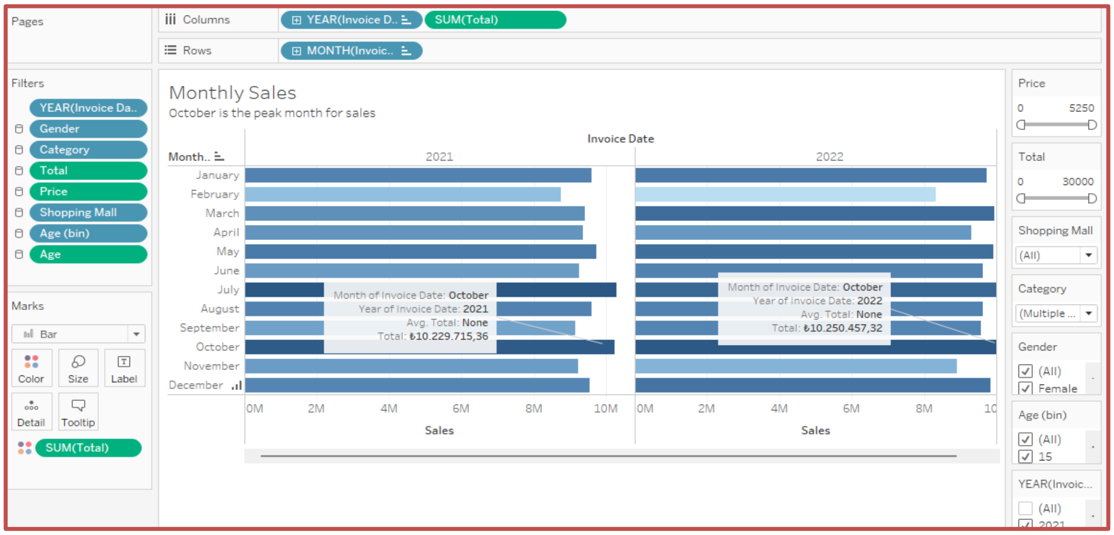
Following the previous analysis, we thought it was important to identify which age group that spends the most and therefore, should be attracted to increase mall sales. For this, we added Shopping mall and Age(bin) in the Columns and CNTD(Customer Id) in Rows. We used ***small multiples*** in the visualization below to show how the customer segments vary across each mall by the various age groups previously defined. ***Consistency*** is maintained below ***via color coding*** for age groups, as mentioned earlier to make the interpretation of visualization intuitive. The older the customers are the darkest the green used to represent each group.



4. Peak Shopping Seasons:

In the visualization below, we are trying to show the seasonality of spending and the months during which there is the highest spending. Each month is shaded with a different blue tone that represents the amount spent. The darkest blue color represents high spending months moving to lighter shades of blue as the spending amount decreases. Also, since we had the data available for 2 years, a comparison is shown side by side to show consistency in spending during similar months across the period of 2 years.

To show the monthly spending over each year, we dropped the invoice date dimension in rows shelf and then sub-filtered it to month. Then dropped the total measure in columns and it automatically got aggregated to the sum. We also added the total measure to Color so that the visualization will show color variation based on the total spent by the month (highest spending month in the darkest blue). The invoice date dimension was also dropped on the column shelf and left it at the year level to split the visualization by year. All the other filters were applied and set to show on the leftmost filter container.



Pages

Columns: YEAR(Invoice D.. | SUM(Total)

Rows: MONTH(Invoic..

Filters: YEAR(Invoice Da.. | Gender | Category | Total | Price | Shopping Mall | Age (bin) | Age

Marks: Bar | Color | Size | Label | Detail | Tooltip | SUM(Total)

Monthly Sales

October is the peak month for sales

Month of Invoice Date: October
Year of Invoice Date: 2021
Avg. Total: None
Total: ₺10.229.715,36

Month of Invoice Date: October
Year of Invoice Date: 2022
Avg. Total: None
Total: ₺10.250.457,32

Price: 0 – 5250

Total: 0 – 30000

Shopping Mall: (All)

Category: (Multiple ...

Gender: ☑ (All) ☑ Female

Age (bin): ☑ (All) ☑ 15

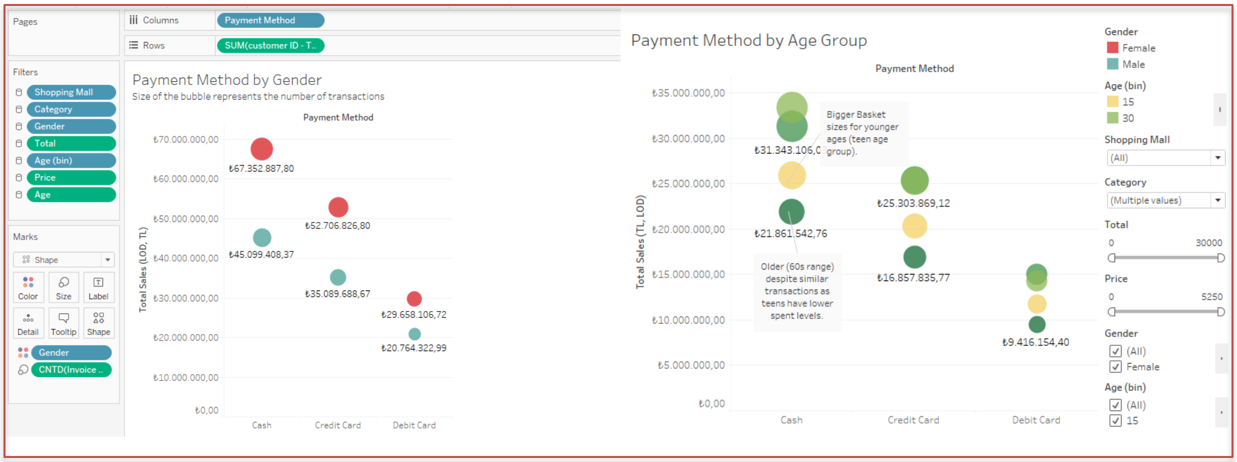
YEAR(Invoic...: □ (All) ☑ 2021

5. Popular Payment Methods by Malls and Customers

This chart was created to depict the pattern of spending between each age and gender group in terms of the number of transactions and the amount of money spent by each group divided by payment method. The visualization below exhibits spending ***patterns highlighted with annotations. Consistency*** is maintained ***via color coding*** for gender as well as for age groups in the below visualization, as mentioned earlier to make the interpretation of visualization very intuitive.

To develop the visualization on the left, we used the calculated field ***customer ID - Total sales LOD*** in row shelf and used the sum aggregation measure. We then added the payment method to the columns shelf. Gender was dropped in Color. Count Distinct for Invoice No was dropped on Size to determine the size of the circle which represents the number of transactions.

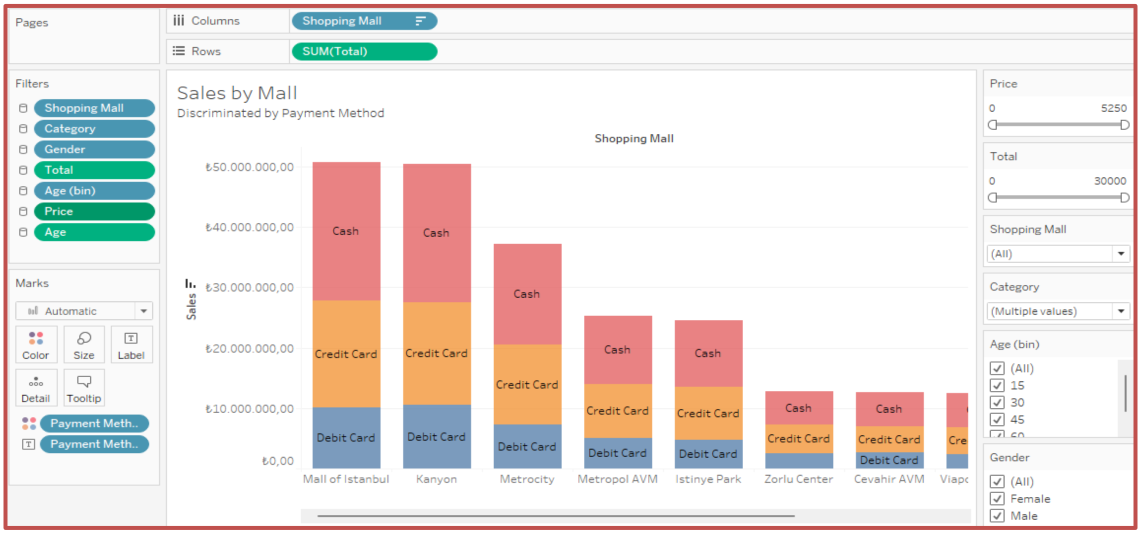
Similarly, for the “Payment Method by Age” visualization on the right, we used the calculated field ***customer ID - Total sales LOD*** in Row shelf and used the sum aggregation measure. We then added the payment method to the columns shelf. The age (bin) dimension was dropped in Color. Count Distinct for Invoice No was dropped on Size to determine the size of the circle which represents the number of transactions.



Columns: Payment Method
Rows: SUM(customer ID - T..

Filters: Shopping Mall, Category, Gender, Total, Age (bin), Price, Age
Marks: Shape — Color: Gender; Size: CNTD(Invoice ..

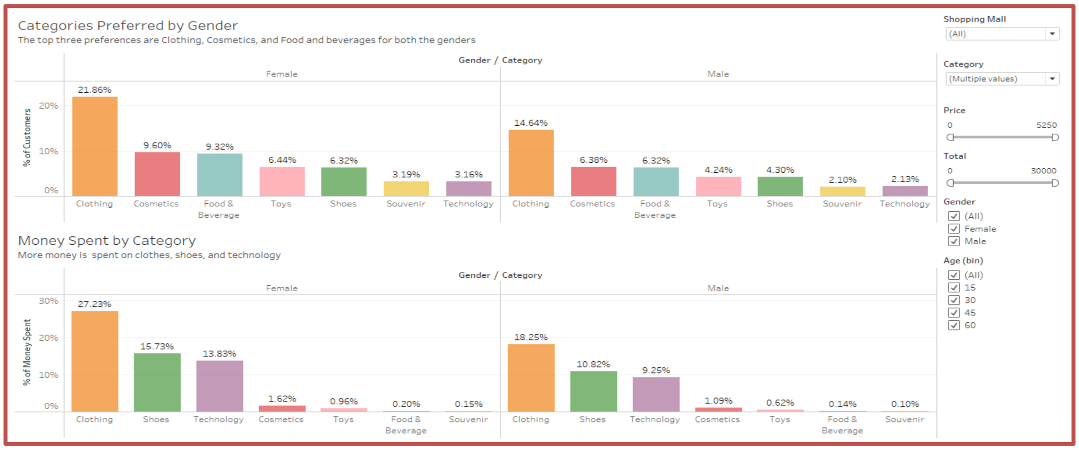
To build this visualization below, we used the total measure in the row shelf and left the default sum aggregation measure. We then added the Shopping Mall dimension to the columns shelf. The payment method was dropped in both colors as well as detail to show color variation by each payment method for each mall and the division of spending for each method by each mall, respectively.



Columns: Shopping Mall
Rows: SUM(Total)

Filters: Shopping Mall, Category, Gender, Total, Age (bin), Price, Age
Marks: Automatic — Color: Payment Meth..; Label: Payment Meth..

6. Spending Categories and by Gender



Categories Preferred by Gender

The top three preferences are Clothing, Cosmetics, and Food and beverages for both the genders

Money Spent by Category

More money is spent on clothes, shoes, and technology

Shopping Mall: (All)

Category: (Multiple values)

Price: 0 – 5250

Total: 0 – 30000

Gender
☑ (All)
☑ Female
☑ Male

Age (bin)
☑ (All)
☑ 15
☑ 30
☑ 45
☑ 60

Categories Preferred by Gender:

This bar chart shows the percentage of customers in different categories based on gender. Female customers prefer clothing the most, followed by cosmetics and food and beverages. Male customers also prefer clothing the most, followed by cosmetics and food and beverages, but in lower percentages compared to females.

To develop the visualization, we used the Customer ID field in the Rows shelf. The Gender field and Category field in the Columns shelf. And aggregated the data by counting the number of customers for each combination of gender and category. We used “Marks” card to customize the color for different categories.

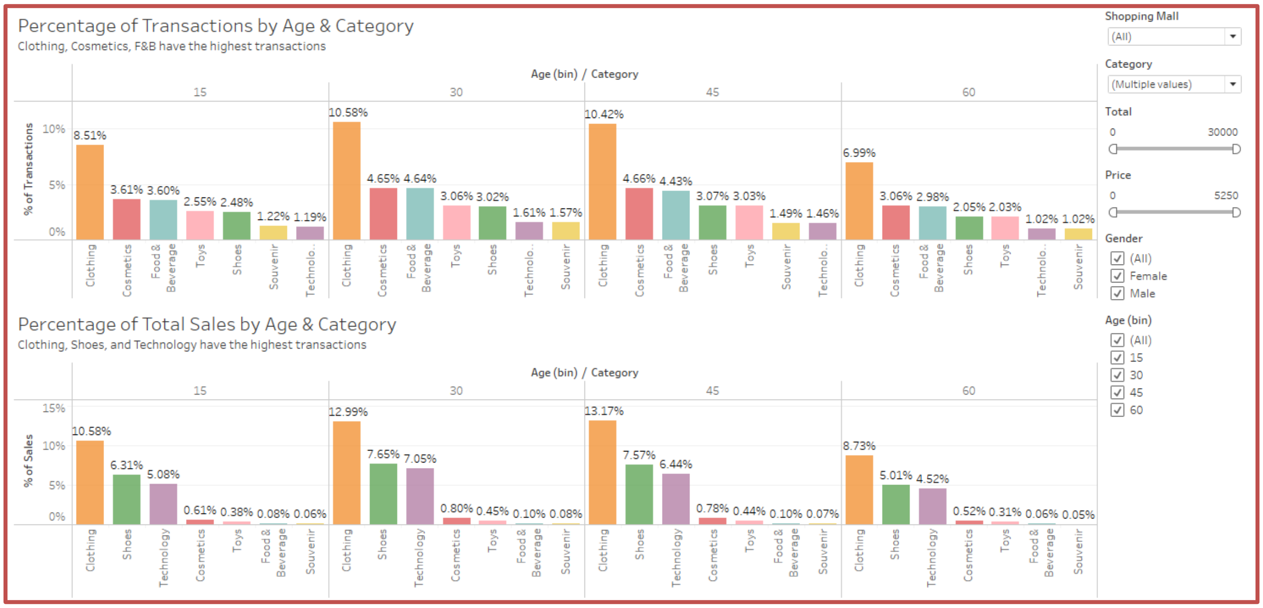
Money Spent by Category:

This bar chart shows the percentage of money spent by male and female customers in different categories. Clothing was the highest category for both genders. For both genders, shoes, technology, and cosmetics were the next most significant categories, with females spending more in all three. Toys, food and beverages, books, and souvenirs were the least significant categories.

To develop the visualization, we used the Sum Basket Size in the Rows shelf. The Gender field and Category field in the Columns shelf. Aggregated the data by summing the basket sizes for each combination of gender and category. We used “Marks” card to customize the color for different categories.

Consistency is maintained via color coding for different categories in the visualization, to make the interpretation of visualization more intuitive. To develop the visualization, we used the Customer ID variable in the Rows shelf. The Gender and Category variables in the Columns shelf. Then we aggregated the data by counting the number of customers for each combination of gender and category.

7. Popular Categories by Age Group



Shopping Mall
(All)

Category
(Multiple values)

Total
0 30000

Price
0 5250

Gender
☑ (All)
☑ Female
☑ Male

Age (bin)
☑ (All)
☑ 15
☑ 30
☑ 45
☑ 60

Percentage of Transaction by Age & Category:

This bar chart shows the percentage of transactions in various categories for different age bins. The categories represented on the X-axis and Y-axis represents the percentage of transactions in each category. There are four age bins represented in this chart: 15, 30, 45, and 60. For the 15-year-old age bin. ***Consistency*** is maintained ***via color coding*** for different categories in the visualization, to make the interpretation of visualization very intuitive.

To develop the visualization, we used the Invoice No field in the Rows shelf. The Age(bin) and Category field in the Columns shelf. We used the "Analysis" menu, to create Bins and dragged the newly created "Age (bin)" field onto the Columns shelf. Aggregated the data by binning the ages and counting the number of unique invoices for each combination of age range and category. We used “Marks” card to customize the color for different categories.

Percentage of Total Sales by Age and Category:

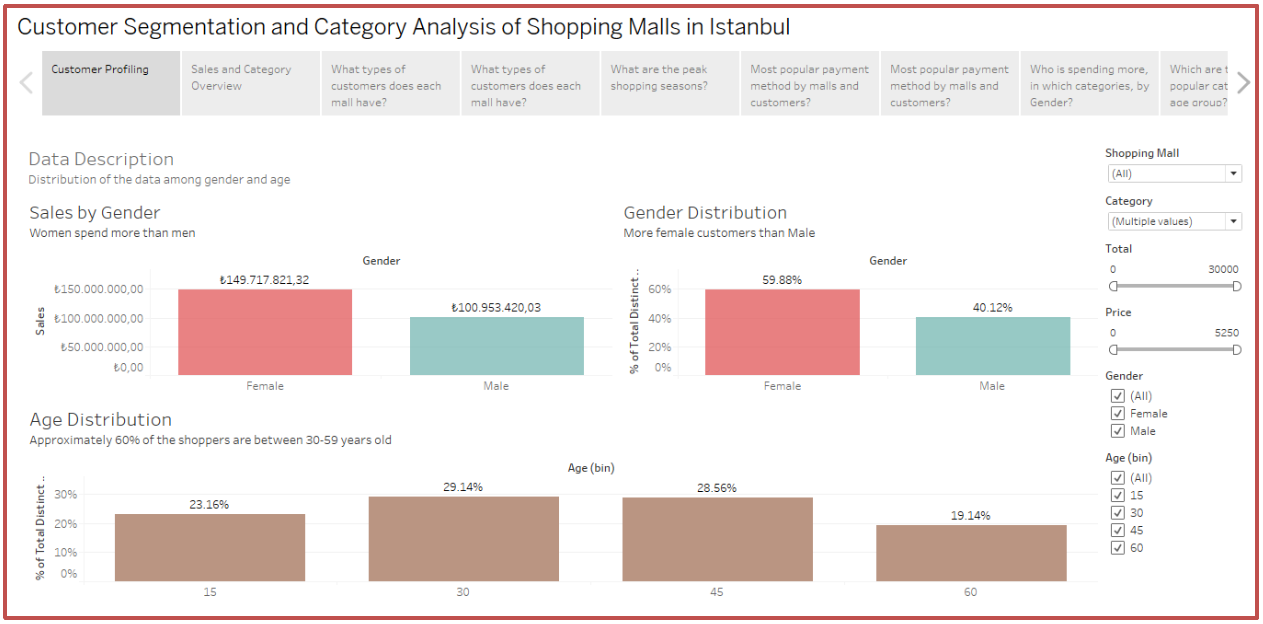
The bar chart displays the percentage of total sales for different categories across four different age bins: 15, 30, 45, and 60. The Y-axis represents the percentage of total sales, while the X-axis represents the different categories. To develop the visualization, we used the SUM(Total) field in the Rows shelf. In the column shelf we used the "Age (bin)" variable. Tableau will automatically aggregate the data by the age bins and add the total sales for each combination of age range and category.

This visualization will display the relative proportion of the total sales for each combination of age range (binned) and category. Overall, clothing and shoes are the highest selling categories across all age groups, with technology also having a significant percentage of total sales. Cosmetics, toys, food and beverages, books, and souvenirs have relatively lower percentages of total sales across all age groups. ***Consistency*** is maintained ***via color coding*** for different categories in the visualization, to make the interpretation of visualization very intuitive.

To develop the visualization, we used the "Total" field in the Rows shelf. Used the "Analysis" menu, to create Bins, and created bins for the "Age" field. In the column shelf we used the newly created "Age (bin)". Right clicked on the "Total" field on the Rows shelf and selected "Measure > Sum" to sum the total sales. Tableau will automatically aggregate the data by binning the ages and summing the total sales for each combination of age range and category. Click on the "Analysis" menu and select "Totals > Show Row Grand Totals". Click on the "Analysis" menu and select "Percentage of Total > Column". We used “Marks” card to customize the color for different categories.

FINAL STORY AND KEY FINDINGS FOR STORYBOARD:

Story:



Customer Segmentation and Category Analysis of Shopping Malls in Istanbul

Customer Profiling
Sales and Category Overview
What types of customers does each mall have?
What types of customers does each mall have?
What are the peak shopping seasons?
Most popular payment method by malls and customers?
Most popular payment method by malls and customers?
Who is spending more, in which categories, by Gender?
Which are t popular cat age group?

Data Description
Distribution of the data among gender and age

Sales by Gender
Women spend more than men

Gender
Sales
₺150.000.000,00
₺100.000.000,00
₺50.000.000,00
₺0,00
₺149.717.821,32
₺100.953.420,03
Female
Male

Gender Distribution
More female customers than Male

Gender
% of Total Distinct ..
60%
40%
20%
0%
59.88%
40.12%
Female
Male

Age Distribution
Approximately 60% of the shoppers are between 30-59 years old

Age (bin)
% of Total Distinct ..
30%
20%
10%
0%
23.16%
29.14%
28.56%
19.14%
15
30
45
60

Shopping Mall
(All)

Category
(Multiple values)

Total
0 30000

Price
0 5250

Gender
☑ (All)
☑ Female
☑ Male

Age (bin)
☑ (All)
☑ 15
☑ 30
☑ 45
☑ 60

Key Findings:

We identified that women spend more than men and clothing is the most popular and high-revenue category across Istanbul for all customers. Shoes and Technology are also the most profitable categories, yielding the highest sales with the lower number of transactions. Cash is the most popular mode of payment type used in all malls while debit card is the least popular.

LIMITATIONS & FURTHER WORK

Data Limitation

Our data doesn't contain information about the same customer doing multiple transactions on different days, as each person was assigned a unique customer ID per transaction they made. Therefore, the customer ID reflects data as if each transaction was made by a different customer every time. It would have been better to visualize the customer preferences if we had returning customers' data as well.

Also, since we see a high purchasing trend during tourist season, we are assuming that a portion of those purchases made are by the tourists. However, we do not have concrete evidence to conclude it and will need to gather more data to verify our hypothesis since the original data did not include any additional details on customer profile other than ages, spending and type of payment method used.

ACKNOWLEDGMENTS

We would like to thank Dr. Kutsal Dogan for all his teaching and support throughout the class. It was a very enjoyable module, and this project gave us the opportunity to apply everything we learned. Thank you for teaching us how powerful data is if visualized in the correct manner and how it can help us identify trends and important information about a business. We would also like to thank our team for all the work and collaboration that lead us to be able to finalize this project successfully.

Our Team:

1. Sarah Alikhan - Viz builder
2. Sri Ramya Simhadri - Viz builder
3. Valeria Latorraca - Designer
4. Swetha Chukka - Evaluator
5. Alejandra Mejia - Project manager

REFERENCES

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