Data Analysis Final Project
Shopping Tendencies Analysis

By:

Sahar Elsersawi

Introduction

- Our project is talking about the analysis of sales in 10 malls in Turkey.
- We used different analysis methodologies and techniques to analyze our dataset like we used
- Python coding and MS Power Bi visualization and reports.
- We set 4 questions to use its answer for good analysis to be shown clearly to the stakeholders :

1- Which gender frequency buy more ? And What is the average age?

2- What is the sales volume of each category? Which category has the majority of sales volume?

3- What is the sales volume for each mall ? Which mall has the majority sales volume?

4- What is the most common payment method?

```
##loading data
#read the dataset csv file
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import numpy as np
df=pd.read_csv("/Users/saharsayed/Downloads/customer_shopping_data
2.csv")
# Display the first few rows of the dataset
#print(df.head())
df.head()
```

	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	5/8/2022	Kanyon
1	I317333	C111565	Male	21	Shoes	3	1800.51	Debit Card	12/12/2021	Forum Istanbul
2	I127801	C266599	Male	20	Clothing	1	300.08	Cash	9/11/2021	Metrocity
3	I173702	C988172	Female	66	Shoes	5	3000.85	Credit Card	16/05/2021	Metropol AVM
4	I337046	C189076	Female	53	Books	4	60.60	Cash	24/10/2021	Kanyon

```
#Clean the data
##finding missing null
valuesdf.isnull().sum()

## Drop rows with missing values
df=df.dropna()

## Drop duplicate
 rowsdf = df.drop_duplicates()
```

Understanding the Dataset

```
#understanding the dataset
# get first 5 rows
rowsdf.head()

##data processing and # check the content

df.shapedf.describe()
```

	age	quantity	price
count	99457.000000	99457.000000	99457.000000
mean	43.427089	3.003429	689.256321
std	14.990054	1.413025	941.184567
min	18.000000	1.000000	5.230000
25%	30.000000	2.000000	45.450000
50%	43.000000	3.000000	203.300000
75%	56.000000	4.000000	1200.320000
max	69.000000	5.000000	5250.000000

Checking number of unique entries

```
# Checking number of unique entries
df.nunique()
```

```
invoice_no.          99457
customer_id          99457
gender                2
age                   52
category               8
quantity               5
price                 40
payment_method         3
invoice_date         797
 shopping_mall        10
year 3
month 12 dtype: int64
```

1- Which gender buy more ?

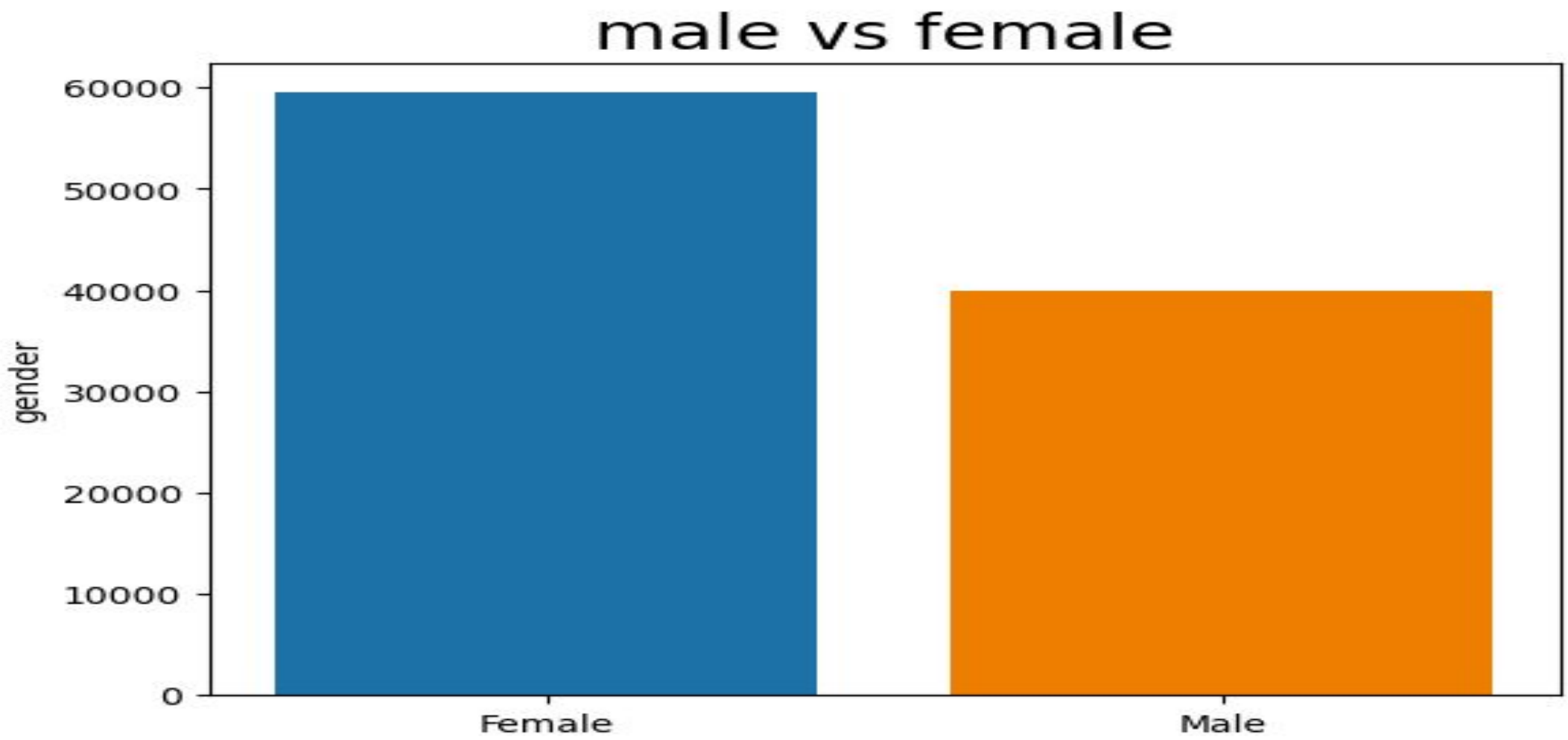
```
# Most frequent entries and their frequencies for 'Gender'
gender_counts = df['gender'].value_counts()
print("Gender Counts:")
print(gender_counts)
```

```
Gender Counts:
Female    59482
Male      39975
Name: gender, dtype: int64
```

- **Females buy with about 20% more than males.**

In [24]:
```
gen_num=df["gender"].value_counts()
sns.barplot(y = gen_num, x = gen_num.index, data = df)
plt.title("male vs female", size=20)
```

Out[24]:
```
Text(0.5, 1.0, 'male vs female')
```



In [25]:
```
##Majority of customers are female
```

sales report 4 - Power BI Desktop

Search

ReDi-60088

File Home Insert Modeling View Optimize Help Format Data / Drill

Paste Cut Copy Format painter — Clipboard

Get data Excel workbook Data hub SQL Server Enter data Dataverse Recent sources — Data

Transform data Refresh — Queries

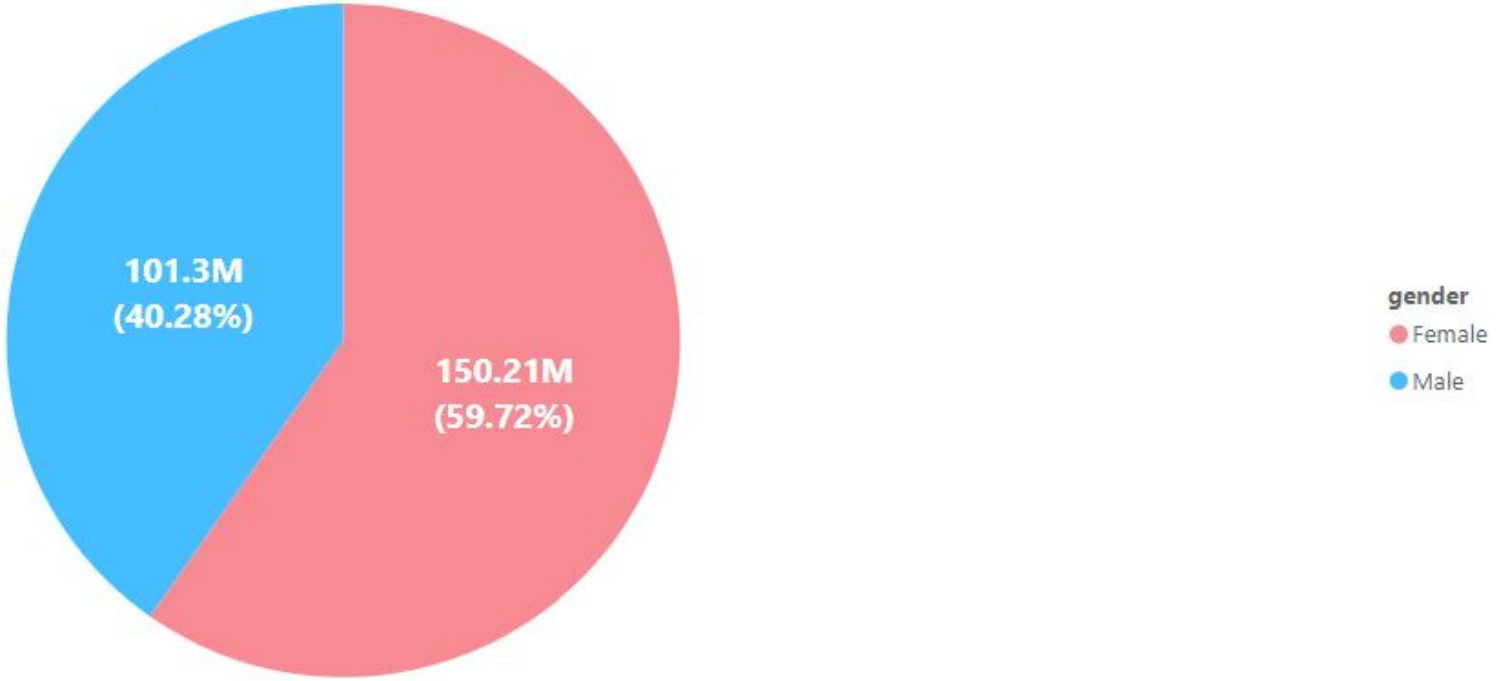
New visual Text box More visuals — Insert

New measure Quick measure — Calculations

Sensitivity — Sensitivity

Publish — Share

Back to report | SUM OF TOTAL_PRICE BY GENDER



Filters Visualizations Data

Page 2 of 2

100%

Type here to search

22°C Soligt

6:44 PM
6/6/2023

What is the average age?

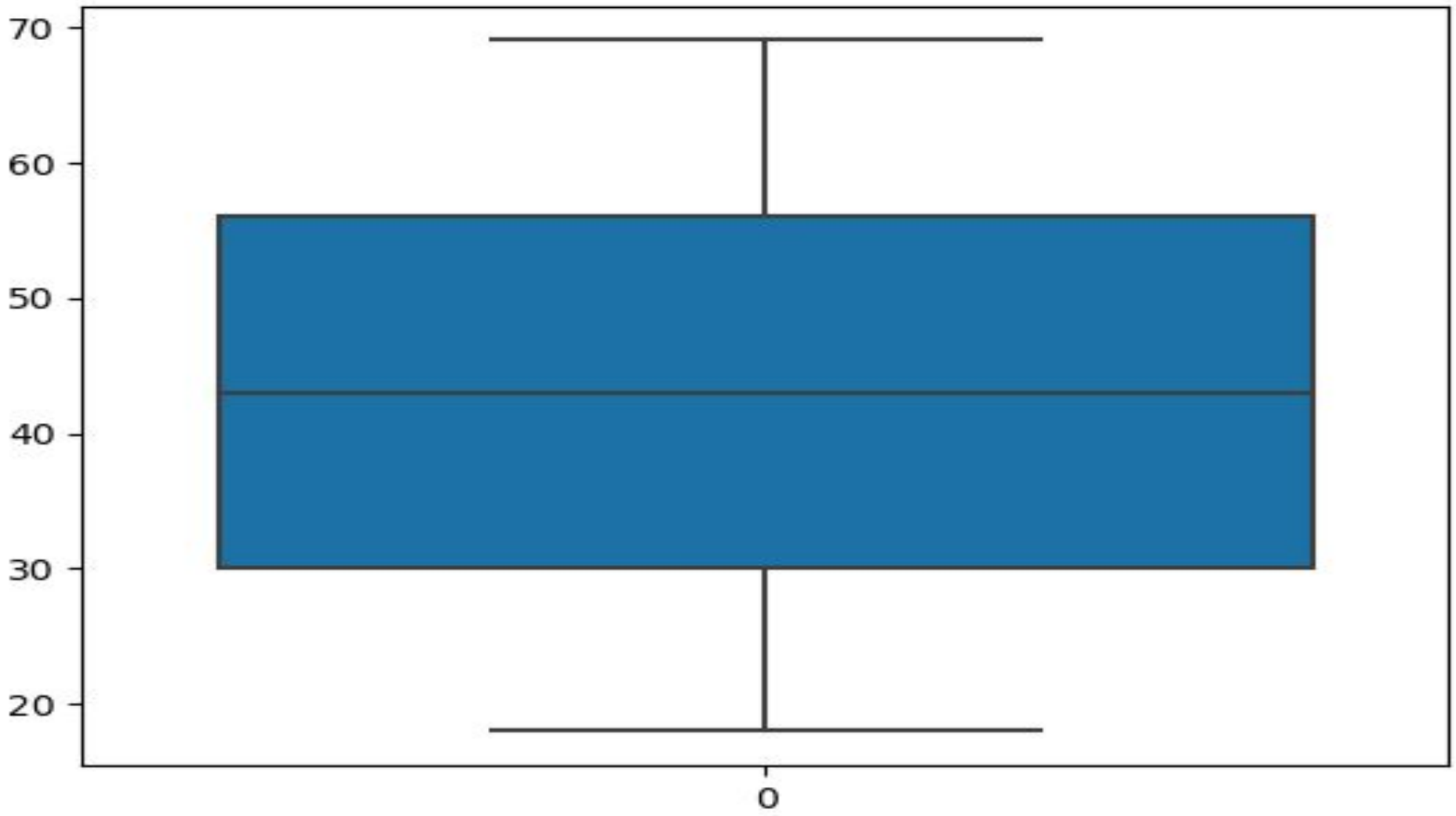
```
# Calculate average age
average_age = df['age'].mean()
print('Average age')
print (average_age)
```

```
Average age
43.42708909377922
```

- **Average age (mean) is 43 years old.**

In [22]:
```
sns.boxplot(df['age'])
```

Out[22]: <Axes: >



In [23]:
```
#So, average customer age is between 30-55 years
```

2- What is the sales volume of each category? Which category has the majority of sales volume?

```
# Group by category and calculate sales volume

category_sales = df.groupby("category")[["quantity", "price"]].sum()
category_sales["sales_volume"] = category_sales["quantity"] * category_sales["price"]
print("SALES VOLUME OF EACH CATEGORY")
print(category_sales["sales_volume"])

# Identify category with the majority sales volume

majority_category = category_sales["sales_volume"].idxmax()
print("Category with the majority sales volume:", majority_category)
```

SALES VOLUME OF EACH CATEGORY

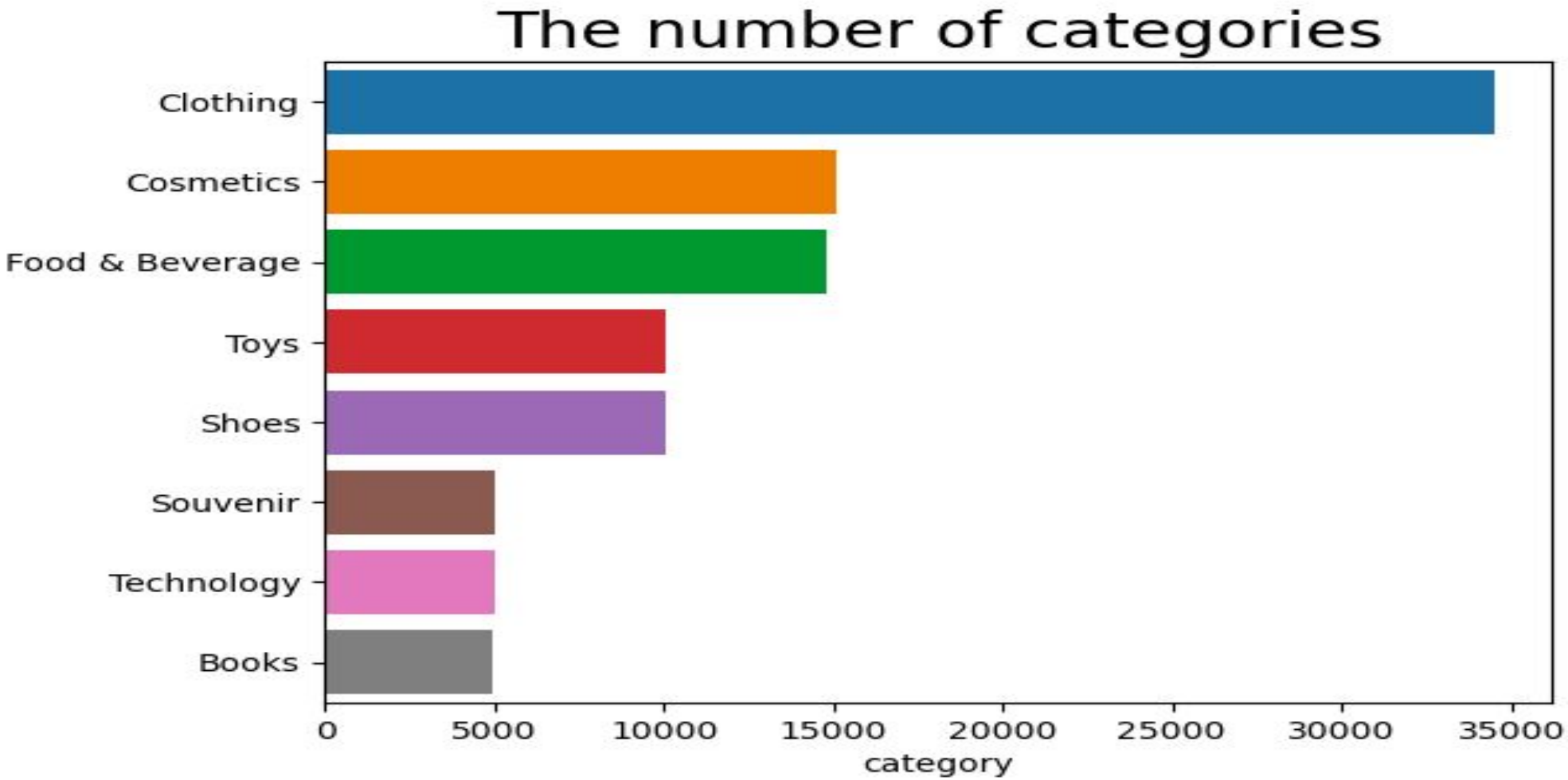
```
category
Books              3.400574e+09
Clothing           3.218136e+12
Cosmetics          8.404691e+10
Food & Beverage    1.025317e+10
Shoes              5.479955e+11
Souvenir           2.594050e+09
Technology         2.369120e+11
Toys               3.294997e+10
Name: sales_volume, dtype: float64
Category with the majority sales volume: Clothing
```

- **Clothes and shoes have the majority of sales volume and books and Souvenir have the minority of sales volume.**

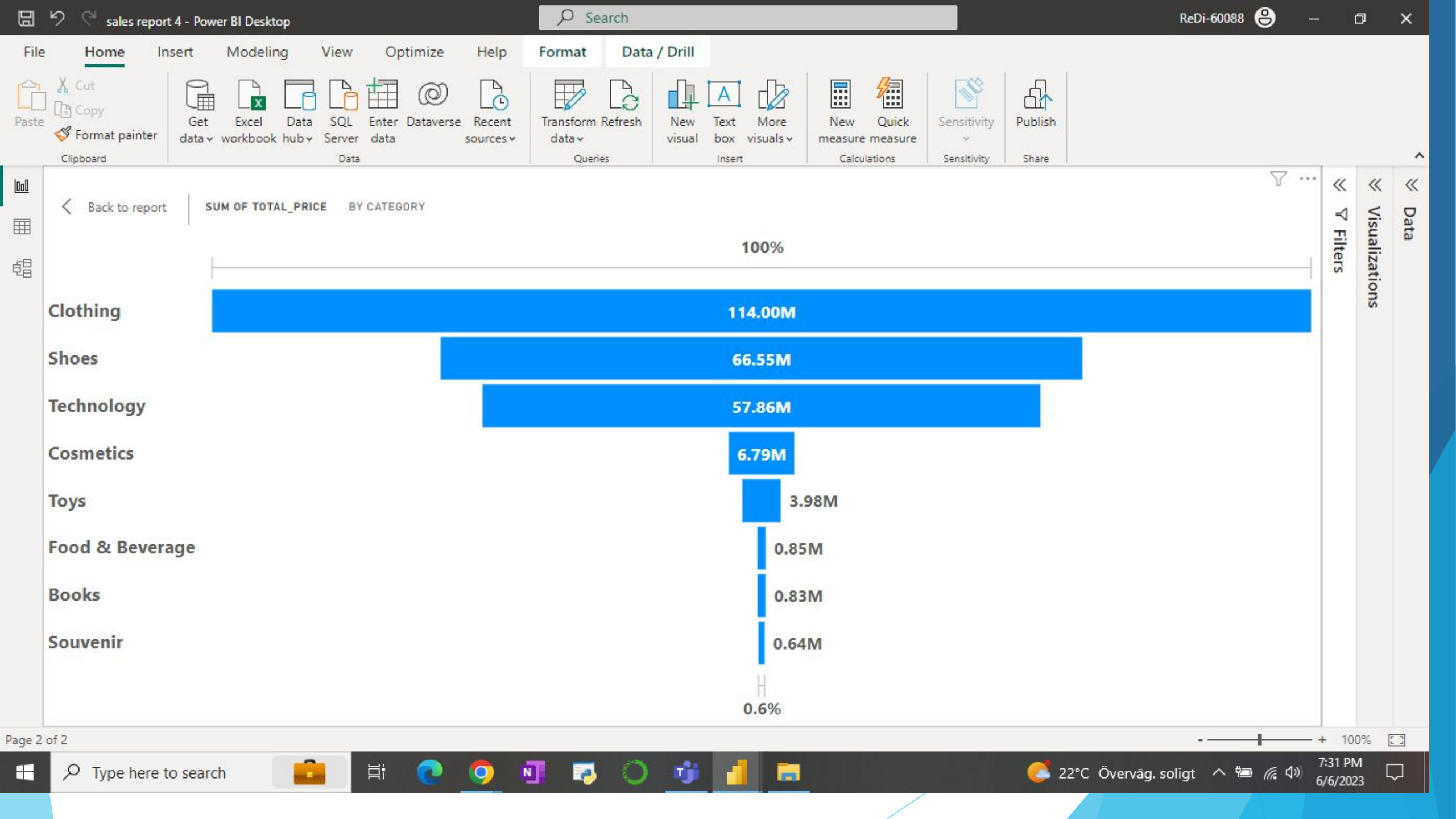
In [ ]:

In [23]:
```
cat_num = df["category"].value_counts()
sns.barplot(x = cat_num, y = cat_num.index, data = df)
plt.title("The number of categories", size=20)
```

Out[23]:
```
Text(0.5, 1.0, 'The number of categories')
```



In [24]:
```
##There are total 8 categories of products and maximum sales by quantity is clothes .
```


sales report 4 - Power BI Desktop
Search
ReDi-60088
File
Home
Insert
Modeling
View
Optimize
Help
Format
Data / Drill
Paste
Cut
Copy
Format painter
Clipboard
Get data
Excel workbook
Data hub
SQL Server
Enter data
Dataverse
Recent sources
Data
Transform data
Refresh
Queries
New visual
Text box
More visuals
Insert
New measure
Quick measure
Calculations
Sensitivity
Sensitivity
Publish
Share
Back to report
SUM OF TOTAL_PRICE
BY CATEGORY
100%
Clothing
114.00M
Shoes
66.55M
Technology
57.86M
Cosmetics
6.79M
Toys
3.98M
Food & Beverage
0.85M
Books
0.83M
Souvenir
0.64M
0.6%
Filters
Visualizations
Data
Page 2 of 2
100%
Type here to search
22°C Överväg. soligt
7:31 PM
6/6/2023

3- What is the sales volume for each mall ? Which mall has the majority sales volume?

```
# Group by shopping mall and calculate sales volume
mall_sales = df.groupby("shopping_mall")[["quantity", "price"]].sum()
mall_sales["sales_volume"] = mall_sales["quantity"] * mall_sales["price"]
print("SALES VOLUME OF EACH MALL:")
print(mall_sales["sales_volume"])

#identify mall with the  majority sales volume
mall_max_sales = mall_sales["sales_volume"] . idxmax()
print ("MALL WITH MAJORITY SALES:")
print (mall_max_sales)
```

SALES VOLUME OF EACH MALL:

shopping_mall

Cevahir AVM 5.132996e+10

Emaar Square Mall 4.916431e+10

Forum Istanbul 4.954737e+10

Istinye Park 1.979187e+11

Kanyon 8.152004e+11

Mall of Istanbul 8.326834e+11

Metrocity 4.601626e+11

Metropol AVM 2.118169e+11

Viaport Outlet 5.024071e+10

Zorlu Center 5.346599e+10

Name: sales_volume, dtype: float64

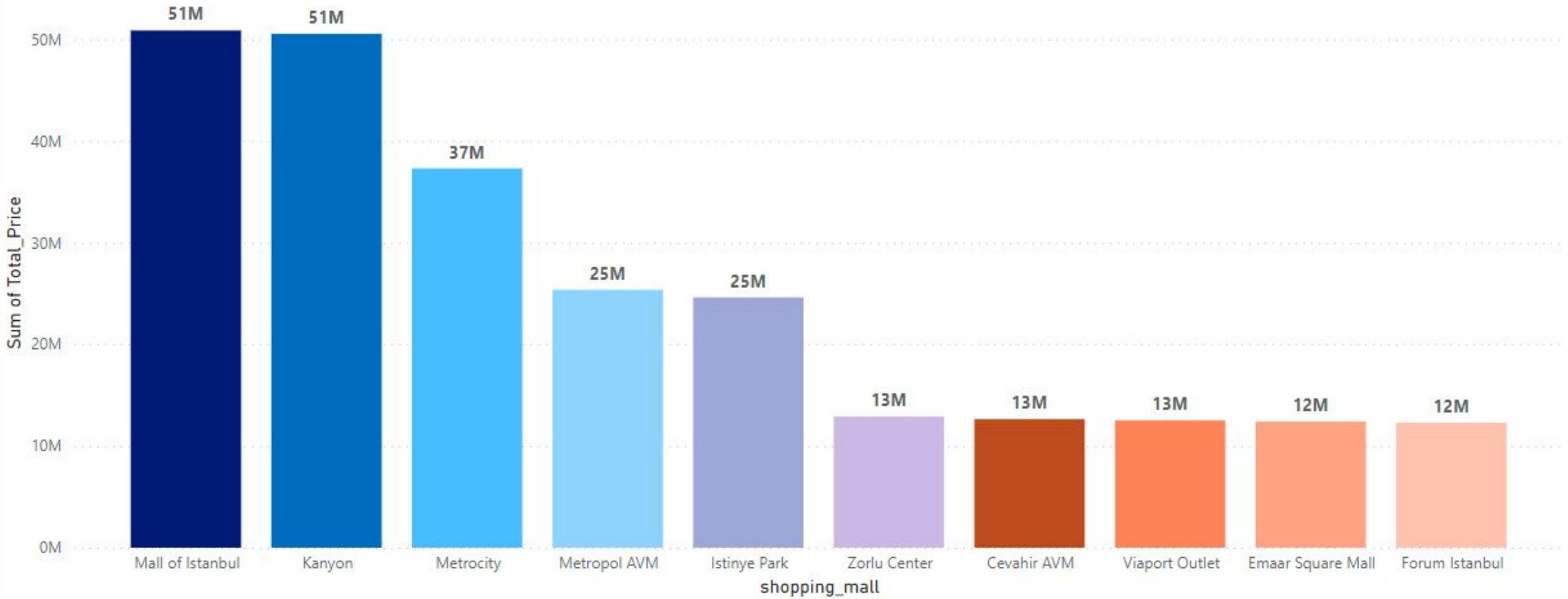
MALL WITH MAJORITY SALES:
Mall of Istanbul

- **Mall of Istanbul has the majority of sales.**
- **Emaar square mall has the minority of sales.**

File Home Insert Modeling View Optimize Help Format Data / Drill

Paste Cut Copy Format painter — Clipboard
Get data Excel workbook Data hub SQL Server Enter data Dataverse Recent sources — Data
Transform data Refresh — Queries
New visual Text box More visuals — Insert
New measure Quick measure — Calculations
Sensitivity — Sensitivity
Publish — Share

Back to report | SUM OF TOTAL_PRICE BY SHOPPING_MALL



Filters Visualizations Data

Type here to search
22°C Överväg. soligt
7:29 PM
6/6/2023

4- What is the most common payment method?

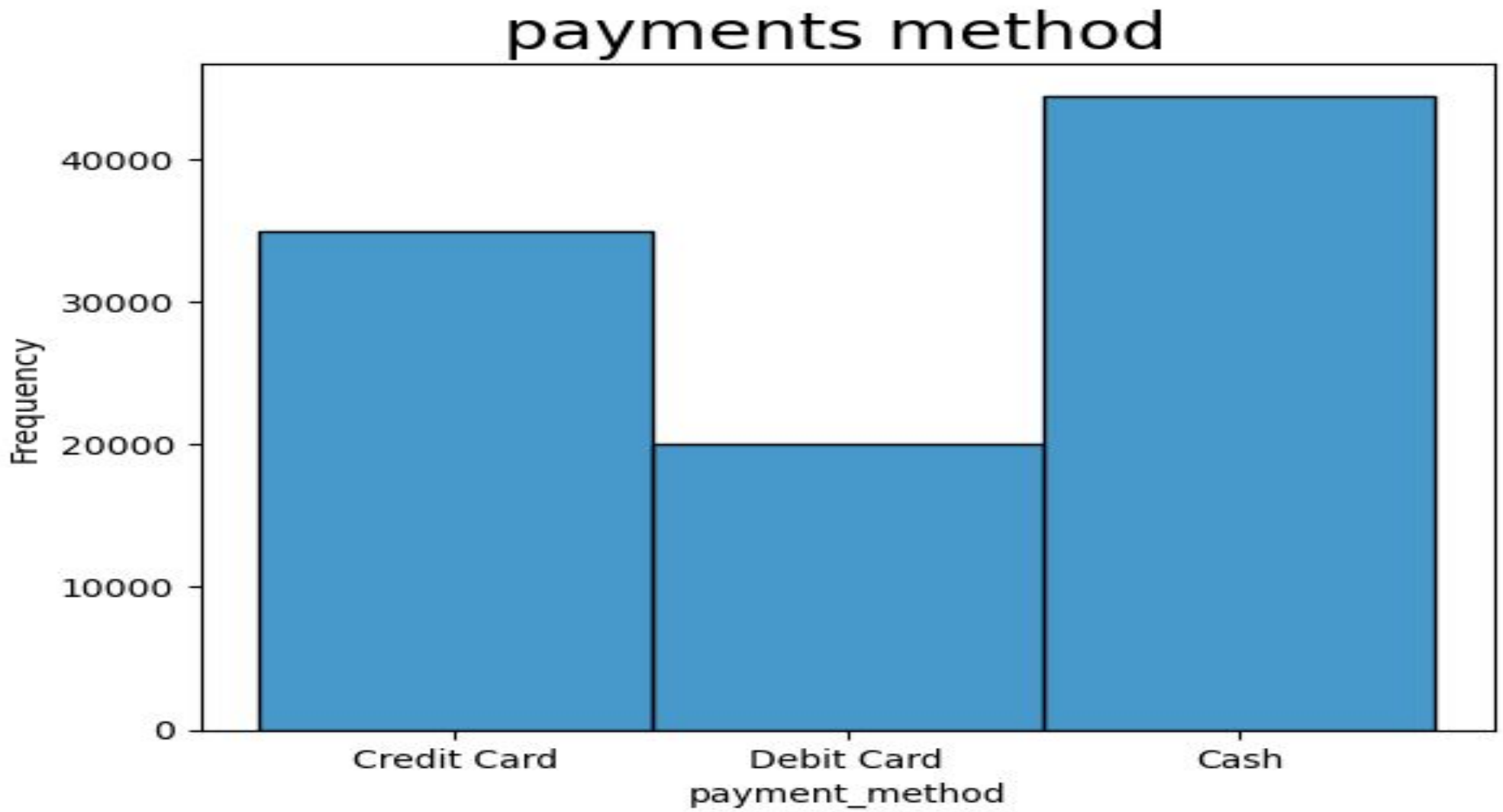
```
### Most frequent entries and their frequencies for 'Payment_Method’
payment_counts = df['payment_method'].value_counts()print("\nPayment Method Counts:")
print (payment_counts)
## identify the common payment method
most_common_payment _method = payment_method_frequency.idxmax()
print (‘MOST COMMON PAYMENT METHOD :’)
print ( MOST_COMMON_PAYMENT_METHOD)
```

```
Payment Method Counts:
Cash          44447
Credit Card   34931
Debit Card    20079
Name: payment_method, dtype: int64
MOST COMMON PAYMENT METHOS:
CASH
```

- **Cash is the common payment method**

In [26]:
```
sns.histplot(df["payment_method"], kde = False,stat='frequency')
plt.title("payments method ", size=20,)
```

Out[26]:
```
Text(0.5, 1.0, 'payments method ')
```



In [27]:
```
## Cash is the most prefferd mode of transaction followed
```

sales report 4 - Power BI Desktop

Search

ReDi-60088

File Home Insert Modeling View Optimize Help Format Data / Drill

Paste Cut Copy Format painter — Clipboard

Get data Excel workbook Data hub SQL Server Enter data Dataverse Recent sources — Data

Transform data Refresh — Queries

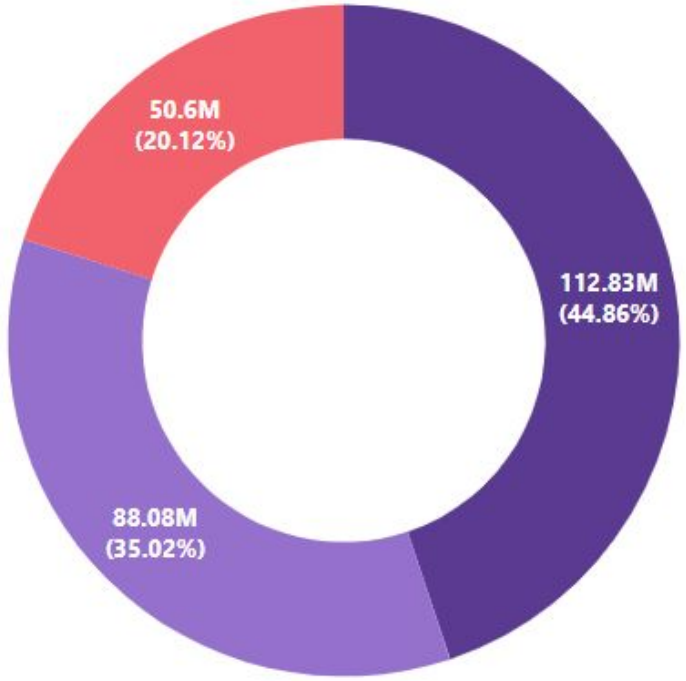
New visual Text box More visuals — Insert

New measure Quick measure — Calculations

Sensitivity — Sensitivity

Publish — Share

Back to report | SUM OF TOTAL_PRICE BY PAYMENT_METHOD



payment_method
Cash
Credit Card
Debit Card

Filters Visualizations Data

Page 2 of 2

100%

Type here to search

22°C Överväg. soligt

6:49 PM
6/6/2023

sales report 4 - Power BI Desktop

Search

ReDi-60088

File Home Insert Modeling View Optimize Help Format Data / Drill

Paste Cut Copy Format painter — Clipboard

Get data Excel workbook Data hub SQL Server Enter data Dataverse Recent sources — Data

Transform data Refresh — Queries

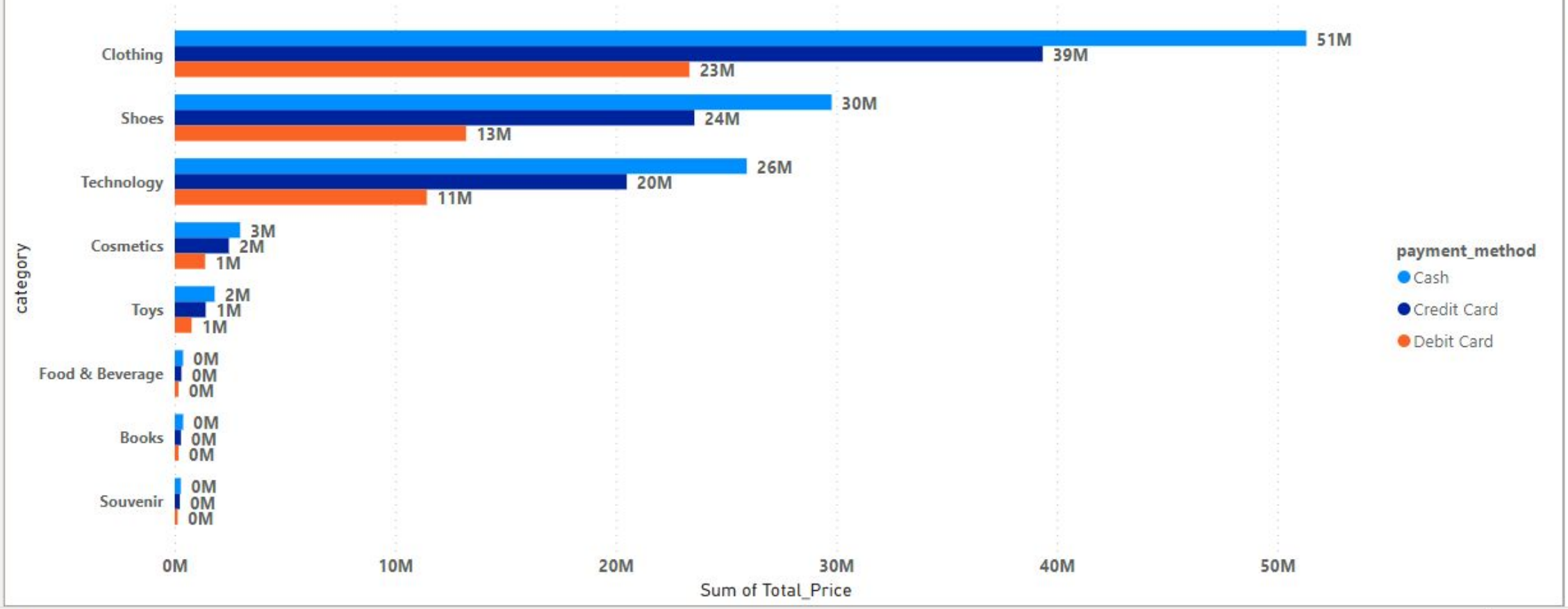
New visual Text box More visuals — Insert

New measure Quick measure — Calculations

Sensitivity — Sensitivity

Publish — Share

Back to report | SUM OF TOTAL_PRICE BY CATEGORY AND PAYMENT_METHOD



Filters Visualizations Data

Page 2 of 2

100%

Type here to search

22°C Överväg. soligt

7:37 PM
6/6/2023

sales report 4 - Power BI Desktop

Search

ReDi-60088

File Home Insert Modeling View Optimize Help Format Data / Drill

Paste Cut Copy Format painter — Clipboard

Get data Excel workbook Data hub SQL Server Enter data Dataverse Recent sources — Data

Transform data Refresh — Queries

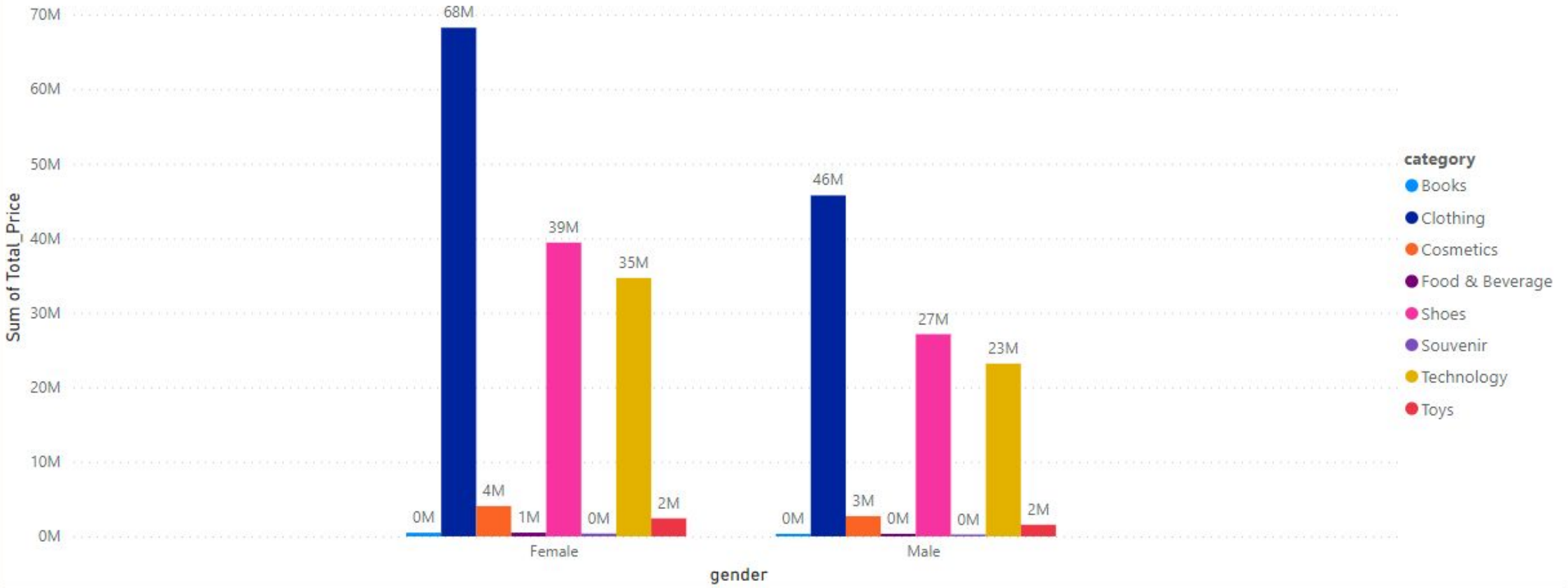
New visual Text box More visuals — Insert

New measure Quick measure — Calculations

Sensitivity — Sensitivity

Publish — Share

Back to report | SUM OF TOTAL_PRICE BY GENDER AND CATEGORY



Filters Visualizations Data

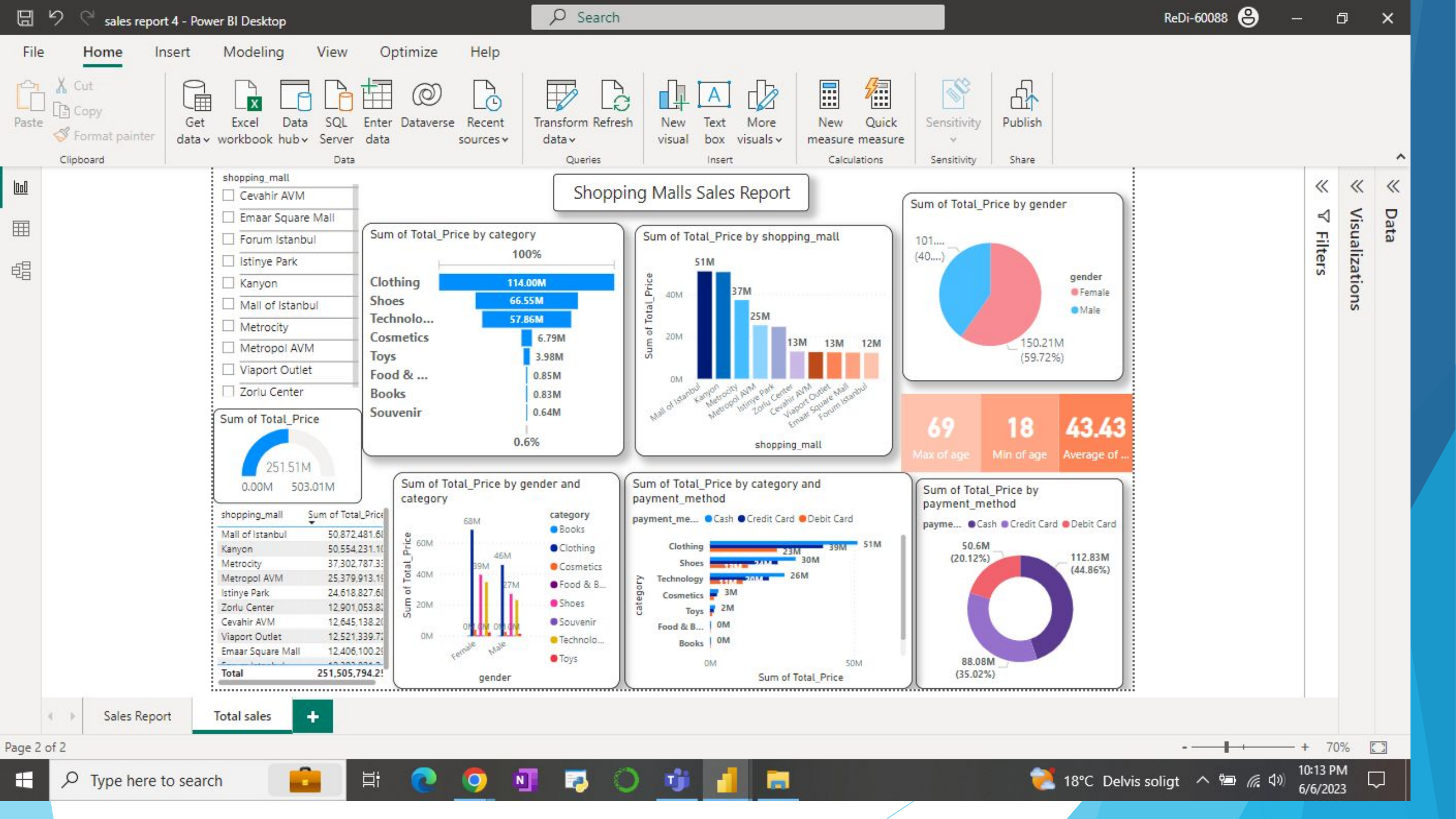
Page 2 of 2

100%

Type here to search

20°C Delvis soligt

9:19 PM 6/6/2023


sales report 4 - Power BI Desktop
Search
ReDi-60088
File
Home
Insert
Modeling
View
Optimize
Help
Paste
Cut
Copy
Format painter
Clipboard
Get data
Excel workbook
Data hub
SQL Server
Enter data
Dataverse
Recent sources
Data
Transform data
Refresh
Queries
New visual
Text box
More visuals
Insert
New measure
Quick measure
Calculations
Sensitivity
Sensitivity
Publish
Share
Shopping Malls Sales Report
shopping_mall
Cevahir AVM
Emaar Square Mall
Forum Istanbul
Istinye Park
Kanyon
Mall of Istanbul
Metrocity
Metropol AVM
Viaport Outlet
Zorlu Center
Sum of Total_Price by category
100%
Clothing 114.00M
Shoes 66.55M
Technolo... 57.86M
Cosmetics 6.79M
Toys 3.98M
Food & ... 0.85M
Books 0.83M
Souvenir 0.64M
0.6%
Sum of Total_Price by shopping_mall
Sum of Total_Price
51M
37M
25M
13M
13M
12M
40M
20M
0M
Mall of Istanbul
Kanyon
Metrocity
Metropol AVM
Istinye Park
Zorlu Center
Cevahir AVM
Viaport Outlet
Emaar Square Mall
Forum Istanbul
shopping_mall
Sum of Total_Price by gender
101....
(40....)
gender
Female
Male
150.21M
(59.72%)
Sum of Total_Price
251.51M
0.00M
503.01M
69
Max of age
18
Min of age
43.43
Average of ...
shopping_mall
Sum of Total_Price
Mall of Istanbul
Kanyon
Metrocity
Metropol AVM
Istinye Park
Zorlu Center
Cevahir AVM
Viaport Outlet
Emaar Square Mall
Total 251,505,794.25
Sum of Total_Price by gender and category
68M
46M
39M
27M
60M
40M
20M
0M
Sum of Total_Price
Female
Male
gender
category
Books
Clothing
Cosmetics
Food & B...
Shoes
Souvenir
Technolo...
Toys
Sum of Total_Price by category and payment_method
payment_me...
Cash
Credit Card
Debit Card
Clothing
Shoes
Technology
Cosmetics
Toys
Food & B...
Books
51M
39M
23M
30M
26M
3M
2M
0M
0M
category
0M
50M
Sum of Total_Price
Sum of Total_Price by payment_method
payme...
Cash
Credit Card
Debit Card
50.6M
(20.12%)
112.83M
(44.86%)
88.08M
(35.02%)
Filters
Visualizations
Data
Sales Report
Total sales
Page 2 of 2
70%
Type here to search
18°C Delvis soligt
10:13 PM
6/6/2023

Conclusion

Data Analytics recommendations

- The malls of minority of sales should make some marketing activities to increase its sales.
- Make promotions for the categories that has minority of sales to attract the customers and increase its sales.
- We recommend to make application to get the customer feedback in each mall for each category to enhance the services and the quality of the products.
- To improve the using of the credit cards the banks who owned the credit cards can make points program for each transaction paid by it and get free gifts from the minority categories of sales.

Thank You