Customers shopping dataset in Turkey

The following dataset represents the shopping habit and customers information for 11 reputed shopping malls in Turkey. This dataset contains customer ID , their age , gender , price spent on shopping , payment method, shopping date , invoice ID as well as their shopping category.

This dataset has been downloaded here from kaggle.com and then analyzed the shopping behavior of turkish customers, their shopping preferences, favourite shopping center, spending amount based on the shopping categories, shopping categories based on shopping centers etc.

For this analysis i used jupyter notebook and python libraries such as numpy, pandas, matplotlib and seaborn etc.

Downloading the Dataset

The shopping dataset has been loaded in dataset_url using opendatasets from the site kaggle and then has been seen how many files it contains. This dataset contains only 1 file named 'customer_shopping_data.csv'. The project has been given the name "zerotopandas-course-project-starter".

```
!pip install jovian opendatasets --upgrade --quiet
```

Let's begin by downloading the data, and listing the files within the dataset.

```
dataset_url = 'https://www.kaggle.com/datasets/mehmettahiraslan/customer-shopping-datas
```

```
import opendatasets as od
od.download(dataset_url)
```

```
Please provide your Kaggle credentials to download this dataset. Learn more: http://bit.ly/kaggle-creds
Your Kaggle username: tarifkhan93
Your Kaggle Key: ·····
Downloading customer-shopping-dataset.zip to ./customer-shopping-dataset
100%|████████████████████| 1.63M/1.63M [00:00<00:00, 50.9MB/s]
```

The dataset has been downloaded and extracted.

Let us save and upload our work to Jovian before continuing.

```
project_name = "zerotopandas-course-project-starter"
```

```
!pip install jovian --upgrade --quiet
```

```
import jovian
```

```
jovian.commit(project=project_name)
```

```
[jovian] Updating notebook "tarif9351/zerotopandas-course-project-starter" on https://jovian.com
[jovian] Committed successfully! https://jovian.com/tarif9351/zerotopandas-course-project-starter
```

```
'https://jovian.com/tarif9351/zerotopandas-course-project-starter'
```

Data Preparation and Cleaning

This dataset contains 99457 rows and 10 columns. Dataset has been prepared for analysis by eliminating the customers having the age 0 and more than 100. Rows also has been deleted having the price of 0 liras. The invoice date has been split into day,weekday,month and year to make it easy for required analysis.

```
!pip install pandas --upgrade --quiet
```

```
import pandas as pd
```

```
shopping_df = pd.read_csv('customer_shopping_data.csv')
```

```
shopping_df
```

	invoice_no	customer_id	gender	age	category	quantity	price	payment_method	invoice_date	shoppin
0	I138884	C241288	Female	28	Clothing	5	1500.40	Credit Card	5/8/2022	K

	invoice_no	customer_id	gender	age	category	quantity	price	payment_method	invoice_date	shoppin
1	I317333	C111565	Male	21	Shoes	3	1800.51	Debit Card	12/12/2021	F Ist
2	I127801	C266599	Male	20	Clothing	1	300.08	Cash	9/11/2021	Met
3	I173702	C988172	Female	66	Shoes	5	3000.85	Credit Card	16/05/2021	Metropo
4	I337046	C189076	Female	53	Books	4	60.60	Cash	24/10/2021	K
...	...	...	...	...	...	...	...	...	...	
99452	I219422	C441542	Female	45	Souvenir	5	58.65	Credit Card	21/09/2022	K
99453	I325143	C569580	Male	27	Food & Beverage	2	10.46	Cash	22/09/2021	F Ist
99454	I824010	C103292	Male	63	Food & Beverage	2	10.46	Debit Card	28/03/2021	Met
99455	I702964	C800631	Male	56	Technology	4	4200.00	Cash	16/03/2021	Istiny
99456	I232867	C273973	Female	36	Souvenir	3	35.19	Credit Card	15/10/2022	M Ist

99457 rows × 10 columns

```
import numpy as np
```

```
shopping_df.dtypes
```

```
invoice_no         object
customer_id        object
gender             object
age                 int64
category           object
quantity            int64
price             float64
payment_method     object
invoice_date       object
shopping_mall      object
dtype: object
```

```
shopping_df.drop(shopping_df[shopping_df.age== 0].index, inplace=True)
```

```
shopping_df.drop(shopping_df[shopping_df.age>= 100].index, inplace=True)
```

```
shopping_df.drop(shopping_df[shopping_df.price== 0].index, inplace=True)
```

```
import pandas as pd
```

```
shopping_df['invoice_date'] = pd.to_datetime(shopping_df.invoice_date, dayfirst = True)
```

```
shopping_df['invoice_year'] = pd.DatetimeIndex(shopping_df.invoice_date).year
shopping_df['invoice_month'] = pd.DatetimeIndex(shopping_df.invoice_date).month
shopping_df['invoice_day'] = pd.DatetimeIndex(shopping_df.invoice_date).day
shopping_df['invoice_weekday'] = pd.DatetimeIndex(shopping_df.invoice_date).weekday
```

```
shopping_df
```

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

99457 rows × 14 columns

```
shopping_df.drop(['invoice_date'], axis =1)
```

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

99457 rows × 13 columns

```
import jovian
```

```
jovian.commit()
```

```
[jovian] Updating notebook "tarif9351/zerotopandas-course-project-starter" on https://jovian.com
[jovian] Committed successfully! https://jovian.com/tarif9351/zerotopandas-course-project-starter
```

```
'https://jovian.com/tarif9351/zerotopandas-course-project-starter'
```

Exploratory Analysis and Visualization

Using python the maximum , minimum and average price of shopping by any individual rows have been found as well as the distribution of shopping price has been shown with histogram. The variation of average price with the shopping category also has been shown below with bar chart.

Let's begin by importing `matplotlib.pyplot and seaborn` .

```
import seaborn as sns
import matplotlib
import matplotlib.pyplot as plt
%matplotlib inline

sns.set_style('darkgrid')
matplotlib.rcParams['font.size'] = 14
matplotlib.rcParams['figure.figsize'] = (9, 5)
matplotlib.rcParams['figure.facecolor'] = '#00000000'
```

```
shopping_df.price.max()
```

```
5250.0
```

Maximum price spent in the given dataset was 5250 turkish liras

```
shopping_df.price.min()
```

```
5.23
```

Minimum price spent in the given dataset was 5.23 turkish liras

```
shopping_df.price.mean()
```

```
689.2563209226097
```

Average price spent in the given dataset was 689 turkish liras

```
shopping_df.quantity.max()
```

```
5
```

Maximum quantity bought by any customer is 5

```
shopping_df.quantity.min()
```

```
1
```

Minimum quantity bought by any customer is 1

```
shopping_df.quantity.mean()
```

```
3.003428617392441
```

Average quantity bought by any customer is 3

```
shopping_df.age.max()
```

```
69
```

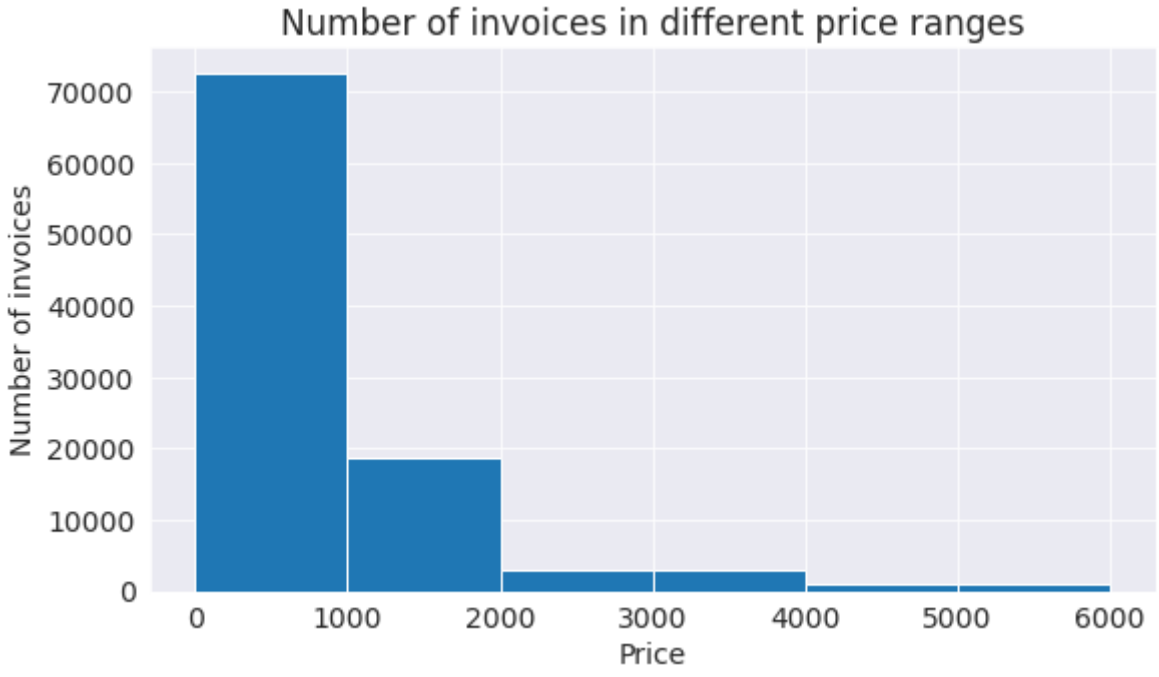
Maximum age of any customer is 69

```
shopping_df.age.min()
```

```
18
```

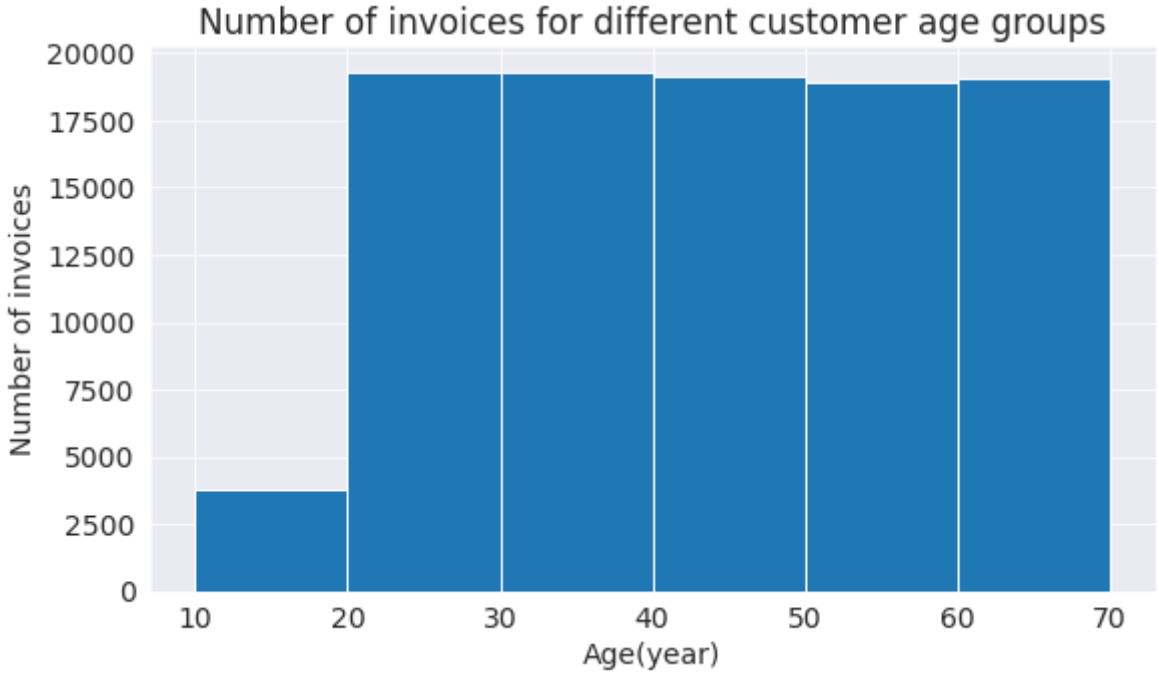
Minimum age of any customer is 18

```
plt.title('Number of invoices in different price ranges')
plt.xlabel('Price')
plt.ylabel('Number of invoices');
plt.hist(shopping_df.price , bins = [0,1000,2000,3000,4000,5000,6000]);
```



This histogram shows the number of invoices between the price groups of 0-1000 , 1001-2000, 2001-3000, 3001-4000, 4001-5000, 5001-6000 liras. Most of the customers lie between 0 to 1000 liras.

```
plt.title('Number of invoices for different customer age groups')
plt.xlabel('Age(year)')
plt.ylabel('Number of invoices');
plt.hist(shopping_df.age , bins = [10,20,30,40,50,60,70]);
```



This histogram shows the number of invoices between the age groups of 10-20 , 21-30, 31-40, 41-50, 51-60, 61-70 years old customers . Most of the customers lie between 20-40 years old in this graph.

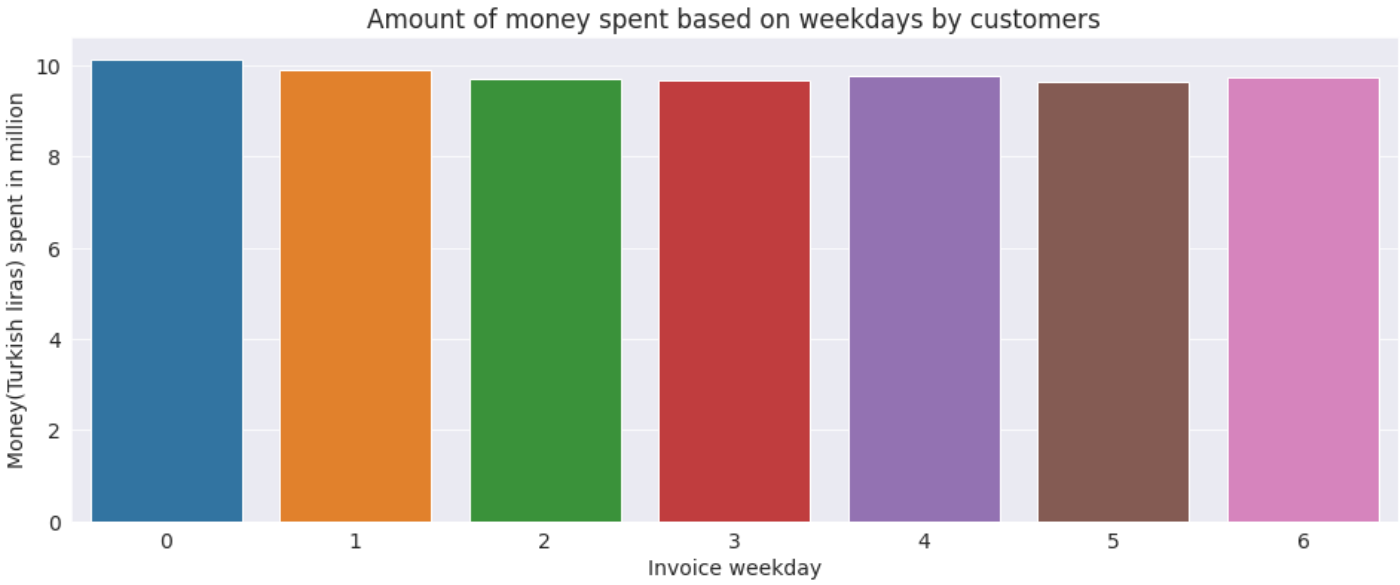
```
plt.figure(figsize=(16,6))
plt.title('Total price spent on various shopping categories')
ax = sns.barplot(x= shopping_df.category, y= shopping_df.price/1000000, estimator = np.
ax.set(xlabel='Shopping category' , ylabel ='Price(Turkish liras) in million')
```

```
[Text(0.5, 0, 'Shopping category'),
 Text(0, 0.5, 'Price(Turkish liras) in million')]
```

This barplot shows the total amount of turkish liras spent on different shopping categories. Most of the money was spent on clothings and the least was in souvenir and books.

```
plt.figure(figsize=(16,6))
plt.title('Amount of money spent based on weekdays by customers')
ax= sns.barplot(x=shopping_df.invoice_weekday, y=shopping_df.price/1000000, estimator =
ax.set(xlabel='Invoice weekday', ylabel ='Money(Turkish liras) spent in million')
```

```
[Text(0.5, 0, 'Invoice weekday'),
 Text(0, 0.5, 'Money(Turkish liras) spent in million')]
```



This barplot shows the total amount of turkish liras spent on different weekdays. Weekday 0 means sunday and it goes to 6 serially as monday , tuesday and so on. The graph says people spent slightly more money on weekend as it is quite normal to go to shopping more on the weekend.

insights:

- Among the shopping category people spent more money on average in buying tech products

- Most of the people spent less than 1000 liras on shopping

- Most of the customers were from 20-30 and 30-40 years age group

- Least amount of money on average was spent on buying souvenir products

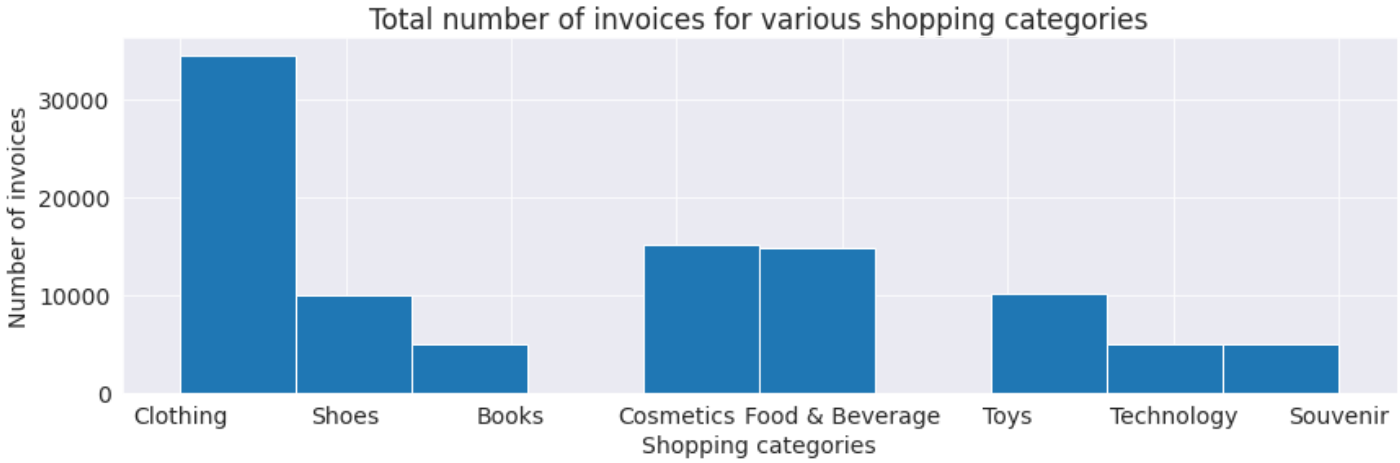
- The highest amount of payment done for shopping was 5250 liras

- People spent more money on technology , shoe and cosmetics on year 2022 than 2021 whereas they spent less on clothing

- People spent slightly more money on weekend as more people tend to go for shopping on weekend

Let us see which category of shopping products had the most customers

```
plt.figure(figsize=(14,4))
plt.title('Total number of invoices for various shopping categories')
plt.xlabel('Shopping categories')
plt.ylabel('Number of invoices');
plt.hist(shopping_df.category);
```

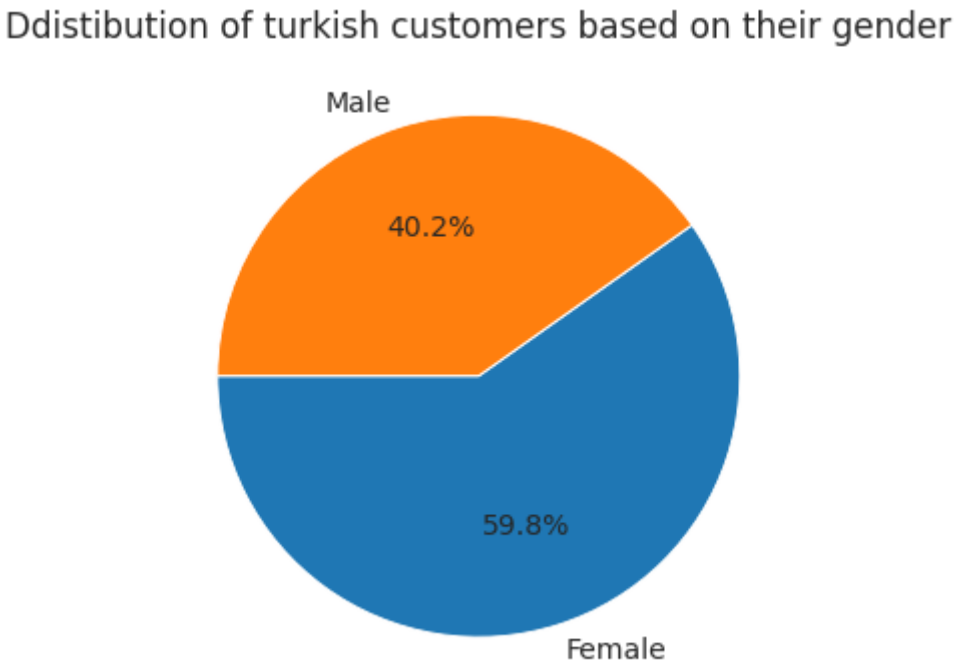


This histogram shows the number of invoices for each shopping categories. Like price, the number of customers is also more in clothing section. Most of the customers went to buy clothing and the least amount of people bought books.

Let's see people from which gender went more on shopping

```
gender = shopping_df.gender.value_counts()
```

```
plt.figure(figsize=(12,6))
plt.title("Ddistibution of turkish customers based on their gender")
plt.pie(gender, labels= gender.index, autopct='%1.1f%%',startangle=180);
```



This piechart shows the percentage of male and female customers in the dataset and it says almost 60% customers are female. We can come to the conclusion that most of the customers were female

Let's see the most polular and least popular shopping category in percentage

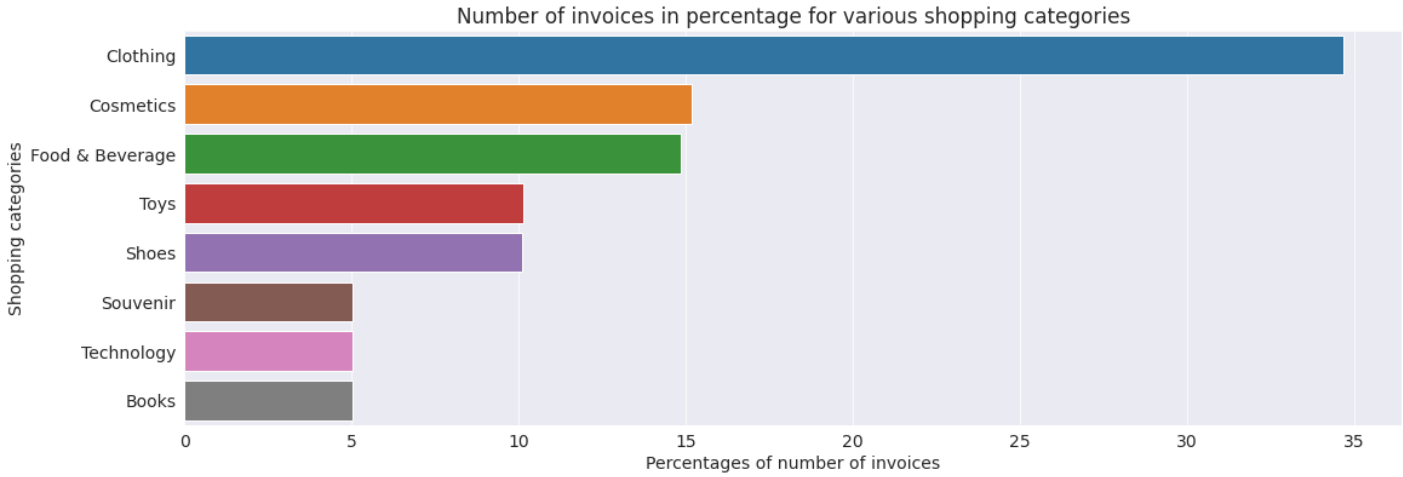
```
shopping_df.category.value_counts().sum()
```

```
99457
```

```
shop_type = shopping_df.category.value_counts() * 100 / shopping_df.category.value_coun
```

```
plt.figure(figsize=(18,6))
plt.title('Number of invoices in percentage for various shopping categories')
ax = sns.barplot(y = shop_type.index, x=shop_type)
ax.set(xlabel='Percentages of number of invoices' , ylabel ='Shopping categories')
```

```
[Text(0.5, 0, 'Percentages of number of invoices'),
 Text(0, 0.5, 'Shopping categories')]
```

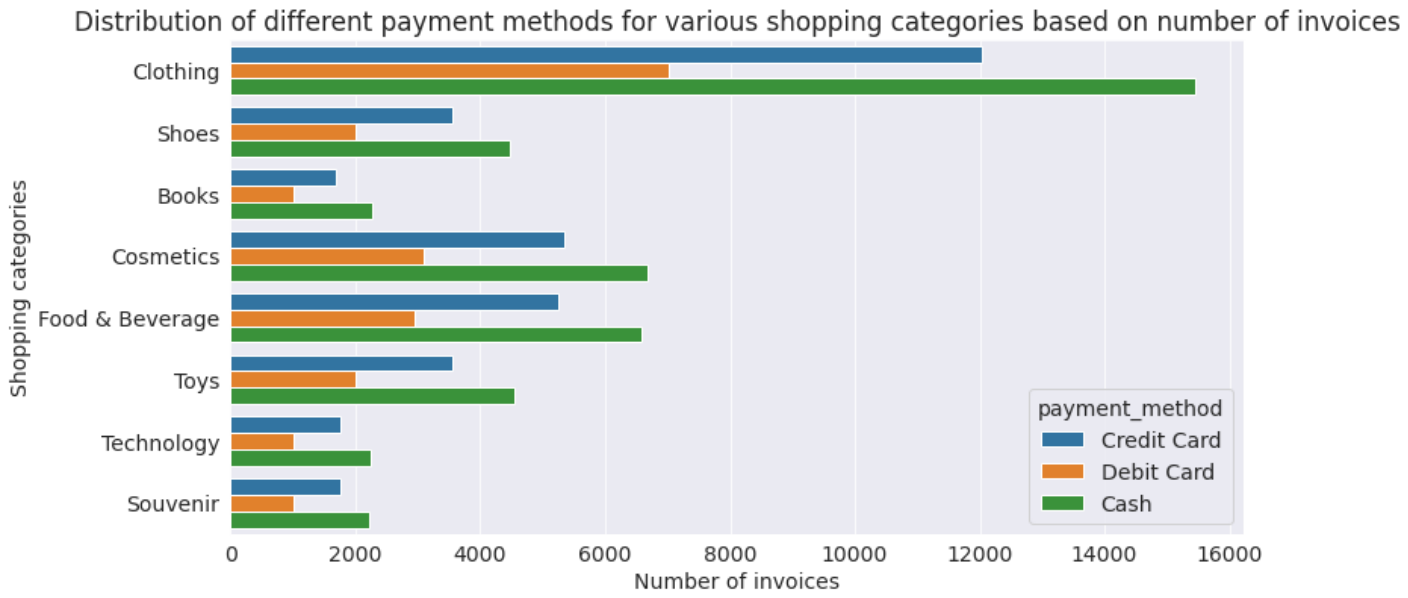


This barplot shows the percentage of number of invoices for each shopping categories. Almost 34% of the shopping products were clothing which is the highest and 5% were books, technology and souvenir separately which are the lowest

Let's see the most popular payment method based on these shopping categories

```
plt.figure(figsize=(12,6))
plt.title('Distribution of different payment methods for various shopping categories ba
ax= sns.countplot(data=shopping_df, x=None, y= shopping_df.category, hue= shopping_df.p
ax.set(xlabel='Number of invoices' , ylabel ='Shopping categories')
```

```
[Text(0.5, 0, 'Number of invoices'), Text(0, 0.5, 'Shopping categories')]
```



This countplot shows the number of invoices for each shopping categories based on customers payment method. Most of the people chose to pay with cash in each category of shopping and least number of payment were done with debit cards

Let's see the distribution of customers gender based on shopping categories

```
plt.figure(figsize=(12,6))
plt.title('Distribution of customers gender for various shopping categories based on nu
ax = sns.countplot(data=shopping_df, x=None, y= shopping_df.category, hue= shopping_df.
ax.set(xlabel='Number of invoices' , ylabel ='Shopping categories')
```

```
[Text(0.5, 0, 'Number of invoices'), Text(0, 0.5, 'Shopping categories')]
```

This countplot shows the number of invoices for each shopping categories based on customers gender. In every category of shopping female customers outnumbered male customers.

Let us save and upload our work to Jovian before continuing

```
import jovian
```

```
jovian.commit()
```

```
[jovian] Updating notebook "tarif9351/zerotopandas-course-project-starter" on
https://jovian.com
[jovian] Committed successfully! https://jovian.com/tarif9351/zerotopandas-course-
project-starter
```

```
'https://jovian.com/tarif9351/zerotopandas-course-project-starter'
```

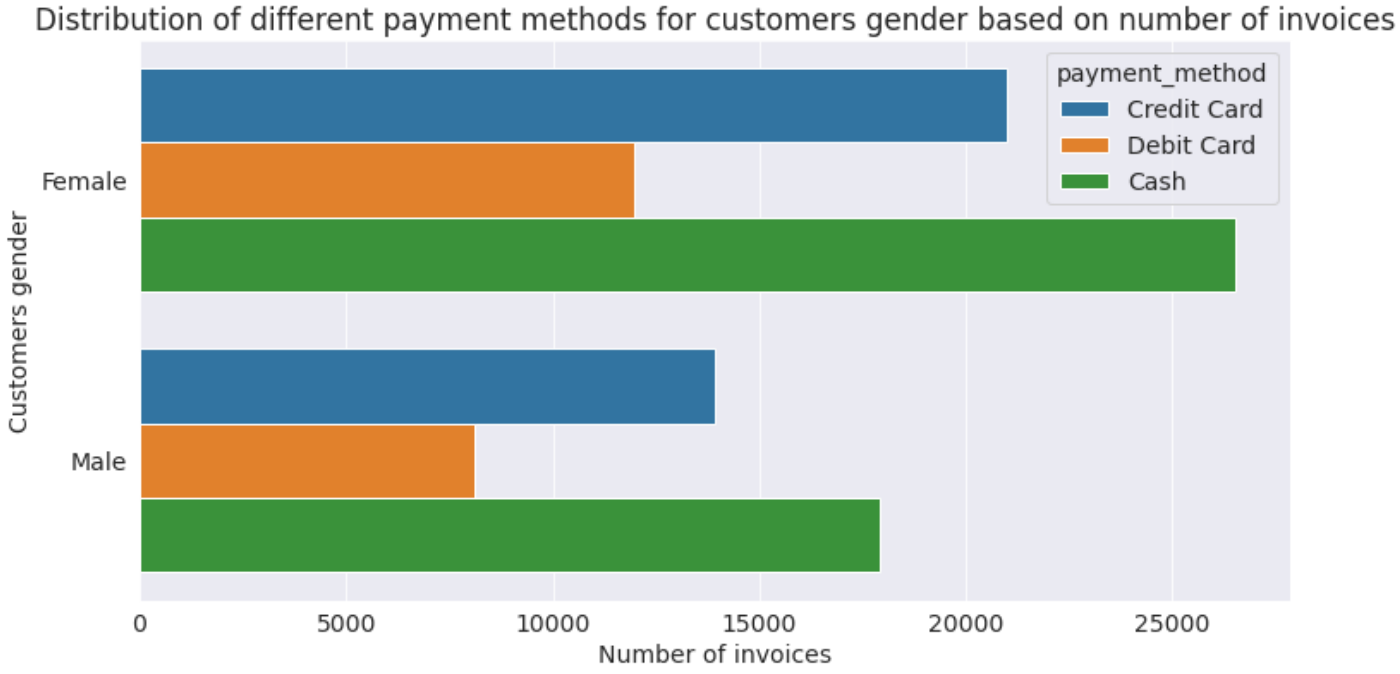
Let's ask some questions and try to give their answer from our dataset using matplotlib and seaborn graphs

Q1: What is/are the most common payment method for male and female customers ?

```
plt.figure(figsize=(12,6))
plt.title('Distribution of different payment methods for customers gender based on numb
```

```
ax = sns.countplot(data=shopping_df, x=None, y= shopping_df.gender, hue= shopping_df.pa
ax.set(xlabel='Number of invoices' , ylabel ='Customers gender')
```

```
[Text(0.5, 0, 'Number of invoices'), Text(0, 0.5, 'Customers gender')]
```

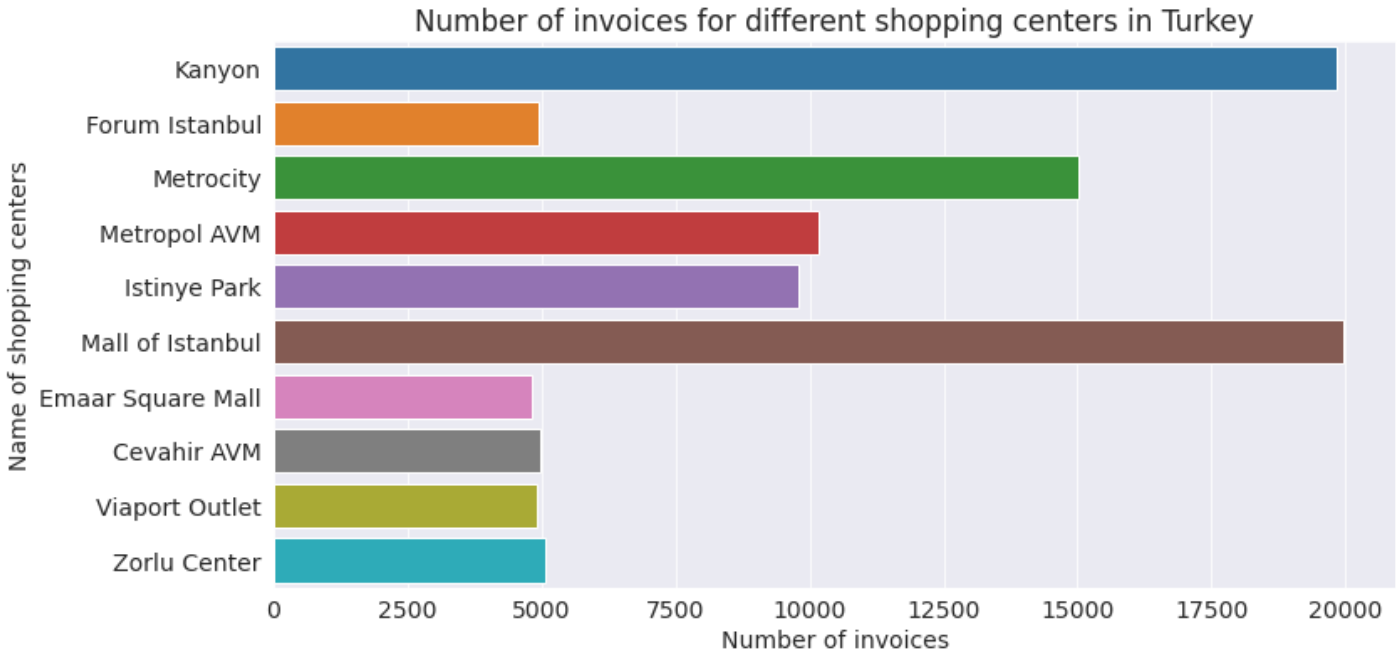


Ans: Most common payment method for both male and female customers are cash followed by credit card

Q2: Which shopping center had the highest number of customers ?

```
plt.figure(figsize=(12,6))
plt.title('Number of invoices for different shopping centers in Turkey')
ax = sns.countplot(data=shopping_df, x=None, y= shopping_df.shopping_mall)
ax.set(xlabel='Number of invoices' , ylabel ='Name of shopping centers')
```

```
[Text(0.5, 0, 'Number of invoices'), Text(0, 0.5, 'Name of shopping centers')]
```

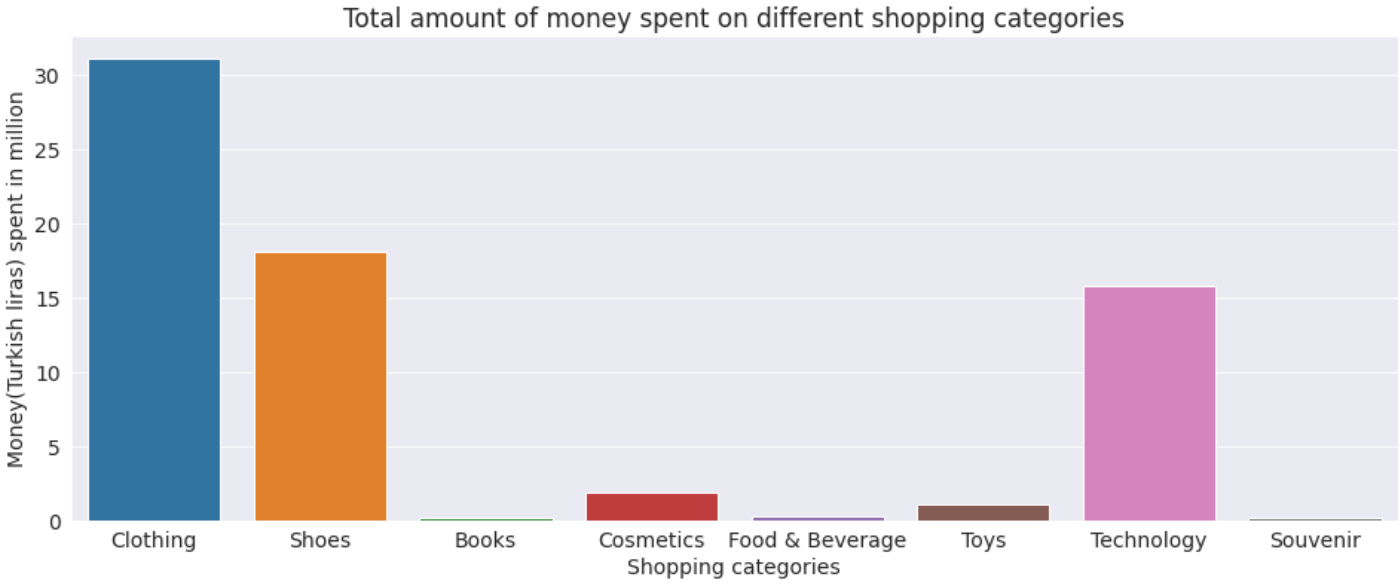


Ans: Mall of Istanbul had the most customers followed by Kanyon shopping center

Q3: In which two categories of shopping most of the money were spent and what are the amount ?

```
plt.figure(figsize=(16,6))
plt.title('Total amount of money spent on different shopping categories')
ax = sns.barplot(x=shopping_df. category, y=shopping_df.price/1000000 ,estimator = np.s
ax.set(xlabel='Shopping categories' , ylabel ='Money(Turkish liras) spent in million')
```

```
[Text(0.5, 0, 'Shopping categories'),
 Text(0, 0.5, 'Money(Turkish liras) spent in million')]
```

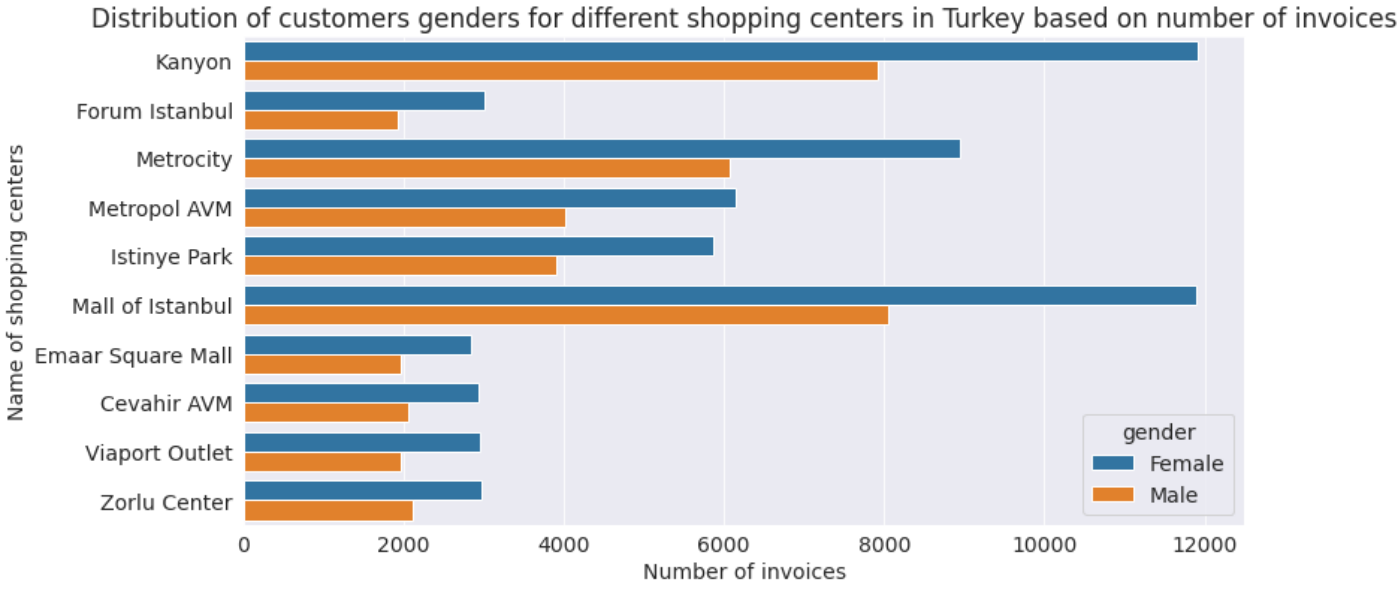


Ans:People spent more than 30 million turkish liras on buying clothes and 18 million turkish liras on buying shoes which are the highest

Q4: Which shopping centers had the highest number of male and female customer respectively ?

```
plt.figure(figsize=(12,6))
plt.title('Distribution of customers genders for different shopping centers in Turkey b
ax=sns.countplot(data=shopping_df, x=None, y= shopping_df.shopping_mall, hue= shopping_
ax.set(xlabel='Number of invoices' , ylabel ='Name of shopping centers')
```

```
[Text(0.5, 0, 'Number of invoices'), Text(0, 0.5, 'Name of shopping centers')]
```

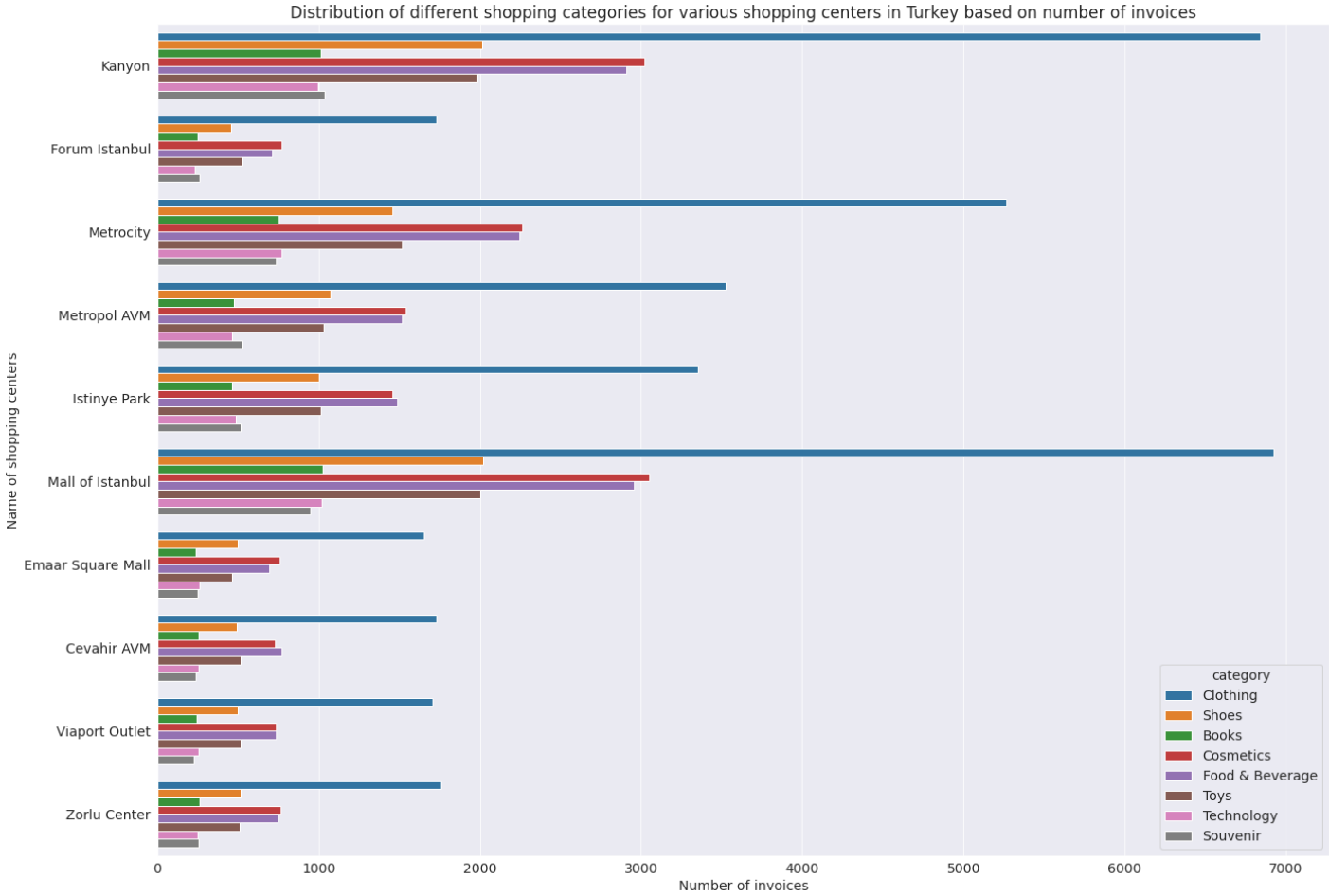


Ans: Mall of Istanbul had the highest male customers while Kanyon had the highest female customers

Q5: Which shopping mall sold the most number of books and which sold the least number of clothes ?

```
plt.figure(figsize=(22,16))
plt.title('Distribution of different shopping categories for various shopping centers i
ax = sns.countplot(data=shopping_df, x=None, y= shopping_df.shopping_mall, hue= shoppin
ax.set(xlabel='Number of invoices' , ylabel ='Name of shopping centers')
```

```
[Text(0.5, 0, 'Number of invoices'), Text(0, 0.5, 'Name of shopping centers')]
```



Ans: Mall of Istanbul sold the highest number of book and Emaar square mall sold the least number of clothes

Let us save and upload our work to Jovian before continuing.

```
import jovian
```

```
jovian.commit()
```

```
[jovian] Updating notebook "tarif9351/zerotopandas-course-project-starter" on https://jovian.com
[jovian] Committed successfully! https://jovian.com/tarif9351/zerotopandas-course-project-starter
```

```
'https://jovian.com/tarif9351/zerotopandas-course-project-starter'
```

Inferences and Conclusion

According to this analysis the most popular shopping centers of turkey are Kanyon and Mall os Istanbull . Turkish females shop more frequently than male and most of them are on clothings. In turkey most of the payments are done with cash followed by credit and debit card .

```
import jovian
```

```
jovian.commit()
```

```
[jovian] Updating notebook "tarif9351/zerotopandas-course-project-starter" on https://jovian.com
[jovian] Committed successfully! https://jovian.com/tarif9351/zerotopandas-course-project-starter
```

```
'https://jovian.com/tarif9351/zerotopandas-course-project-starter'
```

References and Future Work

Using these dataset we can analyze the shopping sentiment of the people of a country. This dataset shows people now buy less books and spend more on other things which is pathetic. By collecting data of coming years we can predict the change in shopping trend of a country. This analysis is done with the help of available online contents, one of which is given below - https://www.youtube.com/watch?v=BaV4PRXYNIY&list=PLyMom0n-MBrpzC91Uo560S4VbsiLYtCwo

```
import jovian
```

```
jovian.commit()
```

```
[jovian] Updating notebook "tarif9351/zerotopandas-course-project-starter" on https://jovian.com
[jovian] Committed successfully! https://jovian.com/tarif9351/zerotopandas-course-project-starter
```

```
'https://jovian.com/tarif9351/zerotopandas-course-project-starter'
```

```
```