Sales and Customer Dataset from Kaggle

Arindam Roy

The below Sales and Customer Data have been collected from Kaggle website. This dataset is about sales transaction data from Istanbul and is publicly available.

https://www.kaggle.com/datasets/dataceo/sales-and-customer-data/data

This dataset has 2 csv files and 11 columns. Out of 11 columns, 8 columns were text based, 2 columns were numeric and 1 decimal number-based column. It had 99,457 sales transactions.

The first dataset, as mentioned below contains customer information.

Customer id: Customer identification number.

Gender: Gender (Male/Female)

Age: Customer age.

Payment method: Payment used by customer.

age	customer_id	gender	payment_method
27	C999853	Female	Cash
27	C999854	Male	Cash
29	C999765	Female	Cash
29	C999974	Female	Cash
30	C999662	Male	Cash
31	C999457	Female	Cash
43	C999574	Female	Cash
46	C999700	Female	Cash
49	C999685	Male	Cash
50	C999687	Female	Cash
57	C999995	Female	Cash
38	C999653	Female	Credit Card
56	C999683	Female	Credit Card
58	C999586	Male	Credit Card
68	C999631	Male	Credit Card
27	C999770	Male	Debit Card
41	C999910	Male	Debit Card
49	C999976	Female	Debit Card
56	C999810	Female	Debit Card
61	C999886	Male	Debit Card

Second dataset is about Sales transactions and below is the column description:

Invoice no: Invoice identification number.

Customer id: Customer identification number.

Category: General item categorization groups

Quantity: Number of products

Price: Price of each product.

Invoice date: Date of purchase.

Shopping mall: Shopping mall location

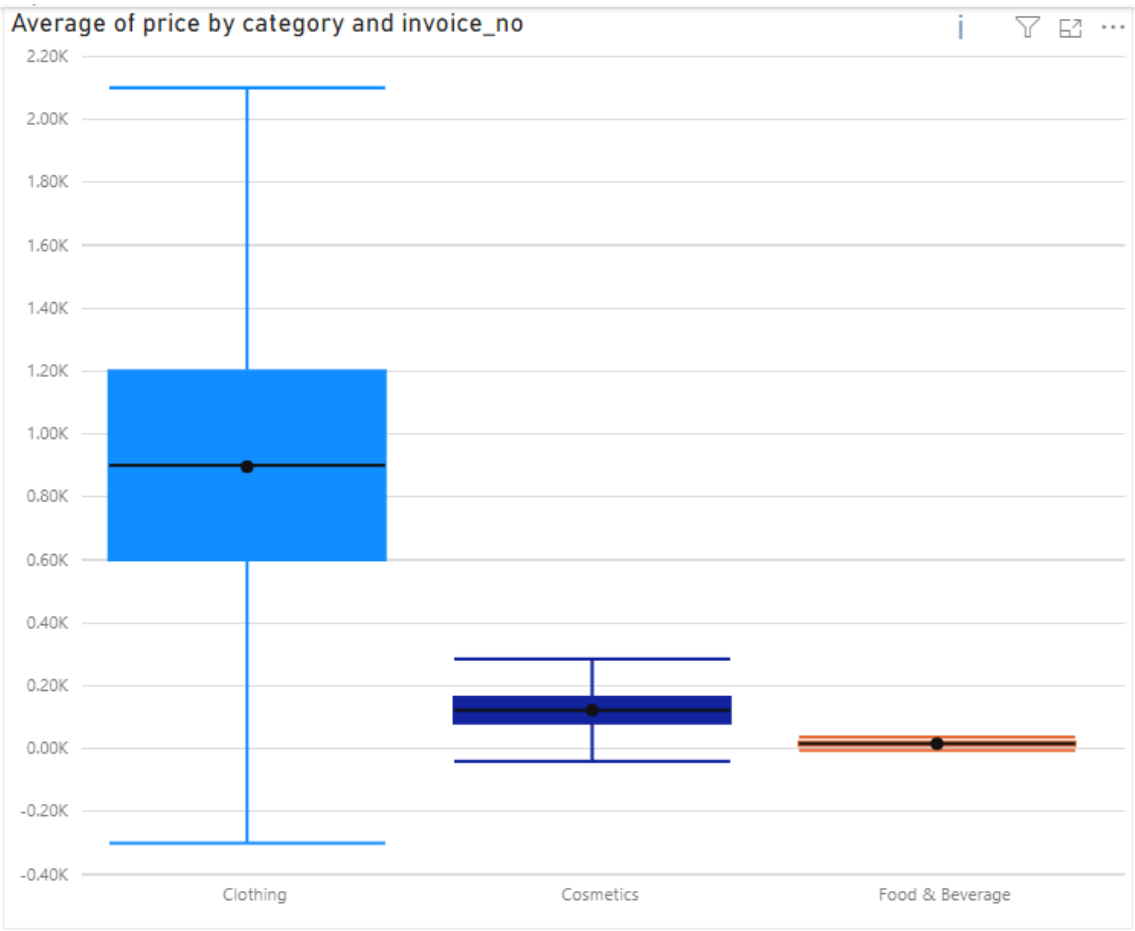
category	customer_id	Year	Quarter	Month	Day	invoice_no	price	quantity	shopping_mall
Books	C631481	2022	Qtr 2	May	14	I999769	60.60	4	Kanyon
Clothing	C190133	2021	Qtr 4	December	3	I999670	900.24	3	Metrocity
Clothing	C249289	2021	Qtr 3	August	28	I999721	600.16	2	Metrocity
Clothing	C286501	2022	Qtr 4	December	23	I999779	900.24	3	Emaar Square Mall
Clothing	C695980	2021	Qtr 4	October	13	I999959	1,200.32	4	Kanyon
Clothing	C983564	2023	Qtr 1	February	1	I999562	300.08	1	Mall of Istanbul
Food & Beverage	C113827	2022	Qtr 4	November	18	I999457	15.69	3	Mall of Istanbul
Food & Beverage	C161331	2023	Qtr 1	January	15	I999948	15.69	3	Mall of Istanbul
Food & Beverage	C265085	2022	Qtr 3	August	20	I999952	15.69	3	Metropol AVM
Food & Beverage	C276179	2022	Qtr 1	January	4	I999692	10.46	2	Metropol AVM
Shoes	C120008	2021	Qtr 3	August	30	I999922	1,800.51	3	Metrocity
Shoes	C200815	2021	Qtr 4	October	24	I999621	1,800.51	3	Kanyon
Shoes	C309718	2023	Qtr 1	January	11	I999868	1,200.34	2	Viaport Outlet
Souvenir	C180131	2021	Qtr 2	June	20	I999852	58.65	5	Metropol AVM
Souvenir	C379550	2021	Qtr 2	May	25	I999572	35.19	3	Mall of Istanbul
Technology	C195931	2021	Qtr 1	January	23	I999972	2,100.00	2	Istinye Park
Technology	C248542	2023	Qtr 1	January	11	I999973	1,050.00	1	Mall of Istanbul
Technology	C299697	2021	Qtr 2	April	9	I999994	3,150.00	3	Kanyon
Toys	C581153	2022	Qtr 3	August	21	I999819	71.68	2	Emaar Square Mall
Toys	C925896	2021	Qtr 3	August	16	I999956	143.36	4	Metrocity

Following ETL Process have been done on the dataset:

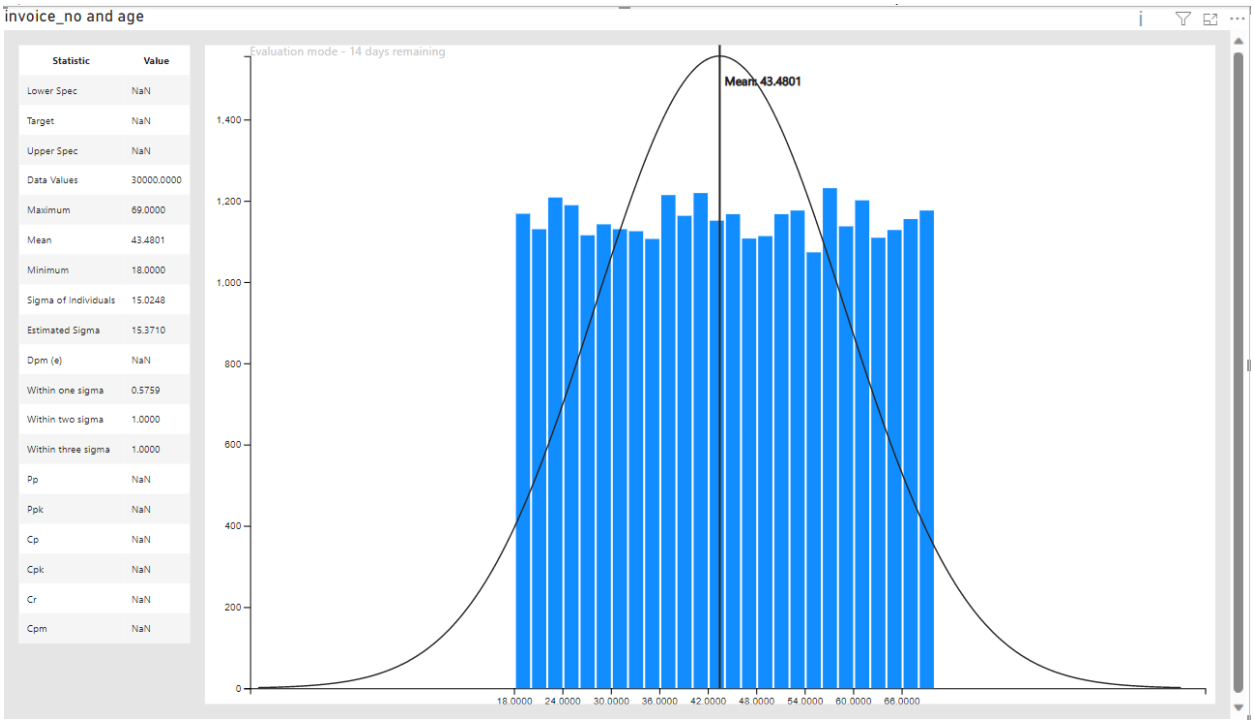
1. The invoice date column was in Text datatype and was formatted like dd-mm-yyyy. The datatype was converted to date datatype with mm/dd/yyyy format.
2. The Age column has 119 missing values. Those missing values were imputed with median values in the Age column.

Following EDA process has been performed on the dataset:

1) Univariate Data Analysis - The box and whisker plot are drawn on 3 categories of data a. Clothing, b. cosmetics and c. Food and Beverage. These are the 3 top categories with the count of sales data.



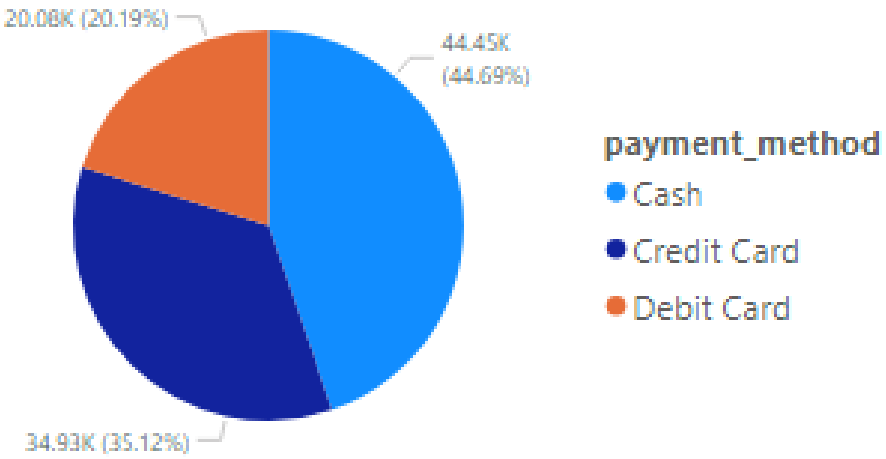
2) The next univariate chart is histogram using additional data visuals from PowerBI. The below chart is the age wise number of transactions in the dataset.



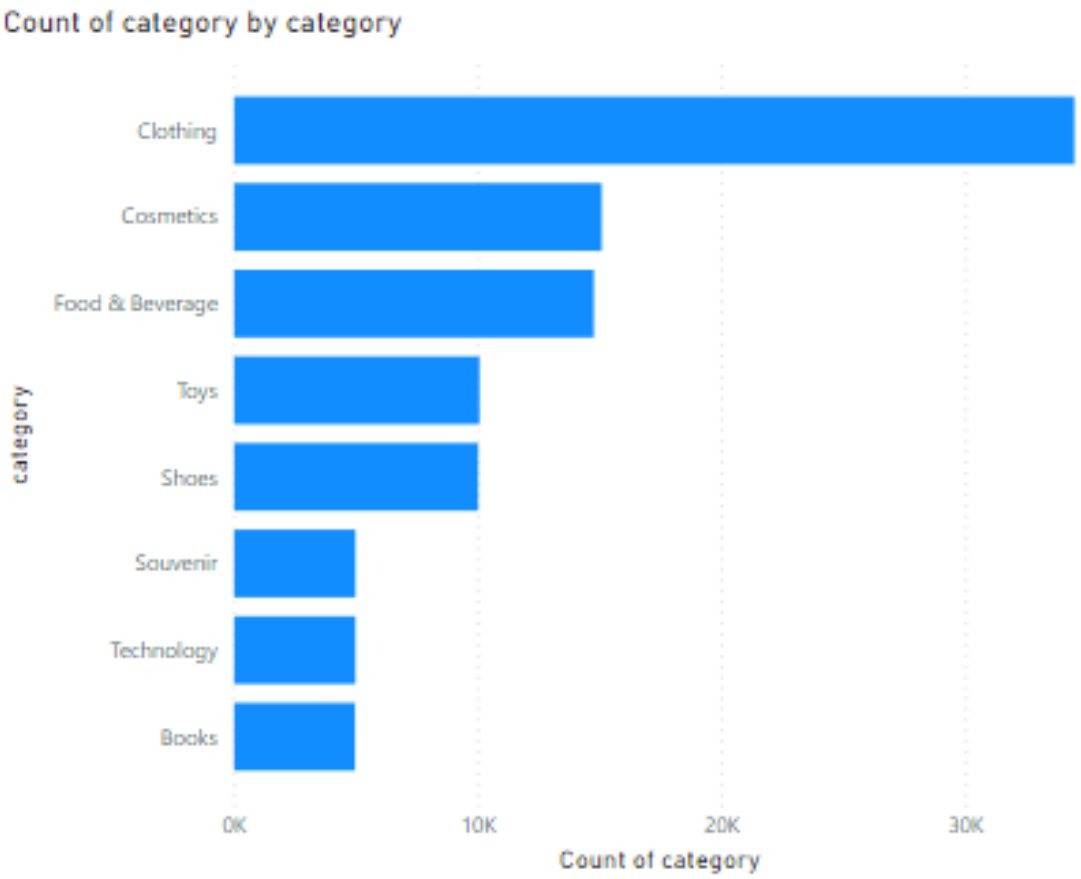
Statistic	Value
Lower Spec	NaN
Target	NaN
Upper Spec	NaN
Data Values	30000.0000
Maximum	69.0000
Mean	43.4801
Minimum	18.0000
Sigma of Individuals	15.0248
Estimated Sigma	15.3710
Dpm (e)	NaN
Within one sigma	0.5759
Within two sigma	1.0000
Within three sigma	1.0000
Pp	NaN
Ppk	NaN
Cp	NaN
Cpk	NaN
Cr	NaN
Cpm	NaN

3) Categorical Data Analysis:
    a. A pie chart was drawn to analyze the payment method frequency in the dataset. It was observed still people in that area still prefer cash transactions (44.69%) followed by credit and debit cards.

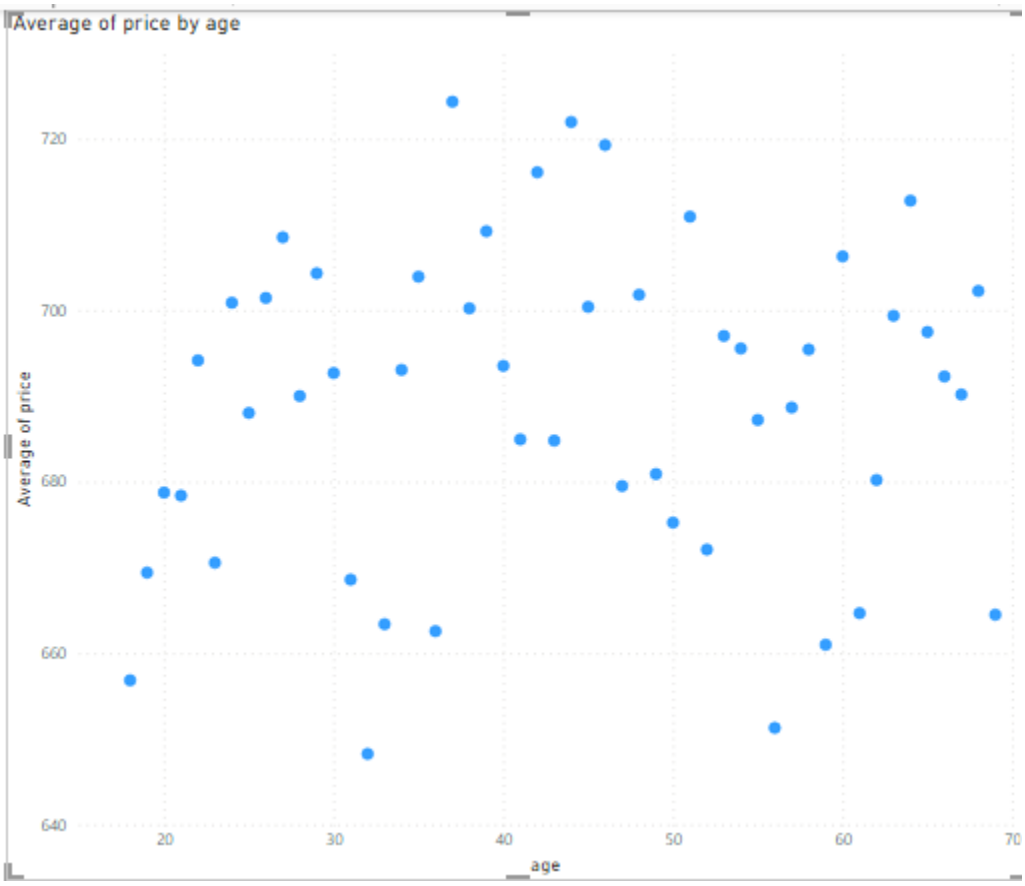
%payment_method by diff payment_method



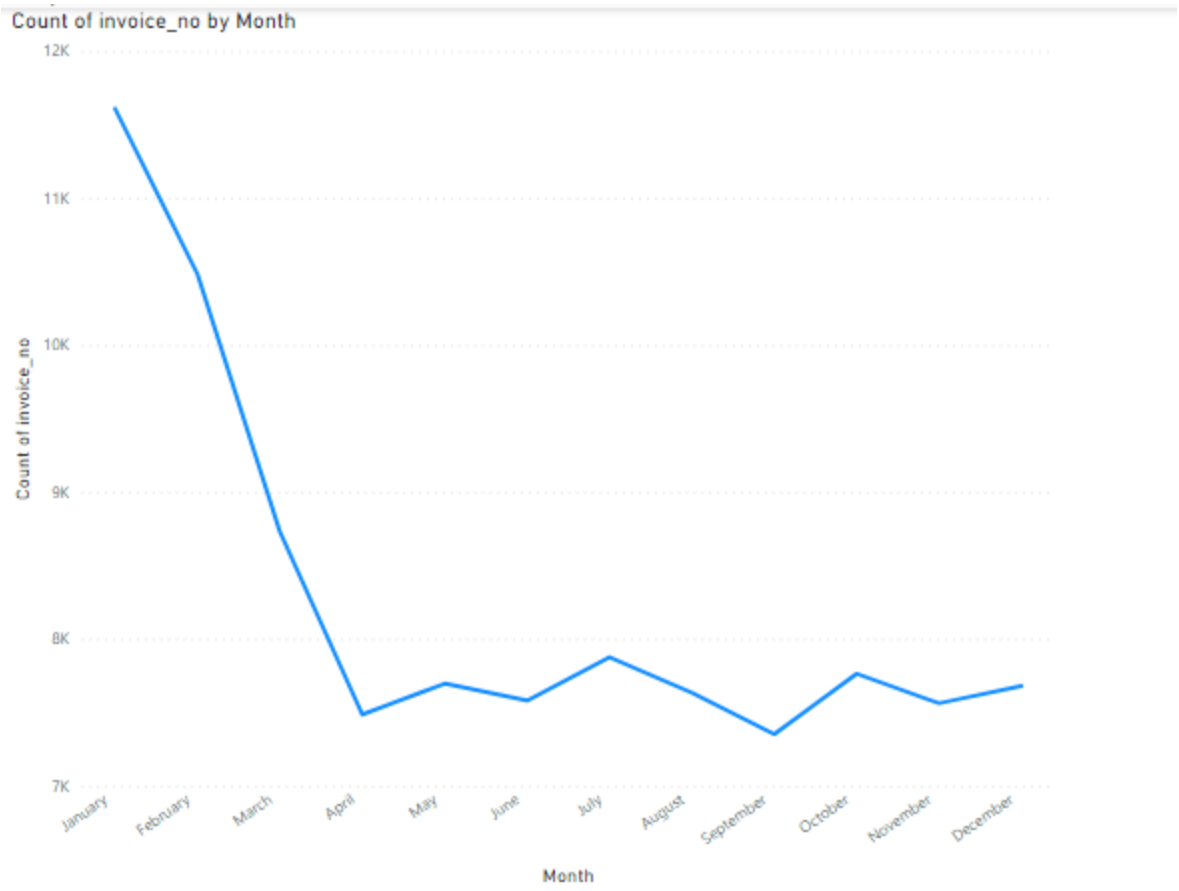
4) The following bar chart was created to analyze which product category has the highest number sales. It was observed that Clothing, cosmetics and Food/beverage have a greater number of sales.



5) Following Scatter plot has been drawn to analyze the spending trends based on the age of the Customers. Although there was no strong relation found between age and spending amount, It was observed the middles aged people likely to spend more than old aged and young people in the store.



6) The following is the line chart drawn to analyze which months the sales are more in the region. It looks like January is the month with the highest selling month followed by February and March. After April, the sales became almost flat.



Summary

1) The Sales and Customer data are collected from dales data in different stored in Istanbul. People are still carrying cash and doing the transactions.
2) People are mostly buying clothing and cosmetics, 50% of total transactions.
3) Month wise, January, February and March are the 3 months where we could see the highest number of sales in the Dataset. There are 3 years of data collected in the dataset. The sales in 2023 was very low, which has been observed too.
4) Middle aged people are likely to purchase more as there is weak relationship found in the dataset.
5) Clothing, Cosmetics and Food and beverages are the 3 top selling categories in the dataset.