Data Scientist II Technical Challenge

Task 1: Exploratory Data Analysis

Cristina Sánchez Maíz | csmaiz@gmail.com | LinkedIn

Table of Contents

Tasks:
You will be provided with a dataset containing information about Customer transactions.

- Load the dataset and perform basic data cleaning (e.g., handling missing values, correcting data types).
- Conduct exploratory data analysis to understand the main characteristics of the data.
- Visualize key insights using appropriate plots (e.g., histograms, bar charts, scatter plots).

Deliverables:

- Summary statistics of the dataset.
- At least three different visualizations with explanations.

Summary statistics of the dataset

Table 1 shows the summary statistics for numeric and datetime columns of the dataset.

	transaction_date	amount	customer_age	customer_income
count	980	980	980	980
mean	23/01/2024 03:11	987.45 €	43.59	70,979.10 €
min	31/07/2023 00:00	248.79 €	18.00	20,111.77 €
25%	22/10/2023 18:00	733.03 €	31.00	46,297.97 €
50%	20/01/2024 12:00	977.32 €	43.00	70,464.16 €
75%	24/04/2024 06:00	1,250.87 €	57.00	95,933.47 €
max	29/07/2024 00:00	1,679.68 €	69.00	119,941.30 €
std		333.31 €	15.05	28,854.58 €

Table 1: summary statistics for numeric and datetime columns

- **Sample Size:** The dataset contains 980 observations.
- **Transaction Date:** The data spans from July 31, 2023, to July 29, 2024.
- **Amount:**
    - Average transaction amount is €987.45.
    - There is a wide range of transaction amounts, from €248.79 to €1,679.68.
    - The distribution is slightly skewed to the right, given the difference between the mean and median.
- **Customer Age:**
    - Average customer age is 43.59 years.
    - The age range is from 18 to 69 years.
- **Customer Income:**
    - Average customer income is €70,979.10.
    - There's a wide range of incomes, from €20,111.77 to €119,941.30.
    - The distribution is likely slightly to the right, given the difference between the mean and median.

Note on customer-related columns

As an example, Table 2 shows the registers of customer with *customer_id*=5. While *transaction_date* spans from September 15, 2023 to July 03, 2024 (about 10 months), *customer_age* seems to be randomly distributed from 24 to 62. Similarly, *customer_income* presents a high random variation, from 34,089.18 to 116,323.81.

	transaction_date	customer_age	customer_income
698	2023-09-15	43	116323.81
180	2023-11-09	24	91336.85
278	2023-12-07	26	22127.70
846	2024-01-16	60	26894.56
446	2024-03-19	62	22516.76
818	2024-04-18	36	100291.40
726	2024-04-25	24	96152.29
864	2024-04-30	37	89794.76
947	2024-05-12	33	34089.18
655	2024-06-05	34	105903.04
882	2024-07-03	28	69336.10

Table 2: Records for customer_id=5

Consequently, I decided not to make calculations that involve the join of *customer_age* and/or *customer_income* with other columns since the random nature of these variables could lead to messy results.

Visualizations with explanations

Numerical variables

Figure 1 is a combination of the histogram (blue) with the correlation plot (pink). The cell (i,j) contains the correlation between variables *i* and *j*, except when *i*=*j*. In that case, the diagonal contains the histogram of the numerical *i*.

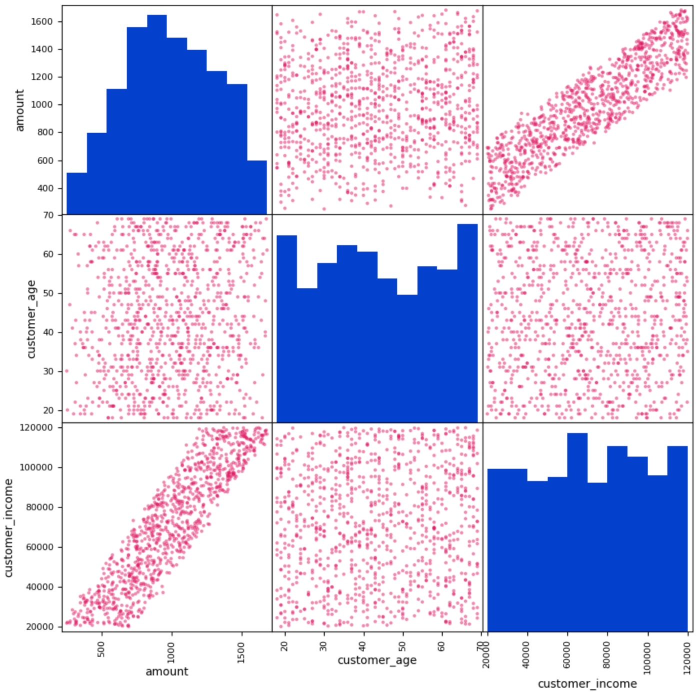

Figure 1: Histogram and correlation plot for numerical variables

Histograms:

- The distribution of *amount* is approximately normal, with a mean of ~1,000 and a standard deviation of ~300.
- There is no apparent mode in the distribution of *customer_income*, indicating a lack of concentration in any particular value.
- The distribution of *customer_age* is relatively flat, suggesting a uniform distribution across the observed range.

Correlation plots:

Note that the variables *customer_income* and *amount* are highly correlated. The correlation matrix shows a value of 0.9 (Table 3).

	amount	customer_age	customer_income
amount	1.000	0.098	0.901
customer_age	0.098	1.000	0.051
customer_income	0.901	0.051	1.000

Table 3: Correlation between numerical variables

Figure 2 shows the boxplots of numerical variables. The median line is centrally located within the interquartile range (IQR), and the upper and lower whiskers are approximately equal in length, suggesting a symmetric distribution, consistent with the findings from the histograms.There are no outliers in the dataset, as confirmed by the boxplot.

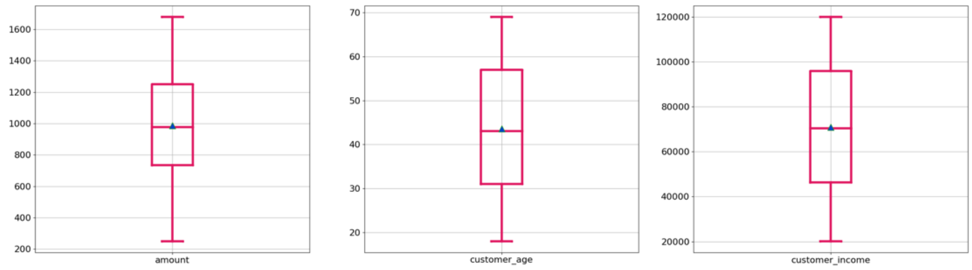

Figure 2: Boxplot of numerical variables

Categorical variables

Imagine we want to know if people prefer to purchase a product category using a specific payment method. A grouped bar graph like Figure 3 is representative of the data and effectively communicates the relationship between payment method and product category. Each group of bars represents a different payment method, and the bars within each group represent the different values for the product category. The bars within each group are of similar height, indicating no apparent differences in usage across payment methods.

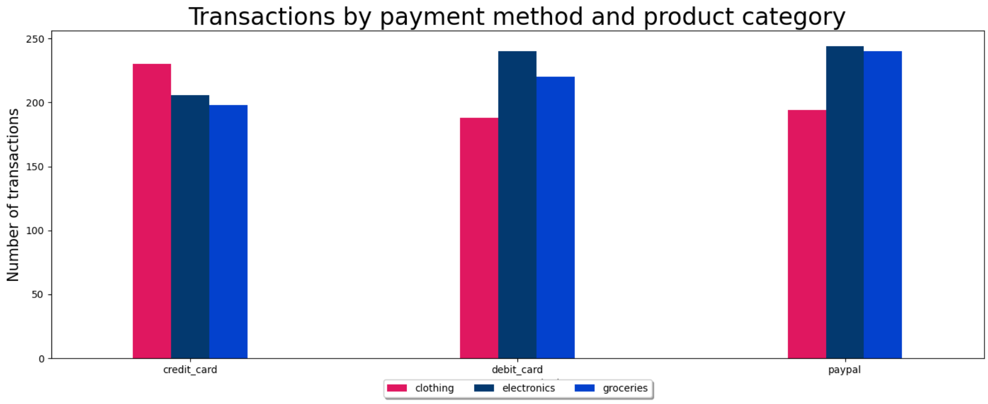

Figure 3: Grouped bar graph for payment method and product category

Time series

Figure 4 is a combined graph (line graph and bar) with two vertical axes.

- Left vertical axis corresponds to the bar graph. It counts the number of transactions made with a certain payment method.
- Right vertical axis corresponds to the line plot and represents the aggregated transactions amount.
- The horizontal axis is shared between both line plot and bar graph, and represents the month in format YYYYMM.

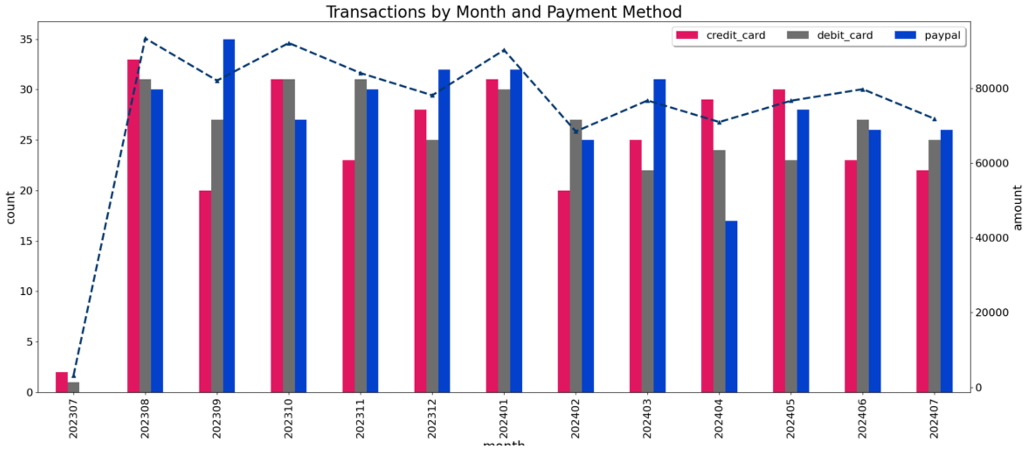

Figure 4: Time series of the transactions by month and the payment method

There are no relevant differences in the number of transactions by payment method along months. The amount does not have peaks either.