Data Scientist II Technical Challenge

Task 1: Exploratory Data Analysis

Cristina Sánchez Maíz | csmaiz@gmail.com | LinkedIn

Summary statistics of the dataset

	transaction_date	amount	customer_age	customer_income
count	980	980	980.0	980
mean	23/1/24 03:11	987.45	43.6	70979.10244
min	31/7/23 00:00	248.79	18.0	20111.77
25%	22/10/23 18:00	733.03	31.0	46297.9725
50%	20/1/24 12:00	977.32	43.0	70464.155
75%	24/4/24 06:00	1250.87	57.0	95933.47
max	29/7/24 00:00	1679.68	69.0	119941.3
std		333.31	15.1	28854.58212

- **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 €1679.68.
  - The distribution is slightly skewed to the right, given the difference between the mean and median.
- **Customer Age:**
  - Average customer age is 43.6 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.

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 each numerical variable.

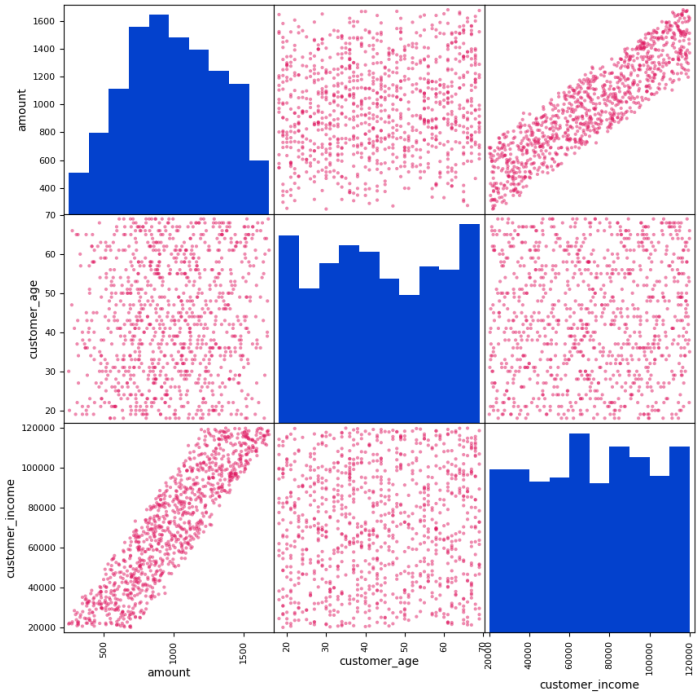

Figure 1: Histogram and correlation plot for numerical variables

Histograms:

- The distribution of *amount* is approximately normal, with a mean of ~1000 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.

	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

Figure 2 show the boxplots of the numerical variables. The median line is centered within the rectangle and the length of the upper and lower whiskers are pretty similar indicating the distribution is not skewed to a certain value (as we already know according to the histograms). There is no outliers since all point lies between the minimum and the maximum.

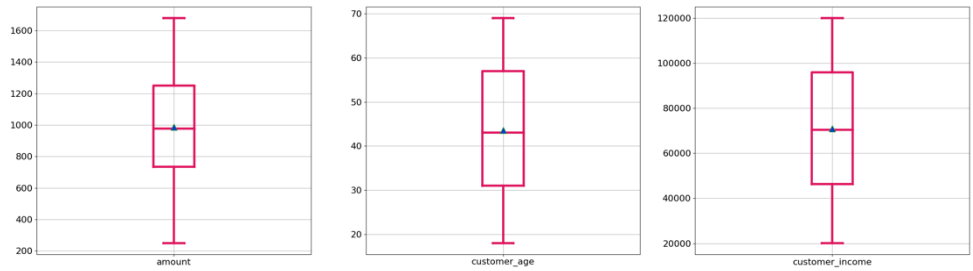

Figure 2: Boxplot of numerical variables

Categorical variables

Imagine we want to know if people prefer to puchase 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 a group are not different heights, this indicates that there are not differences between the groups for a payment method.

The heights of the bars across groups are pretty similar, indicating there are not significant differences between the groups for that category.

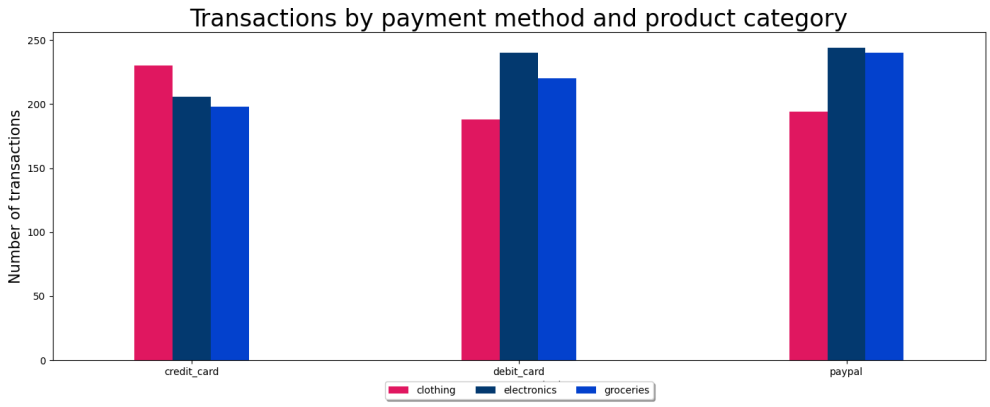

Figure 3: Grouped bar graph for payment method and produc category

Time series

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

- Left axis is for the bar graph. It counts the transactions made with a certain payment method
- Right axis is for the line plot and represents the amount of transactions
-

The horizontal axis is shared and represent the month in format YYYYMM

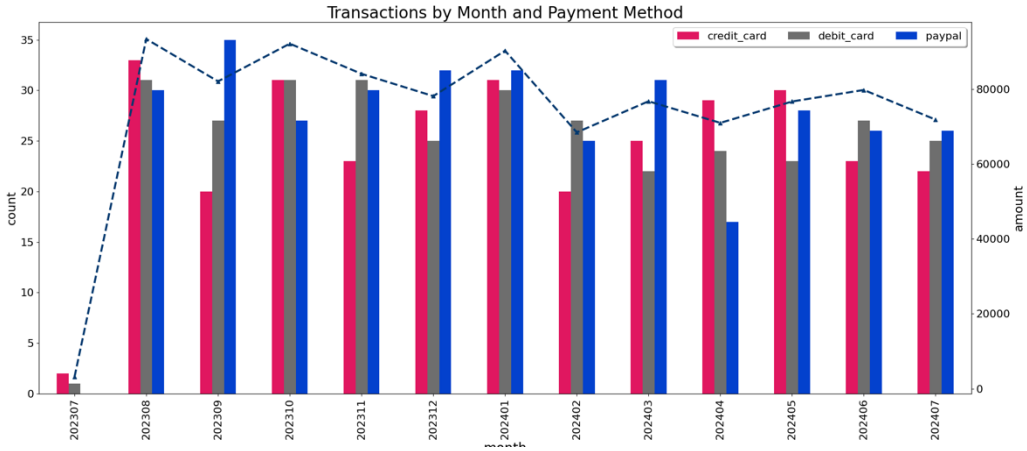

Figure 4: Time series of the transactions by month and the payment method

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