SQL PROJECT VISUALIAZATION REPORT

Goals

Understanding customers is crucial for creating effective marketing strategies and targeting high-potential customers within limited resources. This SQL visualization project aims to understand customers based on their demographics and purchasing behaviors including the amount spent on products and purchasing channels used to identify who is mostly likely to respond to a campaign and what characteristics those customers possess. This information helps companies identify whom to target for implementing a marketing campaign effectively while utilizing limited resources properly.

Dataset Review

The customers are distinguished by customer ID and each ID number is not repeated. The dataset consists of 5 distinct components. First, it includes customers' demographics, such as educational level, income and the number of children. Specifically, there are separate columns for 'Kids at home' and 'Teenagers at home'. I decided to combine them into a single categorical variable by assigning '1' if customers said 'yes' in either column and '0' if they answered 'no' in both columns to avoid overlaps that could occur when customers check both columns, even with only one kid or teenager at home, as no age range is specified in the dataset. Second, there are columns indicating the amount spent on products such as meat, fish, wines and so on. These products are primarily necessities but wines and gold products are categorized as luxury items. Third, there are categorical variables related to marketing campaign acceptance by customers encoded by '1' (Response Yes) and '0' (No response). There are 6 variables and they indicate the customer acceptance for campaigns from the 1st to the 6th (the last) campaign. Fourth, the dataset lists the purchasing channels chosen by customers. The number of purchases through different purchasing channels are specified. Lastly, there are some variables related to this company such as customer enrollment date, the number of days since their last purchase and whether customers complained within the last 2 years.

Explaining Results regarding Business questions

I predefined 8 business questions solved by SQL queries using the dataset.

(1) What is the distribution of customers across different income ranges, and how does it correlate with their education level?

[Visualization Results: **here**]

The education variable is a categorical variable with 5 educational levels such as basic, 2nd cycle, graduation, master and Ph.D. As shown in the results, customers with a basic education level are predominantly in the low-income level. Particularly, customers in the upper-middle income and high-income group are absent from the basic education category. Most customers are classified at the graduation level as shown on the bar chart. Among these, 50% of the customers belong to upper-middle income and high-income groups, possessing the highest numbers in both income levels compared to other education levels. Additionally, the proportions of upper-middle income and high-income groups increase as education level progress from basic and 2nd cycle to Graduation, Master and Ph.D. In conclusion, the correlation between education level and income level exists. When customers have higher education levels, they are more likely to belong to higher income groups. In contrast, when customers have lower education level, they are more likely to fall into lower income groups.

(2) What is the average recency of activity for customers grouped by marital status?

[Visualization Results: **here**]

The marital_status variable is a categorical variable with 8 values. Some customers did not expose their status which is represented as 'Unknown'. Recency means the number of days since a customer's last purchase. This business question investigates whether customers' marital status affect their shopping frequency. If the recency is low, this indicates customers shop more frequently compared to those with higher recency. As shown in the bar chart, customers who are classified as 'Alone' and 'YOLO' have the lowest recency, indicating that they shop more frequently compared to other marital statuses. Particularly, customers who are classified as 'YOLO' exhibit an extremely low recency, shopping approximately every 3 days on average. On the other hand, the average recency for other customer groups ranges between 30 and 50 days. Interestingly, I expected customers with families to shop more frequently than customers living alone. However, customers with partners categorized as 'Together' or 'Married' tend to have longer recency periods, averaging over one month between purchases.

(3) What percentage of customers who complained have higher spending on luxury goods like wines and gold products? [Visualization Results: **here**]

This business question investigates whether customers who complained to this company within the last 2 years exhibit higher spending on luxury goods. I initially hypothesized that there could be a correlation between complaints and higher spending on luxury items because customers purchasing luxury items usually have higher expectations which could increase the likelihood of customers' complaints. I extracted data for customers who complained within the last 2 years and analyzed their total amount of spending on luxury items for each customer. The criteria dividing spending into lower or higher spending is based on the average amount spent on luxury items of non-complaint customers. If their spending is over the average spending of non-complaint customers, they are classified as higher spenders on luxury items and vice versus. Contrary to my expectation, only 23.81% of customers who purchased luxury items complained within the last 2 years, indicating there isn't any strong correlation between complaints and spending on luxury goods.

(4) Does family size affect the likelihood of responding to campaigns? [Visualization Results: **here**]

This business question investigates whether there is a clear correlation between family size and campaign response. To analyze this, I used the marital_status to distinguish family sizes into three groups such as 1, 2, and 3 by combining marital_status with the number of kids in their household. In addition, I created a new 'TotalResponses' variable to sum all campaign acceptance variables and calculate the average responses for each family size group. As shown in the bar chart, when the number of family size increases, the average response to campaigns decreases. The number of family size is inversely proportional to campaign response rates. Customers living alone without kids are more likely to respond campaigns, while larger families are less engaged with marketing campaigns.

(5) Which campaigns are most successful? [Visualization Results: **here**]

This business question investigates which campaign attracted more customer responses among 6 campaign related variables. As shown in the bar chart, the last campaign has the highest response rate compared to all other campaigns. In contrast, the first and second campaigns exhibit the lowest campaign rates, indicating that customers are less likely to respond to earlier marketing campaigns. Their responses gradually increase over campaigns and peak at the last campaign. We cannot guess why this exactly happened but the later marketing campaigns are more effective that the earlier ones.

(6) What is the total spending by customers across all campaigns, and how does it compare by the first campaign acceptance? [Visualization Results: **here**]

This business question investigates which campaign was the most effective based on their total spending for products. I created a new Response variable assigning customers to the campaign which they responded for the first time. For example, if customers responded to the last campaign for the first time, they belonged to the last campaign. As shown in the results, the last campaign achieved the highest total spending on products compared to all other campaigns, followed by the first campaign. This suggests that the last campaign effectively attracted more customer responses and successfully converted them into purchases. Interestingly, the first campaign did not attract a large number of customer responses, many customers purchased products, resulting in high total spending. In contrast, the second campaign exhibited the poorest performance among all other campaigns, which attracted the lowest number of customer responses and conversions.

(7) Which month had the highest number of new customer registrations and how were the registration patterns? [Visualization Results: **here**]

This business question investigates whether there is a recognizable customer registration pattern to assess. If there is a month where customer registration sharply dropped, it should be reviewed internally and identify potential reasons to address this issue. As shown on the line graph, the customer registration dropped sharply between May and July for 2 months and then it went up in Aug, showing the highest number of new customer registrations. The registration patterns exhibit similar patterns during the entire months even with minor fluctuations except for significant declines in both June and July which should be reviewed internally to identify what exactly dropped the customer registrations.

(8) How consistent is the spending behavior of customers across different product categories over the last two years? [Visualization Results: **here**]

This business question investigates whether customers with distinct average spending within the last 2 years on products exhibit different coefficient of variation in their spending patterns. I calculated each customer's average spending on products and their corresponding CV values. The Coefficient of Variation (CV) indicates the consistency of customers' spending patterns. If customers have higher CVs, it indicates they primarily focus on purchasing specific product categories. In contrast, if customers have lower CVs, it indicates they balanced their purchases and are more likely to evenly purchase diverse products. As shown on the graph, X-axis represents the average spending on products and Y-axis represents their CV values. When the average spending is lower, CVs are dispersed whereas the average spending is higher, CVs are more centered and their values become lower, indicating that customer spending more on average tend to evenly purchase diverse products.