Vedant Shah

Winter 2021 Data Science Intern Challenge

Question 1: Given some sample data, write a program to answer the following: [click here to access the required data set](https://docs.google.com/spreadsheets/d/16i38oonuX1y1g7C_UAmiK9GkY7cS-64DfiDMNiR41LM/edit#gid=0)

On Shopify, we have exactly 100 sneaker shops, and each of these shops sells only one model of shoe. We want to do some analysis of the average order value (AOV). When we look at orders data over a 30 day window, we naively calculate an AOV of $3145.13. Given that we know these shops are selling sneakers, a relatively affordable item, something seems wrong with our analysis.

1. Think about what could be going wrong with our calculation. Think about a better way to evaluate this data.

One factor that could be causing this wrong calculation are outliers – products purchased per order. For example, both order\_id 16 and 61 had an order amount of 704000, which in total is 1480000 – these orders are gigantic compared to the majority of order\_amounts which range in the hundreds. Outliers such as these affect the AOV, inflating its value. Another factor is that of the average selling price. Since the average selling price is dictated by the merchant, some merchants can charge higher values which can play a crucial role for the AOV to be $3145.13.

A better way to evaluate this data could be determining the AOV over a 12-month period, which provides a distributed value as there are months where shoe sales flourish and months where they do not; hence, this would be balance out the sales and would provide a more realistic AOV as compared to the one above.

Another way to evaluate this data can be accomplished by dividing the 100 companies into groups based on their revenue history such as Low, Medium, and High and calculate the AOV for each group. This ensures a minimum number (almost zero) of outliers to produce a much more realistic value.

A metric that could be utilized in correlation with AOV is average basket value. This metric determines how many items are sold per transaction. Over the period of a month, you can calculate the average basket value and multiply that value by the price of each product to determine the AOV for each business. Then, the sum of all these values divided by 100 (number of businesses) will provide Shopify a stronger representation of the AOV.

Lastly, the customer lifetime value metric can also be used in tandem with AOV. CLV tells Shopify how much each business owner is valued throughout their relationship with Shopify; this can be calculated over any period and a month-to-month basis might be quite helpful when making decisions. Combining these metrics tells you if businesses are making more or less on subsequent months, which in return provides a value more realistic than $3145.13.

1. What metric would you report for this dataset?

For this dataset, I would report the CLV metric in correlation with the AOV for each of the 100 shoe merchants of Shopify. As stated above, this provides Shopify with a much more accurate representation of each company, which is better than determining the AOV of all 100 companies because the two factors mentioned above, products purchased per order and average selling price, are most likely what caused the poor calculation of the AOV. With this method, a CLV for each company will be generated, helping Shopify make decisions in the future.

1. What is its value?

The formula to calculate this would be as follows:

1. Calculate average purchase value:
2. Calculate average purchase frequency rate:
3. Calculate customer value:
4. Calculate average customer’s lifetime span:
5. Calculate customer lifetime value:

For example, for shop 53 (up to entry 1600), the CLV for a one-month period would be:

With this type of calculation done for every merchant over a certain period, the CLV will ensure that there are no wrong calculations and will provide Shopify with a well-rounded representation of each company’s product and revenue.

Question 2: For this question you’ll need to use SQL. [Follow this link](https://www.w3schools.com/SQL/TRYSQL.ASP?FILENAME=TRYSQL_SELECT_ALL) to access the data set required for the challenge. Please use queries to answer the following questions. Paste your queries along with your final numerical answers below.

1. How many orders were shipped by Speedy Express in total?

SELECT ShipperID

FROM Shippers

WHERE ShipperName = 'Speedy Express';

The above returns 1.

SELECT

COUNT (\*)

FROM [Orders]

WHERE ShipperID = 1;

Final Answer: 54

1. What is the last name of the employee with the most orders?

SELECT EmployeeID

FROM [Orders]

GROUP BY EmployeeID

ORDER BY COUNT (\*) DESC

LIMIT 1;

The above returns 4.

SELECT LastName

FROM [Employees]

Where EmployeeID = 4;

Answer: Peacock

1. What product was ordered the most by customers in Germany?

I am unsure as to how to answer this question. However, my thinking process to solve this question is as follows:

1. Determine which customers are from Germany – SELECT \* FROM [Customers] WHERE Country = ‘Germany’.
2. From the output above, I would determine which CustomerIDs from the outputted table match the CustomerIDs from the Orders table and output the OrderIDs from the matches.
3. I would compare these OrderIDs with those from the OrderDetails table and output the ProductIDs from the matches.
4. Finally, I would determine the ProductID that has the highest frequency and output the name of that product with the following command: SELECT ProductName FROM [Products] Where ProductID = productID\_Outputted;. productID\_Outputted is the output from step 3.