Winter 2021 Data Science Intern Challenge

Brian Jairam

August 26, 2020

1 Average Order Value of Sneakers

1.1 A Brief Discussion On What It Means to Be Wrong

Before jumping into an analysis what went wrong, it would be helpful define what *wrong* means.

Let us examine the dataset first. Assuming that the data in the table is free of any fraudulent transactions, the data itself cannot be wrong. It is after all, a concrete record of what really did happen, whether we like it or not. So, any *wrongness* we find with a metric must stem from whether it addresses the question that we wanted it to address.

So what is the question we are trying to answer with the AOV metric? Note that this challenge does not really specify that. So, perhaps we are jumping to the conclusion that this metric is wrong.

If we want to know what was the expected order value of one additional order in the month of March, assuming the orders placed were a good enough sample of all the total orders that could have been placed for sneakers in March, I argue that the AOV being $3145.13 does not look wrong at all. it is perhaps easier to see this with the help of a few additional metrics.

We can also calculate the mean of the number of shoes sold per order, which is 8.8. If we divide the previously calculated AOV by 8.8, we get a number around $357. We can regard this number as an average of the price of each individual pair of shoes for all transactions. Together with the AOV, these

metrics constitute the expected values of the number, the price, and total value of an additional order in the month of March, conditional on the assumption that all previous orders give a good indication of the nature of all transactions that can happen.

Here is a concrete use case: let us also suppose another data science team has found that if we had spent $\$X$ in March on advertising, we could have brought in 100 more orders for sneakers. We could use the metrics we found above to give a prediction that this effort would have created \$314513 of order value for merchants. We can evaluate whether advertising would have been a good idea by comparing X to the percent of that order value that Shopify earns as revenue.

1.2 Why Does The AOV Look Wrong?

If the AOV looks wrong, either the data is not clean, or the AOV we calculated does not answer whatever question we want to answer.

If the data is not known to be clean, then the AOV metric actually aids us considerably. Taking a look at the table, we note that out of the 5000 orders, 4937 of them have order values less than the AOV. So it would most likely be beneficial to examine the 63 orders that had an order value above the previously calculated AOV.

There are 17 orders originating from the same shop that have order values of \$704000, with 2000 pairs of shoes sold. If we divide the order value by the number of pairs of shoes sold for each of these orders, we get the unit cost of a single pair of shoes, \$352. These orders could be considered suspicious due to the large quantity of shoes ordered and the fact that all of these orders were purchased on credit at precisely the time 4:00 on different days. However, the unit cost of each shoe looks somewhat reasonable. The shop, which has an ID of 42, has sold sneakers at this unit cost before, but with much lower quantities (see orders 1365, 1368, and 1472).

The 46 other orders that had an order value above the AOV were all from one shop, with ID 78. In fact, all of this shops orders were of this nature. All of the unit cost of pairs of sold in these orders were the same, a whopping \$25725 a pair. While this does look unreasonable, perhaps this is Kanye selling autographed shoes...

In any case, if the data is not clean, it would probably be a good idea to investigate the orders that look suspicious.

1.3 Using The Median Instead of The Mean

But what if we cannot clean the data? Then we can use the median, as it is would not be affected by the small amount of these fraudulent transactions.

However, the median can also be used to answer some questions that just using the AOV could not answer (in cases where the AOV looks "wrong").

For example, suppose we want to find an average order value of an *average* buyer or shop. The 67 orders with high order values may not necessarily fit this description. We can may even be inclined to give them a certain name: outliers.

In cases we want to create that metric that characterizes the majority of orders that do occur, the median is great. The median of the order values is $284. The first and third quartiles are $163 and $390 respectively. In fact, over 95% of orders have an order value less than $1000. In terms of characterizing an everyday Joe's order of sneakers, the previously calculated AOV of $3145.13 now just appears absurd. But note that nothing has changed except the questions we are trying to answer.

So what is the practical use case for this new median AOV? Well, suppose Shopify is looking at how to improve customer and merchant experience in orders for sneakers. In this case, an AOV that truly represents how the values of the majority of orders is valuable, so that too much focus is not placed on a small number of orders that generate very high order values. For example, if a benefit is given to consumers or merchants based on the value of orders, setting up based on the AOV calculated with the mean as opposed to the new AOV calculated with the mean would yield very different results in terms of who wins and who loses.

Ideally, we should present this median together with the 1st and 3rd quartile values to both give some context, and to systematically remove outliers (defined as any data point not within 1.5x the inter-quartile range of the median). It would be very nice if we could have the median presented alongside the mean and standard deviation as well, as that would give us a clearer picture of the shape of the distribution.

1.4 Conclusion

In conclusion, here are the three main points of the challenge summarized

a. What is wrong with the calculation depends on the questions you aim to answer with the AOV metric. In some cases, the metric may be right in the sense that it does answer the question you want answered. However, in the particular case of characterizing what orders look like in the majority of cases, there is an issue of 67 orders with very high order values influencing the mean of the data set.

b. As the mean is not affected by outliers (the 67 orders), a good alternative metric would be the median of the order values. However, in an ideal case, the median should be presented along with the other quartiles of the data, and perhaps even alongside the mean together with a standard deviation to grasp what the data is telling us fully.

c. The median order value is $284.

I'd like to thank the reader for reading through my response to the first part of this challenge. I had a lot of fun thinking about this problem. I may be making more changes to this document through the month of August as I find new typos to fix...

I would love to discuss my ideas more in a personal setting... perhaps even in an interview?

Again, thanks to Shopify for fielding applications for this Data Science Internship. I look forward to hearing back from you all!

- Brian

P.S. The second part of the challenge is in section 2 below.

2 SQL Queries

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

```
SELECT
    COUNT(OrderID)
FROM
    Orders
JOIN
    Shippers
        ON Orders.ShipperID = Shippers.ShipperID
WHERE
    Shippers.ShipperName = "Speedy Express"
```

Result: 54

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

```
12  WITH temp AS (
13      SELECT
14          EmployeeID,
15          COUNT(OrderID)
16      FROM
17          Orders
18      GROUP BY
19          EmployeeID
20      ORDER BY
21          COUNT(OrderID) DESC
22      LIMIT 1
23  )
24
25  SELECT
26      Employees.LastName
27  FROM temp
28      JOIN Employees
29          ON temp.EmployeeID = Employees.EmployeeID
30
```

Result: Peacock

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

This question is a little ambiguous. We can either report the product that was in the most orders from customers from Germany, or the product which had the highest unit sales to customers in Germany.

We can solve both.

First, the product that was in the most orders from German customers:

```sql
WITH temp AS (
    SELECT
        u.ProductID,
        u.ProductName,
        COUNT(u.OrderID)
    FROM (
        SELECT s.OrderID, Products.ProductID, Products.ProductName
            FROM OrderDetails
                JOIN Products
                    ON OrderDetails.ProductID = Products.ProductID
                JOIN (
                    SELECT Orders.OrderID
                        FROM Orders
                            JOIN Customers
                                ON Orders.CustomerID = Customers.CustomerID
                    WHERE Customers.Country = 'Germany'
                ) s
                    ON OrderDetails.OrderID= s.OrderID
    ) u
    GROUP BY
        u.ProductID,
        u.ProductName
    ORDER BY
        COUNT(u.OrderID) DESC
    LIMIT 1
)

SELECT temp.ProductName FROM temp
```

Result: Gorgonzola Telino

Next, the product that had the highest unit sales to customers in Germany:

```sql
WITH temp AS (
    SELECT
        u.ProductID,
        u.ProductName,
        SUM(u.Quantity)
    FROM (
        SELECT Products.ProductID, Products.ProductName, OrderDetails.Quantity
            FROM OrderDetails
                JOIN Products
                    ON OrderDetails.ProductID = Products.ProductID
                JOIN (
                    SELECT Orders.OrderID
                        FROM Orders
                            JOIN Customers
                                ON Orders.CustomerID = Customers.CustomerID
                    WHERE Customers.Country = 'Germany'
                ) s
                    ON OrderDetails.OrderID= s.OrderID
    ) u
    GROUP BY
        u.ProductID,
        u.ProductName
    ORDER BY
        SUM(u.Quantity) DESC
    LIMIT 1
)

SELECT temp.ProductName FROM temp
```

Result: Boston Crab Meat