Shopify Data Science Internship Application

Anja Wu

Fall 2022

Part 2

To showcase your skills and to help us get a better understanding of your data science knowledge, please submit the mandatory Data Science Intern Challenge and attach it to the field titled "Shopify Technical Challenge Submission (link)". Please ensure that the challenge is complete before you submit your application, as recruiters will be reviewing applications on a rolling basis. Challenge submission is mandatory in order to be considered:

Question 1: *Given some sample data, write a program to answer the following: click here to access the required data set*

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.

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

I confirmed that he AOV was calculated as sum of order_amount divided by sum of total_items. However, the reason the AOV is so high ($3,145.13) is because there are outliers for both total_items and order_amount. Calculating mean with outliers heavily influences the average.

As can be seen in the Jupyter notebook, for total_items any item amount over 6 is considered an outlier. We can see that there are 17 orders in which the total_items was 2,000, there was also one order that had 8 total_items. The one order with 8 items will not affect the data much, as it is very close to the other total_items, so it was kept in the data, while the orders with total_items of 2000 was removed.

For the order_amount, when we look at the distribution we can see that there is a large jump from just under $2,000 to above $25,000 (this only includes total_items mostly for 1-3). This could be some sort of error in inputting as $25,725 for 1 pair of shoes is strange. If I were working on this and had access to the pipeline or could speak to the process of collecting data, I would dig into the mentioned outliers to see if they are legitimate or if they were mistakes made. Assuming that these points were errors, I removed dollar amounts above $25,000 - this was 46 more points (out of the remaining 4983), leaving 4937 orders.

Now using the three differently adjusted datasets, we can analyze between: mean, median and mode for the AOV.

b) What metric would you report for this dataset?

I would use the adjusted mean versus unadjusted mean, median and mode. The considerations and rationale of adjusted mean is written below.

The metrics to choose from are mean, median, and mode for AOV. Traditionally, AOV uses the arithmetic mean (https://www.bigcommerce.com/ecommerce-answers/what-average-order-value/). Through a source (https://www.shopify.ca/blog/average-order-value) I have found it is important to consider all 3 metrics when looking at improving e-commerce conversions. The mean was calculated with 3 different datasets; first looked at no data points removed (mean of $3145.13), second remove the 17 cases of total_items in order being 2,000 (mean of $754.09), and third on top of the total orders of 2000 there were also orders which had only several items that were valued at over 25,000. This last case seemed like some error, so these were removed. In this last case, we can see the adjusted mean (when outliers are removed both for high total_items and high order_amount) is: $302.58, whereas the median is $284.0 and the mode is $153.

All of these give different facts about what it means for the shoe sales. The mean will give you the classic average of sales, but it is more easily swayed by extremes. The median will give you the middle value of orders, which is better for skewed data. As two of the metrics are close together it would be reasonable to take the classic AOV definition of $302.58. One good point of several articles read was that if the point of the AOV is to figure out a price point for free shipping it would be a good idea to go with the mode, for a company on its own, to increase sales for the most frequent order value. However, since this is not about any one specific store but rather 100 shoes stores run through Shopify, I would use the adjusted mean for the calculation of AOV.

c) What is its value?

Given the fact that we are looking at multiple shoe stores (not just one), the mean and median are close, and the final dataset removed the data that seems (with the information given) unreasonable: the adjusted mean of $302.58 should be used for AOV.

Question 2: *For this question you'll need to use SQL. Follow this link 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.*

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

Query:

```
SELECT
  ord.ShipperID,
  ship.ShipperName,
  COUNT(ship.ShipperName) AS 'total_orders'
FROM
  Orders ord
  JOIN Shippers ship ON ship.ShipperID = ord.ShipperID
WHERE
  ShipperName = 'Speedy Express'
GROUP BY
  1,  2
```

Numerical answer:

Total orders by Speedy Express: 54

ShipperID	ShipperName	total_orders
1	Speedy Express	54

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

Query:

```
SELECT
  o.EmployeeID,
  e.LastName,
  COUNT(o.EmployeeID) AS total_orders
FROM
  Orders o
  JOIN Employees e ON o.EmployeeID = e.EmployeeID
GROUP BY
  1, 2
ORDER BY
  3 DESC
LIMIT 1
```

Numerical answer:
Last name of employee with most orders: Peacock (40 orders)

EmployeeID	LastName	total_orders
4	Peacock	40

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

Query:

```
SELECT
  c.Country,
  od.ProductID,
  p.ProductName,
  SUM(od.Quantity) AS total
FROM
  Customers c
  JOIN Orders o ON c.CustomerID = o.CustomerID
  JOIN OrderDetails od ON o.OrderID = od.OrderID
  JOIN Products p ON od.ProductID = p.ProductID
WHERE
  Country = 'Germany'
GROUP BY
        1,2,3
ORDER BY
        4 DESC
LIMIT 1
```

Numerical answer:
The product ordered most by customers in Germany: “Boston Crab Meat” (ProductID: 40, ordered 160 times)

Country	ProductID	ProductName	total
Germany	40	Boston Crab Meat	160