Shopify Summer 2022 Data Science Intern Challenge

Question 1: 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.
b.) What metric would you report for this dataset?
c.) What is its value?

Answer 1:

a.) As a sneakerhead myself (I'm very happy to see a question about sneakers), the easy answer is that all 100 sneaker shops are selling very rare sneakers, for example the Off-White MCA University Blue Nike Air Force 1 Low (https://stockx.com/nike-air-force-1-low-off-white-university-blue) which, as of writing this, last sold for $4400 in my size. However, assuming that these shops are not all resale market participants, the problem is that the current AOV estimation, the mean is this case, is being impacted by several significant outliers.

Looking deeper into the data we find that customer purchases range between $90 and $704,000. I will admit that my sneaker collection is a bit obsessive, with almost 70 pairs, but I can definitely assure you that it is not $704,000 obsessive. Now that we've identified one of our outliers it may be wise to look into this sneaker purchaser further. Doing so we find that this user (#607) actually made 17 individual purchases in this 30 day period and each purchase was 2000 pairs of sneakers costing them $704,000 each time.

Ultimately user 607 spent $11,968,000 by purchasing 34,000 pairs of sneakers. This user accounted for approximately 76% of all dollars spent, while purchasing approximately 77% of all shoes. These numbers are astounding considering that this user's 17 purchases account for a measly 0.34% of the total purchases made over this 30 day period.

There are another additional 54 orders that result in customers spending more than $1000. Whether this is an optimal point to start counting outliers is debatable and would be getting into to the nitty gritty of sneaker culture. For example, if a shop is exclusively selling white Nike Air Force 1's you would likely find numerous orders of large quantities since this sneaker is a staple of sneaker culture and some sneakerheads will only wear the shoe as long as it is pristine (since it's a white shoe that's not a very long time, so you need many pairs).

Of our 5000 orders only just over 1% I would consider outliers, with this 1% accounting for about 91% of dollars spent and 78% of sneakers purchased. It appears that our outliers can be classified as two types of purchasers. **1.** Those purchasing high quantity and **2.** Those purchasing high quality. For example, user 607 made individual purchases of 2000 pairs of sneakers while user 800 purchased a single pair of sneakers for $25,725. User 800 also made an additional 18 purchases that cost less than $1000, showing some diversity in their sneaker taste. Numerically speaking it may be fair to assume that a **type 1** case may be a big box retailer purchasing sneakers to sell in their store, while the **type 2** case may be an individual looking to fill out their sneaker collection.

Now that we've determined the root cause of our large AOV, we can now attempt to minimize the effect that our outliers are having.

b.) For this dataset I would suggest the use of the median since it is resistant to potential outliers, which we have a plethora of. Holding the number of observations we have constant and adjusting our maximum order amount to, for example, $1,000,000,000 ($1 billion) we would find no change in our current median, however this massive outlier would further bias our mean value upwards. This upwards bias of our mean is what we wish to eliminate by using an outlier resistant measure like the median (I was unsure whether we were expected to develop some sort of novel measure, but I just assumed this was a data exploration/analysis type of question)

c.) Using the median, we find our AOV to be $284, much more reasonable than our original $3145.13.

Question 2: Using SQL query the given dataset to answer the following questions. Provide your queries and final numerical answers.

a.) How many orders were shipped by Speedy Express in total?
b.) What is the last name of the employee with the most orders?
c.) What product was ordered the most by customers in Germany?

Answer 2:

a.) We find that Speedy Express shipped a total of **54** orders using the following SQL query.

```
SELECT ShipperName, COUNT(*) AS SpeedyExpressOrders FROM Orders
LEFT JOIN Shippers
ON Shippers.ShipperID = Orders.ShipperID
GROUP BY ShipperName
HAVING ShipperName = "Speedy Express";
```

b.) We find that the employee with the last name **Peacock** has the most orders using the following SQL query.

```
SELECT ord.EmployeeID, emp.LastName, COUNT(ord.EmployeeID) AS EmployeeOrders
FROM Orders AS ord
JOIN Employees AS emp
ON ord.EmployeeID = emp.EmployeeID
GROUP BY ord.EmployeeID
ORDER BY EmployeeOrders DESC
LIMIT 1;
```

c.) The product ordered the most by customers in Germany is **Boston Crab Meat**, with a total quantity of 256. We can find this result using the following query.

```
SELECT cust.Country, prod.ProductName, SUM(dets.Quantity) AS SumQuantity
FROM OrderDetails AS dets
JOIN Products AS prod ON dets.ProductID = prod.ProductID
JOIN Orders AS ord ON ord.OrderID = dets.OrderID
JOIN Customers AS cust ON cust.CustomerID = ord.CustomerID
GROUP BY dets.ProductID
HAVING cust.Country = “Germany”
ORDER BY SumQuantity DESC
LIMIT 1;
```