Summer 2022 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

> 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.

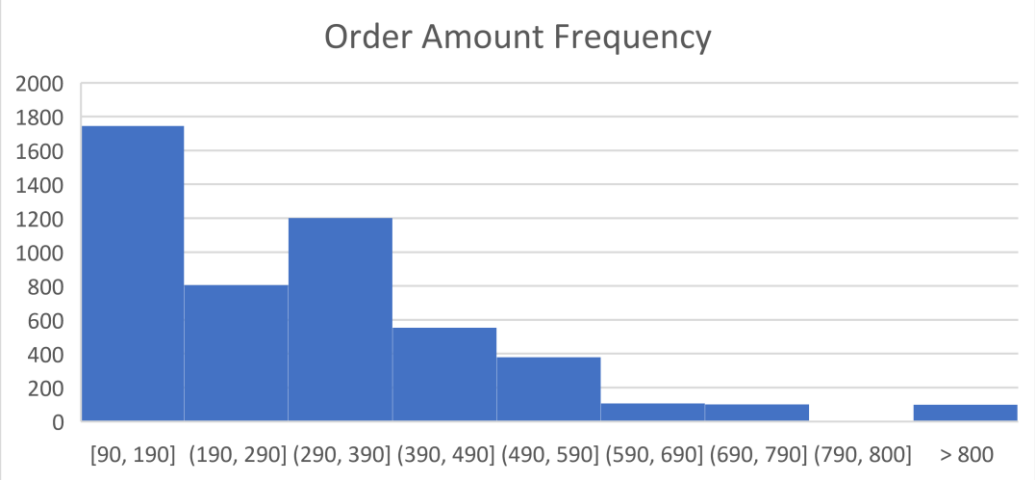
a. Think about what could be going wrong with our calculation. Think about a better way to evaluate this data.

From looking at the data, there are some outliers within the dataset: there's large orders in terms of both size and cost that are skewing the average. This makes sense as

1. some shoes do retail for thousands of dollars, and
2. large orders cost a lot of money.

Given this information, AOV is being calculated accurately; however, we may want to change the metric or our expectations. The high-value purchases might represent a subset of purchasers, such as wholesalers, that we do want to calculate metrics for. It might be better to split the data into two datasets based on types of customers, and calculate averages for both separately.

If we don't want to split the dataset, we might instead calculate the median. By looking at the histogram below, we can see most purchases are under $500:



Therefore, as the large values only represent a very small percentage of the dataset, using the median might better describe typical behavior.

b. What metric would you report for this dataset?

It really depends on what we're trying to accomplish here. If we're trying to describe normal behavior of small purchasers, then I would take the median. If we are more interested in the average, then AOV as calculated previously is perfectly fine. I think it would be worthwhile to differentiate between normal buyers and wholesalers, in which case we could split the dataset and calculate appropriate metrics respectively.

c. What is its value?

Personally, I would choose the median to measure typical buyer behavior, which in this case is $284. This accounts on average for a purchase of 2 pairs of shoes.

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?

Speedy Express shipped 54 orders in total.

Query:

```
SELECT s.ShipperName,
        Count(o.OrderID) as NumOrders
FROM Orders o
JOIN Shippers s
        ON o.ShipperID = s.ShipperID
Where s.ShipperName is 'Speedy Express'
Group by s.ShipperName
```

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

The last name of the employee with the most orders is Peacock, who had 40 orders.

Query:

```
SELECT LastName,
        count(o.OrderID) as NumOrders
FROM Employees e
JOIN Orders o
        on e.EmployeeID = o.EmployeeID
group by LastName
order by 2 desc
limit 1
```

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

More customers in Germany ordered Gorgonzola Telino than any other product.

Query:

```
SELECT p.ProductName, Count(*)
FROM OrderDetails od
JOIN Orders o
        ON o.OrderID = od.OrderID
JOIN Customers c
        ON c.CustomerID = o.CustomerID
JOIN Products p
        ON p.ProductID = od.ProductID
WHERE c.Country = "Germany"
Group by p.ProductName
Order by Count(*) desc
limit 1
```