SAMUEL GARCIA

SamuelCGarcia.3@gmail.com
linkedin.com/in/samuelgarcia3
github.com/samgarcia3

Sunday May 9th, 2021

FALL 2021 DATA SCIENCE INTERN
TECHINCAL CHALLENGE SUBMISSION

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 $3,145.13. 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.

To begin I would like to clearly define the AOV with a statistical metric. The AOV is the Mean (*Total Sum ($) of Order Values / Total Count of Order Values*) of the order values, which is a result of $3,145.13.

At first glance we can see that some of the order values within the dataset are $704,000, this is well above the other order values (~$300). A dataset that contains outliers such as these can heavily skew the Mean of the of the overall order value, resulting in numbers that are unrepresentative of the actual dataset and or intended value.

I can think of a few ways to deal with the outliers – Remove outliers all together and or create a new set of columns for separate AOV ranges (i.e., $0-$1,000, $1,000 – $10,000, …), calculate the Median, and lastly calculate a Trimmed/Truncated Mean.

2. What metric would you report for this dataset?

For this problem specifically I would use the Median as a better metric. The Median is the quickest and easiest solution for an outlier problem such as this one. And in comparison, a dataset that has a Trimmed Mean of 10% (*$287*) and 25% (*$276*), the Median is a good balance between both.

3. What is its value?

Using the Median value for the AOV would give a new result of $284 for an average order value.

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.**

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

```
SELECT COUNT(*)
FROM [Orders]
WHERE ShipperID = 1;
```

Total = 54

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

```
SELECT employees.lastname,
       employees.firstname,
       employee_order_total
FROM   employees
       JOIN (SELECT employeeid,
                    Count(employeeid) AS employee_order_total
             FROM   orders
             GROUP  BY employeeid
             ORDER  BY Count(employeeid) DESC
             LIMIT  1) AS employee_total
         ON employees.employeeid = employee_total.employeeid;
```

Name = Margaret Peacock
Total Orders = 40

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

```
SELECT p.productname,
       Count(ProductName) as Total,
       c.country
FROM   orderdetails AS od
       INNER JOIN products AS p
               ON od.productid = p.productid
       INNER JOIN orders AS o USING (orderid)
       INNER JOIN customers AS c USING (customerid)
WHERE  c.country = 'Germany'
GROUP  BY p.productid
ORDER  BY Count(ProductName) DESC
LIMIT  1;
```

Product Name = Gorgonzola Telino