

Fall 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.
 - b. What metric would you report for this dataset?
 - c. What is its value?
-
- a. It is most likely that the data contains massive outliers, which skews the mean heavily. Some possible reasons include unique sales of large quantities, incorrectly entered data (or improper conversions), or disparate pricing ranges for high-end products. A better way to evaluate the data would be to assess if there are outliers before blindly using the AOV.
 - b. I would report the MOV (median order value) as medians are robust to large outliers and are easy to compute.
 - c. The value of MOV for this data set is \$204.00. which is much closer to what we might expect.

More rigorous analysis can be found in the Jupyter Notebook in this repository.

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?
- b. What is the last name of the employee with the most orders?
- c. What product was ordered the most by customers in Germany?

- a. A total of 54 orders were shipped by Speedy Express:

```
SELECT COUNT(*) AS SpeedyExpressTotalOrders
FROM [ORDERS]
WHERE ShipperID = (
    SELECT ShipperID
    FROM [Shippers]
    WHERE ShipperName = 'Speedy Express');
```

- b. The last name of the employee with the most orders is Peacock (with 40 total orders).

```
SELECT Employees.LastName AS MostOrderEmployee
FROM [Employees]
JOIN [Orders]
ON Employees.EmployeeID = Orders.EmployeeID
GROUP BY Employees.LastName
ORDER BY COUNT(Employees.LastName)
DESC LIMIT 1;
```

- c. The product most ordered by customers in Germany was Boston Crab Meat if quantity was considered (a quantity of 160 items), or Gorgonzola Telino if order number was the concern (5 orders).

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

```
SELECT ProductName, COUNT(*) AS TotalOrders
FROM Orders AS o
```

```
JOIN Customers AS c
ON o.CustomerID = c.CustomerID
JOIN OrderDetails AS od
ON o.OrderID = od.OrderID
JOIN Products AS p
ON od.ProductID = p.ProductID
WHERE c.Country = 'Germany'
GROUP BY p.ProductID
ORDER BY TotalOrders DESC
LIMIT 1;
```