Summer 2022 Data Science Intern Challenge

Question 1:

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

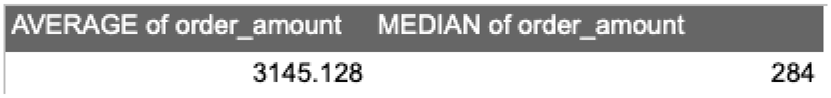
The problem is that this data includes some outliers, and AOV is sensitive to outliers. For example, user 607 always buys 2000 items in one order from shop 42, making the order amount of those orders very high. Also, by calculating unit price = order_amount / total_items, I found that shop 78 is selling a very expensive shoe (unit price = 25725). These orders skew the AOV, which does not represent the whole dataset very well.

To better evaluate the data, we can consider categorizing the orders, and evaluate the orders separately. For example, user 607 might not be an individual customer, because this user always buys 2000 items in one order from shop 42. Also, shop 78 is probably different from other shops, because the shoe it's selling is very expensive. If we filter out these data, AOV can be used.

We can also consider using other metrics to evaluate the data. For example, median order value can better eliminate outliers.

b. What metric would you report for this dataset?
median order value

c. What is its value?
284, calculated by using pivot table



AVERAGE of order_amount	MEDIAN of order_amount
3145.128	284

Question 2:

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

Query:

```
SELECT COUNT(*)
FROM Orders O LEFT JOIN Shippers S ON O.ShipperID = S.ShipperID
WHERE ShipperName = "Speedy Express"
```

Answer: 54

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

Query:

```
SELECT E.EmployeeID, E.LastName, COUNT(DISTINCT OrderID)
FROM Orders O LEFT JOIN Employees E ON O.EmployeeID = E.EmployeeID
GROUP BY E.EmployeeID, E.LastName
ORDER BY COUNT(DISTINCT OrderID) DESC
LIMIT 1
```

Answer:
Peacock

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

Query:

```
SELECT P.ProductName, SUM(Quantity)
FROM OrderDetails OD LEFT JOIN Orders O on OD.OrderID = OD.OrderID
LEFT JOIN Products P on OD.ProductID = P.ProductID
LEFT JOIN Customers C on C.CustomerID = O.CustomerID
WHERE 1=1 AND C.Country = 'Germany'
GROUP BY P.ProductName
ORDER BY SUM(Quantity) DESC
```

Answer:
Gorgonzola Telino