Winter 2022 Data Science Intern Challenge

ANSWERS

Question 1:

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

AOV is calculated by just taking mean of the order_amount on that month that is AOV -$3145.13.

```
count      5000.000000
mean       3145.128000
std       41282.539349
min          90.000000
25%         163.000000
50%         284.000000
75%         390.000000
max      704000.000000
Name: order_amount, dtype: float64
```

From this, we can infer some details

- 50% of the orders are below $284
- 75% of the orders are below $390
- Maximum order amount is $704000
- Standard deviation refers to the dispersion of data. Here, the std deviation is very high (i.e.) $41282 when relatively compared with the mean. So, the data is scattered.

The data is scattered because it has some outliers. These outliers will impact the AOV.

The number of order items are not same for every order. There is an imbalance in the order items. Some records have bulk orders.

So, calculating AOV by also considering the bulk orders will not give correct results.

After some analysis we found that, some of the sneakers are sold for nearly $25000/per item. So, there may be error in the data entry or it may be correct.

- All these high prize sneakers are sold out from the same shop. (i.e.) shop_id - 78
- All the bulk orders are from the same shop (i.e.) shop_id – 42

Shop id - 78. record should not be considered for the AOV calculation. Because, they sold for $25725/sneaker which is a very high value compared to the other shops' rate. So, if we consider them for AOV, then it would give misleading results.

Shop id - 42's sales data shows that, they have taken bulk orders.

Calculating AOV with these outliers gave us the misleading results. So, this could be the thing which drives us to wrong conclusion

b) **What metric would you report for this dataset?**

The metric already used in this calculation is AOV. It is a measure of central tendency of the data. It has some loop holes.

Because AOV will be different for each shop. Some shops sell sneakers for high rate. Some shops sell sneakers for low rate. So, calculating AOV commonly for all the shops will give us misleading results.

We can't rely on a single metric for all the problems. Each and every problem will require different metrics. Each metric will have merits and demerits. It depends on the domain we analyze. We can showcase the importance and drawbacks of such metrics. Finally, Subject matter experts/Domain expertise will decide which metric to be used for the analysis.

Here, I would suggest two ways for the suitable metric.

1) Calculating AOV for each distinct shop on that month

This can be done without removing any outliers.

Particular shops (shop id 42 & 78) which has high sales amount will obviously have higher AOV. Other shops will not be affected by these outlier values.

2) Calculating Median Order Value

c) **What is its value?**

1) Calculating AOV for each distinct shop on that month

	shop_id	total_sales_each_shop	AOV_for_each_shop	median_order_value_for_each_shop
0	1	13588	308.818182	316
1	2	9588	174.327273	188
2	3	14652	305.250000	296
3	4	13184	258.509804	256
4	5	13064	290.311111	284
...	...	...	...	...
95	96	16830	330.000000	306
96	97	15552	324.000000	324
97	98	14231	245.362069	266
98	99	18330	339.444444	390
99	100	8547	213.675000	222

100 rows × 4 columns

Refer **AOV_For_Each_Shop.csv** file for Full answer

2) Calculating Median Order Value with respect to total number of items per order

	total_items	total_sales	count	median_order_value
0	1	763777	1830	153
1	2	1374394	1832	306
2	3	1120803	941	459
3	4	277672	293	592
4	5	58470	77	765
5	6	161460	9	948
6	8	1064	1	1064
7	2000	11968000	17	704000

Refer **Median_Order_Value_for_number_of_items_per_order.csv** file

Question 2:

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

Ans: 54

54 orders were shipped by speedy Express

```
SELECT ShipperName, COUNT(Orders.ShipperID)
FROM Orders
JOIN Shippers ON Orders.ShipperID = Shippers.ShipperID
WHERE Shippers.ShipperName="Speedy Express";
```

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

Ans: Peacock

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

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

Ans: Gorgonzola Telino

```
SELECT Products.ProductName, COUNT(Products.ProductName) AS NumberOfOrders
FROM Orders
JOIN Customers ON Orders.CustomerID=Customers.CustomerID
JOIN OrderDetails ON Orders.OrderID=OrderDetails.OrderID
JOIN Products ON Products.ProductID=OrderDetails.ProductID
WHERE Country = 'Germany'
GROUP BY Products.ProductName
ORDER BY NumberOfOrders DESC
LIMIT 1
```