Fall 2022 Data Science Intern Challenge

Question-1:

Code

```
import pandas as pd

# Reading the data file using pandas library
sneakerData = pd.read_excel('E:/Code-Base/shopify/DS/2019 Winter Data Science Intern Challenge Data Set.xlsx')

# Visual Representation of the data
sneakerData.head()

# Creating a data frame that groups the orders by shops
revenue_per_shop = sneakerData.groupby(['shop_id'])['order_amount'].agg('sum')
items_per_shop = sneakerData.groupby(['shop_id'])['total_items'].agg('sum')

revenue_per_shop.name = 'Total_Revenue'
items_per_shop.name = 'Total_Items'

sneakerDataByShop = pd.concat([revenue_per_shop, items_per_shop], axis = 1)

#  Calculate AOV of all shops
sneakerDataByShop['AOV'] = sneakerDataByShop['Total_Revenue']/sneakerDataByShop['Total_Items']
sneakerDataByShop.head(5)

# Calculating the mean and median values of all the AOV's
meanAOV = sneakerDataByShop['AOV'].mean()
medianAOV = sneakerDataByShop['AOV'].median()
print("Mean AOV is", meanAOV)
print("Median AOV is", medianAOV)
```

Q/A

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

-> The error in calculation most likely happened because of using the count() function on the column 'Total Items'.

-> Instead, the sum() function should be used as shown above.

-> The count function gives the total number of rows instead of total items which is wrong.

b) What metric would you report for this dataset?

=> Looking at the AOV's of all the shops, we can observe some outliers.
=> These outliers have a drastic effect on the mean value which comes out to be 407.99.
=> The median that is 153.0 however gives us a much more accurate description of the data.
=> Thus, instead of using mean, we should use median which is 153.0 as a metric.

c) What is its value?

=> Mean AOV = 407.99
=> Median AOV = 153.0 <-- Appropriate Metric

Question 2

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

```
SELECT count(OrderID) AS "No of Orders" from Orders where ShipperID = (
        select ShipperID from Shippers where ShipperName = 'Speedy Express'
)
```

Output --> 54

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

```
select LastName from Employees where EmployeeID = (
        select EmployeeID from Orders group by EmployeeID order by count(OrderID) desc limit
        1
)
```

Output --> 'Peacock'

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

```
SELECT pd.ProductName
FROM Customers cus, OrderDetails od, Orders ord, Products pd
WHERE cus.CustomerID = ord.CustomerID AND od.ProductID = Pd.ProductID
AND ord.OrderID = od.OrderID AND cus.Country ='Germany'
GROUP BY pd.ProductName
ORDER BY sum(od.Quantity) DESC
LIMIT 1
```

Output --> 'Boston Crab Meat'