Question 1

Code (Was written for a jupyter notebook)

```
import pandas as pd
import numpy as np
from io import StringIO
import io
from google.colab import files
uploaded = files.upload()
import matplotlib.pyplot as plt

df =  pd.read_csv(io.BytesIO(uploaded['2019 Winter Data Science Intern Challenge Data Set - Sheet1.csv']))
df

aov = df['order_amount'].mean()
print(aov)

ind = df.columns.get_loc('order_amount')
orders = df.iloc[:, [ind]]
orders
orders.boxplot(vert=False)
plt.subplots_adjust(left=0.01)
plt.show()

med = df['order_amount'].median()
print(med)
```

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

Answer: If we know that our naive AOV calculation is much higher than what we would expect it to be, it is reasonable to think that the data is skewed due to the prescence of outliers. By plotting our data as a scatterplot, we can confirm this, by seeing the presence of outliers. Because AOV calculations take the average of all orders, outliers with an extremely high value can skew our averages, causing our AOV to be much higher than what we expect.

A better way to evaluate this data would be to either remove all outliers, or to use a different metric which is not so heavily influenced by outliers.

b. What metric would you report for this dataset?

Answer: I would use the median of all the order values (median order value) rather than the average order value. Because the median represents the midpoint of all of our values, it is less influenced by skew and outliers. Very high order values are counted the same as moderately high or even slightly high order values, making the median a more accurate representation of what we might expect a customer to spend on a shoe.

c. What is its value?

Answer: The median is 284.0

Question 2

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

Answer: 54 orders. The query used was

SELECT count(orderID) FROM orders a, shippers b where a.shipperid = b.shipperid and b.shippername = 'Speedy Express'

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

Answer: Peacock. The query used was

SELECT TOP 1 b.LastName FROM orders a, employees b WHERE a.employeeID = b.employeeID GROUP BY a.employeeID,b.LastName ORDER BY COUNT(a.employeeID) desc

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

Answer: Boston Crab Meat. The query used was

SELECT TOP 1 d.productName FROM Customers a, orders b, orderDetails c, products d where a.customerID = b.customerID and b.orderid = c.orderid and d.productID = c.ProductID and a.country = 'Germany' group by c.productID,d.productName order by sum(c.Quantity) desc