Question 1:

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.

It seems that the erroneous AOV was obtained by simply taking the mean of the “order_amount” column. The total items should have been taken into account.

```
[1]: import pandas as pd
     data = pd.read_excel(r'2019 Winter Data Science Intern Challenge Data Set.xlsx')
     data["order_amount"].mean()

[1]: 3145.128
```

In order to evaluate the AOV properly, I used Python to create a dictionary where the keys are the *shop_id*s and the values are arrays with length 2, with the first entry representing the total order amount throughout all orders, and the second entry representing the total number of items throughout all orders for each respective shop_id. This was calculated with the following cell:

```
[2]: shop_dict = {}

     for index, row in data.iterrows():
         if row["shop_id"] not in shop_dict:
             shop_dict[row["shop_id"]] = [row["order_amount"], row["total_items"]]
         else:
             shop_dict[row["shop_id"]][0] += row["order_amount"]
             shop_dict[row["shop_id"]][1] += row["total_items"]
```

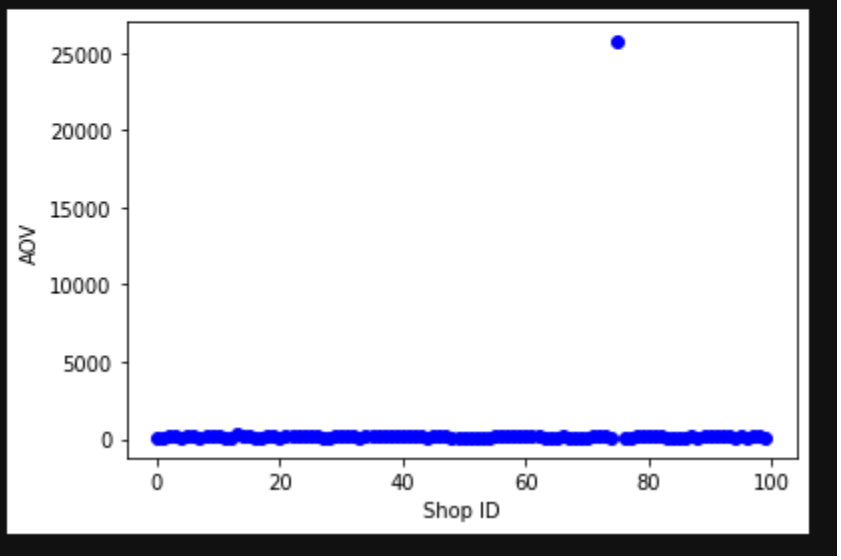
From here, the AOV can be calculated by dividing the order amount by the total items for each shop and appending them to an array and taking the average of the array.

```
[3]: item_price = []
     for key, value in shop_dict.items():
         item_price.append(value[0]/value[1])

     sum(item_price)/len(item_price)

[3]: 407.99
```

b. What metric would you report for this dataset?

I decided to plot the AOVs for every individual shop:

```
[10]: import matplotlib.pyplot as plt
      plt.plot(item_price, 'bo')
      plt.xlabel("Shop ID")
      plt.ylabel("AOV")
      plt.show()
```



We can see that one of the shops has an AOV of around $25000, which is significantly more expensive than all of the other shops (and what one expects of an average shoe price). This means that this particular shop is spiking up the AOV, and the AOV one would expect for each shop is likely much lower than $407.99 ($400 is quite a steep price for shoes!)

I would use the median order value instead because outliers would not have an effect on the median as the actual AOV numbers aren't taken into account when calculating the median.

c. What is its value?

```
[15]: import statistics
      statistics.median(item_price)

[15]: 153.0
```

The median order value is $153.

Question 2:

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

SQL Query:

```
SELECT COUNT(ShipperID)
FROM Orders WHERE ShipperID=
(SELECT ShipperID FROM Shippers
WHERE ShipperName = "Speedy Express")
```

54 orders were shipped by Speedy Express in total.

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

SQL Query:

```
SELECT LastName FROM Employees
WHERE (SELECT EmployeeID
        FROM Orders GROUP BY EmployeeID
        ORDER BY COUNT(EmployeeID)
        DESC LIMIT 1) = EmployeeID
```

The employee with the most orders has the last name Peacock.

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

SQL Query:

```
SELECT ProductName FROM Products WHERE
    (SELECT ProductID FROM OrderDetails WHERE OrderID IN
        (SELECT OrderID FROM Orders WHERE CustomerID IN
            (SELECT CustomerID FROM CUSTOMERS
            WHERE Country = "Germany"))
```

GROUP BY ProductID
ORDER BY SUM(Quantity)
DESC LIMIT 1) = ProductID

The product ordered the most by customers in Germany was Boston Crab Meat.