Winter 2022 Data Science Intern Challenge

Please complete the following questions, and provide your thought process/work. You can attach your work in a text file, link, etc. on the application page. Please ensure answers are easily visible for reviewers!

Question 1: Given some sample data, write a program to answer the following: click here to access the required data set

> 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.
   R: Looking at the statistical analysis of the order values it is clear that while the mathematical average of the value of a sneakers order is 3145.18 dollars, the standard deviation of the mean is significantly higher suggesting that the distribution of the order amounts is very heterogenous. Further looking at the minimum and maximum order amount we can easily see that this is in fact the case. While the minimum order value is 90.00 dollars, the maximum is 704,000.00 dollars. This disparity in order values in turn introduces bias in the average order value causing it to be much higher than the "real" average value of a sneakers order.

b. What metric would you report for this dataset?
   R: Looking at the distribution of each order amount it is clear that there are a few orders that are clear outliers and represent customers that placed orders for an abnormal number of sneakers (e.g. 2000 sneakers in a single order). These are not representative of the majority of the orders but because their values are so high they push the average order value up. Considering the outlier problem, in my view the best way to report the real average order value of this dataset would be to report the median, which is essentially the central number of the order amount values.

c. What is its value?

R: The median order value for this data set is 284.00 dollars

Question 2: For this question you'll need to use SQL. Follow this link to access the data set required for the challenge. Please use queries to answer the following questions. Paste your queries along with your final numerical answers below.

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

```
SELECT ShipperName, count(ShipperID) FROM [Orders]
join [Shippers]
Using(ShipperID)
group by ShipperName
order by count(ShipperID)
```

R: Speedy Express shipped 54 orders

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

```
SELECT LastName, count(EmployeeID) FROM [Employees]
join Orders
using (EmployeeID)
group by LastName
order by count(EmployeeID) DESC
```

R: The last name of the employee with most orders is Peacock

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

```
SELECT ProductName, Quantity , sum(Quantity) as number_items, coun
t(Quantity) as number_orders
FROM Orders, OrderDetails
JOIN Customers ON Orders.CustomerID=Customers.CustomerID
JOIN Products ON OrderDetails.ProductID=Products.ProductID
WHERE Country='Germany'
group by 1
```

```
order by sum(Quantity) DESC
```

R: The product ordered the most in Germany was Gorgonzola Telino, with 350 orders and a total of 11450 items.