Winter 2021 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.

Based to the data, the AOV-Average Order Value of $3145.13 is the average or mean of the attribute order_amount. **Here, mean takes into all the data points in the data set. The mean is affected in case of any outliers, they can increase or decrease the mean depending on the number of outlier values.**
Also, the std(standard deviation) is 41282.53, which is large. It means that the data points are spread indicating that mean is not a good metric to measure this dataset.
So, we will take a closer look at the order_amount values to understand the outliers.

b. What metric would you report for this dataset?

The best metric to report for this dataset is median as the data is skewed, making median a better measure. The data is not normally distributed, so the outliers can skew the mean or average drastically. In this data set we saw there were multiple values over the median which causes the mean to increase unreasonably giving us a wrong representation of the data.

c. What is its value?

The **value is $280** with a standard deviation of 144.45 which shows a smaller spread.

-----Question 2 cont. on next page

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?

Query:

```
Select count(ShipperID) from (Orders)
where ShipperID=1
```

Answer: **count(ShipperID) = 54 . There were 54 shipped by Speedy Express in total.**

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

Query:

```
select o.EmployeeID, e.LastName, count(o.EmployeeID)
from Employees e inner join Orders o on e.EmployeeID = o.EmployeeID
group by(LastName)
order by count(e.EmployeeID) desc
```

Answer: **Peacock is the last name of the employee with the orders.**

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

Query:

```
select c.CustomerID, Country, d.OrderID, p.ProductID, ProductName from
Customers c inner join Orders o on c.CustomerID = o.CustomerID
inner join OrderDetails d on d.OrderID = o.OrderID
inner join Products p on p.ProductID = d.ProductID
where Country='Germany'
group by(c.CustomerID)
```

Answer: **Boston Crab Meat was ordered the most by customers in Germany.**