Fall 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.

The calculation takes outliers into account when determining a central measure. This data can be evaluated by removing the outliers, or the median statistic can be reported. Performing the sensitivity analysis shows that when excluding all order sizes greater than 1000, there is still little variation in the mean. The outlier-adjusted mean is 301.1 (orders < 1000) & 302.6 (orders < 5000 & orders < 10,000). The median is 284 which also paints a more accurate picture.

b. What metric would you report for this dataset?

I would report the median order. The data suggests that there are frequent high and low orders in the system, and they would balance out by using the median, the central measure. Furthermore, when conducting the sensitivity analysis, it became apparent that the adjusted means have no impact on the median and are close to the median (within ~ 6%). This suggests that a median is a strong central measure.

c. What is its value?

Median is 284

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?

54;

```
Select Count(1) from Orders
where ShipperID =
(SELECT ShipperID FROM [Shippers] where ShipperName = "Speedy Express")
```

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

Last Name = Peacock

```
Select LastName From Employees where EmployeeID =
(Select EmployeeID From
(Select EmployeeID, Max(n) From
(Select EmployeeID, Count(*) n From Orders Group By EmployeeID)))
```

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

Product: Gorgonzola Telino

```
Select ProductName From Products where ProductID =
(Select ProductID From
(Select ProductID, Max(n)From
(Select ProductID, Count(*)n From OrderDetails where OrderID in
(Select OrderID From Orders Where CustomerID in
(Select CustomerID From Customers where Country = "Germany"))
Group By ProductID)))
```