Summer 2022 Data Science Intern Challenge

***Sunay Bhat***

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](https://docs.google.com/spreadsheets/d/16i38oonuX1y1g7C_UAmiK9GkY7cS-64DfiDMNiR41LM/edit#gid=0)

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.

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

*The naive Average Order Value calculation is done by taking the mean of all the order values. The issue with this dataset is* ***large outliers skewing the mean*** *- our data is not normally distributed.*

1. What metric would you report for this dataset?

*The simplest option is to* ***take the median or mode****. More complicated options might include assuming a multi-modal distribution and using a cluster method to extract the AOV for each group. This notebook attached below shows both of those methods*

1. What is its value?

* **median** $284.00
* **mode**: $153.00
* **K-means** (3):
  + 1: $302.58 (Majority of orders: 98.75%)
  + 2: $49,213.04 (0.92% of orders)
  + 3: $704,000.00 (0.34% of orders)

**Question 2:** For this question you’ll need to use SQL. [Follow this link](https://www.w3schools.com/SQL/TRYSQL.ASP?FILENAME=TRYSQL_SELECT_ALL) 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.

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

**196**,

Query:

SELECT COUNT(\*) FROM Orders

WHERE ShipperID=3

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

**Handel**

Queries (I know this can be done more efficiently with JOINs, but I have not learned SQL).

SELECT CustomerID, COUNT(\*) FROM Orders

GROUP BY CustomerID

ORDER BY 2 DESC

SELECT \* From Customers

WHERE CustomerID = 20

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

*I don’t know SQL, but I can learn if its core to the job. This query gets you the product ID’s. Need to sort by counts, take the top value, and query the products table for it.*

(SELECT ProductID

FROM OrderDetails

WHERE OrderID IN (SELECT OrderID FROM Orders

WHERE CustomerID IN (SELECT CustomerID FROM Customers

WHERE Country = 'Germany')))