# 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](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.

* We calculated AOV without checking outliers which would turn our calculation wrong.

I found outliers that 17 orders have huge quantity sold at 2000 items ~ $704,000/orders while other orders have quantity from 1 to 8 items only and product in shop number 78 has high price at $25,725 compare to products in others stores which have price in a range from $90 to $352.

* Each store sells only one model of shoe, the price of products are different and user would purchase more than 1 product (place many orders in different stores) so to evaluate this data, I would like to calculate an average user spending.

1. What metric would you report for this dataset?

Average order value and average user spending

1. What is its value?

Average order value in March (exclude orders have 2000 items and orders have items which have price is $25,725): $302.58

Average user spending in March (exclude orders have 2000 items and orders have items which have price is $25,725): $4,979.47

Here is the link of my analysis:

<https://github.com/dinhkimhong/shopify_challenges/blob/master/shopify_challenges_1.ipynb>

**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? - Answer: 54 orders

SELECT COUNT(\*) FROM Orders

LEFT JOIN Shippers ON Shippers.ShipperID = Orders.ShipperID

WHERE ShipperName = 'Speedy Express';

1. What is the last name of the employee with the most orders? - Answer: Peacock

SELECT Employees.EmployeeID, Employees.LastName, COUNT(\*) AS NumberOrders

FROM Orders

LEFT JOIN Employees ON Employees.EmployeeID = Orders.EmployeeID

GROUP BY Employees.EmployeeID

ORDER BY NumberOrders DESC

LIMIT 1;

Or:

SELECT LastName, MAX(NumberOrders)

FROM

(SELECT Employees.EmployeeID, Employees.LastName, COUNT(\*) AS NumberOrders FROM Orders

LEFT JOIN Employees ON Orders.EmployeeID = Employees.EmployeeID

GROUP BY Employees.EmployeeID);

1. What product was ordered the most by customers in Germany? - Answer: Boston Crab Meat

SELECT Products.ProductID, Products.ProductName, SUM(Quantity) as TotalQuantity

FROM OrderDetails

LEFT JOIN Products ON Products.ProductID = OrderDetails.ProductID

LEFT JOIN Orders ON Orders.OrderID = OrderDetails.OrderID

LEFT JOIN Customers ON Customers.CustomerID = Orders.CustomerID

WHERE Customers.Country = 'Germany'

GROUP BY Products.ProductID

ORDER BY TotalQuantity DESC

LIMIT 1;

OR:

SELECT ProductName, MAX(TotalQuantity)

FROM

(SELECT Products.ProductID, Products.ProductName, SUM(Quantity) as TotalQuantity FROM OrderDetails

LEFT JOIN Products ON Products.ProductID = OrderDetails.ProductID

LEFT JOIN Orders ON Orders.OrderID = OrderDetails.OrderID

LEFT JOIN Customers ON Customers.CustomerID = Orders.CustomerID

WHERE Customers.Country = 'Germany'

GROUP BY Products.ProductID);