# 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](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.
2. What metric would you report for this dataset?
3. What is its value?

START OF Q1

My solution and thought process are both documented in the Jupyter notebook, titled Shopify\_Data\_Science\_Intern\_Challenge\_Solution\_Question#1.ipynb. Thank you!

END OF Q1

**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?

START OF Q2

First I see on the Shippers Table that Speedy Express has ShipperID of 1.

Next I switch to the Orders Table, and write the following query

**SELECT SUM(ShipperID) AS Speedy\_Express\_Orders FROM Orders HAVING ShipperID=1;**

The final numerical answer is 54.

Here, I used SUM() since the question asks for a total amount. I summed a random column from the table of all orders, and set the where condition to ShipperID=1 (Speedy Express).

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

First off I’m interested in finding the employee with the most orders. I can see a column labeled EmployeeID in the orders table. What I can do next is group each EmployeeID with the number of orders they fulfilled, sort the result in descending order, and limit the output to 1 so as to get only the employee with the most orders. I know the top 1 has the most orders because I sorted by descending order. I see the top employee is employee 4, with 40 orders.

SELECT TOP 1 EmployeeID FROM Orders a GROUP BY EmployeeID ORDER BY COUNT(EmployeeID) DESC;

Now I simply query the Employees table and get the last name for employee with EmployeeID=4. A simple WHERE clause will constrain the result to the employee with the most orders.

**SELECT LastName FROM Employees WHERE EmployeeID=(SELECT TOP 1 EmployeeID FROM Orders GROUP BY EmployeeID ORDER BY COUNT(EmployeeID) DESC);**

The final answer is ‘Peacock’

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

I go to the Customer table, and query all customers from Germany. I get the CustomerID

SELECT CustomerID FROM Customers WHERE country="Germany";

I now have all the German customers, which I use to filter the Order table using

SELECT OrderID FROM Orders WHERE CustomerID IN (SELECT CustomerID FROM Customers WHERE country="Germany");

I now have the OrderID for all German orders, and I go to the OrderDetails Table and similar to the last question, I sort the ProductID by the SUM of the quantities of those products (since we are trying to find out which product was bought in the largest quantity by Germans), order the results in descending order so that the most bought product is at the top, then just grab the topmost value of ProductID. This is the ProductID of the most bought item.

SELECT TOP 1 ProductID FROM OrderDetails WHERE OrderID IN (SELECT OrderID FROM Orders WHERE CustomerID IN (SELECT CustomerID FROM Customers WHERE country="Germany")) GROUP BY ProductID ORDER BY SUM(Quantity) DESC;

Now I go to the Products table and perform a lookup of the product name based on ProductID.

**SELECT ProductName FROM Products WHERE ProductID IN (SELECT TOP 1 ProductID FROM OrderDetails WHERE OrderID IN (SELECT OrderID FROM Orders WHERE CustomerID IN (SELECT CustomerID FROM Customers WHERE country="Germany")) GROUP BY ProductID ORDER BY SUM(Quantity) DESC);**

The final answer is ‘Boston Crab Meat’

The result is Boston Crab Meat. Yum!

END OF Q2

Thank you for taking the time to read through my solution!