# Summer 2022 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.

**The AOV calculation is being skewed by a small number of very large order amounts (they make up <1.5% of all orders).**

1. What metric would you report for this dataset?

**The quick and easy solution is to use the median order value (MOV). It is not affected by a few outliers, and by inspection we see that the distribution of order amounts are fairly clustered around the MOV.**

**A slightly longer solution is to find that the orders can be grouped into three categories (see supplementary .pynb file):**

**1) Lower number of units sold, lower cost per sneaker.**

**2) Lower number of units sold, higher cost per sneaker.**

**3) Higher number of units sold, lower cost per sneaker.**

**Each of these groups has a distinct MOV, and for each group the MOV and AOV agree well.**

1. What is its value?

**Quick and easy solution: MOV = $284**

**Longer solution:**

**1) MOV = $284**

**2) MOV = $51,450**

**3) MOV = $704,000**

**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.

**The following solutions use SQL Server/MS Access syntax.**

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

**Result:**

**54**

**Query:**

**SELECT COUNT(\*)**

**FROM Orders**

**INNER JOIN Shippers**

**ON Orders.ShipperID = Shippers.ShipperID**

**WHERE Shippers.ShipperName = 'Speedy Express';**

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

**Result:**

**Peacock**

**Query:**

**SELECT LastName**

**FROM Employees**

**WHERE EmployeeID = (**

**SELECT TOP 1 EmployeeID**

**FROM Orders**

**GROUP BY EmployeeID**

**ORDER BY COUNT(\*) DESC;**

**);**

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

**I was unsure if the question referred to highest quantity bought or most orders placed. I chose to answer the former, although determining the latter can be done by replacing the SUM aggregate with COUNT on the ORDER BY statement (last row).**

**Result:**

**Boston Crab Meat**

**Columns needed from specific tables:**

**Country, CustomerID columns in the ‘Customers’ table.**

**CustomerID, OrderID columns in the ‘Orders’ table.**

**OrderID, ProductID columns in the ‘OrderDetails’ table.**

**ProductID, ProductName columns in the ‘Products’ table.**

**Query:**

**SELECT TOP 1 Products.ProductName**

**FROM (((Customers**

**INNER JOIN Orders ON Customers.CustomerID = Orders.CustomerID)**

**INNER JOIN OrderDetails ON Orders.OrderID = OrderDetails.OrderID)**

**INNER JOIN Products ON OrderDetails.ProductID = Products.ProductID**

**)**

**WHERE Customers.Country = 'Germany'**

**GROUP BY Products.ProductName**

**ORDER BY SUM(OrderDetails.Quantity) DESC;**