# 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.

Link to Python work:

<https://colab.research.google.com/drive/10AUGpPJKTvQe74Y6C3HDCghJI3uuuMdT?usp=sharing>

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

Since we are finding an average, outliers could skew the result. After creating a boxplot with the order\_amount values, it becomes apparent that there are some extreme outliers that are throwing off the average. Just in case the data is correct and there were some expensive orders with a lot of shoes, I created another boxplot this time with average price per shoe (order\_amount / total\_items). However, this still showed some outliers so I decided that using the median would be the most accurate measure for this dataset.

1. **What metric would you report for this dataset?**

Depending on what we are really looking for, it should either be median price of order value, or median price of shoes

1. **What is its value?**

Median order value: $284

Median price of shoes: $153

**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?

**54**

SELECT count(OrderID)

FROM [Orders]

Where ShipperID = 1

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

**Peacock**

SELECT Employees.LastName, count(Orders.EmployeeID) as OrdersSold

FROM Orders

Join Employees

on Orders.EmployeeID = Employees.EmployeeID

Group by Orders.EmployeeID

Order by count(Orders.EmployeeID) desc

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

**Boston Crab Meat**

SELECT od.ProductID, p.ProductName, sum(od.Quantity) as Total

FROM Orders as o

join Customers as c

on o.CustomerID = c.CustomerID

join OrderDetails as od

on od.OrderID = o.OrderID

join Products as p

on p.ProductID = od.ProductID

Where c.country = 'Germany'

Group by od.ProductID

Order by Total desc