# 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

> 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.

a. Think about what could be going wrong with our calculation. Think about a better way to evaluate this data.
   Exploring the dataset revealed that the 'order_amount' values were heavily skewed to the right. This was caused by the presence of outliers – data points with abnormally large order_amount values. Failing to account for these outliers resulted in an AOV value that – while technically correct – did not provide the intended insight.

   Identifying and excluding these outliers from the dataset would mitigate their impact on the AOV. I chose to do so by using the Interquartile Range (IQR) approach, and identified 141 outliers. The distribution of the remaining 4,859 data points (~97.18% of the original dataset) had a mean, median, and standard deviation of 293.72, 272.00, and 144.45. The fact that the truncated mean (293.72) is close to the truncated median (272.00) and also to the median of the entire dataset (284.00) provides additional support to the claim that the value of interest is in the correct ballpark.
b. What metric would you report for this dataset?
   I would report the AOV obtained from the 4,859 data points which fell within the IQR fence.
c. What is its value?
   $293.72.

**Question 2:** For this question you'll need to use SQL. Follow this link 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.

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

```
select count(*) as SpeedyExpressOrders from Orders
where Orders.ShipperID = (
        select ShipperID from Shippers
         where Shippers.ShipperName = 'Speedy Express'
);
```

**Answer: 54**

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

```
select LastName from Employees where Employees.EmployeeID = (
        select top 1 EmployeeID from Orders group by EmployeeID order by count(*) desc
    );
```

**Answer: Peacock**

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

```
select Products.ProductName as TopGermanProduct from Products
        where Products.ProductID =
                (
                 select top 1 OrderDetails.ProductId from (
                 Orders
                    inner join (
                       select CustomerID from Customers where Country =
'Germany'
                ) GermanCustomers
                on Orders.CustomerID = GermanCustomers.CustomerID
)
inner join OrderDetails on OrderDetails.OrderID = Orders.OrderID
group by ProductID
order by sum(OrderDetails.Quantity) desc
);
```

**Answer: Boston Crab Meat**