# Winter 2022 Data Science Intern Challenge

**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.
b. What metric would you report for this dataset?
c. What is its value?

**Solution:**

**1(a)** We can observe that the average order value is $3145.13$ which is erroneous. The minimum order amount is 90 and the maximum order value is 704000. The standard deviation is $41282.54$ which extremely high and this is an indication tha something wrong in the observation which is misleading the average value. So, I did some analysis. It can be seen that the order amount is highly skwed towards the lower value (less than $30000$). However, we did found some ouliers ($704000$,$154350$,$102900$, $77175$, $51450$). Among these values $704000$ occured 17 times. So, apparantly these values are affecting the average value. So, **mean is not the perfect metric for this dataset since it is more sensitive to the outliers .** So, two different alternative approach can be used.

**Approach 1 (Immediate approach):**

**1(b) and 1(c)** While taking the mean is misleading. We can use **median** for this data. The median value for this data is $284.00$. It is our observation that 99.1 percent order values are less than 26000. and the mean is $400.04$ for that case. So, keeping this in my consideration **I would like to report the median for this data set and the value is $284.00$.**

Approach 2 (Alternative approach):

We can isolate the outliers. It might not be a very good idea to remove or isolate the outliers since those data might have some other important insight. So, we can isolate the outliers from the original dataset and make statistical analysis for both the data to get any meaningful insight from our data.

**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?**
**Ans. 54**

**Query:**

```
SELECT COUNT(*) as Total_from_speedy_express FROM Orders group by ShipperID
having ShipperID=(SELECT ShipperID FROM Shippers WHERE
ShipperName='Speedy Express');
```

**b. What is the last name of the employee with the most orders?**
**Ans. "Peacock"**

**Query:**

```
SELECT LastName FROM [Employees] WHERE EmployeeID=(SELECT EmployeeID
FROM [Orders] GROUP BY EmployeeID ORDER BY COUNT(*) DESC LIMIT 1);
```

**c. What product was ordered the most by customers in Germany?**
**Ans. Product Name: "Boston Crab Meat"**

**Query:**

```
SELECT ProductName AS Most_Ordered_Product_from_Germany FROM
OrderDetails o join Products p on o.ProductID=p.ProductID

join orders on o.OrderID=orders.OrderID

join customers on orders.CustomerID=customers.CustomerID Group by
o.ProductID having Country='Germany'

ORDER BY SUM(Quantity) DESC limit 1;
```