# Winter 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.

From the dataset, after plotting both the order amount column and the total items column using frequency graph and boxplot, we could see that the data points that are related to the amount 2000 in total items with order amount 704000 is obviously outliers which biased the average order value (AOV) for a huge amount. This happens because AOV is simply the Total Order Amount divided by the Number of Orders and it is sensitive to such huge outliers. In addition, for store number 78, the orders made in this store is extremely pricey than most other stores. This might also be a factor of the naively calculated AOV value.

Since all the 2000 total item data points comes from the same shop with ID 42 and user ID 607, to get a better result using AOV, we could try eliminating those extreme value from this user of this store and then calculate the AOV. (That is being said. It should be better to some investigation on this shop and this user since it is obviously abnormal to order 2000 sneakers in the same shop 17 times using credit card within 30 days.) For orders from store 78, it could also be excluded to improve AOV.

However a better way to evaluate this data might be using the median value since it is not affected by outliers and is more robust.

1. What metric would you report for this dataset?

I would use median to report for this dataset to give a better understanding of the data.

1. What is its value?

By calculating using R the value or order value is 284.

**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?

SELECT count(\*)

FROM Shippers, Orders

WHERE Shippers.ShipperID = Orders.ShipperID

AND Shippers.ShipperName = "Speedy Express"

Solution: 54

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

SELECT LastName

FROM (SELECT Employees.EmployeeID, Employees.LastName, count(\*) as NumOfOrders

FROM Orders, Employees

WHERE Orders.EmployeeID = Employees.EmployeeID

GROUP BY Employees.EmployeeID

ORDER BY NumOfOrders desc)

LIMIT 1

Solution: Peacock

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

SELECT ProductName

FROM Products

NATURAL JOIN (SELECT Customers.CustomerID, Country, Orders.OrderID, ProductID, Sum(Quantity) as NumOfItems

FROM Customers, Orders, OrderDetails

WHERE Customers.CustomerID = Orders.CustomerID

AND Orders.OrderID = OrderDetails.OrderID

AND Country = "Germany"

Group By ProductID

ORDER BY NumOfItems desc

LIMIT 1)

Solution: Boston Crab Meat