# 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.

Under the assumption that shoes on average will not cost more than 500 dollars, examining the dataset we can know that there is one shop (shop\_id = 78) which has an OV (order\_amount/total\_items) of a little bit more of $25000. This means there is either an error with the price of the shoes the store sells, an error with the number of items, or that shop simply sells crazy expensive shoes. The AOV is probably being calculated as the mean value, which is the reason that the AOV appears very high, the single outlier value is skewing the AOV to a higher value than it should be.

1. What metric would you report for this dataset?

Assuming they are normally distributed, a better metric would be the median of all the order values. The median is an accurate replacement for the mean, and it is a way to counter the effect of a few possible outlier values.

1. What is its value?

$153 per shoe

**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 Orders o

INNER JOIN Shippers s

ON o.ShipperID = s.ShipperID

WHERE s.ShipperName = "Speedy Express";

ANS = 54 orders

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

SELECT LastName

FROM EMPLOYEES e

INNER JOIN(

SELECT EmployeeID

FROM Orders

GROUP BY EmployeeID

ORDER BY COUNT(\*) DESC LIMIT 1

)o WHERE o.EmployeeID = e.EmployeeID;

ANS = Peacock

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

SELECT ProductName

FROM Products p

INNER JOIN(Select ProductID, max(SumProducts)

FROM(SELECT ProductID, SUM(Quantity) SumProducts

FROM (SELECT ProductID, Quantity

FROM OrderDetails od

INNER JOIN(

SELECT OrderID

FROM Orders o

INNER JOIN(

SELECT Country, CustomerID

FROM Customers

WHERE Country = "Germany"

)c WHERE o.CustomerID = c.CustomerID

)a WHERE a.OrderID = od.OrderID)

GROUP BY ProductID)

)r WHERE r.ProductID = p.ProductID;

ANS = Boston Crab Meat