# 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.

If we look at the order\_amount field, we observe that there are many orders with incredibly large value. The largest order is worth **$704,000** and had **2000** items purchased. This order is repeated 17 times on different days, all billed at around 4 pm. These could be bulk orders by some other store buying sneakers for reselling them. The next highest order is for six items and is worth **$154,350**, and so on. These orders might be from an affluent customer buying top range products.

We observe that there are certain orders like the above that *do not represent the typical customer*. Such orders may be considered **outliers**. In our analysis, we are taking the average value of all the orders, including the ones for the outliers. Since these outliers are skewing our data towards a higher average order value (AOV), it is also affecting our metric. Thus this metric clearly does not represent a typical daily customer.

A metric that we should use to evaluate this data must not be affected by such outliers.

1. What metric would you report for this dataset?

Since we want our metric to represent a typical customer and not be affected by incredibly large orders, we can use a median order value instead. A median is a measure of central tendency that describes the value at the center, or at the 50% percentile of the entire dataset. In this case, considering the median order value would be appropriate as it would be the order value of the centermost order and is not skewed by any outliers, in this case bulk orders or orders from affluent customers. A median would therefore more accurately represent a typical order value.

1. What is its value?

The median order value is **$284**. This is much more realistic for sneaker orders!

**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?

**Query:** SELECT COUNT(OrderID) AS "Orders shipped by Speedy Express" FROM Orders LEFT JOIN Shippers ON Orders.ShipperID = Shippers.ShipperID WHERE Shippers.ShipperName = 'Speedy Express';

**Answer:** 54

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

**Query:**

SELECT Employees.LastName AS "Employee last name with the most orders" FROM Employees WHERE Employees.EmployeeID = (

SELECT Orders.EmployeeID

FROM Orders

GROUP BY Orders.EmployeeID

HAVING COUNT(Orders.OrderID) =(

SELECT MAX(Order\_Count)

FROM (

SELECT COUNT(Orders.OrderID) AS Order\_Count, Orders.EmployeeID FROM Orders GROUP BY Orders.EmployeeID

)

)

);

**Answer:** Peacock

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

**Query:**

SELECT Products.ProductName AS "Product ordered most by German customers" FROM Products

WHERE ProductID = (SELECT MaxProductID FROM (

SELECT ProductID\_Count, ProductID AS MaxProductID

FROM (

SELECT COUNT(OrderDetails.ProductID) AS ProductID\_Count, OrderDetails.ProductID

FROM Orders

LEFT JOIN OrderDetails ON Orders.OrderID = OrderDetails.OrderID

WHERE Orders.CustomerID IN (SELECT Customers.CustomerID FROM Customers WHERE Customers.Country = 'Germany')

GROUP BY OrderDetails.ProductID

)

)

WHERE ProductID\_Count = (

SELECT MAX(ProductID\_Count) FROM (

SELECT ProductID\_Count, ProductID

FROM (

SELECT COUNT(OrderDetails.ProductID) AS ProductID\_Count, OrderDetails.ProductID

FROM Orders

LEFT JOIN OrderDetails ON Orders.OrderID = OrderDetails.OrderID

WHERE Orders.CustomerID IN (SELECT Customers.CustomerID FROM Customers WHERE Customers.Country = 'Germany')

GROUP BY OrderDetails.ProductID

)

)

)

)

**Answer:** Gorgonzola Telino