# Fall 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.

The issue with the calculation is that the average order value was calculated as the average of the total order amount, not factoring in the average order value relative to the size of the order. The data also shows orders with high unit volumes which is contributing to the large incorrect average order value calculation. To account for this, a new column was created representing the average order amount per order (order\_amount / total\_items). However, the average of this new field will still be positively skewed by sneaker orders with luxury pricing.

1. What metric would you report for this dataset?

Luxury priced sneaker sales are a reality in the sneaker market and will continue to occur and are valid datapoints. Instead of removing these outliers, the metric that I would report for this dataset would be the median order value since it is more robust and not influenced by orders with an extremely high average order values.

1. What is its value?

The median order value across 100 of Shopifys sneaker shops within the 30-day timespan was $153.

**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(OrderID) as "Total Orders Shipped by Speedy Express"

FROM Orders

WHERE ShipperID = 3

**Answer:** In total, 68 Orders were shipped by Speedy Express.

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

SELECT e.LastName, COUNT(OrderID) as "Order Count"

FROM Employees e

JOIN Orders o

ON e.EmployeeID = o.EmployeeID

GROUP BY e.EmployeeID

ORDER BY COUNT(OrderID) DESC

LIMIT 1

**Answer:** The last name of the employee with the most orders was “Peacock” with an order count of 40.

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

SELECT p.ProductName, COUNT(od.OrderID) as “Order Count”, SUM(od.quantity) as “Units Sold”

FROM Products p

JOIN OrderDetails od

ON p.ProductID = od.ProductID

JOIN Orders o

ON o.OrderID = od.OrderID

JOIN Customers c

ON c.CustomerID = o.CustomerID

WHERE c.Country = "Germany"

GROUP BY od.ProductID

ORDER BY SUM(od.quantity) DESC

**Answer:** In terms of units sold, the most ordered item in Germany was the Boston Crab Meat (4 orders, 160 units). In terms of order count, the most ordered item in Germany was the Gorgonzola Telino (5 orders, 125 units).