# 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.

I would start this analysis by checking to ensure I get the same AOV as mentioned above to make sure the error wasn’t a simple typo. I now look to see whether an average is an appropriate metric by looking at the range of the data, the skewness, and a visualization (box-and-whisker plot). Here is the visualization I used, a boxplot with log of order amounts:

Chart, box and whisker chart

Description automatically generated

1. What metric would you report for this dataset?

Please refer to the plot above. Considering I used a log scale, there appear to be some pretty extreme outliers indicating that median might be a more appropriate metric to use thanks to the fact that outliers aren’t as impactful.

1. What is its value?

I calculated a median order value (MOV) of $284, a much more reasonable figure

**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(o.OrderID) as 'total orders shipped by Speedy Express'

FROM Orders o

JOIN Shippers s ON o.ShipperID = s.ShipperID

WHERE ShipperName = 'Speedy Express';

The total number of orders shipped by Speedy Express is 54

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

SELECT COUNT(e.EmployeeID) as 'Number of Orders', e.LastName

FROM Employees e

JOIN Orders o ON o.EmployeeID=e.EmployeeID

GROUP BY e.EmployeeID

ORDER BY COUNT(e.EmployeeID) DESC

LIMIT 1;

The last name of the employee with the most orders is Peacock

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

SELECT ProductName, COUNT(ProductName) as 'total number ordered', c.Country

FROM Products p

JOIN OrderDetails od ON od.ProductID=p.ProductID

JOIN Orders o ON o.OrderID=od.OrderID

JOIN Customers c ON c.CustomerID=o.CustomerID

WHERE Country='Germany'

GROUP BY ProductName

ORDER BY COUNT(ProductName) DESC

LIMIT 1;

The product ordered most by German customers is Gorgonzola Telino