# Fall 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.

Errors in the data could produce results that are largely different from the truth. Therefore, it’s important to see what values cause this unexpected result. Using Pandas data frames, I can filter the data to find the values that are incorrect and then fix them based on logic. Doing this process, I found two stores whose AOV was extraordinarly high and looked at what values caused this. The first store, store 42, had a user who consistently inputted 2000 orders instead of 2 orders, which made the value of the purchases 704000 instead of 704. The second store, store 78, seemed to not use a decimal place when writing all their order amounts.

1. What metric would you report for this dataset?

I look at the sales data mentioned above as incorrect and in need of change instead of dismissing the values entirely. Therefore, I believe it would be useful to use the AOV but with corrected values.

1. What is its value?

The AOV after correcting incorrect values is $443.22

The Jupyter Notebook on which I did the analysis can be found on github.com/vangampalli

**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 sum(quantity)

From Orders o1,orderdetails od1

Where o1.orderid=od1.orderid and ShipperID = 1

Result:

3575

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

Query:

Select a1.employeeID, b1.lastname, sum(quantity) as orders

FROM [Orders] a1,orderdetails od1, Employees b1

where a1.orderid=od1.orderid and a1.employeeID = b1.employeeID

group by a1.employeeID, b1.lastname

order by orders desc

limit 1

Result:

Peacock, with 3232 orders

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

Query:

SELECT productname ,sum(quantity)orders

FROM [Orders] o1, [OrderDetails] od1,[Customers] c1, [Products]p1

where o1.orderID = od1.orderID and o1.customerID = c1.customerID and od1.productID =p1.productID

and country = 'Germany'

group by productname

order by orders desc

limit 1

Result:

Boston Crab Meat, with 160 orders