# 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.
2. What metric would you report for this dataset?
3. What is its value?

**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?
2. What is the last name of the employee with the most orders?
3. What product was ordered the most by customers in Germany?

**Solutions:**

**Ways to solve the problem:**

1. High skewness is observed in the AOV value after calculating; it’s because of the outliers present in the data. Verify the data by understanding the statistics of it with the records. If they are verified as legit, then we can say, these outliers are nothing but from the customers placing bulk orders. One of the ways to deal with this is to remove those outliers and calculate AOV, but before them we need to compare them by grouping order’s and calculating their average amount.

If they are not common and only present in month’s data, they can be removed to calculate the mean value. If they are common in every month’s data and the data is getting affected by these outliers then, we need to consider alternativeaggregate function such as **median** for the given data.

The average value is highly peaked because of the Order 42, Order 78

**Order 42 Amount- 235101.4902**

**Order 78 Amount- 49213.04348**

1. Median.
2. $308.89

**Question: -2**

a)

**SQL Query:**

SELECT count(\*)

FROM [Orders] o

inner join shippers s on o.shipperId = s.shipperid

where shippername ='Speedy Express'

b)

**SQL Query:**

SELECT count(lastName), lastName

FROM [Employees] e

inner join orders o on e.employeeId=o.employeeId

group by lastName

c)

**SQL Query:**

select count(p.productId), p.productName

from products p inner join orderdetails o on p.productId = o.productId

inner join suppliers s on s.supplierId= p.supplierId

where country='Germany'

group by p.productId

order by count(p.productId) desc limit 1