# 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. The AOV value after calculation is highly skewed because of the outliers present in the data. First, we need to verify this data with the records, so that they are entered correctly and there isn’t any error. 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 this given month’s data with rest of the months to check if these outliers/bulk orders are common in every month. If they are not common and only present in this 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.
2. Median.
3. $308.8898

**Question: -2**

a)

**Code**:

SELECT count(\*)

FROM [Orders] o

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

where shippername ='Speedy Express'

**OUTPUT**:

Graphical user interface, application, Teams

Description automatically generated

b)

**Code**:

SELECT count(lastName), lastName

FROM [Employees] e

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

group by lastName

**Output**:

Table

Description automatically generated

c)

**Code**:

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

**Output:**

Graphical user interface, text, application

Description automatically generated