# 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.

* The likely error with the current calculation arises from anomalies in the dataset. In particular from user\_id: 607, which we see make continues purchases of 2000 items at shop\_id: 42. Filtering data which exceeds a threshold order\_amount or total\_items, lets say 20 items. Helps us better evaluate the data as it is likely either the shop or the service would limit the qty of items purchased by a user. Not to mention large order quantities in the sneaker community are generally by scalpers and/or bots.

1. What metric would you report for this dataset?

* I would still proceed with calculating an AOV as the evaluation metric to best understand how sneaker shops generally perform on shopify services. However, noting the above filtering technique to remove extreme anomalies. A potential alternative is to look at the monthly profit distribution by shop. Outside of the scope of this question, I think to improve this analysis in a production setting, taking into consideration return/refunded orders would affect the performance metrics of our sneaker shops. In which case looking at a different performance metric may be beneficial (such as average satisfactory transactions over a 30-day window)

1. What is its value?

* The value of a filtered AOV is:  
  $754.09

Note: Answers to the above are also given by my program submission.

**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(ShipperID)

FROM Orders

WHERE ShipperID == (SELECT ShipperID FROM [Shippers] WHERE ShipperName = 'Speedy Express');

**Answer: 54**

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

SELECT LastName

FROM Employees

WHERE EmployeeID ==

(SELECT `EmployeeID`

FROM `Orders`

GROUP BY `EmployeeID`

ORDER BY COUNT(\*) DESC

LIMIT 1)

**Answer: Peacock**

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

SELECT ProductName FROM [Products]

WHERE ProductID ==

(SELECT ProductID FROM [OrderDetails]

WHERE OrderID IN

(SELECT OrderID FROM [Orders]

WHERE CustomerID IN

(SELECT CustomerID FROM Customers

WHERE Country IN ('Germany')))

GROUP BY `ProductID`

ORDER BY COUNT(\*) DESC

LIMIT 1)

**Answer: Gorgonzola Telino**