# 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.

\*answers are calculated in the ShopifyDataScienceChallenge.ipynb file

1. Think about what could be going wrong with our calculation. Think about a better way to evaluate this data.

The value is being calculated is using the sum of the order amounts divided by the number of orders. This is giving the average of the number of orders, but due to some outliers this value is incorrect, to fix this we can remove the outliers using either the standard deviation or using the percentile points to get calculate an upper or lower bound of the data. This might not be the most accurate method but if you want to calculate the basic mean this would give you a better answer.

1. What metric would you report for this dataset?

A better metric would be to use weighted average that way we know the average value spent on a pair of shows rather than the amount the average per order, since this accounts for large quantity orders. Giving a more realistic value representation for money spent in each order. Along with this another metric that would be useful would the weighted average of each shop to see which shops are creating more value.

1. What is its value?

The value of the resulting value for the average of cost per item is approximately $358.

**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.

\*answers are also in the dataScienceChallengeP2.sql file

1. How many orders were shipped by Speedy Express in total?

Query:

select count(OrderID) from Orders

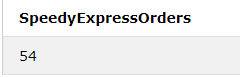
inner join Shippers

on Shippers.ShipperID = Orders.ShipperID

where ShipperName = "Speedy Express"

group by Orders.ShipperName;

Result:



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

Query:

with EmployeeOrders(OrdersCount,EmployeeID) as

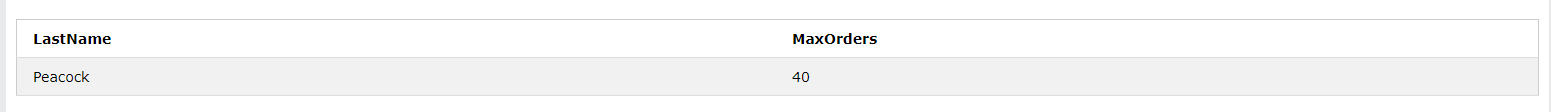
(select (count(OrderID)) as OrderCount,EmployeeID from Orders group by EmployeeID),

MaxOrders(MaxOrders,EmployeeID) as

(select MAX(OrdersCount) as MaxOrders, EmployeeID from EmployeeOrders)

Select LastName, MaxOrders from Employees Join MaxOrders on MaxOrders.EmployeeID = Employees.EmployeeID;

Result:



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

Query:

with GermanCustomers as

(select CustomerID, Country from Customers where Country = "Germany"),

GermanOrders as

(select \* from Orders inner join GermanCustomers on GermanCustomers.CustomerID = Orders.CustomerID inner join OrderDetails on OrderDetails.OrderID = Orders.OrderID),

ProductsOrderedInGermany as

(select count(ProductID) as ProductCount,ProductID,ProductName from GermanOrders inner join Products using (ProductID) group by ProductID order by ProductCount DESC)

select max(ProductCount),ProductID,ProductName from ProductsOrderedInGermany;

Result:

