# 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.

[**HERE**](https://docs.google.com/spreadsheets/d/1qeRZHofc4UlaQ9uQerFutHVtAhBae84y0tDNdbKmlpU/edit?usp=sharing) **is a link to my adjustments to the dataset, as seen on subsequent Sheets.**

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

**By sorting the order\_amount column, we can see that there are some shops with extreme outliers. In particular, Shop 42 and Shop 78 seem to have something interesting going on with them. Given that each shop sells only one sneaker model at a fixed price, I added a column to calculate that shop’s sneaker model cost to review further. Additionally, the Median of the Order Value is only $284.00. Let’s investigate…**

**It appears that Shop 42 has made 17 sales to User 607 of 2000 items for a total of $704,000 each. These very expensive orders are hugely skewing the data, as evidenced by the larger-than-anticipated AOV of $3,145.13, as well as the Standard Deviation of the Order Value of $41,282.54 (wow!). However, because the model\_cost for these orders is reasonable ($352 each), I will look at model\_cost outliers first.**

**If you sort by model\_cost, you find that Shop 78 only sells sneakers that cost $25,725 each. The next-most-expensive sneaker sold by any sneaker store is $352. We can certainly identify that Shop 78’s sneakers are not a reasonable representation of sneaker stores on Shopify. On the next Sheet - “Remove Shop 78” - I show calculations after Shop 78 has been removed from the dataset.**

**On the “Remove Shop 78” Sheet, you see that we now only show shops that host reasonably-priced sneaker models, but we still have an outrageous Standard Deviation of $41,156.00. This tells me that we will need to address User 607’s purchases as well. (User 607 may be a retailer making wholesale purchases, rather than individual purchasers making direct purchases. In any case, User 607 is not representative of the overall Order Values!) On the third sheet - “Then Remove User 607” - you can see the data after User 607’s purchases have been removed. Now, we have an Average Order Value of $302.58, which is much more reasonable compared to the Median Order Value of $284.00.**

1. What metric would you report for this dataset?

**If I am only able to report one value, I would report the Average Order Value of $302.58, which is the AOV after the significant outliers are removed. If I am able to give more context to the AOV of $302.58, I would share the two outlier situations above that were removed to give a more realistic picture. These decisions are highly justified by the change in Standard Deviation from the Full Dataset ($41,282.54) to the edited dataset ($160.80)**

1. What is its value?

**The AOV of all sales after the most egregious outliers are removed is $302.58. (It should be noted that there were additional purchases in range of $730-$1760 range that could have been excluded as per the Interquartile Range Outlier Calculation, but I left them in for the sake of conservative reporting, given that there was not a large gap between those values and the IQR.**

**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.

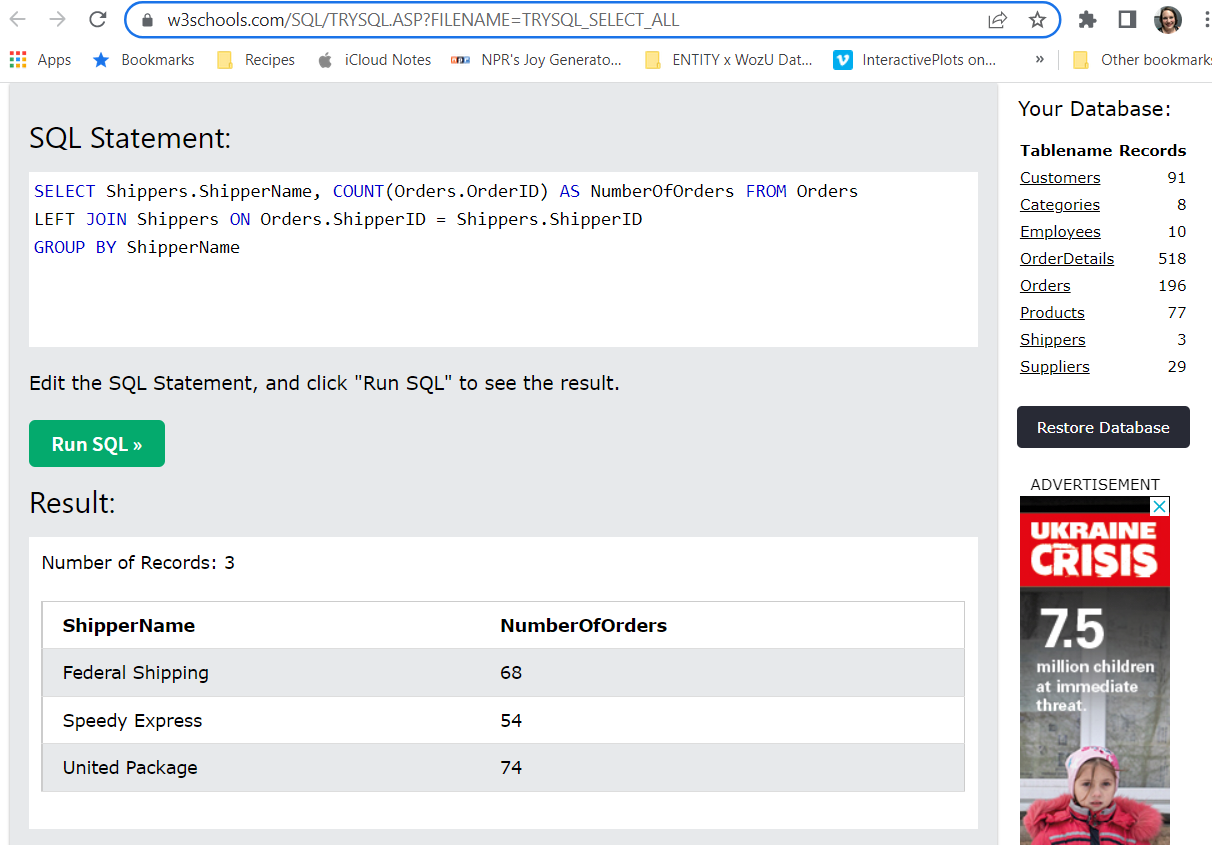
1. How many orders were shipped by Speedy Express in total?

**SELECT Shippers.ShipperName, COUNT(Orders.OrderID) AS NumberOfOrders FROM Orders**

**LEFT JOIN Shippers ON Orders.ShipperID = Shippers.ShipperID**

**GROUP BY ShipperName**

***54 orders were shipped via Speedy Express.***



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

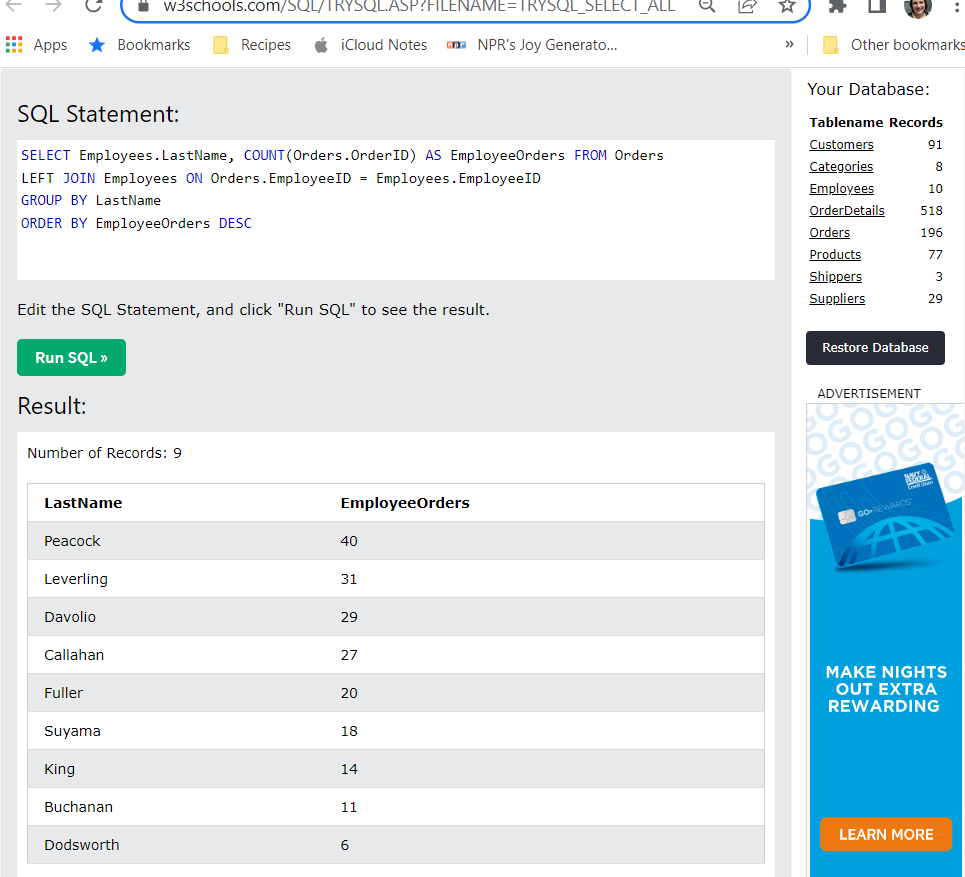
**SELECT Employees.LastName, COUNT(Orders.OrderID) AS EmployeeOrders FROM Orders**

**LEFT JOIN Employees ON Orders.EmployeeID = Employees.EmployeeID**

**GROUP BY LastName**

**ORDER BY EmployeeOrders DESC**

***The Employee with the last name Peacock had the most orders (40 orders).***



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

**SELECT Products.ProductID, Products.ProductName, SUM(OrderDetails.Quantity) FROM ORDERS**

**LEFT JOIN OrderDetails ON Orders.OrderID = OrderDetails.OrderID**

**LEFT JOIN Customers ON Orders.CustomerID = Customers.CustomerID**

**LEFT JOIN Products ON OrderDetails.ProductID = Products.ProductID**

**WHERE Country = 'Germany'**

**GROUP BY Products.ProductID**

**ORDER BY SUM(OrderDetails.Quantity) DESC**

***The Boston Crab Meat was ordered the most by customers in Germany.***

