# Fall 2022 Data Science Intern Challenge

**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 problem was that the AOV was taken from the mean of the *order\_amount* column in the dataset. This means that the AOV didn’t take into account potentially bad data. Specifically, shops 42 and 78 (based on *shop\_id*) were selling sneakers that were wildly outside of the average per sneaker cost.**

**My first inclination was to just use the median, but that was before going through the data.** After looking at the dataset, I noticed the following:

* Shop 78 appeared to have either a currency type problem or data input problem. Being that this was the case, I decided to omit the data into the final model.
* Shop 42’s average sneaker price was closer to other shops, but was still a bit problematic. All of Shop 42’s sales came in March 2017 and based on the average sneaker price $352, I would assume Shop 42 was bundling sneakers, which is very common in the “sneaker world.”
* Furthermore, there was a buyer that bought an incredible amount of sneakers per order, *user\_id* 607. It is possible that this user bought large quantities to resell for their own business. These 2000-item orders came from shop 42– furthering skewing the dataset.

**Ultimately, shop 42 and 78 were removed from the dataset to establish a better model for the AOV.**

1. What metric would you report for this dataset?

Prior to cleaning the dataset, I would have used the median values for AOV; the massive orders by *user\_id* 607 and the extreme pricing by *shop\_id* 78 inherently skewed the data to force this. After cleaning the dataset by removing orders that had an average sneaker value of $25725 (from shop 78) and orders that contained 2000 sneakers per order (by user 607 for shop 42), the model dramatically changed. **The AOV that was most appropriate still came from the median, based on the right skew of the *order\_amount* column**, but the mean value was more acceptable than originally found ($300.16 vs. $3145.13).

1. What is its value?

**The median value I found for the AOV was $284.**

**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.

***FOR FUTURE REFERENCE, I LEARNED THROUGH EXCESSIVE TRIAL AND ERROR THAT SAFARI ON APPLE-BASED OSs YIELD ERRORS WITH webSQL AS HOSTED BY W3SCHOOLS. CHROME WAS ESSENTIAL TO GETTING MY QUERIES TO WORK APPROPRIATELY***

|  |  |
| --- | --- |
| 1. How many orders were shipped by Speedy Express in total? | |
| **Answer: 54** | **SELECT s.ShipperName, COUNT(\*)**  **FROM ORDERS o**  **INNER JOIN Shippers s**  **ON o.ShipperID = s.ShipperID**  **WHERE ShipperName = "Speedy Express"** |
| 1. What is the last name of the employee with the most orders? | |
| **Answer: Peacock** | **SELECT e.LastName, COUNT(e.LastName)**  **FROM Employees e**  **INNER JOIN Orders o**  **ON e.EmployeeID = o.EmployeeID**  **GROUP BY e.LastName**  **ORDER BY COUNT(e.LastName) DESC**  **LIMIT 1;** |
| 1. What product was ordered the most by customers in Germany? | |
| **Answer: Boston Crab Meat** | **SELECT p.ProductName, COUNT(p.ProductID)**  **FROM Customers c**  **INNER JOIN Orders o**  **ON c.CustomerID = o.CustomerID**  **INNER JOIN OrderDetails od**  **ON od.OrderID = o.OrderID**  **INNER JOIN Products p**  **ON p.ProductID = od.ProductID**  **WHERE Country = "Germany"**  **LIMIT 1;** |