# Summer 2022 Data Science Intern Challenge

Josh Hills
josh@joshhills.ca

Shopify — January 19, 2022

**Question 1**

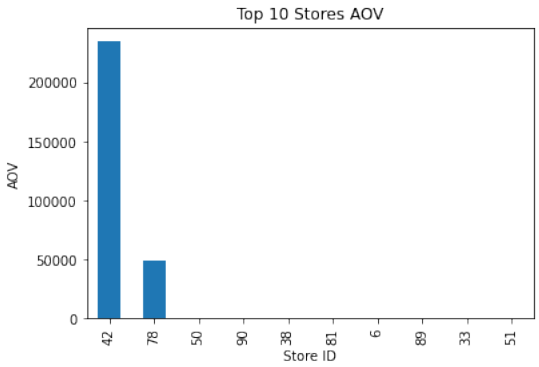
On Shopify, we have exactly 100 sneaker shops, and each of these shops sells only one model of shoe. We want to do some analysis of the average order value (AOV). When we look at orders data over a 30 day window, we naively calculate an AOV of $3145.13. Given that we know these shops are selling sneakers, a relatively affordable item, something seems wrong with our analysis.

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

b. What metric would you report for this dataset?

c. What is its value?

a. The AOV of $3145.13 is clearly too high, if we sort the dataset by the order_amount, we can see some orders with order_amount are way over our mean. These massive sales are coming from two stores, store 42 and store 78, in fact all of store 42's outliers are by the same user_id. Another thing to note is that all of 78's outliers have order value of $25725 per shoe, while 42 has a cost normal cost of $352 a shoe. 78 likely has a mispriced shoe. 42 has a different problem, all sales are completed at 4am on different days with 2000 items an order with total cost of $704000, 42 must have an automatic system malfuncitoning or some other problem. I think a good way to evaluate this data is to drop these outliers, if we look at the mean after droping all outliers theres an AOV of $302.58.




b. Assuming I can't modify the data or add preventative measures, I would use median. The median is relatively close to the AOV after dropping the outliers. I also took a look at the distribution. The median being $284.00 and the AOV after dropping being $302.58. To avoid future descrepancies, I would report the median.

c. The median order value is $284.00.

Question 2

For this question you'll need to use SQL. Follow this link 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.

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

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

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

a. There are 54 orders shipped by Speedy Express.

```
SELECT COUNT(*)
FROM Orders
WHERE ShipperID = 1;
```

b. The last name of the employee with the most orders is Peacock.

```
SELECT Employees.LastName
FROM Employees
INNER JOIN Orders ON Employees.EmployeeID=Orders.EmployeeID
GROUP BY Employees.LastName
ORDER BY COUNT(*) DESC
LIMIT 1;
```

c. The product ordered most by customers in Germany is Boston Crab Meat.

```
SELECT Products.ProductName
FROM (((Customers
INNER JOIN Orders ON Customers.CustomerID=Orders.CustomerID)
INNER JOIN OrderDetails ON OrderDetails.OrderID=Orders.OrderID)
INNER JOIN Products ON Products.ProductID=OrderDetails.ProductID)
WHERE Customers.Country="Germany"
GROUP BY Products.ProductID
ORDER BY SUM(OrderDetails.Quantity) DESC
LIMIT 1;
```