Summer 2022 Data Science Intern Challenge

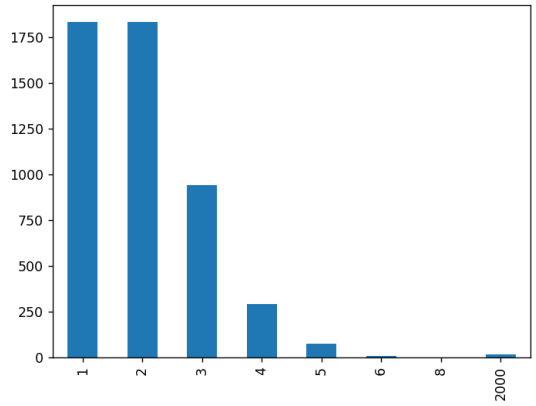
William Yao

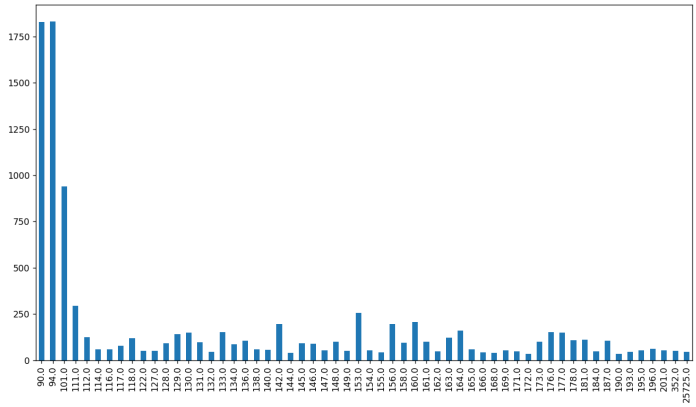
Question 1: *On Shopify, we have exactly 100 sneaker shops, and each of these shops sells only one model of shoe. We want to do some analysis of the average order value (AOV). When we look at orders data over a 30 day window, we naively calculate an AOV of $3145.13. Given that we know these shops are 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 better way to evaluate this data.**

To examine the problem, I wrote a short Python program that loads the .csv file into a dataframe and prints relevant information to the console. Here, my first action was to replicate the problem. Taking the mean value of `order_amounts` yields 3145.13 as expected, and we also find that the time period spans March 1 to March 30, there are 100 stores, and there are no missing entries in the data. This confirms that the problem is not caused by incomplete data or a computational error in Shopify's original program—the mean *is* 3145.13 across all days and stores—but rather with our entire approach to measuring AOV.

Means are easily skewed by outlier values, and there are highly expensive orders positively skewing the AOV. In particular, we observe that orders range from $90 to $704,000—the latter of which is too high for average consumers. Since an order's value is determined by its size * price per sneaker, my program graphs the frequency of each value in these categories to find obvious outliers:





Here, we see that most orders range from 1 to 8 items, with a small minority at 2000. Similarly, most shoes sell for $90 to $352, with a small minority at $25725. Thus, those particular orders will skew the mean, making it an unrepresentative measure of central tendency.

b. **What metric would you report for this dataset?**

Instead of the mean, I would report the median order value (MOV). Medians are less influenced by outliers, and since outliers comprise a minority of our data, the MOV would fall close to the middle order.

Of course, the median does not solve the fundamental issue; there exist two groups of clients: individual consumers, who buy for personal use, and businesses like retailers, who buy in bulk. Individuals outnumber business clients in quantity, but both are equally important to the market. Unfortunately, by

eliminating skew to better represent small customers, the median value now underrepresents trends for larger ones; for instance, if the 2000-item orders were halved to 1000, this would be a critical change in market behavior that is not captured by the median, but would be captured by the mean. The same argument could be made for the orders of $25725 sneakers, which are also underrepresented.

Thus, a more robust system would be to calculate MOVs for small and large clients separately so that trends in each are reflected by their respective metric. This distinction may seem arbitrary for particularly small businesses, but here it is reasonable to separate orders with 1-8 items from those with 2000.

c. What is its value?

The MOV for the entire market 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?

```
SELECT Shippers.ShipperName, COUNT(Orders.ShipperID) AS "Total Orders"
FROM Orders
INNER JOIN Shippers ON Orders.ShipperID = Shippers.ShipperID
GROUP BY Shippers.ShipperID;
```

Answer: 54.

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

```
SELECT Employees.LastName, COUNT(Orders.EmployeeID) AS "Total Orders"
FROM Orders
INNER JOIN Employees ON Orders.EmployeeID = Employees.EmployeeID
GROUP BY Orders.EmployeeID
Order BY COUNT(Orders.EmployeeID) DESC;
```

Answer: Peacock with 40 orders.

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

```
SELECT Products.ProductID, Products.ProductName, SUM(OrderDetails.Quantity)
AS "Total Units", COUNT(Orders.OrderID) AS "Order Count"
FROM Orders
INNER JOIN OrderDetails ON Orders.OrderID = OrderDetails.OrderID
INNER JOIN Products ON Products.ProductID = OrderDetails.ProductID
INNER JOIN Customers ON Orders.CustomerID = Customers.CustomerID
WHERE Customers.Country = "Germany"
GROUP BY Products.ProductID
ORDER BY SUM(OrderDetails.Quantity) DESC;
```

Answer: By number of units sold, Boston Crab Meat was highest with 160 units sold across 4 orders. However, if we are interested in the highest number of times ordered, we can replace the last line with `ORDER BY COUNT(Orders.OrderID) DESC;`, giving Gorgonzola Telino with 5 orders.