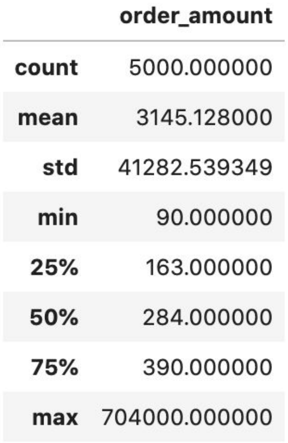
Andrew Lowe
9/2/2020

Shopify Data Science Intern Challenge

Question 1 - See challenge_notebook.ipynb in the repository for code

Over a 30-day window, we can see that the Average Order Value (AOV) is $3,145.13 for sneaker orders on Shopify. While this initial calculation is in fact the mean order amount, it is a misleading representation of the amount spent on most sneaker orders on the platform.

By inspecting the data, shown in the chart to the right, we can see that there is a huge variance in sneaker sales, with the least spent in a single order being $90 and the most spent $704,000. Given that the AOV is so far off from the median—which is $284, shown in the "50%" row on the chart—it is clear that the highest order amount is an extreme outlier.



	order_amount
count	5000.000000
mean	3145.128000
std	41282.539349
min	90.000000
25%	163.000000
50%	284.000000
75%	390.000000
max	704000.000000

If we isolate the outliers on the high-end of the distribution, we can see that there were only two shops represented in the top 1% of the data. One of the shops (shop_id = 42) has 17 orders of the max order sale of $704,000; while the other shop (shop_id = 78) appears to sell sneakers for $25,725 a piece in quantities of 1 to 6.

Solution 1: Find the median

The median is an easy and helpful first solution. We simply order the amount spent on sneaker orders from least to greatest and take the middle number. This middle number is used as a representation of what users are typically spending on the site. As stated before, the median is:

- Solution 1 estimate: $284

Solution 2: Eliminate the extreme values then find the average

The median, however, has a significant drawback: it reflects only the middle value of the dataset. While this is a better representation of the average than the mean in our case, it could give us misleading information when we are trying to, for example, estimate the amount of revenue we would gain if some vendors started making bigger orders.[1] To solve this issue, we could simply eliminate the extreme outliers and take the mean.

After eliminating the two shops with extreme outlier values (shop 42 and shop 78), we calculate an AOV that is much closer to the median:

- Solution 2 estimate: $300.21

Again, the advantage of this solution is that we would be able to make clearer projections with our data. For example, we could now speculate what the changes in revenue for Shopify or a specific vendor might be if they changed their prices or sold in different quantities.

Question 2

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

```
SELECT COUNT(OrderID)
FROM [Orders]
LEFT JOIN Shippers on Orders.ShipperID = Shippers.ShipperID
WHERE Shippers.ShipperName = 'Speedy Express';
```

- Output: 54

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

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

- Output: Peacock

[1] To understand this better, take the set of numbers [2, 5, 6, 10, 11]. The median of this set is 6. But say we increased the last number by ten times, so that the new set was [2, 5, 6, 10, 110]. The median of this set is still only 6 because the middle number in the set remains the same. If we apply this to our problem at hand, then we can see that median doesn't capture differences from one time to another as well as the mean and can't be reliably used to make projections about change.

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

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

- Output: Boston Crab Meat