# Shopify 2022 Data Science Intern Challenge

[Samuel Diaz](https://www.linkedin.com/in/samuel-diaz-data-scientist/)

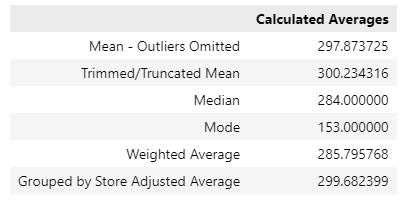
Please complete the following questions, and provide your thought process/work. You can attach your work in a text file, link, etc. on the application page. Please ensure answers are easily visible for reviewers!

**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.

[GitHub Link to my Jupyter Notebook (review my work)](https://github.com/sdman135/DS_Intern_Challenge-S_22/blob/main/DS_Intern_Challenge-S_22.ipynb)

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

 One major issue, after exploring the data, is the appearances of outliers. One store (shop\_id=78) sold each sneaker for $25725.00. Another store (shop\_id=42) had 17 different orders, by the same user (user\_id=607), purchasing 2000 items/sneakers totaling $704,000 (the largest order amount in all the transactions) each transaction.Both very odd behavior. With the calculated AOV of $3145.13 the outliers severely skewed the average. Averages are highly affected by outliers (both large and small). Safest bet would be to work within the range between the 2nd quartile $163.00 or more) and 3rd quartile ($390.00 or less). Outliers will need to be removed or dealt with to correct this issue

.

**𝑂𝑝𝑡𝑖𝑜𝑛 1:** **Manual removal of outliers altogether** Although a viable option, without knowing the full scope of the data or the problem, the removal of outliers, can lead to future problems and/or a large portion of your data being removed.

**Mean - Outliers Omitted = $297.87**

**𝑂𝑝𝑡𝑖𝑜𝑛 2:** **Trimmed/Truncated Mean** (non-permanent removal of some outliers based on a set percentage then calculate the remaining value’s average), Working within 95% of the data (removing top and bottom 2.5% of the data) and calculating the remaining averaging, resulted in the new **Truncated Mean = $300.23**

**𝑂𝑝𝑡𝑖𝑜𝑛 3**: Use a different metric to summarize the data’s central tendencies, by far the easiest solution..

**Median** = the "middle" most value **Median = $284.00**

**Mode** = the most frequently reoccurring value **Mode = $153.00**

**𝑂𝑝𝑡𝑖𝑜𝑛 4**: **Weighted Average**, assigning weights to each order\_amount value based on the scaled values of the order\_amount column. The larger the order\_amount the more it is penalized.

**Weighted Average = $285.80**

***Option 5:* Grouped by Store Adjusted Average** We can bin/group the transactions together based shop\_id and calculate the averages of each group independently from each other. After we can remove the problem stores from these averages (shop\_id=42 & shop\_id=78) then get the collected average of the remaining stores.

**Grouped by Store Adjusted Average = $299.68**

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

I would use Median instead of Average (the easiest of the options). When the numerical data is normally distributed, the sweet spot is right in the middle at the highest point of the bell curve (at 50%), the Median. The median, the "middle" most value when the majority of the values are around the highest point on the bell curve.

Mode (Frequency or most reoccurring value) is another viable option, but the mode=$153.00 (with 87 occurrences) is lower than the 25th percentile (anything less than $163.00); allowing me to disregard Mode as the “best” option to use instead of AOV.

1. **What is its value?**

**Median = $284.00**

**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 ShipperName, COUNT(\*) ‘Total\_Orders\_Shipped’

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?**

SELECT e.LastName, COUNT(\*) as Total\_Orders\_Placed

FROM Employees e

INNER JOIN orders o

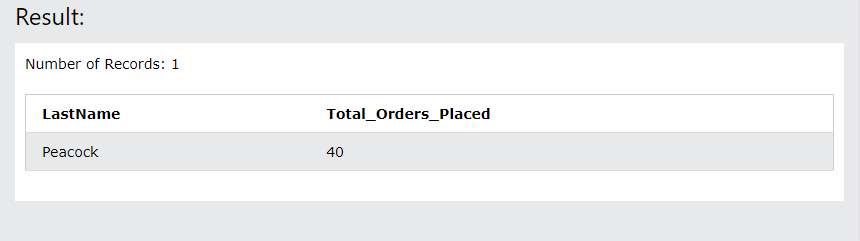
ON e.EmployeeID = o.EmployeeID

GROUP BY e.LastName

ORDER BY Total\_Orders\_Placed desc

LIMIT 1

;



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

SELECT p.ProductName, c.Country, SUM(od.Quantity) Total\_Quantity\_Sold

FROM Products p

LEFT JOIN OrderDetails od

ON p.ProductID = od.ProductID

LEFT JOIN Orders o

ON od.OrderID = o.OrderID

LEFT JOIN Customers c

ON o.CustomerID = c.CustomerID

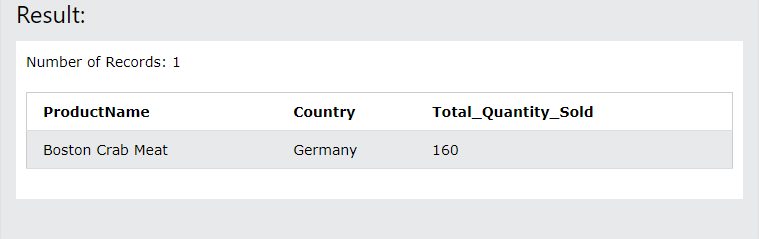
WHERE c.Country = "Germany"

GROUP BY p.ProductName

ORDER BY Total\_Quantity\_Sold desc

LIMIT 1

;

******