# Fall 2022 Data Science Intern Challenge

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.

1. **Think about what could be going wrong with our calculation. Think about a better way to evaluate this data.** 
   * 1. Using the mean as a measure of central tendency here is misleading because it is highly susceptible to the influence of outliers, and this dataset has some noteworthy outliers.
     2. First, shop\_id 42 has a mean order\_amount of $235,101.49. This is due to user\_id 607 placing 17 orders of $704,000 per order (2,000 items at a price of $352.00 per item).
        1. Aside from these 17 orders, the highest number of items sold in a single order is only 8 items.
     3. Second, shop\_id 78 is also worth calling out as an outlier because their shoe price is $25,725.00 - the highest in the dataset. Their mean order\_amount is $49,213.04.
        1. The second highest shoe price in the whole dataset is only $352.00, belonging to shop\_id 42.
     4. If we remove these outliers (orders for shops 42 and 78), the mean order\_amount of the remaining 98 shops is a much more reasonable $300.16, with an average shoe price of $150.40 and an average of 1.997 items sold per order.
        1. Note that this removes a total of 97 rows from the 5000 row dataset, or 1.9% of rows.
     5. Alternatively, we could just look at the median of order\_amount for the whole dataset which is $284.00. The advantage of considering the median from the beginning over the mean is that the median is far less susceptible to the influence of outliers, so we don’t need to do the additional work of removing outliers to arrive at a useful metric.
     6. The mode of order\_amount for the whole dataset is $153.00, followed closely by the second most frequent value for order\_amount: $306.00. This makes sense because $153.00 is also the most frequent individual shoe price among all our stores, and the most frequent values for total\_items (the quantity of items sold in an order) are 2 and 1. It follows that the most frequent order\_amount(s) would be 1 (item) \* $153.00 = $153.00 (order total) and 2 (items) \* $153.00 = $306.00 (order total).
     7. The standard deviation of order\_amount for the whole dataset is a staggering $41,282.54. The values for order\_amount in the original dataset have an excessive amount of variance.
     8. If we remove the outliers (records for shops 42 and 78) the standard deviation of order\_amount for the remaining 98 shops is a reasonable $155.94.
     9. Note that another way we could remove outliers is using the interquartile range (IQR) of order\_amount. In this case, the IQR is $227.00. The first quartile is $163.0 and the third quartile is $390.00. This means that all values less than -$177.50 and all values greater than $730.50 are outliers. If we use this method to remove outliers, we get a mean order\_amount of $293.72, a median order\_amount of $280, a mode order\_amount of $153.00, and a standard deviation of order\_amount of $144.45.
        1. I originally chose not to remove outliers this way because I thought this method assumed normality, and this dataset is heavily skewed to the right. After double-checking, I realize now that this method does not assume normality.
        2. This method removed 141 rows from our 5000 row dataset, or 2.8% of rows.
        3. Overall, while this IQR method is more robust in general, I think I prefer the earlier method of just removing the 2 shops that clearly behave differently from the other 98 for this scenario. I would need more domain knowledge to say with confidence which method I prefer.
2. **What metric would you report for this dataset?**
   * 1. If an immediate metric was necessary, I would provide the median and mode metrics for order\_amount. If I had an hour or so to analyze the data to understand its outliers, I would also provide the mean and standard deviation of order\_amount after removing outliers.
3. **What is its value?**
   * 1. **Please note that I am interpreting this question 2 ways:**
     2. **What is the numerical value of this metric?**
        1. The median order\_amount is $284.00.
        2. The mode of order\_amount is $153.00 and the second most frequent order\_amount is $306.00
        3. The mean order\_amount after removing outliers (records with shop\_id 42 or 78) is $300.16.
        4. The standard deviation of order\_amount after removing outliers (records with shop\_id 42 or 78) is $155.94.
     3. **Why is this metric valuable?**
        1. These measures of central tendency (mean, median, and mode) give us an idea of a central point around which the rest of our data are distributed.
        2. Measures of central variability, such as standard deviation, describe how spread out our data is from its mean.
        3. By combining these measures of central tendency and central variability, we can develop a pretty good intuition of where our data is distributed and how spread out it is.
4. Project Files
   1. Github - <https://github.com/semacki2/20220511_Shopify_Project>
      1. Please take a look at the shopify\_project.ipynb file. This is a python notebook detailing my work.
   2. Kaggle - <https://www.kaggle.com/code/sethmackie/shopify-internship-challenge?scriptVersionId=95625940>
      1. This is the same notebook, but you can run it in the browser if you have a kaggle account.

**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   COUNT(ShipperName)  FROM   Orders  INNER JOIN  Shippers  ON Orders.ShipperID = Shippers.ShipperID WHERE  ShipperName = 'Speedy Express' ; |
| --- |
| // 54 orders were shipped by Speedy Express. |

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

| SELECT  LastName FROM   Orders  INNER JOIN  Employees  ON Orders.EmployeeID = Employees.EmployeeID GROUP BY  Employees.EmployeeID ORDER BY   COUNT(Employees.EmployeeID) DESC LIMIT 1 ; |
| --- |
| // Peacock is the last name of the employee with the most orders. |

1. What product was ordered the most by customers in Germany?
   1. **Please note that I’m interpreting this question 2 ways:**
   2. What product had the most orders placed for it by customers in Germany?

| SELECT  ProductName FROM  Customers  INNER JOIN  Orders  ON Customers.CustomerID = Orders.CustomerID  INNER JOIN  OrderDetails  ON Orders.OrderID = OrderDetails.OrderID  INNER JOIN  Products  ON  OrderDetails.ProductID = Products.ProductID WHERE  Country = 'Germany' GROUP BY  ProductName ORDER BY  COUNT(ProductName) DESC LIMIT 1 ; |
| --- |
| // Gorgonzola Telino had the most orders placed for it by customers in Germany. |

* 1. Which product sold the most quantity to customers in Germany?

| SELECT  ProductName FROM  Customers  INNER JOIN  Orders  ON Customers.CustomerID = Orders.CustomerID  INNER JOIN  OrderDetails  ON Orders.OrderID = OrderDetails.OrderID  INNER JOIN  Products  ON  OrderDetails.ProductID = Products.ProductID WHERE  Country = 'Germany' GROUP BY  ProductName ORDER BY  SUM(Quantity) DESC LIMIT 1 ; |
| --- |
| // Boston Crab Meat sold the most quantity to customers in Germany. |