# Solution for Fall 2022 Data Science Intern Challenge

Shrinivas Sanglikar | <https://www.linkedin.com/in/ssanglikar>

**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.

I am using Jupyter Notebook and Python libraries to analyze this dataset.

Firstly, I checked the mean for the order\_amount. Then I checked for any skewness in that column. It was highly skewed.

I thought there might be some orders with large quantity of total\_items. So, I also checked the skewness for total\_items. It was also highly skewed.

Then I checked for min, max, percentiles and standard deviation for those two columns. The mean for total\_items was acceptable. But the mean for order\_amount is high (as mentioned). The max and standard deviation was also very high for these columns.

Upon checking the unique values for both these columns I observed that there were some outliers for order\_amount. Most orders were in between 90 and 1500.

So I filtered out all data points with order\_amount higher than 1500.

It filtered out 64 entries. The remaining dataset was not highly skewed and the outliers were mostly removed.

1. What metric would you report for this dataset?

To report a metric, I came up with an idea to take average order value per item for every order instead of directly taking mean for all order amounts. I called it Average order Amount Per Item (AAPI). After taking the mean of all the data points the average AAPI value was 302.58. Which is well below the 1500 order\_amount limit.

I also tested this metric without filtering out data and the value was 387.74. Which is still very low compared to previous AOV value.

1. What is its value?

302.58 for filtered data (removing data points where order\_amount > 1500),

387.74 for unfiltered data.

The .pdf file for the jupyter notebook is also attached.

**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?

**Answer: Shipper Name: Speedy Express ; Orders Shipped: 54**

**SELECT t2.ShipperName, COUNT(t1.OrderID) as OrdersShipped**

**FROM Orders t1**

**LEFT JOIN Shippers t2 ON t2.ShipperID = t1.ShipperID**

**WHERE t2.ShipperName IS "Speedy Express";**

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

**Answer: Last Name: Peacock ; Orders: 40**

**SELECT LastName, MAX(number\_of\_orders)**

**FROM (SELECT t1.LastName, COUNT(t2.OrderID) as number\_of\_orders**

**FROM Employees t1**

**JOIN Orders t2 ON t2.EmployeeID = t1.EmployeeID**

**GROUP BY t1.LastName);**

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

**Answer: product name: Boston Crab Meat Quantity: 160**

**SELECT productname, MAX(totalquantity)**

**FROM**

**(SELECT productname, SUM(Quantity) as totalquantity**

**FROM**

**(SELECT productname, quantity**

**FROM products**

**JOIN orderdetails**

**ON products.productid = orderdetails.productid**

**JOIN orders**

**ON orders.orderid = orderdetails.orderid**

**JOIN customers**

**ON customers.customerid = orders.customerid**

**WHERE country IS "Germany"**

**ORDER BY**

**productname)**

**GROUP BY productname**

**ORDER BY totalquantity DESC);**