# 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.
2. What metric would you report for this dataset?
3. What is its value?

**Answer:** a) AOV is calculated as Total Earnings or Revenue/ Total number of Orders and this metric provides us with the average cost of each order. i.e how much a customer will spend on average.

Upon analysis of dataset, I have noticed that the reason for AOV value being very large is that we have not considered quantity in our calculations. Most of the orders were in small quantity ranging from 1-8 items but there were a few orders which had a quantity of 2000 which in turn raises the cost of order and therefore, the average order value. So, AOV is not the right evaluation metric for this scenario. I have suggested better metrics in part b below.

b) There are two evaluation methods that would be suitable in this situation:-

1. Since each shop sells only 1 model of shoe, we can check for the shops where the Order Amount or quantity was relatively high as compared to others. It means that the model of sneakers is very popular and high in demand. We can make use of a sales metric called ‘Product Performance’ here. It is calculated by ranking top few (3, 10, 25) performers by sales revenue.

OR

1. We can make use of ‘Average Purchase value’ which is calculated as Total Sales/ Number of sales/ transactions. This metric returns the average sales value of each sales transaction. So, while calculating number of sales/transactions, we will consider quantity rather than number of orders.

c) **By using MS Excel:**

**Product Performance**

Sort the Column order\_amount from highest to lowest and then expand the selection.

Highest Revenue = 704000 by shop\_id = 42

Second Highest Revenue = 154350 by shop\_id = 78

Third Highest Revenue = 102900 by shop\_id = 78

**Average Purchase value Calculation**

= SUM(order\_amount)/SUM(total\_items)

=SUM(D2:D5001)/SUM(E2:E5001)

=357.92

**By using Python:**

**Product Performance Code:**

Import pandas as pd

df = pd.read\_csv(input\_path) #Input Path is where data set is located

resultant = df.sort\_values(by="order\_amount", ascending=False) #Sorting the data set in descending order

resultant = resultant[['shop\_id', 'order\_amount']]

print(f"{resultant.head(25)}") #This will fetch top 25 entries with highest revenue and corresponding shop\_id

**Average Purchase Value Calculation Code:**

import pandas as pd

df = pd.read\_csv(input\_path)

avg\_purchase\_value = df.order\_amount.sum() / df.total\_items.sum()

print(f"{avg\_purchase\_value}")

**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?
2. What is the last name of the employee with the most orders?
3. What product was ordered the most by customers in Germany?

**Answer:**

1. SELECT COUNT(OrderID) FROM Orders WHERE ShipperID=(SELECT ShipperID FROM Shippers WHERE ShipperName = 'Speedy Express');

Output : 54 orders

1. SELECT LastName from Employees where EmployeeID = (Select EmployeeID from Orders where OrderID = (SELECT OrderID from OrderDetails GROUP BY OrderID ORDER BY SUM(Quantity) DESC LIMIT 1))

O/p: Dodsworth

1. Select ProductName from Products where ProductID = (Select ProductID from OrderDetails where OrderID IN (Select OrderID from Orders where CustomerID IN (Select CustomerID from Customers where Country = 'Germany')) ORDER BY Quantity DESC LIMIT 1)

O/P: Steeleye Stout