# 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!

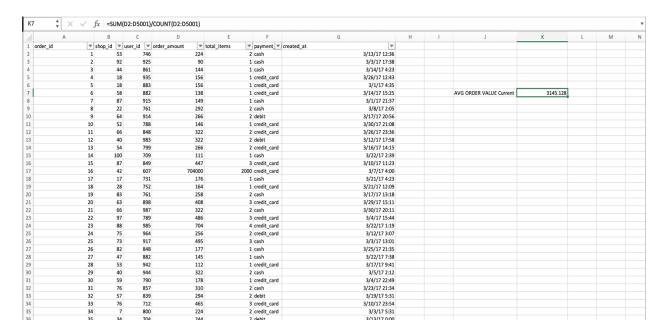
**Question 1:** Given some sample data, write a program to answer the following: <u>click here to access the required data set</u>

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.

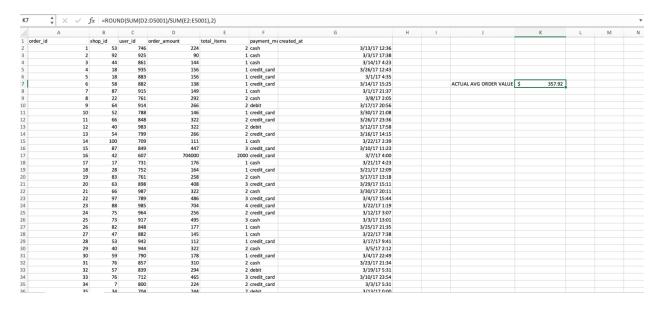
- a. Think about what could be going wrong with our calculation. Think about a better way to evaluate this data.
- b. What metric would you report for this dataset?
- c. What is its value?

#### Answer

a) On observing the data given in the file, the existing formula to calculate the **Average**Order Value is based on the sum of all order\_amount divided by total number of
orders i.e 5000. This formula fails to take the number of items in consideration that are
associated with each order ID and hence we get a bizarre value average order value of
\$3145.13



b) The rectify the above, the correct **average order value** must be calculated as sum(order\_amount)/sum(total\_items)



c) The average order value for the month of March'17 is \$357.92

**Question 2:** For this question you'll need to use SQL. Follow this link 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.

- a. How many orders were shipped by Speedy Express in total?
- b. What is the last name of the employee with the most orders?
- c. What product was ordered the most by customers in Germany?

#### Answer

a) Here I have joined the shippers and orders tables on the common key shipper id.I extracted the shipper's name and took the count of unique order IDs.

#### Number of Records: 1

ShipperName	Packages Shipped
Speedy Express	54

b) Joined orders and employees table on employee id and extracted the last name of the person with the most number of orders placed.

#### Number of Records: 1

lastname	Total Orders Placed
Peacock	40

c) For the most ordered product in Germany , I came up with 2 solutions . The first one is based on the quantity of the product , the second one is based on the number of times a product was ordered.

## Solution 1 (Most Orders Based on Quantity)

```
SELECT c.country
                        AS OrderCountry,
       od.productid,
      p.productname,
       Sum(od.quantity) AS Total_Quantity
      orderdetails od
FROM
       JOIN orders o
         ON od.orderid = o.orderid
       JOIN customers c
         ON o.customerid = c.customerid
       JOIN products p
         ON od.productid = p.productid
WHERE country = 'Germany'
GROUP BY 1,
ORDER BY total quantity DESC
LIMIT
```

#### Number of Records: 1

OrderCountry	ProductID	ProductName	Total_Quantity
Germany	40	Boston Crab Meat	160

# Solution 2 (Based on Number of Orders)

```
SELECT od productid,
      p.productname,
      Count (od. productid) AS Times Ordered,
      Sum(od.quantity) AS Total_Order_Quantity
      orderdetails od
FROM
      JOIN orders o
         ON od orderid = o orderid
       JOIN customers c
         ON o.customerid = c.customerid
       JOIN products p
         ON od.productid = p.productid
WHERE country = 'Germany'
GROUP BY od productid,
         p.productname
      BY times ordered DESC,
ORDER
         total order quantity DESC
       2
LIMIT
```

### Number of Records: 2

ProductID	ProductName	Times_Ordered	Total_Order_Quantity	
31	Gorgonzola Telino	5	125	
40	Boston Crab Meat	4	160	