Fall 2022 Data Science Intern Challenge

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

First, by skimming through the data, we can see that there is a particular store with shop_id 42 that has values that are great outliers – repeated purchases from user_id 607 of 2,000 items worth a total value of 704,000. These skew the results, but doing some basic division, we can equate the unique product price to 352 which is expensive for a pair of shoes but is accurate. This may be a bulk order for a subsidiary shoe seller as it is repeated so we should ask the store or our manager if we should keep the data in our exploration.

Second, we see all sales of shop_id 78 with an item priced at 25,725. This is an absurd price tag and makes us question if a decimal was missed as it is causing a big skew in our results. We must ask our manager if we continue with this data or consult the store to verify the prices of their products.

b. What metric would you report for this dataset?
c. What is its value?

We can choose multiple solutions.

Option A: One solution is to average the values, deleting all values from shop_id 42 that are bulk shipments of 2,000 items. Furthermore, we can also delete all values from shop_id 78 as they significantly skew the data.

Average(D:D) after deleting aforementioned unrelated or badly formatted data.
Average sale is 302.58

Option B: The second solution is to batch sales by the price per item in the sale. We will calculate order_amount divided by total_items. This will give us the average price of a product in an order.

By averaging the new values, we get an updated value of 387.74

If we remove the 46 entries of product price 25,725, this number becomes 152.48 average price per product per order.

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?

We need to find the Shipper ID for Speedy Express

```
SELECT *
FROM [Shippers]
```

Found Shipper ID = 1

Then we have to determine count of orders processed, grouped by shippers

```
SELECT Count(Distinct OrderID) as total_orders, ShipperID
FROM [Orders]
Group by ShipperID
Order by total_orders DESC
```

Now we can query directly for SpeedyExpress with a where clause.

```
SELECT Count(Distinct OrderID) as total_orders, ShipperID
FROM [Orders]
Where ShipperID = 1
Group by ShipperID
Order by total_orders DESC
```

Speedy Express (ShipperID 1) processed 54 orders.

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

First we need to retrieve the number of total orders per employee.

```
SELECT COUNT(OrderID) as total_orders, EmployeeID
FROM [Orders]
GROUP BY EmployeeID
Order BY total_orders DESC
```

Then we can look at the Employees table to determine the last name of the employee with ID number 4.

```
SELECT LastName
FROM [Employees]
Where EmployeeID = 4
```

The last name of the employee with the most orders is Peacock

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

First we must determine CustomerID of all customers ordering to Germany

```
SELECT Distinct CustomerID
FROM [Customers]
Where Country = 'Germany'
```

We get CustomerIDs: 1, 6, 17, 25, 39, 44, 52, 56, 63, 79, 86

Then we must see what these distinct Germany based customers ordered, by looking for order numbers.

```
SELECT *
FROM [Orders]
Where CustomerID in (1, 6, 17, 25, 39, 44, 52, 56, 63, 79, 86)
```

We get OrderIDs: 10267, 10273, 10277, 10279, 10284, 10285, 10286, 10301, 10312, 10313, 10323, 10325, 10337, 10342, 10343, 10345, 10348, 10356, 10361, 10363, 10391, 10396, 10407, 10418, 10438

We can now query to find what ProductIDs were ordered and how much was ordered in total.

```
SELECT ProductID, SUM(Quantity) as total_ordered
FROM [OrderDetails]
Where OrderID in (10267, 10273, 10277, 10279, 10284, 10285, 10286, 10301, 10312, 10313, 10323, 10325, 10337, 10342, 10343, 10345, 10348, 10356, 10361, 10363, 10391, 10396, 10407, 10418, 10438)
Group by ProductID
Order by total_ordered DESC
```

We see that the product ordered the most in Germany was ProductID 40 with 160 orders.

```
SELECT ProductName
FROM Products
WHERE ProductID = 40
```

Boston Crab Meat was the most ordered product.