# QUESTION 1

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

I think the way AOV is calculated by summing order amount of all orders and dividing it by number of orders. But there are some orders that has more than one sneaker so right way to calculate average price for a sneaker is sum(order\_amount)/sum(total\_items). Even with this approach average price comes to be 357.94 which is relatively high for sneakers.

I tried normalizing the dataset by finding average price of a sneaker in an order. Even after normalizing the data the average price is 357.94 which is high. This is because there are 46 data-points where average price of a sneaker is above 25000. These are extreme points (outliers) that is inflating the average price of a sneaker and making it unaffordable. To treat outliers, we can do following things:

1. Option 1 - Remove outlier data points from the data set and calculate average sale price using the dataset without these outlier data points and then calculate Mean

2. Option 2 - We replace outlier data points with some threshold value. Threshold value can be decided by deciding the capped value. For example, in this dataset we can replace all shoes greater than 1000 with 1000 or replace outlier with the median value of the dataset which is 153 and then calculate Mean.

**Question 2- What metric would you report for this dataset?**

Since the dataset has outliers, mean is not the right representative value of the dataset. Mean is inflated because of outliers and not giving right representation of the dataset. We can use median to give the representation value for the data set. And Median won’t be impacted by outliers because median is calculated by ordering the dataset and finding middle value of the dataset. To calculate Median, I have first normalized the data set that is added a column which is avg\_price\_item which is calculated using the formula total\_amount/total\_items for every record. Find the 50th Percentile value of the dataset which is coming as 153. 153 looks reasonable for a sneaker.

**Question 3- What is its value?**

Along with median we can also calculate min, max, 25th Percentile, 50th Percentile, 75th Percentile and 90th Percentile value of the dataset. These values gives good idea about the dataset:

|  |  |
| --- | --- |
| Mean | 90 |
| 25th Percentile | 133 |
| Median | 153 |
| 75th Percentile | 169 |
| Max | 25725 |

# QUESTION 2

**How many orders were shipped by Speedy Express in total?**

Inner query finds shipper id of the Speedy Express. Outer query sees what all orders are using the shipper id and returns number of items

SELECT *Count*(\*)as num\_items  
FROM   orders  
WHERE  shipperid IN (SELECT shipperid  
                     FROM   [shippers]  
                     WHERE  shippername = 'Speedy Express')

**What is the last name of the employee with the most orders?**

Inner query finds the employee id of the employee who has maximum number of orders on his/her name by grouping orders on employee and ordering the dataset in descending order on number of orders. Outer query finds employee last name for the employee id.

SELECT lastname  
FROM   employees  
WHERE  employeeid IN (SELECT employeeid  
                      FROM   orders  
                      GROUP  BY employeeid  
                      ORDER  BY **Count**(\*) DESC  
                      LIMIT  1)

**What product was ordered the most by customers in Germany?**

First, I am finding customers residing in Germany in the innermost query.

SELECT customerid  
FROM   [customers]  
WHERE  country = 'Germany'

Then I am finding orders placed by those customers:

SELECT orderid  
FROM   orders  
WHERE  customerid IN (SELECT customerid  
                      FROM   [customers]  
                      WHERE  country = 'Germany')

Then I am product that was placed maximum time in those orders by using group by on product id and ordering the result in descending order on count(\*):

SELECT productid  
FROM   orderdetails  
WHERE  orderid IN (SELECT orderid  
                   FROM   orders  
                   WHERE  customerid IN (SELECT customerid  
                                         FROM   [customers]  
                                         WHERE  country = 'Germany'))  
GROUP  BY productid  
ORDER  BY **Count**(productid) DESC  
LIMIT  1

Then I am finally finding Product name of that product.

Here is the final query

SELECT productname  
FROM   products  
WHERE  productid IN (SELECT productid  
                     FROM   orderdetails  
                     WHERE  orderid IN (SELECT orderid  
                                        FROM   orders  
                                        WHERE  customerid IN (SELECT customerid  
                                                              FROM   [customers]  
                                                              WHERE  
                                               country = 'Germany'))  
                     GROUP  BY productid  
                     ORDER  BY **Count**(\*) DESC  
                     LIMIT  1)