# Shopify Technical Challenge – Data Science

## Problem 1

The Sneaker Problem

The dataset provided for the challenge is a very intriguing one. We know that a total of 100 stores exist, and every store sells only one model of shoe. The problem, however, states the miscalculation of the AOV (Average Order Value). When calculated based on the dataset, it states that it comes to $3145.13. This basically means that an average cost of a shoe we sell is $3145.13. This is obviously wrong, and some issues persist with the dataset at hand. To move forward with this, I decided to first check and verify the same.

Throughout the problem I have made use of R, and I am also well versed with SAS because of my time working there.

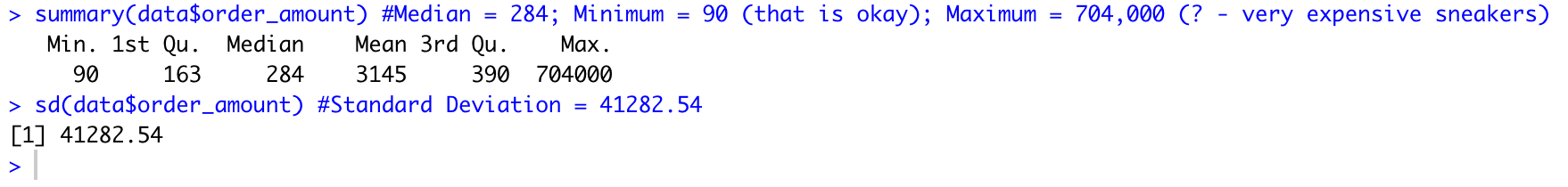
I have attached the R-Script under the same GitHub Repository and would really request you to see it to better understand the workflow which I might not be able to explain through this report.

Firstly, I imported the dataset and make sure that there are no missing values. Following is the screenshot that states all ‘FALSE’ values in all the columns. This is an output of the function ‘is.na()’ that queries the dataset and checks for missing values.   
  
All ‘FALSE’ means there are no missing values in the dataset.

Graphical user interface, text

Description automatically generated with medium confidence

Now, I checked if the AOV is indeed 3145, and to my surprise, it was! Also, the standard deviation is 41282.54.



Now, answering the questions.

### What could be going wrong with our calculation? What is a better way to evaluate this data?

To be able to determine this, we first need to just revisit the summary of the given data. It shows that the maximum value for ‘order\_amount’ in our dataset is $704,000, whereas the lowest is $90 (refer above image). What we need to determine is why is this happening. This would mean that either there are outliers or there is a data entry error.

To do this, I plotted all the orders over days against the order values. The color coding was based on the ‘shop\_id’ just to see if there is anything common.

Following is the output I got:

Graphical user interface, text, application, email

Description automatically generated

Chart, scatter chart

Description automatically generated

The denser color schemes are all the same color and are higher (the $704,000!) and belong to shop with the ID 42. The other outliers come from shop with the ID 42.

To narrow down the search, I created separate data frames for both the shops and evaluated them. The following are the outcomes of both the shops:

Table

Description automatically generatedTable

Description automatically generated

As we can see, Shop 42 has minimum order amount as 352 but maximum as 704,000. Shop 78 has minimum order\_amount as 25725 and maximum as 154,350. Also, the total number of items per order in shop 42 is 1 at the least and 2000 at the most. Secondly, in shop 78, it is 1 at the least and 6 at the most.

To understand it further, I tried to find the mode by writing a function myself (R does not have an in-built function for mode) to determine the most common orders.

Graphical user interface, text, application

Description automatically generated

So, shop 42 mostly has orders with 2000 orders and the amount they get is 704,000$. By simply doing the math, we know that every pair of sneakers at shop 42 costs about $352. On the other hand, the maximum number of items and the maximum amount, when computed, is 6 and 154,350$ in shop 78. This means that they sell extremely expensive sneakers, each worth about $25,725!

In conclusion, the problem with the data is the skew that is caused by the shop 78. Their shoes are priced way higher than all other stores and it cannot be computed with the others and will lead to a major skew always. The data is not normally distributed.

### What metric would you report for this dataset?

AOV (Average Order Value) is a very interesting measure but not suitable when the data is not distributed normally. As we saw, the immense skew in the dataset is leading to extremely wrong assumptions on our part. This is not good to any business analytics and thus needs to be eliminated. To do this, following are a few recommendations:

* Data should be classified based on per unit costs.
* Containers should be created for the shoes in the same price range to be able to accommodate better analytics.
* The more expensive shoes should be considered as a Silo and the development should be done according to that.
* Thresholds must be set before analysis. Though outliers, the data is still important and must be computed.
* If we still want to run an analysis and compute the AOV using this data solely, then three different AOVs need to be calculated based on thresholds.
* A common AOV is not going to be helpful.

### What is its value?

I would recommend calculating the AOV in three segments for now.

1. Premium Sneakers (AOV for Shop 78)
2. Luxury Sneakers (AOV for Shop 42)
3. Casual Sneakers (AOV for the rest of the shops)

The following is the code to do that:

Graphical user interface, text, application

Description automatically generated

Thus, as we can see, the skew gets eliminated and we get 3 separate AOVs. Now, it makes a lot more sense.

## SQL Challenge

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

Graphical user interface, text, application

Description automatically generated

As we can see, Speedy Express shipped a total of 54 orders.

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

Graphical user interface, application

Description automatically generated

As we can see, the Last Name of the employee with the most orders is Peacock.

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

Graphical user interface, application, Teams

Description automatically generated

As we can see, Steeleye Stout was ordered the most (100 times).