Spotify Summer 2022 Data Science Intern Challenge

My first thought would to be how exactly average order value is calculated and what its used for. How could a mistake be made and what effect outliers have on the data, is the data complete? Is there a mistake effecting the calculation, possibly by a vendor mistake? Or possibly an extravagant money laundering scheme? Is there enough data for the AOV to give a reasonable result?

Since I have no knowledge of e-commerce lingo a quick google search yields:

$$\text{Average Order Value (AOV)} = \frac{\text{Revenue}}{\text{Number of Orders}}$$

Revenue here corresponds to the sum of order_amount column
Number of orders here corresponds to the number of rows

A quick look at the data shows that order 161 there was one purchase made of 1 item costing $25725
Order 512 has 2 items costing $51450
Going further down I see shop 78 is selling shoes made of solid gold and managing to sell pairs for $25725 each. Money laundering scheme?

Sorting the data by how many of these shoes were sold at first glance it doesn't look like much to alter the AOV so drastically, but computing them myself I got the same number.

Since most shops aren't selling shoes worth 25k, I'd say a more important metric would be to drop this store and focus on the smaller more affordable shops, if that's what we are interested in analyzing. Considering they are far more common, dropping shop 78 would give us a more focused look on the average shops AOV

Dropping 78 still gives a high AOV, so I'm seeing that the shop 42 seems to be selling 2000 shoes with an order_amount of 704000 at the same time everyday by the same user_id. Suspicious but I can write this off as warehouse purchases for a smaller in-person shop, or possibly a rich someone who just really loves this shoe and wants to wear them for the rest of their life.

Dropping 42 gives a reasonable AOV of ~$300. I'd say this is more valuable as it gives insight into the 98 more typical shops AOV and ignores the high outliers of

only 2 shops.

I’ve attached my rough work in Python using pandas, enjoy!