**Documentation of E-Commerce Project**

**(SPREADSHEETS,PYTHON,R)**

* *Ask & Prepare Phase*

In this phase, I first searched for a relevant database considering today's market. I stumbled upon this one, an interesting e-commerce transaction dataset on Kaggle. After I found it, I decided to imagine a business scenario where I might use this data.

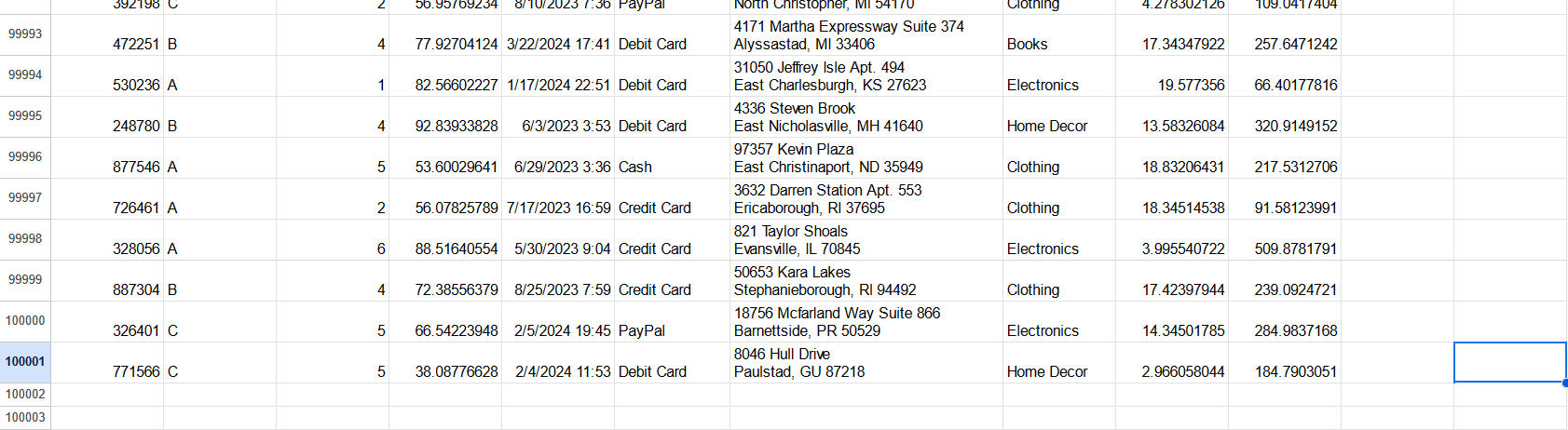
The scenario I imagined is, working as a data analyst I can unlock insights into consumer behavior and retail dynamics from this dataset. Shareholders/managers can use this data to make informed business decisions based on factual data.

This dataset has some key variables such as Order ID, Product ID, Quantity, Price, Customer ID, Location, Demographics that can provide actionable insights.

* *Data Cleaning*

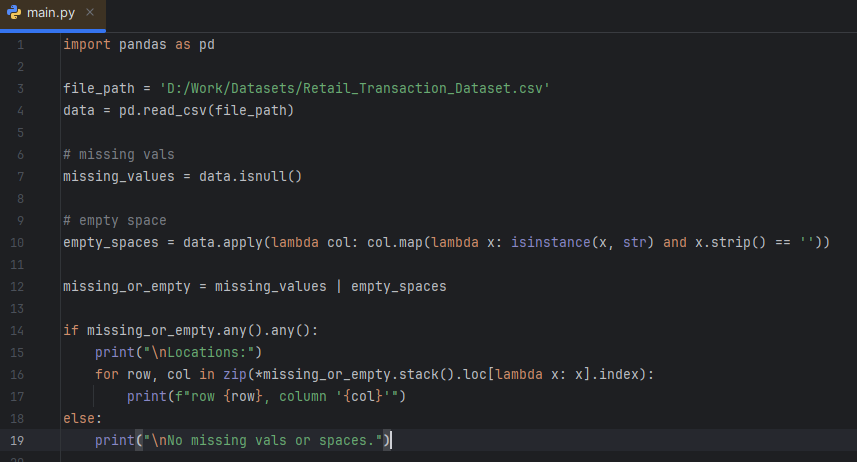
As this dataset is very large, I will be analyzing it with spreadsheets and modifying the incorrect data with python to be efficient and fast.

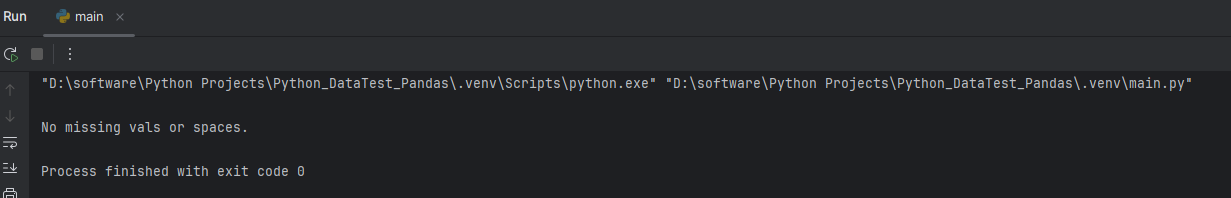
Checking for missing/ empty values with spreadsheets, we can find “missing” values on rows, but after further investigation this is just because the program inserted 2 more lines at the end.

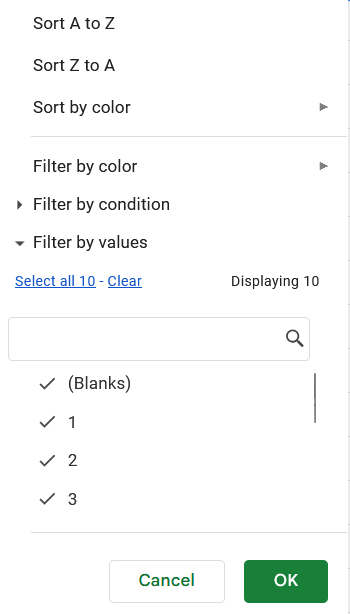
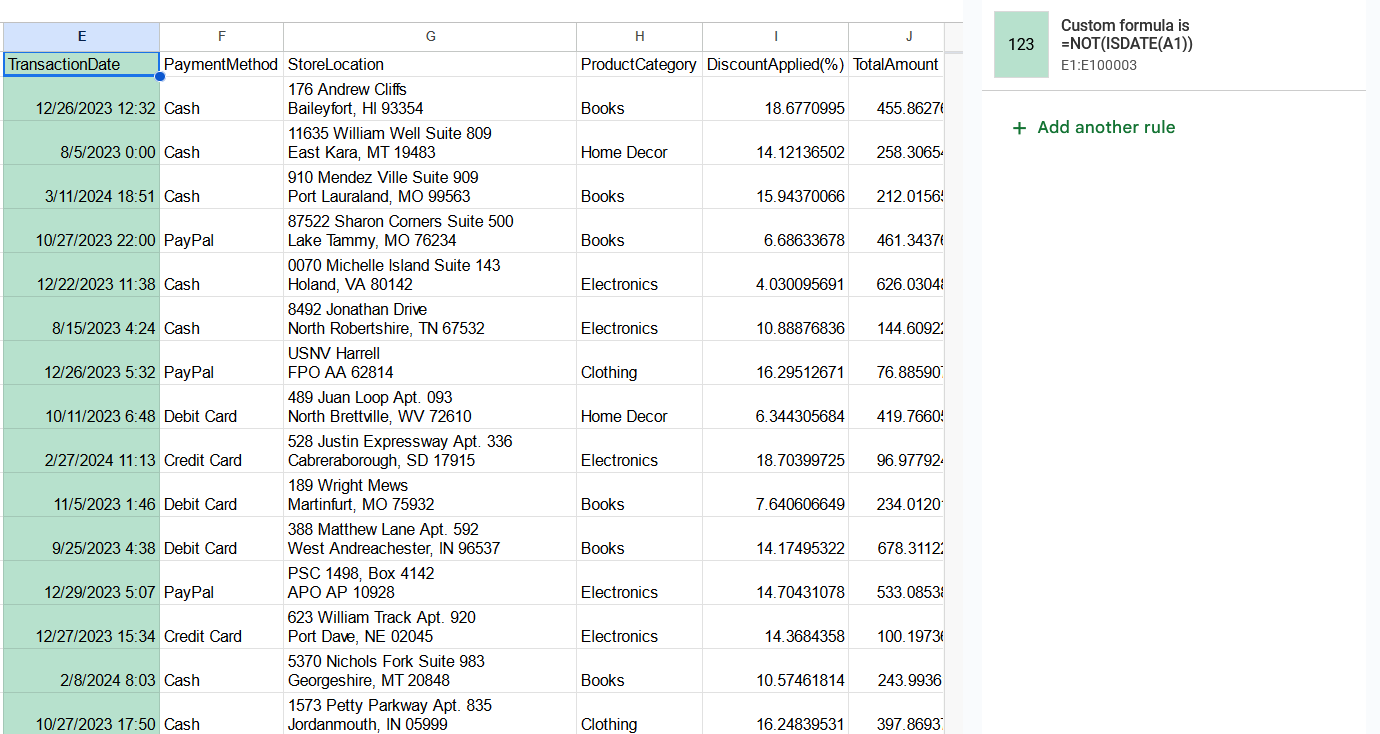


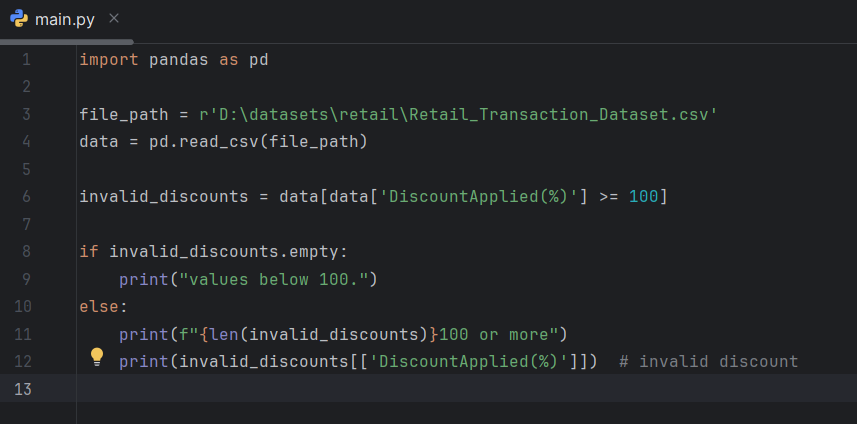
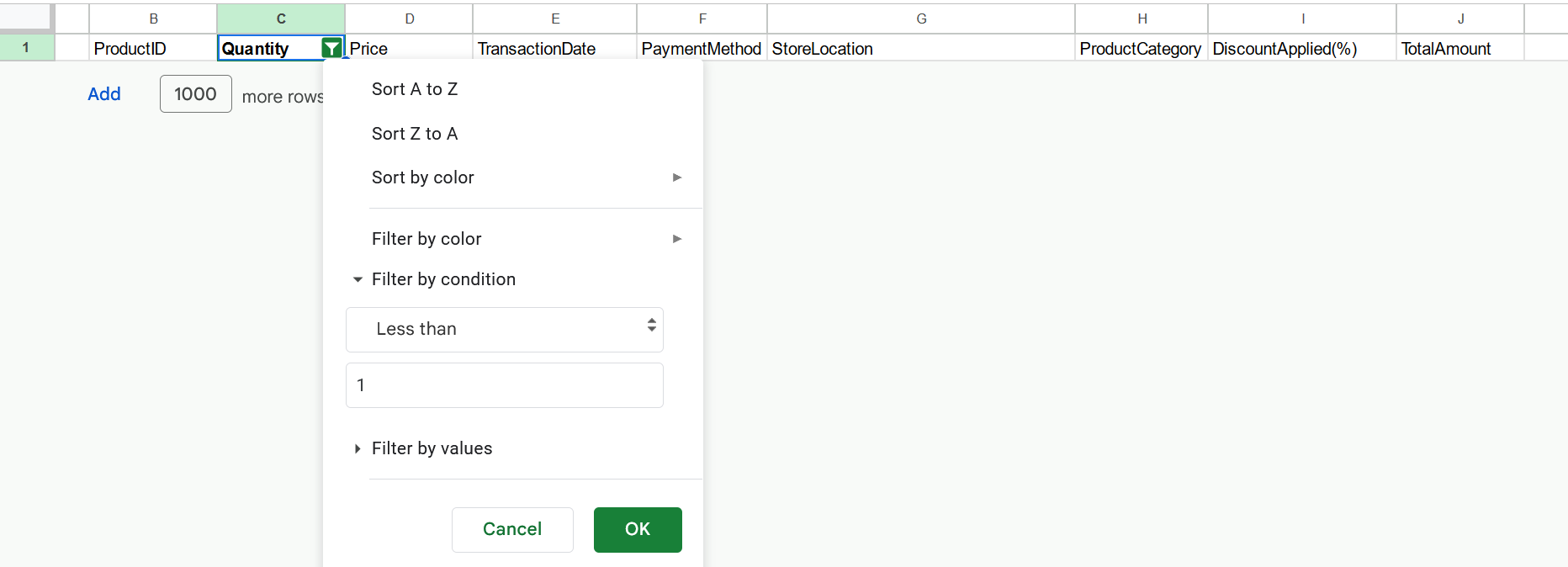
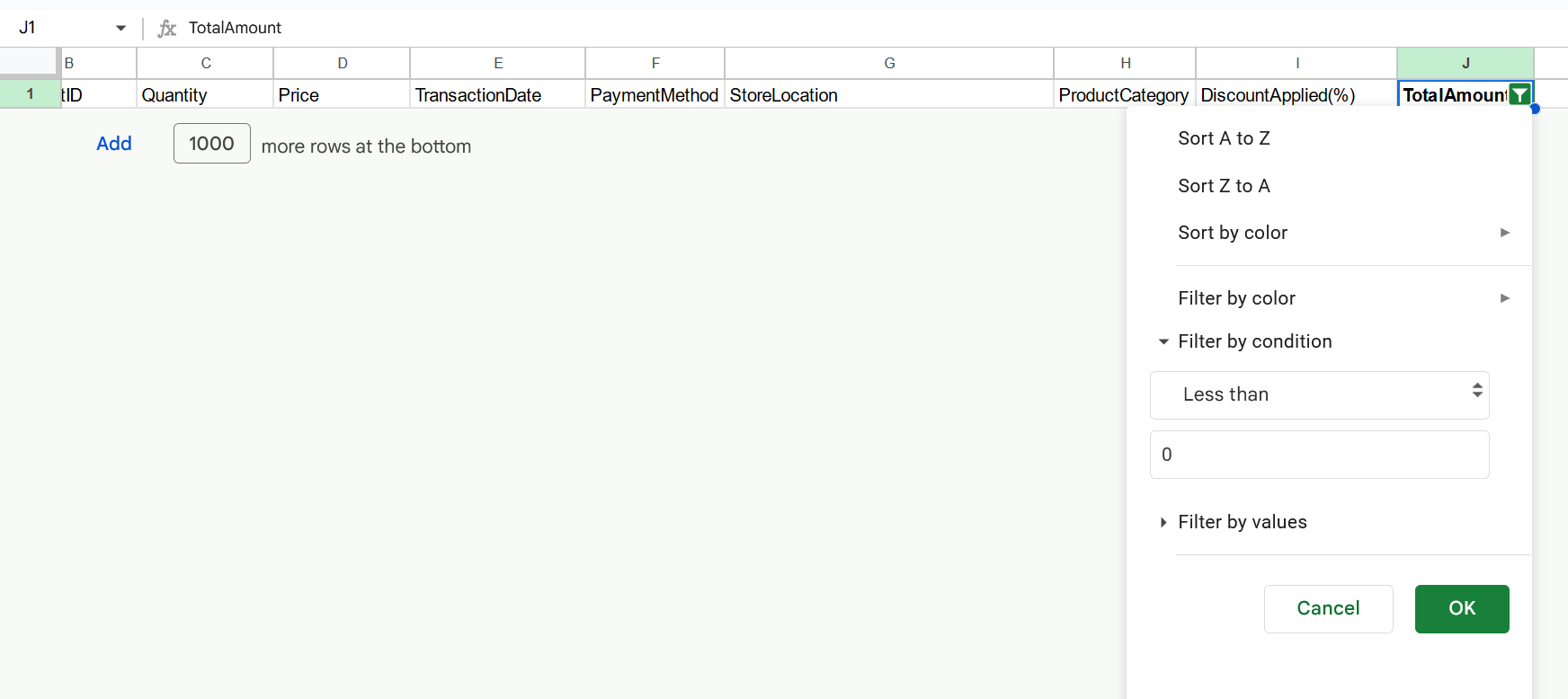
Deleting them, we get a simple and easy fix.

Using this simple python script I searched it and found no missing values here, it must have been a spreadsheets error.





I decided to check for any wrong values, but there weren’t any. I used filters For the columns that had limited values, and tools like conditional formatting or formulas for the more complex ones(like dates):

Now, using filters and python I checked for any outliers, any values that shouldn’t be possible/are wrong. For example, I checked the Discount to not be above 100, the total amount to not be negative etc.

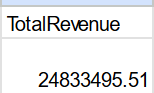
There don’t seem to be any outliers.

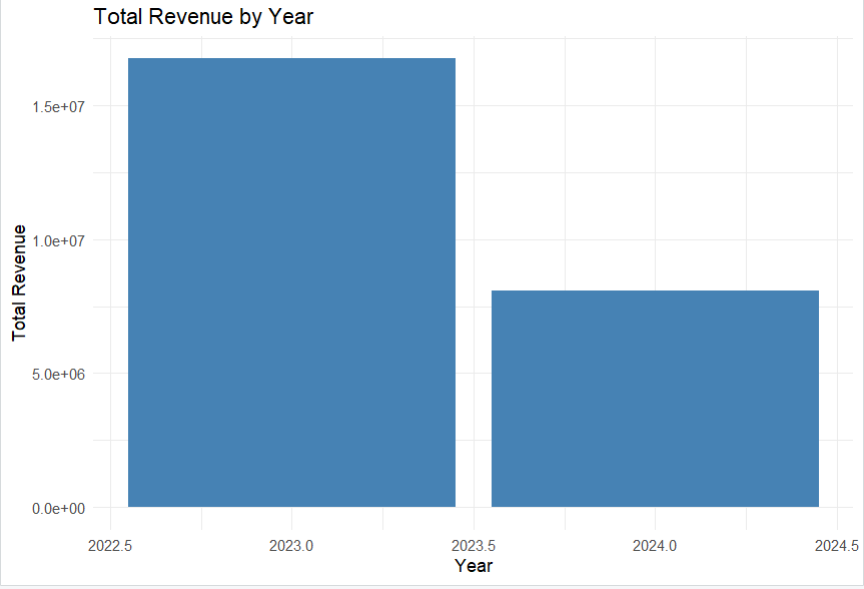
I also checked for data standardization, checked if all values were measured to the same scales, checked for NULLS and everything seems to be correct.

I know that using so much python was unnecessary but I want to improve my skills and also appear more favorable as a candidate, my first project has a dataset that had more things that needed to be fixed :).

* *Data Analysis*

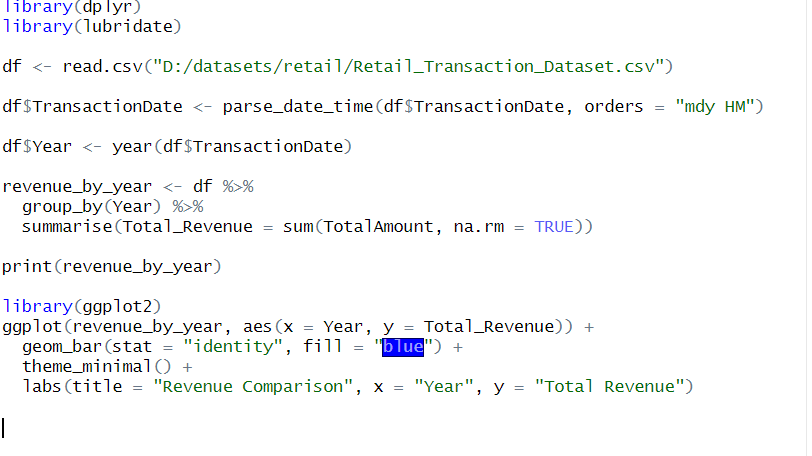
I’ll start with a simple descriptive analysis, so that I can understand the code better.



This is a simple sum of all revenue(sum function in spreadsheets), but it doesn’t tell much. Let’s compare the 2023 to 2024 revenue.

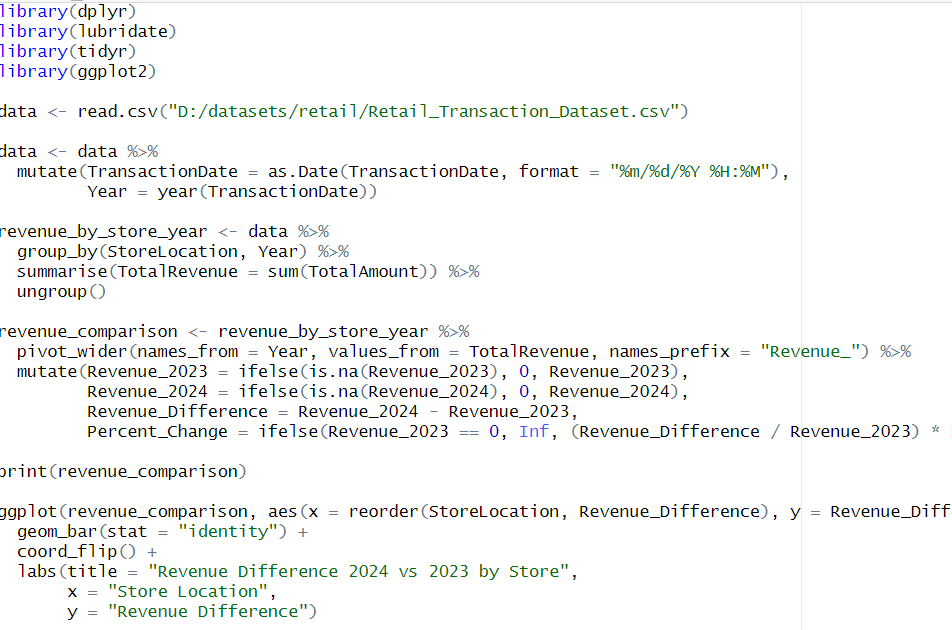
This chart tells a completely

Different story, while the initial revenue seems great, the chart shows us that the revenue in 2024 compared to 2023 declined a lot!



I found this difference using R(Rstudio), here is the code:

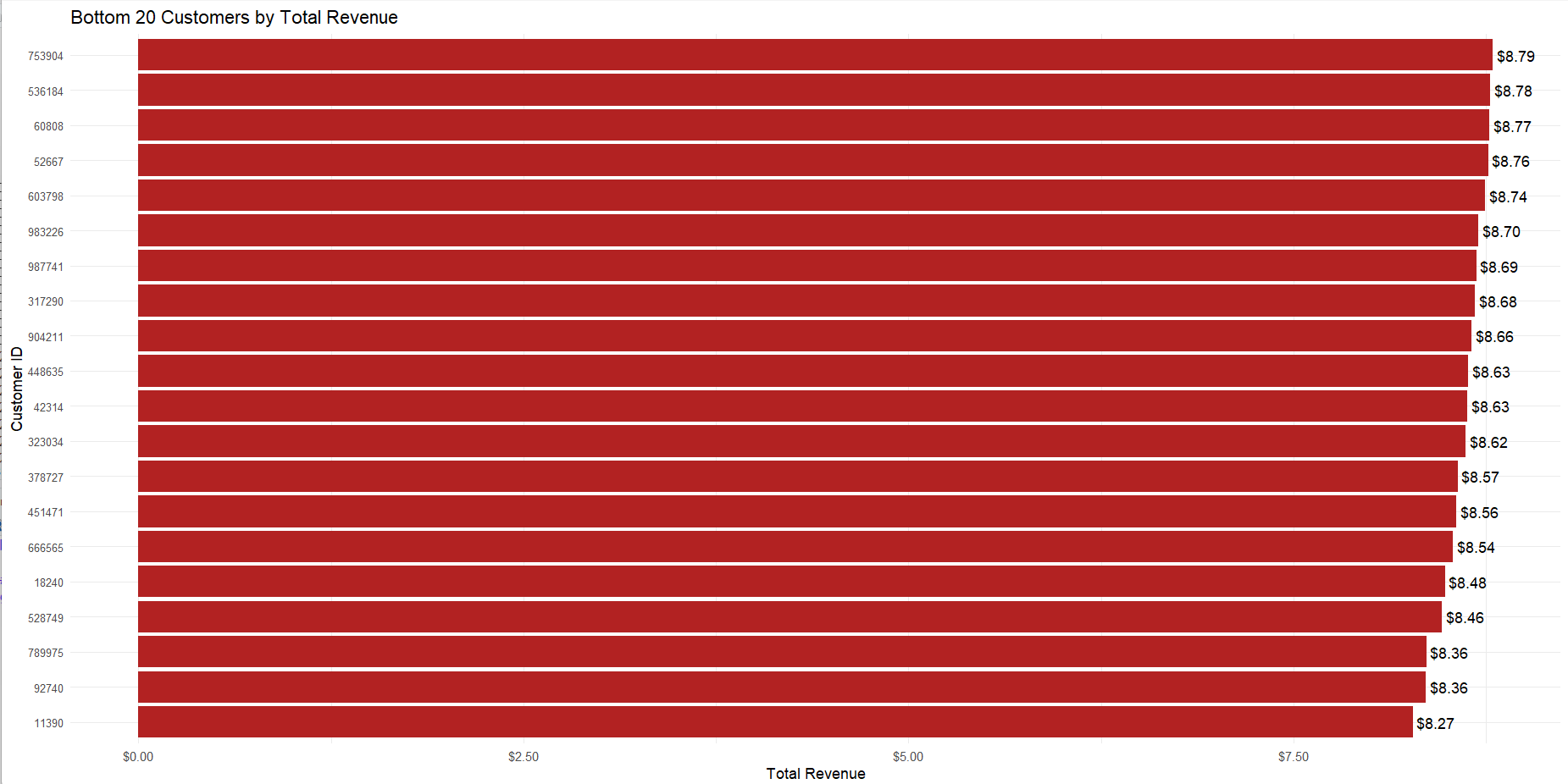
Seeing this disparity, I decided to look further into the problem to find one or more causes that might cause this problem.

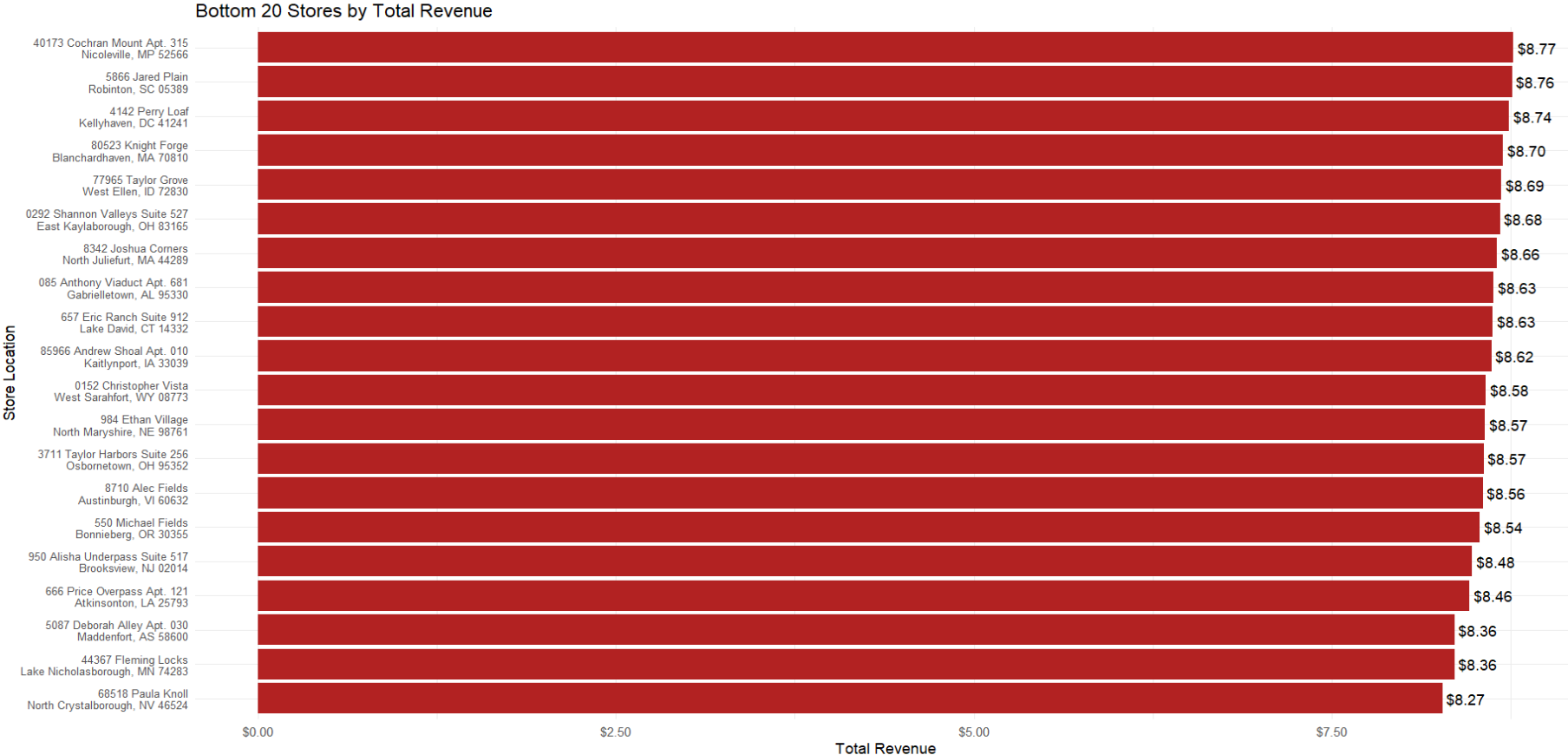
After looking at store locations, nothing stands out. The stores barely repeat, I tried running a script in R comparing the store revenue(2023v2024) but nothing stood out as there are way too many stores and most of them don’t even repeat.

R script:

Therefore, we can conclude that this is probably a mistake in the dataset, it’s probably supposed to be the customer's address! Not the store location.

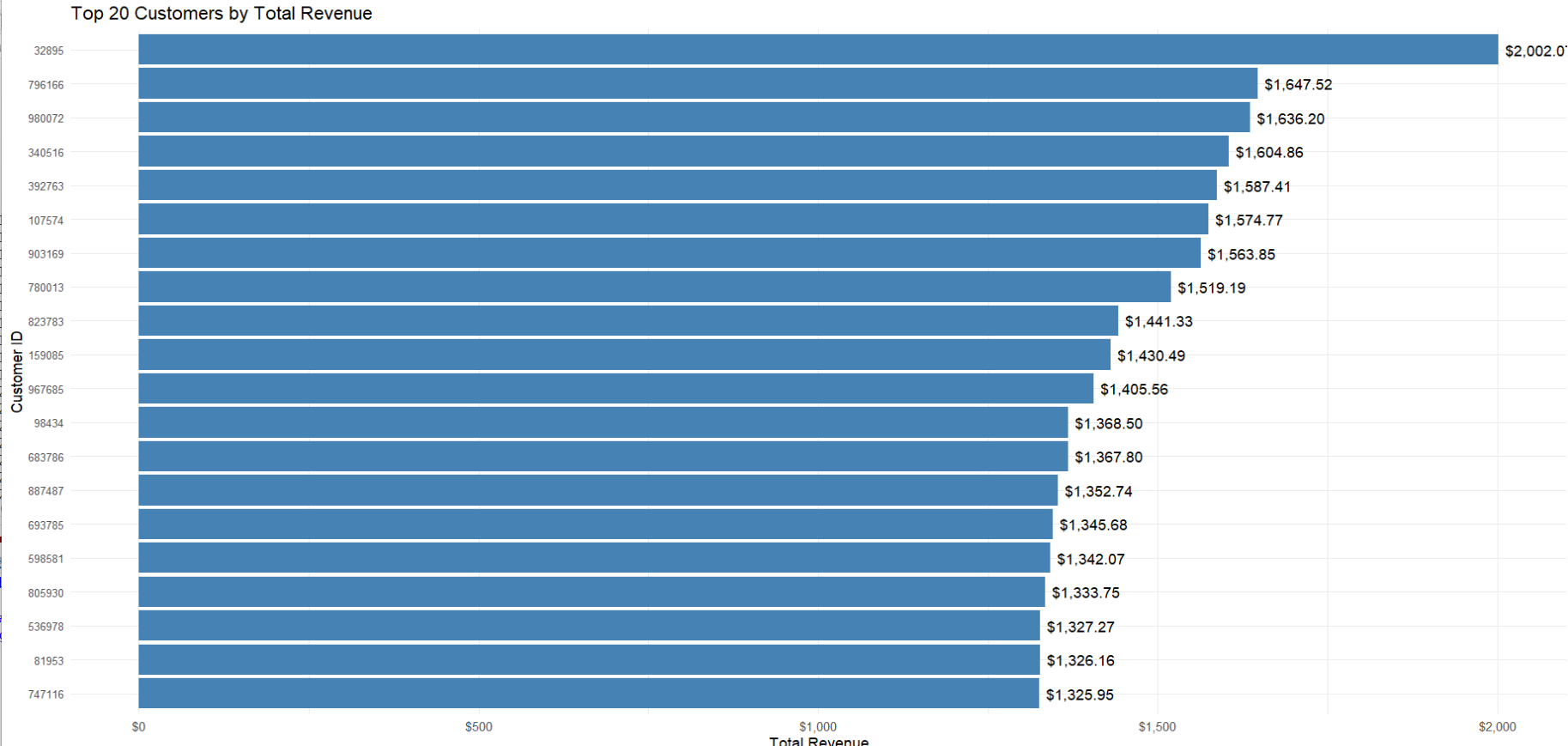
Now, testing this theory out by comparing the highest spending users versus the lowest we can see a correlation.

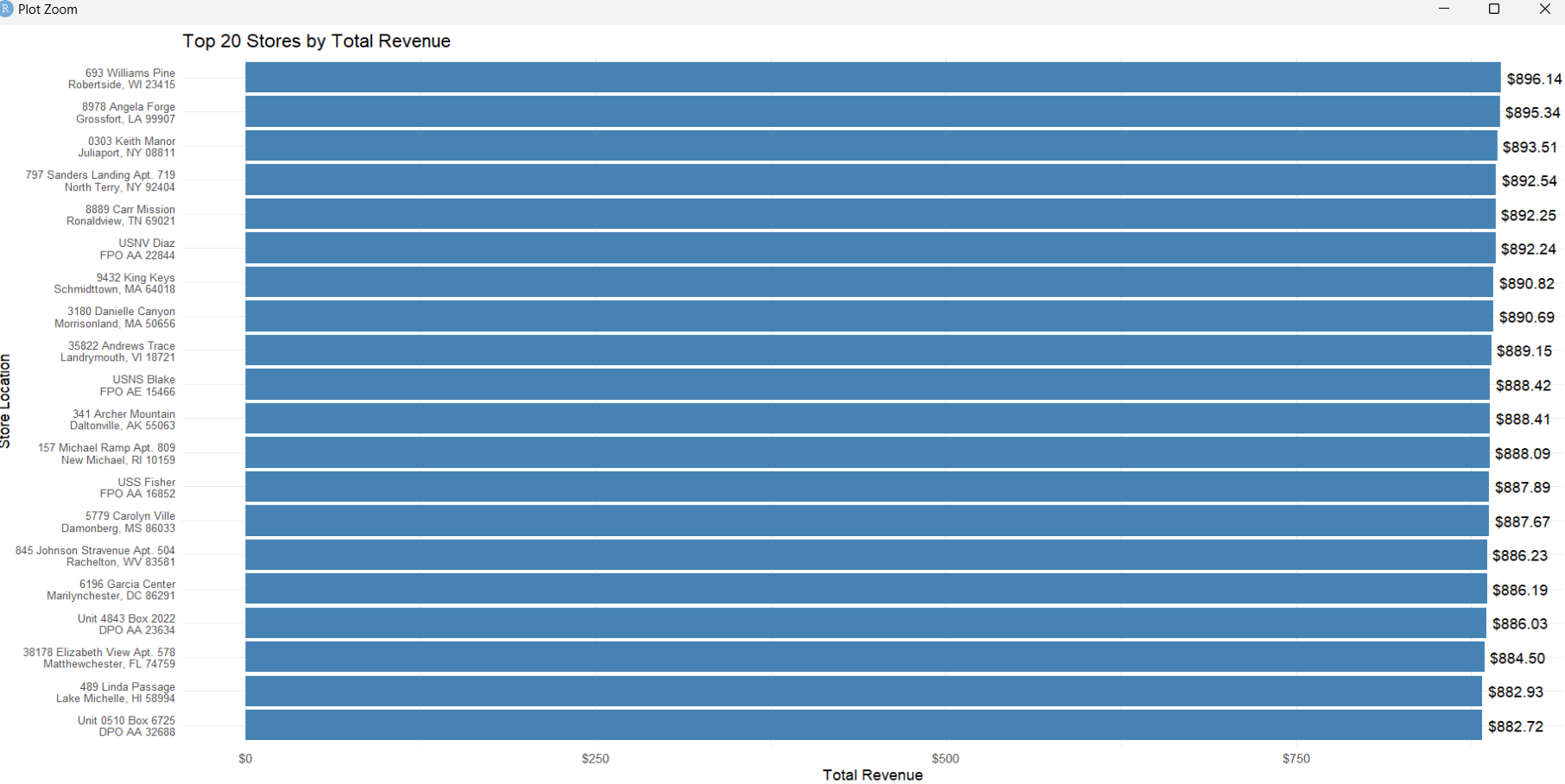




The results are almost identical!

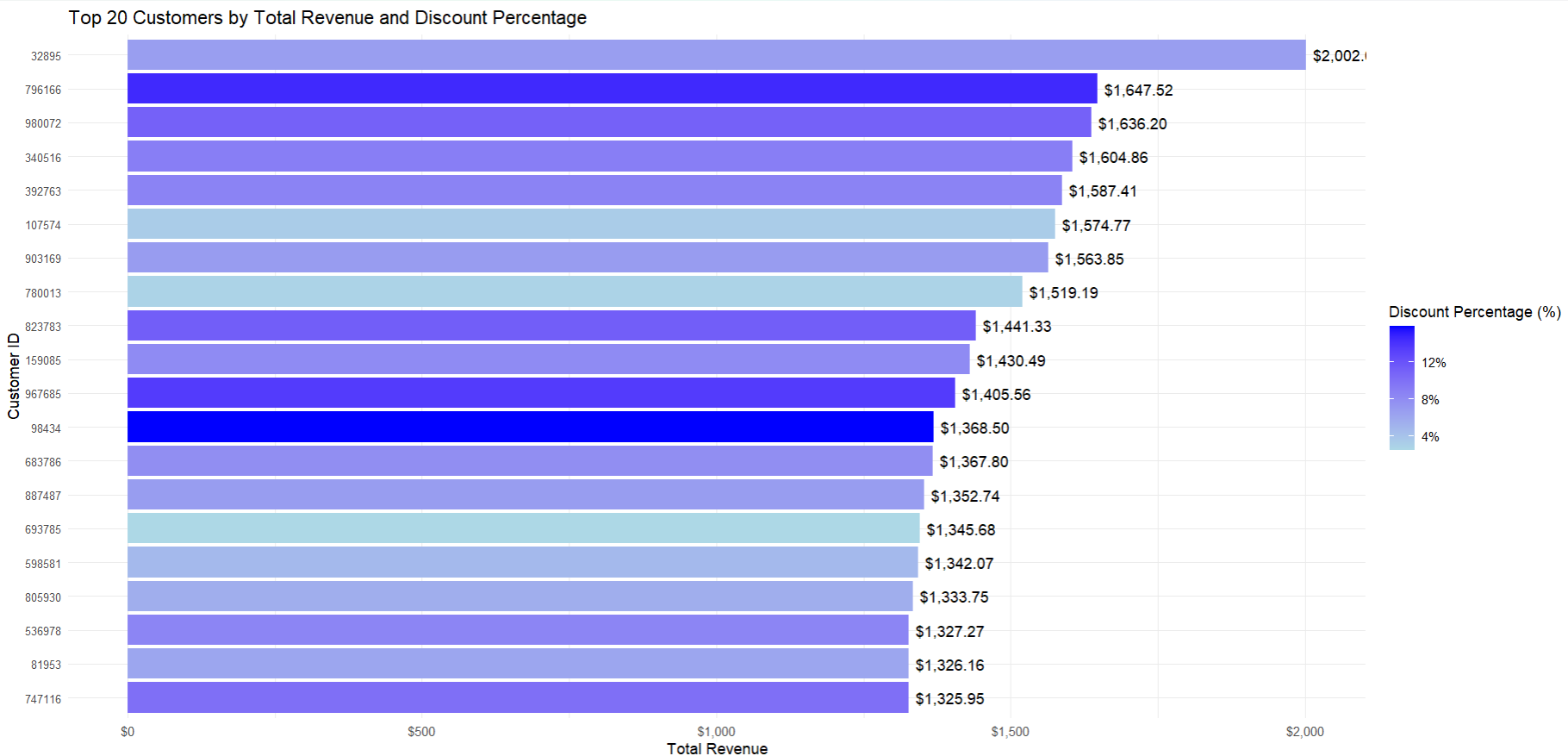
But for the highest spenders the story is a bit different:

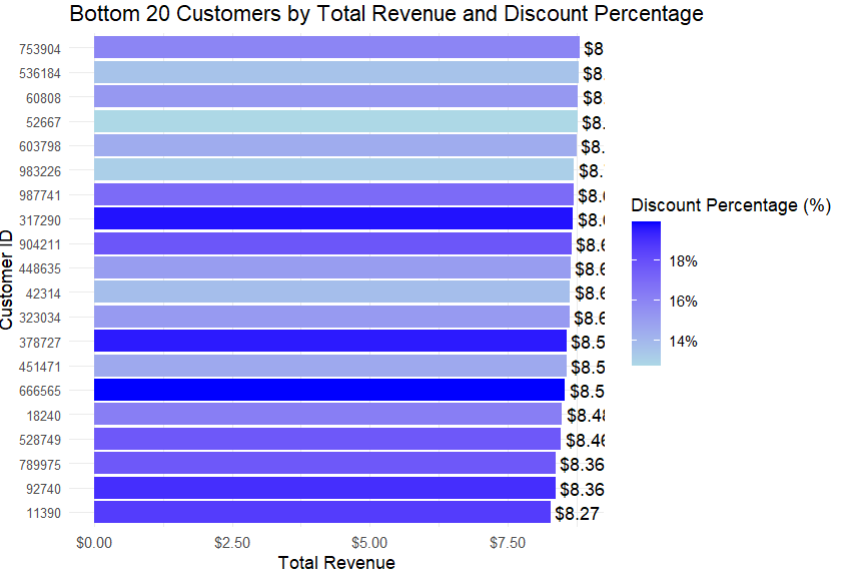




In conclusion, the stores/addresses might appear different for the top 20 because the products might be delivered to different locations. For example some might be delivered at their home, some at their office building etc.

**Now, management might like a list of all the top customers so that they can target them more, or even the lowest, so that they could figure out what they didn’t like. I can provide both of these lists by modifying the code a little and adding lets say the top 500 and bottom 500 to a list that they could use to send targeted emails in which users could describe their experience, what they liked and didn’t like.**

One really important step is seeing how discounts affected this list, here is a look at the data:



This is pretty surprising, even though I would have thought that the customers spending more had a larger discount, this doesn’t seem to be the case!

Probably a business tactic this seller employs, is giving bigger discounts to the bottom customers but this clearly doesn’t seem to be doing the job!

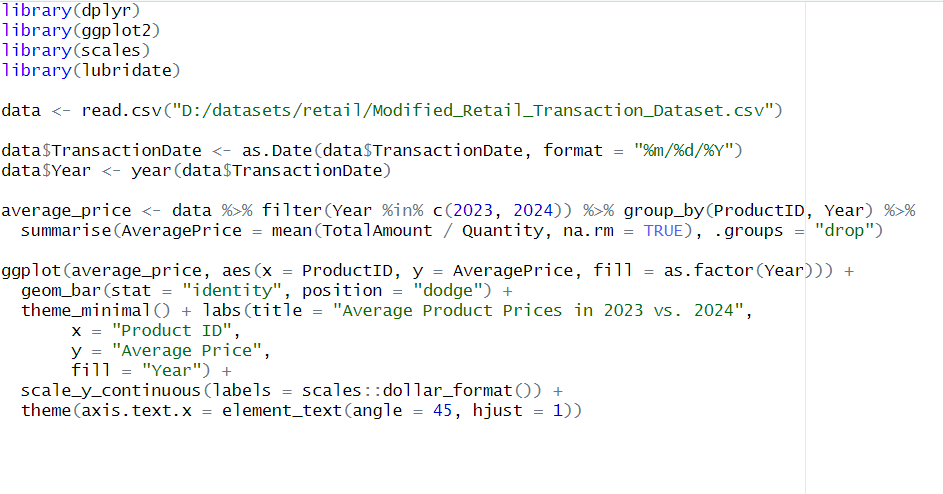
Suggestions:loyalty programs, improve retention rate, don’t focus that much on the discount.

After trying to analyze all the data as detailed as possible I arrived at the conclusion. 

The product prices have increased by 50%!!!!!

All the other columns had data that didn’t show anything important, there is no certain top performing category,preffered payment method etc.

Solution: lower prices, if not possible talk with suppliers and see if lower prices would be a good and sustainable option, company restructure, looking at a different supplier.



Some of the R code that was used to arrive at this conclusion: