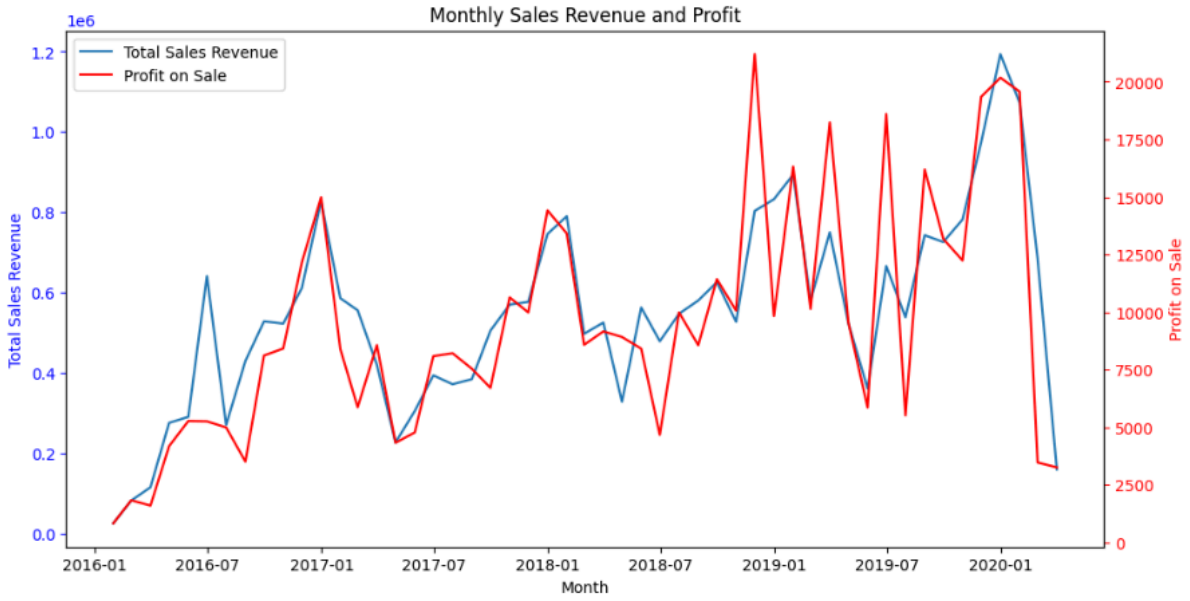
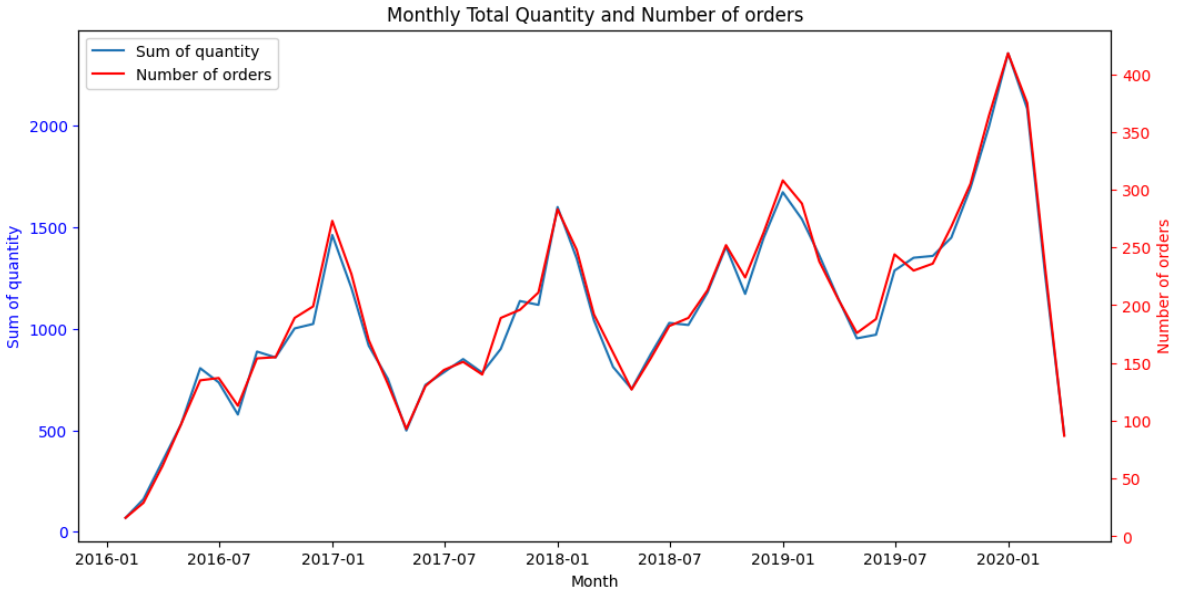
Analysis of the eCommerce Sales_Update dataset

For the basic, preliminary analysis of the dataset and its variables, please refer to the Appendix.

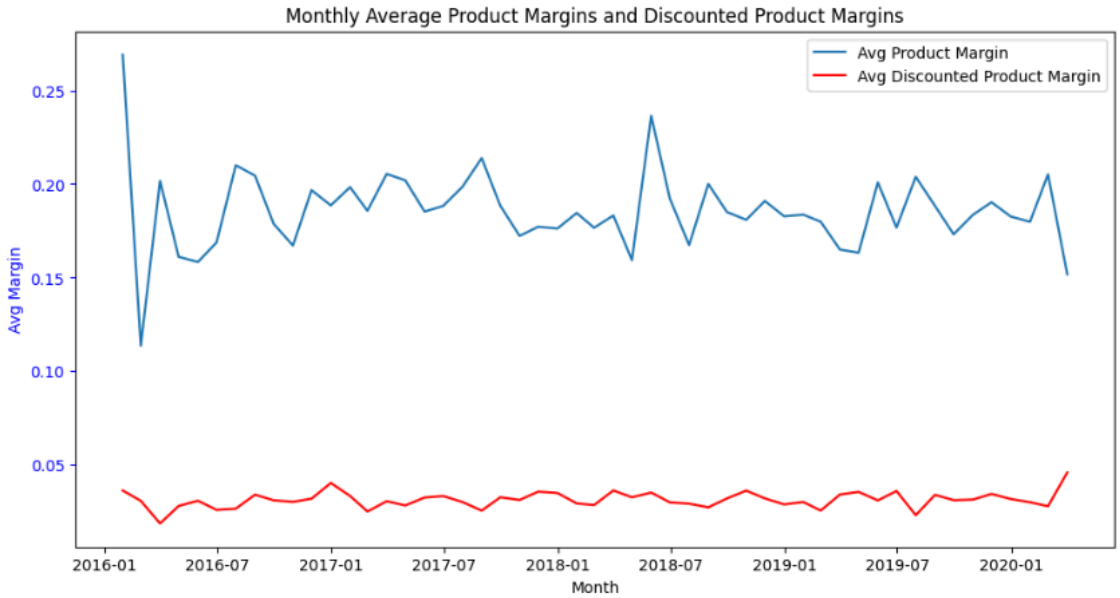
First, we will dive into how the business evolved over time. Let's look at the revenues and profits! To avoid noise in the data, we will aggregate the statistics by month and plot it over the timeframe included in the dataset. To avoid the misleading monthly datapoint for April 2020, we have removed the only 4 orders belonging to this month. We see that in general profits and revenues are going up and are generally in line with each other, but around the end of 2018, we observe large swings in profit that continue over most of 2019.



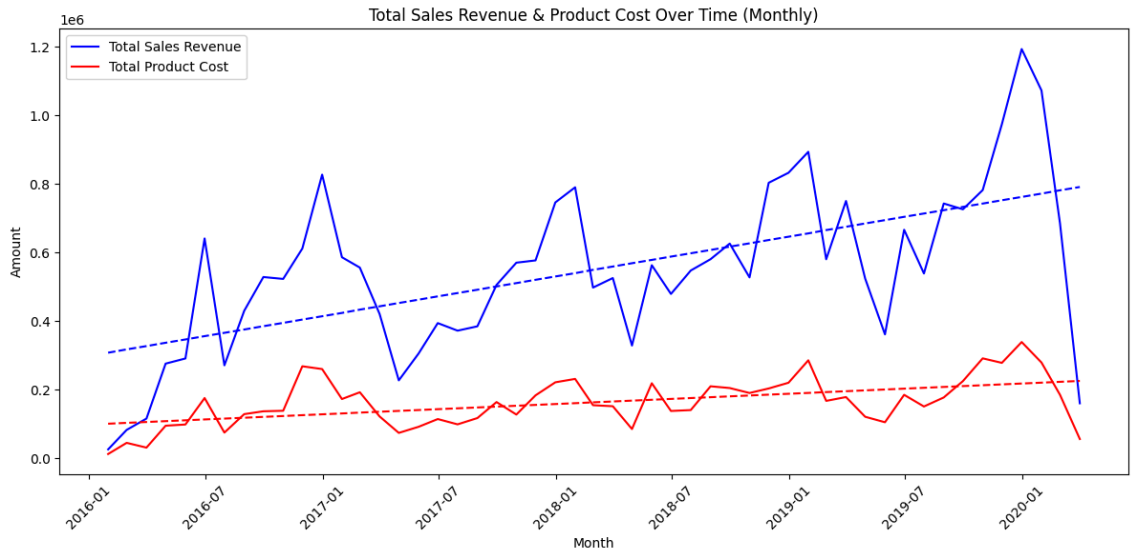
Let's look at the number of orders and quantity! We see that the lines move together indicating that the average quantity per order doesn't really change throughout the time.



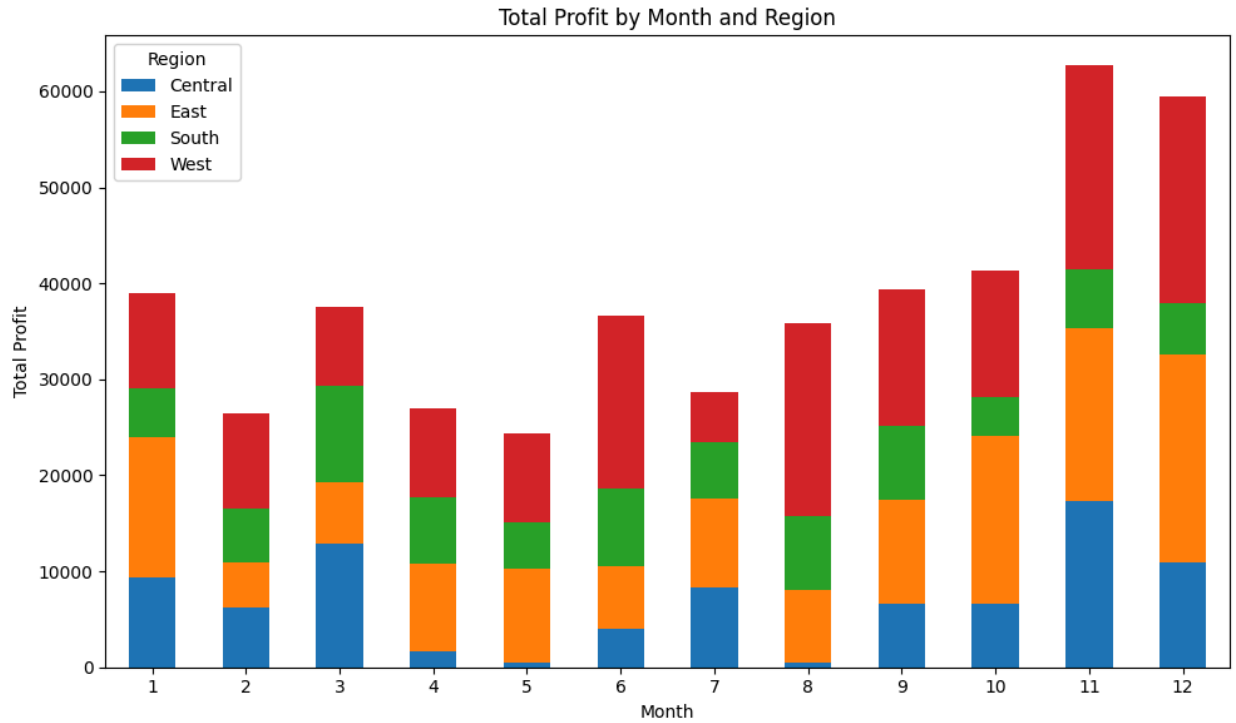
If we look at the margins over time, we see how much the discount drives the profit margins down. Otherwise, the margins seem relatively stable over time.



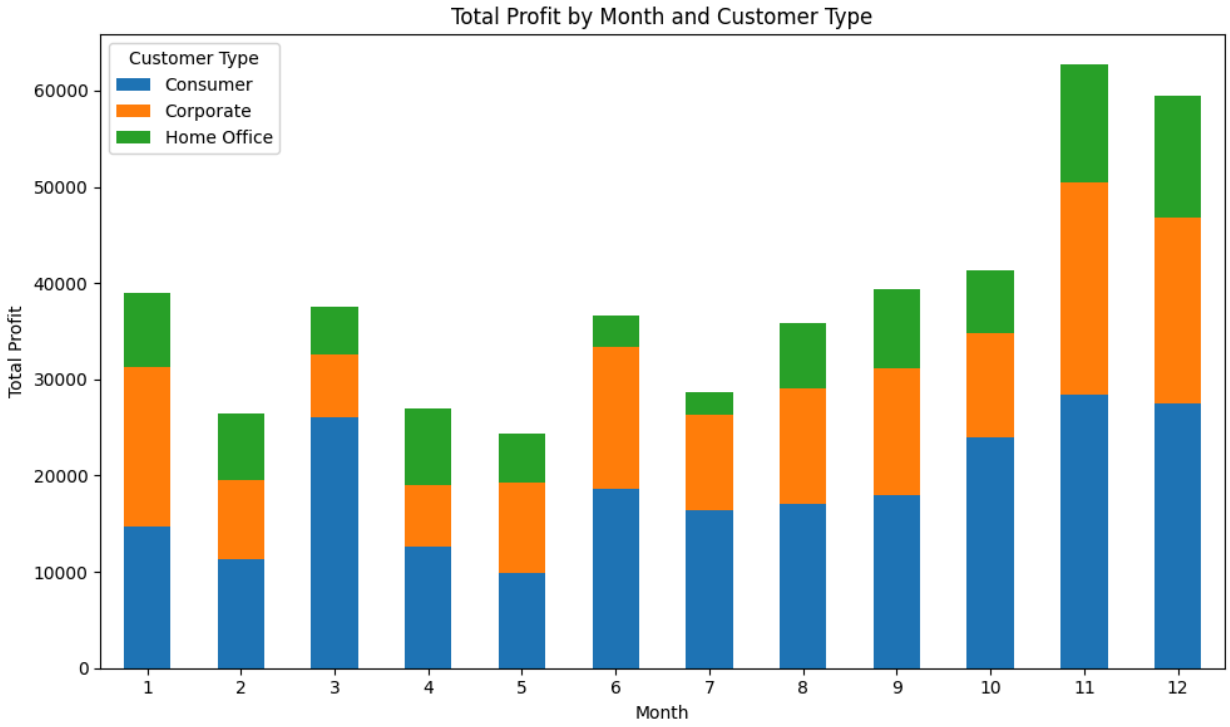
Finally, when looking at the sales revenue vs product cost increase over time, we can clearly see that revenues grow higher than product costs, which is great news for the business. The large drop from 2020 February to March is pretty concerning, however.



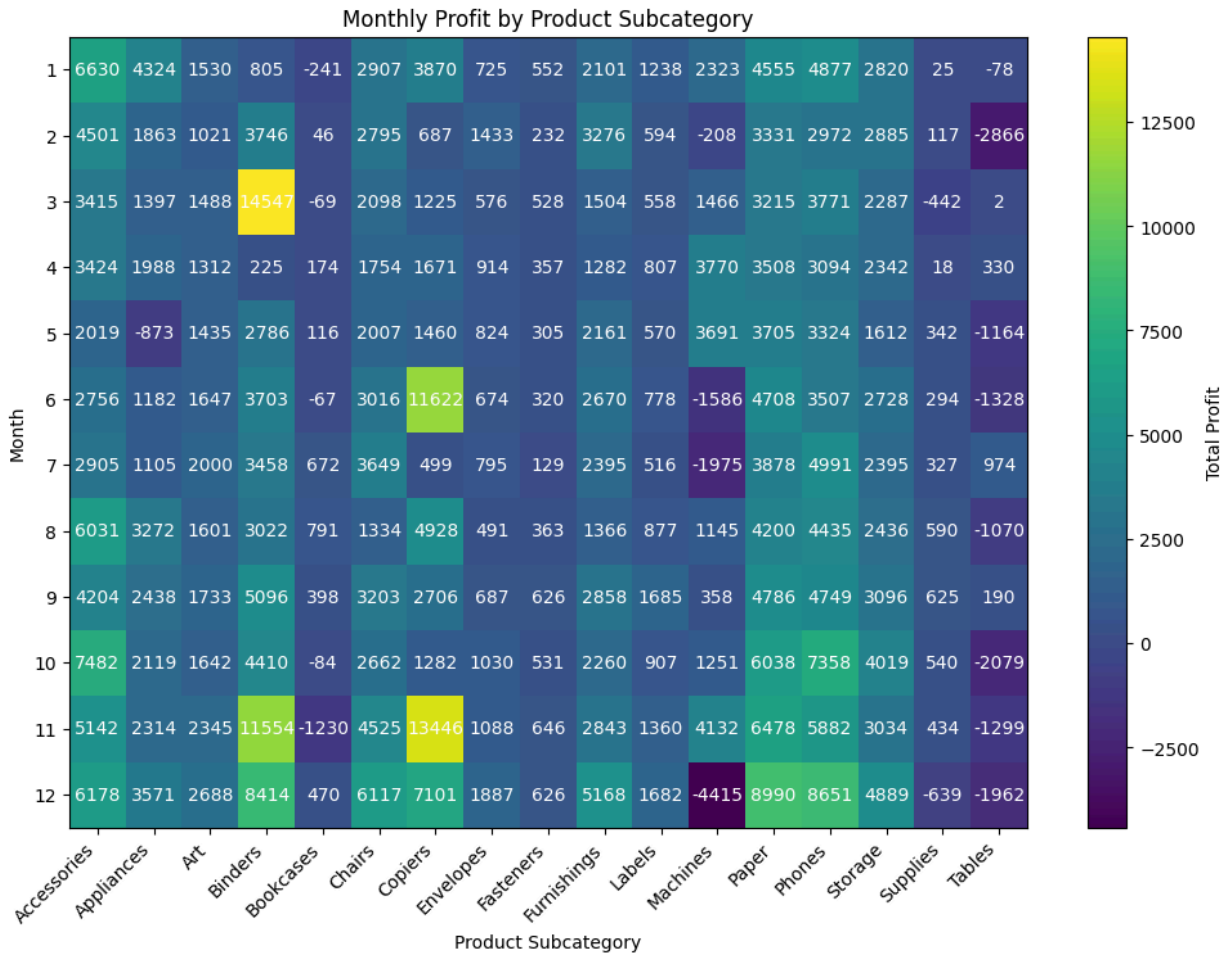
We can clearly see the seasonality trend in the time series chart above. Let's break the data down by month and understand how the total profit varies across the 12 months. (Note: for this graph we removed 2020 data because it was only up to March and would've unfairly inflated the numbers for January to March.) In general, we see higher than average profitability in all regions in November and December, except for the South region.



When it comes to customer types, one interesting fact is that the most profitable month for Consumers is March. In general, for all customer categories, November and December are higher than average profit months.

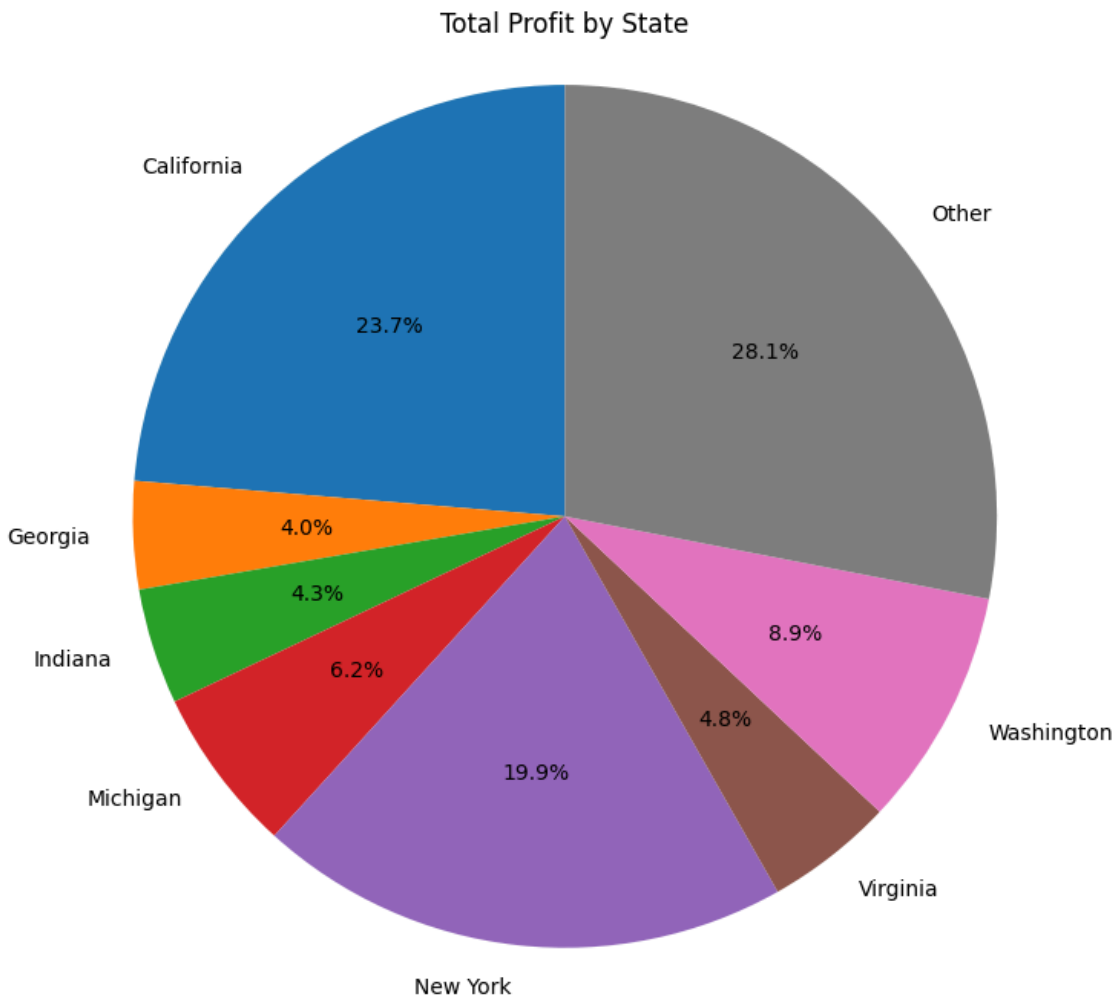


To dive deeper into what is driving the high year-end profitability we make a 2D heatmap between months of the year and product subcategories:

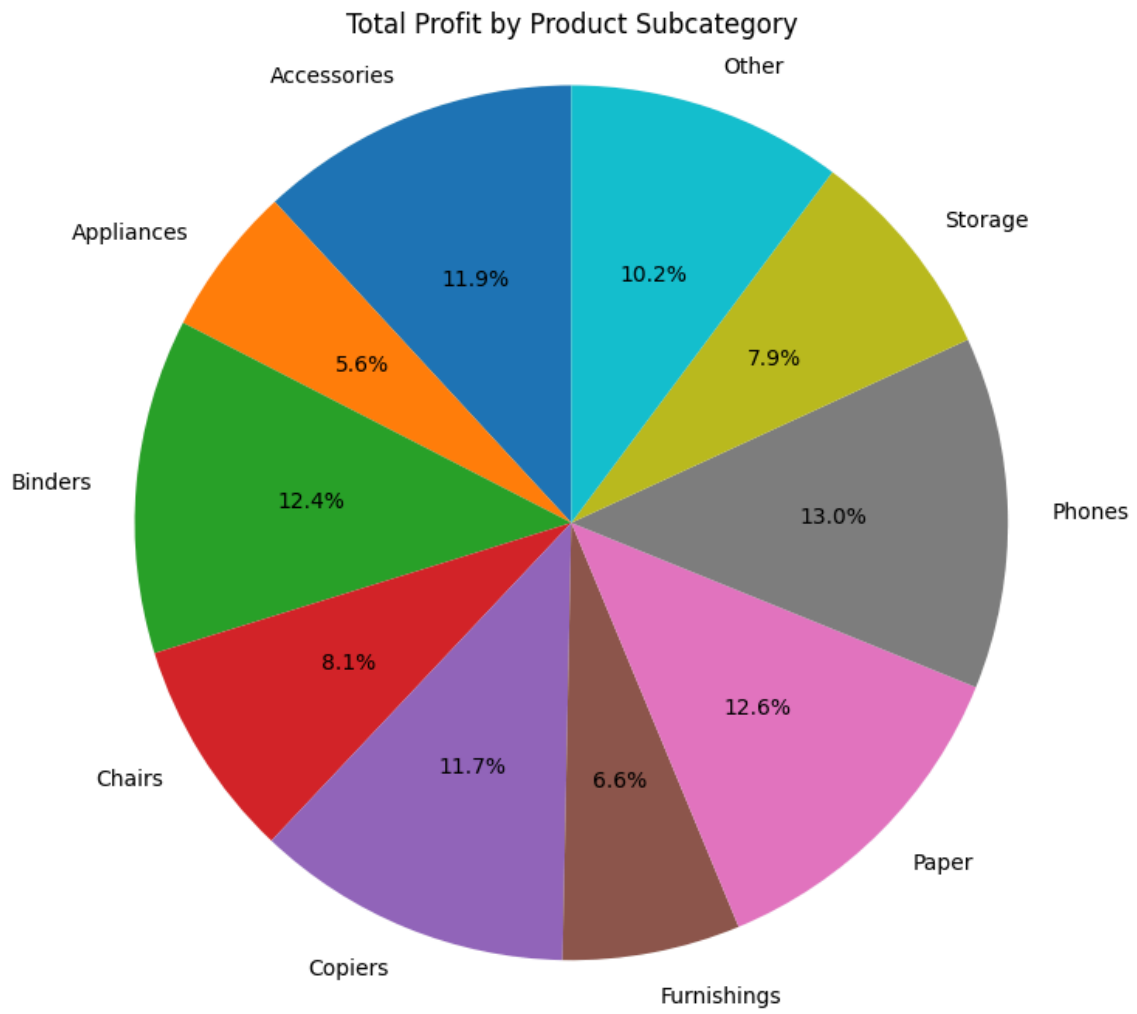


We observe the November profitability is driven by Binders and Copiers and that Binders seem to be oddly profitable in March.

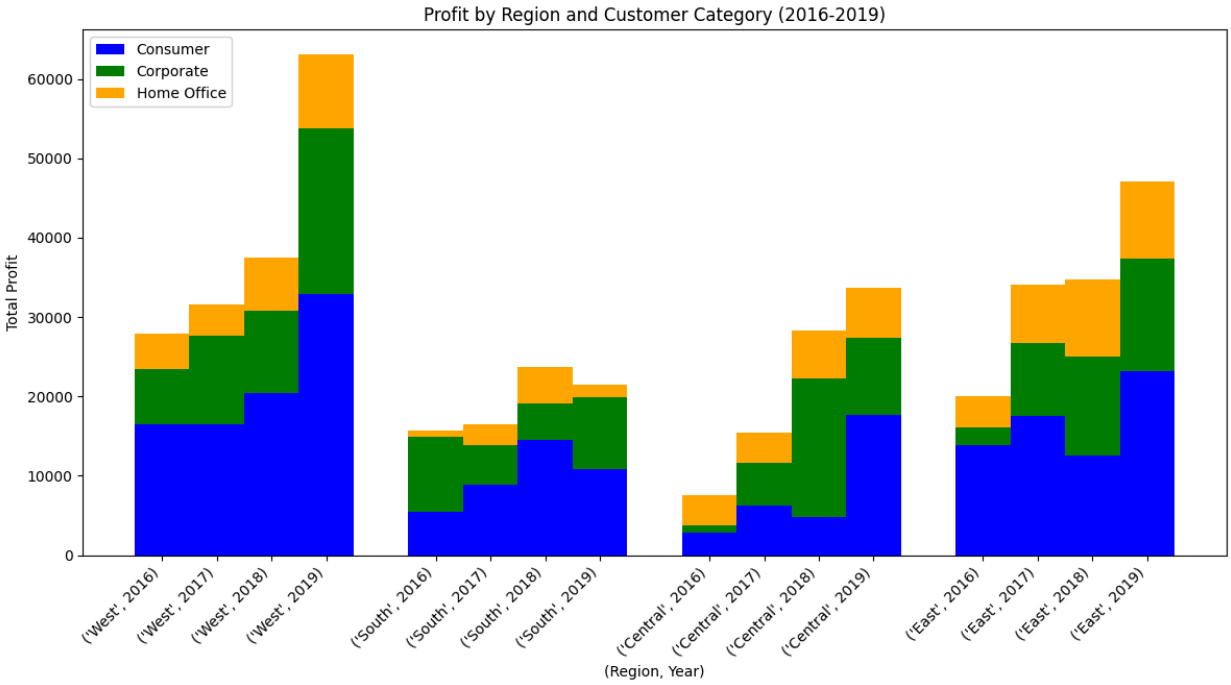
Now let's look at the most profitable states overall. We can easily visualise this using a pie chart:



We see that California is the most profitable state followed by New York.
Equally, let's examine which product subcategories are responsible for the most profits:

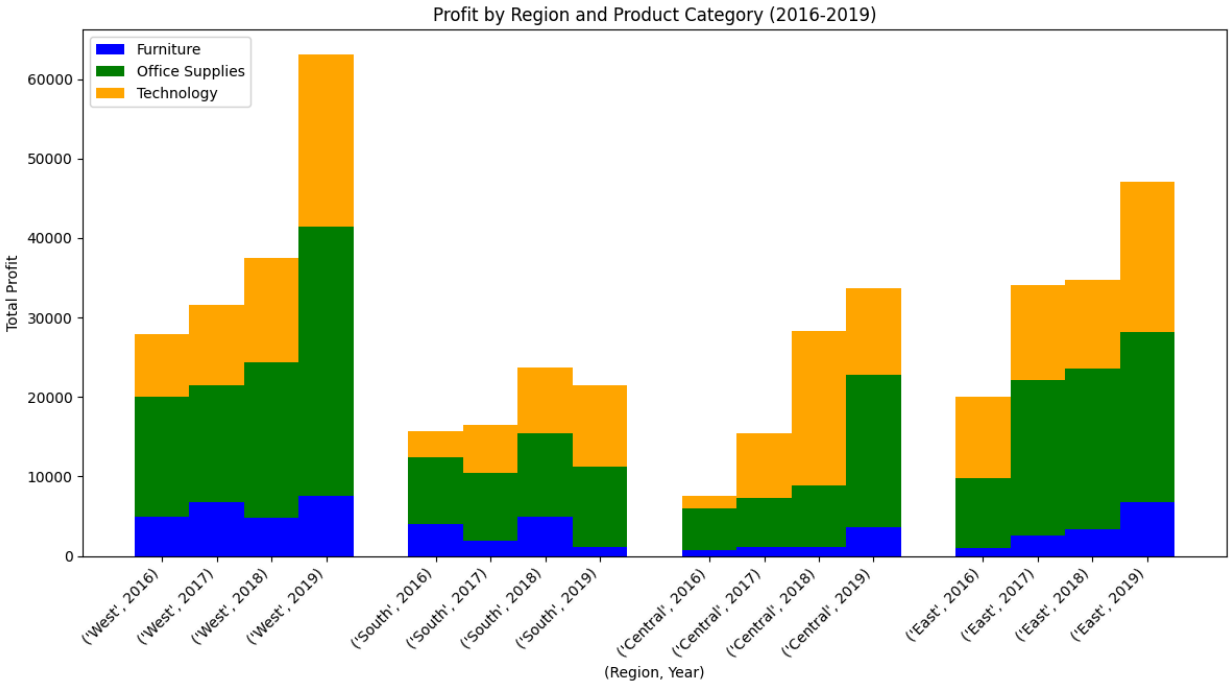


Phones, papers, binders, accessories and copiers seem to be the most profit-generating subcategories in this order.
Our next step is to look at how the profit changes in the 4 regions for each customer type and product category over the years we have complete data for (2016-2019):

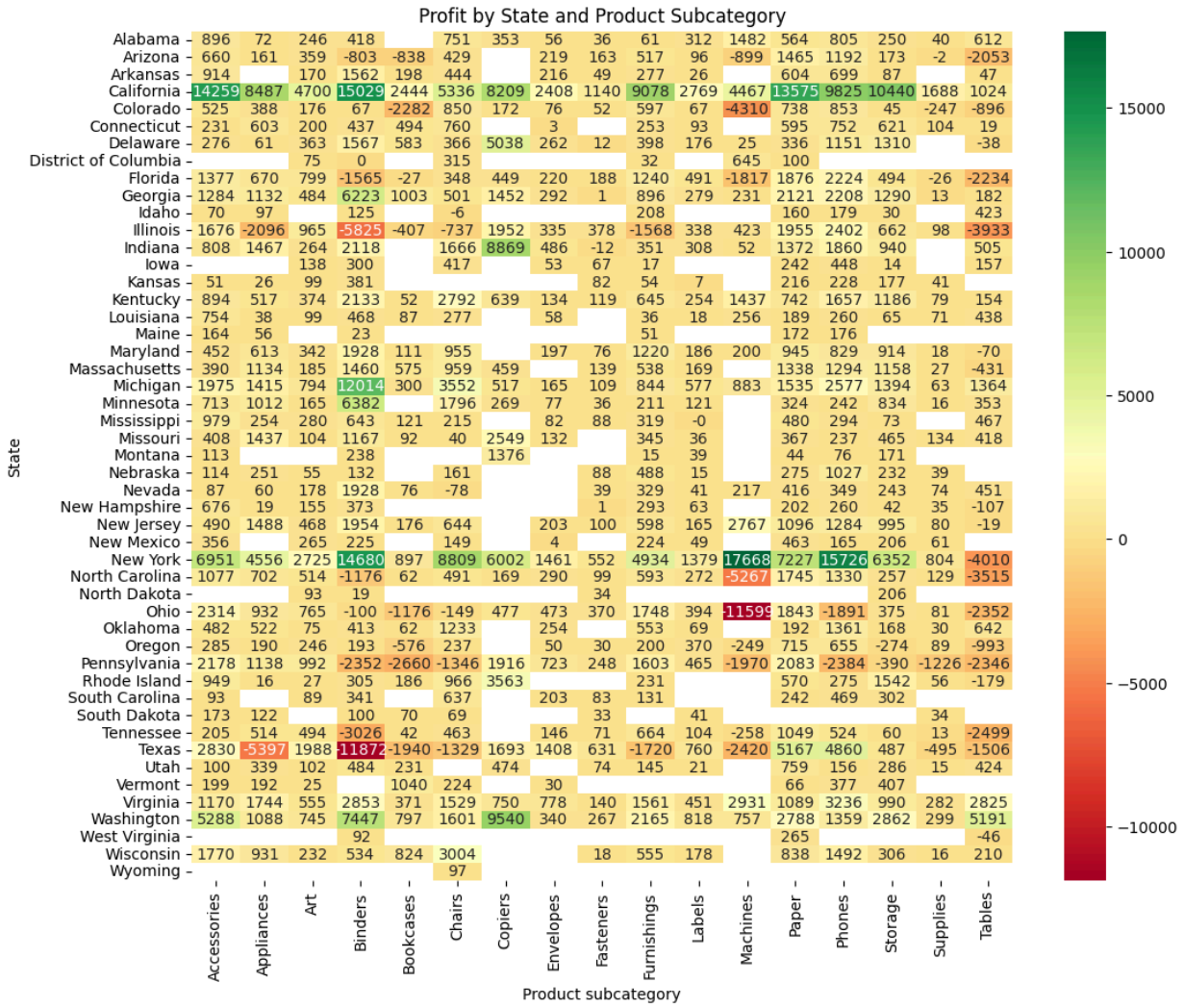


The recent big profit growth in the West region is driven by increases in all 3 customer categories. Recent growth in the East region is fuelled by about a 50% increase in the Consumer category compared to last year. In the Central region, Consumers customer type profits have more than tripled, more than compensating for the lost profits within the Corporate segment. The consumer segment in the West region seems to be the most lucrative business category.

Now, let’s look at product categories!

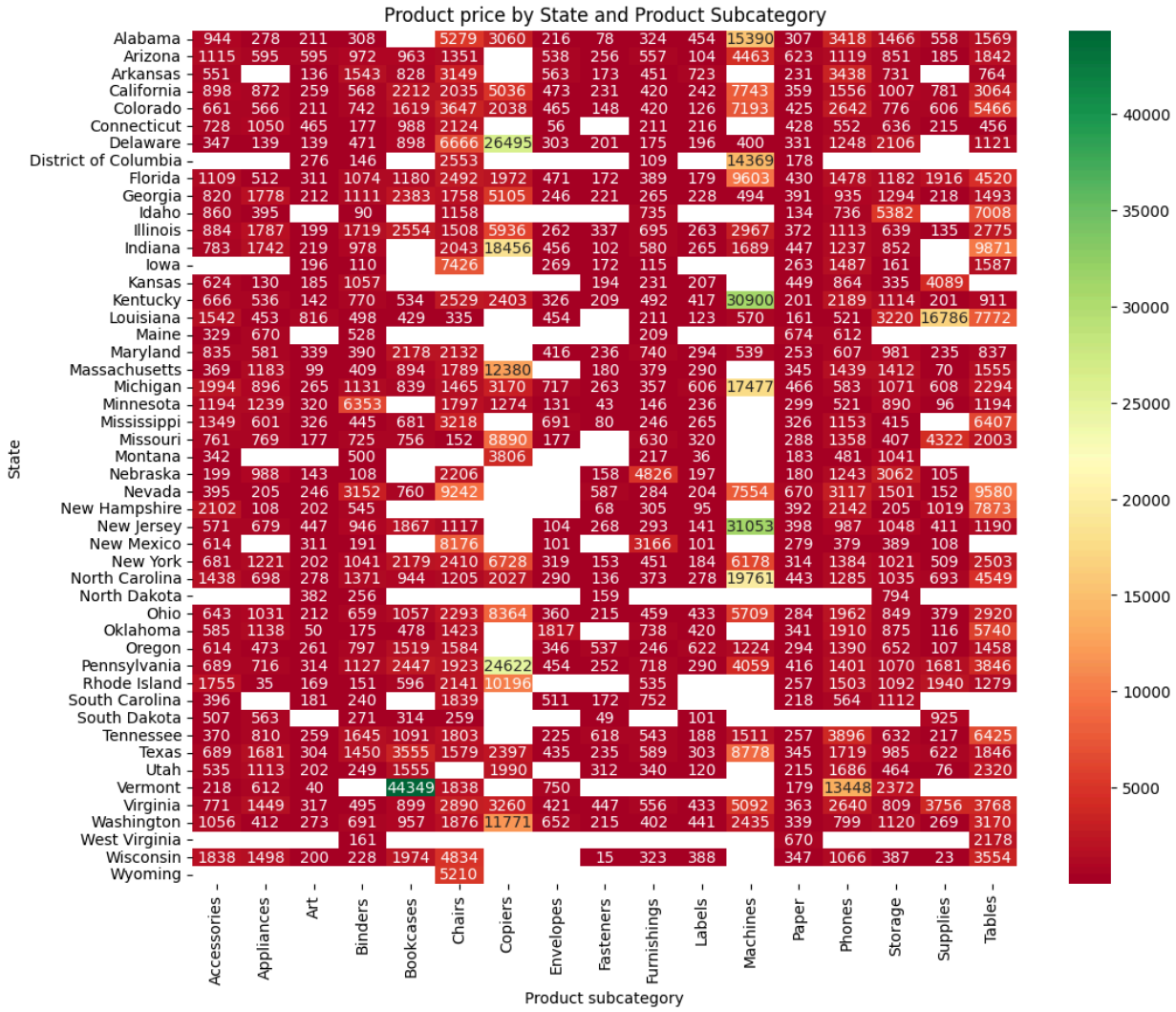


We observe recent big growth in the West region driven by profit increases in Office supplies and Technology. The South region is underperforming with negative growth in the recent year. The Central region started growing rapidly, but growth slowed in the recent year due to decrease in Technology profit. The East region shows more or less steady growth. Throughout the years and across all regions, the Furniture category produces the least amount of profit. Office supplies and the West region seem to be the most lucrative.
Let’s dive deeper by looking directly at States and product subcategories.



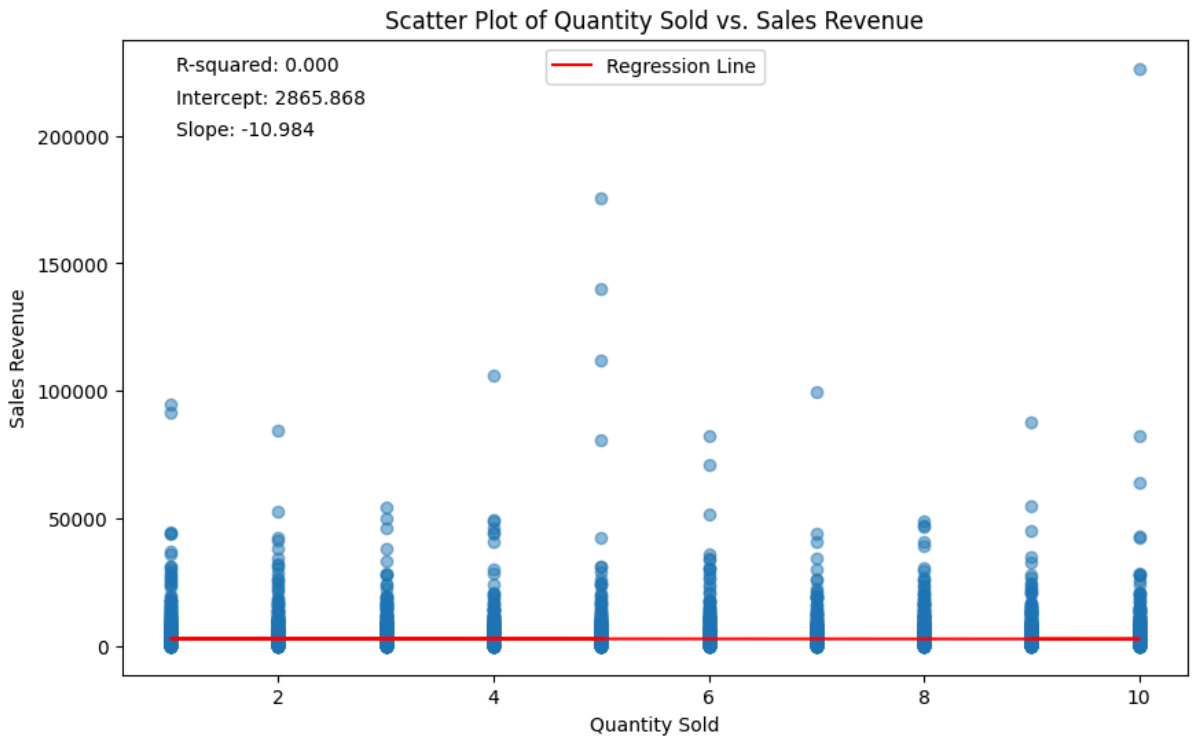
We observe that the most profitable (State, Subcategory) combination is (New York, Machines) and the least profitable is (Texas, Binders), shortly followed by (Ohio, Machines). It is quite interesting to note that machines are highly profitable in New York, but very much lossmaking in Ohio.

Now let's take a look at product prices. This is important to see how consistently the company prices similar products across different states.



We observe high variance in average product subcategory pricing across states. For example, the average original price of bookcases in Vermont was $44,349, but was only $314 in South Dakota. This doesn't seem realistic, even if an 80% discount is offered to buyers in Vermont, unless the distribution of bookcase products that are sold in Vermont are skewed towards the real expensive ones. To investigate this, let's dive deeper and go to the product level: we then discover that only one bookcase was sold in Vermont, which was product FUR-BO-10004834 sold on 2017-12-02 for a price of **$44,349**. Let's see where else this product was sold: for example, there was a sale in New York a few weeks before, on 2017-Oct-26. The product price was **$5,323** and the discounted product price (15% off) was $4,259. It is highly unlikely that the company suddenly increased the product's price over 7 times for a lower average-income state a month later: outliers may be present, but data cleaning was not part of this assignment and hence will not be covered here.

Given the above discrepancy, let's look at whether other generally accepted business principles are satisfied for this dataset. For example, we'd expect the revenue on the order to grow reasonably with the quantities ordered.

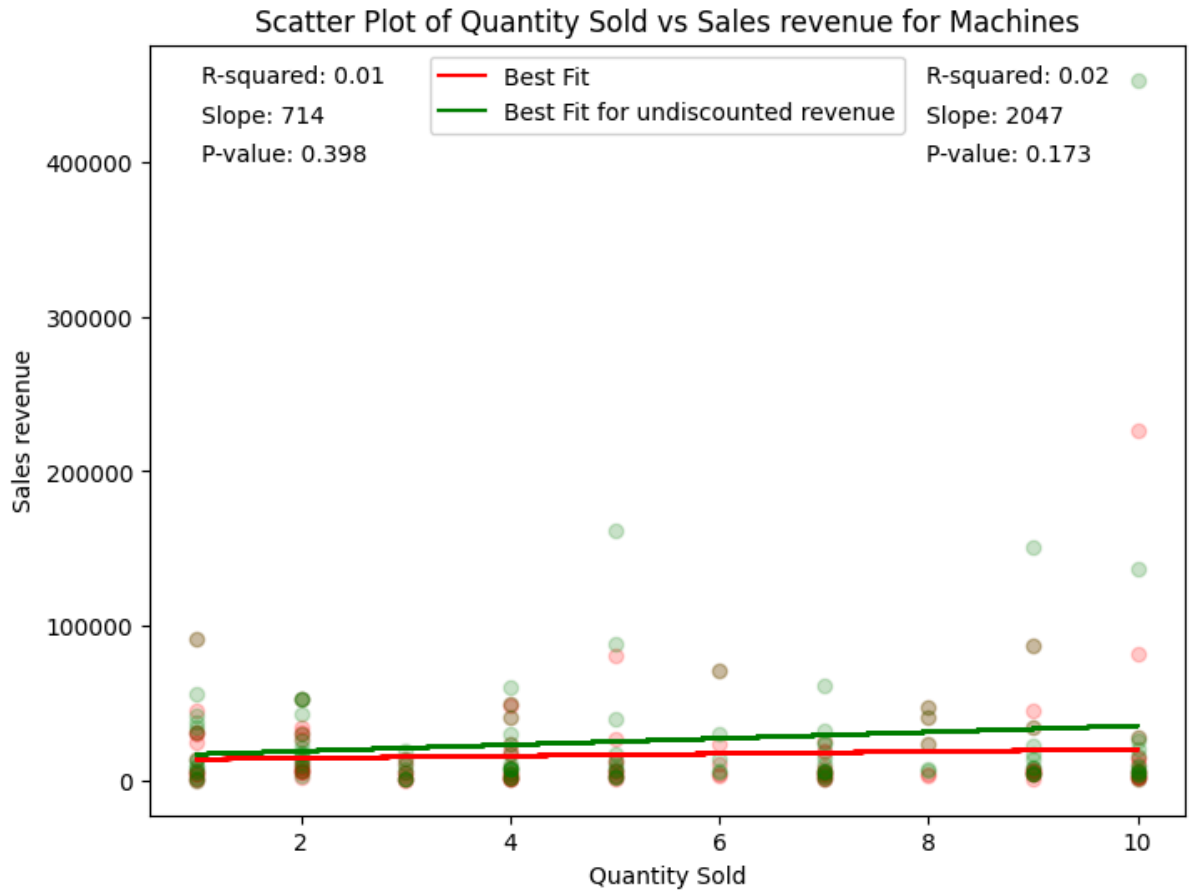


This doesn't seem to be the case across the entire dataset. Of course, there could be multiple valid reasons for this:

(1) there may be confounding, where cheaper products are on average sold at higher quantities driving revenues down at larger quantities. To explore this further, we will make the same plot for product subcategories and for the top products too, to see if we observe the expected linear growth with product price being the slope of the line.

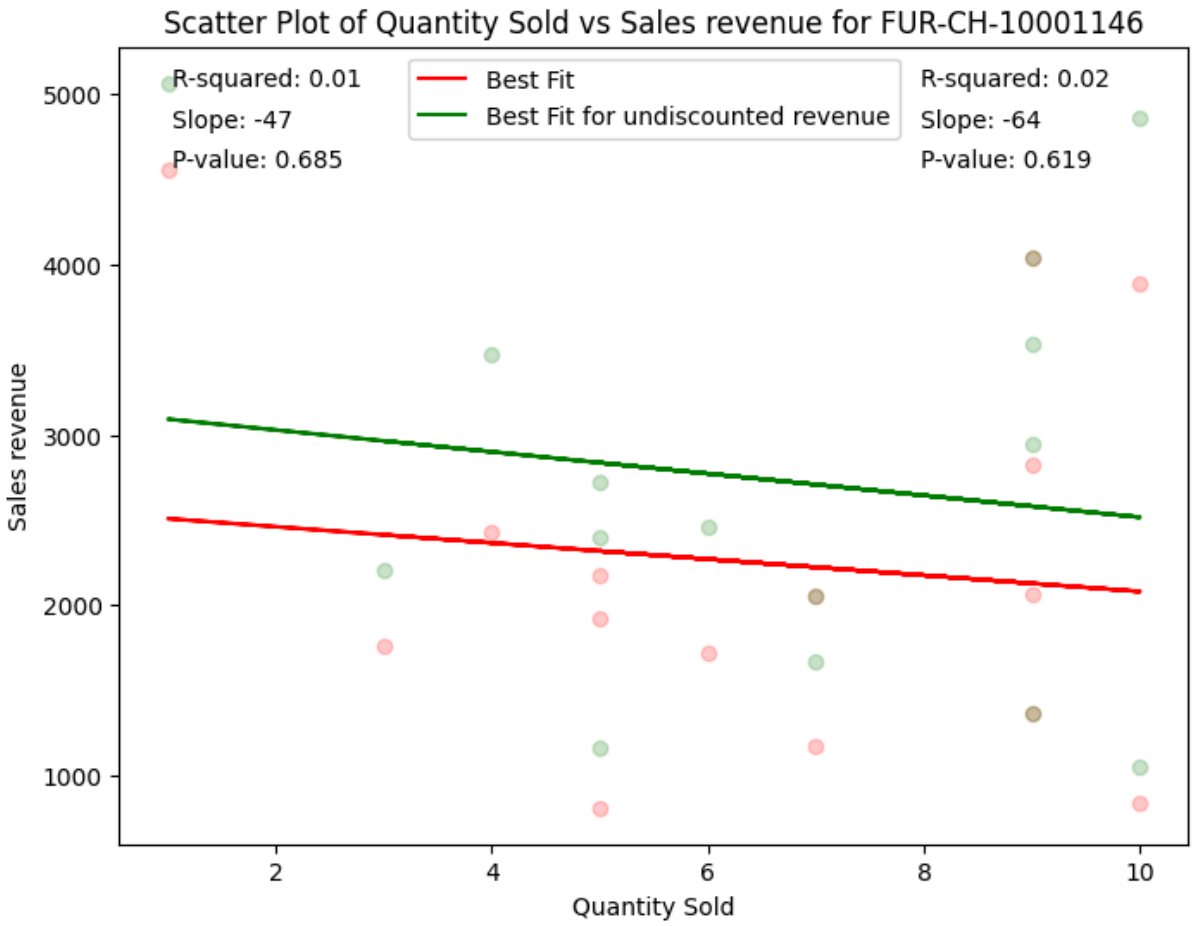
(2) Another reason may be discounting: if multiple purchases are discounted aggressively (let's say 50% discount if a customer buys two products instead of one, essentially giving the second one away for "free") we may observe a smaller increase in revenue than expected when the customer buys more quantities of the same product. To explore this further, we will calculate what the revenue would've been without discounts (simply *product price * quantity*) and see whether plotting that against the quantity reinstates the expected relationship. (Another way of framing this question is asking whether product price is constant with the quantity of the product, which it should be if there are no discounts - but it turns out there's a hyperbolic relationship, see Appendix.)

When plotting sales revenue against Quantity sold for product subcategories, the most significant increase in terms of p-value (0.398) happened for Machines. Plotting the hypothetical undiscounted revenues (obviously in practice, demand would decrease so revenue would be likely lower than this) vs quantity increased the steepness of the slope to more than double and resulted in a p-value of 0.173 (still not significant). Interestingly, for several product subcategories, both regression lines were downward sloping, suggesting a "negative price" solely based on this relationship.



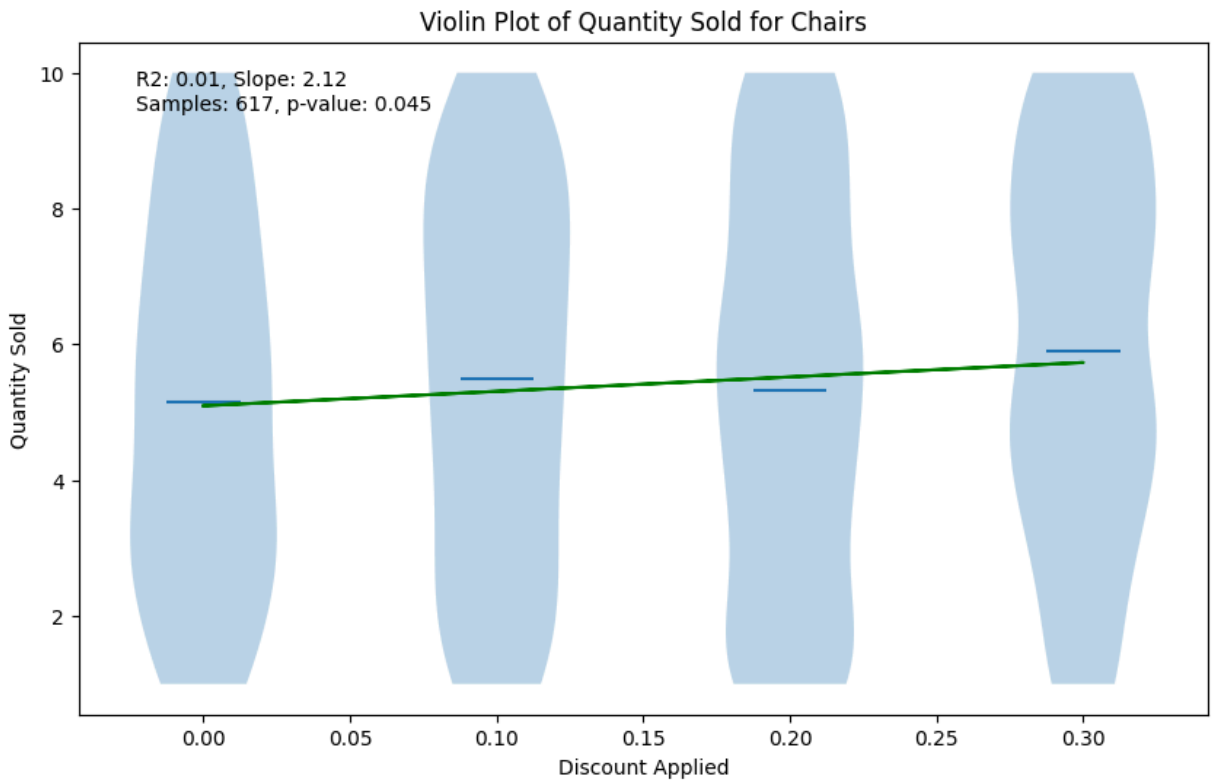
If we now plot the same graph for the most popular product, OFF-PA-10001970 we see that the sales revenue increases with quantity (still not significant association though). Slope is around $200 whereas mean product price is $354 and discounted product price is $343.

The same plot for the second most popular product, FUR-CH-10001146, however, paints a very different picture.

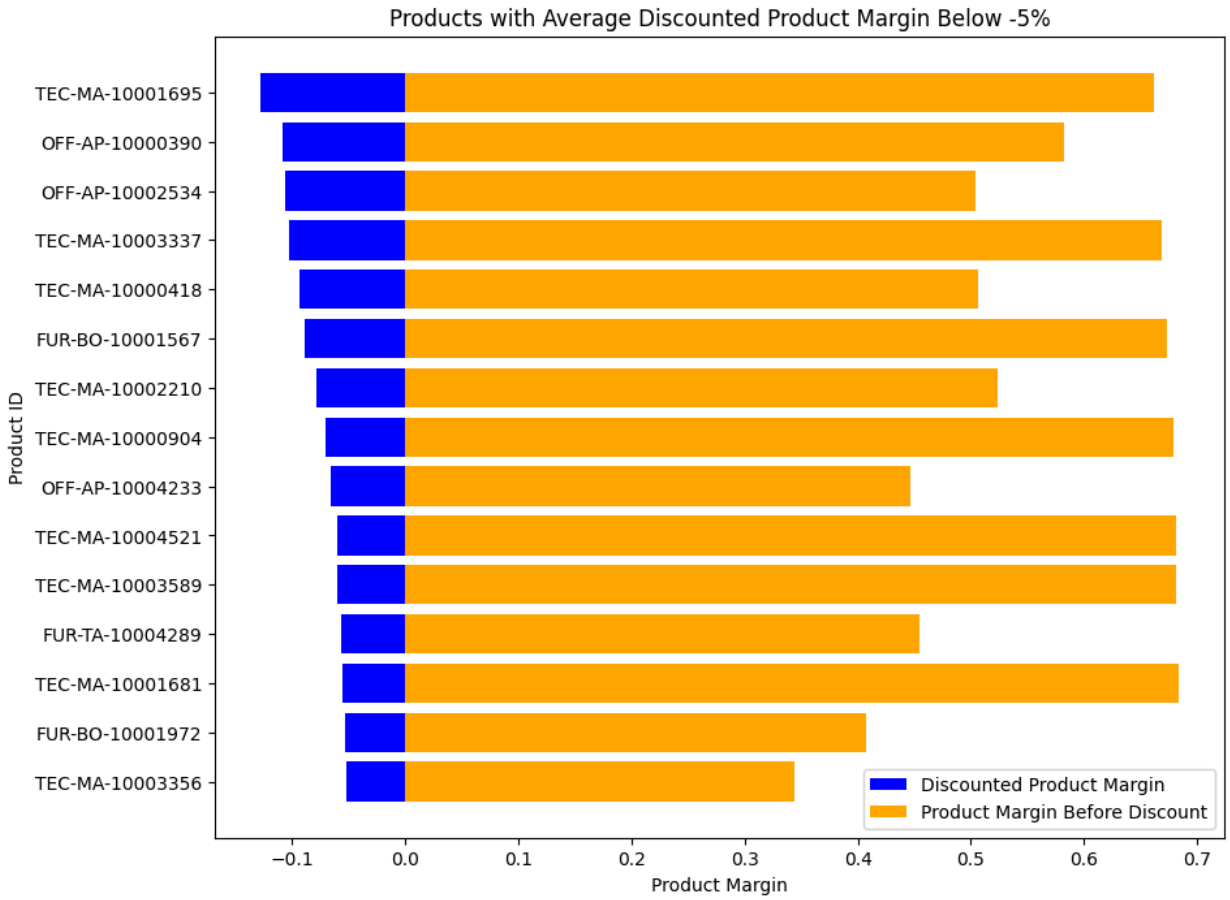


One explanation for the lack of significant increase of sales revenue with respect to quantity sold could be that the original 'Update sale' variable was not actually the revenue on the order but the revenue per product in the order, in other words, the product price given in the order. However, the professor confirmed that 'Update sale' was the revenue on the entire order, so I'll not pursue this idea any further.

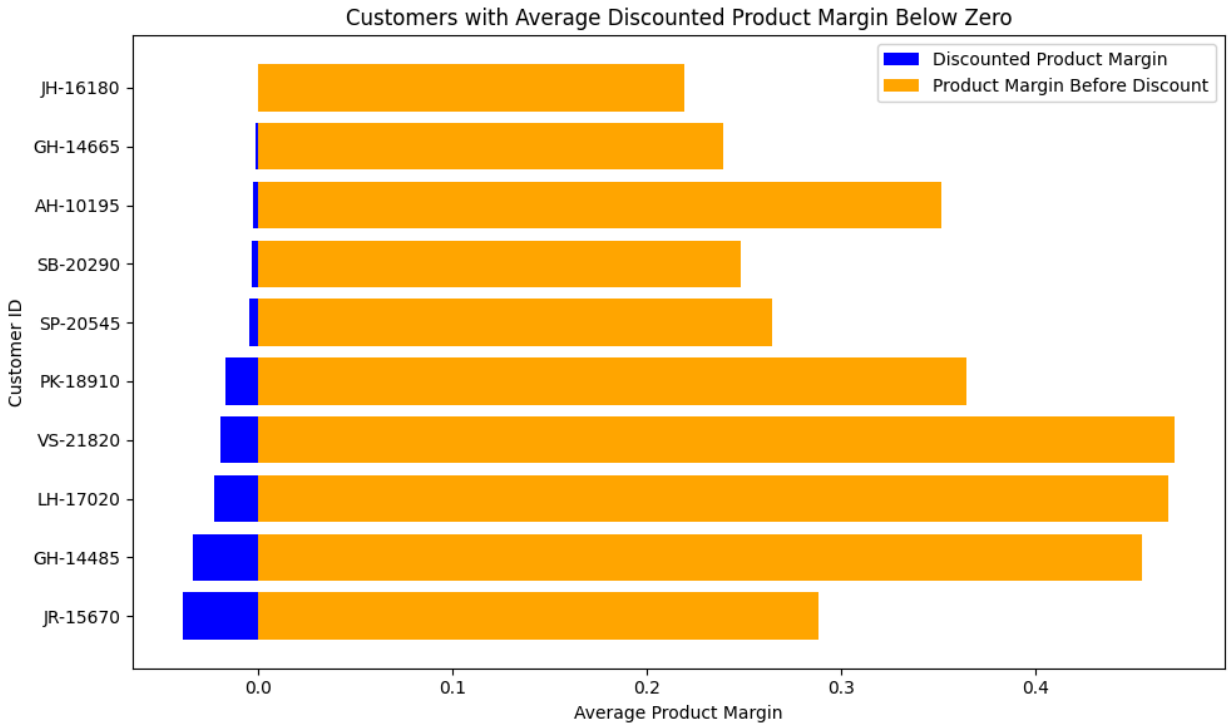
To further elaborate on (2), we can directly examine how quantity sold is related to discount offered in the various product subcategories. In general, quite surprisingly, we've seen no significant relationship between the two, except for the Chairs subcategory (p-value: 0.045, significant, not accounting for multiple hypothesis testing):



We may continue our analysis of discounts by examining what products have been overly discounted with their discounted product margin going well below zero (ideally this should never happen as you're losing money on the sale).



Let’s make the same plot for customers to see who are the best at negotiating:

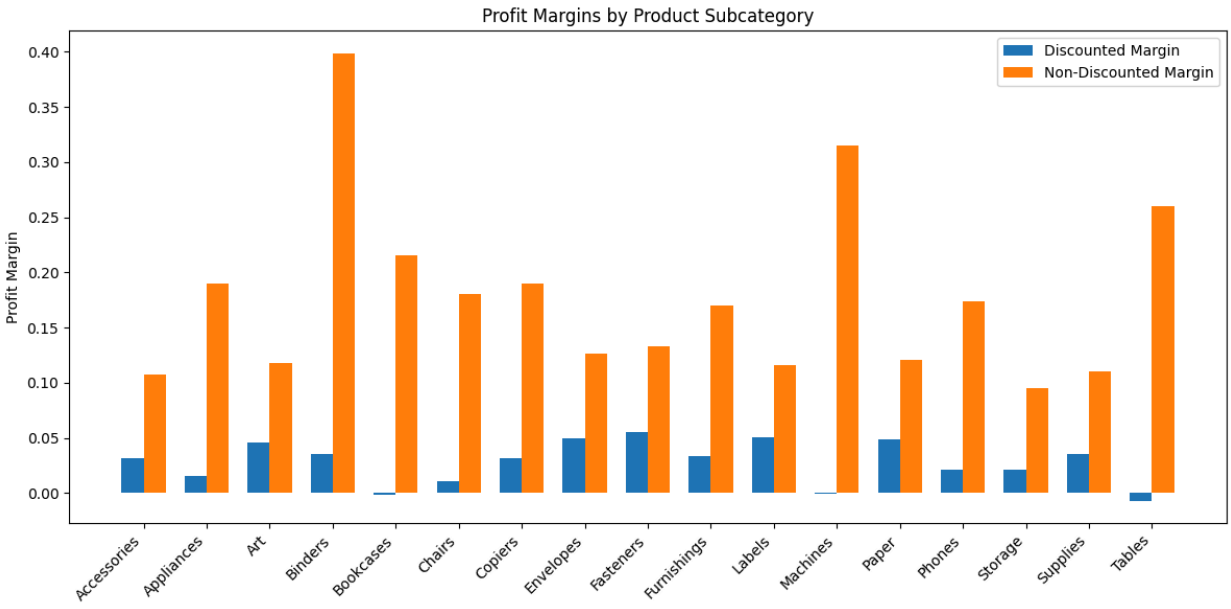


We see many product and customer examples in which an average high margin before discount (40%+) was reduced to a detrimental negative product margin after the discount.

Let's also find information about the least profitable products and customers, in other words those who lost the most money for the company.
There are 134 products that have overall caused financial loss for the company. The following products lost more than $1000 for the company:

Product ID	Total profit	Total quantity sold	Number of unique customers	Avg Product Cost	Avg Product Price	Avg Discount applied	Avg Product Margin	Product Subcategory	Most common City
TEC-MA-10000418	-8840.0	20	3	6959.416667	13612.935406	0.533333	-0.093215	Machines	[Lancaster, Newark, San Francisco]
TEC-MA-10000822	-4480.0	18	4	15176.450000	23039.166667	0.400000	-0.034116	Machines	[Detroit, Houston, Louisville, San Antonio]
TEC-MA-10004125	-3838.0	5	1	16925.600000	32316.000000	0.500000	-0.047506	Machines	Burlington
FUR-TA-10000198	-2770.2	21	5	9112.562667	13097.433333	0.280000	-0.022266	Tables	[Concord, Detroit, Knoxville, Los Angeles, Spr...
FUR-TA-10001889	-1838.3	48	7	2780.618175	4065.102834	0.350000	-0.027934	Tables	New York City
TEC-MA-10002412	-1767.1	10	1	22817.210000	45281.000000	0.500000	-0.007805	Machines	Jacksonville
OFF-BI-10004995	-1723.2	40	6	5198.455556	10893.437500	0.450000	-0.046960	Binders	[Burlington, Houston, Jackson, New York City, ...
OFF-SU-10002881	-1166.1	28	6	9355.875880	9750.272338	0.100000	-0.008274	Supplies	Philadelphia
FUR-TA-10001950	-1128.0	23	4	8252.192708	8987.267361	0.200000	-0.019369	Tables	[Columbia, New York City, Noblesville, Seattle]
FUR-TA-10004154	-1098.4	36	5	1740.984333	2726.734921	0.300000	-0.012228	Tables	[Houston, Marion, New York City, Philadelphia,...
FUR-TA-10004289	-1066.5	14	3	1662.001111	2991.567677	0.483333	-0.056778	Tables	[Chicago, Denver, Jacksonville]
TEC-MA-10002210	-1033.3	11	2	1887.825000	3637.638889	0.550000	-0.078442	Machines	[Houston, Springfield]

On the product subcategory level, the margins look as follows. Bookcases, machines and tables on average seem to be losing money because of the discounts given to customers.



There are 76 customers who have overall caused financial loss for the company. The following customers lost more than $1000 for the company:

Customer_ID	Total profit	Total quantity sold	Number of unique products sold	Most common product subcategories	Avg Discount applied	Customer type	First order	Most common City
CS-12505	-6479.3	45	9	[Machines, Paper]	0.200000	Consumer	2016-05-10	Lancaster
GT-14635	-4002.5	30	6	Binders	0.250000	Corporate	2016-07-02	Long Beach
LF-17185	-3342.0	90	16	[Binders, Furnishings, Paper]	0.318750	Consumer	2016-08-01	San Antonio
SR-20425	-3104.8	36	9	Binders	0.366667	Home Office	2016-07-05	Louisville
HG-14965	-2410.9	84	17	Binders	0.170588	Corporate	2016-10-22	Los Angeles
NC-18415	-2000.9	84	14	Binders	0.264286	Consumer	2016-06-01	Jacksonville
SB-20290	-1819.8	89	17	[Machines, Storage]	0.241176	Corporate	2016-05-19	[Henderson, Philadelphia]
SM-20320	-1675.7	74	15	Paper	0.246667	Home Office	2016-04-30	Jacksonville
CP-12340	-1529.1	89	15	[Binders, Tables]	0.213333	Corporate	2016-03-21	New York City
NF-18385	-1403.0	61	14	Chairs	0.250000	Consumer	2016-03-18	Newark
BM-11140	-1304.9	83	16	Storage	0.168750	Consumer	2016-09-24	San Antonio
TB-21520	-1228.1	119	20	Paper	0.265000	Consumer	2016-10-27	Philadelphia
DB-13120	-1133.8	76	14	Paper	0.142857	Corporate	2017-06-05	[Rochester, San Francisco, Tampa]
DC-12850	-1007.6	106	18	[Art, Storage]	0.211111	Consumer	2016-08-25	Memphis

The average discount in all three customer categories is around 15%, so most of these customers seem to be getting preferential treatment, or they are selectively buying heavily discounted products.

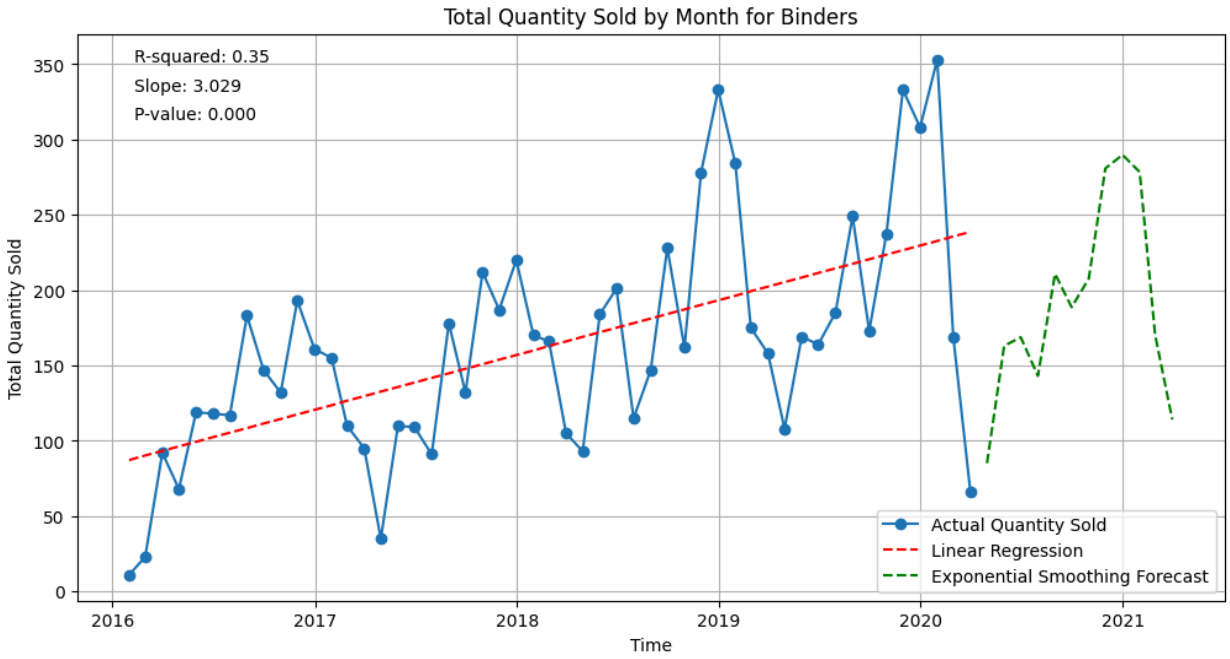
This brings us to profits vs discounts. For every subcategory except for Fasteners, more discount was significantly negatively associated with profit. This was also true for every single region. This implies that the **company's discounting strategy does not work at all**.

Forecasting

To forecast future sales, we will use exponential smoothing with a period of 12 months, which can be clearly observed on the 'Monthly Total Quantity and Number of Orders' figure shown

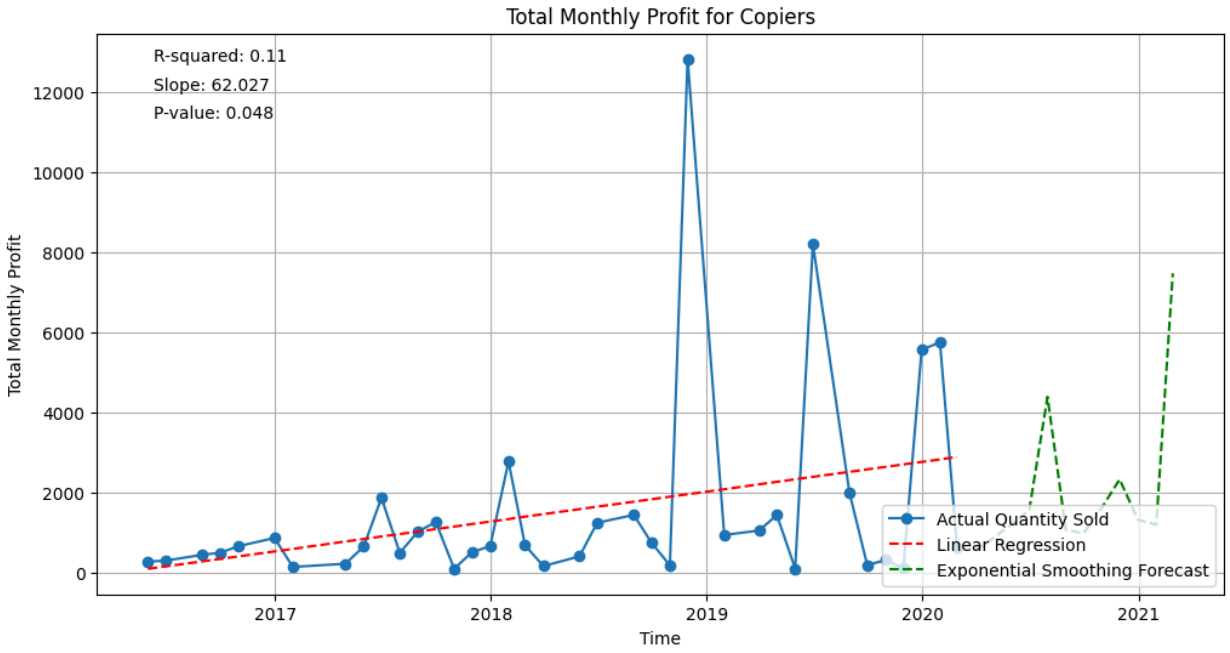
earlier. We will forecast the total quantity sold by month over the next year as well as the total profit by month over the next year separately for each product subcategory. Based on this, we will highlight the quickest growing product and also comment on the ones that are in heavy decline.

For monthly total quantity, Binders have the highest (significant) increase over time with an average slope of an additional 3.03 total quantity sold per month, shortly followed by Paper (3.00 additional sales per month on average).



Good news for the company, all product subcategory sale quantities are increasing over the months. The slowest increase was observed for machines (0.118 / month) and for copiers (0.137 / month).

For monthly total profit, the highest growth slope is for product subcategory Copiers, however the large slope might only be driven by a few outliers (as shown in the graph below). This is also indicated by the just about significant p-value.



The most significant profit increases (p-value of 0.000) were for categories Accessories, Art, Furnishings, Paper and Storage. The trendline was downward sloping for only three categories: Machines, Supplies and Tables, but the trend wasn't significant for either.
The same analysis was also done for average monthly discounted profit margins, however, no significant trends were found for either of the product subcategories.

Suggestions and points for discussion

Based on the above, here are my recommendations and further questions for the company:

1) Stop giving out discounts that drive product profit margins below zero (134 products have had average negative profit margin) and increase your average profit margin after discounts (the average was hovering around 1-4%).
2) If your salespeople give out discounts, it should be related to the quantity the buyer buys - discounts were only (just about) significantly associated with quantity in one of the subcategories: Chairs.
3) Stop giving such high discounts (70% off was given on 418 orders and 80% off was given on 300 orders, together representing over 7% of the orders). Given that none of the products are perishables this high discounting is highly unusual.
4) Start charging people more if they buy more (there was no significant relationship between sales revenue in orders and quantity in orders even when broken down for some of the individual products!)
5) Stop pricing inconsistencies: the most shocking example was for product OFF-BI-10001249 whose minimum list price was $10.4 and maximum list price was $1542.5: a price increase of nearly 150 times within 4 years!

6) Do more customer research to avoid inconsistencies in profits recorded for similar sales. One example is below. It is hard to imagine that the profit increased by 500% in a matter of a few months for the same product, in the same state, sold to the same customer type, in the same quantity, for the same discount, in the same year.

	Order Date	Product ID	Customer_type	State	Region	Quantity_sold	Discount_applied	Profit_on_sale
814	2019-08-19	TEC-AC-10001838	Consumer	California	West	2	0.0	614.0
2730	2019-04-07	TEC-AC-10001838	Consumer	Washington	West	3	0.0	357.0
6633	2019-01-11	TEC-AC-10001838	Consumer	California	West	2	0.0	102.0

7) Focus on what's working: The West region is responsible for highest profits and it has the highest recent growth rate among all regions. California and New York together produce over 40% of profits and each of these states produced more than double the profit overall than the third most profitable state. In California, paper, binders and accessories lead, whereas in New York binders, phones and machines are the top profit-generating products.
8) Customers: Investigate how some customers are losing thousands of Dollars for the firm and why some are getting more than double the average discount for their customer category.
9) South is the only region where profit has decreased from 2018 - 2019 driven by decrease of profits in the Furniture category as well as in the Consumer and Home office customer categories. What happened there? Is the company divesting from that region?
10) Look into why profitability doesn't increase in the South over November and December unlike the other regions. There might be a cheap way to increase profitability there too.
11) Why are Consumers creating so much profit in March? What is special about that month? Why are binders so popular in March?
12) Why was there a large drop in profit from February 2020 to March 2020? Did the business change something or is it related to the effect of COVID?
13) Why are some product subcategories not sold in certain states? Is it the lack of advertisement or the lack of demand there or is it difficulties related to product delivery?
14) How can machines be super profitable in New York but very much lossmaking in Ohio?

Appendix:

The dataset comprises of 9994 rows with each row corresponding to a unique Order ID. Each order corresponds to a type of Product sold to one single customer ordering from one City from the United States. The first order was placed 2016-Jan-08 and the last order was placed 2020-April-06.

Each order contains various amounts of the same product with quantities ranging from 1 to 10 (average 5.47). Each order may have been discounted with discounts ranging from no discount to 80% (average 15.6%). The revenue and the profit on the order was also included in the table (currecy units were not provided but we will assume they are in Dollars): revenues on orders

ranged from $39 to $226405 (average $2806), whereas profits ranged from $-6568 to $8414 (average $49).

In the dataset there are 793 unique customers. Each customer falls under one of the following categories: ‘Consumer’ (total: 409), ‘Corporate’ (total: 236) or ‘Home Office’ (total: 148). On average, customers bought just under 13 times (to be precise 12.7 times from Consumer category, 12.8 types from Corporate category and 12 times from Home office category) with customer ‘WB-21850’ placing the most orders: 37.

In the dataset there are 1862 unique products. Each product falls under one of the following categories: ‘Office supplies’ (total: 1083) or Technology (total: 404) or Furniture (total: 375). The most frequently bought product was OFF-PA-10001970 (category: office supplies, subcategory: paper, 109 total sales from 19 orders) and FUR-CH-10001146 (category: furniture, subcategory: chairs, 99 total sales from 15 orders).

Based on the available information, we can add additional columns to the dataset to help with future calculations. Let’s start with the product price: in general, the price of the product offered to the customer at the time of the order must be the sales revenue from that order divided by the quantity sold.

Discounted product price = revenue / quantity

In this particular case, we also need to undo the discount offered to the customer: if the customer received a 40% discount, it means that they paid for 60% of the original price, so we need to divide the previous result by 0.6 to arrive at the original product price.

Product price = revenue / quantity / (1-sales discount)

The income per product sold from a particular order must be the profit on the sale divided by the quantity sold. Note that here the **product income includes the discount given** - it doesn’t really make sense to undo the discount here and the discount could be the reason why the order was placed in the first place.

Product income = profit / quantity

Now, let’s calculate the cost of making the product. By definition, the cost must be the difference between the revenue and income. After dividing both sides by quantity, we get that the cost per product on the order is the following:

Product cost = Discounted product price - product income

Now we can calculate the margins. The original product margin (before the discount) is by definition the

Product margin = (Product price - Product cost) / Product price

The product margin after the discount can be calculated as:

(Discounted product price - Product cost) / Discounted product price =

Which, can be rearranged as:

Discounted product margin = product income / discounted product price

Code used for analysis and plotting: Final.ipynb