Python for Business Analytics ECO 32500 Fall 2024 Final Project Report

Enmanuel Curiel
Prof. John Droescher
November 2024

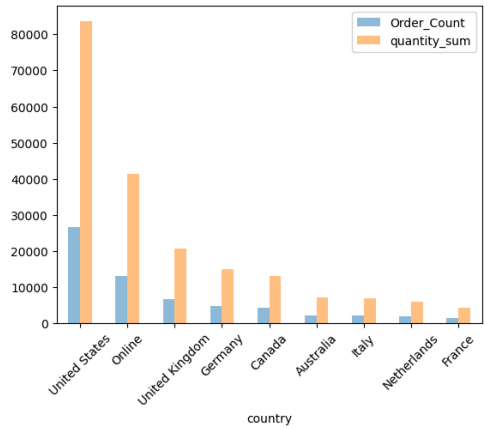
I was given the following problem:
“I am your VP of Sales and have presented you with the following question: I am looking to improve our performance in the next quarter, what suggestions do you have?”

Understanding the question:
To fully grasp what the VP of Sales is asking of me, I need to reframe the question in a different way. A suggestion to improve performance in the next quarter can mean a multitude of things. The way that I understood it is, How can we see improvements in our sales of the company for our next quarter, what did we do right or wrong that have led us to this point. At what point did our company shine the most in terms of sales performance, and at what point did we see our company perform the worst. Comparing these events can tell us where we might’ve gone wrong but also where we can improve and see good results.

Data collection:
With the Visual Studio code program, data collection was done by connecting to the MSSQL server. With integrated extensions to support python and its numerous libraries. Writing queries and extracting the data that was deemed necessary. And using python with the help of pandas to help organize and visualize the data.

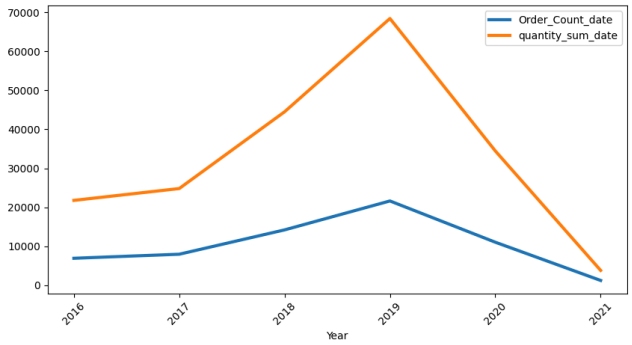
Using a simple SQL join query I was able to join sales and stores on the matching column SoreKey and only use the data that I needed to find the relationships.

Data organization and relationships:
What I first wanted to look at was the relationship between the amount of order numbers ever sold and ranking them by country. Online was included in this to also see its effect compared to the rest of the world. To even further magnify this data I also added the quantity sold of all of the items in all countries and ordered them by smallest to largest.



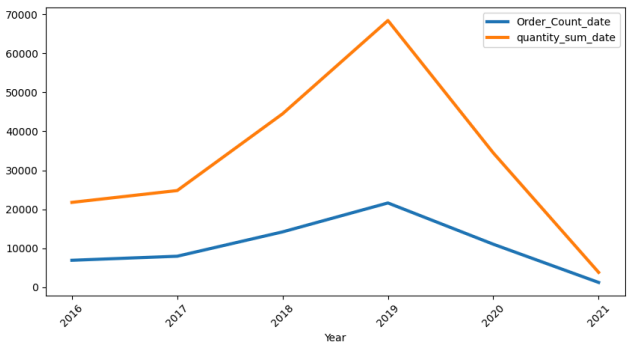
What was found was that the United States took a large margin compared to all the other countries placing it as the capital of sales. It's also noted that naturally the quantity sum will increase as the order count also increases as you cant have one without the other. What also heavily piqued my interest was that online took a huge compared to all other countries besides the U.S. Putting it at just second place this goes to show the huge market that online has as it can be counted from any country that the order was placed in.

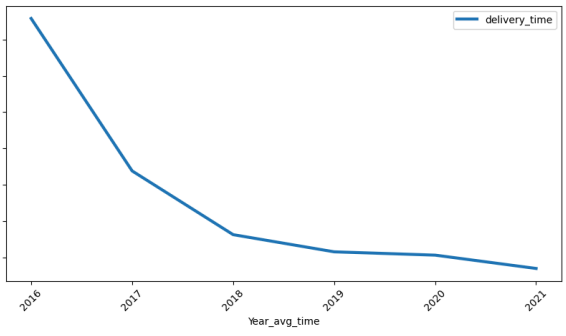
Continuing with the trend of order count and quantity sold, I wanted to look at how that performance was shown with time instead of by location. Creating a line graph with the goal of seeing how order counts dipped and grew over the years these are the results:



This line graph seems simple but actually highlights a lot of what the stores may have been performing poorly on. WIth this graph we can see the highest order counts and quantities sold by year. The most evident point here is that 2019 saw the biggest spike in sales. Following that we see it fall to how it was before pre 2019. Another important note here is that the graph may seem a bit misleading as there is only two months for 2021 making it seem like 2021 had even worse performance of sales which is not the case.

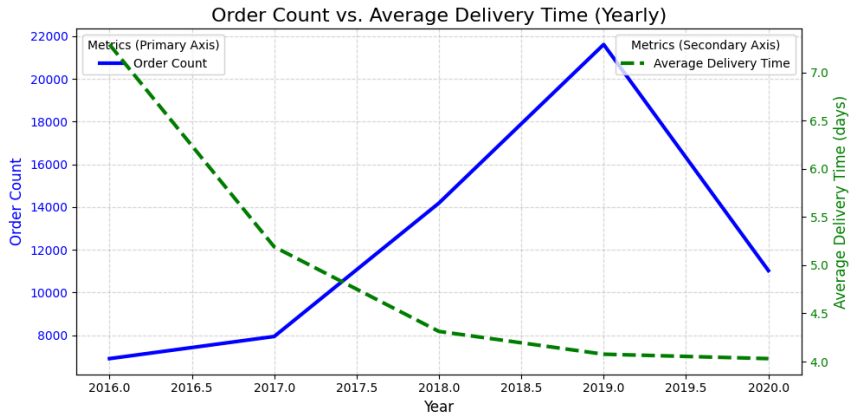
A relationship can be found by comparing the average delivery time to the order and product count. Using a line graph would allow for the concept to be more easily digestible to the eye and show the clear relationships these two graphs share.



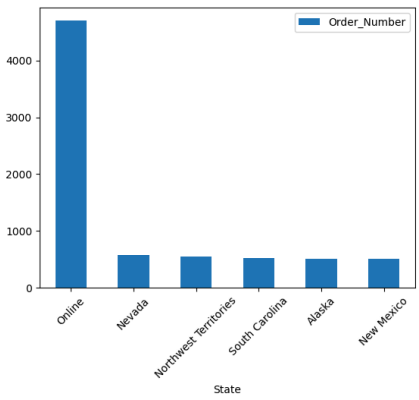


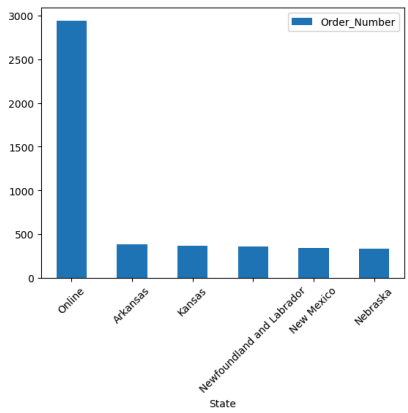
As can be seen with the graph on the right, it displays the average time it takes for a delivery to be completed by year. We can compare that to the graph on the left which displays the earlier line graph of order and product counts. The relationship that is of most importance here is the big order jump seen from 2017 to 2019 also coincided with the decline in average delivery time. It is also important to note that the graph may seem misleading as the year does not have all its data collected for the entire year, however the relationship is still the same.

Overlaying the two graphs above also show that relationship much clearer:



I also wanted to take a look at top 5 states including online sales in the years of 2019 and 2020.

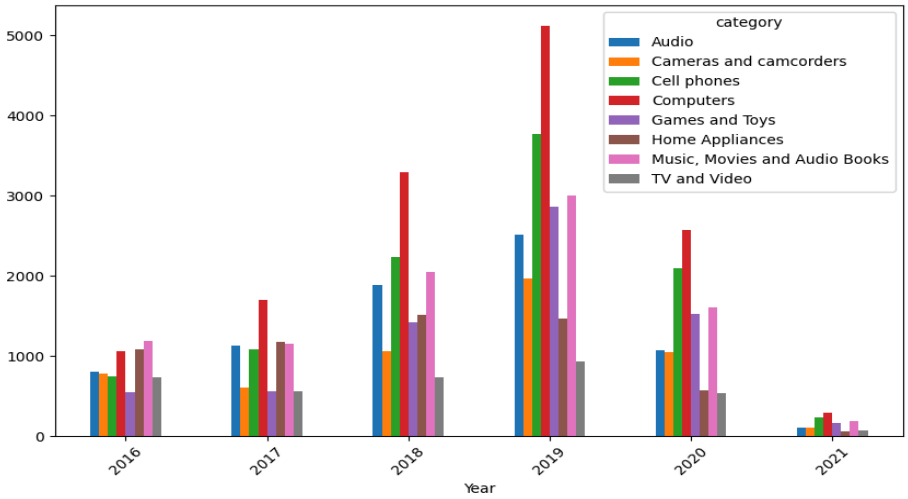




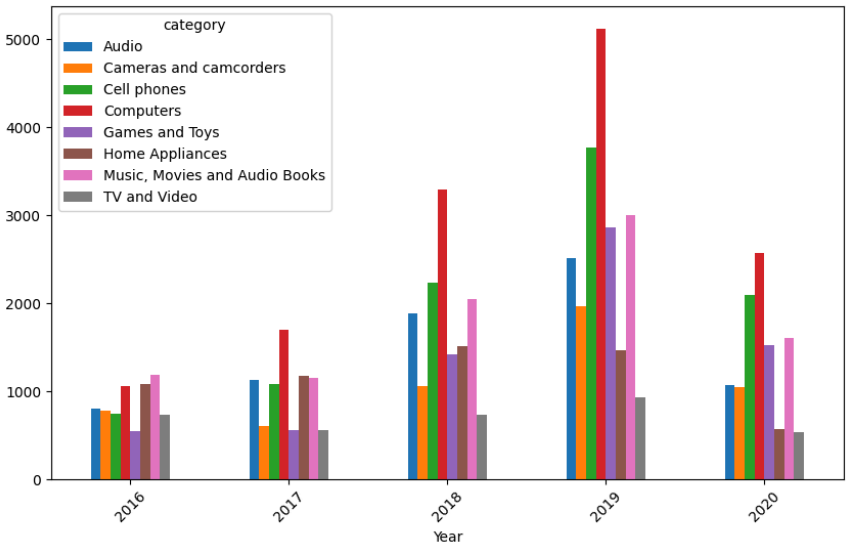
In terms of what can be found most notable in these bar charts, online still held a sizable lead compared to all other countries found in the data pool. The other top 5 states also hold a similar position with each other with no notable gaps in between them. I decided to turn my attention to the products that were being sold.

Using another simple join query I was able to make an inner join with the table sales and products on product key and only keep the information that I needed.

With this specific data set I was able to construct A pivot chart that told me the performance of all the category of items throughout the years. This chart would be able to tell me exactly which category of products are underperforming and which one of them are overperforming.



This bar chart displays a lot of information all at once, but I made the same mistake here again by implying that 2021 is an underperforming year when instead the year has only just begun. A way to fix this is by filtering out the year 2021 as shown below:

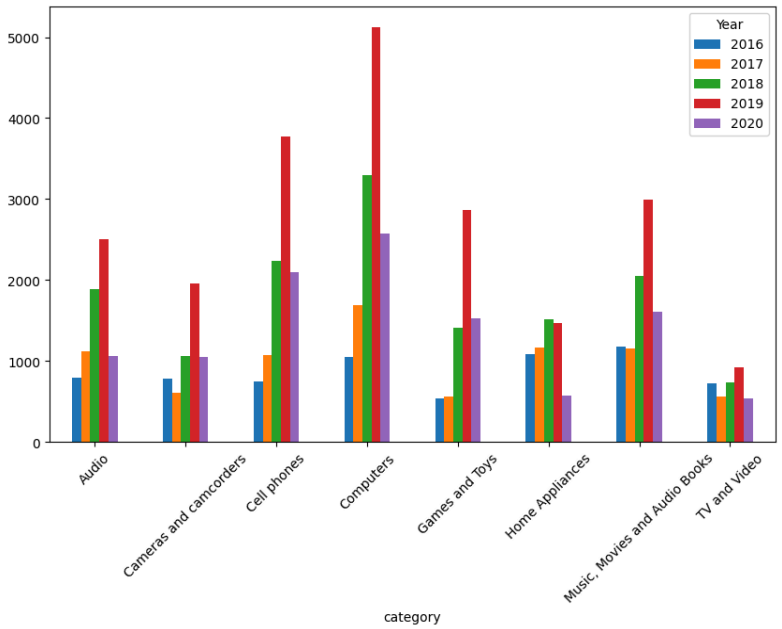


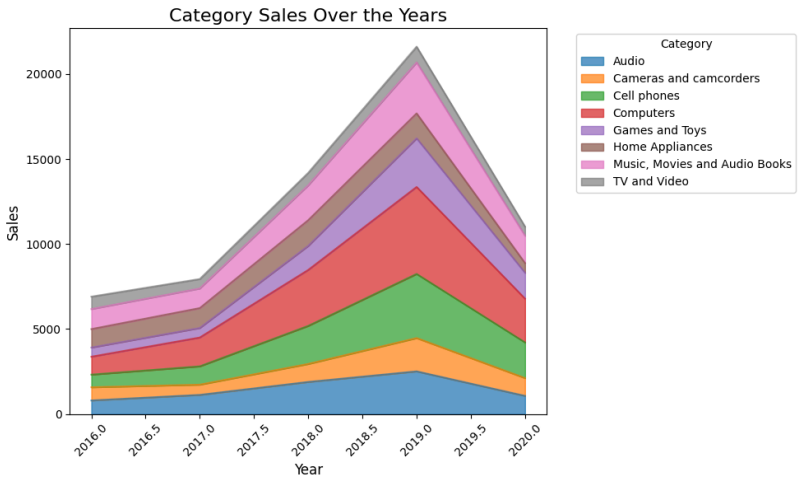
Nevertheless we will take a look at the years 2016 through 2020 which have all of their data recorded in the database. Because there is so much information in this chart, there are lots of relationships that can be of note. The first of that being 2019 holding the number one spot in all the products sold by category. This is a given as we saw in the

other charts before that 2019 was when the company had the best performance. It is still important to see that this trend is still very prominent. The category which saw the biggest jump in sales and has held the number one spot is computers, second to that being cell phones. We can also see how even the category Music, Movies, and Audiobooks has once held the number one spot in 2016.

These three categories tell us that categories pertaining to technology have always been at the top of sales even after the company's decline in sales. The demand for technology has been high for all the years however with one clear exception, that being TV and Video. In 2016 TV and Video did not have gapping performance to all the other categories, but as more products were being sold that gap widened to a much bigger degree and lowered it to the least amount of products sold. Less people are buying TV and Video more than ever and have moved to other more easily accessible alternatives like Computers and Cell phones.

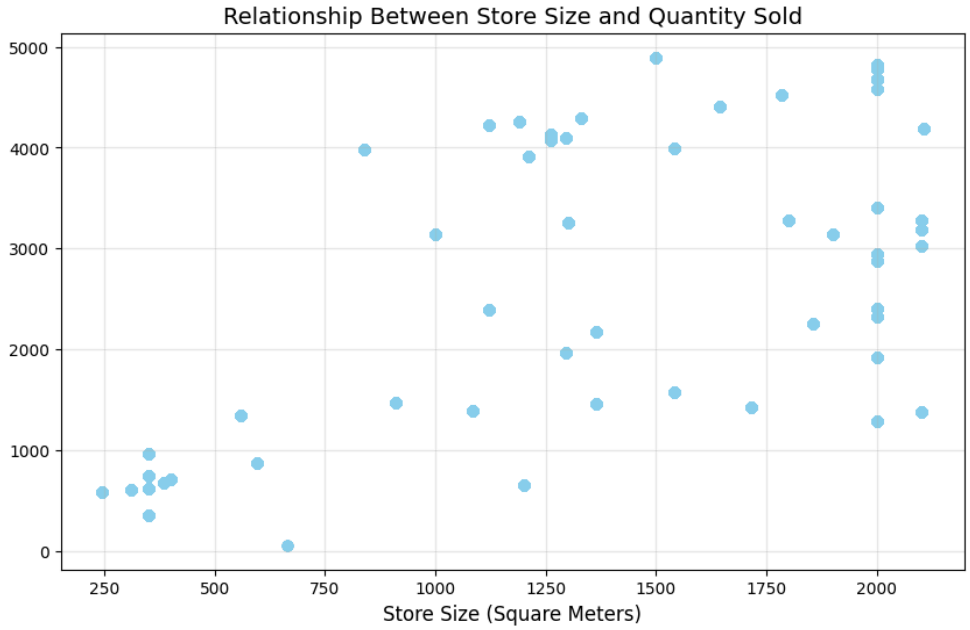
Similarly we can organize the pivot chart in a way so that the categories are on the x axis making it so that we can see the categories progressed over the years:





Another way to represent this data more clearly is by using a stack plot which illustrates all the category sales over the years but in an area depiction. Which more clearly shows how each category performs compared to the rest We can see how the categories that are technology based such as computers, cell phones, and audio hold the top position.

The last avenue that I want to explore in the data was how much space each store took and comparing that to the amount of quantities they have sold. In other words, the square meters and the total quantities sold. Firsty, I created a scatter plot comparing the square meters of each story by the total number quantities they have sold:



This scatter plot tells us that store size and quantity sold do have a positive correlation. That seems very plausible as bigger stores tend to have products available to sell. The correlation coefficient is .37 telling us that there is a less than moderate positive correlation between the two. What is important here, however, are the outliers that are present. Some stores take up a lot of space but are selling less than stores who don't take up as much space. Some stores are underperforming and may have negative effects on the performance as a whole.

To further illustrate this issue I devised a performance metric which is the total quantities sold divided by the meters of the store. I deemed a store as “underperforming” if its performance metric is less than the average and if the square meters is higher than the average. What this achieves is as a store's total quantity increases naturally their performance metric will increase relative to their square meters.

This is not to say that the stores that are highlighted in this grouping are actually underperforming but that these stores are performing less than desirable compared to other stores.

The black line in the middle represents the average performance metric of all the stores. The blue bars are the stores who are performing well enough compared to other stores. The red bars are the stores who compared to other stores are not performing as well. Stores that have more square meters should perform to a certain metric, however there are stores who are not meeting these criterias and are falling behind.

What can be concluded:
2019 has seen the most amount of sales ever made, its lead up to that point is even more important. The rise in online orders and deliveries has shown that they have contributed some part in the performance found in these years. Technology as a whole has also seen rapid growth over the years compared to all other items that have been sold. Factors in technology growth over the years could set that spike in motion. Some stores do not follow in that demand and have been underperforming compared to their sister stores. Whether that be not selling enough items or being too resource heavy because of their size,There is a disconnect highlighted in these stores specifically. Regardless of all of that online shopping has still held its own compared to the rest even after the massive drop off found in 2020 it still held its own weight and garnered the number one spot compared to in person stores.