Theran Meadows

Milestone #4 2/11/2024

DSC 630 Predictive Analytics

Introduction

Predicting sales of products by geographical region can be a powerful and valuable tool for companies. Businesses, and shopping stores, depend on knowing what products will sell more at different times in the year. They must act accordingly to keep up with their supply to make sure that they have enough for the demand. There must be a balance as well between the supply and demand as too much supply leads to wasted product and lower value of sales. Too much demand for not enough products leads to bad customer satisfaction. Customers may go elsewhere to find what they want. The ultimate question I wish to answer is:

What products will sell the most for the next year or maybe 5 years.

I think answering this question can lead to further investigation into what advertisements would need to be planned and what marketing strategies can be implemented. (these are ides for future projects though)

What type of model or models do you plan to use and why?

The type of model I wish to use is a regression model. I would like to see what sales can be predicted to increase or decrease in a given year depending on past sales.
How do you plan to evaluate your results?
I plan to evaluate results starting with simple exploratory data analysis. I would like to see what products were the most profitable over the past years. Simple bar charts help with this step.

I would then like to split the data into training and testing sets where hopefully, we can see the outcome of what product (s) would be selling the most in the next year or maybe even 5 years.

What do you hope to learn?

I hope to learn more about prediction models and see more of the process of deciding whether your results are accurate or not.

Assess any risks and ethical implications with your proposal.

Prediction analysis is simply just that. It is a prediction, not a fact of what will happen in the future. There are many factors that can affect the results of the prediction in real life such as natural disasters, economic crises, etc…

There could also be problems that arise from my own analysis. What if my data doesn't have enough to make a solid prediction? What if the data is too recent or too old? It all affects the accuracy of the prediction.

Identify a contingency plan if your original project plan does not work out.

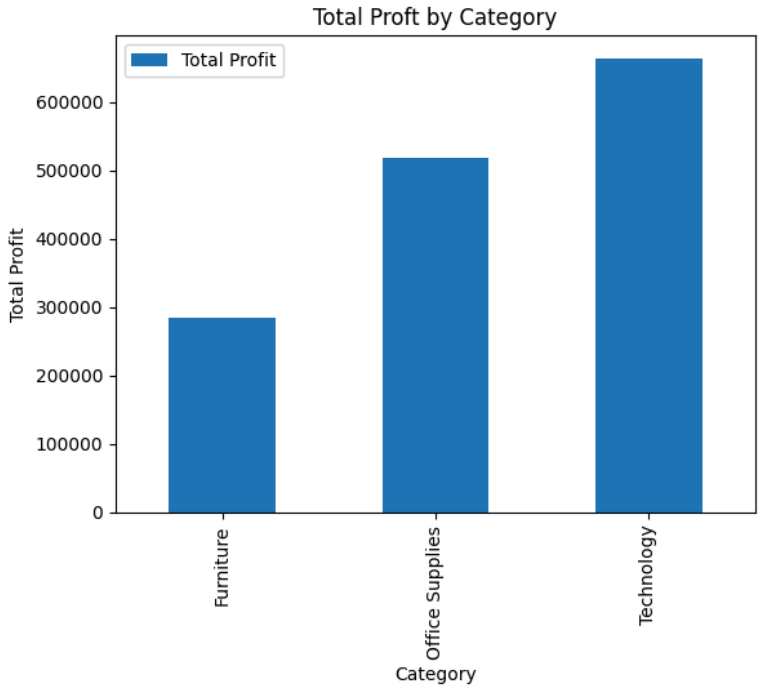
A contingency plan I have is to change or data other data sets to my original. There are plenty of Kaggle that I have already chosen as back-ups. The one I chose now contains 27 columns including Country, Year, Product Category, Sales, and Profit which were all columns I was looking for in a data set to contain.

Include anything else you believe is important.

A backup question I have is who is shown to be the most valuable customer or customers. My data set also contains a Customer name and ID column to see all their orders they have done in a given time period. I think it might be interesting if there was some type of “Rewards for Valuable Members” to earn more of their trust or to even help bring in more customers.

Will I be able to answer the questions I want to answer with the data I have?

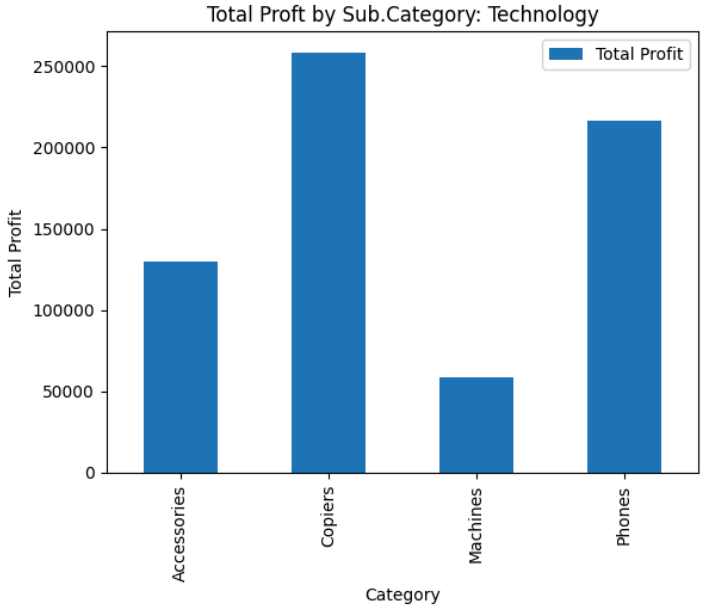
I believe I will still be able to answer the questions I have of wanting to determine what products will have the most sales in the next year. From my preliminary analysis, I have found that there were 3 categories that the Superstore has made to categorize sales. See figure below:

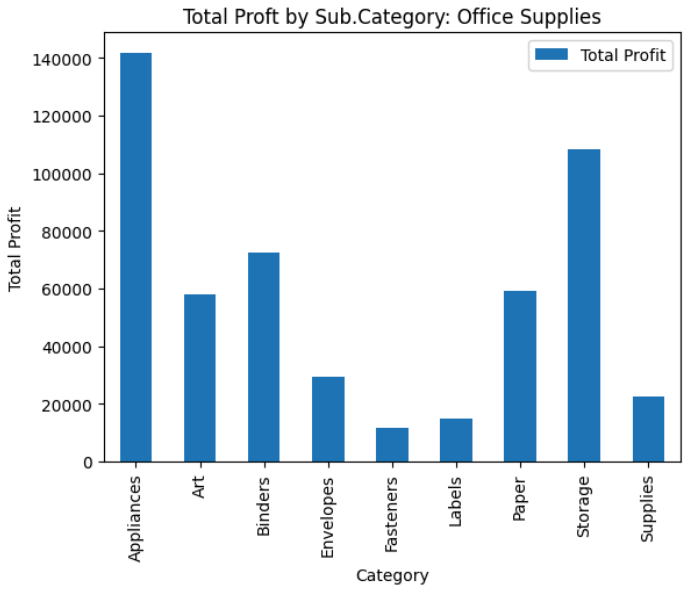


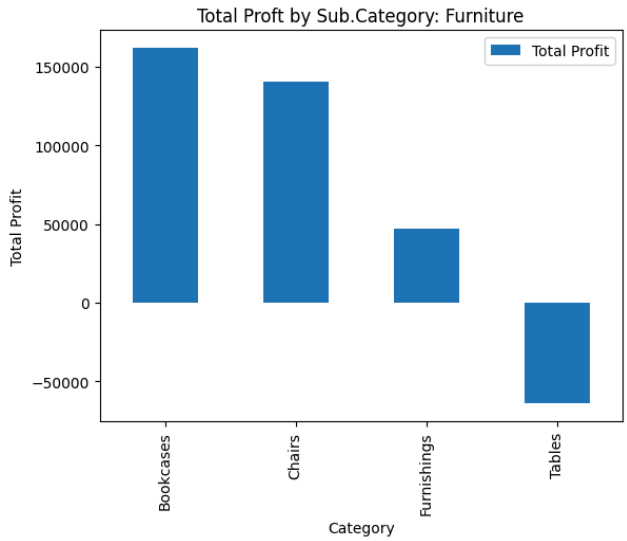
Based on the results shown, it is obvious that the technology category is the most profitable. Not to exclude the other two categories though as they will still be included.

What visualizations are especially useful for explaining my data?

The most important visualizations are the subcategories within the three main categories that give a further insight as to what is the most profitable by category.







Do I need to adjust the data/ and/or driving questions?

I don't think I need to adjust my question. Based on my preliminary analysis, the columns within do show timestamps of when products were ordered. Using the Category sales data from above, I should be able to make another visualization depicting when technology sales are most active in the year.

Do I need to adjust my model/evaluation choices?

I have not fully decided on a model that I want to use. I may use a random first model along with another model to try and validate one another.

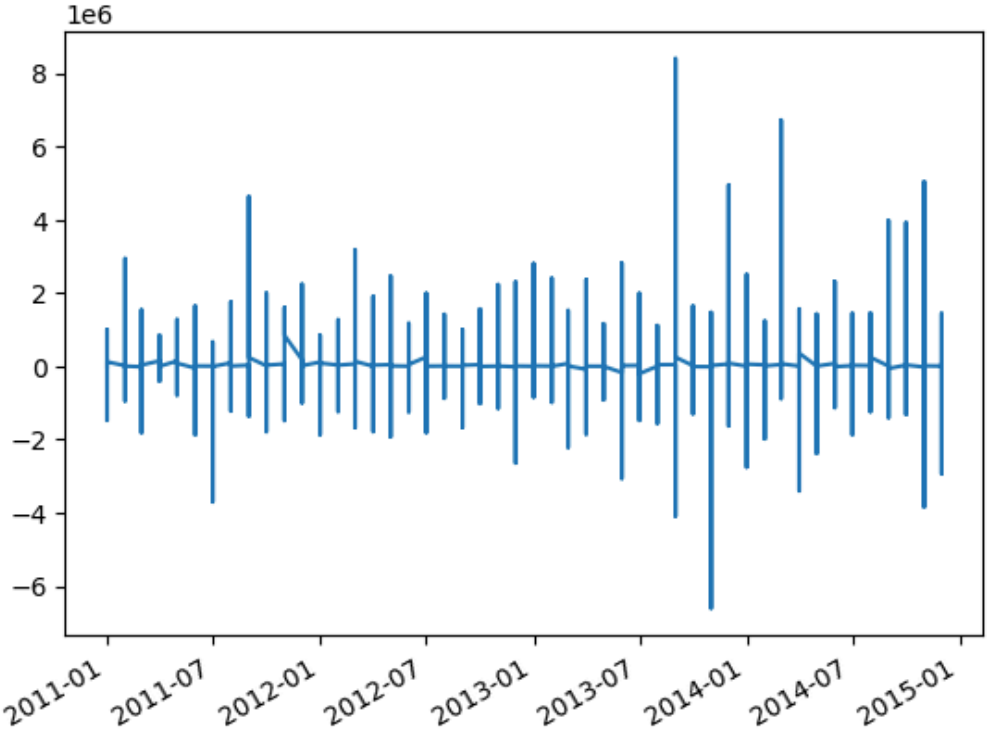
Are my original expectations still reasonable?

I believe my original expectations are still valid and using this data set, I should be able to answer my questions and get the proper prediction I was looking for.

Data Preparation

For preparing my data, I completed some steps to make sure my data would be readable, splittable and ready to be visualized. The first thing I did was remove columns I know I will not need. There were about 12 columns I removed. I then made sure that my ‘Profit’ column was in the right units of measurement by multiplying it by 1000. I then converted the ‘Weeknum’ column into the month the week number fell into. After assigning numbers to the 12 months, I created a ‘Date’ column with the ‘Month’ and ‘Year’ columns.

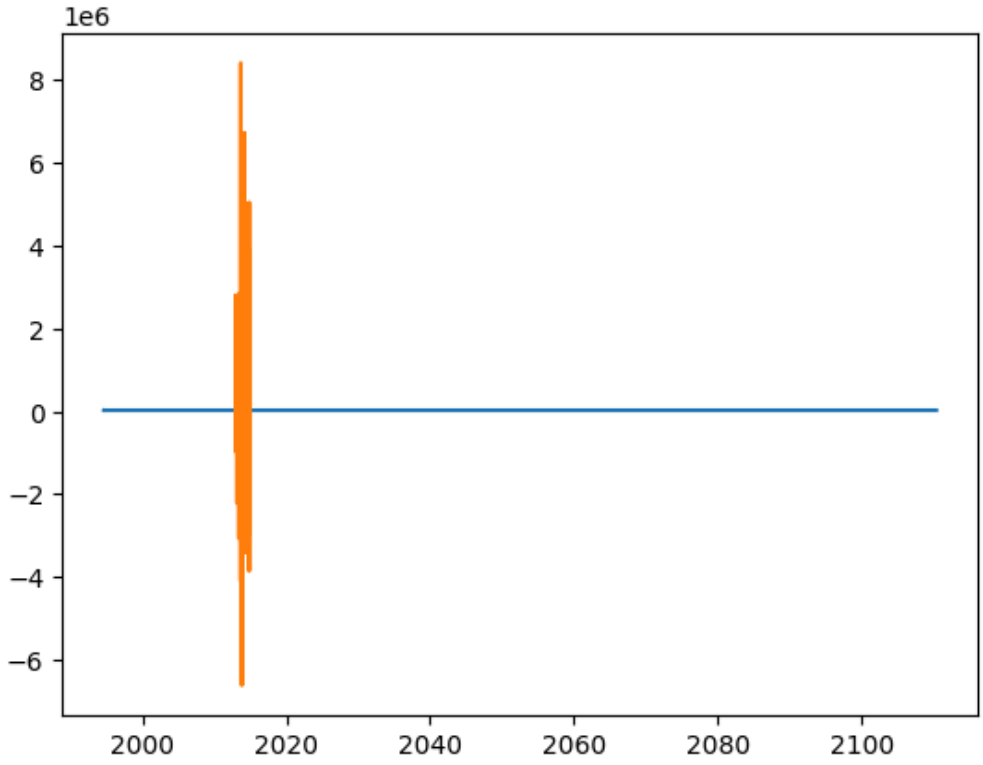
I discovered some problems that I will have to come back to address which was that my data may need to be summed to the proper months to show visuals better. Below can be seen what one of my visuals looks like for the data over time. As you can see, it is very hard to read and understand.



Model Building and Evaluation

The model I chose to do was a predictive time series model. Because my data spanned 5 years of data I started with using the last year of data as my test and the rest as the training. I soo found this as a problem because as it turns out, my data does not have a good distribution across the years. For example, my data has over 50k rows with the year 2011 having only 11k rows. The year 2014 has nearly 18k rows. This makes the test data have a larger portion of the data than it should.

I am having another problem with my predictive results spanning over 100 years when I only want the next year or two. Below is another visualization that is needing work.



With these errors happening, I don't feel ready to make any sort of conclusion or

recommendations. It is still a work in progress.

What I hope to do to fix the errors I am having is find a good splitting point for the data. I can see the different proportions that each year of sales has or I can sum the months together per each year. I may do the latter because that would at least give me the same number of rows each year.