Lucy Moyes
DSP 539: Big Data Analysis
May 8, 2020

Forecasting Analysis

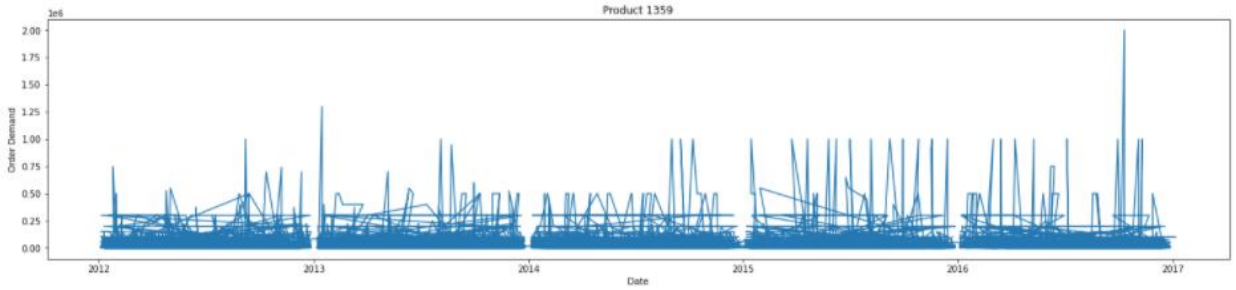
Final Project

This data set, Historical Product Demand .csv file, was downloaded from kaggle.com. It includes orders from six years of historical product demand. From some initial exploration of the data set I found that this company shipped 2,160 unique products across 33 different product categories to four different warehouses – all making up over one million orders between January 8, 2011 and January 9, 2017. The products have numerical codes so we do not know what they are. I want to use this data to be able to forecast a how much of a specific product to order in a specific month. To organize the data, I initially arranged it by ascending dates.

The data set is very big and when I first tried to plot it, it took very long to run. I realized that I would have to filter it down in order to analyze it more clearly. Since I wanted to forecast a specific product, it made sense to filter down to one single product. To figure out which product was the most popular, I used a command to sum all of the order demands for each unique product codes. I ran into some issues here because the items in the "Order_Demand" column were objects and not listed as integers, which is necessary to carry out a sum. I also found out from the Kaggle readme file, linked to the data frame, that some of the order demands included parentheses. This meant that I had get rid of the parentheses and change the "Order_Demand" data type. I used two str.replace commands to replace the "( )." Then I used astype('int64') code to change the order demands into integers (kashdotten, Kaggle). I then used groupby, sum and idmax commands to find the most popular product, which ended up being "Product_59."

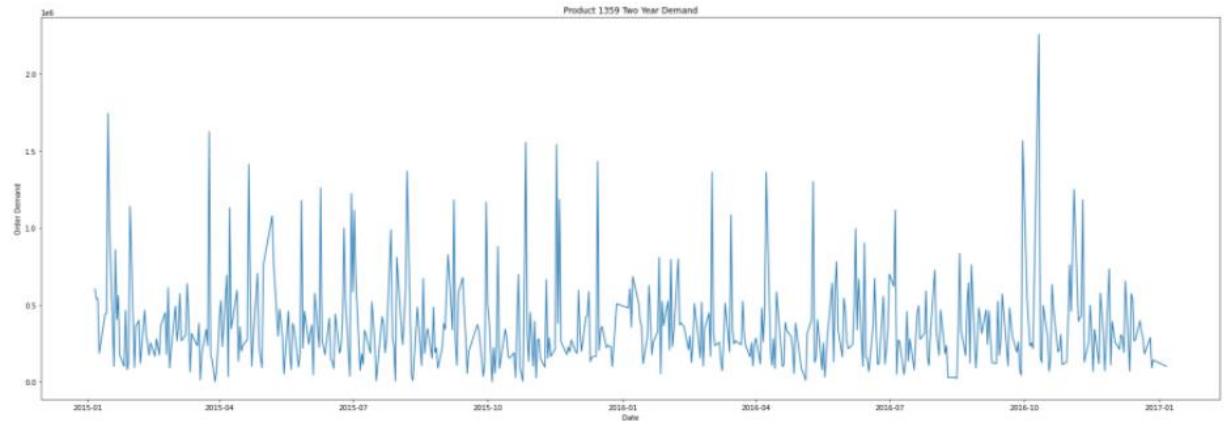
I created a subset dataset which only included this data pertaining to product 1359, I named the new dataset "product_1359." This data set now included 16,936 product orders, with

the same number of rows. I plotted the date and order demand of Product_1359:

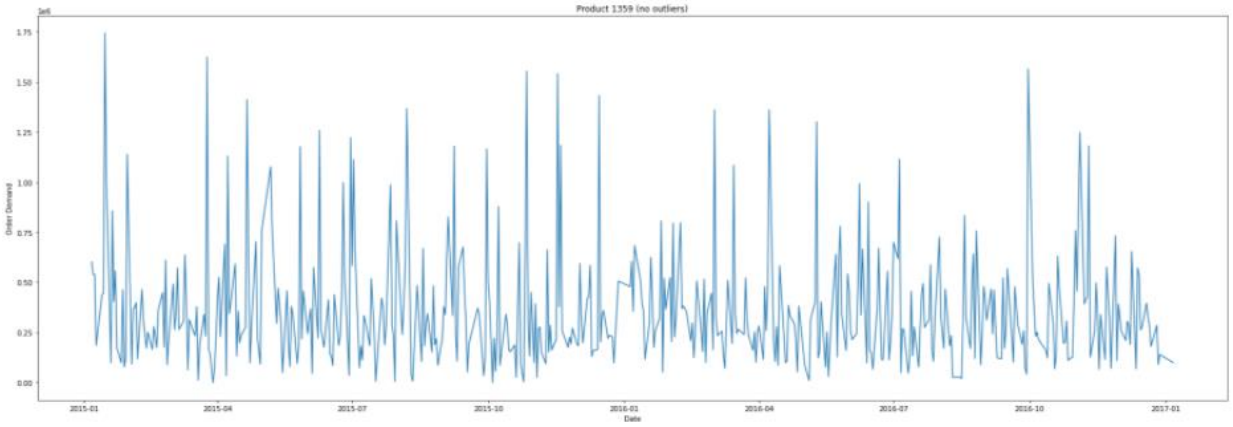


You can see that there are too many data points here to really decipher anything. It also seems as though they have not been plotted in order. Product 1359 has been ordered between January 5, 2012 and January 6, 2017. Five years of data is plotted above. The most relevant information in the dataset, for what I am trying to do, is the date and the order demand, so I got rid of the "Product_Code", "Warehouse" and "Product_Category" columns and made the "Order_Demand" into a daily sum, using the groupby, sum and reset_index commands (kashdotten, Kaggle).

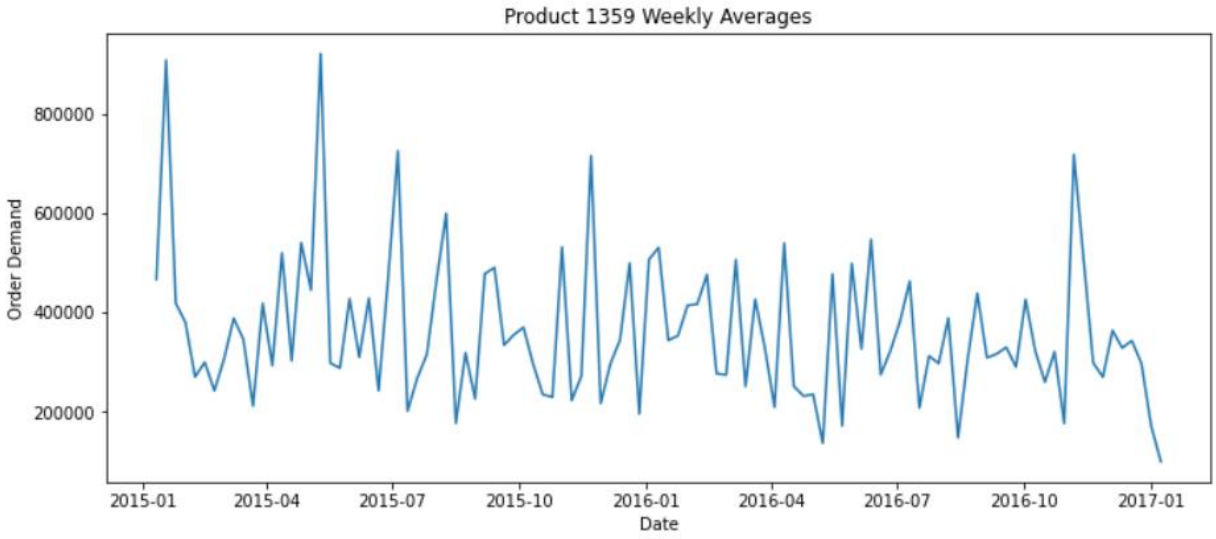
Typically, in business, buyers and planners can use two years of historical data to try to accurately forecast for the coming year. I narrowed the data set down to include only orders from 2015 and 2016 and renamed the new data set "two_yrs_1359." This is what a plot of that data set looked like:



As you can see, there is an outlier of over two million in order demand around October of 2016. I wanted to see what this data would look like without that data point. This is it:



Next, I wanted to see if this might look clearer if I was able to plot the average of each week across these 24 months. In order to do this, I set the date as the index, instead of the row numbers which were the index from the original data set. I used a set_index command (kashdotten, Kaggle) and then a resample and mean command to manipulate the dataset onto weekly averages of the order demand. This is what it looked like plotted:



There seems to be a peak in order demand about once every quarter. It looks like orders for this product have gone down in 2016. This company should consider regular order deliveries of the same amount rather than so many large and small daily orders. After seeing that there is not a clear defined pattern, it might be worth looking at this data with the warehouses in mind, these might have clearer trends. For 2017 I would shoot for forecasting the middle of the corresponding weekly averages of 2015 and 2016, as we cannot count on the demand staying as low as it did on 2016.