Capstone 2: Milestone Report 2

Problem Statement:

Given a data set of one year of sales data, can you make any determinations about patterns within the data? For example: patterns of cancellations by time or product, customer groups most likely to purchase a given product, or times that more customers tend to purchase products.

My client would be the company's marketing department. This data and analysis would potentially allow them to market different products better based on the propensity of customers to purchase. For example, this could manifest by strategically changing the price of a given product at a specific time or offering sales to repeat customers.

Obtaining, cleaning, and wrangling the data:

I was able to download the sales data from a UK-based store from the UCI archive. The excel spreadsheet is available publicly on their website as a free download at: http://archive.ics.uci.edu/ml/datasets/Online+Retail. From that website, I was also given information about each of the columns:

- InvoiceNo: Invoice number. Nominal, a 6-digit integral number uniquely assigned to each transaction. If this code starts with letter 'c', it indicates a cancellation.
- StockCode: Product (item) code. Nominal, a 5-digit integral number uniquely assigned to each distinct product.
- Description: Product (item) name. Nominal.
- Quantity: The quantities of each product (item) per transaction. Numeric.
- InvoiceDate: Invoice Date and time. Numeric, the day and time when each transaction was generated.
- UnitPrice: Unit price. Numeric, Product price per unit in sterling.
- CustomerID: Customer number. Nominal, a 5-digit integral number uniquely assigned to each customer.
- Country: Country name. Nominal, the name of the country where each customer resides.

I converted the excel sheet to a dataframe in my Jupyter notebook using Pandas and called the usual informative commands (.info() and .head()) to get a better idea of the data in front of me. There was some missing data (customer IDs as NaN) but this was not overtly concerning as I do not plan to look at individual customer data per se. I will have to re-evaluate if it becomes relevant to look at individual customer data.

To allow for clustering later on, I created columns to convert the InvoiceDate to the day of the week and an integer associated with that day (Monday=0 to Sunday=6) as well as a column to convert the InvoiceDate to epoch time and a column to assign a number to each of the countries represented. I did this prior to pulling out any subset of data so all of the original data would be treated equally.

I pulled out the cancellation data (denoted by a “C” at the beginning of the InvoiceNo) into its own dataframe to look at only the sales data related to cancellations. I further pulled the “manual” cancellations from that data, as this appeared to be unrelated to a specific product.

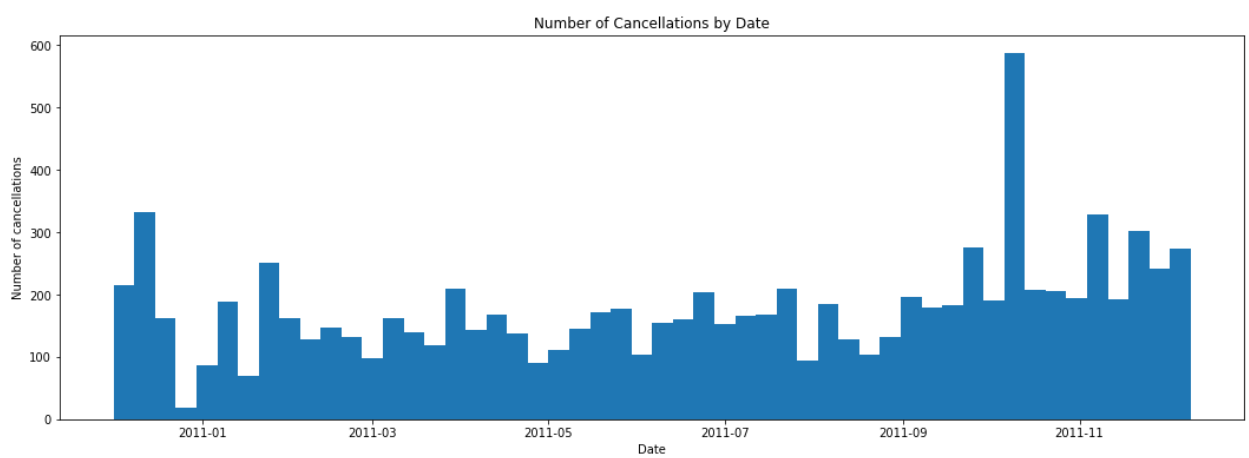
Additionally, I looked at only the purchased products (the full retail data with the cancelled products removed). After dropping the rows containing “adjust bad debt”, amazon fees, postage, dotcom fees, and manual, I was left with over 530,000 data points.

I obtained the holiday data from the website https://www.timeanddate.com/holidays/uk/2011. I created an excel spreadsheet from the available information and converted it to a dataframe using Pandas.

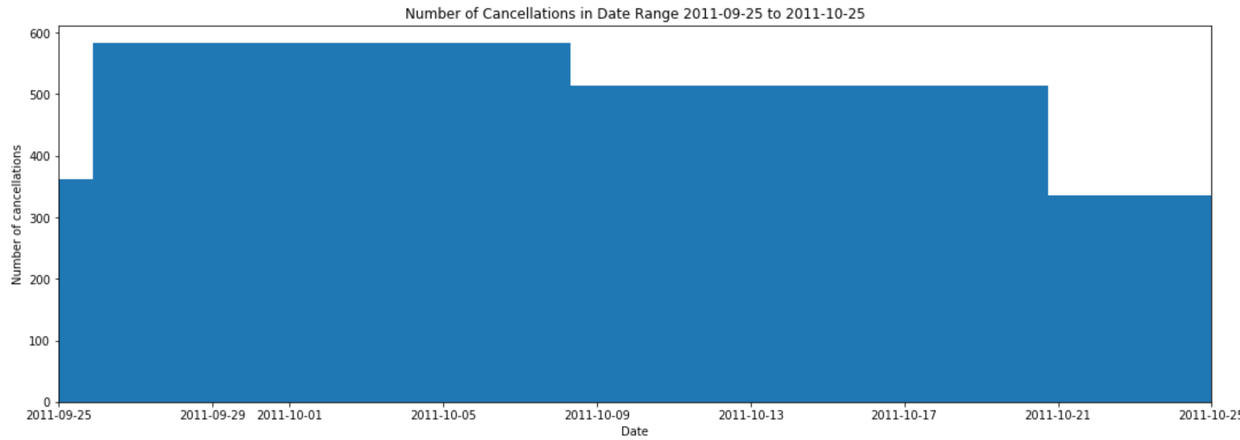
Relevant Exploratory Data Analysis, Statistical Inference, and Clustering:

Cancelled products analysis

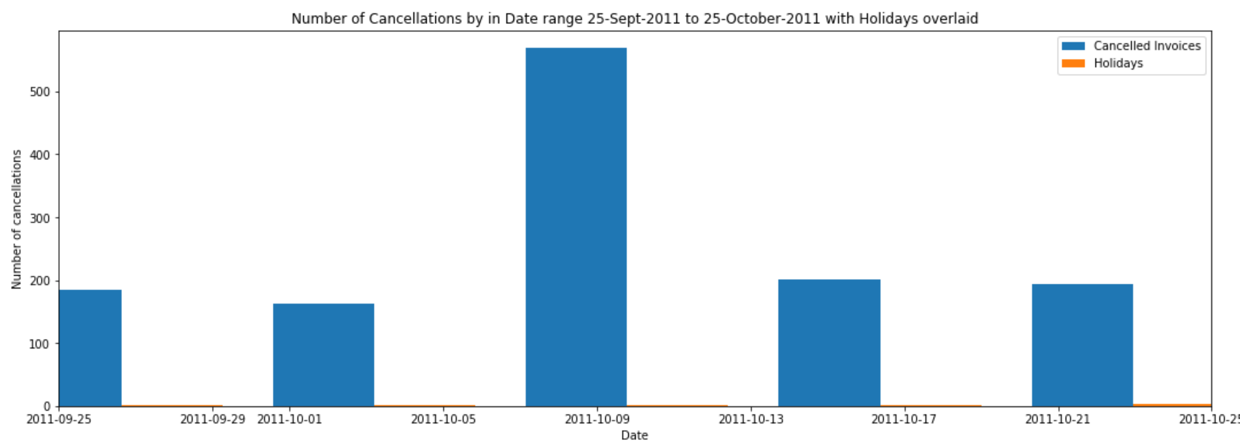
I pulled out the Invoice numbers of cancelled orders into its own dataframe and plotted them by date:



There does appear to be one peak well above the rest at approximately 2011-10. Looking closer, that range appears to be 2011-09-25 to 2011-10-25:



Looking at this with holidays, I got this plot:



The holidays that coincide with that peak are primarily Jewish holidays that are not tied to gift giving:

Date	**Day of Week**	**Name**	**Type**
2011-09-28	Wednesday	Navaratri	Hindu Holiday
2011-09-29	Thursday	Rosh Hashana	Jewish holiday
2011-10-04	Tuesday	Feast of St Francis of Assisi	Christian

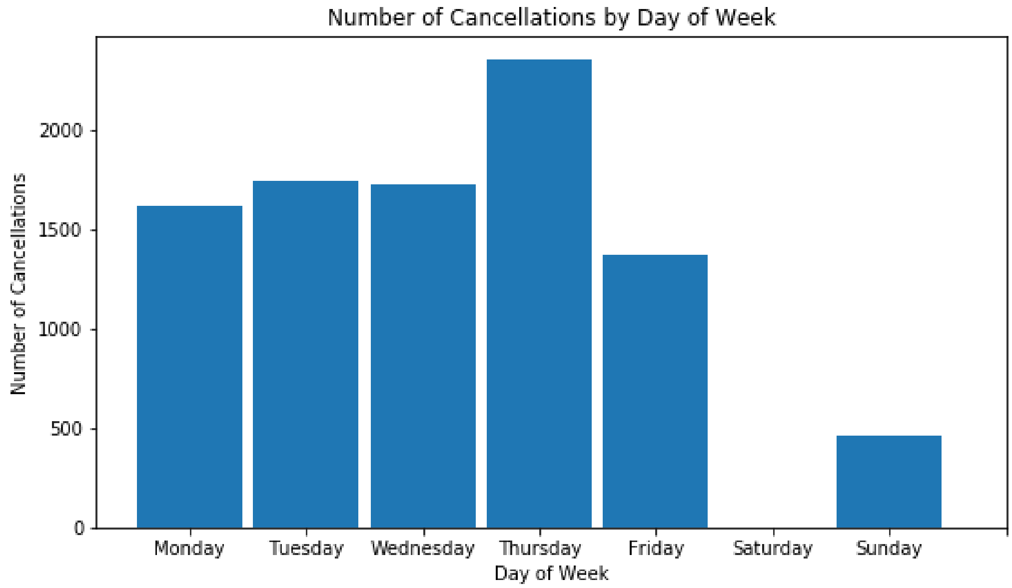
2011-10-06	Thursday	Dussehra	Hindu Holiday
2011-10-08	Saturday	Yom Kippur	Jewish holiday
2011-10-13	Thursday	First day of Sukkot	Jewish holiday
2011-10-19	Wednesday	Hoshana Rabbah	Jewish holiday
2011-10-20	Thursday	Shemini Atzeret	Jewish holiday
2011-10-21	Friday	Simchat Torah	Jewish holiday

I also looked at holidays during the 30 days prior to that surge in returns:

Date	Day of Week	Name	Type
2011-08-26	Friday	Laylatul Qadr (Night of Power)	Muslim
2011-08-29	Monday	Summer Bank Holiday	Common local holiday
2011-08-31	Wednesday	Eid ul Fitr	Muslim
2011-09-01	Thursday	Ganesh Chaturthi	Hindu Holiday
2011-09-23	Friday	September Equinox	Season

None of those seem like likely reasons for the big spike in returns. I completed a brief search for top news stories in the UK around that time and found that unemployment was having “the largest increase in nearly two years” which may be a contributing factor. https://www.bbc.com/news/business-14912236

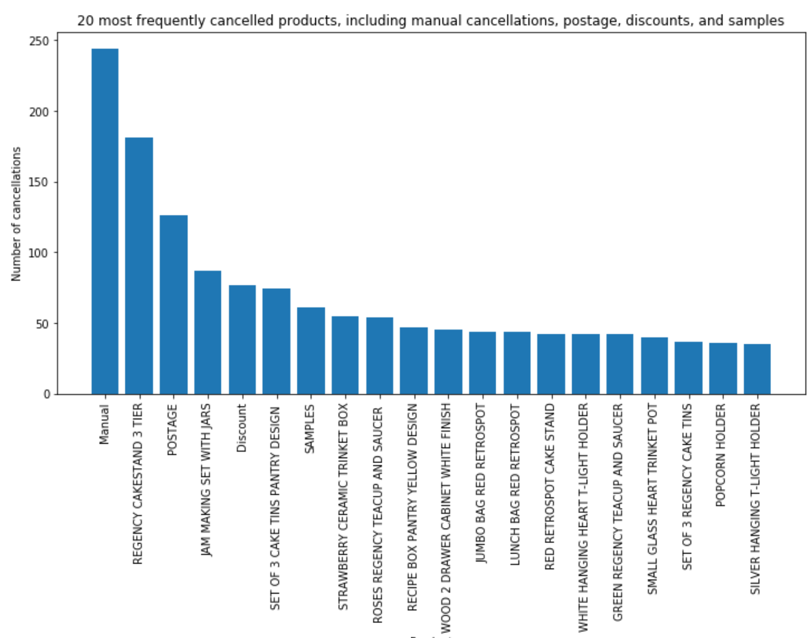
I looked next to see if there was a day of the week that was common for cancellations and produced the following plot:



I noticed that there appears to be a spike on Thursdays and no cancellations at all on Saturdays. (A search using .loc confirmed this).

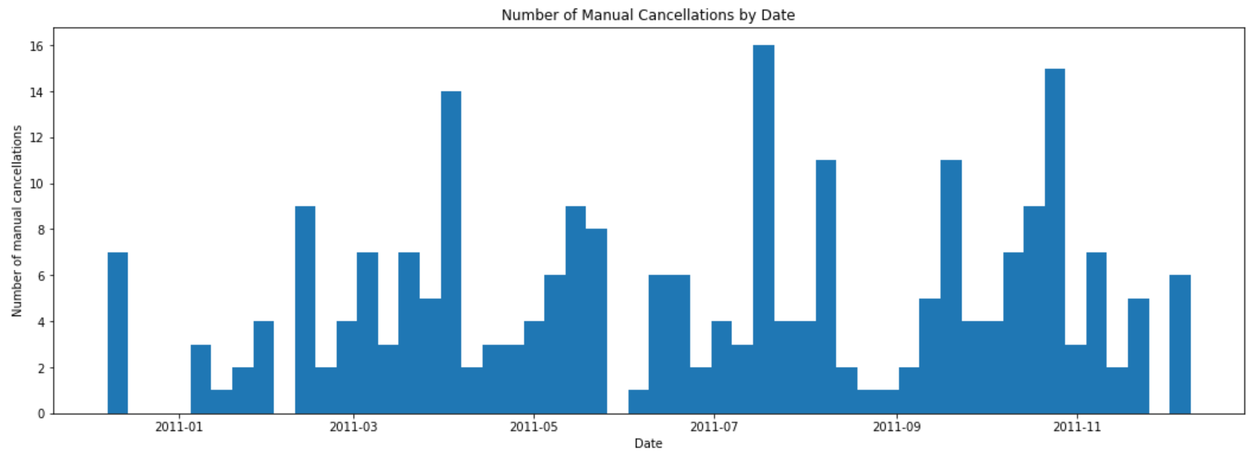
I completed an ANOVA and found that the difference we see is likely not due to chance and is statistically significant.

For a different perspective, I pulled out the top 20 most frequently cancelled products and plotted them:

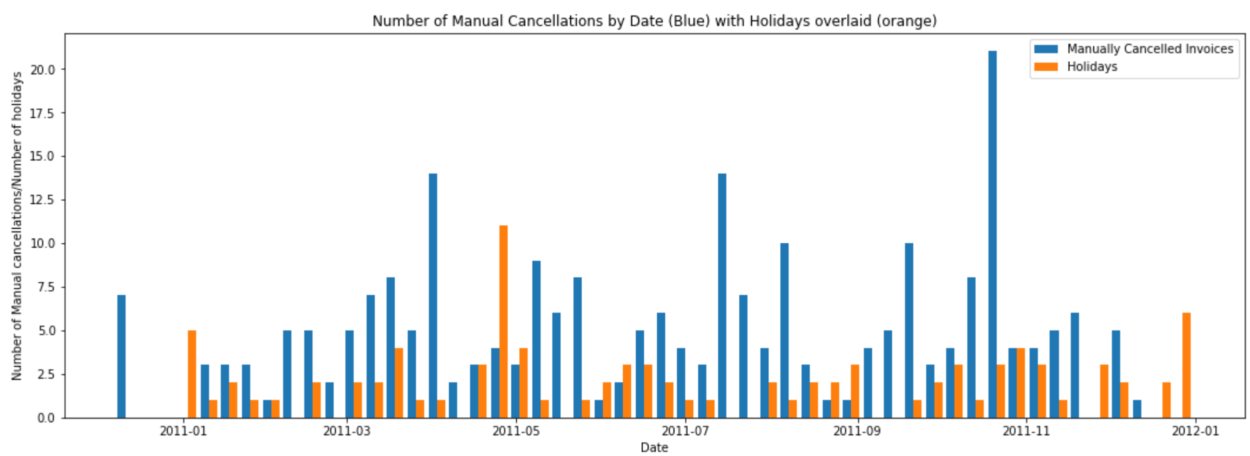


I noticed that 4 of the top 20 were “Manual, POSTAGE, Discount, and SAMPLES”. I then went in two different directions to further explore the data.

First, I looked at just the “Manual” cancellations, as the way this was entered in the dataframe (not in all caps like the other products) suggests to me that this is in reference to a manual override cancellation rather than a product named “Manual”. I made a dataframe of only the manual cancellations and plotted by date in a histogram. (I chose 52 bins for the 52 weeks of the year):

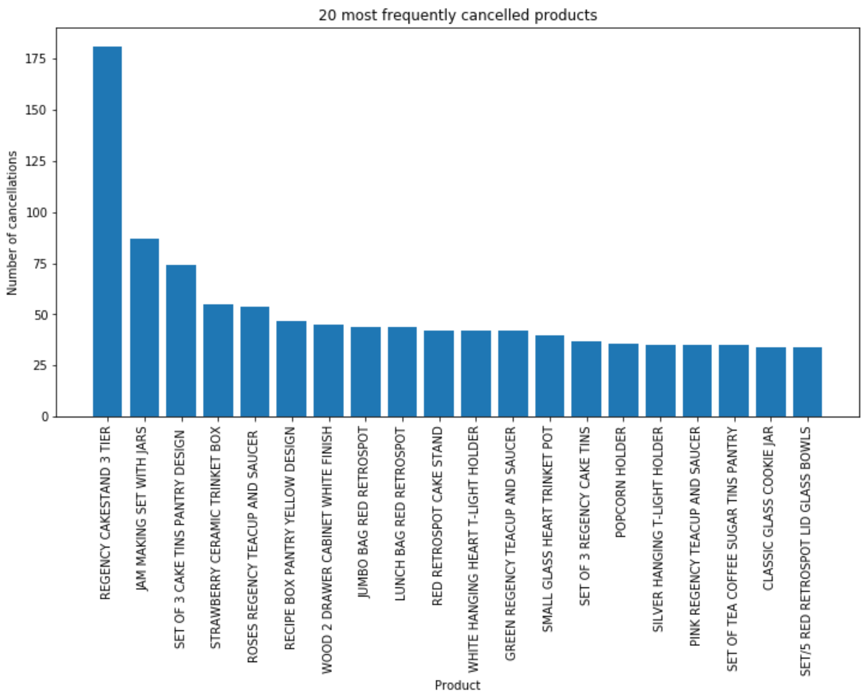


I noticed there are 3 peaks above the rest and definite spikes and lows. I plotted the manual cancellations along with holidays in the UK to see if there is any correlation between cancellations and holidays (when gifts would likely be purchased or returned in higher numbers):



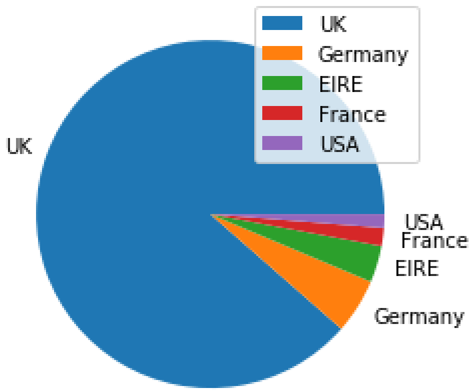
It does appear that two of the peaks with highest cancellation frequency also have holidays at the same time. The first doesn’t appear to coincide with any holidays and the largest peak is consistent in timing with the highest number of manual cancellations explored above.

Second, I looked at the top 20 most commonly cancelled products, excluding the manual cancellations, postage, discounts, and samples. Plotted, the top 20 products looks like this:



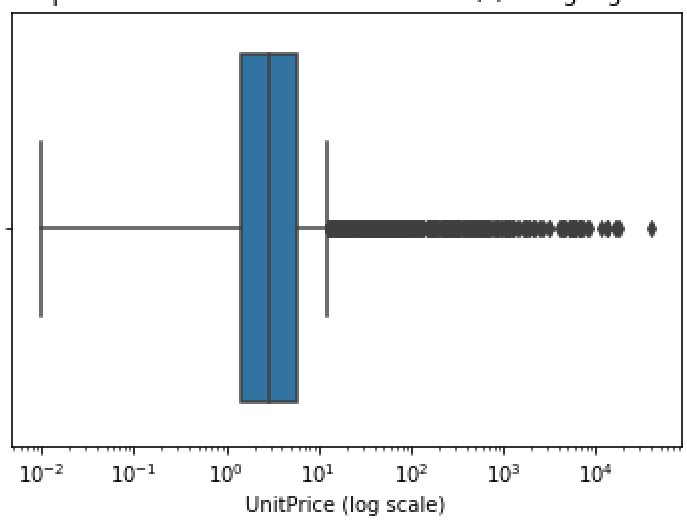
I did look to see if there were any other patterns for cancellations, such as by country. I found that most of the cancelled products (sorted by the top 5 most common) were from the UK, as shown in the pie chart. Given that this is a UK-based company; that makes sense to me.

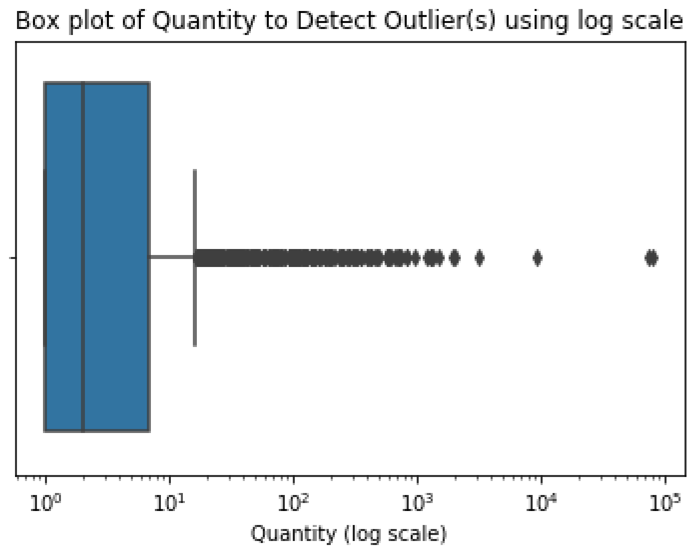
Top 5 Countries of Cancelled Products



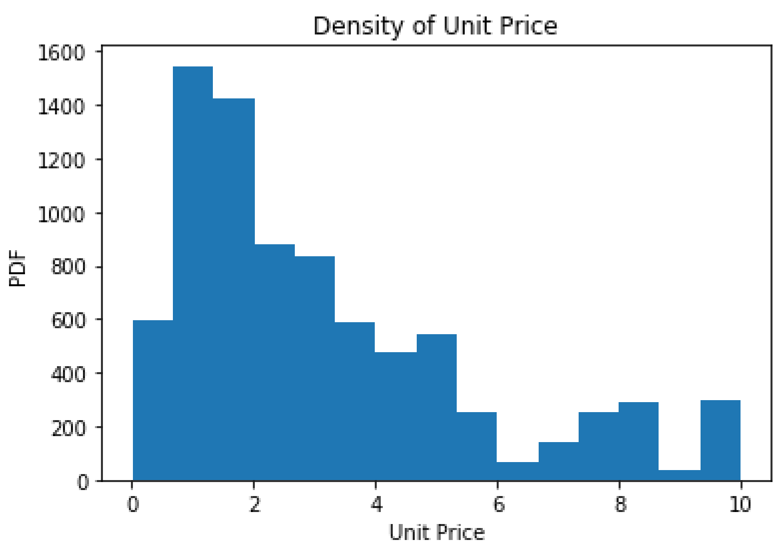
Next I turned to clustering the cancelled products. I checked to see if the data was normally distributed by price and quantity. I looked at a box plot to detect outliers.

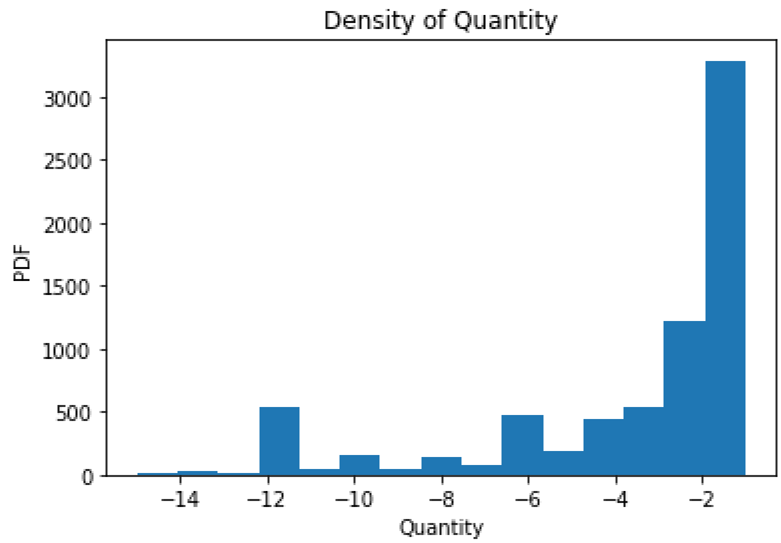
Box plot of Unit Prices to Detect Outlier(s) using log scale





Based on these box plots, I was able to eliminate around 1000 values as outliers for each; those above 10 for price and those above 15 for quantity. The histogram to look for normalcy for each of these looked as follows:

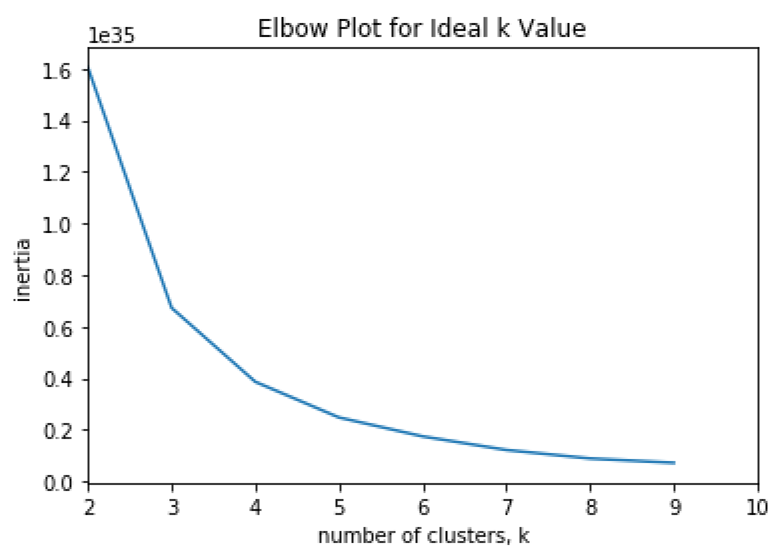




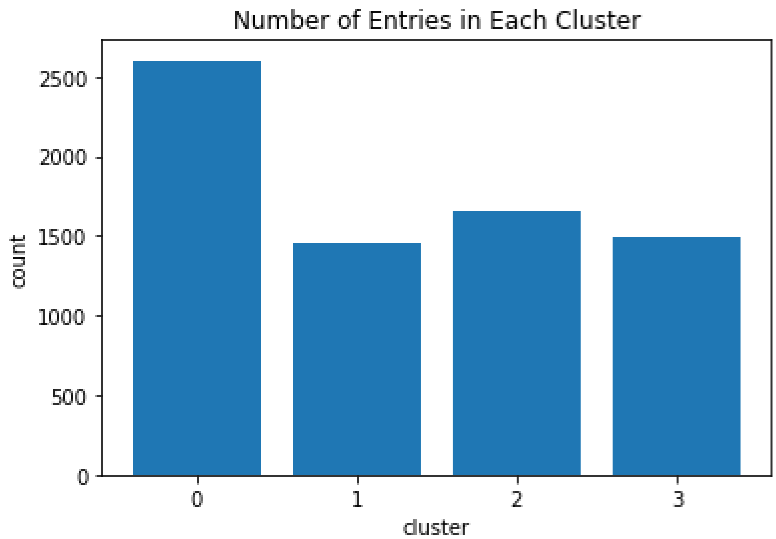
Certainly these are not normal curves, and this will need to be taken into account when reported any results based on normalcy curves.

I completed a KMeans clustering on the cancelled products, using categories of quantity, unit price, weekday, country, and invoice date.

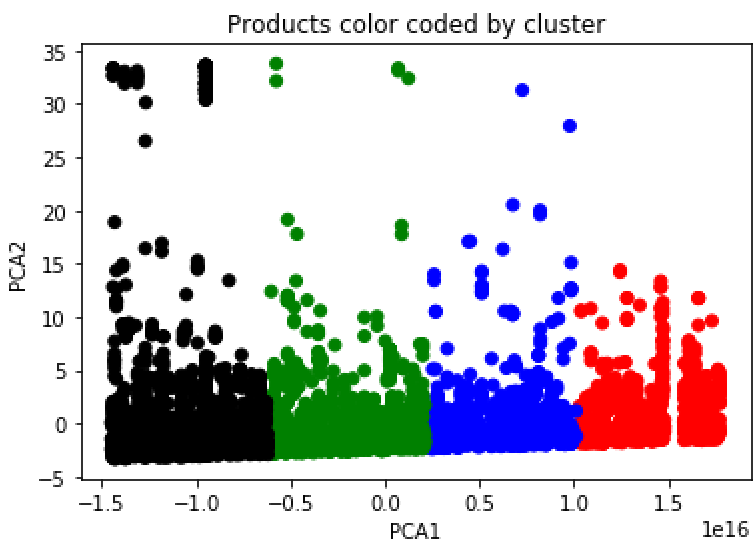
To determine my K value (number of clusters), I created an elbow plot to look for the best point. From the following plot, I selected 4 as my ideal number of clusters.



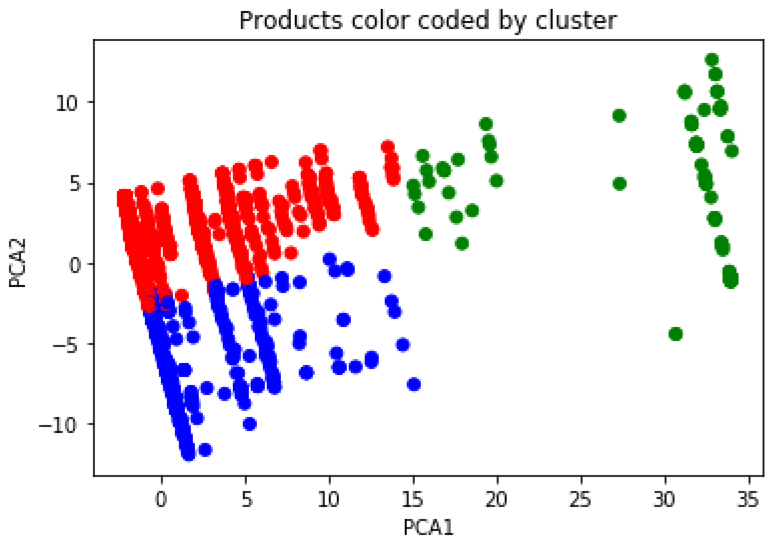
After clustering, I had a reasonably similar number in ¾ clusters and a large number in the first one:



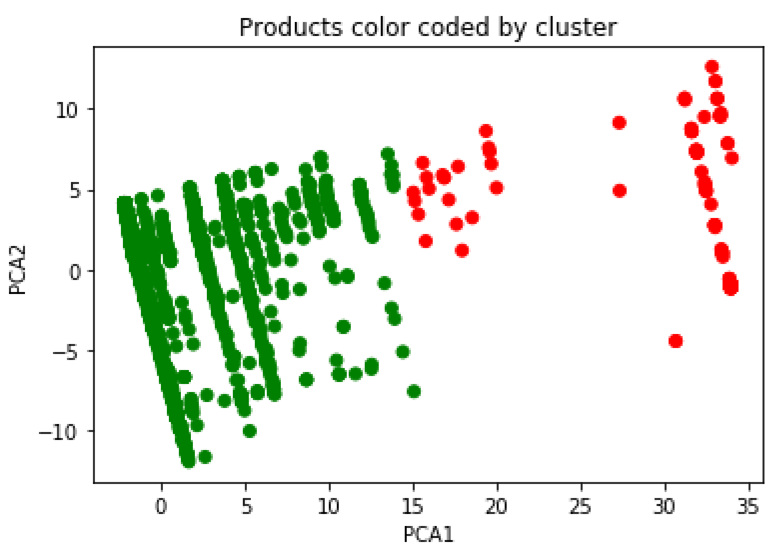
I transformed the clusters into a 2-D representation and color coded based on cluster:



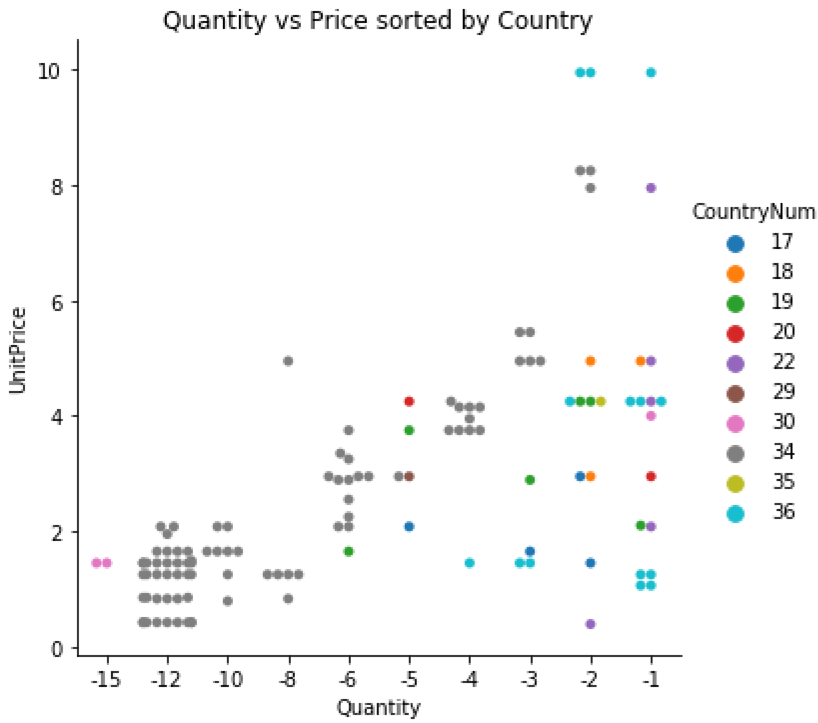
However, when I looked deeper, I realized that the clusters seemed to be solely based on date and didn't appear to have any other discernable pattern. So I removed the date to see if I could find a secondary clustering within that. Again, I created an elbow plot and selected 3 clusters this time. When color coded, they looked like this:



There does seem to be some blurring of the red and blue clusters. I completed a silhouette analysis to attempt to evaluate the effectiveness of the clustering. With that, I found that 2 clusters would be more indicative of a true pattern, so I re-clustered using only 2 with the following result:



These seemed to be divided by country, as the green cluster was countries 0-16 and the red was clusters 17-36. Still, I looked closer at the red cluster to see if there was any other pattern of interest. When I created a categorical plot by quantity and price, sorted by country, I got the following plot:



The same axes plotted by weekday yielded this:

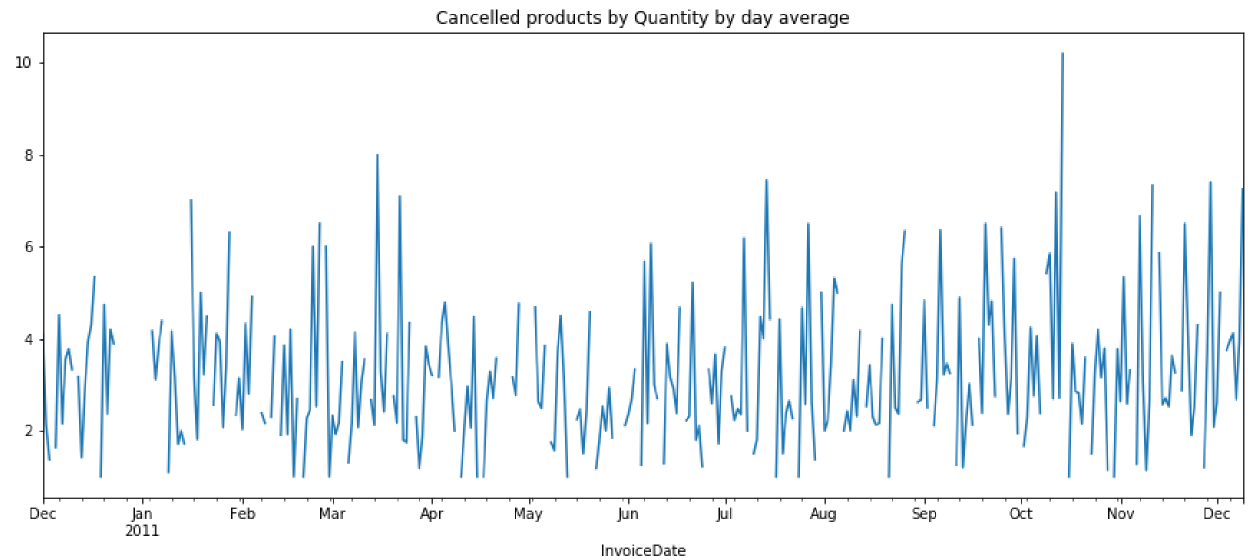
I noticed that the green weekday (2 = Wednesday) seemed to align well with the grey country (34=USA). So I plotted the USA data:

It would seem that customers from the USA tend to order on Wednesdays and as price increases, the quantity they return decreases.

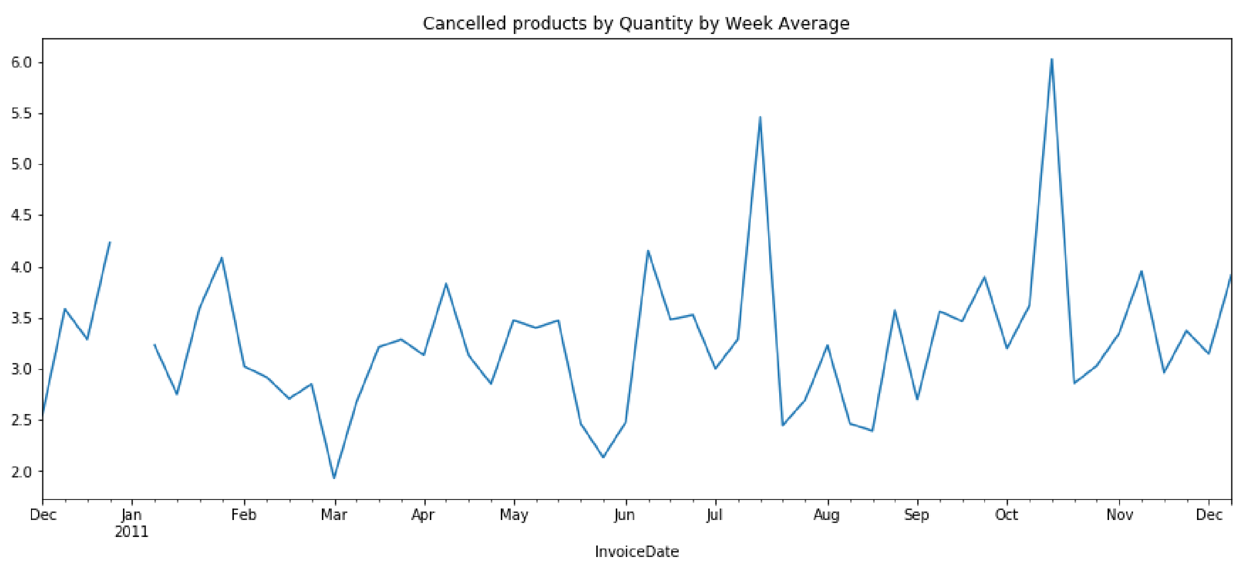
There was no discernable pattern from the other cluster.

Since clustering seemed to yield the best results with two clusters, and there was not a clear element of prediction, I decided to change course and look at a time series analysis.

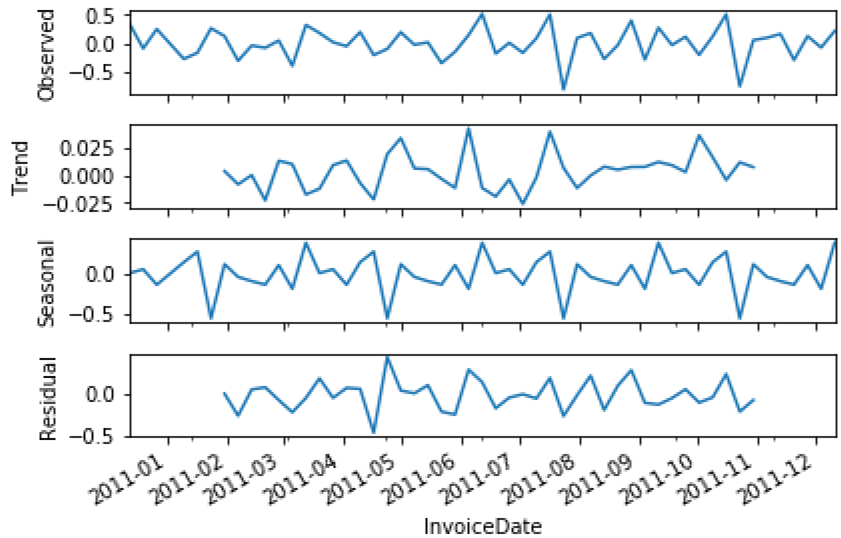
Again, I looked at both price and quantity. By taking an average of each day, I got a very busy plot for quantity:



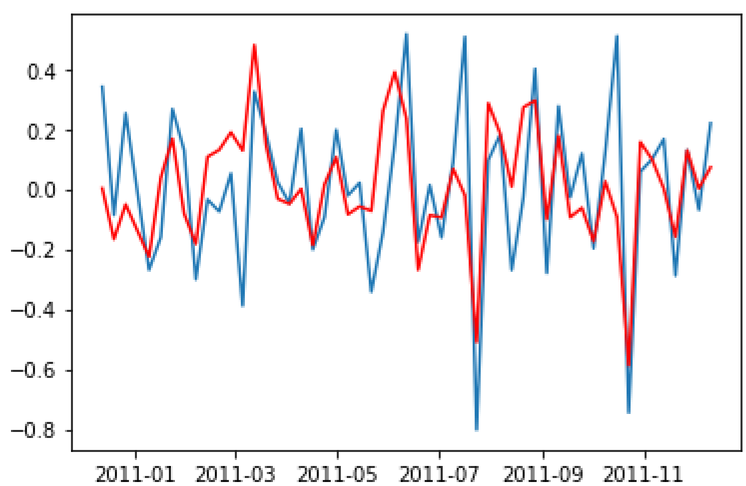
So, I decided to look at an average by week:



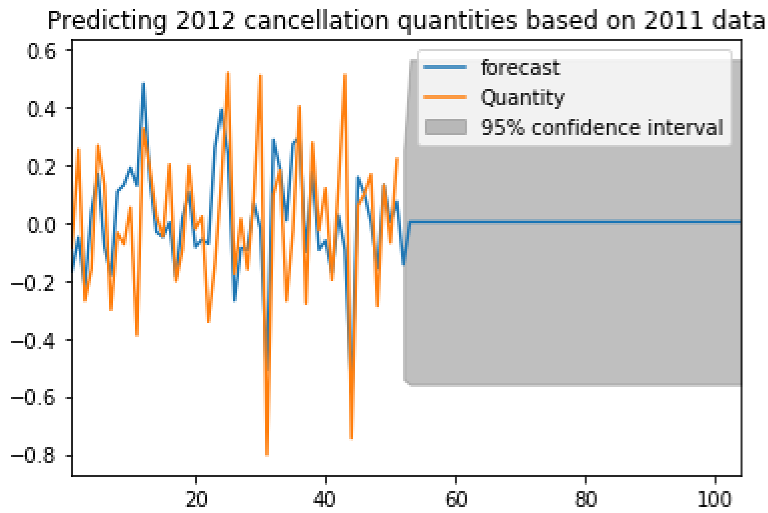
There appears to be a valley in this plot every 3 months or so. I utilized seasonal decomposition to break this plot down into what we observed, the overall trend, the seasonality, and residuals:



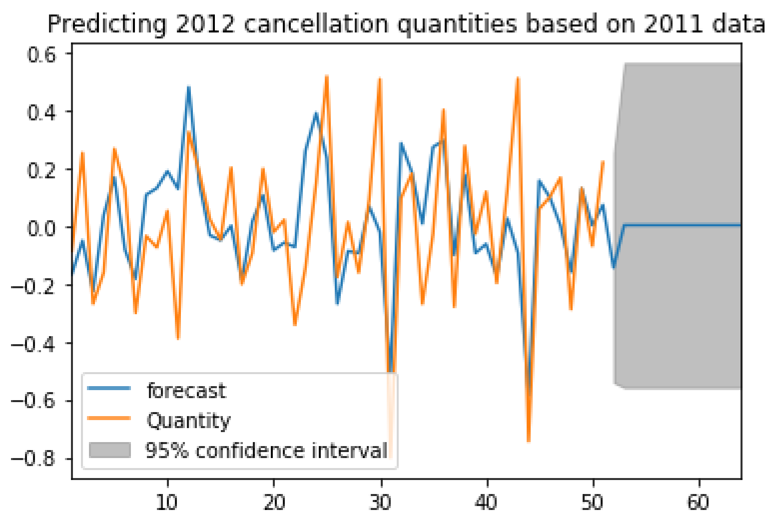
I utilized an ARIMA model (after optimizing parameters) to come up with this model (blue=data, red=model):



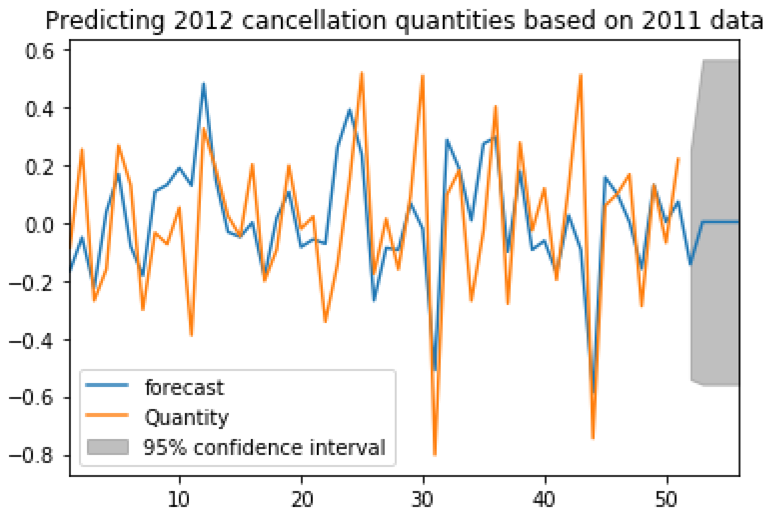
This appears to line up with the data well, however the highly variable nature of it, means that a prediction for 1 year into the future looks like this:



This isn’t terribly predictive. I considered that I was attempting to extrapolate too far into the future given the amount of data I have. I pared it back to a forecast 12 weeks (3 months) in the future and got this plot:

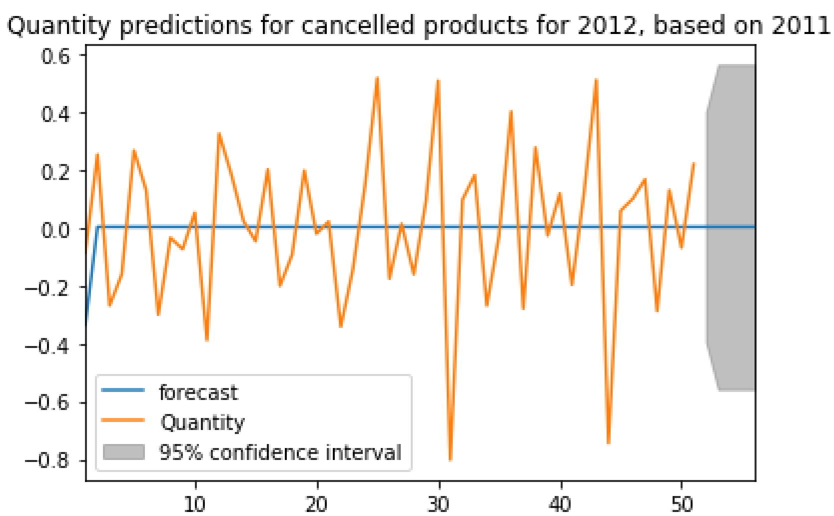


I pared it back further to only look 4 weeks (1 month) in the future:

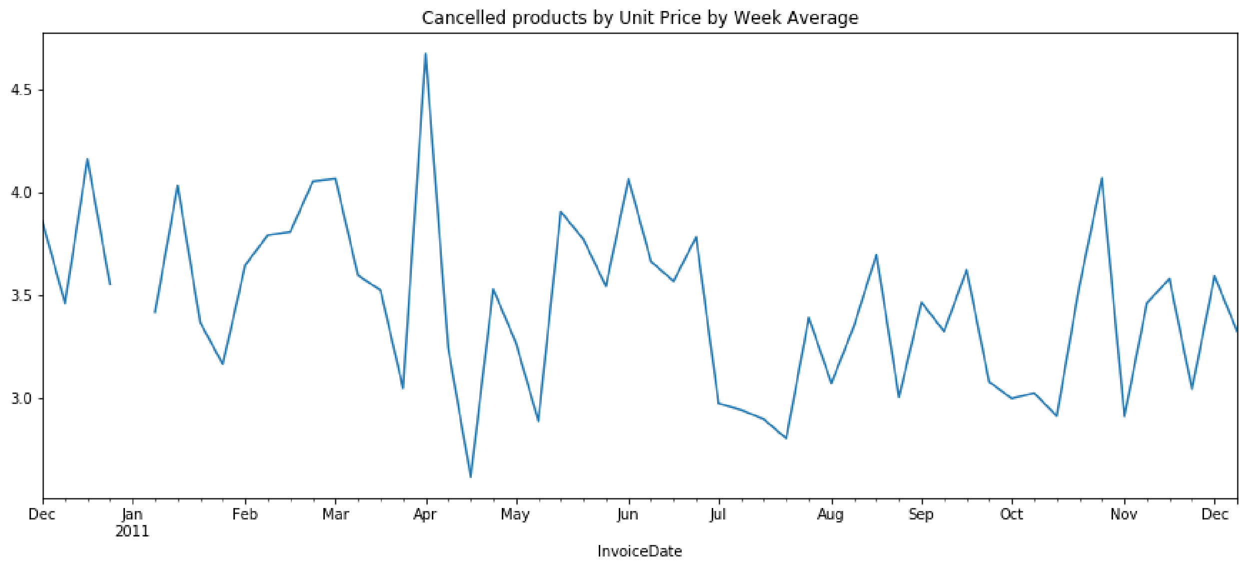


It looks like this feature is not a good one to make predictions on. My model follows the general trend of original data well but doesn't appear to have good predictive power.

Tweaking one of the parameters (dynamic=True) in an attempt to account for the dynamic nature of the plot yielded an even poorer forecast:

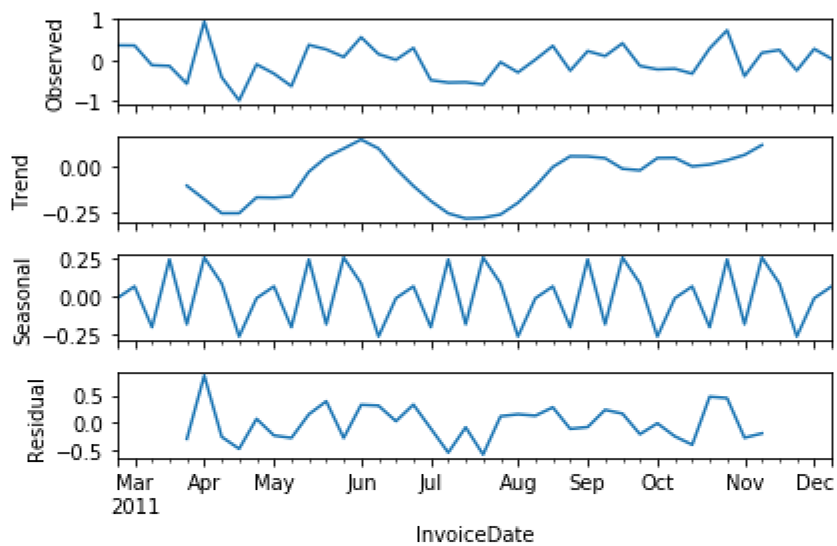


I wanted to look at price as well, to see if there was a clearer predictive model possible from that data. Averaging to weekly values produced the following plot:

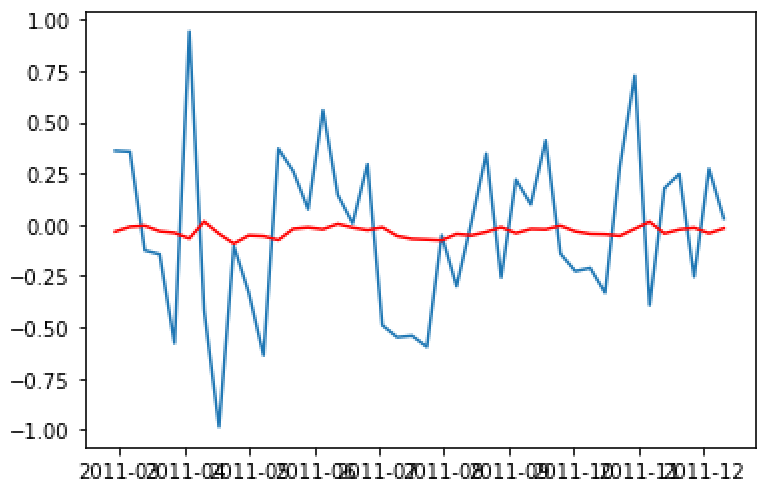


This seems to have a “valley” every 2 months or so.

Again, here is the data broken down into its seasonal components:

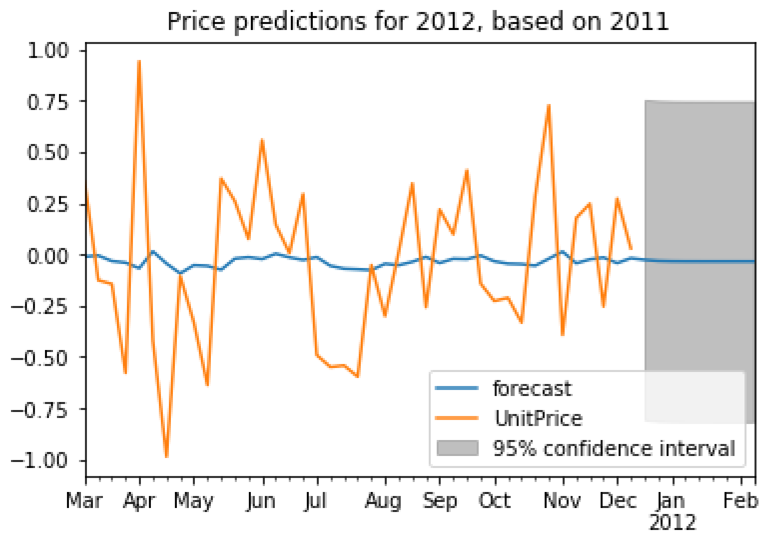


ARIMA (optimized for parameters):



I do note that this model appears to drastically under-fit the original data

Prediction 1 month out:



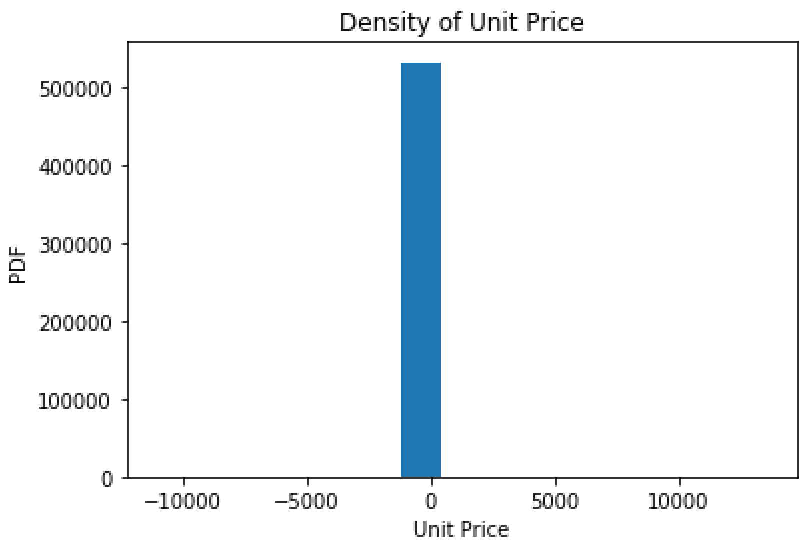
As you might expect from a model that underfit so poorly, the predictive power is not good.

Overall, neither quantity nor price for cancelled products seems like the best way to predict future cancellations, nor is there a clear pattern of when orders are cancelled.

Purchased products analysis

I turned my attention to the purchased products next. First, I removed all the cancelled products to make sure I wasn't counting them in error and did some quick math to make sure I had the number of entries that I expected.

I looked to see if the purchased products followed a normal distribution and found, to my surprise, that the unit prices appeared to skew negative:



This confused me, as I had removed the cancelled products which showed up as negative unit prices. When I looked for those values below 0, I found that there was a stock code "B" that had the description "Adjust bad debt". This corresponded to two entries, each with a unit price of -11062.06. If I were in contact with the company in reality, I would ask what that corresponds to so that I could better account for it in my analysis. As it were, I removed it from the dataframe as it did not appear to correspond to a product.

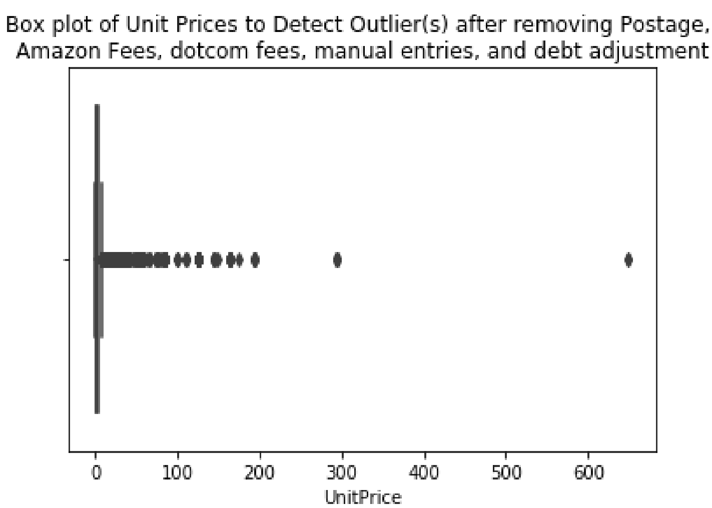
I did create a box plot to check for outliers, suspecting that those values found above would qualify but wanting to be thorough. My box plot looked like this:

The (at least) three points on the high end were found to correspond to Amazon fees, postage, and another, positive, “Adjust bad debt” item.

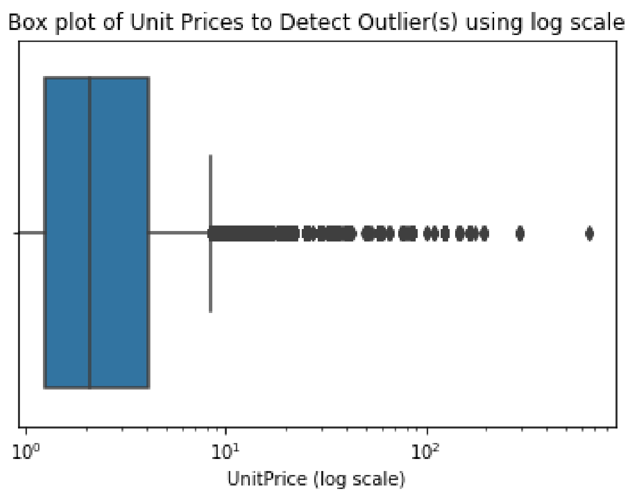
When I completed a statistical analysis to find outliers (using a z-score), I found that every data point was considered an outlier. This seemed suspect to me.

I removed the entries that did not correspond to items for sale: Amazon fees, postage, and bad debt adjustment. Through my exploration, I also found that dot com fees and manual entries were present as well and removed those from the data set too.

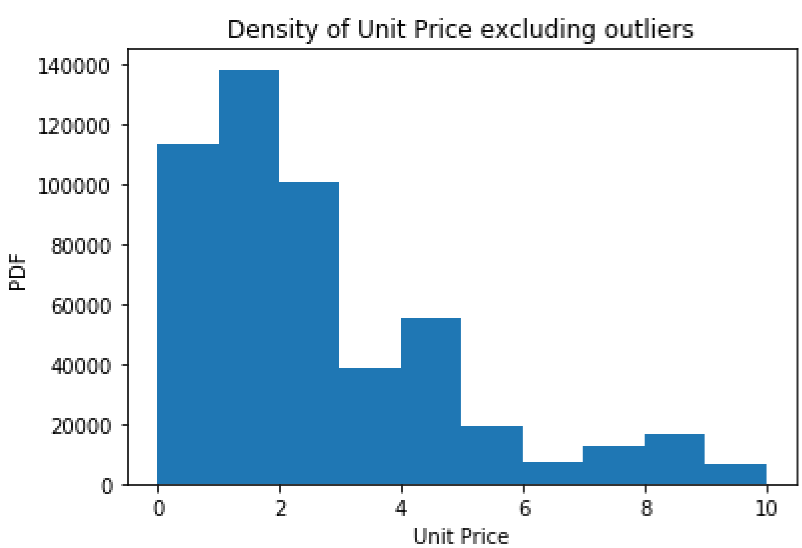
The new box plot still showed a number of outliers, but was at least all positive numbers.



I then created a box plot using a logarithmic scale to better visualize the data and came up with the following, which shows that the majority of products lie within the 0-10 Pound range and also solved my problem from before where all products were outliers:

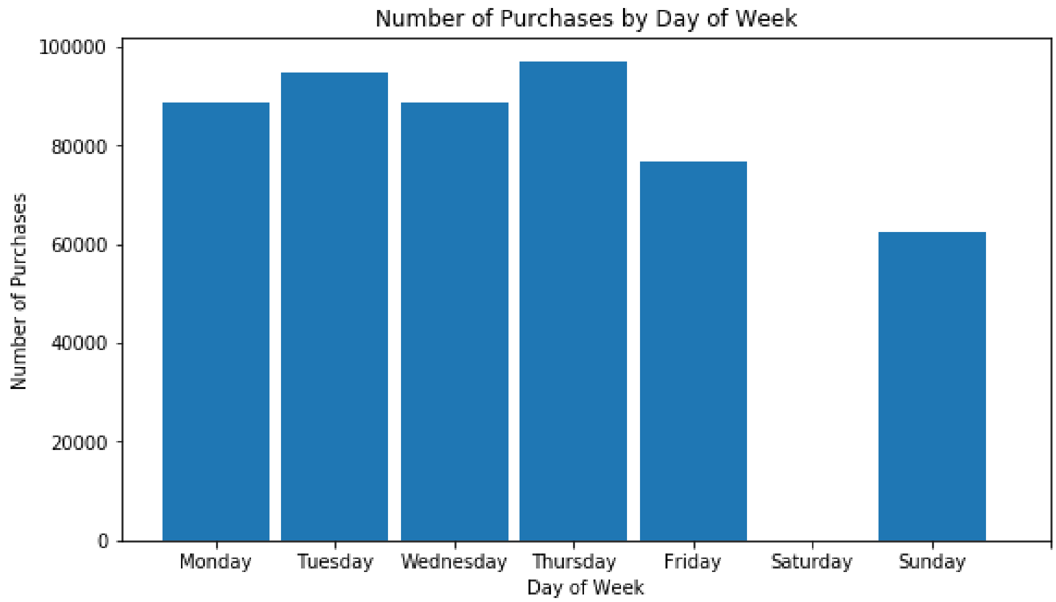


I again looked at the distribution of items for sale by price, excluding the outliers above the 10 Pound price and generated this plot:



This plot suggests to me that any analysis I do that relies on a normal distribution will be skewed high.

I looked at purchased products by day of week, much as I did for cancelled products:

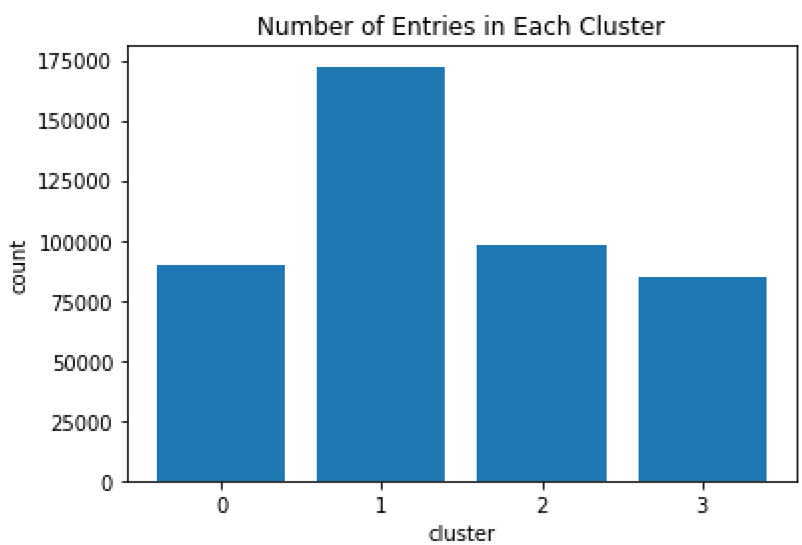


Again, no products were purchased on Saturdays.

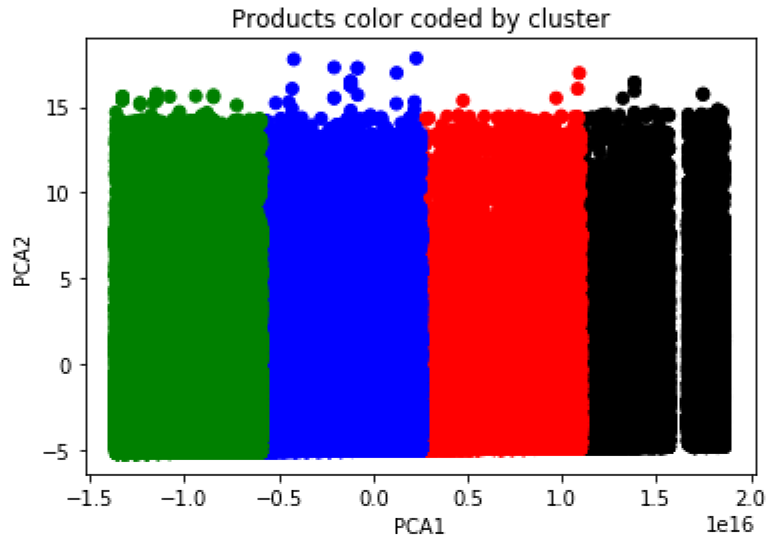
A second ANOVA showed that between all days there is a statistically significant difference and that it is not due to chance.

I completed a kMeans clustering much as I did for the cancelled products. I created an elbow plot to look for the optimal number of clusters and determined that 4 would be most appropriate.

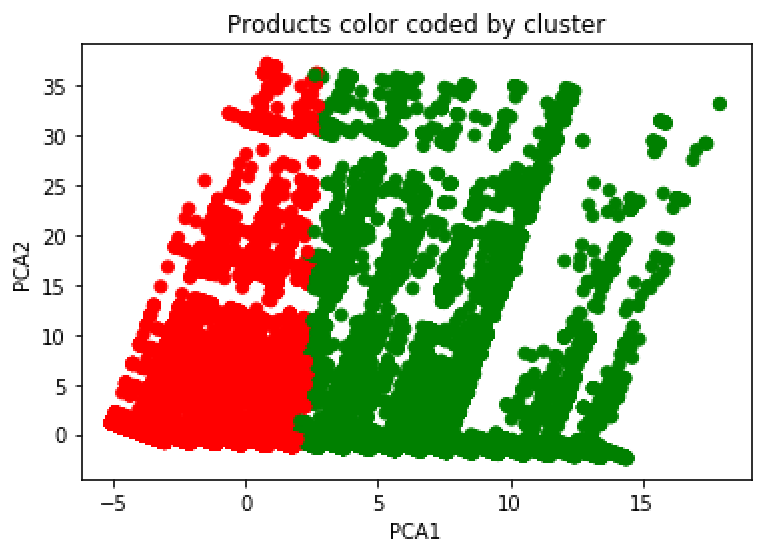
Much as before, there were 3 clusters with similar numbers and one with notably more entries:



And again, when clustered, color-coded, and analyzed more closely, they appear to be separated based solely on date.

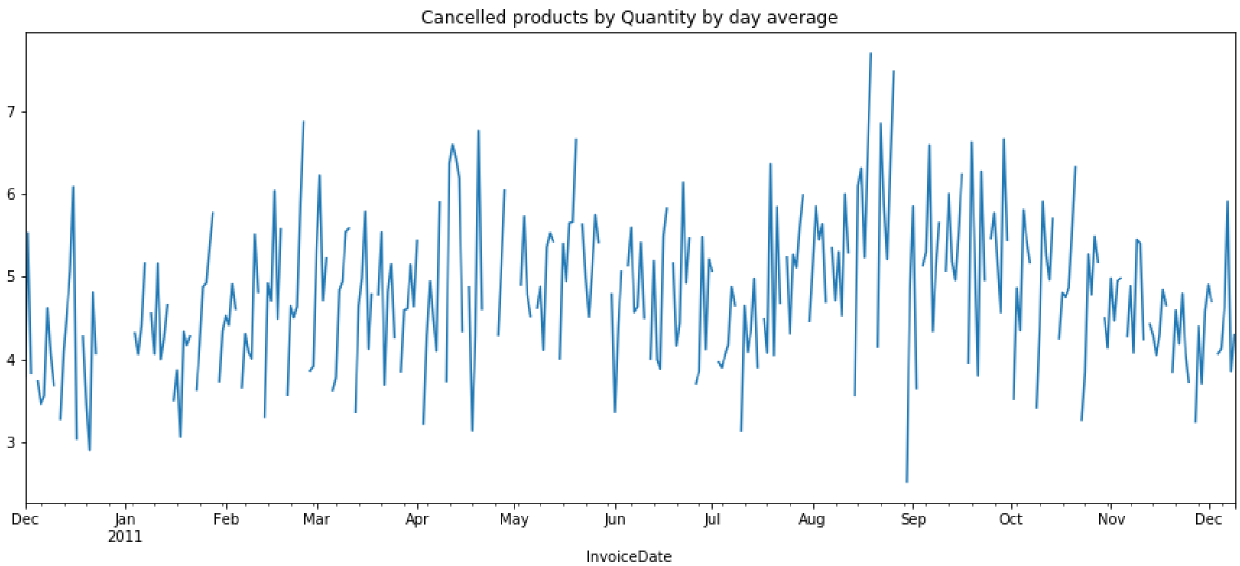


As with the cancelled products, I removed the data and clustered again. This resulted in an unconvincing clustering that did not appear to have predictive value:

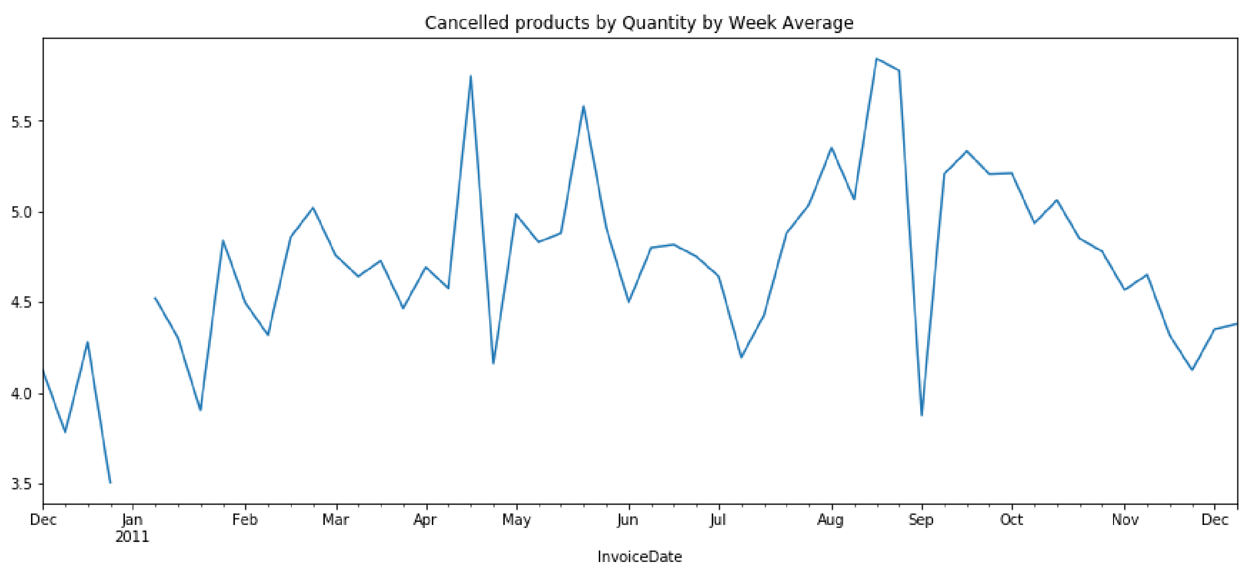


I created a time series plot in a similar fashion to the cancelled products analysis.

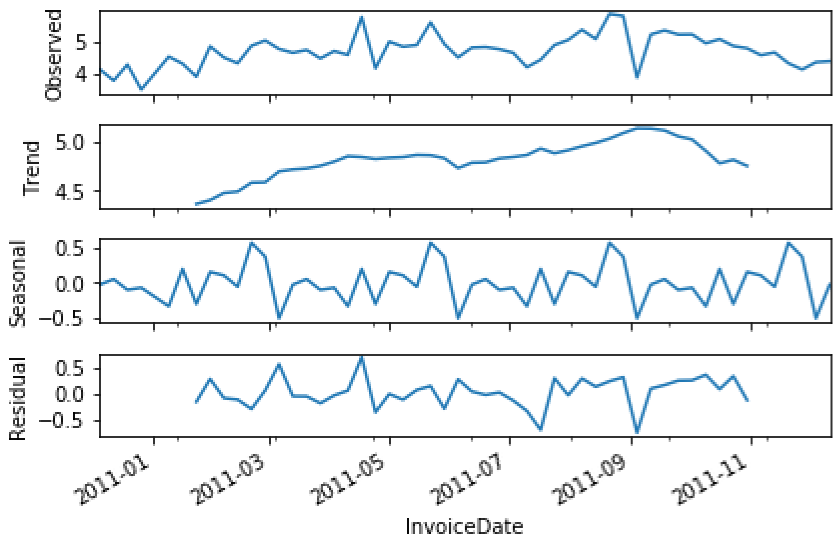
Unsurprisingly, the quantities averaged out by day was difficult to see any trend:



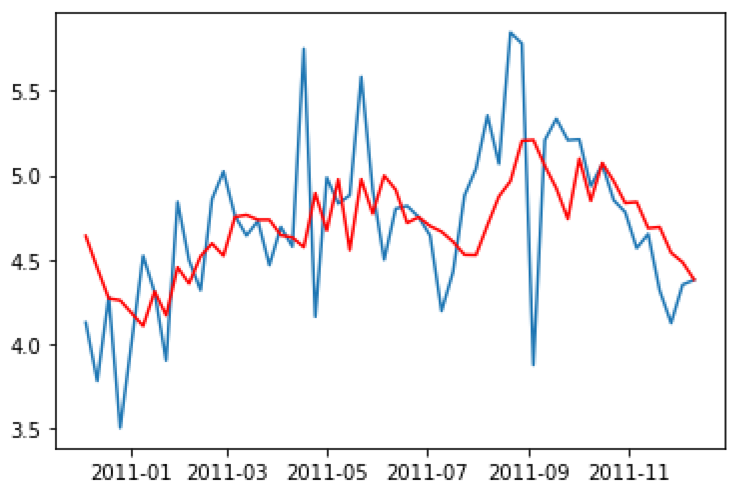
Weekly averages showed more promise:



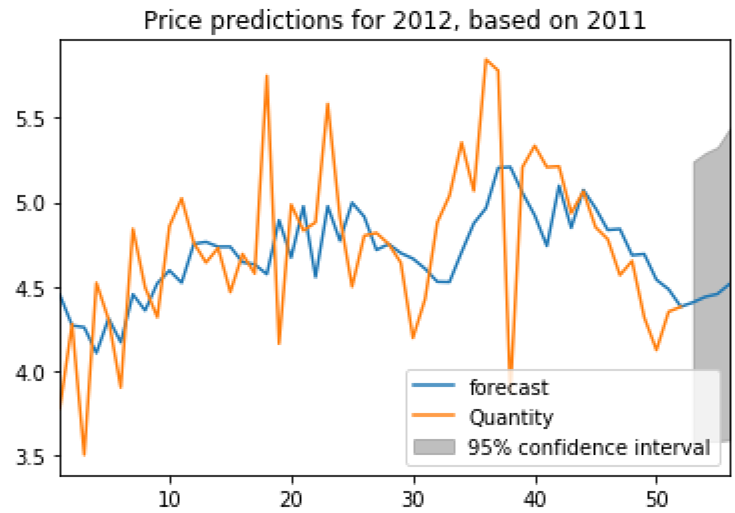
Broken down into its observed, trend, seasonal, and residual components:



ARIMA model:

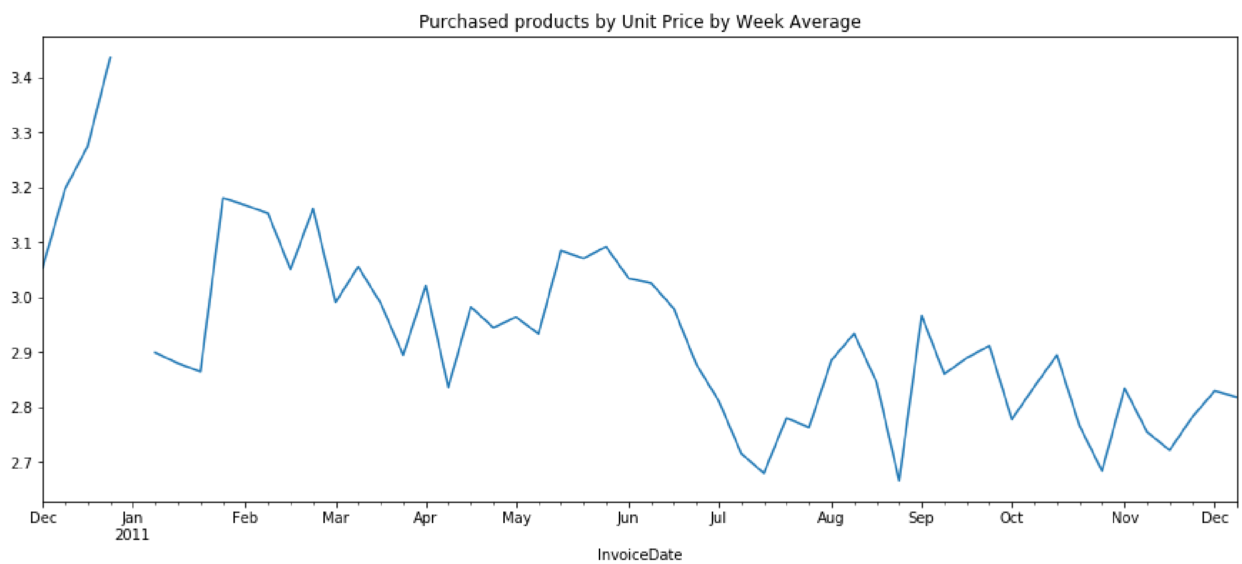


1 month prediction:

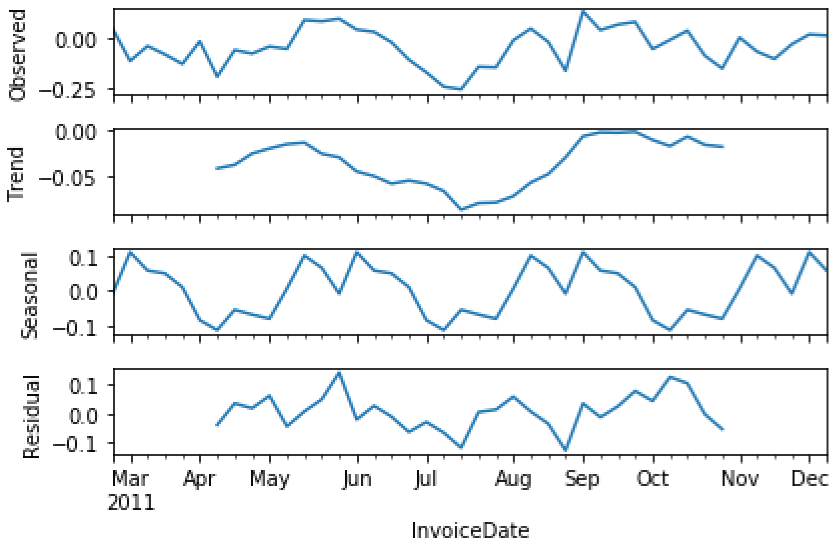


Again, the 95% confidence interval makes this prediction difficult to put much stock in.

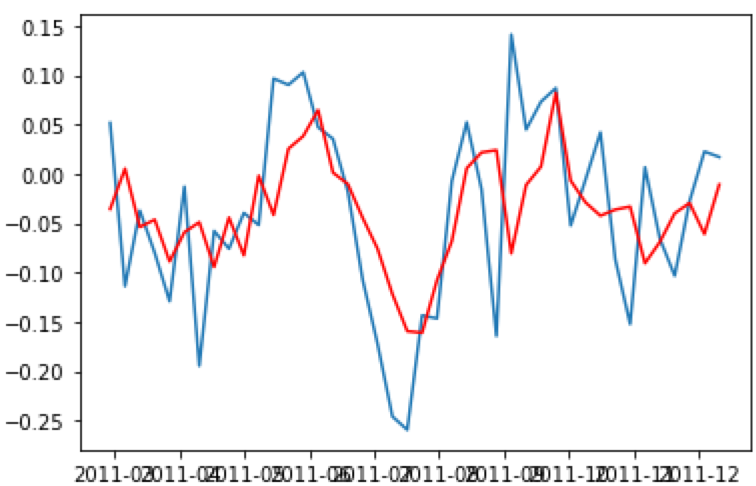
I looked at price of purchased products as well:



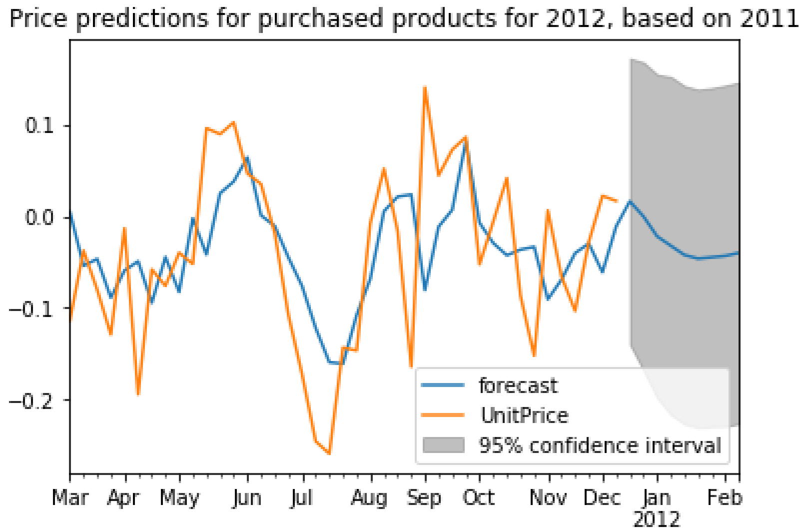
Broken down into components as before:



ARIMA model with optimized parameters:

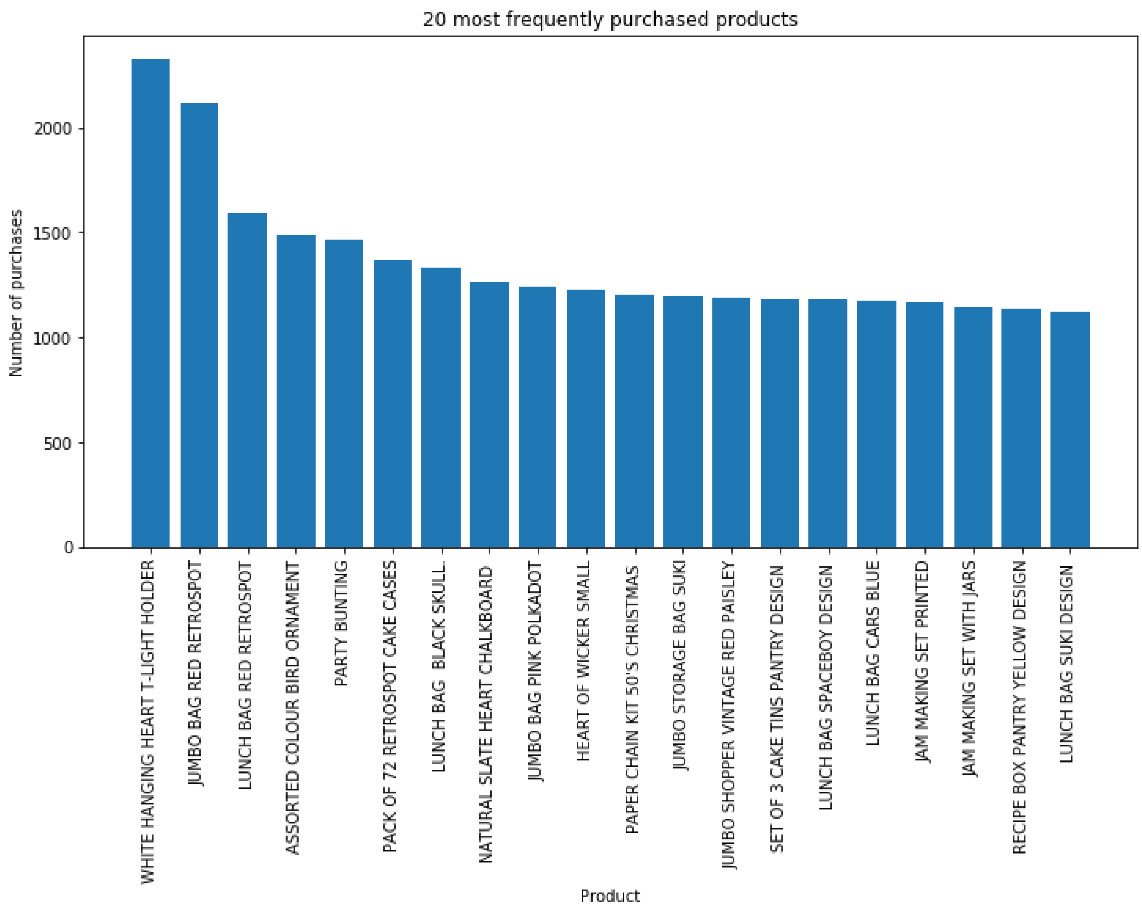


Finally, a prediction 1 month out:



The 95% confidence interval remains quite sizeable.

For an informative look, I plotted the top 20 purchased products:



I also looked to see if there were any common products between the top 20 cancelled products and the top 20 purchased products. There were 6:

Description	**Number cancelled**	**Number purchased**
JAM MAKING SET WITH JARS	87	1142
SET OF 3 CAKE TINS PANTRY DESIGN	74	1182
RECIPE BOX PANTRY YELLOW DESIGN	47	1133
LUNCH BAG RED RETROSPOT	44	1594
JUMBO BAG RED RETROSPOT	44	2115
WHITE HANGING HEART T-LIGHT HOLDER	42	2327

Storytelling:

It is helpful for businesses to know what’s working and what isn’t. One way to look at this is through products sold and returned. Having an understanding of what products are most frequently being sold and when, as well as which products aren’t selling as well can assist you in ordering and providing you the insight to adjust your business as needed. This could look like adjusting prices, creating a targeted ad campaign, or considering no longer stocking a certain product.

I looked at the cancelled products and purchased products separately.

<u>For the cancelled products:</u>

For this business, by far a majority of their cancellations are manual. Perhaps this warrants a deeper look. Is their system not keeping up with requests for cancellations that then need to be manually completed? Are products not being entered into the system?

Excluding the manual cancellations, postage, discounts, and samples, the most commonly cancelled product is the 3 tier Regency Cakestand. Perhaps this would be a product to consider removing from their inventory? Or investigating if there was a common reason (defective? Not as advertised?)

Additionally, more returns happen on Thursdays than any other day. This could direct the company’s energy on that day (maybe allocating more staff to processing returns on Thursdays). Or this could be an opportunity to target increasing sales on Thursdays or surrounding days to offset returns.

When looking to make predictions based on past cancellations, the products (and the habits of people returning them) don’t seem to fit neatly into categories. When attempting to cluster the products into like categories, the results do not lend themselves to prediction.

Looking at the cancelled products over time, in a time series analysis, we can see that the cancellation trends are all over the place and don’t have an obvious frequency or pattern. We can fit a model to the data well enough, but the predictive quality is not high. At best, we can make some pretty broad generalizations about where the trend might head, but the confidence interval is wide.

<u>For the purchased products:</u>

On the purchasing side, almost half of the top 20 products were bags. Perhaps this could be a direction to investigate further for the company to expand sales. Fewer purchases are made on Fridays and Sundays as compared to other days. Perhaps this could be a targeted sale on those two days only to increase sales on those dates.

Again, when attempting to categorize and cluster the products sold, the products did not lend themselves well to this technique. A time series analysis of purchased products by quality and price yields similar results to the cancelled products: there is great variability in the habits of consumers at this particular store. It would be interesting to see if more data (perhaps 5 years instead of 1) would give better results and allow for more accurate and useful predictions.

As an aside, we can also see that there are 6 products that made both the top 20 purchased list and the top 20 cancelled products list. On the highest side of this, 7% of the sales of the “Jam making set with jars” were cancelled. (cancelled/(cancelled + purchased)) = (87/1229). This may not warrant action from the company, but they may be able to utilize this information to adjust return policies, adjust what products they sell, or how they market it.