Crystal Mosley

Predict 413 – Time Series & Forecasting

Midterm – Summer 2018

**Forecasting 1C Company Total Sales for Every Product and Store in Next Month**

**Introduction**

All companies with stores and products to sell need to know if their stores are profitable, or if they’re lagging in certain areas either due to product loss, product popularity, seasonal changes in the needs and wants of their customers, lack of products, etc. Maybe some stores do better in certain geographies versus other geographical areas. Maybe the customer base is significantly larger or smaller in some areas versus others.

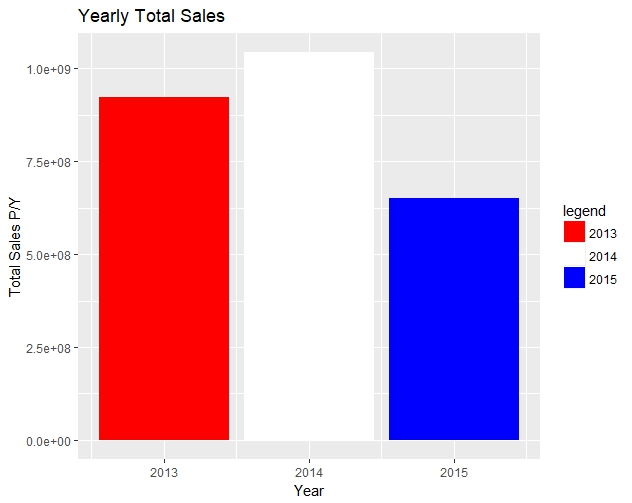
This competition is to predict the total sales for every product and store for the next month in 2015, which was November. Since November is practically the start of the holiday season, depending on where the store is located, it’s important to know whether or not you’ll need more of a certain product or products, and whether or not the stores can improve on sales during the holidays, or if they can find ways to cut back and not take a profit loss – i.e.: shutting down the stores earlier instead of later to save on money; not keeping products that aren’t selling on the shelves.

I didn’t think about getting ahead of this midterm and starting weeks earlier, which is showing to be detrimental because the difficult part of this problem set is the actual data analysis and cleansing. I could not open the csv.gz files, even with a winZip converter because it either came out as nothing but odd characters, or corrupt files. I needed to be able to pull the “item\_cnt\_month” data from the *sample submission* but it too was a csv.gz file that I had issues with. The dataset was the biggest set I’ve ever worked with, and my computer couldn’t even handle such a large set of data when I tried to merge the *sample submission* data into the *sales training* data.

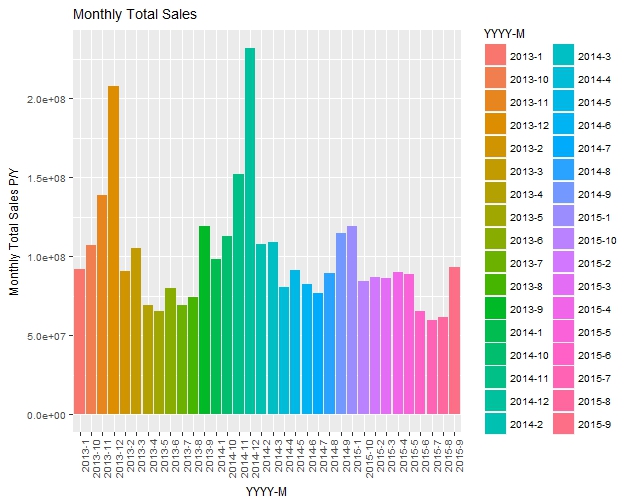
**Data**

To start this analysis of the data, I read in each of the provided datasets: *sales\_train.csv.gz, test.csv.gz, shops.csv, items.csv, item\_categories.csv,* and *sample\_submission.csv.gz*. I reviewed the top 20 rows of each set, as well as looked at descriptions about each set to see the number of observations and variable counts. I decided to join a few columns from the items, shops, and item\_category files into the sales\_train data set, and then confirmed they were merged. I then added “year”, “day”, “weekday”, and “month” columns to the sales\_train dataset.

In reviewing the dataset, I started with the yearly sales from 2013-2015.

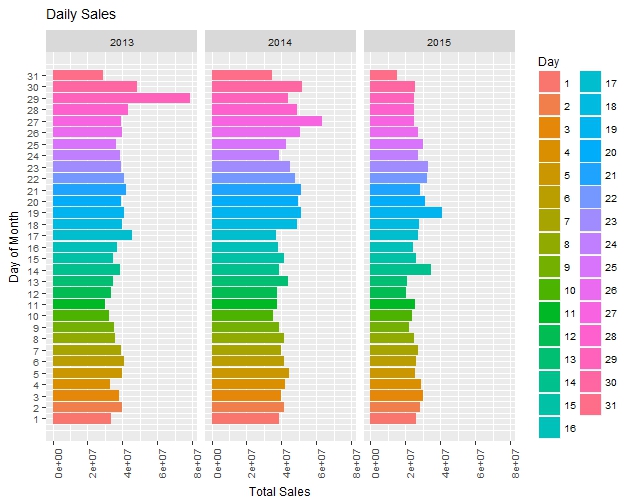


We can see that sales went up by $122 million in 2014 in comparison to 2013; however, in 2015, sales fell by about $393 million.



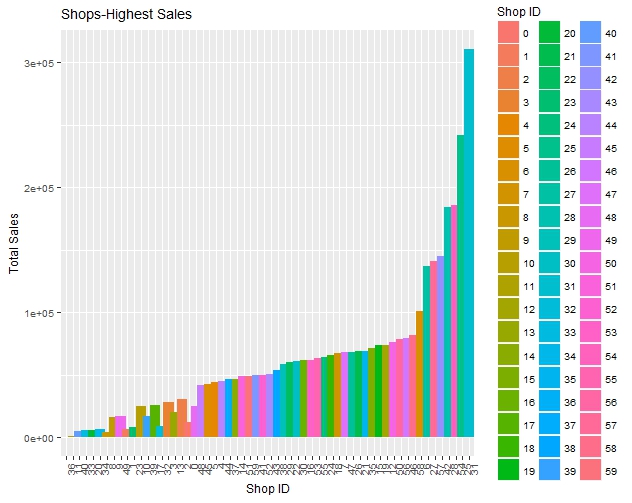
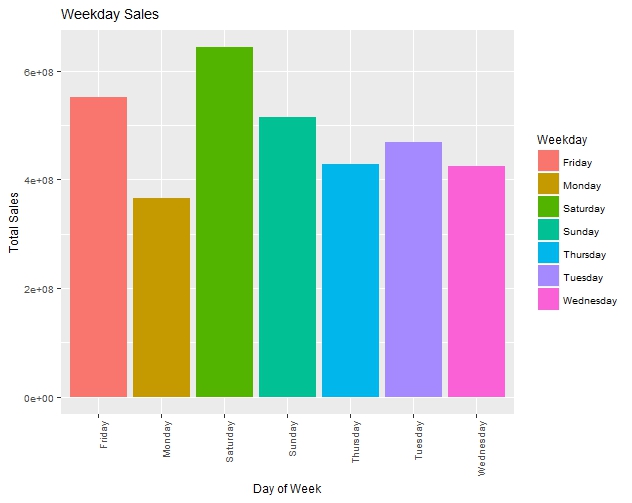
Next, we look at monthly sales. There’s a significant spike in November for both 2013 and 2014, probably due to the holidays. It is likely that November in 2015 will show a spike as well, but not has much as 2014 due to the lack of sales occurring in 2015 prior to November.

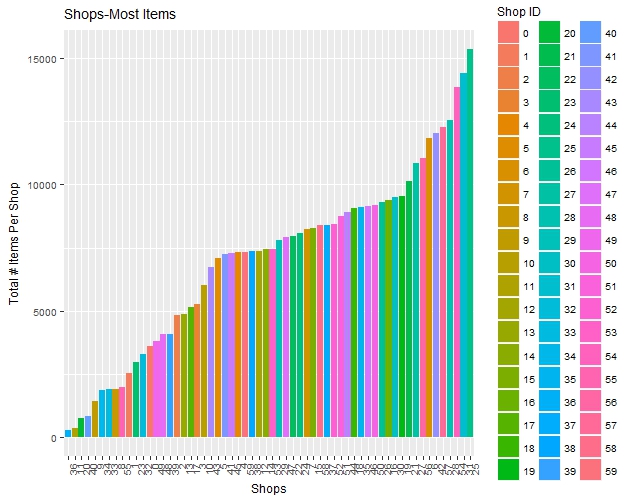
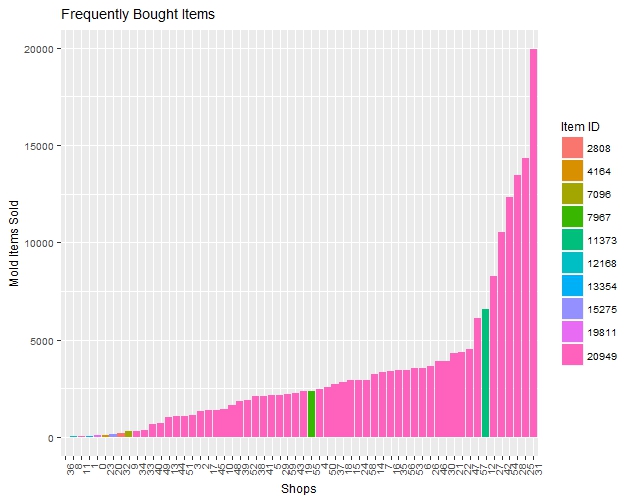
Next, we see which day of the month had the highest sales.

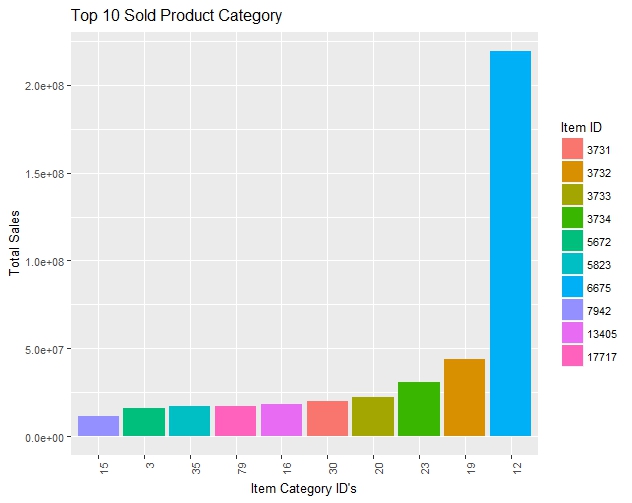
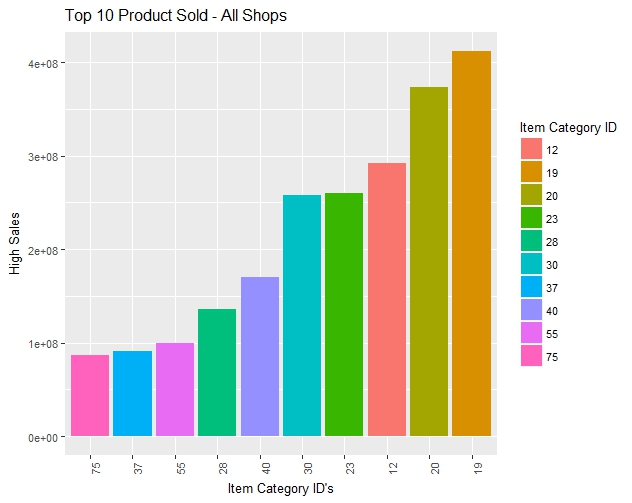
Next we look at more descriptive graphs for:

2013 shows a spike near the end of the month, 2014 shows as nearly equal from the middle to the end of the month, and 2015 shows a decrease as the end of the month with a small increase in the middle.

|  |  |
| --- | --- |
| * Weekday with highest sales | * Frequently bought items per store |
| * Shops with highest sales | * Top 10 sold products – all stores |
| * Shops with most items | * Top 10 sold products – within each category |

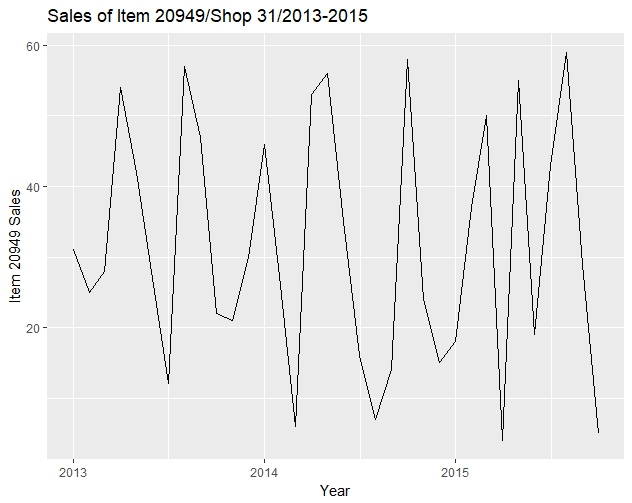
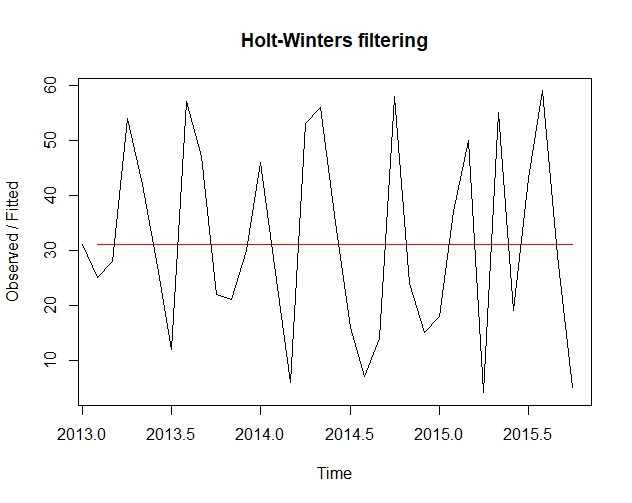




**Models / Formulation / Performance / Limitations**

I created a plot and model for frequently bought items because I felt as though this would have significant impact on sales for each store. In plotting this time series, it looks like nothing but white noise. I then ran a Holt-Winters SES on this data and it shows a straight, smooth line for prediction.

I ran three different methods on this time series: ETS, ETS with “ZZZ”, and an Auto.Arima. Out of the three, ETS and ETS with “ZZZ” were better (and were the same RMSE values. So, I chose to fit and forecast with the above model. The forecast doesn’t seem that great:

Point Forecast Lo 80 Hi 80 Lo 95 Hi 95

Nov 2015 31.45705 8.886689 54.02741 -3.061339 65.97544

Dec 2015 31.45705 8.886689 54.02741 -3.061339 65.97544

Jan 2016 31.45705 8.886689 54.02741 -3.061339 65.97544

Feb 2016 31.45705 8.886689 54.02741 -3.061339 65.97544

Mar 2016 31.45705 8.886689 54.02741 -3.061339 65.97544

Apr 2016 31.45705 8.886689 54.02741 -3.061340 65.97544

May 2016 31.45705 8.886689 54.02741 -3.061340 65.97544

Jun 2016 31.45705 8.886688 54.02741 -3.061340 65.97544

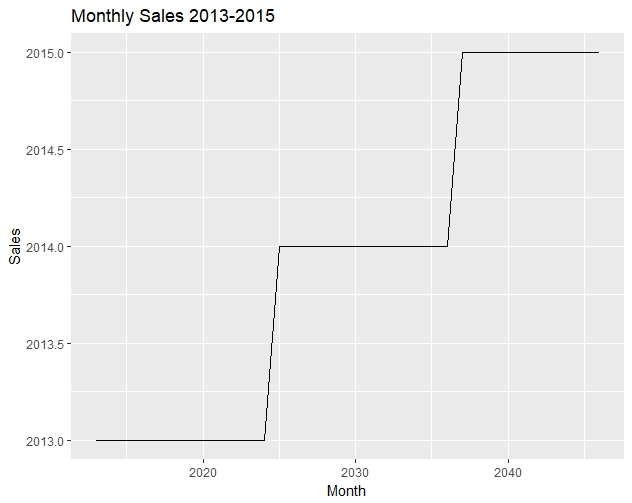
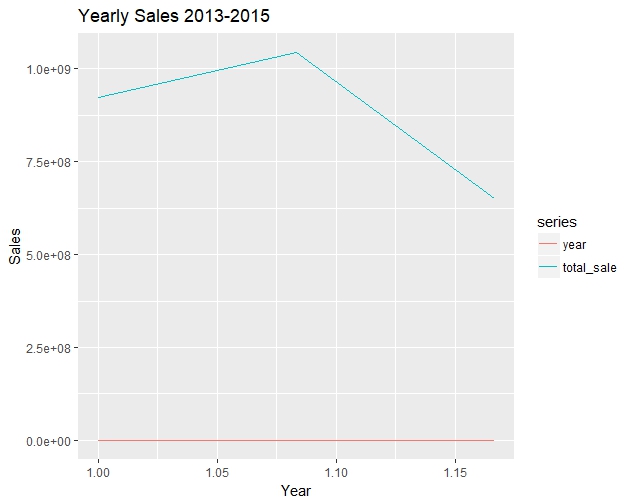
Jul 2016 31.45705 8.886688 54.02741 -3.061340 65.97544

Aug 2016 31.45705 8.886688 54.02741 -3.061340 65.97544

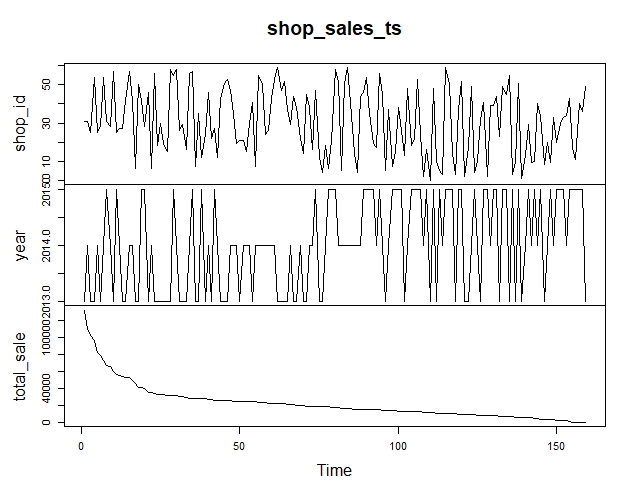
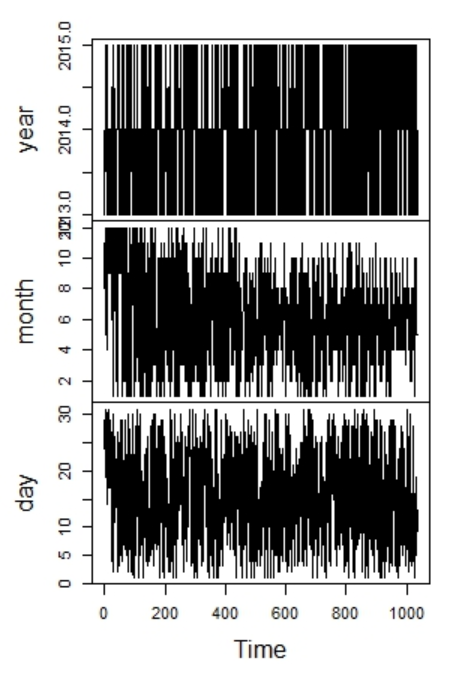
Sep 2016 31.45705 8.886688 54.02741 -3.061340 65.97544

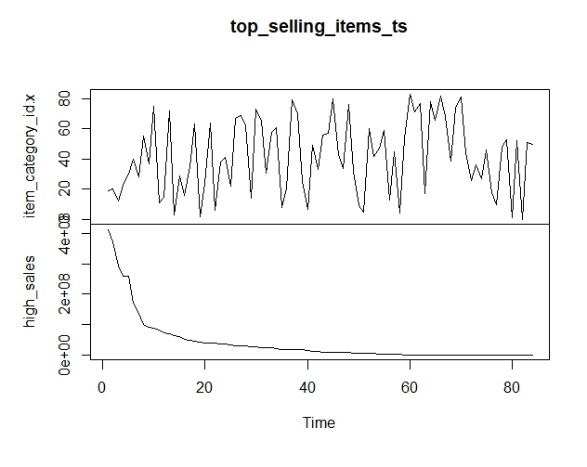
Oct 2016 31.45705 8.886688 54.02741 -3.061341 65.97544

I tried plotting the yearly and monthly sales, but the plots look odd:



I tried to plot and forecast a couple other variables with SES, but they looked odd as well, or didn’t work. It’s probably because some things weren’t seasonal data:





I believe had I given more time to the data cleansing portion, I would’ve been able to utilize what I initially planned on trying out from what we’ve learned so far in class: *meanf*, *naïve*, *snaive*, and *drift* methods to see which would work best. I would then move on to transformation/adjustments with residual diagnostics before moving on to accuracy. Time series decompositions would’ve also been helpful, especially STL decomp because of the below:

* Unlike SEATS and X11, STL will handle any type of seasonality, not only monthly and quarterly data. 1
* The seasonal component is allowed to change over time, and the rate of change can be controlled by the user. 1
* The smoothness of the trend-cycle can also be controlled by the user. 1
* It can be robust to outliers (i.e., the user can specify a robust decomposition), so that occasional unusual observations will not affect the estimates of the trend-cycle and seasonal components. They will, however, affect the remainder component. 1

**Future Work**

For future work (and the final) I plan on starting the data cleansing process immediately (weeks ahead of time) since I struggled with this piece a lot. At this point, anything I do in the future should improve this competition if I were to come back to it since this didn’t turn out so well.

**Learning**

For me, the biggest issue was data cleansing. I can follow the book’s models and other resource methods for forecasting my data, but I must get past the initial cleansing first. I would like to incorporate xboosting and poisson methods as some of the other students did to see how those options would work against the data.

**Literature**

Kapoor, S.G., Madhok, P., Wu, S.M. *Journal of Marketing Research*. Modeling and Forecasting

Sales Data by Time Series Analysis, 1981.

<https://www.jstor.org/stable/3151318?seq=1#page_scan_tab_contents>

Accessed 27 July 2018.

Nunnari, Giuseppe & Nunnari, Valeria. (2017). Forecasting Monthly Sales Retail Time Series: A

Case Study. 1-6. 10.1109/CBI.2017.57.

<https://www.researchgate.net/publication/319224544_Forecasting_Monthly_Sales_Retail_Time_Series_A_Case_Study>

Accessed 28 July 2018.

Sanwlani, Monica & Murlidhar, Vijayalakshmi. (2013). Forecasting Sales Through Time Series

Clustering. International Journal of Data Mining & Knowledge Management Process. 3. 39-56. 10.5121/ijdkp.2013.3104.

<https://www.researchgate.net/publication/269803515_Forecasting_Sales_Through_Time_Series_Clustering>

Accessed 28 July 2018.

Shafik, Engy. *JTATM*. Predictive Modeling of US Winter Apparel Sales Using Time Series

Forecasting Method, 2016.

ojs.cnr.ncsu.edu/index.php/JTATM/article/download/9858/4923

Accessed 27 July 2018.

Ugiliweneza, Beatrice. Use of ARIMA Time Series and Regressors to Forecast the Sale of

Electricity.

<https://analytics.ncsu.edu/sesug/2007/PO10.pdf>

Accessed 29 July 2018.