Disclaimer: this analysis has nothing to do with Corporación Favorita, it is an analysis done for educational purposes only.

Forecasting Sales and Extracting Insights

Introduction

Corporación Favorita is a conglomerate that operates in many industries. The oldest of their lines of businesses is a chain of supermarkets and convenience stores. They are present in many Latin American countries with Ecuador being the place where the headquarters are located and where they conduct most of their activity.

1. Goal setting

- Extract insights from the sales of a particular store of Corporación Favorita
- Create a time series model to forecast sales for families of products for the upcoming 15 days

Check the problem identification document for more information.

2. Data

Our available dataset consists of 6 sub-datasets. Here I specify each of them:

- `train:` 125.497.040 rows and 6 columns
- `items:` 4.100 rows and 4 columns
- `stores:` 54 rows and 5 columns
- `transactions:` 83.488 rows and 3 columns
- `holidays:` 350 rows and 6 columns
- `oil:` 1.218 rows and 2 columns

Here are the first 5 rows of each dataset:

Train

	id	date	store_nbr	item_nbr	unit_sales	onpromotion
0	0	2013-01-01	25	103665	7.0	NaN
1	1	2013-01-01	25	105574	1.0	NaN
2	2	2013-01-01	25	105575	2.0	NaN
3	3	2013-01-01	25	108079	1.0	NaN
4	4	2013-01-01	25	108701	1.0	NaN

Stores

	store_nbr	city	state	type	cluster
0	1	Quito	Pichincha	D	13
1	2	Quito	Pichincha	D	13
2	3	Quito	Pichincha	D	8
3	4	Quito	Pichincha	D	9
4	5	Santo Domingo	Santo Domingo de los Tsachilas	D	4

Items

	item_nbr	family	class	perishable
0	96995	GROCERY I	1093	0
1	99197	GROCERY I	1067	0
2	103501	CLEANING	3008	0
3	103520	GROCERY I	1028	0
4	103665	BREAD/BAKERY	2712	1

Transactions

	date	store_nbr	transactions
0	2013-01-01	25	770
1	2013-01-02	1	2111
2	2013-01-02	2	2358
3	2013-01-02	3	3487
4	2013-01-02	4	1922

Oil

	date	dcoilwtico
0	2013-01-01	NaN
1	2013-01-02	93.14
2	2013-01-03	92.97
3	2013-01-04	93.12
4	2013-01-07	93.20

Holidays

	date	type	locale	locale_name	description	transferred
0	2012-03-02	Holiday	Local	Manta	Fundacion de Manta	False
1	2012-04-01	Holiday	Regional	Cotopaxi	Provincializacion de Cotopaxi	False
2	2012-04-12	Holiday	Local	Cuenca	Fundacion de Cuenca	False
3	2012-04-14	Holiday	Local	Libertad	Cantonizacion de Libertad	False
4	2012-04-21	Holiday	Local	Riobamba	Cantonizacion de Riobamba	False

More information on each dataset can be found on the data wrangling notebook

3. Methodology

I followed a method that consists of 6 steps:

- **Problem Identification:** this step involves identifying the correct problem to solve. Mentioned above.
- **Data Wrangling:** data collection, organization, and definition.
- **Exploratory Data Analysis:** creating plots and charts to understand the relationship between data.
- **Pre-processing and Training Data Development**: standardizing our data for future modeling.
- **Model Creation:** selecting, training and deploying a model to make predictive insights.
- **Presenting results:** creating a final report and presenting the results.

4. Data Wrangling

Data Wrangling Notebook

In terms of data wrangling, I had to do a few corrections:

- The first step was reducing the large dataset. To do this I selected one of the 54 stores. This made our analysis much more manageable. The same process used in our analysis can be used for each of the other stores.
- Deal with some missing values on the oil dataset (I used forward fill to do this)
- Clean some data types in most of the datasets (for example convert dates from strings to datetime formats)
- Creating some new features like day of the week
- Merging the datasets to have everything in one

5. EDA

EDA Notebook

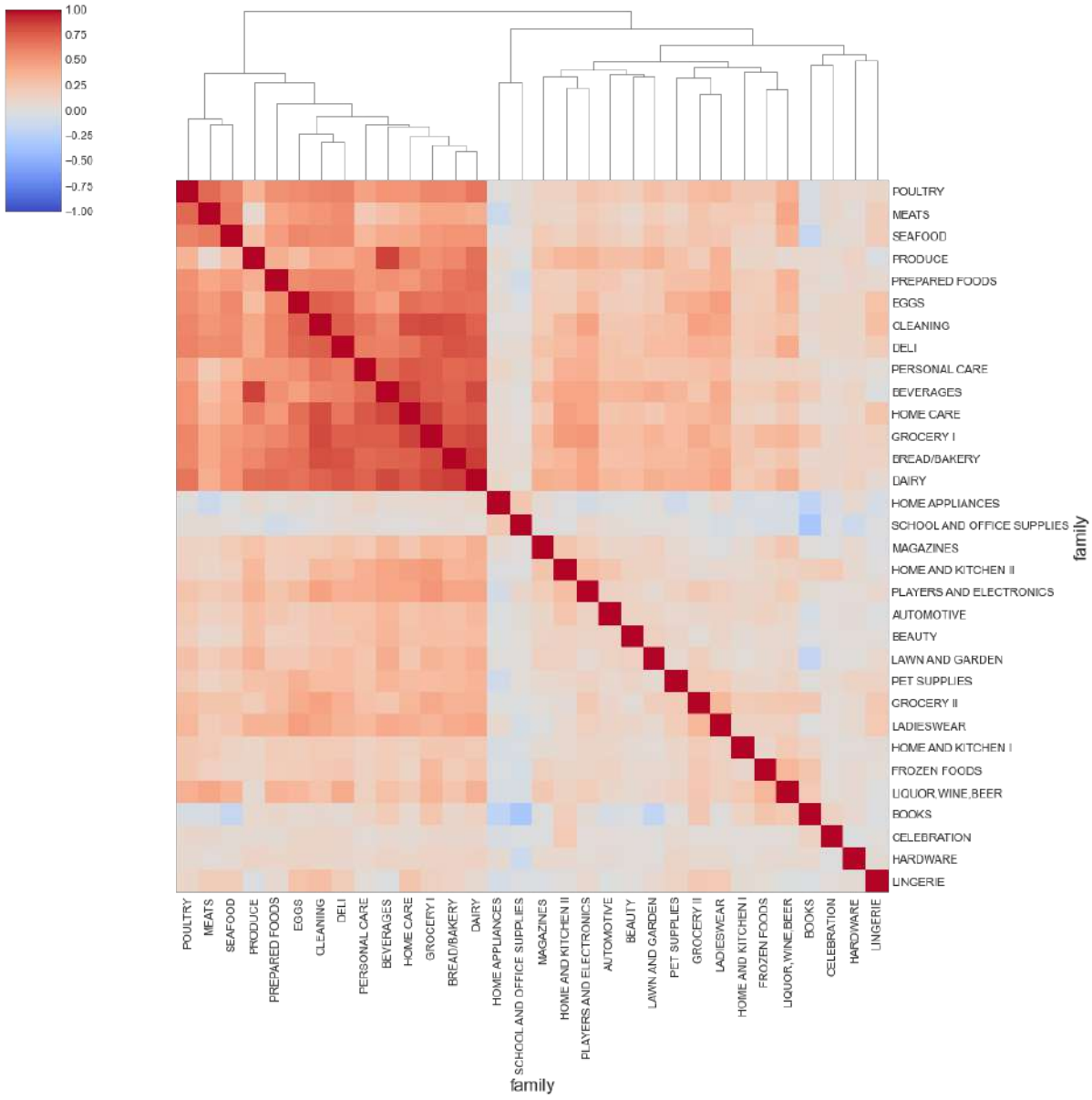
For the Exploratory Data Analysis (EDA) I wanted to be mindful about what our goal was. The dataset is so large that I could have taken many routes to extract insights. For this project I decided to focus on three key business questions that seemed interesting for me:

1. Are sales of families of products correlated between them?
2. What family of products are most influenced by promotions?

3. What family of products are sold more seasonably?

Let's look at them individually:

1. **Are sales of families of products correlated between them?**



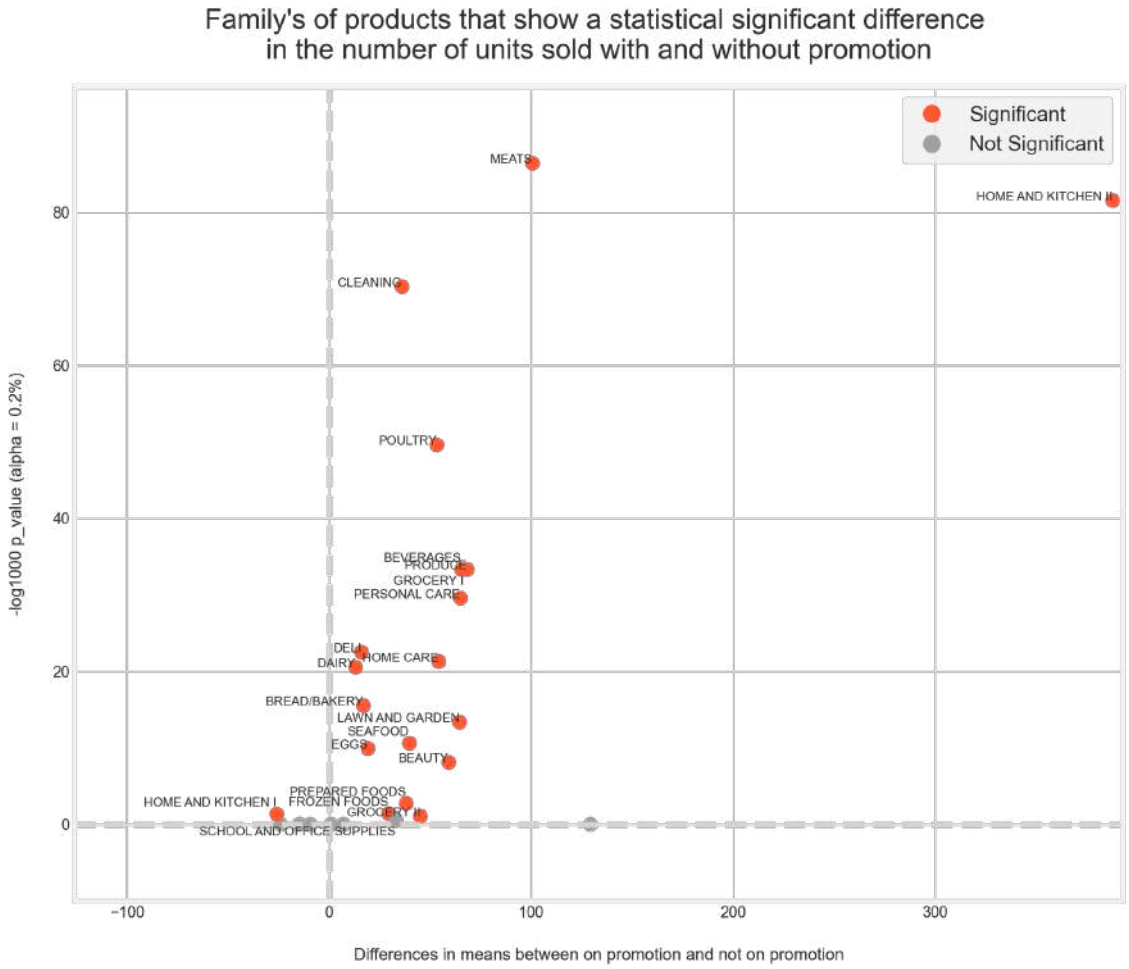
The clustermap shows there is a strong correlation between the unit sales of the first 14 product families of our dataset. As one might expect, these products are mostly perishable and food related products.

However, this information can still be useful. Some supermarkets use the strategy of putting far apart products that sell well together. This, in turn, makes clients walk

more and thus, seeing other products they might not have considered buying on the first hand.

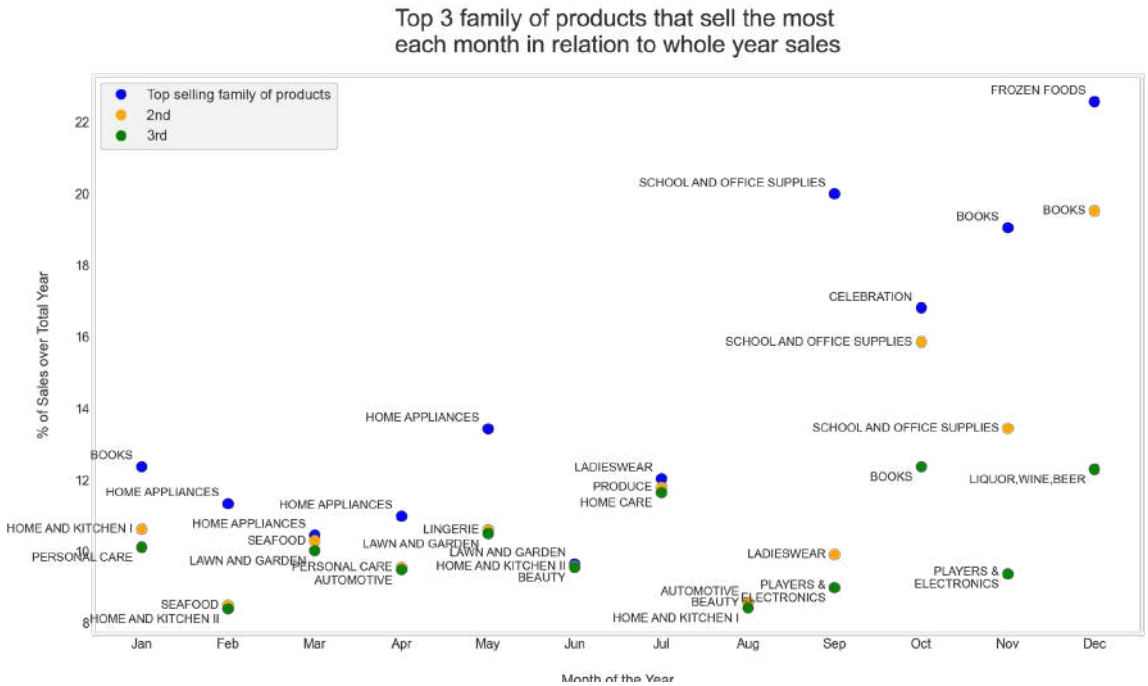
Conversely, if you want to make it easier for clients to find the products that sell well together, this clustermap can also be useful for that.

2. **What family of products are most influenced by promotions?**



This volcano plot shows that some families of products are very responsive to promotions. Product in the categories of “home and kitchen II” and “meats” sell much more on average than when they are not on promotion. In other families of products like ”pet supplies” and “celebration” there is no effect with promotions. Interestingly, products inside the “Home and Kitchen I” category seem to have the opposite effect, when products are in promotion they sell less.

3. **What family of products are sold more seasonably?**



This plot is another way of understanding seasonality. We see that home appliances sell well from February to April, while School and Office Supplies sell a lot from September, to November. Interestingly, almost 25% of the frozen food sales are made in December. This type of information can be useful to decide when to make promotions, when to advertise certain products, and when not to do it.

6. Feature Engineering & Preprocessing

Feature Engineering and Preprocessing Notebook

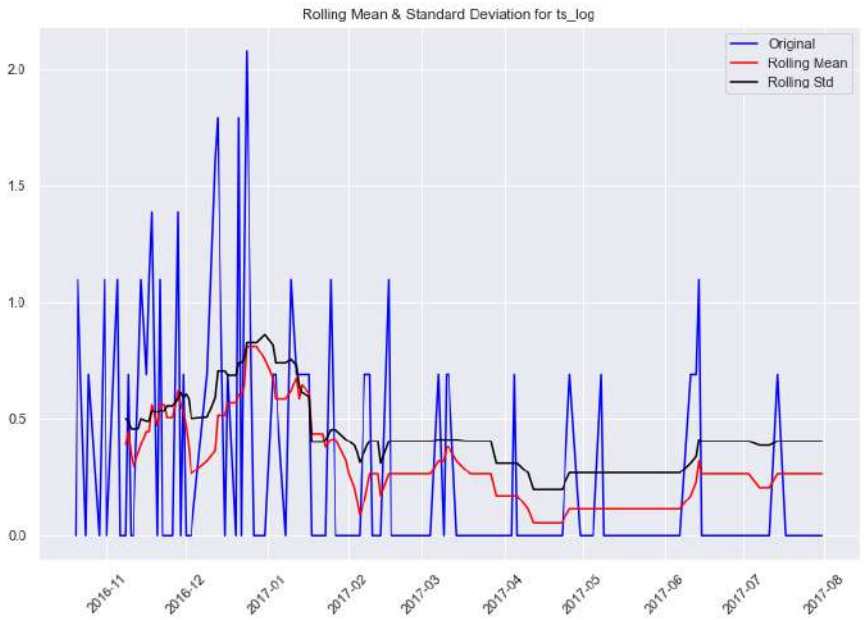
Most time-series models assume that the underlying time-series data is stationary. Stationarity is a statistical assumption that a time-series has:

- Constant mean
- Constant variance
- Autocovariance does not depend on time

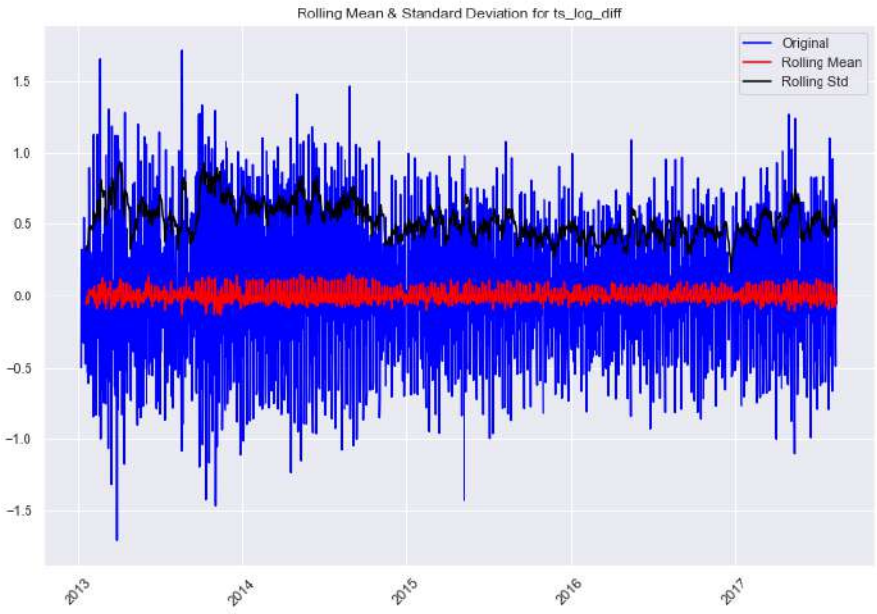
For this section, I focused on one main thing: finding which families of products the data were not stationary and apply the necessary changes to convert them into stationary. This was not very complicated because we only had two families of products that were not stationary, these were: Books and Poultry. To detect this, we did a Dickey-Fuller Test. Here is an example:

```
Results of Dickey-Fuller Test:
Test Statistic                -1.237155e+01
p-value                        5.267526e-23
# Lags Used                    2.200000e+01
Number of Observations Used    1.649000e+03
Critical Value (1%)           -3.434322e+00
Critical Value (5%)           -2.863294e+00
Critical Value (10%)          -2.567704e+00
dtype: float64
```

- **Books**: we did a simple log transformation



- **Poultry**: we did a log transformation and the first difference (n+1) - n.



7. Algorithms & Machine Learning

Modeling Notebook

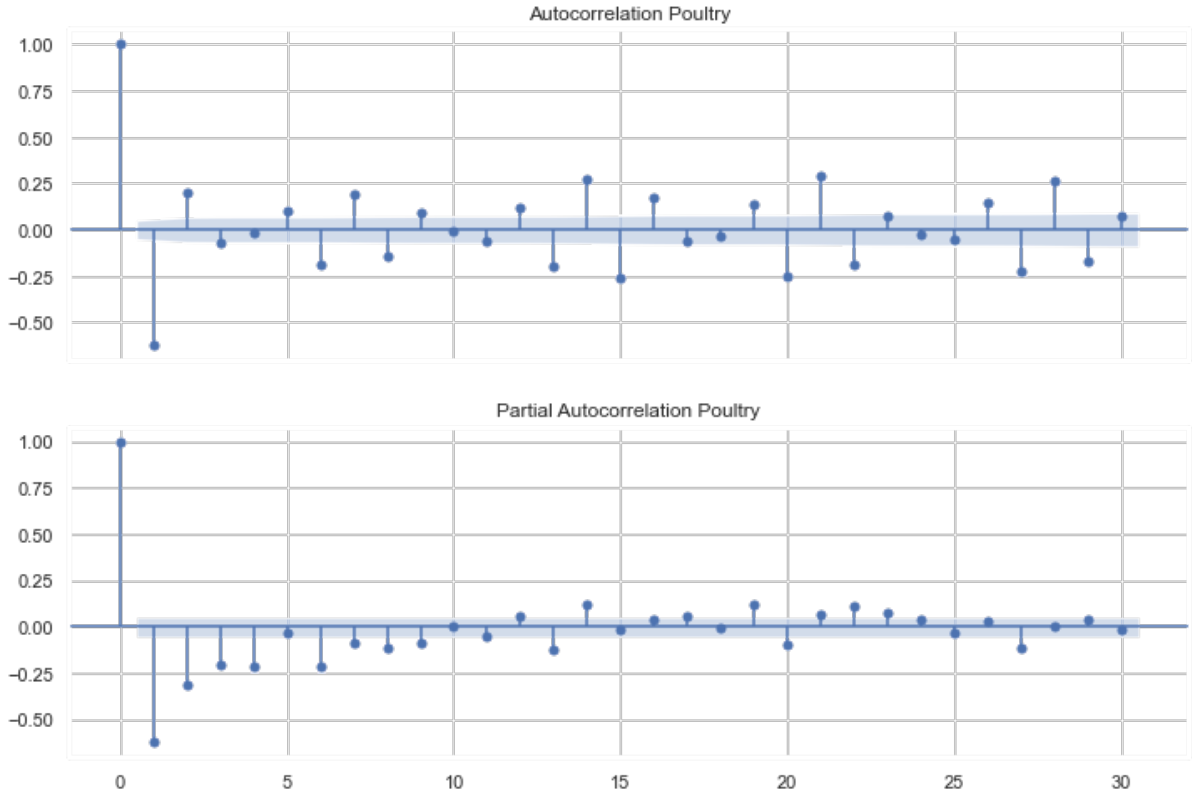
For the modeling section we compared three different models for each family of products, these models were:

- **ARIMA** (Autoregressive Integrated Moving Average) Model (link)
- Facebook's **Prophet** Model (link)
- **LSTM** (Long Short Term Memory) Model (link)

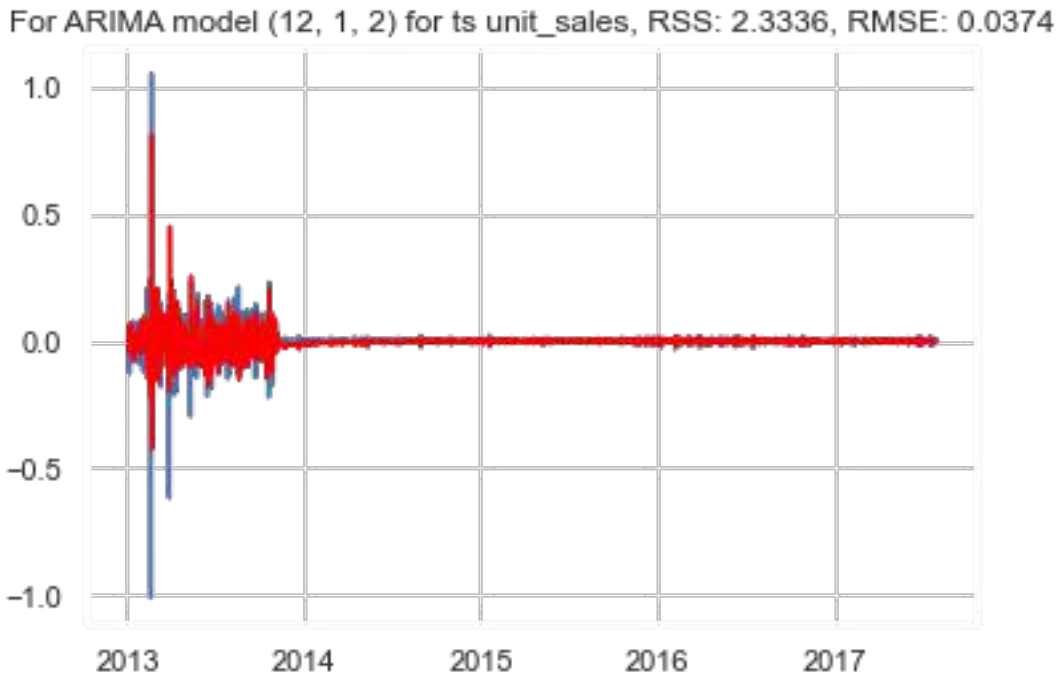
As a reminder, our goal was to predict the sales of the next 15 days (from August 1st to August 15th for each of the families of products. Unfortunately, for the families of products "Home Appliances" and "School and Office Supplies" we did not have data to test our models on (they were thus dropped).

ARIMA MODEL

For the ARIMA model we started by calculating the autoregressive parameter (p) and the moving average parameter (q) through the ACF (Autocorrelation Function) and the PACF (Partial Autocorrelation Function).

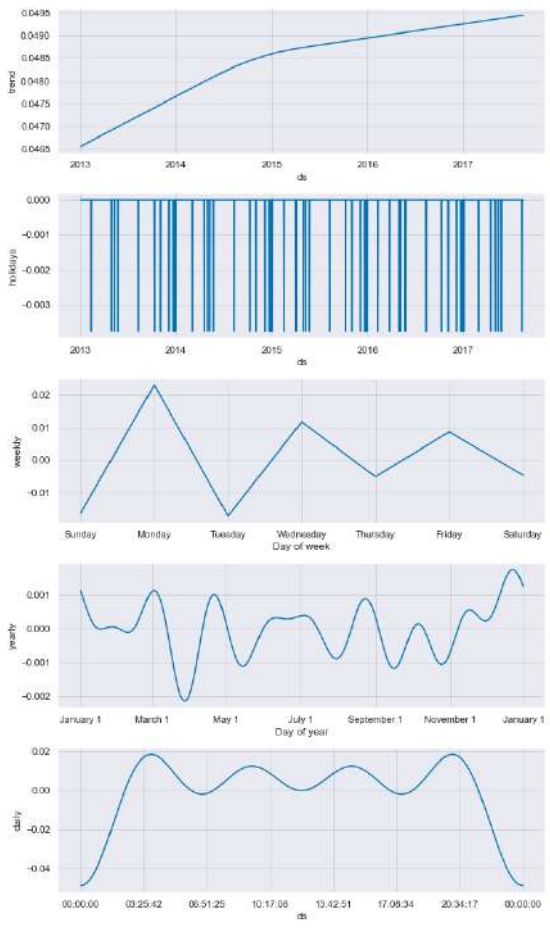


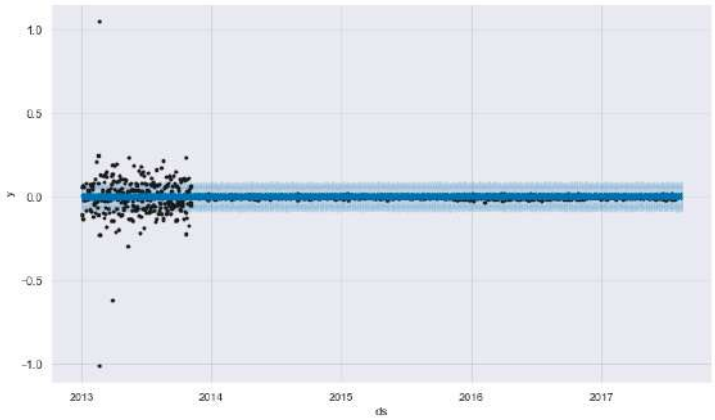
With the newly found parameters for each family of products we constructed the ARIMA models and calculated the RMSE with the test data.



FB'S PROPHET

Facebook's Prophet is a great solution for a quick forecast and easy-to-use procedure. In comparison to the ARIMA model, with FB's Prophet we were able to include information about holidays in our analysis and, as we will see in the final result slide, this model turned out to be the best predictor for most of our families of products.



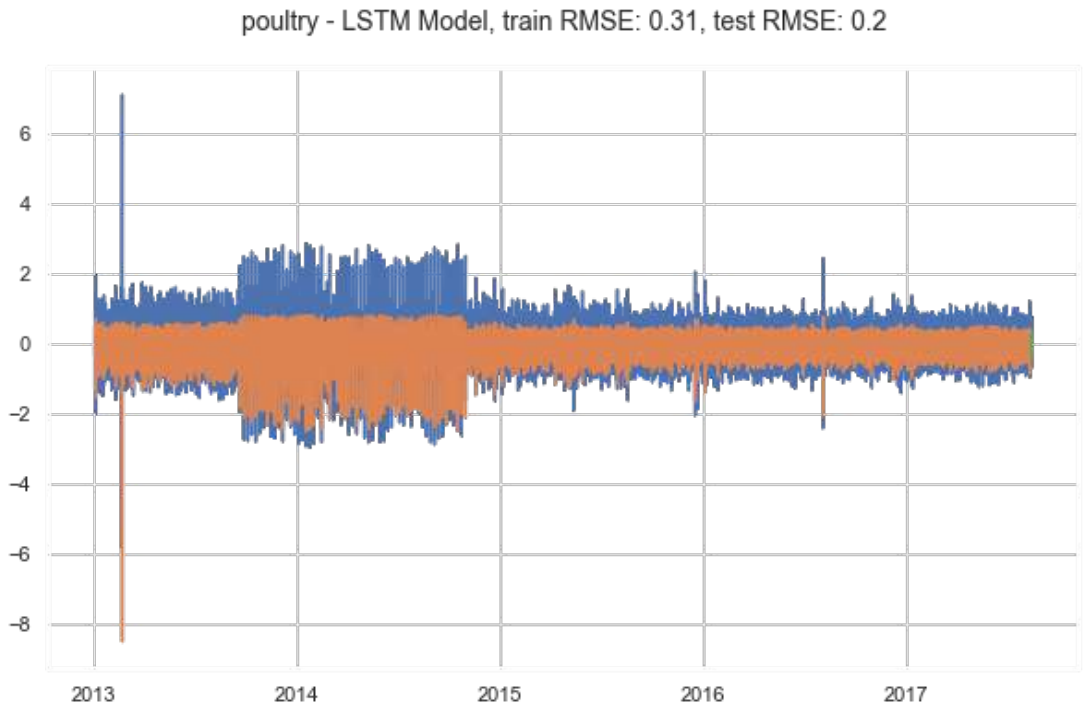


LSTM

The LSTM I built was a simple one, I added 4 units and I used the tanh activation function and the sigmoid recurrent activation function.

The results of the LSTM show a lot of variability from family of products to family of products. For a few product families it worked out well, whilst for others the results were far from optimal.

My LSTM model was very simple, adding more layers and tweaking the other parameters will surely provide better results.



THE RESULTS

In the table below we see a summary of the RMSE results for each of the families of products. As we can observe, the green boxes indicate which model performed the best for each category. In general, it can be observed that FB's Prophet outperformed the other two with some exceptions. The ARIMA models follow in second place and thirdly the LSTM which, as aforementioned, shows very high variability.

	ARIMA	Prophet	LSTM
AUTOMOTIVE	0.525746	0.541472	0.465490
BEAUTY	1.340806	1.198403	1.955809
BEVERAGES	1.466212	0.955745	14.786715
BREAD/BAKERY	1.451920	0.758671	6.682135
CELEBRATION	0.435983	0.685401	1.280236
CLEANING	0.887524	0.497842	1.863060
DAIRY	1.269237	0.550138	1.503198
DELI	0.513029	0.346345	2.088821
EGGS	1.148606	0.660953	6.742007
FROZEN FOODS	1.596824	1.531171	9.257350
GROCERY I	0.962494	0.479726	3.092540
GROCERY II	1.818175	0.883041	2.997685
HARDWARE	0.892242	0.904758	0.598750
HOME AND KITCHEN I	1.495469	1.726396	1.729781
HOME AND KITCHEN II	0.554449	1.603369	2.056229

	ARIMA	Prophet	LSTM
HOME CARE	0.688978	0.443669	2.402029
LADIESWEAR	0.717919	0.757420	0.674770
LAWN AND GARDEN	0.791520	0.853489	2.111415
LINGERIE	2.076734	2.225019	4.171247
LIQUOR,WINE,BEER	1.409355	1.707429	4.563293
MAGAZINES	1.299611	1.227522	1.368692
MEATS	1.868373	1.631309	11.629864
PERSONAL CARE	0.714965	0.706157	0.993362
PET SUPPLIES	0.748577	0.644792	0.720150
PLAYERS AND ELECTRONICS	0.729261	0.681304	3.204853
POULTRY	0.007248	0.015855	0.200142
PREPARED FOODS	3.018114	3.040857	6.140338
PRODUCE	3.891565	3.262213	5.593525
SEAFOOD	2.425937	2.622095	2.723228

8. Future Improvements & Next Steps

- One of the first improvements would be to use all the existing data. I used the data from 1 store but managing it together might shed some interesting insights for the whole company. One way to do this would be to use external servers with more computing power, leverage the power of PySpark or Dask, and be more efficient on my code.

- There are more insights to be extracted from the dataset. I focused on analyzing the promotions and product families but there are also "class" and "individual products". These, however, have no description attached to it, so product 590 we know is from the "poultry" family but nothing else. Getting this information would be ideal.

- Test other forecasting models. We tested three models (ARIMA, FB Prophet, and LSTM) but we have many other tools at our disposal. A first possibility would be to try a multivariate time series forecasting. Only with FB's Prophet model I added the variable "holidays" but we can also consider using the daily oil prices, the day of the week, etc. Also, a lot can be gained from tweaking the LSTM model.

9. Credits

Thanks to Ben Bell for his amazing support and recommendations as a Springboard mentor and to Ines Guix for all her recommendations on statistical analyses.