Springboard – DCS

Capstone Project 3

Forecasting Sales of Multiple Products Across Multiple Favorita Stores

By Michael Bobal

April 2023

(1) Introduction

As a small farmer who supplies produce to a restaurant, I have an intimate understanding and profound respect for the process behind how businesses stock items. The forces of supply and demand are always in play and temporal cycles exist as an important factor, so any ability to predict future sales is essential.

In order to stay in business, restaurants and commercial grocery stores must offer prices that are commensurate with competitors, offer deals to entice customers, and accurately predict which products, and the quantity of those products, to keep in stock. These considerations are confounded by the effect of both seasonal and regional trends.

Especially for grocers, the consequences of poor inventory management are dire. Perishable items like fruits and vegetables can rot before selling if they are overstocked. Conversely, many locations do not have the real estate or capability to store overstocked, low-demand items that are not selling. According to Retail Wire, overstocking costs the average retailer 3.2% in lost revenue, while understocking items can cost 4.1%. A review of the data has shown that overstocks are costing retailers $123.4 billion every year, and understocks remove another $129.5 billion from net inflows.

The goal was to use machine learning time series analysis to forecast sales of different types of items across dozens of stores. I wanted to empower Favorita to become more efficient with its distribution of resources, and inform the company of the best times to offer discounts, whether to stock up on certain items, and knowledge of general market trends.

Our analysis resulted in a model that produces many accurate and actionable forecasts that Favorita could use to ensure that they do not over- or under-stock items.

The notebooks detailing the process which led to our conclusions can be found at the following link:

https://github.com/mikebobal/springboard/tree/main/Capstone%203

(2) Approach

(2.1) Data Acquisition and Wrangling

The data is from a Kaggle competition for forecasting Favorita sales for 54 stores across 33 product families. The total number of time series to forecast numbered 1782. The data came to me fairly organized already, with individual items already aggregated by product family. There were missing dates for every 12/25 in the dataset (presumably due to store closure on the Christmas holiday). For the purpose of having those dates included in the models, we added the dates and inferred the sales using the monthly average.

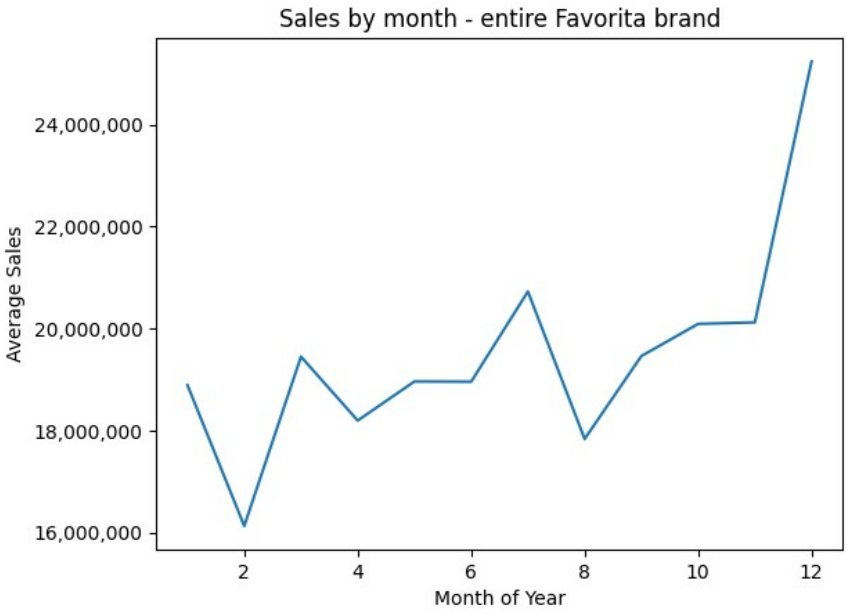
The data files presented us with 3,000,888 data points, comprising every daily sales total across every store and product category over every date from 01/01/2013 to 08/15/2017 – 1688 days.

There were actually multiple CSV files that included exogenous variables such as transaction information, oil prices, and holidays/events. Other data available to the analysis were promotions, store cluster/type, and store city/state location. Missing data in these files were inferred using forward fill and back fill. This cleaned dataset was saved for the next step.

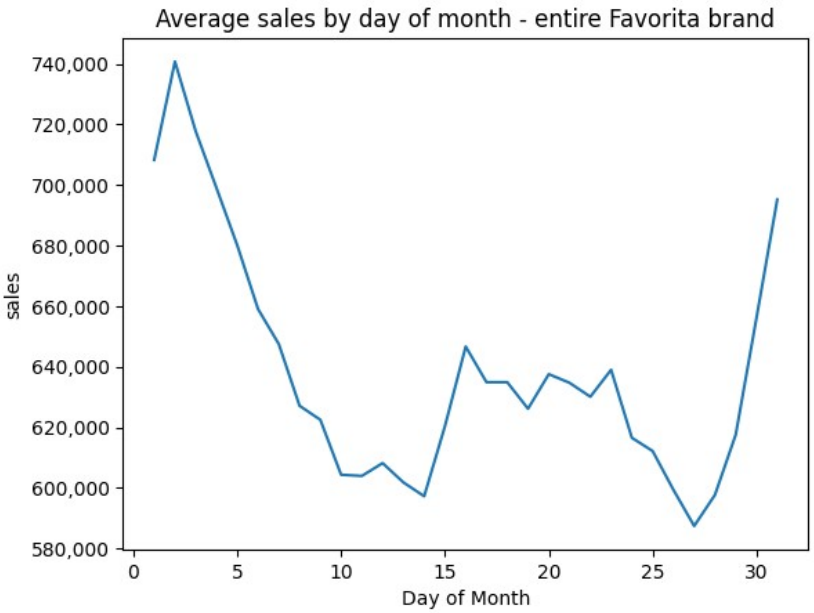
(2.2) Storytelling and Inferential Statistics

During our exploratory data analysis, we saw how there were both missing dates on each 12/25 (time gaps), and missing values (some dataframes with large time periods of zero values). These were issues since we wanted consistent data to make more accurate predictions. Large periods of missing values on any of the 1782 datasets were imputed using exponential smoothing. Then, all datasets had 12/25 added to the index, with forward fill being utilized for the sales information.

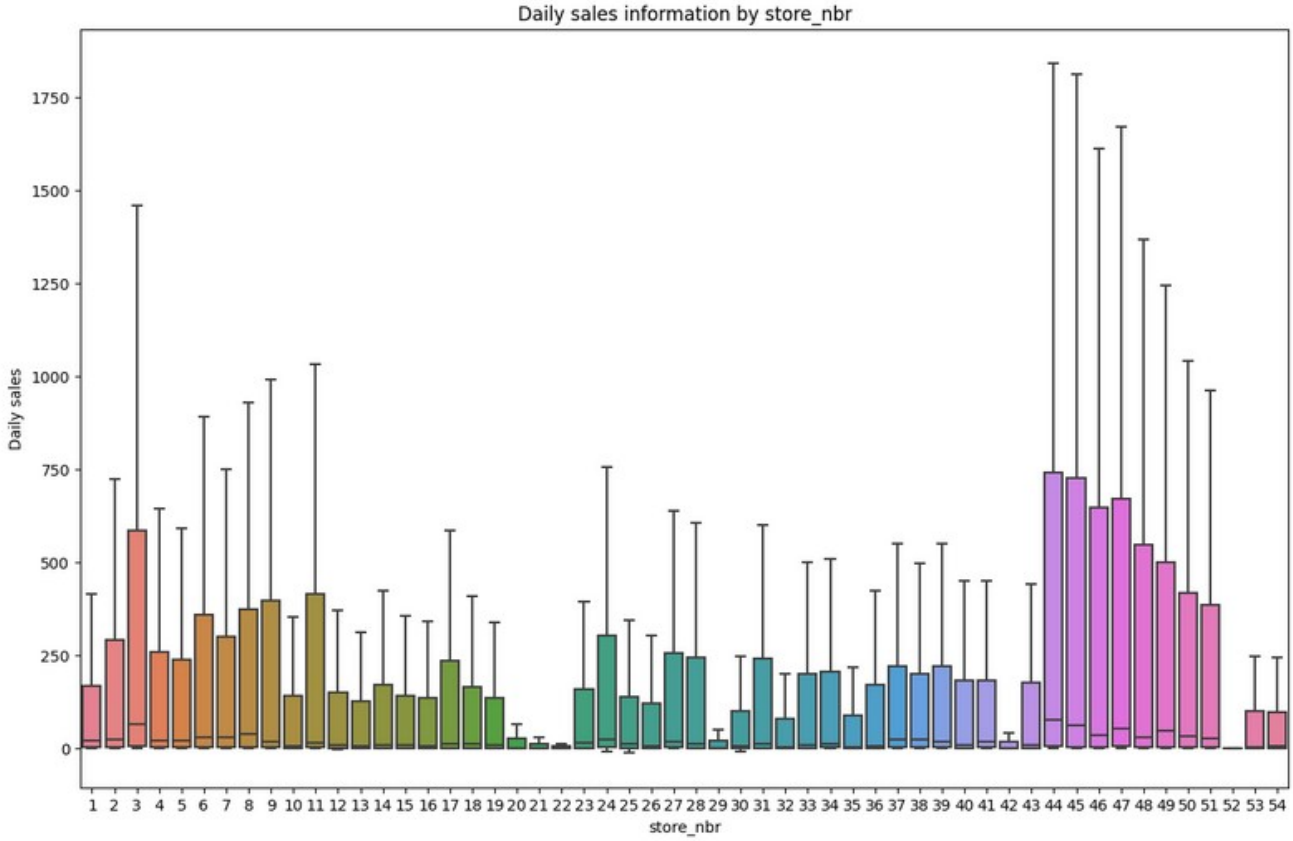
Visualizing the data was informative. Overwhelmingly, the sales in December are the largest, followed by a mid-summer local peak in July. It appears that customers decrease their shopping budgets in February and August. Interestingly, both of these troughs occur within 2 months of the aforementioned peaks. Perhaps there is an element of 'buyers exhaustion' happening, as people could be choosing to stock up some months, causing lower demand in the following months.



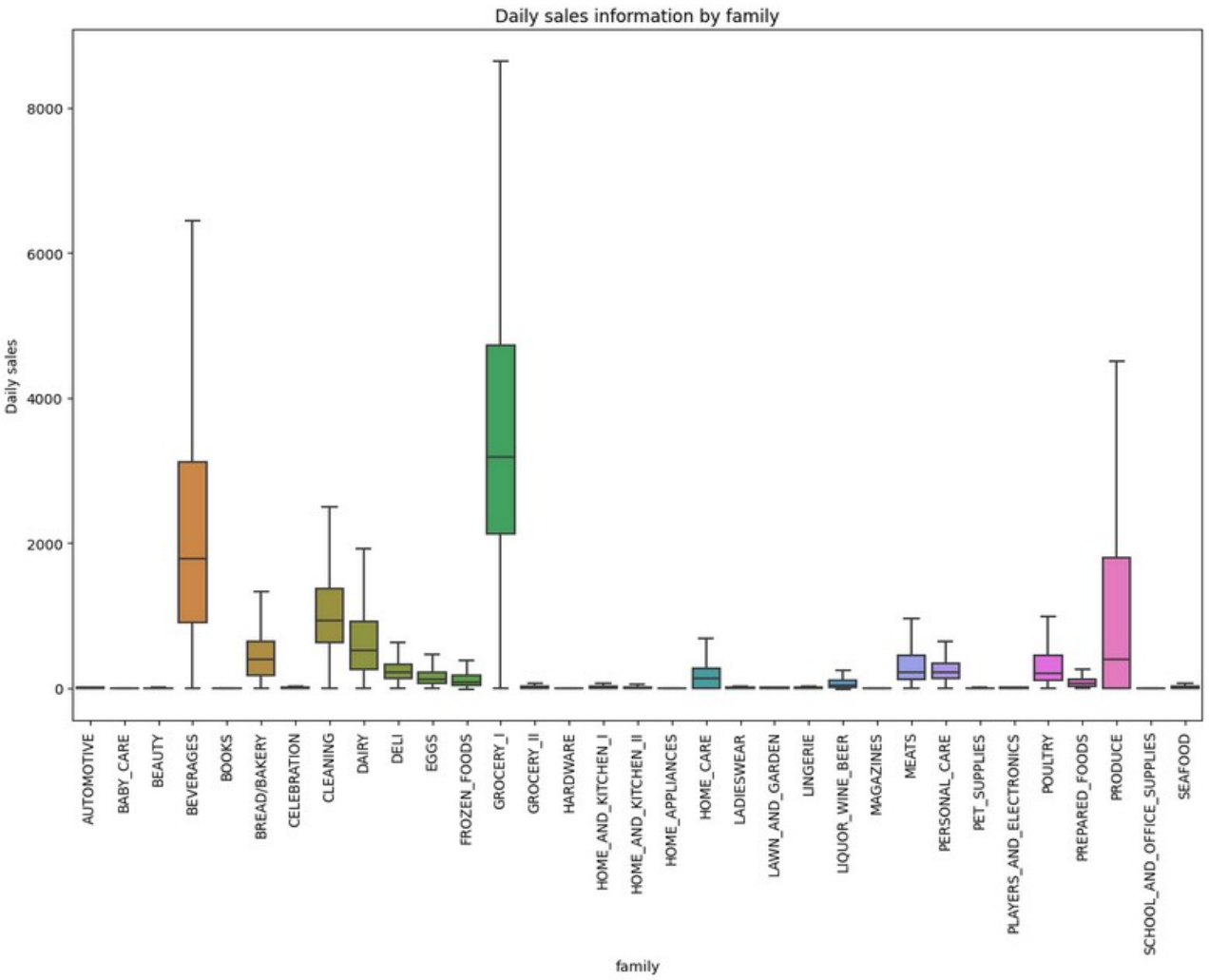
The intra-monthly pattern shows sales starting each month strongly, with a steady decline until the 14th day, a sharp increase which plateaus until the 23rd. Then a trough occurs around the 25th, followed by a steep incline to finish out the month.



Kaggle tells us that public sector employees are paid bimonthly on the 15th and the last day of each month. That would help to explain the sudden uptick in sales on the 15th and final days of each month.

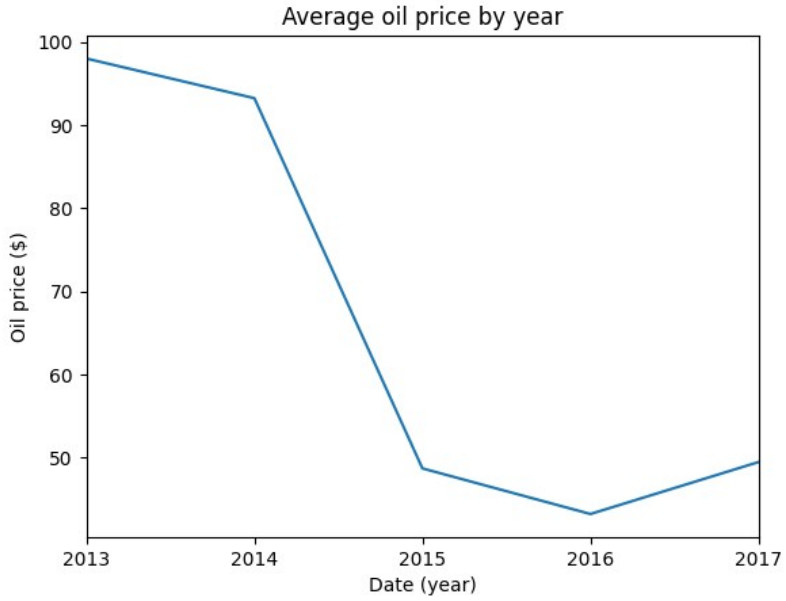


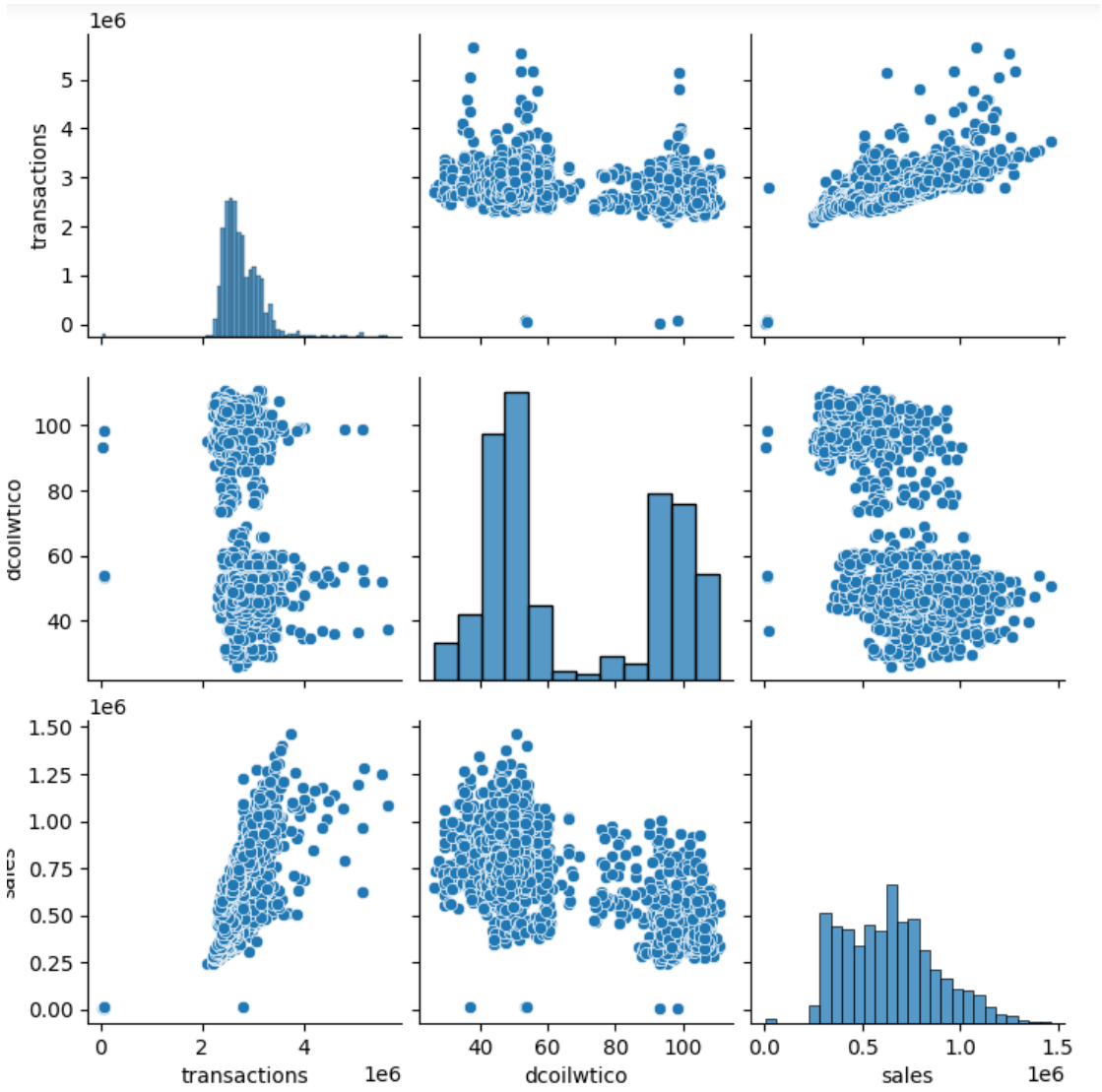
Looking at a breakdown of sales by store number, we can see that stores 3, 44, 45, 46, and 47 see the most volume. Some stores, like 20, 21, 22, 29, 42, and 52, seem to have close to no sales.



For product family, 'GROCERY I' takes the cake, followed by 'BEVERAGES' and 'PRODUCE'. The non-consumable product family with the highest median sales is 'CLEANING'.

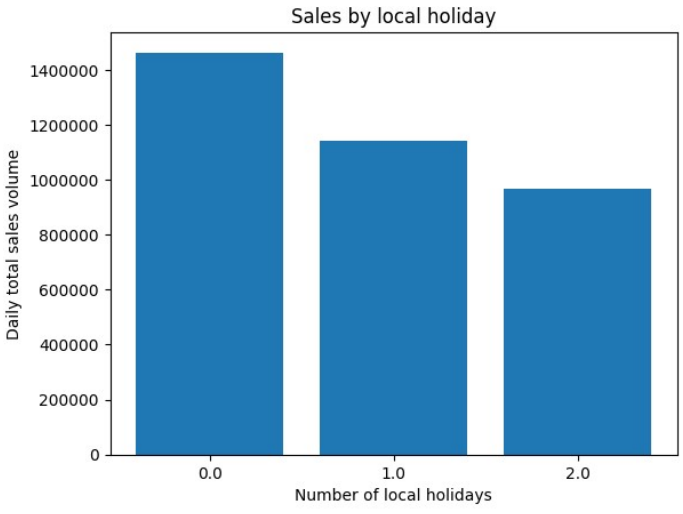
Oil prices experienced a precipitous drop midway through the data, and ended up not being used by our final model.

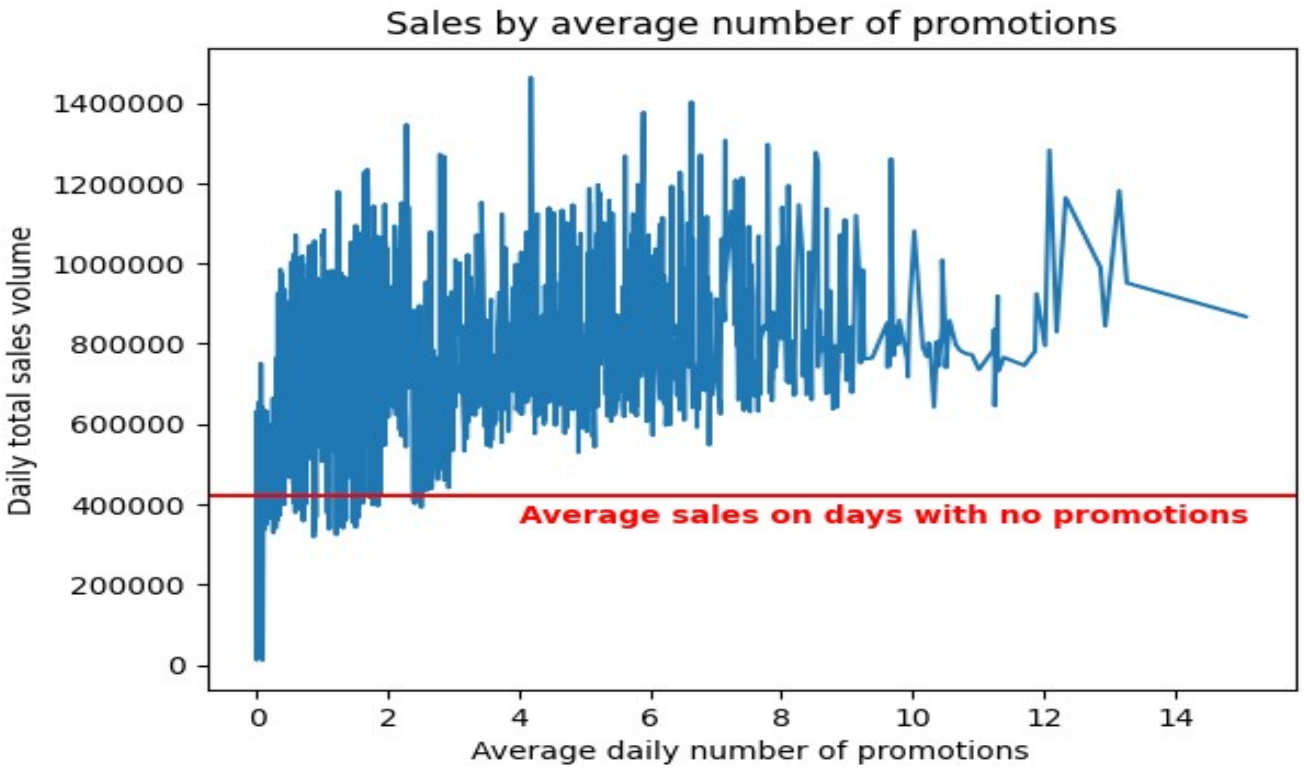




The interesting bimodal-looking distribution of each scatterplot involving oil is likely the result of the precipitous oil price decrease midway through the data. As would be expected, sales correlate strongly positive with transactions: 0.68. There is a strong negative correlation between sales and oil price: -0.624.

The presence of local holidays anywhere in Ecuador severely decreases the national sales for that day. Regional and national holidays do not seem to have much of an effect on sales, though.



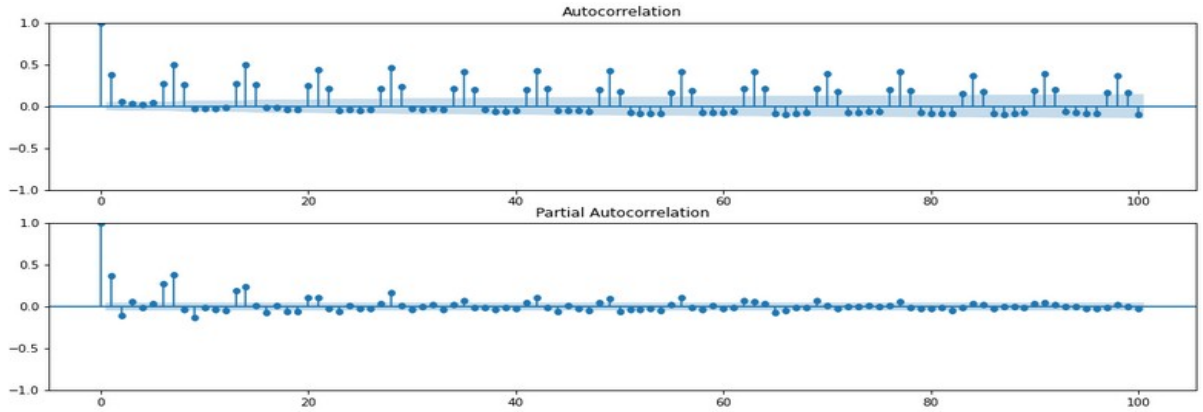


The existence of promotions definitely appears to coincide with an increase in sales. Considering that the daily sales volume is around 40,000 for days with no promotions, most of the time that any promotions are offered, the sales volume is higher.

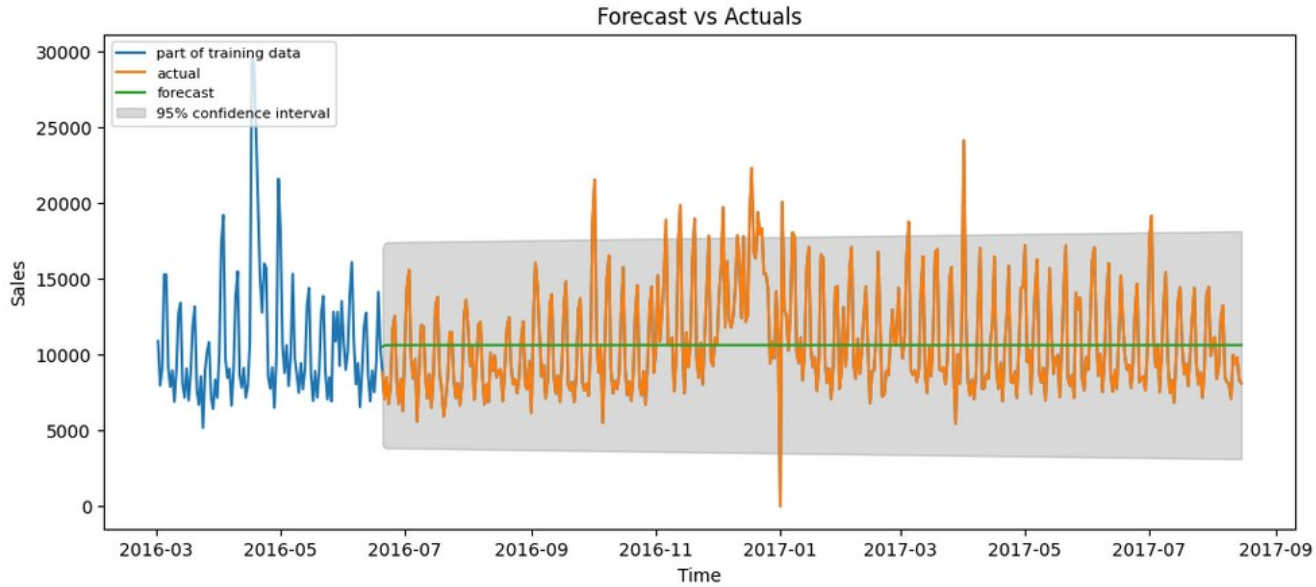
(2.3) Baseline Modeling

The store with the highest sales is Store 44 and the product family with the highest sales is 'GROCERY I', therefore, I decided to use this particular time series as the one to test models on.

We determined the seasonality and periodicity in the data using statistical tests like the Augmented Dickey-Fuller, the KPSS, the OCSB, and the CH tests. Using this knowledge, we visualized the autocorrelation and partial autocorrelation, and determined the p,d,q and P,D,Q for making models using our first algorithm, ARIMA. (p,d,q) was determined to be (1,0,1) and (P,D,Q) was determined to be (0,1,1) with a seasonality of 7. The decision was made to use a combination of Auto Regressive and Moving Average models, ARIMA.



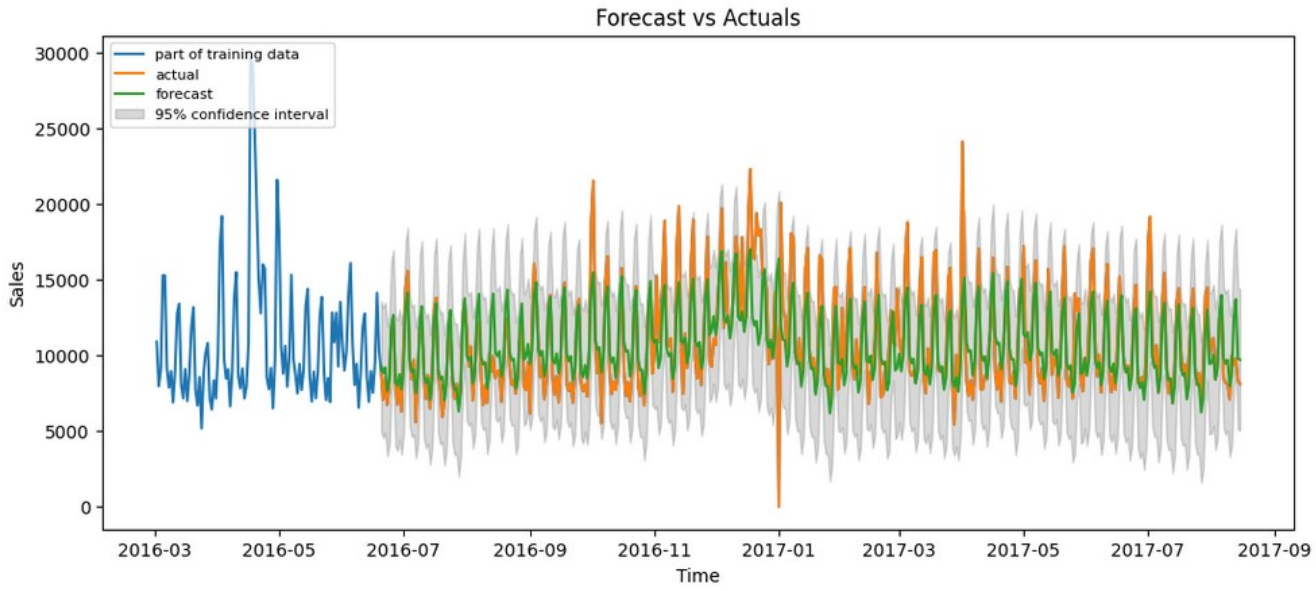
Using the Mean Average Percent Error (MAPE) and the coefficient of determination (R^2) as our performance metrics, we first split the data into training and testing splits. A basic ARIMA with no parameter tuning turned out poorly, as expected.



However, as we tuned the parameters more, the results improved. The seasonal ARIMA done second turned out to have the third-lowest MAPE of the eight models.

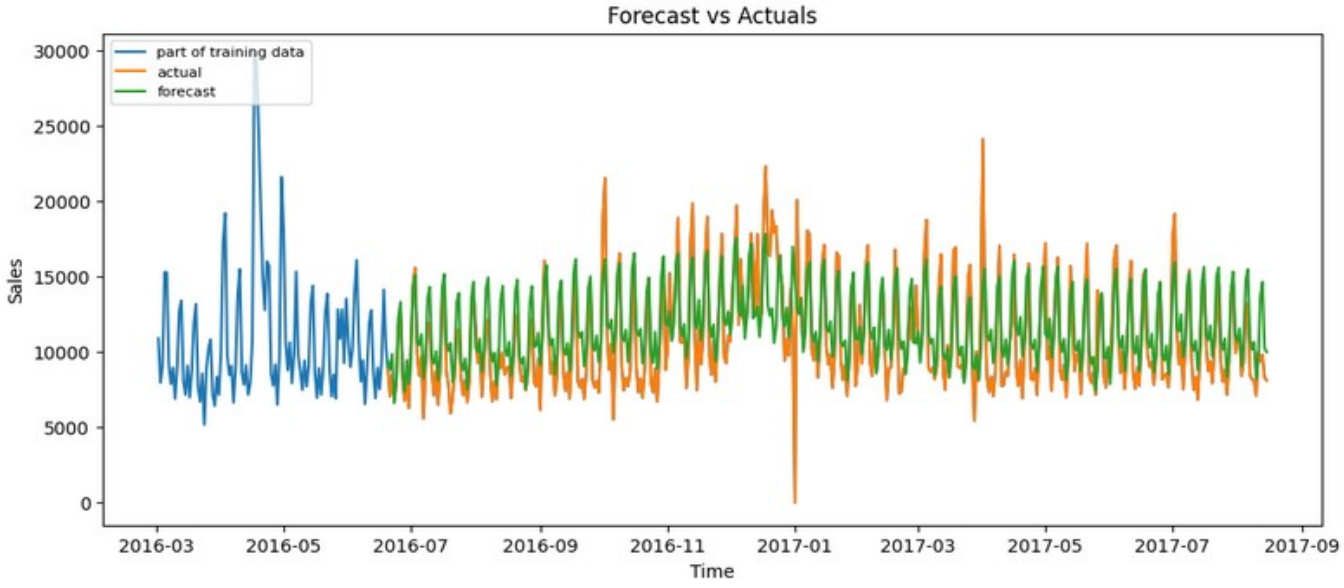
(2.4) Extended Modeling

Following those ARIMAs, we turned to an algorithm which can handle multiple periodicities called TBATS. The original TBATS run scored mediocrely; however, once the precise seasonal periods were used, TBATS quickly jumped to best-performing model, with the lowest MAPE and highest R^2 at the time.



Something that I wanted to try was including exogenous variables in a model. TBATS cannot handle exogenous variables, so we went back to ARIMA and introduced the provided "promotional" items for sale. We also included Fourier transformations to imitate having multiseasonality. These results had varied results.

Using both promotional information and Fourier terms produced the result that ended up with the highest R^2 score, and a very respectable MAPE.



The promotional-info-only ARIMA was terrible. The Fourier-only ARIMA out-performed the Fourier/promotional ARIMA in MAPE, but not R^2.

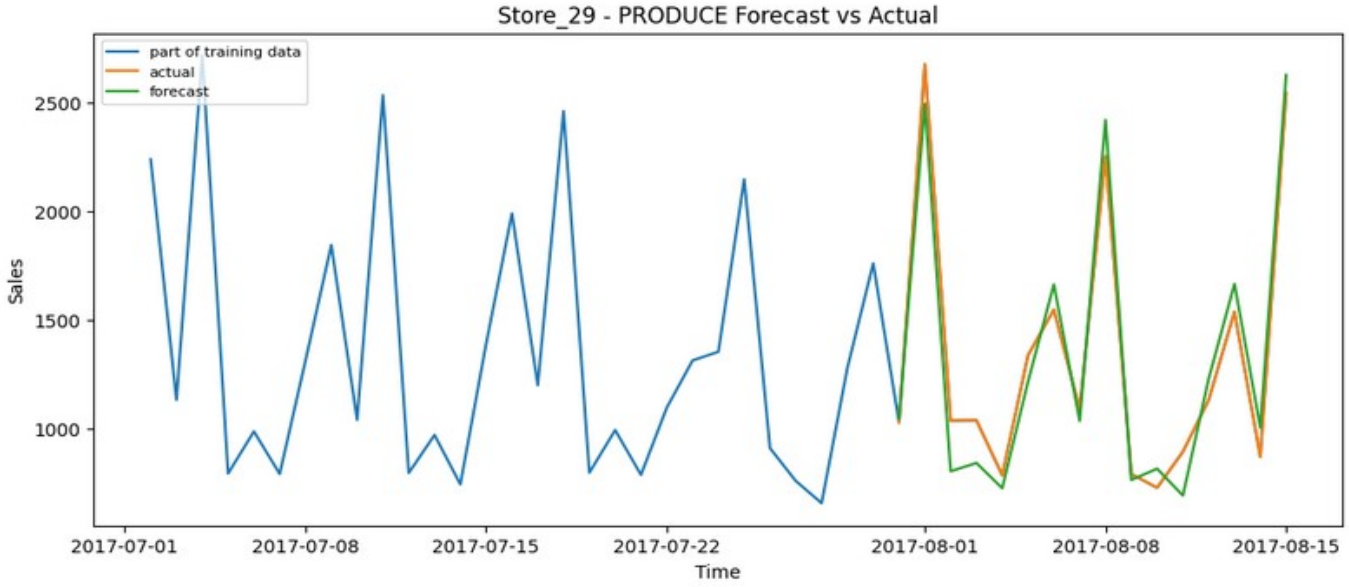
As a final part of the process, we wanted to see the effect of making an ensemble model with the results from our best-performing models for each metric. Combining TBATS and ARIMA with exogenous variables proved to be worth the investment. Winning in MAPE and placing second in R^2, the TBATS/ARIMA model uses calculated weights to give a final prediction.

The major issue with the TBATS/ARIMA model is the inordinate amount of time it takes to run all of our time series (over 3 days). The results would be obsolete by the time the analysis was finished. Therefore, in a business-conscious decision, the exogenous variable inclusive seasonal ARIMA was the chosen model, with a testing MAPE of 16.76, and an R^2 of 0.64902.

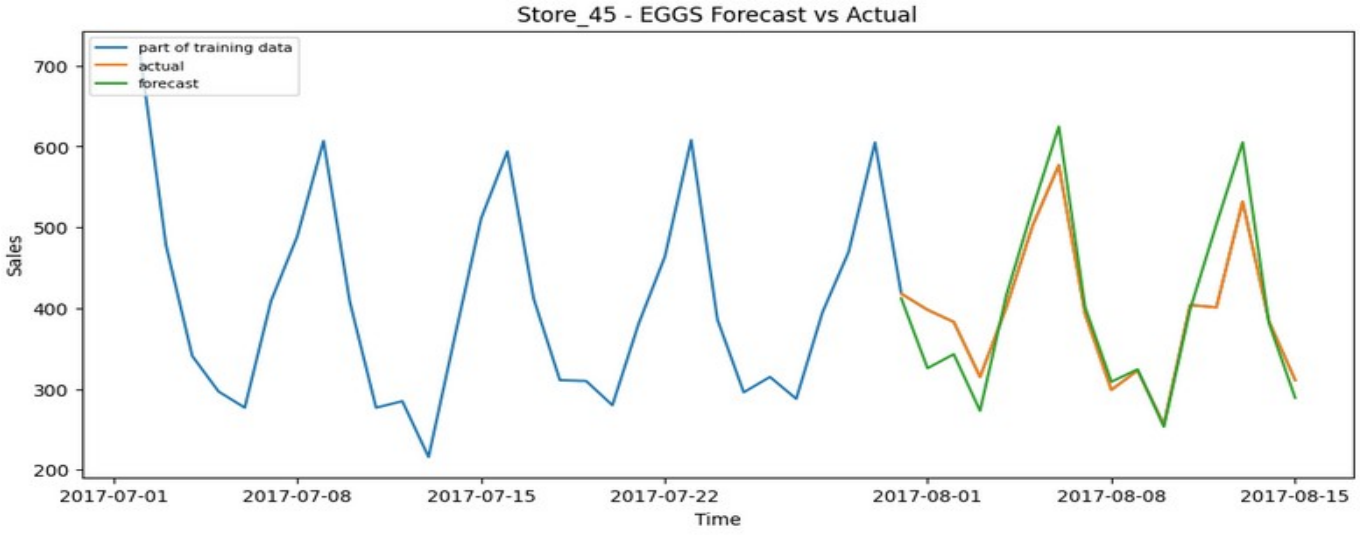
	model_names	MAPE	R^2
4	ARIMA with exogenous variables	16.76	0.64902
7	ARIMA/TBATS combo	13.61	0.63149
3	TBATS2	13.81	0.61295
5	ARIMA with fourier exogenous only	16.82	0.5666
2	TBATS	15.42	0.5561
1	Seasonal_ARIMA	13.93	0.51081
6	ARIMA with promo exogenous only	33.94	0.4344
0	Basic_ARIMA	24.96	0.00222

(3) Findings

I applied the exogenous variable inclusive seasonal ARIMA model selected from the Preprocessing and Training notebook to every one of the 1782 time series in our analysis. Runtime was just over 3 hours, but that number is considerably more time-efficient than other modeling options. The result was 1653 completed forecasts, and 129 time series without enough information to make a proper forecast. Many of the non-analyzable series were simply stores that did not sell certain product types (in which case we would predict zeros for the values).



With the time series that did contain enough information, we were able to see how certain product families had better predictability than others. Specifically, the food and drink items scored much better on our evaluation metrics (R^2 and MAPE) than durable goods items. For our analysis, I chose to trust the R^2 metric over MAPE. According to recent research, while MAPE is more widely accepted as the standard metric, the coefficient of determination (R^2) is the best evaluation metric we could use.



	Store/Product	MAPE	R^2
493	Store_22 - SCHOOL_AND_OFFICE_SUPPLIES	18.09	0.96659
723	Store_29 - PRODUCE	9.23	0.96396
1066	Store_39 - EGGS	52.61	0.95296
822	Store_31 - PRODUCE	7.85	0.95250
789	Store_30 - PRODUCE	12.41	0.95079
1152	Store_40 - PRODUCE	10.00	0.94494
1297	Store_45 - EGGS	7.56	0.93279
1705	Store_7 - LIQUOR_WINE_BEER	15.27	0.92791
855	Store_32 - PRODUCE	10.97	0.92781
1542	Store_51 - MEATS	9.58	0.92661
835	Store_32 - EGGS	25.96	0.92587
750	Store_3 - MEATS	9.87	0.92564
327	Store_18 - PRODUCE	13.36	0.92527
816	Store_31 - MEATS	8.67	0.91414
121	Store_12 - LIQUOR_WINE_BEER	48.85	0.91150
358	Store_19 - POULTRY	20.53	0.90318
420	Store_20 - MEATS	10.58	0.90048
228	Store_15 - PRODUCE	12.43	0.89608
261	Store_16 - PRODUCE	10.80	0.89121
783	Store_30 - MEATS	14.09	0.89065

Overall, my model performed very well, with many of the series having R^2 values greater than 0.90! Looking at the overall performance metrics, I decided to apply weighted averages based on the actual sales volume seen for each product family. A series with a <0.02 R^2 score but sales volume averaging <5 dollars daily is neither informative nor instructive.

Therefore, it was determined that adjusting the overall metrics was the best option. Once the weights were applied, we were able to see overall metrics that were more reflective of the true model accuracy.

	Before weighting:	After weighting:
Whole-project MAPE:	53.2	18.14
Whole-project R^2:	0.327	0.484

(4) Conclusions and Recommendations for the Clients

In conclusion, Favorita daily sales data are able to be accurately predicted using a seasonal ARIMA model which can account for some exogenous variables – at least the consumable product categories. The data provided can be separated into discreet time series based on store and family, and the sales data exhibits a strong multiseasonality, which can be decomposed, modeled, and forecast at least 16 days into the future using models built from multiple algorithms.

Based on this analysis, I can make these recommendations:

1. The seasonal cycles are strong in the food and beverage data. With our model as accurate as it is for consumable items, Favorita can prevent over/under stocking perishable goods by stocking such inventory with the assumption that our model's forecast will become reality (within a reasonable margin of error). This will ensure that customers do not encounter empty shelves, and that Favorita will not lose out on profit with extra items left to rot, unpurchased.

2. The weekly seasonal cycle appears to be the most prominent periodicity, and we know that days with promotions tend to have higher sales. Sundays, in particular, are a peak sales day for consumable items year-round. Since the foot-traffic is already present, offering special Sunday-only promotions on slower-selling non-perishable items could help to sell less-predictable, overstocked inventory. This would allow the stores to allot space more efficiently.

3. Local holidays correlate strongly with decreased sales numbers. Playing into these holidays with well-advertised "special holiday" promotions could help to minimize losses on those days.

(5) Ideas for Future Research

One major pathway which I would like to pursue in a future extension of this analysis is to use different techniques for modeling forecasts that use more of the exogenous variables. Supervised leaning regression algorithms such as Random Forest, XGBOOST, LGBM, etc would be desirable.

Also, it would have been interesting to see how the final R^2 and MAPE metrics would look with a larger forecasting horizon. Instead of only predicting 2 weeks out, I could try to forecast 2 months, 6 months, or 1 year.

(6) Consulted Resources

- Data source: Alexis Cook, DanB, inversion, Ryan Holbrook. (2021). Store Sales - Time Series Forecasting. Kaggle. https://kaggle.com/competitions/store-sales-time-series-forecasting

- Hyndman, R.J., & Athanasopoulos, G. (2021) Forecasting: principles and practice, 3rd edition, OTexts: Melbourne, Australia. OTexts.com/fpp3. Accessed on 03-01-2023

- Mentorship and guidance: AJ Sanchez