Sales Through the Sands of Time: Uncovering Patterns in Retail Data with Time Series Forecasting

Christopher John Apton

Department of Mathematics
University of California, Los Angeles
California
June 2023

Contents

1 Introduction

The data set is a time-series dataset focusing on store sales. We are given multiple data sets on the grocery stores in Ecuador. The machine learning problem is to predict sales for the thousands of product families sold at Favorita stores located in Ecuador. Favorita stores is a large Ecuadorian-based grocery retailer.

We'll have 2 parts, the first one will do some exploratory data analysis and we will better understand the data. This will help into the second portion where we'll use time-series modeling to solve the prediction problem of predicting sales using the test data given from Kaggle.

2 The Data

The first dataset we'll look at is the stores data. This data set contains the primary key of each store number along with some basic information on each store. There are a total of 54 stores in the data set.

	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

The store type is referring to each grocery store chain consisting of key:value pairs of A Megamaxi, B Gran Aki, C SuperMaxi, D Aki, E Super Aki. Also, we have the cluster column which clusters similar stores.

Next, we'll look at the transaction data. This data set ranges between $01/01/2013 - 08/15/2017$ which contains the number of transactions per day for each store.

Additionally, we have the holiday events data. This dataset informs us of important holidays and can help us find unusual activity.

	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

The transferred column represents a holiday that is transferred officially falls on that calendar day, but was moved to another date by the government. A transferred day is more like a normal day than a holiday. The day that it was actually celebrated has the type Transfer.

Looking into the holiday data, we can see how holidays affect sales.

	date	type	locale	locale_name	description	transferred	avg_sales
92	2014-01-01	Holiday	National	Ecuador	Primer dia del ano	False	4.827197
117	2014-07-01	Event	National	Ecuador	Mundial de futbol Brasil: Octavos de Final	False	404.310110
159	2015-01-01	Holiday	National	Ecuador	Primer dia del ano	False	7.168135
211	2016-01-01	Holiday	National	Ecuador	Primer dia del ano	False	9.221882
220	2016-04-17	Event	National	Ecuador	Terremoto Manabi+1	False	713.711414
221	2016-04-18	Event	National	Ecuador	Terremoto Manabi+2	False	755.286535
297	2017-01-01	Holiday	National	Ecuador	Primer dia del ano	True	6.780304
302	2017-04-01	Holiday	Regional	Cotopaxi	Provincializacion de Cotopaxi	False	821.034771
308	2017-05-01	Holiday	National	Ecuador	Dia del Trabajo	False	733.276861

Since this time-series data isn't stationary, I calculated the outliers using a rolling window of size 30 with a z-score threshold of 2.5 to capture most of the outliers in the dataset. As we can see, most of the outliers from holidays occur on new-years which isn't too surprising. There are also some outliers for high sales on the finals of a soccer match, Mundial de futbol Brasil: Octavos de Final. Also, unfortunately, Eduador had an earthquake on Terremoto Manabi so they had abnormally high spending. Provincializacion de Cotopaxi holiday is the anniversary of when Cotopaxi became a province in Ecuador. Lastly, Dia del Trabajo is labor day where they probably spend more.

Looking into the locale column, I observed the Local holidays and prepared them for the train data. Each local holiday corresponds to a city. So, in the train dataset, I made sure to match the local holidays based on the dates where the cities also matching using a left join. Also, since Regional/National holidays aren't city specific and depends only on the date, I created a separate column which signifies if the date is a national holiday or not.

Also, the oil data contains daily oil prices. Specifically, 1 column for the date and 1 for the oil price.

Lastly, we can look at the train/sales data to see what information we are given for the prediction problem later on.

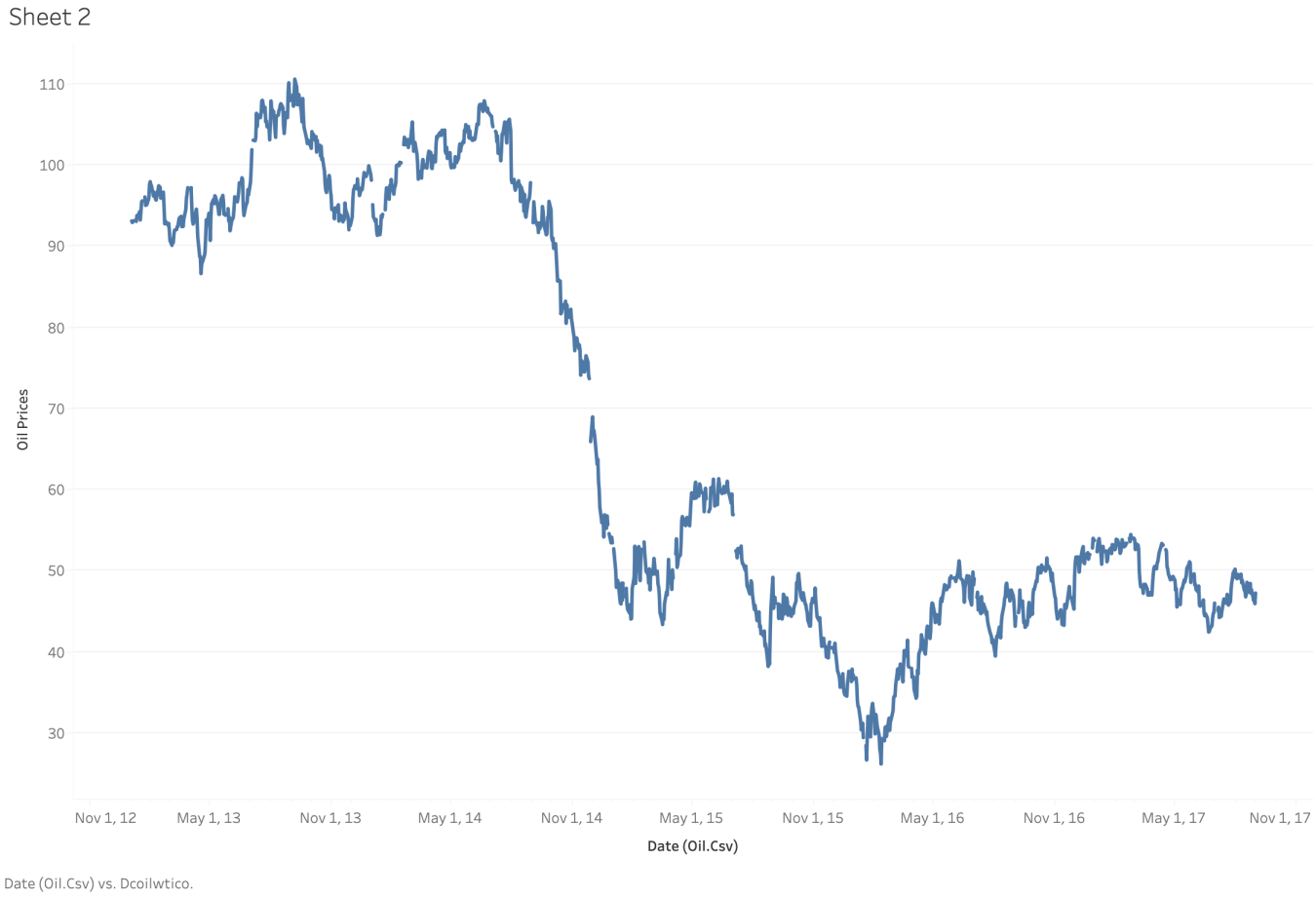
	id	date	store_nbr	family	sales	onpromotion
0	0	2013-01-01	1	AUTOMOTIVE	0.0	0
1	1	2013-01-01	1	BABY CARE	0.0	0
2	2	2013-01-01	1	BEAUTY	0.0	0
3	3	2013-01-01	1	BEVERAGES	0.0	0
4	4	2013-01-01	1	BOOKS	0.0	0

The id column should be dropped or not used as a feature for the ML model. But basically, the train data has key features needed to predict sales. The Sales are the total sales for a product family at a particular store at a given date. Fractional values are possible since products can be sold in fractional units (1.5 kg of cheese, for instance, as opposed to 1 bag of chips). The onpromotion column gives the total number of items in a product family that were being promoted at a store at a given date.

3 Data Analytics

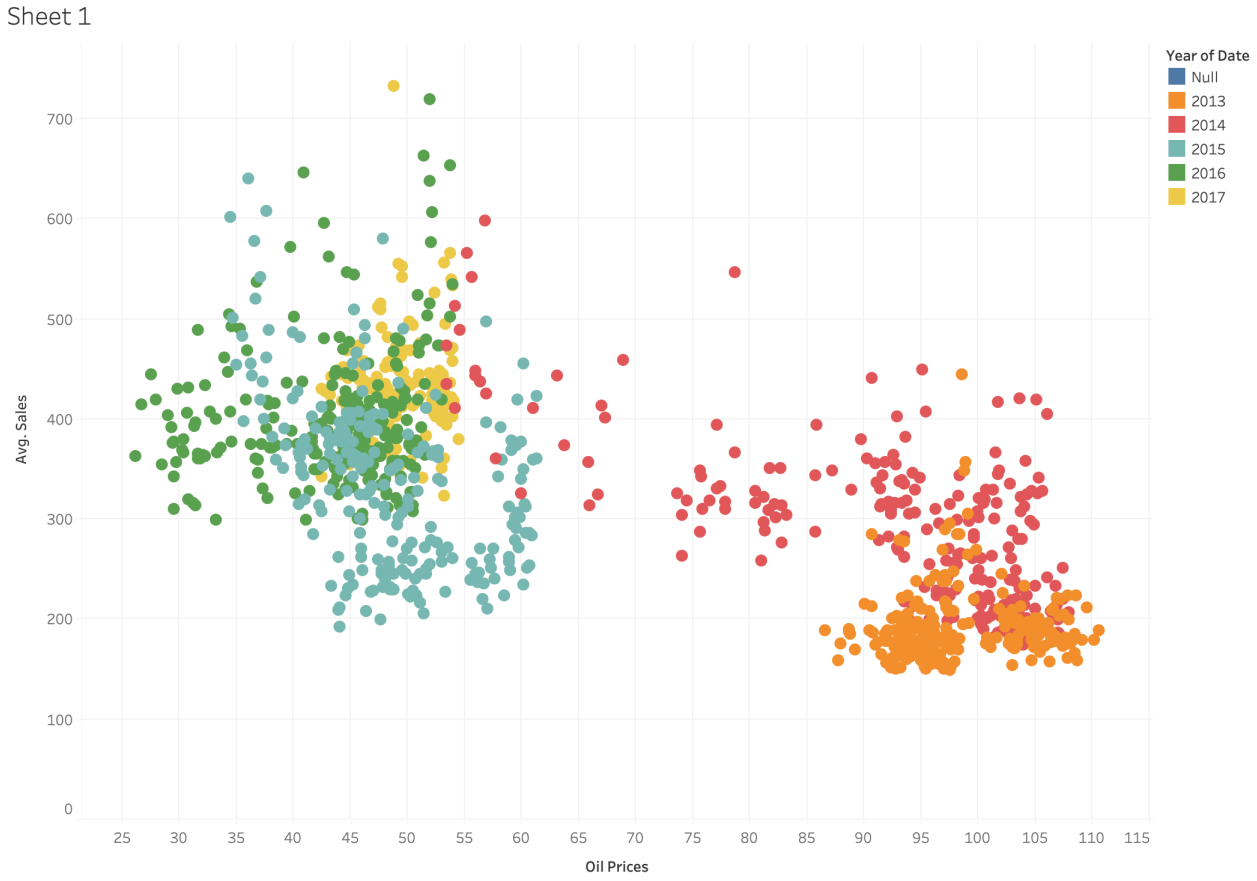
3.1 Oil Prices vs Date

First we can look at oil price to see if there is anything unusual in the data.



There appears to be a dip around 2014.

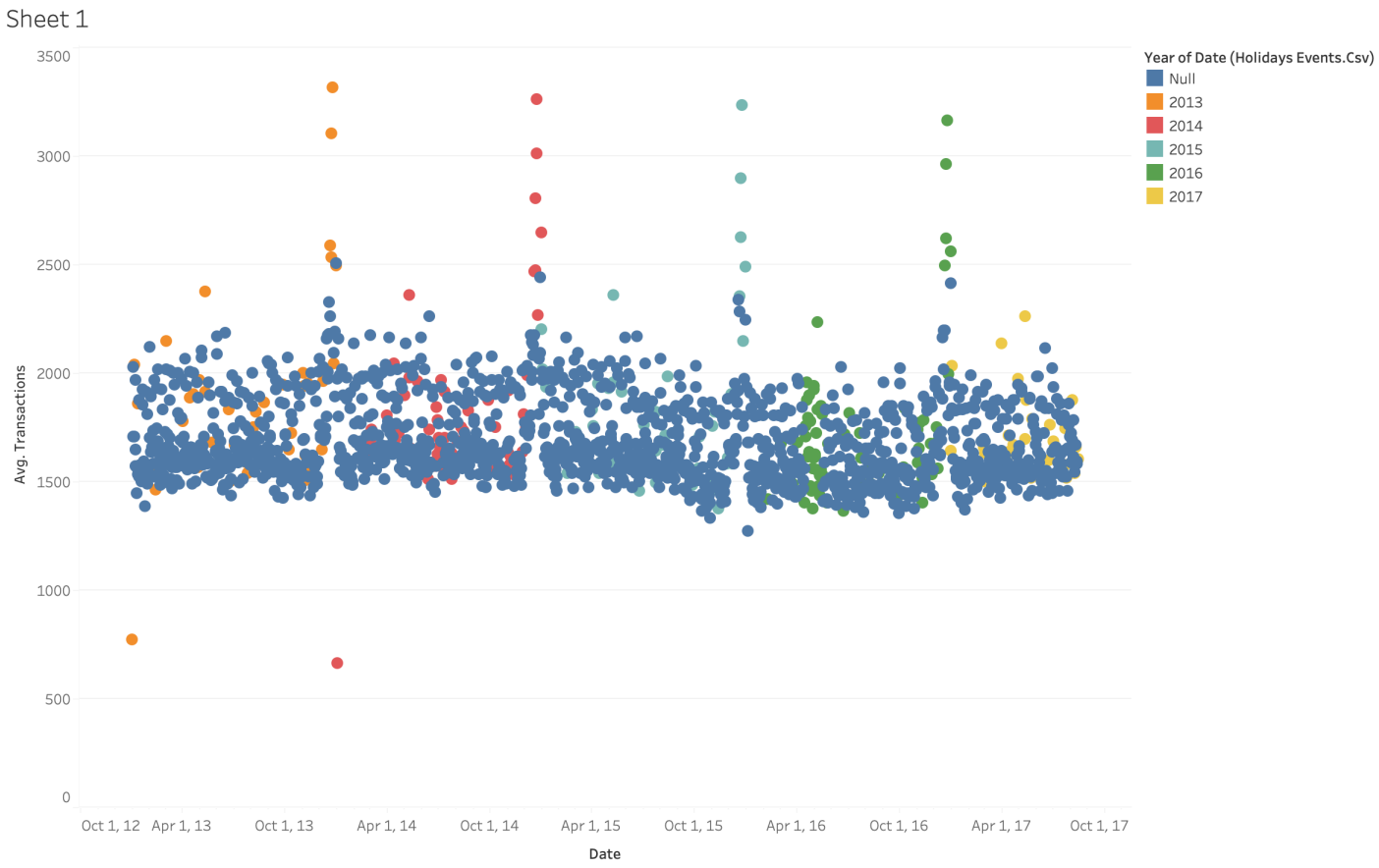
3.2 Average Sales vs Oil Prices



The plot of average of Sales for Dcoilwtico. Color shows details about Date Year.

It appears that Oil Prices have some impact on sales. After 2014, the average number of sales dramatically decreased after the sharp price drop in oil.

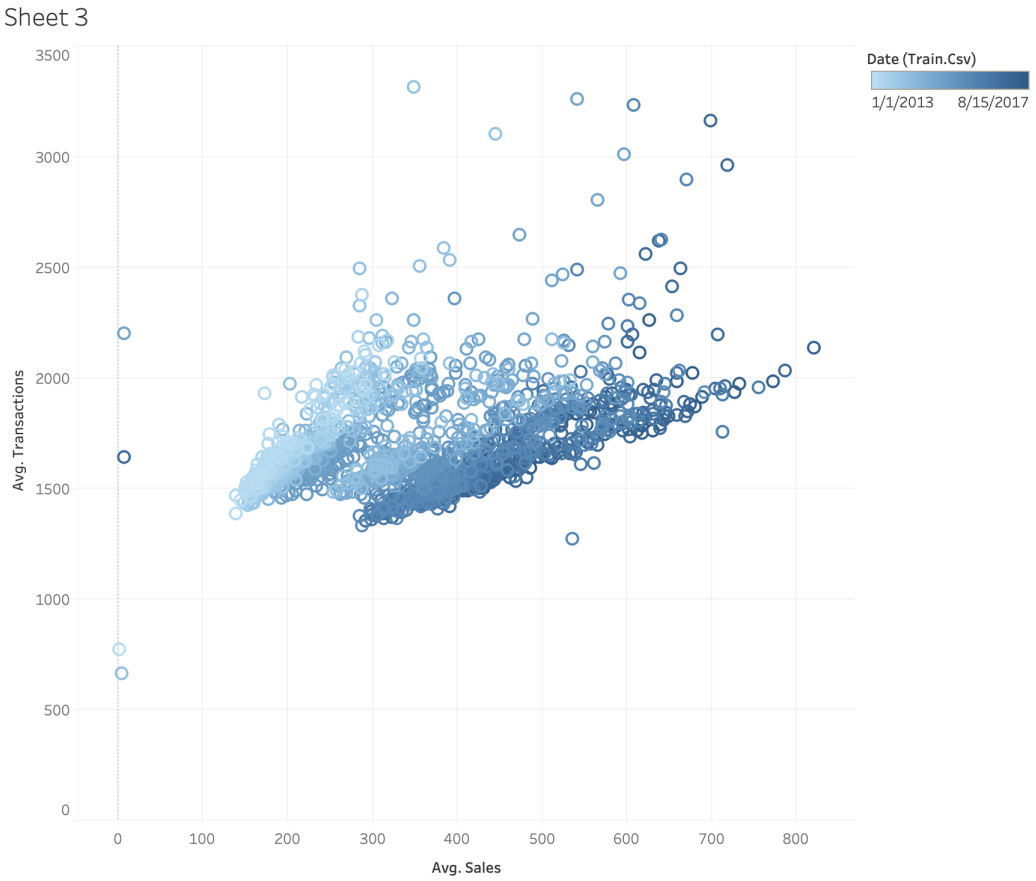
3.3 Average Transactions vs Date



The plot of average of Transactions for Date. Color shows details about Date (Holidays Events.Csv) Year.

This plot shows us seasonality in the data around Christmas and how holidays increase/decrease the average transactions among all the stores.

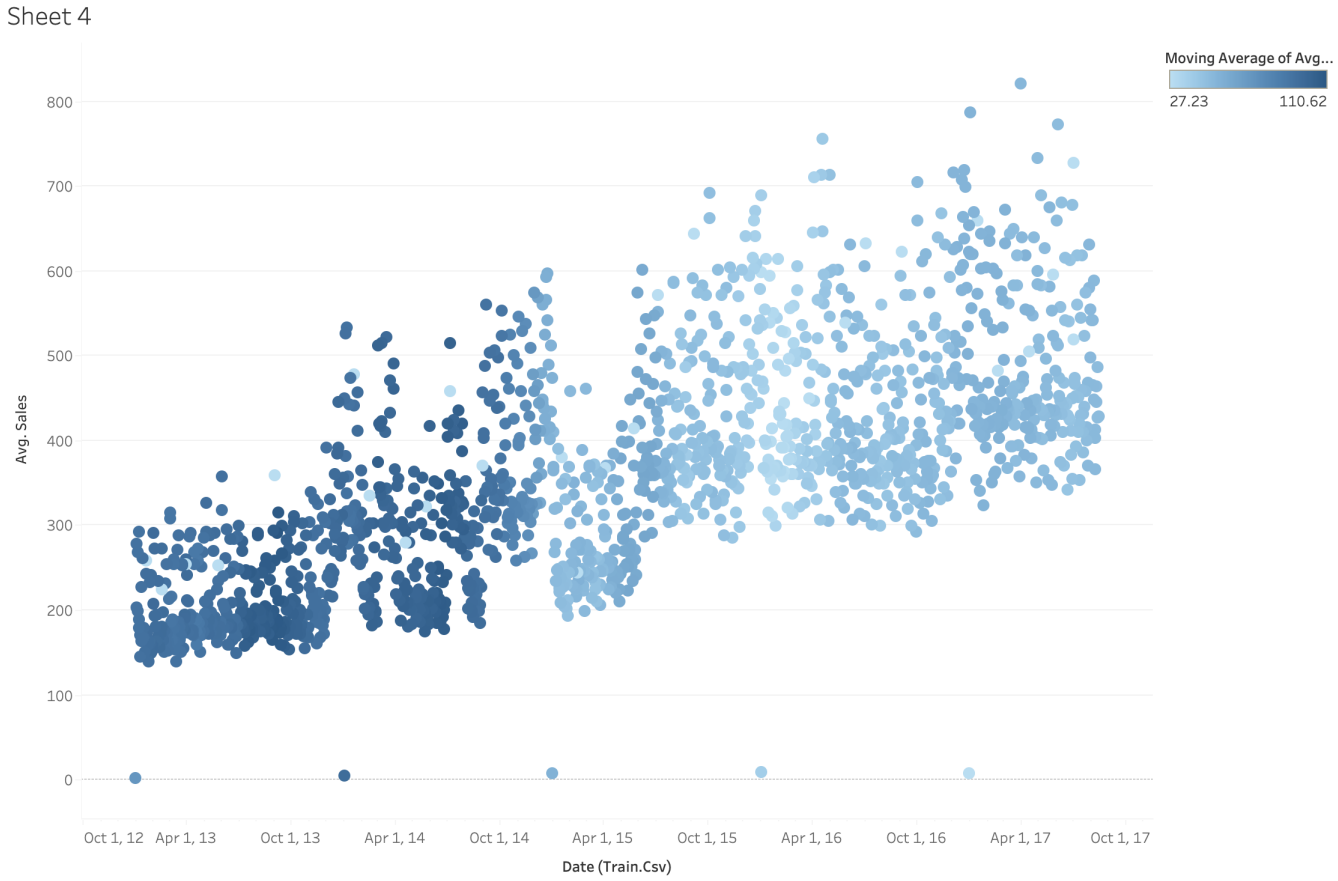
3.4 Average Number of Transactions vs Average Number of Sales



Average of Sales vs. average of Transactions. Color shows details about Date (Train.Csv).

It appears that after the oil price drop, we can see a shift in sales. The number of sales increased, whereas the number of transactions remain roughly the same.

3.5 Average Number of Sales vs Date



The plot of average of Sales for Date (Train.Csv). Color shows Moving Average of Avg. Dcoilwtico.

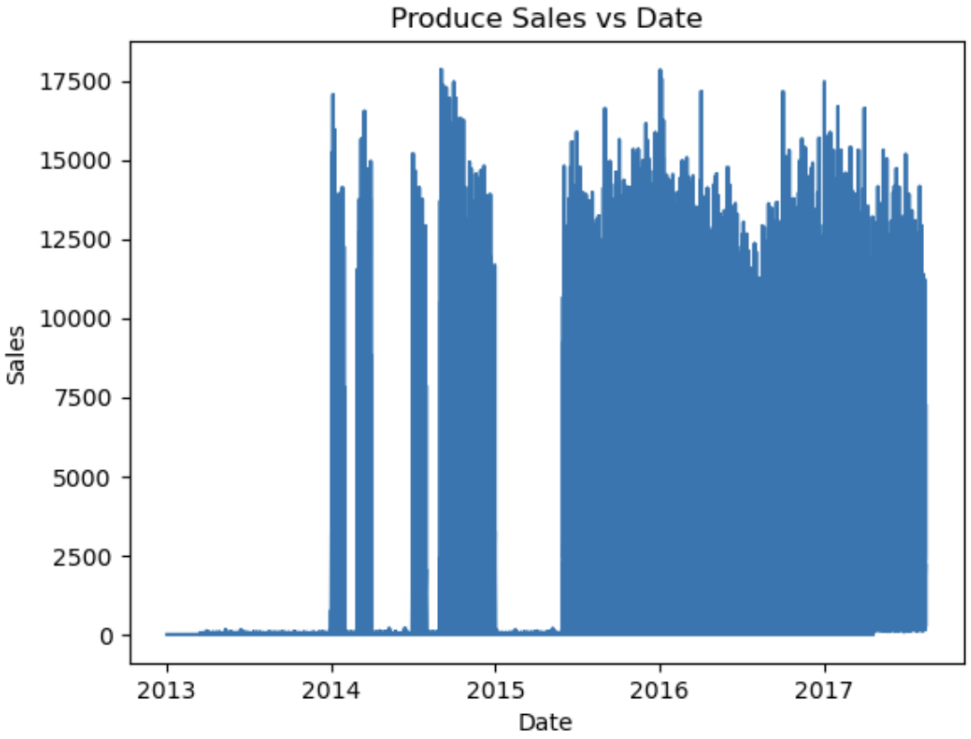
This plot was created using the date and average sales. Each point is colored using the moving average of the average Oil Price averaged from 1 value before and 1 value after to remove most of the NAs in Oil Price. We can observe a negative correlation between oil price and number of sales.

4 Machine Learning

In this task, we'll be predicting sales for the thousands of product families sold at Favorita stores located in Ecuador. Let's start off by cleaning the data.

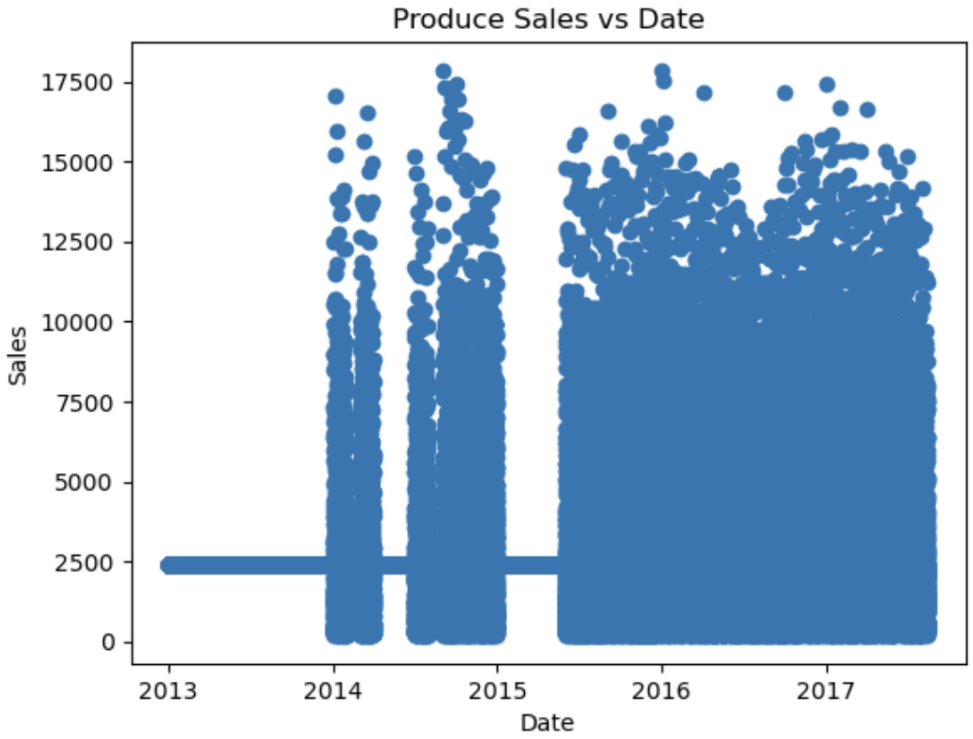
4.1 Data Cleaning

From Kaggle, We are given the train and test data which contains each day which is further split for each store. The data is even further split by product type. So, we end up with a huge dataset of over 3 million rows. The goal for this project is to predict sales for each product on each day. First, after observing the data, we'll clean the data on the produce sales category since it had some missing data.



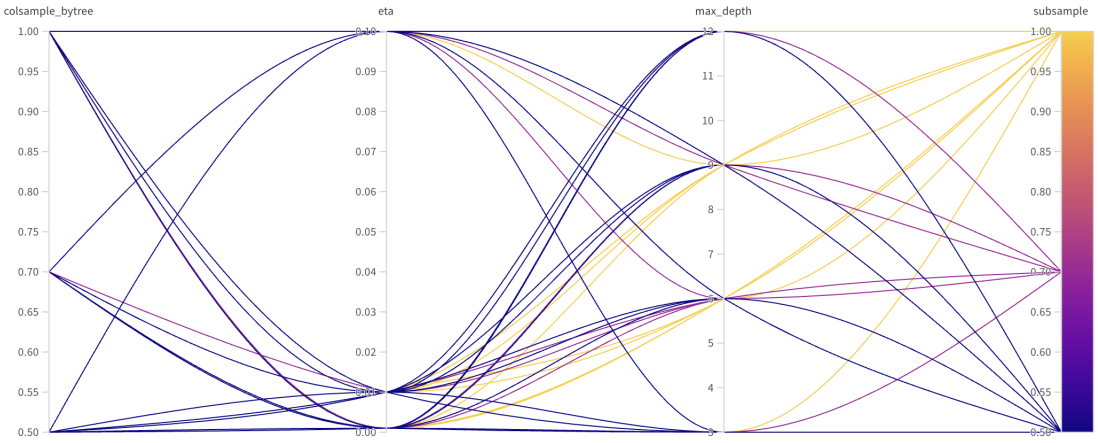
Since the data appeared mostly random, I took the mean produce values from the rows where it appears to have accurate values which was a threshold of 200 to fill in the gaps in the data. Mean produce values are the produce values of each day and taking the mean from all the stores. Here are the

results below from python. I also changed it from lines to points to better visualize the data.

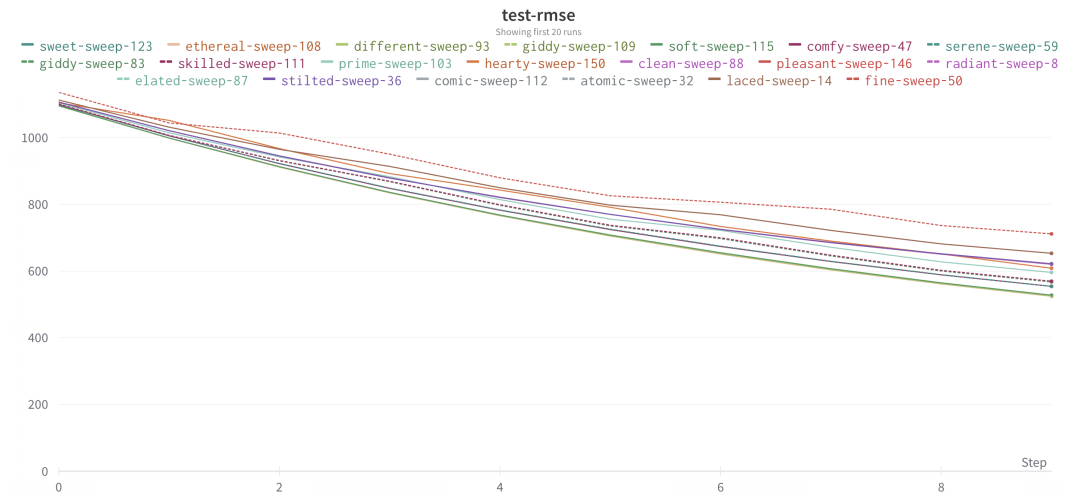


4.2 XGBoost

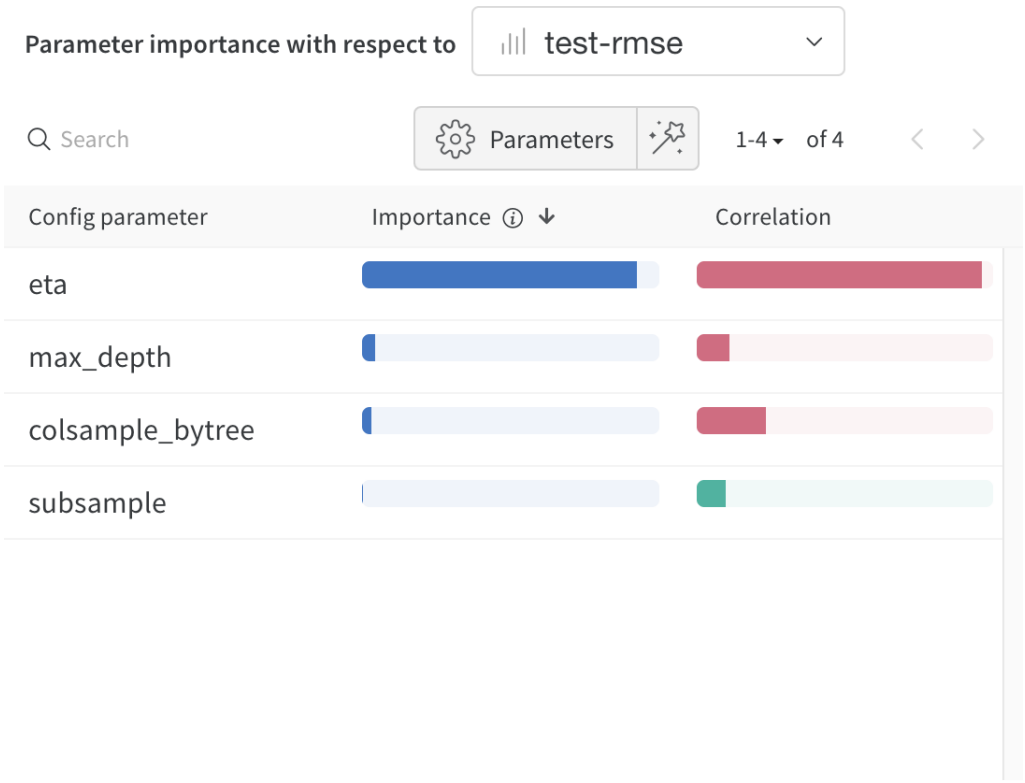
This model is XGBoost also known as Extreme Gradient Boosting is a supervised learning algorithm used for either regression or classification on large datasets. It builds decision trees that also avoids overfitting with other optimization methods. Initially, I trained the model and obtained a score of 1.39. However, the scores weren't good enough.
Tuning the hyperparameters, I used a program called Weights and Biases. It trained randomly over 100 different models with varying learning rate/eta, max depth, sub sample, and col sample by tree. Here's a visual of the different parameters tested.



Below, I'll show the test-rmse for the first 20 runs.



Here's the parameter importance from Weights and Biases below.



The importance measure comes from using the hyperparameters as x values and training a random forest model using the test-rmse as the y variable. There is a correlation that increasing the eta value decreases the test-rmse. Also, it shows that increasing the max depth decreases the test-rmse.

Observing the best model, it has a learning rate of 0.1, max depth of 12, sub sample of 1, and col sample by tree of 1 too. From past experiences, increasing the max depth and decreasing the learning rate tends to improve results, so in my final model I used learning rate of 0.05, max depth of 30, sub sample of 1, and col sample by tree of 1. However, using a max depth of 30 would likely overfit the model. Luckily, scoring the predictions on Kaggle improved the performance to achieve a score of 0.52.

5 Limations

One limitation in the model was not using a method from Kaggle's time-series course where they used the lags as additional features, however, with already 153 columns adding columns from the lags would increase the complexity a bit too much for my computer. Also, I used a grid search for hyperparameter tuning when there are better methods that can further improve my model. Next, I didn't use the oil data in my test or train datasets because I would have to look up the data online which would violate Kaggles rules in the competition, however, it could've slightly improved the results and this restriction typically isn't an issue in practice. Also, another limitation was filling in the produce valeus. I assumed that each store would sell around the same amount of produce for each store which isn't true. I probably should've taken the mean produce for each store and filled each store in separately rather than using the combined stores mean value.

References

[1] Kennytanner. "Store_sales_forecasting_extensive_eda." *Kaggle*, 2023, www.kaggle.com/code/kennytanner/store-sales-forecasting-extensive-eda

[2] "Learn Time Series Tutorials." *Kaggle*, www.kaggle.com/learn/time-series

[3] "Store Sales - Time Series Forecasting." *Kaggle*, www.kaggle.com/competitions/store-sales-time-series-forecasting/overview

[4] "Store Sales - Time Series Forecasting." *Kaggle*, www.kaggle.com/competitions/store-sales-time-series-forecasting/discussion/298626

[5] "XGBoost." *Weights and Biases Documentation*, 2021, docs.wandb.ai/guides/integrations/xgboost