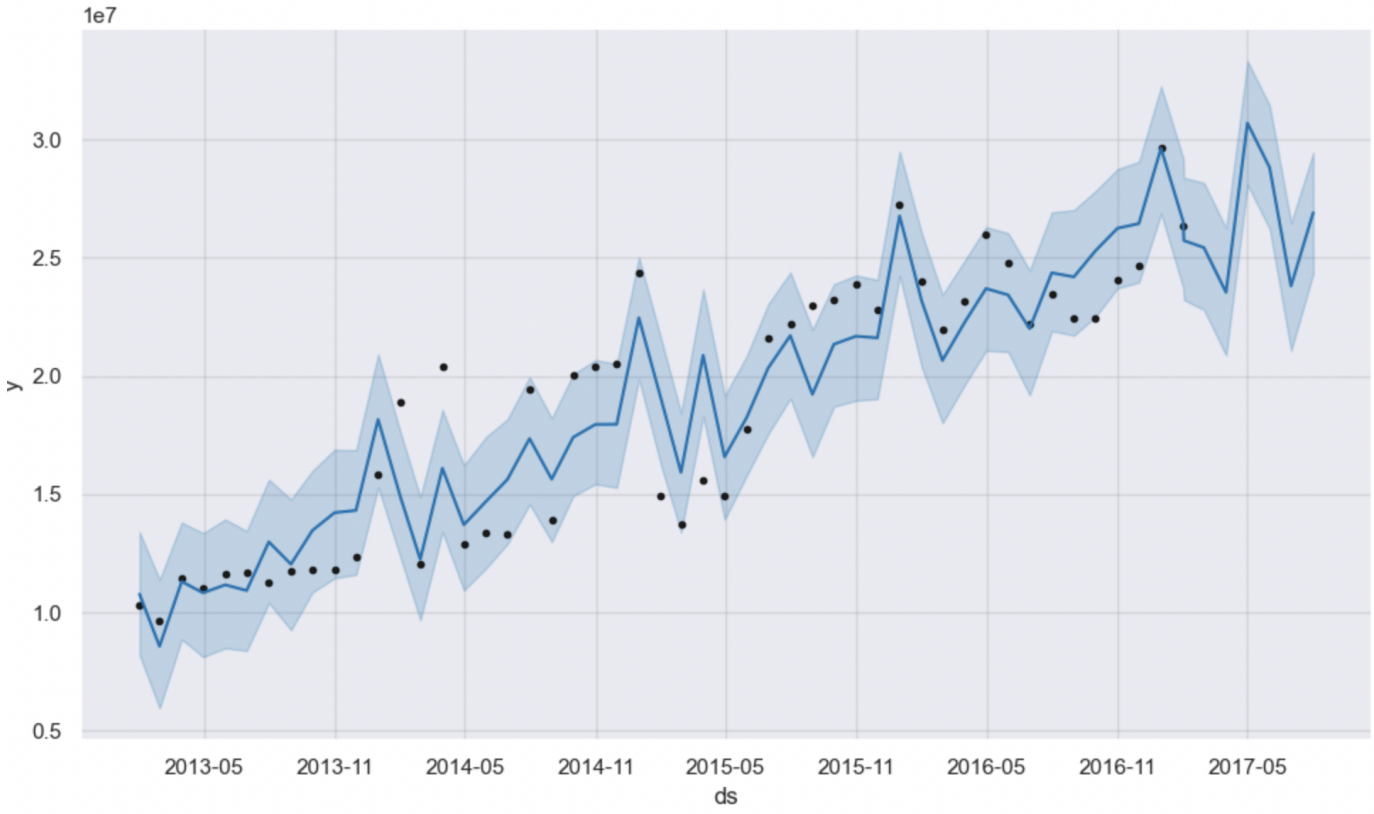
Favorita Sales Forecast: Final Report



By: Hamza Al Bustanji

Mentor: Dipanjan Sarkar

April/2023

Introduction

Goal:

We aimed in this project to develop a forecast of monthly sales from the aggregated sales of different Favorita stores. If accomplished, this would serve as a tool for predicting sales in the future for the corporation.

Data Source:

The data were provided by the giant Ecuadorian retailer, Favorita, through a Kaggle competition. The data is recorded as daily sales by each branch of Favorita, with the sales being further divided by category. The number of rows is 3000888, with 4 columns. Available as well are a file containing oil prices during the same time period, and a record of Ecuadorian holidays.

Audience:

The produced forecast will serve to assist executive-level employees in assessing the expected sales for any specific time period in the future. The forecast will be made based on monthly sales which are aggregated across different stores and different product types.

Data

The training data includes dates, store and product information, whether that item was being promoted, as well as the sales numbers. Additional files include supplementary information that may be useful in building your models.

File Descriptions and Data Field Information

train.csv

- The training data, comprising time series of features **store_nbr**, **family**, and **onpromotion** as well as the target **sales**.
- **store_nbr** identifies the store at which the products are sold.
- **family** identifies the type of product sold.
- **sales** gives 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).
- **onpromotion** gives the total number of items in a product family that were being promoted at a store at a given date.

test.csv

- The test data, having the same features as the training data. You will predict the target **sales** for the dates in this file.
- The dates in the test data are for the 15 days after the last date in the training data.

sample_submission.csv

- A sample submission file in the correct format.

stores.csv

- Store metadata, including **city**, **state**, **type**, and **cluster**.

- **cluster** is a grouping of similar stores.

oil.csv

- Daily oil price. Includes values during both the train and test data timeframes. (Ecuador is an oil-dependent country and it's economical health is highly vulnerable to shocks in oil prices.)

holidays_events.csv

- Holidays and Events, with metadata
- NOTE: Pay special attention to the **transferred** column. 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. To find the day that it was actually celebrated, look for the corresponding row where type is Transfer. For example, the holiday Independencia de Guayaquil was transferred from 2012-10-09 to 2012-10-12, which means it was celebrated on 2012-10-12. Days that are type Bridge are extra days that are added to a holiday (e.g., to extend the break across a long weekend). These are frequently made up by the type Work Day which is a day not normally scheduled for work (e.g., Saturday) that is meant to payback the Bridge.
- Additional holidays are days added a regular calendar holiday, for example, as typically happens around Christmas (making Christmas Eve a holiday).

Additional Notes

- Wages in the public sector are paid every two weeks on the 15 th and on the last day of the month. Supermarket sales could be affected by this.
- A magnitude 7.8 earthquake struck Ecuador on April 16, 2016. People rallied in relief efforts donating water and other first need products which greatly affected supermarket sales for several weeks after the earthquake.

Method

The data is of daily sales divided by category and further divided by branch. To provide a forecast of sales at the level of the entire company we will need to aggregate sales across branches and across products. Further, we will aggregate sales by month instead of by day since the purpose of the forecast is to provide a sense of the sales that are expected.

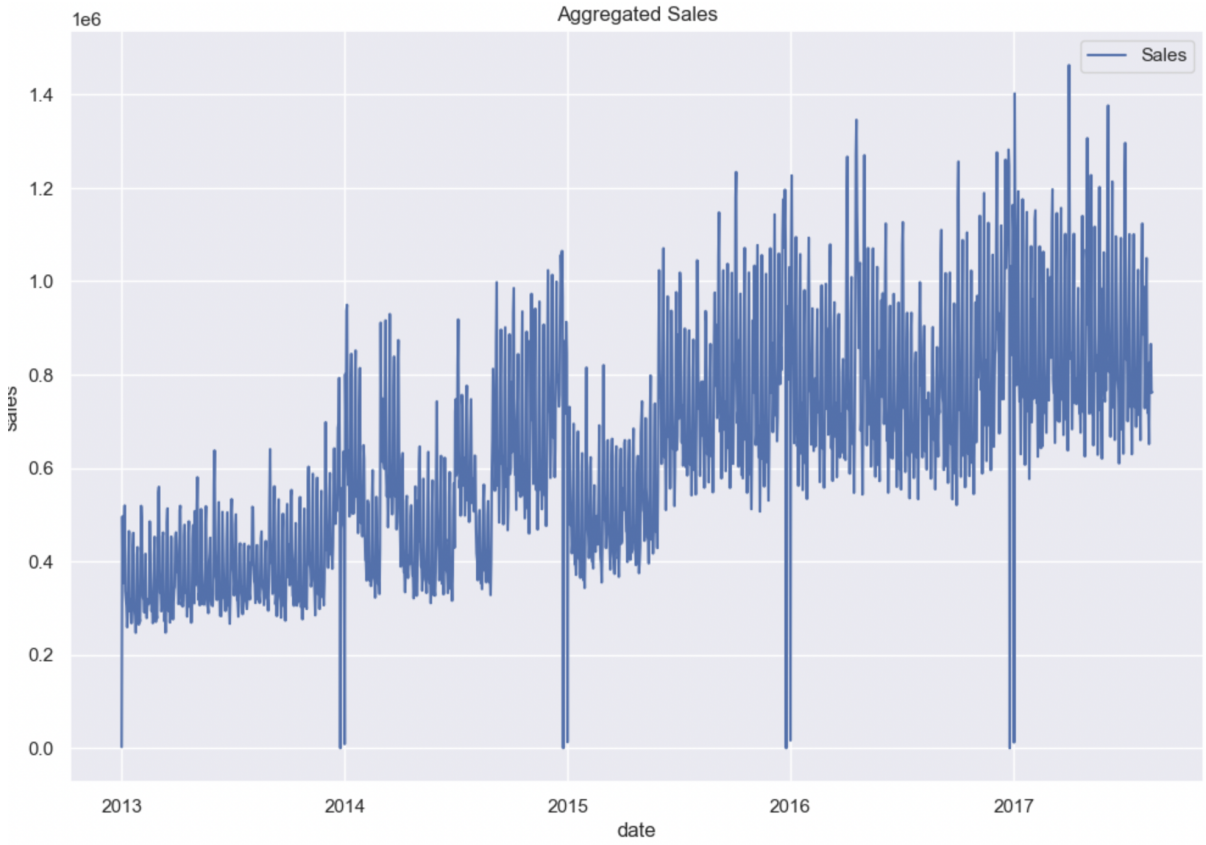
We will start by wrangling the data and cleaning it so that it is in a workable shape for our purposes. Then we will perform an exploratory data analysis to gain insights into our data and see if there are any interesting facts or relationships that we can explore. Lastly, we will pre-process the data for modeling and produce the predictive models for evaluation and final selection.

For modeling, we will consider ARIMA, SARIMA, and Meta's Prophet. We could also use neural networks, but since they generally tend to perform worse than the ARIMA-based models we will skip it.

Data Wrangling

Below are the tasks that were carried out in the Data Wrangling part of project:

- Investigated missing values: missing values were investigated, and none were found in the training data. Though, the oil prices data had some missing values that we could have filled by extrapolating, but since we weren't going to use that data in forecasting, we left it as is.
- We looked into the number of stores that were in the data. 54 stores across different regions in total.
- We aggregated sales across all stores, and plotted the resulting time-series:



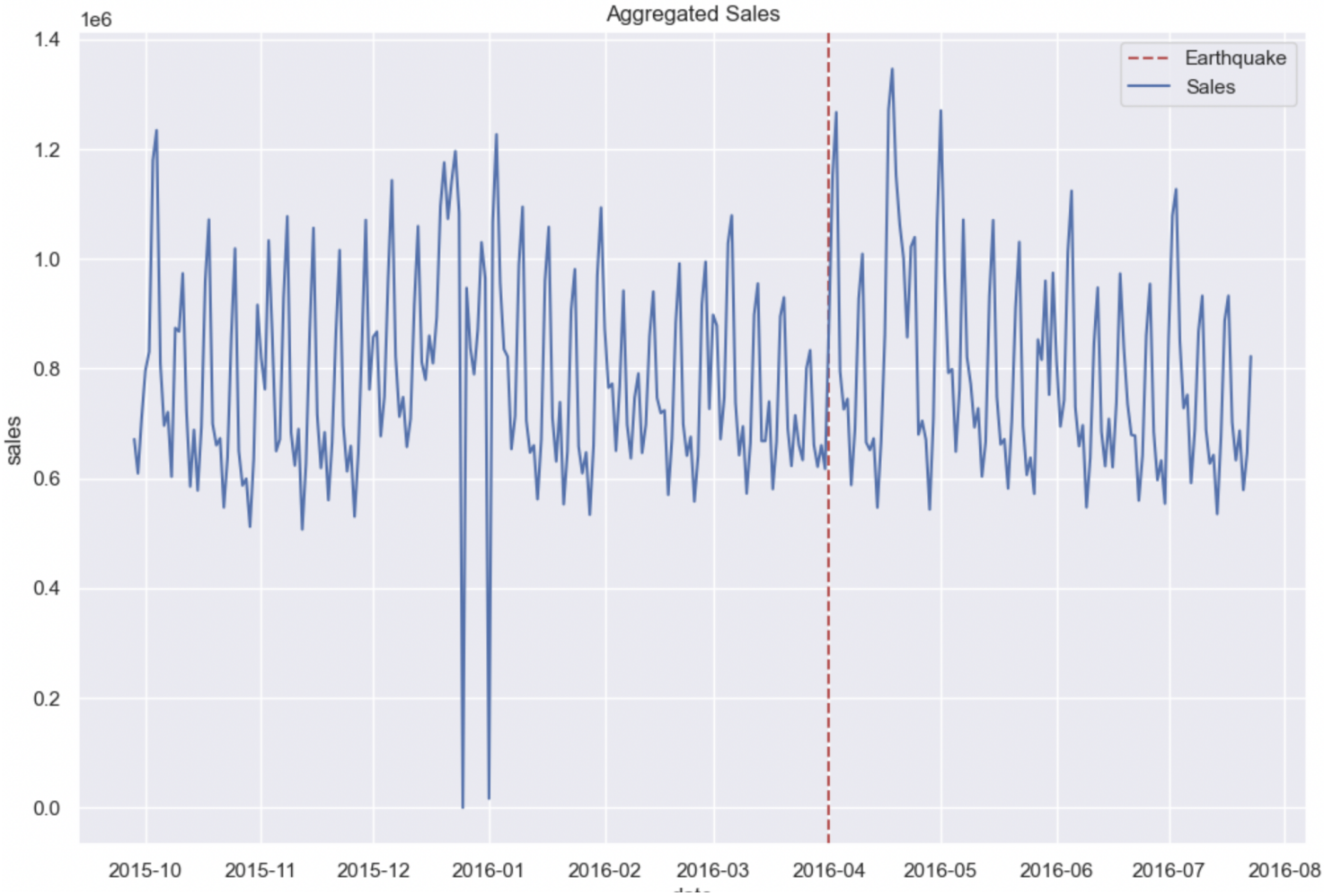
- We created a Data Frame of the aggregated sales and saved it as a csv file for later use.
- We created a Data Frame of total sales by store and saved it as well.
- We created a monthly aggregated Data Frame of sales which will be the one we will use in modeling. We also created one of quarterly sales for EDA.

- We downloaded inflation data from the World Bank's website to use for analysis by adjusting sales for inflation. We cleaned the data and wrangled it into our desired shape and saved it for later use.

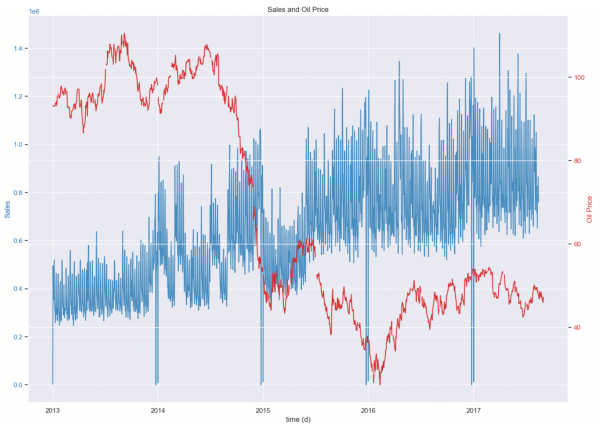
Exploratory Data Analysis

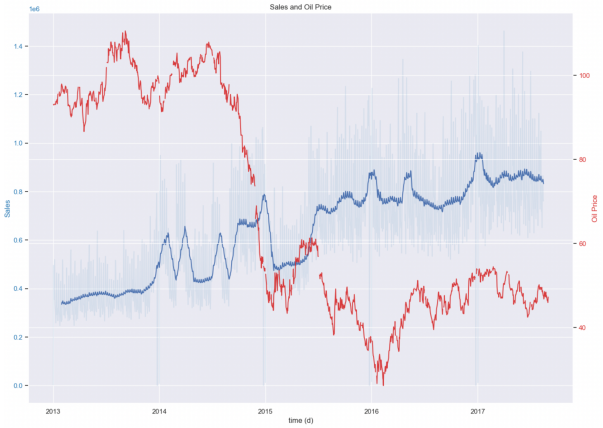
In exploratory data analysis we accomplished the following:

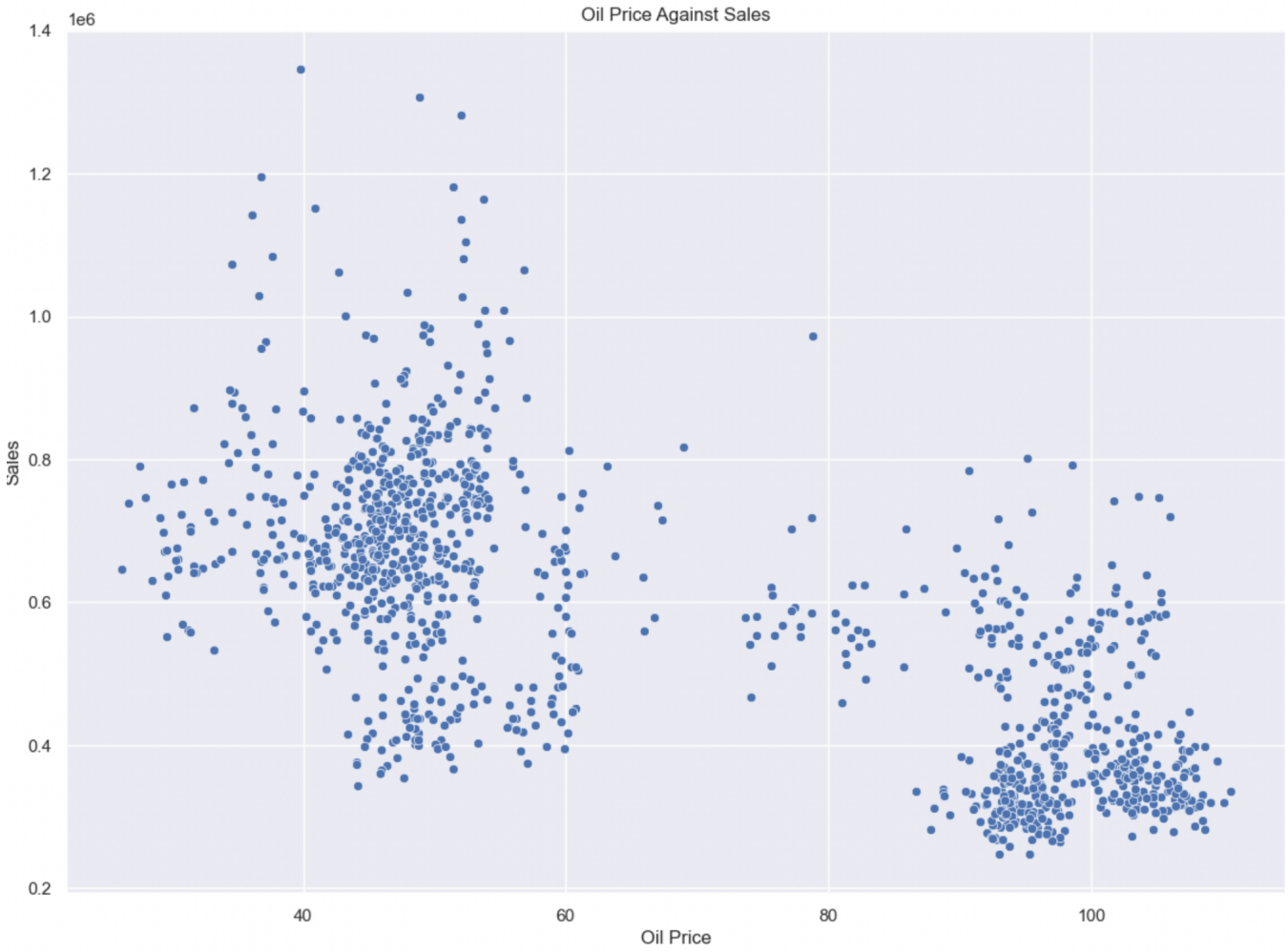
- Explored the effects of an earthquake that hit Ecuador in 2016 on aggregated sales, which weren't apparent:



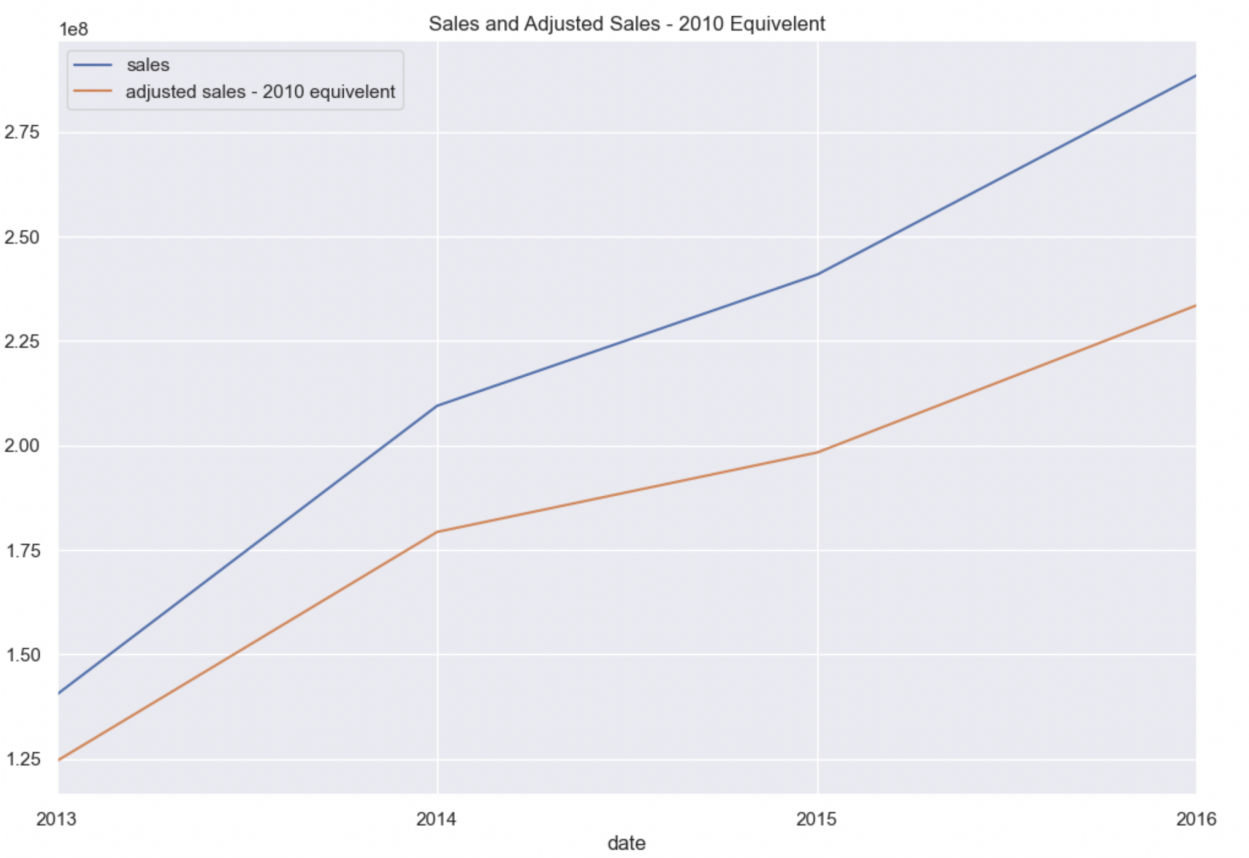
- We explored the relationship between oil (gas) prices and sales, and then calculated the correlation between the two variables:



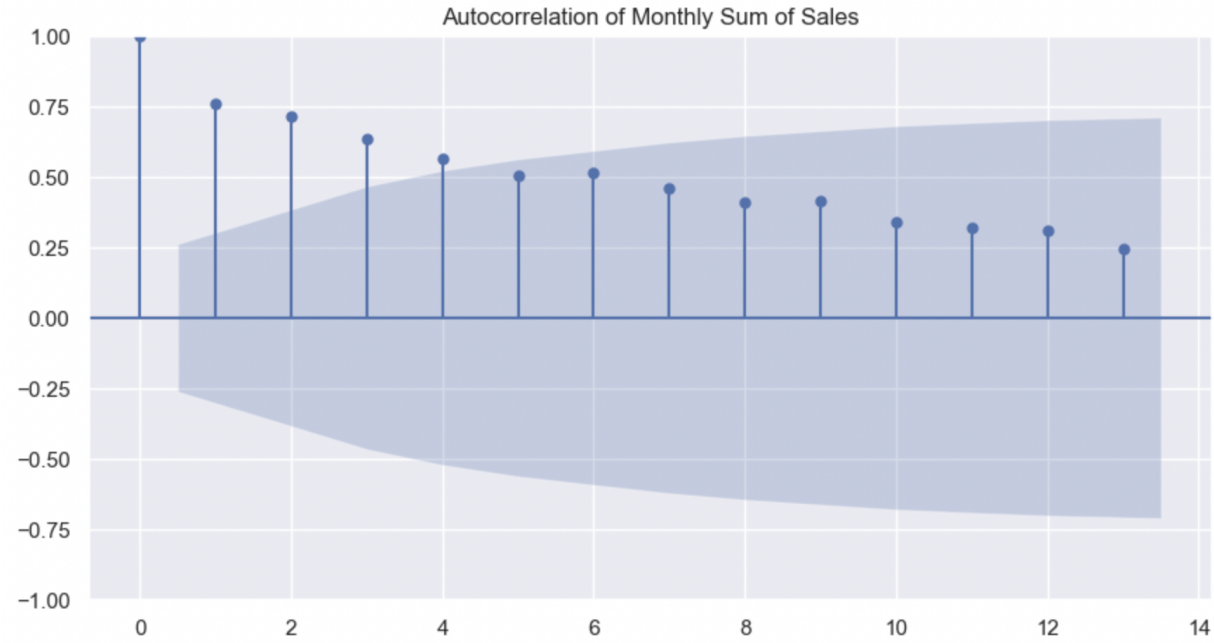




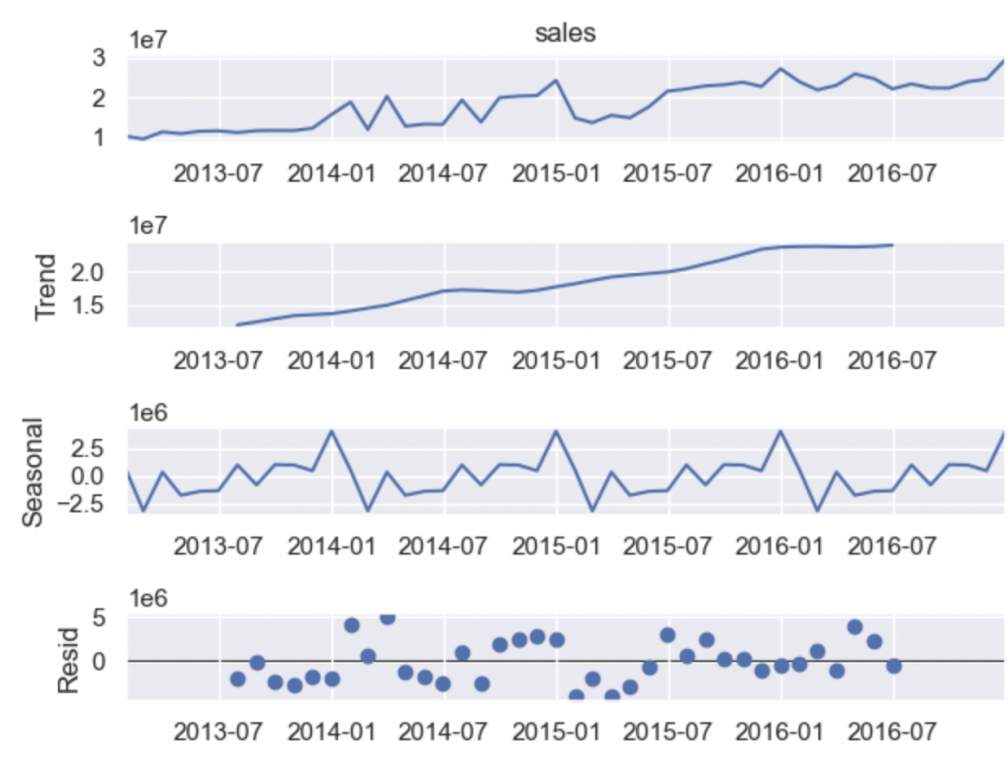
- We calculated the inflation-adjusted annual sales and compared them to the regular sales, the rates of increase seem similar, except for the 2013 - 2014 year:



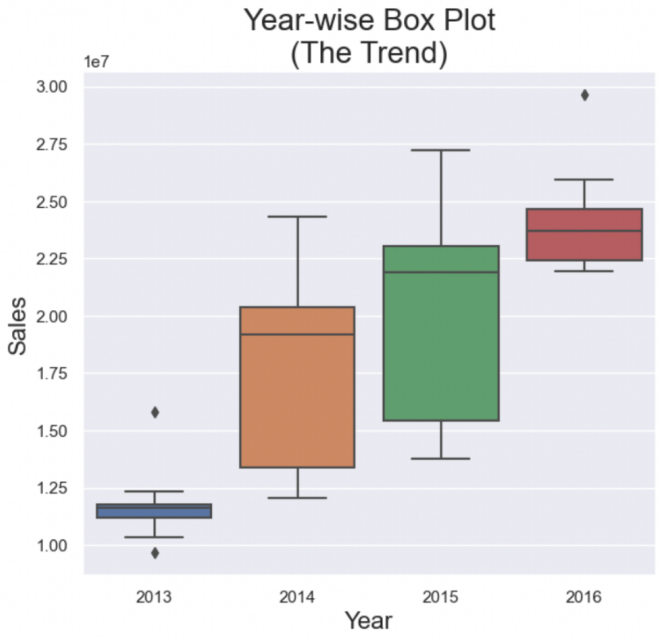
- We calculated ACF:

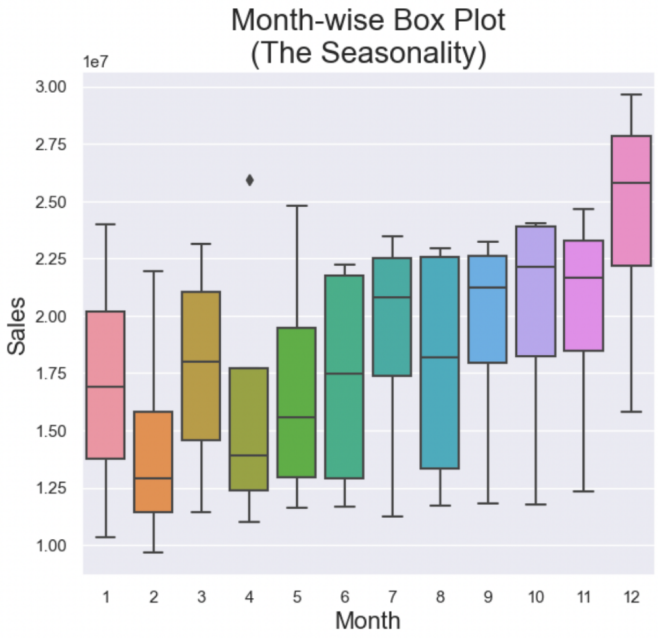


- We performed a trend, seasonal, and residual decomposition:



- We plotted the seasonality and the trend using box plots:





Pre-processing and Modeling

In this step, we did the following:

- Created a train/test split, and used the last available 6 months as a testing subset.
- Performed grid searches for both the ARIMA and SARIMA models.
- Created both an ARIMA and SARIMA model.
- Plotted the results.
- And used the root mean squared error to evaluate each.
- We pre-processed the data to be able to use it in a Prophet model.
- We created a Prophet model.
- We plotted the results and evaluated the model using the root mean squared error as well.

Best performing model

Both the ARIMA and SARIMA models performed the same, which is because the SARIMA model didn't pick up on any seasonality. The Prophet model performed slightly worse than the other two, although it managed to pick up on some of the fluctuations in the data. We chose the ARIMA model as the best-fitting model.

Next steps

At this point, our project came to its conclusion. But, a few recommendations need to be considered for possible revisions: an analysis of the residuals. The distribution of the residuals component is an essential pre-requisite to forecasting, and we should add a comprehensive analysis of it in the future. Next, including confidence intervals with the ARIMA model is essential to be able to convey the uncertainty of our forecasting.