# *Store Sales – Time Series Forecasting*


***RetaiLabs Inc.***

Brian Avila, Zoe Coughlan, Noah Welsh

CS-301 Spring Semester 2022

Ying Wu College of Computing, New Jersey Institute of Technology



**Abstract.** As grocery retailers like Walmart, Costco, Target etc., embark on the adding of new locations with unique needs, new products, ever-transitioning seasonal tastes, and unpredictable product marketing alongside their pre-established locations, they are met with the predictability issue of potentially understocking popular items or overstocking perishable goods. RetaiLabs Inc. pushes to combat the potential issue for retailers through usage of Linear Regression Models in connection with Time Series Forecasting. Through the process of analyzing the data provided by a retailer using statistics and modeling to make predictions and informed strategic decisions alongside proper data manipulation through our models of choice (Linear Regression and Gradient Descent) we strive to ensure retailers please customers by having just enough of the right products at the right time as it further results in decreased food waste and improvements towards customer satisfaction by providing proper forecasting. Through the application of our models of choice with inclusion of proper JAX implementations, using data sets provided by Corporacion Favorita, a large Ecuadorian-based grocery retailer we were able to provide them a display between the relations in sales and decreased oil prices/sales, alongside further developing models for the displaying of correlation between increase in overall sales within the same time frame of oil price drops for the Ecuadorian retailer.


## 1 Introduction

The problem for the chosen Kaggle Competition was centered around the usage of Machine Learning to predict grocery sales through an implementation of forecasting, specifically the forecasting of store sales for the large Ecuadorian-based retailer known as Corporacion Favorita. The problem required the development of a model which accurately predicts the unit sales for thousands of items sold at different Favorita store locations. The importance of this type of sales forecasting comes forth in presenting how it develops a helpful tool which helps in overall business planning, budgeting, and risk management.

Alongside helping businesses to estimate their costs and revenue accurately based on which they are able to predict their short-term and long-term performance. In connection to grocery retailers, more accurate forecasting can decrease food waste related to overstocking and improve customer satisfaction as well as withhold the potential of ensuring many local stores have exactly what buyers need the next time they shop. The results obtained through the completed data manipulation and modeling allowed our company the ability to empirically show the relations in the data such as when a decrease in oil occurs, an increase in overall sales occurs in the same time window.

## 2 Related Works

An article published by Matthew Schneider, an Assistant Professor in the Integrated Marketing Communications Department at the Medill School of Journalism at Northwestern University alongside Sachin Gupta, Henrietta Johnson Louis Professor of Management and Professor of Marketing at the Johnson Graduate School of Management, Cornell University came forth in presenting a varying perspective towards the completion of forecasting. Within their article they state "We consider the problem of predicting sales of new and existing products using both numeric and textual data contained in consumer reviews. Many extant approaches require considerable manual pre-processing of the textual data, making the methods prohibitively expensive to implement and difficult to scale. By contrast, our approach using a bag-of-words method requires minimal pre-processing and parsing, making it efficient and scalable." Which came forth as displaying an enhanced qualitative approach towards forecasting in comparison to the established approach by us combing both quantitative and qualitative data through retailer provider information to predict proper forecasting instead of the usage of customer reviews. They established "Results show that in both tasks the predictive performance of the proposed approach is strong and significantly better than that of models that ignore the textual content of consumer reviews, and a support vector regression machine with the textual content. Further, the approach is easily repeatable across product categories, and readily scalable to much larger datasets." To briefly compare to the obtained results reached through our approach we were able to view how quantitative data variables withhold a stronger affect for the predictability of store sales.

## 3 Data

The data used for the completion of the competition was provided directly from the Kaggle source, which included training data that provided dates, store and product information, whether that item was being promoted, as well as specific sales numbers. This data came forth to being spread through a total of five csv files all which presented both quantitative and qualitative data. Training.csv provided necessary comprising time series of features such as store number, family, and onpromotion as well as the target sales. Store number identifies the store at which the products are sold, family identifies the type of

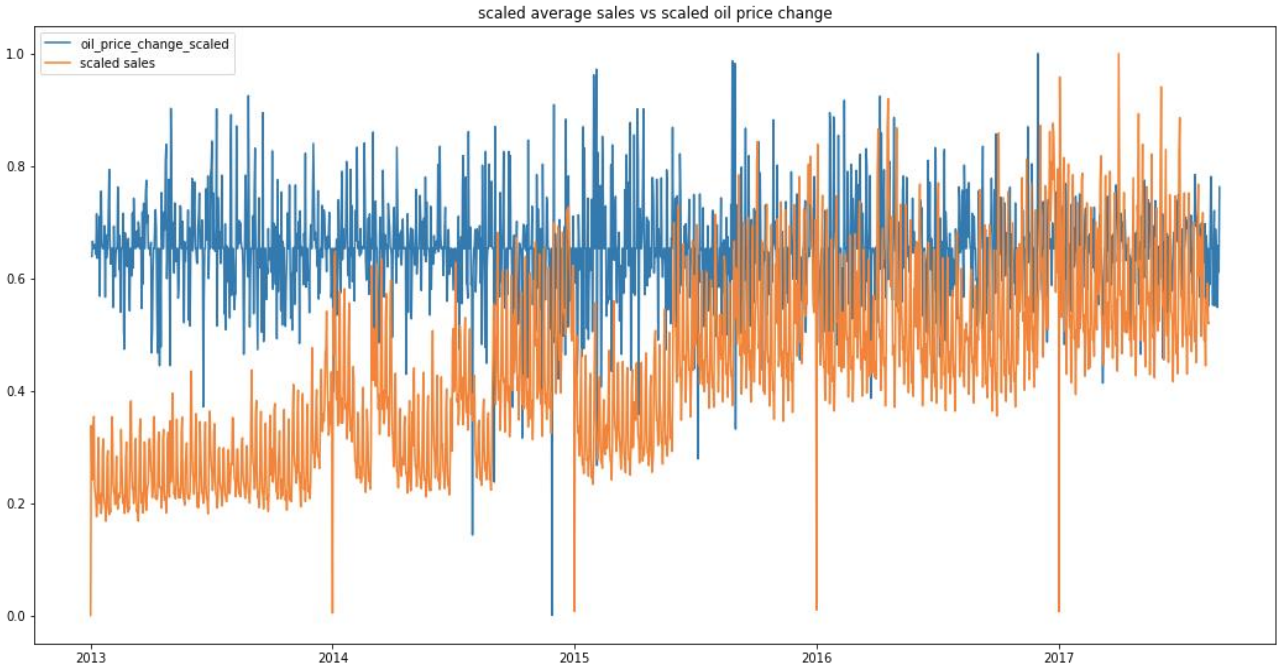
product sold, sales gave the total sales for a product family at a particular store at a given date, and onpromotion gave the total number of items in a product family that were being promoted at a store at a given date. Stores.csv added by providing store metadata, including city, state, type, and cluster, where cluster was a grouping of similar stores. Oil.csv was additional data of daily oil price with inclusion of values during both the train and test data timeframes. While holidays_events.csv provide metadata on holidays and events whether on a national or local level as well as location of the corresponding holiday or event, and transactions.csv provided us with dates and transaction quantity for specific store numbers. As the goal of predicting and further presenting a proper forecast was the focus, the data chosen to be specifically focused on as it presented itself as providing us with the best chance of a proper prediction was the established training and test data sets. These two datasets we worked on were those which efficiently allowed us to view correlation between average sales and outside factors such as holidays, events, and oil price while further extending our capabilities to view the features which provided the most insight towards completing our forecasting models.

## 4 Methods

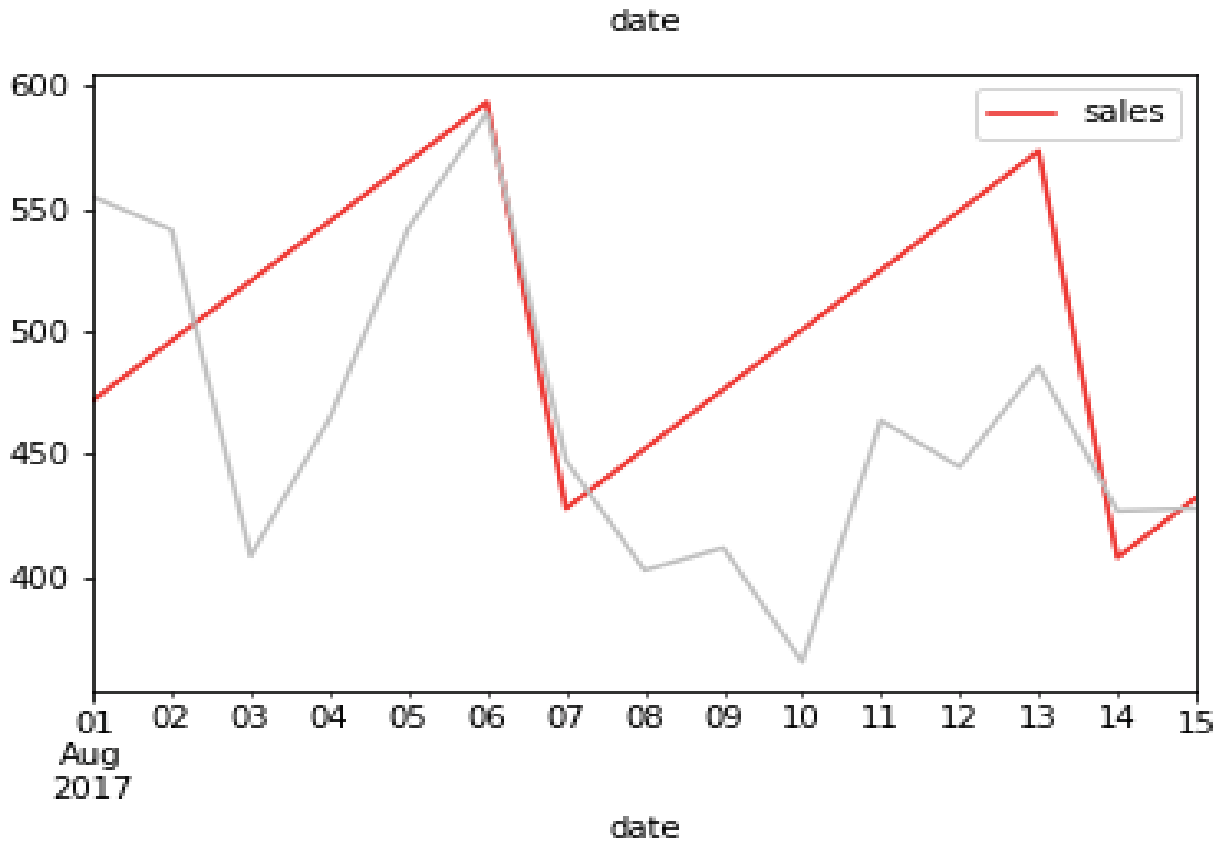
Methodologically, our approach is based on correlation between sales and the economy of the area. We are able to gauge the economy based on oil prices, as this study is based on an Ecuadorian retailer, and the Ecuadorian economy is reliant on oil. Due to that reliance, lower oil costs correlate to a healthier economy. This allows oil prices and economic state to serve as a predictor for sales. Initially, we were not completely positive about what the best approach could be. After successfully manipulating the data, we considered a multitude of approaches considering what the best data to compare would be in a way that could be used to draw definitive conclusions from. From previous assignments displaying the basic parts of the data was simple and something our company had ease doing but applying the best learning method was difficult and accomplished through trial and error. Finally, after much experimentation with the data and trying to draw results from those experiments, our company decided that using models of Linear Regression and Gradient Descent would be the most appropriate approach. We used oil prices as a feature to generate predicted values for sales via linear regression and gradient descent, as we learned to do throughout this semester through our lectures and homework assignments in this course.

## 5 Experiment

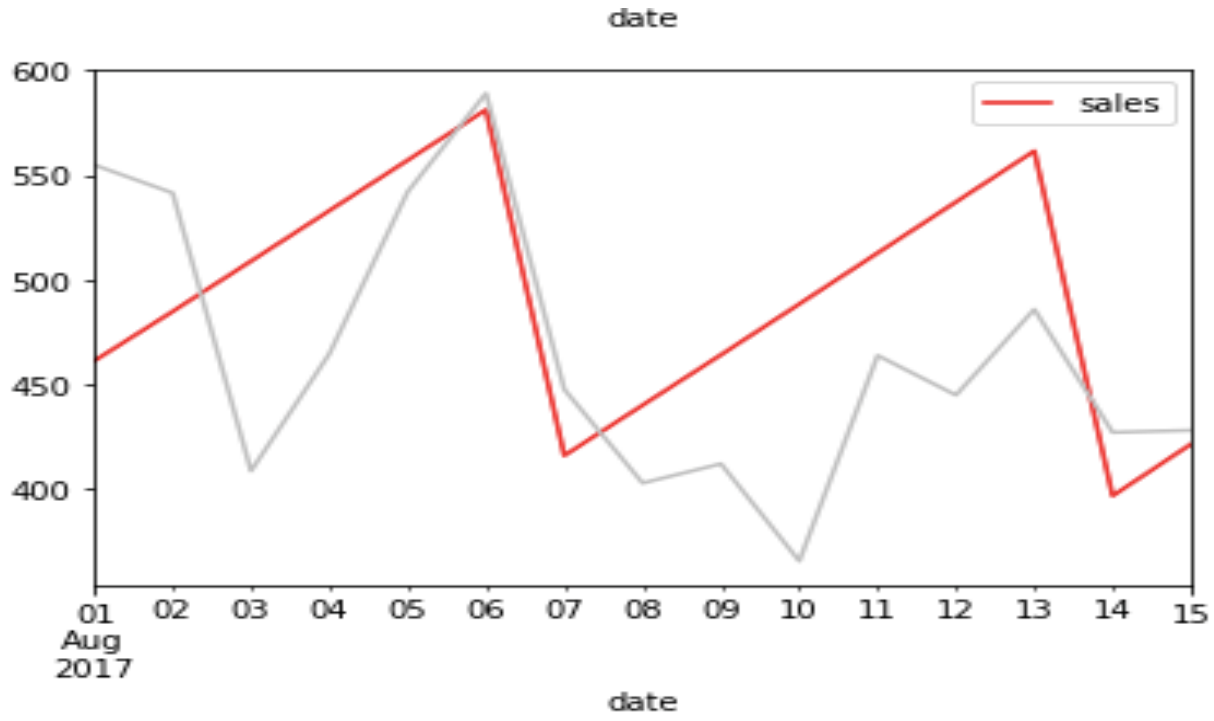
Our initial and most basic experiment involved direct comparison of oil prices and sales over time. This comparison involved scaling these sets of values to allow us to overlay them on the same graph using the same range of times so that they are directly comparable. By viewing the trends visually, we were able to obtain a baseline for what the relationship between oil prices and sales is.




As for linear regression and gradient descent, we applied JAX in tandem with Scikit-learn. We generated predicted and actual values based on two sets of features within the data. One set of features, as shown below, is solely dates and sales, without regard for oil prices.



The other takes oil prices into account to generate a prediction based on average sales over time.




The other takes oil prices into account to generate a prediction based on average sales over time.

## 6 Conclusion

Through the completion of manipulating data through models of Linear Regression and Gradient Descent, the obtained forecast for the Ecuadorian retailer of Corporacion Favorite allowed us as a company to become much further educated on the relationship shared between a variety of given grocery sales data such as past sales, holiday sales, oil sales, and others. Completed by manipulating the data and putting it through the models we developed, our company has been able to empirically show the relations in the data such as when a decrease in oil occurs, an increase in overall sales occurs in the same time window. This was accomplished mainly through the implementation of various libraries. Pandas and matplotlib were necessary to both manipulate the data so that it can be used, and to display the data graphically so that conclusions can then be drawn. Sci-kit learning software, numpy, and jax APIs were all essential and proved crucial in the development of our linear regression model. However, further development and future extensions withhold greater possibilities for the effectiveness towards proper sales forecasting. Due to such the usage of our multivariate model which relies on a variety of inputs, including past sales, holiday calendars, or even economic indicators, may be additionally replicated and analyzed through Neural Networks and Decision Tree-based methods. The usage of these models in comparison to Linear Regression, would further allow us to see direct comparison in pros and cons towards forecasting specifics and allow for their implementations to be retailer specific in deciding the specific multivariable

forecasting in which a company may need. However, forecasting isn't specifically tied to multivariate models and through the usage of auto-regression models and the capability of predicting future sales solely based on past sales values may company sales forecasting present varying methods dependent on that which is most important to a company's economy for predicting the proper forecast.

## Resources


Pathak, P. P. (2021, September 12). *Time Series forecasting - A complete guide*. Medium. Retrieved from https://medium.com/analytics-vidhya/time-series-forecasting-a-complete-guide-d963142da33f#:~:text=In%20statistical%20terms%2C%20time%20series,predictions%20and%20informed%20strategic%20decisions

Mahalingam, K. (2020, September 14). *Importance of sales forecasting & six factors to consider for accurate forecasting*. Chargebee Compass. Retrieved from https://www.chargebee.com/blog/importance-of-sales-forecasting/#:~:text=Sales%20forecasting%20allows%20companies%20to,and%20manage%20its%20cash%20flow.&text=Sales%20forecasting%20also%20helps%20businesses,term%20and%20long%2Dterm%20performance

Gupta, S., & Schneirder, M. (2015, August). *Forecasting Sales of New and Existing Products Using Consumer Reviews: A Random Projections Approach*. ResearchGate. Retrieved from https://www.researchgate.net/publication/281900054_Forecasting_Sales_of_New_and_Existing_Products_Using_Consumer_Reviews_A_Random_Projections_Approach

kalfon, eliott. (2020, July 20). *Using machine learning to predict sales and anticipate demand*. TrueCue. Retrieved from https://truecue.com/resources/blog/using-machine-learning-to-predict-sales-and-anticipate-demand/