Store Sales - Time series Forecasting

🗣️ **Which product family, and from which store, is projected to generate the highest sales over the next 30 days? Well, let's predict!!!**

By

Dudekula Abid Hussain

Email - dabidhussain2502@gmail.com

Introduction

This is a getting started competition from kaggle, where we need to predict store sales for Corporación Favorita, a large Ecuadorian-based grocery retailer. This time series forecasting competition deals with data that changes frequently over time. Our task is to build a model that can accurately predict unit sales for thousands of items across different Favorita store locations.

Before diving into the data, let's understand this through a real-life example. Imagine Corporación Favorita is like a Morrisons store near our homes where we shop daily. Just as we can find Morrisons stores scattered throughout the UK, let's say we buy eggs from our local store every day. One day, when we need a large quantity of eggs, we discover the store has run out—frustrating, right? What if the store could understand our needs and predict egg sales to stock accordingly?

This is exactly what we're doing here: building a model that helps Favorita stores predict unit sales of various products across their locations. This helps the company understand customer needs, manage inventory levels, reduce food waste, and provide better service by having the right products available at the right time.

Understanding the Data

Now that we understand the concept, let's examine the six data files provided: train data, test data, sample submission data, stores data, oil data, and holiday events data. Let's look at each file in detail.

Train data - This CSV file contains time series features with 5 columns: **'date'** indicates the date, **'store_nbr'** identifies the store where products are sold, **'family'** describes the type of product sold, **'sales'** shows the total sales of a particular product at a specific store on a given date (this is our target column), and **'onpromotion'** indicates the total number of items in a product family being promoted at a store on a given date.

Test data - This CSV file contains the same features as the train data, except for the 'sales' column, which we need to predict using our trained model. The test data covers the 15 days following the last date in the training data.

Sample submission - This CSV file provides the correct format for evaluation submissions, showing which columns must be included.

Store data - This CSV file contains store metadata including city, state, type, and cluster.

Oil data - This CSV file contains daily oil prices. Since Ecuador is heavily dependent on oil, and its economic health is vulnerable to oil price shocks, this file includes oil prices during both training and test periods.

Holidays data - This file contains metadata about holidays and events, with 6 columns. The 'transferred' type in the 'type' column indicates holidays that fall on one calendar day but were moved to another date by the government. A transferred day behaves more like a normal day than a holiday. To find the actual celebration date, look for the corresponding 'Transfer' type row. For example, the holiday "Independencia de Guayaquil" was transferred from 2012-10-09 to 2012-10-12, meaning it was celebrated on 2012-10-12. 'Bridge' days are extra days added to holidays (typically to create long weekends), often compensated by 'Work Days'—normally non-working days (like Saturdays) used to make up for the bridge. **Additional holidays are extra days added to regular calendar holidays, such as Christmas Eve.** We'll explore this in more detail during the EDA.

Some additional information: Employees are **paid twice monthly** (on the 15th and the last day of the month), which may affect supermarket sales. Also, a **magnitude 7.8 earthquake struck Ecuador on April 16, 2016**. People participated in relief efforts by donating water and other essential products, which significantly affected supermarket sales for several weeks after the earthquake.

Methodology

One crucial aspect to remember is that we're dealing with **time series data**. Unlike the static, structured datasets we might be used to, where values remain constant, time series data is dynamic; its values continuously change over time. Think of it like stock prices – the unit sales of products won't be the same today as they were yesterday. This inherent variability means we'll need to employ specialized **time series methodologies** for our predictions, which we'll delve into later.

Our immediate focus is to lay the groundwork. First, we'll **import all the CSV files** into our notebook. Following that, the essential step of **data cleaning** begins, involving tasks like meticulously handling and replacing any missing values to ensure data integrity. With our data prepped, we'll move into the core of **Exploratory Data Analysis (EDA)** and **data visualization**. Here, the real insights emerge. We'll meticulously analyze each column across the various files, identifying trends and understanding how different factors —such as **time**, specific **dates**, and other relevant features—impact sales. By thoroughly exploring these relationships, we'll gain a clear picture of the data's characteristics and the stores' historical performance. This comprehensive understanding will then guide us in selecting and applying the most appropriate time series methodologies to accurately predict future sales on our test data.

Important Note - "The code for the Exploratory Data Analysis (EDA) can be accessed either through **my Kaggle account** or **my GitHub repository.**"

Data Importing and Cleaning

Data files are imported from the kaggle competition. Data provided by the competition is already cleaned with no missing values and structured properly, there are only one missing values in the Oil CSV file.

Exploratory Data Analysis

Steps in EDA, EDA is the crutial part in analysing the data, we need to look how the features are playing with the targets columnn and among themselvess as well. To do so from the train data first 3 key features have been focused to analyze the trend between these columns and target column. These key columns include ‘date’, ‘store_nbr’, and ‘Family’.

From the initial view of the train data frame it is observed that the date column is continuous ranging from 1st January 2013 to 15th August 2017 in total 1684 days, 33 unique product families and 54 stores in total, it is observed that each product family data is present across all the 54 stores, because the data showed that the total number data points of each product family is 90936 which is 54*1684 = 90936, and 54*33*1684 = 3000888, that says everything, **the company ‘Corporación Favorita’ have 54 stores across Ecuadorian, with 33 product families in each store, and assuming that the store started from 1st of January 2013.**

When comparing the store vs sales, it is observed that the store number 44, 45, 46, 47, 48 and 49 have a good overall sales in past 4.5 years, and store number 22, 32, 35, 42, and 52 have vey less total sales. Reasons for this may be of various factors which will be discussed in the up coming parts, but overall it is observed that the total sales of particular group is more compared to reset other stores. The graph below is the bar plot of stores vs total sales.

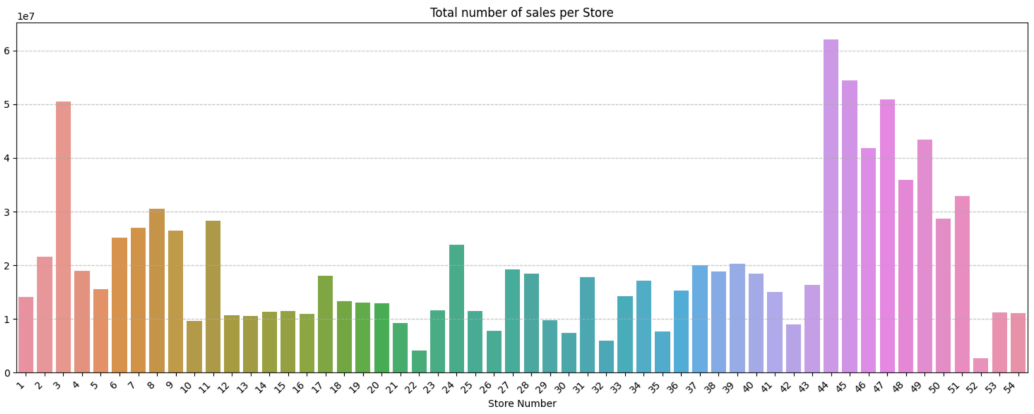


Fig 1 - Total no of sales per store of 54 stores.

Also, when analyzing the relationship between family types and total sales, a surprising pattern emerged: out of the 33 family types, only the ‘Grocery I’ category consistently generated the highest sales across all stores. Even more striking is the fact that, over the past 4.5 years, 13 family types recorded zero sales across all 54 stores. This raises a serious question about their presence in the inventory—maintaining these unproductive categories serves no practical purpose and represents a clear inefficiency in inventory management. It highlights that customers strongly trust and prefer the 'Grocery I' products offered by the company. On a broader level, this suggests that the company should consider focusing more on the 'Grocery I' category and, to a slightly lesser extent, 'Beverages', which also performed better than the rest. The graph below illustrates this insight clearly.

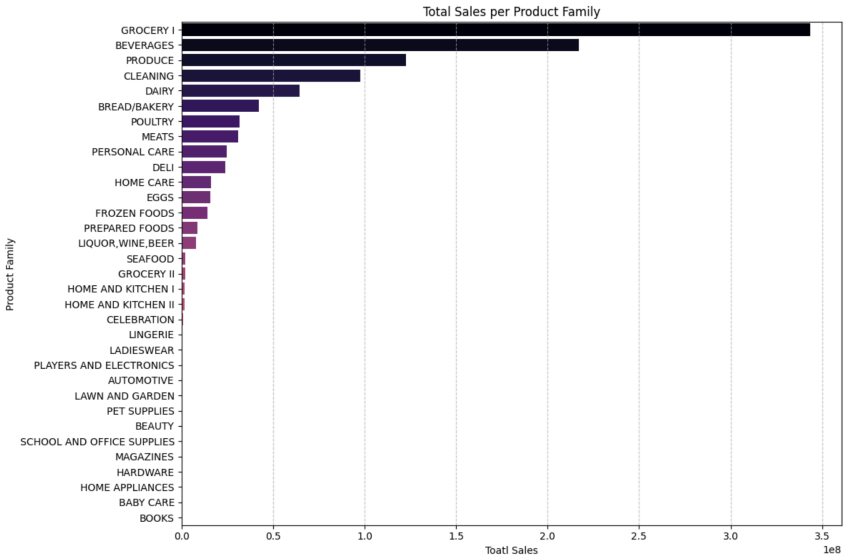


Fig 2

Now, one of the most important aspects to analyze in this time series data is the daily sales trend. We have data spanning 4 years, 7 months, and 15 days, and from this, it's crucial to understand how sales fluctuate on a daily basis. This analysis provides a broader perspective on the overall performance of the company. A graph has been plotted to visualize the trend between daily sales over this 4.5-year period across all 54 stores of the company. It is observed that, overall, the company's sales growth across all stores is steady but not dramatic—there is no drastic surge, but rather a gradual year-by-year increase. Compared to 2013, the sales in 2017 show noticeable improvement. Interestingly, the year 2014 saw highly fluctuating sales, while in 2015, the first half of the year had lower sales compared to the second half. In 2016, as mentioned earlier in the data exploration section, a 7.8 magnitude earthquake occurred in April, which led to a spike in sales, due to people participated in relief efforts by donating water and other essential products. Another trend noticed is that sales tend to peak at the end of each year and dip at the beginning of the following year. The graph below clearly depicts these patterns and provides valuable insights into the company's sales dynamics over time.

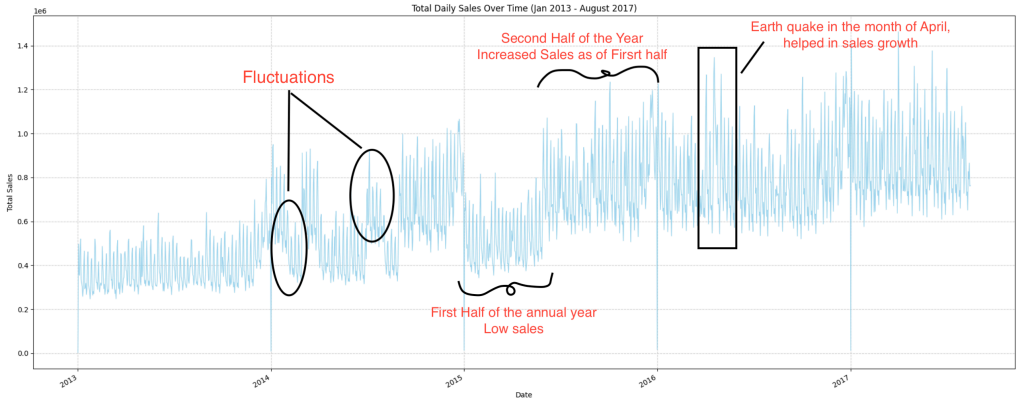


Fig 3 - Daily sales data of all 54 stores over 4.5 years of time

Store Data Analysis

Great, we have analyzed the relationship between store, family type, and day with respect to sales. However, it's important to note that this analysis reflects the overall data — encompassing all 54 stores and all 33 product families over a span of 4.5 years. When it comes to predicting unit sales, the forecast is also made at this overall level, since the training dataset includes all stores and all family types, but only for the next 15 days in August 2017. Therefore, for the purpose of this project, we will maintain our focus on overall comparisons and patterns in sales.

Let's begin with the stores. We have a total of 54 stores, and using the 'stores.csv' file, we extracted detailed information provided about each one. This file contains five columns: 'store_nbr' which identifies the store number, 'city' indicating the city in which the store is located, 'state' representing the state, 'type' which categorizes stores based on format, size, or pricing strategy, and 'cluster', a grouping of similar stores. It is essential to interpret each row carefully to understand the characteristics of each store.

The dataset reveals that there are 54 stores distributed across 22 cities in 16 different states of Ecuador. These stores are categorized into five types — A, B, C, D, and E — and grouped into 17 clusters based on similarity. In Figure 1, we analyzed the sales pattern for each store and observed that store numbers 44 to 49 have notably high sales. To understand this better, we examined what group, cluster, city, and state these stores belong to. Since the target column in this dataset is `store_nbr`, we analyzed the trends of other columns with respect to the store number.

A subplot of store number versus city, state, type, and cluster was created. From the graphs, we gained several insights: the maximum number of stores are located in 'Quito' city and the 'Pichincha' state. It was also observed that type 'D' stores are most common, and cluster number '3' contains the highest number of stores across Ecuador. Notably, as seen in Figure 1, store numbers 44 to 49 — which performed particularly well — are all located in the same city, 'Quito', and the same state, 'Pichincha'. Interestingly, they all fall under type 'A' stores, though they belong to different clusters. Given that Quito city in Pichincha state hosts many stores, and that type 'A' stores here are performing well, it's reasonable to conclude that this store type in this region contributes significantly to the company's sales.

A closer inspection of Figure 1 also allows us to group the stores based on performance trends in sales. We identified four distinct groups:

1. Stores numbered 6 to 9 – located in Quito and of type 'D'.
2. Stores numbered 12 to 21 – spread across various cities, mostly of type 'C'.
3. Stores numbered 36 to 41 – dispersed in different cities, with varied types and clusters, making pattern detection more difficult.
4. Stores numbered 44 to 49 – all in Quito, of type 'A', with strong and consistent performance.

From these groupings, we can draw the insight that among all cities, Quito is outperforming others, and within this city, type 'A' stores are yielding better sales than others, followed by type 'D'. On the other hand, the least performing stores vary significantly in location, type, and cluster. However, two stores in particular — store 22 and store 52 — stand out as clear underperformers. These two stores are in different cities, belong to different types and clusters, and do not follow any recognizable trend, indicating possible internal or operational issues that require attention.

Below, the figure represents the four subplots comparing store number with city, state, type, and cluster — offering a visual understanding of these relationships.

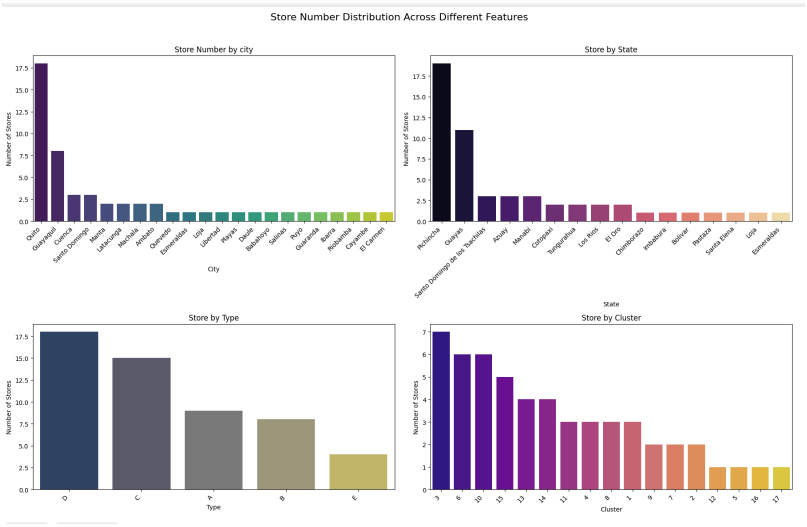


Fig 4

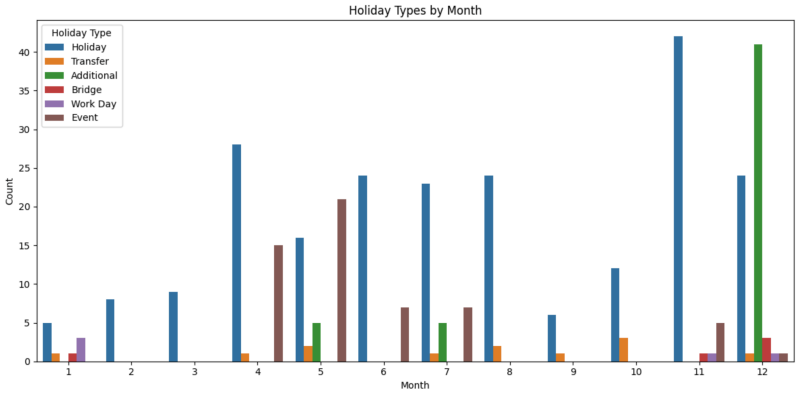
Holidays Data Analysis

Now, let's understand the relationship between holidays and their impact on sales. The holidays dataset contains 350 rows and 6 columns, which implies that there were a total of 350 holidays during the 4.5-year time span covered by our data. To analyze how these holidays have influenced sales, we can examine the `holiday_type` column, which includes six different types of holidays. By identifying what kind of holiday occurred on which date, we can begin to uncover patterns in how sales were affected on those specific days.

Additionally, the dataset provides valuable location-related details such as the `locale` and `locale_name` columns, which indicate the geographical scope of the holiday—whether it was a national event or localized to a specific city or region. This allows us to focus on how certain areas were impacted during holidays. For example, since our previous analysis highlighted that the city of Quito plays a major role in overall sales performance, it would be particularly insightful to look into how holidays in Quito influenced sales trends.

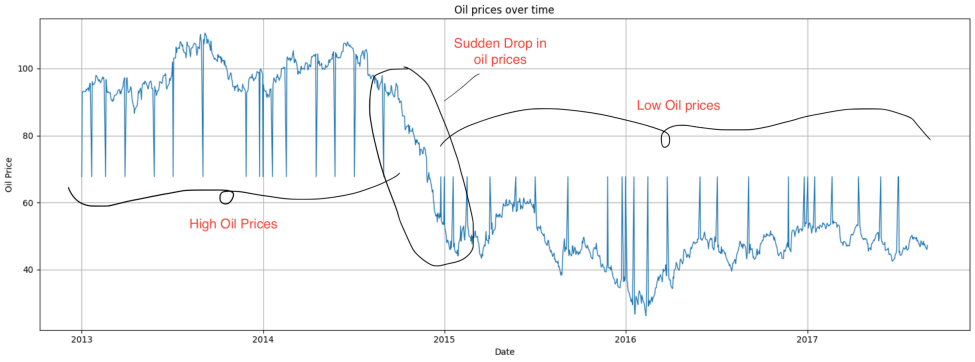
By filtering and analyzing holidays that occurred in Quito and overlaying this with the corresponding sales data, we can investigate whether sales typically increased or decreased on holidays, and if certain holiday types had more impact than others. This will give us a clearer picture of consumer behavior in response to different events and help the company make better planning and inventory decisions during holiday periods.

Monthly holiday trends were analyzed, and it was found that there were significantly more holidays and additional days off in the month of December compared to all other months. An interesting observation emerges when we compare this with the sales data in Fig. 3 — there is a clear trend showing that each year, the last part of the year, especially December, sees a notable spike in sales, whereas the month of January consistently experiences lower sales. From the holiday trends graph, it is also evident that January has the fewest holidays, events, or days off of any month. This correlation strongly suggests that customers tend to shop more during holiday periods than on regular days — which makes sense, especially considering major events like Christmas in December that naturally drive consumer spending. Therefore, the company should focus on ramping up production, introducing new products, and boosting inventory during major holidays and events to capitalize on increased demand. This would not only help maximize sales but also enhance customer satisfaction by ensuring availability during high-demand periods. Below is the graph representing the average monthly holiday trend across the entire 4.5-year span.



Oil Price Analysis

As discussed in the Data Understanding section, Ecuador is a country whose economy is highly dependent on oil prices. To explore this, oil stock price data was analyzed, and it was observed that oil prices were relatively high during the years 2013 to 2014. However, at the end of 2014, there was a sudden and sharp drop in oil prices, and from that point until August 2017, the prices remained consistently low. When we compare this trend with the sales performance shown in Fig. 3, an interesting pattern emerges. During the high oil price period of 2013 to 2014, overall sales were not particularly strong. Conversely, from 2015 to 2017 — the period of low oil prices — sales saw a noticeable improvement. In fact, as oil prices dropped at the end of 2014, there was a clear upward trend in sales beginning shortly afterward, as seen in the graph. This indicates a possible inverse relationship between oil prices and sales: when oil prices go up, sales tend to go down, and when oil prices decrease, sales appear to rise. This pattern suggests the need for deeper investigation into the causal factors behind this correlation. Importantly, if oil prices were to rise again in the future, the data warns of a potential decline in sales. Therefore, the company must prepare and take appropriate strategic measures to mitigate any negative impacts on revenue.



Insights from EDA

1. **Company Overview:** Corporación Favorita is a major grocery retailer operating in Ecuador. It runs 54 stores spread across 22 cities in 16 states. These stores are categorized into five types (A, B, C, D, and E) and grouped into 17 clusters of similar store formats.
2. **Product Assortment Consistency:** Across all 54 stores, the company has maintained a consistent product assortment of 33 family types since January 1, 2013, to August 8, 2017 — a period of 4.5 years. This shows the company's consistency in

offerings, but it also reveals a lack of inventory refresh or adaptation to changing consumer preferences.

3. **Sales Concentration by Product Family:** Over this 4.5-year period, 13 out of the 33 product family types recorded zero sales, indicating a clear disconnect between inventory and customer demand. This suggests an opportunity for the company to phase out non-performing product lines and instead focus on high-performing categories such as 'Grocery I', 'Beverages', 'Produce', and 'Cleaning'. This would lead to cost savings in inventory management and better alignment with customer needs.
4. **Sales Seasonality & Holiday Impact:** A consistent sales pattern is observed year after year, where sales peak in December (likely due to Christmas and holiday shopping) and decline in January. Additionally, sales tend to be higher during the mid-year (summer months) as well. This correlates strongly with the holiday and event calendar, suggesting that holidays significantly boost customer purchases. The company should capitalize on these periods by increasing stock, offering promotions, and improving customer service to maximize revenue and customer satisfaction.
5. **Macroeconomic Influence – Oil Dependency:** As Ecuador is an oil-dependent economy, national economic health is closely tied to global oil prices. It was found that sales performance is inversely proportional to oil prices: high oil prices coincided with reduced sales, while lower oil prices saw an increase in sales. This trend must be monitored continuously, as future increases in oil prices could negatively impact sales, prompting the need for contingency planning and strategic pricing.
6. **Store-Level Performance Patterns:** Among all stores, those located in **Quito city**, particularly of **type A**, exhibited consistently strong sales performance. On the contrary, two stores — located in different cities and store types — were identified as severe under performers. These do not follow any observed trend and might be suffering due to internal operational issues. These outlier stores warrant deeper investigation and potential corrective actions.
7. **Strategic Focus Recommendations:** Moving forward, the company should prioritize stores with high sales volumes, especially type A stores in Quito, as they likely reflect stronger customer loyalty and demand. Forecasting unit sales accurately in these stores is critical. At the same time, performance improvement strategies should be gradually rolled out to underperforming stores, while reevaluating or even closing those with persistently poor sales.