Capstone Three Final Report

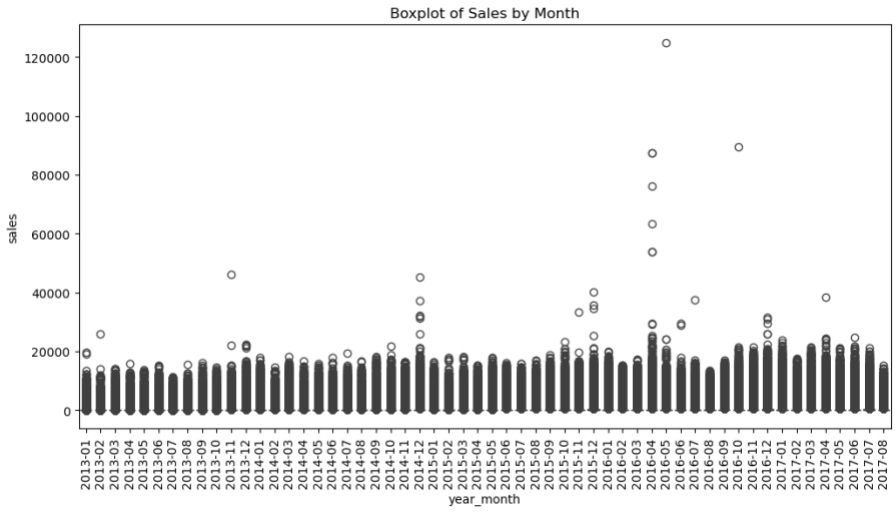
Store Sales Times Series Forecasting

Carl Riemann

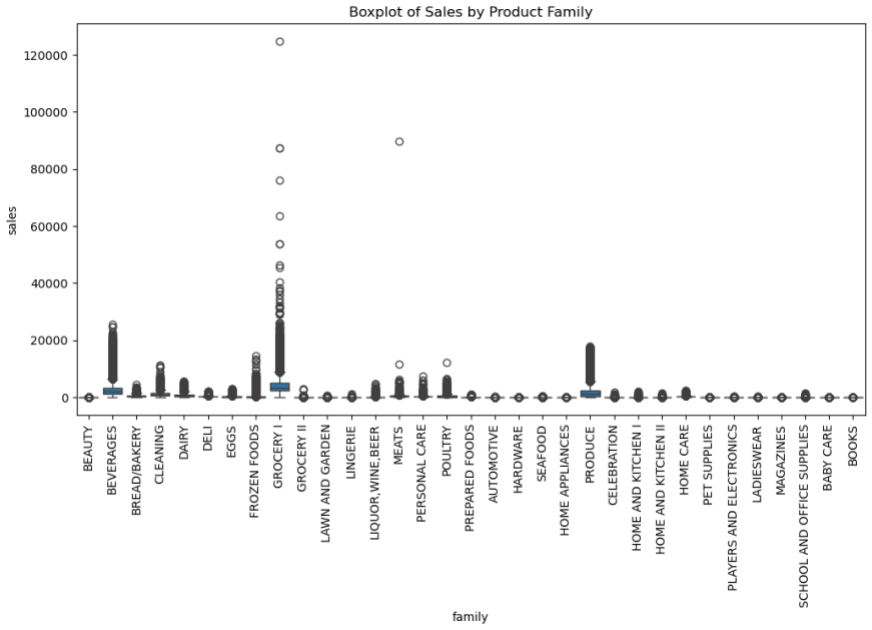
September 27, 2024

Corporación La Favorita is one of Ecuador's largest retail chains. With years of sales data available, building a reliable predictive model can help the company make data-driven decisions for future business strategies.

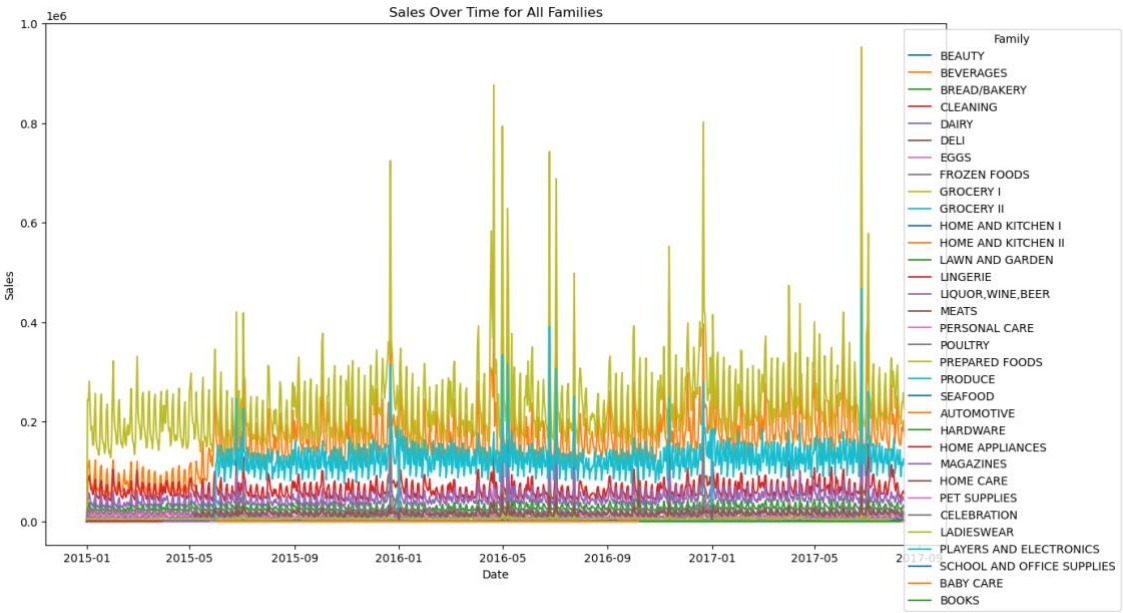
The objective of this project is to develop a time series forecasting model that predicts daily sales for product family of Corporación La Favorita using historical sales data from 2013 to 2017. The predictions aim to support the company's planning, on product sales.



The datasets used for this project includes daily sales data from the company's various stores across Ecuador, spanning from January 1, 2013, to August 15, 2017. There are a total of 6 data frames containing features such as store location, product family, promotions, and sales volumes, along with additional factors like oil prices and holidays.



To prepare the dataset for modeling, the data wrangling process involved several key transformations. First, I focused on cleaning the dataset by removing irrelevant columns and filtering out data before 2015. Additionally, I grouped the data by date and product family to capture seasonal sales trends. New features, such as holiday and promotion effects, were also incorporated to better understand the factors influencing sales fluctuations.



During the EDA I visualized sales trends over time using line plots, which revealed clear seasonal patterns and spikes around holidays. I also explored the distribution of sales across different product families and stores through histograms and box plots, identifying outliers and variations in sales performance. I identified that the 2016 earthquake in Ecuador caused sales to increase, maybe from peoples panic and overstocking on products. After the EDA I created the ARIMA model, the 'sales' column as the target variable, along with 'date' and 'family' as primary grouping factors to capture seasonality and trends at the product family level.

The ARIMA model performed particularly well in predicting sales for categories with relatively stable demand, such as AUTOMOTIVE and HARDWARE. However, for families like BEVERAGES and GROCERY I, larger errors indicate that external factors or a more complex model may be required to improve accuracy. Categories in the middle range, such as CLEANING and FROZEN FOODS, show reasonable performance, but there is still room for improvement in capturing their sales trends more accurately.

	family	MAE	MSE	RMSE
0	BEAUTY	3.524	26.717	5.169
1	BEVERAGES	1441.531	4616627.801	2148.634
2	BREAD/BAKERY	270.864	130389.475	361.095
3	CLEANING	445.531	400698.381	633.007
4	DAIRY	444.819	416453.657	645.332
5	DELI	136.269	34005.661	184.406
6	EGGS	120.203	31123.843	176.420
7	FROZEN FOODS	103.125	67279.670	259.383
8	GROCERY I	1683.797	6795949.540	2606.904
9	GROCERY II	18.576	1199.928	34.640
10	HOME AND KITCHEN I	21.435	1819.764	42.659
11	HOME AND KITCHEN II	20.158	1586.826	39.835
12	LAWN AND GARDEN	9.938	253.654	15.927
13	LINGERIE	5.846	130.731	11.434
14	LIQUOR,WINE,BEER	67.990	15220.767	123.372
15	MEATS	207.055	310519.596	557.243
16	PERSONAL CARE	133.540	39412.394	198.526
17	POULTRY	245.386	153211.961	391.423
18	PREPARED FOODS	64.018	9904.734	99.523
19	PRODUCE	1544.864	5570640.608	2360.220
20	SEAFOOD	23.185	1188.272	34.471
21	AUTOMOTIVE	4.246	36.995	6.082
22	HARDWARE	1.412	4.253	2.062
23	HOME APPLIANCES	0.959	1.843	1.358
24	MAGAZINES	5.726	68.436	8.273
25	HOME CARE	115.238	26322.082	162.241
26	PET SUPPLIES	6.354	88.920	9.430
27	CELEBRATION	9.751	250.855	15.838
28	LADIESWEAR	11.340	258.408	16.075
29	PLAYERS AND ELECTRONICS	7.070	122.453	11.066
30	SCHOOL AND OFFICE SUPPLIES	13.083	1729.836	41.591
31	BABY CARE	1.274	10.628	3.260
32	BOOKS	1.462	6.393	2.528