Store Sales Times Series Forecasting

Capstone Final Presentation

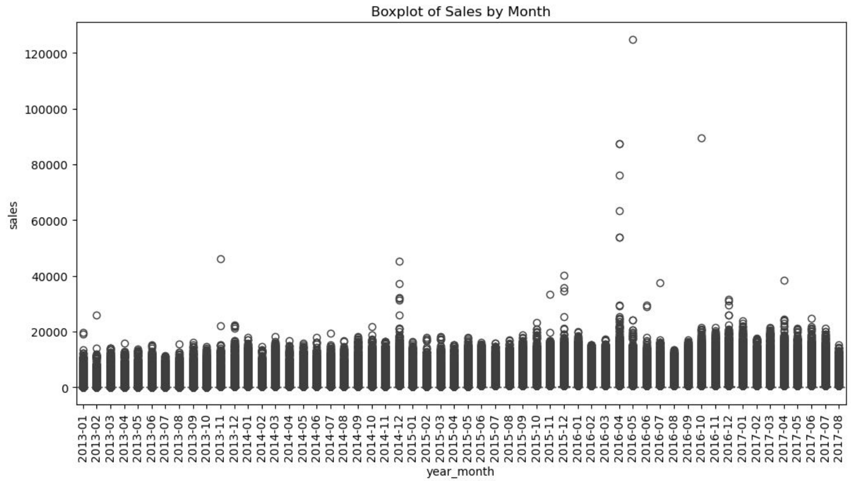
Carl Riemann

Introduction

Corporación La Favorita is one of Ecuador's largest corporations, operating across various industries.

This project focuses on analyzing historical daily sales data (2013-2017) from Corporación La Favorita.

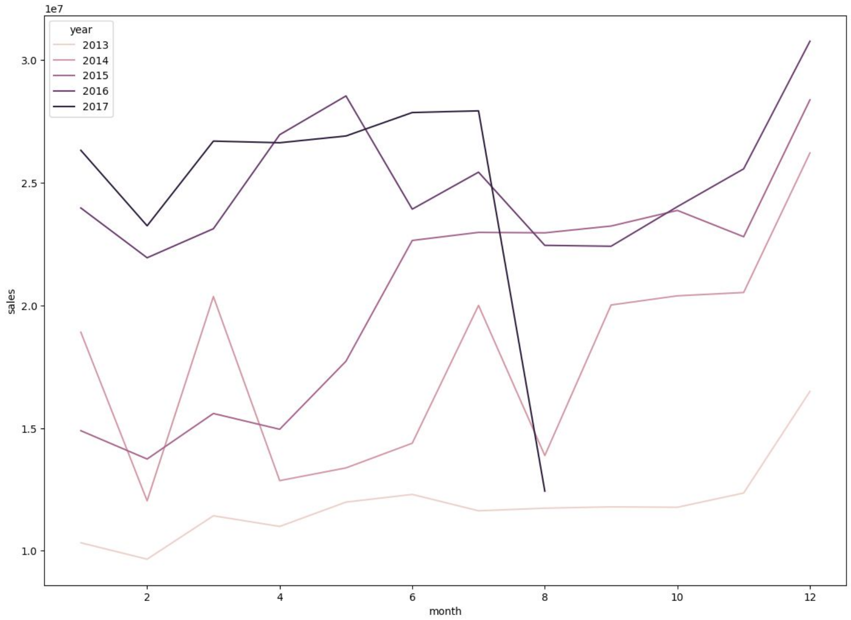
The goal is to build a predictive model to forecast future sales trends



Problem Statement

Can we accurately predict future sales trends using large volumes of historical data?

-Develop a Time Series Model to analyze and forecast sales trends from 2013 to 2017.



Data Overview

- train.csv: Historical sales data
- test.csv: Test data for sales
- oil.csv: Daily oil prices
- holidays_events.csv: Holidays and events metadata
- stores.csv: Store information

```
Missing values in Stores DataFrame:
 store_nbr     0
Missing values in Train DataFrame:
 id             0
date           0
store_nbr      0
family         0
sales          0
onpromotion    0
dtype: int64
ataFrame:
 dcoilwtico    43
 dtype: int64
```

Data Reshaping and Wrangling

The raw data files included columns with varying formats and inconsistencies such as: missing oil prices, holidays_df date format, or unnecessary columns.

Data from the multiple files provided needed merging for consistency and analysis.

```
Columns in train_df: Index(['id', 'date', 'store_nbr', 'family', 'sales', 'onpromotion', 'year',
       'month', 'day', 'day_of_week', 'week_of_year', 'lag_1', 'lag_7',
       'rolling_mean_7', 'rolling_mean_30'],
      dtype='object')
Columns in test_df: Index(['id', 'date', 'store_nbr', 'family', 'onpromotion', 'year', 'month',
       'day', 'day_of_week', 'week_of_year', 'lag_1', 'lag_7',
       'rolling_mean_7', 'rolling_mean_30'],
      dtype='object')
Columns in holidays_df: Index(['date', 'type', 'locale', 'locale_name', 'description'], dtype='object')
Columns in oil_df: Index(['date', 'dcoilwtico'], dtype='object')
Columns in stores_df: Index(['store_nbr', 'city', 'state', 'type', 'cluster'], dtype='object')
```

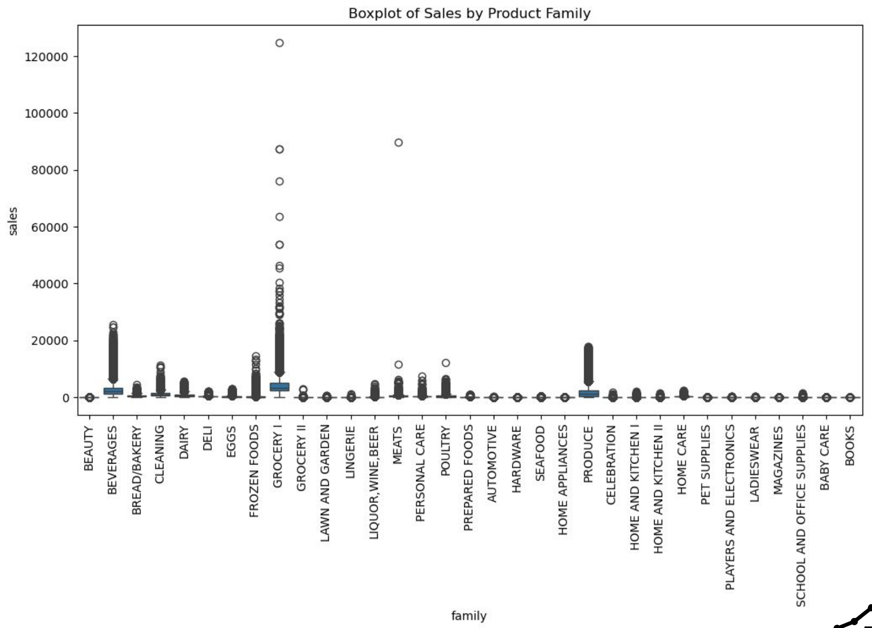
```
Columns in train_df: Index(['id', 'date', 'store_nbr', 'family', 'sales', 'onpromotion', 'year',
       'month', 'day', 'day_of_week', 'week_of_year', 'lag_1', 'lag_7',
       'rolling_mean_7', 'rolling_mean_30', 'city', 'state', 'store_type',
       'cluster', 'dcoilwtico', 'holiday_type', 'has_promotion'],
      dtype='object')
Columns in test_df: Index(['id', 'date', 'store_nbr', 'family', 'onpromotion', 'year', 'month',
       'day', 'day_of_week', 'week_of_year', 'lag_1', 'lag_7',
       'rolling_mean_7', 'rolling_mean_30', 'city', 'state', 'store_type',
       'cluster', 'dcoilwtico', 'holiday_type', 'has_promotion'],
      dtype='object')
```

Exploratory Data Analysis (EDA)

Analyzed the distribution of sales across different product families to identify trends and outliers.

Grocery I and Produce, have significantly higher sales and exhibit notable outliers

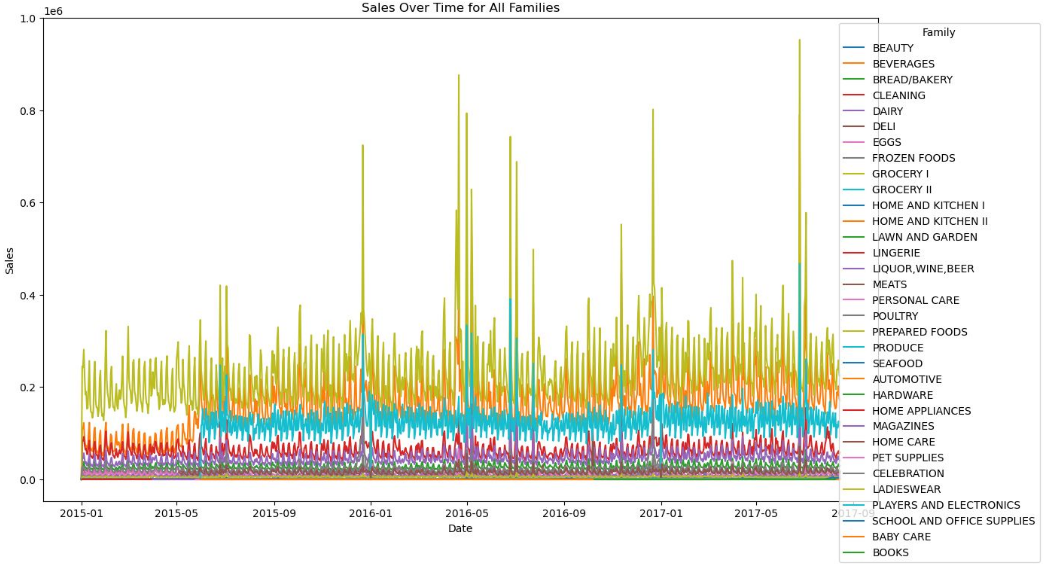
The EDA will guide the model-building process, as families may require separate handling or more detailed feature engineering.



Feature Selection

Grouped data by 'date' and 'family' to handle product categories separately and ensure each family had its unique seasonal trends.

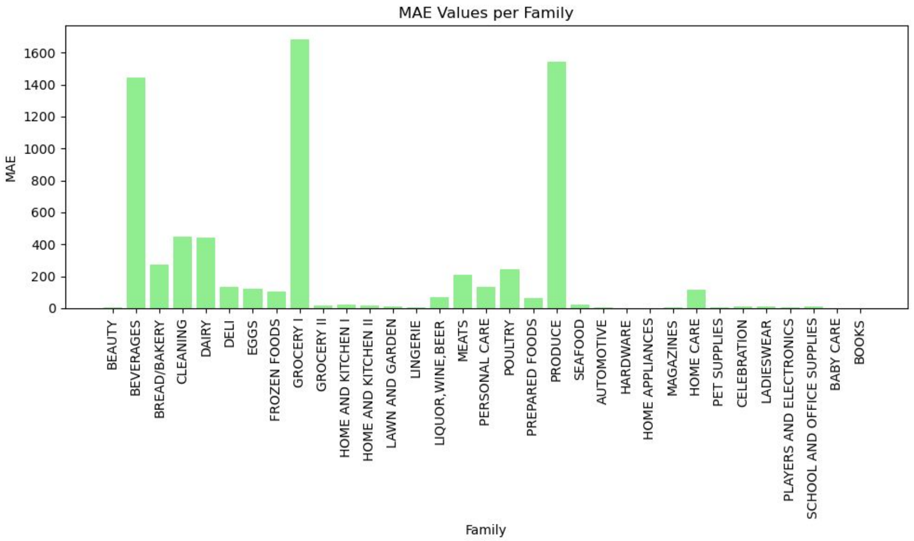
Removed year 2013-2014 for lighter load on model building.



Model Training

The model was trained on the following features:

- Differenced sales data (first-order differencing of the sales)
- Family type (each family of products was modeled individually)
- Time series structure (date components were implicitly captured in the sales sequence, though not as explicit features like year or month)



Model Evaluation

The ARIMA model performs well in predicting sales for categories with relatively stable demand, such as AUTOMOTIVE and HARDWARE.

For families like BEVERAGES and GROCERY I, large errors suggest that additional external factors or a more complex model may be needed for better accuracy.

Families in the middle range, such as CLEANING and FROZEN FOODS, show room for improvement but are reasonably well-captured by the model

	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

Future Work/Recommendations

Several families like FROZEN FOODS, BREAD/BAKERY, CLEANING have moderate error values (e.g., RMSE between 200 and 600), suggesting there is still room for improvement.

BEVERAGES and GROCERY I show the highest errors across MAE, MSE, and RMSE, but they also have the most outliers, also suggesting room for improvement.

Overall there is still room for model improvement some personal suggestions are:

- modeling with more dates
- removing only the earthquake dates (as they caused a lot of outliers)
- incorporating more variables that capture more trends (holidays/events/promotions)