



# Grocery Store Sales – Time Series Forecasting

Data Science Capstone Project 2

by Zhiling Xie


# 1. Problem

- **Client:** Corporación Favorita, a large Ecuadorian-based grocery retailer.
- **Objective:** Increase revenue and please customers by having just enough of the right products at the right time.
- **How:** More accurate forecasts of product sales could help the retailer ensure they have the right, sufficient product available.
- **Project Goal:**
  - Develop time series forecasting models for multiple products sold at multiple Favorita stores.
  - Use the predictive models to provide guidance for Favorita's product stocking strategy.





## 2. Approach

- 2.1 Data Wrangling
  - 2.2 Exploratory Data Analysis (EDA)
  - 2.3 Baseline Modeling
  - 2.4 Extended Modeling
- 



## 2.1 Data Wrangling

### ➤ Data information

#### Product Sales

- Sales of 33 product families and 54 Favorita stores in Ecuador.
- Period: Jan. 2013 – Aug. 2017
- Time series resolution: Daily
- Forecasting horizon: 16 days

#### Other features

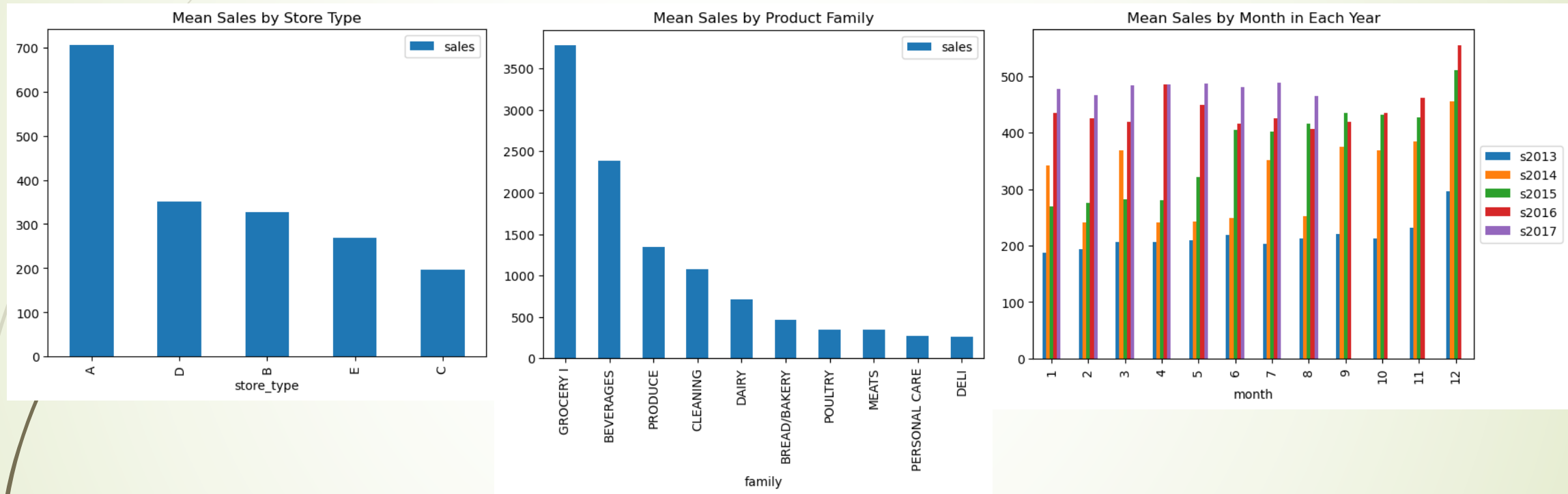
- Categorical : store's city, state, type, and cluster, holiday or weekend.
- Numerical : daily oil price, daily transactions per store, quantity of on-promotion products.

### ➤ Data Cleaning – missing values and time gaps

- Sales do not have missing values but have time gaps on every Christmas day.
- Daily transactions, daily oil price, and holiday columns have missing values & time gaps.
- Missing values and time gaps in this dataset were filled using forward fill method.

## 2.2 EDA

### ➤ Mean Sales by Store, by Product Family, and by Year & Month



- Type-A stores have the highest mean sales.
- Grocery I, Beverages, and Produce are the top 3 product families.
- The sales were increased by year and had an annual cycle.

## 2.3 Baseline Modeling - one time series forecast

- ARIMA model - MA(14): MAPE 10.5%.
- Linear Regression model: MAPE 11.0%.
- Random Forest model: MAPE 12.0%
- **Most important features** identified from both linear regression and random forest models:
  - Weekend or Weekday
  - sales\_lag1 (previous day's sales)
  - sales\_7days\_avg (average sales for the previous 7 days)

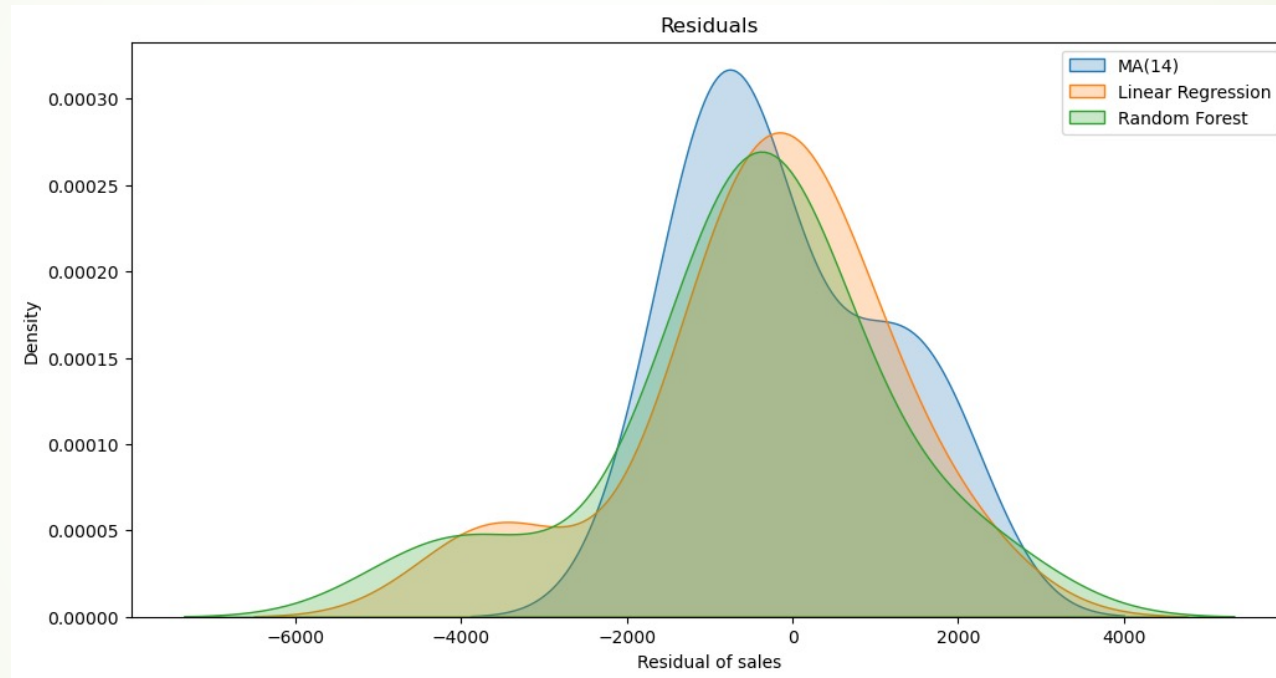


## 2.4 Extended Modeling - multiple time series forecast

- **pmdarima.auto\_arima:** Automatically selected best p, d, q values for each time series.
  - The Best MAPE is about 7%, and the Worst MAPE is about 26%.
- **Prophet:** Got reasonable forecasts on multiple time series with no manual effort.
  - The Best MAPE is about 8%, and the Worst MAPE is about 22%.
- **mljar-supervised AutoML:** Applied multiple model algorithms for each time series and got an ensemble result.
  - The Best MAPE is about 6%, and the Worst MAPE is about 21%.

### 3. Findings

#### ➤ Baseline Modeling Comparison

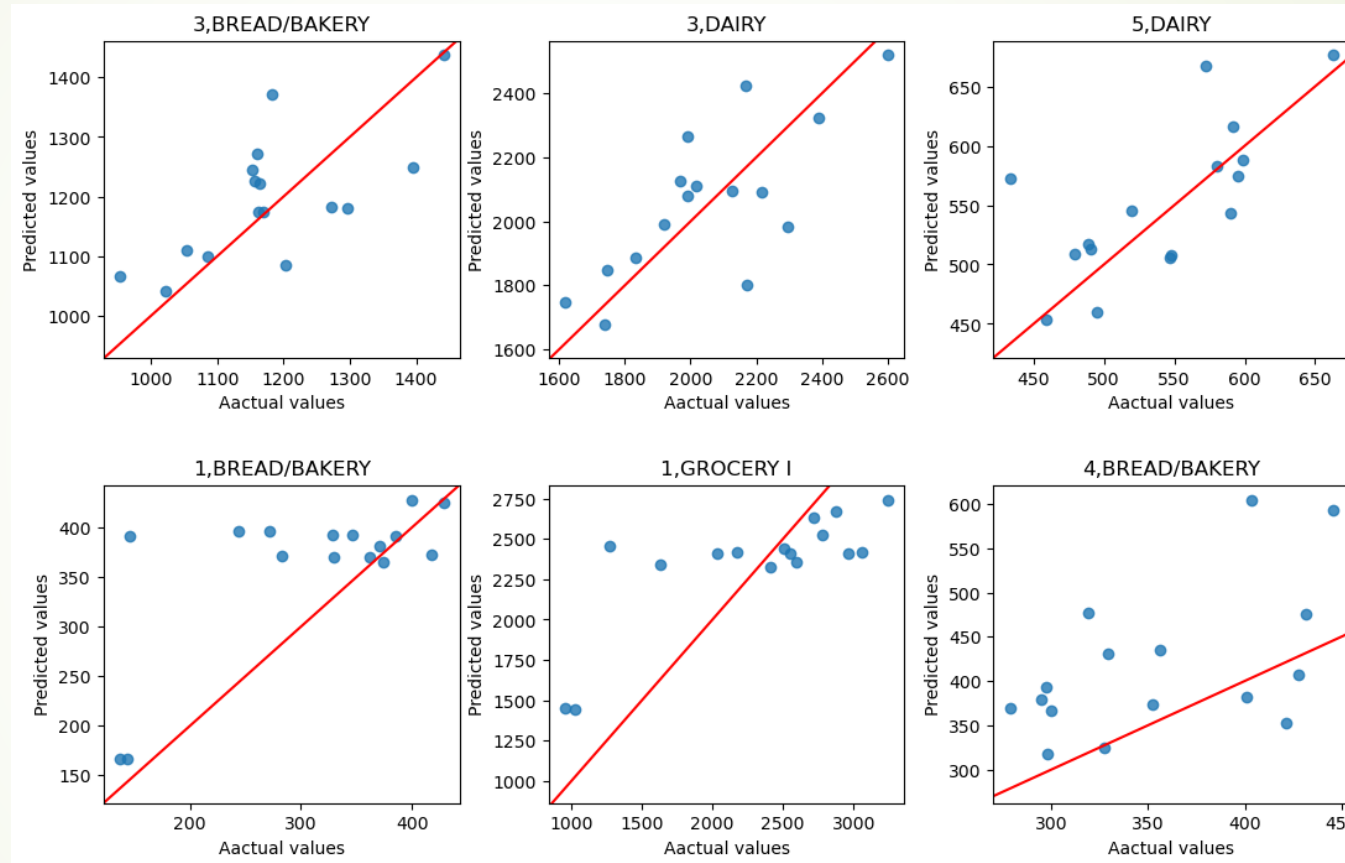


- MA(14) model had a relatively concentrated distribution, but its center was offset from zero.
- Linear regression and random forest had similar distributions, and the linear regression was slightly better.



## Extended Modeling Comparison

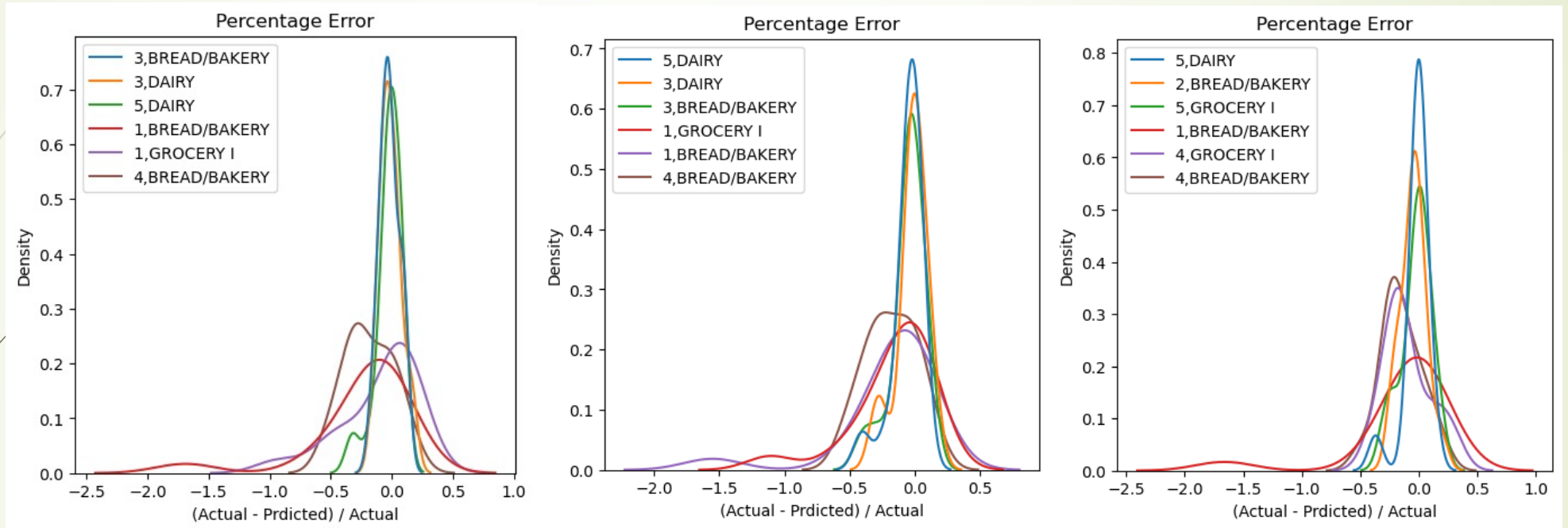
Best 3:



Worst 3:

- The scatter points in the upper Best 3 plots are closer to the  $y=x$  line than those in the lower Worst 3 plots.

## Extended Modeling Comparison



- Distribution of Percentage Error plots (for three extended models respectively):
  - The center lines of the "Best 3" are closer to zero and their distributions are also more concentrated.

## 4. Conclusions and Future Work

- We applied several models for the sales time series forecasting and found that different models had similar results with slight differences.
- Any of these models can be used for the grocery sales predictions.
- We found that large sales amounts were easier to predict, and the predictions of small sales products always had large errors.
- **Current Issue and Future Work:**
  - We forecasted only 18 time series in the Extended Modeling section instead of the thousands provided by the original data. The next step is to compute the forecast of thousands of time series.
  - We may need larger computing sources for large amount of data (such as high- performance computing).
  - Hierarchical forecasting approach: there are similarities between products/stores that might help with the forecasting. Grouping the product/store together along a product/store hierarchy and then use hierarchical forecasting to improve accuracy.
  - Additionally, we can further conduct forecasting in a Cloud-based environment that would allow us to simulate a realistic scenario for which new data is regularly ingested. The models are computed regularly (e.g., weekly), and results are also stored in the Cloud.