Corporación Favorita

Grocery Sales Forecasting

MSCA 31006 – Time Series Analysis and Forecasting

Meet the Team

Aarti Rao

Grace Chai

Jenny Huang

Meghna Diwan

Agenda

- Introduction
  - Background
  - Problem Statement
- Exploratory Data Analysis
  - Data Properties
  - Data Transformations & Assumptions
- Proposed Solution
  - Linear Regression
  - Hierarchical Time Series
  - Prophet
- Results
- Future work

Introduction

Background

Corporación Favorita is an Ecuadorian grocery retailer company based in the city of Quito, Ecuador. It is the largest non-state company in the country with hundreds of supermarkets and over 200,000 products on their shelves.

Corporación Favorita operates*:

- Supermaxi (35 locations)
- Megamaxi (12 locations)
- Akí (38 locations)
- Gran Akí (17 locations)
- Súper Akí (5 locations)

*as of 2016

Problem Statement

Sales forecasting is very important for brick-and-mortar grocery stores as they are always in a delicate dance with demand and supply (purchasing).

Forecasting more than the actual demand leads to grocers being stuck with overstocked, perishable goods.

Forecasting less than the actual demand leads to popular items selling out quickly, leaving money on the table and customers fuming.

Goal - To accurately forecast the item sales for Corporación Favorita across cities, stores and item families using time series analysis to ensure they please customers by having just enough of the right products at the right time.

Exploratory Data Analysis

Overview

Training Dataset

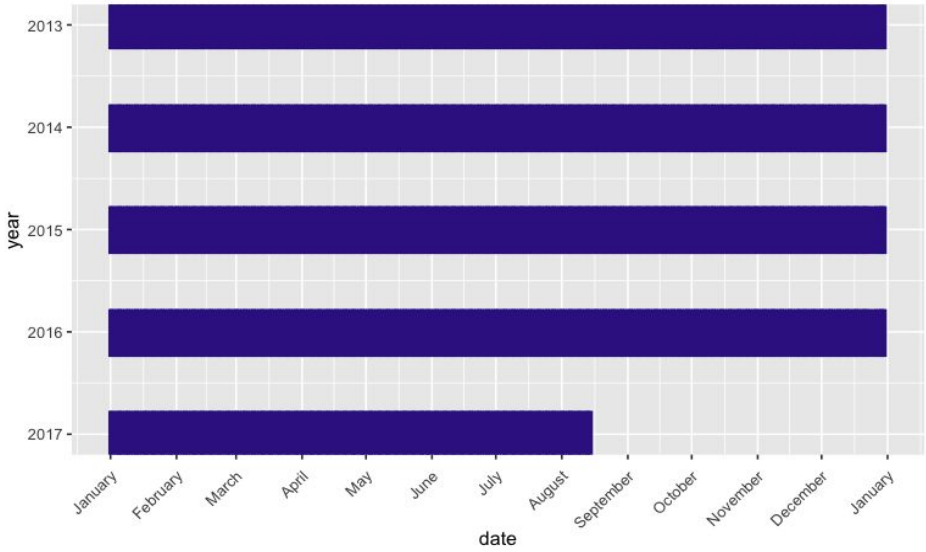
No. of Stores: 54

No. of Store operators: 5

No. of Items: 4100

No. of Item Family: 33

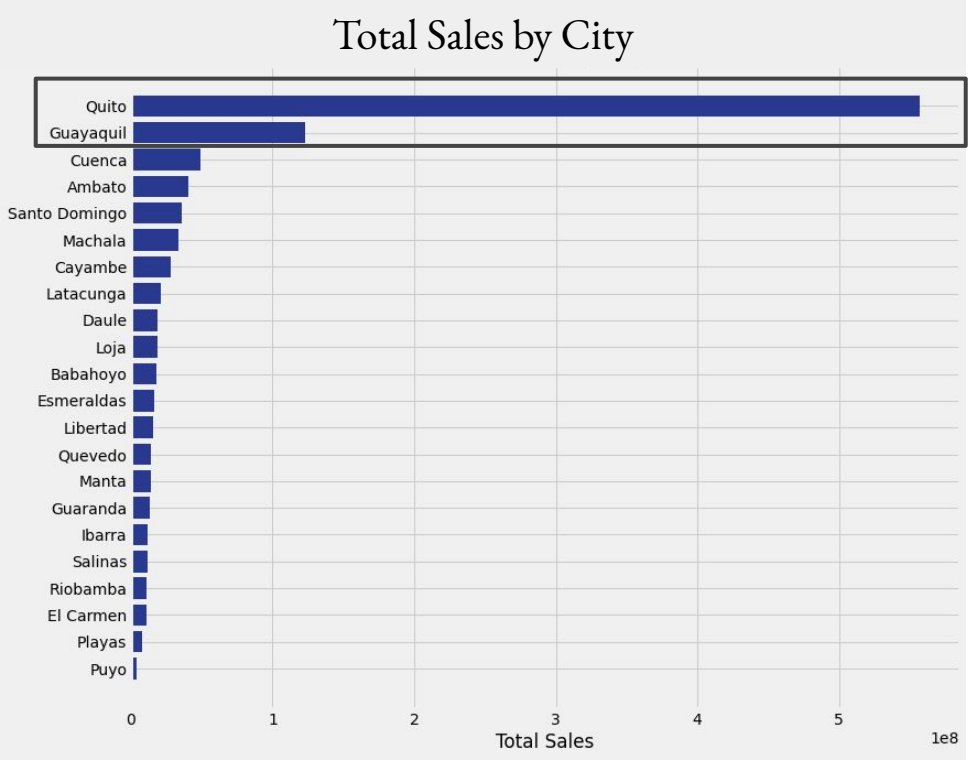
Corporación Favorita Daily Sales



Subset - City

Majority of the sales are in Quito and Guayaquil

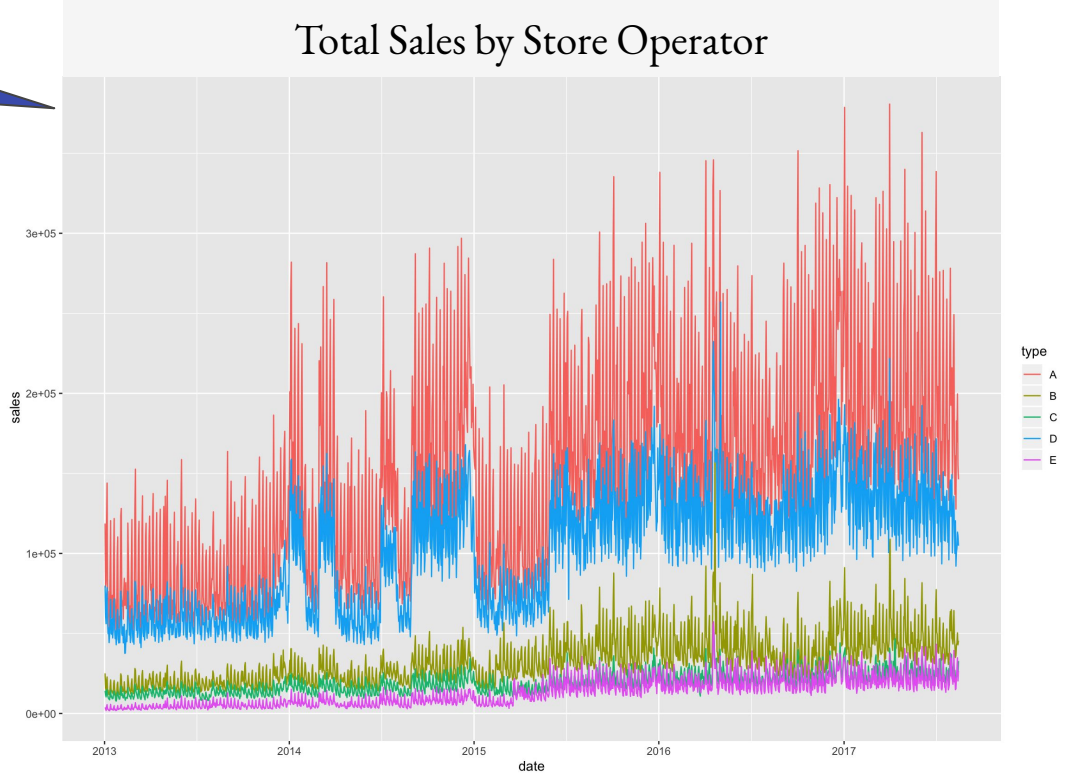
Subset data to only Quito and Guayaquil stores



Subset - Store Operators

Corporación Favorita has 5 different operators

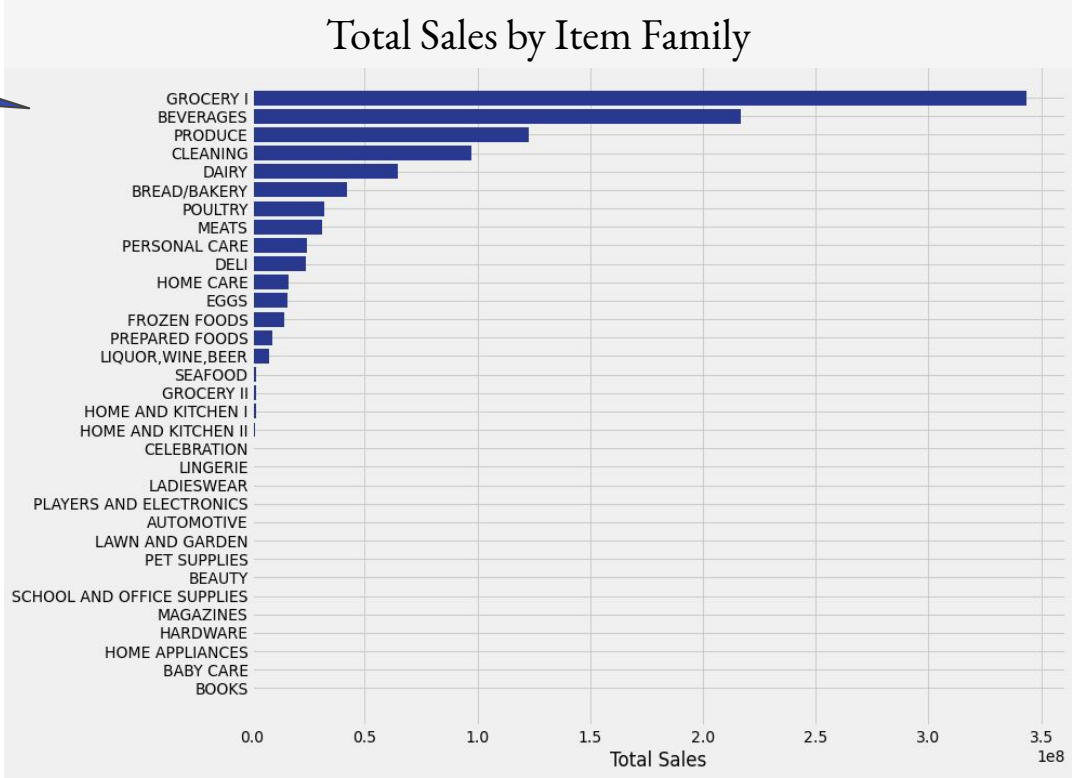
Group stores in each city by operator



Subset - Item Families

Food, Cleaning and Personal Care categories dominate sales

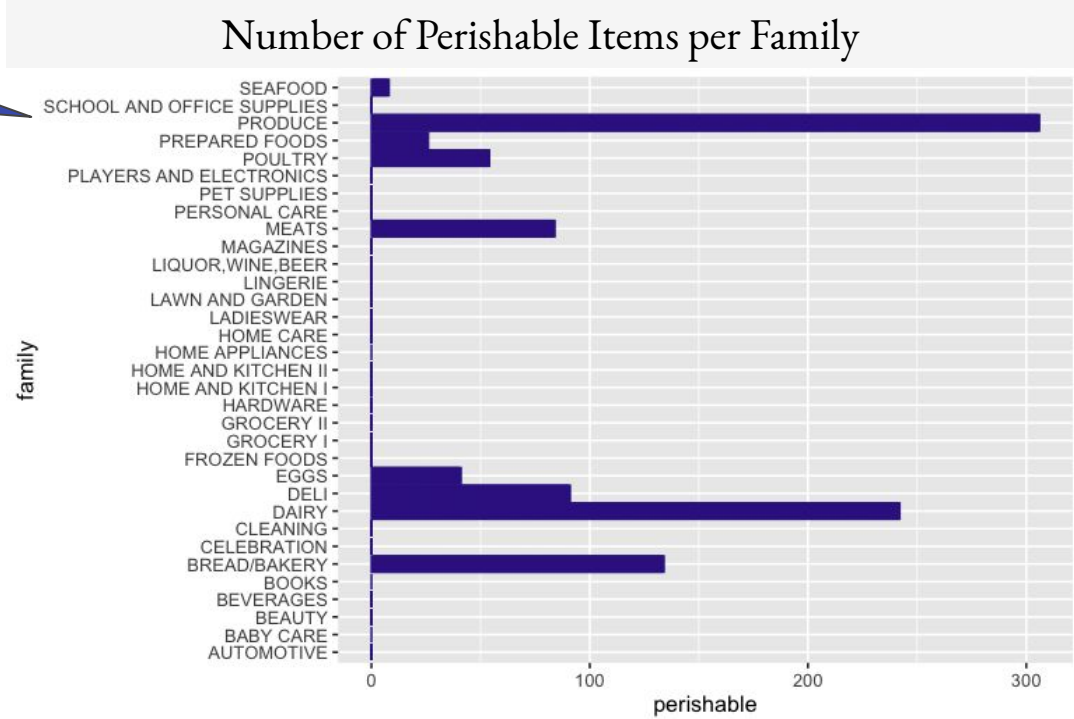
Keep 5 most selling item families per city and store operator



Covariates - Perishable

As expected, only food categories have perishable items confirming their high sales

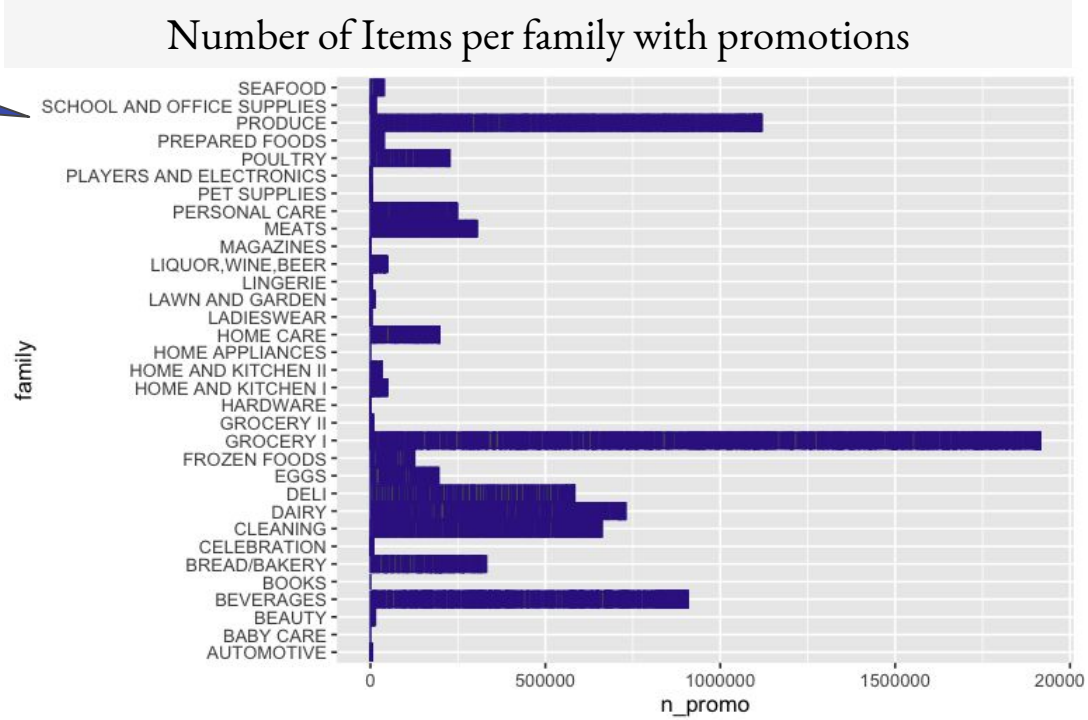
Counted number of perishable items in each family



Covariates - Promotions

Highest number of items with promotions is Grocery, which also has the highest sales

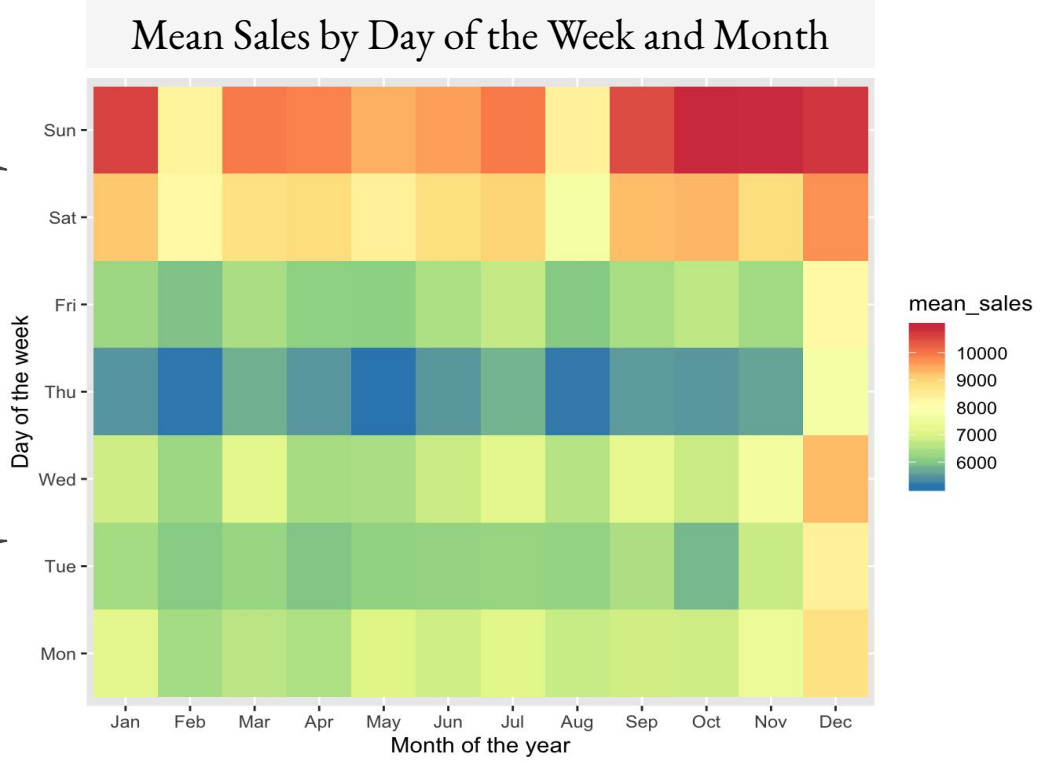
Counted number of items on sale in each family



Covariates - Seasonality

Saturday and Sunday are the most popular days of the week

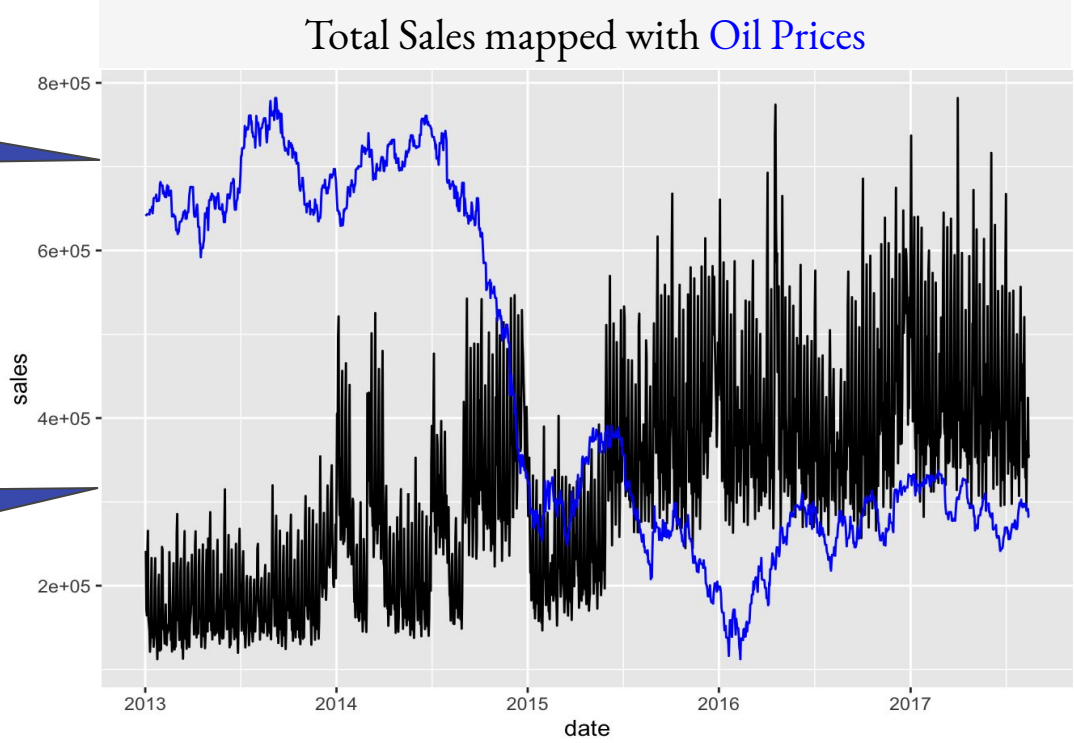
December has the highest sales in each year due to the holiday season



Covariates - Oil Prices

Ecuador has 0.5% of all oil reserves in the world

As oil prices increase, world demand for oil decreases resulting in less spending money in Ecuador

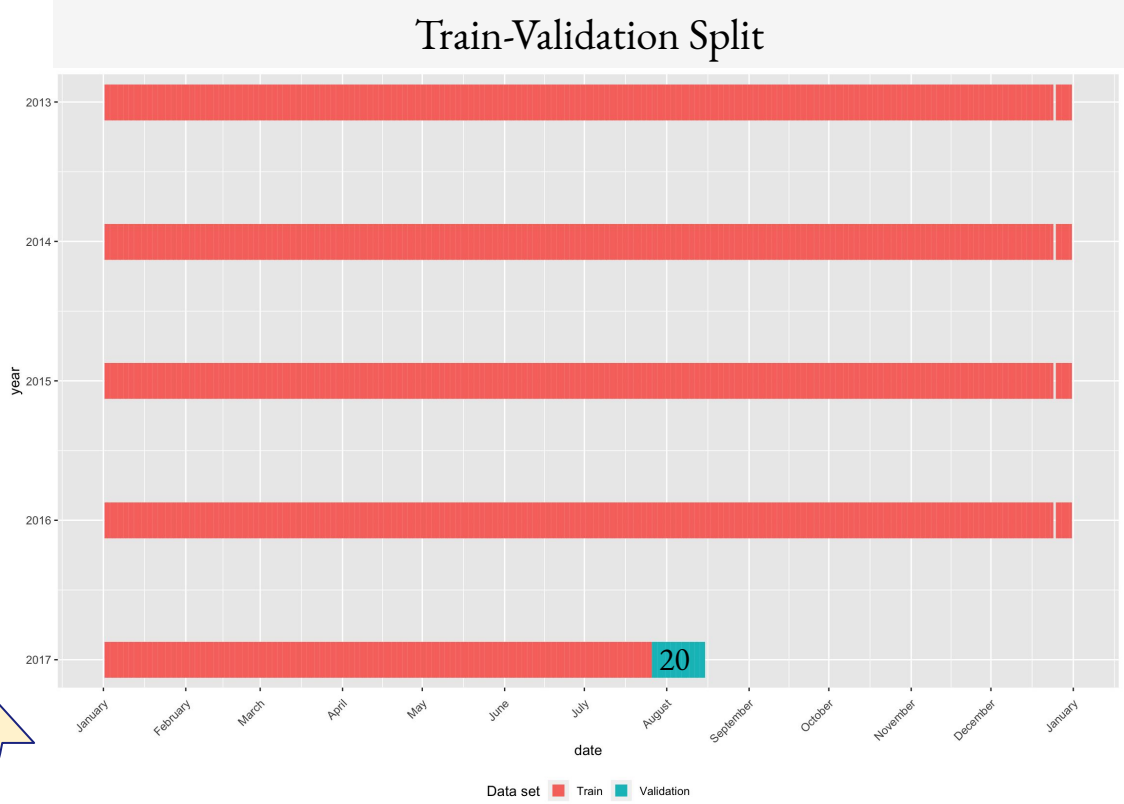


Subset data to only Quito and Guayaquil stores

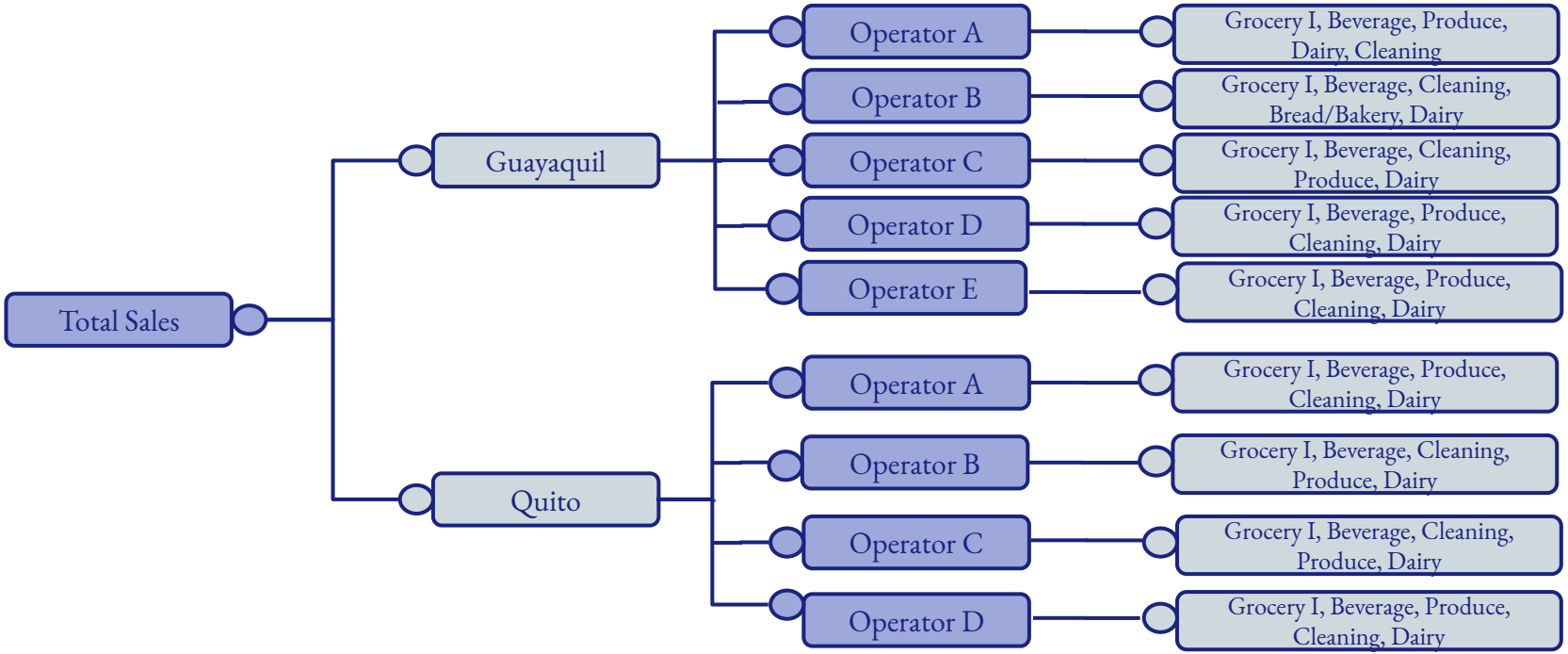
Group stores in each city by operator

Keep 5 most selling item families per city and store operator

Counted number of items on promotion and perishable in each family



Data Structure



LEVEL 1 LEVEL 2 LEVEL 3 LEVEL 4

Proposed Solution

Models Implemented

LINEAR REGRESSION (Top-Down Method)

- Simple Linear Regression
- ARIMA with Xreg
- ARIMA with Fourier & Xreg

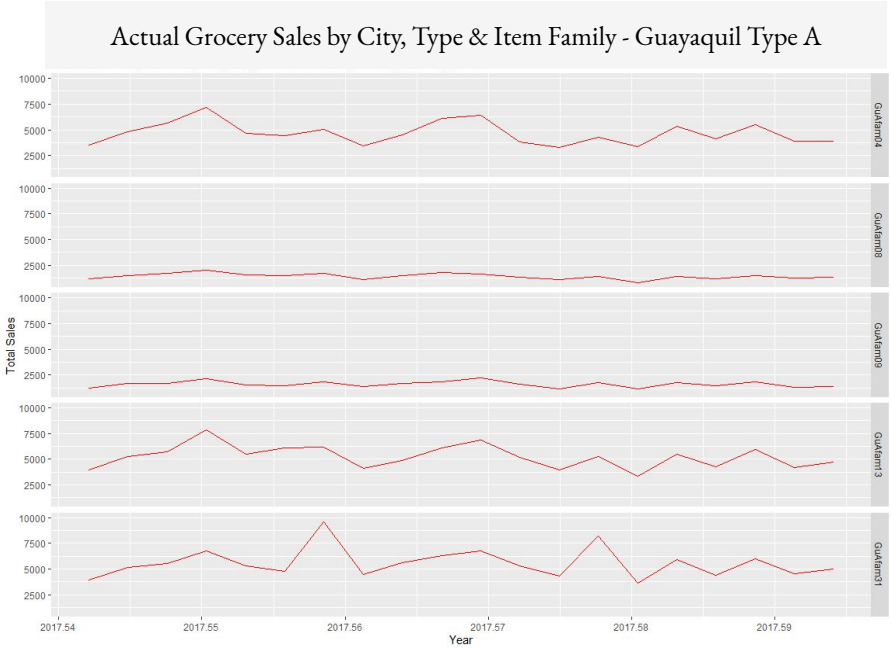
PROPHET (Top-Down Method)

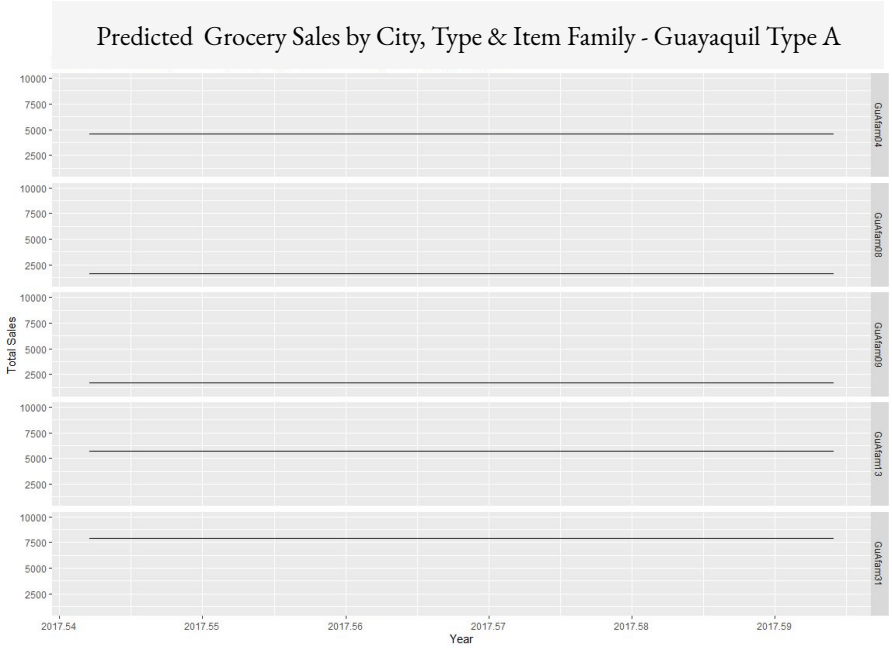
- Prophet with Seasonality
- Prophet with Seasonality & Holidays

HIERARCHICAL

- Forecasting using Random Walk
- Forecasting using ARIMA
- Forecasting using ARIMA with Xreg

Base Model - Forecast with Random Walk





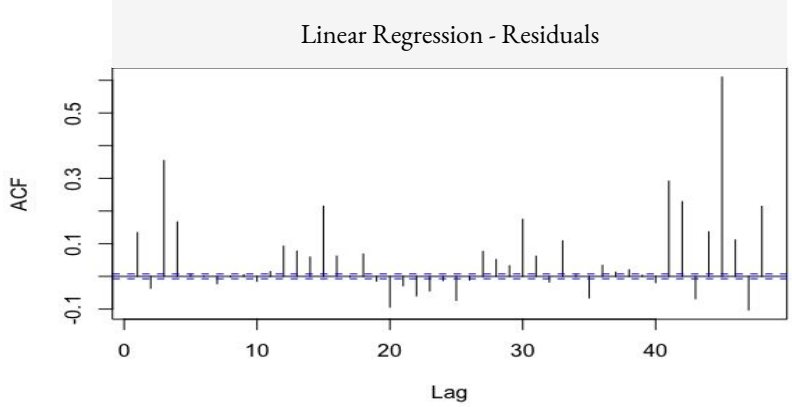
SMAPE on Validation Set: 15.2%

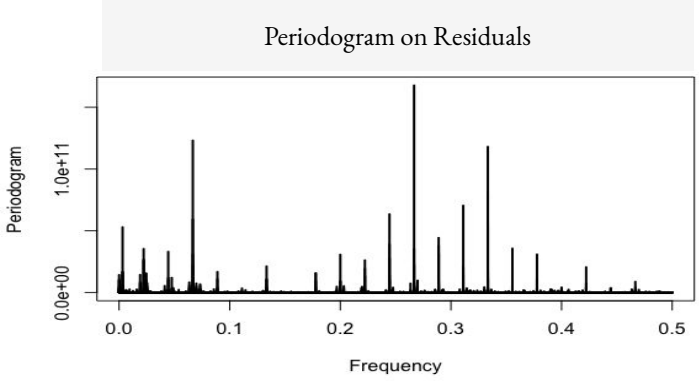
Simple Linear Regression

Predictors: City, Store Operator, Item Family, No. of Perishable Items, No. of Promotional Items. Oil Price, Flag indicating Holiday, Flag indicating if the holiday was national, Flag indicating if the date was a pay day

Response: Total Sales

Adj R- squared: 0.57

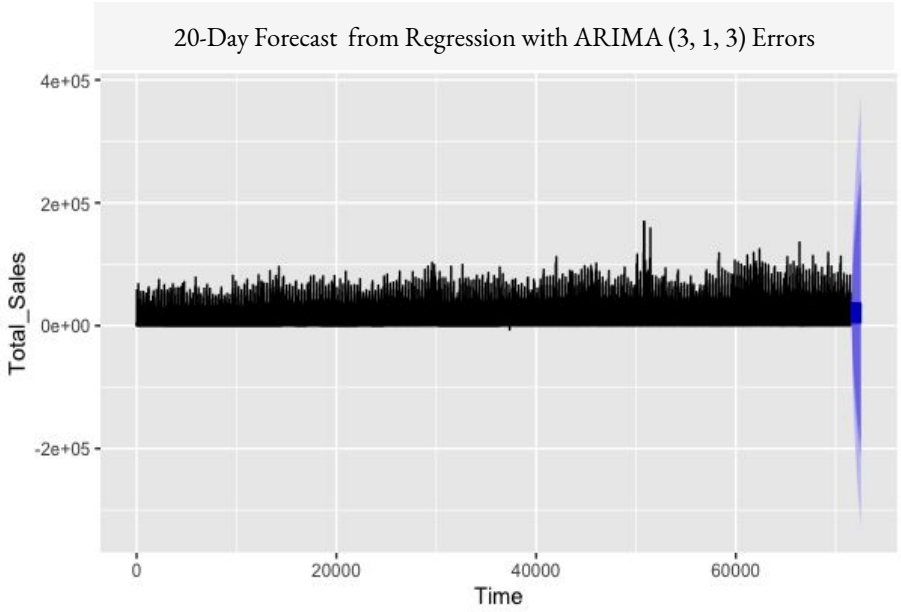


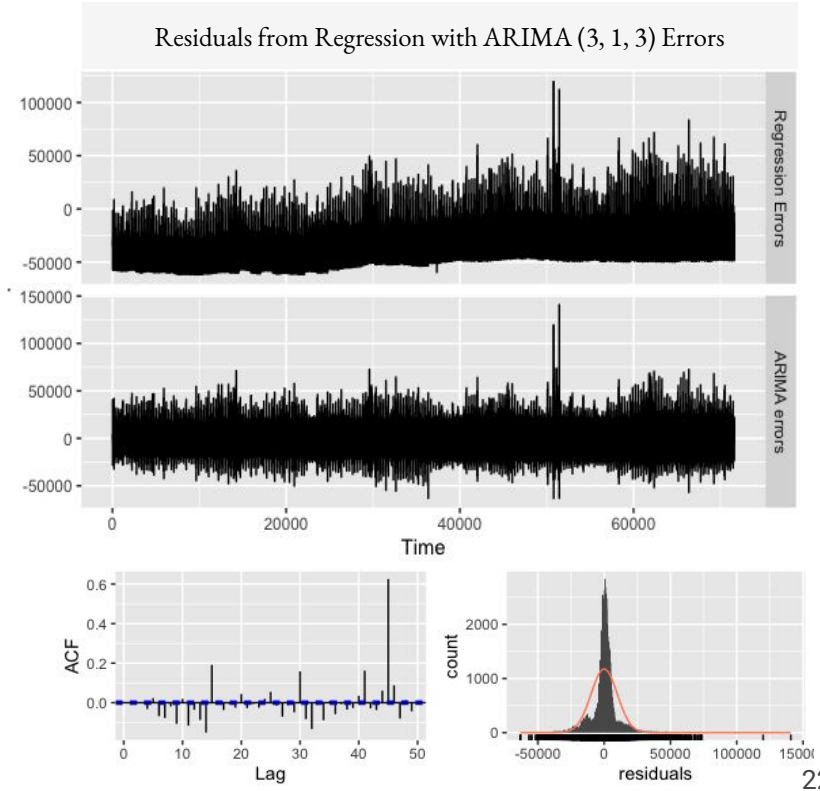


Linear Regression with ARIMA Errors

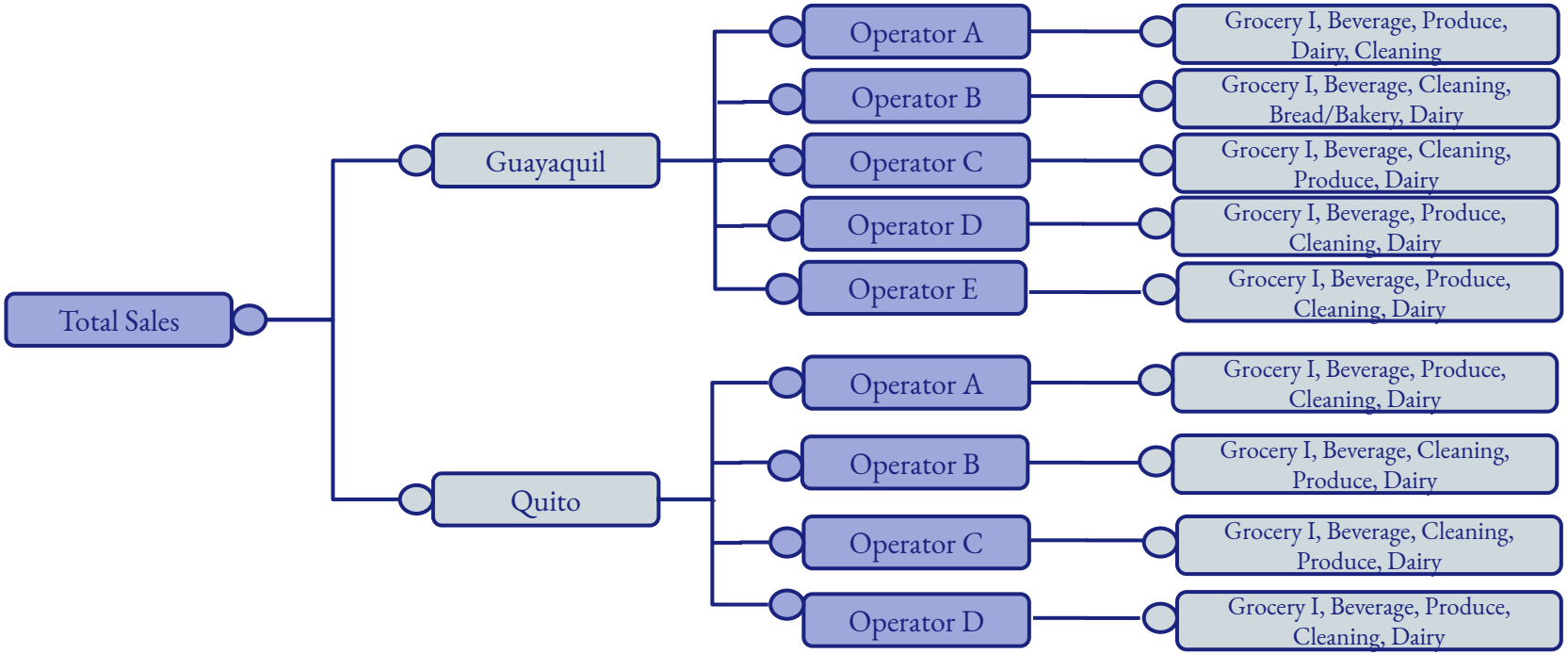
Regressed Auto.Arima on error using Label Encoding on Predictors

SMAPE on Validation Set: 52.8%





Data Structure



LEVEL 1 LEVEL 2 LEVEL 3 LEVEL 4

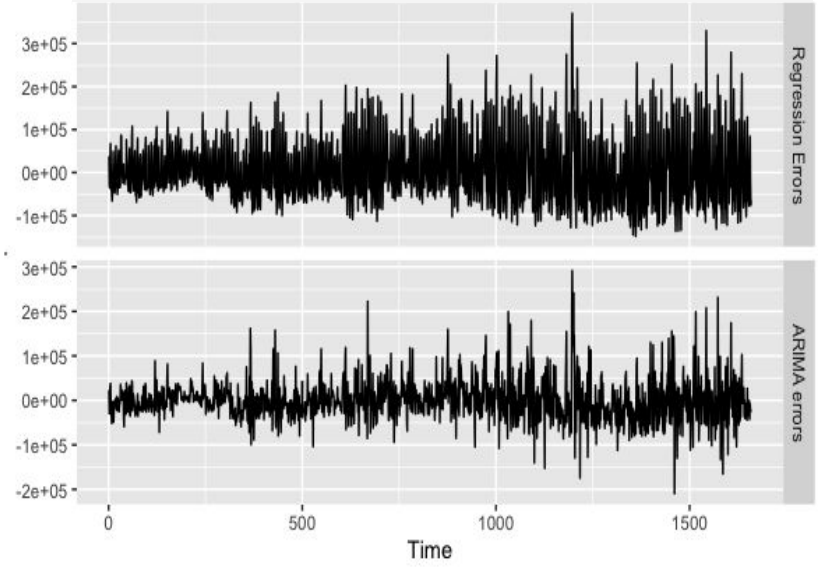
Top-Down Linear Regression with ARIMA Errors

Predictors: No. of Perishable Items, No. of Promotional Items. Oil Price, Fourier Series (K = 5)

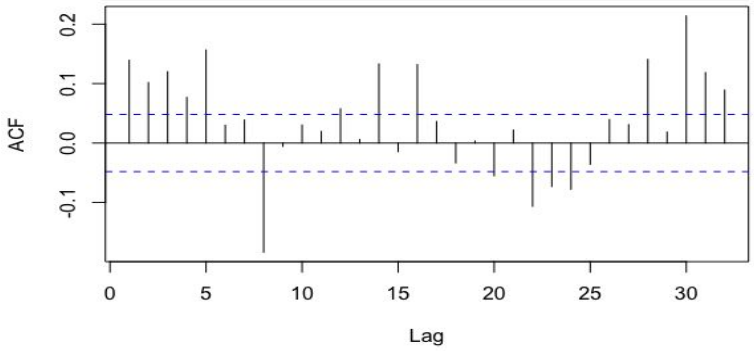
Response: Total Sales across City, Store Operator and Item Family

SMAPE on Validation Set: 4.3%

Comparing ARIMA Residuals against Linear Regression



Residuals from Regression with ARIMA (4, 0, 4) Errors



Top-Down Linear Regression with ARIMA Errors

- Calculated Mean Proportion of Total Sales across the different levels on Train
- Distributed Forecasted Total Sales using the mean proportions across the different levels for Validation

Average SMAPE on Validation Set: 15.3%

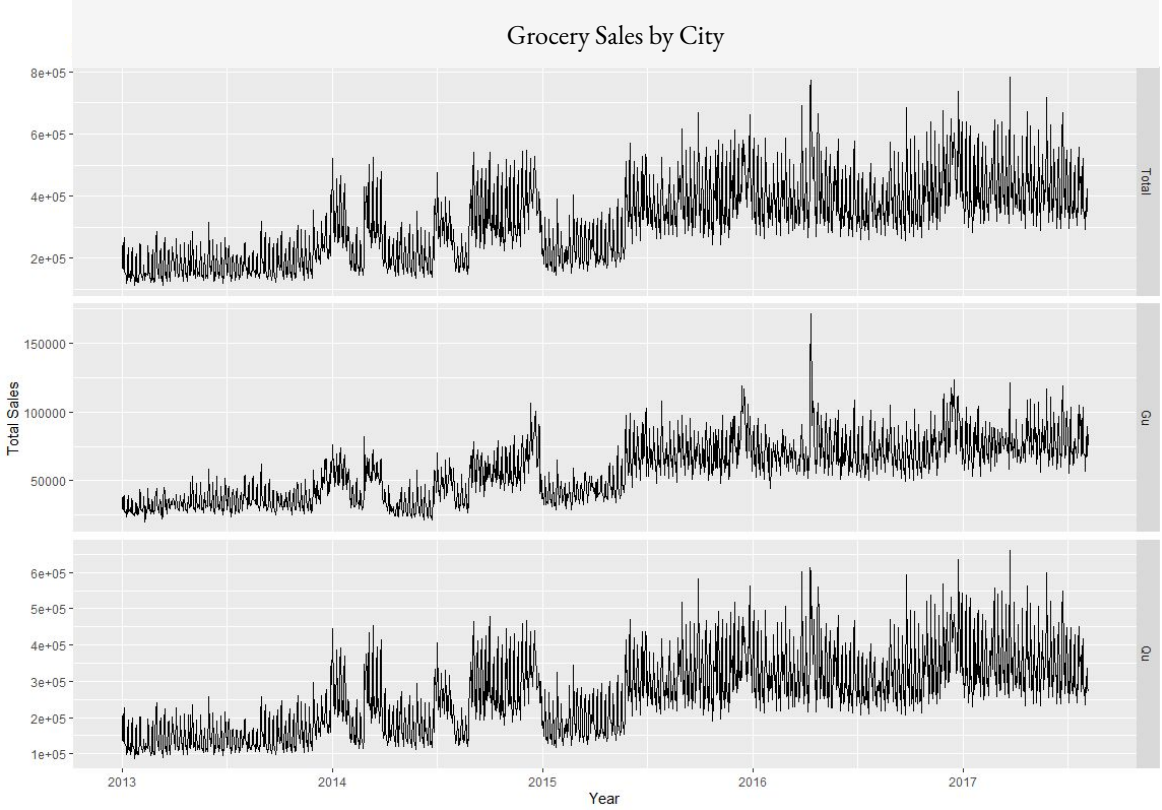
```
 [1] 0.09166888 0.20321652 0.12609089 0.15701901 0.20175805 0.07260083 0.19104399 0.19185711
 [9] 0.10687177 0.12958659 0.19877073 0.23521864 0.09182104 0.11762186 0.25092210 0.06686534
[17] 0.22098269 0.08267449 0.16756755 0.18936793 0.27638371 0.12162177 0.14647787 0.16639709
[25] 0.25705076 0.06921131 0.16086333 0.12037439 0.10591263 0.18274679 0.17980091 0.14747265
[33] 0.06392430 0.07067011 0.27561308 0.06741700 0.16158767 0.09192593 0.09535339 0.27233461
[41] 0.05400647 0.22101729 0.16731320 0.16776608 0.16046413
```

Hierarchical Time Series - Level 1

The data is set up in a 3-level hierarchy with cities (Quito and Guayaquil) at the first level.

The top plot shows total item sales for both cities (for all store types and item families)

The plots below show the total sales disaggregated by city - level 1.

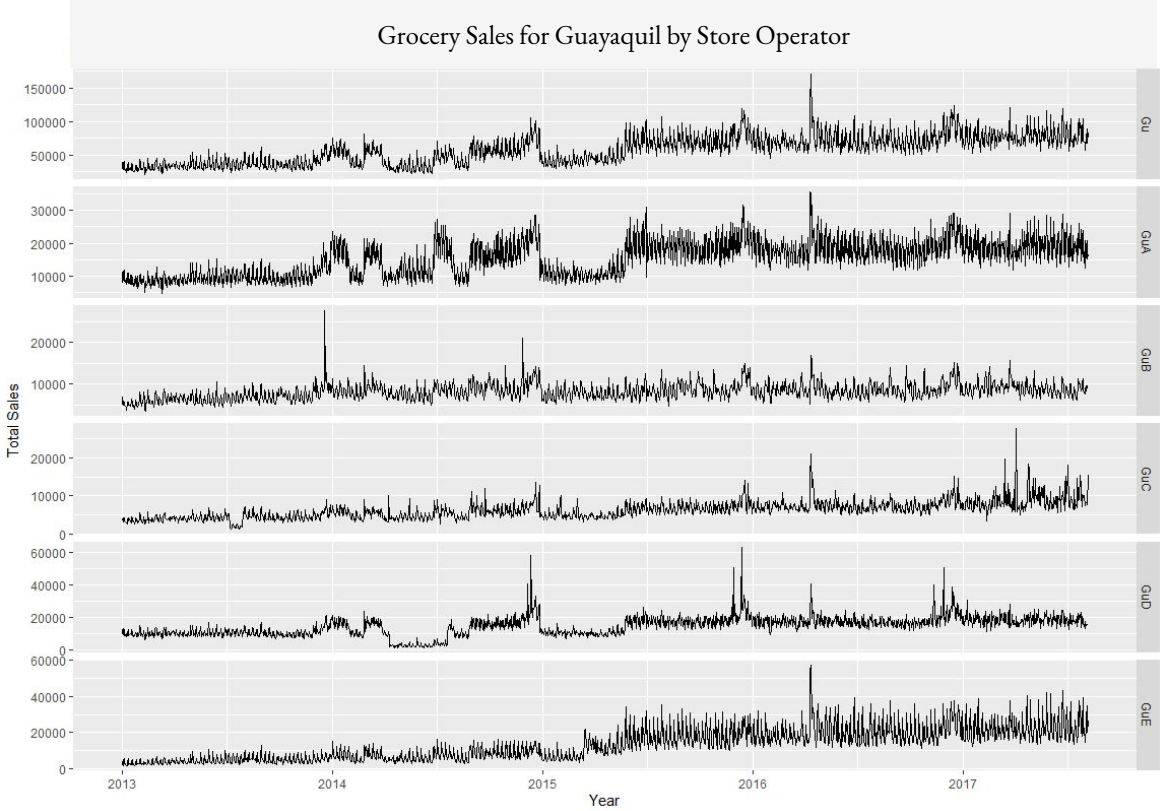


Hierarchical Time Series - Level 2

Type of store (5 types for Guayaquil and 4 types for Quito) make up the second level of the hierarchical time series.

The top plot shows the total sales for the city of Guayaquil.

The plots below show Guayaquil's sales disaggregated by store type - level 2.

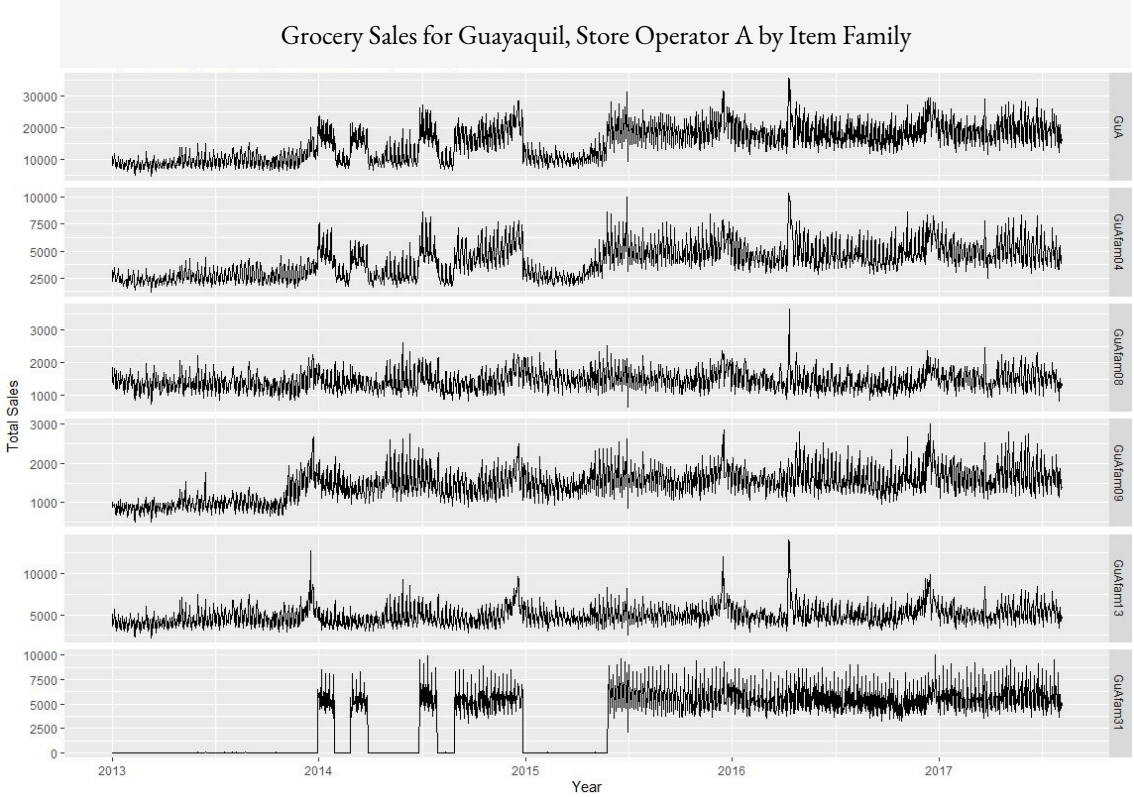


Hierarchical Time Series - Level 3

Item family (5 families for each type of store per city) make up the third level/ bottom level of the hierarchical time series.

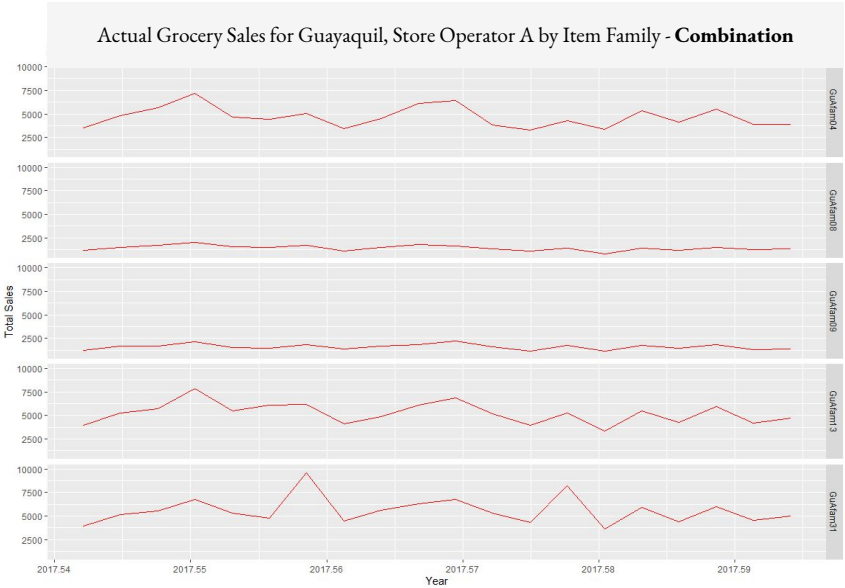
The top plot shows the total sales for store type A in Guayaquil.

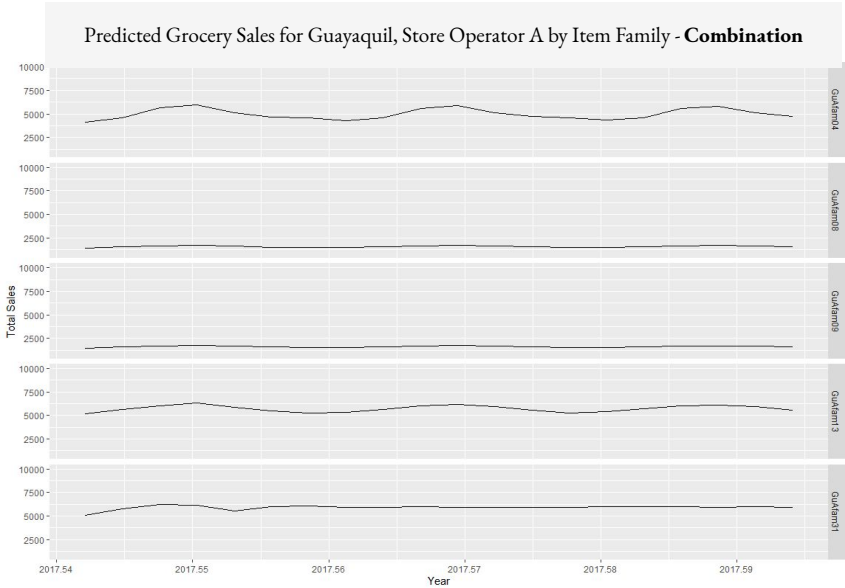
The plots below show Guayaquil's type A stores sales disaggregated by item family - the bottom level.



Hierarchical Time Series - Forecast with ARIMA

Approach	Top Down	Bottom Up	Middle Out	Combination
SMAPE	10.1 %	9.91 %	9.96 %	9.84 %

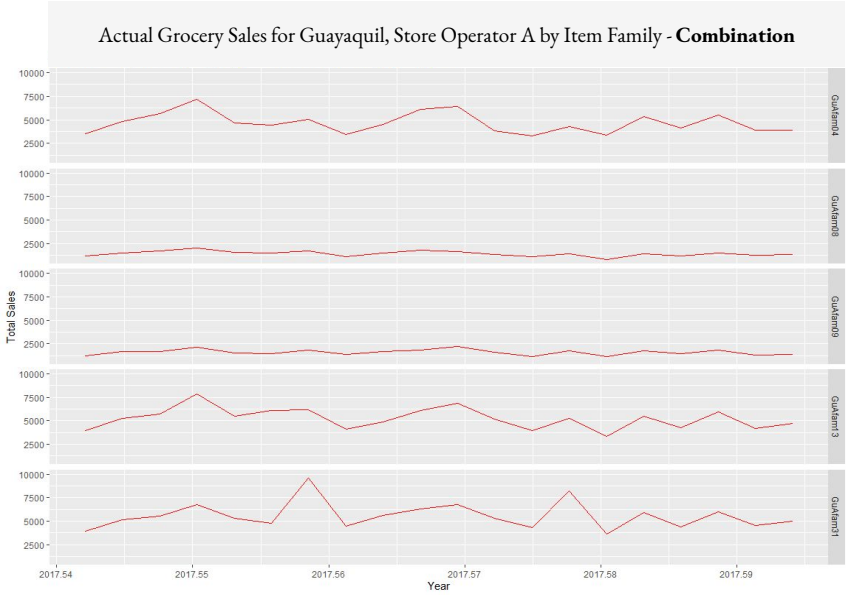


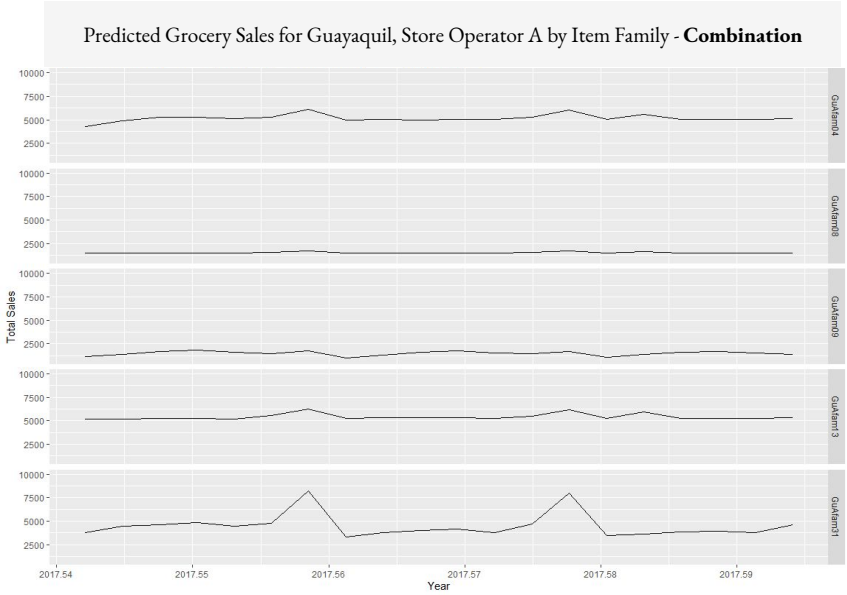


Hierarchical Time Series - ARIMA with Errors

Predictors: No. of Perishable Items, No. of Promotional Items. Oil Price, Flag indicating Holiday, Flag indicating if the holiday was national, Flag indicating if the date was a pay day

SMAPE on Validation Set: 8.73%

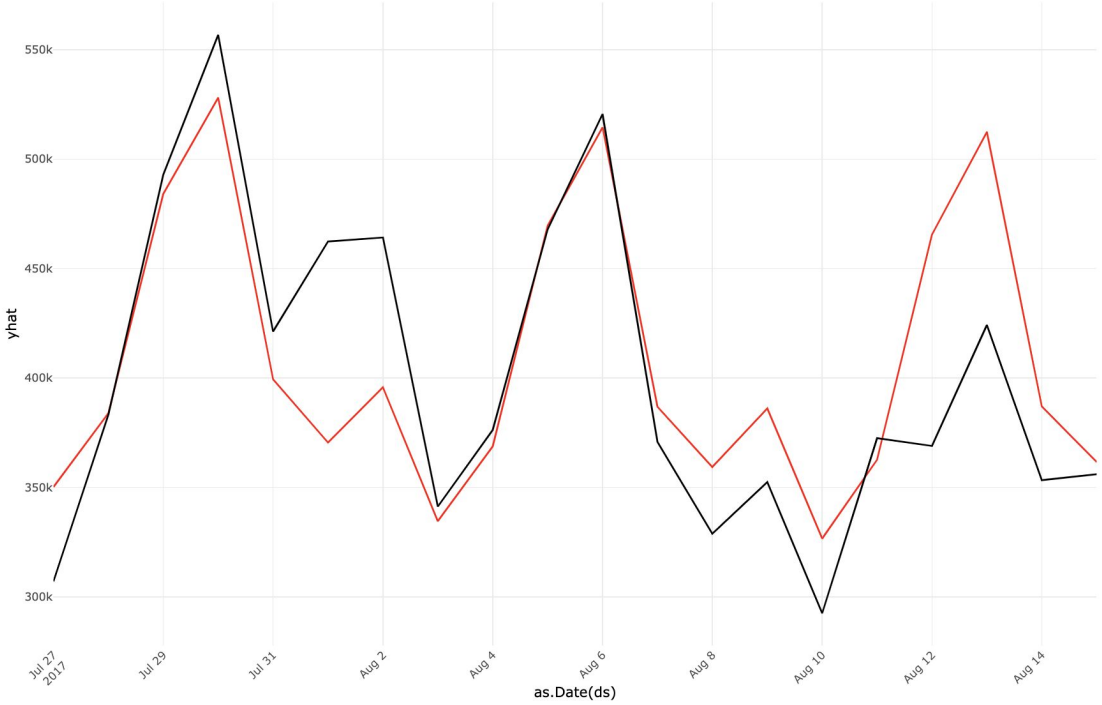




Top-Down Prophet

- Transform data to “ds” and “y” columns before using Prophet
- Forecast on Total Sales with no other parameters set

SMAPE on Validation Set: 3.94%

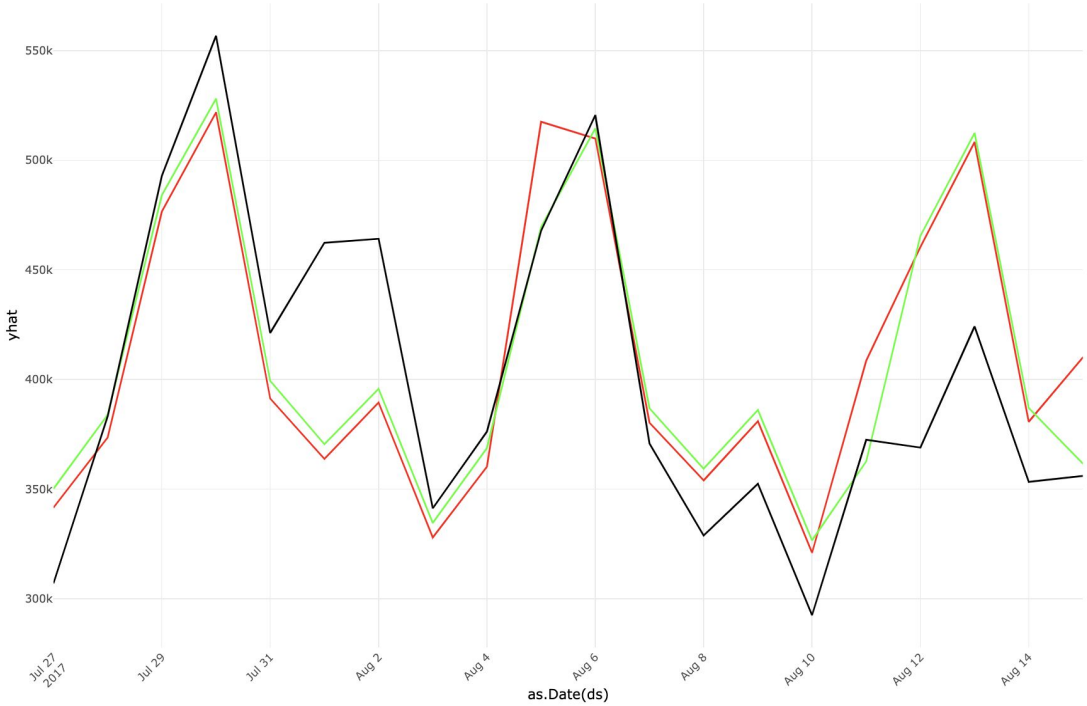


Top-Down Prophet - Add Holidays Parameter

- Constructed a holidays table based on holiday flags in train
- The model did slightly worse
- Removed holidays parameter

SMAPE on Validation Set: 4.75%

Forecast on Total Sales Data without and with Holiday Flags

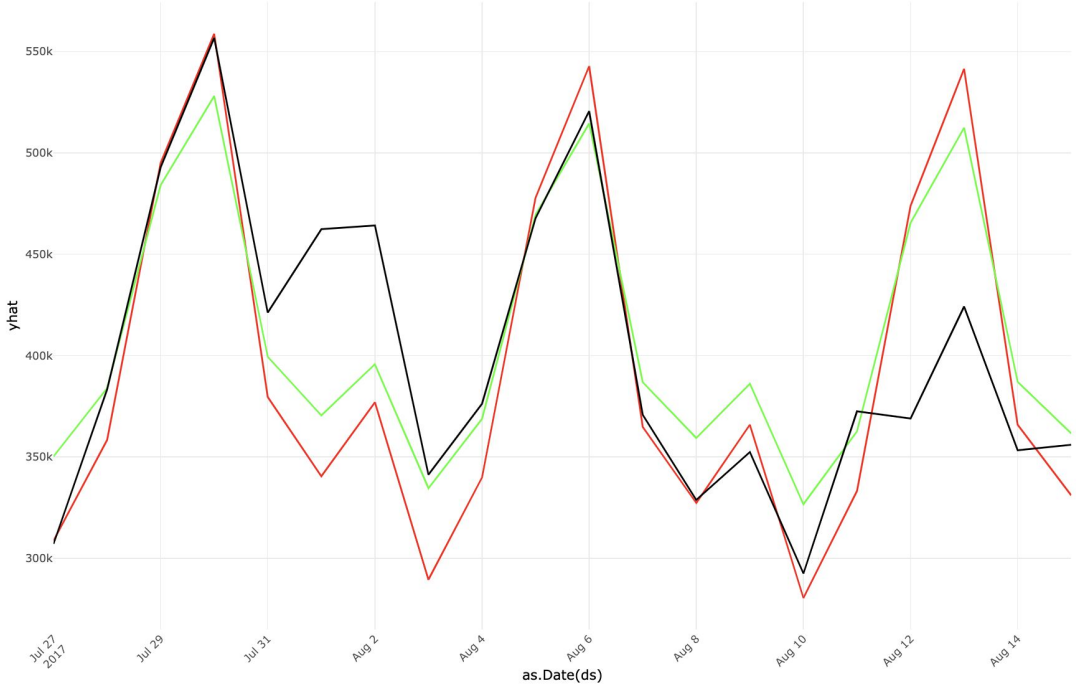


Top-Down Prophet - Change Multiplicative Seasonality

- The model did slightly worse
- Changed Seasonality Mode back

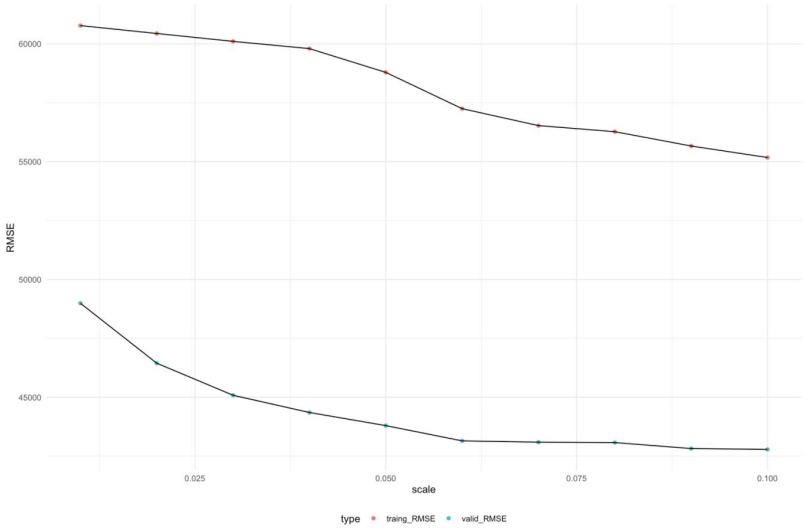
SMAPE on Validation Set: 4.58%

Forecast on Total Sales Data without and with Multiplicative Seasonality Mode



Top-Down Prophet - Change Point Scale

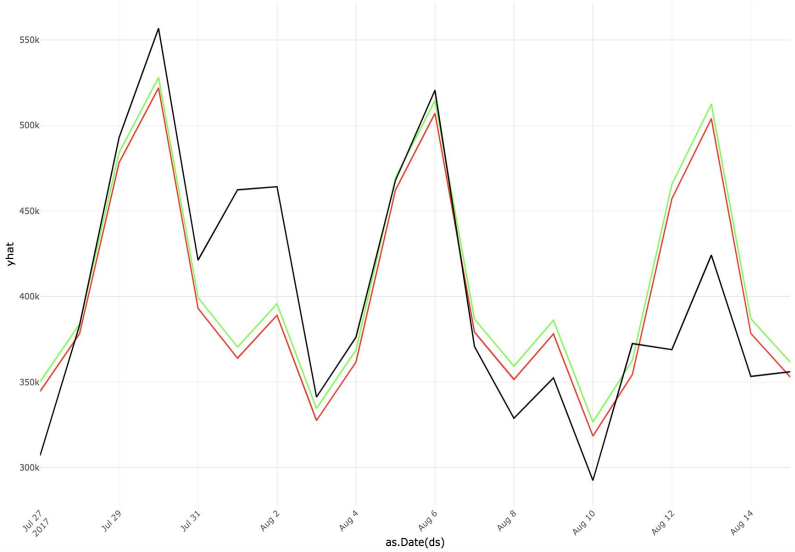
- Changed point scale to values between 0.01 and 0.1 and select the one with lowest RMSE (0.1)



SMAPE on Validation Set: 3.95%

SMAPE slightly worse than model with no parameters set

Forecast on Total Sales Data without and with Changed Point Scale



Top-Down Prophet - Proportion Breakdown

- Calculated Mean Proportion of Total Sales across the different levels on Train
- Distributed Forecasted Total Sales using the mean proportions across the different levels for Validation

Average SMAPE on Validation Set: 15.2%

```
 [1] 0.08885496 0.19897418 0.12165526 0.15269297 0.20605011 0.07857375
 [7] 0.18643999 0.18803863 0.10202856 0.12502798 0.20308374 0.23297403
[13] 0.08535838 0.11851130 0.25516702 0.06264928 0.21655253 0.07784995
[19] 0.16235584 0.19365675 0.28074664 0.11977547 0.14784604 0.17075848
[25] 0.26156002 0.06840504 0.15582465 0.11592195 0.10072611 0.18716478
[31] 0.18419962 0.14289839 0.06402038 0.07606199 0.27997027 0.07326833
[37] 0.15642466 0.09072632 0.09280434 0.27663777 0.04932582 0.21665183
[43] 0.16282448 0.16328499 0.16485811
```

Evaluation & Prediction

Model Performance

	Base Model	Top-Down Linear Regression w Fourier and Xreg	Top-Down Prophet	Combination Hierarchical with Xreg
SMAPE	15.2 %	15.3 %	15.2 %	8.73 %
Pros	Simple and easy to implement	Predicts the top-level very well with SMAPE of 4.3 %	Predicts the top-level very well with SMAPE of 3.9 %	Predicted each level with the same SMAPE
Cons	High SMAPE	Sabotaged at the bottom-level as SMAPE of 15.3%	High SMAPE on lower levels	Long Run Time

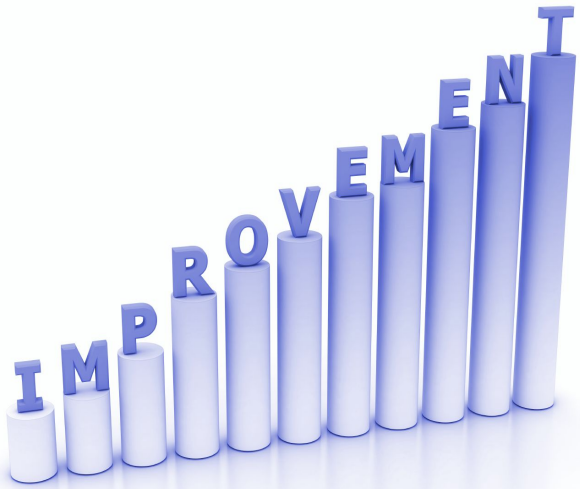
Future Work

- **For Linear Regression** - Use one item family's sales as Y, and all other families' sales as X

- **Use combination methods** instead of top-down for both linear regression and prophet

- **Non-linear regression** approaches such as Deep learning neural networks -
    - CNN - faster feature engineering
    - LSTM - robust to noise and has a multivariate output

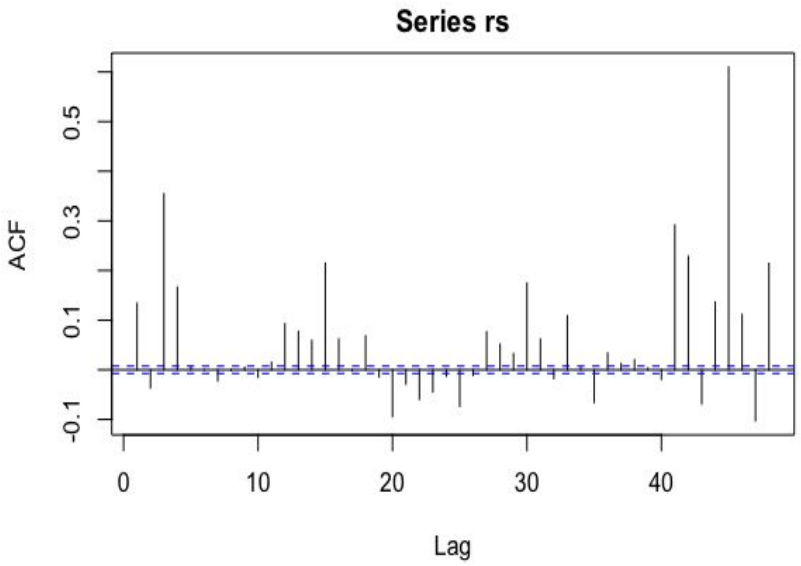
- **For Hierarchical Combination** - Expand to include all stores and items

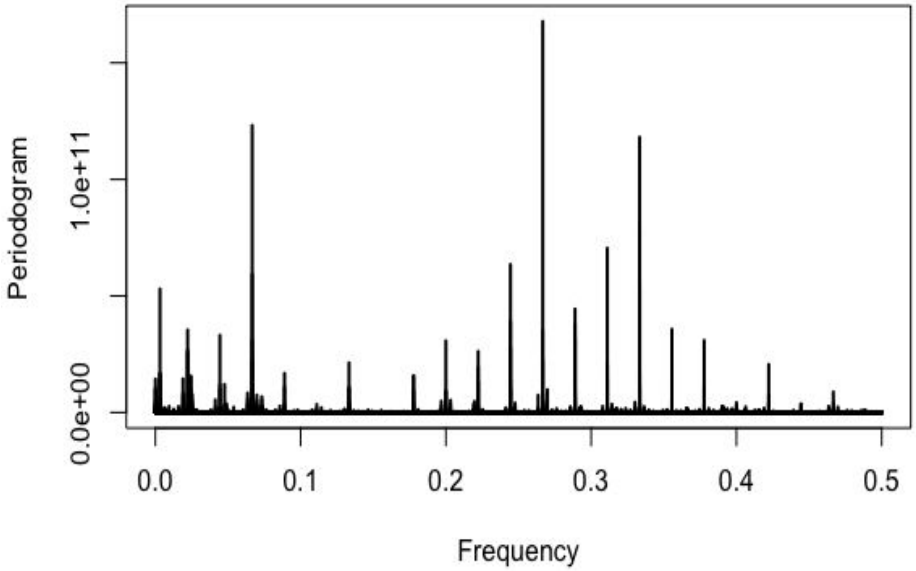


Thank You.

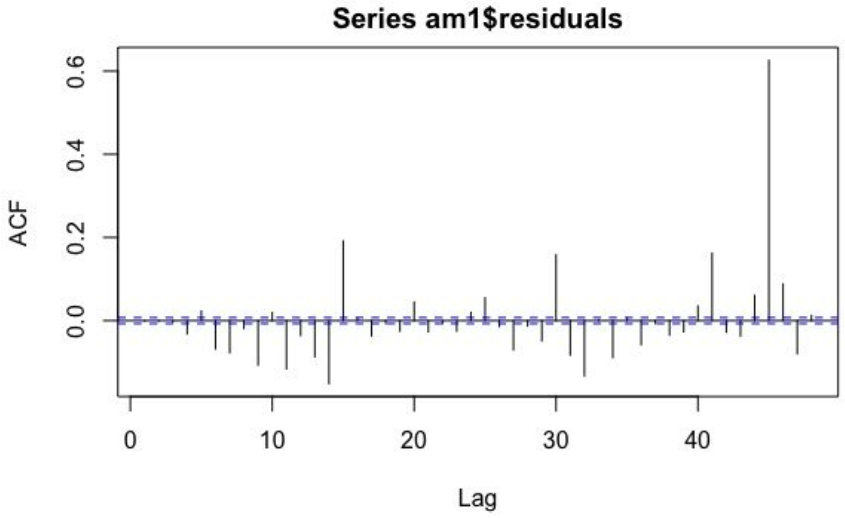
Appendix

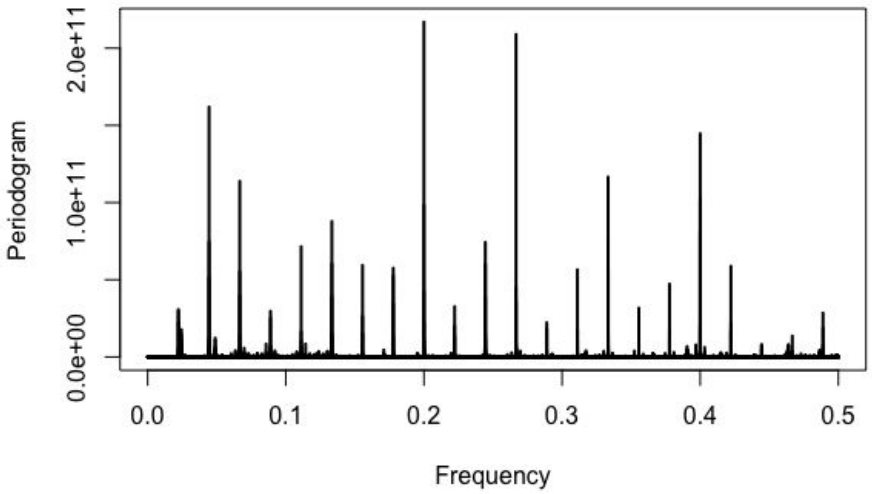
Linear Regression Residuals:



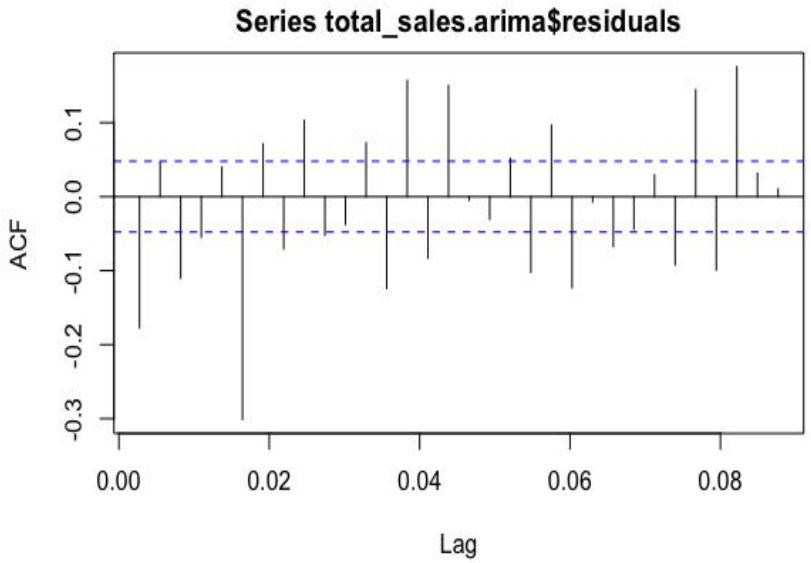


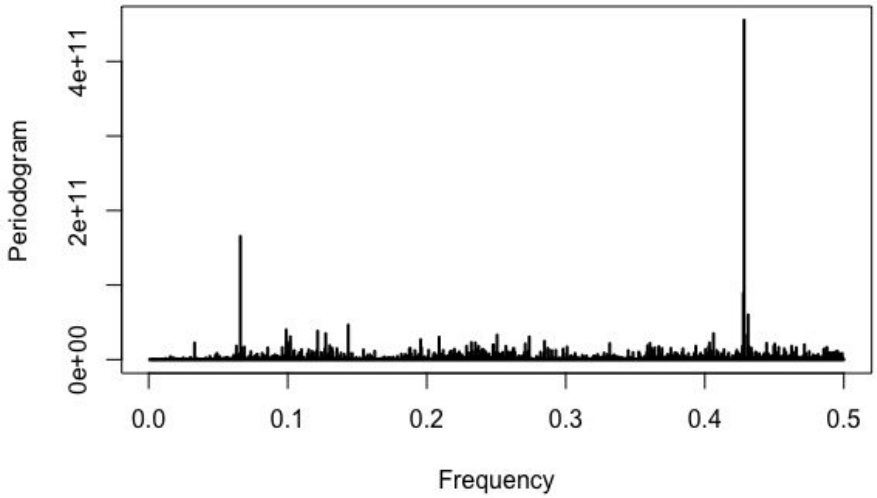
Linear Regression with Auto.Arima Fixed Error



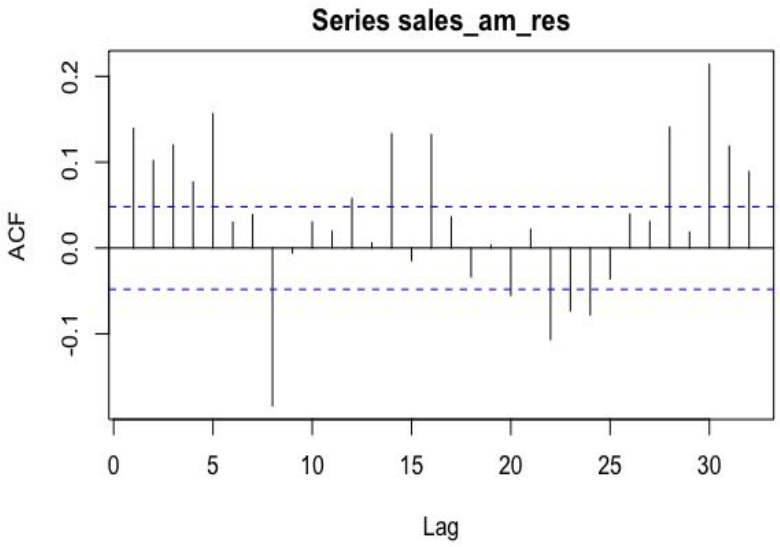


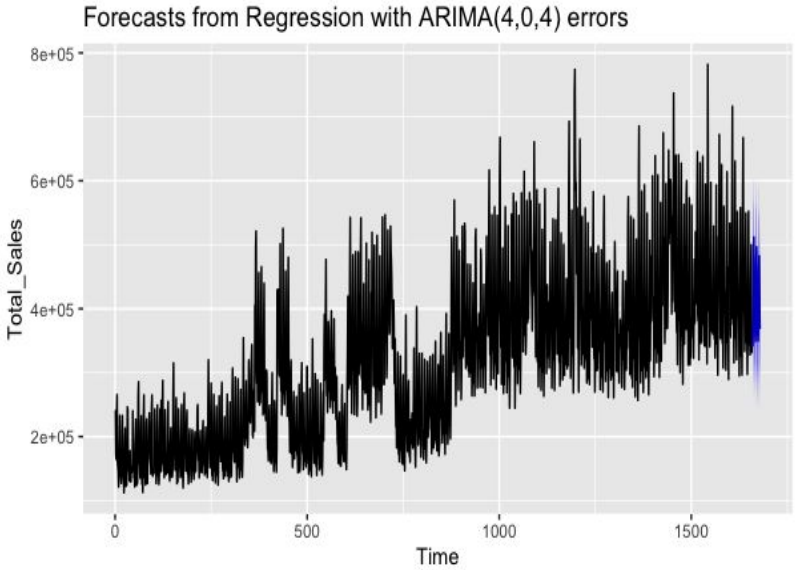
Linear Regression by Total Sales in a day





Linear Regression by Total Sales in a day with Auto.Arima Fixed Error





Hierarchical Time Series - ARIMA with errors variations

Predictors: Oil Price	SMAPE on Validation Set: 9.53%
Predictors: Oil Price, Flag indicating Holiday, Flag indicating if the holiday was national	SMAPE on Validation Set: 9.71%