STORE SALES FORECASTING

Agenda

- Problem Statement
- Approach
- Data Exploration
- Feature Creation
- Modeling
- Outcome
- Next Steps

Problem Statement

What would be the daily sales for 2 weeks ahead by store and product family?

What is the short-term impact of promotions on sales?

Help the store manager plan for stock replenishment and minimize loss of perishable products

What is in the data

- Daily unit sales data of **33 product families** across **54 stores** of Favorita in Ecuador
- Daily **oil prices**
- **Number of products on promotion** for a given product family
- **Holiday**/events metadata
- **Store attributes** – city, state, store type, store cluster
- Major **earthquake in April'16**

Data Source: Grocery Store Sales, Kaggle

Approach

- **Exploratory Data Analysis**
  - Understand data granularity and distribution by visualizing the data
  - Identify irregularities and motive of feature creation – remove some irregularities, or control them by adding dummy variables

- **Feature Engineering**
  - Potential causal relationships, such as increased sales of a product in nearby stores when it is out of stock in other store
  - Attributes of related products may boost sales of other products
  - Use events data intelligently to not include all the available events while modeling

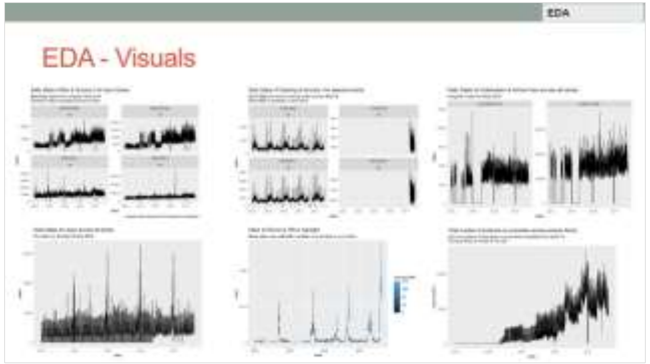
- **Evaluation**
  - Time series cross validation (rolling forecast) to evaluate performance of a model
  - Root mean squared log error (RMSLE) metric to measure goodness of a model

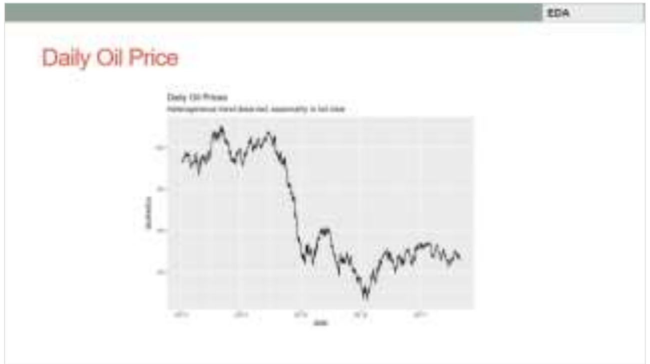
- **Modeling technique**
  - Seasonal ARIMA with regressors
  - Structural additive decomposing with stl function to deseasonalize, and then use auto.arima to fit non-seasonal ARIMA model
  - Additional modeling techniques such as DeepAR* to be explored

*DeepAR: Probabilistic Forecasting with Autoregressive Recurrent Networks, by Salinas et. al, 2017, Amazon Research

Data Exploration - summary

- The train data includes **4.6 years of daily sales data** for **33 product families** sold in **54 stores**, totaling 1,782 combinations
- **31% of the sales records** in the training data **are zero**, not necessarily no sales, can be missing data
- Sales information **abnormally low** for almost every store, product on **1st January**, possible **data collection issue**
- **Top 4 stores** (store number 44, 45, 47, 3) account for **20% of the overall sales***
- **Top 2 product families** (GROCERY I and BEVERAGES) account **for 50% of overall sales***
- Some products had **irregular or zero sales in initial period** of the data
- **Earthquake** in April 2016 had a significant **impact on sales**, positive for some, negative for some
- **Promotion** information available **April 2014 onward**
- **100 unique event** descriptions available, too many for a single model, may be irrelevant for some products



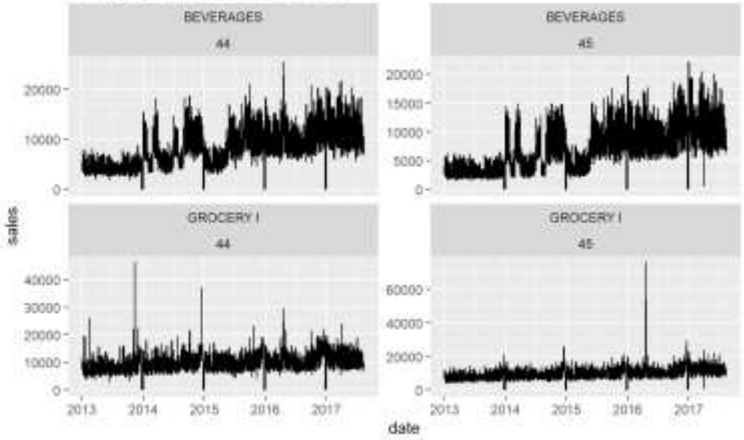


*Unit sales across product families are not summable, but there is no other information available to identify ‘important’ stores, or products

EDA - Visuals

Daily Sales of Bev & Grocery-I for top-2 stores

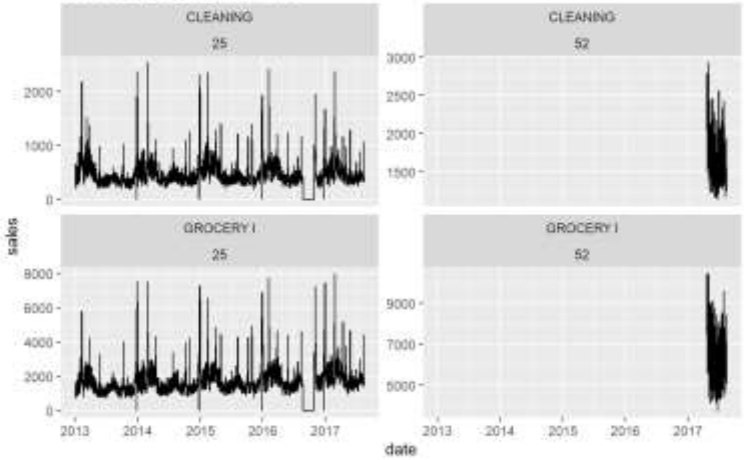
Beverage sales has irregular level shift
Grocery sales increases around Xmas



Irregular spike during April'16 because of earthquake

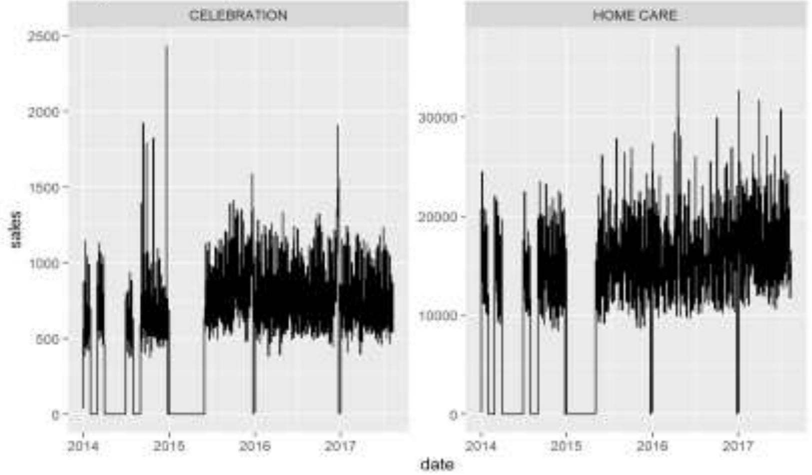
Daily Sales of Cleaning & Grocery-I for selected stores

Store #25 has some missing sales around Sep'16
Store #52 is possibly a new store



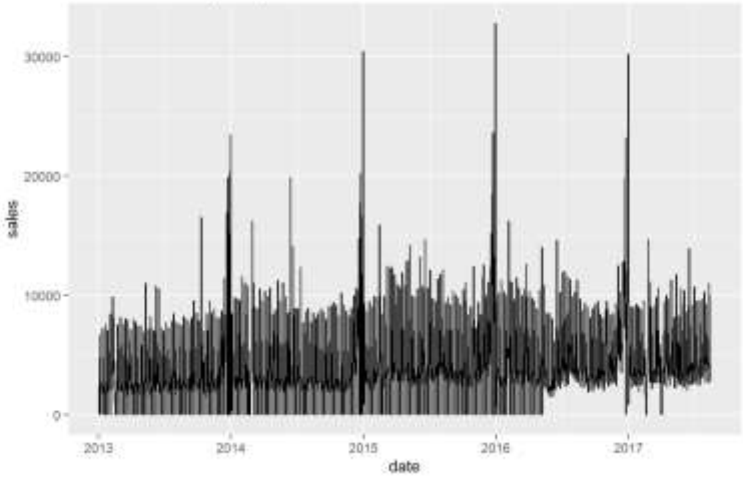
Daily Sales of Celebration & Home Care across all stores

Irregular sales till May 2015



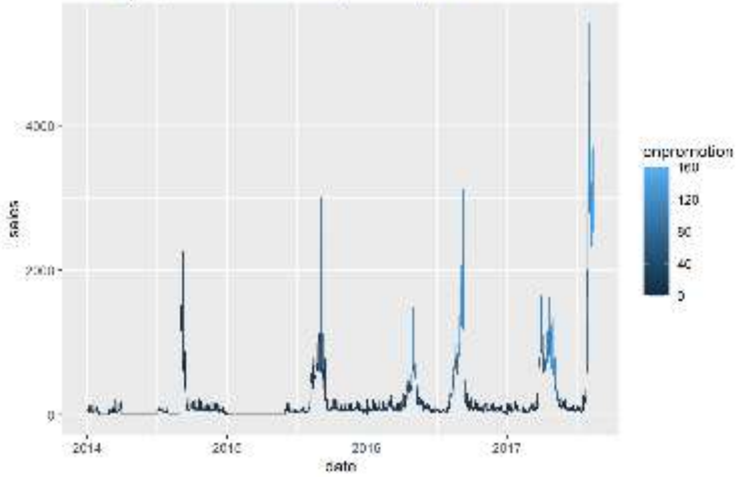
Daily Sales of Liquor across all stores

No sales on Sunday till May 2016



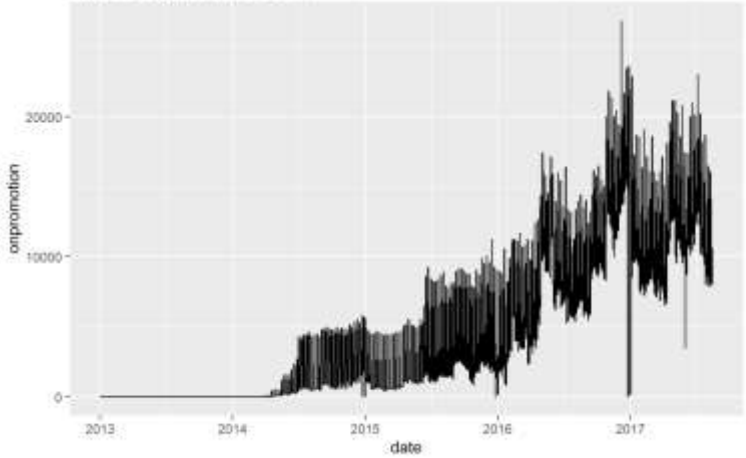
Sales of School & Office Supplies

Sales align very well with numbers of products on promotion



Total number of products on promotion across product family

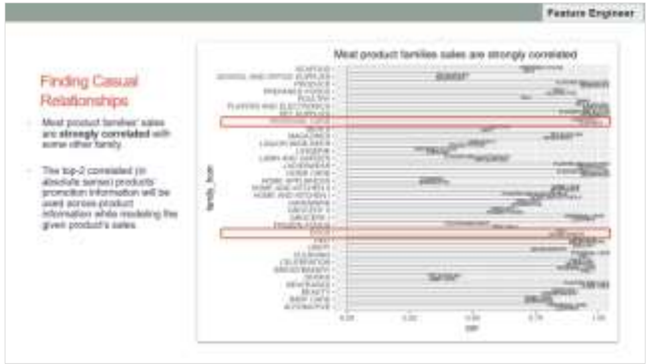
Info on numbers of products on promotion available from April'14
Unusual drop on Xmas & 1st Jan



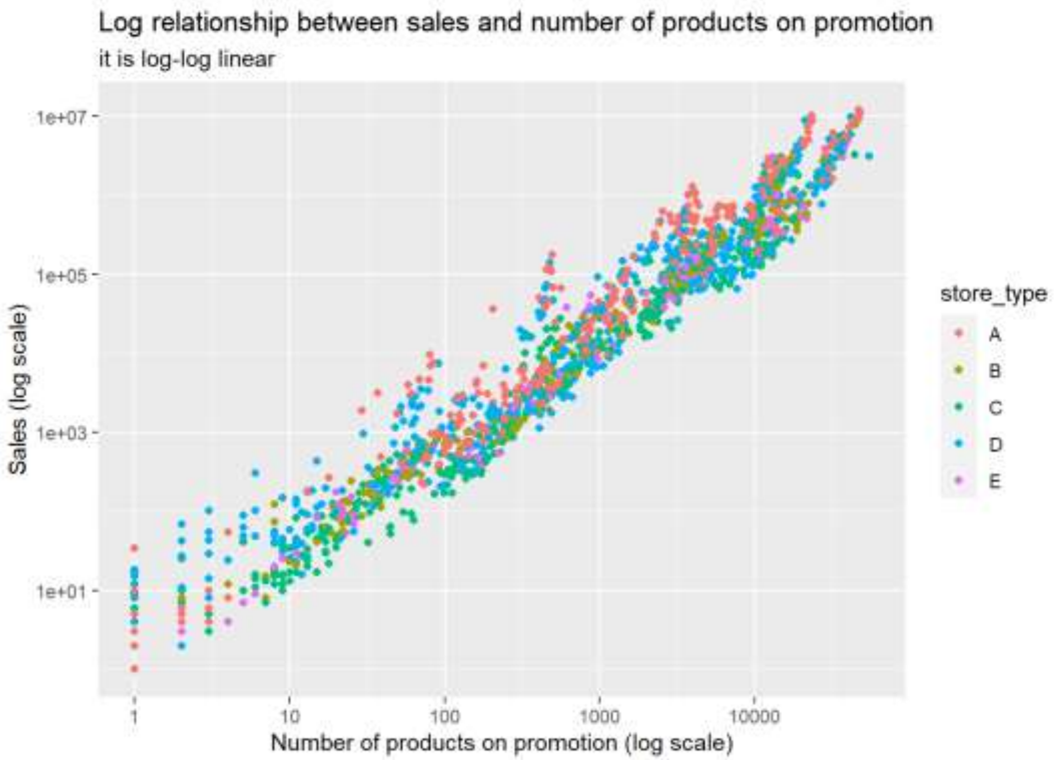
Daily Oil Price

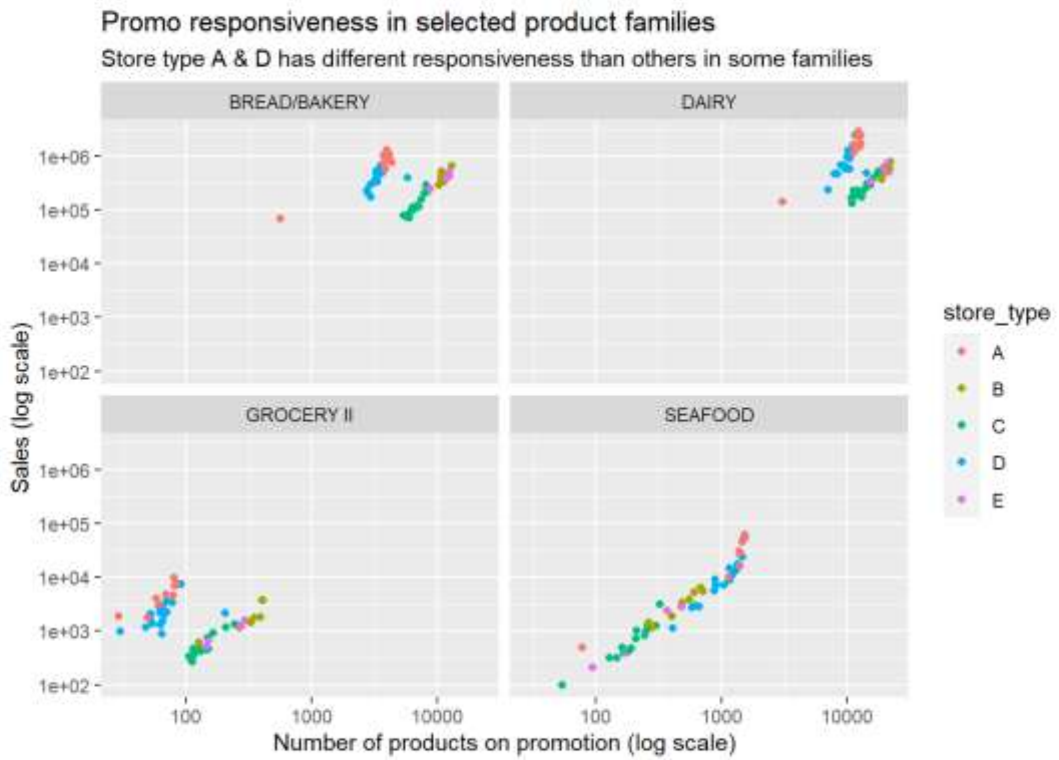
Feature Engineering

- **Truncate irregular/0** sales during **initial periods** for each product
- Dummy variable to **flag 1st January** of every year
- Dummy variable to **flag Sunday for Liquor family** till 8-May-2016
- **Ratio** of **current store promotion** with the **average** number of promotions in the **same state-city** where the store belongs to
- **Top-2 correlated products' promotion** information as cross-product information while modeling the **given product's sales**
  - For example, BABY CARE has top-2 correlated products as HOME CARE & BEVERAGES; test promotion information of HOME CARE & BEVERAGES while modeling the sales of BABY CARE
- **Algorithmically extract important events** in terms of high (or low) sales compared to the sales when there were no events
  - Test only important events relevant for given model using one-hot encoding approach



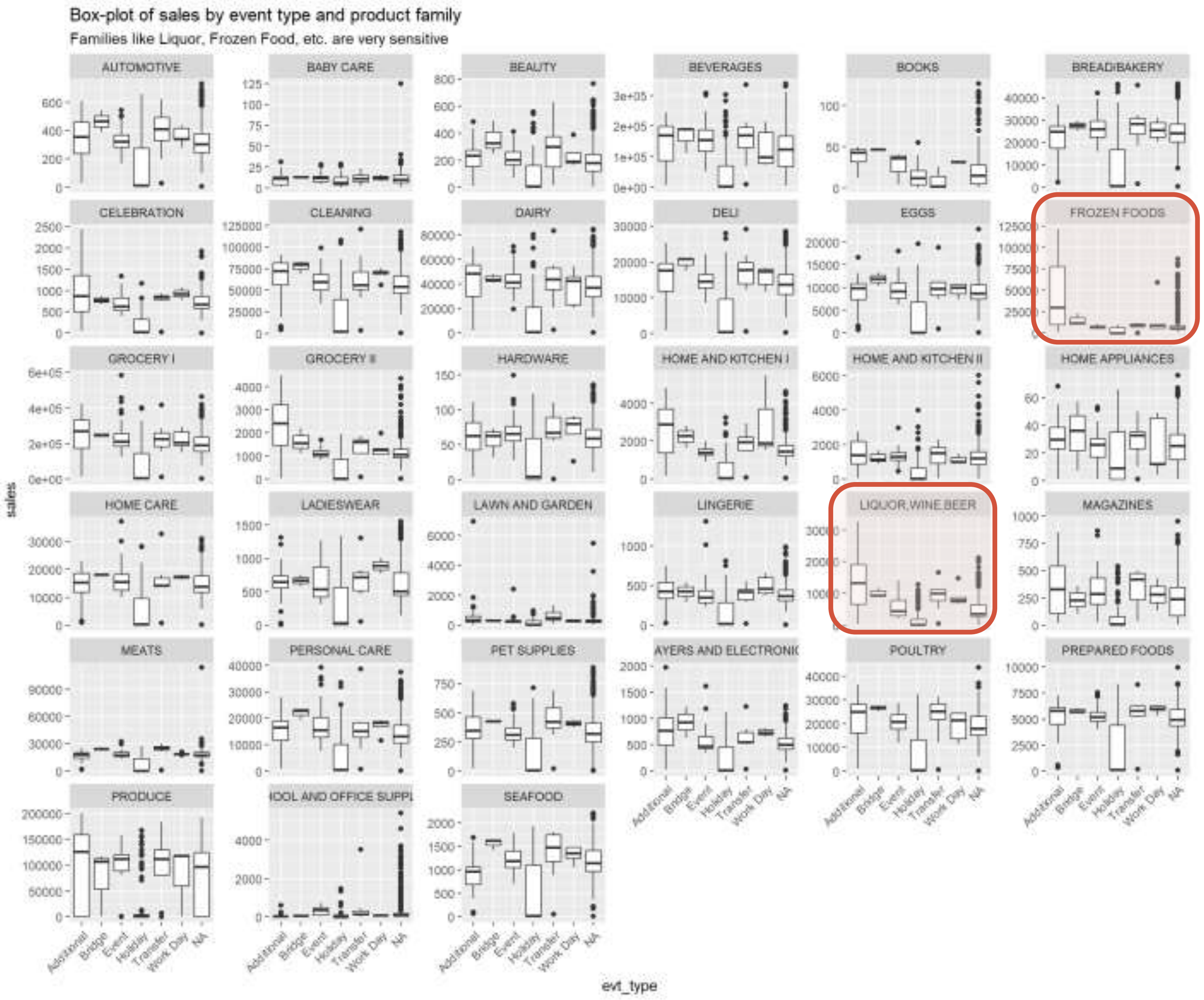
Promotion vs Sales





Sales distribution by event type

- ‘Holiday’ type of event includes 1st January
- Some product families are very sensitive to holidays/ events

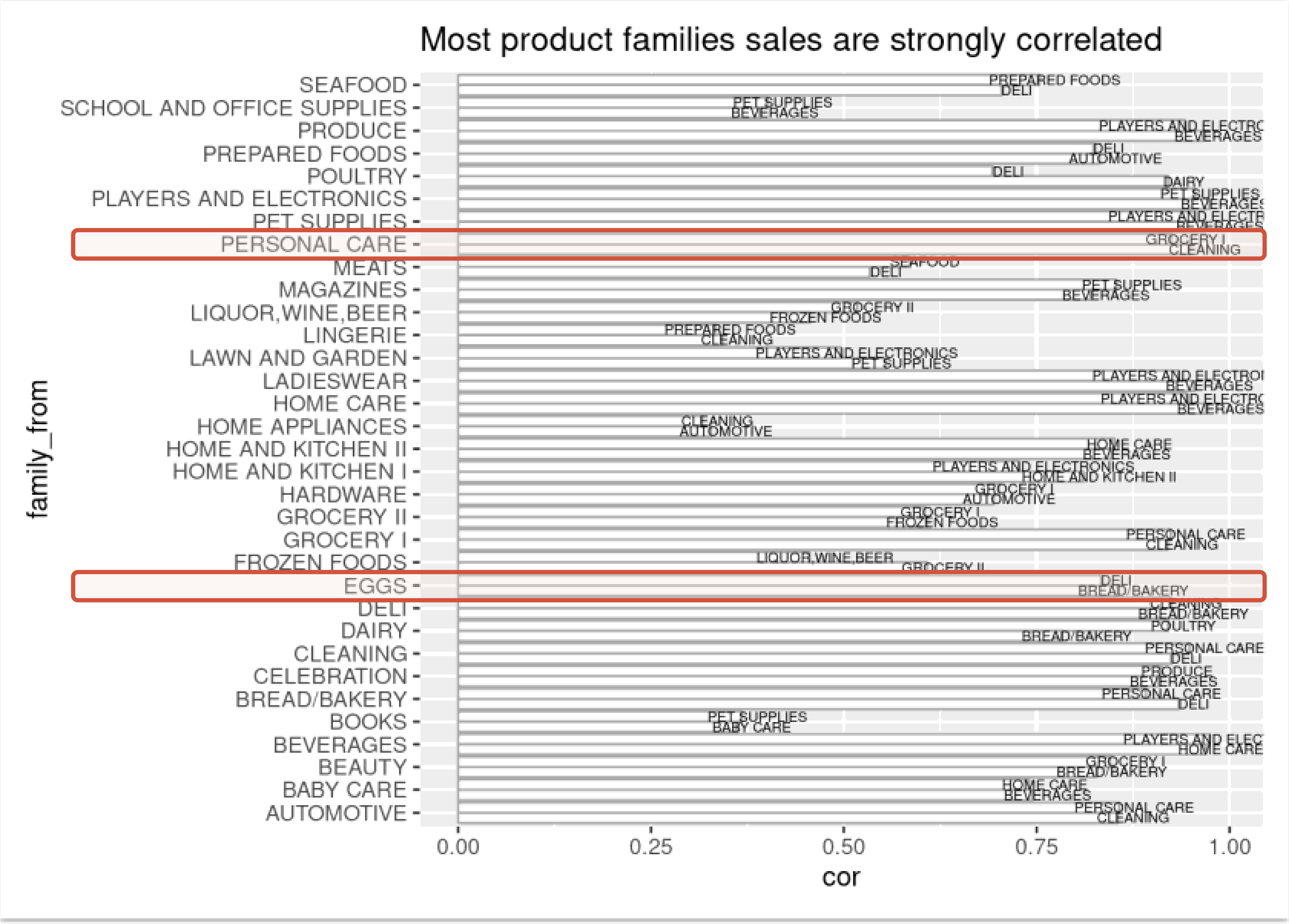


*NA implies no event on that date

Finding Casual Relationships

- Most product families' sales are **strongly correlated** with some other family.

- The top-2 correlated (in absolute sense) products' promotion information will be used across-product information while modeling the given product's sales.

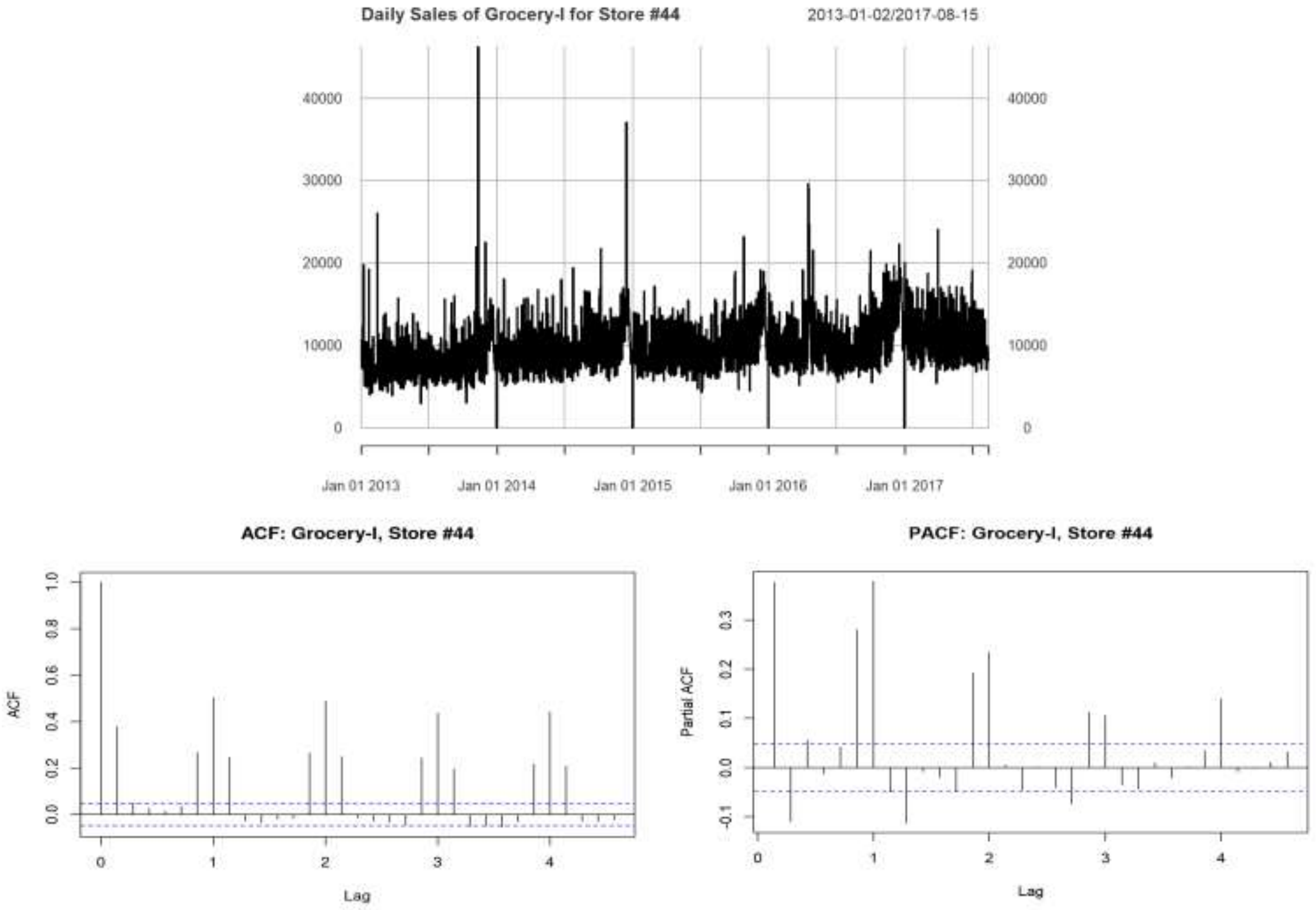


Modeling

Focusing on the **Grocery-I product family for the top grossing store (#44)** to start with a univariate time-series analysis.

Some observations

- There is **noticeable trend**, we may have to difference the data.
- It is **7 period seasonal** as the ACF spikes in the gap of 7 lags.
- **ACF** hints a seasonal differencing might also be necessary as **decay rate** is **not fast** enough.
- **ACF** also suggests we may have to use some **MA order 2**
- **PACF decayed faster than ACF**, possibly there is no need of seasonal AR, only a regular AR will be sufficient.



Modeling flow of biggest store-family

Sales without transformation

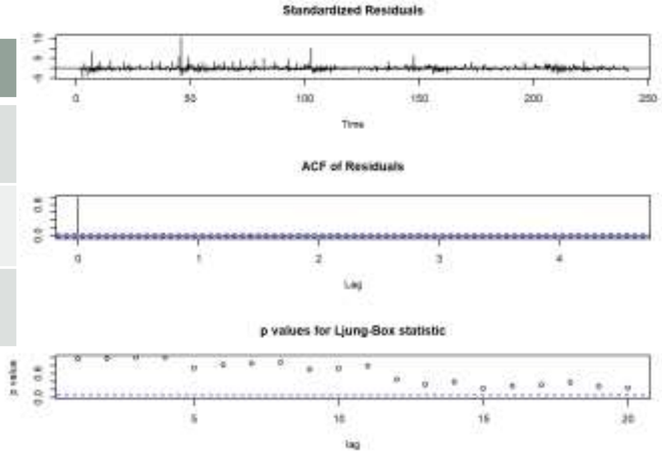
- Initially built univariate model on sales
- Started with auto.arima
- Tuned the order to get clean residuals
- Added all the chosen xreg

With transformation

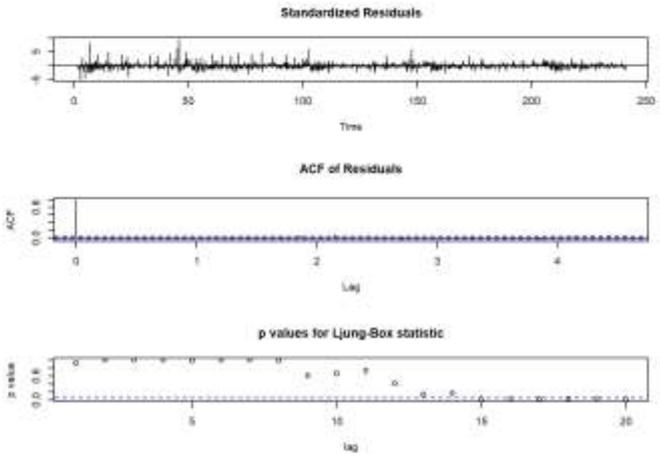
- EDA showed log-log relation between promotion & sales
- Now fit the same model spec (+xreg) on log1p(sales)

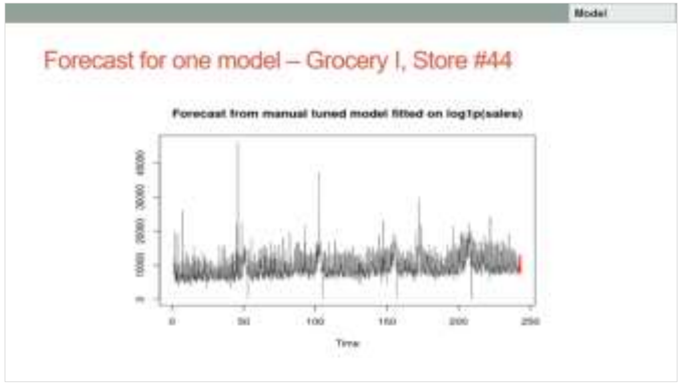
Sales **without transformation**

Model type	Order	AIC
auto.arima (all significant)	(1,1,2) (0,0,2) 7	31,645.7
Manually tuned (all significant)	(1,0,2) (0,1,1) 7	31,236.2
Manually tuned + xreg (~5 insignificant out of ~20)	(1,0,2) (0,1,1) 7	30,995.6

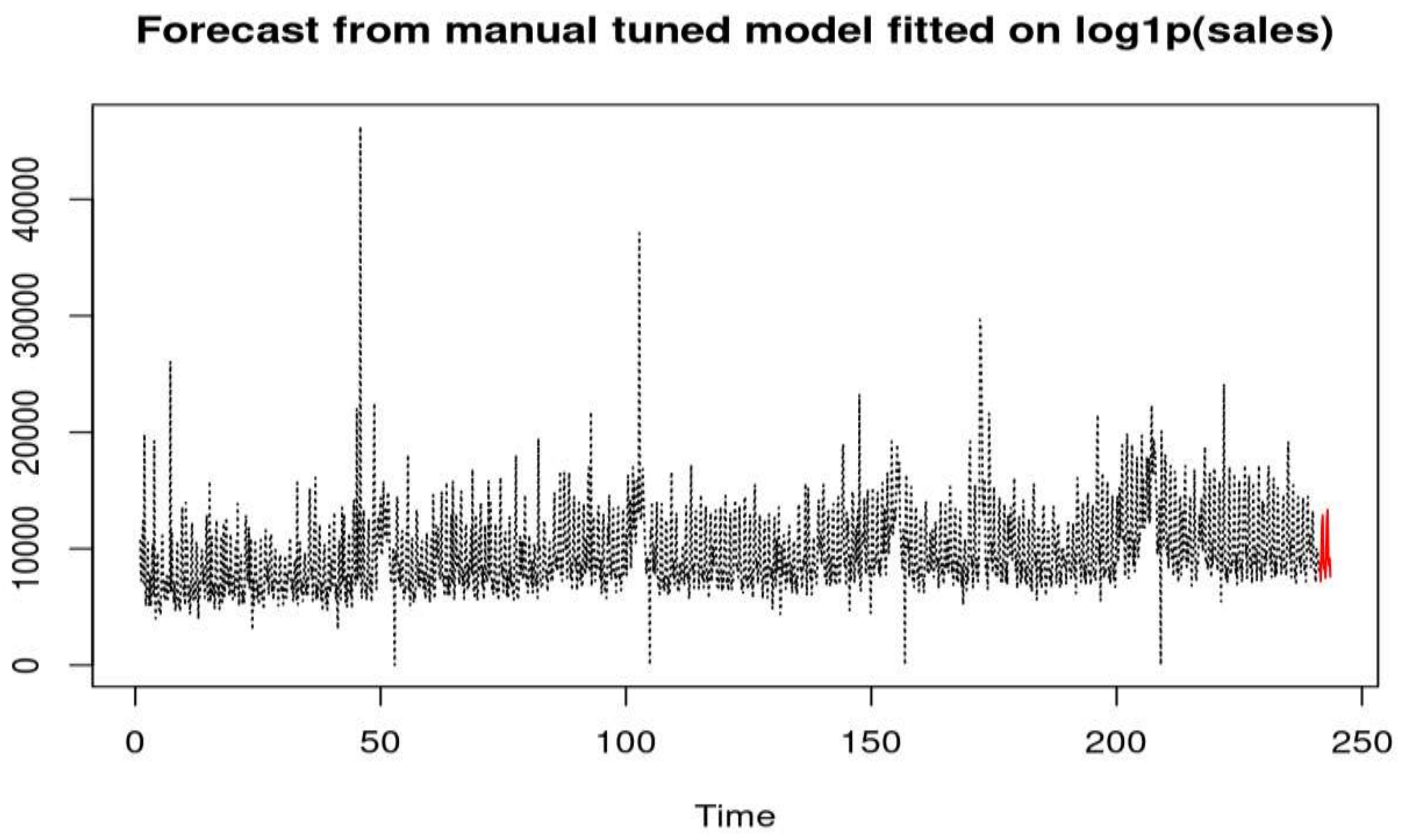


Sales **with log transformation**

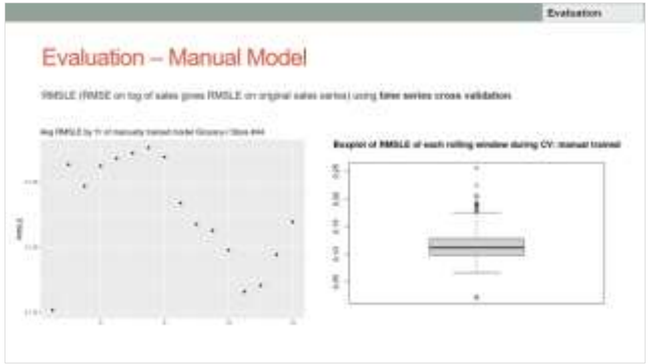


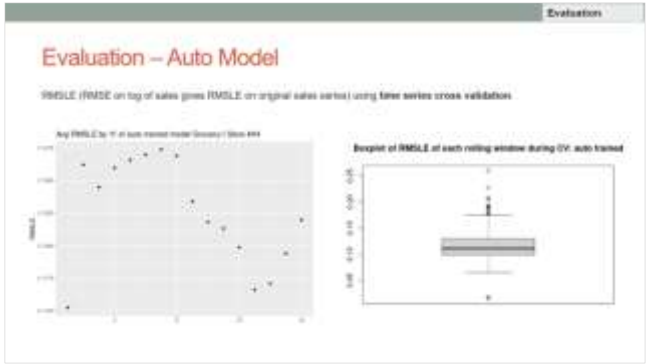


Forecast for one model – Grocery I, Store #44



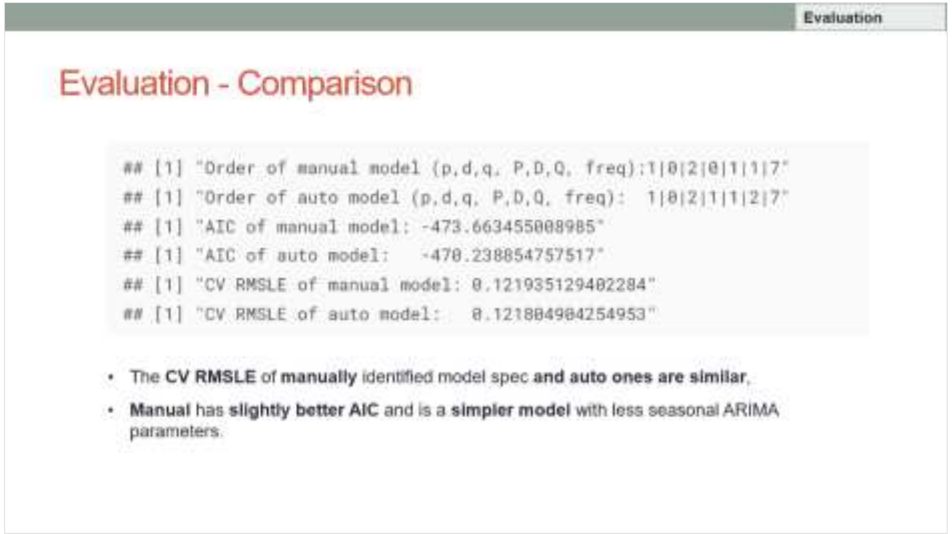
Compare models using cross-validation





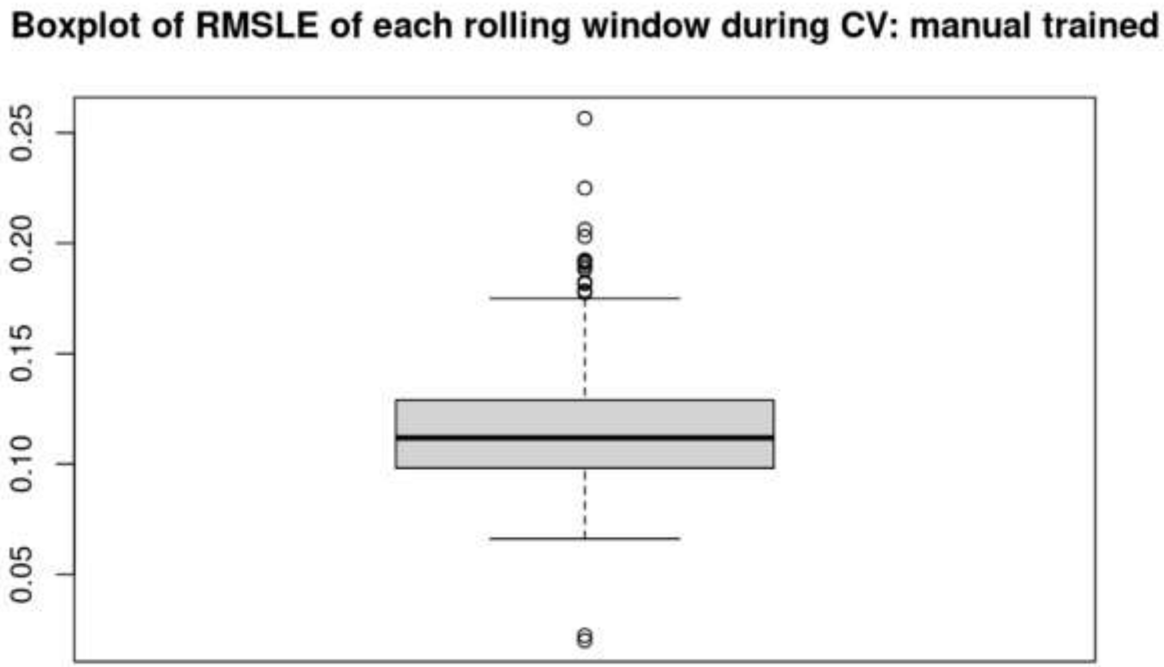
- **Stress testing** of **auto-tuned models** to check their performance

- They seem **comparable with manually** tuned model



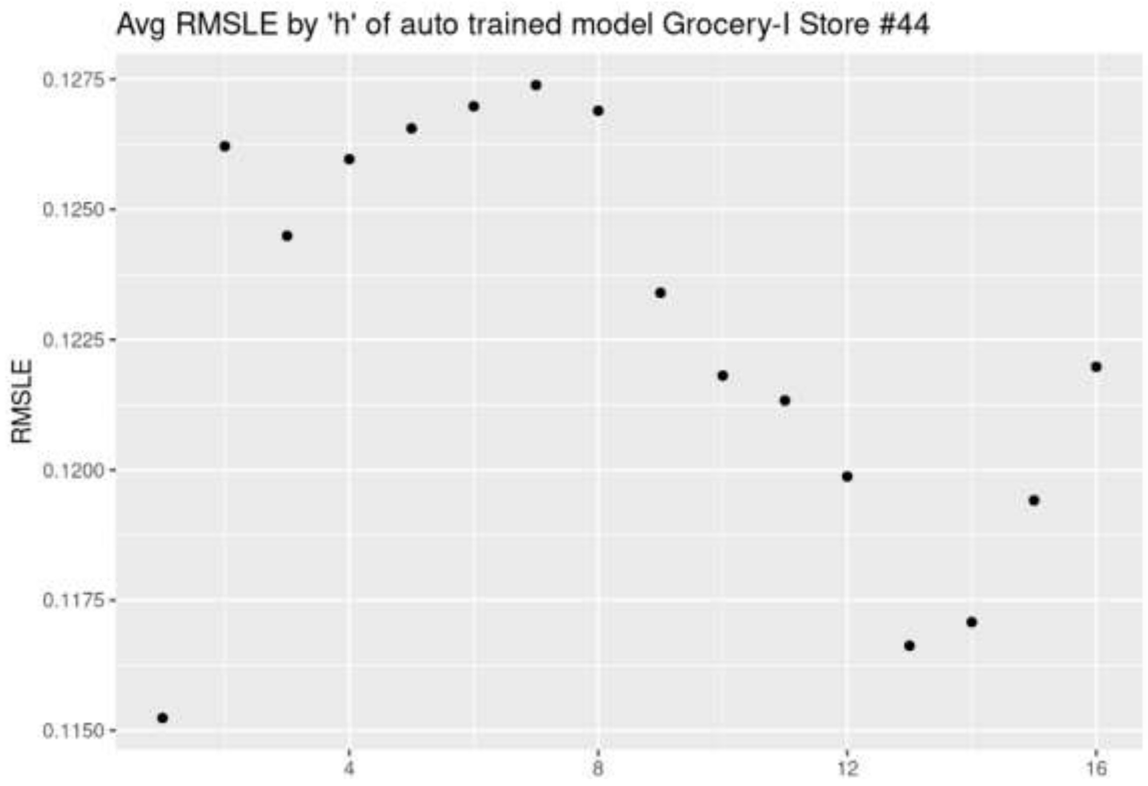
Evaluation – Manual Model

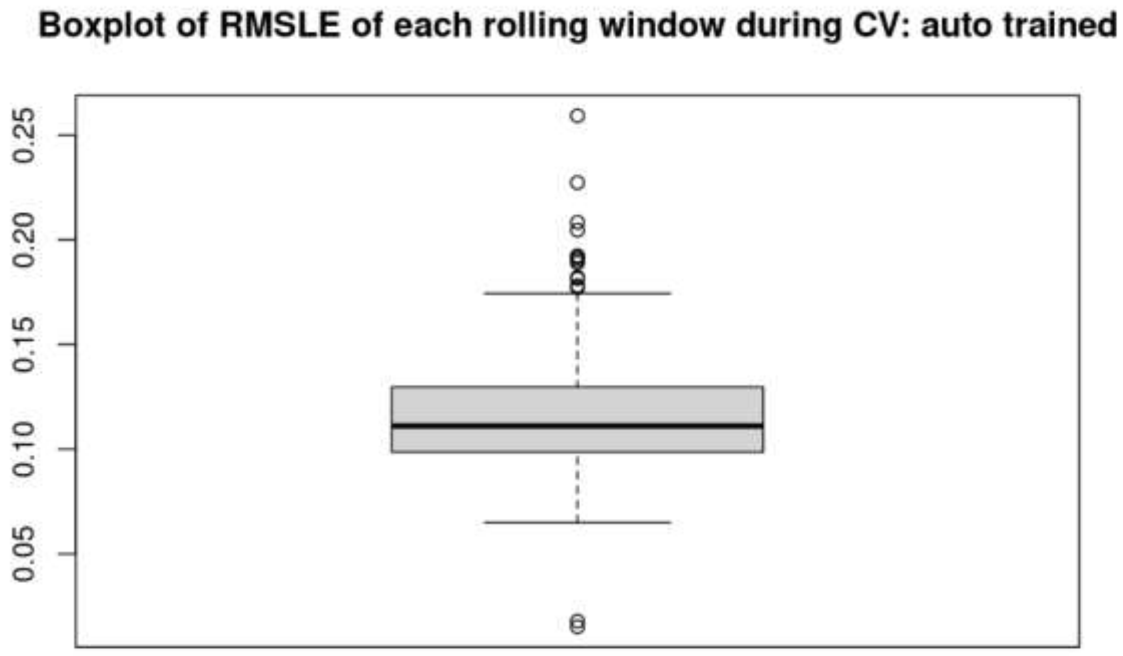
RMSLE (RMSE on log of sales gives RMSLE on original sales series) using **time series cross validation**.



Evaluation – Auto Model

RMSLE (RMSE on log of sales gives RMSLE on original sales series) using **time series cross validation**.





Evaluation - Comparison

```
## [1] "Order of manual model (p,d,q, P,D,Q, freq):1|0|2|0|1|1|7"
## [1] "Order of auto model (p,d,q, P,D,Q, freq):  1|0|2|1|1|2|7"
## [1] "AIC of manual model: -473.663455008985"
## [1] "AIC of auto model:   -470.238854757517"
## [1] "CV RMSLE of manual model: 0.121935129402284"
## [1] "CV RMSLE of auto model:   0.121804904254953"
```

- The **CV RMSLE** of **manually** identified model spec **and auto ones are similar**,
- **Manual** has **slightly better AIC** and is a **simpler model** with less seasonal ARIMA parameters.

Final modeling approach on full data

- It took **significant time** for the cross validation with **pure auto.arima** where seasonal parameters are also estimated
- Try `` `**forecast::stlm**` `` which uses **automated version of `stl`** (forecast::mstl) to find the **optimal `s.window`** and then fits **auto.arima without seasonal parameters**
- **`stlm` resulted CV RMSLE as 0.1240** compared to **pure auto.arima's 0.1218**
- Run time was **~80% less** for Grocery-I store #44 sales data modeling
- `stlm` with ARIMA turns out to be a great time saver with decent model performance
- For simplicity, ignored checking for variable significance
- Resulted in **Kaggle public score** of RMSLE on Kaggle test data as **0.41544**
  - **Ranked at 75 (top 10%)** at the time of submission **out of ~750 submissions**
  - Top score on Kaggle was about 0.37949 at the time of submission.

Next steps

DeepAR implementation on experimental basis

- Major plus point: **single model** as opposed to ~1,780 individual models the auto.arima approach had
- **Run time** for DeepAR for 100 epochs was about **1 hour compared to auto.arima** on full data was about **6.5 hours**
- **RMSLE score 0.54295** on Kaggle test data with DeepAR – **worse than auto.arima**
- It **needs hyperparameter tuning** with cross validation to get better result

Possible improvements

- Apply some **meaningful logic to do imputation of intermittent 0 sales** as there are notable cases like that, which are possibly some data issue
- **Impact of earthquake** on some products that faced sudden increase/decrease in sale before attaining normalcy. Some kind of geometric series could have been created to measure effect of **gradually coming back to normal level**
- **Optimal model selection criteria** in auto.arima could have been **changed to BIC** instead of AIC values as we have ~4.5 years of daily data (~1500 time points)

*__DeepAR__: Probabilistic Forecasting with Autoregressive Recurrent Networks, by Salinas et. al, 2017, Amazon Research