By SZ

Walmart Sales Forecast Using ARIMA and Machine Learning

August 7, 2020

By SZ

WHY RETAIL SALES FORECASTING

- ➤Mass retailing is a high volume / low margin business
- ➤Retailing is a seasonal and cyclical business
- ➤Upfront investments in merchandise and stores
- ➤Fixed costs in store maintenance
- ➤Sales foresting is a business critical factor
  - Overstocking results in markdowns and less profit
  - Under stocking results in loss sales, lower customer satisfactions

DATA PREPARATION

By SZ

Variables	17
Observations	421570
Store	Factor, 45 levels
Dept	Factor, 81 levels
Type	Factor, 3 levels
Size	Interger
Date	Date, format"2010-02-05"
IsHoliday	Logic
Weekly_Sales	Numeric
MarkDown1	Numeric, anonymous
MarkDown 2	Numeric, anonymous
MarkDown 3	Numeric, anonymous
MarkDown 4	Numeric, anonymous
MarkDown 5	Numeric, anonymous
CPI	Numeric
Unemployment	Numeric
Temperature	Numeric
Fuel_Price	Numeric
Week	Numeric

Data background:

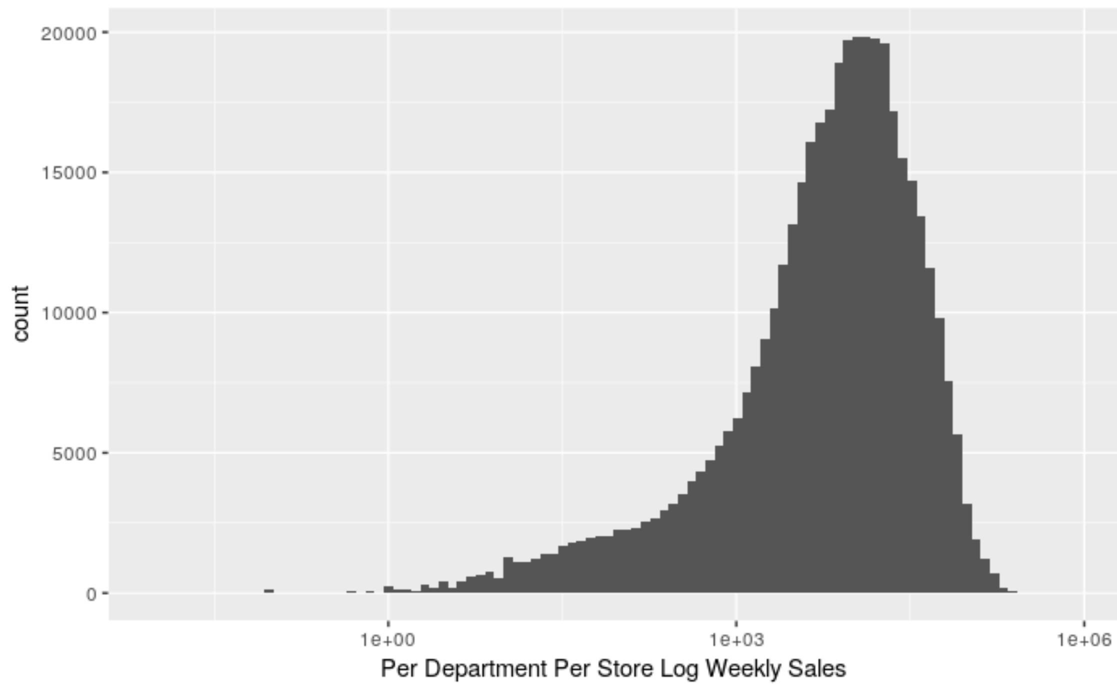
- A Kaggle recruiting competition in 2014
- Test set is withheld
- Time period: weekly sales data from 2010-02-05 to 2012-10-26 (143 weeks)

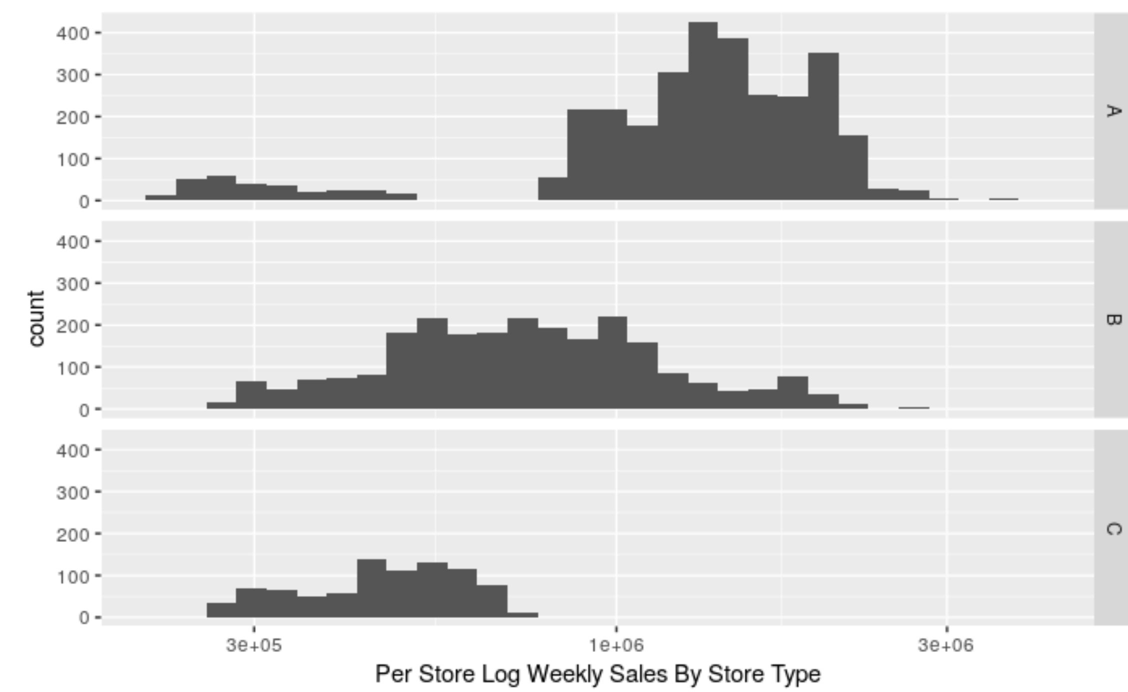
Initial Data Cleaning

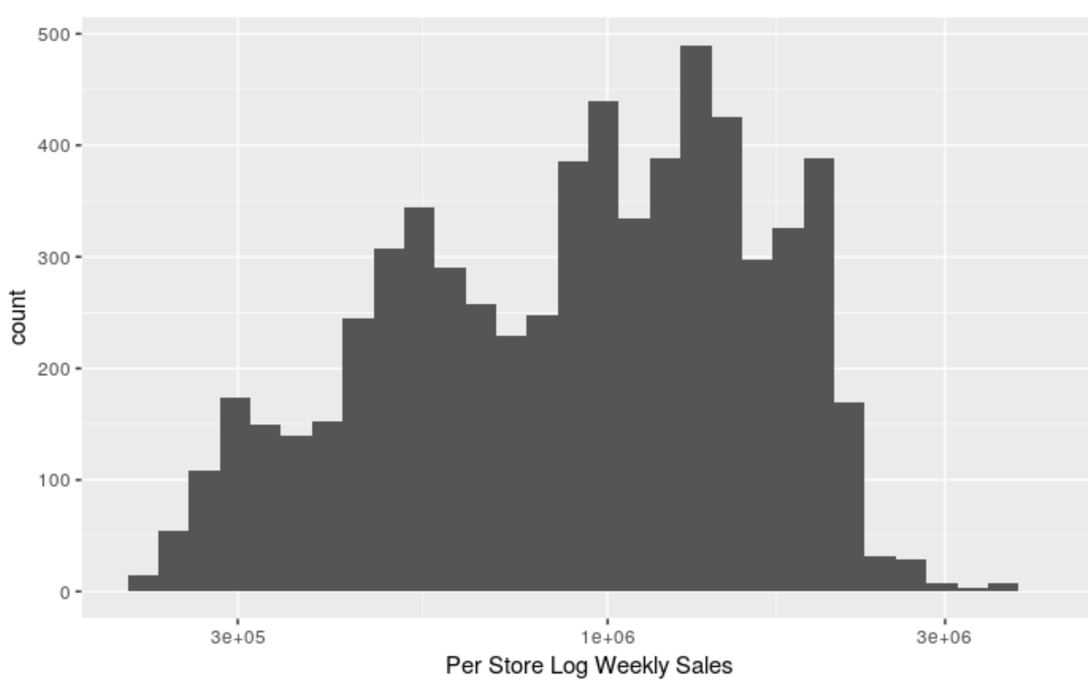
- Change "Store", "Dept" from Integer to Factor
- Change "Date" from Integer to Factor
- Add "Week" variable with isoweek()
- Missing values: missing values only exist in 5 MarkDown variables, accounting for 64% - 74% of variable observations respectively
- Small numbers of negative "Weekly_Sales"

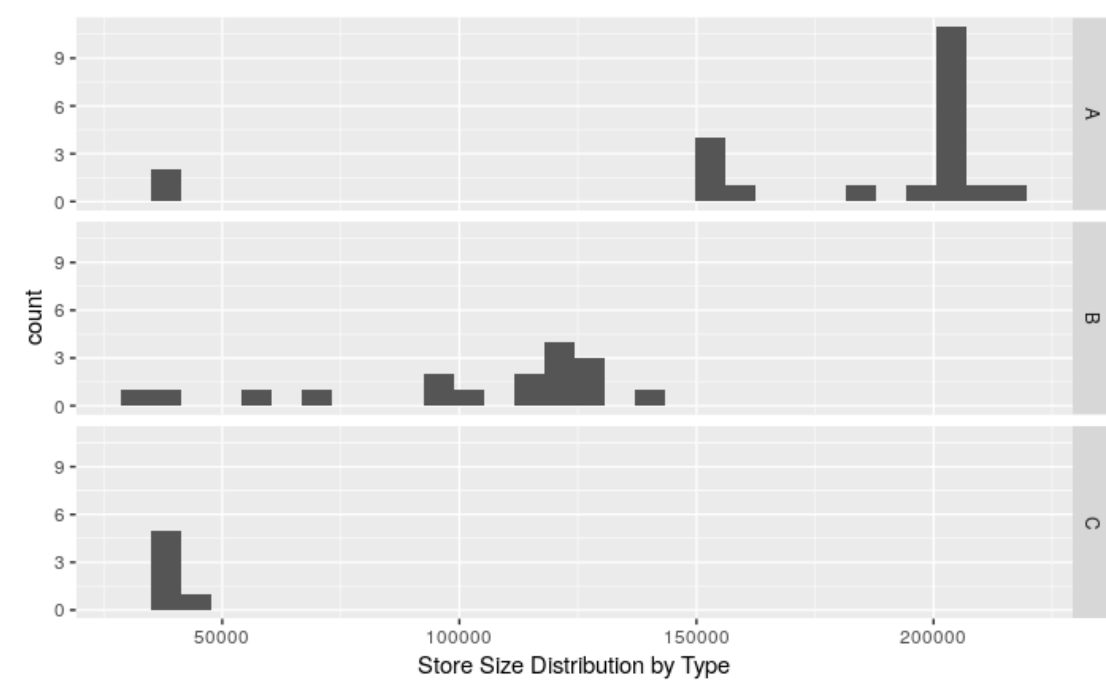
DATA PREPARATION

By SZ



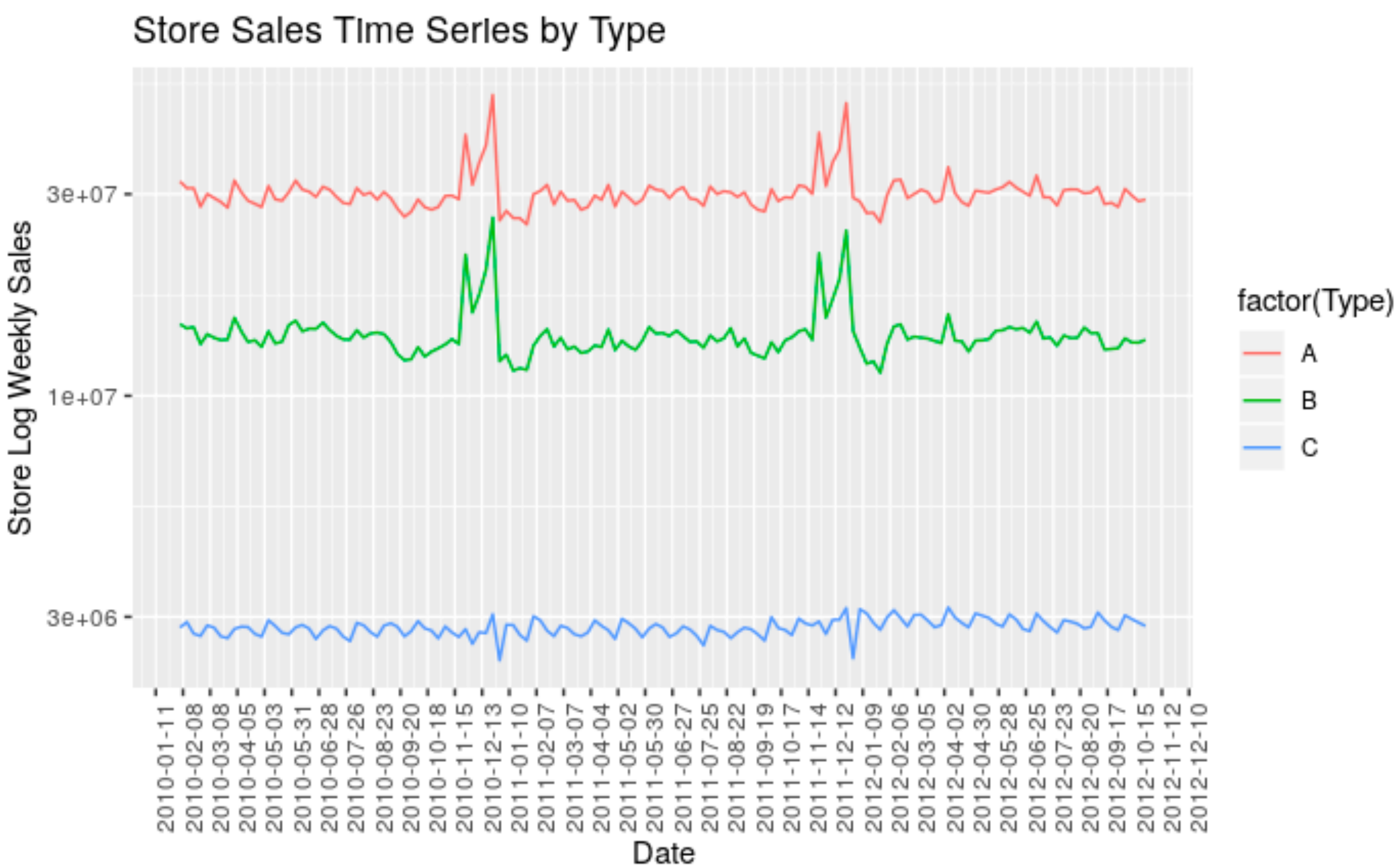


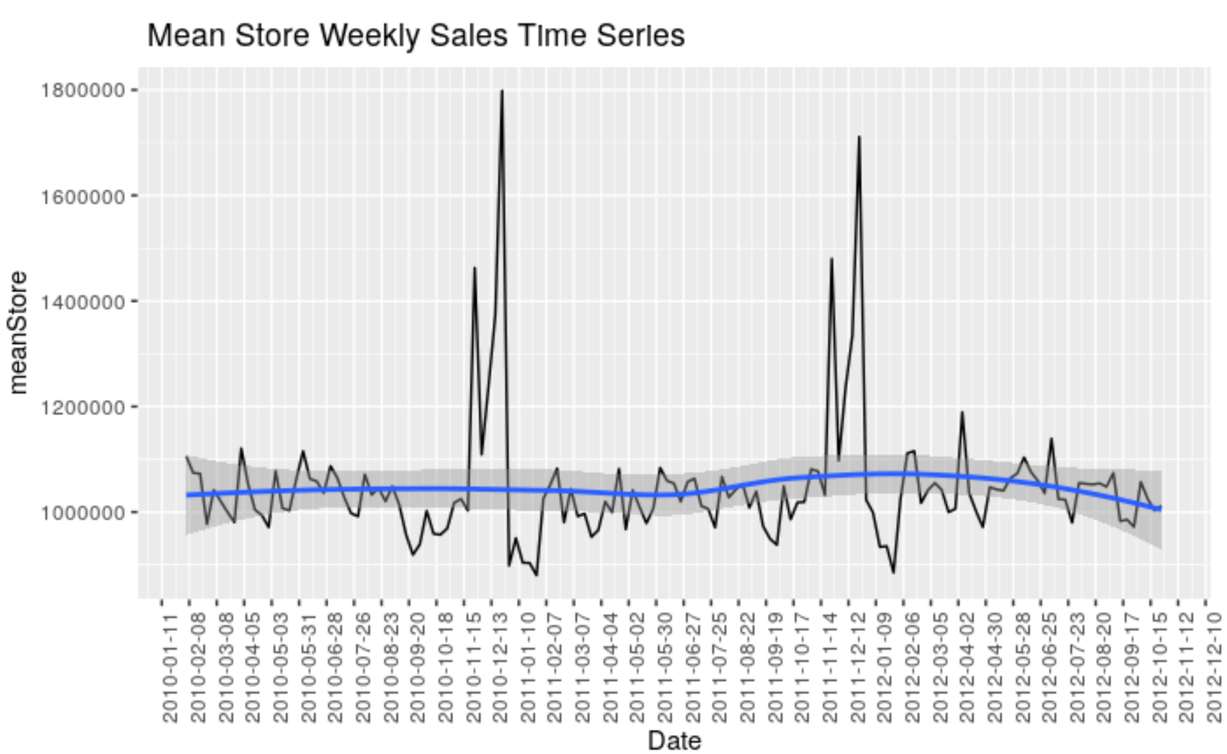




By SZ

DATA PREPARATION





- Time series plotting shows strong seasonality
- Four holidays:

  Super Bowl

  Labor Day

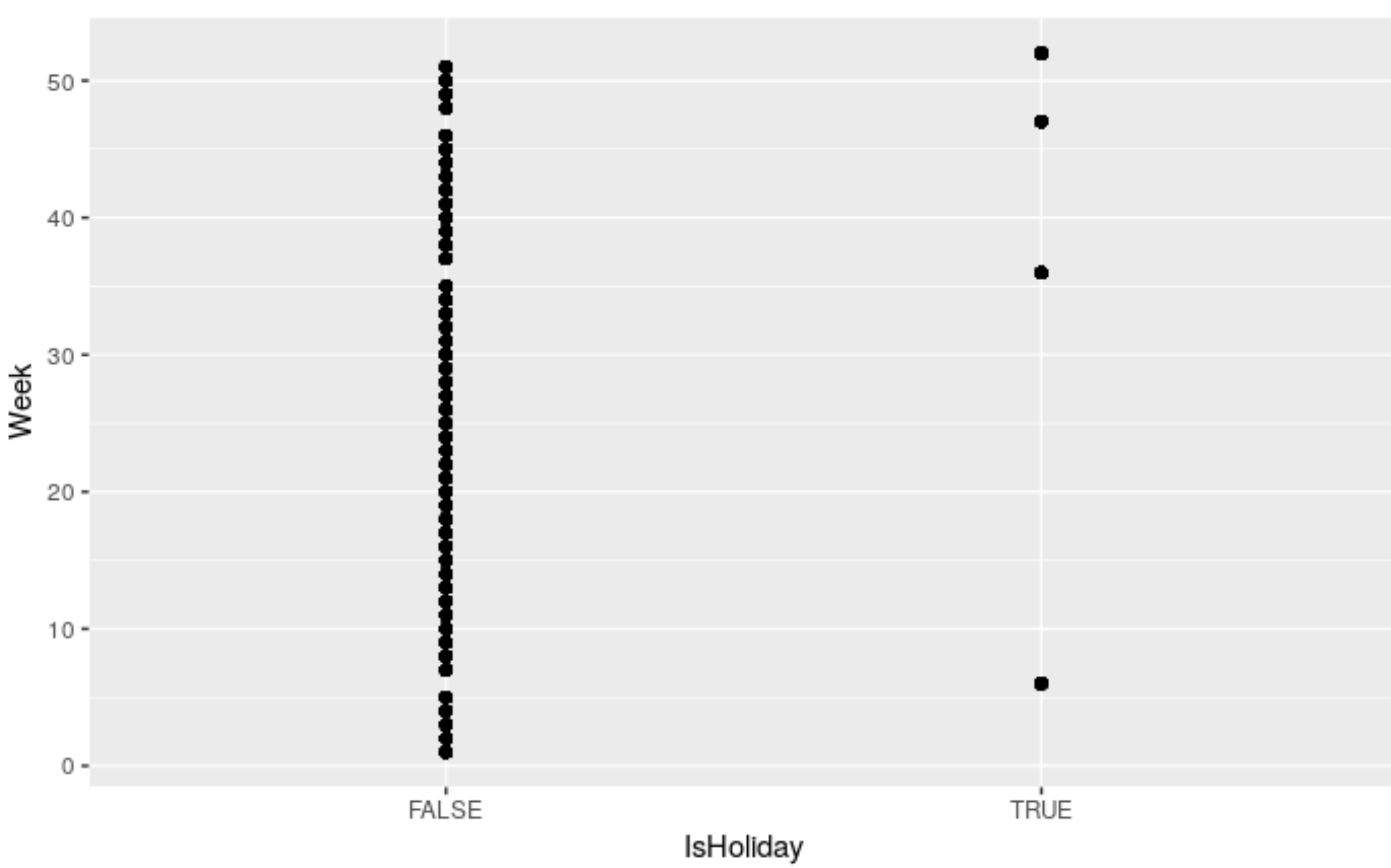
  Thanksgiving

  Christmas
- Sales are dominated by Thanksgiving and Christmas holidays

By SZ

DATA PREPARATION

- Four holidays all appear on the same week of the year
- Super Bowl: week 6, Labor Day: week 36, Thanksgiving: week 47, Christmas: week 52

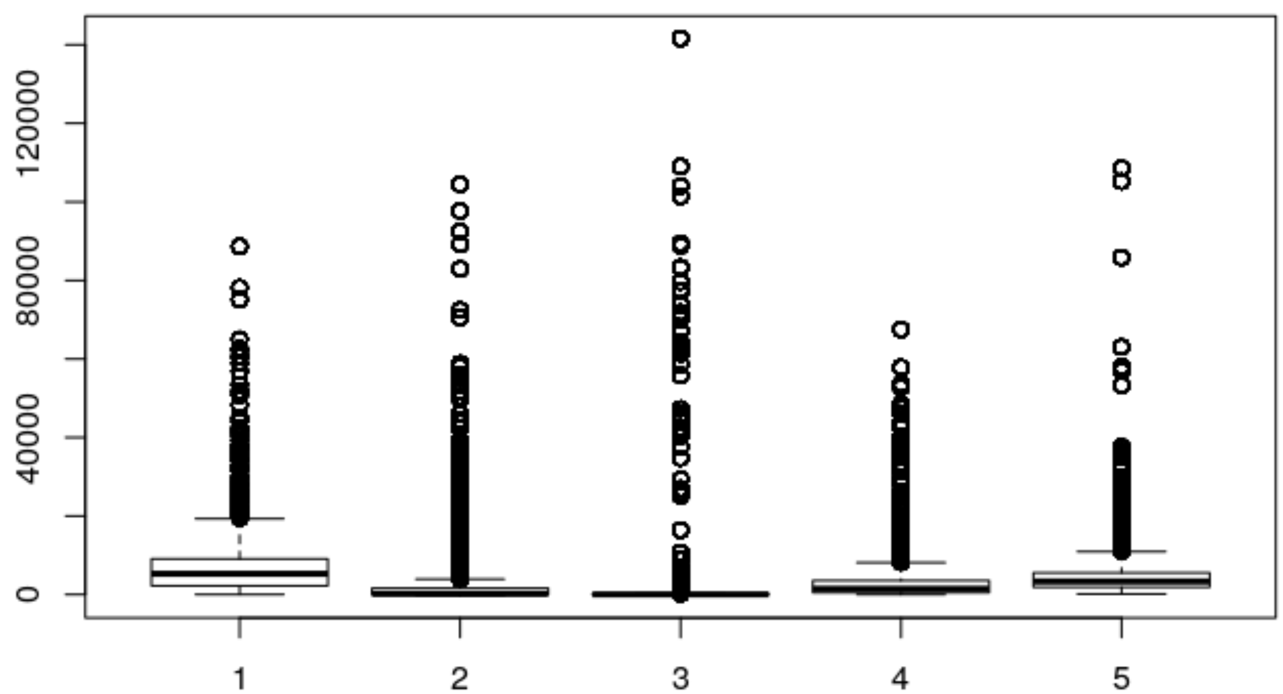


Date	IsHoliday <lgl>	Week <dbl>
2010-02-12	TRUE	6
2010-09-10	TRUE	36
2010-11-26	TRUE	47
2010-12-31	TRUE	52
2011-02-11	TRUE	6
2011-09-09	TRUE	36
2011-11-25	TRUE	47
2011-12-30	TRUE	52
2012-02-10	TRUE	6

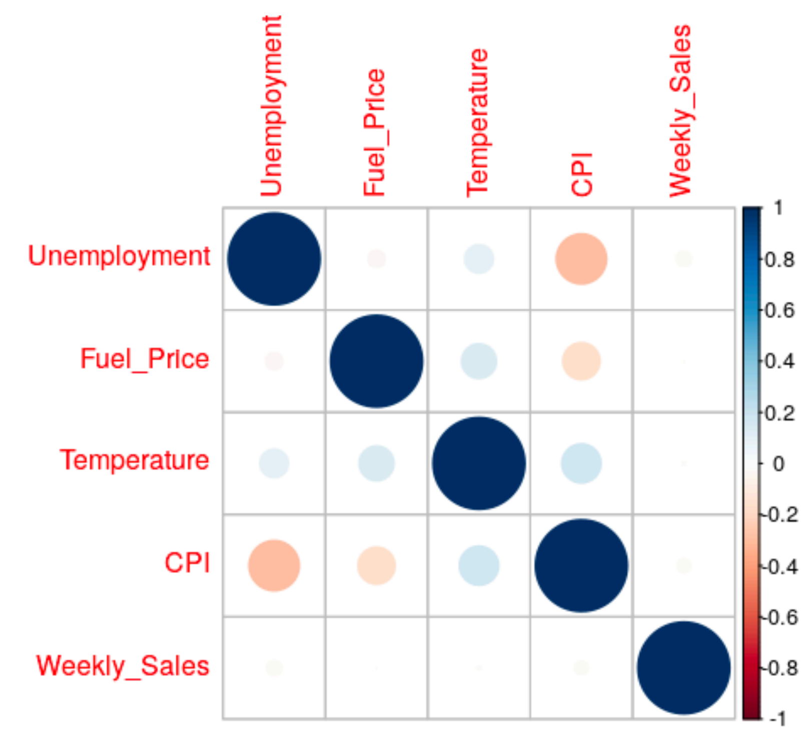
DATA PREPARATION

By SZ

- ➤ Five markdown features are skewed by outliers
- ➤ External indices have almost zero correlation with Weekly_Sales
- ➤ Drop Markdown and external indices features



5 Mardown Features Box Plot

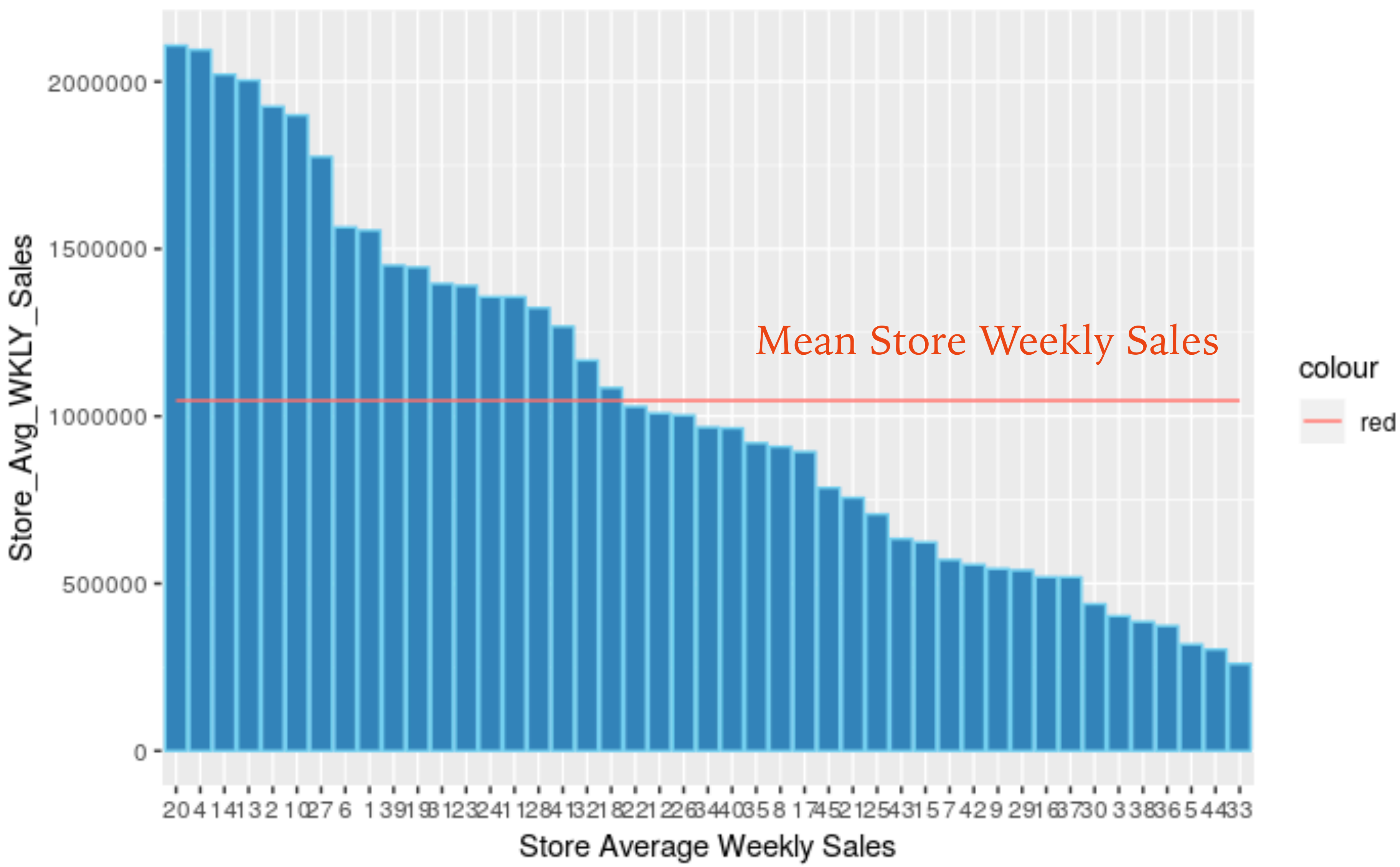


Correlations Matrix

By SZ

DATA PREPARATION

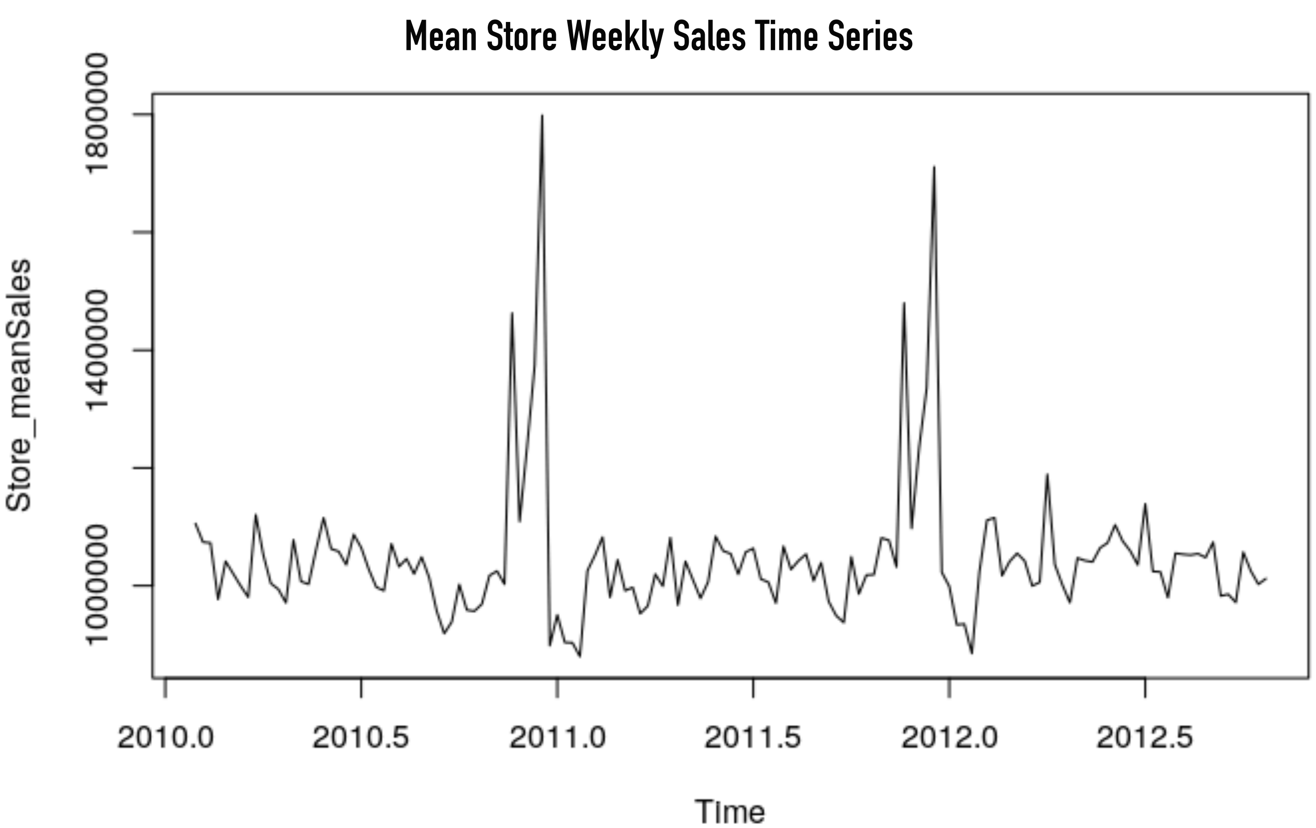
- 3331 Dept-Store ID: use paste() to create ID for every department in every store
- Expensive to compute: 3331 factors X 143 weeks
- This analysis focuses on forecasting 45 stores' aggregated weekly sales

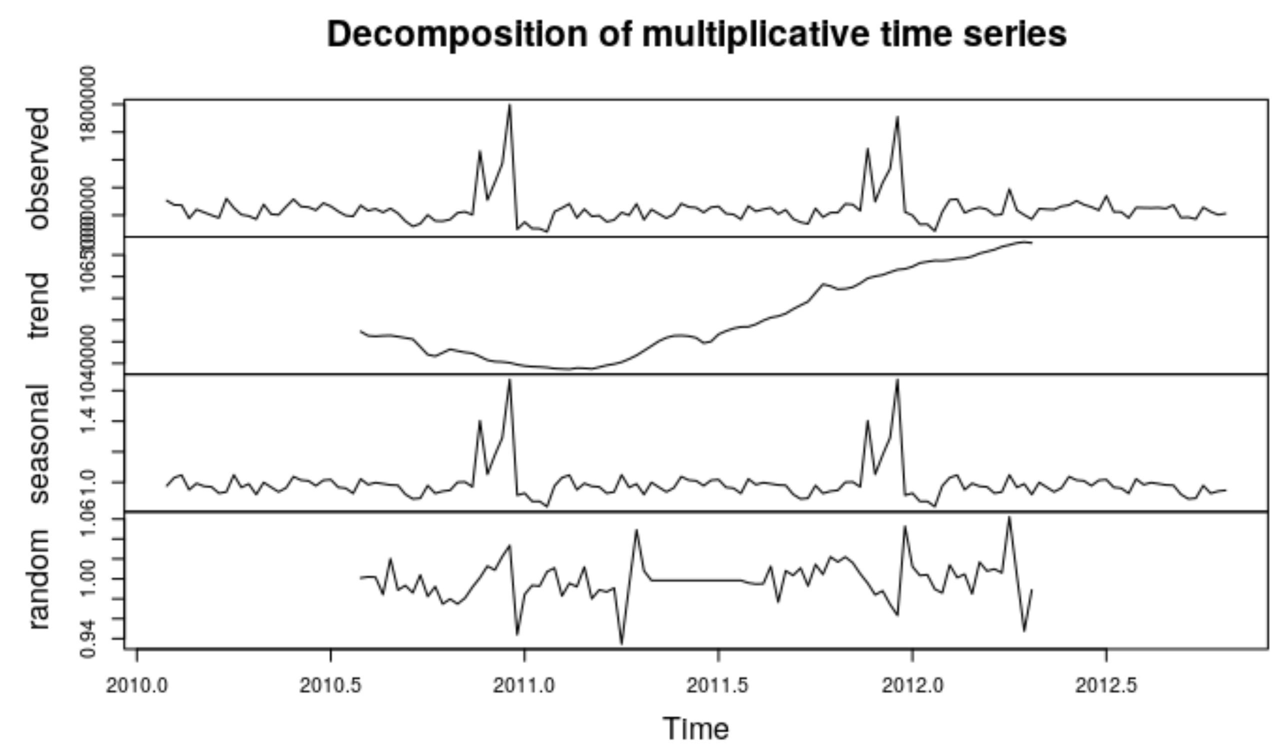


By SZ

DECOMPOSITION OF MULTIPLICATIVE TIME SERIES

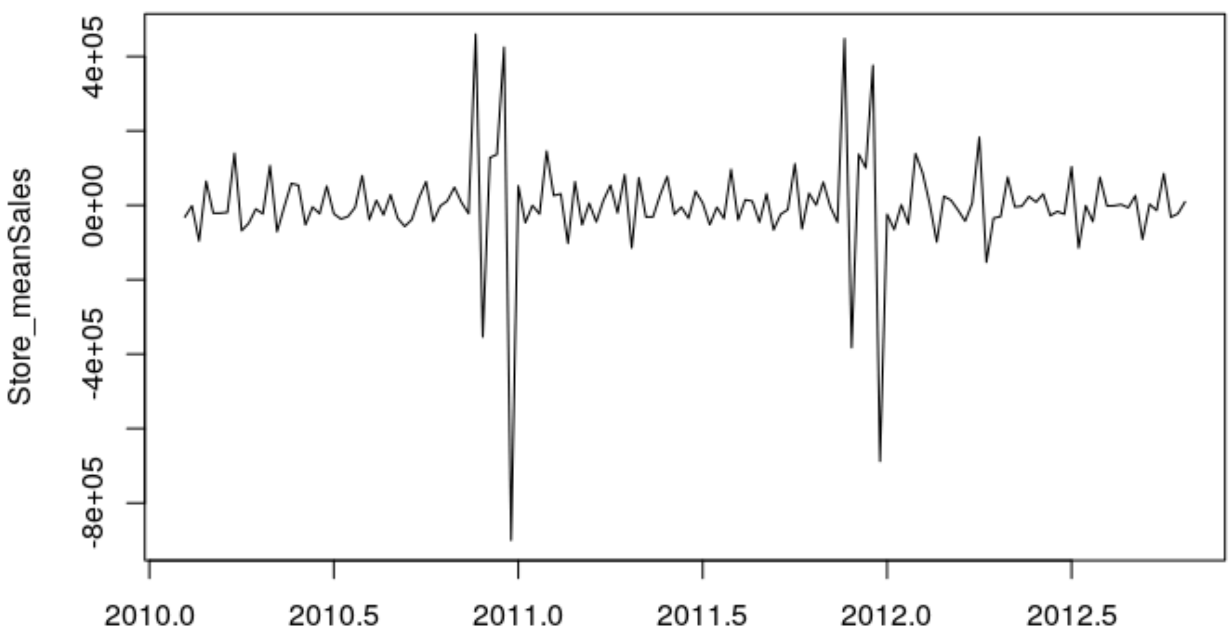
- Decompose mean store weekly sales time series
- Results: strong trend, seasonal components; random component shows randomness



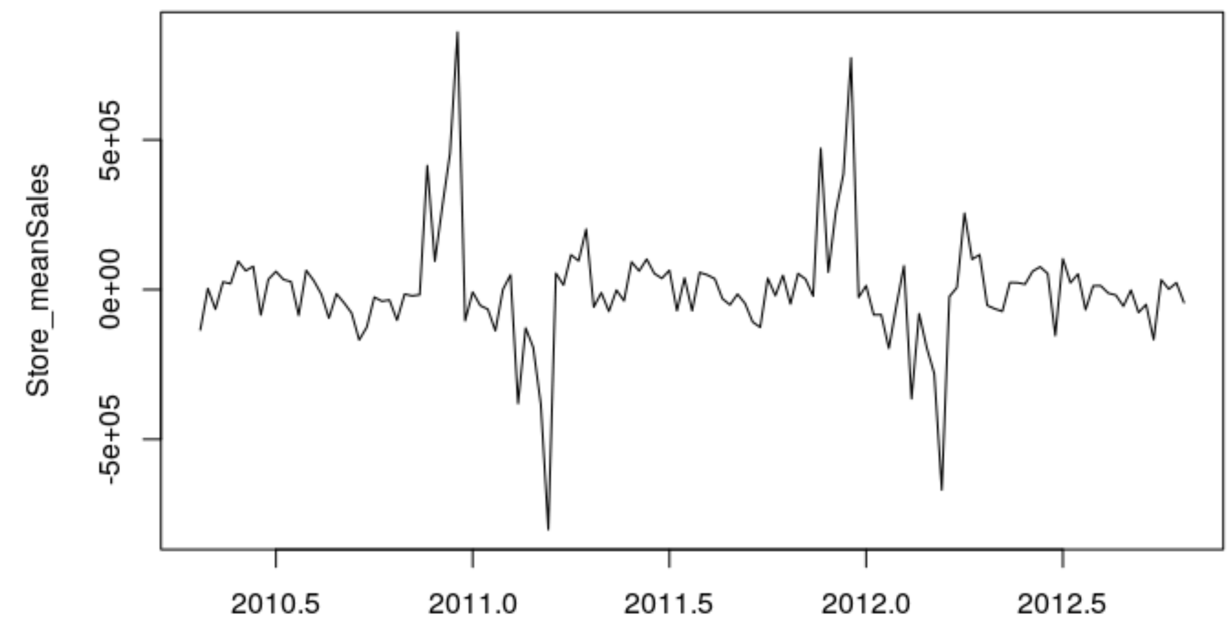


By SZ

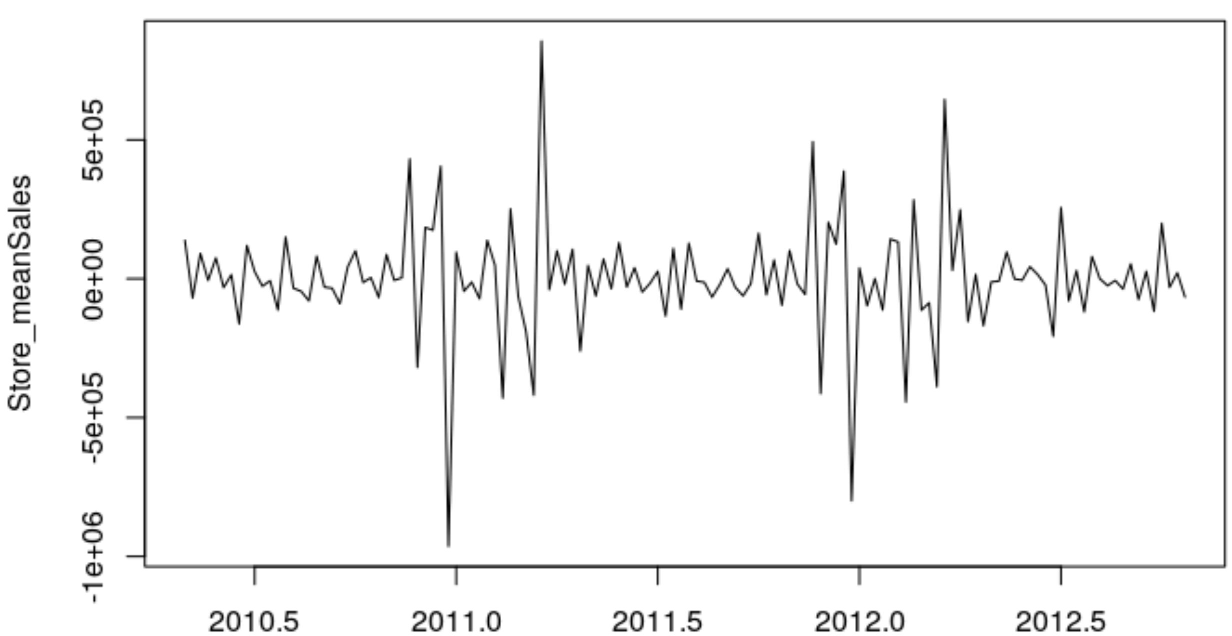
DIFFERENCING TIME SERIES



Time Lag = 1



Seasonal Lag =1



Seasonal Lag = 2

- Differencing removes seasonality in mean store weekly sales time series

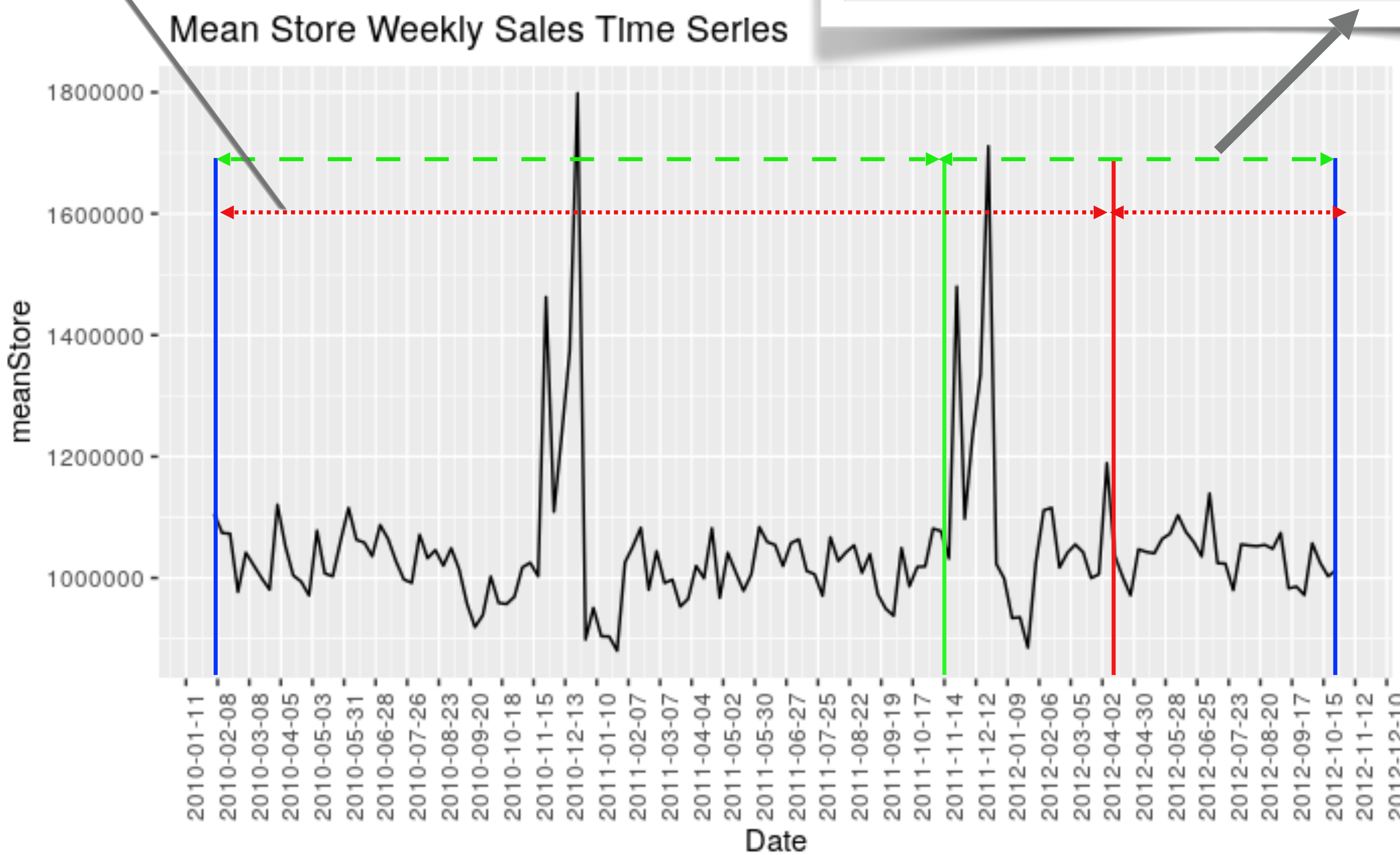
TIME SERIES TRAINING / TEST SETS SELECTION

By SZ

Training/Test option 1: 80/20 split falls on 2012-04-06/2012-04-13

Training/Test option 1: 65/35 split falls on 2011-11-11/2011-11-18

Benefit: include Thanksgiving and Christmas seasonality for test set



TIMES SERIES FORECASTING

By SZ

Methods:

- Seasonal ARIMA
- Random Forests Regression
- XG Boosted Trees
- Neural Networks Regression

Results Diagnostics:

- MAE: scale dependent
- RMSE: scale dependent
- MAPE: non scale dependent

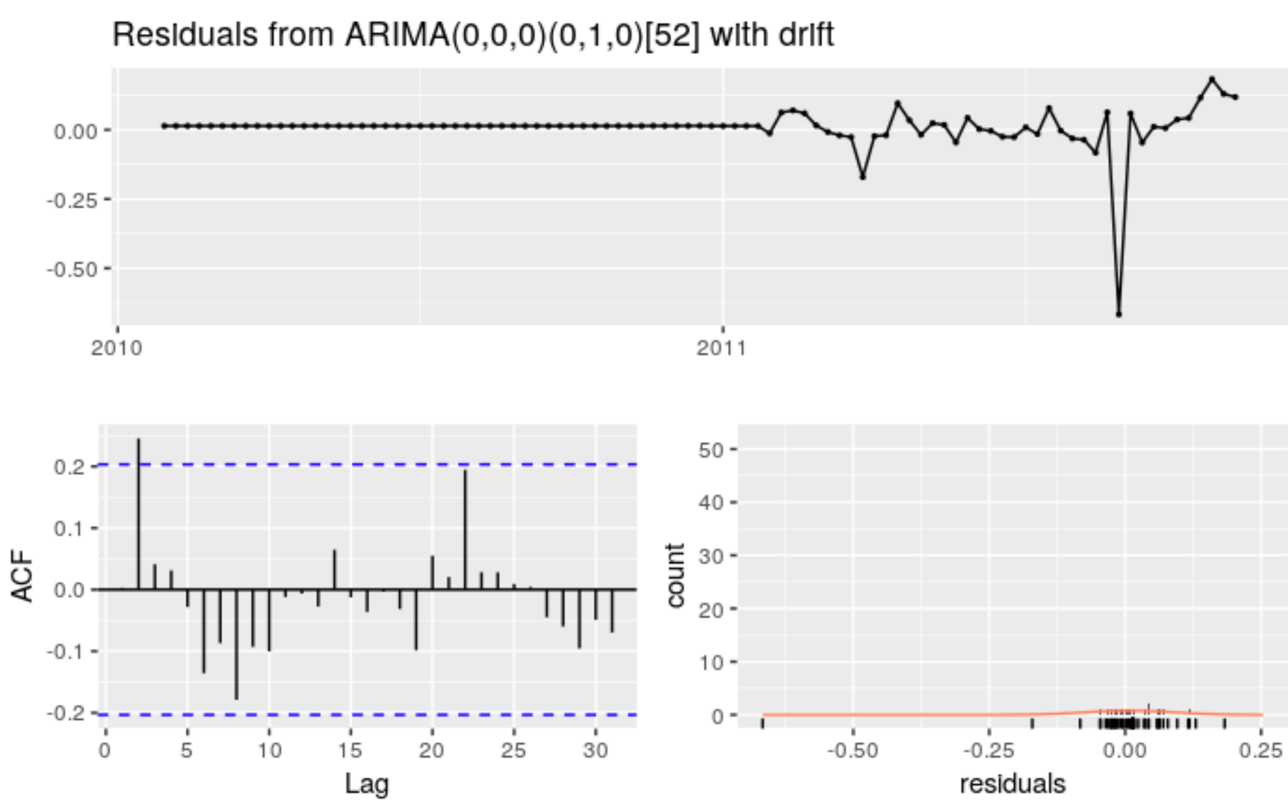
SEASONAL ARIMA MODEL

By SZ

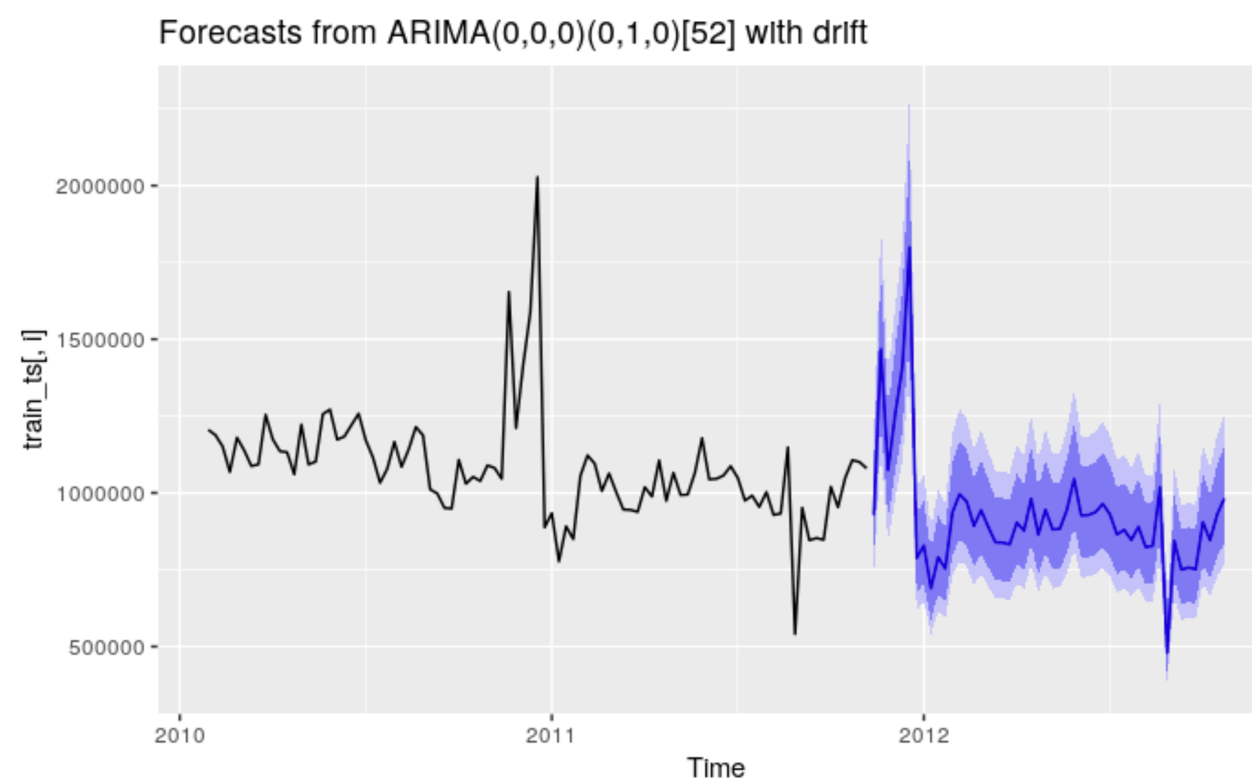
Challenges:

- ➤ Shorter time periods constraint seasonal lag to 1
- ➤ Iterate over 45 time series using for() loop, resulting in slower speed

Residual Plots for Store #18, 65/35 Temporal Split



Forecasts Plot for Store #18, 65/35 Temporal Split

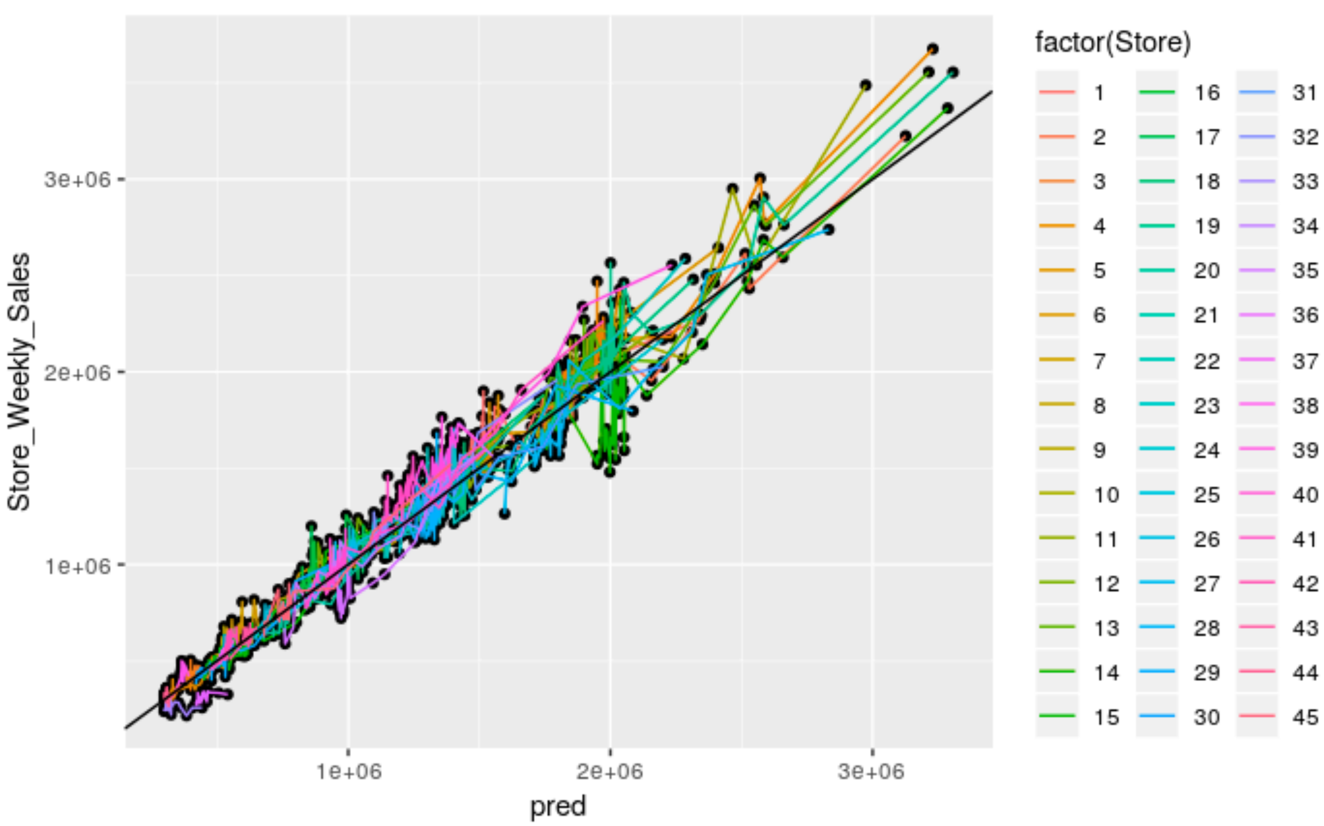


RANDOM FORESTS VS. XG BOOSTED TREES

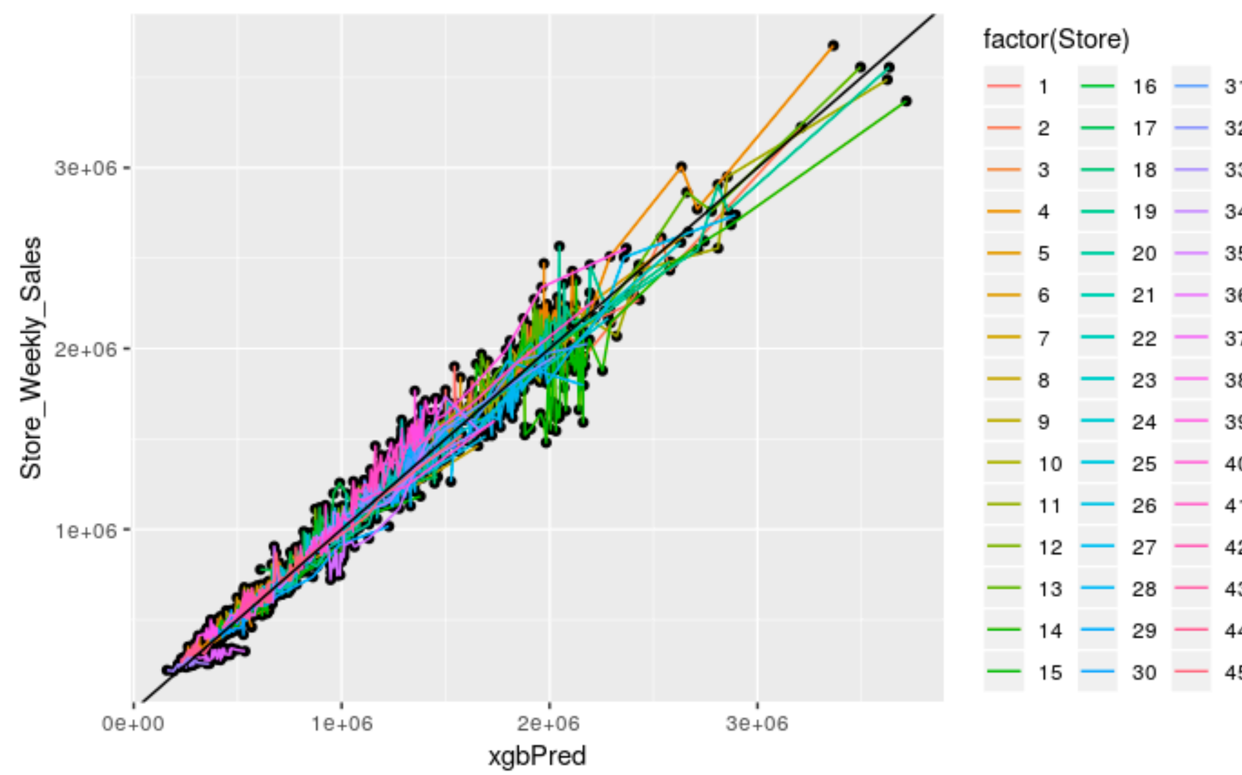
By SZ

XG Boosted Tress has considerably improved prediction accuracy as compared with Random Forests

Random Forests, 65/35 Temporal Split



XG Boosted Trees, 65/35 Temporal Split

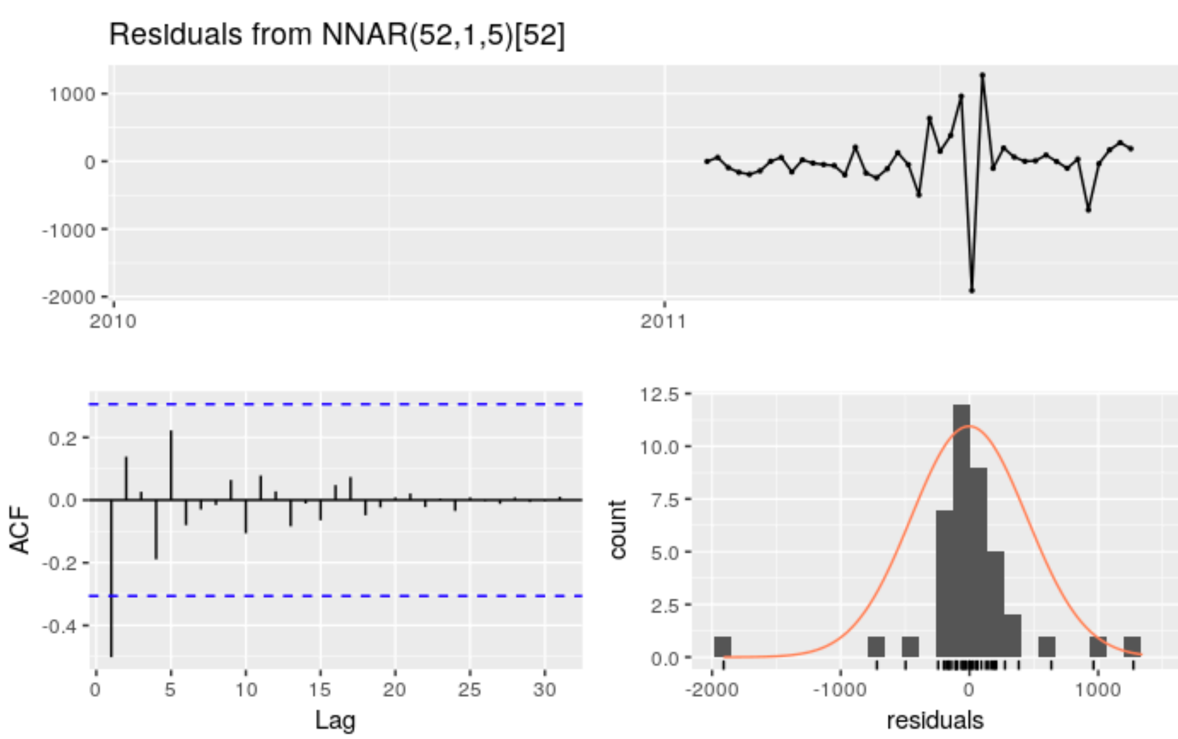


NEURAL NETWORKS

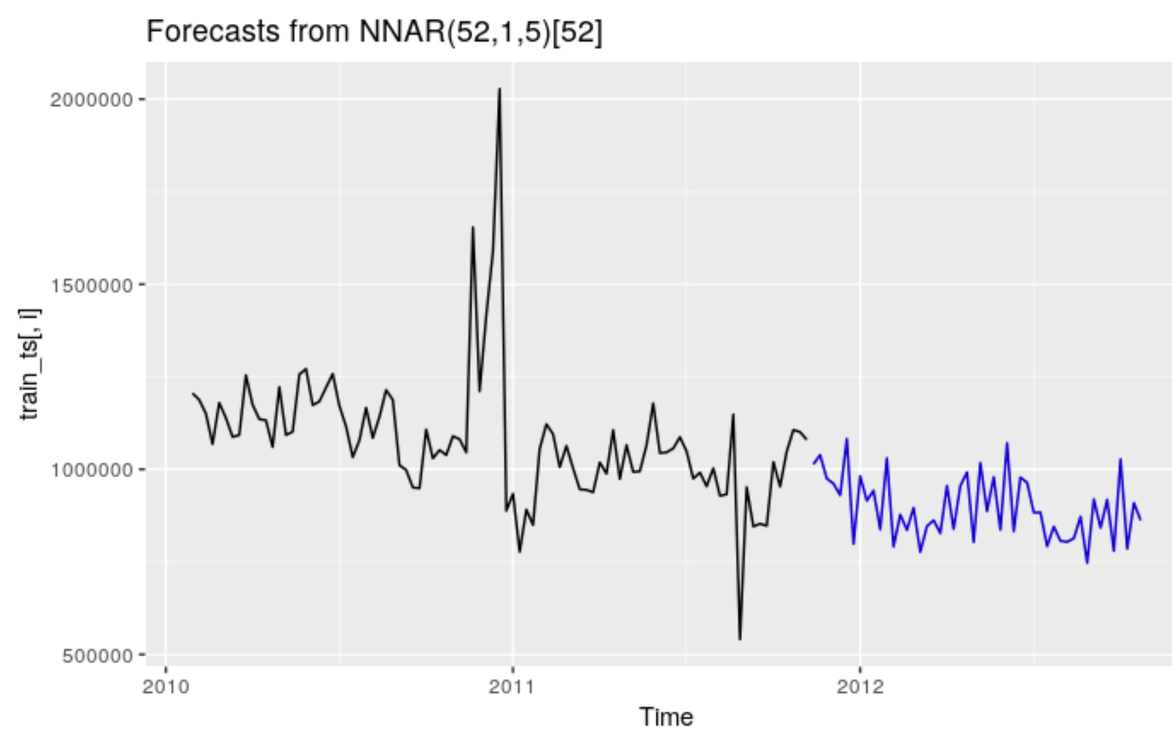
By SZ

- Feed forward nnetar() model with one hidden layer
- 5 neurons in hidden layer, repeat 20 times
- Lambda = 0, no initial data transformation

Residuals Plots for Store # 18, 65/35 Temporal Split



Forecast Plot for Store # 18, 65/35 Temporal Split



MODEL DIAGNOSTICS

By SZ

- Seasonal ARIMA model outperforms in both time split cases
- XG Boost improves tree prediction accuracy notably
- Neural Networks can capture non-linearity better in 65/35 split

80/20 Temporal Split on 2012-04-13

	MAE	RMSE	MAPE
ARIMA	57517	71755	5.53
Random Forests	65746	77079	7.04
Xgboost	63740	73969	6.75
Neural Networks	116660	137922	10.49

65/35 Temporal Split on 2011-11-18

	MAE	RMSE	MAPE
ARIMA	61988	78263	5.76
Random Forests	70209	86753	7.29
Xgboost	66535	80961	6.92
Neural Networks	98224	152123	8.37

By SZ

SHORTCOMINGS

- Lack of true test set data penalizes machine learning models
- Abnormally large missing values force to drop MarkDown features
- Model diagnostics are equal-weight store accuracy statistics
- Overweighting holiday prediction errors should be explored