Retail Data Analysis
CS109a Project

Cathy, Will, Usman, Aryamaan

Walmart

Background

Motivation & Context

- Walmart, one of the top retail chains in the US, wants to estimate sales and profits for earnings forecasts, which is of concern to investors and stakeholders
- Fluctuations in demand is a hurdle for the company make sales projections unstable, especially in an unpredictable macro climate

Description of the Data

- Weekly sales data for 45 Walmart locations nationwide in many geographies from 2010-02-05 to 2012-11-01.
- Also includes variables such as temperature, gasoline prices, holidays, the unemployment rate, and shifts in consumer price indices.

- Store: the store number
- Date: the week of sales
- Weekly_Sales: weekly sales for the given store
- Holiday_Flag: whether the week is a special holiday week 1 – Holiday week 0 – Non-holiday week
- Temperature: temperature on the day of sale
- Fuel_Price: cost of fuel in the region
- CPI: prevailing consumer price index
- Unemployment: prevailing unemployment rate

Description of the Data

First 5 rows:

	Store	Date	Weekly_Sales	Holiday_Flag	Temperature	Fuel_Price	CPI	Unemployment
0	1	05-02-2010	1643690.90	0	42.31	2.572	211.096358	8.106
1	1	12-02-2010	1641957.44	1	38.51	2.548	211.242170	8.106
2	1	19-02-2010	1611968.17	0	39.93	2.514	211.289143	8.106
3	1	26-02-2010	1409727.59	0	46.63	2.561	211.319643	8.106
4	1	05-03-2010	1554806.68	0	46.50	2.625	211.350143	8.106

Summary of data:

	Store	Weekly_Sales	Holiday_Flag	Temperature	Fuel_Price	CPI	Unemployment
count	6435.000000	6.435000e+03	6435.000000	6435.000000	6435.000000	6435.000000	6435.000000
mean	23.000000	1.046965e+06	0.069930	60.663782	3.358607	171.578394	7.999151
std	12.988182	5.643666e+05	0.255049	18.444933	0.459020	39.356712	1.875885
min	1.000000	2.099862e+05	0.000000	-2.060000	2.472000	126.064000	3.879000
25%	12.000000	5.533501e+05	0.000000	47.460000	2.933000	131.735000	6.891000
50%	23.000000	9.607460e+05	0.000000	62.670000	3.445000	182.616521	7.874000
75%	34.000000	1.420159e+06	0.000000	74.940000	3.735000	212.743293	8.622000
max	45.000000	3.818686e+06	1.000000	100.140000	4.468000	227.232807	14.313000

Project Question

To what extent can we predict Walmart's sales performance across 45 representative U.S. stores from macroeconomic and external factors?

Walmart

Exploratory Data Analysis

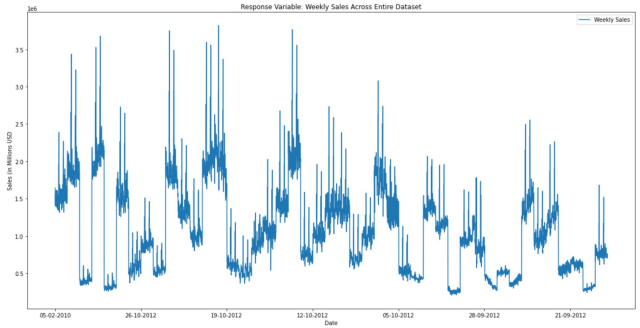
Data Quality

- Mix of continuous, categorical, & binary variables
- No null or duplicate values
- Distributions all within reason

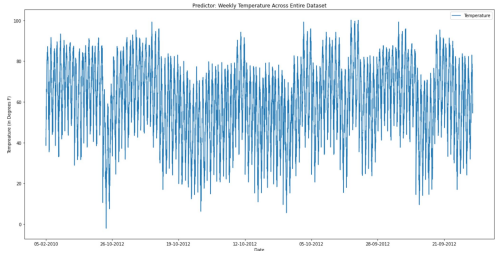
	Weekly_Sales	Holiday_Flag	Temperature	Fuel_Price	CPI	Unemployment
count	6.435000e+03	6435.000000	6435.000000	6435.000000	6435.000000	6435.000000
mean	1.046965e+06	0.069930	60.663782	3.358607	171.578394	7.999151
std	5.643666e+05	0.255049	18.444933	0.459020	39.356712	1.875885
min	2.099862e+05	0.000000	-2.060000	2.472000	126.064000	3.879000
25%	5.533501e+05	0.000000	47.460000	2.933000	131.735000	6.891000
50%	9.607460e+05	0.000000	62.670000	3.445000	182.616521	7.874000
75%	1.420159e+06	0.000000	74.940000	3.735000	212.743293	8.622000
max	3.818686e+06	1.000000	100.140000	4.468000	227.232807	14.313000

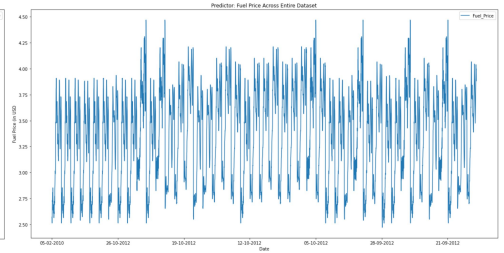
Response and Predictors

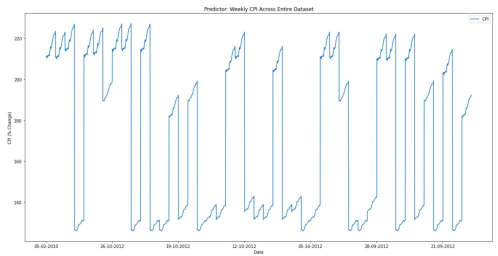
- All exhibit roughly annual seasonality
- Response variable (sales):

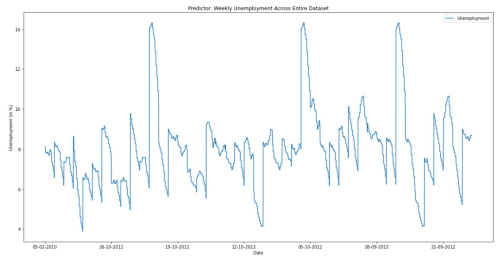


- Predictors:



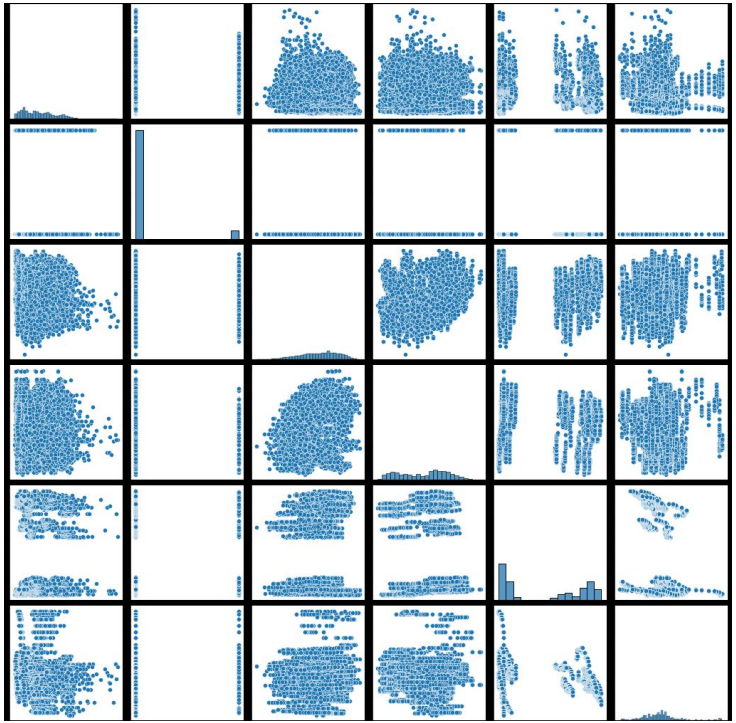






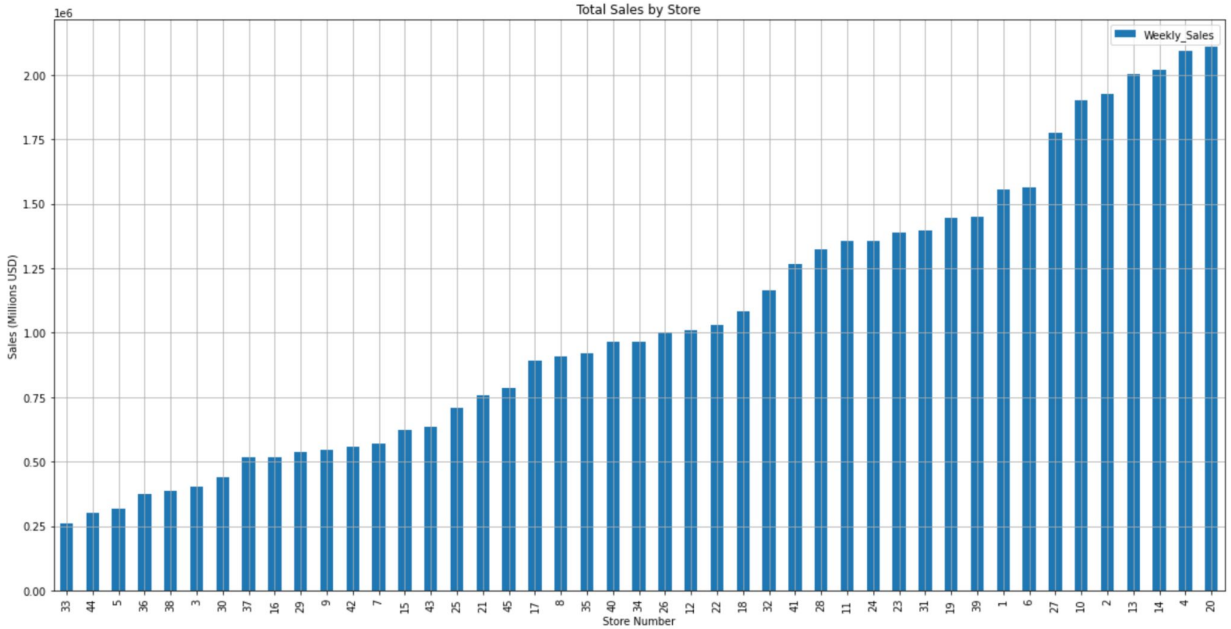
Pairplot

No immediately obvious correlations or relationships between variables



Store-Specific

Large variation in relative total sales of 45 individual stores

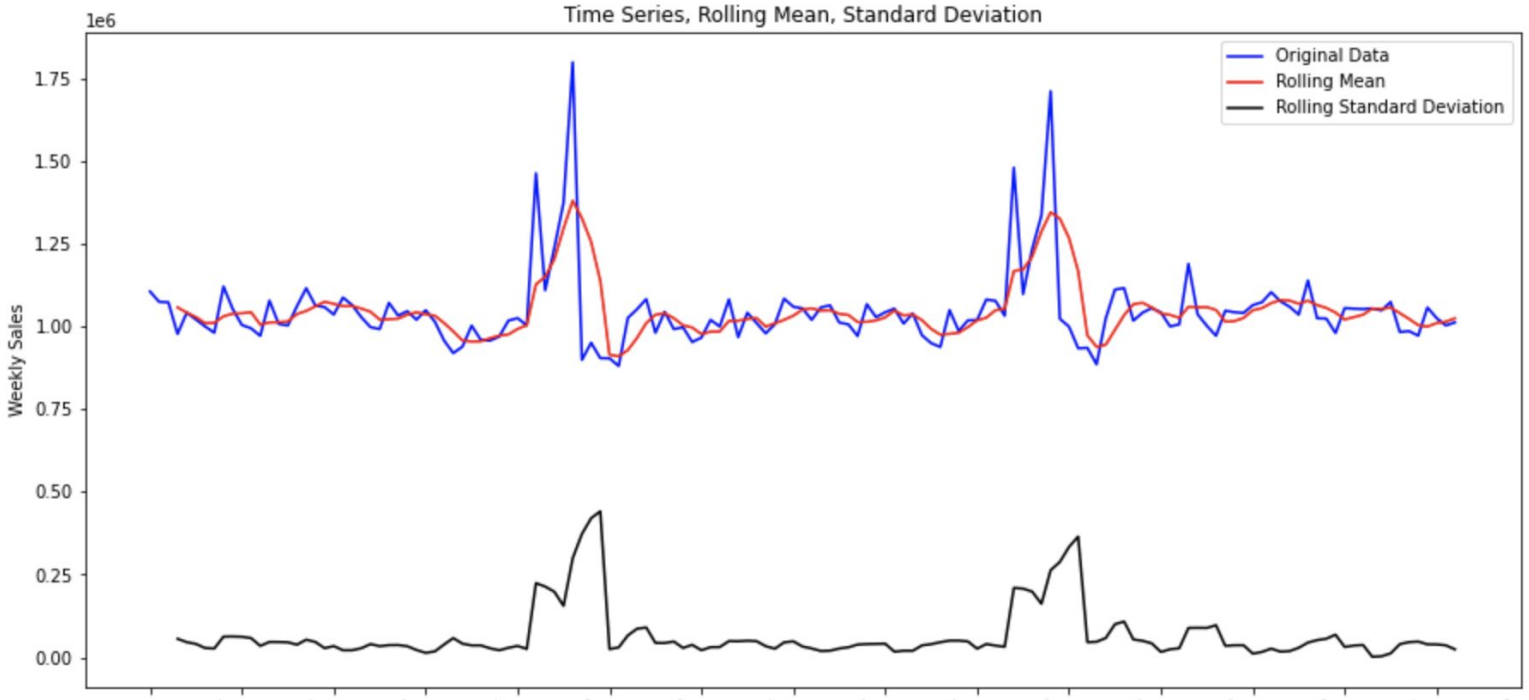


Walmart

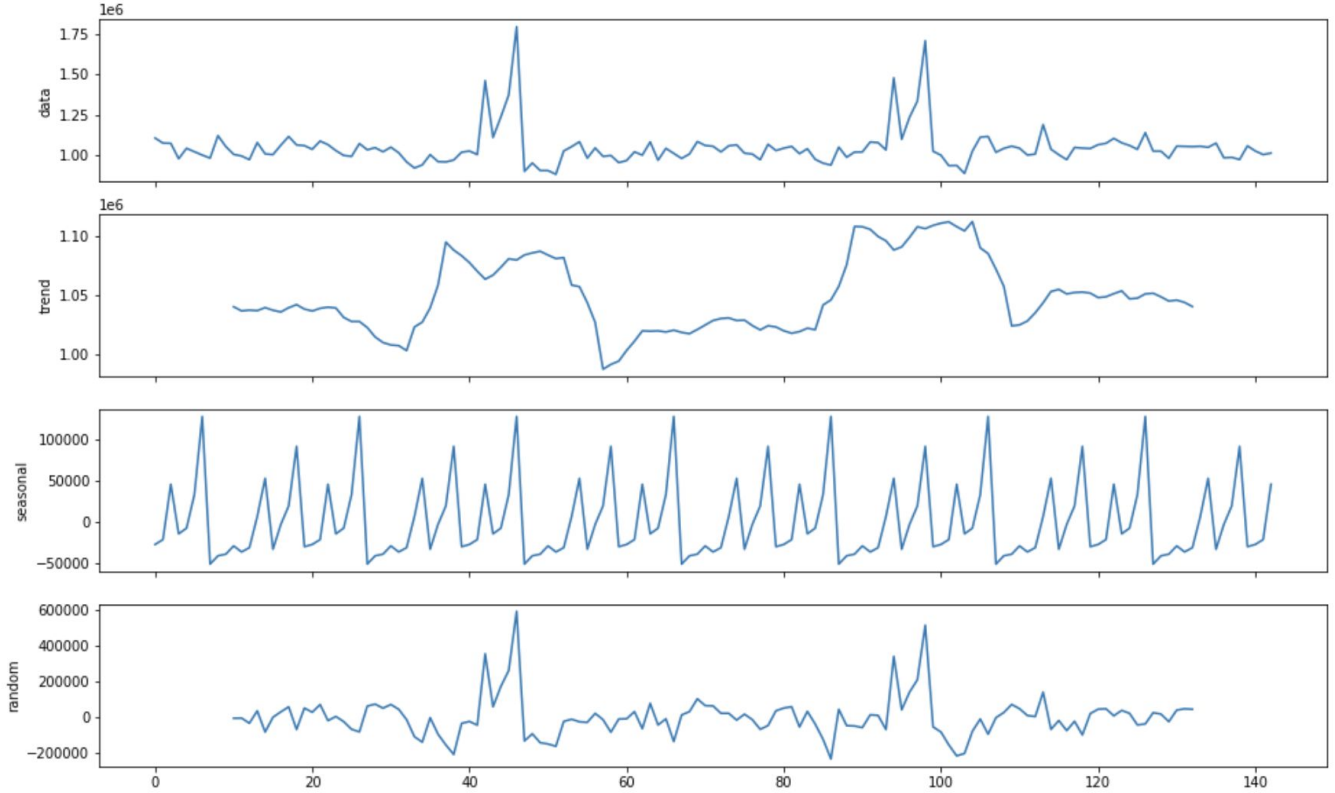
Baseline Model

Stationary Test

	Values	Metric
0	-5.908298e+00	Test Statistics
1	2.675979e-07	p-value
2	4.000000e+00	No. of lags used
3	1.380000e+02	Number of observations used
4	-3.478648e+00	critical value (1%)
5	-2.882722e+00	critical value (5%)
6	-2.578065e+00	critical value (10%)



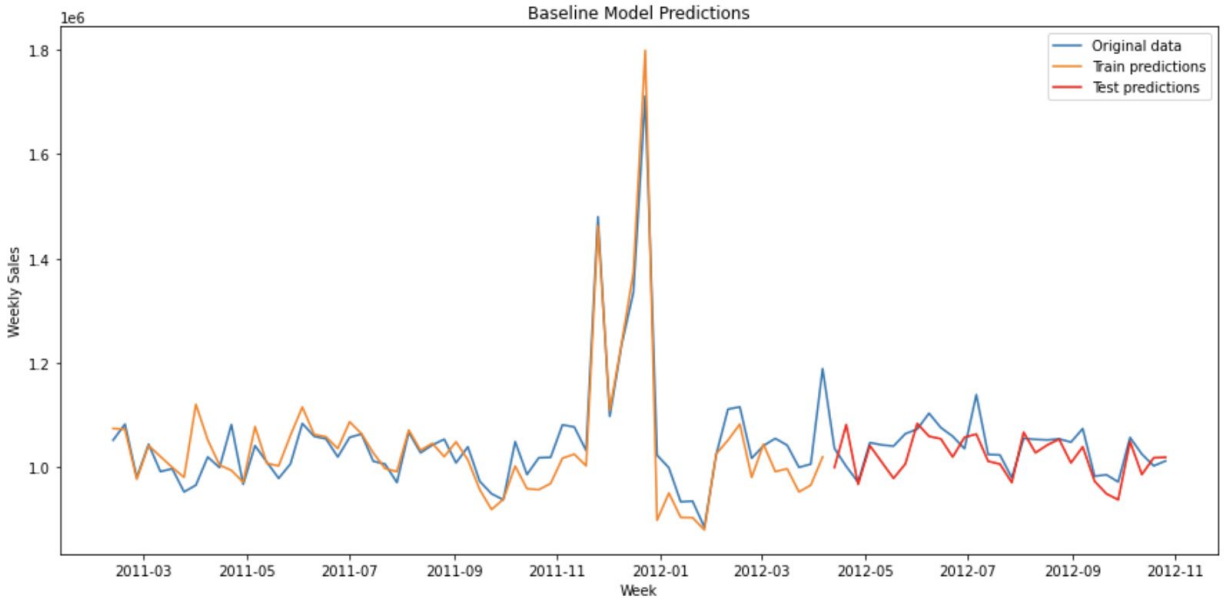
Seasonal Decompose



Baseline Model

Date	t-52	t
2012-04-13	9.994073e+05	1.036206e+06
2012-04-20	1.081704e+06	1.001612e+06
2012-04-27	9.673341e+05	9.714844e+05
2012-05-04	1.041377e+06	1.047204e+06
2012-05-11	1.009914e+06	1.042797e+06

	MAPE
train	0.031124
test	0.026532



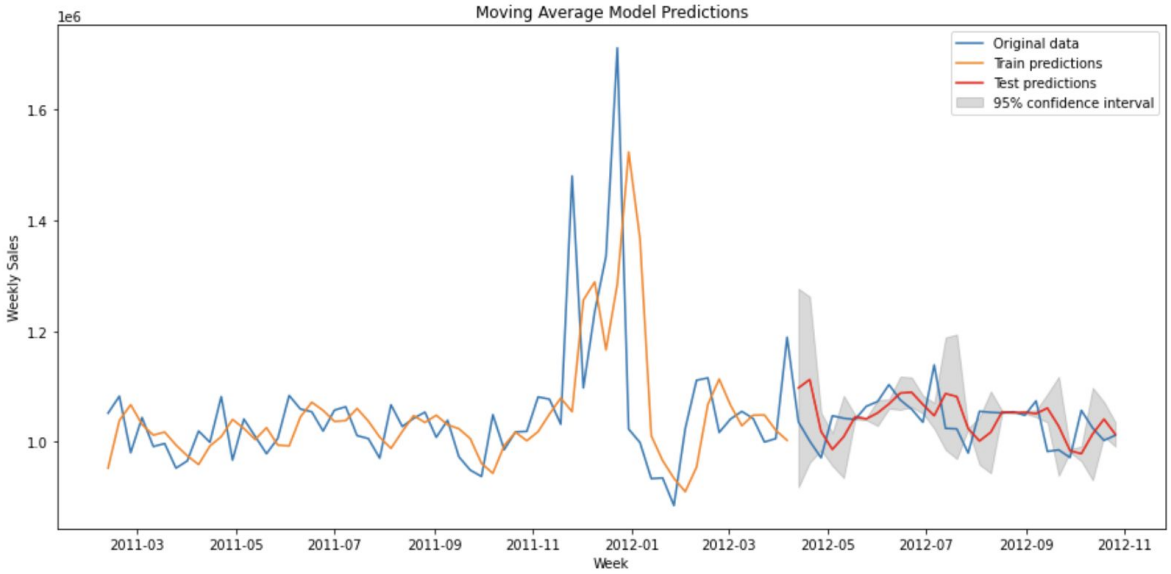
Walmart

Moving Average Models

Moving Average Model

Date	rolling_mean	rolling_std	t
2012-04-13	1.097502e+06	129313.391618	1.036206e+06
2012-04-20	1.112573e+06	107999.631470	1.001612e+06
2012-04-27	1.018909e+06	24461.677601	9.714844e+05
2012-05-04	9.865481e+05	21303.255684	1.047204e+06
2012-05-11	1.009344e+06	53542.110387	1.042797e+06

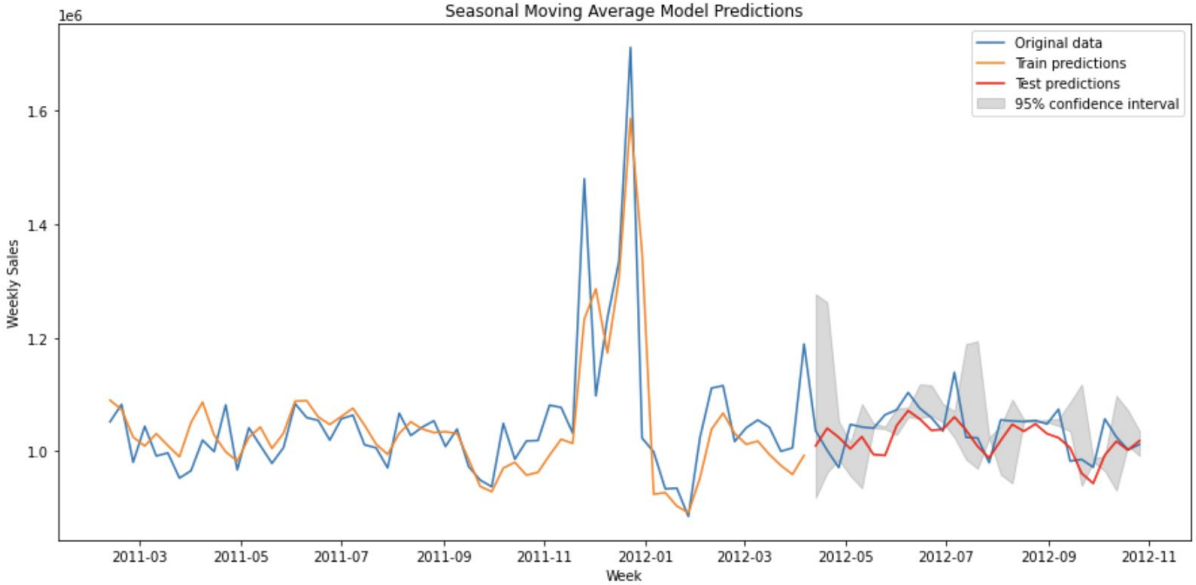
	MAPE
train	0.065877
test	0.037029



Seasonal Moving Average Model

Date	diff_rolling_mean	t
2012-04-13	1.009564e+06	1.036206e+06
2012-04-20	1.040556e+06	1.001612e+06
2012-04-27	1.024519e+06	9.714844e+05
2012-05-04	1.004355e+06	1.047204e+06
2012-05-11	1.025646e+06	1.042797e+06

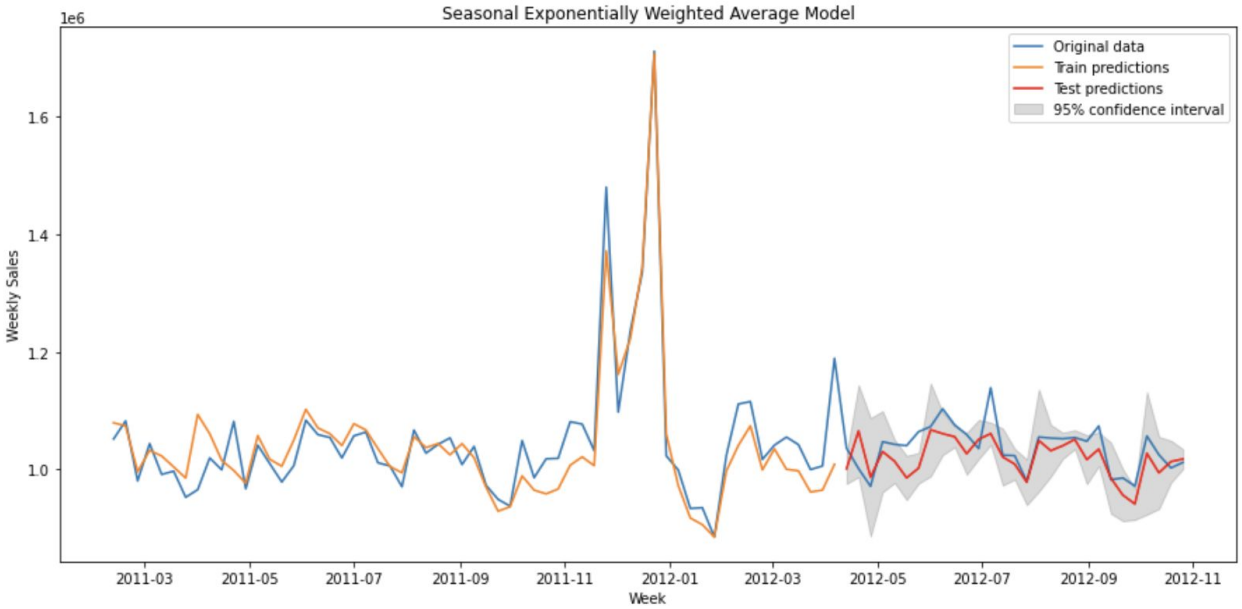
	MAPE
train	0.044108
test	0.026440



Seasonal Exponentially Weighted Model

Date	rolling_ewm	t
2012-04-13	1.001280e+06	1.036206e+06
2012-04-20	1.065619e+06	1.001612e+06
2012-04-27	9.869911e+05	9.714844e+05
2012-05-04	1.030500e+06	1.047204e+06
2012-05-11	1.014031e+06	1.042797e+06

	MAPE
train	0.029934
test	0.024380



Walmart

SARIMA Model

SARIMA Parameter Search

```
Performing stepwise search to minimize aic
 ARIMA(1,0,1)(0,1,1)[15] intercept   : AIC=2677.570, Time=1.14 sec
 ARIMA(0,0,0)(0,1,0)[15] intercept   : AIC=2699.629, Time=0.09 sec
 ARIMA(1,0,0)(1,1,0)[15] intercept   : AIC=2682.396, Time=0.81 sec
 ARIMA(0,0,1)(0,1,1)[15] intercept   : AIC=2688.614, Time=0.93 sec
 ARIMA(0,0,0)(0,1,0)[15]             : AIC=2698.666, Time=0.07 sec
 ARIMA(1,0,1)(0,1,0)[15] intercept   : AIC=2691.319, Time=0.24 sec
 ARIMA(1,0,1)(1,1,1)[15] intercept   : AIC=2679.164, Time=0.76 sec
 ARIMA(1,0,1)(0,1,2)[15] intercept   : AIC=2679.292, Time=1.43 sec
 ARIMA(1,0,1)(1,1,0)[15] intercept   : AIC=2680.570, Time=0.51 sec
 ARIMA(1,0,1)(1,1,2)[15] intercept   : AIC=inf, Time=3.43 sec
 ARIMA(1,0,0)(0,1,1)[15] intercept   : AIC=2680.914, Time=0.39 sec
 ARIMA(2,0,1)(0,1,1)[15] intercept   : AIC=2679.132, Time=0.74 sec
 ARIMA(1,0,2)(0,1,1)[15] intercept   : AIC=2679.679, Time=0.86 sec
 ARIMA(0,0,0)(0,1,1)[15] intercept   : AIC=2697.932, Time=0.43 sec
 ARIMA(0,0,2)(0,1,1)[15] intercept   : AIC=2684.157, Time=0.75 sec
 ARIMA(2,0,0)(0,1,1)[15] intercept   : AIC=2676.627, Time=0.60 sec
 ARIMA(2,0,0)(0,1,0)[15] intercept   : AIC=2691.338, Time=0.12 sec
 ARIMA(2,0,0)(1,1,1)[15] intercept   : AIC=2678.475, Time=0.58 sec
 ARIMA(2,0,0)(0,1,2)[15] intercept   : AIC=2678.494, Time=1.09 sec
 ARIMA(2,0,0)(1,1,0)[15] intercept   : AIC=2680.807, Time=0.38 sec
 ARIMA(2,0,0)(1,1,2)[15] intercept   : AIC=2679.672, Time=2.00 sec
 ARIMA(3,0,0)(0,1,1)[15] intercept   : AIC=2678.851, Time=0.62 sec
 ARIMA(3,0,1)(0,1,1)[15] intercept   : AIC=2678.472, Time=1.26 sec
 ARIMA(2,0,0)(0,1,1)[15]             : AIC=2673.460, Time=0.53 sec
 ARIMA(2,0,0)(0,1,0)[15]             : AIC=2689.452, Time=0.11 sec
 ARIMA(2,0,0)(1,1,1)[15]             : AIC=2675.407, Time=0.48 sec
 ARIMA(2,0,0)(0,1,2)[15]             : AIC=2675.385, Time=1.06 sec
```

```
Best model:  ARIMA(3,0,2)(0,1,1)[15]
Total fit time: 71.008 seconds
```

SARIMAX Results

Dep. Variable:	y	**No. Observations:**	114
Model:	SARIMAX(3, 0, 2)x(0, 1, [1], 15)	**Log Likelihood**	-1323.956
Date:	Sat, 10 Dec 2022	**AIC**	2661.913
Time:	20:56:51	**BIC**	2680.078
Sample:	02-05-2010	**HQIC**	2669.263
	- 04-06-2012		

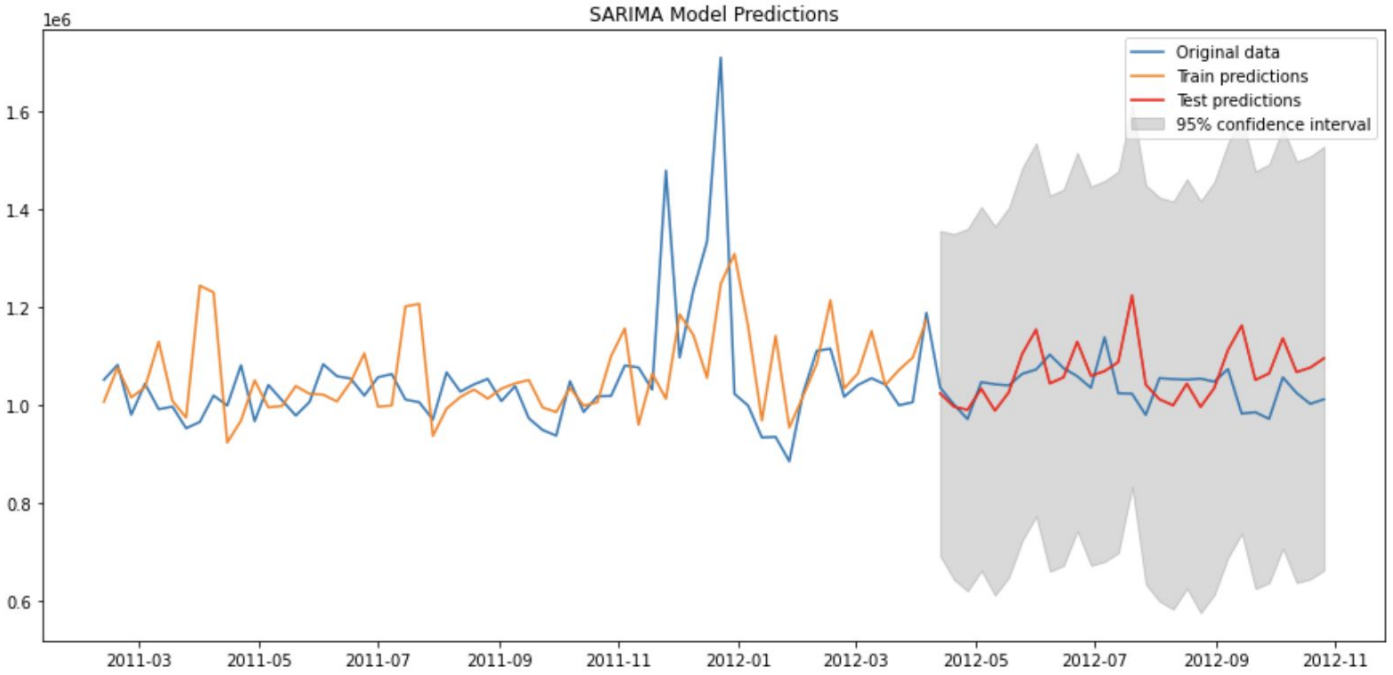
Covariance Type: opg

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	0.5690	0.322	1.770	0.077	-0.061	1.199
ar.L2	0.7861	0.360	2.186	0.029	0.081	1.491
ar.L3	-0.3699	0.132	-2.799	0.005	-0.629	-0.111
ma.L1	-0.2087	0.358	-0.583	0.560	-0.911	0.493
ma.L2	-0.6549	0.332	-1.971	0.049	-1.306	-0.004
ma.S.L15	-0.7346	0.210	-3.496	0.000	-1.146	-0.323
sigma2	2.868e+10	1.81e-11	1.58e+21	0.000	2.87e+10	2.87e+10

Ljung-Box (L1) (Q):	1.09	**Jarque-Bera (JB):**	43.93
Prob(Q):	0.30	**Prob(JB):**	0.00
Heteroskedasticity (H):	0.69	**Skew:**	0.73
Prob(H) (two-sided):	0.29	**Kurtosis:**	5.92

SARIMA Model

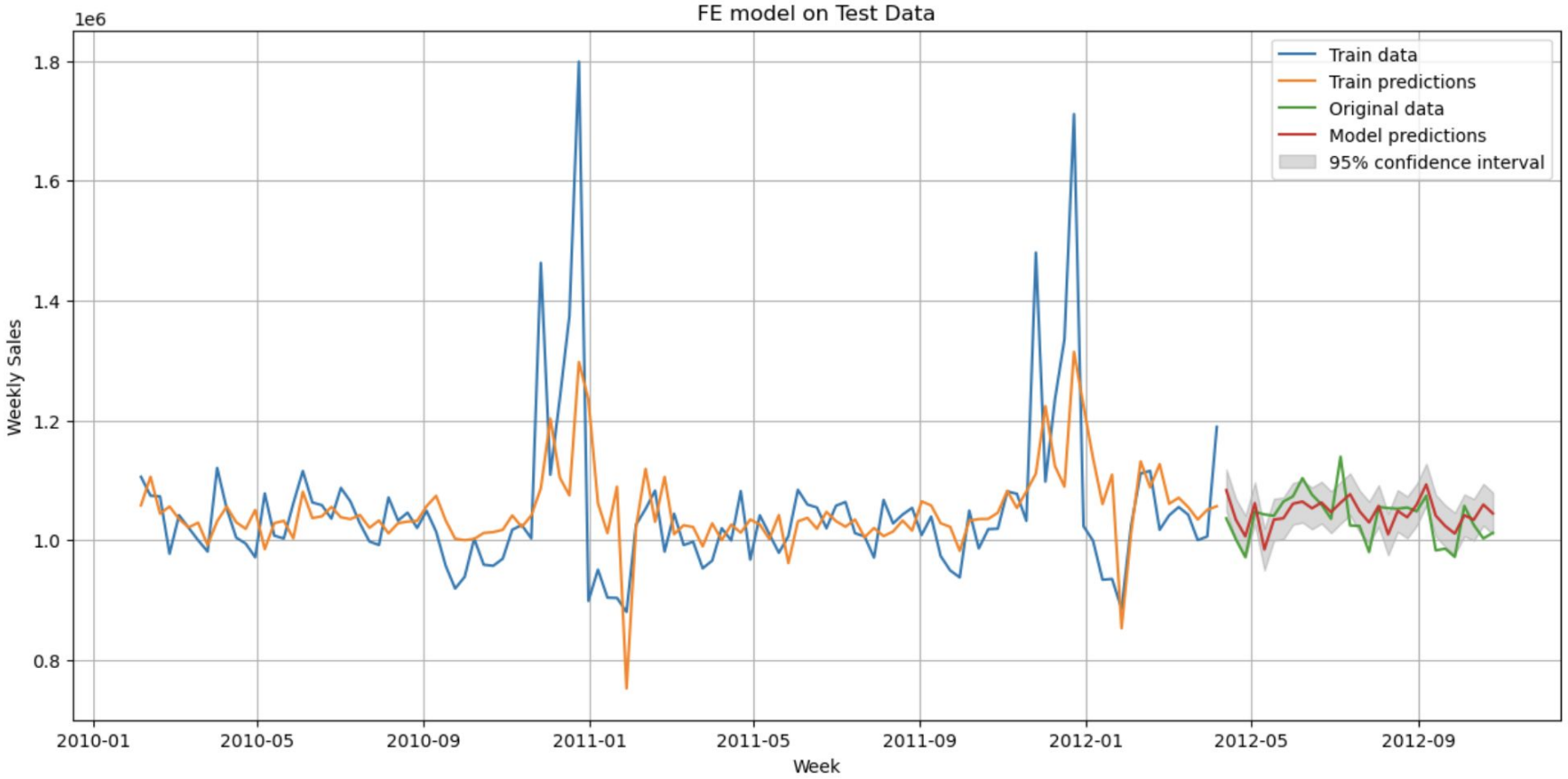
	MAPE
train	0.077369
test	0.055078



Walmart

Fixed and Random Effects

MAPE : 0.028


FE model on Test Data
1e6
1.8
1.6
1.4
1.2
1.0
0.8
Weekly Sales
2010-01
2010-05
2010-09
2011-01
2011-05
2011-09
2012-01
2012-05
2012-09
Week
Train data
Train predictions
Original data
Model predictions
95% confidence interval

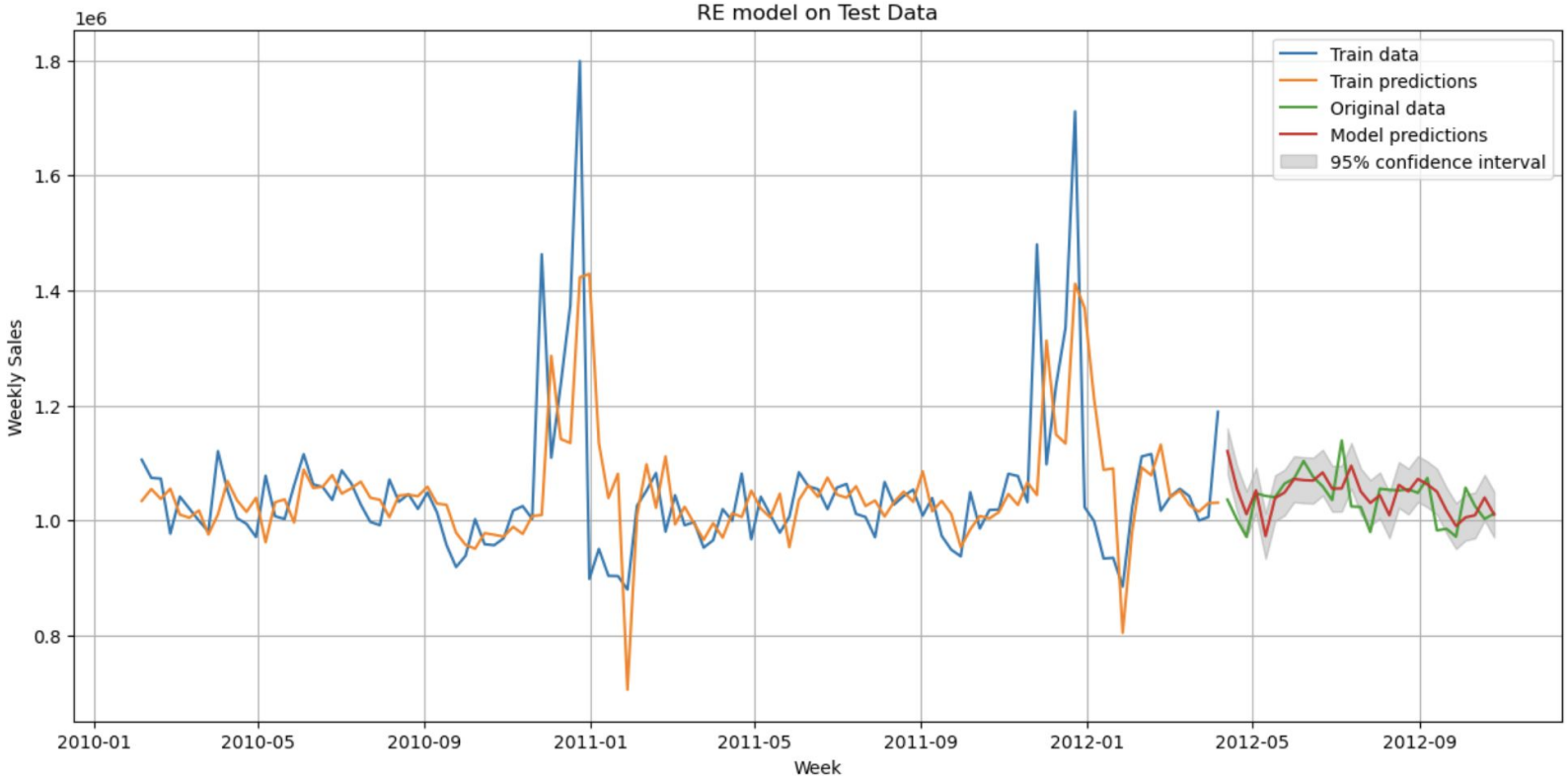
PanelOLS Estimation Summary

Dep. Variable:	Weekly_Sales	R-squared:	0.2623
Estimator:	PanelOLS	R-squared (Between):	0.3792
No. Observations:	5119	R-squared (Within):	0.2623
Date:	Sun, Dec 11 2022	R-squared (Overall):	0.3673
Time:	05:35:10	Log-likelihood	-6.829e+04
Cov. Estimator:	Unadjusted		
		F-statistic:	112.43
Entities:	45	P-value	0.0000
Avg Obs:	113.76	Distribution:	F(16,5058)
Min Obs:	103.00		
Max Obs:	114.00	F-statistic (robust):	112.43
		P-value	0.0000
Time periods:	114	Distribution:	F(16,5058)
Avg Obs:	44.904		
Min Obs:	44.000		
Max Obs:	45.000		

Parameter Estimates

	Parameter	Std. Err.	T-stat	P-value	Lower CI	Upper CI
const	7.266e+05	1.713e+04	42.407	0.0000	6.93e+05	7.602e+05
Holiday_Flag	4.965e+04	8850.3	5.6102	0.0000	3.23e+04	6.7e+04
Temperature	-8461.9	2679.3	-3.1583	0.0016	-1.371e+04	-3209.3
Fuel_Price	-1.448e+04	3544.1	-4.0859	0.0000	-2.143e+04	-7532.8
CPI	1.681e+05	5.424e+04	3.0991	0.0020	6.176e+04	2.744e+05
Unemployment	-4.115e+04	1.154e+04	-3.5652	0.0004	-6.378e+04	-1.852e+04
Lagged_1	0.3399	0.0138	24.666	0.0000	0.3129	0.3670
Lagged_2	0.0630	0.0145	4.3546	0.0000	0.0346	0.0913
Lagged_3	-0.0747	0.0145	-5.1603	0.0000	-0.1030	-0.0463
Lagged_4	0.2851	0.0145	19.721	0.0000	0.2568	0.3135
Lagged_5	-0.3384	0.0149	-22.650	0.0000	-0.3677	-0.3091
Lagged_6	0.0108	0.0158	0.6806	0.4962	-0.0203	0.0418
Lagged_7	0.0604	0.0154	3.9122	0.0001	0.0301	0.0907
Lagged_8	-0.0566	0.0143	-3.9602	0.0001	-0.0846	-0.0286

MAPE: 0.03


RE model on Test Data
1e6
1.8
1.6
1.4
1.2
1.0
0.8
Weekly Sales
2010-01
2010-05
2010-09
2011-01
2011-05
2011-09
2012-01
2012-05
2012-09
Week
Train data
Train predictions
Original data
Model predictions
95% confidence interval

RandomEffects Estimation Summary

Dep. Variable:	Weekly_Sales	R-squared:	0.9055
Estimator:	RandomEffects	R-squared (Between):	0.9980
No. Observations:	5119	R-squared (Within):	0.0111
Date:	Sun, Dec 11 2022	R-squared (Overall):	0.9055
Time:	05:36:27	Log-likelihood	-6.909e+04
Cov. Estimator:	Unadjusted		
		F-statistic:	3054.1
Entities:	45	P-value	0.0000
Avg Obs:	113.76	Distribution:	F(16,5102)
Min Obs:	103.00		
Max Obs:	114.00	F-statistic (robust):	3054.1
		P-value	0.0000
Time periods:	114	Distribution:	F(16,5102)
Avg Obs:	44.904		
Min Obs:	44.000		
Max Obs:	45.000		

Parameter Estimates

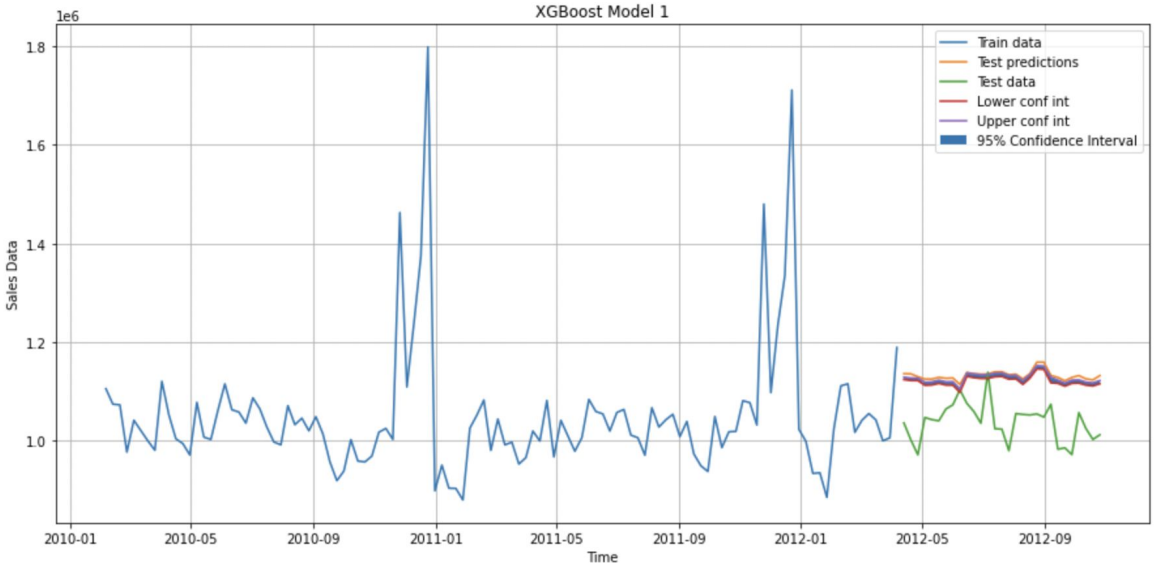
	Parameter	Std. Err.	T-stat	P-value	Lower CI	Upper CI
const	2.432e+04	5477.6	4.4403	0.0000	1.358e+04	3.506e+04
Holiday_Flag	7569.0	1.015e+04	0.7454	0.4561	-1.234e+04	2.748e+04
Temperature	3800.1	2634.9	1.4422	0.1493	-1365.5	8965.6
Fuel_Price	1805.9	2524.2	0.7154	0.4744	-3142.7	6754.5
CPI	-3934.7	2727.2	-1.4428	0.1491	-9281.1	1411.8
Unemployment	-3738.7	2689.3	-1.3902	0.1645	-9010.9	1533.6
Lagged_1	0.6154	0.0142	43.322	0.0000	0.5875	0.6432
Lagged_2	0.1726	0.0166	10.415	0.0000	0.1401	0.2050
Lagged_3	-0.0268	0.0168	-1.5941	0.1110	-0.0597	0.0062
Lagged_4	0.3375	0.0168	20.127	0.0000	0.3046	0.3703
Lagged_5	-0.3513	0.0174	-20.206	0.0000	-0.3853	-0.3172
Lagged_6	0.0627	0.0184	3.4162	0.0006	0.0267	0.0987
Lagged_7	0.1333	0.0179	7.4652	0.0000	0.0983	0.1683
Lagged_8	-0.0384	0.0166	-2.3095	0.0210	-0.0710	-0.0058
Lagged_9	0.1035	0.0168	6.1626	0.0000	0.0706	0.1364

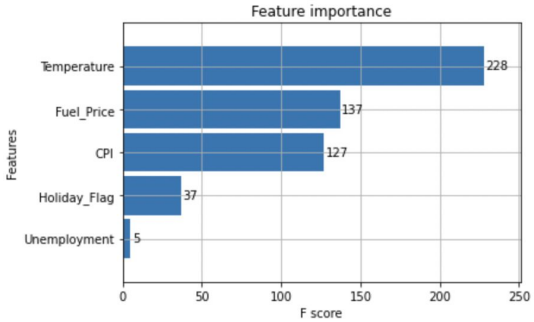
Walmart

Gradient Boosted Trees

XGBoost Model 1

Simple model using XGBRegressor.

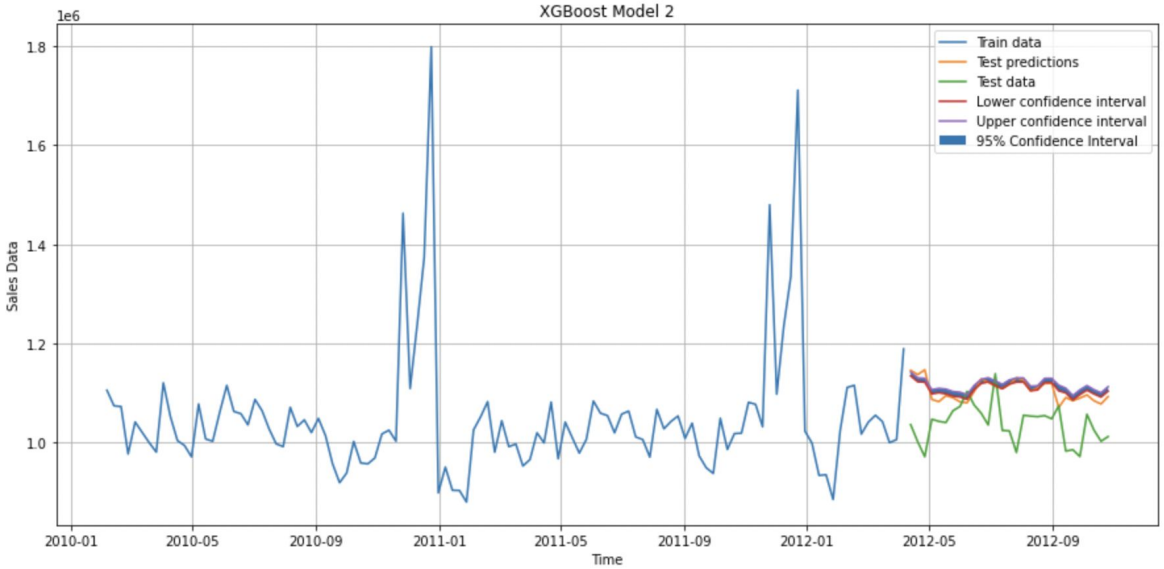


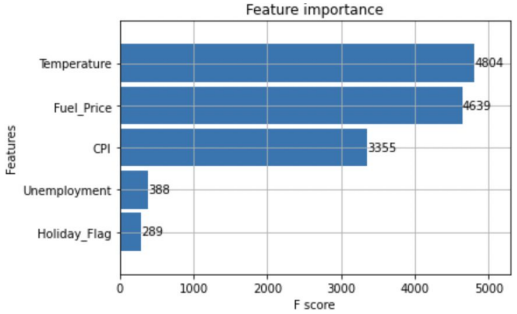



```
MAPE = 0.11378282760159741
```

XGBoost Model 2

XGBRegressor with hyperparameter-tuning using GridSearchCV()





MAPE = 0.06937198322581578

Walmart

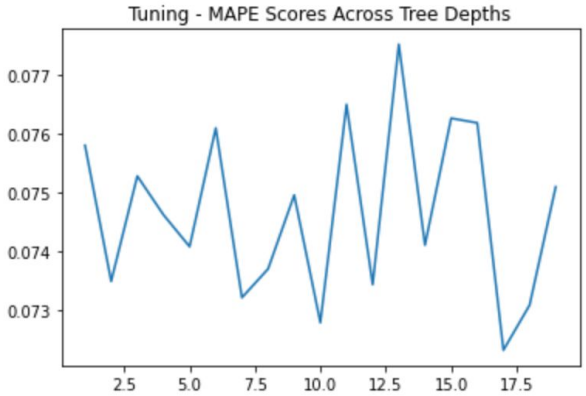
Random Forest Models

Random Forest - Tuning

- Initial performance:

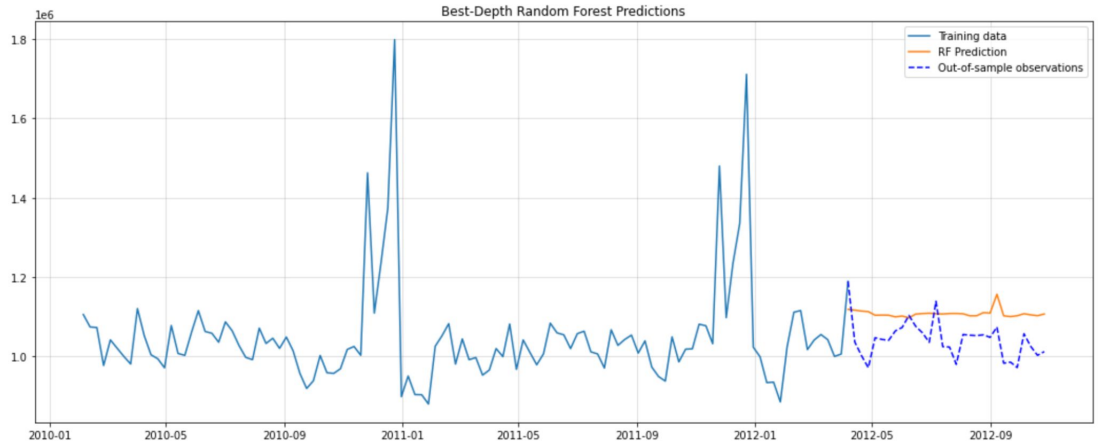
```
The random forest of depth-5 and 1000 trees achieves the following MAPE scores:

        train   0.0354
        TEST    0.0727
```

- Hyperparameter tuning for optimal tree depth:

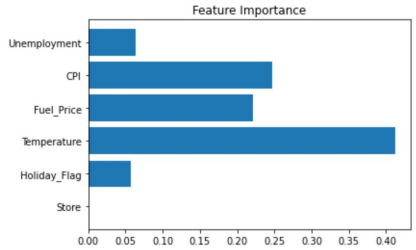


Random Forest - Results

- Tuned model performance
  - MAPE: 0.072



- Feature Importance:

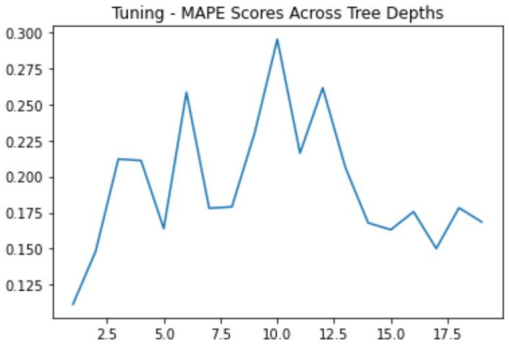


Random Forest - Autoregressive

- New model trained not on features, but past weekly sales
- Initial performance:

```
The ARF model with depth 5 gives a MAPE of 0.164
```

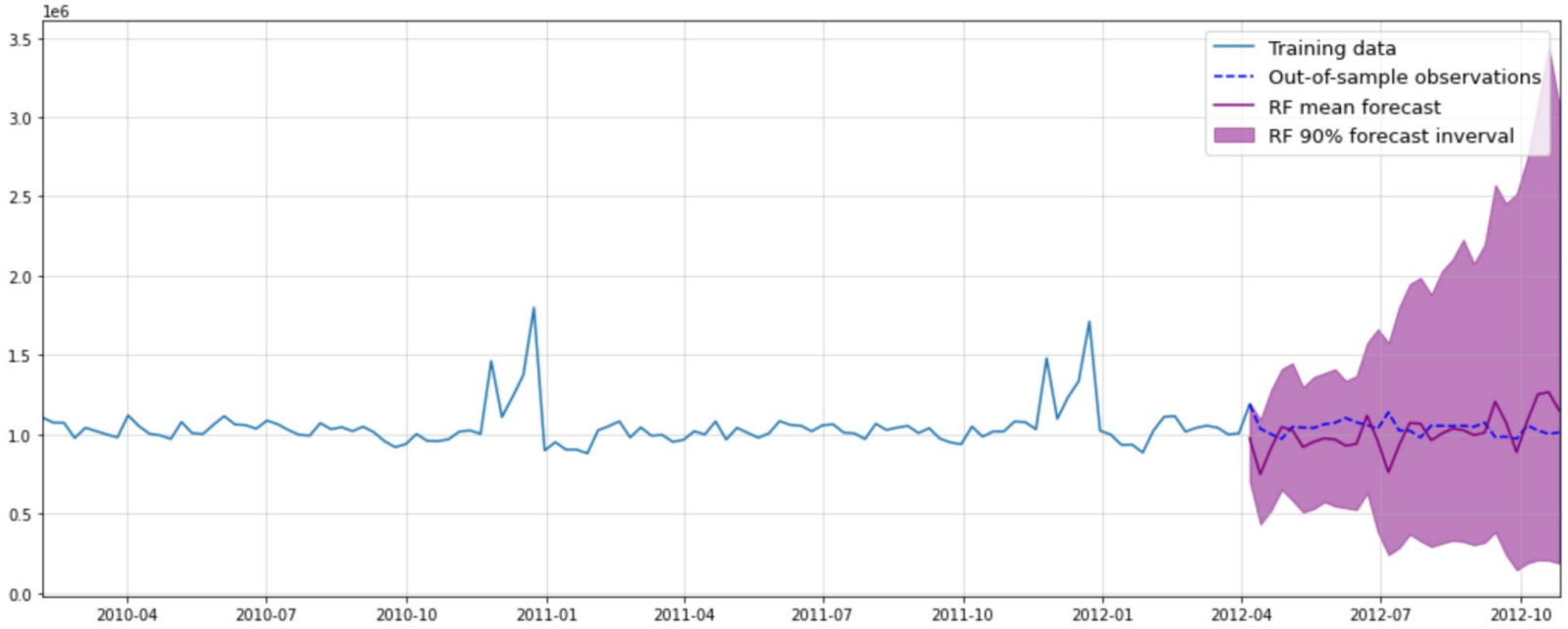
- Tuning:



- Performs best at depth of 1, makes sense since only fitting on one feature

Random Forest - Autoregressive Results

- Tuned model performance
  - MAPE: 0.11



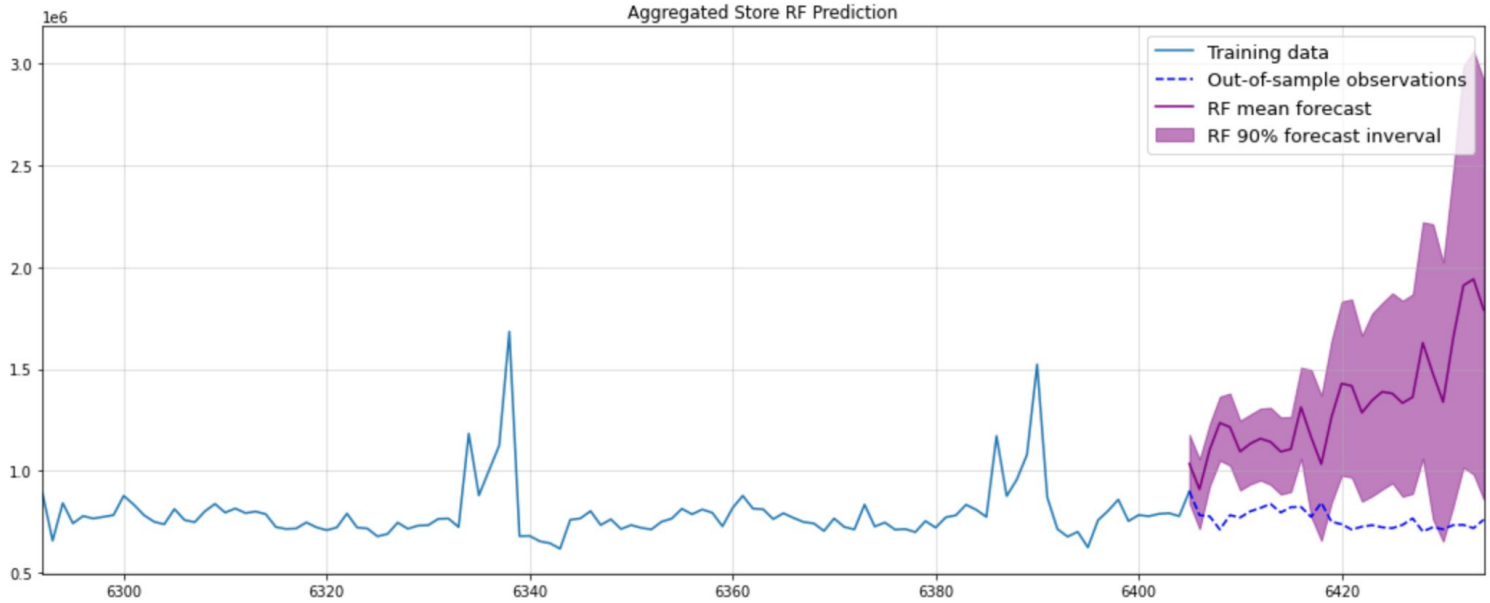
Random Forest - By Store

- Last attempt to fit RF: fit each store to its own model
- MAPEs across stores vary widely:

	store	MAPE
0	1	0.229820
1	2	0.356865
2	3	0.621875
3	4	0.442872
4	5	0.617666
5	6	0.439791
6	7	0.823693
7	8	0.549803
8	9	0.681451
9	10	0.990992

Random Forest - By Store Results

- Aggregated performance of by-store models
- Significant overestimation
  - Likely due to the model being an even worse fit for higher-revenue stores



Walmart

Overall Results

Comparison of Errors

Model	Mean Absolute % Error
Baseline (Seasonal Lag)	**0.026**
Moving Average	0.037
Seasonal Moving Average	0.026
Seasonal Exponentially Weighted	0.024
SARIMA	0.055
Fixed Effects	0.028
Random Effects	0.030
XGBoost	0.069
Random Forest	0.072

Conclusions

- To begin with, the baseline performs extremely well, indicating average weekly sales is driven largely by annual seasonal patterns in this period, and seasonal differencing renders the data extremely stationary
- The seasonal exponentially weighted average model is the best model, it accounts for annual seasonality and further weighs more recent observations more heavily
- Simplicity is king – the simplest prediction models performed much better than more complex models like trees and forests, which makes sense in the context of limited data
- Complex models need more data to train on to capture all relevant patterns and make inferences; over longer horizons they should outperform simple models

Future Work

Recommended Future Work

- Collect more data over a longer period for fitting more complex models – outside of 2010-2012, the seasonality regime is unlikely to hold
- Tune and test more by-store prediction models, ideally with more store-specific features relevant to location and products available
    - It is possible that more “luxury” products are correlated with macroeconomic factors, as opposed to “commodity” products
- Consider regularization strategies to prevent more complex models from overfitting on training data
- Aggregate Walmart data with alternative data on other comparable retail chains such as Target, Costco, etc. to identify correlations