Retail Sales Forecasting: Autoregression Techniques to Predict Store Sales

Victoria Hall

July 3, 2022

Retail Sales Forecasting: Autoregression Techniques to Predict Store Sales

The purpose of this project was to find the best modeling technique to predict sales for a chain of retail stores. The data, which came from Kaggle.com (Singh, 2017) was cleaned and prepped, then used to test a series of time series modeling techniques. The results of the best performing model should be used to help guide recommended actions that will increase bottom line.

Introduction

Business Problem

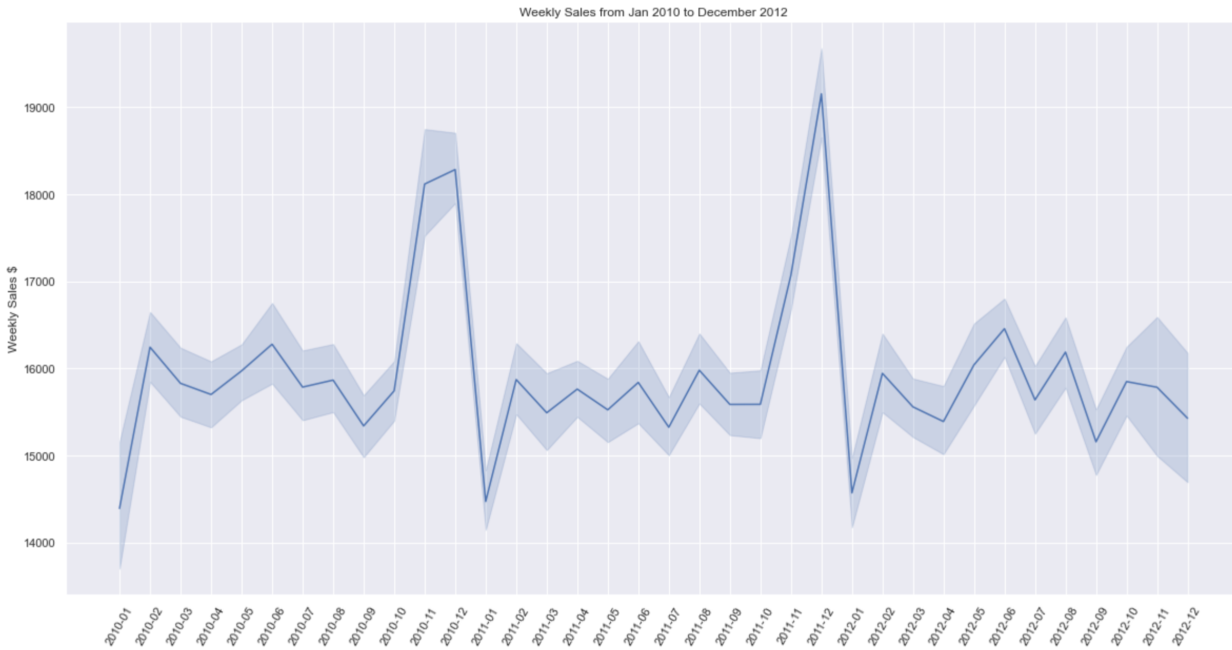
There are several factors that can impact retail sales. Some of these factors are controllable like managing discount events, maintaining proper inventory levels, and initial price point. There are other factors, like weather and the macroeconomic environment that are not so easy to control. For a model to be high performing at predicting retail sales, it will need to take in account the factors that most strongly impact these sales. Without accurate forecasts, a company will be at risk for over or under buying inventory, buying the wrong kind of product, inappropriately running discount events, and ultimately making decisions in the dark that could risk bottom line sales. The purpose of this project is to build, teach and test models so that ultimately, the strongest performing model can be used by stakeholders to better inform their decisions.

History and Data

The data for this project was sourced from Kaggle (Singh, 2017). The dataset was published five years ago with the challenge of modeling retail data in order to make decisions with a limited history of information. The original intention of the challenge was to predict department level sales; however, I have opted to instead to model on a store level. The data is made up of three separate datasets about 45 anonymized stores located in different regions. One dataset features detail about the pricing activities in store as well as exogenous factors like temperature, fuel price, unemployment rate, the consumer price index. The stores dataset contains features about each store and the sales dataset provides sales for each store. These details are provided at a weekly level for each individual store between January 2010 and December 2012.

Before any modeling could be tested, I first prepped the data. Because of the weekly nature of the information, this data needed to be prepped for time series analysis. My first step was to examine the separate datasets and then merge into one data frame for exploratory analysis. The features data had just over 8,00 rows of the data while the store and sales data had 45 and 421,570 rows, respectively. The only columns with any missing information were the columns describing the markdowns taken. These values were filled with zeroes based on the assumption that if the information was missing, there must not have been any markdown activity for that given store in that given week. The last step in the data preparation was to convert any dates to datetime objects.

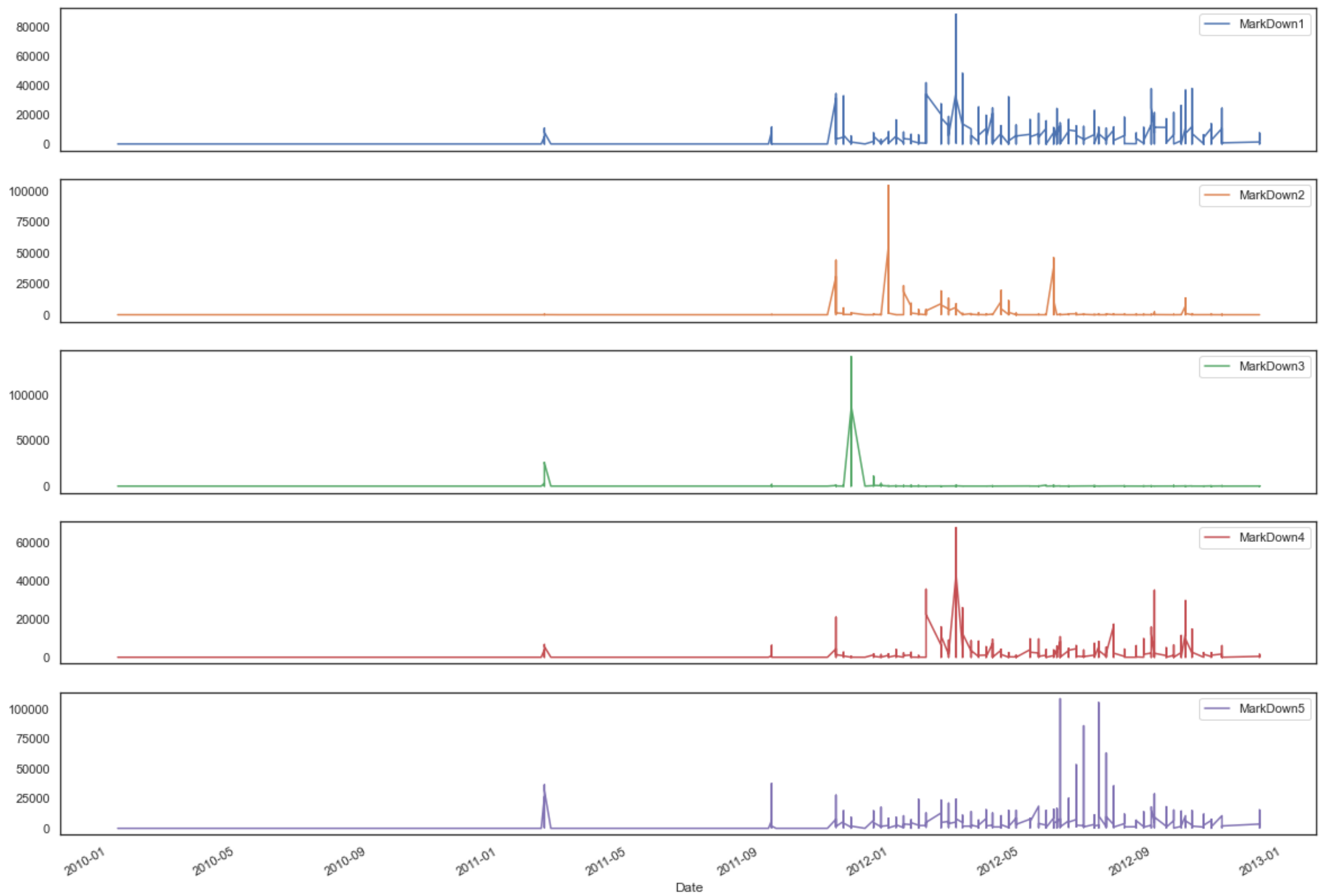
This data preparation allowed for an exploratory phase of analysis. The average weekly sales for each store was $16,000. There were a couple stores that significantly outperformed this, but these stores were almost three times as large as the average store, meaning sales volume was naturally higher with more inventory. Average weekly sales tended to stay between 15 and 17,000 except for a couple large spikes as seen the graph below.



These spikes occur around the holiday seasons in November with sudden drop offs afterwards. Interestingly, this spike does not seem to occur in 2012 as it did in 2011 and 2010.

The trends among the features of the data show what would be expected for the most part. Fuel prices had slight changes week to week but overall increased, temperatures followed seasonality, and unemployment and CPI (consumer price index) mostly moved in tandem with small changes week to week except unemployment rates did seem to be decreasing overall.

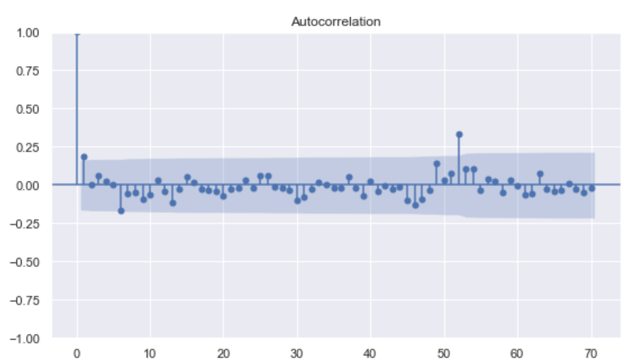
Markdown activity was mostly nonexistent until the beginning of 2012 as demonstrated in the graph below. This activity occurs through the rest of 2012. This markdown activity coincides with a decrease in sales volume. It appears that markdown activity may have increased to try and right size sales.

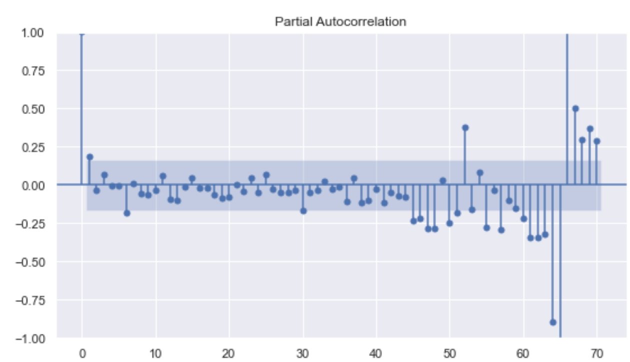


Methods, Testing, and Analysis

There are some common methods when it comes to modeling time series data. The two baseline ideas behind more intensive modeling are autoregression and moving averages. Autoregression assumes that observations at previous time steps will help predict the value at the next step. This type of correlation, autocorrelation, helps the model learn how much weight apply to a variable that correlates to the output variable (Brownlee, 2020). This autocorrelation informs a linear regression model to help forecast future events. A moving average model calculates the residuals of error of the past series of events to calculate the future of the event (Shetty, 2020). More advanced models combine these two methods which is what was used in this time series analysis. SARIMA and SARIMAX modeling techniques take into account seasonality in combination with autoregressive and moving average techniques. SARIMAX goes one step further and includes exogenous variables in this model outside of just the weekly time series data (Brownlee, 2019).

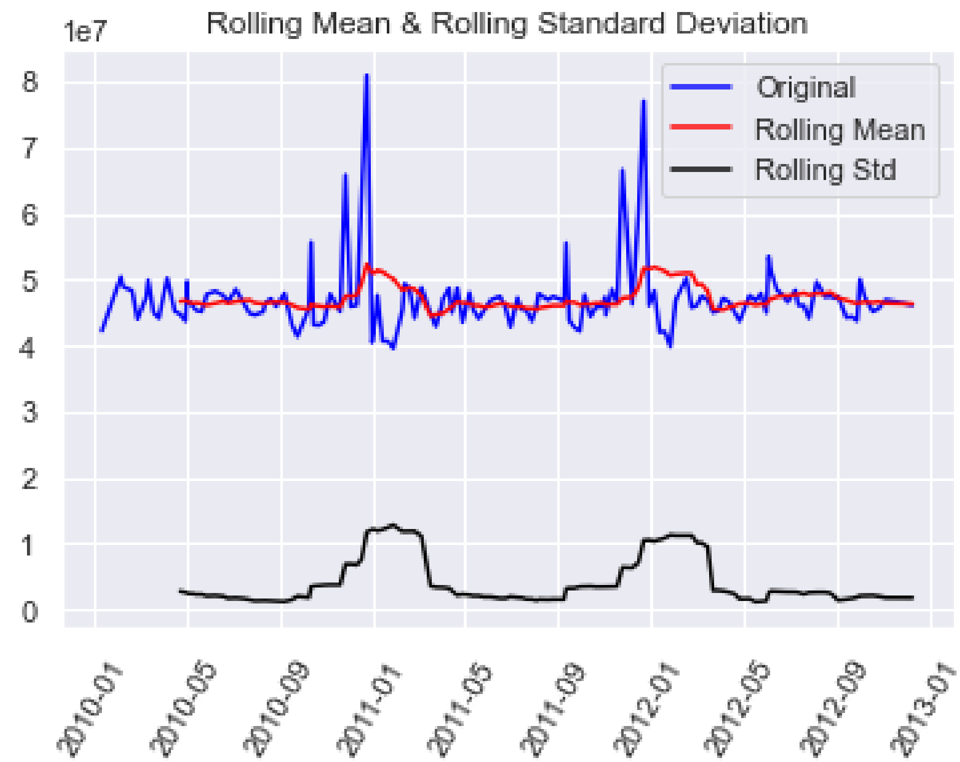
Before implementing the SARIMA or SARIMAX models, I checked to make sure the data would meet the underlying assumptions of these techniques. Autocorrelation and partial autocorrelation can often become apparent by visualizing the functions over the data. Autocorrelation includes direct and indirect correlations between an observation and the past observations in the series. Partial autocorrelation removes these indirect correlations. Both tell information about the strength of a lag of k for the time series, but autocorrelation shows that relationship would carry on for future correlations until the effect weakens. Partial correlation suggests that there is no correlation for lags beyond k (Brownlee, 2021). In this case, the ACF and PACF (functions of autocorrelation and partial autocorrelation) shown below indicate there is a seasonality aspect to this data, which was previously observed in the exploratory analysis process. This indicates these models that consider seasonality are the right choice. These functions also help to inform the order of terms in the model equation.





The final assumption test was checking the series was stationary. A time series is stationary if it has a constant mean, and the variance and covariance is independent of time

(Chhugani, 2020). On visual inspection, the series appeared to be stationary without any overarching trend in one direction or the other. I also ran a test an Adfuller test to ensure this, and the data is stationary (Gupta, 2020).



The validation of these assumptions allowed for model testing to begin. First SARIMA was tested with a series of terms that was determined by reviewing the ACF and PACF plots. The first test was on the whole chain of stores. The scoring factor here which is the R squared value, was .41 and any residuals were not significant. This value shows that the model might not be fitting the best line to the data. The closer the R squared value is to one the better the model explains the data. However, the insignificant residuals show that the errors were independent of one another. I also evaluated this model on the largest store individually rather than the combined chain. The results were worse for this store with an R2 value of .34.

SARIMA modeling was not achieving the results I hoped for, so SARIMAX was tested. This modeling considers exogenous features like markdown activity, fuel prices, etc. Upon initial inspection temperature had the strongest correlation with sales. When running the SARIMAX model, the number of lags was expanded to try and account for more seasonality. The same store that was previously tested was also tested with the SARIMAX model. The model score increased to .58. This value shows that the SARIMAX model outperforms the SARIMA model. The SARIMAX model for the entire chain resulted in an R2 of .69 again with insignificant residuals. I expanded this model and ran trained it individually for each of the 45 stores. The results of the model testing for each store can be found in the appendix (ADD APPENDIX).

When reviewing on a store level, there are certain stores where the model performed better and stores where it did not. In some store instances, the residuals were higher than would be acceptable for implementation of this model. But overall, the SARIMAX model was the better performing of the two.

Conclusion

The final model of choice is the SARIMAX model. The model performance was greatly improved with the inclusion of exogenous variables, allowing the model to explain 70% of the variance in the data. However, the intent of the model was to forecasts stores at a store level. Some stores had low R2 values and high residuals. More than likely, those individual stores are violating one of the assumptions regarding the stationary aspect of the data or the autocorrelation. Some of these stores might have trends occurring that is not reflected in the overall chain. With more data, these problems may or may not resolve. Before any implementation process begins, each individual store should have autocorrelation, seasonality, and stationary aspects of the series explored. If this poses too much of a challenge, similar stores could be grouped, and the models could be tested on the specific groupings to find the appropriate forecasting model. Time series forecasting is tool used in multiple fields. It has applications beyond predicting revenue. It can be used in social sciences to predict behavior or in a field like meteorology to predict temperature. In this instance, it was used for sales forecasting which can have top and bottom-line business implications. Similarly, these forecasts can help plan out receipts and pricing, but it can be difficult to use these models to influence change on the forecast. For instance, temperature played the largest role in explaining variance but that is beyond the company's control. Also, sales are very dependent on human behavior and human behavior can be notoriously difficult to predict.

Overall, I don't see any ethical implications from these results as long the data remains anonymous and when it isn't anonymized, it is only being used by the internal company. As long as the company does not use these forecasts to manipulate investors or violate any SEC guidelines, ethically, it is fine to proceed. After the models are reviewed again to test if specific stores are violating assumptions, an implementation plan can be developed and begin. IT partners to help with building tools and user interface structures so key stakeholders in the business can use this information to make decisions. Similarly, a plan will need to be developed to watch for model degradation and setting thresholds to intervene after a certain level of performance is reached.

References

Brownlee, J. (2019, August 21). *A gentle introduction to sarima for time series forecasting in Python*. Machine Learning Mastery. Retrieved July 8, 2022, from https://machinelearningmastery.com/sarima-for-time-series-forecasting-in-python/

Brownlee, J. (2020, August 14). *A gentle introduction to autocorrelation and partial autocorrelation*. Machine Learning Mastery. Retrieved July 8, 2022, from https://machinelearningmastery.com/gentle-introduction-autocorrelation-partial-autocorrelation/

Brownlee, J. (2021, September 6). *Autoregression models for time series forecasting with python*. Machine Learning Mastery. Retrieved July 8, 2022, from https://machinelearningmastery.com/autoregression-models-time-series-forecasting-python/

Chhugani, R. (2020, December 22). *An overview of autocorrelation, seasonality and stationarity in time series data*. Analytics India Magazine. Retrieved July 8, 2022, from https://analyticsindiamag.com/an-overview-of-autocorrelation-seasonality-and-stationarity-in-time-series-data/

Gupta, M. (2020, April 20). *Adfuller test intuitive explanation for checking time series stationarity*. Medium. Retrieved July 8, 2022, from https://medium.com/data-science-in-your-pocket/adfuller-test-intuitive-explanation-for-checking-time-series-stationarity-1e37ce6d240e

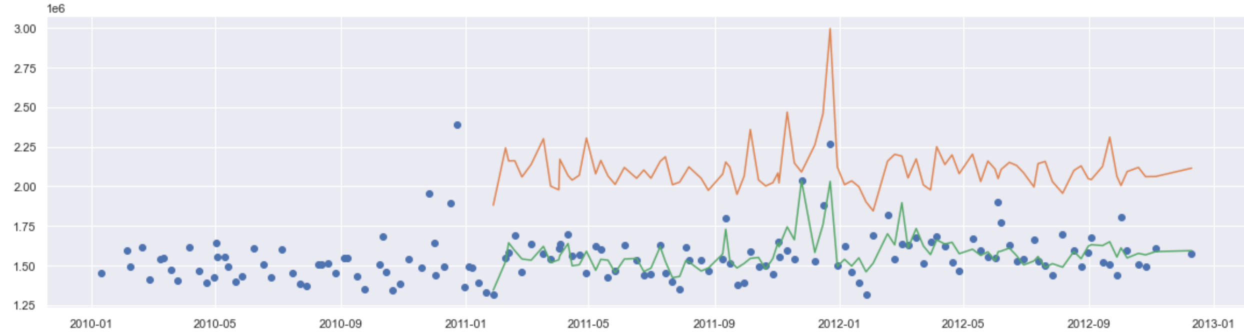
Shetty, C. (2020, September 22). *Time Series models*. Medium. Retrieved July 8, 2022, from https://towardsdatascience.com/time-series-models-d9266f8ac7b0

Singh, M. (2017, September 1). *Retail Data Analytics*. Kaggle. Retrieved July 8, 2022, from https://www.kaggle.com/datasets/manjeetsingh/retaildataset

Appendix

Store Number: 1

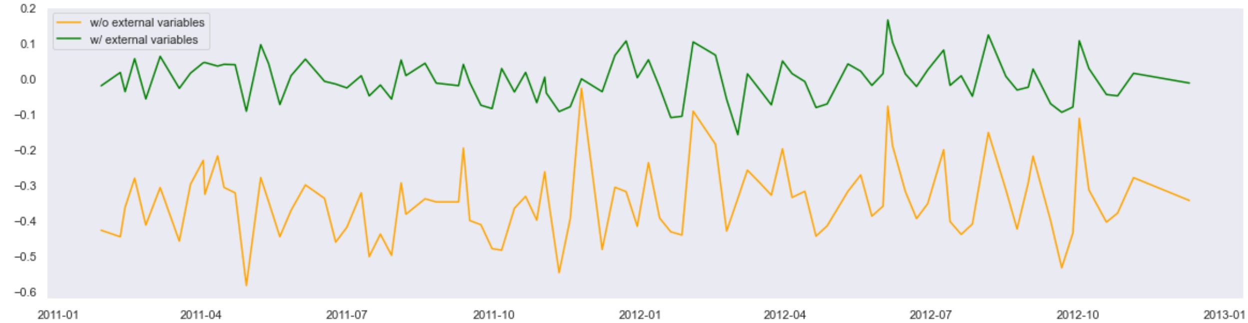
Score factor: 0.54



Store Number: 1

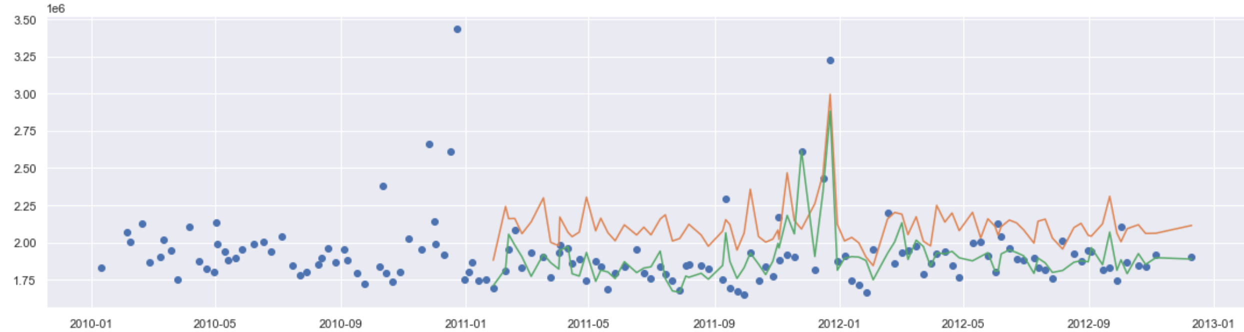
AR Residuals: avg -0.35, std 0.10

AR with Ext Residuals: avg -0.00, std 0.06



Store Number: 2

Score factor: 0.65



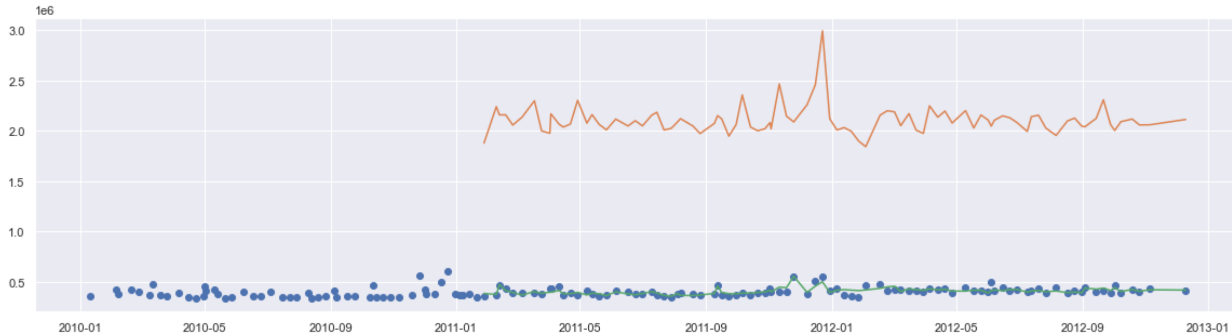
Store Number: 2

AR Residuals: avg -0.12, std 0.09

AR with Ext Residuals: avg -0.00, std 0.06

Store Number: 3

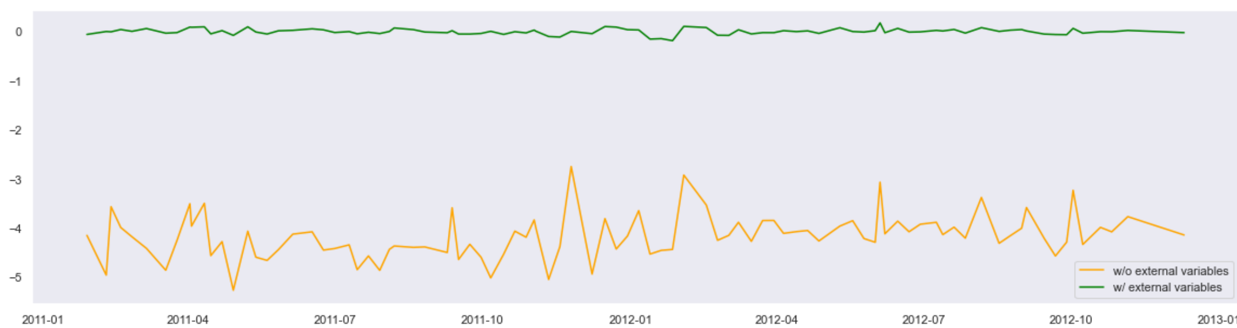
Score factor: 0.60



Store Number: 3

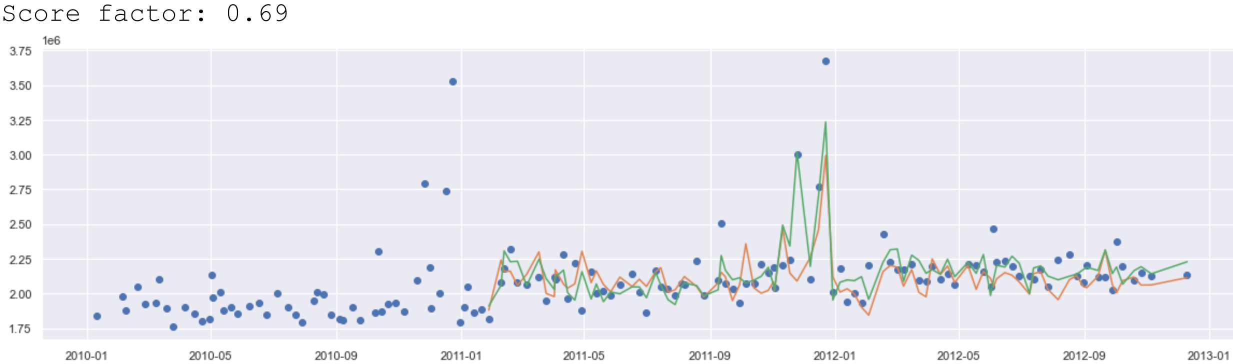
AR Residuals: avg -4.17, std 0.45

AR with Ext Residuals: avg -0.00, std 0.06



Store Number: 4

Score factor: 0.69

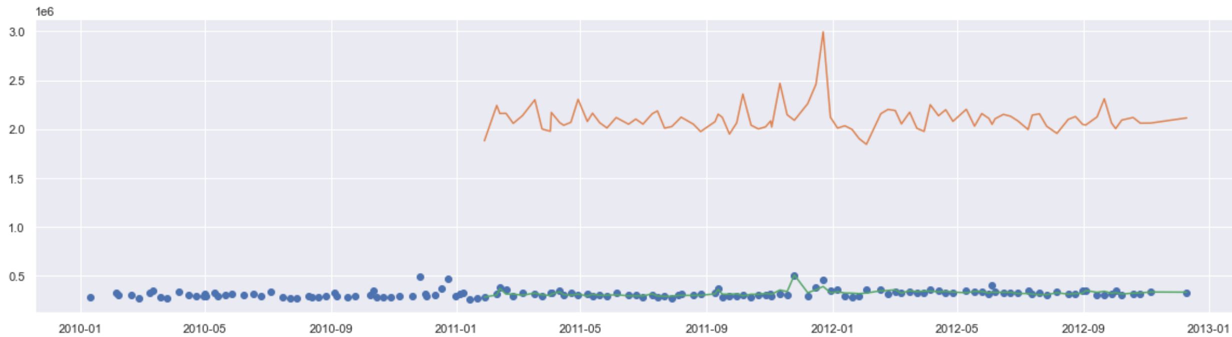


Store Number: 4

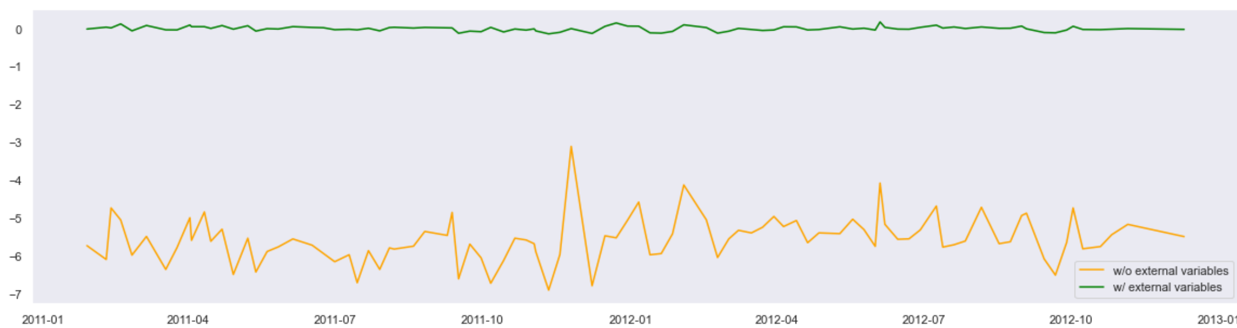
AR Residuals: avg 0.02, std 0.07

```
AR with Ext Residuals: avg -0.00, std 0.06
```

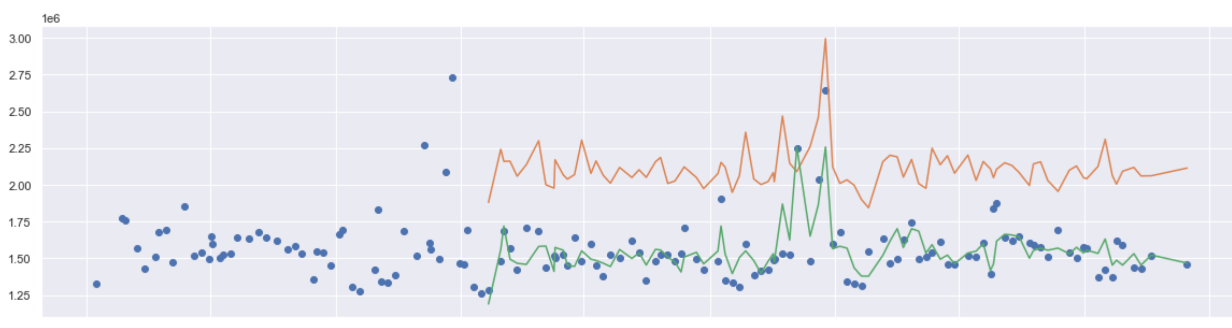

```
Store Number: 5
Score factor: 0.64
```




```
Store Number: 5
AR Residuals: avg -5.57, std 0.61
AR with Ext Residuals: avg -0.00, std 0.06
```



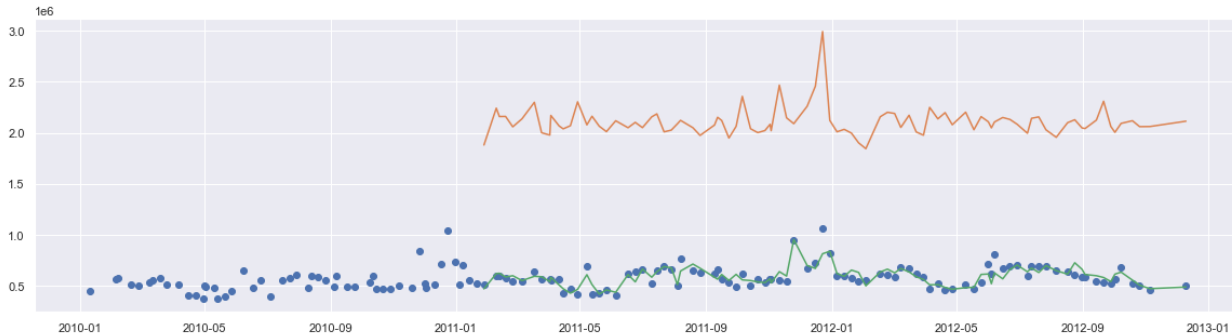

```
Store Number: 6
Score factor: 0.58
```



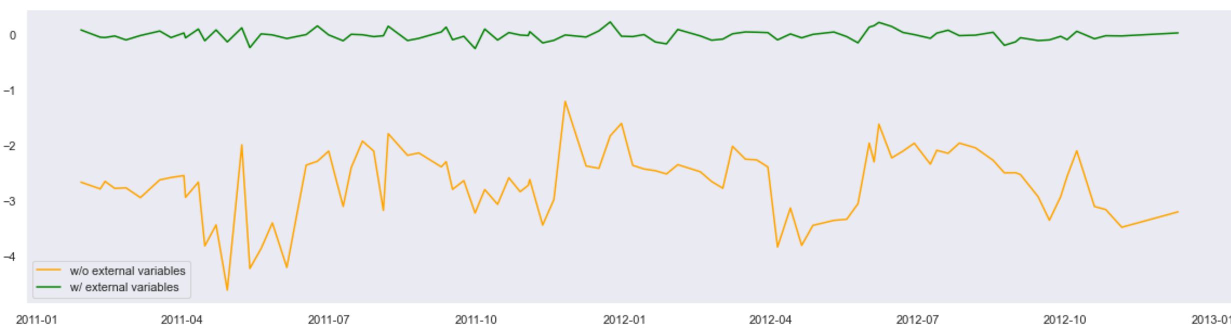

```
Store Number: 6
```

```
AR Residuals: avg -0.37, std 0.12
AR with Ext Residuals: avg -0.01, std 0.08
```

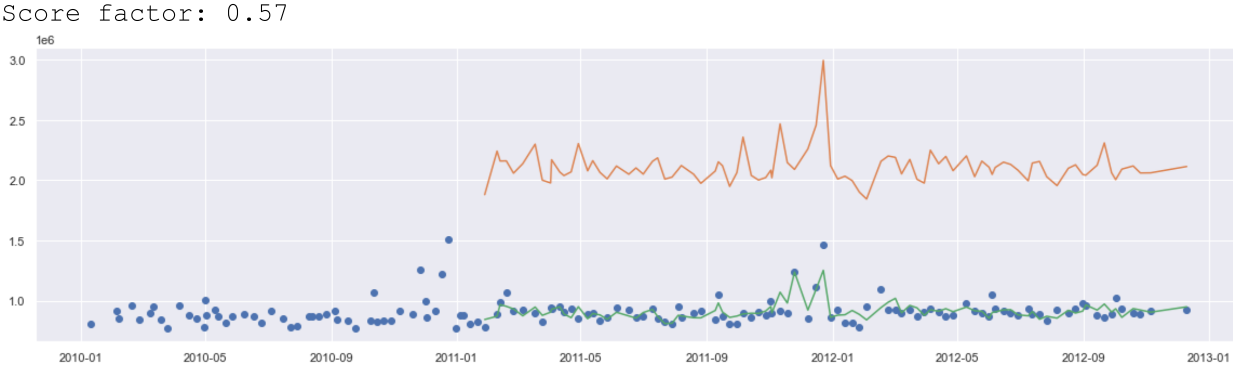

```
Store Number: 7
Score factor: 0.69
```



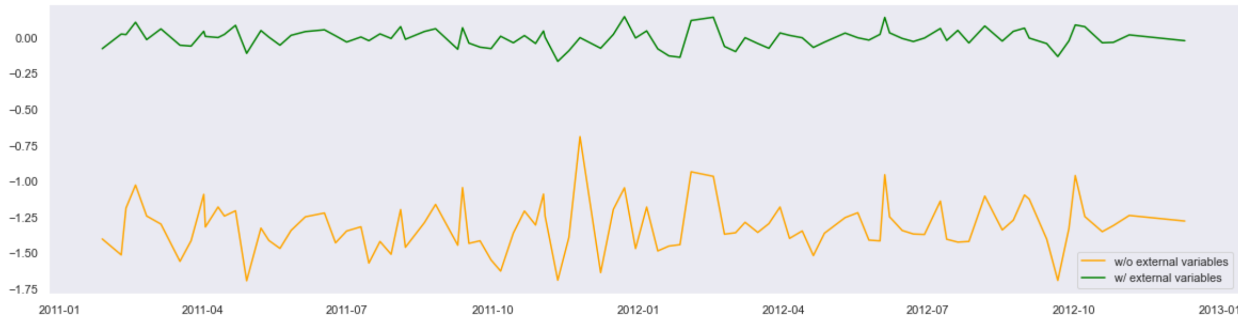
```
Store Number: 7
AR Residuals: avg -2.67, std 0.62
AR with Ext Residuals: avg -0.01, std 0.09
```



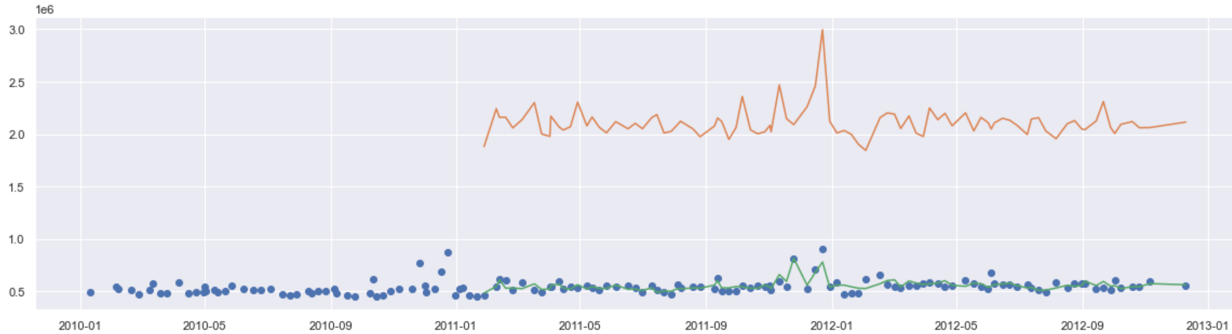

```
Store Number: 8
Score factor: 0.57
```



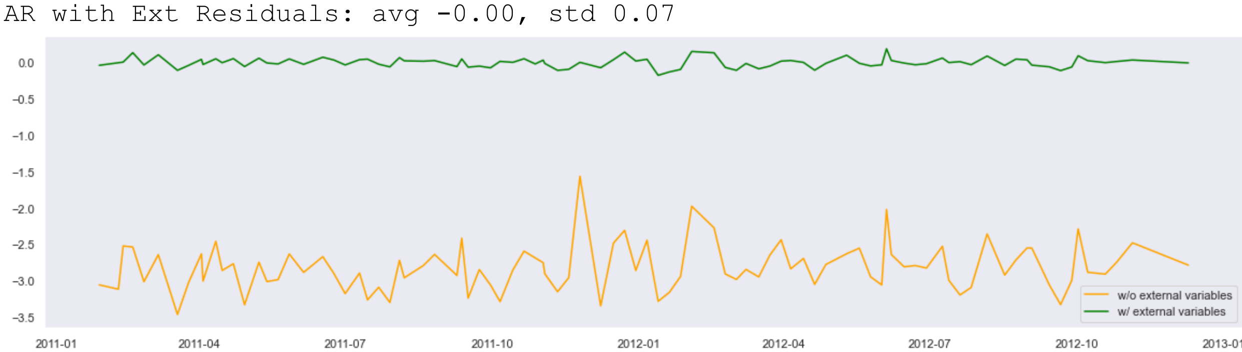
```
Store Number: 8
AR Residuals: avg -1.32, std 0.18
AR with Ext Residuals: avg -0.00, std 0.06
```



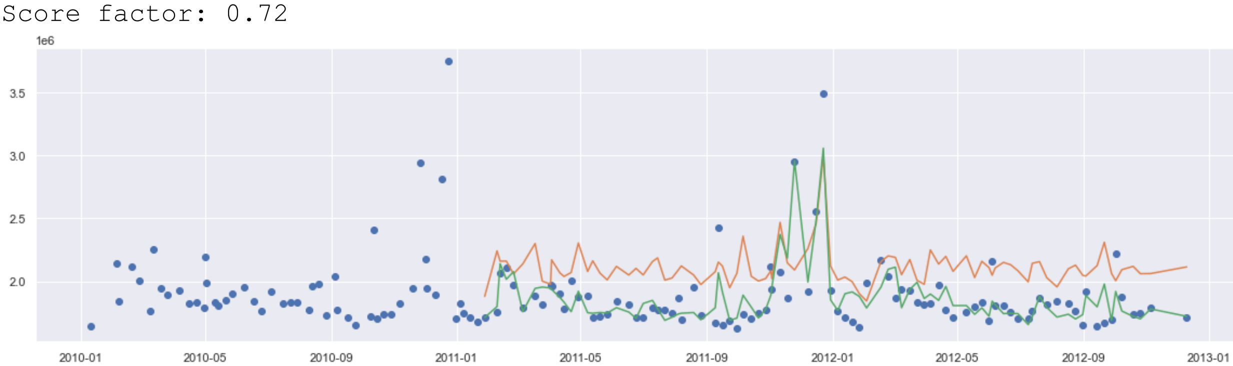

```
Store Number: 9
Score factor: 0.61
```



```
Store Number: 9
AR Residuals: avg -2.81, std 0.32
AR with Ext Residuals: avg -0.00, std 0.07
```

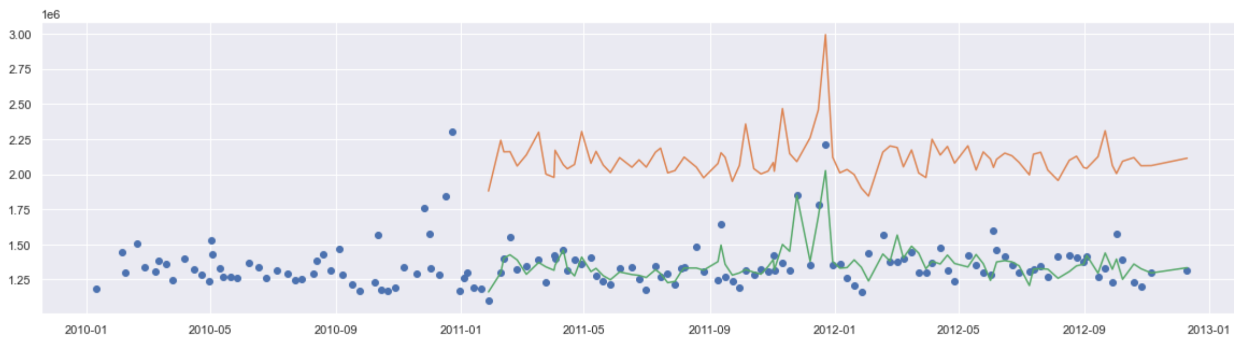



```
Store Number: 10
Score factor: 0.72
```

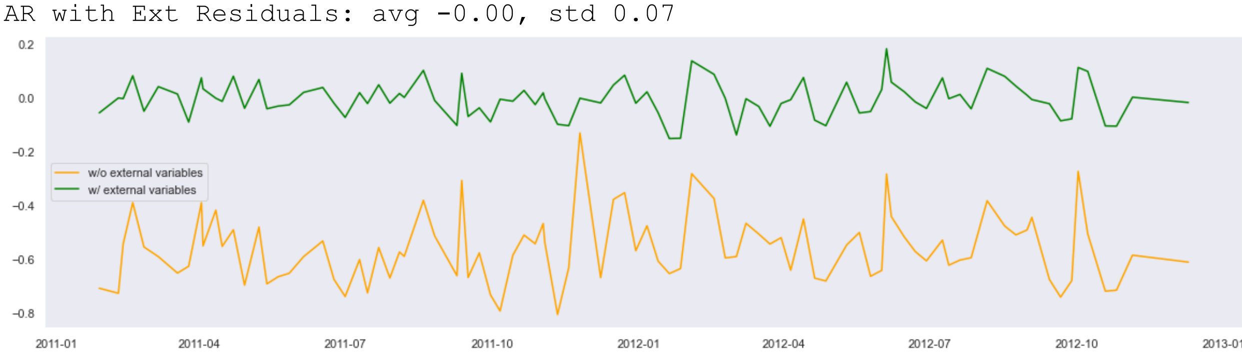


```
Store Number: 10
AR Residuals: avg -0.14, std 0.11
AR with Ext Residuals: avg -0.00, std 0.07
```

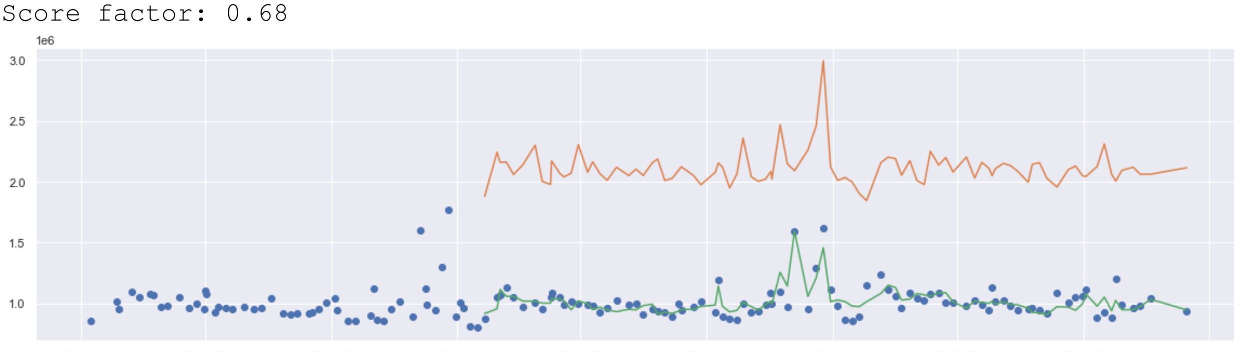

```
Store Number: 11
Score factor: 0.62
```



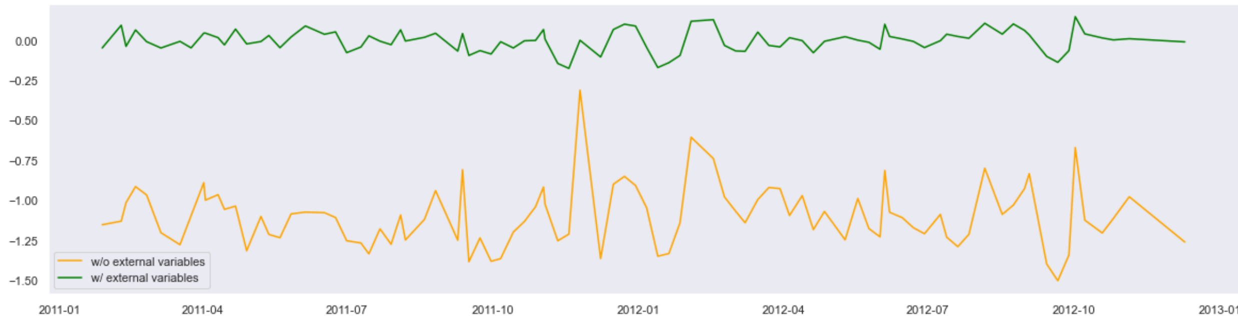

```
Store Number: 11
AR Residuals: avg -0.56, std 0.13
AR with Ext Residuals: avg -0.00, std 0.07
```



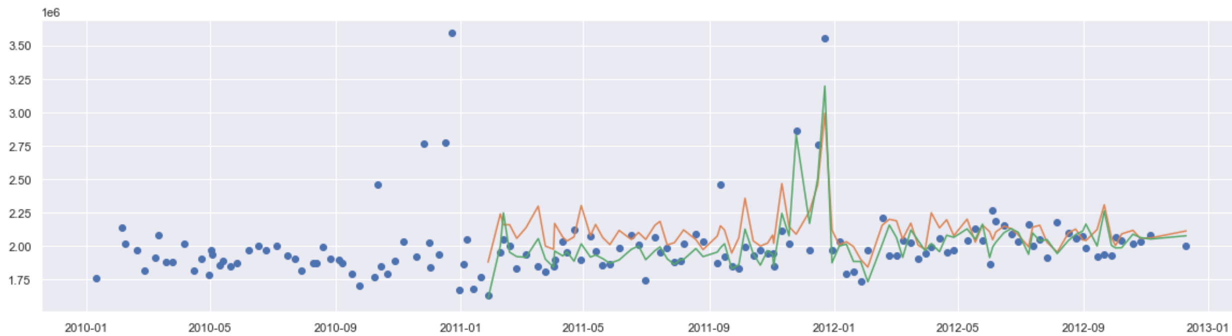

```
Store Number: 12
Score factor: 0.68
```



```
Store Number: 12
AR Residuals: avg -1.10, std 0.19
AR with Ext Residuals: avg -0.00, std 0.07
```

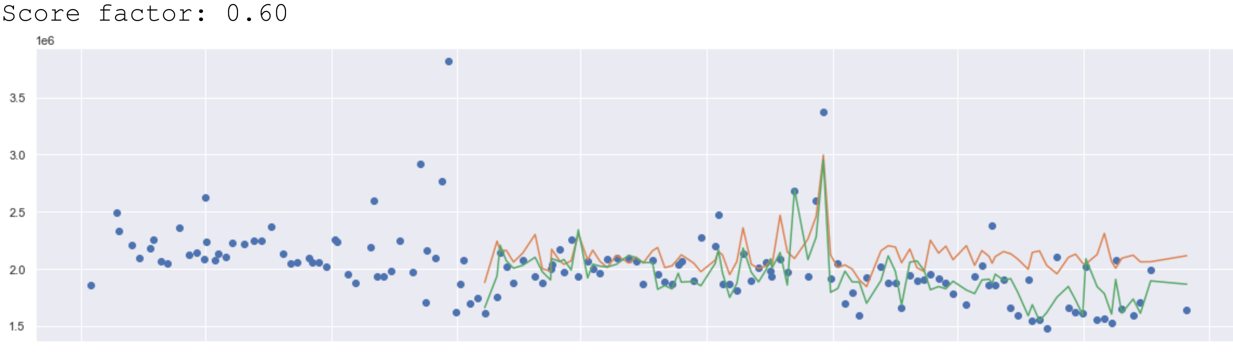



```
Store Number: 13
Score factor: 0.67
```



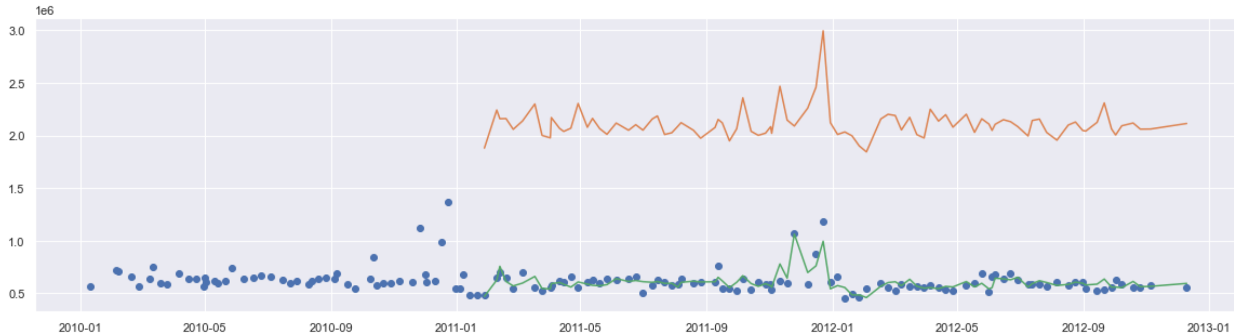
```
Store Number: 13
AR Residuals: avg -0.05, std 0.08
AR with Ext Residuals: avg -0.00, std 0.06
```


```
Store Number: 14
Score factor: 0.60
```

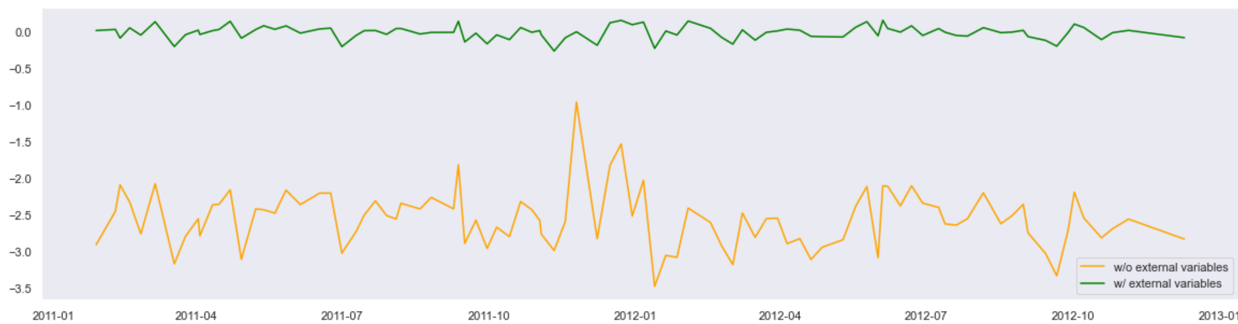


```
Store Number: 14
AR Residuals: avg -0.11, std 0.14
AR with Ext Residuals: avg -0.01, std 0.09
```

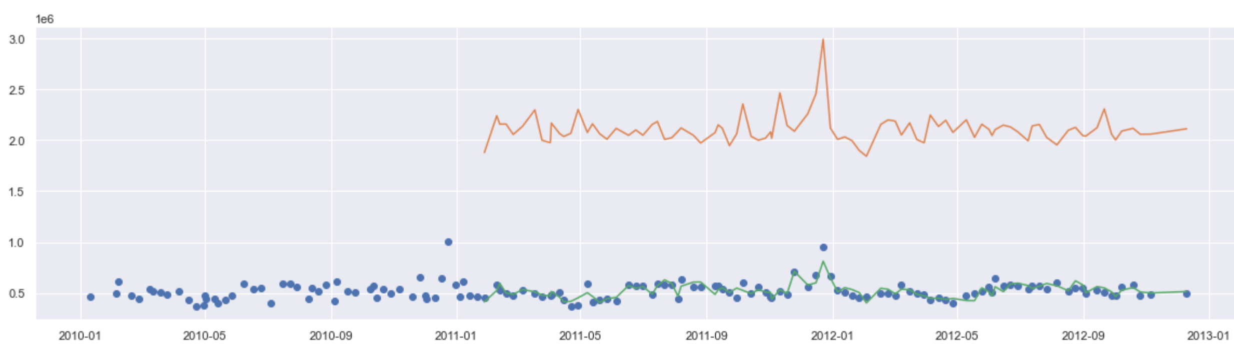

```
Store Number: 15
Score factor: 0.68
```



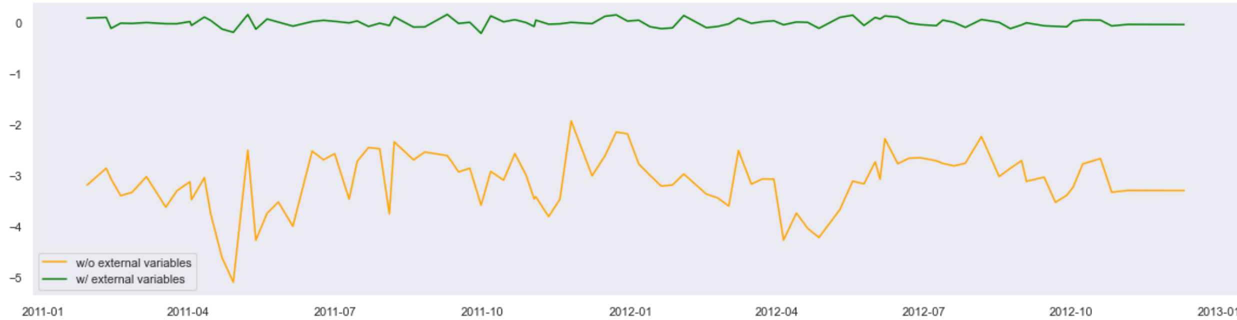
```
Store Number: 15
AR Residuals: avg -2.55, std 0.39
AR with Ext Residuals: avg -0.01, std 0.09
```



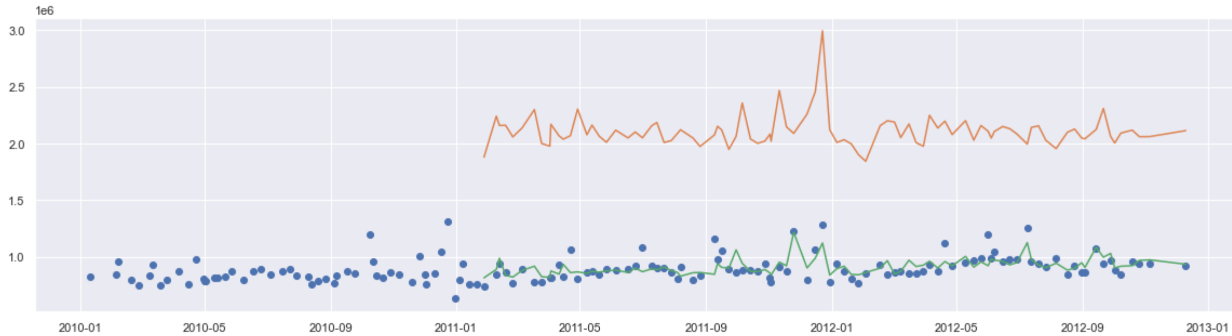

```
Store Number: 16
Score factor: 0.71
```



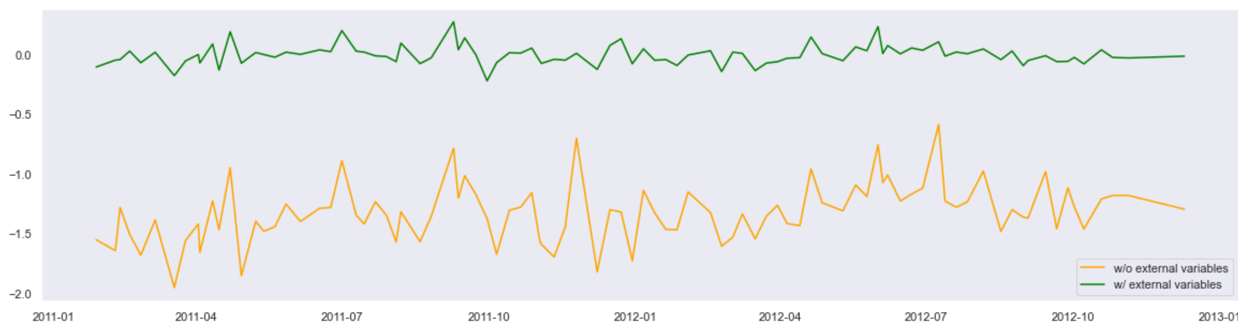
```
Store Number: 16
AR Residuals: avg -3.11, std 0.55
AR with Ext Residuals: avg -0.01, std 0.08
```



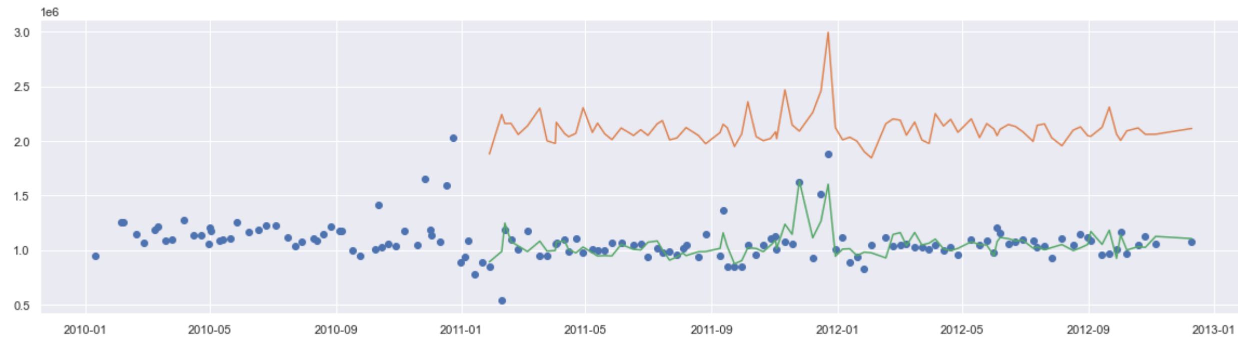

```
Store Number: 17
Score factor: 0.42
```



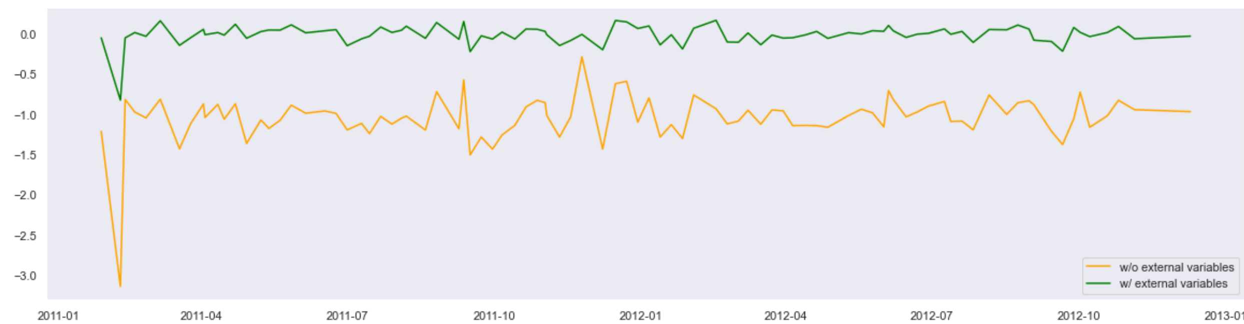

```
Store Number: 17
AR Residuals: avg -1.32, std 0.25
AR with Ext Residuals: avg -0.01, std 0.08
```



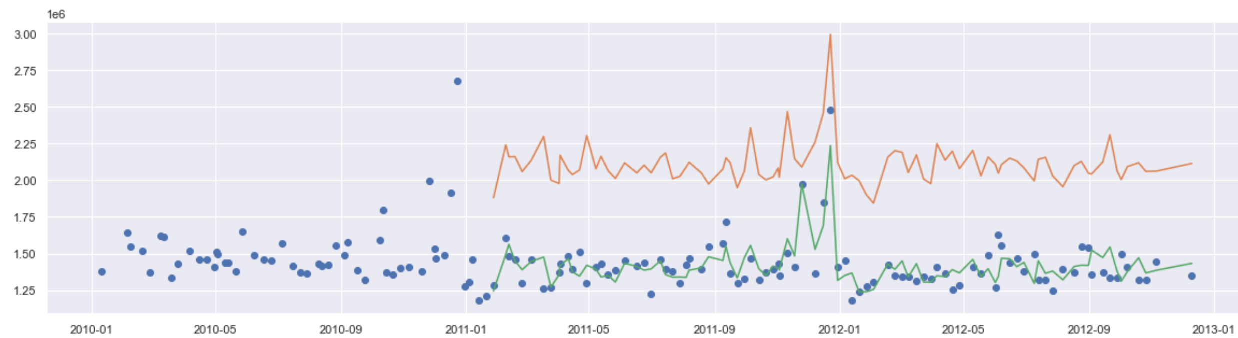

```
Store Number: 18
Score factor: 0.55
```



Store Number: 18
AR Residuals: avg -1.05, std 0.30
AR with Ext Residuals: avg -0.01, std 0.12

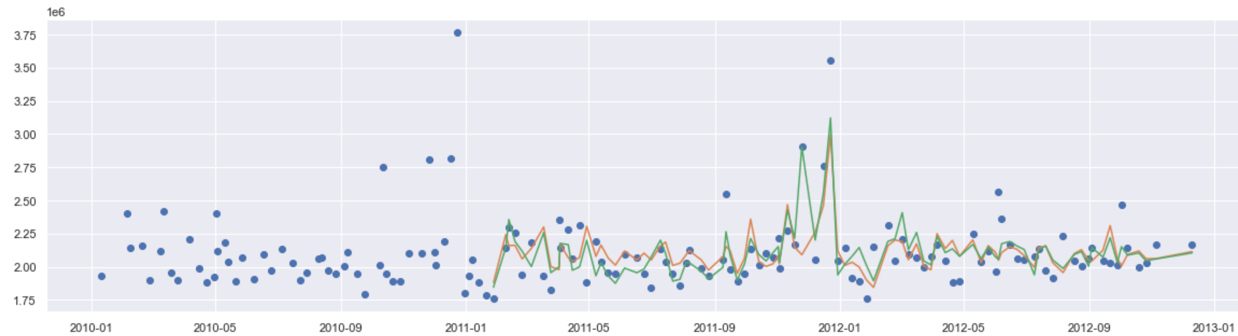


Store Number: 19
Score factor: 0.64



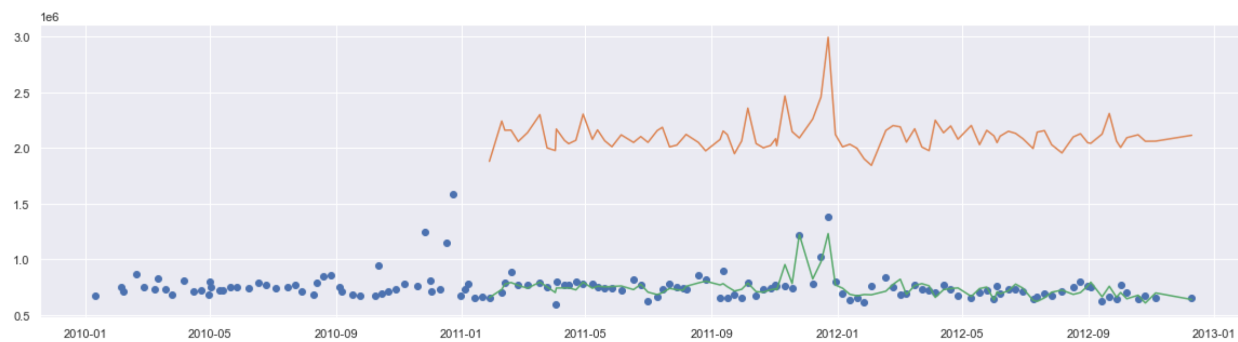
Store Number: 19
AR Residuals: avg -0.50, std 0.13
AR with Ext Residuals: avg -0.00, std 0.07

Store Number: 20
Score factor: 0.58

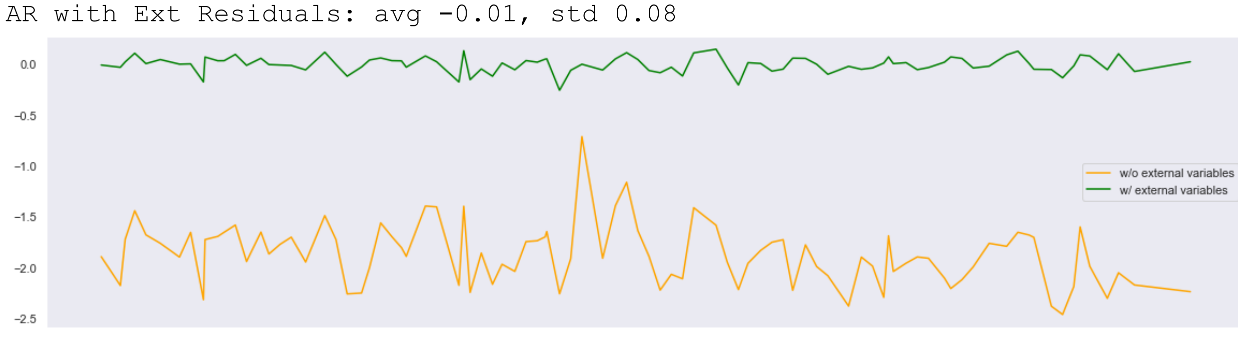



```
Store Number: 20
AR Residuals: avg -0.01, std 0.08
AR with Ext Residuals: avg -0.00, std 0.07
```

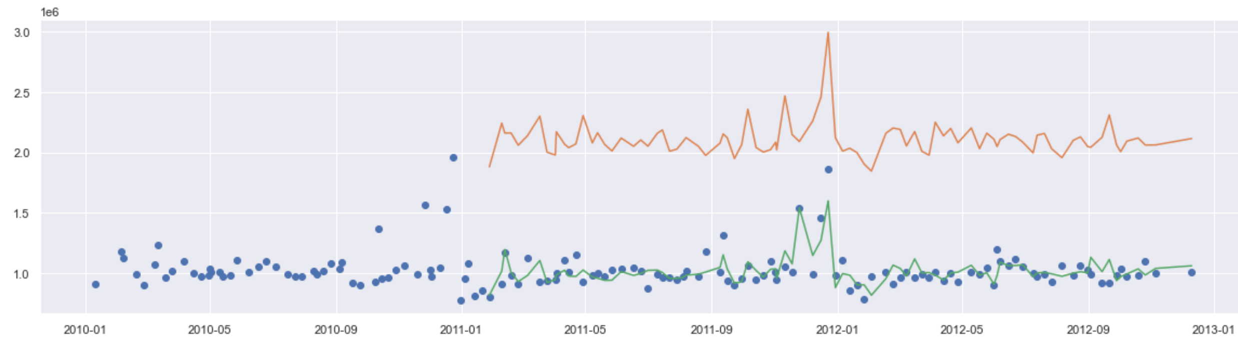

```
Store Number: 21
Score factor: 0.72
```



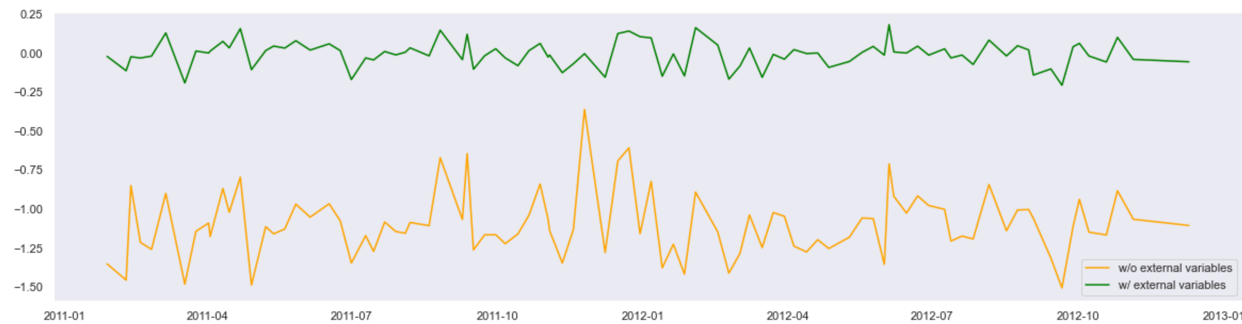

```
Store Number: 21
AR Residuals: avg -1.87, std 0.30
AR with Ext Residuals: avg -0.01, std 0.08
```



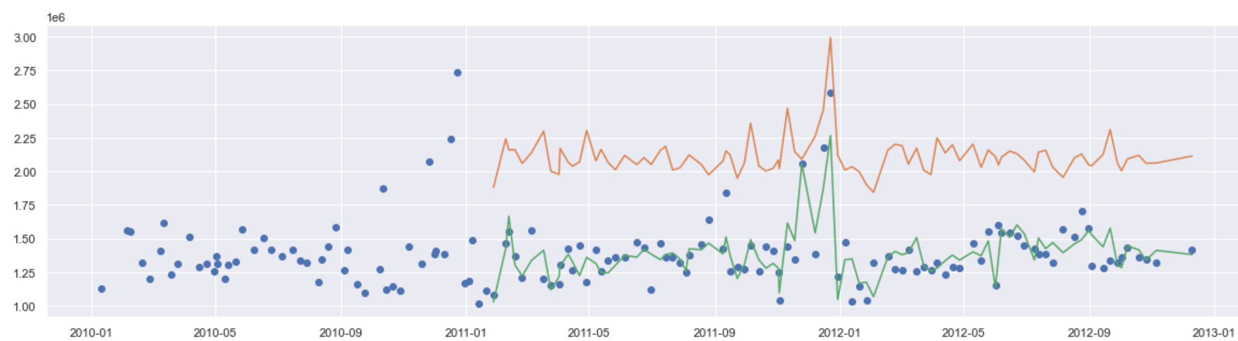

```
Store Number: 22
Score factor: 0.61
```



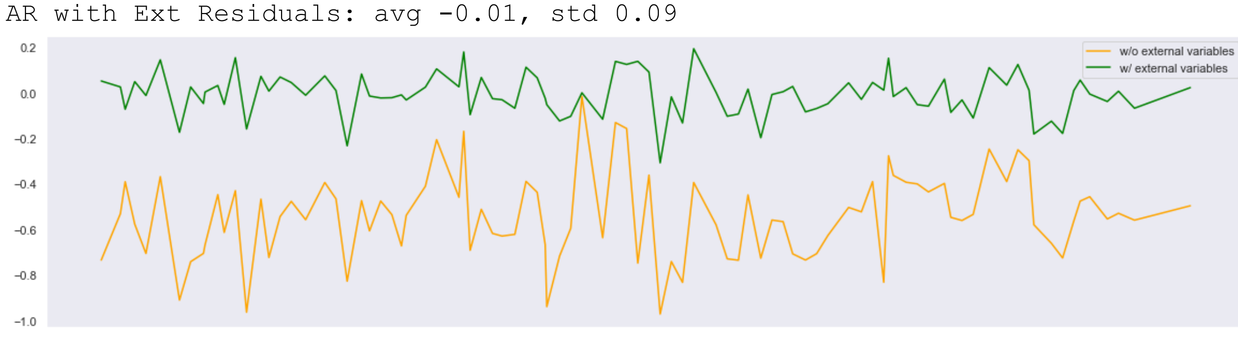
Store Number: 22
AR Residuals: avg -1.10, std 0.21
AR with Ext Residuals: avg -0.01, std 0.08



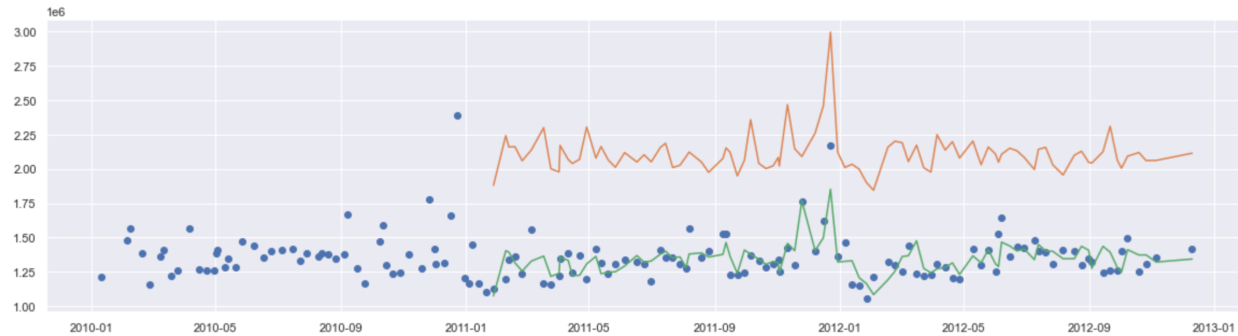
Store Number: 23
Score factor: 0.66



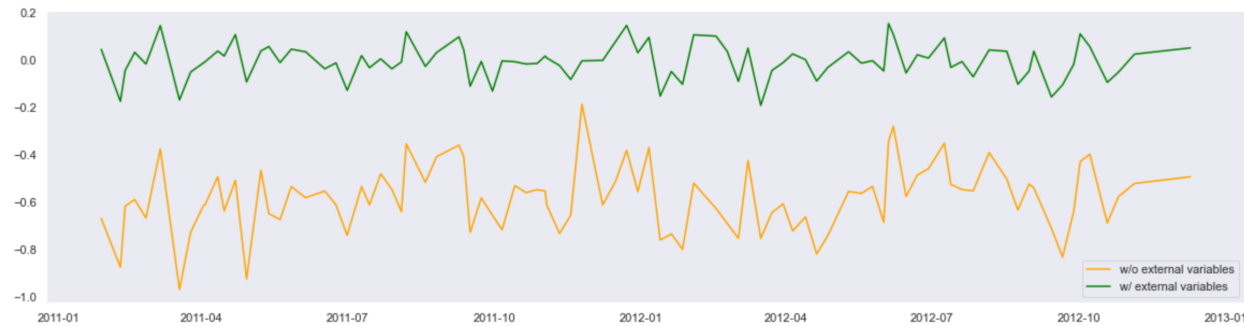
Store Number: 23
AR Residuals: avg -0.55, std 0.19
AR with Ext Residuals: avg -0.01, std 0.09



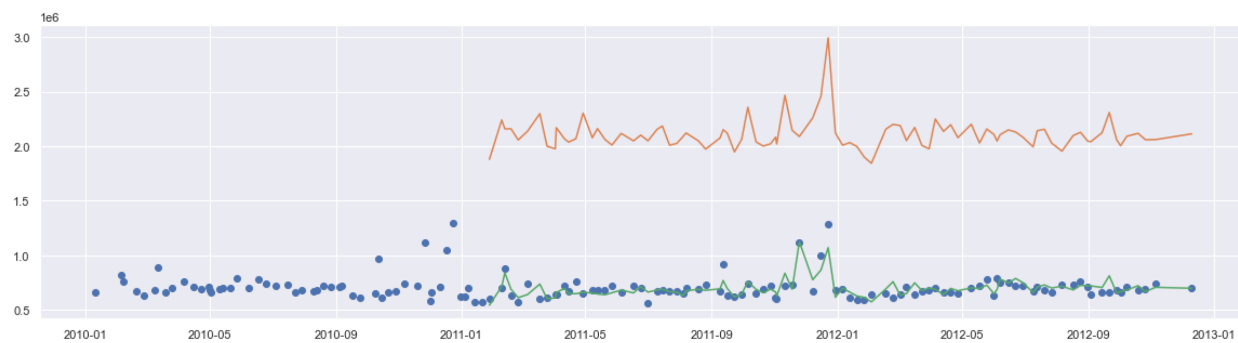
Store Number: 24
Score factor: 0.52



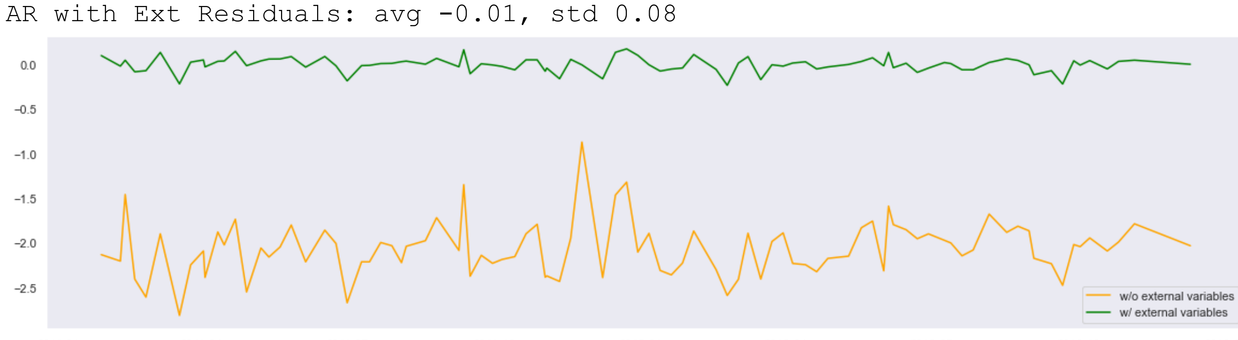
Store Number: 24
AR Residuals: avg -0.58, std 0.14
AR with Ext Residuals: avg -0.01, std 0.07



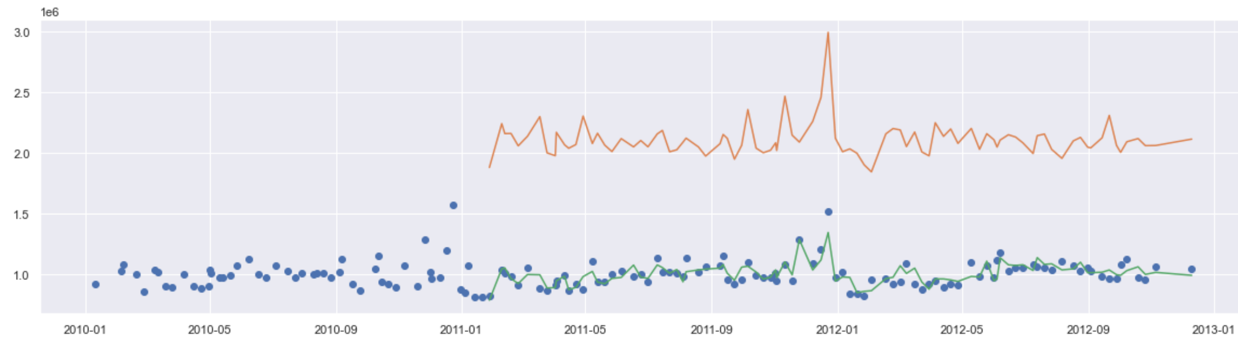
Store Number: 25
Score factor: 0.65



Store Number: 25
AR Residuals: avg -2.06, std 0.31
AR with Ext Residuals: avg -0.01, std 0.08



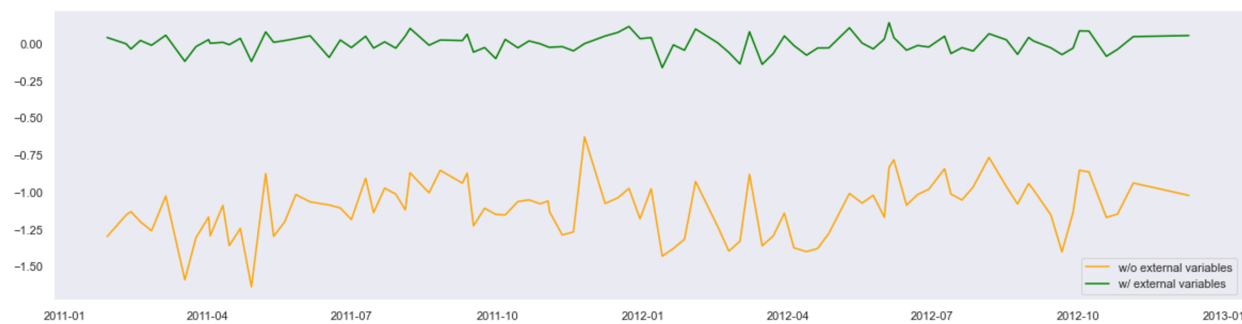
Store Number: 26
Score factor: 0.65



Store Number: 26

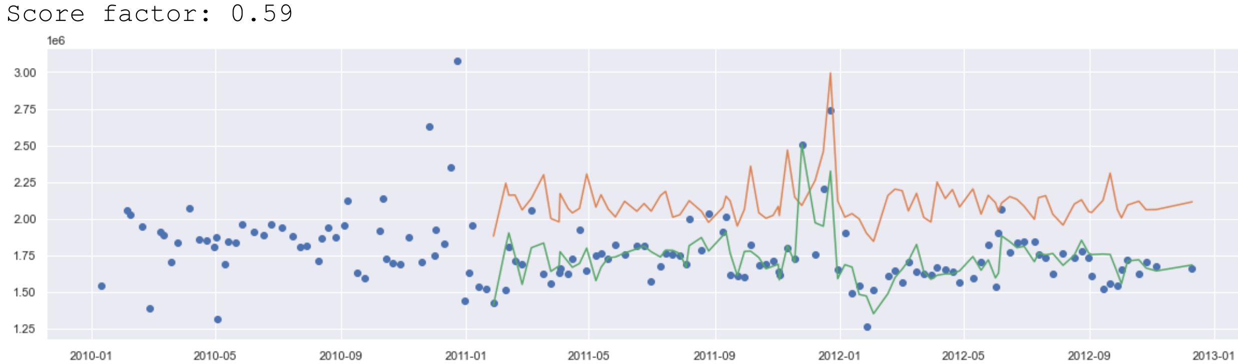
AR Residuals: avg -1.11, std 0.19

AR with Ext Residuals: avg -0.00, std 0.06



Store Number: 27

Score factor: 0.59



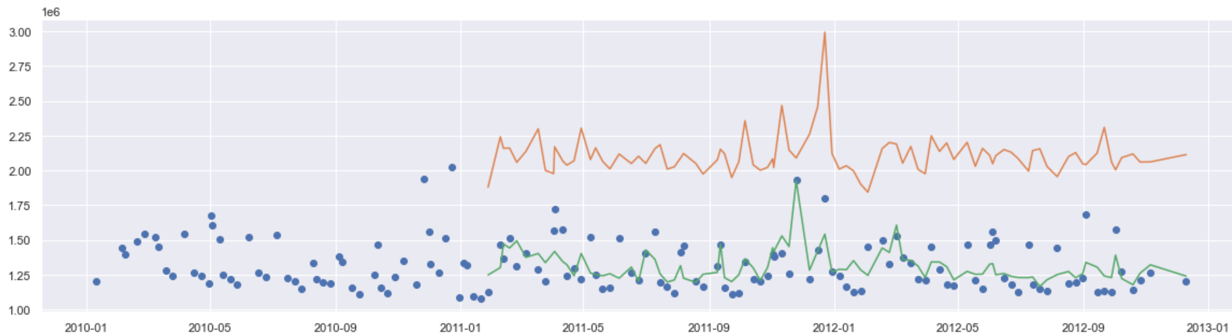
Store Number: 27

AR Residuals: avg -0.23, std 0.11

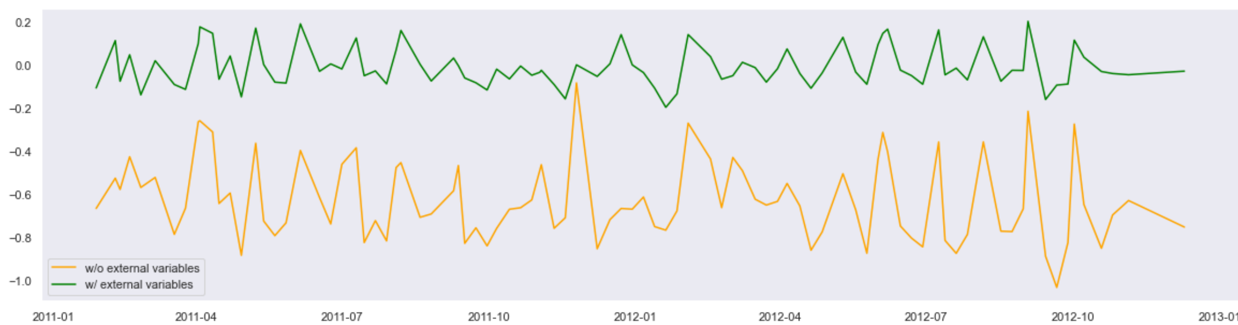
AR with Ext Residuals: avg -0.00, std 0.07

Store Number: 28

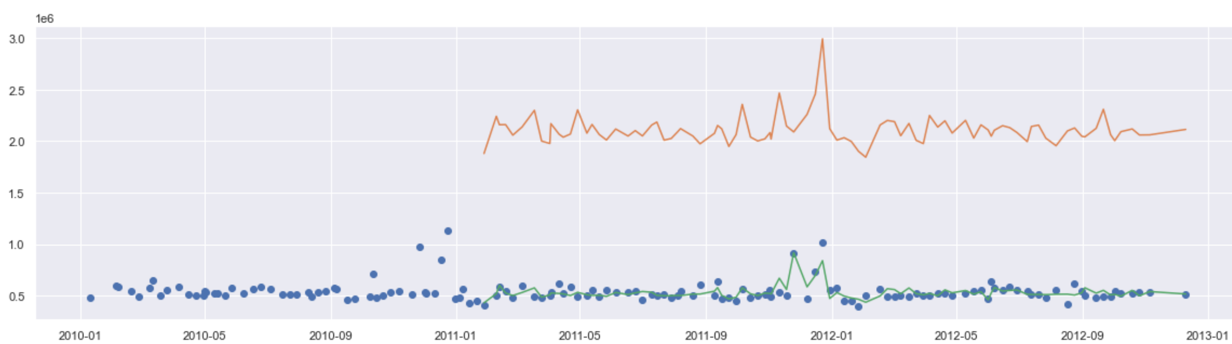
Score factor: 0.42




```
Store Number: 28
AR Residuals: avg -0.63, std 0.19
AR with Ext Residuals: avg -0.01, std 0.09
```



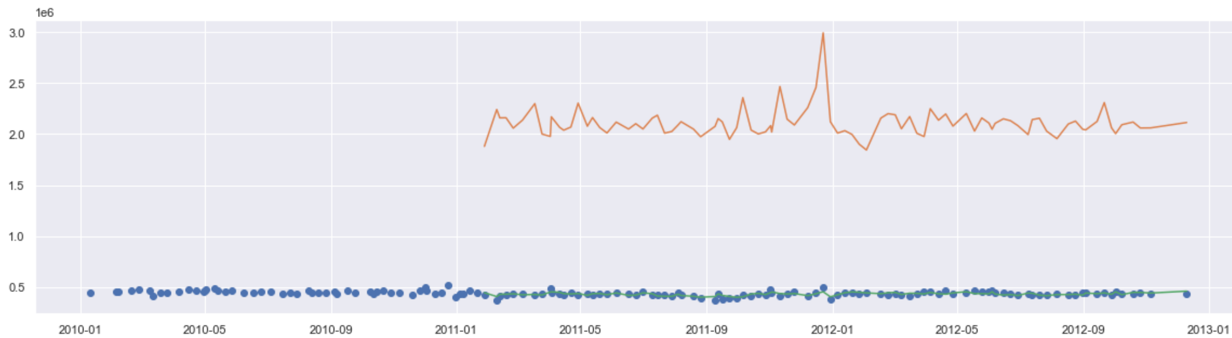

```
Store Number: 29
Score factor: 0.62
```




```
Store Number: 29
AR Residuals: avg -3.03, std 0.44
AR with Ext Residuals: avg -0.01, std 0.09
```


```
Store Number: 30
```

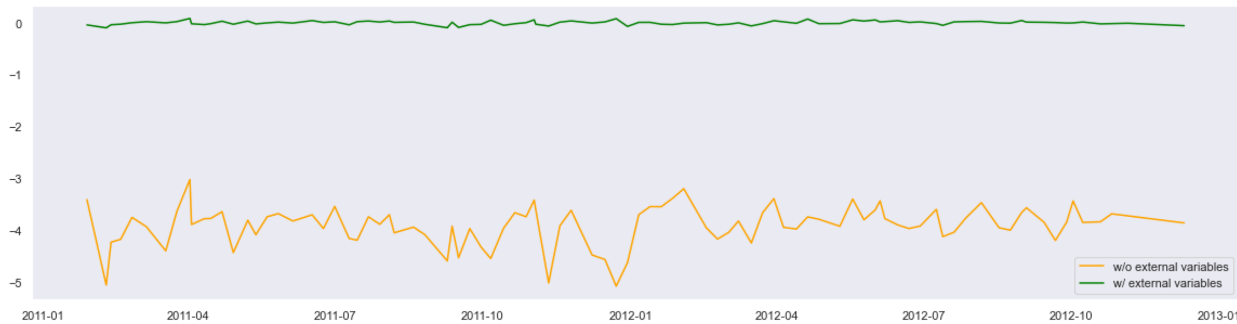
Score factor: 0.45



Store Number: 30

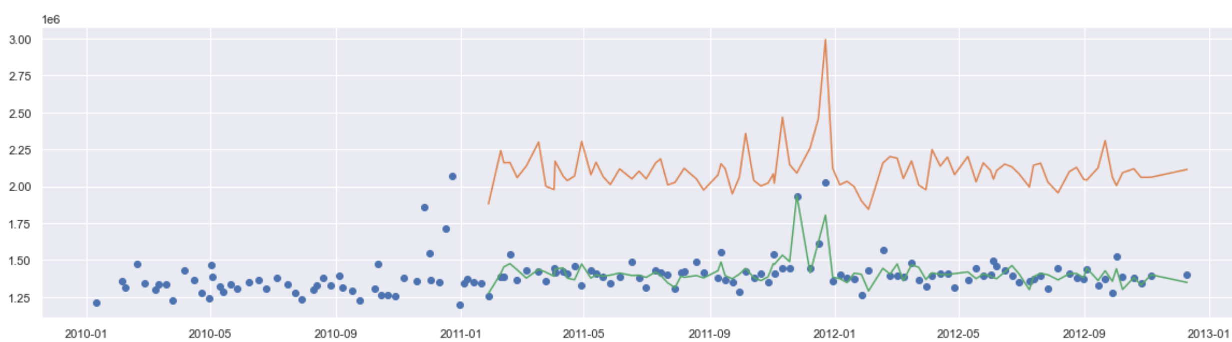
AR Residuals: avg -3.91, std 0.38

AR with Ext Residuals: avg -0.00, std 0.04



Store Number: 31

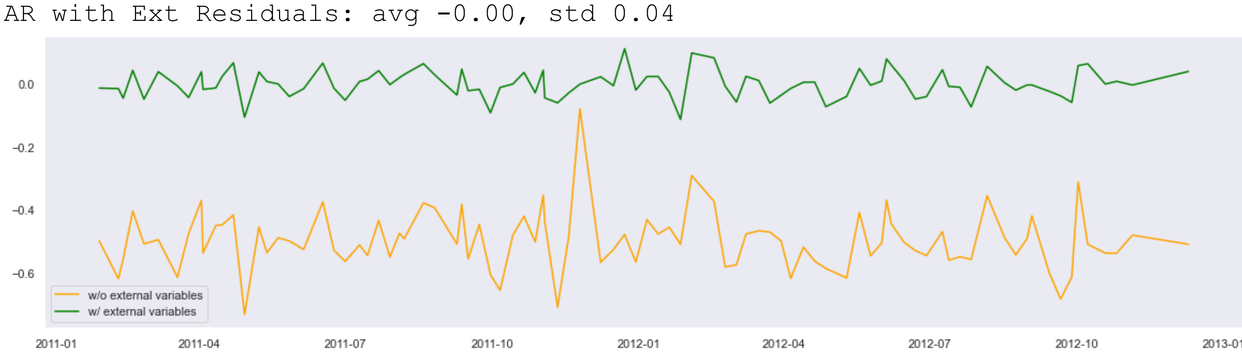
Score factor: 0.65



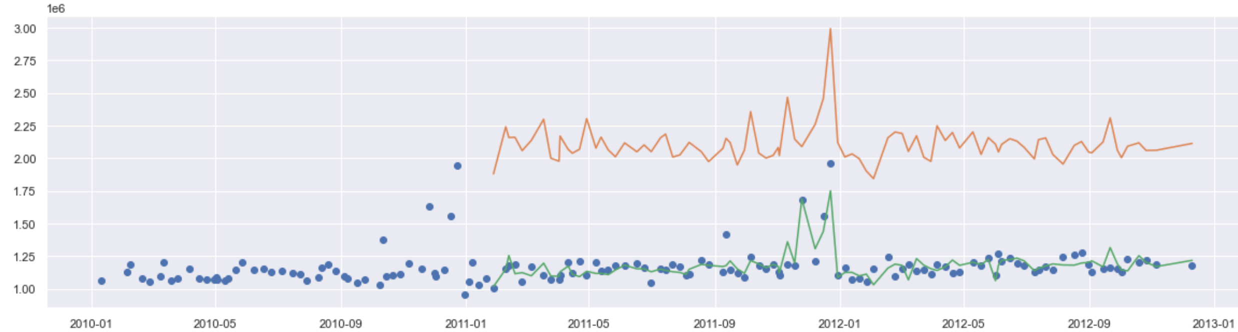
Store Number: 31

AR Residuals: avg -0.50, std 0.09

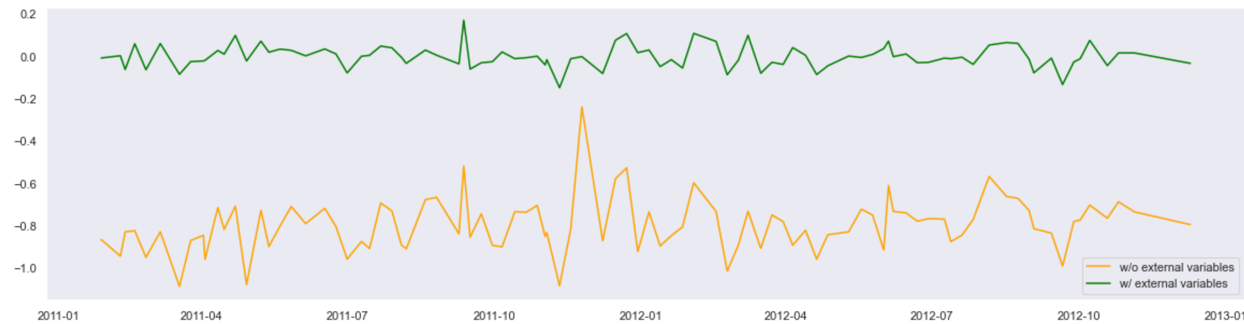
AR with Ext Residuals: avg -0.00, std 0.04



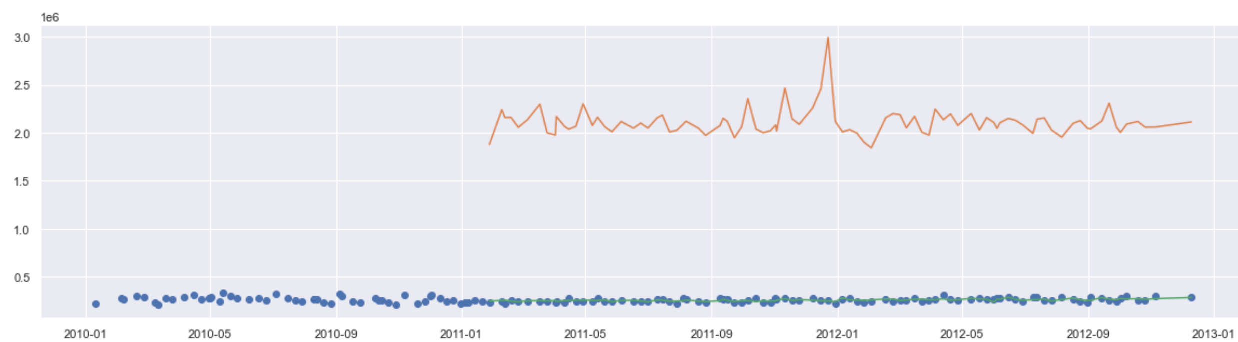
Store Number: 32
Score factor: 0.70



Store Number: 32
AR Residuals: avg -0.80, std 0.13
AR with Ext Residuals: avg -0.00, std 0.05

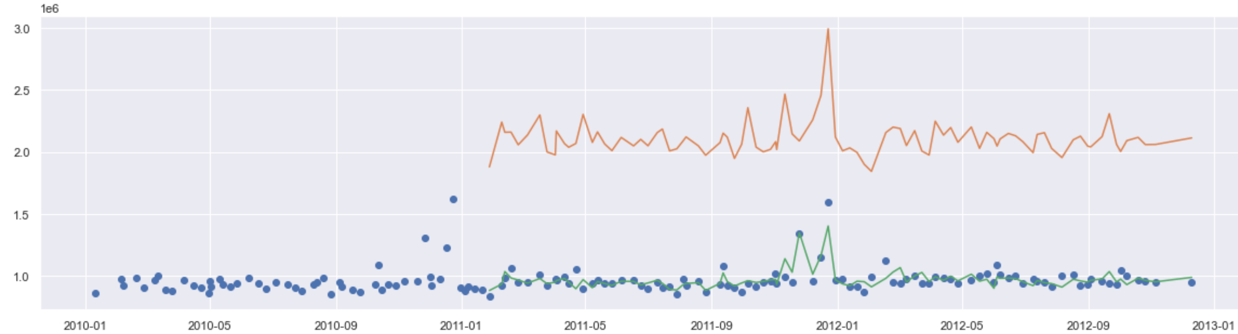


Store Number: 33
Score factor: 0.32

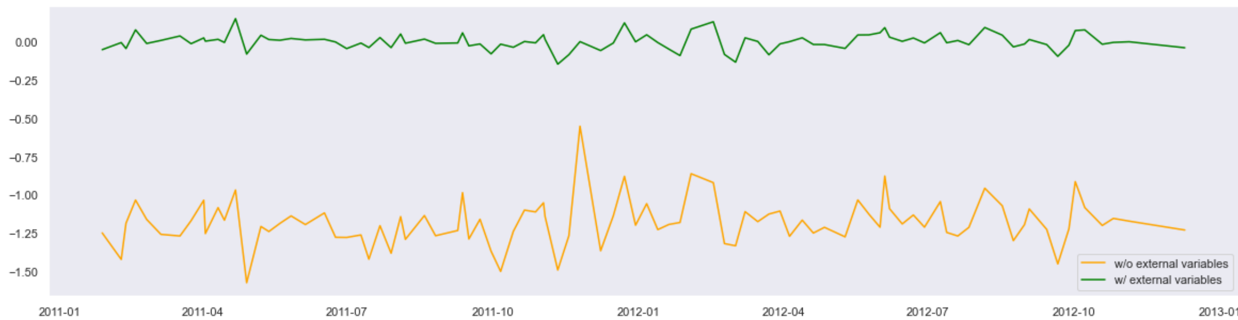


Store Number: 33
AR Residuals: avg -7.23, std 0.76
AR with Ext Residuals: avg -0.00, std 0.06

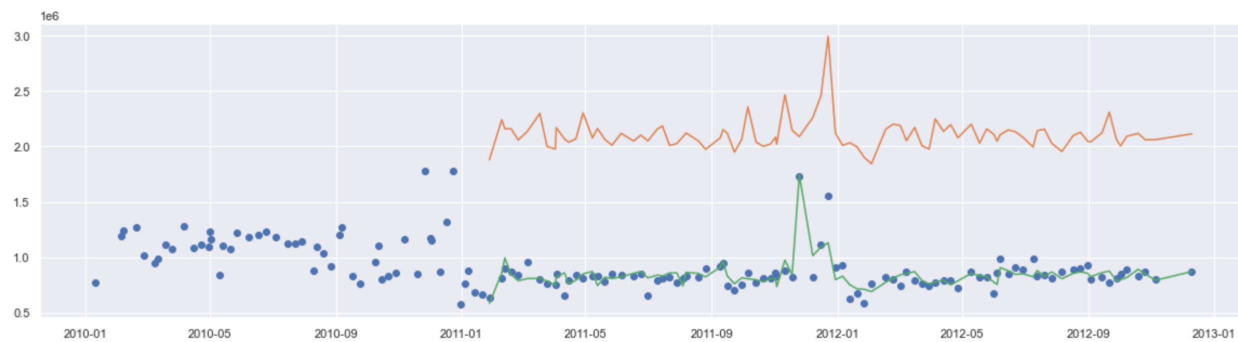
Store Number: 34
Score factor: 0.67



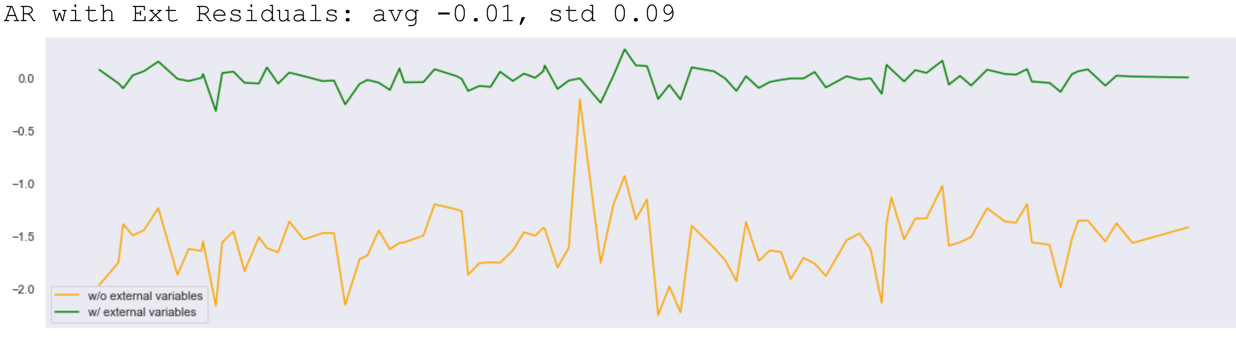
Store Number: 34
AR Residuals: avg -1.18, std 0.15
AR with Ext Residuals: avg -0.00, std 0.05



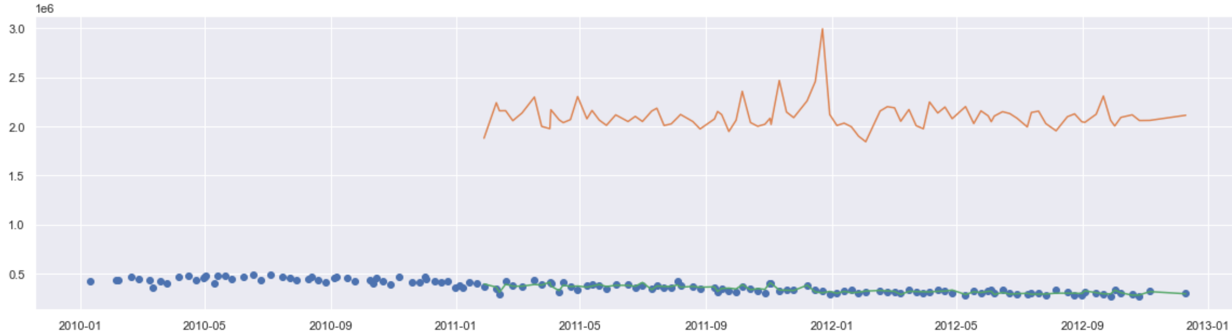
Store Number: 35
Score factor: 0.69



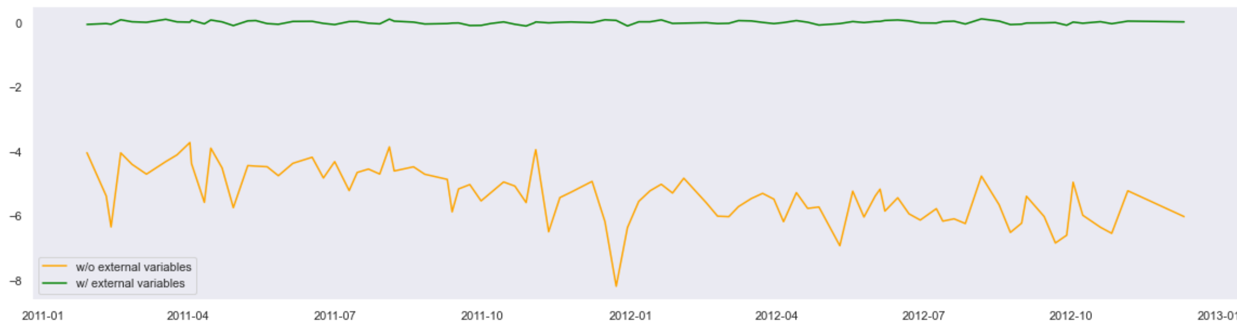
Store Number: 35
AR Residuals: avg -1.56, std 0.30
AR with Ext Residuals: avg -0.01, std 0.09



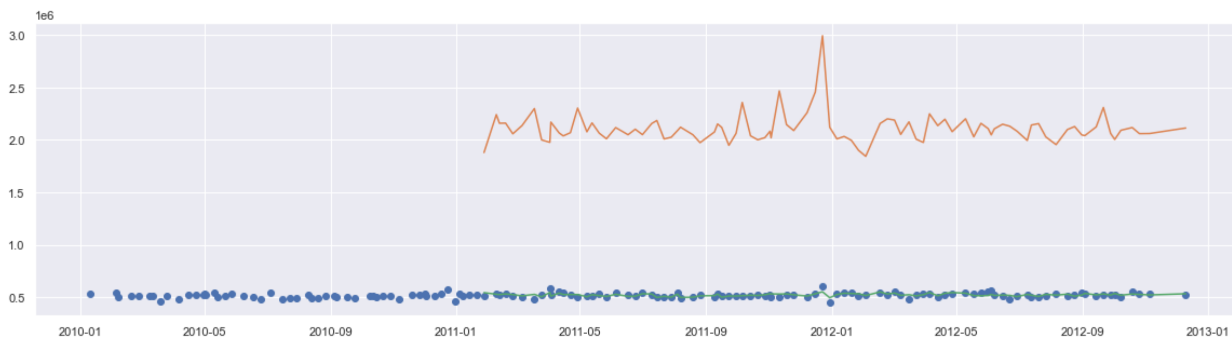
```
Store Number: 36
Score factor: 0.81
```



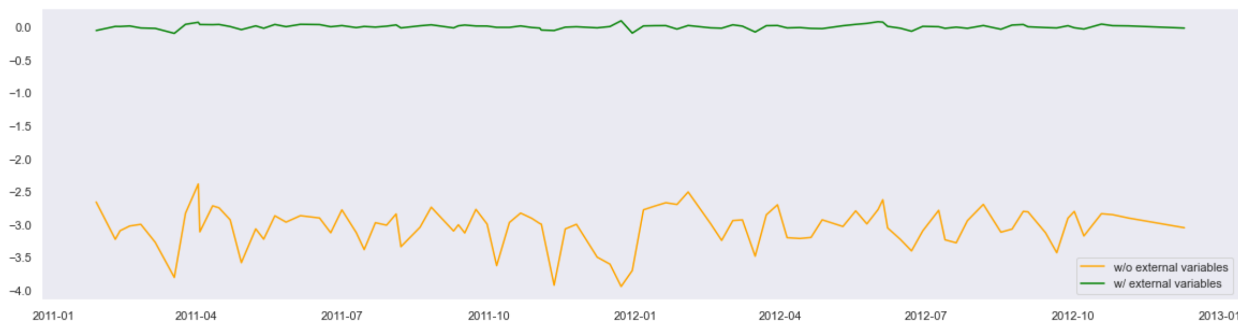

```
Store Number: 36
AR Residuals: avg -5.34, std 0.83
AR with Ext Residuals: avg -0.00, std 0.05
```



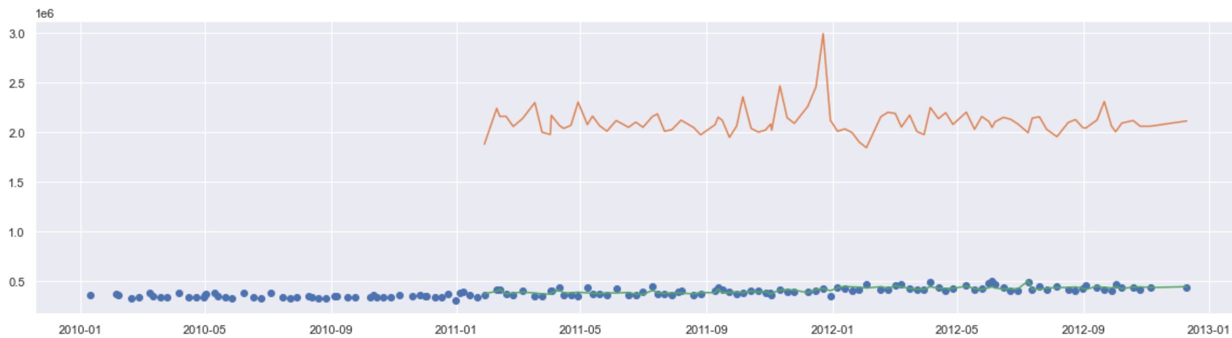

```
Store Number: 37
Score factor: 0.34
```



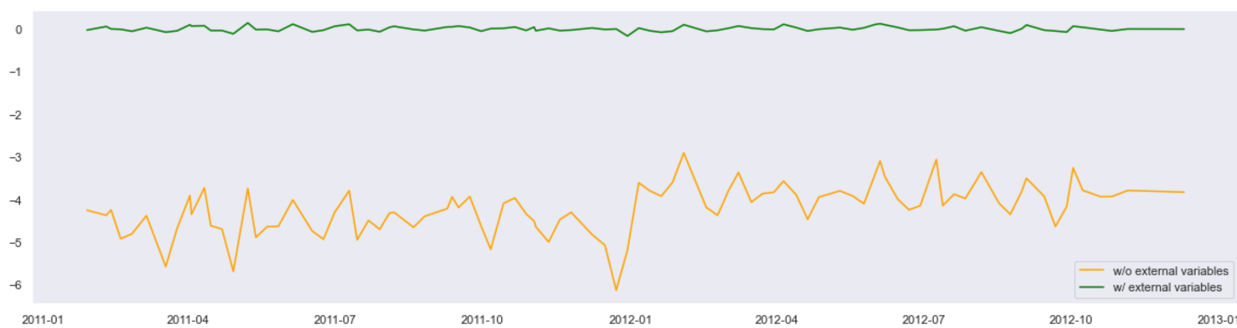

```
Store Number: 37
AR Residuals: avg -3.04, std 0.29
AR with Ext Residuals: avg -0.00, std 0.03
```



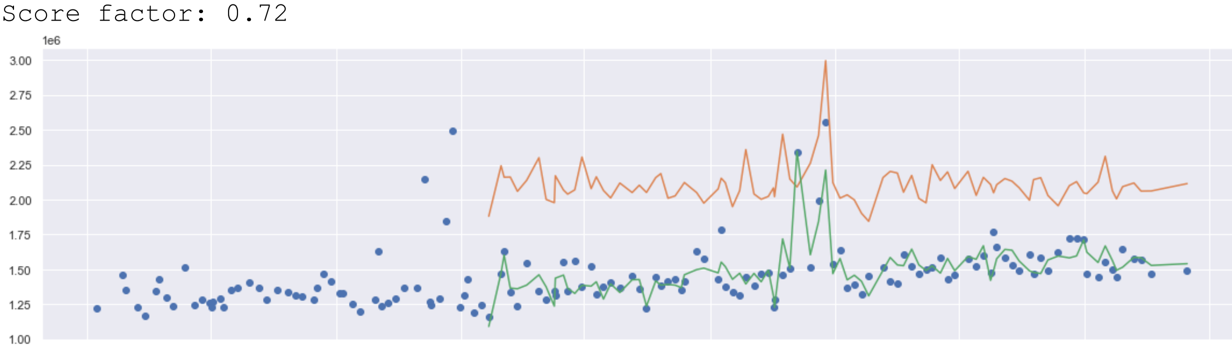

```
Store Number: 38
Score factor: 0.57
```




```
Store Number: 38
AR Residuals: avg -4.22, std 0.57
AR with Ext Residuals: avg -0.00, std 0.06
```

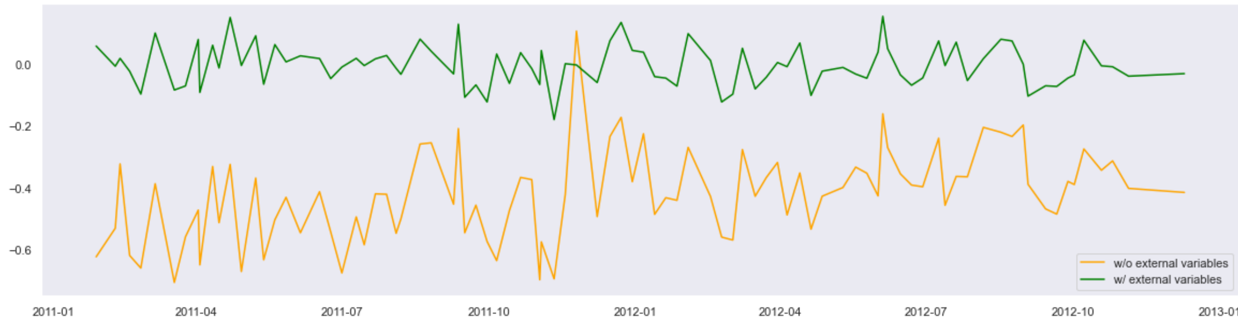



```
Store Number: 39
Score factor: 0.72
```



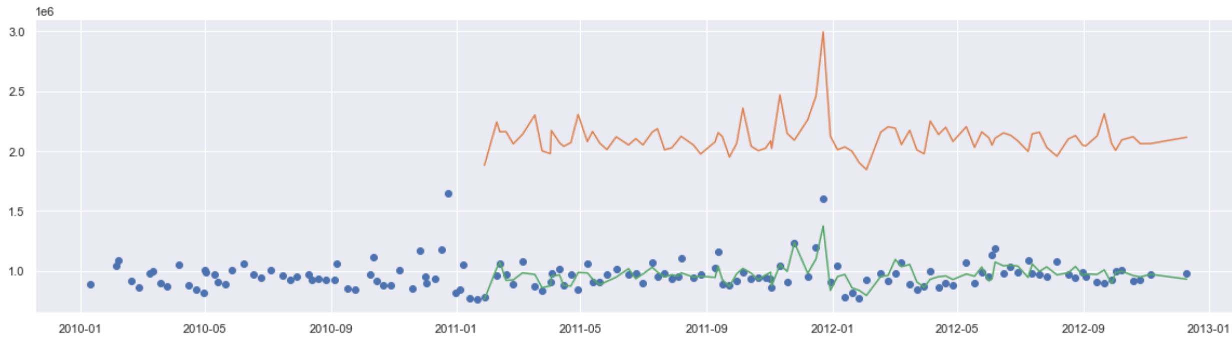

```
Store Number: 39
AR Residuals: avg -0.42, std 0.14
```

AR with Ext Residuals: avg -0.00, std 0.07



Store Number: 40

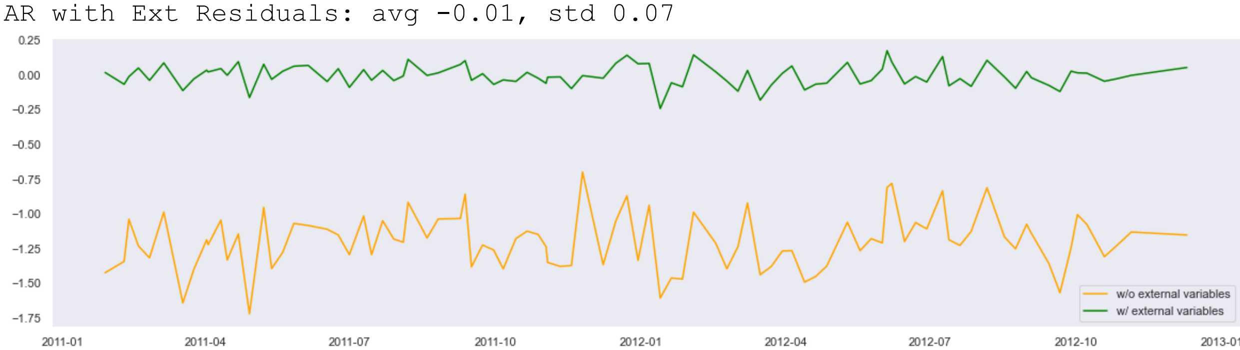
Score factor: 0.55



Store Number: 40

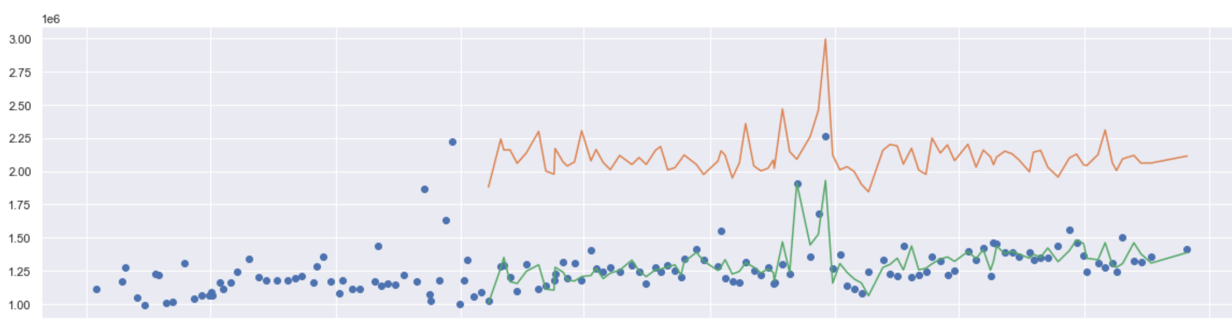
AR Residuals: avg -1.20, std 0.20

AR with Ext Residuals: avg -0.01, std 0.07



Store Number: 41

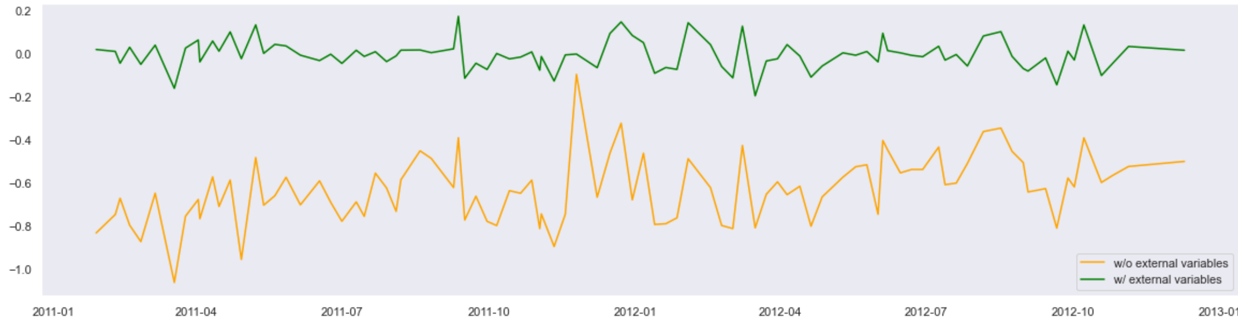
Score factor: 0.66



Store Number: 41

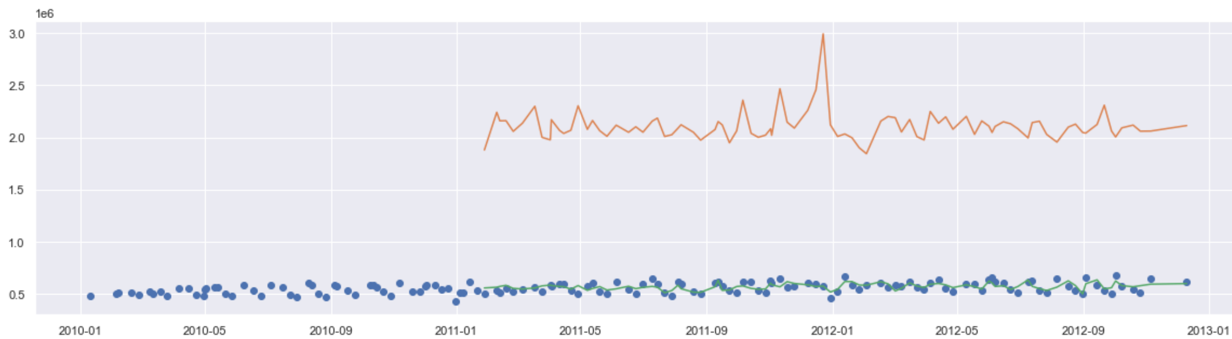
AR Residuals: avg -0.63, std 0.15

AR with Ext Residuals: avg -0.00, std 0.07



Store Number: 42

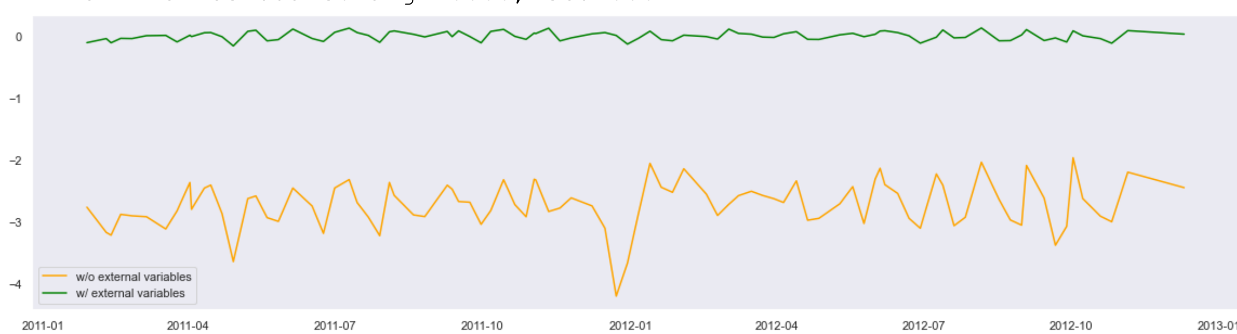
Score factor: 0.36



Store Number: 42

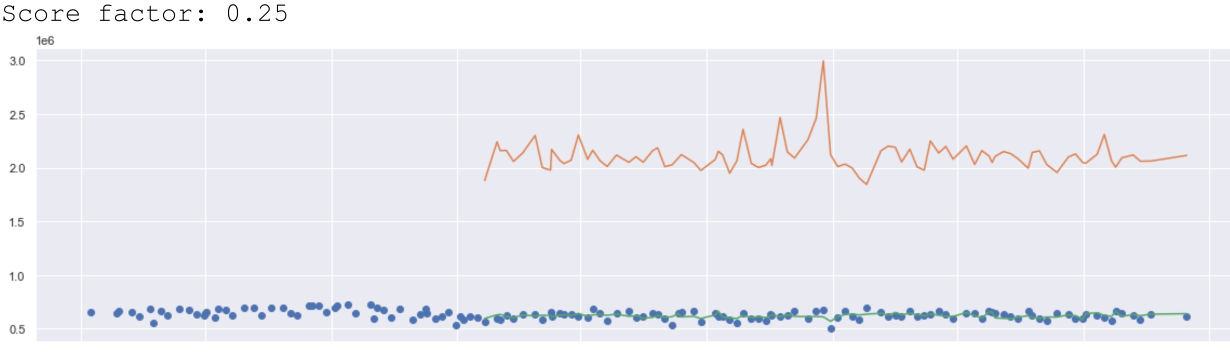
AR Residuals: avg -2.72, std 0.37

AR with Ext Residuals: avg -0.00, std 0.07



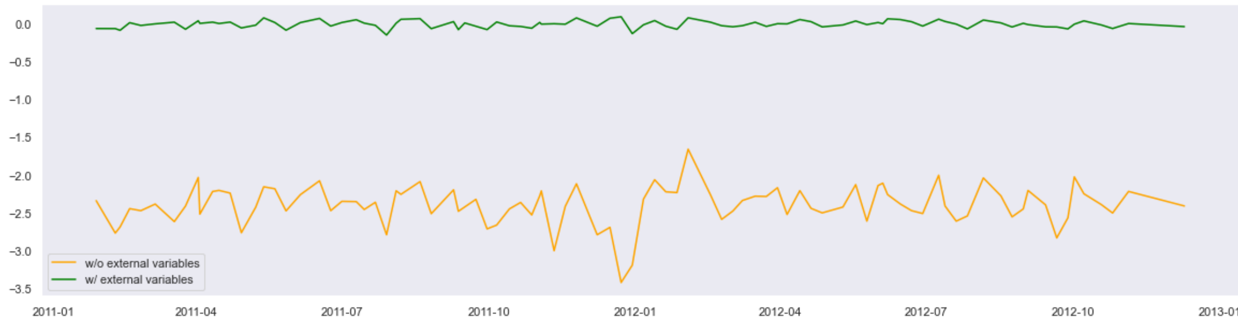
Store Number: 43

Score factor: 0.25

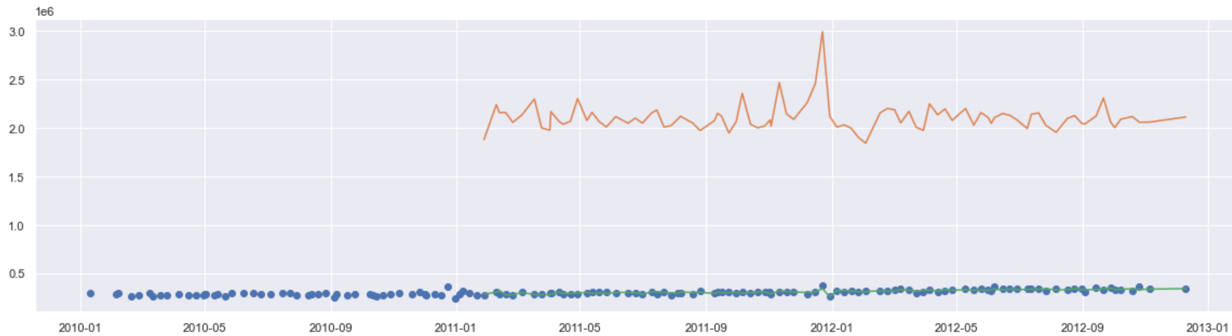


Store Number: 43

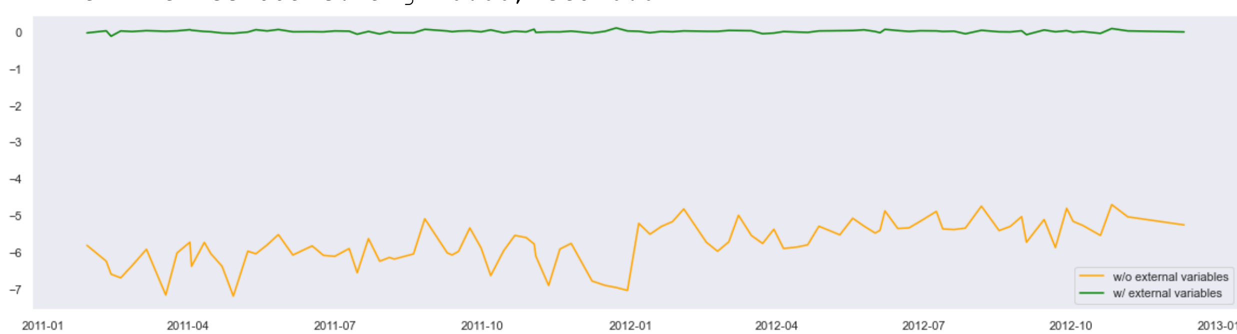
```
AR Residuals: avg -2.40, std 0.26
AR with Ext Residuals: avg -0.00, std 0.05
```



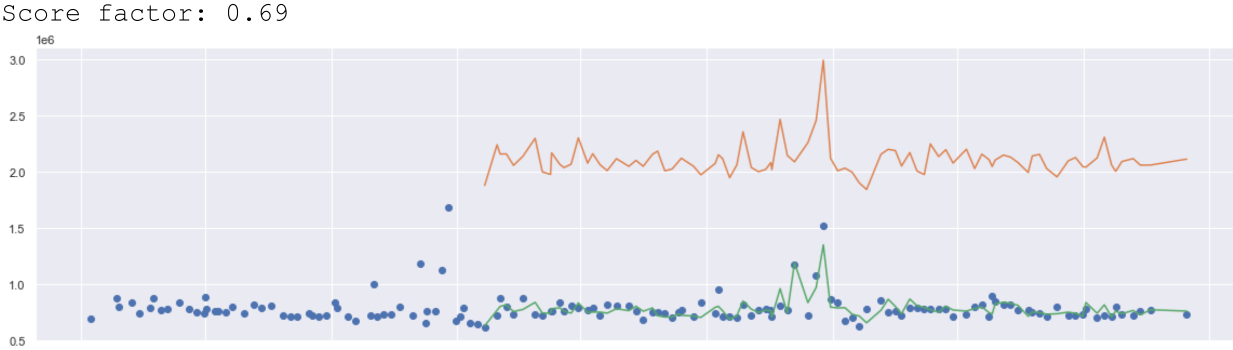

```
Store Number: 44
Score factor: 0.74
```



```
Store Number: 44
AR Residuals: avg -5.76, std 0.58
AR with Ext Residuals: avg -0.00, std 0.04
```




```
Store Number: 45
Score factor: 0.69
```



```
Store Number: 45
AR Residuals: avg -1.73, std 0.25
AR with Ext Residuals: avg -0.00, std 0.07
```