Capstone 2 Walmart sales prediction

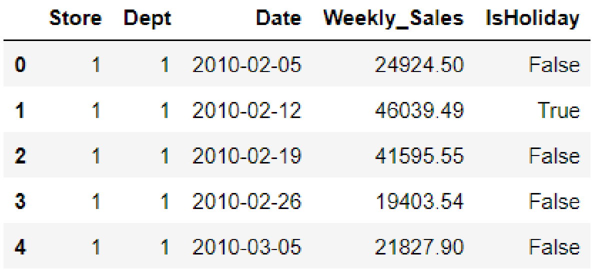
Walmart, as one of the world's largest retail chains, operates thousands of stores across multiple countries, each comprising numerous departments. Accurate sales forecasting for both individual stores and departments within those stores is crucial for optimizing stock levels, manpower, and other resources, while also maximizing revenue and profitability. However, the challenge of sales prediction at Walmart is amplified by various factors such as seasonality, holidays, and promotional events, among others. This report aims to develop a robust model to predict weekly sales across various Walmart stores and departments. Leveraging machine learning algorithms and time series analysis, the objective is to create a predictive framework that is both accurate and scalable, ultimately aiding Walmart in strategic decision-making.

1. Data

This Kaggle project is a sufficient size to develop a good predictor model. There are three datasets in .csv format which have store information for 45 stores. Total number of time series are 3331 with time span between 2010-02-05 and 2012-10-26 on weekly basis. The first dataset contains sales data provided on weekly basis. The second dataset contains features for each store. This dataset include the customer price index(indicator of inflation) , holidays, temperature, unemployment rate and fuel price, as well as the date ranging from 2010-02-05 to 2013-07-26. Last dataset contains basic store information such as store type and store size.

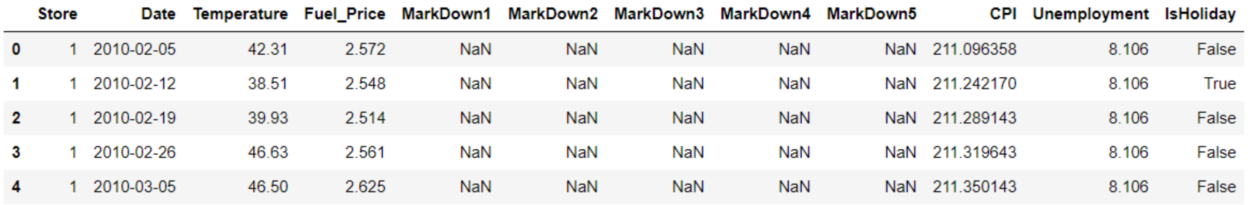
Here are snapshot of all three datasets:

Weekly_sales



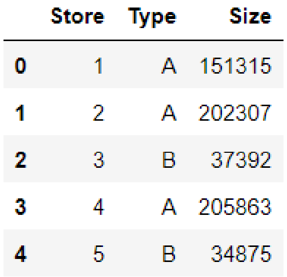
	Store	Dept	Date	Weekly_Sales	IsHoliday
0	1	1	2010-02-05	24924.50	False
1	1	1	2010-02-12	46039.49	True
2	1	1	2010-02-19	41595.55	False
3	1	1	2010-02-26	19403.54	False
4	1	1	2010-03-05	21827.90	False

Features



	Store	Date	Temperature	Fuel_Price	MarkDown1	MarkDown2	MarkDown3	MarkDown4	MarkDown5	CPI	Unemployment	IsHoliday
0	1	2010-02-05	42.31	2.572	NaN	NaN	NaN	NaN	NaN	211.096358	8.106	False
1	1	2010-02-12	38.51	2.548	NaN	NaN	NaN	NaN	NaN	211.242170	8.106	True
2	1	2010-02-19	39.93	2.514	NaN	NaN	NaN	NaN	NaN	211.289143	8.106	False
3	1	2010-02-26	46.63	2.561	NaN	NaN	NaN	NaN	NaN	211.319643	8.106	False
4	1	2010-03-05	46.50	2.625	NaN	NaN	NaN	NaN	NaN	211.350143	8.106	False

Stores



	Store	Type	Size
0	1	A	151315
1	2	A	202307
2	3	B	37392
3	4	A	205863
4	5	B	34875

2. Data Cleaning

1)No missing values found in Store and Weekly Sales data frames. For numerical features, missing values in markdown of feature dataset are about 50% above and all the markdown values from 2010-02-05 to 2011-11-11 are missing mainly because there is no records for markdown values during this period, therefore the values are filled with zeros. Missing values in CPI and unemployment is about 7%, cannot replaced with median value, rolling window strategy was considered to fill in the missing values

2) abnormal(negative) weekly sales are observed (about 1285 out of 421K, 0.3%), observations with negative sales were removed

3) total number of time series with time gap is 695, which takes about 21% of all the time series. Out of all the time series with time gaps, only 18% has missing values less than 10%,27% missing values <20%, 55% has missing value more than 50%. Only time series without time gap were selected for sales predictions

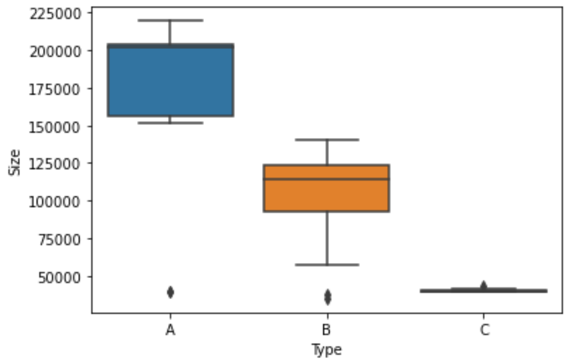
4) additional features like week, season, year, quarter, and holiday names are created to facilitate seasonal analysis for weekly sales

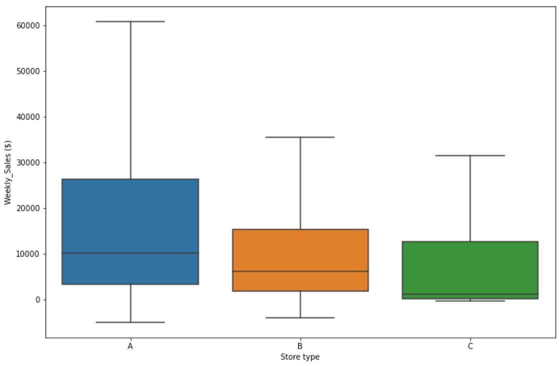
Three data frames were merged to create one single data frame for data analysis later on.

3. Data Exploration

3.1 Store type analysis

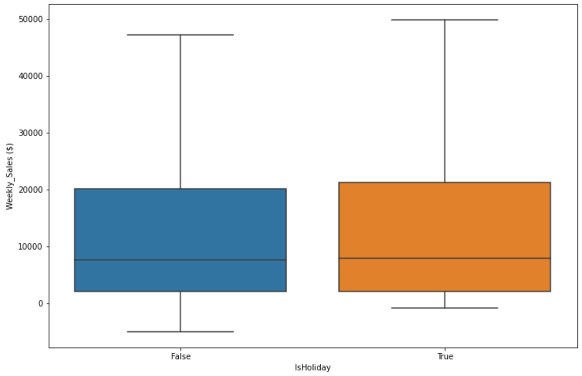
Half of the stores are type A and have largest size in average, type C has smallest size in average and least number of stores; Type A tends to have highest average weekly sales value

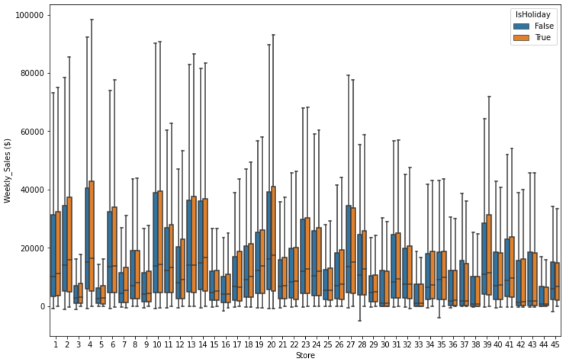




3.2. Holiday Effect

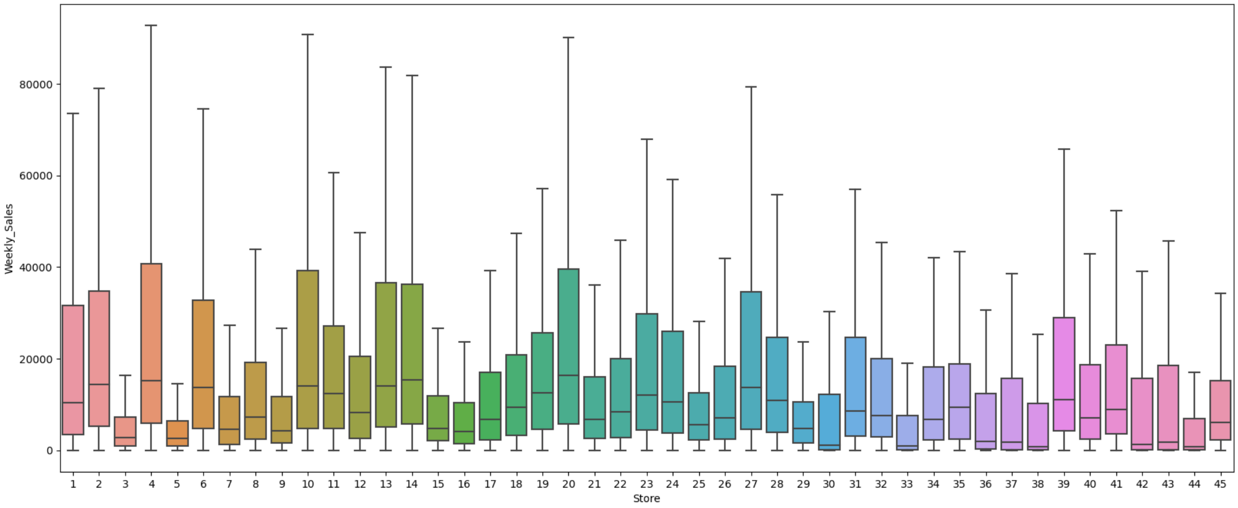
Comparing holidays and nonholidays for weekly sales, the median values overall are similar, maximum or minimum sales values are higher on holidays than that on non-holidays; when looking at each store, the holiday weekly sales is always slightly higher than non-holidays.

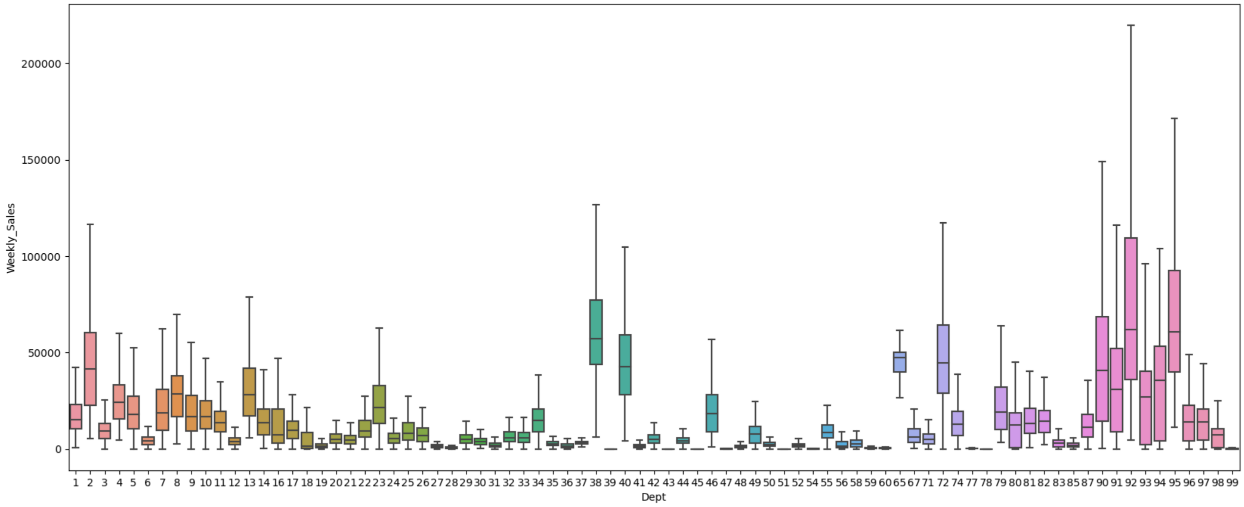




3.3 Weekly Sales Distribution for Each Store/Dept

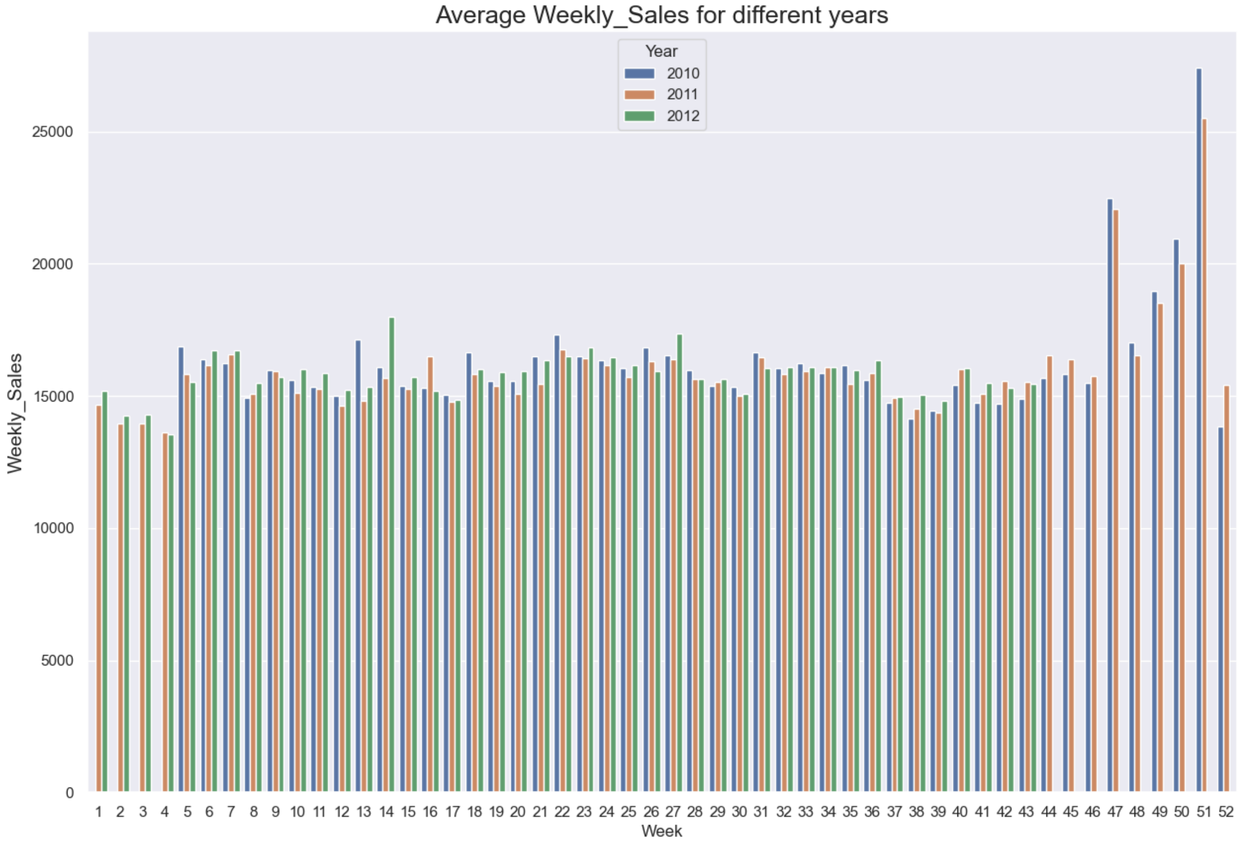
Over the last two years, Store 14's Department 92 has recorded the highest total sales, accumulating a revenue of $26,101,497. When looking at total sales, Department 92 dominates, with Stores 14, 20, and 4 being the top contributors. In terms of average weekly sales, Stores 20, 14, and 4 emerge as the leaders. On the opposite end of the spectrum, Stores 30, 33, 38, and 44 exhibit the lowest average weekly sales. In a department-wise comparison, Department 92 stands out with a higher average in weekly sales compared to other departments.





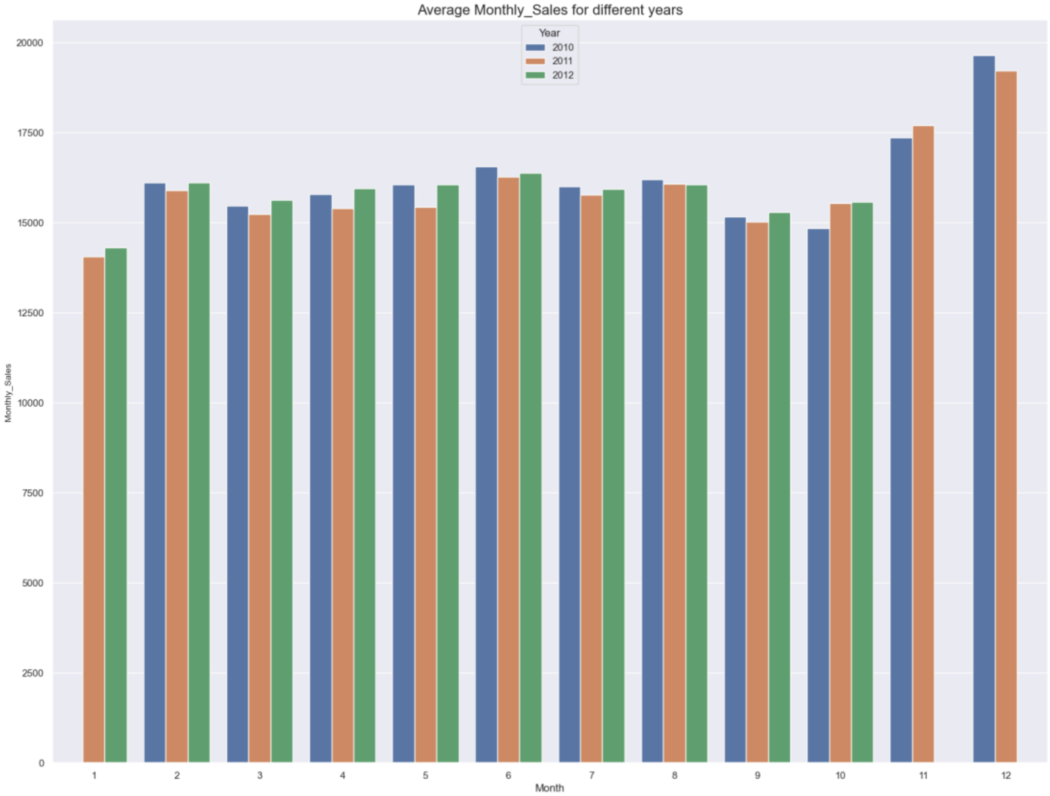
3.4 Periodic Trend of Weekly Sales

Over the course of each year, the sales figures exhibit marked fluctuations, yet these variations follow a remarkably consistent trend across different years. Notably, the end of the year consistently brings a surge in sales, solidifying it as a peak sales period irrespective of the year in question. This annual high point serves as both a consistent pattern and a critical window for revenue generation. Despite these peaks and valleys, the overarching trend remains largely stable from one year to the next, allowing for predictable business planning and resource allocation.



Simultaneously, month-to-month analysis provides a broader view, giving us a deeper understanding of seasonal patterns, long-term growth or decline, and the effectiveness of monthly promotions or events. Building on our annual observations, a more granular monthly analysis further refines our understanding of these fluctuations and trends. Each month presents its own set of challenges and opportunities, yet the overarching consistency of higher sales toward the year-end remains evident. Our monthly review allows us to identify specific periods of increased customer engagement, promotional effectiveness, and inventory needs.

By maintaining these dual lenses, we are better equipped to make informed decisions that cater to both immediate needs and longer-term strategic goals

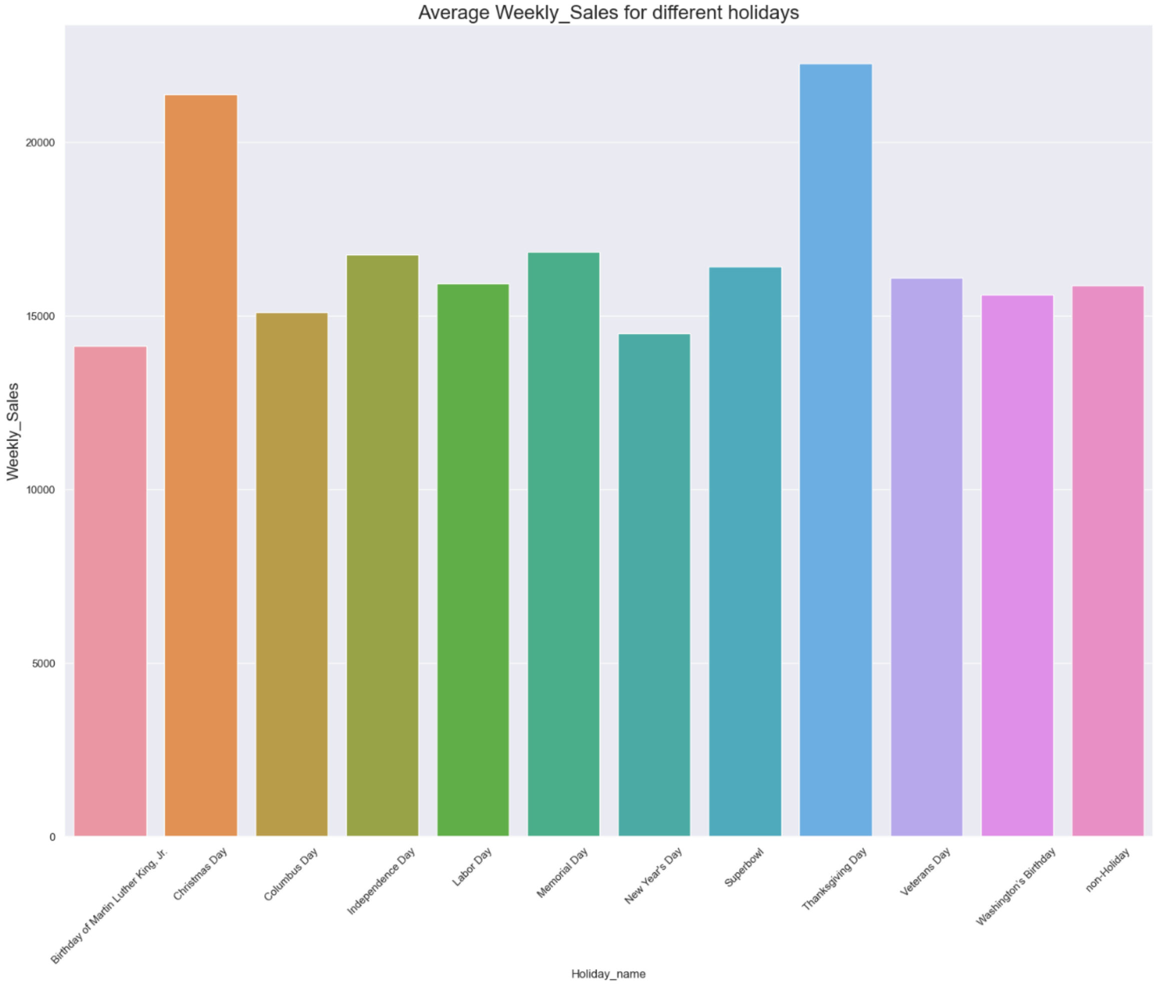


3.5 Holiday Effect on Averaged Weekly Sales

In addition to the weekly and monthly trends, it's imperative to highlight the pronounced impact of holidays on sales performance. The influence of these seasonal events is substantial, often accounting for spikes that align with or even exceed our year-end highs.

The peak in average weekly sales consistently occurs during the weeks of Thanksgiving and Christmas, which accounts for the elevated sales figures we observe at the close of each year. In contrast, other holidays do not appear to significantly impact sales when compared to regular, non-holiday periods.

Whether it's the winter holidays, Black Friday, or other significant public holidays, these periods offer a unique set of opportunities for revenue growth and customer acquisition. Factoring in the 'holiday effect' enhances our data-driven strategy, allowing us to allocate resources more efficiently and capitalize on these peak sales windows.



4. Algorithms & Machine Learning

4.1 Machine Learning for Single Time Series

A representative time series was selected as the initial basis for selecting algorithms. I explored a variety of algorithms to gauge their performance. These algorithms encompass ARIMA, linear regression, random forest regression, XGBoost, and Prophet. Summary of the algorithm performance is listed in the following:

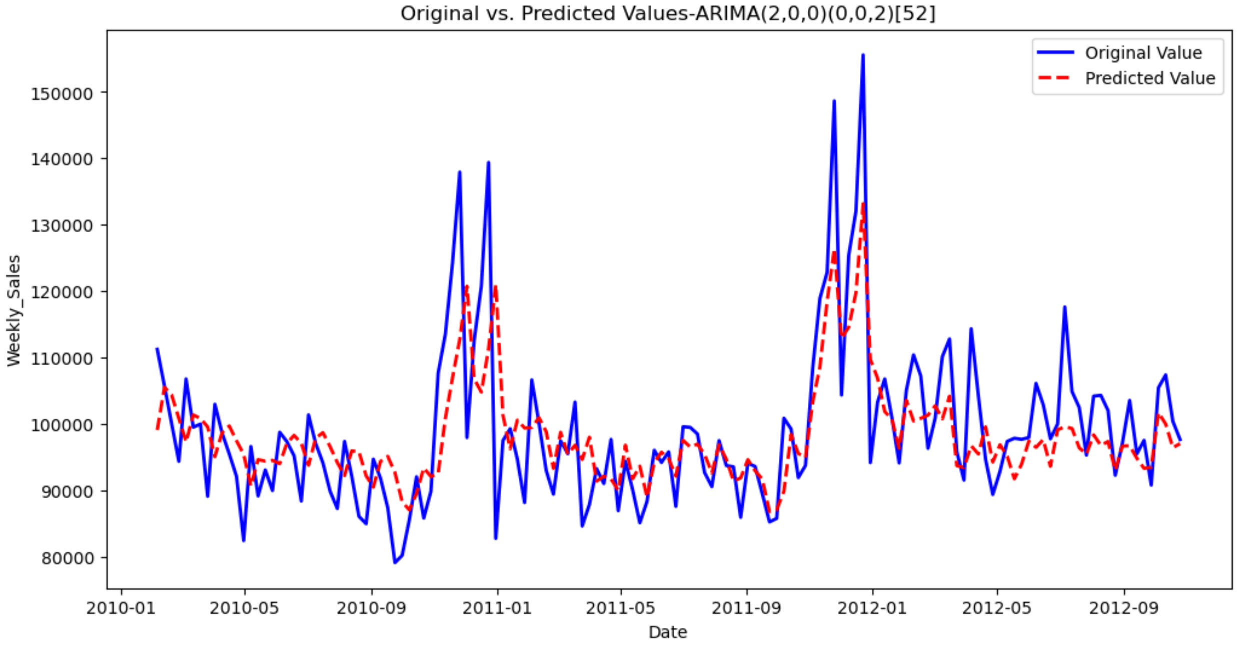
ARIMA algorithm summary:

Simple ARIMA model and grid_search method has been applied to this time series, gridsearch method is used to select optimal (p,q,d) values which give minimum MAPE. Here R^2 is not selected as a criterion for comparing the models because the mean in R^2 calculations does not capture the trend or seasonality for time series, the model may generate negative R^2 values.

MAPE for both training and testing dataset are evaluated and are presented in the following table. MAPE is improved by using gridsearch to optimize parameters (p,q,d) to (1,0,4), as opposed to initial (1,1,1). Furthermore, MAPE for testing dataset is further improved by additional 1% when AUTO_ARIMA is employed, which takes seasonal factors into consideration.

detailed model description	MAPE for training	MAPE for testing
ARIMA(1,1,1)	0.089	0.079
ARIMA(1,0,4)_GridSearch	0.065	0.05268
AUTO_ARIMA((2,0,0)(0,0,2)[52]	0.063	0.04824

As shown in the following plot, the training dataset contains strong seasonality, which is not well captured with ARIMA model, at the same time, testing dataset after 2012-01-20 does not show high seasonality/trending, therefore the MAPE values for testing dataset have lower value comparing to training dataset



Linear regression summary:

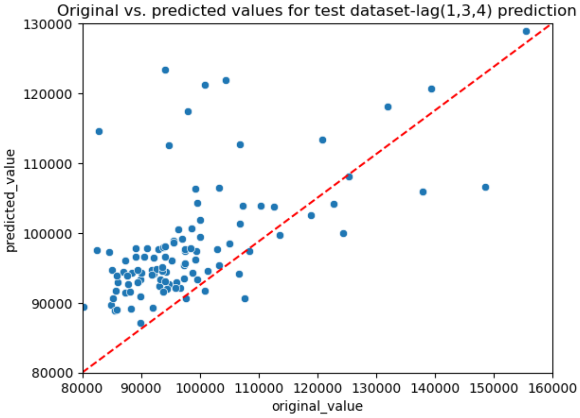
For linear regression, there are 10 different models that have been evaluated.

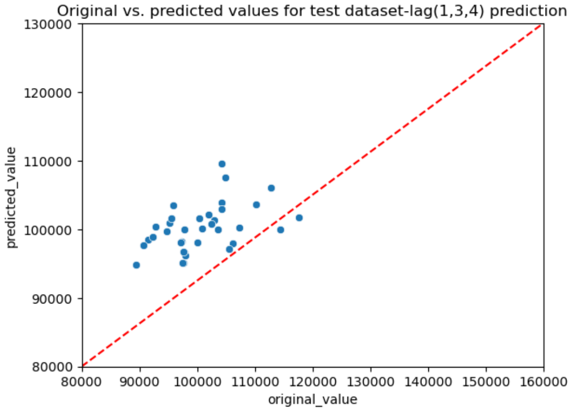
Linear regression with lagged weekly_sales

The first part is to predict weekly sales with different lagged values. The lagged values include first lag of weekly sales (lag_1), lagged values from previous 1-5 time periods(lag_1-5), lagged values from previous 1-7 time periods(lag_1-7) and discrete lagged values from previous 1,3,4 periods(lag_(1,3,4)) using grid search. Results are shown in the following table.

detailed model description	R^2 for training	R^2 for testing	MAPE for training	MAPE for testing
linear regression with lag_1	0.27	-0.1228	0.08	0.0503
linear regression with lag_1-5	0.42	0.0879	0.07	0.0495
linear regression with lag_1-7	0.43	0.1927	0.07	0.0475
linear regression with lag_(1,3,4)	0.37	0.2078	0.073	0.0444

R^2 values for training dataset are always higher than testing dataset suggesting overfitting issues. Below is example of training and test dataset prediction comparison.For most of the data points, the predicted values are higher than the true value. The training dataset show higher proximity to the diagonal line than the test dataset.

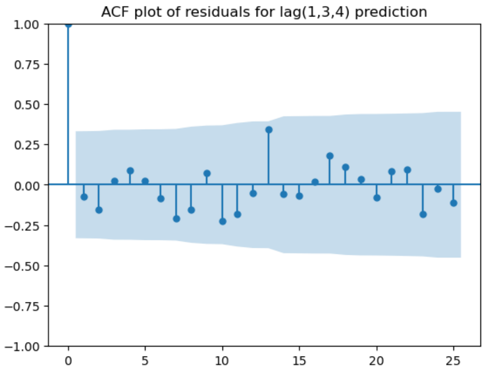


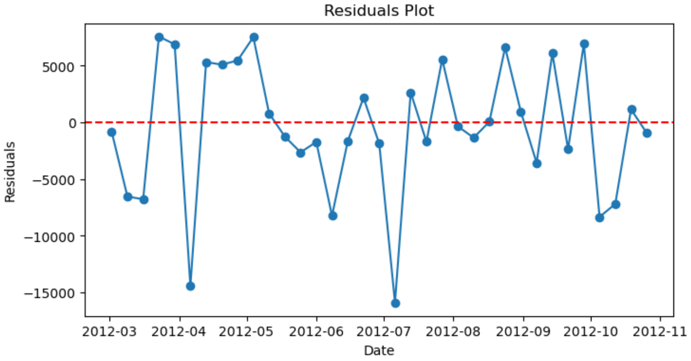


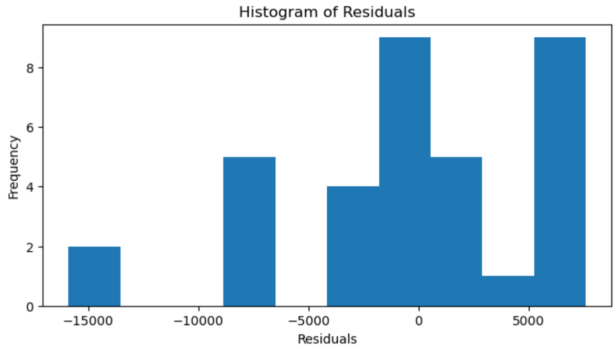
The residual analysis is shown in the following table and example of residual distribution plot is followed.

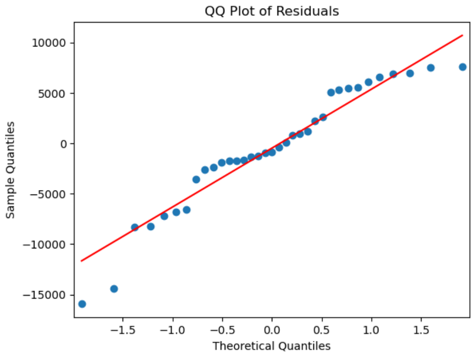
For all the models except “lag_1 model”, the parameters from residuals analysis for test dataset suggests residuals are nearly normally distributed. In lag_1 model, low p_values and high J-B numbers suggest the residuals are not normally distributed. This violates the assumption for R^2 calculations, therefore affecting its validity and leading to negative R^2.

detailed model description	skewness of residuals:	kurtosis of residuals	J_B of residuals:	JB_p_value
linear regression with lag_1	-0.87	0.57	5.04	0.08
linear regression with lag_1-5	-0.46	-0.34	1.43	0.49
linear regression with lag_1-7	-0.49	-0.29	1.53	0.46
linear regression with lag_(1,3,4)	-0.69	0.2	2.84	0.24



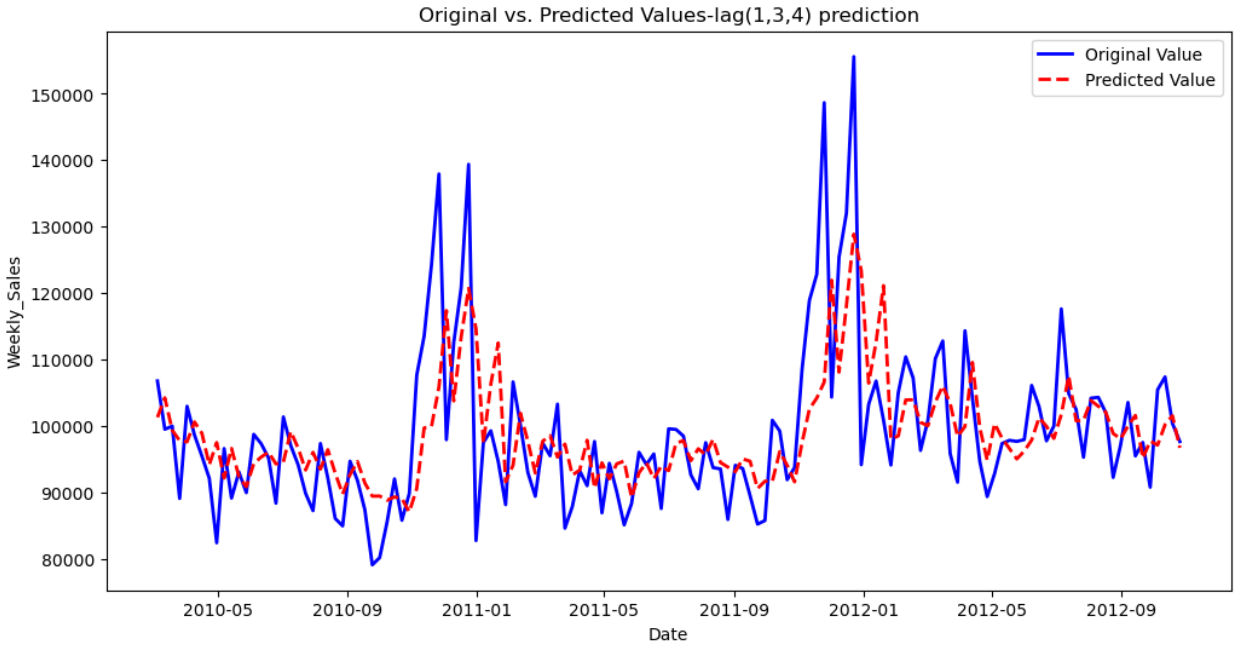






However, MAPE values show different trend from R^2. MAPE values for test dataset is always lower then training dataset, as mentioned earlier in ARIMA model, training dataset has high seasonality which is difficult for the model to capture, whereas the change in test dataset is more moderate.

With the Grid search method, optimal lag values is found to be (1,3,4) with highest R^2 values and lowest MAPE for test dataset. The comparison between predicted value and original value is plotted below.

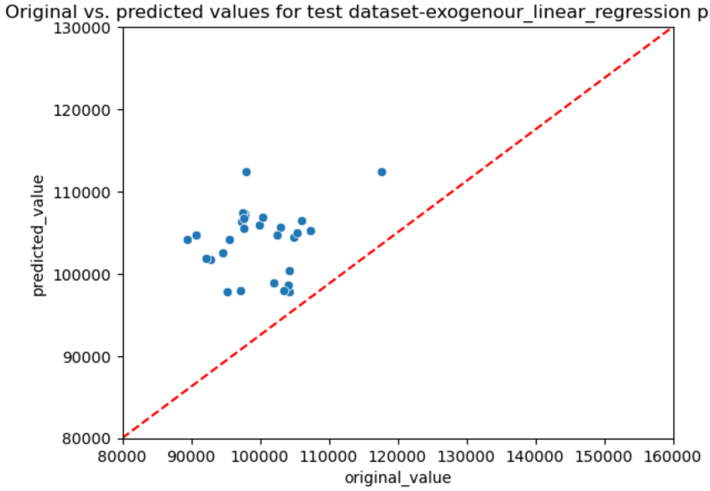


<u>Linear regression with exogenous features</u>

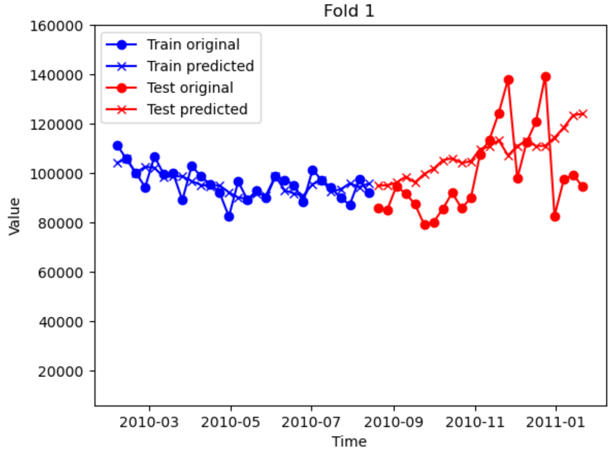
The second part is to predict weekly sales with exogenous features. The features include customer price index(indicator of inflation) , holiday_names, temperature, unemployment rate, fuel price and markdowns. The metrics of different models are listed in the following table.

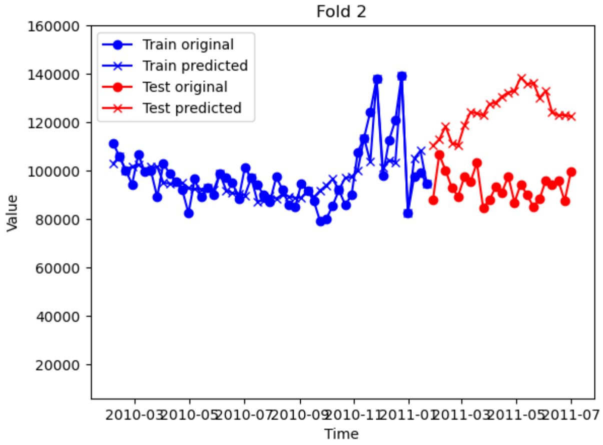
detailed model description	R^2 for testing	MAPE / mean MAPE for training	Variance of MAPE	MAPE/ mean MAPE for testing	Variance of MAPE
linear regression with exogenous features	-0.74	0.06		0.066	
linear regression with exogenous features_5-fold validation	-1349	0.05	0.007	1.1	1.89
linear regression with exogenous_walk_froward validation	-0.74	0.06	0.00	0.066	0.04
linear regression with exogenous-grid search	-1.36	0.074		0.067	
linear regression with exogenous and lagged values	-0.75	0.06		0.067	
linear regression with exogenous and lagged values-grid search	-0.08	-0.06		0.04	

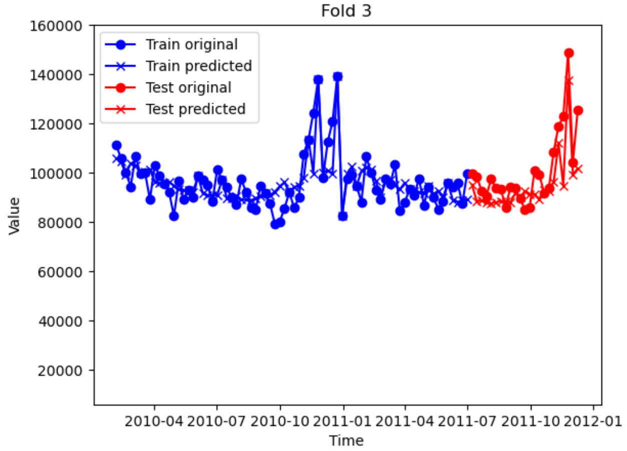
In simple regression, using exogenous features seems to yield worse performance than using lagged values, comparison between predicted values and original values for testing dataset are shown in the following plot.

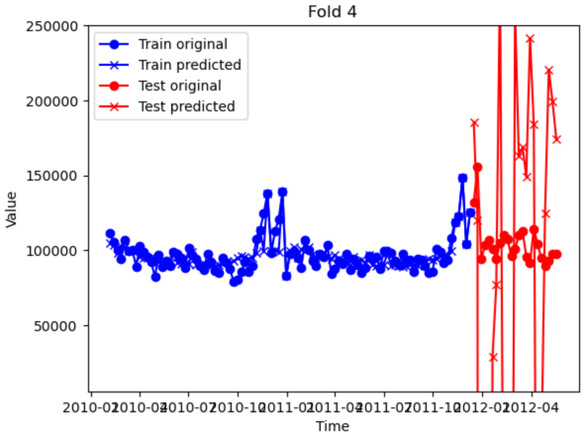


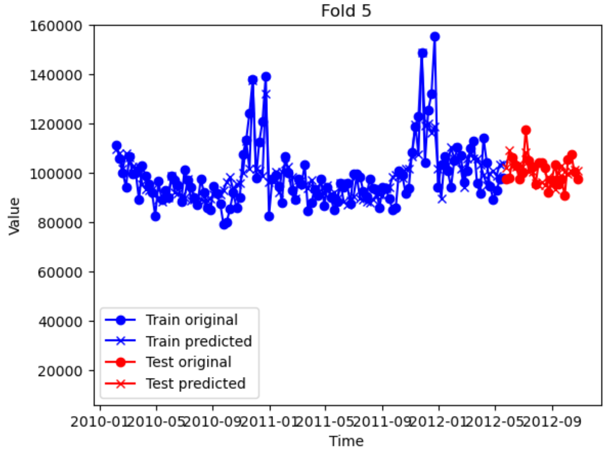
When adding 5-fold-validation step, the model becomes suspectible to underlying trends or seasonal patterns in data subsets. Notably, fourth fold display drastic values due to spike at the end of training data. This anomaly significantly distort the gap between predicted and actual values, skewing the regression model. As a result, R2 for test dataset is unusually high, the average MAPE exceeds 100% with high variance





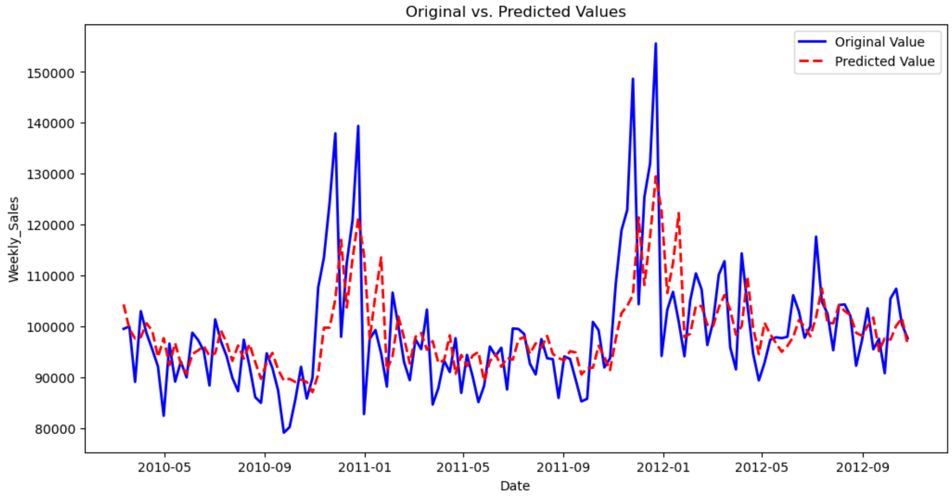






Using walk-forward validation allows the model to update itself with incoming data, capturing the most recent trends that k-fold validation might miss. This method respects the time sequence of the data, making it more suitable for validating time series models. This is proved by small variance observed in MAPE.

Lagged exogenous features is added into the model, the performance of the model is improved as MAPE changes from 0.066 to 0.4 with grid search method. The four key features are Thanksgiving Day, Weekly Sales_lag_4, MarkDown5_lag_1 and Weekly Sales_lag_1, the prediction is shown in the following plots.



Tree based regression

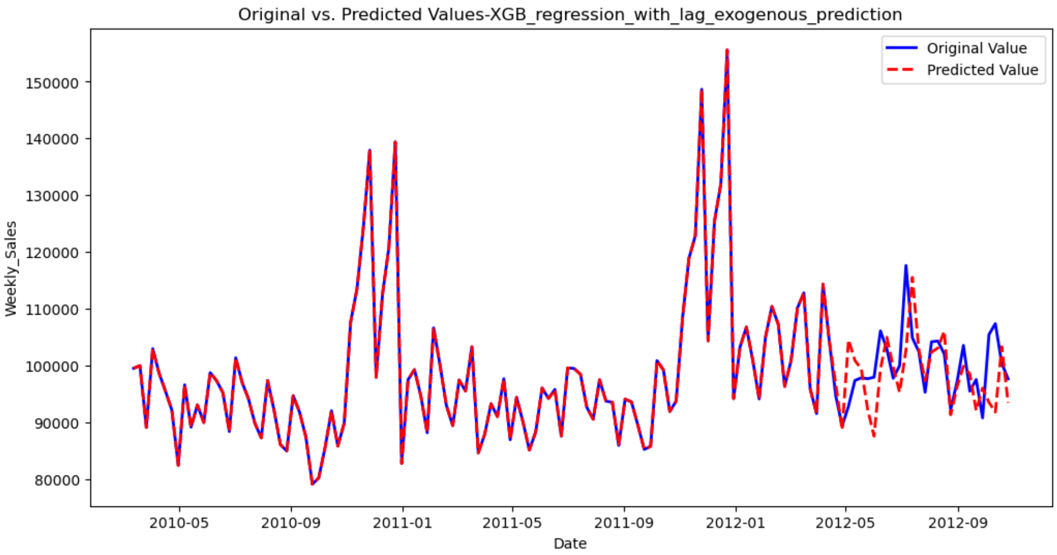
In the initial Random Forest model, the top three features were identified as Christmas, Martin Luther King Day, and Columbus Day. When using a train/test split and one-step validation methods, the training dataset consistently gets a MAPE variance of 0 but the test dataset has a variance of 0.07. This suggests the model is slightly less stable when generalized to unseen data. Upon incorporating lagged exogenous features, I didn't observe a significant improvement in goodness-of-fit or MAPE values.

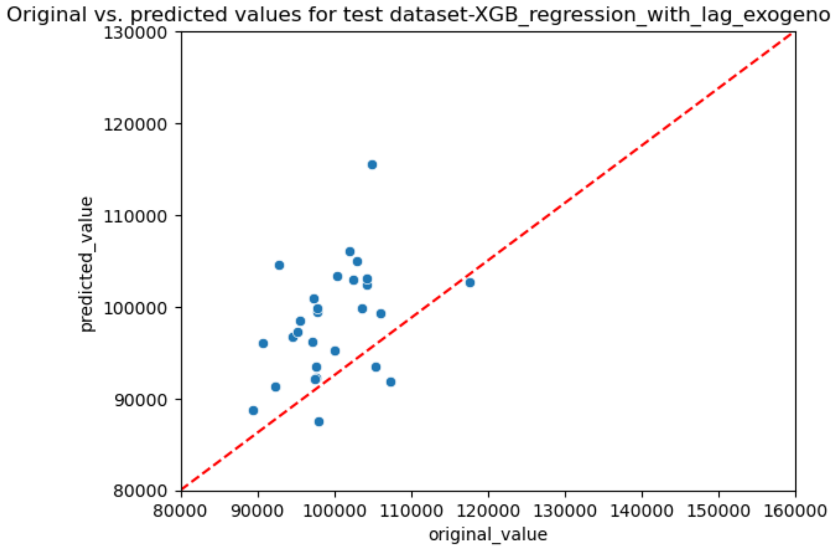
Based on simple XG Boost, the top 4 features are 'New Year s Day', 'MarkDown3', 'MarkDown3_lag_5', 'Temperature_lag_4','Weekly Sales_lag_1'. For all the XG_boost models, they all show perfect fitting with training dataset, but not for testing dataset. The overall performance of XG-boost is better than random forest with lowest MAPE for test dataset at 0.049. With Grid search, I was not able to find better performed model than basic XG-boost model may due to the fact that the parameters in grid search is limited.

detailed model description	R^2 for testing	MAPE / mean MAPE for training	Variance of MAPE	MAPE/ mean MAPE for testing	Variance of MAPE
Random forest regression with exogenous features	-2.2	0.04		0.08	
Random forest regression with walk_froward_validation	-2.2	0.04	0	0.08	0.07
Random forest regression with _lagged_exogenous_features	-3.73	0.03		0.1	
Random forest regression with lagged_exogenous_features_walk _froward_validation	-3.73	0.03	0	0.1	0.08
Simple XG_Boost		0		0.049	

XG_Boost with grid search		0		0.094	
XG_Boost with selected features		0		0.069	

The results of best performed model-simple XG_Boost is shown in the following:





Prophet model

Prophet models have relatively good R2 and low MAPE in testing but varying performance when exogenous variables and holidays are added.

detailed model description	R^2 for testing	MAPE for test
simple_prophet	0.16	0.04
added_holiday_prophet	0.17	0.042
added_exgeneous_holiday_prophet	-0.73	0.06

lagged_exgeneous_holiday_prophet_grid_search	0.14	0.036

Model Complexity: Simple models like 'simple_prophet' are performing similarly to complex grid-search-based models, implying that additional complexity is not necessarily improving performance.

High Variance Models: Some models like 'Random forest regression with walk_froward_lagged_exogenour' have significant discrepancies between their training and testing MAPE, suggesting they might be overfitting.

Best R2 Score: The highest R2 score is for the model 'linear regression with lag(1,3,4)' at 0.2078, followed closely by 'linear regression with lag1-7' at 0.1927. Most algorithms, however, have negative R2 scores, which indicates poor fits.

Overall, the 'lagged_exgeneous_holiday_prophet_grid_search' model has the lowest MAPE for testing at 0.036, closely followed by 'simple_prophet' and 'added_holiday_prophet' both at 0.04.

4.2 Auto Machine Learning for Multiple Time Series

The AutoML package, mljar, will be employed to analyze multiple time series. Due to my computer's CPU limitations, the study will focus on a subset of 50 time series. Mljar offers a free platform for AutoML and includes a diverse range of machine learning models like decision trees, linear regression, random forests, XGboost, LightGBM, Catboost, Neural Networks, and ensemble methods. All relevant outcomes, including the selected models, test dataset MAPE, and predictions, will be compiled in a dictionary named "model_dict".

The results of selected model and MAPE for test dataset is shown in the following table.

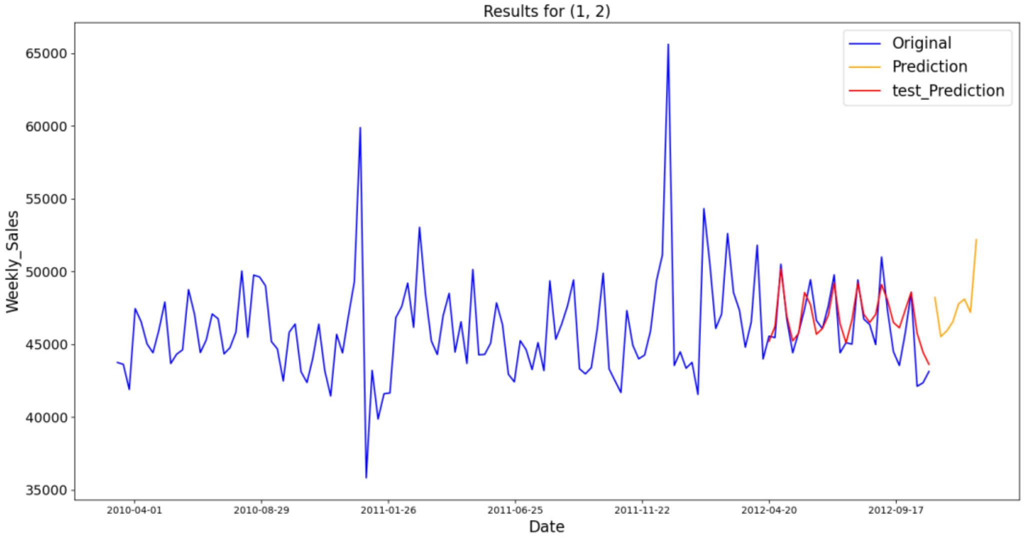
Index	model_name	val_MAPE
(1, 1)	Ensemble	0.036588491
(1, 2)	Ensemble	0.022880428
(1, 3)	Ensemble	0.102711383
(1, 4)	Ensemble	0.034287863
(1, 5)	Ensemble	0.089661202
(1, 7)	Ensemble	0.065940652
(1, 8)	Ensemble	0.032639058

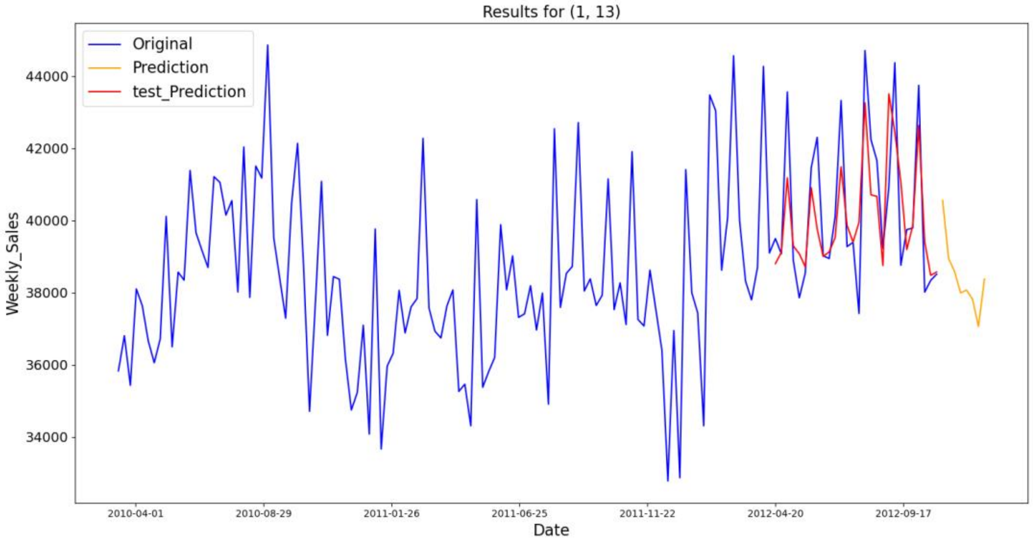
(1, 9)	Ensemble	0.091993969
(1, 10)	Ensemble	0.060693532
(1, 11)	Ensemble	0.11546191
(1, 12)	Ensemble	0.069269739
(1, 13)	Ensemble	0.024706026
(1, 14)	Ensemble	0.088031065
(1, 16)	Ensemble	0.079380506
(1, 17)	Ensemble	0.061286704
(1, 19)	Ensemble	0.167445631
(1, 20)	Ensemble	0.113268197
(1, 21)	Ensemble	0.065800082
(1, 22)	Ensemble	0.086555572
(1, 23)	Ensemble	0.086386558
(1, 24)	Ensemble	0.123138657
(1, 25)	Ensemble	0.0875075
(1, 26)	Ensemble	0.088320692
(1, 27)	Ensemble	0.123534503
(1, 28)	Ensemble	0.130092019

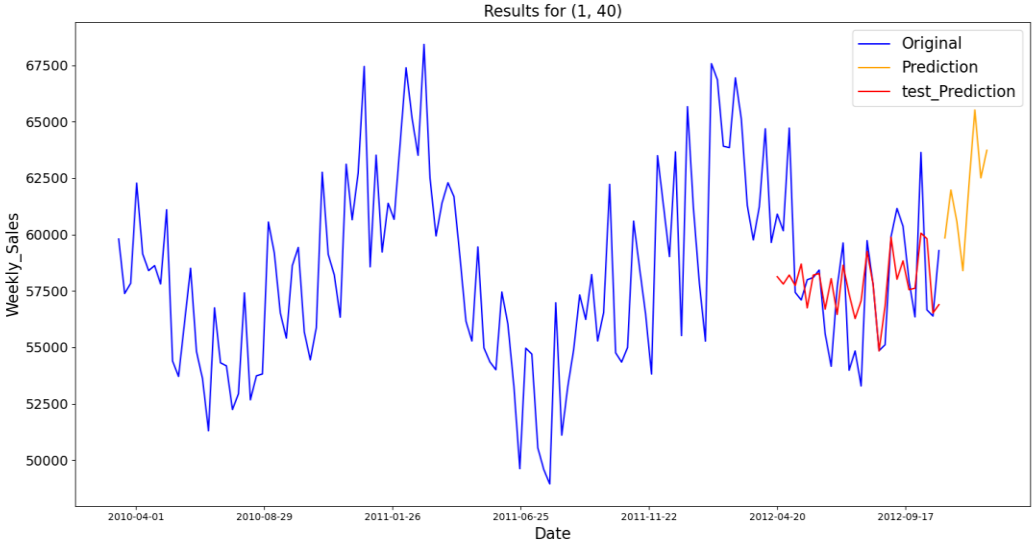
(1, 29)	Ensemble	0.065723621
(1, 30)	Ensemble	0.128446197
(1, 31)	Ensemble	0.190918333
(1, 32)	Ensemble	0.135011641
(1, 33)	Ensemble	0.125898524
(1, 34)	Ensemble	0.059239158
(1, 35)	Ensemble	0.139254297
(1, 36)	Ensemble	0.333464833
(1, 37)	Ensemble	0.067644359
(1, 38)	Ensemble	0.056337
(1, 40)	Ensemble	0.029778584
(1, 41)	Ensemble	0.165471755
(1, 42)	Ensemble	0.075061359
(1, 44)	Ensemble	0.075042024
(1, 46)	Ensemble	0.042453815
(1, 49)	Ensemble	0.121323123
(1, 52)	Ensemble	0.133201003
(1, 55)	Ensemble	0.091233782

(1, 56)	Ensemble	0.107605313
(1, 58)	Ensemble	0.479122888
(1, 59)	Ensemble	0.159957116
(1, 60)	Ensemble	0.087710412
(1, 67)	Ensemble	0.082985515
(1, 71)	Ensemble	0.208093623
(1, 72)	Ensemble	0.080136841

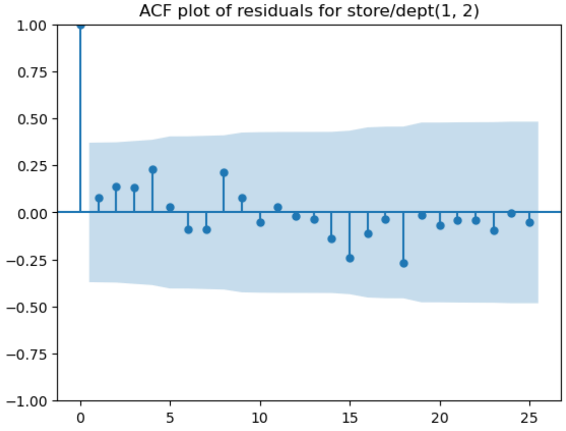
I ranked the top 3 dept/store with lowest MAPE values in the validation dataset. The top 3 store/dept are (1,2), (1,13),(1,40). The predicted values are shown in the plots as well.

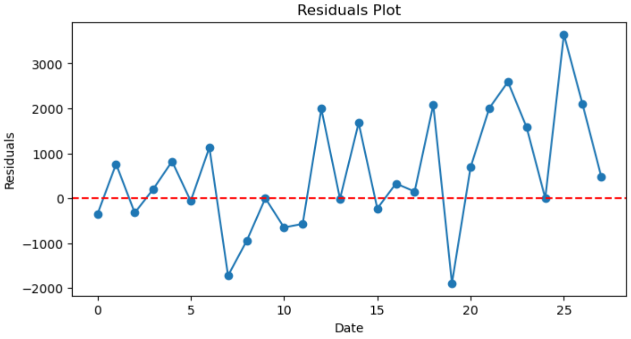


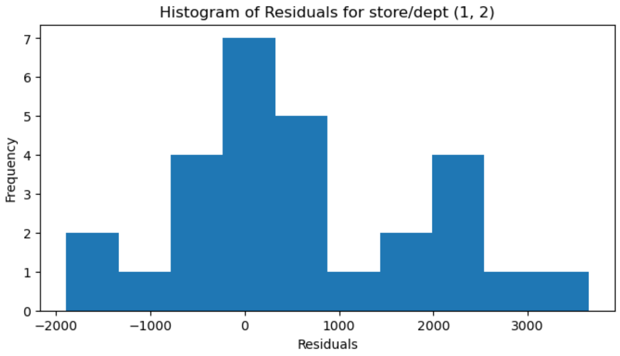


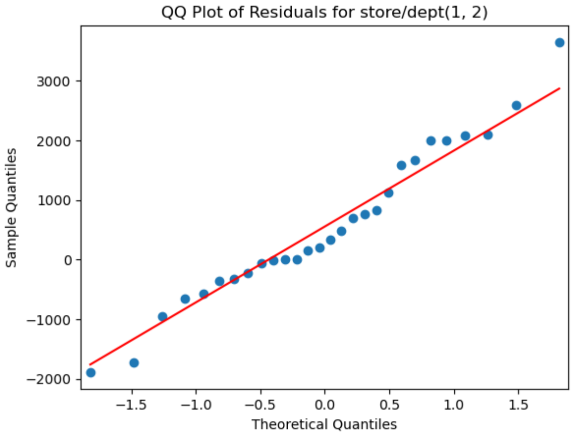


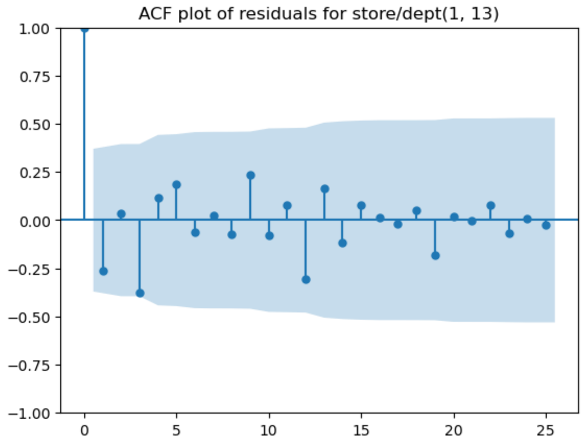
The residual plots for these three store/dept are shown in the following

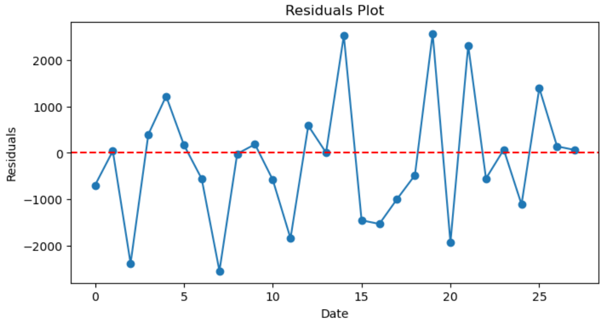


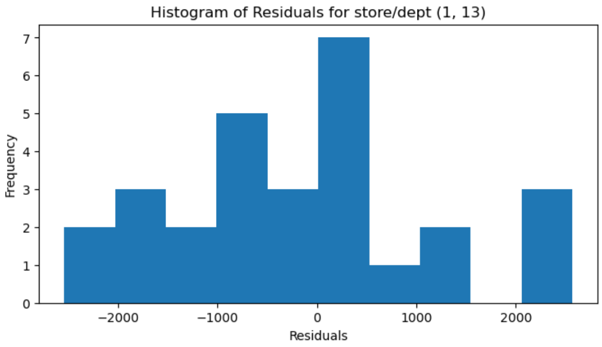


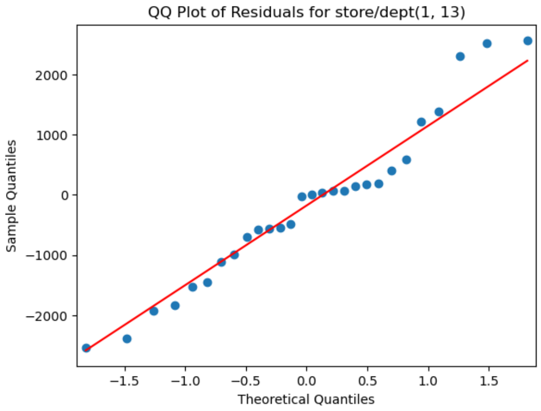

Histogram of Residuals for store/dept (1, 2)

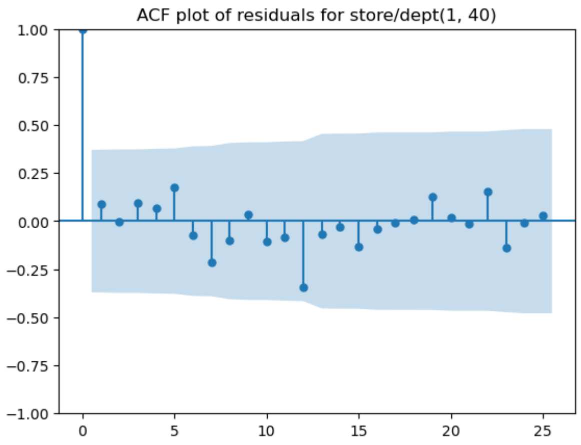

QQ Plot of Residuals for store/dept(1, 2)

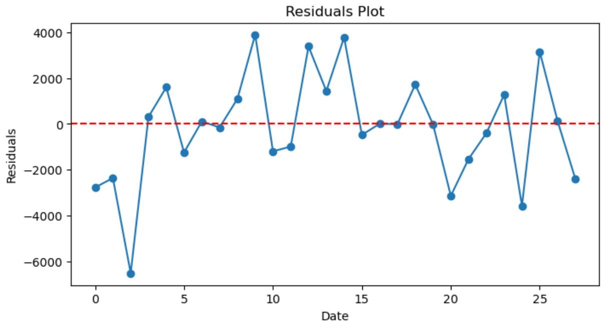

ACF plot of residuals for store/dept(1, 13)

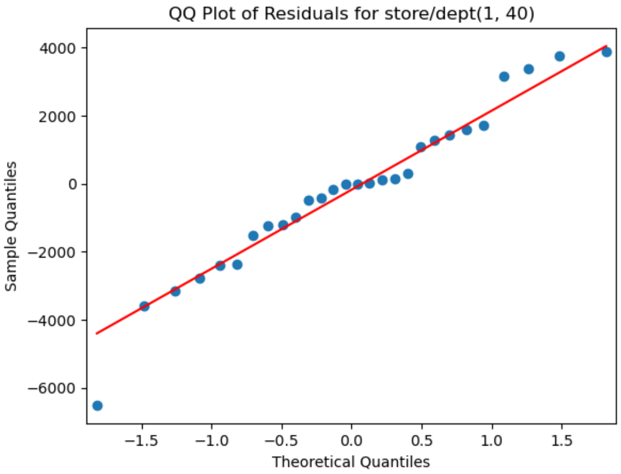

Residuals Plot

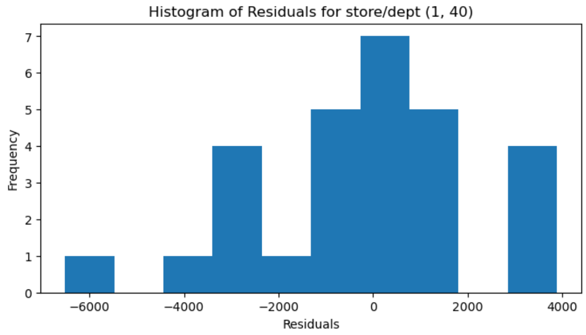

Histogram of Residuals for store/dept (1, 13)


QQ Plot of Residuals for store/dept(1, 13)

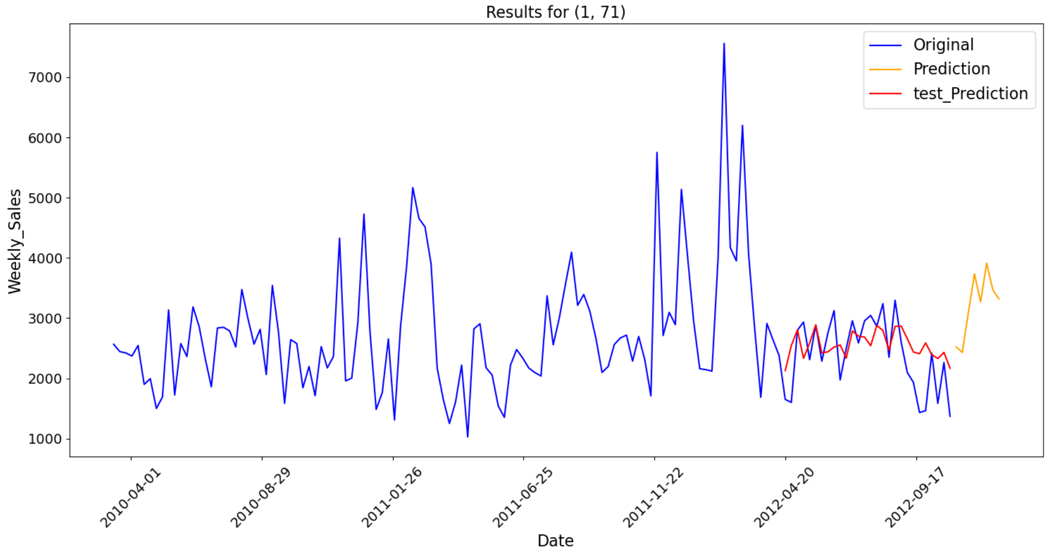


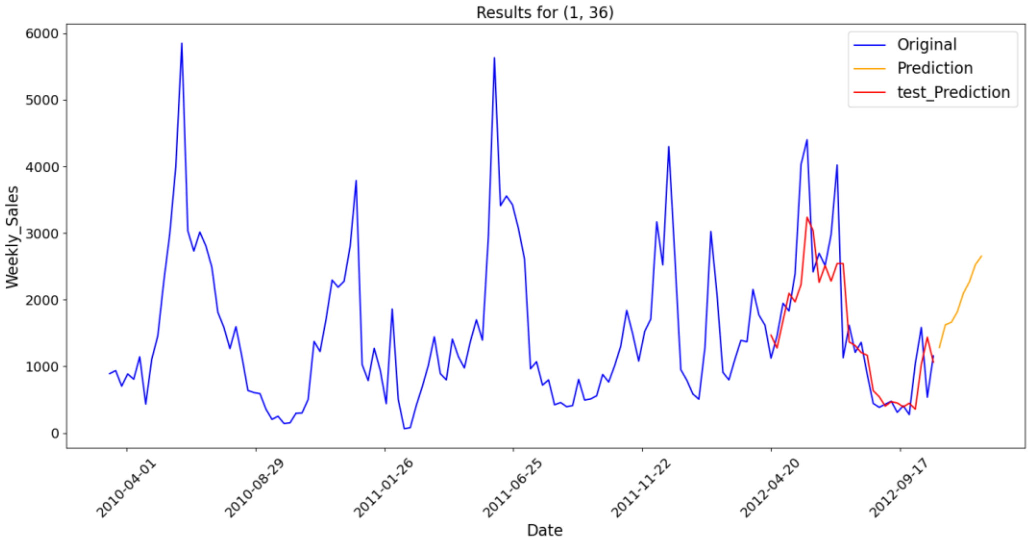


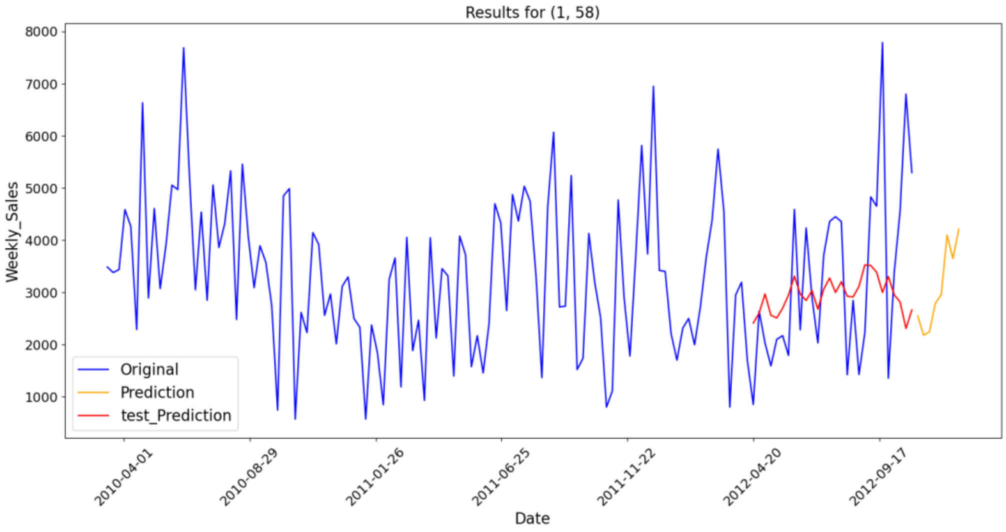




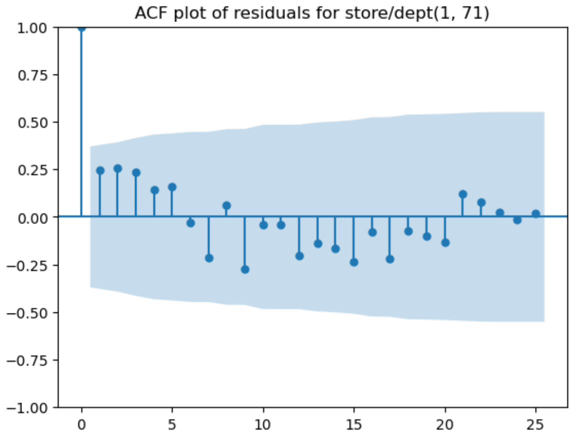
The lowest ranking of store/dept are (1,71),(1,36),(1,58). Results are shown in the following:

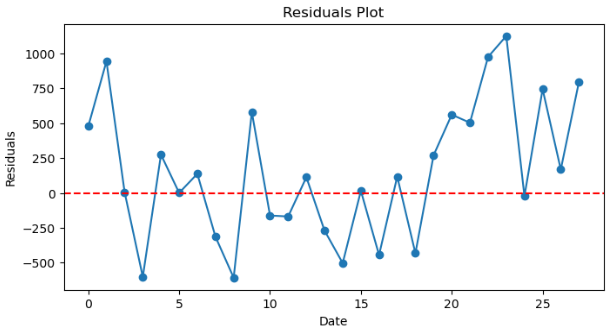


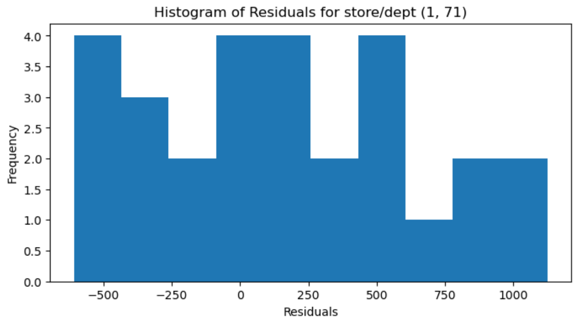


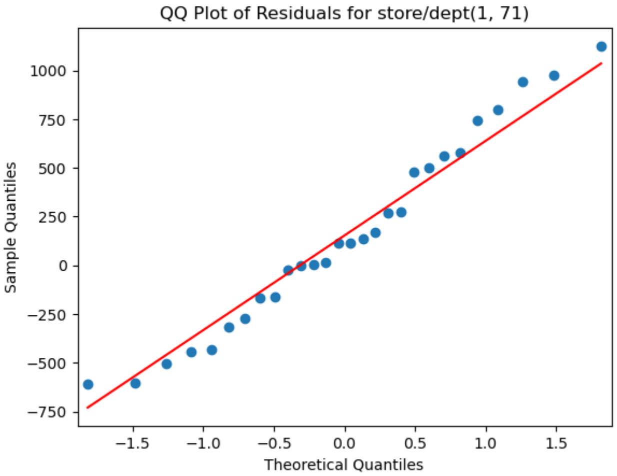


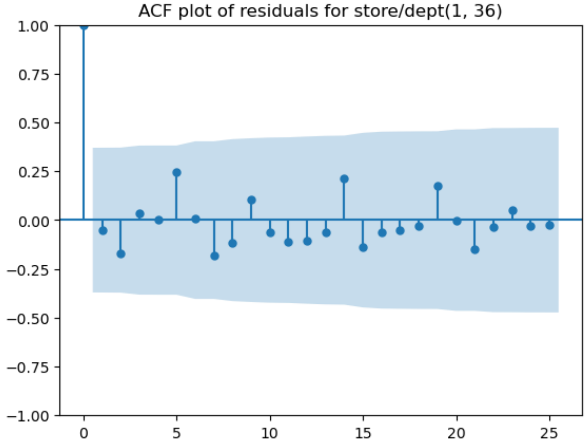
Residual plots for lowest ranking store/dept:

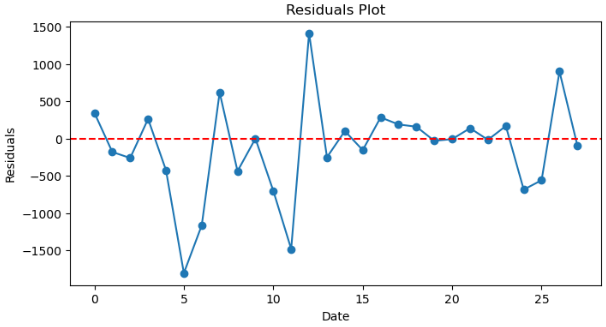


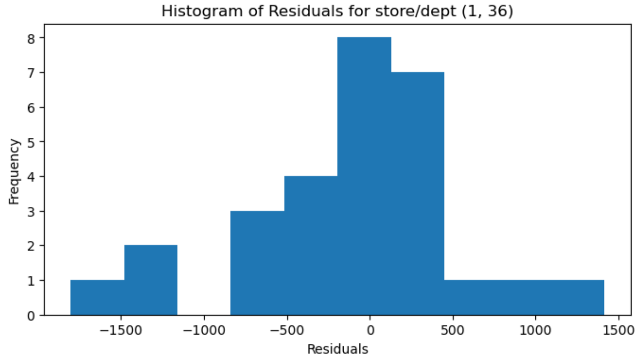


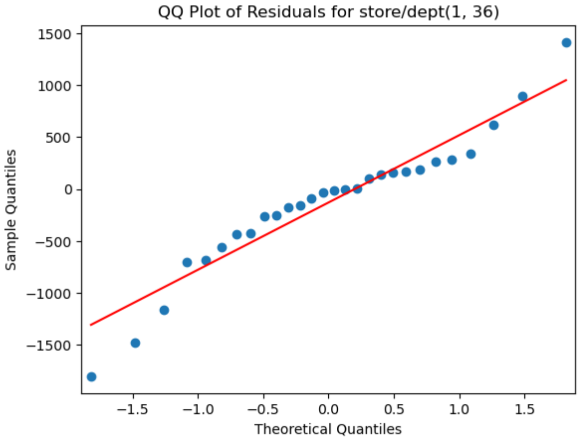

Histogram of Residuals for store/dept (1, 71)
Frequency
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
−500
−250
0
250
500
750
1000
Residuals

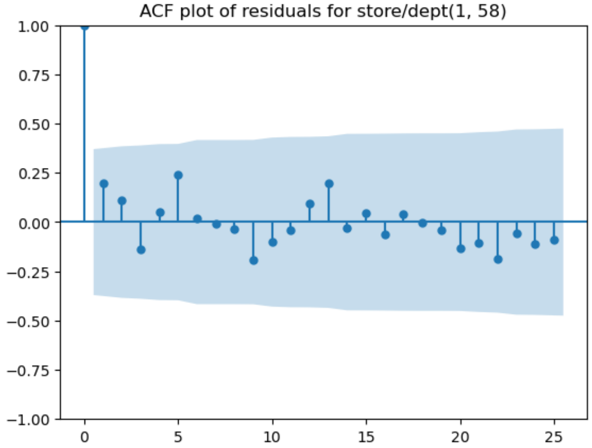

QQ Plot of Residuals for store/dept(1, 71)
Sample Quantiles
1000
750
500
250
0
−250
−500
−750
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
Theoretical Quantiles

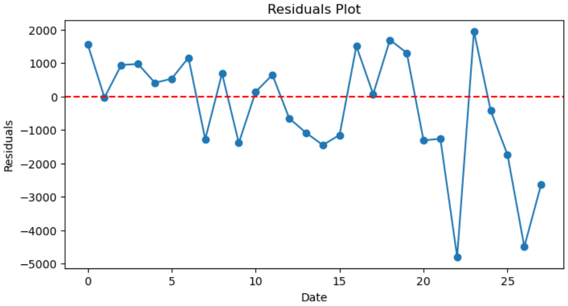

ACF plot of residuals for store/dept(1, 36)
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
0
5
10
15
20
25

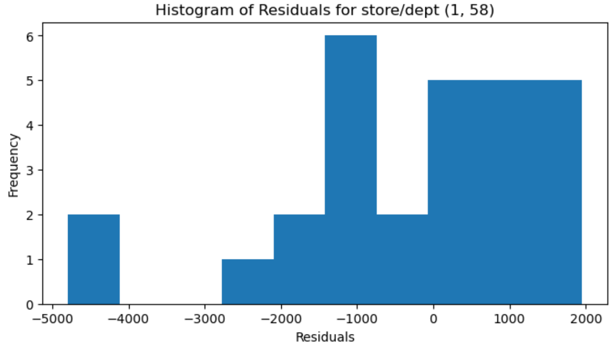

Residuals Plot
Residuals
1500
1000
500
0
−500
−1000
−1500
0
5
10
15
20
25
Date

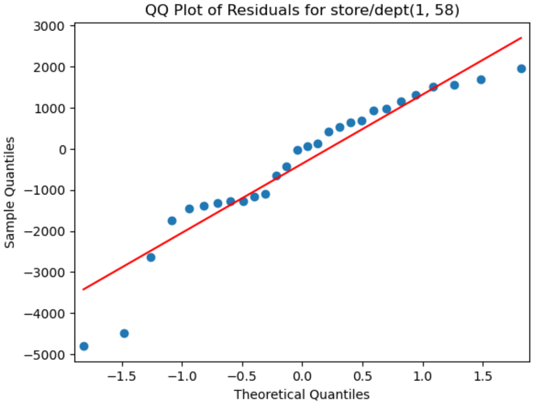

Histogram of Residuals for store/dept (1, 36)
Frequency
0
1
2
3
4
5
6
7
8
−1500
−1000
−500
0
500
1000
1500
Residuals


QQ Plot of Residuals for store/dept(1, 36)
Sample Quantiles
1500
1000
500
0
−500
−1000
−1500
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
Theoretical Quantiles


ACF plot of residuals for store/dept(1, 58)
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
0
5
10
15
20
25


Residuals Plot
2000
1000
0
−1000
−2000
−3000
−4000
−5000
Residuals
0
5
10
15
20
25
Date


Histogram of Residuals for store/dept (1, 58)
6
5
4
3
2
1
0
Frequency
−5000
−4000
−3000
−2000
−1000
0
1000
2000
Residuals


QQ Plot of Residuals for store/dept(1, 58)
3000
2000
1000
0
−1000
−2000
−3000
−4000
−5000
Sample Quantiles
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
Theoretical Quantiles