QBUS2820Group

Assignment2Task2

2017S2

Chenxi Zhou	**450091755**
Yunfeng Guo	**460012953**
Haonan Zhang	**450003747**
Yiran Jing	**460244129**

Part 1: Business context and problem formulation

Rossmann operates over 3000 drug stores in 7 European countries. In this report, we use the dataset from Kaggle competition to forecast six weeks daily sales for several stores by developing a univariate forecasting model.

The objective of this forecast is to ensure that store manager could make an effective staff arrangement that enhances the productivity and motivation. The time series forecasting relies on basic understanding of time series composition and model selection. Therefore, we perform EDA to understand the pattern with the time-varying data and select ARIMA and Exponential moving average method as potential choices to formulate robust model.

Part 2: Exploratory data analysis

Data processing

To begin with, we aggregate all 1115 stores daily sales to obtain the total sales among stores per day, because we try to obtain more information instead of overfitting a particular characteristic on one store by picking a store randomly. Then we discover that the sales on day 7(Sunday) fluctuate within a range between 0 and a small amount compared to the sales on other days. After checking the dataset, nearly all stores have no sales on Sunday, hence we delete the all data for Sunday. Then we check the daily aggregated sales when 80% of stores are closing, by searching the date when open = 0. There are several days that majority stores are not open, which is shown in table 1 below.

Table 1: Dates when majority stores are not open

Date							
01/01/2013	21/04/2014	01/05/2013	03/10/2013	26/12/2013	06/04/2015	01/05/2014	03/10/2014
01/01/2014	01/04/2013	26/12/2014	09/06/2014	25/12/2014	01/05/2015	09/05/2013	25/05/2015
01/01/2015	03/04/2015	29/03/2013	14/05/2015	25/12/2013	20/05/2013	18/04/2014	29/05/2014

Because we only predict future sales by learning the sales itself, we need to minimize the effect from other events. Therefore, these sales need to be replaced by a reasonable number. Because we find the seasonality in EDA part is a week, we replace these numbers by the sales on the same DayOfWeek of the closest cycle.

Our treatment to the dataset is that the data is rescaled by dividing 10^5 so as to maintain the sales data within the range from 60 to 80. This is essential to generate more precise forecasts. We did not perform the log transformation for the sales data, since this will lead to the scale of data relatively small between 10 and 20. Also, there is not a clear increasing trend in the sales data across year, making log transformation irrelevant to make the data stationery. Rescaling the data performs well in our model.

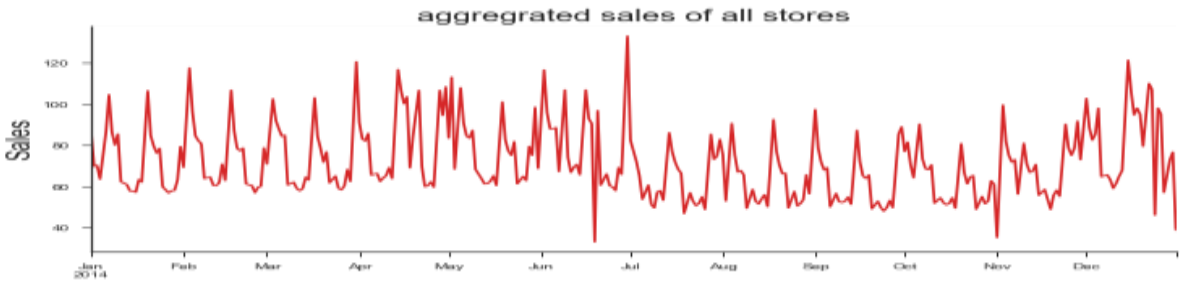


Figure 1: Aggregated Sales of 1115 Stores

From figure 1, there is a descending trend in the second half of 2014. It could be considered as cyclic pattern, but it could not be proved because there is no forward series to confirm it is included in a long cycle. Therefore, we check the dataset, then discover that there are 180 stores closed in that period and reopen in 2015. Because we could not verify it is a trend within a cycle, we determine to exclude the effect resulting from closing of a certain number of stores from dataset. After excluding that 180 stores from our dataset, we obtain figure 2.

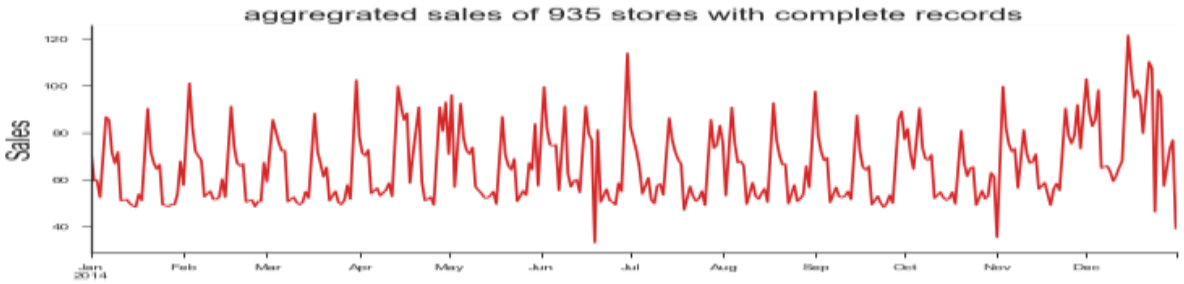


Figure 2: Aggregated Sales of 935 Stores with Complete Records

From figure 2, we think there is no apparent trend as well as cyclic. On the other hand, the seasonality may be more ostensible, but we need to explore further. Therefore, we perform time series decomposition to clearly obtain these components by plots.

Time series decomposition

Figure 4 shows the time series plot, and its trend, seasonal components. We can see that the trend component is not very clear in our case, since there is no significant increasing or decreasing trend over the time series period. In addition, the seasonal component is very clear, since the plot shows the pattern within each week and this pattern continues to present throughout the whole time series period. After the time series component decomposed, the residuals do not show any pattern and therefore we can say that residuals are randomly distributed.

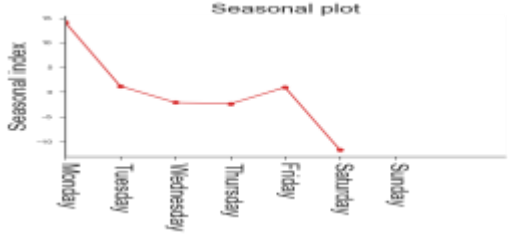


Figure 3: Seasonal Plot Within a Week

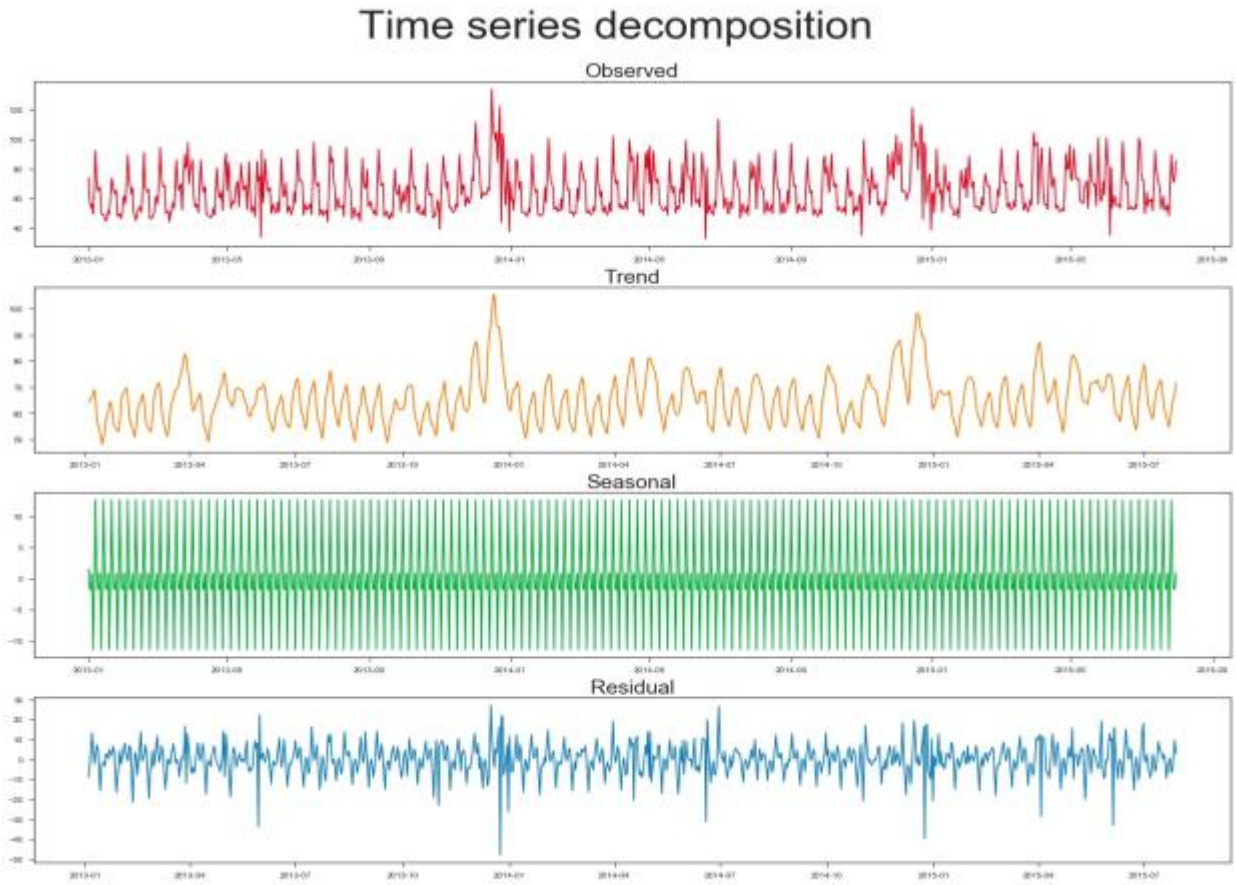


Figure 4: Basic Time Series Decomposition

Part 3: Modelling

Based on our exploratory analysis, we found the seasonal pattern appears within this time series. Our first model is decided to be based on seasonal ARIMA.

$$\text{ARIMA}(p,d,q): (1-\sum_{i=1}^{p}\emptyset_i B^i)(1-B)^d Y_t = c + (1+\sum_{i=1}^{q}\theta_i B^i)\varepsilon_t$$

Where d represents the order of seasonal differencing, while p and q represents orders of non-seasonal AR and MA components.

Seasonal ARIMA model is $\text{ARIMA}(p,d,q)(P,D,Q)_m$ where D is the order of seasonal difference, P and Q are the orders of the seasonal AR and MA components, and m is the number of periods per season.

Seasonal difference

$$\Delta_m Y_t = Y_t - Y_{t-m}$$

$$where\ m\ is\ the\ number\ of\ days\ per\ season$$

The graph for seasonal differences demonstrates a lag of 6, representing a week lag, which is included in the appendix.

The figure 6 below illustrates the autocorrelation and partial autocorrelation for the seasonal differences with a lag of a week. We can figure out the AR or MA model in the seasonal lags of the PACF and ACF. We found that a spike at lag 6 exists but there are still some other lags that are

significantly beyond the confidence interval. The PACF shows that exponentially decay in the seasonal lags at lags 6, 12, 18 and so on. The final values for the parameters are chosen in the model selection that maximizes the selection criteria.

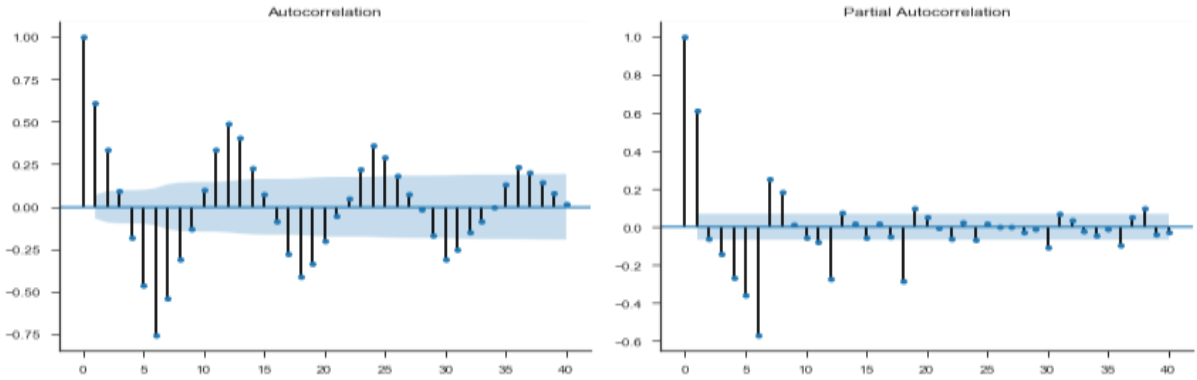


Figure 5: Autocorrelation and Partial Autocorrelation for Seasonal Difference

First difference $\Delta Y_t = Y_t - Y_{t-1}$

The first difference series graph (appendix) also shows the pattern of first differences between the sales of consecutive days.

Figure 8 below shows the autocorrelation and partial autocorrelation graph for the first difference. Autocorrelation graph helps determine whether time series needs differencing. On the one hand, ACF of a nonstationary series will decrease slowly. On the other hand, ACF of a stationery series should drop to zero relatively quickly. In our case, further differencing may be required, and the choice of parameters are determined in the following model selection process. In addition, partial autocorrelations measure the correlation between y_t and y_{t-k} after removing the effects over the lags between 1, 2, 3… k-1. In our case, the partial autocorrelation shows too many spikes that are outside the confidence interval, so the choices of parameters are determined in the following model selection process.

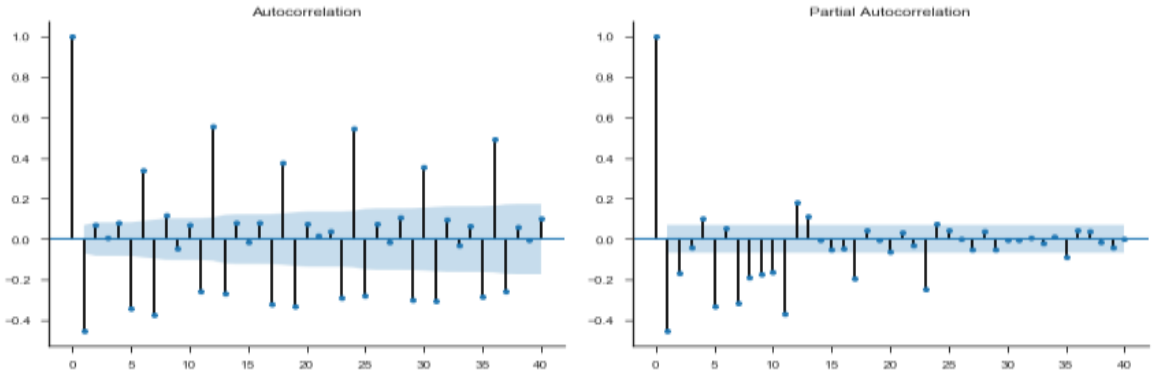


Figure 7: Autocorrelation and Partial Autocorrelation for First Difference

First and seasonal difference

$$\Delta_m (\Delta Y_t) = (Y_t - Y_{t-1}) - (Y_{t-m} - Y_{t-m-1}) \ \ where\ m\ is\ the\ number\ of\ days\ \text{per}\ season$$

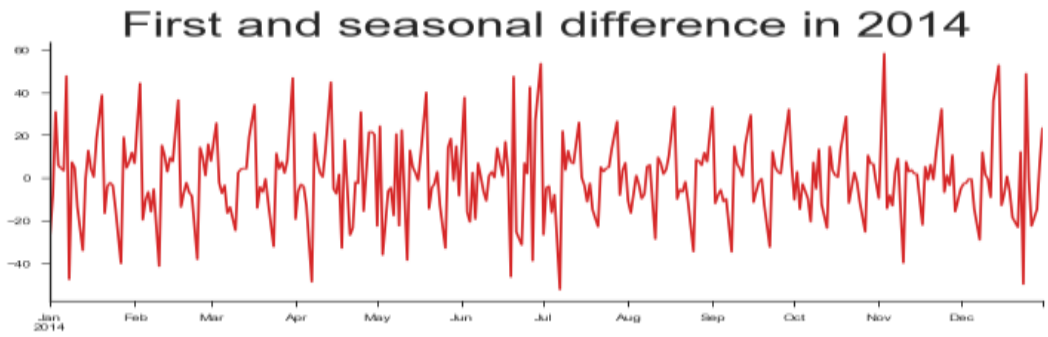


Figure 8: Pattern of First and Seasonal Difference in 2014

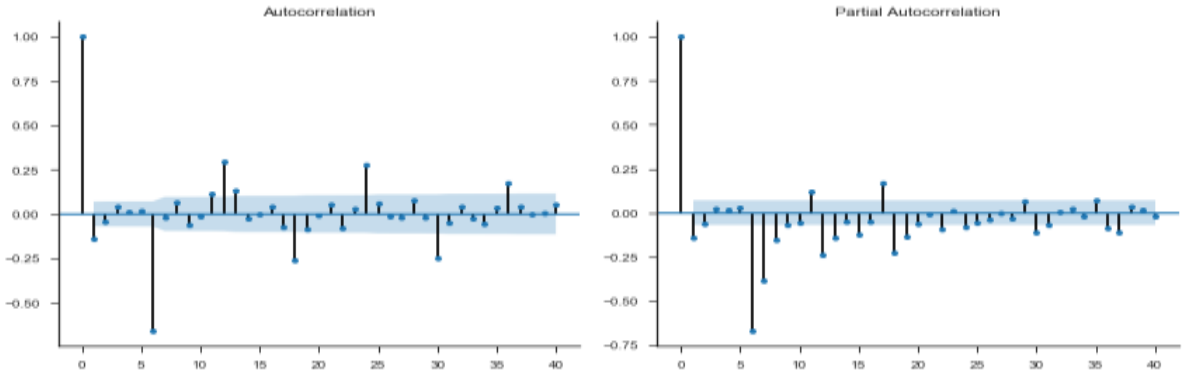


Figure 9: Autocorrelation and Partial Autocorrelation for First and Seasonal Difference

After conducting the first and seasonal difference, the autocorrelation plots still show the exponential decrease from the lag 6, 12, 18 and so on. The partial autocorrelation has a very large strike at lag 6 and after that the autocorrelation is not very large. Based on these fact, we consider that the possible values for number of periods per season is 6. The possible choice for seasonal and non-seasonal AR and MA orders would be 0 or 1 based on the exploratory data analysis result.

Part 4: Modeling

4.1.1 ARIMA model

We perform the automatic selection by considering the possible combination of values for non-seasonal component p, q and seasonal component P and Q. We should also consider seasonal difference and first difference based on EDA, since the difference would give us a relative stationery series compared the series without performing difference.

In the first part, we fix the seasonal component with (1, 1, 1) with m=6, and set the possible range for the non-seasonal components. The model selection is based on AIC. However, the optimal model based on lowest AIC has high degree of orders for AR and MA, whereas the reasonable time series would not have such high orders parameters and complexity. Based on this fact, we manually adjust the parameters and come up with optimal models with relative low AIC and low model complexity.

Model 1:

Table 2: Description for Model 1

Dep. Variable:		Sales	No. Observations:		808
Model:	SARIMAX(1, 1, 1)x(1, 1, 1, 6)		Log Likelihood		-3028.966
AIC	6067.933		BIC		6091.
	coef	std err	z	P>|z|	[0.025 0.975]
ar.L1	0.5907	0.022	26.467	0.000	0.547 0.634
ma.L1	-0.9997	0.155	-6.437	0.000	-1.304 -0.695
ar.S.L6	-0.2482	0.022	-11.314	0.000	-0.291 -0.205
ma.S.L6	-0.9999	3.068	-0.326	0.745	-7.014 5.014
sigma2	106.5631	327.067	0.326	0.745	-534.476 747.602

The model parameters Autoregressive Lag 1, Moving Average Lag 1 and Seasonal Autoregressive. Lag 6 are significantly different from 0 since p-value is less than 5% confidence level. However, the seasonal Moving average lag 6 is not statistical significant from 0 since the standard error is very large and P-value is greater than 0.05.

Model 2:

In this model, we keep the same range for non-seasonal component except for making no difference for the consecutive observations. The optimal model 2 is demonstrated below.

Table 3: Description for Model 2

Statespace Model Results

Dep. Variable:		Sales	No. Observations:		808
Model:	SARIMAX(1, 0, 1)x(1, 1, 1, 6)		Log Likelihood		-3012.546
AIC	6035.091		BIC		6058.564
	coef	std err	z	P>|z|	[0.025 0.975]
ar.L1	0.8054	0.019	42.466	0.000	0.768 0.843
ma.L1	-0.3364	0.031	-10.711	0.000	-0.398 -0.275
ar.S.L6	-0.3561	0.020	-17.584	0.000	-0.396 -0.316
ma.S.L6	-0.9883	0.015	-65.952	0.000	-1.018 -0.959
sigma2	103.7286	3.462	29.965	0.000	96.944 110.513

The model parameters are significantly different from 0 since p-value is less than 5% confidence level. This model, without using the first difference for the non-seasonal component, indicates a better forecasting performance relative to the model with non-seasonal first difference. The validation results for these models are displayed in the mode validation section.

4.1.2 Holt-winters Exponential Smoothing

The second methodology for the time series forecasting related to Holt-winters exponential smoothing. The Holt-winters' model is more appropriate in our case since it captures seasonality. The seasonal method comprises the forecast equation and three smoothing equations related to level, trend and seasonal component. In our case, we figure out the period of seasonality is 6 days.

Additive Holts-Winter Exponential Smoothing

$\hat{y}_{t+1} = l_t + b_t + S_{t+1-L}$, (forecast equation)

$l_t = \alpha(y_t - S_{t-L}) + (1-\alpha)(l_{t-1} + b_{t-1})$, (level)

$b_t = \beta(l_t - l_{t-1}) + (1-\beta)b_{t-1}$, (trend)

$S_t = \delta(y_t - l_t) + (1-\delta)S_{t-L}$ (seasonal indices)

for seasonal frequency L, initial values l_0, b_0 and S_{i-L} for $i = 1, \dots, L$ .parameters $0 \leq \alpha \leq 1, 0 \leq \beta \leq 1, 0 \leq \delta \leq 1$.

Multiplicative Holts-Winter Exponential Smoothing

$\hat{y}_{t+1} = (l_t + b_t) \times S_{t+1-L}$, (forecast equation)

$l_t = \alpha(y_t / S_{t-L}) + (1-\alpha)(l_{t-1} + b_{t-1})$, (level)

$b_t = \beta(l_t - l_{t-1}) + (1-\beta)b_{t-1}$, (trend)

$S_t = \delta(y_t / l_t) + (1-\delta)S_{t-L}$ (seasonal indices)

for seasonal frequency L, initial values l_0, b_0 and S_{i-L} for $i = 1, \dots, L$,parameters $0 \leq \alpha \leq 1, 0 \leq \beta \leq 1, 0 \leq \delta \leq 1$.

In addition, we have several variations related to model choice. In our case, EDA indicates that additive model expects to be a better time series model for forecasting since the seasonal variation is roughly constant through our series plot. However, we construct both additive and multiplicative model and compare the out-of-sample MSE to select the optimal model. The results of smoothing parameters, AIC and BIC for the Holt-winters Exponential smoothing are demonstrated in the table 4 below.

Table 4: Exponential Smoothing Model Results

Exponential smoothing model results					
Model	**alpha (level)**	**beta (trend)**	**delta (seasonal)**	**AIC**	**BIC**
Additive Holt-winters	0.679 (0.042)	0.000 (0.035)	0.216 (0.042)	6333.944	6352.723
Additive Holt-winters (damp=0.000(0.437))	0.433 (0.222)	0.572 (0.437)	0.095 (0.067)	6319.998	6343.471
Multiplicative Holt-winters	0.726 (0.032)	0.000 (0.037)	0.228 (0.043)	6324.236	6343.014
Multiplicative Holt-winters (damp=0.541(0.038))	0.726 (0.043)	0.000 (0.042)	0.227 (0.048)	6327.248	6350.72
Additive Holt-winters (log unadjusted AIC)	0.698 (0.039)	0.000 (0.035)	0.231 (0.016)	-483.331	-464.552

4.2 Residual diagnostics

Residual vs Time

In this section, residual plots versus time indicates that there is not a pattern in the residual changing across the time and the residuals are independent for both models. The residual plots vs Time for two models are in the appendix.

Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF)

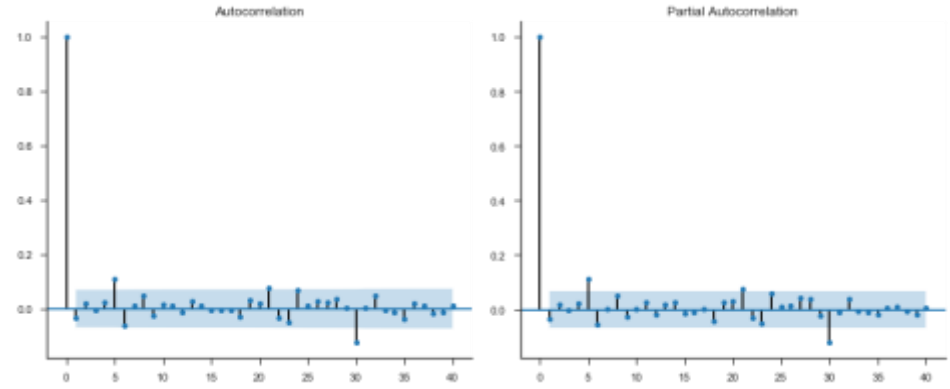


Figure 10: Autocorrelation and Partial Autocorrelation for Seasonal ARIMA Model

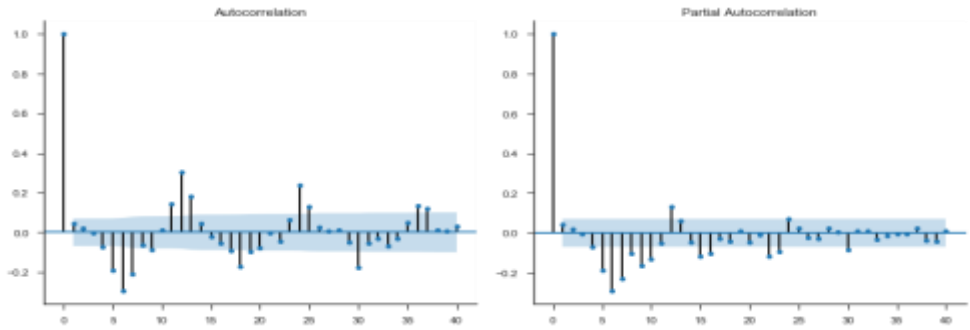


Figure 11: ACF and PACF for Additive Holt Winter Exponential Smoothing Model

The seasonal ARIMA model indicates that there are not many significant spikes in the autocorrelation plot and partial autocorrelation plot. However, there is still a pattern with the autocorrelation plot for the additive Holt winter exponential smoothing model and the autocorrelations are not fully removed.

4.2.3 Residual distribution

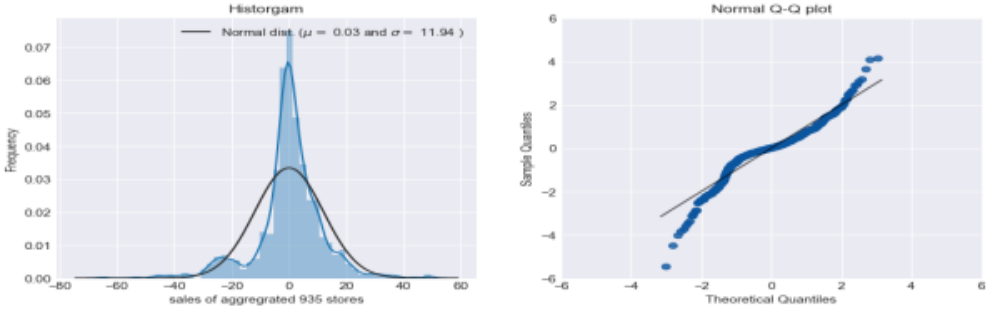


Figure 12 Residuals normality check for seasonal ARIMA

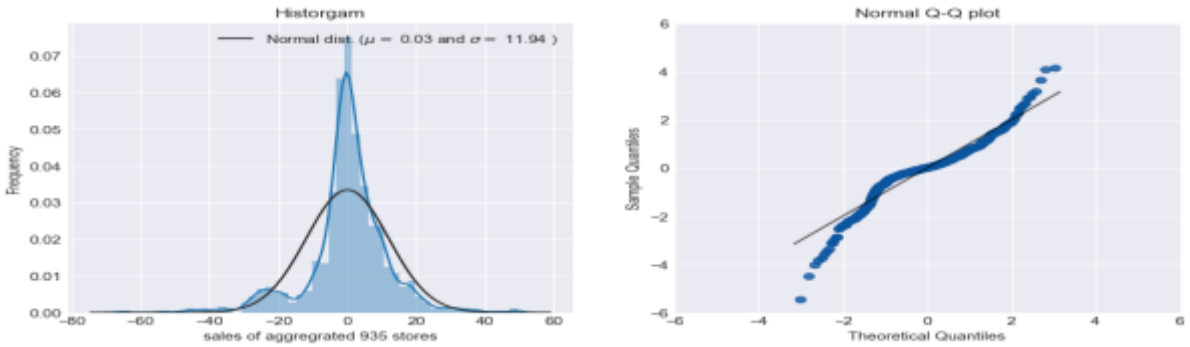


Figure 13 Residual normality check for Additive Holts-Winter damped

The residuals plot for the seasonal ARIMA is quite symmetric but Normal Q-Q plot indicates that normality assumption is violated. The kurtosis is high at 7.41. The violation of normality assumption may lead to inappropriate uses of maximum likelihood estimation. Despite of these, we are still confident that the model is appropriate since it is still a symmetric distribution.

The residual plots for Additive exponential smoothing model is not quite symmetric especially for the negative parts, which violates the use of maximum likelihood estimation. The skewness is -0.636 and the kurtosis is 3.870. We conclude that this model does not have good property.

Part 5: Model Validation

The table below gives us the results for model evaluation based on Root mean square error(RMSE).

Table 5: Model Evaluation for Forecasting Results

Model evaluation for forecasting results			
	RMSE 1 day ahead	**RMSE 1week ahead**	**RMSE 6 week ahead**
Simple exponential smoothing (benchmark)	15.94	15.95	15.98
Additive Holt-winters exponential smoothing	13.86	13.88	13.92
Additive damped Holt-winters exponential smoothing	13.80	13.82	13.84
Multiplicative Holt-winters exponential smoothing	14.09	14.12	14.14
Multiplicative damped Holt-winters exponential smoothing	14.02	14.04	14.08
Log additive Holt-winters exponential smoothing	**14.15**	**14.71**	**15.02**
Seasonal ARIMA $(1,1,1)(1,1,1)_6$	10.04	11.02	12.56
Seasonal ARIMA $(1,0,1)(1,1,1)_6$	9.90	10.95	12.39

Discussion of model evaluation

Our model evaluation considers both the short-term and medium-term forecasting so as to obtain a comprehensive assessment of the model validation results. In our case, the root mean square error based on 1 day ahead,1 week ahead and 6 weeks ahead do not change our choice of optimal model.

Our benchmark model is based on simple exponential smoothing. To improve our forecast, we change the scale of the original sales observation by dividing 10^5, and our comparisons result for different models based on root mean square error (RMSE) do not change, since changing the scale of data is monotonic. Based on this model and our exploratory data analysis results, we improve our model by including the seasonal components and the following model are presented and validated.

Exponential smoothing model

The optimal model is the additive holt winter exponential smoothing with damped trend. It has the lowest root mean square error (13.80) compared with other models. This is followed by the additive holts-winter model, which has the second smallest root mean square error (13.86).

However, the performance for the multiplicative holt winter exponential smoothing is not good enough no matter the model is damped or not. This is consistent with our previous expectation that there is not much variation in the seasonal trend changing with the time.

Seasonal ARIMA model

Our seasonal ARIMA are $Seasonal\ ARIMA\ (1,1,1)(1,1,1)_6$ and $Seasonal\ ARIMA\ (1,0,1)(1,1,1)_6$

The results indicate that the seasonal ARIMA models with or without difference significantly improve the forecast accuracy based on RMSE. The seasonal ARIMA without difference shows better performance with the RMSE at 9.90, followed by the seasonal ARIMA with first difference with RMSE 10.04. However, we can find that there is an increasing trend in the RMSE when we increase the lag for seasonal ARIMA model. Nevertheless, the validation results for seasonal ARIMA model still outperform the corresponding results of Holts-winters exponential smoothing model on average.

In conclusion, we select Additive Holt winters exponential smoothing with damped trend $with\ \alpha = 0.433, \beta = 0.572, \delta = 0.095$, and $Seasonal\ ARIMA\ (1,0,1)(1,1,1)_6$ to forecast sales.

Part 6: Forecast

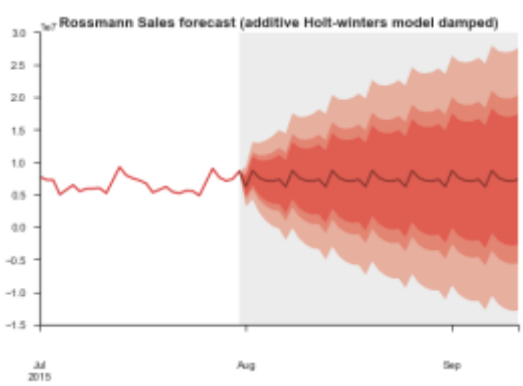


Figure 14 Forecasting for Additive Holt-winters

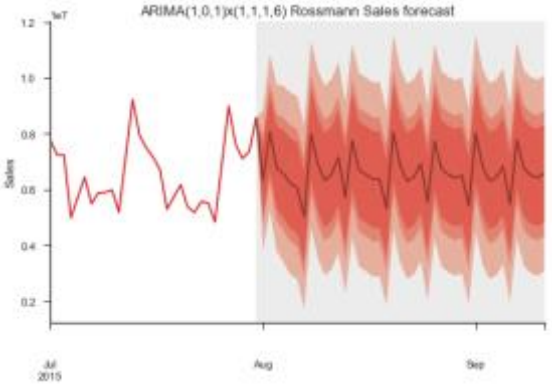


Figure 15 Forecasting for Seasonal ARIMA

In our forecast (figure 14 & 15), the blackline represents the point forecast for each day. Also, the interval forecasts were constructed with α of 0.2, 0.1, 0.05. There are significant differences for the interval forecast between the seasonal ARIMA model and additive Holt-winters model with damped trend. The additive model on the left shows that with greater further to the future estimation, the variance for the forecast becomes larger and larger.

This may potential indicate that the forecast based on additive Holt-winters model damped is not stationery and therefore is not a good forecast. However, ARIMA model indicates that the difference for the estimated variance between the most recent forecast and further forecast is not very significant. Although the scale for the unit of interval in y-axis is different between these two graphs, we could still find a similar trend between two models. Within a season, there is an increasing trend initially, and then decreasing at a higher rate. After the sales slightly reverse, it finally reaches the lowest sales during the week. In summary, the forecasts from ARIMA model is more appropriate and stable. Also, an exact forecast from 2 models for daily sales are demonstrated as a table in appendix.

Analysis of monthly sales

Figure 16 Monthly sales in Rossmann

In this part, we analyze the limitation with the monthly sales forecast based on the monthly data. The dataset contains the daily sales from the beginning of 2013 to the mid-2015. When the forecasting model is constructed based on the monthly data, we can only obtain 30 data points for the stores. The sample size is not enough for formulating the model and the coefficients estimation may have a higher variance due to this limited sample size. We still perform the monthly sales plot, and we found that the season patterns are not clear in this situation.

However, this does not make sense based on our knowledge that the sales should have a seasonal pattern. Therefore, it might not provide sufficient and appropriate data information for estimating the

model that captures the seasonal and trend components.

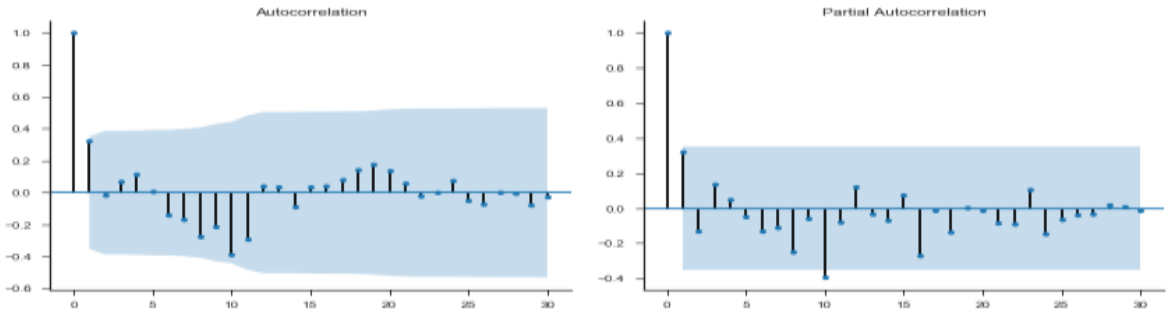


Figure 17 Autocorrelation and Partial Autocorrelation plot

Part 7: Further analysis for bonus marks

We develop a system to automatically generate forecast for all store through a for loop, which includes the following steps:

1. Isolating the stores that have missing data during 2014 from those stores with complete data.
2. For each store indexed in the store list, we fit into our multiple candidate models and select the model that has the lowest AIC, representing our best model.
3. Presenting the best model diagnostics which indicates a fundamental satisfaction of the assumptions. Note that: when there is a happy face, it represents that this store does not contain missing data during 2014, therefore fitting our optimal ARIMA model expects to have a reasonable forecast since the trends for these stores in aggregate are expected to be similar. However, we may not be confident towards the store with missing data during 2014, since our optimal model estimation does not incorporate their information. We expect that the model may not forecast well for these stores.

   However, we have concluded the following interest facts:

   1. The residuals diagnostic for the optimal model of each store have basically satisfied the following assumptions: the residuals are independent over time, and the forecast residuals are relatively normal and the scale of residuals is reasonable. Combining with these facts demonstrated by the plots for each store (included in the appendix), we could reassure that our automatic forecasting systems perform reasonably well, being adaptive to fit the optimal model for each store separately.
   2. After removing some stores that do not have observations during 2014, the optimal model based on 935 stores with complete data would still give a reasonable forecast for the stores with missing data, because the optimal model for approximately 80% of the stores is consistent with our previous optimal and $Seasonal\ ARIMA\ (1,0,1)(1,1,1)_6$

Appendix

A1 Forecast

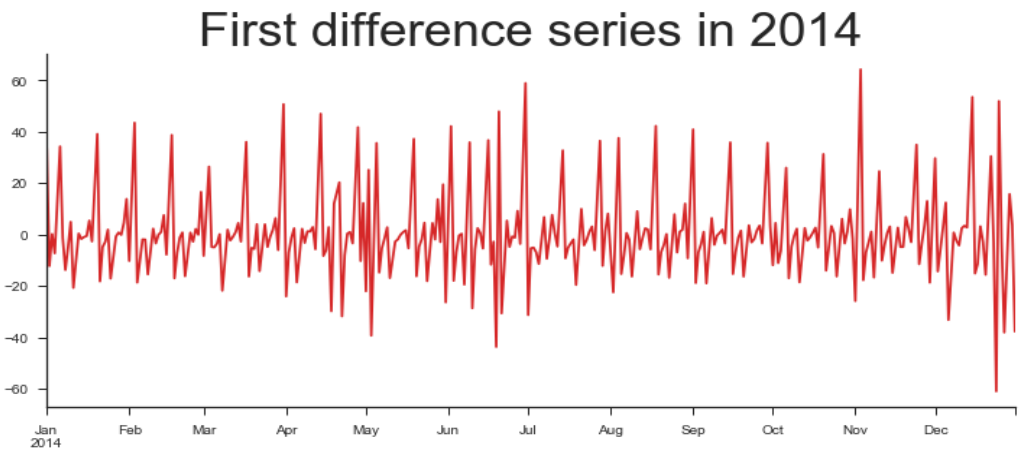
ARIMA forecast	
1/8/15	6.20E+06
2/8/15	0.00E+00
3/8/15	8.67E+06
4/8/15	7.58E+06
5/8/15	7.13E+06
6/8/15	7.04E+06
7/8/15	7.33E+06
8/8/15	6.20E+06
9/8/15	0.00E+00
10/8/15	8.67E+06
11/8/15	7.58E+06
12/8/15	7.13E+06
13/8/15	7.04E+06
14/8/15	7.33E+06
15/8/15	6.20E+06
16/8/15	0.00E+00
17/8/15	8.67E+06
18/8/15	7.58E+06
19/8/15	7.13E+06
20/8/15	7.04E+06
21/8/15	7.33E+06
22/8/15	6.20E+06
23/8/15	0.00E+00
24/8/15	8.67E+06
25/8/15	7.58E+06
26/8/15	7.13E+06
27/8/15	7.04E+06
28/8/15	7.33E+06
29/8/15	6.20E+06
30/8/15	0.00E+00
31/8/15	8.67E+06
1/9/15	7.58E+06
2/9/15	7.13E+06
3/9/15	7.04E+06
4/9/15	7.33E+06
5/9/15	6.20E+06

6/9/15	0.00E+00
7/9/15	8.67E+06
8/9/15	7.58E+06
9/9/15	7.13E+06
10/9/15	7.04E+06
11/9/15	7.33E+06

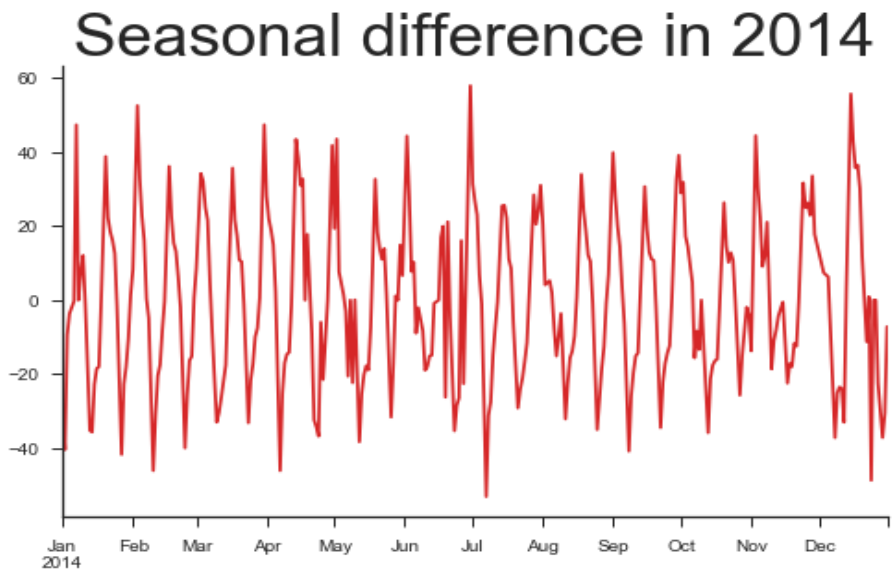
Additive HW forecast	
1/8/15	6.20E+06
2/8/15	0.00E+00
3/8/15	8.67E+06
4/8/15	7.58E+06
5/8/15	7.13E+06
6/8/15	7.04E+06
7/8/15	7.33E+06
8/8/15	6.20E+06
9/8/15	0.00E+00
10/8/15	8.67E+06
11/8/15	7.58E+06
12/8/15	7.13E+06
13/8/15	7.04E+06
14/8/15	7.33E+06
15/8/15	6.20E+06
16/8/15	0.00E+00
17/8/15	8.67E+06
18/8/15	7.58E+06
19/8/15	7.13E+06
20/8/15	7.04E+06
21/8/15	7.33E+06
22/8/15	6.20E+06
23/8/15	0.00E+00
24/8/15	8.67E+06
25/8/15	7.58E+06
26/8/15	7.13E+06
27/8/15	7.04E+06
28/8/15	7.33E+06
29/8/15	6.20E+06
30/8/15	0.00E+00

31/8/15	8.67E+06
1/9/15	7.58E+06
2/9/15	7.13E+06
3/9/15	7.04E+06
4/9/15	7.33E+06
5/9/15	6.20E+06
6/9/15	0.00E+00
7/9/15	8.67E+06
8/9/15	7.58E+06
9/9/15	7.13E+06
10/9/15	7.04E+06
11/9/15	7.33E+06

A2: First difference series

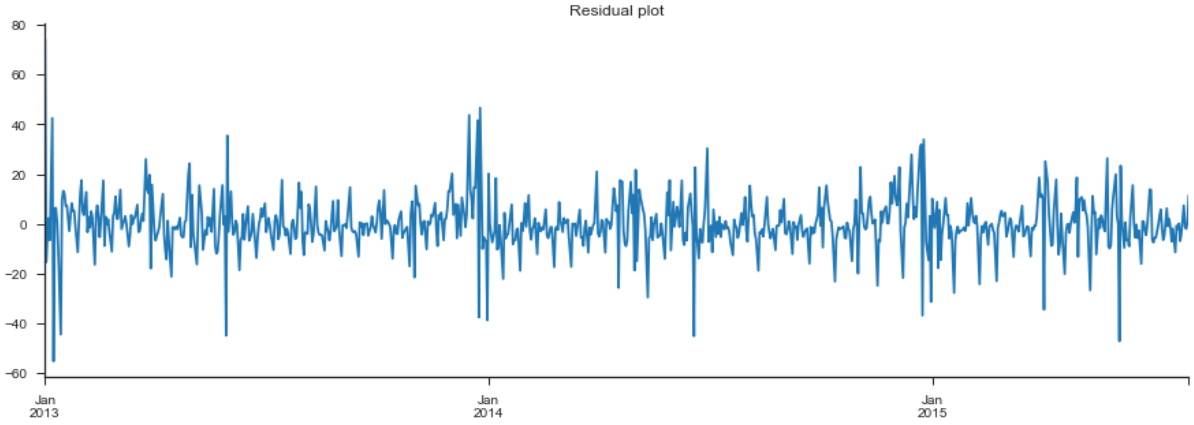


A3: Seasonal difference series

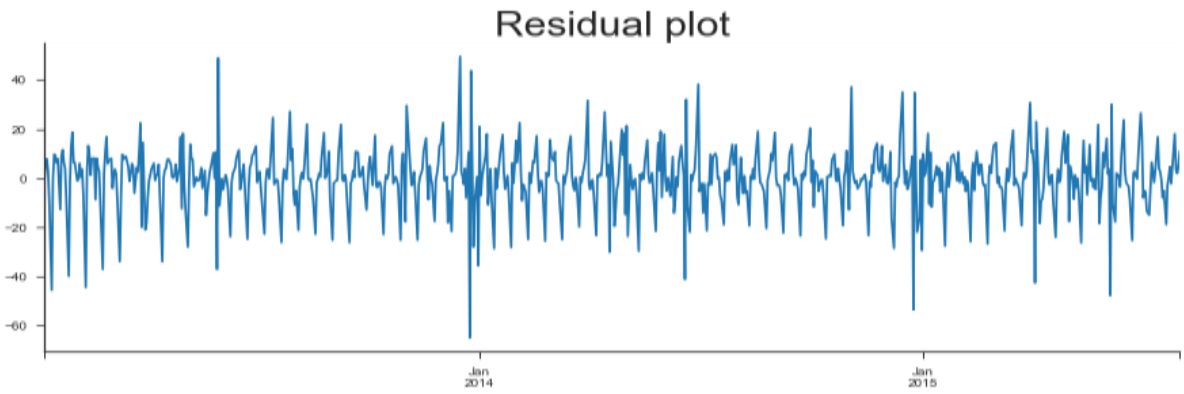


A4: Residual plot vs Time

The figure below shows the residual plot for the seasonal ARIMA model and Additive Holt winter exponential smoothing model.



(figure for the seasonal ARIMA model)



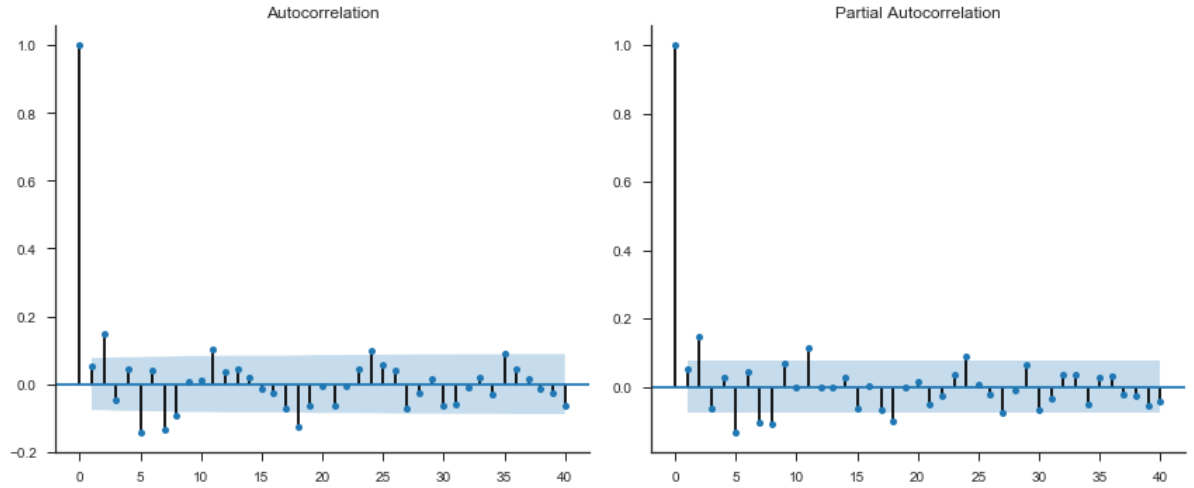
(figure for the Additive Holt winter exponential smoothing model)

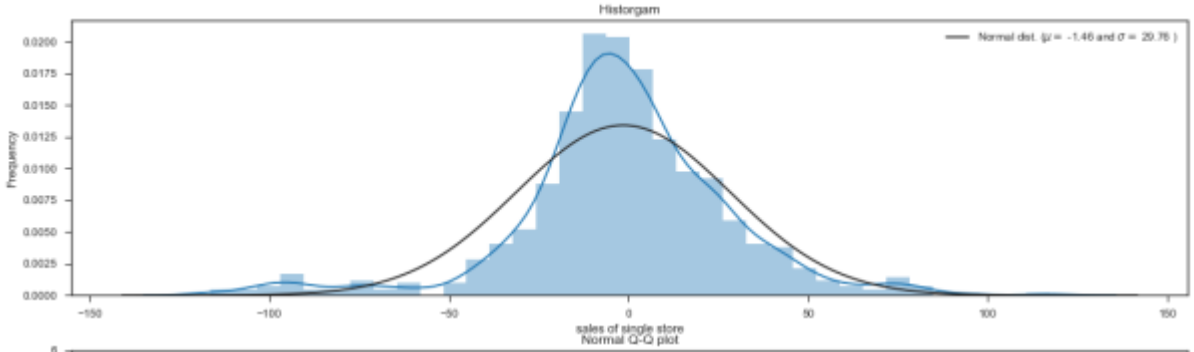
The residual plots versus time indicates that there is not a pattern in the residual changing across the time and the residuals are independent for both models.

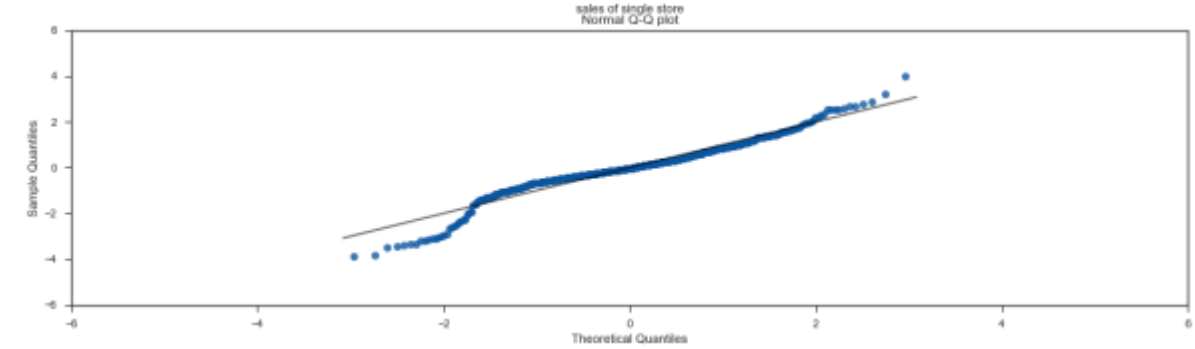
A5: For loop results for the bonus mark

The results indicate that most of the model have the best model selected based on AIC that are the same of seasonal ARIMA.

Store 279 (650 records): Bset model is ARIMA(0,1,1)X(1,1,1,6) with AIC: 6203.594

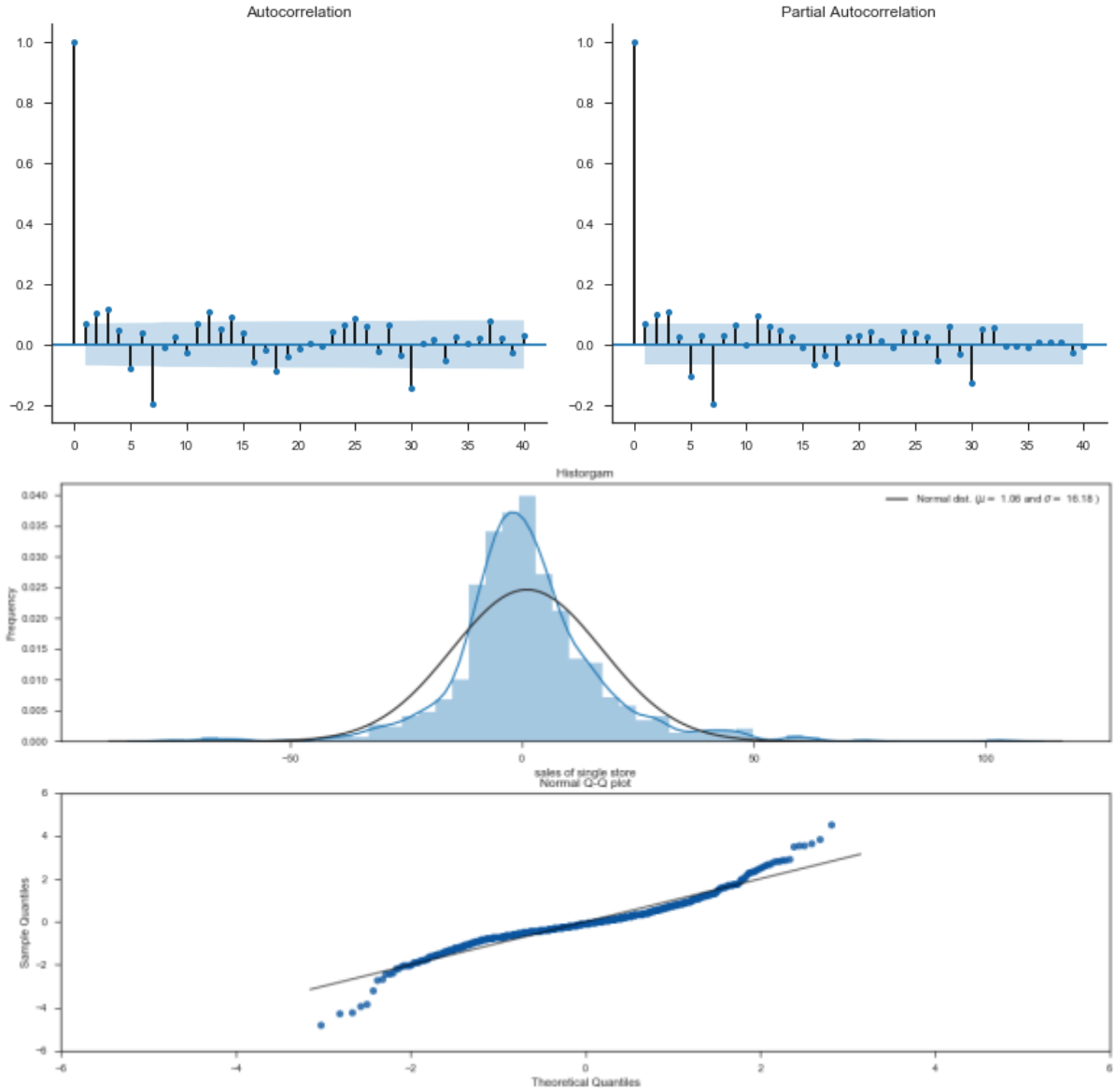




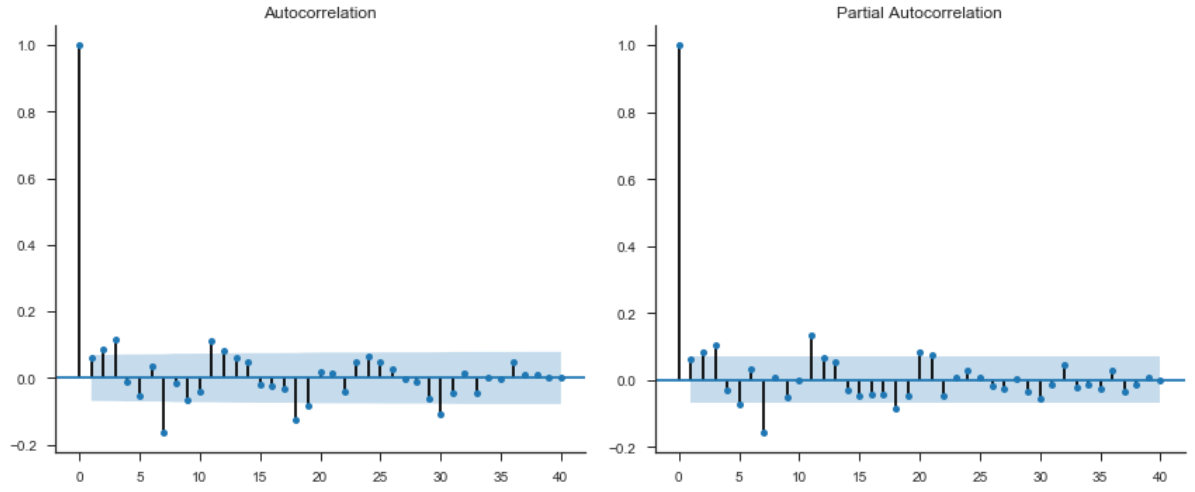


Store 223 (808 records)

Bset model is ARIMA(1,0,1)X(1,1,1,6) with AIC: 6451.052



Store 21(808 records): Bset model is ARIMA(1,0,1)X(1,1,1,6) with AIC: 6615.197




Rossmann Sales forecast
Sales
14000
12000
10000
8000
6000
4000
2000
Jun
2015
Jul
Aug
Sep