# Time Series Analysis and Forecast of Rossman Drug Stores

## Business Context and Problem Formulation

Rossman are a chain of drug stores headquartered in Germany, and operates over 3,000 stores. Like many businesses, it is advantageous to use statistical tools to forecast sales in order to inform decisions including, but not limited to, marketing strategies, managing supply to meet demand in a cost-effective way, and ensuring price stability. Overall, being able to forecast sales will allow a business to run more efficiently.

The goal of this report is to provide an in depth analysis of given time series data to create models to forecast six weeks of future daily sales for Rossman drug stores. This will be done through building univariate forecasting models from the provided data. This is a model in which the future values of the time series are assumed to be based only on the past values. In this case, our variable of interest is *Sales*. The models will be fit over the entire data set, and the performance will be measured by calculating the root mean squared error (RMSE) over a validation set, which will be taken as a subset of the entire data.

Rossman is spread over Germany and a few countries in Europe. Because of these location differences many of the stores differ on when they are open and how many records are available. So for purposes of this analysis, we focus on stores that are open all seven days of the week and have complete records for the entire dataset. The techniques explored will consist of exponential smoothing and autoregressive integrating moving average (ARIMA) models, to find a univariate model that best fits the data. By the end of the analysis, we will identify a single univariate forecasting model, as well as the six weeks of daily forecasts.

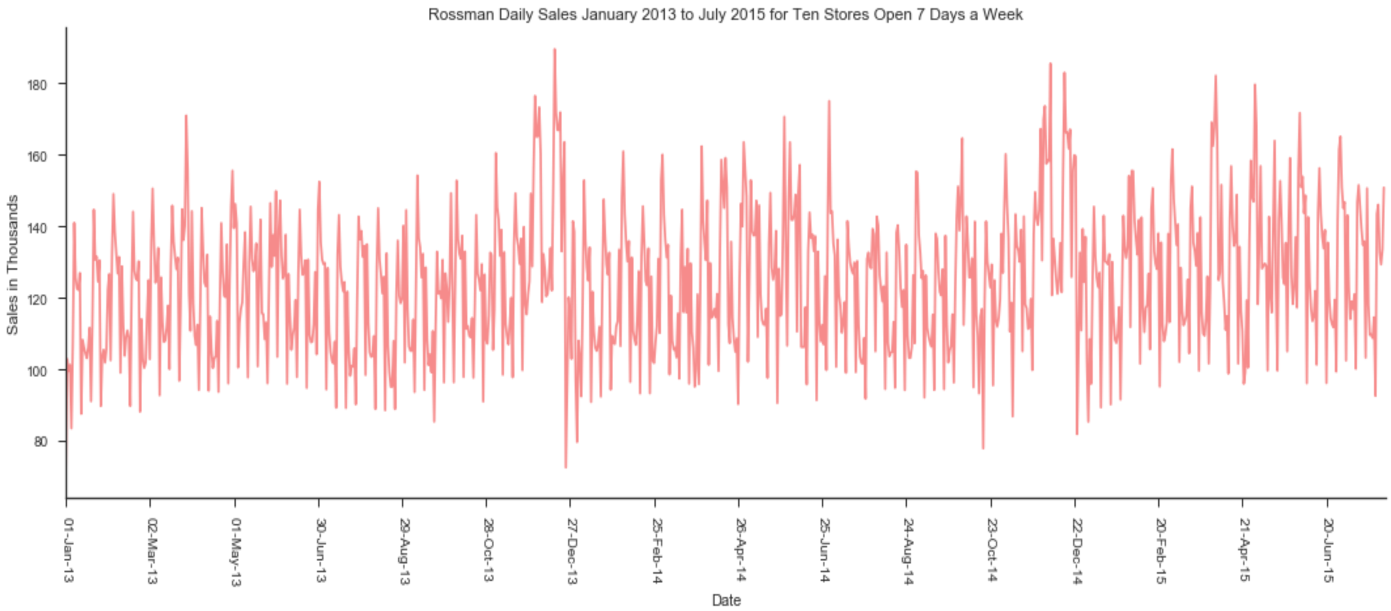
## Data Processing

The data range is from the beginning of January in 2013 to the end of July in 2015, which adds up to 31 months of data. In total, there are 1,017,209 rows of data and 1,115 unique stores, but not all stores have the same amount of records. Our first step is ensuring the data is as clean and error free as possible for the subset we are interested in. We filter our data to find there are ten stores that are open all seven days of the week, including Christmas, for the entire length of the data set. These store numbers are as follows: 85, 262, 335, 423, 494, 562, 682, 733, 769, 1097. We find that each of these ten stores have 942 non-zero data points for *Sales*, a point for every day in the data range, which adds up to 9,420 total points. The data processing is slightly simplified because there are no missing values. We then drop the *Store* variable and group the data by *Date,* being sure to sum up the *Sales* of each of the ten stores for every day. This brings us back to 942 data points, with each point being the aggregate of sales of the ten chosen stores on that date.

The sum of *Sales* is now a relatively large number, so we divide the values by a thousand to scale it down. We do this because it is generally better to work with smaller numbers for time series, as large numbers could lead to imprecision and accumulate numerical errors. We then convert *DayOfWeek* from a numerical to a categorical variable, and replace the numbers with the day names for EDA purposes.

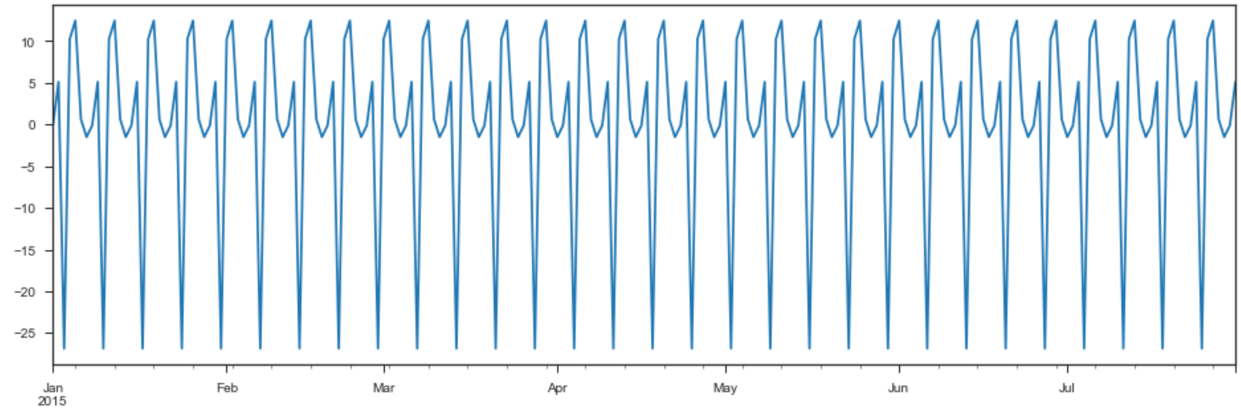
# Exploratory Data Analysis

We are only interesting in univariate models, so we focus our EDA on the *Sales* variable. Our first step is to plot the time series of the data over the entire range of the data set to see if there are any clear patterns such as trend or seasonality. As seen in Plot 1, there is no obvious trends, but there does appear to be some seasonality in the tight fluctuations. We also notice there to be a slight increase in sales during December due to the Christmas season. These two observations could be indicative of multiple seasonalities, as any store sales would generally be expected to fluctuate through the year. However, capturing this would require more complex models which we do not consider.



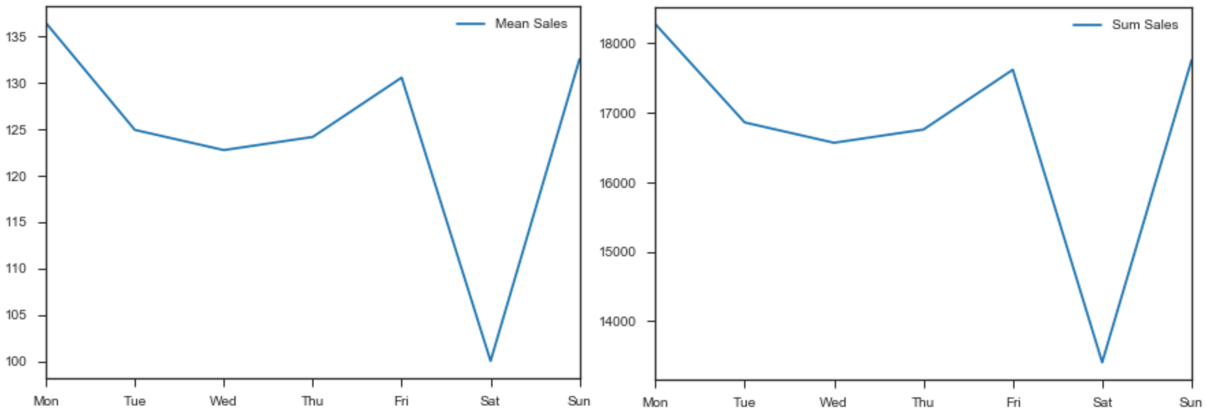
**Plot 1. Time series of *Sales* over entire time period for 10 chosen stores**

We decompose the time series into the trend, seasonality, and irregular components to do further analysis. There still does not appear to be any trend, but there is a clear seasonality component that goes through a full “season” every week. Hence, the seasonality is 7 days. For better visualization, Plot 2 shows the seasonality component for only 2015, as the pattern is the same for the previous years. There are about 4 seasonality cycles per month, which would support a weekly seasonality. Please refer to Appendix (i) for full decomposition.



**Plot 2. Seasonality component of *Sales* for 2015**

In Plot 2, we notice there are two peaks within a season, and this behavior does not change for the entire time period. We then consider Plots 3 and 4 which show the average sales and total sales by day over the entire period, to further explore and gain a sense of the general seasonality within a week. We notice that the weekday sales decline in the middle of the week and pick back up right before the weekend. Then there is a big drop on Saturday before it rises back up on Sunday. Ths behavior through the week and on Sunday is expected, but the drop in Saturday seems unusual since most people tend to go shopping on the weekend. This is likely due to Rossman having shorter store hours on Saturdays. Regardless of the reason, the pattern is consistent.



**Plots 3 and 4. Mean of *Sales* and sum of *Sales* by day of the week**

For the models we are going to use, it is important to make sure the variance in *Sales* is stable throughout the time period. This seems to be the case as seen in Plot 1, but as a simple check, we calculate the variance of the first half and second half of the data. They are 355.517 and 394.982 respectively. This relatively small difference and visual analysis of Plot 1 supports leaving the time series untransformed, as it will not significantly affect the model. It follows that we will likely use an additive model specification, which is relevant for exponential smoothing. Table 1 summarizes the distribution of the data.

| Mean | Std. Dev. | Min. | 25% | 50% | 75% | Max |
|---|---|---|---|---|---|---|
| 124.463 | 19.561 | 70.278 | 109.323 | 124.920 | 137.102 | 189.543 |

**Table 1. Key metrics of the distribution of *Sales***

For this analysis, outliers are considered any data point that is more than three standard deviations from the mean. This translates to a range between 65.780 and 183.146. However, since there are only 2 points outside this range with the max being 189.543, we will leave them. What we are more interesting in seeing are the *Sales* on Christmas Eve and Christmas Day. Since the chosen stores are open every day, including Christmas, we expect the behavior to be different on these days than for the rest of the month. Looking at these two days for 2013 and 2014 there is a large drop in *Sales* compared to the rest of the month, as show in Appendix (ii).

While these drops don’t seem significant, because they occur during a month where the rest of the days have relatively high sales, they will affect the fitting of the model and cause the residuals to be very large at these points. This will in turn affect the residual diagnostics. We replace the 24th and 25th of December in both years with the one week lagged values on the 17th and 18th respectively to fix this.

The EDA conducted looks at both models generally, but we will need to consider additional analysis with differencing and autocorrelations, which will be addressed in the Modelling phase.

# Modelling

## Benchmark Model - Seasonal Random Walk

We start off by defining a benchmark model. For this analysis, our benchmark will be a seasonal random walk. The seasonal random walk is a simple model where an observation at time t is equal to the observation at the previous seasonal lag. In this case, the seasonal lag is 7, so the fitted value of $y_t$is simply the value at $y_{t-7}$. For forecasting future values, the seasonal random walk simply sets all the forecasts equal to the value of the last observation in the data, which in this case is 150.821. Seasonal random walks can be useful in capturing the seasonality in a time series, but will not be very effective in identifying any other patterns or changes. We will use this model for comparison when evaluating our more sophisticated models.

## Exponential Smoothing

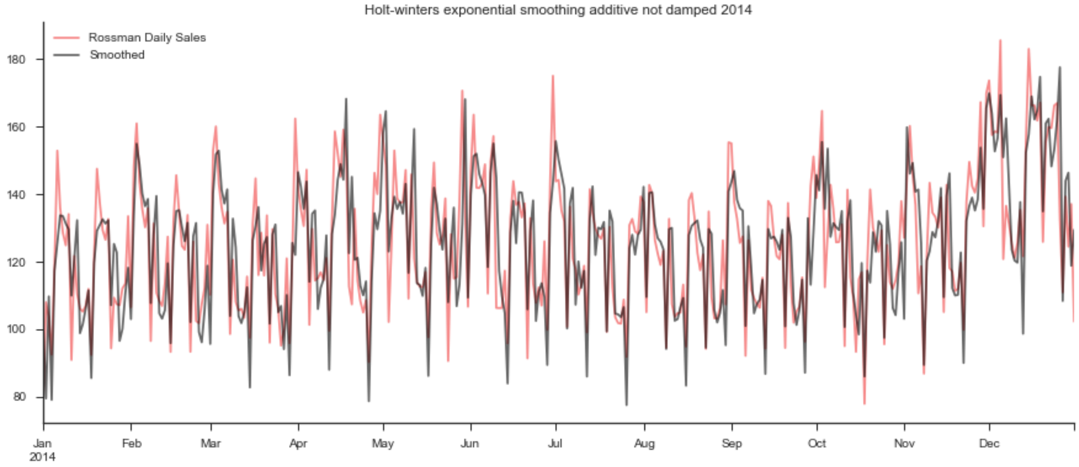
Exponential smoothing models work in an intuitive way. Forecasts for exponential smoothing models are a weighted average of past observations where the weights exponentially decay the further back the observation is. This means more recent observations have more weight in forecasting. A few well known methods are simple exponential smoothing, trend corrected and

Holt-Winters seasonal method. Holt-Winters is the only one that takes seasonality into account by considering the weighted averages of the values at the seasonal lags. Because we identified a weekly seasonality as well as an additive specification during the EDA process, we will first fit additive models with a seven day lag with and without damping. Additionally, we try multiplicative models to see how they compare. A summary of the models for comparison is in Table 2.

| Model | Additive No Damp | Additive Damping | Mult. No Damp | Mult. Damping |
|---|---|---|---|---|
| In-sample RMSE | 12.406 | 12.397 | 12.463 | 12.469 |
| AIC | 7425.523 | 7426.122 | 7434.134 | 7437.145 |

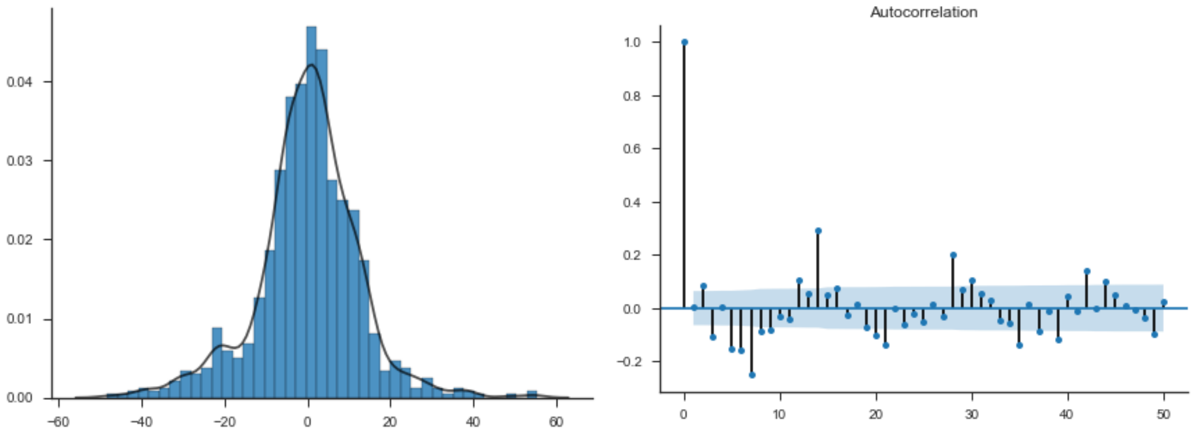
**Table 2. Holt-Winters short model summary**

We select the model based on the lowest Akaike Information Criteria (AIC) and choose the additive non-damped model. The AIC values are very similar, so the differences in the models will be small. For better visualization, we only show the smoothing for 2014 in Plot 5 (See Appendix (iii) for full time period). The smoothed line fits the actual data quite closely, but there are points where it greatly underestimates or overestimates the values. This is likely due to cases where the increase or decrease between an actual value and it’s value from the previous week is large. The previous week’s value carries more weight in the exponential smoothing causing the smoothed value of the current week to be drastically off.



**Plot 5. Holt-Winters additive exponential smoothing for 2014**

While never perfect in practice, normality of residuals is an important assumption check for time series, so residual diagnostics are included below for further analysis of the model. The distribution of the residuals in Plot 6 is slightly right skewed with a skewness of -.221 and an unusual amount of residuals around -20. This could be indicative of a larger than expected drop in daily sales between seasonal periods. The autocorrelation (ACF) of the residuals in Plot 7 contain a lot of significant spikes, especially around the season periods of multiples of 7, that are slowly decreasing. While this ACF plot is a vast improvement on simple exponential smoothing (Appendix (iv)), there is still some behaviour not being captured. These effects will be addressed with differencing and an ARIMA model.



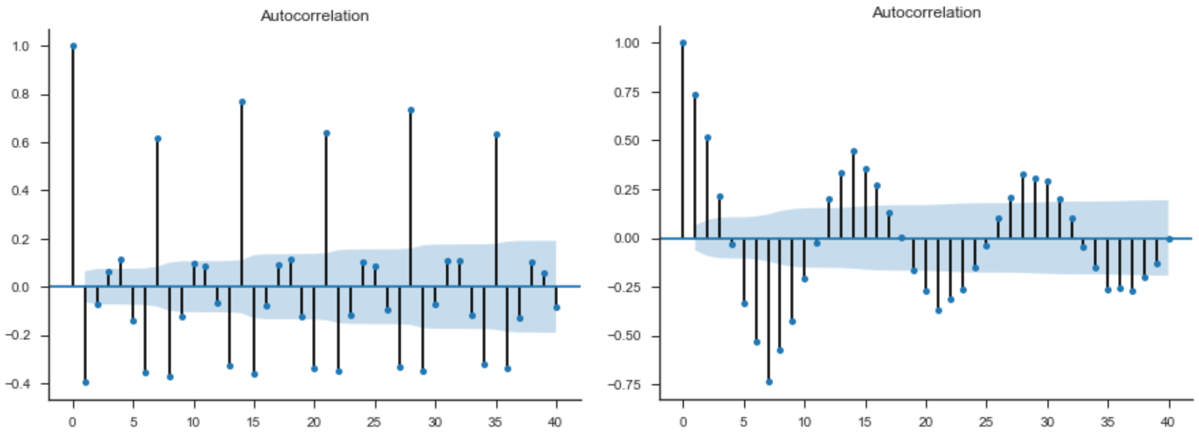
**Plots 6 and 7. Distribution and autocorrelations of Holt-Winters additive model**

Please see Appendix (v) for residual plot and QQ plot.

## ARIMA

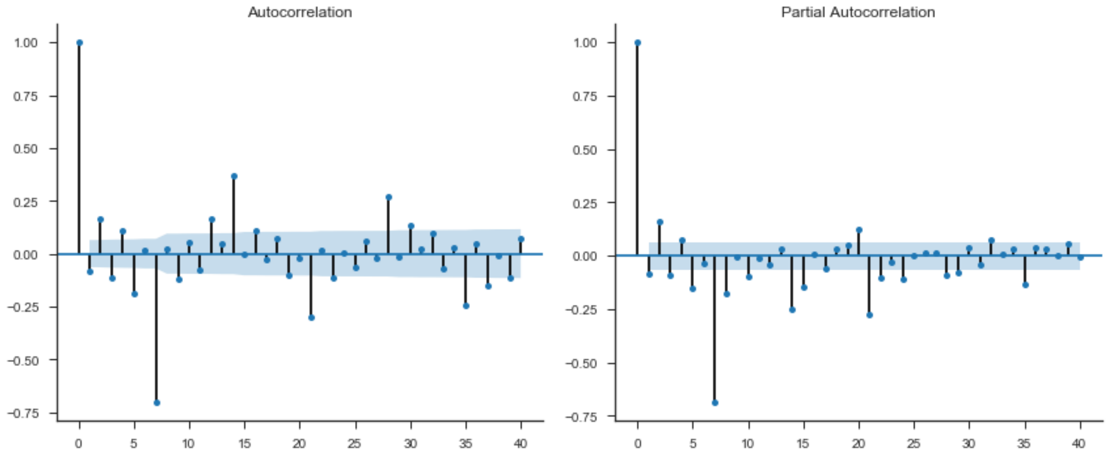
ARIMA models differ from exponential smoothing because they are models built off the autocorrelations within the data, rather than directly on the trend and seasonality. The first part of an ARIMA model is the autoregressive part in which the forecasts of a variable are a linear combination of past values of that variable. The second part, the moving average, is more abstract in that it uses a linear combination of past forecast errors. In this sense, a value at time t is thought of as a moving weighted average of past forecast errors. For this analysis, we focus on a seasonal ARIMA model with the specifications ARIMA(p,d,q)(P,D,Q,m). Here, p is the order of the non-seasonal autoregressive order AR(p). The non-seasonal moving average order is q, or MA(q). And d is the non-seasonal order of differencing. The capital letters are the seaonal counterparts, and m represents the season, so in this case 7.

The first step is to make sure the data is stationary. In other words, it does not depend on the time in which it was observed. Non-stationary data is made stationary by differencing. This is done by taking the difference between a value at time t and its value at an earlier time period, usually t-1 or t-m for first differencing and seasonal differencing. By definition, data with seasonality is not stationary, so we have to difference our sales data to make it fit for the ARIMA model. We take the first and seasonal difference of our data separately to determine what kind of differencing is necessary. Plots 8 and 9 show the ACF of the differenced data. The partial autocorrelations (PACF) and differenced data series are in Appendix (vi).



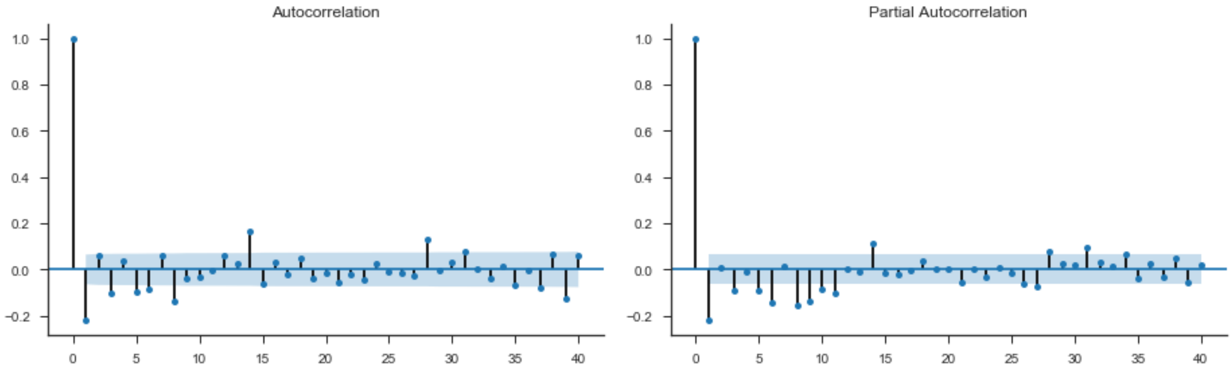
**Plots 8 and 9. ACF of first differenced (left) and seasonal differenced (right) data**

The spikes at the seasonal lags in the first difference and the sinusoidal pattern in the seasonal difference indicate a necessity for both first and seasonal differencing. Plots 10 and 11 show the ACF and PACF for the data when it is both first and seasonally differenced. Refer to Appendix (vii) for the data series. We see a slow decrease in the seasonal lags of both the ACF and PACF. This indicates both an autoregressive and a moving average model of specification AR(1) and MA(1) for the seasonal component of the ARIMA model.



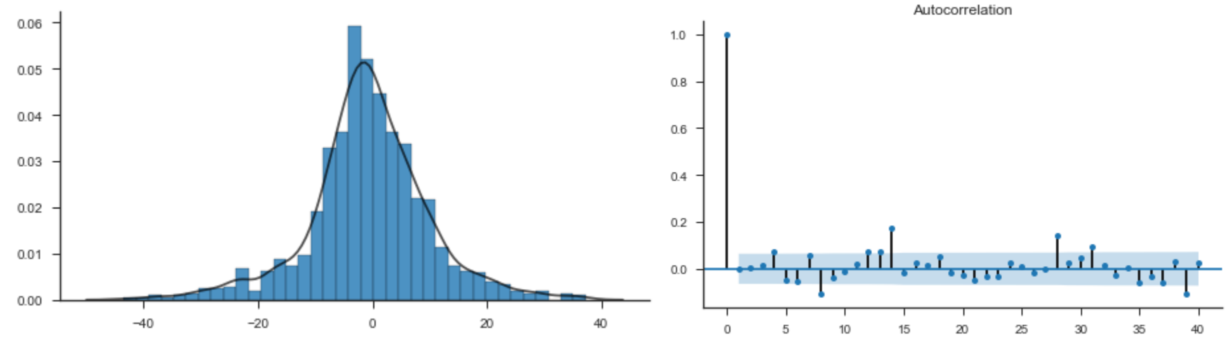
**Plots 10 and 11. ACF and PACF of data when applying both first and seasonal difference**

It is difficult to see the behavior of the non-seasonal component, so we fit an ARIMA(0,1,0)(1,1,1,7) model and plot the ACF and PACF of the residuals in Plot 12 and 13 to identify any other patterns. The plots looks better, but we still see a slow decrease in both the ACF and PACF, suggesting an AR(1) and MA(1) in the non-seasonal component as well. Because no clear patterns are visible, this method is not entirely reliable, so we try various other specifications of the non-seasonal part and select a model based on the AIC. We then do residual diagnostics to check the normality of the residuals, and try to identify any behavior we are not already capturing in our model. This is show in Plots 14 and 15.



**Plots 12 and 13. ACF and PACF of residuals of ARIMA(0,1,0)(1,1,1,7) model**

As mentioned above, we try various specifications of non-seasonal orders and select a model based on the AIC value. This is summarized in the Appendix (viii). The model chosen is an ARIMA(3,1,1)(1,1,1,7).



**Plots 14 and 15. Residual distribution and autocorrelations of ARIMA(3,1,1)(1,1,1,7)**

The distribution of the residuals looks fairly normal with a skew of -.145 and a kurtosis of 1.774. The ACF of the residuals (PACF in Appendix (ix)) are mostly non-significant and tend to zero aside from a couple significant spikes at some seasonality lags like 14 and 28. Additionally, the Ljung-Box p-value is less than 0.05, indicating the residuals are not entirely independent of one another. While this model does capture a good amount of the behaviour in the time series, this could be indicative of some additional factors that are not captured in the model. This is a limitation as it would require more complex modelling that is not considered in this analysis.

# Model Validation

As a way to further compare the sales forecasting performance of our two models, we will conduct a real time forecasting exercise for model validation. We select the validation set (out of sample data) as the last 7 months of data (roughly 23% of the data), beginning on the 1st of January 2015, but we will consider another validation size as it relates to Rossman store sales. The real time forecasting is done by fitting the models on a consistently updating in sample (training) data to forecast the next value in the validation set. What this means is in order to forecast $y_t$, the model is estimated on the available values up to t-1, then used to forecast $y_t$. For $y_{t+1}$, the model is re-estimated on the available values up to t, then used to forecast $y_{t+1}$. This is repeated for all the values in the validation data, then the RMSE is calculated to estimate and compare the performance of various models. The Jackknife resampling method is then used to estimate the standard errors.

We first do validation on the last 7 months of data. We choose this simply because it is the beginning of the year and the time series has gone through two full years. We identified a weekly seasonality, but in the context of store sales, it is not unreasonable to say there is some degree of yearly fluctuations. The RMSE results are in Table 3. The seasonal random walk model is included for reference.

| Model | RMSE (thousands) | SE |
|---|---|---|
| Seasonal Random Walk | 26.683 | 1.272 |
| HW Additive Non-Damped | 15.362 | .783 |
| ARIMA(3,1,1)(1,1,1,7) | 11.042 | .774 |

**Table 3. Validation results for last 7 months**

As seen in the EDA, sales are slightly increased at the end of the year due to the holidays. As such, the second validation set we consider starts at the beginning of November (roughly 30% of the data) to see how the models performs during this time. The results are in Table 4.

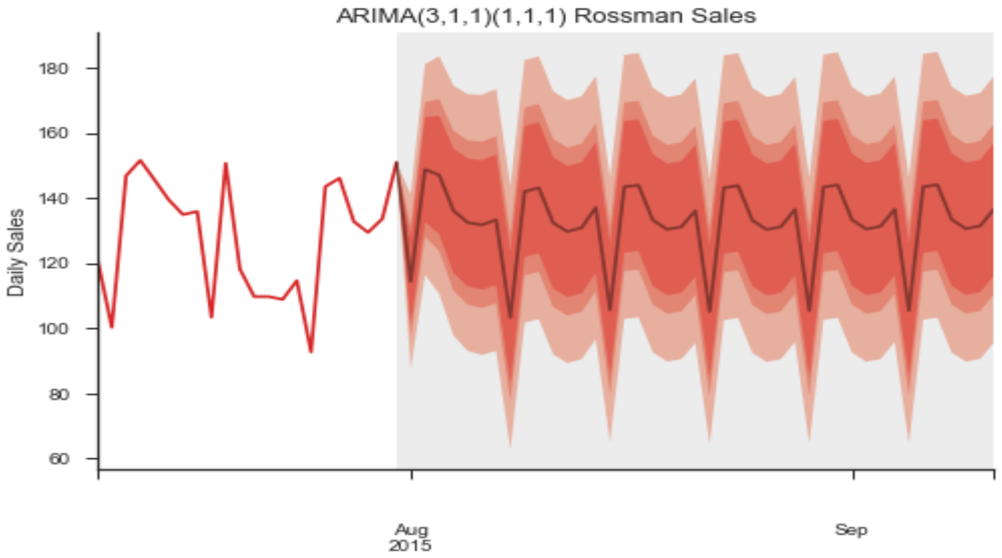
| Model | RMSE (thousands) | SE |
|---|---|---|
| Seasonal Random Walk | 26.968 | 1.063 |
| HW Additive Non-Damped | 15.696 | .722 |
| ARIMA(3,1,1)(1,1,1,7) | 11.653 | .700 |

**Table 4. Validation results from November 2014 onward**

The results of the two validation periods do not differ by much. The results in Table 3 are slightly better in RMSE, but have higher SE. This is a factor of the larger in-sample data and smaller validation data. The fact that the results in Table 4 are very close to the ones in Table 3 suggest that the models are adjusting for the end of the year and capturing that little bit of change in *Sales*. In both cases, ARIMA unsurprisingly performs the best and we select it for our final forecasts.

# Forecast

From the chosen ARIMA(3,1,1)(1,1,1,7) model, Plot 16 shows the fan chart of forecasts for 6 weeks of daily sales. The forecast horizon is therefore 42 days. The actual point estimates and intervals are given in Appendix (x).

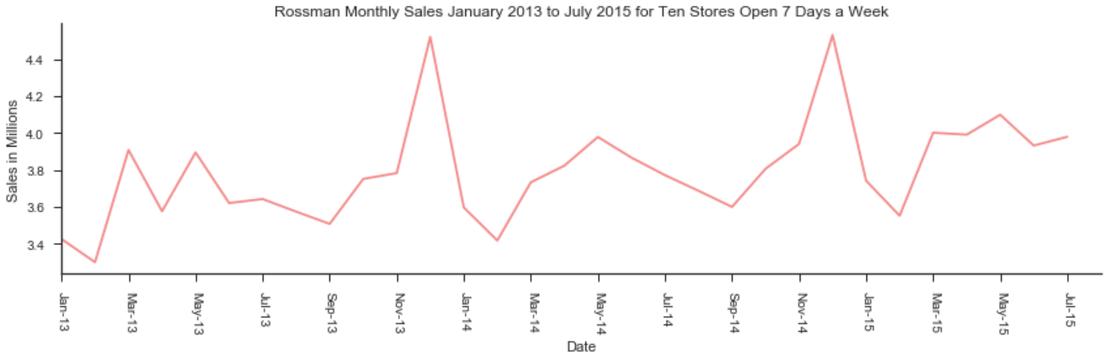


**Plot 16. Six week forecast for daily sales**

The forecasts are more or less as expected. The data follows a pretty similar pattern where there are two “peaks” in a week, for the entire data set. There is no reason to believe this pattern would suddenly change in the six week forecast. Additionally, these forecasts are for August and September where the daily sales in 2013 and 2014 hovered around the average as seen in the EDA. This year is no different, with the forecasted sales following the same behaviour as previous years. The intervals do not get significantly wider for forecasts further in the future, so the variance is not changing. Rossman can therefore expect their daily sales to remain stable and would not need to adjust any of their operational strategies based on sales. For reference, the intervals are at 80%, 90% and 99%.

# Monthly Analysis

It is of interest to conduct a short monthly analysis as it has been shown there is some seasonality through the year, especially the increase in *Sales* in December. A more complex model is necessary to capture both the weekly and yearly seasonality, so for this analysis we model them separately. We start with a time series of the aggregate sales per month in Plot 17. Note the sales have been scaled down to be in millions.



**Plot 17. Total monthly sales for entire time period**

The seasonality in the year is more clear. There is an increase in sales sometime at the beginning in of Summer, then a decline before spiking up in November and December. Doing a decomposition also reveals there is a very slight upward trend between July 2013 and January 2015. Please see Appendix (xi) for decomposition. We can see the seasonality is 12, so we fit a seasonal random walk as before for a benchmark, and try various Holt-Winters models. A table of the models by AIC is given in Appendix (xii). As with the daily analysis, the best Holt-Winters model based on the AIC is the additive non-damped model.

For ARIMA, nearly all of the autocorrelations in the data fall within the non-significant range for the non-differenced data. While there is a slightly larger spike at lag 12 for the ACF, the PACF doesn't show any discernible seasonal pattern. When taking the first and seasonal differences, the ACF and PACF appear to behave more or less like white noise. We know there is some degree of seasonality by looking at Plot 17, but it does not appear to be very strong. Nevertheless, we fit some ARIMA models to the data and select a (1,1,1)(1,1,0,12) model based on AIC. Table 5 shows the results based on a validation set from March 2015. Note we use this as the date so the ARIMA model can be properly implemented.

| Model | RMSE (millions) | SE |
|---|---|---|
| Seasonal Random Walk | .180 | .034 |
| Additive Holt-Winters | .124 | .045 |
| ARIMA(1,1,1)(1,1,0,12) | .222 | .134 |

**Table 5. Validation of monthly models**

Not unexpectedly, the seasonal random walk actually performs rather well because there is not enough data for there to really be any distinct patterns outside of the spike in December, so the simpler models are able to capture this. Ultimately, additive Holt-Winters performs the best for the monthly analysis. ARIMA breaks down since it is based off modelling the autocorrelations, which were not very strong for the monthly data. Table 6 contains the point forecasts and intervals for 4 months of the Holt-Winters model.

| Month | Point Forecast | 80% LB | 80% UB | 90% LB | 90%UB | 99%LB | 99%UB |
|---|---|---|---|---|---|---|---|
| Aug-15 | 3815938 | 3665754 | 3966122 | 3623179 | 4008697 | 3514078 | 4117797 |
| Sep-15 | 3761854 | 3546456 | 3977252 | 3485394 | 4038314 | 3328920 | 4194789 |
| Oct-15 | 4017902 | 3750397 | 4285406 | 3674563 | 4361240 | 3480236 | 4555567 |
| Nov-15 | 4062059 | 3748885 | 4375234 | 3660104 | 4464014 | 3432601 | 4691518 |

**Table 6. Point forecasts and interval forecasts for monthly Holt-Winters additive**

The results are again not unexpected, with sales remaining rather constant and then increasing toward the end of the year as the holiday season approaches.

# Final Remarks

This analysis could of course be extended to forecast the daily sales at any number of the Rossman stores as long as the store hours and openings are accounted for, given that many of the stores seems to follow the same patterns. As seen in the daily sales analysis, the ARIMA model is superior to the other methods used. However, it did break down in the monthly analysis when the sample size was very small and there were less clear patterns in the autocorrelation. These models are of course limited in that they are univariate, while the data contains other variables such as school holidays, state holidays, and promotion information. Additionally, only the seasonality in the weeks was captured, even though there was indications of seasonality in the months. If these factors were incorporated, these is no doubt the forecasting would improve, but these types of models are far more complex and were not used for this analysis.

In regards to Rossman, sales are forecasted to remain on track, so there is not a necessity for any changes to improve efficiency. If more complex models were used, however, further conclusions could be drawn about the effectiveness of promotions or holidays.