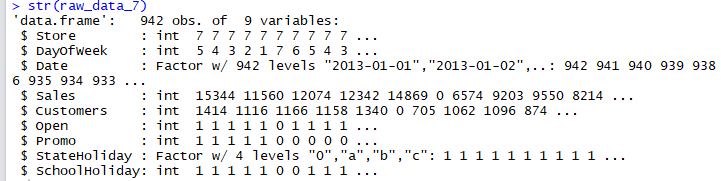
**Data Cleansing:**

**Data Extraction:**

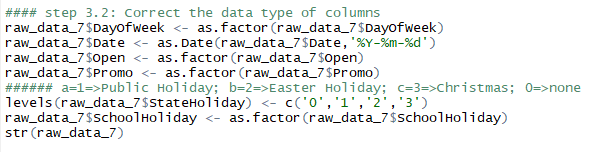
Given dataset has sales of 1115 stores. As a first step in our goal, one store must be selected for forecasting the sales for 10 weeks. In this case, store number 7 is selected. Data for store 7 is extracted from the whole file.

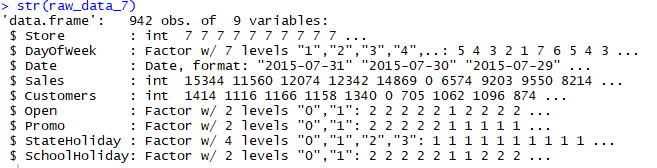




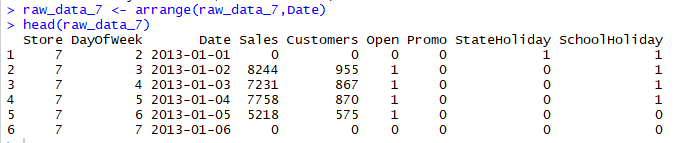
**Datatype Checking:**

Data imported from CSV file does not always have the appropriate datatypes for columns. Checking and changing the column datatypes is a vital step before proceeding with any data analysis. In this case, below chunk of R code is used to convert to datatype of columns as per requirement.

‘StateHoliday’ column has alphanumeric characters with ‘0’ representing ‘Not a State Holiday’, ‘a’ representing Public Holiday, ‘b’ and ‘c’ denoting Easter and Christmas holiday respectively. To make this column of same type, a, b and c are coded as 1 ,2 and 3 respectively and qualified as a factor.



**Sorting Data:**

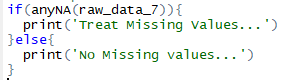
Since its time series data, data is sort as per date of sales.

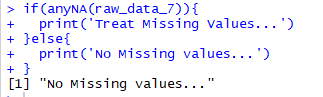
**Check for NA values**

NA Values and Outliers can potentially affect the coefficients determined by the model. Lot of NA values and Outliers make our model unreliable and produce biased results. Below are few ways by which NA values can be handled.

* Remove the rows that has NAs values exceeding the threshold, say 20% of values in a row.
* If the column is normally distributed, use mean of the column to replace the NA value
* If the column has high correlation with other column, conduct a regression between two to determine the possible missing values.

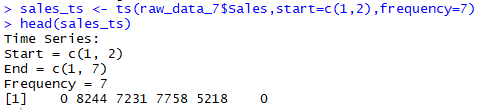
In this case, there is no NA/missing values in sales data.





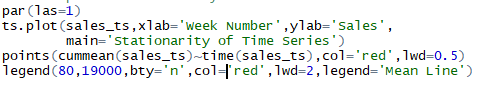
**Create Time series sales data**

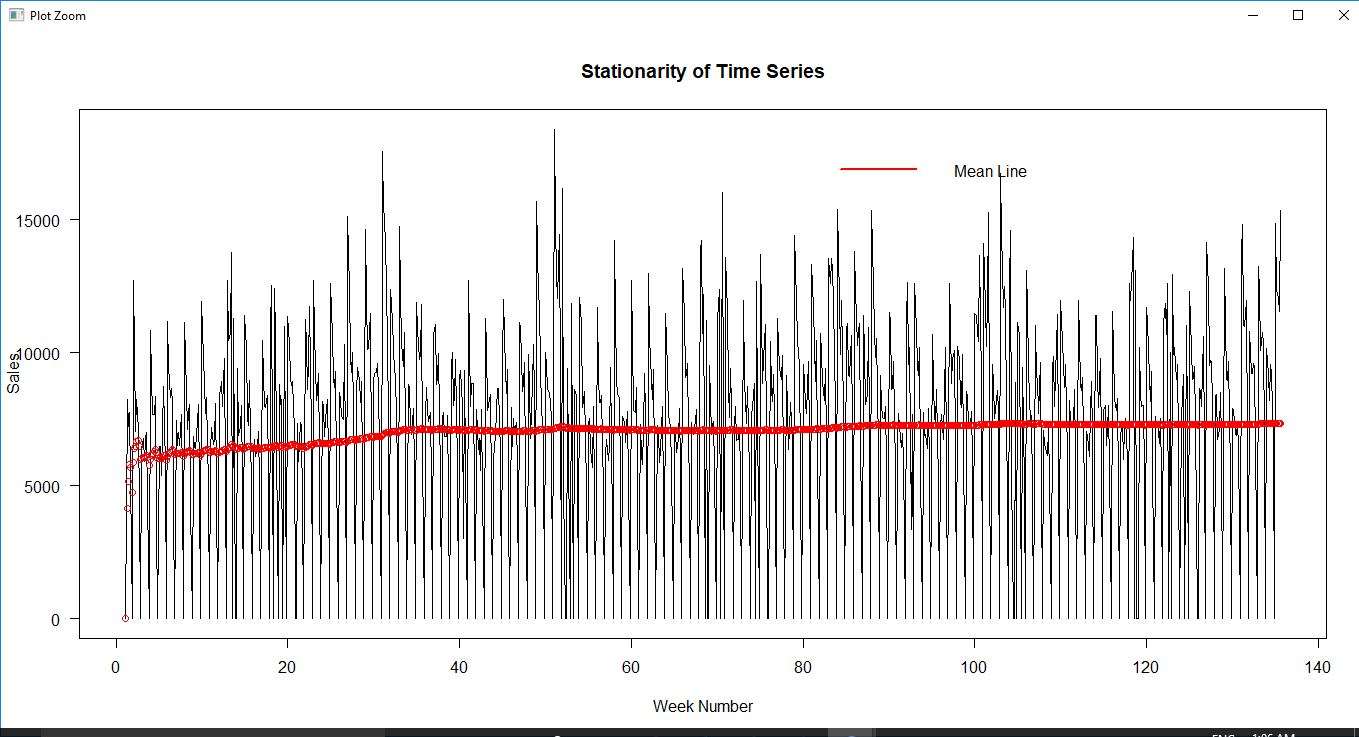
Sales data presented so far is a numeric data. Sales data is converted to time series data to conduct time series analysis. Since sales was recorded for each day, time series of sales is created with weekly seasonality and with frequency of 7 days a week.



**Stationarity in Time series:**

Times series should meet stationary property before applying any time series model. By Stationary property, time series should have its statistical properties such as mean, and variance should be constant over time.

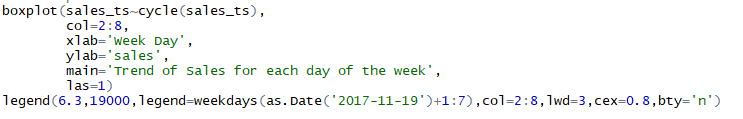


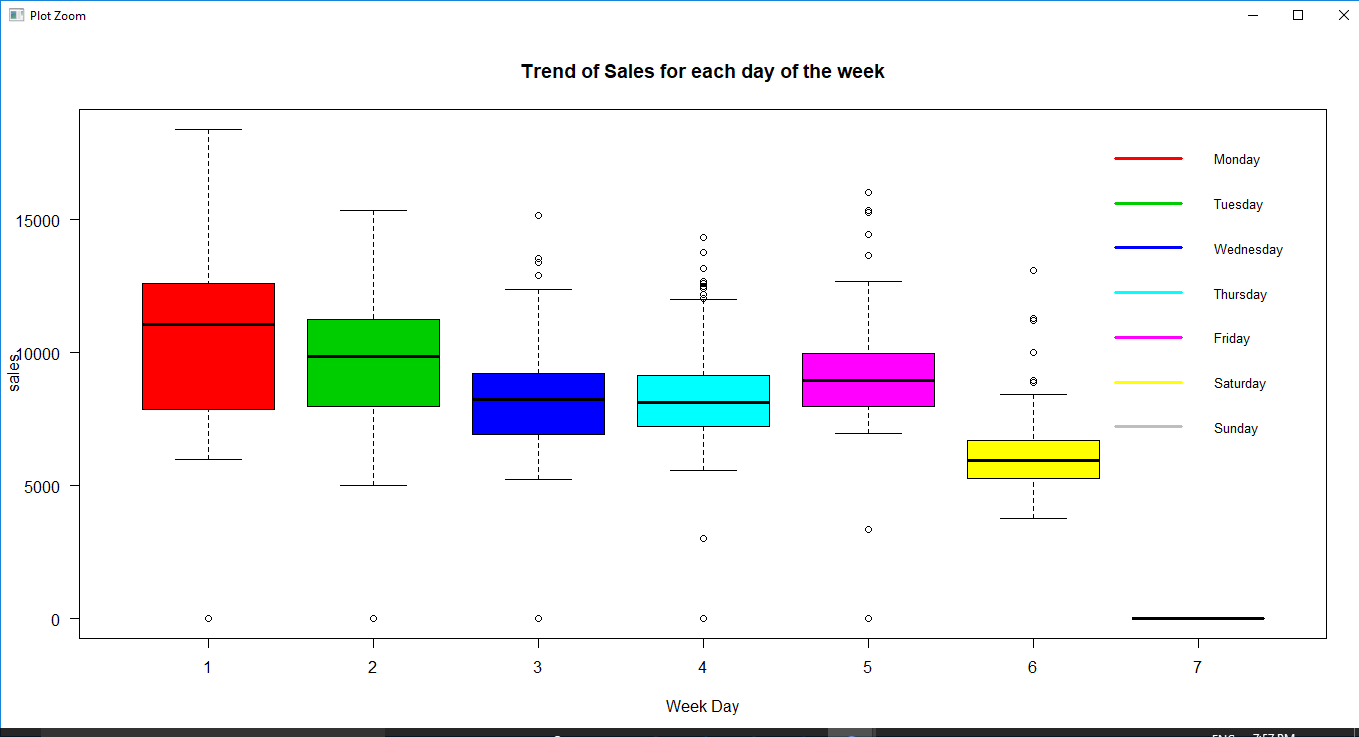


From the graph, it can be inferred that mean is almost constant over time with no considerable trend and variance of the sales also remain relatively same over the time. Hence the Times Series follows stationary property

**Check for Outliers**

Unlike NA values, outliers cannot be removed or replaced without exploring its effects on the data. Box plot of data is one of ideal ways to see possible outliers.

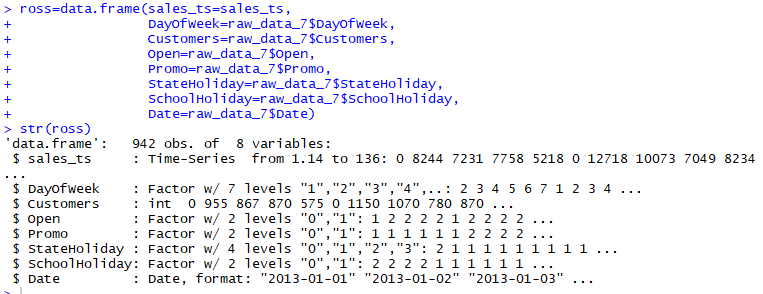




In this case, Eyeballing the data and boxplot indicates potential outliers in the data that falls outside the 1.5 times IQR (Interquartile Range) of sales. Here, Outliers are not so remote to exclude or handle it. It is presumed all the data points in the sales are vital. So, no sales values are removed as outliers.

**Create Data Frame with Sales Time Series and predictor variables**

Created time series sales data is combined with other potential predictors for conducting further analysis.

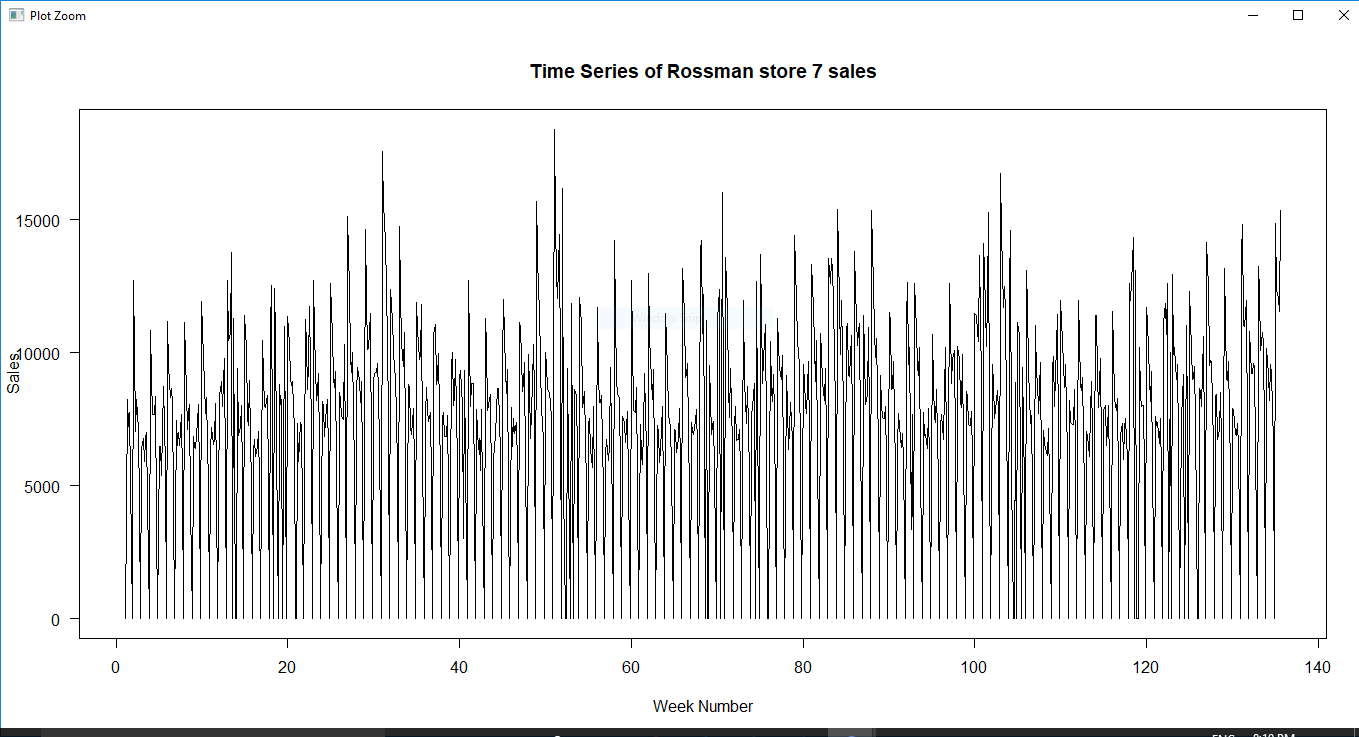


**Data Visualization**

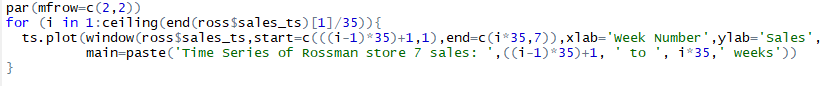
Visualizing the data gives clear indication of patterns and abnormalities of data. It helps to understand the trend of data and to decide on the approach to analyze further.

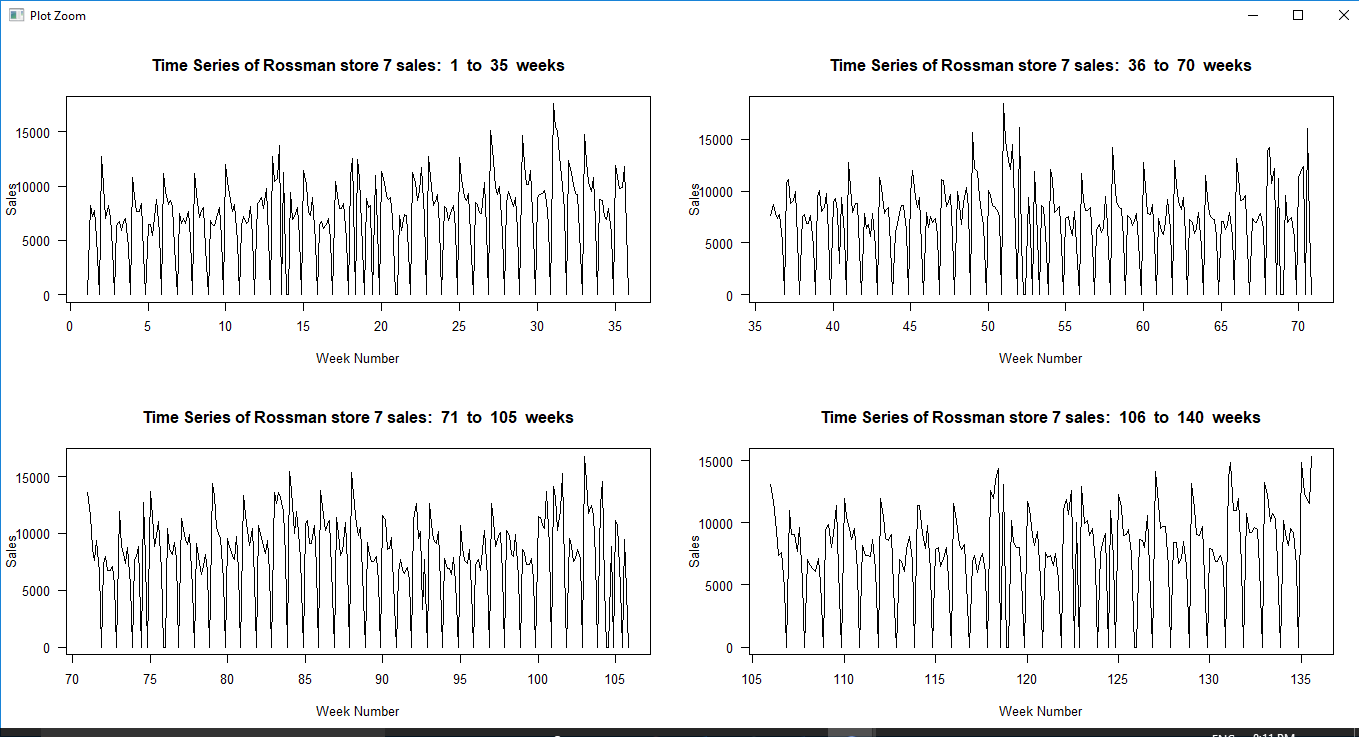
**Time Series plot:**





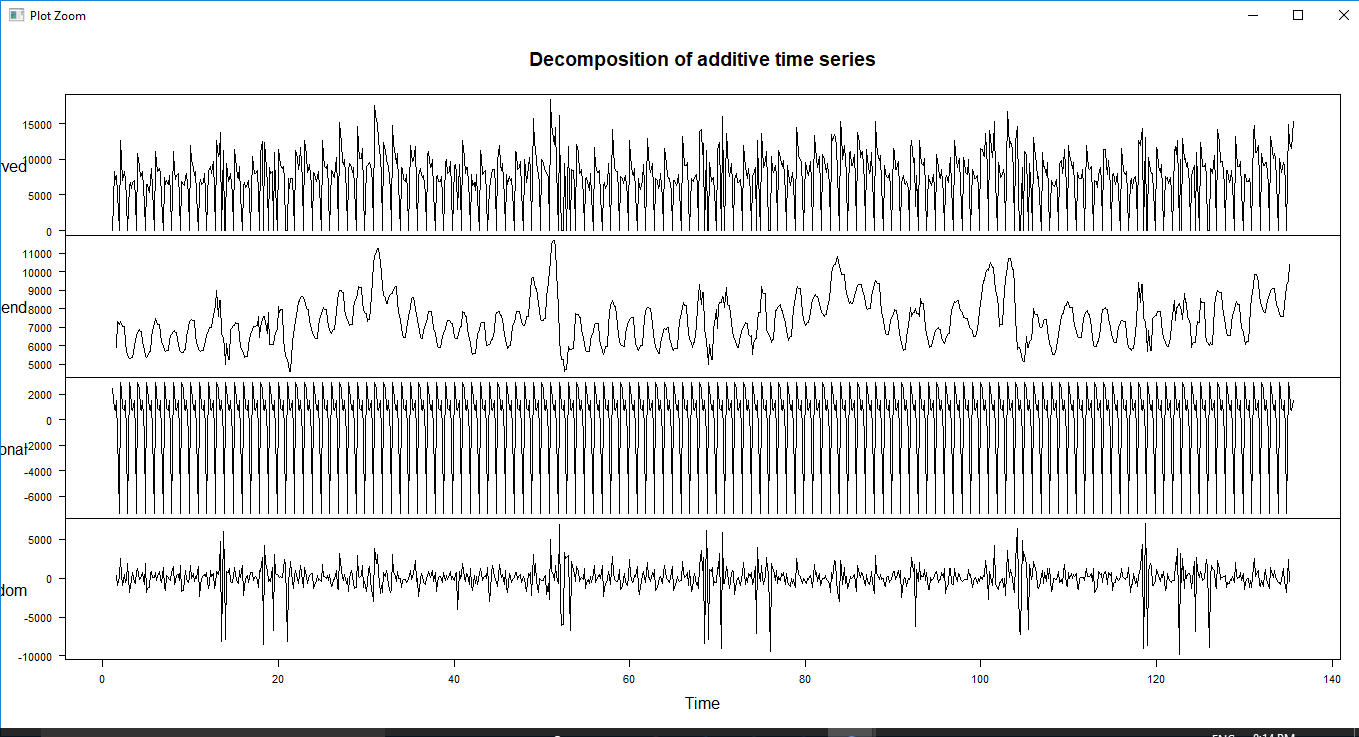
**Time Series plot (split by 35 weeks):**



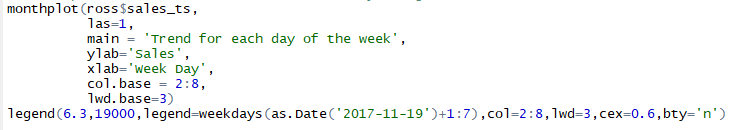


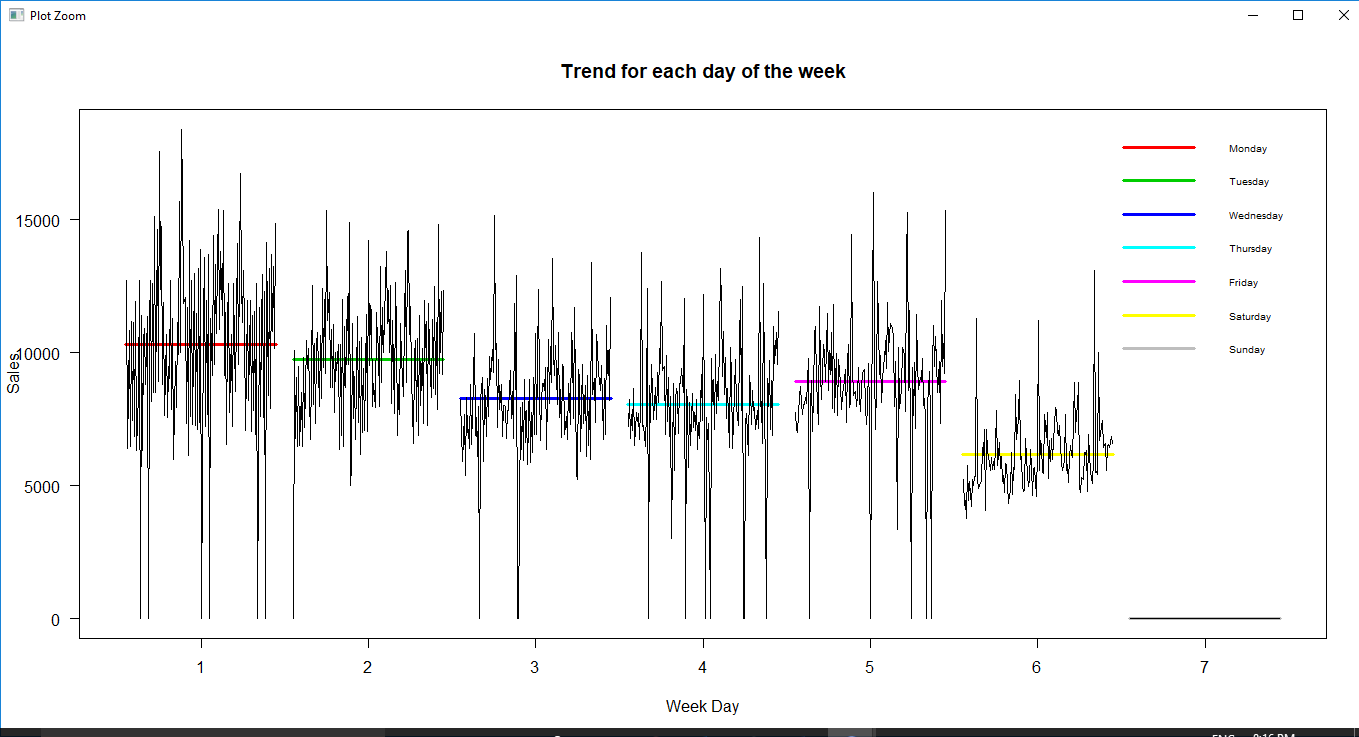
**Decomposition of Time Series Elements**





**Trend of Sales for each weekday using Month Plot**





From the above plots, it can be inferred,

* Average Sales is high on Monday and low on Saturday
* Sales has weekly seasonality pattern with minor trend

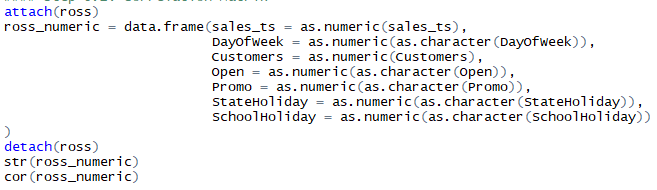
**Application of predictive techniques**

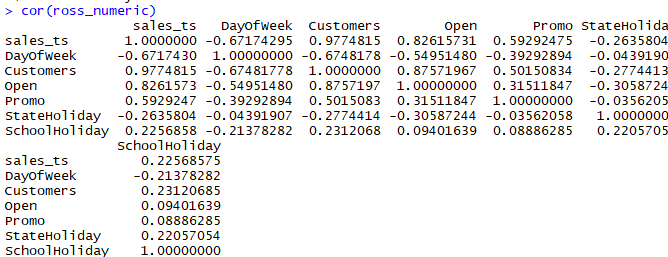
**Correlation between Predictors and response variable**

To identify the potential predictors for the model to be fitted, it is necessary to understand the correlation between all the predictor variables with the response variable (sales). Potential predictors are identified using two steps.

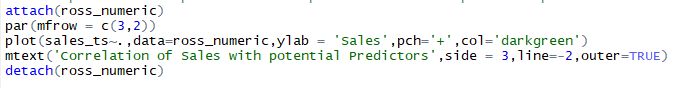
* Correlation Matrix
* Scatterplot

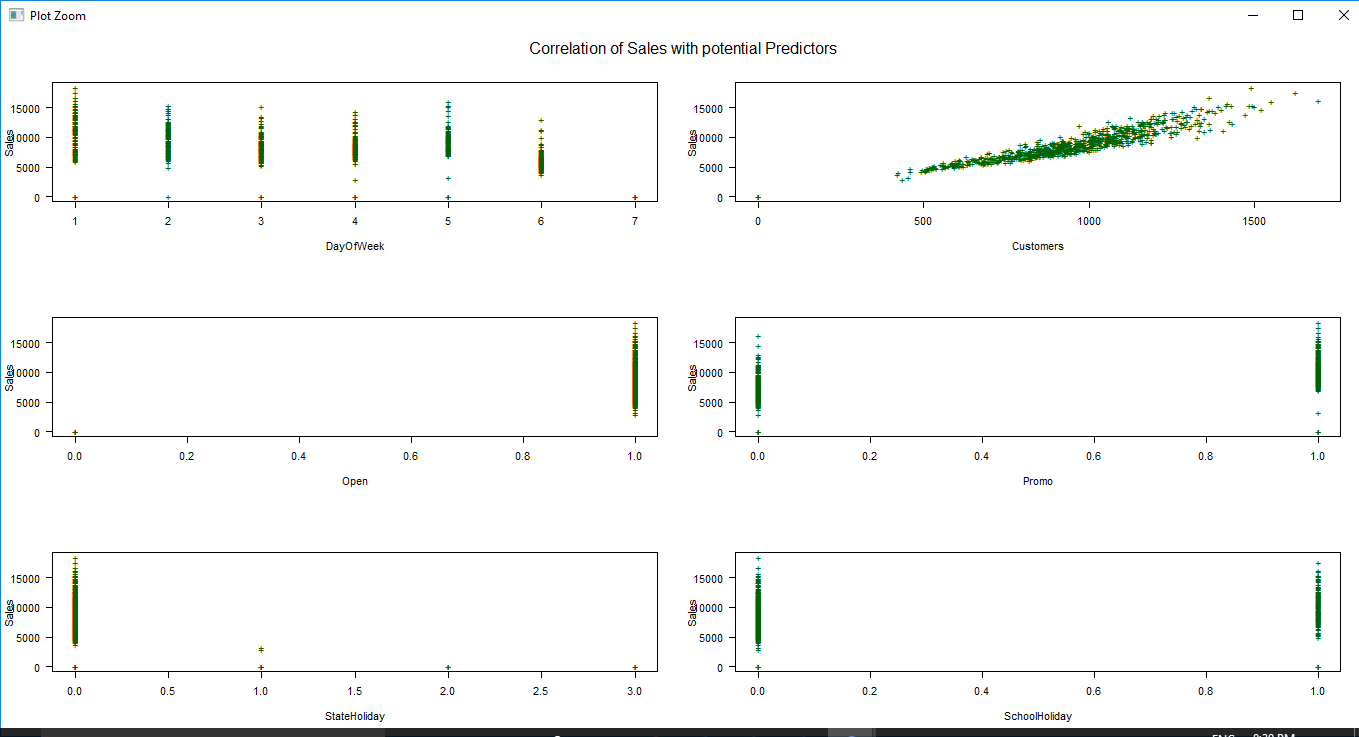
**Correlation Matrix**





**Scatter Plot:**

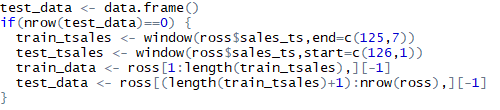




From the correlation matrix and scatter plot of the predictor variables and Sales, it is inferred that Sales is highly correlated to number of Customers, Week day, store functioning and promotional offers. It should also be noted that number of customers is strongly correlated with Week day and store functioning. It can lead to potential multicollinearity if all three data columns are used in the model.

**Splitting entire data to training and testing set**

Sales data is available for 135 weeks. In this case, 125 weeks are used for training the model and 10 weeks for testing the performance of the model.



**Model Fitting**

Time Series linear model is fitted for this data as the data involves both time series and regressors.

**Predictor Variables:**

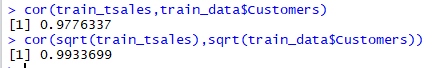
Below variables are selected as regressors for this model.

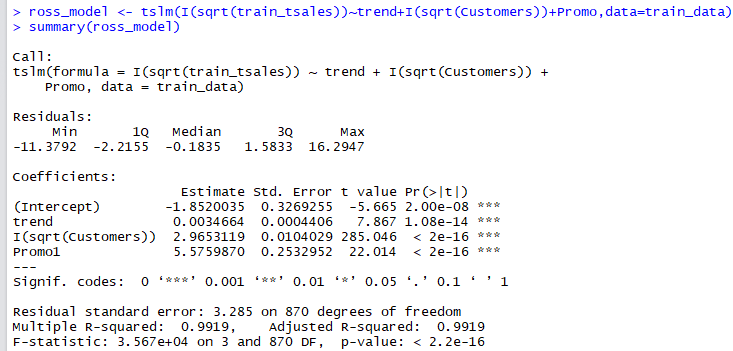
* Customer
* Promo

Note that Sales is highly correlated to ‘DayOfWeek’. ‘DayOfWeek’ is highly correlated to Customers. To avoid multicollinearity among 2 columns, only Customer is considered as potential regressor. Sales is weekly correlated to ‘StateHoliday’ and ‘SchoolHoliday’ which indicates these two terms don’t play vital role in determining sales for any given day. Hence, these 2 fields are ignored.

**Model 1: With Transformed Predictors and Response variable.**

To further improve the linearity between sales and number of customers, both the fields are transformed using square root operator. This improves the correlation between 2 further by closing packing the data values.





Considering 5% significance level, all the coefficients used in the model are significant and model is significant enough based on p-value.

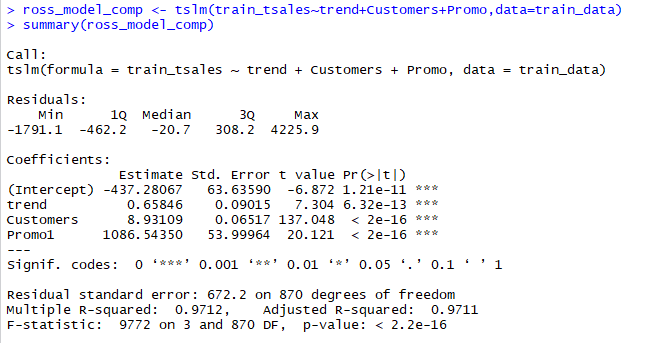
**Model 1 Coefficient understanding:**

Coefficients obtained here are with respect to transformed variables

* When there is no customers and no promotional offers, Sales value would be -1.852. It does not make sense in this context as ideally sales should be 0 with no customers. But it is important to keep the intercept in the model to ensure the proper orientation of regression line.
* Keeping promotional offers constant, if the number of customers increases by 1 unit, sales value would increase by 2.965 units
* Keeping the number of customers constant, if there are promotional offers announced for sales, then it would increase the sales by 5.576 units

**Model 2: With untransformed Predictors and Response variable.**

To understand the efficiency of model with transformed variable, one more model with same predictors but untransformed is created.

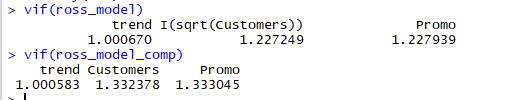


It should be noted that coefficients here are with respect to untransformed variable.

Considering 5% significance level, all the coefficients used in the model are significant and model is significant enough based on p-value.

**Multicollinearity check:**

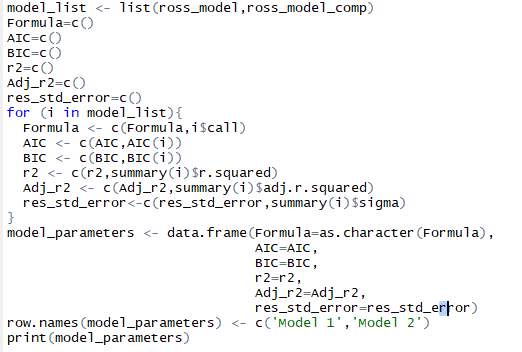
Variance Inflation Factor(VIF) is used to compute the multicollinearity among predictors in the models. VIF value of less than 2.5 indicates no or negligible multicollinearity.

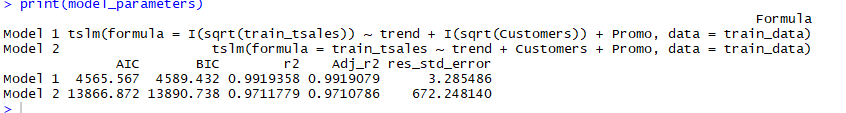


Based on VIF values for the predictors, there is no multicollinearity issue in both the models

**Goodness of Fit Evaluation**

Quality and efficiency of each model is determined by comparing AIC, BIC, R-squared, Adjusted R-squared and residual standard error values for the models.



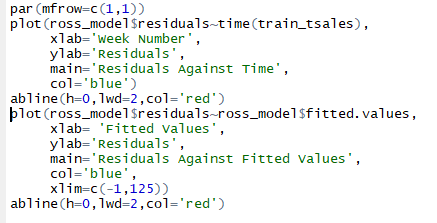


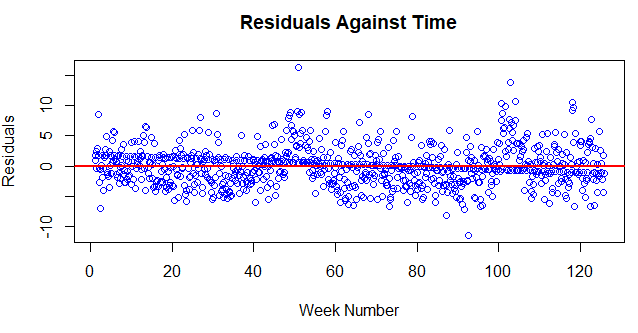
Comparing aforementioned parameters, Model 1 (with transformed variables) performs better in all sections and can be considered as the best fit for this dataset. Adjusted R-squared value for model 1 is 0.9919 which indicates model has greater explanatory power utilizing all the information from the predictors and lesser error terms.

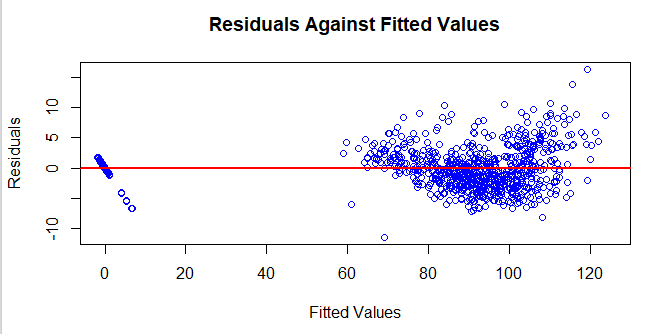
**Residual Diagnostics:**

For an efficient model, residuals generated by the model should represent the white noise characteristics. It should have 0 mean, constant variance, 0 correlation with fitted values or Time and be completely random with no visual patterns in its plot. It can be normally distributed as well.





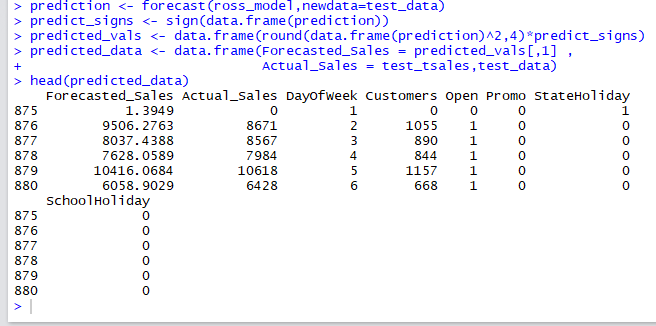




Model 1’s residual plots shows the white noise characteristics and it can be concluded that no significant information is left over in the residual that is left unused in the model.

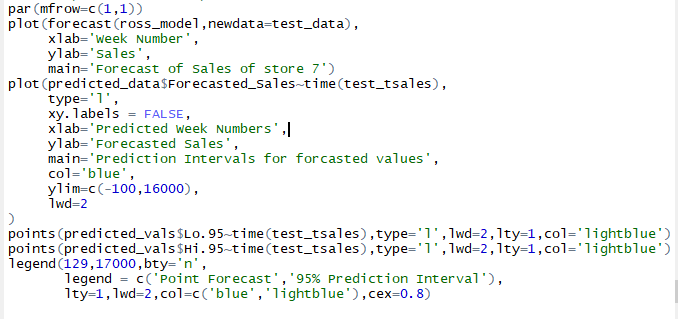
**Prediction of Sales with Model 1:**

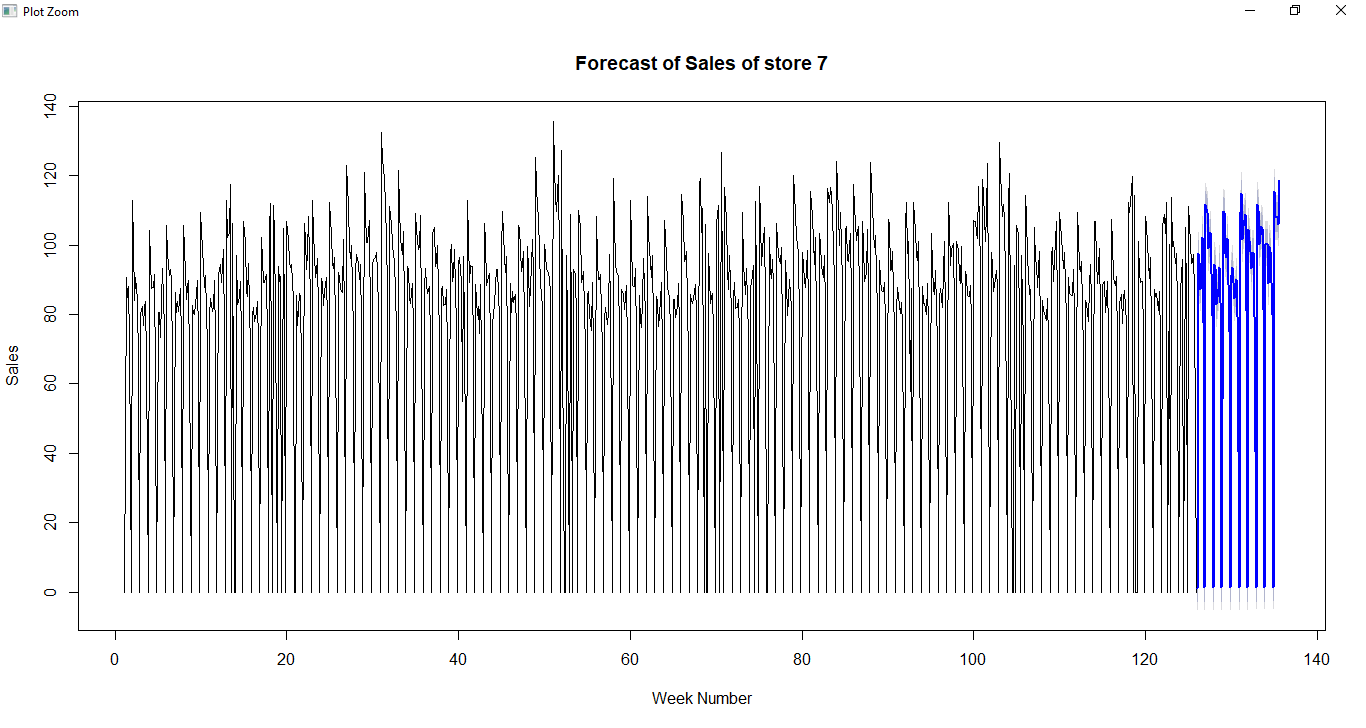
Using Sales data is forecasted for the test data which comprises of 10 weeks. Since response variable is transformed in the model, it is necessary to inverse the transformation to get the original value. In this case, square root is used as transformation. Hence after forecasting, predicted values is squared to get the desired sales forecast.

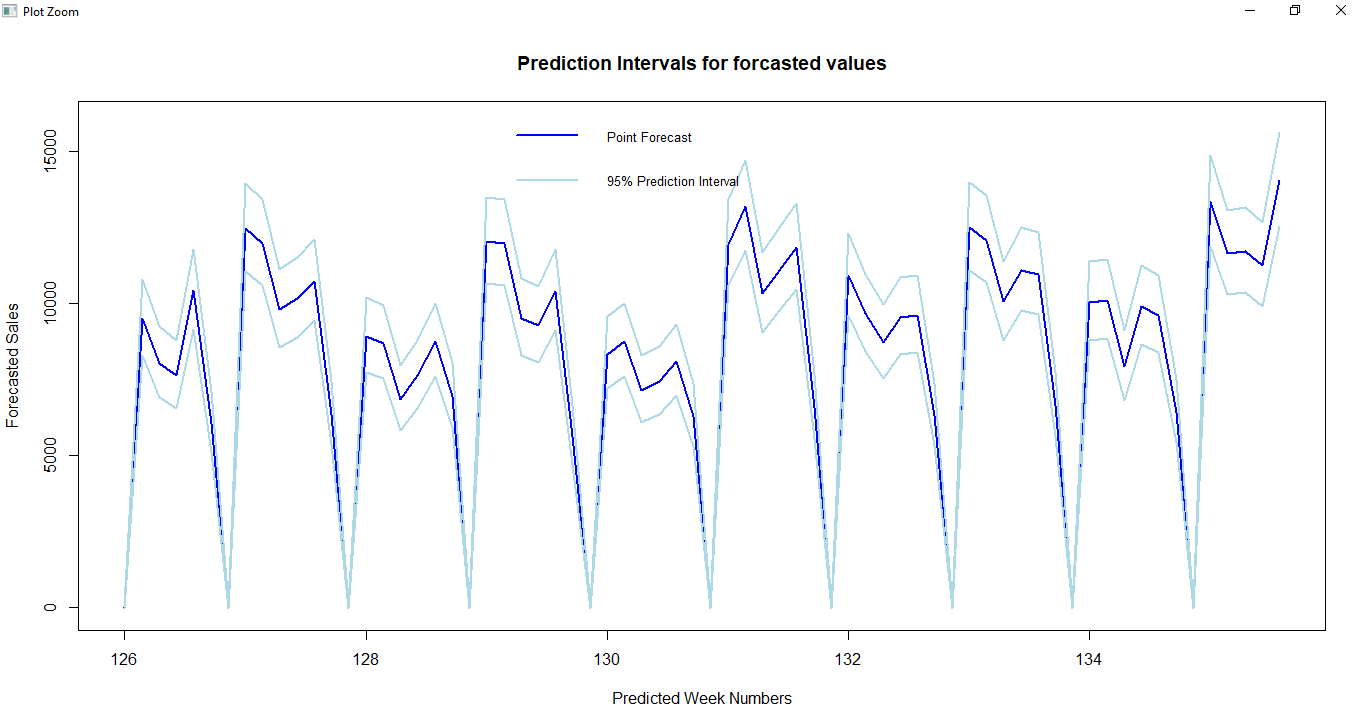


**Plot of Predicted values**

Forecasted sales and prediction intervals of sales are plotted again the time to visualize the forecasted sales trend.



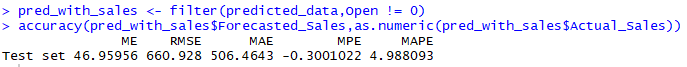




From the plots, it can be inferred that predicted sales closely follows with past sales values. Prediction interval for the sales is not much wider.

**Accuracy of Predicted Values**

Mean Error, Root Mean Square Error, Mean Absolute Error, Mean Percentage Error and Mean Absolute Percentage Error are calculated for the predicted values



To understand the magnitude of error, Root Mean Square Percentage Error is calculated.

Low value of RMSPE indicates the very less error in predicted values. Hence predicted values are aligned closely with actual values using Model 1 fit.

**Conclusion**

Time Series Linear regression model with transformed predictors and response variable gives better accuracy for this dataset. Hence this model can predict the sales of Rossman store 7 efficiently.