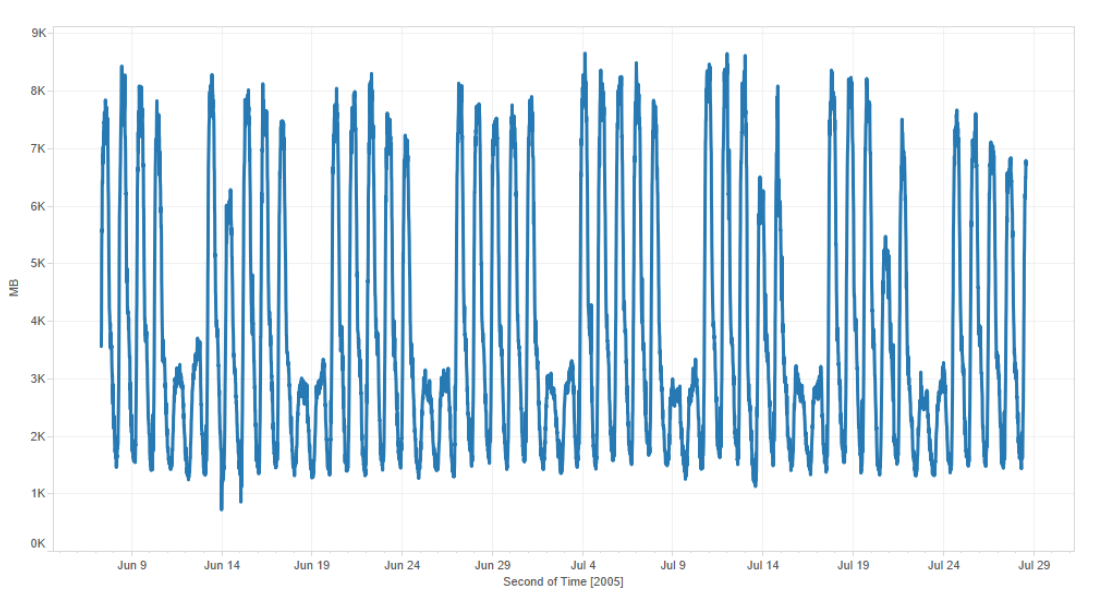
**Advances in Data Sciences and Architecture  
Internet Usage- Time Series Analysis- Report**



**Report Prepared by**

**Anirudha Deepak Bedre**

**Guruprasath**

**Balamanikandan Gopalakrishnan**

**Vijayakumar Jayaseelan**

**INDEX**

**TOPIC PAGE NO:**

* **INTRODUCTION 3**

**-**Time Series **3**

* **EXECUTIVE SUMMARY**

**-**Problem Statement **4**

**-**Approach **4**

**-**Methods Used **4**

**-**Insights Obtained **5**

**-**Recommendations **6**

* **ANALYSIS 7**

**-**Prediction Model **7**

**-**Trend Analysis **8**

**-**Seasonality Analysis **8**

**-**Initial Prediction Model **9**

**-**ARIMA Analysis **12**

**-**Improved Final Model (Full) **14**

**-**Training and Validation Forecast  **15**

**-**R Script Model in Microsoft Azure

* **CONCLUSION**
* **REFERENCES**

**INTRODUCTION:**

In this report we will look at the basics of time series data, study their behavior, look at their various characteristics in detail and practically apply the knowledge to a real-world dataset and provide results and recommendations. We begin by looking at the basics of time series data.

**TIME- SERIES DATA:**

A string of data points that have been collected across a common time interval is what constitutes a time series plot. A time series plot helps in understanding the behavior of a process as time goes on. Time series analysis and its benefits are spread across multiple areas of application such as statistics, signal processing, pattern recognition, econometrics, mathematical finance, weather forecasting, intelligent transport and trajectory forecasting, earthquake prediction, electroencephalography, control engineering, astronomy, communications engineering, and largely in any domain of applied science and engineering which involves temporal measurements.

Time series analysis comprises methods for analyzing time series data in order to extract meaningful statistics and other characteristics of the data. Time series forecasting is the use of a model to predict future values based on previously observed values. While regression analysis is often employed in such a way as to test theories that the current values of one or more independent time series affect the current value of another time series, this type of analysis of time series is not called "time series analysis", which focuses on comparing values of a single time series or multiple dependent time series at different points in time.

Time series analysis are used for:

* Forecasting in statistics, econometrics, quantitative finance, seismology, meteorology, and geophysics.
* Signal detection and estimation in signal processing, control engineering and communication engineering
* Clustering, classification, forecasting, outlier detection in data mining, pattern recognition and machine learning areas.

There are two major applications from time series data analysis:  
They are:

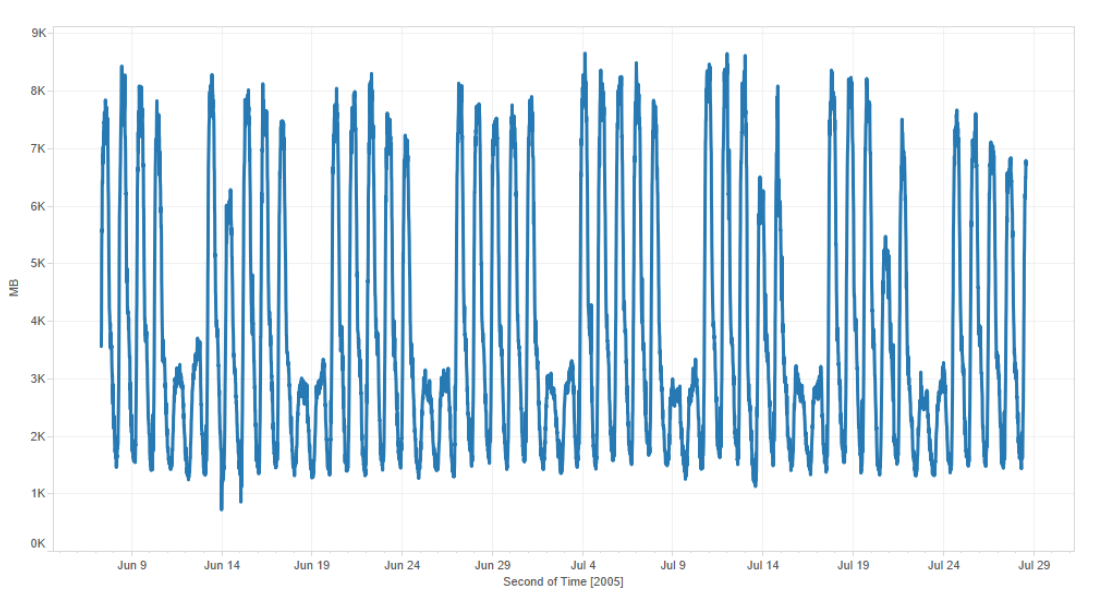
* Exploratory Analysis- To Understand behavior of process under study and learn
* Prediction and Forecasting- To look at the data and design a mathematical model to predict behavior of process for now and for the future.

Exploratory Analysis methods include:  
  
 - Autocorrelation analysis to examine serial dependence.  
 - Spectral analysis to examine cyclic behavior which need not be related to seasonality.  
 - Separation into components representing trend, seasonality.

**EXECUTIVE SUMMARY:**

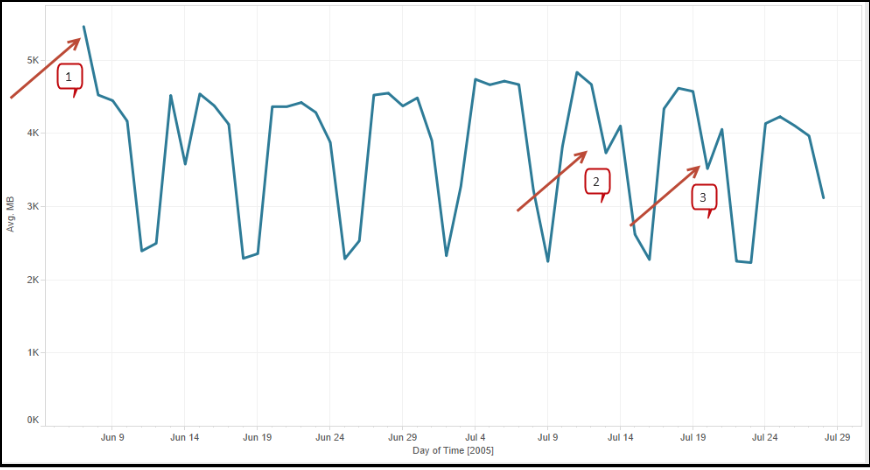
**PROBLEM STATEMENT:**The given dataset provides us the internet usage in MB (Megabytes) for 11 European countries using the same ISP provider observed for a time period of 2 months with data being recorded at every 5 minutes.  
Our approach aims to apply both the concepts of time series analysis i.e. exploratory analysis of the given data to learn and understand its behavior in detail and make recommendations to the ISP provider and also in building a prediction model that will help in predicting model behavior.

**APPROACH:**   
The data set was looked at first for irregularities in the provided data and an exploratory analysis was conducted to see if any transformation had to be applied to the data. The time parameter was specified in timestamp format and was converted into a quantitative unit of time (t=1, 2, 3…etc.). The data was given in number of bytes consumed and was converted into MB (Megabits) of data for easier approach to analysis.  
  
To better understand the behavior, the time series was then visualized as a two dimensional plot using the data visualization software tableau. The plot is shown below.



**METHODS USED:**As shown above, our methods begin with visualizing the given time series and looking at it to find any points from which we can learn any useful aspects about the data that can be later used for prediction or forecasting or an anomaly point which may have occurred due to an unusual event taking place which we have to ignore when we begin building our mathematical model for prediction. We used XLMiner to data transformation such as converting the time to quantitative units and conversion of bytes to Megabytes and also to do a dry run of the multi-linear regression model that will be improved upon at the end. From looking at the two dimensional plot that we framed above we were able to identify three unusual events that have affected the behavior of the time series plot at three points they are given below. These are some of the insights that we gathered looking at the plot and checking for corresponding major events in the participating countries for that day that could have had a significant effect on the internet usage.

**INSIGHTS OBTAINED:**



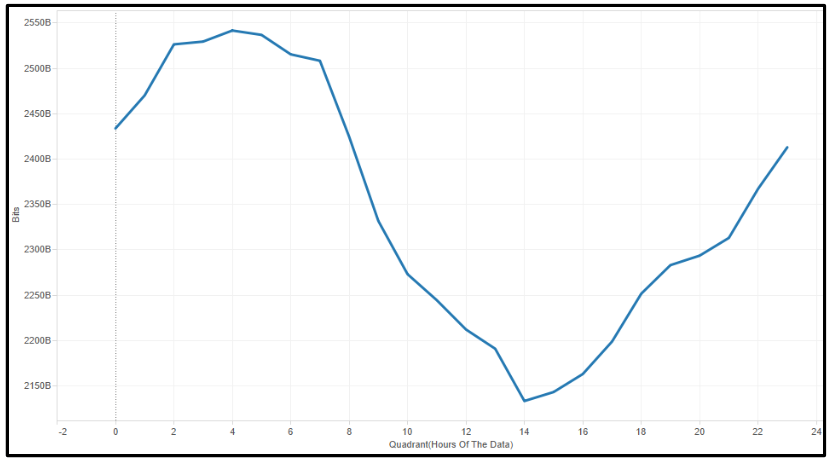
The first point indicated in the time series plot above indicated the point in time in which there were elections conducted in one of the participating countries which resulted in many subscribers using up internet data to catch up with the latest news and electoral analysis for the day. The second point indicated in the plot above is the point in time in which there was a dip on online activity due to the fact that one of the most biggest selling books, Harry potter and the half-blood prince was released which resulted in the biggest section of internet users (people below the age of 30) being occupied in purchasing and reading the new book rather than spend time on the internet. The third point in the time series plot indicates a drop in the internet usage due to the fact that there was a terrorist plot in one of the participating countries which resulted in a sheer drop in the internet usage for that day before going up the next day as people clambered online to know more information about the event.

In order to get our next insight into the data provided we aggregate the time aspect of our data to reflect the time component in the form of hours of the day rather than every 5 minutes of the day. The reason why we did this is we aimed to look at how the usage of internet is carried out in different hours of the day which could help in the following:

* Helps schedule daily maintenance on hours of low activity
* Helps the ISP provider market offers to customers to encourage usage during specific times of the day.
* Helps the ISP provider know how many servers need to be online in order to meet the huge demands for data at a particular time of the day

The time component was adjusted to hours of the day in the plot and the data consumed in megabytes was accumulated for each hour of the day.

The time series plot for this is shown below:



From the above plot, we were able to visualize the fact that the pattern of the internet usage of the participating countries was reflecting the method of data plans that existed during the time when ISP providers had Dial-up modems which could only have either the phone or the internet connection working at a given time and not both at the same time which is why we see internet usage predominant in early morning and late night. This also reflects the usage of internet by office-goers who checked their emails and messenger accounts every beginning and end of the day.

**RECOMMENDATIONS:**

Based on the initial analysis conducted we arrived at some inferences and recommendations for the ISP provider to improve his quality of service and his economic profitability.

The following inferences are arrived at:

* The internet has been used predominantly in the non-business hours. This has been due to the fact that phones took more precedence than internet usage in office hours which resulted in minimal internet usage at that time because of the use of Dial-up modems.
* The network traffic has been high during weekdays as people in Europe based on their lifestyle usually are outdoors during weekends.

Factoring these inferences, the following recommendations are made:

* Promotional offers can be introduced during weekends to encourage users to access the internet in weekends and business hours.
* Daily maintenance can be conducted on the servers and load balancing (Shutting off unneeded servers when not used) can be done during afternoons as usage is low.

These recommendations will help the ISP provider efficiently use his resources to provide the best service to his customers across the 11 European countries while also finding newer sources of profitable income.

**ANALYSIS:**

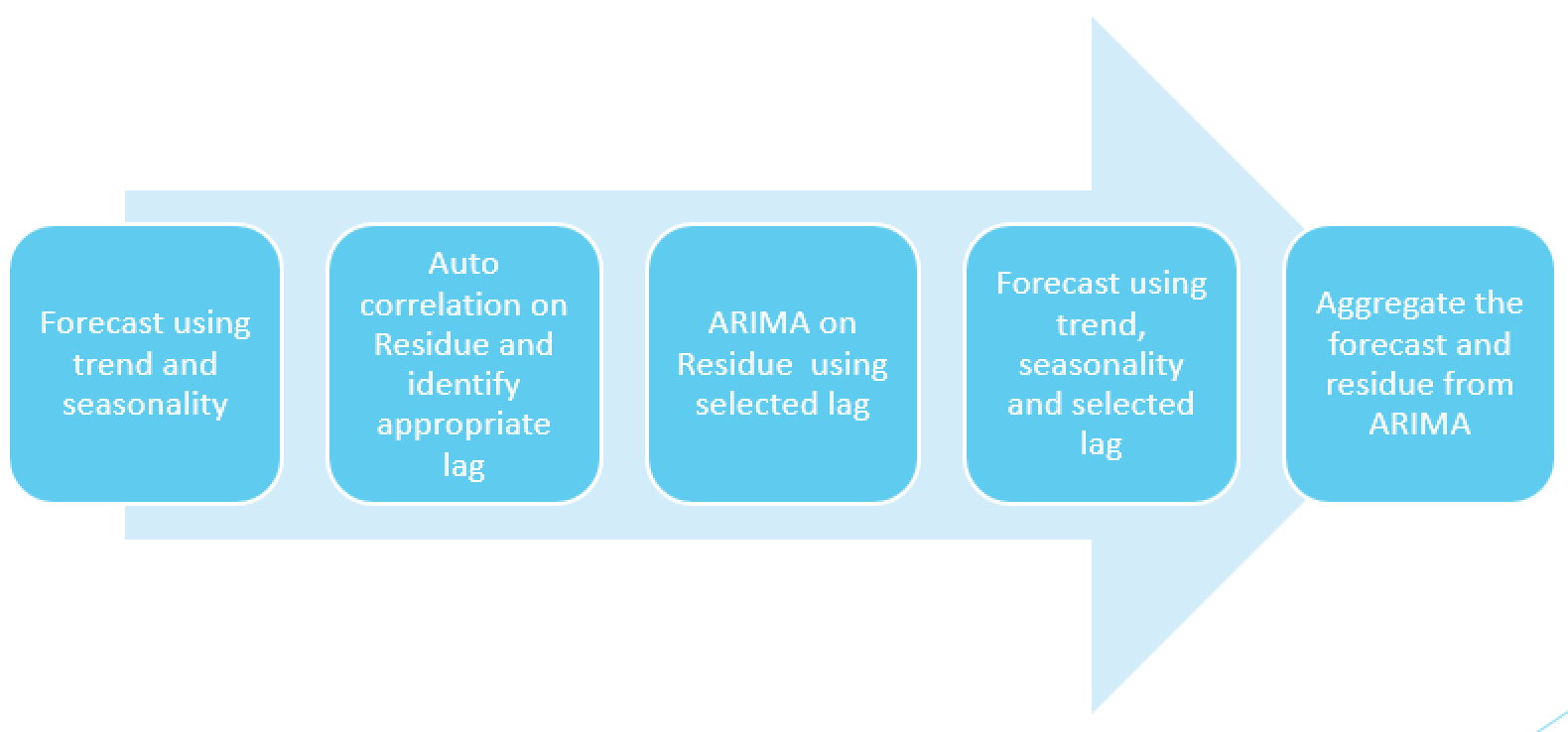
**PREDICTION MODEL:**

We now look to do a prediction from the given data set using regression analysis. We will first look at the four components of a time series plot and how the time series prediction will be conducted in a step by step process below.

The four components of a time series that we will be looking at in our prediction model are:

* Level
* Trend
* Seasonality
* Noise

Our approach in prediction is explained in the step-by-step sequence shown in the diagram below:



As initially stated, our approach will combine all the results obtained from analyzing each component individually and then accumulate them to get one final improved model which will then be improved by advanced time series analysis techniques such as smoothing and ARIMA analysis. Which will provide us the best prediction model that can be obtained with the given dataset.  
  
Let us look at analyzing the first component of the time series i.e. Level

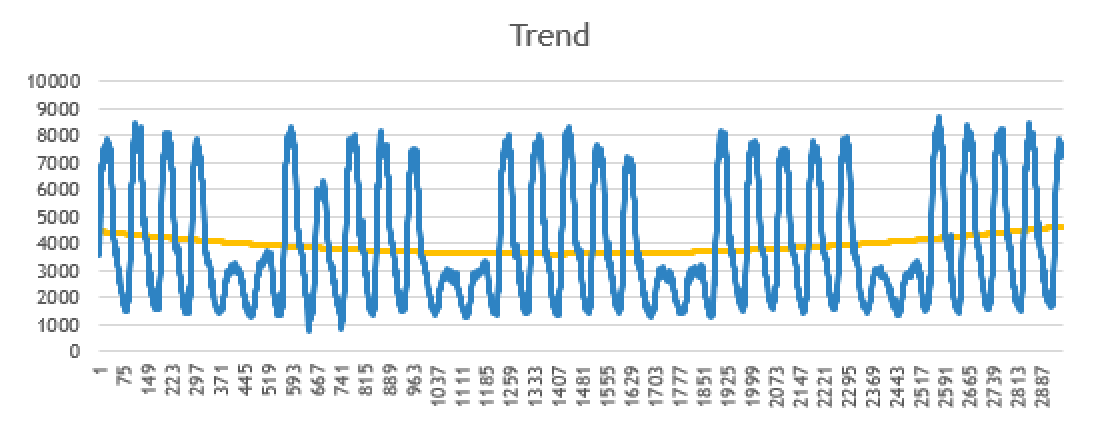
**Level:**Level is defined as the component that is obtained by taking the average of all the values and then drawing a line in the plot.

**TREND ANALYSIS:**

Trend is the component of either a steady increase or decrease that is usually seen in many time series data. A time series dataset is supposed to have positive trend if it increases with time and it is supposed to have negative trend if it has constant decrease in value with respect to time.

In order to find this component in our dataset, we plotted the data as a two dimensional plot and observed the data’s behavior.

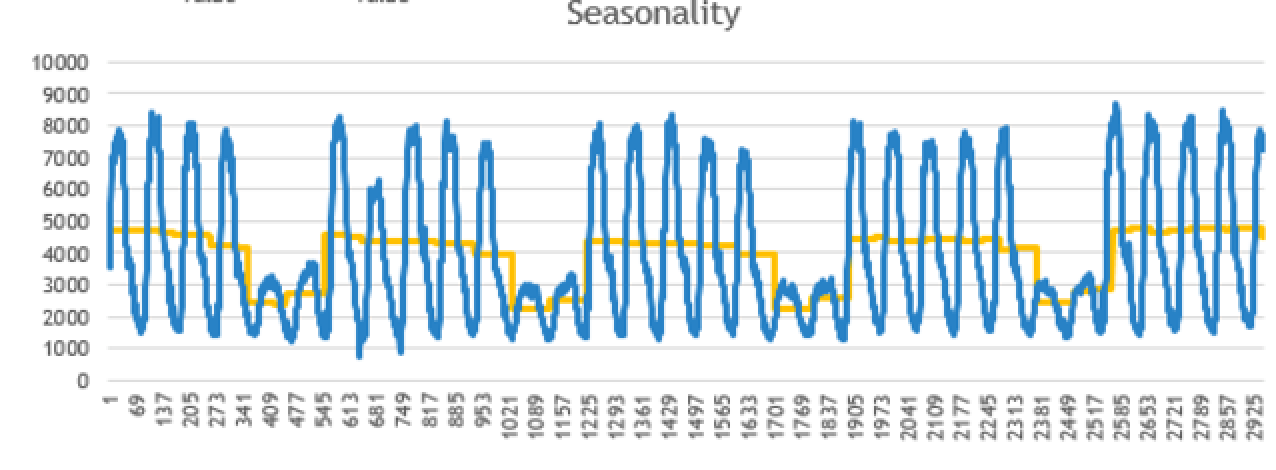
The plotted data and its study is shown below:



From the representation shown above in which the trend in the data is shown as the orange curve, we find that there is no particular type of trend that is seen in the given dataset, so from doing a trend analysis of our data we conclude that we do not have any particular trend that is evident in the data. This will later play an important role in deciding which method of smoothing that we will apply to the data.

**SEASONALITY ANALYSIS:**

Seasonality is the component of a time series plot in which we identify repetitions of patterns at certain intervals (Ex: Rise in travelling ticket sales in April and December due to summer and winter holidays). From the two dimensional plot we have for the dataset we found the seasonality pattern as shown below:

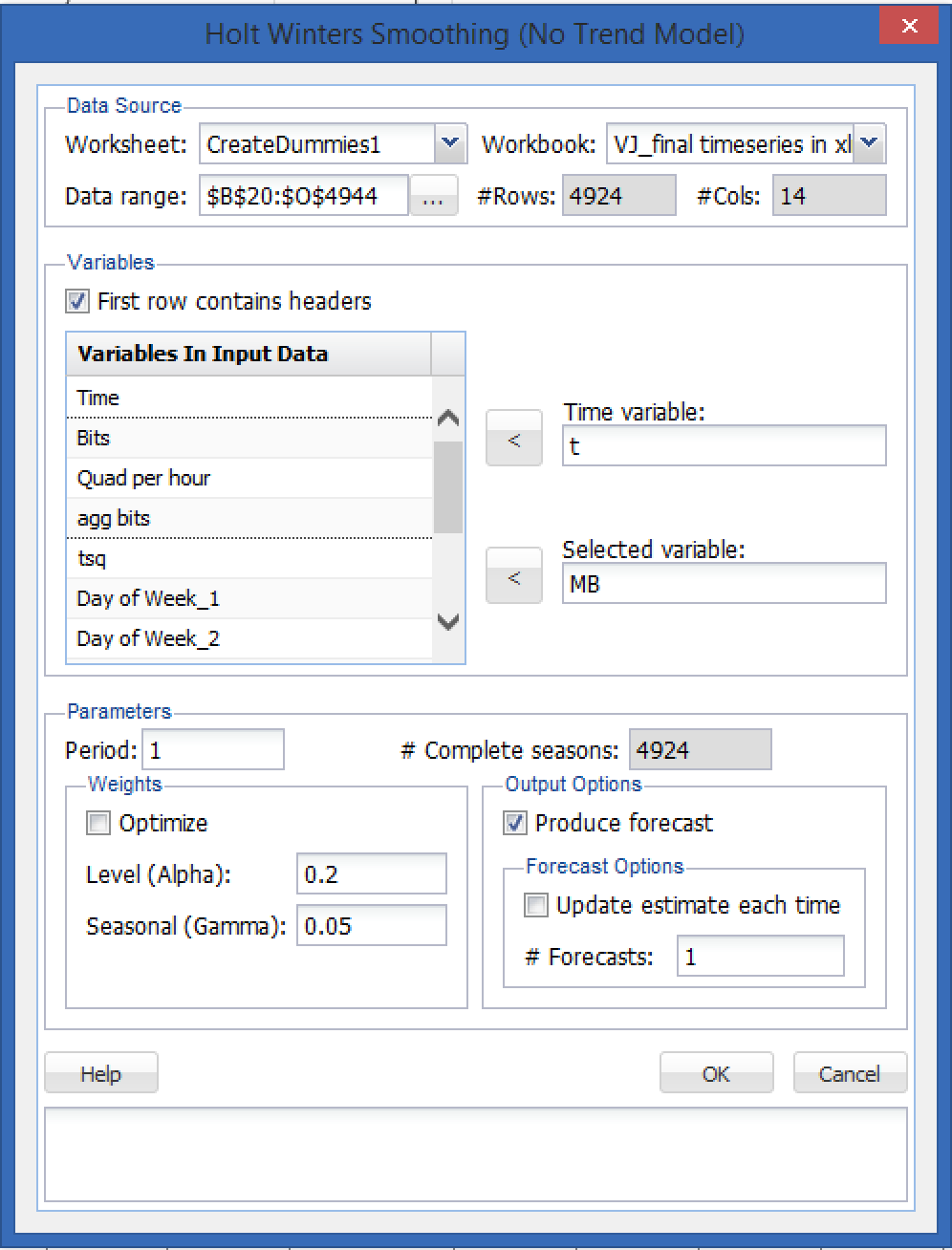


The yellow line depicts the seasonality observed in the data. In order to factor in the effect of seasonality into the prediction model that we will build, we will introduce dummy variables into the data model that will help us observe the seasonal behavior of data in every day of data consumption.

**INITIAL PREDICTION MODEL:**

We will now look at screenshots observed when the data model factoring in the three components of time series was built using XLMiner:

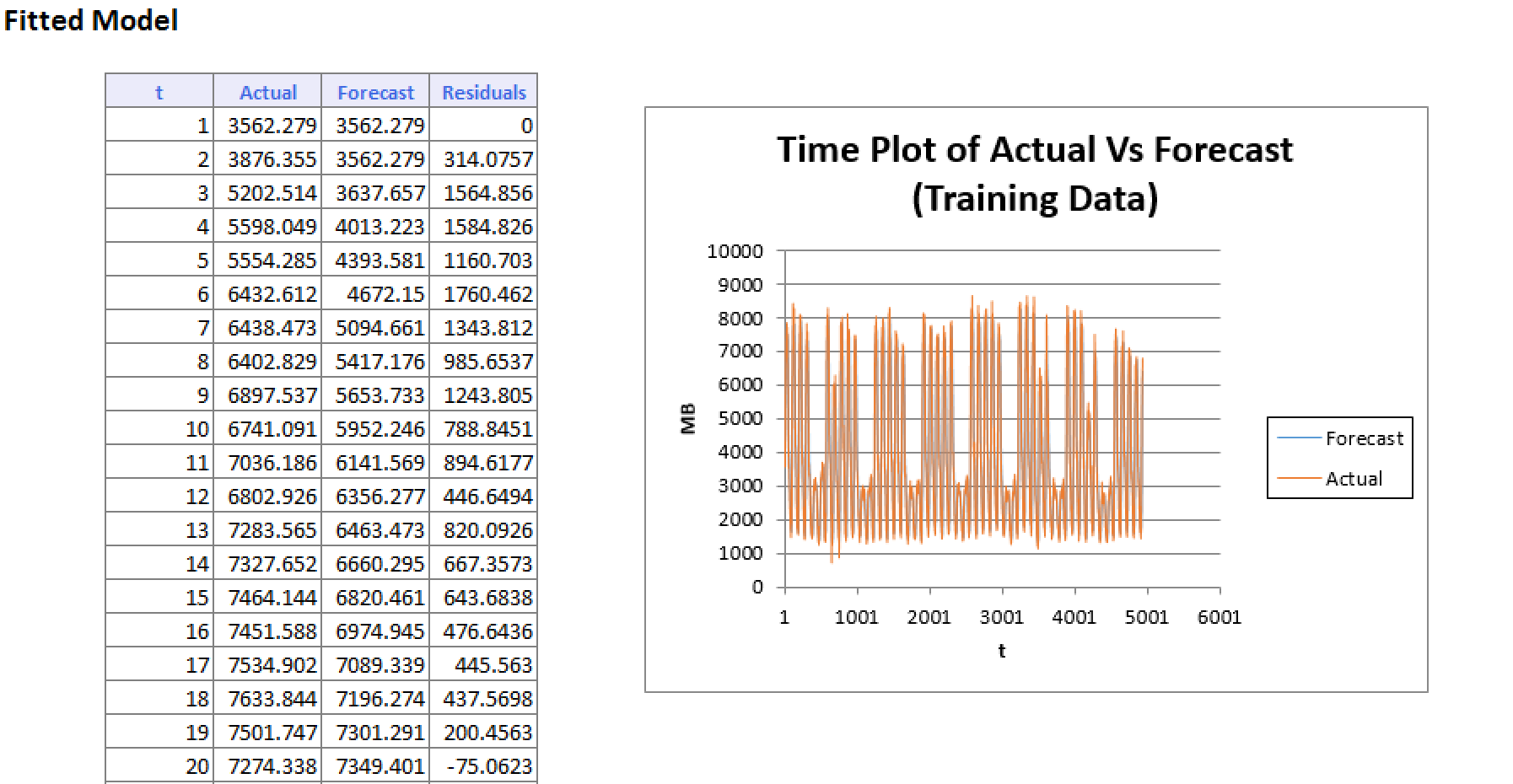
We began with applying smoothing techniques to the data. We took the squared value of time (tsq) into the list of parameters that will influence our model. We applied the Holtwinter smoothing technique to the data as our dataset has no trend but seasonality. We applied the smoothing principle using XLMiner.

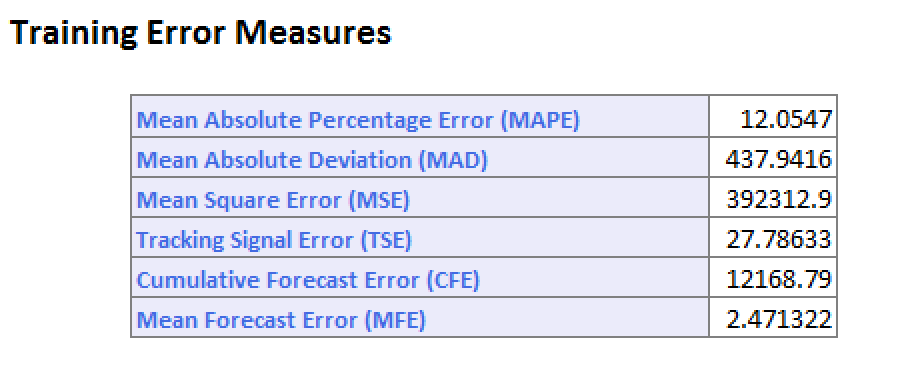


Our selected variable was the Megabytes of data consumed (MB) and the time variable was “t”. XLMiner automatically chose an alpha value of 0.2 to compensate for the level component of the time series and a seasonal value of 0.05 for the seasonality in the data. After applying the smoothing to the data the following output plot was observed.

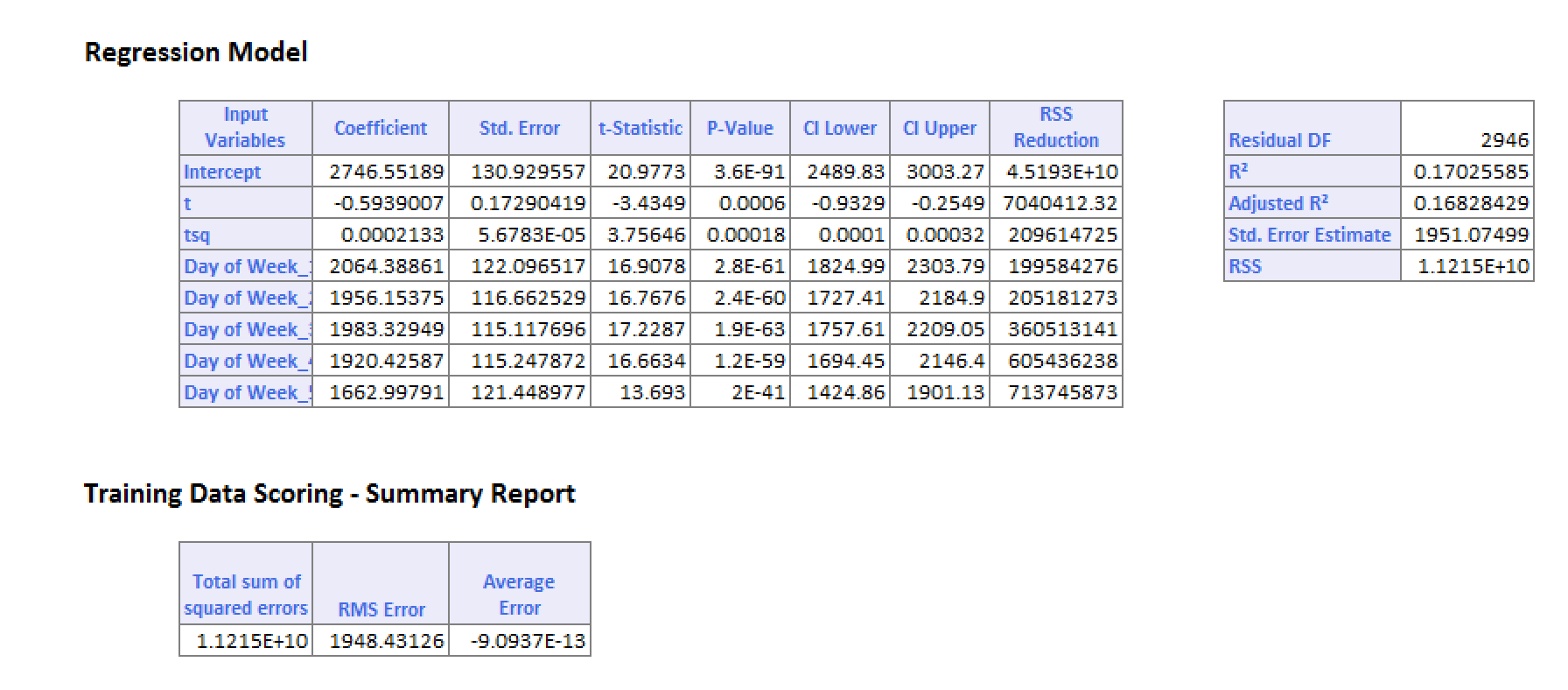
The following images shown the output observed after the smoothing techniques were applied.

**HOLT-WINTER SMOOTHING TECHNIQUE**



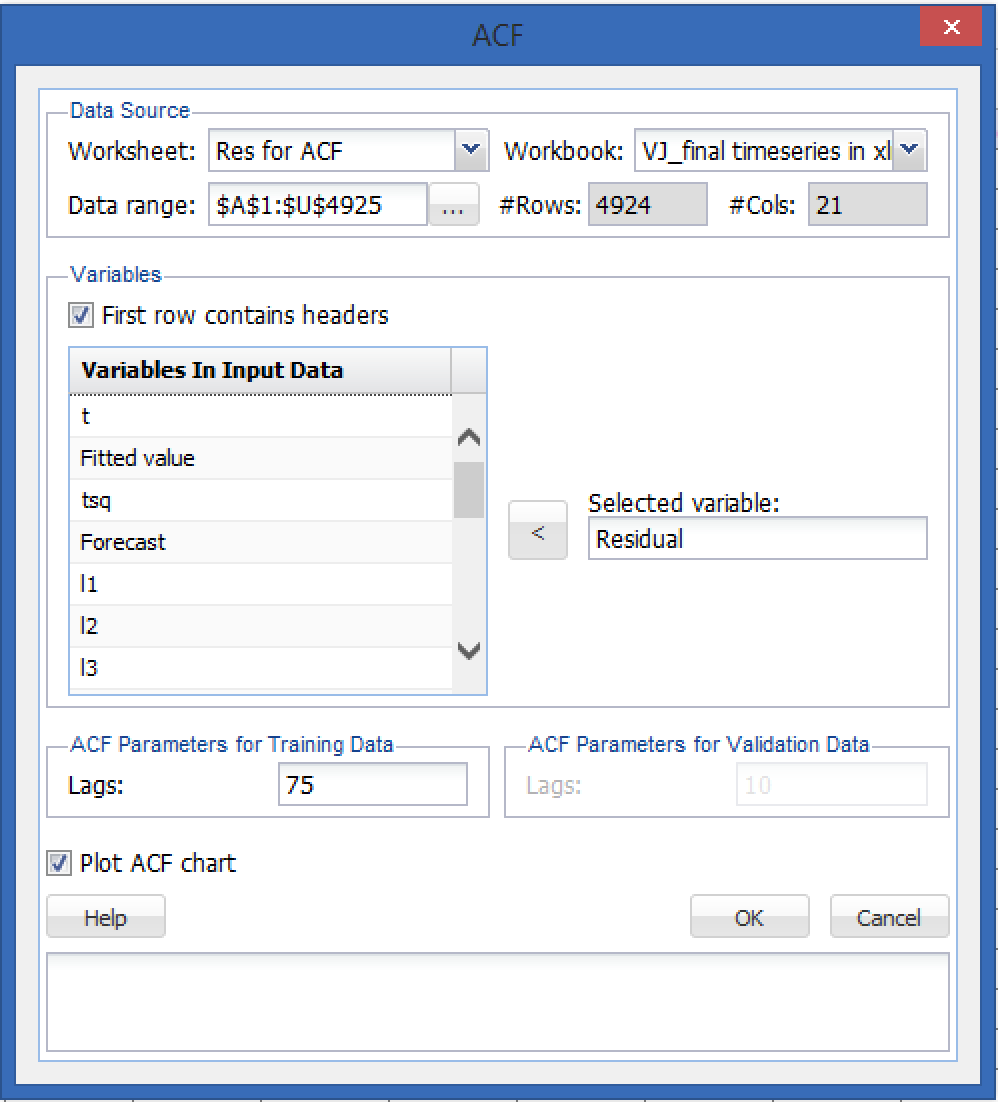


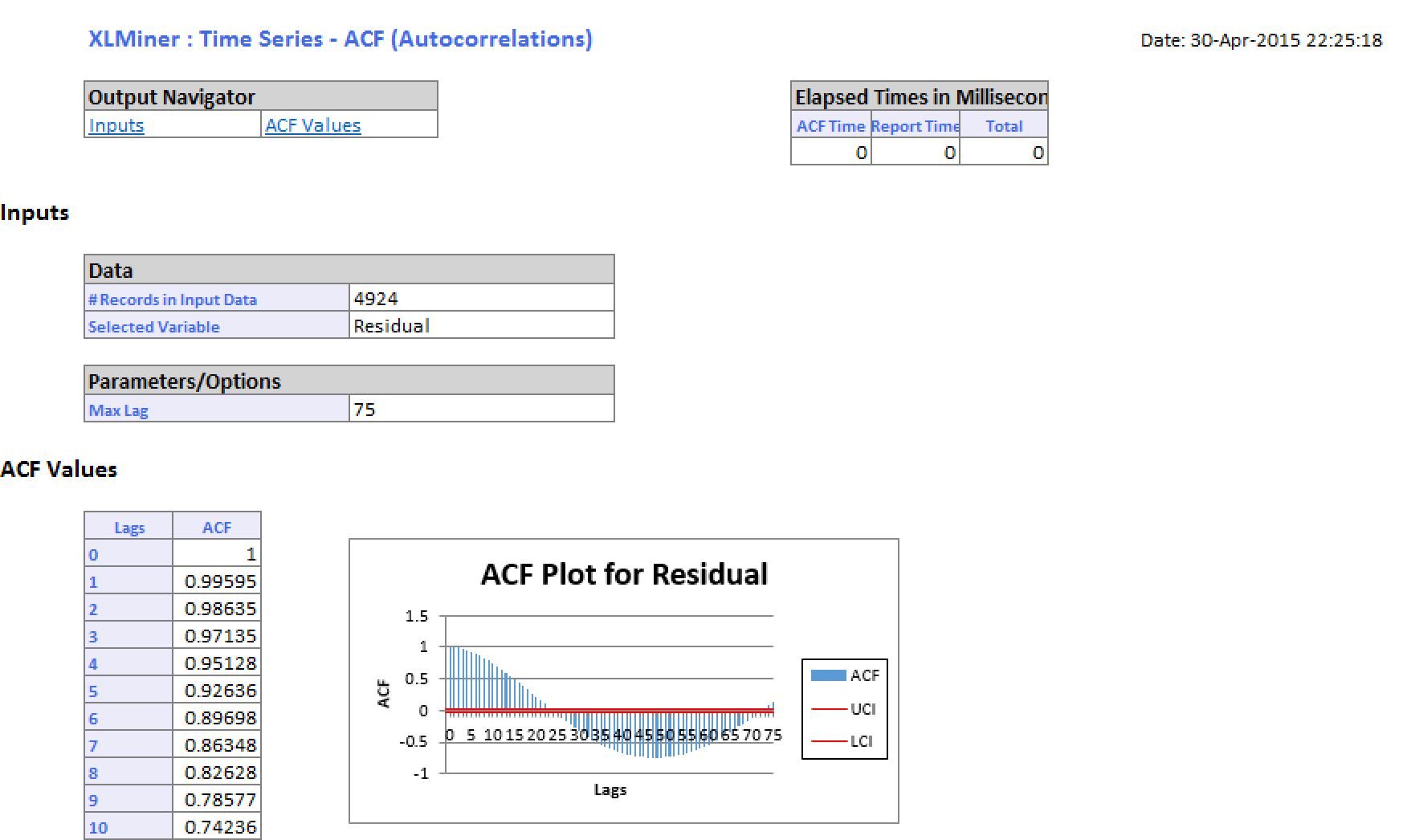
Upon completion we ran a dry run of the Multi-linear regression algorithm factoring in the trend and seasonal components that we had analysed and the following RMS error value was reported



Upon identifying that the initial MLR Run presented an RMS error value of 1948 MB which is high we decided to improve upon on this model by doing the first step of data model improvement which is the application of Autocorrelation Function technique (ACF).

Autocorrelation function looks at the correlation between data points if they are present at any given time. This is done by introducing lag into the time component of the time series data. A lag is basically the data being shifted down by one time unit. We applied the Autocorrelation into our dataset and visualized the results in the dataset below.

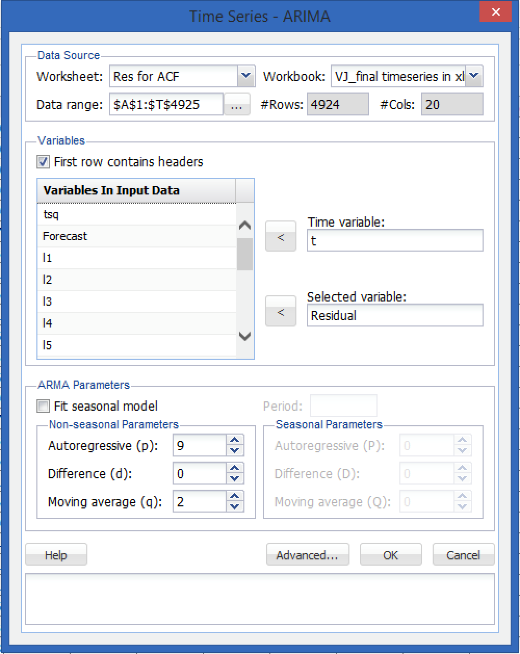


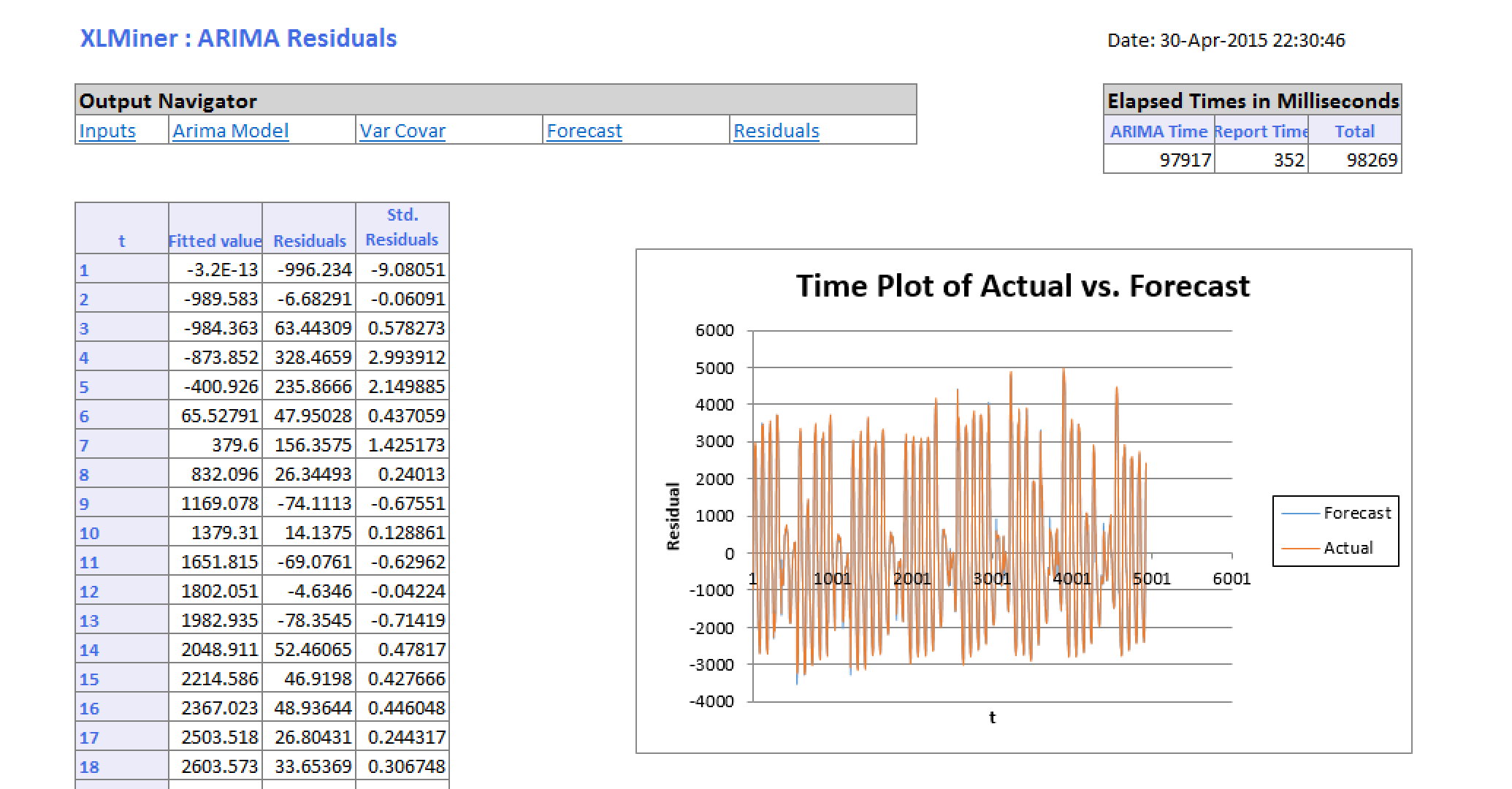


From this plot, we find that until 9 lags we have significant correlation that we can use to improve our data model. Hence we factor in 9 lags in the time part of the time series data that we have. This now provides a changed dataset with the lags factored in.

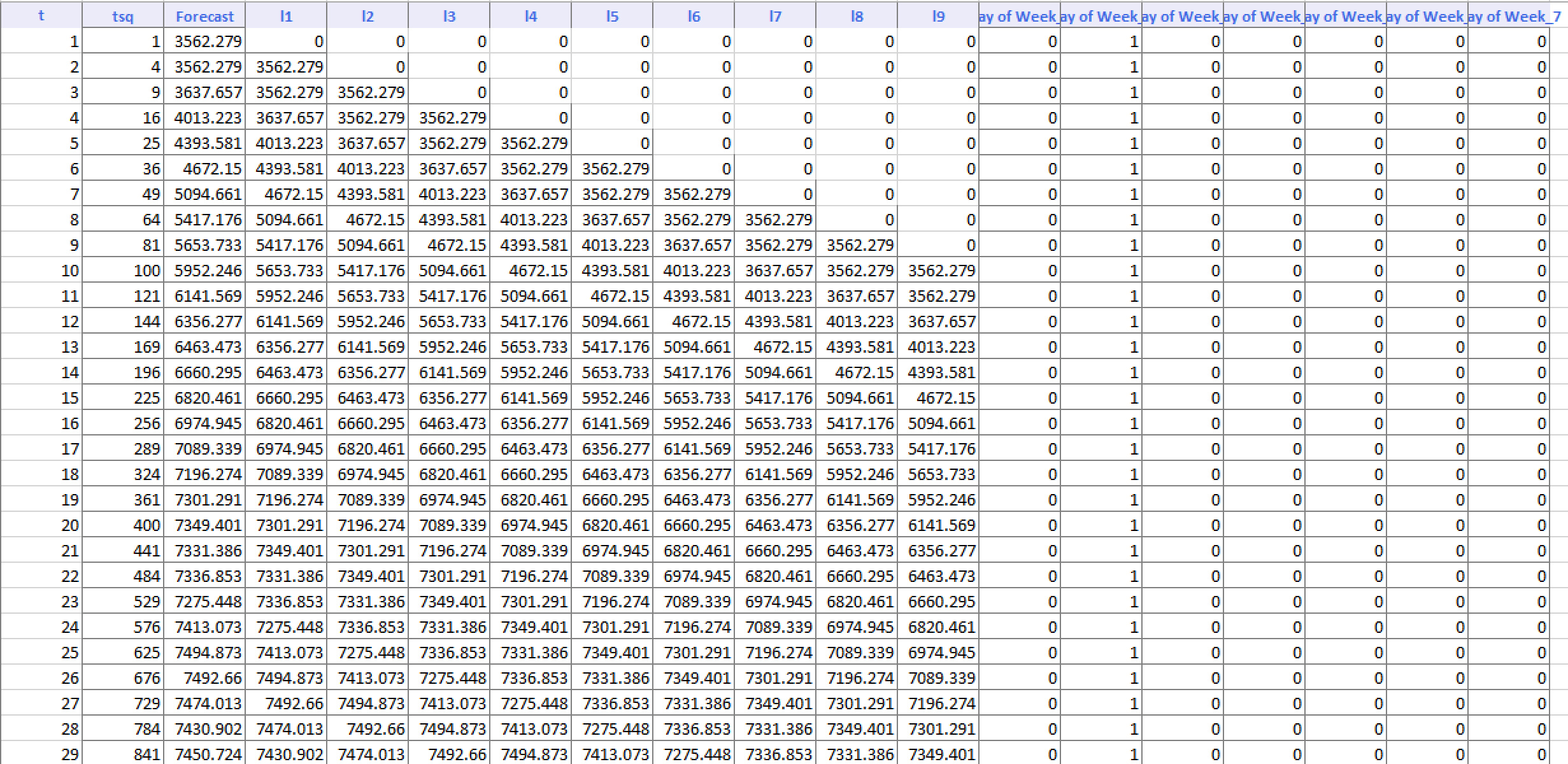
**ARIMA ANALYSIS:**

ARIMA stands for Autoregressive Integrated Moving Average. It is a technique by which we hope to extract as much useful information as possible from the residual values that are obtained from the actual versus predicted values. The model applies techniques such as auto regression, integration or difference and moving average that are applied to the two dimensional plot of the residues to extract useful information which can then be feedback into the model to improve on points that have either been overshot or under predicted by the model the remaining points in the residue are points of random noise from which no useful information can be obtained.

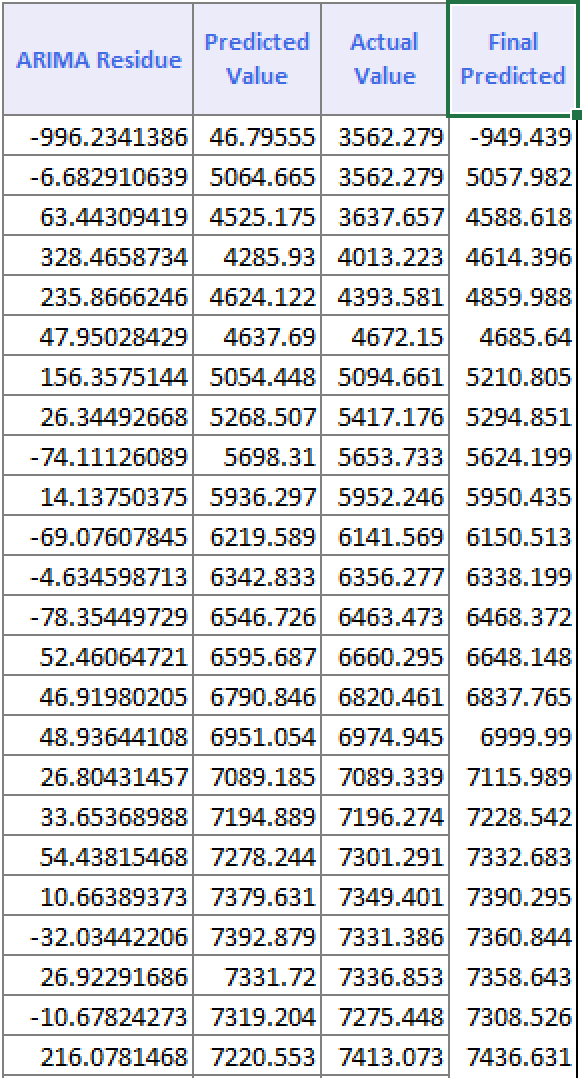
9



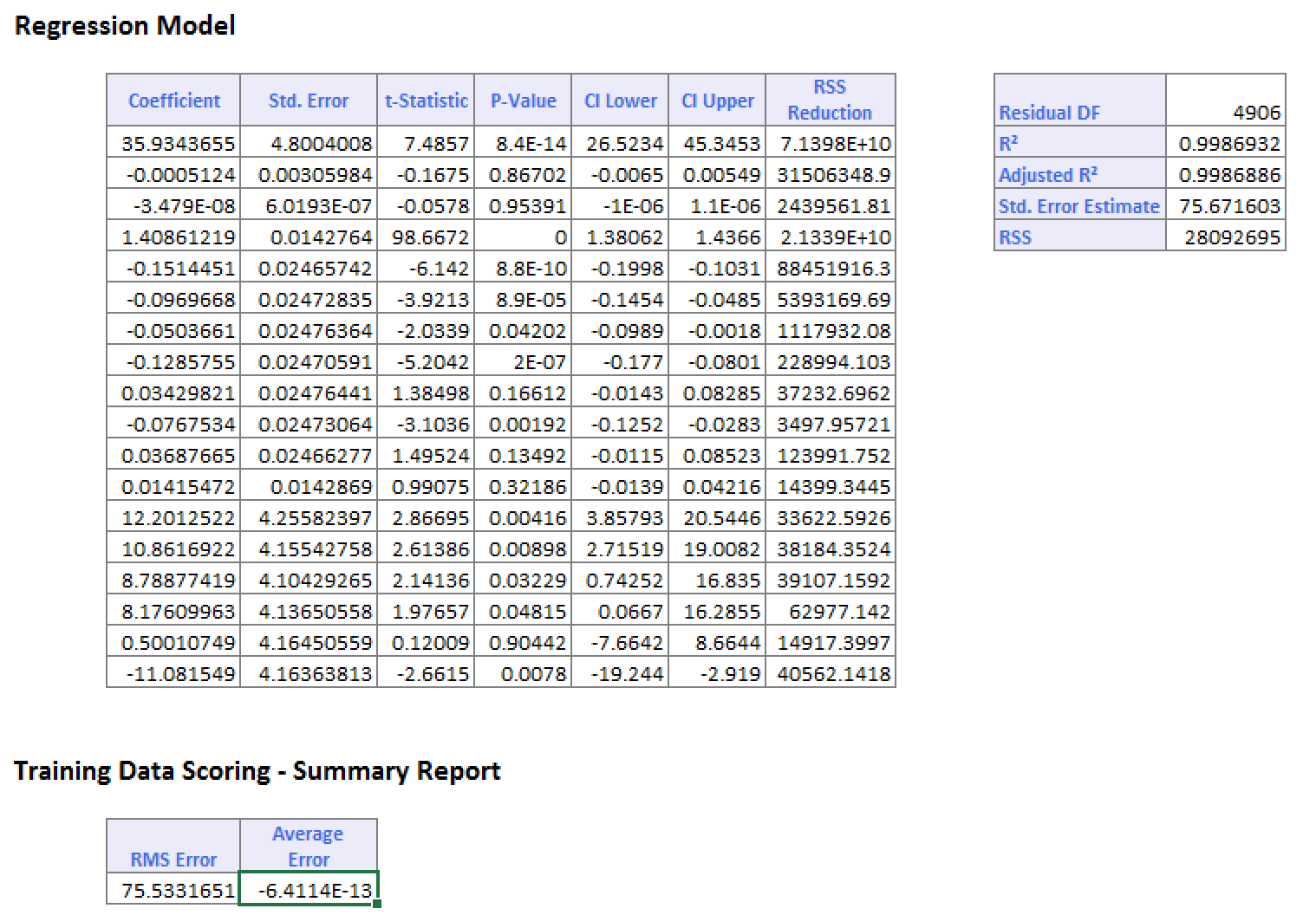
As we found out from the autocorrelation function, up to 9 lags there was a major effect on the ACF value which we will now factor into our new MLR model.



Upon running the new MLR model, we get new predicted and actual values. Now we use the ARIMA residuals that we obtained and feed them back into the MLR output.

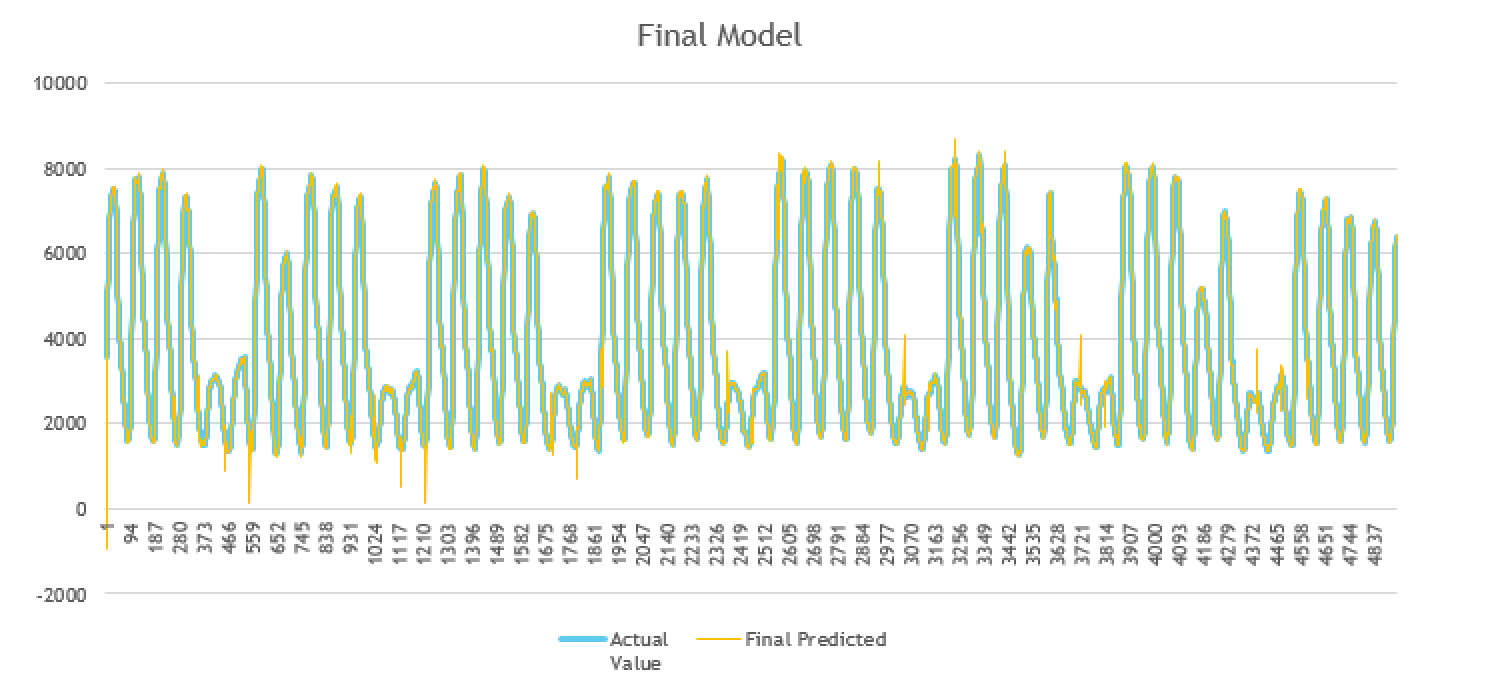


Here the predicted value is the sum of ARIMA residual and the predicted value of the MLR Model. We now plot the new predicted value against the actual value



**IMPROVED FINAL MODEL (Full):**

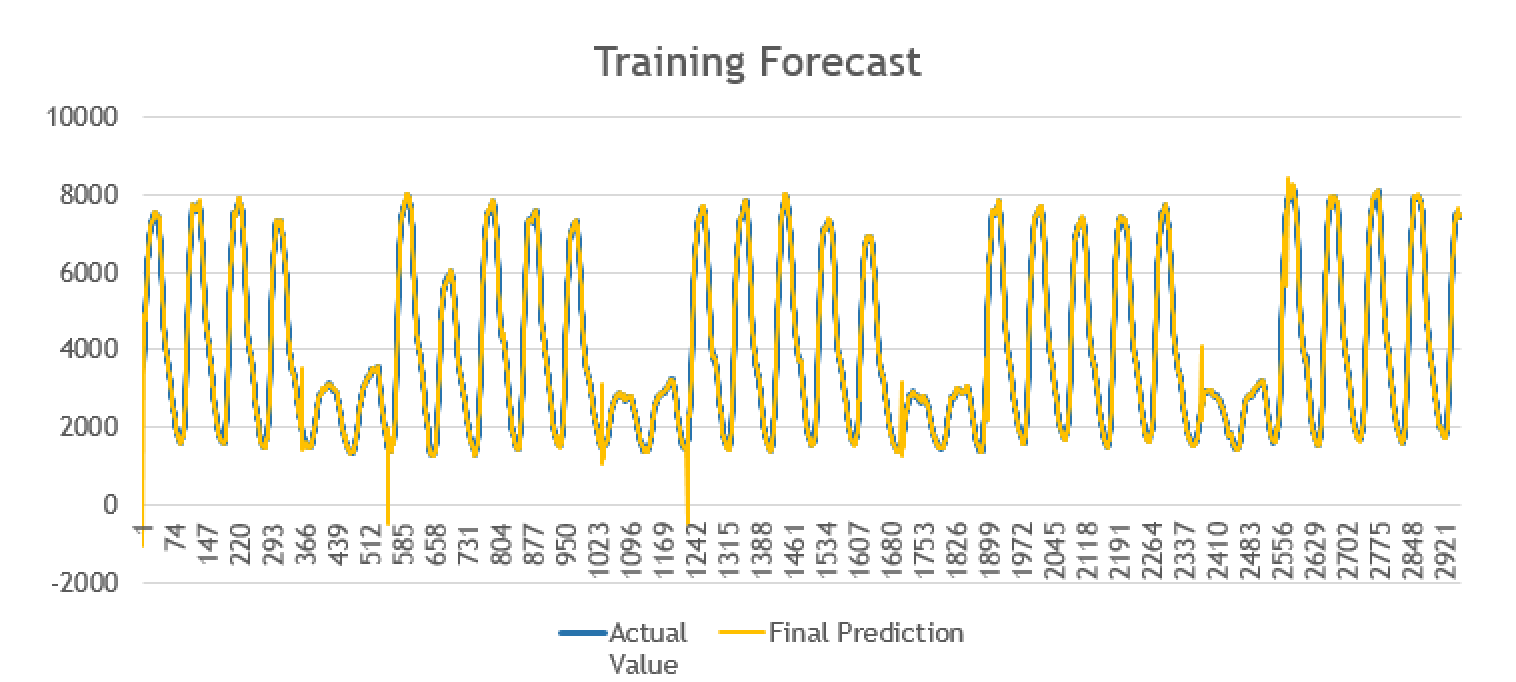
We now begin building the final and improved model which comprises of the features that were developed in the trend and seasonality analysis conducted with the Autocorrelation factor entered in the ARIMA analysis conducted on the residual data that was obtained upon comparing the actual and predicted values from the last MLR model output and upon combining these we get the final MLR predicted model. The two dimensional plot for which is shown below.



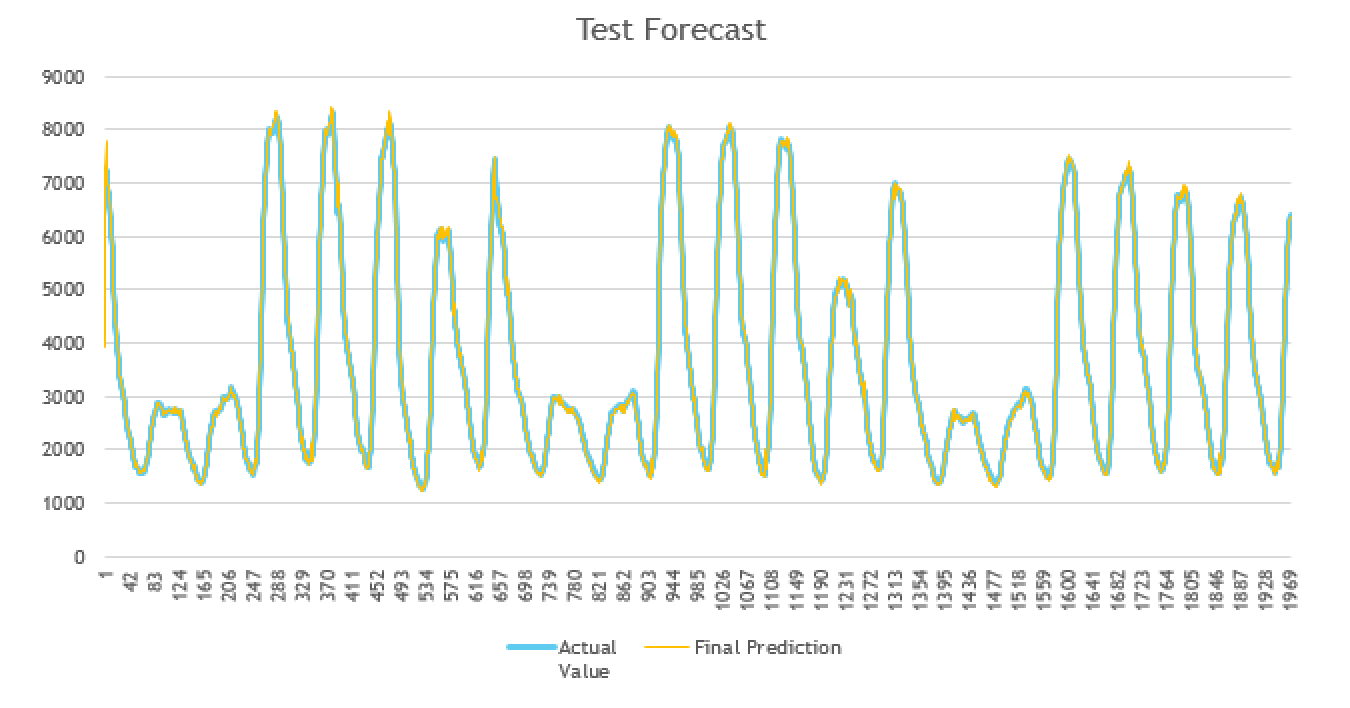
**TRAINING AND VALIDATION FORECAST:**

When the dataset was partitioned into training and validation dataset on a 60-40 ratio the following two dimensional plots were achieved.

The RMS value returned for the training dataset was 87 MB

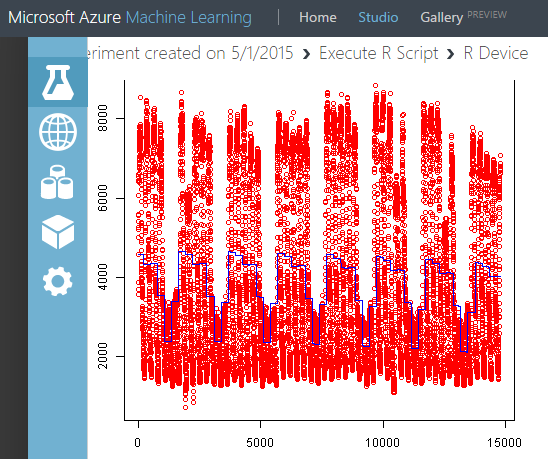


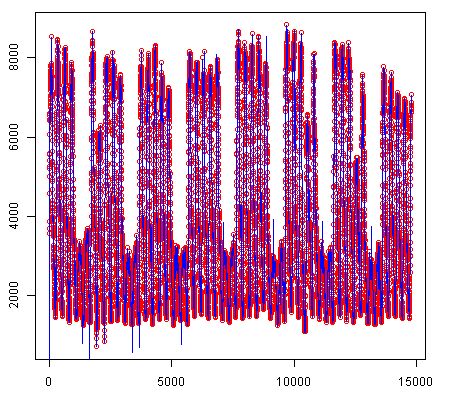
The RMS value returned for the validation dataset was 137 MB



**R SCRIPT MODEL IN MICROSOFT AZURE:**

We then scripted the data model using R Programming language and then ran the script using the cloud based data mining tool Microsoft Azure. We observed the following results.





**CONCLUSION:**

Hence we were able to look into the given dataset on an exploratory basis to gain insights and provided recommendations and also analyzed the given data set on the impact of each of the four components in the time series such as level, trend, seasonality and noise then built a model that in a step by step approach accumulated the lessons learnt from them into building the model which was further improved by the principles of smoothing and ARIMA analysis to give a final model which was then scripted using R programming language and run on the cloud platform of Microsoft Azure.

**REFERENCES:**

* [www.datamarket.com](http://www.datamarket.com)
* Little book for R
* Wikipedia