Web Traffic Forecasting

MSCI – 718 Final Project

By – Group 17

Introduction

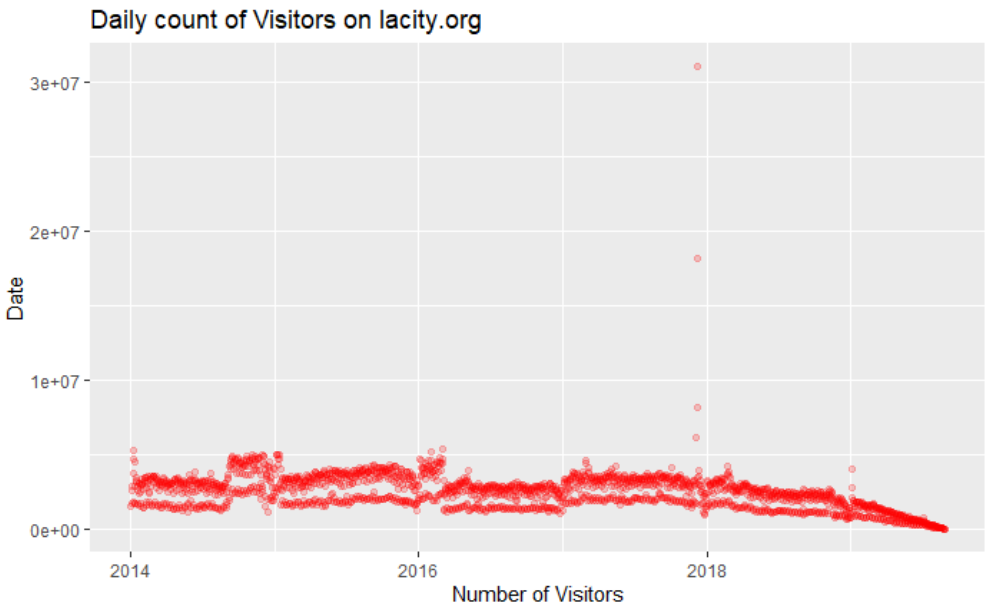
Web sites must forecast Web page views in order to plan computer resource allocation, estimate upcoming revenue and advertising growth (if website is for business), and to ensure secure, reliable and qualitative networking. This project aims to give a detailed step-by-step analysis of such a time series data, collected from the lacity.org website in an attempt to analyze its web traffic pattern.

We begin with Exploratory Data Analysis, which highlights some characteristics and components of our data, followed by fitting an appropriate model to finally forecasting and testing the model's accuracy.

Examining the data-set - EDA

The data-set contains more than 8 million observations described by 6 variables. Starting from 01/01/2014 to 07/31/2019, we have the record for less than 6 years. This data frame contains the data on the number of visitors, number of sessions, bounce rate, date, browser information and the device category.

We are only concerned with the total number of unique visitors to the website and their respective dates. So, we pre-process the data frame to include the univariate predictor variable - "X..of.Visitors" and outcome variable - "Date". We start by plotting the series and visually inspecting for any outliers.



We see that we have incomplete data for 2019, so we use data from 2014-01-01 to 2018-12-31 further on by creating a time series object.

Another assumption that we make here is, setting frequency as 365, despite the fact that 2016 is a leap year. We could possibly reason it as - because we aim to make forecasts on a daily basis, it won't affect our modeling.

We see that nearing the end of 2017, the number of visitors is usually high for 3-4 days. This could be observed as a suspected outlier, as it could be due to spammers attack. As this would bias our model we remove the outlier using the function: tsclean() which replaces the outlier using series smoothing and decomposition.

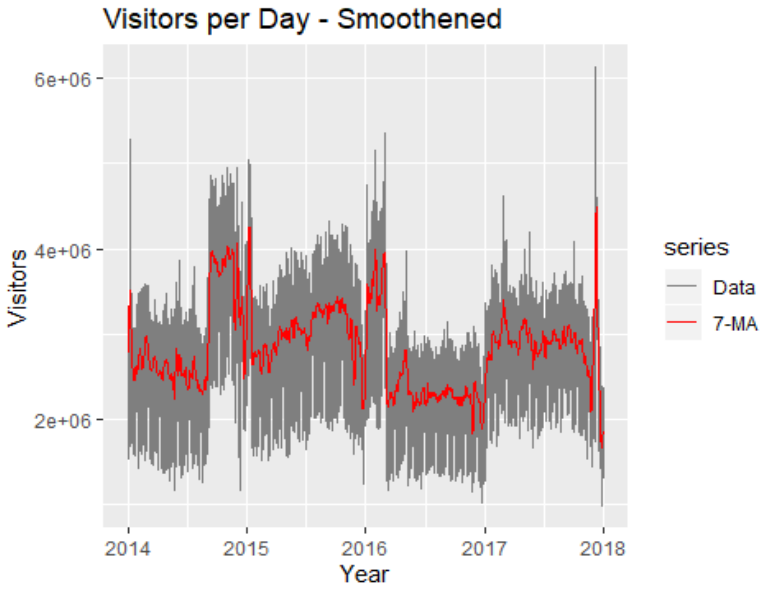
We split the data to train and test data where records after 01/01/2018 is used to test the model and the records from 01/01/2014 to 31/12/2017 is used in creating a model.[Plot-1]

Decomposition of Time Series into Components

We start examining the series by first zooming in, then later visualizing a zoomed out variations. Drilling down to more granular level i.e. Weekdays, we see that there is a Weekly cycle. 'lacity.org' is the government's official website. As the government offices' working hours are from Monday to Friday, we see a high web traffic on weekdays, and could say that visitors include government people and a general public which accounts for the weekly pattern of the series. [Plot-3]

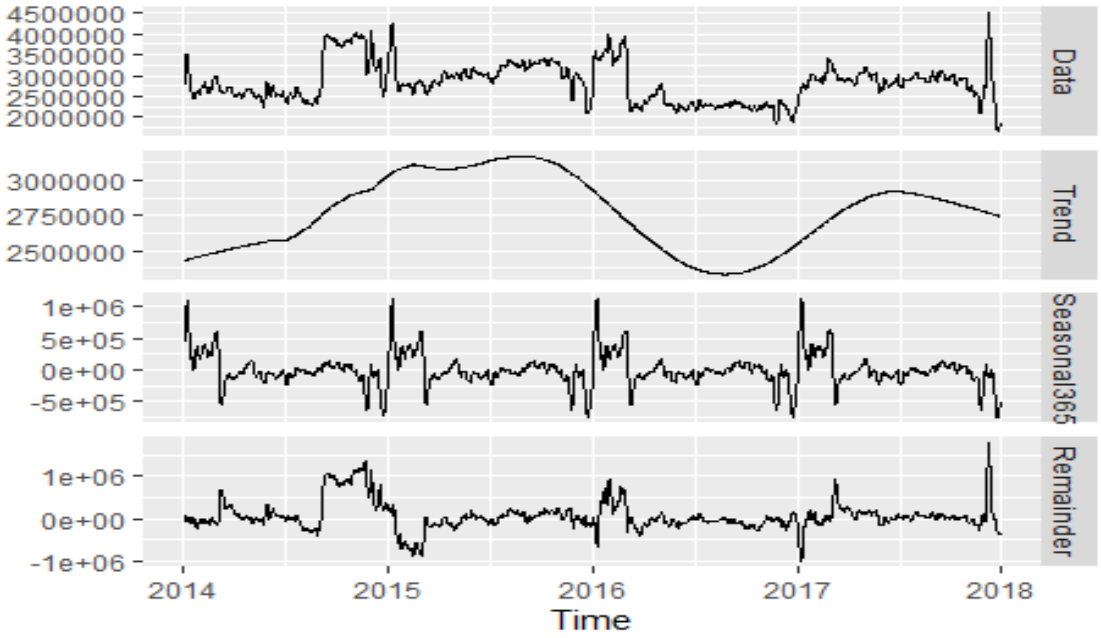
A time series is made up of seasonality, trend, cycles and error. Seasonality refers to the fluctuations in the series in accordance with the calendar cycle whereas trend is the overall increasing or decreasing trend of the series. A cycle occurs when the data exhibit rises and falls that are not of a fixed frequency. The error or randomness in a series is the uncorrelated data points contrary to the obvious correlated adjacent data points which accounts for the unexplained fluctuations in the series.

This series seems additive in nature as we see that the variation in seasonal pattern is cyclic and does not increase or decrease as the trend vary[Plot-2]. Also another seasonal component is weekly seasonality which we identified earlier. Thus we can say that the series is the sum of its components, trend, seasonality and randomness. The trend does not show a constant increase or decrease, however it increases initially, fluctuates and then decreases for some time, after which it increases linearly. We could sense a cyclic nature - that any two consecutive years have similar patterns, but years far apart seem to have different patterns.



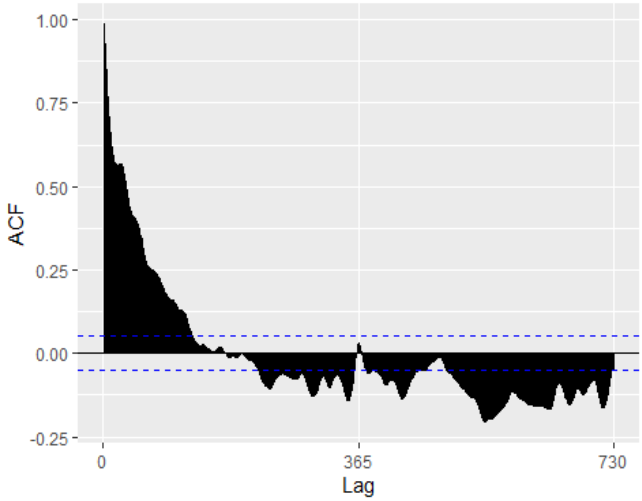
To avoid imposing of multiple seasonal patterns - i.e. weekly cycles and the yearly cycles, we average out the weekly cycles by using Moving average technique - a data smoothing technique, thus making the series more stable and predictable. Here, we take a period of 7 giving us a weekly moving average (Visitors per Day – Smoothened)

Now we decompose Smoothened series using STL, an acronym for "Seasonal and Trend decomposition using Loess" This decomposition method assumes that the seasonal component repeats from year to year, however we control the trend-cycle, which here we choose to be yearly (365). Decomposing the smoothened series gives almost the same trend, but seasonality can be better visualized. Hence, we conclude there is a seasonal effect with a varying trend.



Stationarity

Owing to the fact, that there is some seasonality and variation in trend, we can say that the data is not stationary. The series that we have in our hand does not have its statistical properties same in the future as they have been in the past i.e. the mean and variance function are a function of time.



To further support our claim we check acf plots (correlogram) to see if there is any underlying pattern for trends or seasons. The trend is shown as a slow decay and the seasonal spikes can be seen superimposed on the plot. Most of the correlation values are significant i.e. most of the values lie outside the dotted line which maps 5% significance level.

Next we do some statistical test to confirm our claim that data is not stationary.

1. ADF (Augmented Dickey Fuller) Test -

Null Hypothesis: The series has a unit root (value of a =1)(p>0)
Alternate Hypothesis: The series has no unit root. (p=0)

The p-value is very less for short lags (<30) but not equal to zero. We could say for short periods our series is weakly stationary. However, our data has yearly trends so performing an augmented dickey-fuller test with lag=365 we see that p is not equal to zero(p=0.7593) which suggests that our time series data can be a non-stationary data as it has a unit root.

2. KPSS (Kwiatkowski-Phillips-Schmidt-Shin) Test

Null Hypothesis: The process is trend stationary.
Alternate Hypothesis: The series has a unit root (series is not stationary).

The p-values are less than 0.01, hence we reject null hypothesis in favor of alternate hypothesis that our data is Non-stationary. Summing up from both the test, we can say that our data is non-stationary.

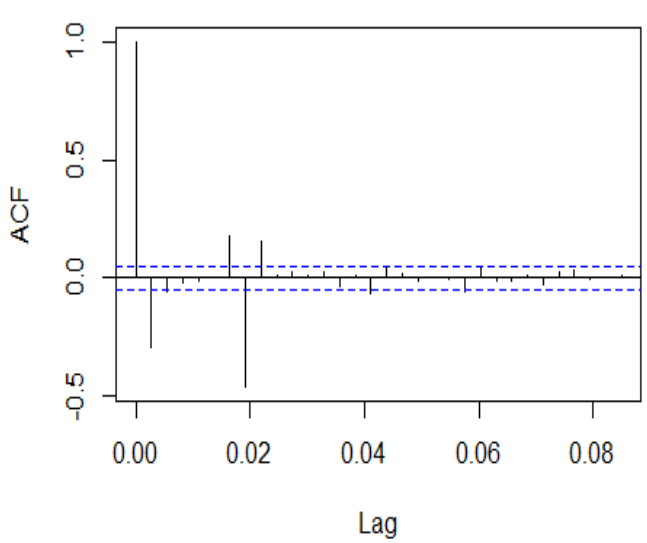
Fitting a model

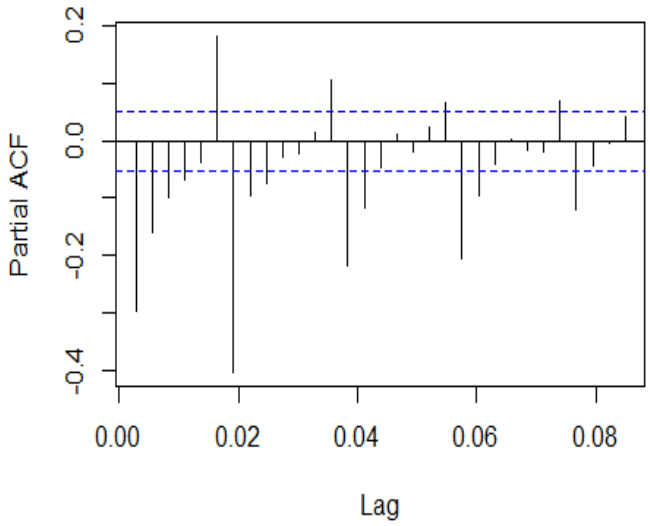
We know that our data is non-stationary and that it has varying trend and seasonal fluctuations. Hence, a valid reasoning would be to fit a non-stationary Time-Series model.

ARIMA models

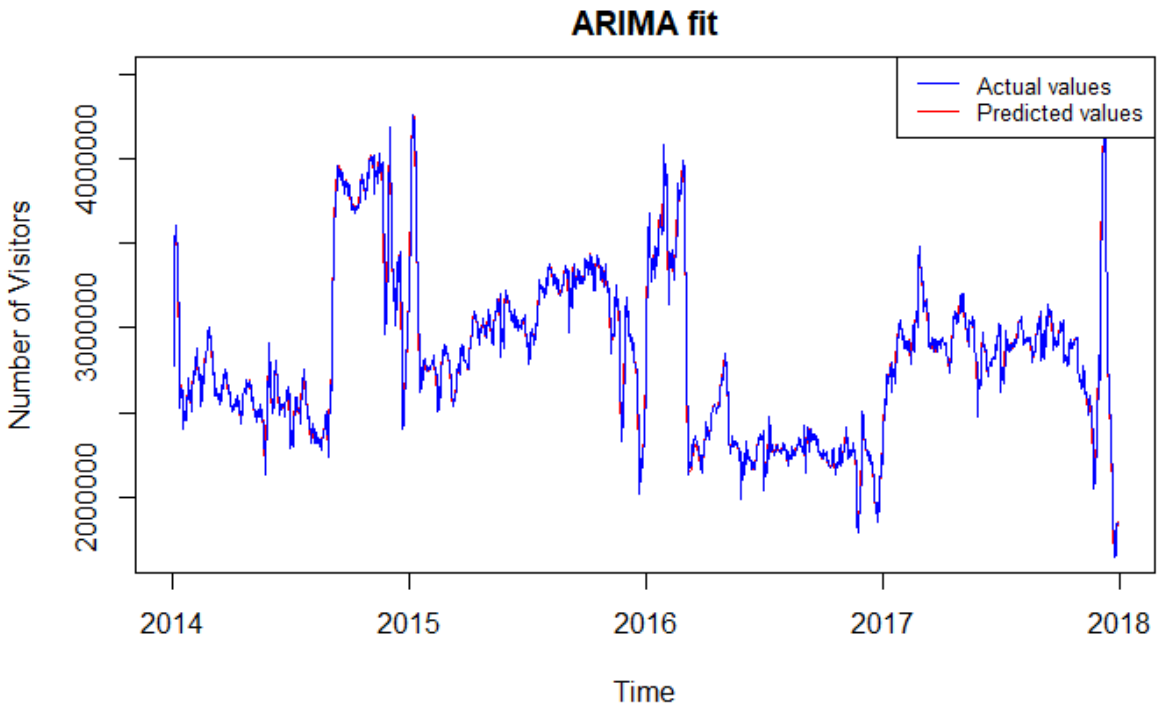
A non-stationary series can be differenced to form a stationary series and could be modeled using a stationary process by combining Auto-Regression and Moving Average processes. The resulting ARIMA process follows the model - ARIMA(p,d,q), which has 3 hyper parameters - P(auto regressive lags),d(order of differentiation),Q(moving average) which respectively comes from the AR, I & MA components. To identify p,d,q first we plot acf and pacf correlograms and the point where correlation drops to zero in acf plot gives q value, in pacf plot gives q value and the order of the difference is d, to make the series stationary. We do a log transformation to average out the variance.

In the acf plot, the value in x-axis where graph line drops to 0 in y-axis for 1st time is the q value, which is 2. In the pacf plot, following the same procedure, we get p value as 4. So we try to over fit and under fit some more models around the estimated p,d,q values and choose the model with least AIC.



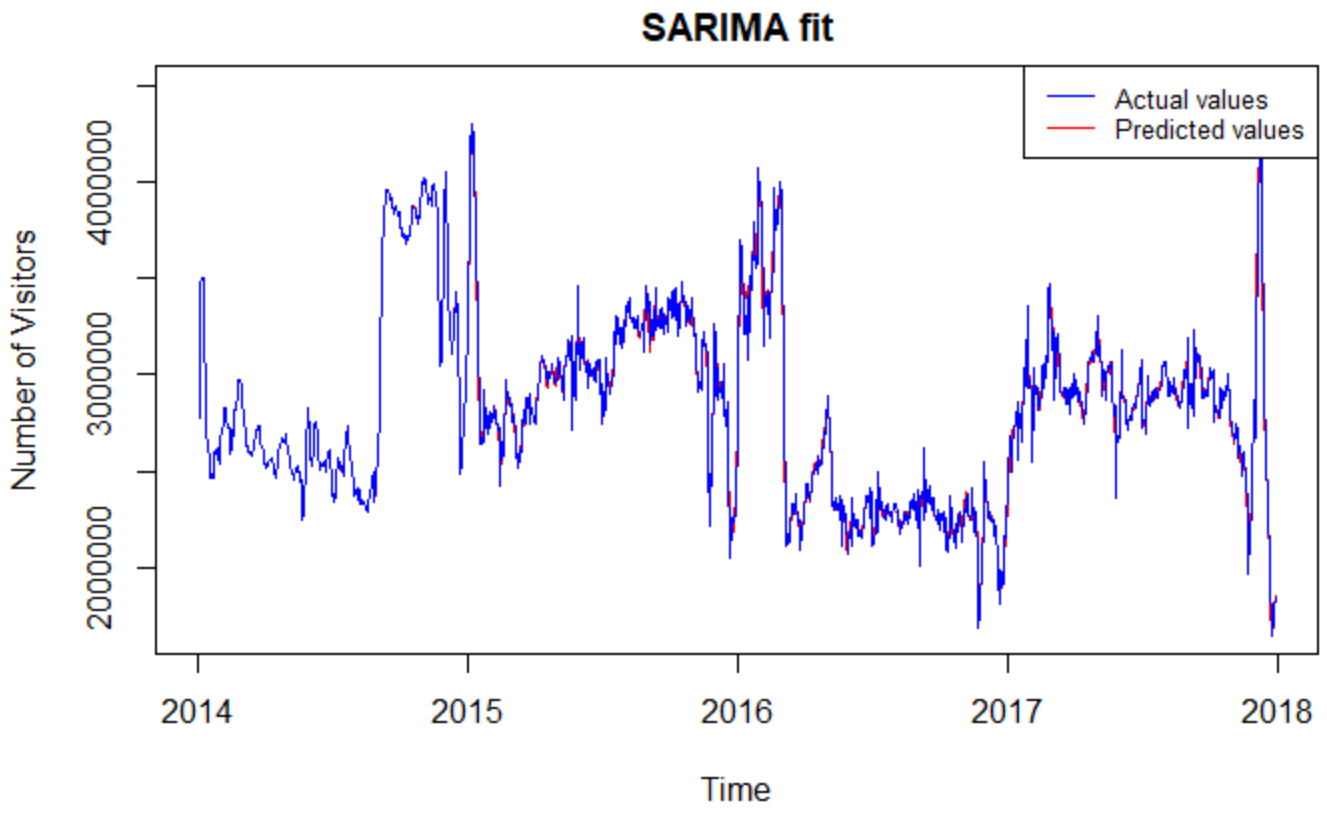


We get ARIMA model with p=5,d=1 and q=3 with least AIC, MAPE and RMSE value - which we consider as the best fit. The residuals[Plot-4] follows a constant mean of 0, a normal distribution and ACF plot shows an approximate white noise distribution with few significant correlations. To check there is no actual correlation we do a Ljung-Box test which gives a very high p-value (0.5731), thus retaining null hypothesis that observations are independent.

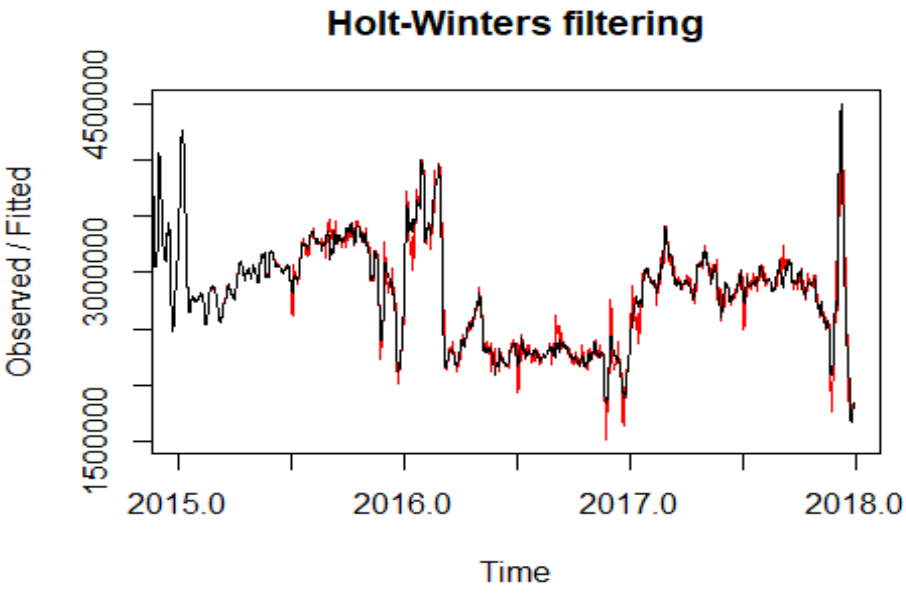


The ARIMA process can be extended to include seasonal terms, giving a non-stationary seasonal ARIMA (SARIMA) process. As our data has some seasonality, to create the model, we are going to make use of auto.arima function. It has the capability to create multiple models with different p,d,q parameters and it then picks the model with the least AIC value. Since, we need a seasonal ARIMA model, we set D=1.[Plot-5]

The result is an SARIMA model with (p,d,q)=(5,1,1),(P,D,Q)=(0,1,0) and m=365 (yearly seasonality). The residuals follow the process of a White noise and seem to be normally distributed. This model seems to be a better fit for our data, as also the AIC value is less compared to the previous ARIMA(5,1,3) model. Upon doing the Ljung box test, the p value is considerably greater than 0.05, thus not rejecting the null hypothesis of independence.



Holt-Winters Seasonal model - Triple Exponential Modelling:



Our data does not have a constant linear trend i.e. the level of trend changes. For this we can use exponentially weighted moving averages to update estimates of the seasonally adjusted mean (called the level), slope, and seasonal.

The optimum values for the smoothing parameters, based on minimizing the one-step ahead prediction errors, are 1, 0, and 1 for alpha, beta and gamma respectively. It follows that the level and seasonal variation adapt very rapidly whereas the trend does not. The residuals follow a normal distribution with high positive kurtosis, and is conditional heteroskedastic. The residuals also have a constant mean 0 and acf plot seems to follow the distribution of white noise roughly. However, upon doing the Ljung box test, the p value is very low, which rejects the null hypothesis of independence. Thus, this model could not be considered a good fit.[Plot-7]

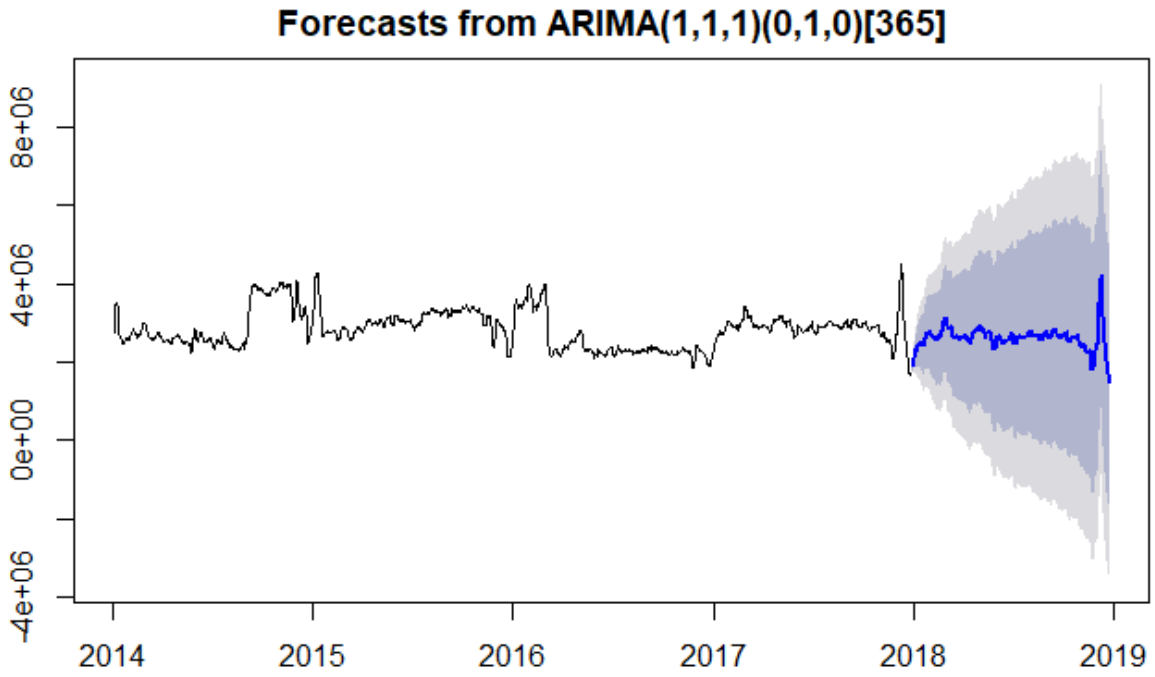
Forecasting

Forecasting daily number of visitors from a dataset which contains 4 years of data can be tricky as it contains multiple seasonal cycles, despite the smoothening. It should also be noted that the extrapolated forecasts are based entirely on the trends in the period during which the model was fitted and would be a sensible prediction assuming these trends continue. So we choose to use Holt-Winters model with alpha=1 and gamma=1 for forecasting long term predictions and ARIMA(5,1,3) could be used for short term forecasts.

ARIMA forecasting

ARIMA models has a MAPE value of 1.23, which means on average, the model fitting is off by 1.23%. A Model having MAPE value less than 20 is considered a good model. MASE is 0.05915596, our goal is to have a MASE value lesser than 1, and that we have achieved. MAE is 36396.82. This means that on average our fitted data is off by 1.2% which is actually very good.

Forecasting for 365 days gives almost identical pattern of that of previous years' value, which could be interpreted as ARIMA models are better suited for short term forecasting as it captures short term fluctuations better. Upon comparing with our test data, we find out that MAPE value is 20.098 i.e. forecasts are off by 20.1%. MAE is 497553.5, which gives 17.54% of variability, which is reasonable considering we are doing a long term forecast.



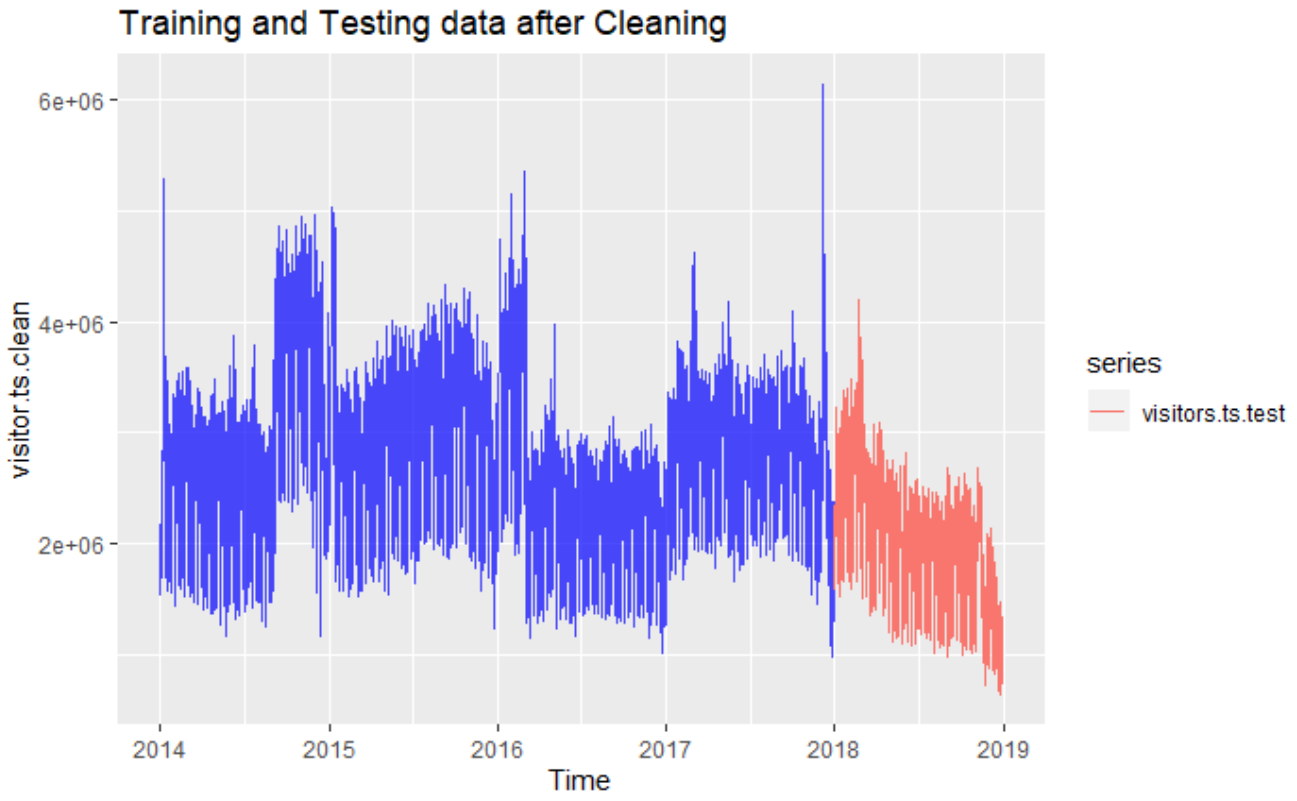
Conclusion

Our models only allow for regular seasonality, despite the fact that we have smoothened the data. Capturing seasonality associated with moving events such as Easter, Christmas, or the New Year is more difficult. If our time-series were relatively short which captured a single seasonality our fitted models would have worked better, as it could be seen from the forecasts of SARIMA model.

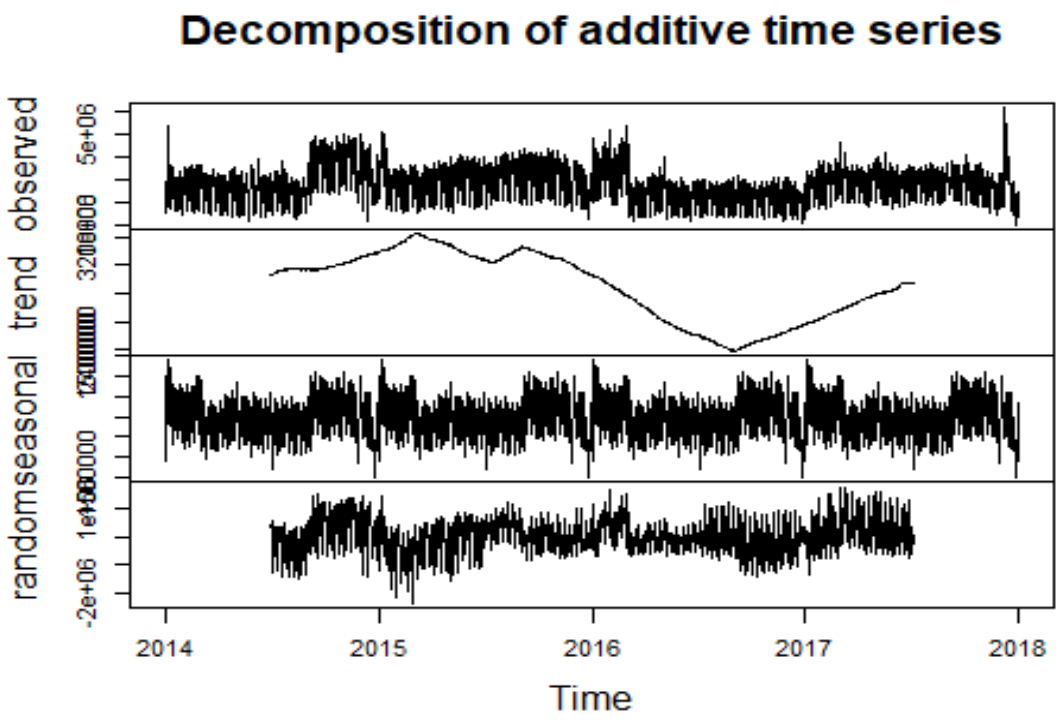
Appendix – References and Additional Plots

- Time Series Analysis and Its Applications With R Examples - by Robert. H. Shumway David S.
- Stoffer Forecasting: Principles and Practice - by Rob J Hyndman and George Athanasopoulos
- https://www.analyticsvidhya.com/blog/2018/09/non-stationary-time-series-python/
- https://machinelearningmastery.com/white-noise-time-series-python/
- https://towardsdatascience.com/the-complete-guide-to-time-series-analysis-and-forecasting-70d476bfe775

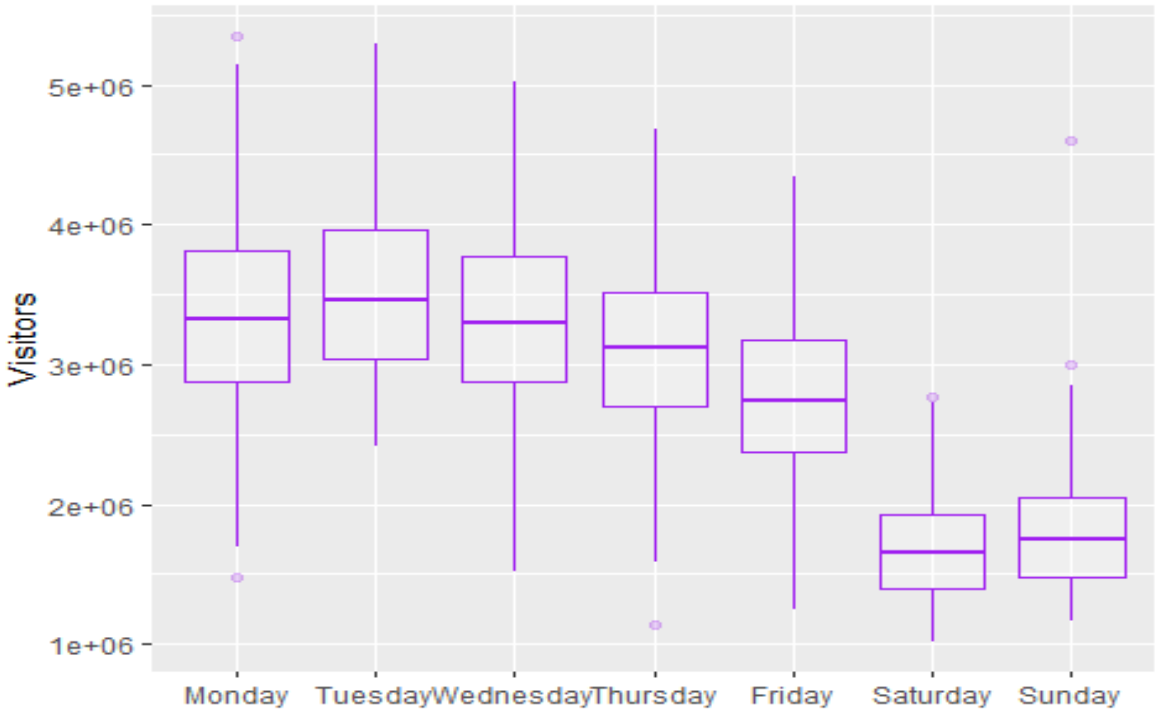
Plot -1



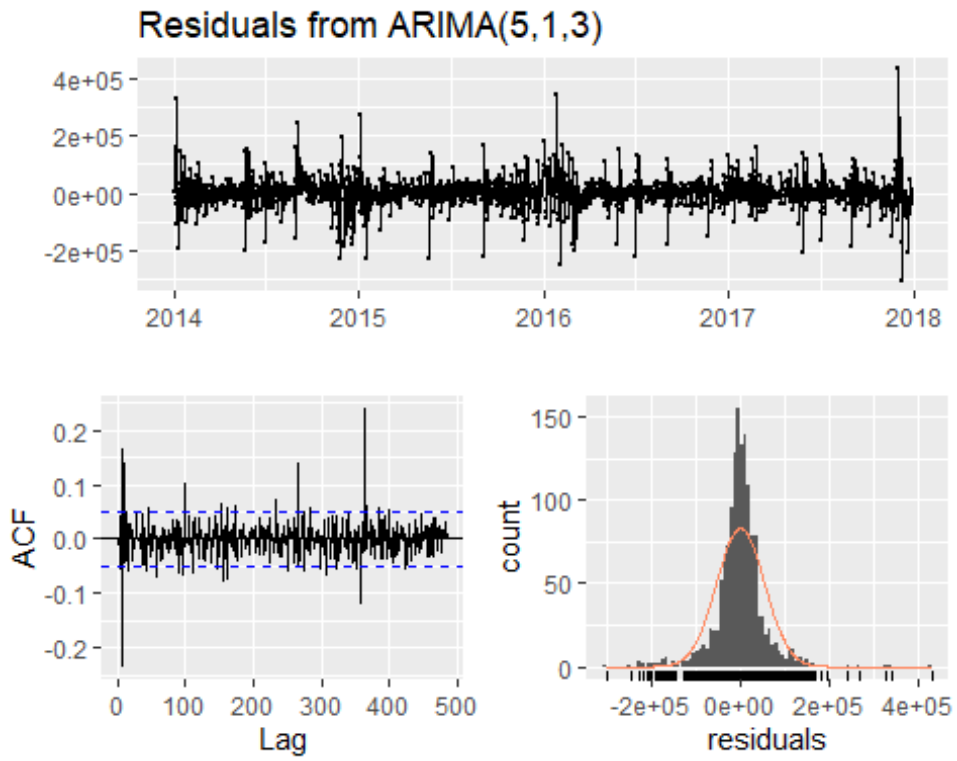
Plot -2


Decomposition of additive time series
observed
trend
seasonal
random
2014
2015
2016
2017
2018
Time

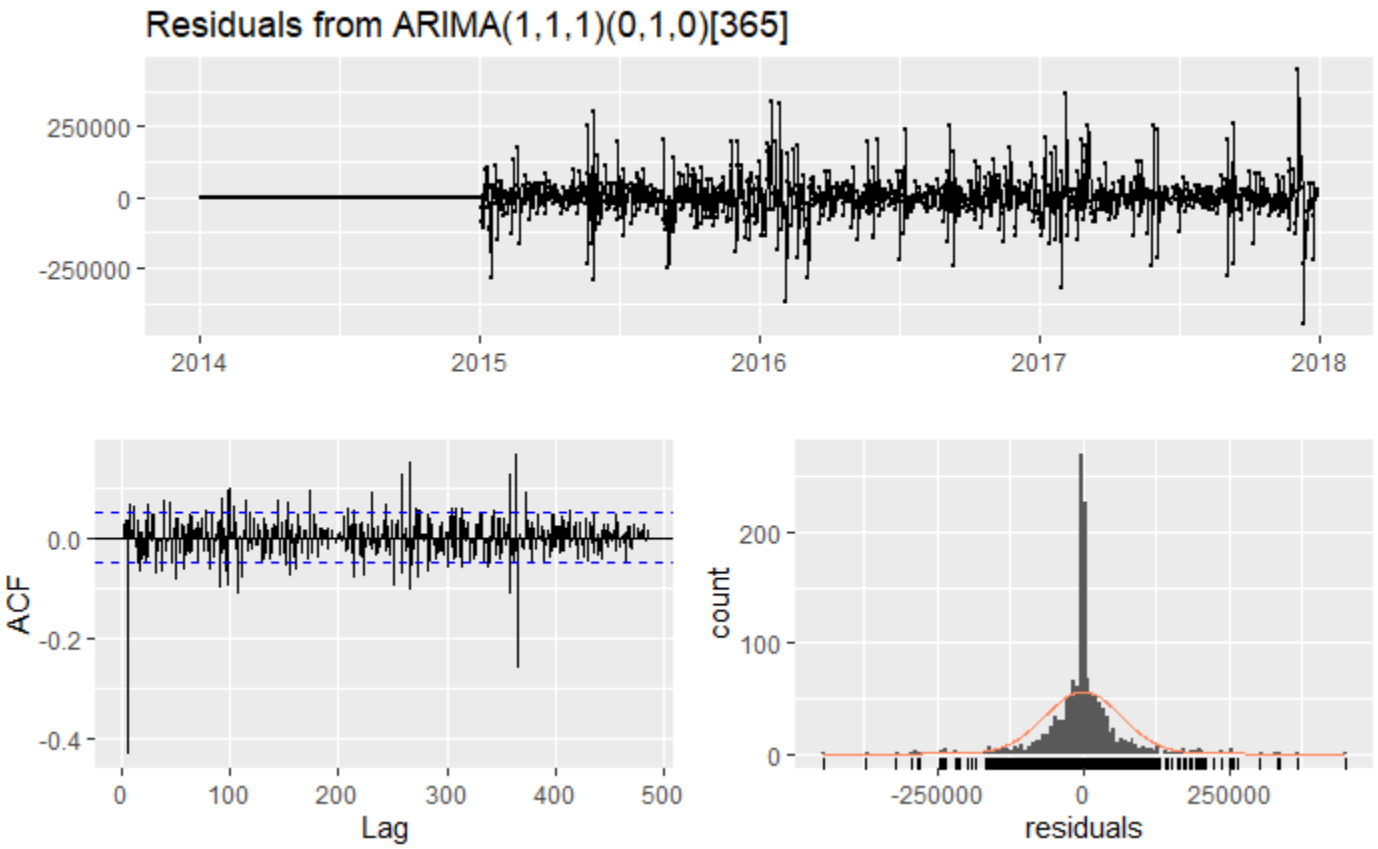
Plot -3


Visitors
5e+06
4e+06
3e+06
2e+06
1e+06
Monday
Tuesday
Wednesday
Thursday
Friday
Saturday
Sunday

Plot -4



Plot- 6 SARIMA residuals



Plot – 7 Holt – Winters Residuals