Web Traffic Time Series Forecasting

MSCA 31009

Chicago | December 7, 2020

Agenda

Introduction

Proposed Approach

Model Selection

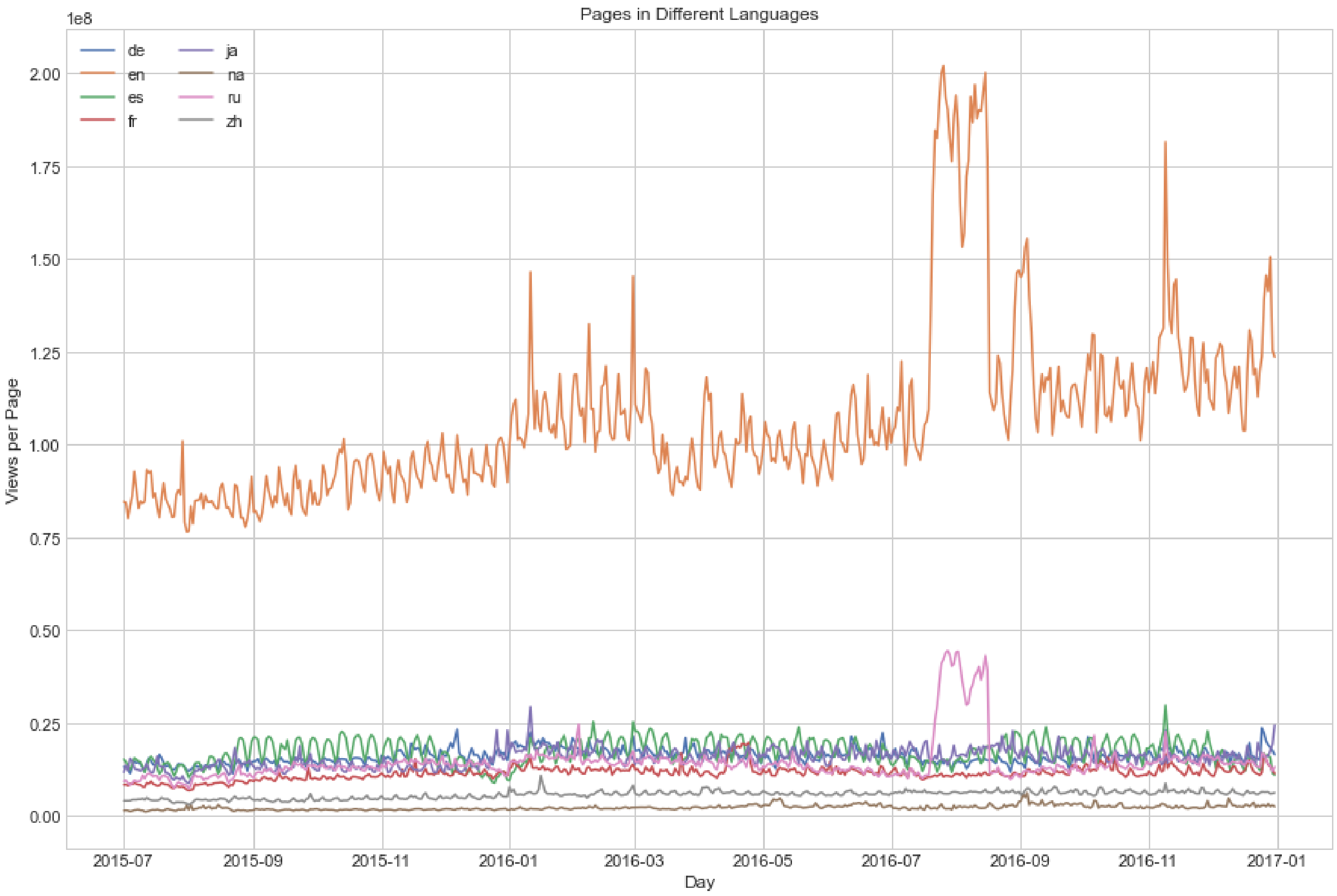
Intervention Analysis

Problem Statement - predicting web traffic

This forecasting can help website servers a great deal in effectively handling outages.

The technique we employed by product companies to better understand user behavior as to how a user interacts with their product and improve user experience.

The technique we implemented can be extended to diverse applications in financial markets, weather forecasts, audio and video processing. Not just that, understanding your website's traffic trajectory can open business opportunities too!

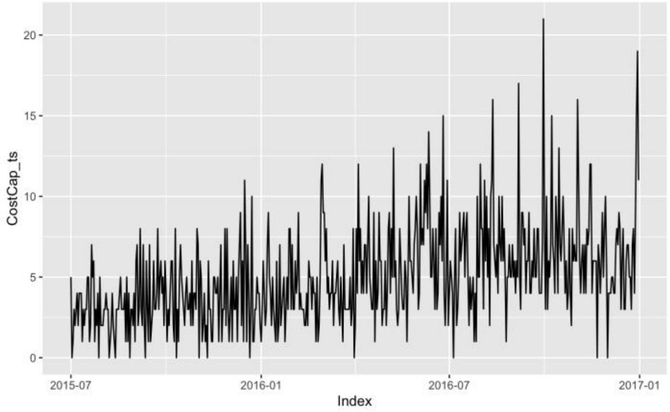


Training Data

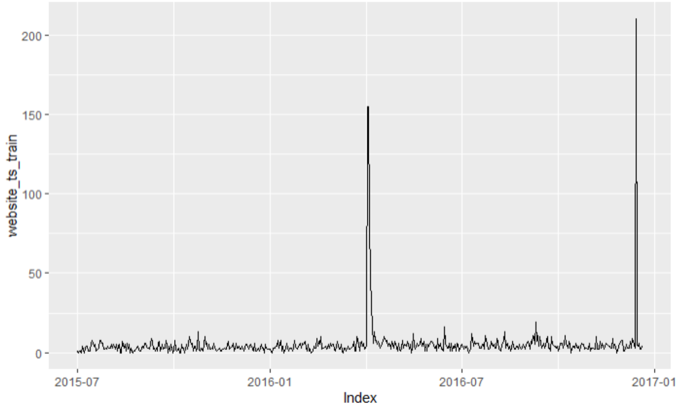
- *Number of Time series:* 145k Wikipedia Pages
- *Frequency* : Daily
- *Span:* **July 1st, 2015 - December 31st, 2016**
- *Languages*: 7

Data

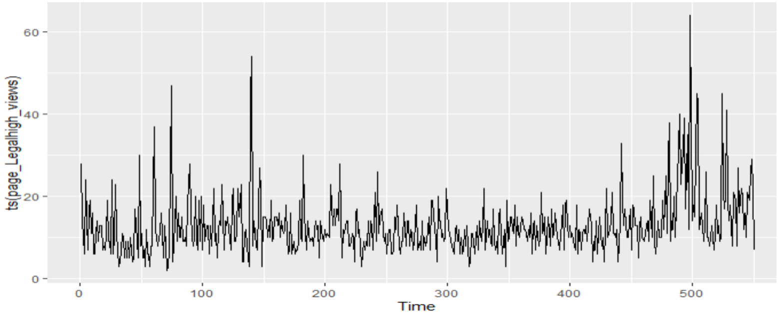
Weighted average cost of capital



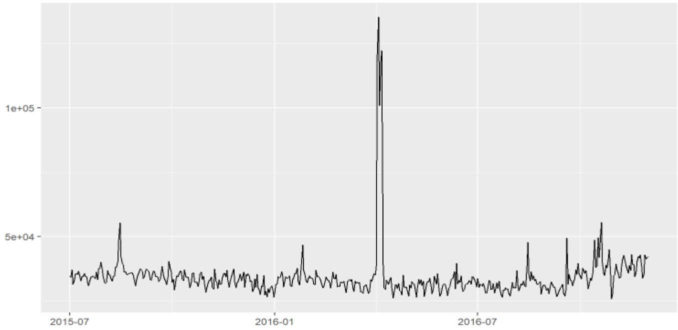
CYP2R1 in complex with vitamin D2



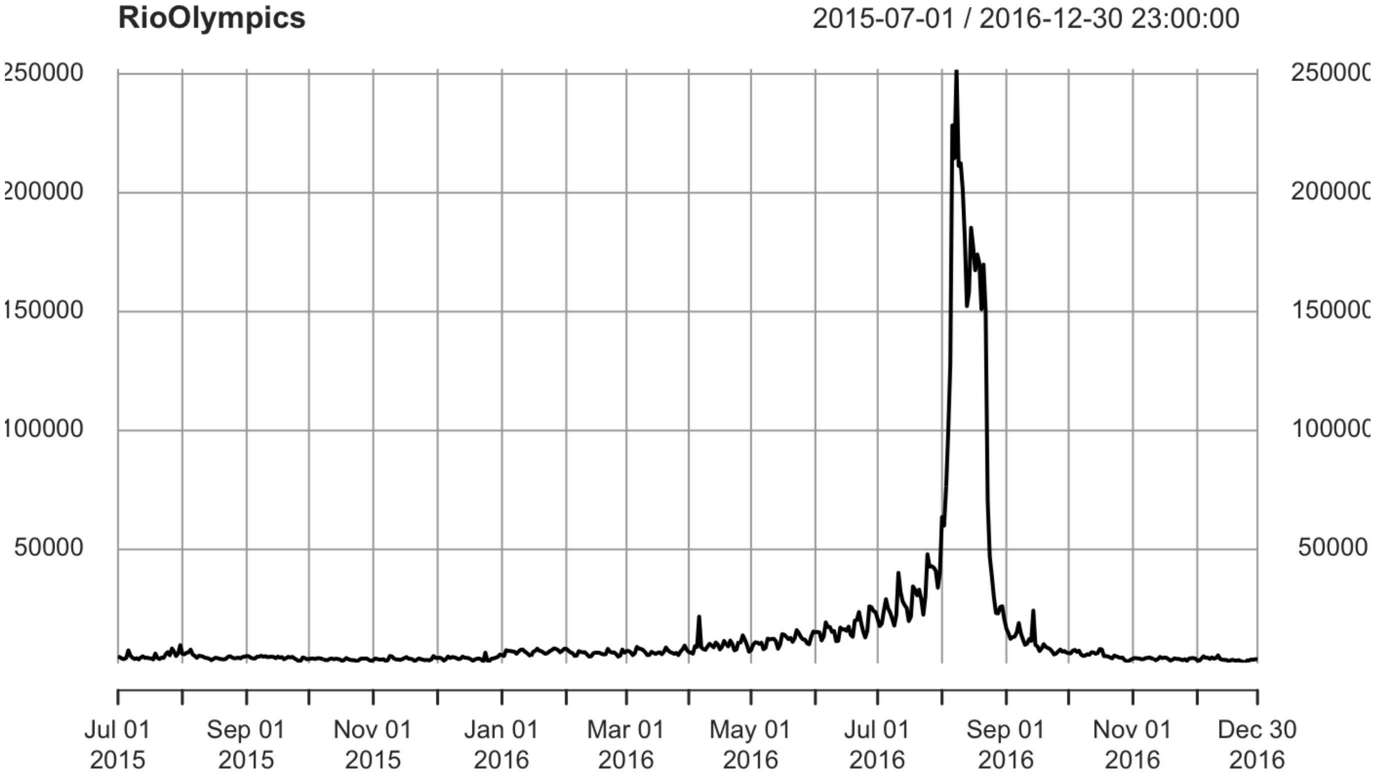
Legal High – Japanese TV Series



India



RIO Olympics



- Span: 5 to 21 August 2016
- Frequency : Daily
- Dec 2015: Human right violations in Rio
- April 2016: Torch Relay

Down Sampling the data from daily to weekly

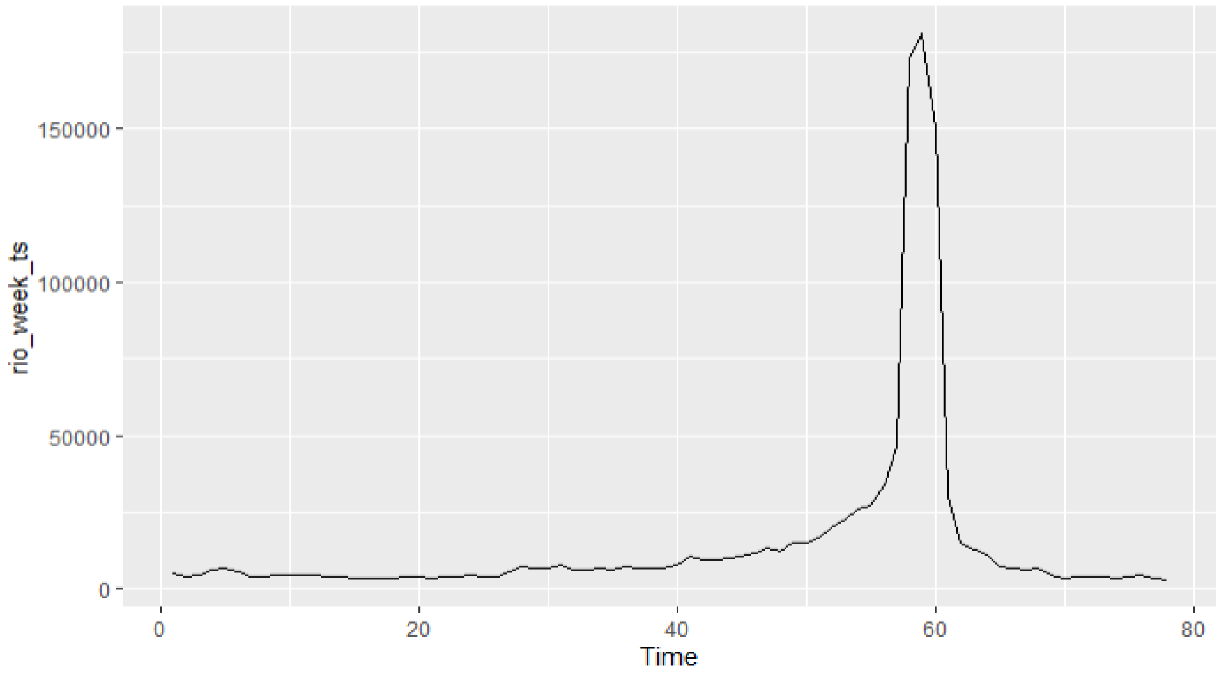
Code

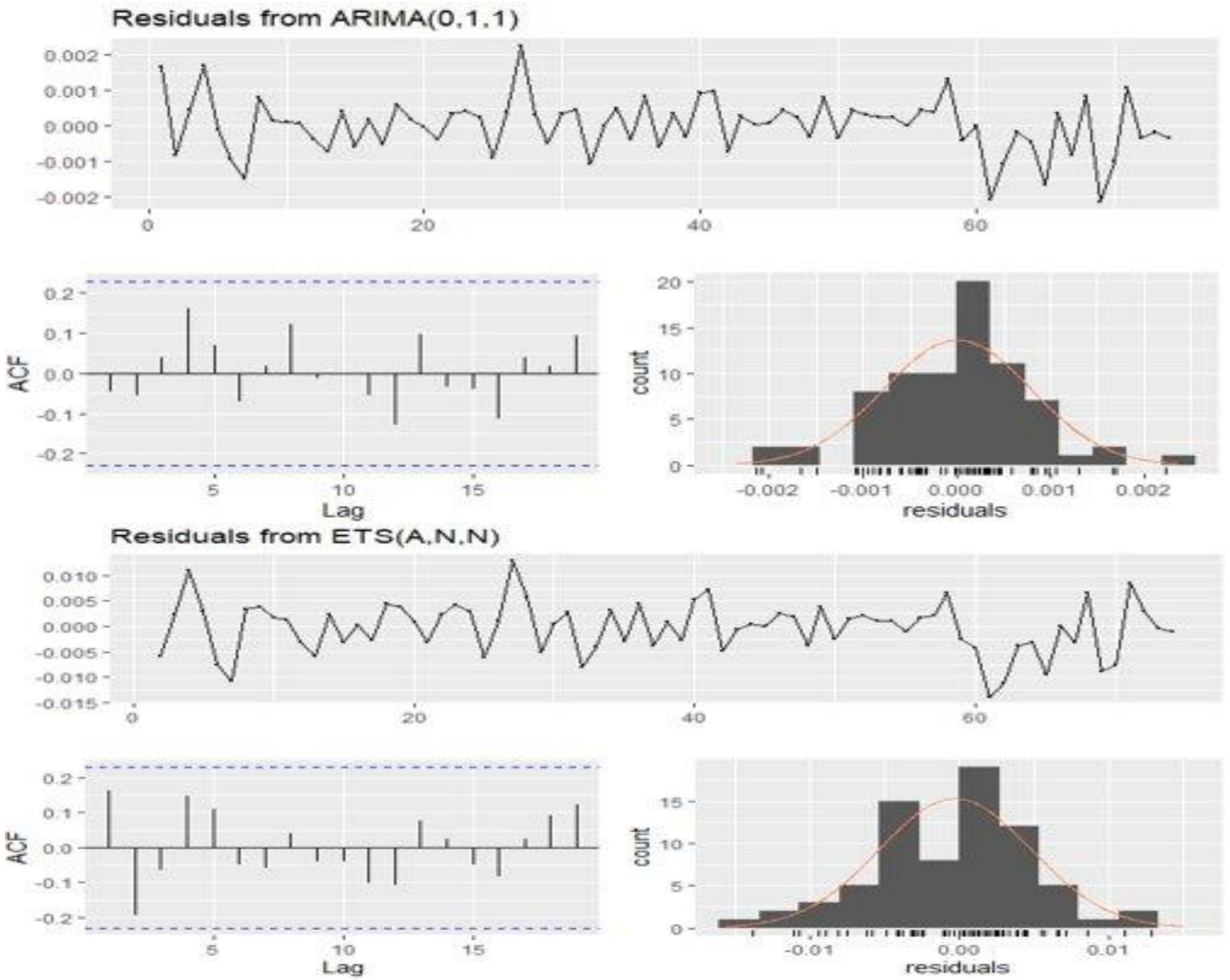
```{r}
rio_week = rollapply(rio_ts[,1], 7, mean, by=7)
```

```{r}
head(rio_week, 14)
```

```
                [,1]
2015-07-01        NA
2015-07-02        NA
2015-07-03        NA
2015-07-04        NA
2015-07-05        NA
2015-07-06        NA
2015-07-07  4944.143
2015-07-08        NA
2015-07-09        NA
2015-07-10        NA
2015-07-11        NA
2015-07-12        NA
2015-07-13        NA
2015-07-14  4308.714
```

Plot





Model Results

Preprocessing Techniques

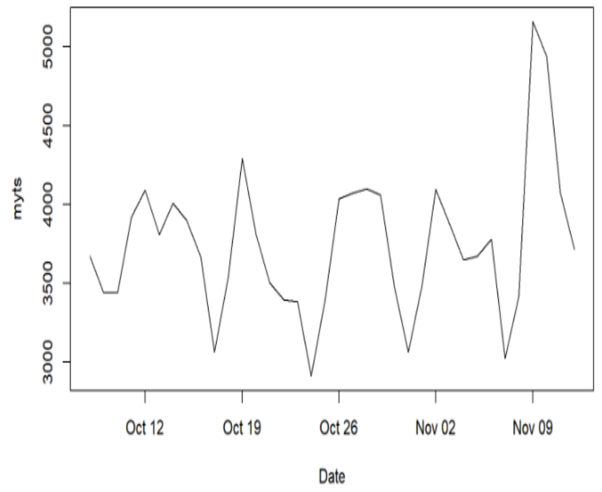
- Box-Cox Transformation
- Differencing

Models Applied

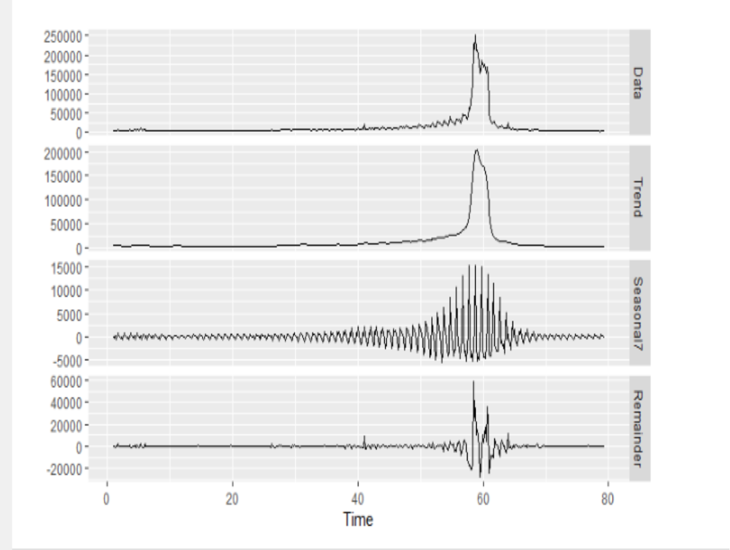
- **ARIMA:** Auto Regressive Integrated Moving Average' explains' a given time series based on its own past values
- **ETS,** (Error, Trend, Seasonal) used to handle the combination of trend, dam ping and seasonality.

TS Models with Weekly Time Period

Actual Data

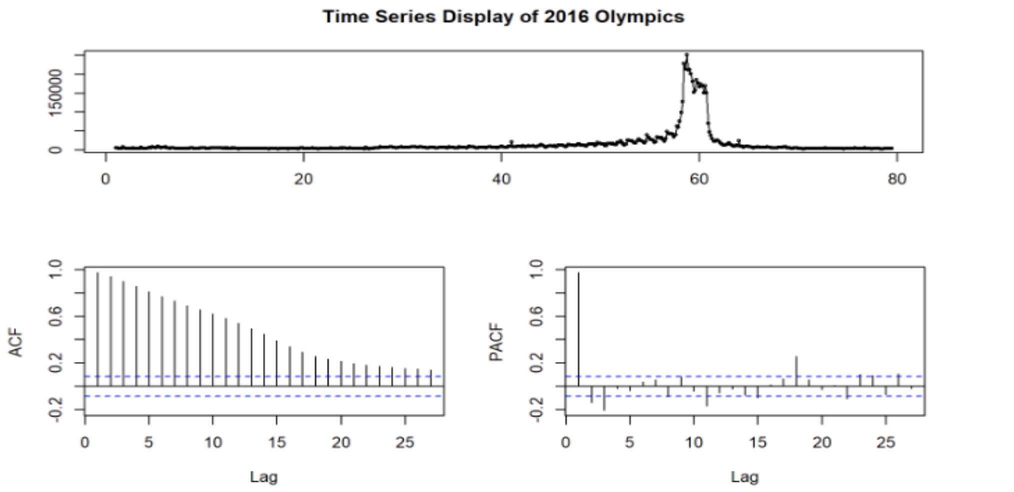


Complex Seasonality

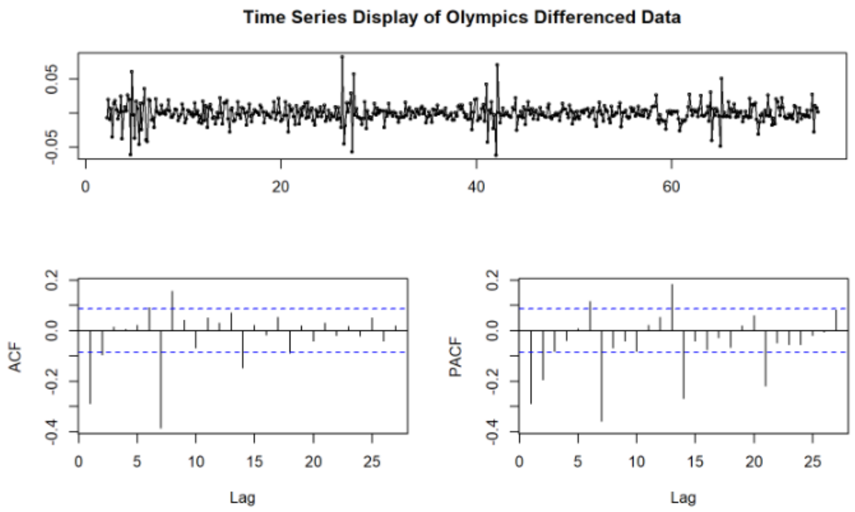


TS Models with Weekly Time Period

Actual Data

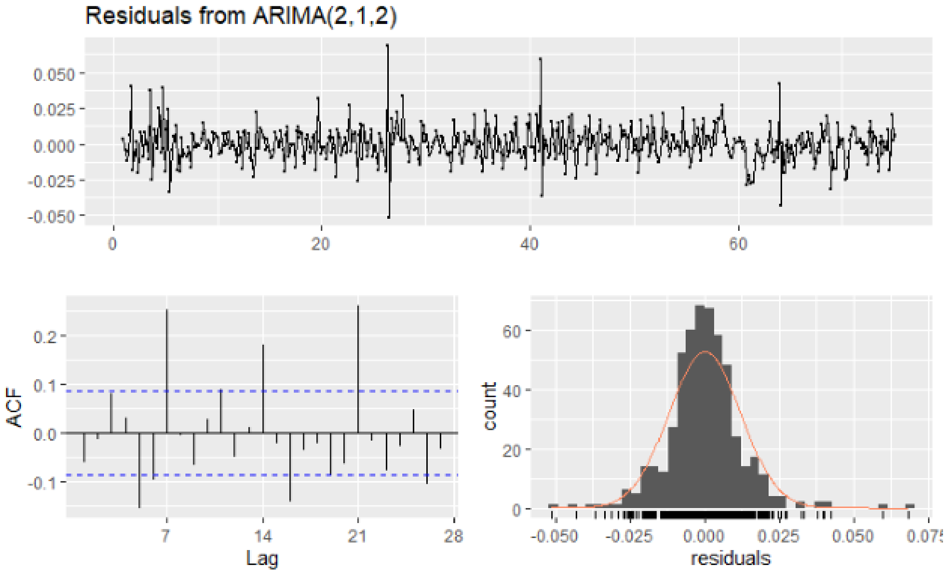


Differenced Data

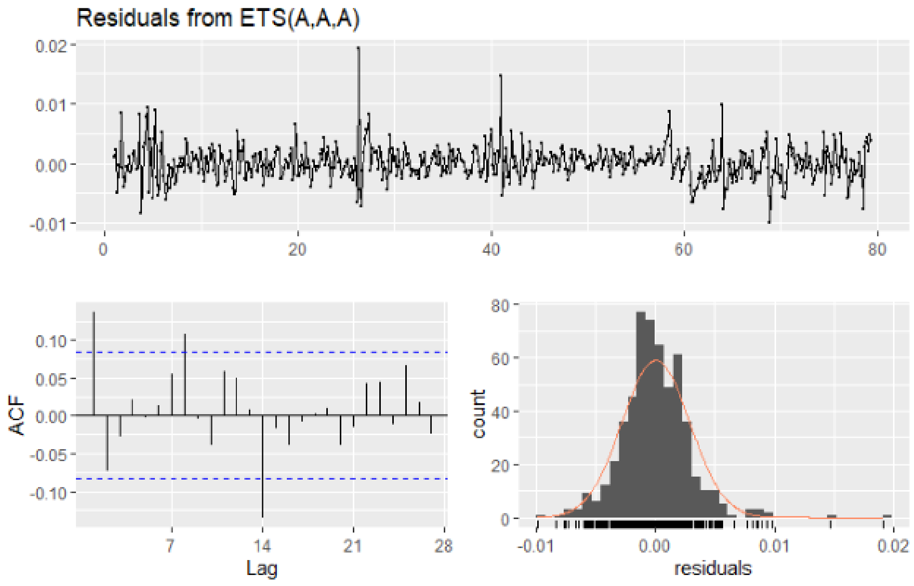


TS Models with Weekly Time Period

Non-Seasonal ARIMA model

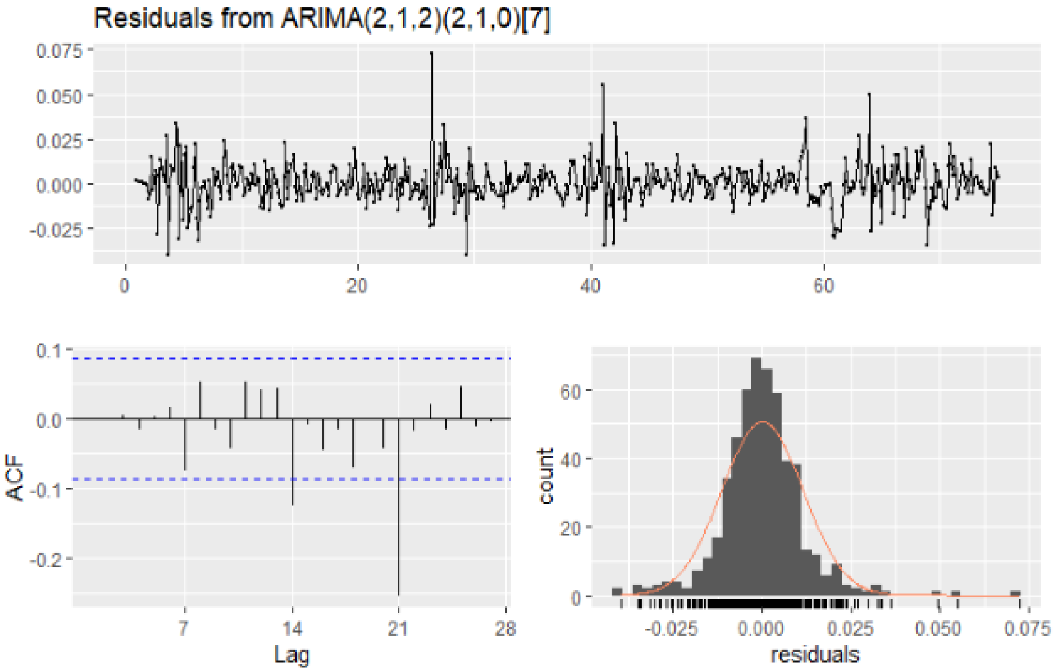


ETS model

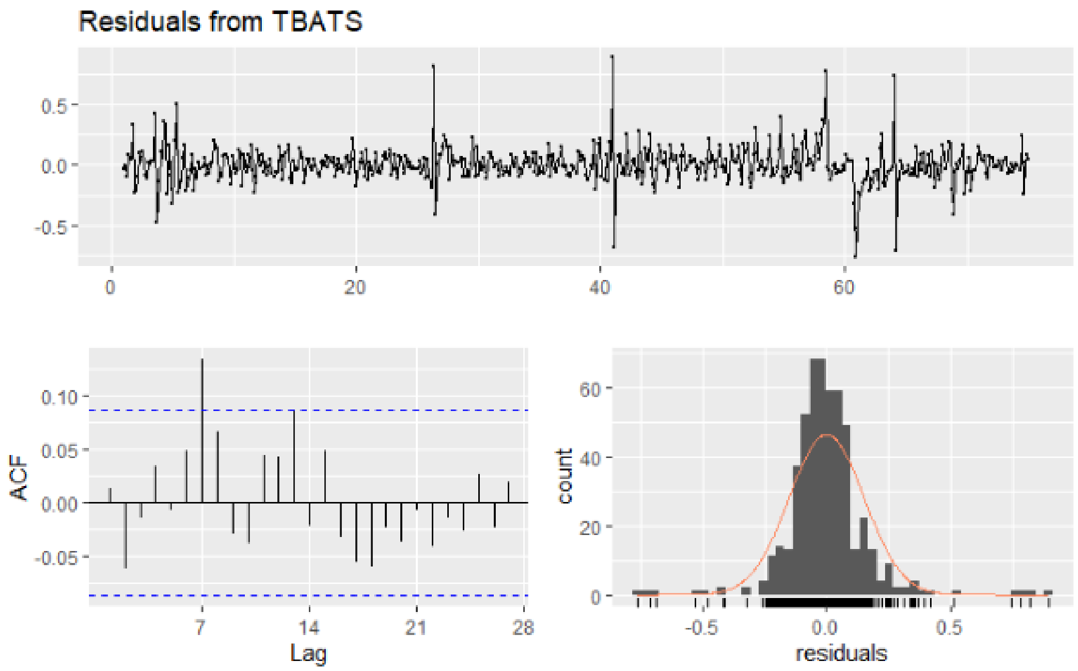


TS Models with Weekly Time Period

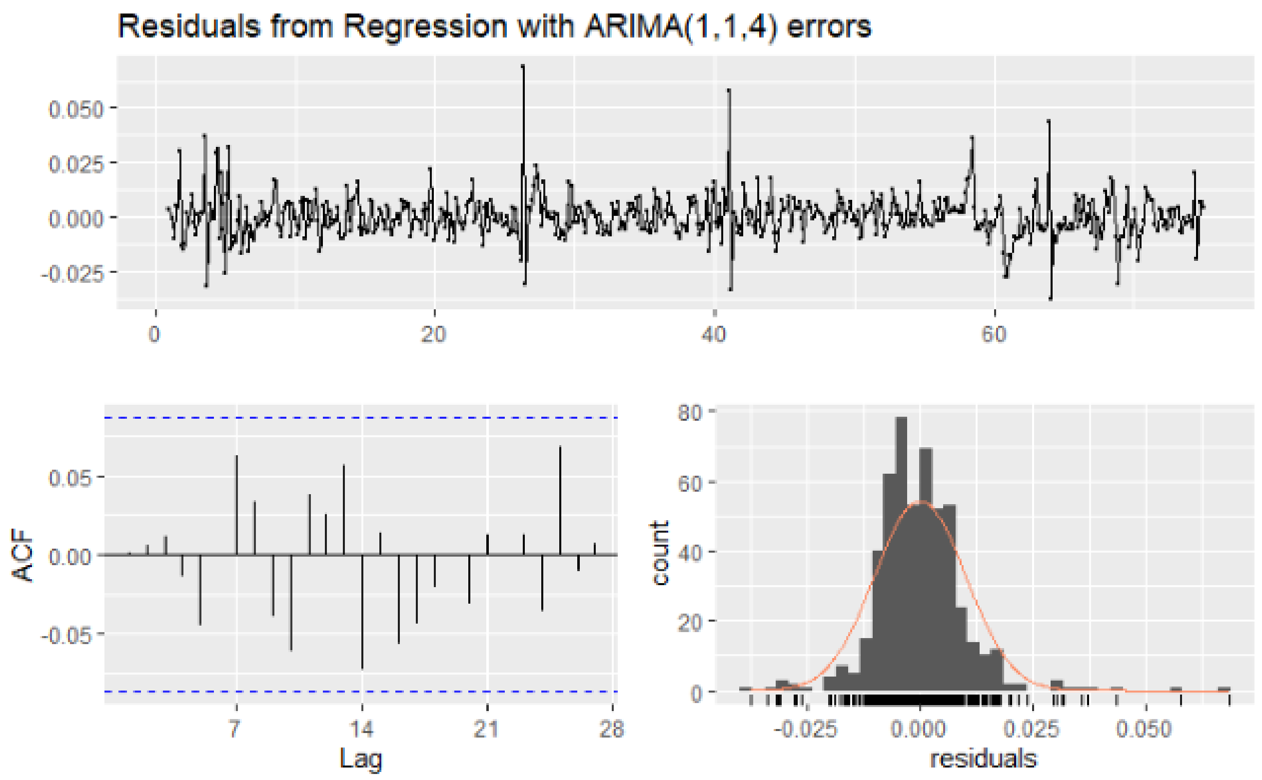
Seasonal ARIMA model



TBATS(0, {0,0}, -, {<7,3>})



Dynamic Harmonic Regression



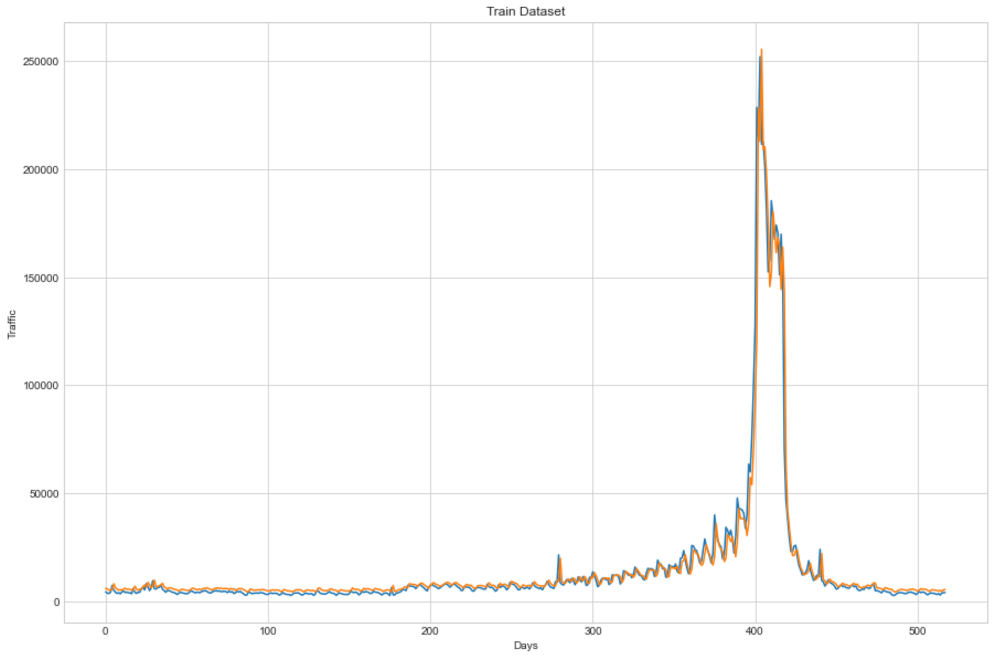
- Fourier Transform with K=3
- ARMA Error Residuals looks like white noise
- Outliers due to change in variance

LSTM

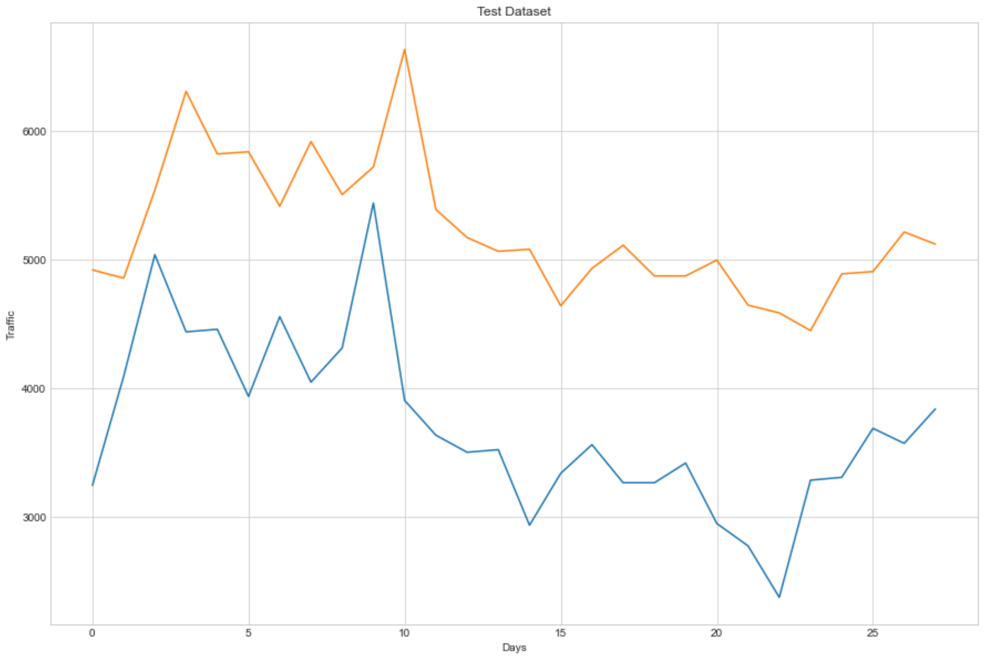
1.
   - Min Max Scaling
2.
   - Test Train Split
3.
   - Restructure the data
     - *Lookback window*
     - *Data Format: samples, **time steps, features***
4.
   - LSTM Model
5.
   - Invert the predictions

LSTM

Train



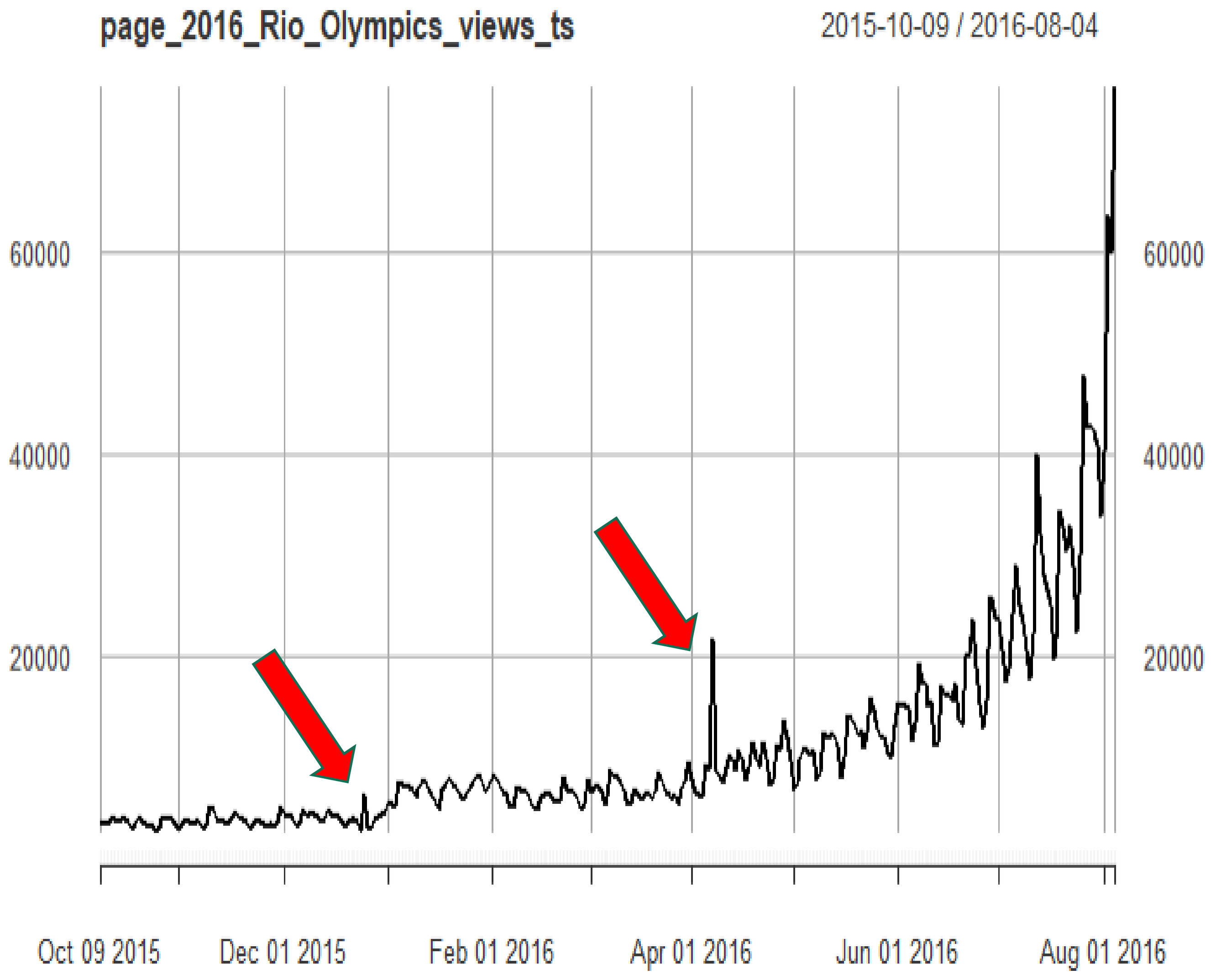
Test



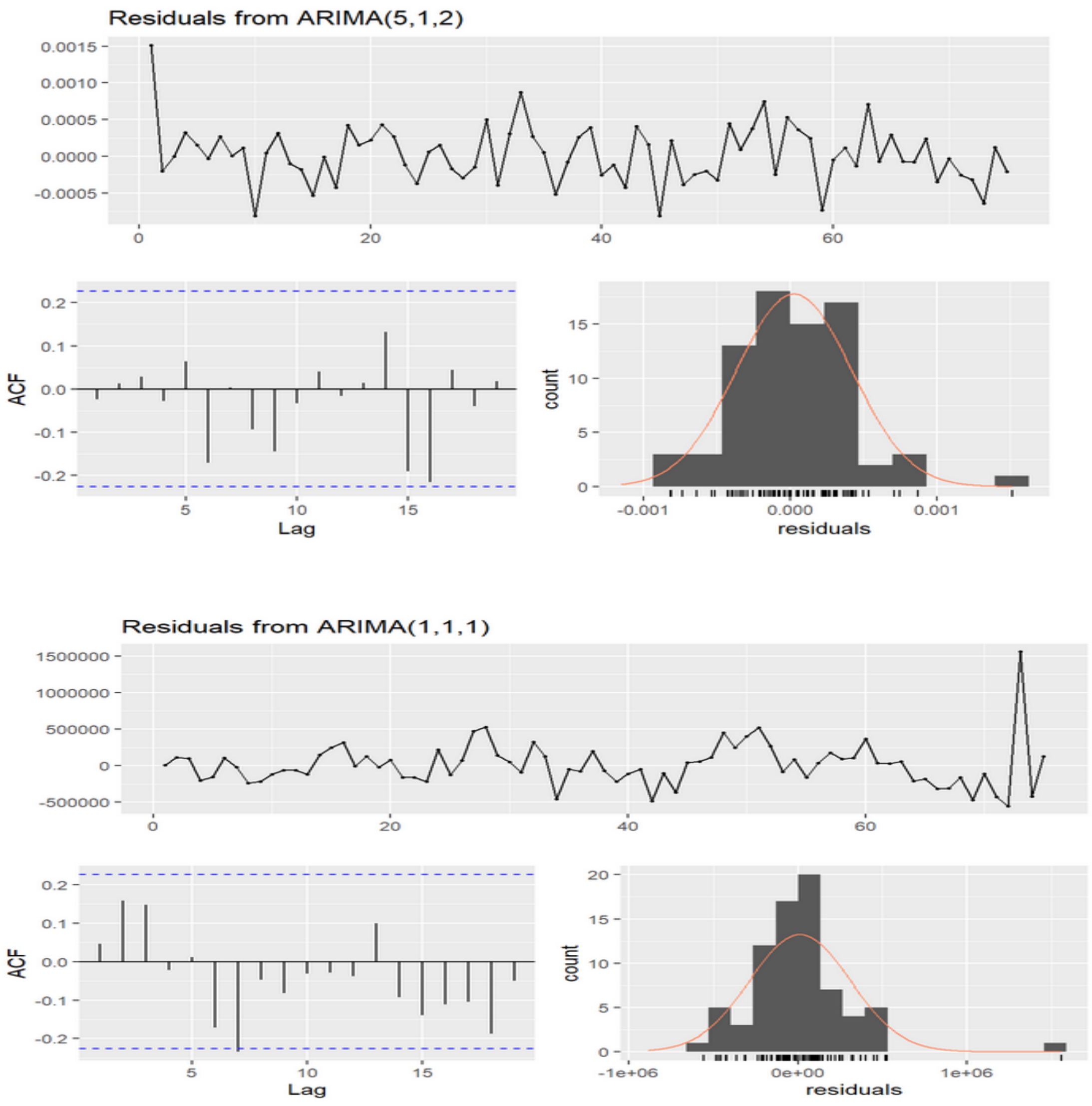
Model Comparison

Approach	Down Sampling ARIMA	Regression with ARIMA Error	LSTM
Train RMSE	30228.8853	14911.731	7756.18
Test RMSE	528.3937	620.672	1610.41
Pros	Smooth variance	Periodic seasonality	Lower generalization gap
Cons	Loss of Information	Capture complex seasonality	No long-term dependency in data

Intervention Analysis



- Taking a subset of observations
- Intervention on 23rd Dec 2015
- Dividing to pre- and post-intervention
- Trying to find the underlying process before intervention



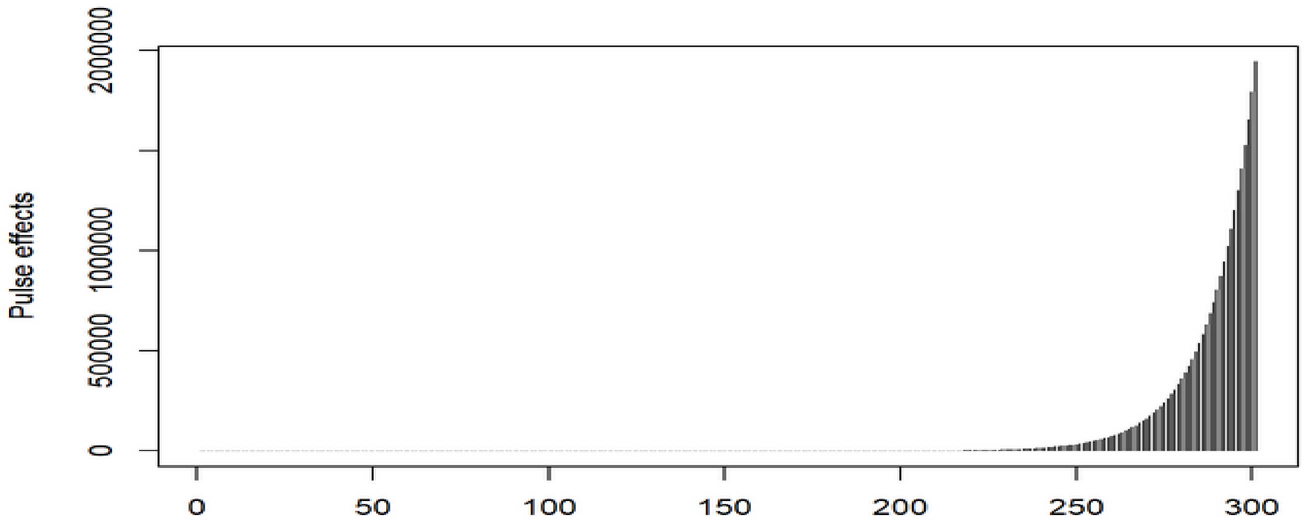
Intervention Analysis

Model fitted pre-intervention – ARIMA(5,1,2)

Residuals look good but model complex

Model fitted after down sampling – ARIMA (1,1,1)

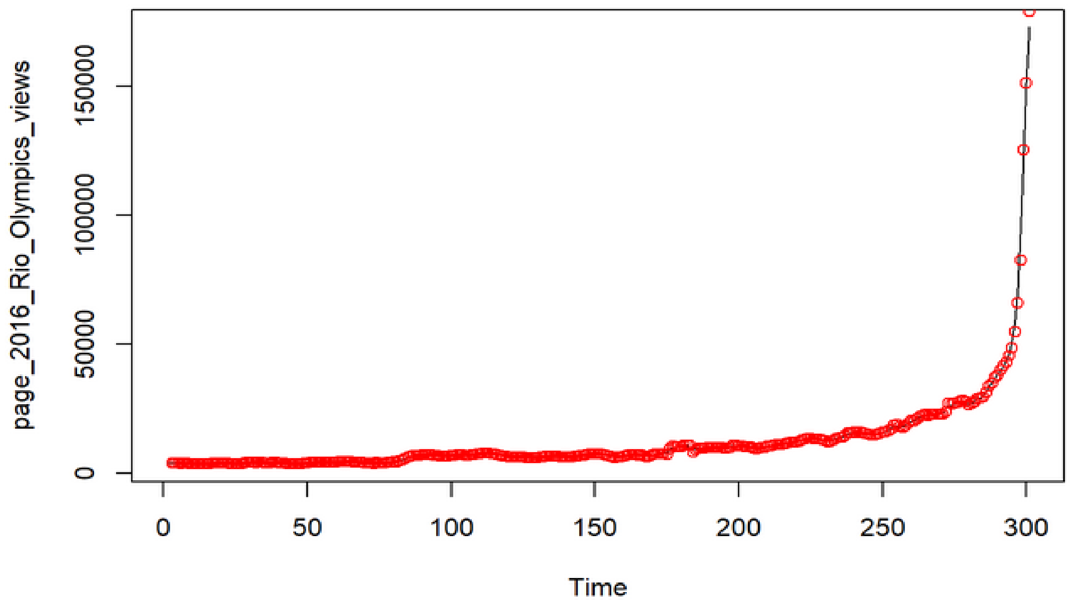
Intervention Analysis




```
##
## Call:
## arimax(x = log(page_2016_Rio_Olympics_views), order = c(1, 1, 1), xreg = 1 *
##     (seq(page_2016_Rio_Olympics_views) == 175), method = "ML", xtransf = data.frame(IHRV),
##     transfer = list(c(2, 2)))
##
## Coefficients:
##          ar1      ma1     xreg  IHRV-AR1  IHRV-AR2  IHRV-MA0  IHRV-MA1
##       0.8637  -0.2707  -0.0927    1.0836   -0.1840   -0.0425   -0.0210
## s.e.  0.0456   0.0810   0.0148    0.8783    0.7929    0.0268    0.0394
##       IHRV-MA2
##        -0.0256
## s.e.    0.0304
##
## sigma^2 estimated as 0.00082:  log likelihood = 635.53,  aic = -1255.05
```

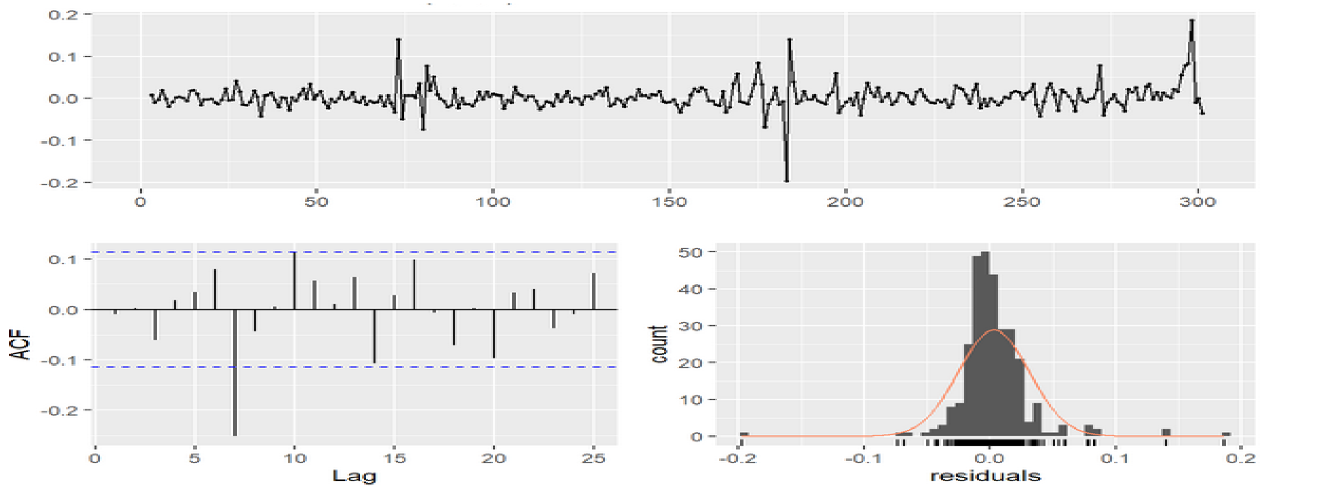
Assumed that intervention was ARMA(2,2) process

ARIMAX model to capture underlying process and intervention



Intervention Analysis

- Model doing for fitted values and actual values
- Residuals not white noise
- Model needs improvement



Future Work

- Better estimation of transfer function
- Requirement of cyclical data: Last 8 years
- Additional covariance

Meet the Team

Aakash Pahuja

Devanshi Verma

Prafulla Ranjan Dash

Surendiran Rangaraj

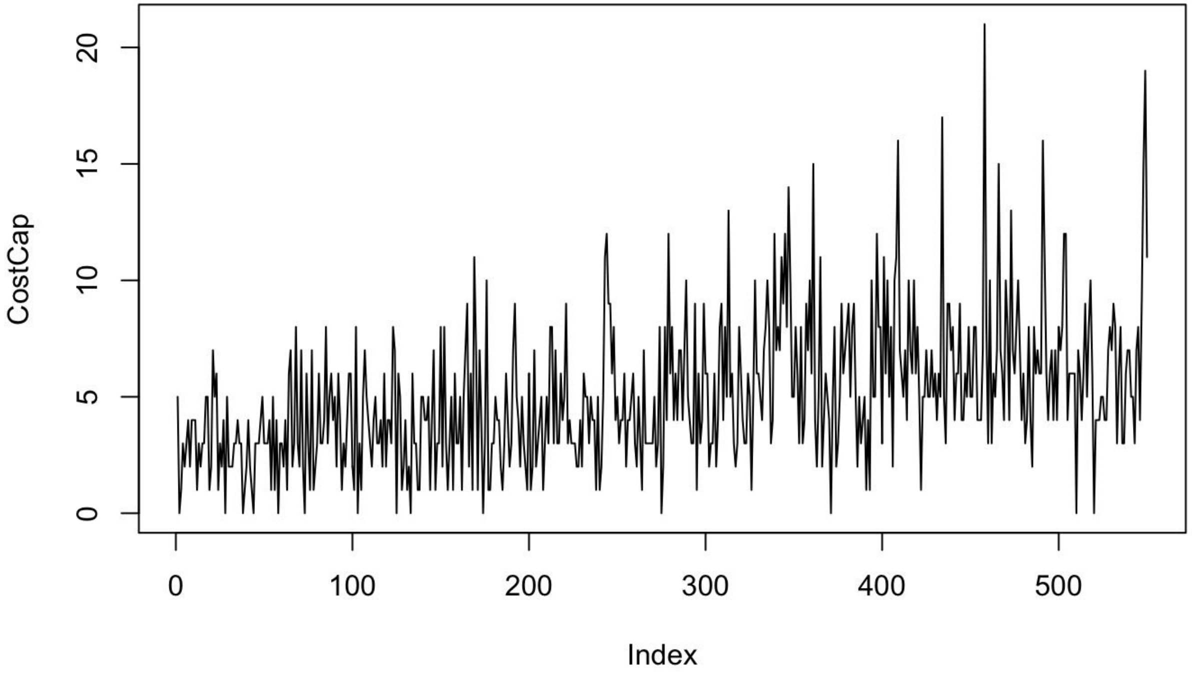
THE UNIVERSITY OF CHICAGO

Thank you!

Questions?

Appendix

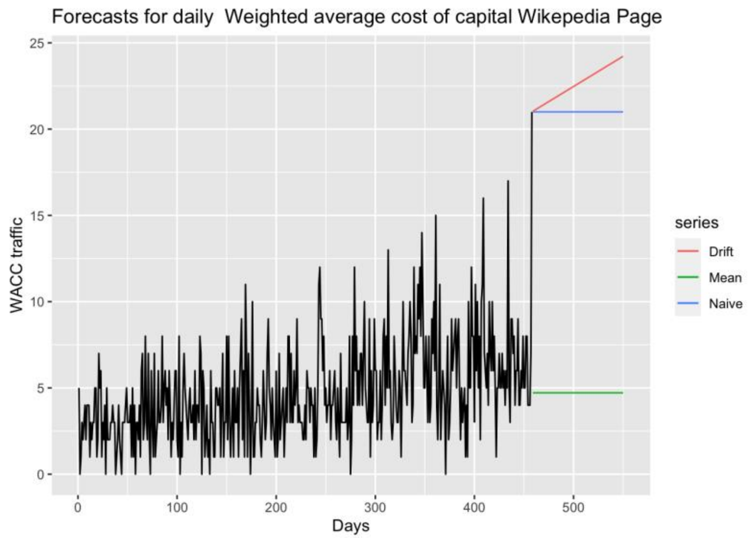
Data



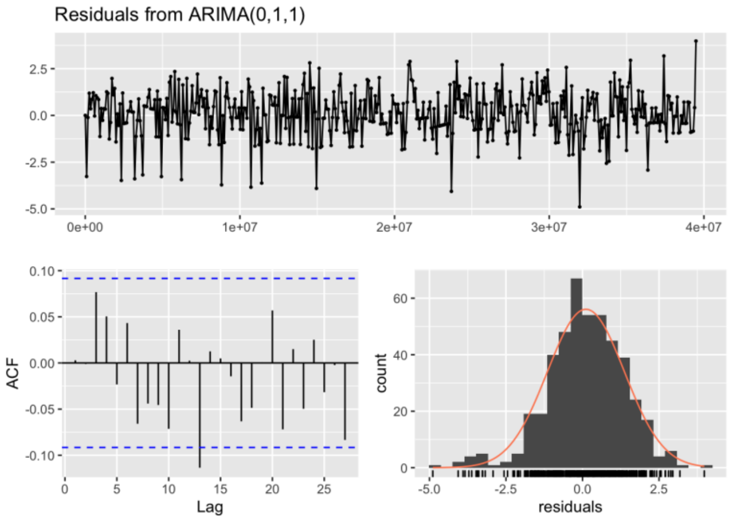
Before delving deep into the data, the web page let's know about its story. It's a Chinese Wikipedia page for Weighted average cost of capital which means measure of the cost of capital of a company. Because financing cost is seen as a logical price tag, it was used by many companies as the discount rate for a financing project in the past. It's not a much-researched topic and the range is usually in the range of 0-20 views per day but an extremely varying number.

Models and Results

Naïve Forecasts



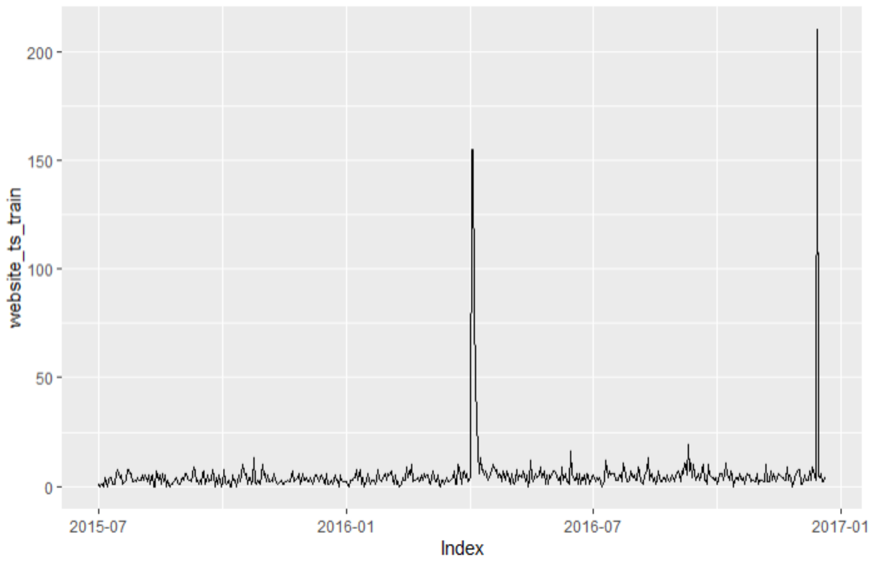
ARIMA



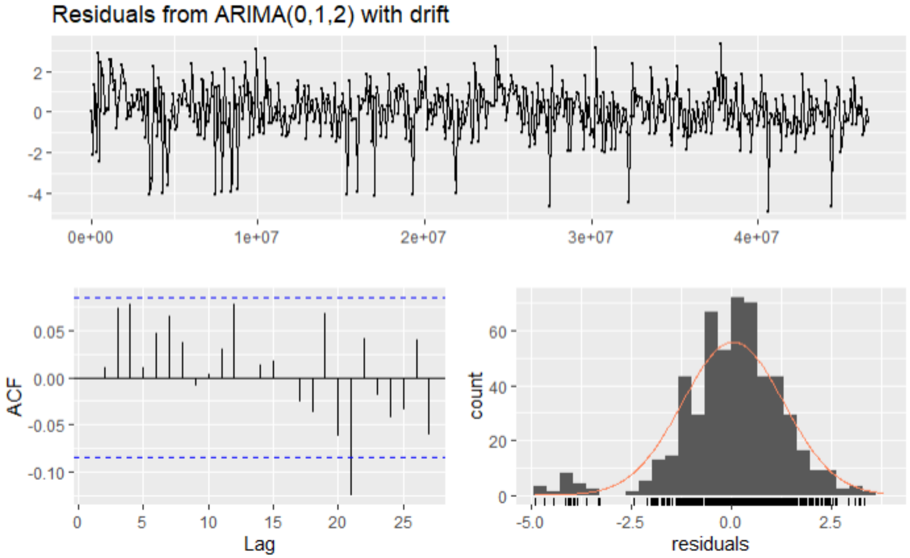
Approach	Mean	Naïve Model	ARIMA
Train RMSE	2.920003	3.540327	2.736433
Test RMSE	3.725011	14.720586	3.213281

3C_zh Models and Results

Unfiltered Time Series



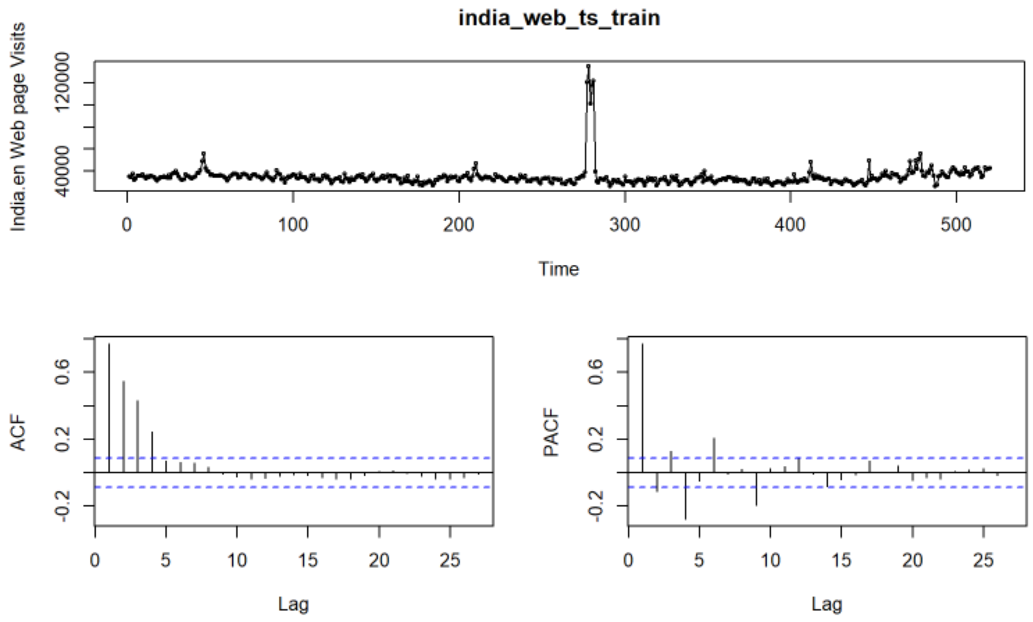
Residuals from ARIMA



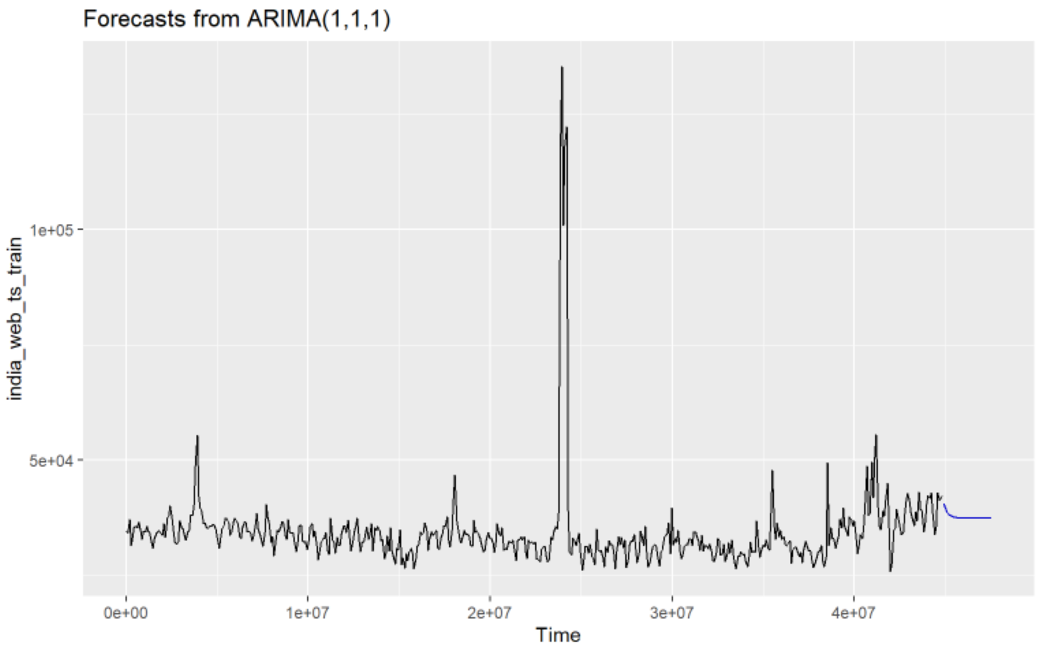
Approach	Mean	Naïve Model	ARIMA
Train RMSE	2.650704	1.994064	2.675221
Test RMSE	4.496604	2.601935	4.460078

India.en Models and Results

Time Series

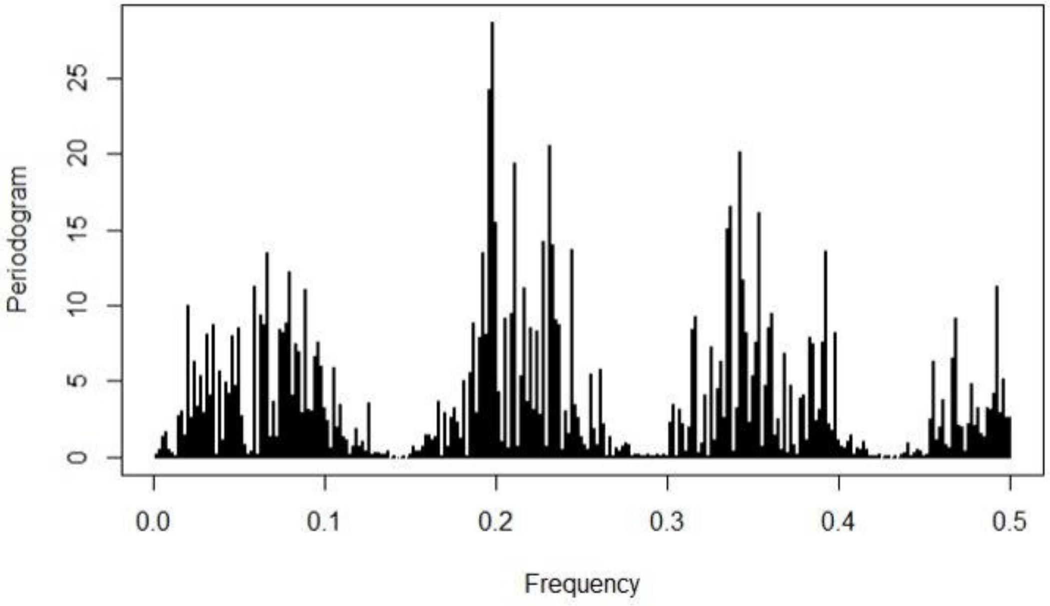


ARIMA

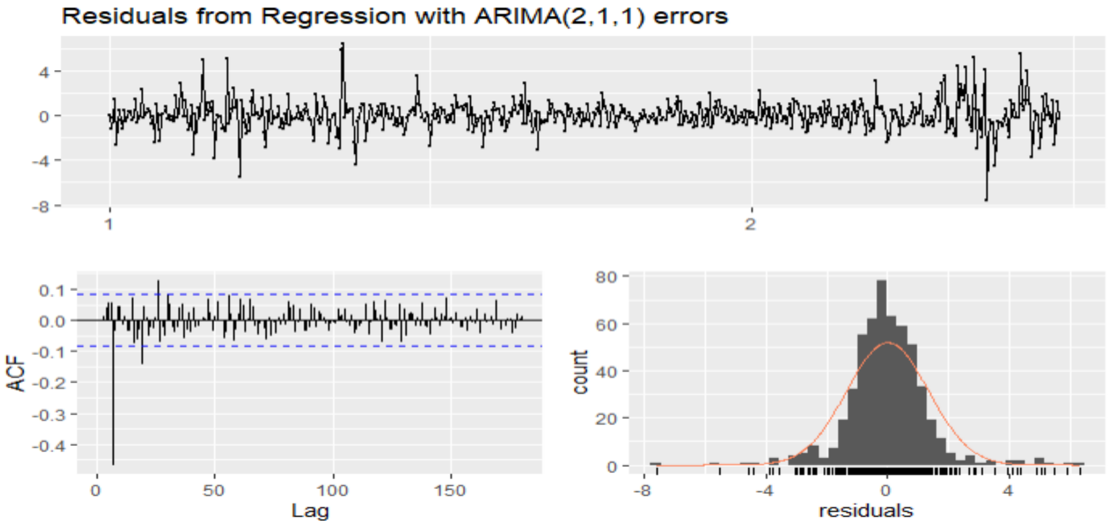


Legal High – Models and Results

Periodogram

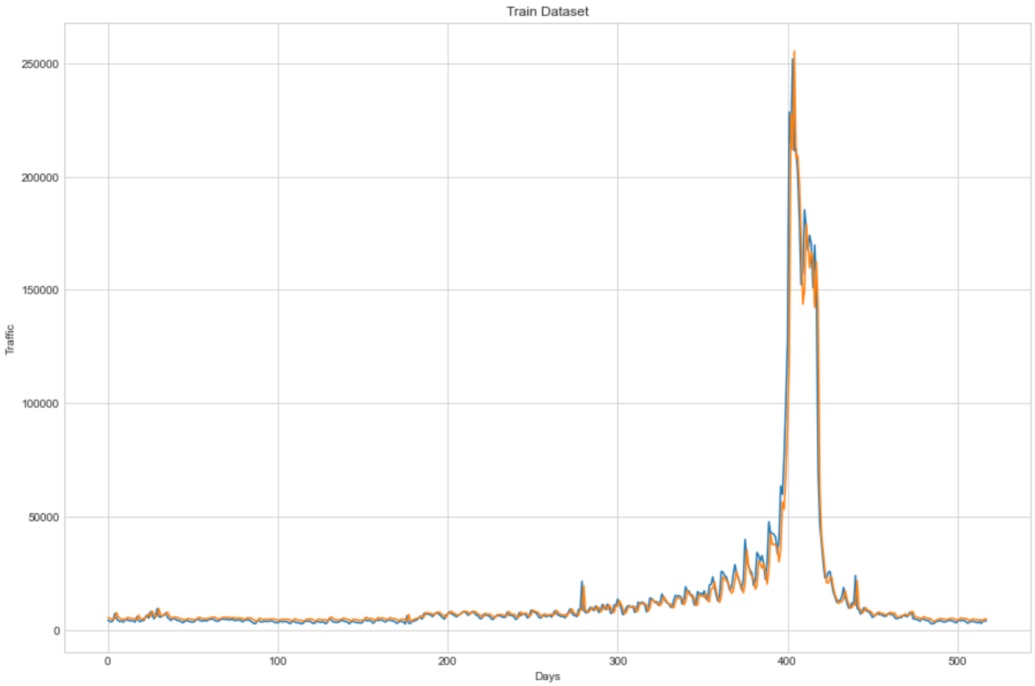


Residuals

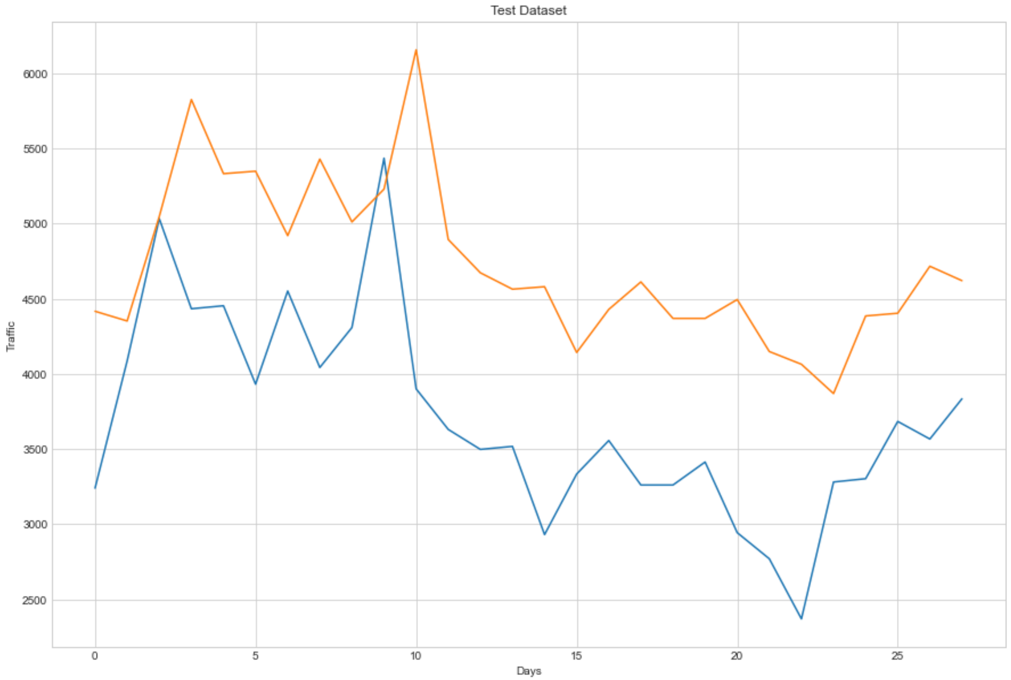


LSTM

Train



Test



Questions