5/10/2020

Forecast Future Traffic to Wikipedia Pages using Time Series Models

STAT 5414: Time Series Analysis I

Submitted By: Sonam Devadiga

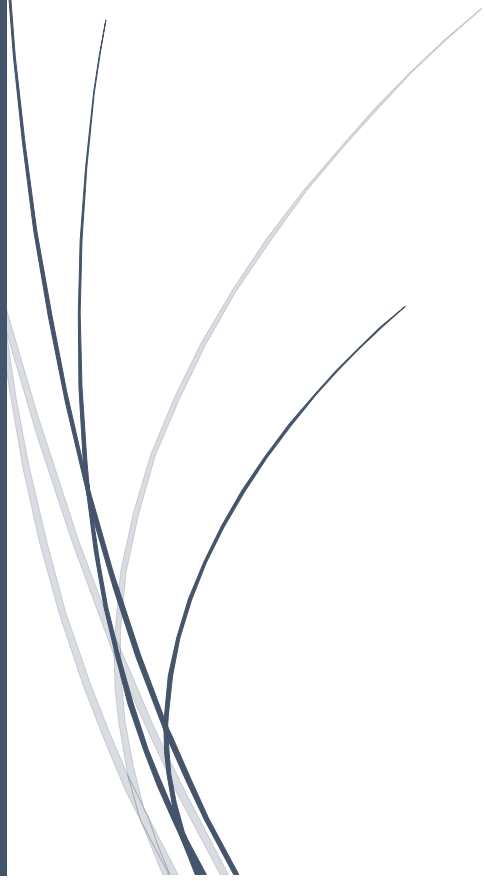

Table of Contents

1. Introduction:

Sequential or temporal observations emerge in many key real-world problems, ranging from biological data, financial markets, weather forecasting, to audio and video processing. The Web-traffic to Wikipedia pages also tend to follow these sequential or temporal patterns. The problem of forecasting the future values of multiple time series, has always been one of the most challenging problems in the field.

This project aims at selecting and analyzing select Wikipedia pages and using Box-Jenkins and State of the Art models to forecast future values . The project objective is to compare the traditional time series models as well as state of art models viz. DLM and Time varying regressive models to forecast the future web-traffic for select Wikipedia pages.

The web-traffic can be classified under two categories:

1. Topic-wise : Prediction forecast based on the different topics in each page.
2. Devices-wise : The web traffic is generated through use of Mobile Devices, Desktop Devices and Autobots. Thus, a further forecast will be made based on the type of device.

The web-traffic forecasting is useful in real-world to gauge and create budgets for storage or web server resources as per the time of the day.

2. Data Source:

The forecasting is done using the data obtained from Kaggle[1] on 'Web Traffic Time Series Forecasting'. The data consists of 804 data points of web-traffic recorded daily from 7/1/2015 to 9/10/2017.

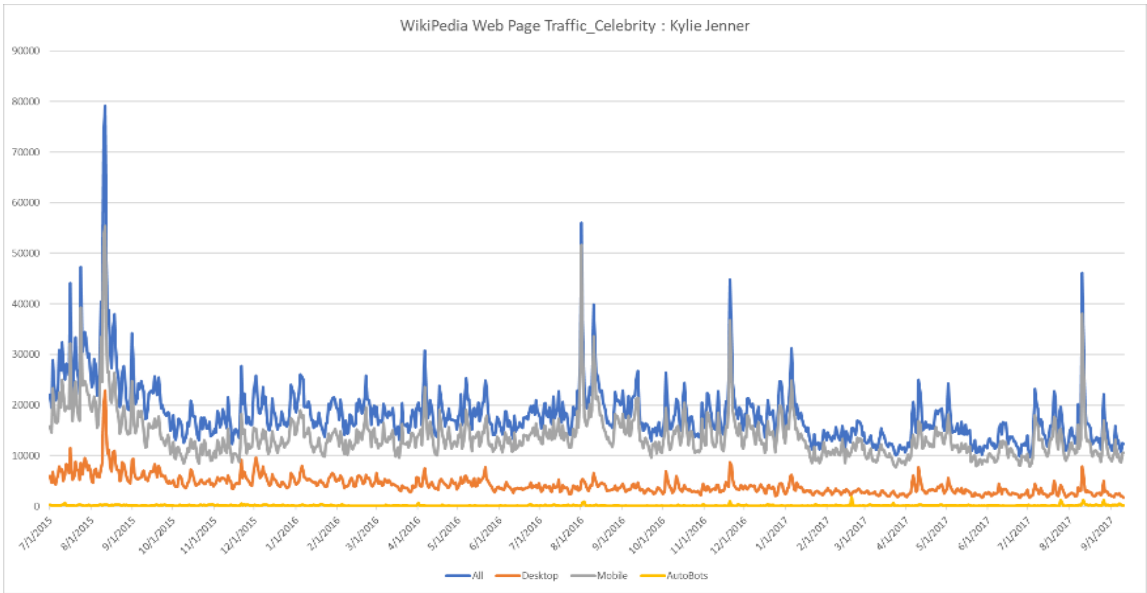

Fig. 1: Raw Data for Dataset 'Kylie Jenner'

3. Project Methodology:

The dataset Kylie Jenner will be compared against three of the following time-series models:

3.1. ARIMA (p,d,q) model :

ARIMA is a Box-Jenkins model. An auto regressive (AR(p)) component is referring to the use of past values in the regression equation for the series Y. The auto-regressive parameter p specifies the number of lags used in the model. The d represents the degree of differencing in the integrated (I(d)) component. A moving average (MA(q)) component represents the error of the model as a combination of previous error terms e_t. The order q determines the number of terms to include in the model Differencing, autoregressive, and moving average components.

The ARIMA model is represented by the following equation

$$Y_t = c + \phi_1 y_{d\ t-1} + \phi_p y_{d\ t-p} + \ldots + \theta_1 e_{t-1} + \theta_q e_{t-q} + e_t$$

where,

y_d is Y differenced d times

ϕ1, ϕ2 are parameters for the model

c is a constant.

The model above assumes non-seasonal series

These models directly rely on past values, and therefore work best on long and stable series. ARIMA simply approximates historical patterns and therefore does not aim to explain the structure of the underlying data mechanism.

3.2. SARIMA (p,d,q)(P,D,Q) model:

ARIMA models can be also specified through a seasonal structure. In this case, the model is specified by two sets of order parameters: (p, d, q) as described for ARIMA and (P, D, Q)m parameters describing the seasonal component of m periods

$$(1-\phi_1 B)\ (1-\Phi_1 B^4)\ (1-B)\ (1-B^4) y_t \ = \ (1+\theta_1 B)\ (1+\Theta_1 B^4) e_t.$$

(Non-seasonal AR(1)) (Seasonal AR(1)) (Non-seasonal difference) (Seasonal difference) (Non-seasonal MA(1)) (Seasonal MA(1))

Thus a multiplicative seasonal ARMA (p,q)×(P,Q)s model with seasonal period s is a model with AR characteristic polynomial φ(x)Φ(x) and MA characteristic polynomial θ(x)Θ(x).

3.3. Dynamic Linear Model (DLM):

In theory, all of system noise, observation noise, state equation and observation equation could change over time. At any time, we have a current best estimate on where the state-space is (given past observations). We compute the next true position (state) based on the state equation (this is the prior). We predict the corresponding observation based on the observation equation (this is the likelihood). As soon as the new observation arrives, we can correct our estimate and compute the posterior on the state.

- state equation $\theta_t = G_t\theta_{t-1} + w_t$, where $w_t \sim \mathcal{N}(0, W_t)$
- observation equation $Y_t = F_t\theta_t + v_t$, where $v_t \sim \mathcal{N}(0, V_t)$
- prior $\theta_0 \sim \mathcal{N}_p(m_0, C_0)$

Kalman filter is used to correct our prediction by how much it differs from the incoming observation.

Posterior estimate $E(\theta_t|y_{1:t}) = m_t = a_t + R_t F_t' Q_t^{-1} e_t$

Kalman gain $R_t F_t' Q_t^{-1}$

4. Model Building:

The Kylie Jenner dataset consists of 800 points. The data set is plotted initially as shown in Fig. 2. This series has an observed mean of 18267.66 and an observed variance of 38217289. Thus, this dataset is not stationary and has seasonality. To make the time series stationary log transformation of the dataset is taken as shown in Fig 3. The observed mean is 9.771454 while the observed variance is 0.07450695 of the datasets. Thus, the data is still not stationary. To make the data stationary the first log differential is taken as shown in Fig 5. and then to remove seasonality the twelfth differential of the log differential function is taken as shown in Fig 6.

The ACF and Partial ACF functions for the log transformation functions are shown in Fig 4. The PACF plot shows around 11 peaks thus the AR term is around 11. Using this data the next ARIMA, SARIMA and DLM models are built. These models have different outcomes where the Box Jenkins models use AIC and BIC for model selection while DLM uses Fourier Transformation and Posterior Probability for estimations.

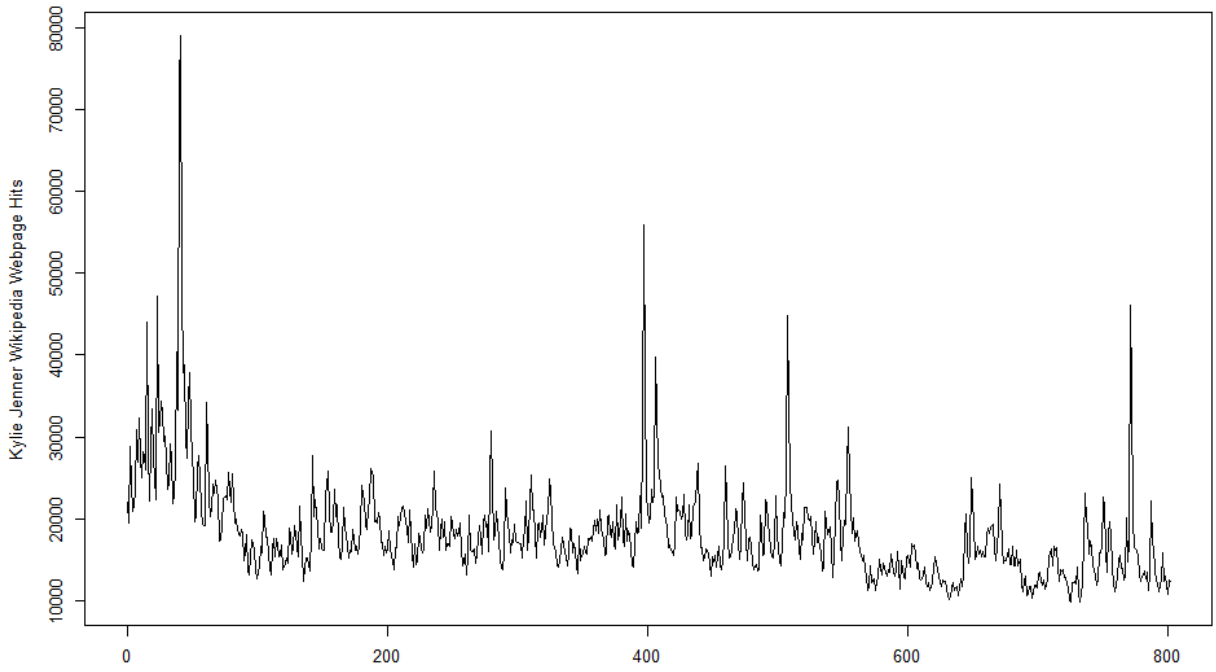

Fig. 2: Wikipedia Page Hits for 'Kylie Jenner'

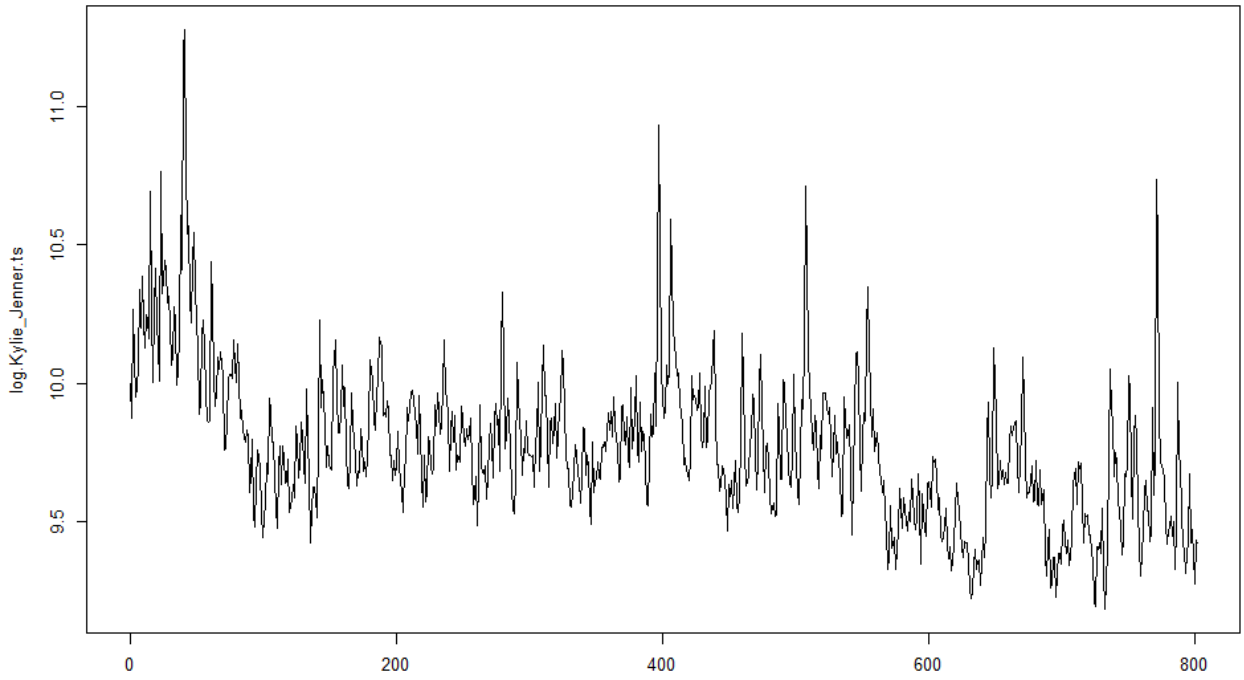

Fig. 3: Log Transformation of the dataset

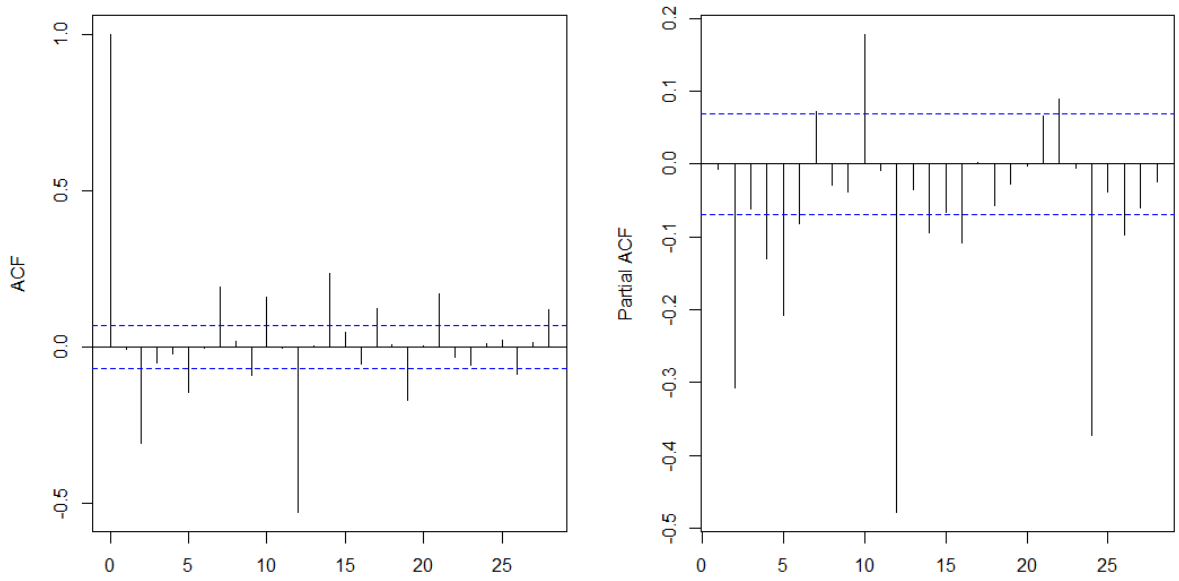

Fig. 4: ACF and PACF plots

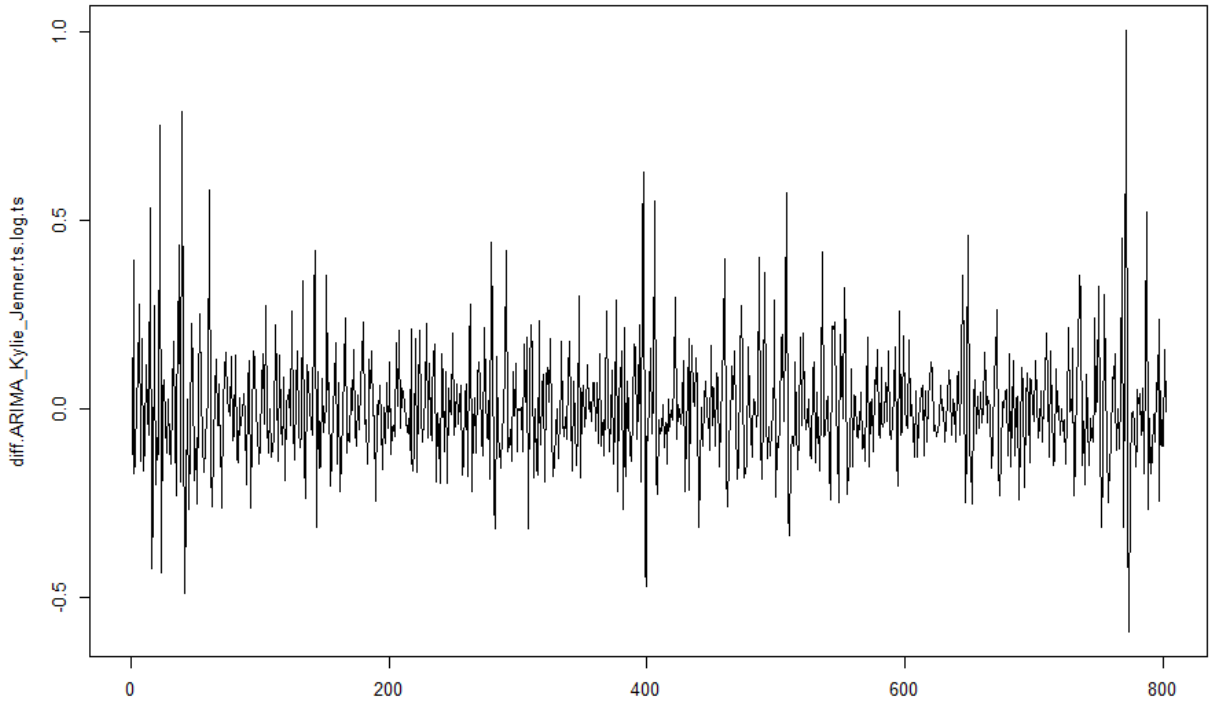

Fig. 5: 1st Differential of Log Transformation of the dataset

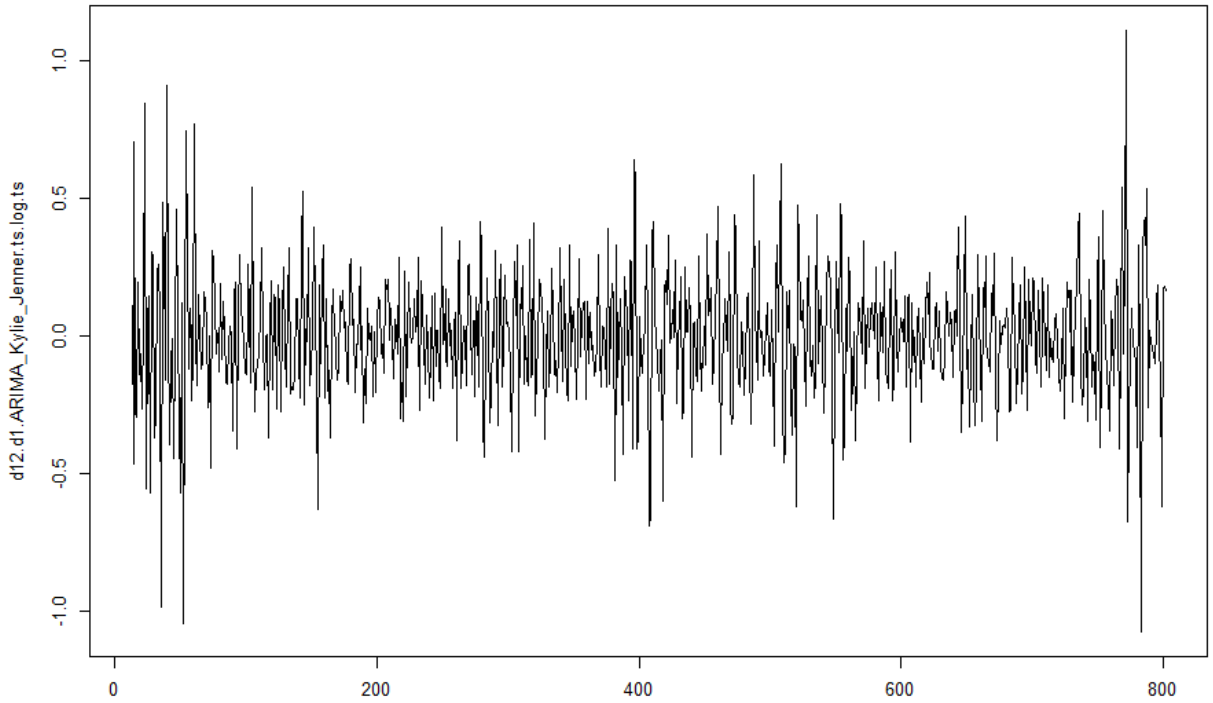

Fig. 6: 12th lag and 1st Differential of log Transformation of the dataset

4.1 ARIMA model Analysis

The ARIMA model is estimated through iteration using ML method where p= 11 , d= 1 and q =2. The best fit model obtained is with the least BIC of -783.5349 and log likelihood of 405.14. Thus ,the best model as per ARIMA analysis is (1,1,2).

```
> best.bic
[1] -783.5349
> best.fit

Call:
arima(x = x1.ts, order = c(p, d, q), method = "ML")

Coefficients:
          ar1      ma1      ma2
      0.4318  -0.5736  -0.3049
s.e.  0.0703   0.0718   0.0498

sigma^2 estimated as 0.0213:  log likelihood = 405.14,  aic = -802.29
> best.model
[1] 1 1 2
```

The (1,1,2) model was used to build and prediction for 36 lags was obtained as shown in Fig 7 and Fig 8. However, the model didn't represent the true trend in the data as the predictions were pretty flat.

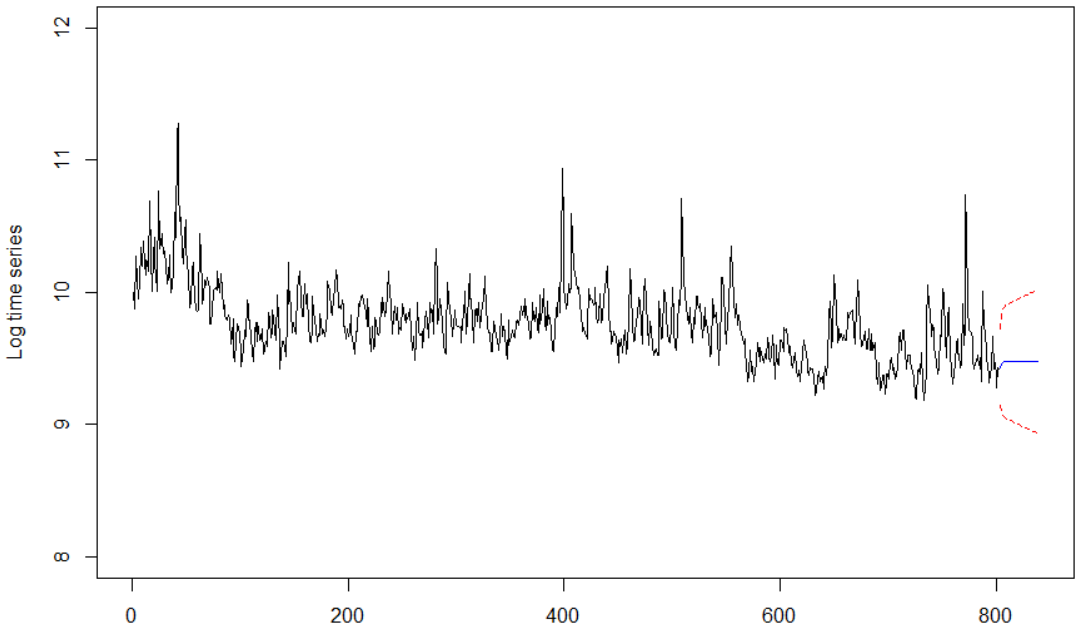

Fig 7 : Prediction for Log time series for h=36

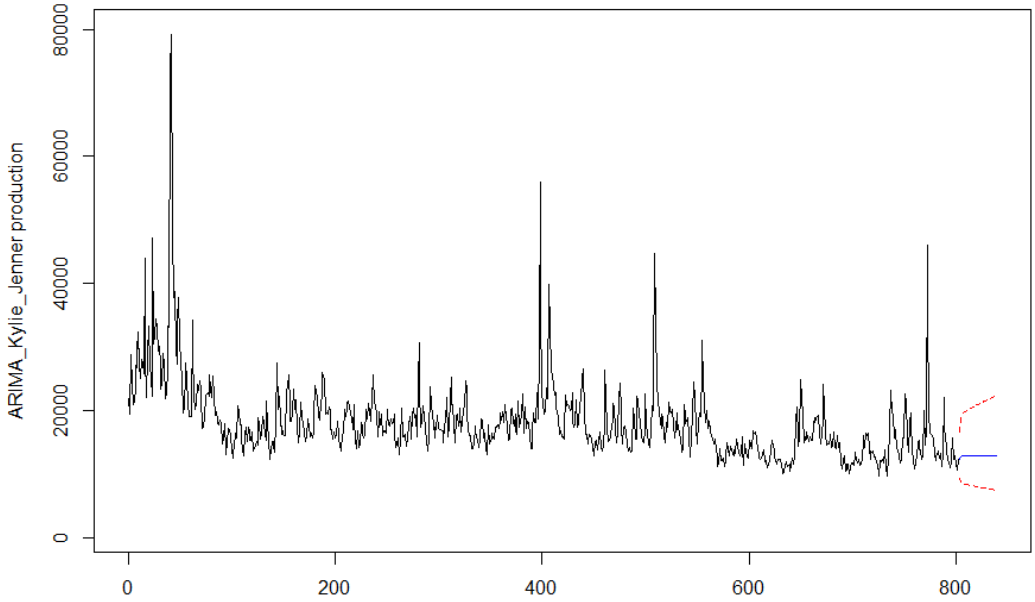

Fig 8 : Prediction for actual time series for h=36

So, in order to verify this phenomena the forecast library was used and auto.ARIMA function was used on the Kylie Jenner time series which gave the ARIMA(4,1,4) as true model with BIC of 15501.63 and log likelihood of -7720.72. The prediction for this function is represented in Fig 9 which gives a slightly better estimation for the time series compared to (1,1,2) model. However the improvement is still minor and not the true representation of time series.

```
> auto.arima(Kylie_Jenner.ts)
Series: Kylie_Jenner.ts
ARIMA(4,1,4)

Coefficients:
          ar1      ar2     ar3     ar4     ma1     ma2      ma3      ma4
      -0.1431  -0.7050  0.3492  0.1301  -0.059  0.2984  -0.6334  -0.3954
s.e.   0.1927   0.0719  0.1195  0.1299   0.187  0.1000   0.0821   0.1518

sigma^2 estimated as 13579316:  log likelihood=-7720.72
AIC=15459.45   AICc=15459.67   BIC=15501.63
```

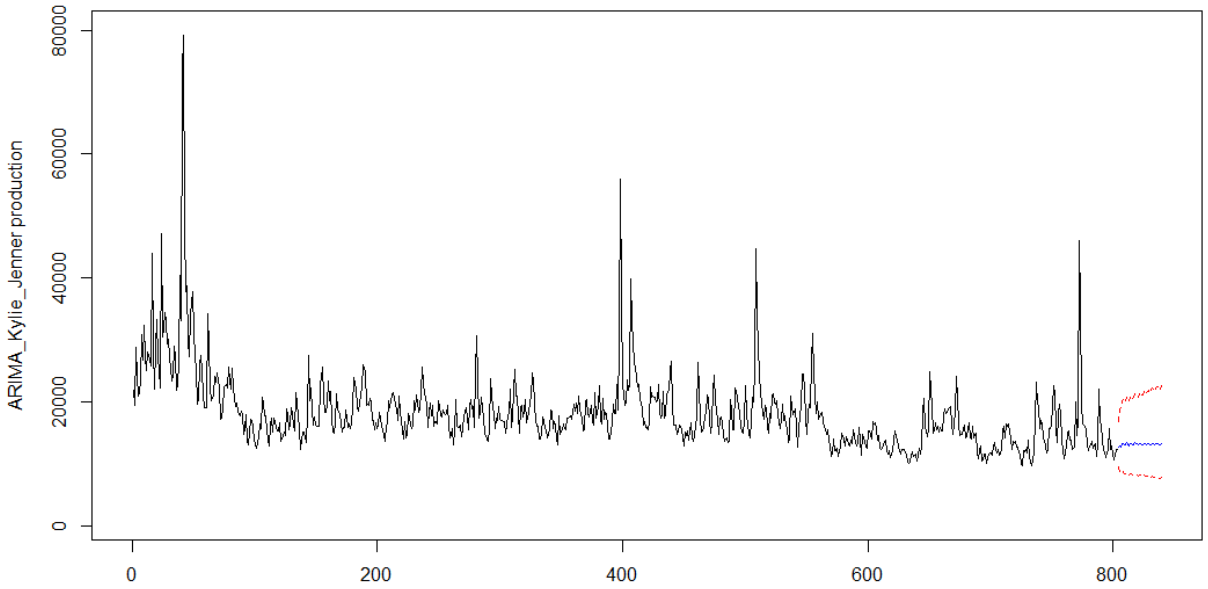

Fig 9 : Prediction for actual time series for h=36 using Auto.ARIMA function

4.2 SARIMA model building & analysis

The ARIMA model is estimated through iteration using ML method where p= 11 , d= 1 and q =2 and P= 2 , D= 1 and Q=2. The best fit model obtained is with the least BIC of -783.5349 and log likelihood of 405.14. Thus ,the best model as per ARIMA analysis is (0,0,0,1,1,2).

```
> best.bic
[1] -783.5349
> best.fit

Call:
arima(x = x.ts, order = c(p, d, q), seasonal = list(order = c(P, D, Q), frequency(x.ts)),
    method = "ML")

Coefficients:
         sar1     sma1     sma2
       0.4318  -0.5736  -0.3049
s.e.   0.0703   0.0718   0.0498

sigma^2 estimated as 0.0213:  log likelihood = 405.14,  aic = -802.29
> best.model
[1] 0 0 0 1 1 2
```

The SARIMA(0,0,0,1,1,2) model was used to build and prediction for 36 lags was obtained as shown in Fig 10 and Fig 11. This model gave much better predictions for the time series as can be seen from the nature of the plots in Fig. 10 and Fig. 11.

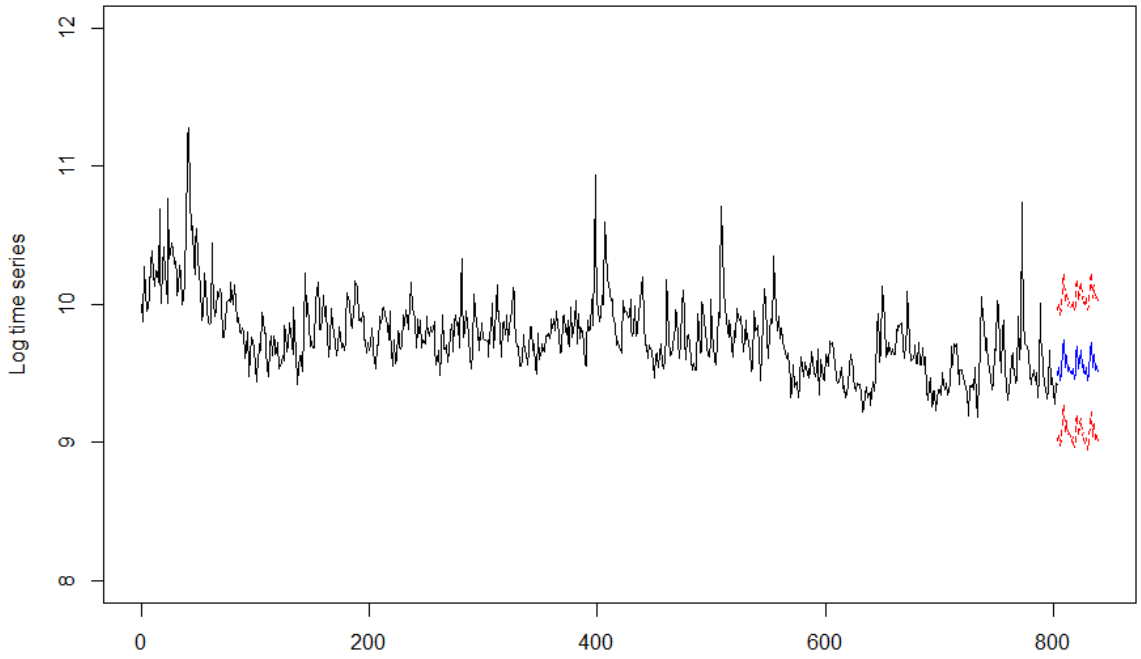

Fig 10 : Log Prediction for time series for h=36

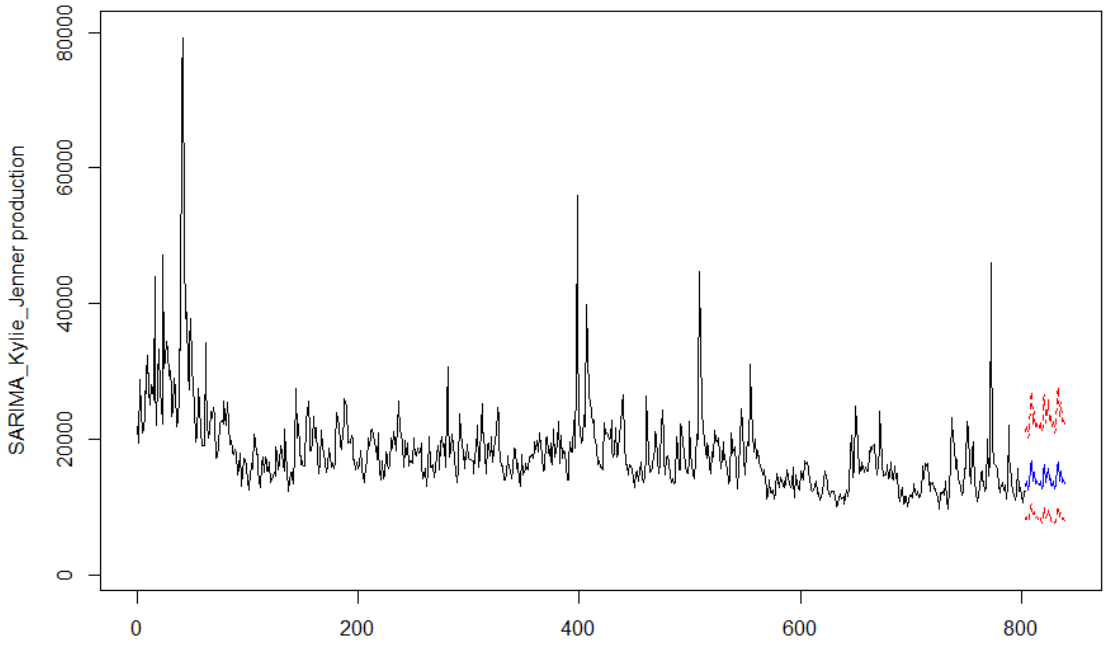

Fig 11 : Actual Prediction for time series for h=36

4.3 Dynamic Linear Model Build and Analysis:

We are going to model this time series using 2nd order polynomial trend and a Fourier representation with all Fourier frequencies. The model is defined with the MLE of observational and system variances using the dlmModPoly function and dlmModTrigs for seasonality. Fig 12 shows the one step ahead forecast errors for the residuals while Fig 13 shows the Q-Q plot for one step ahead forecast errors.

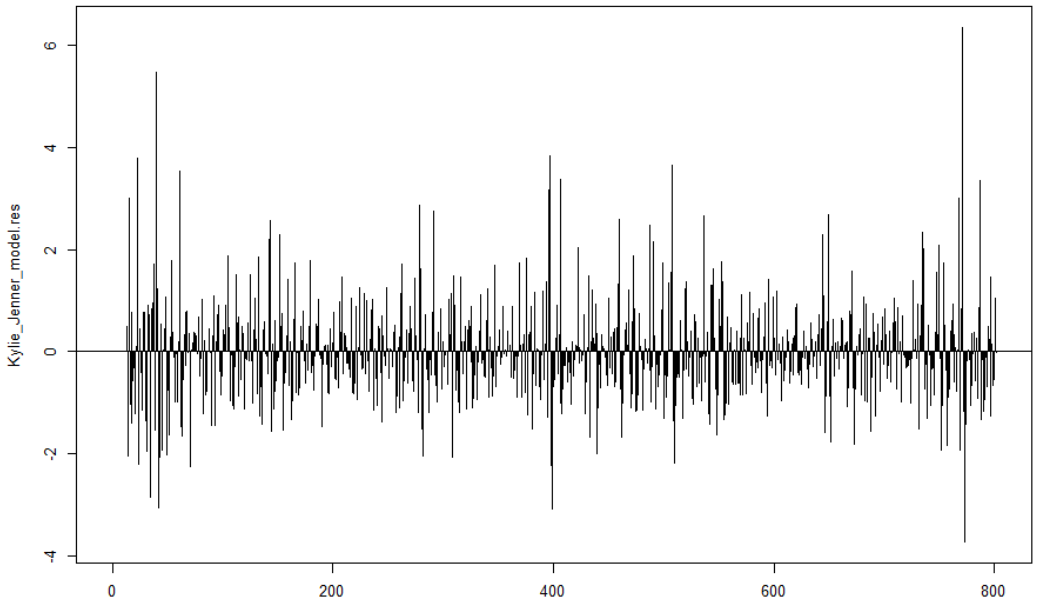

Fig 12 : One-step ahead forecast errors

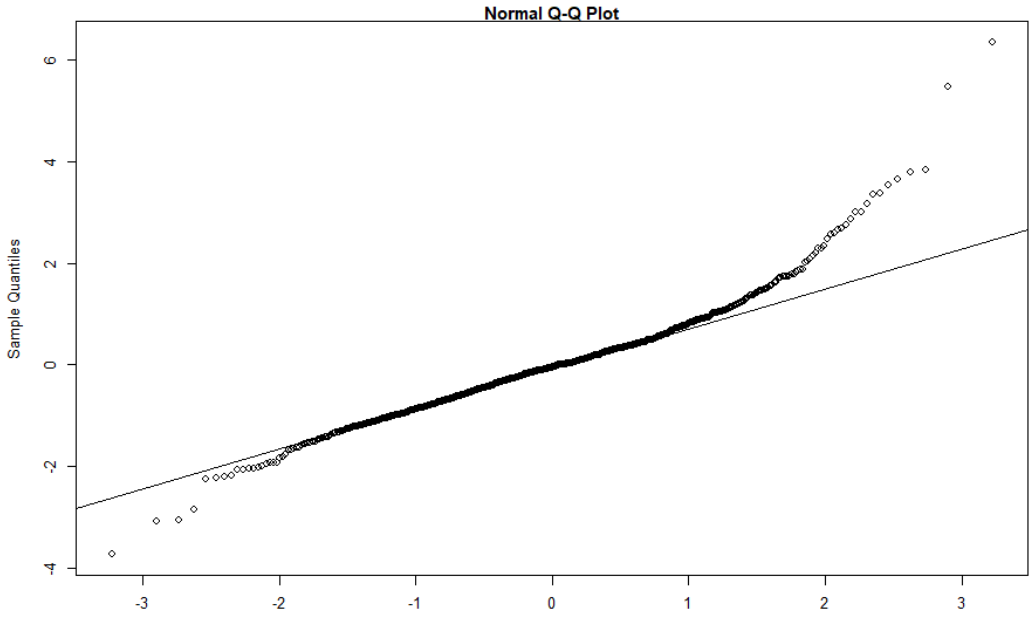

Fig 13 : Q-Q plot for one-step ahead forecast errors

The diagnostic test for normality was made with Shapiro-Wilk normality test where the null hypothesis was that errors are normally distributed. However, the p-value obtained was very low which proved that the errors were not normally distributed.

```
> shapiro.test(Kylie_Jenner_model.res)

        Shapiro-Wilk normality test

data:  Kylie_Jenner_model.res
W = 0.94703, p-value = 2.32e-16
```

Again, the Ljung-Cox test was conducted to test the autocorrelation with the null hypothesis that the errors are independent. However again the p-value was low indicating that the errors were co-related . Thus the time series is not suitable for DLM prediction

```
> Box.test(Kylie_Jenner_model.res, lag=20, type="Ljung")

        Box-Ljung test

data:  Kylie_Jenner_model.res
X-squared = 127.23, df = 20, p-value < 2.2e-16
```

Nonetheless the predictions for lag of 36 was plotted as shown in Fig 14 and Fig 15. Thus, the nature of the graph is not representative of the nature of the time-series. Hence the DLM is not a good predictor for this time-series.

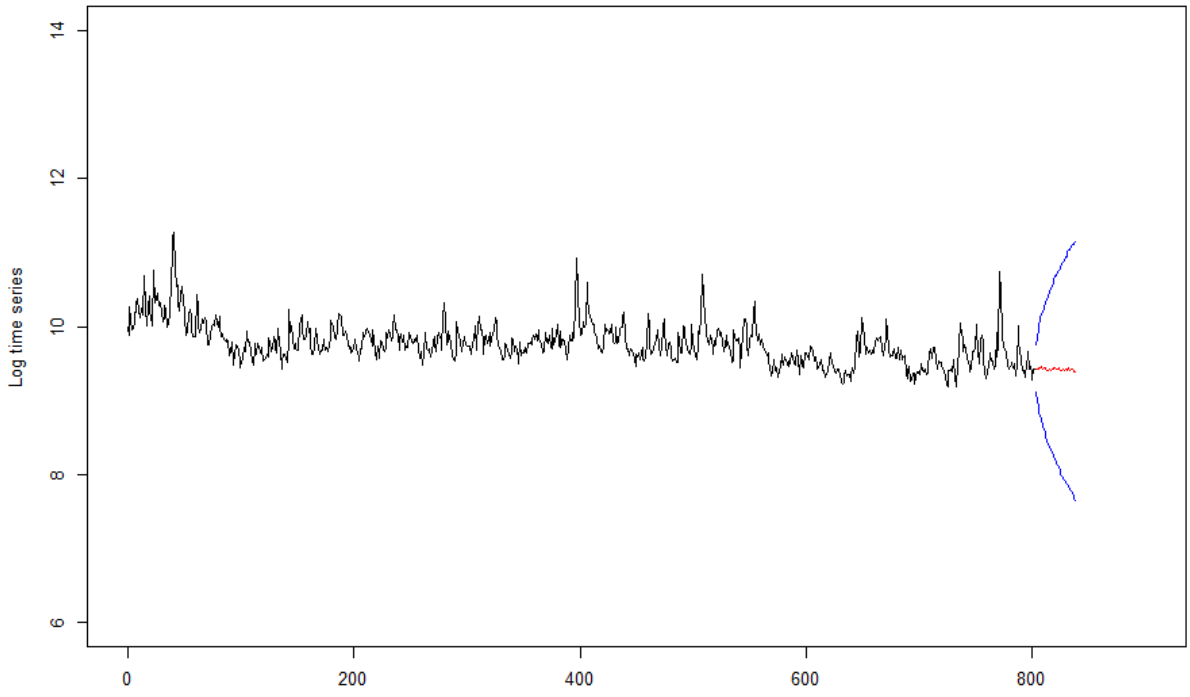

Fig 14 : Log Prediction of Time Series with h=36

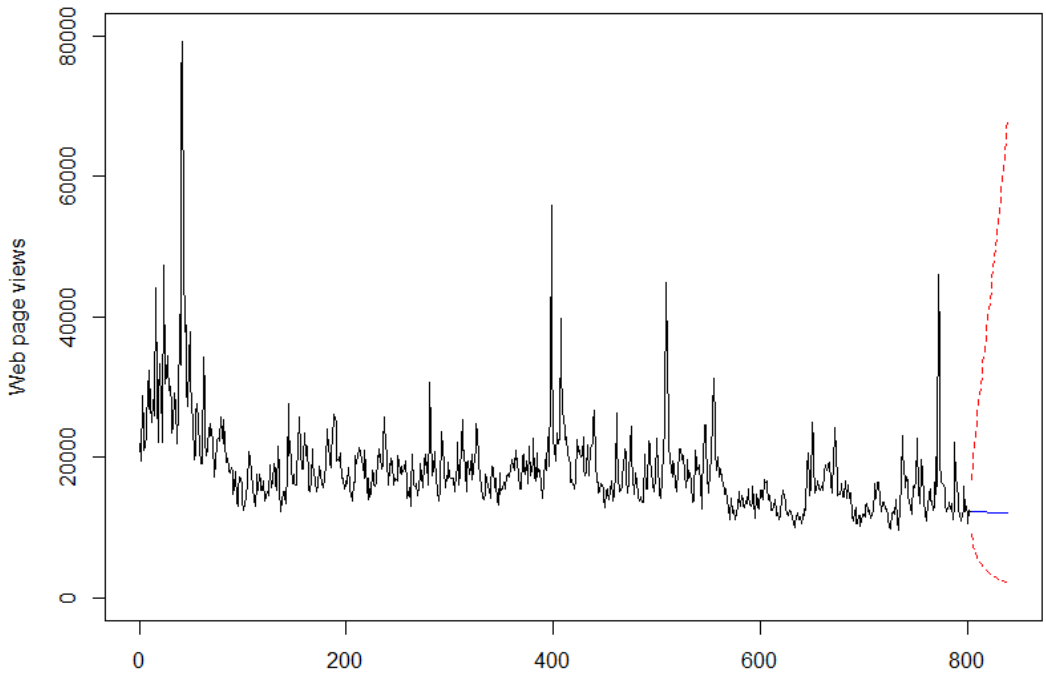

Fig 15 : Actual Prediction of Time Series with h=36

5. Model Comparison:

The SARIMA models gave the best results followed by Auto.ARIMA and ARIMA modles with prediction closer to the historical profile. The DLM model was not successful in predicting the time forecast for the data. This might be due to underlying asymmetry of the time series model.

6. Conclusion:

A case study of Dynamic Linear Modeling vs Box Jenkins Methods was made in this study using the webpage traffic data for Wikipedia pages. Time Series models can be successfully used for web-traffic forecasting by fitting the right model. One should start at simple models and then proceed towards more complex models to find the best fit analysis. The supervised classification algorithm can be applied to further classify each of the time-series datasets.

This work can be further extended in the future to classify other web-page traffic with an algorithm that iterates from simplest time-series model to more complex model with a top-down approach. The Machine Learning algorithm if based on Time Series models should thus take a simple model to complex model evaluation approach to classify the data

7. References:

1. https://www.kaggle.com/c/web-traffic-time-series-forecasting/data.
2. 'Time Series Modeling Computation and Inference' by Raquel Prado and Mike West
3. Bojer, Casper Solheim & Meldgaard, Jens Peder. (2020). Learnings from Kaggle's Forecasting Competitions. 10.13140/RG.2.2.21579.75046.
4. https://docs.aws.amazon.com/forecast/latest/dg/related-time-series-datasets.html

8. Appendix:

8.1 ARIMA Code

```
################################################################################
#########################
# Definig and plotting time series
#Plot the model
ARIMA_Kylie_Jenner.ts <- ts(read.table("Kylie_Jenner_Data.txt",header = TRUE)[,1],start=7/1/2015,freq=1)

par(mar=c(3.1,4.1,1.1,2.1),cex=0.8)
plot.ts(ARIMA_Kylie_Jenner.ts, xlab = "Days", ylab = "Kylie Jenner Wikipedia Webpage Hits")
#help(plot.ts)


#par(mfrow=c(1,1))
#ARIMA_Kylie_Jenner.ts <- ts(production$ARIMA_Kylie_Jenner,start =1958,frequency = 12)
#plot(ARIMA_Kylie_Jenner.ts, ylab = "ARIMA_Kylie_Jenner production")

# Log of time series transformation
ARIMA_Kylie_Jenner.ts.log <- log(ARIMA_Kylie_Jenner.ts)
plot(ARIMA_Kylie_Jenner.ts.log, ylab = "log ARIMA_Kylie_Jenner production")

# Converting to a stationary time series
diff.ARIMA_Kylie_Jenner.ts.log.ts <- diff(ARIMA_Kylie_Jenner.ts.log, differences = 1)
plot(diff.ARIMA_Kylie_Jenner.ts.log.ts)

# Removing the seasonality
d12.d1.ARIMA_Kylie_Jenner.ts.log.ts <- diff(diff.ARIMA_Kylie_Jenner.ts.log.ts, lag = 12)
plot(d12.d1.ARIMA_Kylie_Jenner.ts.log.ts)


# ACF and PACF values
par(mfrow=c(1,2))
acf_ARIMA_Kylie_Jenner <- acf(d12.d1.ARIMA_Kylie_Jenner.ts.log.ts)
pacf_ARIMA_Kylie_Jenner <- pacf(d12.d1.ARIMA_Kylie_Jenner.ts.log.ts)
#

# Fitting Best Model

n = length(ARIMA_Kylie_Jenner.ts.log)
max.p = 7
```

```
max.d = 2
max.q = 2
#max.P = 2
#max.D = 1
#max.Q = 2
BIC.array =array(NA,dim=c(max.p+1,max.d+1,max.q+1))
AIC.array =array(NA,dim=c(max.p+1,max.d+1,max.q+1))
best.bic <- 1e8
x1.ts = ARIMA_Kylie_Jenner.ts.log

for (p in 0:max.p) for(d in 0:max.d) for(q in 0:max.q)
 #for (P in 0:max.P) for(D in 0:max.D) for(Q in 0:max.Q)
 {
  cat("p=",p,", d=",d,", q=",q,"\n")

  fit <- arima(x1.ts, order = c(p,d,q), method="ML")
  number.parameters <- length(fit$coef) + 1
  BIC.array[p+1,d+1,q+1] = -2*fit$loglik + log(n)*number.parameters
  AIC.array[p+1,d+1,q+1] = -2*fit$loglik + 2*number.parameters

  if (BIC.array[p+1,d+1,q+1] < best.bic)
  {
   best.bic <- BIC.array[p+1,d+1,q+1]
   best.fit <- fit
   best.model <- c(p,d,q)
  }

 }

best.bic
best.fit
best.model


# Best BIC model is [111102] with BIC of -662.7087
best_ARIMA_Kylie_Jenner_model <- arima(ARIMA_Kylie_Jenner.ts.log, order =
c(1,1,2),method = "ML")

number.parameters <- length(best_ARIMA_Kylie_Jenner_model$coef) + 1
-2*best_ARIMA_Kylie_Jenner_model$loglik +
log(length(ARIMA_Kylie_Jenner.ts.log))*number.parameters
```

```
# Prediction of 36 lags using the best model
h <- 36
forecast <- predict(best_ARIMA_Kylie_Jenner_model,n.ahead = h)

n <- length(ARIMA_Kylie_Jenner.ts.log)
plot(c(ARIMA_Kylie_Jenner.ts.log,rep(NA,h)),type="l",ylim=c(8,12),ylab = "Log time series")
lines((n+1):(n+h),forecast$pred,col="blue")
lines((n+1):(n+h),forecast$pred+1.96*forecast$se,lty=2,col="red") # Confidence Interval
lines((n+1):(n+h),forecast$pred-1.96*forecast$se,lty=2,col="red")


# Plotting the actual forecast
par(mfrow=c(1,1))
n <- length(ARIMA_Kylie_Jenner.ts)
plot(c(ARIMA_Kylie_Jenner.ts,rep(NA,h)),type = 'l',ylab = "ARIMA_Kylie_Jenner production",
ylim=c(1000,80000))
lines((n+1):(n+h),exp(forecast$pred),col="blue")
lines((n+1):(n+h),exp(forecast$pred+1.96*forecast$se),lty=2,col="red")    #    Confidence
Interval
lines((n+1):(n+h),exp(forecast$pred-1.96*forecast$se),lty=2,col="red")
```

8.2 SARIMA Code

```
##############################################################################
########################
# Definig and plotting time series
#Plot the model
SARIMA_Kylie_Jenner.ts    <-    ts(read.table("Kylie_Jenner_Data.txt",header    =
TRUE)[,1],start=7/1/2015,freq=1)

par(mar=c(3.1,4.1,1.1,2.1),cex=0.8)
plot.ts(SARIMA_Kylie_Jenner.ts, xlab = "Days", ylab = "Kylie Jenner Wikipedia Webpage Hits")
#help(plot.ts)


#par(mfrow=c(1,1))
#SARIMA_Kylie_Jenner.ts <- ts(production$SARIMA_Kylie_Jenner,start =1958,frequency =
12)
#plot(SARIMA_Kylie_Jenner.ts, ylab = "SARIMA_Kylie_Jenner production")
```

```
# Log of time series transformation
SARIMA_Kylie_Jenner.ts.log <- log(SARIMA_Kylie_Jenner.ts)
plot(SARIMA_Kylie_Jenner.ts.log, ylab = "log SARIMA_Kylie_Jenner production")

# Converting to a stationary time series
diff.SARIMA_Kylie_Jenner.ts.log.ts <- diff(SARIMA_Kylie_Jenner.ts.log, differences = 1)
plot(diff.SARIMA_Kylie_Jenner.ts.log.ts)

# Removing the seasonality
d12.d1.SARIMA_Kylie_Jenner.ts.log.ts <- diff(diff.SARIMA_Kylie_Jenner.ts.log.ts, lag = 12)
plot(d12.d1.SARIMA_Kylie_Jenner.ts.log.ts)


# ACF and PACF values
par(mfrow=c(1,2))
acf_SARIMA_Kylie_Jenner <- acf(d12.d1.SARIMA_Kylie_Jenner.ts.log.ts)
pacf_SARIMA_Kylie_Jenner <- pacf(d12.d1.SARIMA_Kylie_Jenner.ts.log.ts)
#

# Fitting Best Model

n = length(SARIMA_Kylie_Jenner.ts.log)
max.p = 7
max.d = 1
max.q = 2
max.P = 7
max.D = 1
max.Q = 2
BIC.array =array(NA,dim=c(max.p+1,max.d+1,max.q+1,max.P+1,max.D+1,max.Q+1))
AIC.array =array(NA,dim=c(max.p+1,max.d+1,max.q+1,max.P+1,max.D+1,max.Q+1))
best.bic <- 1e8
x.ts = SARIMA_Kylie_Jenner.ts.log

for (p in 0:max.p) for(d in 0:max.d) for(q in 0:max.q)
 for (P in 0:max.P) for(D in 0:max.D) for(Q in 0:max.Q)
 {
  cat("p=",p,", d=",d,", q=",q,", P=",P,", D=",D,", Q=",Q,"\n")

  fit <- arima(x.ts, order = c(p,d,q),
           seas = list(order = c(P,D,Q),
                  frequency(x.ts)),method="ML")
```

```
    number.parameters <- length(fit$coef) + 1
    BIC.array[p+1,d+1,q+1,P+1,D+1,Q+1] = -2*fit$loglik + log(n)*number.parameters
    AIC.array[p+1,d+1,q+1,P+1,D+1,Q+1] = -2*fit$loglik + 2*number.parameters

    if (BIC.array[p+1,d+1,q+1,P+1,D+1,Q+1] < best.bic)
    {
      best.bic <- BIC.array[p+1,d+1,q+1,P+1,D+1,Q+1]
      best.fit <- fit
      best.model <- c(p,d,q,P,D,Q)
    }

  }

best.bic
best.fit
best.model

# Best BIC model is [111102] with BIC of -662.7087
best_SARIMA_Kylie_Jenner_model <- arima(SARIMA_Kylie_Jenner.ts.log, order = c(0,0,0),
seasonal = list(order=c(2,1,2),period=12),method = "CSS-ML")

number.parameters <- length(best_SARIMA_Kylie_Jenner_model$coef) + 1
-2*best_SARIMA_Kylie_Jenner_model$loglik                                    +
log(length(SARIMA_Kylie_Jenner.ts.log))*number.parameters


# Prediction of 36 lags using the best model
h <- 36
forecast <- predict(best_SARIMA_Kylie_Jenner_model,n.ahead = h)

n <- length(SARIMA_Kylie_Jenner.ts.log)
plot(c(SARIMA_Kylie_Jenner.ts.log,rep(NA,h)),type="l",ylim=c(8,12),ylab = "Log time series")
lines((n+1):(n+h),forecast$pred,col="blue")
lines((n+1):(n+h),forecast$pred+1.96*forecast$se,lty=2,col="red") # Confidence Interval
lines((n+1):(n+h),forecast$pred-1.96*forecast$se,lty=2,col="red")


# Plotting the actual forecast
par(mfrow=c(1,1))
n <- length(SARIMA_Kylie_Jenner.ts)
```

```
plot(c(SARIMA_Kylie_Jenner.ts,rep(NA,h)),type = 'l',ylab = "SARIMA_Kylie_Jenner production", ylim=c(1000,80000))
lines((n+1):(n+h),exp(forecast$pred),col="blue")
lines((n+1):(n+h),exp(forecast$pred+1.96*forecast$se),lty=2,col="red") # Confidence Interval
lines((n+1):(n+h),exp(forecast$pred-1.96*forecast$se),lty=2,col="red")
```

8.3 DLM Code

```
#########################################################################
#DLM

library('ggplot2')
library('forecast')
library('tseries')
require(dlm)

DLM_Kylie_Jenner.ts <- ts(read.table("Kylie_Jenner_Data.txt",header = TRUE)[,1],start=7/1/2015,freq=1)

#DLM_Kylie_Jenner.ts <- tsclean(Raw_DLM_Kylie_Jenner.ts)


par(mar=c(3.1,4.1,1.1,2.1),cex=0.8)
plot(DLM_Kylie_Jenner.ts)

#Using log transformation to account for seasonal pattern

log.DLM_Kylie_Jenner.ts <- log(DLM_Kylie_Jenner.ts)
par(mar=c(3.1,4.1,1.1,2.1),cex=0.8)
plot(log.DLM_Kylie_Jenner.ts)

adf.test(log.DLM_Kylie_Jenner.ts)
decomp = stl(log.DLM_Kylie_Jenner.ts, s.window = "periodic")


# The series is non-stationary and has seasonality

# We are going to model this time series using 2nd order polynomial trend and a Fourier representation with all Fourier frequencies
#Define the model and MLE of observational and system variances
```

```
model_Kylie_Jenner <- function(parm) {
 dlmModPoly(order = 2, dV = exp(parm[1]), dW = c(exp(parm[2]),exp(parm[3]))) + dlmModTrig(s = 12, dV = 0, dW=exp(parm[4]))
}
fit.model_Kylie_Jenner <- dlmMLE(log.DLM_Kylie_Jenner.ts, rep(0.1,4), model_Kylie_Jenner)
fit.model_Kylie_Jenner$convergence
unlist(model_Kylie_Jenner(fit.model_Kylie_Jenner$par)[c("V","W")])

#Define model using MLE

Kylie_Jenner_mod.MLE <- model_Kylie_Jenner(fit.model_Kylie_Jenner$par)

#Kalman filter
Kylie_Jennerfilt <- dlmFilter(log.DLM_Kylie_Jenner.ts, Kylie_Jenner_mod.MLE)

Kylie_Jennercov.filt <- with(fit.model_Kylie_Jenner, dlmSvd2var(U.C,D.C))

seas.term = 3

#Analysis of the seasonal Fourier frequencies indicates
#that we may need 5 or 6 harmonics. So little or no gain would be obtained from removing model.

#One-step ahead forecast errors
Kylie_Jenner_model.res <- residuals(Kylie_Jennerfilt, sd = FALSE)


#Plot one-step ahead forecast errors
plot(Kylie_Jenner_model.res, type='h'); abline(h=0)

#Plot ACF and PACG on one-step ahead forecast errors
acf(Kylie_Jenner_model.res, na.action = na.pass)
pacf(Kylie_Jenner_model.res, na.action = na.pass)

#Plot qq plot of one-step ahead forecast errors
qqnorm(Kylie_Jenner_model.res);qqline(Kylie_Jenner_model.res)
# Results show some lags are statistically significant

#Diagnostic Test
#Normality test with Shapiro-Wilk normality test
```

```
#Null hypothesis : errors are normally distributed

shapiro.test(Kylie_Jenner_model.res)

#P-value is large . So no departure form normality

#Ljung-Cox test to test the autocorrelation
#Null hypothesis : errors are independent

Box.test(Kylie_Jenner_model.res, lag=20, type="Ljung")
sapply(1:20,function(i)
  Box.test(Kylie_Jenner_model.res, lag = i, type ="Ljung-Box")$p.value)


# Prediction of 36 lags using the best model
h <- 36
Kylie_Jenner_forecast <- dlmForecast(Kylie_Jennerfilt,n = h)

plot(Kylie_Jenner_forecast$f)

n <- length(Kylie_Jennerfilt)
plot(log.DLM_Kylie_Jenner.ts, type="l",ylim=c(6,14),,xlim =c(1,900), ylab = "Log time series")
lines(Kylie_Jenner_forecast$f, col="red")
lines(Kylie_Jenner_forecast$f+1.95*sqrt(unlist(Kylie_Jenner_forecast$Q)), col="blue")
lines(Kylie_Jenner_forecast$f-1.95*sqrt(unlist(Kylie_Jenner_forecast$Q)), col="blue")

# Plotting the actual forecast
par(mfrow=c(1,1))
n <- length(DLM_Kylie_Jenner.ts)
plot(c(DLM_Kylie_Jenner.ts,rep(NA,h)),type = 'l',ylab = "Web page views", ylim=c(0,80000))
lines((n+1):(n+h),exp(Kylie_Jenner_forecast$f),col="blue")
lines((n+1):(n+h),exp(Kylie_Jenner_forecast$f+1.95*sqrt((unlist(Kylie_Jenner_forecast$Q))))
,lty=2,col="red") # Confidence Interval
lines((n+1):(n+h),exp(Kylie_Jenner_forecast$f-
1.95*sqrt((unlist(Kylie_Jenner_forecast$Q)))),lty=2,col="red")
```