IOSR Journal of Engineering (IOSRJEN) **www.iosrjen.org**
ISSN (e): 2250-3021, ISSN (p): 2278-8719
Vol. 12, Issue 8, August 2022, ||Series -I|| PP 25-37

Time Series Analysis – Daily Website Visitors

DEBMALYA RAY

Received 22 August 2022; Accepted 10 September 2022

ABSTRACT
This file contains 5 years of daily time series data for several measures of traffic on the statistical forecasting teaching notes website whose alias is statforecasting.com. The variables have complex seasonality that is keyed to the day of the week and to the academic calendar. The patterns you see here are similar in principle to what you would see in other daily data with day-of-week and time-of-year effects. Some good exercises are to develop a 1-day-ahead forecasting model, a 7-day ahead forecasting model, and an entire-next-week forecasting model (i.e., next 7 days) for unique visitors.

I. INTRODUCTION

A time series is a series of data points or observations recorded at different or regular time intervals. In general, a time series analysis is a sequence of data points taken at equally spaced time intervals. The frequency of recorded data points can be hourly, daily, weekly, monthly, quarterly or annually.

Time Series Forecasting is the process of using a statistical model to predict future values of a time series based on past results.

A time series analysis encompasses statistical methods for analyzing time series data. These methods enable us to extract meaningful statistics, patterns and other characteristics of the data. Time series are visualized with the help of line charts. So, time series analysis involves understanding inherent aspects of the time series data so that we can create meaningful and accurate forecasts.

The variables are daily counts of page loads, unique visitors, first-time visitors and returning visitors to an academic teaching notes website. There are 2167 rows of data spanning the date range from September 14, 2014, to August 19, 2020. A visit is defined as a stream of hits on one or more pages on the site on a given day by the same user, as identified by IP address. Multiple individuals with a shared IP address (e.g., in a computer lab) are considered a single user, so real users may be undercounted to some extent. A visit is classified as "unique" if a hit from the same IP address has not come within the last 6 hours. Returning visitors are identified by cookies if those are accepted. All others are classified as first-time visitors, so the count of unique visitors is the sum of the counts of returning and first-time visitors by definition. The data was collected through a traffic monitoring service known as StatCounter.

1.1. Problem Statement

Problem:

To predict the future values of daily website visitors based on past results.

Background:
There are various techniques and methods through which the time series data prediction is performed. We tried to explore various techniques to understand the evolution of such techniques using simple ML to time series algorithms in the best possible ways.

1.2. Aim and Objectives

Aim of the study:
To understand the evolution of the time series algorithm and how we can apply such techniques in time series problems.

Based on the aim, we have created a set of objectives as mentioned below:

- To use the traditional ML approaches
- To understand if the time series is additive or multiplicative in nature.
- To use simple models like a naive or moving average.
- To use smoothening techniques like Exponential, Holt's smoothing and Winter's smoothing.

- To understand if the time series is stationary or non-stationary
- To find out the trends and seasonality of time series data
- Use of Auto-Correlation and Partial Auto-Correlation.
- Use of Lag plots, ARIMA and SARIMAX.

1.3. Significance of Study

As those trends become stronger and stronger, there is much need to study web user behaviour to better serve the users and increase the value of institutions or enterprises.

Today, understanding the interests of users is becoming a fundamental need for Website owners in order to better serve their users by making adapting the content and usage, the structure of the website to their preferences.

1.4. Scope of Study / Challenges Involved

The work will include:

i) Finding the total number of visitors coming to the website

ii) Forecasting the future prediction of the visitors

Challenges involved:

All the data has to be collected before the study can be performed.

Since these data are in a public format, so no legal authorization is required to collect them but a certain amount of effort will be utilized in scrapping and formatting them in a structured manner.

II. LITERATURE REVIEWS

2.1. Introduction

The internet is growingly rapidly and has a great impact on many businesses. Thousands of companies now own a website and websites have become an integrated part of the business.

Furthermore, many companies have employed many technologies which are available through the web such as online services. With web information, web developers and designers can improve user interfaces, search engines, navigation features, online help and information architecture and have happier visitors/customers. One of the most popular ways that most frequented websites use to collect data and information about their websites is through web analytics. Web analytics collects a large amount of data from users such as browser type, connection speed, screen size, visitors' type etc. The collected data are usually large in quantity and type that need to be further processed to become useful information or knowledge.

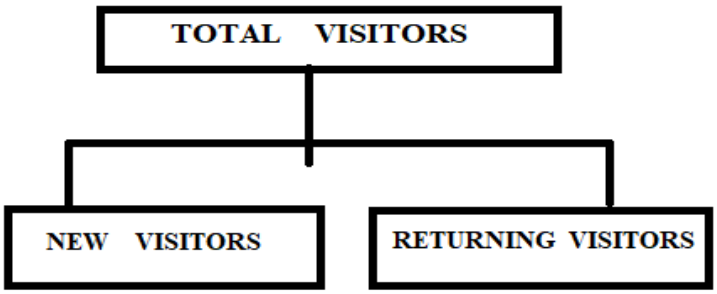


Figure: Independent Variables

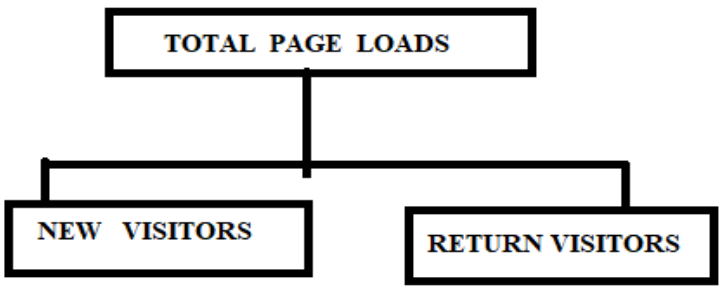


Figure: Dependent Variables

Time Series Definition

A time-series data is a series of data points or observations recorded at different or regular time intervals. In general, a time series is a sequence of data points taken at equally spaced time intervals. The frequency of recorded data points may be hourly, daily, weekly, monthly, quarterly or annually.

Components of a Time-Series

Trend - The trend shows a general direction of the time series data over a long period of time. A trend can be increasing(upward), decreasing(downward), or horizontal(stationary).

Seasonality - The seasonality component exhibits a trend that repeats with respect to timing, direction, and magnitude. Some examples include an increase in water consumption in summer due to hot weather conditions.

Cyclical Component - These are the trends with no set repetition over a particular period of time. A cycle refers to the period of ups and downs, booms and slums of a time series, mostly observed in business cycles. These cycles do not exhibit a seasonal variation but generally occur over a time period of 3 to 12 years depending on the nature of the time series.

Irregular Variation - These are the fluctuations in the time series data which become evident when trend and cyclical variations are removed. These variations are unpredictable, erratic, and may or may not be random.

ETS Decomposition - ETS Decomposition is used to separate different components of a time series. The term ETS stands for Error, Trend and Seasonality.

2.2. Summary Of The Past Studies

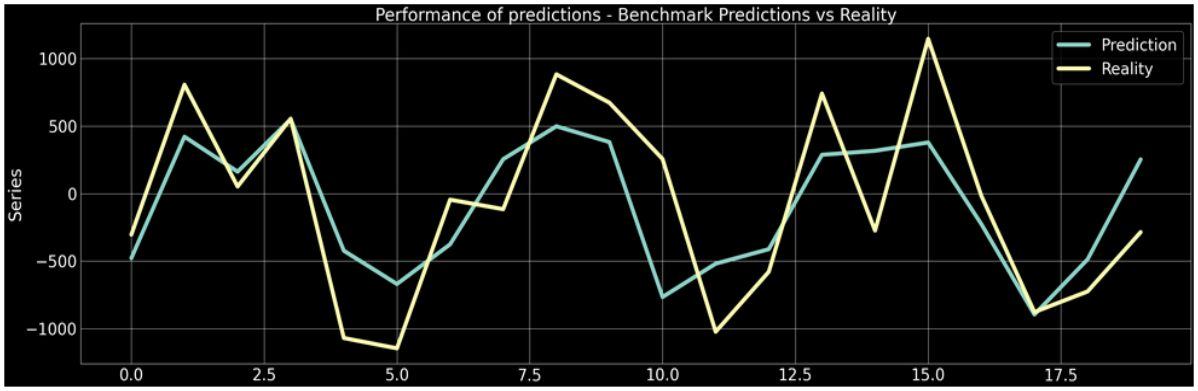
Topics	Author	Machine Learning
Identifying the Number of Visitors to improve Website Usability from Educational Institution Web Log Data	Arvind K. Sharma, P.C. Gupta	Augmented Dickey-Fuller Test, Autoregressive Moving Average (ARMA)
Analyzing The Effects of Sessions on Unique Visitors and Unique Page Views with Google Analytics: A case study of a Tourism Website in Thailand	Jiaranai Awichanirost & Naragain Phumchusri	Augmented Dickey-Fuller Test, Autoregressive Moving Average (ARMA)
Data mining approach for predicting the daily Internet data traffic of a smart-university	Aderibigbe Israel Adekitan, Jeremiah Abolade & Olamilekan Shobayo	Random Forest Predictor, Naïve Bayes Predictor, KNN

Table 1

2.3. Time-Series – Traditional ML Approaches

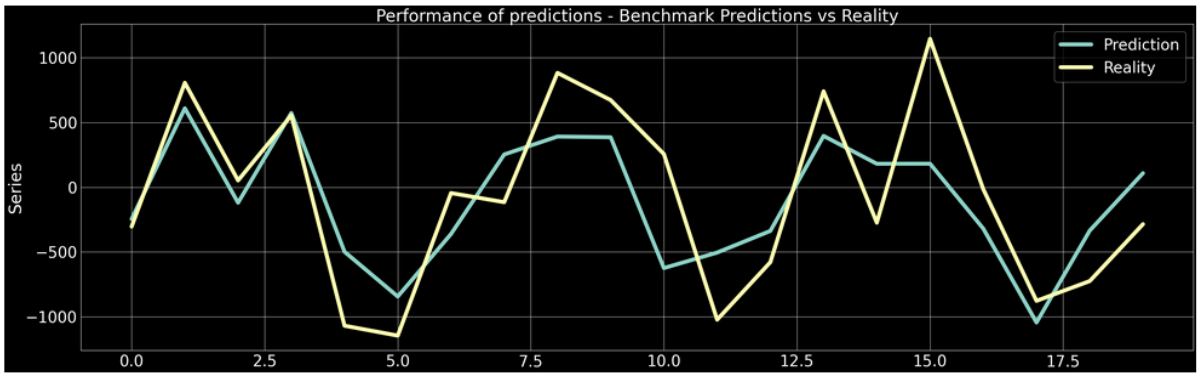
The use of machine learning is described as below :

Adaboost Regressor



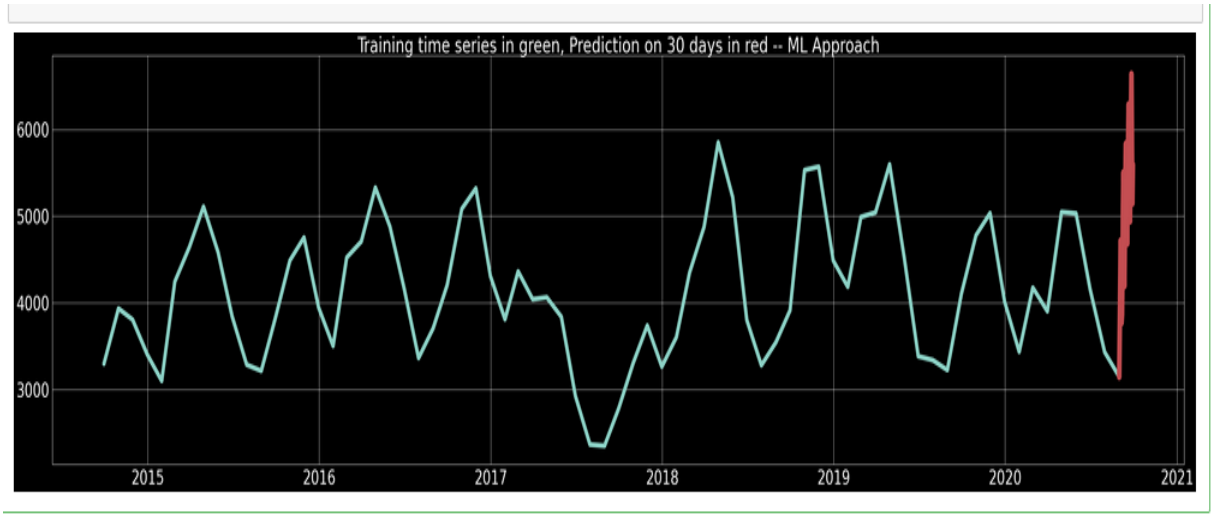

```
------------------------------------------------
mae with 70% of the data to train: 383.9013841337698
------------------------------------------------
R-score with 70% of the data to train: 0.5570560128546717
```

Gradient Boosting Regressor




```
------------------------------------------------
mae with 70% of the data to train: 371.8701889598243
------------------------------------------------
R-score with 70% of the data to train: 0.5932987491660813
```

Training time series in green, Prediction on 30 days in red -- ML Approach



2.4 Additive and Multiplicative Time Series

Additive and Multiplicative Time Series

We may have different combinations of trends and seasonality. Depending on the nature of the trends and seasonality, a time series can be modelled as an additive or multiplicative time series. Each observation in the series can be expressed as either a sum or a product of the components.

Additive time series:
Value = Base Level + Trend + Seasonality + Error
Multiplicative Time Series:
Value = Base Level x Trend x Seasonality x Error

III. RESEARCH METHODOLOGY

Please find below the required methodology to be performed for achieving the aims and objectives:

3.1 Understanding The Data:

Please find below the dataset collected and referred for our analysis:

daily-website-visitors.csv

The variables described are mentioned in Table 3:

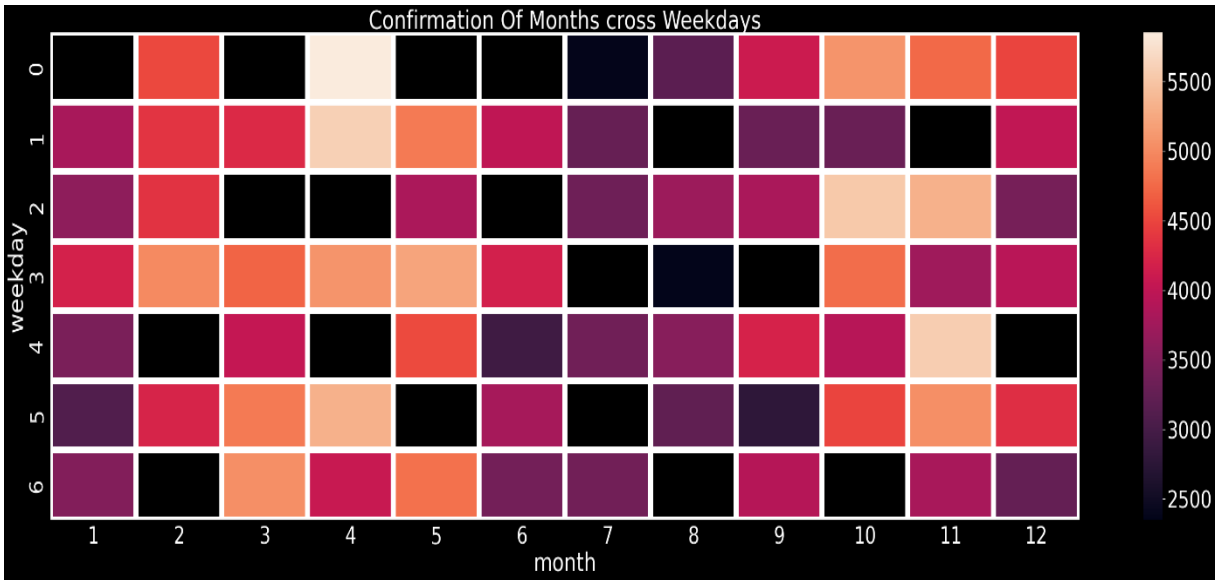
Variable Name	Variable Description	Label
Row	Row	Feature
Day	Day	Feature
Day. Of.Week	Day of the week	Feature
Date	Visitor's Date	Feature
Page.Loads	Number of page loading	**Target**
Unique.Visits	Number of unique visitors	Feature
First.Time.Visits	Number of first-time visitors	Feature
Returning. Visits	The number of visitors returned	Feature

Table: Description of the variables

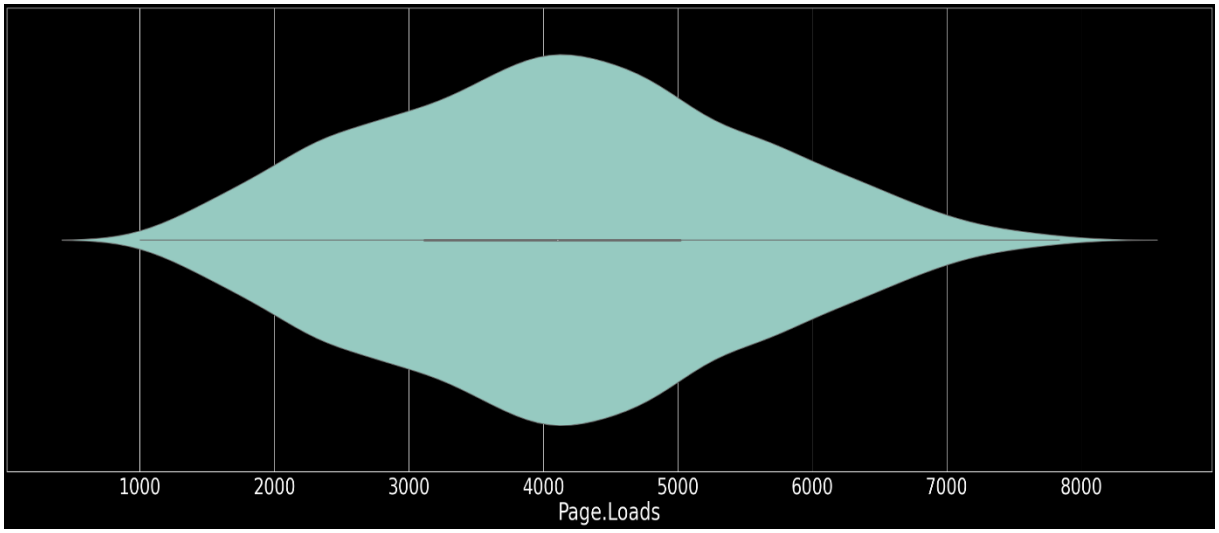
3.2 Visualizing Time Series Data:

Before we proceed into the next step, it is very important that the data should be explored further. Visualizing the time series data will help us to understand its pattern and bring sense out of it. Every time series data works based on data time format. In this document, we will try to visualize the trends, and patterns based on the date represented as weekly, monthly or yearly wise.

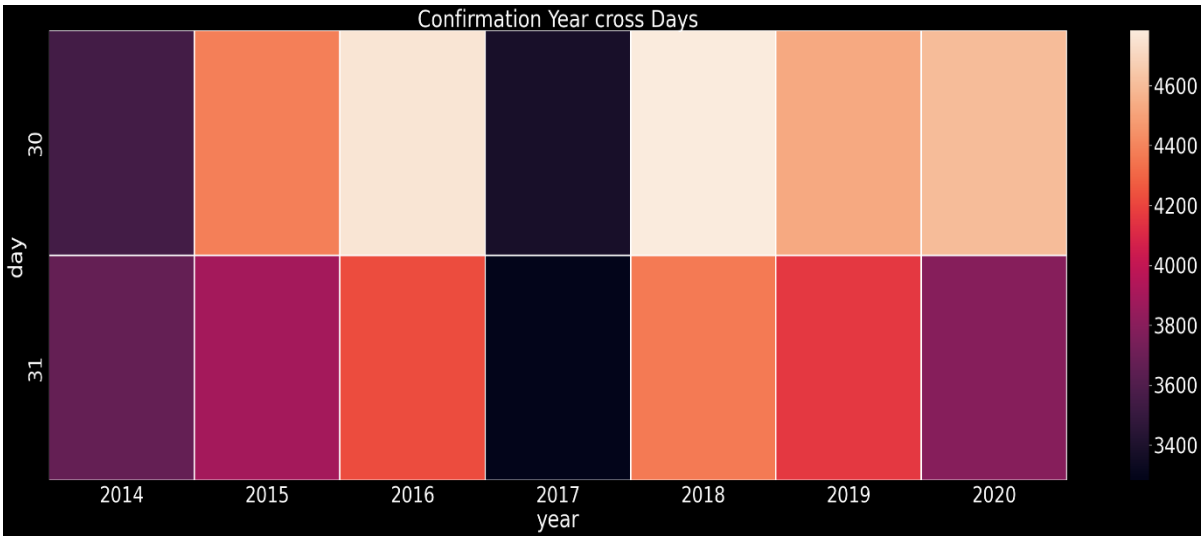
The intensity of data for months cross weekdays



Violin Plot for dataframe – Page Loads



The intensity of data for Year cross Days



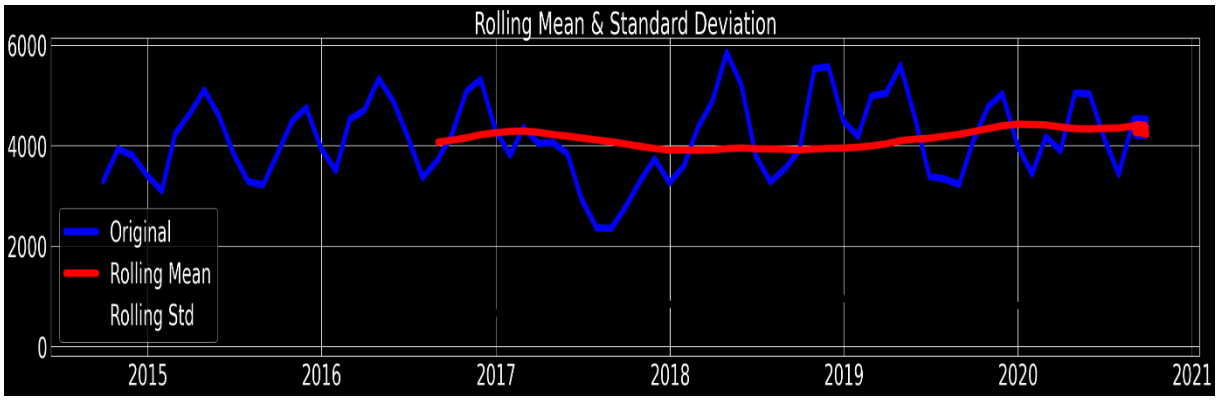
Time Series Analysis to check if the data is Stationary or not

Stationarity is a property of a time series. A stationary series is one where the values of the series are not a function of time. So, the values are independent of time.

The statistical properties of the series like mean, variance and autocorrelation are constant over time. The autocorrelation of the series is nothing but the correlation of the series with its previous values.

Dickey-Fuller TEST

ADF (Augmented Dickey-Fuller) test is a statistical significance test which means the test will give results in hypothesis tests with null and alternative hypotheses. As a result, we will have a p-value from which we will need to make inferences about the time series, whether it is stationary or not.




```
Results of Dickey-Fuller Test:
Test Statistic                  -3.151029
p-value                          0.023000
#Lags Used                      12.000000
Number of Observations Used     88.000000
Critical Value (1%)             -3.506944
Critical Value (5%)             -2.894990
Critical Value (10%)            -2.584615
dtype: float64
```

From the above observation, we can find out that the value of p is less than 0.05. So, it is stationary in nature. To convert any non-stationary data to stationary, we will follow the mentioned below steps.

Introduction to Differencing

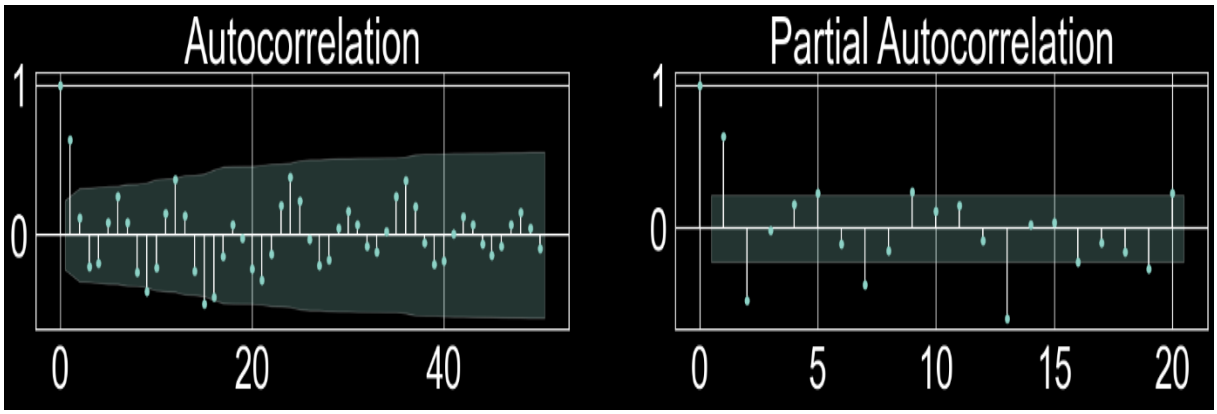
- If Y_t is the value at time t, then the first difference of Y = Yt – Yt-1. In simpler terms, differencing the series is nothing but subtracting the next value by the current value.
- If the first difference doesn't make a series stationary, we can go for the second differencing and so on.
- For example, consider the following series: [1, 5, 2, 12, 20]
- First differencing gives: [5-1, 2-5, 12-2, 20-12] = [4, -3, 10, 8]
- Second differencing gives: [-3-4, -10-3, 8-10] = [-7, -13, -2]

Autocorrelation and Partial Autocorrelation Functions

Autocorrelation is simply the correlation of a series with its own lags. If a series is significantly autocorrelated, that means, the previous values of the series (lags) may be helpful in predicting the current value.

Partial Autocorrelation also conveys similar information but it conveys the pure correlation of a series and its lag, excluding the correlation contributions from the intermediate lags.

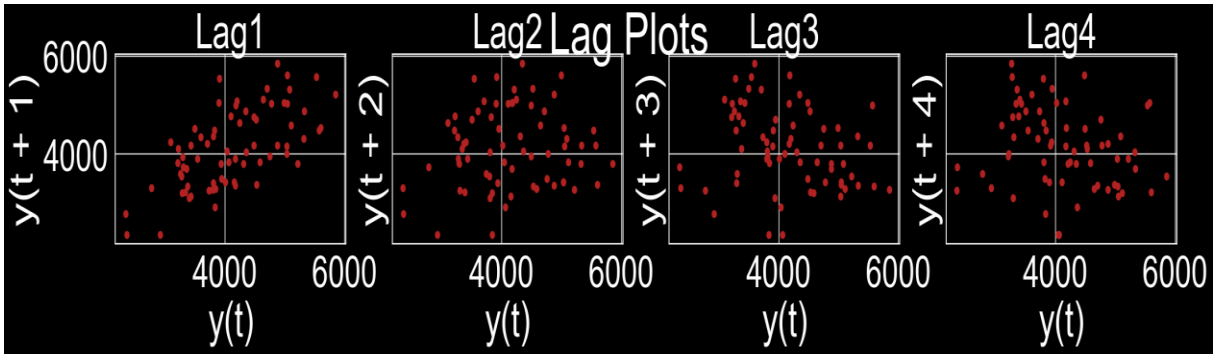
From the above Autocorrelation plot, the value of p derived should range between 8 to 10.



From the above Partial-autocorrelation plot, the value of q is considered as 2. Since the daily website visitors, data provided is stationary in nature, so we considered the value of d as 0.

Lag Plots

A Lag Plot is a scatter plot of a time series against a lag of itself. It is normally used for autocorrelation. If there is any pattern existing in the series, the series is autocorrelated. If there is no such pattern, the series is likely to be a random white noise



3.3 Use of ARIMA and SARIMAX:

ARIMA MODEL

AR: Autoregression. A model that uses the dependent relationship between an observation and some number of lagged observations.

I: Integrated. The use of differencing of raw observations in order to make the time series stationary. MA: Moving Average. A model that uses the dependency between an observation and a residual error from a moving average model applied to lagged observations.

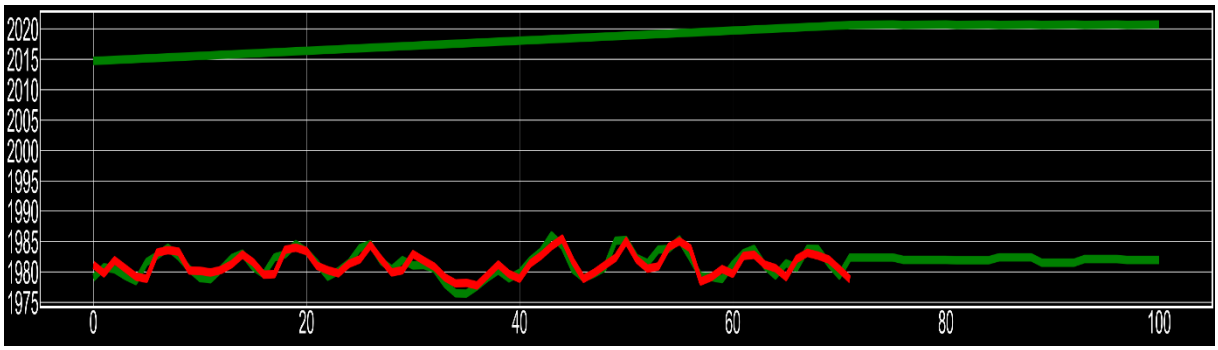
The parameters of the ARIMA model are defined as follows:

p: The number of lag observations included in the model, also called the lag order.

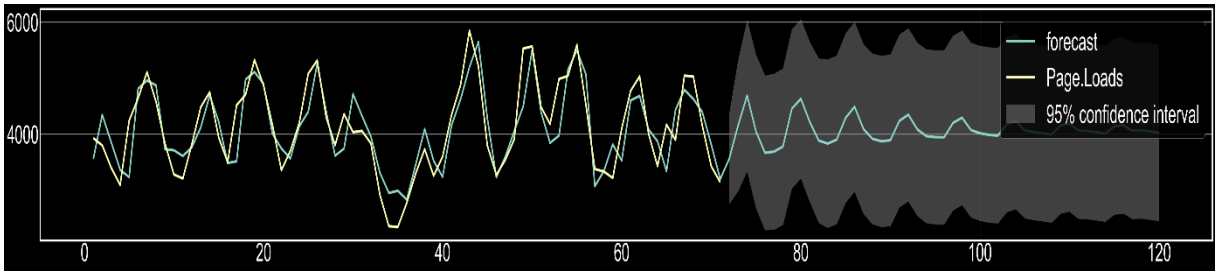
d: The number of times that the raw observations are differenced, also called the degree of difference.
q: The size of the moving average window, also called the order of moving average.

ARIMA
An autoregressive integrated moving average (ARIMA) model is a generalization of an autoregressive moving average (ARMA) model. Both of these models are fitted to time series data either to better understand the data or to predict future points in the series (forecasting). ARIMA models are applied in some cases where data show evidence of non-stationarity, where an initial differencing step (corresponding to the "integrated" part of the model) can be applied one or more times to eliminate the non-stationarity. ARIMA model is of the form: ARIMA(p,d,q): p is the AR parameter, d is a differential parameter, q is the MA parameter

Model Fitting – Time Series



Forecasting the Page Loads




```
The root means the squared error is 500.7702837813342.
```

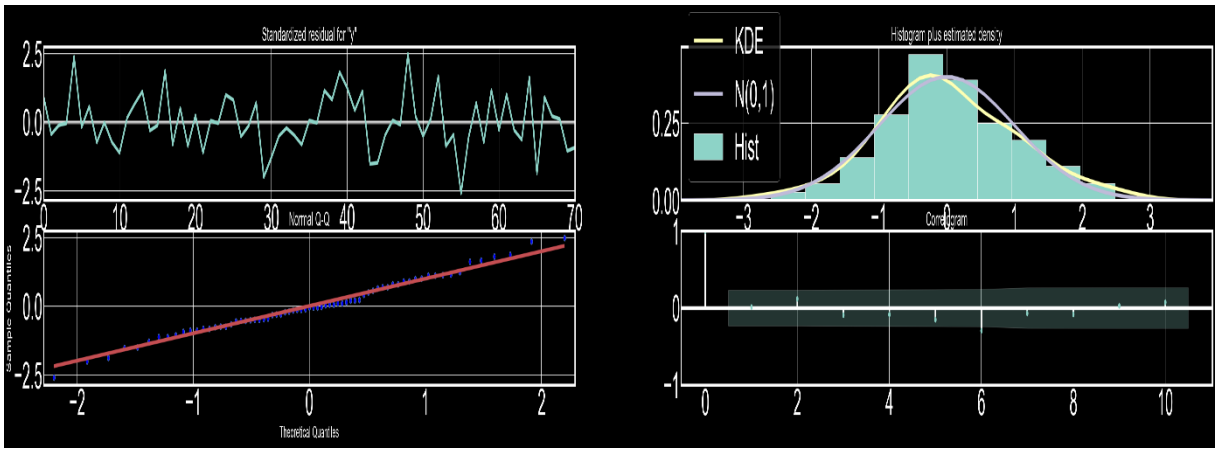
3.4 SARIMA Models

SARIMA models are useful for modelling seasonal time series, in which the mean and other statistics for a given season are not stationary across the years. The SARIMA model defined constitutes a straightforward extension of the nonseasonal autoregressive-moving average (ARMA) and autoregressive integrated moving average (ARIMA) models presented

3.5 SARIMAX Results

```
SARIMAX Results
==========================================================================
===
Dep. Variable:                          y   No. Observations:
71
```

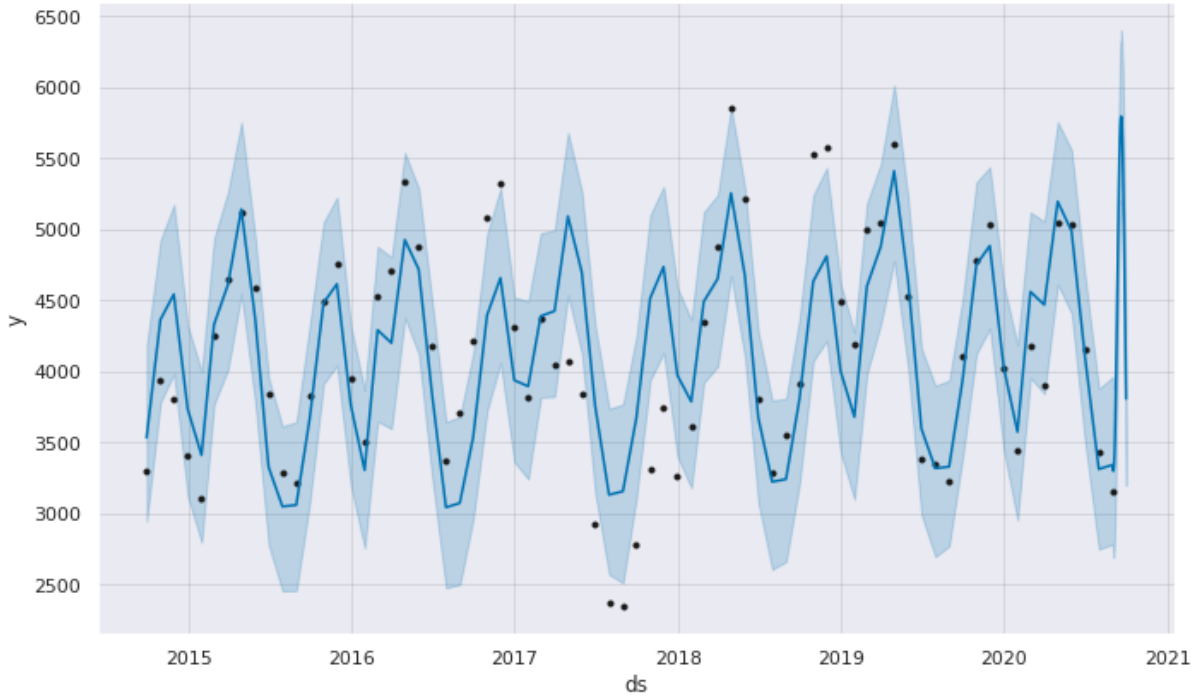
```
Model:                SARIMAX(4, 0, 4)   Log Likelihood                -
528.986
Date:                 Mon, 05 Sep 2022   AIC
1077.972
Time:                         13:08:50   BIC
1100.599
Sample:                              0   HQIC
1086.970
                                 - 71
Covariance Type:                   opg
==============================================================================
===
                 coef    std err          z      P>|z|      [0.025
0.975]
------------------------------------------------------------------------------
---
intercept     -1.7779    152.339     -0.012      0.991    -300.356
296.800
ar.L1         -0.0208      0.070     -0.296      0.767      -0.159
0.117
ar.L2         -0.9629      0.068    -14.171      0.000      -1.096      -
0.830
ar.L3         -0.0207      0.068     -0.304      0.761      -0.154
0.113
ar.L4         -0.9998      0.014    -70.458      0.000      -1.028      -
0.972
ma.L1          0.0382      1.240      0.031      0.975      -2.392
2.468
ma.L2          0.9810      0.203      4.823      0.000       0.582
1.380
ma.L3          0.0382      1.232      0.031      0.975      -2.376
2.453
ma.L4          0.9920      0.227      4.373      0.000       0.547
1.437
sigma2      1.458e+05      0.006   2.53e+07      0.000    1.46e+05
1.46e+05
==============================================================================
========
Ljung-Box (L1) (Q):                   0.01   Jarque-Bera (JB):
0.33
Prob(Q):                              0.92   Prob(JB):
0.85
Heteroskedasticity (H):               1.71   Skew:
0.16
Prob(H) (two-sided):                  0.20   Kurtosis:
3.12
==============================================================================
========


Warnings:
[1] Covariance matrix calculated using the outer product of gradients
(complex-step).
[2] The covariance matrix is singular or near-singular, with condition
number 8.93e+24. Standard errors may be unstable.
The root means the squared error is 658.8184177799501.
```



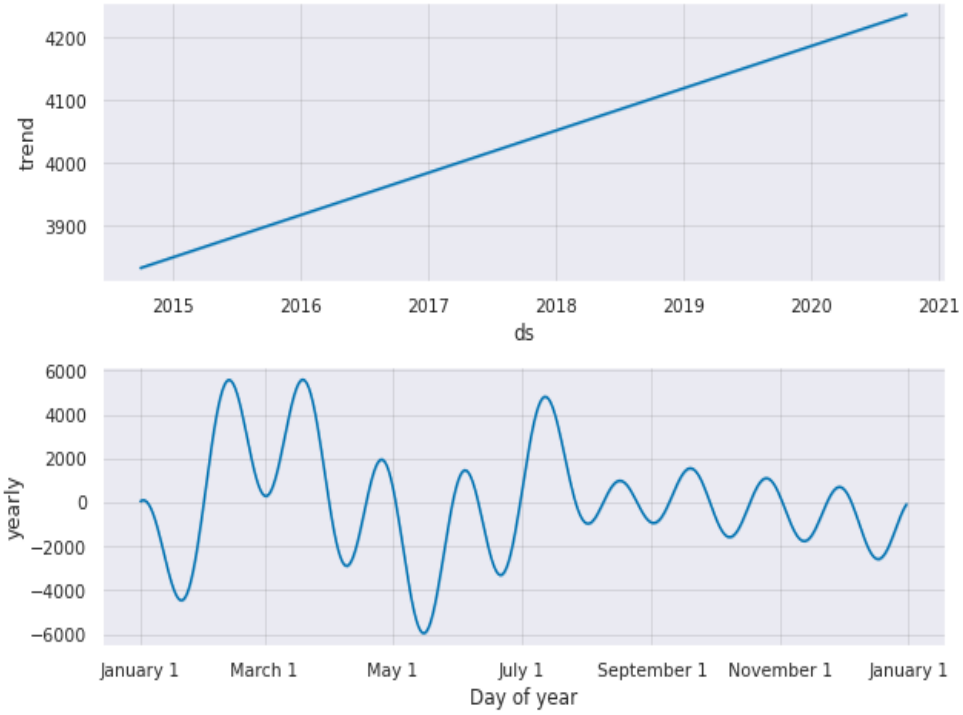
IV. SARIMAX RESULTS

FBProphet

Fbprophet is a python library that can also be used for the development of time series. Here we have considered the "weekly_seasonality=True" and "daily_seasonality=True"



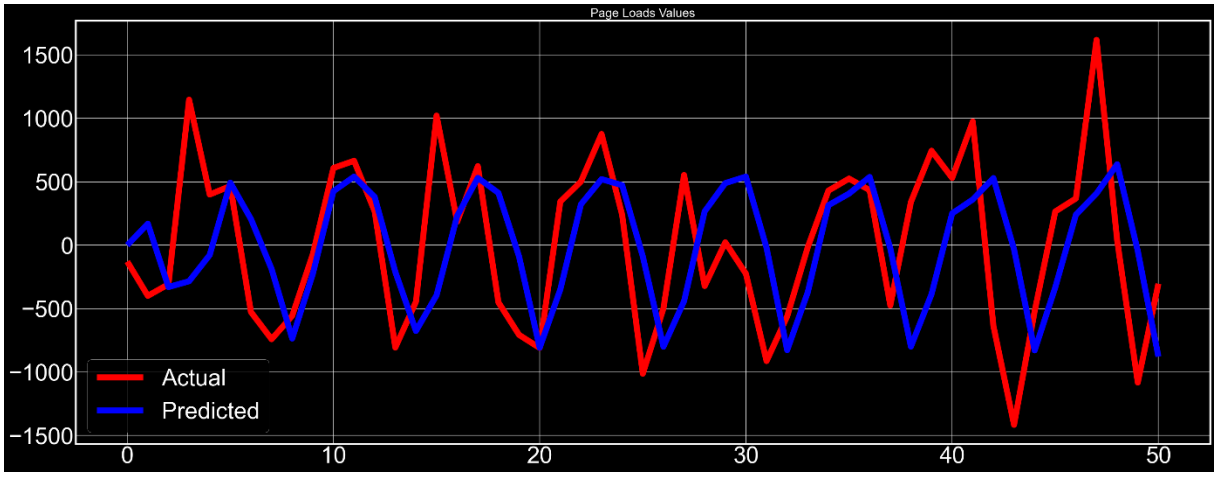
Plot components of Fb Prophet



Forecasting :

	ds	yhat	yhat_lower	yhat_upper
127	2020-10-26	5330.834419	4751.788556	5925.022366
128	2020-10-27	5273.023309	4683.035709	5854.614992
129	2020-10-28	5173.757456	4570.216846	5780.155457
130	2020-10-29	5035.992618	4443.000401	5633.666262
131	2020-10-30	4863.888660	4288.295257	5446.763848

Final Predict :



4. Required Resources:

Software used:
Jupyter Notebook, Anaconda Navigator

Hardware used:

Parameter	Minimum Configuration
CPU Frequency	2.30 GHZ
Number of CPU Cores	2
RAM	26.75GB

Table 5: Required Resources

Please Note:
The above configuration is chosen based on the basic classical algorithms used in this problem statement.

Programming Language used: Python
Important Python Libraries:

1. Data Visualization and Feature Engineering:
Pandas, Numpy, Matplotlib, Seaborn
2. Time Series: statsmodels
3. Data Modelling: Sklearn

5. Research Plan
Please find below Gantt Chart explaining the plan and timeline:


Gantt_Chart_Web_s
ite_visitors_2014-202

V. CONCLUSION

For all simple and centred moving averages, last n observations are treated equally. All observations before that are ignored. In some scenarios, all of the past data must be given gradual weightage. Maximum weightage should be given to the most recent data while minimum weightage should be given to the least recent data or vice-versa.

The introduction of ARIMA and SARIMAX helped us to identify the actual trends and seasonality in such a way that all the observations are considered. The **SARIMAX** is used as seasonal ARIMA that can be used for non-stationary data as well.

Future Work
Time Series Applications can be created in future which involve the mentioned below steps :

1) Creation of .pickle file as a stored as a binary data
2) Creation of web files or front page
3) Required python libraries
4) Procfile creation

REFERENCES

[1]. Arvind K. Sharma, P.C. Gupta , Identifying the Number of Visitors to improve Website Usability from Educational Institution Web Log Data
[2]. Jiaranai Awichanirost & Naragain Phumchusri , Analyzing The Effects of Sessions on Unique Visitors and Unique Page Views with Google Analytics: A case study of a Tourism Website in Thailand
[3]. Aderibigbe Israel Adekitan, Jeremiah Abolade & Olamilekan Shobayo, Data mining approach for predicting the daily Internet data traffic of a smart-university