Web Traffic Forecasting

Time Series Analysis

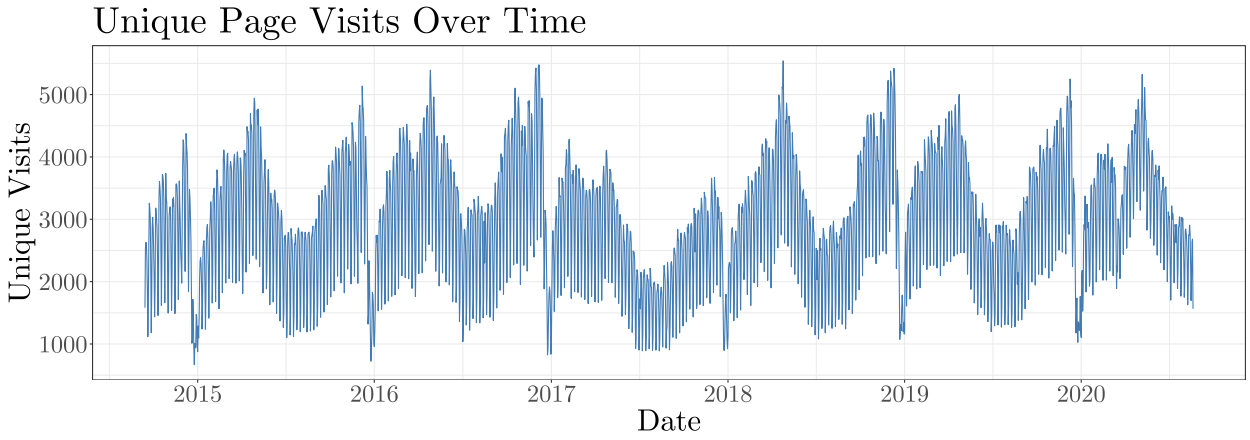
Shree Karimkolikuzhiyil; Programming; Statistics, M.S. (Distance); shreejesh@tamu.edu
Jingcheng Xia; Computations; Computer Science, B.S. (On-Campus); sixtyfour64@tamu.edu
Jackson Smith; Analysis; Statistics, M.S. (Distance); jackson.t.smith@tamu.edu
Samuel Burge; Writing; Statistics, M.S. (Distance); samuelburge@tamu.edu
Max Kutschinski; Theory; Statistics, M.S. (Distance); mwk556@tamu.edu

---

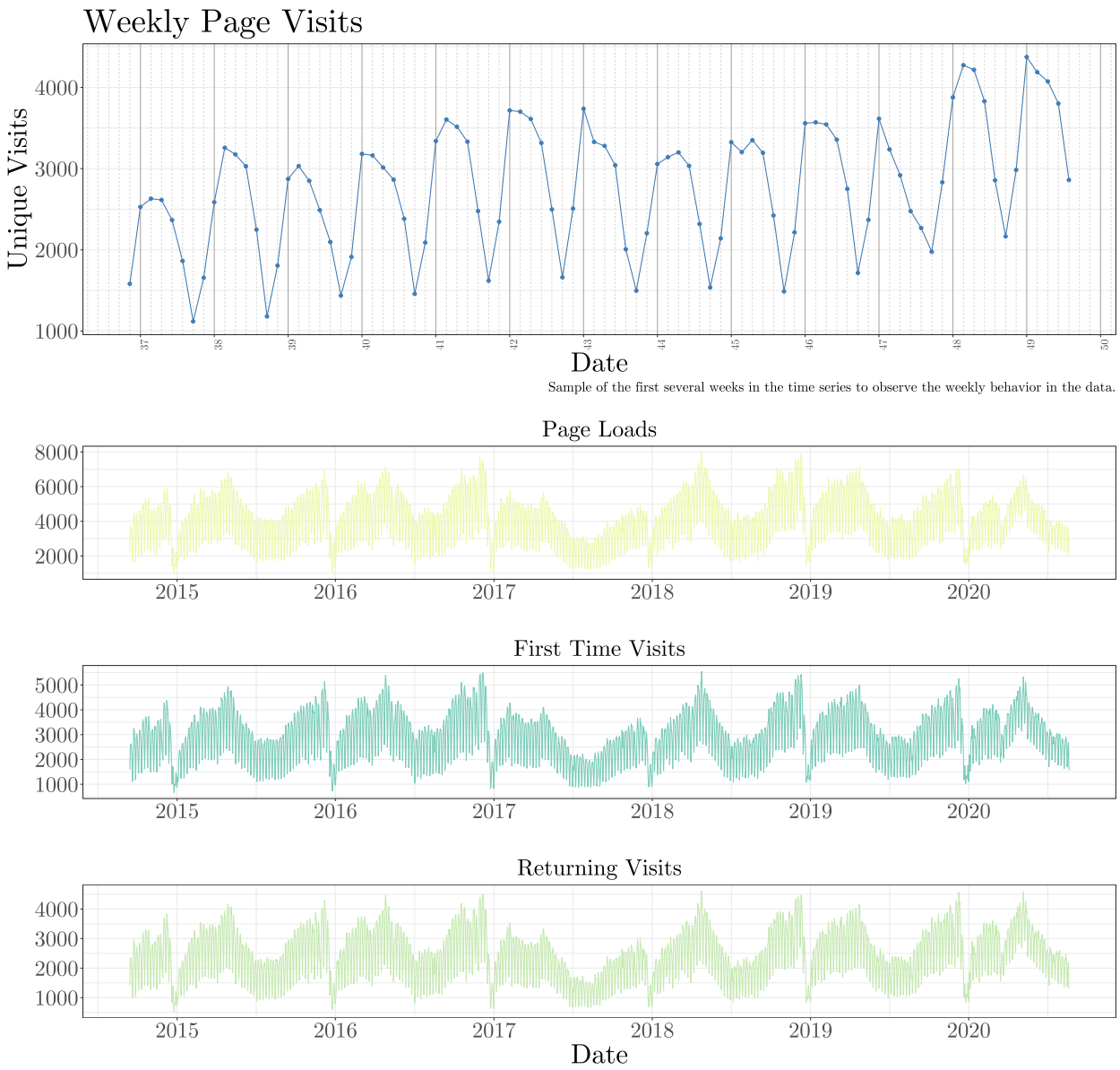
Introduction and Motivation

The objective of this analysis is to forecast daily unique visitors to an academic website containing lecture notes and supplemental material related to statistics. Predicting website traffic allows IT departments to manage project throughput and prioritize maintenance and enhancements to website functionality and effectively allocate web server resources. Web traffic is also a key indicator of customer growth and expansion, as well as sustaining recurring customers and ingrained growth. The details provided by web traffic throughput reports contain many metrics, including page loads, returning visitors, and unique visits, each of which convey a different picture and set of information for an organization. As well, having a picture of expected throughput and confirming (or denying) expectations with reality allows a business to understand unexpected growth and/or unexpected decay in business development.

The data contains five years of daily time series data of user visits. There are four features in the data set, which include daily counts for the number of page loads, first-time visitors, returning visitors, and unique visitors.[1] An initial plot of the data shows strong seasonality and volatility, but doesn't appear to have any discernible trend or cyclical behavior. An explanation for this could be due to the nature of the website. Students would likely be the largest share of users for a website of this nature, and the seasonality seems associated with the academic calender typically seen at academic institutions.

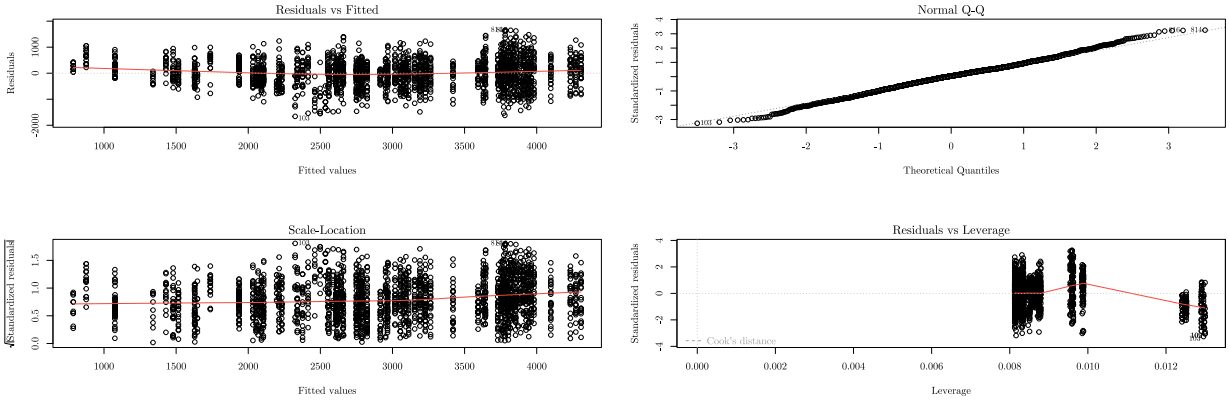


[1] A visit is defined as a stream of hits on one or more pages on the site on a given day by the same user within a 6 hour window, identified by the IP address of the specific device. Returning visitors are identified through allowed cookies on a user's device, and the total returning and first-time visitors is, by definition, the number of unique visitors.




```
##
## Call:
## lm(formula = UniqueVisits ~ DayOfWeek + MonthCat + SummerBreak +
##     WinterBreak, data = webData)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1658.46  -322.42    22.63   308.09  1657.20
##
## Coefficients: (1 not defined because of singularities)
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)      2749.94      57.23  48.050  < 2e-16 ***
## DayOfWeek2       1121.56      41.09  27.296  < 2e-16 ***
```

```
## DayOfWeek3       1195.50      41.09  29.095  < 2e-16 ***
## DayOfWeek4       1159.07      41.09  28.208  < 2e-16 ***
## DayOfWeek5        985.62      41.12  23.968  < 2e-16 ***
## DayOfWeek6        307.82      41.12   7.486 1.03e-13 ***
## DayOfWeek7       -552.87      41.12 -13.445  < 2e-16 ***
## MonthCat2        -141.63      63.88  -2.217   0.0267 *
## MonthCat3         -92.14      62.83  -1.467   0.1427
## MonthCat4         362.26      63.20   5.732 1.13e-08 ***
## MonthCat5          17.30      62.83   0.275   0.7830
## MonthCat6        -717.46      63.20 -11.352  < 2e-16 ***
## MonthCat7       -1121.59      62.83 -17.851  < 2e-16 ***
## MonthCat8       -1120.34      63.60 -17.616  < 2e-16 ***
## MonthCat9        -681.01      64.09 -10.625  < 2e-16 ***
## MonthCat10       -105.07      62.83  -1.672   0.0946 .
## MonthCat11         35.77      63.20   0.566   0.5714
## MonthCat12        -89.70      54.91  -1.634   0.1025
## SummerBreakTRUE       NA         NA      NA       NA
## WinterBreakTRUE -1320.41      54.94 -24.035  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 511.5 on 2148 degrees of freedom
## Multiple R-squared:  0.7286, Adjusted R-squared:  0.7264
## F-statistic: 320.4 on 18 and 2148 DF,  p-value: < 2.2e-16
```



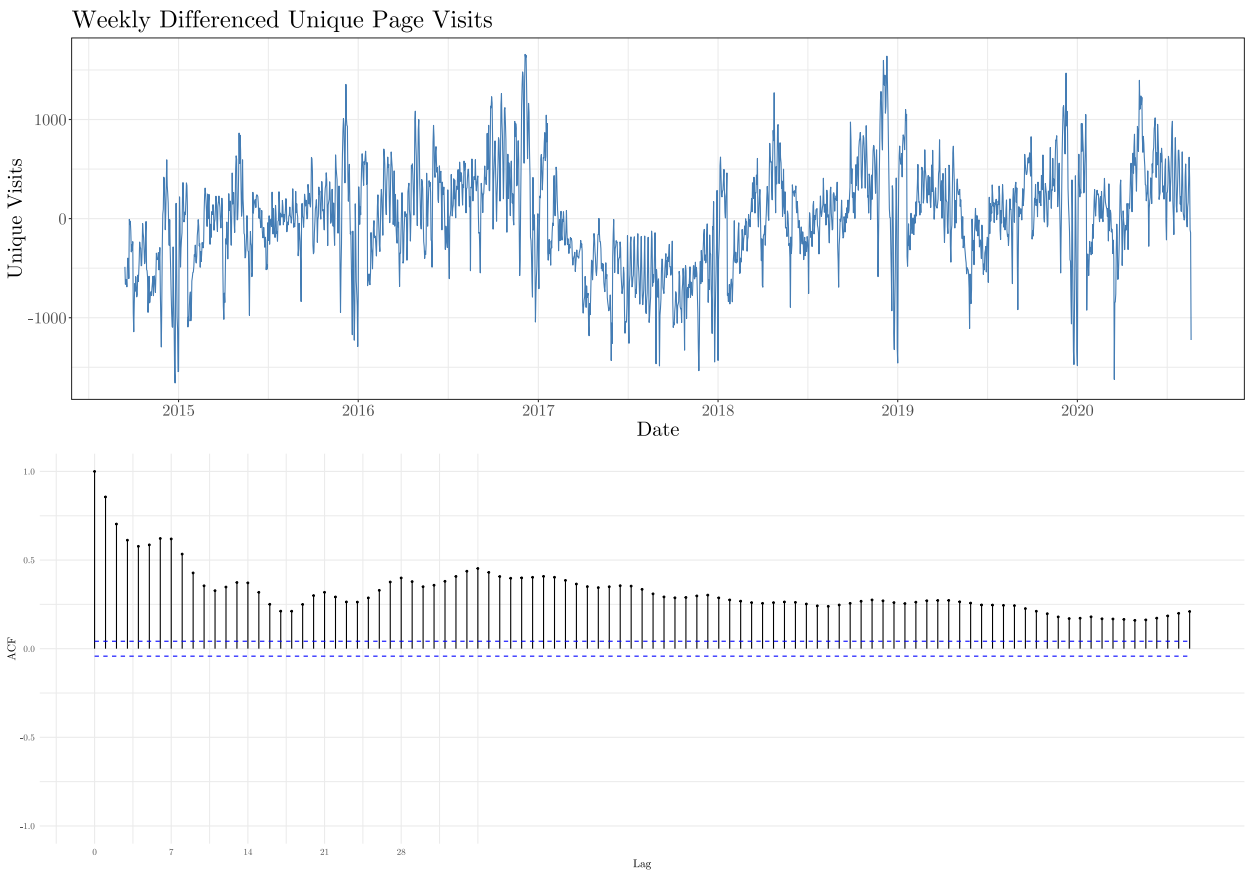


Figure 3: Sample autocorrelation function (ACF) of differenced time series with 30 lags, showing the weekly seasonality (and associated autocorrelation) for weekly time periods.