MASTER OF SCIENCE IN DATA SCIENCE
UNIVERSITY *of* WASHINGTON

Wikipedia Article Web Traffic Time Series Analysis

Forecasting future views of Wikipedia pages by total traffic, types, and topics

Created Date: June 8, 2022

Authored By: Aaliyah Hänni

Table of Contents

Overview

Wikipedia is a free online encyclopedia that has over 6 million articles accessed by users all over the world (Wikimedia Foundation, 2022). This project outlines the exploration of web traffic for a sample of approximately 145,000 Wikipedia articles over a two year period. The work below outlines a forecast for the weekly number of page views over time, and explores potential relationships between page views, traffic types and article topics.

Background

Personal Motivation

The initial motivation for this analysis was to obtain more experience working with large datasets and to compete in my first Kaggle Competition. This specific Kaggle Competition is closed, but Kaggle members are still able to create their own submission to obtain a personal score on how well their forecasting predictions compare to others. Experience competing in a competition using a large dataset, will allow me to grow my professional portfolio of projects that I have completed.

Forecasting and Modeling

The ability to develop a model that forecasts web traffic for a website has many benefits related to networking and resource allocation. Knowing if there are trends or seasonality in website traffic, allows web administrators to allocate the needed bandwidth to optimize performance, while minimizing costs.

Identifying Relationships

Identifying patterns in page views can provide interesting sociological information about potential trends or seasonality in the topics that people are interested in. This can be the beginning of motivation for additional research in why people are interested in specific topics. For example, in the article, ‘Why do people search Wikipedia for information on multiple sclerosis?’, researchers found peaks in google searches for multiple sclerosis correlated with celebrities mentioning it on television, or when new treatments were released (Brigo, et. al, 2018).

Project Goals

1. Wikipedia Article Page Views Over Time

   Using daily web traffic information for a sample of Wikipedia articles, we will explore several key properties, such as identifying potential trends or seasonality in the total number of page views of varying articles over time. We will then generate forecast predictions of web traffic for the different articles, and compare predictions against the true number of page views.

2. Wikipedia Article Page Views Over Time by Traffic Type

   For a given Wikipedia article site visit there are three types of traffic that Wikipedia measures: mobile, desktop, and spider. Mobile and desktop traffic are site visits from individual users through different interfaces, while spider traffic is access to a site from a web-crawler that was developed to collect data. For this analysis, we will explore differences between the types of traffic in Wikipedia articles over time, and identify the types of relationships that exist between them.

3. Wikipedia Topics Page Views Over Time

   There are natural groupings of Wikipedia articles by topic, such as articles about flowers, sports, or historical events. For this analysis, we will explore the page traffic over time for varying Wikipedia topics. To determine a Wikipedia article topic, we will develop a simple classification model to group articles based on tokenized article titles. Then we will explore relationships in the aggregated sum of page views over time for the varying topics. The topics were determined using Wikipedia's defined WikiProject Lists of topics (source: https://en.wikipedia.org/wiki/Wikipedia:WikiProject_Lists_of_topics), which are:

   - Art and culture
   - Geography and places
   - Health and fitness
   - History and events
   - Mathematics and abstractions
   - Natural sciences and nature
   - People and self
   - Philosophy and thinking
   - Religion and spirituality
   - Social sciences and society
   - Technology and applied sciences

Data

The data set used for this analysis will be the Web Traffic Time Series Forecasting provided by Google Inc, and hosted on Kaggle (Google Inc., 2017). This dataset contains the daily count of unique user visits to 145,063 Wikipedia articles from July 1st, 2015 to September 10th, 2017. These articles are in multiple languages, have varying grades (i.e article content assessment score), and contain multiple topics. The dataset also groups the counts of the number of articles accessed by each traffic type: all, mobile, desktop, and spider.

The data was aggregated to give weekly totals of traffic views, for all articles visited, as well by type. To obtain the article topics, a clustering algorithm was applied on the article titles to group them into 1 of 11 defined Wikipedia topics of Art and culture, Geography and places, Health and fitness, History and events, Mathematics and abstractions, Natural sciences and nature, People and self, Philosophy and thinking, Religion and spirituality, Social sciences and society, and Technology and applied sciences

Limitations

The dataset used for this analysis does not differentiate between missing data and zero page views. To overcome this limitation, careful considerations need to be made for how to handle missing data, such as by performing missing value imputations or removing specific rows of data with too many missing values, as it could skew analysis.

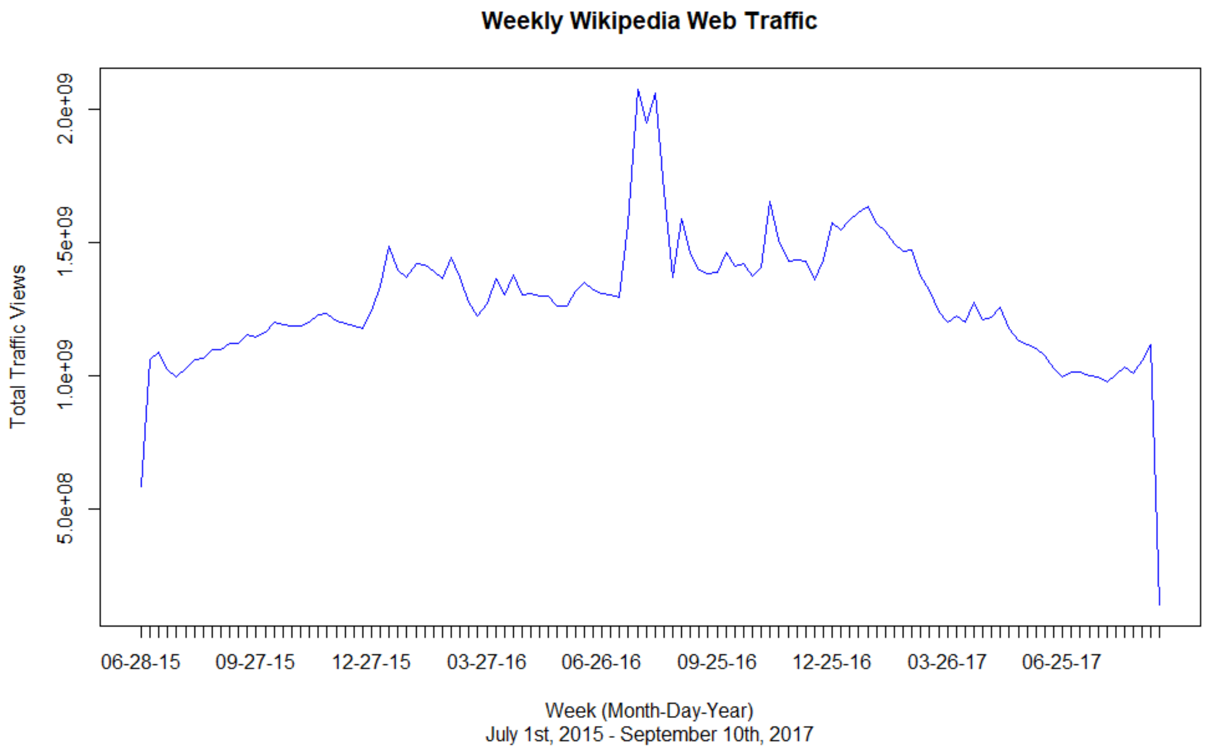
Another concern about the data is the overall size of the dataset and lack of needed computation power. Opening the dataset in excel caused the program to crash several times. If this is an issue for analysis, a potential solution is to limit the number of websites used by random selection.

Generating the Wikipedia article topics by using the title generates another set of potential biases and issues. GIven that the titles are in different languages, it will make generalizations of the topic difficult to verify. Also, not all groupings of topics may be intuitive or ideal. A potential way to overcome this problem, is to incorporate an additional dataset that includes article summaries (Scheepers, 2018), and perform the classification algorithm on article summary rather than title.

Exploratory Analysis

1. Wikipedia Article Weekly Traffic

Initial Observations



From the time series above, we can see a significant decrease in the total views in both the beginning and ending weeks. This is likely due to problems with the data not differentiating between missing or NULL values, and zero page views. Since these events only occur in the beginning and ending of the time series, it is likely due to how the data was collected.

There is a peak in the total traffic around late summer of 2016, and this is potentially due to the November 2016 US presidential election, which was very controversial and resulted in large views in pages of the candidates and other political topics included in this data set.

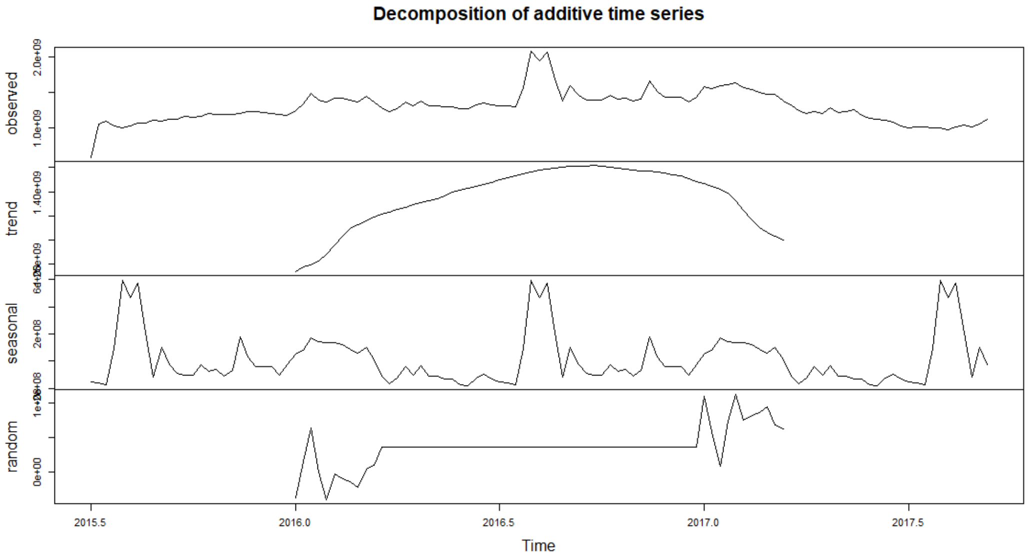
Evaluating Stationarity

```
> kpss.test(weekly_ts, null = "Trend") #p < 0.05

        KPSS Test for Trend Stationarity

data:  weekly_ts
```

```
KPSS Trend = 0.44291, Truncation lag
parameter = 4, p-value = 0.01

> kpss.test(diff(weekly_ts, lag = 1), null = "Trend") #p > 0.05

        KPSS Test for Trend Stationarity

data:  diff(weekly_ts, lag = 1)
KPSS Trend = 0.040709, Truncation lag
parameter = 4, p-value = 0.1
```

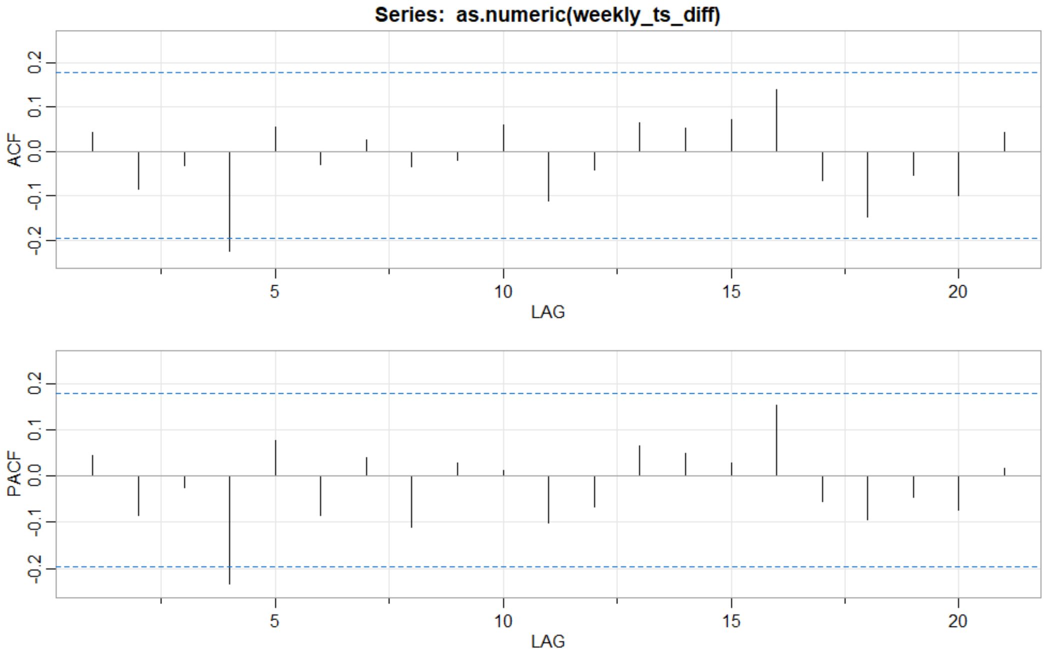
Using a Kwiatkowski–Phillips–Schmidt–Shin hypothesis test for stationary series, we observe that the p value is less than the significance threshold of α = 0.05. Thus, we reject the null hypothesis that the time series is stationary. Reapplying the stationary test on the difference set, we obtain a p-value greater than our significance level, and thus there is significant evidence to conclude that the differenced set is stationary.

Investigate Seasonality



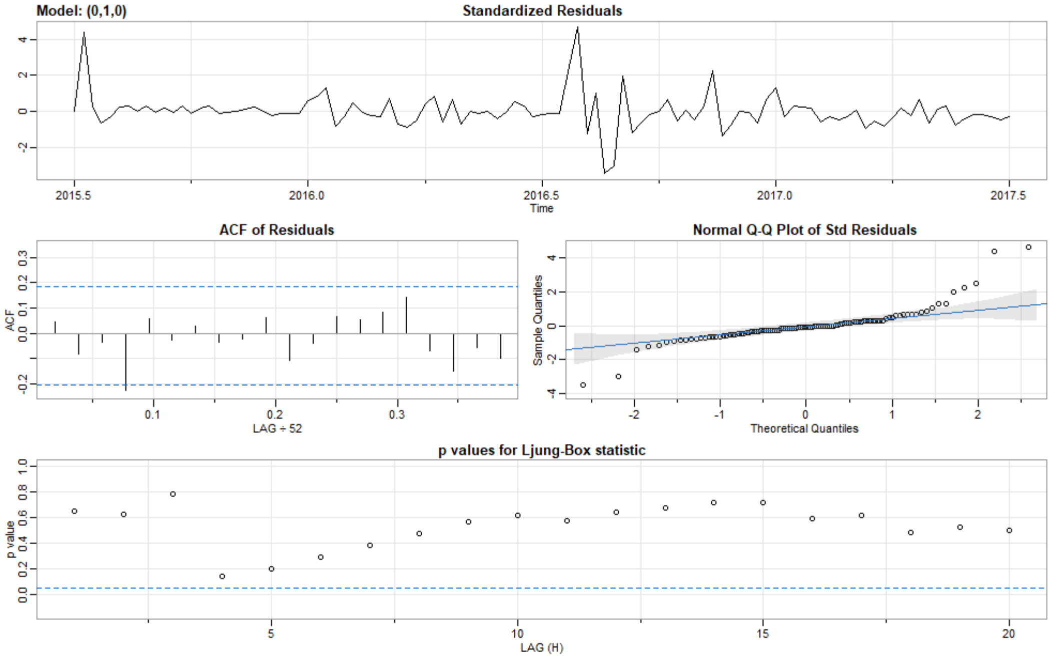
From the plots above, we can see that there does not appear to be any strong seasonality in the time series. Reviewing the repetitions in the seasonality, it does appear that it is largely being impacted by the extreme points observed in the data, such as the low views in the beginning and the peak in mid 2016.

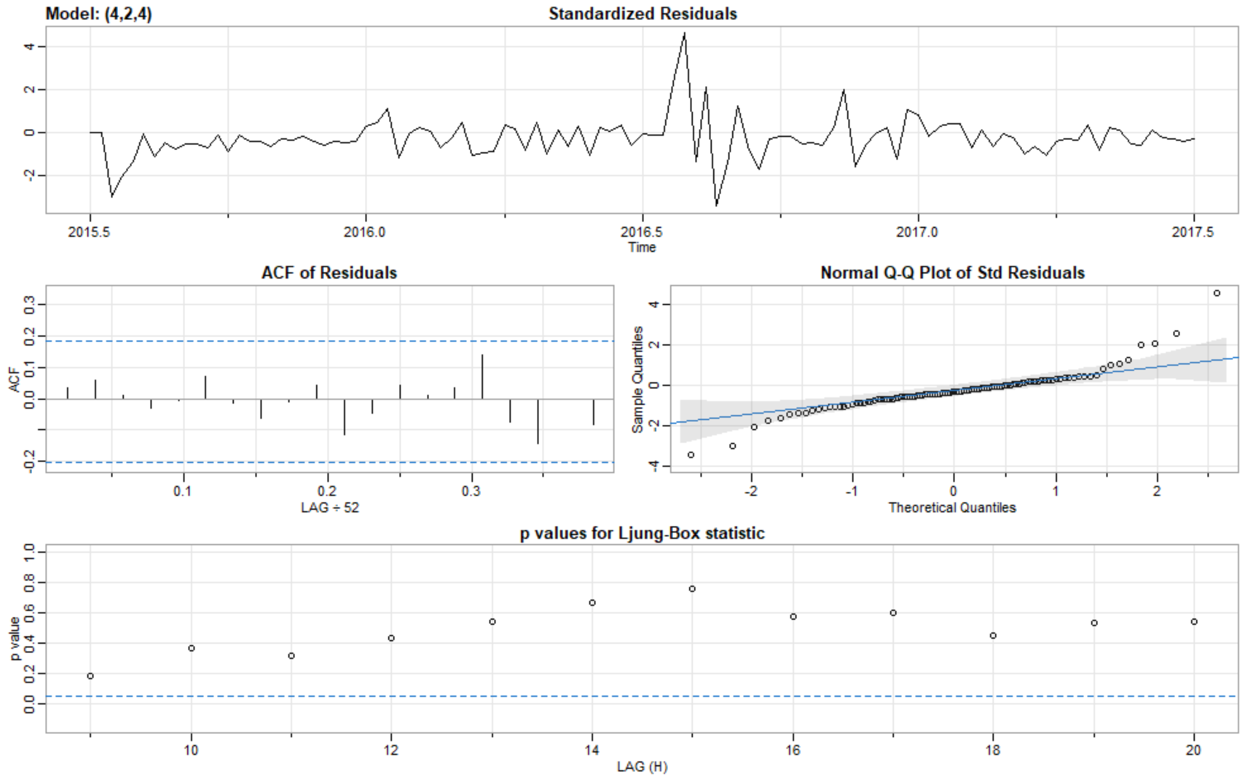
ACF/PACF



In the ACF and PACF, we see a significant lag at lag 4, which is a multiple of our weekly seasonality, meaning that we will likely need to include a moving average and autoregressive seasonality term in our ARIMA model.

ARIMA Modeling




Model: (4,2,4)

Standardized Residuals

ACF of Residuals

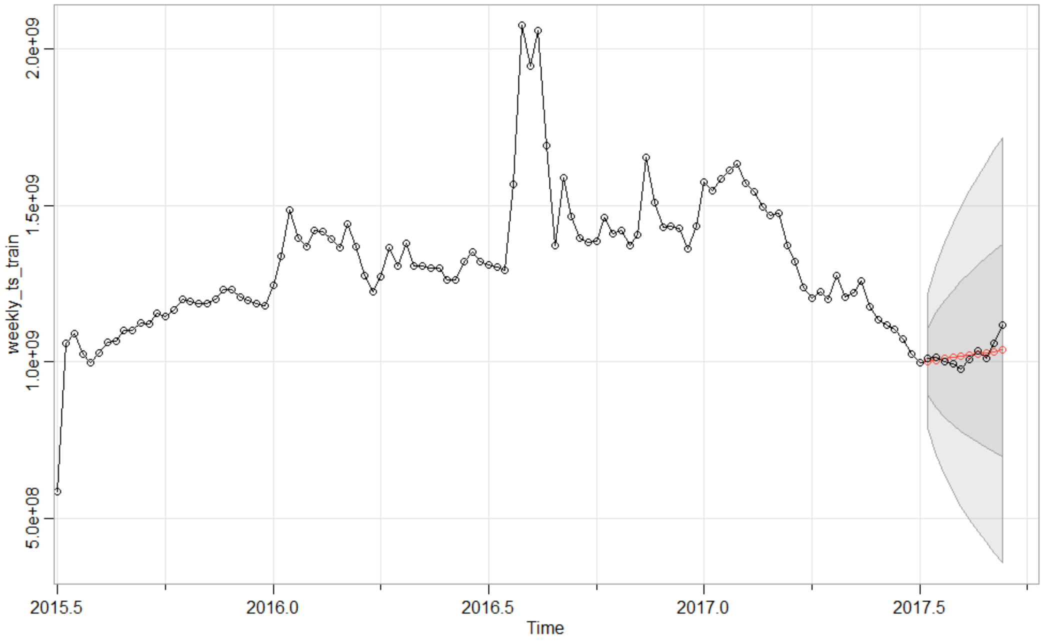
Normal Q-Q Plot of Std Residuals

p values for Ljung-Box statistic

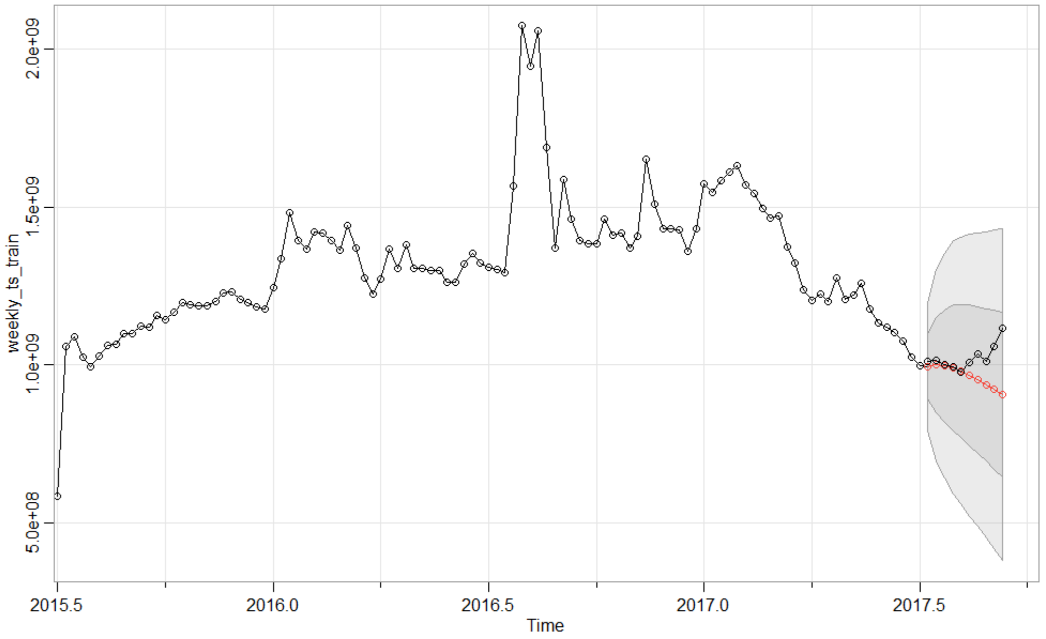
The auto ARIMA function predicts that the best model is ARMIA(0,1,0). Reviewing the residual plots of this ARIMA model, we obtain a random ACF of residuals, significant p-values for the Ljung-Box, and a Q-Q plot that suggest some normality of standard residuals.

In contrast, including a lag of 4 to both the moving average and autoregressive terms, as suggested by the ACF/PACF plots, we obtain a beautifully random ACF and standardized residuals, all of the p-values in the Ljung-Box are significant, and we observe slightly better normality of standardized residuals in our Q-Q plot.

ARIMA(0,1,0)



ARIMA(2,4,2)



Comparing the models generated above, for the ARIMA(0,1,0) model, the test values appear near the predicted line and all values are within a single standard error of our predictions. In contrast, the ARIMA(4,2,4) model appears to be a poor fit towards the end of the predictions.

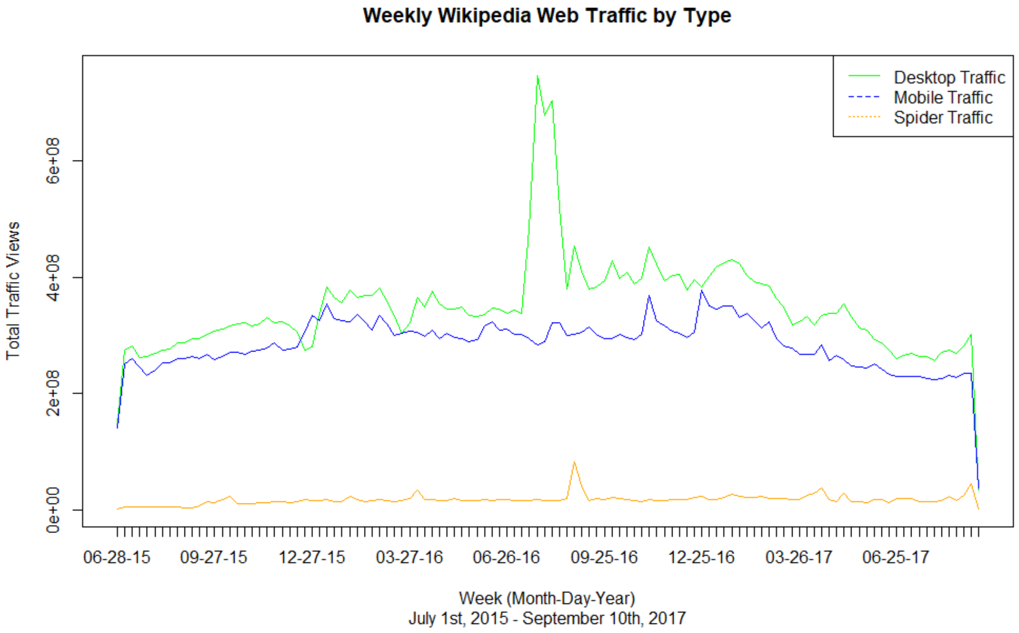
```
>#Model 1
> weekly_ts_for1$pred
Time Series:
Start = c(2017, 28)
```

```
End = c(2017, 37)
Frequency = 52
 [1] 1001362729 1005327674 1009292620 1013257565 1017222511 1021187456
1025152402 1029117347 1033082293
[10] 1037047238

> weekly_ts_for1$se
Time Series:
Start = c(2017, 28)
End = c(2017, 37)
Frequency = 52
 [1] 107162581 151550775 185611035 214325162 239622815 262493643 283525539
303101550 321487743 338877835

> accuracy(weekly_ts_for1$pred, weekly_ts_test)
              ME     RMSE      MAE       MPE    MAPE      ACF1 Theil's U
Test set 3615929 31413006 23344146 0.2509568 2.23801 0.2998092  1.065328


> #Model 2
> weekly_ts_for2$pred
Time Series:
Start = c(2017, 28)
End = c(2017, 37)
Frequency = 52
 [1] 993362519 999284725 997319712 991696392 981115775 966706314 952718703
936435344 921508956 905996596

> weekly_ts_for2$se
Time Series:
Start = c(2017, 28)
End = c(2017, 37)
Frequency = 52
 [1] 101199111 151239973 177904913 200405153 211123325 223129750 232404413
241821040 252619102 262255153

> accuracy(weekly_ts_for2$pred, weekly_ts_test)
               ME     RMSE      MAE     MPE     MAPE      ACF1 Theil's U
Test set 58206409 88142731 58899814 5.48109 5.552016 0.5634913   3.00091
```

For the first model we obtain a ME = 3,615,929, RMSE = 31,413,006, MAE = 23,344,146.
For the second model we obtain a ME = 58,206,409, RMSE = 88,142,731, MAE = 58,899,814.
From these evaluation metrics we can see that the first model ARIMA(0,1,0) performs significantly better than the second.

2. Wikipedia Article Weekly Traffic Types

Initial Observations



Although the large peaks and valleys are consistent between the three types of web traffic, there is a large magnitude in range and variance of each traffic type. Desktop is the largest web traffic, with few exceptions, while the spider traffic is very minimal and consistently magnitudes of views lower than the other two traffic types.

In all three types, there exists the same patterns observed above in the overall traffic views, of steep declines in the beginning and ending of the time series, and a large peak around the middle of 2016.

Evaluating Stationarity

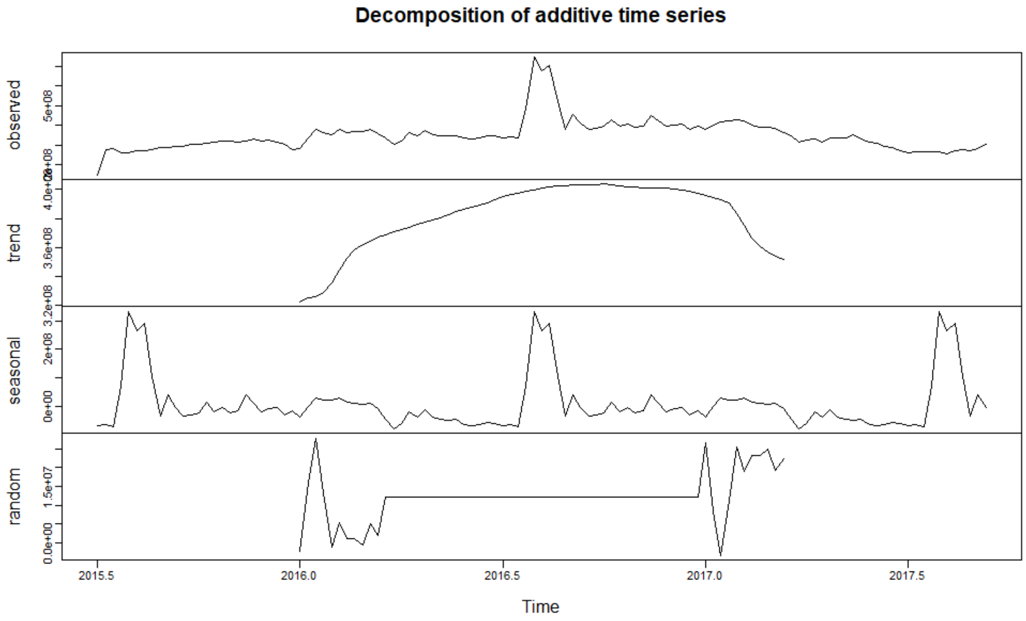
```
> #DESKTOP
> kpss.test(desktop, null = "Trend") #p < 0.05

        KPSS Test for Trend Stationarity

data:  desktop
KPSS Trend = 0.38568, Truncation lag parameter = 4, p-value = 0.01

> #check the difference set for stationary
> kpss.test(diff(desktop, lag = 1), null = "Trend") #p > 0.05

        KPSS Test for Trend Stationarity
```

```
data:  diff(desktop, lag = 1)
KPSS Trend = 0.026958, Truncation lag parameter = 4, p-value = 0.1


> #MOBILE
> kpss.test(mobile, null = "Trend") #p < 0.05

        KPSS Test for Trend Stationarity

data:  mobile
KPSS Trend = 0.43006, Truncation lag parameter = 4, p-value = 0.01

> #check the difference set for stationary
> kpss.test(diff(mobile, lag = 1), null = "Trend") #p > 0.05

        KPSS Test for Trend Stationarity

data:  diff(mobile, lag = 1)
KPSS Trend = 0.073952, Truncation lag parameter = 4, p-value = 0.1


> #SPIDER
> kpss.test(spider, null = "Trend") #p < 0.05

        KPSS Test for Trend Stationarity

data:  spider
KPSS Trend = 0.1844, Truncation lag parameter = 4, p-value = 0.02185

> #check the difference set for stationary
> kpss.test(diff(spider, lag = 1), null = "Trend") #p > 0.05

        KPSS Test for Trend Stationarity

data:  diff(spider, lag = 1)
KPSS Trend = 0.052361, Truncation lag parameter = 4, p-value = 0.1
```

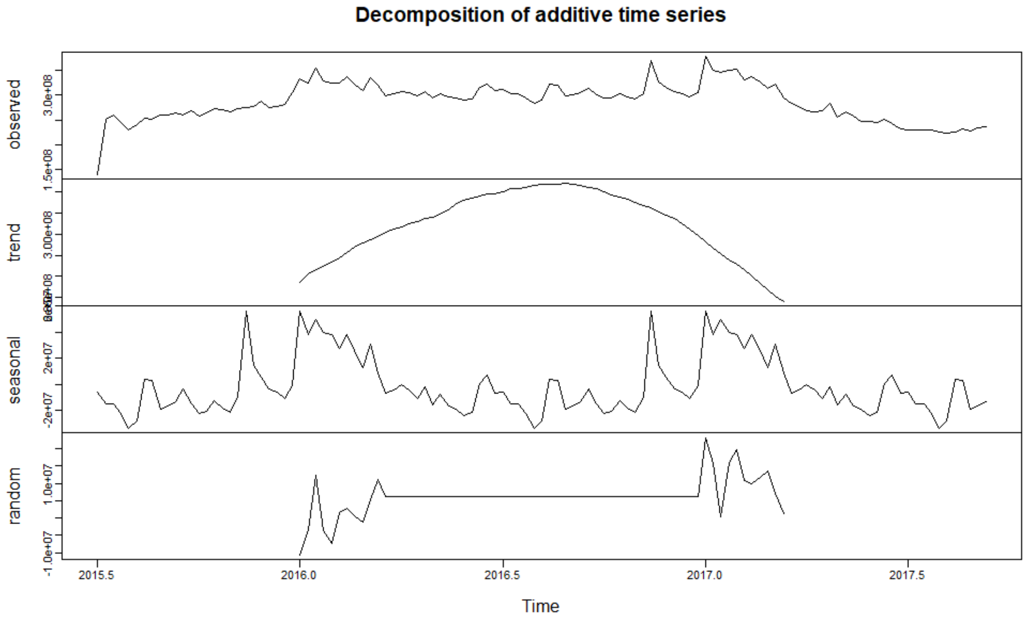
Using a Since Kwiatkowski–Phillips–Schmidt–Shin hypothesis test for all three stationary series, we observe that the p value is less than the significance threshold of $\alpha = 0.05$. Thus, we reject the null hypothesis that the time series are stationary. Reapplying the stationary test on the difference series, we obtain a p-value greater than our significance level, and thus there is significant evidence to conclude that the difference series are each stationary.

Investigate Seasonality

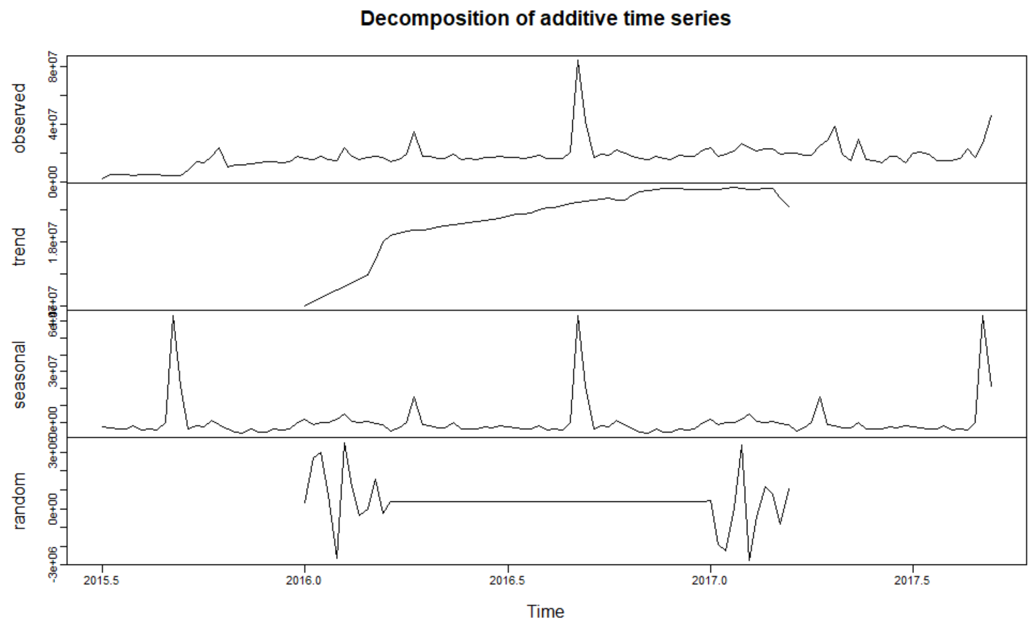
Desktop



Mobile



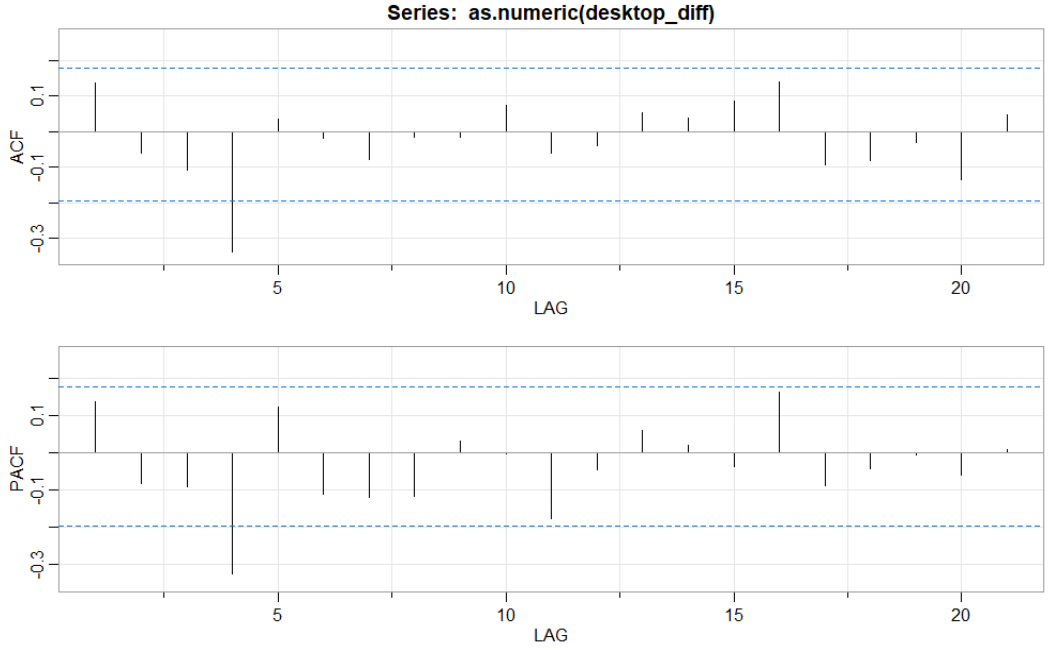
Spider



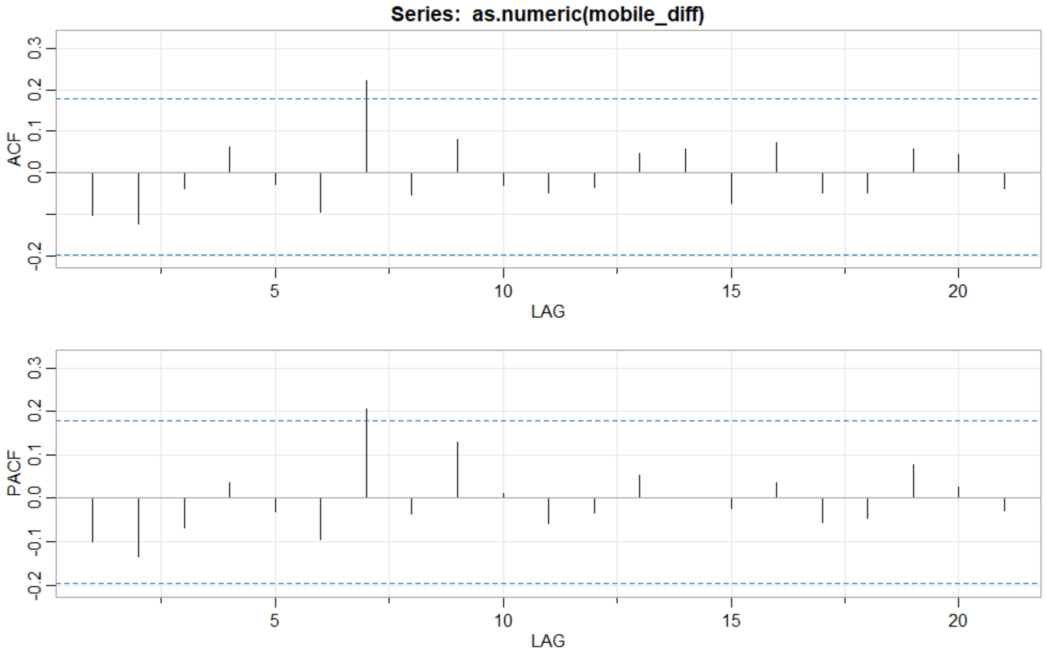
From the plots above, we can see that there does not really appear to be any seasonality. In the desktop traffic time series, the pattern is similar to the total traffic views, in that it is sensitive to the extreme observations. The mobile traffic shows a different pattern that is less sensitive to the extremes, but still nonseasonal. The spider traffic pattern appears to be the most seasonal, but still does not appear to show a strong consistent seasonality.

ACF/PACF

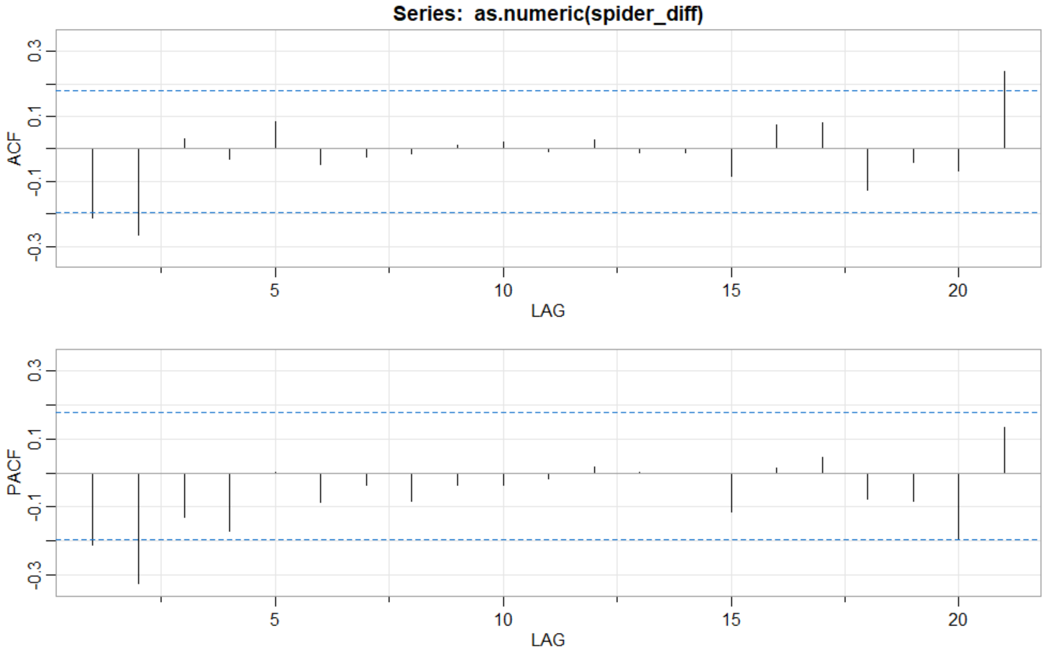
Desktop



Mobile



Spider



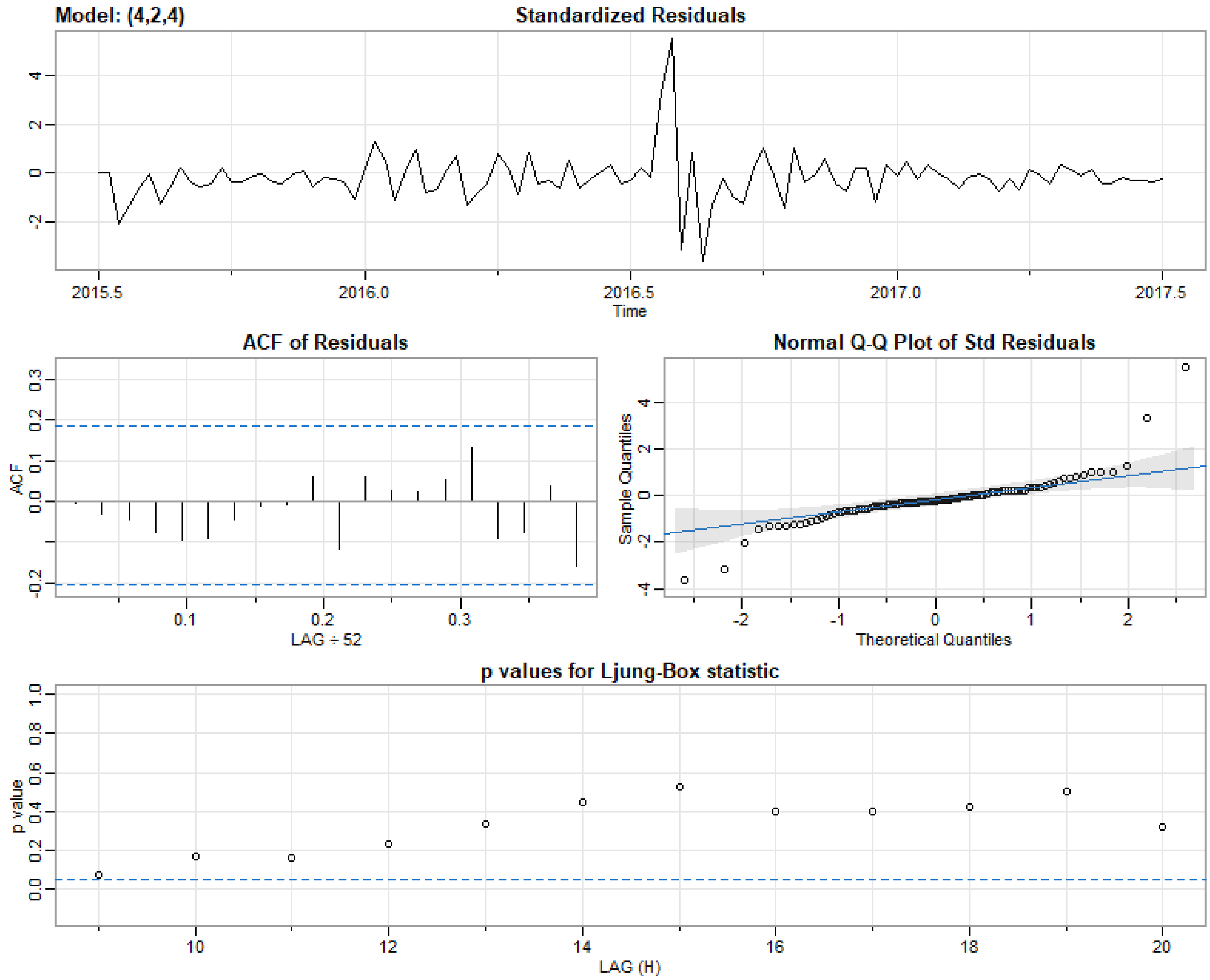
For desktop traffic the ACF and PACF shows a significant lag at lag 4, which implies we will likely need to include a moving average and autoregressive seasonality term in our ARIMA model.

For mobile traffic the ACF and PACF show a significant lag at lag 7, which implies we will likely need to include a moving average and autoregressive seasonality term in our ARIMA model.

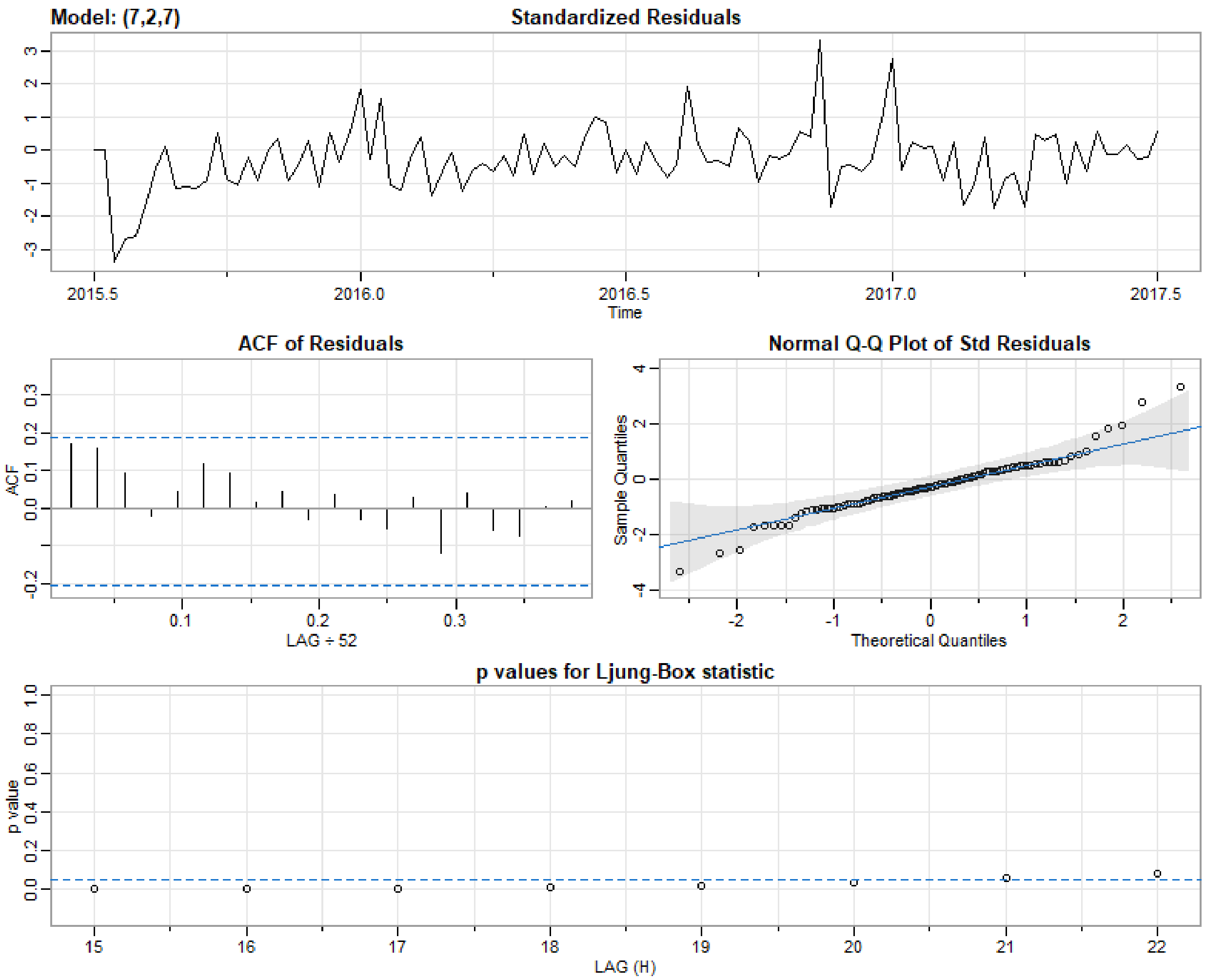
For spider traffic the ACF and PACF show a significant lag at lag 1 and 2, which implies we will likely need to include a moving average and autoregressive seasonality term in our ARIMA model.

ARIMA Modeling

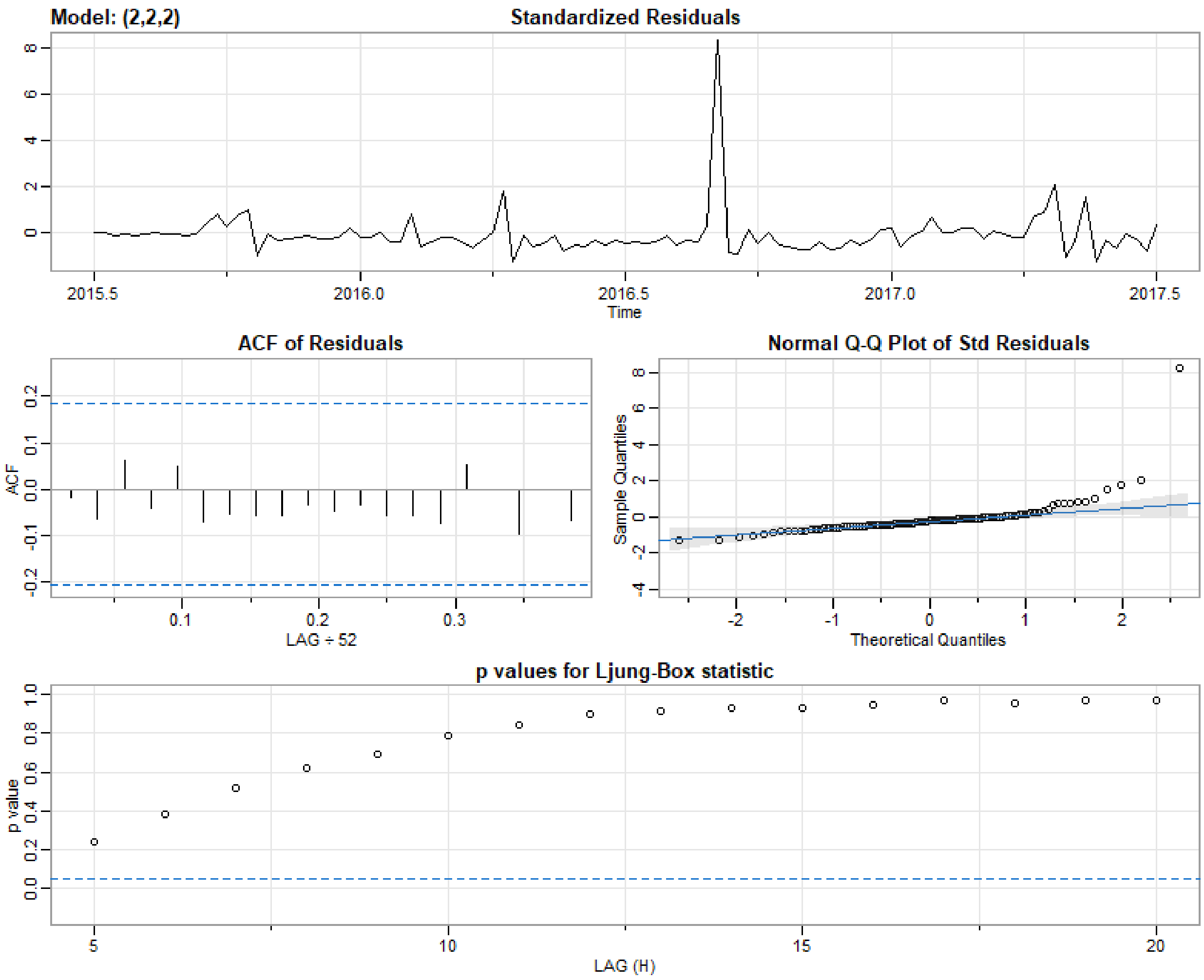
Desktop



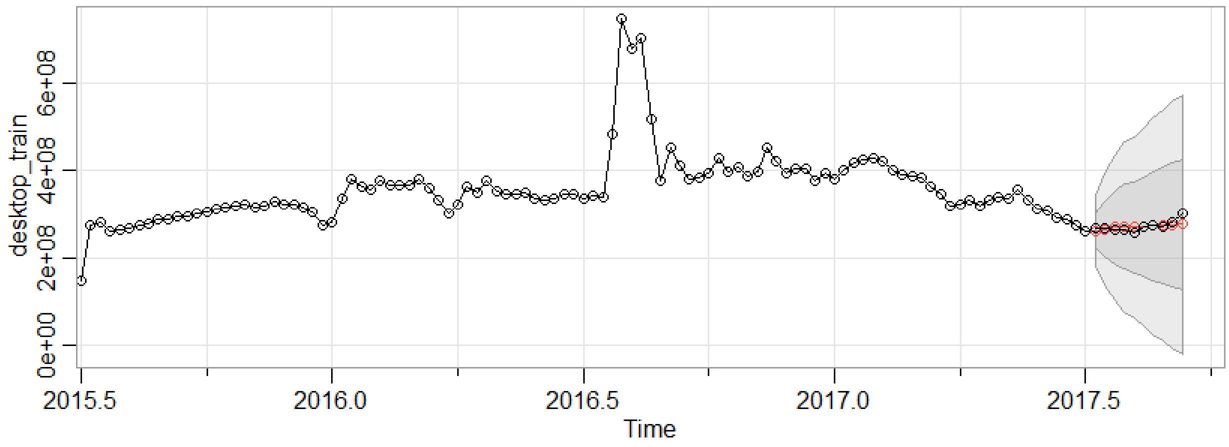
Mobile


Model: (7,2,7)
Standardized Residuals
3
2
1
0
-1
-2
-3
2015.5
2016.0
2016.5
2017.0
2017.5
Time
ACF of Residuals
ACF
0.3
0.2
0.1
0.0
-0.2
0.1
0.2
0.3
LAG ÷ 52
Normal Q-Q Plot of Std Residuals
Sample Quantiles
4
2
0
-2
-4
-2
-1
0
1
2
Theoretical Quantiles
p values for Ljung-Box statistic
p value
1.0
0.8
0.6
0.4
0.2
0.0
15
16
17
18
19
20
21
22
LAG (H)

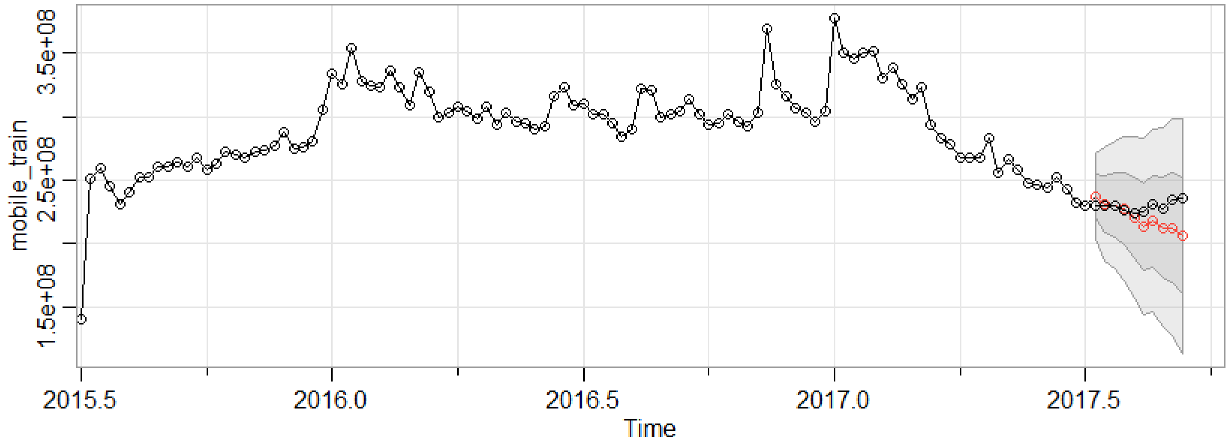
Spider



Desktop

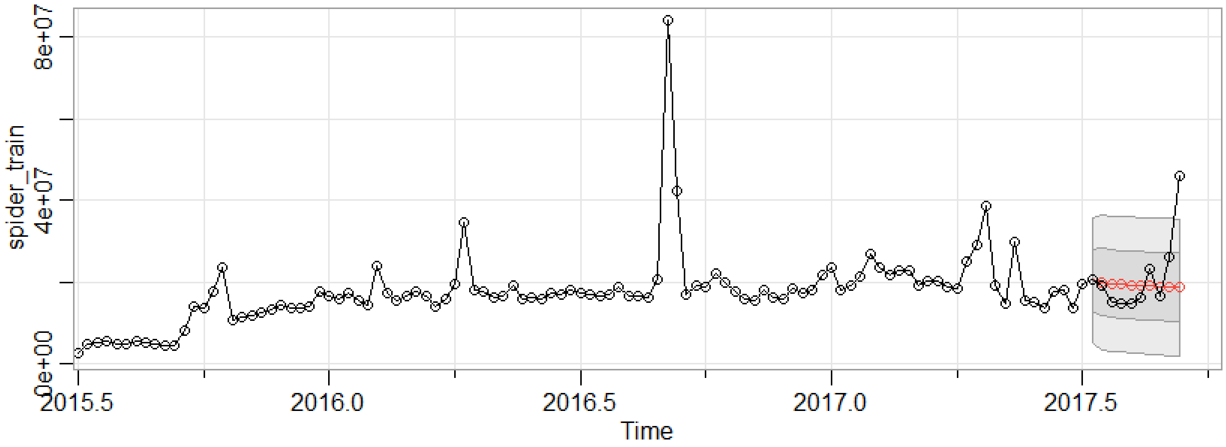


Mobile



Spider

```
> #DESKTOP
> accuracy(desktop_for$pred, desktop_test)
             ME    RMSE     MAE       MPE
Test set 1275340 9965231 7177156 0.3293039
            MAPE      ACF1 Theil's U
Test set 2.580173 0.2890519  1.064817

> #MOBILE
> accuracy(mobile_for$pred, mobile_test)
             ME     RMSE      MAE      MPE
Test set 8571501 14110709 10651123 3.692157
            MAPE      ACF1 Theil's U
Test set 4.597406 0.6293226  4.060513

> #SPIDER
> accuracy(spider_for$pred, spider_test)
             ME    RMSE     MAE       MPE
Test set 1989710 9519916 5897163 -1.960016
            MAPE      ACF1 Theil's U
Test set 23.43718 0.2783601  1.168997
```



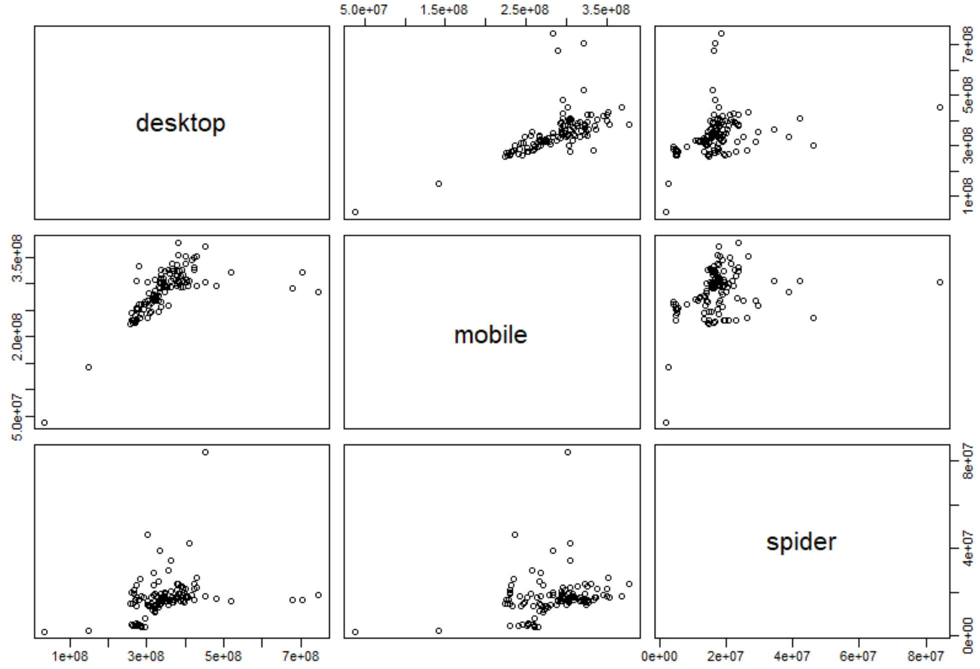
For the desktop model, visualizing the model predictions appears to be fairly accurate. Evaluating the error metrics we obtain the following: ME = 1,275,340, RMSE = 9,965,231, MAE = 7,177,156.

For the mobile model, the model visually appears to poorly estimate the long term shape of the date. Reviewing the error metrics we obtain the following: ME = 8,571,501, RMSE = 14,110,709, and MAE = 10,651,123.

For the spider model, the predictions fail to follow the pattern of the actual data, and even fall outside the two standard error ranges. The error metrics yield the following: ME = 1,989,710, RMSE = 9,519,916, and MAE = 5,897,163.

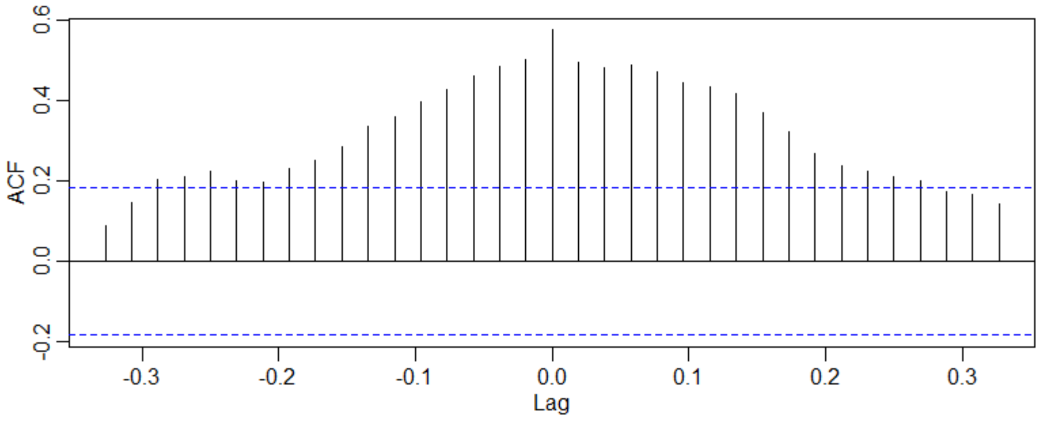
Multivariate Modeling

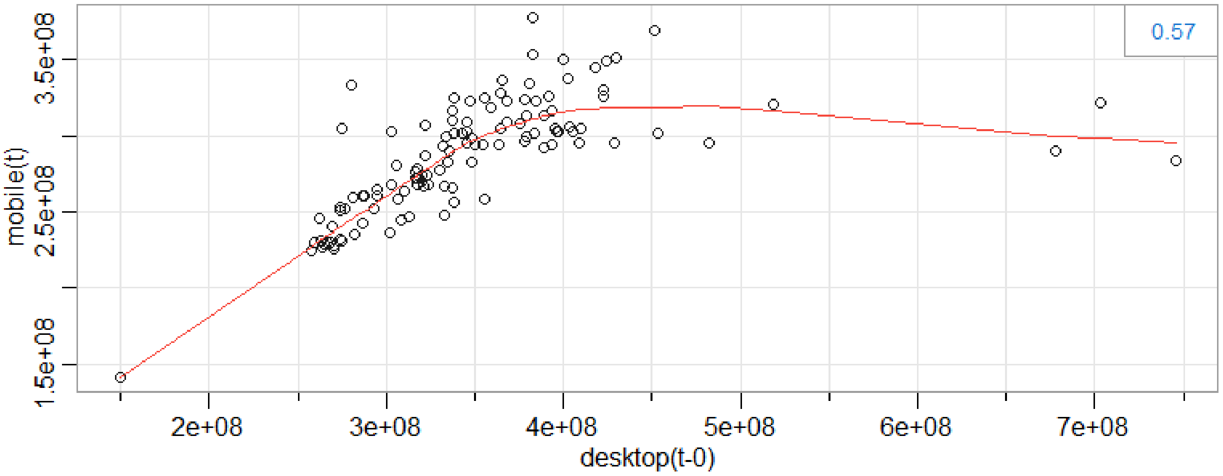
Wikipedia Type Page View Scatterplot Matrix

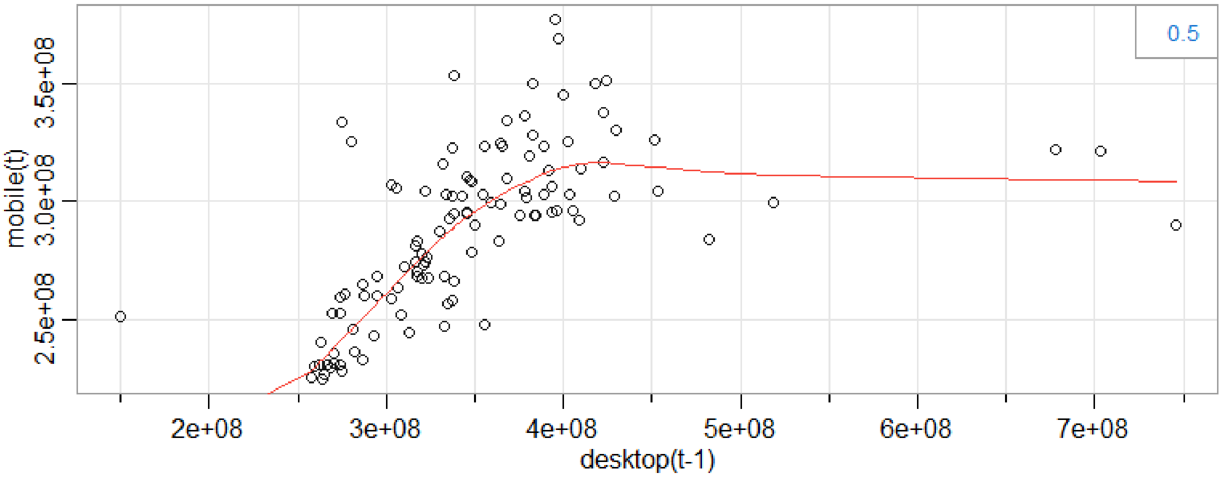


From the plot above, we observe that there appears to be a relationship between mobile and desktop. Exploring this in more detail, we will observe the correlations and develop a model to evaluate if the views in desktop can predict views in mobile.

Cross Correlation of Desktop and Mobile Page View







Given that the correlation coefficient is higher for the non lagged time series model, we will generate a forecasting model with 0 lagged variables.

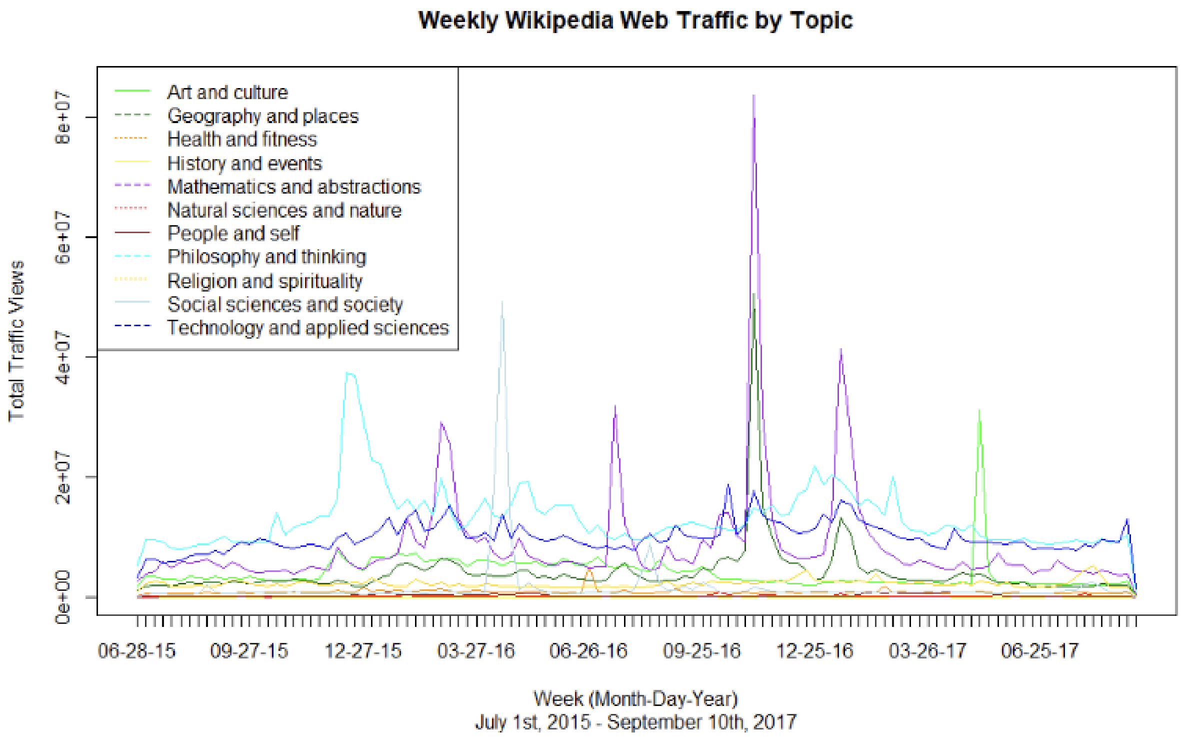
```
> summary(desktop_lm)

Call:
lm(formula = desktop ~ mobile, data = traffic_type_week)

Residuals:
      Min         1Q     Median         3Q        Max
-128936925  -27698535  -10682534    6426296  399106465

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -6.733e+06  4.049e+07  -0.166    0.868
mobile       1.245e+00  1.405e-01   8.857 1.26e-14 ***
---
Signif. codes:
```

```
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 66470000 on 114 degrees of freedom
Multiple R-squared:  0.4076,  Adjusted R-squared:  0.4024
F-statistic: 78.45 on 1 and 114 DF,  p-value: 1.258e-14
```

Since the p-value for the mobile variable is less than the threshold of 0.5, we determine that there is a significant relationship between mobile and desktop views. Given the low R-squared values, this model does not appear to be a good fit for predicting Desktop views.

3. Wikipedia Article Weekly Traffic Topics

Initial Observations



Given the amount of data and variation that is shown, it is difficult to observe any significant patterns. There does appear to be a somewhat consistent range of views depending on topic, with some peaks occurring in mathematics, technology, philosophy and art.

Evaluating Stationarity

```
> #check each type for stationary
> for (i in 1:length(ts)) {
```

```
+   print(topic_list[i])
+   print(kpss.test(ts[[i]], null = "Trend")$p.value)
+   #kpss.test(i, null = "Trend") #p < 0.05
+ }
[1] "Art and culture"
[1] 0.01977933
[1] "Geography and places"
[1] 0.0217544
[1] "Health and fitness"
[1] 0.01653109
[1] "History and events"
[1] 0.04868433
[1] "Mathematics and abstractions"
[1] 0.02184341
[1] "Natural sciences and nature"
[1] 0.1
[1] "People and self"
[1] 0.1
[1] "Philosophy and thinking"
[1] 0.03888734
[1] "Religion and spirituality"
[1] 0.1
[1] "Social sciences and society"
[1] 0.1
[1] "Technology and applied sciences"
[1] 0.01


> #check the difference set for stationary
> for (i in 1:length(ts)) {
+   print(topic_list[i])
+   print(kpss.test(diff(ts[[i]], lag = 1), null = "Trend")$p.value)
+ }
[1] "Art and culture"
[1] 0.1
[1] "Geography and places"
[1] 0.1
[1] "Health and fitness"
[1] 0.1
[1] "History and events"
[1] 0.1
[1] "Mathematics and abstractions"
[1] 0.1
[1] "Natural sciences and nature"
[1] 0.1
[1] "People and self"
[1] 0.1
[1] "Philosophy and thinking"
[1] 0.1
[1] "Religion and spirituality"
```

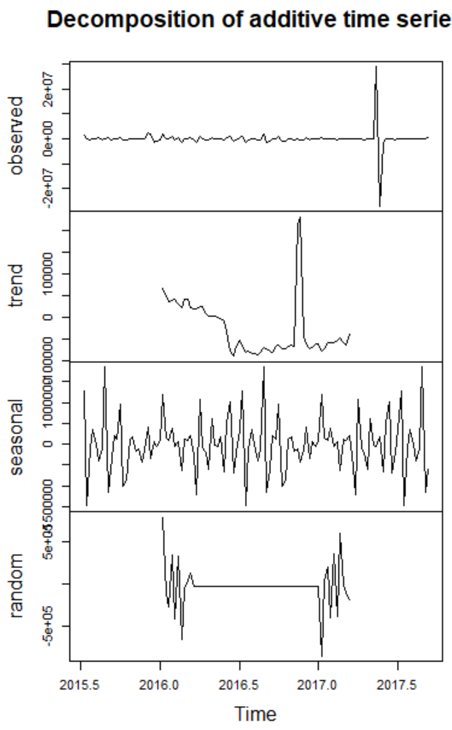
```
[1] 0.1
[1] "Social sciences and society"
[1] 0.1
[1] "Technology and applied sciences"
[1] 0.1
```

Using a Kwiatkowski–Phillips–Schmidt–Shin hypothesis test for stationary series, we observe a split in the two topics. In the first group, we find that the p-value > 0.5, and thus we conclude that the following topics are stationary: Natural sciences and nature, People and self, Religion and spirituality, Social sciences and society.

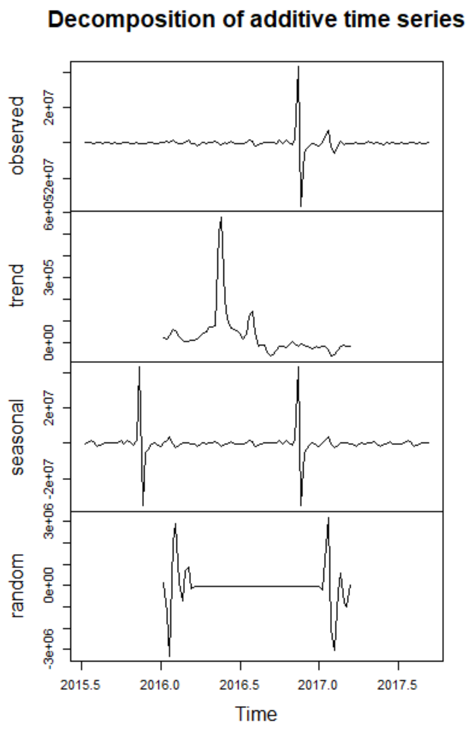
For the second group, we observe that the p-value < 0.5, thus apply a differencing and retest the stationarity and find that for the following topics there is significant evidence to conclude that the differenced time series are stationary: Art and culture, Geography and places, Health and fitness, History and events, Mathematics and abstractions, Philosophy and thinking, Religion and spirituality, Social sciences and society, Technology and applied sciences.

Investigate Seasonality

Art and culture

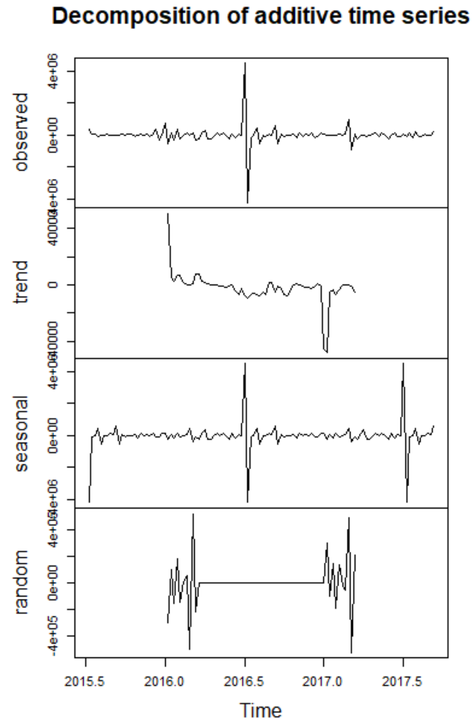


Geography and places

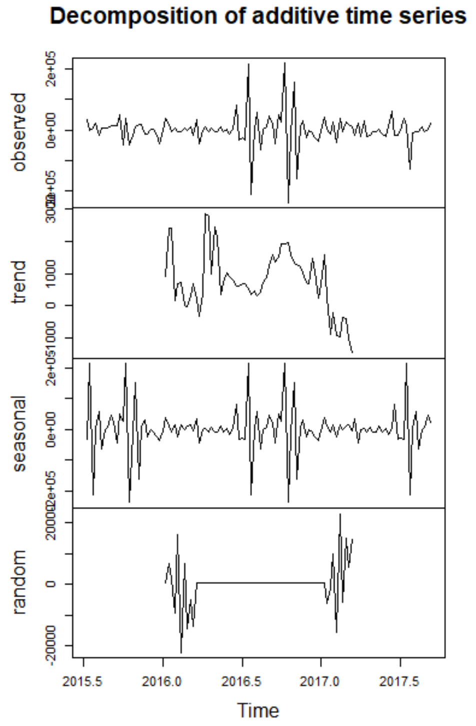


Health and fitness

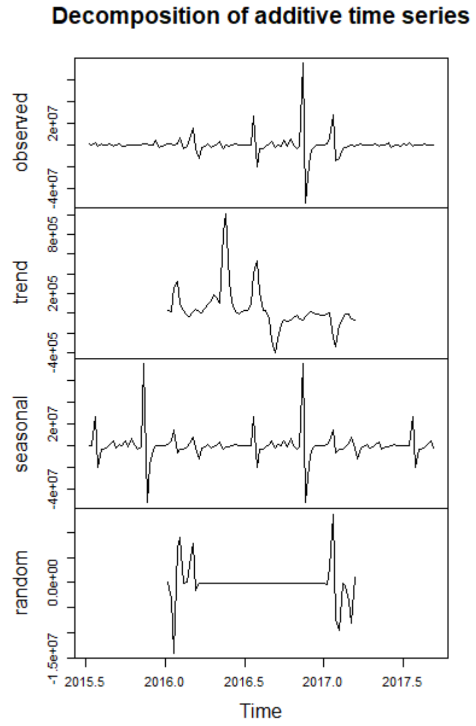
History and events



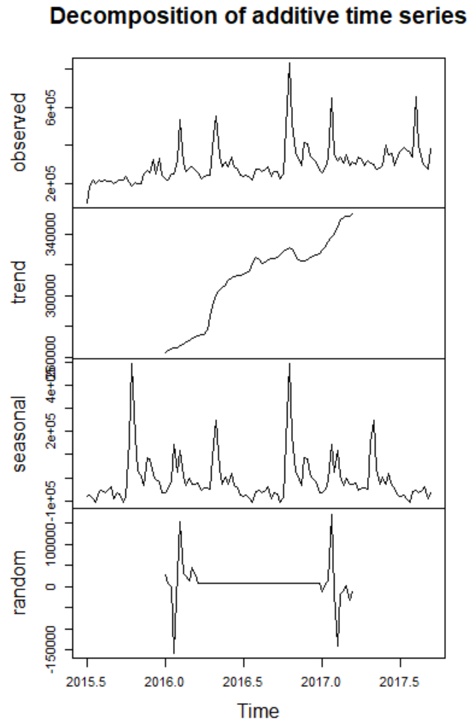
Mathematics and abstractions



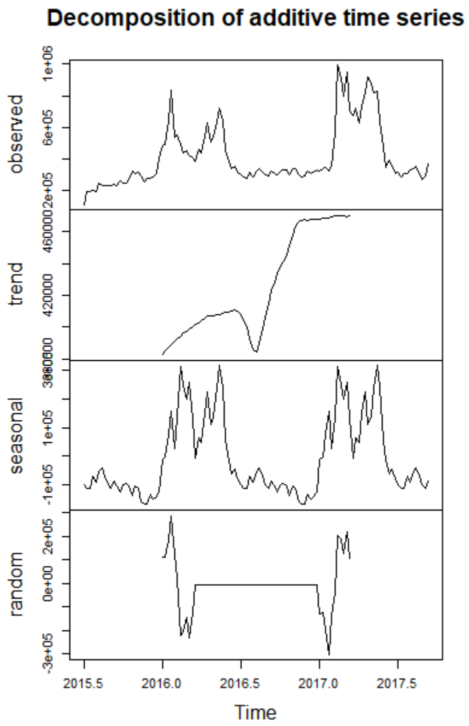
Natural sciences and nature



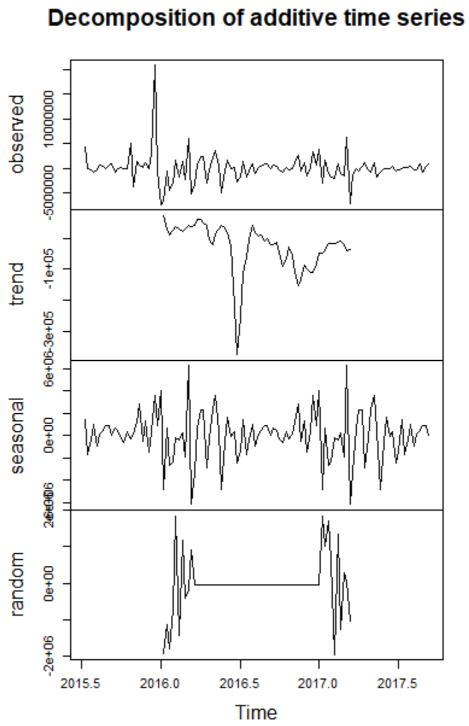
People and self



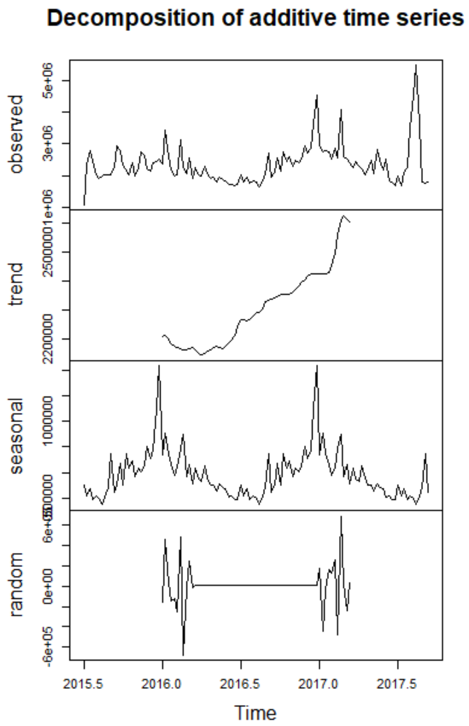
Philosophy and thinking



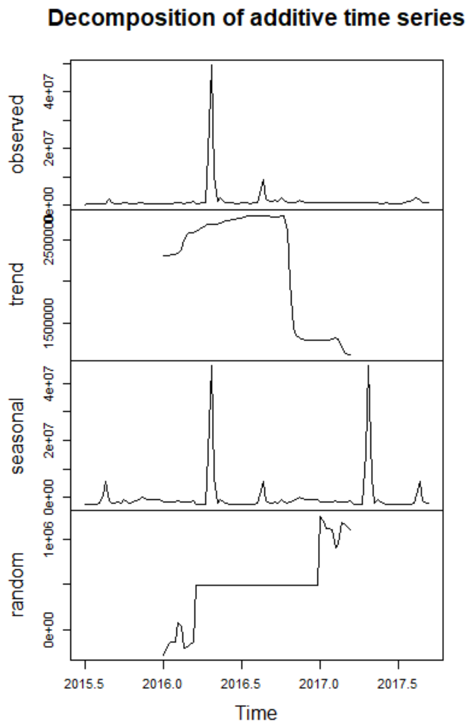
Religion and spirituality

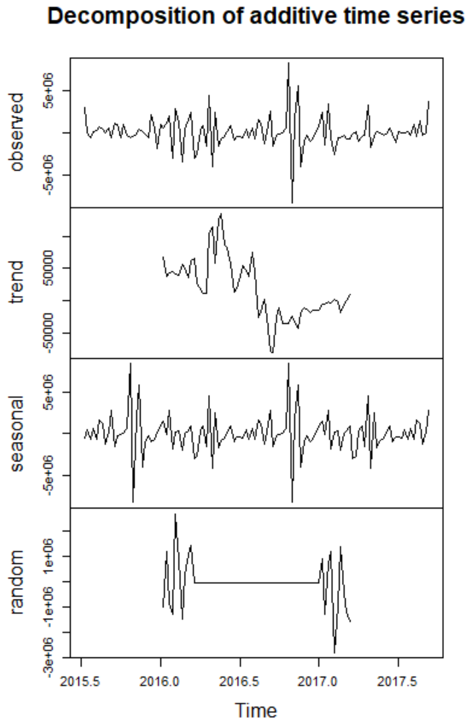


Social sciences and society



Technology and applied sciences





Using decomposition to observe if there is seasonality within the various time series, we observe another splitting between the two groups, of almost half that display strong evidence of seasonality, and half that do not appear to have any seasonality. Observing the graphs above, we split the topics into two groups:

Stationary Topics:

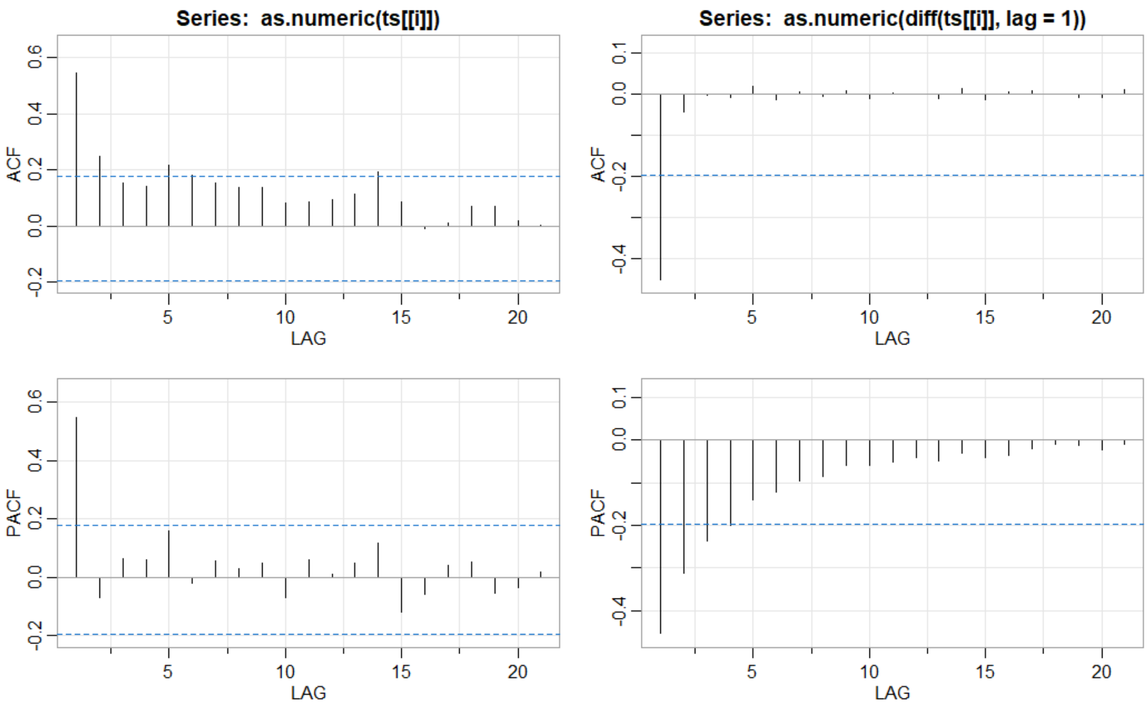
- History and events
- Mathematics and abstractions
- Natural sciences and nature
- People and self
- Philosophy and thinking
- Religion and spirituality

Non-stationary Topics:

- Art and culture
- Geography and places
- Health and fitness
- Social sciences and society
- Technology and applied sciences

ACF/PACF

Natural Science and nature **Arts and culture**



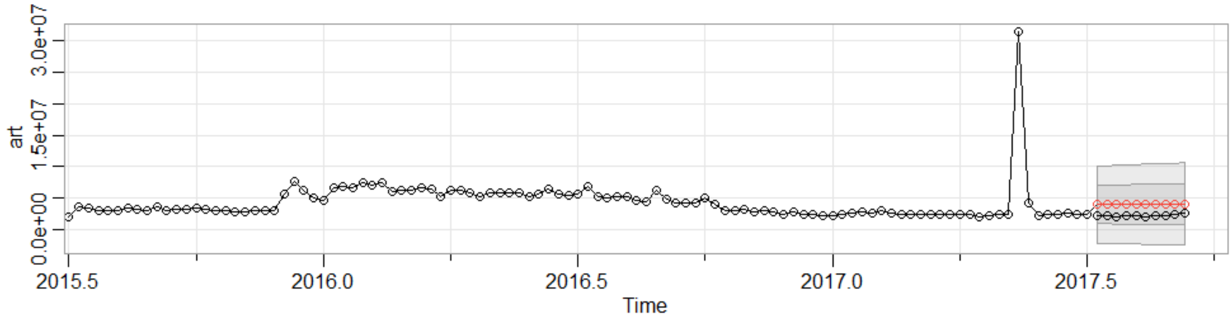
The ACF/PACF plots for the differenced and non-differenced topic series show a lot of variation in significant lags. The significant lags are noted for each topic below:

- Art and culture: ACF is sinusoidal 1- 10, PACF is 1, 2, 3
- Geography and places: ACF 1, PACF is sinusoidal 1, 2, 3
- Health and fitness: ACF 1, PACF 3
- History and events: ACF 1, 3, 4, PACF 2
- Mathematics and abstractions: ACF 1, 10, 16, PACF is sinusoidal 1, 2, 9, 15
- Natural sciences and nature: ACF 2, PACF 1
- People and self: ACF is descending 1-9, indicative that perhaps some stationary exists in plot, PACF: 1
- Philosophy and thinking: ACF 0, PACF 0
- Religion and spirituality: ACF 2, PACF 1
- Social sciences and society: ACF 1, PACF 2
- Technology and applied sciences: ACF 1, PACF 2

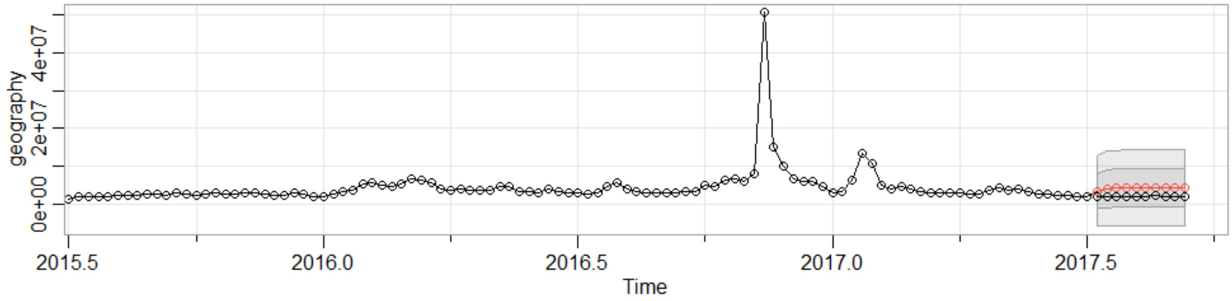
ARIMA Modeling

For simplicity, an auto ARIMA function was applied to extract the best models for all 11 topics. The below list summarizes the best ARIMA model estimated for each topic.

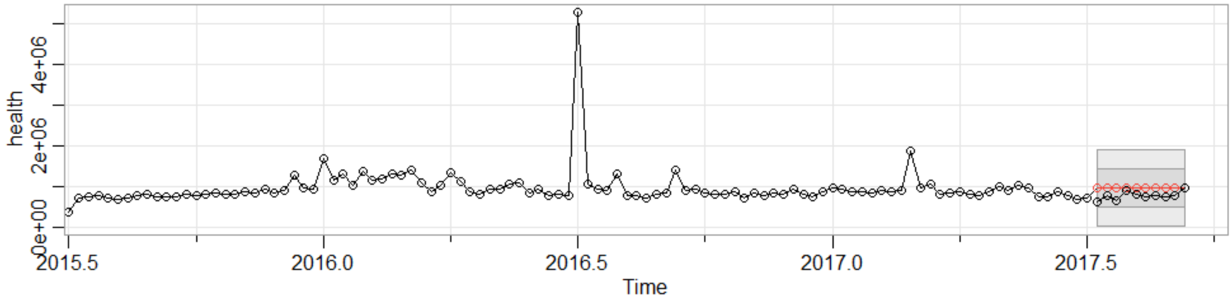
- Art and culture: ARIMA(0,1,1)



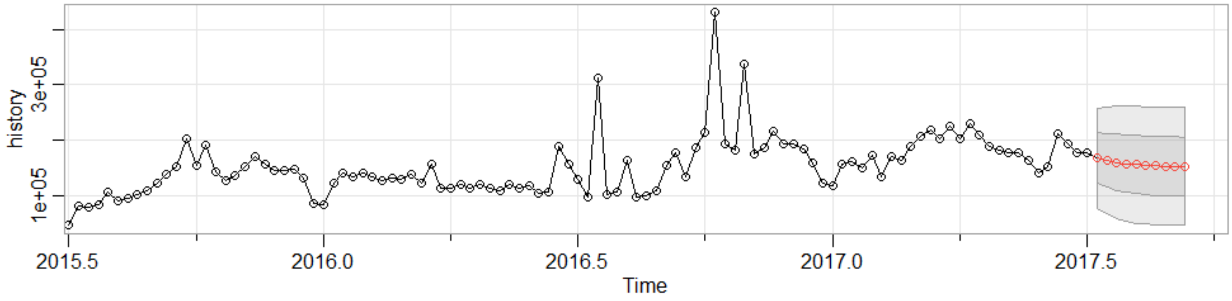
- Geography and places: ARIMA(1,0,0)



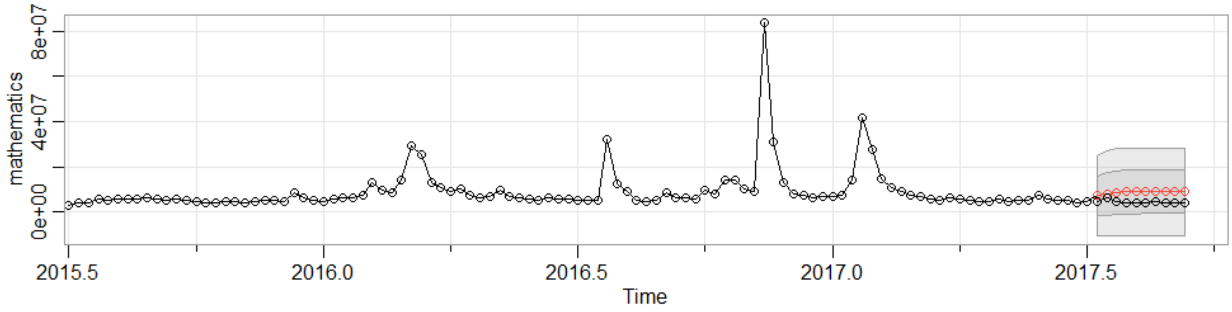
- Health and fitness: ARIMA(0,0,0)



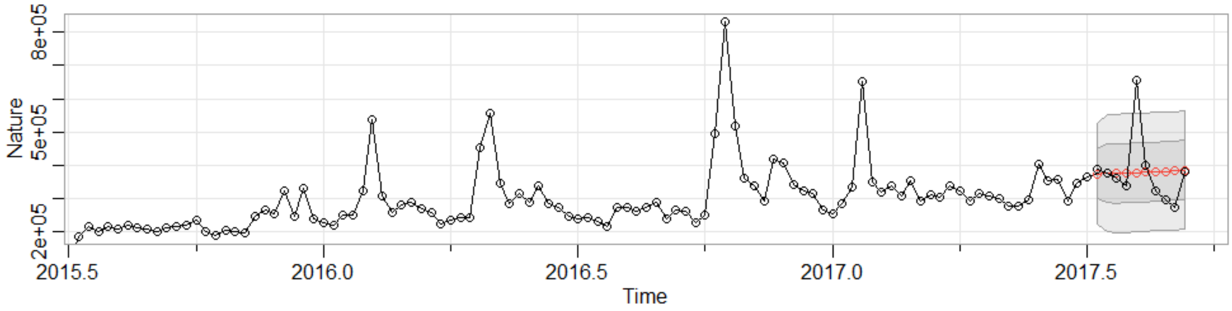
- History and events: ARIMA(2,0,0)



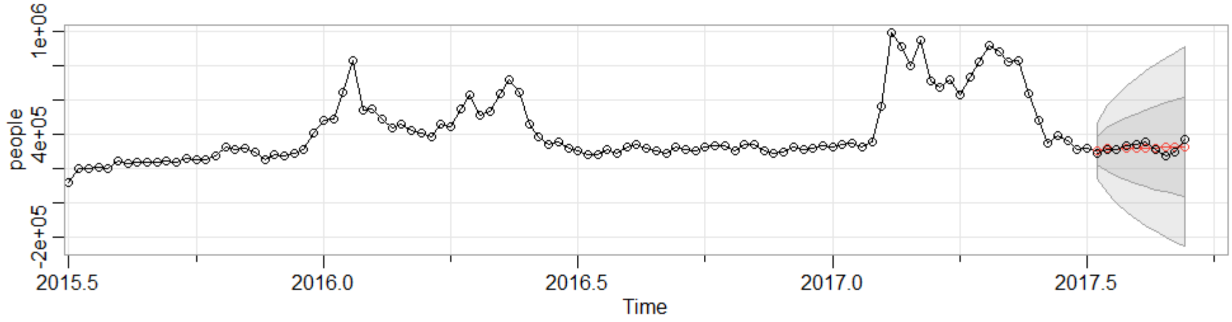
- Mathematics and abstractions: ARIMA(1,0,0)



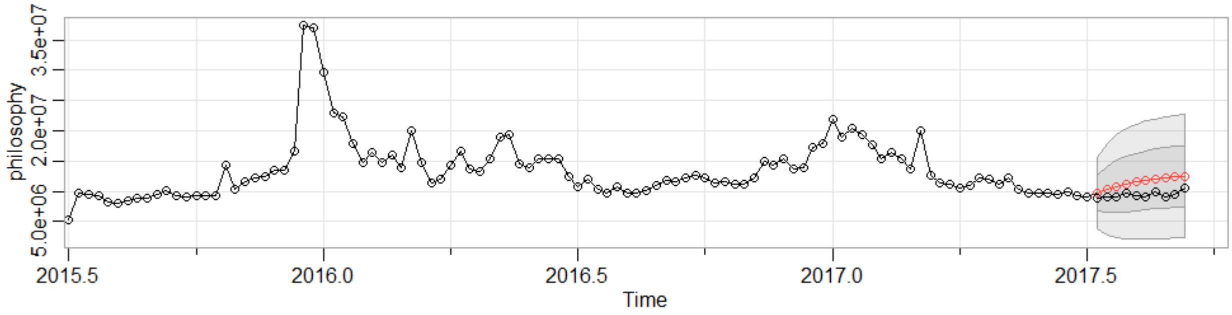
- Natural sciences and nature: ARIMA(2,1,1)



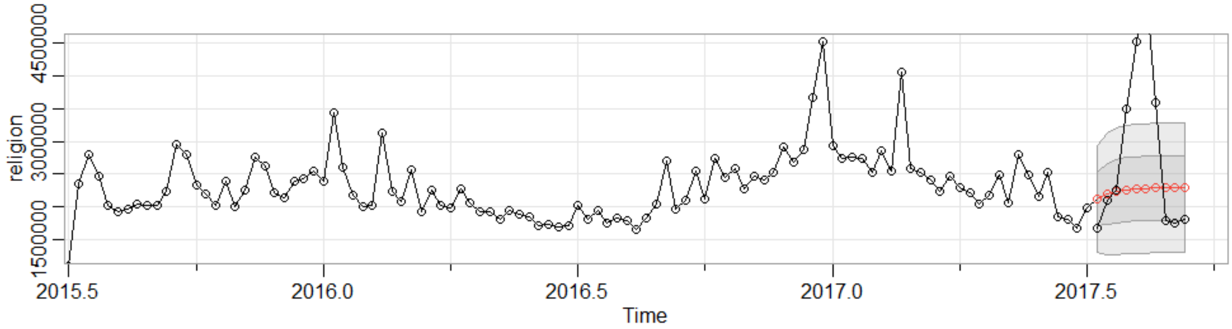
- People and self: ARIMA(1,1,1)



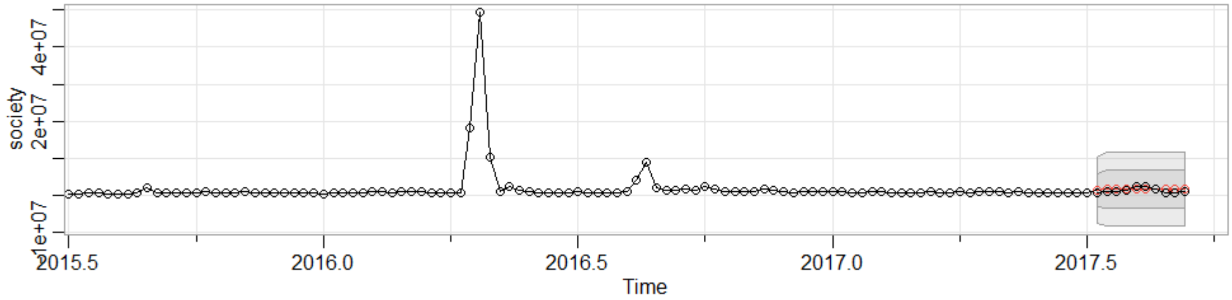
- Philosophy and thinking: ARIMA(1,0,0)



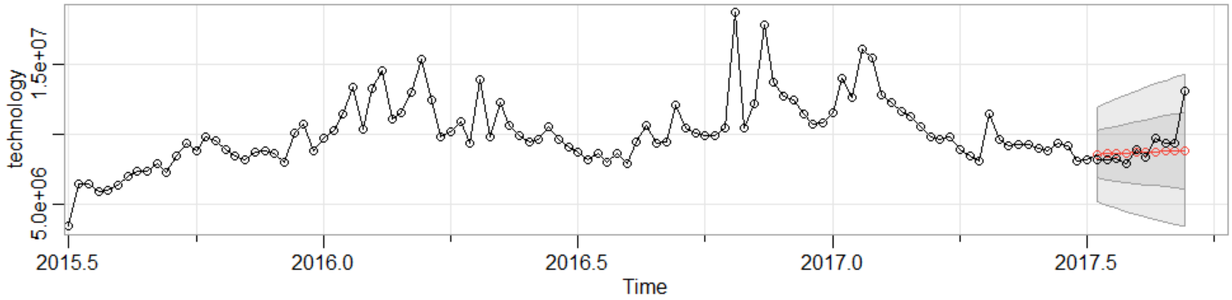
- Religion and spirituality: ARIMA(1,0,0)



- Social sciences and society: ARIMA(0,0,1)

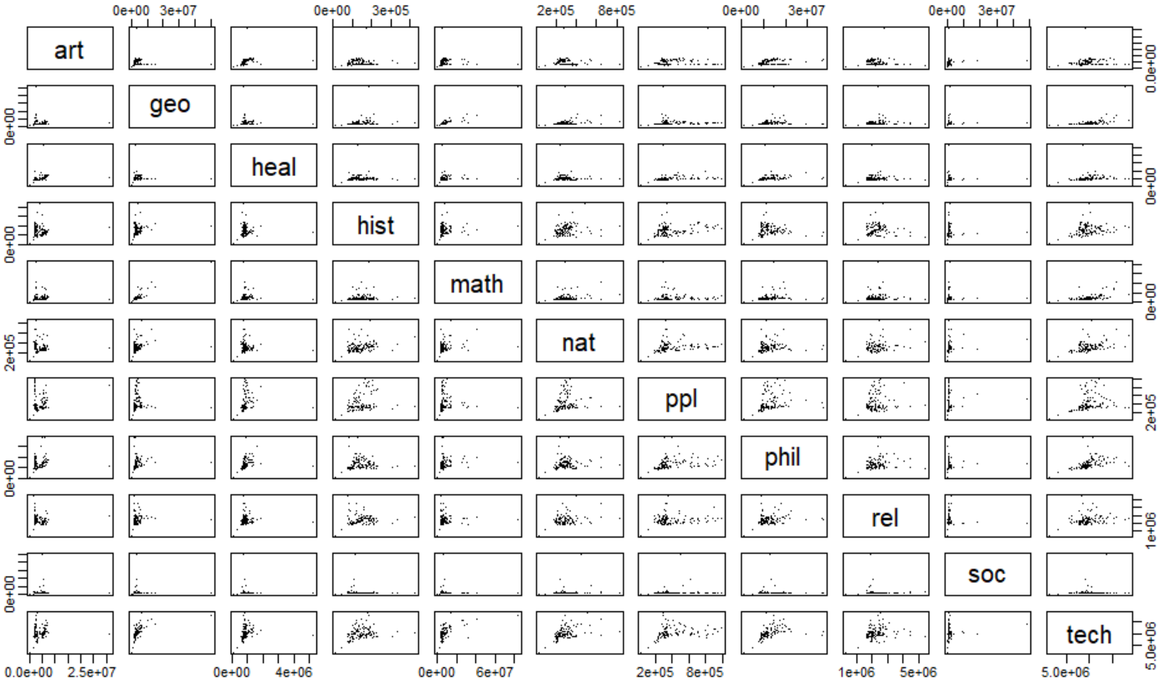


- Technology and applied sciences: ARIMA(0,1,1)



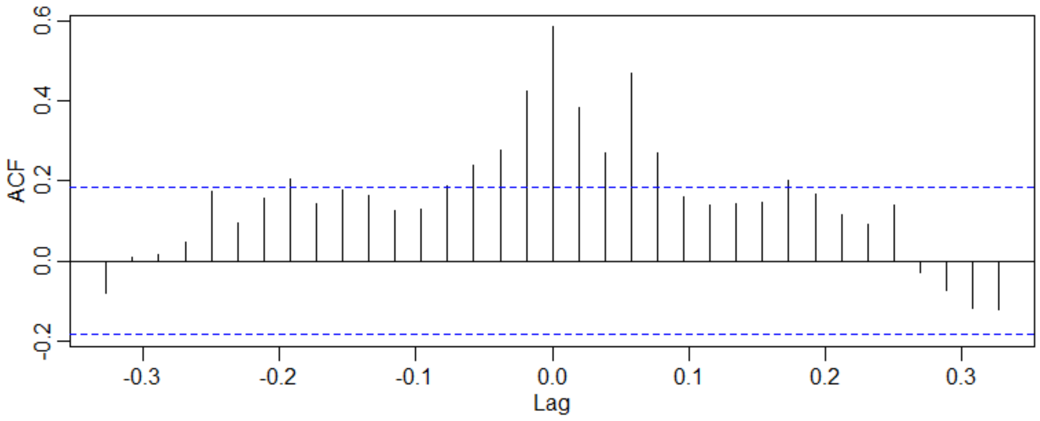
Multivariate Modeling

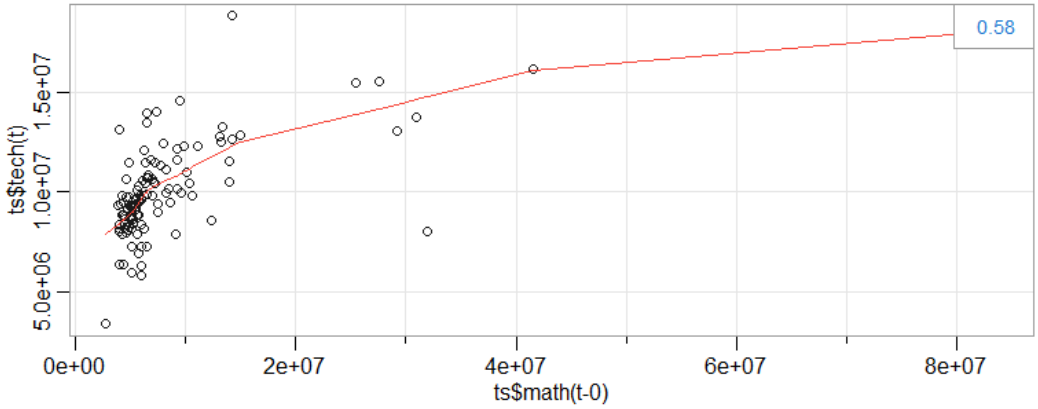
Wikipedia Topic Page View Scatterplot Matrix

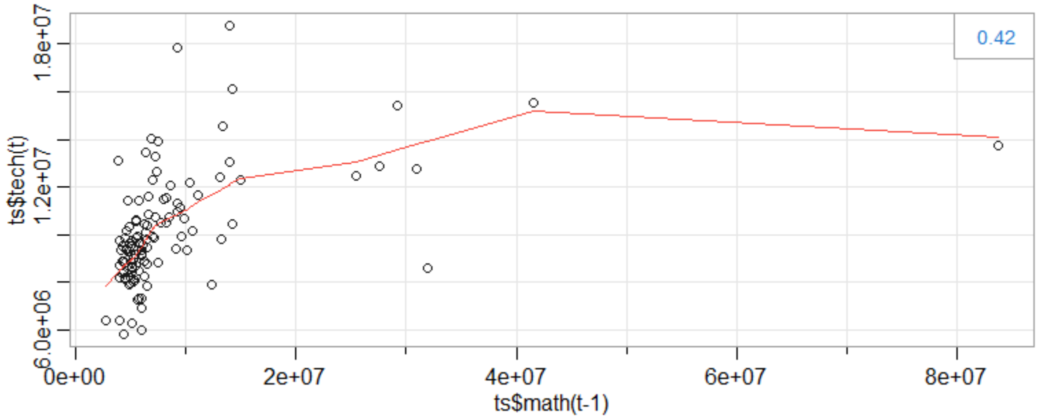


Reviewing the matrix of all topic combinations, we see an interesting correlation between Mathematics and Technology, and Art and Philosophy. Thus we will explore these two topic pairings in greater detail, to see if there are any correlations or models that can be developed from them.

Mathematics ~ Technology





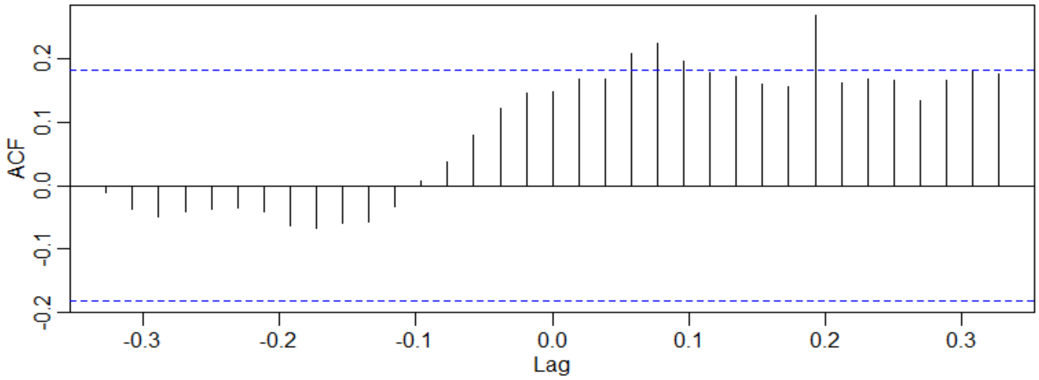


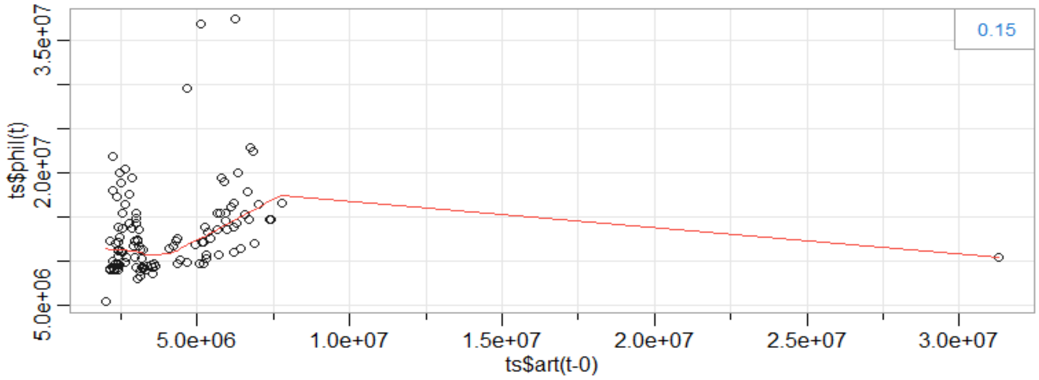
```
> #regression
> math_lm = lm(math ~ tech, data = ts)
> summary(math_lm)

Call:
lm(formula = math ~ tech, data = ts)

Residuals:
      Min        1Q    Median        3Q       Max
-14349179  -3027008  -1019722    800952  57266527

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.436e+07  3.097e+06  -4.635 9.64e-06
tech         2.286e+00  2.994e-01   7.637 7.77e-12

(Intercept) ***
tech        ***
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

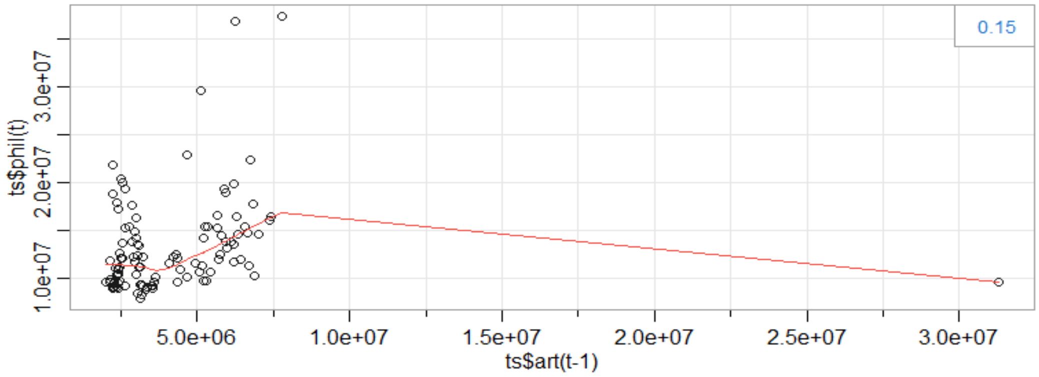
```
Residual standard error: 7645000 on 113 degrees of freedom
Multiple R-squared:  0.3404,  Adjusted R-squared:  0.3346
F-statistic: 58.32 on 1 and 113 DF,  p-value: 7.771e-12
```

Note that technology is a significant variable for determining mathematics page views, but that the R-Squared value is very low, signifying that this is again a poor fit.

Art ~ Philosophy







```
> #regression
> art_lm = lm(art ~ phil, data = ts)
> summary(art_lm)

Call:
lm(formula = art ~ phil, data = ts)

Residuals:
     Min       1Q   Median       3Q      Max
-2730331 -1487629  -698392  1240627 27355362

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 3.009e+06  7.957e+05   3.781 0.000251
phil        8.936e-02  5.630e-02   1.587 0.115210

(Intercept) ***
phil
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3003000 on 113 degrees of freedom
Multiple R-squared:  0.02181,	Adjusted R-squared:  0.01316
F-statistic:  2.52 on 1 and 113 DF,  p-value: 0.1152
```

Contrary to my original hypothesis, there does not appear to be any significant correlations between art and philosophy article page views. This can be shown by the low correlation coefficient, as well as the philosophy variable being determined to be insignificant in the model that was attempted.

Future Works

There are several different models and techniques that could be applied to this dataset to provide better results, with more accurate forecasts. The web traffic appears to follow a random walk distribution, so remodeling all forecasts under the assumption that they are a random walk may provide better results. It would also be interesting to see if including more variables in multivariate analysis could help to make better predictions, such as using both mobile and spider views to predict desktop views.

Unfortunately in every model applied, the Q-Q plot shows large tails that display significance in the residuals following a non-normal distribution. Since forecasting an ARIMA model assumes that residuals are uncorrelated and normally distributed, and since we are not meeting this assumption here with some significance, there will be additional error introduced into our forecasting that is not accounted for. This means that applying different models that do not require the assumption that residuals are normal would provide more accurate evaluations.

Conclusions

Analyzing the time series of total weekly Wikipedia Article traffic, the ARIMA model with a single difference value performed the best in comparison to other models applied. With this model, we were able to achieve reasonable predictions that all fell within the standard error.

Surprisingly, of the different types of web traffic analyzed, the ARIMA model for the desktop had the best predictions, even though the desktop time series contained the most variation in the data. In contrast the spider traffic type had the worst predictions, and obtained predictions even outside of the prediction intervals. Although desktop and mobile views were significantly correlated, the model generated did not have great evaluation metrics when compared to the test set.

The topic time series was the most difficult to model, and due to the large number and variation of topics. The ARIMA models all appeared to provide fairly accurate forecasts, with the exception of the three topics: technology, religion, and nature. The multivariate analysis of mathematics and technology proved fruitful, as there was a strong correlation identified, while the relationship explored between art and philosophy appeared nonexistent.

Overall, there does appear to be a pattern in Wikipedia web traffic with relation to time, although it is not always clear and highly sensitive to large peaks and valleys. The total number of views appears to be the easiest to model, but there were some interesting correlations found between page view types and topics.

Cited Sources

- Brigo, F., Lattanzi, S., Bragazzi, N., Nardone, R., Moccia, M., & Lavorgna, L. (2018, February 5). *Why do people search Wikipedia for information on multiple sclerosis?* Multiple Sclerosis and Related Disorders. Retrieved April 9, 2022, from https://www.sciencedirect.com/science/article/pii/S2211034818300476?casa_token=12JD_y_NnVIAAAAA%3AHs6WgvUIGzhk3kSQToSki32Pz5ksktGCqrNUKUZRe1bnbwqOT9TG_MaB_ZZaNGrd_T7SAgCHlw

- Google Inc. (2017, September 12). *Web traffic time series forecasting*. Kaggle. Retrieved April 9, 2022, from https://www.kaggle.com/competitions/web-traffic-time-series-forecasting/overview

- Scheepers, T. (2018, January 16). *Wikipedia Summary Dataset*. GitHub. Retrieved April 9, 2022, from https://github.com/tscheepers/Wikipedia-Summary-Dataset

- Wikimedia Foundation. (2022, April 5). *Size of Wikipedia*. Wikipedia. Retrieved April 9, 2022, from https://en.wikipedia.org/wiki/Wikipedia:Size_of_Wikipedia