Time Series of Google Domestic Trends: Data Science Searches

Stewart Bobbitt - Time Series - Spring 2016

Introduction

The burgeoning field of data science has become ever more visible over the last decade. To get an idea of just how visible, we can measure public interest by the number of Google searches over time. This analysis employs a variety of R packages to derive a model structure for "data science" searches in the United States and then forecasts the future growth of such searches.

Data

The data is freely available at Google Domestic Trends - Data Science. The data contains three base variables: the starting date of the week, the ending date of the week and the number of searches in thousands for the period. Data was taken from the first week in 2006 through the last week in 2015. Weeks are considered within a year if their start date was within the year.

```
library(gtrendsR)
library(ggplot2)
library(dplyr)
library(TSA)
```

The gtrendsR package allows us to attach to the Google Search API, providing a portal for extract raw and metadata about search patterns. I've extracted only the trend attribute from the gtrends object as it contains the tabular data we'll want to examine.

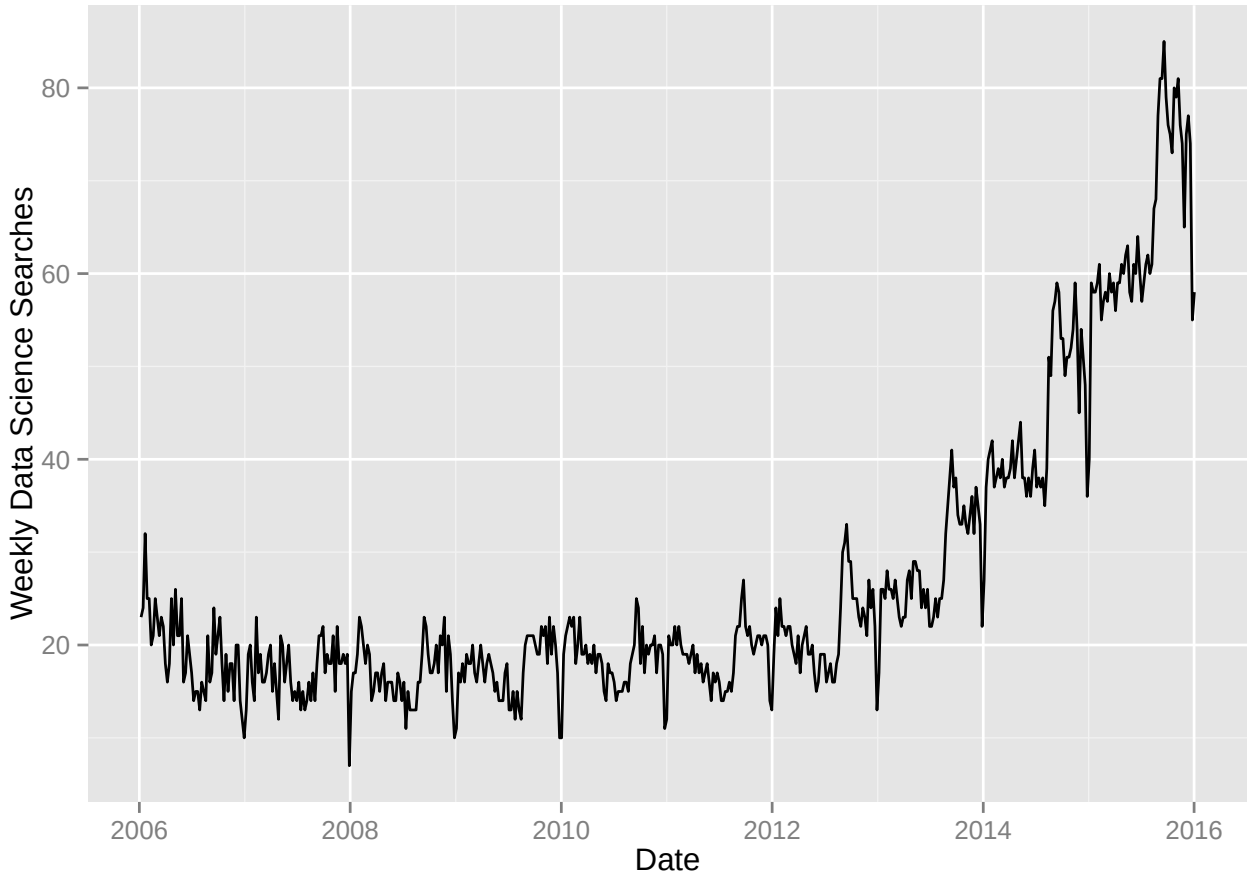
```
data_science_search_data <- gtrends(query = "data science",
                                    start_date = "2006-01-01",
                                    end_date = "2015-12-31")$trend
```

In this step, I've formalized the variable names slightly, presented a glimpse of how the data looks in its tabular form, then illustrated the pattern of searches over the time axis.

```
data <- transmute(.data = data_science_search_data,
                  Start = start,
                  End = end,
                  Searches = data.science.)

head(data)
```

```
##        Start        End Searches
## 1 2006-01-01 2006-01-07       23
## 2 2006-01-08 2006-01-14       24
## 3 2006-01-15 2006-01-21       32
## 4 2006-01-22 2006-01-28       25
## 5 2006-01-29 2006-02-04       25
## 6 2006-02-05 2006-02-11       20
```

```r
ggplot(data = data,
       aes(x = End,
           y = Searches)) +
  geom_line() +
  xlab("Date") +
  ylab("Weekly Data Science Searches")
```



This graphic illustrates the number of searches in thousands over time. From the graph of the data, two distinguishing characteristics are readily visible: there appears to be an upward trend in searches over time, an illustration of the increase in popularity, and searches follow an annual cycle, likely the product of seasonal academic and career searches.

Outline of Method

1. Assess Autocorrelation and Stationarity

- autocorrelation plot
- first differences and corresponding autocorrelation plot
- first differences of seasonal differences with corresponding autocorrelation plot

2. Fit an ARIMA model

- fit ARIMA to first differences of seasonal differences
- QQ plot of residuals

3. Develop a forecast

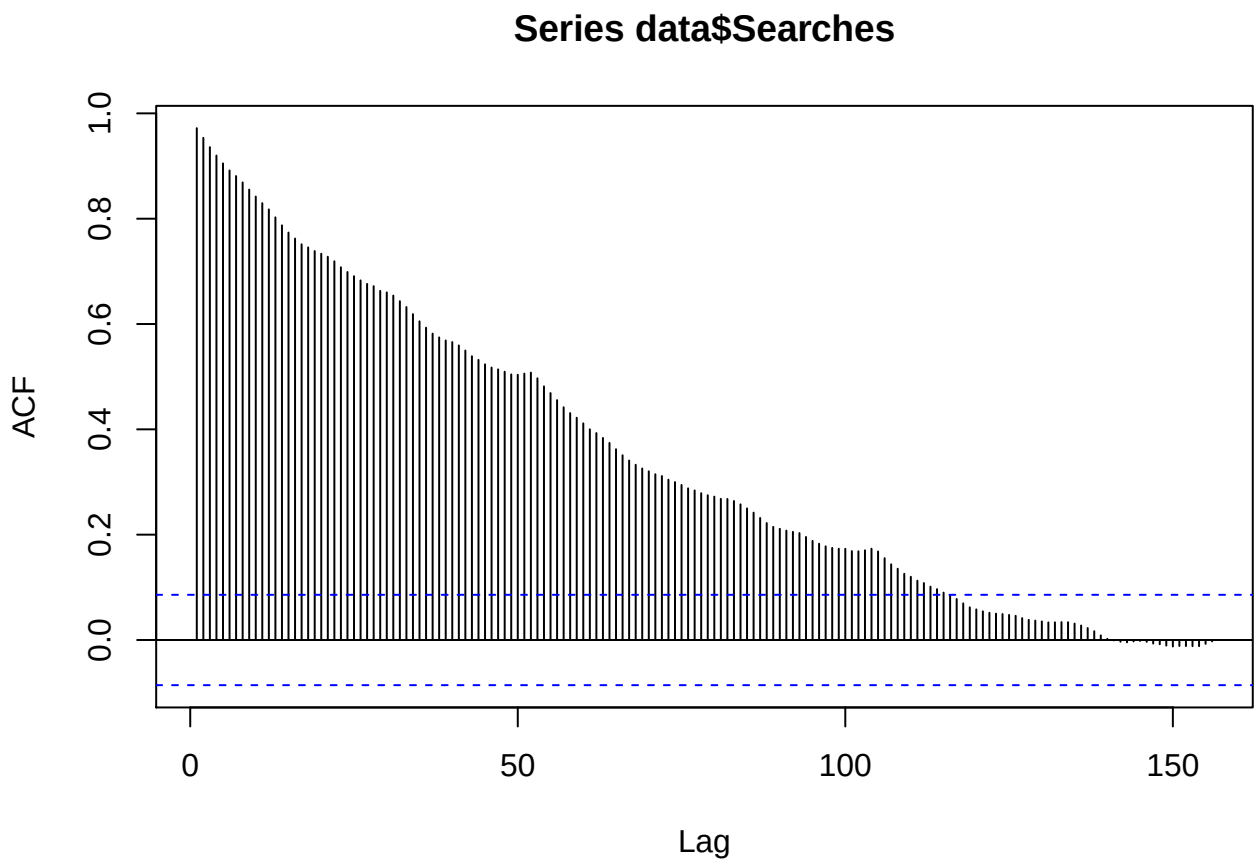
- plot forecast with confidence intervals

Analysis

Assess Autocorrelation and Stationarity

My first step is to determine if autocorrelation even exists within the data:

```
acf(data$Searches, lag.max = 156)
```

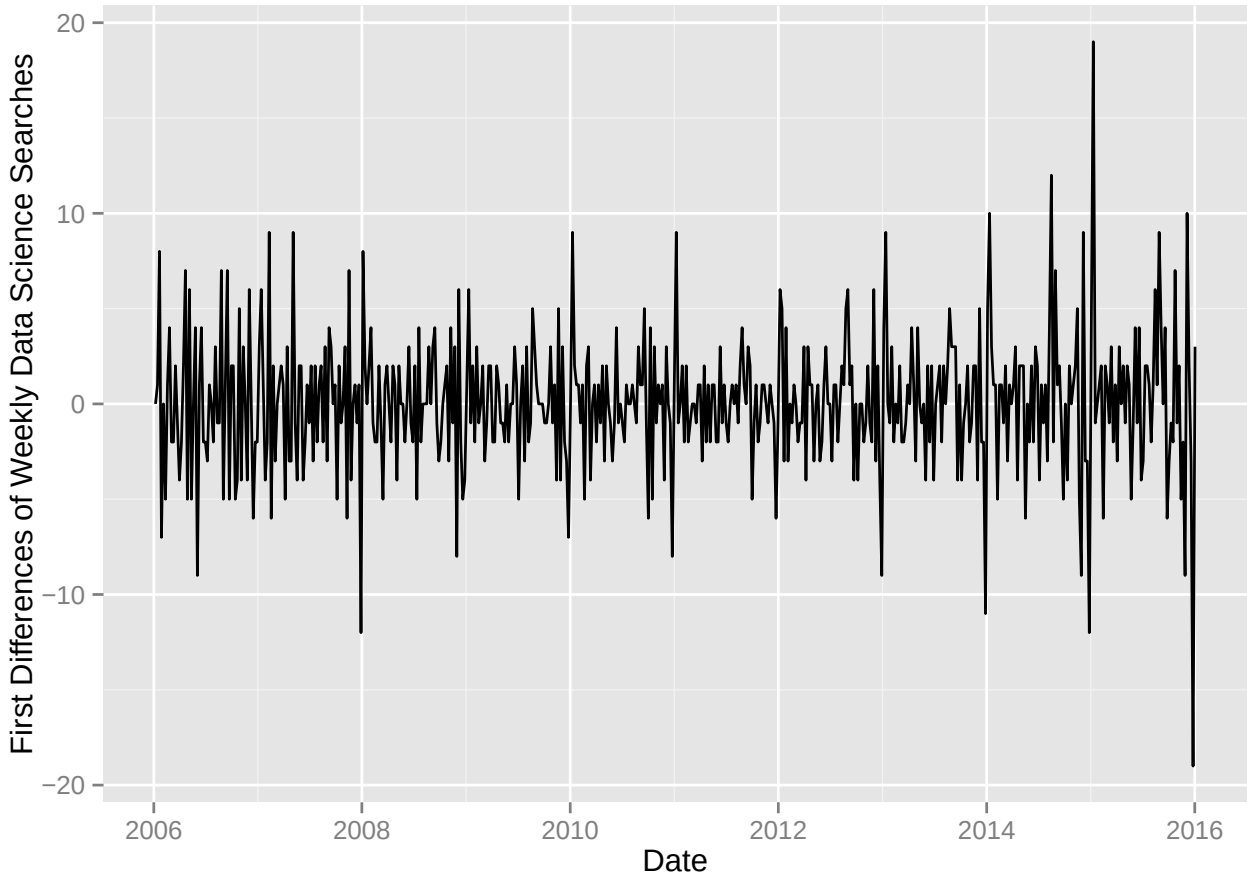


This graphic illustrates the autocorrelation of searches within a week to prior weeks with up to a 3 year lag. Clearly searches over time are highly correlated to the recent past. The 95% confidence intervals are provided by the blue-dashed lines however these are consistent only under the assumption that the series is uncorrelated which is clearly violated in this case. Since the autocorrelation dies off over time, my first inclination is that their may be an autoregressive component to the process.

From the plot of searches over time in the prior section, the process does not appear to be stationary so I've taken the first differences of search counts and provided the following illustration. Note that we lose the first observation by taking these differences, which I have replaced with a zero for the graphical representation.

```
ggplot(data = data,
       aes(x = End,
           y = c(0,diff(Searches)))) +
```

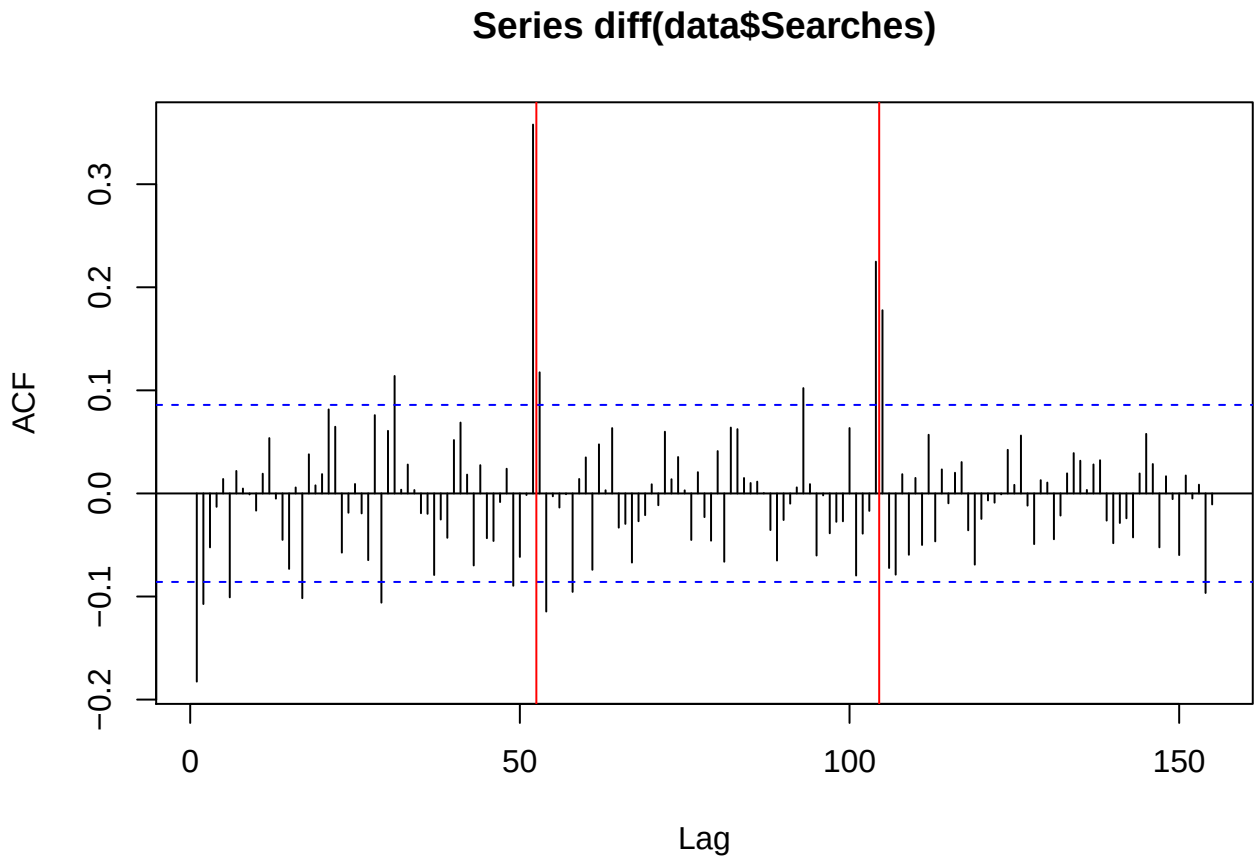
```
  geom_line() +
  xlab("Date") +
  ylab("First Differences of Weekly Data Science Searches")
```



The first differences actually appear to be relatively stationary but a pattern exists that is most easily noticed by inspecting high positive and negative differences at the end and beginning of each year respectively.

To measure the patterns found in the first differences I've generated a plot of the autocorrelation between the first differences and their lags.

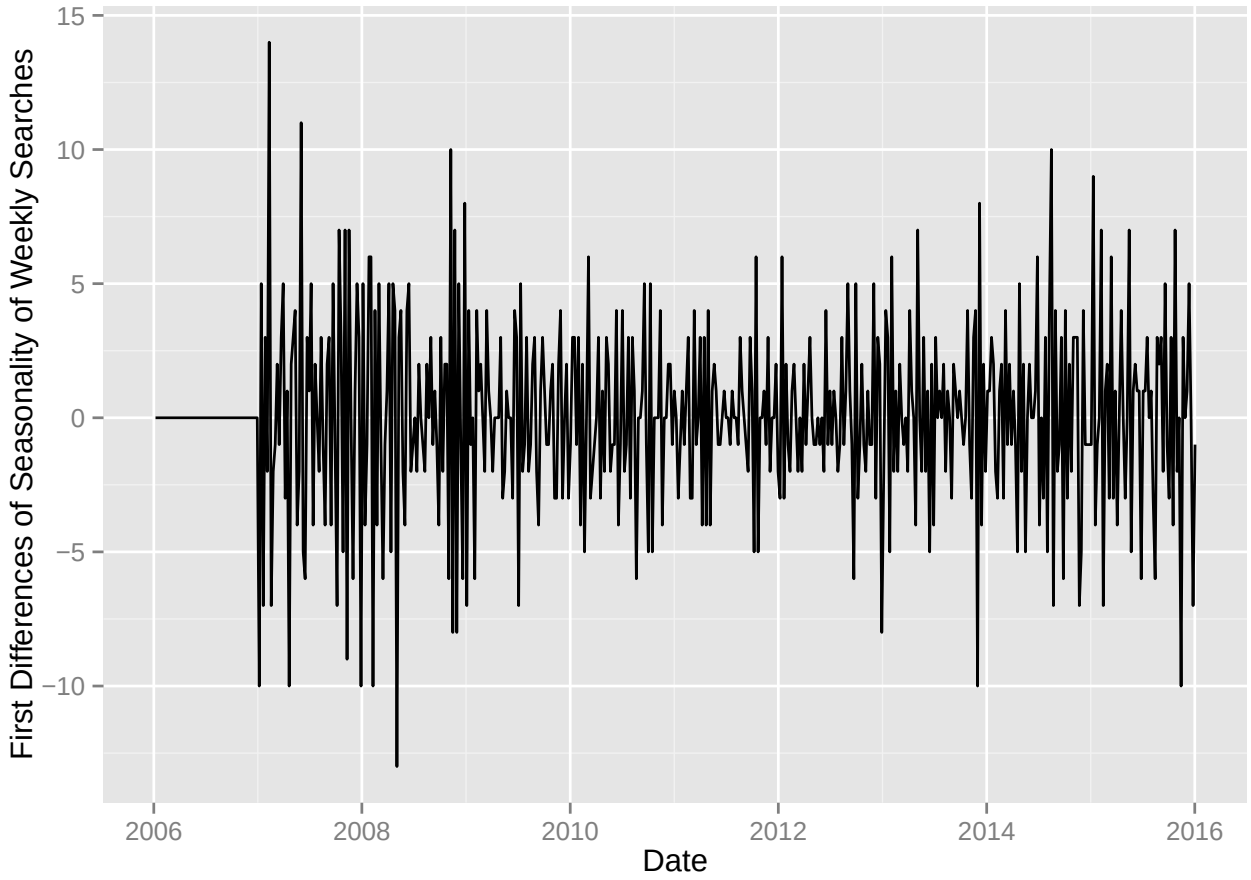
```
acf(diff(data$Searches),
    lag.max = 155)

abline(v = c(52.5, 104.5),
       col = "red")
```



In this graphic, the red line characterizes the transition of lags to prior years. From the autocorrelation plot of first differences, using the confidence intervals to guide our inference, we see a significant negative autocorrelation with the first lag and a significant positive autocorrelation with the 52nd lag. This is confirmatory of two prior observations. First, the current difference appears to be correlated to last week's difference as well as the corresponding difference from last year. Second, negative autocorrelation suggests a moving average process rather than an autoregressive one.

In the presense of the information conveyed by the prior graphic, it is useful to inspect the stationarity of first differences of the seasonal, 52nd differences. The following graphic illustrates these differences over time. Note that we lose 53 observations from the beginning of the series are lost, one for the first differences and another 52 for the 52nd differences.

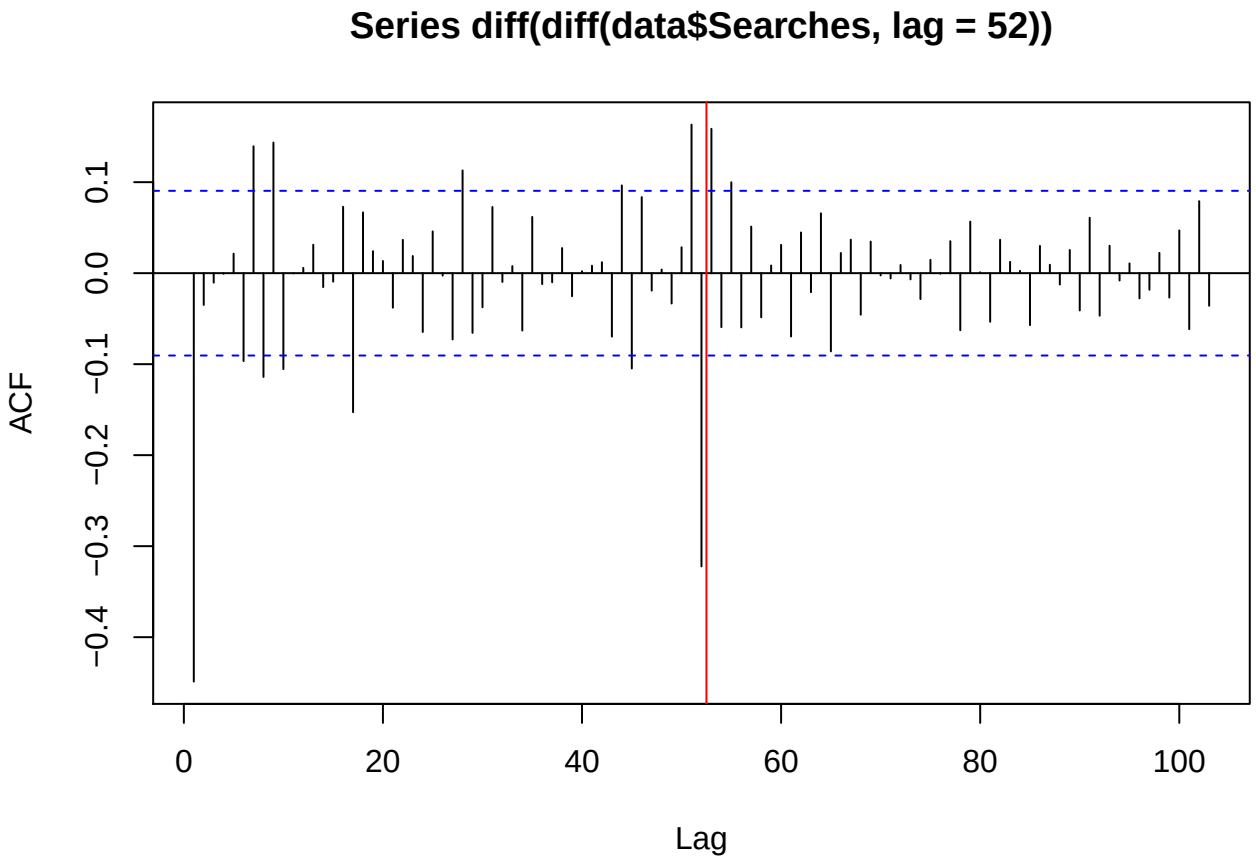
```
ggplot(data = data,
       aes(x = End,
           y = c(0,
                 diff(c(rep(0,
                            52),
                        diff(Searches,
                             lag = 52)))))) +
  geom_line() +
  xlab("Date") +
  ylab("First Differences of Seasonality of Weekly Searches")
```



The first differences of the seasonal, 52nd differences also appear to be stationary with almost no distinguishable pattern.

To help specify exactly what effect the seasonal differences have, I have generated a graphic of the autocorrelation of the corresponding differences from the prior illustration.

```
acf(diff(diff(data$Searches,
              lag = 52)),
    lag.max = 103)

abline(v = 52.5,
       col = "red")
```



In this graphic, the red vertical characterizes the transition between lags in different years. From this graphic, we see that the seasonal effect of first differences is has a significant negative autocorrelation which suggests another moving average effect. Armed with this knowledge, we can move forward and generate a model which incorporates the moving average and seasonal effects.

Fit an Appropriate ARIMA Model

Given the results of the prior section, it seems reasonable to consider the process to be well fit by an ARIMA model taking into account the effect of the first lag and the lag from the prior year. The model is specified as follows.

```
arima.mod <- arima(data$Searches,
                   order = c(0,1,1),
                   seasonal = list(order = c(0,1,1),
                                   period = 52))
print(arima.mod)
```

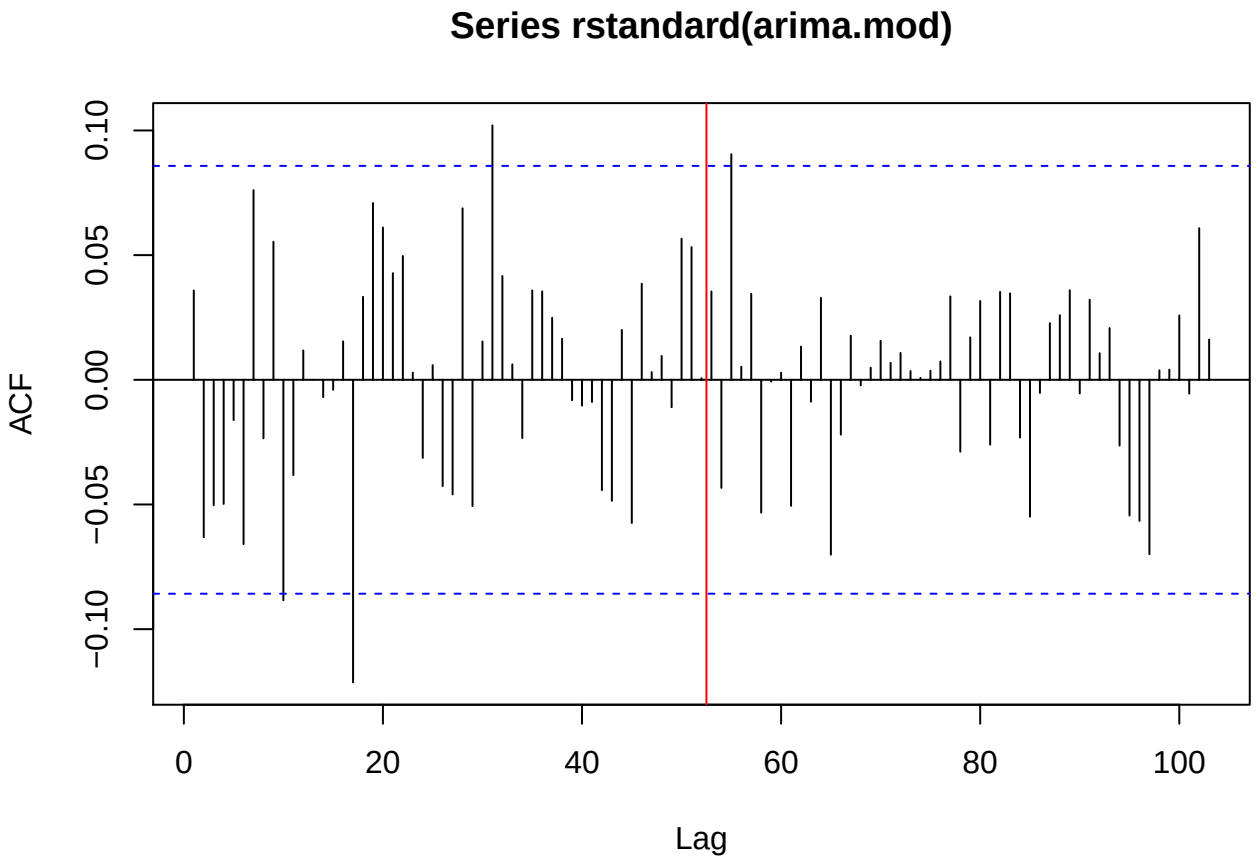
```
##
## Call:
## arima(x = data$Searches, order = c(0, 1, 1), seasonal = list(order = c(0, 1,
##     1), period = 52))
##
## Coefficients:
##           ma1     sma1
##       -0.7634  -0.3771
## s.e.   0.0322   0.0504
```

```
##
## sigma^2 estimated as 7.068:  log likelihood = -1128.48,  aic = 2260.95
```

On the basis of the printed model results the model provides a fit to the data that is significantly different from no effect. The coefficient for the first lag effect, ma1 = -0.7634, is roughly 24 standard errors away from zero, while the coefficient for the seasonal lag effect, sma1 = -0.3771, is more than 7 standard errors from zero, both of which are highly significant under normal theory.

Inspection of the autocorrelations of the standardized residuals helps to confirm whether significant patterns still exist. I've provided the corresponding graphic as follows.

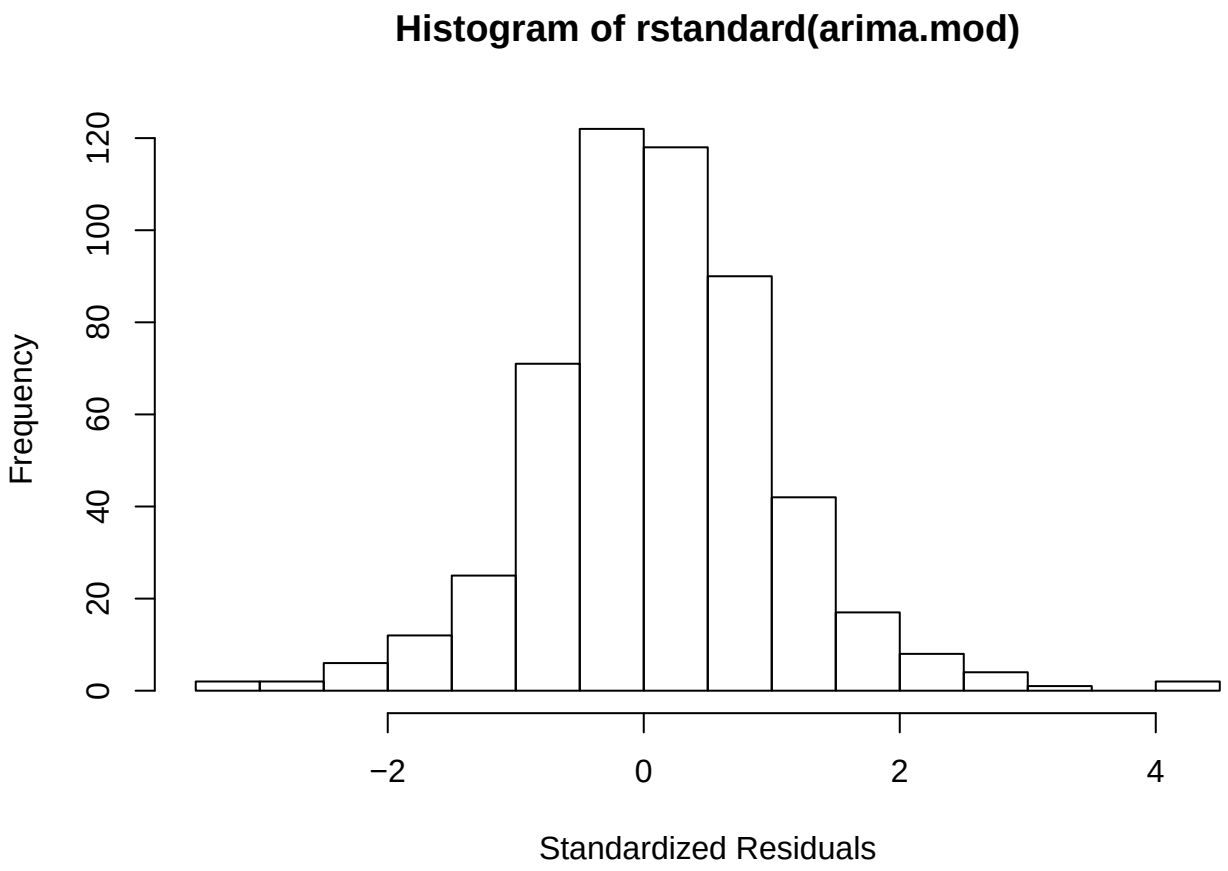
```
acf(rstandard(arima.mod),
    lag.max=103)
abline(v = 52.5,
       col = "red")
```



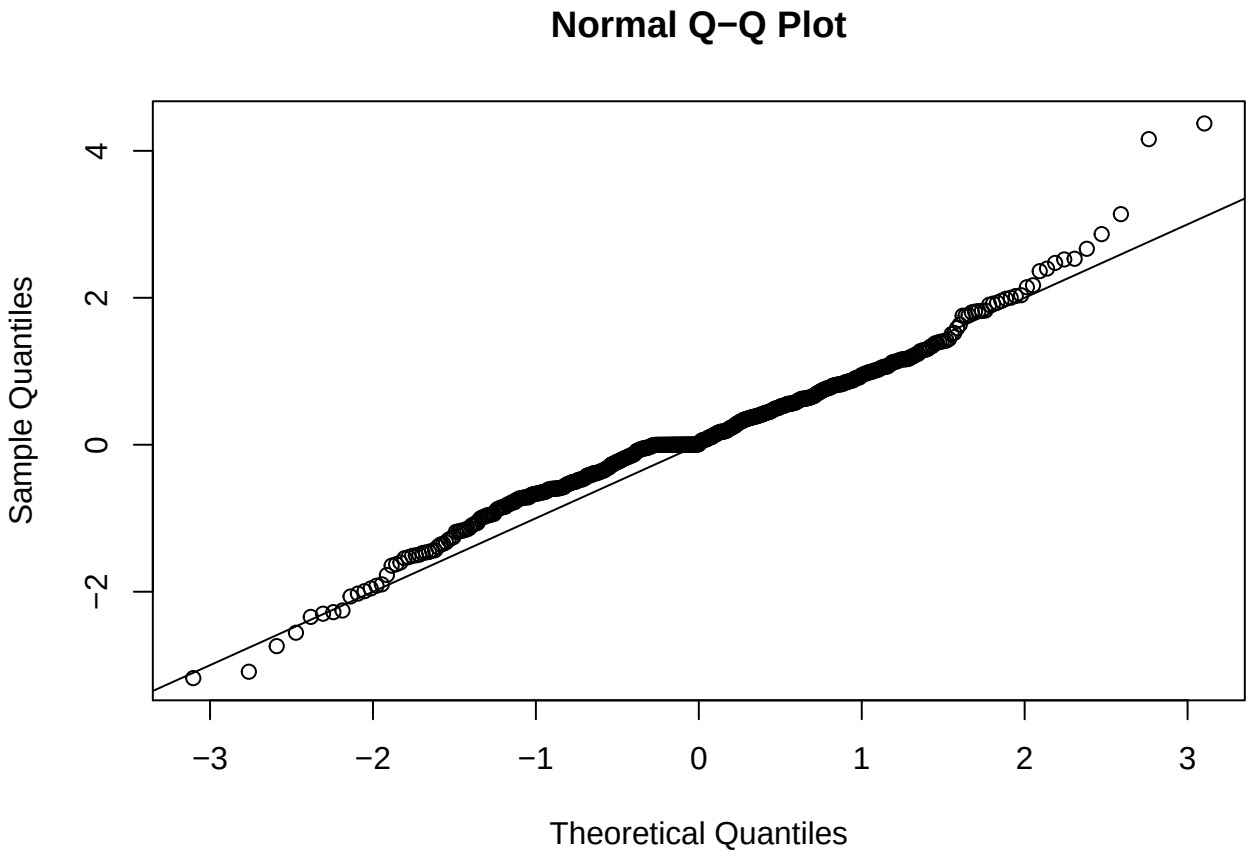
From this plot, only four residuals exhibit marginally significant autocorrelation which can occur merely by chance. This leads me to believe that though some pattern may still exist it does not appear to be statistically significant, given the data.

It is also important to assess the normality of residuals. The following illustrations are provided as useful tools from which we can infer whether the assumption of normality was reasonable.

```
hist(rstandard(arima.mod), xlab = "Standardized Residuals")
```




```
qqnorm(rstandard(arima.mod))
abline(a = 0,
       b = 1)
```

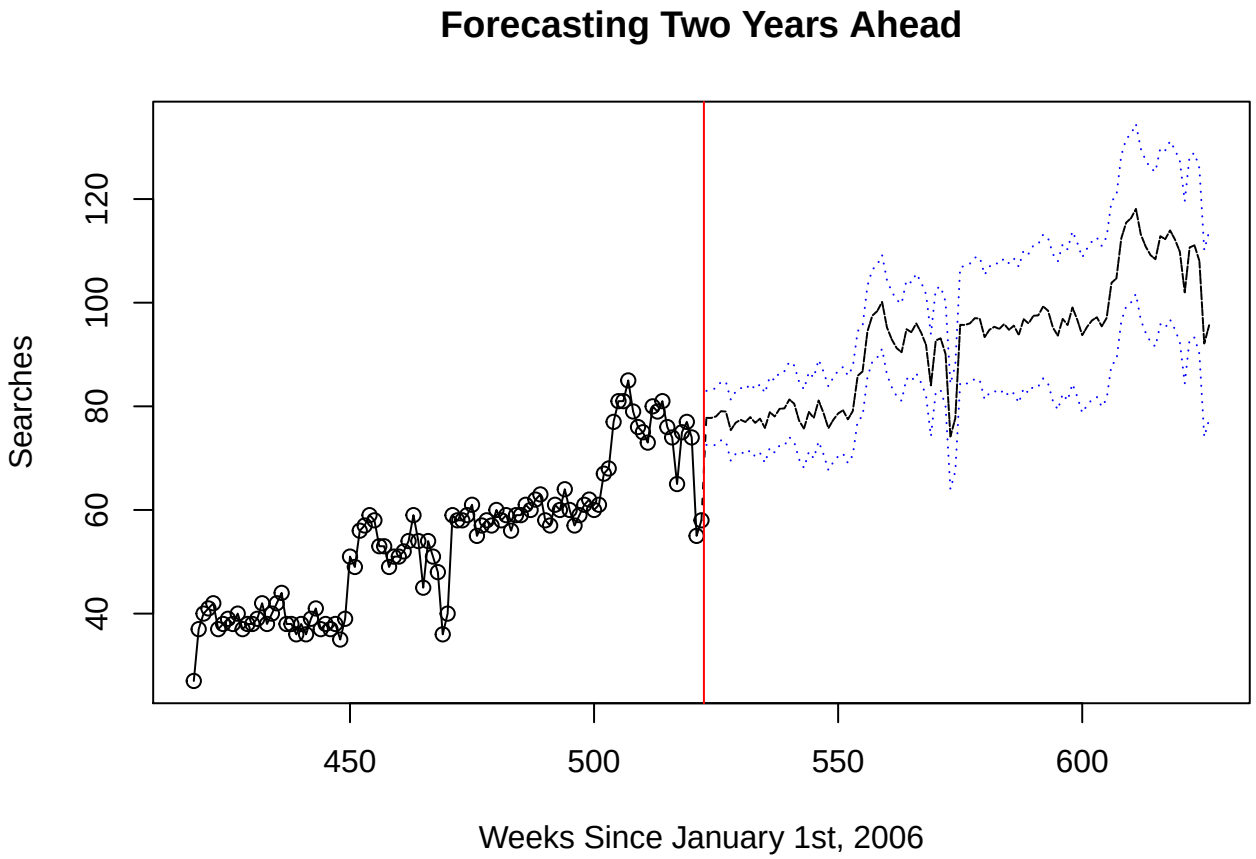


The first graphic is a histogram of the standardized residuals which illustrates that they are symetric and that the tails are not too fat. From the second, quantile-quantile, plot we see that there is evidence of non-normality, given the slight bias above the normal line left of zero and below right of zero with some curvature in the tails. However, given that the residuals are symetric, regression techniques are very robust with respect to the normality assumption. The quantiles are still very close to consistent, the autocorrelations between residuals were not significant and our current specification is relatively parsimonious. From these results, It seems reasonable to conclude that the process is sufficiently specified by the ARIMA(0,1,1)(0,1,1)[52] model.

Forecast the Next Year's Query Results

As a final step, I want to generate a projection of how searches should progress in the near future. The following graphic illustrates this projection by forecasting the specified model over the next two years of cycle.

```
plot(arima.mod,
     n1 = 522-104,
     n.ahead = 104,
     main = "Forecasting Two Years Ahead",
     type = "l",
     col = "blue",
     ylab = "Searches",
     xlab = "Weeks Since January 1st, 2006")
abline(v = 522.5, col = "red")
```



The points left of week 522, the red vertical, represent the actual peformance in 2014 and 2015 while the dashed lines left of week 522 is the forecasted mean with its corresponding 95% confidence interval. These forecasts provide further evidence that the model is a reasonable approximation of the process, given that the cycles apparent in the prior two years and the forecasted two years appear very similar with a consistent trend of growth.

Conclusion

In this project, I have explored the nature of Google searches for "data science" over time. We have seen that it is well estimated by a stationary moving average model which incorporates first and seasonal effects. After specification we were able to forcast into the future two years, generating confident approximations of the growth of the subject over that period. It will be very interesting to see how well this approximation agrees with the actual results but in any case, we should expect a steady, strong growth of the topic in public search patterns.