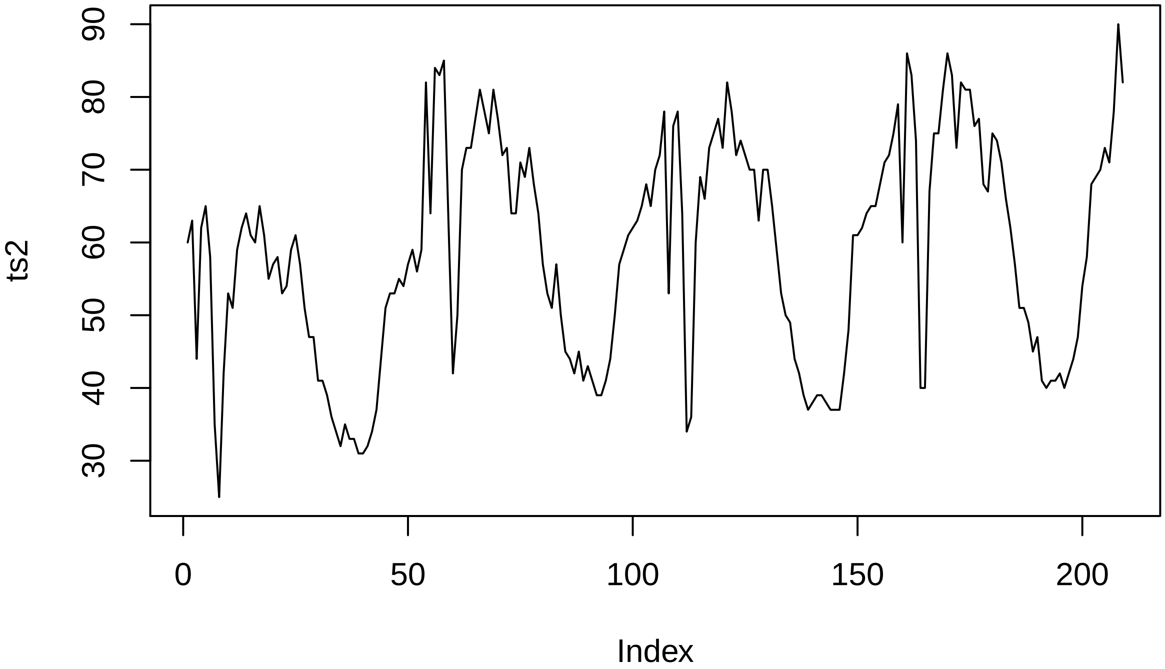
Dataset2 Report

Shicheng Huang

November 12, 2016

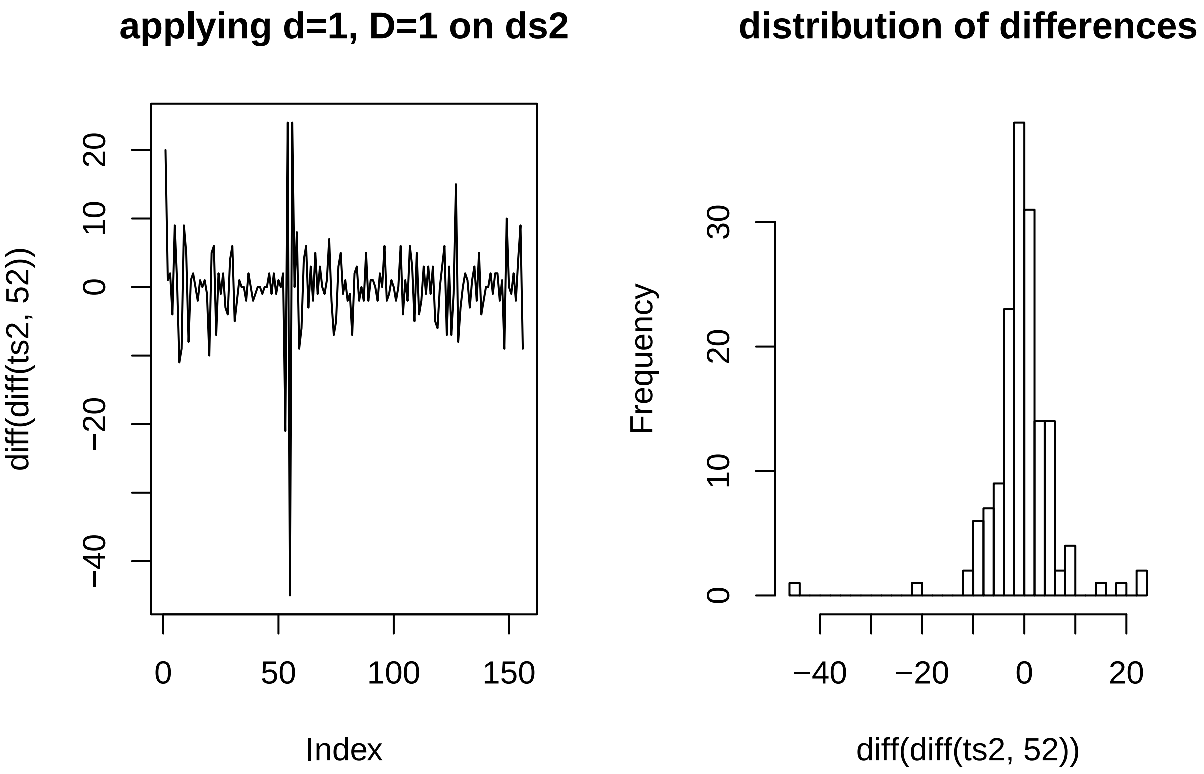
1.Exploratory Data Analysis



1.The data set exhibits really strong seasonality, and a slightly upward trend.
2.I don't think I need to use log transform since the data has very little increasing-variance-trend. 3.The first year seems to lower than the rest, maybe something happened that makes the key word gains lots of popularity.

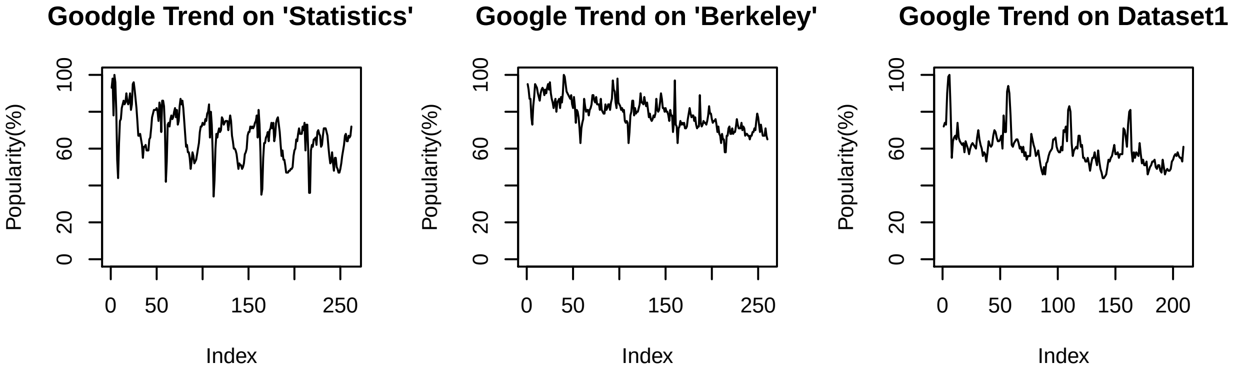
ARIMA model

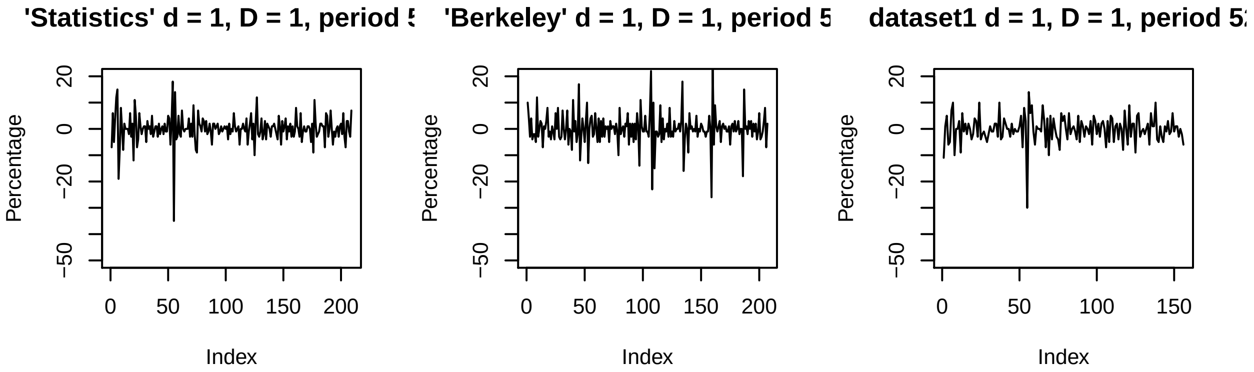
Differencing



After seasonal differencing and ordinary differencing, the histgram of the differences look apprximately normal so kind of like white noise except the sharp decrease around index 52. Recalling that dataset one also shows similar abnormality, **I suspect there is some weird event that happened at 2013 around November that people search less in general.**

To investigate this issue, I deliberate downloaded two extra google trend datasets, ‘Statistics’ and “Berkeley” and check if they have similar behaviors(maybe it is due to events happened during the specific week).





It seems only ‘Statistics’ has similar behavior(the sharp decrease in differencing). I hypothesize that this decrease has something to do with the strong volatile seasonal behavior of the dataset(the popularity changes by a hugh persontage with seasonal changes), which is a characteristic of all three datasets(‘Statistics’, dataset1 and dataset2).

```
## [1]   3  10 -13   0   4   5  -3
```

```
## [1]   4  -6   3  18 -35  14  -4  -3
```

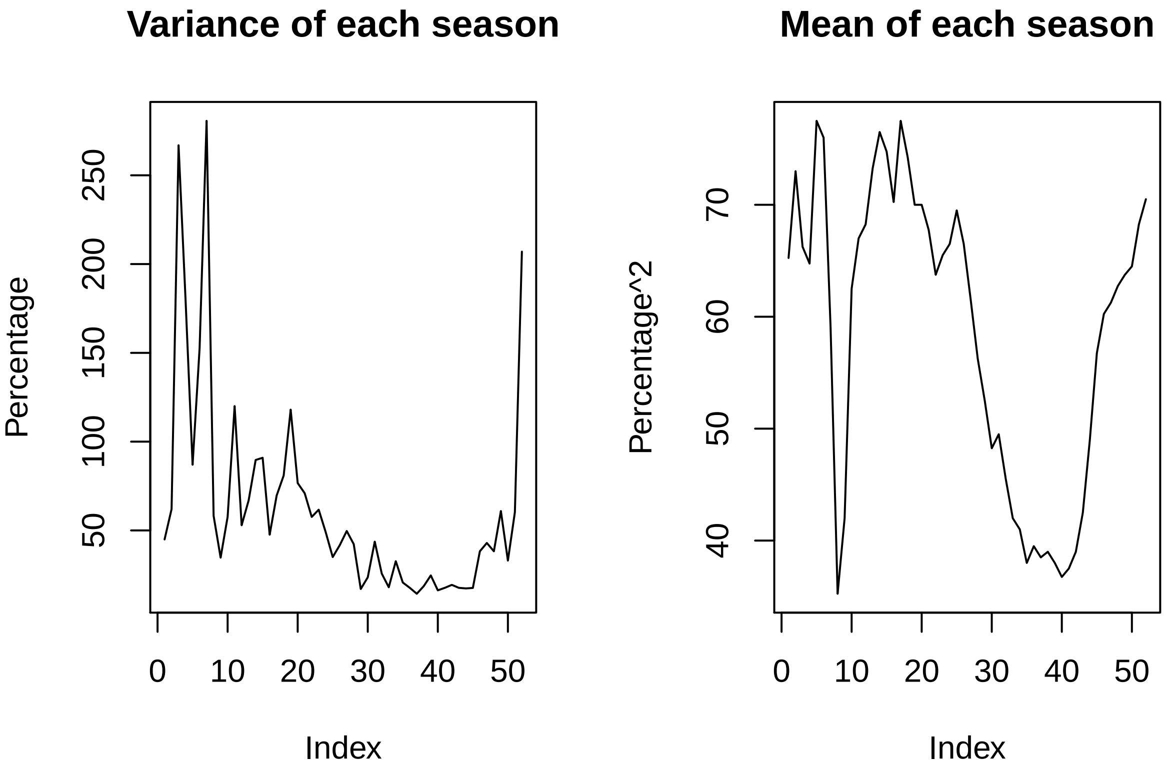
```
## [1]   5  -7   8   3 -30  14   6   9
```

```
## [1]   0   2 -21  24 -45  24   0   8
```

Above are ‘berkeley’, ‘statistics’, dataset1, and dataset2’s 51 to 58 th data of diff(diff(dataset, 52)). We can see the sharp decrease all come from the same index. I just want to argue that this sharp decrease has little to do with the keyword but the some behavior of the data/weekly data. Below is the correlation of dataset2 and dataset1, ‘Statistics.’ At least dataset2 and dataset1 is not really correlated but have the same sharp decrease index.

```
## [1] 0.2856464
```

```
## [1] 0.5962833
```



```
##   year1 year2 year3 year4      var  mean
## 3    44    64    78    79 266.9167 66.25
## 4    62    84    53    60 179.5833 64.75
## 6    58    85    78    83 152.6667 76.00
## 7    35    63    64    74 280.6667 59.00
```

I make a new dataframe whose first four columns are each year's data, and the fifth, sixth columns are the variance and mean of each year respectively. Observing the graphs of variance and mean, I find that period 2,3,6 and 7 have abnormally high variance. And 3, 6 and 7's abnormality comes from period 1 being really low. I decide not to think about the weird behavior of year1; I just want to capture the regular behavior in year2, 3, 4 in my model and predict based on that. These analysis inspires me to make a model only based on year2, 3 and 4.

Since I already invested a good amount of time to this relatively to the amount of time I spent on this proj, I decided to stop my investigation of the sharp decrease and focus a bit more on modeling. I wonder if other students have extracted better information from this and use that to improve their models.

2.Modeling

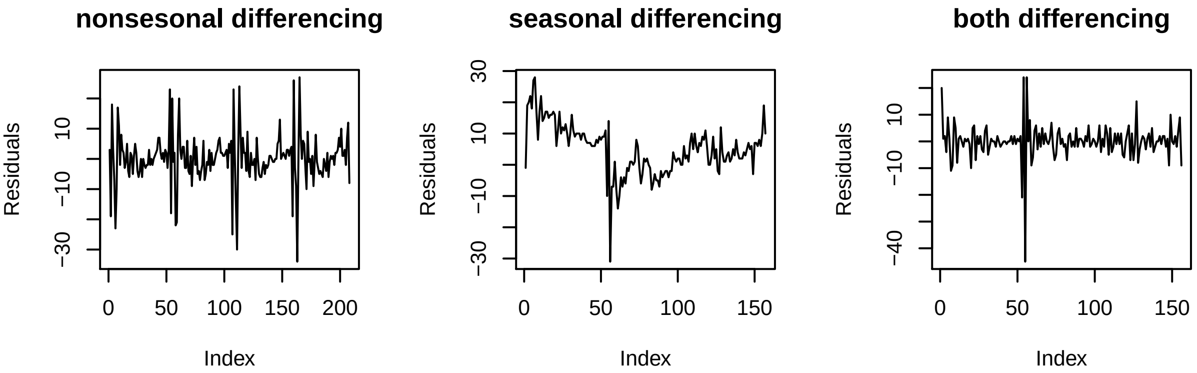
Instinct

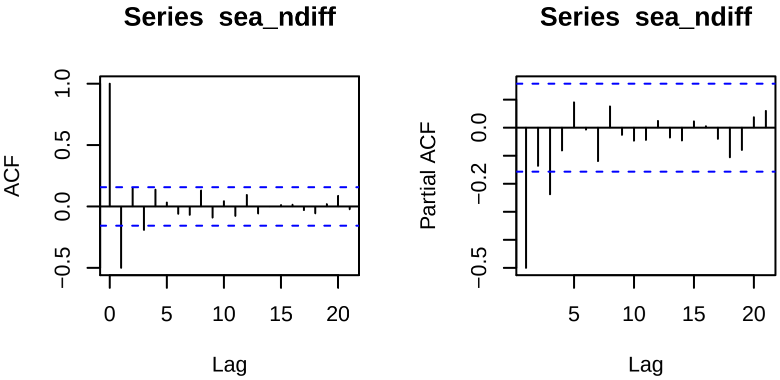
I normally would like to use last year's data to predict the next year's but not for this data since it has an upward trend. So I decide to use the general seasonal $arima(0,0,1)(0,1,1)_{52}$ as my base model.
Reasons:
1.From graphs below, only differencing both could give me a close to white noise residuals plot.

2.After deciding to two seasonal and non-seasonal difference, I would like to add an MA term in both seasonal and nonseasonal model because I could think of it as the difference is the the weighted past error. Using AR here is not as intuitive to me since I don't think the lags differences could correlate with each other.

3.I read from different sources that 011,011 is a commonly used seasonal model for datasets with strong seasonal patterns(and negative acf/pacf)

4.We will see how well my intuition/readings do via mse testing use previous three years to predict the fourth year.





Using my general model(predict the 4th year using the previous 3), I get the mse and aic as follows

```
##
## Call:
## arima(x = ts2[yr123_index], order = c(0, 1, 1), seasonal = list(order = c(0,
##     1, 1), period = 52))
##
## Coefficients:
##           ma1     sma1
##       -0.7044  -0.2491
## s.e.   0.0718   0.1681
##
## sigma^2 estimated as 33.09:  log likelihood = -331.52,  aic = 669.05
```

```
## [1] 16.67774
```

It seems my base model is pretty decent comparing to the test mse of 14-16 I have for dataset1. Now I may need to check out some other orders to see if I can have beter results.
Note:
1. I have tested to predict using an additive linear model + arima model for residuals, but linear model has too much prediction variance that makes it inferior to arima.
2. I don't want to use an additive sinusoid model because I don't understand it well enough and time constraint.

Searching for Better Models

Tuning/Tweaking the general models

Below are the mse using previous 3 years predicting the 4th, using different arima

```
## arima(1, 1, 0)(1, 1, 0), switching ma to ar
```

```
## [1]  18.60141 677.83803
```

```
## arima(0, 1, 1)(0, 1, 2), adding an MA term at the seasonal model
```

```
## [1]  16.62495 671.04741
```

```
## arima(0, 1, 2)(0, 1, 1), adding an MA term at the nonseasonal model
```

```
## [1]  16.43692 670.50431
```

AIC-wise all pretty much the same. Ar model does a slightly worse job than MA, adding extra term in the seasonal or nonseasonal model does make the error slightly smaller but not signficant. I stick with my general model for now.

Benchmarking

An important thing we should consider is time complexity and space complexity. Lets see how much time it take for each model(including my base model) to run.(would have shown a complete profiling including memory usage if no page limit.)

```
## Unit: milliseconds
##                                                                     expr
##  mod_mse(ts2[yr123_index], ts2[yr4_index], c(0, 1, 1), c(0, 1,      1))
##  mod_mse(ts2[yr123_index], ts2[yr4_index], c(1, 1, 0), c(1, 1,      0))
##  mod_mse(ts2[yr123_index], ts2[yr4_index], c(0, 1, 1), c(0, 1,      2))
##  mod_mse(ts2[yr123_index], ts2[yr4_index], c(0, 1, 2), c(0, 1,      1))
##  mod_mse(ts2[yr123_index], ts2[yr4_index], c(0, 1, 2), c(0, 1,      2))
##        min        lq      mean    median        uq       max neval
##   1232.632  1232.632  1232.632  1232.632  1232.632  1232.632     1
##   1581.922  1581.922  1581.922  1581.922  1581.922  1581.922     1
##   8091.403  8091.403  8091.403  8091.403  8091.403  8091.403     1
##   1818.996  1818.996  1818.996  1818.996  1818.996  1818.996     1
##  11263.914 11263.914 11263.914 11263.914 11263.914 11263.914     1
```

Adding a second MA term in the seasonal model will make the model run significantly slower! While doing it on the nonseasonal model doesn't really change the speed much. So my general model thrives in both runtime and accuracy. So simple so strong!

Taking out the first year

The first year is quite low, comparing to others, lets see if our errors get better if we take it out.

```
## arima(0, 1, 1)(0, 1, 1), general model with first year off
```

```
## [1]  34.3227 350.9780
```

```
## arima(0, 1, 2)(0, 1, 1), extra nonseasonal ma
```

```
## [1]  22.88719 351.13826
```

```
## arima(0, 1, 1)(0, 1, 2), extra seasonal ma
## [1]  34.38194 352.97800
```

I bring down the AIC but the accuracy is not so unstable. That may have to do with insufficient years. After long considerations, I decide to stick with my general model. Because the weird year(s) may not be always at the first year, if in the mid, taking it(them) out would be intuitively weird.

3.Final Notes

1. after all the work I have done I am still unable to beat my very first model $arima(0,1,1)(0,1,1)_{52}$.
2. technical-wise I have learned a lot: using Rproject, practicing using Vim, benchmarking and profiling, some rmarkdowns tricks, and some technical things about arima.
3. I have spent much more than on this project than I expected, and only approximately 20% of the effort is producing 80% of the result. I should really have set a time limit beforehand so that I can prioritize on the important tasks.