Memo for Search trends: Diet and Gym

Part I: Data Exploration

There are some patterns in the data for the Google search score on diet and gym. The more obvious peaks and lows in the search scores for both diet and gym occur roughly around the beginning and ending of each year. We can also see a periodic trend in both data, which has about a 1 year cycle. Since the period of cycle for both data are roughly the same, the peaks and valleys for the diet and gym searches roughly match to each other.

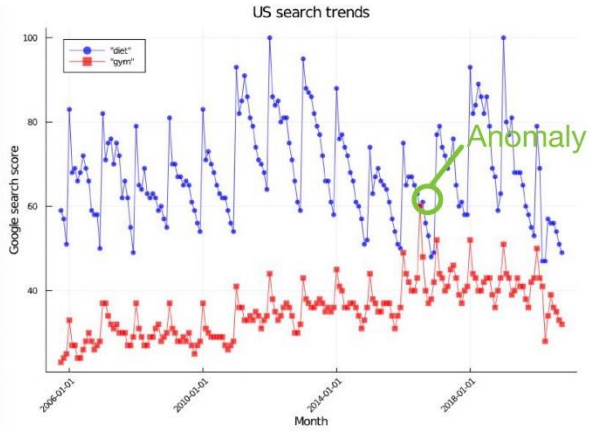

Figure 1. Example Anomaly in Google Search Trend

The patterns shown above do make sense empirically. People usually make their new year's challenge about maintaining their health and losing weight.

Therefore, the google search score on both diet and gym start to dramatically rise up in the beginning of each year. However, as time goes by, people start to lose interest or motivation in continuing their new year's challenge. This is the reason why we can observe the valleys at the end of each year.

Even if the data show a consistent trend in the Google search score on diet and gym, there are still some anomalous peaks or valleys. For example, there is a moment when the Google search score for gym reaches to peak while that of diet is not in July, 2016 as shown in the plot below. This is probably because people start to get more motivation again in exercising during summer.

Part II: Model fitting implementation details

The function 'lsqfit_poly_periodic' is used to fit the data. For "gym" trend data, the optimal d_poly, d_periodic and T choices are 6, 7, 12. For "diet" trend data, the choices are 6, 7,12.

By creating a plot of squared error versus polynomial degree, we can get the optimal d_poly, which minimizes the error. As shown in the figure below, the optimal d_poly for "gym" and "diet" are both 6. In this way, we can get the best choice for the other two parameters. When using the above parameters, the predicted trends have the smallest squared errors.

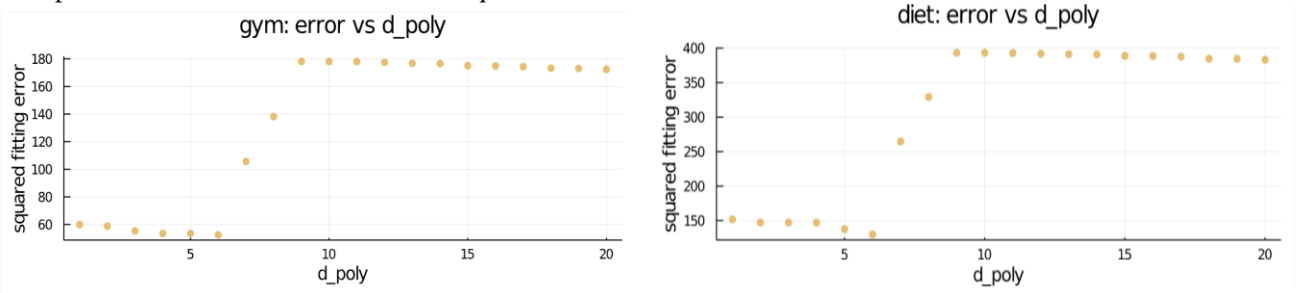

Figure 2. Error for Gym data and diet data by d_poly

The choice of T is made based on the unit of data collection period (month) and periodic pattern in the data. Since the similar periodic pattern happens throughout each year, it is reasonable to set T = 12.

The following two plots show the results of fitted trend versus data being fit for "gym" and "diet" respectively.

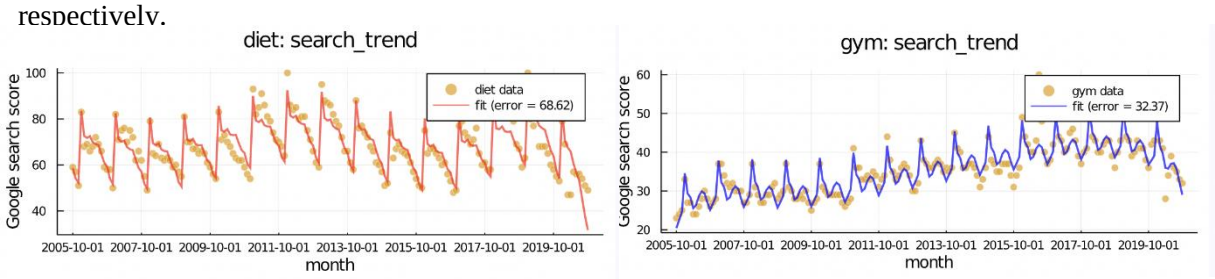

Figure 3. Fitted model for Gym data and Diet data

Part III: Forecasting search trends for a year into future

The computation for the forecasting is performed by using the function poly_periodic which generates the y-value for inputted dates, namely we used the 13 future dates to conduct the prediction. The plot of the forecast in dashed linestyle is in a way that "follows" the fit in a solid line. The forecast is not quite "good" from our perspective. Our definition of "good" is a prediction that truly reflects the cyclic pattern as well as the trend that appeared in previous times. As shown in the plot, the prediction is downward to the bottom which has nothing to do with the previous trend. There are indeed some "good" points in the prediction. For example, there is a peak at the beginning of the predictions for both Gym and Diet data. This part reflects a bit of the pattern that appeared in the raw data of Gym and Diet, which is seemingly believable. We also used the ARIMA model on the data to make predictions. The result is shown below. Compared with the model using polynomial and triangular functions combined, the prediction made by the ARIMA model is much more plausible. The prediction made by our polynomial model should probably be attributed to the amount of parameters. The redundancy of the parameters made the model easy to oscillate with small disturbance and hard to make extrapolation. In other words, it is an over-fitted model.

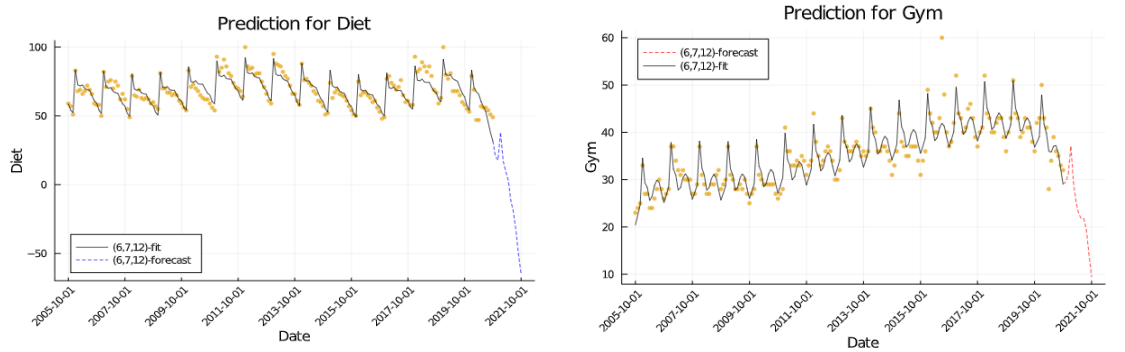

Figure 4. Prediction model for diet and gym data

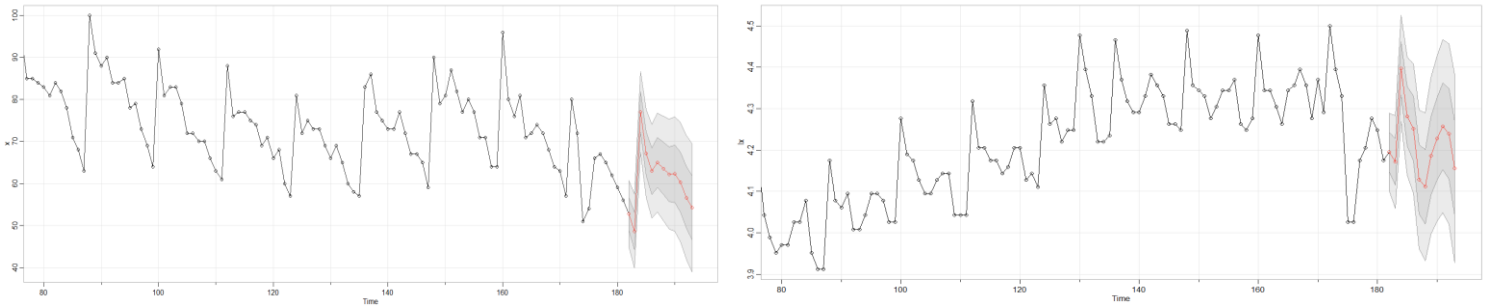

Figure 5. ARIMA(0,1,1)*(0,1,1)$_{12}$ seasonal model for Diet data. ARIMA(0,1,1)*(0,1,1)$_{12}$ seasonal model with a log transformation for Gym data. (Generated using R)