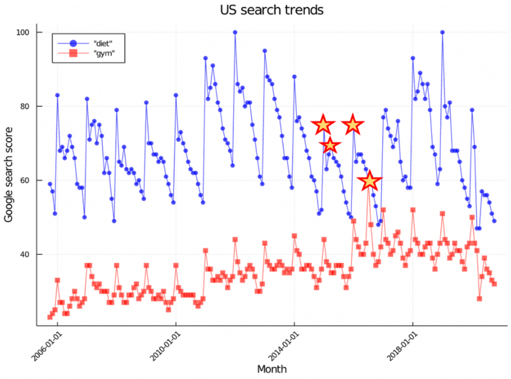
Predicting search trends for “diet” and “gym” - Yutong Zhong, Bryan Tapnio, Yuewei Wen

Data exploration (by Yutong Zhong)



Several global maxima and minima showed up in the search score data for both “diet” and “gym”. In January 2012 and January 2019, search scores for “diet” peaked. January is wintertime for most parts of the US and is the beginning of the year. These peaks can be explained by people’s unwillingness to go to the gym in the cold as well as habits of writing new year resolutions. Search scores for “gym” peaked in July 2016. Summer can be an energetic time and people may care more about body shapes when going to the beach. The lowest search for “diet” is in March 2020 and April 2020 since the Covid-19 pandemic grabs the most attention from individuals. The lowest one for “gym” is in October 2005 since more and more people recognized the benefits of exercises after 2005.

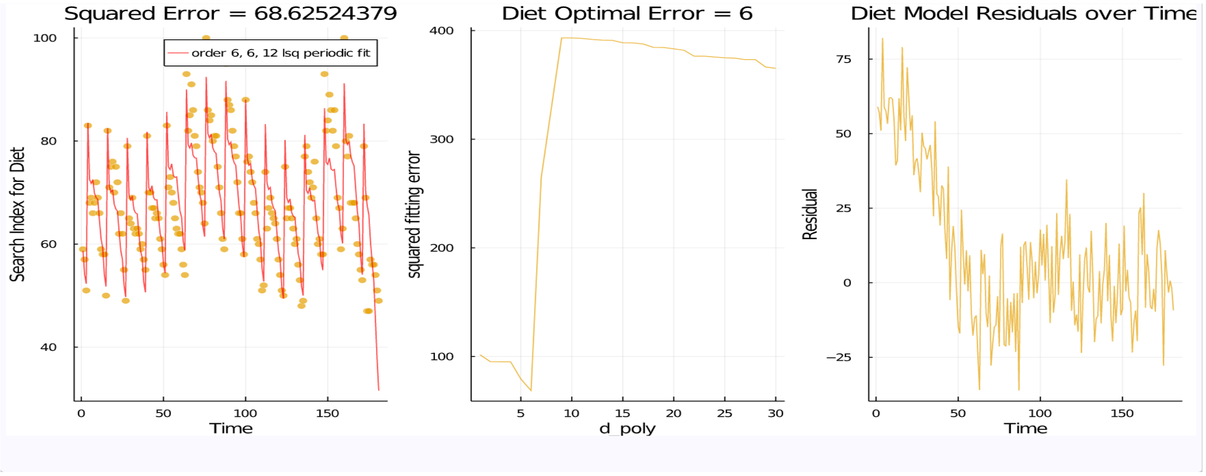
In addition, we found a 2 to 3-month cycle between local peaks and lows for both words. Overall, search scores for “diet” and “gym” always peak annually in January. The general trends of “diet” and “gym” are positively correlated most of the time. Data for “gym” has an abnormally tall peak in July 2016 that cannot be spotted in any other years. Another anomaly is a valley in the search score of “diet” from 2015-2017 when the local maxima is lower than other years.

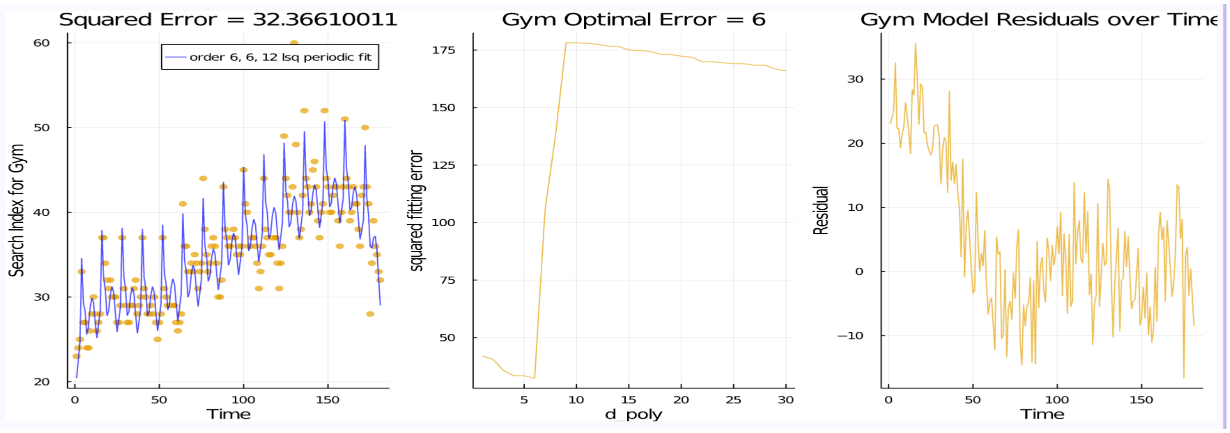
Implement a least-square model fitting (by Bryan Tapnio)

We fit the diet and gym search index trends data using a modified polynomial-periodic least squares function that consists of polynomial terms in time as well as sine and cosine functions w.r.t. time.

$$search\ score\ =\ \sum_{i=i}^{d_{poly}} c_i t^i\ +\ \sum_{i=1}^{d_{perodic}} \alpha_i sin(\frac{2\pi it}{T})\ +\ \beta_i cos(\frac{2\pi it}{T})$$

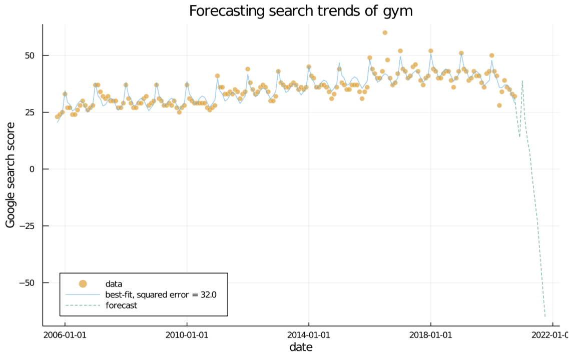
This model has three parameters: the degree of the polynomials (d_poly), the number of periodic terms in the equation (d_periodic), and the fundamental period (T). We found that the values that will minimize





the difference between the best-fit model and the data are d_poly = 6, d_periodic = 6, and T = 12 for both the "diet" and "gym" search score data.

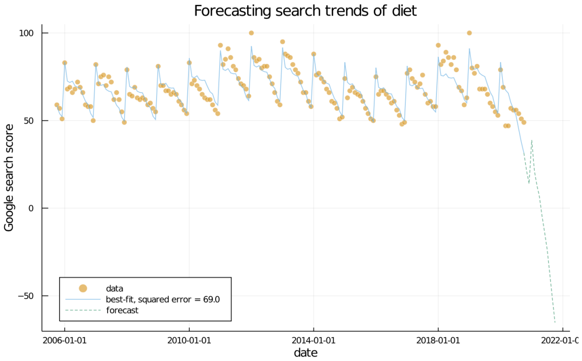
We pick T = 12 because the unit of time in the data is a month, and the seasonal trend seems to repeat itself every year. For a fixed value of d_poly, the parameter value of d_periodic does not have a significant impact, and we arbitrarily set it to be 6. We determined the best-fit value for d_poly by scanning through a range of integer values from 1 to 30 and found that d_poly = 6 minimizes the l2 error. Plots of the results, as well as residuals and best-fit plots, are below. The model generally fits data well. For both words, residuals are higher in the earlier years and fluctuate around 0 in later years.



Search trend forecast (by Yuewei Wen)

To forecast the future search scores of "gym" and "diet", we generated some future dates, converted the data type from "date" to a list of numbers, and plugged the "converted dates" into the best-fit model (i.e. the poly-periodic function) to compute the Google search score for these future dates.

Results are shown in the next two figures, The green dashed line in each plot represents predicted future trends. The predicted trend, for both words, went through a period and then plunged downwards. As a quantity like Google search score is not physically defined in the range of negative numbers, the overall prediction is not reliable.



Limitations of extrapolating with a least-square fit are demonstrated clearly in these results. The best-fit technique we employed only concerns itself about the oscillations in the existing data but has very limited information after the end of the last cycle. The portion where the Google search score went negative is unusable, and the only part that is somewhat usable is perhaps the last oscillation cycle in the existing data.