There are 2 methods in this analysis:

1- Google Prediction API (GPA):
GPA is an online analytics offered by Google. It aims for building predictive models for 2 major tasks: Regression and Classification. However, there's a number of limitation that the tool offers.

First, it requires many observations in order to make good prediction. However, it took lots of time to build the model when more data is added. Second, they need the input data to be in an exact format: response be the first columns, while the remaining be the predictors. Third, there is no functionality of data partition as well as dimension reduction (even simple task like ignoring some predictors in building the model). Last but not least, there's no option for performing analysis with cross-validation.

The model built by GAP does not provide good result at all. It's off to hundreds of value and sometimes return negative ones, which is illogical.

2- R Script:
Before going straight to building models, there are a couple things to do:

a. Initial data processing: remove string predictor, and other irrelevant/duplicate predictors.
b. Sampling data: since the data is big, I randomly partitioned the data into training (5000) and test set (1000).
c. Data exploration:
   - Compute correlation between response and predictors: each predictor seems to be not really correlated with the response as the highest correlation score is only: 0.135 (predictor: kw_avg_avg).
   - Visualize relationship between response and its highest correlated predictor.
   - The visualization display some data points to be outliers; so, to avoid being influenced by them, I removed those data points.
d. Approach (using train/test set and cross-validation):
   - Linear Regression with the most correlated variable: I tried polynomial models and selected degree = 3 as the best according to p-value. And the RMSE for test data is 225.41.
   - Since there're a lot of predictors, I tried some methods to reduce the number of predictors: Forward subset, Backward subset. Both suggest slightly different number of predictors in term of different metrics: Adjusted R^2, Cp, BIC.
   - Since the subset selection suggest to use more predictors, I tried a bunch of Multiple Linear Regression models with a bunch of different predictors accordingly.
   - Then I tried Ridge and Lasso which are very applicable for data with many predictors. But, they did not really get a good result comparing to previous models.
   - Next, I tried PCR, another regression technique for data with many predictors. Again, the result was not very helpful.
   - Last, I tried cross-validation with Lasso, Ridge and PCR.
   - The best result so far is the Linear Regression with one most correlated predictor. Yet, it doesn't mean that the model is a good one because its RMSE is still big: 225.41.

Overall, I think the models built in R script is better than Google Prediction API. However, the result is not very good yet. Perhaps, the dataset is not a good match for Regression Analysis. Also, it'd be a lot more helpful if we have more domain knowledge on the dataset so that we can incorporate it with the methods that we consider as more appropriate for the analysis.