Analyzing the Popularity and Attitudes of “Fad” Diets

Vinh Tran
Capstone #1
5/31/2019
https://github.com/vinhttran/fad_diets

Project Description

Analyzing tweets mentioning “fad” diets over Memorial Day weekend by assessing their popularity and analyzing sentiment. The diets studied were: Keto, paleo, gluten-free, Whole30, low-fat, and Mediterranean.

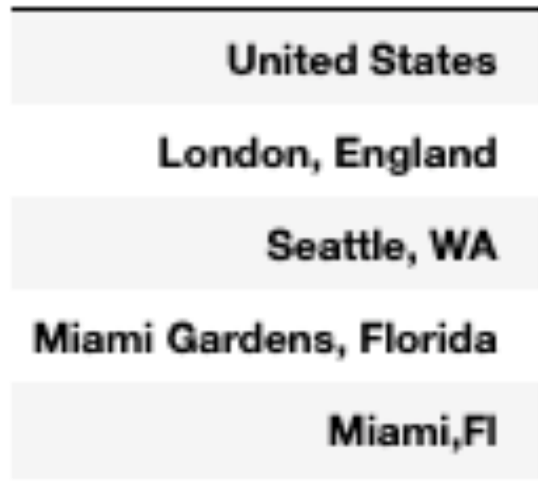
Motivation

My background is in nutrition and nutrition research. When I bring this up one of the first questions I always get is, “What do you think about <***insert fad diet***>. One limitation of traditional nutrition research is that it requires a big investment in money and time, involves human subjects, and also can’t keep up with trends.

Measuring sentiment on social media is the next frontier of guiding research. From a public health perspective, we can invest dollars into diets that are making a big impact or showing signs of promise. From a marketing perspective, companies can glean insights on what consumers are interested in.

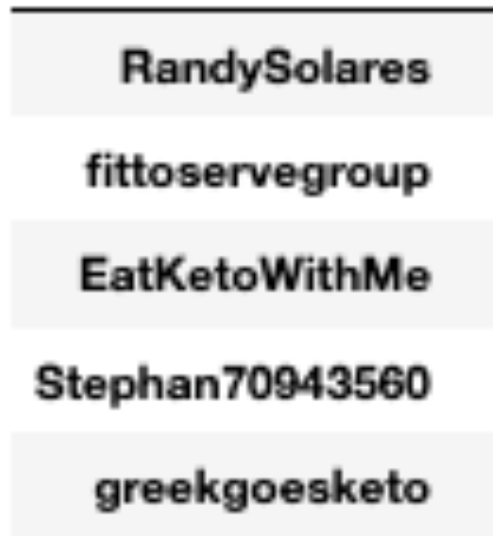
Pipeline and EDA

Top Locations

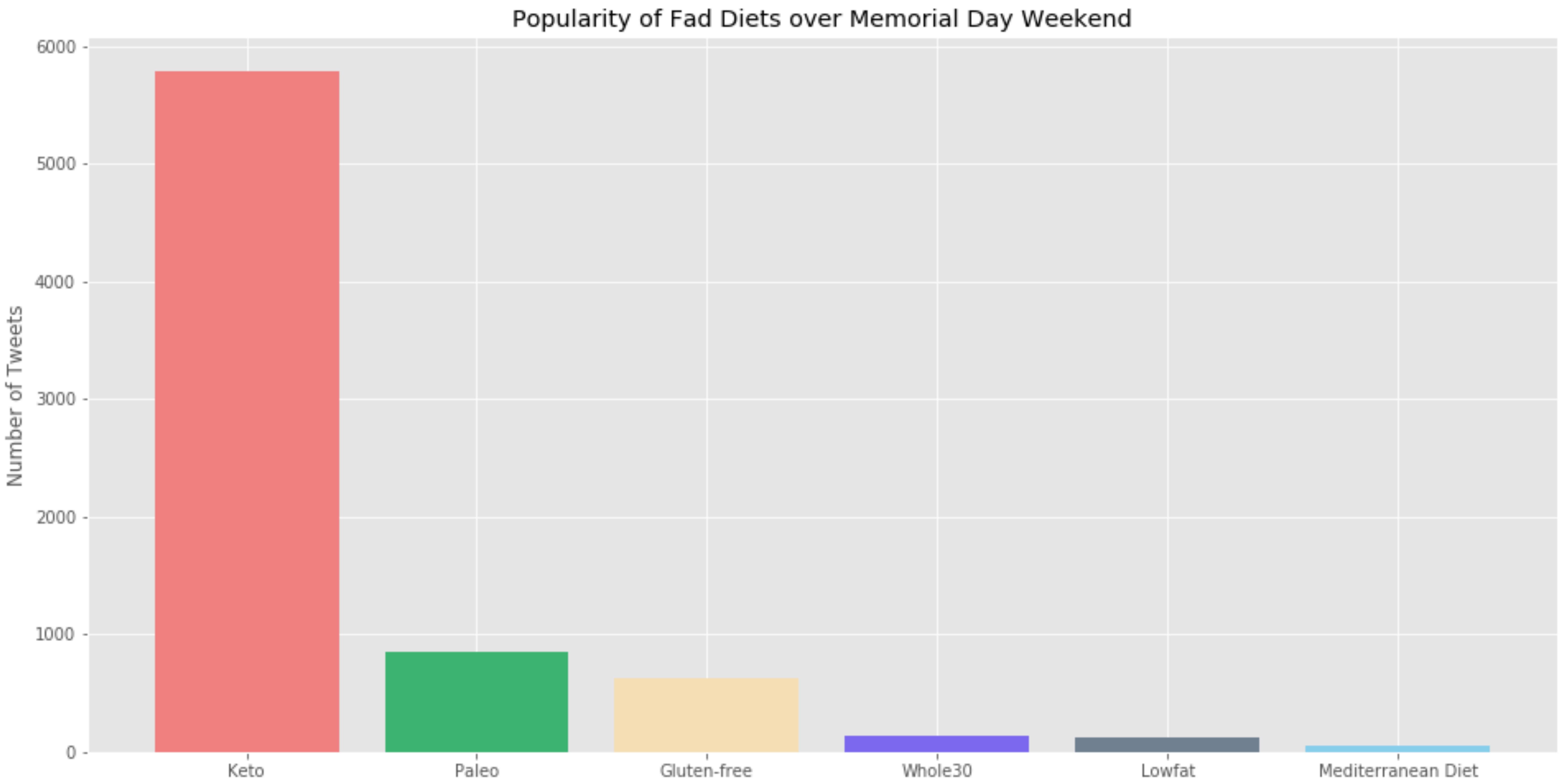


United States
London, England
Seattle, WA
Miami Gardens, Florida
Miami,Fl

Top Users



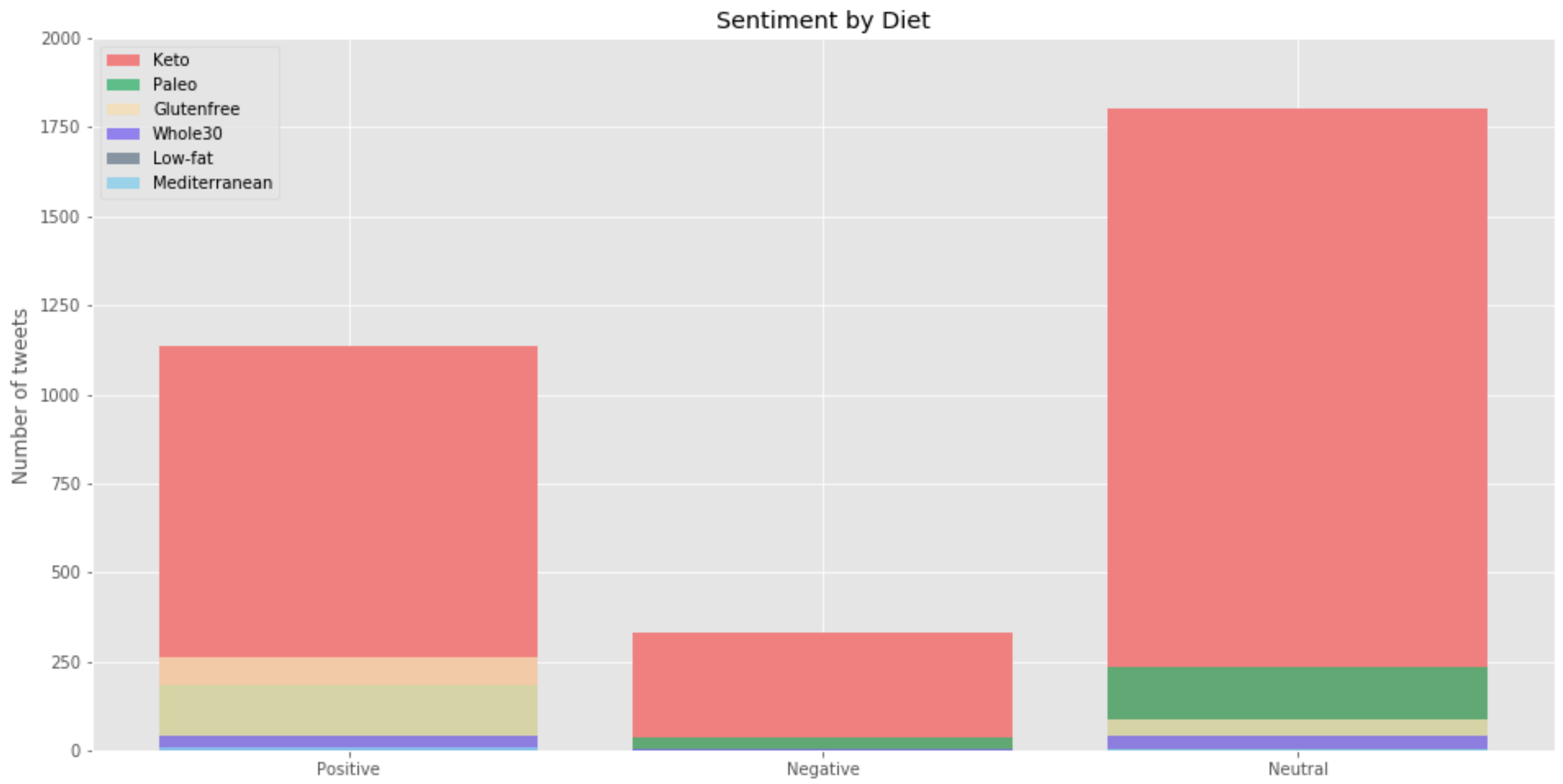
RandySolares
fittoservegroup
EatKetoWithMe
Stephan70943560
greekgoesketo

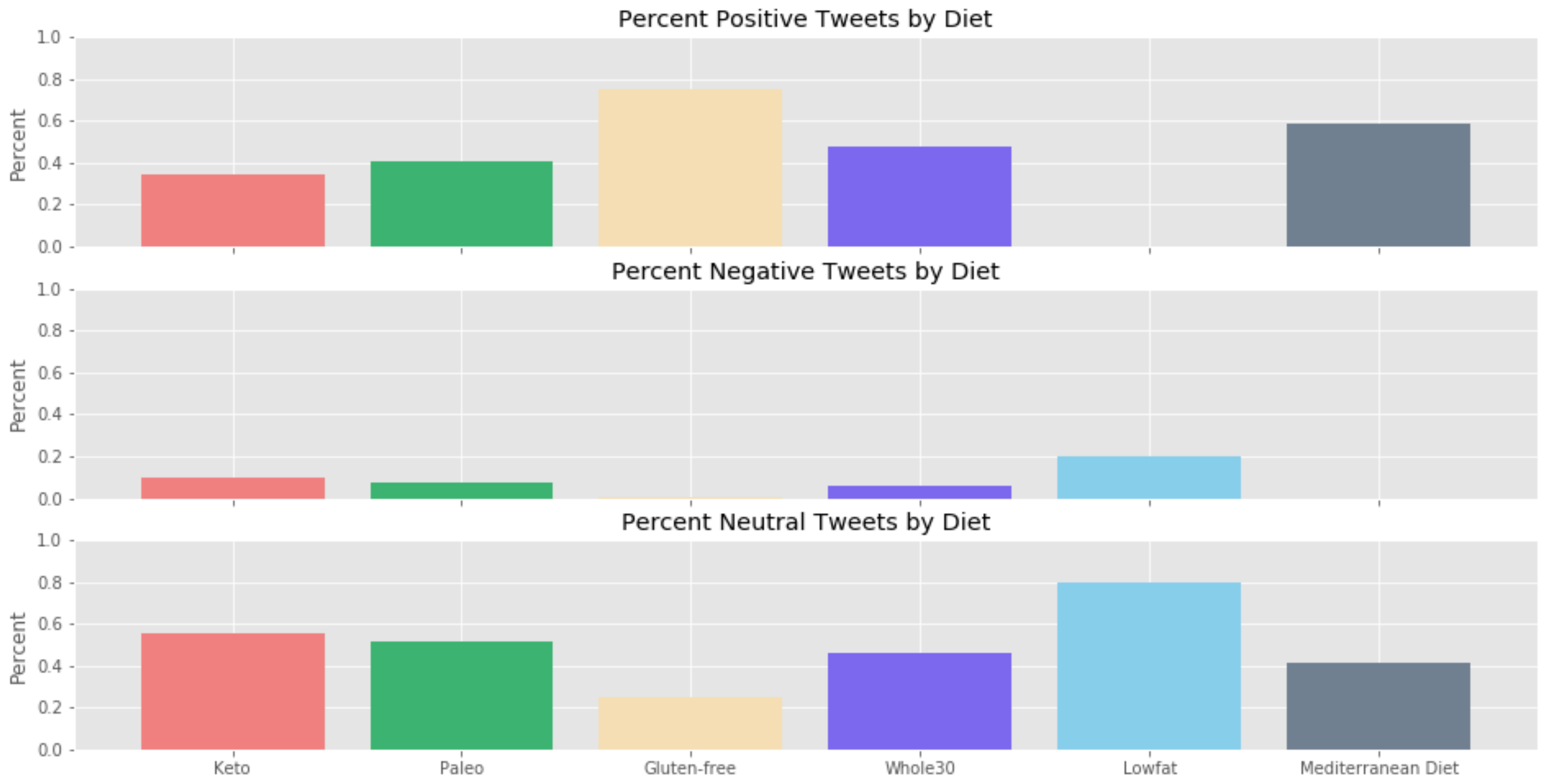


Popularity of Fad Diets over Memorial Day Weekend

This trend generally holds when using “lost” as a proxy for weight loss mentions (i.e. I lost 10 lbs. on Keto) and “start” as a proxy for starting a diet (i.e., I just started paleo!”), and for retweets

Pipeline and EDA



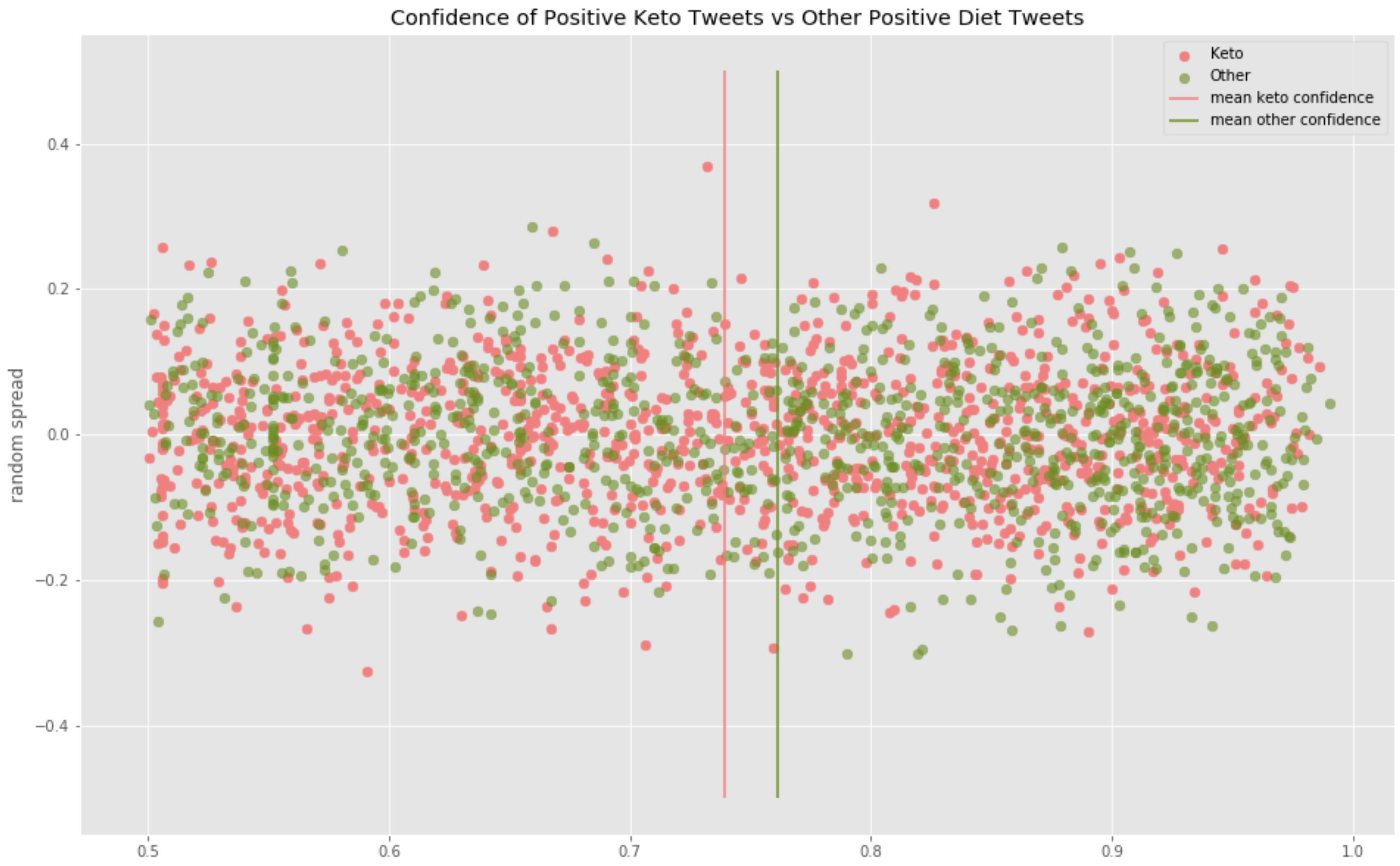


By number of positive, negative and neutral tweets, Keto has the most tweets overall (as expected)

By percentage, it appears that gluten-free has the greatest percentage of positive tweets. It also seems like people are very neutral about low-fat diet….*now time for some hypothesis testing!*

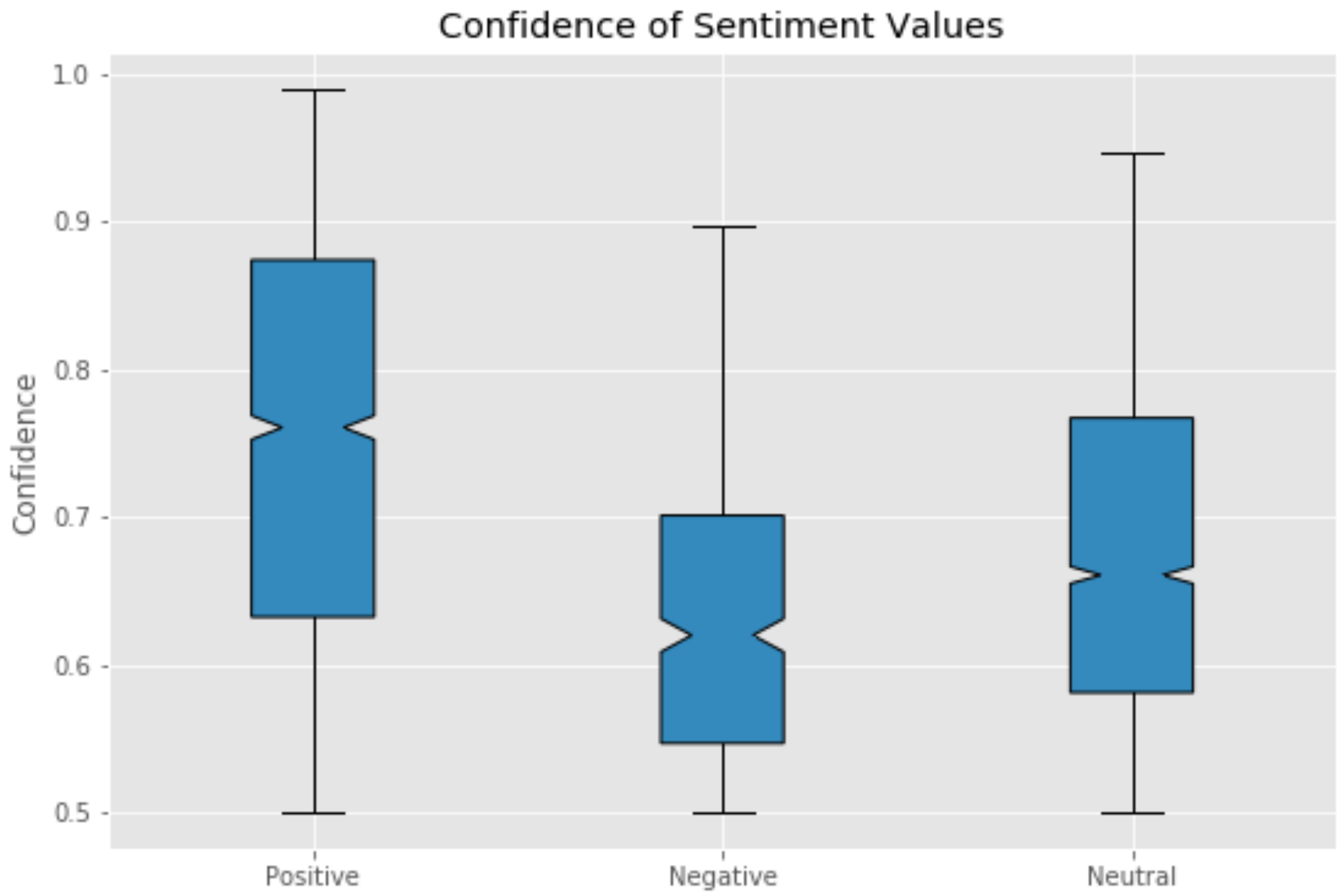
Hypothesis Testing

Hypothesis Test 1: Since the Keto diet appears to be most popular, let’s test if positive tweets about the Keto diet are more confidently positive.
H0: Positive Keto tweets are no more confidently positive than positive tweets about other diets.
H1: Positive Keto tweets are more confidently positive than positive tweets about other diets. Performed: Welch’s, U-Test, two-sample approximate test of population proportions. For all, p = 1.00.
Conclude - No difference! Not surprising looking at the scatter plot.



Hypothesis Test 2: Looking at previous slide, gluten-free has the greatest percentage of positive tweets and no negative tweets. Let’s see if gluten-free tweets are more confidently positive than positive tweets about other diets. Following the same steps as Hypothesis Test 1, I get a p-value of < 0.001

Hypothesis Test 3: As part of the EDA process I noticed that negative tweets had lower confidence than positive tweets. I tested to see if there was a difference using Welch’s and p< 0.001. Positive tweets have a significantly higher mean confidence than negative tweets.



Conclusions

In terms of popularity, the Keto diet is by far the most mentioned diet over Memorial Day weekend. This is followed by paleo, gluten-free, then Whole30. There were barely any mentions of the Mediterranean diet (considered a “gold-standard” of diets), and low-fat diets (whose time may have passed due to recent research refuting its benefits).

Positive tweets about Keto were no more confidently positive than tweets about other diets. When people are tweeting positively about Keto they may not be doing so more enthusiastically than other diets. This suggests that although Keto is the most popular, the enthusiasm is not greater than other diets. However, positive tweets about gluten-free diet is more enthusiastic. One explanation is that people on gluten-free diets do so to address a specific problem (i.e. an allergy to gluten) and the effect of going on this diet is more dramatic and instant which is reflected in the enthusiasm of the tweet.

Positive tweets are more confidently classified as positive than negative tweets. This may be a function of the Wit.ai analyzer being better at identifying positive tweets.

Next Steps

1. The tweets were collected over a very limited and specific time period - Memorial Day weekend. After filtering out retweets and non-english tweets, the sample became a quarter of the size. To perform more meaningful analysis more tweets need to be captured. This can be done using AWS and Spark to better learn these data science skills.

2. Dealing with retweets by matching on “RT” took time and was clunky. Capturing tweets with retweet_count will be easier for next time.

3. I only looked at tweets but not extended tweets. I did not realize this until I was too far into the project. Extended tweets may have increased confidence or changed the outcome of the sentiment analyzer.

4. I used Wit.ai as a “black-box.” To better understand the sentiment analysis I will need to understand how it works and perhaps explore other tools. Wit.ai only returns one value (positive, neutral, negative) and a confidence value which measures how confident the designation is. More robust output will give me more dimensions to analyze.

5. I intend on continuing to exploring this topic. I will need to refine the fad diets I am examining (i.e. maybe drop low fat and Mediterranean) and perhaps add celery juice, which has been getting a lot of buzz. I also want to be able to use geo-location data and relate this to county-level health outcomes from CDC, NHANES, or USDA data.