Mikee Jazmines
Cher Panlilio
Leina Santiago
CS 129.1 A

I. Big Data problem

Given tweets regarding pineapple on pizza, how many users believe that pineapple should or should not belong on pizza?

II. Description of source dataset

A. The following phrases were used as a datasource:
   1. Positive:
      a) “I like pineapples on pizza”
   2. Negative:
      a) “I don’t like pineapples on pizza”
      b) “I do not like pineapples on pizza”

B. Tweets will be scraped and the following data will be taken from each tweet:
   1. Username and Full Name
   2. Tweet-id
   3. Tweet text
   4. Tweet timestamp
   5. No. of likes
   6. No. of replies
   7. No. of retweets

C. Each like and each retweet of a tweet may also be recorded to see how many people agree with the tweet

D. How to Obtain the Data Set
   1. At first, the group wanted to use the API of twitter to retrieve the data set, however, twitter’s API allows the user to scrape tweets within 7 days only.
   2. To acquire the data set, we used a twitterscraper found on github by taspinar. It can be found here: https://github.com/taspinar/twitterscraper.
   3. We then went to twitter and searched for the queries: “I like pineapples on pizza”,“I don’t like pineapples on pizza”, and “I do not like pineapples on pizza”. We then copied a part of the url of each of the searched queries

on the terminal and ran the twitterscraper. The data for each query was saved as a json file.

E. Sample query codes:

1. `twitterscraper "i%20like%20pineapples%20on%20pizza"%20since%3A2017-01-01%20until%3A2017-11-30 like2017.json`
2. `twitterscraper "i%20like%20pineapples%20on%20pizza"%20since%3A2016-01-01%20until%3A2016-12-31 like2016.json`

III. Description of the output, explanation and discussion of the output in relation to the Big Data problem

A. Output:

I LIKE PINEAPPLES ON PIZZA

Yearly

YEAR	COUNT
2016	112347
2017	270910

Same month

MONTH	COUNT
January	19375
February	41314
March	66418
April	60365
May	16221
June	15105
July	33079
August	19075
September	17879

October	16063
November	73242
December	5121

Monthly (2016-2017)

YEAR	MONTH	COUNT
2016	January	659
2016	February	361
2016	March	1108
2016	April	483
2016	May	771
2016	June	872
2016	July	969
2016	August	2512
2016	September	1586
2016	October	1278
2016	November	46236
2016	December	1646
2017	January	4718
2017	February	9535
2017	March	6764
2017	April	7346

2017	May	3793
2017	June	4135
2017	July	4691
2017	August	3362
2017	September	5475
2017	October	2484
2017	November	14125

I DON'T LIKE PINEAPPLES ON PIZZA

Yearly

YEAR	COUNT
2016	58481
2017	66246

Same month

MONTH	COUNT
January	5377
February	9714
March	7872
April	7829
May	4564
June	5007

July	5660
August	5874
September	7061
October	3762
November	60361
December	1646

Monthly (2016-2017)

YEAR	MONTH	COUNT
2016	January	2401
2016	February	2185
2016	March	23298
2016	April	1978
2016	May	2085
2016	June	2719
2016	July	3380
2016	August	7571
2016	September	4586
2016	October	5016
2016	November	52367
2016	December	5121

2017	January	17334
2017	February	39129
2017	March	43120
2017	April	58387
2017	May	14136
2017	June	12386
2017	July	29699
2017	August	11504
2017	September	13293
2017	October	11047
2017	November	20875

B. Analysis:

1. The trend of when people post a lot about it (high point and low point)
    a) Compare by year (2016 vs 2017), month
        (1) In 2017, there were generally more tweets about the pineapples on pizza debate, with the peak month being **April** of that year. For 2016, more people tweeted about pineapples belonging or not belonging on pizza in **November**.

        The first surge in tweets occurred in February, and the number of tweets continued to increase until it reached it's peak in April. This is possibly because articles were being produced at the time talking about the pineapples on pizza debate. With more articles, comes more discussion,

therefore the debate really sparked at this time perhaps because news outlets were giving it more attention.

(2) Together, there were more people participating in this debate during the month of **November**.

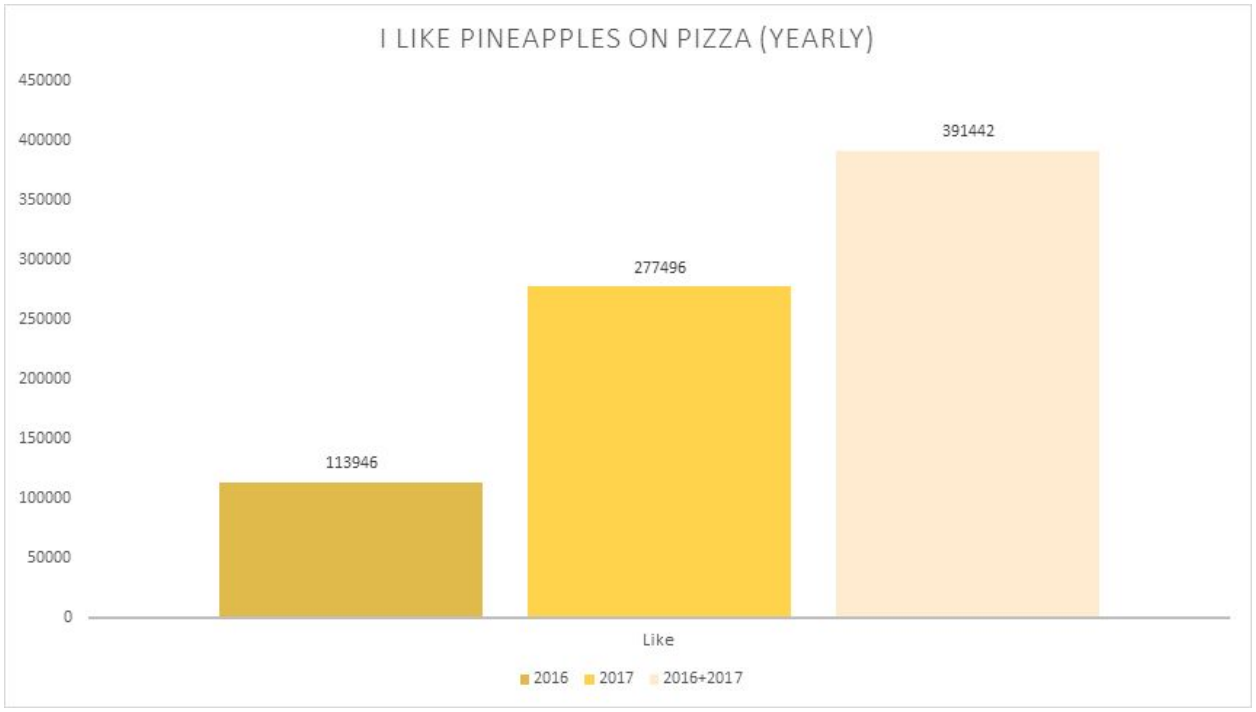
b) Compare the trends

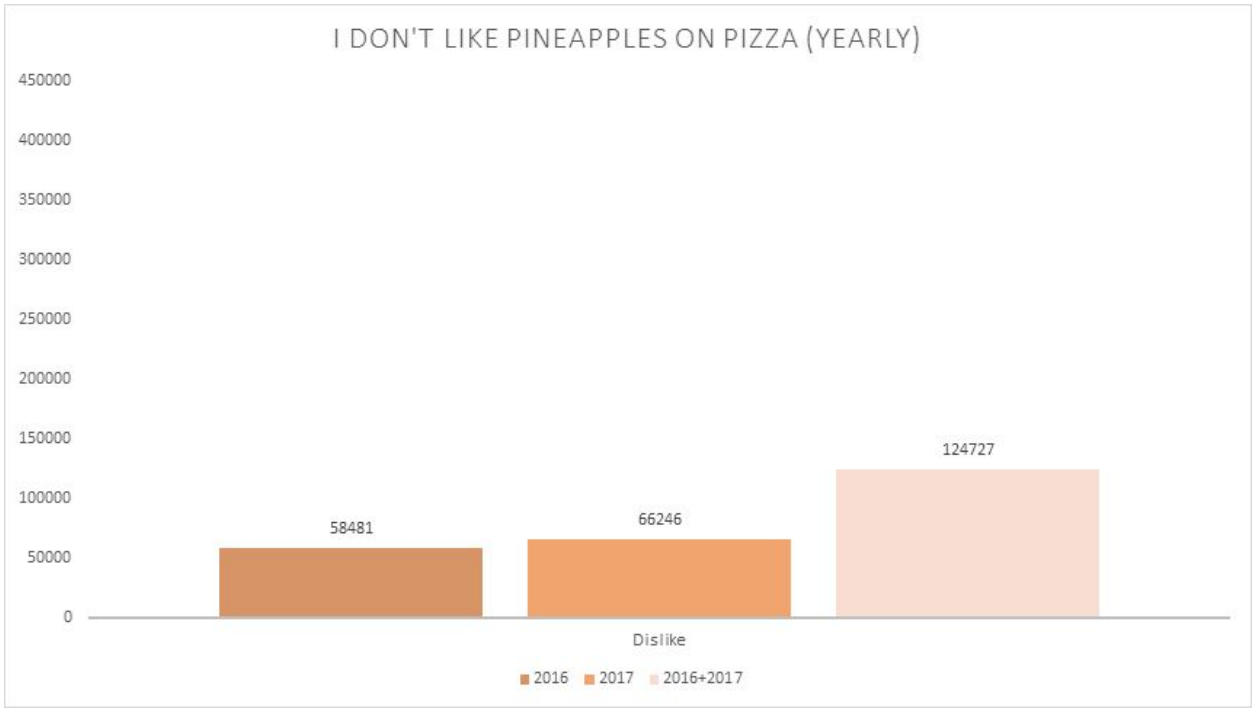
(1) Check if there is a direct relationship

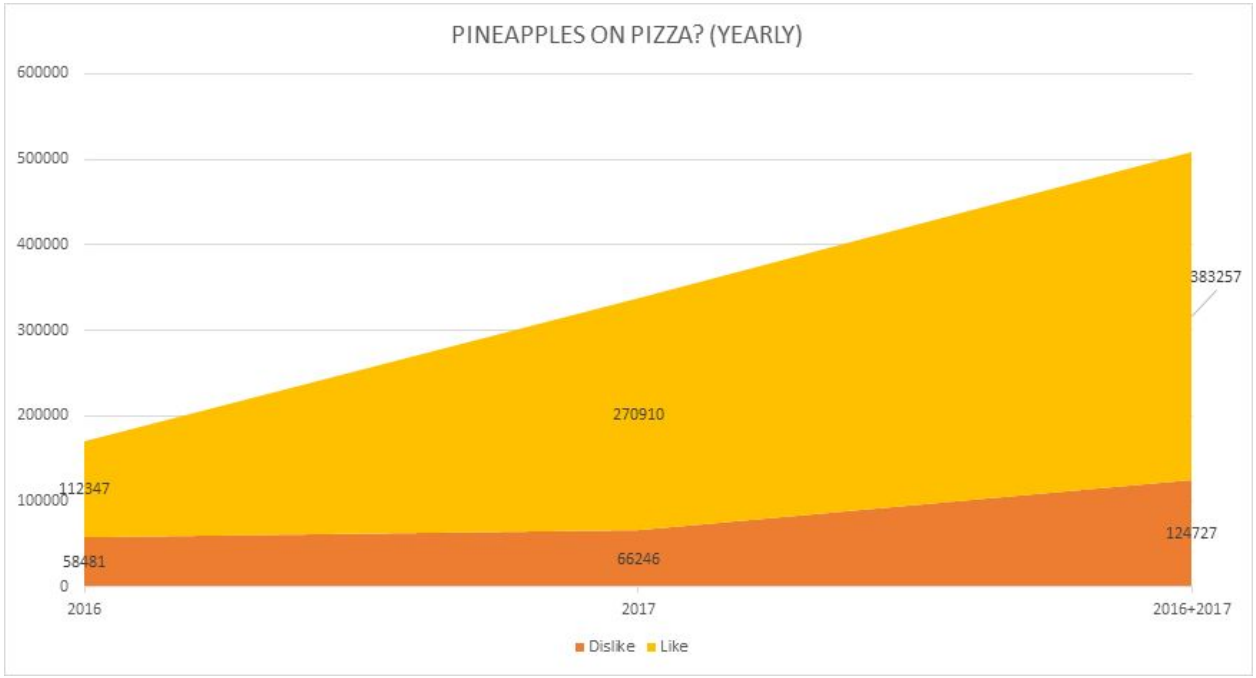
(a) In almost all the months, the tweets of those who liked pineapples on pizza, and the tweets of those who disliked pineapples on pizza had a **direct relationship**. As seen in the data, for every increase or decrease in tweets from the previous month for one side of the users, the other side will also increase or decrease depending on the trend.

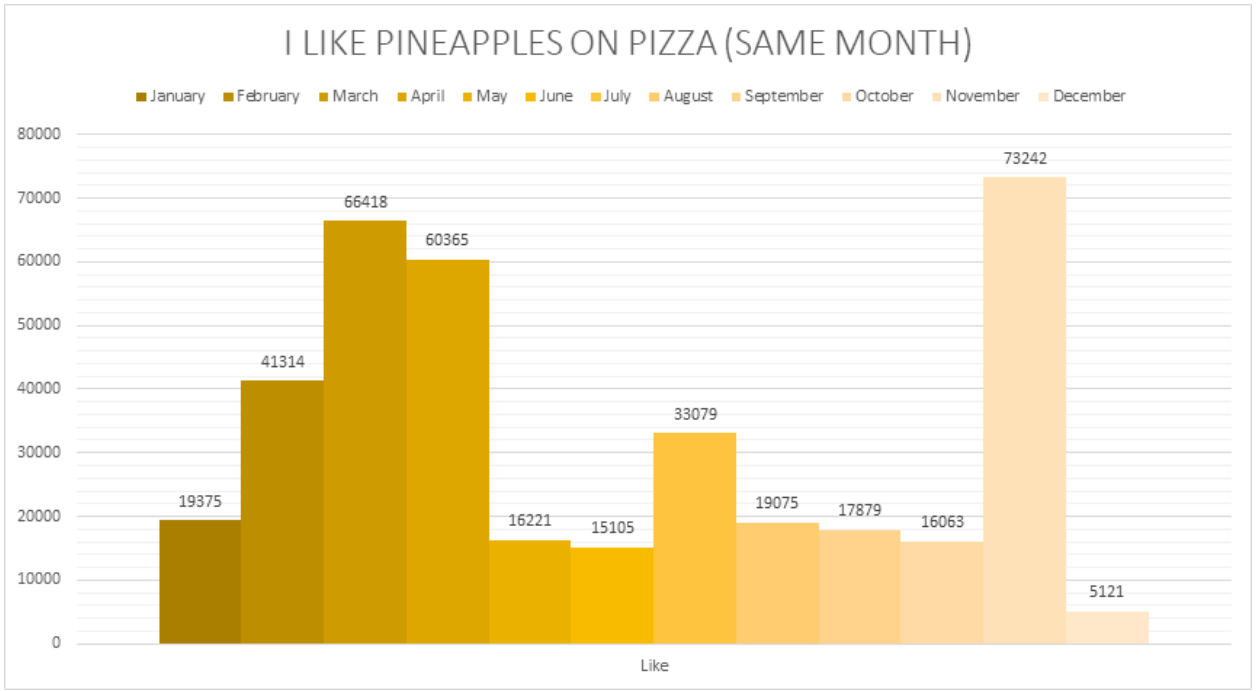
This could have been made possible because of the conflicting sides always trying to gain the upperhand in the debate. If one party tweets about liking pineapples on pizza, this will in turn cause those who dislike pineapples on pizza to tweet as well.

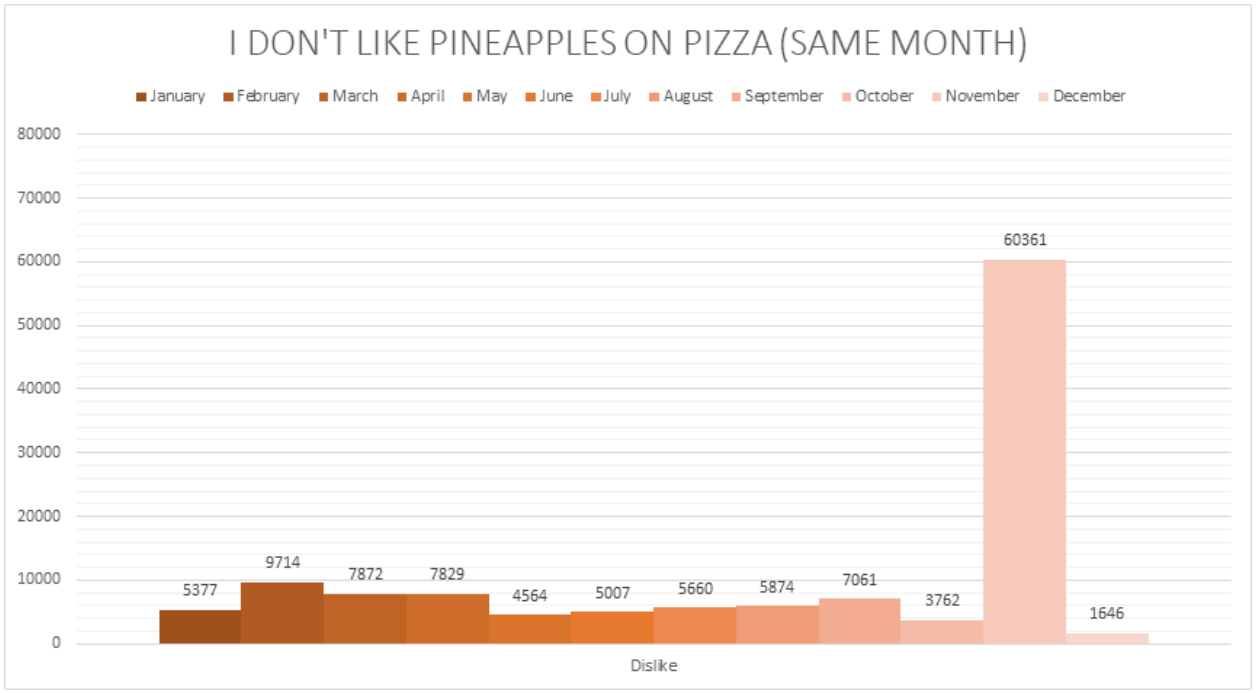
IV. Visualization of the output

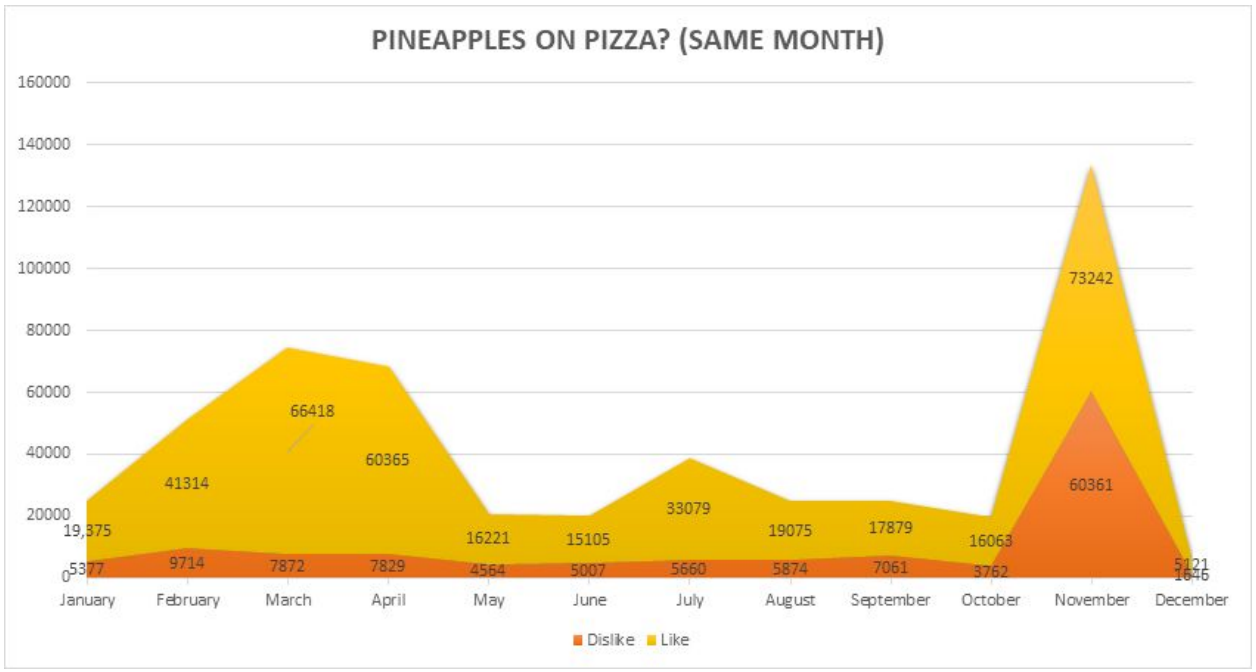


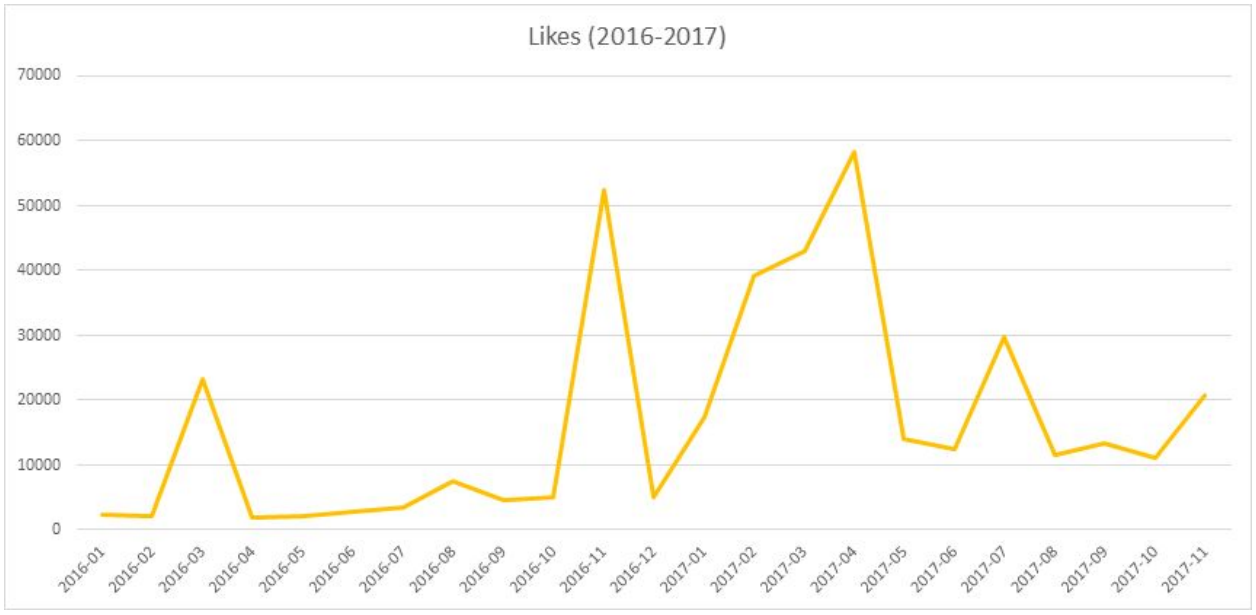


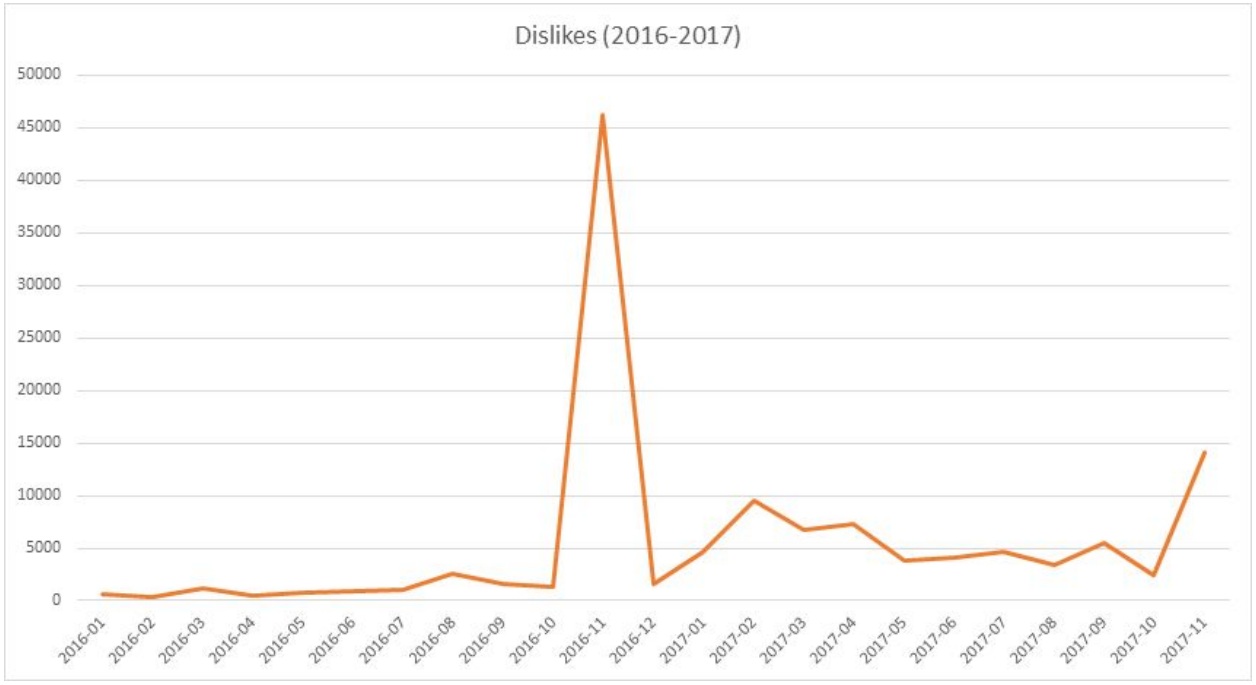

PINEAPPLES ON PIZZA? (YEARLY)
600000
500000
400000
300000
200000
100000
0
383257
270910
112347
124727
58481
66246
2016
2017
2016+2017
Dislike
Like

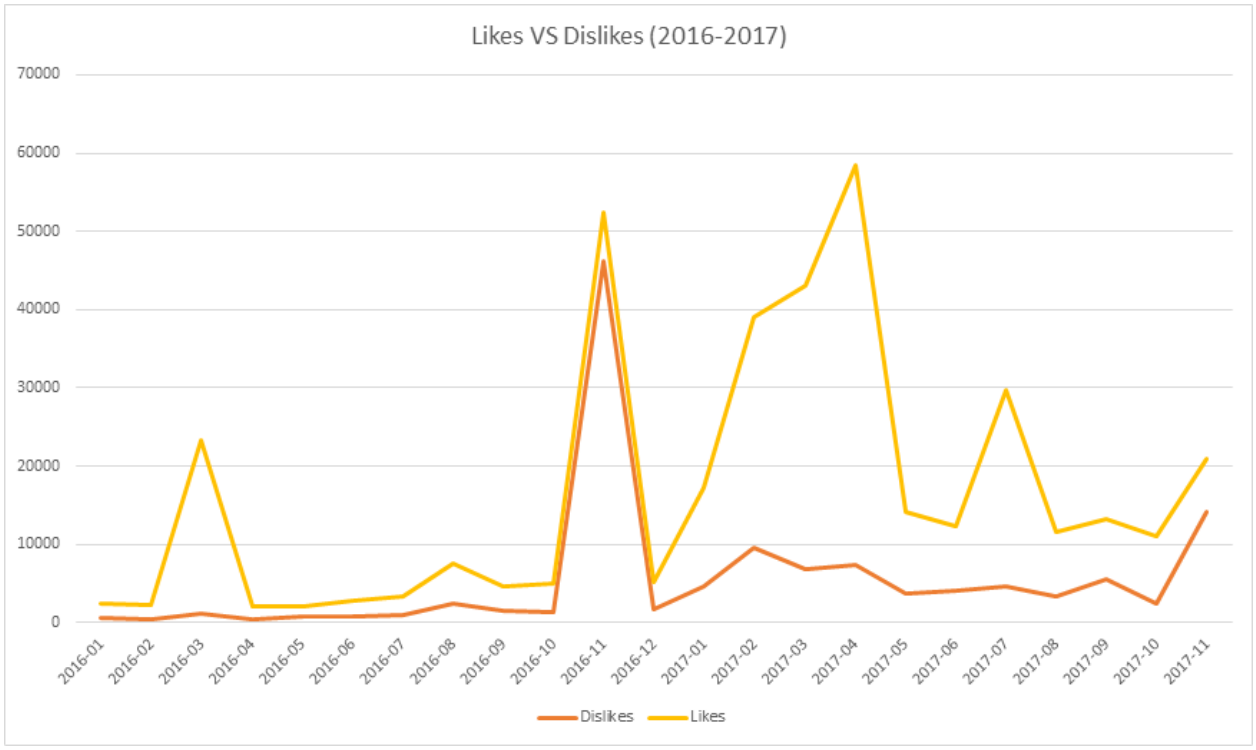

I LIKE PINEAPPLES ON PIZZA (SAME MONTH)
January
February
March
April
May
June
July
August
September
October
November
December
80000
70000
60000
50000
40000
30000
20000
10000
0
19375
41314
66418
60365
16221
15105
33079
19075
17879
16063
73242
5121
Like


I DON'T LIKE PINEAPPLES ON PIZZA (SAME MONTH)
January
February
March
April
May
June
July
August
September
October
November
December
80000
70000
60000
50000
40000
30000
20000
10000
0
5377
9714
7872
7829
4564
5007
5660
5874
7061
3762
60361
1646
Dislike


PINEAPPLES ON PIZZA? (SAME MONTH)
160000
140000
120000
100000
80000
60000
40000
20000
0
19,375
41314
66418
60365
16221
15105
33079
19075
17879
16063
73242
5377
9714
7872
7829
4564
5007
5660
5874
7061
3762
60361
5121
1646
January
February
March
April
May
June
July
August
September
October
November
December
Dislike
Like


Likes (2016-2017)
70000
60000
50000
40000
30000
20000
10000
0
2016-01
2016-02
2016-03
2016-04
2016-05
2016-06
2016-07
2016-08
2016-09
2016-10
2016-11
2016-12
2017-01
2017-02
2017-03
2017-04
2017-05
2017-06
2017-07
2017-08
2017-09
2017-10
2017-11


Dislikes (2016-2017)
50000
45000
40000
35000
30000
25000
20000
15000
10000
5000
0
2016-01
2016-02
2016-03
2016-04
2016-05
2016-06
2016-07
2016-08
2016-09
2016-10
2016-11
2016-12
2017-01
2017-02
2017-03
2017-04
2017-05
2017-06
2017-07
2017-08
2017-09
2017-10
2017-11


Likes VS Dislikes (2016-2017)
70000
60000
50000
40000
30000
20000
10000
0
2016-01
2016-02
2016-03
2016-04
2016-05
2016-06
2016-07
2016-08
2016-09
2016-10
2016-11
2016-12
2017-01
2017-02
2017-03
2017-04
2017-05
2017-06
2017-07
2017-08
2017-09
2017-10
2017-11
Dislikes
Likes