
ISyE 2028 – Basic Statistical Methods - Fall 2015
Bonus Project: “Big” Data Analytics
Proposal (or Final Report)
“Yik Yak Traffic” by Mary Alyce Martin

Abstract:

For my bonus project, I decided to conduct a study that would help provide insight into the times when the app Yik Yak, popular among college students, is used here on Georgia Tech’s campus. The application is based on the idea of an anonymous feed of funny statements and statuses submitted by users and anonymously “up-voted” and “down-voted” by other users. Each user can up-vote any post that they like or support and likewise down-vote those that they oppose. The number of votes are then tallied and displayed to the right of each post respectively. These posts are often called “yaks.” I carried out my study by recording the number of up-votes awarded to the top three “yaks” four times a day for thirty days. This data would help to determine the busiest time of day for the app. I expected the app to be more heavily used in the later part of the day and late at night, considering the strange sleep schedules typical of college students. With almost nocturnal sleeping patterns, I was expecting to see the midnight traffic to be heavier than any other time of day.

Methods:

I checked the app at 10:00 am, 2:00 pm, 6:00 pm, and 12:00 midnight. Originally I planned to also check the number of yaks on the entire feed at each of these times as well, instead of just the number of up-votes on the most popular posts, but I then realized that this information is not readily available or displayed on the app. To remind myself to check on this data four times a day, I set daily alarms on my phone that would keep me from forgetting to gather my data. It was vitally important for me to collect plenty of data for each time of day, specifically for at least 30 days. Thanks to the central limit theorem, this would also ensure that I could use the standard Z table for finding test statistics and interpreting my findings, simplifying my calculations and deeming the student t statistic unnecessary except when dealing with the isolated sample of weekend data. Being a college student myself, I also found no problem with staying awake for the necessary times. I myself am on a sleep schedule where I wake up relatively late and go to bed past midnight, a pattern atypical of a working adult or grade-school student. I also had to be very careful with any inference I made about weekend Yik Yak traffic as opposed to weekday traffic considering the very small sample size of weekend data that I was able to collect in this time frame. I stored my observations in a spread sheet, which I updated daily, and found this to be an effective method for organizing the numbers.

Results:

My results were slightly different from what I expected to observe.

Up-Votes for Most, 2nd Most, and 3rd Most Popular Yaks

Date	10 AM	2 PM	6 PM	12 AM
10/30	118, 109, 75	203, 200, 165	122, 117, 104	97, 88, 84
10/31	205, 201, 167	199, 194, 188	130, 120, 112	102, 96, 88
11/1	145, 131, 105	186, 145, 143	109, 100, 98	125, 113, 100
11/2	189, 175, 166	214, 168, 115	120, 118, 101	112, 101, 93
11/3	213, 133, 119	174, 162, 136	130, 117, 109	98, 77, 60
11/4	200, 112, 95	229,185,93	151, 132, 119	133, 129, 127
11/5	144, 121, 97	212, 160, 154	125, 122, 121	140, 120, 114
11/6	156, 129, 110	178, 105, 105	132, 125, 103	149, 140, 120
11/7	170, 145, 99	146, 68, 52	128, 116, 104	130, 105, 78
11/8	149, 144, 132	169, 145, 129	130, 127, 115	99, 89, 85
11/9	187 ,177 ,168	227, 187, 133	122, 121, 97	68, 57, 57
11/10	146, 131, 114	205, 142, 113	116, 112, 107	103, 86, 62
11/11	95, 75, 71	186, 119,114	137, 110, 108	114, 84, 73
11/12	97, 77, 68	153, 142, 130	109, 102, 86	113, 95, 60
11/13	110, 108, 100	195, 130, 98	129, 120, 111	95, 87, 82
11/14	99, 93, 89	157, 134, 128	119, 103, 101	148, 122, 101
11/15	112, 90, 84	217, 202, 123	140, 116, 98	125, 116, 107
11/16	150, 145, 110	196, 149, 141	153, 140, 112	110, 107, 98
11/17	121, 100, 92	185, 170, 120	211, 167, 155	177, 103, 87
11/18	134, 96, 84	173, 111, 93	130, 130, 116	128, 90, 81
11/19	144, 116, 100	205, 103, 99	114, 113, 109	115, 86, 58
11/20	187, 101, 94	160, 159, 118	162, 120, 100	97, 95, 82
11/21	102, 57, 46	177, 145, 133	154, 150, 114	100, 87, 65
11/22	143, 136, 130	212, 134, 124	148, 142, 105	136, 122, 101
11/23	97, 90, 82	207, 195, 156	135, 121, 111	140, 109, 96
11/24	135, 114, 105	201, 188, 167	127, 125, 113	110, 90, 80
11/25	102, 98, 81	199, 145, 139	141, 129, 120	120, 111, 110
11/26	73, 63, 56	200, 197, 184	124, 99, 90	114, 69, 48
11/27	95, 91, 68	189, 176, 162	113, 110, 102	117, 70, 65
11/28	118, 92, 90	195, 173, 171	290, 199, 190	104, 91, 86

To carry out my calculations, I summed the number of up-votes for the three top yaks at each time on each day and found an average for each of the four time periods, as well as the sample standard deviation.

Sum of Up-Votes for Top 3 Most Popular Yaks

Date	10 AM	2 PM	6 PM	12 AM
10/30	302	568	343	269
10/31	573	581	362	286
11/1	381	474	307	338
11/2	530	497	339	306
11/3	465	472	356	235
11/4	407	507	402	389
11/5	362	526	368	374
11/6	395	388	360	409
11/7	414	266	348	313
11/8	425	443	372	273
11/9	532	547	340	182
11/10	391	460	335	251
11/11	241	419	355	271
11/13	318	423	360	264
11/14	281	419	323	371
11/15	286	542	354	348
11/16	405	486	405	315
11/17	313	475	533	367
11/18	314	377	376	299
11/19	360	407	336	259
11/20	382	437	382	274
11/21	205	455	418	252
11/22	409	470	395	359
11/23	269	558	367	345
11/24	354	556	365	280
11/25	281	483	390	341
11/26	192	581	313	231
11/27	254	527	325	252
11/28	300	539	679	281
Mean	352.8	476.933	373.5	300.067
Sample SD	94.846	70.786	72.077	53.902

I then constructed four 95% two-sided confidence intervals using these statistics, as well as the z-score 1.96 that corresponds with this level of confidence.

	10 AM	2 PM	6 PM	12 AM
Confidence Interval for Sum of Up-Votes for Top 3 Yaks	(318.86, 386.74)	(451.603, 502.263)	(347.708, 399.292)	(281.381, 319.959)

I next took the data from the weekend days and found 95% two-sided confidence intervals for these to compare to the original confidence intervals. With a sample size of only 9 weekend days, I had to use the t-table in my calculations with 8 degrees of freedom and .025 area, which yielded a t-value equivalent to 2.31 for this case.

	10 AM	2 PM	6 PM	12 AM
Mean	363.778	465.444	395.333	313.444
Sample SD	108.723	91.697	111.59	42.430
Confidence Interval	(280.06, 447.49)	(394.84, 536.05)	(309.41, 481.26)	(280.77, 346.12)

Conclusion:

I was surprised to see, when comparing these confidence intervals, that they indicated midnight is the least popular time to use Yik Yak. I was also not expecting the data to suggest 2 PM is a time of high traffic for this app. These intervals were also fairly wide, which is a result of the large sample standard deviations and moderate sample size. To reduce the width of the intervals, I could either increase the sample size or decrease the level of confidence. I found that the intervals for the weekend days were really too wide to be of use for me and showed no clear pattern that would indicate behavior on Yik Yak usage is different from weekdays to weekends. One of the major assumptions of my study was that the number of up-votes was directly proportional to the number of users active on the app at a given time, but this might not be accurate. I also must take caution in reporting these results due to a smaller sample size than what would be ideal for such a study as a result of a limited time frame for collecting my data.