

## Capstone 2 Project Milestone 1 Report

To say that 2020 has been an exceptional year would be an understatement. Covid19 put the whole world in lockdown, derailing economies while, at the same time, turning masks into fashion statements. There hasn't been a time when people all over the world go into self-seclusion for an entire year such as we have done.

### Some interesting questions to explore are:































- *What happens when we live in seclusion for an extended period of time?*
- *What behavioral challenges and changes occurred during this time?*
- *What new trends emerged, either positive or negative?*
- *Was there an increase in stress? Depression? Fear? Or Loneliness?*
- *Do people experience different stages in seclusion similar to the stages of grief? If so, what are these stages?*

Data collected in 2020 could be a useful source to answer these questions. This could be useful information to help discover issues or attitudes and reshape policies or procedures that can help **corporations** or **agencies** as they consider work from home as a viable temporary or permanent mode of operation.

### Initial Exploration of the Dataset

The **data** we will start with is from a kaggle dataset: <https://www.kaggle.com/smid80/coronavirus-covid19-tweets>

We have a set of 30 CSV files **scraped from twitter from March 30th to April 30th** - the period of time immediately following the shut down due to increases in Coronavirus transmission.

 2020-03-30 Coronavirus Tweets.CSV	204,112 KB	 2020-04-16 Coronavirus Tweets.CSV	194,240 KB
 2020-03-31 Coronavirus Tweets.CSV	234,792 KB	 2020-04-17 Coronavirus Tweets.CSV	181,944 KB
 2020-04-01 Coronavirus Tweets.CSV	207,864 KB	 2020-04-18 Coronavirus Tweets.CSV	134,571 KB
 2020-04-02 Coronavirus Tweets.CSV	204,713 KB	 2020-04-19 Coronavirus Tweets.CSV	120,619 KB
 2020-04-03 Coronavirus Tweets.CSV	189,412 KB	 2020-04-20 Coronavirus Tweets.CSV	147,660 KB
 2020-04-04 Coronavirus Tweets.CSV	162,010 KB	 2020-04-21 Coronavirus Tweets.CSV	150,119 KB
 2020-04-05 Coronavirus Tweets.CSV	158,178 KB	 2020-04-22 Coronavirus Tweets.CSV	149,129 KB
 2020-04-06 Coronavirus Tweets.CSV	205,846 KB	 2020-04-23 Coronavirus Tweets.CSV	140,497 KB
 2020-04-08 Coronavirus Tweets.CSV	185,464 KB	 2020-04-24 Coronavirus Tweets.CSV	154,604 KB
 2020-04-09 Coronavirus Tweets.CSV	169,416 KB	 2020-04-25 Coronavirus Tweets.CSV	114,756 KB
 2020-04-10 Coronavirus Tweets.CSV	152,229 KB	 2020-04-26 Coronavirus Tweets.CSV	110,532 KB
 2020-04-11 Coronavirus Tweets.CSV	115,286 KB	 2020-04-27 Coronavirus Tweets.CSV	142,502 KB
 2020-04-12 Coronavirus Tweets.CSV	90,263 KB	 2020-04-28 Coronavirus Tweets.CSV	127,750 KB
 2020-04-13 Coronavirus Tweets.CSV	87,025 KB	 2020-04-29 Coronavirus Tweets.CSV	140,805 KB
 2020-04-14 Coronavirus Tweets.CSV	159,867 KB	 2020-04-30 Coronavirus Tweets.CSV	128,251 KB

The data itself contains almost 4 GB of tweets.

We **begin by combining the csv** files into one dataframe to begin exploration of the data.

```
extension = 'CSV'
all_filenames1 = [i for i in glob.glob('*.{}'.format(extension))]

#combine all files in the list
combined_csv1 = pd.concat([pd.read_csv(f) for f in all_filenames1])
```

**Initial exploration** of the data columns give us a few ideas:

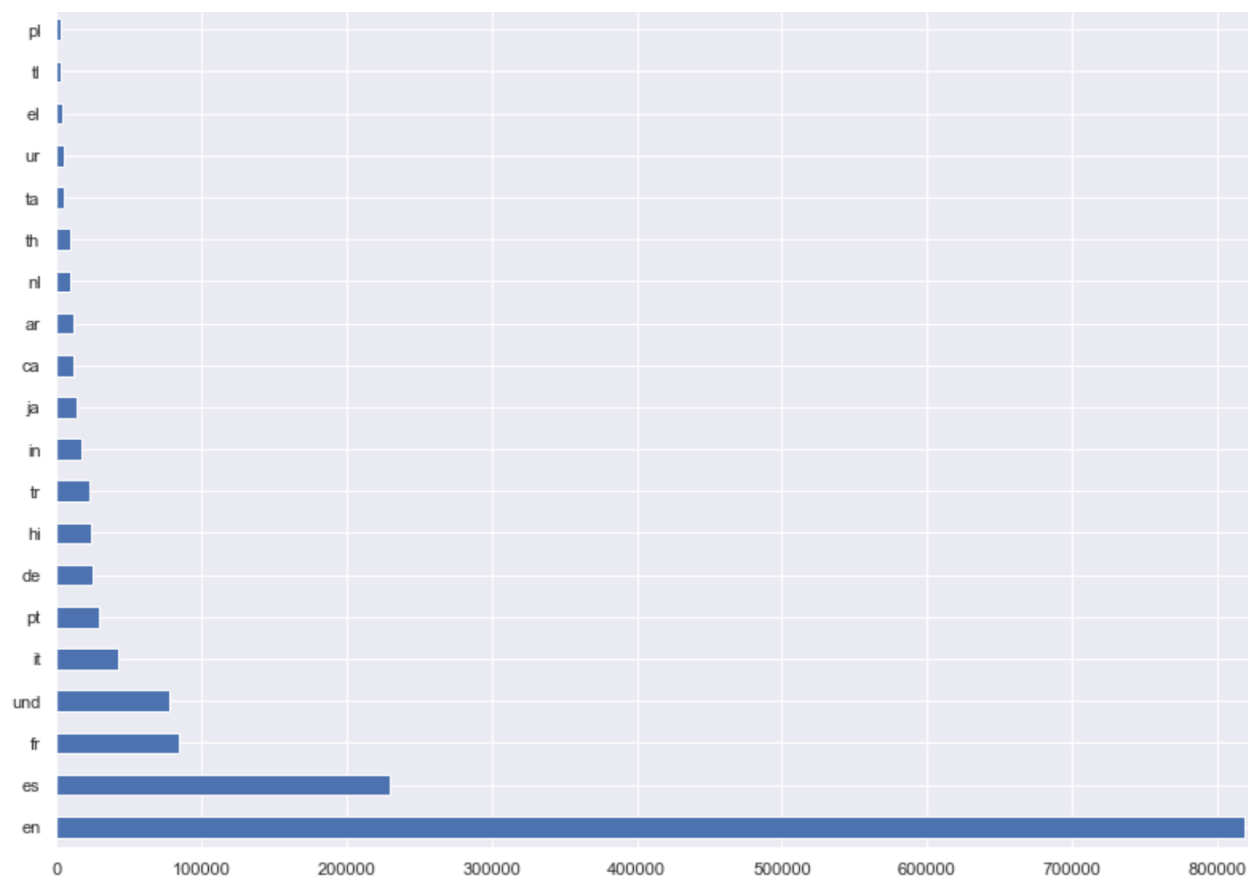
```
combined_csv1.info()

<class 'pandas.core.frame.DataFrame'>
Int64Index: 7160008 entries, 0 to 449491
Data columns (total 22 columns):
 #   Column                Dtype
---  -
 0   status_id             int64
 1   user_id               int64
 2   created_at            object
 3   screen_name           object
 4   text                  object
 5   source                object
 6   reply_to_status_id    float64
 7   reply_to_user_id      float64
 8   reply_to_screen_name  object
 9   is_quote              bool
10  is_retweet             bool
11  favourites_count       int64
12  retweet_count          int64
13  country_code           object
14  place_full_name        object
15  place_type             object
16  followers_count        int64
17  friends_count          int64
18  account_lang           float64
19  account_created_at     object
20  verified               bool
21  lang                   object
dtypes: bool(3), float64(3), int64(6), object(10)
memory usage: 1.1+ GB
```

**First observations**, we have almost 7.16 million rows of data taking 1.1 GB of memory. The files are large and may present some challenges with computational resources. The tweets are in the **'text'** column.

We take a look at the *'lang'* column to see the distribution of languages:

```
combined_csv1['lang'].value_counts()[:20].plot(kind='barh')
```



Most of the tweets are in english with spanish being a distant second with the rest trailing behind.

```
#cleaning up the data a bit
truncated_df_eng = combined_csv1[combined_csv1.lang == 'en']
truncated_df_eng = truncated_df_eng[['user_id', 'created_at', 'screen_name', 'text', 'is_quote',
                                     'is_retweet', 'favourites_count', 'retweet_count', 'followers_count', 'friends_count']]
```

We isolate english language tweets for analysis and drop unused columns.

```
truncated_df_eng.describe()
```

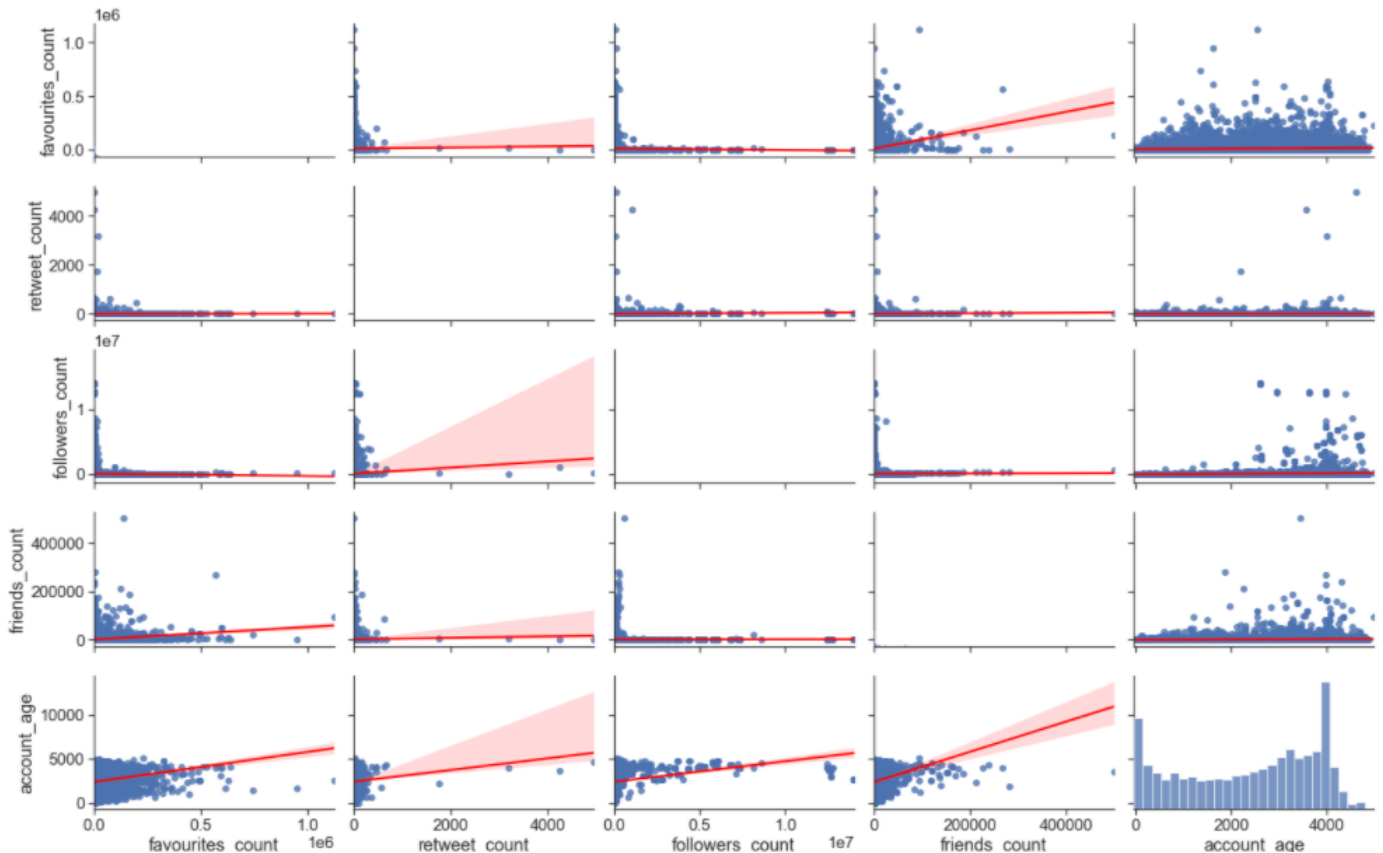
	user_id	favourites_count	retweet_count	followers_count	friends_count
count	8.187250e+05	8.187250e+05	818725.000000	8.187250e+05	8.187250e+05
mean	3.333511e+17	1.323988e+04	3.588336	5.730710e+04	2.273526e+03
std	4.912573e+17	3.982695e+04	82.064074	5.894429e+05	1.156747e+04
min	2.650000e+02	0.000000e+00	0.000000	0.000000e+00	0.000000e+00
25%	1.377155e+08	2.550000e+02	0.000000	1.410000e+02	1.580000e+02
50%	1.151877e+09	1.877000e+03	0.000000	7.870000e+02	5.480000e+02
75%	8.779354e+17	9.266000e+03	1.000000	4.072000e+03	1.658000e+03
max	1.250570e+18	1.989070e+06	26508.000000	8.121826e+07	1.496242e+06

We drop down to a little over 800,000 rows of data. Let's see what correlations exist between the data.

## Checking for Correlations

We plot a simple regression line  $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1$  to see how well correlated each column is to the other.

```
# Scanning for correlations on a small sample of the data
sns.set(style="ticks", color_codes=True, font_scale=1.5)
data = truncated_df_eng.sample(n = 20000)
data = data[['favourites_count', 'retweet_count', 'followers_count', 'friends_count', 'account_age']]
sns.pairplot(data, kind="reg", aspect=16/10, plot_kws={'line_kws':{'color':'red'}})
plt.show()
```



We see a **correlation** between *friends\_count* and *favourites\_count*. What's interesting is there is **no correlation** between *friends\_count* and *retweets\_count*. Also interesting is there's **no correlation** between *followers\_count* and *favourites\_count* or *retweets\_count*. What we see is friends will most likely favorite a tweet, which makes sense, but not necessarily retweet. Followers, on the other hand, are less likely to favorite a tweet but more likely to retweet than friends. These are insightful, however we need to delve into the text column in order to answer some of the questions we started with.

First, let's take a look at the top retweeted and favorited tweets and see if we can find topics that are coming to the top of the list of conversations.

## Most Retweeted and Most Favorited Tweets

```
df_toptweets = truncated_df_eng[['retweet_count', 'favourites_count', 'screen_name', 'text']]
top_retweets = df_toptweets.sort_values(by = 'retweet_count', ascending = False)
top_retweets[:16]
```

screen_name	text
AdamMilstein	#coronavirus News Alert: Dr. Vladimir Zelenko, a board-certified family practitioner in NY, has now treated 699 Covid-19 patients with 100% success using Hydroxychloroquine Sulfate, Zinc and Z-Pak. All symptoms of shortness of breath resolved within 4-6 hr <a href="https://t.co/siCvNg845Q">https://t.co/siCvNg845Q</a> <a href="https://t.co/lOrUvBHp0A">https://t.co/lOrUvBHp0A</a>
BenjAlvarez1	This is how Angela Merkel explained the effect of a higher #covid19 infection rate on the country's health system.\n\nThis part of today's press conf was great, so I just added English subtitles for all non-German speakers. #flattenthecurve <a href="https://t.co/VzBLdh16kR">https://t.co/VzBLdh16kR</a>
DineshDSouza	Something doesn't add up. The global panic and extreme action seem massively disproportionate to what we know about #CoronaVirus (for instance, its mortality rate). So either we are overreacting to a ridiculous degree, or they are not telling us something
eileenguo	#Wuhan residents estimate, based on calculations of cremations and urns now being returned to families, that between 42k-46k (!) died in city + surrounding areas in the 2.5 months of lockdown. Far more than official figure of 2535 deaths. #COVID19 \n\n <a href="https://t.co/L1OsFv0VEf">https://t.co/L1OsFv0VEf</a>
gautam_adani	ADANI FOUNDATION is humbled to contribute Rs. 100 Cr to the #PMcaresfund in this hour of India's battle against #COVID19. ADANI GROUP will further contribute additional resources to support the GOVERNMENTS and FELLOW CITIZENS in these testing times.
marcorubio	Some in our media can't contain their glee & delight in reporting that the U.S. has more #CoronaVirus cases than #China\n\nBeyond being grotesque, its bad journalism\n\nWe have NO IDEA how many cases China really has but without any doubt its significantly more than why they admit to
RealCandaceO	The number one killer in America is Heart disease. 1,002 people a day. \n\nDid you know that if you die from heart disease right now, and they determine you to be an asymptomatic carrier of Covid-19 in your post-Mortem, they legally add your death to the #Coronavirus death toll?
LindseyGrahamSC	If it were up to me the whole world would send China a bill for the #CoronavirusPandemic. \n\nThis is the third pandemic to come out of China.
DavidLat	As someone who recently spent almost a week hooked up to a ventilator (and probably wouldn't be alive today without one), I can attest to the importance of ventilators and the need for an adequate supply. #coronavirus #CoronavirusUSA #COVID19 #Covid_19 <a href="https://t.co/03J82jk9QK">https://t.co/03J82jk9QK</a>
PrakritiGaba	Last night in the ICU of a #NYC hospital, I cared for 20 patients who were all on breathing machines due to #COVID19. Some REALLY young (20s), without comorbidities. Everyone is extremely sick. But sicker patients keep flooding in...
ACNH_support	Because we are all dealing with the #coronavirus, I thought I'd spread some positivity!❤️\n\nI'm giving 2 #AnimalCrossingNewHorizons bundles\n\nRetweet\nFollow @ACNH_support \n\nGood Luck To Everyone!🍀\n\n#StayAtHomeAndStaySafe \n\n#animalcrossing <a href="https://t.co/Rjtzcw7Fhj">https://t.co/Rjtzcw7Fhj</a>
Jim_Jordan	My friend @replouiegohmert says it best: \n\n"We don't have to pay organizations to lie to us. They'll probably do it for free!" \n\nHe's right. The #WHO wasn't honest about #coronavirus. And I applaud President Trump's decision to halt funding.
NCDCgov	Thirty-four new cases of #COVID19 have been reported as follows:\n\n18 Lagos\n12 in Kano\n2 in Katsina\n1 in Delta\n1 in Niger\n\nAs at 11:20 pm 15th April there are 407 confirmed cases of #COVID19 reported in Nigeria. 128 have been discharged with 12 deaths\n\n#TakeResponsibility <a href="https://t.co/oxM9pVb9QQ">https://t.co/oxM9pVb9QQ</a>
Jim_Jordan	There are #coronavirus task forces doing great work. \n\nBut there's one task force that's missing in action: the U.S. Congress! \n\nWhere's Speaker Pelosi's plan to get back to work?
JoeBorelliNYC	A sad scene at Brooklyn Hospital #Covid_19 #coronavirus <a href="https://t.co/dRUNE61yPs">https://t.co/dRUNE61yPs</a>
NCDCgov	Fourteen new cases of #COVID19 have been reported in Nigeria; 9 in Lagos and 5 in FCT\n\nAs at 09:30 pm 29th March there are 111 confirmed cases of #COVID19 reported in Nigeria with 1 death. <a href="https://t.co/qowl0bEPag">https://t.co/qowl0bEPag</a>



We see that the **most retweeted** news is about the **positive results** of Hydroxychloroquine Sulfate, Zinc and Z-Pak on Covid19 patients. This was early on in the year and everyone was hoping for a quick solution to the virus. There was some **blame for China's** initial cover-up. **Some doubted** the seriousness of the virus while others reported deaths from the frontline. The 2nd most retweeted tweet was Angela Merkel's explaining the importance of **flattening the curve**. Then there were **concerns** about **the economic effects** of going into a shutdown.

Looking at the **top favorite tweets** we find firstly 14 of the 16 tweets are from one poster: David Levitt. His tweets are mostly political in nature. There isn't anything wrong with this but it blocks our view from seeing other things people are talking about. On the following page we will look at top favorite tweets not authored by David Levitt.

```
top_favorites = df_toptweets.sort_values(by = 'favourite_count', ascending = False)
top_favorites[:16]
```

ChelseaAMusic	When we get over this #COVID19 virus I think it's time movie shoots, filming TV shows, premiere events, Award events, TV & Music events/festivals do more to protect the actors, actresses & musicians more so this doesn't happen again ever!! 🙏🙏
ChelseaAMusic	.@PrimeMinisterGR: I know you are committed to keeping all kids in #Greece healthy & safe through #COVID19. Hundreds of migrant kids are locked up for no reason. You can save their childhoods. #FreeTheKids, put them in child-friendly housing now! <a href="https://t.co/8nugDsDfo5">https://t.co/8nugDsDfo5</a>
David_Leavitt	W.H.O. didn't blow it.\n\nYOU BLEW IT.\n\nYou gave faulty recommendations, telling the public that #COVID was a hoax, that #COVID19 was just a flu, that #coronavirus was contained, that the #CoronavirusPandemic would just go away, and now these #CoronavirusUSA deaths are your fault. <a href="https://t.co/tSezI6NxWi">https://t.co/tSezI6NxWi</a>
David_Leavitt	"You're not gonna die from this drug," @realDonaldTrump says of hydroxychloroquine\n\nYesterday the president of AMA said there's fatal side effects.\n\n#COVID #COVID19 #coronavirus #CoronavirusPandemic
David_Leavitt	The frozen meat brand Steak-umm is doing a better job than the president.\n\n#COVID #COVID19 #coronavirus <a href="https://t.co/fBoicFlpMa">https://t.co/fBoicFlpMa</a>
David_Leavitt	You can't shoot the #coronavirus with a gun.\n\nThe @NRA is a terrorist organization.
David_Leavitt	Fox News is being sued for peddling #coronavirus misinformation <a href="https://t.co/dQioMcC6xn">https://t.co/dQioMcC6xn</a>
David_Leavitt	Seeing your friends after the #coronavirus quarantine ends <a href="https://t.co/AQRU0s0ScV">https://t.co/AQRU0s0ScV</a>
David_Leavitt	ALL non-Rhode Island plated cards are being pulled over by cops upon entering Rhode Island now #coronavirus
David_Leavitt	People are complaining about another 30 days of lockdown when it's really gonna be another 90+ days\n\n#COVID19 #Covid_19 #coronavirus
David_Leavitt	The ratings-tweet alone should be grounds for @realDonaldTrump's removal with the 25th amendment.\n\n#COVID19 #Covid_19 #coronavirus
David_Leavitt	Literally within hours of each other:\n\n"We're going to have at least 100,000 to 200,000 deaths and millions of cases" says Dr. Fauci\n\n"Sure, but have you seen my tv ratings?" asks Trump\n\n#COVID19 #Covid_19 #coronavirus
David_Leavitt	@realDonaldTrump @nytimes Your loved ones might be dead or dying, but my ratings are through the roof!!!\n\n#COVID19 #Covid_19 #coronavirus
David_Leavitt	WTF are you rambling about @realDonaldTrump?\n\nIt's when the PEOPLE no longer have a voice then it's no longer a democracy. \n\nIt has nothing to do with you, you sexist racist lying cheating orange fat fuck\n\n#25thAmendment #COVID19 #Covid_19 #coronavirus
David_Leavitt	Why are gun stores are considered "ESSENTIAL" and allowed to be open while book stores are not?\n\nWe need less guns and more books.\n\n#COVID19 #Covid_19 #coronavirus
David_Leavitt	"We're going to have at least 100,000 to 200,000 deaths and millions of cases" says Dr. Fauci #COVID19 #Covid_19 #coronavirus <a href="https://t.co/hxhPyOGQKI">https://t.co/hxhPyOGQKI</a>

We **drop duplicates** from individual posters:

```
top_favorites = df_toptweets[['favourites_count', 'screen_name',
                             'text']].sort_values(by = 'favourites_count',
                                                  ascending = False).drop_duplicates(subset='screen_name')
top_favorites[:20]
```

favourites_count	screen_name	text
1995152	ChelseaAMusic	It's time @SmithfieldFoods took responsibility for what they did & how they didn't protect their workers when the first case of #COVID19 came out in their factory like so not right at all!!! 🤔🤔🤔 <a href="https://t.co/8YVWBA2HPX2">https://t.co/8YVWBA2HPX2</a>
1562269	David_Leavitt	The @WhiteHouse Gift Shop is selling "World vs. Coronavirus" coin for \$100...and the description says there's only 1,000 but somehow I feel like this is another @realDonaldTrump scam to cash in on #COVID19 \n\nSurprised it's not a "Trump vs. Coronavirus" coin tbh <a href="https://t.co/FTEKXaEKBb">https://t.co/FTEKXaEKBb</a>
1422809	MiguelCalabria3	People under lockdown are showing their gratitude to front-line healthcare workers worldwide by applauding them.\n\n#coronavirus #COVID19\n#QuedateEnCasa #StayAtHome #RestezChezVous #sanitarios @famartinez2001 <a href="https://t.co/FvYzuvv1V1">https://t.co/FvYzuvv1V1</a>
1311806	littlebytesnews	Terrible, may she RIP. Hopefully this doesn't become a trend and medical professionals get the mental healthcare they need to overcome so much suffering and death.\n#Coronavirus: Top NYC doctor takes her own life <a href="https://t.co/YLgP04CxxM">https://t.co/YLgP04CxxM</a> <a href="https://t.co/DnTxwJWksQ">https://t.co/DnTxwJWksQ</a>
1266919	SueRMichael	This is a wonderful story of surviving #COVID19 and of hope <a href="https://t.co/2fhFVPV8BV">https://t.co/2fhFVPV8BV</a>
1258128	hazelglasgow	"Coronavirus: WHO warns 'the worst is yet ahead of us' in outbreak" #Coronavirus <a href="https://t.co/mdtDhJlvq">https://t.co/mdtDhJlvq</a>
1251292	madanabhat	Listen to the most recent episode of my podcast: Digital democracy as a consequence of the #Coronavirus crisis <a href="https://t.co/116Y5G3UWK">https://t.co/116Y5G3UWK</a>
1186068	amor_vuelveTX	@MissClioMurray @come_for_t @setzcat @SamusAran2020 @LauraEastlick1 We're waiting for one tonight. We're far down south, not even #Coronavirus #COVID19 won't come around mol #BabyYvette lubz frenz <a href="https://t.co/xE0Djuh7un">https://t.co/xE0Djuh7un</a>
1131771	fahma311	@OntHospitalAssn Stay safe heroes #StayHomeStaySafe #FlattenTheCurve
1126530	paoloigna1	Continueremo a parlare per esigenze Geopolitiche ed elettorali di #Trump #covid19:BBC News - #Coronavirus: US intelligence debunks theory it was 'manmade'\n <a href="https://t.co/dDZVtm0SPg">https://t.co/dDZVtm0SPg</a>
1065477	ben10dinosaur	That concludes the One World: #TogetherAtHome benefit concert! Thanks to @jimmykimmel, @jimmyfallon and @StephenAtHome for hosting! And thanks to the artists for their great performances! And most of all, thanks to the healthcare workers in the front lines against #COVID19!
1059413	geoffrey_payne	gee! did you compliment the President on the US #COVID19 death toll? #Auspol #Covid19usa <a href="https://t.co/Ug7HpOPdiA">https://t.co/Ug7HpOPdiA</a>
1048085	Solutioneer72	@RonaldKlain ARBs* - esp valsartan/sacubitril\n#ARDS** caused by #COVID19 #SARSCoV2\n\n(Q: what abt losartan (reduces clotting)?\n\n#Angiotensin Receptor Blockers\n\n**Acute Respiratory Distress Syndrome \n\n\n <a href="https://t.co/4SZC2AbOHD">https://t.co/4SZC2AbOHD</a>
1012747	jordanshirumat2	HAPPENING NOW\n—\nH.E. @KagutaMuseveni is addressing the nation about general Covid19 updates and interstate strategy to managing #Covid19 infections among truck drivers.\n\nJoin Conversation 🗨️ #M7Address #StaySafeUG \nRemarks (READ THREAD)🗨️ \n\nWATCH LIVE 📺 <a href="https://t.co/XLoYX8imLF">https://t.co/XLoYX8imLF</a> <a href="https://t.co/6gqsj7AdXA">https://t.co/6gqsj7AdXA</a>
1004022	i_am_IBBU	Can someone help me understand, if the\nHydroxychloroquine can cure or can prevent #Covid19 then why this medicine has not been made widely available at every chemist shop with doses prescription? \n\nAny doctor who can enlighten me? 🤔🤔🤔 copied bhavika
981249	segmentis	New #Trump "accomplishment:" Fastest retreat from rallying around the flag in history.\n\nPoll: Majority of Americans now disapprove of @realDonaldTrump's #coronavirus response <a href="https://t.co/y6mxj9XIs4">https://t.co/y6mxj9XIs4</a>
977141	JRoc23	This is so sad.\n#COVID19 #CoronavirusPandemic <a href="https://t.co/qAgjGul2ht">https://t.co/qAgjGul2ht</a>
975529	growingaway	🤔🤔🤔🤔 #Repost gunprideworldwide with @get_repost\n · · · \nIt's not even November yet. \n#billgates #virus #wuhan #5g #africa #saudiarabia #india #iran #syria #middleeast #corona #coronavirus #epstein #vaccine #NWO... <a href="https://t.co/Zplclwdx5J">https://t.co/Zplclwdx5J</a>
975513	collectibulldog	#collectibulldogs has an original organic white hat links page ! This page was found by a specialist after reading @MoishesMom article on #Covid19 They love the style and wish to sponsor the page ! If you'd like to swap #backlinks free just get in touch
971907	ThePerezHilton	PLEASE smart scientists, come through!!! <a href="https://t.co/mnYadp2CH">https://t.co/mnYadp2CH</a> #Coronavirus

Here have more diverse conversations ranging from **gratitude** for front line workers, **survival** stories, **political criticism**, etc ... There were some requests for and discussions about medical information about Covid19 treatment, etc ... prove to be more informative and helpful.



## In Conclusion

The dataset we're exploring were from a month's worth of data scraped from twitter in the end of March to the end of April 2020 right at the beginning of lockdown protocols in the US. The initial cleaning and exploration of the dataset reveals it to be a sizable dataset and may present a challenge computationally as well as a potential challenge of time and resources. This has to be taken into consideration as we tackle the questions we are looking to answer: What trends / sentiments / behaviors / challenges arose during our period of self-seclusion?

We will have to look into the "text" column and look at the tweets themselves. The initial idea is to use NLTK, spaCy or Gensim to do **Natural Learning Process** on the tweets. We could explore clustering using KMeans Clustering to see if any clusters of topic are apparent and see if there are patterns that arise.