1.Overview

The COVID-19 has affected the economy and people's lives across the globe. Vaccines can effectively reduce the spread of COVID-19. People's willingness to get vaccinated would be affected by many different variables. We collected data from four different data sets, some of their data would be used to measure the severity of the epidemic, and others would be used as the variables that influence people's willingness. We measured the cleanliness level on those data by identifying most of the missing values and some noise.

2.Background

COVID-19 is a global virus, which has seriously affected our lives. This virus can affect our upper and lower respiratory tract by infecting our cells. Thus, it does great harm to our health. And it is not only very contagious(evidence suggests that the virus can spread from an infected person's mouth or nose in small liquid particles when they cough, sneeze, speak or breathe), but also very tricky, some of the infected persons are asymptomatic. According to these features, it seems unexpected, but it makes sense that it rapidly spreads around the world. Many people directly or indirectly suffered from this disaster. Some people died from this disaster, some people are unemployed and on the streets. Besides, COVID-19 has a great negative impact on the global economy and almost any field, such as industrial production, service industry, entertainment...

3.Data Science Problem

Since the start of COVID-19 vaccination, it has demonstrated a good immune effect and has shown that vaccination is the most reliable way to avoid a large-scale outbreak of the epidemic. So if the public is willing to be immunized is an important influencing factor of controlling the pandemic[1]. There are so many factors that influence people's willingness to vaccinate: whether the vaccines will clear safety trials, the Role of politics, the dynamics of antivaccine groups on social networks, and so on[2] .People's enthusiasm for vaccinations fluctuates at different stages of the development of the epidemic. After discussion, our group put forward many novel ideas.

First of all, we believe that the emotional orientation from people's tweets about vaccines has a significant impact on the willingness of others. For example, when people generally suspect that vaccination will have serious side effects, their willingness will decrease. When the results of experiments on vaccine effects are released, people will be more convinced of the effectiveness of the vaccine.

Secondly, we raised an original point of view, that is, the remaining local medical resources will influence people's decisions. During the severe period of the epidemic, we observed a shortage of hospital beds, especially ICU beds. Although COVID-19 is not a disease with a high fatality-rate, the lack of adequate medical resources leading to the inability to receive treatment in time means that getting the illness will cause more serious consequences. At that time, people will be more inclined to take prevention through various means, for example, getting vaccinated.

The topic of this project is the influence of various factors on the population's willingness to vaccinate in the United States. We think that lots of subjective and objective factors can cause changes in the vaccination rate. For example, the significantly increasing number of positives, local policies for vaccines, public sentiment about the COVID-19 on social media, and whether sufficient medical equipment is available. This project will focus on finding out what factors will dominate the changes and predicting the trend of people being vaccinated under current conditions. We are convinced that our views will bring new and different inspirations from the research results, and we cannot wait to complete the project and bring help to the school and the community.

4.Collecting Data and Analyzes

4.1The Twitter-COVID-dataset

The Twitter-COVID-dataset, collected from 28 January 2020 to now (11 September 2021), records the emotional intensity of Twitter users from their responses to the COVID-19 pandemic in the United States[3]. Tweets are collected by using four keywords ("corona", “wuhan", "nCoV" and "covid"). The variables can be divided into four parts:

(1) three information attributes of the post: tweet ID, user ID and the posted time.
(2) one keyword attribute: the keywords used when collecting data
(3) five emotional intensity attributes: the degree of intensity of the valence, fear, anger, happiness and sadness emotions.
(4) two qualitative attributes: the sentiment category and dominant emotion category

Since the values in keyword attributes are similar, the emotional intensity mentioned above will be more useful to this project. Emotions can directly reflect people's attitudes towards COVID-19 and vaccines. For example, If the fear and sadness intensities are very high during a certain period of time, then we guess that the vaccination rate might also increase. Besides, we can put these attributes together with factors such as death, increasing positive rate or hospital capacity in other data sets based on time, and then, some relationships between people's reactions and the objective factors at different time periods can be built.

4.2 The COVID Tracking Project

The data collected from The COVID Tracking Project contains the Covid 19 data of 50 US states, the District of Columbia and 5 other US territories, and their summary.The main variables it contains are Deaths, Pending, Tests, Recovered, Hospitalized, ICU, States, Country, and so on.Deaths, Pending, Tests, Recovered can be used to analyze the severity of the pandemic.

Hospitalized, ICU can be used as a variable to analyze people's willingness to vaccinate.For example, would the number of ICU patients increase encourage people to want to get vaccinated?

Therefore, we intend to analyze the pandemic situation in each state and the changes in various variables, as well as the impact of different variables on the pandemic situation, and determine which variables have greater influence.

4.3 The Covid Act Now API

Through the Covid Act Now API, we can obtain a comprehensive COVID19 data set. What excites us most is that it contains data related to medical resources, which will help us verify the idea that medical resources will affect people's willingness to vaccinate. As mentioned above, we believe that when medical resources are scarce, people's willingness will increase. In addition to common information, this data set also contains the attribute of risk level, which will better help us judge the trend of the epidemic.

4.4 MongoDB COVID-19 Open Data Set

We use Rest API to retrieve data from MongoDB COVID-19 Open Data cluster. It provides many basic information about COVID-19, like daily death, daily infected person and so on. Compared with other datasets, it is divided by county, not state, which enables us to analyze in a smaller scope. The main variables in this dataset are state, county, location, deaths, daily_deaths, confirmed, confirmed_daily, date, population. These variables enable us to know the situation in a special period on a special date. In this case, we can relate this information to vaccination willingness. For example, people will receive vaccination due to the increased number of daily deaths..

5.Data Issues and Cleanliness

5.1 The Twitter-COVID-dataset

In Twitter-COVID-dataset, the range of emotional intensity is between [0,1]. Values close to 0 mean that the tweet does not express too much of this sentiment, and values close to 1 means that it contains an extremely intense sentiment. During the cleanliness, fortunately, no data is empty. However, 2763 tweets contain at least one kind of score out of this rage. It is found that these data are in the range of [-1,0] or [1,2]. We guess that it is because the sentiment intensity score is obtained from CrystalFeel[4], so a small part of the data will exceed the given range. A negative score or score greater than 1 might indicate no relationship or unusually extreme emotions, so we currently treat those data as noise, unless this part of extreme data is needed in future study.

Table 1: Quality evaluation for Twitter-COVID-dataset

Values	**Values Number**	**Missing Value**	**Noise**	**Missing Quality**	**Noise Quality**
information	54726351	0	0	0	0
keyword	54726351	0	0	0	0
emotional intensity	54726351	0	2673	0	4.88E-05
sentiment and dominant emotion category	54726351	0	0	0	0

5.2 The COVID Tracking Project

Most of the data of the COVID Tracking Project API are numbers. By observation we find there are many null values(The official docs explains that returns null means there is no data available).Some of the number data should not be negative. So we compute the fraction of null values of some important varabies. And count the number of negative values.

Now we haven’t found noise in this data set.

Table 2: the COVID Tracking Project

dataset	**Values**	**Values Number**	**Missing Value**	**Noise**	**Missing Value Quality**	**Noise Quality**
states_daily	positive	20780	188	0	0.009	0
states_daily	negative	20780	7490	0	0.360	0
states_daily	hospitalizedCurrently	20780	3441	0	0.165	0
states_daily	inIcuCurrently	20780	9144	0	0.440	0

states_daily	recovered	20780	8777	0	0.422	0
states_daily	death	20780	850	0	0.040	0
US_daily	positive	420	1	0	0.002	0
US_daily	negative	420	48	0	0.114	0
US_daily	hospitalizedCurrently	420	64	0	0.152	0
US_daily	inIcuCurrently	420	73	0	0.173	0
US_daily	recovered	420	420	0	1	0
US_daily	death	420	28	0	0.067	0

5.3 The Covid Act Now API

Since the data sources of the Covid Act Now API are relatively official, the proportion of noise in the data set is very low. On the contrary, when analyzing the response, we found that the phenomenon of missing values is more common. For data loss, we think there are two important reasons. First, variables such as vaccinations_initiated_ratio, vaccinations_completed_ratio, vaccines_distributed，vaccinations_initiated，vaccinations_competed，vaccines_administeredare statistical data collected sometime after the outbreak of COVID19, therefore, there is no relevant information in the early data, which affects The missing rate. Second, part of the data returned by the API actually comes from other data sources. These data sources may have missing data or do not support searching for historical data. In summary, although the calculated fraction of missing values is relatively low, the data is relatively complete and will not affect subsequent work.

5.4 MongoDB COVID-19 Open Data Set

The MongoDB COVID-19 open dataset provides fairly clean data. We write a program to collect the missing values and there is no missing value. And according to the data attributes such as deaths, daily_deaths, confirmed, and daily confirmed, firstly, all of these attributes must be positive and the number of deaths should be bigger than the number of daily deaths (same in daily confirmed and confirmed). These are some noises in the dataset.

Table 3: Quality evaluation for MongoDB-COVID-dataset

Values	Values Number	Missing Value	Noise	Missing Quality	Noise Quality
confirmed	824966	0	0	0	0
confirmed daily	824966	0	13408	0	0.0163
deaths	824966	0	0	0	0
deaths_daily	824966	0	4129	0	0.0050

6.Citation

[1] *Mallapaty, S., & Ledford, H. (2020). COVID-vaccine results are on the way - and scientists' concerns are growing. Nature, 586(7827), 16–17. https://doi.org/10.1038/d41586-020-02706-6*

[2] *Cornwall, warren. (2020, June 3). Just 50% of Americans Plan to Get a COVID-19 Vaccine. Here’s How to Win over the Rest. Science. https://www.sciencemag.org/news/2020/06/just-50-americans-plan-get-covid-19-vaccine-here-s-how-win-over-rest*

[3] *Gupta, R. K., Vishwanath, A., & Yang, Y. (2021, September 14). COVID-19 Twitter Dataset with Latent Topics, Sentiments and Emotions Attributes. OPENICPSR. https://doi.org/10.3886/E120321V10-97100*

[4] *Institute of high performance computing, a*star. (2021). CrystalFeel - Multidimensional Emotion Intensity Analysis from Natural Language. https://socialanalyticsplus.net/crystalfeel*