COMP7507 Project Report: COVID-19 Vaccination

Group 19

Name	Student ID
Shen, Jian	3035803774
Wang, Xinyu	3035802366
Zhao, Wantong	3035802536
Yang, Ze	3035803451

Table of Content

Foreword

The World Health Organization has launched the "Strategy to Achieve Global COVID-19 Vaccination by mid-2022" on October 7th, 2021. The new strategy outlines the plan that: by mid-2022 year, 70% of the region's population could be vaccinated. According to the data shown online on November 18: "52.4% of the world population has received at least one dose of a COVID-19 vaccine.". There are still some people who do not get vaccinated. To meet the vaccination target as soon as possible, our project suggested the number of intuitive visualizations for the people who are not vaccinated, we hope to enhance their motivation and willingness to get vaccinated. In addition, we also provided visualized data support for various countries or regions to formulate their own vaccination's policy and strategy to prevent the spread of the COVID-19.

Data

We used various data from multiple data sources, the brief description of the data that we used are showed as the following:

1. COVID-19 Confirmed, Death cases and Vaccinations Dataset (Numerical)
   Our dataset covers the counties/regions all over the world from January 22nd, 2020, to November 14th, 2021, for a total of 224 counties and regions. This data source is coming from the website "Our World in Data". Some variables in this dataset are list below:
   - Counties/regions
   - Total_cases: Total confirmed cases of COVID-19
   - New_cases: New confirmed cases of COVID-19
   - Population
   - People_vaccinated: Total number of people who received all doses prescribed by the vaccination protocol
   - New_vaccinations: New COVID-19 vaccination doses administered (only calculated for consecutive days)

2. Vaccination Policy Data (Numerical, Categorical)
   This dataset contains the policies on the availability of vaccinations for each county/region starting from January 1st,2020 to November 14th,2021.
   - Counties/regions are group into 6 categories:
     0: vaccination policy with no availability
     1: Available for ONE of the following: key workers/ clinically vulnerable groups / elderly groups
     2: Available for TWO of the following: key workers/ clinically vulnerable groups / elderly groups
     3: Available for ALL of the following: key workers/ clinically vulnerable groups / elderly groups
     4: Available for all three plus partial additional availability (select broad groups/ages)
     5: Universal availability

3. Social Media- Twitter Tweets Data (Text)

The Original dataset contains the daily tweets started on August 1st, 2020, which have #CovidVaccine hashtag from Twitter. The data totally consists of 1 lakh+ records. We find this data from the Kaggle website. Some of the variables are list below:

- user_name
- date
- text: actual text of the Tweet
- hashtags

4. Vaccine Adverse Reaction Data (Text)

This dataset was created by the Food and Drug Administration and Centers for Disease Control and Preventions to receive reports about adverse events which are associated with vaccines. Some variables in this dataset are list below:

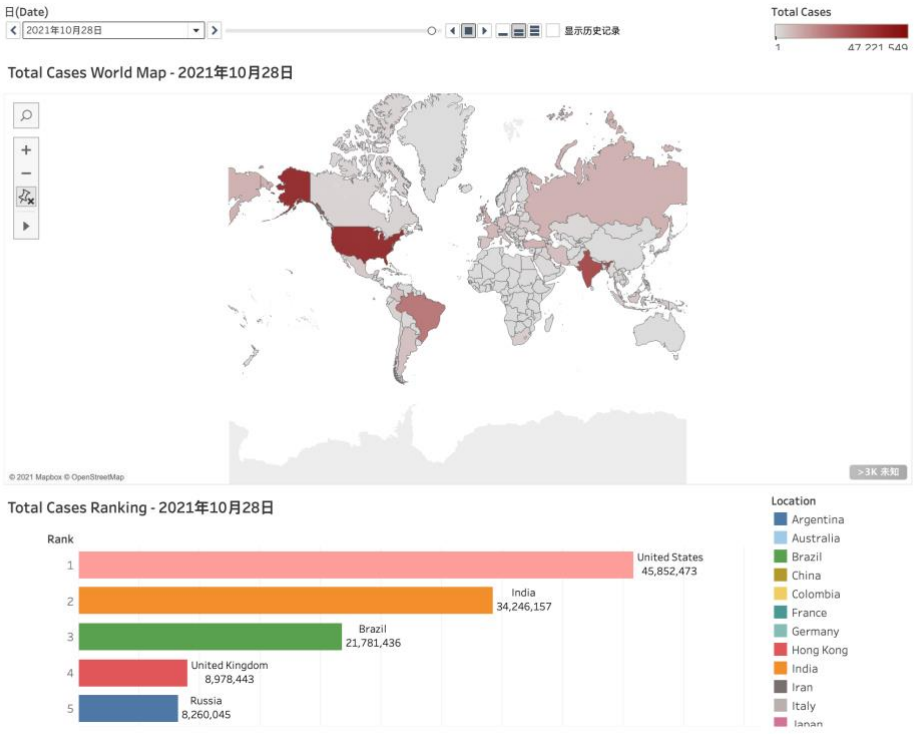
- SYMPOTM_TEXT: Symptom text record
- Patient outcome:
  i. DIED: If the vaccine recipient died a "Y" is used; otherwise, the field will be blank.
  ii. HOSPITAL: Whether the vaccine recipient was hospitalized.
  iii. DISABLE: Whether the vaccine recipient was disabled.
  iv. Doctor or other healthcare professional office/clinic visit.
  v. X_STAY: Whether a patient's hospitalization is prolonged as a result of the adverse event associated with the vaccination.

Dashboard 1 & Dashboard 2: Background

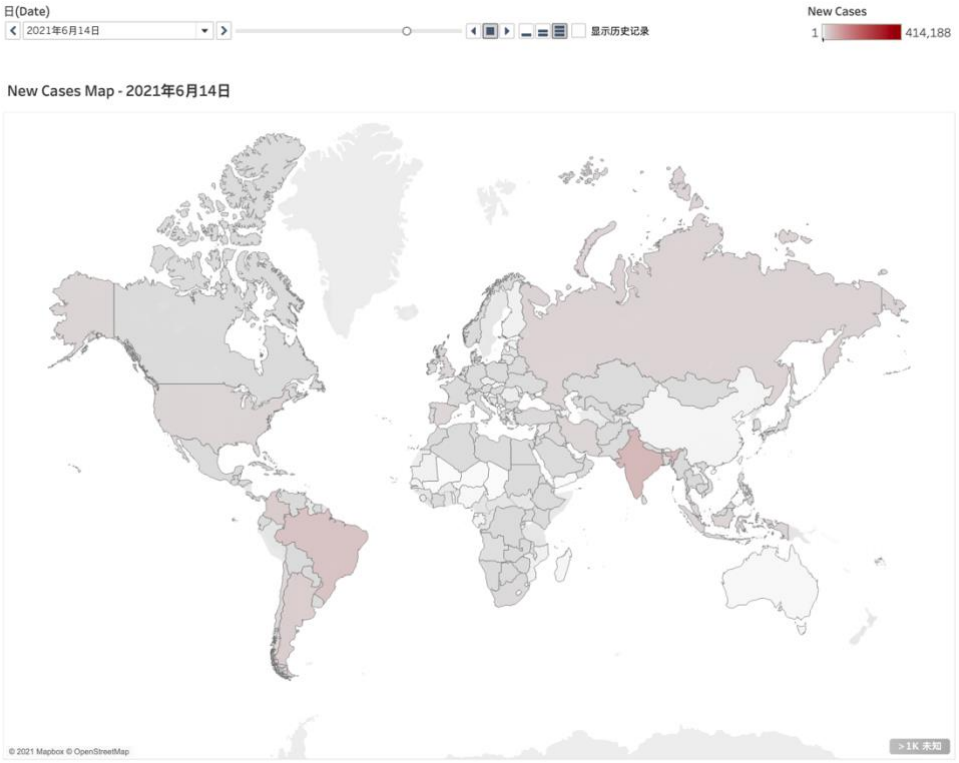
Design

Dashboard 1 and Dashboard 2 are the background of the entire storyline. The aim is to give viewers a sense of the context of the COVID-19, so we apply geospatial data to achieve this goal. In dashboard 1, we use the number of total confirmed cases to mark the color of each country/region on the map. In addition, we break a view into a series of pages to better analyze how the number of total confirmed cases in each country/region change over time. Considering the size of the map, it is not reasonable to show the number of total confirmed cases on the map because there are 224 countries/regions. Suppose we choose to indicate the number and the name of the counties/areas on the map simultaneously. In that case, the information could be overwritten, and the viewers will not capture any information. We added one racing bar chart. Here, we chose the top 5 countries and created an interactive relationship between these two sheets to move together within time. Since the first dashboard is a cumulative number, we hope that the people can capture through our dashboard that the epidemic situation in some countries is changing. In the second dashboard, we chose the number of new confirmed cases as the mark of color and added the date into the “pages” function in the tableau. Viewers can compare the COVID-19 situation within and between the countries/regions intuitively. Some cons exist when we use tableau to show the information. The darkness of the color in the tableau was created by the number of the new confirmed cases; since the range of this number was huge, which is between 0 to 414188, in the tableau, we could not set up the range for each steeped color. Tableau divided the stepped color equally so that the differences between colors may not be as intuitive as the numbers.

Final Visualization



(Dashboard 1 Overview)



(Dashboard 2 Overview)

Dashboard 3: Vaccinations

Introduction

This dashboard aims to provide a view of people vaccinated all over the world. We focus on the vaccination policy and the relation between new cases and people vaccinated in each country. We would like to provide evidence that the vaccination is an effective way to reduce progress of the pandemic. We divide this dashboard into two parts: Political analysis and Vaccination rate analysis.

For the political analysis, in order to achieve the goal of vaccination as soon as possible, it is necessary to have a view of the changes of vaccine policies in different regions over time. The more active the government's vaccine policy is, the more people are willing to vaccinate, and the easier it is to promote the global goal of 70% vaccination rate in 2022. On the contrary, the more negative the government's vaccine policy is, the fewer people are willing to be vaccinated and the less helpful it will be to achieve the goal of global vaccination rate.

The main purpose of our analysis on vaccine policies in different regions is to identify areas with negative vaccine policies, encourage their governments in these areas to improve vaccine policies, so that vaccines can be popularized to more people in policies, and achieve the global vaccination goal of 2022 as soon as possible.

For the Vaccination rate analysis, in particular, this visualization is designed to answer the following question:

Whether vaccination is one of the factors that effectively contain the epidemic?

Data Processing

- **Political Analysis**
  We directly use the data downloaded from the Internet as our data set without any modification. The time interval is from January 1st, 2020 to November 14th, 2021. The dataset has four columns. The first column is "Entity", which represents the name of each region, the second column is "Code", which is the abbreviation of first column, the third column is "Day", and the fourth column is "vaccination_policy", which is a number in 0-5 representing the degree of the population targeted by the vaccine.

- **Vaccination Rate Analysis**
  First, we drop all the values of which location is a continent or “world”. Then, since the population of each country varies greatly, we use the proportion of people vaccinated or the proportion of new cases instead of the people vaccinated and new cases.

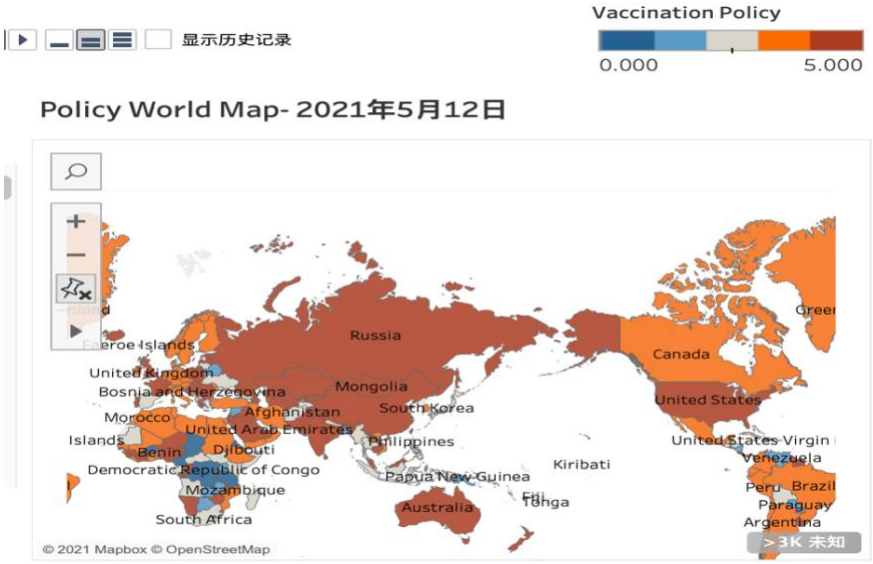
Design

- **Political Analysis**
  The following figure maps government policies on COVID-19 vaccination. Note that this only tracks policies on the *availability* of vaccinations. It does not track the number of people who have been

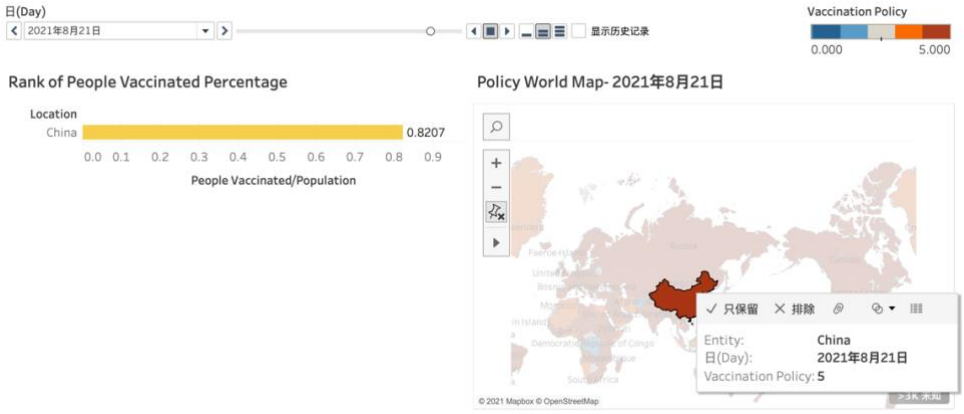
vaccinated. We can click the auto play button to observe the policy changes of different regions in the whole period of time. We can also choose a specific day to see the vaccine policies of different regions on that day. 0-5 take different colors from dark blue to dark red, dark blue represents vaccine within non availability, and dark red represents vaccine is universal availability. If we are interested in a specific area in the map, we can click the area, and the static map on the left will show the vaccination rate in the area up to November 15, 2021. Therefore, the map can also be associated with the static map on the left.

The reason why we choose to use the map is that we have the data of a great number of regions. Using the map allows us to have all the horizons, and we can observe and analyze them according to their geographical location. If the rank chart is used, the page may only display a small number of regions, and it is impossible to compare and analyze the relationship between policies in a region and adjacent regions according to geographical location.

Because we only have six different numbers to represent different vaccine policies, we use six separate different color changes instead of continuous color changes to represent different numbers, so we can more intuitively see the specific policy situation.



（Political Analysis）



（Interaction of Political Analysis）

- **Vaccination Rate Analysis**

  There are 5 dimensions of data involved in this visualization:
  - Continent (Category)
  - Location (Category)
  - New Cases / Population (Numerical)
  - People Vaccinated / Population (Numerical)
  - Time (Numerical)

  We came up with the design based on the data types and how they are usually visualized.
  - Temporal: since the new cases and people vaccinated are time series data, we would like to use the animation to show the changing of the data.
  - Multivariate: since the numerical data(time) can be shown by animation, the new cases and people vaccinated can be shown using colors and size.

  Therefore, we explore 2 sheets in this part, People Vaccinated VS New Cases (World Wild Tendency), and People Vaccinated VS New Cases (Countries and Continent Tendency).

  - People Vaccinated vs New Cases (World Wild Tendency)

    To build this visualization, we choose to use a scatter plot to show the relationship between the vaccine and the new confirmed case based on the dot's position and size. We established a dynamic daily scatter plot to give viewers a clear sense of the changes vaccines have brought to the epidemic. The viewer can press the bottom to see the changing of the dots in the scatter plot. We put the proportion of people vaccinated (People Vaccinated / Population) as x-axis and proportion of new cases (New Cases / Population) as y-axis. To be noticed: the dataset contains 224 counties/regions. In other words, there are 224 dots in one scatter plot, and it may be challenging to capture some of the changes by observing these dots, so we add one trend line to demonstrate the relationship between the vaccine and new confirmed cases. The trend line was added using the method of least square, which is a statistical procedure to find the best fit for these data points by minimizing the sum of the courts.

    The pros and cons of the scatter plot are listed as the following. By considering these, we ended up using a scatter plot to answer this question.

    Pros:
    - Indicates data correlation; this is one of the main reasons that we choose the scatter plot.
    - Can show the spread of data and can capture some outliers.
    - Can illustrate non-linear patterns.

    Cons:
    - Could not label the data points. Since there are too many dots in one plot, some of them are crowded together, we could not label out the name of each country/region, and it is hard to view.
    - Could not show the relationship between more than two variables at once.

- People Vaccinated vs New Cases (Countries and Continent Tendency)
  We use the scatter plots with animation in this part to show the relationship between vaccination and new cases. The size and color of each dot represents the number of new cases and the continent of the country. The x-axis represents the proportion of people vaccinated (People Vaccinated / Population) and the y-axis represents the proportion of new cases (New Cases / Population). Moreover, we split the plot into four quadrants by the average of People Vaccinated / Population and average of New Cases / Population to classify the situation of each country. The first quadrant stands for the country have a high new case rate and a high vaccinated rate, the second quadrant stands for the high new case rate and a low vaccinated rate, the third quadrant stands for a low new case rate and low vaccinated rate, the fourth quadrant represent the low new case rate and high vaccinated rate. We use scatter plots with animation since it can show the changing of each country or continent.

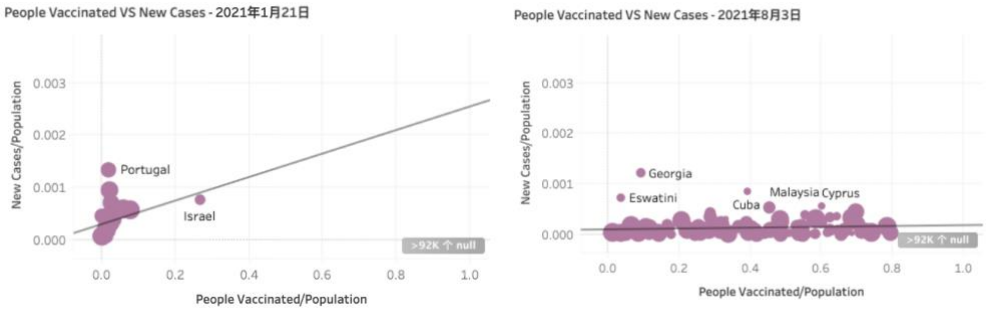
  Pros and cons of animation in trend:

  Pros:
  - Easy to show trends in multi-dimensional data.
  - The fastest technique for presentation.
  - Participants find it enjoyable and exciting.

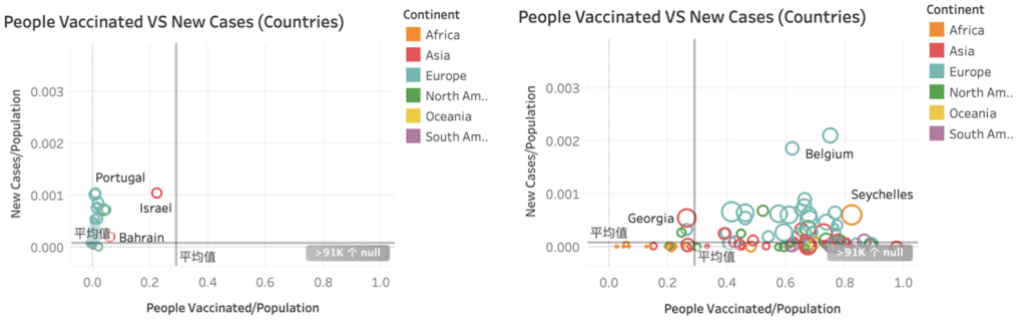
  Cons:
  - Lead to participants errors.

Final Visualization

(Dashboard 3 Overview)



(People Vaccinated vs New Cases (World Wild Tendency) on Different Date)



(People Vaccinated vs New Cases(Countries and Continent Tendency) on Different Date)

Analysis and Insights

- **Political Analysis**

  It can be observed that before January 2021, the vaccine policies in various regions have not been opened or are only aimed at a class of people. After January 2021, vaccine policies in various regions began to gradually target more people. By October 2021, it can be observed that the vaccine policy in most areas has basically faced all citizens. But there are still some blue, gray and orange areas where the vaccine policy is not universal enough. Therefore, we encourage and hope that the governments in these regions can pay more attention and promulgate vaccine policies for a wider population, so as to promote the completion of the global goal of 70% vaccination rate in 2022.

- **Vaccination Analysis**

  - People Vaccinated vs New Cases (World Wild Tendency)

    At the beginning of the period, the dots in the scatter plot are all concentrated on the y-axis. The slope of the trend line in the early period is high, which is reasonable since the vaccine has not yet been developed. No country/region's people were getting vaccinated. As time goes on, people worldwide get vaccinated, the dots are moving in the y-axis direction, and the slope of the trend line is getting lower, which means that vaccination can slow down the epidemic.

  - People Vaccinated vs New Cases (Countries and Continent Tendency)

    In the beginning, all the dots are in the second or the third quadrant since there is no vaccine. After a few months, the dots start to move rightward, especially for the Asian and European countries. Therefore, we can say that vaccination does help to control the progress of the pandemic in Europe and Asia. However, in the end, most European countries move to the first quadrant due to Delta bringing the outbreak of pandemic in Europe.

Dashboard 4: Symptoms and Attitudes

Introduction

The above chapters we introduced the relationship of vaccination and the total confirmed new cases, shows that the vaccination is very important for the control of the COVID-19 new cases. In this visualization, we would like to introduce the possible symptoms and reactions after the vaccination, also the people's attitudes towards vaccinations on social networks like Twitter.

In particular, this visualization is designed to answer the following question:

Is vaccination safe enough? Would some symptoms occur after the vaccination? For a particular person, does he need to pay more attention to the vaccination? Do people support vaccination or not?

In answering this particular question, we need to know what are the symptoms performed after the vaccination, which group or which cluster of people are more sensitive to the symptoms after the vaccination, and what is the main attitude of people towards vaccination.

Data Processing

This visualization uses two types of dataset. The first one is the "COVID-19 World Vaccine Adverse Reactions" from Kaggle, which comes from The Vaccine Adverse Event Reporting System (VAERS), which contains the reports about adverse events that may be associated with vaccines. The second one is the "Covid Vaccine Tweets" from Kaggle, which comes from some social media like Twitter to get people comments on the vaccination, the dataset contains the twitter tweets that with the hashtag #CovidVaccine.

For the first dataset, we need to extract the symptoms after vaccination and reactions after symptoms. We use an unsupervised machine learning algorithm (clustering) to segment adverse reactions into groups, and give back the most common symptom of each group. Then make use of the Making inferences, which is the most interesting part in the usage of Bayesian networks. We can easily compute the probability of an event given prior knowledge.

For the second dataset, we need to extract the attitude towards the vaccination from the tweets. We use the sentiment analysis in R package (a predefined natural language processing method) to give the score of each tweet, which would contain the score of eight emotions, and conclude to a either positive or negative side.

Design

To build this visualization, we choose a work cloud to show the different symptoms based on the word's size and colour, and a stacked bar chart to show the percentage and the comparison of the later development of the symptoms if you have, and a bubble chart to show the people's attitude towards the vaccination from the social media like Twitter.

The following lists the pros and cons of the stacked bar chart and the bubble chart, the comparison ability within and between bars in the stacked bar chart, the visualization relationship between variables of the bubble chart are what we need, so that we pick them.

Stacked bar chart:

Pros

- Multiple categories and data series in a compact space.
- Able to show change over time of category sub-components.
- Not only compare within bar but also compare between bars.

Cons

- Difficult to compare all but first series.
- Become visually complex as categories or series are added.

Bubble chart:

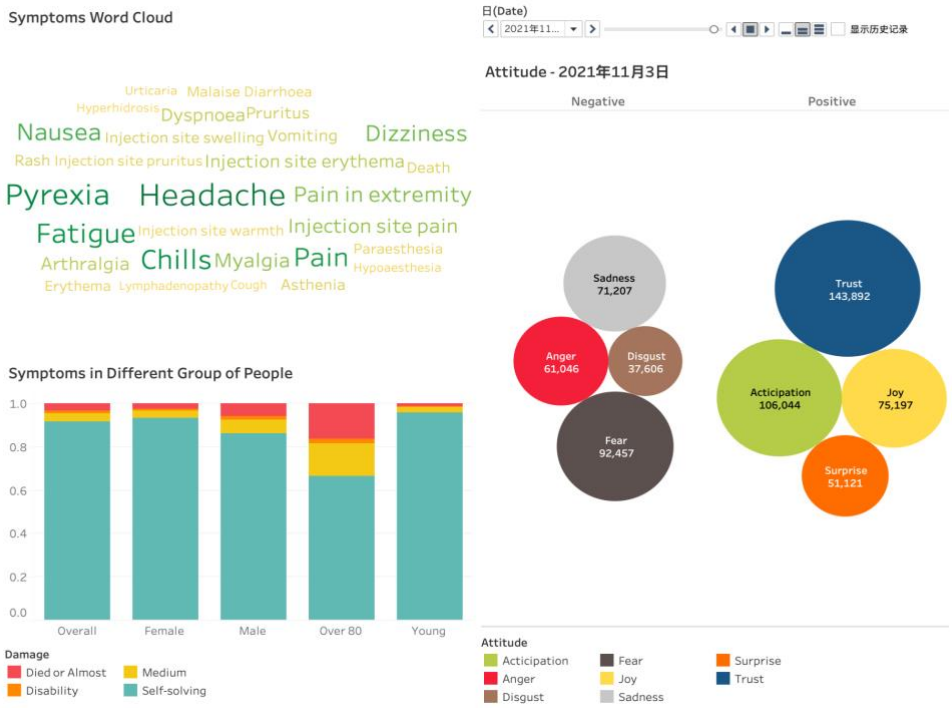
Pros

- Can show the relationship of one variable to another.
- Unique ability to show data about a third dimension.
- Can visually display correlation.

Cons

- Cannot be used to display a lot of data.

Final Visualization



(Dashboard 4 Overview)

Analysis and Insights

In the word cloud, we can see that most people report that the symptoms after vaccination are not very serious, such as headache, fatigue, fever, and rarely have some serious feedback symptoms.

The next stacked bar chart shows the follow-up impact of symptoms reported by people for the symptoms. It is also extracted through data and divided into four categories: died or almost, disability, medium, and self-solving. It can be seen that most of the symptoms are self-solving and only a few will result in serious conditions. At the same time, four types of data are classified according to gender and age. But we need to pay more attention to the situation of people over 80 years old. Compared with young people, we need to be cautious in dealing with symptoms after vaccination.

The final bubble chart shows the classification of people's attitudes towards vaccination on Twitter. The classification is divided into eight attitudes and emotions, of which four are trust, anticipation, and joy, the surprise is positive, and the other four kinds of sadness, anger, fear and disgust are negative. It can be seen that most public opinions have a positive attitude towards vaccination.

Contribution

	Wang, Xinyu	Zhao, Wantong	Yang, Ze	Shen, Jian
Proposal	✓	✓	✓	✓
Visualization (Tableau)	✓	✓	✓	✓
Data Collecting	✓	✓	✓	✓
Report	✓	✓	✓	✓