Corona virus outbreak

Guangming Chen 21664707

Table of content

1 Background 2

2 Introduction to experimental data 3

3 Experiment 4

4 Discussion and summary 10

5 Future 11

6 Reference 11

1. Background

As 2020 is approaching, a new epidemic caused by a new coronary pneumonia suddenly erupts. People around the world are affected by this sudden epidemic. Many companies have to stop business, and countries around the world have issued prevention and control measures against the epidemic. Relevant measures, the world seems to have fallen into a tense atmosphere. In this epidemic-resistant battle against smokeless smoke, a moving scene emerged: some medical staff concealed their family members and took the initiative to invite them to the battlefield; some were sealed under protective clothing for 8 hours continuously, letting any sweat drip They did not go down the line of fire; some of them stayed for too many hours and treated patients day and night ... but under heavy protective clothing, they were also ordinary people like us, and they also had their own warm homes, but they still chose to go to the front line To move forward with a heavy load in order to protect the safety of the broad masses of people. It was in this epidemic that I deeply realized my insignificance. Looking at these medical staff on TV, all I could do was cheer up and encourage, without even a little substantial support. This course mentioned the topic of data analysis on new coronary pneumonia. I chose this topic without hesitation. I also hope that through my professional knowledge, I can analyze even a little bit of information related to new coronary pneumonia, avoid Many people are tortured by the virus.

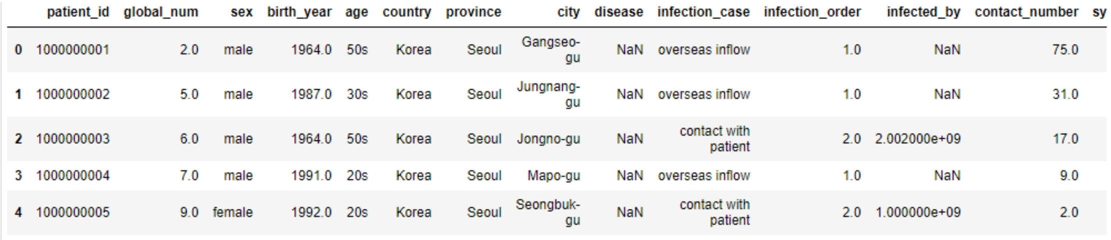
When analyzing the infection of the new coronavirus, there are many factors that may cause people to be infected with the virus. Moreover, people's gender and age are different, and the risk of infection is different. We count the infection status of different genders and ages in different periods, and draw relevant images to visually show the proportion of infected people under various factors Analyze the potential susceptibility factors.

2. Introduction to experimental data

In this experiment, the dataset we used came from Kaggle [1]. This dataset contains a total of 11 CSV files. Since this experiment only used PatientInfo.csv and PatientRoute.csv, we will only introduce the related information of these two tables here.

PatientInfo.csv records the basic information of patients infected with new coronavirus. Each row represents the basic information of a confirmed patient. Table 1 shows the data in the first 5 rows of the table:

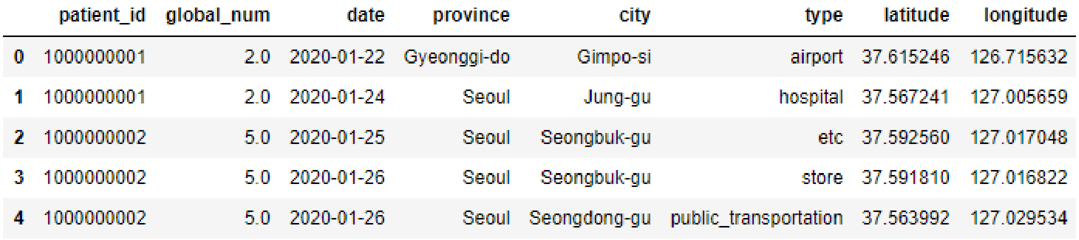
Table 1 PatientInfo.csv information display



	patient_id	global_num	sex	birth_year	age	country	province	city	disease	infection_case	infection_order	infected_by	contact_number	sy
0	1000000001	2.0	male	1964.0	50s	Korea	Seoul	Gangseo-gu	NaN	overseas inflow	1.0	NaN	75.0	
1	1000000002	5.0	male	1987.0	30s	Korea	Seoul	Jungnang-gu	NaN	overseas inflow	1.0	NaN	31.0	
2	1000000003	6.0	male	1964.0	50s	Korea	Seoul	Jongno-gu	NaN	contact with patient	2.0	2.002000e+09	17.0	
3	1000000004	7.0	male	1991.0	20s	Korea	Seoul	Mapo-gu	NaN	overseas inflow	1.0	NaN	9.0	
4	1000000005	9.0	female	1992.0	20s	Korea	Seoul	Seongbuk-gu	NaN	contact with patient	2.0	1.000000e+09	2.0	

PatientRoute.csv is the course of action within a certain period of time before and after the diagnosis of the new coronavirus infection. Each row indicates that a certain confirmed patient arrives at a certain place and related time and other information. These two tables can be linked through the "patient_id" column to establish a patient action roadmap information. Table 2 shows the first 5 rows of data of PatientRoute.csv:

Table 2 PatientRoute.csv information display

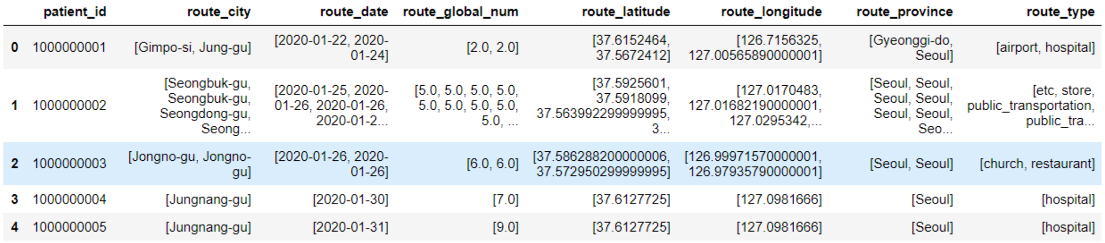


	patient_id	global_num	date	province	city	type	latitude	longitude
0	1000000001	2.0	2020-01-22	Gyeonggi-do	Gimpo-si	airport	37.615246	126.715632
1	1000000001	2.0	2020-01-24	Seoul	Jung-gu	hospital	37.567241	127.005659
2	1000000002	5.0	2020-01-25	Seoul	Seongbuk-gu	etc	37.592560	127.017048
3	1000000002	5.0	2020-01-26	Seoul	Seongbuk-gu	store	37.591810	127.016822
4	1000000002	5.0	2020-01-26	Seoul	Seongdong-gu	public_transportation	37.563992	127.029534

3. Experiment

In this experiment, we first merge the two tables, PatientInfo.csv and PatientRoute.csv, to merge the data with the same patient_id, and combine their other attributes (city date global_num latitude longitude province type) The elements are combined in order to form the corresponding list, and the compressed new DataFrame is obtained. Except for the patient_id in the columns, the other correspondences become: (route_city route_date route_global_num route_latitude route_longitude route_province route_type) Five-line data:

Table 3 Merged table



	patient_id	route_city	route_date	route_global_num	route_latitude	route_longitude	route_province	route_type
0	1000000001	[Gimpo-si, Jung-gu]	[2020-01-22, 2020-01-24]	[2.0, 2.0]	[37.6152464, 37.5672412]	[126.7156325, 127.00565890000001]	[Gyeonggi-do, Seoul]	[airport, hospital]
1	1000000002	[Seongbuk-gu, Seongbuk-gu, Seongdong-gu, Seong...	[2020-01-25, 2020-01-26, 2020-01-26, 2020-01-2...	[5.0, 5.0, 5.0, 5.0, 5.0, 5.0, 5.0, 5.0, 5.0, ...	[37.5925601, 37.5918099, 37.563992299999995, 3...	[127.0170483, 127.01682190000001, 127.0295342,...	[Seoul, Seoul, Seoul, Seoul, Seoul, Seoul, Seo...	[etc, store, public_transportation, public_tra...
2	1000000003	[Jongno-gu, Jongno-gu]	[2020-01-26, 2020-01-26]	[6.0, 6.0]	[37.586288200000006, 37.572950299999995]	[126.99971570000001, 126.97935790000001]	[Seoul, Seoul]	[church, restaurant]
3	1000000004	[Jungnang-gu]	[2020-01-30]	[7.0]	[37.6127725]	[127.0981666]	[Seoul]	[hospital]
4	1000000005	[Jungnang-gu]	[2020-01-31]	[9.0]	[37.6127725]	[127.0981666]	[Seoul]	[hospital]

Through the combined table, we can view the action trajectory of any person in the table. Based on this information, we have constructed a graph network based on the action trajectory of a person. Here, we take piatent_id = 1000000008 as an example to show the diagnosis of the diagnosed patient. The course of action is as follows:

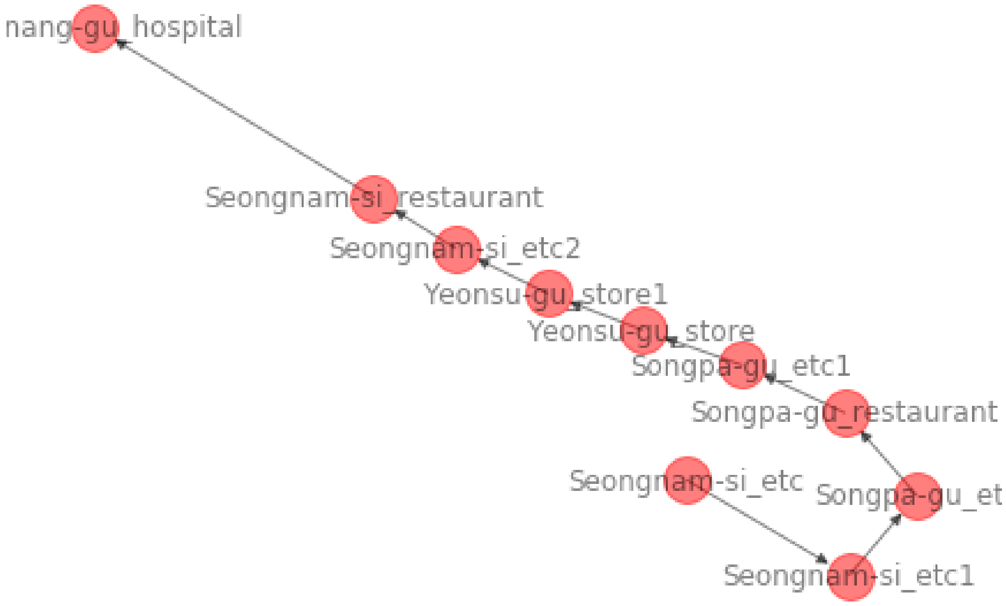


Figure 1 The course of action of a diagnosed patient, where each red dot represents a place, named city_type

In addition, we established a line chart of the number of people diagnosed, died, and recovered on the day according to the time of diagnosis and recovery time (time of death) of each patient after the merger, and a line chart of the cumulative number of people diagnosed, died, and recovered. Among them, the line chart of the currently diagnosed patients is shown in Figure 2:

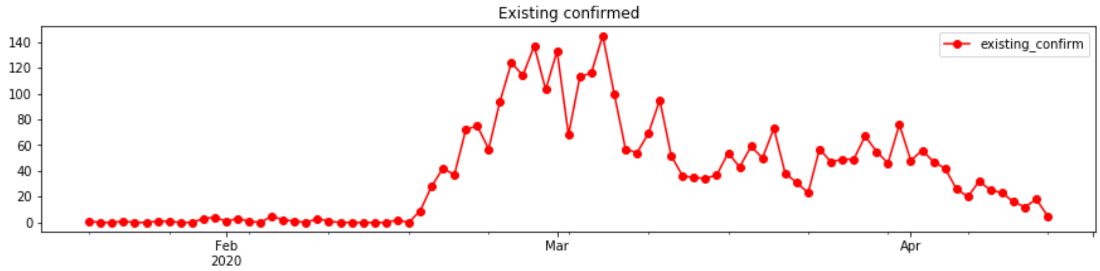


Figure 2 Change curve of existing diagnosed patients

Figure 3 shows the changes in the number of diagnoses, rehabilitation and deaths on the day after the merger. Each point represents the situation on the day, blue indicates the number of diagnoses, orange indicates the number of deaths, and green indicates the number of recovery:

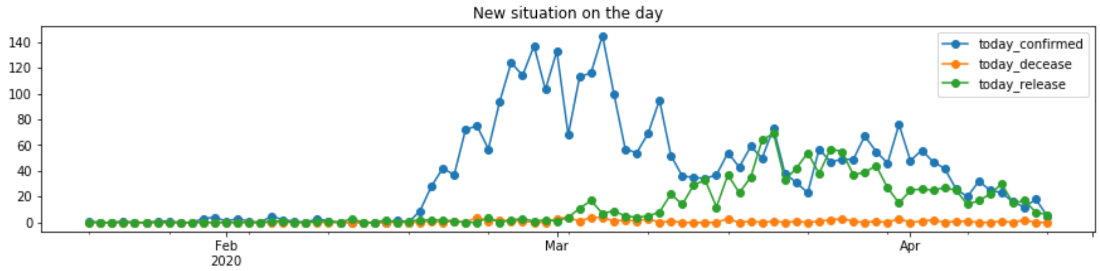


Figure 3 Changes in the number of diagnoses, rehabilitation and deaths on that day

Figure 4 shows the changes in the cumulative number of diagnoses, rehabilitation and deaths after the merger, where each point represents the current day, blue represents the number of diagnoses, orange represents the number of deaths, green represents the number of rehabilitation:

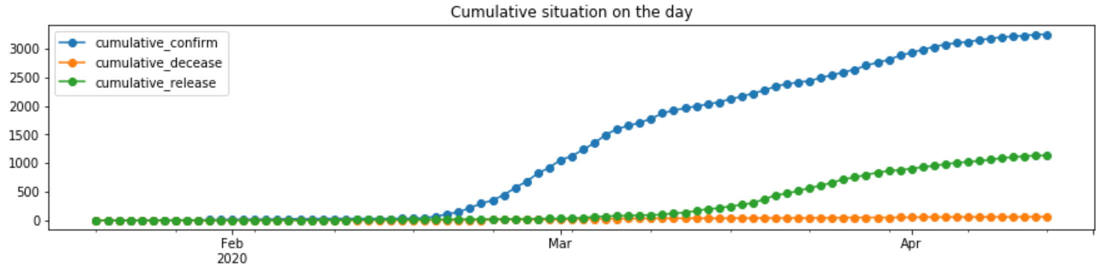


Figure 4 Change curve of cumulative number of diagnosed patients,

number of recovered patients and number of deaths

In addition to showing the changes in the number of diagnosed people, the number of recovered people and the number of deaths over time, we also counted the distribution of the cumulative number of diagnosed people by age and gender. Before this, we first counted the different types of genders and different ages in the table based on the information in the PatientInfo.csv table. There are three cases of ['nan', 'female', 'male'] in the gender column. There are a total of [nan, '0s', '10s', '20s', '30s', '40s', '50s', '60s', '70s', '80s', '90s', '100s'] 11 situations. Figure 5 shows the cumulative number of diagnoses of different genders from January 21, 2020 to April 13, 2020:

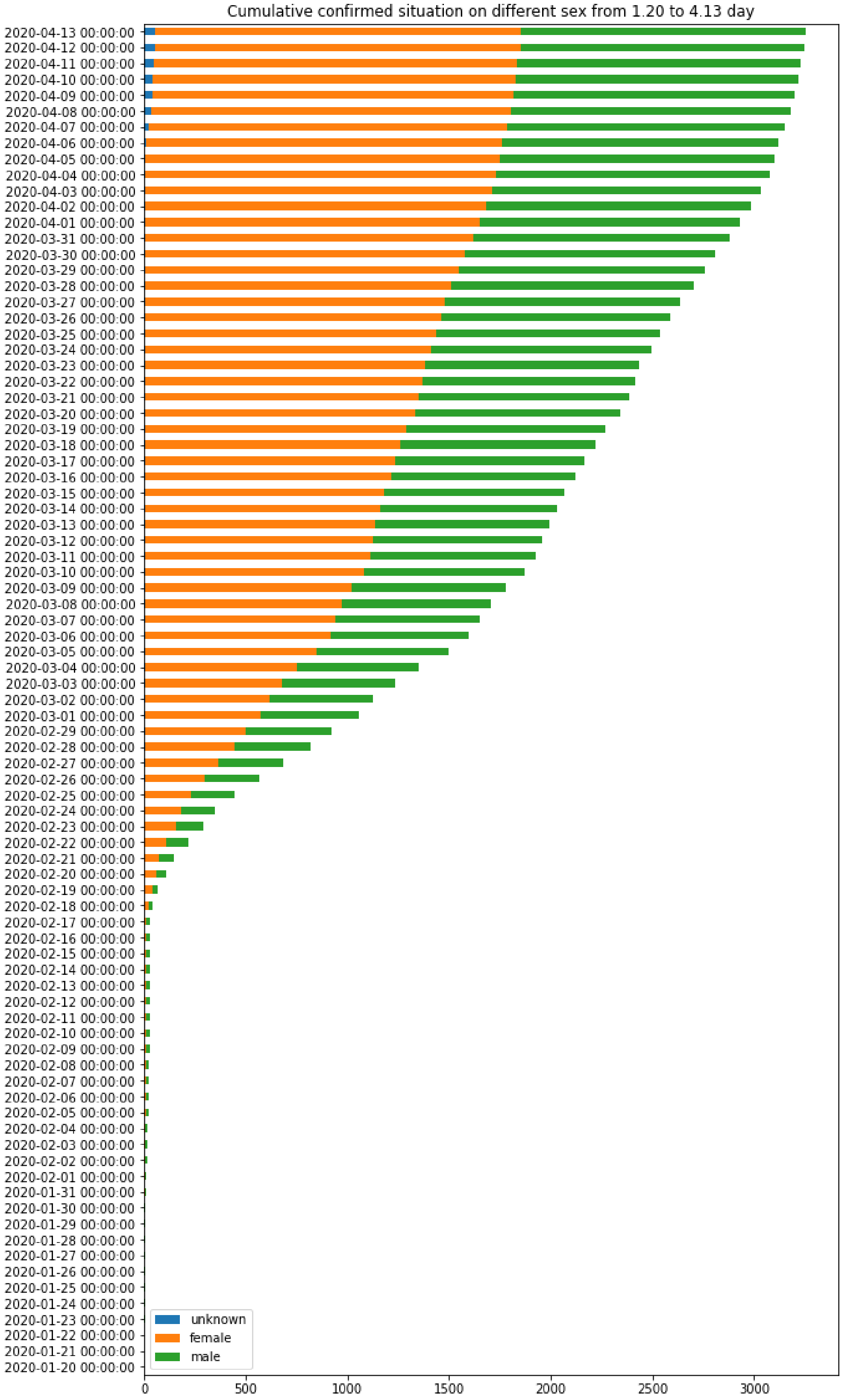


Figure 5 Accumulated number of diagnoses by gender

Figure 6 shows the cumulative number of diagnoses at different ages from January 21, 2020 to April 13, 2020:

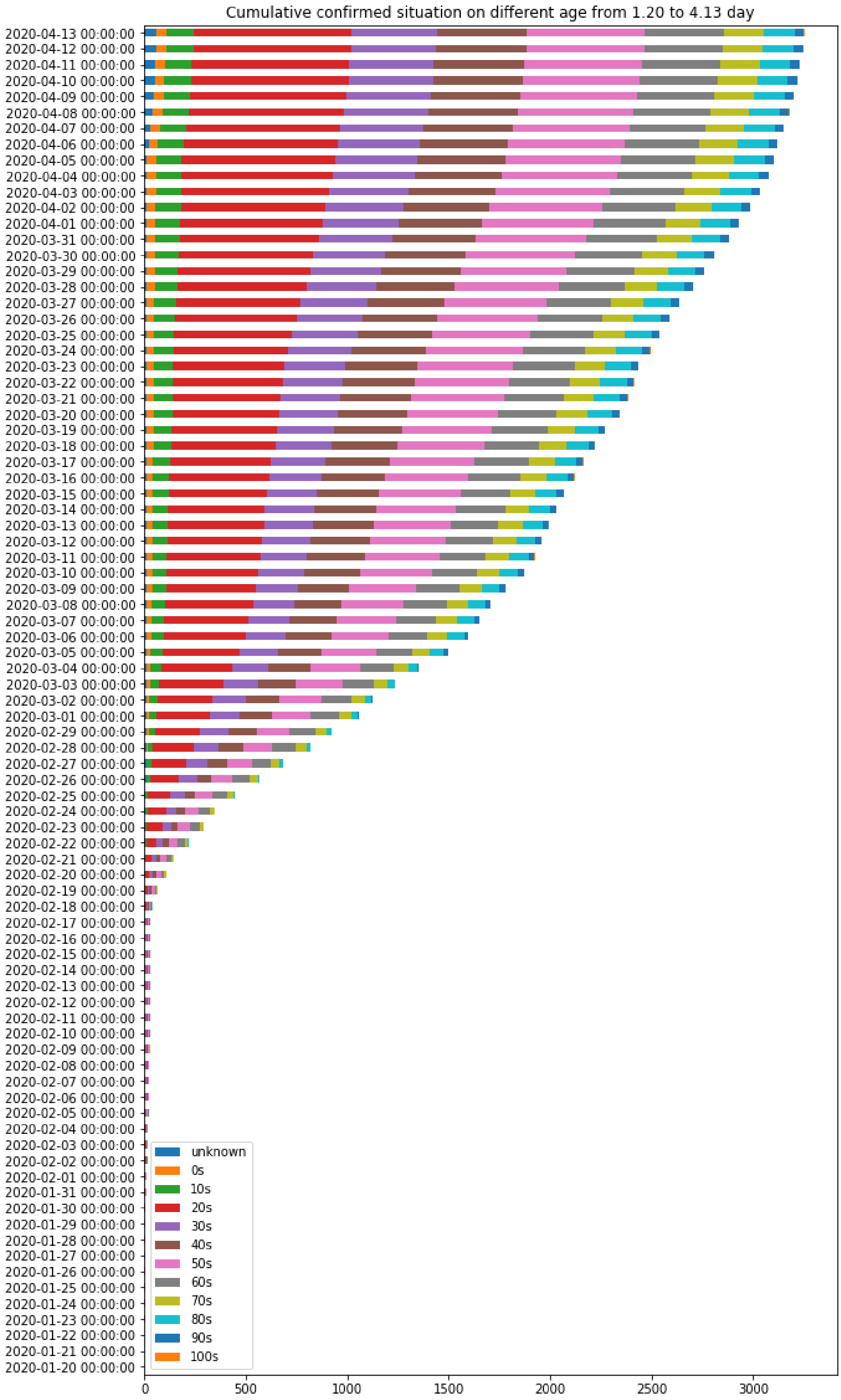


Figure 6 Accumulated number of diagnoses in different age groups

4. Discussion and summary

As can be seen from Figure 2, since February 20th, the number of diagnoses has increased rapidly, and the number of diagnoses on the day in Figure 3 and the cumulative number of diagnoses in Figure 4 also increased rapidly at this time, indicating that patients at this time point May have clinical manifestations, based on which the outbreak time of epidemic transmission can be estimated forward based on the incubation period of the new coronavirus. The green dot in Figure 3 began to rise rapidly around March 11. Although there were small fluctuations afterwards, the overall increase continued, indicating that many patients have recovered. Based on this and the time of the outbreak, it can be estimated from The time period from diagnosis to treatment and recovery. The orange curve in Figure 3 has been relatively stable, indicating that the mortality rate has been kept at a relatively low level, which also indicates that although the new coronavirus infection rate is high and the spread rate is fast, the mortality rate is not very high, which may also be It is because of the improvement of medical level that patients have higher recovery rate and lower mortality rate.

Figure 5 shows the change of the cumulative number of diagnoses under different genders. From the figure, it can be seen that except for a small number of patients who do not know the gender, the ratio of male patients to female patients is basically close to 1: 1, which also shows the difference The risk of being infected with the new coronavirus is almost the same by gender.

Figure 6 shows the change of the cumulative number of diagnosed people at different ages. From the figure, it can be seen that this part of the population at the age of '20s' is the largest proportion of people infected with new coronavirus, followed by the population of the' 50s' age group, and '0s The number of people diagnosed in the three groups of ',' 90s', and '100s' is the least. These three types of people may be due to less outings at ordinary times, and the risk of contact with the diagnosed personnel is lower, so the reason for less infection, and the' 20s' Social is more extensive, so it is most susceptible to new coronavirus.

5. Future works

Although the age and gender of the diagnosis of the new coronavirus is analyzed, we still need more effective data information to analyze the possible trend of the new coronavirus. In the future, we are going to analyze in which areas may be higher according to the geographic location of the diagnosed person Infection rate to remind people to go to these places as little as possible, thereby reducing the number of newly diagnosed.

6.Reference

[1] https://www.kaggle.com/vanshjatana/analysis-on-coronavirus