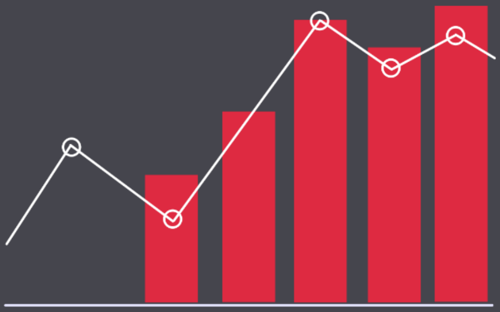
COVID - 19

Data Analysis



The coronavirus COVID-19 pandemic is the defining global health crisis of our time and the greatest challenge we have faced since World War Two. Since its emergence in Asia late last year, the virus has spread to every continent except Antarctica. Cases are rising daily in Africa the Americas, and Europe.[1]

Data Wrangling

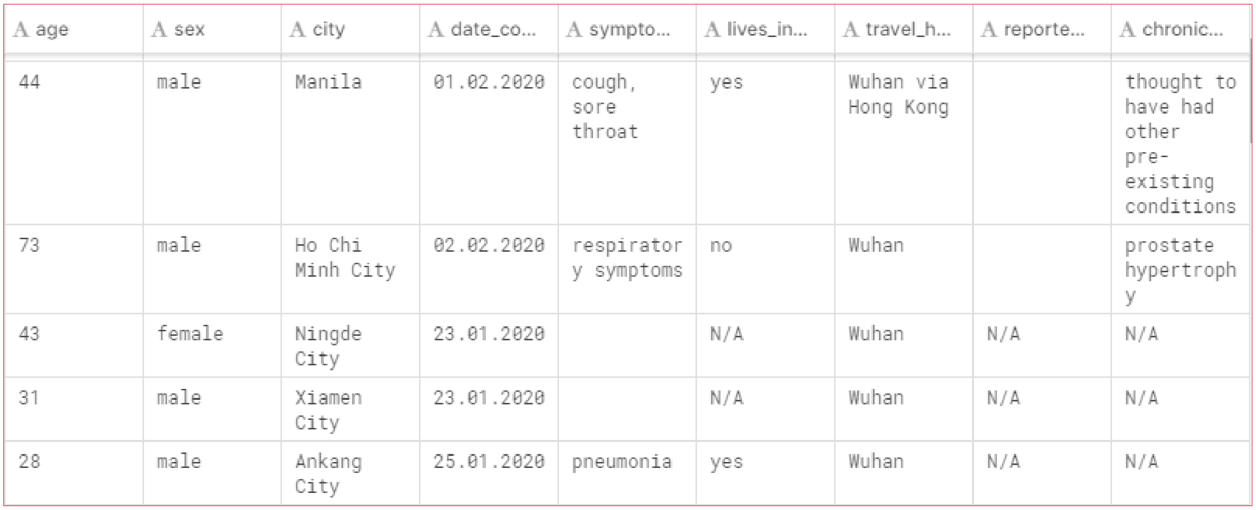
Dataset used in this work is found on Kaggle website on the following link:

https://www.kaggle.com/sudalairajkumar/novel-corona-virus-2019-dataset#COVID19_open_line_list.csv

Dataset has a large amount of missing values for every attribute and was used for the purpose of practicing SQL and Python only. Missing values were filled in with randomly chosen ones for specific attribute.

Data itself includes only one table and it required a lot of cleaning and arranging for it to be in a state that is going to ease working on issues outlined in the Data Analysis section.

Below is the example of few records from the mentioned dataset:



A age	A sex	A city	A date_co...	A sympto...	A lives_in...	A travel_h...	A reporte...	A chronic...
44	male	Manila	01.02.2020	cough, sore throat	yes	Wuhan via Hong Kong		thought to have had other pre-existing conditions
73	male	Ho Chi Minh City	02.02.2020	respiratory symptoms	no	Wuhan		prostate hypertrophy
43	female	Ningde City	23.01.2020		N/A	Wuhan	N/A	N/A
31	male	Xiamen City	23.01.2020		N/A	Wuhan	N/A	N/A
28	male	Ankang City	25.01.2020	pneumonia	yes	Wuhan	N/A	N/A

Figure 1 Kaggle Covid-19 Dataset Snippet

Structured Query Language (SQL) was used for cleaning and arranging data.

Columns that required processing, other than normalizing N/A, NULL and ' ' values, are:

[1] "Coronavirus (COVID-19)." Google News. Google. Accessed April 29, 2020. https://news.google.com/covid19/map?hl=en-US&gl=US&ceid=US:en.

- Column Age:
  Included not valid data such as '19-Oct', 'Belgium', '80-80'. Records that had range instead of age were used to create and fill new columns: AgeUpperLimit and AgeLowerLimit.

- Column City
  Data was not standardized. It included excessive information such as block and site number. One example is: Block 1| Site 11| Whampao Garden.

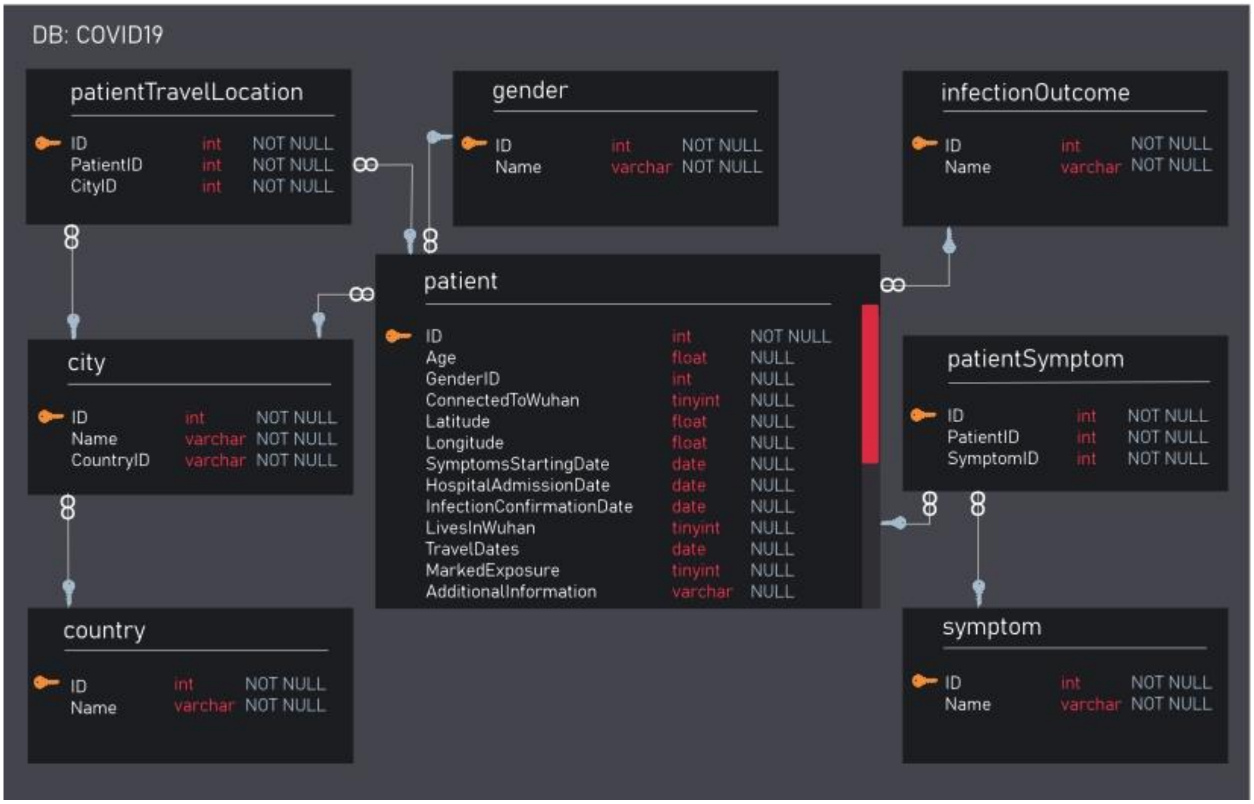
- Column Gender
  New table Gender was created and filled up with male and female values.

- Column HospitalAdmissionDate
  Data in the column contained some erroneous data like partial and not-existing dates. Those values were dropped.

- Column Symptoms
  It contained comma separated, not standardized records of symptoms.

- Column TravelLocations
  Contained comma separated location names in single cell.

The output is a database comprised of multiple tables showed on figure below.



DB: COVID19

patientTravelLocation

ID	int	NOT NULL
PatientID	int	NOT NULL
CityID	int	NOT NULL

gender

ID	int	NOT NULL
Name	varchar	NOT NULL

infectionOutcome

ID	int	NOT NULL
Name	varchar	NOT NULL

patient

ID	int	NOT NULL
Age	float	NULL
GenderID	int	NULL
ConnectedToWuhan	tinyint	NULL
Latitude	float	NULL
Longitude	float	NULL
SymptomsStartingDate	date	NULL
HospitalAdmissionDate	date	NULL
InfectionConfirmationDate	date	NULL
LivesInWuhan	tinyint	NULL
TravelDates	date	NULL
MarkedExposure	tinyint	NULL
AdditionalInformation	varchar	NULL

city

ID	int	NOT NULL
Name	varchar	NOT NULL
CountryID	varchar	NOT NULL

patientSymptom

ID	int	NOT NULL
PatientID	int	NOT NULL
SymptomID	int	NOT NULL

country

ID	int	NOT NULL
Name	varchar	NOT NULL

symptom

ID	int	NOT NULL
Name	varchar	NOT NULL

Figure 2 Database Diagram

EXPLORATORY DATA ANALYSIS

Technologies that have been used for data analysis are:

- Python,
- Pandas,
- Matplotlib,
- SQL.

Short data summary gathered through simple SQL queries is:

Table 1 Data summary

	Number of records	13 159
	Age range	0.08 - 96
	Mean age	44
	Median age	21.5

	Number of countries of residence	41
	Most frequent countries of residence	China (7890 patients) Japan (637 patients)
	Number of cities of residence	418
	Most frequent cities of residence	Yokohama (545 patients) Wenzhou (448 patients)
	Number of patients that live in Wuhan	489
	Number of cities traveled	126

	Date of first symptom reported	2019-12-29
	Number of symptoms reported	79
	Patient temperature range	37 - 40.3

	Date of first admission to hospital	2020-01-04
	Date of first infection confirmation	2020-01-12
⊕	Possible infection outcomes	Death, discharged, recovered, stable
	Date of first death case	2020-01-15

Other issues that have been analyzed to give some insight in trends regarding spread and outcomes of COVID-19 are:

- Number of confirmed cases per day
- Number of confirmed cases per city and country
- Distribution of age in confirmed cases
- Number of patients who were discharged relative to the age of patient
- Number of confirmed cases in Chinese cities per date
- Growth in number of stable, discharged and death cases through days

- Gender ratio in confirmed cases
- Density estimation for infection outcome
- Ratio of total cases and death cases per country
- The most common and the most rare infection symptoms
- Most travelled in world places by infected people

Concise conclusion arrived from data analysis process:

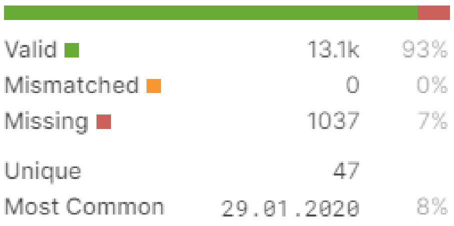
> First reported case of Covid19 infection was on 2019-12-29, if the first symptom reporting date is considered, or on 2020-01-04, if the first admission to hospital is considered. First infection confirmation was reported 2020-01-15. More than half of infected patients were males. Patients' state was categorized into four categories: stable, discharged, recovered and dead. First death case was reported on 2020-01-15.
>
> Number of different symptoms that have been reported by infected patients is 79 with a temperature range from 37C to 40.3C. The most common symptom was fever, followed by cough and sore throat. Youngest patient was less than one year old, while oldest was 96 years old. Most of the patients were in middle ages of life. Patients were from 41 different country, mostly from China and Japan. Significant number of patients have travelled abroad, visiting in total 126 cities, including Wuhan that was visited by the largest number of patients.
>
> Growth in number of confirmed infection cases had its peek point at the end of January 2020. The highest drop in number of confirmed cases per day was in the second week of February 2020.

Each issue was analyzed separately in the following sections.

Number of confirmed cases per day

Number of confirmed cases per month was calculated based on InfectionConfirmedDate column.

Below is the column summary given on the Kaggle website.



Valid 🟩	13.1k	93%
Mismatched 🟧	0	0%
Missing 🟥	1037	7%
Unique	47	
Most Common	29.01.2020	8%

Figure 3 InfectionConfirmedDate Column Summary

The following plot shows trend of change in number of confirmed cases through days. As seen from the plot, the peek point is between 2020-01-22 and 2020-02-01 dates, that is, on the date 2020-01-29, as shown on the detailed plot under.

Number of countries included is 53.

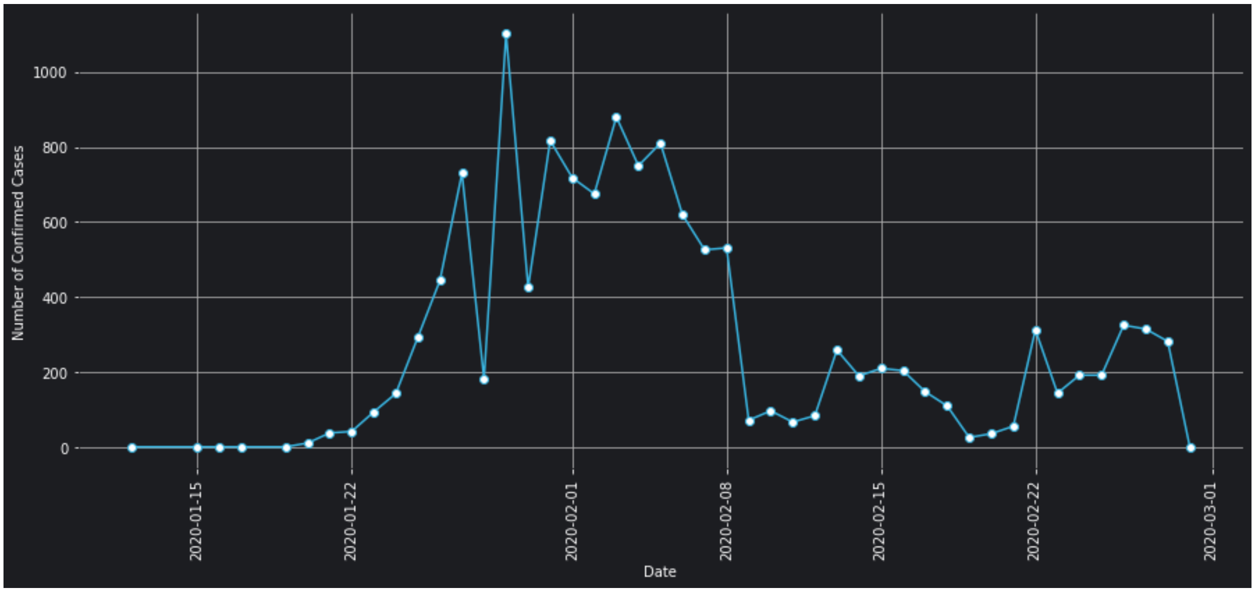


Figure 4 Plot: Number of confirmed cases per day

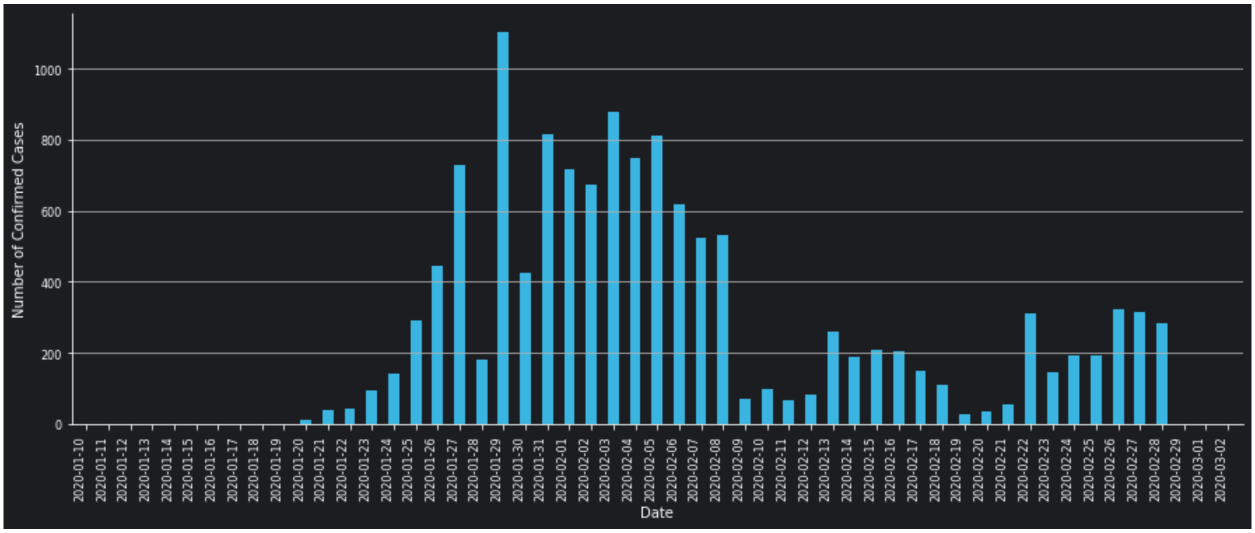


Figure 5 Plot: Number of confirmed cases per day - Detailed version

Number of confirmed cases per city and country

Plot below shows eight cities with highest number of infected cases. All cities are Chinese, with an accent on Wenzhou and Yokohama as the most infected cities. At the same time, Wenzhou is one of the first cities to report confirmed cases of Covid19 infection, while Yokohama is one of the last.

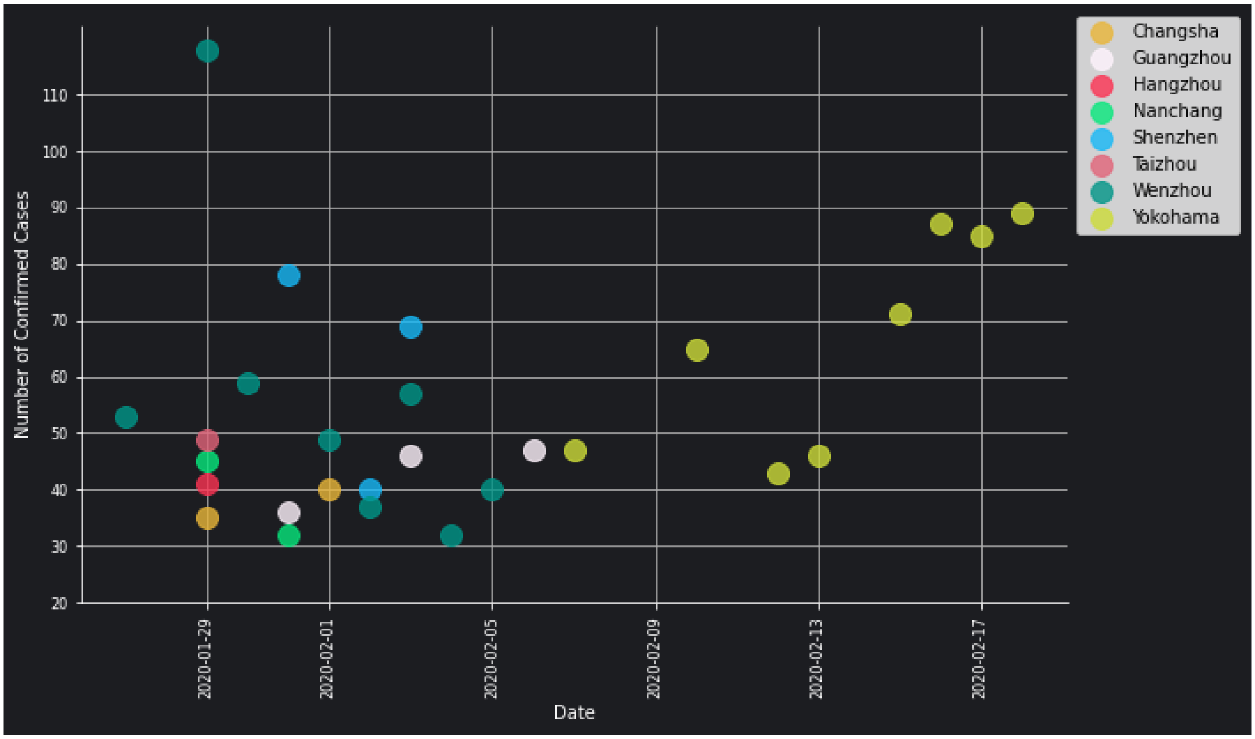


Figure 6 Number of confirmed cases per city

Plot below shows countries with the highest number of infected cases. As expected, China is on the first place, followed by Iran Italy and Japan. In the time of first infection in Japan, number of infected cases in China started to decrease after the peak point in the end of January.

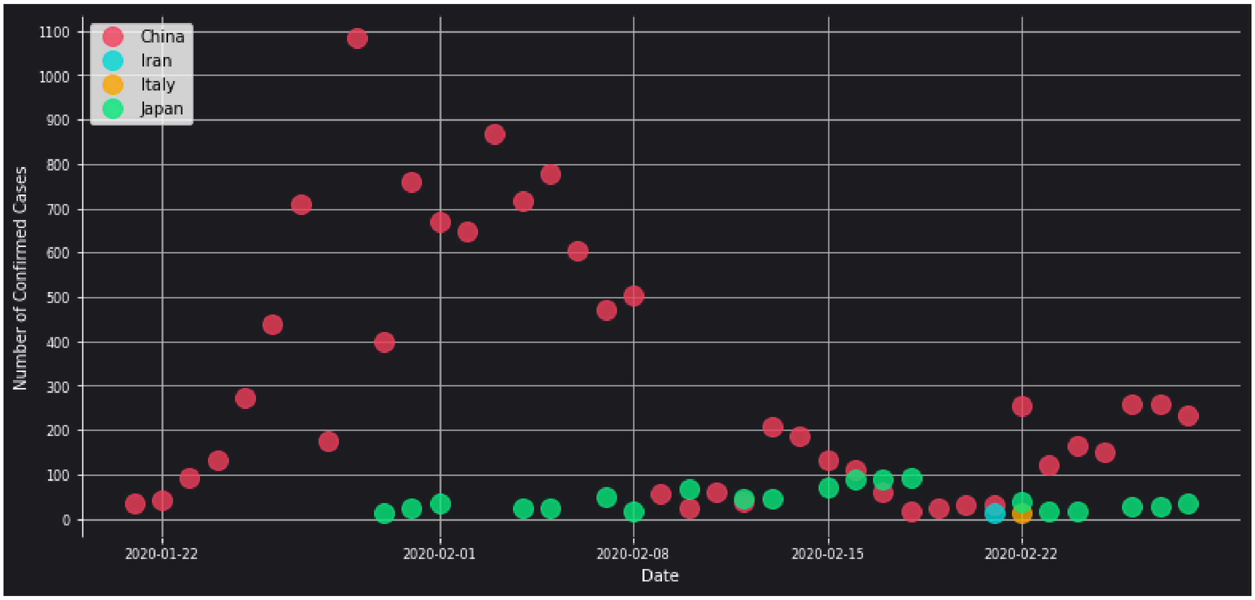


Figure 7 Number of confirmed cases per country

Distribution of age in confirmed cases

Distribution of age in shown on the plot below. As seen from the plot, the distribution is quite symmetrical. Most of the patients were in the middle ages of life, that is around 50 years old, while there are less than 100 cases whose age was less than 20 or greater than 80.

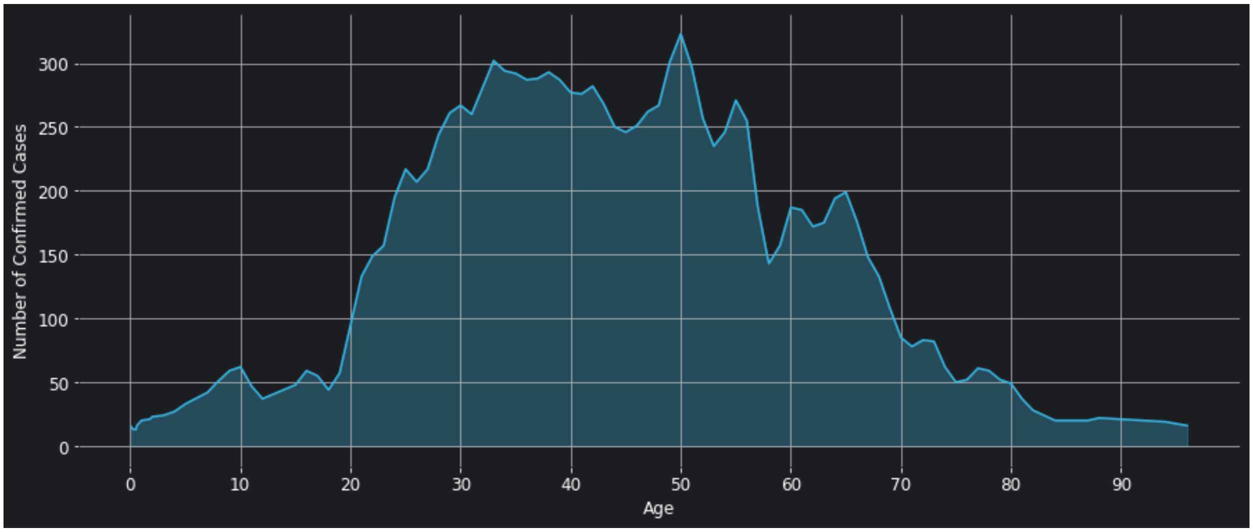


Figure 8 Distribution of age In confirmed cases

Number of patients who were discharged relative to the age of patient

The plot below shows the number of patients who were discharged relative to the age of the patient. As seen from the plot, roughly half of the patient of every age was discharged.

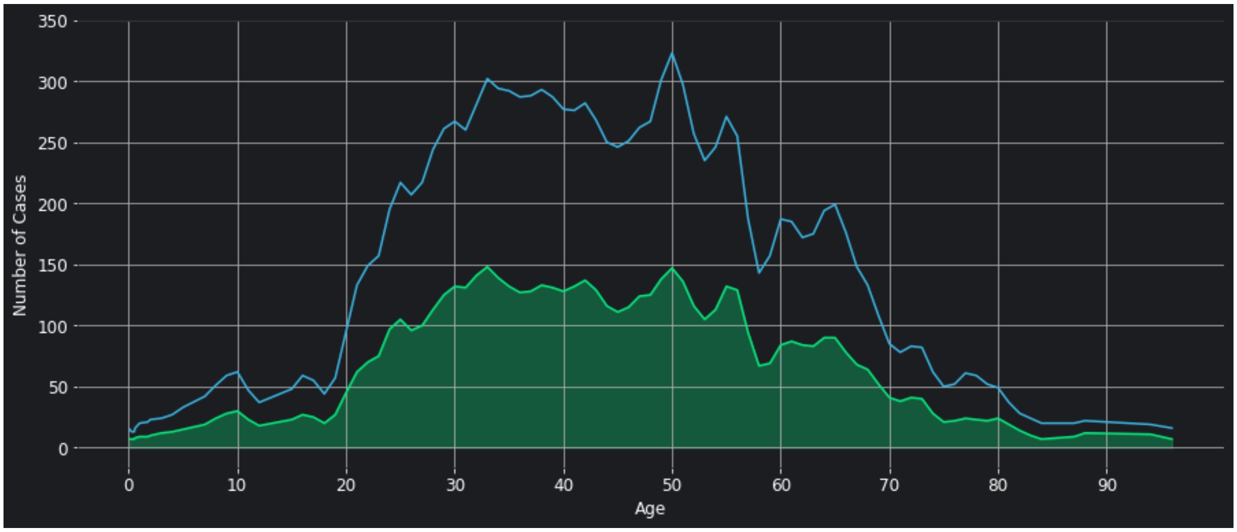


Figure 9 Number of patients who were discharged relative to the age of patient

Gender ratio in confirmed cases

Plot below shows that the total number infected cases includes greater number of male than female persons, that is, male patients are around 53% and female patients are around 47%.

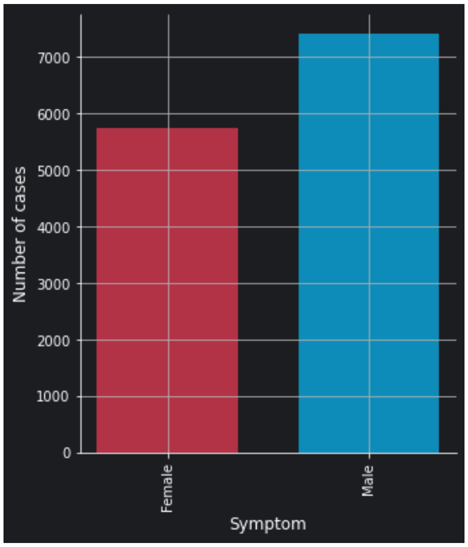


Figure 10 Gender ratio in infected cases

The most and the least common infection symptoms

Most of the COVID-19 patients experienced similar symptoms in the time they were positive on the virus. Fever was reported as the symptom by most of the infected patients, followed by cough, sore throat and fatigue.

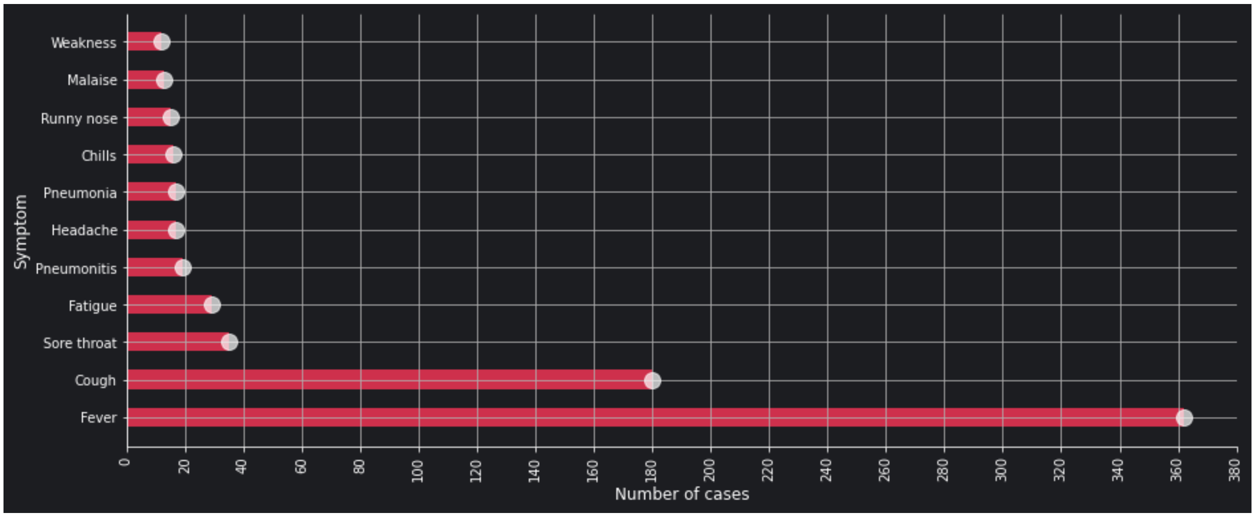


Figure 11 Most common symptoms of infection

The plot below shows symptoms reported by less than eight infected patients. Vomiting and phlegm are the symptoms reported by the smallest number of patients.

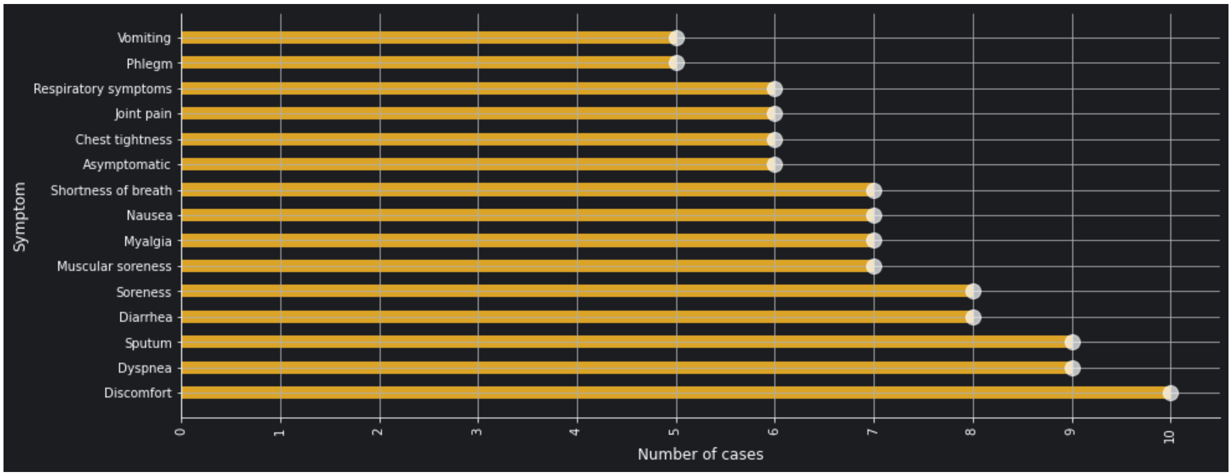


Figure 12 Most rare symptoms of infection

Most travelled in world places by infected people

Plot below shows the most travelled locations by infected people. As expected, Wuhan is on the first place. It is followed by Milan and Beijing.

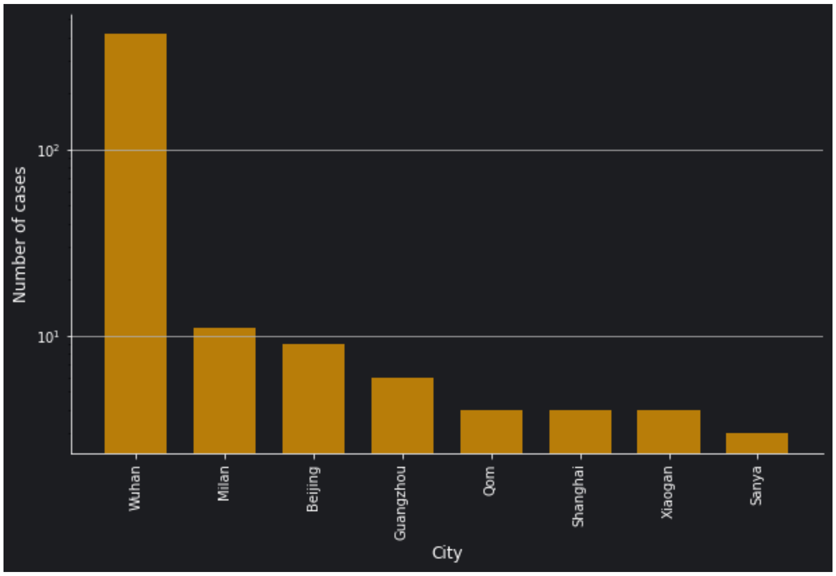


Figure 13 Most travelled locations