PM566 Final Project

Sherry Shen

1. Introduction

1.1 Background

When talking about 2020, "COVID-19", a novel coronavirus firstly found in Wuhan, China, in the end of 2019 (Mclntosh, 2020), must be the most popular word that appears on the Internet, newspapers and daily conversations. Since COVID-19 has been discovered and claimed as human pandemic by Chinese government, there are more than 36 millions of people being confirmed with COVID-19 and about 1 million deaths by now (2020). COVID-19 is definitely raising a huge concern of public health around the world.

1.2 Problem Statement

There are many opinions about the origin of COVID-19. Some suggests that the sequence of COVID-19 is similar to bat virus (Shin, 2020), while others do not support that idea. There are also many unclear characteristics about this new virus. Therefore, this report will mainly focus on discovering main and critical characteristics of COVID-19, including incubation period, COVID-19 distribution of age, gender.

1.3 Data Description

This data is obtained from Kaggle: Novel Corona Virus 2019 Data set. These COVID-19 case data are collected from 20th January to 28th February, 2020, which is the early pandemic period of COVID-19. This data set consists of 10 qualitative attributes, 7 quantitative attributes and 3 other attributes (Table1.1).

Table 1.1 Description of COVID-19 data set

Variable	Description
id	Patient ID
case_in_country	Number of case in country
reporting date	Case reporting date
summary	Patient summary description
location	City of the case
country	Country of the case
gender	Gender
age	Age
symptom_onset	Symptom onset date
if_onset_approximated	Whether symptom onset is approximated
hosp_visit_date	Hospital visiting date
exposure_start	Exposure start date
exposure_end	Exposure end date
visiting Wuhan	Whether the patient visit Wuhan
from Wuhan	Whether the patient is from Wuhan
death	Whether the case die
recovered	Whether the patient recover
symptom	Symptom
source	Source of the patient's information
link	Source link

2. Methodology

The data set that is used in this report is downloaded from a Kaggle page. It offers day-level information and data set on COVID-19 affected cases. First, basic statistics of variables will be computed. Second, basic plots, for instance, box plot, line plot and histograms will be drawn to give ideas of how each variables are distributed. Then, missing value will be evaluated and necessary substitution or deletion of missing value will be used for each variables. Specifically, the following method will be used to deal with specific variable:

1. If a row has all NA in most of the columns, it will be deleted;

2. Missing values in "location", "gender" and "country" will not be imputed;

3. Missing values in "age" will be generated using the mean of "age";

4. Missing values in "hosp visit date" will be imputed by it's "reporting date";

5. Missing value in "symptom onset date" will be imputed by the average of the difference between "symptom onset data" and "hosp visit date";

6. Missing value in "exposure start data" will be imputed by the average of the difference between "exposure start data" and "symptom onset date";

7. Missing value in "exposure end data" will be imputed by "hosp visit data".

8. Missing value in "symptoms" will be imputed by the most frequent symptoms.

In the later steps, time-series plot of COVID-19 cases will be drawn to see how the case number increases by time. Age-specific, gender-specific, country-specific confirm rate will be compared.

3. Results

In the beginning, 3 descriptive variables "summary", "source" and "link" will be excluded and the main data set will have 17 variables and 1085 observations. The following table listed some descriptive statistics about main variables in COVID-19 data set. In order to avoid problems when analyzing sparse data set, different methods of imputing missing values were used to deal with variables individually.

Table3.1 Descriptive statistics of main variables in COVID-19 data set

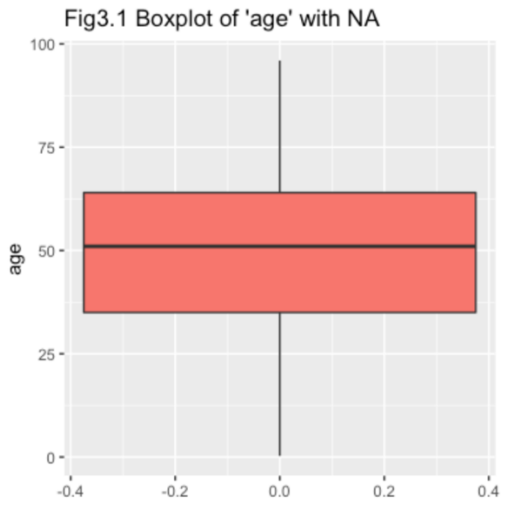
Variable	**Obs/Freq**	**#NA**
country	38	183
gender	F/M=382/520	242
age	25-96: mean = 49.48	222
If_onset_approximated	Yes: 24 No: 536	577
visiting Wuhan	Yes: 192 No: 893	
from Wuhan	Yes: 156 No: 925	4
death	Yes: 63 No: 1022	
recovered	Yes: 159 No: 926	
symptom		815

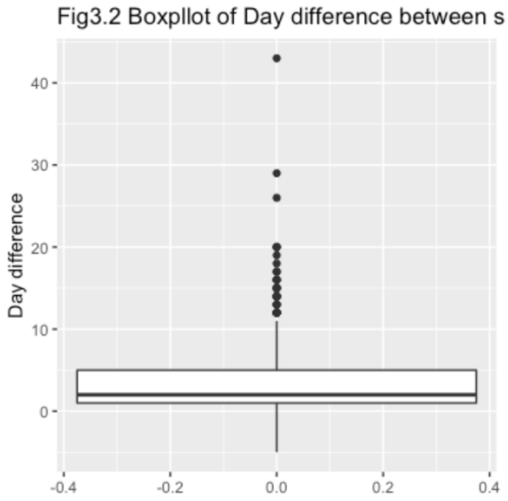
1. When checking the missing values in "reporting data", we found that 262 row has all other attributes equal to "NA". Thus the 262th row is deleted.

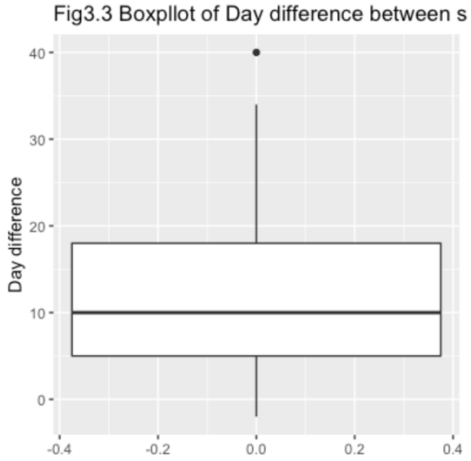
2. The mean of age is around 49 (Fig3.1). Thus 49 is used for imputing the missing values in "age".

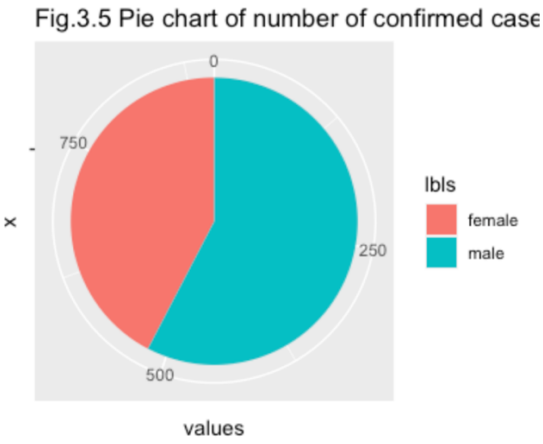
3. The mean difference between hospital visiting date and symptom onset date is 4 (Fig3.3). Then, the missing value in "symptom_onset" is substitute by "hosp_visit_date" plus 4 days.

4. The mean difference between exposure start date and symptom onset date is 12 (Fig3.3). Then, the missing values in "exposure start" are imputed by "symptom onset" minus 12 days.









5. By computing the frequency for the main symptoms, it is easy to find that "fever" is the most common symptom that shows among COVID-19 cases. Thus, the missing values in "symptom" are imputed by "fever".

6. The difference in numbers infected cases between male and female can be seen in Fig3.5. We can see that there are more males than females who are confirmed with COVID-19.

4. Conclusion

First, by computing the mean age of patient, we can see that most of the confirmed cases are around 49 years old, which can be explained by the reason that older people tend to be more infected by virus. Then, we found that the difference between the date of exposure and date of symptom onset is around 14 days. This results is consistent with the claimed incubation period of COVID-19. Third, fever is the most frequent symptom that

shows up on the confirmed cases. It is why fever is concerned as official symptom of COVID-19. Even though there are difference in confirm numbers between gender, the there is not enough evidence show that the infected rate is difference between Male and Female. However, this COVID-19 data is only collected in the early period of the pandemic, which is not enough to draw conclusion outside this data set. Thus, more statistical tests should be done and a updated data set is required in future steps.

Reference

https://www.uptodate.com/contents/coronavirus-disease-2019-covid-19-epidemiology-virology-and-prevention

https://medium.com/microbial-instincts/the-latest-theory-that-may-answer-the-origin-of-covid-19-d9efbe7072ae