1.Introduction

COVID-19 is spreading in the world, and many people in different countries use the open to try to forecast and analysis in order to control it as soon as possible. I will use three different data sets including Korea patient information, Canada patient information and the cases number of all over the world to do COVID-19 data visualization from January to April 2020 in this paper, answer the questions and try to find the increasing rate between countries and who be most affected by this virus as well, mostly by R and some by Tableau.

2.Data Wrangling

2.1Data format

The three separate csv data sets are read in R as data frames. I find the format of three data frames looks correctly after checking the data type of each column, column name and each column presents one type of information of COVID by seeing the first few rows of data. Then I use “str” function to see the type of value of each column. Such as Korea dataset shown as figure1.

```
> str(df_k)
'data.frame':   3326 obs. of  18 variables:
 $ patient_id        : num  1e+09 1e+09 1e+09 1e+09 1e+09 ...
 $ global_num        : int  2 5 6 7 9 10 11 13 19 21 ...
 $ sex               : Factor w/ 3 levels "","female","male": 3 3 3 3 2 2 3 3 3 2 ...
 $ birth_year        : int  1964 1987 1964 1991 1992 1966 1995 1992 1983 1960 ...
 $ age               : Factor w/ 12 levels "","0s","100s",..: 8 6 8 5 5 8 5 5 6 9 ...
 $ country           : Factor w/ 11 levels "Canada","China",..: 6 6 6 6 6 6 6 6 6 6 ...
 $ province          : Factor w/ 17 levels "Busan","Chungcheongbuk-do",..: 16 16 16 16 16
16 16 16 16 16 ...
 $ city              : Factor w/ 152 levels "","Andong-si",..: 40 90 88 91 119 88 88 32 1
26 119 ...
 $ disease           : logi  NA NA NA NA NA NA ...
 $ infection_case    : Factor w/ 24 levels "","Bonghwa Pureun Nursing Home",..: 19 19 5 1
9 5 5 5 19 19 5 ...
 $ infection_order   : int  1 1 2 1 2 3 3 1 2 3 ...
 $ infected_by       : num  NA NA 2e+09 NA 1e+09 ...
 $ contact_number    : int  75 31 17 9 2 43 0 0 68 6 ...
 $ symptom_onset_date: Factor w/ 70 levels "","  ","2020-01-19",..: 4 1 1 5 1 1 1 1 1 1 ..
.
 $ confirmed_date    : Factor w/ 78 levels "","2020-01-20",..: 3 6 6 6 7 7 7 9 11 11 ...
 $ released_date     : Factor w/ 67 levels "","2020-02-05",..: 2 19 10 7 13 10 5 13 12 17
...
 $ deceased_date     : Factor w/ 34 levels "","2020-02-19",..: 1 1 1 1 1 1 1 1 1 1 ...
 $ state             : Factor w/ 3 levels "deceased","isolated",..: 3 3 3 3 3 3 3 3 3 3 .
..
```

Figure

In the three data sets, the columns won’t be used to do exploration or analysis are deleted, such as “global_num”, “symptom_onset_date” and “case_source”.

2.2 Missing Value

I find there are many missing values (“NAN” or “Not reported”) in Korea and Canada dataset, so I plot the missing value percentage for the columns having missing value by ggplot, show as figure 2 and figure3.

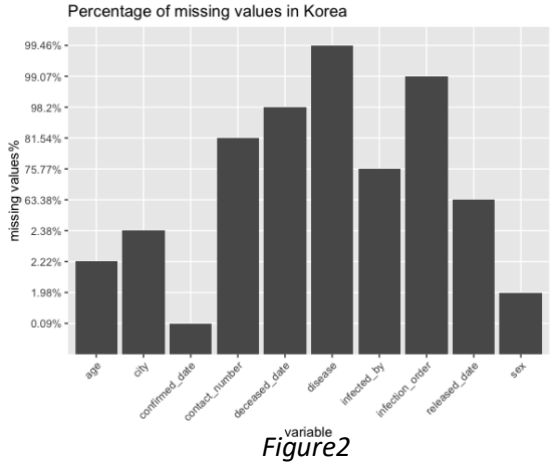

Figure2

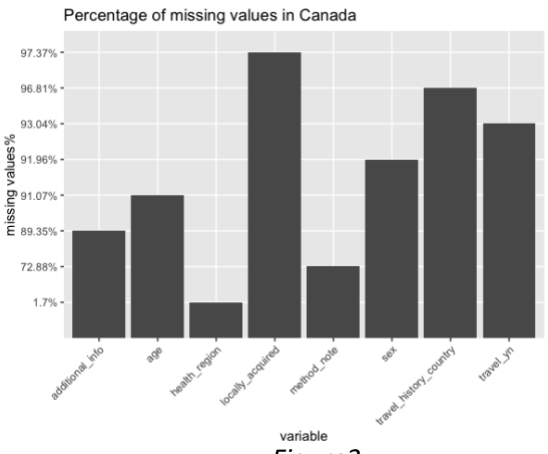

Figure3

The plots show that some columns miss many values, and most of them I cannot fill them since the data is about patients' personal information. For the world case number data set, there is only one column "province" have missing value, the percentage is 45%.

2.3 Outlier and error

Then checking the outlier and errors for the three datasets, such as the age range, date range, the name of province and city of country, whether date format are same, whether the released or deceased date after the confirmed date and 'travel or not' corresponding with 'travel country'.

After the data wrangling, the format and information for the three dataset looks all good. For instance, the world case number as shown in figure4.

```
> head(df_w_all)
  SNo ObservationDate Province.State Country.Region Confirmed Deaths Recovered
1   1      01/22/2020          Anhui Mainland China         1      0         0
2   2      01/22/2020        Beijing Mainland China        14      0         0
3   3      01/22/2020      Chongqing Mainland China         6      0         0
4   4      01/22/2020         Fujian Mainland China         1      0         0
5   5      01/22/2020          Gansu Mainland China         0      0         0
6   6      01/22/2020      Guangdong Mainland China        26      0         0
```

Figure4

3.Data Exploration

3.1 Q1

Q1: Which country or area have the fastest growing case number in recent three month?

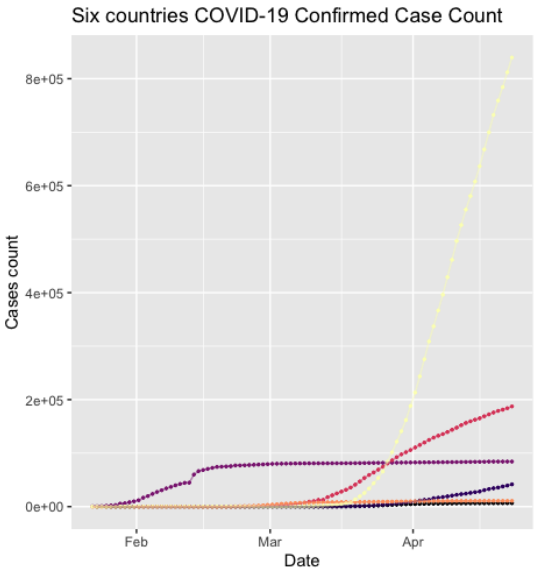

Figure5

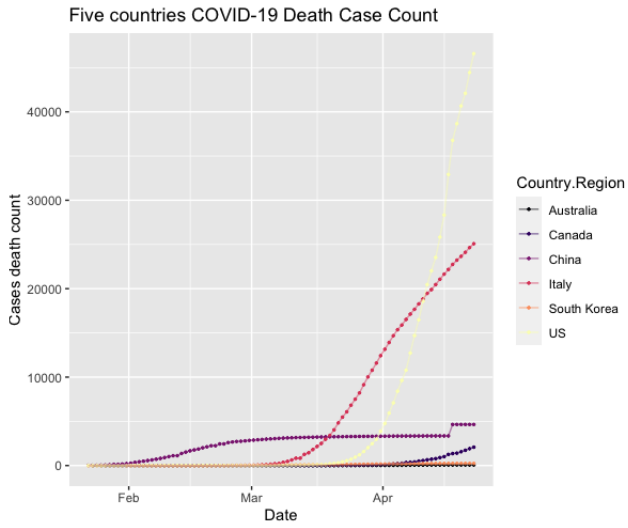

Figure6

Not just increasing rate of confirmed case. As figure5 and figure6 shows, I plot the number of six countries confirmed cases and death case increasing line trends with the date in R, the fastest growing cases country is US starting from Mar 15th and the increasing rate keeping raise nearly vertical. China grows fast before Feb 15th, but the increasing rate not so much larger as US and it control the spreading of virus. The number of cases in Italy larger than US before the intersection point which is between Mar 15th and Apr 1st, but the increasing rate seems same after the intersection. South Korea, Canada and Australia do very good work

control the spreading of virus since the number of cases lines and death lines don't grows much than other three countries. The recovery rate line corresponding with the confirmed plot, China has fastest recovery rate since they control the spreading of virus after Feb 15th, as figure7 shown.

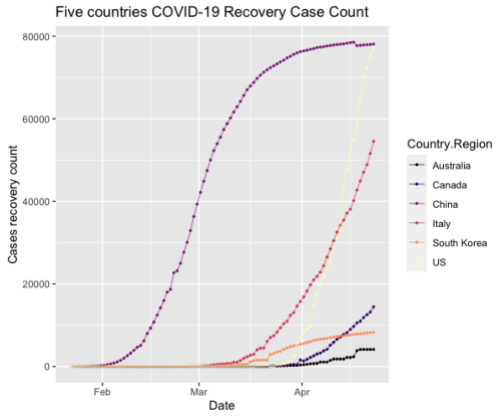

Figure7

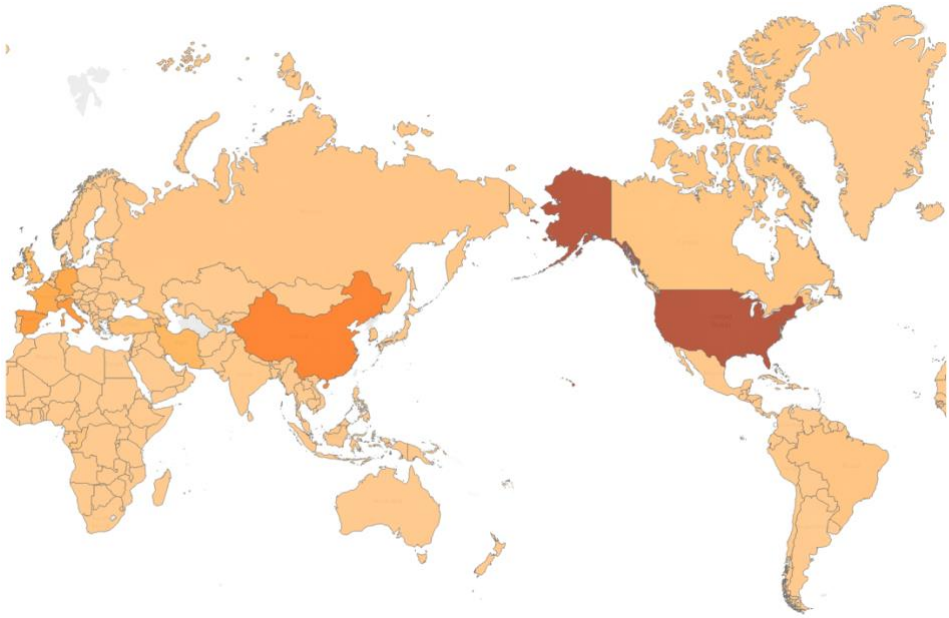

Figure8

Apart from question1, I could also see the cases of countries comparison in map. Since there are many missing value in "province" and no longitude and latitude data for each province, I use Tableau to draw the number of patients of countries all over the world by colour and the colour more shadow, the more number of patients.

3.2 Q2

Q2: Among age, sex, travel history and disease condition, which factor was the most influential factor in whether or not to have COVID in Korea?

In order to answer this question, I analysis each factor firstly, then try to find the relationship between them, all images created by R except figure13 by Tableau. Since the disease condition only has 12 records of "True" and others are null value instead of "False" in total of 3326 records, I don't analysis this factor in this question.

The plot of age, sex and infection place (include travel) shown in figure9, figure10 and figure11.

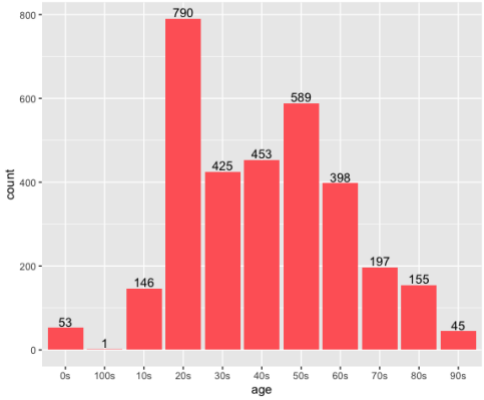

Figure9

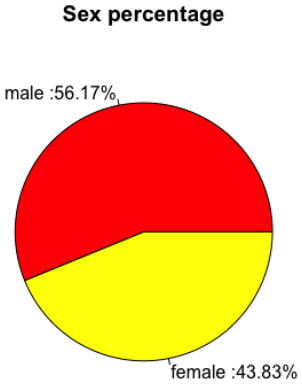

Figure10

In figure9, there are 790 South Korean aged 20s people has COVID, they are the main part of patient, the second large part of patients are aged 50s. The South Korean aged 20s and 50s might be the people most impacted by COVID, but I need to consider with infection place later.

In figure10, the male and female patient nearly have the same percentage of the total patient, so gender is not the main factor affecting whether South Koreans get sick or not.

In datasets, there are many kinds of places in the column of “infection case” and they are both the public places of Korea, so I sign them as “public place”. The figure 11 shows that there are 22.16% of patient have travel history and 77.84% of patients are infected in their country, including contacting with confirmed patient in home or public.

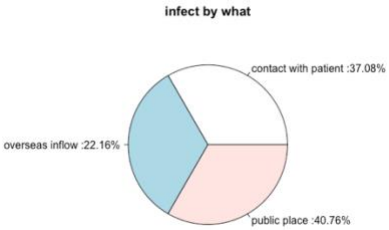

Figure11

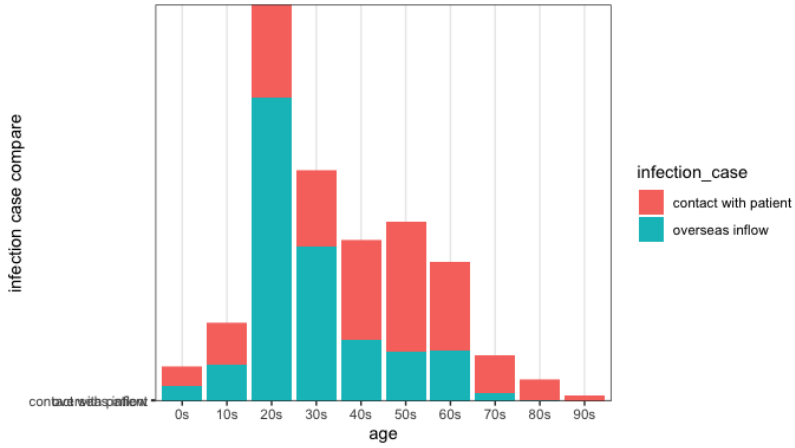

Figure12

As I said above, I need analysis the age with infection case to see why 10s and 50s age are the main patients. Figure12 shows that, in the group of 20s patients, the number of people having overseas history more than ‘contact with patient’ people very much. Young person have energy to travel or they travel due to work, they would contact with patients of other countries, this let 20s aged person becoming the main patient in South Korea.

Apart from the question, in Figure13, the point represents the patients’ location, the size of each point represents the number of patients in one place, different colour with the different province. It shows that, many patients are in “Gyeongsangbuk-do”(yellow) and Soul(pink), there are many points and the points are large.

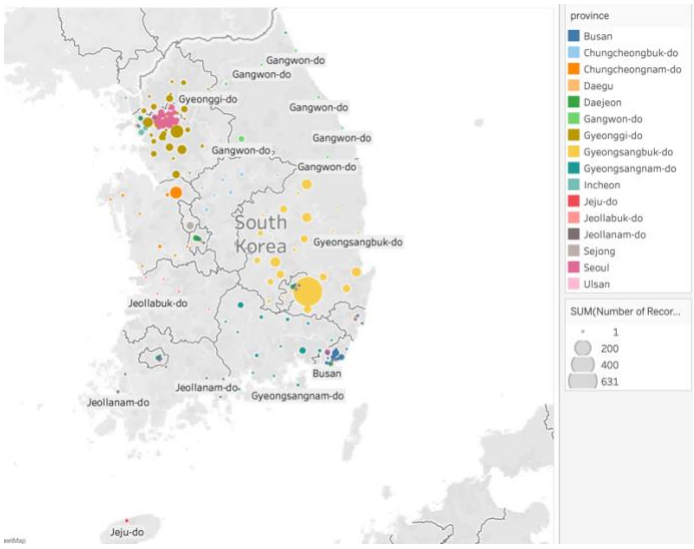

Figure13

3.3 Q3

Q3: Do age, sex and travel history increase the probability of having COVID in Canada and Korea?

Similar with q2, I create the image of each factor in R and analysis their relationship.

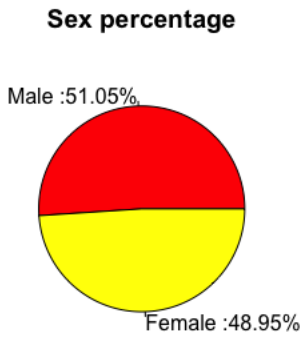

Figure14

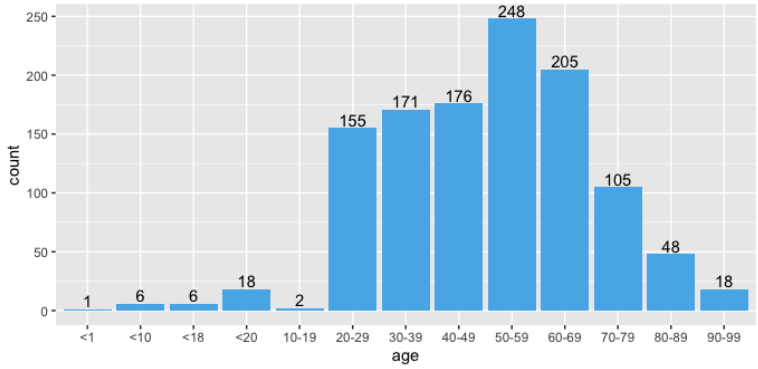

Figure15

Figure14 shows that the gender doesn't impact people whether will be infected, it is same as South Korea. But in Figure15, different from South Korea, the largest part of patients is aged 50s, and the second age group patient is '60-69'. I will analysis the age with other factor below.

Figure16

I created a waffle plot in R, the figure 16, it shows that most patients in Canada had travel history. Similar with South Korea, travel or not is a factor impacting people be infected.
In figure 17 below, I list the count of travel history countries that the patient visited in bar plot, it shows that US are most Canada patient travelled, it might be caused by the country location.

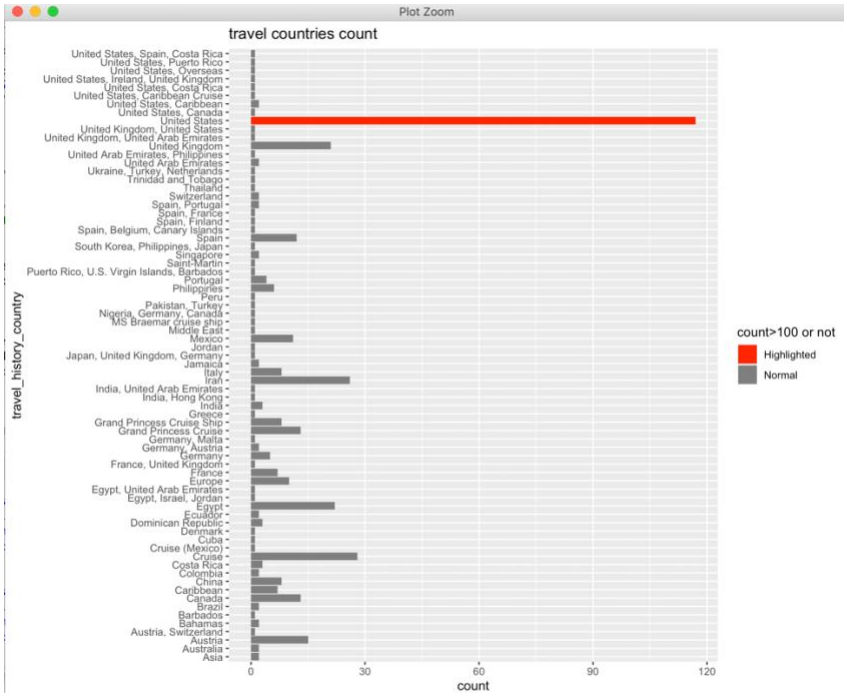

Figure17

In the figure 18, I use Tableau to show how many Canada patient have the travel history in each age group. It is clearly that, most of patient has travel history except the young child age less than 1and elder people aged more than 90s. Not similar with South Korea, the most largest aged group people have overseas flow, in Canada almost every age group have travel history.

Figure18

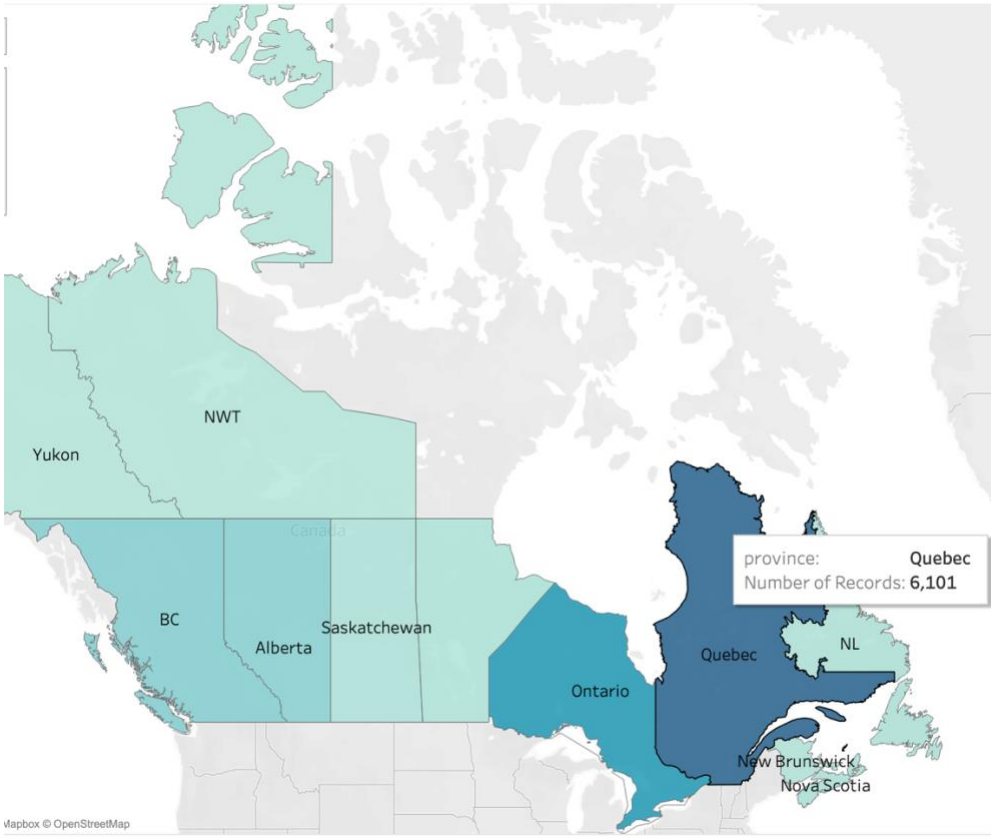

Figure19

The firgure19 show how many Canada patients in each province on map, the province has the most numbers of patients is Quebec, there are 6101 patients.

4.Conclusion

In the six countries, the US has the fastest confirmed and death increasing rate. For the factors sex age, travel history or not, sex or gender don't impact the people are infected or not, the percentage of female and male are same. But age is the impact factor in South Korea and Canada. In South Korea, the 20s aged patient almost have travel history, other age are infected by travel and contact patient in their country equally. In Canada, most patients all have travel history, so the age become the main infection reason, since 50s aged people's health or immune system might not good as youngers.