STAT184 Final project

Code

Analysis for coronavirus data.

The 2019–20 coronavirus pandemic is an ongoing pandemic of coronavirus disease 2019 (COVID-19) caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). The outbreak was identified in Wuhan, China, in December 2019. The World Health Organization declared the outbreak to be a Public Health Emergency of International Concern on 30 January 2020, and recognised it as a pandemic on 11 March 2020. As of 23 April 2020, more than 2.62 million cases of COVID-19 have been reported in 185 countries and territories, resulting in more than 183,000 deaths. More than 709,000 people have recovered, although there may be a possibility of relapse or reinfection.

Alt text

What's more, as of now, coronaviruses have infected more than 900,000 people in the United States and caused more than 50,000 deaths. So I downloaded the data files of the number of diagnoses and deaths in each state from the Internet.

I think I have a few things to analysis this data.

1. exploring the case data of us by each state, find the current coronavirus cases. And draw the graph about the growing trend of Pennsylvaina, then compare it to other state in one graph. Ranking the state by case number, I will find the top 5 states in rank and draw the growing trend of their case number in one graph.

First, I load the package into r session.

Hide

```
#loading the packages into r notebook
library(readr)
library(tidyverse)
library(ggplot2)
library(rvest)
library(party)
```

Second, I load the cvs file and webpage into data frames.

Hide

```
#loading the cvs file from same root
us_states_cases <- read_csv("us-states.csv")
us_doctornum <- read_csv("us_doctor.csv")
#loading the url link
page <- "https://en.wikipedia.org/wiki/List_of_U.S._states_and_territories_by_GDP"
#formating the url link to datatable
tableList <- page %>%
  read_html() %>%
  html_nodes(css = "table") %>%
  html_table(fill = TRUE)
gdp_table <-tableList[[3]]
```

Third, I check these data frames by head, tail and glimpse functions.

Hide

```
#checking the table us_states_cases
us_states_cases%>%
  tail(10)
```

date	state	fips	cases	deaths
<date>	<chr>	<chr>	<dbl>	<dbl>
2020-04-18	Tennessee	47	6637	152

date <date>	state <chr>	fips <chr>	cases <dbl>	deaths <dbl>
2020-04-18	Texas	48	18927	487
2020-04-18	Utah	49	2942	25
2020-04-18	Vermont	50	803	37
2020-04-18	Virgin Islands	78	53	3
2020-04-18	Virginia	51	8053	258
2020-04-18	Washington	53	11802	629
2020-04-18	West Virginia	54	825	18
2020-04-18	Wisconsin	55	4199	212
2020-04-18	Wyoming	56	309	2

1-10 of 10 rows

Hide

```
us_states_cases%>%
  glimpse()
```

```
Observations: 2,609
Variables: 5
$ date   [3m[38;5;246m<date>[39m[23m 2020-01-21, 2020-01-22, 2020-01-23, 2020-01-24, 2020-01
-24, 2020-01-25, 2020-01-25, 2020-01-25,…
$ state  [3m[38;5;246m<chr>[39m[23m "Washington", "Washington", "Washington", "Illinois", "W
ashington", "California", "Illinois", "W…
$ fips   [3m[38;5;246m<chr>[39m[23m "53", "53", "53", "17", "53", "06", "17", "53", "04", "0
6", "17", "53", "04", "06", "17", "53", …
$ cases  [3m[38;5;246m<dbl>[39m[23m 1, 1, 1, 1, 1, 1, 1, 1, 1, 2, 1, 1, 1, 2, 1, 1, 1, 2, 1,
1, 1, 2, 1, 1, 1, 2, 2, 1, 1, 3, 2, 1, …
$ deaths [3m[38;5;246m<dbl>[39m[23m 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, …
```

Hide

```
#checking the table us_doctornum
us_doctornum%>%
  head(10)
```

state <chr>	doctor_pre_100000 <dbl>
Massachusetts	449.5
Maryland	386.0
New York	375.1

state <chr>	doctor_pre_100000 <dbl>
Rhode Island	370.0
Vermont	367.1
Connecticut	352.1
Maine	330.2
Pennsylvania	320.5
New Hampshire	315.1
Hawaii	314.1

1-10 of 10 rows

Hide

```
us_doctornum%>%
  glimpse()
```

```
Observations: 50
Variables: 2
$ state             [3m[38;5;246m<chr>[39m[23m "Massachusetts", "Maryland", "New York", "Rho
de Island", "Vermont", "Connecticut", "M…
$ doctor_pre_100000 [3m[38;5;246m<dbl>[39m[23m 449.5, 386.0, 375.1, 370.0, 367.1, 352.1, 33
0.2, 320.5, 315.1, 314.1, 306.5, 303.4, 3…
```

Hide

```
#checking the table gdp_table
gdp_table%>%
  tail(10)
```

R…	Statefederal district or territory <chr>	2019 Q4[note 1] <chr>	% of Nation <chr>	GDP per capita <chr>
48 47	North Dakota	57,400	0.3	75,321
49 48	Alaska	55,759	0.3	76,220
50 49	South Dakota	54,057	0.3	61,104
51 50	Montana	52,948	0.2	49,540
52 51	Wyoming	39,794	0.2	68,757
53 52	Vermont	35,271	0.2	56,525
54 53	Guam	5,859	0.03[A]	35,665
55 54	U.S. Virgin Islands	3,855	0.02[A]	36,802

R…	Statefederal district or territory	2019 Q4[note 1]	% of Nation	GDP per capita
<chr>	<chr>	<chr>	<chr>	<chr>
56	55 Northern Mariana Islands	1,593	0.007[A]	28,164
57	56 American Samoa	634	0.003[A]	11,399

1-10 of 10 rows

Then use filter function to get the newest data of each state and add the case number of all state. Then we get the total case number of US.

Hide

```
new_data<-us_states_cases%>%
  filter(date == "2020-04-18") #filter the specific date
sum(new_data$cases) #get the total case number
```

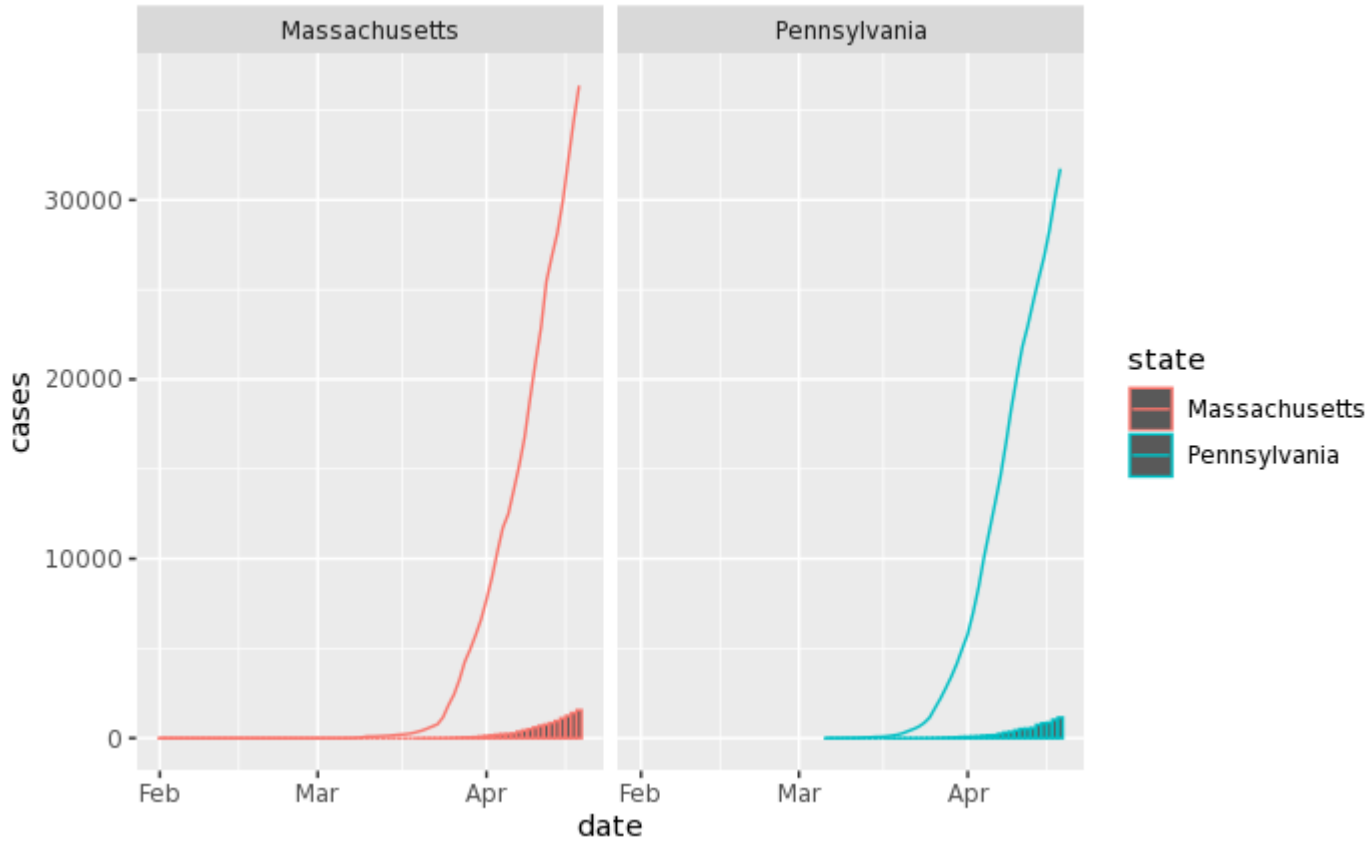
```
[1] 728094
```

I use the filter funtion to get the row that belong to Pennsylvania and Massachusetts, and I assign these rows to an new data frame.

Then I draw the graph that compare the case and deaths number of two states.

Hide

```
penn_cases<- us_states_cases%>% #assign the new data frame
  filter(state == "Pennsylvania"|state == "Massachusetts") #filter the two states
penn_cases%>%
  ggplot(aes(x= date, y = cases))+ #use the ggplot package
  geom_col(aes(y=deaths, color=state))+ #use the column to express the death number
  geom_line(aes(y=cases, color=state))+ #use the line to express the case number
  facet_wrap(~state) #use the facet to define two states
```



Then, I use arrange function to explore what are top 5 five states in case number.

Hide

```
new_data%>%
  arrange(desc(cases))%>% #arranging the column by cases
  distinct(state)%>%
  head(5)
```

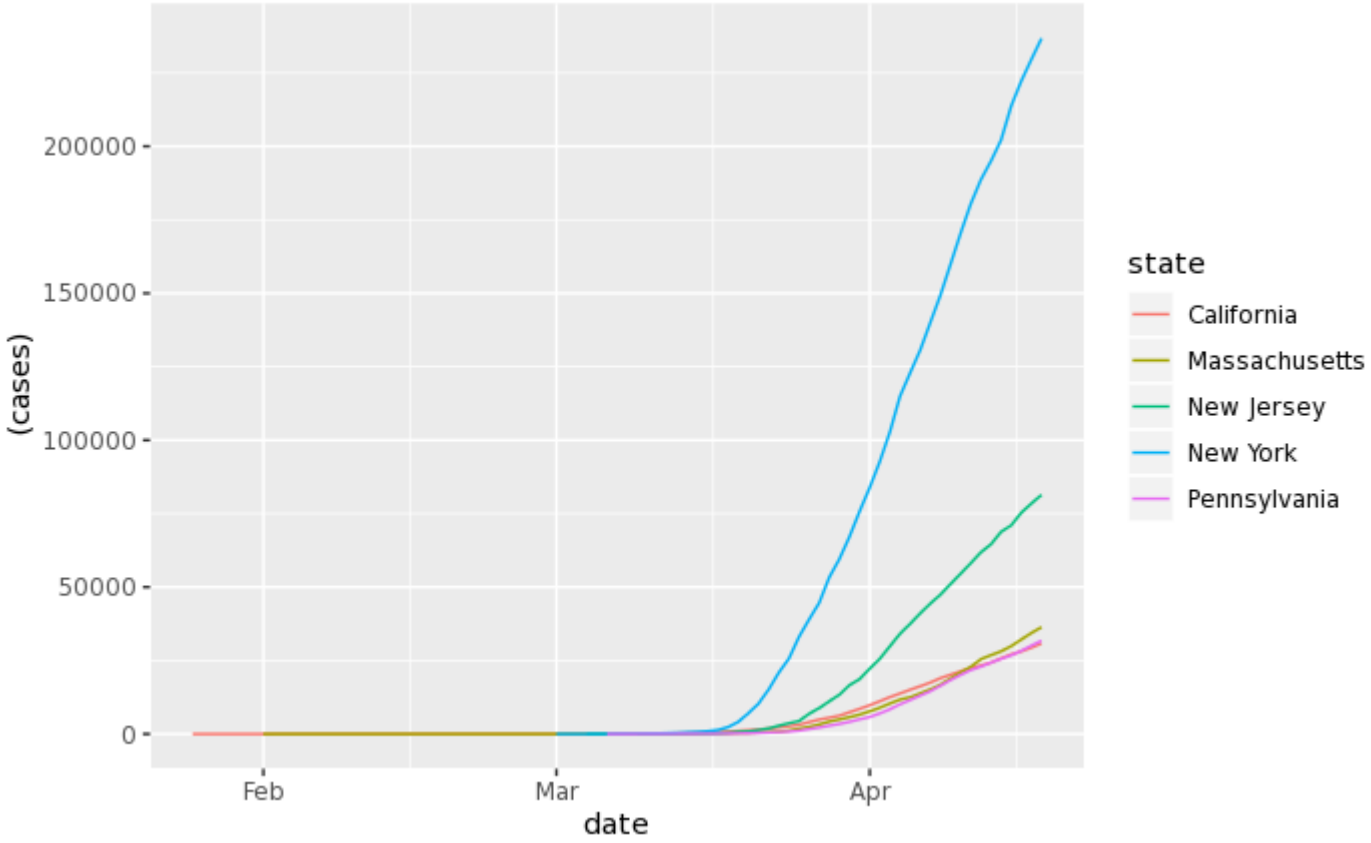
state
<chr>
New York
New Jersey
Massachusetts
Pennsylvania
California

5 rows

Then I draw the graph to show the growing of cases of top 5 states, and I use different color to show the different states.

Hide

```
top5_states <-us_states_cases %>% #assiging the new table.
  filter( grepl( "^New York",state) |grepl( "^New Jersey",state)|grepl( "^Pennsylvania",state)|g
repl( "^Massachusetts",state)|grepl( "^California",state)) #using the regular expressions to fil
ter the top 5 states.
top5_states %>%
  ggplot(aes(x=date, y = (cases), group = state)) + #using ggplot function.
  geom_line(aes(color=state)) #using different color to express different states.
```



Hide

```
NA
```

2. I find the death numbers of different states have very large difference. So, I want to find the factor that can affect the death rate of each state. Thus, I think there might be some variables that attribute to the difference of death rate of different states. If we can find the relation behind it, we can probably reduce the death rate of some region which can save many lives.

From my pointview, I think that the wealth level and the proportion of doctors in the population may be related to the mortality rate. So, I collect other data source. One is the data table about the

doctor number in 100000 people for each state. And another one is the data table about the gdp per-capite for each state in US. I want to draw some graphs to investigate the relation between the deaths rate and these variable.

I transfer the number of doctors in per 100000 people into an new variable “doctor_condition”. The “doctor_condition” has 4 grades which are high, medium, low, and poor.

Hide

```
us_doctornum[,"doctor_condition"] <- NA #creating a empty column
for(i in 1:nrow(us_doctornum)) { #setting the for loop
    if (us_doctornum[i,2]>350){ #setting the if condition
      us_doctornum[i,3]<-"high" #assign the condition to the table
    }
    else if(us_doctornum[i,2]>270 && us_doctornum[i,2]<=350){ #setting the if condition
      us_doctornum[i,3]<-"medium" #assign the condition to the table
    }
    else if(us_doctornum[i,2]>220 && us_doctornum[i,2]<=270){ #setting the if condition
      us_doctornum[i,3]<-"low" #assign the condition to the table
    }
    else{
      us_doctornum[i,3]<-"poor" #assign the condition to the table
    }
}
```

To draw the graph, I combine the newest data with “us_doctornum” doctor table. Then, I add two variables. First one called “death_rate” which is the ratio of death. Second, one called “death_rate” which is the ration of patients and doctor rate.

Hide

```
data_with_doctor_inf <- new_data%>% #assigning the new table
  inner_join(us_doctornum, by = "state")%>% #joining the table
  mutate(death_rate = deaths/cases, doctor_rate = cases/doctor_pre_100000) #mutating the new var
iable

data_with_doctor_inf%>%
  head(10)
```

date	state	fips	ca…	deat…	doctor_pre_100000	doctor_condition	death_rat
<date>	<chr>	<chr>	<dbl>	<dbl>	<dbl>	<chr>	<dbl>
2020-04-18	Alabama	01	4723	147	217.1	poor	0.0311242
2020-04-18	Alaska	02	312	7	276.9	medium	0.0224359

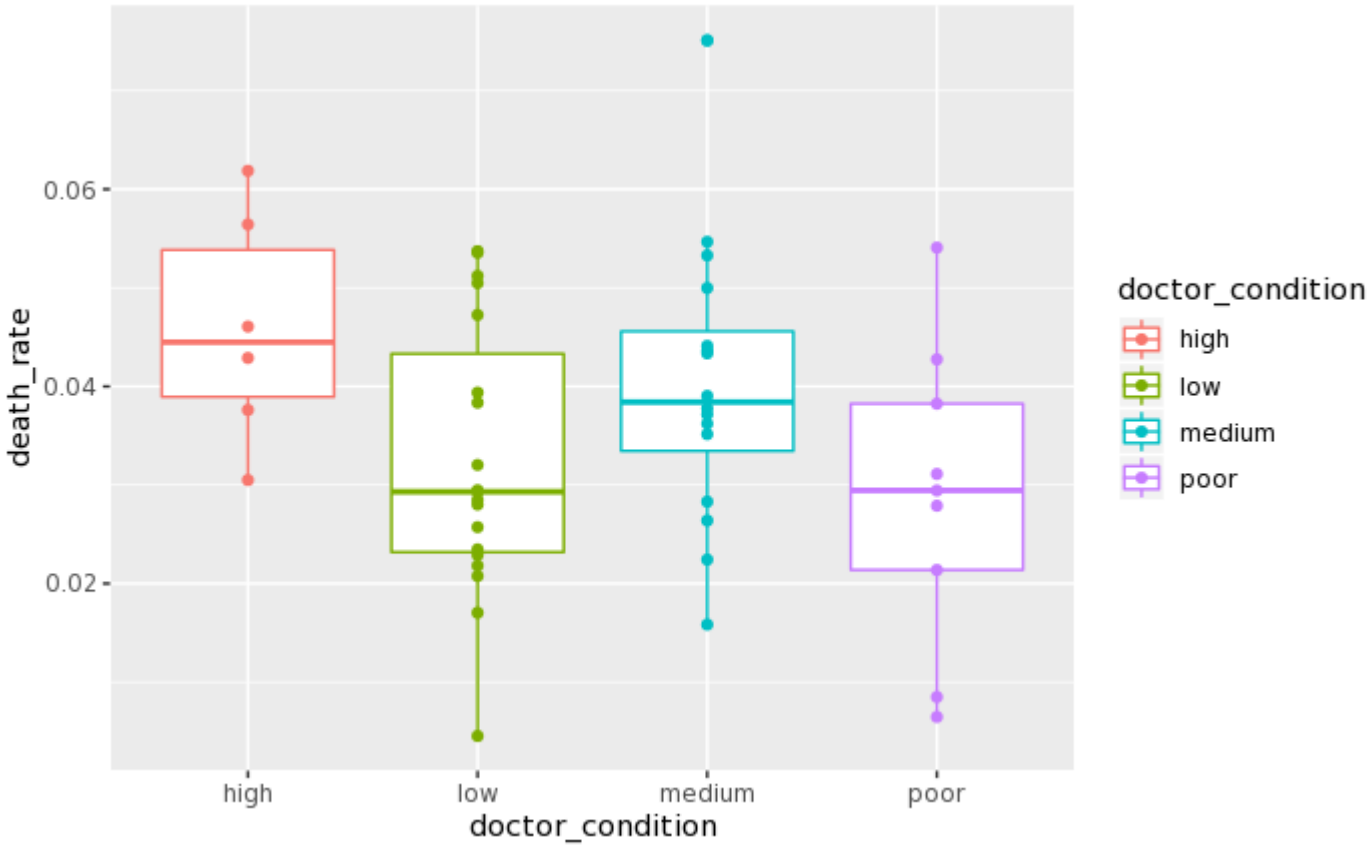
date	state	fips	ca…	deat…	doctor_pre_100000	doctor_condition	death_rat
<date>	<chr>	<chr>	<dbl>	<dbl>	<dbl>	<chr>	<dbl>
2020-04-18	Arizona	04	4719	181	242.0	low	0.0383555
2020-04-18	Arkansas	05	1777	38	207.6	poor	0.0213843
2020-04-18	California	06	30829	1146	279.6	medium	0.0371727
2020-04-18	Colorado	08	9433	409	285.7	medium	0.0433584
2020-04-18	Connecticut	09	17550	1086	352.1	high	0.0618803
2020-04-18	Delaware	10	2538	67	284.6	medium	0.0263987
2020-04-18	Florida	12	25484	747	265.2	low	0.0293125
2020-04-18	Georgia	13	17014	670	228.7	low	0.0393793

1-10 of 10 rows

From the boxplot and point graph, we can see the state with high and medium dotcor condition have higher average death rate. But, the state with low and poor doctor condition have large death rate range.

Hide

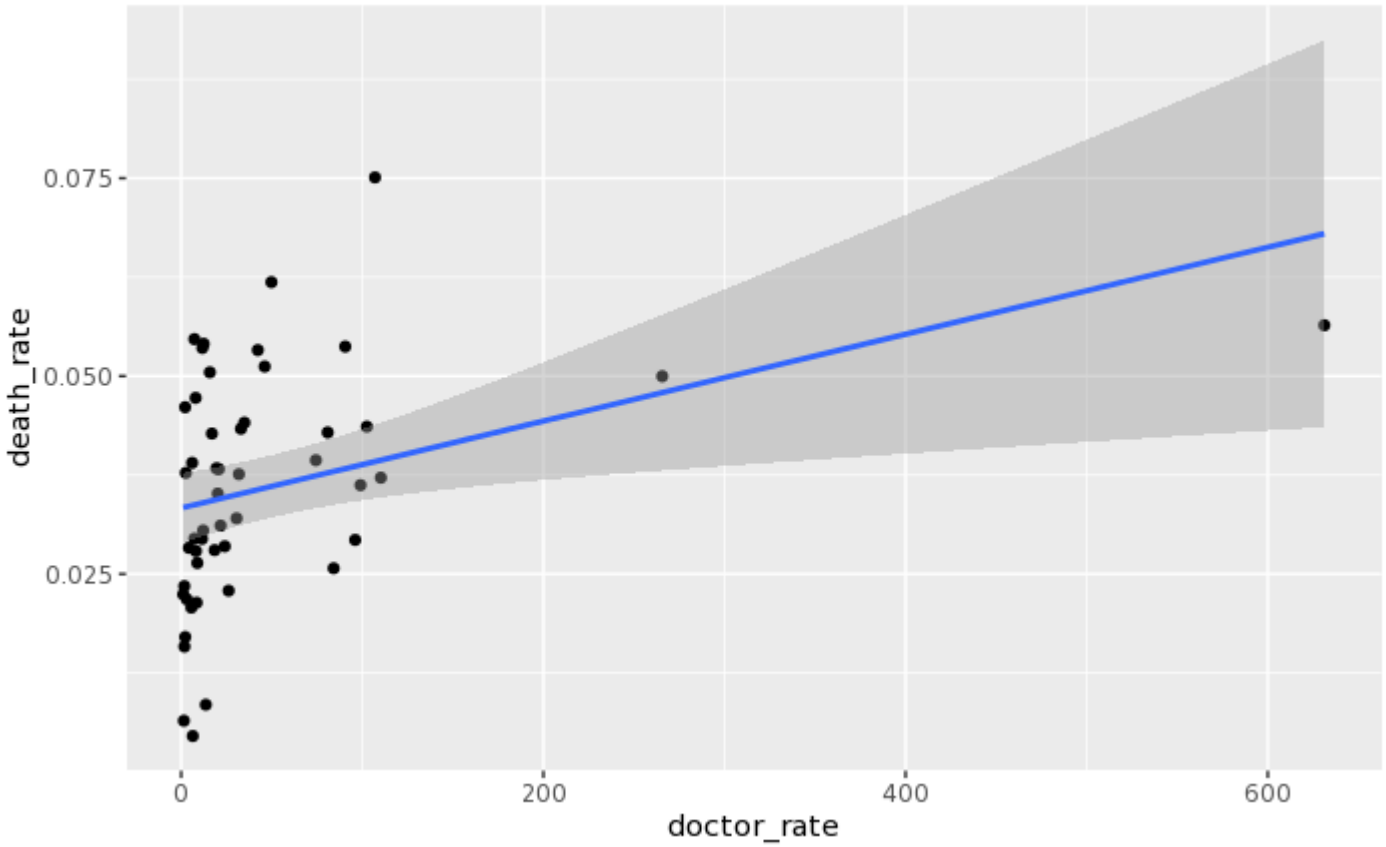
```
data_with_doctor_inf%>%
  ggplot(aes(x=doctor_condition , y = death_rate, color = doctor_condition))+ #setting the aes
  geom_boxplot()+ #drawing the boxplot
  geom_point() #drawing the points
```



But, If we look the realtion between the deaths rate and doctor rate, we can find there are position reation. It's means the death rate will be higher, if doctor have to treat too many patients.

Hide

```
data_with_doctor_inf%>%
  ggplot(aes(x=doctor_rate , y = death_rate))+ #setting the aes
  geom_point()+ #drawing the points
  geom_smooth(method='lm', formula= y~x) #drawing the regression line
```



Then, I clean the gdp table and only select the state name and GDP per capita in this table. Then I assign these two variables to a new table.

Hide

```
per_gdp_table <- gdp_table%>%
  rename(per_gdp = "GDP per capita", #changing the columns names
         state = "Statefederal district or territory")%>%
  select(state,per_gdp) #selecting the variables
per_gdp_table%>%
  tail(10) #checking the table
```

	state <chr>	per_gdp <chr>
48	North Dakota	75,321
49	Alaska	76,220
50	South Dakota	61,104
51	Montana	49,540
52	Wyoming	68,757
53	Vermont	56,525
54	Guam	35,665

	state <chr>	per_gdp <chr>
55	U.S. Virgin Islands	36,802
56	Northern Mariana Islands	28,164
57	American Samoa	11,399

1-10 of 10 rows

I join the GDP per capita table with original table.

Hide

```
complete_data<- data_with_doctor_inf%>%
  left_join(per_gdp_table, by = "state") #joining the table
complete_data$per_gdp<- as.numeric(gsub(",", "", complete_data$per_gdp )) #changing the format o
f per_gdp variable
complete_data%>% #checking the table
  glimpse() #checking the table
```

```
Observations: 50
Variables: 10
$ date              [3m[38;5;246m<date>[39m[23m 2020-04-18, 2020-04-18, 2020-04-18, 2020-04-
18, 2020-04-18, 2020-04-18, 2020-04-18, …
$ state             [3m[38;5;246m<chr>[39m[23m "Alabama", "Alaska", "Arizona", "Arkansas",
"California", "Colorado", "Connecticut", …
$ fips              [3m[38;5;246m<chr>[39m[23m "01", "02", "04", "05", "06", "08", "09", "1
0", "12", "13", "15", "16", "17", "18", "…
$ cases             [3m[38;5;246m<dbl>[39m[23m 4723, 312, 4719, 1777, 30829, 9433, 17550, 25
38, 25484, 17014, 568, 1577, 29160, 1064…
$ deaths            [3m[38;5;246m<dbl>[39m[23m 147, 7, 181, 38, 1146, 409, 1086, 67, 747, 67
0, 9, 44, 1272, 545, 74, 86, 145, 1267, …
$ doctor_pre_100000 [3m[38;5;246m<dbl>[39m[23m 217.1, 276.9, 242.0, 207.6, 279.6, 285.7, 35
2.1, 284.6, 265.2, 228.7, 314.1, 192.6, 2…
$ doctor_condition  [3m[38;5;246m<chr>[39m[23m "poor", "medium", "low", "poor", "medium", "m
edium", "high", "medium", "low", "low", …
$ death_rate        [3m[38;5;246m<dbl>[39m[23m 0.03112429, 0.02243590, 0.03835558, 0.0213843
6, 0.03717279, 0.04335842, 0.06188034, 0…
$ doctor_rate       [3m[38;5;246m<dbl>[39m[23m 21.754952, 1.126761, 19.500000, 8.559730, 11
0.261087, 33.017151, 49.843794, 8.917779,…
$ per_gdp           [3m[38;5;246m<dbl>[39m[23m 47735, 76220, 51179, 44808, 80563, 68828, 810
55, 78468, 51745, 58896, 69593, 46043, 7…
```

Hide

```
complete_data%>%
  head(10) #checking the table
```

date <date>	state <chr>	fips <chr>	ca… <dbl>	deat… <dbl>	doctor_pre_100000 <dbl>	doctor_condition <chr>	death_rat <dbl>

date	state	fips	ca…	deat…	doctor_pre_100000	doctor_condition	death_rat
<date>	<chr>	<chr>	<dbl>	<dbl>	<dbl>	<chr>	<dbl>
2020-04-18	Alabama	01	4723	147	217.1	poor	0.0311242
2020-04-18	Alaska	02	312	7	276.9	medium	0.0224359
2020-04-18	Arizona	04	4719	181	242.0	low	0.0383555
2020-04-18	Arkansas	05	1777	38	207.6	poor	0.0213843
2020-04-18	California	06	30829	1146	279.6	medium	0.0371727
2020-04-18	Colorado	08	9433	409	285.7	medium	0.0433584
2020-04-18	Connecticut	09	17550	1086	352.1	high	0.0618803
2020-04-18	Delaware	10	2538	67	284.6	medium	0.0263987
2020-04-18	Florida	12	25484	747	265.2	low	0.0293125
2020-04-18	Georgia	13	17014	670	228.7	low	0.0393793

1-10 of 10 rows | 1-8 of 10 columns

By using the correlation function, I find the GDP per capita has strong positive relation with cases number and weak relation with death rate.

Hide

```
cor(complete_data$per_gdp,complete_data$cases) #finding the correlation index
```

```
[1] 0.4576911
```

Hide

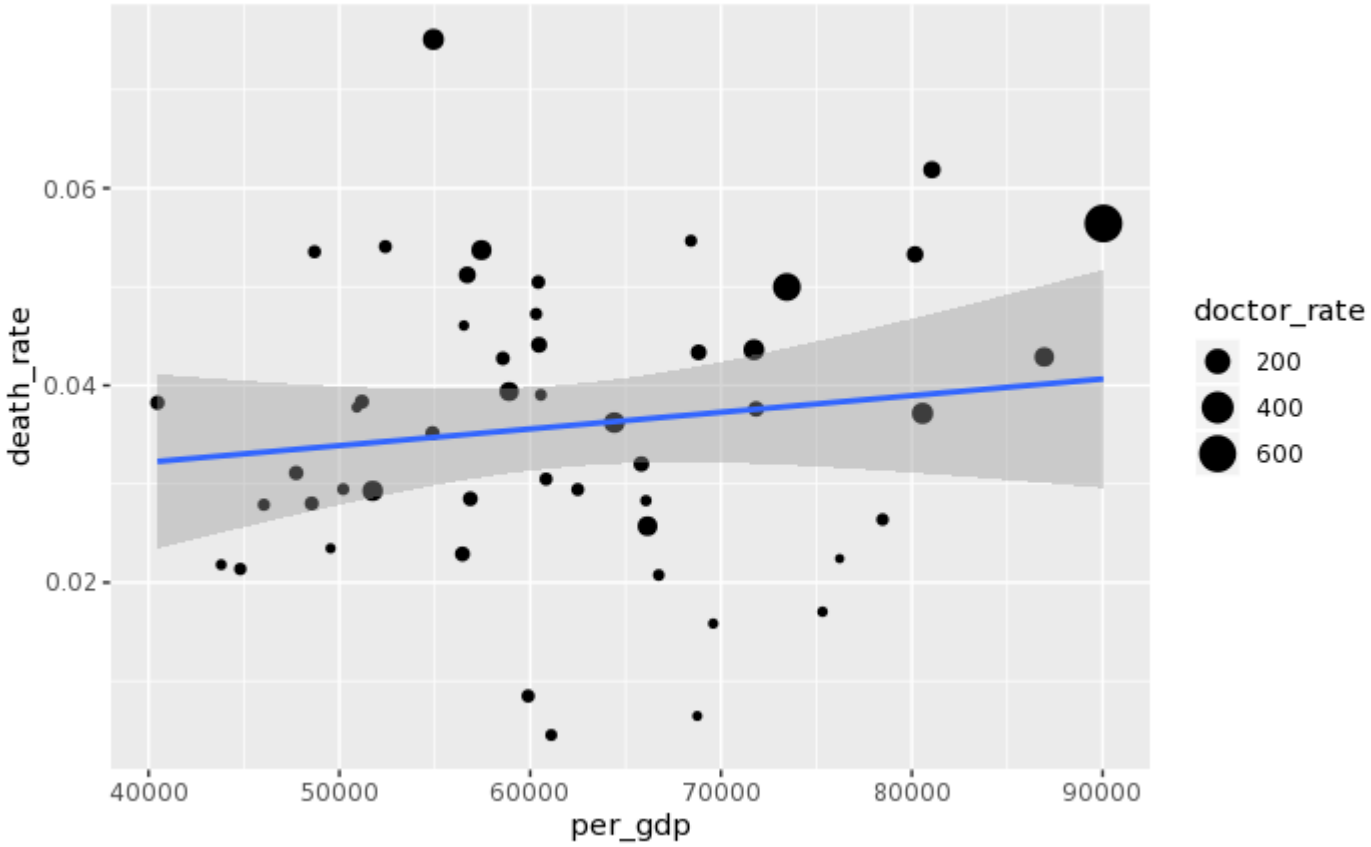
```
cor(complete_data$per_gdp ,complete_data$death_rate) #finding the correlation index
```

```
[1] 0.133436
```

Then I draw the point graph. The x axis is GDP per capita and y axis is the death rate. Then the regression line shows that there are weak relation between death rate and GDP per capita. And the dot size represents the doctor rate. Thus we can see the trend that high doctor rate (meaning a doctor needs to take care of more patients) tend to have higher death rate.

Hide

```
complete_data%>%
  ggplot(aes(x=per_gdp, y = death_rate ))+ #using ggplot to draw the points
  geom_point(aes( size = doctor_rate ))+
  geom_smooth(method='lm', formula= y~x) #drawing the regression line
```



Finally, although I used graph to explore the relationship between these variables and mortality, if I want to be able to clearly show the relationship between these variables and mortality, I think I need to use machine learning methods to define and clear the relationship between the death rate and these variables.

using the party package to get the mod for the relationship between the death rate and these variables.

Hide

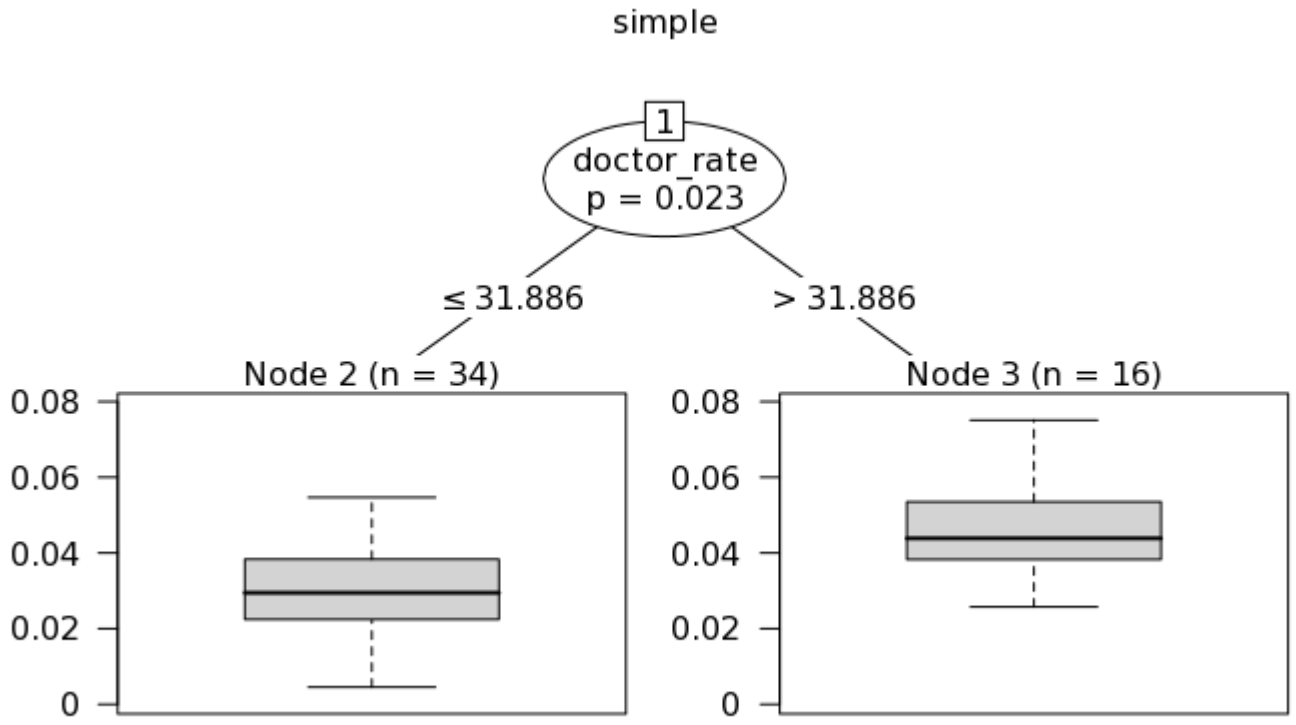
```
mod1 <- party::ctree(death_rate ~  per_gdp + doctor_rate, data = complete_data) #use ctree to dr
aw the mod between variables
mod1
```

```
     Conditional inference tree with 2 terminal nodes

Response:  death_rate
Inputs:  per_gdp, doctor_rate
Number of observations:  50

1) doctor_rate <= 31.88601; criterion = 0.977, statistic = 6.364
  2)*  weights = 34
1) doctor_rate > 31.88601
  3)*  weights = 16
```

using the plot function to draw the node graph of the mod1. We can find that the result of machine learning indicates that the doctor rate is more related to death rate. In node 2 which the doctor below than 31.886, the death rate is lower with the range from 0.05 to 0.5 and the mean of death rate is 0.3. In node 3 which the doctor greater than 31.886, the death rate is higher with the range from 0.25 to 0.75 and the mean of death rate is 0.45.

Hide

```
plot(mod1, "simple") #ploting the mod
```



Conclusion: Through the exploration in the first part, we have seen an increase in the number of cases and deaths in each state. In the second part, I introduced several variables, including death rate, doctor ratio, GDP per capita, and doctor status. Then I drew a graph to study the relationship between death rate and other variables. The conclusion was that there was a very weak positive relationship between gpd per capita and death rate; there was a relatively strong positive relationship between the proportion of doctors and mortality; There is a tendency to have a higher mortality rate for these states with good doctors.

Thinking about the project: Although I studied the relationship between these variables and mortality through drawing and machine learning. But the conclusion I got is very counter-intuitive, because in common sense, we think the state that has more wealthy and more doctors will have higher death rates. But in this series of data analysis, we see that doctors and states with high GDP per capita will have more mortality. However what is in line with our common sense is that states with doctor rate (meaning a doctor needs to take care of more patients) will have a higher death rate. But in the future improvement of the project, I will add more variables and draw more detailed images to explain the reasons behind the mortality.