Is there an association between people's income and deaths caused by Covid-19?

Wenbo Jiang

Introduction:

At the beginning of 2020, Covid-19 affects people's life globally. Every countries enacted several policy to deal with this disease. The increasing unemployment, decreasing GDP, higher inflation and so on are signs to reflect that the economic market is under tremendous risk. For this project, I am wondering whether the different in people's income would influence the death cases by Covid-19 in the US. Also, I would consider GDP level during Covid-19 period as the confounding variable in our analysis.

Method:

Variable Description:

- State: 51 States in the US
- State_full_name: Full name of each state
- Lon:Longitude
- Lat:Latitude
- Income: Median Household Income in United States
- Urban_rural_code:a classification scheme distinguishes counties by the population
- Covid_death: Death caused by Covid-19
- All-Causes death: All death during analysis
- total_covid_death_instate: total number of death caused by Covid-19 in each state
- total_all_death_instate: total number of death in each state
- death_mean_urban: Average number of death caused by Covid-19 in different type of counties.

For the first dateset, I choose to use Median Income for each state in the US provided by United State Census and the link is 'https://www.census.gov/search-results.html?q=Median+income+&page=1&stateGeo=none&searchtype=web&cssp=SERP&_charset_=UTF-8'. For the second dateset, I choose to use the collection of Covid-19 cases and all-causes death cases in each state and county in the US provided by the CDC and the link is 'https://data.cdc.gov/NCHS/Provisional-COVID-19-Death-Counts-in-the-United-St/kn79-hsxy'. For the third dataset, I found the GDP level across each state in the US on the website "https://worldpopulationreview.com/state-rankings/gdp-by-state".

I need to merge two datasets which contain our main effects variables: Income and death caused by Covid-19 by the variable 'State' to get a full dataset which is helpful for the further analysis. Then, I delete the comma occurred in some numerical number such as changing 14,500 to 14500 in order to better run the data in

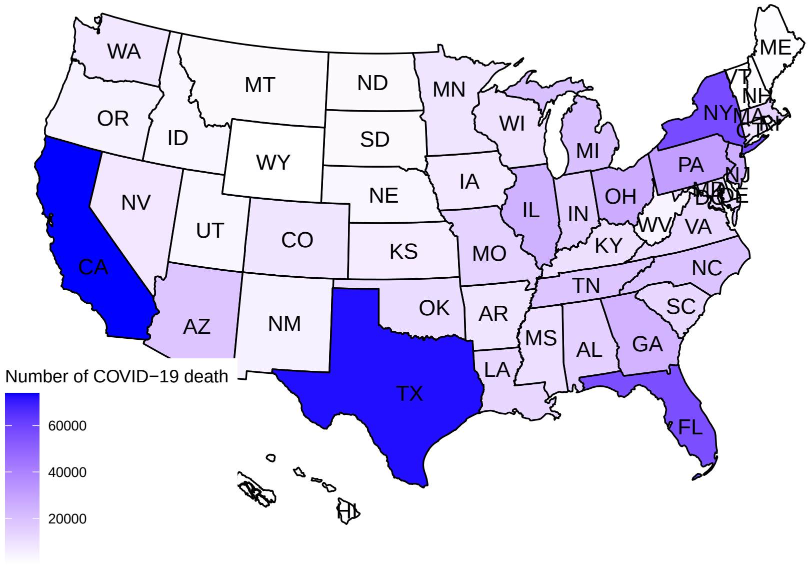
R. For the next step, I renamed certain variables that include ‘space’ like changing “urban rural code” to “urban_rual_code” as a whole word. Before providing some statistical result, the most important step is to check the missing value occurs in our data. For any observations with the missing value for the death cases, I just replaced them with 0. In order to better summary the key outcome by the variable ‘state’, I created new variables to reflect the total death cases in each state. For analyzing our confounding variable, we just combined our existing date ‘covid1’ with the GDP data and for a new dateset called ‘gdp_incme_covid’. For this combined data, we would measure the association between GDP level and Covid-19 deaths and the association between GDP level and Income. Since the GDP data we choose is distince enough, so we don’t need to clean this combined dataset anymore. Then, I created a table to show the details of each key variable. The table contains six variables which classified by State: the full name of the state, number of counties, GDP, Income, COVID-19 death cases and all-caused death cases. For the data visualization, I plotted 4 graphs to show the association between each key variables. For example, I used draw a US map to show the density of COVID-19 death in each state and draw a scatter plot to reflect the linear association between Income and number of Covid-19 death cases.

State	State_full_name	Number_of_County	GDP	Income	Covid_death	All_death
AK	Alaska	19	50413	77640	601	8363
AL	Alabama	67	228062	50536	14867	114311
AR	Arkansas	73	130709	47597	8307	68003
AZ	Arizona	15	378297	58945	18609	138229
CA	California	56	3120386	75235	73920	573696
CO	Colorado	57	394271	72331	8148	82474
CT	Connecticut	8	283601	78444	8666	61268
DC	District of Columbia	1	143389	86420	1587	12746
DE	Delaware	3	76468	68287	1975	19245
FL	Florida	67	1111614	55660	56495	450857
GA	Georgia	155	627667	58700	24129	185638
HI	Hawaii	4	89866	81275	866	21717
IA	Iowa	99	195353	60523	6836	60088
ID	Idaho	42	85552	55785	3169	29676
IL	Illinois	101	875671	65886	24747	217893
IN	Indiana	92	379293	56303	15866	135459
KS	Kansas	103	175465	59597	5899	54058
KY	Kentucky	119	213169	50589	10120	98291
LA	Louisiana	62	244577	49469	12717	98182
MA	Massachusetts	14	590307	81215	14323	116827
MD	Maryland	24	427616	84805	11323	103050
ME	Maine	15	67129	57918	1192	27590
MI	Michigan	82	524828	57144	20073	197070
MN	Minnesota	84	379388	71306	8403	88465
MO	Missouri	114	325841	55461	14532	132315
MS	Mississippi	81	115900	45081	10187	70288
MT	Montana	50	51934	54970	2111	20864
NC	North Carolina	99	594126	54602	19228	177232
ND	North Dakota	43	54044	64894	1856	14232
NE	Nebraska	79	129761	61439	3166	33065
NH	New Hampshire	10	86319	76768	1512	23853
NJ	New Jersey	21	625659	82545	25922	158180
NM	New Mexico	29	100777	49754	4807	39700
NV	Nevada	16	175509	60365	7686	56609
NY	New York	61	1705127	68486	57508	338892
OH	Ohio	88	683460	56602	26585	248104
OK	Oklahoma	75	186883	52919	10986	82606

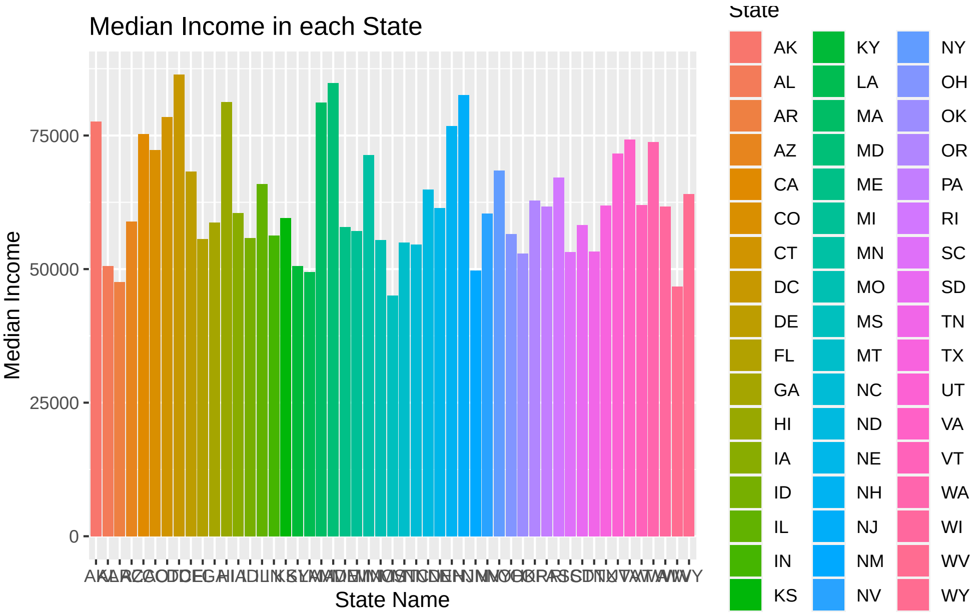
State	State_full_name	Number_of_County	GDP	Income	Covid_death	All_death
OR	Oregon	34	253849	62818	4307	72731
PA	Pennsylvania	67	788500	61744	31082	269248
RI	Rhode Island	5	61081	67167	2812	20327
SC	South Carolina	46	245473	53199	12924	107898
SD	South Dakota	57	55243	58275	2138	16492
TN	Tennessee	95	369063	53320	18262	159485
TX	Texas	245	1772132	61874	72436	457910
UT	Utah	25	198630	71621	3233	39219
VA	Virginia	131	557986	74222	13117	142750
VT	Vermont	13	33278	61973	283	10983
WA	Washington	38	632013	73775	7739	113328
WI	Wisconsin	71	344500	61747	9450	106051
WV	West Virginia	52	74511	46711	3752	41299
WY	Wyoming	23	36000	64049	951	9748

Preliminary Results:

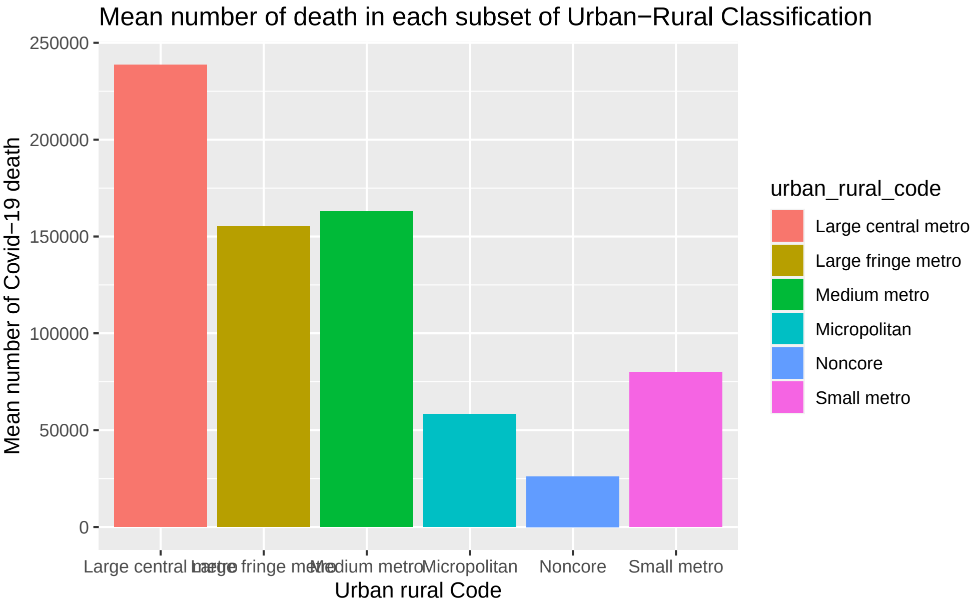
We checked the dimension of our data and noticed that there are 3023 total observations and 17 different factors for each of our observation. Then, I did some summaries for the key variables such as Income, GDP, Covid-19 death cases and all caused death cases. I found people living in Mississippi has the the lowest median income which is $45081 and people living in District of Coloumbia has the the highest median income which is $86420. Also, I noticed that the lowest death cases caused by COVID-19 is in Vermont which equals to 283 and highest death cases caused by COVID-19 in California which equals to 73920 and mean death cases caused by COVID-19 in the US is 20504. For the variable GDP, I found that the state Vermont also has the lowest GDP which equals to 33278 million dollar and the state California has the highest GDP which equals to 3120386 million dollar.



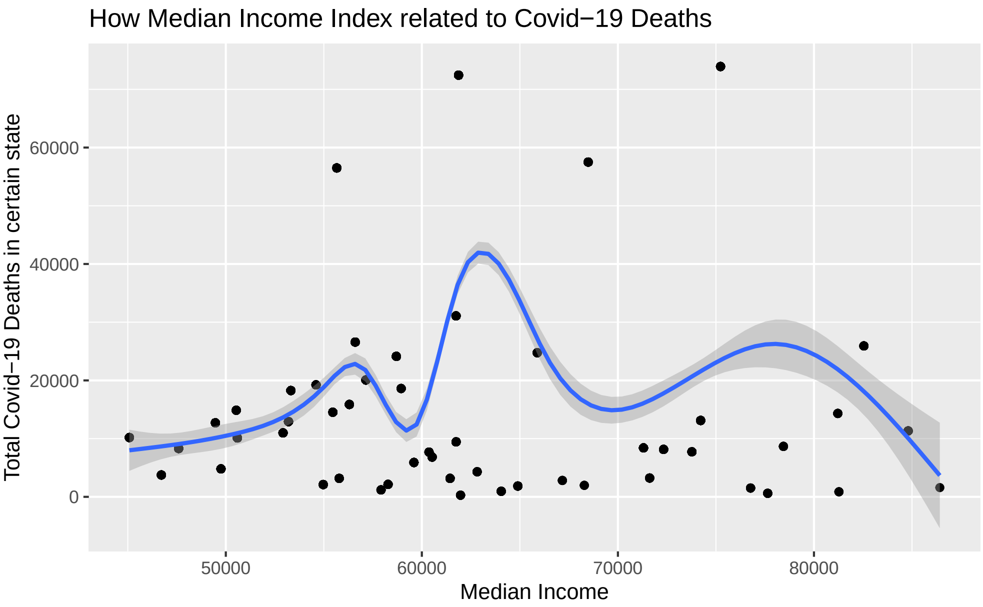
This graph provide the distribution of Covid-19 death cases visualized by US map. If the state contains more cases, the color of that state would more closely tend to blue. We noticed that California, Florida, New York and Texas contains much more COVID-19 death than other states. To be detail, during the period from 01/01/2020 to 10/20/2021 California has 73920 Covid-19 death cases, Texas has 72436 Covid-19 death cases, New York has 57508 Covid-19 death cases and Florida has 56496 Covid-19 death cases.



For the second plot, we measured the distribution of Income classified by state visualized by bar plot. We noticed that the range of Income between each state is relatively large which equals to 41339. The state Mississippi with the lowest median income which equals to 45081 and the state District of Columbia with the highest median income which equals to 86420.

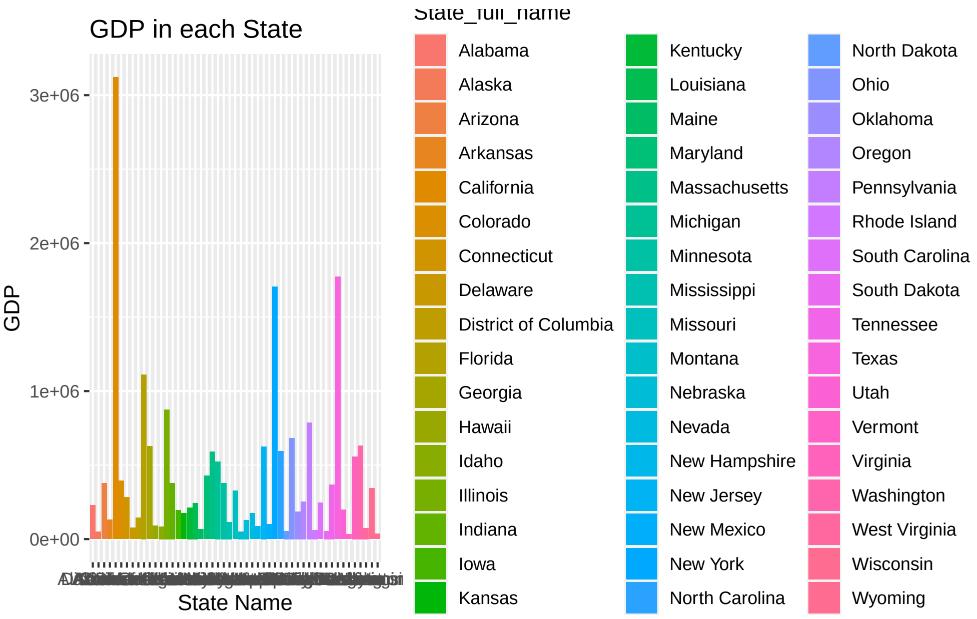


This graph is about the association between different urban-rural classification and COVID-19 death cases. We found that there is not a clear linear association. We cannot say that if the counties contains more population, It would be more COVID-19 death cases. It's clear to notice that as counties defined as 'Median metro' have more Covid-19 deaths than counties defined as 'Large fringe metro'. Also, there is a larger amount Covid-19 death cases in counties defined as 'small metro' than counties defined as 'micropolitan'.

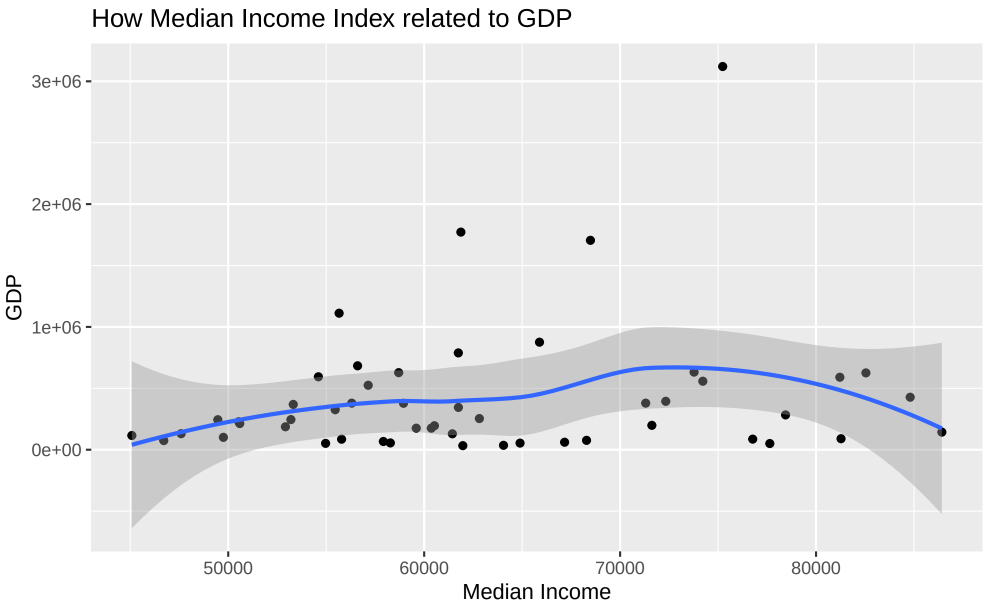


This graph is the reflection of the relationship between our two main variable: Median Income and death due to Covid-19. We used scatter plot with a smooth line to detect the association. However, the pattern is not clear and looks like a normal distribution since those 4 states which contain especially high value of Covid-19 death cases affect a lot to the overall association. For the next step, we would consider GDP as a confounding variable.

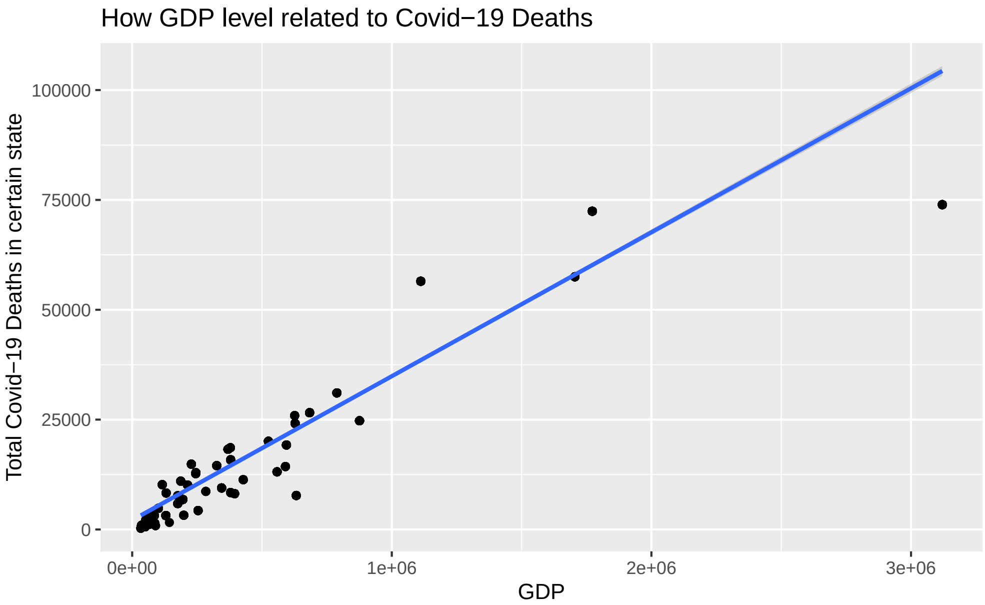
Find out whether GDP is a confounding variable and affect the association between Income and death caused by Covid-19



We use the bar chart to find out the GDP level in each State. From the graph, we noticed that Top 3 high GDP state is California, New York and Texas. California has the highest GDP which equals to 3120386 million dollar. GDP in Texas equals to 1772132 million dollar and in New York equals to 1705127 million dollar.



The scatter plot with a smooth line measures the association between Median income and GDP Level. It looks like a positive linear assciation, but the slope is very small.



This scatter plot with smooth line measures the association between GDP and Covid-19 death cases. We can easily find that there is a strong positive linear association between GDP and Covid-19 death cases. The variable GDP is associated with both Median Income and Covid-19 death cases, so we would say GDP is a confounding variable for our main analysis. This is a very important find since we would do furthur analysis after controlling the variable GDP.

Conclusion

We collect the information about the median Income and COVID-19 death for all 50 States in the US. Four of those state which are LA,TX,NY and FL have the higher COVID-19 death cases than other states. For the Median income for people living in CA,TX and NY are over $60,000 which is a relative large value, but for the linear association between income and COVID-19 deaths, there is not a clear pattern. Also, GDP is considered as confounding variable in our analysis and needed to be controlled. For the further analysis, I would introduce more variables like race, gender to show whether they confounded the association between income and COVID-19 deaths.