Link to Databricks notebook:
https://databricks-prod-cloudfront.cloud.databricks.com/public/4027ec902e239c93eaaa8714f173bcfc/2601820831055742/4472107455055949/749842193659695/latest.html

Effectiveness of Social Distancing on Covid-19 Transmission in the United States

Table of Contents:

Effectiveness of Social Distancing on Covid-19 Transmission in the United States

Summary:

Most of the World, including the United States, has been severely impacted by Covid-19 or coronavirus. Many countries have started to mandate some sort of social distancing to help stop the spread of Covid-19. Many citizens in the U.S. have started to become restless having to stay isolated in their homes and more seriously many citizens have lost their jobs due to businesses having to shut down due to government regulations. This has led to some states opening up more of their economies in late April and early May of 2020. Is this wise? This paper will try and figure out the effective social distancing has had in the United States up until April 28th. We will look at data on social distancing measures each state had in place, how seriously the citizens of each state took those regulations, and the number of cases and deaths of Covid-19 to see what correlations they had up until April 28th, 2020. Hopefully, this paper will help the public get a better understanding of the growth of Covid-19 and how social distancing may be making an impact. We will next move into the primary objectives of this project followed by the project plan and methodologies used and finish up with the results and conclusions based on those results.

Project Objectives:

This project was created to give the general public a better understanding of what social distancing measures are in place in each state, how well citizens in each state are abiding by those restrictions, and most importantly how well social distancing is working in each state. To get a more in-depth understanding of what this paper will try to answer, key questions were identified to address the concerns laid out above. The paper will try to address these questions:

- How well has the implementation of social distancing stopped the spread of Covid-19 on a state-by-state basis? More specifically, has the rate of infection and death gone down since implementing social distancing?
- When will it be safe to stop social distancing?
- How can we measure how well each state's citizens are abiding by the rules?
    - Which state's citizens are best distancing themselves and staying isolated?
    - Is there a correlation with how well a state's citizens are staying isolated and the spread of Covid-19?
- What are the social distancing measures being taken in each state and when will they end?
- Do the states with more severe social distancing measures have less cases? Or is Covid-19 spreading less rapidly in those states?

These questions will serve as a basis to begin the project planning and execution. The data found may or may not be able to perfectly address each of the questions and as a byproduct, some of the questions may not be answered completely with high certainty. The paper will only try to draw conclusions based on the data found and results that have reasonable certainty. To get more clear and significant conclusions, this project should be run again with updated data in two to three weeks when many of the states are off of social distancing and the impacts of leaving social distancing restrictions will be clearer. As most states got off restrictions in late April or will get off in early-to-mid May (April 27th – May 20th), it will be hard to conclude with certainty the effects of leaving social distancing based on data up until April 28th.

Project Plan & Methodologies:

To ensure organization and proper methodologies, this project was started with the Deployment and Life Cycle Management methodology which was outlined by Arnie Greenland of the University of Maryland's Robert H. Smith School of Business [1]. The major steps of using the Deployment and Life Cycle Management methodology that were used were (Appendix A1 – page 8 for Diagram):

1. Getting a Business Understanding – Defining the Questions in the Project Objectives section

2. Data Understanding & Preparation – This step included identifying data needs and sources for the key questions, acquiring the data, cleansing the data and joining the data to condense the data into one cohesive dataset.
3. Modeling – Displaying data outputs from simple queries on the data as well as more advance Machine Learning techniques like linear regression to predict values.
4. Evaluation & Deployment – How well did the above models perform? Does anything need to be redefined? Do the models address the key questions? Draw conclusions to the public.

Data Understanding & Preparation:
Since the business understanding was addressed in the project objectives section, we will start looking at the data understanding and preparation. When looking at the business questions it is clear a dataset was needed on the number of cases and deaths which was done with a Kaggle dataset [a]. After finding an adequate dataset on Covid-19 cases, the type of social distancing for each state was needed. There weren't any readily available datasets that were found with this type of information, so a dataset was created from references that can be seen in the references section ([2], [3], [4], [5], [6]). The dataset created had data on each state's current social distancing restriction rules, religious restrictions, end date of the restriction in place, and the current population. The social distancing restrictions were put into seven different categories.
1 - Stay At Home
2 - 10 or Fewer
3 - 20 or fewer
4 - safer at home
5 - Closed for Nonessential business, schools, dining
6 - Opening of some small businesses, schools & large gatherings still banned
7 - Social Distancing of 6 feet everywhere but no restrictions

With the types of social distancing restrictions for each state in place, the last piece of data needed to answer the questions was on how well the state's citizens were abiding to the rules. To do this mobility data was used that anonymously tracked people's cellphones to see how much they were moving. Two different datasets were used. The first was put together by Descartes Labs [b]. This mobility dataset included the state, county, and a m50 and m50_index measure. The m50 measure was the median max-distance of travel for each sample in a region (again could be a state or county region). The m50_index was the percent of travel each sample (sample being person or cell phone) had compared to a normal value which was set with data from February 17th up until March 7th. This m50_index would be a value like 55% which means a region as a whole traveled 55% of the normal 100% during a specific date. The second mobility dataset was put together on Kaggle based on Google's mobility dataset [c]. This dataset contained mobility data for countries around the world and drilled down into state and county locations in the United States. It included a mobility type which had retail, grocery, parks, transit, workplace, and residential types of mobility. It also included a mobility change which is similar to the m50 index described above which gives a percentage of the change in mobility compared to normal days before the spread of Covid-19.

With the datasets now created, the next step was to load them into a single place for analysis. The technology chosen was Databricks which has a free version for members to use up to 15 GB of data. The datasets were placed in an AWS S3 bucket which is a fancy way of saying amazon cloud storage. This storage can be free for a new user up to a certain GB threshold. Once in the amazon storage, it was then pulled in through Databricks and saved there. On Databricks, the data was next looked at to create a schema or design. This design helps structure the data for the analysis to come. Once each dataset had a schema created, the data was loading into the schemas and cleansed. What is meant by cleansing is that the data was looked at for any potential errors or unwanted data for the analysis. Errors might have included missing data which needs to be accounted for, wrong types in certain columns (types meaning, for example if there was an integer or number where there needed to be a state name which is called a string in programming), duplications of data which can throw off results, and more. The cleansing step also included filtering down to states and counties in some

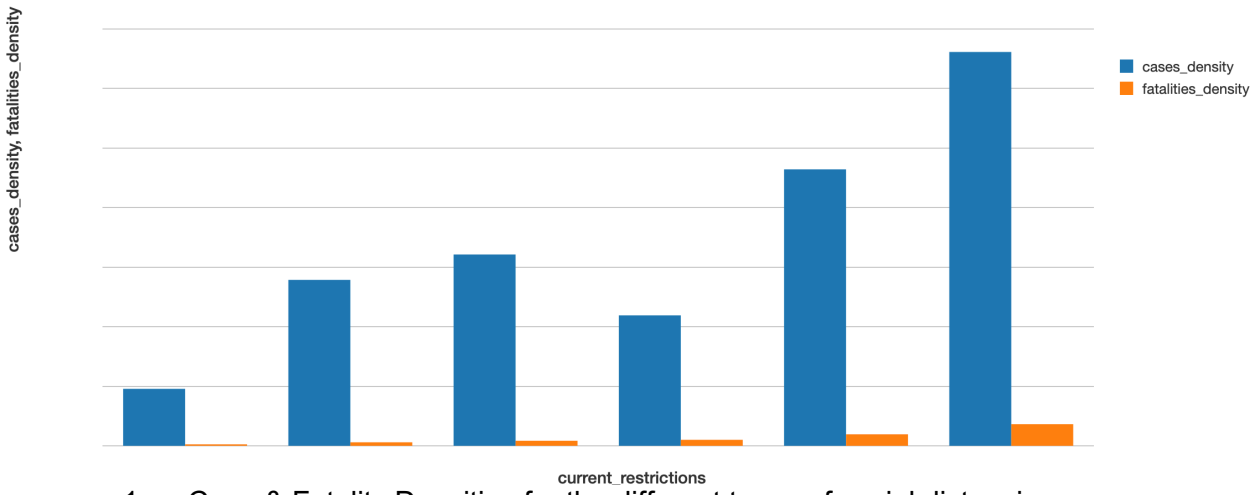
cases. The cases and deaths data included data from around the world when only the United States was being analyzed.

With the dataset schemas in place and the data cleansed for analysis, the last step here was to join the data together. What is meant by joining is allowing us to analyze the data together. You join on certain objects of the data. E.g. the cases and deaths dataset were cleansed to include the state, date, confirmed cases, and fatalities. The social distancing type dataset had the state, religious and social restrictions, end date of restrictions, and current population. To join the datasets, we join them on state and create a bigger dataset that would now contain state, date, confirmed cases, fatalities, religious and social restrictions, end date of restrictions and current population. A process like this was done with every combination of the datasets to give us the ability to analyze different trends between the different types of data – mobility, social distancing, and the cases and deaths of Covid-19. All of the datasets were also combined in order to answer any questions that involved all three of the types of data. The third and fourth steps (Modeling and Evaluation & Deployment) of the Deployment and Life Cycle Management methodology that were laid out will next be addressed in the Project results section.

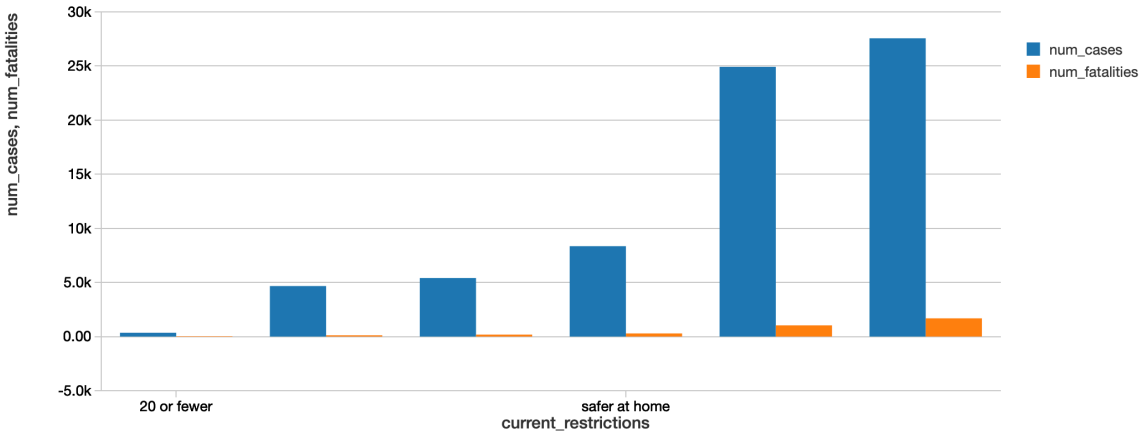
Project Results:

With the datasets joined the analysis could now begin. The key questions were kept in mind as the analysis began with the core concept of finding out how well social distancing had stopped the spread of Covid-19 as of April 28th. The first model created was to see how well each of the social distancing categories was doing in terms of cases and fatalities. The seven types of social distancing were outlined in the data understanding section. The output of the model (1a) below shows the Covid-19 cases and fatalities densities for the six different types of social distancing that appeared as well as the actual number of cases and deaths in the second output (1b). The bars from left to right are (the "10 or fewer" social distancing type had 0 states and thus was not in any of the model outputs):

1. 20 or fewer
2. closed nonessential businesses, schools, dining,
3. opening of some small businesses
4. safer at home
5. social distancing of 6 feet but no restrictions
6. stay at home



1a – Case & Fatality Densities for the different types of social distancing



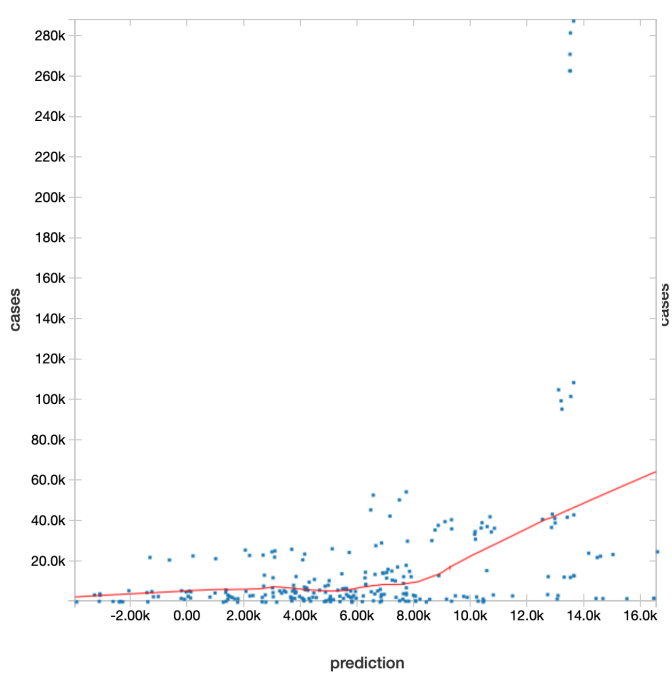
1b – Actual Number of Cases & Fatalities for the different types of social distancing

Based just on the outputs above, it appeared as though the stay at home and social distancing of 6 feet with no restrictions had the most cases and highest case density. The social distancing of 6 feet restriction was the most lenient restriction to date and shows that social distance may indeed be working as the states that had that type of restriction had high case density and number of cases. The stay at home order on the other hand was the most severe type of social distancing, yet it had the most cases and highest case density. Saying that the stay at home order wasn't working is probably presumptuous because the states that had the stay at home order were the ones who were hit the hardest initially and have high population densities. This makes it easier for the virus to spread even if everyone was at home. States like Washington, New York, Massachusetts, and New Jersey were all stay at home states who were hit early and continue to have high numbers of new cases. These states were thus making the stay at home order appear to not perform as well as the other types. Another deduction that was made when comparing the graphs was that the states that have the restriction on "opening of some of the small businesses" have a much higher case density and fatality that expected. When looking at the number of cases, this restriction was the third lowest and was less than a fifth of the values of the stay at home order and 6 feet restriction states. But when looking at the density, the opening of small businesses group was nearly half the stay at home order and 6 feet restriction states. This means that the states that opened small businesses as of April 28th were seeing a much higher case density than they should have. Most of the states in this category are very sparsely populated states or southern states. This conclusion shows us that opening up restrictions too soon or potentially never having strict restrictions has allowed Covid-19 to spread more than it should in those less densely populated states. This trend can also be seen in the appendix diagrams (appendix B1 & appendix B2 on page 9) that shows every states case density and actual number of cases of Covid-19.

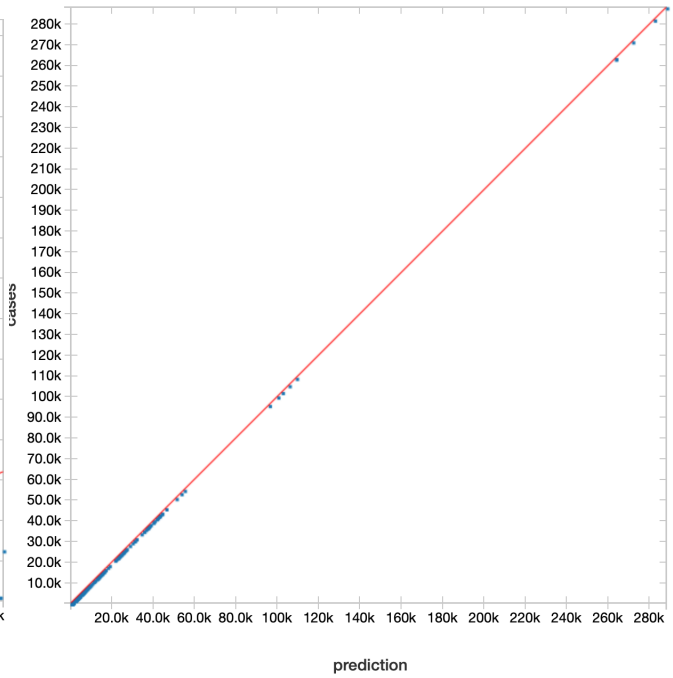
The next question that needed to be answered was how well each state's citizens are abiding by the restrictions in place? To do this the mobility data was used along with the cases and deaths and social distancing restrictions for each state. Several models were run with the two different types of mobility data described in the data understanding section. The outputs were a little too large to fit in the summary of results, but you can see them in the appendix C1, C2, C3 on pages 10-11. The most interesting output came from the m50_index which is a percentage value that shows us the percentage everyone in the state was moving compared to their usually movements. The m50_index itself showed us different restriction levels at play. The states that had the strictest social distancing measures indeed had citizens moving the least. New Yorker's were moving only 3.85% of their normal average. Meanwhile people in Wyoming were moving about 70% of their normal values. This makes sense as many people in Wyoming are on farms/ranches and there also was not strict restrictions on movements of individuals. The m50_index seem to be a great indicator of how well each state is abiding by the rules, and based on the results, citizens are generally following the rules of their respective states. Even in the states where there were not strict rules, it was encouraging to see that the mobility values were all much less than normal. Visit Appendix C1 (p.10) for table output of m50 index by state.

With the questions on the spread of social distancing and how well each state is abiding by their respective restrictions addressed, the final part of the analysis involved working with trying to predict when it will be safe to end social distancing and how well we can predict the number of cases based on the data. To accomplish this, a Machine Learning algorithm was used to attempt to predict the number of cases based on social distancing measures, current cases, and mobility data. If the model could predict this number accurately, we could hopefully see how long it will be until the number of cases and deaths starts to decrease and when it will be safe to start opening up parts of the economy again.

The model used is called a linear regression. The general purpose of a linear or multilinear regression is to use variables (in this instance the data on social distancing, current cases, and mobility) to try and predict a value for an outcome variable, in this instance the number of new Covid-19 cases. The data had to be slightly altered for the linear regression which included taking out the mobility data from the Google dataset as well as reassigning current restriction names to numbers. E.g. Stay at Home would now be represented as a 0, safer at home a 1, and so on. This is done because a linear regression can only predict off of integers or decimal values rather than strings. Once the data was transformed, a training and testing set was made. The training set is used to train our model to make predictions. The testing set was then used to test whether or not the model performed well on new data it hadn't seen before. With the training and testing sets created, the model was ready to be produced and give us some predictions. Two different types of models were created. The first had data on social distancing types and mobility data. From there the model would have to try and predict the number of cases just based on social distancing types and mobility data without any previous knowledge on the number of cases. The second model had the social distancing types, and mobility data along with the previous number of cases. It would then try to predict the future number of cases. The outputs of the predictions vs actual number of cases are below for each model.



Model 1



Model 2

The first model shows a huge underestimation on the number of cases. The model continued to predict the total number of cases to be much lower than they actually were. It also failed to try and create a best fit line which is the red line in the models above. This model actually had a negative R-squared value which means the model fits worse than a horizontal line. As such, there really was not any statistical significance or conclusions we could draw from this model other than that social distancing types and mobility data alone cannot predict the future number of Covid-19 cases.

Model 2 on the other hand had a R-squared value of 0.999998 which means the model was very accurate and carried statistical significance. This model was able to correctly predict the number of cases of a state based on the previous number of cases, social distancing measure, and mobility of the state's citizens. The prediction shows us that the number of cases in the near future will continue to climb based on its projections. This tells us that if we keep the current course, the cases will continue to grow based on the data. This model could possibly show a decline in cases if it were run again in a couple weeks for the states that have continued to practice social distancing. Social distancing's impact is said to not be seen immediately, but rather over a period of time a state's cases will decrease as less individuals will be going out and being exposed. The model would also perhaps show us that the number of cases in states that are opening up will not be decreasing. The only way to know for certain is to run the model again in mid-May and see how well the initial prediction was holding up.

To view a fitted vs residual plot for each of the models that shows how far off the predictions were see the appendix D1 on pages 12-13.

Based on model 2's predictions alone, it appears that Covid-19 will continue to spread at its current rate and it is unclear how much social distancing is at play to stop it. This means the study has been inconclusive on how well each method of social distancing is stopping the spread of Covid-19. What is clear from the models is that Covid-19 is not going away anytime soon and may continue to spread at the same rate. This means people need to continue to try and distance themselves to not capture and spread the disease. As for the question on when to end social distancing, the results are again inconclusive. The model shows that it could keep spreading at the same rate as of April 28th, so staying distant from other people will continue to be key until at least mid-May. The model could then be run again to see how states that have opened up as well as states that have stayed with strict social distancing rules have fared. To learn more about the actual work performed and the Databricks notebook with all the calculations, queries, models, and outputs, visit the following link: https://databricks-prod-cloudfront.cloud.databricks.com/public/4027ec902e239c93eaaa8714f173bcfc/2601820831055742/4472107455055949/749842193659695/latest.html

References:

[1] Greenland, Arnie. "Chapter 8." *INFORMS Analytics Body of Knowledge*, by James J. Cockran, Wiley, 2019, pp.275-310.
[2] https://www.latimes.com/politics/story/2020-04-22/states-without-coronavirus-stay-at-home-order
[3] https://www.msn.com/en-us/news/us/every-states-rules-for-covid-19-social-distancing/ssBB12sr0O#image=5
[4] https://www.pewresearch.org/fact-tank/2020/04/27/most-states-have-religious-exemptions-to-covid-19-social-distancing-rules/
[5] https://wallethub.com/edu/states-where-social-distancing-is-most-difficult/73336/
[6] https://www.huschblackwell.com/newsandinsights/50-state-update-on-expirations-of-shelter-in-place
[7] http://www.healthdata.org/sites/default/files/files/Projects/COVID/Estimation_update_041720.pdf

Data Sources (original source – not Amazon S3 bucket storage link):

[a] https://www.kaggle.com/c/covid19-global-forecasting-week-4/data - Covid-19 cases and deaths each day for each state
[b] https://github.com/descarteslabs/DL-COVID-19 - Mobility Data m50 and m50_index with county and state levels
[c] https://www.kaggle.com/lanheken0/community-mobility-data-for-covid19 - mobility based on Google's mobility dataset - https://www.google.com/covid19/mobility/data_documentation.html?hl=en
[d] Databricks notebook: https://databricks-prod-cloudfront.cloud.databricks.com/public/4027ec902e239c93eaaa8714f173bcfc/2601820831055742/4472107455055949/749842193659695/latest.html

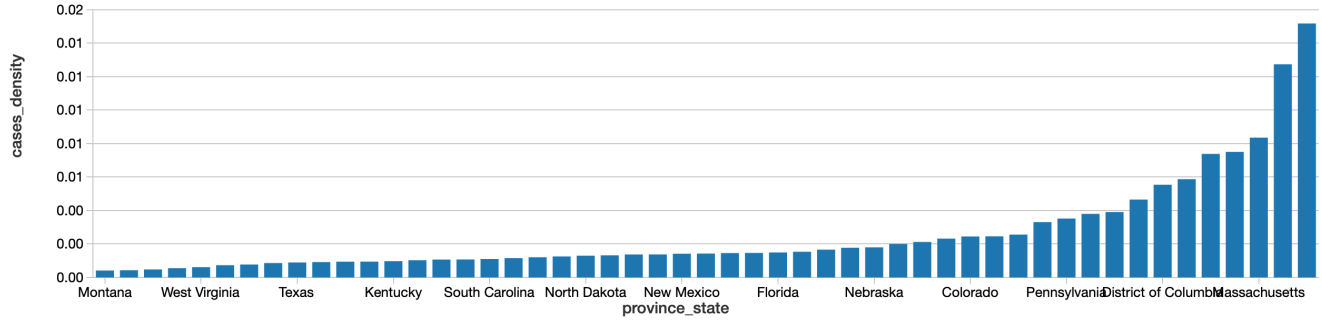
Appendix:

A1: Deployment and Life Cycle Management methodology

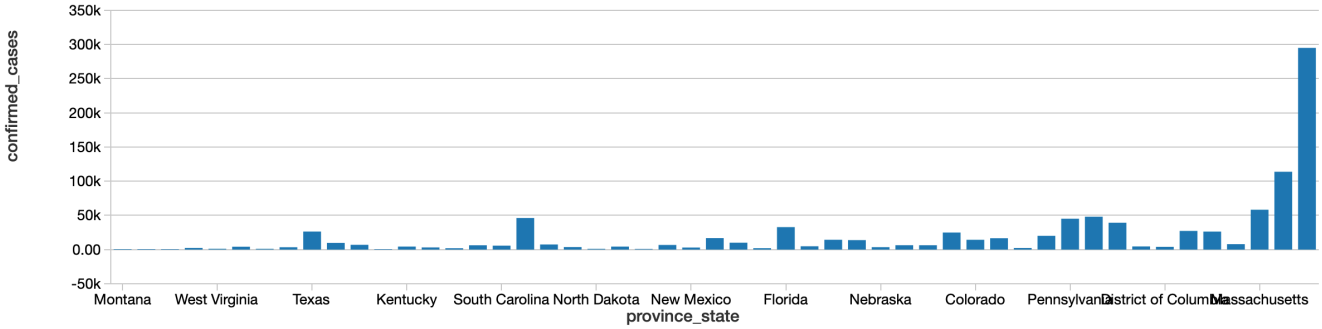
Figure 8.1 CRISP-DM diagram.

B1 & B2:

Case Density per state



Cases per state below



The Case Density per state and Cases per state diagram again illustrate the trend mentioned in the first part of the project results section which is that some of the states who have a lower number of cases still have a disproportionate case density. New York has by far the most cases, but in terms of the cases per population, the other states are a lot closer. This illustrates that New Yorkers have probably

done a better job of staying isolated than the other states, but yet still has the most cases due to the initial output and the amount of people that live so closely together.

C1: State by state output in order of lowest m50 index to highest

state	state_avg_m50	state_avg_m50_index	state_avg_cases	state_avg_fatalities	current_restrictions
New Jersey	0.4178879071984746	3.857771421508908	0.0055162113041215466	0.00017778911159923907	Stay At Home
Massachusetts	0.40083281969560713	8.106309185992151	0.0023444809939716415	0.000060380060823515286	Stay At Home
New York	0.6379710015099147	8.40318430411154	0.007441389182867524	0.00031444304271668503	Stay At Home
Connecticut	0.9222408854166667	9.100260416666668	0.002520096218801895	0.00009890101366973923	Stay At Home
Alaska	0.567521739130435	12.130434782608695	0.0003086711723306527	0.00000924058612326965 8	20 or fewer
Pennsylvania	1.0212647899986989	13.313785567467198	0.0012185101908124643	0.000023251266915610845	Stay At Home
Delaware	1.3936215277777773	13.809027777777779	0.0011358417678841019	0.0000229127560251434	Stay At Home
New Hampshire	1.3795074712643676	15.780128465179178	0.0005330190683378582	0.000010733770077214099	Stay At Home
Michigan	1.3473803899932577	16.236628093747726	0.001624222141362445	0.00008085050161583273	Stay At Home
Rhode Island	1.388219837914024	17.41902748414376	0.0012019048932840463	0.000029150210063210825	Stay At Home
Hawaii	1.2747598039215682	18.09558823529412	0.00030126672662471874	0.000004117101998240399	Stay At Home
Maryland	1.918611084589543	18.610730004930634	0.0007479169170842292	0.000017623895950628338	Stay At Home
Washington	0.936361748135051	19.311114639257795	0.0010843171895022047	0.000050662984674695396	Stay At Home
Vermont	1.5477682051076023	19.818152308927672	0.0006955340075775972	0.00002680385179042114	safer at home
Maine	1.4406890715131198	19.956579290053426	0.0003294420261843161	0.00000849930952894458	Stay At Home
California	1.334351250451128	20.7036211687486	0.0004560290065647329	0.000012052796020460411	Stay At Home
Arizona	1.2404456582633052	28.528384687208213	0.0003912834940524808	0.00001097314250122621	Closed Nonessential business, schools, dining
Wisconsin	2.09888939787056	29.517866379906117	0.000364001576766403	0.000013395683811379526	Stay At Home
Ohio	3.3203647630557924	30.00412168851408	0.0003868261355204713 6	0.00001425685634894108 3	Stay At Home
Oregon	1.389421382639147	30.175640910414348	0.0002401312495075584	0.00000726005471032278 2	Stay At Home
Florida	4.259133784406322	32.59744528379094	0.0006198241855229365	0.000012934107902869682	Stay At Home
Nevada	1.9682485401027152	33.3959170138518	0.0005843480264314788	0.00001856386840790916	Stay At Home
Minnesota	2.2730301854505948	34.00579344506213	0.0001547078902335966	0.000005635486328796552	Stay At Home
Colorado	2.8429856148463295	36.742997481879044	0.0007082831769422582	0.00002375334142942407	Opening of some small businesses, schools & large gatherings still banned
Indiana	4.629589913484754	37.46692786827741	0.0006437344990674049	0.00002362020874309245 7	Stay At Home
Utah	3.443898328216844	40.03282465909903	0.0005217091185143866	0.00000391689164525177 8	Closed Nonessential business, schools, dining

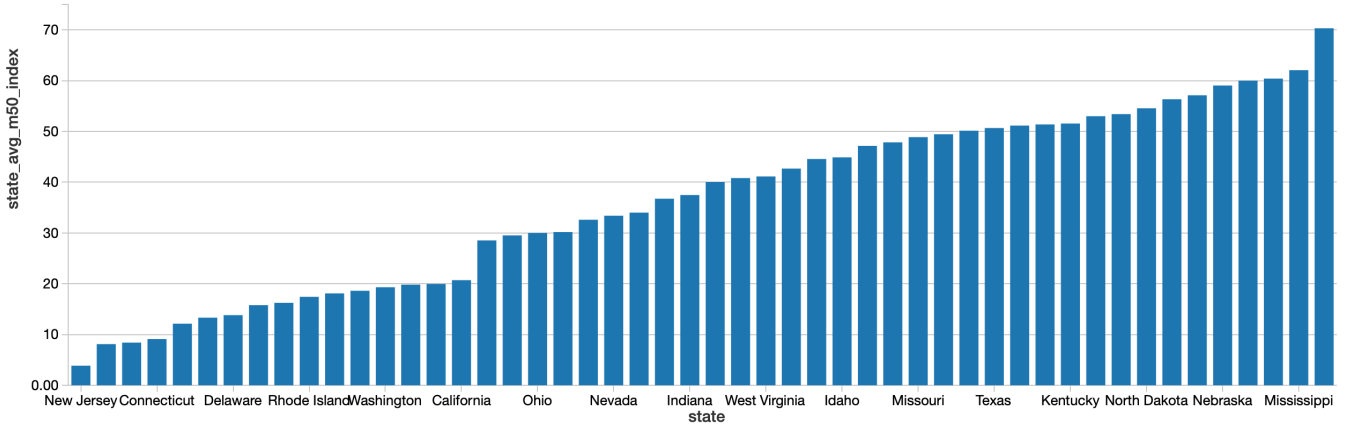
state	state_avg_m50	state_avg_m50_index	state_avg_cases	state_avg_fatalities	current_restrictions
Illinois	4.833412791476085	40.8111604517089	0.0008891504083282727	0.00002650236296156276 7	Stay At Home
West Virginia	3.536325745772316	41.12725664869788	0.00015590190391241617	0.00000141164179295378 45	Stay At Home
Virginia	6.485273753850709	42.67489846327046	0.0002949798833610746	0.00000693847807109373 7	Stay At Home
North Carolina	5.502686393466434	44.56631475418557	0.00028110245261806876	0.00000520584208142941 9	Stay At Home
Idaho	2.852892071603304	44.88584911174361	0.0004646499954794314	0.00000644350508108320 4	Stay At Home
Kansas	3.755799252238916	47.1500089616207	0.000260058599423158	0.00000902610465909265 6	Stay At Home
South Carolina	6.200695210683813	47.84317274094742	0.0004184235851226956 5	0.00000950456338092727 3	Opening of some small businesses, schools & large gatherings still banned
Missouri	5.566102888813599	48.87308879247993	0.0003803571366267570 6	0.00000901158584538912 3	Stay At Home
Louisiana	7.435887088001887	49.43800140046356	0.002436133304992655	0.00000920889393942638 6	Stay At Home
Montana	2.0983824688768786	50.139075911436166	0.0002085653517428842 4	0.00000369783544210572 54	Opening of some small businesses, schools & large gatherings still banned
Texas	9.463011457302148	50.66443059896104	0.0002372446306868398	0.00000448816465064197 5	safer at home
Tennessee	6.9178092920921275	51.14522430291961	0.0004564490794630329 4	0.00000827460726957329 9	Opening of some small businesses, schools & large gatherings still banned
Iowa	5.308644643104531	51.37131974627677	0.0002447567944408161 4	0.00000556804918258218 3	Closed Nonessential business, schools, dining
Kentucky	6.3661989133817105	51.55161081243778	0.0001815722439394891 5	0.00000936366281445538	safer at home
Oklahoma	6.734696153584004	52.99694001156632	0.0002814681221152867 4	0.00001307233713581509 5	safer at home
Georgia	7.550630581191345	53.406692315758924	0.000643722013691468 1	0.00002271721726235554	Social Distancing of 6 feet everywhere but no restrictions
North Dakota	2.736581114730866	54.55393262410174	0.0002380810081743517 8	0.00000425968383201145 6	Closed Nonessential business, schools, dining
New Mexico	43.411630663338364	56.33034667696973	0.0003234138113593354 5	0.00000519461447836624 05	Stay At Home
Alabama	8.508583916173482	57.116584065002954	0.0003722529408075500 3	0.00000998458983053369 2	safer at home
Nebraska	4.615916846215979	59.038250261037966	0.0001945521660378311 5	0.00000444750439446531 2	Opening of some small businesses, schools & large gatherings still banned
Arkansas	7.8978931722812336	59.99974690141612	0.0002302503547197616 3	0.00000416245796298400 3	Stay At Home
South Dakota	2.679021078628468	60.400613101512306	0.0003289101920371235 3	0.00000448477928400628 6	Opening of some small businesses, schools & large gatherings still banned
Mississippi	9.524568849226341	62.05948216308812	0.000411932312585195 2	0.00001277663262718205 2	Opening of some small businesses, schools & large gatherings still banned
Wyoming	3.3787026819866566	70.30897956754328	0.0002929894891622619 5	0	Opening of some small businesses, schools & large gatherings still banned

The above output in C1 shows that states that have the stay at home restriction are indeed moving much less than the other states. Near the bottom are states that never even enforced a strict social distancing restriction or states that have recently started to open up more of their economies.

C2:
Statewide m50_index for each state. Just a visualization of C1.

```
1 display(statewide_m50_ordered_df)
```

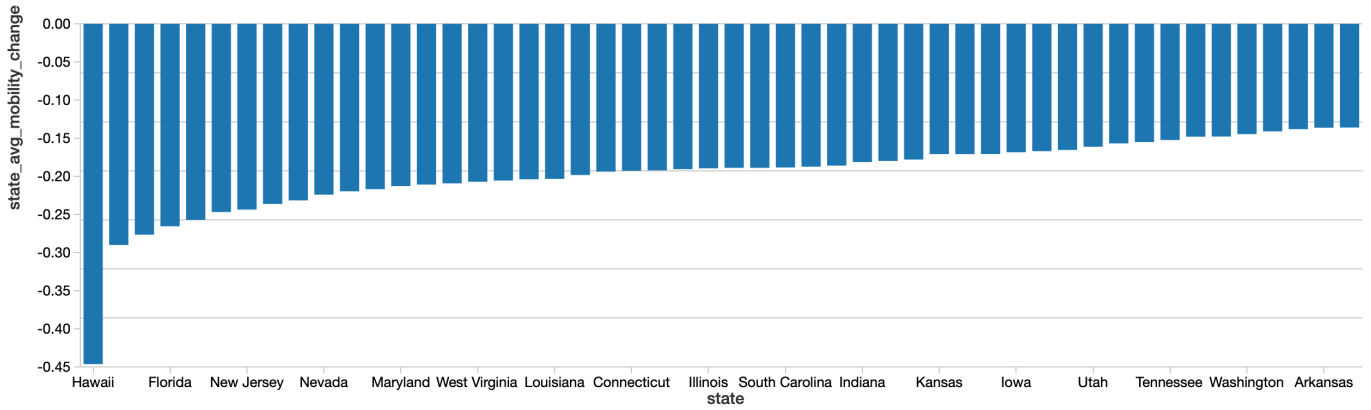
(3) Spark Jobs



C3:
Statewide mobility change based on Google's mobility data. This chart shows the total change in mobility as a percentage of each state.

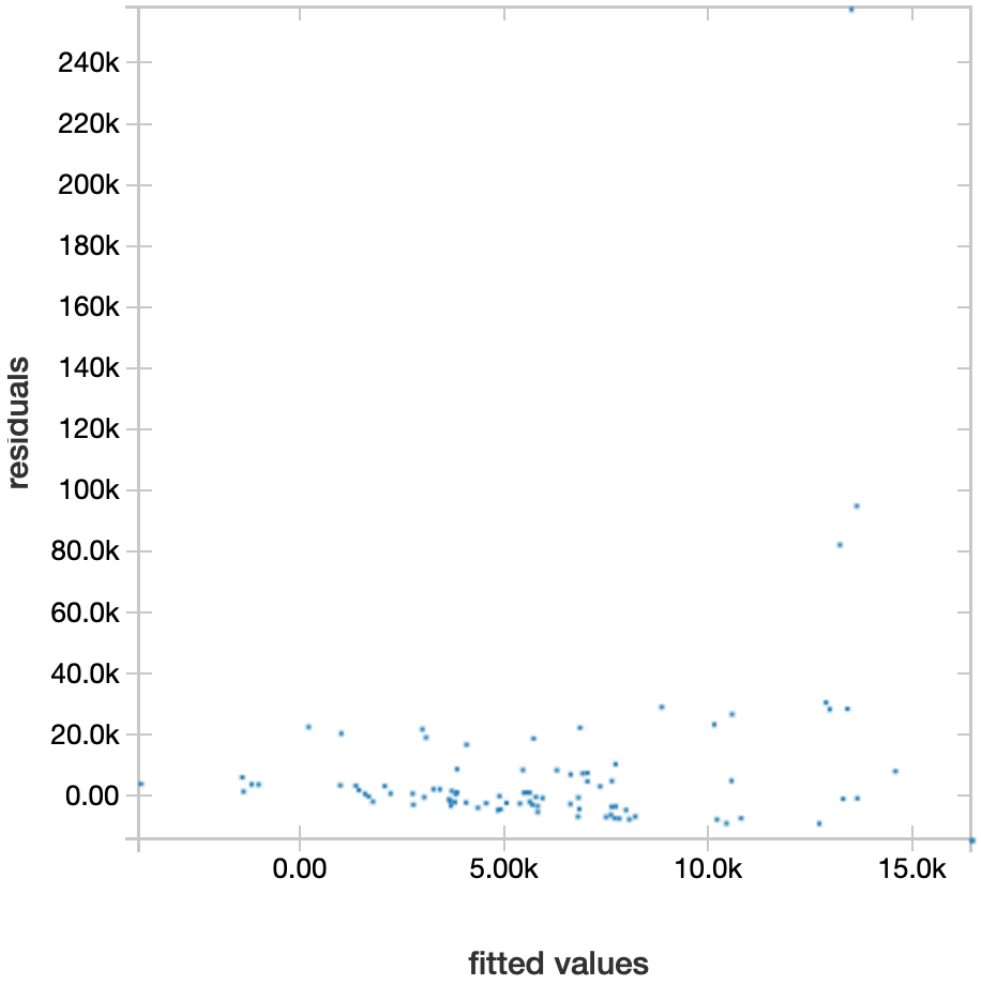
```
1 display(statewide_mobility_ordered_df)
```

(3) Spark Jobs

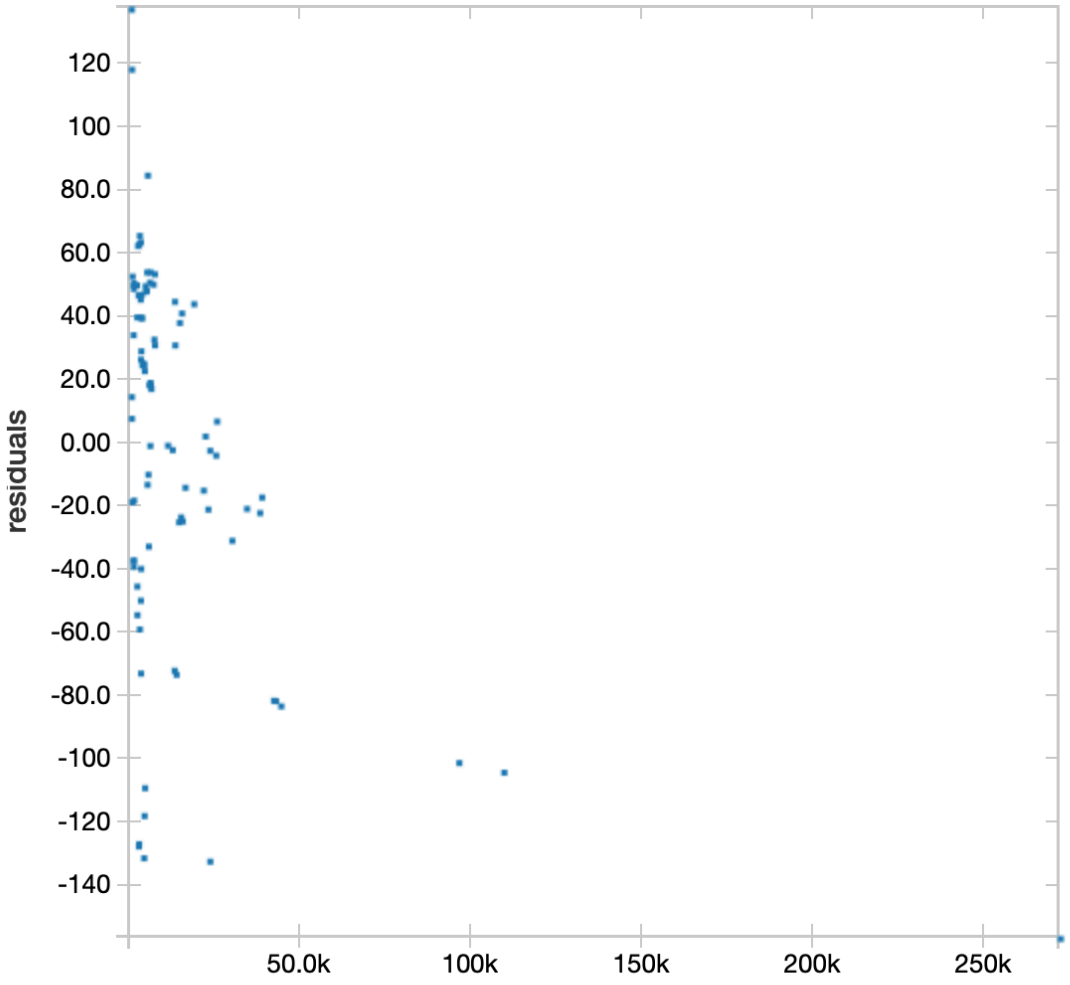


D1: Linear Regression fitted values vs residuals graph.

Residuals are difference between actual value and predicted value. The 2nd model has much lower residuals showing smaller error, it both over and underestimates at points. The first model underestimates the actual values as can be seen by the large positive residual values for the corresponding fitted values. The range for the second model is -140 to 130 which means the models predictions were all very accurate given some of the values approached 300,000 cases like in New York. The range for the first model was -10,000 to 250,000. This means the model overestimated by 10,000 on a data point and underestimated by 250,000 on another. Overall, from the plot you can see that the first model underestimated a lot more than it overestimated and by larger margins. The first models plot is below, and the second model is on the next page.



First Model Fitted vs Residuals output



Second Model Fitted vs Residuals output