Sar Velick, Henry Basu, Quinn Frankovsky, and North Carpenter
Final Project Report
Stat 253-01
May 4, 2022

Opportunity Insights Economic Tracker: Drawing Conclusions From Pandemic-Era Economic, Health, and GPS Movement Data

The index for R code can be found in the accompanying PDFs, organized by section.

1. Data Context

For this project, we used a dataset called the "Opportunity Insights Economic Tracker." This data was aggregated, anonymized, and cleaned by the Harvard-based Opportunity Insights team. They are a non-profit non-partisan organization that works to apply "big data" research to policy issues. Their team is led by several economists and their director is Harvard economics professor Raj Chetty. The purpose of this public data set was to facilitate research on the impact of COVID-19. It is the data behind their "track the recovery" project. This data set uses data from a large variety of organizations and contains more variables than we ended up using.

Each case in our final clean data set represents a day in one state. We have 50 states and the District of Columbia included in the data set. There is data for 483 days in each state. We used fewer days than the full time span we could have used. Based on some earlier experimentation, the data values before 2020-04-13 were so volatile that they dramatically impacted all our models' performances. This period was during the early days of the pandemic when the change in several variables was more than ten times the normal level. All the variables we did include are specific to that day in that state although some are moving averages from that specific day.

The data we used falls into three broad categories. We have a number of spending variables based on credit card data including change in total spending and change in spending on online services. We used a set of employment variables broken down by income quartile in each state. We have a number of relative COVID-19 related variables involving cases, and the final variable we used was relative time spent outside the home; this was a movement variable.

These are the variables we used with a brief description:

- `fullvaccine_rate`, Vaccine series completed per 100 people.
- `case_rate`, Confirmed COVID-19 cases per 100,000 people, seven day moving average.
- `hospitalized_rate`, New patients currently hospitalized in an inpatient bed who have suspected or confirmed COVID-19 per 100,000 people, seven day moving average
- `emp_incq1`, the employment level for the bottom quartile of the income distribution (<27,000$)
- `emp_incq2`, the employment level for the second quartile of the income distribution (>27,000$, <37,000$)
- `emp_incq3`, the employment level for the third quartile of the income distribution (>27,000$, <60,000$)
- `emp_incq4`, the employment level for the top quartile of the income distribution (>60,000$)
- `spend_remoteservices`, is a measure of the change in statewide spending in remote services January 4-31 2020 seasonally averaged and as a 7-day moving average.
- `gps_away_from_home` was our outcome variable. It is statewide time spent outside residence relative to Jan 3-Feb 6 2020 adjusted for seasonal variation.

Our data was collected and provided by a variety of firms and organizations. The spending variables we used came from Affinity Solutions, a large marketing agency that collects consumer data and uses it to target messages to consumers. In our original data set, we had data from Jan 3, 2020, to August 10, 2021. The data was handed to Opportunity Insite voluntarily to aid in their track of the recovery project. Presumably, Addfinity Solutions purchased this data from credit card companies or has access to some other method of tracking credit card purchases.

The unemployment data was collected from Jan 4-31 2020 to August 10, 2021. The data was provided by Paychex, Intuit, Earnin, and Kronos. Paychex, Intuit, and Kronos are all, at least in part, payroll service companies. Earnin is an online payday loan service. These for-profit companies provided the raw daily average data to opportunity Insite over the course of their track of the recovery project.

All of the COVID-19 data was collected from the Center for Disease Control. The CDC aggregated national and state-level data publicly on the COVID-19 pandemic. It covered the whole period from January 2020 to August 10, 2021.

The mobility data was collected from Google's public Google COVID-19 Community Mobility Reports. It was collected from Jan 3 to August 10, 2021, as well. Google indexed their relative variables from Jan 3-Feb 6 2020. They made this information available at a daily level throughout the pandemic as an aid to research and policymaking.

2. Research Questions

Regression Task

Our investigation into our `gps_away_from_home` variable led us to regression analysis. We sought to understand the relationship between the outcome variable and predictive variables such as COVID case rates, hospitalization rates, income, and expenditure on remote services. Our outcome variable is measured by time spent outside the place of residence.

Classification Task

Instead of designating the `gps_away_from_home` as being the outcome variable, we chose to explore a different side of the data for classification purposes. We classified each of the data points into its corresponding region in the United States, since each of the data points reflects a particular state. Hence, we used `gps_away_from_home`, case rates, hospitalization rates, expenditure on remote services, expenditure on health care services, and income to predict the specific region that that particular data set is in. We set the `Region` variable containing 4 possible categories (West, South, Midwest, and Northeast.)

We also categorized COVID restriction policy levels in each of the states. Based on the categories described above, we wanted to analyze the probability of a data point being in a state with extensive COVID policies and we did such using a hard threshold.

Unsupervised Learning Task

For the unsupervised learning task, our goal was to see if the clusters created based on a number of public health and economic variables aligned with the state regions (i.e. South, Midwest, Northeast, etc.) If the clusters did align with regions, this means that the variables are likely distinct among these regions.

3. Regression: Methods

The first model we used was Ordinary Least Squares (OLS). For this model, we used cross-validated data with 6 folds. Our outcome variable was `gps_away_from_home` and our predictor variables were `case_rate, hospitalized_rate, emp_incq1, emp_incq2, emp_incq3, emp_incq4`, and `spend_remoteservices.` Originally we also included `fullvaccine_rate` as a predictor, but there were so many N/A data points that we ultimately decided to remove it. Thus the recipe was

```
gps_away_from_home ~  case_rate + hospitalized_rate + emp_incq1 + emp_incq2 + emp_incq3 + emp_incq4 + spend_remoteservices
```

We created a workflow for OLS using the above recipe and `lm_spec`, which was set to use `'lm'` in `'regression'` mode. This workflow was fit to the `'minnesota'` data and stored as `'mn_mod.'`

The second model we used was LASSO, which was very similar to the OLS model. For LASSO, we again used 6-fold cross-validation with the same predictor and outcome variables. The outcome variable was `gps_away_from_home` and the predictor variables were `case_rate, hospitalized_rate, emp_incq1, emp_incq2, emp_incq3, emp_incq4`, and `spend_remoteservices.`We aimed to make the variables the same for each model to allow for easier comparisons. Thus the recipe was the same as for OLS:

```
gps_away_from_home ~  case_rate + hospitalized_rate + emp_incq1 + emp_incq2 + emp_incq3 + emp_incq4 + spend_remoteservices
```

We created a workflow using the above recipe and `lm_lasso_spec`, which was set to use `glm` in `'regression'` mode.

The third model we used was a GAM model. We did this by first creating a new `gen_additive_mod`, and setting the engine to `'mgcv'` and the mode to `'regression.'` We then created an object `gam_mod` using the fit function with `gam_spec` and the following variables: The outcome variable was again `gps_away_from_home,` and for the GAM model we used the following predictors: `case_rate, hospitalized_rate, emp_incq1, emp_incq2, emp_incq3, emp_incq4`, and `spend_remoteservices.` Again, we omitted vaccine rate, one of the original predictors, because it had a lot of N/A data points. The GAM model used a

`step_ns` for each variable as well in order to account for nonlinearity. Thus the final fit function for the GAM model used this formula:

```
gps_away_from_home ~ s(case_rate) + s(hospitalized_rate, k=20) + s(emp_incq1, k=20) + s(emp_incq2, k=20) + s(emp_incq3, k=20) + s(emp_incq4) + s(spend_remoteservices)
```

Model Evaluation

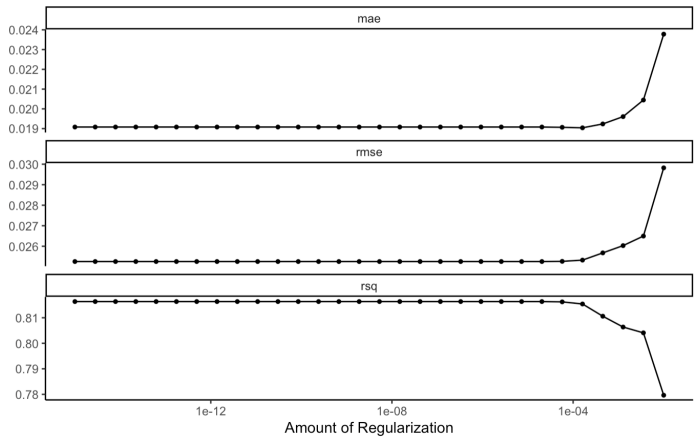
For the OLS model, we used `collect_metrics` for the `mn_mod` with the following results:

- `MAE: 0.01559419`
- `RMSE: 0.02071264`
- `Rsq: 0.86862455`

We did the same for the LASSO model with the following outcomes:

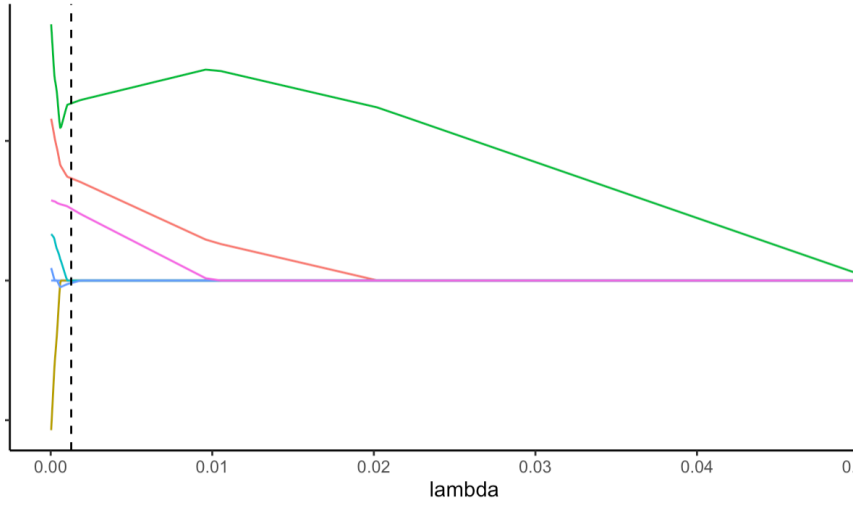
- `MAE: 0.01908087`
- `RMSE: 0.02525511`
- `Rsq: 0.81634721`

For the LASSO model, we used `autoplot` to plot error metrics (rsq, mae, rmse) against Amount of Regularization.



Based on these plots, it appears that the penalty has very little effect on the RMSE until it gets quite high. We hypothesized that this is because some predictors are quickly eliminated and the more important ones are not removed until much later.

For the LASSO model, we also found the best penalty value by using ggplot to plot the coefficients of each variable against possible lambda values. The plot is shown below:



Overall, 0.001268961 was the best value for lambda, indicated by the vertical line on the plot.

For the GAM model, we used `collect_metrics` to record the various error metrics with the following results:

- `MAE: 0.004224356`
- `RMSE: 0.005462357`
- `Rsq: 0.988170058`

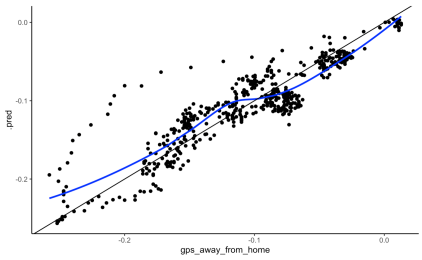
We then used `gam_mod %>% pluck('fit') %>% summary()` to find the significance (EDF) of each variable:

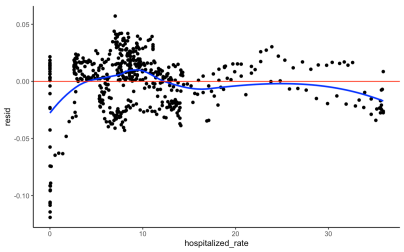
```
                          k'   edf k-index p-value
s(case_rate)            9.00  4.95    0.51  <2e-16 ***
s(hospitalized_rate)   19.00 16.21    1.03    0.65
s(emp_incq1)           19.00 18.33    0.76  <2e-16 ***
s(emp_incq2)           19.00 17.58    0.80  <2e-16 ***
s(emp_incq3)           19.00 15.07    0.82  <2e-16 ***
s(emp_incq4)            9.00  7.52    0.81  <2e-16 ***
s(spend_remoteservices) 9.00  1.00    0.97    0.21
```

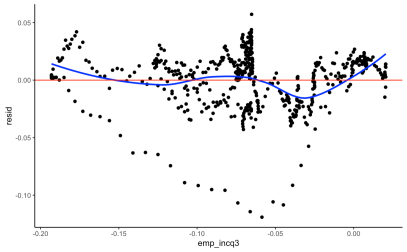
Based on the p-values in this output, the most important predictors are the income variables and `case_rate`, whereas `hospitalized_rate` and `spend_removeservices` are less important to the model's accuracy. For all variables, EDF is lower than k, indicating that there are enough knots to account for the nonlinearity. `spend_remoteservices` had an EDF of 1 which is an unexpected outcome that indicates that the relationship between it and `gps_away_from_home` is perfectly linear. The `case_rate` and `inc_q4` variables also had relatively low EDFs of 4.95 and 7.52 respectively. Other EDFs range from 15.07 to 18.33.

To assist with model evaluation, we made residual plots of each variable for each of the models. For OLS, most of the residual plots were randomly distributed. The only noteworthy aspect of the residual plots was that two variables, `hospitalized_rate`, and `spend_remoteservices` had higher residuals for lower values.

For the LASSO model, the residuals contained some strange trends. Many of the predictors had strong curves that looked like some kind of spline. Since some of them are eliminated very quickly by our LASSO method we did not worry about several of these but `emp_incq3` seems to have some strange patterns in it that we didn't fully understand the source of. Also, there were significantly more outliers in most variables' residual plots using LASSO compared to OLS. Below are three examples of residual plots from the LASSO model:







Overall, all of the regression models were made to answer our original point of inquiry, which is the connection between various predictor variables and time spent away from home (`gps_away_from_home`). We used similar predictors in each model and collected error metrics with the aim of discovering which model was the most accurately able to connect the predictors to `gps_away_from_home`.

4. Regression: Results

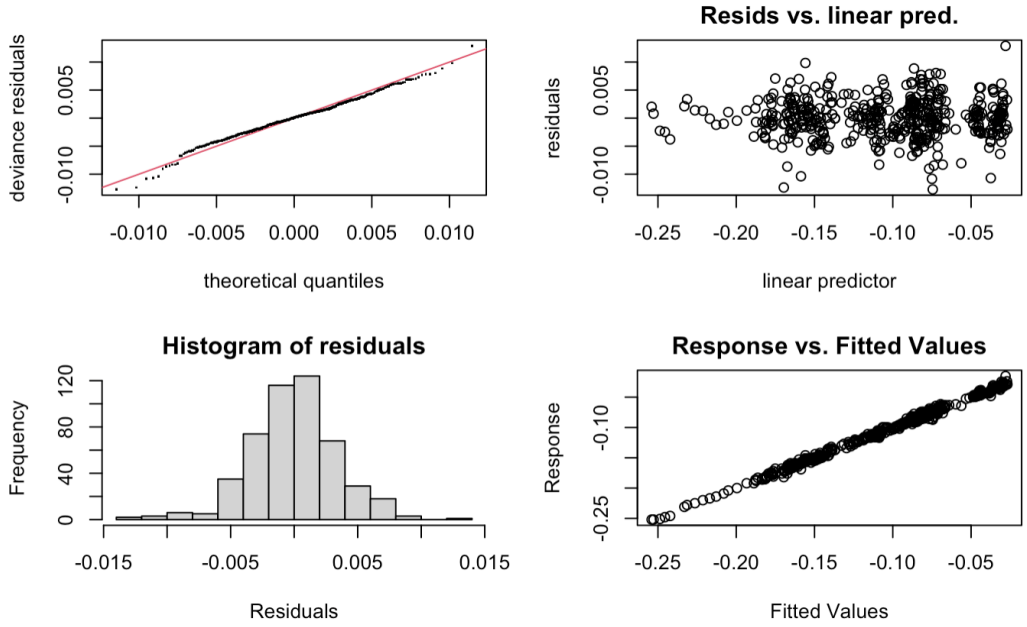
For our final model, we chose the GAM model because it has the best standard error metrics (MAE, RMSE, R-squared). Additionally, it's reasonable to assume that some of the predictor variables have a nonlinear relationship with time spent away from home because of the large and complex nature of the data, the pandemic, and the polarized behavior of different residents of the country.

Here is a comparison of the error metrics of each model:

	Error Metric						
Model		MAE	MAE Standard Error	RMSE	RMSE Standard Error	R-squared	R-squared Standard Error
	OLS	0.0156	0.0009	0.0207	0.0011	0.8686	0.0126
	LASSO	0.0191	0.0011	0.0253	0.0013	0.8163	0.0209
	GAM	0.0042	0.0004	0.0055	0.0005	0.9881	0.0067

The GAM model performs the best by all three of these metrics, having the lowest MAE and RMSE as well as the highest R-squared value. This aligns with the hypothesis that some of the variables may have a nonlinear relationship with `gps_away_from_home`, making a GAM model more ideal to show the relationships than OLS or LASSO models. However, it should be noted that all three models performed very well by the recorded error metrics, and the differences are minute. All models do a good job of predicting `gps_away_from_home`.

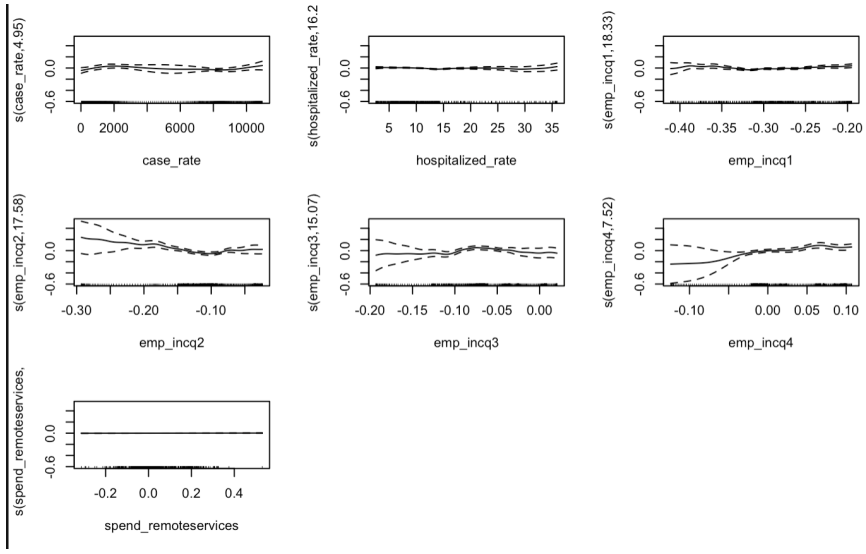
We plotted residual-based metrics using `gam_check`. The results are below:



There are no discernable patterns in the Q-Q plot. They are close to the line, indicating that the residuals are approximately Normal. There is also no obvious pattern in the residuals vs. linear pred. plot, which indicates the random distribution of residuals. The histogram of the residuals is unimodal and symmetric. Finally, the response vs. fitted values are positively correlated as expected.

5. Regression: Conclusions

Final Model Interpretations



The relationships between `gps_away_from_home` and some of the predictors look to be at least slightly nonlinear. The nonlinear predictors include all of the `emp_inc` variables and `case_rate`. Specifically, the `emp_inc` variables appear nonlinear at low values, whereas `case_rate` appears to be nonlinear at values between 3000 and 6000. These patterns indicate that forcing linear relationships would not have been ideal.

Evaluation Metric Interpretation

Error metrics like MAE, RMSE, and R-squared do not have exact values that are "good enough" or "not good enough." It depends on the units of measurement, the tolerance for error, and comparison to a baseline error, in this case the error of the OLS model. In this case, the GAM model performs better than the baseline in all three relevant error metrics. The MAE reveals that on average, the GAM model's predictions of `gps_away_from_home` is 0.4% away from the true percent change in time spent away from home. The RMSE shows that the average squared distance between a GAM model prediction and a true data point is 0.5%. The R-squared tells us that 98.8% of the variation in the data can be explained by the GAM model. All of these metrics place the GAM model well within reasonable error margins.

The mean `gps_away_from_home` value is -0.1051686, meaning that if the GAM model is off by 0.004, the error is very small relative to the total variable value. Similarly, the RMSE value of 0.005 is not a significant error value.

6. Classification: Methods

To practice using different kinds of classification methods we chose to answer two different questions that fit a random forest and logistic regression model respectively. For classification, we decided to omit the employment level variables except for middle employment, because the quartiles did not seem to add much insight to our classification models.

Random Forest

We chose to use random forest to predict the region that that particular state is in based off of the spending, income, and gps data. For the random forest, we can see its performance by evaluating the OOB error metrics such as sensitivity, specificity, and overall model accuracy. We also calculated the AUC of ROC for each of the regions in the United States (Midwest, Northeast, West, and South) to see how well our model did in predicting each of these regions. In addition, we estimated the variables, and order, that were most important in determining the forest. This was beneficial to evaluate our model in the context of determining regionality because it showed us which variables helped explain our outcome variable the best.

Logistic Regression

We used logistic regression to establish the predicted probability of a data point being in a state with an extensive COVID restriction policy, relative to states with a less restrictive policy. To evaluate our model, we imposed soft and hard thresholds on our outcome variable (`day_with_allCloser`). This binary variable indicates whether or not the state that it represents on a given day had nonessential business closure and stay at home orders in place. We used boxplot visualizations to set a hard threshold, from which we computed accuracy measures, such as sensitivity, specificity, and overall accuracy. We also estimated the AUC of ROC for the hard threshold.

7. Classification: Results

The random forest algorithm gives us pertinent information about our dataset. We are able to see our model accuracy (in terms of OOB), our most important variable, and measures of sensitivity and specificity. The OOB prediction error is 0.0246, which means our model is about 99.97 % accurate. The sensitivity measure was 97.8 percent, while the specificity measure was 99.9 percent, a little bit better than the sensitivity measure. These two measures, sensitivity and specificity, measure how well our model performs. Our model was able to correctly identify data points 99.97 percent of the time, with an overall accuracy of 99.7 percent. Since the largest group we had in our data was less than 25% of the data this is a significant improvement. Additionally, the most important variable is the variable income below the median level. This threshold was best, out of all of our predictors, at determining our outcome variable. In other words, this threshold gave us the most amount of information if the datapoint was or was not in this geographic division.

Our logistic regression model gave us a sensitivity measure of 93.5, a specificity measure of 80.8, and an overall accuracy of 88.6. In other words, our model was able to correctly predict data points in the correct category 93.5 percent of the time, with an overall accuracy of 88.6. This is a ~20-30% improvement on our data's no information rate.

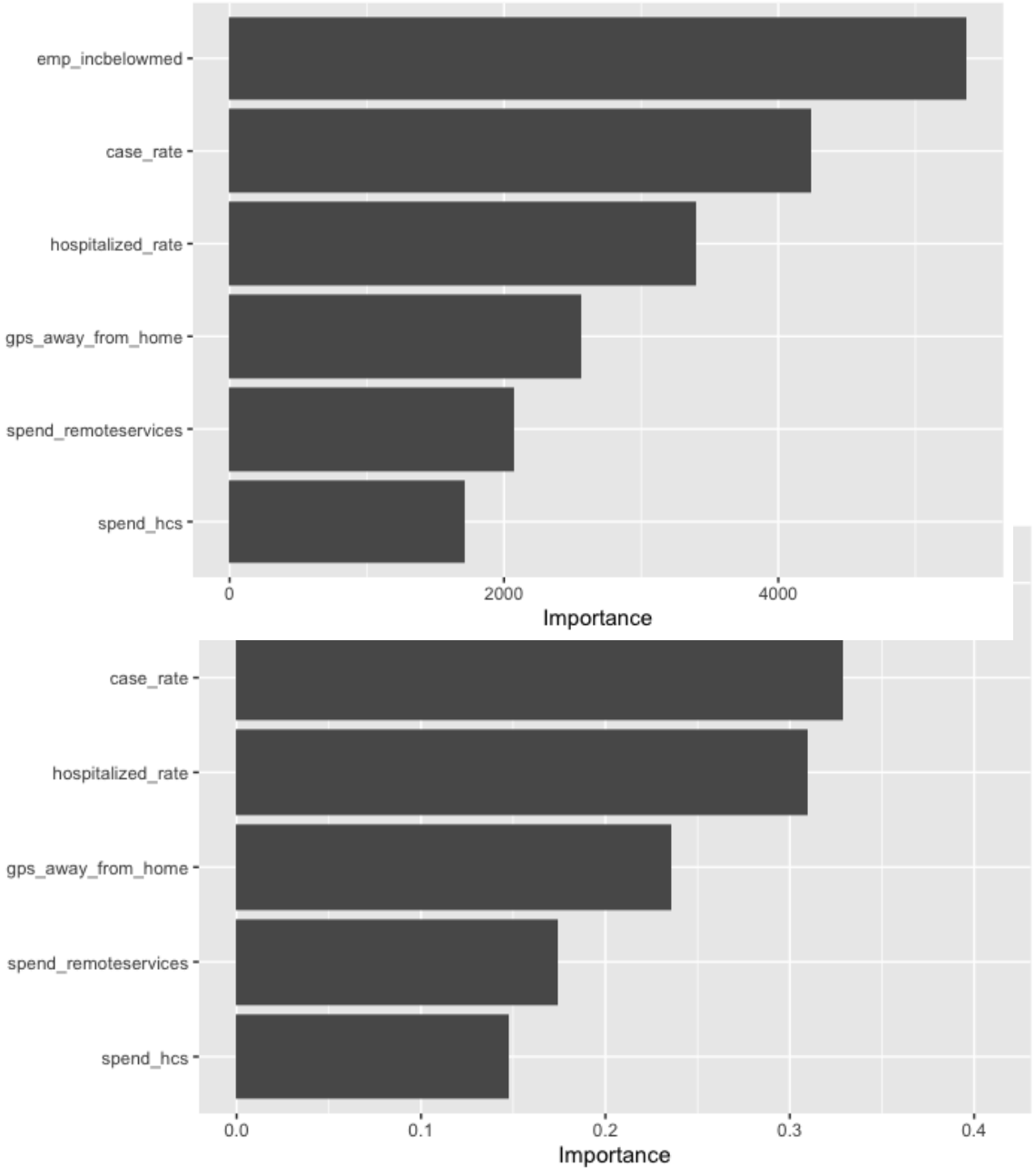
8. Classification: Conclusions

As was stated, for our regional classification model we had a very high predictive accuracy. In 99.97% of cases, our model given a day of data would classify that data into the correct region of the country. Given the low stakes nature of this prediction and the model's good performance that is certainly an acceptable level of error. In the model's confusion matrix we can see this accuracy in effect.

	Truth Midwest	Northeast	South	West
Prediction Midwest	5266	36	123	48
Northeast	56	6563	26	44
South	427	115	9471	131
West	59	62	60	7031

The model most often confuses the Midwest and the South but that is still very rare. Overall the vast majority of guesses are quite accurate. It is a bit concerning that the incorrect guesses are not quite evenly distributed. That could be a result of either bias in our model or real significant differences between certain regions that are easier to detect.

To get a little more insight into this model's performance we can look at variable importance based on impurity reduction and permutation (or the greatest error reduction).



Both of these graphs show the same pattern of variables. The most important to our prediction is the level of employment for middle income populations. Then the most significant drop in importance occurs to case rate, then hospitalization rate and so on. None of these variables are unimportant for our prediction. It is interesting that the employment variable and two covid variables prove so important in this scenario. It is possible that regions in the US had very pronounced and separate case rates which is possible given the geographic component of the spread of COVID.

While the regions are sorted correctly the sensibility of this sorting and its usefulness of it is questionable. Given any normal data set with state-level data, it is likely that you could very easily figure out what region that state is in without using a predictive model. It also makes relatively little sense to sort regions with the data we used. State-level data can be very indicative of what state is represented. There is no strong overarching theoretical framework to suggest why state-level data on employment, spending, or COVID-19 rates should predict what region of the country that state is in. The one relatively simple answer to why our model can predict is that given the data it can determine which state any given day came from and then just sort it into that state's given region. Our minimum number of cases in each node is twenty. It would probably be relatively easy for our trees to make many branches that just classify twenty out of thousands of days as being from one state and then asinine a region. Given these constraints, this model seems to prove that we can easily classify regions with state-level macro-economic factors but it doesn't definitively prove why that is possible.

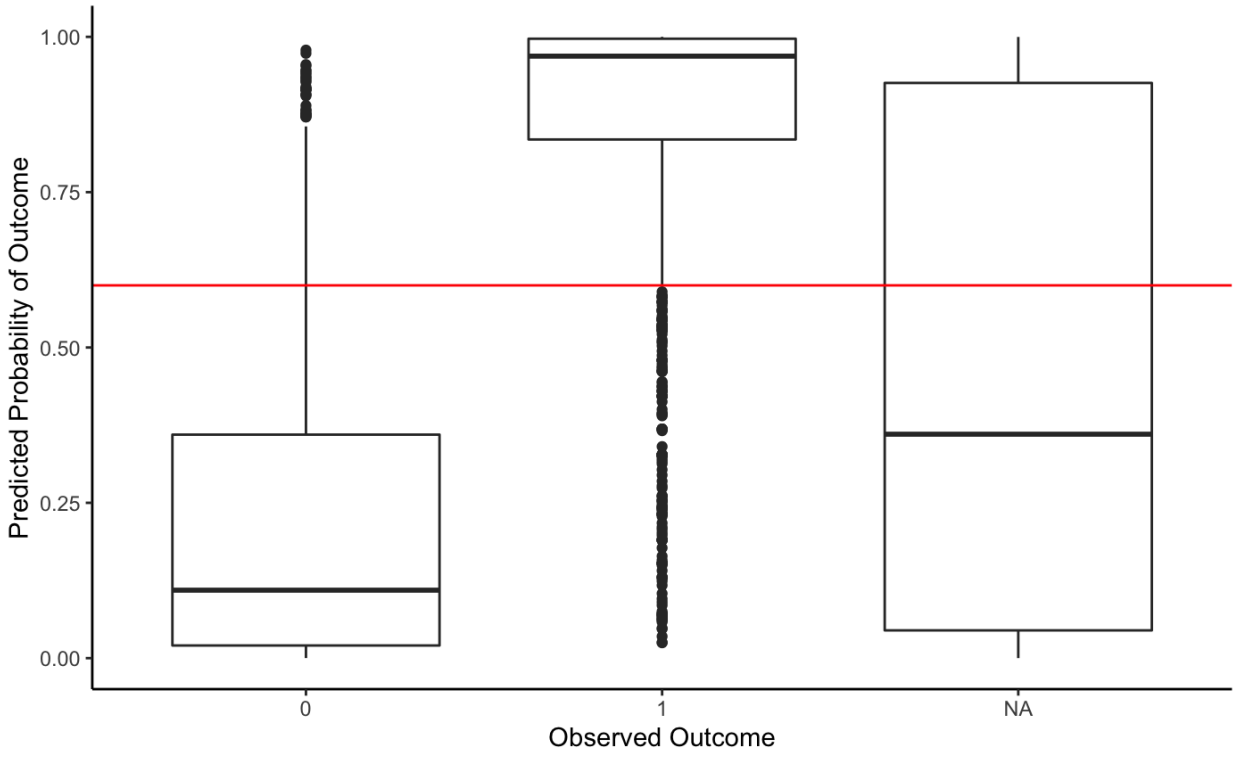
Our logistic regression model also performed relatively well as mentioned. It provided a decent improvement in consistently predicting lockdown effects in states. We decided to limit the data to only the period of time in which lockdowns were in effect so we could compare states to other states rather than states to themselves. This made it so all days were classified under similar general macroeconomic and social conditions during the pandemic. Since predicting if a state or region is in lockdown could theoretically be a useful tool if local data is absent or if you wanted to automatically trigger some kind of lockdown system, I think it is important for this model to perform well.

Our confusion matrix was as follows:

	Truth No Lockdown	LockDown in Effect
Prediction No Lockdown	819	104
Lockdown in Effect	194	1505

Sensitivity 93%, Specificity 81%, Accuracy 89%

In this model we correctly identify 93% of days that actually were lockdowns and 78% of days that were not lockdowns. In both cases, this is a 25%+ improvement over just guessing true or false. In 12% of no lockdown cases, however, the model incorrectly predicted a lockdown. That is not good for an accurate prediction of whether or not a region has gone into lockdown. At least in, for example, a finance or policy context, incorrectly predicting lockdowns 11% of the time is probably unacceptable.



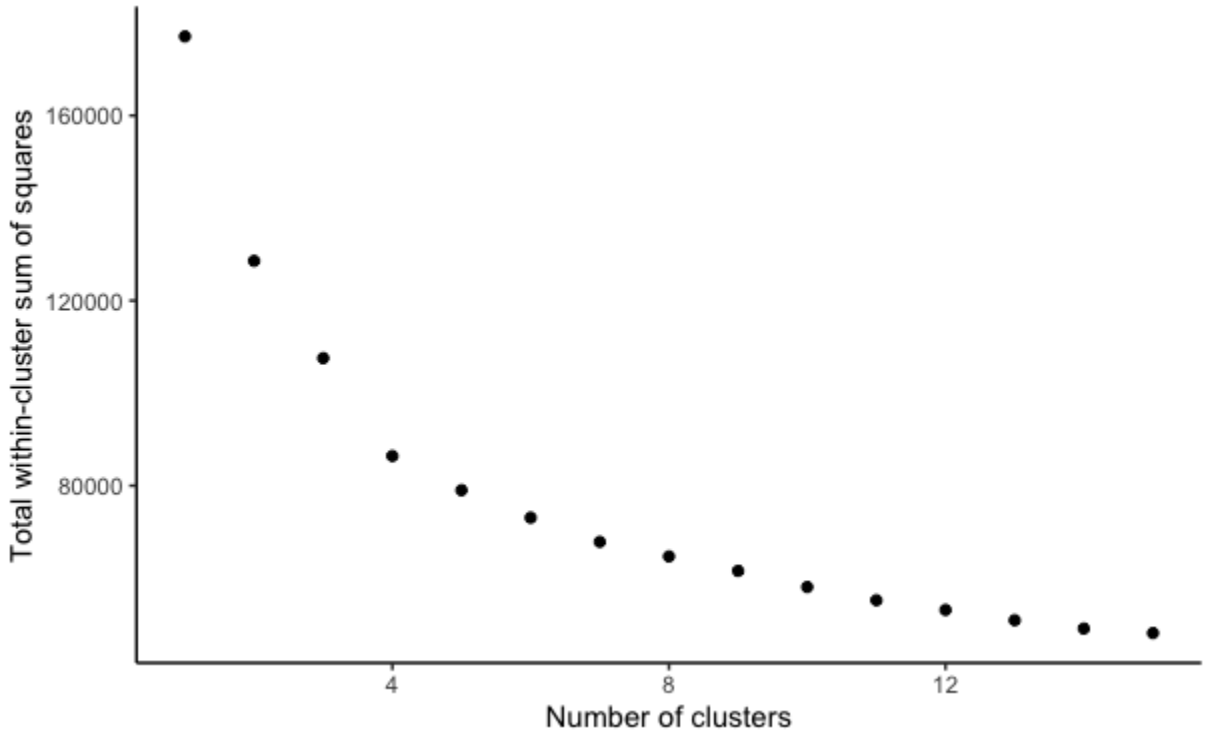
Taking a look at the predictions of our model we can see that it is impossible for our model to not have both false positives and false negatives without always predicting one or the other. There are too many highly predicted non-lockdown days for us to avoid this. This model, while an improvement on guessing, is probably not acceptably accurate in any real context.

9. Unsupervised Learning: Clustering

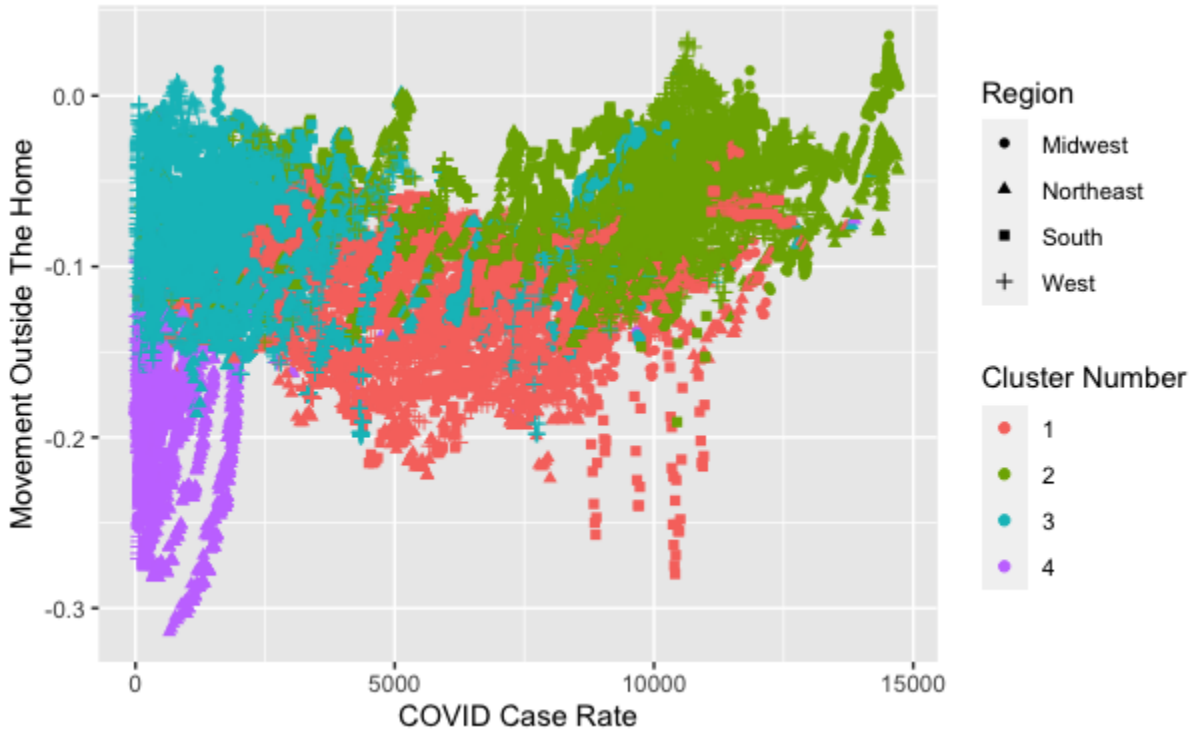
For unsupervised learning, we used both k-means clustering and hierarchical clustering to determine whether the natural clusters created from the cases follow the regional groupings of states (i.e. South, West, Midwest, and Northeast).

For k-means clustering, we used Euclidean distance in order to calculate the distance from each case to the centroid of each cluster. We normalized all the variables in order to ensure that no one variable was given more weight in creating the clusters.

In addition, we chose a k-value of 4 for two reasons: it matches the number of regions (South, West, Midwest, and Northeast), and the within-cluster sum of squares starts leveling off when k=4. Here is a graph of the within-cluster sum of squares for different values of k:



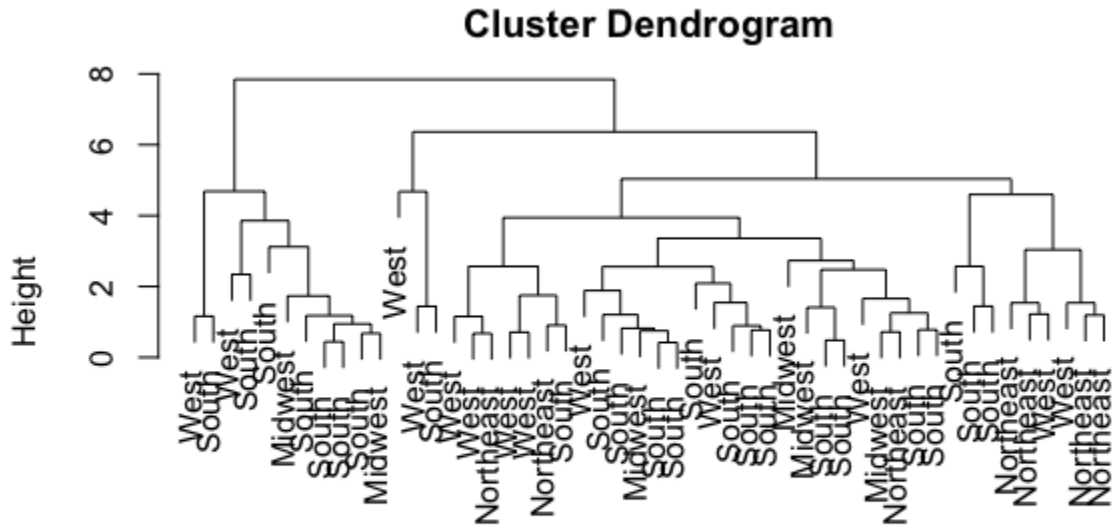
The results of the k-means clustering analysis are shown below:

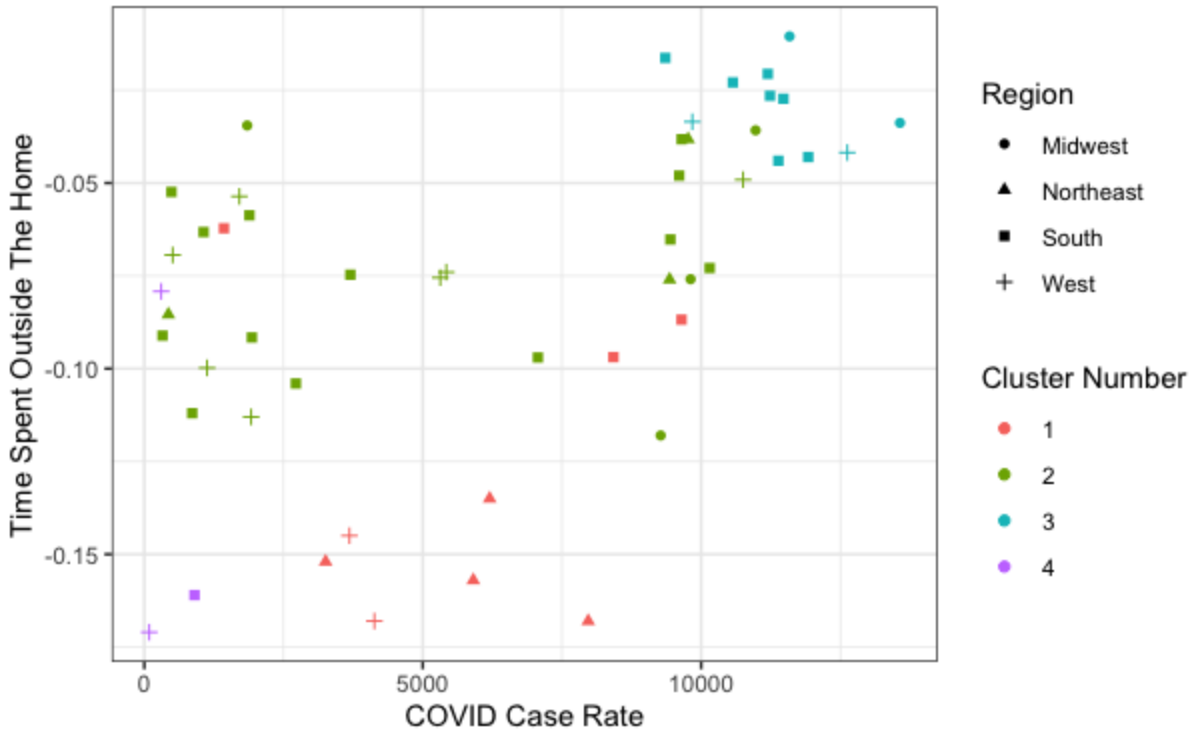


In summary, there did not seem to be a strong relationship between the four clusters and the four US regions. We created a confusion matrix (shown below) based on which regions the four clusters aligned the most with. However, there does not seem to be as much overlap between the clusters and regions, which is confirmed by its overall accuracy of only 34%. Indeed, the adjusted Rand index, which compares two sections of data, was only -0.016, indicating that the two sections were not very similar.

	Truth			
Prediction	Midwest	Northeast	South	West
Midwest	2	1	7	5
Northeast	1	4	5	2
South	2	0	7	2
West	1	2	5	4

For hierarchical clustering, we used complete linkage in order to create clusters, as this method created the most dense clusters. We hoped that these dense clusters would more closely align with the four regions, rather than other forms of linkage in which the clusters are not as tight. We normalized the variables before performing the analysis to ensure that the variables were weighted equally. We chose to cut the dendrogram at the height of 5, as this was the height at which there were four clusters.





In summary, we found that there did not seem to be a strong relationship between the hierarchical clusters and the US regions. We did not create a confusion matrix for hierarchical clustering, because we assumed that the results would be similar to that of k-means clustering.

Overall, our unsupervised learning analysis does not show there to be a relationship between the natural clusters and the US regions. One reason for this may be that we only looked at four regions, but there is likely an extreme amount of diversity among our variables within those regions. For instance, both Delaware and Oklahoma are considered to be a part of the “South,” but those states have very different political, health, and economic situations. In addition, there are huge differences between urban and rural areas within each state. All of these differences mean that the “averages” we use for our data are not representative of the regions as a whole. Because of this, it makes sense that our natural clusters did not closely align with the four regions that we chose.

10. Big Takeaways

Our analysis has provided us with several lessons about state level data and modeling. Our main takeaways were as follows:

- All three of our regression models (Ordinary Least Squares, LASSO, and GAM) were very accurate at predicting `gps_away_from_home`. GAM was slightly better than the other two.
- Based on our GAM model, the case rate and the income variables are the most important predictors in the model.
- Given the random forest model's constraints, it proves that we can easily classify regions with state-level macro-economic factors; it doesn't definitively prove why that is possible.
- Predicting on a given day if a lockdown is in effect is a difficult task that cannot be done accurately.
- Within our data, there does not seem to be any natural clustering that corresponds with the US regions.

While exploring this data, we have created a number of different models that answer a variety of questions. We have shown that we cannot find natural regional clusters, can classify states into regions, predict if a state was in lockdown on a given day, and estimate how much people will move. These takeaways tell us about each individual prediction but also combined tell us a little bit about using state level data. We have used a variety of state level economic factors like movement, unemployment, and expenditure to answer a number of questions. It seems that state-level predictors are decently suited to classification and estimation of other economic factors. State-level economic factors do not, however, lend themselves to easy explanations of why trends may occur or tell a story.

11. Limitations of Project

This project has many limitations. For example, we learned a lot about the limitations of state level data's explanatory power. To answer our questions, we always use states and regions within our analysis. This hides the extreme variation within those states and regions. Statewide predictors should not be used to predict at a smaller or individual scale. We have shown that they seem to be able to predict and classify trends at the state level, but that does not mean they can be specified.

In answering our questions, we necessarily suggested both that these predictors have real causal relationships to one another, we do not have the expertise to provide a theory that backs up these relationships. We also use statewide averages to draw conclusions about populations within those states. For example, our predictive models predicted the average relative change in state-level population movement. Our model is not trained and can not be used to predict why or how a given individual would move. It does, however, show that at this macro-level, economic variables can predict movement relatively accurately. When applying or analyzing all of our conclusions, we should not take these average numbers as a prediction of individual behavior in the future.

Finally, an underlying issue with our data is that for some variables states may collect data differently. This means that comparing or classifying by state may not be entirely accurate. Quirks in how states collected data could help in classifying them, and using the model to make state-to-state comparisons can not be done, because we are unsure of the level of accuracy in each state's reporting of COVID-19 statistics. These limitations are key in adding context to our data, models, and conclusion.