# Predicting Daily COVID-19 Rates in U.S. Counties: Visuals

labeledData.csv: contains every record with all the features and labels (the number of daily cases). This complete dataset has 354,977 records and 32 features.

county	day	mobility	population	density	num_househ	income	poverty	white	black	native	asian	hispanic	no_college	blue	emergency	election_pos	school_closu
1001	39	7.194	55869	35.853419	21313	59338	3.6	76.7252388	19.6309419	0.48020719	1.22479812	2.965774	43.9	-49.48	0	0	0
1001	40	14.587	55869	35.853419	21313	59338	3.6	76.7252388	19.6309419	0.48020719	1.22479812	2.965774	43.9	-49.48	0	0	0
1001	41	13.865	55869	35.853419	21313	59338	3.6	76.7252388	19.6309419	0.48020719	1.22479812	2.965774	43.9	-49.48	0	0	0
1001	42	13.88	55869	35.853419	21313	59338	3.6	76.7252388	19.6309419	0.48020719	1.22479812	2.965774	43.9	-49.48	0	0	0
1001	43	14.659	55869	35.853419	21313	59338	3.6	76.7252388	19.6309419	0.48020719	1.22479812	2.965774	43.9	-49.48	0	0	0
1001	44	17.066	55869	35.853419	21313	59338	3.6	76.7252388	19.6309419	0.48020719	1.22479812	2.965774	43.9	-49.48	0	0	0
1001	45	11.562	55869	35.853419	21313	59338	3.6	76.7252388	19.6309419	0.48020719	1.22479812	2.965774	43.9	-49.48	0	0	0
1001	46	8.51	55869	35.853419	21313	59338	3.6	76.7252388	19.6309419	0.48020719	1.22479812	2.965774	43.9	-49.48	0	0	0
1001	47	15.882	55869	35.853419	21313	59338	3.6	76.7252388	19.6309419	0.48020719	1.22479812	2.965774	43.9	-49.48	0	0	0
1001	48	14.429	55869	35.853419	21313	59338	3.6	76.7252388	19.6309419	0.48020719	1.22479812	2.965774	43.9	-49.48	0	0	0

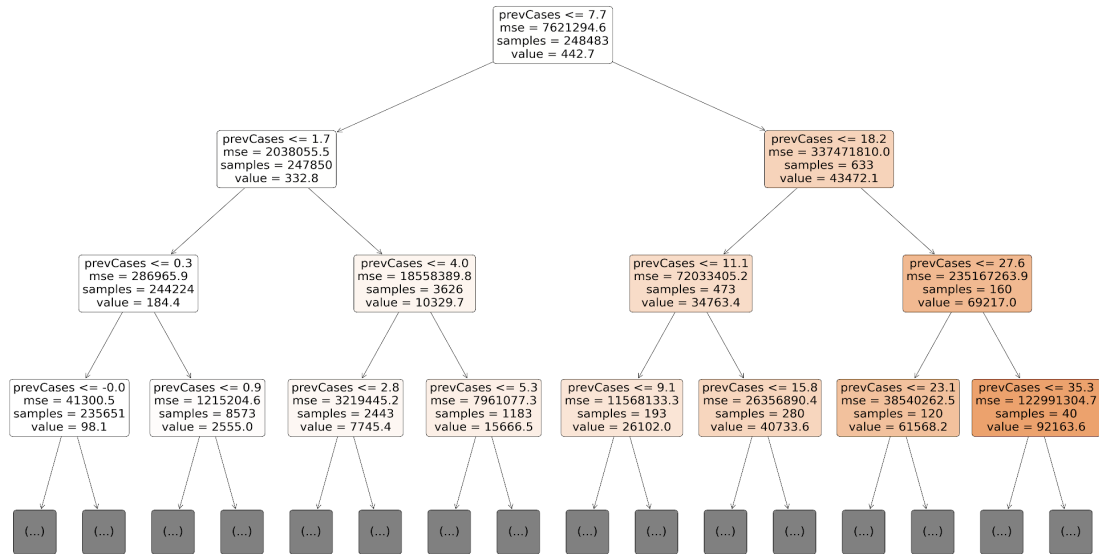*Visualization of Government Policy Implementation Over Time*
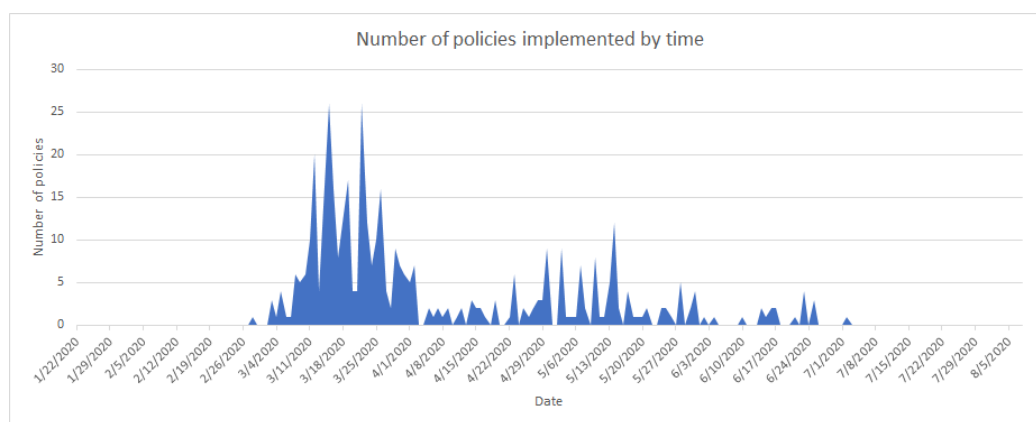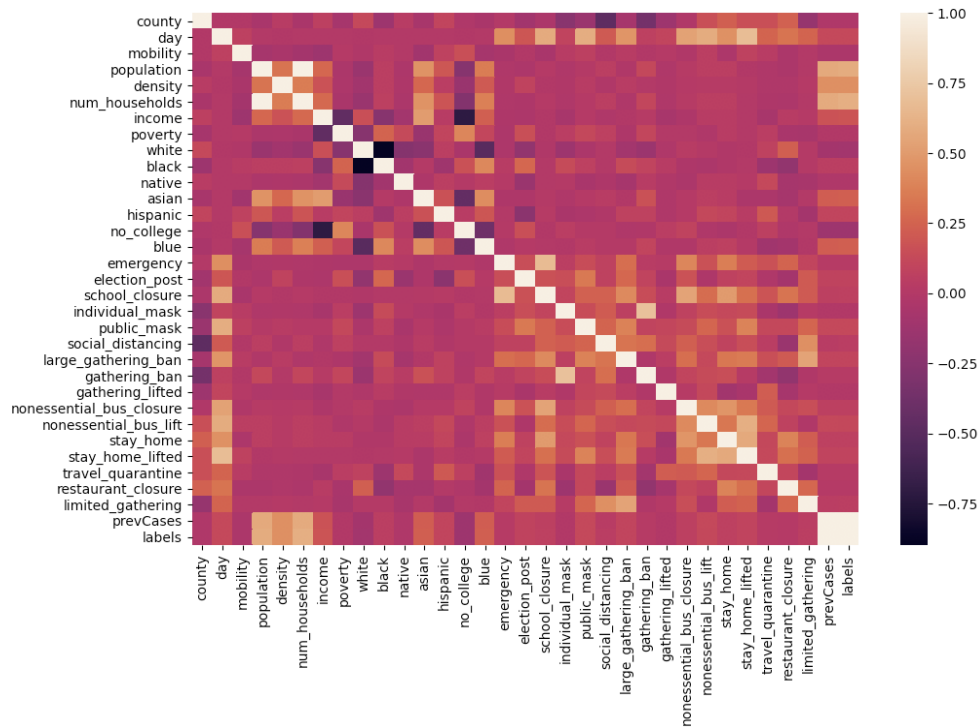

Visualization of Linear Regression Coefficients
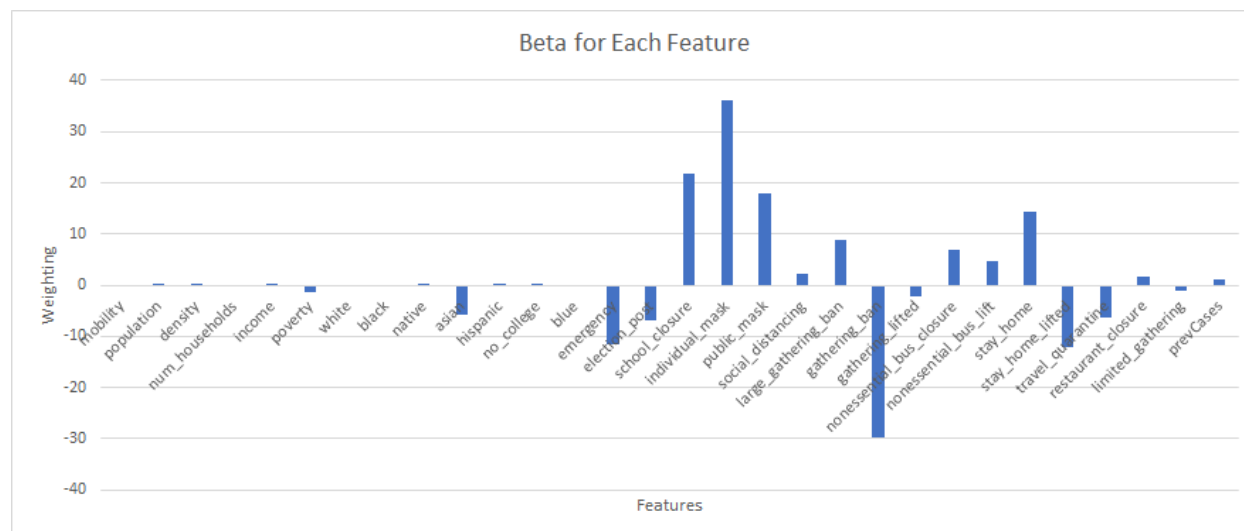
Visualization of Decision Tree

Feature Selection Methods

1. *Pearson Correlation:* Drop highly correlated features

Feature 1 (keep)	Feature 2 (drop)	Pearson Correlation
num_households	**population**	0.9956
income	**no_college**	0.7103
school_closure	**emergency**	0.6688
gathering_ban	**individual_mask**	0.7088
nonessential_bus_lift	**stay_home_lifted**	0.6041

2. *PCA Dimension Reduction:* Keeping 95% variance, reduce 32 features to 24 components

1st Component	2nd Component	3rd Component
day (0.387)	num_households (0.422)	black (0.430)
school_closure(0.332)	population (0.416)	white (-0.408)
stay_home_lifted (0.320)	asian (0.366)	poverty (0.291)
stay_home (0.305)	blue (0.323)	county (0.287)
nonessential_bus_closure (0.296)	prevCases (0.295)	income (-0.271)

3. *Ensemble Method*: Keep only the features with significant coefficients from the linear regression model results as the features for decision tree regressor and random forest regressor.

Empirical Results and Comparisons

1. *Linear Regression*

	Train R^2	Train RMSE	Test R^2	Test RMSE
Non-regularized	0.99678	156.61270	0.99704	153.89906
Pearson Dropped Feats Lasso (alpha = 0.001)	0.99677	156.83426	0.99705	153.89909
Pearson Dropped Feats Ridge (alpha = 10.0)	0.99677	156.83426	0.99678	153.89912

2. *Decision Tree Regressor (*Highest R^2 and lowest RMSE)*

	Train R^2	Train RMSE	Test R^2	Test RMSE
Without PCA (MD = 11, MLS = 1)	0.99986	42.80797	0.99903	88.17512
With PCA (MD = 23, MLS = 1)	0.99986	32.94302	0.99578	184.07828
Without PCA, Pearson Dropped Feats (MD = 20, MLS = 1)	**0.99999**	**7.55379**	**0.99904**	**87.61181**
Pearson Dropped Feats With PCA (MD = 20, MLS = 1)	0.99974	44.69032	0.99694	156.71367

3. *Random Forest Regressor*
   a. row 1,2,4: Nest = 50, MaxFeat = $\sqrt{numFeat}$, MD=10, MLS=3
   b. row 3: Nest = 50, MaxFeat = $\sqrt{numFeat}$, MD=12, MLS=3

	Train R^2	Train RMSE	Test R^2	Test RMSE
Without PCA	0.99466	196.50934	0.99382	222.67817
With PCA	0.97283	435.26068	0.97550	443.32335
Pearson Dropped Feats Without PCA	0.99797	122.58274	0.99713	151.84601
Pearson Dropped Feats With PCA	0.96767	472.18065	0.97062	485.51301

4. *Ensemble Methods*

	Train R^2	Train RMSE	Test R^2	Test RMSE
Decision Tree (MD =20 , MLS = 1)	0.99999	8.82178	0.99873	101.06618
Random Forest (Nest = 50, MaxFeat = $\sqrt{numFeat}$, MD=12, MLS=3)	0.96388	483.08772	0.96830	480.33673