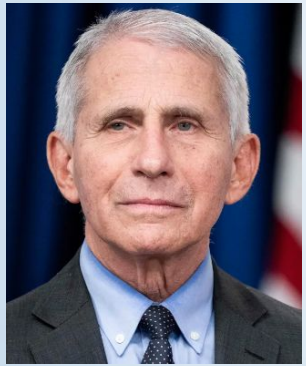
Future Faucis

Covid Lessons, Pandemic Predictors



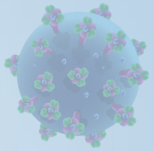
McBossy - Katie Ellsweig
And the Minions:
John Gilheany
Juston Dea
Rafi Rahman
Sulaiman Bah

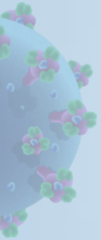
Problem Statement

The Department of Health and Human Services has hired our firm to study the covid-19 pandemic in order to make recommendations to prepare for future pandemic.

- Which variables are most strongly linked to covid outcomes?
- Which policies if any were significant?
- How did demographics and pre-existing conditions pay a role?

Presentation Overview





Problem Statement

Background

Previous Research

Project Flow

Data Collection & Cleaning

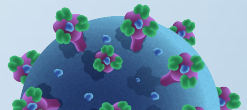
EDA

Tableu

Modeling

Sources

Conclusion



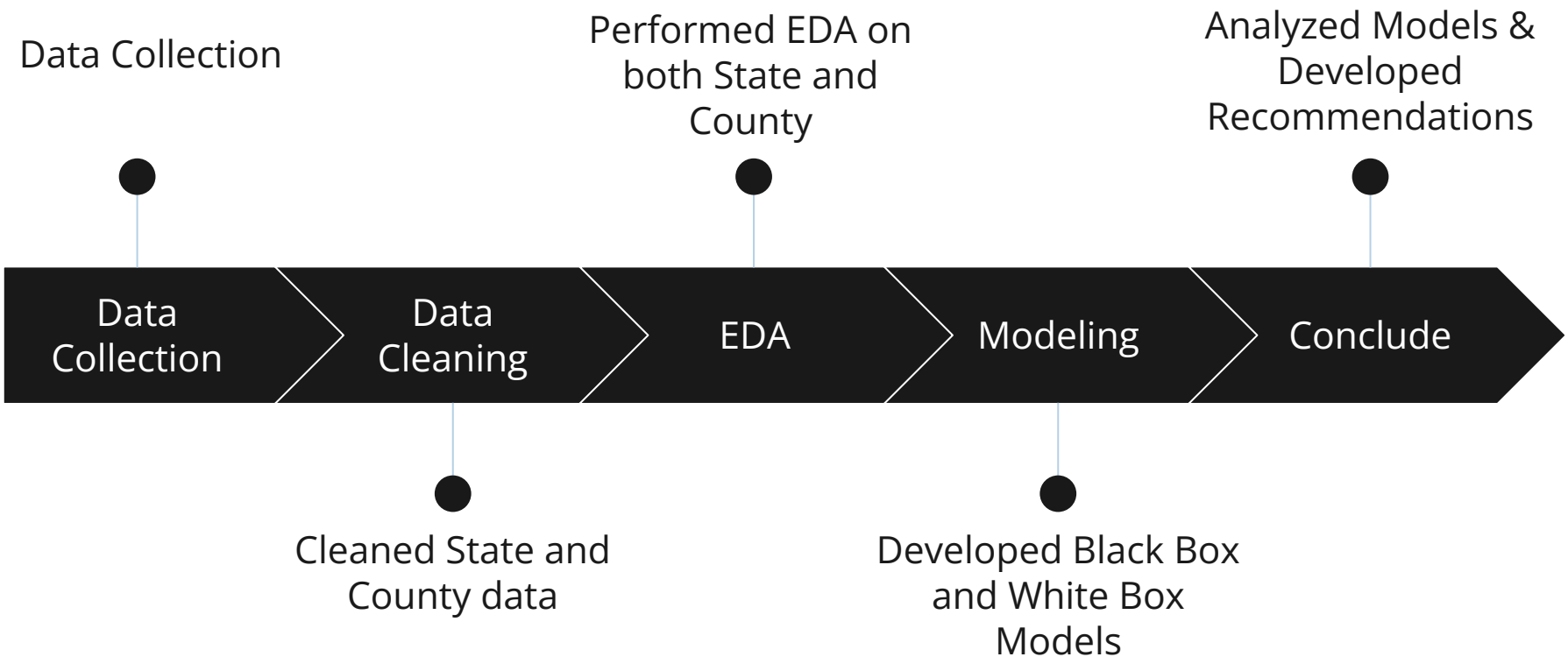
Background

- The United States fared poorly during the Covid-19 pandemic
- Only Peru had more deaths per capita than the United States
- The United States had the most overall deaths

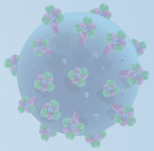
Previous Research

- Lancet Study- 50 States
  - Inequality
  - Mask policy and stay at home orders
  - Access to healthcare
- UK study
  - Racial disparities in outcomes
- Ohio vs Florida County-Level Comparison
  - Political leanings of population affected outcomes

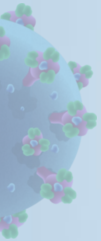
Project Flow

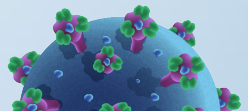


Data Collection & Cleaning

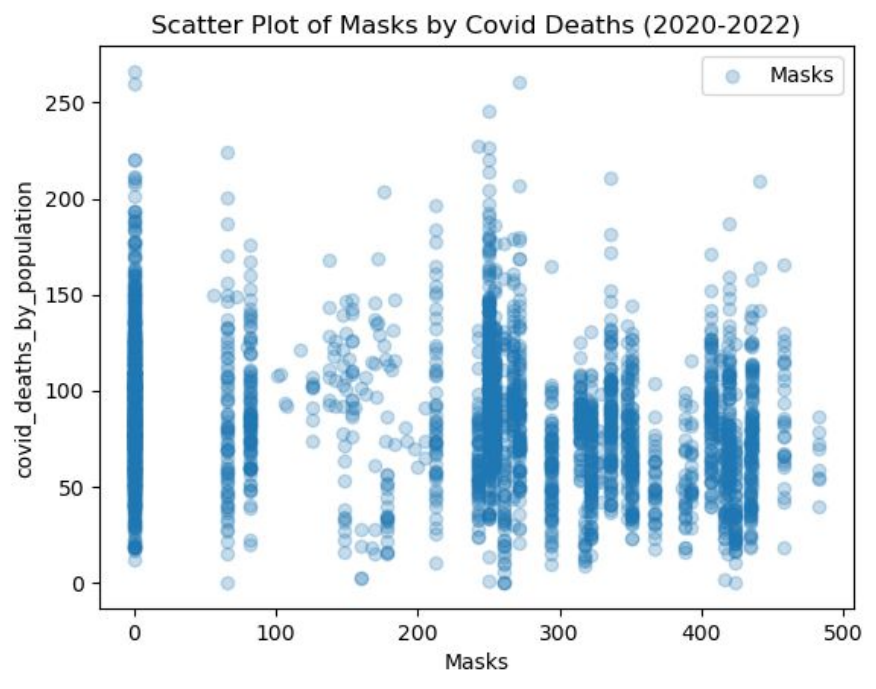


- Web Scraping / Downloaded CSVs
- Ignore files for large data
- Began pulling in data by State, after modeling, we decided to also pull in County data
- Filled missing data
  - 0 for masks
  - Substituted 2018 pre-existing conditions stats for NJ (2019 missing)
- Dropped duplicates
- Delegated unallocated values, county population percentage, distributed State values to the county
- Concatenated and merged data collections to make a State and County dataframe





EDA


Scatter Plot of Masks by Covid Deaths (2020-2022)
Masks
covid_deaths_by_population
Masks
0
50
100
150
200
250
0
100
200
300
400
500

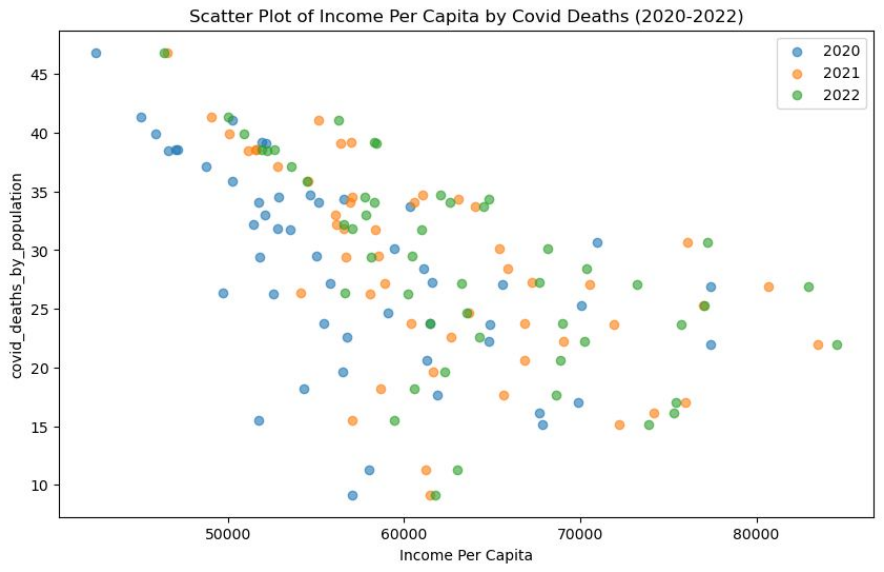
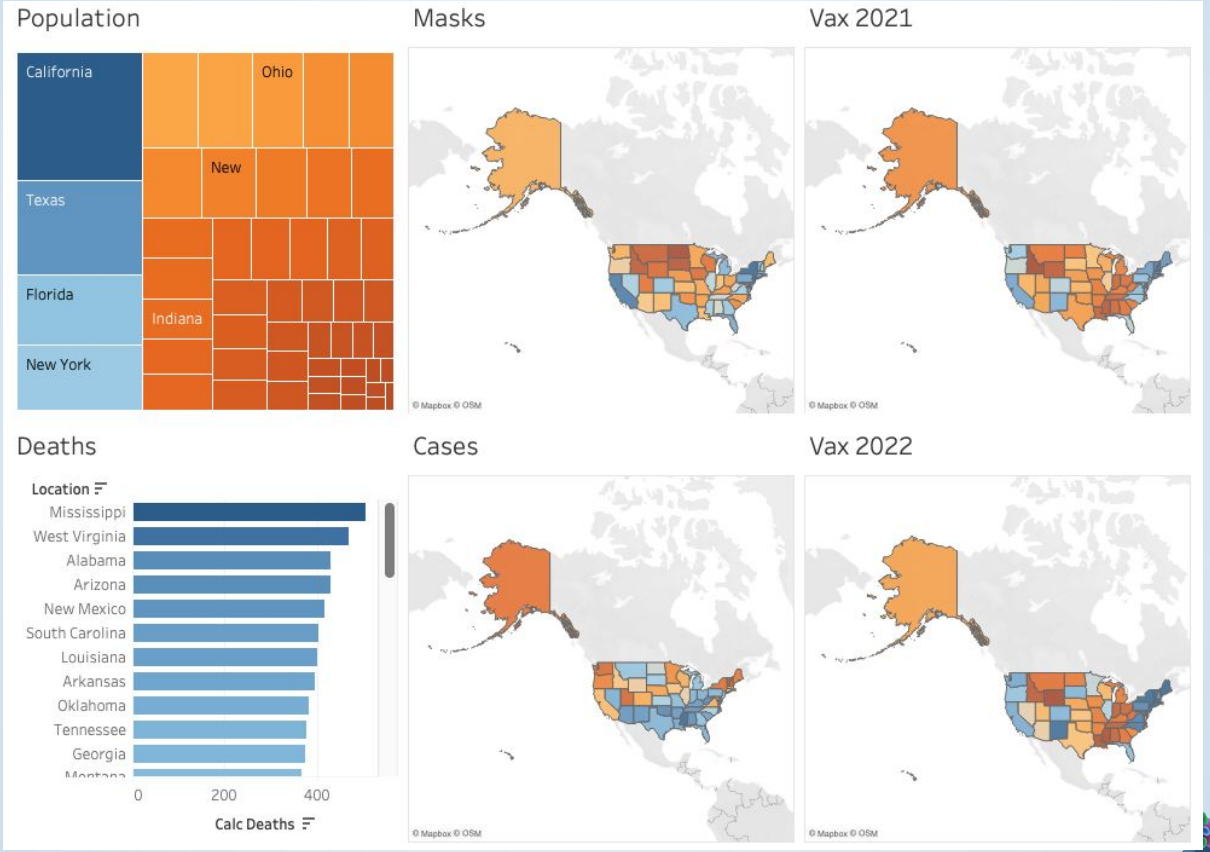

Scatter Plot of Income Per Capita by Covid Deaths (2020-2022)
2020
2021
2022
covid_deaths_by_population
Income Per Capita
10
15
20
25
30
35
40
45
50000
60000
70000
80000

Tableau Demo



Model Summaries

State Level Models	Y = Excess Deaths Count						Y = Number of Covid Deaths						Y = (Covid Deaths / Population) *100			
Model Name	Model Type	Key Features	Train 1	Test 1	Train 2	Test 2	Model Type	Key Features	Train 1	Test 1	Train 2	Test 2	Model Type	Key Features	Train 1	Test 1
Model 1	Linear Regression		1	0.7934779884	1	-1.966949035	Linear Regression		1	0.8804713015			Linear Regression		1	-3.74982683
Model 2	Linear Regression	Standard Scaler	1	0.7934779884	1	-0.08104481218	Linear Regression	Standard Scaler	1	0.9133207097			Linear Regression	Standard Scaler	1	-1.576498624
Model 3	KNN		1	0.7694275677	0.7474726986	0.7752529507	KNN		0.7791424748	0.6836087418			KNN		0.2828520321	-0.5384576466
Model 4	KNN	Standard Scaler	0.7597748656	0.6530279454	0.625531001	0.4591595909	KNN	Standard Scaler	0.6966822173	0.4383678725			KNN	Standard Scaler	0.6721761969	0.1596908953
Model 5	Decion Tree		1	0.7252620627	1	0.6490272631	Bagging		0.9352376396	0.7768267249			Bagging		0.8983256166	-0.1034077588
Model 6	Decion Tree	Standard Scaler	1	0.190149813	1	0.7454567967	Bagging	Standard Scaler	0.9167215447	0.6405498403			Bagging	Standard Scaler	0.8622280312	0.03385608897
Model 7	Bagging		0.9171158987	0.6501179269	0.9671084244	0.7756343758	RF		0.9494492395	0.7206860207			RF		0.9134075592	0.1047253986
Model 8	Bagging	Standard Scalar	0.9343257762	0.5798671247	0.9412297644	0.6779138483	RF	Standard Scaler	0.9604545741	0.7500553107			RF	Standard Scaler	0.9224546568	0.1442125894
Model 9	Random Forest		0.9642629021	0.7628486851	0.9405926051	0.8261410141										
Model 10	RandomForest	Standard Scalar	0.9490787867	0.7168981406	0.9452305168	0.8151190555										

	Y = Excess Deaths Count						Y = Number of Covid Deaths					
	Model Type	Key Features	Train 1	Test 1	Train 2	Test 2	Model Type	Key Features	Train 1	Test 1	Train 2	Test 2
	Random Forest	Poly & SS	0.9556624784	0.6596390176			Random Forest	Poly & SS	0.9592415099	0.8145570907		
	Bagging	Poly & SS	0.9668165468	0.7742026379			Bagging	Poly & SS	0.9684471928	0.8230630848		
	KNN	Poly & SS					KNN	Poly & SS	0.7436503808	0.5625614946		
	RF	Ada+Multi+Rand+Poly+SS	0.9581031719	0.7994069265			RF	Ada+Multi+Grid+Poly+SS	0.9615903661	0.7842375911		
	Gradient Boosting	Grid, KFold, MultiOutput	0.9996712213	0.9039638429			Gradient Boosting	Grid, KFold, MultiOutput	0.9996712213	0.9034508645	0.9999918873	0.9182707939
	RF	Grid, KFold	0.7930247826	0.6592708755			RF	Grid, KFold	0.7930247826	0.6592708755		
	Linear Regression	Poly & SS	1	0.9170538077			LASSO		0.99999914	0.9226665135		
	LASSO		0.9999983098	0.8926730136			Ridge		0.9999991019	0.8806077278		
	Ridge		0.9999995948	0.8031595859								

County Level	Y = Number of Cases						County Level	Y = Number of Deaths					
Model Name	Model Type	Key Features	Train 1	Test 1	Train 2	Test 2	Model Name	Model Type	Key Features	Train 1	Test 1	Train 2	Test 2
lr	Linear Regression		0.9681483874	0.9325206419	0.9639786858	0.9325020607	lr	Linear Regression		0.926815388	0.8859051229	0.923403073	0.9044371814
knn	K Nearest Neighbors		0.6724283833	0.6910618917	0.6883310563	0.7165832065	knn	K Nearest Neighbors		0.656143894	0.7031525856	0.6639593331	0.711999107
tree	Decision Tree		1	0.8802491736	1	0.8587801828	tree	Decision Tree		1	0.6266117387	1	0.8918904809
tree_bagged	Decision Tree	Bagged	0.9459740639	0.8251022411	0.8911918973	0.8185710802	tree_bagged	Decision Tree	Bagged	0.9516149889	0.8733045203	0.8827291186	0.7922803276
rf	Random Forest		0.9466441925	0.8239315925	0.9469735268	0.8434835697	rf	Random Forest		0.9516391408	0.8768628588	0.9468652333	0.8121540004
rf_rs	Random Forest	Random Search	0.8491720297	0.9052849824	0.9475963064	0.8313269453	rf_rs	Random Forest	Random Search	0.9530671119	0.8763718398	0.9442887487	0.767885037
rf_gb	Random Forest	Gradboost	0.9981189105	0.8795899755	0.9973426354	0.8861501802	rf_gb	Random Forest	Gradboost	0.9969479146	0.8670715481	0.9938920089	0.9222647585
rf_gb_rs	Random Forest	Gradboost + RS	0.9999999993	0.8126106046	0.9946659198	0.8852267062	rf_gb_rs	Random Forest	Gradboost + RS	0.9999999993	0.7860479483	0.9939825576	0.8959240127
et	Extra Trees		1	0.8448992488	1	0.8914158581	et	Extra Trees		1	0.9219399605	1	0.885150115
et_rs	Extra Trees	Random Search	0.8960953582	0.9256104083	0.89852462	0.8759890389	et_rs	Extra Trees	Random Search	1	0.9224885287	1	0.8717618973
vr	Voting Regressor		0.956784952	0.9067405328	0.9805737869	0.8835549657	vr	Voting Regressor		0.983714747	0.8986686758	0.9692824747	0.8914604279

County Level	Y = Number of Cases			
Model Name	Model Type		Train	Test
lr_county_cases	Linear Regression (All variables)		0.9649955997	0.9301377526
lr_county_cases_lim	Linear Regression (8variables)		0.9639786858	0.9325020607
rf_gb_county_cases	Random Forest (RF) w/ GradBoost & Random Search CV		0.9592546254	0.9240205265
rf_gb_county_cases	Extra Trees (ET) w/ Random Search CV		0.901956254	0.9039337135

County Level	Y = Number of Deaths			
Model Name	Model Type		Train	Test
lr_county_deaths	Linear Regression (All variables)		0.9311904155	0.9076985159
lr_county_deaths_lim	Linear Regression (8variables)		0.923403073	0.9044371814
rf_gb_county_deaths	Random Forest (RF) w/ GradBoost & Random		0.9944518849	0.909777691

State WhiteBox Vs BlackBox Models

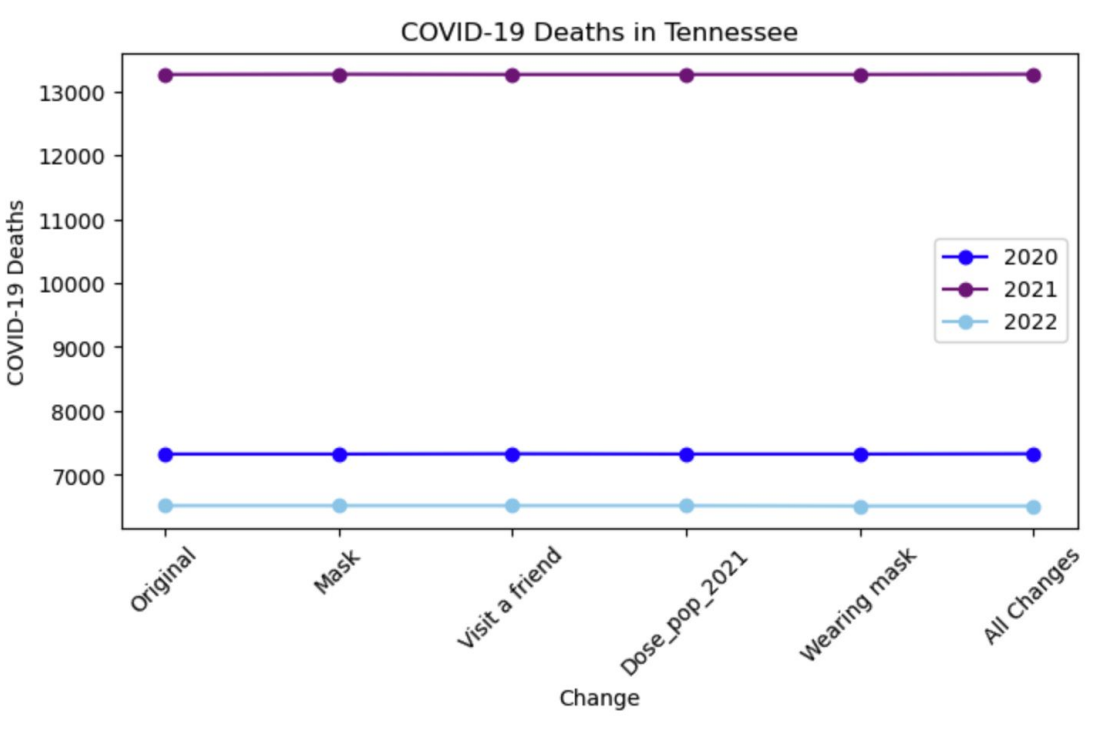
- Two Sets of y variables - Excess Deaths and Number of Covid Deaths
- Key Restriction of only 50 States
- Wide variety of features and each only have a small impact on the predicted deaths
- Use of of MultiOutput Regressor

Model Type	Train Score	Test Score
Linear Regression	1	0.8804713015
Random Forest	0.9592415099	0.8145570907
Bagging	0.9684471928	0.8230630848
Gradient Boosting	0.9999918873	0.9182707939
LASSO	0.99999914	0.9226665135

Production Model Testing

- Taking the 5 States with the highest number of Covid deaths.
- Chose a few parameters to change that had the highest feature importance.
- Taking values from the States that had the least amount of deaths

State	Mask Mandatory	Administered Dose	Wearing mask	Go Visit a Friend



County White Box Models: Linear Regression

Cases (Test Score: 0.93)

Key Variable	Coefficient
percent Native Hawaiian/Other Pacific Islander	-3,673.5
Percent Unemployed	-1,178.9
Average Daily PM2.5	-1,033.1
percent_smokers	-771.1
percent Excessive Drinking	-409.1
percent African American	1,746.4
percent Hispanic	1,876.8
percent Non-Hispanic White	1,960.5
percent American Indian/Alaskan Native	2,064.7
water	2,682.2

Deaths (Test Score: 0.90)

Variable Name	Coefficient
percent Native Hawaiian/Other Pacific Islander	-196.4
Food Environment Index	-30.6
Percent Food Insecure	-23.0
% Fair/Poor Health	-17.7
Inadequate Facilities	-11.3
Percent Insufficient Sleep	27.3
percent Non-Hispanic White	28.1
percent Hispanic	29.2
percent American Indian/Alaskan Native	33.7
water	50.3

County Black Box Models

Cases (Test Score: 0.90)
Extra Trees (ET) w/
Random Search CV

Features	Importance
Population	0.691067
percent Asian	0.059093
Percent Severe Housing Problems	0.055242
Severe Housing Cost Burden	0.023799
Average Daily PM2.5	0.018630
percent Not Proficient in English	0.018506
Overcrowding	0.016577

Deaths (Test Score: 0.91)
Random Forest (RF) w/
GradBoost & Random Search CV

Features	Importance
Population	0.725828
Inadequate Facilities	0.196904
Percent Severe Housing Problems	0.018299
Segregation index black/white	0.012491
Severe Housing Cost Burden	0.009958
% Physically Inactive	0.009101
Income Ratio	0.003917

Initial Findings & Analysis

- No solid takeaways yet from the state-level analysis
- Policies such as mask mandates did have an impact though limited
- The average behaviors and choices of individuals seemed to impact outcomes
- Regional differences
- Inequality and Access

Recommendations

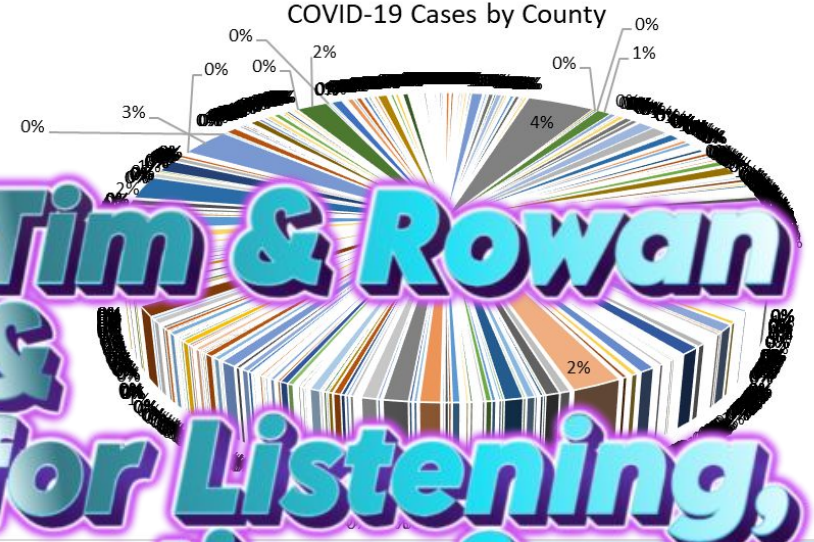
- Reducing racial segregation and inequality
- Increasing healthcare access
- Outreach and education programs to increase trust in public health officials and organizations
- Vaccine Public Health Campaign
- Subsidizing insurance further to reduce rates of uninsured

Next Steps

- Further Data Engineering:
  - Combining similar/collinear variables into index variables
  - Use PCA to reduce dimensionality & combine factors into 1 to 3 variables
- More Data Collection
- Future Models:
  - Time-Series Analysis
  - Clustering

Sources

- Education by state
  https://worldpopulationreview.com/state-rankings/educational-attainment-by-state
  https://data.census.gov/table/ACSST5Y2021.S1501?g=010XX00US$0400000]
  https://data.ers.usda.gov/reports.aspx?ID=17829
- County level data
  https://usafacts.org/visualizations/coronavirus-covid-19-spread-map/state/texas/
- Mask Policy Implementation Dates
  https://www.kaggle.com/datasets/manuvarghese98/impact-of-mask-mandate-on-covid19
- Death Counts by week and state
  https://data.cdc.gov/NCHS/Provisional-COVID-19-Death-Counts-by-Week-Ending-D/r8kw-7aab
- Excess Deaths
  https://www.cdc.gov/nchs/nvss/vsrr/covid19/excess_deaths.htm
- Population per State
  https://www.census.gov/quickfacts/fact/dashboard/US/
- Brainstorm site w Population Density:
  https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9022803/
- State Healthcare Compare
  https://statehealthcompare.shadac.org/Bulk#1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21,22,23,24,25,26,27,28,29,30,31,32,33,34,35,36,37,38,39,40,41,42,43,44,45,46,47,48,49,50,51,52
- Three Datasets- Health Behaviors, Public Health Measures, and Executive Approval
  https://lazerlab.shinyapps.io/Behaviors_During_COVID/
- Covid Vaccines by week by state
  https://data.cdc.gov/Vaccinations/COVID-19-Vaccinations-in-the-United-States-Jurisdi/unsk-b7fc



Thank you Tim & Rowan & Thank you for Listening, Any Questions?