

Philip Ballentine 2020-12-14





**03** FINAL DATASET

04 FINDINGS

05 TABLEAU



## **COVID-19 IMPACTS ON THE US: 2020-12-14**

>15M

Infections

~300K

Deaths



210%

Black Americans' relative risk of COVID-19 death

## **COVID-19 IMPACTS ON THE US: 2020-12-14**

- Close to 300,000 Americans have died from COVID-19, with 2,500-3,000 dying each day from the disease as part of a fall/winter "spike"
- In June 2020, the Federal Reserve forecasted a 6.5% decline in GDP in 2020 amid a "collapse in employment" as a result of the pandemic and public health measures taken to address it (CEPAL, 2020).
- In addition to the death and morbidity caused by the disease, many have pointed to the devastating effects of the economic damage, closed schools, isolation due to public health measures, etc.
- Impacts have fallen disproportionately on Black, Hispanic, immigrant, and other disadvantaged communities due to
  the intersection of environmental racism, occupational hazards, inequalities in social determinants of health, and other
  factors.

"The COVID-19 crisis has also had differential impacts among various racial and ethnic groups. Inequities in the social determinants of health—income and wealth, health-care access and utilization, education, occupation, discrimination, and housing—are interrelated and put some racial and ethnic minority groups at increased risk of contracting and dying from COVID-19.... Black and Hispanic people are dying at much higher rates relative to their share of the U.S. population."

(O'Donnell et al, 2020)

## **COVID-19 QUESTIONS OF INTEREST**

- How much of the mortality toll of COVID-19 can be predicted by county-level characteristics such as the state, the geographic region, how rural or urban the county is, racial/ethnic composition, county median income and poverty rates, how educated the county inhabitants are, etc. while addressing multicollinearity? How have these trends changed over time?
- What is the "best model" to predict "lagged" COVID-19 deaths, or the rate of COVID-19 deaths per 100K people in a county population? What is the relationship between reported cases and "lagged" COVID-19 deaths?
- Is there any evidence in Google Mobility data at the county level that increased time spent at home may be related to reported deaths from COVID-19 per capita at the county level?









### **COVID-19 DATA**

Time series of total reported cases and total reported deaths from COVID-19 at the county level

Characteristics of US counties from the US Census, including ethnic/racial population percentage, income, education, etc.

### **COUNTY DATA**





### **MOBILITY DATA**

Google has made mobility data at the county level available as a time series that shows percentage from baseline activity.

### **COVID-19 DATA**

- Acquire data that is used by the COVID Tracking Project, Johns Hopkins University, using the <u>Datahub</u> package and infrastructure.
- Data is updated daily
- Data is not geocoded, and only has the county and/or city **text name**
- Data is collated by volunteers, and is sometimes inconsistently reported/has corrections
- Variables of interest are date, location, total cases as of that date, and total deaths as of the date

### **COUNTY DATA**

- Data sources from the US Census, particularly the 5-Year Annual Community Survey, last performed in 2018
- Some data sourced using Census-provided API to retrieve information using FIPS codes; other data sourced from the USDA Economic Research Service via flat files
- All data is at county level, as defined by FIPS Code
- 64,000+ variables in the ACS alone
- I selected the variables total population, Black population, Native American population, Hispanic population (not mutually exclusive), percent in poverty, median income for county, and education variables (percent with high school diploma only, percent with Bachelor's degree or higher)

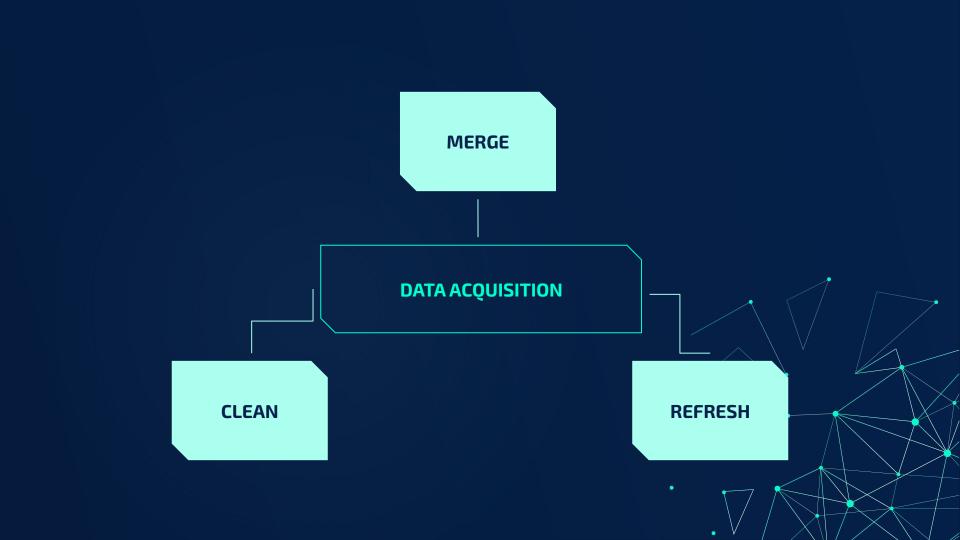
### **GOOGLE MOBILITY DATA**

- Reports from <u>Google from Google maps data</u>, reported aggregated at county level as a time series
- Data is updated daily, I hand-downloaded the CSV
- Data is geocoded using FIPS
- Data is not reported for all counties, and this changes over time
- Data gap for ~50% of counties in September that I reached out to Google to ask about
- Variables of interest are percent change in activity from baseline in various settings (work, shopping, home, etc.) The most consistently reported was time at home spent relative to baseline, which was generally elevated (although not everywhere!)

### **COUNTY SHAPE FILES (FOR TABLEAU)**

- Shape file that uses FIPS to code counties
- Provides polygons and locations for US counties and states
- I hand loaded the .CSV and imported it into Tableau





## **DATA ISSUES**

- Geocoding the county names in the COVID-19 Dataset
  - ◆ Some counties have the same name as the state that they are located in (New York County, Utah County, etc.) and geocoders tended to code these as states, not counties
  - Some areas were included in the COVID-19 dataset that are not at the county level at all (such as at the city level), and in some cases are places that used to be part of different FIPS codes, so the geocoder would sometimes get these codes wrong
- Data types
  - ◄ FIPS codes should be coded as strings and sometimes contain leading zeroes, but they are often implicitly read as integers/floats in both Excel, Python, and R leading to leading zero loss without explicitly altering their data type during import
- Geographic Data consistency
  - ◆ Although Puerto Rico was included in the COVID-19 data, and is available in many Census datasets, I was only able to find comprehensive data for educational attainment, poverty, and rural-urban divides in flat file format from the USDA Economic Research Service, which omitted Puerto Rico in their poverty estimates, which I used as the backbone of my counties dataset.

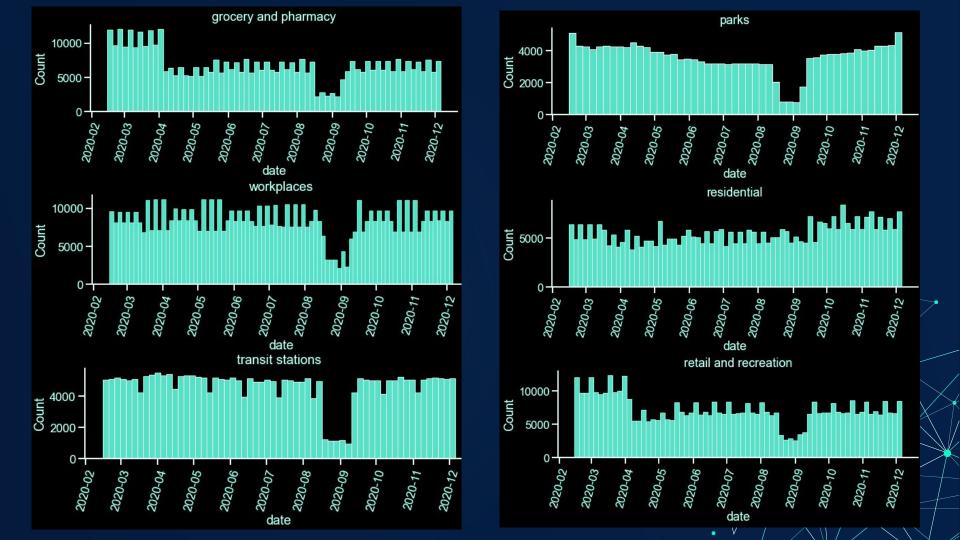
## **DATA ISSUES**

### API performance

- Using the Census API, the average call/response time for a single variable for a single county that I clocked was about 1.5 seconds, so for roughly 3300 counties, it took about 80 minutes to look up a single variable
- Looking up several variables took many hours (overnight)
- ◆ This illustrates the performance benefit of using a .CSV for this purpose, but finding .CSVs was difficult.

### Google Mobility Data

- Google mobility data was not available for every county for every day
- ◆ This may have been because there were not enough Google users in that county on a specific day to a) protect the privacy of users and b) calculate valid measurements
- ◆ There was also a "revision" made to the files which I found may have resulted in a fair amount of missing data during mid-August into early September
- I reached out to Google via Twitter and email (thanks Vivian!) but since this was late in the game, was never able to determine what the issue was/is
- ◆ The residential change from baseline was a) the most reliably populated, and b) unaffected by this issue, so I ended up using it in my regression to avoid dropping tons of data



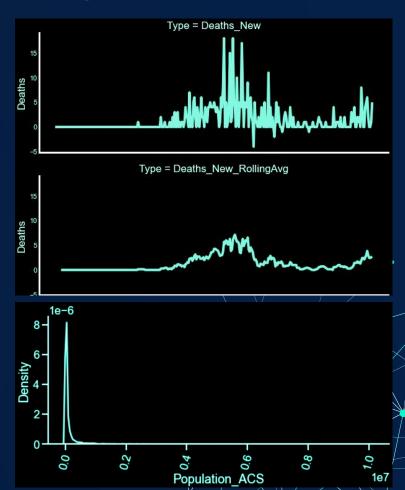
### **TOOLS USED**

- Data Wrangling Python Jupyter Notebooks
  - Download the day's COVID data (final analysis done on 2020-12-09)
  - Import mobility data using flat files
  - Import county data using flat files and API
  - Geocode and clean the COVID dataset
  - Merge the datasets together
  - Unit tests using Pytest to ensure that counties weren't duplicated or missing
  - Subset different datasets to feed into Tableau
- Analysis Python JuPyter Notebook using R-style regression
  - Assess multicollinearity using VIF and corrplot
  - Python implementation of all-subsets regression
  - Selecting models using BIC and Adjusted R-Squared as well as logic and judgment
- Visualization Python using Seaborn, Tableau
  - Custom color scheme to match this presentation

```
for i in list_frames:
    plt.rcParams.update({"grid.linewidth":0.5, "grid.alpha":0.5})
    sns.set(style="ticks", context="talk")
    plt.style.use("dark_background")
    sns.displot(i[0].date, height=3, aspect=3, color="#00ffcd")
    plt.xticks(rotation=75)
    plt.title("{x}".format(x=i[1]))
```

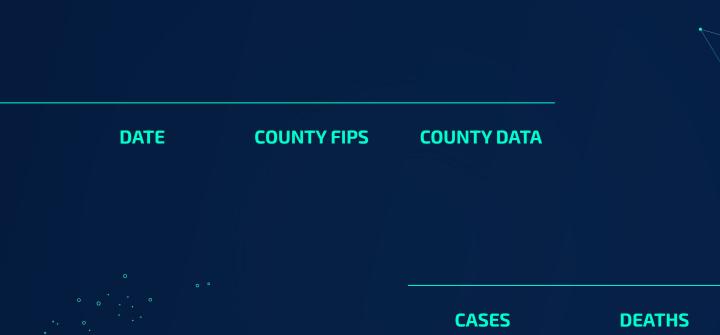
## **DATA TRANSFORMATION**

- Data change over time
  - New deaths
  - New cases
- Rolling Averages
  - Helps to eliminate spikiness and data anomalies (negative counts due to revisions, holidays, weekend effects, etc.)
- Data lag
  - Deaths 7, 14, 21 days after the cases reported on day T
- Per capita measures
  - Deaths, cases, by 100K population
  - ◆ Population by county is very right-tailed, with a median of 25K, a mean of 101K, and a standard deviation of 325K
  - Raw measures by themselves are not always very helpful, but inherent tension here because reporting by 100K population can also skew human impact





## **VARIABLES**





**MOBILITY** 

## 1.01 M

Rows / Observations

68

Columns / Variables

3130

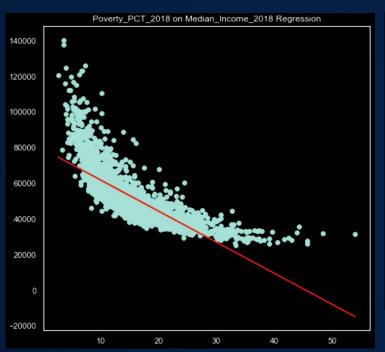
Counties represented in the data



# O4 FINDINGS



## **MULTICOLLINEARITY / NONLINEARITY**



		OLS Regre	ssion Re	sults		
Dep. Variable:		y	R-squ	======= ared:		0.594
Model:		OLS	_	R-squared:		0.594
Method:	Lea	st Squares		tistic:		4574.
Date:	Mon, 1	.4 Dec 2020	Prob	(F-statisti	c):	0.00
Time:		10:25:36	Log-L	ikelihood:		-32886.
No. Observations:		3131	AIC:			6.578e+04
Df Residuals:		3129	BIC:			6.579e+04
Df Model:		1				
Covariance Type:		nonrobust				
	oef st	d err	t	P> t	[0.025	0.975]
const 7.914e	+04 42	20.792 1	88.083	0.000	7.83e+04	8e+04
x1 -1739.3	686 2	25.718 -	67.632	0.000	-1789.795	-1688.942
======================================	=======	 1509.945	Durbi	======= n-Watson:	=======	1.563
Prob(Omnibus):		0.000	Jarqu	e-Bera (JB)	:	10807.420
Skew:		2.180	Prob(	JB):		0.00
Kurtosis:		10.990	Cond.	No.		43.8
Kurtosis:	======	10.990	Cond.	No.		43.8

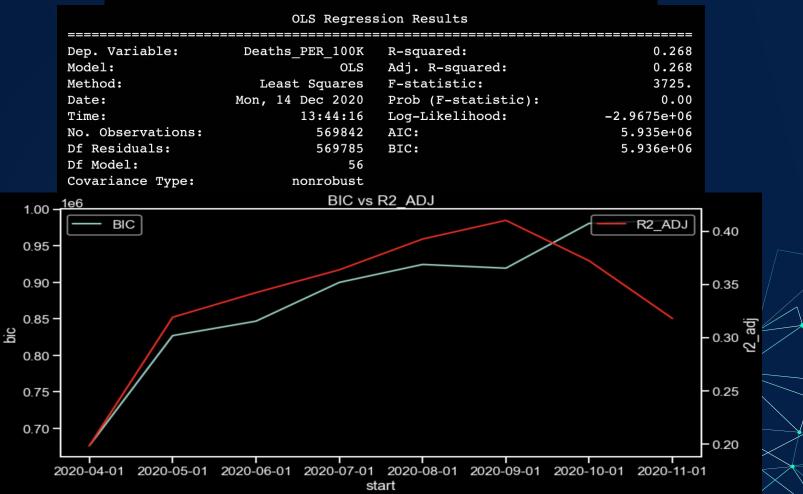
								- 1.0
PCT_Black_ACS	1.00	-0.10	-0.11	-0.10	0.02	-0.26	0.49	- 0.8
PCT_Native_ACS	-0.10	1.00	-0.03	-0.08	0.02	-0.10	0.24	- 0.6
PCT_Hisp_ACS	-0.11	-0.03	1.00	-0.01	-0.30	0.04	0.09	- 0.4
BACHELORS_PLUS_2018	-0.10	-0.08	-0.01	1.00	-0.77	0.72	-0.48	- 0.2 - 0.0
HS_DIPLOMA_ONLY_2018	0.02	0.02	-0.30	-0.77	1.00	-0.54	0.29	0.2
Median_Income_2018	-0.26	-0.10	0.04	0.72	-0.54	1.00	-0.77	0.4
Poverty_PCT_2018	0.49	0.24	0.09	-0.48	0.29	-0.77	1.00	0.6
	PCT_Black_ACS	PCT_Native_ACS	PCT_Hisp_ACS	BACHELORS_PLUS_2018	HS_DIPLOMA_ONLY_2018	Median_Income_2018	Poverty_PCT_2018	_

features	VIF Factor	
HS_DIPLOMA_ONLY_2018	13.00	62
BACHELORS_PLUS_2018	12.70	61
Median_Income_2018	8.80	63
Poverty_PCT_2018	5.40	64
PCT_Hisp_ACS	3.10	60
PCT_Black_ACS	2.70	58
Metro[T.Smaller_metro]	2.10	4
Metro[T.Micropolitan]	2.10	1
PCT_Native_ACS	1.60	59
residential_PCT_CFB_RollingAvg	1.40	65
Metro[T.Noncore_adjacent]	1.20	2
Metro[T.Noncore_remote]	1.10	3
Intercept	0.00	0

## **COVID-19 QUESTIONS OF INTEREST**

■ How much of the mortality toll of COVID-19 can be predicted by county-level characteristics such as the state, the geographic region, how rural or urban the county is, racial/ethnic composition, county median income and poverty rates, how educated the county inhabitants are, etc. while addressing multicollinearity? How have these trends changed over time?

	bic	aic	r2_adj	num_var
formula				
Deaths_PER_100K ~ C(Metro) + PCT_Black_ACS + BACHELORS_PLUS_2018 + Poverty_PCT_2018 + Median_Income_2018 + C(StateCD)	5934307 63	5933643.70	0.27	6
Deaths_PER_100K ~ C(Metro) + PCT_Black_ACS + Poverty_PCT_2018 + Median_Income_2018 + C(StateCD)	5935424.33	5934771.65	0.27	5
Deaths_PER_100K ~ C(Metro) + PCT_Black_ACS + BACHELORS_PLUS_2018 + Poverty_PCT_2018 + C(StateCD)	5935511.80	5934859.12	0.27	5
Deaths_PER_100K ~ PCT_Black_ACS + BACHELORS_PLUS_2018 + Poverty_PCT_2018 + Median_Income_2018 + C(StateCD)	5935485.40	5934866.48	0.27	5
Deaths_PER_100K ~ C(Metro) + PCT_Black_ACS + Poverty_PCT_2018 + C(StateCD)	5935811.75	5935170.32	0.27	4
Deaths_PER_100K ~ PCT_Black_ACS + Poverty_PCT_2018 + Median_Income_2018 + C(StateCD)	5936775.65	5936167.98	0.27	4
Deaths_PER_100K ~ PCT_Black_ACS + BACHELORS_PLUS_2018 + Poverty_PCT_2018 + C(StateCD)	5938020.65	5937412.98	0.27	4
Deaths_PER_100K ~ PCT_Black_ACS + Poverty_PCT_2018 + C(StateCD)	5938067.90	5937471.48	0.26	3



### Deaths\_PER\_100K ~ C(Metro) + PCT\_Black\_ACS + Poverty\_PCT\_2018 + C(StateCD)

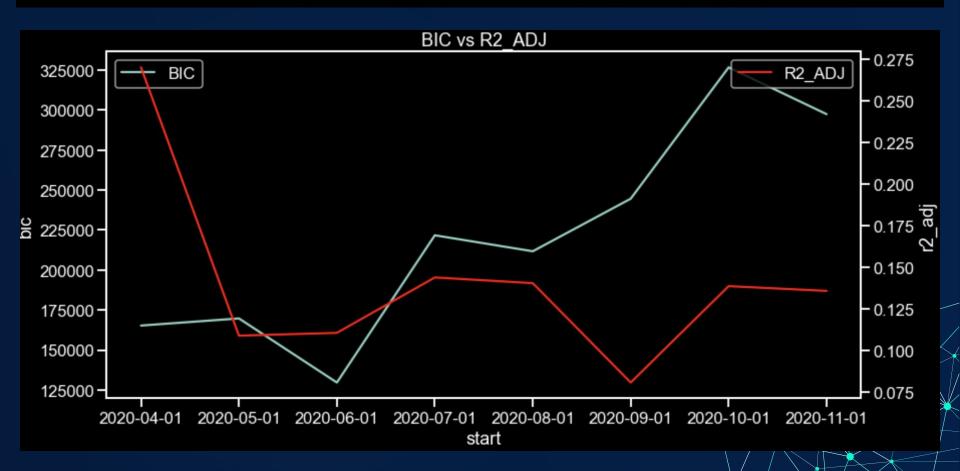
Dep. Variable:	Deaths_PER_100					268	
Model:	OL.		Adj. R-squared:			268	
Method:	Least Square		tistic:			725.	
Date:	Mon, 14 Dec 202			istic):		0.00	
Time:	16:51:3		ikeliho	ood:	-2.9675		
No. Observations: Df Residuals:	56984				5.935		
Df Model:	56978 5				5.936	+00	
Covariance Type:	nonrobus						
	C	oef st	d err	t	P> t	[0.025	0.975]
Intercept	-2.0	887 (	0.667	-3.132	0.002	-3.396	-0.782
C(Metro)[T.Micropoli	itan] -7.7	215 (	2.229	-33.735	0.000	-8.170	-7.273
C(Metro)[T.Noncore_a	adjacent] -7.7	845 (	0.220	-35.367	0.000	-8.216	-7.353
C(Metro)[T.Noncore_r	remote] -12.7	244 (	0.291	-43.683	0.000	-13.295	-12.153
C(Metro)[T.Smaller_m	netro] -8.7	221 (	0.213	-40.913	0.000	-9.140	-8.304
C(StateCD)[T.AZ]	65.9	668 :	1.065	61.915	0.000	63.879	68.055
C(StateCD)[T.CT]	81.9	<b>050</b> :	1.325	61.817	0.000	79.308	84.502
C(StateCD)[T.DE]	45.1	777 2	2.001	22.575	0.000	41.255	49.100
C(StateCD)[T.LA]	56.5	112	<b>0.</b> 782	72.233	0.000	54.978	58.045
C(StateCD)[T.MA]	91.8		1.084	84.682	0.000	89.693	93.943
C(StateCD)[T.MS]	46.3		0.772	60.107	0.000	44.869	47.894
C(StateCD)[T.NJ]	136.4		0.968	140.872	0.000	134.521	138.317
C(StateCD)[T.NY]	43.7		764	57.273	0.000	42.256	45.251
C(StateCD)[T.RI]	48.7	492	1.607	30.328	0.000	45.599	51.900
OMMITTING STAT	TES WITH COEFS LE	SS THAN +4	40				
Poverty_PCT_2018	1.1		0.014	82.477	0.000	1.112	1.166
PCT_Black_ACS ========	0.8 	315 ( =======	0.007 	124.151 	0.000 	0.818 ====	0.845
Omnibus:	371955.91		n-Watso			030	
Prob(Omnibus):	0.00		e-Bera	(JR):	7991721		
Skew:	2.82					0.00	
Kurtosis: =======	20.45 	8 Cond.	NO.		1.70e	====	

Notes: [1] Standard Errors assume that the covariance matrix of the errors is correctly spec [2] The condition number is large, 1.7e+03. This might indicate that there are strong multicollinearity or other numerical problems.

## **COVID-19 QUESTIONS OF INTEREST**

■ What is the "best model" to predict "lagged" COVID-19 deaths, or the rate of COVID-19 deaths per 100K people in a county population? What is the relationship between reported cases and "lagged" COVID-19 deaths?

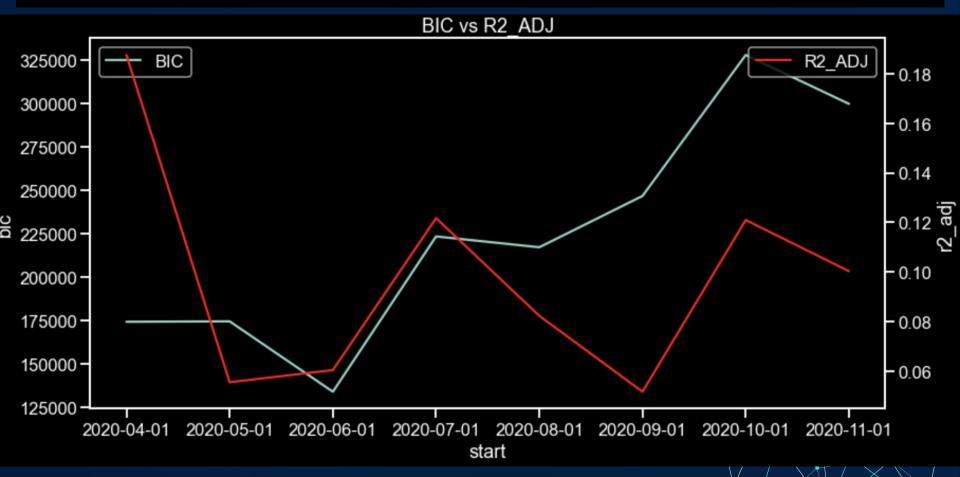




### Deaths\_New\_14\_RollingAvg\_PER\_100K ~ C(Metro) + Confirmed\_New\_RollingAvg\_PER\_100K + C(StateCD)

Dep. Variable: De	aths_New_14_Rolli	ngAvg_Pl	ER_100K	R-squared:		0.14	0
Model:		J J_		Adj R-squared	l:	0.14	0
Method:		Least S		F-statistic:		1644	•
Date:	Mo	n, 14 De	ec 2020	Prob (F-statis	tic):	0.0	0
Time:		10	6:23:51	Log-Likelihood	l:	-8.3027e+0	5
No. Observations:			554187	AIC:		1.661e+0	6
Df Residuals:			554131	BIC:		1.661e+0	6
Df Model:			55				
Covariance Type:		noi	nrobust				
		coef	std err	======== t	P> t	[0.025	0.975]
Intercept		 -0.2587	0.016	-15.772	0.000	-0.291	 -0.227
C(Metro)[T.Micropolit		0.1088	0.005	20.355	0.000	0.098	0.119
C(Metro)[T.Noncore_ad		0.1732	0.005	34.638	0.000	0.163	0.183
C(Metro)[T.Noncore_re		0.1790	0.007		0.000	0.166	0.192
C(Metro)[T.Smaller_me		0.0577	0.005		0.000	0.048	0.068
Confirmed_New_Rolling	Avg_PER_100K	0.0121	4.53e-05	266.321	0.000	0.012	0.012
	======================================	====== Durbiı	======= n–Watson:	=========	0.256		
Prob(Omnibus):	0.000		e-Bera (JB	): 8660	35413.769		
Skew:	9.092	Prob(			0.00		
Kurtosis:	195.807	Cond.			3.13e+03		

- [1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
- [2] The condition number is large, 3.13e+03. This might indicate that there are strong multicollinearity or other numerical problems.



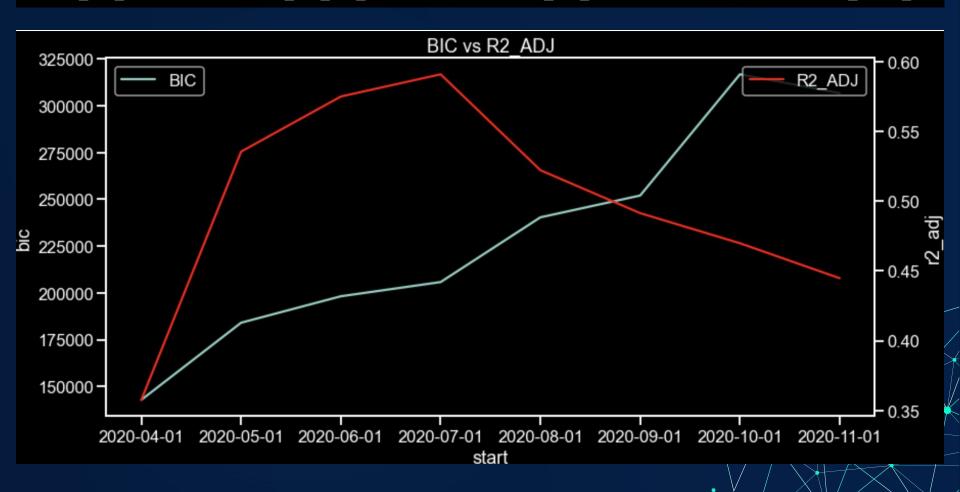
### Deaths\_New\_14\_RollingAvg\_PER\_100K ~ C(Metro) + Confirmed\_New\_RollingAvg\_PER\_100K

	OL	S Regre	ssion Resu	lts 			
Dep. Variable: De	aths_New_14_Rolli	naAva P	ER 100K	R-squared:		0.127	
Model:		J J		Adj. R-squared	:	0.127	
Method:		Least		F-statistic:		1.619e+04	
Date:	Mo			Prob (F-statis	tic):	0.00	
Time:		1		Log-Likelihood		-8.3438e+05	
No. Observations:			554187	AIC:		1.669e+06	
Df Residuals:			554181	BIC:		1.669e+06	
Df Model:			5				
Covariance Type:		no	nrobust				
============	=======================================		=======	=========	=======		======
		coef	std err	t	P> t	[0.025	0.975]
Intercept		0.0482	0.004	11.982	0.000	0.040	0.056
C(Metro)[T.Micropolit		0.0861	0.005	16.801	0.000	0.076	0.096
C(Metro)[T.Noncore_ad	ljacent]	0.1587	0.005	33.618	0.000	0.149	0.168
C(Metro)[T.Noncore_re	emote]	0.1256	0.006	20.436	0.000	0.114	0.138
C(Metro)[T.Smaller_me	tro]	0.0495	0.005	9.924	0.000	0.040	0.059
Confirmed_New_Rolling	Avg_PER_100K	0.0123	4.41e-05	278.886	0.000	0.012	0.012
Omnibus:	833597.351	Durbi	n-Watson:		0.253		
Prob(Omnibus):	0.000	Jarqu	e-Bera (JB	): 8351	28245.965		
Skew:	9.047	Prob(.	JB):		0.00		
Kurtosis:	192.313	Cond.	No.		255.		

## **COVID-19 QUESTIONS OF INTEREST**

■ Is there any evidence in Google Mobility data at the county level that increased time spent at home may be related to reported deaths from COVID-19 per capita at the county level?

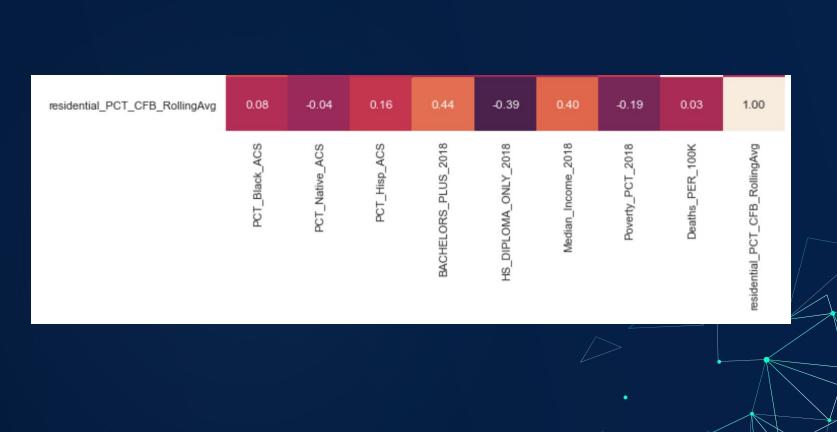




Dep. Variable:	Deaths_PER_100K		quared:		0.39		
Model:	OLS		. R-squared	:	0.39		
Method:	Least Squares		tatistic:		2011		
Date:	Mon, 14 Dec 2020		b (F-statis		0.0		
Time:	16:43:59 161394		-Likelihood -	:	-8.0615e+0		
No. Observations: Df Residuals:	161344	AIC BIC			1.612e+0 1.613e+0		
Df Model:	161340	DIC			1.0136+6	00	
Covariance Type:	nonrobust						
	CC	oef	std err	t	P> t	[0.025	0.975]
Intercept	-24.48	891	1.573	-15 <b>.</b> 570	0.000	-27 <b>.</b> 572	-21.406
C(StateCD)[T.CT]	79.33	385	1.805	43.964	0.000	75.802	82.876
C(StateCD)[T.DE]	44.63	101	2.177	20.491	0.000	40.343	48.877
C(StateCD)[T.LA]	<b>57.</b> 53		1.686	34.107	0.000	54.209	60.819
C(StateCD)[T.MA]	103.49		1.734	59.696	0.000	100.092	106.888
C(StateCD)[T.NY]	57.13		1.602	35.663	0.000	53.994	60.274
C(StateCD)[T.RI]	63.33	194	2.093	30.248	0.000	59.217	67.422
OMMITTING STATE	ES WITH COEFS LESS TH	HAN +	40				
residential_PCT_CFE	3_RollingAvg 1.56	540	0.032	48.717	0.000	1.501	1.627
Poverty_PCT_2018	2.1	519	0.024	89.496	0.000	2.105	2.199
PCT_Black_ACS	0.53	107	0.011	47.544	0.000	0.490	0.532
Omnibus:	59679 <b>.</b> 520	Dur	======= bin-Watson:		 0.02	== 25	
Prob(Omnibus):	0.000	Jaro	que-Bera (J	B):	419925.59	92	
Skew:	1.610	Pro	b(JB):		0.0	00	
Kurtosis:	10.216	Cond	d. No.		2.57e+6	93	

#### Notes:

- [1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
- [2] The condition number is large, 2.57e+03. This might indicate that there are strong multicollinearity or other numerical problems.

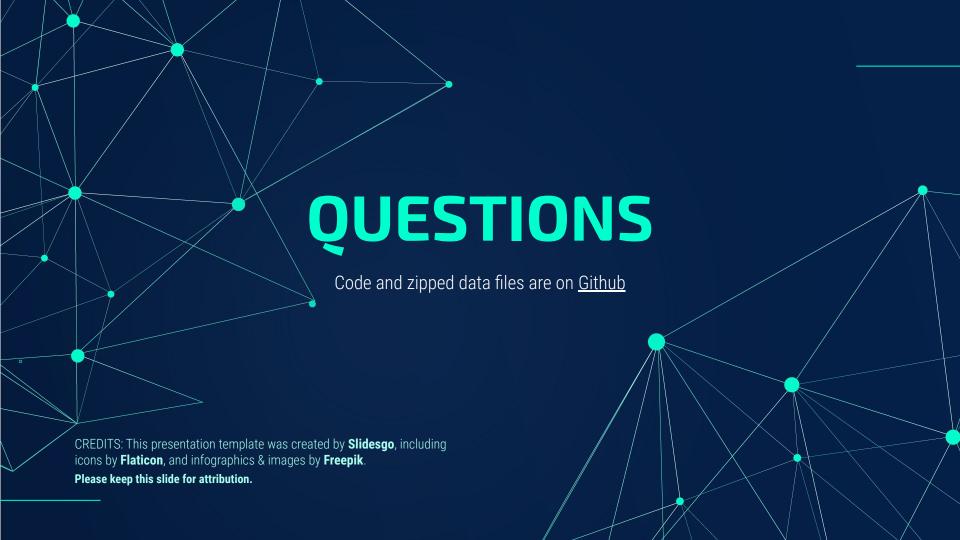


## **SUMMARY OF FINDINGS**

- Multicollinearity between many measures poses challenges to analysis
- Data at a county level is rather imprecise
  - Ex: Boston, Newton, Lowell, and Weston are in the same county but are very different communities
- Percentage of population identifying as Black is a significant predictor of death toll for COVID-19, although this
  measure is correlated with other measures that also act as predictors for higher per capita death toll from COVID-19
  - A model with multiple interactions would be a good next step to analyze this
- There are omitted variables in this analysis, such as the rate of true disease prevalence, the rate of hospitalization, different rates of reporting cases and deaths, and many others
- 100 reported cases seem to predict roughly .5 1.5 deaths from COVID reported 14-21 days later, and this seems pretty stable over time since approximately June 1
- Many models of deaths (either new deaths or total death toll from COVID-19 show an increase in R squared up until the early Fall of 2020, and then falls in R squared later. This may be due to a more widespread disease toll that is affecting nearly all US counties more "equally" than before (at the county level)

# O5 TABLEAU





## **slides**go

## Fonts & colors used

This presentation has been made using the following fonts:

### Exo 2

(https://fonts.google.com/specimen/Exo+2)

### **Roboto Condensed**

(https://fonts.google.com/specimen/Roboto+Condensed)

