EFFECTS OF AGE, LOCATION AND ETHNICITY ON CHRONIC DISEASES

KARMISHTHA SETH

JULY 18, 2020

1. INTRODUCTION

The CDC classifies Chronic Diseases as "...conditions that last 1 year or more and require ongoing medical attention or limit activities of daily living or both", common examples being cancer, heart disease, diabetes and asthma. Common causes for such chronic diseases are tobacco usage, excessive alcohol intake, poor nutrition etc. Chronic diseases impact almost every human being's life, and in a major way. Studies have shown that chronic diseases are quite biased in terms of the groups that are targeted by them. A study done in the UK, testing effects of race on diabetes and chronic kidney disease showed that there was a significant increase in the aforementioned diseases in the Black and South Asian community when compared to Caucasians[1]. Another study performed by Italian General Practitioners looked at how ageing had an effect on chronic disease, and noted that there is a significant increase in the prevalence of chronic diseases after the age of 60[2]. Looking at such data, I was motivated to look at what effect age, ethnicity and location have on chronic disease prevalence.

COVID-19 has taken over the entire world in a mere 6 months, has caused multiple countries to shut down, lock borders, caused the market to be lower than it has been in decades and has even accelerated the speed of recession. While the effects of it will be long lasting, I am very intrigued to see how this virus disproportionately affects minority communities, and older age-d people.

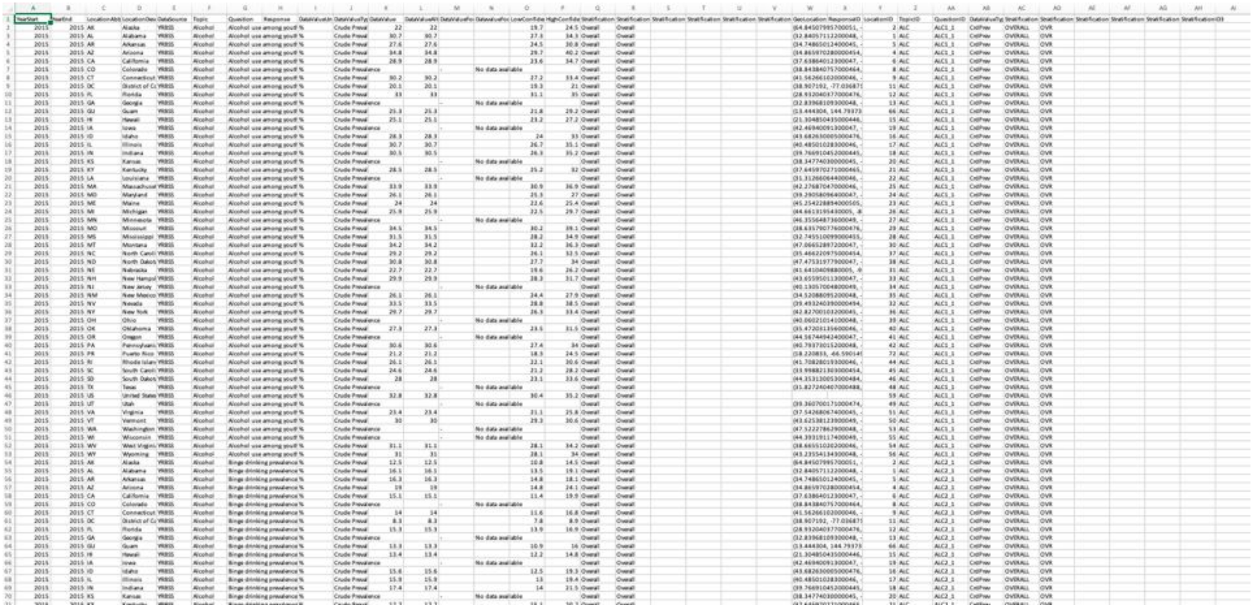
Dr Chaand Nagpaul, the British Medical Association (BMA) council chair and a general practitioner (GP) in north London, UK said, "However, it is a clear and consistent theme from the reports and what we know about those who have died so far, that a disproportionate number of those health-care workers who have tragically lost their lives are from BAME communities.[3]"

In addition COVID-19 shows an increased number of cases and a greater risk of severe disease with increasing age[4] with it disproportionately fatally affecting people over the age of 50 more than those younger. Inspired by this, I wanted to apply my results from analysing the effects of stratification on chronic diseases to see if it followed a similar format to deaths due to COVID-19.

2. METHODS

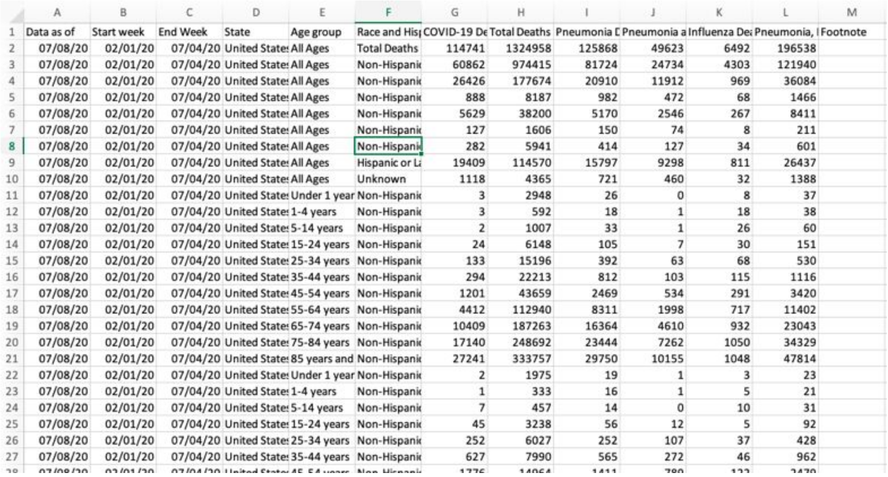
I collected my initial data set for chronic diseases was a CDC dataset collected from Kaggle, labelled "U.S._Chronic_Disease_Indicators"[5] (Figure 1.) from which I used the variables "LocationAbbr", "Stratification1", "Topic", "LocationDesc", "DataValue".

The second dataset, i.e. my COVID-19 dataset was another CDC dataset collected from Kaggle, labelled "Provisional COVID-19 Death Counts by Sex, Age, and State"[6] (Figure 2) from which I used the variables "Age group", "Race and Hispanic Origin Group", "State" and "COVID-19 Deaths".



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB	AC	AD	AE	AF	AG	AH
1	YearStart	YearEnd	LocationAbb	LocationDesc	DataSource	Topic	Question	Response	DataValueUn	DataValueTyp	DataValue	DataValueAlt	DataValueFo	DatavalueFoo	LowConfiden	HighConfiden	Stratification	Stratification	Stratification	Stratification	Stratification	Stratification	GeoLocation	ResponseID	LocationID	TopicID	QuestionID	DataValueTyp	Stratification	Stratification	Stratification	Stratification	Stratification	StratificationID3
2	2015	2015	AK	Alaska	YRBSS	Alcohol	Alcohol use among youth		%	Crude Preval	22	22			19.7	24.5	Overall	Overall					(64.84507995700051, -		2	ALC	ALC1_1	CrdPrev	OVERALL	OVR				
3	2015	2015	AL	Alabama	YRBSS	Alcohol	Alcohol use among youth		%	Crude Preval	30.7	30.7			27.3	34.3	Overall	Overall					(32.84057112200048, -		1	ALC	ALC1_1	CrdPrev	OVERALL	OVR				
4	2015	2015	AR	Arkansas	YRBSS	Alcohol	Alcohol use among youth		%	Crude Preval	27.6	27.6			24.5	30.8	Overall	Overall					(34.74865012400045, -		5	ALC	ALC1_1	CrdPrev	OVERALL	OVR				
5	2015	2015	AZ	Arizona	YRBSS	Alcohol	Alcohol use among youth		%	Crude Preval	34.8	34.8			29.7	40.2	Overall	Overall					(34.865970280000454,		4	ALC	ALC1_1	CrdPrev	OVERALL	OVR				
6	2015	2015	CA	California	YRBSS	Alcohol	Alcohol use among youth		%	Crude Preval	28.9	28.9			23.6	34.7	Overall	Overall					(37.63864012300047, -		6	ALC	ALC1_1	CrdPrev	OVERALL	OVR				
7	2015	2015	CO	Colorado	YRBSS	Alcohol	Alcohol use among youth		%	Crude Prevalence			-	No data available			Overall	Overall					(38.843840757000464,		8	ALC	ALC1_1	CrdPrev	OVERALL	OVR				
8	2015	2015	CT	Connecticut	YRBSS	Alcohol	Alcohol use among youth		%	Crude Preval	30.2	30.2			27.2	33.4	Overall	Overall					(41.56266102000046, -		9	ALC	ALC1_1	CrdPrev	OVERALL	OVR				
9	2015	2015	DC	District of Co	YRBSS	Alcohol	Alcohol use among youth		%	Crude Preval	20.1	20.1			19.3	21	Overall	Overall					(38.907192, -77.036871		11	ALC	ALC1_1	CrdPrev	OVERALL	OVR				
10	2015	2015	FL	Florida	YRBSS	Alcohol	Alcohol use among youth		%	Crude Preval	33	33			31.1	35	Overall	Overall					(28.932040377000476,		12	ALC	ALC1_1	CrdPrev	OVERALL	OVR				
11	2015	2015	GA	Georgia	YRBSS	Alcohol	Alcohol use among youth		%	Crude Prevalence			-	No data available			Overall	Overall					(32.83968109300048, -		13	ALC	ALC1_1	CrdPrev	OVERALL	OVR				
12	2015	2015	GU	Guam	YRBSS	Alcohol	Alcohol use among youth		%	Crude Preval	25.3	25.3			21.8	29.2	Overall	Overall					(13.444304, 144.79373		66	ALC	ALC1_1	CrdPrev	OVERALL	OVR				
13	2015	2015	HI	Hawaii	YRBSS	Alcohol	Alcohol use among youth		%	Crude Preval	25.1	25.1			23.2	27.2	Overall	Overall					(21.304850435000446,		15	ALC	ALC1_1	CrdPrev	OVERALL	OVR				
14	2015	2015	IA	Iowa	YRBSS	Alcohol	Alcohol use among youth		%	Crude Prevalence			-	No data available			Overall	Overall					(42.46940091300047, -		19	ALC	ALC1_1	CrdPrev	OVERALL	OVR				
15	2015	2015	ID	Idaho	YRBSS	Alcohol	Alcohol use among youth		%	Crude Preval	28.3	28.3			24	33	Overall	Overall					(43.682630005000476,		16	ALC	ALC1_1	CrdPrev	OVERALL	OVR				
16	2015	2015	IL	Illinois	YRBSS	Alcohol	Alcohol use among youth		%	Crude Preval	30.7	30.7			26.7	35.1	Overall	Overall					(40.48501028300046, -		17	ALC	ALC1_1	CrdPrev	OVERALL	OVR				
17	2015	2015	IN	Indiana	YRBSS	Alcohol	Alcohol use among youth		%	Crude Preval	30.5	30.5			26.3	35.2	Overall	Overall					(39.766910452000445,		18	ALC	ALC1_1	CrdPrev	OVERALL	OVR				
18	2015	2015	KS	Kansas	YRBSS	Alcohol	Alcohol use among youth		%	Crude Prevalence			-	No data available			Overall	Overall					(38.34774030000045, -		20	ALC	ALC1_1	CrdPrev	OVERALL	OVR				
19	2015	2015	KY	Kentucky	YRBSS	Alcohol	Alcohol use among youth		%	Crude Preval	28.5	28.5			25.2	32	Overall	Overall					(37.645970271000465,		21	ALC	ALC1_1	CrdPrev	OVERALL	OVR				
20	2015	2015	LA	Louisiana	YRBSS	Alcohol	Alcohol use among youth		%	Crude Prevalence			-	No data available			Overall	Overall					(31.31266064400046, -		22	ALC	ALC1_1	CrdPrev	OVERALL	OVR				
21	2015	2015	MA	Massachuset	YRBSS	Alcohol	Alcohol use among youth		%	Crude Preval	33.9	33.9			30.9	36.9	Overall	Overall					(42.27687047000046, -		25	ALC	ALC1_1	CrdPrev	OVERALL	OVR				
22	2015	2015	MD	Maryland	YRBSS	Alcohol	Alcohol use among youth		%	Crude Preval	26.1	26.1			25.3	27	Overall	Overall					(39.29058096400047, -		24	ALC	ALC1_1	CrdPrev	OVERALL	OVR				
23	2015	2015	ME	Maine	YRBSS	Alcohol	Alcohol use among youth		%	Crude Preval	24	24			22.6	25.4	Overall	Overall					(45.254228894000505,		23	ALC	ALC1_1	CrdPrev	OVERALL	OVR				
24	2015	2015	MI	Michigan	YRBSS	Alcohol	Alcohol use among youth		%	Crude Preval	25.9	25.9			22.5	29.7	Overall	Overall					(44.66131954300005, -8		26	ALC	ALC1_1	CrdPrev	OVERALL	OVR				
25	2015	2015	MN	Minnesota	YRBSS	Alcohol	Alcohol use among youth		%	Crude Prevalence			-	No data available			Overall	Overall					(46.35564873600049, -		27	ALC	ALC1_1	CrdPrev	OVERALL	OVR				
26	2015	2015	MO	Missouri	YRBSS	Alcohol	Alcohol use among youth		%	Crude Preval	34.5	34.5			30.2	39.1	Overall	Overall					(38.635790776000476,		29	ALC	ALC1_1	CrdPrev	OVERALL	OVR				
27	2015	2015	MS	Mississippi	YRBSS	Alcohol	Alcohol use among youth		%	Crude Preval	31.5	31.5			28.2	34.9	Overall	Overall					(32.745510099000455,		28	ALC	ALC1_1	CrdPrev	OVERALL	OVR				
28	2015	2015	MT	Montana	YRBSS	Alcohol	Alcohol use among youth		%	Crude Preval	34.2	34.2			32.2	36.3	Overall	Overall					(47.06652897200047, -		30	ALC	ALC1_1	CrdPrev	OVERALL	OVR				
29	2015	2015	NC	North Caroli	YRBSS	Alcohol	Alcohol use among youth		%	Crude Preval	29.2	29.2			26.1	32.5	Overall	Overall					(35.466220975000454,		37	ALC	ALC1_1	CrdPrev	OVERALL	OVR				
30	2015	2015	ND	North Dakot	YRBSS	Alcohol	Alcohol use among youth		%	Crude Preval	30.8	30.8			27.7	34	Overall	Overall					(47.47531977900047, -		38	ALC	ALC1_1	CrdPrev	OVERALL	OVR				
31	2015	2015	NE	Nebraska	YRBSS	Alcohol	Alcohol use among youth		%	Crude Preval	22.7	22.7			19.6	26.2	Overall	Overall					(41.64104098800005, -9		31	ALC	ALC1_1	CrdPrev	OVERALL	OVR				
32	2015	2015	NH	New Hampsl	YRBSS	Alcohol	Alcohol use among youth		%	Crude Preval	29.9	29.9			28.3	31.7	Overall	Overall					(43.65595011300047, -		33	ALC	ALC1_1	CrdPrev	OVERALL	OVR				
33	2015	2015	NJ	New Jersey	YRBSS	Alcohol	Alcohol use among youth		%	Crude Prevalence			-	No data available			Overall	Overall					(40.13057004800049, -		34	ALC	ALC1_1	CrdPrev	OVERALL	OVR				
34	2015	2015	NM	New Mexico	YRBSS	Alcohol	Alcohol use among youth		%	Crude Preval	26.1	26.1			24.4	27.9	Overall	Overall					(34.52088095200048, -		35	ALC	ALC1_1	CrdPrev	OVERALL	OVR				
35	2015	2015	NV	Nevada	YRBSS	Alcohol	Alcohol use among youth		%	Crude Preval	33.5	33.5			28.8	38.5	Overall	Overall					(39.493240390000494,		32	ALC	ALC1_1	CrdPrev	OVERALL	OVR				
36	2015	2015	NY	New York	YRBSS	Alcohol	Alcohol use among youth		%	Crude Preval	29.7	29.7			26.3	33.4	Overall	Overall					(42.82700103200045, -		36	ALC	ALC1_1	CrdPrev	OVERALL	OVR				
37	2015	2015	OH	Ohio	YRBSS	Alcohol	Alcohol use among youth		%	Crude Prevalence			-	No data available			Overall	Overall					(40.06021014100048, -		39	ALC	ALC1_1	CrdPrev	OVERALL	OVR				
38	2015	2015	OK	Oklahoma	YRBSS	Alcohol	Alcohol use among youth		%	Crude Preval	27.3	27.3			23.5	31.5	Overall	Overall					(35.47203135600046, -		40	ALC	ALC1_1	CrdPrev	OVERALL	OVR				
39	2015	2015	OR	Oregon	YRBSS	Alcohol	Alcohol use among youth		%	Crude Prevalence			-	No data available			Overall	Overall					(44.56744942400047, -		41	ALC	ALC1_1	CrdPrev	OVERALL	OVR				
40	2015	2015	PA	Pennsylvani	YRBSS	Alcohol	Alcohol use among youth		%	Crude Preval	30.6	30.6			27.4	34	Overall	Overall					(40.79373015200048, -		42	ALC	ALC1_1	CrdPrev	OVERALL	OVR				
41	2015	2015	PR	Puerto Rico	YRBSS	Alcohol	Alcohol use among youth		%	Crude Preval	21.2	21.2			18.3	24.5	Overall	Overall					(18.220833, -66.590149		72	ALC	ALC1_1	CrdPrev	OVERALL	OVR				
42	2015	2015	RI	Rhode Islan	YRBSS	Alcohol	Alcohol use among youth		%	Crude Preval	26.1	26.1			22.1	30.6	Overall	Overall					(41.70828019300046, -		44	ALC	ALC1_1	CrdPrev	OVERALL	OVR				
43	2015	2015	SC	South Caroli	YRBSS	Alcohol	Alcohol use among youth		%	Crude Preval	24.6	24.6			21.2	28.2	Overall	Overall					(33.998821303000454,		45	ALC	ALC1_1	CrdPrev	OVERALL	OVR				
44	2015	2015	SD	South Dakot	YRBSS	Alcohol	Alcohol use among youth		%	Crude Preval	28	28			23.1	33.6	Overall	Overall					(44.353130053000484,		46	ALC	ALC1_1	CrdPrev	OVERALL	OVR				
45	2015	2015	TX	Texas	YRBSS	Alcohol	Alcohol use among youth		%	Crude Prevalence			-	No data available			Overall	Overall					(31.827240407000488,		48	ALC	ALC1_1	CrdPrev	OVERALL	OVR				
46	2015	2015	US	United States	YRBSS	Alcohol	Alcohol use among youth		%	Crude Preval	32.8	32.8			30.4	35.2	Overall	Overall							59	ALC	ALC1_1	CrdPrev	OVERALL	OVR				
47	2015	2015	UT	Utah	YRBSS	Alcohol	Alcohol use among youth		%	Crude Prevalence			-	No data available			Overall	Overall					(39.360700171000474,		49	ALC	ALC1_1	CrdPrev	OVERALL	OVR				
48	2015	2015	VA	Virginia	YRBSS	Alcohol	Alcohol use among youth		%	Crude Preval	23.4	23.4			21.1	25.8	Overall	Overall					(37.54268067400045, -		51	ALC	ALC1_1	CrdPrev	OVERALL	OVR				
49	2015	2015	VT	Vermont	YRBSS	Alcohol	Alcohol use among youth		%	Crude Preval	30	30			29.3	30.6	Overall	Overall					(43.62538123900049, -		50	ALC	ALC1_1	CrdPrev	OVERALL	OVR				
50	2015	2015	WA	Washington	YRBSS	Alcohol	Alcohol use among youth		%	Crude Prevalence			-	No data available			Overall	Overall					(47.52227862900048, -		53	ALC	ALC1_1	CrdPrev	OVERALL	OVR				
51	2015	2015	WI	Wisconsin	YRBSS	Alcohol	Alcohol use among youth		%	Crude Prevalence			-	No data available			Overall	Overall					(44.39319117400049, -		55	ALC	ALC1_1	CrdPrev	OVERALL	OVR				
52	2015	2015	WV	West Virgini	YRBSS	Alcohol	Alcohol use among youth		%	Crude Preval	31.1	31.1			28.1	34.2	Overall	Overall					(38.66551020200046, -		54	ALC	ALC1_1	CrdPrev	OVERALL	OVR				
53	2015	2015	WY	Wyoming	YRBSS	Alcohol	Alcohol use among youth		%	Crude Preval	31	31			28.1	34	Overall	Overall					(43.23554134900048, -		56	ALC	ALC1_1	CrdPrev	OVERALL	OVR				
54	2015	2015	AK	Alaska	YRBSS	Alcohol	Binge drinking prevalence		%	Crude Preval	12.5	12.5			10.8	14.5	Overall	Overall					(64.84507995700051, -		2	ALC	ALC2_1	CrdPrev	OVERALL	OVR				
55	2015	2015	AL	Alabama	YRBSS	Alcohol	Binge drinking prevalence		%	Crude Preval	16.1	16.1			13.5	19.1	Overall	Overall					(32.84057112200048, -		1	ALC	ALC2_1	CrdPrev	OVERALL	OVR				
56	2015	2015	AR	Arkansas	YRBSS	Alcohol	Binge drinking prevalence		%	Crude Preval	16.3	16.3			14.8	18.1	Overall	Overall					(34.74865012400045, -		5	ALC	ALC2_1	CrdPrev	OVERALL	OVR				
57	2015	2015	AZ	Arizona	YRBSS	Alcohol	Binge drinking prevalence		%	Crude Preval	19	19			14.8	24.1	Overall	Overall					(34.865970280000454,		4	ALC	ALC2_1	CrdPrev	OVERALL	OVR				
58	2015	2015	CA	California	YRBSS	Alcohol	Binge drinking prevalence		%	Crude Preval	15.1	15.1			11.4	19.9	Overall	Overall					(37.63864012300047, -		6	ALC	ALC2_1	CrdPrev	OVERALL	OVR				
59	2015	2015	CO	Colorado	YRBSS	Alcohol	Binge drinking prevalence		%	Crude Prevalence			-	No data available			Overall	Overall					(38.843840757000464,		8	ALC	ALC2_1	CrdPrev	OVERALL	OVR				
60	2015	2015	CT	Connecticut	YRBSS	Alcohol	Binge drinking prevalence		%	Crude Preval	14	14			11.6	16.8	Overall	Overall					(41.56266102000046, -		9	ALC	ALC2_1	CrdPrev	OVERALL	OVR				
61	2015	2015	DC	District of Co	YRBSS	Alcohol	Binge drinking prevalence		%	Crude Preval	8.3	8.3			7.8	8.9	Overall	Overall					(38.907192, -77.036871		11	ALC	ALC2_1	CrdPrev	OVERALL	OVR				
62	2015	2015	FL	Florida	YRBSS	Alcohol	Binge drinking prevalence		%	Crude Preval	15.3	15.3			13.9	16.9	Overall	Overall					(28.932040377000476,		12	ALC	ALC2_1	CrdPrev	OVERALL	OVR				
63	2015	2015	GA	Georgia	YRBSS	Alcohol	Binge drinking prevalence		%	Crude Prevalence			-	No data available			Overall	Overall					(32.83968109300048, -		13	ALC	ALC2_1	CrdPrev	OVERALL	OVR				
64	2015	2015	GU	Guam	YRBSS	Alcohol	Binge drinking prevalence		%	Crude Preval	13.3	13.3			10.9	16	Overall	Overall					(13.444304, 144.79373		66	ALC	ALC2_1	CrdPrev	OVERALL	OVR				
65	2015	2015	HI	Hawaii	YRBSS	Alcohol	Binge drinking prevalence		%	Crude Preval	13.4	13.4			12.2	14.8	Overall	Overall					(21.304850435000446,		15	ALC	ALC2_1	CrdPrev	OVERALL	OVR				
66	2015	2015	IA	Iowa	YRBSS	Alcohol	Binge drinking prevalence		%	Crude Prevalence			-	No data available			Overall	Overall					(42.46940091300047, -		19	ALC	ALC2_1	CrdPrev	OVERALL	OVR				
67	2015	2015	ID	Idaho	YRBSS	Alcohol	Binge drinking prevalence		%	Crude Preval	15.6	15.6			12.5	19.3	Overall	Overall					(43.682630005000476,		16	ALC	ALC2_1	CrdPrev	OVERALL	OVR				
68	2015	2015	IL	Illinois	YRBSS	Alcohol	Binge drinking prevalence		%	Crude Preval	15.9	15.9			13	19.4	Overall	Overall					(40.48501028300046, -		17	ALC	ALC2_1	CrdPrev	OVERALL	OVR				
69	2015	2015	IN	Indiana	YRBSS	Alcohol	Binge drinking prevalence		%	Crude Preval	17.4	17.4			14	21.5	Overall	Overall					(39.766910452000445,		18	ALC	ALC2_1	CrdPrev	OVERALL	OVR				
70	2015	2015	KS	Kansas	YRBSS	Alcohol	Binge drinking prevalence		%	Crude Prevalence			-	No data available			Overall	Overall					(38.34774030000045, -		20	ALC	ALC2_1	CrdPrev	OVERALL	OVR				

Figure 1. U.S._Chronic_Disease_Indicators dataset



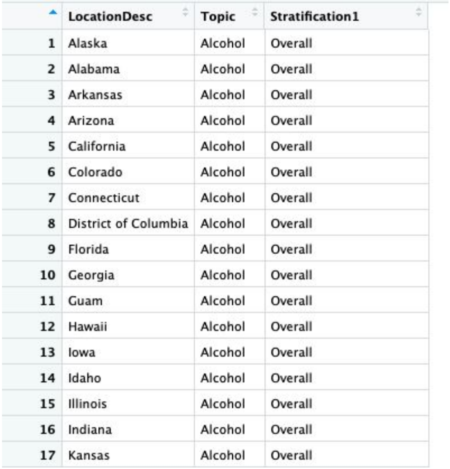
	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Data as of	Start week	End Week	State	Age group	Race and His	COVID-19 De	Total Deaths	Pneumonia D	Pneumonia a	Influenza Dea	Pneumonia, I	Footnote
2	07/08/20	02/01/20	07/04/20	United State	All Ages	Total Deaths	114741	1324958	125868	49623	6492	196538	
3	07/08/20	02/01/20	07/04/20	United State	All Ages	Non-Hispani	60862	974415	81724	24734	4303	121940	
4	07/08/20	02/01/20	07/04/20	United State	All Ages	Non-Hispani	26426	177674	20910	11912	969	36084	
5	07/08/20	02/01/20	07/04/20	United State	All Ages	Non-Hispani	888	8187	982	472	68	1466	
6	07/08/20	02/01/20	07/04/20	United State	All Ages	Non-Hispani	5629	38200	5170	2546	267	8411	
7	07/08/20	02/01/20	07/04/20	United State	All Ages	Non-Hispani	127	1606	150	74	8	211	
8	07/08/20	02/01/20	07/04/20	United State	All Ages	Non-Hispani	282	5941	414	127	34	601	
9	07/08/20	02/01/20	07/04/20	United State	All Ages	Hispanic or L	19409	114570	15797	9298	811	26437	
10	07/08/20	02/01/20	07/04/20	United State	All Ages	Unknown	1118	4365	721	460	32	1388	
11	07/08/20	02/01/20	07/04/20	United State	Under 1 year	Non-Hispani	3	2948	26	0	8	37	
12	07/08/20	02/01/20	07/04/20	United State	1-4 years	Non-Hispani	3	592	18	1	18	38	
13	07/08/20	02/01/20	07/04/20	United State	5-14 years	Non-Hispani	2	1007	33	1	26	60	
14	07/08/20	02/01/20	07/04/20	United State	15-24 years	Non-Hispani	24	6148	105	7	30	151	
15	07/08/20	02/01/20	07/04/20	United State	25-34 years	Non-Hispani	133	15196	392	63	68	530	
16	07/08/20	02/01/20	07/04/20	United State	35-44 years	Non-Hispani	294	22213	812	103	115	1116	
17	07/08/20	02/01/20	07/04/20	United State	45-54 years	Non-Hispani	1201	43659	2469	534	291	3420	
18	07/08/20	02/01/20	07/04/20	United State	55-64 years	Non-Hispani	4412	112940	8311	1998	717	11402	
19	07/08/20	02/01/20	07/04/20	United State	65-74 years	Non-Hispani	10409	187263	16364	4610	932	23043	
20	07/08/20	02/01/20	07/04/20	United State	75-84 years	Non-Hispani	17140	248692	23444	7262	1050	34329	
21	07/08/20	02/01/20	07/04/20	United State	85 years and	Non-Hispani	27241	333757	29750	10155	1048	47814	
22	07/08/20	02/01/20	07/04/20	United State	Under 1 year	Non-Hispani	2	1975	19	1	3	23	
23	07/08/20	02/01/20	07/04/20	United State	1-4 years	Non-Hispani	1	333	16	1	5	21	
24	07/08/20	02/01/20	07/04/20	United State	5-14 years	Non-Hispani	7	457	14	0	10	31	
25	07/08/20	02/01/20	07/04/20	United State	15-24 years	Non-Hispani	45	3238	56	12	5	92	
26	07/08/20	02/01/20	07/04/20	United State	25-34 years	Non-Hispani	252	6027	252	107	37	428	
27	07/08/20	02/01/20	07/04/20	United State	35-44 years	Non-Hispani	627	7990	565	272	46	962	
28	07/08/20	02/01/20	07/04/20	United State	45-54 years	Non-Hispani	1776	14064	1411	780	122	2470	

Figure 2.

Deaths_involving_coronavirus_disease_2019__COVID-19__by_race_and_Hispanic_origin_group_and_age__by_state dataset

2 a. Chronic Diseases Dataset

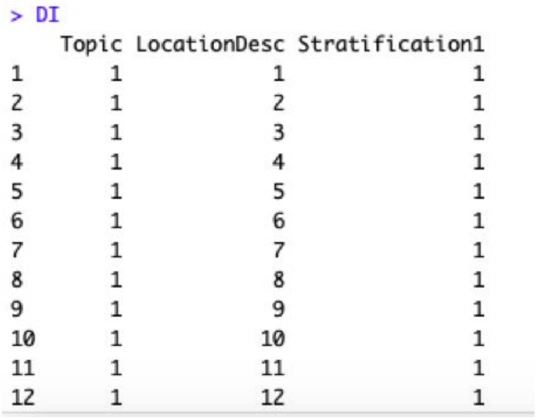
The U.S._Chronic_Diseases_Indicators dataset consists of 403985 rows of data and 34 variable columns. Out of the 34 variables, I used 3 of them, specifically "Stratification1", "Topic", and "LocationDesc" (FIGURE 3).



	LocationDesc	Topic	Stratification1
1	Alaska	Alcohol	Overall
2	Alabama	Alcohol	Overall
3	Arkansas	Alcohol	Overall
4	Arizona	Alcohol	Overall
5	California	Alcohol	Overall
6	Colorado	Alcohol	Overall
7	Connecticut	Alcohol	Overall
8	District of Columbia	Alcohol	Overall
9	Florida	Alcohol	Overall
10	Georgia	Alcohol	Overall
11	Guam	Alcohol	Overall
12	Hawaii	Alcohol	Overall
13	Iowa	Alcohol	Overall
14	Idaho	Alcohol	Overall
15	Illinois	Alcohol	Overall
16	Indiana	Alcohol	Overall
17	Kansas	Alcohol	Overall

Figure 3. Required columns from dataset

To analyse my question, how likely is someone belonging to a certain ethnicity and location to have a chronic disease, I decided to use a decision tree. Before creating my decision tree, random forest results and predictions, I looked at the distributions of the different variables and their density plots to look at how much overlap there was. Decision trees help lay the problem out in a very simple manner, and allow us to fully analyse the consequences of any decision. It provides a framework to quantify the values of outcomes and the probabilities of achieving them. For my decision tree, I used Topic as my dependent variable, and Stratification and Location as my predictor variables. I then performed Random Forest on my data. I used random forest because since it consists of multiple decision trees, it provides higher accuracy, especially with larger data[7]. Since all my variables of interests were categorical variables, I factored them out, where each category is represented by a certain number (Figure 4). Based on my random forest results, I created predictions for how likely someone is to have a chronic disease based on their location and stratification.

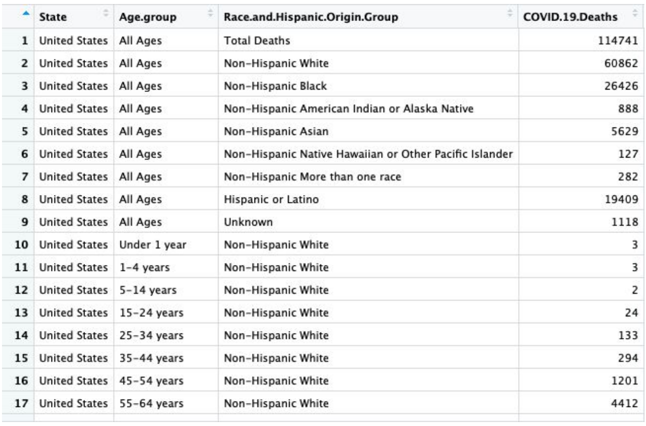


```
> DI
   Topic LocationDesc Stratification1
1      1            1               1
2      1            2               1
3      1            3               1
4      1            4               1
5      1            5               1
6      1            6               1
7      1            7               1
8      1            8               1
9      1            9               1
10     1           10               1
11     1           11               1
12     1           12               1
```

Figure 4. Factored variables

2 b. COVID-19 Dataset

The Deaths_involving_coronavirus_disease_2019__COVID-19__by_race_and_Hispanic_origin_group_and_age__by_state.csv dataset consists of 4762 rows of data and 13 variable columns. Out of the 13 variables, I used 4 of them, specifically, “Age group”, “Race and Hispanic Origin Group”, “State” and “COVID-19 Deaths” (Figure 5).



	State	Age.group	Race.and.Hispanic.Origin.Group	COVID.19.Deaths
1	United States	All Ages	Total Deaths	114741
2	United States	All Ages	Non-Hispanic White	60862
3	United States	All Ages	Non-Hispanic Black	26426
4	United States	All Ages	Non-Hispanic American Indian or Alaska Native	888
5	United States	All Ages	Non-Hispanic Asian	5629
6	United States	All Ages	Non-Hispanic Native Hawaiian or Other Pacific Islander	127
7	United States	All Ages	Non-Hispanic More than one race	282
8	United States	All Ages	Hispanic or Latino	19409
9	United States	All Ages	Unknown	1118
10	United States	Under 1 year	Non-Hispanic White	3
11	United States	1-4 years	Non-Hispanic White	3
12	United States	5-14 years	Non-Hispanic White	2
13	United States	15-24 years	Non-Hispanic White	24
14	United States	25-34 years	Non-Hispanic White	133
15	United States	35-44 years	Non-Hispanic White	294
16	United States	45-54 years	Non-Hispanic White	1201
17	United States	55-64 years	Non-Hispanic White	4412

Figure 5. Required columns from dataset

To analyse my question, how likely is someone to die from COVID based on their age and ethnicity, I initially created distribution charts to see what percent of a certain age, location and ethnicity was represented in the data. Due to the large data size, I created a density plot, to look at the overlap of variables. Finally, to look at how age and ethnicity was related to the death count, I created a scatterplot comparing the two variables. I created a decision tree to help answer my question, then performed random forest, and based on the results of the random forest, created predictions for how likely someone of a certain age and ethnicity is to die from COVID.

3. RESULTS

3 a. Chronic Diseases Dataset

My distribution charts showed a higher prevalence of chronic diseases in more minority ethnic groups than caucasians (Figure 5) and I expected cardiac diseases and diabetes to be the most common chronic diseases, which is also shown (Figure 6). Since I have such large data, I performed hexagonal binning to look at how much overlap there is between stratification and location, and as shown in Figure 7, there is significant overlap between the two. Based on my decision tree I created (Figure 8), by just looking at it, I am able to predict which disease one is more likely to have based on their stratification and location. However, I did encounter an issue, where my decision tree is extremely condensed, which is primarily due to the large number of categories in each variable. On performing my random forest (Figure 9), I did notice a large OOB estimate of error rate, which could be attributed to the large number of categorical data. To better understand my random forest results, I ran a prediction and confusion matrix, examples of which can be found in Figure 10.

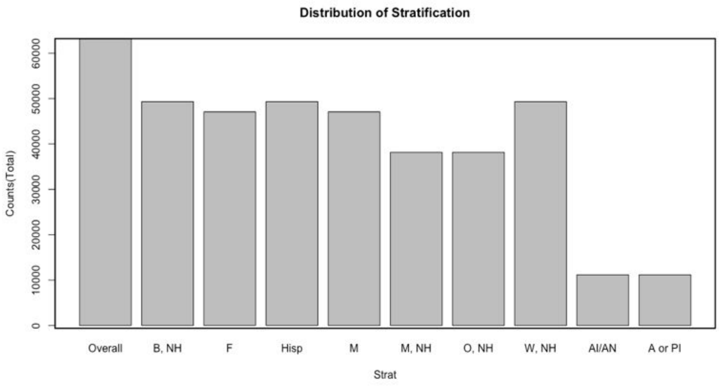


Figure 5. Distribution of Stratification

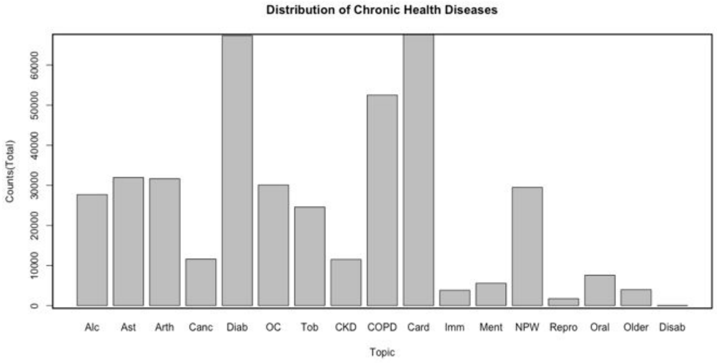


Figure 6. Distribution of Chronic Health Diseases

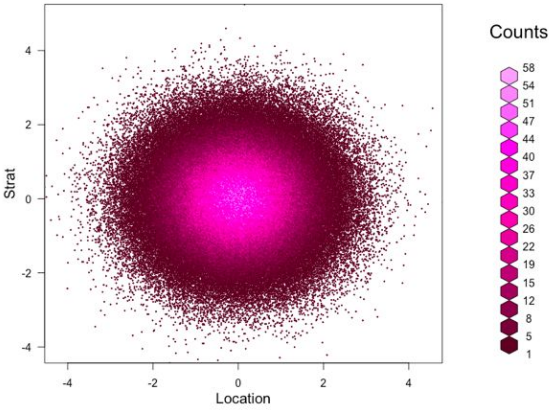


Figure 7. Hexagonal Binning showing Overlap between Location and Stratification.

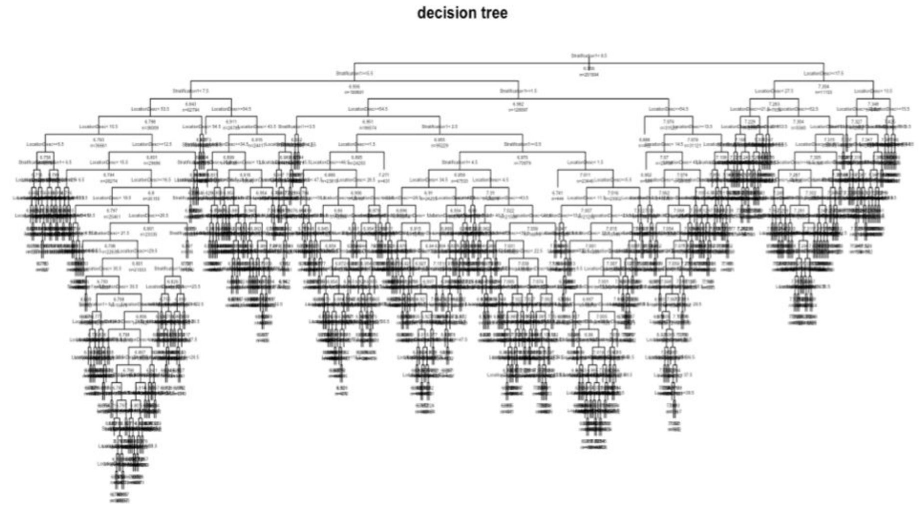
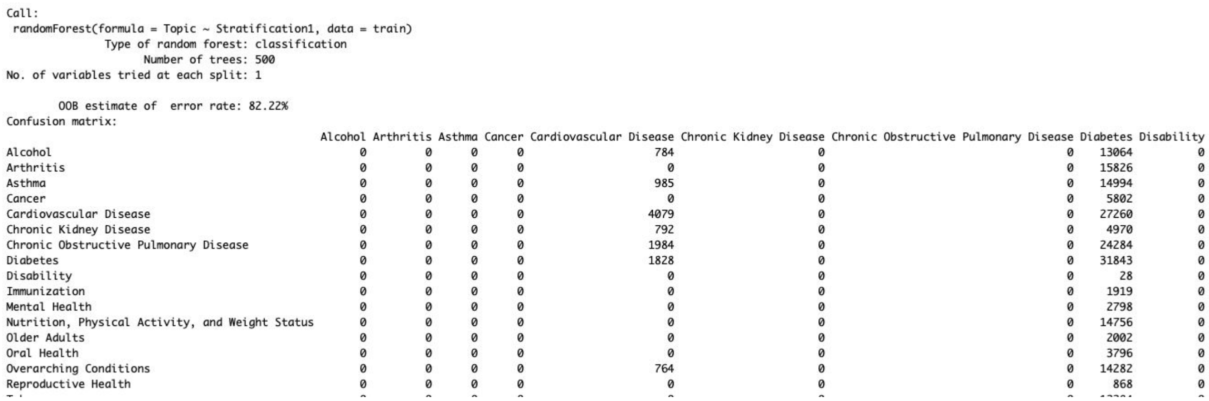


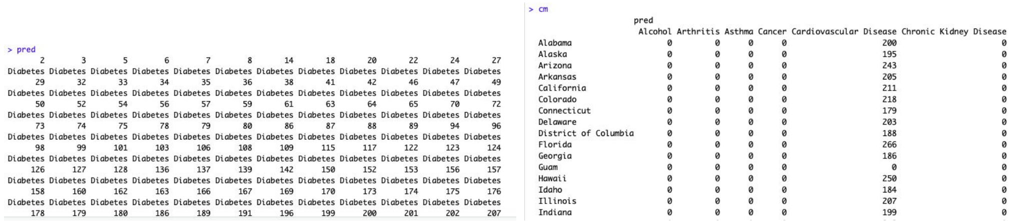
Figure 8. Decision Tree



```
Call:
 randomForest(formula = Topic ~ Stratification1, data = train)
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 1

        OOB estimate of  error rate: 82.22%
Confusion matrix:
                                                Alcohol Arthritis Asthma Cancer Cardiovascular Disease Chronic Kidney Disease Chronic Obstructive Pulmonary Disease Diabetes Disability
Alcohol                                               0         0      0      0                    784                      0                                     0    13064          0
Arthritis                                             0         0      0      0                      0                      0                                     0    15826          0
Asthma                                                0         0      0      0                    985                      0                                     0    14994          0
Cancer                                                0         0      0      0                      0                      0                                     0     5802          0
Cardiovascular Disease                                0         0      0      0                   4079                      0                                     0    27260          0
Chronic Kidney Disease                                0         0      0      0                    792                      0                                     0     4970          0
Chronic Obstructive Pulmonary Disease                 0         0      0      0                   1984                      0                                     0    24284          0
Diabetes                                              0         0      0      0                   1828                      0                                     0    31843          0
Disability                                            0         0      0      0                      0                      0                                     0       28          0
Immunization                                          0         0      0      0                      0                      0                                     0     1919          0
Mental Health                                         0         0      0      0                      0                      0                                     0     2798          0
Nutrition, Physical Activity, and Weight Status       0         0      0      0                      0                      0                                     0    14756          0
Older Adults                                          0         0      0      0                      0                      0                                     0     2002          0
Oral Health                                           0         0      0      0                      0                      0                                     0     3796          0
Overarching Conditions                                0         0      0      0                    764                      0                                     0    14282          0
Reproductive Health                                   0         0      0      0                      0                      0                                     0      868          0
```

Figure 9. Sample of Random Forest Results



```
> pred
       2        3        5        6        7        8       14       18       20       22       24       27 
Diabetes Diabetes Diabetes Diabetes Diabetes Diabetes Diabetes Diabetes Diabetes Diabetes Diabetes Diabetes 
      29       32       33       34       35       36       38       41       42       46       47       49 
Diabetes Diabetes Diabetes Diabetes Diabetes Diabetes Diabetes Diabetes Diabetes Diabetes Diabetes Diabetes 
      50       52       54       56       57       59       61       63       64       65       70       72 
Diabetes Diabetes Diabetes Diabetes Diabetes Diabetes Diabetes Diabetes Diabetes Diabetes Diabetes Diabetes 
      73       74       75       78       79       80       86       87       88       89       94       96 
Diabetes Diabetes Diabetes Diabetes Diabetes Diabetes Diabetes Diabetes Diabetes Diabetes Diabetes Diabetes 
      98       99      101      103      106      108      109      115      117      122      123      124 
Diabetes Diabetes Diabetes Diabetes Diabetes Diabetes Diabetes Diabetes Diabetes Diabetes Diabetes Diabetes 
     126      127      128      136      137      139      142      150      152      153      156      157 
Diabetes Diabetes Diabetes Diabetes Diabetes Diabetes Diabetes Diabetes Diabetes Diabetes Diabetes Diabetes 
     158      160      162      163      166      167      169      170      173      174      175      176 
Diabetes Diabetes Diabetes Diabetes Diabetes Diabetes Diabetes Diabetes Diabetes Diabetes Diabetes Diabetes 
     178      179      180      186      189      191      196      199      200      201      202      207 
```

```
> cm
                      pred
                       Alcohol Arthritis Asthma Cancer Cardiovascular Disease Chronic Kidney Disease
  Alabama                    0         0      0      0                    200                      0
  Alaska                     0         0      0      0                    195                      0
  Arizona                    0         0      0      0                    243                      0
  Arkansas                   0         0      0      0                    205                      0
  California                 0         0      0      0                    211                      0
  Colorado                   0         0      0      0                    218                      0
  Connecticut                0         0      0      0                    179                      0
  Delaware                   0         0      0      0                    203                      0
  District of Columbia       0         0      0      0                    188                      0
  Florida                    0         0      0      0                    266                      0
  Georgia                    0         0      0      0                    186                      0
  Guam                       0         0      0      0                      0                      0
  Hawaii                     0         0      0      0                    250                      0
  Idaho                      0         0      0      0                    184                      0
  Illinois                   0         0      0      0                    207                      0
  Indiana                    0         0      0      0                    199                      0
```

Figure 10. Prediction and Confusion Matrix

3 b. COVID-19 Dataset

Based on my distribution charts, there is a very equal distribution of Age, Ethnicity and Location in the data, which implies there might not have been random sampling (Figure 11, 12, 13). However, based on my scatterplot comparing death counts based on Age and Ethnicity (Figure 14) we can see that there is a much higher death count for minority ethnic groups when compared to Cauasians. In the density plot, looking at the overlap of Age and Ethnicity (Figure 15), there is not much overlap, and where the overlap is could be the minority ethnic groups and the older age groups. I created three different decision trees, the first one predicting probability of death based on age and ethnicity (Figure 16), the second one predicting death based on age (Figure 17) and the last on predicting death based on ethnicity (Figure 18). These decision trees are quite clear and provide the expected probabilities. The random forest results (Figure 19) are also much cleaner and provide a better prediction of probability of death. Figure 20 shows an example of the prediction for death based on ethnicity. Since I suspected bias in data collection, I also wanted to see how much importance was given to each of my predictor variables, and the importance plot (Figure 21) shows that both variables were similar in treatment, which proves my theory of there being a bias.

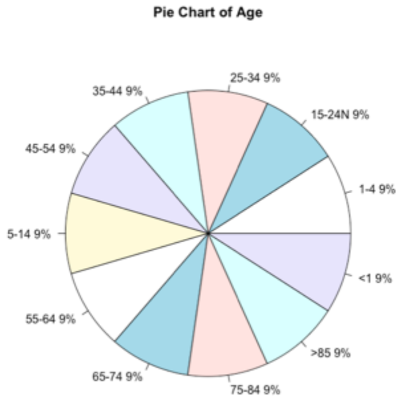


Figure 11. Age Distribution

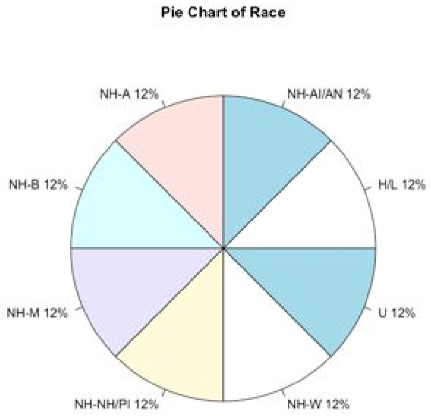


Figure 12. Ethnicity Distribution

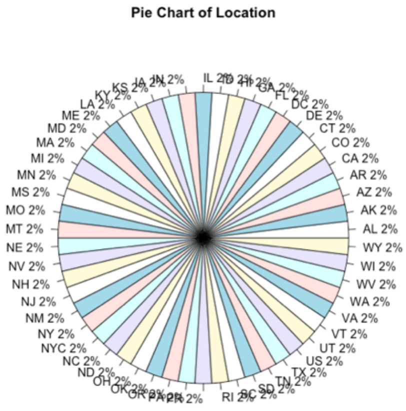


Figure 13. Location Distribution

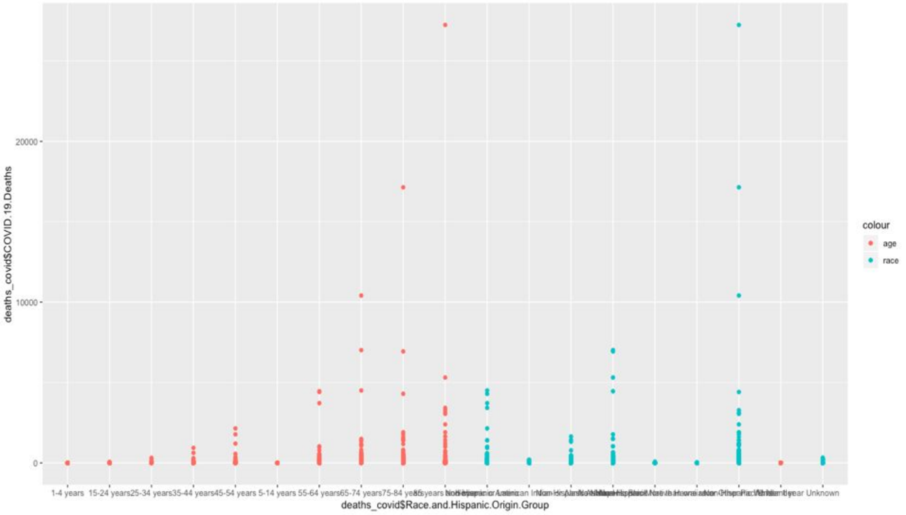


Figure 14. Scatterplot Comparing Age and Ethnicity Death Counts

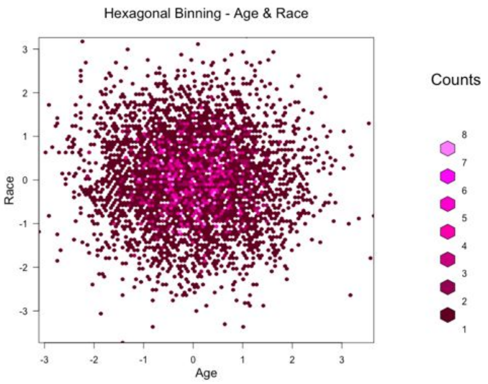


Figure 15. Density Plot of Age and Ethnicity

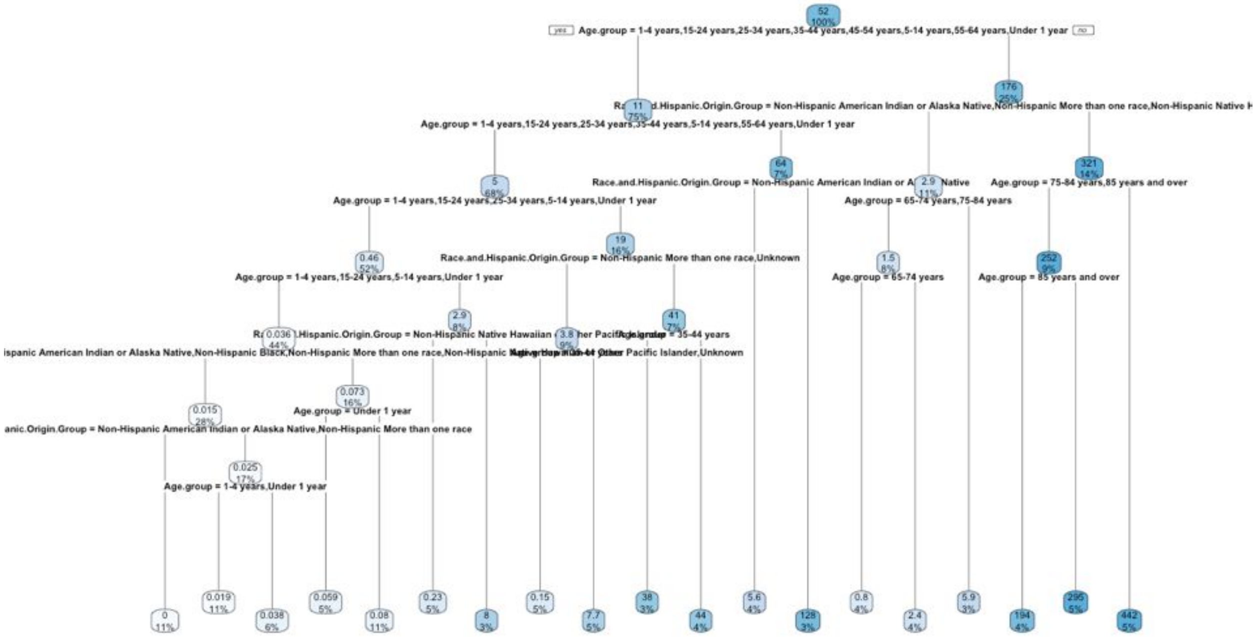

52
100%
yes
Age.group = 1-4 years,15-24 years,25-34 years,35-44 years,45-54 years,5-14 years,55-64 years,Under 1 year
no
11
75%
Race.and.Hispanic.Origin.Group = Non-Hispanic American Indian or Alaska Native,Non-Hispanic More than one race,Non-Hispanic Native H
176
25%
Age.group = 1-4 years,15-24 years,25-34 years,35-44 years,5-14 years,55-64 years,Under 1 year
5
68%
64
7%
Race.and.Hispanic.Origin.Group = Non-Hispanic American Indian or Alaska Native
2.9
11%
Age.group = 65-74 years,75-84 years
321
14%
Age.group = 75-84 years,85 years and over
Age.group = 1-4 years,15-24 years,25-34 years,5-14 years,Under 1 year
19
16%
Race.and.Hispanic.Origin.Group = Non-Hispanic More than one race,Unknown
0.46
52%
Age.group = 1-4 years,15-24 years,5-14 years,Under 1 year
1.5
8%
Age.group = 65-74 years
252
9%
Age.group = 85 years and over
2.9
8%
41
7%
Age.group = 35-44 years
0.036
44%
Race.and.Hispanic.Origin.Group = Non-Hispanic Native Hawaiian or Other Pacific Islander
3.8
9%
ispanic American Indian or Alaska Native,Non-Hispanic Black,Non-Hispanic More than one race,Non-Hispanic Native Hawaiian or Other Pacific Islander,Unknown
Age.group = 35-44 years
0.073
16%
Age.group = Under 1 year
0.015
28%
anic.Origin.Group = Non-Hispanic American Indian or Alaska Native,Non-Hispanic More than one race
0.025
17%
Age.group = 1-4 years,Under 1 year
0
11%
0.019
11%
0.038
6%
0.059
5%
0.08
11%
0.23
5%
8
3%
0.15
5%
7.7
5%
38
3%
44
4%
5.6
4%
128
3%
0.8
4%
2.4
4%
5.9
3%
194
4%
296
5%
442
5%

Figure 16. Decision Tree: Death ~ Age + Ethnicity

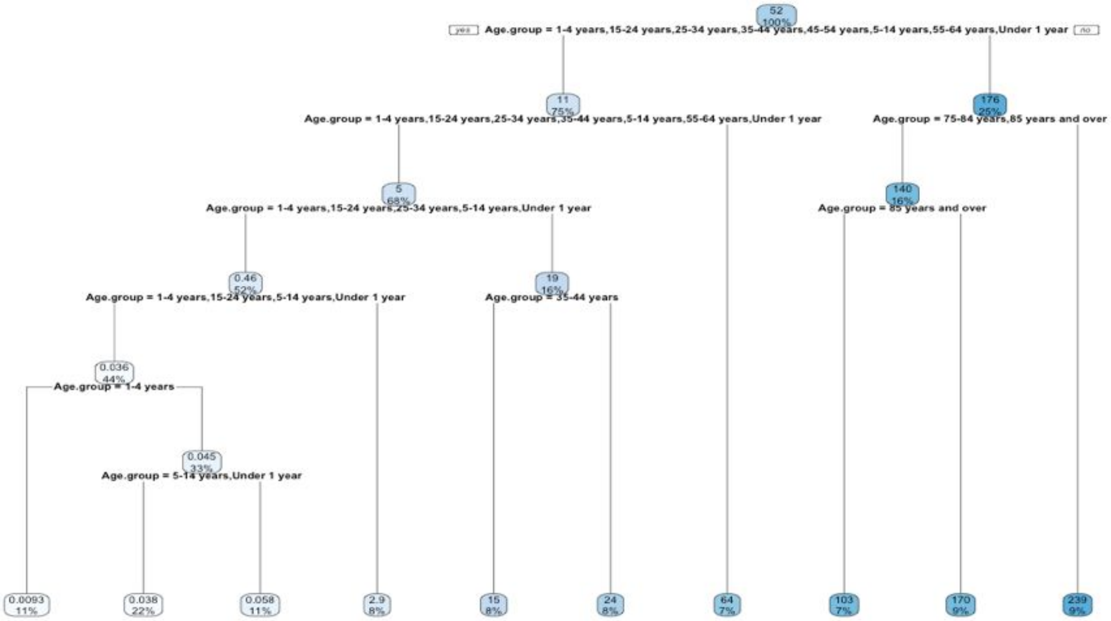

52
100%
yes
Age.group = 1-4 years,15-24 years,25-34 years,35-44 years,45-54 years,5-14 years,55-64 years,Under 1 year
no
11
75%
Age.group = 1-4 years,15-24 years,25-34 years,35-44 years,5-14 years,55-64 years,Under 1 year
176
25%
Age.group = 75-84 years,85 years and over
5
68%
Age.group = 1-4 years,15-24 years,25-34 years,5-14 years,Under 1 year
140
16%
Age.group = 85 years and over
0.46
52%
Age.group = 1-4 years,15-24 years,5-14 years,Under 1 year
19
16%
Age.group = 35-44 years
0.036
44%
Age.group = 1-4 years
0.045
33%
Age.group = 5-14 years,Under 1 year
0.0093
11%
0.038
22%
0.058
11%
2.9
8%
15
8%
24
8%
64
7%
103
7%
170
9%
239
9%

Figure 17. Decision Tree: Death ~ Age

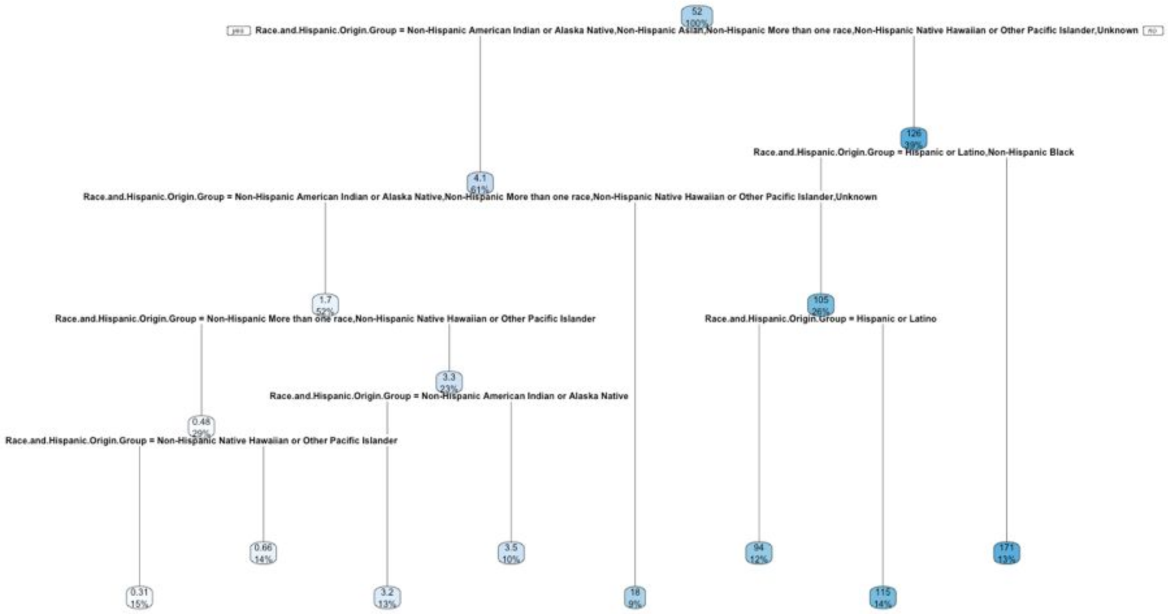
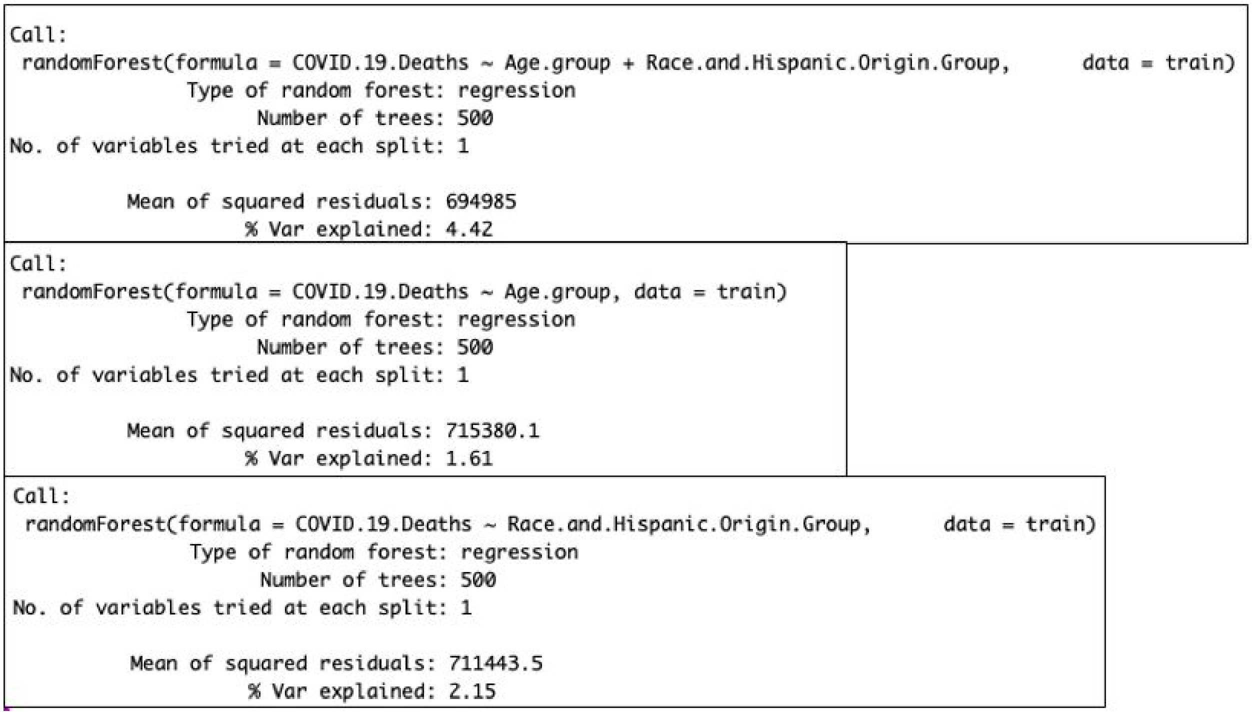

52
100%
yes
Race.and.Hispanic.Origin.Group = Non-Hispanic American Indian or Alaska Native,Non-Hispanic Asian,Non-Hispanic More than one race,Non-Hispanic Native Hawaiian or Other Pacific Islander,Unknown
no
126
39%
Race.and.Hispanic.Origin.Group = Hispanic or Latino,Non-Hispanic Black
4.1
61%
Race.and.Hispanic.Origin.Group = Non-Hispanic American Indian or Alaska Native,Non-Hispanic More than one race,Non-Hispanic Native Hawaiian or Other Pacific Islander,Unknown
1.7
52%
Race.and.Hispanic.Origin.Group = Non-Hispanic More than one race,Non-Hispanic Native Hawaiian or Other Pacific Islander
105
26%
Race.and.Hispanic.Origin.Group = Hispanic or Latino
3.3
23%
Race.and.Hispanic.Origin.Group = Non-Hispanic American Indian or Alaska Native
0.48
29%
Race.and.Hispanic.Origin.Group = Non-Hispanic Native Hawaiian or Other Pacific Islander
0.66
14%
3.5
10%
94
12%
171
13%
0.31
15%
3.2
13%
18
9%
115
14%

Figure 18. Decision Tree: Death ~ Ethnicity



```
Call:
 randomForest(formula = COVID.19.Deaths ~ Age.group + Race.and.Hispanic.Origin.Group,      data = train)
               Type of random forest: regression
                     Number of trees: 500
No. of variables tried at each split: 1

          Mean of squared residuals: 694985
                    % Var explained: 4.42
```

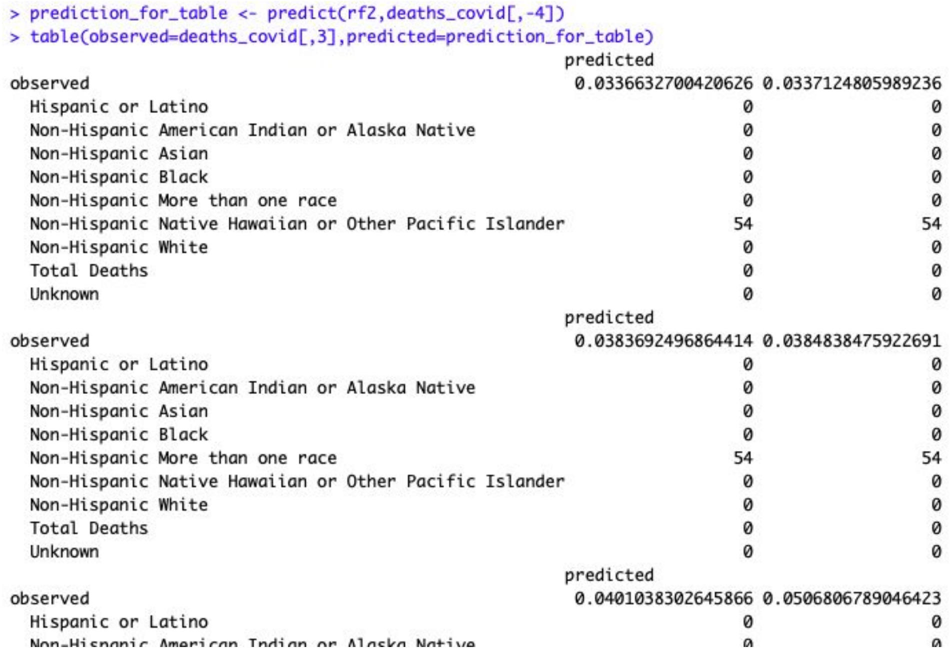
```
Call:
 randomForest(formula = COVID.19.Deaths ~ Age.group, data = train)
               Type of random forest: regression
                     Number of trees: 500
No. of variables tried at each split: 1

          Mean of squared residuals: 715380.1
                    % Var explained: 1.61
```

```
Call:
 randomForest(formula = COVID.19.Deaths ~ Race.and.Hispanic.Origin.Group,      data = train)
               Type of random forest: regression
                     Number of trees: 500
No. of variables tried at each split: 1

          Mean of squared residuals: 711443.5
                    % Var explained: 2.15
```

Figure 19. Random Forest Results for each situation



```
> prediction_for_table <- predict(rf2,deaths_covid[,-4])
> table(observed=deaths_covid[,3],predicted=prediction_for_table)
                                                            predicted
observed                                                      0.0336632700420626 0.0337124805989236
  Hispanic or Latino                                                           0                  0
  Non-Hispanic American Indian or Alaska Native                                0                  0
  Non-Hispanic Asian                                                           0                  0
  Non-Hispanic Black                                                           0                  0
  Non-Hispanic More than one race                                              0                  0
  Non-Hispanic Native Hawaiian or Other Pacific Islander                      54                 54
  Non-Hispanic White                                                           0                  0
  Total Deaths                                                                 0                  0
  Unknown                                                                      0                  0
                                                            predicted
observed                                                      0.0383692496864414 0.0384838475922691
  Hispanic or Latino                                                           0                  0
  Non-Hispanic American Indian or Alaska Native                                0                  0
  Non-Hispanic Asian                                                           0                  0
  Non-Hispanic Black                                                           0                  0
  Non-Hispanic More than one race                                             54                 54
  Non-Hispanic Native Hawaiian or Other Pacific Islander                       0                  0
  Non-Hispanic White                                                           0                  0
  Total Deaths                                                                 0                  0
  Unknown                                                                      0                  0
                                                            predicted
observed                                                      0.0401038302645866 0.0506806789046423
  Hispanic or Latino                                                           0                  0
  Non-Hispanic American Indian or Alaska Native                                0                  0
```

Figure 20. Sample Confusion Matrix of Prediction for Death based on Ethnicity

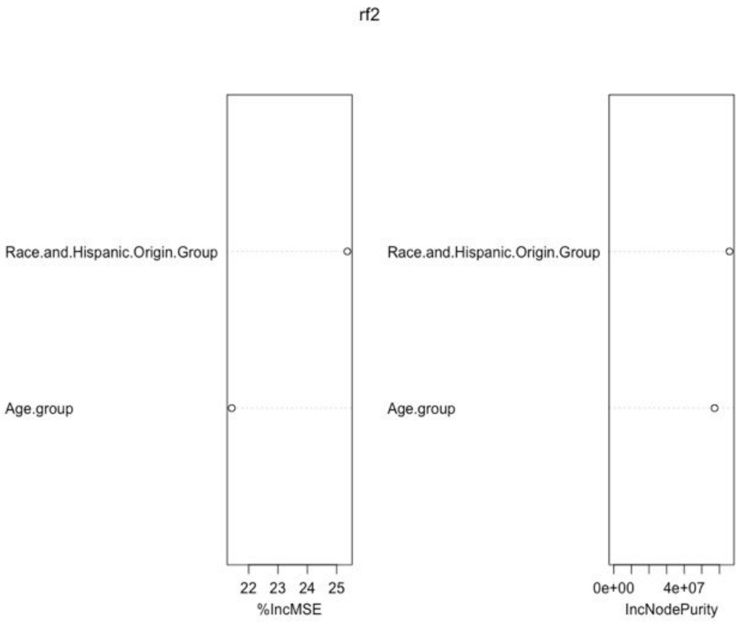


Figure 21. Importance Plot

4. CONCLUSION

There is a significant interaction between the type of chronic disease, and location and stratification. Using my decision tree and prediction matrix, I am able to gauge what chronic disease someone of a certain stratification and location is most likely to have. I applied this knowledge on my COVID dataset, and was able to predict how likely someone is to die based on their age and ethnicity. My data and results supported my hypothesis of there being a greater prevalence of chronic diseases in minority ethnic groups when compared to white, non hispanic. I also saw a similar trend in death counts due to COVID-19. I could further apply this understanding into looking at how other pandemics like Pneumonia and Influenza disproportionately affect minority ethnic groups.

5. BIBLIOGRAPHY

1. Dreyer, G., et al. "Effect of Ethnicity on the Prevalence of Diabetes and Associated Chronic Kidney Disease." *OUP Academic*, Oxford University Press, 15 Jan. 2009, academic.oup.com/qjmed/article/102/4/261/1550086.

2. Atella, Vincenzo, et al. "Trends in Age-Related Disease Burden and Healthcare Utilization." *Aging Cell*, John Wiley and Sons Inc., Feb. 2019, www.ncbi.nlm.nih.gov/pmc/articles/PMC6351821/.

3. Kirby, Tony. "Evidence mounts on the disproportionate effect of COVID-19 on ethnic minorities." *The Lancet. Respiratory medicine* vol. 8,6 (2020): 547-548. doi:10.1016/S2213-2600(20)30228-9

4. Davies, Nicholas G., et al. “Age-Dependent Effects in the Transmission and Control of COVID-19 Epidemics.” *Nature News*, Nature Publishing Group, 16 June 2020, www.nature.com/articles/s41591-020-0962-9.

5. Centers for Disease Control and Prevention. “Chronic Disease Indicators.” *Kaggle*, 17 Aug. 2017, www.kaggle.com/cdc/chronic-disease.

6. “Provisional COVID-19 Death Counts by Sex, Age, and State.” *Data.gov*, Publisher Centers for Disease Control and Prevention, 25 June 2020, catalog.data.gov/dataset/provisional-covid-19-death-counts-by-sex-age-and-state.

7. Deng, Houtao. “Why Random Forests Outperform Decision Trees.” *Medium*, Towards Data Science, 12 Dec. 2018, towardsdatascience.com/why-random-forests-outperform-decision-trees-1b0f175a0b5.