MODELLING THE PANDEMIC

Sociodemographic predictors of COVID-19 impact in Chicago neighborhoods
by
Bored Grads Yacht Club

Christopher Owen
cowen20@uic.edu
https://github.com/antennarius

Kazi Shahrukh Omar
komar3@uic.edu
https://github.com/komar41

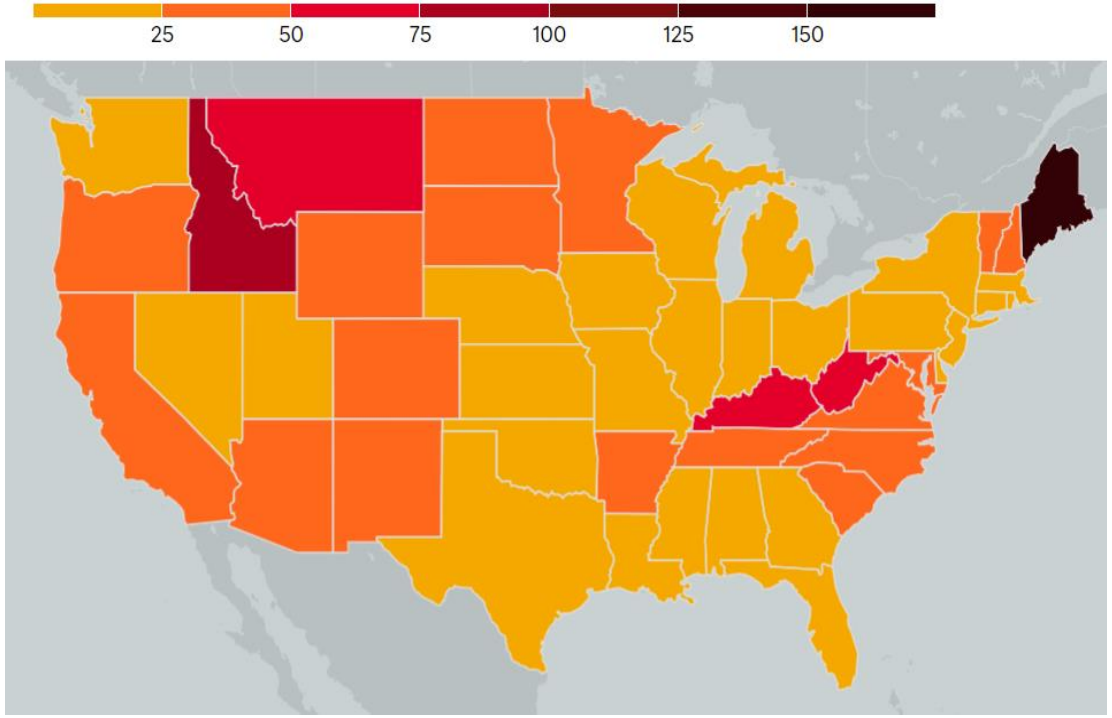
Abdul Rafey Siddiqui
asiddi73@uic.edu
https://github.com/rafeyyyyy

Nguyen Hoa Pham
npham30@uic.edu
https://github.com/nhpham27

Gautam Kushwah
gkushw2@uic.edu
https://github.com/gautam-kushwah

Project repository: https://github.com/uic-cs418/cs418-spring22-bored-grad-yacht-club

MOTIVATION



Average daily cases per 100,000 people (last updated: 2/23/2022)

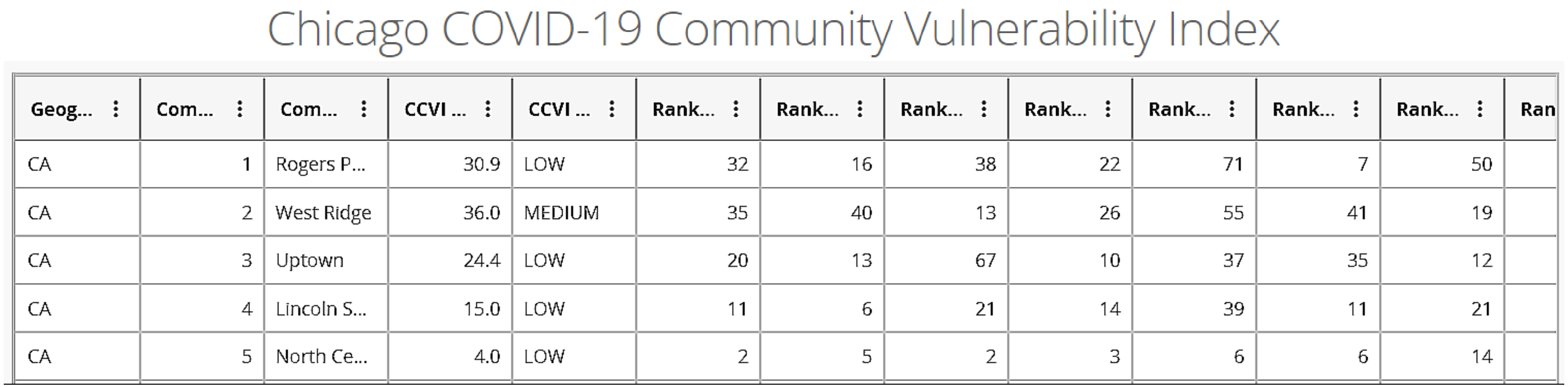
- The rapid outbreak of COVID-19 and its impact.
- Widely available COVID-19 data.
- Curiosity in finding a way to link socio-demographic data and COVID-19 impact.

DEFINITIONS

- How do we define sociodemographic data?
  - Physical factors like age, gender, ethnicity etc.
  - Social factors like income, level of education, time spent on public transit etc.
- How do we define COVID-19 impact?
  - Number of COVID-19 cases, deaths and hospitalizations.

WHERE WE STARTED OFF

- Focused on COVID-19 data in Chicago.
- Aimed to improve the existing CCVI ranking model.



Chicago COVID-19 Community Vulnerability Index

Geog...	Com...	Com...	CCVI ...	CCVI ...	Rank...	Rank...	Rank...	Rank...	Rank...	Rank...	Rank...	Ran
CA	1	Rogers P...	30.9	LOW	32	16	38	22	71	7	50	
CA	2	West Ridge	36.0	MEDIUM	35	40	13	26	55	41	19	
CA	3	Uptown	24.4	LOW	20	13	67	10	37	35	12	
CA	4	Lincoln S...	15.0	LOW	11	6	21	14	39	11	21	
CA	5	North Ce...	4.0	LOW	2	5	2	3	6	6	14	

EXPECTATION

- With our model, we aim to achieve:
  - Quantifiability of COVID-19 impact
  - Accuracy and uniformity

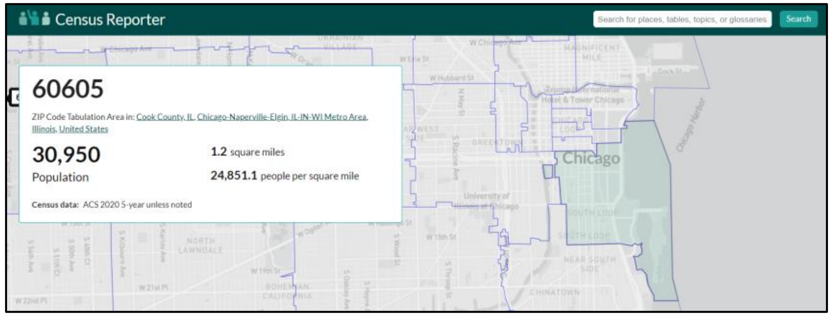
- Why is this important?
  - Distributing healthcare resources more equitably.
  - Targeting vaccinations.
  - Designing policy to help areas most in need.

GATHERING DATA

- Gathered COVID-19 data and socio-demographic data for Chicago.
- COVID-19 data was collected from the Chicago Data portal:
  - Included COVID-19 case/death data along with the victim's ZIP code.
    - Link: https://data.cityofchicago.org/browse?limitTo=datasets&sortBy=alpha&tags=covid-19
- Socio-demographic data was collected from the CensusReporter website:
  - Scraped ZIP code-based data to match granularity of COVID-19 data.
    - Link: https://censusreporter.org/profiles/86000US60607-60607/

CHICAGO DATA PORTAL

+



Census Reporter

60605

ZIP Code Tabulation Area in: Cook County, IL, Chicago-Naperville-Elgin, IL-IN-WI Metro Area, Illinois, United States

30,950
Population

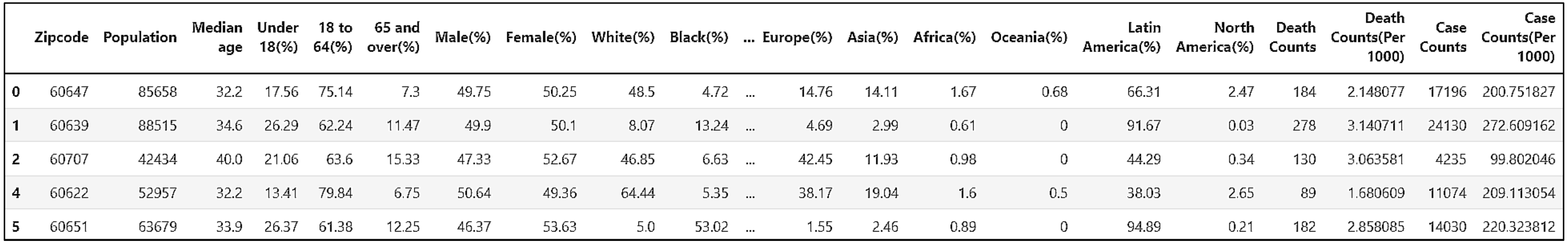
1.2 square miles

24,851.1 people per square mile

Census data: ACS 2020 5-year unless noted

CLEANING DATA

- Removed instances of Covid death where:
  - manner of death was accident or suicide
  - ZIP code was outside of Chicago
- Removed unneeded columns
- Merged the datasets:
  - ○ Each line represents a ZIP code with its socio-demographic and COVID-19 data.
- Normalized Covid deaths and cases by each ZIP code's population:
  - ○ Cases/deaths per 1000.

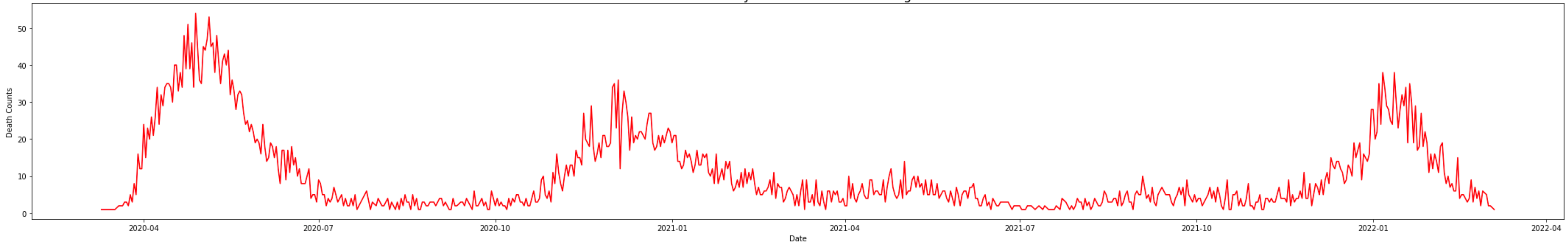


	Zipcode	Population	Median age	Under 18(%)	18 to 64(%)	65 and over(%)	Male(%)	Female(%)	White(%)	Black(%)	...	Europe(%)	Asia(%)	Africa(%)	Oceania(%)	Latin America(%)	North America(%)	Death Counts	Death Counts(Per 1000)	Case Counts	Case Counts(Per 1000)
0	60647	85658	32.2	17.56	75.14	7.3	49.75	50.25	48.5	4.72	...	14.76	14.11	1.67	0.68	66.31	2.47	184	2.148077	17196	200.751827
1	60639	88515	34.6	26.29	62.24	11.47	49.9	50.1	8.07	13.24	...	4.69	2.99	0.61	0	91.67	0.03	278	3.140711	24130	272.609162
2	60707	42434	40.0	21.06	63.6	15.33	47.33	52.67	46.85	6.63	...	42.45	11.93	0.98	0	44.29	0.34	130	3.063581	4235	99.802046
4	60622	52957	32.2	13.41	79.84	6.75	50.64	49.36	64.44	5.35	...	38.17	19.04	1.6	0.5	38.03	2.65	89	1.680609	11074	209.113054
5	60651	63679	33.9	26.37	61.38	12.25	46.37	53.63	5.0	53.02	...	1.55	2.46	0.89	0	94.89	0.21	182	2.858085	14030	220.323812

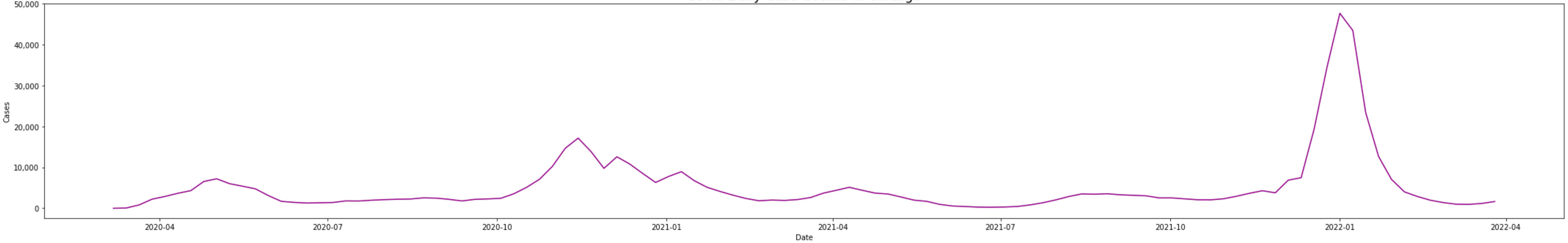
EDA AND VISUALIZATIONS

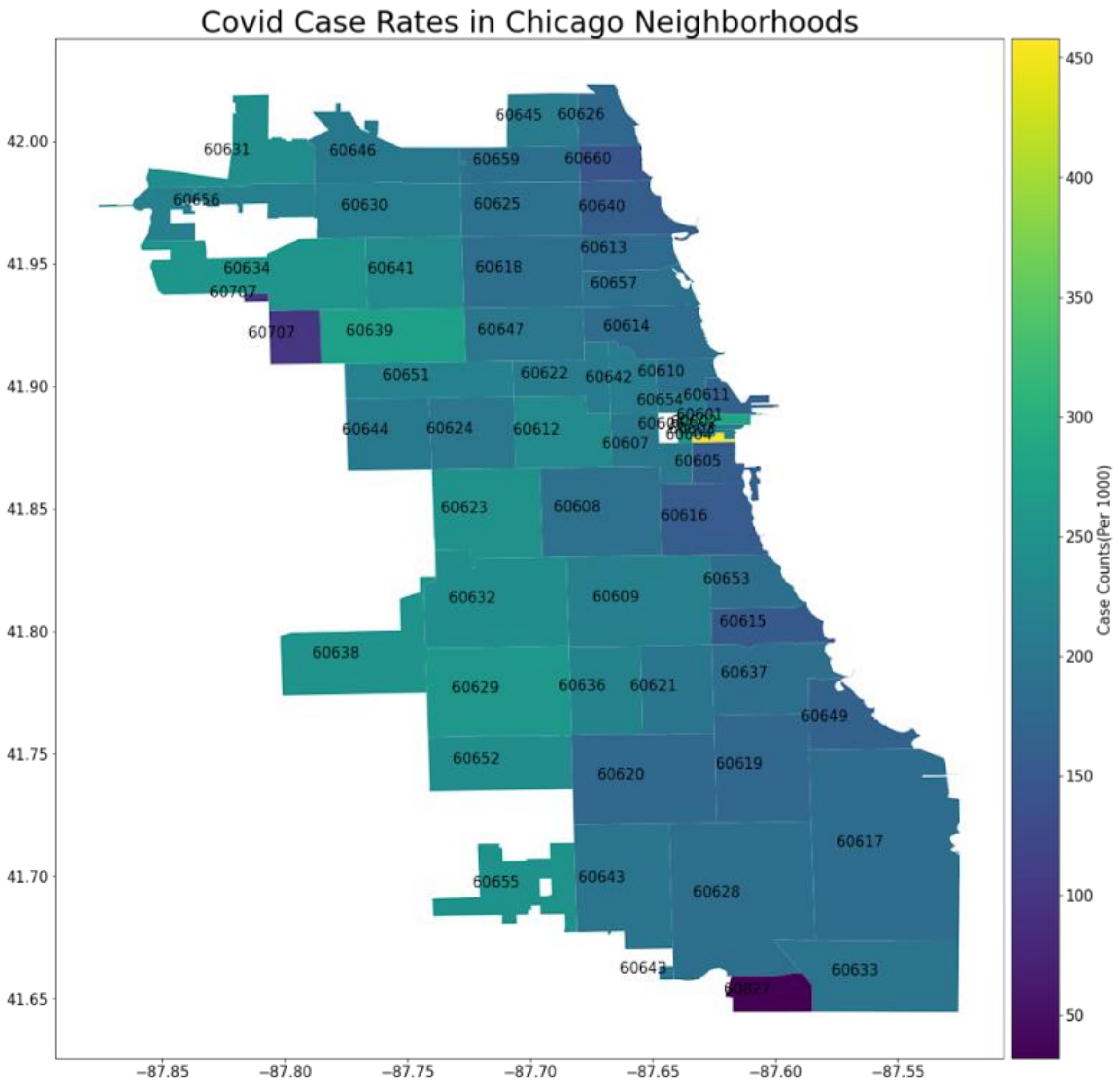
- For EDA, we looked at the correlations between different socio-demographic factors and COVID-19 data.
- We created some visualizations to better understand these relationships.

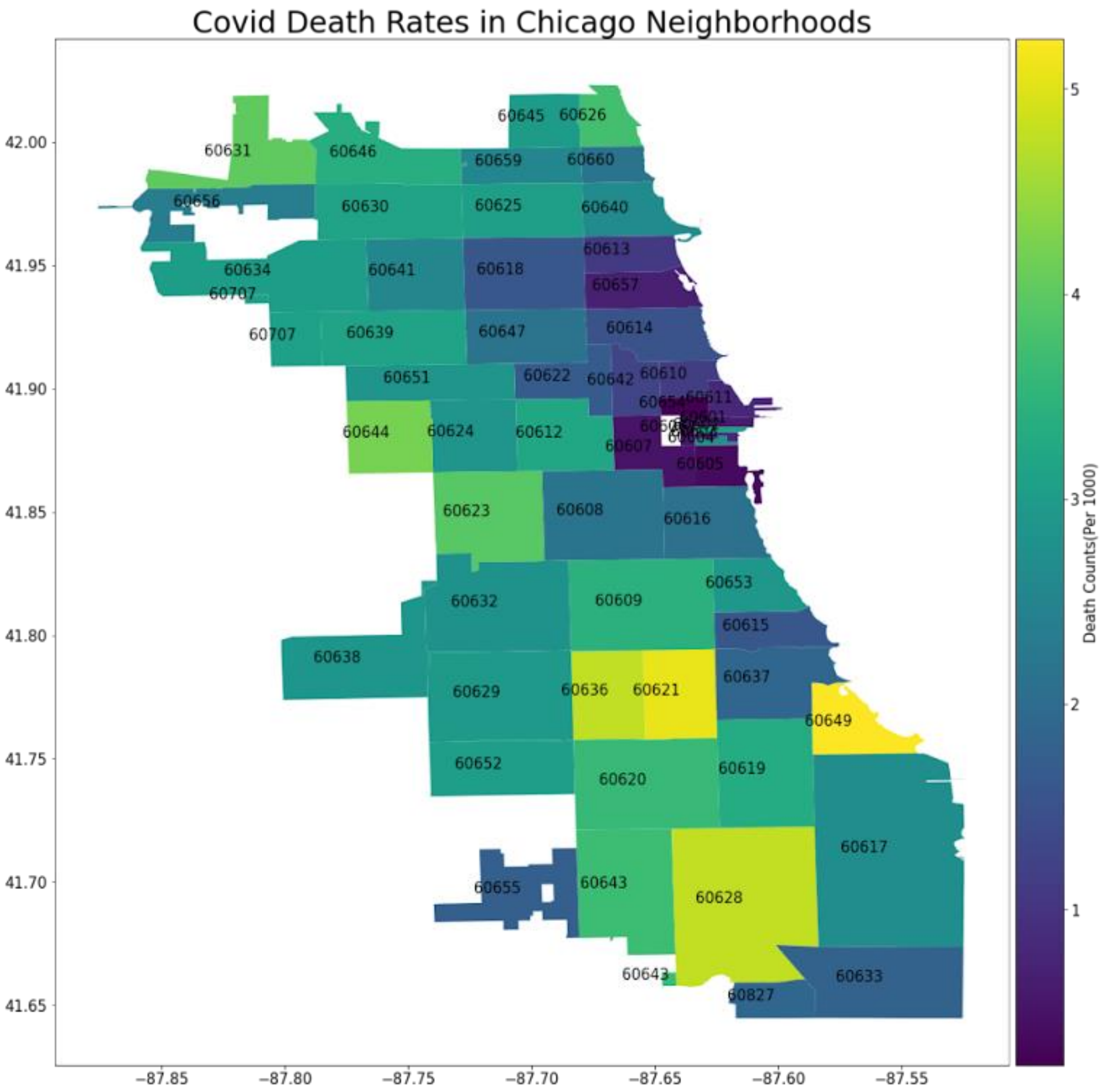
Covid Daily Death Counts in Chicago



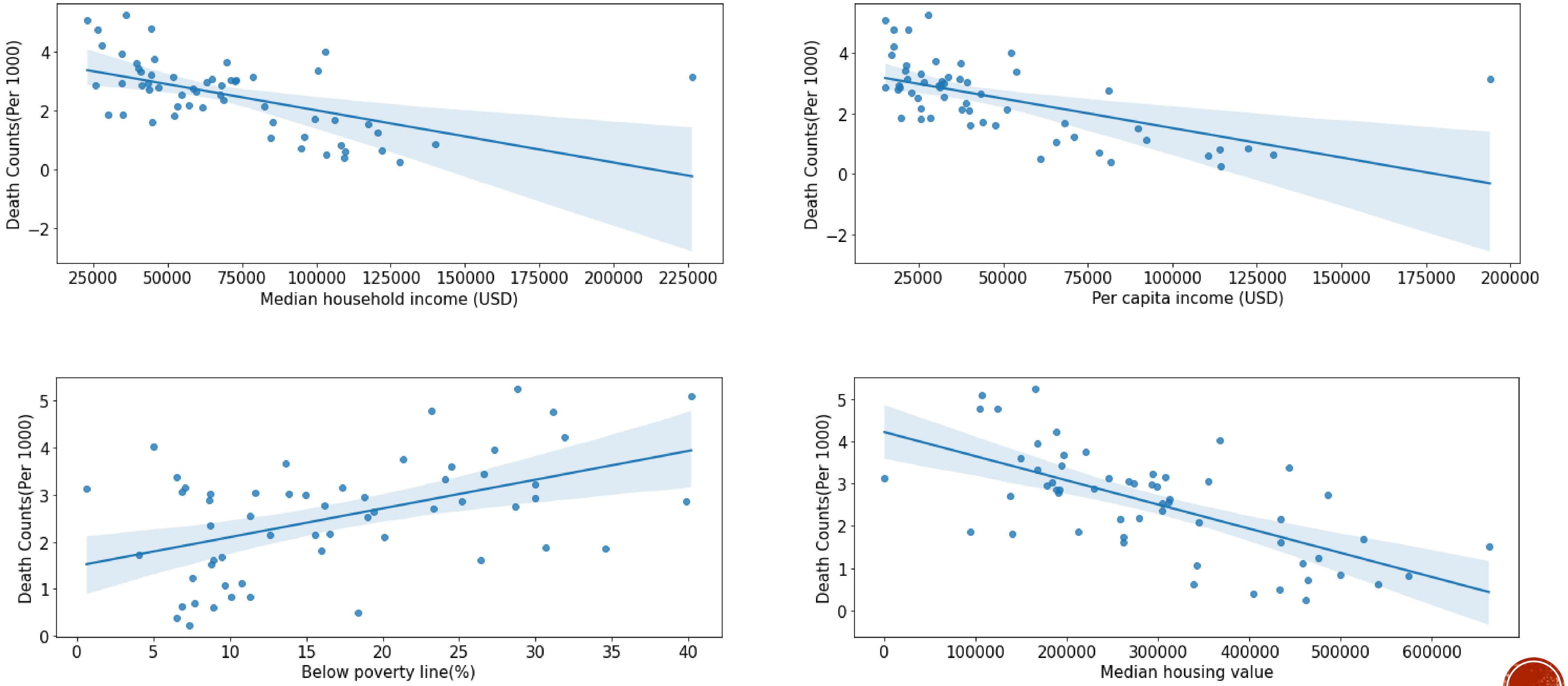
Covid Daily Case Counts in Chicago




Covid Case Rates in Chicago Neighborhoods

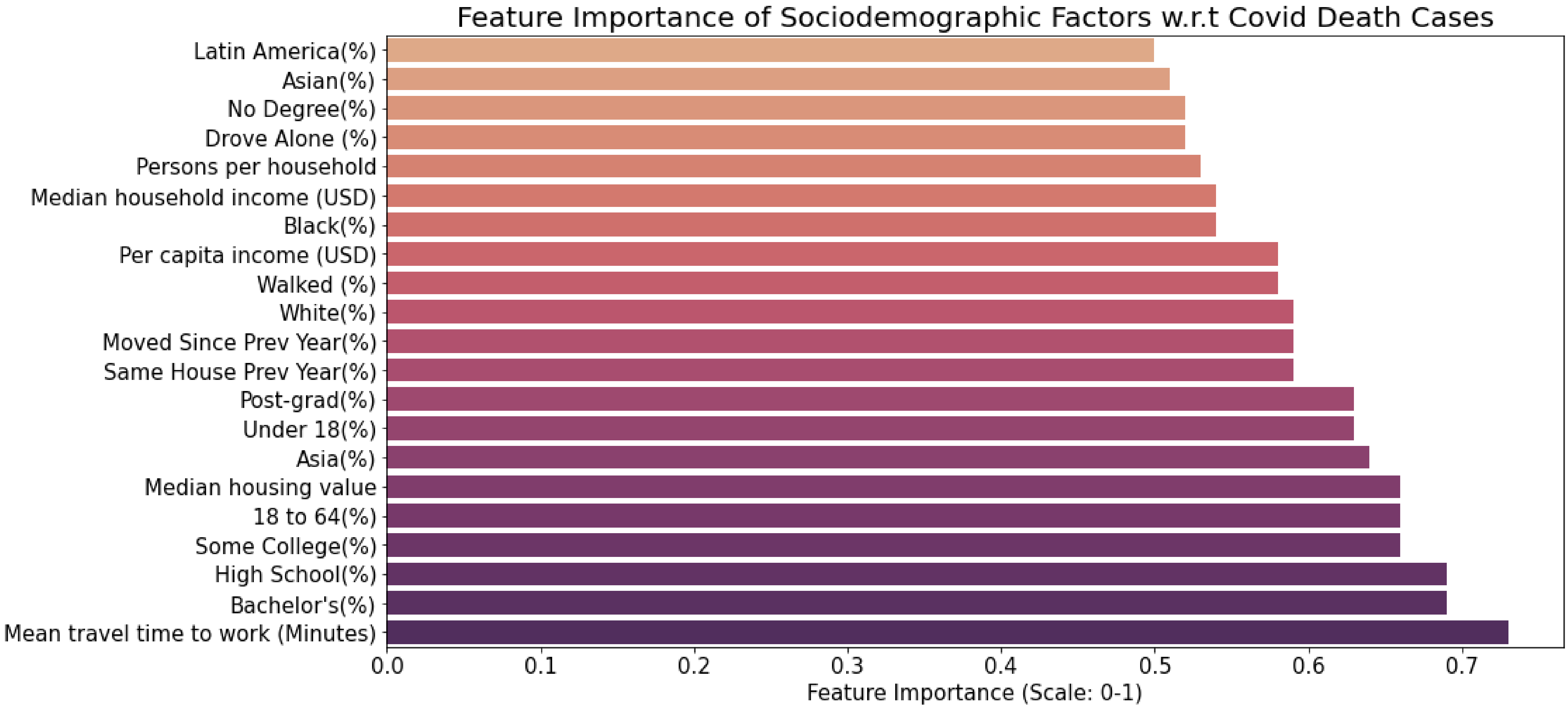

Covid Death Rates in Chicago Neighborhoods

Correlation of Sociodemographic factors with Covid Death Cases



SELECTING SOCIO-DEMOGRAPHIC FACTORS

- Find most important features for predicting COVID-19 vulnerability.
- The importance (on a 0-1 scale) indicates a correlation between a socio-demographic factor and COVID-19 death rate(1 being the highest correlation).
- Selected features with an importance of above 0.5.



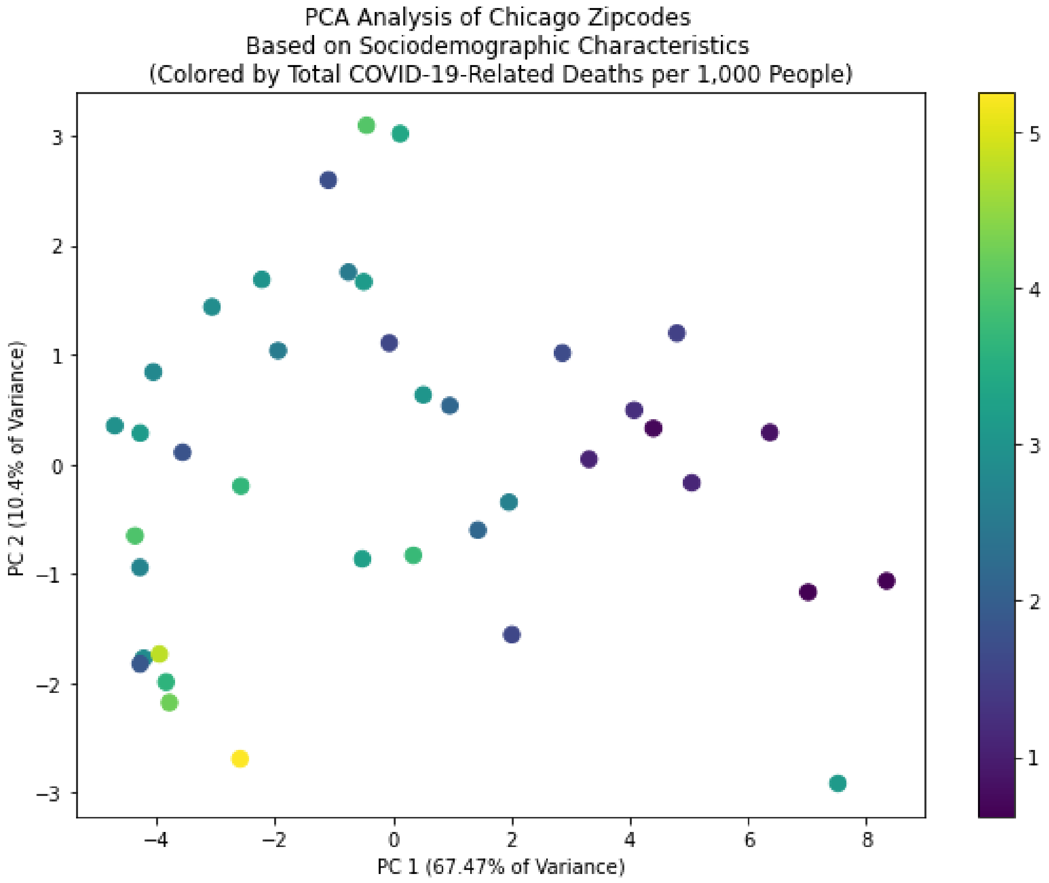
Feature Importance of Sociodemographic Factors w.r.t Covid Death Cases

RANDOM FOREST REGRESSION MODEL

Trained baseline model and RFR model using selected features with importance > 0.5

	Baseline model	**RFR model**
Data splitting(train:test)	70:30	70:30
Definition	Predict all as median death rate	Random forest regression
Hyper parameter tuning	N/A	Randomized search on hyper parameters
Cross validation	N/A	5 folds of 2 splits
Average absolute error	1.03 deaths/1000 people	0.62 deaths/1000 people

PRINCIPAL COMPONENT ANALYSIS



- PCA to visualize the distribution of COVID-19-related death rates across factors.
- Only training data from RFR model was used for this analysis.
- We found a pattern between socio-demographic factors and COVID-19 deaths.
- Substantial amount of noise present in the data.

XGBOOST MODEL

- 70% training data, 30% testing data
- Socio-demographic factors with correlation coefficients >0.5 were selected.
- Average absolute baseline error = 0.96 deaths per 1000.
- Average absolute model error = 0.63 deaths per 1000.

KEY TAKEAWAYS

- 21 of the 48 socio-demographic factors from census data showed strong correlation to COVID-19 impact.
- Some of the most important indicators for COVID-19 impact were:
  - Travel time to work
  - Education level
  - Age
- Principal Component Analysis showed pattern between COVID-19 deaths and socio-demographic factors.
- Our RFR model predicted COVID-19 death rate with an error rate of 0.62 deaths/1000.
- Our XGBoost model predicted COVID-19 death rate with a model error rate of 0.63 deaths/1000

IMPROVING THE MODEL

- Our current model only incorporates 60 ZIP codes.
- We are currently in the process of incorporating more ZIP code based data into our model.
- This new data is from different cities and states in America.
- More data points will allow us to train a more accurate model and reduce our model error rate.

THANK YOU