

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 Rafeey 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>

Group repository: <https://github.com/uic-cs418/cs418-spring22-bored-grad-yacht-club>

Our motivation for this project

- The COVID-19 pandemic has been a historic, life-changing and terribly unfortunate event in our lives.
- As data science students, we were very interested in the widely available data for this pandemic.
- The main question we landed on through our initial research was:
 - **Is there a way to link COVID-19 impact to socio-demographic data?**
- With our project, we hope to answer this question!

The background of the slide is a dense, colorful illustration of a crowd of people. Each person is depicted from the chest up, wearing a white face mask. The individuals have various hairstyles, colors of hair, and are wearing different colored clothing (shirts, sweaters, etc.). The overall style is a soft, painterly illustration. The word "Definitions" is centered in the upper half of the image in a dark blue, sans-serif font.

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.
 - Number of COVID-19 deaths.

Where we started off

- We focused on COVID-19 data for the city of Chicago.
- We found an existing COVID-19 risk metric called the CCVI index:
 - This model ranks Chicago neighborhoods on sociodemographic data.
 - Assigns a COVID risk score for each neighborhood based on the rankings.
- A potential flaw with this model is that the score relies on neighborhood rankings.
- For example, 2 neighborhoods could be one rank apart in terms of income, but have vastly different average income compared to the next ranked neighborhood.
- Our model aims to improve upon this index.

Our COVID-19 model

- With our model, we aim to achieve:
 - Quantifiability of COVID-19 impact
 - The model should be able to predict the number of COVID-19 cases/deaths based on socio-demographic data.
 - Accuracy and uniformity
 - The predictions should be based on exact variable values, not rankings.
- 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.
- Socio-demographic data was collected from the CensusReporter website:
 - Scraped ZIP code-based data to match granularity of COVID-19 data.

Cleaning data

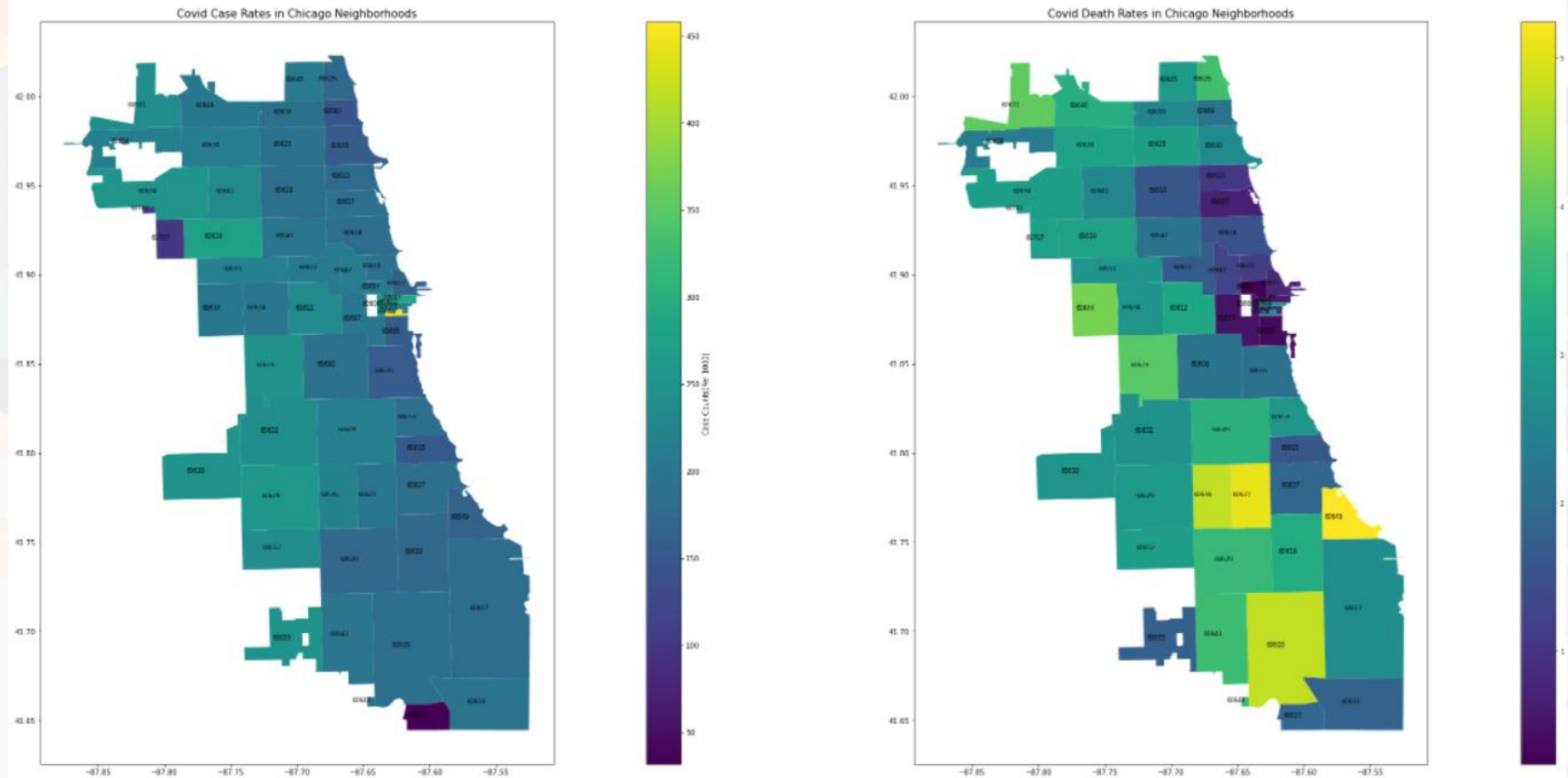
- **Cleaned COVID-19 data for Chicago:**
 - Removed instances of death where manner of death was accident or suicide.
 - Removed ZIP codes outside of Chicago.
 - Removed unneeded columns.
 - Aggregated cases and deaths per ZIP code.
- **Cleaned socio-demographic data for Chicago:**
 - Removed unneeded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

The background of the slide is a dense, colorful illustration of a diverse crowd of people. Each person is wearing a white face mask. The individuals have various hairstyles, skin tones, and are wearing different colored clothing. The overall style is a flat, modern illustration with soft colors like pinks, yellows, and blues.

EDA and Visualizations

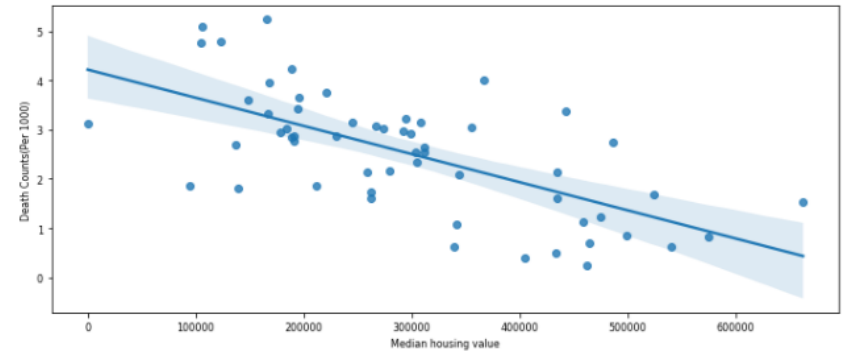
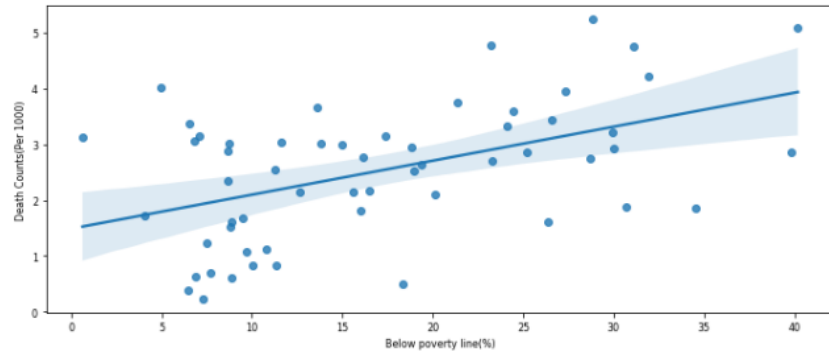
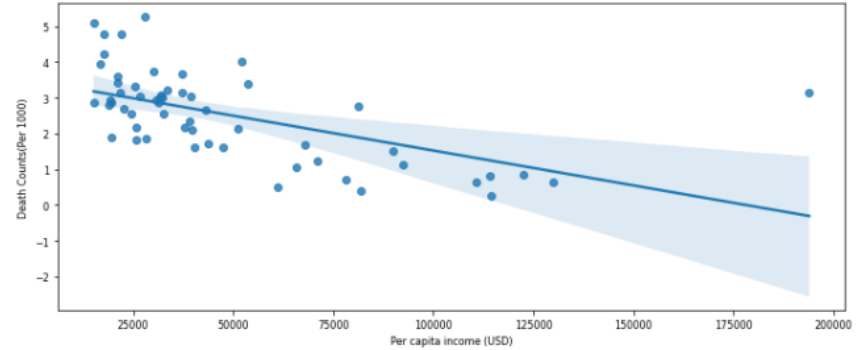
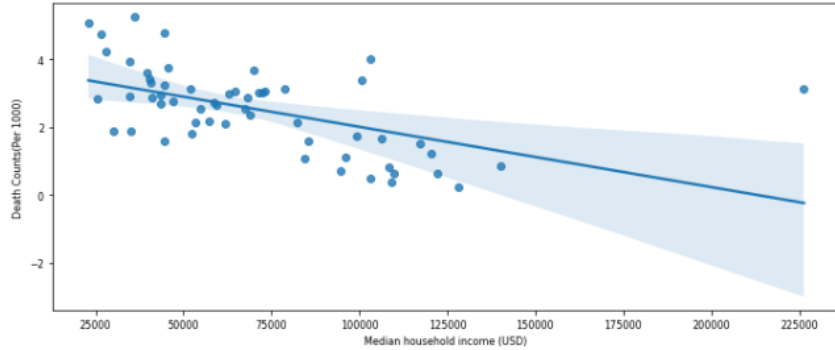
- For our EDA, we looked at the correlations between different socio-demographic factors and COVID-19 data.
- We created some visualizations to better understand these relationships.

EDA and Visualizations



EDA and Visualizations

Correlation of Sociodemographic factors with Covid Death Cases



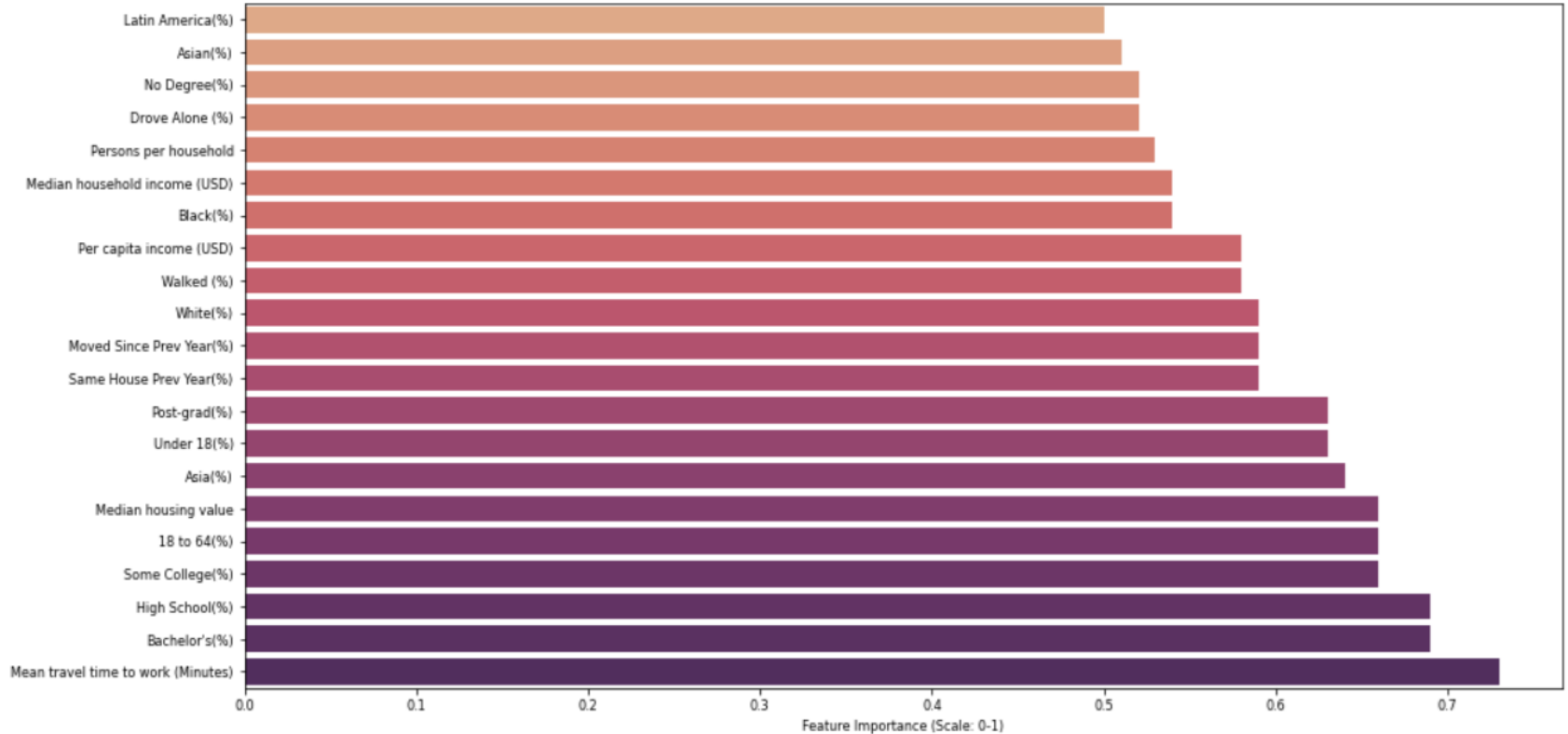
The background of the slide features a repeating pattern of stylized, colorful illustrations of people of various ethnicities and ages, all wearing white face masks. The colors used for the people include shades of blue, green, yellow, orange, and red, creating a vibrant yet muted backdrop.

Selecting socio-demographic factors

- We wanted to see which features were most important to predicting COVID-19 cases/deaths.
- The importance (on a 0-1 scale) indicates a correlation between that socio-demographic factor and COVID-19 (1 being the highest correlation).
- We selected features with an importance of above 0.5.

Selecting socio-demographic factors

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



Random forest regression model

- Baseline model: deaths/1000 equal to median for all ZIP codes.
- Average absolute baseline error = 1.03 deaths per 1000.
- Data split into 70% training data, 30% testing data.
- Socio-demographic factors with correlation coefficients >0.5 were selected.
- Optimizing hyperparameters for RFR:
 - Randomized search strategy across a grid of possible hyperparameter values.
 - Repeated K-fold cross validation, 5 repeats of 2 splits for each randomly-selected combination.
- Average absolute model error = 0.62 deaths per 1000.

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 an 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.

The background of the slide is a dense, repeating pattern of stylized human faces. Each face is depicted with simple, flat colors and is wearing a white surgical-style face mask. The faces vary in hair color (brown, blonde, black, grey) and style (short, long, curly, straight), as well as facial features like glasses. The overall effect is a sense of a large, diverse group of people, all united by the common action of wearing a mask.

Thank you!

Questions?