How racism and neighbors tell the covid-19 risk for Chicago citizen

CAPP 30122, Winter 2021, Final Project

Qiwei Lin (qiweilin) Yangzhou Ou (yangzhou) Javier Rojas (jarojasag) Xiaoting Sun (sunxt)

(Ordered by Last name)

Overview

Motivated by the surging pandemic in Chicago caused by the covid-19, we reckon with the complexity of races and socioeconomic conditions on community levels. According to IDPH statistics, 8.2% of Illinois vaccine doses have been given to Black residents--who currently make up 14% of the state's population. Thus, this project aims to predict the different outcomes with different racial composition in a specified zip code area in Chicago. Also, we provide users with information of their surroundings by comparing a specified zip code area with its neighbors on covid-19 cases, deaths, tests and doses. The product of this project is a user-interactive front-end dashboard which enables an individual using community-level variables to create insights into the ways that better his/her community.

Software Architecture

To launch the project, just run **run.py**, this file will run the following three files: **i) data_collect.py, ii) data_processing.py, iii) dashboard_pca.py**. The remainder of this section will describe the software architecture.

We collected data from three APIs: Chicago Data Portal API, Chicago Health Atlas API, City Health Dashboard API. Data from these three sources was put into Pandas dataframes through **data_collect.py**.

After data collection, we ran **data_processing.py** to clean and transform the raw data into some tabular data available for analysis. Before we launched the interactive dashboard, we built two other helper modules to facilitate modeling and prediction: i) **data_analyzing.py**: comparing a zip code area with its neighbors on a variable selected by users by K-nearest neighbors algorithm, ii) **data_modeling.py**: performing principal component analysis and generalized linear regression for inference and prediction

The interactive dashboard can be launched by running ipython **dashboard_pca.py**. The dashboard has four panels: i) Top Left: a time series plot of a variable of interest in a zip code specified by the user; ii) Top Right: a bar chart representation of prediction (conditional on the majority race) of a variable selected by users in the same zip code as the Top Left panel; iii)

Bottom Right: a PCA correlation plot (the first principal component is fixed but users can select the second one) and iv) Bottom Left: a PCA scatter plot with two principal components on x- and y-axis and a highlighted red dot representing the selected zip code.

We also tried to build a sqlite3 database: we cleaned through **data_cleaning.py**. Our database, **covid_research.sqlite3**, was created through **build_sql_database.py** by calling the SQL setup script **create_table.sql**. This step is unfinished before the deadline but it will not affect the users' experience using the interactive dashboard.

Accomplishments, Shortcomings and Future Directions

We achieved the main goal of creating an interactive data visualization dashboard that takes users' input of community-level features and presents the predicted indicators associated with the community with those features. We also build the predicting model with principal community public health and demographic indicators.

This was achieved thanks to the use of analysis in principal components and generalized linear models. The first, allowed us to summarize the joint behavior of all the variables in the dataset, whereas the second allows us to quantify the effect that race composition within a zip code has on outcomes. We found that the number of cases correlate with the number of young persons in each zip code and thus, the vaccine efforts are being focused there, however, this is done at the expense of older people who are not being vaccinated in numbers that resemble the contagions around them. Furthermore, we find that testing doesn't drive vaccination nor is a useful predictor for either confirmed cases or deaths. Finally and thanks to the generalized linear models used, we found substantial evidence of heterogeneity in outcomes driven by race composition within each zip code.

Even though we wanted to include more socioeconomic variables and produce a balanced panel data for all the zip code areas, data limitations prevented us from doing so. The incompleteness of the data affects both visualization and prediction. Future work can be built upon datasets with fewer missing values and generate insights by applying other advanced machine learning algorithms.

Appendix

- **Dashboard Variables**

Column name	Type	Description
zip_code	int64	Zip code
tests_weekly	int64	Number of tests in the week
tests_cumulative	int64	Number of tests through the week
test_rate_weekly	float64	Test rate per 100,000 population of the week
test_rate_cumulative	float64	Cumulative test rate per 100,000 population
percent_tested_positive_weekly	float64	Percentage of per 100,000 population who tested positively
percent_tested_positive_cumulative	float64	Cumulative percentage of per 100,000 population who tested positively
deaths_weekly	int64	Number of deaths in the week
deaths_cumulative	int64	Cumulative number of deaths
death_rate_weekly	float64	Death rate per 100,000 population of the week
death_rate_cumulative	float64	Death rate per 100,000 population through the week
population	int64	Zip code population
cases_weekly	float64	Number of cases in the week
cases_cumulative	float64	Total number of cases through the week
case_rate_weekly	float64	Case rate per 100,000 population in the week.
case_rate_cumulative	float64	Total case rate per 100,000 population through the week.
total_doses_daily	int64	Number of doses administered on the date.
total_doses_cumulative	int64	Total number of doses administered through the date.
_1st_dose_daily	int64	Number of people who received a first vaccine dose on the date.
_1st_dose_cumulative	int64	Total number of people with at least one vaccine dose through the date.
_1st_dose_percent_populat	float64	Percentage of population with at least

ion		one vaccine dose on the date
vaccine_series_completed_daily	int64	Daily sum of completed vaccine series
vaccine_series_completed_cumulative	int64	Cumulative sum of completed vaccine series
vaccine_series_completed_percent_population	float64	Percentage of population which have completed vaccine series
vaccionation_sites	float64	Number of vaccination sites in each zip code zones
health_centers	float64	Number of health centers in each zip code zones
number_of_hospitals	float64	Number of hospitals in each zip code zones