CDC County/DC Vulnerability - Housing Dashboard

Amani Desormeaux, Emily Wilt, and Berkeley Poulsen

Project Github:

https://github.com/BerkeleyPoulsen/447-Project

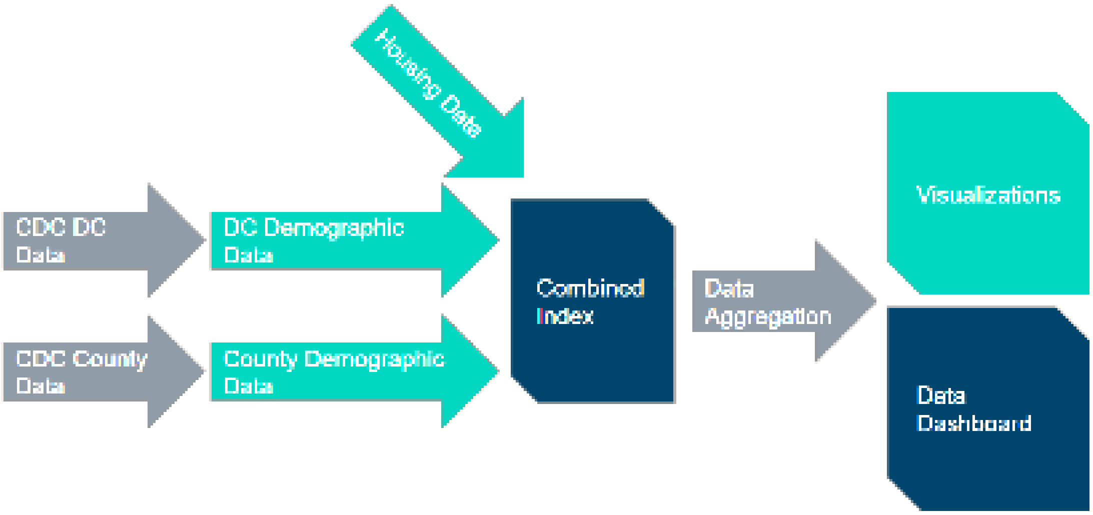
Introduction

Our project is the culmination of research involving real estate data and CDC gathered data involving the vulnerability of communities. Communities that are necessarily vulnerable, as we found, often have higher populations and are greater in urban density. A dashboard was made to provide an interface for users to access basic information about every country as well as the District of Columbia, and let a user find data about those regions in an informative way.

Data was collected through CDC data primarily and was merged with data found through Maryland and DC open data resources. Z scores were found with the CDC and housing real-estate variables, allowing us to see that more populated regions is correlated with higher scores on the vulnerability score and with real-estate development. This correlation may show that people realize that a higher population means that people involve themselves in new real-estate, and that vulnerability is a result of more people in an area. Urbanization is perhaps correlated with a higher community vulnerability.

Method

Our Data Pipeline was as follows: Gather, clean and build Data from CDC and open data resources and find areas where the data could fit together. The main goal of our data manipulation was to build a dataframe that could be read into a dashboard.The dashboard was then built off of a modified dataframe that included aggregations of data. Openrefine was not used to an extent because almost every variable in our final data set is a numerical variable.

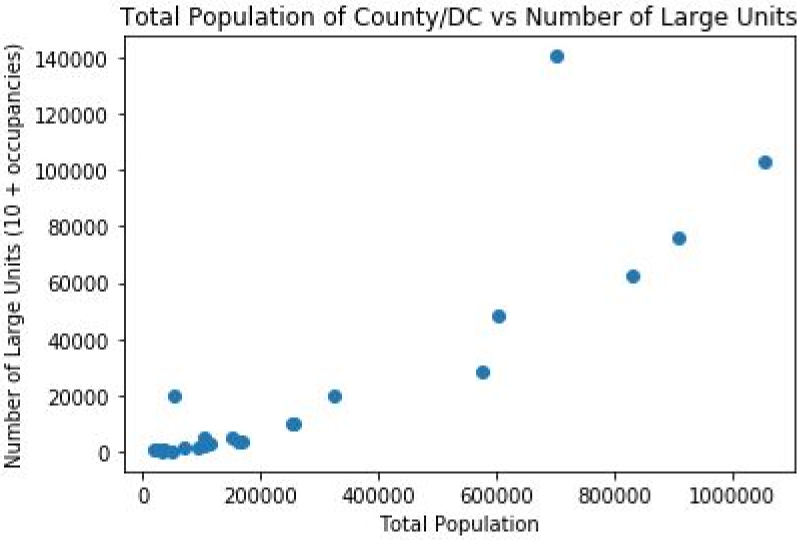


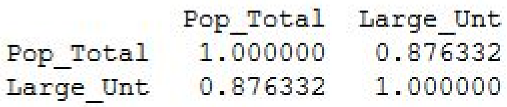
Above is a summary illustration of the general method we used to refine data into a dashboard as well as visualizations. Our main question and problem solution involved combining DC CDC and Demographic Data, as well as County CDC and Demographic Data, and then finding our housing information about those regions. A combined index was made with all the data possible in one dataframe. The data was then sent to a Data Dashboard and Visualizations were made from the main index. Z scores for CDC and housing Data were also found and placed into the same index, allowing us to see particularly high disparities at a county level.

Preprocessing

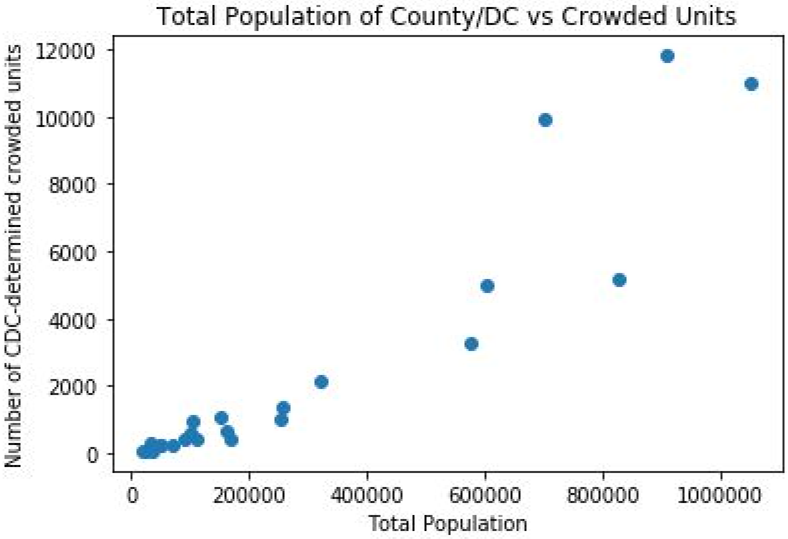
Because our project was mostly based on numerical variables involving multiple datasets, our preprocessing was limited, and not much was done besides changing variables based on the merging-availability at the time. Most processing was done after the fact with the inclusion of merging, finding a z score, and changing the state of dataframes to match those in other parts of our analysis.

Results





```
            Pop_Total  Large_Unt
Pop_Total    1.000000   0.876332
Large_Unt    0.876332   1.000000
```



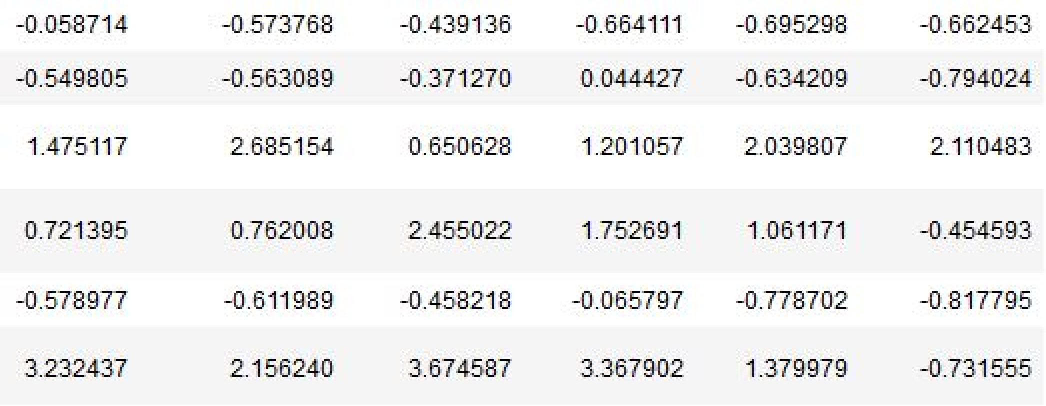
Our results are a conclusion of finding multiple correlations given our final global analysis and getting multiple variables together to see which ones had the highest correlations.

What we found is as follows:

The higher a population in a region, the higher it scores in CDC vulnerability.

The figure below demonstrates the disparity of certain variable values, such as DC being 3.6 standard deviations from the mean in terms of the variable “not having a vehicle”.

Figure: Z scores of county and DC Vulnerability.



-0.058714	-0.573768	-0.439136	-0.664111	-0.695298	-0.662453
-0.549805	-0.563089	-0.371270	0.044427	-0.634209	-0.794024
1.475117	2.685154	0.650628	1.201057	2.039807	2.110483
0.721395	0.762008	2.455022	1.752691	1.061171	-0.454593
-0.578977	-0.611989	-0.458218	-0.065797	-0.778702	-0.817795
3.232437	2.156240	3.674587	3.367902	1.379979	-0.731555

Our findings suggest that higher regional vulnerabilities are correlated with the total population and population density of a region, due to correlations found in the two graphs showing similar trends. Data put into our dashboard suggests that almost every region held a similar order of magnitude for their data index values, and most counties are within one standard deviation of the mean in terms of CDC vulnerability. Our group believes that the inherent differences in the infrastructure and population densities of cities, compared to the suburbs or rural areas, can explain this disparity. The lowest record on the graph above is the z scores found for variables explained by DC vulnerability.

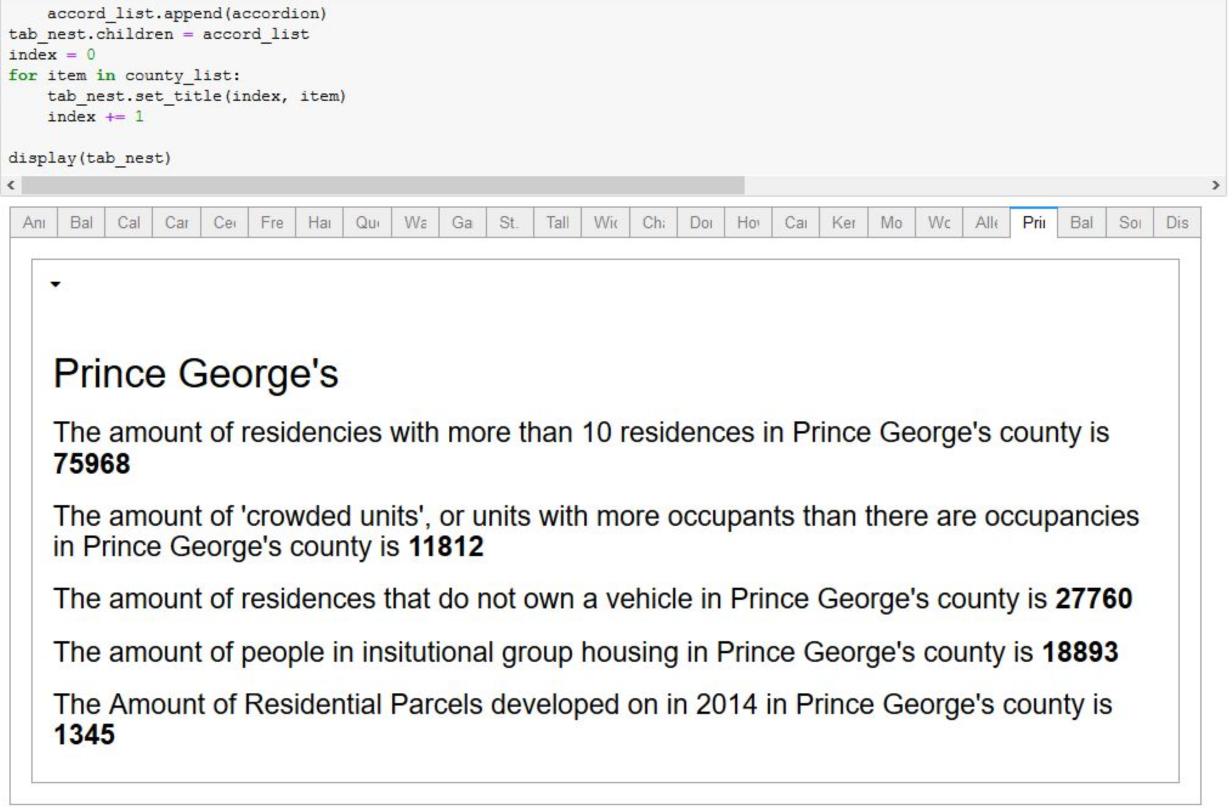
Our findings also suggest that there is no correlation with the amount of residential parcels built in a region and the population of that region. We think this is due to the fact that there is a large disparity in the amount of units built when a residential parcel is built on. For instance, a DC residential parcel may be an apartment block that can hold 1000 residents, but the same parcel can be seen in a more rural county, and will only have one house built on it, holding 3 to 5 people.

DC and PG county being large population zones relate to how vulnerable communities are when it comes to higher population density and higher total population. This makes sense from a practical standpoint as more people in a region necessarily means that there are more chances for disease spreading, and roads are more crowded in disaster situations.

To conclude, population is a higher indicator of vulnerability than anything else as it is the most implicated in the entire study. Our dashboard provides some insight into this as we found that the numbers correlated in the CDC index are correlated with the population of a region.

Data Dashboard

Part of our project included a data dashboard that can show a user different counties as well as DC in context, and provide information about them in a user-friendly way. We used ipython widgets, as well as a custom made HTML-ui to develop the data dashboard for our project. Future directions may include more information sent to the HTML front end, but at this stage summary information for every county and DC is shown. Looking at this summary information actually gave some insight as to why we think population necessarily implicates CDC vulnerability.



```
    accord_list.append(accordion)
tab_nest.children = accord_list
index = 0
for item in county_list:
    tab_nest.set_title(index, item)
    index += 1

display(tab_nest)
```

Prince George's

The amount of residencies with more than 10 residences in Prince George's county is **75968**

The amount of 'crowded units', or units with more occupants than there are occupancies in Prince George's county is **11812**

The amount of residences that do not own a vehicle in Prince George's county is **27760**

The amount of people in insitutional group housing in Prince George's county is **18893**

The Amount of Residential Parcels developed on in 2014 in Prince George's county is **1345**

Limitations

Real Estate data was hard to come by, and often correlated with every variable in the CDC vulnerability index. More pieces of the index can be implicated in this study to determine more implications from the CDC index.

Confounding variables, namely those implicated with the population of a region, can most likely explain our project findings.

Combining DC data and Maryland County data produced some disparity.

Z scores don’t explain everything. We had to really think about why DC, PG county and others were heavily beyond the standard deviation. The explanation seems to be that population and density are implicated with this score.