Project Healthy Homes

By Stacey Beck and
Bradley Thompson

Professor Chris Teplovs, Ph.D
SIADS 591 & 592
First Milestone Project
University of Michigan MADS Program

May 21th, 2020

Acknowledgements

We want to thank Sherri Dixon, PhD, Senior Biostatistician at National Center for Healthy Housing, for her guidance on understanding available housing data. We also want to thank Chase Haller, Neighborhood Christian Legal Clinic, for all of his help in creating appropriate weights for the factors that determine whether an apartment is legally substandard.

Table of Contents

Appendices:

Motivation

Conditions in places where people live, learn, work, and play affect a wide range of health risks and outcomes. These conditions are known as the social determinants of health, or SDoH. We know that health is compromised in communities with poor SDoH such as inadequate housing, low income, unsafe neighborhoods, or substandard education. To improve health, therefore, both policymakers and healthcare professionals are focused on improving the SDoH for the most vulnerable.

As a concrete example, in the current COVID-19 crisis, experts are concerned that substandard housing will cause its inhabitants to be more susceptible to the virus. An obvious example is the lack of running water, because washing hands is an important step to reducing the risk. Substandard housing conditions also make it simply more difficult logistically to shelter in place. At least one author has asserted that substandard housing can substantially increase the risk of COVID-19.[1]

As a part of an overall project involving other MADS students called Project Protect, we are assessing the need for legal aid to assist indigent Americans struggling with the SDoH. The broader group is working to compile data that will show the number of instances Americans need access to legal aid to help improve the SDoH, and consequently health itself. Our part of that overall project is to calculate the number of Americans who need legal aid to compel landlords to make improvements to substandard apartments. Our analysis on substandard housing will feed into Project Protect's larger work product, as other students look at the need for legal aid to address such SDoH as violence against women, the rate of foreclosures and evictions, and access to needed government benefits.

In summary, in this phase of that broader project, we tried to answer two specific questions:

1) Estimating the number of Americans living below 150% of the Federal Poverty Level who likely have a legal claim against a landlord for substandard housing. This number will be used for Project Protect and combined with other analysis on the other SDoH by other MADS students, and
2) Understanding the relationship between the existence of substandard housing and the risk of contracting COVID-19. This second goal reflects our desire to better understand the importance of adequate housing, as a SDoH, to protect from COVID-19.

[1] "America's inequitable housing system is completely unprepared for coronavirus" by Jenny Schuetz of Brookings. https://www.brookings.edu/blog/the-avenue/2020/03/12/americas-inequitable-housing-system-is-completely-unprepared-for-coronavirus/amp/

Results

Measure of Substandard Housing

Based on our data analysis, we estimate that 4,840,000 households in 119 large metropolitan areas across United States have household incomes under 150% of the Federal Poverty Guidelines and live in apartments under such conditions that a legal aid attorney would likely (greater than roughly 30 percent chance)[2] be willing to represent the household to assert a claim that the apartment is unfit or in violation of local building codes. These households therefore would benefit from having access to a legal aid attorney to help them improve their living conditions.

The map below in the visualization section shows where those households are located. Our analysis was limited to 119 metropolitan areas because those are the areas where we had data from the American Housing Survey to analyze over the last seven years. The analysis therefore excludes 34.65 % of the US population.

Relationship between Substandard Housing Units and Incidence of COVID-19

We failed to confirm our hypothesis that there is a relationship between these substandard housing units and the incidence of COVID-19, taking into account such other likely variables such as population size. When we removed population from the equation, the data showed essentially no correlation between the prevalence of substandard housing units and the incidence of COVID-19. Testing this hypothesis more thoroughly would require finding housing data that is more granular than the public use database of the American Housing Survey.

The public use data version of that survey, to protect the confidentiality of survey subjects, includes geographic units only down to the metropolitan area. Those metropolitan areas, as you likely know, include typically many counties. Our understanding of the science and public health characteristics of COVID-19 suggest that the incidence varies considerably ZIP Code to ZIP Code. Thus, geographic measurements that are aggregated to the metropolitan area level are unlikely to demonstrate any useful information about the correlation between substandard housing--which also varies considerably ZIP Code to ZIP Code – and the incidence of COVID-19.

Companion Files

Our code can be accessed here.[3]

[2] We say roughly both because this is based on a subjective assessment by a legal aid attorney, but also because we fully recognize that simply adding three different conditions each with a 10% chance a legal aid housing professional would take the case does not yield exactly a 30% chance. We discuss the significance of the weights below.

[3] Link to our Google Colab notebook:
https://colab.research.google.com/drive/1TYDlHqTie7KsYjmIGaR2PD7UGQvKzINZ

Data Sources

We used the American Housing Survey from the years 2017, 2015, and 2013 created by the Department of Housing and Urban Development (HUD) and contained in census data because these versions of the survey include very specific and pointed questions about the condition of homes in which people dwell, and includes different metro areas for each of those years surveyed. Every other year, for this housing data the census surveys a group of the largest metropolitan areas, and then on a rotating basis surveys smaller metropolitan areas. Therefore, by combining multiple years, we obtained data on 119 different metropolitan locations. For each metropolitan area, we calculated a percentage of homes surveyed that were substandard.

To estimate the total number of substandard apartments in a given region, we multiplied that percentage by the total number of apartments in the Metropolitan Areas. We obtained the total number of apartments from the 2018 census data.

To discern which surveyed apartments were likely out of compliance with building codes, we consulted an experienced legal aid housing expert who helped us construct an elaborate weighting system.

To determine how many apartment dwellers would qualify for legal aid, we needed to apply the Federal Poverty Guidelines.

To stitch the housing data with the COVID-19 data on geography, we had to develop our own table that correlates the geography units used for COVID-19 with the geography units used in the housing database.

And finally, realizing that population accounted for likely much of the correlation we were seeing, we needed to normalize on population, so we obtained population data for each of the geographic areas.

In order to access all of our data sets from a remote server while using Google Colab, we took full use of Amazon's S3 simple storage service, making those data sets public.

1) American Housing Survey Data for 2017, 2015, 2013:
    - In each case, the following AHS data were:
        - Comma Separated Values
        - All Accessed using Pandas
        - All AHS variables of interest are contained in Appendix C.

    - AHS 2017 National PUF v3.0 CSV, unzipped located in household.csv
    - 66752 rows, 1090 columns, 463 MB
    - Direct CSV

- AHS 2017 Metropolitan PUF v2.0 CSV, unzipped located in household.csv
- 114035 rows, 897 columns, 160.2 MB
- Direct CSV

- AHS 2015 Metropolitan PUF v3.0 CSV, unzipped located in household.csv
- 24886 rows, 1114 columns, 176 MB
- Direct CSV

- AHS 2013 National PUF v1.3 CSV, unzipped as newhouse.csv
- 84355 rows , 760 columns, 294 MB
- Direct CSV

- AHS 2013 Metropolitan PUF v1.2 CSV, unzipped as newhouse.csv
- 83430 rows, 735 columns, 278.4 MB
- Direct CSV

2) Variable weights assigned to the American Housing Survey Data
   - This is not a public database, but rather weights developed by an outside legal expert, Chase Haller, as discussed more below.
   - We developed the data with Mr. Haller using an Excel spreadsheet that contained the actual language from the variables, both the questions and the answers. Mr. Haller specified the answers that he was interested in, and the weights associated. Given its nonstandard nature in the Excel spreadsheets, we then retyped the information manually into a pandas data frame.
   - The 2017/2015 table is attached as Appendix A.
   - The 2013 table is substantially the same, but the survey developers changed the variable names.

3) Federal Poverty Guidelines for years 2013, 2015 and 2017
   - The federal poverty guidelines are developed and released each year by the United States Health and Human Services Department, and guidelines can be obtained here: https://aspe.hhs.gov/prior-hhs-poverty-guidelines-and-federal-register-references (These guidelines are only applicable to the 48 contiguous states, and we do have one data point for Hawaii. But we did not adjust the Hawaiian numbers.)
   - The guidelines are not in downloadable form, but contain only eight rows of two columns, so we simply retyped the data manually.
   - The 2017 Federal poverty guidelines are attached as Appendix B.
   - The 2015 and 2013 guidelines are substantially the same but slightly lower.

4) Rental data was taken from the 2018 American Community Survey data
    - Census website
    - 3220 lines
    - Read as a CSV into a Pandas DataFrame.
    - The columns of interest were 'id' and the "Estimate!!HOUSING TENURE!!Occupied housing units!!Renter-occupied".

5) COVID-19 data (We used a COVID-19 API to access confirmed cases developed by George Saieed and Ellen Kendall which takes data from the New York Times GitHub repository):
    - COVID-19 API that uses NYT github COVID-19 Data
    - JSON dictionary
    - Access through API then read into PostgreSQL table
    - The length of this API is 2886 rows and the time period of data collected starting on Jan. 21, 2020 and we last updated it through May 17th, 2020.
    - The variables of interest in this API are the number of confirmed cases, county, date, FIPS and state.

6) Population data for the American metropolitan areas from 2018
    - Fetched from an API located at Population API on the census website
    - Contained in a list of lists
    - Inserted one line at a time into a PostgreSQL table we created
    - The length of this API is 3220 rows.
    - The variables of interest were the population value and the state FIPS code as well as the county FIPS code.

7) Bridge Table: We created our own unique crosswalk table, specific for our needs. This table was created using outside sources. We used resources provided by the United States Patent and Trademark Office (USPTO) to get the associated counties and Federal Information Processing Standards (FIPS) codes for each corresponding Metropolitan Area's Core-Based Statistical Area (CBSA) code. Additionally, we used the AHS codebook document in conjunction with The National Bureau of Economic Research's (NBER) crosswalk table to develop the association between CBSA, FIPS and Standard Metropolitan Statistical Area (SMSA) codes to ensure quality control of our work. In order to prevent overlapping of metropolitan areas between the 2015-2017 AHS and the 2013 AHS, we used the AHS 2011 and 2013 Tech Document. We developed this table in Excel and read it into a Spark data frame for further manipulation before inserting it into a table in PostgreSQL for merging with the COVID-19 and population tables. After data manipulation, the table contained 610 rows.

Data Manipulation Methods

The COVID-19, population and rental data were joined with the AHS substandard housing data using codes for geographic areas. The 2015 and 2017 AHS data sets are organized by CBSA codes (formerly referred to as the SMSA codes) and the 2013 AHS data sets are organized by SMSA codes. In order to join each of the data sets, we created a bridge table with a column of CBSA codes, a column of the counties that make up the CBSA with the corresponding five-digit codes (FIPS) (a national code system used to keep track of counties throughout the country)[4], and a column of SMSA codes. The COVID-19, rental and population data sets are organized using FIPS codes, which uniquely identify the counties[5] in the United States[6] for specific metropolitan areas.

Preparing the American Housing Survey Data

For the AHS data, the basic task was filtering. We went through each of the relevant variables and filtered on threshold values to estimate the number of homes that do not meet building codes. Since not all variables are of equal importance, we weighted the results using a scale created by a legal expert in substandard housing.

For that, we consulted with an attorney, Chase Haller, Senior Staff Attorney, Helping Hoosier Homeowners Program, Neighborhood Christian Legal Clinic in Indianapolis, Indiana. Mr. Haller is an expert with 10 years helping indigent clients with substandard housing issues. One of the study authors, Brad Thompson, himself an attorney who worked with Haller for 10 years, assisted by reviewing all of the variables and identifying the ones that Mr. Haller should examine more closely and consider weighting. He determined the weights on a scale of 1 to 10 by assessing the likelihood that he as a legal aid housing professional would take the case based on the answer, assuming that he has room on his docket to do so (refer to Appendix D). We have an excel spreadsheet that lists the variables we used, and the weights that Mr. Haller assigned them.

After assigning the weights, we produced a bar chart of the total number of violations and observed that there was a natural breaking point between a weight of two and a weight of three. We decided that we would make the cut off there as to the apartments we would include in the final tally because that means that a household would have a roughly 25% to 30% chance of getting an attorney to take their case. We say "roughly" because we recognize both the subjectivity of the weighting and that these percentages cannot be simply added together because they are each established independently. Indeed, some apartments based on the survey results were so unfit that they had weights well above 10. But a cut off between two and three seemed to represent an efficient use of everyone's time, both legal aid attorneys and prospective clients.

[4] https://en.wikipedia.org/wiki/Federal_Information_Processing_Standards
[5] https://en.wikipedia.org/wiki/County_(United_States)
[6] https://en.wikipedia.org/wiki/United_States

We also filtered for renters, as opposed to homeowners, because the issue is whether people have a claim against a landlord. We then estimated the number of substandard homes in a given metropolitan area by multiplying the percent of apartment units not meeting code in the survey by the number of apartments in the metropolitan area.

Specifically our manipulation steps included:

1) Import survey results
2) Choose the variables of interest
3) Select just the rows that represent apartments
4) Create and apply weights to the variables of interest
    a. Create table of weights from expert assessment
    b. Apply through a mask
    c. Apply zeros to those remaining
5) Sum weighted values to a total violations column
6) Select just those rows with total values that exceeded two, on the theory that minor violations should not be included
7) Create a table to show percent substandard
    a. Left side is total substandard surveyed
    b. Right side is total surveyed
    c. Divide to get the percent in the survey that were substandard

As a matter of quality control, we also vertically summed each individual column of housing condition variables to see which ones were contributing to the total violations. We did this to make sure that we weren't picking violations that were too common and therefore less impactful, and also to assess the impact of the weighting.

As observed at the beginning, we actually had two different goals, one is the comparison of the prevalence[7] of legally substandard housing with the incidence of COVID-19. The other is to make a heat map of the United States to show the prevalence of substandard housing where the occupant needs legal aid. For the second purpose, we created an additional filter of whether the apartment occupant qualified for legal aid as earning less than 150% of the federal poverty guidelines. That trigger is a common one applied by legal aid societies. (We did not, however, use this federal poverty guideline filter for the housing data for the COVID-19 analysis because it would have interjected an additional variable – namely poverty – as a possible confounding variable.)

As a result, for purposes of the Project Protect, we did the sequential analysis again but added in a new step, after prior step number three, to filter on poverty. The filtering on poverty applied the Federal Poverty Guidelines as a screen, using the income reported for the household, and the total

[7] Measures of morbidity frequency characterize the number of persons in a population who become ill (incidence) or are ill at a given time (prevalence). So for housing, we describe everything as a measure of prevalence. Below, for COVID-19, we measure everything in terms of incidence, although given the short duration there would not be much difference between incidence and prevalence.

number of people reported in the household. Those two variables were enough to pick the threshold Federal Poverty Level that applied and filter on it. We created parallel variables with a 'p' at the end to designate that they were part of the poverty calculation. We wanted to keep the two sets of variables completely apart.

We noted that actually quite a few people who live above poverty live in substandard housing. The poverty variable reduced by about half of the substandard apartments on average.

The other major task was to do this same work for enough different biannual American Housing Survey data sets over time to have a sufficient number of data points. The survey is done every other year, and while the census includes the top 15 Metropolitan areas each time, the census also includes another perhaps 15 Metropolitan areas on a rotating basis. Therefore, we considered data collected in five different surveys as follows:

1) 2017 National
2) 2017 Metropolitan
3) 2015 Metropolitan
4) 2013 Metropolitan
5) 2013 National

We included the 2013 national survey because we noticed that it had a few additional metropolitan areas that had not been previously surveyed.

When manipulation was completed for each of the years, we needed to:

1) Concatenate the tables to include all years and the associated percentages
2) Join that table, as an inner, to a table of the total apartments by Metropolitan region
3) Multiply the percent substandard by the total number of apartments to get the total substandard apartments by metro area

As already suggested, we repeated this process twice, once for use with the COVID-19 calculation, and once for use in Project Protect with the additional poverty filter.

Challenges Presented By the AHS Data Set

Overall, the AHS data sets were high-quality with no missing values so we did not need to fill in anything. The biggest challenge, as it were, was the repeated changes in the variables over the years, mostly simply changing names but also combining multiple variables into one, and the changes in geography designations over the years that had to be reconciled. We spent a lot of time researching the history of individual variables and setting up a separate excel spreadsheet that tracked the changes in the variable names over time. This was important both so that we could

incorporate the appropriate variables, but also so that we could update the masking table that we used to apply the weights. We also had to make our basic code flexible enough that it could simply incorporate updated tables without having to rewrite the code each time.

Preparing the COVID-19 and Population Data Sets

For the COVID-19 data set, the basic task was calculating the incidence of COVID-19 cases in the same metro areas as were studied in the AHS by using current case counts.

The data we used is updated daily and received through an API that uses current COVID-19 case counts from the New York Times. The first step in preparing the data was to read the data in as a JSON file and then load it into a PostgreSQL database. The data here were not particularly large, but working with a database would prove to be an efficient way to eventually combine all of the other elements of our data, like the population data set and the bridge table, together.

We first created a PostgreSQL table with all of the columns of interest in the data set, which we felt would be "cases", "county", "date", "fips" and "state". Since the COVID-19 data was a list of dictionaries, each item had to be read row for row into the database table, accessing each key, which corresponded to the column name, and inserting each value. One challenge we faced was keeping leading zeros intact for our integer variables, like "fips". In order to preserve the full five-digit code, we inserted the value as a VARCHAR into the database.

Next, we created the population table. This also used data from an API source, which came from census data and needed to be accessed using a unique key (once requested). This data was in the form of a list of lists and the first row had to be popped off before reading into the population table we created in PostgreSQL. The columns we created were "pop", "state" and "county". The last two columns were actually state and county FIPS codes that needed to be combined. The table was created assigning those columns as a VARCHAR type to preserve the five-digit number and not to lose leading zeros. Populating the table in PostgreSQL for this data was more straightforward. Using the command 'executemany,' each line of the list of lists was read into the table. Lastly, we altered the table by concatenating the "state" and "county" columns and calling it "FIPS" to match all of the other columns from other data sets that we would later be joining together.

From here, the COVID-19 and population tables could then be merged together using a natural join.

Preparing the Rental Data Set

The rental data set – namely the total number of occupied apartment units in each metropolitan area – required some visual inspection in excel prior to deciding how to best handle proceeding with the manipulation. Due to the nature of the data set's complicated headers, we decided to read it into Pandas as a CSV. The column names were lengthy and very similar so to identify the columns of interest, Pandas offered the best interface to provide ease for further inspection.

The data set had a multiindex and so we read the CSV into Pandas by accepting the second header index and ignoring the top index. The header that we were left with was the index of column names. Using a regex, we were able to find the particular column of interest which was the estimated occupied rentals. We also used the "id" column to extract the FIPS code contained in the columns' values using the str.replace() method and creating a new column called 'FIPS' in order to match the bridge table for later merging. We were left with only two columns, the fips code and the estimated number of occupied rentals, after saving the columns to a new data frame.

Preparing the Bridge Table and Merging Steps for all Data

The goal of this phase of the analysis was to create a final table including three key columns as follows:

1. Metro area codes
2. Number of substandard apartment units in that metro area
3. Number of COVID-19 cases in that metro area

To build that table, we had to settle upon a standard geographic unit of measure. Of necessity (because it is the only geographic unit released as part of the public use 2017 and 2015 AHS data sets), that unit of measure was the CBSA code. We completed the table by estimating the number of COVID-19 cases per CBSA for the units we gathered from the AHS.

A CBSA is a U.S. geographic area defined by the Office of Management and Budget (OMB) that consists of one or more counties (or equivalents) anchored by an urban center of at least 10,000 people plus adjacent counties that are socioeconomically tied to the urban center by commuting. Areas defined on the basis of these standards applied to Census 2000 data were announced by OMB in June 2003. These standards are used to replace the definitions of metropolitan areas that were defined in 1990. The OMB released new standards based on the 2010 Census on July 15, 2015.

Note that prior to 2015, CBSAs were called Standard Metropolitan Statistical Area (SMSA), so the earlier data will make reference to that name.[8] The term "CBSA" refers collectively to both metropolitan statistical areas (MSA) and micropolitan areas. So MSAs are the more populated regions, and micropolitan are the smaller. We are focusing on the larger, so all of our regions are called MSAs. That's just a subset on CBSAs. For simplicity, we will just use the broader term CBSA, except obviously the actual variables may have different names.

We needed to create a bridge table that was unique to our needs that provided the link between FIPS, SMSA and CBSA codes in order to join all of our data sets. In order to provide useful visualizations for the bubble, scatterplot and choropleth charts, it was also important for us to have the geographic code name translation for metropolitan areas as well as counties. We gathered this data from a few various sources, which provided quality control between the data set CBSA and SMSA definitions of metropolitan areas.[9]
The data was prepared in excel using CBSA, metro area, county and SMSA columns. The data was entered into the spreadsheet to avoid tedium and therefore, required further manipulation once loaded into the notebook. The metro area column contains the names of the metropolitan areas corresponding to a particular CBSA code. The county column contains either single or multiple county names as well as the corresponding FIPS code(s), with each county separated by a semicolon. This data was copied and pasted from the United States Patent and Trademark Office web resource. The SMSA code contained the codes for metro areas included in the 2013 AHS data sets and corresponding to a particular CBSA code which was provided by The National Bureau of Economic Research's crosswalk table (refer to Appendix E, TABLE A).

We chose to use a PySpark data frame to expand our county data by turning the values for each row into a list, then expanding each county out in the column as a new row per its given CBSA code and metro area name. We then extracted the FIPS codes and states from the county column using split and regex methods and saved those to new columns labeled as "FIPS" and "state", respectively. Lastly, this data was saved as a dictionary to be later loaded into a PostgreSQL table called "bridge". The final bridge table contained CBSA, metro area, county, SMSA, FIPS and state columns (refer to Appendix E, TABLE B).

We first merged the COVID-19 and population tables together using a left inner join in PostgreSQL on the common column of FIPS codes. Second, we used PostgreSQL to perform a natural join between the COVID-19 and population table and the bridge table (refer to Appendix F, TABLE C). From PostgreSQL, we read that table into a Pandas DataFrame and merged that with the rental data on the FIPS code as a left join.

Once the rental data was included in the bridge table, we merged the bridge table using a left join (on SMSA codes) with the AHS 2013 metro and the AHS 2013 national data sets after these tables

[8] https://en.wikipedia.org/wiki/Metropolitan_statistical_area

[9] The four sources used to provide quality control were: (a) https://www.census.gov/data-tools/demo/codebook/ahs/ahsdict.html, (2) https://www.uspto.gov/web/offices/ac/ido/oeip/taf/cls_cbsa/cbsa_countyassoc.htm
(3) Census Technical Document and (4) CBSA-FIPS Crosswalk CSV

were concatenated together. The data frame at this point included the following columns: CBSA, metro area, county, SMSA, FIPS, state, county_bridge, population, COVID-19 cases, rental occupied units, control_x, control_y, and percent violative columns (refer to Appendix F, TABLE D). The latter three columns are from the 2013 AHS data. There were many rows with missing data for control_x, control_y and percent violative due to those rows not being contained in the 2013 AHS data, so we had to drop all 'NaN' values.

We concatenated the national and metro 2017 AHS as well as the metro 2015 AHS final data frames and performed a left inner join with the bridge table, dropping 'NaN' values to remove the rows not contained in these data sets. We also dropped the "SMSA" column (which was the geographic variable to join on for the 2013 AHS data frames).

To get the metro area totals for population, cases, and renter occupied units, the 2015 and 2017 AHS post-bridge data frame was grouped by the CBSA variable as well as the control_x and control_y columns, and the 2013 metro and national AHS post-bridge data frame was grouped by SMSA as well as control_x and control_y columns, and the rest of the columns were summed. Lastly, the two data frames, 2015 and 2017 AHS post-bridge and 2013 metro and national AHS post-bridge tables, were concatenated to get 119 rows of data which included the CBSA, the control variables, COVID-19 cases, population totals, renter occupied units and percent violative (refer to Appendix F, TABLE E).

Challenges Presented by Combining All Data Sets

The biggest challenge with using the bridge table was ensuring quality control and making sure we were not capturing a metropolitan area twice with a translation loss between CBSA and SMSA codes. Based on the AHS manipulation, we should have ended up with roughly 183 data points, yet, we were left with 119 unique CBSA codes in our final table and 118 in the final table that included poverty levels. Prior to creating the bridge table, we did a deep investigation on the unique CBSA and SMSA codes for each data set where we used the .unique() method on each data frame's geographic series to get those values. From there we checked the unique codes between each of the AHS data sets that we used by locating the rows in the geographic series where that code did not exist in the other data frame. By using this logic, we determined that there was actually only one usable SMSA code unique to the 2013 metro AHS data frame compared to the 2013 national data frame.

We also checked whether the 2013 SMSA metropolitan locations overlapped with the 2015 and 2017 CBSA locations, using four separate sources to conduct quality control.[8] We started building out the bridge table using the newest data first, which was the 2017 metropolitan and national data and then the 2015 metropolitan data. Once those values were in our table, we used the AHS data codebooks to check a SMSA code's given metropolitan name and then cross referenced that with what we had already included in our Excel bridge table. If that metropolitan name was already contained in the 2015 or 2017 data, that SMSA would not be added to the bridge table (which could otherwise cause problems when merging the table with the AHS data later on). There were

other instances where the SMSA was part of a cluster of SMSA codes making up one CBSA already contained in the bridge table. In those cases, the SMSA code was not included in the bridge table unless any of the counties included in the SMSA were not a part of the corresponding CBSA. Lastly, all CBSA and SMSA codes of "99998", "99999" and "9999" were labeled as data points not in metropolitan areas or had the metropolitan area name withheld. In those cases, those data points were not captured in our final table.

Overall, we tried our best not to capture data from a metropolitan area more than once. The AHS documentation was limited in describing which counties specifically were a part of each SMSA or CBSA. We had to use the USPTO reference to come up with the county make-up of each CBSA area and the NBER's crosswalk table along with the AHS code book's metro name to ensure continuity between the SMSA code, the metro name and the CBSA code.[8]

Finally, our last challenge rested on ensuring that each data point that was meant to be captured was actually included during the bridge table merge. Instead of group merging or merging on a specific column, neither of which methods worked due to issues of having multiple columns in common between each of the AHS data frames, we needed to concatenate like-tables (concatenate 2013 national and metro data frames, and concatenate 2015 metro with 2017 metro and 2017 national data frames), then merge each of the concatenated data frames with the bridge table, then perform a final concatenation between 2013 and 2015-2017 (refer to Appendix F, all TABLES).

The Final Table

As we discussed more below in connection with the final visualizations, after joining the data, we wanted to remove population as a factor since after visually inspecting the results in a bubble chart, population obviously had an overwhelming impact on both the number of substandard apartment units and the number of COVID-19 cases. We wanted the final table to contain columns of substandard apartments per capita and COVID-19 cases per capita.

The consequence of all the manipulation was a table with the following columns:

1. Metro codes
2. Number of substandard apartment units in that metro area
3. Number of COVID-19 cases in that metro area
4. Population per Metro codes
5. COVID-19 cases per capita
6. Estimated substandard homes per capita

Analysis and Visualization

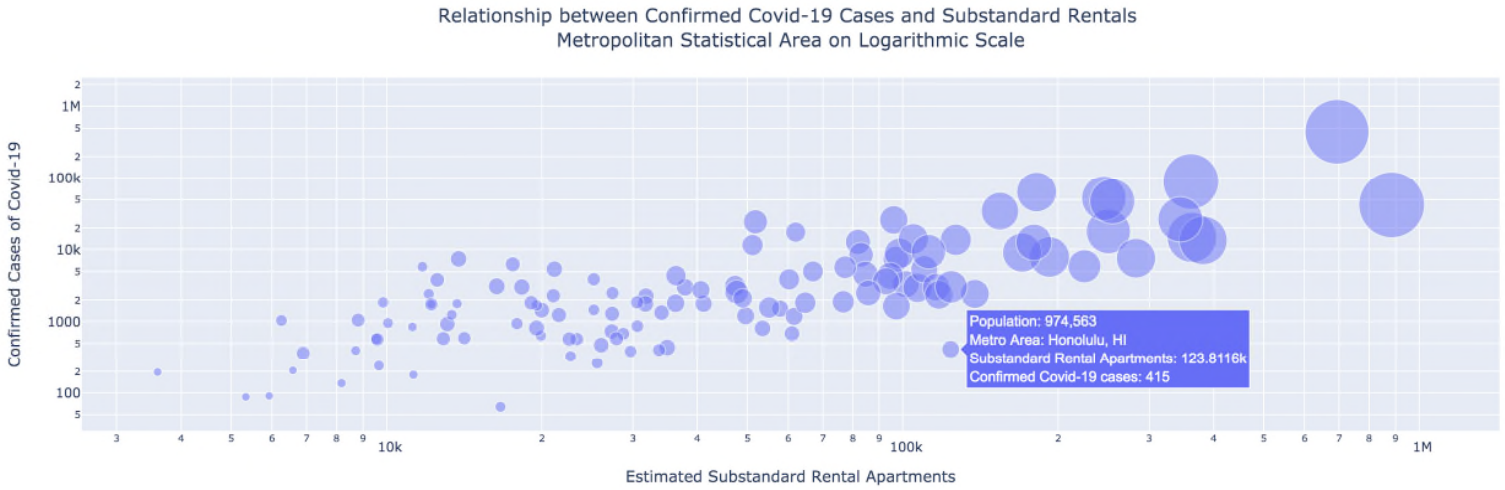
Visualizations

1. Bubble graph with population as the third variable

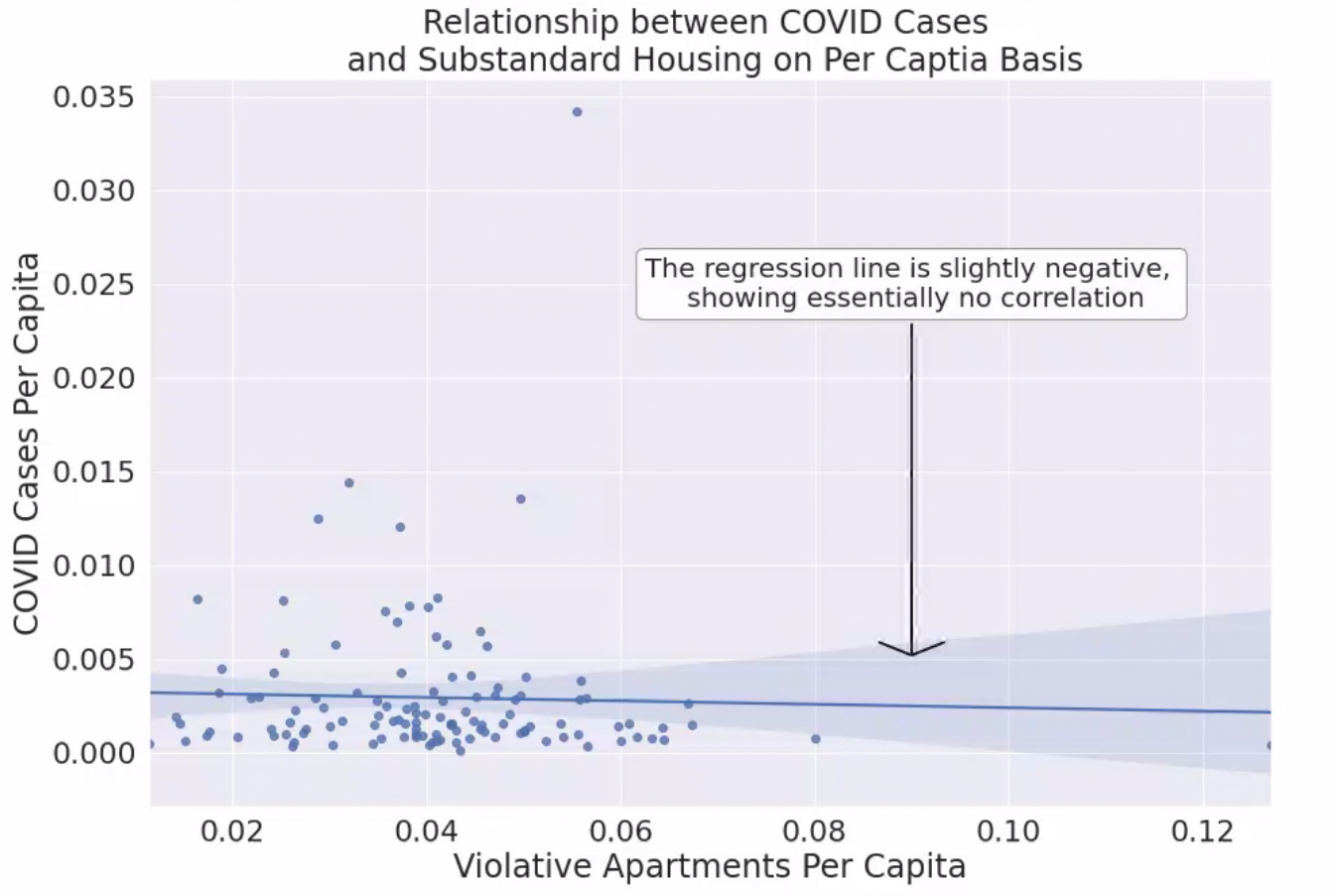
Below we present two columns of our final table in a visualization by creating a bubble chart where the x-axis is the dependent variable of substandard housing and the y-axis representing the number of COVID-19 cases, as the independent variable. The third variable represented by the size of the marks is the population for each metropolitan area, because we suspected that population was highly relevant to both variables. There are roughly 604 data points at the FIPS (i.e. county) level corresponding to the 119 CBSAs (metro areas) on which we can find those data over the years 2013 through 2017. We use the CBSAs for the chart. Since some of the data were clustered and others were outliers, we used a logarithmic scale to better visualize the data. The graph clearly demonstrates a positive relationship between the number of COVID-19 cases and estimated substandard rental apartments where population size is smallest at the lower left corner of the graph and is largest at the upper right corner.

2. Scatter chart created by normalizing both the X and Y axis by dividing them by their respective populations to produce per capita numbers

We realized from the bubble chart that population was a large factor in the correlation between COVID-19 case numbers and the number of substandard rentals, so we created a scatterplot with a regression line, depicting on the x-axis substandard housing per capita and on the y-axis COVID-19 cases per capita. By dividing each of the axes in the bubble chart by population, we took population at least nominally out of the visualization. That left us with a truer depiction of the relationship between substandard housing and COVID-19, which reveals here a slightly negative correlation. Of course we did not remove the impact of population entirely. From a scientific,



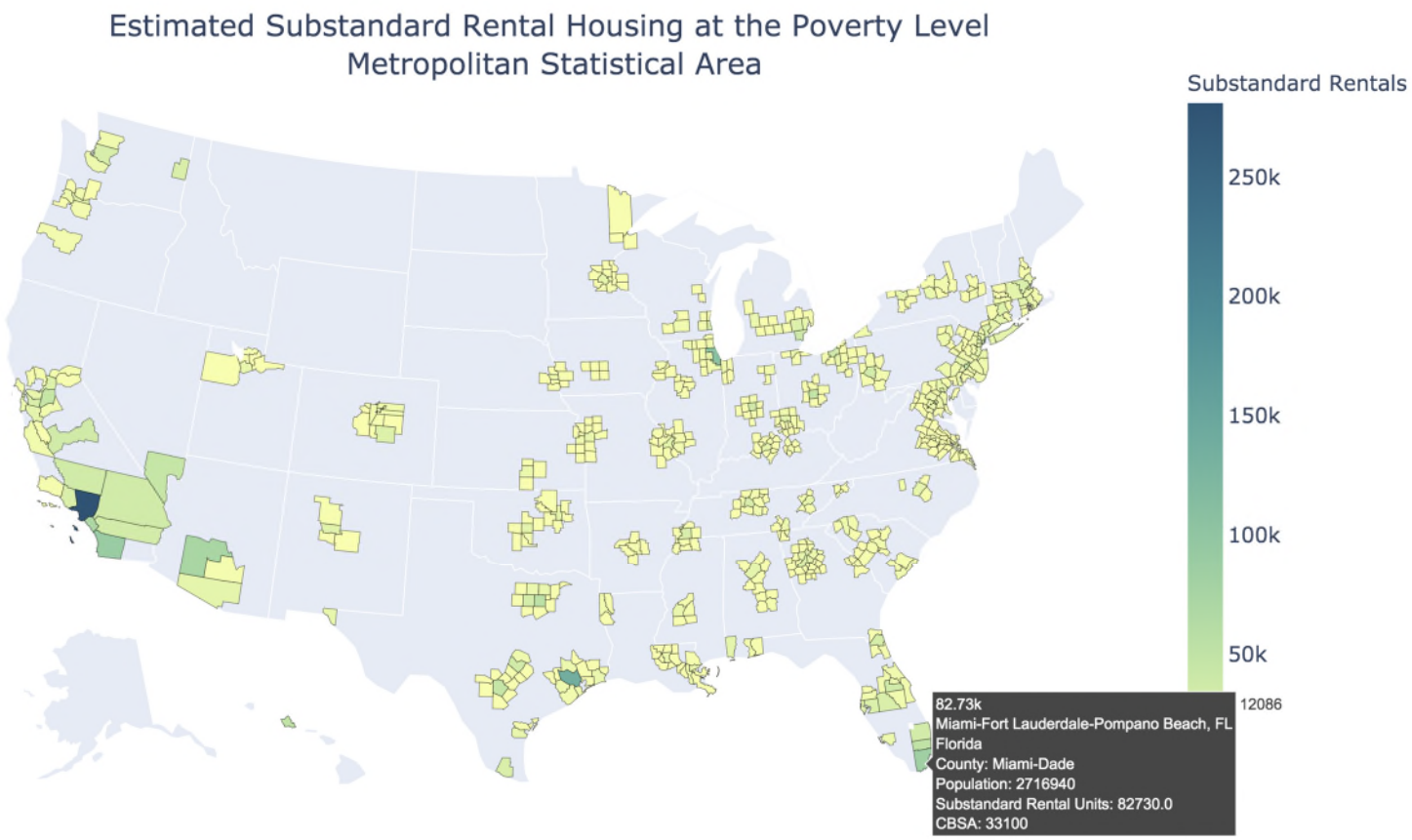
public health standpoint, greater populations are likely to be more densely compact, and that will impact the spread of a communicable disease. That effect is still in the data.



3. Choropleth showing the prevalence of substandard housing

For purposes of linking our substandard housing analyses to the work product of other teams as a part of the legal aid analysis under the name Project Protect, we produced a US choropleth of the total number of US apartments that likely do not meet legal standards rented by people earning less than 150% of the Federal Poverty Guidelines.

For the choropleth, we used the variables associated with the poverty screen since the objective is to see where the legal aid cases are. We used the same 2013 AHS post-bridge table and the raw bridge table merged with the 2015 and 2017 AHS final data frames. The SMSA columns were dropped for both sets of data frames before using a final outer join of the post-bridge 2013 AHS data frames and the post-bridge 2015-2017 data frames. The reason to go back to previous merge points was to capture the categorical data for the data points to display on the map.



Analysis

By being able to combine the data sets on the basis of geography, we tested whether failure to meet housing codes is associated with contracting COVID-19. We are carefully choosing our words to merely look for an association, because we know causation is well beyond our ability to demonstrate using these data. Indeed it's possible that failure to meet housing codes simply indicates the prevalence of low-income people, and there are other aspects of low income that make someone vulnerable to the virus. That is one reason we did not filter for poverty the data that we planned to use to evaluate association with COVID-19. We will not be able to investigate or account for these other associations.

1. Bubble graph with population as the third variable

The bubble graph does purport to show a statistical significance between the prevalence of substandard housing and the incidence of COVID-19.

However, as you can see in the visualization, all of the data points in the upper right-hand corner are large populations and all of the data points in the lower left are smaller populations. Therefore, we are skeptical that this positive trend is purely related to COVID-19 cases and substandard rentals alone. We might be able to intuitively say that the more people you have, the more opportunity you have for substandard housing and for increased presence of COVID-19. Both of those should be greatly influenced by the size of the population. As a result, we decided to create another visualization where we took population out of the equation by dividing both of the axes by population, producing substandard housing per capita and COVID-19 per capita as our new variables.

Separate from the visualization, we calculated the Pearson regression coefficient for number of COVID-19 cases per location versus number of substandard housing units per location. That regression coefficient is 0.5942. We then ran a separate Pearson regression coefficient calculation on number of cases per location per capita, versus number of substandard apartments per location per capita. There the regression coefficient is -0.0345. Those two calculations do suggest that much of the initial correlation was related to population differences among locations. We discuss more the implications of the latter slightly negative coefficient below in connection with the scatter chart.

2. Scatter chart with population factored in by normalizing both the X and Y axis by dividing them by their respective populations

Our suspicions were correct, in that when we took population out of the equation by dividing each data point by its associated population, essentially the correlation coefficient between substandard housing and COVID-19 went to zero. In other words, the data on its face would suggest that there is no correlation between substandard housing and COVID-19.

We are still not satisfied and have doubts about this conclusion as well as from what we know about substandard housing and disease in general. We believe that the data possibly failed to show a correlation simply because of the limitations that we further describe below regarding the data, most notably that we were unable to conduct the apartment analysis on granular data, other than by metropolitan statistical area. We believe the scientific evidence suggests that COVID-19 actually varies significantly from ZIP code to ZIP code, but because the American Housing Survey only discloses metropolitan statistical area data publicly due to privacy concerns, we were unable to do analysis to the level of granularity that we would have desired.

3. Choropleth showing the prevalence of substandard housing

In support of the overall Project Protect, we were able to calculate the number of substandard housing units per metropolitan area occupied by people living below 150% of the federal poverty guidelines. Specifically, we found **there are 4,840,000 households in the 119 metro areas who need an attorney to help them improve their living conditions. The map shows where those households are located.**

But here again, our objective was not entirely met, because we were only able to do analysis in 119 metropolitan areas, and we have no reason to believe that in the rest of the United States there are no other people living in substandard housing units earning less than 150% under the Federal Poverty Guidelines. For example, evidence suggests that college towns are typically filled with substandard housing units,[10] but communities such as Champaign Urbana, Illinois, Lincoln, Nebraska, and West Lafayette, Indiana would not be included in this analysis. We have no data on the prevalence of substandard apartments in smaller town and rural areas. The analysis thus excludes 34.65 % of the U.S. population. So our analysis undoubtedly significantly under reports the number of substandard housing legal matters in need of legal aid representation.

Limitations of Analysis and Uncertainty

We provide a more comprehensive list of the limitations of this analysis as follows:

- ✓ The timing of our data is not in sync. The housing data is from 2013 through 2017, and the COVID-19 data is from 2020.
- ✓ Our analysis is not as granular as we would like it. The housing data in the later years was restricted to aggregates for the entire metropolitan areas. We could not break it down into finer geographic units. We could have used more granular housing data through the American Community Survey, but that data was limited with roughly only 4 variables to assess a home's livability and would be less accurate for assessing access to legal aid. The COVID-19 data, on the other hand, is available in fine geographic units. Restricting ourselves to the larger units inhibited our ability to truly assess correlation.
- ✓ The weights assigned are the opinions of two experts. Mr. Haller is an expert with much very relevant experience, but there may well be a diversity of views. The AHS questions were not drafted with compliance with legal building codes in mind, so in many cases the questions were written imprecisely for our purposes. We thus had to apply subjective judgment in several cases.
    - We would note that our analysis is not overly sensitive to the weights, in that apartments ended up counting as substandard as long as there were a total of three points for a given apartment unit. So the main issue is whether we accurately identified the lower end of the weight scale.
- ✓ As a matter of quality control, we closely looked at the impact of the weights on the findings of individual surveys, and found that no individual variable accounted for more than 10% of the violations, and the vast majority were well below that. Consequently, if there is debate about a single variable, a single variable would not significantly impact the findings.
- ✓ The housing analysis is significantly limited to only the 119 metropolitan areas included in the last three years of AHS surveys. This leaves out not only the entire rural areas of

[10] Daniel E. Wenner, Note, *Renting in Collegetown*, 84 Cornell L. Rev. 543, 544 (1999) : https://scholarship.law.cornell.edu/cgi/viewcontent.cgi?article=2752&context=clr

the United States, but it also leaves out many significant but slightly smaller cities and college towns across the country. The analysis excludes 34.65 % of the US population.

- ✓ There is limited information on the AHS website as to how they defined each SMSA and CBSA (although, CBSA definitions are more clearly defined elsewhere) as to which counties were included in those areas for those survey years. It would take much more extensive research to accurately uncover each SMSA and CBSA definition as it was used in the AHS.
- ✓ In our final graph on housing versus COVID-19, we included specifically the area of uncertainty based on our data.

Statement of Work

Stacey and Brad jointly developed the proposal and the overall strategy. Stacey researched and found the data (COVID-19 prevalence, population, apartment prevalence and housing survey data), did the COVID-19 data manipulation, learned geographic codes, and created the table for joining the COVID-19 data with the housing data. Brad worked on filtering the housing data, and worked with the outside legal expert to develop the weights for the variables as well as the overall project analyses. Each person then checked the others' work. They jointly worked on the visualizations and the final report.

Appendix A

Variable Weights Assigned To the American Housing Survey Data

	Variable	Trigger	Weight
0	SUPP1HEAT	'01'	10
1	SUPP1HEAT	'05'	10
2	SUPP1HEAT	'10'	10
3	SUPP1HEAT	'11'	10
4	SUPP2HEAT	'01'	10
5	SUPP2HEAT	'05'	10
6	SUPP2HEAT	'10'	10
7	SUPP2HEAT	'10'	10
8	SUPP2HEAT	'11'	10
9	COOKTYPE	'2'	1
10	COOKTYPE	'3'	1
11	COOKTYPE	'4'	6
12	FRIDGE	'2'	7
13	KITCHSINK	'2'	4
14	NOWIRE	'2'	3
15	NOWIRE	'3'	9
16	FUSEBLOW	'1'	1
17	FUSEBLOW	'2'	1
18	FUSEBLOW	'3'	2
19	FUSEBLOW	'4'	3
20	PLUGS	'2'	1
21	ADEQUACY	'3'	5
22	HEATTYPE	'07'	8
23	HEATTYPE	'08'	8
24	HEATTYPE	'09'	8
25	HEATTYPE	'10'	8
26	HEATTYPE	'11'	8
27	HEATTYPE	'13'	8
28	HEATTYPE	'14'	8
29	COLDHTCAP	'1'	3
30	COLDINSUL	'1'	5
31	BATHROOMS	'07'	8
32	BATHROOMS	'08'	7
33	BATHROOMS	'09'	3
34	BATHROOMS	'10'	9
35	BATHROOMS	'11'	9
36	BATHROOMS	'12'	8
37	BATHROOMS	'13'	10
38	BATHEXCLU	'2'	10
39	MOLDKITCH	'1'	2
40	MOLDBATH	'1'	4
41	MOLDBEDRM	'1'	5
42	MOLDLROOM	'1'	5
43	MOLDBASEM	'1'	3
44	MOLDOTHER	'1'	5
45	RODENT	'1'	7
46	RODENT	'2'	4
47	RODENT	'3'	2
48	ROACH	'1'	7
49	ROACH	'2'	4
50	ROACH	'3'	2
51	ROACH	'4'	1
52	SEWBREAK	'3'	5
53	SEWBREAK	'4'	5
54	FNDCRUMB	'1'	3
55	ROOFHOLE	'1'	7
56	ROOFSHIN	'1'	3
57	ROOFSAG	'1'	3
58	WALLSIDE	'1'	2
59	WALLSLOPE	'1'	3
60	WINBOARD	'1'	1
61	WINBROKE	'1'	6
62	FLOORHOLE	'1'	6
63	WALLCRACK	'1'	3
64	PAINTPEEL	'1'	2
65	HOTWATER	'7'	10
66	COLDEQFREQ	1	3
67	COLDEQFREQ	2	3
68	COLDEQFREQ	3	9
69	COLDEQFREQ	4	9
70	COLDEQFREQ	5	9
71	COLDEQFREQ	6	9
72	COLDEQFREQ	7	9
73	COLDEQFREQ	8	9
74	NOTOILFREQ	1	3
75	NOTOILFREQ	2	3
76	NOTOILFREQ	3	6
77	NOTOILFREQ	5	6
78	NOTOILFREQ	5	6
79	NOTOILFREQ	6	6
80	NOTOILFREQ	7	9
81	NOTOILFREQ	8	9
82	NOWATFREQ	'1'	3
83	NOWATFREQ	1	3
84	NOWATFREQ	'2'	3
85	NOWATFREQ	'3'	8
86	NOWATFREQ	'4'	8
87	NOWATFREQ	'5'	8
88	NOWATFREQ	'6'	8
89	NOWATFREQ	'7'	9
90	NOWATFREQ	'8'	8

Appendix B

2017 Federal Poverty Guidelines

The income level is dictated by those guidelines based on the number of people in a home. We then added a column called cut off that represented 150% of that level, which corresponds to the level typically applied by legal aid groups to determine eligibility for services.

	People_in_House	Income	Cutoff
0	1	12060	18090.0
1	2	16240	24360.0
2	3	20420	30630.0
3	4	24600	36900.0
4	5	28780	43170.0
5	6	32960	49440.0
6	7	37140	55710.0
7	8	41320	61980.0

Appendix C
AHS Variables of Interest

The first group of variables are the condition variables broken out by year of survey. These are the variables that, according to our legal experts, directly relate to whether a given apartment complies with legal standards.

The second group of variables relate to the geographic location of the apartments, again broken out by year of survey.

The final group of variables related to demographics, including number of people and income level necessary to determine whether a given household was under the federal poverty guideline threshold, and whether the unit was a rental unit.

- cond_var_11 = ['EXPOSE', 'ELEV', 'KITCHEN', 'BSINK', 'TOILET', 'TUB', 'HEQUIP', 'REFR', 'SINK', 'BURNER', 'COOK', 'NOWIRE', 'PLUGS', 'NUMBLOW', 'NUMCOLD', 'WHYCD2', 'WHYCD3', 'NUMTLT', 'NUMDRY', 'HOTPIP', 'NUMSEW', 'PLUMB', 'ECRUMB', 'EHOLER', 'EMISSR', 'ESAGR', 'EMISSW', 'ESLOPW', 'EBOARD', 'EBROKE', 'HOLES', 'CRACKS', 'ZADEQ', 'WATERS', 'RATFREQ', 'ROACHFRQ', 'MOLD', 'ASTHEMR', 'MUST', 'SECSMK', 'STAIRRL', 'STAIRMIS', 'STAIRBRK', 'SMOKE']

- cond_var_13 = ['EXPOSE', 'ELEV', 'KITCHEN', 'BSINK', 'TOILET', 'TUB', 'HEQUIP', 'REFR', 'SINK', 'BURNER', 'COOK', 'NOWIRE', 'PLUGS', 'NUMBLOW', 'NUMCOLD', 'WHYCD2', 'WHYCD3', 'NUMTLT', 'NUMDRY', 'HOTPIP', 'NUMSEW', 'PLUMB', 'ECRUMB', 'EHOLER', 'EMISSR', 'ESAGR', 'EMISSW', 'ESLOPW', 'EBOARD', 'EBROKE', 'HOLES', 'CRACKS', 'ZADEQ', 'WATERS', 'RATFREQ', 'ROACHFRQ']

- cond_var_15 = ['COOKTYPE', 'FRIDGE', 'KITCHSINK', 'ASTHEMR', 'MUST', 'SECSMK', 'NOWIRE', 'FUSEBLOW', 'PLUGS', 'ADEQUACY', 'HEATTYPE', 'COLDEQFREQ', 'COLDHTCAP', 'COLDINSUL', 'BATHROOMS', 'BATHEXCLU', 'MOLDKITCH', 'MOLDBATH', 'MOLDBEDRM', 'MOLDLROOM', 'MOLDBASEM', 'MOLDOTHER', 'RODENT', 'ROACH', 'NOTOILFREQ', 'NOWATFREQ', 'SEWBREAK', 'STAIRRL', 'STAIRMIS', 'STAIRBRK', 'FNDCRUMB', 'ROOFHOLE', 'ROOFSHIN', 'ROOFSAG', 'WALLSIDE', 'WALLSLOPE', 'WINBOARD', 'WINBROKE', 'FLOORHOLE', 'WALLCRACK', 'PAINTPEEL', 'HOTWATER', 'WATSAFE']

- cond_var_17 = ['SUPP1HEAT', 'SUPP2HEAT', 'COOKTYPE', 'FRIDGE', 'KITCHSINK', 'NOWIRE', 'FUSEBLOW', 'PLUGS', 'ADEQUACY', 'HEATTYPE', 'COLDEQFREQ', 'COLDHTCAP', 'COLDINSUL', 'BATHROOMS', 'BATHEXCLU', 'MOLDKITCH', 'MOLDBATH', 'MOLDBEDRM', 'MOLDLROOM', 'MOLDBASEM', 'MOLDOTHER', 'RODENT', 'ROACH', 'NOTOILFREQ', 'NOWATFREQ', 'SEWBREAK', 'FNDCRUMB', 'ROOFHOLE', 'ROOFSHIN', 'ROOFSAG', 'WALLSIDE', 'WALLSLOPE', 'WINBOARD', 'WINBROKE', 'FLOORHOLE', 'WALLCRACK', 'PAINTPEEL', 'HOTWATER']

- location_var_11 = ['SMSA', 'COUNTY', 'STATE']
- location_var_13 = ['SMSA']
- location_var_15 = ["OMB13CBSA"]
- location_var_17 = ["OMB13CBSA"]
- rent_var_1715 = ['CONTROL', 'TENURE', 'HINCP', 'FINCP', 'NUMPEOPLE']
- rent_var_1311 = ['CONTROL', 'TENURE', 'ZINC2', 'ZADULT', 'KIDU18']

Appendix D

Criteria Used In Developing Survey Weights

Here is the full text of the instruction given to Mr. Haller, asking him to create the weights:

"Specifically, I'm asking if you would pick a number between 1 and 10 that corresponds to the likelihood that you as a legal aid housing professional would take the case based on that answer, assuming that you have room on your docket to do so. So in that sense, it's not comparative or relative to other cases that are vying for your time. It's a more absolute scale. Conceptually I understand that an attorney would never simply take a case or not based on one multiple-choice question. I'm
asking you to step back and assess the importance of the topic to you as a legal aid professional in deciding which cases you take.

More specifically, conceptually please think of it as:

10 means based on that factor you would be 100% likely to take the case from a significance standpoint (you would always have to explore the facts more deeply: we are just conceptually asking how important the factor is in your determination to take a case)

5 means that you would be 50% likely to take the case

1 means that you would be 10% likely to take the case

0, which is not one of your options, means that you would not take the case and those are the numbers that I'm applying to all the other answers not in my suggested list for you.

And obviously you can choose any number from 1 to 10, I just wanted to show you the pattern."

Appendix E

Background on Creating the Bridge Data Frame

TABLE A) The following table is an image of the bridge data frame loaded into PySpark before manipulation.

OMB13CBSA	metro_area	county_bridge	SMSA
36420	Oklahoma City, OK	Canadian County, OK (4	null
13820	Birmingham-Hoover, AL	Bibb County, AL (01007	null
40060	Richmond, VA	Amelia County, VA (510 Petersburg city, VA (51730); Richmond city, VA (51760)	null
45300	Tampa-St. Petersburg-Clearwater, FL	Hernando County, FL (1	null
29820	Las Vegas-Paradise, NV	Clark County, NV (3200	null
12580	Baltimore-Towson, MD	Anne Arundel County, M	null
33460	Minneapolis-St. Paul-Bloomington, MN-WI	Anoka County, MN (2700	null
40380	Rochester, NY	Livingston County, NY	null
41700	San Antonio-New Braunfels, TX	Atascosa County, TX (4	null
41940	San Jose-Sunnyvale-Santa Clara, CA	San Benito County, CA	null

TABLE B) Below is the final bridge data frame before reading it as a dictionary and inserting it into a PostgreSQL table. This shows the table with the county_bridge column expanded and the fips and state values added from the county_bridge into their own respective columns.

OMB13CBSA	metro_area	county_bridge	SMSA	fips	state
36420	Oklahoma City, OK	Canadian County	null	40017	OK
36420	Oklahoma City, OK	Cleveland County	null	40027	OK
36420	Oklahoma City, OK	Grady County	null	40051	OK
36420	Oklahoma City, OK	Lincoln County	null	40081	OK
36420	Oklahoma City, OK	Logan County	null	40083	OK
36420	Oklahoma City, OK	McClain County	null	40087	OK
36420	Oklahoma City, OK	Oklahoma County	null	40109	OK
13820	Birmingham-Hoover...	Bibb County	null	01007	AL
13820	Birmingham-Hoover...	Blount County	null	01009	AL
13820	Birmingham-Hoover...	Chilton County	null	01021	AL

Appendix F

Background on Merging the Bridge Data Frame

TABLE C) Below is the bridge table merged with the COVID-19 and population data. We kept two county columns to provide one more layer of quality control and decided to drop the state column from the original bridge table while keeping the state column from the COVID-19 table.

	fips	cases	county	date	state	pop	omb13cbsa	metro_area	county_bridge	smsa
0	01001	74	Autauga	2020-05-10	Alabama	55869	33860	Montgomery, AL	Autauga County	5240
1	01007	46	Bibb	2020-05-10	Alabama	22394	13820	Birmingham-Hoover, AL	Bibb County	None
2	01009	44	Blount	2020-05-10	Alabama	57826	13820	Birmingham-Hoover, AL	Blount County	None
3	01021	65	Chilton	2020-05-10	Alabama	44428	13820	Birmingham-Hoover, AL	Chilton County	None
4	01051	150	Elmore	2020-05-10	Alabama	81209	33860	Montgomery, AL	Elmore County	5240

TABLE D) This next table is what resulted from merging the bridge with the rental data and then with both 2013 AHS data sets then dropping 'NaN' values.

	fips	cases	county	date	state	pop	OMB13CBSA	metro_area	county_bridge	SMSA	renter_occupied_units	CONTROL_x	CONTROL_y	Percent Violative
154	18019	382	Clark	2020-05-10	Indiana	118302	31140	Louisville, KY-IN	Clark County	4520	13082	1188.0	331.0	0.27862
156	18043	233	Floyd	2020-05-10	Indiana	78522	31140	Louisville, KY-IN	Floyd County	4520	8091	1188.0	331.0	0.27862
161	18061	158	Harrison	2020-05-10	Indiana	40515	31140	Louisville, KY-IN	Harrison County	4520	2547	1188.0	331.0	0.27862
176	18175	48	Washington	2020-05-10	Indiana	28036	31140	Louisville, KY-IN	Washington County	4520	2556	1188.0	331.0	0.27862
201	21029	82	Bullitt	2020-05-10	Kentucky	81676	31140	Louisville, KY-IN	Bullitt County	4520	5691	1188.0	331.0	0.27862

TABLE E) This last table demonstrates what the final table looked like after the bridge table was merged with the 2015 and 2017 AHS data tables.

	OMB13CBSA	CONTROL_x	CONTROL_y	cases	pop	renter_occupied_units	Percent Violative
0	12060	781.0	257.0	16650	6001088	760053	0.329065
1	12580	622.0	186.0	12131	2800053	350557	0.299035
2	13820	503.0	181.0	1831	1153956	136075	0.359841
3	14460	764.0	199.0	58725	4873019	697398	0.260471
4	16980	695.0	202.0	74368	9458539	1243348	0.290647

*TABLE D was reused again as the 2013 data that would be combined with the 2015 and 2017 AHS data to create the data frame that would make the choropleth map and bubble chart.