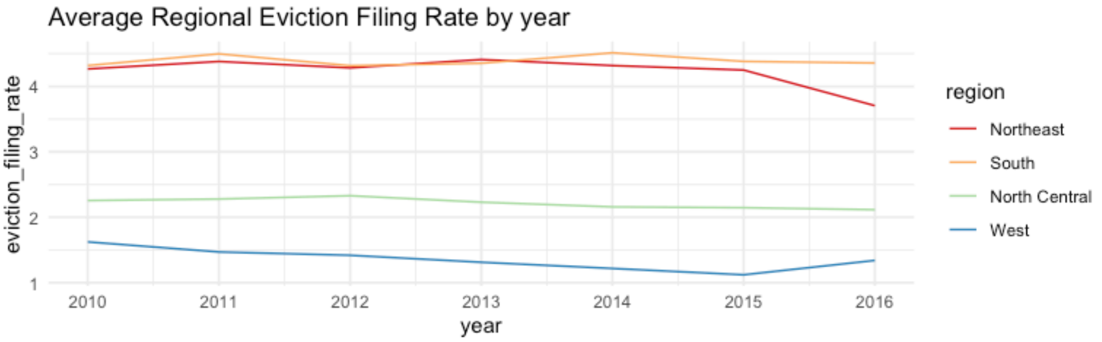
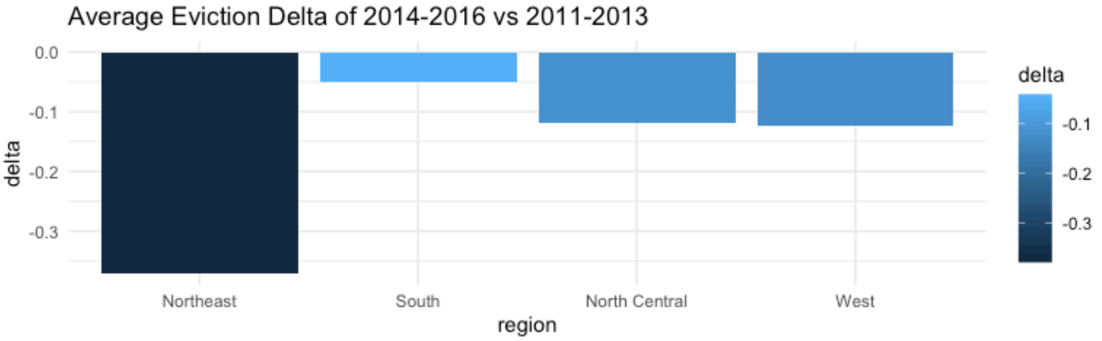
Eviction Lab Data Exploration

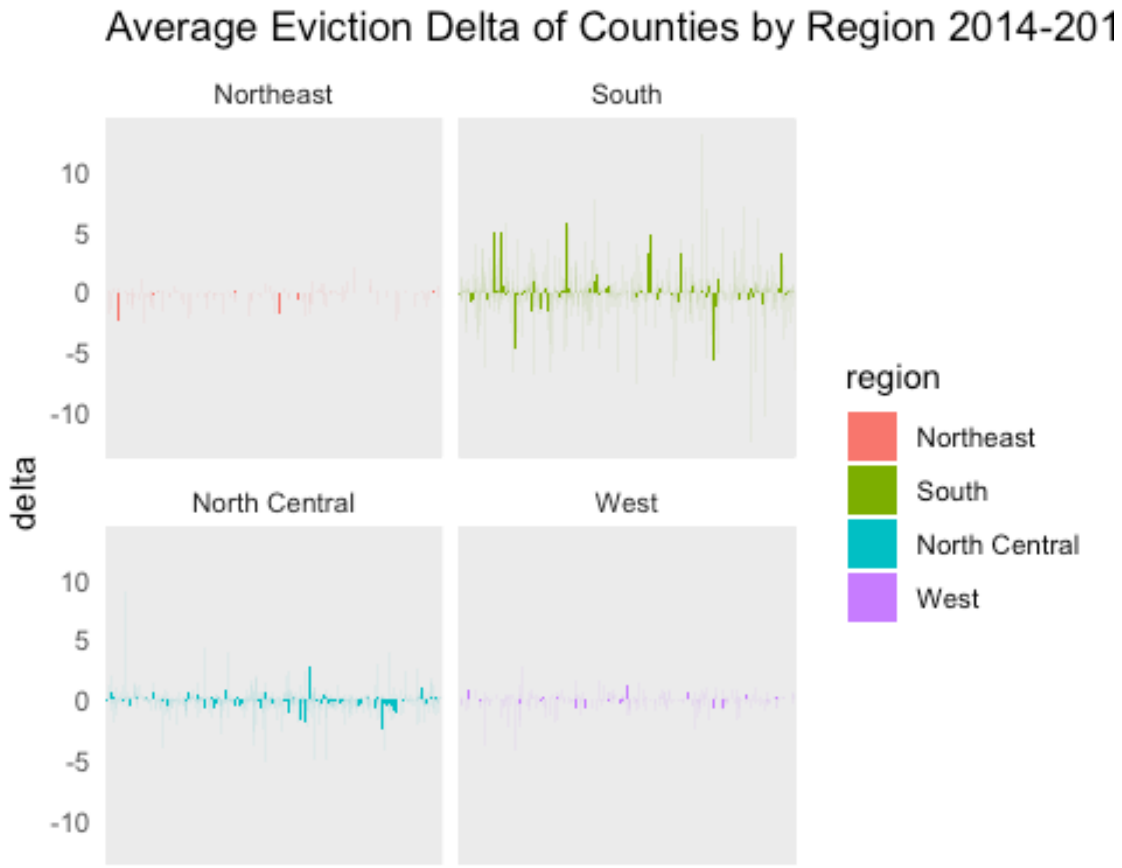
Allison Shafer, Monica Puerto, Allison Ragan

12/1/2019

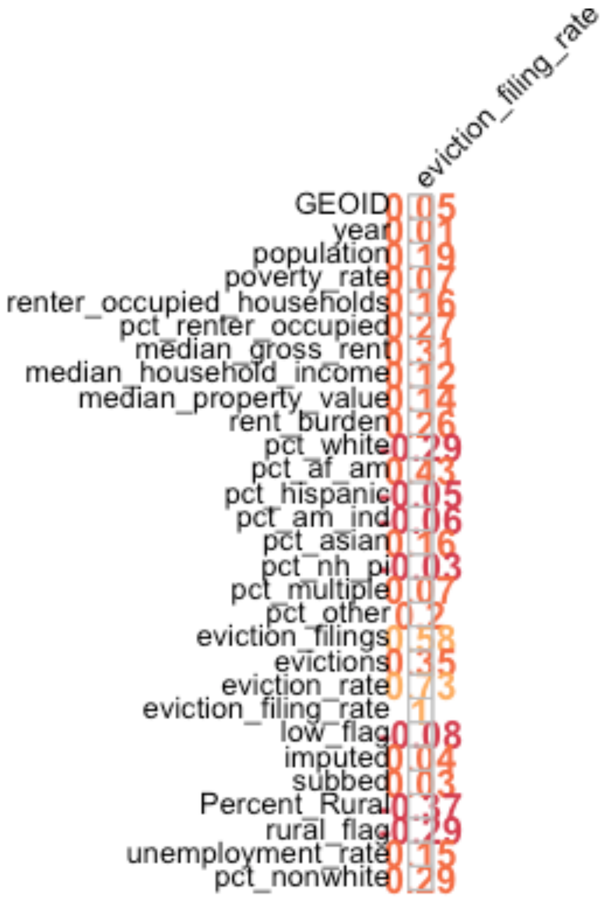
For our group project, we analyzed data from The Eviction Lab to answer our posed questions regarding which attributes were most associated with eviction filing rates and whom is most affected by evictions. Eviction filing rates varied greatly by region with the Northeast and the South regions experiencing one to two times higher eviction filing rates compared to the West and North Central regions. Comparing the eviction filing rate averages, during the last few years of the dataset ending in 2016, the Northeast region has seen the steepest decline starting in 2013 compared to the rest of the regions, whereas the South has remained flat, and in the North Central and West regions have declined since 2012, but the West saw a slight increase in 2016. Drilling into the change by counties, the South as a region remained flat because there are many counties who experience an increase being offset by many counties in the South decreasing.



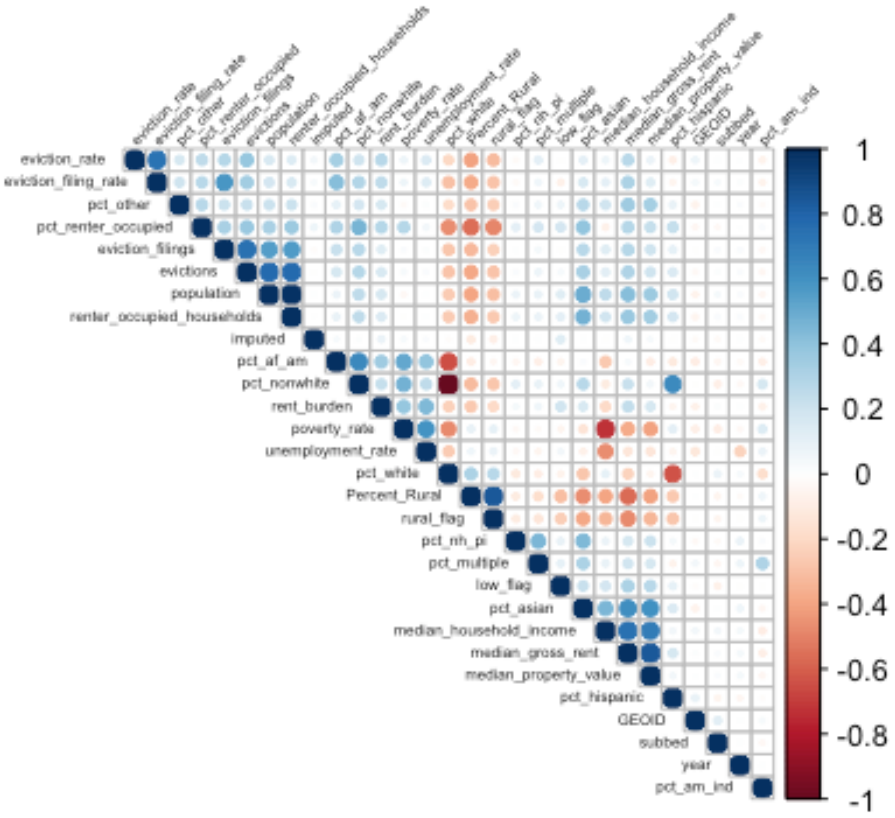




Additionally, the attributes that had the highest positive correlations with eviction filing rates were the percent of African American population and median gross rent. Eviction filing rates were also positively correlated with rent burden, renting costs, and percent of renter occupied. Eviction filing rates were negatively correlated with the percent of white population and the percent rural area. Areas that were more rural had less diversity and lower renter occupancy rates therefore saw less filing eviction rates. Areas that had more ethnic diversity, were more populated, and had higher renter occupancy percentages with higher rental costs saw higher eviction filing rates.

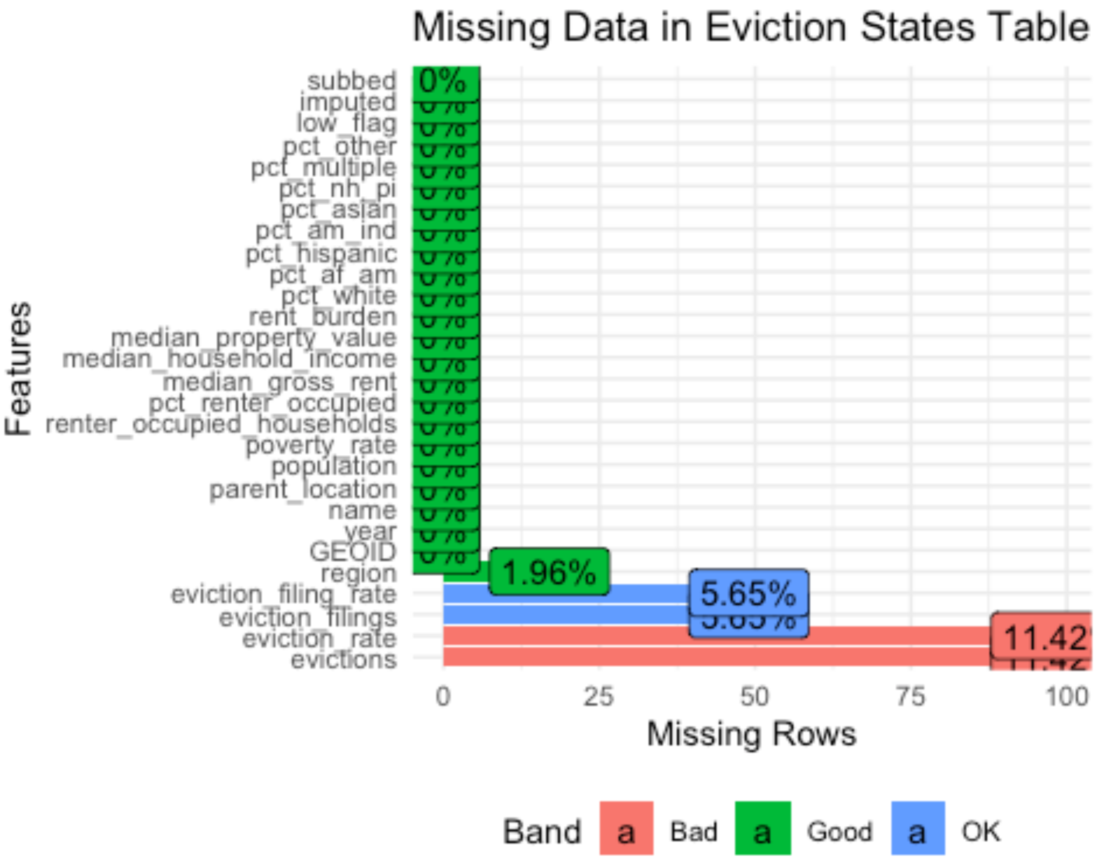


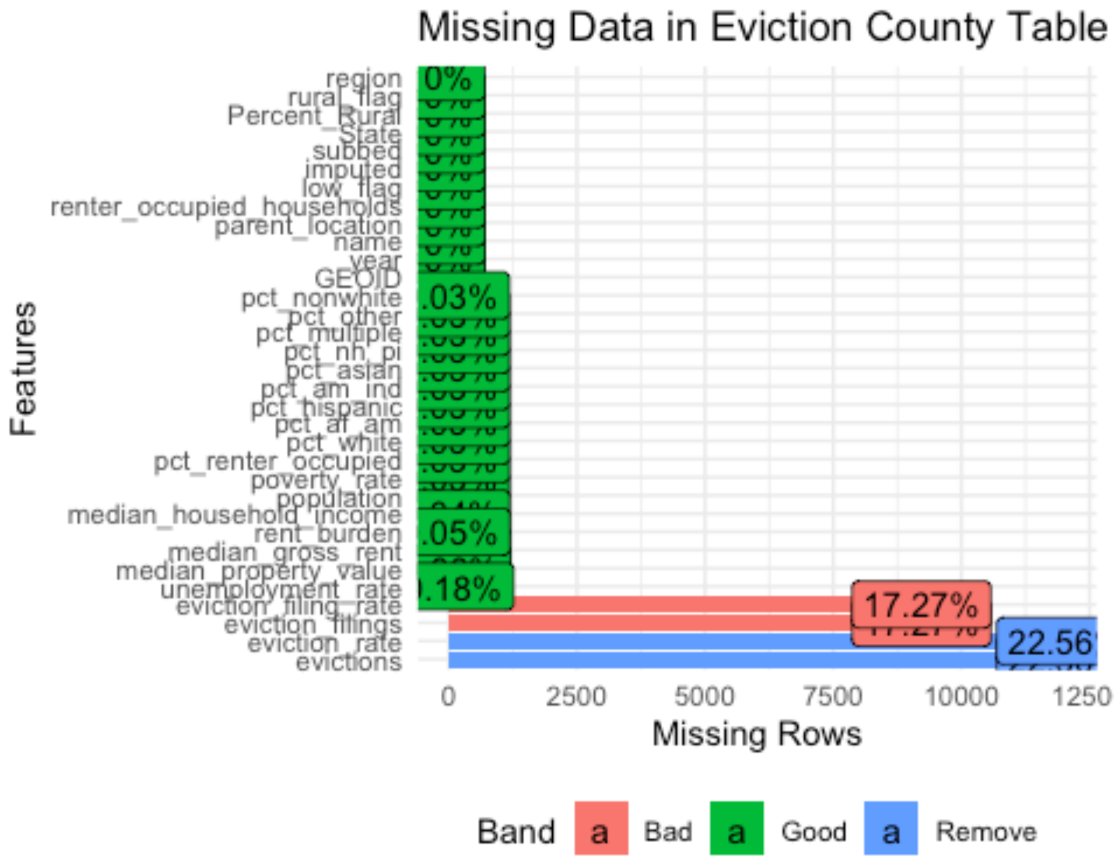
eviction_filing_rate
GEOID 0.05
year 0.01
population 0.19
poverty_rate 0.07
renter_occupied_households 0.16
pct_renter_occupied 0.27
median_gross_rent 0.31
median_household_income 0.12
median_property_value 0.14
rent_burden 0.26
pct_white -0.29
pct_af_am 0.43
pct_hispanic -0.05
pct_am_ind -0.06
pct_asian 0.16
pct_nh_pi -0.03
pct_multiple 0.07
pct_other 0.2
eviction_filings 0.58
evictions 0.35
eviction_rate 0.73
eviction_filing_rate 1
low_flag -0.08
imputed 0.04
subbed 0.03
Percent_Rural -0.37
rural_flag -0.29
unemployment_rate 0.15
pct_nonwhite 0.29



The Eviction Lab dataset came from a department at Princeton University called The Eviction Lab. They collected, geocoded, aggregated, cleaned, and publicized a massive dataset consisting of 82.9 million court records that occurred in the United States during 2000-2016. The Freedom of Information Act (FOIA), allows any citizen to request access to records from any federal agency. Through bulk data requests from 13 states, the Eviction Lab was able to acquire 12.8 million individual court-ordered eviction records. The rest, around 66 million records were purchased from the services of LexisNexis and about 11 million from American Information Research Services Inc (AIRS). The Eviction Lab then treated each observation at the household level versus the individual level, so if there were two people in the case record, it would be counted as one case record per household versus one case record per person in the eviction case record. Also, commercial evictions were excluded from these court records and focused solely on cases that occurred in addresses with rental leases. In the end, the Eviction Lab team ended with around 38 million unique case records. The Eviction Lab then supplemented the eviction data with Census Bureau data by joining the datasets by zip code. The Census Bureau data included were variables like population, Geographic ID (GEOID), demographic percentages, income levels, rental variables (percent renter-occupied, median gross rent), etc.

The Eviction Lab dataset we started our analysis with included 5,501,283 observations with 27 variables, and included data for the years 2000-2016. The data included information for all levels of governments including state, county, incorporated place, and minor civil divisions, as well as for Census tracts and block groups. The dataset was mostly ‘tidy’, but we ended up adding more variables from different datasets that needed tidying to be joined and we then segmented further for analysis. After our initial look at the data, we decided as a group that we wanted to focus our analysis on the state level and county level. We based this decision on a few observations from our initial exploratory data analysis. First, the more granular the data were the more null values we encountered; at the county level, our original variable of interest, eviction rate, had around 23% null values. If the data was anymore granular, we would encounter higher percentages of null values in our variables of interest for areas when looking at lower counties levels. For the purposes of analyzing data correlations and or using the data for modeling, we did not want a lack of data to be an issue.


Missing Data in Eviction States Table
Features
subbed
imputed
low_flag
pct_other
pct_multiple
pct_nh_pi
pct_asian
pct_am_ind
pct_hispanic
pct_af_am
pct_white
rent_burden
median_property_value
median_household_income
median_gross_rent
pct_renter_occupied
renter_occupied_households
poverty_rate
population
parent_location
name
year
GEOID
region
eviction_filing_rate
eviction_filings
eviction_rate
evictions
0%
1.96%
5.65%
11.42
0
25
50
75
100
Missing Rows
Band
a
Bad
a
Good
a
OK



Second, we decided we were interested in incorporating additional data that was not included in our dataset, such as unemployment rates, regional information, and an indicator of whether a county was considered rural, to determine if these variables impacted the eviction related data. During our research for additional data, we noticed that most of the datasets we were searching for were most frequently tabulated at the county level. Additionally, most of the resources included the same primary key representing the geographic code for the counties, the GEOID, that was present in the Eviction Lab dataset.

To obtain the unemployment rate data, we used webscraping methods and built functions to download and aggregate unemployment rate data which was organized in multiple Excel workbooks by year, from the United State Bureau of Labor Statistics website. In order to join the unemployment rate data to our dataset, we had to create the primary key of GEOID for the unemployment rate data. We did this by concatenating the state and county FIPS codes. For the rural indicator, we obtained data that featured the percent rural area from the Census Bureau which was tabulated based off of 2010 Census data. Using the provided percent rural field, we created a field to indicate whether we would consider the counties as rural or not. We used a threshold of 50% or more for the percent rural field to classify the counties as rural. We then joined this data to our dataset using the primary key, GEOID. Lastly, we used a built in R dataset called United States Figures and Facts containing region names (Northeast, North Central, South, and West) to identify the regions for analysis, which we joined into our dataset by state name.

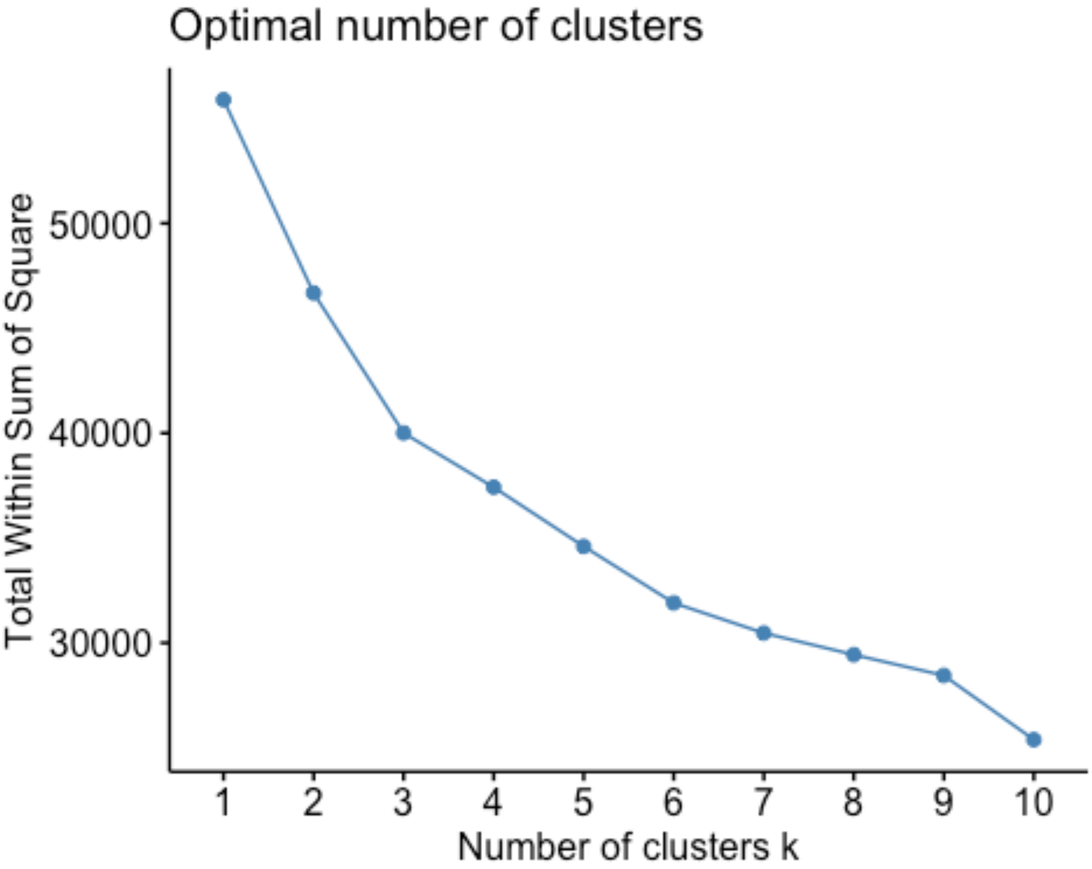
We also decided that since we were analyzing data for the various levels of governments in the country, we wanted to visualize our dataset using maps. In order to create the maps, we needed to acquire geospatial data or shapefiles. We used a package in R called Tigris to obtain the shapefiles for the states, however the Tigris option for using the county level data was not working properly so we had to seek alternative resources. For the county level spatial data, we downloaded shapefiles from the Census Bureau. We then had to join the spatial datasets to our datasets to be able to display the data using a choropleth map. We completed additional spatial data manipulations to get the data in a SpatialPolygonDataframe dataframe to utilize the data with the Leaflet package for our Shiny application.

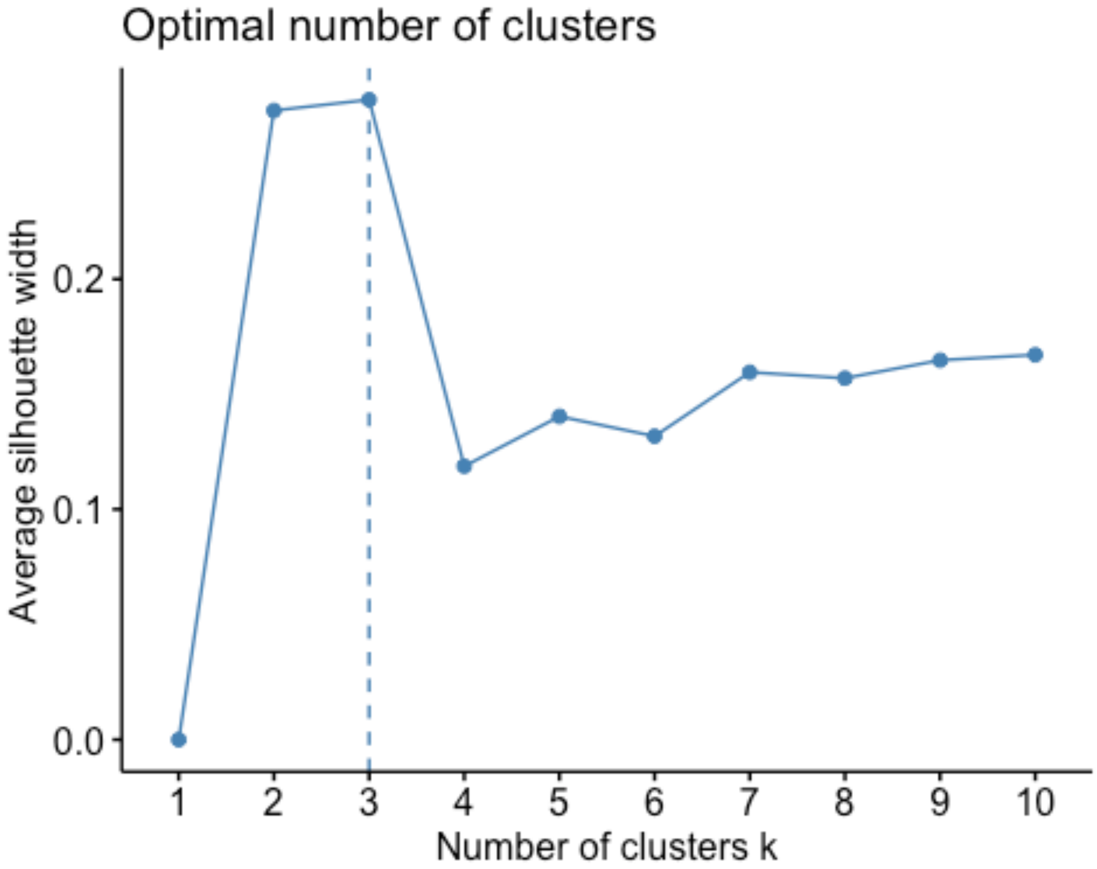
In addition to joining outside datasets to our dataset, we added a handful of extra columns using the existing data. We created a column called 'percent non-white' which subtracted one minus the percentage of white population. We also created a column that assigned a cluster to a county using a K-means model.

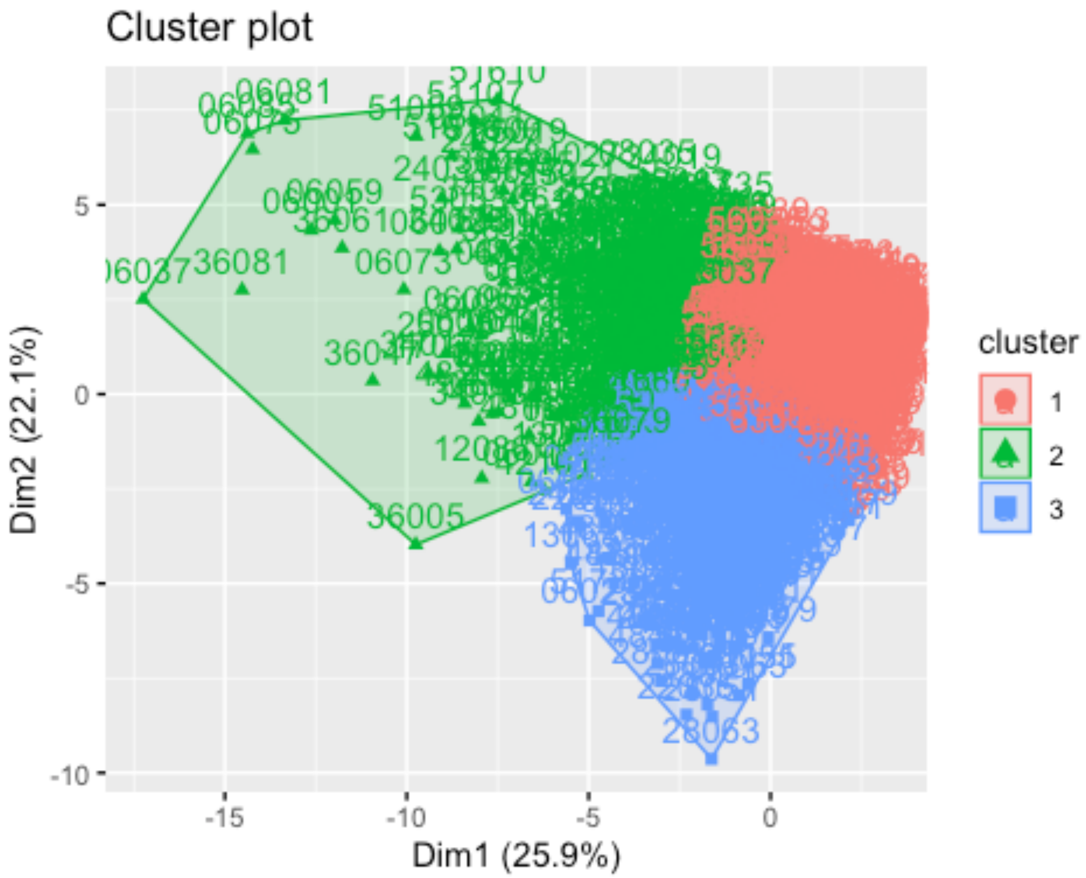
Another main decision we made early on was to split the dataset into two, one representing the state level data and one representing the county level data. We planned on using the state level data for higher level analysis as well as summary and descriptive statistics and wanted to use the county level data for a K-means model. Our goal was to see if the counties that were in similar clusters also share similar eviction filing rates, or if there were any unexplained relationships not captured by our data.

For the K-means model, we averaged the data over the years by county so that we were left with one unique row for each county. We then scaled the data to make the variables comparable across the same scale. We excluded eviction filing rates and location data from the model, feeding in the Census Bureau data, the unemployment rate, and housing data.

One downside of using K-means to cluster counties is that the k, or number of clusters, must be specified in advance, but we used both the elbow method and the silhouette method to recommend the optimal amount of clusters. Both determined that 3 is the optimal amount of clusters so we built a K-means model using the Hartigan-Wong algorithm with 3 clusters


Optimal number of clusters
Total Within Sum of Square
50000
40000
30000
1
2
3
4
5
6
7
8
9
10
Number of clusters k

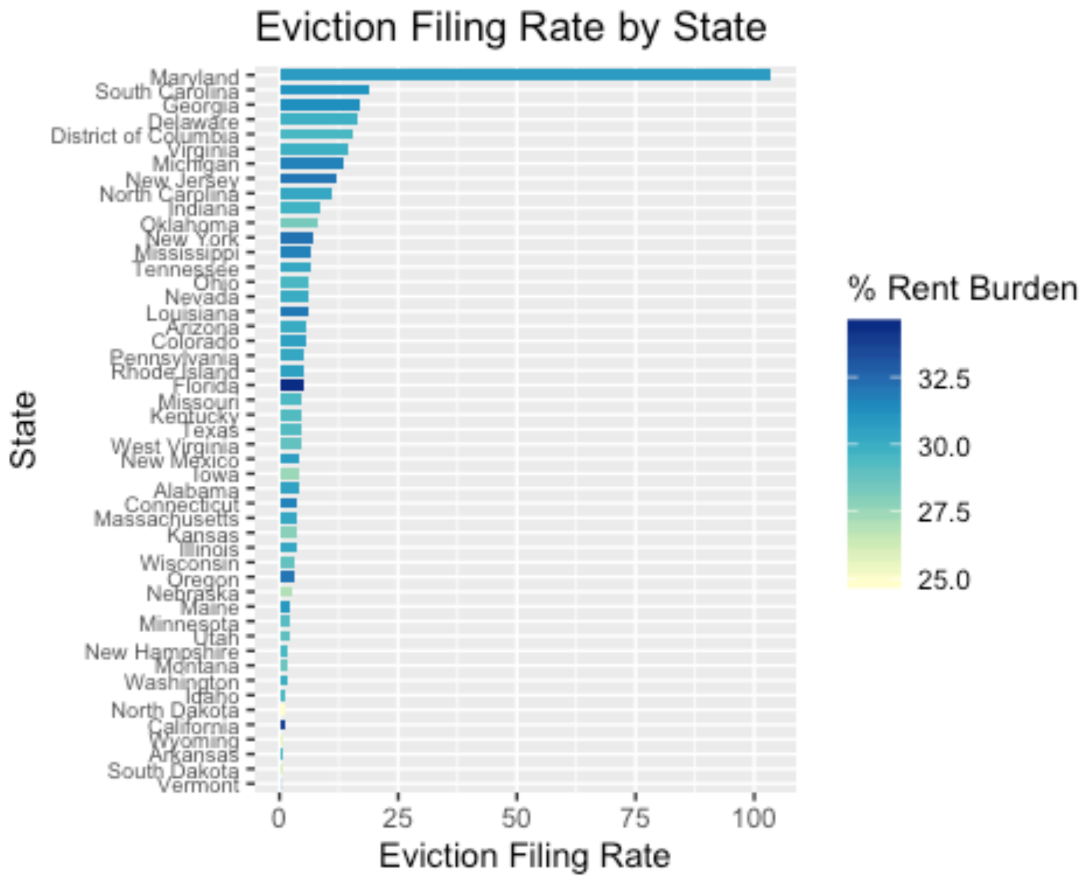

Optimal number of clusters
Average silhouette width
0.0
0.1
0.2
1
2
3
4
5
6
7
8
9
10
Number of clusters k

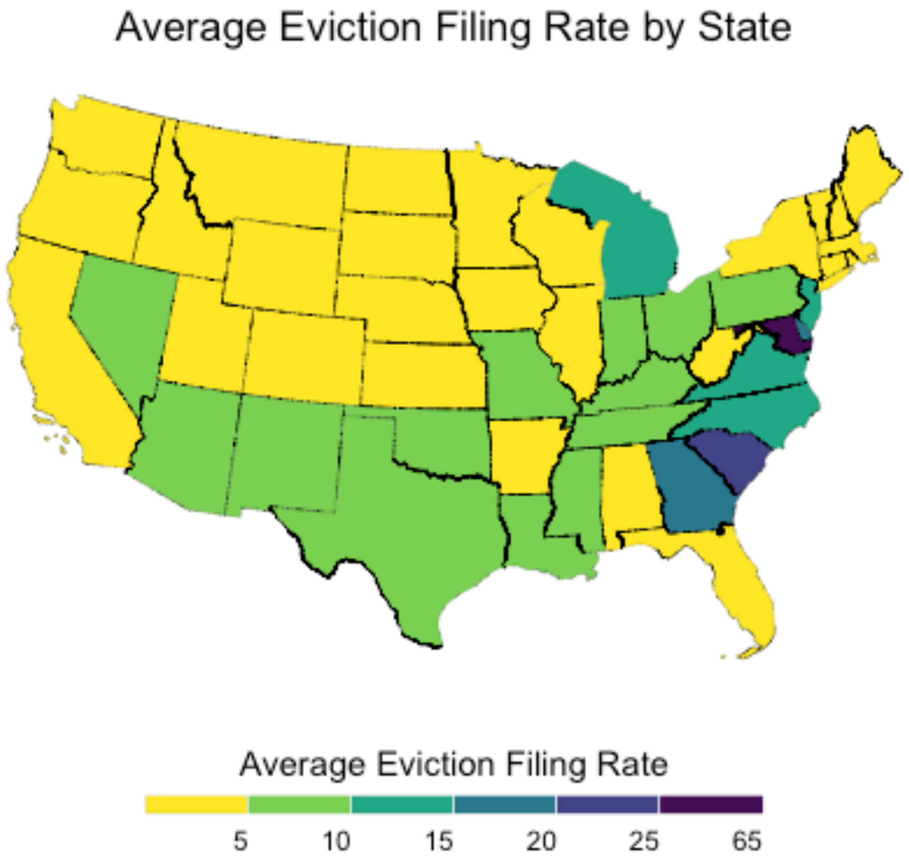


When exploring the data by visualizing and summarizing data, one of our initial discoveries was that four states, South Dakota, North Dakota, Alaska, and Arkansas, did not have eviction rate data because the Eviction Lab team decided to only include data and derive estimates from counties that had at least two consecutive years worth of data. Because of this, we decided to focus our analysis on the eviction filing rate instead.

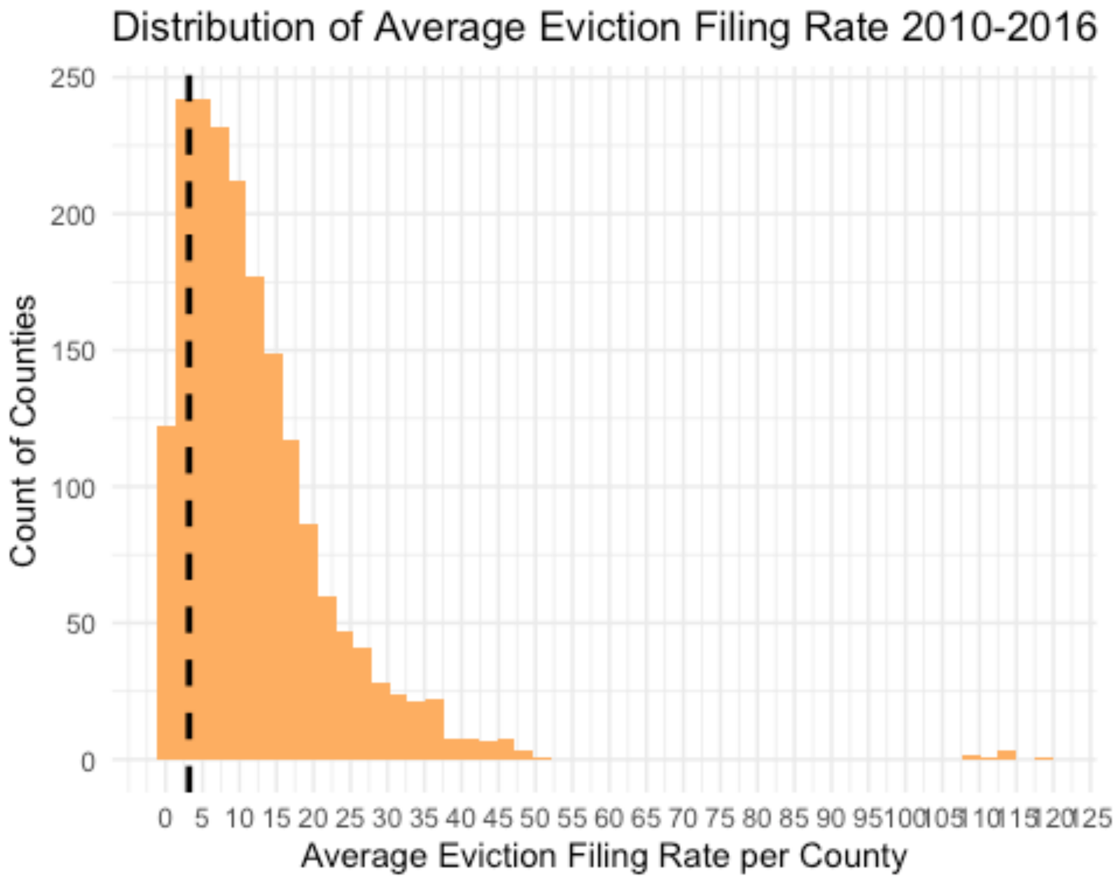
After analyzing our main variable of interest, eviction filing rate, we realized that the percentage of null eviction filing rate values at the county level averaged around 17% from 2000 to 2010, and dropped to 13% after 2010, while the state level averaged around 6% null values for 2000-2010, and dropped to less than 1% of null values for data from 2010-2016. Therefore, we decided to subset the data from 2010 onwards for both state and counties.

Starting with a higher level of analysis by reviewing the state level data, we used the newly altered dataset to analyze the eviction filing rate by state. It was apparent that the highest eviction filing rate for all seven years was in Maryland, with South Carolina having the second highest eviction filing rate all seven years.

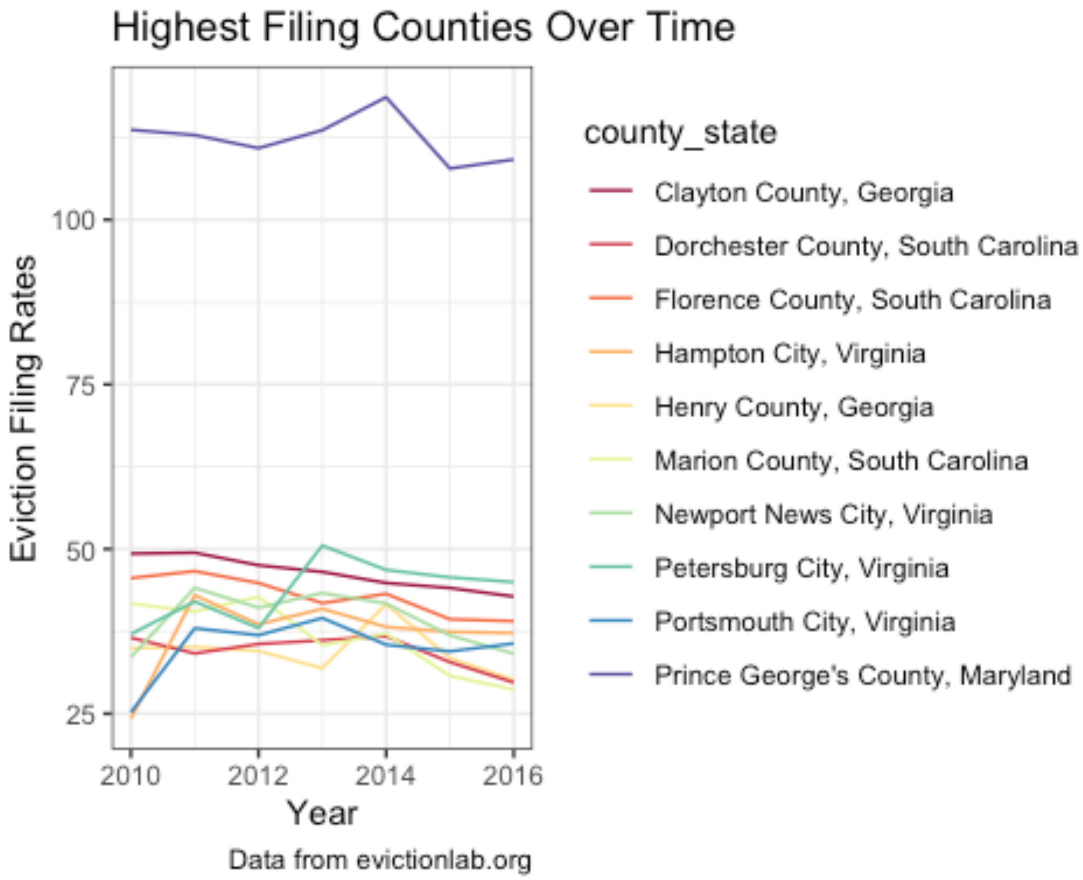

Eviction Filing Rate by State
State
Maryland
South Carolina
Georgia
Delaware
District of Columbia
Virginia
Michigan
New Jersey
North Carolina
Indiana
Oklahoma
New York
Mississippi
Tennessee
Ohio
Nevada
Louisiana
Arizona
Colorado
Pennsylvania
Rhode Island
Florida
Missouri
Kentucky
Texas
West Virginia
New Mexico
Iowa
Alabama
Connecticut
Massachusetts
Kansas
Illinois
Wisconsin
Oregon
Nebraska
Maine
Minnesota
Utah
New Hampshire
Montana
Washington
Idaho
North Dakota
California
Wyoming
Arkansas
South Dakota
Vermont
0
25
50
75
100
Eviction Filing Rate
% Rent Burden
32.5
30.0
27.5
25.0


Average Eviction Filing Rate by State
Average Eviction Filing Rate
5
10
15
20
25
65

Looking at the distribution of eviction filing rates, the median average eviction filing rates per county from 2010-2016 were around 3.00 percent with a heavy right tail skew with a frequency of about 250 counties. There were over 200 counties with averages during this time period that experienced average eviction filing rates that ranged from 20 to 50.



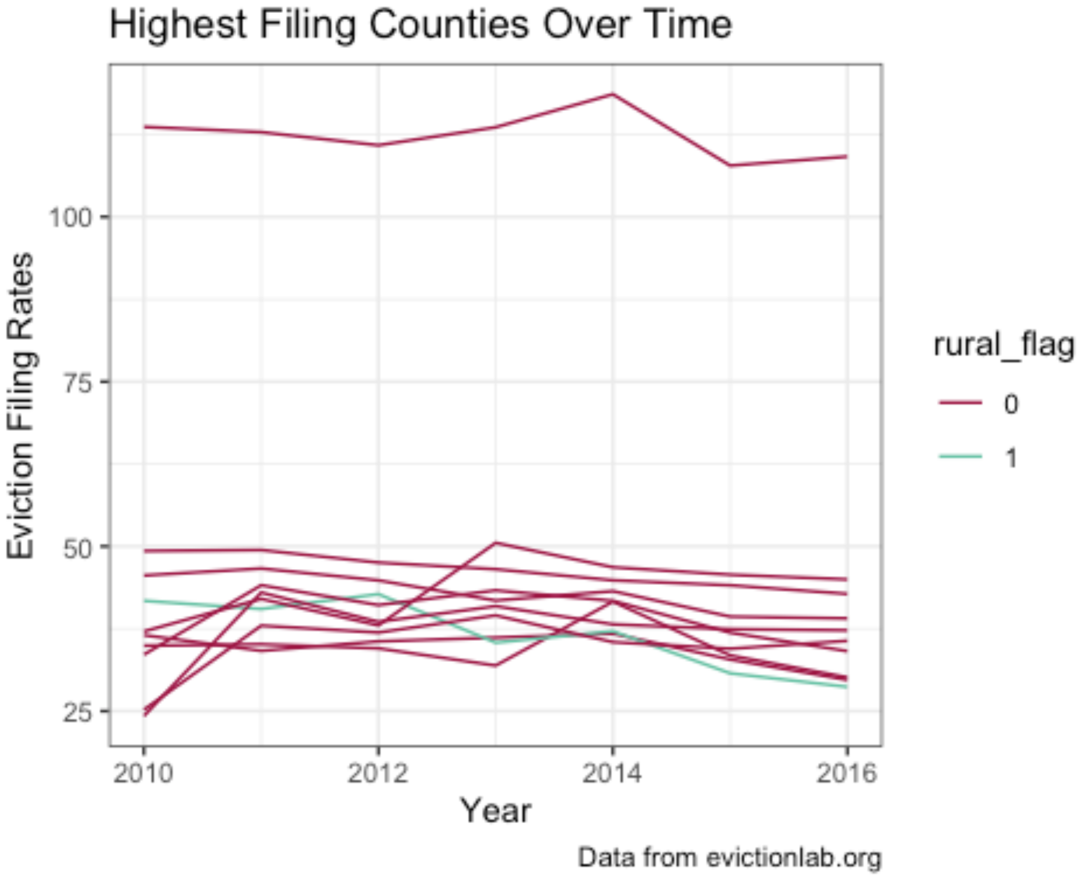
The top ten counties with average eviction rates over the span of the seven years, all belonged to the Southern region with Prince George's County, Maryland experiencing over 100.00 average eviction filing rates. There are several reasons why we think Maryland has the highest and eviction filing rates over 100. One, according to the Eviction Lab's Methodology Report most jurisdictions start with an out of court notice, whereas Maryland starts immediately with an eviction filing in court. Additionally, as of 2006 landlords can electronically start an eviction filing in Maryland on Maryland's District Courts website. Lastly, Maryland has greater diversity than most of the country with the state being home to ~50% non-whites compared to the national average of 24% non-white.



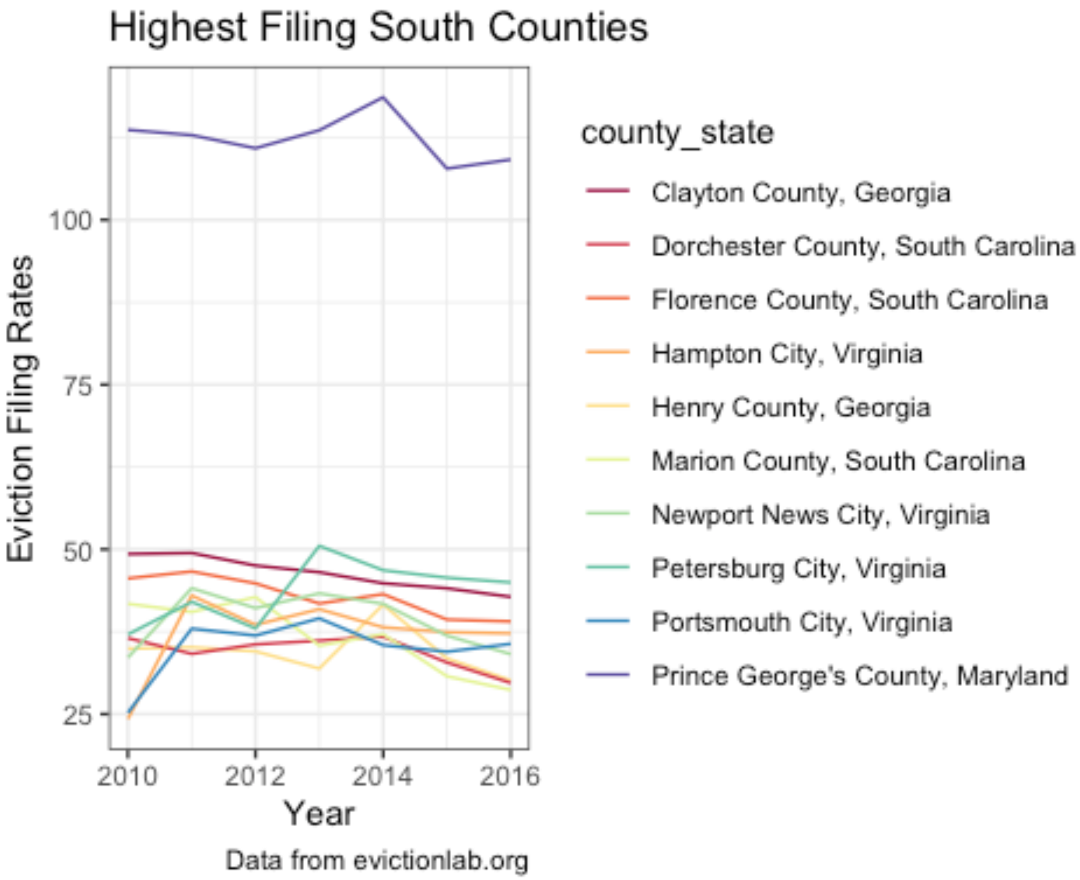
Maryland also experienced a 19% increase in poverty rate from the 1990s to 2016, while seeing a 25% increase in population during that same time according to Washington Top News. Prince George County's according to our K-means model is similar to San Bernardino County, CA with similar percentages of renter occupied rates, median gross rent, rent burden rates, and property values. San Bernardino County, however, had an even higher poverty rate and higher population, but had a much smaller average eviction filing rate with 6% versus 112%. This shows that there might be something missing from the data that is not being captured in this disparity. The only stark difference in this data from these counties is Prince George's County has about ⅓ more of people of color.

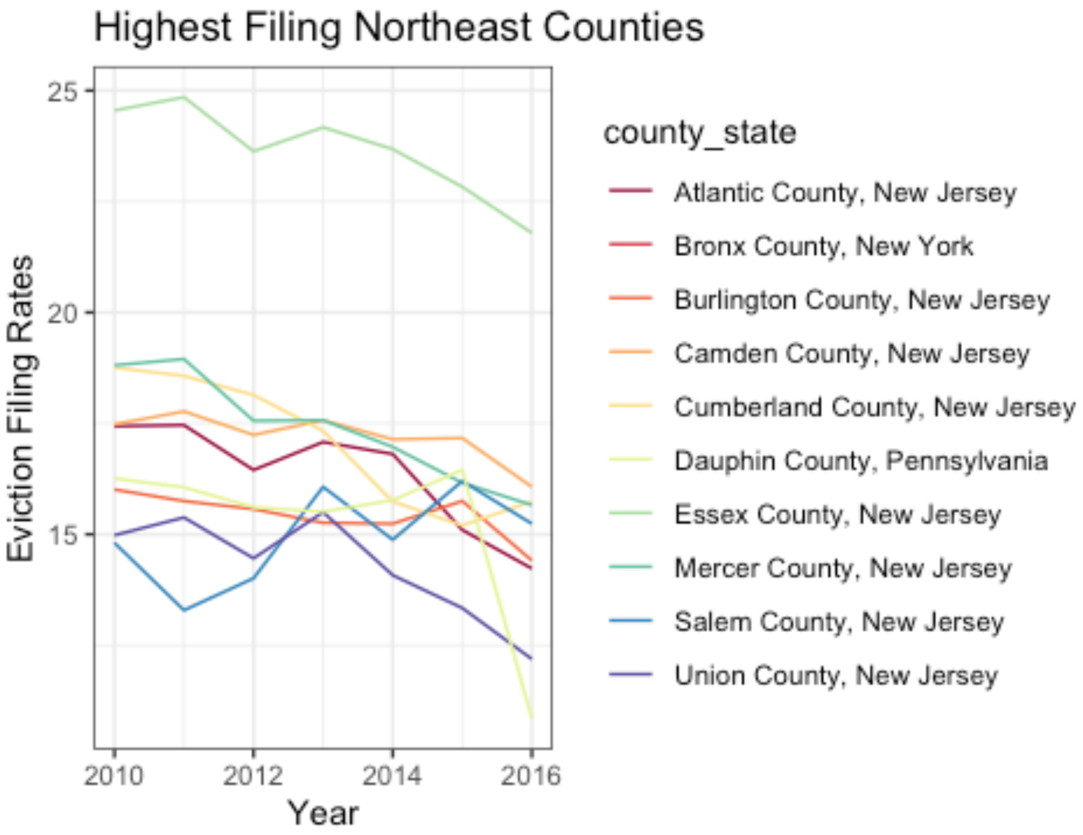
According to an article in The News & Advance, Lynchburg, Virginia experienced high eviction rates. The article explained that most tenants do not go through with the court order and end up leaving because they are worried about the uphill legal battle or the repercussions–most landlords will refuse to rent to formerly evicted residents. From 2008 to 2017, 60% of 21,000 eviction suits were in favor of the landlord whereas the rest were dismissed and less than 1% were won in favor of the tenants. After the Eviction Lab data came out, according to the same article, laws have been passed to help tackle the high eviction rates Virginia has been experiencing. In the above chart visualizing the top ten counties by average eviction filing rates during 2010-2016, four of them belonged to Virginia.

Marion County, South Carolina was the only one out of the top highest average eviction filing rate counties that was predominantly rural. Marion County, similar populations, poverty rates, unemployment rates, and renter occupied rates had stark different eviction rates.

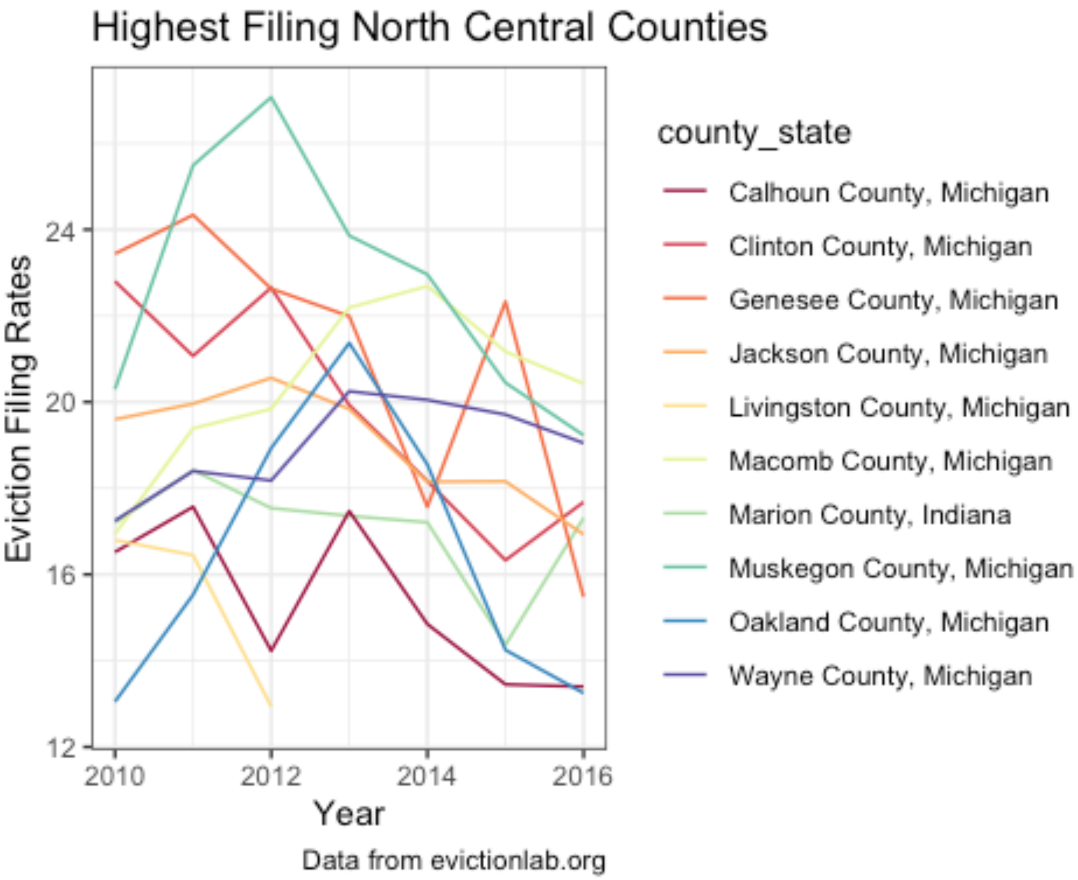


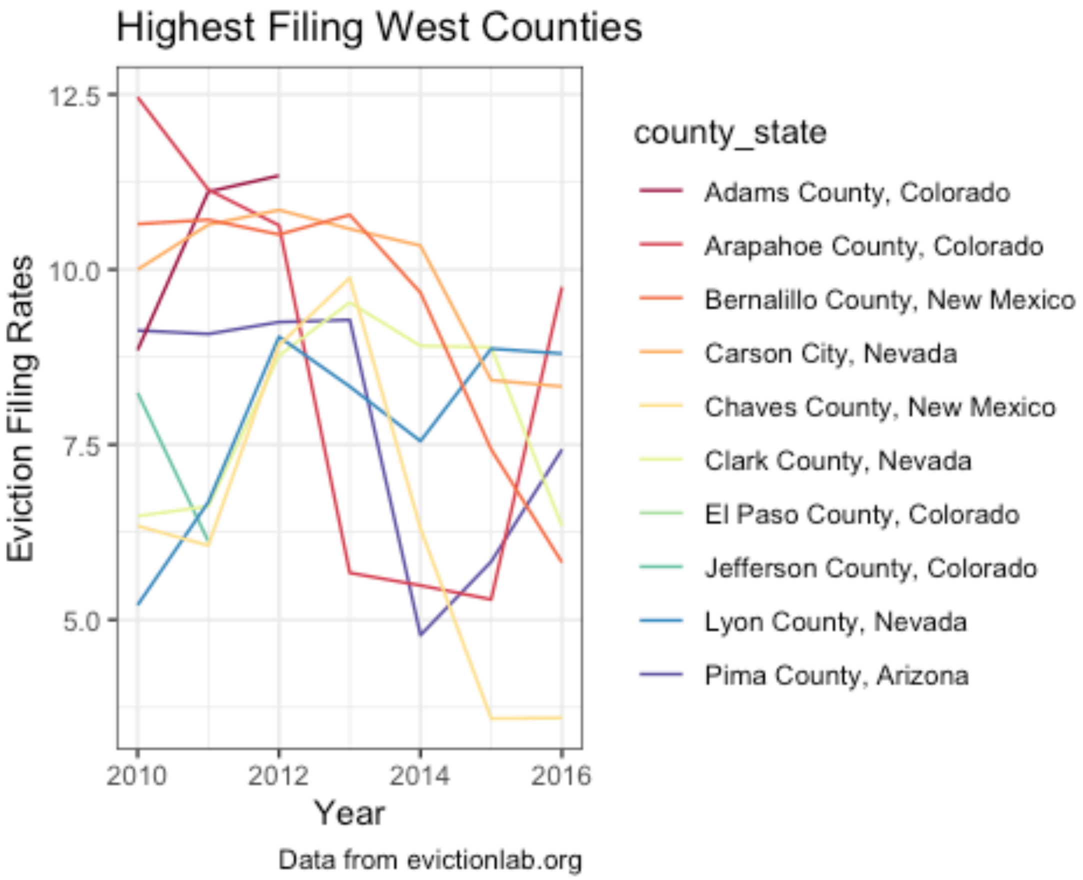
The top ten filing counties by regions also seemed to center around the same state, with nine out of the ten highest eviction filing North Central counties all belonging to Michigan. Nine out of the top counties in Northeast all belonged to New Jersey. The top ten western counties were spread out across three states, Colorado, New Mexico, and Nevada, versus a common one. Across the regions, most of the counties with the top eviction filing rates have experienced a decline since 2010, except Lyon County in Nevada, Arapahoe County in Colorado, and Marion County in Indiana.

Highest Filing South Counties
Eviction Filing Rates
100
75
50
25
2010
2012
2014
2016
Year
county_state
Clayton County, Georgia
Dorchester County, South Carolina
Florence County, South Carolina
Hampton City, Virginia
Henry County, Georgia
Marion County, South Carolina
Newport News City, Virginia
Petersburg City, Virginia
Portsmouth City, Virginia
Prince George's County, Maryland
Data from evictionlab.org

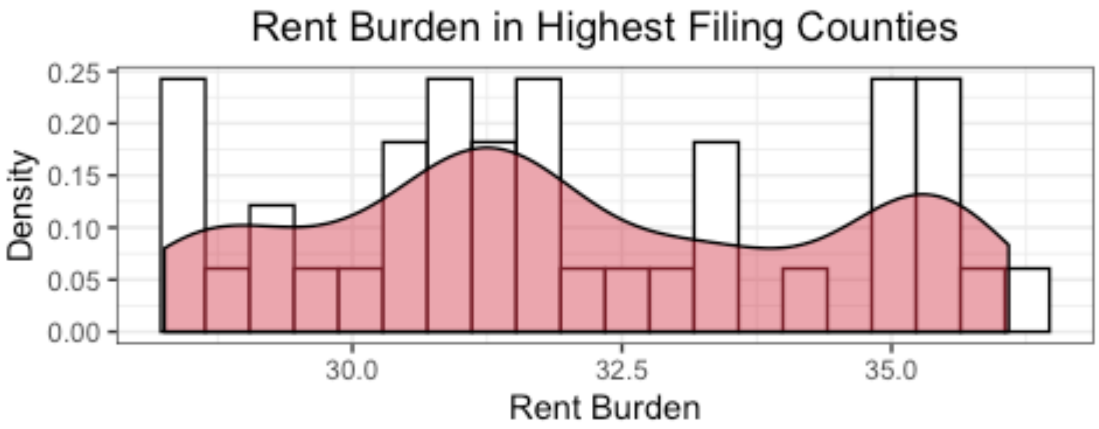


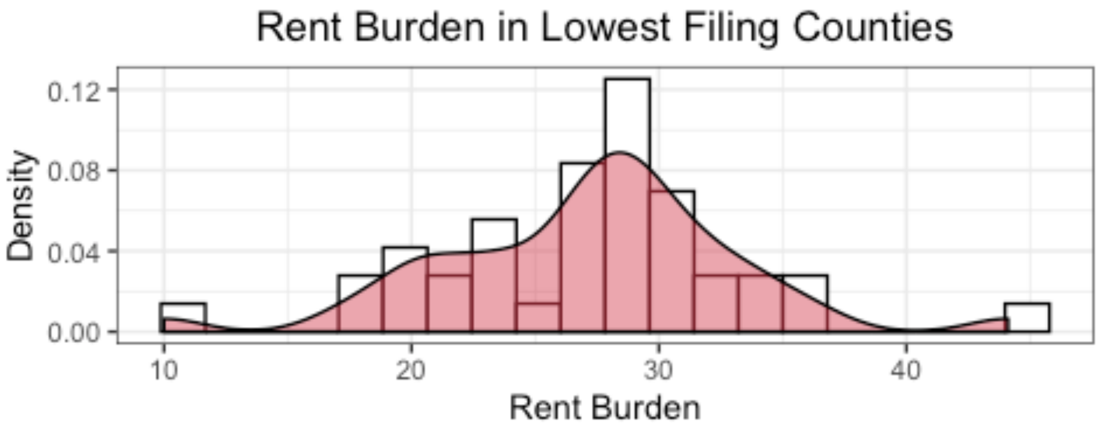
Data from evictionlab.org

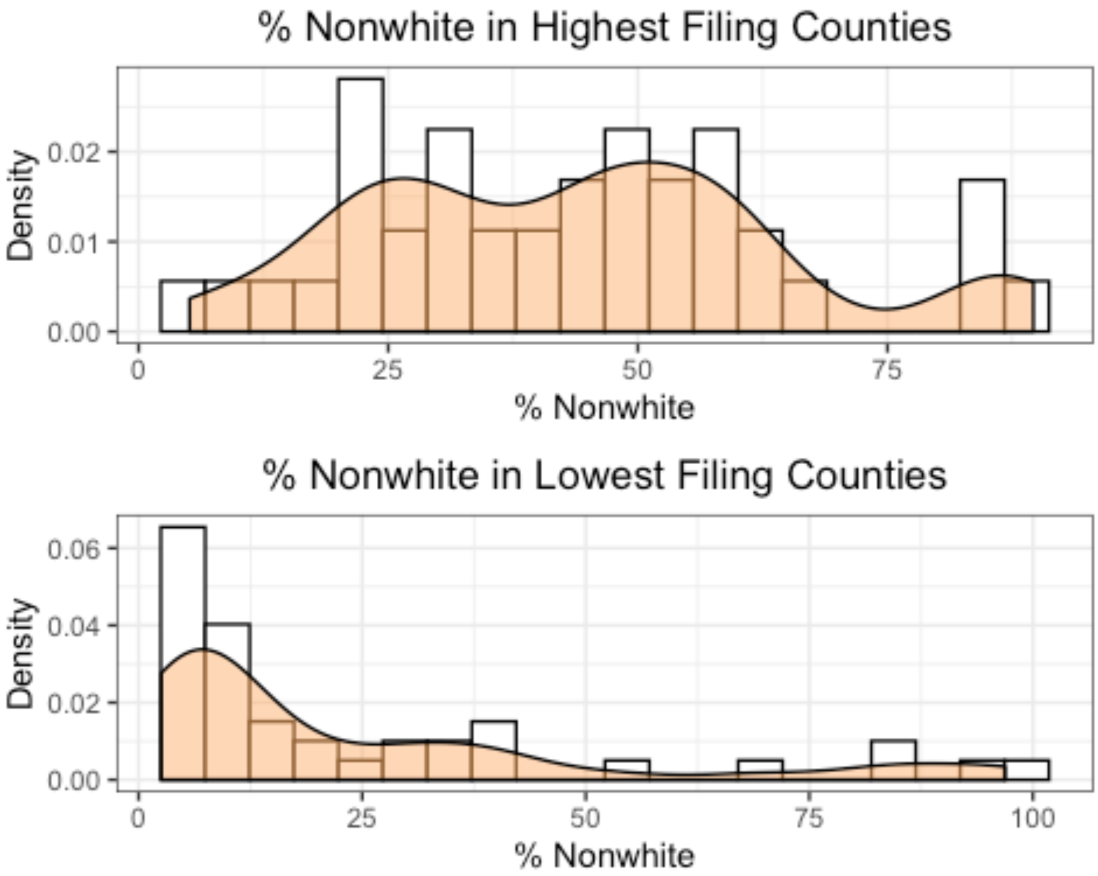

Highest Filing North Central Counties
Eviction Filing Rates
24
20
16
12
2010
2012
2014
2016
Year
county_state
Calhoun County, Michigan
Clinton County, Michigan
Genesee County, Michigan
Jackson County, Michigan
Livingston County, Michigan
Macomb County, Michigan
Marion County, Indiana
Muskegon County, Michigan
Oakland County, Michigan
Wayne County, Michigan
Data from evictionlab.org

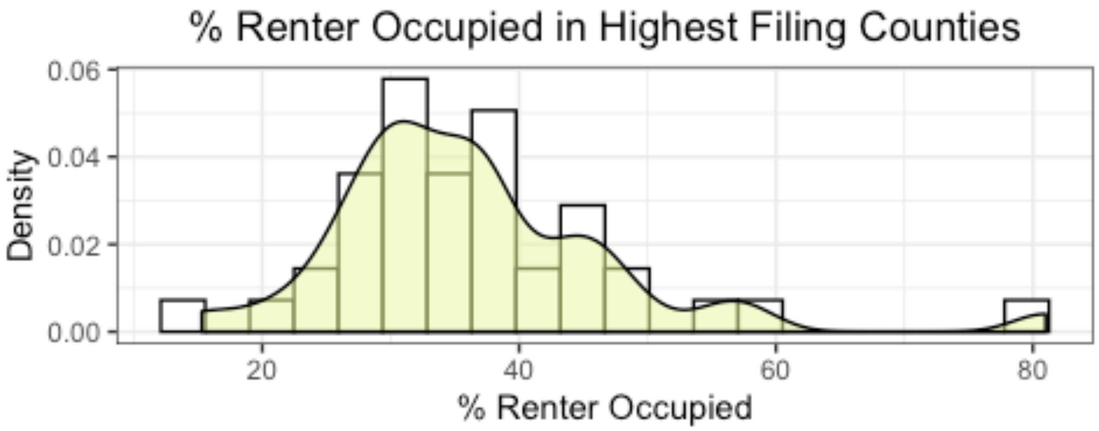


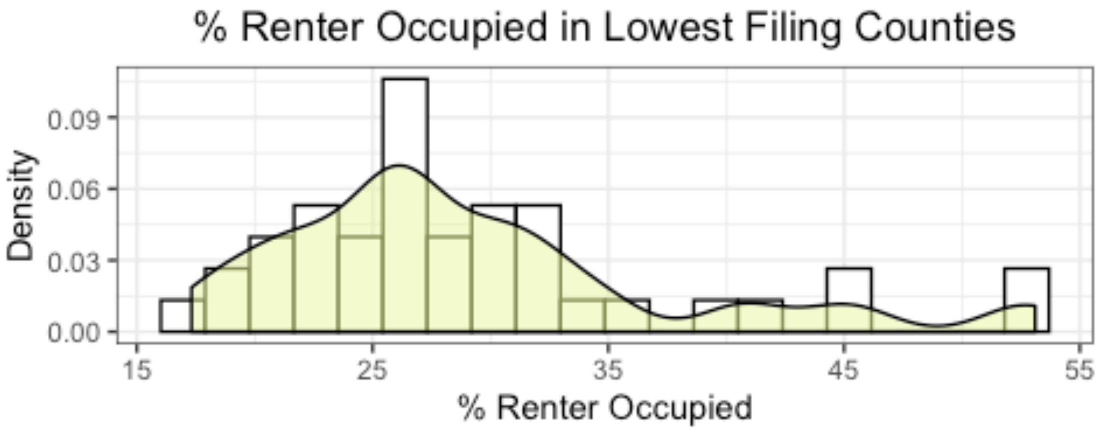
We then compared some attributes in the highest filing counties versus the lowest filing counties utilizing histograms and density area graphs overlapping the histograms. The attributes that seemed to differ across these two segments were rental costs, percentage of rural population, the percentage of white population, rent burden, and median household income. The top ten filing counties had lower percentages of rural, white population, and higher concentrated areas of rent burden and a larger spread of median income.

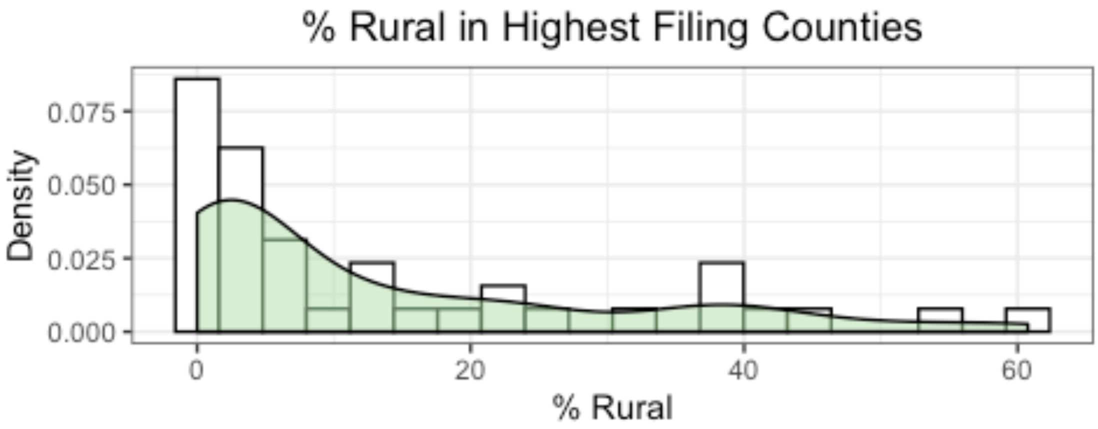

Rent Burden in Highest Filing Counties
Density
0.25
0.20
0.15
0.10
0.05
0.00
30.0
32.5
35.0
Rent Burden


Rent Burden in Lowest Filing Counties
Density
0.12
0.08
0.04
0.00
10
20
30
40
Rent Burden

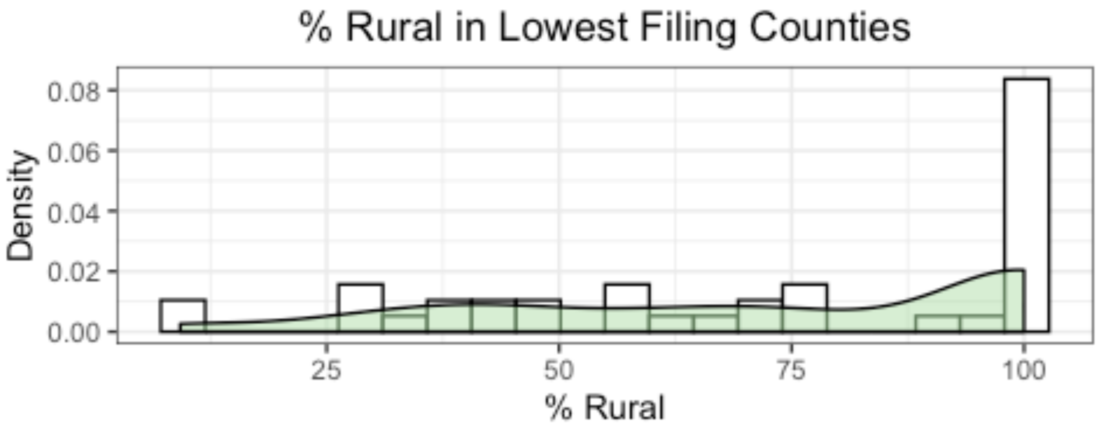

% Nonwhite in Highest Filing Counties
Density
0.00
0.01
0.02
0
25
50
75
% Nonwhite
% Nonwhite in Lowest Filing Counties
Density
0.00
0.02
0.04
0.06
0
25
50
75
100
% Nonwhite


% Renter Occupied in Highest Filing Counties

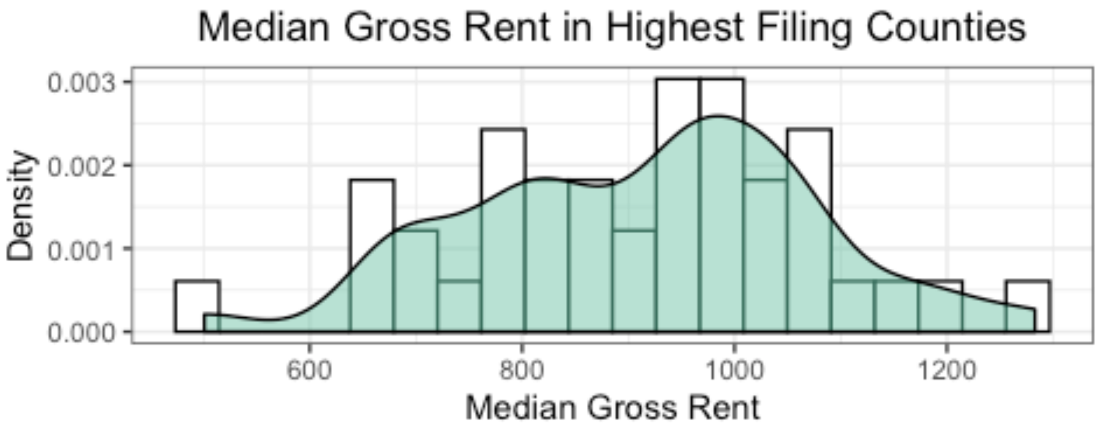

% Renter Occupied in Lowest Filing Counties



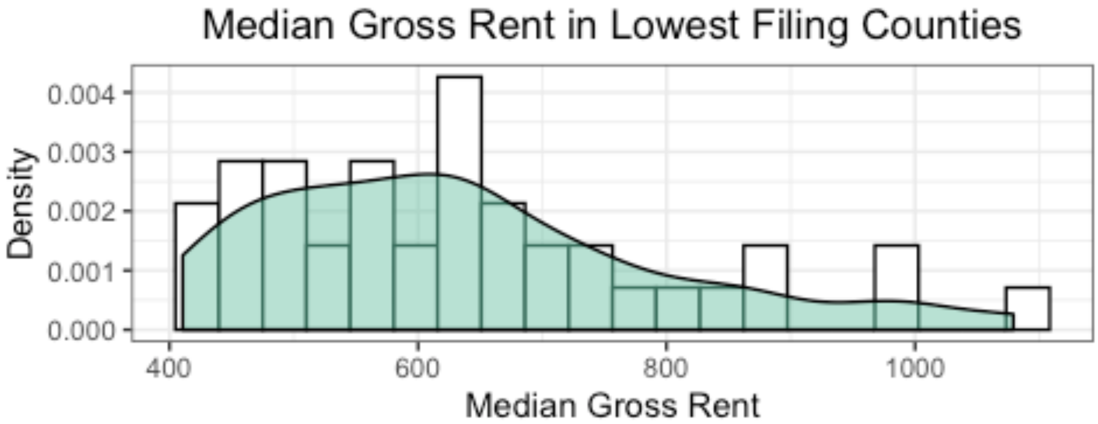
% Rural in Highest Filing Counties



% Rural in Lowest Filing Counties

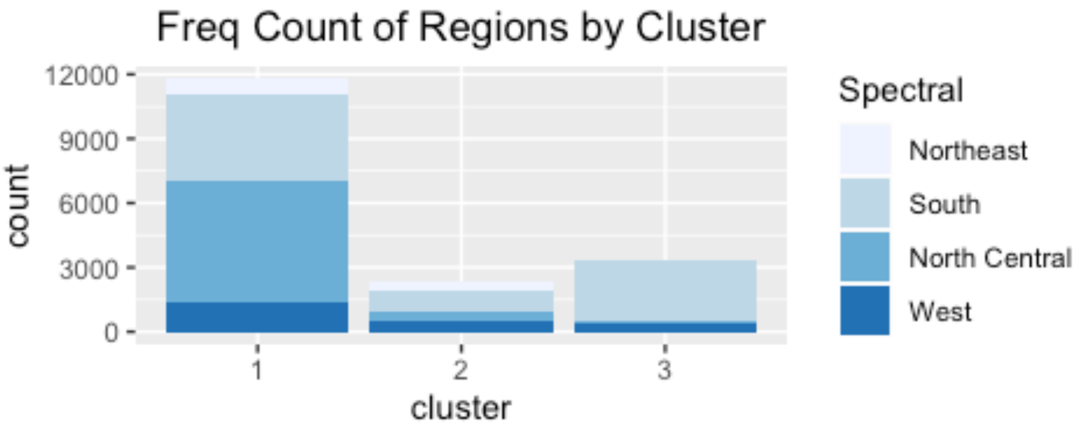


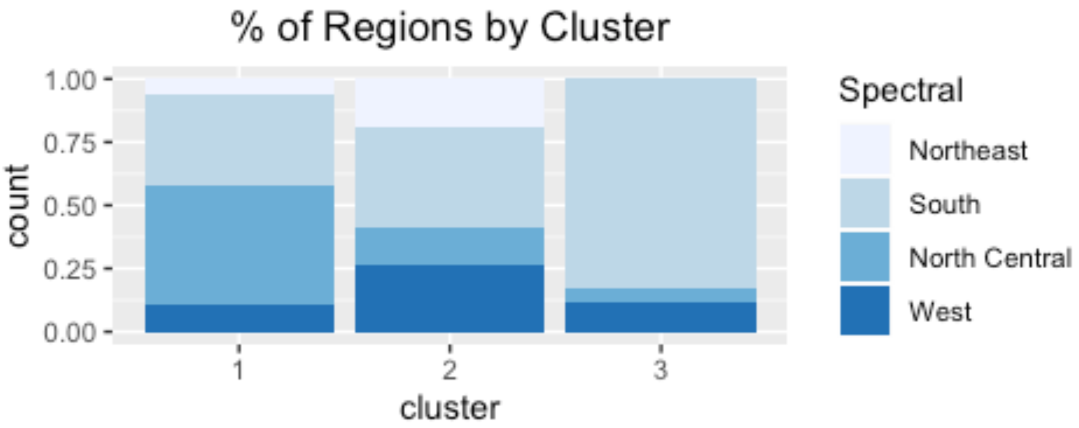
Median Gross Rent in Highest Filing Counties



Median Gross Rent in Lowest Filing Counties

For our Shiny app, we wanted to have two main components; a map that the user can use to visualize the eviction filing rate data or percent renter occupied data by state for each year, and an analysis component that a visitor to the app can use to find eviction filing rate data for a county of their choice, and the ability to easily and compare their selection to other counties with similar demographic make-up. The descriptive stats are time series data and the user is able to look at eviction filing rates and X by state from 2010 to 2016. The county level data has a time series graph that trends eviction filing rates or unemployment rates by year of the county selected from the dropdown menu. In the counties tab it also populates a table with averages of some attributes of that county along with 4 other randomly selected cities pertaining to the same cluster such as rural percentage, monthly gross rent, median income, median property value, some demographic percentages, etc. Using this table the user can compare the average eviction filing rate across the county selected and the randomly four other selected from the same cluster to see if they also share similar eviction rates. Most of the counties were mapped to cluster one, whereas the South makes up the majority of cluster 2, and cluster with the smallest frequency of counties are evenly distributed across regions.


Freq Count of Regions by Cluster


% of Regions by Cluster

After analyzing the Eviction Labs dataset we were able to conclude there may have been some bias and issues within the dataset. The dataset had instances of omitted variable bias. One of them was the information Eviction Labs aggregated from case records did not have demographics or other information of the exact people that were filed against for eviction, except for where the filing occurred which was geocoded by the Eviction Labs from addresses from the case records which then the frequencies were assigned to counties. We had to use Census Bureau data which was a combination of demographic and economic variables at the county level to see if we could find a pattern within eviction filing rates that were filed in the same county across all the counties in the United States during 2010-2016. The makeup/characteristics of the county where the eviction filings occurred were the best estimates on trying to extrapolate patterns across counties. Ideally we would have the demographics from the people evicted.

The second omitted variable bias was access to data in certain states and/or counties so estimates/imputations were created by the Eviction Lab. According to the Eviction Lab's methodology report, any states that did not have statewide coverage statistics of county level eviction data were not included in the calculations. States whose counties were estimated from state level data were Alaska, North Dakota, South Dakota, and Arkansas. There were also counties that had underestimated eviction counts, including counties that were in the urban parts of New York, California, New Jersey, Maryland, etc. Rural areas also experienced data collection difficulties in states such as Kentucky, Louisiana, Tennessee, Wyoming, and the District of Columbia. For example, in California it was stated in the methodology report that many cases that ended in eviction were not publicly accessible and it was hard to access data as a whole. This solidified our decision to look at eviction filing rates versus actual eviction rates, because data collection can be quite cumbersome across states. Which is why there are higher eviction filing rates and lower eviction rates in some areas such as Prince George's County in Maryland. Additionally, there were missing data for counties in specific years only. For instance, the eviction related data for the District of Columbia was missing for three years in the middle of the collection period.

These data omissions noted could impact the conclusions drawn from the dataset, particularly when comparing the eviction-related data among multiple counties. As a result, even in counties with demographic and economic makeups consistent with higher filing rates we may not have the full picture of evictions based on the collection methods and availability of data. However, with the data that was available to us, we can conclude that eviction filing rates positively correlated with higher diversity rates, a larger population, higher rental occupancy, and higher rent burden.

References

Data: Matthew Desmond, Ashley Gromis, Lavar Edmonds, James Hendrickson, KatieKrywokulski, Lillian Leung, and Adam Porton. Eviction Lab National Database: Version1.0. Princeton: Princeton University, 2018,www.evictionlab.org.

Methodology Report: Matthew Desmond, Ashley Gromis, Lavar Edmonds, James Hendrickson, Katie Krywokulski, Lillian Leung, and Adam Porton. Eviction Lab Methodology Report: Version 1.0. Princeton: Princeton University, 2018, www.evictionlab.org/methods.