Analysis of Evictions

January 2019

Executive Summary

This document presents an analysis of data concerning the prediction of evictions at the county level from socioeconomic and demographic indicators. The analysis is based on 2546 observations of eviction data, each containing specific characteristics of housing, ethnicity, economic, health and demographic categories.

After exploring the data by calculating summary and descriptive statistics, and by creating visualizations of the data, several potential relationships between socioeconomic and demographic characteristics, and evictions were identified. After exploring the data, a regression model to predict the number of evictions from its features was created.

After performing the analysis, the author presents the following conclusions:
While many factors can help indicate the number of evictions at county level, significant features found in this analysis were:

- **population** - Total population. The number of evictions tends to increase as the population size increases.
- **renter_occupied_households** - Count of renter-occupied households. The number of evictions tends to increase as more households are occupied by renters.
- **median_property_value** - Median property values in the lower range of $137000 - $142000 tend to have the highest number of evictions.
- **pct_af_am** - Percent of population that is Black or African American alone and not Hispanic or Latino. There is a positive association between the percent of the population that is African American and the number of evictions, on average.
- **pct_asian** - Percent of population that is Asian alone and not Hispanic or Latino. There is a positive association between the percent of the population that is Asian and the number of evictions, on average.
- **pct_other** - Percent of population that is other race alone and not Hispanic or Latino. There is a positive association between the percent of the population that is of Other race and the number of evictions, on average.
- **pct_female** - Percent of population that is female. The number of evictions tends to be slightly higher for the female gender of the population.
- **pct_civilian_labor** - Civilian labor force, annual average, as percent of population. There is a positive association between the percent of the population that is in the Civilian Labor Force and the number of evictions, on average.
- **pct_below_18_years_of_age** - Percent of population that is below 18 years of age. There is a positive association between the percent of the population that is Below 18 years of Age and the number of evictions, on average.
- **pct_adults_bachelors_or_higher** - Percent of adult population which has a bachelor's degree or higher as highest level of education achieved. There is a positive association between the percent of the population that has a Bachelors Degree and the number of evictions, on average.
- **economic_typology** - County Typology Codes "classify all U.S. counties according to six mutually exclusive categories of economic dependence and six overlapping categories of policy-relevant themes. Non-Specialized counties have a much higher number of evictions than other economic typologies on average.

Analysis of Evictions

January 2019

Initial Data Exploration

The initial exploration of the data began with summary and descriptive statistics.

Individual Feature Statistics

Summary statistics for minimum, maximum, mean, median, standard deviation, and distinct count were calculated for numeric columns, and the results taken from 2546 observations are shown below:

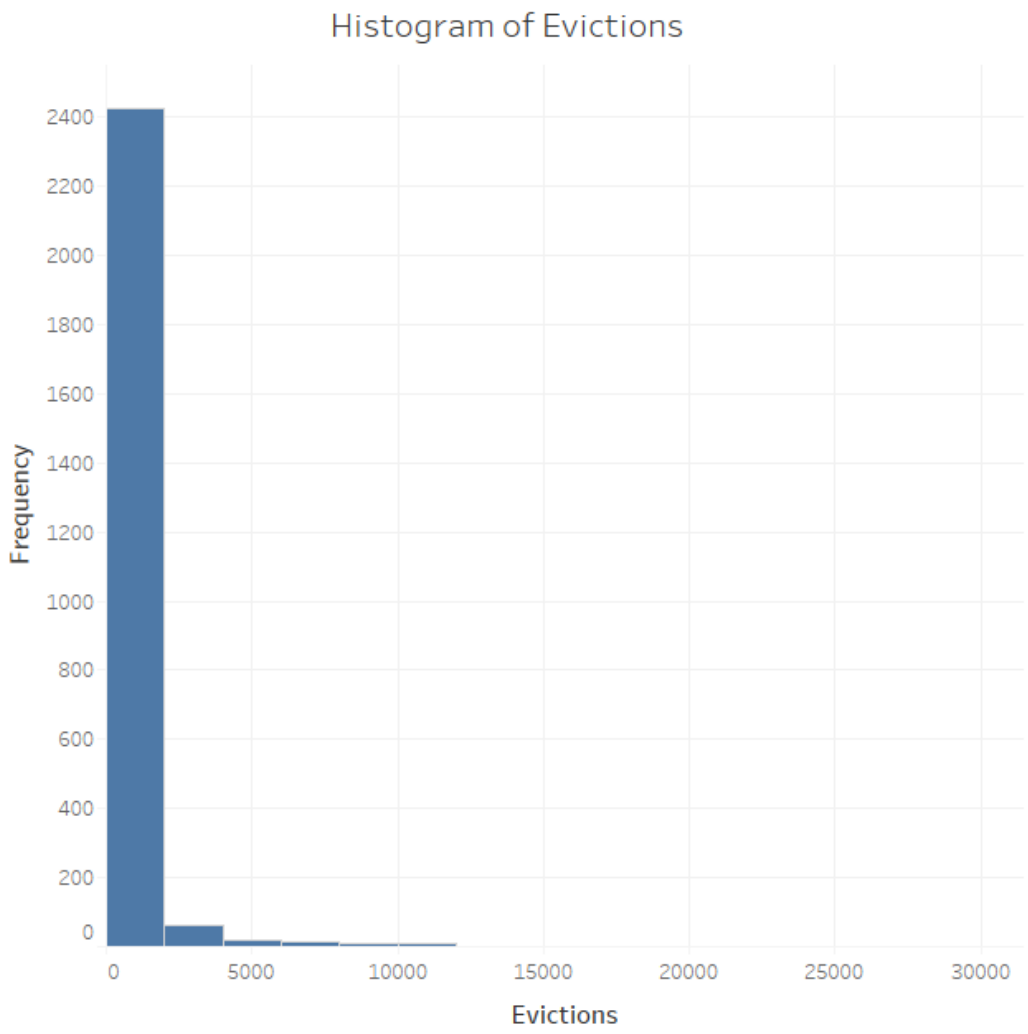
Column	Min	Max	Mean	Median	Std Dev	DCount
population	116	5279852	106245.938	23863	322852.005	2495
renter_occupied_households	14	882101	15008.009	2580.5	53333.6842	2222
pct_renter_occupied	7.305	70.61	28.1473904	26.866	7.94014021	2425
median_gross_rent	336	1728	688.838178	642	183.722492	606
median_household_income	19328	123452	46050.6014	44480	11584.6272	1887
median_property_value	32287	904937	129609.579	108844	76236.6063	2520
rent_burden	9.986	49.535	28.5205613	28.78	4.4531646	2352
pct_white	0.0509	0.99511401	0.77627228	0.85547755	0.20114925	2546
pct_af_am	0	0.85899743	0.08977388	0.02186448	0.14555027	2459
pct_hispanic	0	0.93620077	0.09060446	0.03606029	0.14227411	2537
pct_am_ind	0	0.80136429	0.01246696	0.00238699	0.05140591	2318
pct_asian	0	0.33767244	0.01165279	0.00496056	0.02437189	2338
pct_nh_pi	0	0.09652725	0.00064493	0	0.00256015	1187
pct_multiple	0	0.20847454	0.01769836	0.01456118	0.01607246	2501
pct_other	0	0.01982206	0.00088634	0.00020171	0.00176253	1469
poverty_rate	0	44.732	12.3698559	11.543	5.65447649	2353
pct_civilian_labor	0.213	1	0.46768853	0.469	0.0738127	358
pct_unemployment	0.019	0.182	0.05942302	0.057	0.02095314	121
pct_uninsured_adults	0.051	0.495	0.21590888	0.214	0.06551049	319
pct_uninsured_children	0.014	0.283	0.08638452	0.077	0.0403137	204
pct_adult_obesity	0.151	0.471	0.30665632	0.308	0.04192981	239
pct_adult_smoking	0.046	0.511	0.214645	0.211	0.06086331	303
pct_diabetes	0.041	0.198	0.1096469	0.109	0.02232066	133
pct_low_birthweight	0.04	0.231	0.08406529	0.08	0.02143776	44
pct_excessive_drinking	0.042	0.309	0.16327419	0.163	0.05018052	238
pct_physical_inactivity	0.12	0.441	0.27615593	0.278	0.05209924	268
air_pollution_particulate_matter_value	7.5425	14.8809467	11.7031246	12.0164565	1.5516251	2545
homicides_per_100k	-0.4	50.49	5.8475	4.5	5.05731005	527
motor_vehicle_crash_deaths_per_100k	3.09	76.05	20.9227659	19.5	10.1341447	1644
heart_disease_mortality_per_100k	109	482	279.705813	276	57.1508487	275
pop_per_dentist	490	28130	3504.29457	2694.5	2635.39261	1066
pop_per_primary_care_physician	189	23399	2587.69675	1980	2216.14734	855
pct_female	0.285	0.572	0.49912647	0.504	0.02424743	158
pct_below_18_years_of_age	0.088	0.359	0.2261791	0.225	0.03272536	202
pct_aged_65_years_and_older	0.063	0.345	0.17158327	0.168	0.04192787	231
pct_adults_less_than_a_high_school_diploma	0.0160	0.46593186	0.14789148	0.13086913	0.06807717	1857
pct_adults_with_high_school_diploma	0.1271	0.55034895	0.35319772	0.35657371	0.07016671	1920

Analysis of Evictions

January 2019

Column	Min	Max	Mean	Median	Std Dev	DCount
pct_adults_with_some_college	0.137	0.44869215	0.30091055	0.3013013	0.05181136	1763
pct_adults_bachelors_or_higher	0.0188	0.5840796	0.19800025	0.17667677	0.08641523	1845
birth_rate_per_1k	3.6121	28.9228678	11.4819226	11.3060369	2.56597868	2542
death_rate_per_1k	0	27.3972603	10.4071343	10.4780881	2.72013523	2544
evictions	0	29251	378	29	1405.277	646

Since **evictions** is of interest in this analysis, it was noted that the mean and median of this value are significantly different and that the comparatively large standard deviation indicates that there is considerable variance in the number of evictions. A histogram of the **evictions** column shows that the evictions values are right-skewed – in other words, most evictions occur at the lower end of the evictions range, as shown below:



Analysis of Evictions

January 2019

In addition to the numeric values, the evictions observations include categorical features, including:

- **year** - Year, denoted as a or b
- **state** - Unique identifier for each state
- **rucc** - Rural-Urban Continuum Codes "form a classification scheme that distinguishes metropolitan counties by the population size of their metro area, and nonmetropolitan counties by degree of urbanization and adjacency to a metro area".
  Metro - Counties in metro areas of 1 million population or more
  Metro - Counties in metro areas of 250,000 to 1 million population
  Metro - Counties in metro areas of fewer than 250,000 population
  Nonmetro - Completely rural or less than 2,500 urban population, adjacent to a metro area
  Nonmetro - Completely rural or less than 2,500 urban population, not adjacent to a metro area
  Nonmetro - Urban population of 2,500 to 19,999, adjacent to a metro area
  Nonmetro - Urban population of 2,500 to 19,999, not adjacent to a metro area
  Nonmetro - Urban population of 20,000 or more, adjacent to a metro area
  Nonmetro - Urban population of 20,000 or more, not adjacent to a metro area
- **urban_influence** - Urban Influence Codes "form a classification scheme that distinguishes metropolitan counties by population size of their metro area, and nonmetropolitan counties by size of the largest city or town and proximity to metro and micropolitan areas." Large-in a metro area with at least 1 million residents or more
  Micropolitan adjacent to a large metro area
  Micropolitan adjacent to a small metro area
  Micropolitan not adjacent to a metro area
  Noncore adjacent to a large metro area
  Noncore adjacent to a small metro and does not contain a town of at least 2,500 residents
  Noncore adjacent to a small metro with town of at least 2,500 residents
  Noncore adjacent to micro area and contains a town of 2,500-19,999 residents
  Noncore adjacent to micro area and does not contain a town of at least 2,500 residents
  Noncore not adjacent to a metro/micro area and contains a town of 2,500 or more residents
  Noncore not adjacent to a metro/micro area and does not contain a town of at least 2,500 residents
  Small-in a metro area with fewer than 1 million residents
- **economic_typology** - County Typology Codes "classify all U.S. counties according to six mutually exclusive categories of economic dependence and six overlapping categories of policy-relevant themes.
  Farm-dependent
  Federal/State government-dependent
  Manufacturing-dependent
  Mining-dependent
  Nonspecialized
  Recreation

Bar charts were created to show frequency of these features, and indicate the following:

- The average number of evictions across states varies widely, ranging from 5 to more than 2,000.
- The most common state is 930f257, which has almost twice as many observations as the next most common states.
- Counties with a Non-Specialized economic typology are the most common have a significantly high number of evictions that other economic typologies.
- Counties with rucc codes in metro areas have a significantly high number of evictions that those in non-metro areas.
- Counties with Urban Influence codes in a metro are the most common and have a higher number of evictions that those in non-core areas.

Analysis of Evictions

January 2019

Data Cleaning and Transformation

The dataset is visualized to identify possible problems with the data, after which cleansing and transformation is applied to address the problems identified. Visualizations are performed once more to verify that the cleansing and transformation had the desired effect.

The following features had missing values:

homicides_per_100k

pct_adult_smoking

pct_excessive_drinking

median_household_income

median_property_value

pct_low_birthweight

air_pollution_particulate_matter_value

motor_vehicle_crash_deaths_per_100k

pop_per_dentist

pop_per_primary_care_physician

The visualizations performed have conveyed that most the features above have a skewed distribution. Thus, the median value of the features was used to replace the missing values in the dataset to reduce the influence of outliers. The features homicides_per_100k, pct_excessive_drinking and pct_adult_smoking were removed from the dataset as they had a significant number of missing values.

Analysis of Evictions

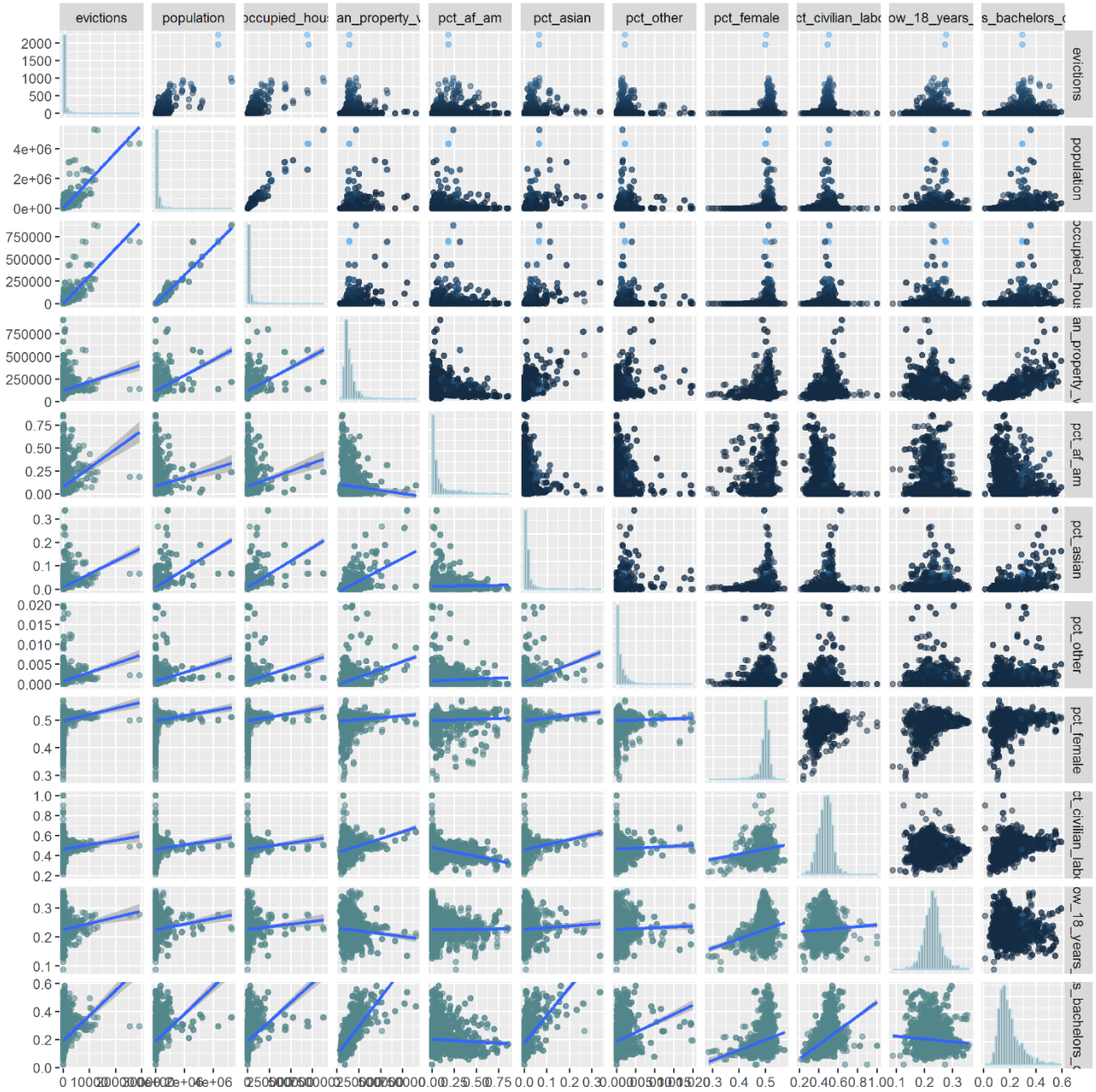
January 2019

Correlation and Apparent Relationships

After exploring the individual features, an attempt was made to identify relationships between features in the data – in particular, between **evictions** and the other features.

Numeric Relationships

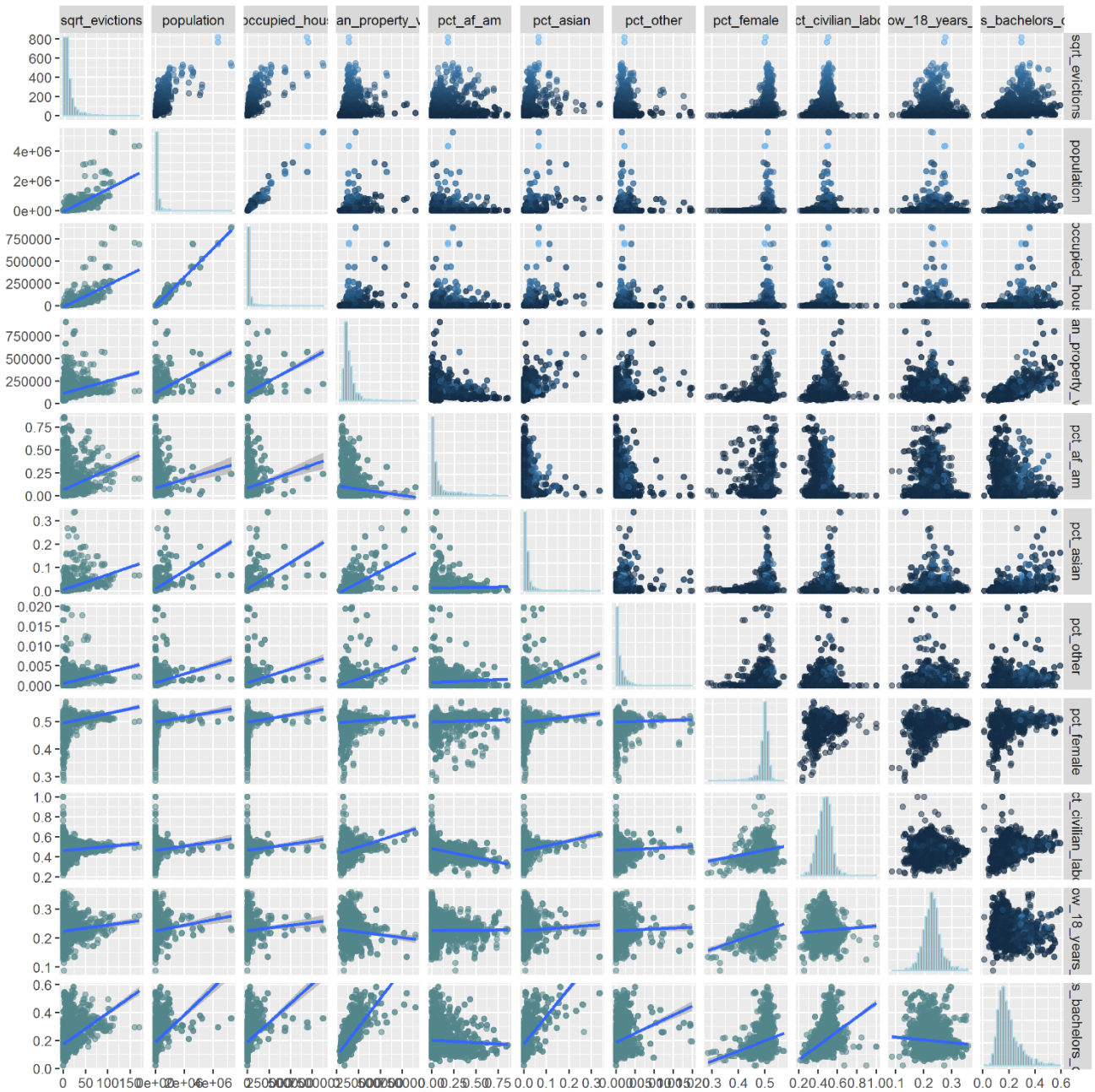
The following scatter-plot matrix was generated initially to compare numeric features with one another. The key features in this matrix are shown here:



Analysis of Evictions

January 2019

Viewing plots in the top row or the left-most column of this matrix shows an apparent relationship between evictions and other numeric features. As each of the features increase, so does evictions.
The evictions distribution can be seen to be significantly right-skewed in the bar chart on the top left of the matrix. This results in relationships between the numeric features and evictions that exhibit a nature that is not quite linear. In an attempt to improve the fit of the features to evictions, the square root value for evictions was calculated. The resulting scatter-plot matrix shows increased linearity in the relationships between square root of evictions and the other numeric features:



Analysis of Evictions

January 2019

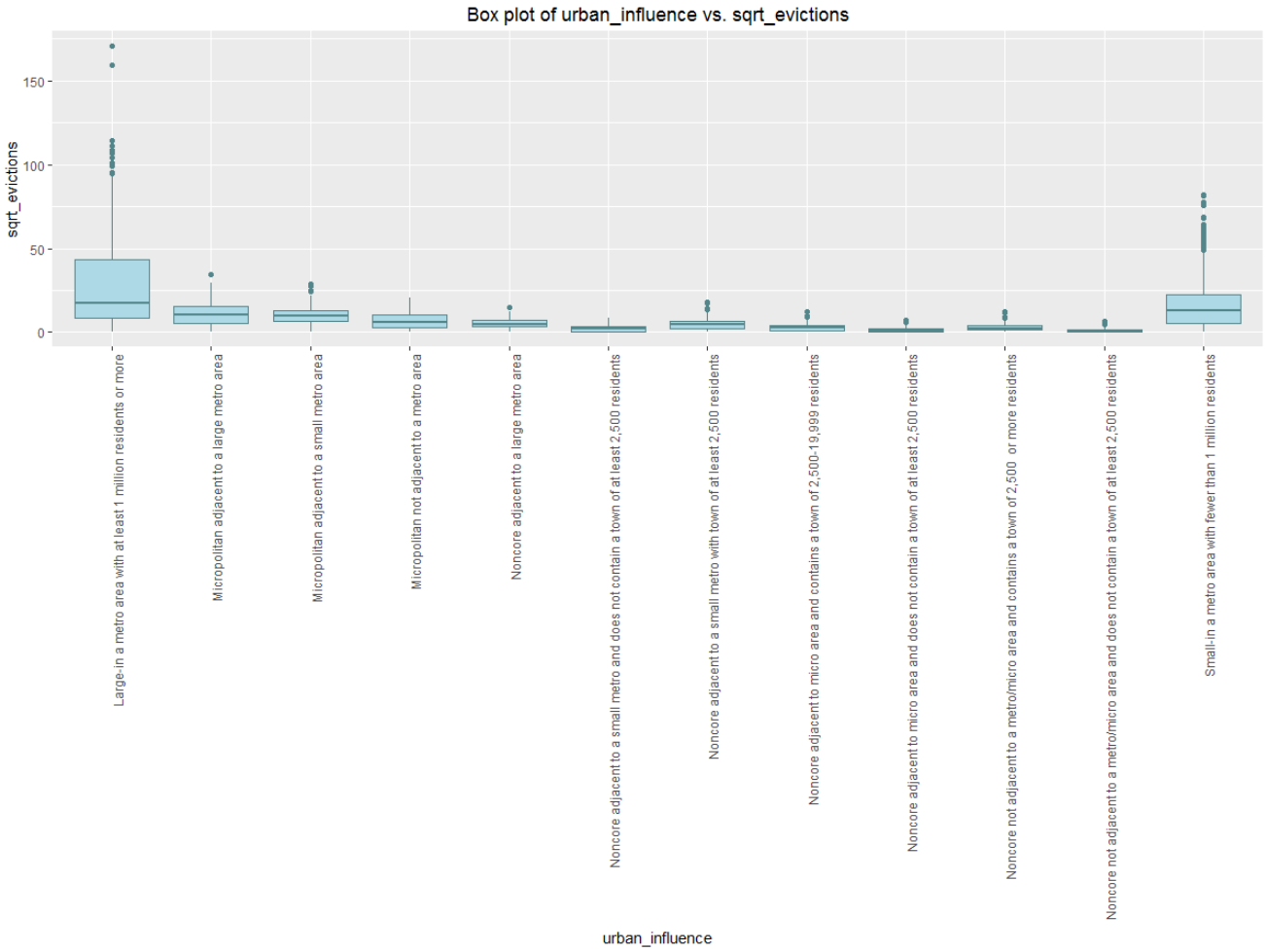
The correlation between the numeric columns was then calculated with the following results:

	sqrt_evictions	population	renter_occupied_households	median_property_value	pct_af_am	pct_asian	pct_other	pct_female	pct_civilian_labor	pct_below_18_years_of_age	pct_adults_bachelors_or_higher
sqrt_evictions	1.000000	0.757541	0.738662	0.289954	0.244607	0.429772	0.251155	0.230747	0.092306	0.104906	0.418409
population	0.757541	1.000000	0.974264	0.362970	0.105124	0.510165	0.201869	0.122986	0.094409	0.095099	0.365148
renter_occupied_households	0.738662	0.974264	1.000000	0.360436	0.123370	0.496126	0.207787	0.116302	0.089337	0.060910	0.342575
median_property_value	0.289954	0.362970	0.360436	1.000000	-0.072690	0.609933	0.337574	0.085903	0.287679	-0.091494	0.725775
pct_af_am	0.244607	0.105124	0.123370	-0.072690	1.000000	0.039322	0.078883	0.061940	-0.355371	0.009781	-0.061505
pct_asian	0.429772	0.510165	0.496126	0.609933	0.039322	1.000000	0.303772	0.095580	0.162945	0.045161	0.561068
pct_other	0.251155	0.201869	0.207787	0.337574	0.078883	0.303772	1.000000	0.031042	0.046828	0.032486	0.267214
pct_female	0.230747	0.122986	0.116302	0.085903	0.061940	0.095580	0.031042	1.000000	0.168741	0.238467	0.205670
pct_civilian_labor	0.092306	0.094409	0.089337	0.287679	-0.355371	0.162945	0.046828	0.168741	1.000000	0.065089	0.434801
pct_below_18_years_of_age	0.104906	0.095099	0.060910	-0.091494	0.009781	0.045161	0.032486	0.238467	0.065089	1.000000	-0.074634
pct_adults_bachelors_or_higher	0.418409	0.365148	0.342575	0.725775	-0.061505	0.561068	0.267214	0.205670	0.434801	-0.074634	1.000000

These correlations validate the plots by showing positive correlations between the numeric features and the square root of evictions.

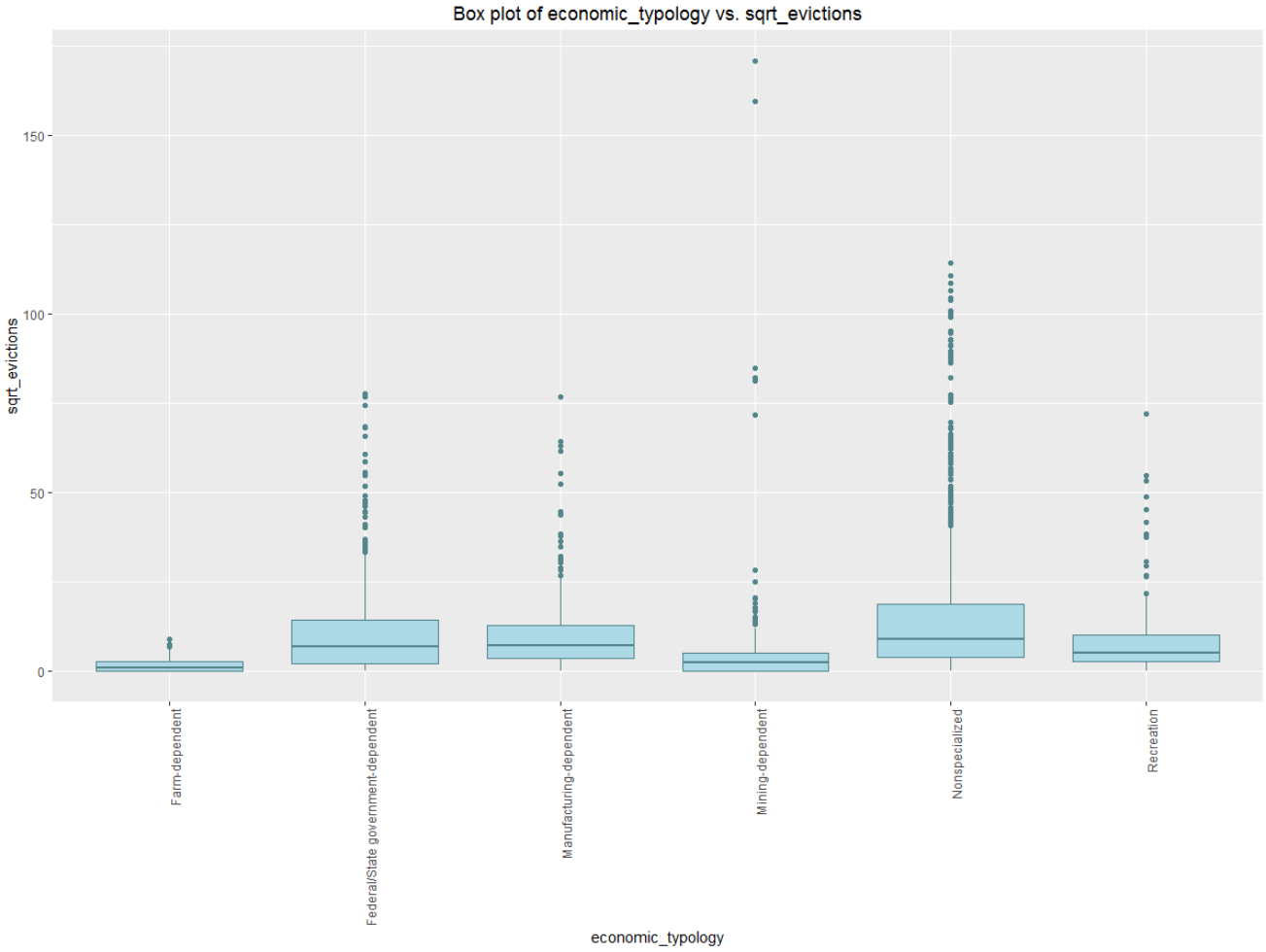
Categorical Relationships

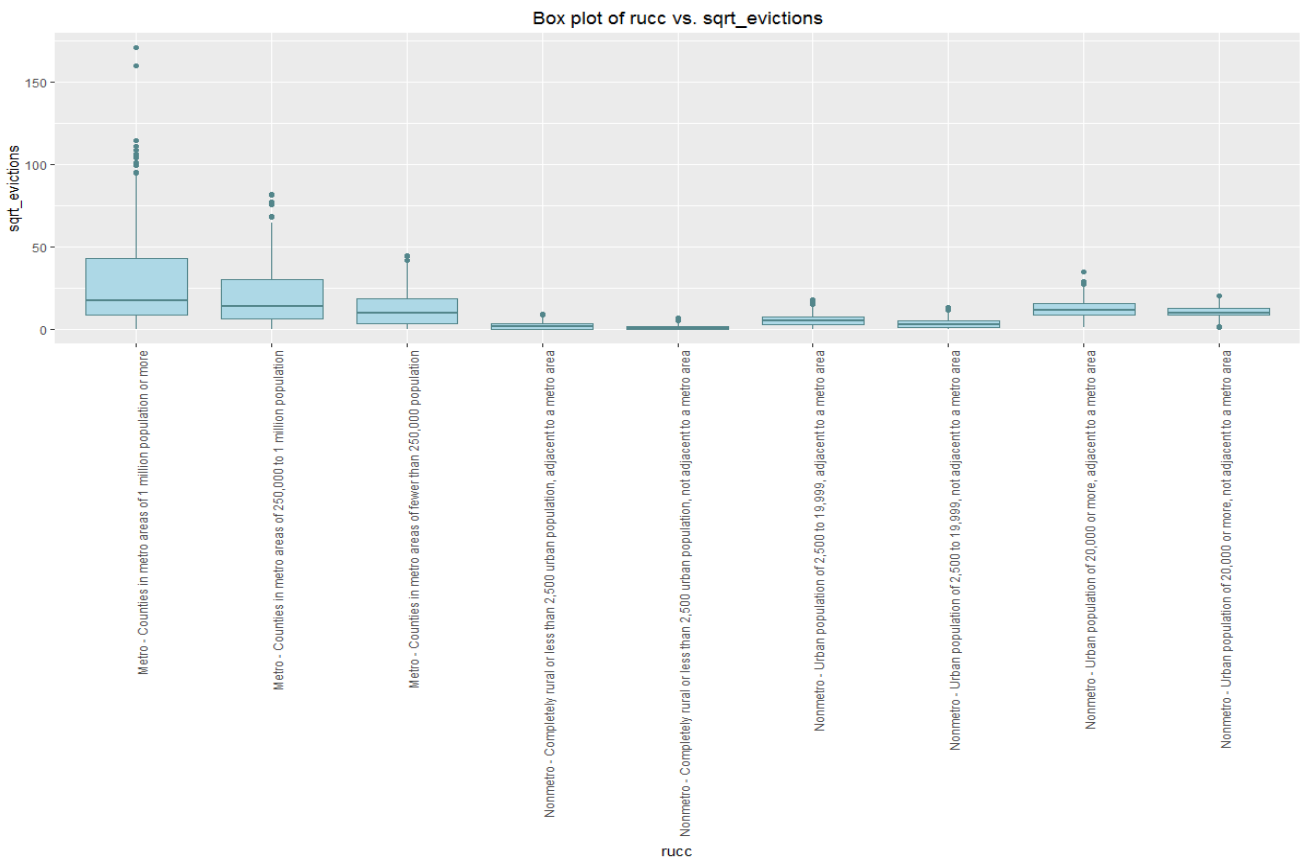
Having explored the relationship between evictions and numeric features, an attempt was made to discern any apparent relationship between categorical feature values and evictions. The following box-plots show the categorical columns that seem to exhibit a relationship with the square root of evictions:



Analysis of Evictions

January 2019





Analysis of Evictions

January 2019

The box plots show some clear differences in terms of the median and range of evictions values for different categorical features.
For example:

• urban_influence: In metro areas have higher rates of evictions, while the rest of the areas have fairly low rates of evictions.

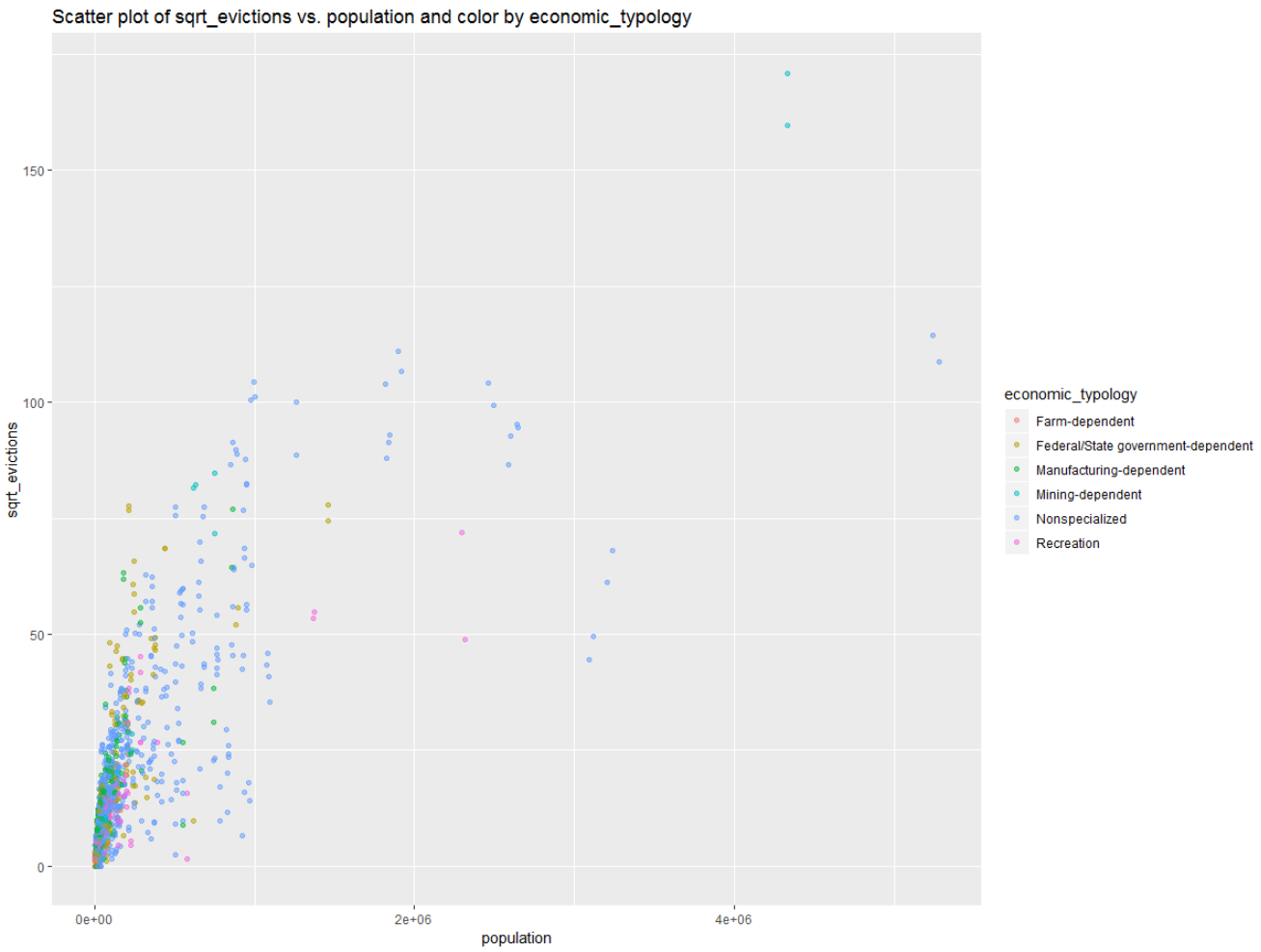
• rucc: Metro areas have higher rates of evictions, while non-metro areas have fairly low rates of evictions.

• economic_typology: Nonspecialized counties have the highest rate of evictions.

Multi-faceted Relationships

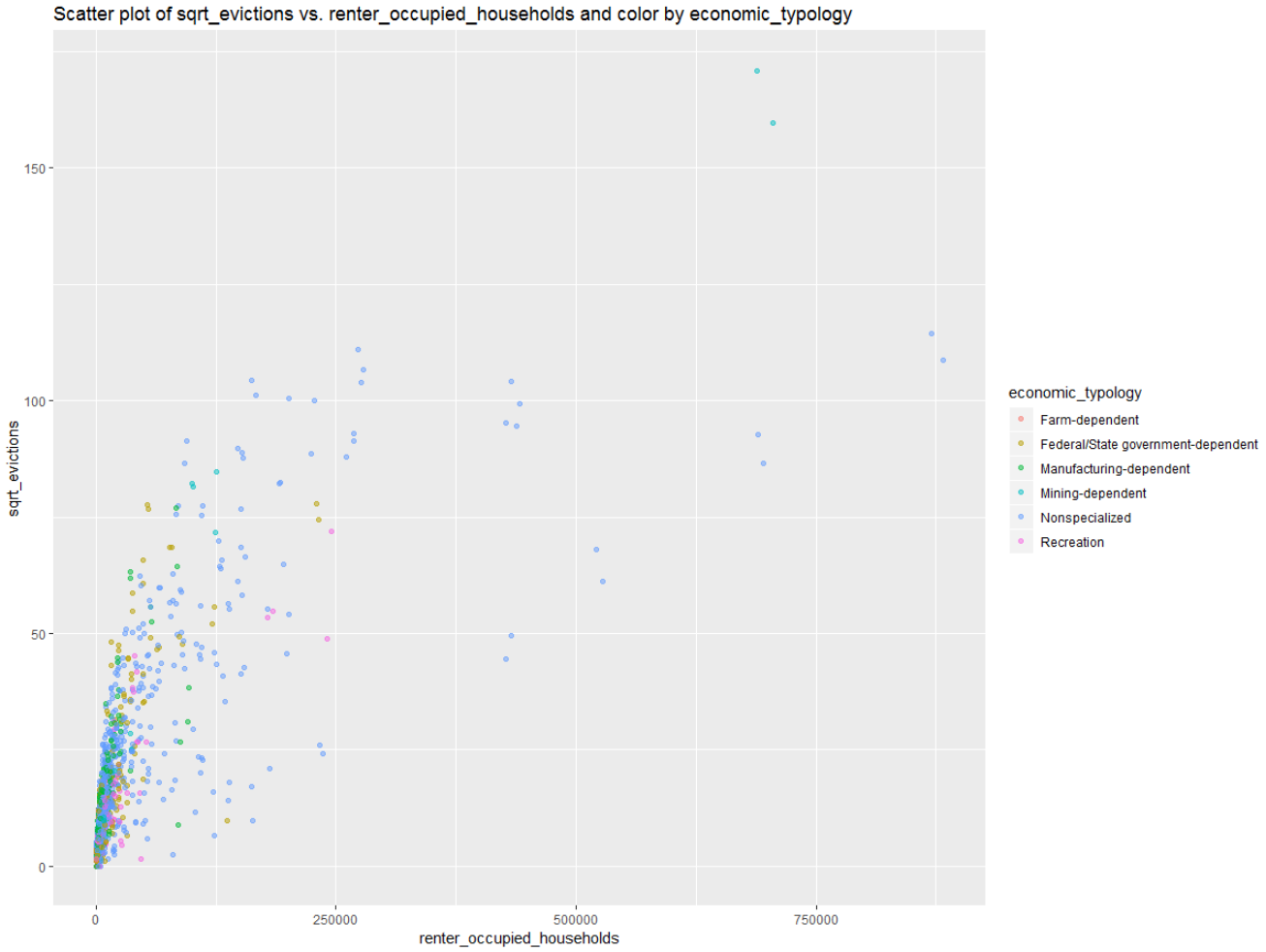
Apparent relationships between evictions and individual features are helpful in determining predictive heuristics. However, relationships are often more complex, and may only become apparent when multiple features are considered in combination with one another. To help identify these more complex relationships, some faceted plots were created.

The following plots show some interesting aspects of economic_typology. It can be seen from these plots that economic_typology types can be indicative of population and renter_occupied_households, both of which are typically predictive of evictions.



Analysis of Evictions

January 2019



From these plots, it can be seen that nonspecialized areas tend to have a higher population and renter_occupied_households than recreation, manufacturing-dependent and mining-dependent areas

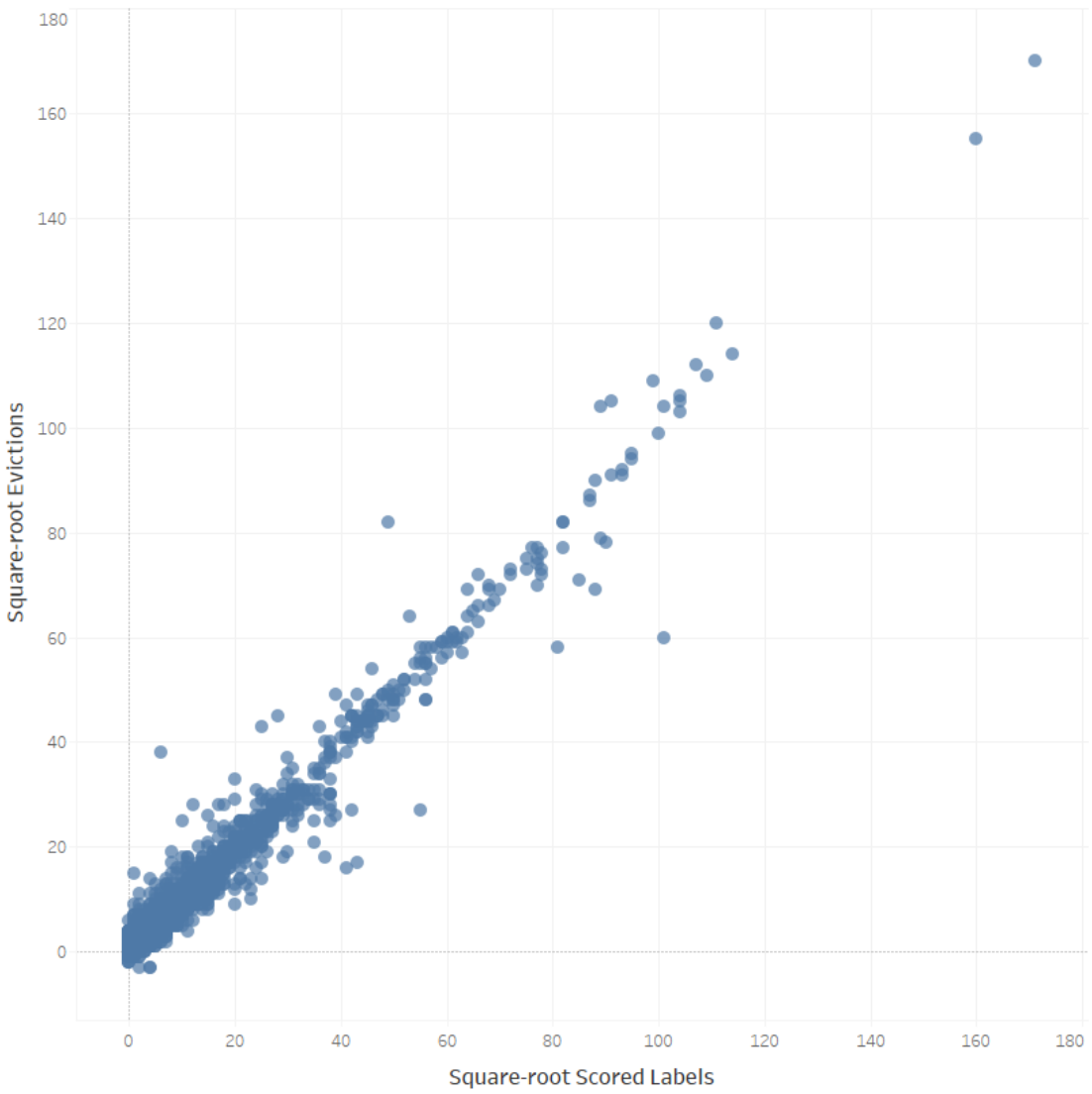
Analysis of Evictions

January 2019

Regression

A regression model to predict the actual number of evictions was created. Based on the apparent relationships identified when analyzing the data, a multiple linear regression model was created to predict the square-root value for evictions, from which the predicted evictions can be calculated.
The XGBoost algorithm with xgbTree parameters was used to perform a linear regression to improve execution speed and model performance with a 10-fold cross validation.

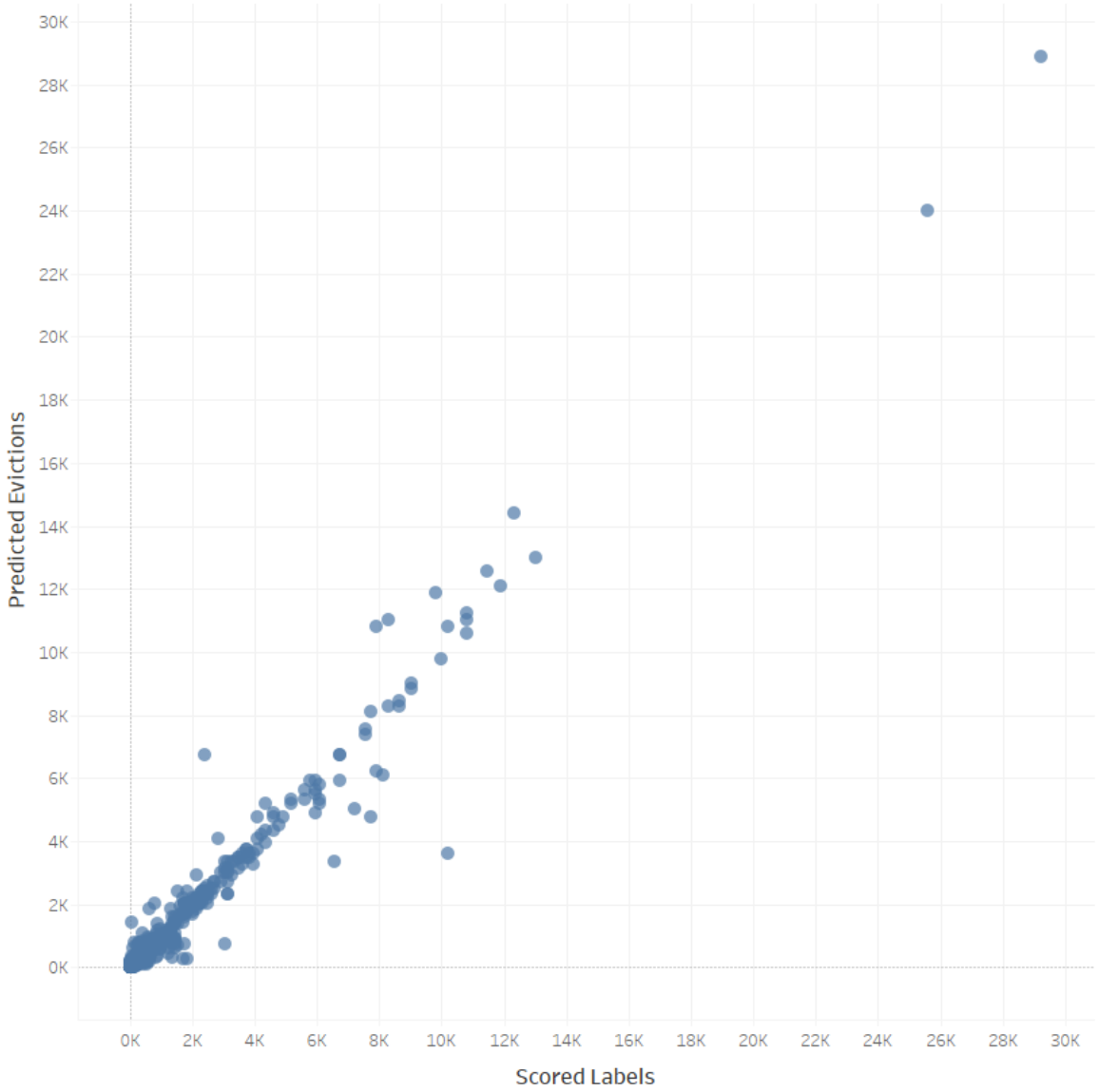
The model was trained with 70% of the data and tested with the remaining 30%. A scatter plot showing the predicted evictions and the actual evictions is shown below:



Analysis of Evictions

January 2019

This plot shows a clear linear relationship between predicted and actual values in the test dataset. The R-squared R^2 for the test results is 0.9234388, indicating that the model performs very well. When the predicted square-root evictions is converted back to its actual value, the following scatter plot shows the results.



Conclusion

This analysis has shown that the number of evictions can be confidently predicted from its characteristics. In particular, the population, renter_occupied_households, median_property_value, pct_af_am, pct_asian, pct_other, pct_female, pct_civilian_labor, pct_below_18_years_of_age, pct_adults_bachelors_or_higher and nonspecialized economic typology have a significant effect on the number of evictions.