Predicting County Level Rents

George Pinto 12/2019

Executive Summary

This document presents an analysis of data concerning county level rents from socioeconomic and demographic indicators gathered from the Economic Research Service (ERS), U.S. Department of Agriculture (USDA), https://www.ers.usda.gov/ and Eviction Lab. https://evictionlab.org/. The analysis is based on 1562 observations each containing information on 43 variables. After exploring the data by calculating summary and descriptive statistics and by creating visualizations of the data, it was determined that a regression model could be created to predict the median gross rent, and that the results could be further improved by using a support vector machine algorithm.

The following features were found to be the most informative for the purpose of this analysis (creating a generalized model predicting gross rent):

- **State and County Codes** –These location variables provide useful variance between median gross rent values and assist with the analysis and pattern recognition, but individual state and county codes are not identified.
- **Percentage Adult Bachelors or Higher** – Bachelor's degree or higher tends to increase as the median gross rent increases. This is a strong correlation.
- **Motor Vehicle Crash Deaths per 100K** – Bachelor's degree or higher decreases as motor vehicle crash deaths increase. This is also the effect on median gross rent. This is a strong negative correlation.
- **Percentage of Excessive Drinking** – Bachelor's degree or higher increases as percentage of excessive drinking increases, this occurs as population increases. Median gross rent also increases. This is a significant correlation
- **Percentage of Adult Smoking** – Bachelor's degree or higher decreases as percentage of adult smoking increases. Median gross rate also decreases.
- **Obesity, Diabetes, Physical Inactivity** – Obesity, diabetes, and physical inactivity all have strong negative correlation to median gross rent values. They have strong correlations to each other, and they are strongly positively correlated to percentage high school diploma.
- **Percentage High School Diploma** – Bachelor's or higher decreases as percentage high school diploma increases, median gross rate also decreases. This is a significant negative correlation.

- **Evictions** – Bachelor's degree or higher increases as evictions increase, this occurs as population increases. Median gross rent also increases. This is a significant correlation

- **Population** – Bachelor's degree or higher increases as the population increases. Median gross rent also increases. This is a significant correlation.

- **RUCC, Urban Influence and Economic Typology** – In order to maintain this as an executive summary we'll need to just give a general overview of these variables that were substantially important in this analysis. Basically, they segregate regions into metropolitan, rural and urban areas, amount of urban influence, and specific economic types. The main insight is that these categories exert unique influences on median gross rent values depending on their combinations within individual counties and states and are crucial for understanding these values and predicting them. RUCC is an acronym for Rural-Urban Continuum Codes.

Initial Data Exploration

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

Individual Feature Statistics

Individual summary feature statistics for minimum, maximum, mean, median, standard deviation and distinct count were calculated for numeric columns and the results taken from 1562 observations shown here (variables that could create bias towards a group or groups were removed per ethical guidelines) also note the lower counts for some variables which are nulls, empty spaces or nan values, whichever terminology is preferred – those have to be filled in a way that doesn't affect the stats negatively:

```
row_id
count    1562.000000
mean      780.500000
std       451.054875
min         0.000000
25%       390.250000
50%       780.500000
75%      1170.750000
max      1561.000000
Name: row_id, dtype: float64

county_code
count        1562
unique       1562
top       bcbd98f
freq            1
Name: county_code, dtype: object

state
count        1562
unique         50
top       1b0d913
freq          131
Name: state, dtype: object
```

population

```
count    1.562000e+03
mean     1.083407e+05
std      3.745229e+05
min      2.690000e+02
25%      1.045275e+04
50%      2.528200e+04
75%      6.836150e+04
max      1.002029e+07
Name: population, dtype: float64
```

renter_occupied_households

```
count    1.562000e+03
mean     1.490462e+04
std      6.255947e+04
min      6.400000e+01
25%      1.078250e+03
50%      2.754000e+03
75%      7.987000e+03
max      1.760277e+06
Name: renter_occupied_households, dtype: float64
```

pct_renter_occupied

```
count    1562.000000
mean       28.525570
std         8.121601
min         7.279000
25%        22.874000
50%        27.199500
75%        32.341500
max        73.008000
Name: pct_renter_occupied, dtype: float64
```

evictions

```
count     1235.000000
mean       397.410526
std       1522.801074
min         -1.000000
25%          3.000000
50%         27.000000
75%        162.500000
max      29251.000000
Name: evictions, dtype: float64
```

rent_burden

```
count    1562.000000
mean       28.537526
std         4.670467
min         9.909000
25%        25.869000
50%        28.768000
75%        31.307250
max        49.665000
Name: rent_burden, dtype: float64
```

poverty_rate
```
count    1562.000000
mean       12.182780
std         5.783889
min         0.000000
25%         7.991000
50%        11.173500
75%        15.018500
max        38.792000
Name: poverty_rate, dtype: float64
```

rucc
```
count                                                  1562
unique                                                    9
top       Nonmetro - Urban population of 2,500 to 19,999...
freq                                                    301
Name: rucc, dtype: object
```

urban_influence
```
count                                                  1562
unique                                                   12
top       Small-in a metro area with fewer than 1 millio...
freq                                                    358
Name: urban_influence, dtype: object
```

economic_typology
```
count              1562
unique                6
top       Nonspecialized
freq                631
Name: economic_typology, dtype: object
```

pct_civilian_labor
```
count    1562.000000
mean        0.470535
std         0.070936
min         0.186000
25%         0.425000
50%         0.470500
75%         0.515000
max         0.996000
Name: pct_civilian_labor, dtype: float64
```

pct_unemployment
```
count    1562.000000
mean        0.062551
std         0.022615
min         0.012000
25%         0.046000
50%         0.061000
75%         0.076000
max         0.242000
Name: pct_unemployment, dtype: float64
```

pct_uninsured_adults
```
count    1560.000000
mean        0.220037
std         0.067500
min         0.053000
25%         0.171750
50%         0.216000
75%         0.265000
max         0.520000
Name: pct_uninsured_adults, dtype: float64
```

pct_uninsured_children
```
count    1560.000000
mean        0.088844
std         0.041480
min         0.018000
25%         0.059000
50%         0.079000
75%         0.109000
max         0.327000
Name: pct_uninsured_children, dtype: float64
```

pct_adult_obesity
```
count    1560.000000
mean        0.304546
std         0.043550
min         0.133000
25%         0.283000
50%         0.306000
75%         0.331250
max         0.474000
Name: pct_adult_obesity, dtype: float64
```

pct_adult_smoking
```
count    1344.000000
mean        0.211682
std         0.064045
min         0.031000
25%         0.169000
50%         0.206000
75%         0.249000
max         0.513000
Name: pct_adult_smoking, dtype: float64
```

pct_diabetes
```
count    1560.000000
mean        0.106566
std         0.022521
min         0.033000
25%         0.092000
50%         0.105000
75%         0.122000
max         0.180000
```

```
Name: pct_diabetes, dtype: float64
```

pct_low_birthweight

```
count    1446.000000
mean        0.083326
std         0.021239
min         0.030000
25%         0.070000
50%         0.080000
75%         0.091000
max         0.182000
Name: pct_low_birthweight, dtype: float64
```

pct_excessive_drinking

```
count    1100.000000
mean        0.164818
std         0.051483
min         0.032000
25%         0.129000
50%         0.163000
75%         0.196000
max         0.419000
Name: pct_excessive_drinking, dtype: float64
```

pct_physical_inactivity

```
count    1560.000000
mean        0.276981
std         0.053098
min         0.104000
25%         0.243000
50%         0.281000
75%         0.312000
max         0.446000
Name: pct_physical_inactivity, dtype: float64
```

air_pollution_particulate_matter_value

```
count    1542.000000
mean       11.637336
std         1.534144
min         7.209413
25%        10.432910
50%        11.907300
75%        12.885680
max        14.992477
Name: air_pollution_particulate_matter_value, dtype: float64
```

homicides_per_100k

```
count    613.000000
mean       5.752414
std        4.297808
min       -0.080000
25%        2.710000
50%        4.540000
75%        7.600000
```

```
max       26.920000
Name: homicides_per_100k, dtype: float64
```

motor_vehicle_crash_deaths_per_100k

```
count    1372.000000
mean       21.715153
std        10.721369
min         3.140000
25%        14.047500
50%        20.285000
75%        27.072500
max       110.450000
Name: motor_vehicle_crash_deaths_per_100k, dtype: float64
```

heart_disease_mortality_per_100k

```
count    1562.000000
mean      275.482714
std        57.827540
min        76.000000
25%       233.000000
50%       270.000000
75%       311.000000
max       511.000000
Name: heart_disease_mortality_per_100k, dtype: float64
```

pop_per_dentist

```
count     1447.000000
mean      3421.828611
std       2538.670834
min        340.000000
25%       1859.000000
50%       2730.000000
75%       4064.500000
max      25169.000000
Name: pop_per_dentist, dtype: float64
```

pop_per_primary_care_physician

```
count     1448.000000
mean      2508.303867
std       1960.312344
min        279.000000
25%       1389.750000
50%       1969.500000
75%       2840.000000
max      16740.000000
Name: pop_per_primary_care_physician, dtype: float64
```

pct_below_18_years_of_age

```
count    1560.000000
mean        0.228672
std         0.034732
min         0.082000
25%         0.208000
```

```
50%         0.227000
75%         0.246000
max         0.415000
Name: pct_below_18_years_of_age, dtype: float64
```

pct_aged_65_years_and_older

```
count    1560.000000
mean        0.167707
std         0.044555
min         0.036000
25%         0.140000
50%         0.164000
75%         0.191000
max         0.488000
Name: pct_aged_65_years_and_older, dtype: float64
```

pct_adults_less_than_a_high_school_diploma

```
count    1562.000000
mean        0.145666
std         0.067483
min         0.019000
25%         0.094226
50%         0.129388
75%         0.187719
max         0.535750
Name: pct_adults_less_than_a_high_school_diploma, dtype: float64
```

pct_adults_with_high_school_diploma

```
count    1562.000000
mean        0.346271
std         0.071152
min         0.074297
25%         0.300601
50%         0.352176
75%         0.396556
max         0.535536
Name: pct_adults_with_high_school_diploma, dtype: float64
```

pct_adults_with_some_college

```
count    1562.000000
mean        0.303030
std         0.052236
min         0.114458
25%         0.269461
50%         0.303303
75%         0.337672
max         0.477341
Name: pct_adults_with_some_college, dtype: float64
```

pct_adults_bachelors_or_higher

```
count    1562.000000
mean        0.205033
std         0.092001
min         0.064128
```

```
25%          0.142142
50%          0.182365
75%          0.241234
max          0.788153
Name: pct_adults_bachelors_or_higher, dtype: float64
```

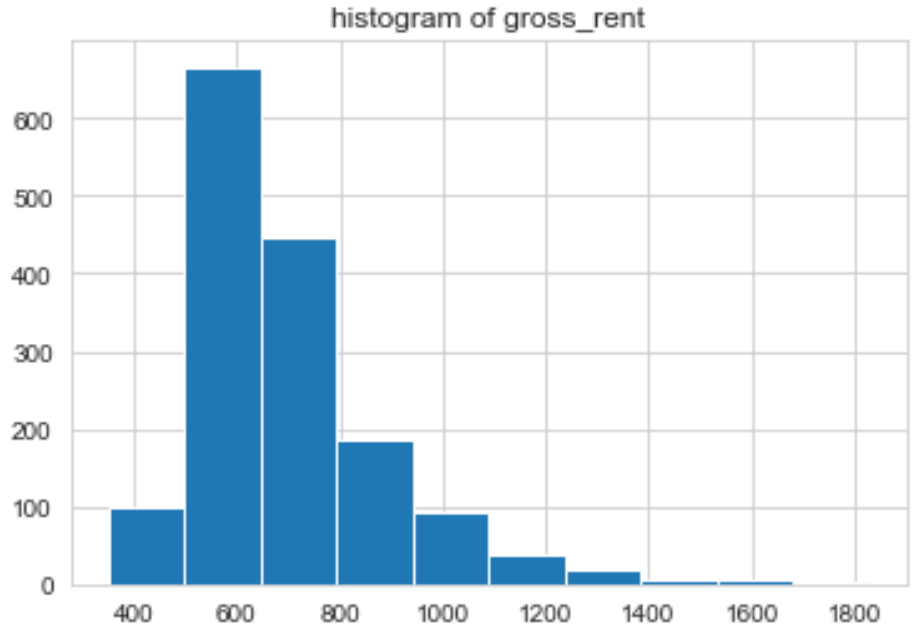
birth_rate_per_1k

```
count    1562.000000
mean       11.621356
std         2.756009
min         3.654080
25%        10.014818
50%        11.435441
75%        12.940235
max        29.034900
Name: birth_rate_per_1k, dtype: float64
```

death_rate_per_1k

```
count    1562.000000
mean       10.415138
std         2.772070
min         0.961076
25%         8.613691
50%        10.396898
75%        12.250655
max        24.281150
Name: death_rate_per_1k, dtype: float64
```

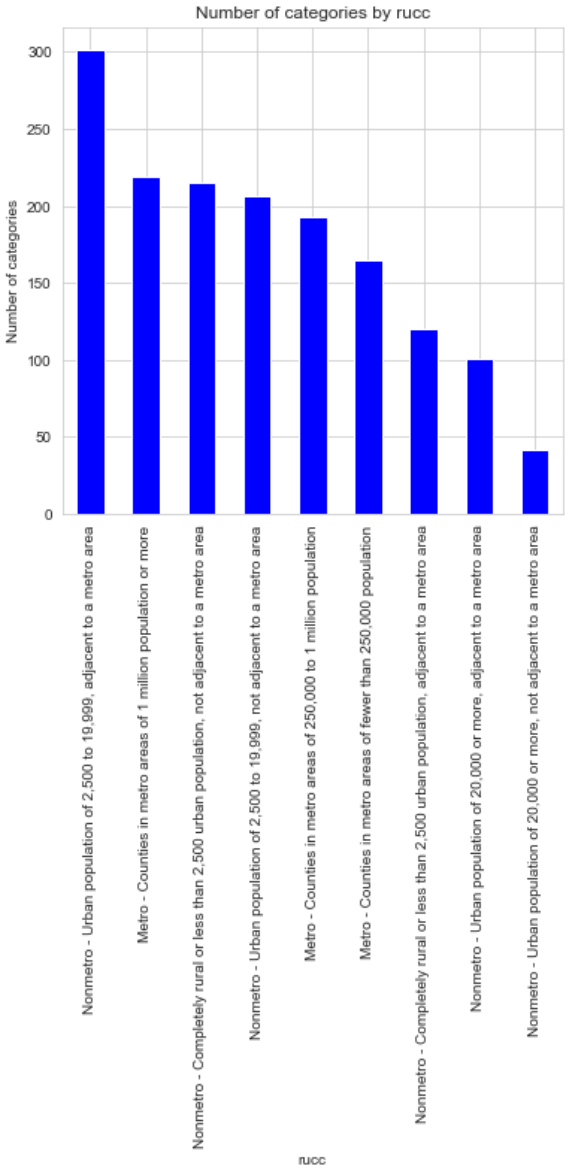
Since we are interested in gross rent, it was noted that the distribution was right skewed, with a mean around 701 and a median of 650, and outliers going up to a bit higher than 1800, this significantly wide standard deviation tells us there is considerable variance between gross rent values across counties:

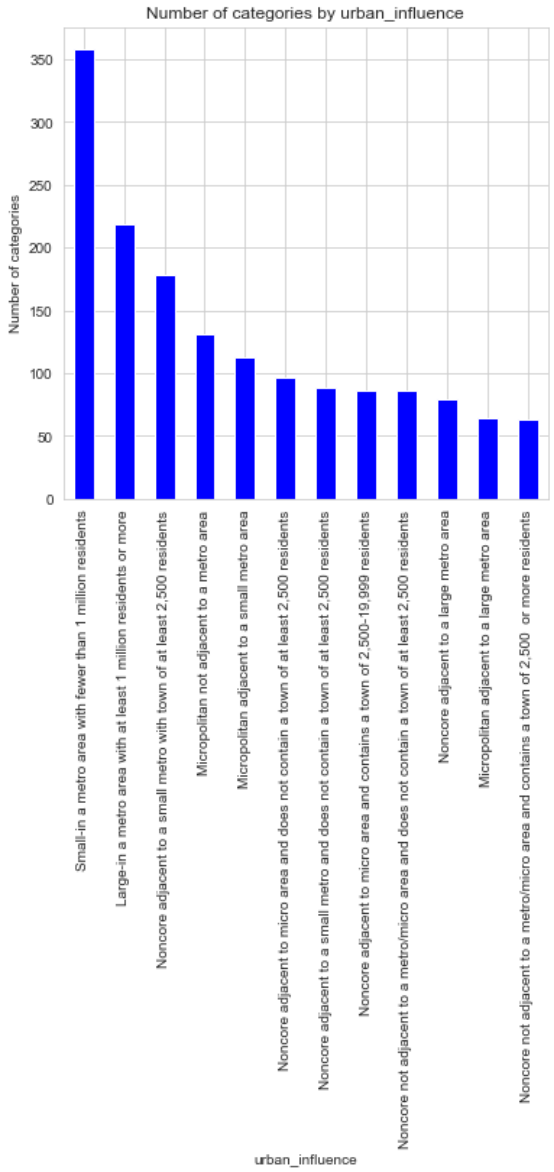


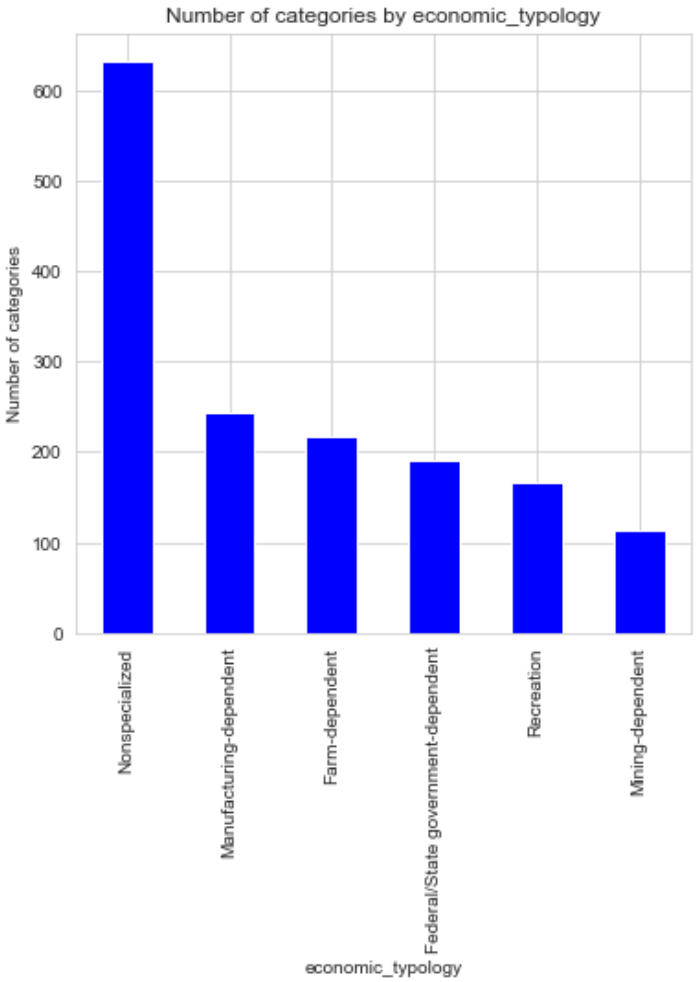
gross_rent

```
count    1562.000000
mean      701.142125
std       192.883110
min       351.000000
25%       578.000000
50%       650.000000
75%       773.750000
max      1827.000000
Name: gross_rent, dtype: float64
```

Let's look at how the distribution of categorical variables looked, recall these were RUCC, Urban Influence and Economic Typology:







This is how the actual numbers of each categorical type were distributed:

For column rucc

```
Nonmetro - Urban population of 2,500 to 19,999, adjacent to a metro area:
301
Metro - Counties in metro areas of 1 million population or more:
219
Nonmetro - Completely rural or less than 2,500 urban population, not adjacent
to a metro area:
215
Nonmetro - Urban population of 2,500 to 19,999, not adjacent to a metro area:
206
Metro - Counties in metro areas of 250,000 to 1 million population:
193
Metro - Counties in metro areas of fewer than 250,000 population:
165
Nonmetro - Completely rural or less than 2,500 urban population, adjacent to a
metro area:
120
Nonmetro - Urban population of 20,000 or more, adjacent to a metro area:
101
```

Nonmetro - Urban population of 20,000 or more, not adjacent to a metro area:
42
Name: rucc, dtype: int64

For column urban_influence
Small-in a metro area with fewer than 1 million residents:
358
Large-in a metro area with at least 1 million residents or more:
219
Noncore adjacent to a small metro with town of at least 2,500 residents:
178
Micropolitan not adjacent to a metro area:
131
Micropolitan adjacent to a small metro area:
113
Noncore adjacent to micro area and does not contain a town of at least 2,500
Residents:
97
Noncore adjacent to a small metro and does not contain a town of at least 2,500
residents:
88
Noncore adjacent to micro area and contains a town of 2,500-19,999 residents:
86
Noncore not adjacent to a metro/micro area and does not contain a town of at
least 2,500 residents:
86
Noncore adjacent to a large metro area:
79
Micropolitan adjacent to a large metro area:
64
Noncore not adjacent to a metro/micro area and contains a town of 2,500 or more
residents:
63
Name: urban_influence, dtype: int64

For column economic_typology
Nonspecialized:
631
Manufacturing-dependent:
244
Farm-dependent:
217
Federal/State government-dependent:
191
Recreation:
166
Mining-dependent:
113
Name: economic_typology, dtype: int64

The focus that was maintained throughout the analysis was to not eliminate a subgroup while cleaning the data. A good sense of the magnitude of each category was necessary, since the analysis of the categorical data was very detailed, especially when considering 43 variables, I will give an overview without covering all the details to give a strong intuition of what was kept in mind during the analysis :

- **RUCC:** categories range from 300 to 42, so the largest category (Nonmetro - Urban population of 2,500 to 19,999, adjacent to a metro area) is more than 5 times larger than the smallest one (Nonmetro - Urban population of 20,000 or more, not adjacent to a metro area). Each category has a distinct median and standard deviation that provides unique information on the value of median gross rent

- **Urban Influence:** categories range from 358 to 63, so the largest category (Small-in a metro area with fewer than 1 million residents) is more than 5 times larger than the smallest one (Noncore not adjacent to a metro/micro area and contains a town of 2,500 or more residents). Each category has a distinct median and standard deviation that provides unique information on the value of median gross rent

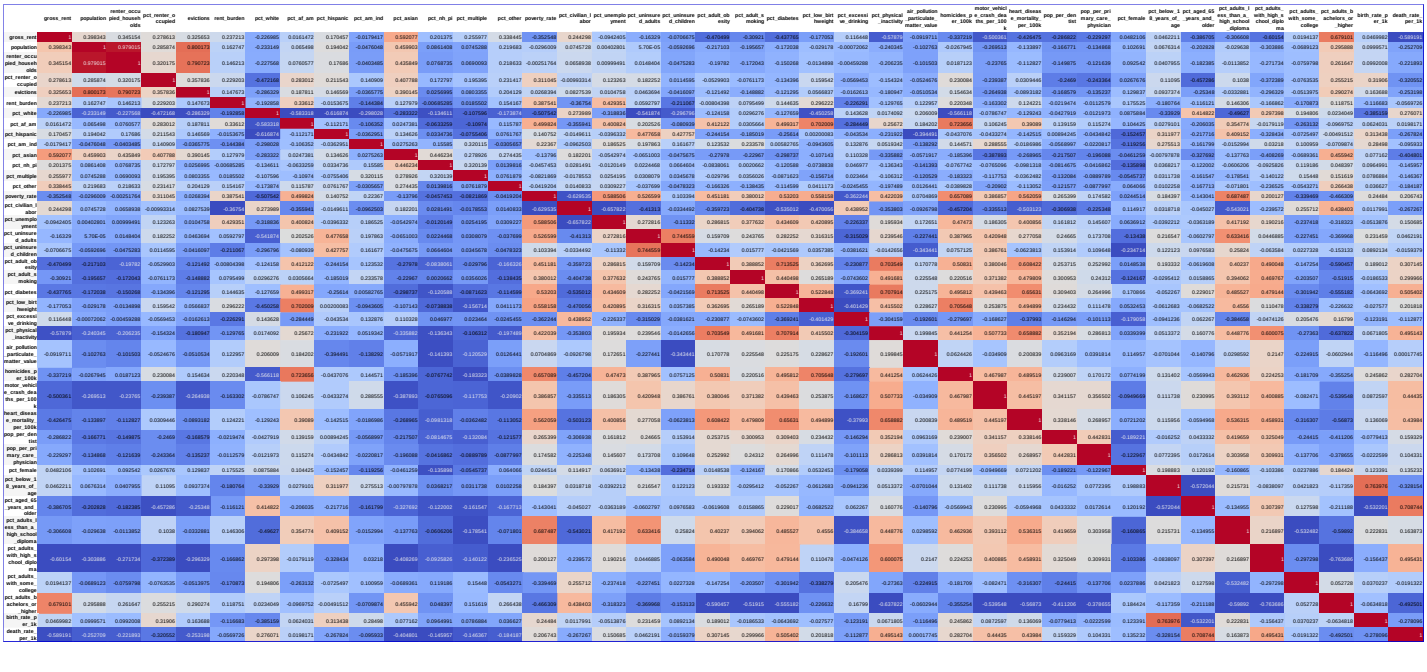
- **Economic Typology:** categories range from 631 to 113, so the largest category (Nonspecialized) is more than 5 times larger than the smallest one (Mining). Each category has a distinct median and standard deviation that provides unique information on the value of median gross rent

Correlation and Apparent Relationships

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

Numeric features

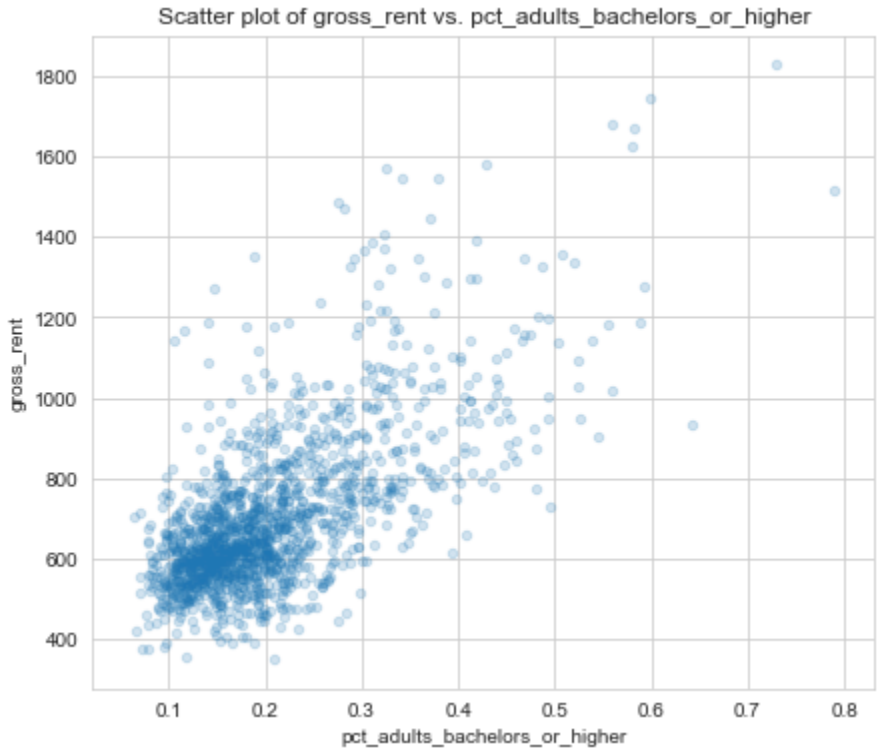
Here is our correlation matrix. There are so many features that it's very hard to see the details, but it is basically a cool/warm color map so that the strongest positive correlations hint towards red and the strongest negative correlations hint towards dark blue:



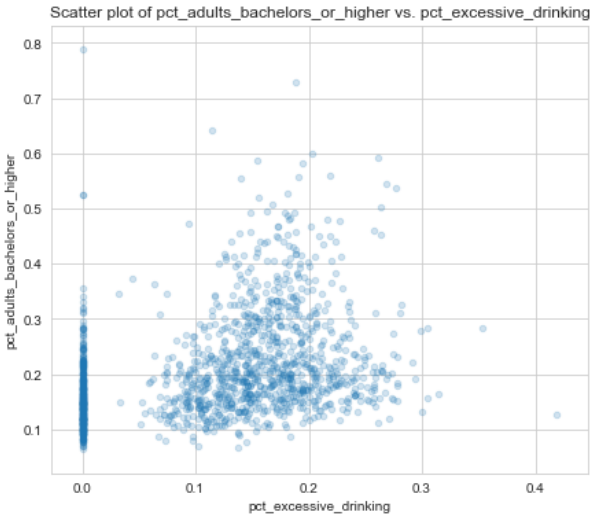
Note that the strongest positive correlation to the median gross rent is percentage adults with a bachelors or higher at 0.679101 and the strongest negative correlation is percentage adults with high school diploma at -0.60154. I think gross rent will be predicted very well by examining these positive and negative relationships from education and the variables affecting them. The variables affecting them will turn out to be the other strongly correlated variables of gross rent that will make strong predictors:

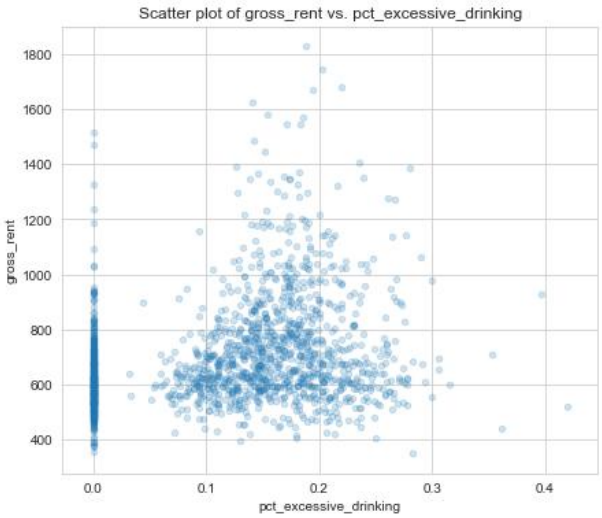
pct_adults_with_high_school_diploma	-0.60154	-0.303886
pct_adults_with_some_college	0.0194137	-0.0689123
pct_adults_bachelors_or_higher	0.679101	0.295888

Let’s start with the highest correlation to median gross rent, percent adults’ bachelors or higher:

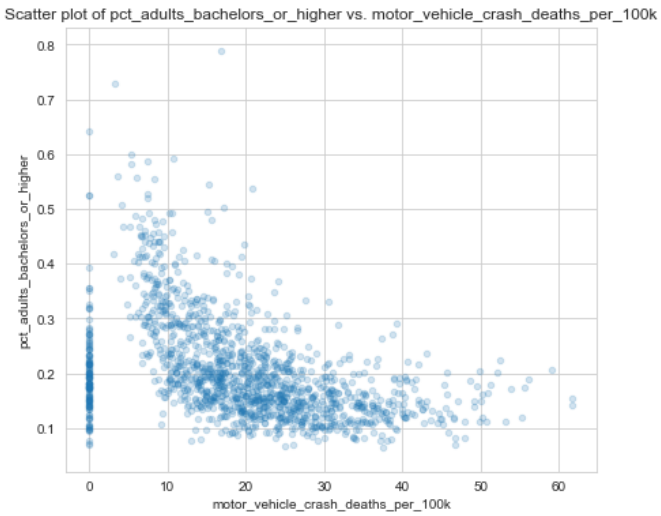


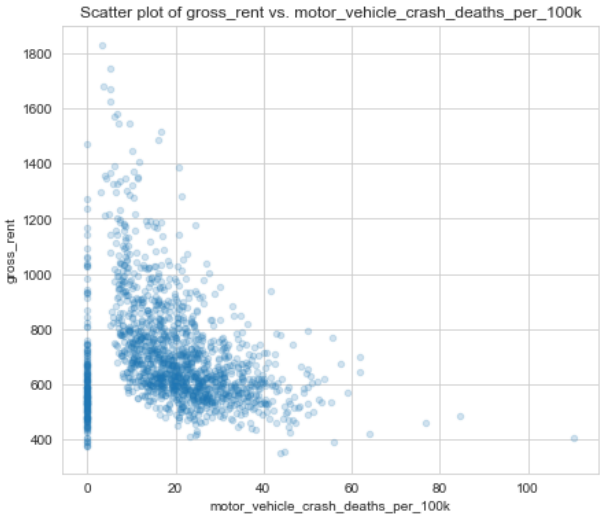
Percent adults bachelors or higher is positively correlated to percentage excessive drinking which aligns with median gross rent (note the noise in the data, from when empty spaces where filled, it was deemed appropriate to use zeros; a median or mean fill would make the median of the variable more prone to error and we need the ‘right’ median even if we made percentage excessive drinking distribution bimodal. Also note the density of the error at zero:



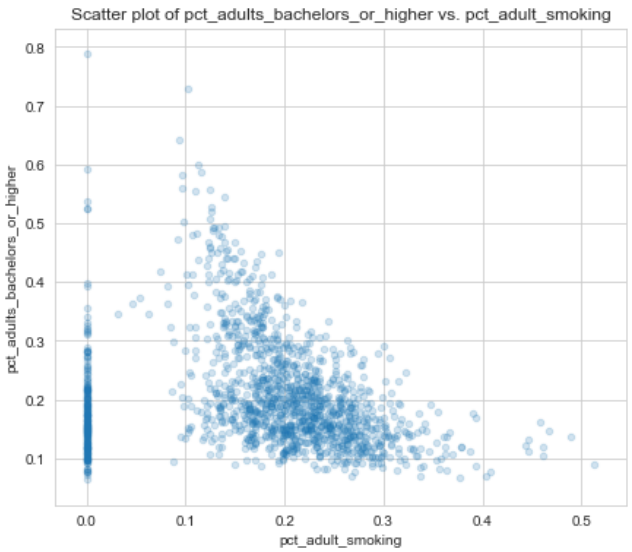


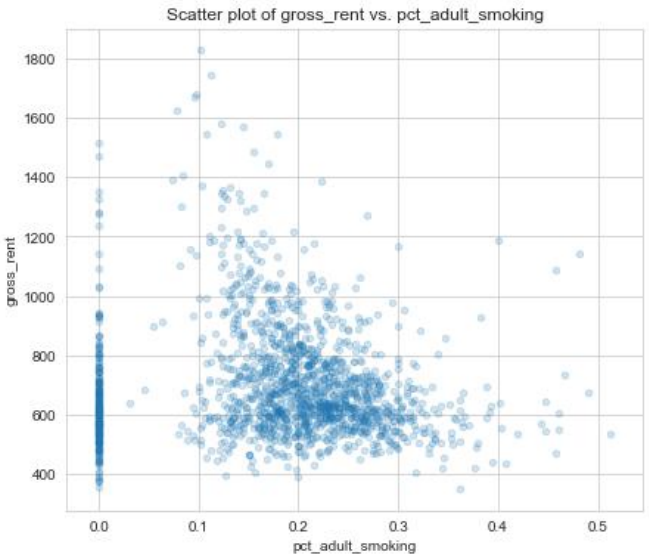
Percentage adults bachelors or higher is negatively correlated to percentage motor vehicle crash deaths, which is also the relationship with median gross rent (note the noise in the data, from when empty spaces where filled, it was deemed appropriate to use zeros; a median or mean fill would make the median of the variable more prone to error and we need the ‘right’ median even if we made percentage motor vehicle crash deaths distribution bimodal. Also note the density of the error at zero:



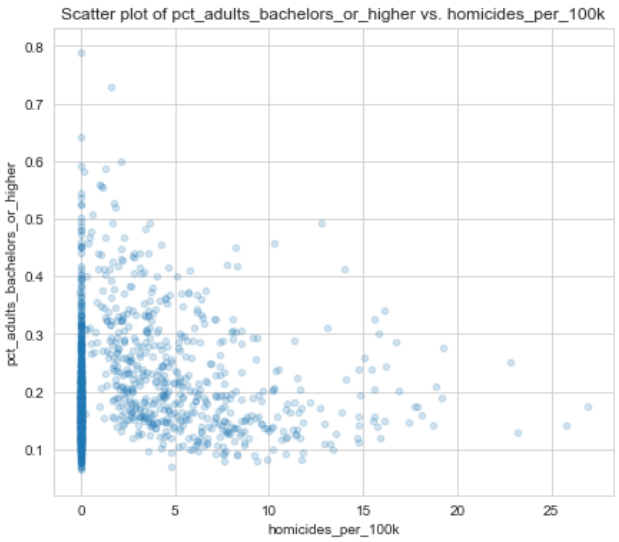


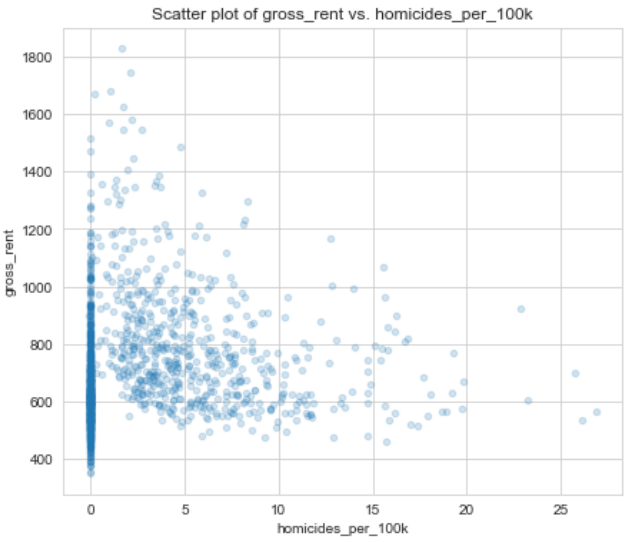
Percentage adults bachelors or higher is negatively correlated to percentage adult smoking per 100k, this is also the relationship with median gross rent (note the noise in the data, from when empty spaces where filled, it was deemed appropriate to use zeros; a median or mean fill would make the median of the variable more prone to error and we need the ‘right’ median even if we made the percentage adult smoking distribution bimodal:



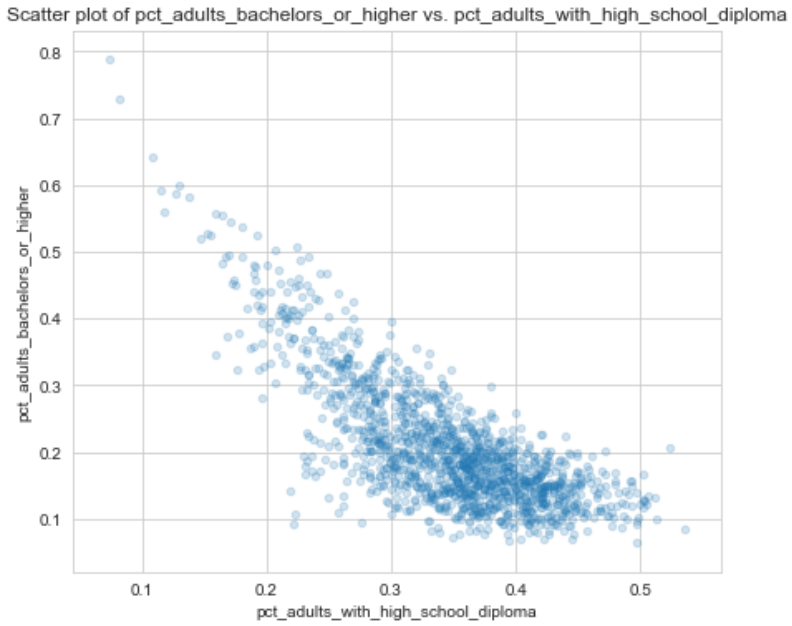


Percentage adults bachelors or higher is negatively correlated to homicides per 100k, this is also the relationship with median gross rent. Note the noise in the data, from when empty spaces where filled, it was deemed appropriate to use zeros; a median or mean fill would make the median of the variable more prone to error and we need the 'right' median even if we made the homicides per 100k distribution bimodal:

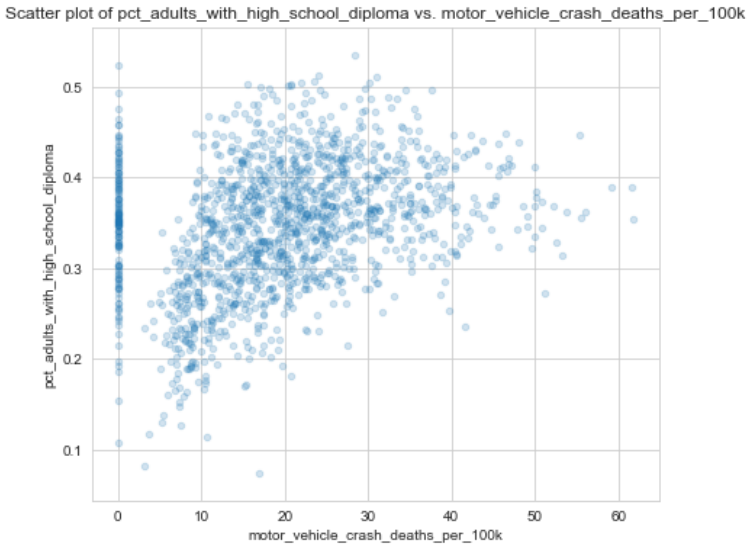




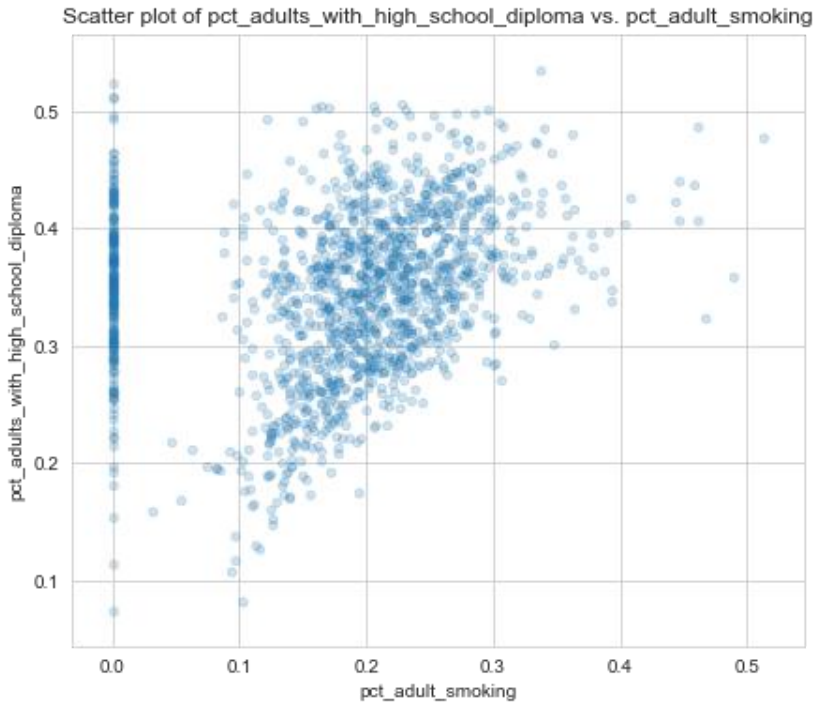
Percentage bachelors or higher is also very strongly negatively correlated to percentage adults with a high school diploma as is median gross rent:



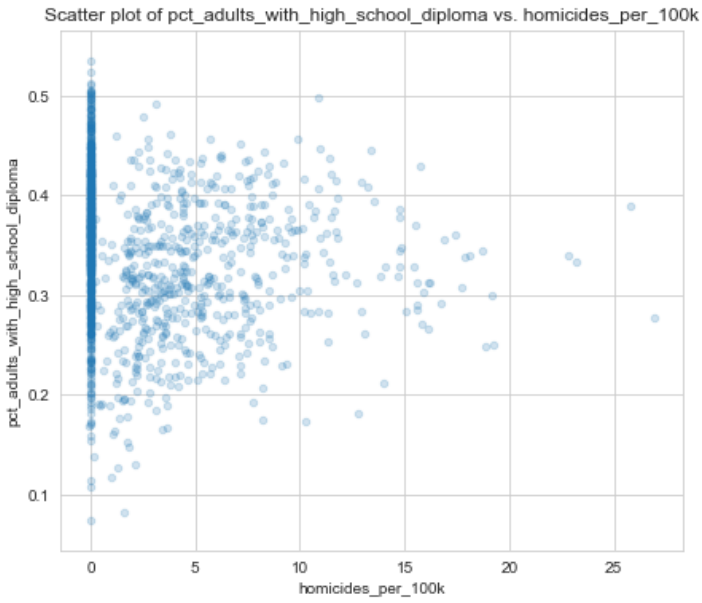
Percentage adults with a high school diploma have a positive correlation with motor vehicle crash deaths. Note the same noise in the data we mentioned before:



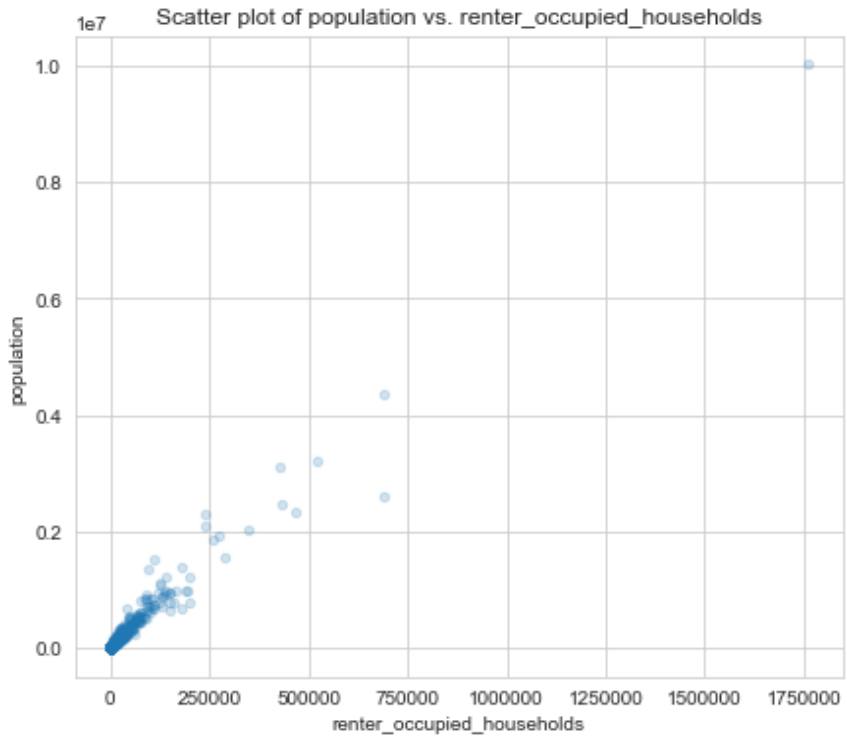
Percentage adults with a high school diploma have a positive correlation with percentage adult smoking. Note the same noise in the data we mentioned before:



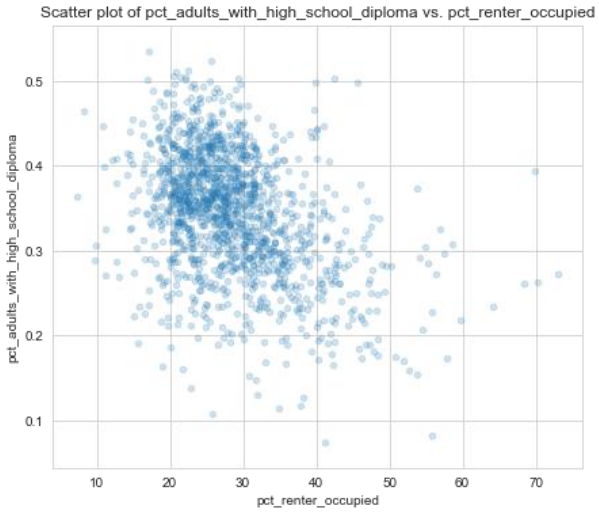
Percent adults with high school diploma have a positive correlation with homicides per 100 k, note the same noise in the data we mentioned before:

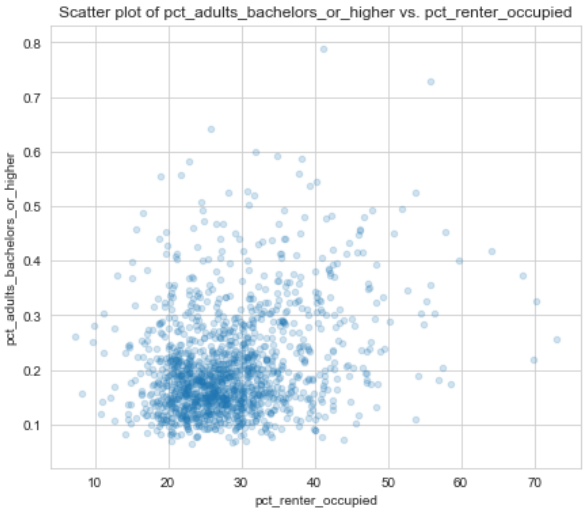


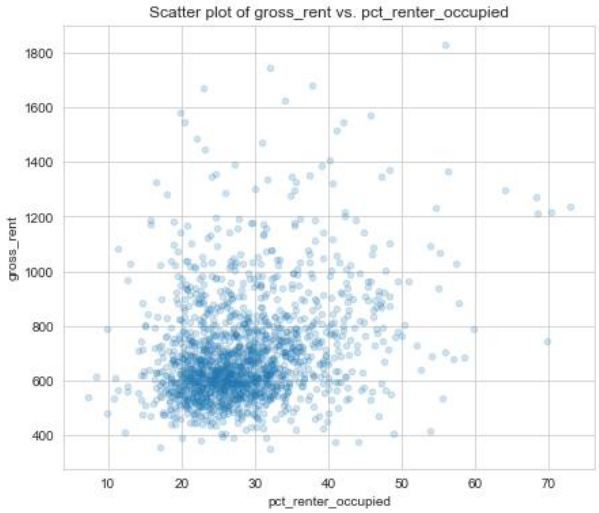
Next, we have population and percentage renter occupied households as you can see, they are very strongly correlated.

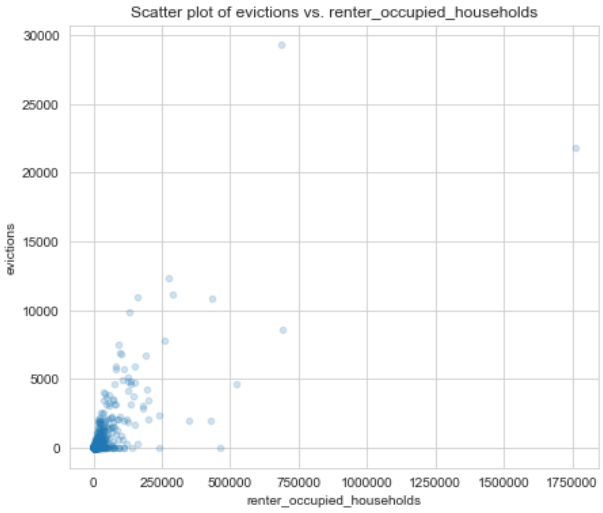


Then we check percent rented occupied vs percent adults with high school diploma and we see a negative correlation. Which is a positive correlation for percent with bachelors or higher and median gross rent (shown here), and finally evictions have a positive correlation with renter occupied households (and population because of the correlation above):









Percent adults with a high school diploma also have positive correlations with obesity, physical inactivity, heart disease, diabetes and death rate, all negatively correlated with percent adults with bachelors or higher and the median gross rent, making them also strong predictors for the median gross rent. There are also other helpful relationships with 43 variables interacting with each other and we won't be able to discuss them all. In order to focus our discussion on prediction accuracy we've selected the variables above because of their strong correlation with gross rent. We need to address the noise in the data where it was shown on the visualizations. We can certainly see the strength of the correlation of these predictors with median gross rent, and we need to make sure our algorithm sees it.

In summary, we have strong correlations to median gross rent using the following variables:

	gross_rent	pct_adults_bachelors_or_higher	pct_adults_with_high_school_diploma	pct_adult_obesity	pct_diabetes	pct_physical_inactivity
gross_rent	1	0.710152	-0.634984	-0.482307	-0.438652	-0.577041
pct_adults_bachelors_or_higher	0.710152	1	-0.761688	-0.594428	-0.559656	-0.63927
pct_adults_with_high_school_diploma	-0.634984	-0.761688	1	0.49217	0.480347	0.602733
pct_adult_obesity	-0.482307	-0.594428	0.49217	1	0.731772	0.705268
pct_diabetes	-0.438652	-0.559656	0.480347	0.731772	1	0.724041
pct_physical_inactivity	-0.577041	-0.63927	0.602733	0.705268	0.724041	1

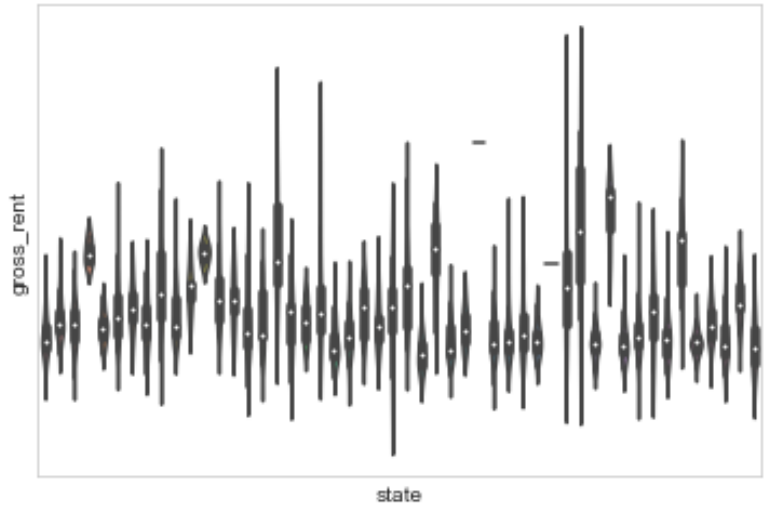
The above predictors are affected by excessive drinking, smoking and motor vehicle crash deaths:

	pct_excessive_drinking	pct_adult_smoking	motor_vehicle_crash_deaths_per_100k	pct_adult_obesity	pct_diabetes	pct_physical_inactivity
pct_excessive_drinking	1	0.191254	-0.246779	-0.250937	-0.390954	-0.385806
pct_adult_smoking	0.191254	1	0.258843	0.256894	0.251431	0.289787
motor_vehicle_crash_deaths_per_100k	-0.246779	0.258843	1	0.343313	0.393515	0.405683
pct_adult_obesity	-0.250937	0.256894	0.343313	1	0.731772	0.705268
pct_diabetes	-0.390954	0.251431	0.393515	0.731772	1	0.724041
pct_physical_inactivity	-0.385806	0.289787	0.405683	0.705268	0.724041	1

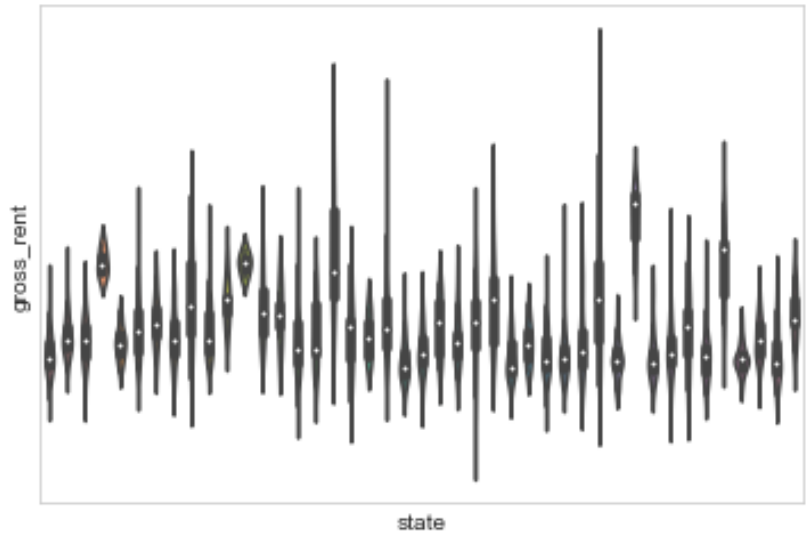
We can see that this is a powerful combination of predictors but recall that motor vehicle crash deaths, excessive drinking, and adult smoking had some noise that would make the distribution bimodal, one mode will be noise and the other one the correct distribution.

Categorical Relationships

We'll do a run through the categorical relationships. I don't think we would argue about nice variance between states, let's check it out:

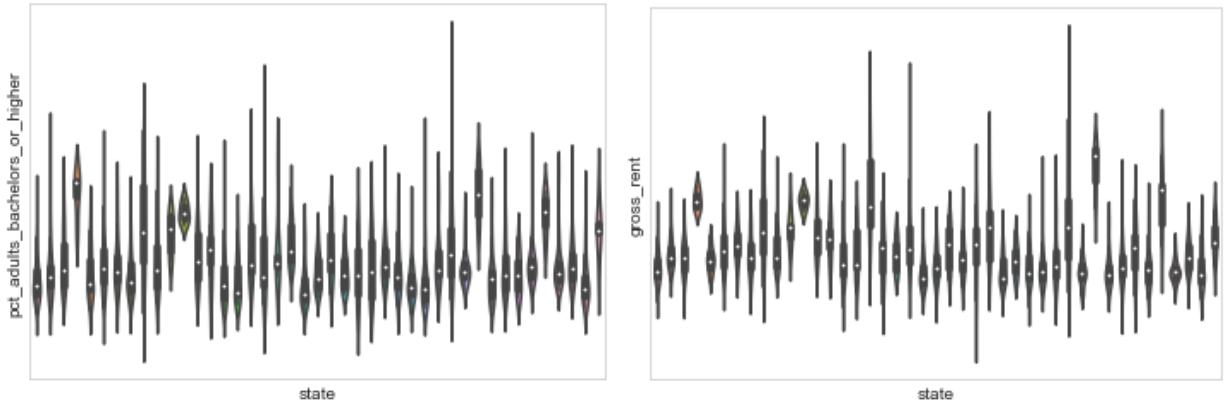


No standard deviation could be a problem--the two flat points. If we are looking at predictors for gross rent, we need the standard deviation and the mean to provide useful variance. We will remove them.

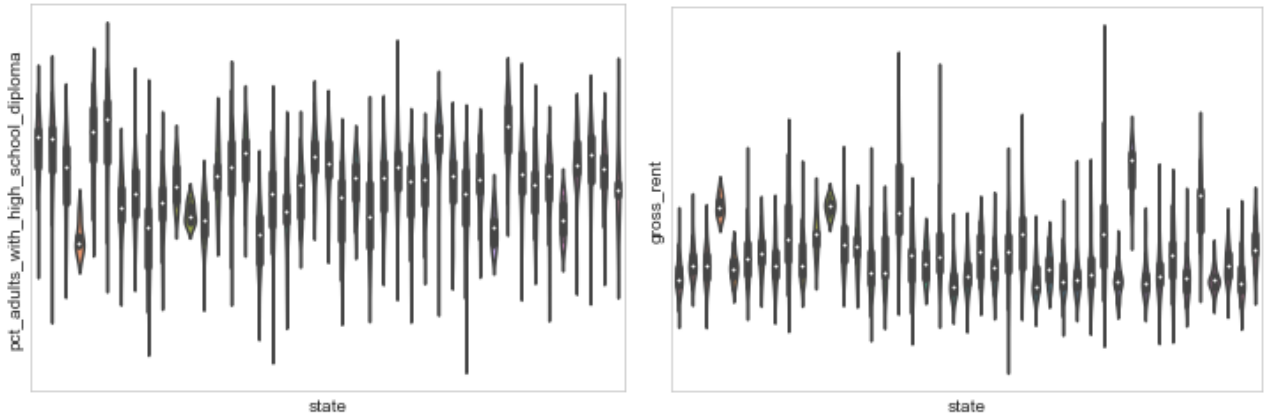


This looks much better. In this case evictions were not growing by population as in the rest of the data. Let's check the trends on the relationships we've discussed that will help us predict gross rent.

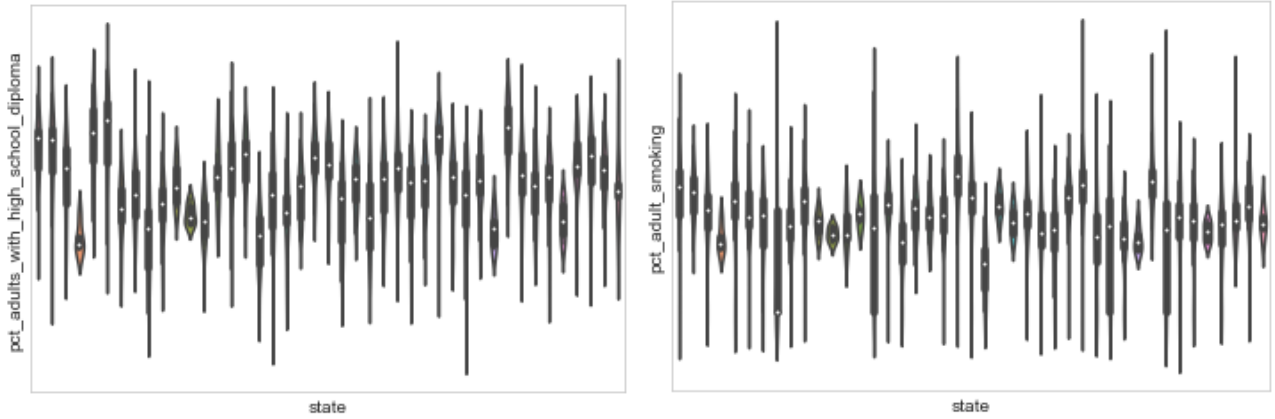
Percent adult bachelors or higher aligns closely with gross rent:



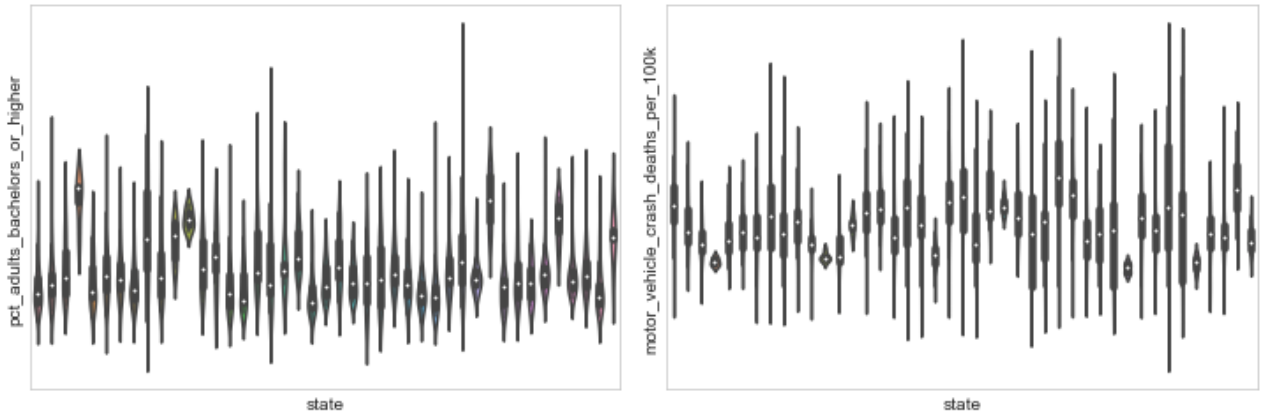
Now let's check percent adult high school diploma: you can see the negative correlation with gross rent opposing the values rather than following them like percent adult bachelors or higher.



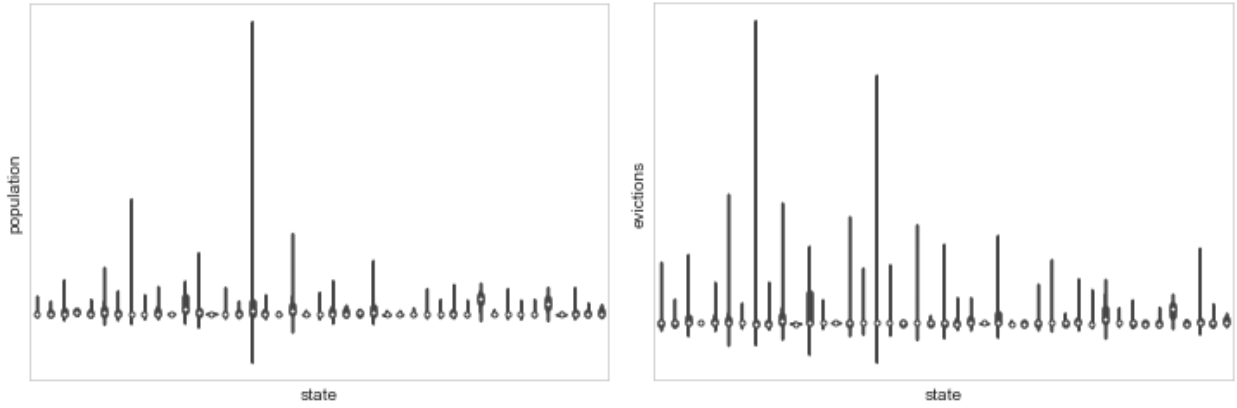
Negative correlations should line up with each other, here are motor vehicle crash deaths and percent adults with high school diploma, look at the median values in the violin plots, they do indeed follow each other. There are some slight variations that will be explained as we dig deeper:



We can also check that negative correlations also oppose the positive correlations that we explored before. Here are percentage adults with bachelor's degree or higher and motor vehicle crash, see how they oppose rather than follow each other, again, for the most part, we will certainly dig deeper:

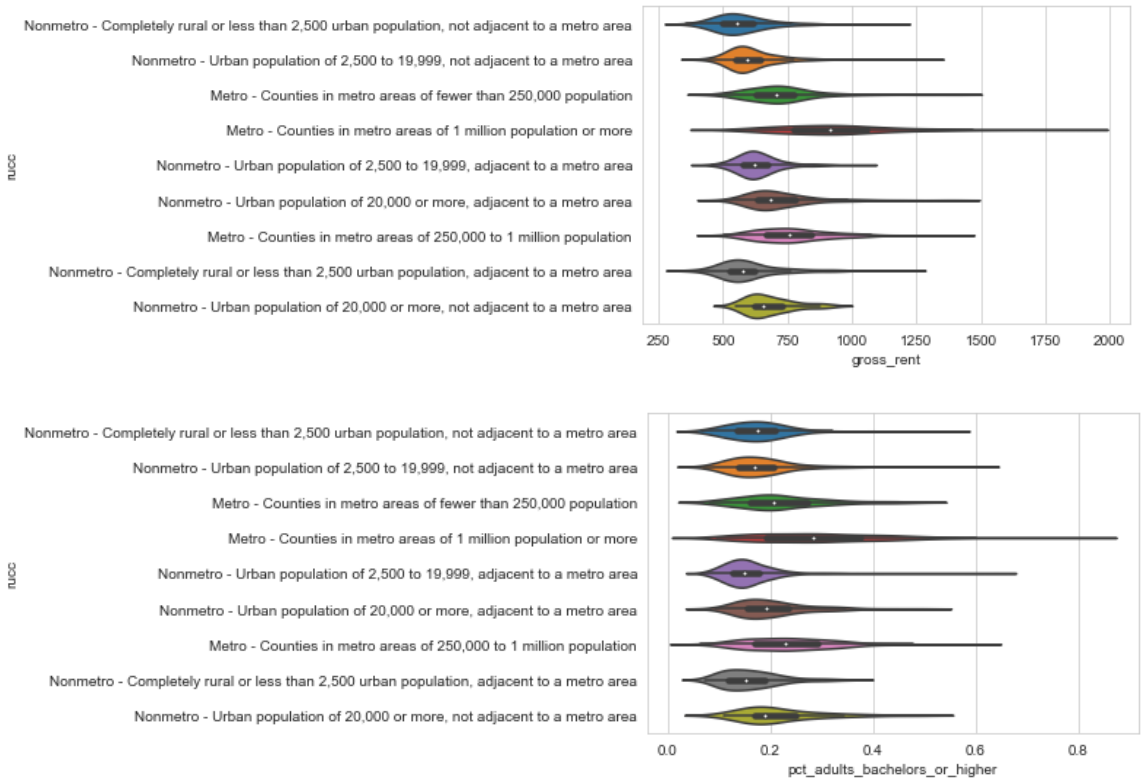


Now that we know the relationships we unveiled are being maintained, let's look at population and evictions, note states 7 and 16 with the largest populations and the number of evictions, start from 0 on the left – the two largest populations have the largest number of evictions, although if the relationship was more linear you would have expected 16 to be greater than 7 in the second plot, but there will be more insights soon for us, there are more complex relationships in the data:



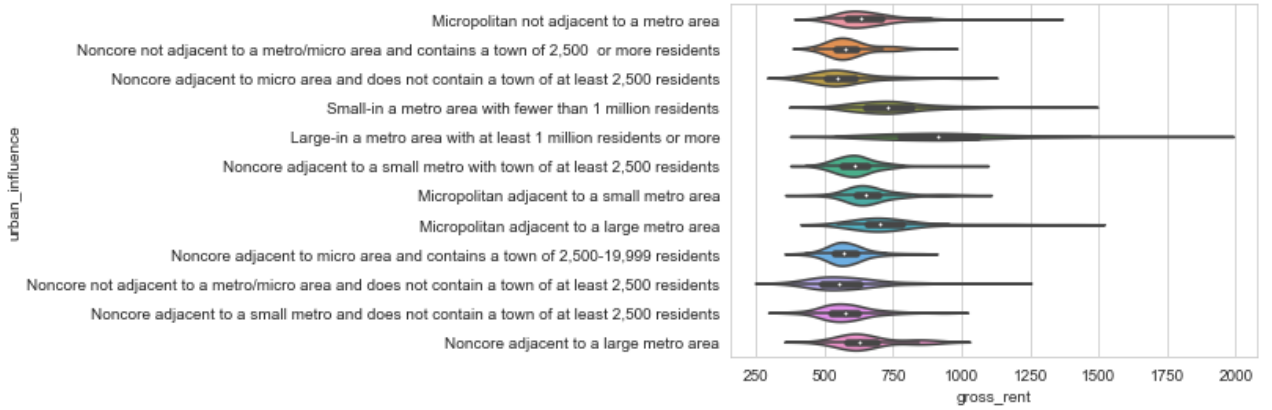
Visualizing state and county codes help us make sure we are maintaining uniform distributions throughout counties and any situation where we have a flat standard deviation, like the one we corrected, can be found by looking at the forest for the trees per say through visualization. We can also make sure that relationships we established are being maintained. Now, in this case we can see there are some other influences that we have not explained yet.

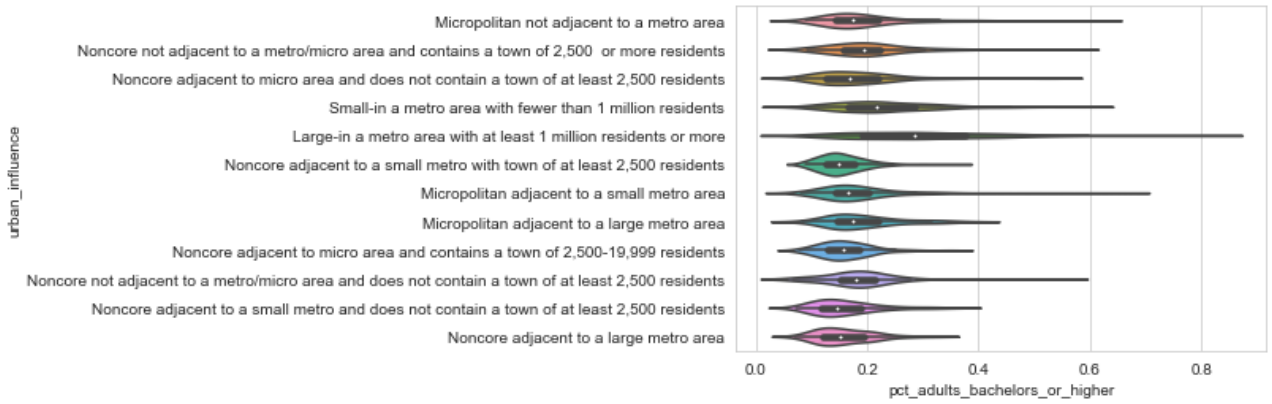
Let’s get even more granular by looking at these relationships through our categorical variables RUCC, urban influence and economic typology, so that we can understand the other influences that are occurring here. We’ll start with gross rent and RUCC, and compare to adults with bachelors or higher and RUCC:



Note the median gross rent cost and the percentage of the population with a bachelor’s degree. Populations with lower percentage of bachelor’s degree or higher tend to have lower rent prices and vice versa, following the same positive and negative correlation relationships we’ve followed throughout this report, but this happens when they are rural or urban with small populations except for the first two populations from the top, you would’ve thought the first one from the top would have had a lower median percentage bachelors or higher. Clearly RUCC is exerting an influence based on its categories. Metro areas with larger populations, let’s say 250,000 and up, tend to have flatter modes and very widespread distributions, that spread into very high rent values (right skewed) but the median remains relatively close to the medians of the other populations that are much less wide spread, have tighter modes and less of a standard deviation. This effect gets magnified even further on very large populations.

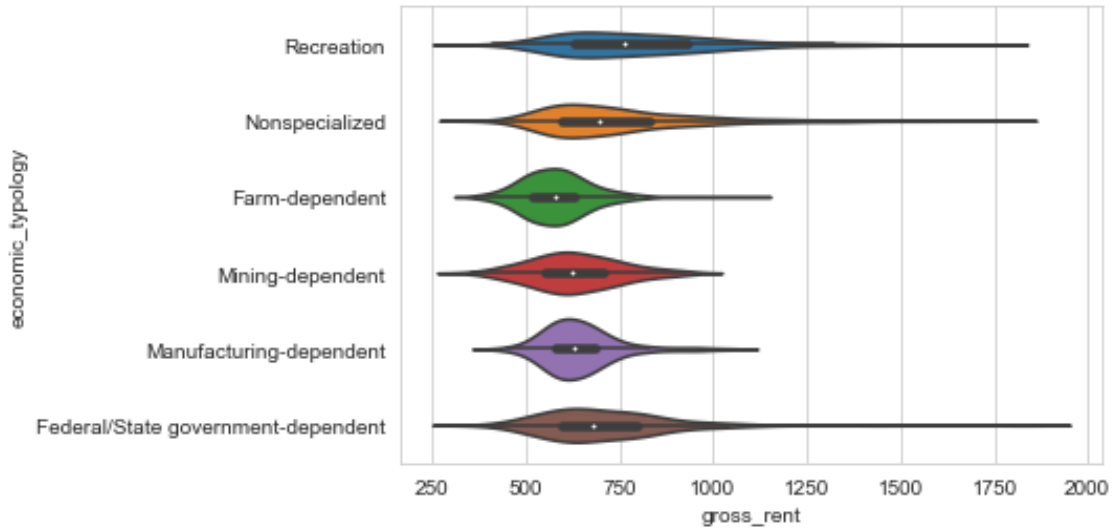
Let’s look at gross rent and percent bachelors or higher and compare urban influence this time:

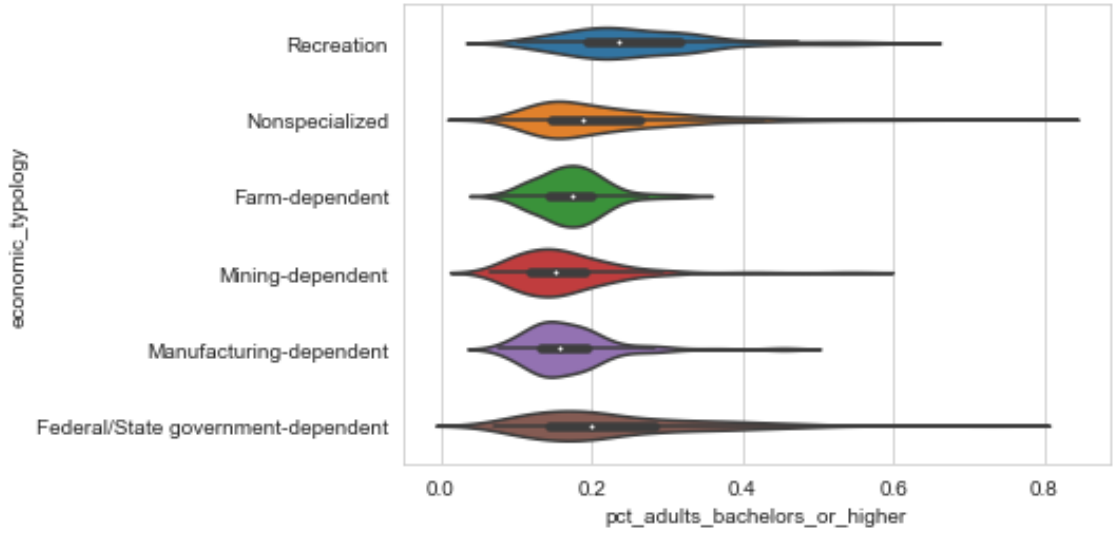




Once the urban influence is added, the higher percentage bachelors or higher does not necessarily correspond to the higher median gross rent within each urban influence group, also, again the distributions are widely dispersed in some categories, you clearly see a correspondence between the size of the population and the flattening of the mode and also an increase in standard deviation for both the median gross rent and the bachelors or higher just like with the RUCC plots.

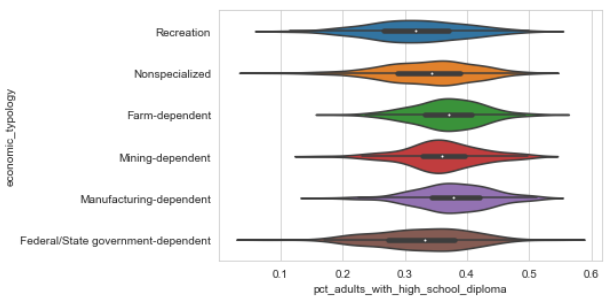
Let’s look at economic typology, once more we will use gross rent and bachelor’s or higher and compare them:

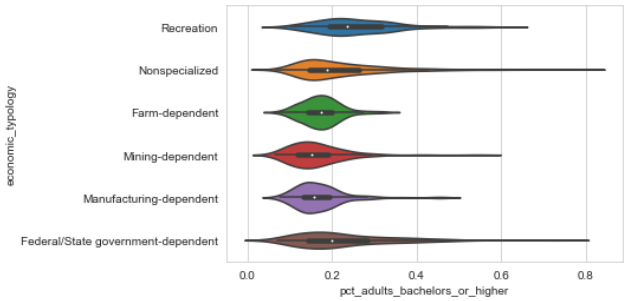




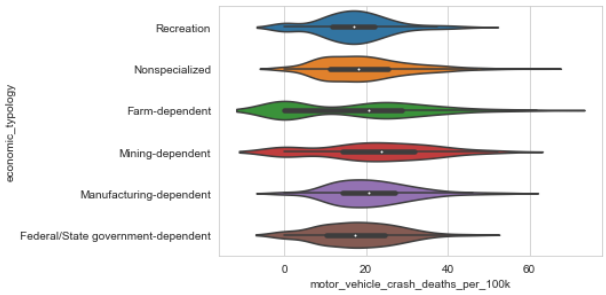
That is aligned with our numeric variables in terms of the correlation between bachelors or higher and the median gross rate, but it also gives us further insight as to the percentage of bachelors or higher within the economic types in these regions.

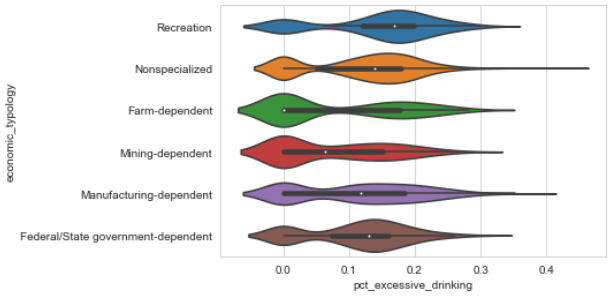
Now, now considering the implications of the correlations we've studied, note that there is a bigger percentage of adults with high school diploma for each of the economic typology categories:

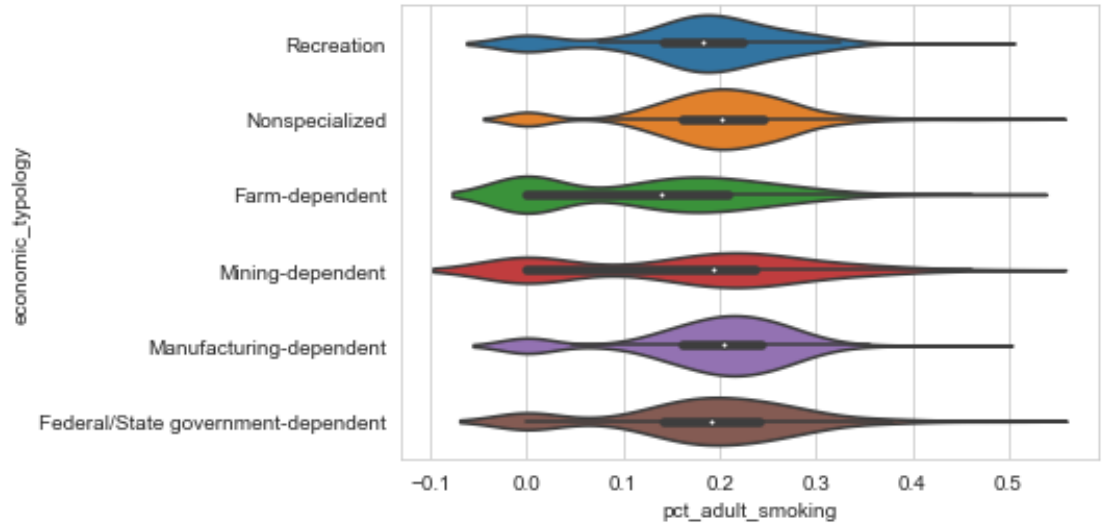




And adults with a high school diploma are correlated to the following variables we are focusing on and we showed on our correlation matrices, you can clearly see the bimodal distributions affecting these categories. This is the noise we recognized earlier caused by missing value replacement. This will affect regression:

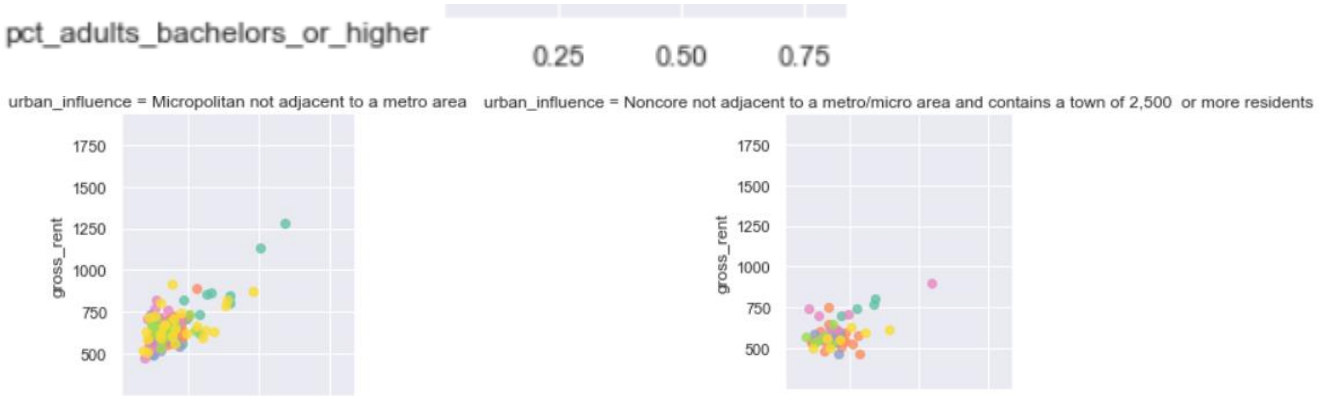


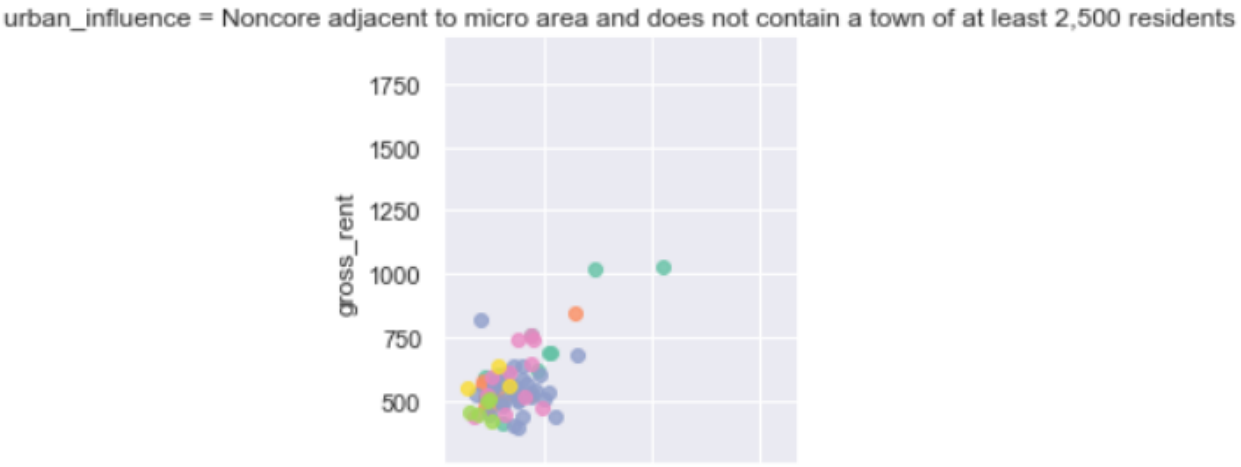




Multi-faceted Relationships

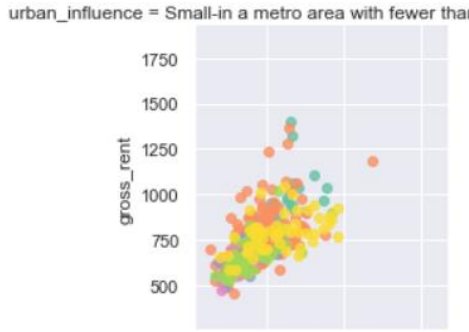
Let's explore the relationships we've discussed in a multidimensional faceted plot so we can more clearly see the interaction of our numeric variables with the categorical variables. We want to establish the importance of the categorical variables. We'll select urban influence since we potentially saw more variance there (we can see as in all research there's a lot of room for further exploration, both RUCC and urban influence show useful variance for our model). We'll use gross rate and bachelor's or higher for numerical variables again since they give us a good glimpse of positive correlations and we'll use urban influence on the columns while adding economic typology as hue:



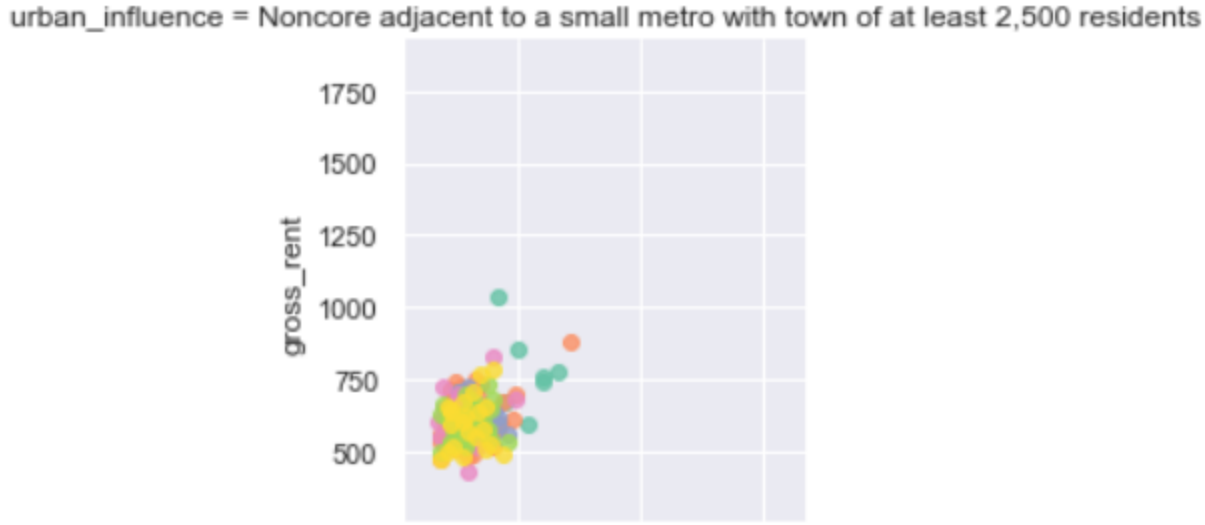


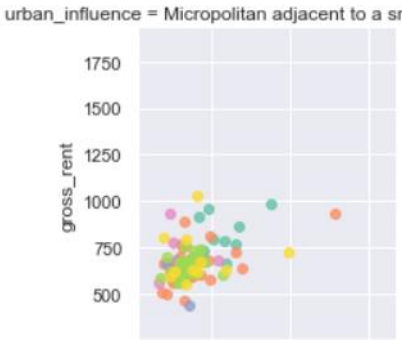

economic_typology
Recreation
Nonspecialized
Farm-dependent
Mining-dependent
Manufacturing-dependent
Federal/State government-dependent


pct_adults_bachelors_or_higher
0.25
0.50
0.75


urban_influence = Small-in a metro area with fewer than 1 million residents
gross_rent
1750
1500
1250
1000
750
500


urban_influence = Large-in a metro area with at least 1 million residents or more
gross_rent
1750
1500
1250
1000
750
500

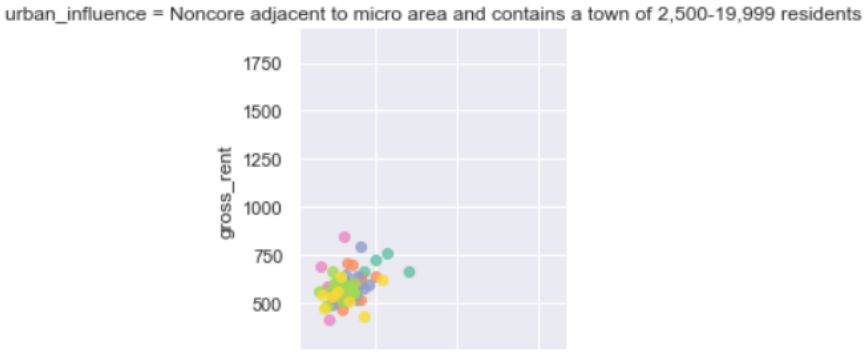

urban_influence = Noncore adjacent to a small metro with town of at least 2,500 residents
gross_rent
1750
1500
1250
1000
750
500

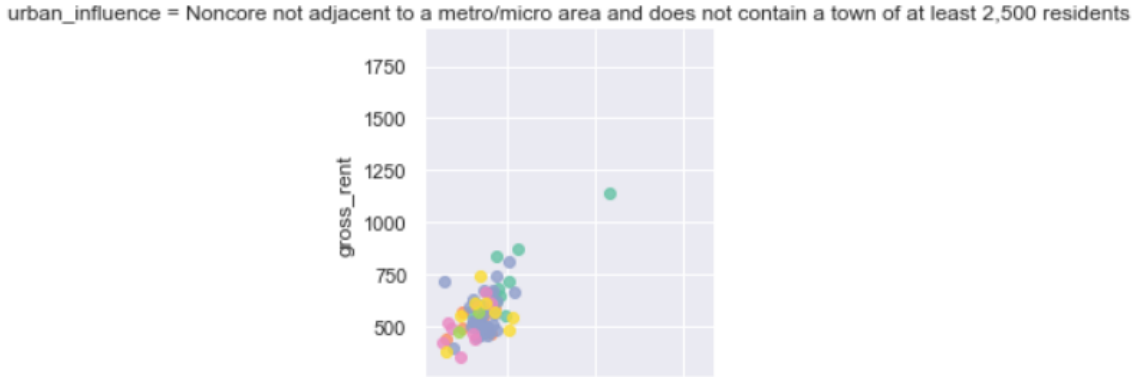

urban_influence = Micropolitan adjacent to a small metro area
gross_rent
1750
1500
1250
1000
750
500

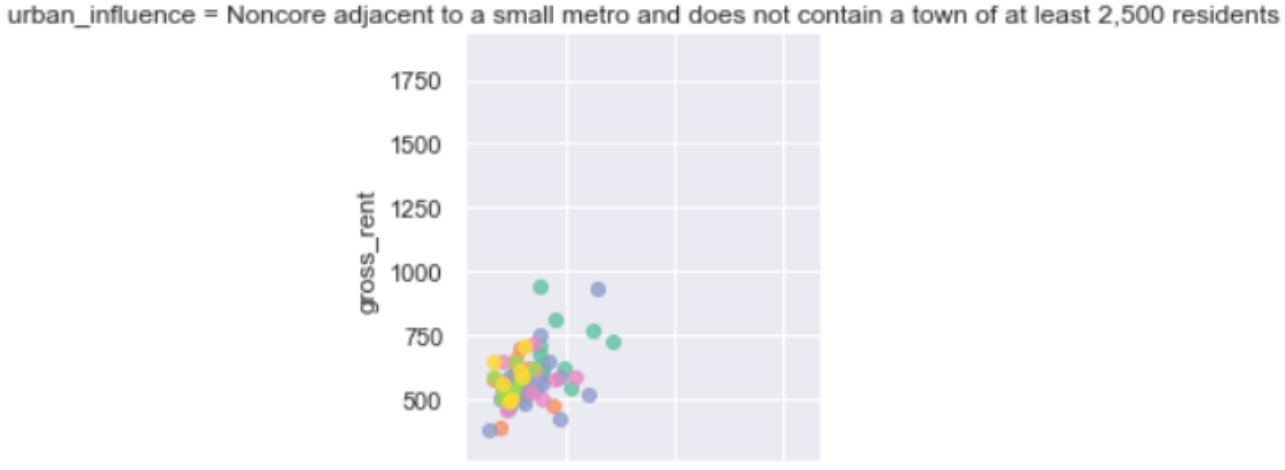

urban_influence = Micropolitan adjacent to a large metro area
gross_rent
1750
1500
1250
1000
750
500

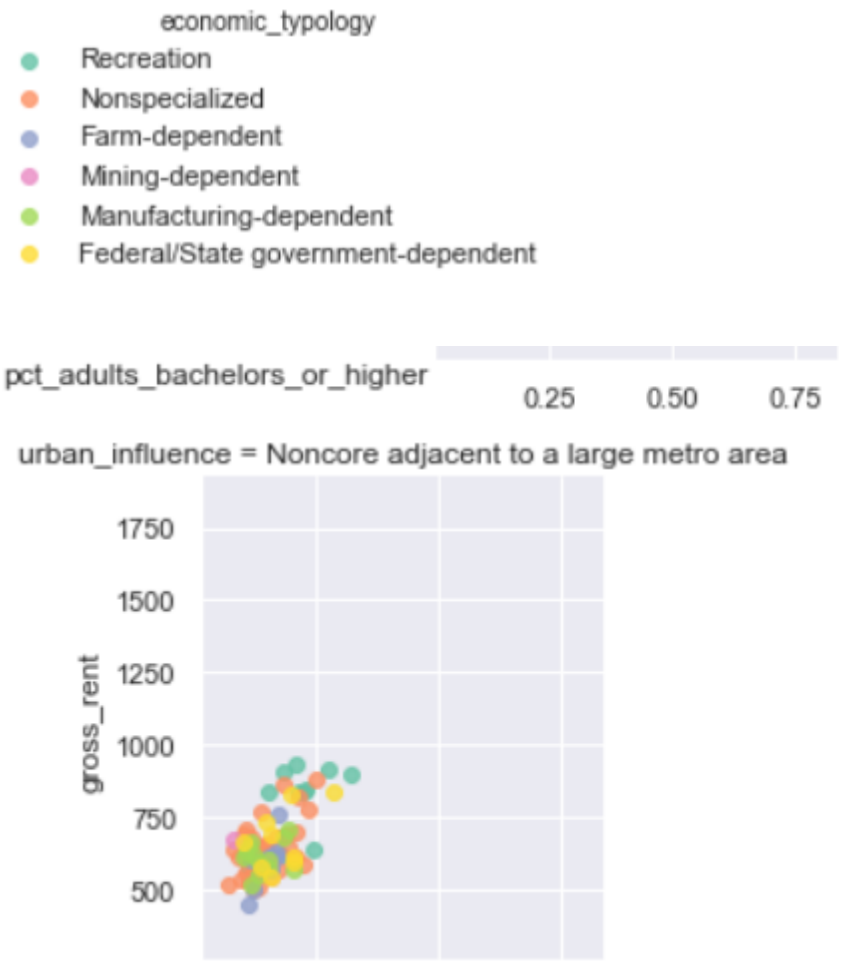

economic_typology
Recreation
Nonspecialized
Farm-dependent
Mining-dependent
Manufacturing-dependent
Federal/State government-dependent


pct_adults_bachelors_or_higher
0.25
0.50
0.75


urban_influence = Noncore adjacent to micro area and contains a town of 2,500-19,999 residents
gross_rent
1750
1500
1250
1000
750
500

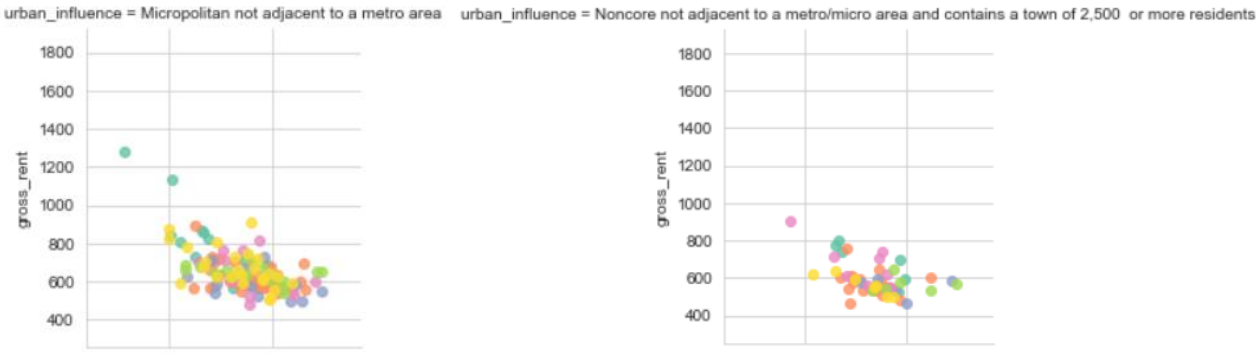

urban_influence = Noncore not adjacent to a metro/micro area and does not contain a town of at least 2,500 residents
gross_rent
1750
1500
1250
1000
750
500

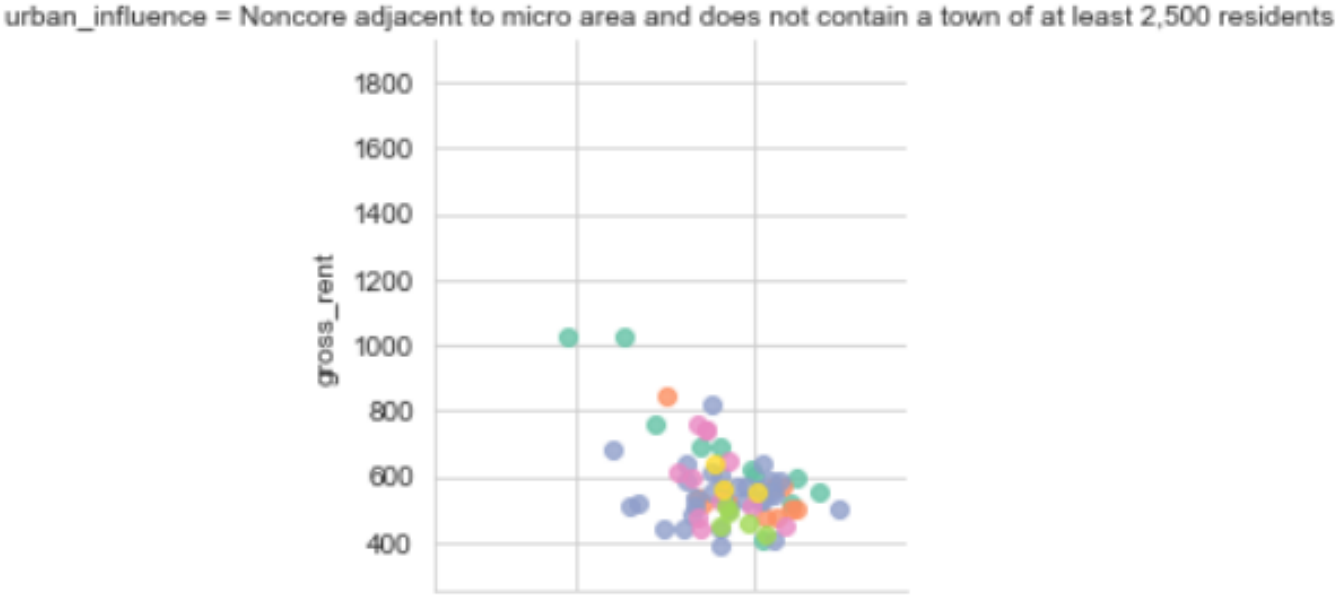

urban_influence = Noncore adjacent to a small metro and does not contain a town of at least 2,500 residents
gross_rent
1750
1500
1250
1000
750
500



Fascinating isn't it? The highest rents might be determined by economic typology, which is itself affected by urban influence and the highest gross rents might not be correlated with the highest bachelor's degree after these influences, even if overall we do still see the positive correlation we've studied, which is evident, these other influences can be observed even in the larger metro areas, as evidenced by the large metro area with over one million population having the greatest gross rent being of economic typology 'federal/state government dependent' as we might have expected from urban influence rather than the correlation between 'bachelors or higher' and gross rent, we see that the highest rent does not necessarily correspond to the highest percent bachelor's or higher. The data that we have on these variables is crucial for our model to predict correctly. They are creating additional variance useful for our model.

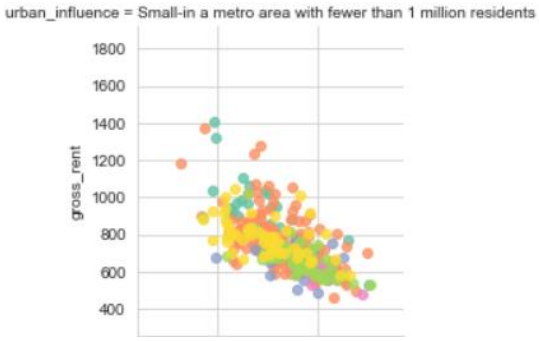
Let’s examine how the negatively correlated variables react with the categorical ones, from our exploration, we know they are associated with percent adults with high school diploma, which is correlated to the variables that are negatively correlated with gross rent, so we will just swap that variable for percentage bachelors or higher and re-examine the faceted plots:

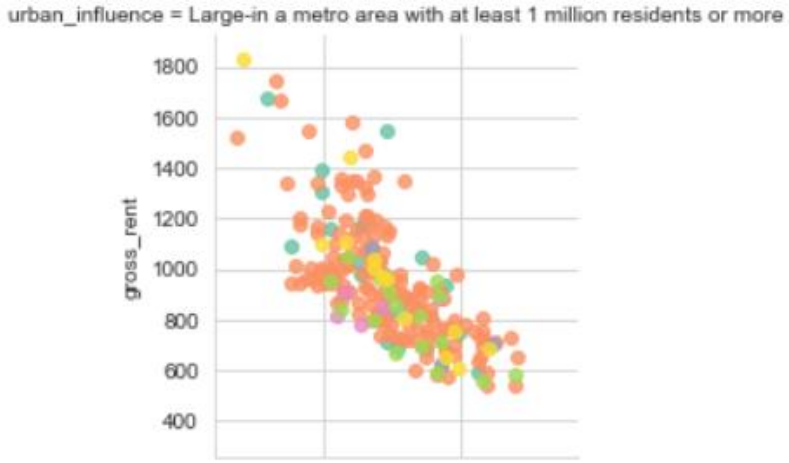


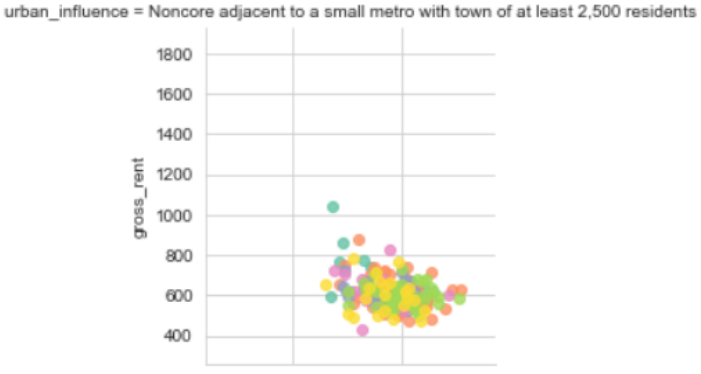



economic_typology
Recreation
Nonspecialized
Farm-dependent
Mining-dependent
Manufacturing-dependent
Federal/State government-dependent


pct_adults_with_high_school_diploma
0.2
0.4

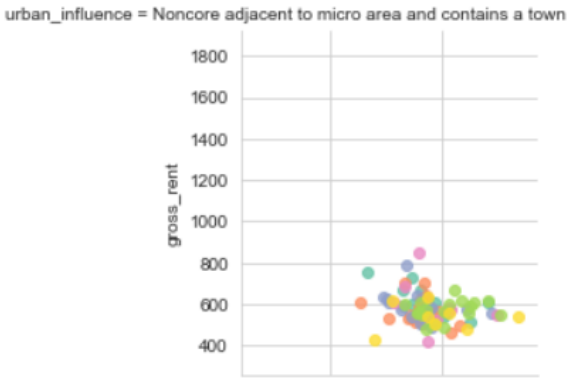

urban_influence = Small-in a metro area with fewer than 1 million residents
gross_rent
1800
1600
1400
1200
1000
800
600
400

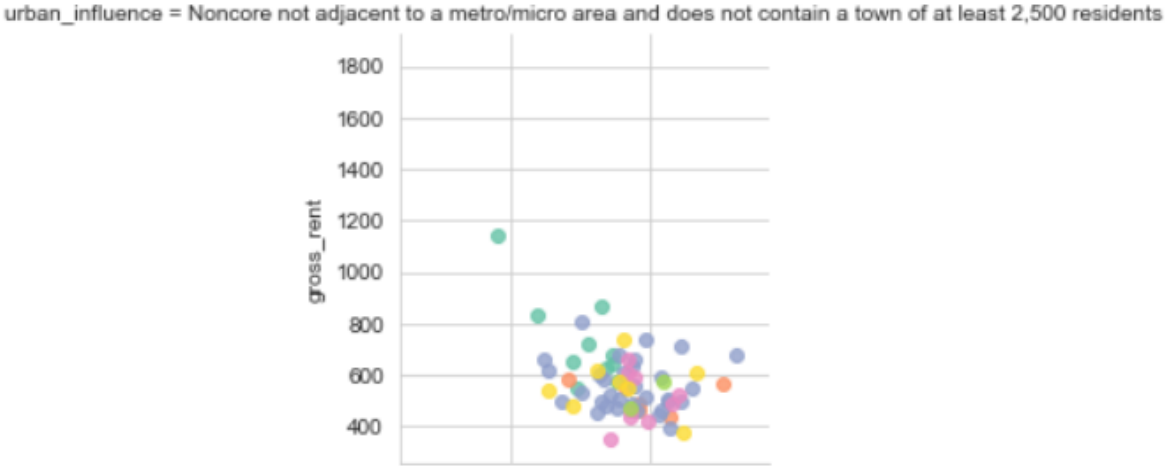

urban_influence = Large-in a metro area with at least 1 million residents or more
gross_rent
1800
1600
1400
1200
1000
800
600
400

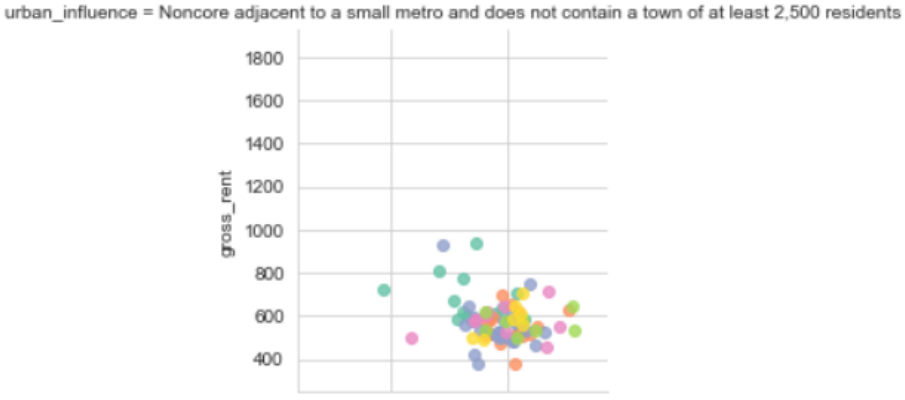

urban_influence = Noncore adjacent to a small metro with town of at least 2,500 residents
gross_rent
1800
1600
1400
1200
1000
800
600
400

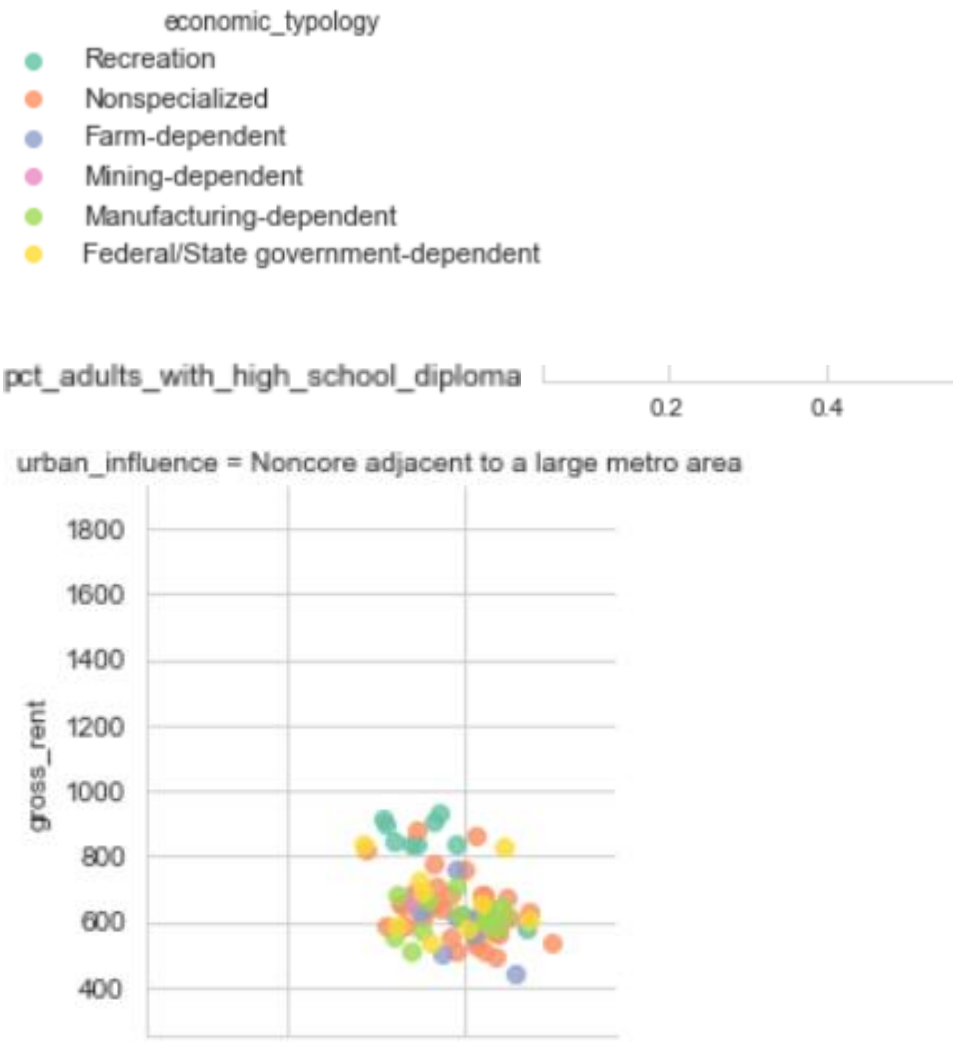

economic_typology
Recreation
Nonspecialized
Farm-dependent
Mining-dependent
Manufacturing-dependent
Federal/State government-dependent


pct_adults_with_high_school_diploma
0.2
0.4


urban_influence = Noncore adjacent to micro area and contains a town of 2,500-19,999 residents
gross_rent
1800
1600
1400
1200
1000
800
600
400


urban_influence = Noncore not adjacent to a metro/micro area and does not contain a town of at least 2,500 residents
gross_rent
1800
1600
1400
1200
1000
800
600
400


urban_influence = Noncore adjacent to a small metro and does not contain a town of at least 2,500 residents
gross_rent
1800
1600
1400
1200
1000
800
600
400

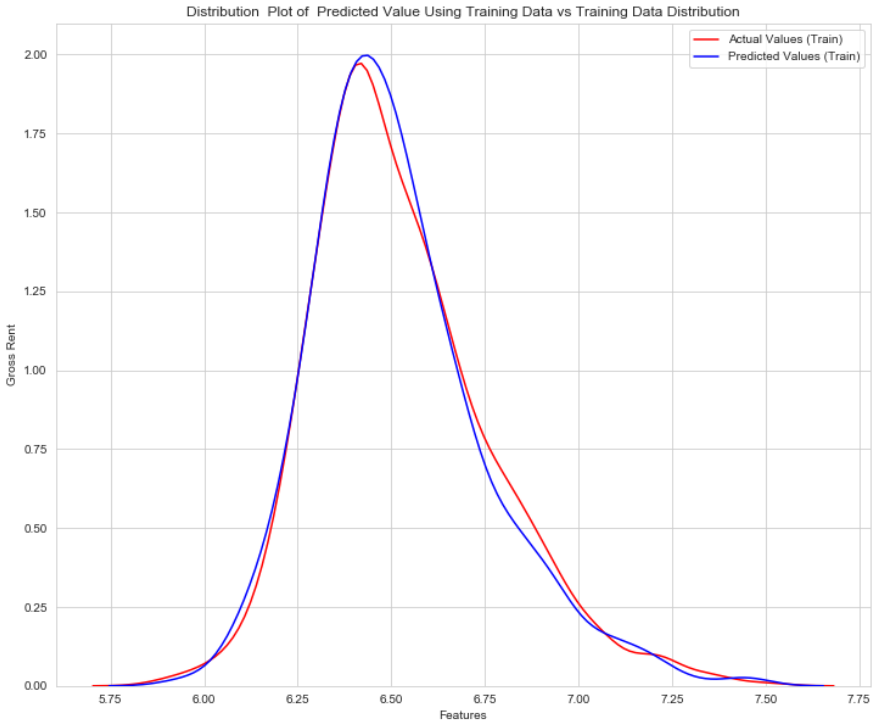


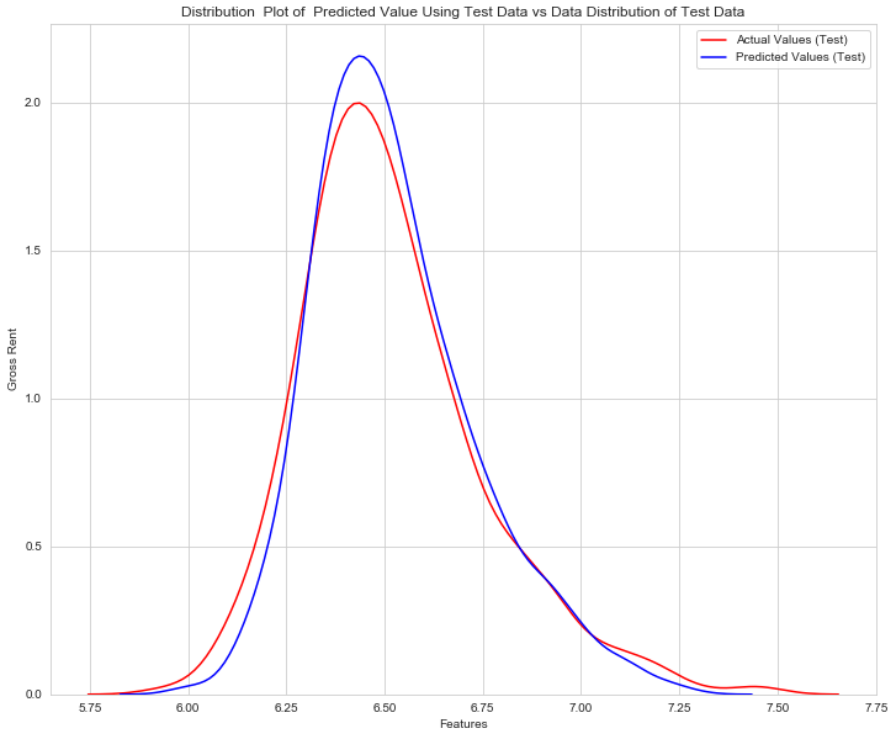
Higher rents appear to be more correlated with recreation and nonspecialized economic typologies and mining for some smaller regions, depending on the urban influence. These categories appear to be more correlated with a bachelor's degree or higher percentage, but we must be careful to check each region as there are stronger influences depending on the size of the population, the amount of urban influence and overall amount of particular economic typologies from the region. Lower rents for the most part, appear to be correlated to farming, mining or manufacturing depending on the urban influence and all these appear to be correlated with percentage high school diploma, we also see a correlation with federal state dependent economic typology in some of the urban influence regions correlated with percentage high school diploma depending on the urban influence. Clearly a lot of useful variance is created from these features. Urban Influence, combined with RUCC and economic typology are indeed providing unique, useful and completely necessary information for accurate predictions. It would be very rewarding to explore these influences further, for example, the same type of faceted plots using RUCC, time allotted, but we have explored enough to show the importance of these categorical variables in creating important influences and more complex relationships in the data.

Regression

It was determined that regression could be performed after doing some feature engineering to improve the shape of the distributions of numeric features. All the work here was performed using Jupyter Notebooks and Python libraries. As mentioned during the study, some of the features most correlated with median gross rent and in turn the features most correlated with those features had a lot of missing values and distribution shapes that would cause issues. It was determined log normal median gross rent would be best for the label, the result could be transformed back by exponentiation, and features that needed some engineering were reshaped to the log of those features or log+1 in the case of features that would include the log of zero in the resulting values. After one hot encoding the categorical variables, it was also necessary to reduce the numeric variables and do some feature selection due to the large number of features. The VarianceThreshold function from the Feature Selection package from Sci-kit Learn was used and set to select features with an 80% or higher variance. The selected features were split into training and testing sets at 70 and 30 percent consecutively for each set through random sampling. Several algorithms from the Sci-Kit Learn library were tested. The first algorithm chosen was basic linear regression with the following results:

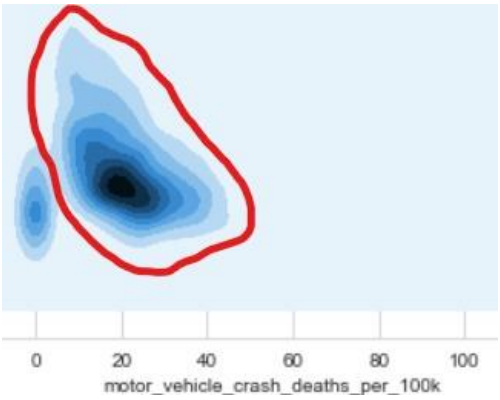
```
Mean Square Error      = 0.014236399579523135
Root Mean Square Error = 0.11931638437164921
Mean Absolute Error    = 0.09130573283864644
Median Absolute Error  = 0.06955492997663004
R^2                    = 0.7376117678489192
Adjusted R^2           = 0.7215471822070163
```





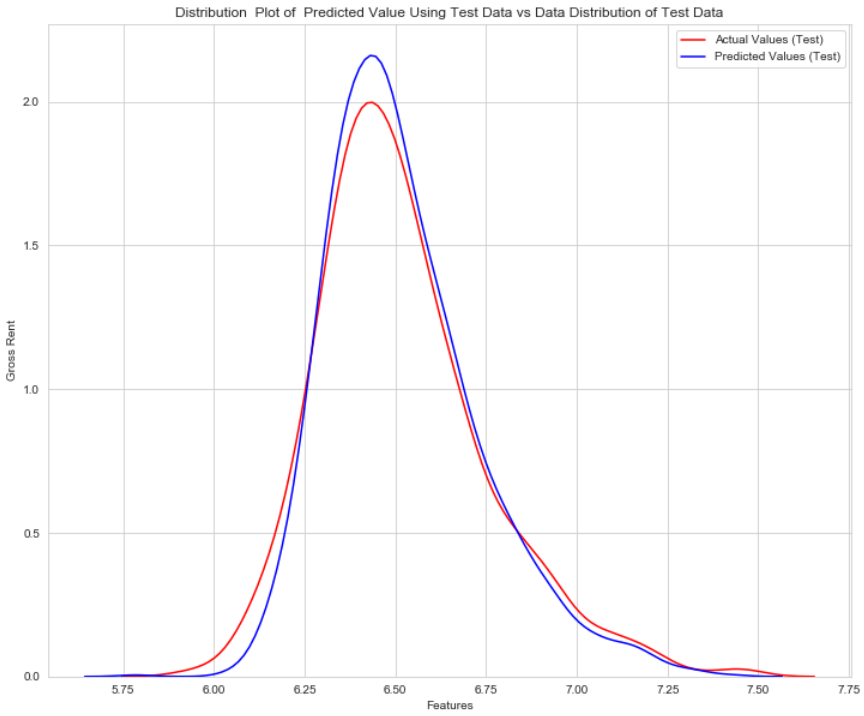
SVM (support vector machines)

We know that a glaring issue during data exploration was bimodal distributions due to noise. We mentioned an algorithm that could see that fact is what was needed. Linear regression is not the best algorithm to find probability distribution densities of different shapes, this would be a nonlinear issue, so support vector machines were selected so we could find a hyper plane cutting across the distribution densities and an RBF (radial basis function) kernel was selected so that through tuning we could wrap around the correct distributions. It was important during scaling of the features in this case, to make sure the bimodal characteristics did not disappear, so standard scaling was used. It would not be good to transform the features and lose the different distribution shapes of the noise and the correct data, so that the radial basis function could find and select the right one. The algorithm would work sort of like this intuitively; it would first create the hyper plane with boundaries, then it would find the correct distribution and approximate the shape of the distribution:



The RBF kernel does indeed help wrap around those distributions more tightly, look at the right and right bottom portion of the distribution density of predicted vs test below and compare it to the linear model's predicted vs actual test distribution plot above. You can see how much better the fit is with this model in terms of the distribution shape of the test data and there is also a significant improvement on the results:

```
Mean Square Error      = 0.011694825731745122
Root Mean Square Error = 0.1081426175554537
Mean Absolute Error    = 0.08335513894390041
Median Absolute Error  = 0.06924876656249168
R^2                    = 0.7844550069048879
Adjusted R^2           = 0.7712583746745749
```



Conclusion

This analysis has shown that we can certainly create a prediction model for median gross rent from the features available from other socioeconomic and demographic indicators. With improved data handling and collection, I can only see the model improving in performance, even with the missing data and considerable noise, we were able to achieve some more than decent measurements for prediction. The available categorical and continuous data, specifically the features with strong correlation to our target, have a significant effect on median gross rent and that is why we needed to include them, represented as cleanly as possible through careful selection of machine learning tools. There is a nonlinear component that needs to be considered as a linear model alone would not achieve the highest performance. The SVM algorithm with an RBF kernel is a highly recommended machine learning

algorithm for this model, since it can detect the shape of the distributions and capture the nonlinear aspect, and it can be then tuned to find an optimal solution.