Predicting County-Level Rents

Jean-Thierry BONTINCK – December 2019

This analysis has been done as part of the Capstone Project of the Microsoft Professional Program for Data Science.

Executive Summary

Context

The current housing market in USA tend to exacerbate the gap between the rich and poor. Over the past five decades, housing costs have risen faster than incomes, low-cost housing has been disappearing from the market, and ethnic disparities in homeownership rates have deepened. Indeed, the rent is typically the biggest part of a household spending. Therefore, the "rental burden" can produce a snowball effect and impact other parts of the household budget. When the part of the monthly incomes taken by the rental grows, it can preclude spending like food or health care. When it lowers, household can invest in long term wealth like education or health. More, housing issues are intertwined with many other social problems and, combined, produce a growing need for help. But housing assistance is not an entitlement. Even if eligibility is based on US citizenship, incomes and family size, only 1 over 4 who is eligible gets assistance. Nevertheless, this assistance can have a chain effect and really benefit to those who need protection the most, like the children, the elder or the disabled.

The goal of the project is to analyze how poverty, income, health, ethnicity, and other sociodemographic factors are related to rent costs by building a model which predicts median gross rent values in counties across the United States of America. The dataset is composed a train and a test set. They are 43 features grouped in several categories: Housing, Ethnicity, Economic, Health, Demographic. The target, the gross rent, is known for the train set only.

Key findings

We have covered a detailed statistical analysis over the all data provided and built a regression model to predict the median gross rent using those data. Among the variables included the dataset, some seems to have a stronger relationship with the target and have a greater influence on the prediction of the model. Those important features are spread over almost all categories provided and are presented in the summary below:

1. **Greater metropolitan (more than 1 million inhabitant).** The typology of urban area seems to have a high influence on our prediction model, especially when it comes to the bigger one. The top 2 features come from the **rucc** and the **urban_influence** categorical features and represents the big metropolitan of more than 1 million population (respectively, "*Metro - Counties in metro areas of 1 million population or more*" and "*Large-in a metro area with at least 1 million residents or more*").

2. **Education.** The education level is also well represented across 2 features present in top 10 features (**pct_adults_bachelors_or_higher** and **pct_adults_with_high_school_diploma)** and were identified at the very early stage of the analysis as correlated with the target.

3. **Ethnicity**. Apart from the **pct_asian** that seems to have a stronger relationship with the target, all the other ethnicity related variables seem to have a moderate relationship. However, they are widely present and seems to have, all together, a significant influence.

4. **Death_rate_per_1k.** This individual feature is at the 5th position in our important features list. It is a variable from the demographic category, but our hypothesis is that it could be linked to several other features that we can observe as important to moderately important: elder-related (pct_aged_65_years_and_older) or health-related (pct_adult_obesity, pct_physical_inactivity).

Data Description

Target Variable

We're trying to predict the variable **gross_rent** (a positive integer) for each row of the test data set.

Features

There are 43 variables in this dataset. Each row in the dataset represents a United States county in a single year. We provide a unique identifier for an individual county, but the counties in the test set are distinct from counties in the train set. In other words, no county that appears in the train set will appear in the test set. Thus, county-specific features (i.e. county dummy variables) will not be an option. However, the counties in the test set still share similar patterns as those in the train set and so other feature engineering will work the same as usual.

The variables are as follows:

ID

- **county_code** - Unique identifier for each county
- **state** - Unique identifier for each state
- **population** - Total population

HOUSING

- **renter_occupied_households** - Count of renter-occupied households
- **pct_renter_occupied** - Percent of occupied housing units that are renter-occupied
- **evictions** - Number of eviction judgments in which renters were ordered to leave in a given area and year
- **rent_burden** - Median gross rent as a percentage of household income

ETHNICITY

- **pct_white** - Percent of population that is White alone and not Hispanic or Latino
- **pct_af_am** - Percent of population that is Black or African American alone and not Hispanic or Latino
- **pct_hispanic** - Percent of population that is of Hispanic or Latino origin
- **pct_am_ind** - Percent of population that is American Indian and Alaska Native alone and not Hispanic or Latino
- **pct_asian** - Percent of population that is Asian alone and not Hispanic or Latino
- **pct_nh_pi** - Percent of population that is Native Hawaiian and Other Pacific Islander alone and not Hispanic or Latino
- **pct_multiple** - Percent of population that is two or more races and not Hispanic or Latino
- **pct_other** - Percent of population that is other race alone and not Hispanic or Latino

ECONOMIC

- **poverty_rate** - Percent of the population with income in the past 12 months below the poverty level

- **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. The official Office of Management and Budget (OMB) metro and nonmetro categories have been subdivided into three metro and six nonmetro categories. Each county in the U.S. is assigned one of the 9 codes." (USDA Economic Research Service)
- **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." (USDA Economic Research Service)
- **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. The economic dependence types include farming, mining, manufacturing, Federal/State government, recreation, and nonspecialized counties. The policy-relevant types include low education, low employment, persistent poverty, persistent child poverty, population loss, and retirement destination." (USDA Economic Research Service)
- **pct_civilian_labor** - Civilian labor force, annual average, as percent of population.
- **pct_unemployment** - Unemployment, annual average, as percent of population

HEALTH

- **pct_uninsured_adults** - Percent of adults without health insurance
- **pct_uninsured_children** - Percent of children without health insurance
- **pct_adult_obesity** - Percent of adults who meet clinical definition of obese
- **pct_adult_smoking** - Percent of adults who smoke
- **pct_diabetes** - Percent of population with diabetes
- **pct_low_birthweight** - Percent of babies born with low birth weight
- **pct_excessive_drinking** - Percent of adult population that engages in excessive consumption of alcohol
- **pct_physical_inactivity** - Percent of adult population that is physically inactive
- **air_pollution_particulate_matter_value** - Fine particulate matter in $\mu g/m^3$
- **homicides_per_100k** - Deaths by homicide per 100,000 population
- **motor_vehicle_crash_deaths_per_100k** - Deaths by motor vehicle crash per 100,000 population
- **heart_disease_mortality_per_100k** - Deaths from heart disease per 100,000 population
- **pop_per_dentist** - Population per dentist
- **pop_per_primary_care_physician** - Population per Primary Care Physician

DEMOGRAPHIC

- **pct_female** - Percent of population that is female
- **pct_below_18_years_of_age** - Percent of population that is below 18 years of age
- **pct_aged_65_years_and_older** - Percent of population that is aged 65 years or older

- **pct_adults_less_than_a_high_school_diploma** - Percent of adult population that does not have a high school diploma
- **pct_adults_with_high_school_diploma** - Percent of adult population which has a high school diploma as highest level of education achieved
- **pct_adults_with_some_college** - Percent of adult population which has some college as highest level of education achieved
- **pct_adults_bachelors_or_higher** - Percent of adult population which has a bachelor's degree or higher as highest level of education achieved
- **birth_rate_per_1k** - Births per 1,000 of population
- **death_rate_per_1k** - Deaths per 1,000 of population

Individual Feature Statistics

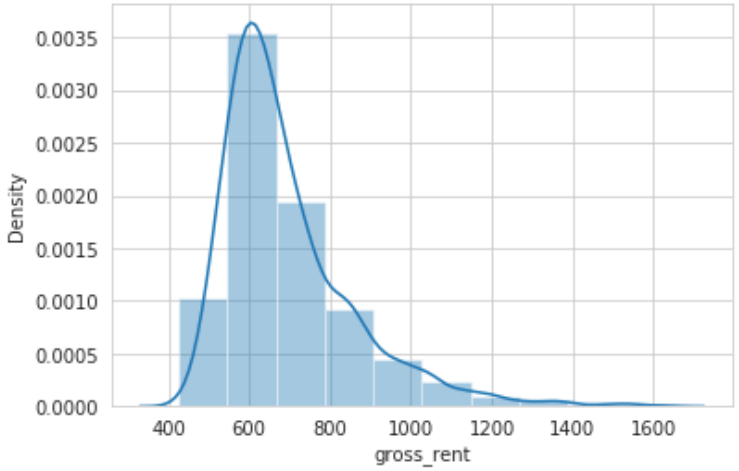
To start our analysis, we will first have a look at the basic descriptive statistics and the distribution of our features.

Target Variable

	count	mean	std	min	median	max
gross_rent	1562	701.14	192.883	351	650	1 827.00

Observations:

- From the difference between the mean and the median, we can suspect there is some positive skewness in the distribution of data. This is confirmed by the distribution plot below. This is especially important in this case as it is the data that we are willing to predict.



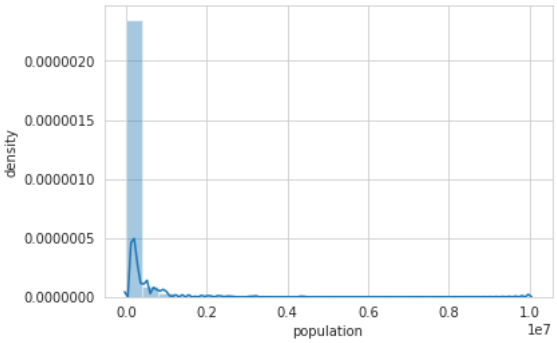
Features

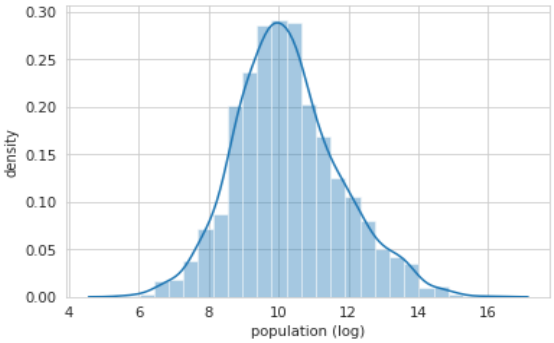
ID

	count	mean	std	min	median	max
population	1562	108 341	374 523	269	25 282	10 020 290

Notes:

- We can already detect a strong positive skewness.



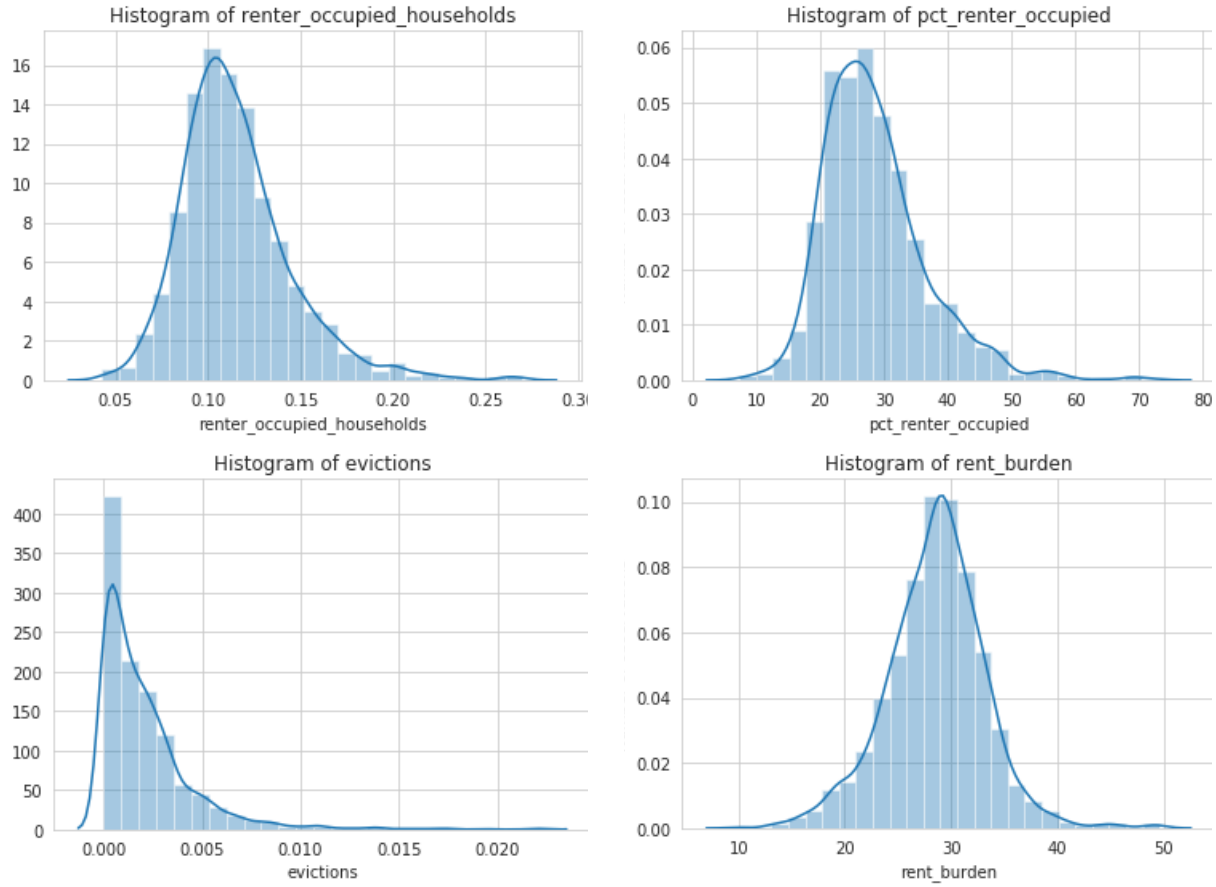


HOUSING

	count	mean	std	min	median	max
renter_occupied_households	1562	14 905	62 559	64	2 754	1 760 277
pct_renter_occupied	1562	28.53	8.12	7.28	27.20	73.01
evictions	1235	397.41	1 522.80	-1.00	27.00	29 251.00
rent_burden	1562	28.54	4.67	9.91	28.77	49.67

Notes:

- We can notice that **renter_occupied_households** and **evictions** are strongly positively skewed. It's also important to consider that we are dealing with full values for those variables (and not percentage or rate like the 2 others). So, we must normalize both by divided them by the population number in order to extract value that we can use in comparison with different councils. Then we notice a high correlation between **renter_occupied_households** and **pct_renter_occupied**.
- There are some inconsistent data for **evictions** (negative values)
- There are also an important number of missing values for **evictions.**



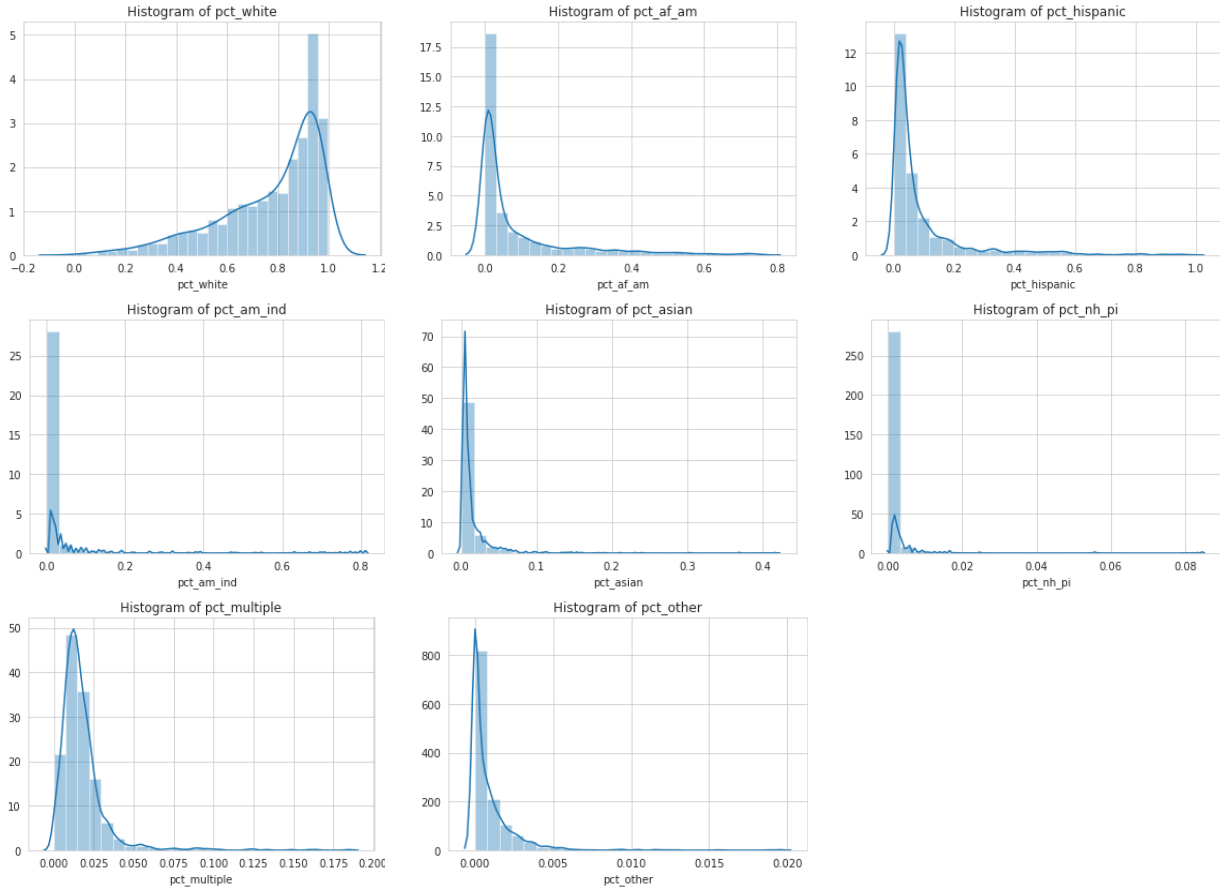
Note: renter_occupied_households and evictions are divided by population

ETHNICITY

	count	mean	std	Min	median	max
pct_white	1562	0.7690	0.2033	0.0095	0.8465	1.0000
pct_af_am	1562	0.0886	0.1435	0.0000	0.0206	0.7600
pct_hispanic	1562	0.0920	0.1416	0.0000	0.0376	0.9900
pct_am_ind	1562	0.0184	0.0748	0.0000	0.0026	0.8200
pct_asian	1562	0.0127	0.0267	0.0000	0.0053	0.4200
pct_nh_pi	1562	0.0007	0.0031	0.0000	0.0000	0.0900
pct_multiple	1562	0.0176	0.0158	0.0000	0.0144	0.1800
pct_other	1562	0.0009	0.0017	0.0000	0.0003	0.0200

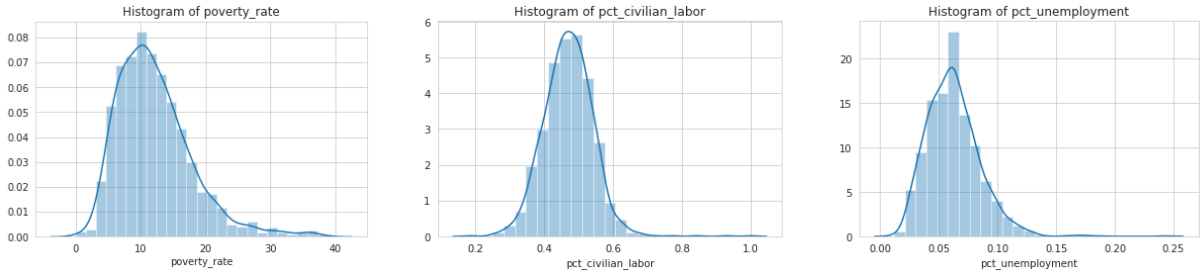
Notes:

- We must be caution with this kind of data as it can be source of discrimination. This is a question that needs to be raised and put in relationship with the goal and use of our prediction. In our case, the goal is to understand how those characteristics are in relation with the gross rent and analyze the ethnic discrimination. Note that as it is aggregated data, the risks related to collection and storage are out of our scope.
- Most of those features show a strong skewness (positive or negative)



ECONOMIC (Numeric)

	count	mean	std	min	median	max
poverty_rate	1562	12.183	5.784	0.000	11.174	38.790
pct_civilian_labor	1562	0.471	0.071	0.186	0.471	1.000
pct_unemployment	1562	0.063	0.023	0.012	0.061	0.240

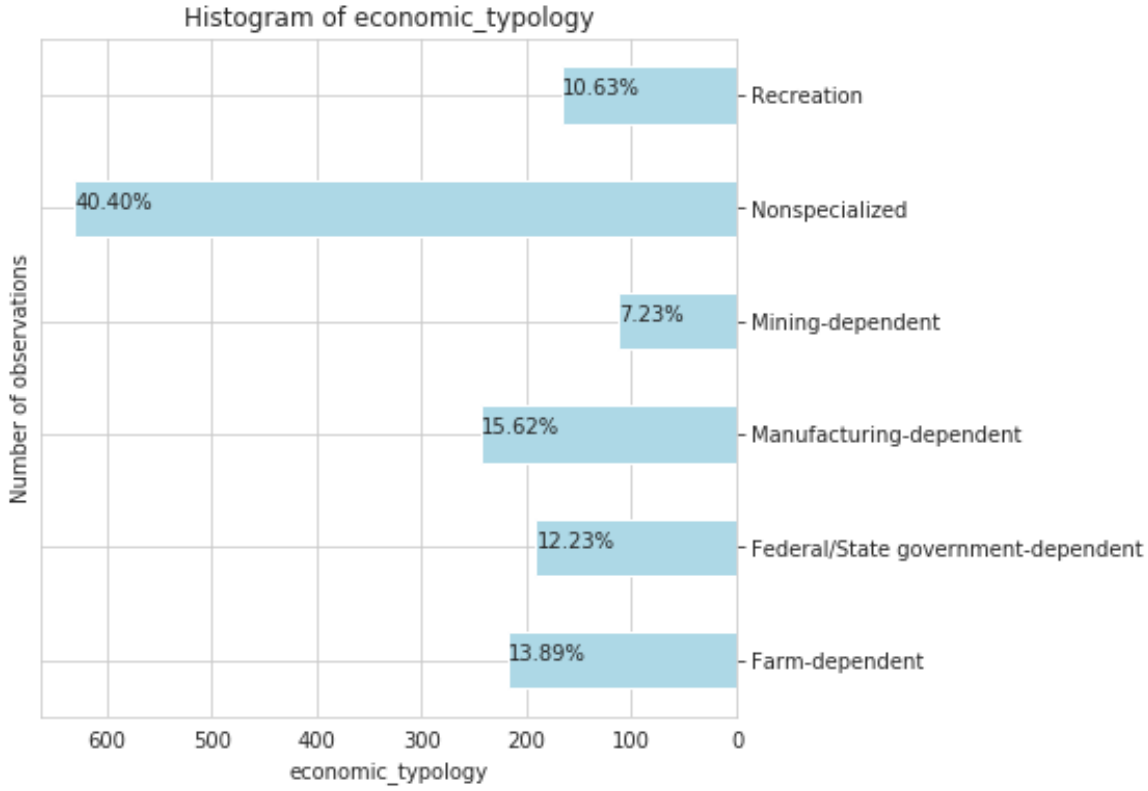


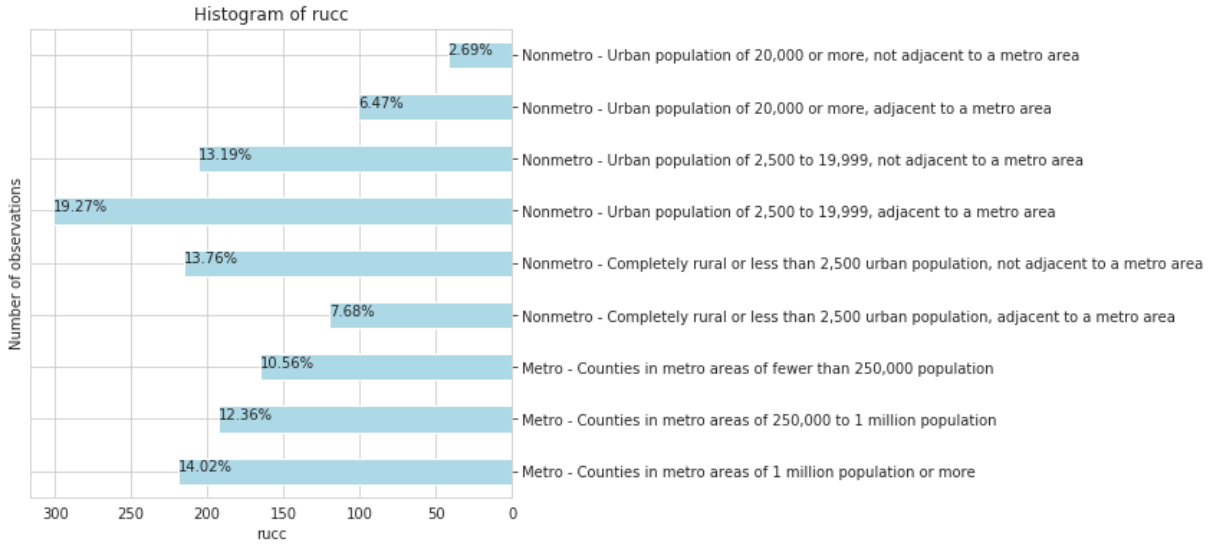
Notes:

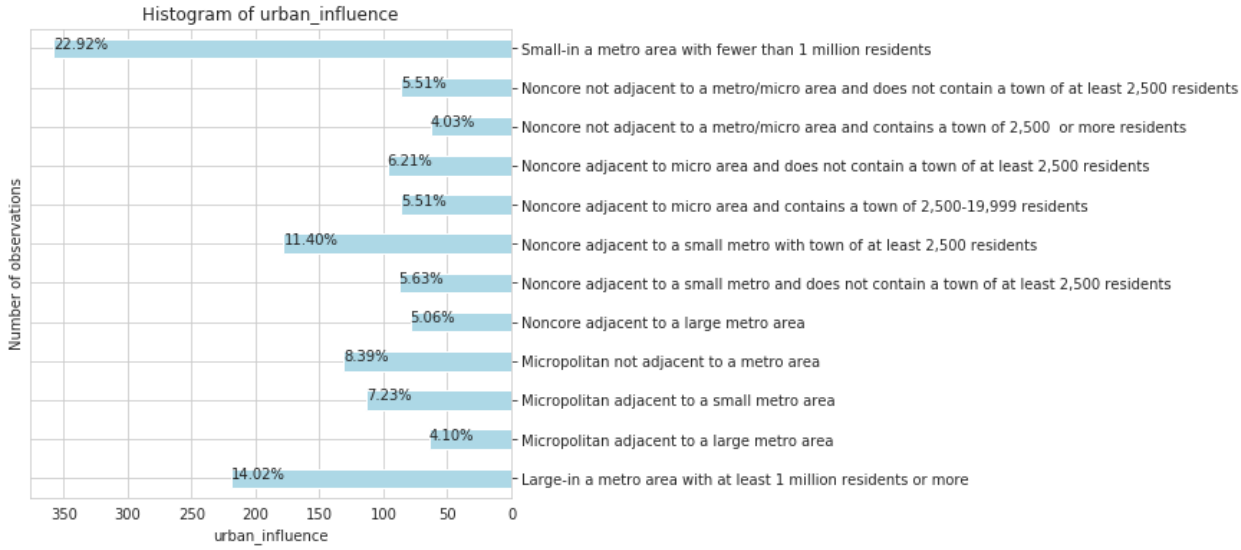
- The numeric features are close to normal distributed.

ECONOMIC (Categorical)

economic_typology	6 categories
Rucc	9 categories
urban_influence	12 categories







Notes:

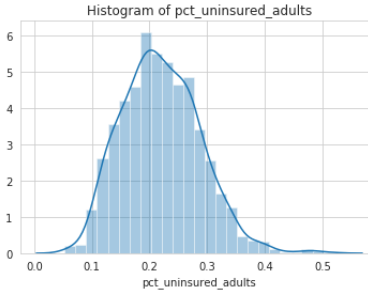
- The numeric features are close to normal distributed.
- The **economic_typology** categorical feature shows a clear unbalance distribution with one category having a frequency over 40%
- The **urban_influence** categorical feature has one predominant category with more than twice more observations than the second most represented category

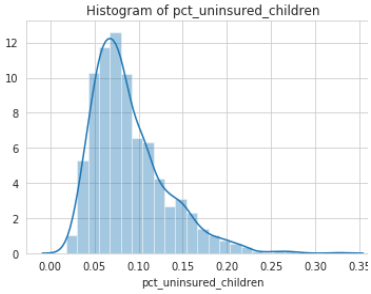
HEALTH

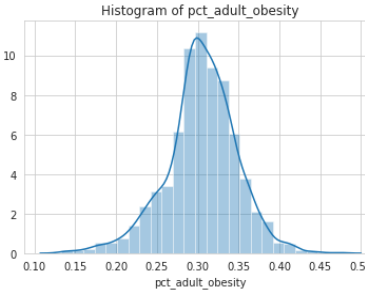
	count	mean	std	min	median	max
pct_uninsured_adults	1560	0.220	0.068	0.053	0.216	0.520
pct_uninsured_children	1560	0.089	0.041	0.018	0.079	0.330
pct_adult_obesity	1560	0.305	0.044	0.133	0.306	0.470
pct_adult_smoking	1344	0.212	0.064	0.031	0.206	0.510
pct_diabetes	1560	0.107	0.023	0.033	0.105	0.180
pct_low_birthweight	1446	0.083	0.021	0.030	0.080	0.180
pct_excessive_drinking	1100	0.165	0.051	0.032	0.163	0.420
pct_physical_inactivity	1560	0.277	0.053	0.104	0.281	0.450
air_pollution_particulate_matter_value	1542	11.637	1.534	7.209	11.907	14.990
homicides_per_100k	613	5.752	4.298	-0.080	4.540	26.920
motor_vehicle_crash_deaths_per_100k	1372	21.72	10.72	3.14	20.29	110.45
heart_disease_mortality_per_100k	1562	275.48	57.83	76.00	270.00	511.00
pop_per_dentist	1447	3 422	2 539	340	2 730	25 169
pop_per_primary_care_physician	1448	2 508	1 960	279	1 970	16 740

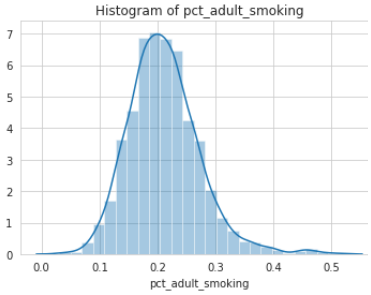
Notes:

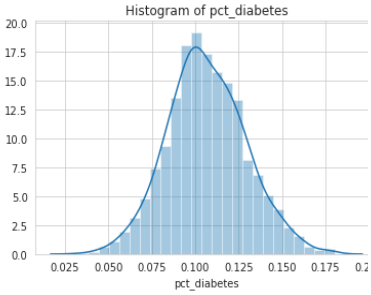
- There are missing values almost for each feature.
  - Some with a very high number of missing values (i.e. **homicides_per_100k**) and we might have to simply ignore the entire feature.
  - Some with more acceptable number of missing value (i.e. **air_pollution_particulate_matter_value**) and we should decide of an imputation policy for those (i.e. replace by the mean value when missing).
  - There are a repetitive number of features where only 2 values are missing. We can suspect to have 2 rows with a lot of missing values, and it may then be more convenient to simply remove those 2 rows for the training.
- There are some inconsistent values for **homicides_per_100k** (negative values)

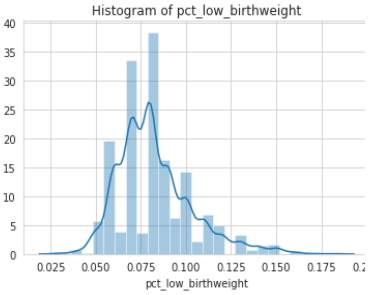

Histogram of pct_uninsured_adults
pct_uninsured_adults

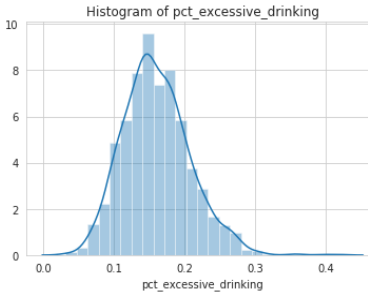

Histogram of pct_uninsured_children
pct_uninsured_children

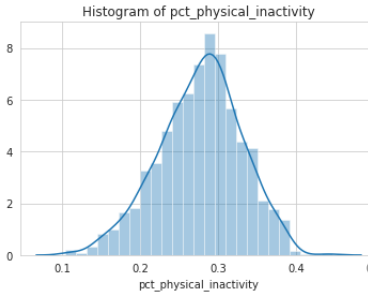

Histogram of pct_adult_obesity
pct_adult_obesity

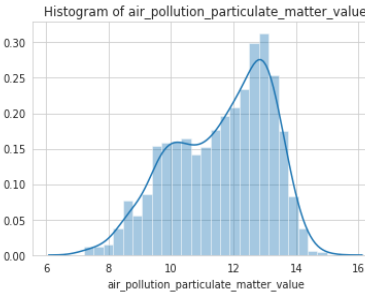

Histogram of pct_adult_smoking
pct_adult_smoking

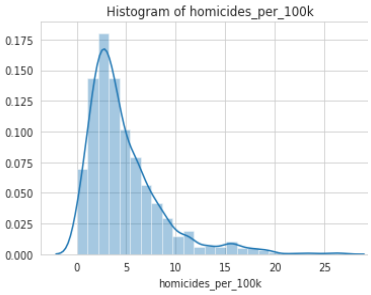

Histogram of pct_diabetes
pct_diabetes

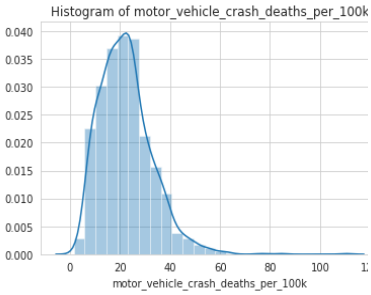

Histogram of pct_low_birthweight
pct_low_birthweight

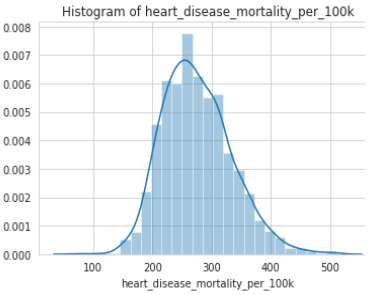

Histogram of pct_excessive_drinking
pct_excessive_drinking

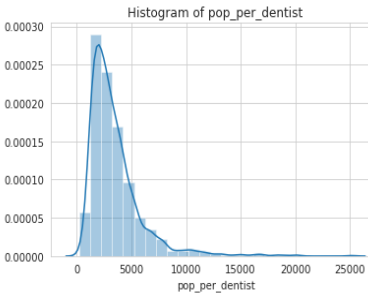

Histogram of pct_physical_inactivity
pct_physical_inactivity

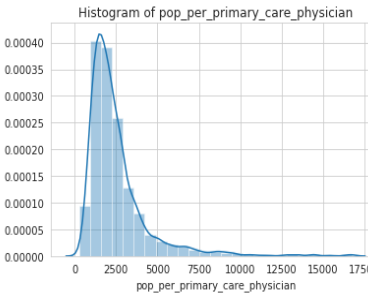

Histogram of air_pollution_particulate_matter_value
air_pollution_particulate_matter_value


Histogram of homicides_per_100k
homicides_per_100k


Histogram of motor_vehicle_crash_deaths_per_100k
motor_vehicle_crash_deaths_per_100k


Histogram of heart_disease_mortality_per_100k
heart_disease_mortality_per_100k


Histogram of pop_per_dentist
pop_per_dentist

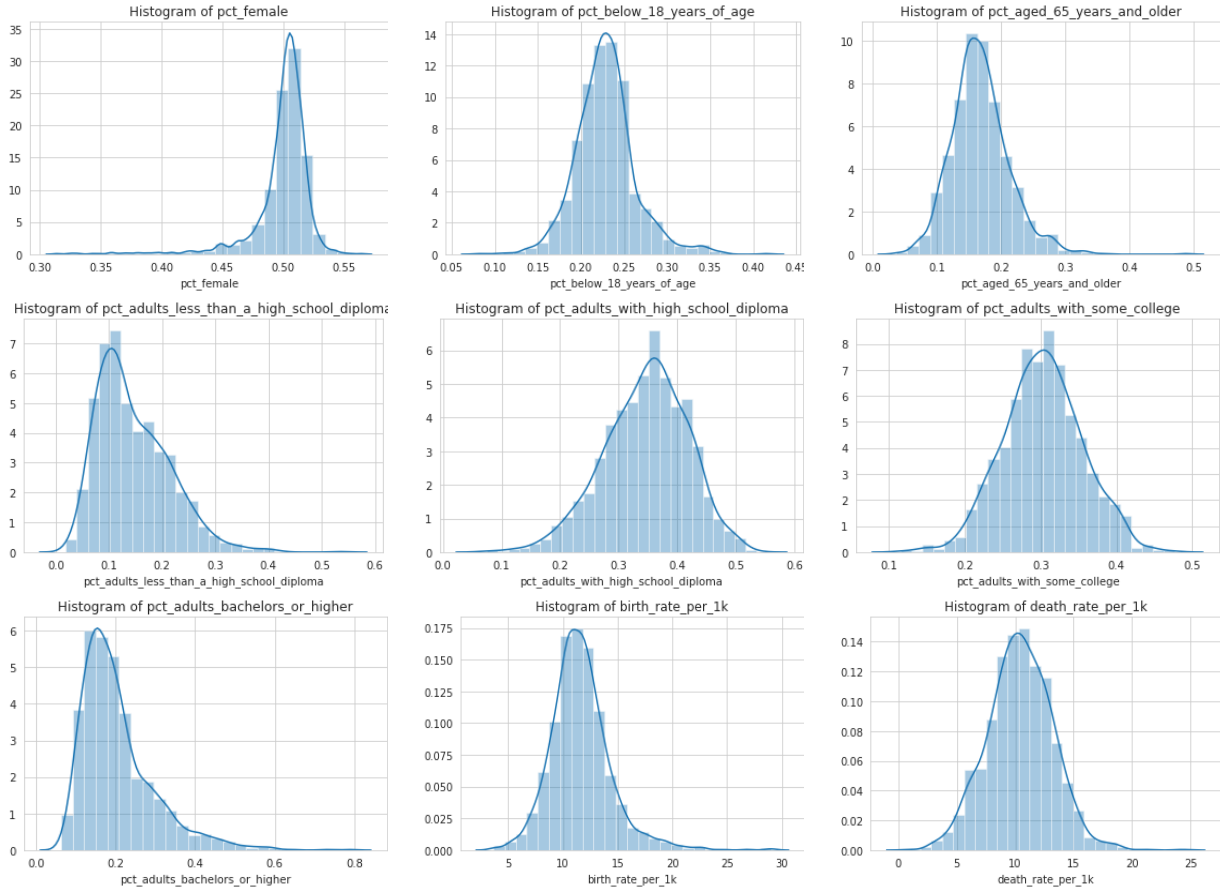

Histogram of pop_per_primary_care_physician
pop_per_primary_care_physician

DEMOGRAPHIC

	count	mean	std	min	median	max
pct_female	1560	0.499	0.024	0.314	0.504	0.560
pct_below_18_years_of_age	1560	0.229	0.035	0.082	0.227	0.420
pct_aged_65_years_and_older	1560	0.168	0.045	0.036	0.164	0.490
pct_adults_less_than_a_high_school_diploma	1562	0.146	0.067	0.019	0.129	0.540
pct_adults_with_high_school_diploma	1562	0.346	0.071	0.074	0.352	0.540
pct_adults_with_some_college	1562	0.303	0.052	0.114	0.303	0.480
pct_adults_bachelors_or_higher	1562	0.205	0.092	0.064	0.182	0.790
birth_rate_per_1k	1562	11.62	2.76	3.65	11.44	29.03
death_rate_per_1k	1562	10.42	2.77	0.96	10.40	24.28

Notes:

- Same observation than for the features from the health category. There are a repetitive number of features where only 2 values are missing. We can suspect to have 2 rows with a lot of missing values, and it may then be more convenient to simply remove those 2 rows (for training phase).



Correlation between numeric features

The figure below shows the heatmap of the correlation between the different features. Red represents a high positive correlation et green a moderately high negative correlation. A pale color represents a weak correlation.

Notes:

- The heatmap shows some strong positive and negative correlation between features. This observation will lead us to consider some dimensionality reduction and/or features selection before we feed our model. Also, this observation will influence the model type that we will choose.

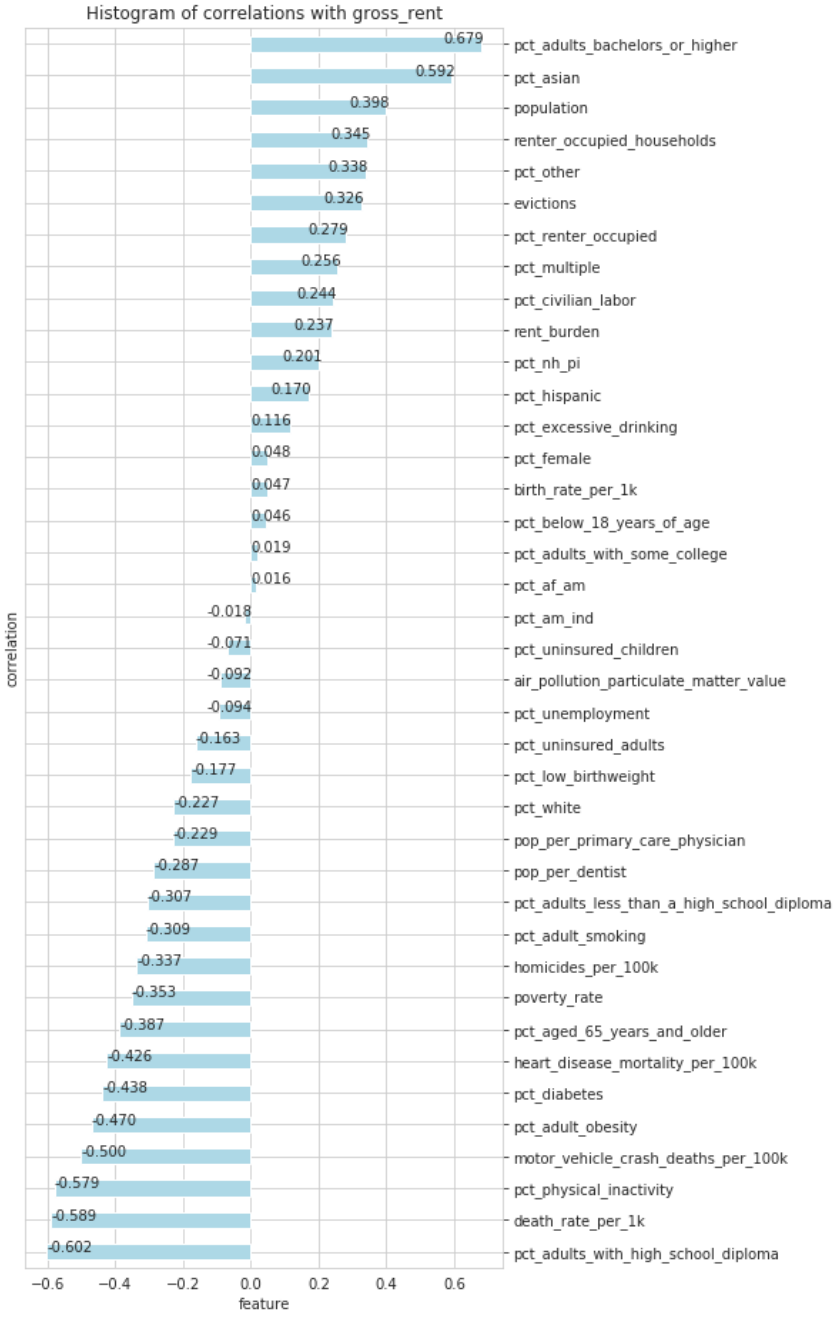
Relationships of features with the target

Correlations of numerical features with the target

The figure beside shows the correlation between each numerical feature and the **gross_rent**.

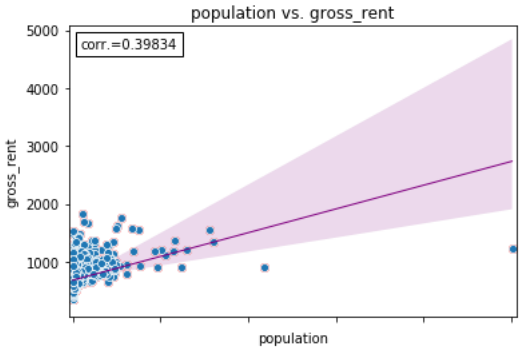
We can observe that 2 education related features are the most positively and negatively correlated.

Also, there are several slightly negatively correlated features that are related to health.

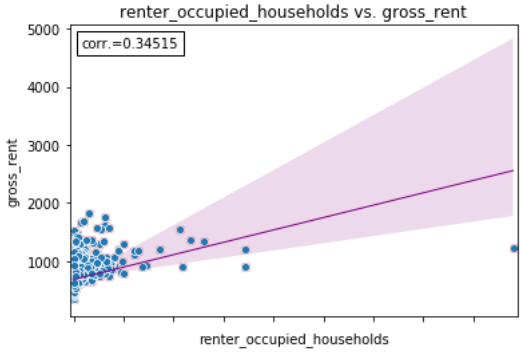


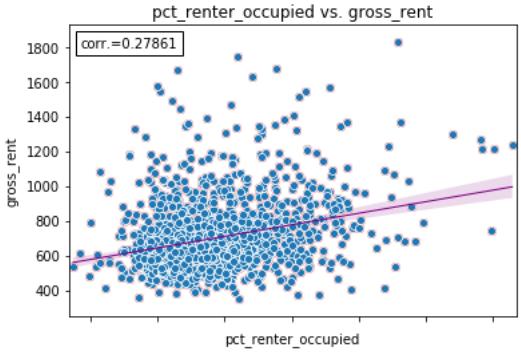
Visualization of features against the target variable

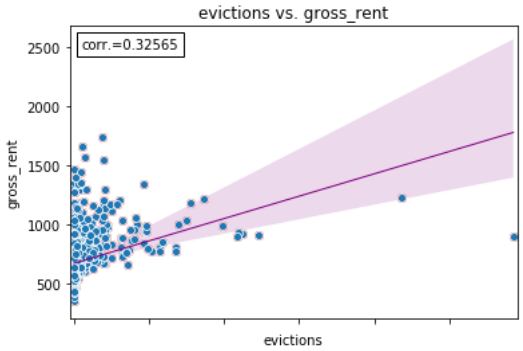
ID

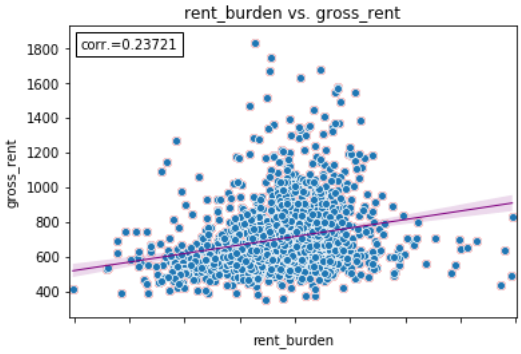


HOUSING

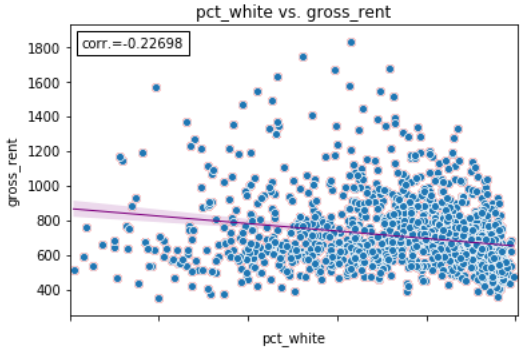


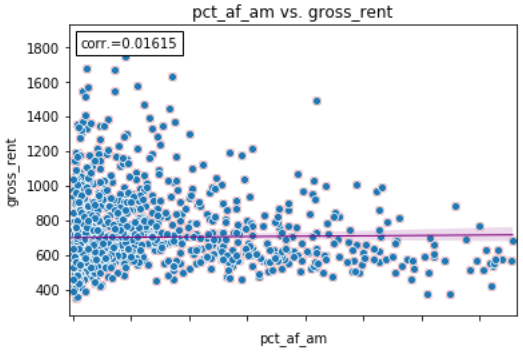


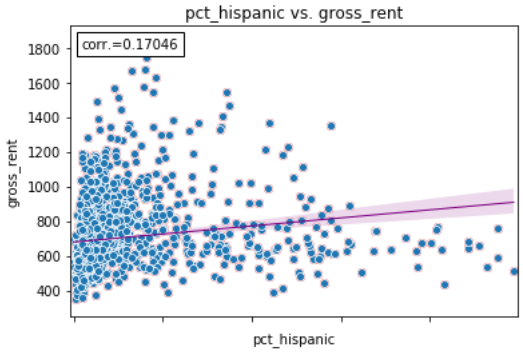


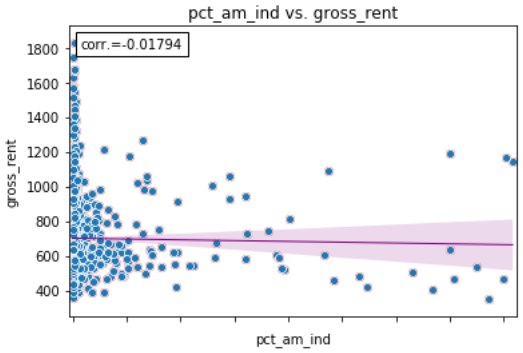


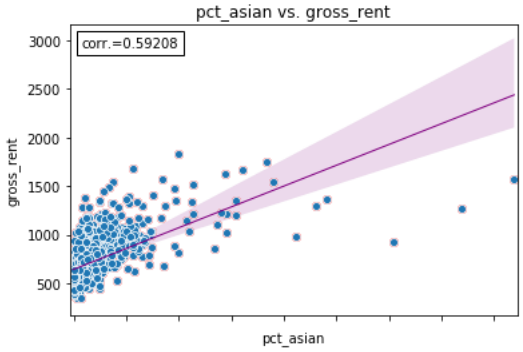
ETHNICITY

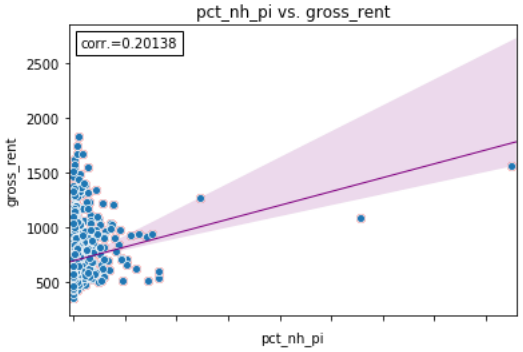

pct_white vs. gross_rent
corr.=-0.22698
gross_rent
pct_white

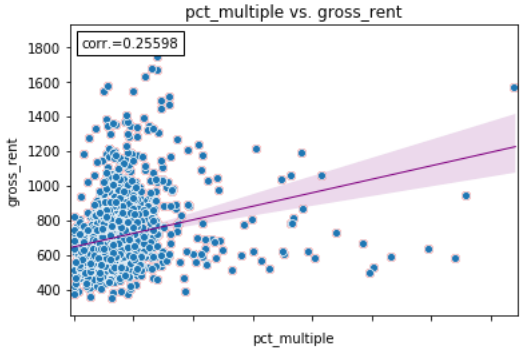

pct_af_am vs. gross_rent
corr.=0.01615
gross_rent
pct_af_am

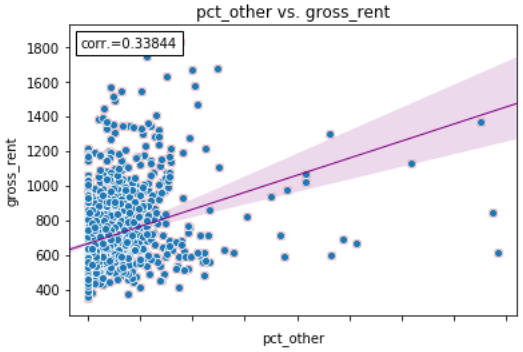

pct_hispanic vs. gross_rent
corr.=0.17046
gross_rent
pct_hispanic


pct_am_ind vs. gross_rent
corr.=-0.01794
gross_rent
pct_am_ind

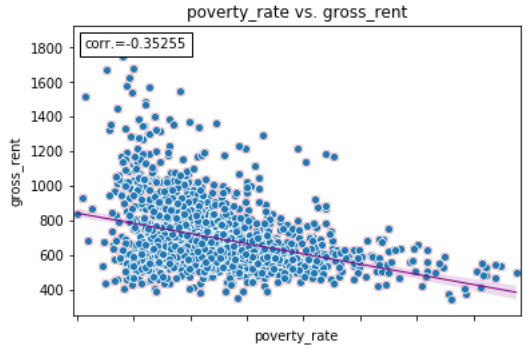

pct_asian vs. gross_rent
corr.=0.59208
gross_rent
pct_asian

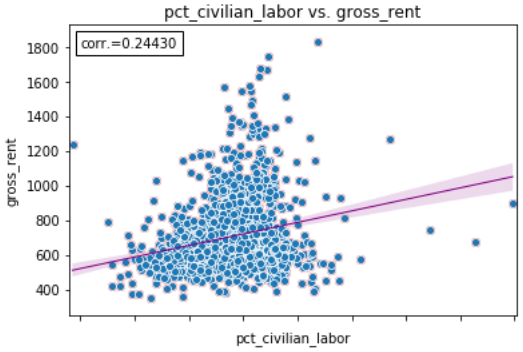

pct_nh_pi vs. gross_rent
corr.=0.20138
gross_rent
pct_nh_pi

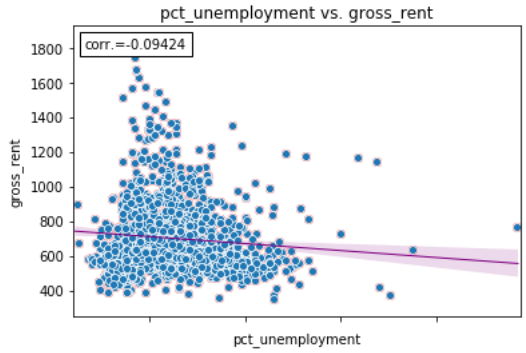

pct_multiple vs. gross_rent
corr.=0.25598
gross_rent
pct_multiple


pct_other vs. gross_rent
corr.=0.33844
gross_rent
pct_other

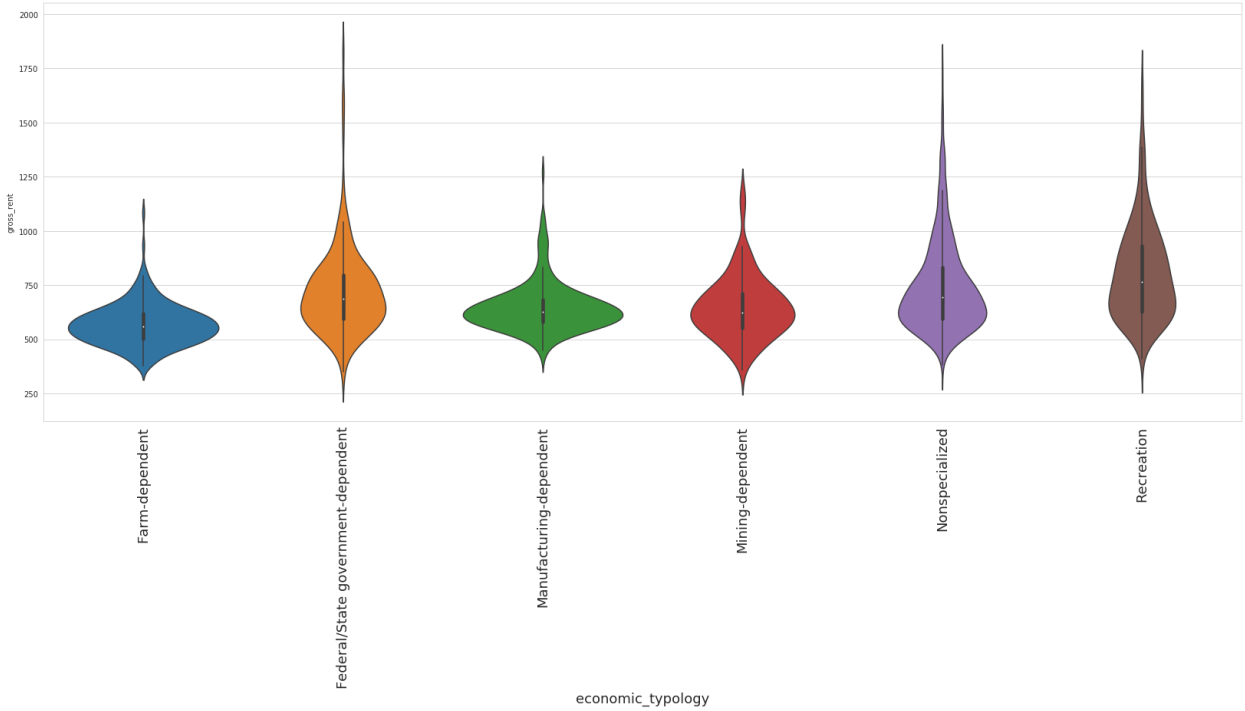
ECONOMIC (Numeric)

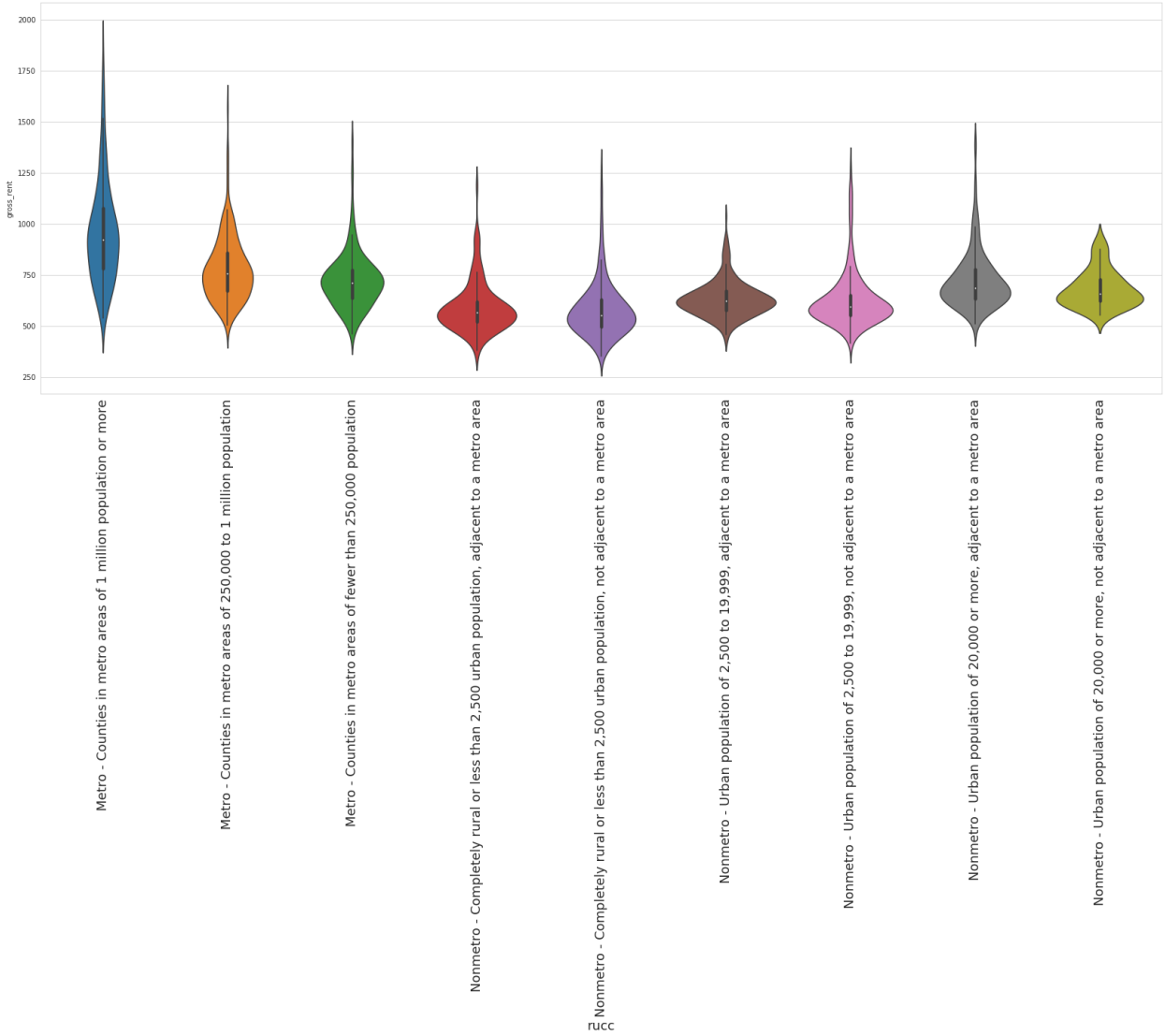


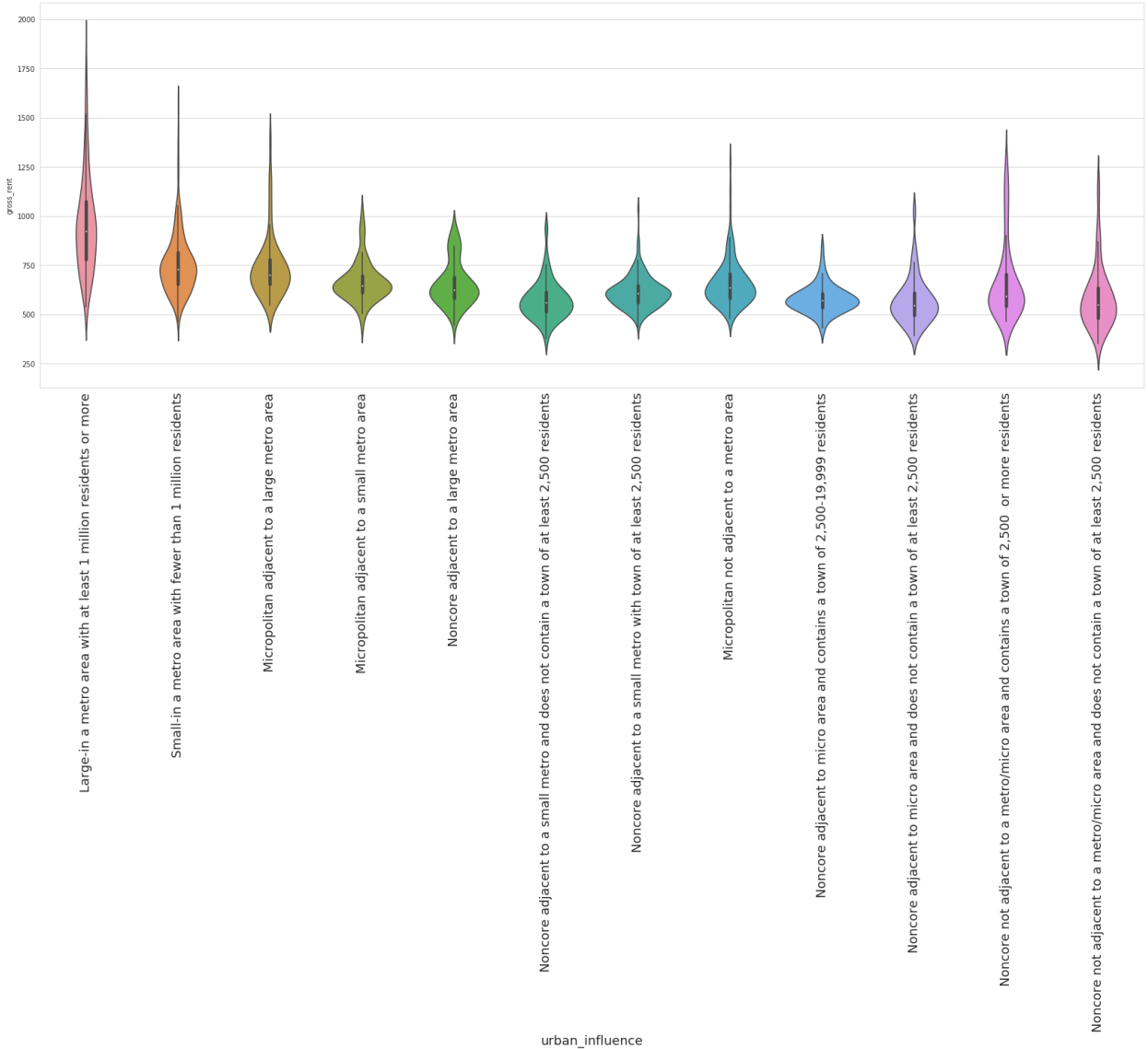




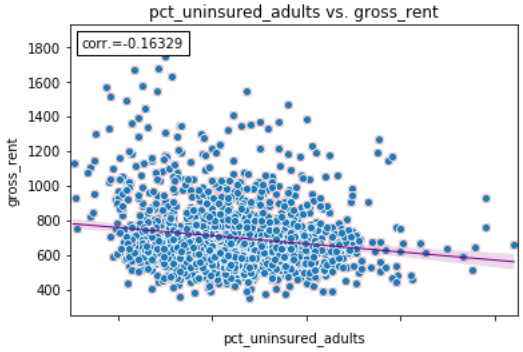
ECONOMIC (Categorical)

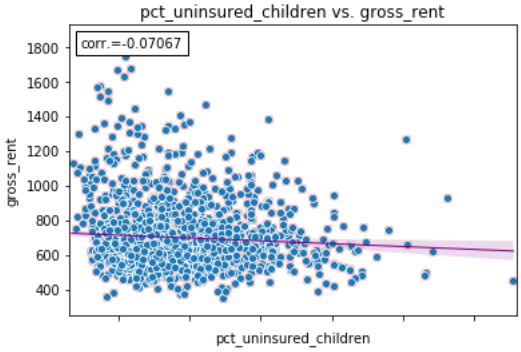


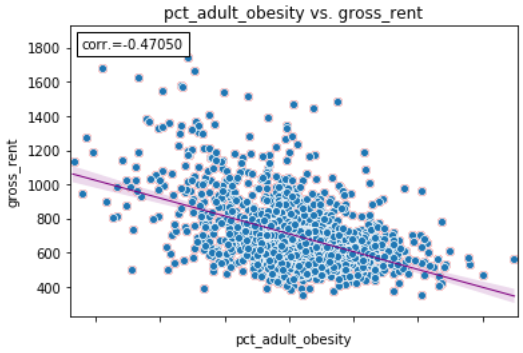

gross_rent
2000
1750
1500
1250
1000
750
500
250
Metro - Counties in metro areas of 1 million population or more
Metro - Counties in metro areas of 250,000 to 1 million population
Metro - Counties in metro areas of fewer than 250,000 population
Nonmetro - Completely rural or less than 2,500 urban population, 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
rucc

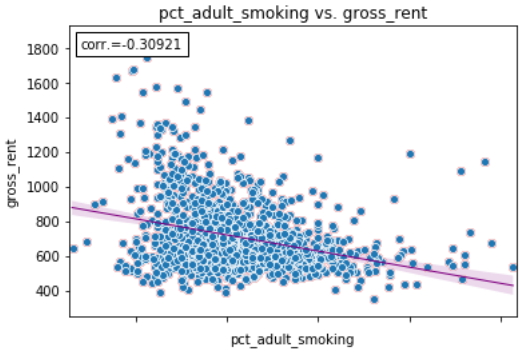

gross_rent
2000
1750
1500
1250
1000
750
500
250
Large-in a metro area with at least 1 million residents or more
Small-in a metro area with fewer than 1 million residents
Micropolitan adjacent to a large metro area
Micropolitan adjacent to a small metro area
Noncore adjacent to a large metro area
Noncore 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
Micropolitan not adjacent to a metro area
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
urban_influence

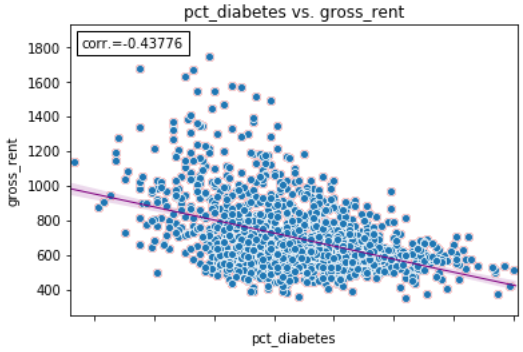
HEALTH

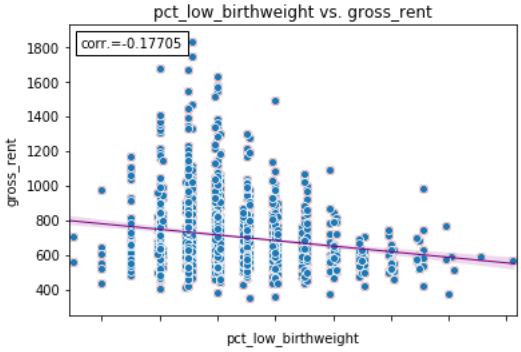

pct_uninsured_adults vs. gross_rent
corr.=-0.16329
gross_rent
pct_uninsured_adults

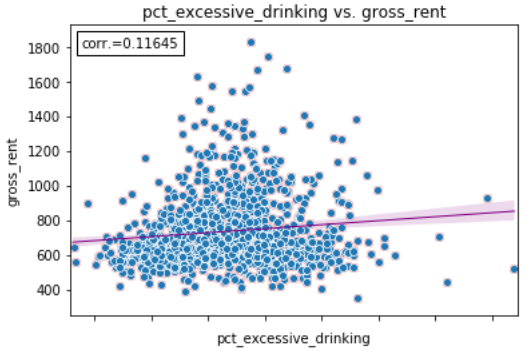

pct_uninsured_children vs. gross_rent
corr.=-0.07067
gross_rent
pct_uninsured_children

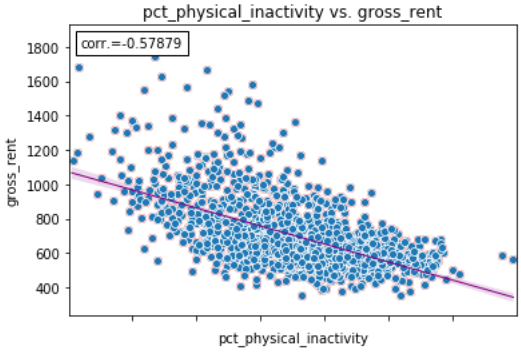

pct_adult_obesity vs. gross_rent
corr.=-0.47050
gross_rent
pct_adult_obesity

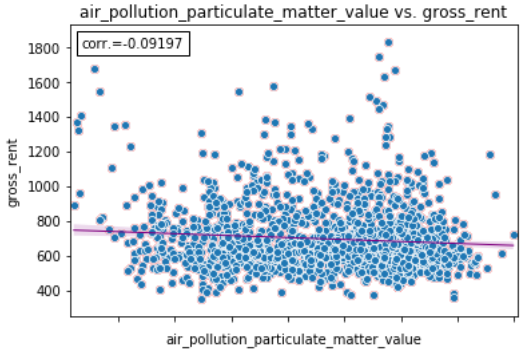

pct_adult_smoking vs. gross_rent
corr.=-0.30921
gross_rent
pct_adult_smoking

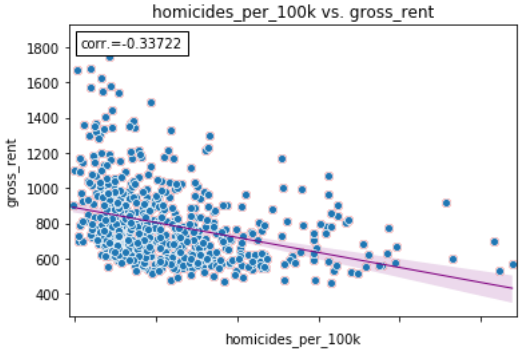

pct_diabetes vs. gross_rent
corr.=-0.43776
gross_rent
pct_diabetes

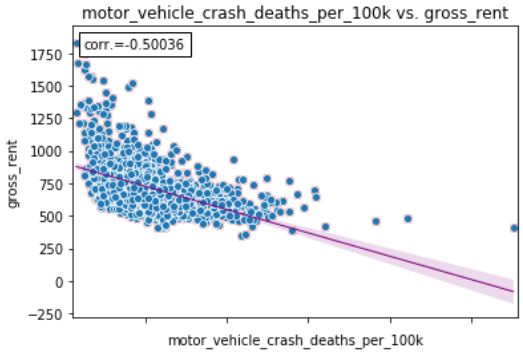

pct_low_birthweight vs. gross_rent
corr.=-0.17705
gross_rent
pct_low_birthweight

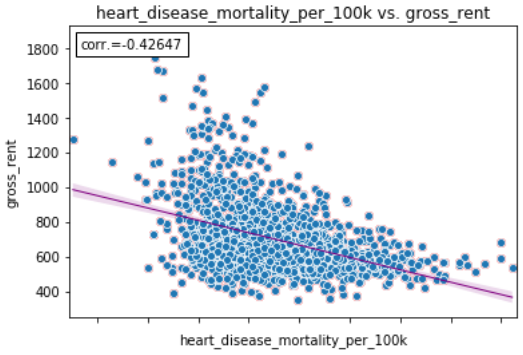

pct_excessive_drinking vs. gross_rent
corr.=0.11645
gross_rent
pct_excessive_drinking

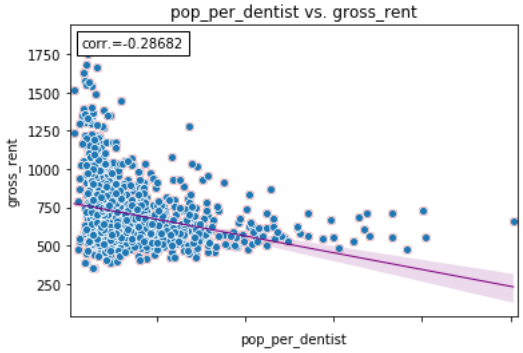

pct_physical_inactivity vs. gross_rent
corr.=-0.57879
gross_rent
pct_physical_inactivity

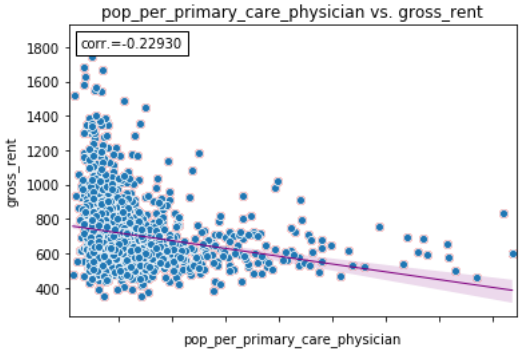

air_pollution_particulate_matter_value vs. gross_rent
corr.=-0.09197
gross_rent
air_pollution_particulate_matter_value


homicides_per_100k vs. gross_rent
corr.=-0.33722
gross_rent
homicides_per_100k

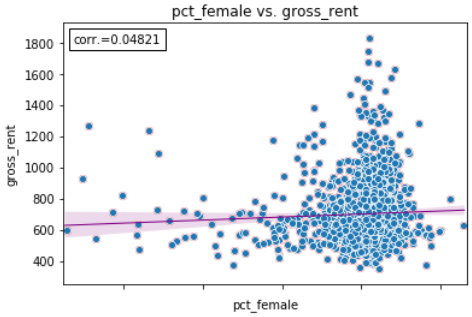

motor_vehicle_crash_deaths_per_100k vs. gross_rent
corr.=-0.50036
gross_rent
motor_vehicle_crash_deaths_per_100k

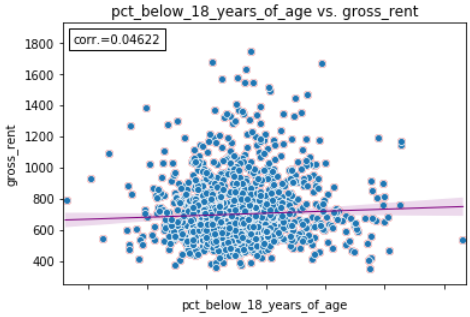

heart_disease_mortality_per_100k vs. gross_rent
corr.=-0.42647
gross_rent
heart_disease_mortality_per_100k

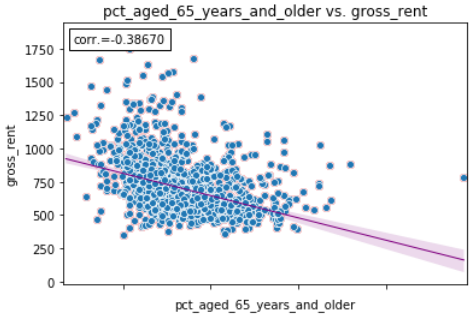

pop_per_dentist vs. gross_rent
corr.=-0.28682
gross_rent
pop_per_dentist

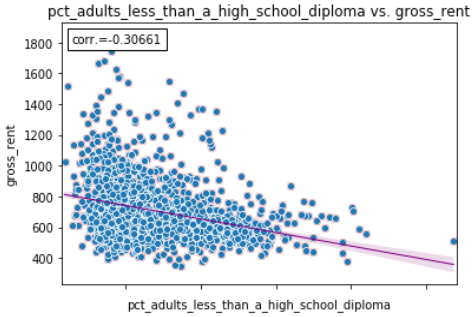

pop_per_primary_care_physician vs. gross_rent
corr.=-0.22930
gross_rent
pop_per_primary_care_physician

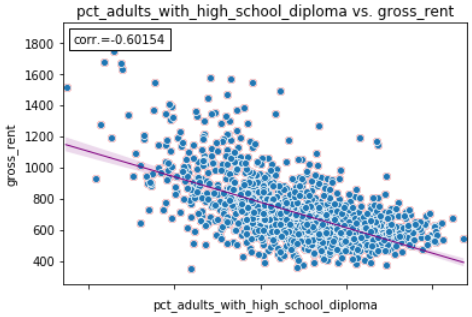
DEMOGRAPHIC

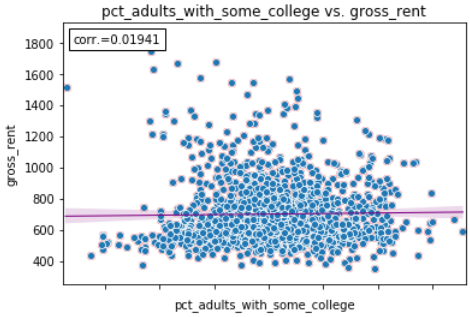


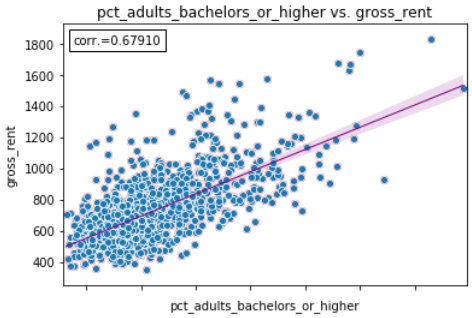


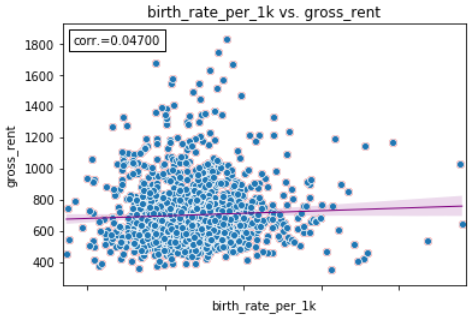


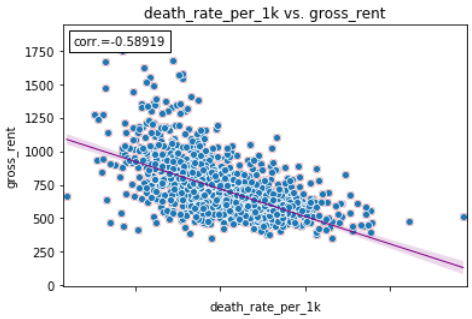












Notes

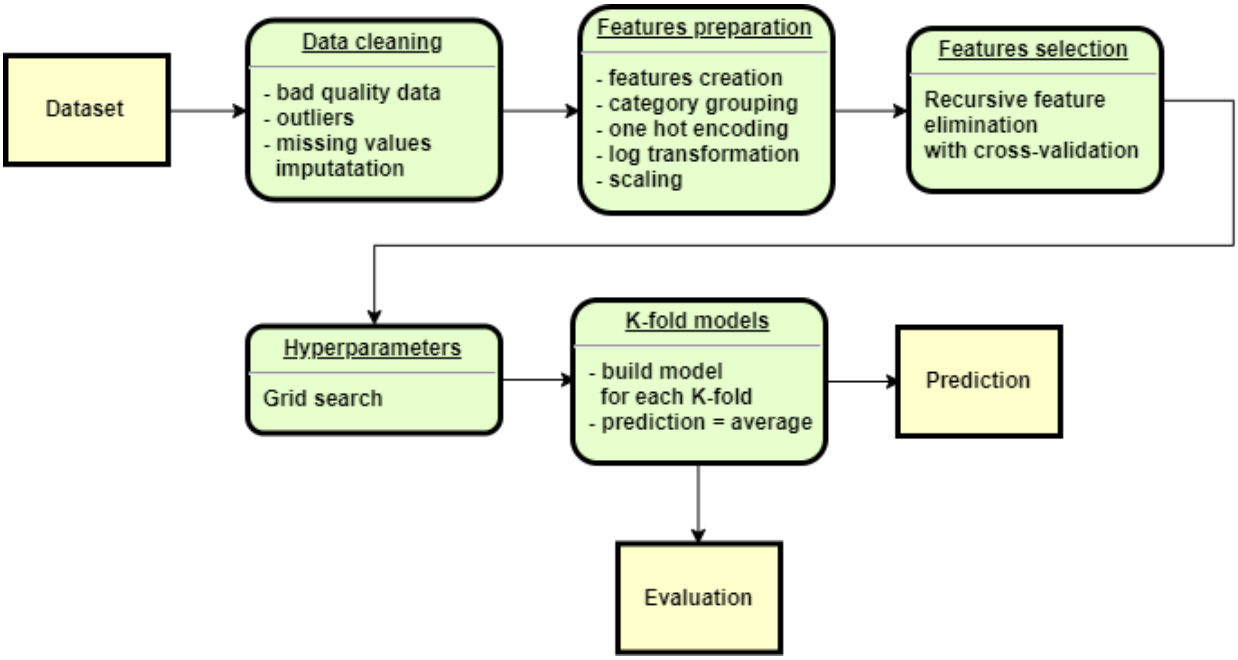
1. For some features like **motor_vehicle_crash_deaths_per_100k, pct_aged_65_years_and_older**, or **pct_nh_pi** we can spot a couple of outliers. This can act like noise for our model and should be addressed. There are different ways to fix this: in the choice of a model and or a scaling method that is outliers-resistant, or simply by removing the outliers.
2. For the 3 categorical features, considering our previous observation that the features are not well balanced, our proposal is to group categories that have a similar relationship with the target.
3. In each category of data, we can already detect promising features. Especially, there are features that have a good correlation with the target (negative or positive) and, when plot against our **gross_rent**, they show a clear apparent relationship. Some examples per category are given in table below:

Category	*Feature*
Ethnicity	pct_asian
Economic	rucc: Metro - Counties in metro areas of 1 million population or more urban_influence: Large-in a metro area with at least 1 million residents or more
Health	pct_adult_obesity pct_physical_inactivity
Demographic	pct_adults_with_high_school_diploma pct_adults_bachelors_or_higher death_rate_per_1k

Regression modeling

Data Science Process

The diagram below describes our data science process from the dataset to the model. It shows a very linear flow from one step to another, but, in reality, there could be several iterations of each step with forward and backward movement in the process. In fact, some findings while working on one step, can bring new insights that requires to modify a previous step. This kind of process is common in Data Science methodology and is also how more findings and knowledge are extracted from the data.



Data cleaning and features preparation

This part is more important that it could seem at first glance as good data is the ground and what our model is built. Important steps we have covered here are:

- **Data cleaning**
  - Incorrect data: we had some negative values for variables that can't be negative. We have applied the same treatment than for missing values.
  - Outliers: In a first attempt, they were removed from dataset, but we finally kept them as the dataset is quite small. We have favored the use of a "robust scaler" later in the process that is more resistant to outliers.
  - Missing data imputation: we have used a tool that implement multivariant imputation. Each missing value has then been estimated given all the other data available.

- **Features preparation**
    - Features creation: from the existing features we have computed several other variables: we divided by the population for features that were given in full value and not in proportion, we took the complementary values (1-p) for rate values that were highly left skewed, and we log transformed all variables with a right skewed distribution.
    - Category grouping and one-hot-encoding: For all categorical features, we made some grouping using the category maps detailed below:

economic_typology

Original categories	*Categories group*
Manufacturing-dependent	Manufacturing_ Mining _Farm-dependent
Mining-dependent	
Farm-dependent	
Nonspecialized	Nonspecialized
Federal/State government-dependent	
Recreation	Recreation

urban_influence

Original categories	*Categories group*
Large-in a metro area with at least 1 million residents or more	Large-in
Small-in a metro area with fewer than 1 million residents	Small-in and adjacent
Micropolitan adjacent to a large metro area	
Noncore not adjacent to a metro/micro area and contains a town of 2,500 or more residents	Not adjacent with metro/micro
Noncore not adjacent to a metro/micro area and does not contain a town of at least 2,500 residents	
Micropolitan not adjacent to a metro area	
Noncore adjacent to a large metro area	Adjacent with metro/micro
Micropolitan adjacent to a small metro	
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 a small metro and does not contain a town of at least 2,500 residents	not at least 2,500 residents
Noncore adjacent to micro area and does not contain a town of at least 2,500 residents	

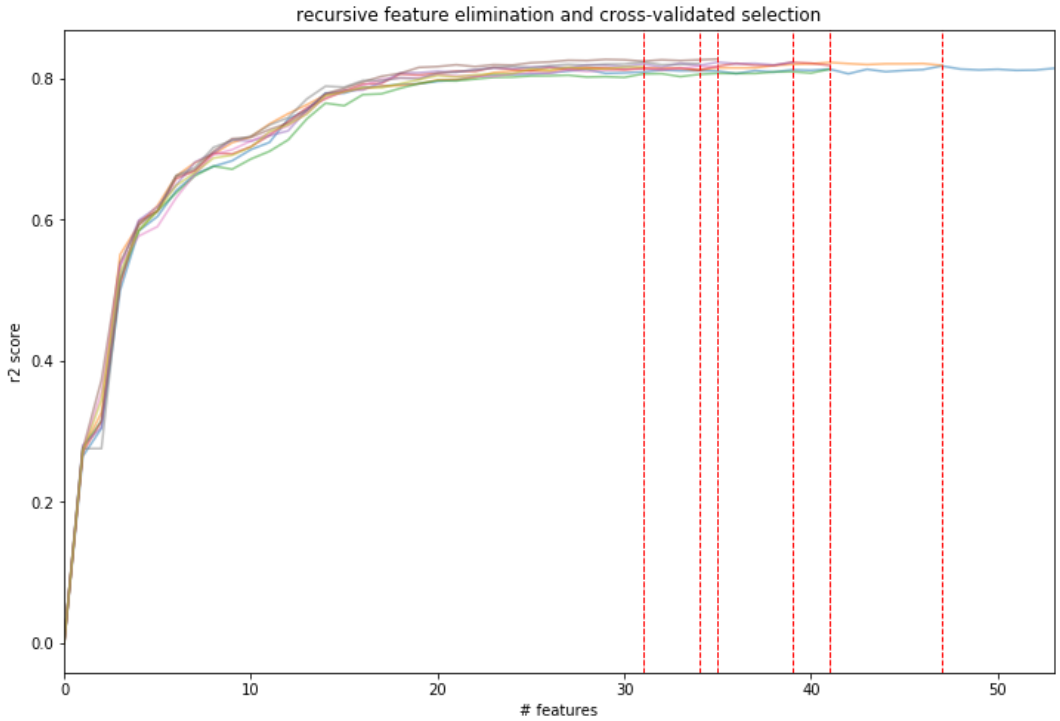
rucc

Original categories	*Categories group*
Metro - Counties in metro areas of 1 million population or more	Metro big
Metro - Counties in metro areas of 250,000 to 1 million population	Metro medium
Metro - Counties in metro areas of fewer than 250,000 population	Metro small Nonmetro - 20
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	
Nonmetro - Urban population of 2,500 to 19,999, adjacent to a metro area	Nonmetro - 2,5- not adjacent
Nonmetro - Urban population of 2,500 to 19,999, not adjacent to a metro area	Nonmetro - 2,5 - adjacent
Nonmetro - Completely rural or less than 2,500 urban population, not adjacent to a metro area	Nonmetro - rural
Nonmetro - Completely rural or less than 2,500 urban population, adjacent to a metro area	

Features selection

To select our features, we used recursive feature elimination and cross validation (RFECV) selection. We have performed several iterations recursively also until no features were eliminated anymore. Our hypothesis is that, as we choose a non-deterministic model, there was a bit of randomness in the process and we had re-apply the whole process several time until a final and stable set of features was selected. The graph below represents each RFECV process and the R^2 score progression vs the number of features. The final set of features is composed of, by category:

- **ID**: population
- **Housing**: rent_burden
- **Ethnicity**: pct_white, pct_af_am, pct_hispanic, pct_am_ind, pct_asian, pct_nh_pi, pct_multiple, pct_other
- **Economic**: poverty_rate, rucc ("*Metro big*", "*Nonmetro - rural*"), urban_influence ("*Large-in*", "*Small-in and adjacent*"), economic_typology ("*Manufacturing_ Mining _Farm-dependent*", "*Recreation*"), pct_civilian_labor, pct_unemployment
- **Health**: pct_uninsured_children, pct_adult_obesity, pct_adult_smoking, pct_diabetes, pct_physical_inactivity, air_pollution_particulate_matter_value, motor_vehicle_crash_deaths_per_100k, heart_disease_mortality_per_100k
- **Demographic**: pct_aged_65_years_and_older, pct_adults_with_high_school_diploma, pct_adults_bachelors_or_higher, death_rate_per_1k



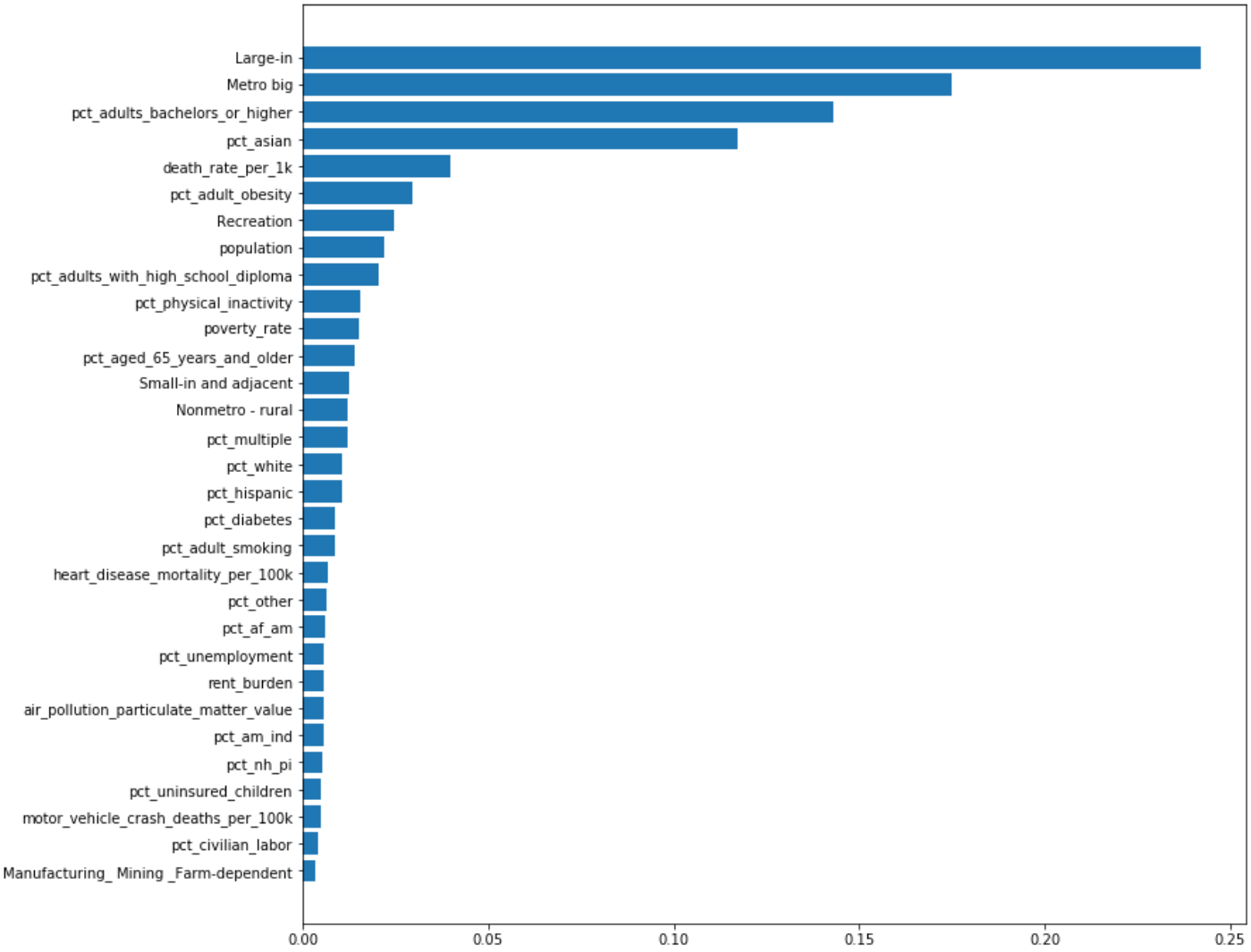
Regression model

Creation

We have chosen a gradient boosted tree model ("XGBoost") as it is a very powerful and reliable algorithm. It has been widely used in Data Science competition and seems to give great results even on a small dataset like ours. We have performed grid search to optimize the hyperparameters, and also, we have generated 5 parallel models using a 5 folds cross-validation. This has been done to extract most of the value of our small dataset and also to remove a bit of variance by averaging the results of the 5 models.

Features importance

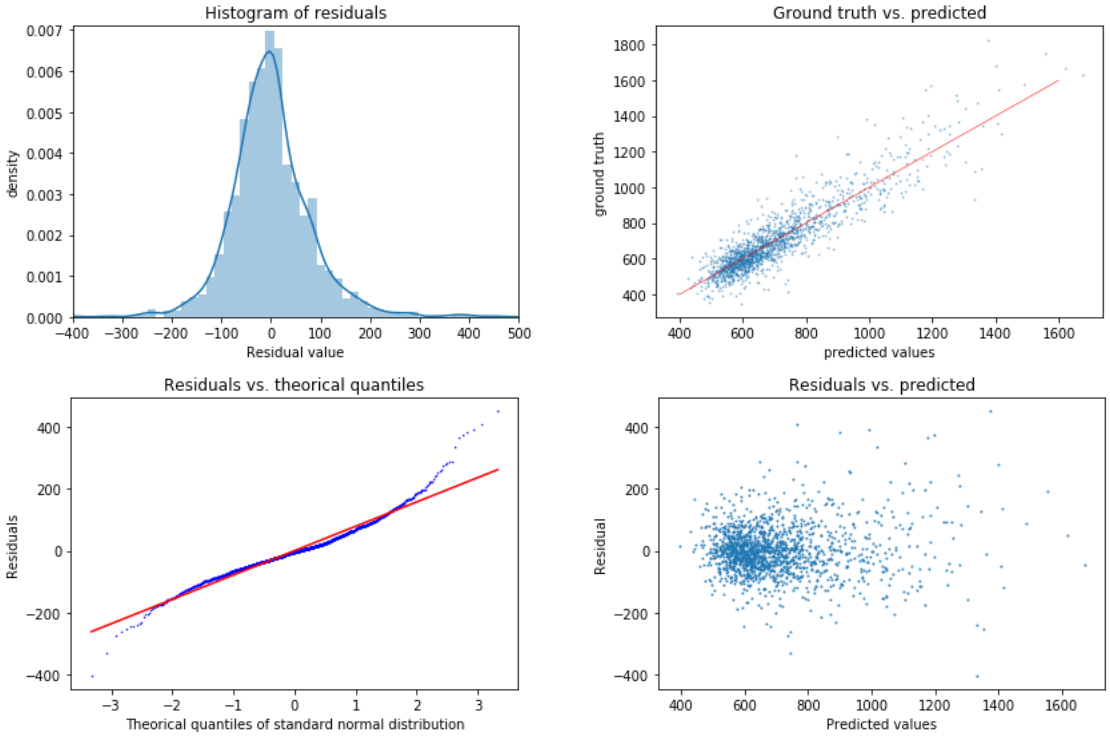
One usefull property of this kind of model is that you can extract the relative importance of each feature in the model. Even if this can of analysis should be consider with caution, the results found tend to match well with what we observed previously. The histogram of the features and their relative importance is given below.



Evaluation

Evaluated using the R^2 metric and a 5-fold cross-validation, the model scores 0.817 on a mean value (with a standard deviation of 0.031). The R^2 score is of a value of 1 at the maximum and of a value of 0 for a random guess. Therefore, we are confident to interpret this score as a quite good capacity of the model to predict the target from the given variables.

To go further, we have analyzed the residuals using in the plots below. The histogram of residuals shows a close to normal distribution of the residuals that is confirmed by the QQ plot just under. There is a slight right skewness in the distribution with a little more and little higher positive residuals. We can also sport a little bump on the right side around +100. All this means that the model tends to underpredict the target more (in value and frequency). We can also notice that on the plot showing ground truth against predicted values (more points are over the red line). Finally, the last plot of residuals against the predicted values at the bottom right, shows a homogenous spread that confirm our first analysis of the residuals.



Conclusion

With this project, we have seen that the median gross rent in a council can be predicted using variables related to this council taken from various categories like housing, ethnicity, economic, health and demographic. The statistical analysis gave us very interesting insights at first, but we have seen that building the regression model and the whole Data Science process help us build more knowledge about the data all along the way.

It's not surprising to observe that being in presence of a great metropolitan area or not, has a major influence on the predicted value. We have seen that through the importance given by the regression model to the belonging of urban area category ("Large-in a metro area with at least 1 million residents or more", "Metro - Counties in metro areas of 1 million population or more").

We have also seen that variables related to education are important for our model through the importance given to percentage of adults bachelors or higher, and less the percentage of adult with high school diploma.

Further, we have seen a group of features that could be linked to health and life expectancy starting with the death rate, including obesity and physical inactivity to the percentage of people aged more than 65 years old.

Lastly, we have seen that variables from the ethnicity category are present. We have noticed that the percentage of Asian people is among the more important features to predict the median gross rent. One hypothesis could be that the variable acts as a proxy to some other variable. We have noticed that this the percentage of Asian is also correlated with population. But even if the other ethnicity variables have a moderate influence, the presence of all of them in the list of important features give them a global significant importance.

Table of Contents