Exploring Gentrification in NYC

Gentrification

Gentrification is the process of change in a neighborhood through increase in more affluent residents & businesses. It is a controversial topic in urban planning as it can lead to population migration and displacement.

Predictors used in studies on Gentrification

University of Colorado (2019)	Kings College of London (2018)	RentCafe (2018)
Access to jobs	Household Income	Home Value Change
Proximity to transit	Real Estate Values	Household Income Change
Quality of housing	Occupational Share	Higher Education Change
Diversity of neighborhood	Job Qualifications	

Median Income

N = 821   Y = Median Income

X Variables	Description
Rent	Median household rent
Severe Rent Burden	Percentage of Renter households who spend 50% or more of their income on rent
Overcrowded Rental Housing	Percentage of Occupied rental housing units with more than one person per room
Home Ownership	Percentage of residents who own their own unit
Employment Ratio	Proportion of the population aged 16 and over that are currently employed.
Unemployment*	Percentage of residents 16 years and older actively seeking employment
Population*	Number of residents in each neighborhood

*Removed from the model due to lack of statistically significant coefficient
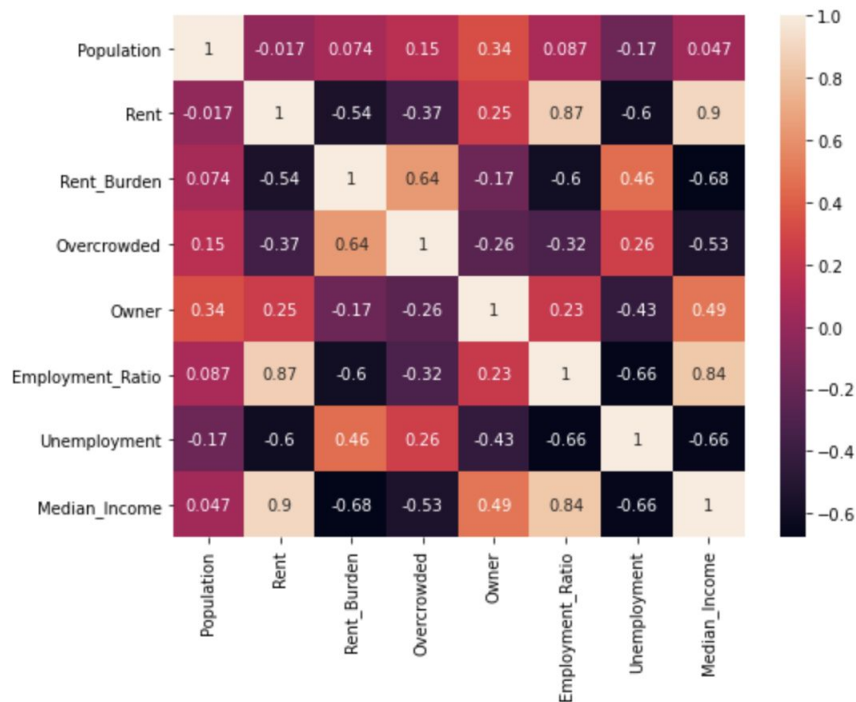
Median Income

Model: Multiple Linear Regression

Data Exploration:
- Correlation matrix and scatterplots to reveal relationships among predictors

Target: Median Income
Predictors:
- Rent
- Rent Burden
- Overcrowded Rental Housing
- Home Ownership
- Employment Ratio

Median Income

Findings:

- **Rent**, **Rent Burden**, **Overcrowded Rental Housing, Home Ownership**, and **Employment Ratio** were statistically significant predictors of the median income for a NYC neighborhood in a given year, and the model was statistically significant at the .001 level ($F(5, 569) = 1190$). The R^2 value for both the test and training sets was .94.
- A model with percent change in these predictors was statistically significant ($F(5, 569) = 53.54$, $p < .001$) but explained less variance in median income ($R^2 = .38$).
- Finally, I used the stepAIC function in the MASS package in R to select an optimal model according to Bayesian Information Criterion (BIC) tuning. This model included one predictor, **Rent**, with an R^2 of .83 for the training set and .77 for the test set.

Rent and Median Income in NYC Neighborhoods
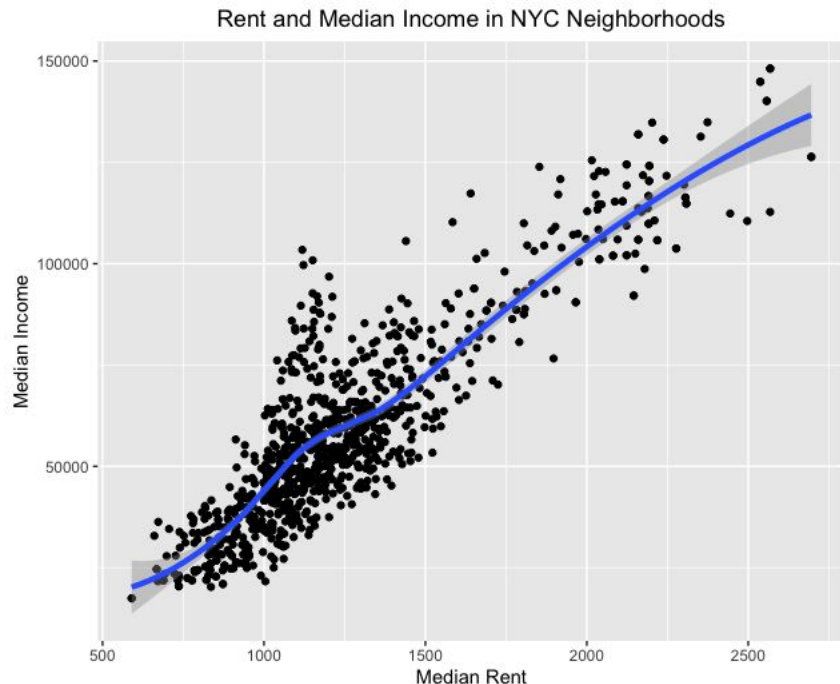
Educational Attainment

N = 826   Y = Percentage of residents in each neighborhood with a Bachelor's Degree or higher

| X Variables	Description
Rent	Median household rent
Severe Rent Burden	Percentage of Renter households who spend 50% or more of their income on rent.
Overcrowded	Percentage of Occupied rental housing units with more than one person per room.
Home Ownership	Percentage of residents who own their own unit.
Employment Ratio	Proportion of the population aged 16 and over that are currently employed.
Unemployment	Percentage of residents 16 years and older actively seeking employment
Median Income	Median household income
Uninsured	Percentage of residents without health insurance
SNAP	Percentage of residents receiving food stamps

Educational Attainment

Method: Linear Regression
1. Explore relationship across all variables

Strong positive correlation between Rent and Median_Income (r = 0.92) as well as Rent and Employment_Ratio (r = 0.86).

Strong negative correlation between Median_Income and Snap_Pct (r = -0.82).
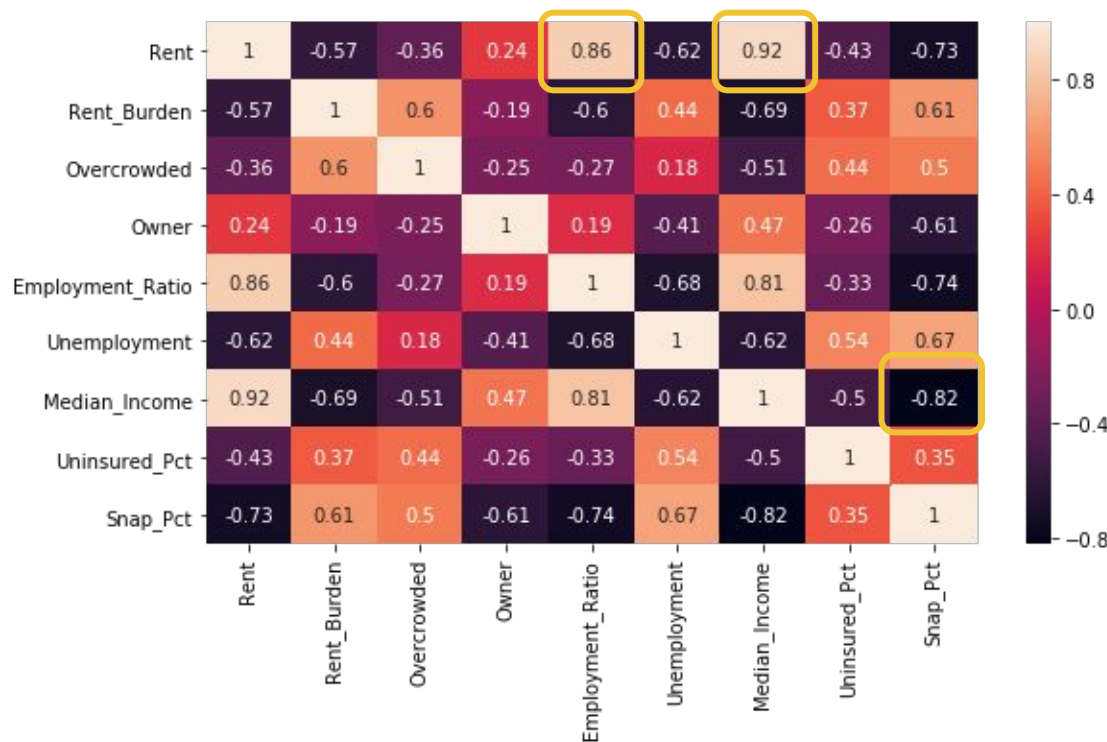
Educational Attainment

Method: Linear Regression
1. Explore relationship across all variables
2. Run regression model

Rent_Burden and Unemployment are not significant and are dropped from model.

R^2 trained= .9099 (trained); R^2 test= .8915

Factors	Coefficient	Std Err	p-value
Rent	0.2770	0.083	0.001
Rent_Burden	-0.0614	0.037	0.101
Overcrowded	-0.0870	0.034	0.011
Owner	-0.3923	0.041	0.000
Employment_Ratio	-0.1254	0.054	0.020
Unemployment	-0.0234	0.040	0.560
Median_Income	0.4691	0.098	0.000
Uninsured_Pct	-0.0917	0.031	0.003
Snap_Pct	-0.3568	0.054	0.000

Median Rent

N = 910   Y = Median Rent in each neighborhood

Feature Importance	X Variables	Description
0.8044	Employment Ratio	Proportion of the population aged 16 and over that are currently employed.
0.0172	White (Race/Ethnicity)	Number of residents of White Race/Ethnicity in each neighborhood
0.0164	Unemployment	Number of residents 16 years and older actively seeking employment
0.0126	Severe Rent Burden	Percentage of Renter households who spend 50% or more of their income on rent.
0.0123	$200,000 or more	Number of Renter households who earn $200,000 or more income

Median Rent

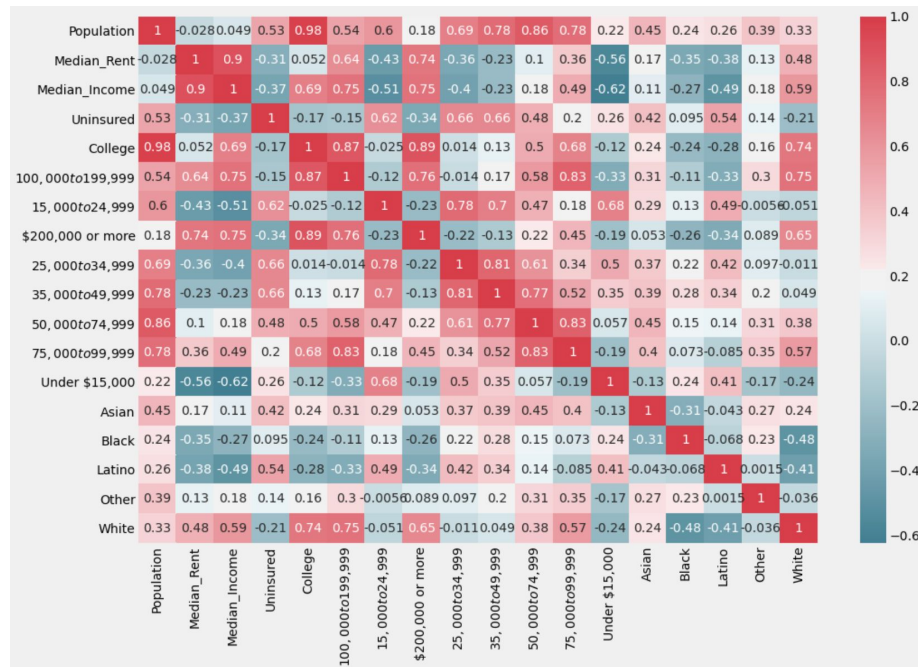**Model:** Random Forest Regression

Data Exploration:
- Correlation matrix and scatterplots to reveal relationships among predictors

Target: Median Rent
Predictors:
- Employment Ratio
- Median Income
- $200,000 or more
- White
- Rent Burden
- Unemployment

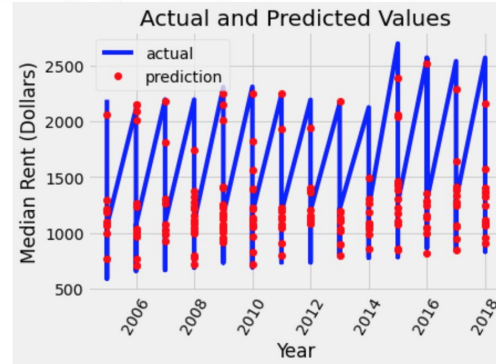# Machine Learning Models
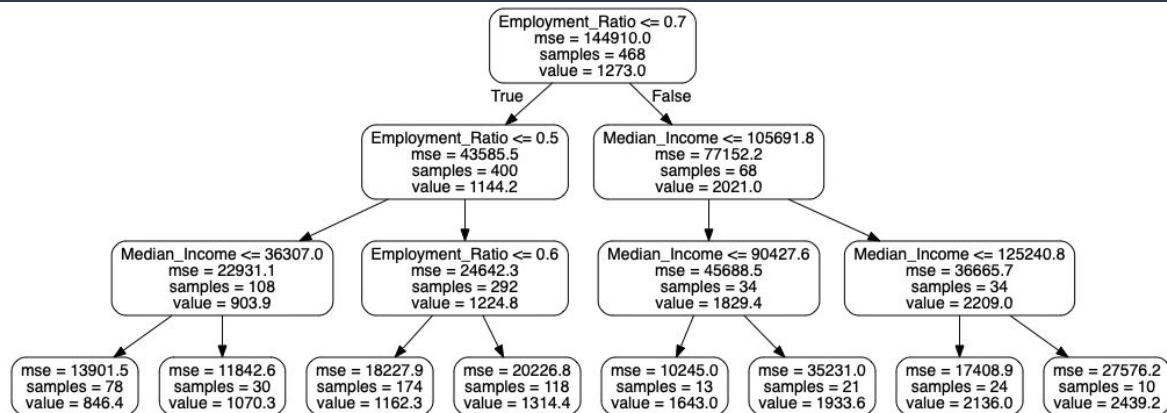
Method 1: Random Forest Regression
- 72 features
- 1000 trees
- MAE: 54.25
- RMSE: 72.21
- Accuracy: 95.49 %

Method 2: Random Forest Regression
- 6 features
- 1000 trees
- MAE: 66.92
- RMSE: 86.03
- Accuracy: 94.44 %

Method 3: GridSearchCV (Fine Tuned Model)
- 8 features
- 100 trees
- RMSE: 13.95

Dataset & ETL Process

Datasource: [Citizens' Community for Children](#)

- Combined all factors by merging data from separate csv files on neighborhood id (FIP) and Year (Battery Park 2005, Battery Park 2006, etc.)

- Cleaned data to remove city and borough level to only use data by neighborhood

- Limitation of the dataset: all data categorized by neighborhoods producing 65 different locations (if categorized by zip code, dataset would have expanded to 177 locations)

Factors	Years
Total Population	**2005-2018**
Median Incomes	**2005-2018**
Median Monthly Rent	**2005-2018**
Severe Rent Burden	**2005-2018**
Homeownership	**2005-2018**
Employment Population Ratio	**2005-2018**
Educational Attainment	**2006-2018**
Unemployment Rate	**2005-2018**
Uninsured	**2008-2018**
Overcrowded Rental Housing	2005-2018
SNAP (Food Stamps)	2001-2019
Public Assistance	2005-2017

Closing

Limitations
- Neighborhood-level data
- Small dataset

Future Directions:
- Time series analysis
- Classification based on a tagged dataset of gentrified neighborhoods
- Include other data associated with gentrification such as changes in number & type of businesses and impact on education