The Impact of Housing Conditions on Monthly Utility Costs in New York City: An Analysis of the 2021 NYCHVS Microdata

Jiayu (Jade) Gu

Internship

APSTA-GE_2310

05/05/2024

Abstract

This study explores the influence of housing conditions on monthly utility costs in New York City by analyzing the 2021 New York City Housing and Vacancy Survey (NYCHVS) Microdata. Utilizing a dataset with 7089 observations and 150 variables, this study employs a range of statistical techniques to offer insights into the complex dynamics between housing attributes and utility costs. The analytical approach includes Lasso Regression for feature selection, General Linear Models (GLMs) to quantify effects, and advanced machine learning methods such as Random Forests and Gradient Boosting Machines for predictive modeling. Additionally, Propensity Score Matching (PSM) is used to examine the impact of tenure status on monthly utility costs. The goal of this study is to provide a detailed analysis that not only enhances understanding of housing-related expenses but also informs policy decisions aimed at improving residential living conditions in urban settings. This report reveals significant predictors of utility costs and discusses the implications of these findings for stakeholders and policymakers.

1. Introduction

The quality of housing is a critical component of urban life, profoundly impacting residents' economic stability, comfort, health, and access to care. In New York City, where the urban landscape is as diverse as its population, understanding how various housing conditions influence utility costs is important for both economic and environmental policy making. Housing instability, which includes challenges such as frequent moves, overcrowding, and spending a significant portion of household income on housing, 'may negatively affect physical health and make it harder to access health care' [1].

1.1. Purpose of the Study

This study aims to identify key housing-related factors that significantly affect utility costs in New York City and seeks to highlight potential areas for policy intervention that could lead to more sustainable and cost-effective housing solutions. The findings are intended to assist policymakers, stakeholders, and the public in understanding the dynamics between housing conditions and utility expenditures, thus informing decisions that enhance the quality of life for New York City' s residents.

1.2. Background Information on NYCHVS

The New York City Housing and Vacancy Survey (NYCHVS) is conducted every three years by the U.S. Census Bureau to gather data on the city's housing stock, occupancy status, and resident demographics. The survey provides a comprehensive dataset that reflects the conditions and challenges of housing in one of the world's most dynamic urban environments [2]. The 2021 survey iteration, which serves as the data foundation for this study, includes detailed information on over

7000 housing units, covering a wide range of variables from structural conditions to occupancy details.

1.3. Importance of Analyzing Housing Conditions and Their Impact on Monthly Utility Costs

Analyzing how housing conditions affect monthly utility costs is crucial for several reasons. Firstly, these costs represent a significant portion of monthly expenses for many households, influencing economic stability and disposable income levels. Understanding the variations in monthly utility costs between rented and owned properties can provide deeper insights into the economic pressures faced by different segments of the population. Secondly, this knowledge can lead to more targeted, energy-efficient, and cost-effective housing policies, which are essential in the context of global environmental challenges and the push towards sustainable urban living. Lastly, such analysis helps to identify disparities in housing quality that may disproportionately affect lower-income households, guiding targeted interventions to improve living conditions for vulnerable populations.

2. Data & Data Processing

The data source used for this study is the 2021 New York City Housing and Vacancy Survey (NYCHVS) Microdata [3]. This survey is conducted every three years by the U.S. Census Bureau to provide a detailed view of the living conditions across the city's diverse housing stock. For the 2021 cycle, the dataset encompasses information from 7089 housing units, featuring over 150 variables that cover various aspects of housing conditions, occupancy details, and demographic information. This comprehensive dataset is crucial for analyzing the relationships between housing conditions and utility costs in New York City.

2.1. Description of Variables

The dataset includes a broad array of variables that focus on attributes potentially impacting utility costs, including tenure status (owned or rented), household composition (number of residents, age distribution), building conditions (heating issues, presence of pests), and detailed monthly utility costs (MUTIL).

2.2. Data Processing

1. Data Cleaning: Non-relevant variables such as control numbers, flag indicators, and full weighting variables were removed. Codes indicating non-applicable, not reported, or no response (-3, -2, -1) were converted to NA to streamline the dataset for accurate analysis.
2. Handling Missing Data: Median imputation was used for numeric columns with missing data to preserve data integrity without being influenced by outliers.
3. Outlier Management: The Interquartile Range (IQR) method identified and addressed extreme values in utility cost variables, ensuring robustness and reliability.

4. Final Validation: The dataset underwent a final review to reassess missing values and outliers, confirming its readiness for the analytical processes.

The final dataset “occupied_puf_21_cleaned.csv” contains 7029 observations of 70 variables. This dataset is saved in project folder ready for detailed statistical analysis, providing a solid foundation for uncovering insights into how housing conditions influence utility expenses in New York City.

3. Methodology

The methodology of this study incorporates a combination of statistical and machine learning techniques to analyze the 2021 NYCHVS Microdata. The focus is on identifying the most influential housing-related factors that affect monthly utility costs in New York City. Each technique was selected to complement the data’s characteristics and the study’s objectives, ensuring a comprehensive analysis of the complex relationships within the data.

3.1. Lasso Regression for Variable Selection

Lasso Regression, was employed as the initial step in the analysis to handle variable selection and regularization. At the beginning of the analysis, all categorical variables are converted to factors to ensure they are appropriately handled in the analysis. To avoid data leakage and multicollinearity, variables directly related to utility costs and redundant variables are removed. A 10-fold cross-validation is implemented to optimize the selection of the regularization parameter, lambda. The lambda that minimizes the cross-validation error is chosen for the final model. After that, check for correlation among the selected variables. For numeric predictors, a correlation matrix is generated to look for any highly correlated variables. For categorical variables, a Chi-squared test is conducted to ensure the independence of features; variables with high p-values suggesting independence are considered for exclusion.

At last, a final selection of variables that are included in the subsequent analyses. Update the data with only selected variables, and name the updated dataset as ‘new_data’. This new dataset will be used in the following sections.

3.2. General Linear Models (GLMs)

Following variable selection using Lasso Regression, General Linear Models (GLMs) were utilized to analyze the relationship between the selected predictors and the target variable, ‘MUTIL’. GLMs helped to estimate the effects of various housing conditions on utility costs while accommodating the potentially non-normal distribution of errors.

3.3. Advanced Machine Learning Models

In addition to traditional statistical methods, advanced machine learning models including Random Forests and Gradient Boosting Machines were applied. These methods are particularly adept at handling complex interactions and non-linear relationships.

- **Random Forests**: The rf_model is constructed using the random forest function with new_data, where ‘MUTIL' serves as the dependent variable and all selected variables are predictors. This model builds 100 trees (ntree = 100), allowing for robust ensemble learning through decision trees that reduce variance and potential overfitting. Post-training, the variable importances are extracted using the importance() function, which calculates the significance of each predictor in the model. This analysis is critical for identifying which variables most influence monthly utility costs. A plot of the importance provides a visual representation, making it easier to discern the most impactful predictors.

- **Gradient Boosting Machines (GBMs)**: A gradient boosting model gbm_model is configured with new_data, setting ‘MUTIL’ as the response variable and using all selected variables as predictors. The model is specified to use a Gaussian distribution with parameters including 1000 trees (n.trees = 1000), an interaction depth of 3 (interaction.depth = 3), and a learning rate (shrinkage) of 0.01. Cross-validation with five folds (cv.folds = 5) is implemented to assess model robustness and prevent overfitting. After training, the summary(gbm_model) function is used to examine the performance and significance of the model features, aiding in understanding their impacts on predicting monthly utility costs.

At the end of modeling, calculate the predictions and RMSE for each model and make comparisons.

3.4. Propensity Score Matching (PSM)

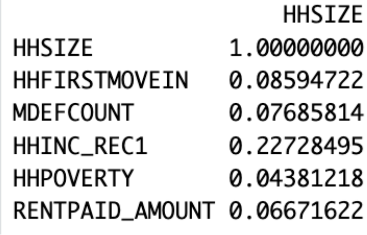
To assess the effect of tenure status on monthly utility costs, Propensity Score Matching (PSM) was utilized. In new_data, the ‘TENURE’ variable, indicating whether a household rents or owns their residence, was converted to a factor with levels labeled ‘Renter’ and ‘Owner’. A propensity score model matchit_model was then created using the matchit() function, with TENURE as the treatment variable and all other selected variables as covariates, employing the nearest neighbor matching method.

After matching, the dataset matched_data containing only the matched observations was extracted. A logistic regression model glm_matched was subsequently fitted on this matched data, with MUTIL as the dependent variable and TENURE as the independent variable, to investigate the impact of renting versus owning on utility costs. The results were standardized to allow for better comparison and interpretation of the effect size.

4. Results

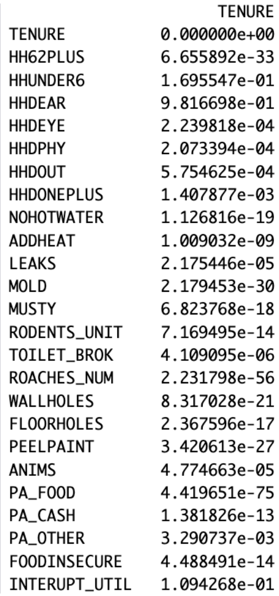
4.1. Lasso Regression for Variable Selection

The correlations between all numerical covariates from the Lasso Regression analysis indicate that there are no strong relationships, with the highest observed correlation being approximately 0.227 between household size (HHSIZE) and household income (HHINC_REC1). This moderate correlation suggests that multicollinearity is not a significant concern, so I can include all numeric variables in further analyses.

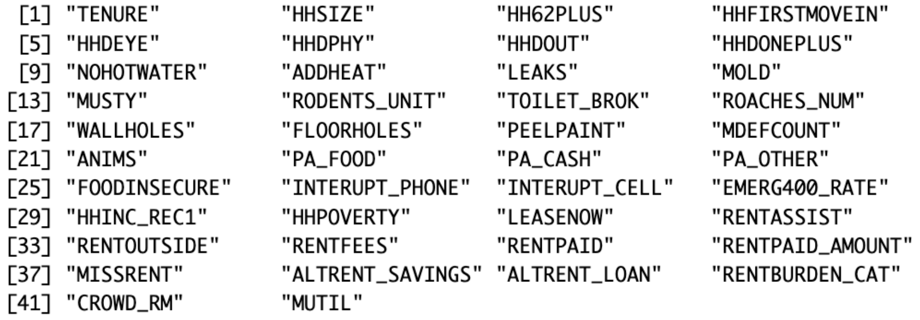


```
                    HHSIZE
HHSIZE          1.00000000
HHFIRSTMOVEIN   0.08594722
MDEFCOUNT       0.07685814
HHINC_REC1      0.22728495
HHPOVERTY       0.04381218
RENTPAID_AMOUNT 0.06671622
```

The p-values for all categorical covariate pairings from the chi-squared tests indicate various degrees of association between these variables, with the most significant relationships generally having p-values less than 0.05. This suggests relevant associations without extremely strong dependencies. To prevent potential multicollinearity, I can exclude variables with p-values greater than 0.05. Thus, 'HHUNDER6', 'HHDEAR', and 'INTERUPT_UTIL' are excluded. Additionally, 'CROWD_BDRM' is also excluded because it is entirely dependent on 'CROWD_RM'.



```
                    TENURE
TENURE        0.000000e+00
HH62PLUS      6.655892e-33
HHUNDER6      1.695547e-01
HHDEAR        9.816698e-01
HHDEYE        2.239818e-04
HHDPHY        2.073394e-04
HHDOUT        5.754625e-04
HHDONEPLUS    1.407877e-03
NOHOTWATER    1.126816e-19
ADDHEAT       1.009032e-09
LEAKS         2.175446e-05
MOLD          2.179453e-30
MUSTY         6.823768e-18
RODENTS_UNIT  7.169495e-14
TOILET_BROK   4.109095e-06
ROACHES_NUM   2.231798e-56
WALLHOLES     8.317028e-21
FLOORHOLES    2.367596e-17
PEELPAINT     3.420613e-27
ANIMS         4.774663e-05
PA_FOOD       4.419651e-75
PA_CASH       1.381826e-13
PA_OTHER      3.290737e-03
FOODINSECURE  4.488491e-14
INTERUPT_UTIL 1.094268e-01
```



```
 [1] "TENURE"         "HHSIZE"          "HH62PLUS"       "HHFIRSTMOVEIN"
 [5] "HHDEYE"         "HHDPHY"          "HHDOUT"         "HHDONEPLUS"
 [9] "NOHOTWATER"     "ADDHEAT"         "LEAKS"          "MOLD"
[13] "MUSTY"          "RODENTS_UNIT"    "TOILET_BROK"    "ROACHES_NUM"
[17] "WALLHOLES"      "FLOORHOLES"      "PEELPAINT"      "MDEFCOUNT"
[21] "ANIMS"          "PA_FOOD"         "PA_CASH"        "PA_OTHER"
[25] "FOODINSECURE"   "INTERUPT_PHONE"  "INTERUPT_CELL"  "EMERG400_RATE"
[29] "HHINC_REC1"     "HHPOVERTY"       "LEASENOW"       "RENTASSIST"
[33] "RENTOUTSIDE"    "RENTFEES"        "RENTPAID"       "RENTPAID_AMOUNT"
[37] "MISSRENT"       "ALTRENT_SAVINGS" "ALTRENT_LOAN"   "RENTBURDEN_CAT"
[41] "CROWD_RM"       "MUTIL"
```

After excluding variables with higher p-values and redundant variables, there are 41 predictors and 1 outcome variable (as shown above) that are selected for further modeling.

4.2. General Linear Models (GLMs)

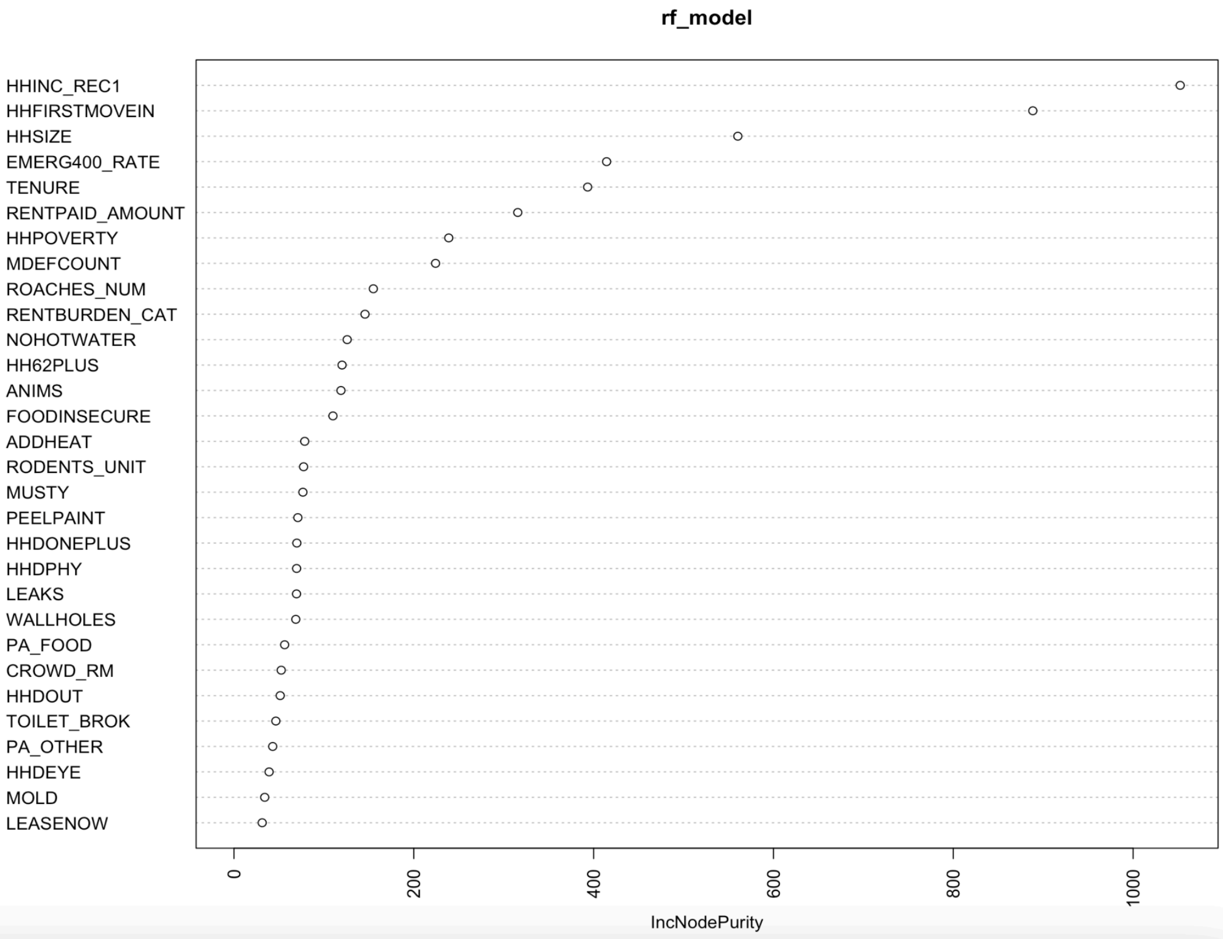
The GLM analysis using the Gaussian family to test factors affecting monthly utility costs (MUTIL). I check for their coefficients and significance. Variables with a p-value less than 0.05 are considered statistically significant, suggesting a strong evidence against the null hypothesis, thus indicating a significant effect on the dependent variable at a 95% confidence level. Here are the variables with a p-value less than 0.05: 'TENURE2', 'HHSIZE', 'HH62PLUS2', 'HHFIRSTMOVEIN', 'HHDPHY2', 'NOHOTWATER2', 'NOHOTWATER3', 'ADDHEAT2', 'LEAKS2', 'TOILET_BROK2', 'ROACHES_NUM2', 'ANIMS2', 'PA_OTHER2', 'INTERUPT_CELLS2', HHINC_REC1', 'RENTBURDEN_CAT4', and 'CROWD_RM3'. The coefficient for 'TENURE2' in the GLM model is 0.525330 which indicates that, holding all other variables constant, the expected monthly utility costs for those in the TENURE2 category (owned) are significantly higher compared to the base category (rented). Besides, most bad housing conditions such as maintenance problems and pest issues show negative coefficients, suggesting that these houses tend to have lower monthly utility costs. This could indicate that properties with such issues, which often belong to lower socioeconomic sectors, are either smaller or have less functional utilities, leading to reduced utility usage. This finding underscores the connection between lower housing quality and reduced utility expenses among economically disadvantaged groups.

In addition, I calculated the prediction accuracy and the Root Mean Squared Error (RMSE) for the GLM model. Its RMSE is 0.926, indicating a moderate level of prediction error.

4.3. Advanced Machine Learning Models

- **Random Forest**

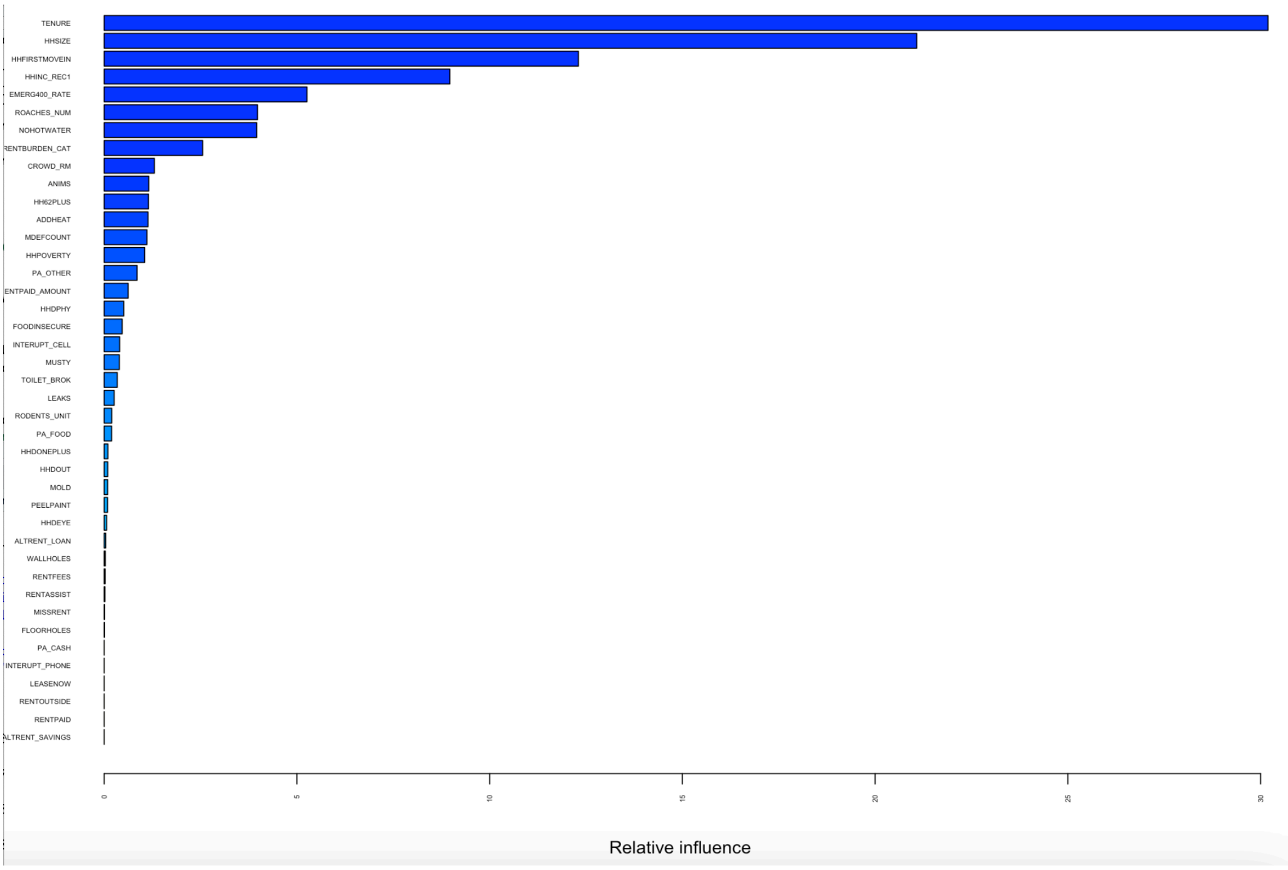
Next, the Random Forest model were employed to further analyze the factors affecting monthly utility costs. I tested all selected covariates' effect on the outcome 'MUTIL'. Based on the result, Key variables such as 'HHINC_REC1' (household income), 'HHFIRSTMOVEIN' (move-in year), 'HHSIZE' (household size), 'EMERG400_RATE' (ability to pay for emergency savings expense), and 'TENURE' (rented vs. owned) were identified as the top 5 most influential factors. These factors emphasizes the strong relationship between income levels and utility costs.



In addition, I calculated the prediction accuracy and the Root Mean Squared Error (RMSE) for the Random Forest model. Its RMSE is significantly lower at 0.489, suggesting a high accuracy in predicting monthly utility costs. Thus, I considered the Random Forest as the best model here.

- **GBM**

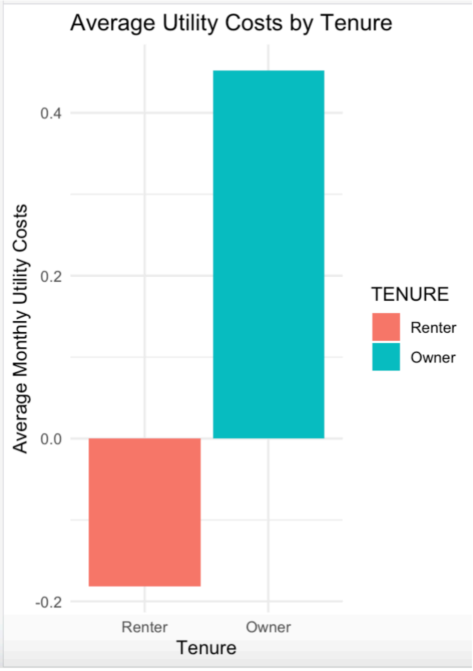
Besides, I built a gradient boosting model to analyze the factors influencing monthly utility costs. The top five most influential factors identified are 'TENURE', 'HHSIZE', 'HHFIRSTMOVEIN', 'HHINC_REC1' , and 'EMERG400_RATE', as shown in the Relative Influence Plot below. The GBM results are similar to the Random Forest results.

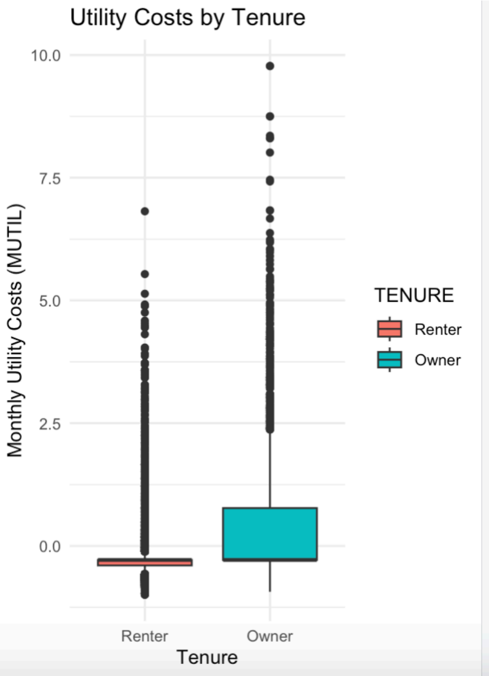


I also calculated the prediction accuracy and the Root Mean Squared Error (RMSE) for the Gradient Boosting model. Its RMSE is 0.890, which, while slightly better than the GLM, shows that it still holds substantial predictive capability, but less accurate than the Random Forest model.

4.4 Propensity Score Matching (PSM)

Propensity Score Matching (PSM) was employed to balance the dataset based on the tenure of housing, comparing renters to owners. I matched the units using the nearest neighbor method to ensure the best match for each unit based on their propensity scores calculated from all other variables in the dataset. Then I built a GLM model on matched data. Based on the summary of the matched GLM, the coefficient for 'TENUREOwner' is 0.57842 with a standard error of 0.03692 and a p-value less than 2e-16. This indicates a strong statistical evidence that, holding all other variables constant, the expected monthly utility costs for owners are significantly higher compared to renters. The Intercept, representing the baseline category of renters, also shows significance, suggesting the model's intercept is different from zero.



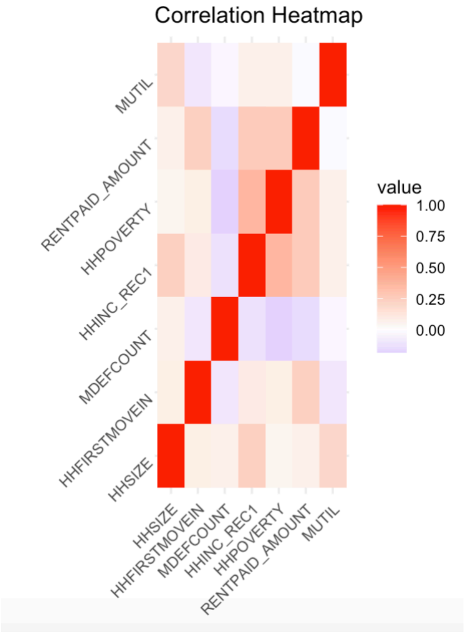


Additionally, one bar chart and one box plot of monthly utility costs by tenure are created. Both plots show that owners generally face higher utility expenses. Based on the box plot, owners have a broader range of costs indicated by a larger interquartile range, while renters have lower and more consistent utility costs. And outliers in the owner group suggest that some owners experience exceptionally high utility costs, likely due to larger or more energy-consuming properties.

4.5 Other Interesting Plots

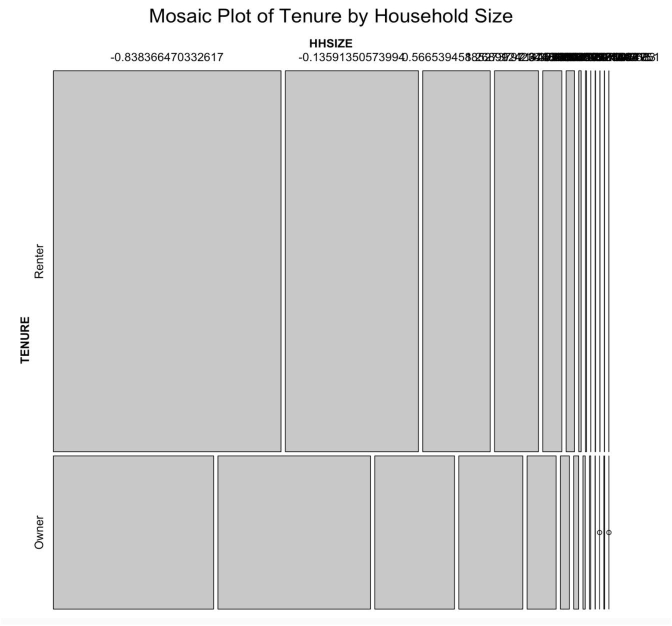
Correlation Heatmap

This is a correlation heatmap that visualizes the relationships among numerical predictors related to utility costs. From the plot, I observe several strong correlations. Notably, the association between household size (HHSIZE) and the number of household defects (MDEFCOUNT) suggests that larger households may experience more maintenance issues, potentially affecting utility costs.



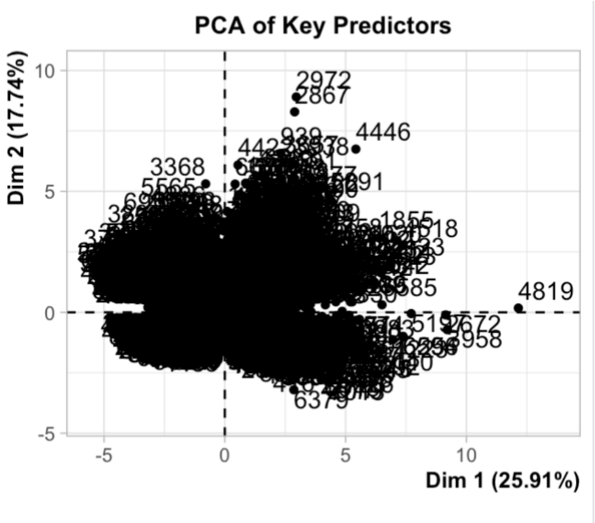
Mosaic Plot of Tenure by Household Size

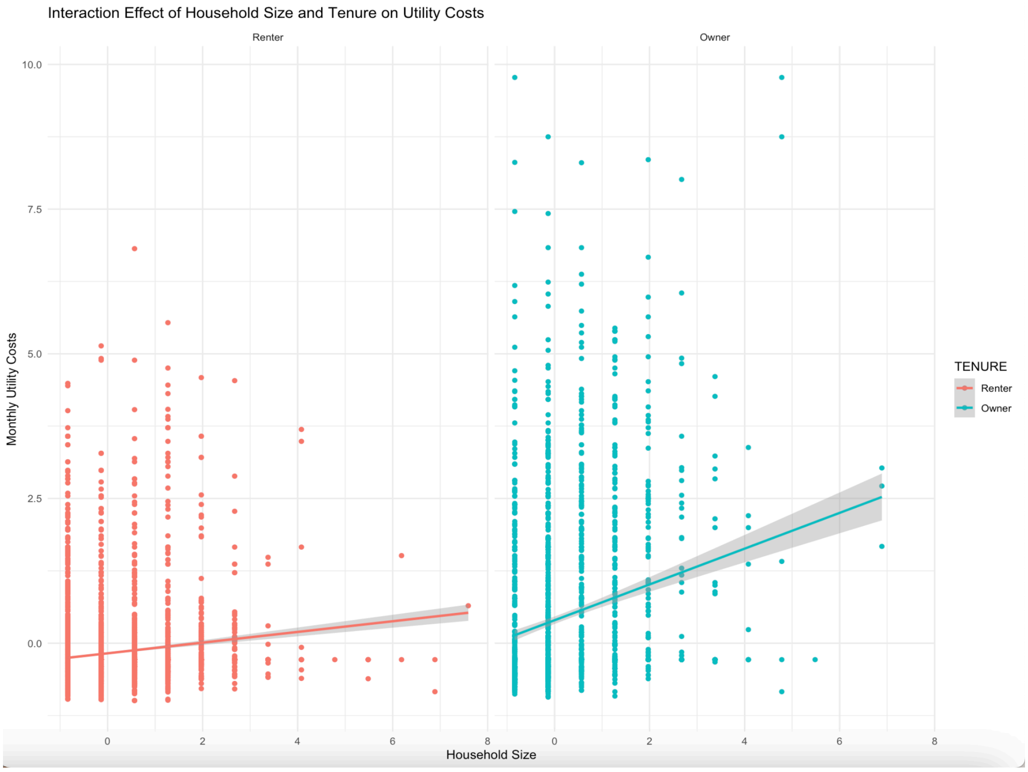
This is a mosaic plot of tenure by household size which graphically represents the categorical relationship between tenure status and household size. From the plot, it is apparent that larger households predominantly belong to renters rather than owners, which could be indicative of urban housing trends where larger families opt for rental accommodations.



PCA of Key Predictors

This is a PCA of key predictors that simplifies the multivariate dataset into principal components, highlighting the most influential factors on utility costs. The plot demonstrates that few principal components explain a significant portion of the variability, indicating that only a handful of variables significantly impact utility costs.



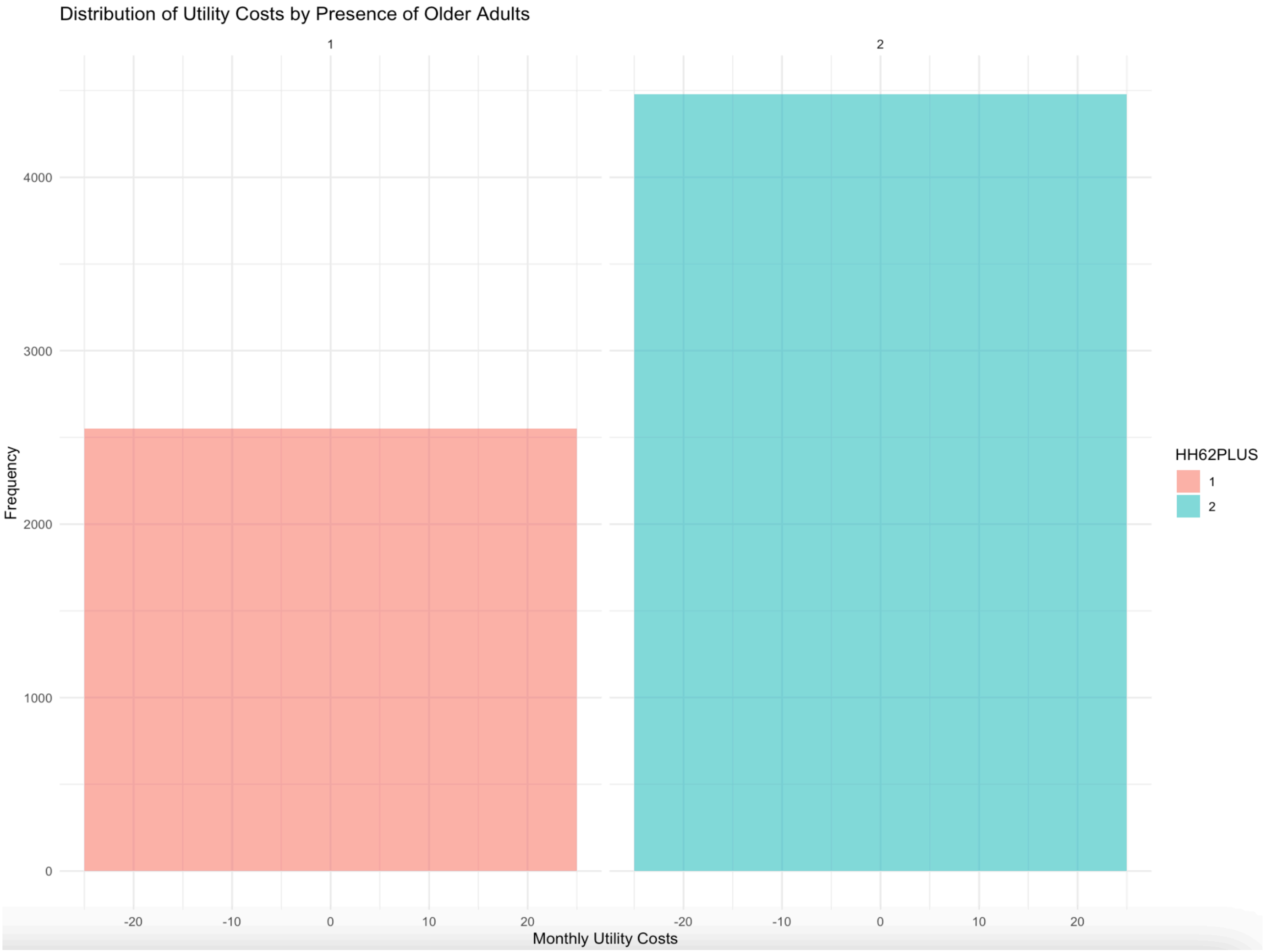


Interaction Effect of Household Size and Tenure on Utility Costs plot

This plot examines how these two variables together influence MUTIL. From the plot, I can see a clear trend where utility costs increase with household size, and this effect is more pronounced among homeowners than renters.

Distribution of Utility Costs by the Presence of Older Adults plot

I am also interested in whether the presence of older adults affect the monthly utility cost. This plot shows how utility spending distributes between households with and without elderly members. The plot reveals a stark difference in utility cost distribution, suggesting that households with older adults likely have higher utility costs, possibly due to increased heating and cooling needs.



5. Conclusion & Limitation

In this study, I explored the complex dynamics between housing conditions and their impact on monthly utility costs in New York City, using data from the 2021 NYCHVS Microdata. My analysis revealed that homeowners tend to incur higher utility costs than renters, likely due to the larger sizes and greater energy demands of owned properties. Additionally, the presence of older adults, higher household incomes, and larger household sizes are associated with increased utility expenses, likely reflecting higher energy use and specific needs such as consistent heating and cooling. Interestingly, my findings also suggest that suboptimal housing conditions, marked by issues like pests and poor maintenance, are linked to lower utility costs. This could be due to these properties being smaller or having less functional utility systems, which naturally leads to lower consumption.

While these insights are valuable, it is important to acknowledge the limitations of my study. The scope of the data may not fully capture nuances such as the energy efficiency of appliances or individual consumption behaviors. Additionally, despite employing Propensity Score Matching to approximate causal relationships, this method does not completely mimic the rigor of experimental designs, so caution must be exercised in drawing causal conclusions.

Based on my findings, I recommend several policy measures. Promoting energy efficiency improvements in rental properties could help alleviate utility costs for economically disadvantaged groups. Support programs specifically designed for households with older adults could assist in managing their typically higher utility costs, perhaps through subsidies or energy-efficient upgrades. Enforce maintenance standards in rental accommodations could not only improve living conditions but also help regulate utility expenses. Educate the public about energy conservation and efficient practices could further help reduce utility costs across various housing types.

I believe this study contributes to a broader understanding of urban housing challenges and underscores the need for informed policy interventions that could lead to more equitable and sustainable living conditions in New York City.

References

[1] U.S. Department of Health & Human Services. (n.d.). *Housing Instability as a Social Determinant of Health*. Healthy People 2030. Retrieved from https://health.gov/healthypeople/priority-areas/social-determinants-health/literature-summaries/housing-instability

[2] City of New York. (n.d.). *OneNYC 2050: Building a Strong and Fair City*. Retrieved from https://www.nyc.gov/site/sustainability/onenyc/onenyc.page

[3] U.S. Census Bureau. (2021). *2021 New York City Housing and Vacancy Survey Microdata*. Retrieved from https://www.census.gov/data/datasets/2021/demo/nychvs/microdata.html

Appendices

https://github.com/GJYCS/Internship_Project.git