Dataset:
The dataset consists of the 2021 New York City Housing and Vacancy Survey (NYCHVS) Microdata from the U.S. Census Bureau, which includes 7089 observations and 150 variables. This comprehensive dataset provides insights into the housing conditions across New York City.

Research Question:
How do various housing conditions, including tenure status and household characteristics, influence monthly utility costs in New York City?

Variables:
The dataset encompasses a broad range of 150 variables covering diverse aspects of housing conditions:

- **Tenure**: TENURE (Owner/Renter status)
- **Household Characteristics**: HHSIZE, HH62PLUS, HHUNDER18, HHUNDER6, HHFIRSTMOVEIN, HHDPHY, HHDOUT, HHDONEPLUS, ANIMS, PA_ANY
- **Maintenance Issues**: NOHEAT, NOHOTWATER, LEAKS, MOLD, MUSTY, ELEVATOR_BROK, TOLIET_BRO K, WALLHOLES, FLOORHOLES, PEELPAINT
- **Pest Issues**: RODENTS_UNIT, ROACHES_NUM
- **Cost**: MUTIL (Monthly Utility Costs)
- **Other variables**: Includes additional factors like UTIL_ELECTRIC, UTIL_GAS, UTIL_HEAT, UTIL_WATER, UTIL_INCLUDED, and many more, capturing detailed aspects of housing utilities and conditions.

Analyses (>=3):

- **Variable Selection via Lasso Regression**: I will utilize Lasso regression to refine variable selection by identifying the most significant predictors of monthly utility costs. The purpose of this step is to reduce the complexity of the model and improve interpretability. I plan to standardize the variables and apply Lasso to eliminate the less significant predictors by shrinking their coefficients towards zero. This will help in focusing subsequent analyses on the most impactful variables.
- **General Linear Models (GLMs)**: After narrowing down the predictors using Lasso, I will employ General Linear Models to analyze how these selected variables influence the monthly utility costs (MUTIL). This stage will concentrate on understanding the specific impacts of key predictors, particularly tenure, on utility expenses. This approach will allow for a detailed exploration of linear relationships within the data.
- **Advanced Machine Learning Models (Random Forests and Gradient Boosting Machines)**: I plan to use advanced machine learning techniques, specifically Random Forests and Gradient Boosting Machines, to model and predict MUTIL based on the variables identified by the Lasso regression. These methods are particularly effective in capturing complex interactions and non-

linear relationships among predictors, providing a comprehensive understanding of the factors driving utility costs.

- **Propensity Score Matching (PSM)**: Through PSM, I will estimate propensity scores for each housing unit to match units with similar probabilities of being owner-occupied versus renter-occupied, based on their observed characteristics. This matching will allow for a balanced comparison to rigorously assess the impact of tenure on utility costs, controlling for confounding variables.

Visualizations (>=9):

- **Variable Importance Plot (RF)**: Display the most influential variables from the RF model.
- **Relative Influence Plot (GBM)**: Highlight key predictors from the GBM model.
- **Correlation Heatmap**: Show correlations between all numerical predictors and MUTIL.
- **Mosaic Plot**: Explore the relationship between tenure status and household size.
- **Box Plots by Tenure**: Compare utility costs distributions between renters and owners.
- **PCA Plot**: Reduce dimensionality to visualize significant variables impacting utility costs.
- **Bar Chart**: Illustrate average utility costs by tenure.
- **Interaction Effects Plot**: Analyze how tenure and household size jointly affect utility costs.
- **Distribution Plot**: Examine utility costs distribution across households with and without older adults.

Objective:

The objective is to provide a comprehensive analysis that can inform policymakers, stakeholders, and the public about the current state of housing in NYC. This research aims to leverage sophisticated statistical techniques to focus on the most impactful variables, thereby ensuring a robust analysis of housing-related utility expenses.