Final Paper

STOR 320.01 Group 14

April 28, 2024

INTRODUCTION

Public housing was a grassroots initiative officially established in 1937 as part of the New Deal to provide a solution to the large proportion of the population living in slums as a result of the rapid urbanization after the Great Depression. This Housing Act created the United States Housing Authority, whose main job is to reallocate federal subsidies to local housing authorities so they can build public housing units. Then, in 1968, the Fair Housing Act of 1968 was implemented to eliminate discrimination toward those seeking federally-assisted housing. This act was a turning point in public housing developments and laid the foundation for the current Department of Housing and Urban Development which handles all public housing developments in the United States today. While HUD was established and grew to have a significant impact on the country, there were limitations placed on the department from certain presidential administrations. The most significant was the Clinton Administration through the Faircloth Amendment which prohibits HUD from funding new developments with specific funds if it exceeds the number that the Public Housing Authority owned at the date of October 1st, 1999. This imposed a constraint on HUD's activities, halting the creation of new public developments, even in cases where local demand for such projects was evident.

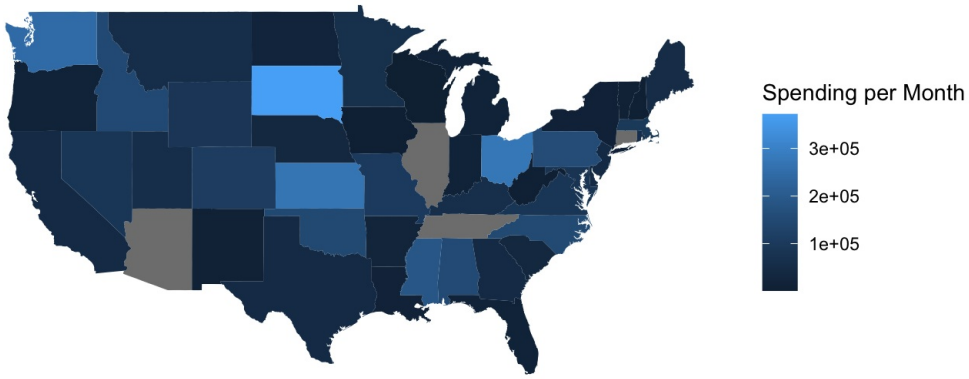
Now, in 2024, U.S. Department of Housing and Urban Development (https://www.hud.gov/) continues this effort, creating safe and decent rental housing for low-income families, individuals with disabilities and the elderly population. Currently, there are around 1.2 million households living in public housing units. Our group wanted to investigate the relationship and trends between different variables and the government spending/rent prices within the developments. We hypothesized that there would be specific variables that would be better at predicting government spending and rent prices, and those variables would be different between the two models. Predicting government spending based on certain variables can be useful to determine the amount of public housing units that can be built in an area. The specific variables will provide insight on the specific demographic data of the surrounding location, like minority proportions or individuals with disabilities amounts, that can predict the government spending in the area. Furthermore, it can provide knowledge on what the government constitutes to be important when deciding how to allocate money to each development zone. Predicting the rent could be especially useful to individuals looking to rent housing, as they can decide which development in their area matches their thresholds for paying certain rent prices. To maximize the affordability of rent prices, specific predictors can provide insight on the best locations (with certain features) to build housing units.

DATA

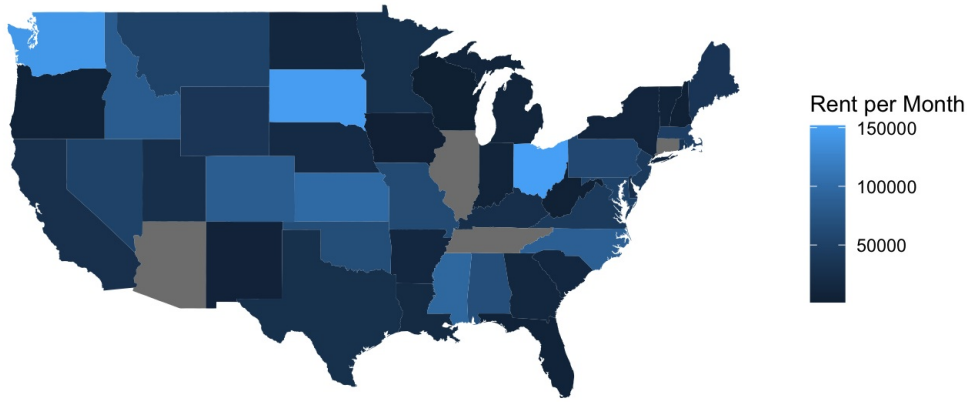
Our data was derived from the Office of Policy Development and Research (https://hudgis-hud.opendata.arcgis.com/datasets/HUD::public-housing-developments/about) on the Public Housing Developments of HUD, and was published on 12/06/2023. The location of each unit is determined by the location of the building with the largest number of units that are managed by HUD. The dataset itself is comprised of many different variables including percent of minorities (`PCT_MINORITY`), percent of women as the head of households (`PCT_FEMALE_HEAD`), percent of households with income below $5,000 per year (`PCT_LT5K`), etc.. For a more complete list, visit this site (https://www.arcgis.com/sharing/rest/content/items/5c96143f79c940a0a8cedae99a1ac562/info/metadata/metadata.xml?format=default&output=html). The numerical variables are the most relevant to our discussion because most are proportions that can be also serve as predictors of government spending per month (`SPENDING_PER_MONTH`) and rent prices per month (`RENT_PER_MONTH`). There were over 150 variables in the dataset, so we subsetted the variables to include most of the numerical variables and the categorical variables that could be relevant. For example, we kept `STATE2KX` , the state variable, because that could provide insight on differences between states in relation to our response variables. Furthermore, when building the models for government spending and rent prices, we removed the other variable to understand their relationships separately. For example, in the `SPENDING_PER_MONTH` models, we removed `RENT_PER_MONTH` , and vice versa.

The graph below shows the government spending per month and rent prices per month around the United States (and its territories).

Average Government Spending per Month by State



Average Rent per Month for Public Housing Developments by State



Looking into the government

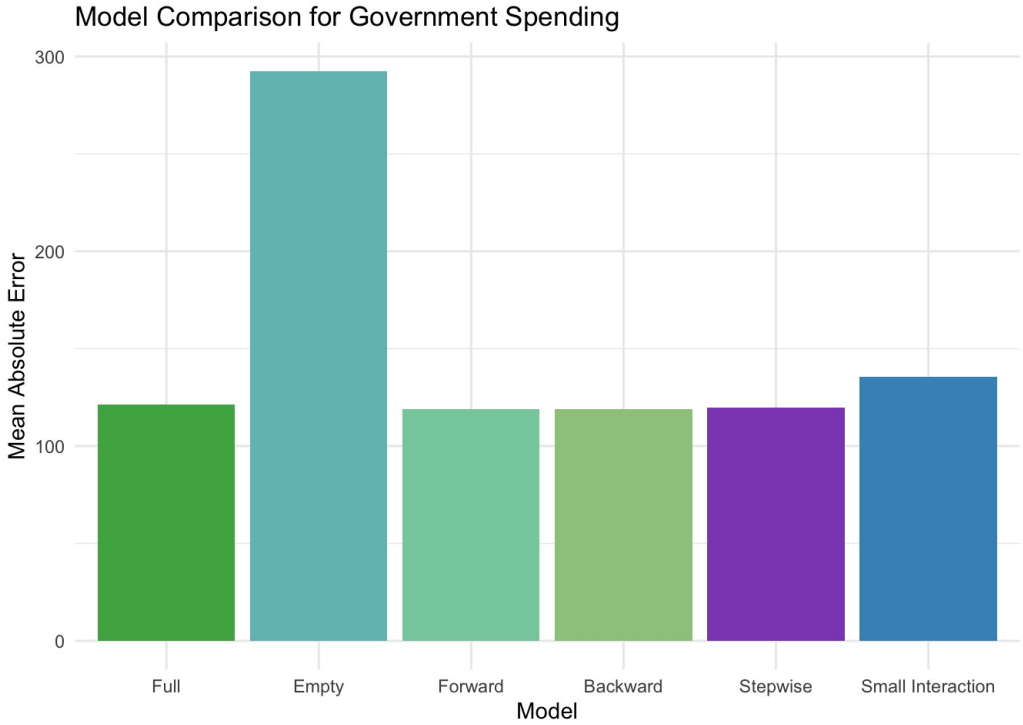
spending, we decided to merge a dataset from Pew Research Center (https://www.pewresearch.org/religious-landscape-study/database/compare/party-affiliation/by/state/) of the political leaning of each state into the original HUD dataset to determine whether there was a statistically significant relationship. Below is a table of the first ten states of the table and their political leanings.

State	Republican Leaning	No Lean	Democratic Leaning
Alabama	52%	13%	35%
Alaska	39%	29%	32%
Arizona	40%	21%	39%
Arkansas	46%	16%	38%
California	30%	21%	49%
Colorado	41%	17%	42%
Connecticut	32%	18%	50%
Delaware	29%	17%	55%
District of Columbia	11%	15%	73%
Florida	37%	19%	44%

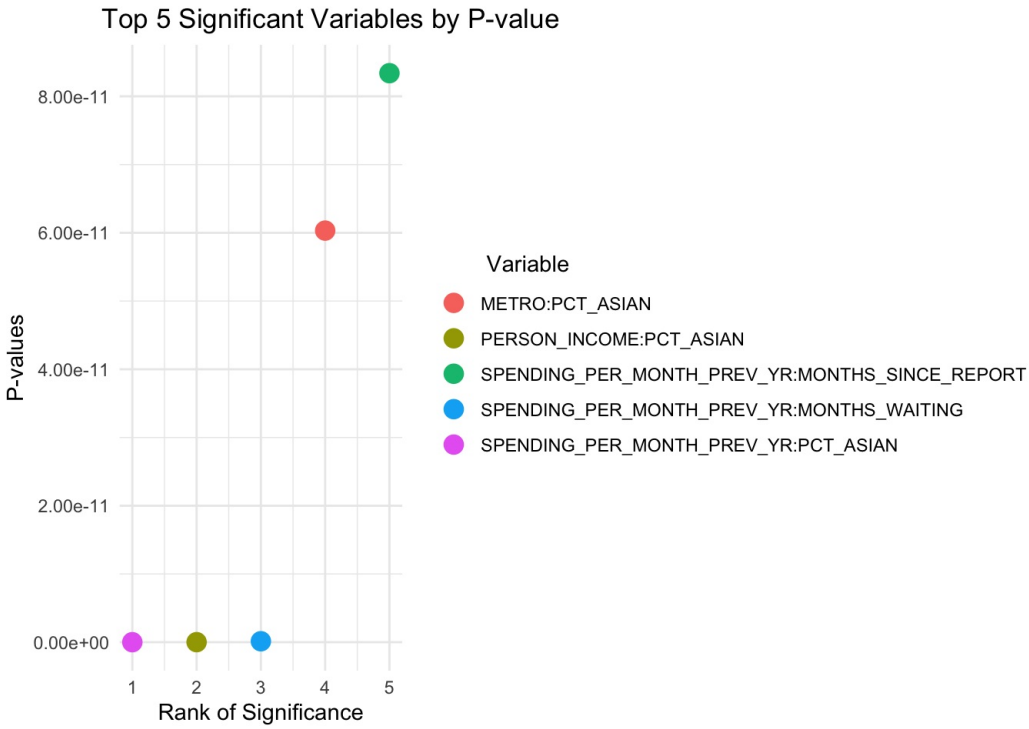
RESULTS

Question One: What Factors Predict Government Spending?

To determine what variables predicted government spending, we began with building five models: full, empty, forward regression, backward selection, and stepwise regression. All the models were statistically significant. After running a 10-fold cross-validation with an MAE function and calculating the adjusted R^2, it is hard to strictly conclude the best-fitted model due to differing statistical metrics, specifically Mean Absolute Error, adjusted R^2, and the p-values. The Mean Absolute Error (MAE) is a metric used to measure the average absolute difference between the predicted values and the actual values in a dataset. In simpler terms, MAE tells you, on average, how much the predictions of a model deviate from the actual values, without considering the direction of the deviation. As you can see in the visual below, the first 5 bars do not have significantly different Mean Absolute Error values other than the Empty model, which is significantly larger. However, to fit an interaction model, which accounts for the relationship between each variable with another, we needed to reduce the number of variables. To select the best model, we opted for the one with the fewest variables. We decided this because the models with the lowest mean absolute errors showed only slight differences, varying by tenths, indicating their close similarity in statistical significance. Therefore, we chose the stepwise function because it had the least amount of variables.

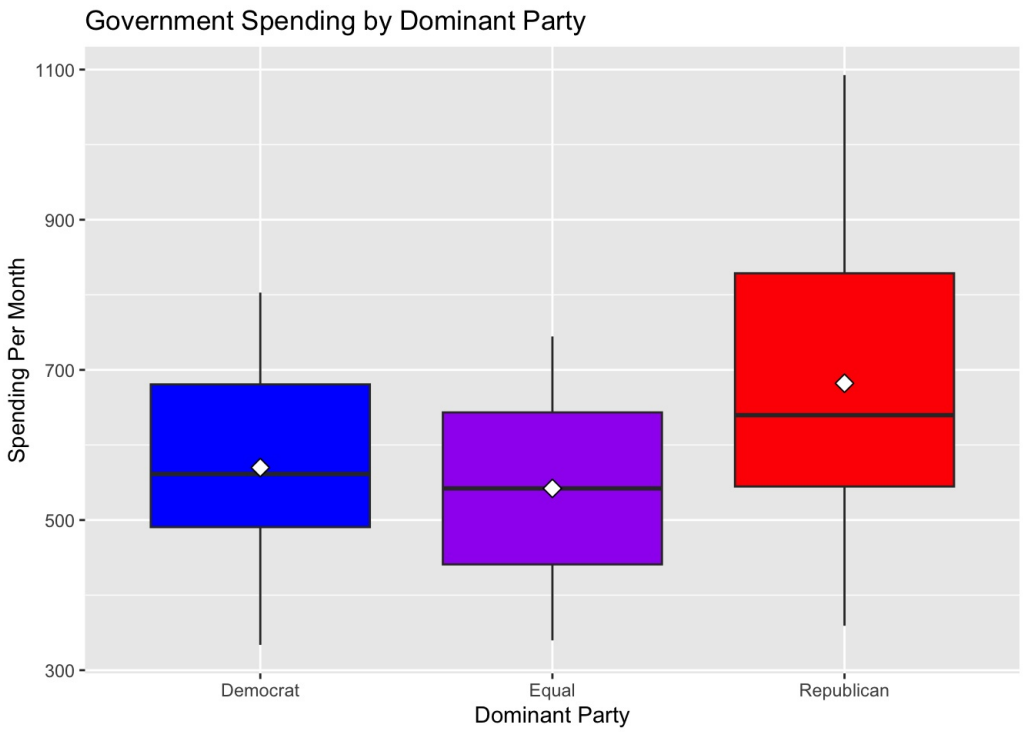


To narrow down the variables for the full interaction model, we subsetted the stepwise model to only include statistically significant variables (p-values less than 0.05). After fitting the interaction model, we found it to have the highest adjusted R^2 value, with it being 0.97, compared to the previous models' 0.93. The adjusted R^2 tells us how well the independent variables explain the variation in the dependent variable, considering the number of predictors in the model, with higher values indicating a better fit. However, we know that the more predictors you add to the model, the higher the adjusted R^2 will be, which is why we decided to fit a model with the five lowest p-values for the interaction variables and see if this smaller model had the same predictive accuracy based on MAE (a more accurate measure of prediction). From the plot below, you can see the five variables and their respective p-values. However, after fitting this model, and completing a 10-fold cross-validation, we found the MAE to be higher than the first five models built, which can be seen in the bar graph above with the bar labeled "Small Interaction". This indicates that the interaction models are not significantly better at predicting government spending. *The table below provides descriptions of each variable.*



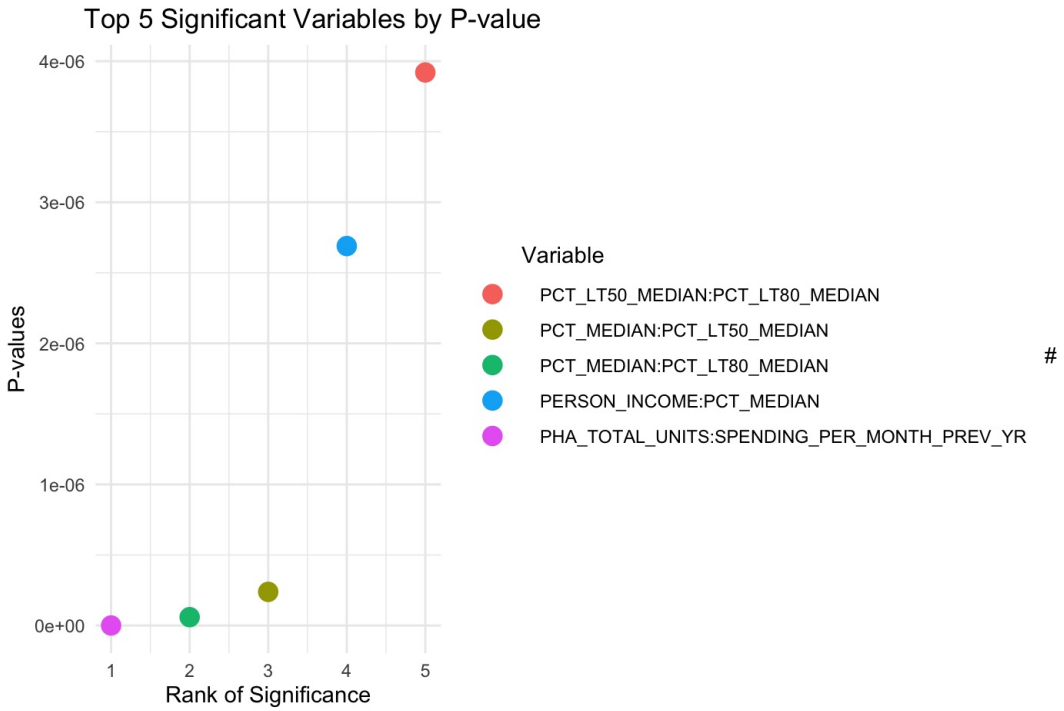
Variable	Description
METRO	Metropolitan Area Indicator
PCT_ASIAN	Percent Asian or Pacific Islander
PERSON_INCOME	Average Household Income per Person per Year
SPENDING_PER_MONTH_PREV_YEAR	Previous Year Spending per Month
MONTHS_SINCE_REPORT	Average Number of Months since Manager Reported on Household
MONTHS_WAITING	Average Number of Months on Waiting List among Admissions

After fitting models, we decided to investigate the relationship between states' dominant political party and their government spending because the stepwise model (the best basic model) had some of those variables (`Moderate_Percent`) in its model. To do this, we used the merged data set that contains each state's dominant political leaning, utilizing this metric to predict the government spending per month in each state. We created a model that displays the distribution of average government spending per month based on the dominant political leanings of the states, which is displayed below. Based off this model, it is evident that there is larger variation within Republican-dominated states for government spending per month, with their median being higher than Democratic-dominated and equal leaning states. Therefore, there seems to be a trend of Republican states spending more money on their public housing developments. Republican States also have, on average, fewer people. The p-value for the model that took into account state politics was less than 2.2e-16, which indicates statistical significance.



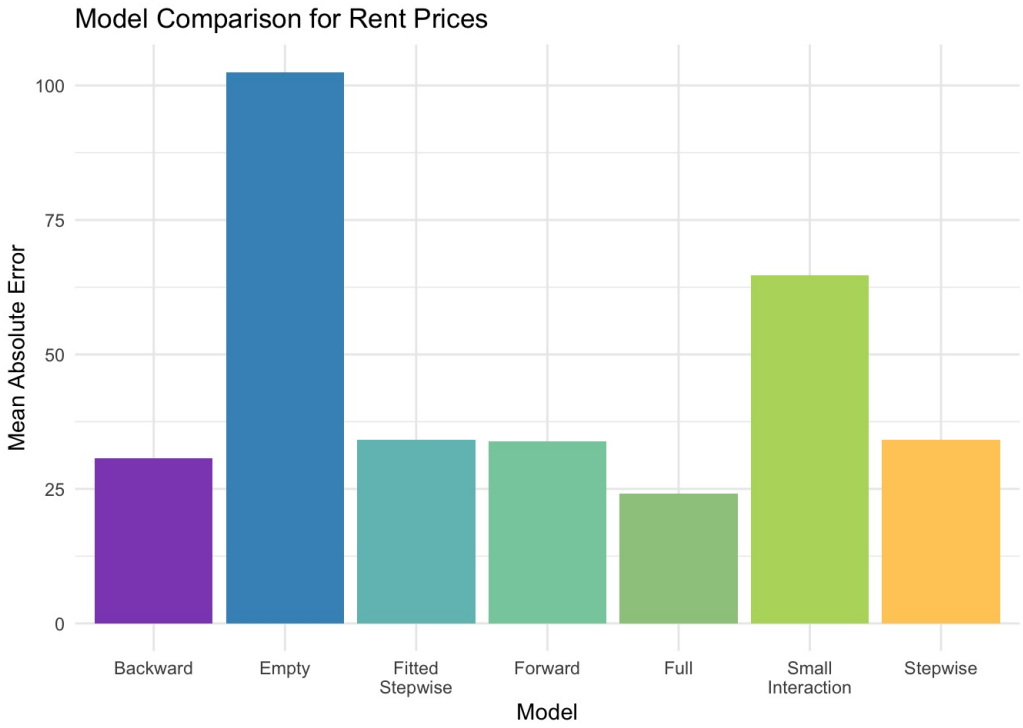
Question Two: What Factors Predict Rent Prices?

When predicting rent, we wanted to determine whether building its models or translating government spending models would prove to be better at predicting rent prices. Therefore, we constructed five models: full, empty, forward regression, backward regression, and stepwise regression, like before, and completed 10-fold cross-validation. We found that the full model had the lowest MAE. So, like above, we created a subsetted data set that only contained the variables from the stepwise regression model. Once again, the adjusted R^2 was higher, but we knew a smaller model could have the same predictive quality. To do this, we created another plot (pictured below) that displayed the five interactions with the lowest p-values, fitting a linear model using only those variables and `RENT_PER_MONTH` as the response variable. *The table below provides descriptions of each variable.*

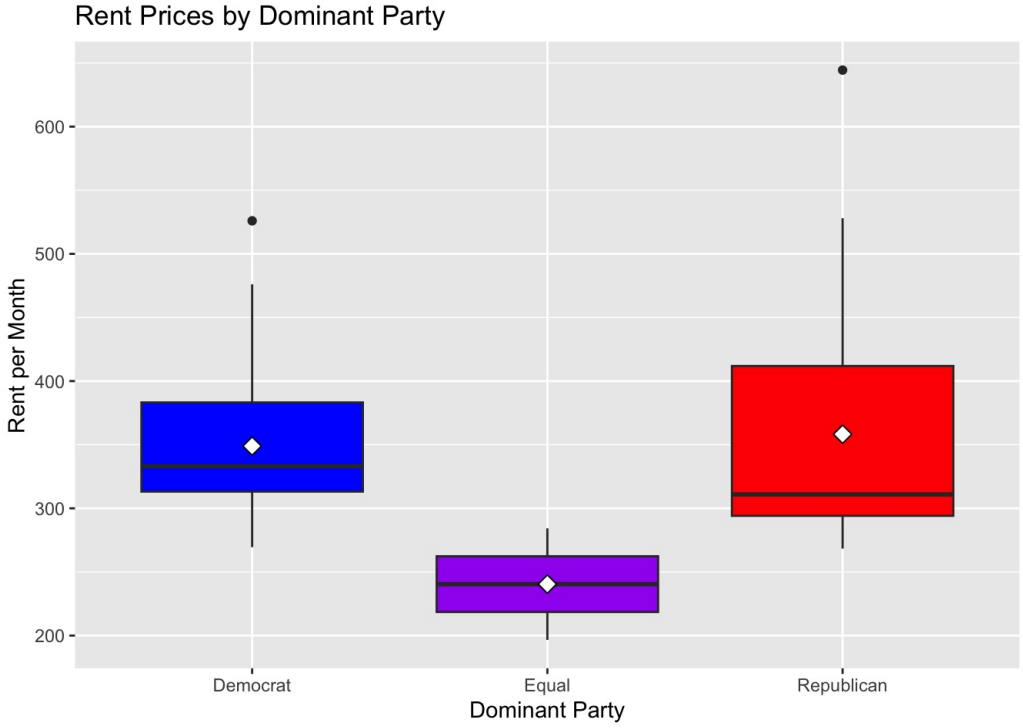


Variable	Description
PCT_LT50_MEDIAN	Percent of Households below 50% median local area Income
PCT_LT80_MEDIAN	Percent of Households below 80% median local area Income
PCT_MEDIAN	Household income as a percent of local area median family income
PERSON_INCOME	Average household income per person per year
PHA_TOTAL	Number of units under contract for federal subsidy and available for occupancy
SPENDING_PER_MONTH_PREV_YR	Previous Year Spending per Month

After fitting cross-validation and computing the MAE for this smaller model, the MAE from the “Small Interaction” model was significantly higher than the forward, backward, and stepwise models. To look at the relationship between government spending and rent, we fit the stepwise model that predicted government spending the best to rent prices to see if the same predictors were just as effective for rent prices. However, the MAE for this model, “Fitted Stepwise”, was not as effective, which can be seen in the bar plot below that compares all the MAE values. The Full Model has the lowest MAE, with the Fitted Stepwise and Small Interaction not proving to have a lower MAE.



After finding the full model to have the predictive ability for rent prices, we decided to compare the political leanings of each state, to see if its effect is similar or different to government spending. After examining the confidence intervals below, it appears that the Democratic and Republican party's average rent prices are more similar than government spending, indicating that political leanings have a less significant relationship with `RENT_PER_MONTH` than `SPENDING_PER_MONTH`.



CONCLUSION

The variables utilized to predict the rent prices did not correlate to the variables that best predicted the government spending. Considering government spending (`SPENDING_PER_MONTH`), the lowest p-values were seen in variables such as government spending in the year before, the percent of Asian or Pacific Islander descent in the community, and location in the country of the community as can be seen in the scatter plot. However, the most significant predictors for rent price (`RENT_PER_MONTH`) were variables such as average household income in the community or the percent of households below a certain income percentile compared to the average income in the local area. In essence, our government spending model could not be used to predict rent and our rent model would not be able to be used to predict government spending as they simply had different predictor variables and factors. Moreover, when we took our most accurate government spending model and fitted it to predict rent, the mean absolute error that was generated was larger than that of the full model we created for rent price prediction. This again shows how the differing predictors are what caused our models for government spending and rent prices to behave differently.

When examining the political affiliation in a state, this metric had clear effects on government spending; the same can not be said for rent prices. Red States had more money spent on them on average, but the same trends were not found when it came to rent. You could also accurately predict a state's expenditure knowing its politics. This highlights a tie between our political system's parties and our federal allocation of public housing.

Moving forward, we recommend conducting further comparisons between models predicting government spending and rent prices. Additionally, exploring additional interaction variables could uncover predictors with better predictive capabilities for each response variable. Incorporating historical data from previous years could offer valuable insights into the correlations guiding government spending allocation and rent price determination. Finding variables that are stronger at predicting rent prices and government spending is beneficial to policymakers and stakeholders because they're involved in housing policy and resource allocation. For example, the government spending in the previous year provides insight into the spending of the current year. While this may seem obvious, that relationship gives insight that changes in the demographics of the housing site are not as important as the previous spending values. Therefore, the future of public housing sites, including their funding and affordability, can be determined by these models.