Socioeconomic Expressions of Energy Poverty

Team 6: Sean Franco, Linsie Zou, and Sreya Rapolu

Introduction

Most households in the United States rely on diverse energy sources for their daily needs, and these energy sources vary seasonally. Fluctuating energy prices exacerbate the cost burdens on households' energy consumption and choice of energy (Chester and Morris, 2016). This inquiry explores the sociological household makeup related to energy poverty. Expression of household demographics, including race, gender, income, and education, are analyzed by their energy usage. By analyzing data from a national survey on energy consumption, we examine energy poverty as expressed through heating and cooling costs proxies and medical responses due to the lack of household climate regulation. Our goal is to accurately map the extent of energy poverty, highlighting both clear-cut cases and energy poverty's nuanced edges. Aligning with the basic premise that marginalized groups are affected the most, income status directly impacts energy poverty. The inquiry attempts to touch on techniques to elicit nuances in detecting energy poverty that may initially be overlooked.

Literature

Looking at the previous literature, many articles discuss the relationship between energy poverty, GDP, prices, and health. A 2021 study on the nexus between energy poverty and energy efficiency found that energy poverty reduces gross domestic product (GDP) (Li et al., 2021). It also leads to dramatic decreases in a country's social welfare in the long run. Most often it is low-income families who are directly targeted by energy poverty as they are unable to successfully eliminate the causes of energy poverty.

One of the main drivers of energy poverty in European countries is energy prices (Halkos and Gkampoura, 2021). People facing energy poverty in European countries were unable to keep up with price increases. In addition to energy prices, unemployment and the percentage of people at risk of poverty also were linked to energy poverty. The authors also found that GDP is inversely related to energy poverty, which corroborates the study done by Li et al. (2021).

As people encounter energy poverty, their risks for negative environmental conditions and health issues become apparent. A 2019 paper created a structural energy poverty vulnerability (SEPV) index to study the relationship between SEPV and energy poverty in the European Union (Recalde et al., 2019). Using a hierarchical cluster analysis, they grouped countries by their SEPV and then fitted a Poisson regression model to the data. Analyzing the model results showed countries with low SEPV scores had statistically significantly higher frequencies of energy poverty and higher risk for excess winter mortality. When people are unable to pay the expenses for energy, they face increased risk of poor quality of life conditions and even death.

The European Union is not the only region that faces energy poverty issues. Globally, residents are affected by energy poverty. Pan, Biru, and Lettu examined the impact of energy poverty on public health in a 2021 study. The authors found that “energy poverty has a detrimental effect on public health.” Specifically, it is through living standards that energy poverty impacts people's health. Countries that have higher standards of living see weaker effects of energy poverty on health.

Understanding energy poverty is important for maintaining a country's living standards, its management of public health issues, and maintaining its social contract of protection and prosperity to its citizens. In this paper, we explore the relationship between household income and nuanced measures of energy usage and how medical attention changes depending on household income levels and the type of energy needed.

Research Questions

To guide our exploration, we have formulated the following SMART questions:

- How do poverty status in 2020 and other socioeconomic factors impact people’s ability to manage their energy consumption and, consequently, their health?
- Does energy poverty status relate to demographics such as race and income status?
- Do respondents require medical attention due to extreme heat or cold in their homes?

Data Source

To address these questions, we utilized the US Energy Information Administration’s (EIA) Residential Energy Consumption Survey (RECS). The RECS survey is conducted every five years. It encompasses over 700 variable columns and 18,500 observations. Each observation is a household sample. Given the number of observations close to twenty thousand, our group determined this is sufficient to sample national household energy usage. Access our GitHub repository for our datasets and our Python script for data analysis.

Data Overview

The correlation matrix of our key variables of interest serves as a guide to inform how we could construct our data mining models. Based on this correlation plot, employment status, household age (EMPLOYHH and HHAGE), education and income (EDUCATION and MONEYPY), and lack of electric air conditioning and electric heating (NOACEL and NOHEATEL) are correlated. Our matrix also shows how employment status and income (EMPLOYHH and MONEYPY), employment and education level (EMPLOYHH and EDUCATION), and employment and ‘received some social program assistance’ (MONEYPY and PAYHELP) are relatively negatively correlated. These notable positive and negative correlations showcase how social data is interrelated and might introduce endogeneity into models. But given our large dataset, we are proceeding with caution when introducing variables that are correlated, in either way, into our models, which serves as a check on creating overfitted models.

We also explored our dataset and found that it generally samples more households in higher income brackets with higher education levels, as observed in Figures 2 and 3.

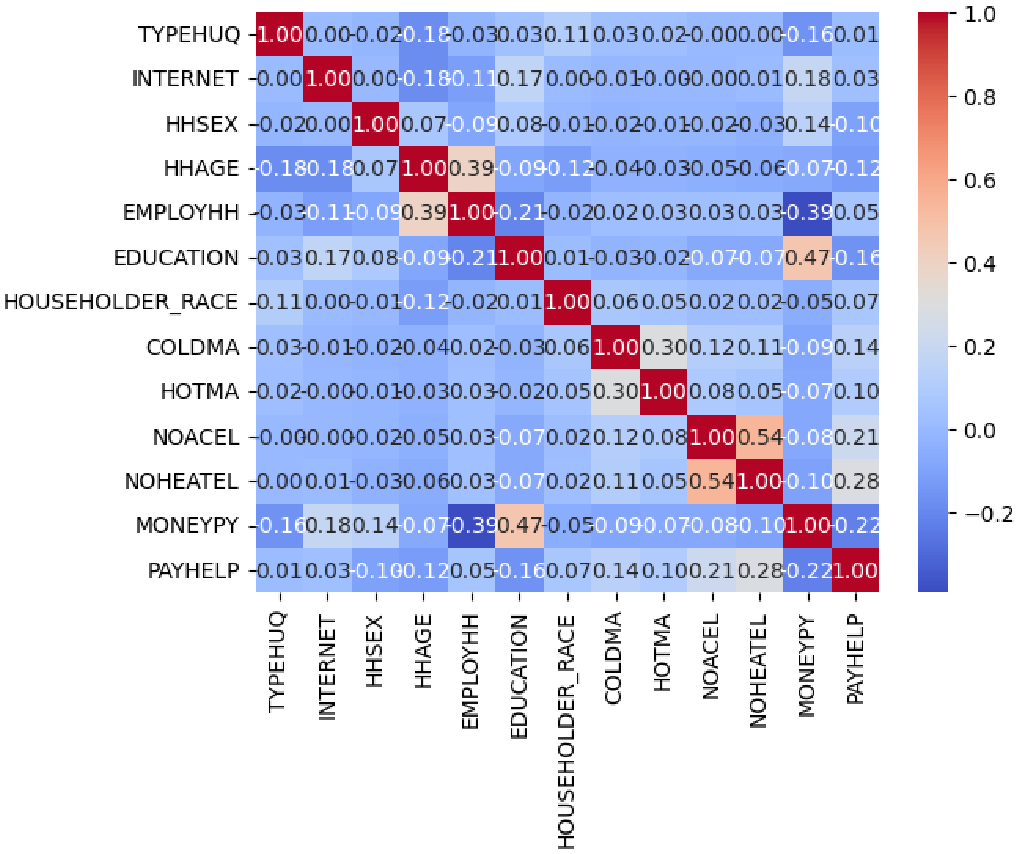
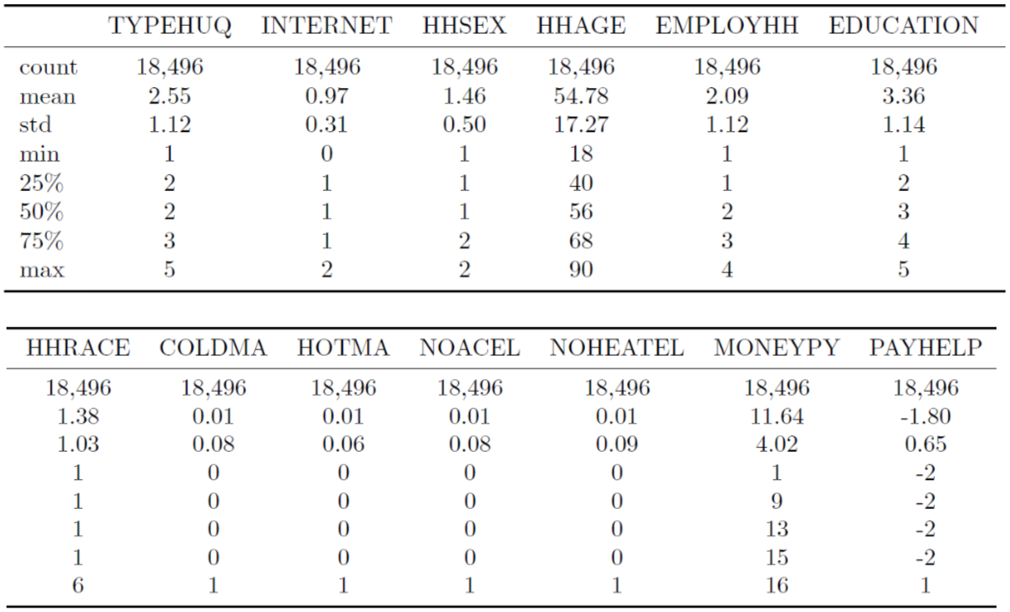


Figure 1: Correlation matrix plot of our variables of interest.

Table A:



	TYPEHUQ	INTERNET	HHSEX	HHAGE	EMPLOYHH	EDUCATION	HHRACE	COLDMA	HOTMA	NOACEL	NOHEATEL	MONEYPY	PAYHELP
count	18,496	18,496	18,496	18,496	18,496	18,496	18,496	18,496	18,496	18,496	18,496	18,496	18,496
mean	2.55	0.97	1.46	54.78	2.09	3.36	1.38	0.01	0.01	0.01	0.01	11.64	-1.80
std	1.12	0.31	0.50	17.27	1.12	1.14	1.03	0.08	0.06	0.08	0.09	4.02	0.65
min	1	0	1	18	1	1	1	0	0	0	0	1	-2
25%	2	1	1	40	1	2	1	0	0	0	0	9	-2
50%	2	1	1	56	2	3	1	0	0	0	0	13	-2
75%	3	1	2	68	3	4	1	0	0	0	0	15	-2
max	5	2	2	90	4	5	6	1	1	1	1	16	1

Table A: A table detailing the count, mean, standard deviation, and other descriptive statistics of some of our variables of interest. It is important to note that all of these variables are categorical.

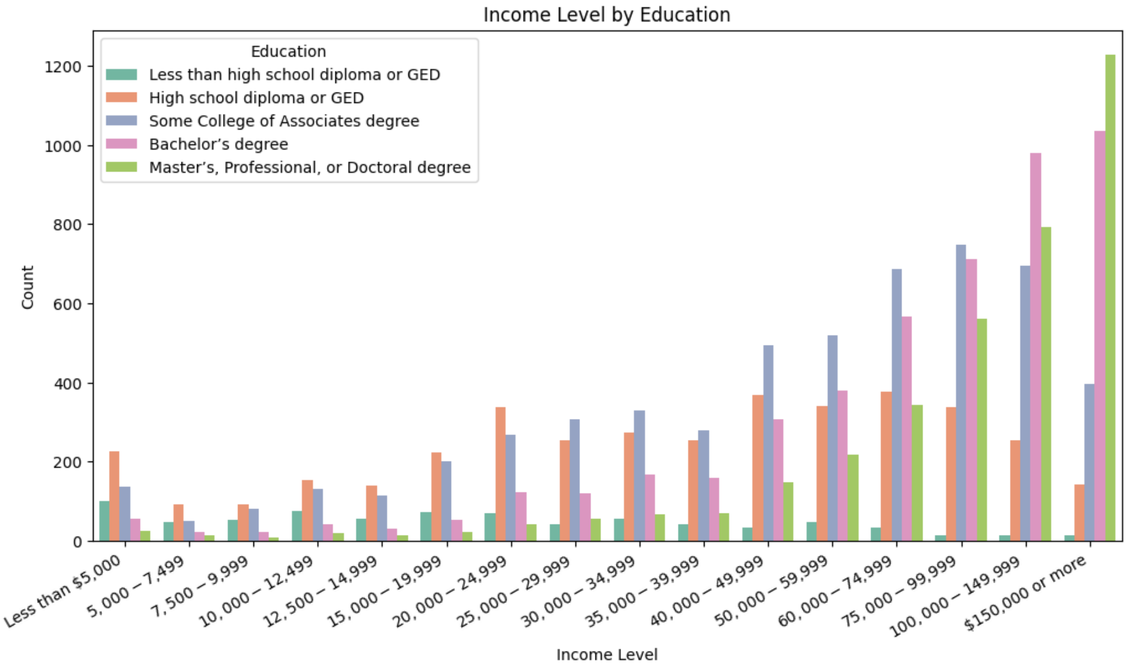


Figure 2: Census income brackets for each household and their count of education status. Higher incomes correlate with higher educational attainment, typically at the graduate level.

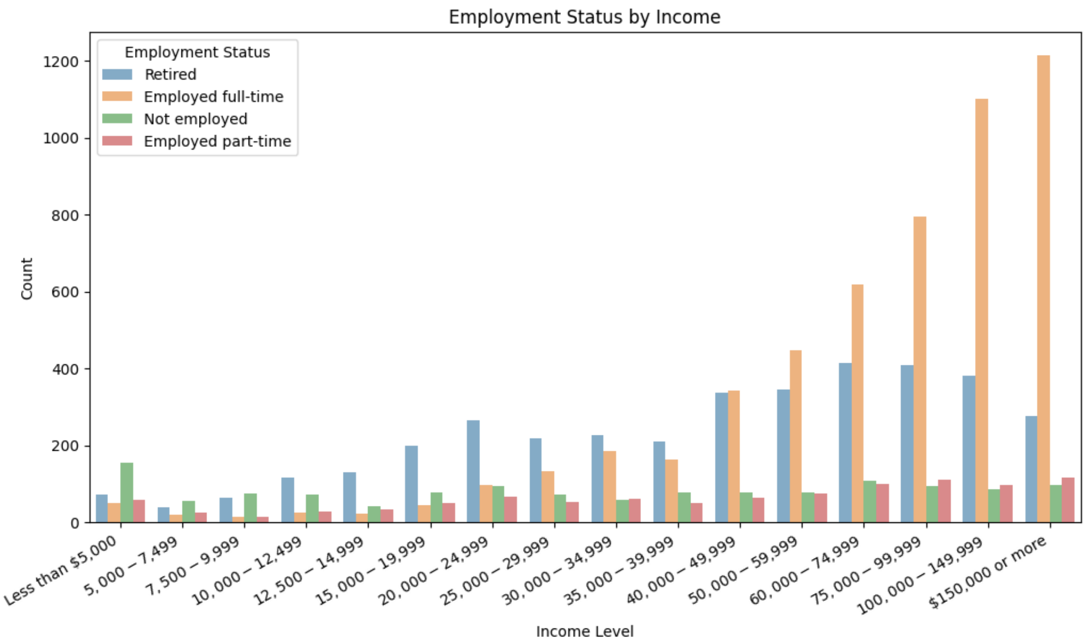


Figure 3: Census income brackets for each household and their count of employment status. Overall, higher incomes are correlated with Employed full-time status and retired status.

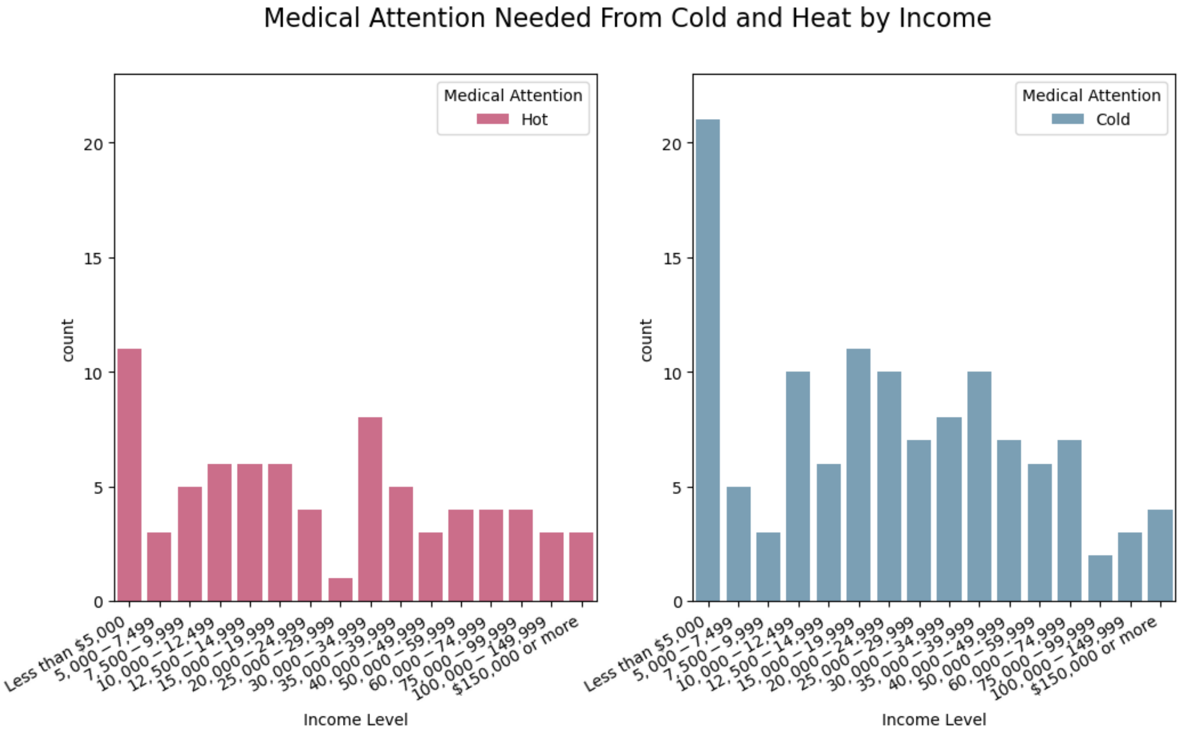


Figure 4: Medical attention needed for extremely hot and extremely cold homes, shown by income status. HOTMA and COLDMA

Methodology

Our analysis will employ various machine learning models, including:

- **Random Forest:** To identify significant predictors of energy consumption and how key variables of interest for use in other models perform at predicting true positives and negatives.
- **Correlation Matrix:** This is our starting point to explore relationships of variables in our dataset, as a guide for performing impactful inquiries of variables.
- **Multivariate Regression:** To quantitatively assess the impact of socioeconomic factors on energy usage.
- **Hypothesis Testing:** Chi-squared independence test of multiple categorical variables.
- **Logistic regression**: For hot and cold medical attention status respectively and for PAYHELP.
- **Support Vector Machine (SVM)**: For PAYHELP on income, employment status, type of home, and no central air or heating.
- **K-Nearest Neighbors Classifier**: For PAYHELP on income, employment status, type of home, and no central air or heating.
- **Gradient Boosting:** for PAYHELP on income, employment status, type of home, and no central air or heating.

Results

We performed a logistic regression of PAYHELP status on income (MONEYPY), household age (HHAGE), and hot and cold medical attention (HOTMA, COLDMA) (see Figure 5). Our rationale was to determine which variables and their corresponding levels are more likely to predict the use or non-use of PAYHELP.

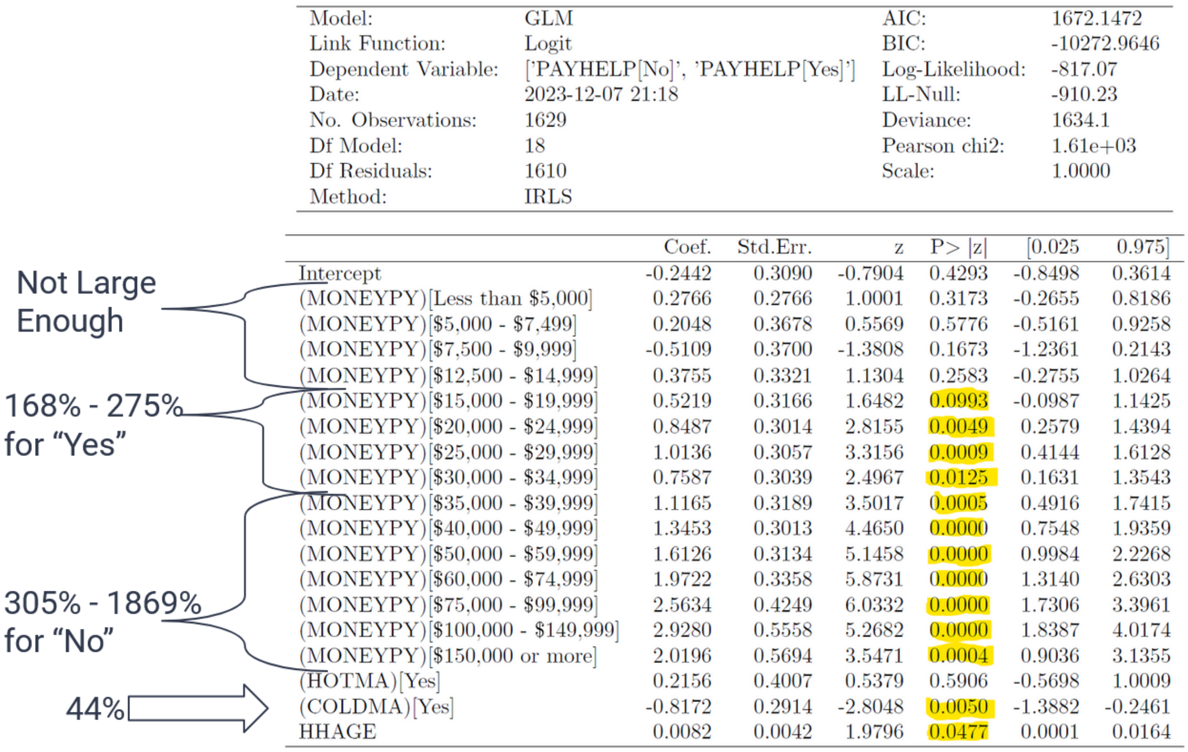
Our RECS survey had low observations for households below $15,000, so our model reflects this as insignificant coefficients. This insight addresses the challenge of adequately surveying low-income households due to factors such as transient housing, among other factors. Conversely, households with incomes ranging from $15,000 to $35,000 had significant coefficients, with the odds ratio between 168% and 275%. At this income range, we can expect more stable long-term housing, which means it is easier for surveyors to collect data. Incomes above $35,000 also have significant coefficients, but we are interpreting them as significant for true negatives.

Household age (HHAGE) coefficients are significant, at an odds ratio of 8%. Since this is continuous data, as households include older age members, the odds ratio increases for positive PAYHELP status to help with social programs. This intuitively makes sense since older-aged household members are generally at risk for associated medical concerns, and often are retired. Additionally, this variable may account for social factors related to intergenerational households needing social program assistance. This poses this variable as a possibility for suffering from energy poverty.

Among hot and cold medical assistance, the latter is significant in our model. However, in our data, cold medical attention has more observations than hot medical attention. We suspect that it's easier to affordably address heat-related medical needs. Most households can affordably address heat medical needs like hydrating with cool or electrolyte waters, taking a cool shower, or wearing cool rags to cool one's body temperature. These examples represent more affordable options to cool oneself if cool central air and its associated electrical costs are unaffordable. These examples rely on a relatively inexpensive price of water considering the type of immediate cooling they provide.

Addressing cold-related medical attention is more of a challenge since it takes more time and energy for the body to warm up versus the immediate effects of cooling down. Common ways to heat a home range from oil, kerosene, gas, wood, and heated air. Even though these fuels offer an increased variety of central heating, their common denominator is high costs. These methods of central heating are more expensive, and, likely, households with low incomes would also have social programs to help assist with these costs. In our model, COLDMA represents the effects of households who are inelastic to heating costs, and choose to forgo incurring the costs. COLDMA has a significant odds ratio of 44% on PAYHELP, which presumably means that households that forgo heating also receive other social program assistance such as food vouchers, heating subsidy assistance, etc. This insight means that these social programs miss the gaps in fully addressing heating and cooling needs.

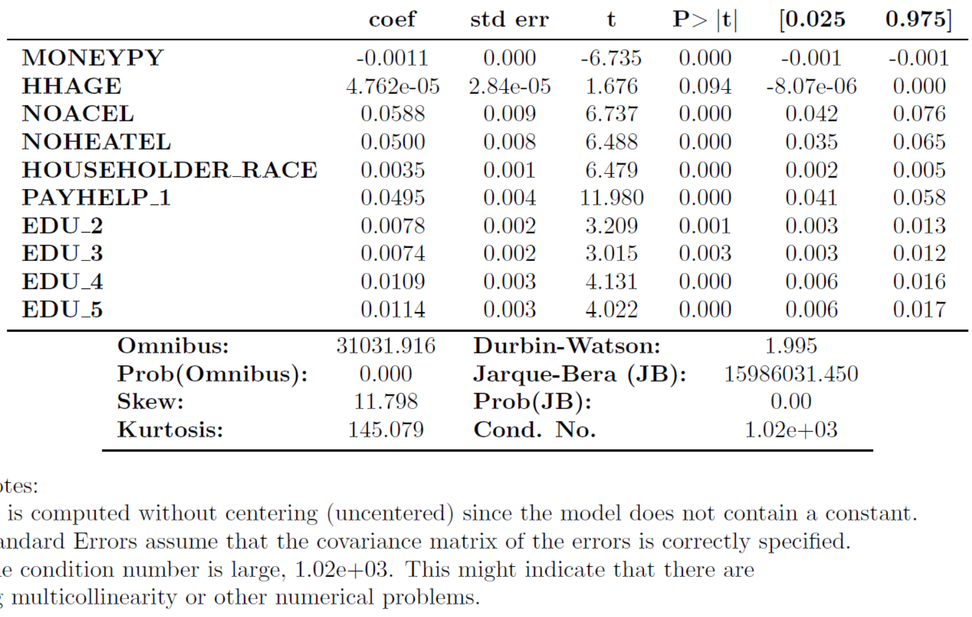
Running a multilinear regression, the explanatory variables explain 3.6% of the variation in the COLDMA variable (see Figure 6). Compared to the Linear Regression conducted previously, this model is better. However, the r-squared value of 3.6% is still very small. Nonetheless, the multilinear model shows when people seek medical attention when the home is too cold. They can also not use other temperature systems like AC and heating equipment. Some people who sought medical attention because the home was too cold also received assistance to pay energy bills. This finding is concerning as it indicates welfare assistance may not do enough to reduce mortality risks from experiencing energy poverty.



Model:	GLM	AIC:	1672.1472
Link Function:	Logit	BIC:	-10272.9646
Dependent Variable:	['PAYHELP[No]', 'PAYHELP[Yes]']	Log-Likelihood:	-817.07
Date:	2023-12-07 21:18	LL-Null:	-910.23
No. Observations:	1629	Deviance:	1634.1
Df Model:	18	Pearson chi2:	1.61e+03
Df Residuals:	1610	Scale:	1.0000
Method:	IRLS		

	Coef.	Std.Err.	z	P> |z|	[0.025	0.975]
Intercept	-0.2442	0.3090	-0.7904	0.4293	-0.8498	0.3614
(MONEYPY)[Less than $5,000]	0.2766	0.2766	1.0001	0.3173	-0.2655	0.8186
(MONEYPY)[$5,000 - $7,499]	0.2048	0.3678	0.5569	0.5776	-0.5161	0.9258
(MONEYPY)[$7,500 - $9,999]	-0.5109	0.3700	-1.3808	0.1673	-1.2361	0.2143
(MONEYPY)[$12,500 - $14,999]	0.3755	0.3321	1.1304	0.2583	-0.2755	1.0264
(MONEYPY)[$15,000 - $19,999]	0.5219	0.3166	1.6482	0.0993	-0.0987	1.1425
(MONEYPY)[$20,000 - $24,999]	0.8487	0.3014	2.8155	0.0049	0.2579	1.4394
(MONEYPY)[$25,000 - $29,999]	1.0136	0.3057	3.3156	0.0009	0.4144	1.6128
(MONEYPY)[$30,000 - $34,999]	0.7587	0.3039	2.4967	0.0125	0.1631	1.3543
(MONEYPY)[$35,000 - $39,999]	1.1165	0.3189	3.5017	0.0005	0.4916	1.7415
(MONEYPY)[$40,000 - $49,999]	1.3453	0.3013	4.4650	0.0000	0.7548	1.9359
(MONEYPY)[$50,000 - $59,999]	1.6126	0.3134	5.1458	0.0000	0.9984	2.2268
(MONEYPY)[$60,000 - $74,999]	1.9722	0.3358	5.8731	0.0000	1.3140	2.6303
(MONEYPY)[$75,000 - $99,999]	2.5634	0.4249	6.0332	0.0000	1.7306	3.3961
(MONEYPY)[$100,000 - $149,999]	2.9280	0.5558	5.2682	0.0000	1.8387	4.0174
(MONEYPY)[$150,000 or more]	2.0196	0.5694	3.5471	0.0004	0.9036	3.1355
(HOTMA)[Yes]	0.2156	0.4007	0.5379	0.5906	-0.5698	1.0009
(COLDMA)[Yes]	-0.8172	0.2914	-2.8048	0.0050	-1.3882	-0.2461
HHAGE	0.0082	0.0042	1.9796	0.0477	0.0001	0.0164

Figure 5: Logistical regression results for income, household age, and hot and cold medical attention on PAYHELP status.



	coef	std err	t	P> |t|	[0.025	0.975]
MONEYPY	-0.0011	0.000	-6.735	0.000	-0.001	-0.001
HHAGE	4.762e-05	2.84e-05	1.676	0.094	-8.07e-06	0.000
NOACEL	0.0588	0.009	6.737	0.000	0.042	0.076
NOHEATEL	0.0500	0.008	6.488	0.000	0.035	0.065
HOUSEHOLDER_RACE	0.0035	0.001	6.479	0.000	0.002	0.005
PAYHELP_1	0.0495	0.004	11.980	0.000	0.041	0.058
EDU_2	0.0078	0.002	3.209	0.001	0.003	0.013
EDU_3	0.0074	0.002	3.015	0.003	0.003	0.012
EDU_4	0.0109	0.003	4.131	0.000	0.006	0.016
EDU_5	0.0114	0.003	4.022	0.000	0.006	0.017

Omnibus:	31031.916	**Durbin-Watson:**	1.995
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	15986031.450
Skew:	11.798	**Prob(JB):**	0.00
Kurtosis:	145.079	**Cond. No.**	1.02e+03

Notes:
[1] R^2 is computed without centering (uncentered) since the model does not contain a constant.
[2] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[3] The condition number is large, 1.02e+03. This might indicate that there are
strong multicollinearity or other numerical problems.

Figure 6: Multilinear regression of income, household age, no central cooling or heating, household race, PAYHELP status, and education level on cold medical attention.

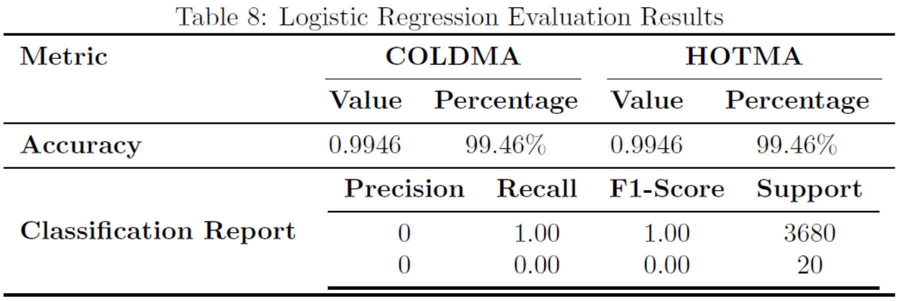
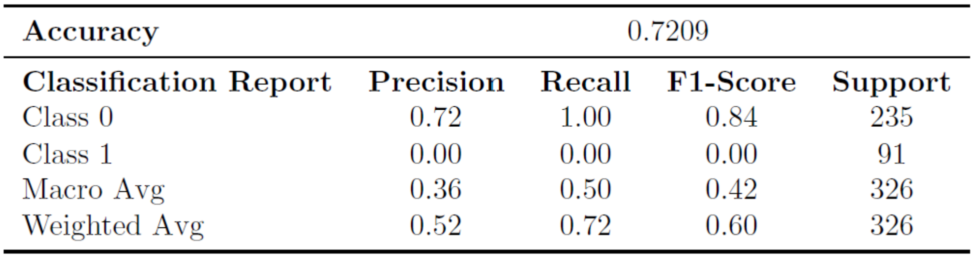


Table 8: Logistic Regression Evaluation Results

Metric	COLDMA		HOTMA	
	Value	Percentage	Value	Percentage
Accuracy	0.9946	99.46%	0.9946	99.46%
	Precision	Recall	F1-Score	Support
Classification Report	0	1.00	1.00	3680
	0	0.00	0.00	20

Figure 7: Logistic regression of cold and hot medical attention, respectively, controlling for household race, education level, income, and no central air conditioning status.

The confusion matrix, in Figure 7, shows that the model does not detect any true positives. Instead, our model has high scores in precision and recall for identifying negatives. The model is accurate in detecting our 3.6k true negatives, which increases the model's accuracy. However, our model does not detect any true positive values, which showcase its failure. Our model over-fitted at a 99.4% accuracy. It inflates the false positives higher than the actual true positives.

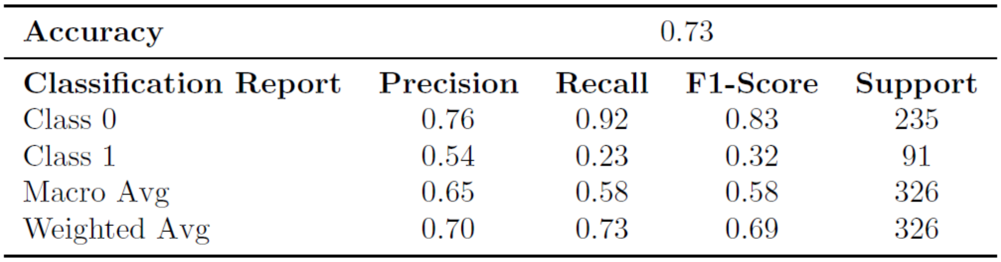


Accuracy	0.7209			
Classification Report	Precision	Recall	F1-Score	Support
Class 0	0.72	1.00	0.84	235
Class 1	0.00	0.00	0.00	91
Macro Avg	0.36	0.50	0.42	326
Weighted Avg	0.52	0.72	0.60	326

Figure 8: SVM results of PAYHELP status controlling for income, employment status, type of home, and no central air or heating.

The model performs well for predicting Class 0, with high detection of negatives of PAYHELP status, as seen in Figure 8. But even with the support vector, it cannot classify Class 1, or true positives, which informs us that there are not enough Class 1 responses in the dataset to infer meaningful predictions.

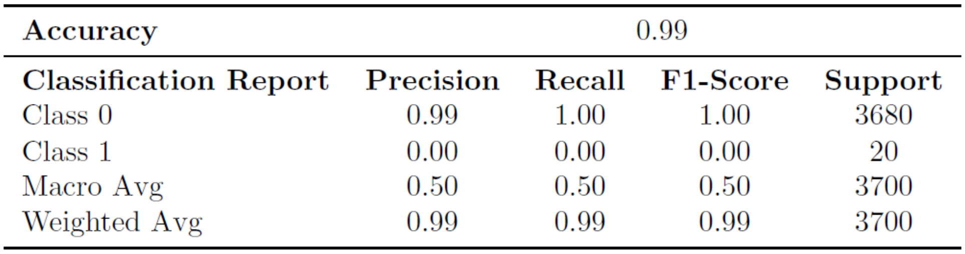
The KNN Model for the 'PAYHELP' variable has an accuracy of 73% (see Figure 9). In particular, the model performs well for Class 0 in all 4 evaluation metrics. KNN provides higher results than the other models for detecting false positives, but not for false negatives. It does not perform well in detecting false positives and false negatives.



Accuracy	0.73			
Classification Report	**Precision**	**Recall**	**F1-Score**	**Support**
Class 0	0.76	0.92	0.83	235
Class 1	0.54	0.23	0.32	91
Macro Avg	0.65	0.58	0.58	326
Weighted Avg	0.70	0.73	0.69	326

Figure 9: KNN results of PAYHELP status controlling for income, employment status, type of home, and no central air, or heating.

The KNN models for COLDMA and HOTMA variables both exhibit a high accuracy of 99.46% (see Figure 10). However, the models perform well only for Class 0, with precision and recall of 99% and 100%, respectively. For Class 1, the precision, recall, and F1-score are all 0%, indicating the model struggles to predict this class. The high accuracy score is primarily driven by correct predictions for Class 0, while Class 1 predictions are ineffective.



Accuracy	0.99			
Classification Report	**Precision**	**Recall**	**F1-Score**	**Support**
Class 0	0.99	1.00	1.00	3680
Class 1	0.00	0.00	0.00	20
Macro Avg	0.50	0.50	0.50	3700
Weighted Avg	0.99	0.99	0.99	3700

Figure 10: KNN results of cold and hot medical attention, respectively, controlling for income, employment status, type of home, and no central air, or heating.

Conclusion

The interactions between various energy sources, seasonality, and fluctuating energy prices significantly impact U.S. households. This inquiry delves into the sociological dynamics of energy poverty, through demographic factors such as race, gender, income, and education on household energy usage.

In conclusion, our analysis of a national energy consumption survey provides a revealing glimpse into the manifestations of energy poverty and explores the intricate link between energy poverty and sociodemographic factors, emphasizing race, gender, income, and education. We examine proxies such as heating and cooling costs, along with medical responses arising from inadequate household climate regulation. Our findings highlight the universal health implications of energy poverty, particularly in regions with lower living standards. Our overarching objective is to enhance our understanding of the evident instances and nuanced boundaries of energy poverty.

From our Logistic regression examining how explanatory variables impact people's need for energy payment assistance–the PAYHELP variable–we found that households that suffer from extreme cold are the households that have a higher likelihood of asking for medical assistance and energy payment assistance. Medical assistance and energy payment assistance were found to be more important for extremely cold homes compared to extremely hot homes. This approach builds upon the foundational notion that energy poverty disproportionately impacts marginalized groups and exhibits a direct correlation with income levels.

Some limitations we faced include the fact that the data we used came from a survey. Many issues can happen with surveys, such as non-response bias, response order bias, social desirability bias, survey fatigue, and generalizability limits. Respondents could choose to not respond to some questions in the survey or not respond to the survey at all, leading to non-response bias. The order of questions and the phrasing of questions can lead to response order bias. Respondents may also feel like they need to conform to certain social standards and may not give true answers to their circumstances. This is known as social desirability bias. Long surveys like the RECS survey can also result in response discrepancies that occur because of survey fatigue. Finally, even though the RECS survey is conducted every 5 years and samples households across the nation, the conclusions drawn using the survey data may not be completely generalizable.

Future studies could consider examining the data across time and utilizing survey measures accommodating the COVID-19 pandemic. Data measured by county or state would also give another layer of analysis. Looking at energy poverty and socioeconomic indicators at these levels instead of nationally would help local and state policymakers address energy poverty issues with more precision and effectiveness. New policies and welfare programs could decrease the risk of harm and mortality and increase people's quality of life.

References

Chester, Lynne, and Alan Morris. “A New Form of Energy Poverty is the Hallmark of Liberalised Electricity Sectors.” *The Australian journal of social issues* 46.4 (2011): 435-59. *CrossRef.* Web.

Halkos, George E., and Elena-Christina Gkampoura. “Evaluating the Effect of Economic Crisis on Energy Poverty in Europe.” *Renewable and Sustainable Energy Reviews* 144 (2021): 110981. Web.

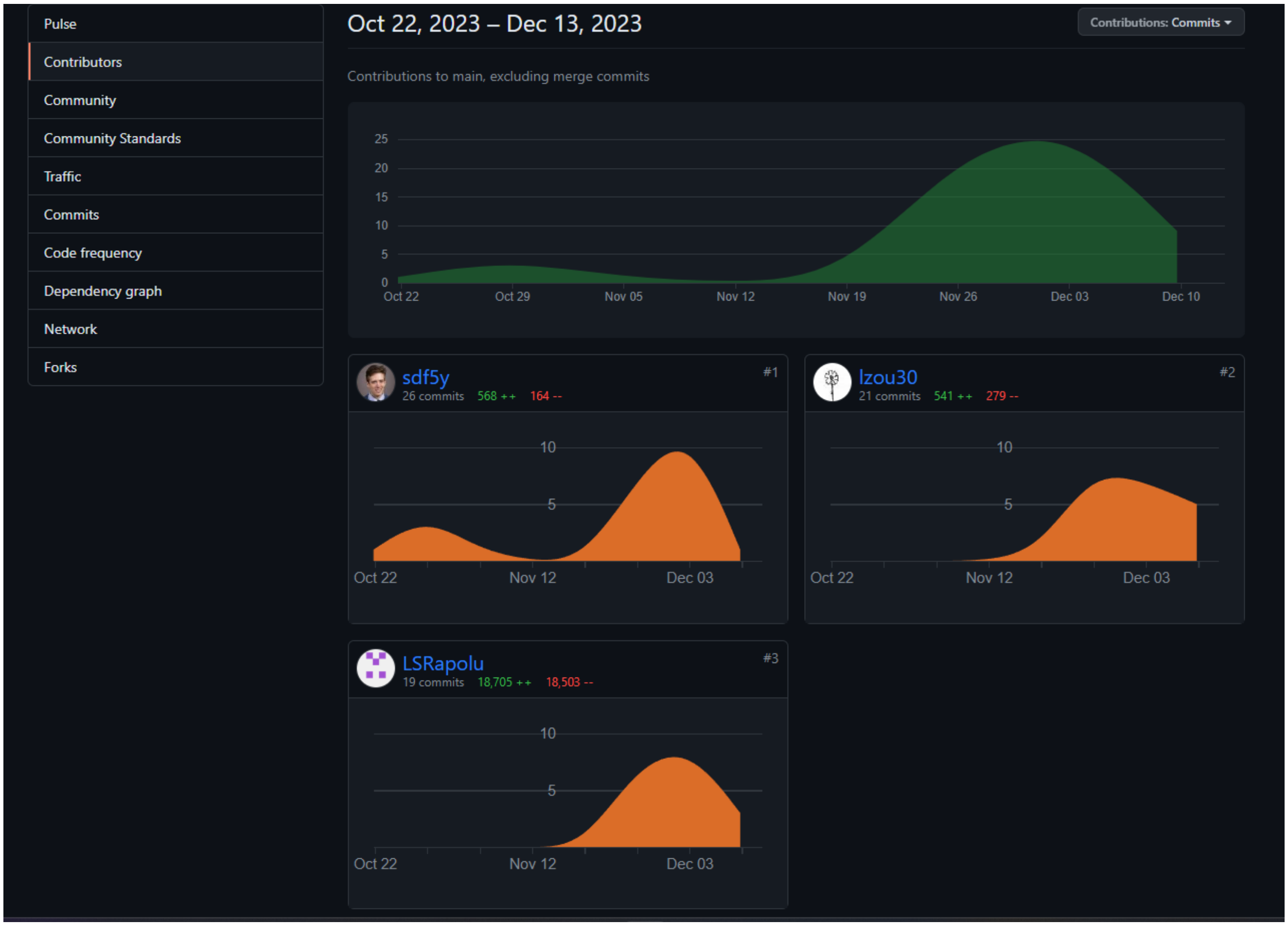
Li, Weiqing, et al. “Nexus between Energy Poverty and Energy Efficiency: Estimating the Long-Run Dynamics.” *Resources policy* 72.2 (2021): 102063. *CrossRef.* Web.

Pan, Lei, Ashenafi Biru, and Sandra Lettu. “Energy Poverty and Public Health: Global Evidence.” *Energy Economics* 101 (2021): 105423. Web.

Recalde, Martina, et al. “Structural Energy Poverty Vulnerability and Excess Winter Mortality in the European Union: Exploring the Association between Structural Determinants and Health.” *Energy Policy* 133 (2019): 110869. Web.

U.S. Energy Information Administration (EIA). “Residential Energy Consumption Survey (RECS).” Web. <https://www.eia.gov/consumption/residential/data/2020/>.

GitHub Contributions



Oct 22, 2023 – Dec 13, 2023

Contributions to main, excluding merge commits

sdf5y
26 commits 568 ++ 164 --

lzou30
21 commits 541 ++ 279 --

LSRapolu
19 commits 18,705 ++ 18,503 --