Data 102 Group 46: Final Report

Authors: August Arneson, Reily Fairchild, Pranav Walimbe, Rohan Zeller

1. Data Overview

Our analysis included a variety of datasets which can be separated by research question:

- RQ 1: EPA eGrid dataset and EPA's Environmental Justice Screening and Mapping Tool (EJScreen).
- RQ 2: EIA oil / electricity data, Federal Reserve Bank Of St. Louis economic datasets, NOAA weather data, OPEC oil supply data

1.1 EPA eGRID:

The EPA eGRID database is a census of all electricity-generating power plants in the U.S. that report data to the U.S. government, which represents almost all electricity-generating power plants. We used two of the tables: the first aggregated demographic information in a 3-mile radius surrounding plants, and the second described each plant's fuel type and output. Each row represents one power plant, and contains percentiles of demographic factors. We dropped rows that were empty in any of the columns of interest, which shrank the dataset from 12,612 to 7,756 rows. The eGRID documentation cites inconsistencies between data sources, missing data, and ambiguous data for null values. Because there was no further explanation provided, we cannot safely claim the plants used for analysis are independent and identically distributed thus it is possible that excluding plants with missing data introduced bias into the analysis.

1.2 EJScreen:

To conduct A/B testing, we needed the same demographic data as with eGRID, but for areas in the US without plants. We used the eGRID's data source for all demographic metrics, the EPA's EJScreen dataset, to calculate demographics for non-host communities. One row in the EJScreen data represents a census block group. There are seven socioeconomic indicators featured in eGRID and thus queried from EJScreen: People of Color, Low Income, Unemployment Rate, Limited English Speaking Household, Less Than High School Education, Under Age 5, and Over Age 64. The socioeconomic data for EJScreen is sourced from the U.S. Census Bureau's American Community Survey (ACS) 2018-2022 5-Year Estimates (ACS 2022), with Block groups as the smallest granularity publicly available. The EJScreen data was removed from the EPA site under the Trump Administration, but the tool is being patronized by

The Public Environmental Data Partners, and the data by The Environmental Data & Governance Initiative (EDGI) (Data).

1.2.1 EJScreen Data Cleaning and Preprocessing:

To create a dataset of non-host communities, we first used separate census data to query each block's latitude and longitude, and scikit-learn's nearest neighbor BallTree package to identify the nearest neighboring plant to each block group. If the distance was greater than 3 miles or less than 50 miles, we included the block and aggregated them by county. We then calculated the seven socioeconomic factors as detailed in the EJScreen Technical Documentation, dividing by the appropriate population for percentiles; for example, Less Than High School Education was divided by the number of people over 25 rather than total population.

Defining Host and Non-Host Communities of Utility Plants

Following the eGRID classification, any block group within 3 miles of a plant is considered included in a host community, aggregated by county. Non-host communities were considered block groups within 3 to 50 miles of a plant in the eGRID data, aggregated by county. Block groups whose block centroid was beyond 50 miles were excluded from the analysis. The design choice to limit nonhost communities to a defined radius around hosting communities was informed by a similar study, which argued this would control (partially) for potential confounding factors such as 'regional differences in demographics, urban development, and energy resources,' (Cranmer). While these boundaries were empirically informed, they introduce a degree of arbitrary thresholding that may impact our model's decision outcomes based on the refutable definitions of host versus non-host versus excluded areas.

1.3 EIA:

We used the US Energy Information Administration (EIA) datasets "Domestic Crude Oil First Purchase Price by Area" (1), "Fossil Fuel Consumption For Electricity Generation by Year, Industry Type, and State" (2), and "Total Energy Prices and Expenditures (total, per capita, and per GDP)" (3). In the context of the causal inference research question, the crude oil price data was used to acquire treatment data, the fossil fuel consumption data was used to acquire outcome data, and the energy price data was part of our confounding variable data. All of these datasets would be most accurately categorized as regulatory census data. The fossil fuel consumption and crude oil price datasets are at the granularity of state-level and monthly frequency, while the energy price data is at the state-level and yearly frequency. We dealt with NaN values in the crude oil price data by dropping rows that contained such entries. With regards to data accuracy, these datasets represent the gold standard in reliability due to the fact that they come from government censuses.

1.4 Federal Reserve Bank of St. Louis:

We used the Federal Reserve Bank of St. Louis datasets "10-Year Treasury Constant Maturity Minus 2-Year Treasury Constant Maturity" (1), "Inflation, Consumer Prices in United States" (2), "Industrial Production: Total Index" (3). These data sources were included to account for macroeconomic confounding influences in our causal analysis (e.g., production levels would in

theory influence fossil fuel consumption). The treasury yield spread and industrial production index data are available at the monthly frequency for nation-level, and the inflation data is available at the annual frequency for nation-level. These data sources would classify as economic indices published by government entities (like the Bureau of Labor Statistics). These datasets contained some missing values that had to be removed.

1.5 NOAA:

We used average temperature data at the state-level and monthly frequency from NOAA. The goal was to account for another potential confounder in our causal analysis (e.g., temperature anomalies in theory would influence fossil fuel consumption). This data source would be classified as a quantitatively measured dataset published by a government agency, which means it represents a high degree of reliability. There were no missing values in this dataset.

OPEC:

We used the “World Crude Oil Production By Country” dataset from the Organization of Petroleum Exporting Countries, which is a regulatory census of total oil production by annual frequency, country-level granularity. The reasoning behind including this data source was that global oil supply could be another potential confounder in our causal analysis.

2. Research Questions

Research Question 1 (Multiple Hypothesis Testing):

Are power plants disproportionately located in areas with marginalized populations? This question about environmental justice has the potential to inform policy on regulations, allocation of funding for public health initiatives, and the placement of new plants. We will use multiple hypothesis testing to answer this question, which allows us to examine several facets of marginalization both separately and aggregately, and fits our imperative to compare two populations. Multiple hypothesis testing is limited to a binary decision, to reject the null hypothesis or not, and does not allow us to make inferences about causation. It is also subject to false positives and p-hacking, which could lead us to conclude that a result is statistically significant when in fact it occurred by chance.

Research Question 2 (Causal Inference):

Does monthly state-level crude oil first purchase price ($/barrel) causally affect monthly state-level petroleum-derived electricity consumption (MWh)? Understanding this casual relationship can influence real-world efforts related to transitioning from fossil fuel derived energy use to clean energy use. Causal inference is relevant here because it's important to understand how energy supply chain factors (such as oil price) directly affect fossil fuel consumption behavior; the magnitude of said relationship would be an important datapoint for shaping how we approach clean energy efforts. A major limitation of this approach is accurately tracking all confounding variables related to oil price and fossil fuel consumption, given that a vast range of economic and supply chain variables could influence this relationship.

3. Prior Work

Research Question 1 Sources:

1. Cranmer, Z., Steinfield, L., Miranda, J., & Stohler, T. (2023). Energy distributive injustices: Assessing the demographics of communities surrounding renewable and fossil fuel power plants in the United States. Energy Research & Social Science, 100, 103050. https://doi.org/10.1016/j.erss.2023.103050

   In this study, the researchers look at a utility plant's fuel type and socioeconomic demographic data to empirically review the distributive injustices related to energy siting, inquiring whether renewable utility plants are continuing a long history of utility plant developments in marginalized communities. Their methodology to compare 'host' and 'nonhost' communities' association with plants was a two-sided Mann-Whitney U test with a Bonferroni correction. The researchers also used the American Community Survey for demographic information, and power plant data from the EIA.

Research Question 2 Sources:

1. Barrales-Ruiz, Jose, and Pablo Neudörfer. "The oil price (Ir) relevance for global CO2 emissions." *Energy Reports* 11 (2024): 3016-3021.

   This study analyzed the relationship between oil price changes and global CO2 emissions. The researchers employed an instrumental variable approach using oil supply shocks and economic shocks to reduce confounding influences in their analysis. They found that specifically shocks in price due to oil supply had the most significant effect on global emissions increases. Based on this finding, we opted to include data related to oil supply in our dataset as a confounding variable in our causal inference.

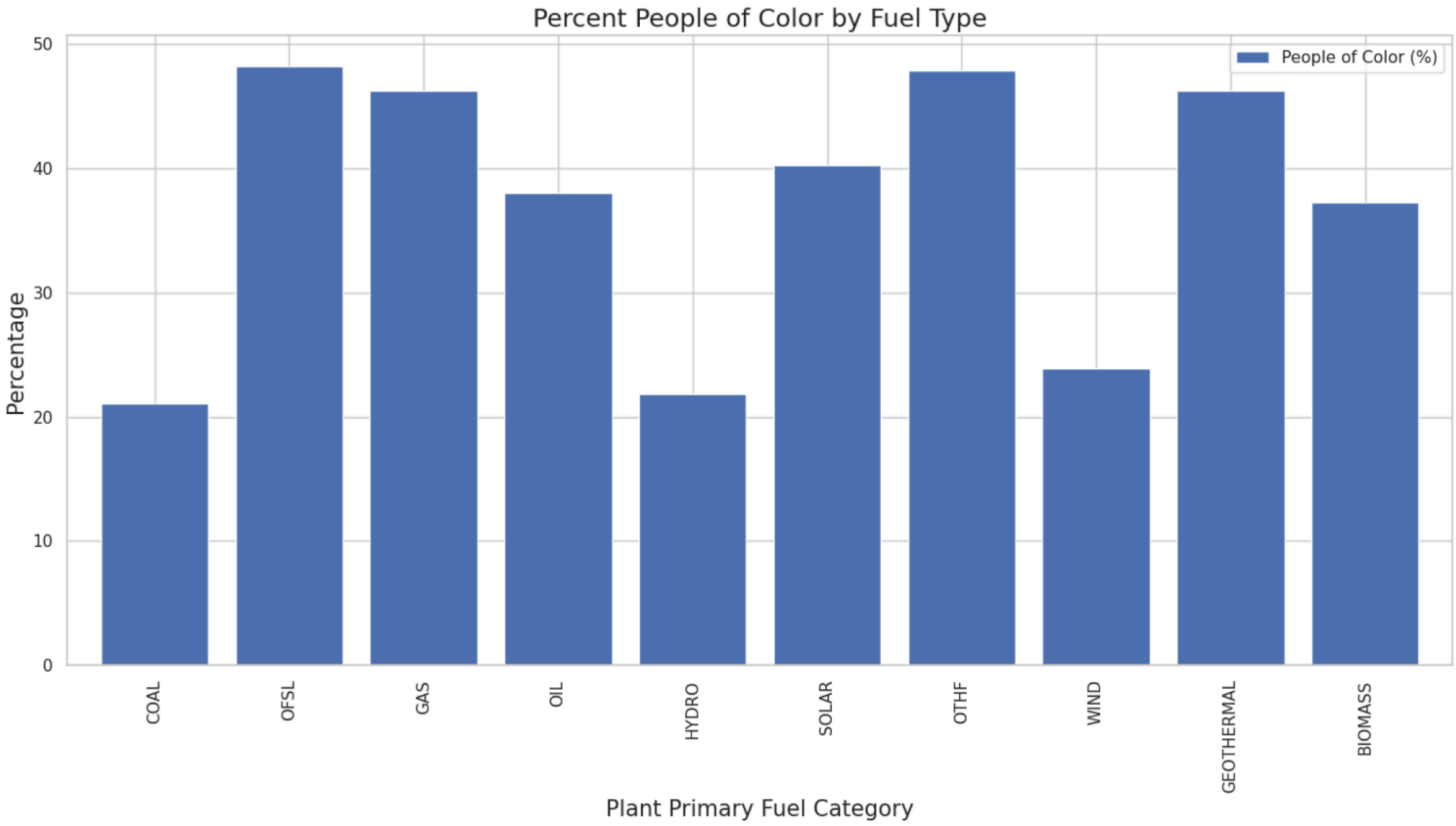
2. Haque, Mohammad Imdadul. "Oil price shocks and energy consumption in GCC countries: a system-GMM approach." *Environment, Development and Sustainability* 23.6 (2021): 9336-9351.

   This study analyzed what factors (among energy prices, income levels, etc) most strongly predicted energy consumption in Gulf Coast Cooperation Countries. They found the strongest positive relationship between income level and energy consumption and strongest negative relationship between crude oil price and energy consumption. This study was relevant because it showed us that including economic variables that gauge macroeconomic circumstances and industrial output are important for our causal inference.

4. EDA

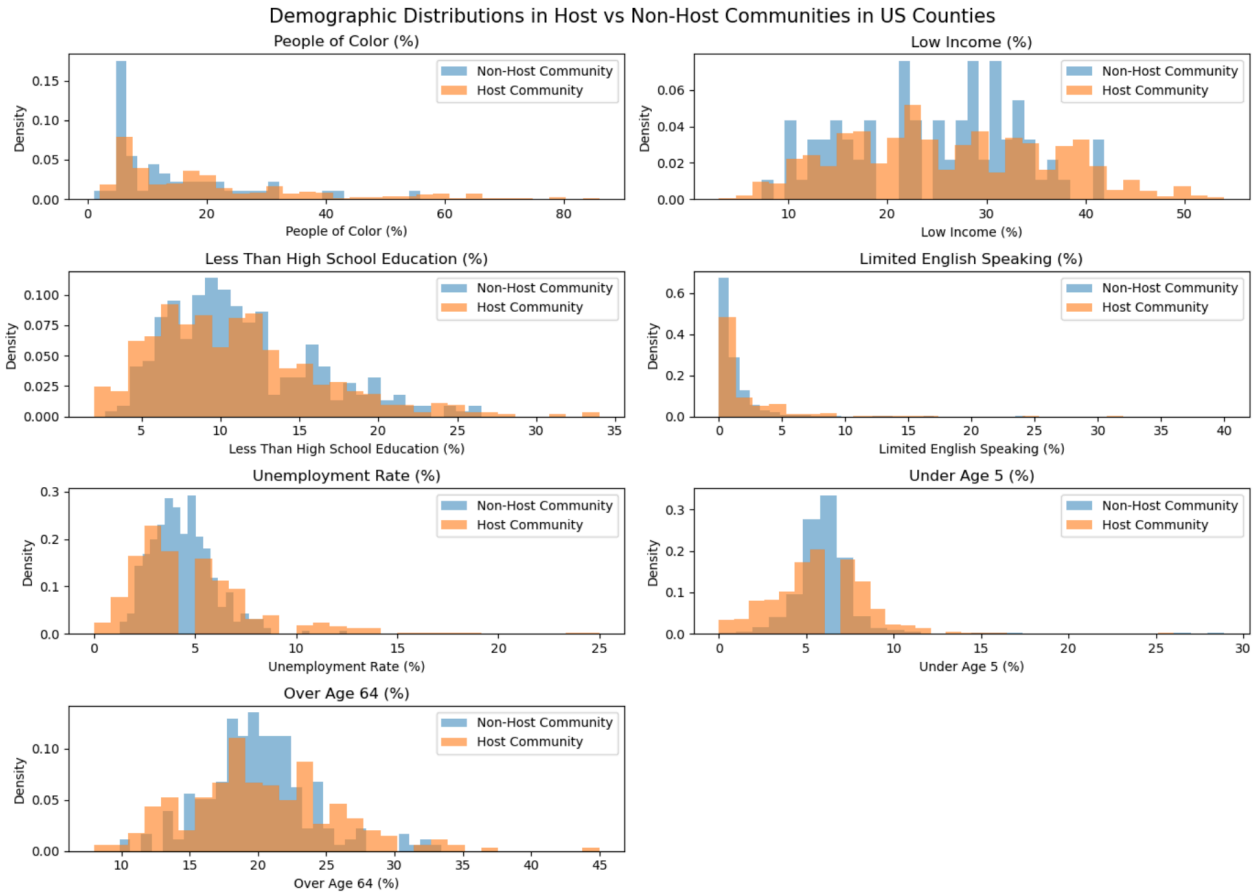
Research Question 1:

Fig. 1: Percent People of Color by Plant Primary Fuel Type



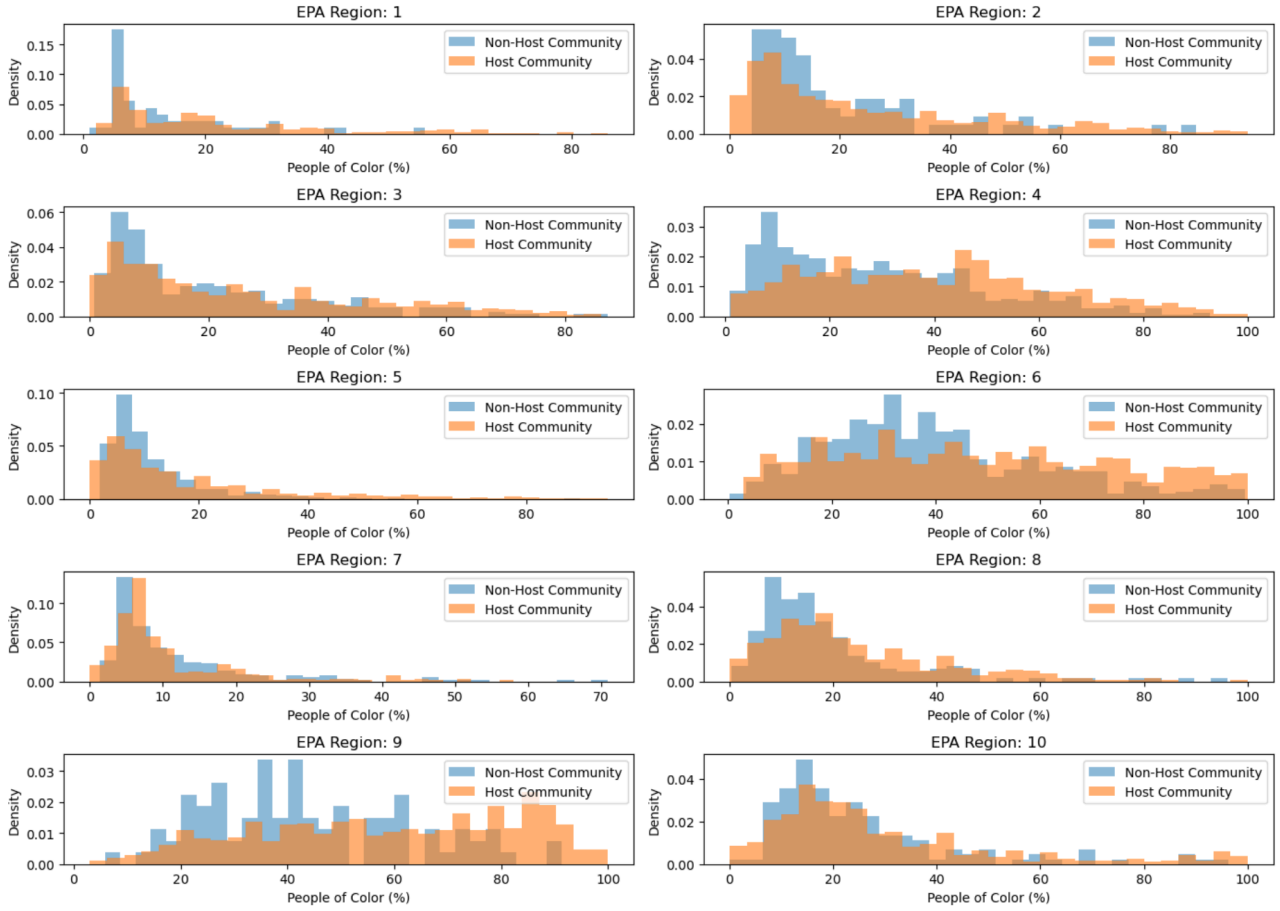
We wanted to investigate if one of our variables of interest, Percent People of Color, varied by the fuel type of the plant. We observe that coal, hydro, and wind plants have significantly lower percentages than other fuel types. This suggests that the disparities in demographic features between areas with and without plants are nuanced, our results without stratifying by fuel type may be overly generalized.

Fig. 2: Distributions of Demographic Factors Among Areas With and Without Plants



As we are interested in the difference in demographic factors between counties with plants and those without, we have visualized each factor of interest, stratified by whether there is a plant, as a histogram. For all of the factors, the distributions of counties with and without plants are quite similar, though for percent low income, percent over age 64, and especially percent under age 5, the spread of the counties without plants is larger than counties with plants. For percent people of color, the distribution of counties without plants is more strongly skewed right. With the exception of percent people of color and percent limited English speaking, the differences in means are not immediately apparent, signaling the need for numerical analysis.

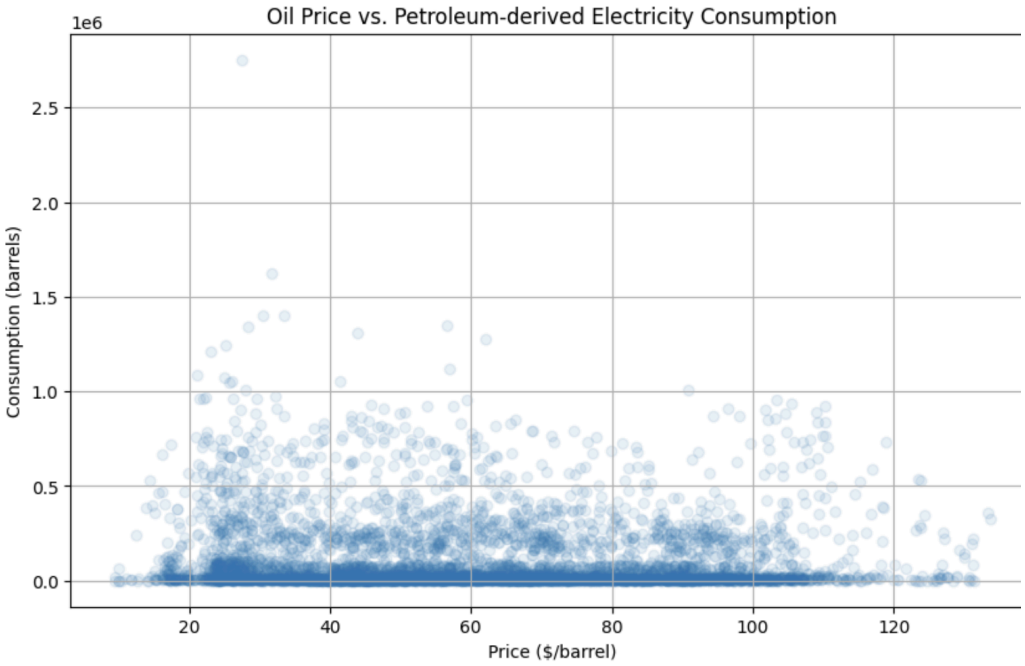
Fig. 3: Percent People of Color Among Areas With and Without Plants by EPA Region



In order to further investigate the disparity in percent of people of color in host and non-host areas, we visualized the same histogram from above separated by EPA region. This reveals that while some of the distributions look nearly identical between areas with plants and without, in some (4, 6, 9) the difference is more apparent and there is a visible difference in means with host communities having on average a greater percentage of people of color. These are the three regions covering the southeast, south central, and southwest states in the country, respectively, suggesting that states with larger shares of people of color tend to have larger disparities in plant location.

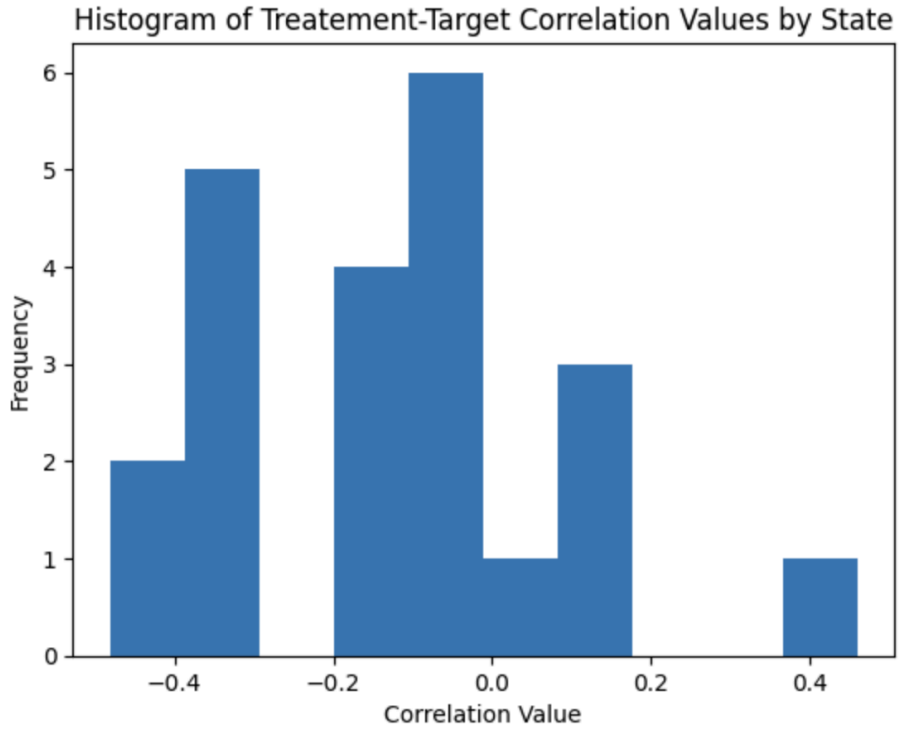
Research Question 2:

Fig. 1:



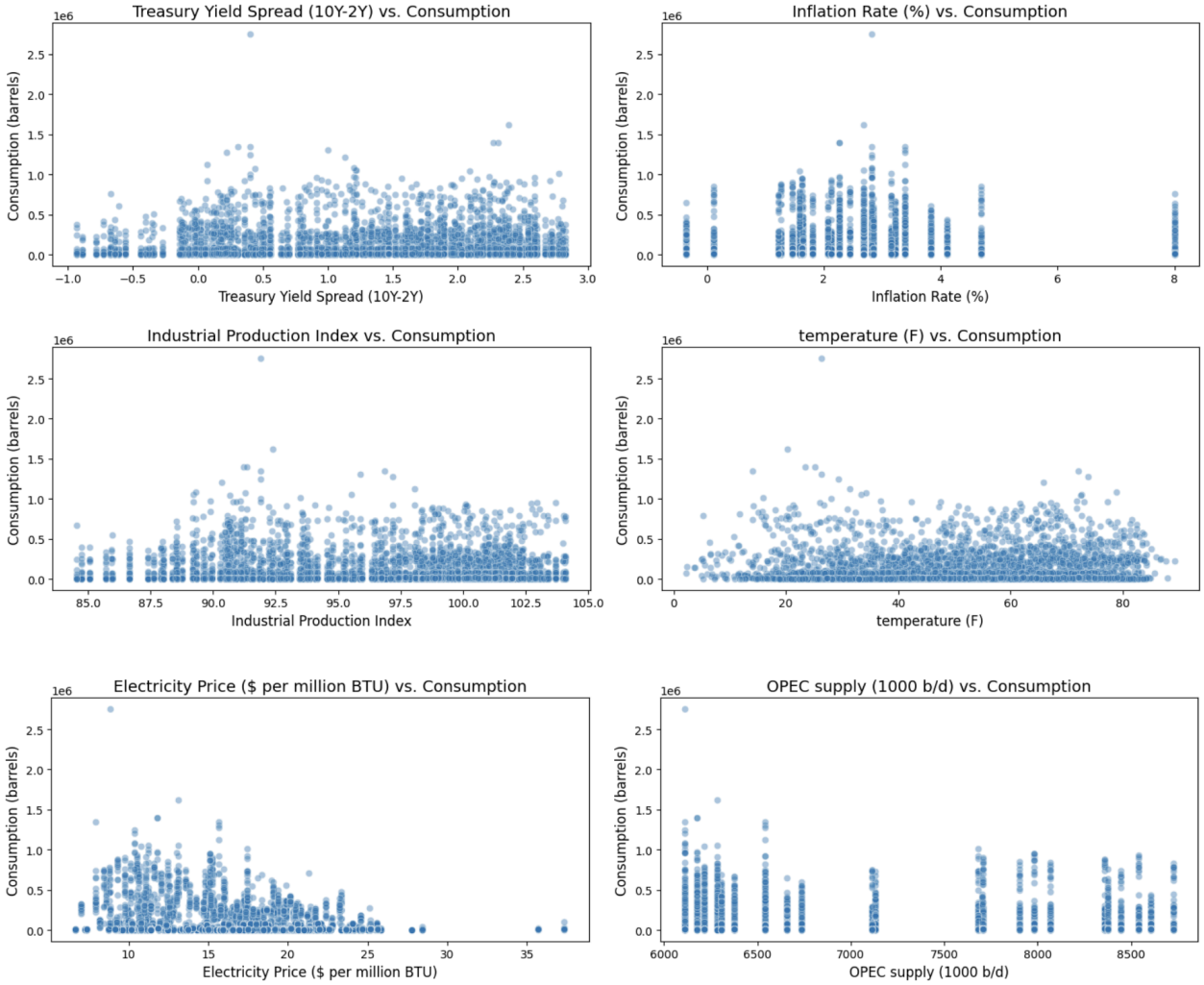
The above is a scatterplot of the treatment and target variables in our data. This visualization showed us that there was only a slightly negative correlation between the treatment and target variables, which indicated to us that extensive feature engineering to identify confounding variables would be critical for discerning a meaningful causal relationship between the variables.

Fig. 2:



The goal of this visualization was to understand how the correlation coefficient values between the treatment oil price variable and target consumption variable varies state-by-state. While correlation is different from causal relationship, understanding the associations between these two variables was an important context for our analysis. The main takeaway here was that we should one-hot encode the state column to enable our linear regression model to learn state-by-state differences in consumption patterns (confounding variable), though even this may only be partially effective at mitigating the effect of state-by-state consumption pattern differences in our model given our dataset size (~6000 cumulative datapoints).

Fig. 3:



The above set of visualizations are scatterplots between our confounding variable features and the target variable. The goal here is to understand the associations between these variables and the target variable. Identifying features that correlate well with the target variable is highly challenging in this domain, likely due to the large number of variables that can influence consumption. Even after multiple iterations of feature selection, our feature set showed weak association with the target variable, which is an important piece of context when evaluating our modeling results. Essentially, there's little evidence to support the assumptions we've made specifically with respect to confounding variables.

5. Inference and Decisions

Research Question 1: Multiple Hypothesis Testing

Methods:

Our general hypothesis is that plants are disproportionately placed in marginalized and under-resourced communities. The characteristic of a marginalized community is not singular, which elicits the need to conduct multiple hypothesis testing to observe potential inequalities across multiple socioeconomic demographic metrics. EJScreen identified seven general socioeconomic metrics as indicators of a community's potential susceptibility to environmental factors. These indicators form the basis of our hypothesis formulization: Low Income %, People of Color %, Less Than High School Education %, Limited English Speaking Households %, Unemployment Rate %, People Under Age of 5 %, and People Over Age 64%.

Assumptions:

- The plants left after removing those with missing values are representative of the population.
- The seven socioeconomic metrics are not independent of one another.
- Our definition of host community and non-host community are representative of the areas implicated in utility plant site allocations.
- Confounding factors such as regional differences in demographics, urban development, and energy resources are (partially) controlled for by limiting the non-host communities to within 50 miles of a utility plant.

Null Hypothesis:

In the United States, the distribution of key socioeconomic demographic indicators is the same for host and non-host communities of utility energy plants.

Alternative hypotheses:

1. Plants are disproportionately located in areas with higher % People of Color.
2. Plants are disproportionately located in areas with higher % Low Income.
3. Plants are disproportionately located in areas with higher % Less Than High School Education.
4. Plants are disproportionately located in areas with higher % Limited English speaking households.
5. Plants are disproportionately located in areas with higher % Unemployment Rate.
6. There is a difference in the percentage of people under age 5 for counties with plants vs. without.
7. There is a difference in the percentage of people over age 64 for counties with plants vs. without.

Specific Alternative Hypothesis:

- **Null:** In the United States, the distribution of People of Color % is the same for host communities as for non-host communities, on average.
- **Simple Alternative:** In the United States, host communities have 3% more People of Color, on average, than non-host communities.

The 3% difference for the simple alternative was chosen based on a similar study which found that communities with oil plants were 2.6% more Black, 4.2% more Latinx, and 1.4% more Asian than communities without oil plants (Cushing 2023).

We assumed a Normal distribution for the difference in People of Color % between host and non-host communities, centered at a 0% difference under the null hypothesis and 3% for the alternative hypothesis. The standard deviation for the normal distributions of the null and alternative was determined using the standard error of the difference between means ("Difference in Means").

With a significance level of 0.05, an observed difference of 11.89, and threshold of 0.82, we rejected the null. The power (TPR) of the test was 0.99.

A/B Testing for Multiple Hypotheses

To test our multiple hypotheses, we conducted A/B testing against each hypothesis. We created a binary treatment variable, where host communities are assigned to the 'treatment' group (value=1), and non-host communities to the 'control' group (value=0). We chose A/B testing because the treatment and the control was a well-defined binary variable, and our interest was in the difference in means.

To correct for the multiple hypothesis tests error rates, we will apply two different methods:

1) To control for the False Discovery Rate (FDR) at a threshold of 0.05, we use the Benjamini–Yekutieli Procedure, which controls the FDR under arbitrary dependence assumptions across the hypothesis tests. This procedure is more appropriate than the Benjamini-Hochberg Procedure because the seven socioeconomic metrics are not independent (Massey and Denton 2018).

2) To control for the FWER at 0.05, we apply the Bonferroni correction.

Controlling for FDR is more appropriate for our research question because the consequences of a false positive are not extremely serious – the aim of the research is to explore any sign of inequitable plant siting practices. Controlling FWER favors more false negatives over false positives, so it is less appropriate given the intentions of the hypothesis formulation.

Results

After conducting our A/B permutation tests, the p-values for the difference in means of each demographic factor between host and non-host communities are as follows:

People of Color (%): 0.0,
Low Income (%): 0.0,
Less Than High School Education (%): 0.0064,
Unemployment Rate (%): 0.0,
Limited English Speaking (%): 0.0,
Over Age 64 (%): 1.0,
Under Age 5 (%): 0.994

The statistically significant p-values (below 0.05) include People of Color (%), Low Income (%), Less Than High School Education (%), Unemployment Rate (%), and Limited English Speaking (%). This suggests that the observed difference of these metrics between host and non host communities are unlikely under the null hypothesis. In real world terms, this suggests marginalized communities are disproportionately located near a utility-level energy plant. The two age metrics, Under Age 5 (%) and Over Age 64 (%), attained a p-value close to 1, suggesting the observed difference was due to chance and reasonable under the null hypothesis.

Discussion

It is important to distinguish what decisions should be made from individual tests compared to aggregate hypotheses results. While the individual tests reveal some insight into the likelihood of each hypothesis under the null, decisions based on these answers should not be made in isolation. It is important to consider the results in aggregate and the cumulative chances of false positive rates with multiple hypothesis testing when using the results to draw conclusions.

- Controlling FWER with the Bonferroni correction, the adjusted threshold is 0.0071, with five discoveries made: People of Color (%), Low Income (%), Less Than High School Education, Unemployment Rate (%), and Limited English Speaking Households (%). This threshold is set to control the probability of making any false discoveries among any of the seven tests to be below 0.05.
- Controlling FDR with the Benjamini–Yekutieli procedure, the adjusted threshold is 0.0064, with five discoveries made: People of Color (%), Low Income (%), Less Than High School Education, Unemployment Rate (%), and Limited English Speaking Households (%).

To ensure we avoided p-hacking, we did not stray from our original hypothesis formulation – once our p-values had been calculated, we did not augment any of our prior theories or methodology, or add any additional tests.

If additional data were available, we would conduct further analysis into the disaggregated general demographic indicators. While the general socioeconomic indicators tested in this research managed to capture a general picture of the disparities in utility plant locations, it failed to explore nuance within each demographic. Relying on a binary ‘White’ versus ‘non-White’ framework generalizes experiences of diverse racial groups that are by no means monolithic. If the eGRID data included more specific racialized demographic breakdowns, we would have run our hypothesis testing across these groups instead of a summarized ‘People of Color’ percentage.

The findings of our work were consistent with the similar study cited, although our findings are less detailed, having no differentiation between fuel type, region, or race. Cranmer et al. likewise found evidence of site distribution inequalities, but these fell along varied demographic lines when accounting for fuel type; for example, coal and wind plants tended to be located in whiter communities with high rates of English proficiency and lower income. They also found variation across states; for example, states with higher proportions of people of color showed more discrepancy in plant locations.

Research Question 2: Causal Inference

Methods:

The treatment variable is monthly state-level first contact crude oil price ($/barrel). The outcome variable is monthly state-level consumptions of petroleum-derived electricity (barrels).

The confounders included as features are monthly treasury yield spread (10Y-2Y), annual inflation rate (%), monthly industrial production index, state-level monthly average temperature (F), state-level yearly average electricity price ($ per million BTU), and yearly OPEC total oil supply (1000 b/d). The objective with this selection of confounding variables was to account for the main external factors that would influence the treatment and target (namely economic conditions, industrial output, temperature, and oil supply). We can't say with certainty that the unconfoundedness assumption is fully met — there could realistically be hundreds of confounding variables related to the factors we identified, but it wouldn't be feasible to control for each one in the scope of this project.

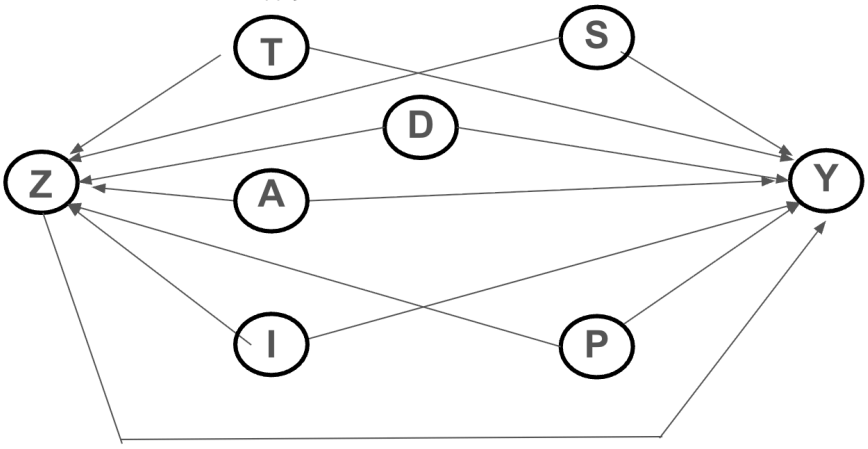
To adjust for confounders, we chose to use inverse propensity weighting. We chose this method because satisfying the conditions for IPV (e.g, correlate with treatment but not outcome, etc) were difficult and applying a meaningful matching mechanism to many continuous variables would not have been feasible. We chose to binarize our treatment variable based on the median value (with the goal of choosing a value robust to outliers). We fit a logistic regression model on our confounding variables to generate propensity scores for our treatment values. We trimmed the bottom 2.5% and top 2.5% of propensity scores to handle extreme values in our dataset. Then we used the Horvitz Thompson estimator (ATE formula from class) to measure the ATE.

There are no obvious colliders among our feature set. One could make the argument that the feature electricity price could affect the outcome variable, we feel that this feature is better kept as a confounder. The logic behind this is that we assume that electricity price would influence people's willingness to consume crude oil and petroleum-derived electricity, thus affecting the treatment price and outcome consumption variables.

Causal DAG:

Key:

- Z: monthly state-level crude oil first purchase price (treatment)
- Y: monthly state-level consumption of petroleum-derived electricity (outcome)
- T: monthly treasury yield spread
- A: annual inflation rate
- I: monthly industrial production index
- S: monthly state-level avg. temperature
- P: annual state-level energy prices
- D: OPEC oil supply



Results:

We used the Horvitz-Thompson ATE to estimate the causal effect of the treatment. As described above, we binarized our treatment based on the median value, derived propensity scores using logistic regression, and used the Horvitz-Thompson estimator to find the ATE. The result of the ATE estimate was roughly -6915, which can be interpreted as each increase in state-level monthly first contact crude oil price by $1 per barrel led to an decrease in state-level monthly consumption of petroleum-derived electricity of about 6915 barrels. Caveats to this estimate though are the fact that we used weight trimming (remove bottom and top 2.5% weights) to prevent extreme weight values, we binarized the treatment column relative to median value, and were not able to include every confounder in our feature set (due to impossibility).

Discussion:

Our model yielded a negative causal effect between the treatment and outcome variables, which aligns with findings from studies (such as study 2 in previous work) that demonstrate a negative relationship between oil price and energy consumption. While the direction of the causal effect aligns with economic intuition and previous studies, we are not highly confident in the magnitude of our computed causal effect.

There are multiple potential sources of noise that could have influenced the accuracy of our computed causal effect. One source could be incorrect assumptions regarding casualty between variables. While our current feature set represents our best attempt at capturing the sources of influence given the time provided, there are likely more features, such as clean energy indices, which may have been beneficial additions to our feature set that more accurately fit the confounder requirements or provided necessary exposure to influencing factors (economic,

industrial, etc). We saw in our EDA that there were weak associations between the features and the target variable, which reduces our confidence in our model results. Furthermore, we included a range of states in our dataset, which we demonstrated in the EDA section following somewhat different consumption patterns evidenced by the oil price-consumption correlations between states. This would be a significant source of noise even when using a method like one-hot encoding to make state-by-state behavior distinguishable in our model. Binarizing the treatment variable in our ATE estimate is also a possible source of noise given that the magnitude of the treatment value would be important in estimating accurate propensity scores. ATE is an unbiased estimator when dealing with binary treatments; when binarizing a continuous treatment variable to use this approach, some accuracy in propensity scores is lost. The magnitude of the treatment should, in theory, be an important factor in the propensity score calculation, but excluding the magnitude was a necessary tradeoff to find our estimated ATE. There are continuous IPW methods that fit a density function to the treatment variable, though these methods are highly complex and were not included in this analysis due to class scope.

Another important limitation is the potential for reverse causality in our treatment and outcome variables. We chose these variables under the assumption that oil price would influence energy consumption behavior and be primarily influenced by oil supply chain conditions (following economic principles). However, after going through the modeling process, it's possible that energy consumption, specifically driven by local economic activity, drives oil price changes when we look at the granularity of individual states in the US. Most studies look at oil supply from a global perspective, which introduces different dynamics than state-level granularity. Reverse causality in this case would lead to incorrect conclusions in our causal inference.

6. Conclusion

Conclusion - Research Question 1: Multiple Hypothesis Testing

Outcomes summary:

RQ1:
We found evidence to support our hypothesis that power plants are disproportionately located in areas with marginalized populations. After controlling FDR, we still find a statistically significant difference in means of People of Color (%), Low Income (%), Less Than High School Education, Unemployment Rate (%), and Limited English Speaking Households (%) among host and non-host communities.

Critical Evaluation:

The limitation in the data we could not account for was missing data values - there were approximately 5,000 records with important missing values of a dataset of 12,612 plants, and we omitted these plants entirely. This may have introduced bias if plants with missing values were non-random or of a certain characteristic. Additionally, these results are not generalizable to specific regions of the country or to specific racial and ethnic groups - they only point broadly to systemic inequalities in the locations of power plants across the United States.

We are missing domain knowledge about how power plants are regarded in society and their unique impact on the surrounding communities. We would ask an expert: Are there types of power plants which are more desirable than others? Should host and non-host community radii be defined dependent on plant type (e.g. do wind plants have a 1 mile impact, while oil-ran utility plants have more?) How is the location of utility plant siting regulated by the government, and in what scenarios are community concerns considered? Further insight into the regulatory, public perception, and health outcomes of specific utility plant fuel types would have enabled us to run more informed and nuanced hypothesis tests to investigate whether the disparities in socioeconomic factors vary based on the desirability of plant type and community resources.

While the conclusions are robust given our modeling choice, the radius threshold of host and non-host communities we defined could be refuted and/or altered, potentially impacting the conclusions. Studies measuring the impact of the proximity to a plant vary from 1 to 5 miles, and the 50 mile cutoff for non-host communities is a generalization given fuel types differ in potential site location radii (Cranmer 2023). Usage of our findings should consider these informed boundary thresholds to hold a degree of arbitrary, and further studies should be made to explore the appropriate geographic boundaries of the direct impacts of proximity to utility plants.

Recommendations:

We propose a future study that considers intersectionality, which could reveal that some of the marginalized identities we found to be associated with power plant locations may vary across identities or only maintain that association in combination with other demographic metrics. To further assess the nature and severity of these environmental inequities, the study could account for types of power plant, available community resources, and more specified demographic indicators to assess the burden placed on the community around it and produce more actionable insights for policy coordination.

Based on our findings, we recommend that governments prioritize marginalized communities in the allocation of funds for air quality monitoring and regulation, and better regulate and formalize how utility locations are chosen. Power plants are regulated by the government, but the study shows it is not adequate enough to protect under-resourced communities from predatory utility development practices. Underserved communities are subjected to the placement of utility plants nearby while wealthier areas have the political and economic resources to resist them

(Cranmer 2023). In practice, the onus is often imposed on community organizations to advocate against developments, and policy should be developed to support such organizations for improved monitoring and stricter regulations. Further, the location selection of new plants should be embedded in a stricter framework that accounts for more equitable distribution across community demographic group profiles. The needs of the most vulnerable populations should be at the forefront of conversation to combat historical patterns of injustice.

Conclusion - Research Question 2: Causal Inference

Outcomes summary:

The interpretation of our ATE is that an increase in state-level monthly first contact crude oil price by $1 per barrel leads to an decrease in state-level monthly consumption of petroleum-derived electricity of about 6915 barrels, which corroborates the direction of the causal relationship that previous studies had found.

Critical Evaluation:
While we did use an extensive selection of confounders in our analysis, the sheer complexity of the topic means absolute confidence in the comprehensiveness of our list would be impossible. It was critical to include every confounding variable such that our causal inference assumptions would be satisfied, but overloading our model with parameters could easily lead to overfitting, especially given the amount of noise in the data.

None of us are professionals in the field of contemporary petroleum economics, which is certainly an incredibly complex and nuanced area of study. If given the opportunity to solicit a domain expert for advice, we would ask for a **list of outside factors that frequently affect both oil supply and demand** for use in our model. Having this for our model would provide a much greater confidence that we had achieved a sufficient threshold of confounders to make causal inferences.

We used a pure ATE estimation to model the impact of oil prices on oil consumption. We could've used a generalized linear model, and we indeed applied one to our data, but conclusions from it were inconsistent and tended to overfit on our training set. Going with a direct ATE estimation means the model is more robust, but relies on fewer assumptions and uses inverse proportional weighting, which may bias the conclusions. If a potential user is concerned about the impact of weighting, then they should revert to the linear model instead.

Our findings performed well under cross-validation, and the variables are not too specific to our situation, which means the data is likely generalizable across the United States and over the last several decades. It is possible that nations outside the US (which we did not include data for) may have different relationships between oil price and consumption, which narrows the breadth of our findings.

Recommendations:

A future study could gather additional data from before 2000 after 2025, allowing them to expand the range of the data. Alternatively, they could analyze data from one or more other countries with differing national relationships with oil. Either of these paths could support or refute our conclusions.

Policymakers seeking to improve price conditions for their constituents should use oil price predictions to inform oil reserve rates — by predicting when and where oil prices will spike or drop, they can “smooth the edges” by directly growing or shrinking oil supply. The United States Department of Energy already has a Strategic Oil Reserve, which has been in place since the 1973-1974 oil embargo. However, individual states do not often hold reserves of petroleum at that scale. If US states imitated the DOE’s practice in this way, this would likely supplement the list of government benefits that American energy companies receive.

The results of this analysis are also relevant to efforts related to pivoting from fossil fuels to clean energy. This specific causal relationship provides insight into the magnitude of the effect that oil price changes have on behavior related to fossil fuel use. For instance, this could inform the estimated decrease in oil reserve supply needed to cause a 20% reduction in fossil fuel use at a local or global level.

Works Cited

EJScreen Environmental Justice Mapping and Screening Tool EJScreen Technical Documentation for Version 2.3. 2024. https://www.epa.gov/system/files/documents/2024-07/ejscreen-tech-doc-version-2-3.pdf

American Lung Association. (n.d.). *Disparities in the impact of air pollution*. https://www.lung.org/clean-air/outdoors/who-is-at-risk/disparities#:~:text=Some%20studies%20have%20found%20that:%20*%20Poorer,live%20in%20communities%20that%20are%20predominately%20white.

Cranmer, Z., Steinfield, L., Miranda, J., & Stohler, T. (2023). *Energy distributive injustices: Assessing the demographics of communities surrounding renewable and fossil fuel power plants in the United States*. Energy Research & Social Science, 100, 103050. https://doi.org/10.1016/j.erss.2023.103050

Cushing, L.J., Li, S., Steiger, B.B. et al. *Historical red-lining is associated with fossil fuel power plant siting and present-day inequalities in air pollutant emissions*. Nat Energy 8, 52–61 (2023). https://doi.org/10.1038/s41560-022-01162-y

Data. “Data + Screening Tools.” Data + Screening Tools, 2025, https://screening-tools.com/epa-ejscreen.

Massey, D. S., & Denton, N. A. (2018). *American Apartheid: Segregation and the Making of the Underclass*. In *Inequality in the 21st Century: A Reader* (pp. 142-150). Taylor and Francis. https://doi.org/10.4324/9780429499821-27

“Difference in Means.” Stattrek.com, https://stattrek.com/sampling/difference-in-means.

Liu X, Lessner L, Carpenter DO. *Association between residential proximity to fuel-fired power plants and hospitalization rate for respiratory diseases*. Environ Health Perspect. 2012 Jun;120(6):807-10. doi: 10.1289/ehp.1104146. Epub 2012 Feb 27. PMID: 22370087; PMCID: PMC3385425.

Barrales-Ruiz, Jose, and Pablo Neudörfer. "The oil price (Ir) relevance for global CO2 emissions." *Energy Reports* 11 (2024): 3016-3021.

Haque, Mohammad Imdadul. "Oil price shocks and energy consumption in GCC countries: a system-GMM approach." *Environment, Development and Sustainability* 23.6 (2021): 9336-9351.

Data Access Links

Research Question 1:

- FITZGERALD, W., & Gehrke, G. (2025). EPA Environmental Justice Screening Tool (EJ Screen) data, 2015-2024 [Data set]. Zenodo. https://doi.org/10.5281/zenodo.14767363
- US Census Bureau. Accessed December 9, 2025. https://www2.census.gov/geo/docs/reference/cenpop2020/blkgrp/
- U.S. Environmental Protection Agency. (2023). Emissions & Generation Resource Integrated Database (eGRID2023). ST23 and DEMO23 Tables. https://www.epa.gov/egrid/detailed-data

Research Question 2:

- "Domestic Crude Oil First Purchase Prices by Area." *Www.eia.gov*, www.eia.gov/dnav/pet/PET_PRI_DFP1_K_M.htm.
- "Detailed State Data." *Www.eia.gov*, 16 Oct. 2025, www.eia.gov/electricity/data/state/.
- Federal Reserve Bank of St. Louis. "10-Year Treasury Constant Maturity Minus 2-Year Treasury Constant Maturity." *FRED, Federal Reserve Bank of St. Louis*, 1 June 1976, fred.stlouisfed.org/series/T10Y2Y.
- ---. "Inflation, Consumer Prices for the United States." *Stlouisfed.org*, 16 Apr. 2025, fred.stlouisfed.org/series/FPCPITOTLZGUSA.
- FRED. "Industrial Production Index." *Stlouisfed.org*, 2025, fred.stlouisfed.org/series/INDPRO.
- "Climate at a Glance | National Centers for Environmental Information (NCEI)." *Www.ncei.noaa.gov*, www.ncei.noaa.gov/access/monitoring/climate-at-a-glance/statewide/time-series.
- "United States - SEDS - U.S. Energy Information Administration (EIA)." *Www.eia.gov*, www.eia.gov/state/seds/seds-data-complete.php.
- "OPEC Digital Publications - Annual Statistical Bulletin." *Opec.org*, 2023, publications.opec.org/asb/Download.