

# Know the Signs: *Housing Choice and the Lived Environment*

*Empowering consumers to make informed housing decisions  
by making environmental data centralized and accessible*



Team 61- DS4A Empowerment

Taylor Brown, Tinika McIntosh-Amouzouvi, Luis Esparza, Ashley Aviles, Chioma Dunkley

# Table of Contents



Introduction.....	3-5
Data Analysis & Computation.....	6-9
Description of Dashboard.....	10-13
Conclusions and Future Work.....	14-15
Contact the Team.....	16
Acknowledgements.....	17



## INTRODUCTION

# Overview of the Problem

The first environmental justice court case to receive national attention took place in the early 1980s in Warren County, North Carolina. One major outcome of this case showed the relationship between hazardous waste sites and the racial and economic status of the surrounding communities. This study revealed something that communities of color and lower income residents knew all along that they are disproportionately impacted by hazardous environmental conditions. However, this national case helped to reveal many other offenses that exist in the environmental sector against people of color and lower income individuals. As race was the greatest predictor of where these hazardous environments were located, we saw the convergence of the environmental movement and the social justice movement into what we know now as environmental justice. It is in this framework that we were able to find the support as well as the need for this environmental scorecard project.

The environment in which we live plays a major role in our physical and mental health and well-being. However, when looking for a place to live it is often very difficult to locate information about important environmental indicators for a specific neighborhood or residential area. This difficulty is even more exasperated for communities of color and for those of lower socioeconomic status, as other factors such as affordability and credit may influence housing choices more than environmental factors.

Compiling pertinent environmental data and making it accessible for everyday consumers would allow them to make more informed decisions around housing. The project aims to compile data related to but not limited to; air quality, noise pollution, water sources, alternative energy use, toxic metal presence, and the presence of green spaces. Additionally, the availability and access to data around environmental factors will also benefit policy makers with decision making around environmental issues as well as those looking to understand the relationship between population health and the environment.



## INTRODUCTION

# Solving the Problem

There is an absence of transparency around environmental factors that have the ability to impact the health and well-being of homeowners and renters. A direct impact of this failed transparency is the concentration of communities of color and those of lower socioeconomic status in areas that contain multiple poor environmental conditions.

This project aims to compile environmental data in order to create an environmental score for all counties in the United States to be used by consumers on housing location platforms. This environmental score would be available on consumer sites in an easy-to-read format for everyday consumers to aid in their home buying or rental decisions. Publicly compiling environmental data in order to make this information more transparent for consumers is also aimed to aid in fair pricing and accountability for landlords and jurisdictions

As environmental factors don't rank high on the list for many people searching for housing it is important for us to make it easier to access and more transparent. This pertinent information directly and indirectly impacts the lives of home-seekers. A central location that contains information about environmental factors can be used by policy makers in order to ensure fair practices around zoning, pricing, and other policies related to housing





# Environment Matters

Transparency in the realm of environmental factors and housing will open the door for better affordable housing that does not adversely impact the lives of the inhabitants. The environmental justice framework supports addressing each of these items as an attempt to decrease disparities that exist around environmental conditions. The environmental justice framework also removes responsibility for proof from harm to the "protected groups" instead of those typically impacted groups. This maintains that already vulnerable groups do not have to prove that something has harmed them before mitigation occurs



There are policies around housing, but these policies typically impact housing at the individual level such as the 1979 ban on lead-based paint. However, there are many factors that have the potential to impact entire neighborhoods as the hazards is not inside of homes but outside in a shared space. As a result, for this project we are looking at community level environmental indicators instead of those at the individual level. The environmental indicators that this project aims to examine are the air quality index, arsenic in community water, greenhouse gas emissions, power plants in the community. We also look at relationships between these indicators and the social vulnerability index and cancer prevalence.





# Datasets

Our team gathered data from the Centers for Disease Control and Prevention (CDC), U.S. Environmental Protection Agency (EPA), and the Power Plants and Neighboring Communities Mapping Tool. We identified a total of 13 data sources, with eight sources containing environmental data and five containing health data. After an extensive cleaning process and significant down scaling of the project scope we only create our final project. using five of those original data sources. Those data sources are greenhouse gas emissions, air quality index, power plant mapping, and social vulnerability index which are listed below.

## **Green House Gasses (GHG) Emissions**

This dataset comes from the EPA's Greenhouse Gas Reporting Program (GHGRP) It contains facility-level greenhouse gas emissions data from large industrial sources across the U.S. The data that we used contained data gathered from 2010-2020 across most green house gas emitters that produced more than 25,000 metric tons of CO2 per year

## **Air Quality Index (AQI)**

This dataset comes from the EPA's Air Quality System (AQS). The data is then collected in the AQS by agencies that submit their data. This data is aggregated and stored in various formats for EPA's internal use. The data that we used was the annually reported AQI by CBSA (metro area)

## **Power Plants**

This dataset identifies the locations of power plants and highlights the key demographics of people living within three miles of those plants. It also displays all fossil fuel-fired power plants that supply electricity to the grid.

## **Arsenic in Community Water Systems**

This dataset comes from the CDC and it is reported as the mean concentration of Arsenic in drinking water (in micrograms per liter) per county at the national level, grouped by county and community water system.

## **Social Vulnerability Index**

This dataset is from the Agency for Toxic Substances and Disease Registry (ATSDR) and the CDC as an effort to identify socially vulnerable populations during public health emergencies. The assumption is that these vulnerable populations are even more at risk than others, so this data helps to identify them and allow proactive interventions.

# Data Wrangling & Cleaning

## Data Cleaning

Our data came from well-established governmental sources, and as a result, our data was mostly well organized and did not possess many missing values. Therefore, when we did see missing values, we realized that they were the result of our own processing and data aggregation, and hence needed further investigation. We realized far too late that our merge was not a quality merge when removing the missing values left us with an empty dataset. However, we solved this issue by using the same python package, `addfips`, to calculate the FIPS codes for every dataset. This ensures that the FIPS codes were in the same format and consistent across datasets and then we merged them one by one into one final dataset containing all the relevant features of the project. Prior to combining the datasets, we also performed some preprocessing on the individual datasets to ensure that all the irrelevant columns were moved, and ambiguous columns were labeled so that the combination of the datasets would still be clear. However, we did still perform quality assurance checks on the final data to ensure there weren't any missing values that could have impacted the results.

## Data integration

To explore our data further, we combined our data using Python. The most granular level in which all our data was available, was the county level. At first, we thought of combining our data sets using the county name itself, but we later realized that there exist multiple counties across different states with the same names. Given that our data also included the uniquely- identified, FIPS code, we decided to merge our data this way to perform our analysis. We used a Python library, `addfips`, to accomplish this task. After having a variable to merge our datasets, we joined each one in pairs to prepare for further analysis.

## Data Transformation

In this initial analysis, we performed minor transformations to our datasets. GHG data contained observations by facility. We therefore combined each observation by county to match the rest of our datasets. As we also noted above, when we aggregated our data, we noticed some NaN values, in particular with the GHG dataset. Some observations contained zeros for the reported CO2 emissions, and this was due largely because facilities that produced less than 25,000 metric tons per year did not have to report their emissions. When our data was imported into python, those zeros were converted to NaN values. For the rest of our datasets any missing values were a small proportion compared to the rest of our data. We concluded that removing them would not have an impact on further analysis for the purpose of our model.

## Data Reduction

We would like to note that we went through this process several times given that our project scope needed to be narrowed down to create a model that the team could complete in the given time period. We started with 13 unique data sources (eight environmental, five health), and after performing some initial analysis and modelling we were advised to pick a few features and their key values to add to the final dataset. This reduction in scope would make it easier for us to understand the data and allocate our time and resources to creating our scorecard. Our methodology for picking the features was based on selecting different domains of environmental health and the data we already had on hand as well as an all-encompassing adverse health indicator. This process was the final process once we narrowed down key indicators that we would eventually be used in our final model.

# Data Wrangling & Cleaning Visualizations

These images display some our data wrangling & cleaning findings for our datasets.

In Figure 1, we were trying to explore our GHG dataset by using the .info() function in Python. In Figure 2, we display our data after we were able to add the FIPS code in order to merge our datasets further.

## GHG



Figure 1

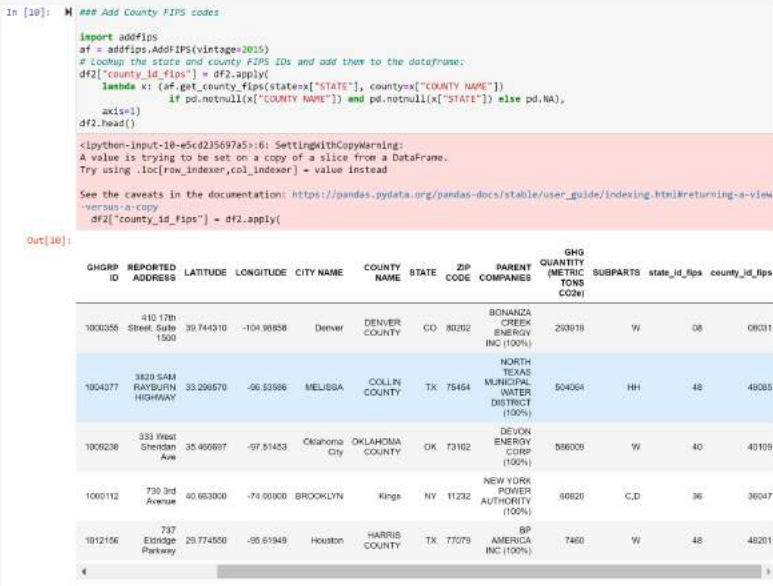


Figure 2

## Ideal format for cleaned dataset

FIPS	Co2 level	Cancer rate in 2012	Asthma in 2012
1001	0.5	3%	12%
48127	10	10%	5%

Figure 4

In Figure 4, we identify the ideal format for our merged datasets. We used this table as a guide when we were conducting our data cleaning.

In Figure 3, we found that some our merging attempts yielding a large amount of NaN values. We were able to identify that this issue resulted from the merging of our datasets. Although we were merging our data using FIPS code, we found that the leading zeros for certain FIPS codes were missing and thus resulted in this error for all of the datasets that we tried merging. This led us to look further exploring other ways to merge our data.

## Too many missing Values

763	13-00430	F	1	2013	...	NaN	NaN
2064	1ST RESPONDER	F	1	2013	...	NaN	NaN
1242	X2013080371	T	1	2013	...	NaN	NaN
2121	2013100343	T	1	2013	...	NaN	NaN
3096	WI DNR	T	1	2013	...	NaN	NaN
1942	DEC1303429	F	1	2013	...	NaN	NaN
2201	1ST RESPONDER	F	1	2013	...	NaN	NaN
1217	NRC # 1055301	T	1	2013	...	NaN	NaN
3009	WDNR	F	1	2013	...	NaN	NaN

	RELS2CHEM5	CHEM6	CHM_QCAT6	CHM_UNIT6	RELS1CHEM6	RELS2CHEM6	TOT_VICT
1752	NaN	NaN	NaN	NaN	NaN	NaN	0
763	NaN	NaN	NaN	NaN	NaN	NaN	0
2064	NaN	NaN	NaN	NaN	NaN	NaN	2
1242	NaN	NaN	NaN	NaN	NaN	NaN	0
2121	NaN	NaN	NaN	NaN	NaN	NaN	0
3096	NaN	NaN	NaN	NaN	NaN	NaN	0
1942	NaN	NaN	NaN	NaN	NaN	NaN	0
2201	NaN	NaN	NaN	NaN	NaN	NaN	0
1217	NaN	NaN	NaN	NaN	NaN	NaN	0
3009	NaN	NaN	NaN	NaN	NaN	NaN	0

Figure 3



# Exploratory Data Analysis

We performed a series of visualizations for our datasets, from graph charts to maps in order to find some insights into our data. With the amount of data that we had at the beginning of our project, we were hoping to find correlations within our data. We discovered that some of the correlations that we were hoping would be apparent did not exist. Because of this finding we decided to reevaluate the data that we had as well as go back to the beginning and narrow down our scope. The following shows some of the insights that we discovered using the Asthma Prevalence in Adults ,GHG, and the AQI.

## Average Percentage of Asthma Prevalence in Adults by State

```
[148]: <function matplotlib.pyplot.show(close=None, block=None)>
```

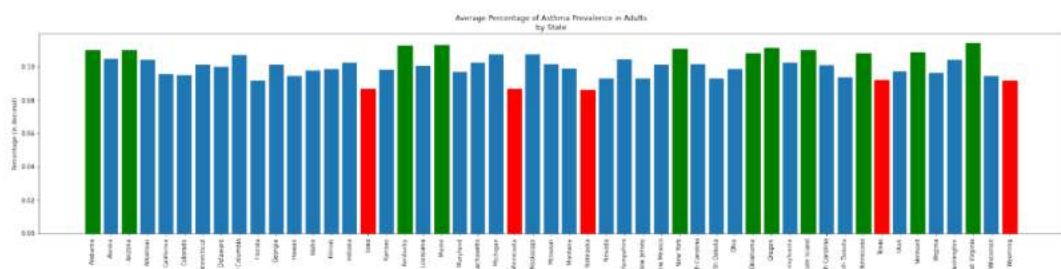


Figure 5

## Greenhouse Gas Ranking Map

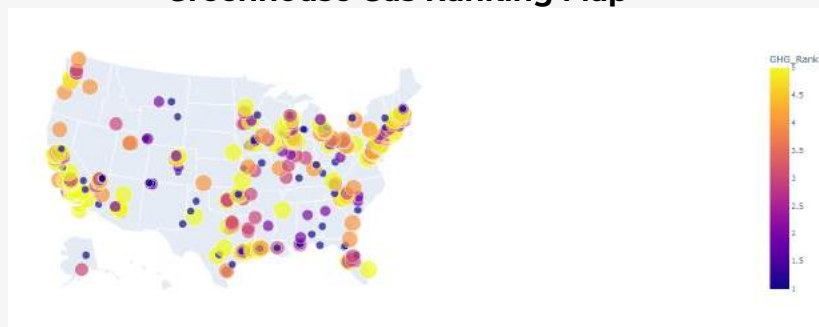


Figure 6

## Density plots for AQI variables

<---90th Percentile AQI

Days Ozone-->

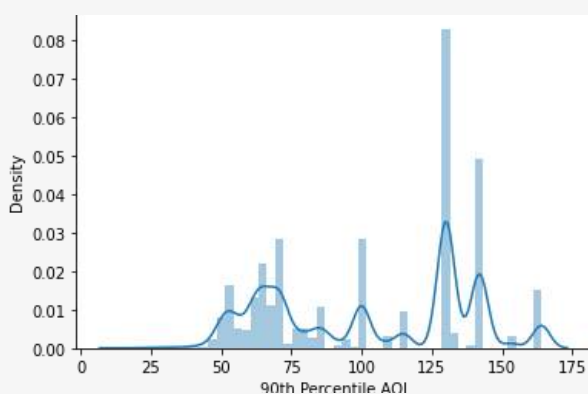


Figure 7

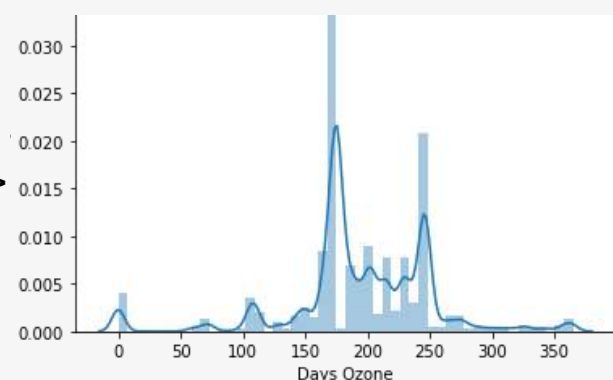


Figure 8

# Dashboard

Link to Dashboard:

<https://datastudio.google.com/s/ks1Q5UDVn-A>

## Summary Dashboard

Filter: State, City, Year

REPORTING YEAR

STATE

COUNTY NAME

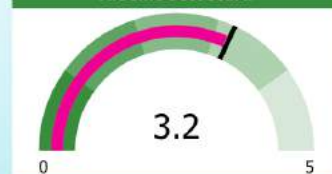
### GHG Scorecard



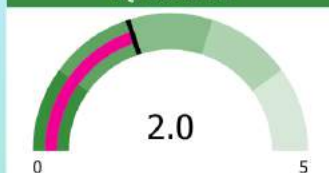
### Overall Environmental Scorecard



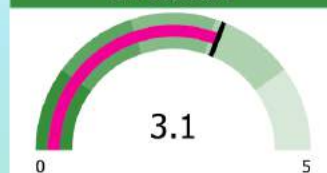
### Arsenic Scorecard



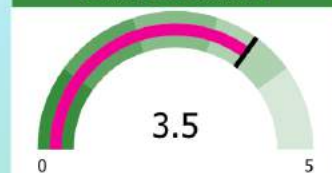
### AQI Scorecard



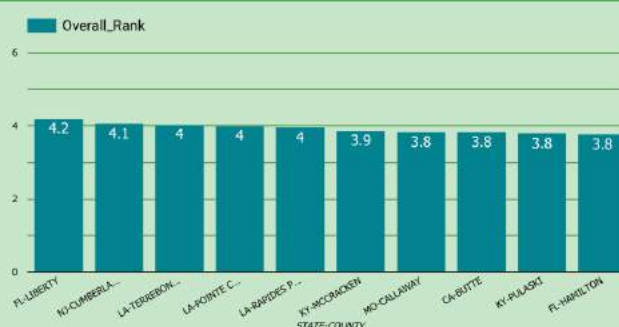
### SVI Scorecard



### Power Plant Scorecard



### Top 10 Highest Counties





# Dashboard Use Case



\_\_\_\_\_

The end users for our dashboards will be individuals who want to understand how environmentally safe the communities that they live in or may want to live in are. We also understand that this environmental data may be: a) hard to find and b) in various places. We hope that our dashboards can aggregate and summarize information in one place. By using our environmental scorecard, we believe that users can make informed decisions on where they buy property and choose to make a home. Additionally we would like the scorecard to provide transparency around environmental indicators that will eventually impact pricing and housing policies.

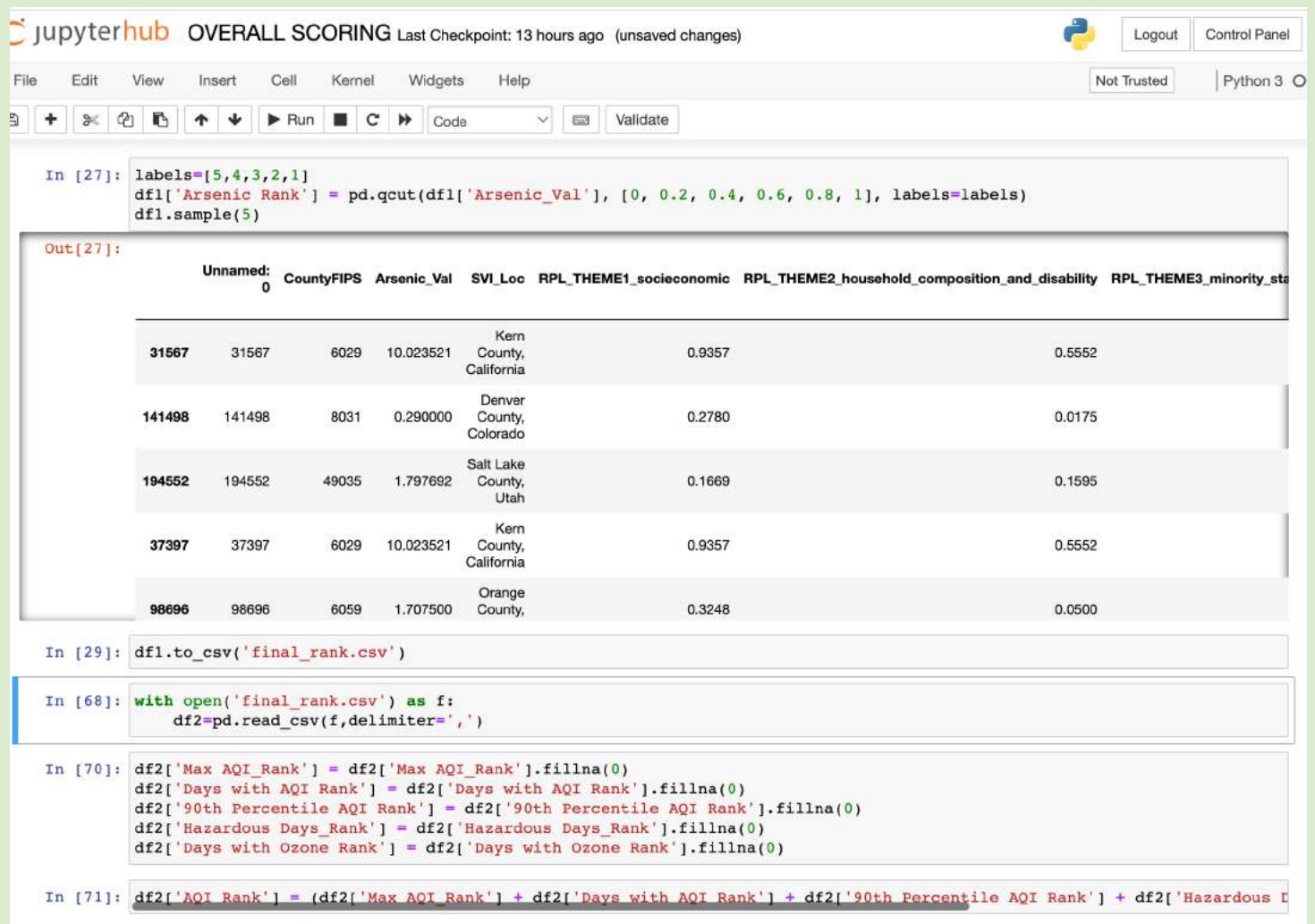


# Dashboard

## Data Engineering

We were inspired by the engineering that created the Walk Score® and found a source that involved someone using reverse engineering to predict a walk score to determine which features had the greatest impact on the walk score. This was a good example for us to model our own environmental score after. As a result, we set out to create a predictive model for cancer prevalence in each location using the features that we pre-selected. We wanted to train the model to learn how to predict cancer prevalence and measure the impact of each of the feature on cancer. To begin such a task, we needed to observe a linear relationship between the factors and the adverse health event which was cancer in this case. We were not able to determine a linear relationship and did not have time to explore non-linear relationships that may have done a better job at explaining the relationship between the features. However, given the opportunity we would have explored kmeans and other regression types and selected a machine learning model based on the strength of the relationships determined there

As we mentioned in our Data Wrangling and cleaning section, we were able to create 5 different cleaned data sets for our environmental factors. From there we used pandas qcut and cut functions, to rank each of the values from 1 - 5. One issue we faced during this process was that the qcut function only works if there are enough distinct values to categorize into your given quantiles. Since we used a ranking metric of 1-5, some of our datasets, like the Power plant dataset, did not have enough unique values for qcut to aggregate into quintiles. We had to use a more manual work around where we assigned the numerical ranges for each of the five bins using the cut function. After assigning the individual rank for each metric, we used a simple average to create the overall rank for the purpose of the first iteration of the scorecard. Looking forward it would be interesting to see how we could assign different weights to each of the factor based on how likely it is to effect health outcomes



The screenshot displays a Jupyter Notebook titled "OVERALL SCORING" with a last checkpoint 13 hours ago. The interface includes a menu bar (File, Edit, View, Insert, Cell, Kernel, Widgets, Help) and a toolbar with icons for file operations, running, and code execution. The notebook is running on Python 3. The code in the notebook performs the following steps:

- In [27]:** Defines labels [5, 4, 3, 2, 1] and uses `pd.qcut` to rank 'Arsenic\_Val' into five bins. It then samples 5 rows from the resulting DataFrame.
- Out[27]:** Displays a preview of the sampled data as a table.
- In [29]:** Saves the DataFrame to a CSV file named 'final\_rank.csv'.
- In [68]:** Reads the 'final\_rank.csv' file back into a DataFrame using `pd.read_csv`.
- In [70]:** Fills missing values (NaN) in several columns: 'Max AQI Rank', 'Days with AQI Rank', '90th Percentile AQI Rank', 'Hazardous Days Rank', and 'Days with Ozone Rank'.
- In [71]:** Calculates the 'AQI Rank' by averaging the ranks of the five AQI-related metrics.

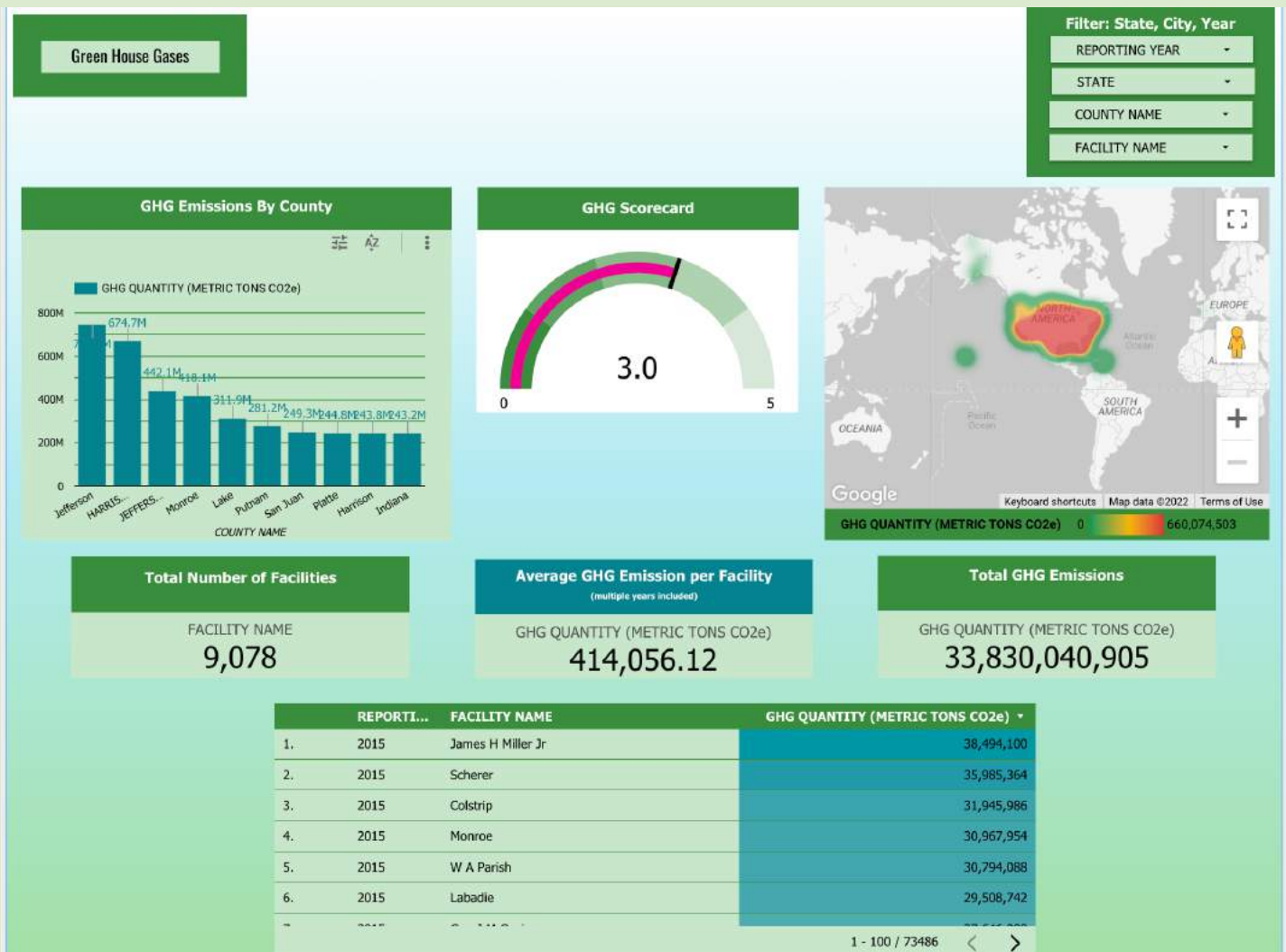
Unnamed: 0	CountyFIPS	Arsenic_Val	SVI_Loc	RPL_THEME1_socioeconomic	RPL_THEME2_household_composition_and_disability	RPL_THEME3_minority_sta
31567	31567	6029	10.023521	Kern County, California	0.9357	0.5552
141498	141498	8031	0.290000	Denver County, Colorado	0.2780	0.0175
194552	194552	49035	1.797692	Salt Lake County, Utah	0.1669	0.1595
37397	37397	6029	10.023521	Kern County, California	0.9357	0.5552
98696	98696	6059	1.707500	Orange County,	0.3248	0.0500

# Dashboard

## Data Engineering

To visualize our scorecard, we utilized Google Data Studio. We wanted the dashboard to be dynamic and show not only the scorecard but the underlying detail to the score. Further we also wanted our dashboard to be easy to navigate as part of the goal is this project is. to make information on the environmental factors accessible to all.

The general outlay of each dashboard is the scorecard front and center, a chart of the top 10 counties with the highest environmental factor values, a plotting of the values on a geographic map, and various summary statistic highlighted below. Users are then able to filter by State, County, and Year in the drop-down list in the top right.



# Conclusions

"YOU CANNOT GET THROUGH A SINGLE DAY WITHOUT HAVING AN IMPACT ON THE WORLD AROUND YOU." - JANE GOODALL

---

While we could not find a correlation between our environmental factors and our health factor using the statistic models we learned in class. We believe that with more advanced statistical modeling we could start to identify key trends. Therefore, we are currently still working to see if we can progress our model and create a predictive scorecard that could correlate a specific environmental factor to the likelihood of cancer prevalence. However, we are able to see here a clear picture of where there are more negative environmental factors and of the potential side effects of these factors on the health of residents around those areas.

Researching different environmental factors that can cause health problems such as respiratory diseases, heart disease, and some types of cancer is important. This topic is also more important when you consider how people with low incomes are more likely to live in polluted areas and have unsafe drinking water. And people of color are more likely to experience housing discrimination that exposes them to more negative environmental factors such as living closer to power plants. We hope that our work contributes to transparency and aids in the creation of empowered individuals who understand how their housing impacts their health and quality of life.





# Future Work

## RECOMMENDATIONS TO CONTINUE BRINGING HEALTH AND THE ENVIRONMENT TO THE FOREFRONT

The next steps for our model would be to include predictive elements such as using machine learning to predict the likelihood of health adverse outcomes in specific counties based on multiple environmental factors. Moreover, we would love to scale up the model to include other environmental factors to increase the accuracy of scorecards for different states and counties.

Additionally, because environmental factors are more specific and unique beyond county areas, we would like to look into integrating zip code level data. With this data, we can map exact areas where negative environmental factors are present and where there are higher rates of illness. Thus, the accuracy would be increased with our model.

Finally, with this level of data, we can begin to map how lower-income and people of color are disproportionately affected by environmental factors. This can further housing justice by demonstrating how redlining and other forms of housing discrimination can be playing a factor in which healthy or unhealthy areas lower-income and people of color are pushed to live in.



# Contact Us

Name	Location	Email	Linkedin
Tinika McIntosh-Amouzouvi	Richmond, VA	tinika.mcintosh@gmail.com	<a href="https://www.linkedin.com/in/tinika-mcintosh-amouzouvi">www.linkedin.com/in/tinika-mcintosh-amouzouvi</a>
Taylor Brown	Oakland, CA	taybrown545@gmail.com	<a href="https://www.linkedin.com/in/taylorebrown/">https://www.linkedin.com/in/taylorebrown/</a>
Luis Esparza	Los Angeles, CA	luisfesparza@gmail.com	<a href="https://www.linkedin.com/in/luisfesparza01/">https://www.linkedin.com/in/luisfesparza01/</a>
Ashley Aviles	Minneapolis, MN	ashleyaviles1@gmail.com	<a href="https://www.linkedin.com/in/ashley-aviles-brizuela/">https://www.linkedin.com/in/ashley-aviles-brizuela/</a>
Chioma Dunkley	Philadelphia, PA	cdunkley@terpmail.umd.edu	<a href="https://www.linkedin.com/in/chiomadunkley/">https://www.linkedin.com/in/chiomadunkley/</a>

# Acknowledgements

- DS4A Empowerment Cohort 3
- TAs- Scott and Revell
- Mentors-Khalil and Maggie

Thank you all so much, this project  
would not have been possible  
without your support

-----Team 61!