An Exploration of Climate Equity in California

By: Tara Verma and Godsee Joy
https://github.com/UC-Berkeley-I-School/Project2_Verma_Joy.git

Questions

Our primary questions driving this exploration are as follows:

1. How do measures of income and wealth (poverty and median home value) correlate with various demographic factors (race, county) and health outcomes (life expectancy)?
2. How do different green space metrics (IPUMS land coverage, share of tract land, and impervious or crop land) relate to one another? What can they together tell us about green space access?
3. How do green spaces relate to risk for different climate challenges (extreme heat and fire risk)?
    a. Which geographic areas are disproportionately affected?
    b. How do these areas of impact coincide with socioeconomic and demographic factors?

Data

Data Sources

Our primary dataset comes from the Climate and Economic Justice Screening Tool[1]. This is a government website with data at the census tract level across the United States with indicators of burdens across climate change, energy, health, housing, legacy pollution, transportation, water and wastewater, and workforce development. The purpose of the tool is to help federal agencies identify disadvantaged communities to prioritize funding related to climate and clean energy. There are about 74K rows in the national dataset, covering nearly all census tracts, with over 30 columns. We will be focusing on California (around 8K rows).

For supplemental data, we also merged data from the IPUMS National Historical Geographic Information System (NHGIS)[2]. Specifically, their 2015 census tract-level data (about 72K rows) on different types of land coverage and 2014 county-level data on mean temperatures (about 3,100 rows). While an account is needed to access IPUMS data, it is free to use. To build geospatial views, we also used 2019 shapefiles from the U.S. Census[3] at the census tract and county levels and merged with our main datasets on the Census GEOIDs.

[1] Climate and Economic Justice Screening Tool
[2] IPUMS NHGIS
[3] Census Shape Files 2019

Sanity Checks

Merging Data & Missing Data

To join the three datasets, we used the Census FIPS codes[4] as our primary joining key. Our main dataset is at the census tract level, as was the NHGIS data on types of land coverage. We had to modify the GISJOIN variable in the original dataset to remove placeholder zeros breaking up the state, county, and census tract codes to match the NHGIS data census tract codes. The NHGIS average temperature data is at the county level so we had to take a subset of the census tract geoid (state + county) to join the 2014 temperature data.

Given our focus on access to green space and intersections with socioeconomic disparities, we scanned through the available variables and data documentation to prioritize non-percentile data points and a single metric when several were available (e.g., adjusted 200% FPL instead of binary is low income), and dropped columns related to other areas of climate equity (e.g., proximity to waste management facilities). When limiting to California, the size of our data dropped to 8,057 rows. The county level NHGIS temperature data had mean, min, max, and standard deviation of temperatures annually for each county going back to 1895, so we limited this dataset to just 2014. Since the main dataset is at the census tract level, tracts in the same county will have the same temperature data.

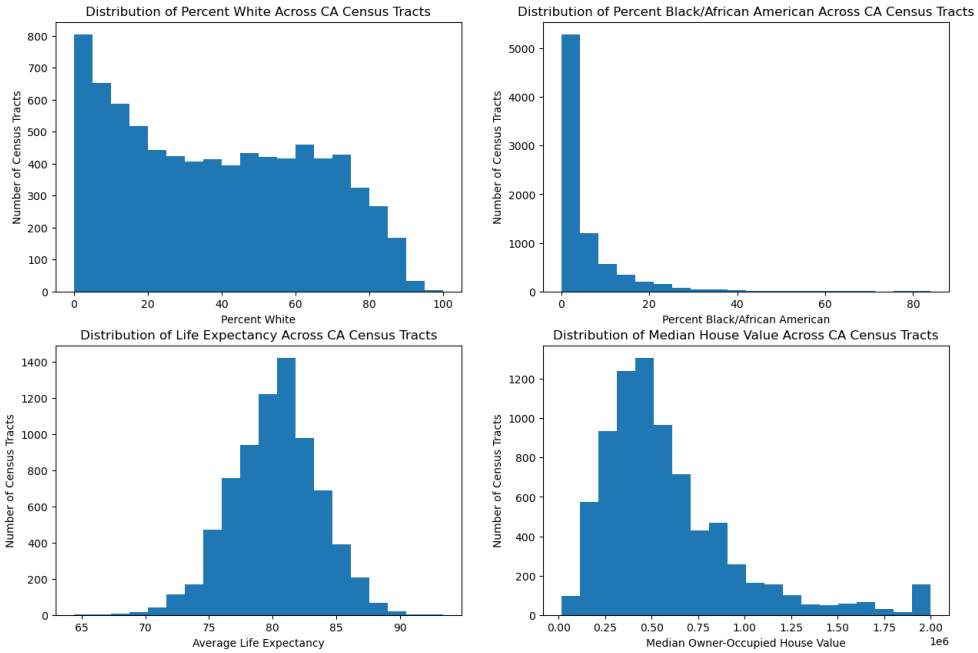
We first checked to see that all census tracts and counties existed in the data. From cross-referencing 2010 Census reports[5], we confirmed our main dataset has 8,057 rows (one per census tract) for California. We saw that 99% (8,038) of California's census tracts were in the NHGIS land coverage data as well. For the temperature data, we confirmed that all 58 California counties[6] were represented.

For missing data values, only 14 columns in our main dataset had missingness rates of 5% or more. However, the NHGIS land coverage data had very high rates of missing data across most columns except the four developed areas (open space, low intensity, medium intensity, high intensity). The NHGIS temperature data is available for all counties in California. See Appendix Table 1 for more detailed information on missingness of data in the main dataset and the additional NHGIS data as well as the final columns we used for this exploratory analysis.

Distribution and Data Type Checks

As part of the first question, we aim to examine relationships between several of the demographic variables in our primary dataset. We created a series of histograms to build a high-level understanding of the underlying distribution of key demographic variables across California census tracts, and to gut check the values to ensure nothing seemed unreasonable or unexpected. The following figures show distributions of the variables we will be using as representations and proxies for race, health, and wealth information.

[4] Census GeoID reference; state is 2 digits + county: 3 digits + census tract: 6 digits
[5] https://www2.census.gov/geo/pdfs/reference/guidestloc/ca_gslcg.pdf
[6] https://www.counties.org/californias-counties



We examined distributions for all of our variables and checked that they were the correct data types, and that the distributions were reasonable. Looking at a subset of these provides some initial insights to explore further. We also checked the same key data points for the national dataset (Appendix Figure 1) for a comparison point.

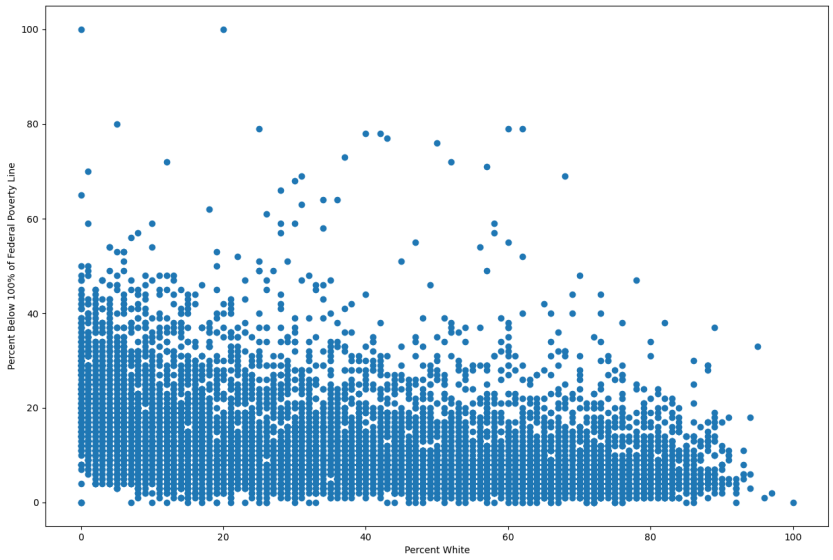
Our Findings

Demographic and Socioeconomic Relationships

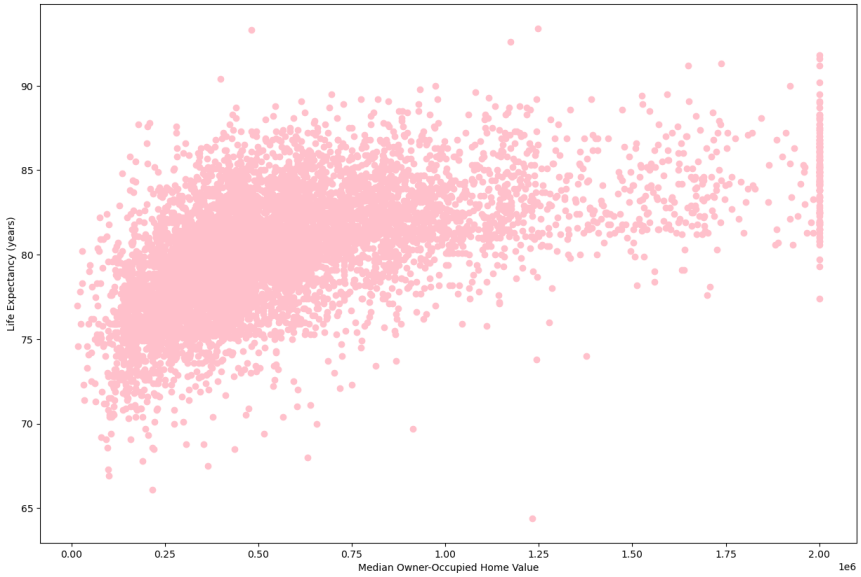
Our primary question aims to understand climate inequity in the state of California by examining how green space access varies across census tracts with different socioeconomic makeups. In order to fully understand this, it's important to first develop an understanding of the interrelationships between different factors (focusing primarily on life expectancy, race, and income/wealth) and build a high-level picture of California's demographic and socioeconomic landscape.

We started by examining a correlation matrix of relevant demographic, socioeconomic, and health variables (Appendix Figure 2), and noticed a few prominent lines of negative correlation along percent white, median house value, and life expectancy, and a few lines of positive correlation across percent of individuals living below 100% and 200% of the federal poverty line. This is an initial indication that many health and energy burdens seem less prominent in affluent, white census tracts.

To dig further into this trend, we examined a few of these relationships more directly. The first is the relationship between the percentage of the census tract that's white and the percentage living below 100% of the federal poverty line (correlation coefficient: -0.41). Although not particularly dramatic, we do see a decrease in poverty in predominantly white census tracts.



Another piece that is prudent to examine is the relationship between health (represented by life expectancy) and wealth (represented by median owner-occupied house value). Given that in the US, many of the factors that allow for long, healthy lives (access to nutritious food, clean water, and healthcare, free time to exercise, safe and stable housing, etc.) can be prohibitively expensive, it would make sense to see a general increase in life expectancy as median house value increases, which is evident in the figure below.



Another key data point in our primary dataset records whether or not that census tract is identified as disadvantaged. This is determined as part of the Climate and Economic Justice Screening Tool process, using set threshold criteria (Appendix Table 2) for various socioeconomic, climate, and health burdens. Splitting the dataset into two groups based on whether or not they are identified as disadvantaged, and then taking the mean values of our demographic variables also yields interesting results. Since these are mean percentages, the way to interpret the information below is, as an example, the average proportion of Black residents across census tracts that are classified as disadvantaged is 7%.

	Mean Values	
Variable	Not Disadvantaged	Disadvantaged
percent_black_or_african_american_alone	4%	7%
percent_american_indian_alaska_native	0%	1%
percent_asian	15%	12%
percent_native_hawaiian_or_pacific	0%	0%
percent_two_or_more_races	5%	3%
percent_white	50%	20%
percent_hispanic_or_latino	25%	57%
percent_other_races	8%	22%
percent_age_under_10	10%	13%
percent_age_10_to_64	72%	73%
percent_age_over_64	16%	12%
adjusted_percent_of_individuals_below_200pct_federal_poverty_line	11%	39%
median_value_usd_of_owner-occupied_housing_units	$707,380	$388,763
life_expectancy_years	81	79
unemployment_percent	4%	8%
percent_of_individuals_below_100pct_federal_poverty_line	8%	21%

As we might expect, the average proportion of white residents is much higher in census tracts that did not identify as disadvantaged, and conversely, the proportion of Hispanic/Latino residents is much higher in tracts that do identify as disadvantaged. The other rows of note in this table pertain to wealth; we see a large difference between groups in the percentage of residents below 200% of the poverty line and in median home value. On the flipside, it seems as though age remains relatively constant, and thus, does not seem to be a variable of interest when looking for disparity.

The figures above provide an overview of the socioeconomic stratification evident in California, but our core question pertains to climate equity (green space and high temperatures). Given that the most stark differences we see seem to stem from percent Hispanic, percent white, percent of individuals living in poverty, and median home value, we will aggregate these metrics by county using the mean value, and pay particular attention to the counties with extreme values.

Counties with the highest percentages of Hispanic/Latinx residents, and of white residents

county_name	percent_hispanic_or_latino
Imperial County	77%
Tulare County	63%
Merced County	59%
San Benito County	57%
Kings County	55%

county_name	percent_white
Sierra County	87%
Nevada County	86%
Plumas County	85%
Calaveras County	84%
Trinity County	82%

Counties with the highest and lowest median house values

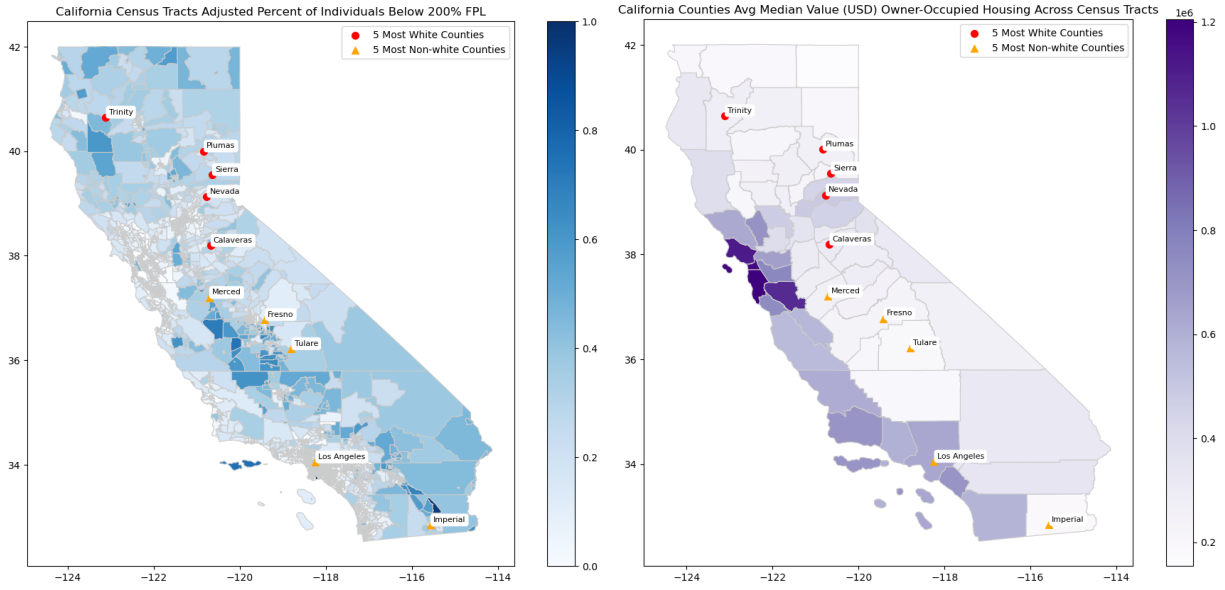
county_name	median_value_usd_of_owner-occupied_housing_units
San Mateo County	$ 1,204,862.29
San Francisco County	$ 1,179,213.10
Marin County	$ 1,111,072.31
Santa Clara County	$ 1,060,131.98
Alameda County	$ 762,488.46

county_name	median_value_usd_of_owner-occupied_housing_units
Modoc County	$ 154,250.00
Imperial County	$ 180,876.67
Siskiyou County	$ 187,685.71
Kern County	$ 194,067.59
Tulare County	$ 196,888.31

Counties with the highest percentage of residents living below 200% of the federal poverty line

county_name	adjusted_percent_of_individuals_below_200pct_federal_poverty_line
Del Norte County	44%
Tulare County	43%
Modoc County	42%
Imperial County	40%
Trinity County	40%

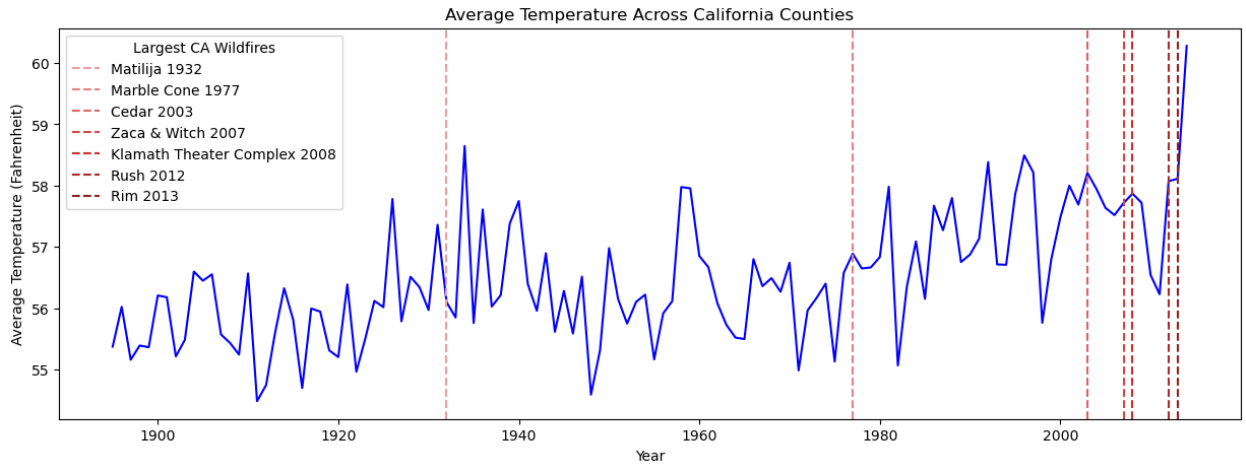
For geospatial visuals, we used the GeoPandas and Matplotlib Python libraries. Three of the top five most counties are also among the top five most Hispanic/Latinx counties. When we map the top five most white counties and the top five most non-white counties, and look at FPL status and median housing unit value, we see that the census tract view (left-side) shows higher poverty levels closer to the top non-white counties. However, notice the level of detail for large cities like Los Angeles and San Francisco are such that without zooming in, we are unable to see the breakdown for census tracts in those areas. Similar observations regarding granularity were seen in the median home value visual (right-side), so this one we aggregated up to the county level. Here we see there is a strong skew towards the coastal area counties.



Adding Climate to the Mix

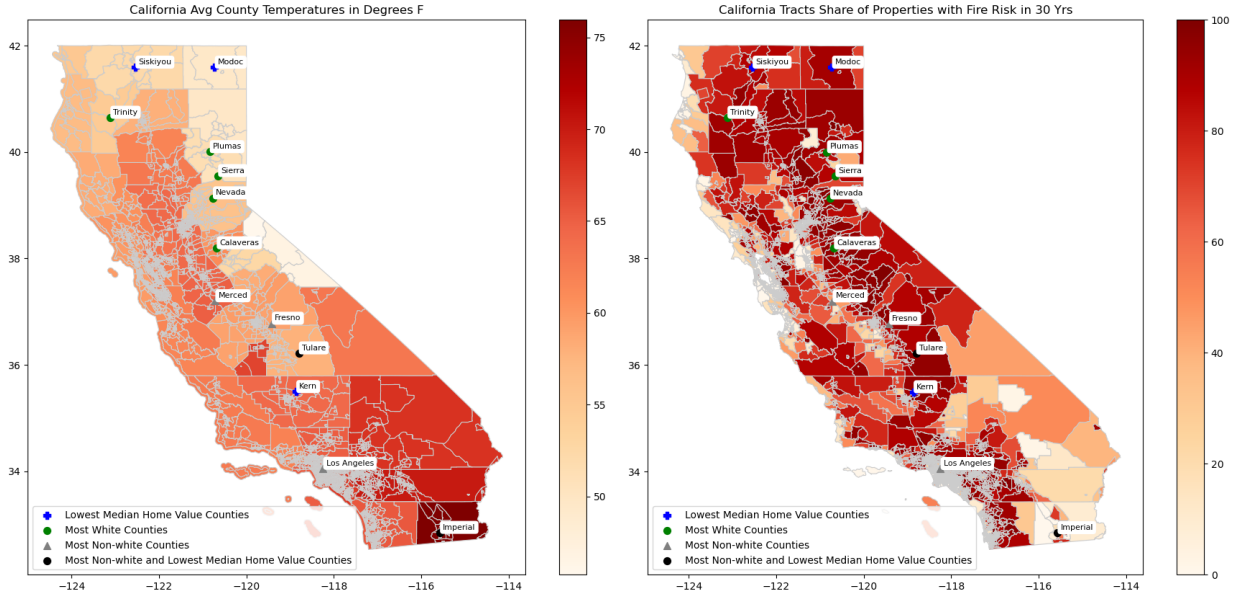
Temperature and Green Space

The NHGIS temperature dataset has average temperatures at the county level going back to 1895. Given this was our only temporal dataset, we decided to explore time series trends before exploring the larger merged dataset that used 2014 only data. When looking at average temperatures across counties in California and highlighting the years where one of the top 20 largest wildfires in California occurred, we see a noticeable two-fold trend: first, that temperatures have been on an upward trend and second, that most of the largest California wildfires occurred after 2000. In fact, *12 out of the 20 largest fires are not even displayed on the graph* below because they happened after 2014.[7]



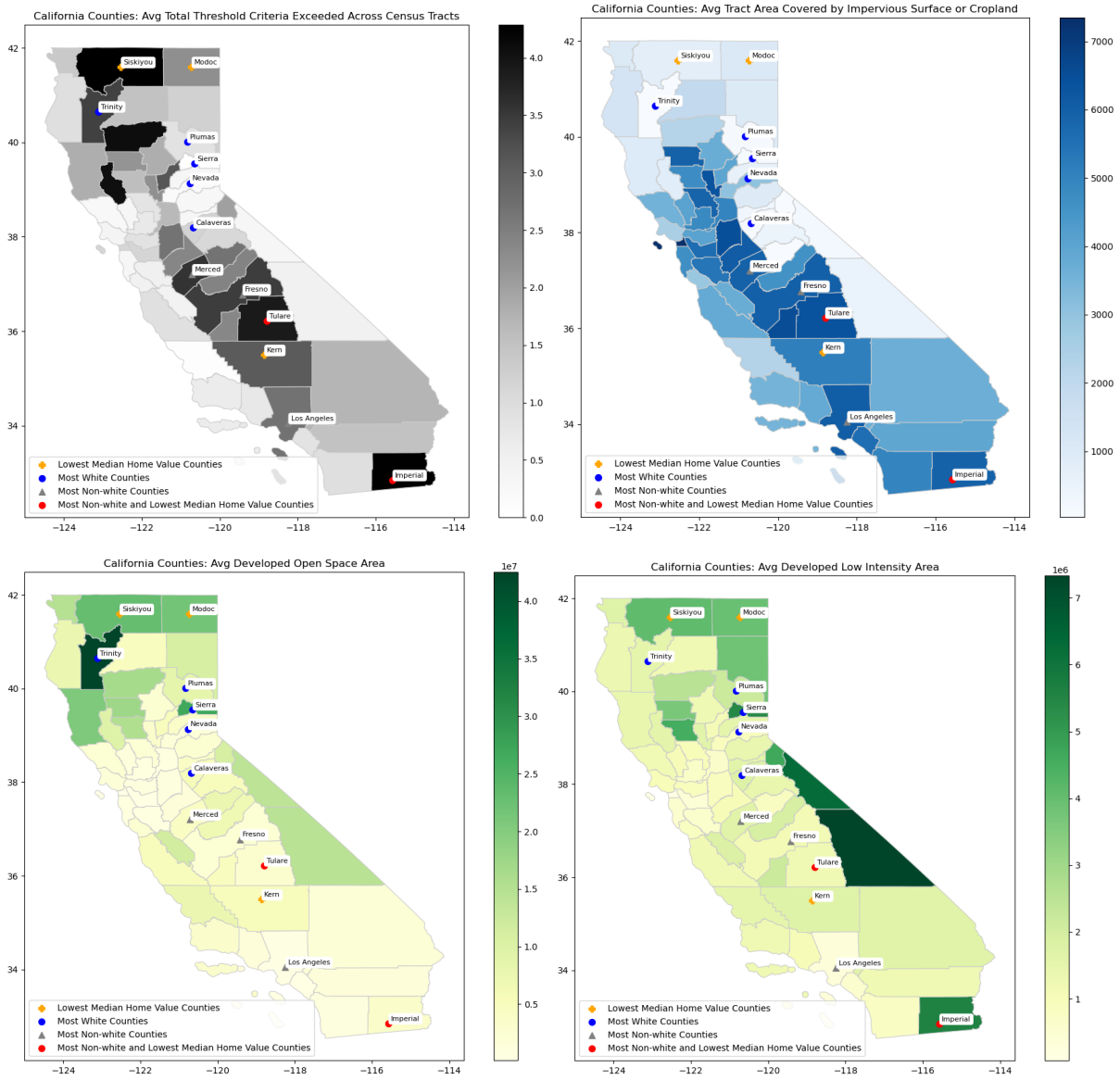
[7] https://www.fire.ca.gov/our-impact/statistics; note, this ranking is based on fires from 1932 onwards, due to poor data quality prior

Interestingly, when we look at 2014 average temperatures (originally in Celsius, which we converted to Fahrenheit for easier interpretation) across counties and compare that geospatial visual with the share of census tracts with fire risk for their properties in 30 years, it is not an immediate 1-to-1 correlation. This is likely because of other climate factors affecting fire risk. While southern California is hotter, for example, San Diego is on the coast and, further inland, the climate is arid and desert-like suggesting lower fire risk. We do observe, however, that while fire risk affects most of California, Imperial County is particularly affected by higher temperatures and Siskiyou, Tulare, Modoc Counties by fire risk while also being in the top five counties with the lowest median home value in California.



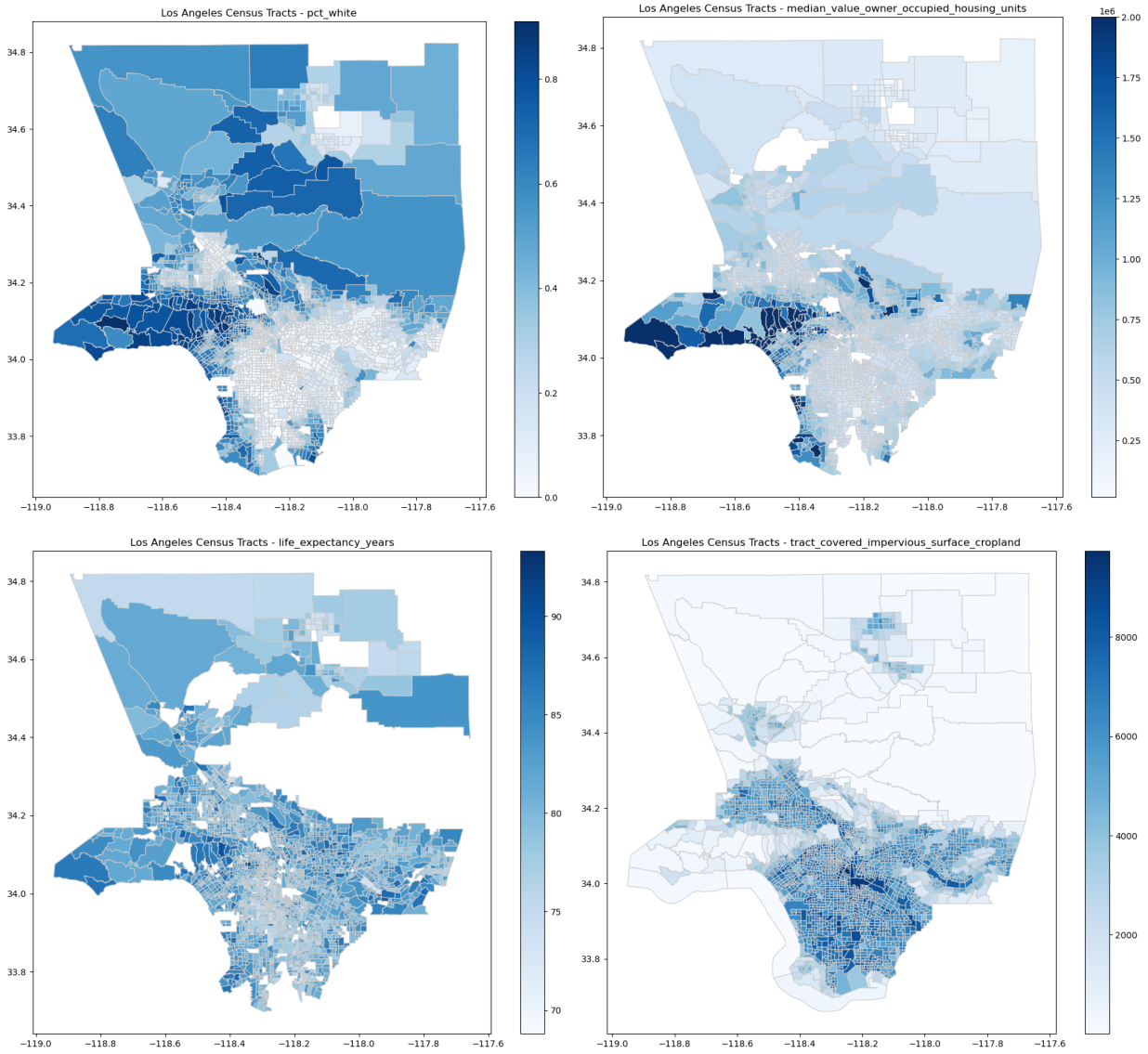
When we looked at the correlations between the variables related to green space (Appendix Fig. 3), surprisingly, the developed high intensity variable from NHGIS was not that correlated with our main dataset's tract area covered by impervious surfaces or cropland variable (.16 correlation coefficient) but the proportion version was (.64 correlation coefficient). We decided to use the variable in our main dataset (impervious surfaces and cropland), developed open space since that had the highest correlation in the negative direction with our main dataset variable (-.45 correlation coefficient) and the developed low intensity to augment the open space variable.

In the next geospatial views below, we see a clear inverse relationship between the counties with larger areas of impervious surfaces aka less green space (top right) and those with more (bottom left and right). In the top left, we see that the majority of top five most white counties exceeded fewer barrier criteria and live in areas with more green space. However, we also see a trend here where more rural areas (Siskiyou, Modoc, Trinity counties) might have more access to green spaces, but exceed other climate equity barrier thresholds.



Let's explore how things might change when we zoom into a city - Los Angeles - so we can examine census tract level of detail more clearly[8]. Here we see relationships that were masked by the state-wide visual. The census tracts with the highest proportion of white residents (top left) also have higher median home values (top right) and less strongly but still a slight correlation with higher life expectancy (bottom left) and a clear trend with more green space (bottom right). See Appendix Figure 4 for additional plots focused on Los Angeles.

[8] We removed two census tracts that represented a U.S. Naval base.



Conclusion

As part of this project, we learned a lot in two main areas: insights about our dataset and the people it represents, and properties of our dataset that made it easy in some ways, and difficult in others, to work with. Starting with insights about our dataset, our driving questions sought to examine the relationships between various demographic factors (including health, race, and wealth) and the prevalence of green spaces to determine whether access was disproportionately available to certain groups. Through our exploration, we saw that one of the major themes in the social discourse zeitgeist–particularly that affluence and health are more available to white people–seemed to be well-represented in our dataset. We also found that green spaces seem to fall into that same category of disproportionate access. These same insights and ideas are reflected in the raw data–with total criteria exceeded and total categories exceeded providing an aggregate measure of disadvantage across different domains. However, these takeaways come with some caveats, rooted in the properties of our datasets.

It is important to remember that communities face disadvantages across many domains in different ways - an interactive visual dashboard would be more helpful for policymakers and community advocates looking to understand the greatest barriers facing their communities. While we did find trends in the kinds of areas that get labeled disadvantaged, there is far more nuance than what was captured in this data and our exploration. We also found that the granular look at the census tract level reveals disparities that are masked by county-level aggregations–which means that the insights we gained about the counties in our datasets cannot be assumed to represent every census tract or resident in that county. A dynamic visualization tool with the ability to add multiple filters at once would be key to gaining a deeper understanding of the variance. Finally, we only looked at a small subset of the data available, and at one state: California. There is ample opportunity for further research and exploration into other states and variables within our datasets.

Appendix

Regarding missingness, a majority of the columns have most data available from our main dataset (missing less than 2% for the nationwide and CA specific data). There were some exceptions related to tribal areas, Formerly Used Defense Sites, and wastewater for example, but we did not use these columns in our analysis. For the IPUMS land coverage dataset unfortunately, most of the columns had very high levels of missingness for California. We thus limited usage of columns from this dataset to just four. The final list of columns and their missingness rates for California:

Table 1: Data Missingness

Final Column	Source	Granularity	% Missing in California
census_tract_2010_id	Climate and Economic Justice Screening Tool (main dataset)	Census tract	0
county_name	Climate and Economic Justice Screening Tool (main dataset)	Census tract	0
state_territory	Climate and Economic Justice Screening Tool (main dataset)	Census tract	0
percent_black_or_african_american_alone	Climate and Economic Justice Screening Tool (main dataset)	Census tract	0.57
percent_american_indian_alaska_native	Climate and Economic Justice Screening Tool (main dataset)	Census tract	0.57
percent_asian	Climate and Economic Justice Screening Tool (main dataset)	Census tract	0.57
percent_native_hawaiian_or_pacific	Climate and Economic Justice Screening Tool (main dataset)	Census tract	0.57
percent_two_or_more_races	Climate and Economic Justice Screening Tool (main dataset)	Census tract	0.57
percent_white	Climate and Economic Justice Screening Tool (main dataset)	Census tract	0.57
percent_hispanic_or_latino	Climate and Economic Justice Screening Tool (main dataset)	Census tract	0.57
percent_other_races	Climate and Economic Justice Screening Tool (main dataset)	Census tract	0.57
percent_age_under_10	Climate and Economic Justice Screening Tool (main dataset)	Census tract	0.57
percent_age_10_to_64	Climate and Economic Justice Screening Tool (main dataset)	Census tract	0.57

percent_age_over_64	Climate and Economic Justice Screening Tool (main dataset)	Census tract	0.57
total_threshold_criteria_exceeded	Climate and Economic Justice Screening Tool (main dataset)	Census tract	0
total_categories_exceeded	Climate and Economic Justice Screening Tool (main dataset)	Census tract	0
identified_as_disadvantaged	Climate and Economic Justice Screening Tool (main dataset)	Census tract	0
adjusted_percent_of_individuals_below_200pct_federal_poverty_line	Climate and Economic Justice Screening Tool (main dataset)	Census tract	0.57
is_low_income	Climate and Economic Justice Screening Tool (main dataset)	Census tract	0
share_of_properties_at_risk_of_fire_in_30_years	Climate and Economic Justice Screening Tool (main dataset)	Census tract	0.29
energy_burden	Climate and Economic Justice Screening Tool (main dataset)	Census tract	0.96
housing_burden_percent	Climate and Economic Justice Screening Tool (main dataset)	Census tract	1.07
median_value_usd_of_owner-occupied_housing_units	Climate and Economic Justice Screening Tool (main dataset)	Census tract	2.87
share_of_the_tract's_land_area_that_is_covered_by_impervious_surface_or_cropland_as_a_percent	Climate and Economic Justice Screening Tool (main dataset)	Census tract	0
life_expectancy_years	Climate and Economic Justice Screening Tool (main dataset)	Census tract	6.71
unemployment_percent	Climate and Economic Justice Screening Tool (main dataset)	Census tract	0.83
percent_of_individuals_below_200pct_federal_poverty_line	Climate and Economic Justice Screening Tool (main dataset)	Census tract	0.91
percent_of_individuals_below_100pct_federal_poverty_line	Climate and Economic Justice Screening Tool (main dataset)	Census tract	0.91
merge_id	IPUMS NHGIS land coverage	Census tract	0
AREA21	IPUMS NHGIS land coverage	Census tract	6.98
AREA22	IPUMS NHGIS land coverage	Census tract	0.97
AREA23	IPUMS NHGIS land coverage	Census tract	0.27

AREA24	IPUMS NHGIS land coverage	Census tract	1.19
merge_id_county	IPUMS NHGIS temperature	County	0
MEAN	IPUMS NHGIS temperature	County	0
TIME	IPUMS NHGIS temperature	County	0

Figure 1: National Distributions of Race, Health, and Wealth Variables

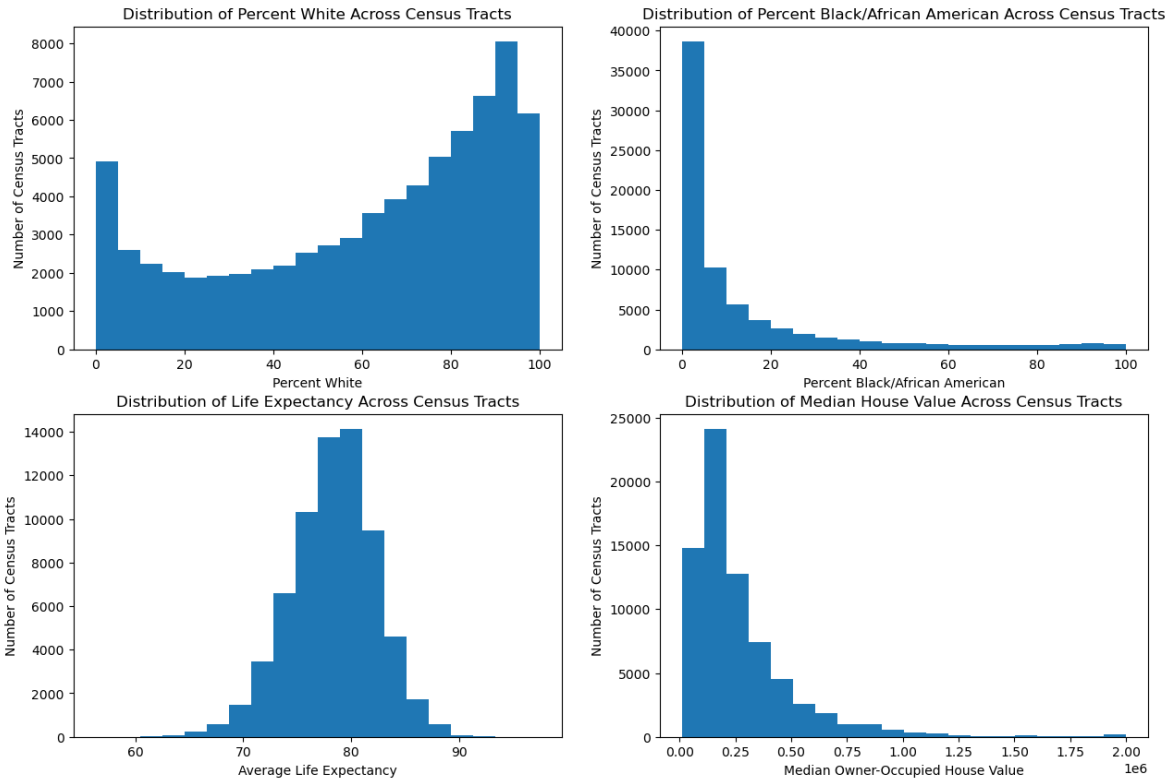


Figure 2: Correlation Heatmap of Demographic Variables

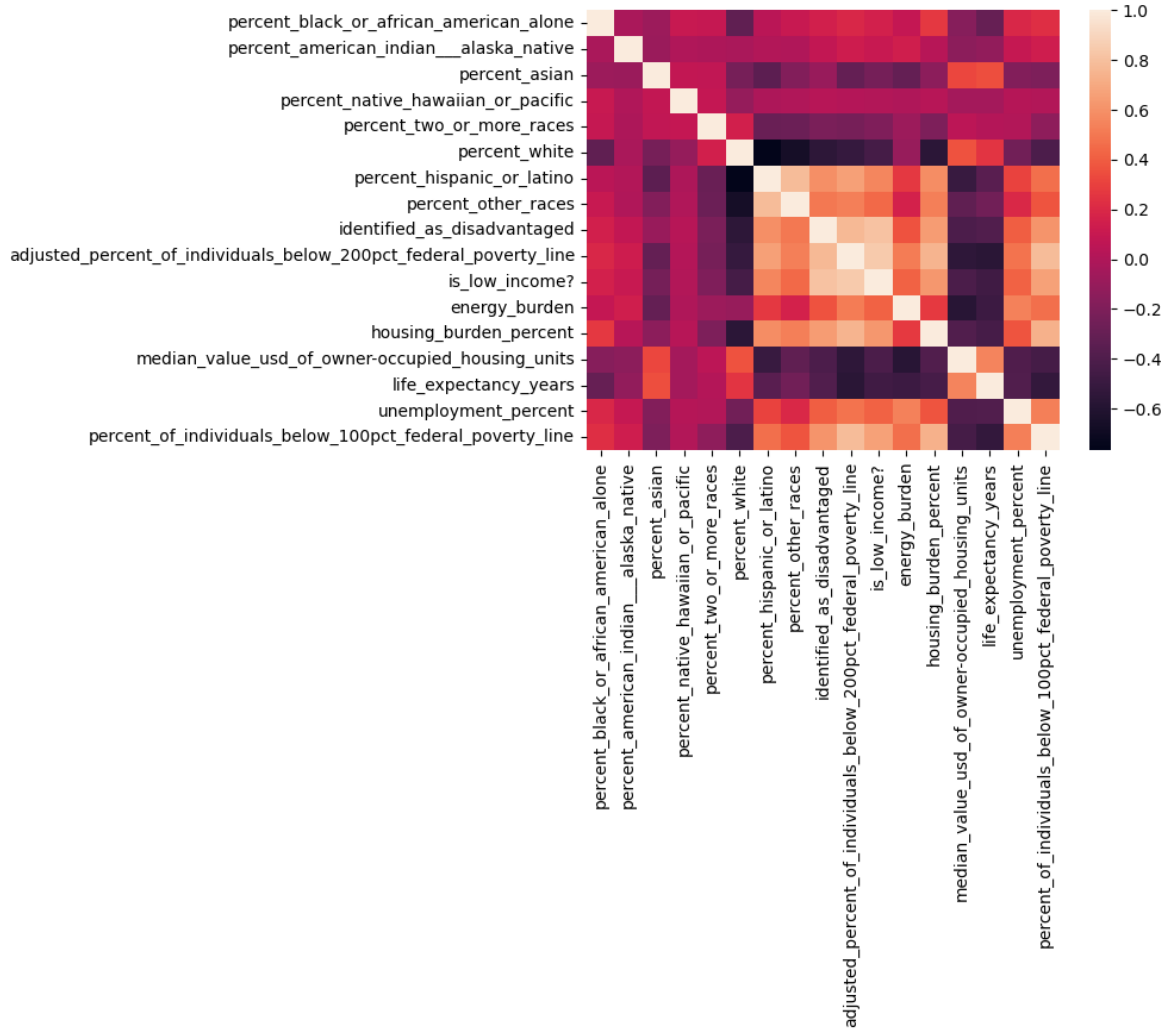


Table 2:. Overview of methodology used in version 1.0 of the CEJST (from technical documentation)

(Items marked as NEW are changes made from the beta version to version 1.0) Communities are considered disadvantaged:

- if they are located in a census tract that meets the thresholds for at least one of the tool's categories of burden, or;
- if they are on land within the boundaries of Federally Recognized Tribes **(NEW)**
- Census tracts that are completely surrounded by disadvantaged communities are also considered disadvantaged if they meet an adjusted low income threshold (≥ 50th percentile). **(NEW)**

Category	Environmental, climate, or other burdens	Socioeconomic burden
Climate change	Expected agriculture loss rate ≥ 90th percentile OR Expected building loss rate ≥ 90th percentile OR Expected population loss rate ≥ 90th percentile OR Projected flood risk ≥ 90th percentile **(NEW)** OR Projected wildfire risk ≥ 90th percentile **(NEW)**	Low income*
Energy	Energy cost ≥ 90th percentile OR PM 2.5 in the air ≥ 90th percentile	Low income*
Health	Asthma ≥ 90th percentile OR Diabetes ≥ 90th percentile OR Heart disease ≥ 90th percentile OR Low life expectancy ≥ 90th percentile	Low income*
Housing	Historic underinvestment = Yes **(NEW)** Housing cost ≥ 90th percentile OR Lack of green space ≥ 90th percentile **(NEW)** OR Lack of indoor plumbing ≥ 90th percentile **(NEW)** OR Lead paint ≥ 90th percentile	Low income*
Legacy pollution	Abandoned mine land present = Yes **(NEW)** OR Formerly Used Defense Site (FUDS) present = Yes **(NEW)** OR Proximity to hazardous waste facilities ≥ 90th percentile OR Proximity to Superfund or National Priorities List (NPL) sites ≥ 90th percentile OR Proximity to Risk Management Plan (RMP) sites ≥ 90th percentile	Low income*
Transportation	Diesel particulate matter ≥ 90th percentile OR Transportation barriers ≥ 90th percentile **(NEW)** OR Traffic proximity and volume ≥ 90th percentile	Low income*
Water and wastewater	Underground storage tanks and releases ≥ 90th percentile **(NEW)** OR Wastewater discharge ≥ 90th percentile	Low income*
Workforce development	Linguistic isolation ≥ 90th percentile OR Low median income ≥ 90th percentile OR Poverty ≥ 90th percentile OR Unemployment ≥ 90th percentile	High school education < 10%

* Low Income = 65th percentile or above for census tracts that have people in households whose income is less than or equal to twice the federal poverty level, not including students enrolled in higher education **(NEW method of calculation)**

Figure 3: Correlation Heatmap of Green Space Variables

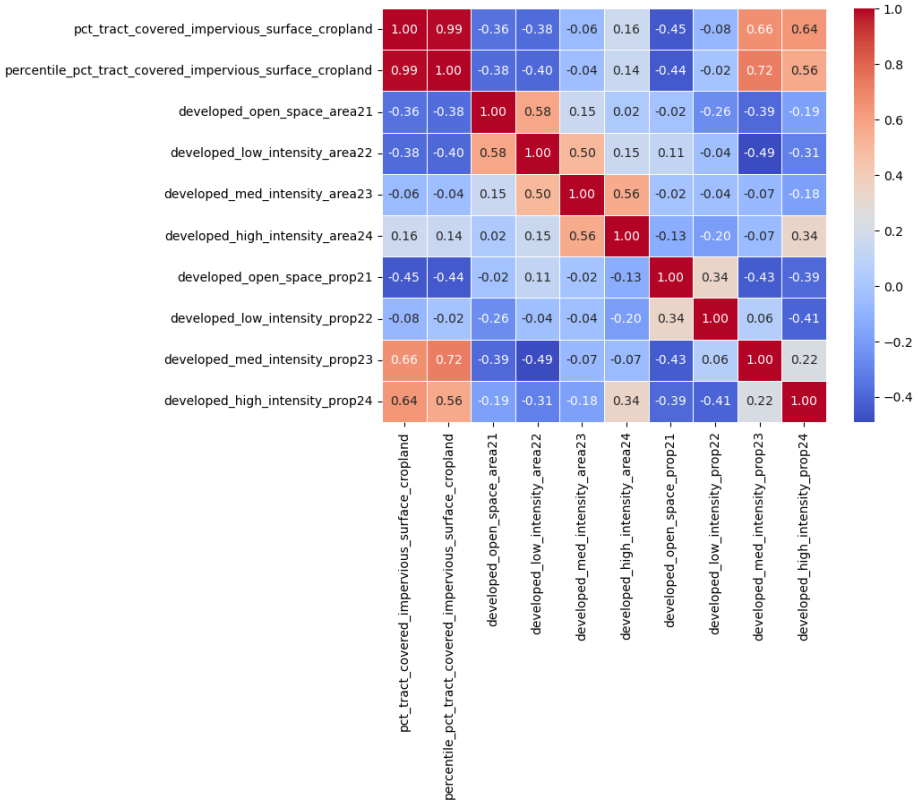
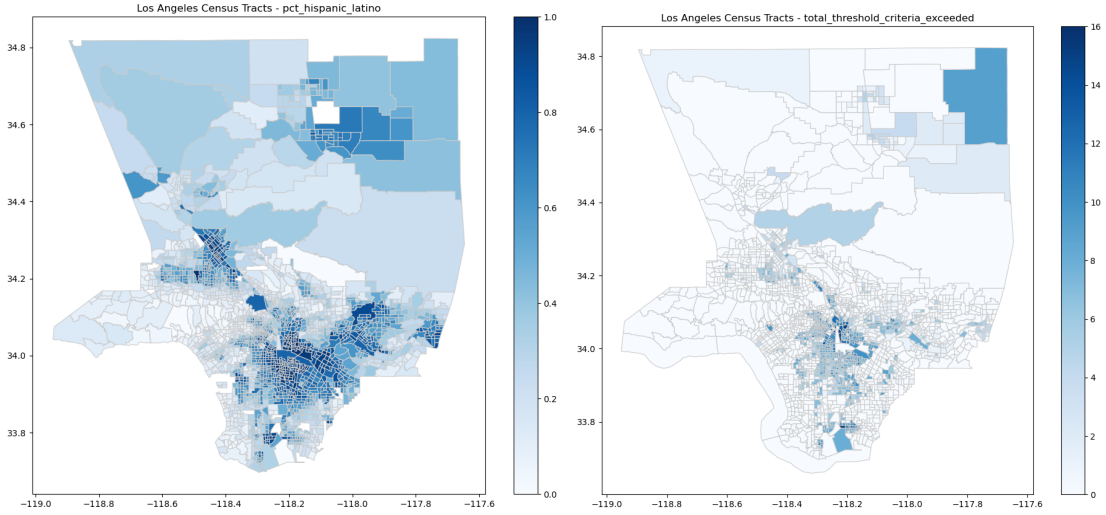
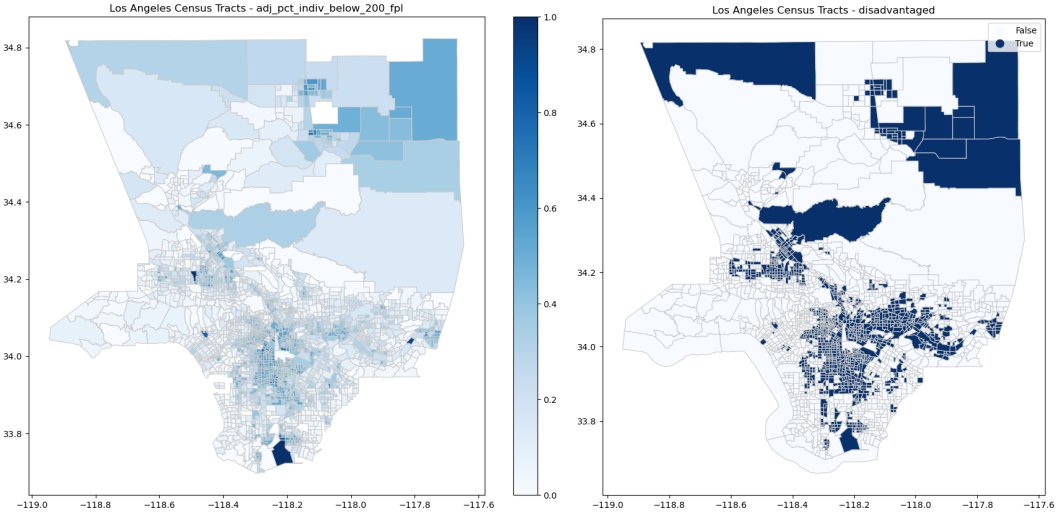
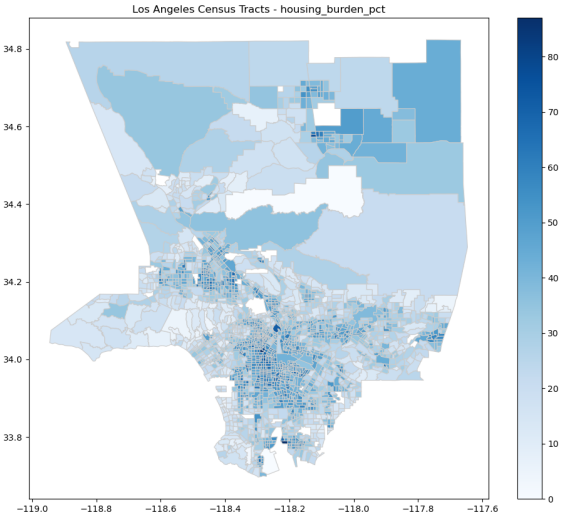


Figure 4: Los Angeles Additional Geospatial Plots




Los Angeles Census Tracts - adj_pct_indiv_below_200_fpl
Los Angeles Census Tracts - disadvantaged
False
True


Los Angeles Census Tracts - housing_burden_pct