Milestone 1: Project Abstractions

1. Basic Info

Project Title: Mapping the Dark: An Interactive Visualization of U.S. Power Outages (2019–2023)

Group members:

Name	Student ID	Email
Alethea Kramer	15219141	aletheakramer@outlook.com
Muna Ibrahim	35763119	muna.ibr08@gmail.com
Mo Fardinzaman	57779167	mfardinzaman@gmail.com
William Ho	42307041	who64@students.cs.ubc.ca

2. Overview

Power outages are a widespread issue in the United States, affecting one in four households in 2023 [1]. However, the scale, frequency, and distribution of outages may vary significantly across regions and demographics. This project aims to provide an interactive data visualization platform to explore power outages across U.S. counties from 2019 to 2023. Through an interactive choropleth map, time-based exploration, and statistical charts, users will be able to examine outage trends, identify anomalies such as during the COVID-19 pandemic, and analyze correlations between outages, income levels, and racial demographics.

This visualization is designed with policymakers, infrastructure planners, and advocacy groups in mind, hoping to provide insights that could help identify regions most in need of power grid improvements. Additionally, it serves as a resource for the general public to understand the intersectionality between power outages and their impact on different communities. By making complex outage data more accessible, this project helps drive informed discussions around energy infrastructure and identify imbalances in consistent access to energy among various populations.

3. Data

3.1 Power outage dataset

Source: Power Outages in the United States from 2019 to 2023

Attributes	Category	Cardinality/Range
fips_code	Categorical	~3000
run_start_time (Date of Data Recorded)	Quantitative	[2019-01-01 00:00:00, 2023-12-31 23:45:00]
County	Categorical	~2000
State	Categorical	53
sum (Customers Without Power)	Quantitative	[0, ~500 000]
Racial Demographics for Each County for Each Race (Population)	Quantitative	[0, ~5 000 000]

3.2 Map of the US Counties Dataset

Source: Map of the US Counties in geojson format

Attributes	Category	Cardinality/Range
STATEFP	Categorical	53
COUNTYFP	Categorical	~300
COUNTYNS	Categorical	~3000
AFFGEOID	Categorical	~3000
GEOID	Categorical	~3000
NAME	Categorical	~2000
LSAD	Categorical	9
ALAND	Categorical	~3000
AWATER	Categorical	~3000
geometry	Categorical	~3000

3.3 US Decennial Census for Race (2020)

Source: Government Decennial Census Data for Race

Attributes	Category	Cardinality/Range
County (3221 county attributes in total)	Quantitative	[0, ~10 000 000] depending on the county

The census dataset has over 3000 county attributes (columns) and 71 race groups (items/rows). The row labels include but are not limited to: 'Total', 'Population of One Race' (which is then further broken down into specific races such as 'White' or 'Black/African American'), and 'Population of Two or More Races'. The 'Population of Two or More Races' is also broken down into multiple categories, depending on the race combination. Overall, this dataset indicates the total count of each ethnicity in each county.

3.4 Derived Variables

Derived Variables	Category	Cardinality/Range
Month	Ordinal	12
Year	Ordinal	5
Total Power Outage Events	Quantitative	[0, ~35 000]
Racial Demographics for Each County for Each Race (Percentage)	Quantitative	[0, ~100]

3.5 Data Preprocessing

All data pre-processing including dataset joins, clean up, imputation, and calculations, will be conducted using a python script.

3.5.1 Power outage & Geometric Map Dataset Processing

The power outage data is logged once every 15 minutes which is too specific for our purposes. Thus, the outage data will be aggregated by month. The values will be grouped together by counting the number of outage events to highlight how disruptive the outages can be precisely because of their frequency.

Furthermore, this dataset provides each year's outage records in a separate file, so each of the 5 years of data (i.e. 2019 to 2023) will be merged for further convenience. However, much of the power outage data is inconsistent across years, and these inconsistencies must be handled prior to merging the data. For example, the variable name for the number of customers missing power in 2023 was changed from `customers_out` to `sum` which must

have consistent naming before merging. Additionally, the county names in the power outage data have incomplete naming. For example, Baltimore City and Baltimore County are both listed as “Baltimore” in the county attribute, causing a gap between the cardinality of county and FIPS codes. As a result, all joins must be done on the FIPS code.

The geometric map data’s GEOID column exactly matches the FIPS code values, allowing for easy joins. Only the geometry attribute is needed for the map data; all the other attributes from the map data can be ignored.

3.5.2 Census Race Dataset Processing

The government census data, on the other hand, doesn’t have a variable for FIPS codes clearly linked. It has the racial demographic data in the form of rows and each column denotes a single county while the power outage data has an attribute denoting the county and each power outage is represented by a row. This means the government census data needs to be transposed to be consistent with the power outage data, where each row represents a single event. The census and outage data can be subsequently merged. Furthermore, the population column is read as a string object when imported using pandas and must be converted to integer.

Finally, due to the mixed race individuals comprising a small portion of the population, individuals identifying as more than one race will be aggregated into a mixed-race category. Single race individuals will be counted as-is. The racial makeup which will be considered are as follows:

- White
- Black or African American
- American Indian and Alaska Native
- Asian
- Native Hawaiian and Other Pacific Islander
- Population of Two or more races

All other rows will be filtered out, in addition to counties that do not correspond to the 50 American states, such as those in Puerto Rico and American Samoa.

4. Tasks

#	Domain-specific task	Abstract task	Datasets/attributes
1	A policy maker or energy provider wants to identify which counties experience the highest number of power outages and whether certain counties are more prone over time.	{discover distribution} (spatial) {identify trends} over time	Dataset: power outage records, map of US counties Attributes: `fips_code`, `Total Power Outage Events`, `year`, `month`, `geometry`
2	An advocacy group wants to investigate the correlation between power outages and racial composition to see if certain communities are disproportionately affected	{identify correlation}	Dataset: power outage records, census data Attributes: `fips_code`, `Total Power Outage Events`, `Racial Demographics`
3	A public health advocacy group wants to compare power outage patterns during the COVID-19 pandemic against normal years and identify anomalies.	{detect anomalies}	Dataset: power outage records,, map of US counties Attributes: `fips_code`, `Total Power Outage Events`, `year`, `month`, `geometry`

5. Visualizations (EDA)

This analysis examines power outage records from 2019 to 2023 alongside census data to explore potential links between infrastructure reliability and socio-economic conditions, namely race. Separate exploratory data analyses (EDA) will be conducted for each dataset, first identifying trends in power outages across different counties, and then analyzing key demographic factors from the census data.

5.1 EDA of Recorded Electricity Outages 2014-2023

5.1.1 Exploratory Data Analysis of Power Outage Data (2019-2023)

Power outage records from 2019 to 2023 cover all U.S. counties, with over 126 million recorded events. The total number of outages increased yearly, from 24 million in 2019 to 26 million in 2023. To ensure consistency, `sum` (used in 2023) was merged with `customers_out` from previous years.

5.1.2 Distribution of Outages

Outages are concentrated in specific counties, with Wayne County, Michigan, recording the highest number of affected customers in 2023, followed by Oakland (MI), Los Angeles (CA), and Shelby (TN). In contrast, the most affected counties in 2022 were primarily in Puerto Rico, with Mayagüez, Ponce, and Caguas leading.

At the other end of the spectrum, rural counties such as Skamania (WA), Box Butte (NE), and Chautauqua (KS) reported significantly fewer outages in 2023. These counties consistently rank among the least-affected over time.

The urban-rural divide remains stark, with highly populated areas experiencing more frequent and widespread outages due to infrastructure demand and climate-related vulnerabilities.

5.1.3 Missing Data and Anomalies

The dataset contained 4.57 million missing values in `customers_out` (3.6%), likely due to incomplete reporting. The dataset also excludes locations that encountered no outages throughout the year. This results in around 100 counties being missing from the dataset. There were no negative values, indicating no major data corruption. The corrected dataset now maintains consistency across all years.

5.1.4 Trends and Outliers

The most outage-prone counties have shifted over time. Puerto Rican counties dominated in 2022, likely due to extreme weather events, whereas in 2023, major outages were recorded in mainland urban centers such as Wayne and Los Angeles.

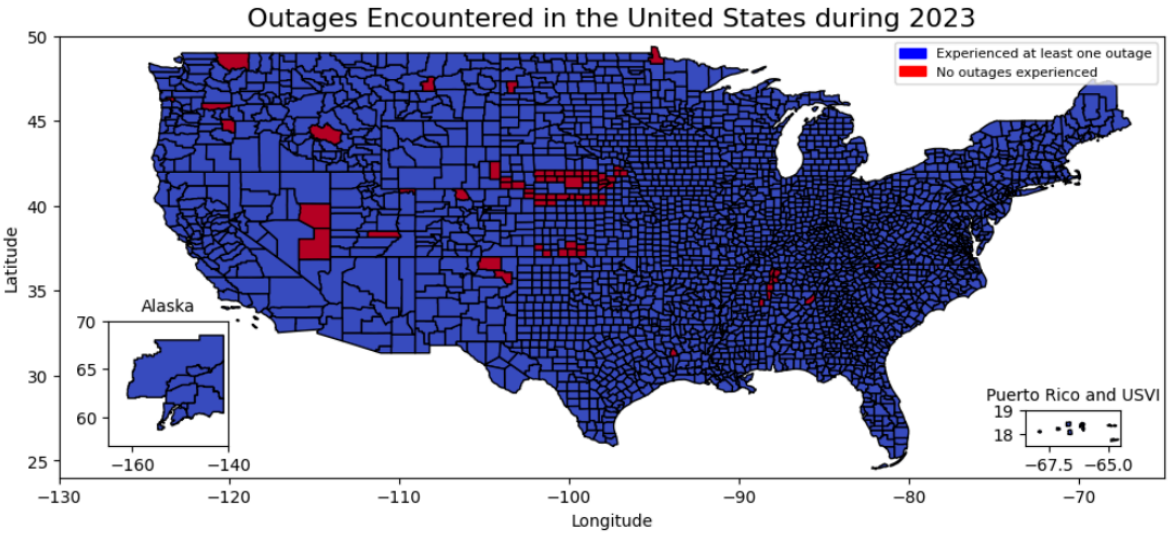
Conversely, counties with the fewest reported outages have remained relatively stable over time, with many experiencing only sporadic incidents.

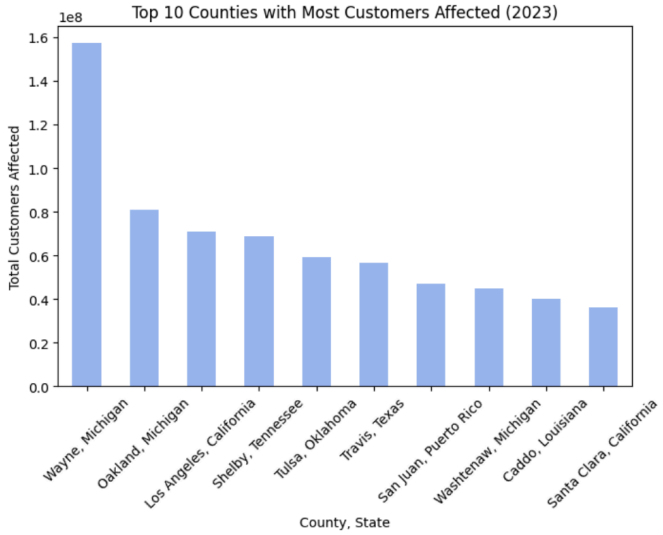
Long-term trends indicate an overall rise in outages, with spikes in certain years corresponding to extreme weather events or infrastructure failures.

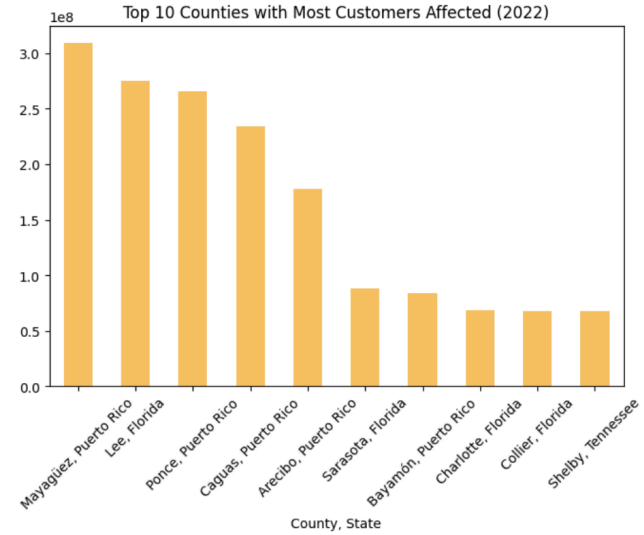
5.1.5 Conclusions

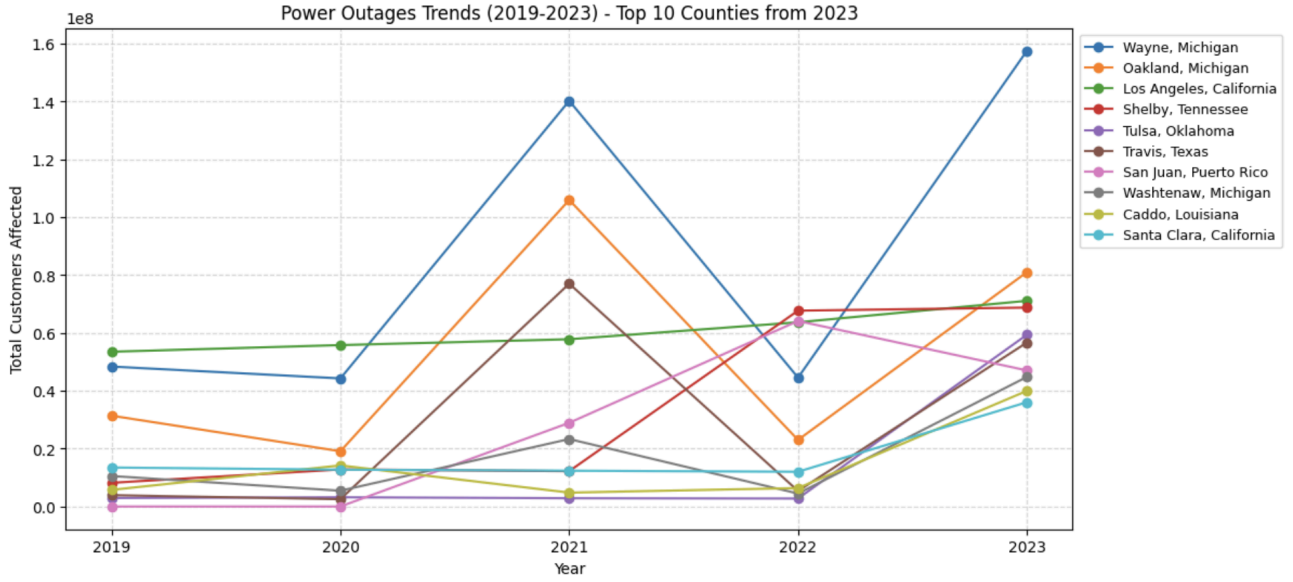
Power outages are increasing over time, with significant variation based on geography. Urban counties with high population density and aging infrastructure experience frequent disruptions, while rural counties see fewer outages.

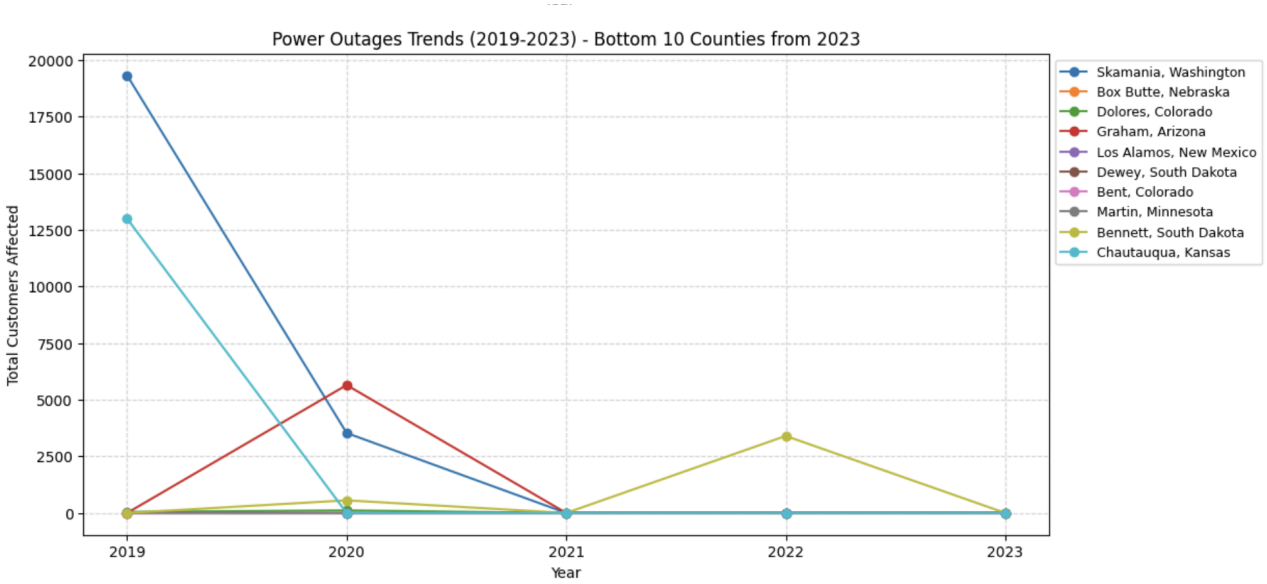
Puerto Rico experienced the most widespread outages in 2022, whereas in 2023, the highest-impacted areas were in the mainland U.S. To better understand these trends, we will integrate census data to analyze demographic, economic, and infrastructural factors contributing to outage frequency. This approach will help identify communities most at risk and inform strategies for improving grid resilience and outage mitigation.


Outages Encountered in the United States during 2023
Experienced at least one outage
No outages experienced
Latitude
Longitude
Alaska
Puerto Rico and USVI


Top 10 Counties with Most Customers Affected (2023)
1e8
Total Customers Affected
Wayne, Michigan
Oakland, Michigan
Los Angeles, California
Shelby, Tennessee
Tulsa, Oklahoma
Travis, Texas
San Juan, Puerto Rico
Washtenaw, Michigan
Caddo, Louisiana
Santa Clara, California
County, State


Top 10 Counties with Most Customers Affected (2022)
1e8
Mayagüez, Puerto Rico
Lee, Florida
Ponce, Puerto Rico
Caguas, Puerto Rico
Arecibo, Puerto Rico
Sarasota, Florida
Bayamón, Puerto Rico
Charlotte, Florida
Collier, Florida
Shelby, Tennessee
County, State


Power Outages Trends (2019-2023) - Top 10 Counties from 2023
1e8
Total Customers Affected
Year
Wayne, Michigan
Oakland, Michigan
Los Angeles, California
Shelby, Tennessee
Tulsa, Oklahoma
Travis, Texas
San Juan, Puerto Rico
Washtenaw, Michigan
Caddo, Louisiana
Santa Clara, California



5.1.6 Analysis of Power Outage Trends 2019-2023

The data shows that power outages are highly concentrated in a few urban counties, while many rural areas remain largely unaffected. Counties like Baltimore, Maryland, and Wayne, Michigan consistently rank among the most outage-prone, suggesting persistent infrastructure or environmental challenges rather than isolated events. The least-affected counties show stability, with minimal outages year after year, indicating more resilient grid systems or fewer environmental risks.

Outages in the most vulnerable areas are increasing over time, pointing to worsening grid reliability, rising energy demand, or more extreme weather events. Some counties see temporary spikes in outages, suggesting occasional disasters rather than systemic issues, but overall trends show that disruptions are becoming more frequent in high-risk areas.

Certain states, like Maryland, Michigan, and Florida, consistently have multiple counties in the most-affected lists, while others experience almost no outages. This suggests that regional factors such as climate, aging infrastructure, or urbanization contribute to repeated disruptions.

The persistence of high outage numbers in the same counties year after year suggests that underlying weaknesses are not being addressed effectively. The increasing frequency of outages in some regions raises concerns about whether grid resilience is improving or deteriorating. Targeted infrastructure upgrades in high-risk areas could help reduce outages, while further research could explore whether policy changes or climate adaptation strategies are making an impact. Understanding these trends is essential for strengthening power grid stability.

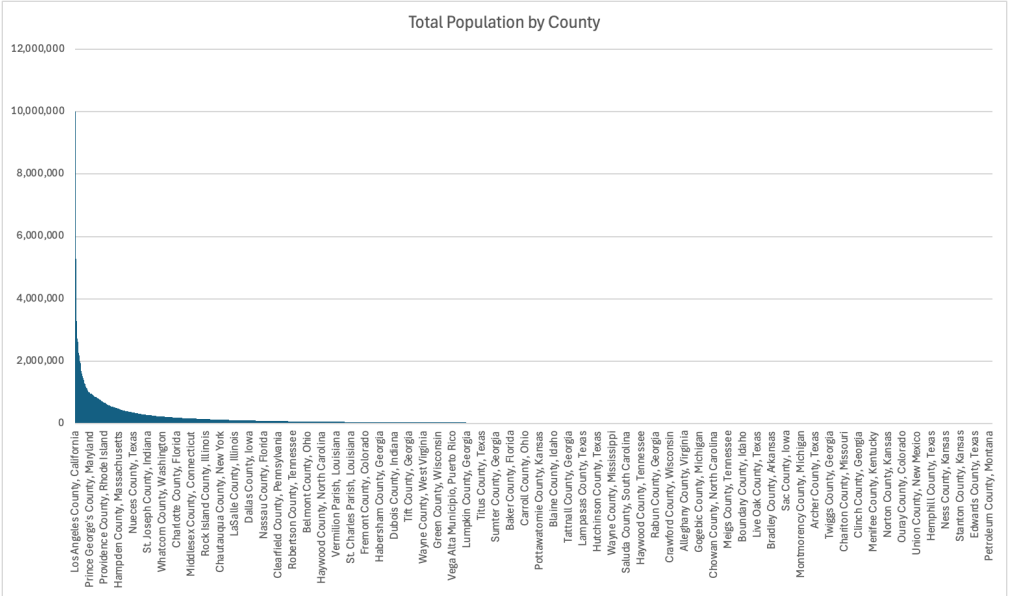
5.2 EDA of 2020 Decennial Census Race Dataset

5.2.1 Data Overview

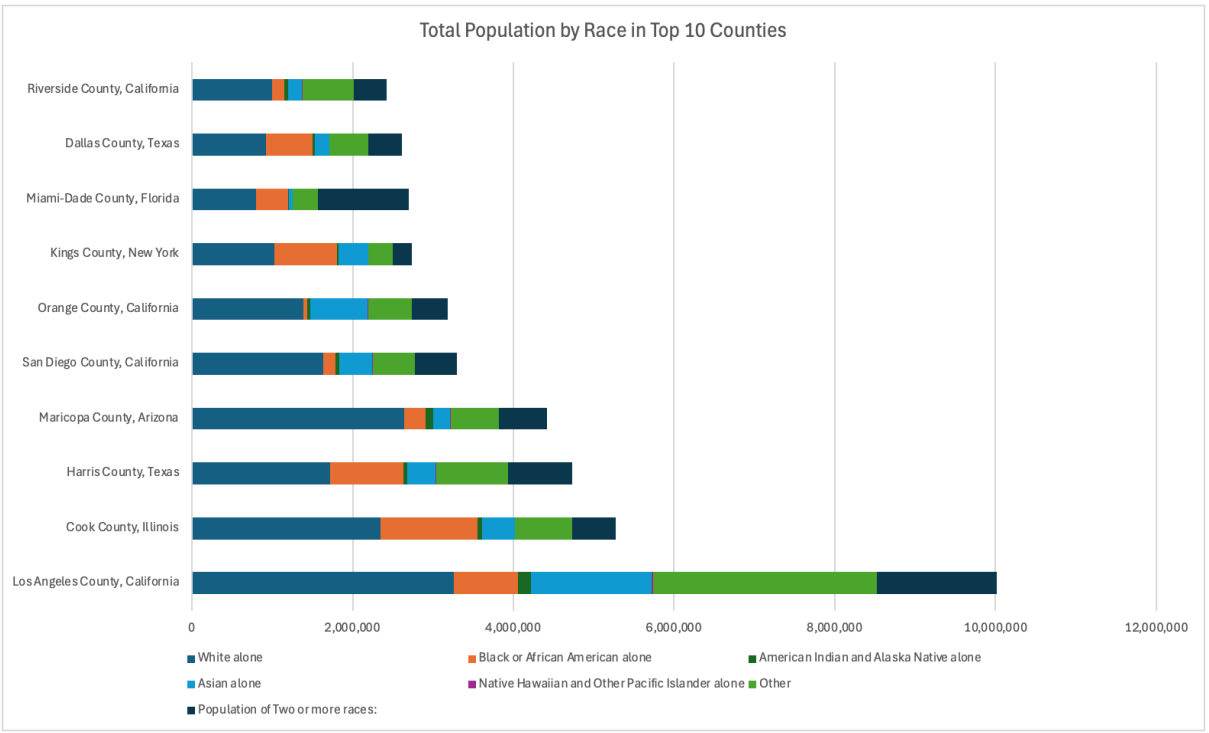
No missing data or outliers were detected in the 2020 census dataset. Furthermore, the primary focus will be on single race and the aggregate data for mixed-race categories. That is to say, we will not consider the different race-combinations for populations of two or more races.

5.2.2 Population Distribution

There is a very broad spread in the total population distribution across counties, with LA County having the highest total population at approximately 10 million people. This significantly surpasses the next most populated county, Cook County, which has a population of around 5 million. With 100 people, Loving County has the smallest population.

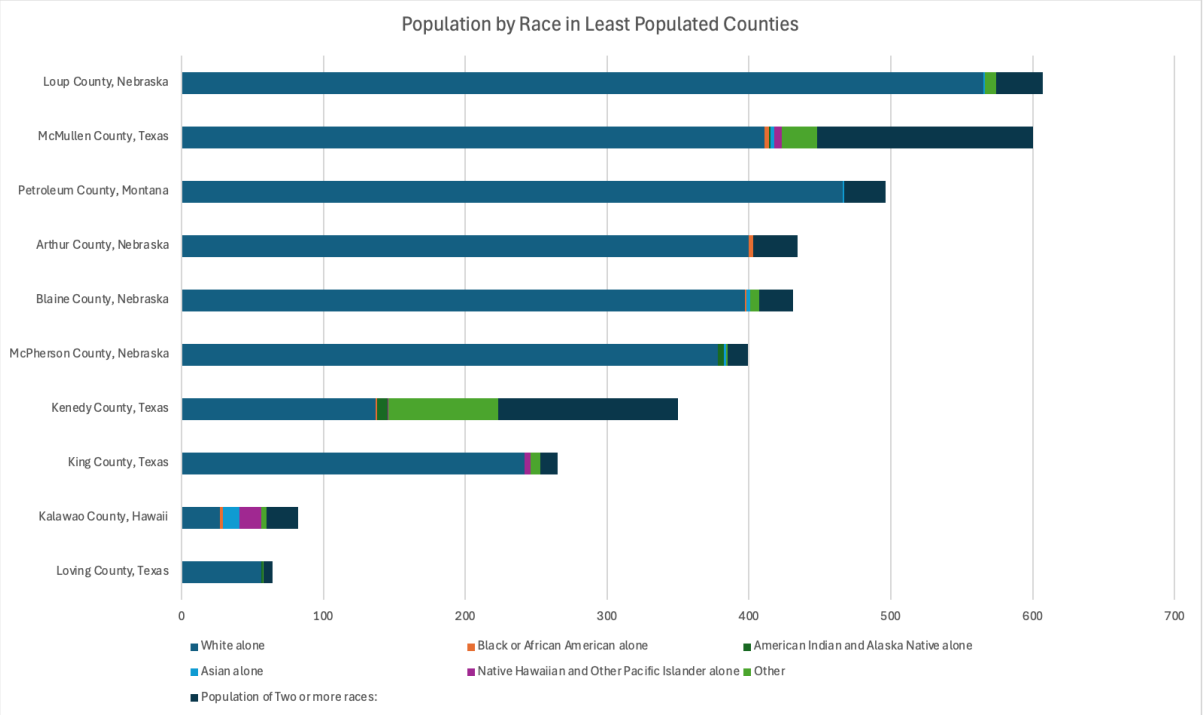


In the top ten most densely populated counties, the combined 'non-white' population generally exceeds the 'white-only' population. The most common racial categories are 'White,' 'Other,' 'Asian,' 'Black/African American,' and 'Two or More Races.'

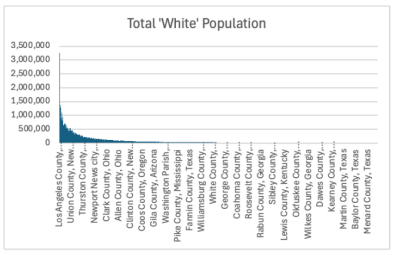


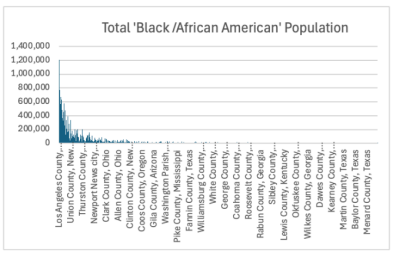
5.2.3 Racial Composition Across Counties

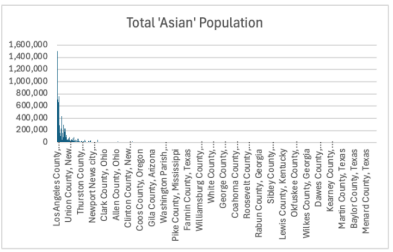
In the least populated counties, except for Kalawao County and Kenedy County, the 'White-only' population is more common than any other 'non-white' racial groups. Notably, there is a significant proportion of indigenous populations, as well as people identifying with two or more races. Asian and Black/African American populations are much smaller in these counties compared to the most populated counties.

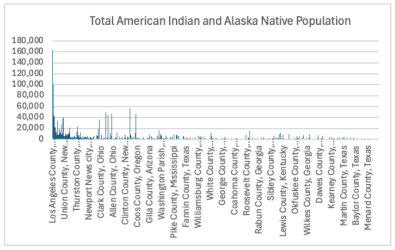


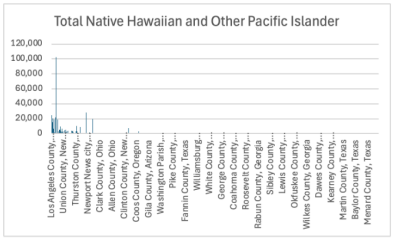
In terms of geographic distribution, the largest counties have the most diverse populations. However, indigenous populations, including significant numbers of American Indian/Alaska Native and Native Hawaiian people, are spread across smaller counties.

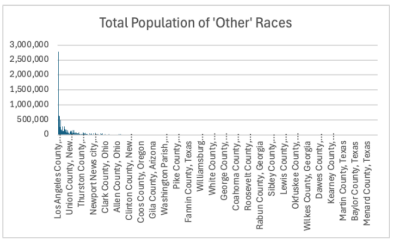


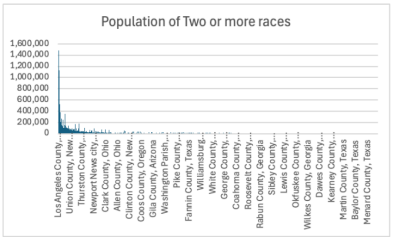












5.2.4 Anomalies and Unknowns

There is a lack of clarity regarding the ethnicity represented by the 'Other' category. Since the dataset does not include a "Latin American" category, it is likely that individuals in the 'Other' group are primarily Latin American.

Additionally, there is no clear definition of which ethnicities are categorized as 'White.' For example, it is unclear whether individuals of Middle Eastern or North African descent are considered 'White' or categorized under 'Other.

6. Usage Scenarios

[leave blank for now]

7. Team Communication Plan

Our team will use Facebook chat as our primary communication platform and meet in person every Tuesday after class, with additional meetings scheduled as needed. If necessary, we will also use Zoom for remote check-ins. We have also created a Google drive so we can all contribute to planning and writing the report asynchronously.

For asynchronous communication, team members are expected to respond within a reasonable timeframe (24 hours), while also being understanding of each other’s schedules and availability. The goal is to keep communication open and efficient without creating unnecessary pressure. If a team member becomes inactive and does not respond to messages for more than four days, the rest of the team will check in again through Facebook. If there’s still no response, we will notify the course staff.

So far, we have met once in person, once on Zoom, and have been using Facebook chat for ongoing collaboration. This setup has worked well, and we are aligned on keeping communication clear and reliable throughout the project.

8. Work Breakdown (To Date)

Each project component had designated contributors/owners responsible for ensuring its accuracy and completeness. Contributors were also responsible for writing the corresponding sections in the report. Any additional report components not explicitly listed below were completed by their respective contributors (e.g., William wrote the 'Data' section, and Mo wrote the 'Task' section). All team members participated in editing the report. The table below outlines the primary contributors for each component, along with completion dates and hours spent:

Component	Contributor(s)	Completion Date	Time Spent
Dataset selection	All	2025/01/30	0.5 hr per person
Brainstorm purpose/ domain tasks	All	2025/01/30	1.5 hours
Task Abstraction	Mo	2025/02/02	1 hr
Data Characterization *(including starting on data preprocessing)*	William	2025/02/02	3 hours
EDA for power outages dataset	Alethea	2025/02/01	2 hours
EDA for map in tandem with outages dataset	William	2025/02/03	1 hour
EDA for census data *(including preliminary data transformation)*	Muna	2025/02/02	3 hours
Report Writing - Overview	Mo	2025/02/02	45 mins
Report Writing - Team communication Plan	Alethea	2025/02/01	15 mins
Report Writing - Work Breakdown	Muna	2025/02/03	15 mins
Report Editing	Muna	2025/02/03	45 mins
	Mo	2024/02/03	30 mins

Hourly contribution per team member:

Name	Total Contribution (hr)
Alethea Kramer	2.25 + 2hr group work = **4.25 hrs**
Muna Ibrahim	4 + 2hr group work = **6 hrs**
Mo Fardinzaman	2.25 + 2hr group work = **4.25 hrs**
William Ho	4 + 2hr group work = **6 hrs**

Total time spent on M1: 22.5 hours

9. Credits

[leave blank for now]

10. Bibliography

[1] P. Madamba, “About 1 in 4 Households Experienced a Power Outage in the Span of a Year,” census.gov. Accessed: Feb. 02, 2025. [Online]. Available: https://www.census.gov/library/stories/2024/10/power-outages.html