Competitive Discretionary Grant Allocation Equity Analysis

Fall 2024

https://github.com/BU-Spark/ds-sen-markey-earmarks/tree/Final_Report/fa24-team-b

Project Members:

Data Engineering and Dashboard Creation:
Anthony Huang, 2025, tsehou26@bu.edu
John Salloum, 2025, jsbu2024@bu.edu

Data Visualizations:
Atul Das, 2025, atuladas@bu.edu
Carlos Garcia, 2025, cfg001@bu.edu
Daniel Strick, 2025, dstrick@bu.edu

Table of Contents

Create your final report and include the following:

1. **Cover page** - project name; team members names, year, email
2. **Table of contents**
3. **Introduction** - project goal & overview; what is the big picture/impact; who is the client
4. **Dataset Description** - Summarize the data, where the data originated from, any content or privacy restrictions. Include some exploratory data analysis and visualizations.
5. **Data Analysis** - Explain your methodology. What analytical techniques did you use? Provide answers to the key project questions. Attempt to answer as many overarching project questions as possible.
6. **Conclusions and recommendations** for next steps.

Additional guidance:

- All visualizations must be labeled properly (x-axis, y-axis,) and have a description below the figure explaining what it means.
- Team members must indicate in the report where they contributed work.

Introduction

Project Goal and Overview

Our team worked with Massachusetts Senator Ed Markey on data analysis for federal grant funding. Our task was to create an interactive dashboard that would help the Senator's team examine the equity of this allocation process on a local level, especially amongst BIPOC, low-income, and underserved communities.

Client

Our client was Senator Ed Markey. Ed Markey has served in the House of Representatives for 37 years, and served in senate since 2013. Senator Markey has been a leader involved in clean energy, environmental protection, and championing for policy that fosters equity and the improvement of the lives of the Massachusetts residents.

Another stakeholder within the project and who was more directly involved in the project was Liam Horsman. Liam has served as the Regional Director for the Office of Senator Ed Markey since September 2021. He has addressed regional priorities such as environmental and transit justice, energy sustainability, and affordable, accessible healthcare by facilitating communication between government offices and agencies, legislative staff, advocacy groups, and key local stakeholders.

Dataset Description

Summarize the data

Our team received multiple excel files where each row represented a federal grant. The main columns of interest included the name of the grant, the government agency that funded the grant, the total amount of the grant, and the city that received the grant.

Where the data originated from

The original grant information was provided to us from Senator Ed Markey's office, which was produced by last year's interns. However, there were some issues with this dataset which led to a new dataset being provided to us which was also provided by Senator Ed Markey's office.

Any content or privacy restrictions

The federal grant data we received has no known restrictions on its privacy that were made known to the team. The data seems to be readily available from invest.gov

Include some exploratory data analysis and visualizations

We investigated potential differences in funding between predominantly white and BIPOCcommunities as well as low-income and high-income communities.

Data Analysis

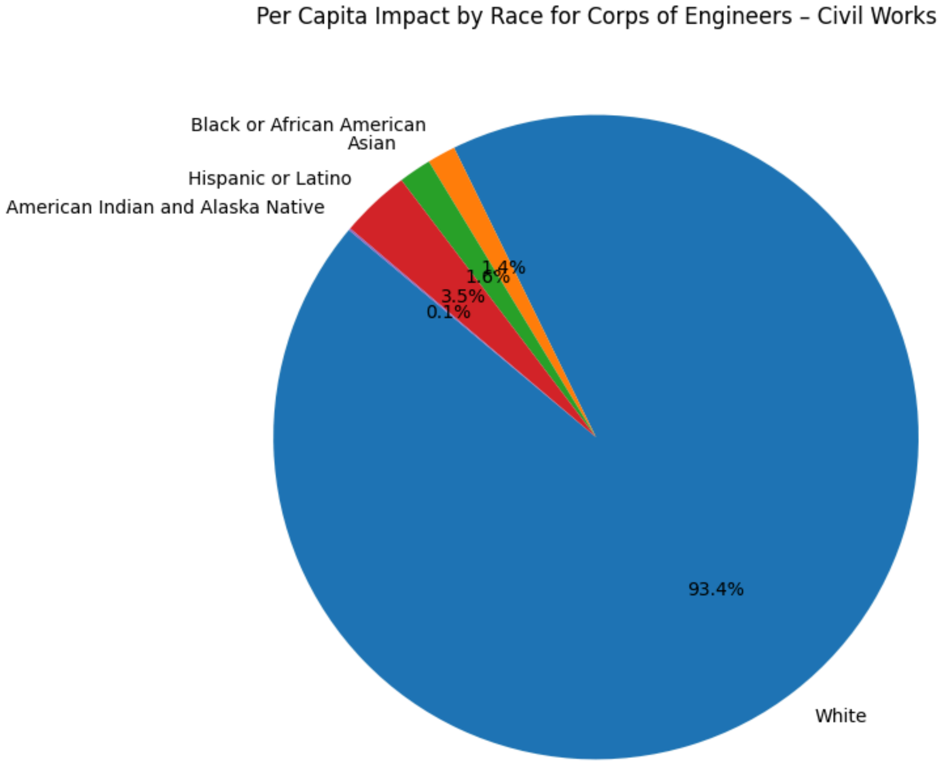
Explain your methodology

After receiving an initial csv file from our client, we analyzed it and found many faults. Particularly, the file was poorly formatted, with certain rows containing data that didn't fit the schema's structure. This seemed to indicate that the previous group's data was the result of several very faulty JOIN statements, something which our project managers agreed with.

Because of this, we started anew with Invest.gov data with each row indicating a funding initiative for a given sub county tract of Massachusetts. This came with several features such as a given agency, program name, and the funding amount given in each initiative, excluding loans. Critically, this was lacking in background information on the given area, such as demographic breakdowns or house incomes. In addition, a decent amount of regex was required to process city names as different naming schemes, cases, and spellings made each area somewhat challenging to narrow down for further analysis. Furthermore, certain cases stood out as being particularly difficult to understand in the context of our study. Many initiatives were directed statewide throughout Massachusetts; these initiatives were eventually handled by removing them as they did not serve a strong purpose in an examination of metrics in comparative funding throughout specific Massachusetts areas. In a similar sense, inter-area initiatives were removed because they were difficult to process in comparison to the rest of the data and relatively small in funding quantity compared to the rest of the data.

We performed basic exploratory analyses with the given data in order to get a sense of the data that we were working with. We first started with creating charts to observe the different types of projects and the impact they had on people of different communities. The communities were initially defined as White, American Indian and Alaska Native, Hispanic or Latino, Asian, and Black or African American. We aimed to get the per capita impact per race by category of project, which resulted in different pie charts illustrating the different projects and what communities they primarily affected. Attached is an example of what these visualizations look like.

Figure 1: Pie Chart showing the race of people impacted by grants



When approaching data enrichment, a GEO_ID value was appended to each column based on the contents of that column's city name; these GEO_IDs were derived from a sub county census tract API call which had matches for the majority of our cities. From this, we were then able to use the list of unique GEO_ID values occurring within our dataset to query the 2020 Decennial census data from the census.gov API. From the API, we specifically requested Total Population, Population consisting of one race, White Population, Black or African American Population, American Indian and Alaskan Native Population, Asian Population, Native Hawaiian and Other Pacific Islander Population, Population of other race, and Population of two or more races. This was fairly comprehensive in providing information on demographics, but it was still limited in terms of insight regarding economic factors. This necessitated a separate census call to the American Community Survey wherein we were able to retrieve Median Household income (Within the last 12 months).

From this, we grouped several features as an initial analysis. Particularly, we grouped rows based on a 79% threshold on said area's white population percentage. Similarly, we grouped income based on <$65,000, $65,000-$90,000, $90,000-$115,000, and >$115,000 groups. While these groupings were ultimately not as essential to our data analysis, they served to inform our data analysis and bridged the gap between the enrichment and proper analysis phases of this overall project.

What analytical techniques did you use?

We applied K-means clustering to each town that had funding data in order to learn more about the racial demographics and median household income. Each data point was a city in Massachusetts that received federal funding and they were grouped by median annual household income and percentage of population that identified as white. We used the elbow method to analyze the within sum of squares distance for various numbers of clusters and identified an optimal number of 3 clusters for our data: Higher income and predominantly white, lower-income and predominantly white, and lower income and predominantly BIPOC.

We transposed those three clusters on an actual map of Massachusetts to see geographically where the wealthier and predominantly white cities were located as well as the lower-income and BIPOC cities and the lower-income predominantly white cities. There were distinct patterns in the geographic locations of these cities. This map was interactive so the user can click on a city and see insights into the city's total population, total federal funding, percent of population that identifies as white, funding per capita, and which cluster they belong to. We used Nominatim to get the latitudes and longitudes of every city and county in Massachusetts. After obtaining the data, we used Folium to plot the map which initially resulted in a map similar to the ones in Google Maps. We then used the MASSGIS shape file to get the boundaries of all cities and counties in Massachusetts and then plotted the resultant maps.

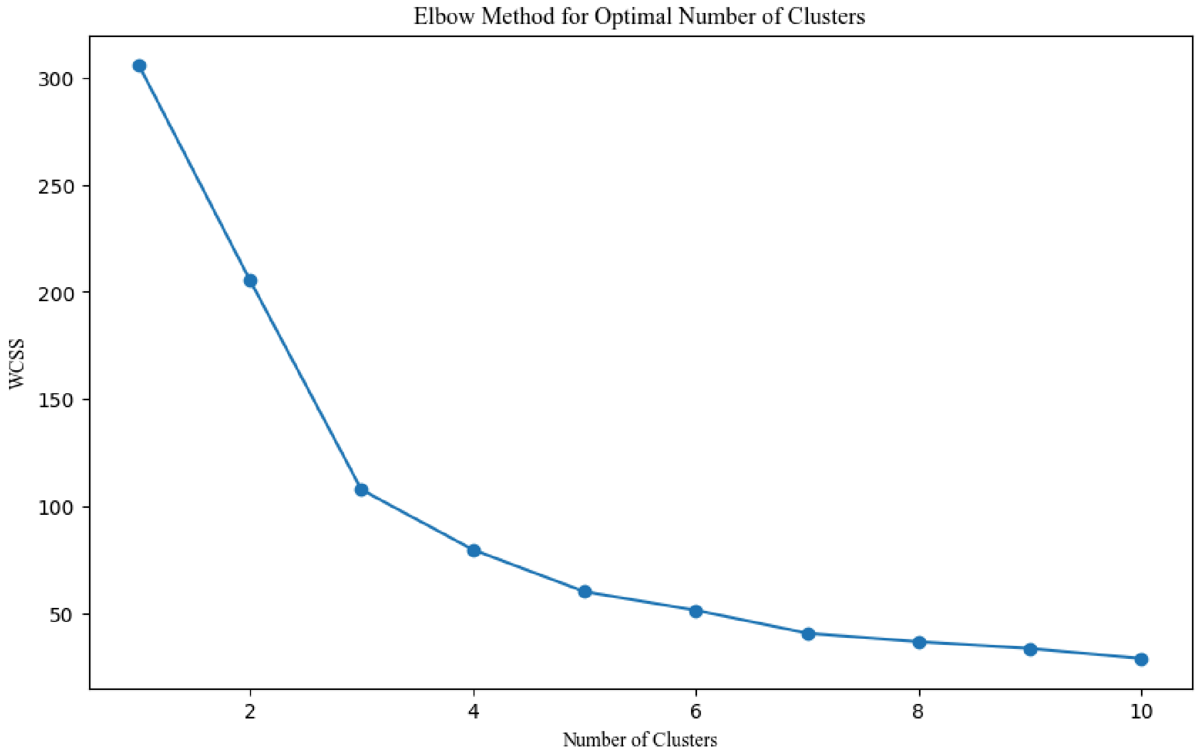


Figure 2: Elbow Method for Optimal Number of Clusters
This elbow curve illustrates the within-cluster sum of squares (WCSS) for different numbers of clusters (ranging from 1 to 10) used in the K-means clustering algorithm. The x-axis represents the number of clusters, while the y-axis shows the WCSS. The "elbow" point, where the rate of decrease sharply shifts, helps determine the optimal number of clusters. In this case, three clusters are identified as optimal, as the WCSS reduction slows significantly beyond this point.

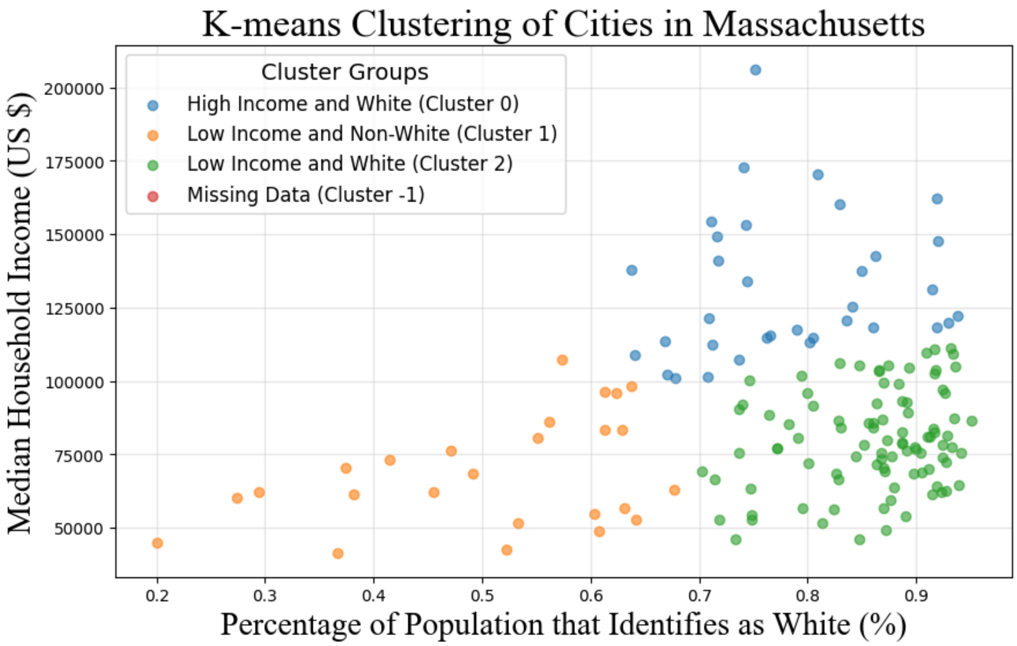
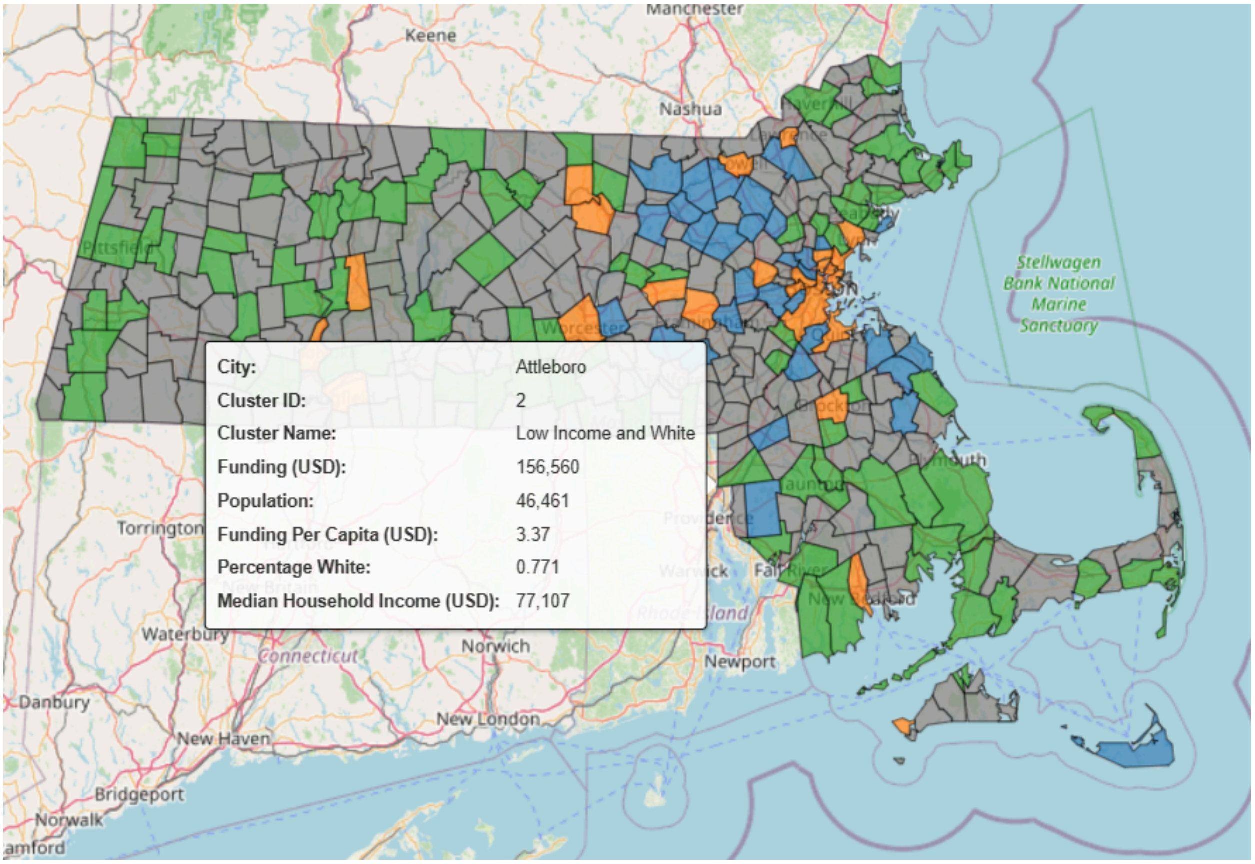


Figure 3: K-means Clustering of Cities in Massachusetts

This scatter plot visualizes the clustering of cities in Massachusetts based on their median household income and the percentage of the population identifying as white. The cities are categorized into three clusters:

- **Cluster 0 (Blue)**: High Income and Predominantly White
- **Cluster 1 (Orange)**: Low Income and Predominantly Non-White
- **Cluster 2 (Green)**: Low Income and Predominantly White

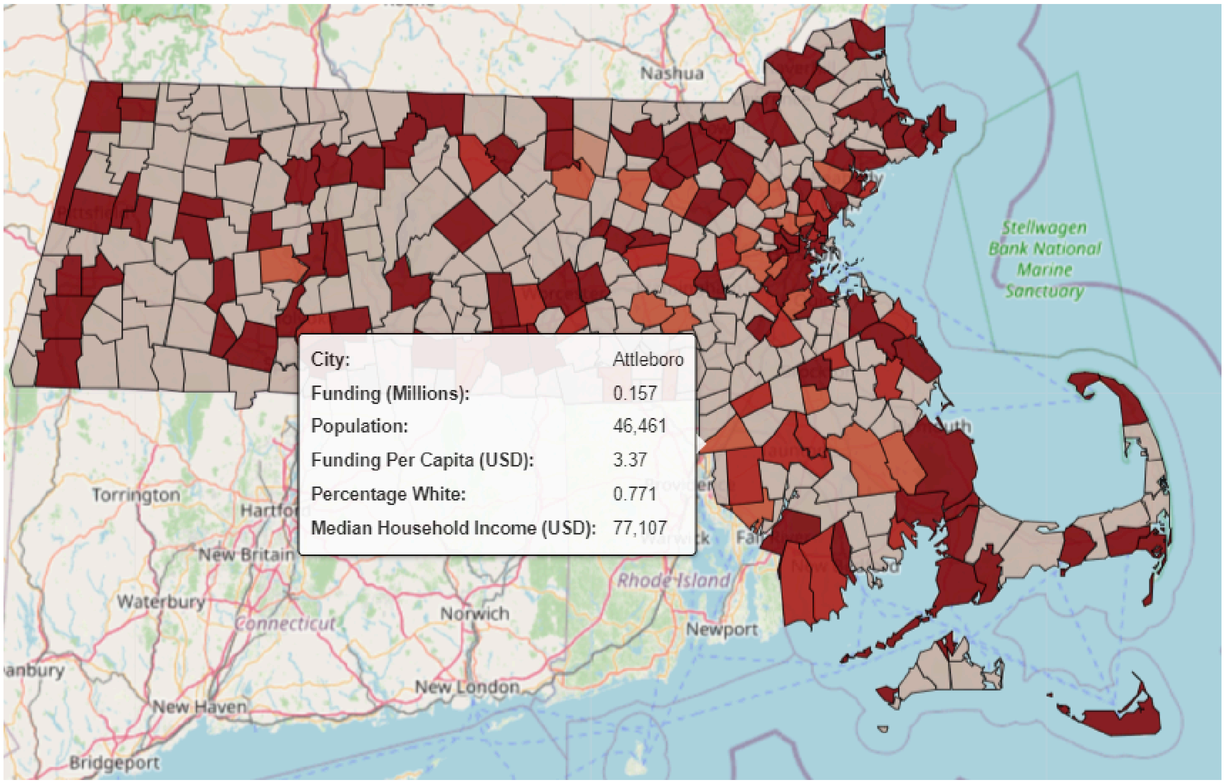
Each data point represents a city, with the x-axis showing the percentage of the white population and the y-axis indicating the median household income. The clusters are color-coded for clarity, highlighting distinct socio-economic groups within the state.



City:	Attleboro
Cluster ID:	2
Cluster Name:	Low Income and White
Funding (USD):	156,560
Population:	46,461
Funding Per Capita (USD):	3.37
Percentage White:	0.771
Median Household Income (USD):	77,107

Figure 4: K-means Clustering of Cities in Massachusetts
This map visualizes the clustering of cities in Massachusetts by their actual geographic locations. Each city is color-coded according to its cluster membership. Hovering over a city displays detailed information, including the city's name, cluster ID, cluster label, total funding amount, population, funding per capita, percentage of white population, and median household income. This interactive map allows for an intuitive understanding of the socio-economic distribution across Massachusetts cities. Cities in grey did not have any funding information and demographic information was also unavailable.

We also created three heat maps of the state of Massachusetts including cities that received federal funding. Each map uses a darker shading to highlight cities with higher values, providing a clear visualization of the variable's distribution across the state. The three maps represented funding per capita, percentage of population identifying as white and median annual household income for each city. We were interested to see how median household income and racial demographics correlated with funding per capita as well as the geographic distributions of funding per capita, income and race.

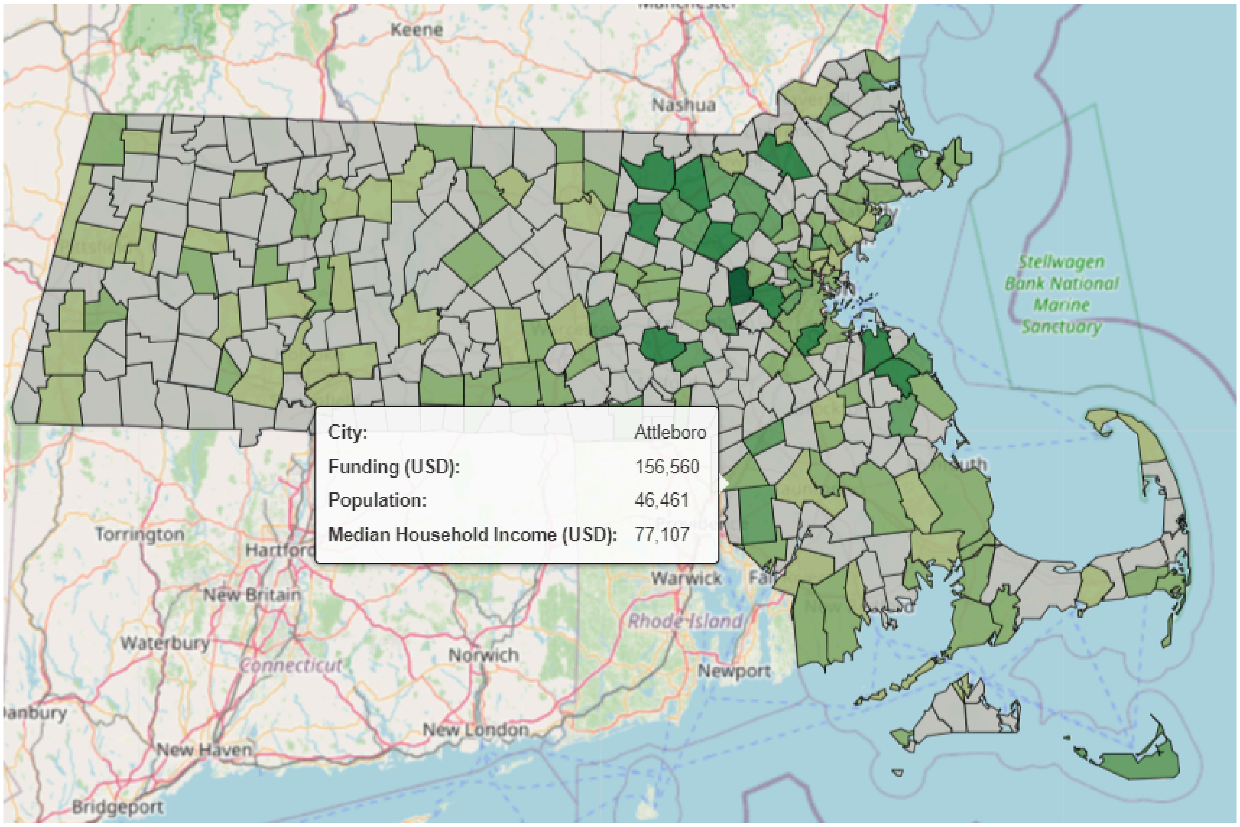


City:	Attleboro
Funding (Millions):	0.157
Population:	46,461
Funding Per Capita (USD):	3.37
Percentage White:	0.771
Median Household Income (USD):	77,107

Figure 5: Funding Per Capita Distribution Across Massachusetts Cities

This interactive choropleth map visualizes the distribution of funding per capita (excluding loans) across cities in Massachusetts. The map uses a logarithmic scale for funding per capita to normalize the data, providing a clearer representation of funding distribution. The color gradient from light to dark red indicates increasing levels of funding per capita, with darker shades representing higher funding. Cities in grey did not receive any funding.

Hovering over each city reveals detailed information including the city name, total funding (in millions of dollars), population, funding per capita, percentage of the population identifying as white, and median household income. This map helps to understand the geographic and socio-economic distribution of funding across the state.

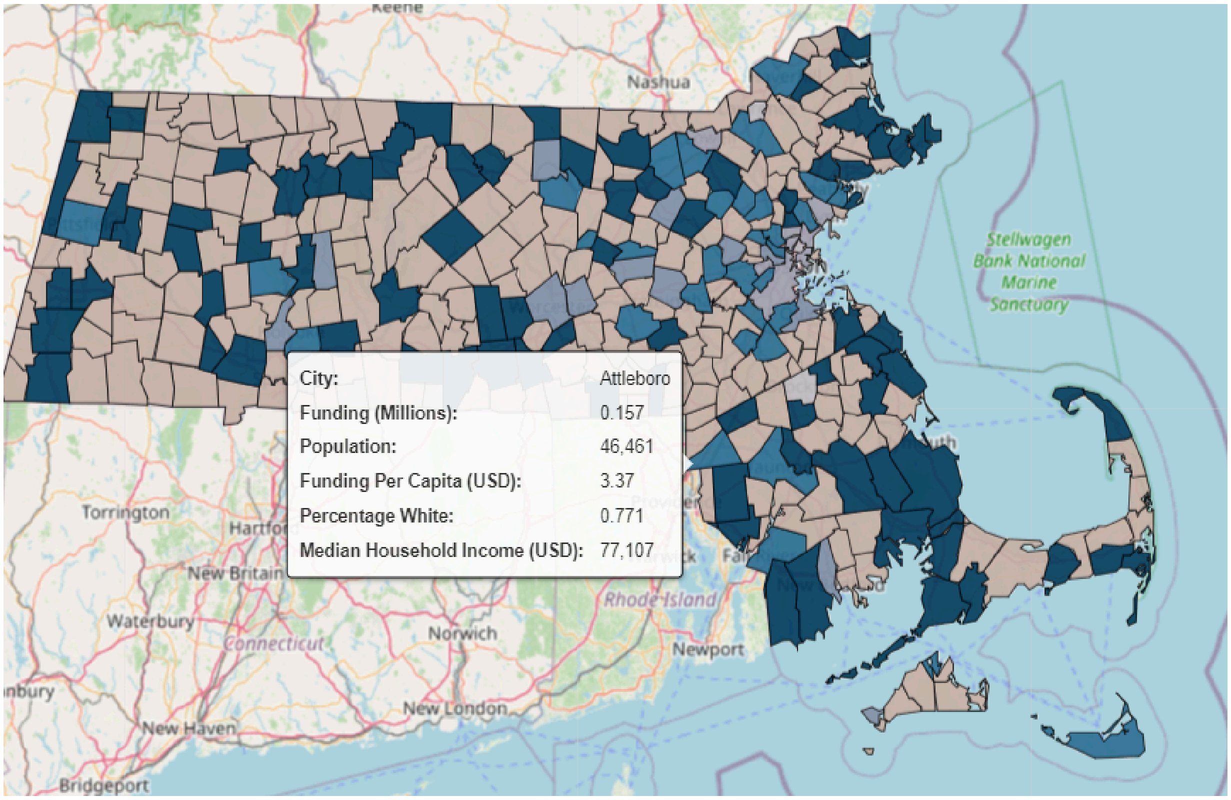


City: Attleboro
Funding (USD): 156,560
Population: 46,461
Median Household Income (USD): 77,107

Figure 6: Median Household Income Across Massachusetts Cities

This interactive choropleth map visualizes the median household income across cities in Massachusetts. The map uses shades of green to represent varying levels of median household income, with darker shades indicating higher incomes. Each city is color-coded based on its median household income, allowing for a visual comparison of economic status across the state. Cities in grey did not have any funding information and demographic information was also unavailable.

Hovering over each city provides detailed information, including the city name, total funding amount (excluding loans), population, and median household income. This map offers insights into the distribution of household income across Massachusetts cities, highlighting areas of economic disparity and prosperity.

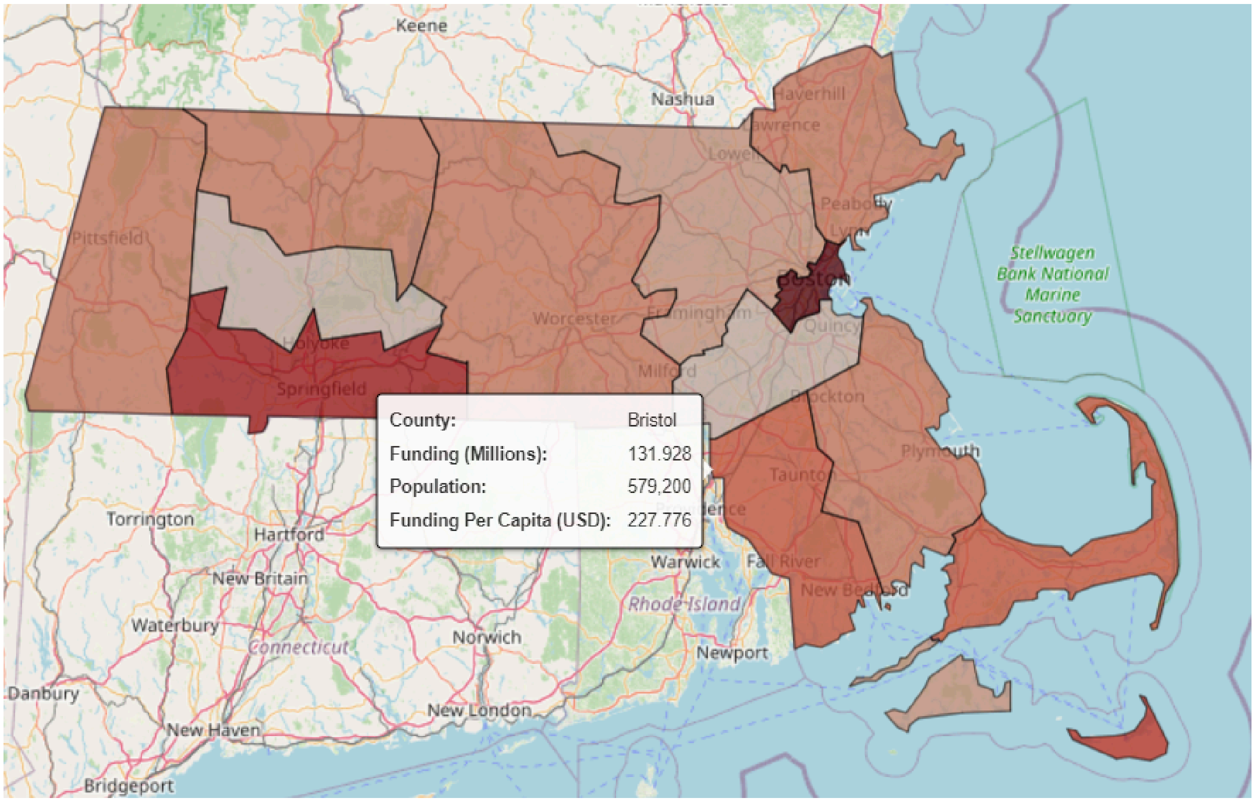


City:	Attleboro
Funding (Millions):	0.157
Population:	46,461
Funding Per Capita (USD):	3.37
Percentage White:	0.771
Median Household Income (USD):	77,107

Figure 7: Percentage of White Population Across Massachusetts Cities
This interactive choropleth map visualizes the percentage of the population identifying as White across cities in Massachusetts. The map uses a color gradient from light blue to dark blue, representing increasing percentages of the White population. Each city is color-coded based on its percentage, with darker shades indicating a higher percentage.

Hovering over each city reveals detailed information, including the city name, total funding (in millions of dollars), population, funding per capita, percentage of the population identifying as White, and median household income. This map provides an intuitive understanding of the demographic distribution across the state. Cities in grey did not have any funding information and demographic information was also unavailable.

We also looked at funding per capita on the county level to see how funding is distributed across wider regions of the state. We noticed Suffolk county has by far the highest funding per capita of any county and suspect that population density is correlated with funding per capita.



County: Bristol
Funding (Millions): 131.928
Population: 579,200
Funding Per Capita (USD): 227.776

Figure 8: Funding Per Capita Distribution Across Massachusetts Counties
This interactive choropleth map visualizes the distribution of funding per capita across counties in Massachusetts. The map uses a color gradient from light red to dark red, representing increasing levels of funding per capita. Each county is color-coded based on its funding per capita, with darker shades indicating higher funding levels.
Hovering over each county provides detailed information, including the county name, total funding amount (in millions of dollars), population, and funding per capita. This map offers insights into the geographic distribution of funding across Massachusetts counties, highlighting areas with varying levels of financial support.

To further investigate equity, we created scatter plots with the variables of funding per capita vs median annual income for each city as well as funding per capita vs percentage of population identifying as white for each city. Within each scatter plot we created a line of best fit and calculated a correlation coefficient.

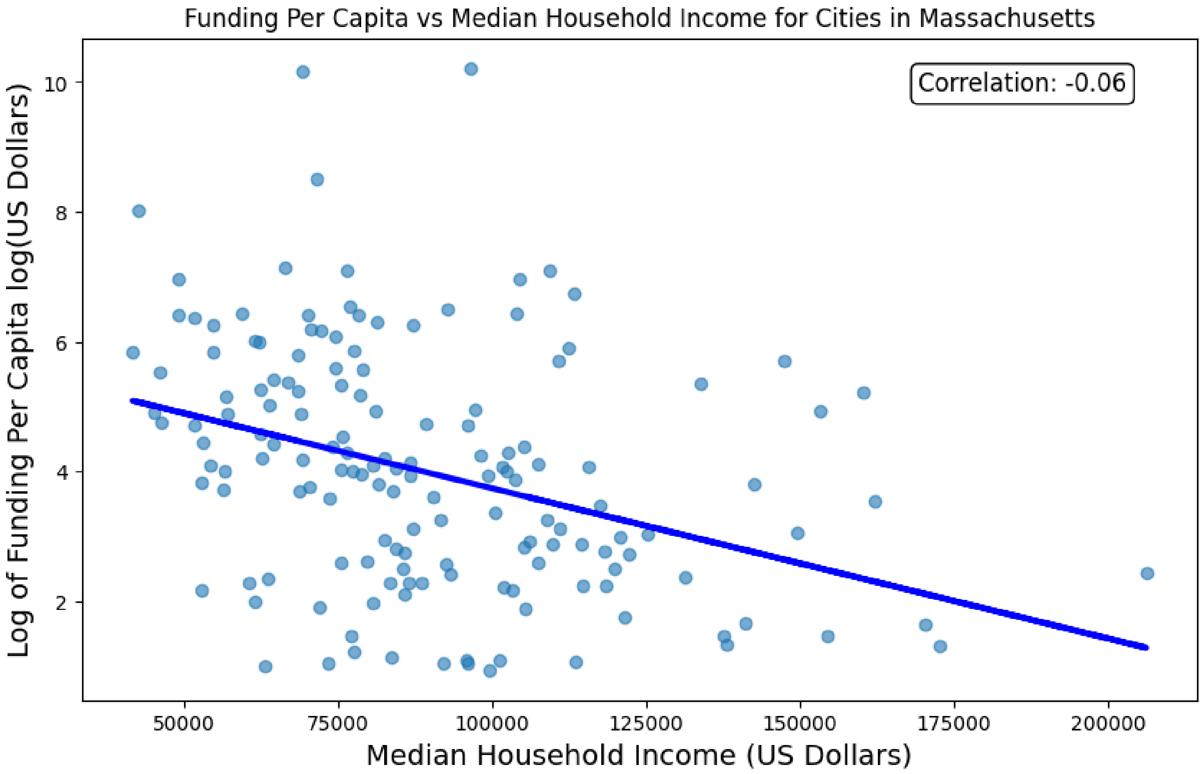


Figure 9: This scatter plot illustrates the relationship between the median household income and the logarithm of funding per capita for cities in Massachusetts that received federal grants.

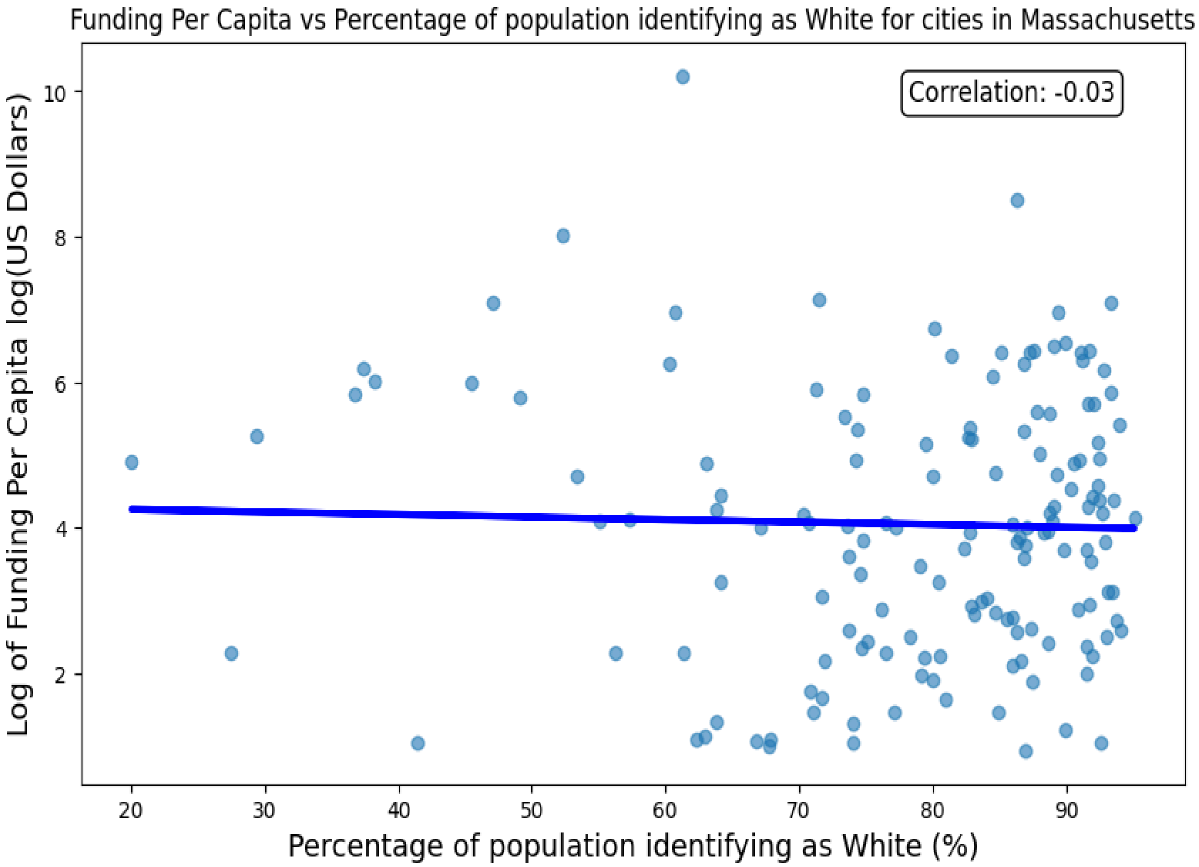


Figure 10: This scatter plot illustrates the relationship between the percentage of the population identifying as white and the logarithm of funding per capita for cities in Massachusetts that received federal grants.

Provide answers to the key project questions. Attempt to answer as many overarching project questions as possible

As for the final deliverable, we created a Looker Studio dashboard that allows the client to better visualize the distribution of the grants and filter out specific demographic information. The client can retrieve information for specific cities, race majorities and income levels that allow for a finer understanding of the current climate. The maps and charts in the dashboard adjust to the specific information the client would like to review, allowing them to make more informed decisions for the following year's grants. From the visualization themselves we saw that low income, highly diverse cities generally received more funding. Likewise clean energy initiatives such as environmental remediation, and clean water and energy comprised the second and third highest funding grant categories. These grants were notably focused on supporting gateway communities.

Population metrics are also crucial in our analysis. Of note, population density changes drastically across the state and may play a pivotal role in funding allocation. This issue is particularly relevant for transportation funding as these generally go to large urban areas. A general view of our analysis would demonstrate the population density is correlated with total funding, however, this trend is largely due to transportation funding alone.

Conclusions and recommendations for next steps

In conclusion, the analysis shows that grants are predominantly allocated to low-income, highly diverse areas, which aligns with their intended purpose. Environmental remediation grants were the most common in these regions, indicating significant efforts to address issues related to rundown properties. Similarly, transportation funding accounted for approximately 42% of all grants, reflecting its prominence in big cities where public transportation is more accessible; this category could be excluded in future analyses to prioritize other types of grants. The dashboard provided a comprehensive overview of the data and offers potential for further exploration with grants for future years.

Population density data was difficult to obtain but may play a key role in determining funding allocation. Notably, urban areas received significantly more funding related to public transportation. It would be interesting to explore how the funding distribution would look if public transportation was removed.