Team 079 Final Report
Erick Garcia, Saisuriya Paranthaman, Viet Nguyen, Yuchen Wang, Binh Vu, Yong Yan Zhu

Bridging the Gap between Citizens and Local Governments: A Democratized Data Solution

Introduction

Many regions of the United States require government assistance, but government resources are finite. The Federal Reserve uses quantitative easing to stimulate economic activity and help provide assistance to individuals and businesses(Goldrick-Rab(7)). Watkins(20) argued against this approach as it exacerbates the wealth and income gaps. Another approach is to identify individuals/families in lower income brackets using unemployment rates (Muller(14), Bartik(2)) or Earned Income Tax Credit (EITC). However, Shaefer(18) and Falcettoni(4) discovered that not everyone benefited equally from unemployment insurance. EITC is a tax break for low-to-moderate income workers and Hoynes(8) found it helps its target demographic, but it too has its flaws.

Problem definition

There is a lack of transparency in government decision-making for aid distribution and biases (i.e. intrusion of corporate interests) (Tenney & Sieber(19)). Increasing transparency using ground-truth (i.e., government-produced) datasets will increase citizen input, which in turn allows local governments to better represent their constituents (Franklin & Ebdon(5)). We aim to use a clustering algorithm with multiple government data sources to develop a democratized tool that allows both local governments and citizens equal access to ready-to-consume data that will help address these issues. Our tool will allow the user to explore clustered data, apply human intuition, and identify areas that require the most aid and potentially create more equal policies.

Literature survey

Research shows poorer households pay more income taxes than the ultra wealthy (Institute on Taxation and Economic Policy (9), Loughead(13)), but the benefits do not translate proportionally (Glomm (6), Baicker(1)). Therefore, we will use income tax statistics data including UI and EITC for our clustering algorithm. Our tool can identify which regions pay the most taxes and can support policy changes that help minimize wealth inequality by making state and local tax systems more progressive and increasing the effectiveness of social welfare programs. Users can determine which regions require the most aid based on their criterias since there are many solutions. Buchanan(3), Joliffe(10), and Mueller(14) agree providing support to low-income areas equalizes wealth. However, Joliffe(10) suggested reducing poverty when adjusted for cost of living in metro areas while Mueller(14) argued for more support for rural areas due to the lack of economic activities there.

We will implement an unsupervised clustering algorithm k-means to segment income data in an unbiased manner, one where part of our research is incorporated. K-means is perfect for our case because each assigned cluster will be characterized by the greatest similarity within the same cluster and the greatest dissimilarity between different clusters (Sinaga and Yang (11)). That means, based on the user-selected cluster, a dynamic allocation of a cluster will be assigned to a particular zip code.

Proposed method

Intuition

Our approach to aid distribution is based on solid academic research and uses a clustering algorithm with numerous government data sources. It might offer a solution that is more efficient than current methods. We can partition income data objectively by utilizing k-means clustering, a commonly used and tested unsupervised clustering procedure, and guaranteeing the most similarity inside the cluster and the most remarkable dissimilarity between distinct clusters. This method is good at seeing patterns and similarities in data, making it an excellent tool for allocating aid. We seek to empower local governments and communities to make more informed decisions and advance more substantial equity in aid distribution by increasing openness and granting access to this information.

Testbed - Hardwares/Software Tools

For our project, we used a standard quad-core PC with no special graphics card running Windows 11. For data/algorithm, we used python 3.11 working with a mixture between Visual Studio Code and Jupyter Notebook. We used Tableau for visualization and sharing those results with Tableau Public.

Data/Algorithm

Our primary focus involves utilizing two data sets: (a) IRS - Individual Income Tax Statistics - ZIP Code Data and (b) USASpending.gov - Financial Assistance - ZIP Code Data. The income dataset gives total income, taxes paid, unemployment payment, net investment, and real estate info based on zip code. We utilize this dataset to gain insights into income per zip code, along with our visualization, enabling us to identify zip codes with higher income and target them for increased government funding collection. For each zip code, total income is broken down into six brackets. (**Innovation 1**) In order to compare different zip code incomes, we have to calculate the weighted income average for each zip code.

$$\sum_{b=1}^{b}(\text{Income Amount}_b * \text{Num of return}_b) / \sum_{b=1}^{b}(\text{Num of return}_b)$$

where b = bracket classification. The award dataset shows how much the US government has given in financial assistance to each zip code. Based on this info, the location and zip code where we can see which areas receive more awards compared to others. Furthermore, this same equation is applied to other factors such as unemployment, investment, state tax, and real estate tax.

(**Innovation 2**) The unsupervised clustering algorithm - K-means. Each iteration, the centroid is updated based on the following algorithm.

$$CP(x_1, x_2, ..., x_k) = (\frac{\sum_k^{i=1} x^{1st}}{k}, \frac{\sum_k^{i=1} x^{2nd}}{k}, ..., \frac{\sum_k^{i=1} x^{nth}}{k})$$

where n dimensional centroid points amid k n-dimensional points. Each point is categorized based on the euclidean distance to the closest centroid. The algorithm will start with randomly selected centroids used as points for every cluster. It then performs iterative calculations to select the positions of the centroids optimally. The algorithm continually calculates the position of the centroids until it reaches the defined number of iterations or experiences no change in the centroid values.

We are using K-means clustering on this data set because it is relatively simple to implement. It's the most well studied clustering algorithm that is easy for other contributors to assist with. K-means is also very fast compared to other clustering algorithms and doesn't require calculating pairwise distances between points. This means the performance will scale linearly with the number of points in the data set. The more data points, the longer it will take for the algorithm to run. It scales to large data sets and guarantees convergence.

This ML model also adapts easily to new examples and datasets as the positions of the centroids adjust according to the data set so not much work is needed to adjust the algorithm. K-means can also adapt to naturally imbalanced clusters of different shapes and sizes. The number of clusters can be increased or cluster width can be changed so the problem of different cluster sizes can be mitigated.

There are some disadvantages to K-means as it is dependent on initial values. As the number k increases, you need advanced versions of k-means to pick better values of the initial centroids which is called k-means seeding. K-means is very sensitive to initial starting conditions as the same results may not be obtained if changes are made to the initial conditions. K-means does have trouble clustering data where clusters are varying sizes and density. To cluster this type of data, we would need to generalize K-means by varying widths and dimensions of the cluster. Centroids can also be dragged by outliers or in some cases might get their own cluster. To address this issue it's important to remove outliers when cleaning the data set. Another disadvantage is if the number of dimensions increases, a distance-based similarity converges to a constant value. To reduce the dimensionality, PCA can be used on the feature data or spectral clustering to change the clustering ML algorithm.

K-means is also sensitive to the scale of the variables. If one of the variables is on a much larger scale than the others, the variable will have an outsized effect on the distance calculated from the data points to the centroid. It's also difficult to incorporate categorical variables as K-means is meant to be used for numeric features.

With the cleaned income dataset, we execute the K-means clustering model with a range between 1-20 clusters to examine potential similarities between data points within the data. The attributes mentioned in income data are iteratively added to the training set. For example, the first model trained only on weighted average, the second model on weighted average and taxes paid, etc. From those trained dataset, we hand selected the best combinations of clusters visually. Ultimately, we decided to use K-mean clusters with all factors as our main dataset for visualization.

Visualization

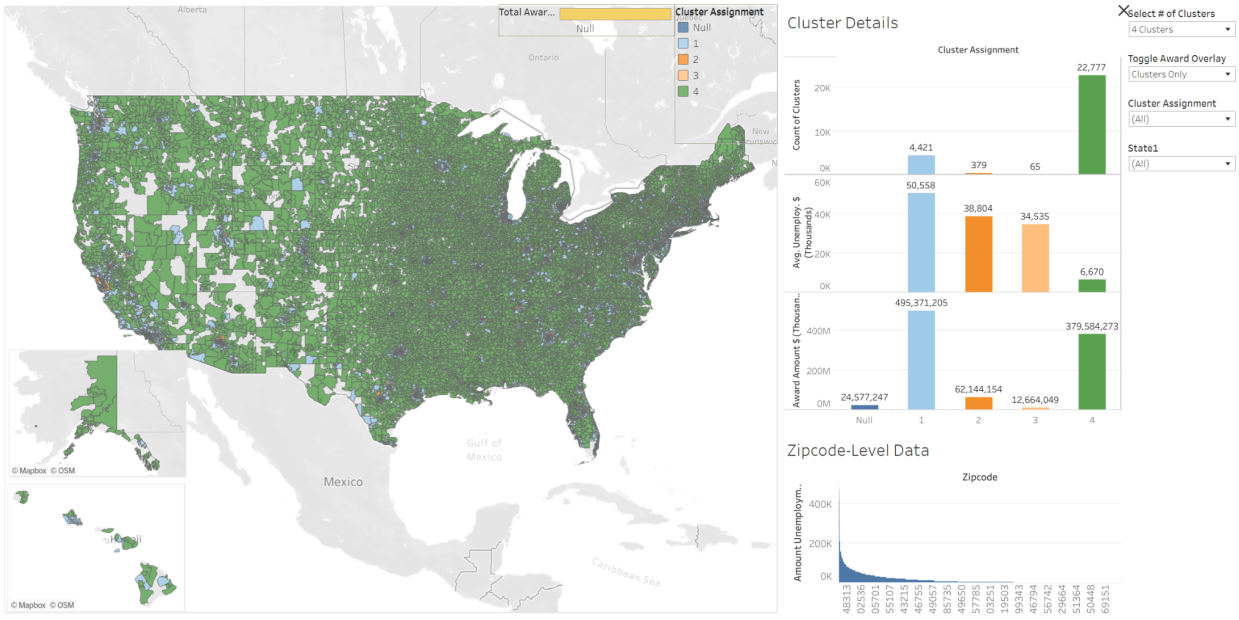
Our visualization was developed in Tableau and focuses primarily on displaying geospatial data and various overlaid attributes. Tableau was selected as a platform for several reasons:

- Popularity for data visualization;
- Relative ease of working with geospatial datasets;
- Availability of licenses for students (low-cost); and
- Portability for hosting in Tableau Public

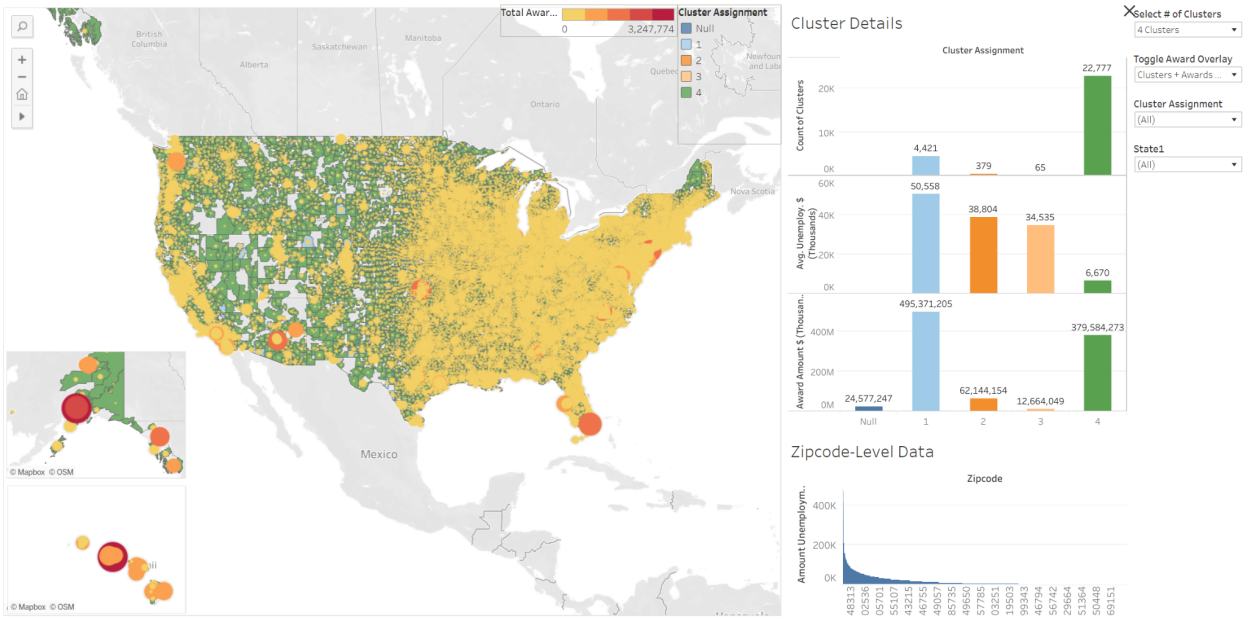
The visualization focuses on one dashboard that features a map of the United States. The map is partitioned based on states and zip codes. The user can toggle between two views: “Clusters Only” and “Clusters + Awards Merged”. Each zip code will have a color coding based on the cluster assigned by the K-means algorithm. The K-means algorithm generated 2-20 clusters, and in both views, the user can choose how many clusters they want to explore in the “Select # of Clusters” option. There are accompanying bar graphs which inform about 1) number of zip codes per cluster, 2) award amount (thousands) per cluster, and 3) unemployment amount(thousands) per cluster. To show a bigger picture of the unemployment data, we also created a bar graph to display the unemployment amounts for all zip codes. Additionally, users can select a state to focus on a specific state’s zip codes or view the entire United States as a whole. If the user chooses to only view certain cluster assignments (i.e. cluster 2 in California), they can do so by selecting the state and changing the “Cluster Assignment”.

The “Clusters” Only view displays a color coded map of the United States based on its cluster assignment. The “Clusters + Awards” view has an additional layer of information about the award amounts per zip code. The awards amount is represented as a circle overlaid on top of the cluster assignments. The larger the circle, the more the government has awarded aid to this zip code. The circles are also color coordinated, with yellow representing smaller award amounts and red representing larger amounts. In this view, when the user hovers over the circles or zip codes, they will find additional summary statistics such as total income, total tax paid, total amount in real estate, total amount in net investments, etc.

Our goal is to enable the user to explore the clustering assigned by the K-means algorithm and apply human heuristics to find patterns. For example, the user may zoom in on specific zip codes based on their familiarity or proximity. They can see how the clustering compares to other geographically close zip codes or compare it to other, further zip codes that with which they are familiar. **(Innovation 3)** - In this regard, one of our key visualization innovations is to allow users to compare ground-truth data by seeing discrepancies in identified clusters (that may correspond with low-income or high-need areas) compared to actual federal award amounts given.



Example: “Clusters Only” view with “Select # of Clusters” (4)



Example of “Clusters + Awards” view with Select # of Clusters (4)

Experiments/ Evaluation

Our testbed intends to assess how well our solution works to increase aid distribution equality. The experiment aims to determine if our tool can reveal the current status of aid distribution, the shortcomings of current methods, and visualize potential areas requiring aid using our clustering methodology. We also want to gauge user knowledge changes, perceptions of equity, and satisfaction with the tool by combining KPIs, user surveys, interviews, and quantitative and qualitative analyses. Our primary evaluation method for our tool will be user feedback. We have designed the following detailed approach for evaluation.

1. Key performance indicators (KPIs): Baseline KPIs such as changes in the distribution of funds or the number of underfunded areas that receive aid.
2. Test user selection: Select a diverse group of users, including citizens, local government officials, and other stakeholders, ensuring adequate representation for testing.
3. Pre-test survey: Administer a pre-test survey to gather users' baseline knowledge of aid distribution and their perceptions of equity in their localities.
4. Tool training: Provide users with training on how to use the tool, interpret the results from the k-means clustering, and identify patterns among zip codes.
5. User testing phase: Allow users to interact with the tool for a predetermined period, encouraging them to explore federal fund allocation and identify patterns that could improve equity.
6. Post-test survey and interviews: Conduct semi-structured interviews (e.g., through a Google form) with a subset of users to gather in-depth feedback and insights into their experiences.
7. Quantitative analysis: Analyze the KPIs as they change over time to evaluate the tool's impact on users' (i.e., citizens, government officials) ability to identify patterns that could promote equity.

8. Qualitative analysis: Analyze interview transcripts to identify recurring themes, challenges, and suggestions for improvement, providing valuable insights into the user experience and the tool's overall effectiveness.
9. Synthesize findings: Combine the quantitative and qualitative results to assess the tool's ability to achieve its goal and gather user input for future enhancements.

This full process was not feasible during the project timeline. As such, our experimentation and evaluation was limited to internal feedback between our "data team" and "visualization team". We focused on steps **4, 5, 6,** and **8** between both internal teams to evaluate our tool and determine if it addresses questions from the testbed. We noted the following observations from our testing:

1. The clustering algorithm provided an intuitive way to explore trends in the zip code data. By design, it did not inherently tell a user which clusters actually needed aid.
2. The interactivity of the visualization made it easy to identify familiar zip codes, compare them to unfamiliar zip codes, and see what attributes made them similar to be included in the same cluster
3. Areas that were previously identified by federal or state governments as in high need for aid could be used as a baseline to give clusters a meaning. The cluster for these baseline areas could then be filtered to inspect other zip codes that shared the same cluster.
4. Using this methodology, we were able to determine high federal award/assistance areas such as El Segundo, CA (90245, $413M in awards) that appeared to be disproportionately awarded compared to members of the same cluster like Venice, CA (90291, $179M in awards).

Conclusions and discussion

In conclusion, our project used zip code-level data analysis and K-means clustering to create a tool to improve the allocation of government funds in underfunded areas. Our tool displays the data in a consumable format for all users, it can bolster public trust and confidence (Krah & Mertens (12), increase community participation in government decision-making processes and foster a sense of community between citizens and government (New Zealand Department of Internal Affairs(15)), and improves tax administration and compliance (Nicholson-Crotty(16), OECD(17)).

We have proven the usefulness of our tool through self-testing and collective gathering of observations. We also recognize that our strategy has several drawbacks, including the necessity for additional data sources and potential bias in the clustering process. For example, we could add to visualization to increase its effectiveness (a) Cost of Living and (b) Distinction between rural and urban area. Adjusted income for cost of living would allow us to compare state-by-state in an equal plane field while visual representation of urban and rural areas would allow us to see where most of the funding is allocated. Furthermore, improvements can be made to allow the user to select which factor(s) is important to them which can be incorporated as multiple check boxes that can be toggled on and off in Tableau. We have decided against these implementations in this project due to time and resource constraints. Therefore, we advise more studies to solve these drawbacks and broaden the tool's applicability to more fields and funding sources.

Distribution of team effort: All team members contributed a similar amount of effort.

References

1. Baicker, K., & Skinner, J. (2011). Health care spending growth and the future of US tax rates. *Tax Policy and the Economy*, *25*(1), 39-68.
2. Bartik, A. W., Bertrand, M., Cullen, Z. B., Glaeser, E. L., Luca, M., & Stanton, C. T. (2020). Employment Effects of Unemployment Insurance Generosity During the Pandemic. National Bureau of Economic Research. https://tobin.yale.edu/sites/default/files/files/C-19%20Articles/CARES-UI_identification_vF(1).pdf
3. Buchanan, James M. "Federal Grants and Resource Allocation." Journal of Political Economy, vol. 60, no. 3, University of Chicago Press, June 1952, pp. 208–17. https://doi.org/10.1086/257209.
4. Falcettoni, E., & Nygaard, V. M. (2021). A literature review on the impact of increased unemployment insurance benefits and stimulus checks in the United States. *Covid Economics*, *64*, 186-201.
5. Franklin, A., & Ebdon, C. (2004). Aligning priorities in local budgeting processes. *Journal of Public Budgeting, Accounting & Financial Management*.
6. Glomm, G., & Ravikumar, B. (1998). Flat-rate taxes, government spending on education, and growth. *Review of Economic Dynamics*, *1*(1), 306-325.
7. Goldrick-Rab, S., Kelchen, R., & Houle, J. (2014). Expanding Enrollments and Contracting State Budgets: The Effect of the Great Recession on Higher Education. The ANNALS of the American Academy of Political and Social Science, 655(1), 163-180. https://doi.org/10.1177/0002716213500035
8. Hoynes, H. W., & Patel, A. J. (2018). Effective policy for reducing poverty and inequality? The Earned Income Tax Credit and the distribution of income. *Journal of Human Resources*, *53*(4), 859-890.
9. Institute on Taxation and Economic Policy. (2018). Who Pays? A Distributional Analysis of the Tax Systems in All 50 States. American Economic Review, 108(4-5), 1007-1044. https://www.itep.org/wp-content/uploads/whopaysreport.pdf
10. Jolliffe, Dean. "The Cost of Living and the Geographic Distribution of Poverty." Economic Research Report, Jan. 2006, https://doi.org/10.22004/ag.econ.7254.
11. K. P. Sinaga and M. -S. Yang, "Unsupervised K-Means Clustering Algorithm," in *IEEE Access*, vol. 8, pp. 80716-80727, 2020, doi: 10.1109/ACCESS.2020.2988796.
12. Krah, R. D. Y., & Mertens, G. (2020). Transparency in local governments: Patterns and practices of twenty-first century. *State and Local Government Review*, *52*(3), 200-213.
13. Loughead, K. (2020). *State individual income tax rates and brackets for 2020*. Washington, DC: Tax Foundation.
14. Mueller, J. T., McConnell, K., Burow, P. B., Pofahl, K., Merdjanoff, A. A., & Farrell, J. (2021). Impacts of the COVID-19 pandemic on rural America. *Proceedings of the National Academy of Sciences*, *118*(1), 2019378118.
15. New Zealand Department of Internal Affairs. (2018). How digital can support participation in government. Retrieved from https://www.digital.govt.nz/assets/Standards-guidance/Engagement/How-digital-can-support-participation-in-government.pdf

16. Nicholson-Crotty, S. (2008). Fiscal Federalism and Tax Effort in the U.S. States. State Politics & Policy Quarterly, 8(2), 109-126. doi:10.1177/153244000800800201
17. OECD. (2019). Transparency principles for tax policy and administration. Retrieved from https://fiscaltransparency.net/wp-content/uploads/2022/08/Transparency-Principles-for-Tax-Policy-and-Administration-Approved02Aug22-1.pdf
18. Shaefer, H. L. (2010). Identifying key barriers to unemployment insurance for disadvantaged workers in the United States. *Journal of social policy*, *39*(3), 439-460.
19. Tenney, M., & Sieber, R. (2016). Data-driven participation: Algorithms, cities, citizens, and corporate control. *Urban Planning (ISSN: 2183-7635)*, *1*(2), 101-113.
20. Watkins, John. "Quantitative Easing as a Means of Reducing Unemployment: A New Version of Trickle-Down Economics." Journal of Economic Issues, vol. 48, no. 2, Taylor and Francis, Dec. 2014, pp. 431–40. https://doi.org/10.2753/jei0021-3624480217.