US Ignite National Broadband Tool

Empowering Policymakers to Close the Digital Divide

By

Kishor Mannur

Marc Edwards

Shweta Sampath Kumar

Swathi Ganesan

Supervisor: Don Patchell

A Capstone Project

Submitted to the University of Chicago in partial fulfillment
of the requirements for the degree of

Master of Science in Applied Data Science

Division of Physical Sciences

December 2023

Abstract

Existing efforts demonstrate the impact of broadband on communities in the US are limited in approach and impact. US Ignite partners with communities, businesses, local governments, and federal agencies to aid community leaders in securing funding and to effectively manage broadband transformation projects for widespread access and utilization of essential connectivity. The project utilizes a model that predicts the socioeconomic effect an increase in broadband accessibility has on communities and utilizes large language models to generate data and insight driven grant proposals to apply to and secure government funds.

Keywords: broadband, clustering, government funding, XGBoost, explainable AI, large language models

Executive Summary

Broadband is increasingly a requirement for modern life in America, yet many communities lack the infrastructure necessary to provide broadband to their residents. US Ignite partners with these communities to improve broadband access. US Ignite works with community and government leaders to prioritize broadband transformation based on the benefit to the communities they lead. The research described in the paper enables US Ignite to compare socioeconomic metrics based on recent broadband growth (or lack thereof), controlling for demographic factors such as race, age, veteran status, and disability status.

The communities seeing the biggest benefit from broadband are the ones lacking effective broadband in general, defined as “underserved” or “unserved.” These communities have less access to quality jobs, education resources, and key government services because of the digital divide they face as compared to better-served communities. To address the digital divide related to broadband, the Infrastructure Investment and Jobs Act of 2021 allocates about $15 billion in funding to underserved and unserved communities (NTIA, 2022) and in June 2023 the Department of Commerce allocated over $42 billion for universal broadband access by 2030 (The White House, 2023).

Using vast volumes of data sourced from open-access datasets submitted by Internet Service Providers (ISPs) to the FCC every six months, combined with ACS Census Data to support broadband policy and investment decision-making and helping policymakers direct funding and programmatic investments to communities most impacted by the digital divide, US Ignite can demonstrate the socioeconomic benefits of broadband transformation by modeling the impact of an increase in broadband on communities and can utilize large language models to generate data and insight driven grant proposals to apply to and secure government funds.

Table of Contents

List of Figures

List of Tables

Introduction

US Ignite is accelerating the smart city movement and assisting with broadband transformations across the United States through research, public/private partnerships, and analytics. They are a 501(c)(3) research organization based in Washington, DC, with major corporate sponsorships across the telecommunications and technology industries (ProPublica, 2022). This project aims to combine demographic, socioeconomic, and broadband transformation data across the United States to assist US Ignite with educating local policymakers on the benefits of broadband transformation and the resources available to support the transition. We aim to identify the infrastructure needs of small towns and underserved communities to help them access funding grants and resources to improve their broadband connectivity.

Problem Statement

Small and rural towns throughout the United States are confronted with the daunting task of identifying available grants that align with their infrastructure needs, compounded by a lack of resources. Consequently, these communities face difficulties in competing for funds and resources, resulting in missed opportunities to enhance their broadband accessibility and overall infrastructure. Insufficient tools, research, and expertise hinder their ability to present compelling cases and prepare grant applications, further exacerbating the challenge. To address this issue, a comprehensive solution is required to enable small and rural towns to effectively identify their broadband and broadband infrastructure requirements while providing them with the necessary resources to access grants and funding opportunities. This solution should empower these

communities to bridge the digital divide, establish sustainable broadband connectivity, and stimulate economic growth and community development.

Within this context, US Ignite assumes a vital role in the process. As a prominent organization advancing smart communities and next-generation applications, US Ignite possesses the expertise and extensive network necessary to collaborate with small towns in identifying their specific broadband needs, formulating effective strategies, and connecting them with relevant grants and resources. By working closely with these communities, US Ignite can provide valuable assistance in the development of compelling grant applications, conducting comprehensive research, and offering technical support. Through this collaborative effort, small towns can effectively compete and secure the investments required to improve their broadband infrastructure. Ultimately, empowering small towns to identify their broadband infrastructure needs, access essential grants and resources, and leverage the support and expertise provided by organizations like US Ignite is essential in bridging the digital divide and fostering comprehensive development within these communities.

Analysis Goals

To build leaders' confidence in broadband transformation metrics, a trustworthy data-backed tool is needed. Previous phases of this project used clustering to establish groups of peer census tracts based on demographic factors unrelated to broadband, such as race, income, and population density. The project also applied advanced data exploration and machine learning techniques to understand how the model is formed. The analysis established trust by articulating how the clusters were formed and using the clusters to control demographic factors.

Previous iterations identified the differences between well-served and underserved census tracts. It reported broadband access data to define each tract's starting point and subsequent transformation. The defined labels allow users to assess the impact of broadband transformation by comparing changes in socioeconomic metrics for untransformed versus transformed communities across and within the census tract clusters. US Ignite can illustrate how much broadband correlates with the socioeconomic health of each census tract, highlighting patterns and key metrics to help communicate the impact of broadband transformation to community leaders.

This project phase is aiming to take all the previous analyses a step further. The previous models will be run on additional data to further validate the previous results. Once validated, additional confidence is established and will be used to geographically cluster and segment the tracts over time. This will lead to the development of a predictive modeling tool to estimate potential growth with proposed infrastructure investments. With these new results and predictions, the team can bring these results to US Ignite and work with developers to incorporate these findings into their National Broadband Tool dashboard. The dashboard will allow local officials and the US Ignite team to overlay and understand multiple variables at once, all while being periodically updated as new data becomes available. This will give the local officials a chance to understand the effects of policy changes within their neighborhood.

Scope

Due to the impact of COVID-19 on various socioeconomic and demographic factors, there have been delays in updating the data provided by the American Community Survey's Census Bureau. To address this, the project acknowledges the need to work with available data sources that are relevant and up to date. As a result, the analysis focuses on pre-2020 data,

specifically from 2015 and 2019, to ensure a comprehensive understanding of the socioeconomic and demographic landscape.

Being a US-based non-profit organization, US Ignite recognizes the importance of considering data from all 50 US states, including Alaska and Hawaii. This approach ensures that the project's insights and recommendations are inclusive and representative of the entire country. By leveraging data from diverse geographic regions, the project aims to capture the unique characteristics and challenges faced by different communities, facilitating the development of tailored and effective commercial strategies in the emerging smart community market.

Background

US Ignite, as an organization, is highly focused on closing the digital divide by advancing the adoption of communications technologies such as broadband. The digital divide refers to the gap between demographics and regions that have access to modern technology and those that are unserved or underserved (Daley, 1999). While this term previously referenced telephones, televisions, and personal computers, the project focuses on the digital divide specific to reliable broadband internet connectivity in the United States.

Today, broadband is increasingly a requirement to participate in employment opportunities, access necessary education resources, benefit from equitable services (e.g., telemedicine, government, emergency), boost economic growth, and strengthen social connectivity. Broadband communication has direct economic benefits for communities. Based on a study from 2001-2006 across 39 counties in California, as the share of broadband increased by 1%, employment growth increased by 13% (Nazareno & Jose, 2021). The positive economic impact on communities persists beyond the transformation process. Broadband’s educational

impact is apparent as well. 59% of lower-income families with students in primary education tend to face one of three obstacles to completing daily tasks: unreliable internet, no computer, or no smartphone (Vogels, 2020). These communities that struggle to keep up with high-quality jobs and education have challenges with access to services others take for granted. Without the internet, some people did not have the option to fill out the 2020 Census online. Many could not register to vote or request a mail-in ballot, given that some government agencies offering in-person registration were closed to the public or had limited hours due to the pandemic (Sharpton et al., 2020). These communities are most often lower-income rural and urban communities. A 2019 report shows that "approximately 5 million rural American households and 15.3 million urban or metro areas still do not access broadband internet." (Horrigan, 2019).

Broadband is often expensive for companies and communities to deploy and maintain and expensive for customers to buy, especially when they cannot rely upon it consistently. Politics, geography, and technical expertise are all potential impediments to sustainable broadband transformation. Companies often lack sufficient economic incentives to pursue these customers, especially when fiber infrastructure costs between $44,000 and $55,000 per mile (OTELCO, 2018). Often, the biggest challenge is geography. Rivers, mountains, and trees are just some challenges to facilitating broadband for everyone. Building the infrastructure does not make sense for them without some form of incentive or coercion from the government.

Challenges to these transformations can also directly result from politics. The economic benefits of the communities do not align with the interests of political leaders and their major donors, often internet service providers (ISPs). In 2019 and 2020, internet service providers spent almost $235 million on political contributions and lobbying expenditures on issues such as net neutrality and more directly relevant issues such as rural broadband deployment and adoption

(Brodkin, 2021). These political contributions create an environment where it is against legislators' and regulators' political self-interest to hold ISPs accountable for anticompetitive and antitrust behaviors, such as the fight against accurate broadband maps waged by AT&T and other corporations (Brodkin, 2020). These conditions make it nearly impossible for the government to coerce ISPs from connecting remote communities without substantial incentives.

Despite the above-mentioned challenges, there has been significant progress in reducing costs and increasing broadband technologies' reach. Wireless telecommunications companies such as T-Mobile and Verizon have centered recent advertising campaigns around the buildout of their 5G networks (T-Mobile, 2021) (Verizon, 2022). These new networks offer substantially increased speed over short distances, making them ideal for urban deployment. In some of America's largest cities, 5G networks are a viable alternative to wired internet services (Verizon, 2022). Alternatively, the Starlink service deployed by SpaceX is a satellite internet service. SpaceX's approach of deploying a constellation of low-earth-orbit satellites is ideal for providing service to rural and remote communities since no equipment is needed other than a receiver (Starlink, 2022). Starlink internet has proved highly successful and adaptable. During the 2022 Russian invasion of Ukraine, Russian forces disabled internet services to various parts of the country to disrupt Ukrainian communications (Pearson & Satter, 2022). In response, SpaceX deployed Starlink over Ukraine to undermine Russia's tactic and restore internet services in Ukraine, proving Starlink's viability to rapidly deploy broadband to new and remote areas of the world (Massie, 2022).

The Infrastructure Investment and Jobs Act (November 15, 2021) passage is a significant inflection point for efforts to close the digital divide. Of the $1.2 trillion allocated in the new law, Congress earmarked $65 billion for broadband deployment (The White House, 2021). According

to the NTIA, about $15 billion of the funding is available directly to communities via grants (NTIA, 2022). Additionally, in June of 2023, the Department of Commerce announced funding for each state, territory and the District of Columbia for high-speed internet infrastructure deployment through the Broadband Equity Access and Deployment (BEAD) program – a $42.45 billion grant program created in the Bipartisan Infrastructure Law and administered by the Department of Commerce (The White House, 2023). With over 72,000 census tracts in the US potentially vying for funding, policymakers across the country work with US Ignite to access funding and other resources.

Literature Review

Key partners and project objectives are both focused on broadband and telecommunications, exploring the broadband industry's application of machine learning would be tangential to the project objective. Modeling approaches from major telecommunications companies such as Verizon and AT&T focus on solving problems within their enterprise, such as predicting demand for content and detecting security issues. Rather, the project's industry is utility transformation. The research also draws insights from public infrastructure transformation.

The main applications for modeling focus on making metrics available to show utility (broadband) disparities across the United States. The University of Chicago's Data Science Institute has an ongoing Internet Equity Initiative separate from the project described in the paper. The Internet Equity Initiative seeks to compare households with similar internet providers in different neighborhoods and how broadband results can vary among census tracts through their "Netrics" platform (The University of Chicago, 2022). To classify broadband performance, the Netrics platform identifies latency, download and upload speed, bandwidth, and subfactors within census tracts. These factors vary significantly for different neighborhoods using similar

providers. The limitation of simple dashboard approaches is that they do not provide a trustworthy baseline for leaders to compare their communities with other similar communities. Without that base, policymakers are hesitant to make decisions based on the metrics the tool provides.

Since community leaders are naturally skeptical of the data, the project aims to understand demographics by using demographic data to cluster census tracts into peer groups. Other research groups at The Universities of Clemson and South Carolina ran a study summarizing survey statistics on the most significant socioeconomic effects when trying to cluster broadband non-adoption in specific counties within South Carolina (Dickes et al., 2017). They concluded that county classification, age, education, income, and race were all important factors in broadband adoption. Rural settings, ages 70 and older, African Americans, and incomes between $15K and $25K were all examples of certain groups of people with significant characteristics of broadband non-adoption. Building on that research's conclusions, a team from New York University conducted a similar project, also sponsored by US Ignite, to cluster each census tract in the contiguous United States and detect correlations between broadband expansion and economic benefits (Levine et al., 2022).

Decision trees can help validate and improve cluster interpretability. Using decision trees for cluster interpretability is a two-step process. First, cluster the dataset with any clustering algorithm (e.g., K-Means, DBSCAN, Hierarchical Clustering). Then, using the clusters as labels, train a decision tree. Users then interpret the clusters by examining the tree's splits, or decision points (Zuccarelli, 2020). Building from the cluster interpretability concept, MIT developed an all-inclusive approach called Interpretable Clustering for Classification Trees ("ICOT"). ICOT reduces the technique to a one-step process, training the interpretability model to best fit the

clusters (Bertsimas et al., 2018). Using the decision tree, the model generates rule lists to show what decision points apply to each cluster, providing more transparency to community leaders.

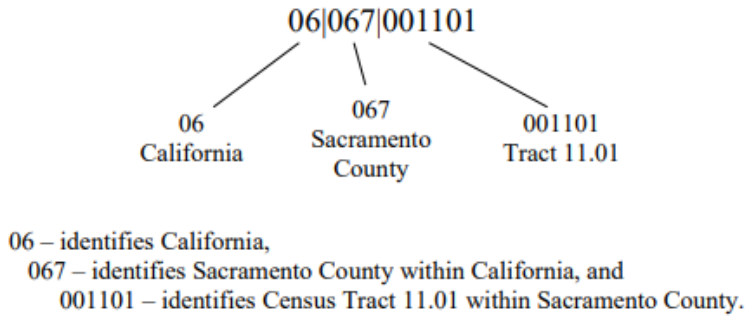
The research also explored approaches to clustering for demographics outside the utility and public infrastructure transformation domains. For instance, marketing and product placement teams can leverage peer census tract clusters to better experiment and scale product placement strategies. Sintec's Business Analytics team tackled the problem of finding and keeping "the right product at the right place on the right shelf at the lowest cost and the least disruption." (Herrera, 2020) Essentially, Sintec tried to solve the product optimization problem where the goal was to find which products would sell more at specific stores within municipalities in Mexico City. In the public health domain, The Academy of Finland studied neighborhood characteristics and effects of behavior-related risk health factors. The study ran interaction tests with neighborhood advantages against socioeconomic factors, such as occupational position (high, intermediate, or low), residence size, smoking, excessive drinking, marriage status, unemployment, population density, and more. The conclusion from the study stated that neighborhood differences could not be explained through socioeconomic factors alone, but neighborhood covariates also need to be incorporated (Halonen et al., 2021). As a result, the project incorporated additional factors such as housing costs and population density.

Data

The project uses data from the American Community Survey specifically demographics for clustering and income, jobs, and education for socioeconomic benchmarking from 2017 through 2021 (US Census, 2022) as well as 5-year Summary Files from 2019; Equivalent Summary Files from 2015 also support the benchmarking. To augment the demographic data, the project also includes land area data from the Census Bureau's 2019 Planning Database. The

project also leverages data from the FCC Form 477 to identify broadband transformation (or lack thereof) across each census tract from 2017 – 2021, and data from June 2015 through June 2019 to create labels for later in the analysis. There are over 84,000 census tracts across the fifty US states and sixteen territories. Figure 1 below illustrates how census tracts are labeled (FCC, 2020).

Figure 1. *Diagram of Census Tracts Identifier, Using Census Tract 11.01 within Sacramento County, California as an example*



Data Sources

American Community Survey (ACS)

The primary data source for demographics and benchmarking was the American Community Survey 5-year Summary Files from 2019. Equivalent Summary Files from 2015 also support benchmarking. The 5-year Summary Files "contain all of the Detailed Tables for the ACS data releases." US Census Detailed Tables state that "contain the most detailed cross-tabulations, many of which are published down to block groups. The data are population counts." ACS also describes the 5-year data as "estimates that represent data collected over a period of time" (US Census, 2022). The primary advantage of using multiyear estimates is the increased statistical reliability of the data for less populated areas and small population subgroups. The

increased reliability is welcome since the data is analyzed at the census tract level. Based on this information, the 2015 5-year Summary File represents 2011-2015, and the 2019 Summary File represents 2015-2019. ACS provides guidance not to use overlapping estimate ranges but given the target window (2015-2019) and target granularity (census tract), the project accepts the overlapping year of 2015. The other option was to shift to 1-year data, but 1-year data is limited to counties and places with populations greater than 65,000 (US Census, 2022).

From the ACS Summary Files, the research sourced over 60,000 variables reported across 84,414 census tracts, selecting and filtering variables relevant to the demographic and socioeconomic factors our project targets. Selected variables included race, ethnicity, population, population density, migration, jobs, education, and income, aligning with the factors identified during the literature review. In most cases, each variable represented a cross-section of one or more facts, which mapped back to over 100 top-line variables. ACS reports each variable as a count of people or households. The research included each relevant top-line variable and their relevant groups, using 115 variables from the data source in demographic clustering and 40 variables from the data source in socioeconomic benchmarking (see Appendix A for variable information)

Census Planning Database

While the ACS data source contained most of the variables used for demographic modeling; it did not contain population density information. Population density information is crucial for analysis because it allows the model to incorporate population scaled by one of the most important determining factors for broadband strategy, cost, and success: land area. Since census tracts are not uniform in geographic size, the “area of population” features does not definitively tell the population density of a community. “Area of population” reports the

population of the larger statistical area the census tract is in but divides it into three features based on whether the census tract is part of a metropolitan area, micropolitan area, or other (likely rural area) (US Census, 2021). To include population density in clustering, the research included separate data from the 2019 Census Tract Planning Database, also provided by the US Census Bureau (US Census, 2021). The population density feature incorporates the population, divided by the land area for each census tract, in square miles, to include in the model.

FCC Form 477

The project uses FCC Form 477 data from June 2015 through June 2019 to create these labels, data from 2017 through 2021 was also utilized during the modelling phase of the project. The FCC website states, "All facilities-based broadband providers are required to file data with the FCC twice a year (Form 477) on where they offer Internet access service at speeds exceeding 200 kbps in at least one direction." (FCC, 2022) The FCC aggregates this data on its website to create maps and analyze changes in broadband availability over time. The FCC broadband subscribership data excludes all connections identified as business connections and necessarily excludes residential mobile wireless connections (which are reported for the state but not for individual census tracts).

Internet service providers (ISPs) report through FCC Form 477, for each census tract, where they offered speeds that meet or exceed 10 Mbps download speed and 1 Mbps upload speed (10/1). This standard is a limitation because the 10/1 measure in this data set is outdated for adequate internet speed. The latest FCC guidance defines a "broadband" internet connection as one that provides at least 25 Mbps for download speed and 3 Mbps for upload, and there is a push by the FCC chair to push this standard even higher to 100/20 Mbps (Velazco, 2022).

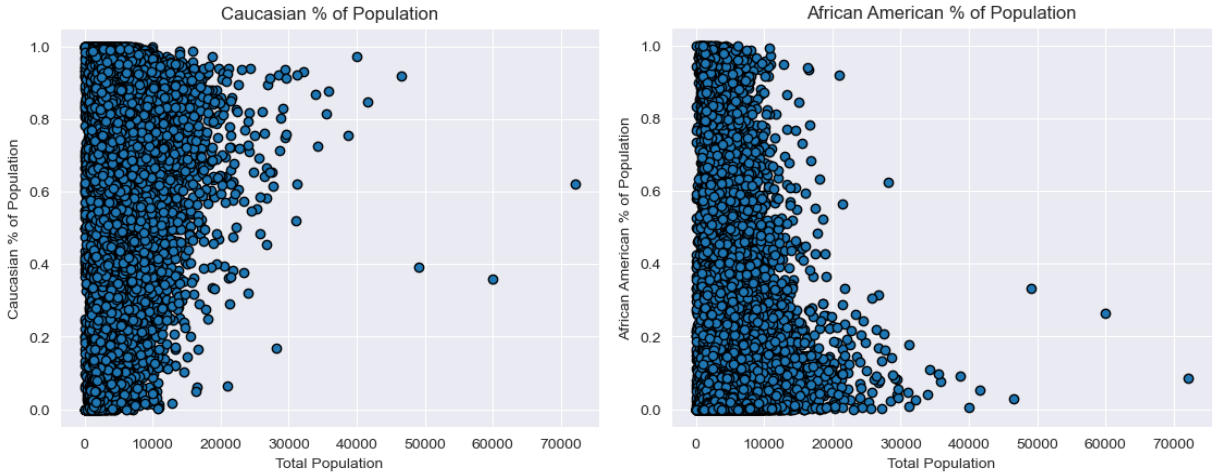
The project uses residential fixed broadband subscribership as a proxy measure for public investment in broadband transformation. While increased broadband investment aims to increase broadband adoption, the two measures are not perfectly correlated, as there is a delayed impact expected from infrastructure investment, including broadband. Ideally, actual infrastructure investment dollars would be used as the benchmark measure, but this data is not publicly available or easy to source for all census tracts.

Descriptive Analysis

American Community Survey (ACS)

The demographic data from ACS shows a normal distribution of population across census tracts with outliers having higher than normal population totals. White or Caucasian populations in census tracts are left-skewed with many tracts at 100% white population, while non-white populations are right-skewed. The 2020 Census data shows that non-Hispanic whites are the racial and ethnic majority, followed by Hispanic and Latino Americans and Black or African Americans. The negative correlation between Caucasian and African American populations is -79%, while Hispanic populations have a high positive correlation of 72% with populations reported as Other. This suggests a prevalence of multiple races and ethnicities in the communities at least two of three types of communities: White, Black, and Diverse. Figures 2 and 3 below show these dynamics.

Figures 2 and 3. *Scatter Plots of Caucasian and African American Populations, reported by count within census tract vs percentage of total census tract population.*

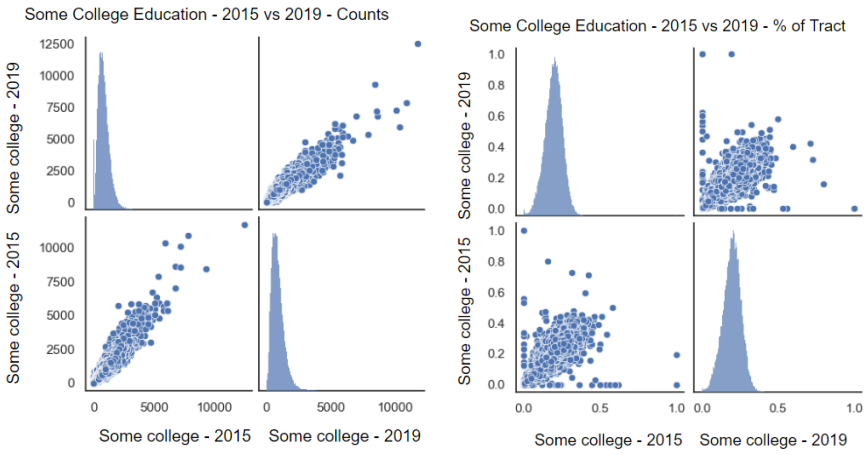


In terms of socioeconomic benchmarks, the raw counts of each selected variable displayed a right-skewed distribution. However, there was a strong positive correlation between the 2015 and 2019 results for almost every feature, which is consistent with the population growth during the same period. When it came to median income, approximately 40% of the census tracts did not report it, resulting in its exclusion from the benchmarking metrics instead of imputing values.

We proceeded to scale the metrics for the remaining variables according to the population of each census tract that was obtained from the FCC census tract data. While some right skew remained, especially for minority class variables such as higher income, immigration, higher educational attainment, and non-private work class, the normalization of distributions across the board was possible through population adjustment. It was observed that people who did not immigrate to the census tract within the past year were also right skewed, implying that many communities remained largely the same over the 2015-2019 period. Furthermore, poverty and school enrollment data peaked at zero, indicating that several census tracts had no poverty or no populations of children. This gives us the motive to dive deeper into data from 2016 through 2018.

Despite these observations, there still exists a strong positive correlation between most variables' 2015 and 2019 results. As a result, the previous implementation of the Broadband tool concludes that population-adjusted growth is a far better metric than raw growth, a conclusion that is verified and supported by the changes in the distribution. Based on this finding, the socioeconomic benchmarking will include these population-adjusted growth metrics. Figures 4 and 5 demonstrate the impact of population adjustment on residents with some college education but have not attained a bachelor's degree or higher.

Figures 4 and 5. *2015 and 2019 - Some College - Before and After Population Adjustment*



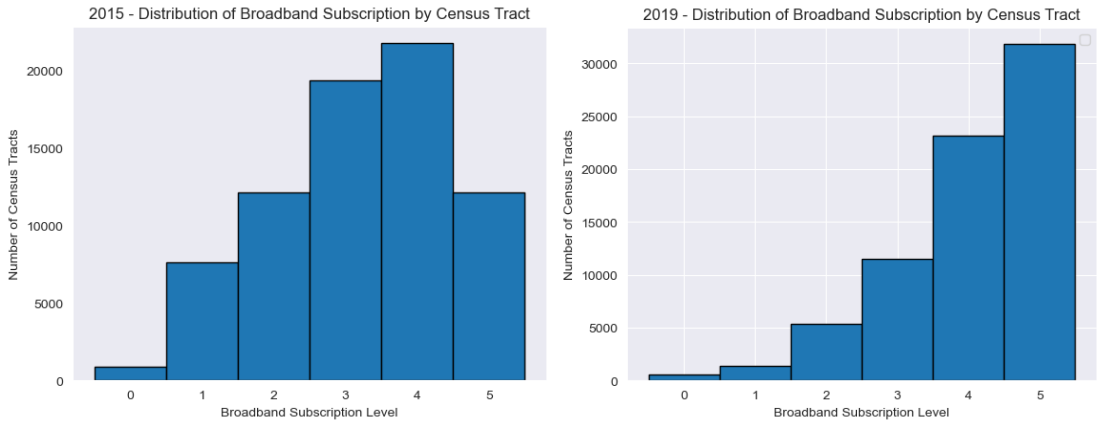
Census Planning Database

In 2019, the US had 360 census tracts with no population. The project excludes empty census tracts from the remaining analysis. Both land area and population density (population adjusted for the geographic size of the tract) are right-skewed. Only certain rural census tracts contain large areas of unpopulated land, and most other census tracts are comparatively much smaller overall in land area. The mean census tract is about 47.8 square miles, but the median census tract is only about 1.8 square miles.

FCC Form 477

When exploring the FCC Form 477 data, the aim is to understand the overall transformation trends across census tracts. In 2015, the average census tract had 40-60% of households meeting the 10/1 standard. By 2019, the average shifted to the next highest band, with the average census track having 60-80% of households meeting the same standard. From 2015 to 2019, the number of census tracts with 80% or more households meeting the 10/1 standard doubled. Figures 6 and 7 below illustrate the adoption rates in 2015 and 2019.

Figures 6 and 7. *Census Tract Broadband Adoption by Level based on 10/1 standard, 2015 and 2019, respectively.*



Analyzing this transformation by state, states with less broadband adoption in 2015 continue to lag based on 2019 data. However, states such as Montana, New Mexico, and Kentucky ranked towards the bottom in 2015 but showed strong growth based on 2019 data. Conversely, states with solid adoption as of 2015, such as Rhode Island, Delaware, and Hawaii, were among the lowest in broadband growth based on the analyzed data. Table 1 below details top and bottom states for 2015 and 2019 adoption rates, as well as the growth rate over the same period.

Table 1. *Broadband Adoption by State based on 10-down/1-up standard.*

Top States by 10-down/1-up	2015 Adoption	2019 Adoption	2015-2019 Growth
Best	1. Rhode Island 2. New Jersey 3. Delaware 4. Hawaii 5. Massachusetts	6. Rhode Island 7. New Jersey 8. New Hampshire 9. Massachusetts 10. Connecticut	11. Montana 12. Kentucky 13. Maine 14. Utah 15. New Mexico
Worst	16. Iowa 17. Montana 18. Idaho 19. Mississippi 20. New Mexico	21. Oklahoma 22. Idaho 23. New Mexico 24. Mississippi 25. Alaska	26. Hawaii 27. Rhode Island 28. Delaware 29. Alaska 30. New York

The findings suggest broadband growth according to the 10/1 standard does not necessarily correlate with broadband adoption. Instead, growth is influenced heavily by the baseline broadband adoption rate in 2015. Since the project aims to benchmark peer census tracts based on their transformation, the script filters census tracts that had already achieved 80-100% broadband adoption at the 10/1 standard based on 2015 data. By removing these census tracts, the script prevents census tracts that are already "post-transformation" from being labeled as "no growth" or "moderate growth" and skewing the benchmarking.

Methodology

Feature Engineering

ACS 5-year Subject Files for each state were downloaded and census-tract level variables were extracted, which were then transformed into data frames and combined for all 50 states. To adjust for population, raw growth was computed and adjusted for population for benchmarking variables. Benchmarking variables were selected based on background research and converted

race/ethnicity and benchmarking variables to a percentage of the total population to compare census tracts with different populations. To avoid convergence effects, the K-Means model was used to bin each demographic variable into three bins (i.e., low, medium, and high). Housing-related features were converted to three buckets using a weighted mean and K-Means binning approach. The broadband growth was stratified into four distinct levels (negative, zero, moderate, and significant) by calculating the growth for each tract and labeling census tracts with negative and zero growth based on this growth score. Similarly, feature engineering will be applied for additional years of data to validate the clusters created previously.

Modeling Frameworks

The modeling framework in the previous study involved clustering demographic data to generate peer census tracts for the analysis of the socioeconomic impact of broadband. Initially, Agglomerative Clustering was explored, but due to the large number of census tracts, it was not feasible to analyze the splits and set a cutoff point to define the clusters. Therefore, DBSCAN was used to cluster the data into groups, which group data points based on distance from their neighbors and the number of neighbors. K-Means was also explored to minimize outliers and partition the dataset into a specified number of groups.

The remaining implementation focused on optimizing the clusters based on a few evaluation metrics, such as minimizing outliers, keeping the number of clusters manageable, and maximizing the training data fit. The project used the ICOT method for the validation process, which trained a decision tree classifier on the clustering results to establish rule lists that explain how the algorithm formed each cluster.

The project explored modifying other default parameters, such as setting the minimum number of samples, applying PCA, using Manhattan distance, and using K-Nearest Neighbors to

find the optimal value for EPS. Ultimately, the script sets EPS to 6 to create the optimal number of clusters.

The current project will further explore different clustering methods to try including Agglomerative Clustering, DBSCAN, and K-Means while fine-tuning which can involve modifying default parameters such as PCA, distance metrics, and K-Nearest Neighbors. We will be comparing and evaluating the models using the distance between clusters (Manhattan, Euclidean, etc.), the number of clusters, the number of outliers within the clusters.

In addition to trying different clustering algorithms and fine-tuning the parameters, there are several other techniques that will be explored to improve the quality of clustering. First, the most relevant features were selected for clustering, which can help improve the quality of clustering. This can be achieved by using methods like Principal Component Analysis (PCA), which reduces the number of features while retaining the most relevant information. Second, normalizing the data is useful in cases where the range of values across different features is significantly different and can help avoid situations where certain features are given more weight than others. Third, combining the results of different clustering algorithms can help improve the overall quality of clustering which is done by either using a majority voting approach or using a weighted approach where the weights are assigned based on the performance of each clustering algorithm. Next, identifying and removing outliers from the data can improve the quality of clustering. This can be achieved by using methods like DBSCAN, which can detect outliers as noise points. Finally, fine-tuning the hyperparameters of clustering algorithms can significantly improve their performance, which include the number of clusters, the distance metric, and the similarity measure.

These techniques will be used in combination or individually to improve the quality of clustering. After iterating various model parameters to identify models that best represent each community group while achieving a balance between accuracy and succinctness in their descriptions, we will validate the model by evaluating its ability to best represent the broadband influence on communities over time.

Findings

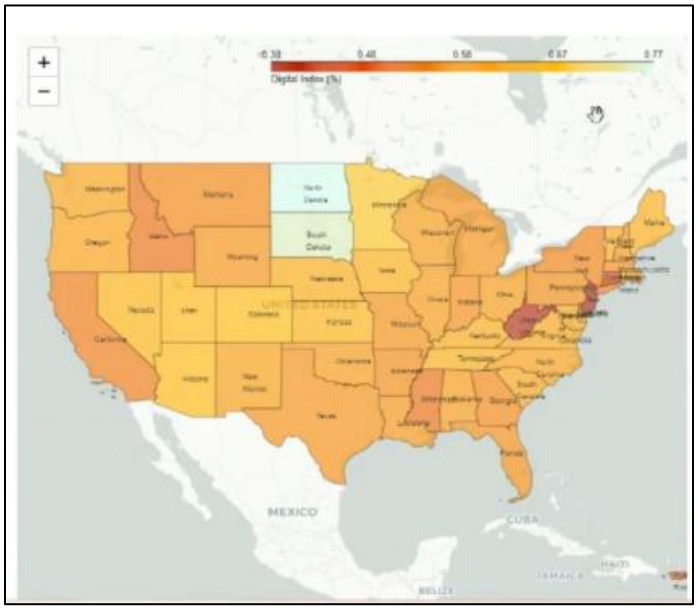
During this research, an extensive exploration of various clustering algorithms, including Agglomerative Clustering, DBSCAN, and K-Means was conducted. This exploration involved a critical evaluation and fine-tuning of default parameters such as PCA, distance metrics, and K-Nearest Neighbors, alongside the implementation of advanced techniques like feature selection, normalization, ensemble clustering, outlier detection, and hyperparameter tuning. These methods significantly enhanced the quality of clustering, crucial for identifying communities for targeted broadband infrastructure enhancements.

The primary output of this project includes a sophisticated prediction model and an intuitive dashboard. These tools are designed to not only identify but also estimate the growth trajectory of communities in need of infrastructural investments. By establishing a benchmark based on the performance of digitally well-connected communities, a standard for assessing the development potential of emerging communities has been created.

The dashboard visualizations (Figures 1 and 2) present compelling evidence of the transformative power of broadband access, showing its correlation with improvements in socioeconomic conditions and education over time. These visualizations are instrumental in

highlighting the potential for broadband access to enhance quality of life and opportunities in key areas such as education, healthcare, employment, and social interaction.

Figures 8. *Digital Index Scores Map of the United States*



Figures 9. *Comparison of Digital Index and Socioeconomic Factors in Similar Tracts*

tract_code	county_name	connections_above25	income_score	employment_score	education_score
04013612700	Maricopa County	1	94.9	48.04	56.66
19079960500	Hamilton County	1	76	51.16	45.27
27111960500	Otter Tail County	1	70.82	49.14	48.09
35001000112	Bernalillo County	1	78.87	50.29	51.51
38067950500	Pembina County	1	75.35	50.14	47.17

The following table provides the succinct snapshot of a 20% increase in digital index score within rural tracts and how the increase affects the socioeconomic scores across education, employment, and income. The results come from the XGBoost Regression with Random Search using Cross Validation. The predictors in this model included Demographic (Gender, Age,

Veteran Status, Disability), 2 Yr. Lagged Dependent Variables for Causality (Education, Employment, Income) , % of broadband connections at least 25 mbps, percent of broadband connections less than 25 mbps, and percent type of connection (Broadband, Other, No internet). The response variables were the scores of Education, Employment, and Income.

Table 2. *Rural Socioeconomic Scores with a 20% Increase in Digital Index*

Socioeconomic Factor	Percent Change with 20% Increase in Digital Index	XGBoost Model Accuracy and Reliability (RMSE)
Education	4.9%	0.066538
Employment	0.7%	0.015154
Income	7.6%	0.036163

Additionally, through the strategic use of our clustering outputs, our project is positioned to perform effective community segmentation. This allows for the identification of communities with similar characteristics to those in past data, thereby enhancing the planning, application, and allocation of resources based on proven success models.

Discussion

In furthering the modeling framework for identifying areas with broadband usage for US Ignite, the project will investigate additional clustering algorithms, including Agglomerative Clustering, DBSCAN, and K-Means, while also fine-tuning default parameters such as PCA, distance metrics, and K-Nearest Neighbors. To improve the quality of clustering, the project will employ various techniques such as feature selection, normalization, ensemble clustering, outlier detection, and hyperparameter tuning.

Ensemble clustering, which combines multiple clustering algorithms to achieve a better clustering result, can provide a higher level of accuracy than single clustering algorithms. Feature selection will be used to identify the most significant variables and reduce the dimensionality of the data. Normalization techniques will be employed to transform variables into a common scale to prevent variables with larger values from dominating the analysis. Outlier detection will also be applied to remove noise and improve the clustering quality. Finally, hyperparameter tuning will optimize the parameters of each clustering algorithm to improve performance. The goal of exploring these methods is to achieve a higher quality of clustering that can better identify areas with broadband usage for US Ignite. This will provide more accurate information to community leaders to help them make informed decisions about broadband infrastructure investments. The results of this project will provide valuable insights into the performance of different clustering algorithms and techniques for broadband mapping, which can be applied in similar contexts.

Conclusion

In conclusion, a modeling framework that combines demographic clustering with broadband and socioeconomic metrics to identify areas with low broadband usage in the US is proposed. The framework uses unsupervised learning techniques, such as clustering algorithms and dimensionality reduction, to identify patterns and relationships between variables. The evaluation metrics used to validate the models are the silhouette score and the ICOT rule lists, which provide a balance between accuracy and interpretability.

The analysis and modelling find that the DBSCAN clustering algorithm, combined with feature scaling, outlier detection, and hyperparameter tuning, performs the best for the project's use case. While the use of PCA initially seemed to improve the model's performance, the

validation method revealed that it decreased interpretability and was not necessary for accurate clustering.

The analysis shows that areas with broadband growth experience significant improvements in education and income metrics, while areas without broadband growth see a decline in these metrics. These findings can inform policymakers and community leaders about the benefits of expanding broadband infrastructure and increasing access to technology.

The development of the National Broadband Tool Dashboard combines the results of the analysis and visually represents the areas of broadband need in the United States. Being able to find and recognize these areas is the first step in getting them the funding they need to increase their broadband speed and infrastructure. The next step in acquiring grant funds is creating a proposal, backed by data and insights from the dashboard. Data from the dashboard can be uploaded to a GPT that is trained, prompted, and designed to help public officials and homeowners petition the government for funds from the Broadband Provisions of the 2021 Infrastructure Investment and Jobs Act and the Department of Commerce's announced funds from June 2023.

In future research, the modeling framework can be further improved by exploring different clustering algorithms and fine-tuning the default parameters. The model can incorporate other variables, such as health outcomes or environmental factors, to provide a more comprehensive understanding of the impact of broadband access on communities. Additional opportunities also lie within collating enhanced broadband data for the model by refining the identification of Broadband Serviceable Locations (BSL) by distinguishing between served, underserved, and unserved areas using data from additional partner states, federal agencies, industry, and accessible commercial datasets. Next, a fully automated community report

generation tool can be created and incorporated into the dashboard to automate the end-to-end grant proposal generation by creating community reports; offering proposed broadband and socio-economic metrics to bridge digital equity gaps. Finally, planning for the sustainability of the tool can be attained by exploring business models or funding strategies to ensure the long-term success and use of the tool, which includes potential partnerships with broadband providers.

Overall, the study demonstrates the potential of data-driven approaches to inform policies and investments that promote equity and opportunity for all.

References

Barrero, J. M., Bloom, N., & Davis, S. J. (2021). *Why Working From Home Will Stick*. National Bureau of Economic Research. https://www.nber.org/system/files/working_papers/w28731/w28731.pdf

Bertsimas, D., Orfanoudaki, A., & Wiberg, H. (2018). *Interpretable Clustering via Optimal Trees*. NeurIPS 2018. https://arxiv.org/pdf/1812.00539.pdf

Brodkin, J. (2021). *ISPs spent $235 million on lobbying and donations, "more than $320,000 a day"*. Ars Technica. https://arstechnica.com/tech-policy/2021/07/isps-spent-235-million-on-lobbying-and-donations-more-than-320000-a-day/

Brodkin, J. (2020). *AT&T hopes you'll forget its years-long fight against accurate broadband maps.* Ars Technica. https://arstechnica.com/tech-policy/2020/09/att-hopes-youll-forget-its-years-long-fight-against-accurate-broadband-maps/

Dabbaru, I. (2018*). K-means Clustering: Algorithm, Applications, Evaluation Methods, and Drawbacks.* Towards Data Science. https://towardsdatascience.com/k-means-clustering-algorithm-applications-evaluation-methods-and-drawbacks-aa03e644b48a

Daley, W. M. (1999*). Falling Through the Net: Letter From William M. Daley*. National Telecommunications and Information Administration. https://www.ntia.doc.gov/legacy/ntiahome/fttn99/daley.html

Dickes, L., Crouch, E., & Walker, T. (2017). *Socioeconomic determinants of broadband non-adoption among consumer households in South Carolina, USA.* Redalyc.org. https://doi.org/10.4422/ager.2018.17

do Prado, K. S. (2018). *How DBSCAN works and why should we use it?* Towards Data Science. https://towardsdatascience.com/how-dbscan-works-and-why-should-i-use-it-443b4a191c80

FCC (2020). *More About Census Tracts*. https://transition.fcc.gov/form477/Geo/more_about_census_tracts.pdf

FCC (2022). *Fixed Broadband Deployment Data from FCC Form 477.* Federal Communications Commission. https://www.fcc.gov/general/broadband-deployment-data-fcc-form-477

Herrera, H. (2020*). Store Clustering: Applying socio-demographic data to create clusters.* LinkedIn. https://www.linkedin.com/pulse/store-clustering-applying-socio-demographic-data-create-hugo-herrera/?trk=public_profile_article_view

Horrigan, J. B. (2019). *Analysis: Digital Divide Isn't Just a Rural Problem.* The Daily Yonder. https://dailyyonder.com/analysis-digital-divide-isnt-just-a-rural-problem/2019/08/14/

ITC Holdings (2022). *Power Grid History*. https://www.itc-holdings.com/a-modern-power-grid/power-grid-history

Levine, C. (2018). *POLL: Paramus Mayor Won't Change The Blue Laws*. Daily Voice. https://dailyvoice.com/new-jersey/paramus/politics/poll-paramus-mayor-wont-change-the-blue-laws/731965/

Levine, D., Magid, M., Zhang, K., & Yan, Z. (2022). *Mapping connections between broadband expansion and equitable economic prosperity.* GitHub. https://github.com/dlevine01/broadband-economy

Lilly, C. M., MD, Cody, S., MSN/MBA, & Zhao, H., PhD (2011). *Hospital Mortality, Length of Stay, and Preventable Complications Among Critically Ill Patients Before and After Tele-ICU Reengineering of Critical Care Processes*. JAMA. https://doi.org/10.1001/jama.2011.697

Maklin, C. (2018). *Hierarchical Agglomerative Clustering Algorithm Example In Python.* Towards Data Science. https://towardsdatascience.com/machine-learning-algorithms-part-12-hierarchical-agglomerative-clustering-example-in-python-1e18e0075019

Massie, G. (2022). *Elon Musk helps Ukraine with SpaceX's Starlink satellites.* The Independent. https://www.independent.co.uk/news/world/europe/elon-musk-helps-ukraine-satellites-b2024893.html

McKinsey & Company (2022). *Americans are embracing flexible work-and they want more of it.* https://www.mckinsey.com/industries/real-estate/our-insights/americans-are-embracing-flexible-work-and-they-want-more-of-it

Nazareno, L., & Jose, J. (2021). *The Effects of Broadband Deployment in Rural Areas: Evaluating the Connect America Fund Program*. SSRN. https://doi.org/https://ssrn.com/abstract=3897867

NRECA (2022). *History. National Rural Electric Cooperative Association.* https://www.electric.coop/our-organization/history

NTIA (2022). *Notice of Funding Opportunity - Broadband Equity, Access, and Deployment Program.* https://broadbandusa.ntia.doc.gov/sites/default/files/2022-05/BEAD%20NOFO.pdf

North, A. (2020). *Hybrid school might be the worst of both worlds.* Vox. https://www.vox.com/21515864/covid-hybrid-school-learning-remote-plan-pandemic

OTELCO (2018). *High Speed Fiber Infrastructure Where, when, why, and how.* GoNetspeed. https://www.otelco.com/fiber-infrastructure/

Pearson, J., & Satter, R. (2022). *Internet in Ukraine disrupted as Russian troops advance.* Reuters. https://www.reuters.com/world/europe/internet-ukraine-disrupted-russian-troops-advance-2022-02-26/

Powers, J. (2022). *A Step-by-Step Explanation of Principal Component Analysis (PCA).* Built In. https://builtin.com/data-science/step-step-explanation-principal-component-analysis

ProPublica (2022). *US Ignite Inc*. Nonprofit Explorer - ProPublica. https://projects.propublica.org/nonprofits/organizations/453943413

Sharpton, A., Rev., Starks, G., Gupta, V., Morial, M., & Coley, M. (2020). *Broadband Access Is a Civil Right We Can't Afford To Lose-But Many Can't Afford To Have.* Essence. https://www.essence.com/news/broadband-access-is-a-civil-right-we-cant-afford-to-lose-but-many-cant-afford-to-have/

Singer, H. J., & West, J. D. (2010). *Economic Effects of Broadband Infrastructure Deployment and Tax Incentives for Broadband Deployment*. Neoconnect. https://neoconnect.us/wp-content/uploads/2015/09/Economic_Effects_of_FTTH.pdf

Speed Matters (n.d.). *Economic Growth & Quality Jobs.* SpeedMatters.org. https://speedmatters.org/economicgrowthqualityjobs

Starlink (2022). *Satellites*. https://www.starlink.com/

Steiner, C. (2010). *Wall Street's Speed War*. Forbes. https://www.forbes.com/forbes/2010/0927/outfront-netscape-jim-barksdale-daniel-spivey-wall-street-speed-war.html?sh=7ce8c117741a

Szabo, F. E., PhD (2015). *Manhattan Distance.* Science Direct. https://www.sciencedirect.com/topics/mathematics/manhattan-distance

T-Mobile. (2021, January 4). *Taking Our 5G Network to the Next Level [Video]*. YouTube. https://www.youtube.com/watch?v=_VvxMZJn7MY

The University of Chicago (2022). *A Tale of Two Gigs.* The University of Chicago Data Science Institute. https://internetequity.uchicago.edu/data-story/a-tale-of-two-gigs/

The White House (2021). *President Biden's Bipartisan Infrastructure Law.* https://www.whitehouse.gov/bipartisan-infrastructure-law/

The White House (2023). *Fact Sheet: Biden-Harris Administration Announces Over $40 Billion to Connect Everyone in America to Affordable, Reliable, High-Speed Internet.* https://www.whitehouse.gov/briefing-room/statements-releases/2023/06/26/fact-sheet-biden-harris-administration-announces-over-40-billion-to-connect-everyone-in-america-to-affordable-reliable-high-speed-internet/

US Census (2022). *American Community Survey 5-Year Data (2009-2020).* United States Census Bureau. https://www.census.gov/data/developers/data-sets/acs-5year.html

US Census (2021). *ACS Summary File Sequence-Based Format (2005-2021).* United States Census Bureau. https://www.census.gov/programs-surveys/acs/data/summary-file/sequence-based.html

US Census (2022). *American Community Survey 1-Year Data (2005-2021).* United States Census Bureau. https://www.census.gov/data/developers/data-sets/acs-1year.html

US Census (2021). *Metropolitan and Micropolitan - About.* United States Census Bureau. https://www.census.gov/programs-surveys/metro-micro/about.html

US Census (2021). *Planning Database.* United States Census Bureau. https://www.census.gov/topics/research/guidance/planning-databases.2019.html#list-tab-Y441DKRP0EE0GCAQFR

US Census (2021). *Racial and Ethnic Diversity in the United States: 2010 Census and 2020 Census.* United States Census Bureau. https://www.census.gov/library/visualizations/interactive/racial-and-ethnic-diversity-in-the-united-states-2010-and-2020-census.html

Velazco, C. (2022). *FCC calls 25 Mbps 'broadband' speed. The push is on to up it to 100.* Washington Post. https://www.washingtonpost.com/technology/2022/07/19/fcc-broadband-new-definition-100mbps/

Verizon (2022). *Verizon 5G Ultra Wideband Super Bowl 2022 TV Spot, 'Going Ultra'*. iSpot.tv. https://www.ispot.tv/ad/qiWW/verizon-5g-ultra-wideband-super-bowl-2022-going-ultra

Verizon (2022). *5G Home.* https://www.verizon.com/5g/home/?CMP=crm_h_p_lob_dm_acq_2022_99_gm_5ghome

Vogels, E. A. (2020). *59% of U.S. parents with lower incomes say their child may face digital obstacles in schoolwork.* Pew Research Center. https://www.pewresearch.org/fact-tank/2020/09/10/59-of-u-s-parents-with-lower-incomes-say-their-child-may-face-digital-obstacles-in-schoolwork/

Vogels, E. A. (2021). *Some digital divides persist between rural, urban and suburban America.* Pew Research Center. https://www.pewresearch.org/fact-tank/2021/08/19/some-digital-divides-persist-between-rural-urban-and-suburban-america/

Zimiles, A. (2020). *Four new statistics that prove that telemedicine isn't just a pandemic fad.* Medical Economics. https://www.medicaleconomics.com/view/four-new-statistics-that-prove-that-telemedicine-isn-t-just-a-pandemic-fad

Zuccarelli, E. (2020). *Interpretable Clustering*. Towards Data Science. https://towardsdatascience.com/interpretable-clustering-39b120f95a45

Appendix A: List of American Community Survey Variables Used

Table A1. *List of variables used in the project from the American Community Survey 2015 and 2019 5-year Subject Tables, including definitions and groupings provided in the reporting.*

Purpose		Variable Category	
Demographic Clustering		Total Population	
	Population by Gender	Number of people by gender	Male, Female
	Population by Race/Ethnicity	Number of people by reported race/ethnicity	White, Black, Native American, Asian, Hawaiian, Other
	Population by Hispanic Origin	Number of people by reported Hispanic origin	Hispanic, Not Hispanic
	Population by "area of population"	Number of people by "area of population" - whether the census tract is part of a metropolitan or micropolitan area	Metropolitan, Micropolitan, Other
	Housing Units by Ownership Age	Number of housing units by year when the most recent owner took ownership	1989 or earlier, 1990s, 2000s, 2010-2014, 2015-2016, 2017 or later
	Rental units by move-in	Number of rental units by year when the most recent renter moved in	1989 or earlier, 1990s, 2000s, 2010-2014, 2015-2016, 2017 or later
	Mortgages by Monthly Cost	Number of mortgages by monthly cost in US dollars	Less than $200, then $100 increments to $999, then $250 increments to $1500, then $500 increments to $3999, over $4000
	Rent by Monthly Cost	Number of rental units by monthly cost in US dollars	Less than $100, then $50 increments to $399, then $100 increments to $1499, over $1500
	Mortgage Tax Dollars	Aggregate real estate taxes paid for units with a mortgage	N/A
Socioeconomic Benchmarks		Total Population	
	Population Ratio to Poverty	Number of people by ratio of income to poverty line	Under 1, 1-1.99, 2 and over
	Educational Attainment by Level	Number of people by highest level of schooling achieved	Less than high school (HS), HS grad, some college/associate degree, bachelor's degree, Graduate or Professional degree
	Income by Amount	Number of people by annual income range in US dollars	None, Under $15K, $10-$15K, $15-$25K, $25-$35K, $35-$50K, $50-$65K, $65-$75K, $75K+
	Immigration within Past Twelve Months	Number of people by whether/where they moved from in the past twelve months	Did not move, moved from within the same county, moved from within the same state, moved from another state, moved from another country
	Work class	Number of people by the class of work they do	Work for private company, corporate business owner, work for not-for-profit, work for local gov, state gov, federal gov, unincorporated business owner, family/stay-at-home unpaid work
	School Enrollment by Level	Number of students enrolled by grade levels	Nursery, kindergarten, 1-4, 5-8, 9-12, undergrad, graduate/professional degree