DEEP DIVE NEIGHBORHOOD-LEVEL ANALYSIS OF 12 FLORIDA COUNTIES IMPACTED BY CENSUS UNDERCOUNT

Florida Philanthropic Network

Florida Data Science for Social Good (FL-DSSG)

University of North Florida

Jacksonville, FL

CONTENTS

Executive Summary

The U.S. Census Bureau conducts a nationwide count every decade to estimate the population and demographic characteristics of the United States. These estimates are crucial, as they guide the distribution of federal funding to each state. An undercount in any state, including Florida, could limit the resources needed to address important social and economic issues.

The **Florida Philanthropic Network (FPN)** has been working to understand and address the risk of a potential undercount in Florida for the upcoming 2030 census. Leveraging data from **Florida Tax Watch**, **U.S. Census Bureau**, **Opportunity Atlas**, **Florida voter registration and participation**, **Florida Nonprofit Alliance**, and the **Texas Census Institute's** methodology, the **Florida Data Science for Social Good (FL-DSSG)** has estimated undercount risks for selected counties in Florida.

The analysis focuses on three key dimensions: **Personal**, **Geographic**, and **Census Campaigning**. Findings indicate that rural counties generally face higher undercount percentages compared to larger counties, with **Miami-Dade County** being a significant outlier. Despite its size, Miami-Dade not only had a high undercount percentage but also a high absolute number of undercounted individuals. Other large counties with notable undercounts included **Broward**, **Palm Beach**, **Orange**, **Hillsboro**, and **Duval**.

The analysis also explored issues such as the **digital divide** in rural areas and the impact of **social integration** in larger counties. To support ongoing efforts, FL-DSSG has developed an interactive dashboard with data estimates for FPN aimed at helping with strategic planning and mitigation of undercounts in the 2030 census.

DSSG

The ability to affect change and do good in one's community increasingly depends on having the right information at the right time to make the right decisions about the most important. Directors of community programs and funding agencies want evidence of impact and demonstrated efficiency in programs that serve our communities. Often, the information available to meet these needs is not well organized, poorly understood, and not packaged in a way that helps those working in the community do their best.

The Florida Data Science for Social Good (FL-DSSG) runs a summer internship program that matches data science expertise with real-world problems. The FL-DSSG program works with community organizations who are trying to affect change in their communities and who have data management, analysis, and data visualization projects that have the potential to shift understanding around a community issue, influence planning, revise practices, or see efforts in supporting community initiatives more focused or renewed. The goal of the FL-DSSG is to promote data-informed approaches and partner with organizations that use these approaches

to solve wicked social problems while creating educational programs for aspiring data scientists.

"We are social trustees of knowledge with a unique capacity to do social good."

Wicked Problem

The Florida Philanthropic Network (FPN) is working with the Florida Data Science for Social Good (FL-DSSG) to tackle the persistent and complex issue of census undercounts, this time focusing on the zip code level within 12 selected undercounted counties. Accurate census counts are crucial for ensuring equitable distribution of federal funds, which directly impact public services such as healthcare, education, and infrastructure. An undercount can lead to significant financial shortfalls and disparities in resource allocation, as evidenced by Florida's estimated 3.48% undercount in the 2020 Census, potentially resulting in up to $11 billion in lost federal funding over a decade. This year's project seeks to uncover the specific variables contributing to undercounts at a more granular level, aiming to understand the nuanced factors that lead to certain populations being missed in the census.

The social issue at hand is the systematic underrepresentation of certain demographic groups in the census, which exacerbates existing inequalities. These undercounts often disproportionately affect marginalized communities, leading to a cycle of underfunding and neglect. By identifying the variables that correlate with high undercount rates at the zip code level, the project aims to inform targeted interventions that can address these disparities. Understanding the root causes of undercounts will enable the development of more effective communication strategies, community partnerships, and resource allocation to ensure a more accurate and inclusive census in the future. This effort is essential for promoting social equity and ensuring that all communities receive the support and resources they deserve.

Partner Perspectives

Through a collaborative effort to increase Florida residents' awareness of the importance of the Census, the Florida Philanthropic Network (FPN) increased the representation of historically underrepresented groups in the 2020 Census. FPN continues to fundraise and engage with the community to encourage residents to complete the 2030 Census.

The 2023 Florida Data Science for Social Good (FL-DSSG) team provided FPN with crucial information about which counties were most severely undercounted during the 2020 Census. FPN used this information to identify which areas to target their philanthropic focus and help

counties and municipalities understand the importance of accurately counting their residents during the 2030 Census.

To further inform FPN's efforts, with the help of the FL-DSSG team, FPN plans to explore who lives within some of the most undercounted counties in Florida. With this information, FPN can collaborate with groups to create a strategic awareness campaign based on each area's demographics. Additionally, the group can use this information to emphasize the importance of accurate Census counting for potential grantees.

Data Sources

Our project began with an in-depth review of the Texas Census Research Institute's methodology and prior reports. This helped us identify the critical factors related to Hard-to-Count (HTC) populations, which we then aligned with the specific goals of our FPN project. The primary dataset was the American Community Survey (ACS) from the U.S. Census Bureau[1], along with several other data sources. Our focus was on 12 counties across 4 Core-Based Statistical Areas (CBSAs). Key factors influencing our analysis included:

- Internet Accessibility and Digital Device Availability (DP02): Given that the 2020 U.S. Census was fully online, internet access became a crucial variable.
- Poverty (DP03), Employment Status (DP03), and Total Household Income (DP04): These socio-economic indicators significantly impact internet use.
- Rent as a Percentage of Household Income (DP04): High rent burdens can correlate with a lower likelihood of census participation.
- Disability Status (DP02): This factor is also critical in understanding accessibility barriers.
- English Language Proficiency (DP02): Non-fluent English speakers may struggle with online forms, making this a barrier to census participation.

For residents without home internet access who may rely on public resources, we included variables like Vehicle Availability and Public Library Access from Career One Stop[2]. Recognizing the importance of internet performance, we sourced internet speed data from Ookla Speed Test[3]. To further refine our analysis, we collected Area Density data from Fourfront[4], as some areas with lower undercount percentages still had a large number of individuals underrepresented. Finally, we incorporated voter data from the Florida Division of Elections[5] to explore correlations between civic engagement and the variables affecting census participation. This multi-faceted approach enabled a robust and comprehensive understanding of the factors influencing census participation in our selected counties.

1. U.S. Census Bureau: https://data.census.gov/table
2. Public Library: https://www.careeronestop.org/LocalHelp/CommunityServices/find-libraries.aspx?location=florida
3. Ookla Speed Test: https://registry.opendata.aws/speedtest-global-performance/
4. Population Density: https://www.fourfront.us/data/datasets/us-population-density/
5. Voter Data: https://dos.fl.gov/elections/data-statistics/elections-data/

Data Cleaning

The data cleaning process was essential to ensure the dataset was accurate, consistent, and relevant for the analysis. The following steps were taken to prepare the data:

Filtering: The analysis focused exclusively on Standard zip codes, which represent areas with residential populations. By filtering out PO Box and Unique zip codes, we ensured that the analysis remained relevant to the issue of undercounts, as these Standard zip codes are where the population resides and where accurate census data is most crucial.

Handling Missing Data: The dataset was largely complete, with minimal missing values. In cases where data was missing for a particular variable, it often indicated that the estimate for that variable was zero in that zip code. Given this context, missing data was not imputed or removed, as the absence of data was itself meaningful and indicative of the underlying demographics.

Outlier Detection: Outliers were not removed from the dataset, as they were considered potentially significant findings rather than errors. These outliers may represent unique characteristics or patterns in specific zip codes that could contribute valuable insights to the analysis. Instead of removal, these data points were carefully examined and retained to better understand their implications.

Annotation: Detailed annotations were added throughout the dataset to ensure transparency and consistency. Each variable was thoroughly documented, including the source of the data, any transformations applied, and the rationale behind those transformations. For example, notes were made on why certain variables were included or excluded and how missing data was treated. These annotations serve as a guide for anyone reviewing the dataset, ensuring that all data preparation decisions are clearly understood and traceable.

Master Data Files

Two master files were prepared—one containing estimate data and the other containing percentage data. This separation allowed for more focused analysis depending on the type of data being considered.

Merging and Cleaning: The cleaned datasets were consolidated into the two master files. Each file included all relevant variables and zip codes, ensuring consistency across the dataset.

Additional Columns: New columns were added to enrich the dataset, including the County, Core Based Statistical Area (CBSA) associated with each zip code and a column to filter out zip codes with less than 1,000 population. These additions were crucial for refining the analysis and focusing on areas with significant populations.

Transformation and Formatting: Data was formatted and transformed to meet the requirements of the analysis tools. Since all variables in the dataset are continuous, there was no need for categorical encoding. Normalization or scaling was applied where necessary to prepare the data for further analysis.

Feature Selection: Variables were carefully selected based on their relevance to the analysis. Unnecessary variables were excluded to streamline the dataset and focus on the most impactful features. This step was guided by both domain knowledge and exploratory data analysis, ensuring that the final dataset was optimized for uncovering the factors influencing undercounts.

Descriptive Analysis

Descriptive statistics were calculated for key variables to understand their central tendencies and distributions. These statistics provided a foundational understanding of the dataset, highlighting trends and patterns that could inform further analysis.

For instance, one key variable analyzed was the percentage of households without internet access. The statistics revealed significant variation across zip codes, with some areas showing a high proportion of households lacking internet access. This finding was crucial, as it suggested potential barriers to census participation in these areas, aligning with broader research on the digital divide.

Example Variable: Internet Access

Exploratory Insights:

Visual tools such as Histograms and Bar Graphs were used to explore the distribution of internet access across different zip codes. These visuals highlighted the disparity in internet access, with some zip codes showing notably high percentages of households without connectivity. This metric was particularly revealing, as it pointed to areas where a lack of digital access might hinder census participation.

New Characteristics Discovered: Further exploration revealed that zip codes with low internet access also tended to have higher percentages of Poverty and Non Fluent English Speakers. This correlation suggested that these areas might face compounded challenges in census participation, including language barriers and digital exclusion.

Comparison with Research: These findings were consistent with existing research, which underscores the importance of internet access in achieving accurate census counts. The insights gained from this analysis reinforce the need for targeted interventions in areas with low digital connectivity to improve census accuracy.

Methodology

We collected data from multiple sources to examine potential factors contributing to census undercount in 12 selected counties in Florida. The data was consolidated into a dataset comprising hundreds of variables representing different socio-economic, infrastructural, and demographic factors. Ensured all datasets were properly formatted and cleaned by handling missing values and ensuring consistent data types

Given the large number of variables, it was essential to reduce the dimensionality of the dataset before performing the final analysis. For this purpose, we conducted **Spearman correlation analysis** to identify relationships between variables and remove redundant or less significant variables.

Spearman's rank correlation coefficient is a non-parametric measure of the strength and direction of association between two ranked variables. It assesses how well the relationship between two variables can be described using a monotonic function.

The formula for Spearman correlation is:

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

Where:

- ρ = Spearman rank correlation coefficient.
- d_i = Difference between the ranks of corresponding variables.
- n = Number of data points.

Spearman correlation was chosen due to its robustness in handling non-linear relationships and ordinal data, which was critical given the nature of the variables. After calculating the Spearman correlation between pairs of variables, we selected variables that had a high correlation with census undercount while minimizing multicollinearity. Variables with high inter-correlations were evaluated, and redundant variables were removed to create the final set for our master file.

The reduced dataset, now containing only the most relevant variables, was used for further analysis to identify key factors contributing to census undercount. These included:

- Demographic characteristics
- Socio-economic factors
- Infrastructure and access indicators like internet speed and library resources

For our final analysis, we conducted a **cluster analysis** to group zip codes into clusters based on their similar risk factors. These clusters represented groups of zip codes that shared similar characteristics impacting census undercount. This analysis allowed us to identify High-Risk Areas in which clusters were formed based on variables such as population density, internet speed, and economic conditions.

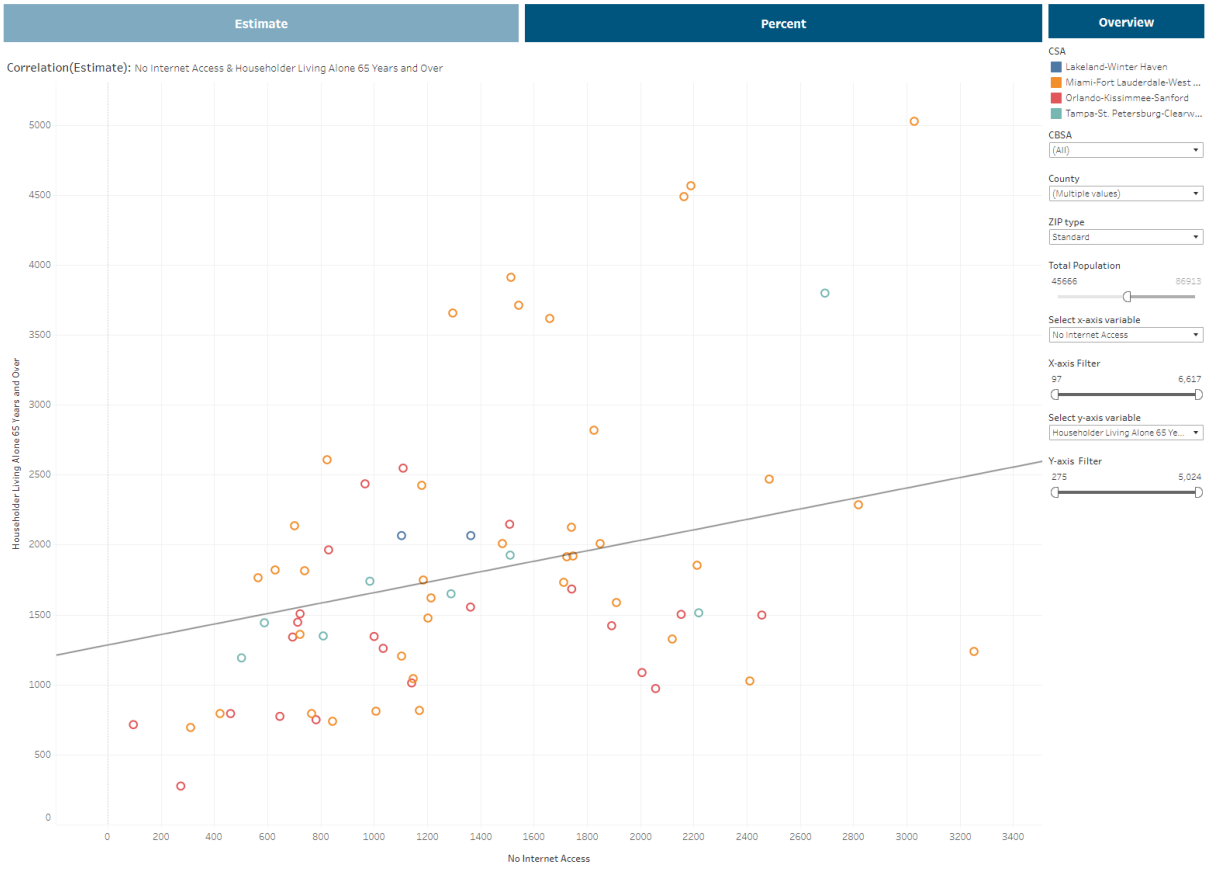
This methodology allowed us to focus on the most impactful variables, ensuring a data-driven approach to addressing the issue of census undercount in the selected counties.

Analysis and Findings

Our first goal for the project was to reduce the amount of variables to analyze. To identify redundant variables, we conducted a series of Spearman's correlations. If two variables were highly correlated, we removed the variable that had less variability. However, if two variables were statistically redundant but conceptually different (e.g., Latin American immigrants and Spanish non-fluent English speakers), we retained both variables.

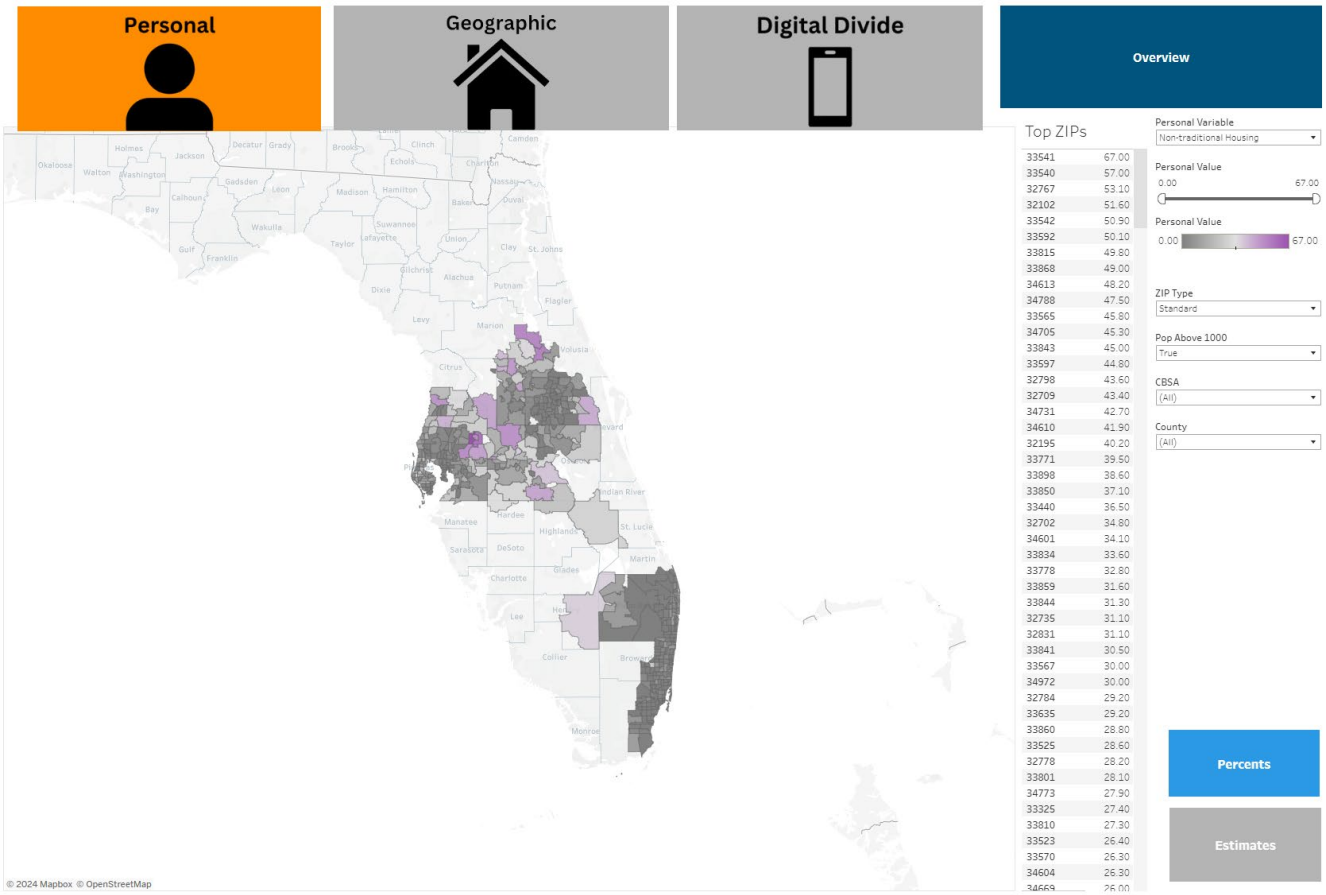
Correlations:

To identify ZIP codes that have multiple risk factors for being undercounted, we created a dynamic correlation plot for every variable in our dataset. For example, as can be seen in the figure below from the percentage data of our master file.



Dimensions:

After reducing the variables in each dimension, we created a dashboard to explore which risk factors are present in each ZIP code. For example, examining non-traditional housing within the personal dimension, the map shows that non-traditional housing is an important risk factor in central Florida. Additionally, according to the “Top ZIPs” table, 33541 has the highest percentage of non-traditional housing in our selected counties.



Cluster Analysis:

After reducing dimensions, we conducted a K-means cluster analysis to identify ZIP codes that share similar features. For example, one cluster of ZIP codes has many residents who are unemployed and have no health insurance. From our cluster analysis, we discovered that ZIP codes can have similar features even if they are geographically distant from each other.

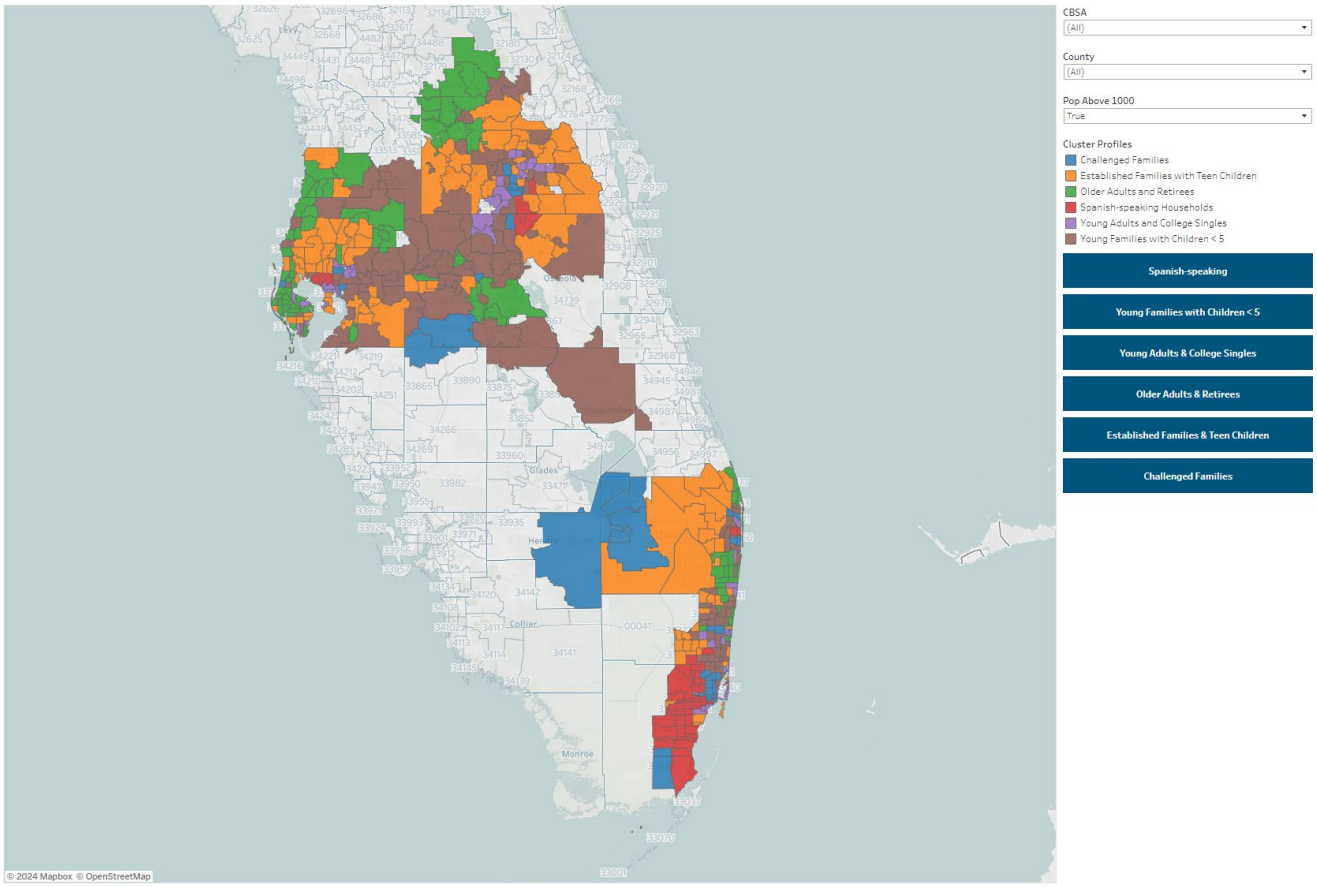


Tableau Dashboard:

We have developed and deployed a Tableau dashboard that provides access to correlation, dimension, and clustering analysis results. FPN can utilize interactive filter options provided in the dashboard to focus on variables, counties, or zip codes of their interest. The Tableau dashboard can be accessed at https://tabsoft.co/4dQrRf2.

Conclusion

With the insights derived from our analysis and data visualizations at the zip code level, the Florida Philanthropic Network (FPN) is now equipped with detailed, localized information to inform strategic interventions ahead of the 2030 Census. This granular view allows FPN to identify specific communities at the highest risk of undercounting, enabling targeted efforts to address the factors contributing to the undercount.

The availability of zip code-level visualizations will allow FPN to refine its outreach strategies in the most affected areas and facilitate more impactful communication with partners across the state. These efforts will help build a robust and dynamic network of local, regional, and state organizations dedicated to mitigating undercounting risks.

Recommendations And Next Steps

By addressing the underlying factors—such as population density, internet access, and voter registration trends—at the zip code level, FPN can better allocate resources, improve public engagement, and reduce the potential for undercounting in 2030. This data-driven approach will ensure that resources are used efficiently and effectively to maximize participation in the Census and minimize the risk of future undercounts across Florida. By understanding the shared risk factors within these clusters that we have created, FPN can develop targeted, region-specific strategies that address the underlying causes of census undercounting in each group of zip codes.

Using the findings from the cluster analysis, FPN can identify and prioritize the zip code clusters that are most affected by census undercount risk factors. These clusters should be the focus of immediate strategic planning and outreach efforts.

Design intervention programs tailored to the specific risk factors identified within each cluster. For instance, areas with low internet access may benefit from digital literacy campaigns or offline census participation drives.

Engage local organizations and stakeholders within each cluster to ensure culturally relevant and accessible outreach efforts. Collaborating with trusted community partners will help amplify FPN's efforts.

Get More Information

For additional information about the FL-DSSG program, its process, methodologies, tools, and techniques, and how to participate in the program, contact FL-DSSG program directors. Visit the FL-DSSG website for contact information.

FL-DSSG interns work on wicked problems that require technical know-how and relevant domain knowledge. To help interns complete the projects, we obtain mentors from industry professionals and research faculty members. All mentors offer their assistance voluntarily. We deeply appreciate the assistance provided by mentors.

FPN DSSG Project Resources

Project Presentation YouTube Video: https://youtu.be/VwM9DGhaqxk
2024 Big Reveal Presentation Slides: https://bit.ly/24FLDSSGBigRevealSlides
Zip-code Level Analysis Tableau Dashboard: https://tabsoft.co/3TH3t8q

2024 FL-DSSG Interns

Venkata Hema Abhinav Aaduru, Master of Science – Business Analytics and Information Systems, University of South Florida
Ziena Baker, Master of Science – Psychological Science, University of North Florida
Rebecca Burstein, Bachelor of Science – Health Science, University of North Florida
Devan Kreitzer, Master of Science – Business Analytics and Information Systems, University of South Florida
Md Anwar Hossain, Master of Science – Statistics, University of North Florida
Kevin Luhrs, Master of Public Administration – Public Policy, University of North Florida
Shanmukh Sai Madhu, Master of Science – Computer Science, University of South Dakota
Ula McCarthy, Master of Science – Psychological Science, University of North Florida
Michael Mezzano, Master of Science – Data Science, University of West Florida
Lissa Nzirlu, Master of Science – Statistics, University of North Florida
Samuel Pearl, Master of Science – Psychological Science, University of North Florida
John Reddick, Bachelor of Science – Computer Science, University of North Florida

2024 Sponsors

University of North Florida, Jacksonville Jaguars, Cathedral Arts Project, Florida Philanthropic Network, NLP Logix, PGA Tour, Jacksonville Area Legal Aid, Florida Health Justice Project, ADVOS Legal, Miller Electric Company, MarketOnce, iVenture Solutions, Leanovation Labs, and Smith Gambrell Russell

FL-DSSG Program Directors

Dr. Dan Richard, Professor, Psychology Department
Dr. Karthikeyan Umapathy, Professor, School of Computing

Florida Data Science for Social Good (FL-DSSG)
University of North Florida
December 2024
https://dssg.unf.edu/