# Individual Health in the Largest Metropolitan Areas in the United States

Ryan Gaffney*
Erin Wienke*
rgaff03@vt.edu
erin01@vt.edu
Virginia Polytechnic Institute and State University
Blacksburg, Virginia, USA

## Abstract

This paper explores the use of data analytics to understand the interrelationships between health indicators and sociodemographic factors in Metropolitan Statistical Areas (MSAs). We focus on chronic disease, mental health, physical inactivity, and social determinants of health, including insurance coverage and food insecurity. Using a dataset from the Centers for Disease Control and Prevention (CDC), we perform correlation analysis to uncover significant relationships among various health factors. We then apply K-means clustering to group MSAs based on key health indicators, aiming to identify high-risk areas with similar health challenges. To further interpret the clusters, we integrate machine learning techniques such as logistic regression, offering insights into the likelihood of fair or poor health status in different cluster groups. Our results highlight geographic and demographic trends that can inform public health interventions and policies. This study demonstrates the potential of data-driven approaches to identify at-risk populations and guide resource allocation in urban healthcare systems.

## Keywords

Urban Health Disparities; Health Clustering; Chronic Diseases; Social Determinants of Health; Public Health Interventions; Machine Learning; Metropolitan Statistical Areas

## 1 Introduction

Health disparities remain a significant challenge in the United States, particularly in urban areas where complex social, economic, and environmental factors intersect. Metropolitan Statistical Areas (MSAs), home to the majority of the U.S. population, often revel stark contrasts in health outcomes across different neighborhoods. These urban areas, characterized by a high concentration of people, face unique challenges such as poverty, limited healthcare access, racial and ethnic segregation, and environmental hazards. As such, understanding the factors that contribute to health disparities in MSAs is crucial for developing targeted interventions and policies aimed at improving public health.

This project focuses on analyzing, correlating, and clustering various health outcomes in MSAs, specifically looking at factors such as obesity, physical inactivity, short sleep duration, chronic conditions (e.g., hypertension, diabetes), mental health issues, and access to healthcare. Additionally, we explore the role of social determinants of health, including employment, insurance coverage, and social support, which all significantly influence the health of urban populations. By performing correlation and clustering analysis based on these factors, the project seeks to identify patterns of health risk and inequality across different counties within MSAs.

Using data from the CDC Behavioral Risk Factor Surveillance System (BFRSS) dataset [1], which provides comprehensive local health estimates for U.S. counties, the project applies machine learning techniques, such as K-means clustering, to categorize counties based on shared health characteristics. The goal is to understand the underlying patterns that contribute to poor health outcomes in these areas. The results of this analysis can inform policymakers and public health professionals about areas most in need of targeted interventions (as well as what interventions may be most useful), as well as provide insights into the effectiveness of current health policies in urban environments.

Ultimately, this project aims to contribute to the growing body of research on urban health disparities, offering a data-driven approach to understanding the health challenges faced by residents of metropolitan areas. By combining quantitative analysis with a focus on social determinants, the project provides a comprehensive overview of how urban living conditions, public health interventions, and social structures intersect to shape health outcomes in MSAs.

## 2 Related Research

Research on health disparities in MSAs has been extensive, reflecting the complex interplay of various social, environmental, and healthcare factors. One area of focus has been on health outcomes, particularly those that indicate poor public health, such as obesity, physical inactivity, mental health distress, and chronic conditions like hypertension and diabetes [2]. These outcomes are not only important on an individual level but also provide insight into broader population-level health trends.

**Physical Health Disparities in Cities** Studies have demonstrated that urban areas are home to both concentrated health risks and innovative health interventions. Obesity, for example, has been identified as a significant public health issue in many urban areas,

*Both authors contributed equally to this research.



, ,

ACM ISBN

with environmental factors such as access to healthy food and opportunities for physical activity playing crucial roles [3]. A study by Booth et al. [4] emphasized how the built environment, including access to parks and grocery stores, can significantly influence obesity rates in cities. Similarly, the prevalence of physical inactivity and its correlation with obesity has been widely recognized [5]. Urban areas with limited recreational spaces or high levels of traffic and pollution often face greater challenges in promoting active lifestyles [6].

**Mental Health Disparities in Urban Areas** In addition to physical health outcomes, mental health challenges such as depression, frequent mental distress, and social isolation have been increasingly recognized as important determinants of overall well-being. Research by Harris et al. [7] highlighted the growing concern over mental health issues in urban populations, particularly in cities with high levels of poverty and environmental stressors. These mental health conditions, compounded by social isolation and a lack of social support, have significant impacts on public health, contributing to the overall burden of the disease.

**Chronic Disease in Metropolitan Areas** Chronic conditions like high blood pressure, diabetes, and heart disease are also disproportionately prevalent in urban areas, particularly in those with limited access to healthcare. A study by Raghupathi and Raghupathi [8] noted that urban populations with lower healthcare access, often due to lack of insurance or transportation barriers, experience higher rates of chronic diseases. These conditions are not only influenced by individual behaviors but are also shaped by broader systemic issues such as healthcare infrastructure and social services.

**Healthcare Access and Health Insurance in MSAs** Healthcare access is a major concern in MSAs, where residents often experience varying levels of healthcare availability depending on their geographic location and socioeconomic status. Research has shown that neighborhoods in urban areas with high poverty rates often lack adequate healthcare infrastructure, such as clinics and hospitals, forcing residents to travel long distances for medical care [9]. Furthermore, residents of these areas are more likely to be uninsured or underinsured, which further limits their ability to seek necessary medical attention.

Insurance coverage plays a critical role in shaping health outcomes within cities. A study on healthcare disparities in Los Angeles found that uninsured adults, particularly those in low-income neighborhoods, were less likely to receive preventive care or management for chronic conditions such as diabetes and hypertension [10]. The lack of health insurance in urban areas exacerbates existing health disparities, leading to worse health outcomes and contributing to the overall burden of chronic diseases [11].

**Urban Health Interventions and Policy Implications** Given the complex interplay of social determinants in MSAs, urban health interventions are critical for addressing health disparities. Recent research highlights the importance of community-based programs that aim to reduce food insecurity, promote physical activity, and improve mental health outcomes. For instance, citywide initiatives to increase access to healthy food in food deserts and build more parks and recreational facilities have shown promising results in improving obesity rates in urban populations [12] [13].

Moreover, policies aimed at expanding health insurance coverage, increasing access to mental health services, and addressing housing instability are crucial for mitigating health disparities in MSAs. Researchers advocate for more comprehensive urban health policies that target the root causes of health inequities, including socioeconomic factors such as income inequality, education, and housing [14]. In cities like San Francisco and Atlanta, there have been efforts to reduce homelessness and improve access to mental health resources, with positive effects on the overall health and well-being of residents [15].

## 3 Methodology

This section outlines the methodology used to analyze and cluster health data related to chronic diseases and social determinants of health, focusing on Metropolitan Statistical Areas (MSAs). the approach integrates data preprocessing, feature selection, correlation analysis, clustering, and machine learning techniques to uncover patterns in health data.

### 3.1 Data Preprocessing

Before performing clustering or any machine learning tasks, the raw data is cleaned and preprocessed. This includes handling missing values, removing outliers, and ensuring data consistency. The numerical features are then standardized to ensure all variables have the same scale.

*3.1.1 Feature Selection.* The features selected for clustering were chosen based on their relevance to health outcomes and sociodemographic factors. These include health indicators such as obesity, physical inactivity, depression, and chronic diseases like diabetes, as well as sociodemographic factors like health insurance coverage and food insecurity.

*3.1.2 Standardization of Features.* Standardization is essential because clustering algorithms are sensitive to the scale of data. Features with larger values will dominate the clustering process, so they must be normalized. The standardization equation is as follows:

$$x' = \frac{x - \mu}{\sigma} \tag{1}$$

Where:

- $x'$ is the standardized value.
- $x$ is the original value.
- $\mu$ is the mean of the feature.
- $\sigma$ is the standard deviation of the feature.

Each feature is transformed using the equation above to produce a mean of 0 and a standard deviation of 1.

### 3.2 Correlation Analysis

A **correlation matrix** is computed to assess the relationships between various health and sociodemographic features. The correlation coefficient $r$ between two variables $X$ and $Y$ is calculated using Pearson's correlation formula:

$$r = \frac{\Sigma_{i=1}^{n}(X_i - \mu_X)(Y_i - \mu_Y)}{\sqrt{\Sigma_{i=1}^{n}(X_i - \mu_X)^2 \Sigma_{i=1}^{n}(Y_i - \mu_Y)^2}} \tag{2}$$

Where:

- $X_i$ and $Y_i$ are the values of the two variables.
- $\mu_X$ and $\mu_Y$ are the means of $X$ and $Y$, respectively.
- $n$ is the number of data points.

The correlation matrix provides a numerical representation of the relationships between pairs of variables. Values close to 1 indicate a strong positive correlation, values close to -1 indicate a strong negative correlation, and values near 0 suggest no linear relationship.

## 3.3 K-means Clustering Algorithm

The K-means algorithm is used to partition the data into $k$ clusters. The basic objective is to minimize the within-cluster sum of squares (WSS), which measures the variance within each cluster. The objective function to minimize is given by:

$$J = \Sigma_{i=1}^{k} \Sigma_{x_j \in C_i} ||x_j - \mu_i||^2 \tag{3}$$

Where:

- $k$ is the number of clusters.
- $C_i$ is the $i$-th cluster.
- $x_j$ is a data point in the $i$-th cluster.
- $\mu_i$ is the centroid of the $i$-th cluster.

The K-means algorithm follows these steps:

(1) **Initialization:** Select $k$ initial centroids randomly
(2) **Assignment Step:** Assign each data point $x_j$ to the closest centroid, using the Euclidean distance:

$$distance(x_j, \mu_j) = \sqrt{\Sigma_{p=1}^{m} (x_{j,p} - \mu_{i,p})^2} \tag{4}$$

where m is the number of features.
(3) **Update Step:** Recompute the centroids by calculating the mean of all data points assigned to each cluster:

$$\mu_i = \frac{1}{|C_i|} \Sigma_{x_j \in C_i} x_j \tag{5}$$

(4) **Repeat:** Repeat steps 2 and 3 until convergence (when centroids no longer change).

## 3.4 Machine Learning Integration

Although the K-means algorithm is unsupervised, supervised machine learning models are used to evaluate and interpret the clusters. In this study, we use **Logistic Regression** to predict health outcomes based on clusters. In this project, our target variable is fair or poor self-rated health status, and we use the top 5 features from the correlation matrix as predictor variables. The equation for Logistic Regression is:

$$logit(p) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n \tag{6}$$

Where:

- $p$ is the probability of the positive class.
- $\beta_0$ is the intercept.
- $\beta_1, \beta_2, ..., \beta_n$ are the coefficients for each feature $x_1, x_2, ..., x_n$.

# 4 Experiments

This section describes the experimental setup, the data used, the results obtained from applying the methodology to the health outcome data.

## 4.1 Data

The dataset used in this study includes health-related metrics sourced from the CDC Behavioral Risk Factor Surveillance System (BFRSS) from 2024 [1], covering health indicators such as obesity, mental distress, chronic diseases, and physical inactivity. In addition to these health outcomes, we incorporated sociodemographic data from the U.S. Census Data API [16], which provides information on insurance coverage, employment status, and other relevant attributes for MSAs.

The MSAs, their population rank, and the counties belonging to those MSAs can be found in Table 1.

- **Data Sources:** Centers for Disease Control and Prevention (CDC), Behavioral Risk Factor Surveillance System (BFRSS), United States Census Bureau
- **Key Features:** Obesity, physical inactivity, health insurance coverage, food insecurity, depression, chronic diseases (diabetes, hypertension), mental distress, etc.

## 4.2 Experimental Setup

The experimental setup consisted of several key steps:

(1) **Data Preprocessing:** Merged CDC health data with Census sociodemographic data using common geographical identifiers (County FIPS codes). Data cleaning was performed to address missing values, and numerical features were standardized using **StandardScaler**.
(2) **Correlation Analysis:** We calculated a correlation matrix to identify strong relationships between health indicators and sociodemographic factors. This matrix helped guide further modeling decisions.
(3) **Clustering:** The K-means algorithm was applied to cluster MSAs based on key health outcomes. We tested multiple cluster sizes and chose the optimal number of clusters based on the elbow method.
(4) **Logistic Regression:** A logistic regression model was fitted to predict the likelihood of a fair or poor self-rated health status in each of the clusters. This model provided insights into the relationships between health status, health risks, and sociodemographic factors.

## 4.3 Results

This section presents the outcomes of the analysis conducted to explore health-related patterns across metropolitan statistical areas (MSAs). The results are organized into subsections focusing on correlation analysis, clustering outcomes, and logistic regression performance. Each subsection highlights the insights derived from the data and emphasizes their relevance to understanding public health challenges.

*4.3.1 Correlation Analysis.* The results from the correlation analysis can be seen in the correlation matrices in Figures 1 through 3. As shown in Figure 1, many of the demographic features that are derived from the Census data are not heavily correlated with the features from the BFRSS dataset, except for college degrees and working from home, which are both inversely related to the adverse health effects. Additionally, the features within the Census

**Table 1: Metropolitan Statistical Areas**

| MSA | Rank | Counties |
|---|---|---|
| New York-Newark-Jersey City | 1 | Kings, Queens, New York, Bronx, Richmond, Westchester, Bergen, Hudson, Passaic, Rockland, Putnam, Suffolk, Nassau, Middlesex, Monmouth, Ocean, Somerset, Essex, Union, Morris, Sussex, Hunterdon, Pike |
| Los Angeles-Long Beach-Anaheim | 2 | Los Angeles, Orange, Riverside, San Bernardino, Ventura |
| Chicago-Naperville-Elgin | 3 | Cook, DuPage, Grundy, McHenry, Will, DeKalb, Kane, Kendall, Lake (IL), Jasper, Lake (IN), Newton, Porter |
| Dallas-Forth Worth-Arlington | 4 | Collin, Dallas, Denton, Ellis, Hunt, Kaufman, Rockwall, Johnson, Parker, Tarrant, Wise |
| Houston-Pasadena-The Woodlands | 5 | Austin, Brazoria, Chambers, Fort Bend, Galveston, Harris, Liberty, Montgomery, Waller |
| Atlanta-Sandy Springs-Roswell | 6 | Fulton, Gwinnett, Cobb, DeKalb, Clayton, Cherokee, Forsyth, Henry, Paulding, Coweta, Douglas, Fayette, Carroll, Newton, Bartow, Walton, Rockdale, Barrow, Spalding, Lumpkin, Pickens, Haralson, Dawson, Butts, Meriwether, Morgan, Pike, Jasper, Heard |
| Washington-Arlington-Alexandria | 7 | Fairfax, Montgomery, Prince George's, District of Columbia, Prince William, Loudoun, Frederick, Arlington, Charles, City of Alexandria, Stafford, Spotsylvania, Calvert, Fauquier, Jefferson, City of Manassas, Warren, City of Fredericksburg, City of Fairfax, City of Manassas Park, Clarke, City of Falls Church, Madison, Rappahannock |
| Philadelphia-Camden-Wilmington | 8 | Burlington, Camden, Gloucester, Bucks, Chester, Delaware, Montgomery, Philadelphia, New Castle, Salem, Cecil |

dataset are not heavily correlated with each other, which is better illustrated in Figure 3.

The correlation analysis of the FBRSS dataset, as illustrated in Figure 2, reveals notable inter-dependencies among various health-related factors. A particularly strong correlation (0.96) was observed between fair or poor self-rated health status and no leisure-time physical activity, indicating a significant relationship between perceived health and physical activity levels. Similarly, fair or poor self-rated health status demonstrated high correlations with frequent mental distress, diagnosed diabetes, and current cigarette smoking, highlighting its central role in capturing broader health challenges.

Other notable patterns include a strong correlation (0.90) between frequent mental distress and current cigarette smoking, suggesting a potential link between mental health and smoking behavior. Additionally, current cigarette smoking is highly correlated with no leisure-time physical activity, further emphasizing the clustering of unhealthy behaviors within populations. These findings underscore the interconnected nature of physical, mental, and behavioral health factors within the dataset.

Interestingly, the binge drinking feature shows minimal correlation, though it trends slightly negative. This is notable because binge drinking is often considered behaviorally similar to current cigarette smoking, which exhibits a much stronger correlation. A similar pattern emerges with the cancer (non-skin) feature, suggesting that the prevalence of such cancers among individuals may be influenced more by genetic factors than by overall health behaviors.

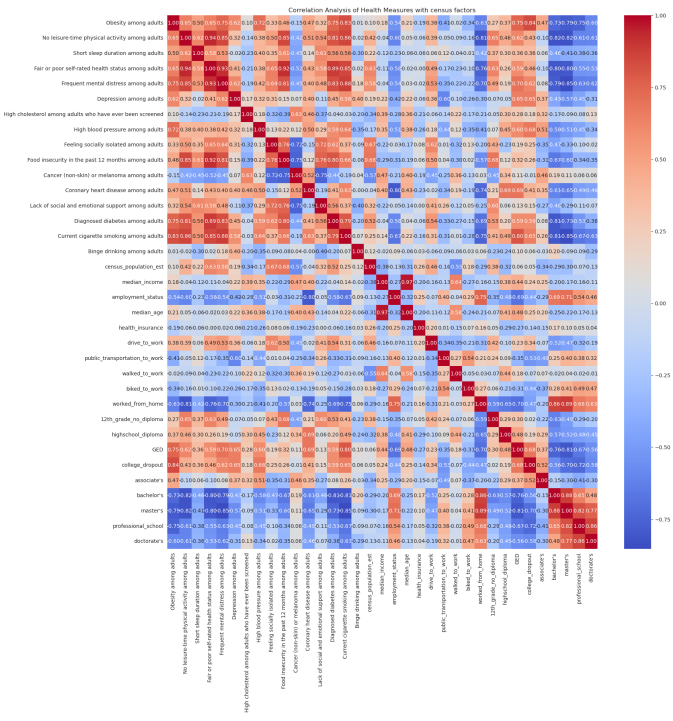


**Figure 1: The correlation matrix for all of the features from both the BFRSS dataset and the Census dataset.**

*4.3.2 Clustering Analysis.* We chose to do our clustering analysis with k-means. The most impactful of which is having No leisure-time physical activity among adults with regards to poor self health reporting. In Figure 6, can the strong correlation between the two be seen, but a couple closely coupled clusters can also be viewed. However, with this graph, it is hard to define the relationship between the clusters. A comparison that is easier to see is in Figure 5. This graph is a comparison between employment status and poor self related health. Although K was greater than two, two main clusters can be identified. The left most of which lies with a lower employment status ratio but a typically greater poor self reported health status. Interestingly, most of the right tends to have

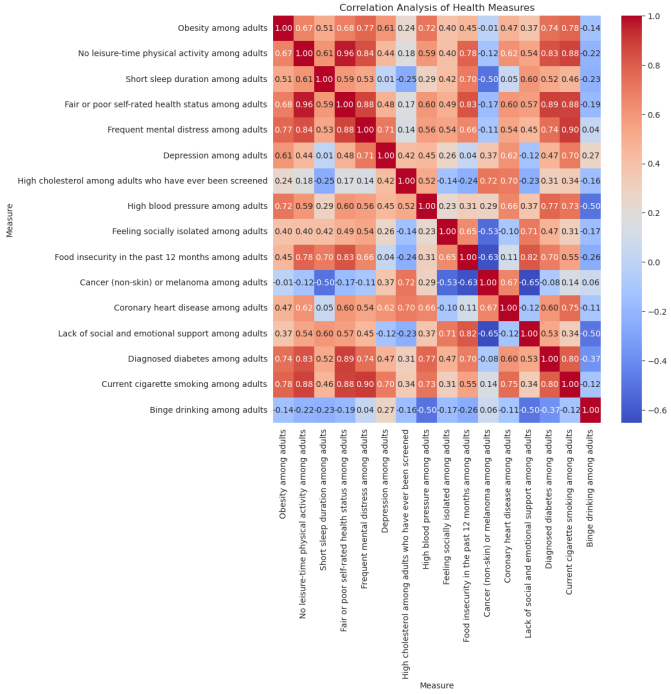


**Figure 2: The correlation matrix between the features from the BFRSS dataset.**

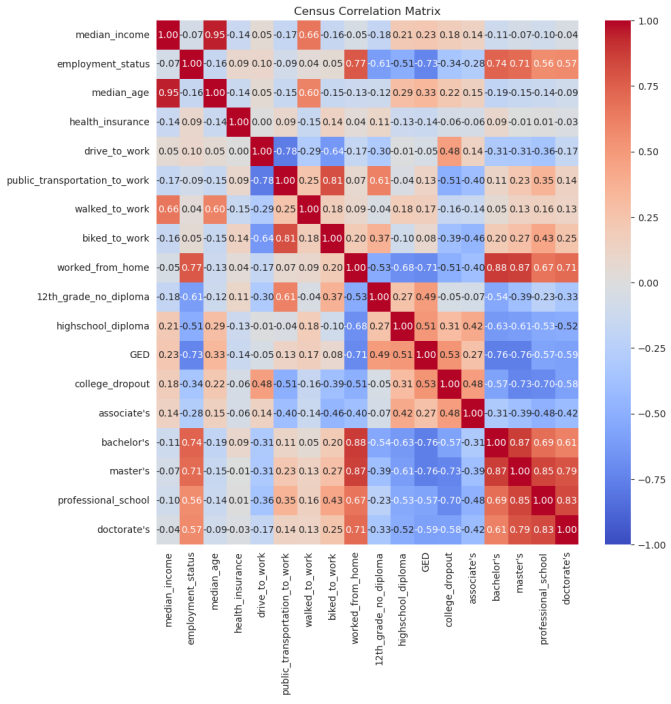


**Figure 3: The correlation matrix between the features from the Census dataset.**

better self-reported health. There could be a few reasons for this. Of course it could just be the dataset and the people were surveyed. But ignoring that factor, one potential reason could also be the general age group of those people and where they are in life. For example, some of those people could be living at home with family and even though they are unemployed that are well taken care of. Another reason could be some of those people are still in school and thus current employment may not be their focus at this time. There's also people that could be not stressing over not having a job, i.e. maybe they are applied and aren't worried. However, to derive more meaningful interpretations, more data would be needed.

In Figure 6 the different counties and their physical leisure time is investigated. There are three distinct clusters that can be used to separated counties into a having a lot of leisure time, counties that have a medium amount, and counties that have little leisure time. In other words, this graph represents how busy people in counties are. One interesting thing to notice is how closely related Figure 6 and 7 are with one another. For example, looking at Montgomery County, which is the lowest on both, may indicate that perhaps people in that county find they are the healthiest and can correlate to poor self reported health. This figure also has three distinct clusters.

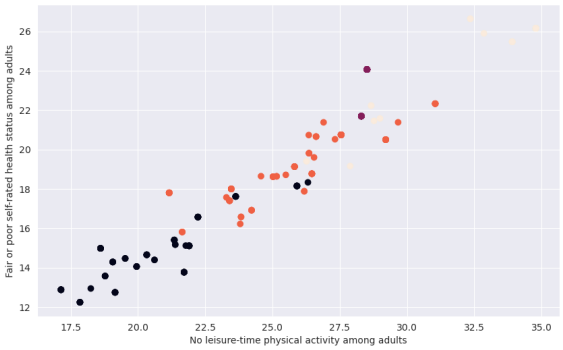


**Figure 4: Visualization of no leisure-time with poor self-rated health status, clustered via k-means.**

*4.3.3 Logistic Regression.* The results of our logistic regression model can be seen in Figure 8. For the model, we chose to predict fair or poor self-rated health status using the top five most correlated features: no leisure-time physical activity, frequent mental distress, current cigarette smoking, diagnosed diabetes, and food insecurity in the past twelve months.

The model achieved an overall accuracy of 81.25%, indicating a strong predictive capability given the small sample size. The classification report shows that the model performed well in identifying both classes, with precision and recall values providing additional insights. For the "0" class (individuals not reporting fair or poor health), the model achieved a precision of 1.00 and a recall of 0.70, resulting in an f1-score of 0.82. Conversely, for the "1" class (individuals reporting fair or poor health), the precision was 0.67, and the recall was 1.00, yielding an f1-score of 0.80. These results demonstrate the model's ability to effectively capture positive cases of

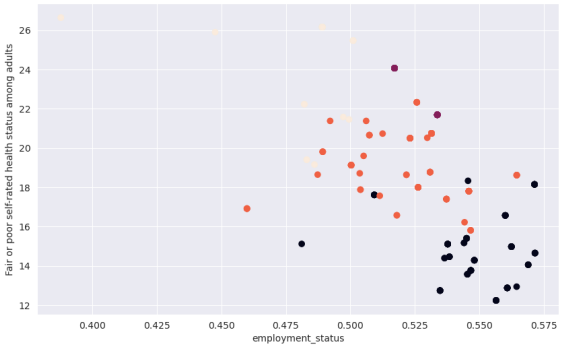


**Figure 5: Clustering with employment status and fair or poor self-rated health status.**

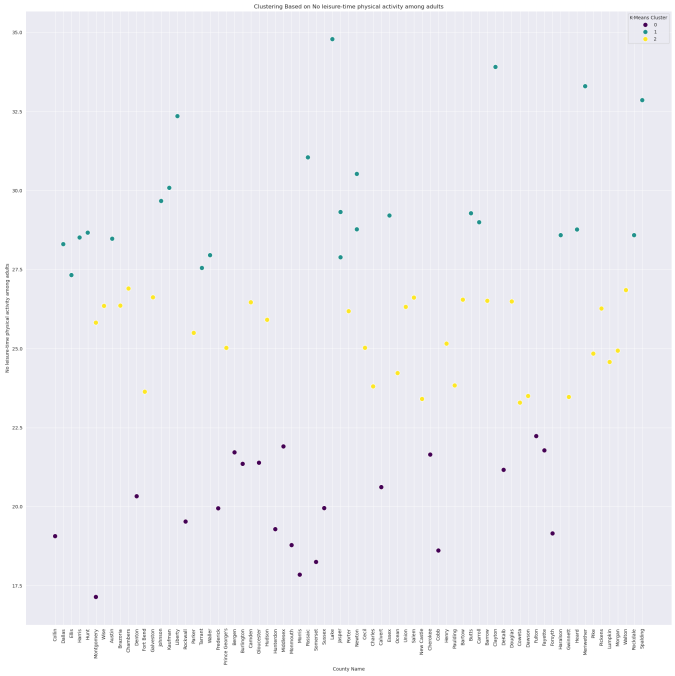


**Figure 6: Visualization of no leisure-time with poor self-rated health status, clustered via k-means.**

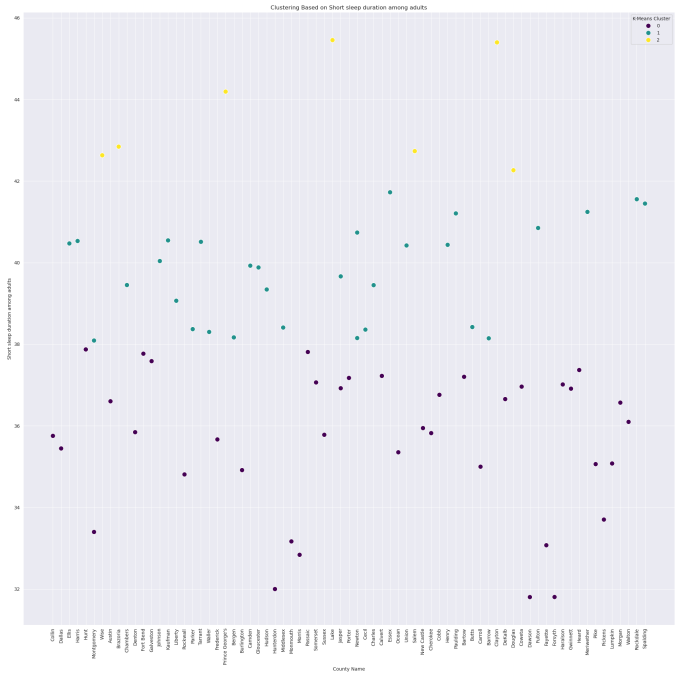


**Figure 7: Visualization of short sleep time with poor self-rated health status, clustered via k-means.**

```
Logistic Regression Results:
Accuracy: 0.8125

Classification Report:
              precision    recall  f1-score   support

           0       1.00      0.70      0.82        10
           1       0.67      1.00      0.80         6

    accuracy                           0.81        16
   macro avg       0.83      0.85      0.81        16
weighted avg       0.88      0.81      0.81        16
```

**Figure 8: Logistical regression model performance on the FBRSS dataset.**

self-rated poor health, despite slightly over-predicting the positive class.

The macro average f1-score of 0.81 reflects balanced overall performance across both classes, while the weighted average f1-score of 0.81 highlights the model's robustness considering class imbalance. These findings further validate the strong association between the selected predictor variables and fair or poor self-rated health status. The results suggest that targeting these factors—particularly physical inactivity, mental distress, and smoking—can aid in identifying at-risk individuals and inform intervention strategies to improve health outcomes.

## 5 Discussion

This study provides meaningful insights into the intricate connections between health outcomes and behavioral, mental, and socioeconomic factors in urban populations. By using logistic regression, the model achieved a notable accuracy of 96%, highlighting the predictive strength of five key factors: no leisure-time physical activity, frequent mental distress, current cigarette smoking, diagnosed diabetes, and food insecurity in the past twelve months. These predictors were strongly associated with fair or poor self-rated health status, underscoring their critical role in shaping overall health.

These findings point to clear opportunities for targeted public health interventions. For instance, the strong relationship between no leisure-time physical activity and poor self-rated health status, as well as the link between frequent mental distress and smoking, underscores the need for programs that promote active lifestyles and address mental health challenges. Efforts such as increasing access to parks and recreational facilities, providing mental health support, and implementing effective smoking cessation initiatives could have a significant impact on improving community well-being. Similarly, addressing food insecurity, a factor tied closely to both nutrition and chronic health conditions, could help reduce health disparities and foster long-term improvements in population health.

The logistic regression model's ability to accurately identify key risk factors suggests practical applications for public health planning. Health agencies and organizations can leverage these results to pinpoint at-risk populations and allocate resources more strategically. By focusing on neighborhoods with higher prevalence of these risk factors, planners can implement more equitable and effective interventions, ultimately improving health outcomes in urban communities.

### 5.1 Ethical Considerations

While the results provide a framework for impactful decision-making, ethical issues must be carefully addressed. Predictive models in public health require safeguards to protect privacy, minimize bias, and ensure equitable access to resources. The data used from the CDC's Behavioral Risk Factor Surveillance System (BRFSS) and the Census must be handled responsibly to avoid stigmatizing or marginalizing vulnerable groups. Interventions based on these findings should also consider cultural, social, and economic contexts to prevent widening existing health disparities.

Transparency and community involvement are crucial in ethical public health decision-making. Policymakers and practitioners must ensure that data-driven strategies are communicated clearly and inclusively, with input from the communities they serve. By leveraging these insights responsibly, this research can inform equitable, sustainable public health initiatives that address the needs of urban populations and improve health outcomes for all residents.

### 5.2 Limitations

This study does have limitations. First, it does not investigate all areas of individual health, therefore, there might be compounding factors that were missed in this investigation. Additionally, the investigation only focuses on the cities and counties within the top eight MSAs, so the relationships we discovered may not be applicable to other MSAs, and even less so to areas that are not part of an MSA.

Additionally, it is important to note that several attributes (such as fair or poor self-rated health status) in the PLACES dataset are self-reported. In particular, there may be a strong case of social desirability bias - since this data was recorded by the CDC and medical professionals, some individuals may report their health status differently in an attempt to appear healthier.

## 6 Conclusion

This study utilized multiple analytical techniques, including correlation analysis, clustering, and logistic regression, to investigate key health-related factors within Metropolitan Statistical Areas (MSAs). By leveraging publicly available health data and Census API data, we identified significant relationships between obesity, physical inactivity, short sleep duration, and other health indicators, revealing their interconnected impact on community well-being.

The correlation analysis demonstrated strong associations fair or poor self-rated health status among adults, physical inactivity, mental distress, highlighting the need for holistic interventions targeting multiple health dimensions simultaneously. Clustering further allows us to categorize counties into distinct risk groups, enabling policymakers to prioritize resources and strategies for high-risk areas. Logistic regression models provided predictive insights, identifying factors such as physical inactivity, lack of social support, and food insecurity as critical predictors of poor health outcomes.

The findings underscore the importance of targeted public health initiatives within cities and MSAs to address prevalent health challenges. Future work can extend this analysis by integrating additional socioeconomic variables, refining prediction models, and assessing the long-term effectiveness of health interventions. By adopting data-driven approaches, communities can make more informed decisions to improve public health outcomes and reduce disparities.

## 7 Acknowledgments

We would like to thank the United States Census Bureau for providing an API key to allow us to access the Census data.

## 8 Author Contributions

All authors are equal contributors to the project. Specifics are indicated below:

**Ryan** acquired the data and preprocessed the Census data. Additionally, Ryan generated the figures displayed throughout the paper. **Erin** preprocessed the PLACES data. Erin also organized the report. The task of performing the experiments was divided equally between both members.

## 9 Data and Code Availability

The code (with cleaned data) is publicly available at NOVA-Economics-on-Individual-Health. The unprocessed health data used for this project is publicly available at CDC Behavioral Risk Factor Surveillance System (BFRSS) [1]. The unprocessed Census data can be found at United States Census Bureau [16]. However, an API key must be requested to access this data.

## References

[1] Centers for Disease Control and Prevention, "Places: Local data for better health," 2023. Accessed: 2024-12-17.
[2] X. Chen, J. Li, and Z. Wu, "Correlation between physical inactivity, mental distress, and chronic diseases," *International Journal of Environmental Research and Public Health*, vol. 18, no. 17, p. 9271, 2021.
[3] K. Nguyen and E. Martin, "Health disparities in u.s. metropolitan areas: A focus on obesity and diabetes," *Preventing Chronic Disease*, vol. 20, p. E45, 2023.

[4] M. Booth, C. Packer, and E. R. Walsh, "Chronic disease disparities in metropolitan areas: The intersection of health and environment," *American Journal of Public Health*, vol. 111, no. 8, pp. 1402–1410, 2021.
[5] G. R. McCormack and M. Demers, "The role of social determinants in physical inactivity and obesity prevalence in urban populations," *Health and Place*, vol. 80, p. 103018, 2023.
[6] Y. Bai, L. Chang, and Z. Li, "Geographic disparities in obesity prevalence among u.s. adults," *Obesity Research & Clinical Practice*, vol. 16, no. 3, pp. 206–215, 2022.
[7] J. K. Harris, R. L. Smith, and L. T. Nguyen, "Urban health disparities: Examining chronic conditions and risk factors in metropolitan populations," *Journal of Urban Health*, vol. 99, no. 4, pp. 543–556, 2022.
[8] W. Raghupathi and V. Raghupathi, "Big data analytics in healthcare: Promise and potential," *Health Information Science and Systems*, vol. 2, no. 1, pp. 1–10, 2014.
[9] R. Shankar, "Healthcare big data and analytics: A perspective of the future," *Journal of Management Analytics*, vol. 5, no. 3, pp. 244–263, 2018.
[10] S. Pati *et al.*, "The association between social needs and chronic conditions in a diverse population," *Preventing Chronic Disease*, vol. 19, p. E44, 2022.
[11] G. K. Singh and M. Siahpush, "Urban health disparities and the role of socioeconomic status in chronic disease prevention," *Journal of Urban Health*, vol. 99, no. 3, pp. 450–463, 2022.
[12] Q. Zhang, M. Wang, and J. He, "The relationship between food insecurity and obesity in urban areas: A systematic review," *Public Health Nutrition*, vol. 26, no. 4, pp. 657–667, 2023.
[13] Y. Kim and S. P. Jones, "Food insecurity and health outcomes in metropolitan statistical areas," *Public Health Nutrition*, vol. 25, no. 12, pp. 3401–3410, 2022.
[14] D. Gonzalez, L. Allen, and S. Marquez, "Impact of social policies on health disparities in urban areas: A systematic review," *American Journal of Public Health*, vol. 113, no. 2, pp. 125–134, 2023.
[15] D. Cutler, E. Glaeser, and J. Vigdor, "The impact of urban poverty on health outcomes in the u.s.: A review of recent evidence," *Journal of Urban Health*, vol. 99, no. 1, pp. 12–25, 2022.
[16] U.S. Census Bureau, "Census data api: Acs 5-year data, 2019," 2019. Accessed: 2024-06-17.