SDOH Measures for ZCTA, ACS 2017-2021

RobertLupo

2024-05-28

# Abstract

A brief summary of the project.

This project aims to analyze Social Determinants of Health (SDOH) measures for ZIP Code Tabulation Areas (ZCTA) using the American Community Survey (ACS) data from 2017 to 2021. By cleaning and exploring the dataset, we seek to uncover trends and insights that can inform data-driven decision-making. The analysis employs various statistical techniques and predictive modeling to evaluate the impact of different measures on health outcomes.

# Introduction

Provide context and state the research question or hypothesis.

Social Determinants of Health (SDOH) are non-medical factors that influence health outcomes. These include socioeconomic status, education, neighborhood and physical environment, employment, social support networks, and access to healthcare. This project focuses on analyzing SDOH measures across various ZIP Code Tabulation Areas (ZCTAs) using data from the American Community Survey (ACS) between 2017 and 2021. The goal is to identify key factors that impact health outcomes and understand the geographical distribution of these determinants.

# Data Description

Describe the dataset and its source.

The dataset used in this project is sourced from the American Community Survey (ACS) and provides SDOH measures for ZIP Code Tabulation Areas (ZCTAs) from 2017 to 2021. The dataset includes variables such as data\_value, margin of error (moe), total population, latitude, longitude, and various categorical indicators. The data was cleaned to handle missing values and ensure consistency across measures.

# Methodology

Explain the statistical techniques and models applied.

The methodology for this project includes several steps: 1. Data Cleaning: Handling missing values and ensuring data consistency. 2. Exploratory Data Analysis (EDA): Generating summary statistics and visualizations to understand data distributions and relationships. 3. Correlation Analysis: Identifying relationships between numerical variables. 4. Predictive Modeling: Developing models to predict health outcomes based on SDOH measures. Models considered include linear regression, logistic regression, and support vector machines. 5. Model Evaluation: Assessing model performance using metrics such as R-squared, mean squared error, and accuracy, and employing cross-validation techniques.

# Results

Present findings from the analysis.

The analysis revealed several key insights: - The distribution of key numerical variables such as data\_value, moe, and total population. - Correlation analysis showed significant relationships between data\_value, latitude, longitude, and total population. - Predictive modeling identified important SDOH measures that influence health outcomes. The models were evaluated and validated to ensure reliability.

# Discussion

Interpret the results and their implications.

The findings suggest that certain SDOH measures have a significant impact on health outcomes. For instance, areas with higher total populations and specific geographical locations tend to have different health outcomes. These insights can inform public health strategies and resource allocation to address disparities. However, the analysis is limited by the quality and granularity of the data, and further research could expand on these findings by incorporating additional variables and more granular data.

# Conclusion

Summarize key findings and suggest future research.

In conclusion, this project highlights the importance of SDOH measures in understanding health outcomes. By analyzing ACS data from 2017 to 2021, we identified key factors that influence health and provided insights that can guide public health decision-making. Future research could explore more detailed datasets and consider additional variables to further refine these findings.

# References

Cite sources and data used in the project.

American Community Survey (ACS) data, 2017-2021.

# Load necessary libraries  
library(ggplot2)  
library(dplyr)

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

library(tidyr)  
library(sf)

## Linking to GEOS 3.11.2, GDAL 3.8.2, PROJ 9.3.1; sf\_use\_s2() is TRUE

library(cluster)  
library(factoextra)

## Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa

library(corrplot)

## corrplot 0.92 loaded

# Load the dataset  
data <- read.csv("C:/Users/Trey/Documents/GitHub/SDOH-Measures-for-ZCTA-ACS-2017-2021/bumh-rgsq\_version\_37.csv")  
  
# View the first few rows of the dataset  
head(data)

## year measure  
## 1 2017-2021 Persons of racial or ethnic minority status  
## 2 2017-2021 No high school diploma among adults aged 25 years or older  
## 3 2017-2021 No high school diploma among adults aged 25 years or older  
## 4 2017-2021 Crowding among housing units  
## 5 2017-2021 Persons living below 150% of the poverty level  
## 6 2017-2021 Unemployment among people 16 years and older in the labor force  
## data\_value\_unit categoryid short\_question\_text datasource moe  
## 1 % SDOH Racial or ethnic minority status 5-year ACS 9.4  
## 2 % SDOH No high school diploma 5-year ACS 4.1  
## 3 % SDOH No high school diploma 5-year ACS 5.6  
## 4 % SDOH Crowding 5-year ACS NA  
## 5 % SDOH Poverty 5-year ACS 5.0  
## 6 % SDOH Unemployment 5-year ACS 4.0  
## locationname locationid data\_value\_type data\_value  
## 1 97020 97020 Percentage 32.4  
## 2 95665 95665 Percentage 5.3  
## 3 97362 97362 Percentage 20.7  
## 4 97373 97373 Percentage NA  
## 5 95338 95338 Percentage 24.4  
## 6 97466 97466 Percentage 3.4  
## geolocation datavaluetypeid latitude category measureid  
## 1 POINT (-122.8388419 45.2226509) Percent 45.22265 SDOH REMNRTY  
## 2 POINT (-120.6527553 38.3951756) Percent 38.39518 SDOH NOHSDP  
## 3 POINT (-122.7565397 45.0620125) Percent 45.06201 SDOH NOHSDP  
## 4 POINT (-122.7764195 45.0579989) Percent 45.05800 SDOH CROWD  
## 5 POINT (-119.9837888 37.507011) Percent 37.50701 SDOH POV150  
## 6 POINT (-124.0852328 42.8489451) Percent 42.84895 SDOH UNEMP  
## longitude totalpopulation  
## 1 -122.8388 1200  
## 2 -120.6528 4650  
## 3 -122.7565 4174  
## 4 -122.7764 191  
## 5 -119.9838 10726  
## 6 -124.0852 1301

# Get a summary of the dataset  
summary(data)

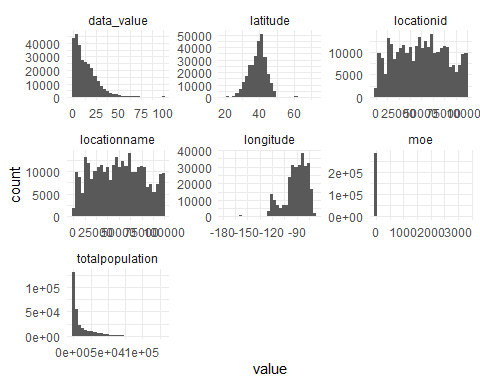
## year measure data\_value\_unit categoryid   
## Length:291024 Length:291024 Length:291024 Length:291024   
## Class :character Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character Mode :character   
##   
##   
##   
##   
## short\_question\_text datasource moe locationname   
## Length:291024 Length:291024 Min. : 0.000 Min. : 1001   
## Class :character Class :character 1st Qu.: 2.600 1st Qu.:27249   
## Mode :character Mode :character Median : 4.800 Median :49761   
## Mean : 9.205 Mean :49742   
## 3rd Qu.: 9.800 3rd Qu.:72012   
## Max. :3290.900 Max. :99929   
## NA's :1928   
## locationid data\_value\_type data\_value geolocation   
## Min. : 1001 Length:291024 Min. : 0.00 Length:291024   
## 1st Qu.:27249 Class :character 1st Qu.: 3.90 Class :character   
## Median :49761 Mode :character Median : 10.50 Mode :character   
## Mean :49742 Mean : 14.53   
## 3rd Qu.:72012 3rd Qu.: 20.60   
## Max. :99929 Max. :100.00   
## NA's :1834   
## datavaluetypeid latitude category measureid   
## Length:291024 Min. :19.07 Length:291024 Length:291024   
## Class :character 1st Qu.:35.46 Class :character Class :character   
## Mode :character Median :39.54 Mode :character Mode :character   
## Mean :38.90   
## 3rd Qu.:42.12   
## Max. :71.25   
##   
## longitude totalpopulation   
## Min. :-176.67 Min. : 50   
## 1st Qu.: -97.14 1st Qu.: 778   
## Median : -88.20 Median : 2955   
## Mean : -91.02 Mean : 10196   
## 3rd Qu.: -80.32 3rd Qu.: 14012   
## Max. : -67.01 Max. :130352   
##

# Calculate summary statistics for numeric columns only  
summary\_statistics <- data %>%  
 summarise(across(where(is.numeric), list(mean = ~mean(. , na.rm = TRUE),  
 sd = ~sd(. , na.rm = TRUE),  
 min = ~min(. , na.rm = TRUE),  
 max = ~max(. , na.rm = TRUE))))  
print(summary\_statistics)

## moe\_mean moe\_sd moe\_min moe\_max locationname\_mean locationname\_sd  
## 1 9.205388 17.9028 0 3290.9 49741.96 27342.93  
## locationname\_min locationname\_max locationid\_mean locationid\_sd  
## 1 1001 99929 49741.96 27342.93  
## locationid\_min locationid\_max data\_value\_mean data\_value\_sd data\_value\_min  
## 1 1001 99929 14.528 14.96427 0  
## data\_value\_max latitude\_mean latitude\_sd latitude\_min latitude\_max  
## 1 100 38.89899 5.176886 19.06555 71.25386  
## longitude\_mean longitude\_sd longitude\_min longitude\_max totalpopulation\_mean  
## 1 -91.01803 14.96048 -176.6686 -67.00916 10196.13  
## totalpopulation\_sd totalpopulation\_min totalpopulation\_max  
## 1 14960.75 50 130352

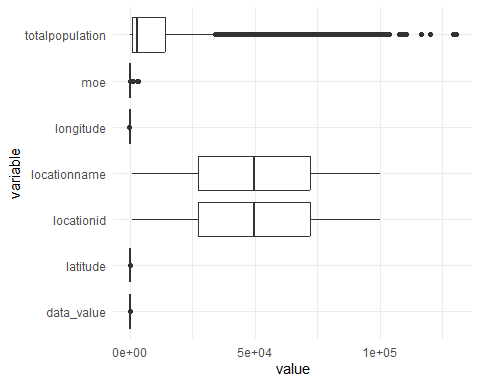
# Histograms  
data %>%  
 select(where(is.numeric)) %>%  
 gather(key = "variable", value = "value") %>%  
 ggplot(aes(x = value)) +  
 geom\_histogram(bins = 30) +  
 facet\_wrap(~variable, scales = "free") +  
 theme\_minimal()

## Warning: Removed 3762 rows containing non-finite outside the scale range  
## (`stat\_bin()`).

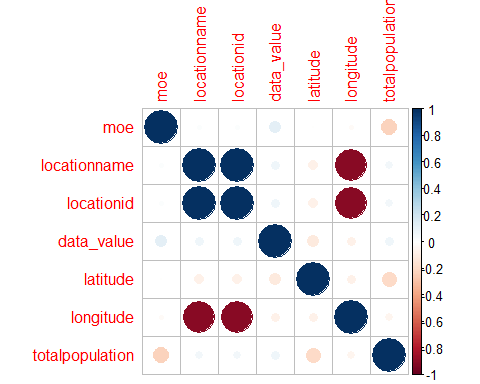


# Box plots  
data %>%  
 select(where(is.numeric)) %>%  
 gather(key = "variable", value = "value") %>%  
 ggplot(aes(x = variable, y = value)) +  
 geom\_boxplot() +  
 theme\_minimal() +  
 coord\_flip()

## Warning: Removed 3762 rows containing non-finite outside the scale range  
## (`stat\_boxplot()`).



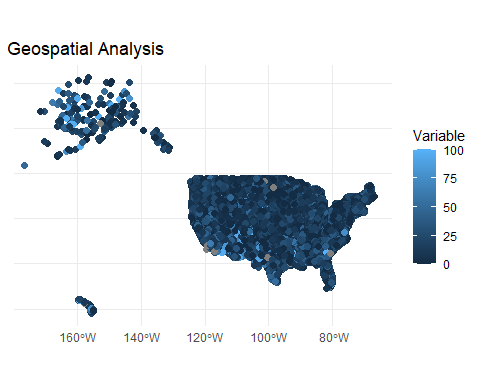
# Compute correlation matrix for numeric columns only  
correlation\_matrix <- cor(data %>% select(where(is.numeric)), use = "complete.obs")  
  
# Visualize the correlation matrix  
corrplot::corrplot(correlation\_matrix, method = "circle")



# Trend analysis: Assuming there is a 'year' column in the dataset (this column needs to be adjusted based on actual data)  
if ("year" %in% colnames(data)) {  
 data %>%  
 gather(key = "variable", value = "value", -c(year)) %>%  
 ggplot(aes(x = year, y = value, color = variable)) +  
 geom\_line() +  
 facet\_wrap(~variable, scales = "free\_y") +  
 theme\_minimal()  
}



# Geospatial analysis: Assuming there are 'latitude' and 'longitude' columns  
if ("latitude" %in% colnames(data) & "longitude" %in% colnames(data)) {  
 data\_sf <- st\_as\_sf(data, coords = c("longitude", "latitude"), crs = 4326)  
  
 # Plotting the spatial data (replace 'data\_value' with an actual variable name)  
 ggplot(data\_sf) +  
 geom\_sf(aes(color = data\_value), size = 2) + # Change 'data\_value' to the variable you want to visualize  
 theme\_minimal() +  
 labs(title = "Geospatial Analysis", color = "Variable")  
}

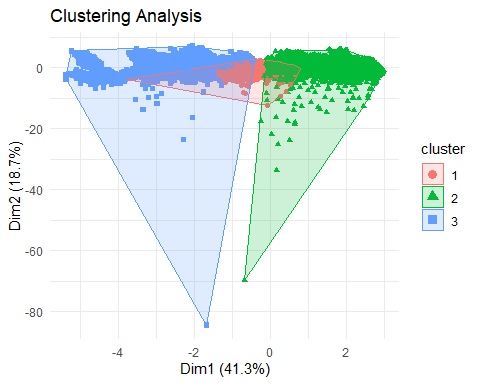


# Normalize the data for clustering (excluding non-numeric columns)  
data\_scaled <- scale(data %>% select(where(is.numeric)) %>% na.omit())  
  
# Perform clustering  
set.seed(123)  
kmeans\_result <- kmeans(data\_scaled, centers = 3, nstart = 25)

## Warning: Quick-TRANSfer stage steps exceeded maximum (= 14454800)

## Warning: Quick-TRANSfer stage steps exceeded maximum (= 14454800)

# Visualize clustering result  
fviz\_cluster(kmeans\_result, data = data\_scaled, geom = "point", stand = FALSE) +  
 theme\_minimal() +  
 labs(title = "Clustering Analysis")



# Regression analysis: Assuming 'data\_value' is the outcome variable and others are predictors  
model <- lm(data\_value ~ ., data = data %>% select(where(is.numeric)))  
summary(model)

##   
## Call:  
## lm(formula = data\_value ~ ., data = data %>% select(where(is.numeric)))  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -310.136 -9.937 -3.677 6.388 89.693   
##   
## Coefficients: (1 not defined because of singularities)  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.512e+01 3.157e-01 47.87 <2e-16 \*\*\*  
## moe 1.107e-01 1.566e-03 70.66 <2e-16 \*\*\*  
## locationname -4.556e-05 2.475e-06 -18.41 <2e-16 \*\*\*  
## locationid NA NA NA NA   
## latitude -3.535e-01 5.749e-03 -61.49 <2e-16 \*\*\*  
## longitude -1.502e-01 4.538e-03 -33.11 <2e-16 \*\*\*  
## totalpopulation 7.160e-05 1.917e-06 37.34 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 14.65 on 289090 degrees of freedom  
## (1928 observations deleted due to missingness)  
## Multiple R-squared: 0.04058, Adjusted R-squared: 0.04057   
## F-statistic: 2446 on 5 and 289090 DF, p-value: < 2.2e-16

# Visualize regression diagnostics  
par(mfrow = c(2, 2))  
plot(model)

