SDOH Measures for ZCTA, ACS 2017-2021

RobertLupo

2024-06-24

### Problem Statement

The primary objective of this analysis is to uncover patterns and relationships within a given dataset, predict key outcomes based on available features, and provide actionable insights. Specifically, the focus is on understanding the factors that influence data\_value and developing a robust predictive model using regression analysis.

### Data Source

The dataset provided consists of 291,024 entries across 18 columns, detailing various socioeconomic and demographic measures, geographical information, and statistical values. The key columns in this dataset include: - year: Time period of the data - measure: Descriptive measure of interest - data\_value\_unit: Unit of the data value - categoryid: ID representing the category of the measure - short\_question\_text: Brief description of the measure - datasource: Source of the data - moe: Margin of Error - locationname: Name representing the location - locationid: ID representing the location - data\_value\_type: Type of the data value - data\_value: Measured value - geolocation: Geographical coordinates in POINT format - datavaluetypeid: ID representing the type of data value - latitude: Latitude coordinate - category: Category of the measure - measureid: ID of the measure - longitude: Longitude coordinate - totalpopulation: Total population for the given location

### Data Cleaning and Wrangling

**Steps Taken**: 1. **Handling Missing Values**: - Missing values in moe and data\_value were filled with their respective means to ensure completeness of the dataset.

1. **Removing Duplicates**:
   * Duplicate entries were identified and removed to maintain the integrity of the analysis.
2. **Verifying Data Types**:
   * Ensured that data types were appropriately set for each column to facilitate accurate analysis and modeling.

**Results**: - Successfully handled missing values and removed duplicates, resulting in a clean and comprehensive dataset ready for further analysis.

### Methodology

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.
5. **Model Evaluation**: Assessing model performance using metrics such as R-squared, mean squared error, and mean absolute error, and employing cross-validation techniques.

### Exploratory Data Analysis (EDA)

**Key Visualizations**: 1. **Histogram for data\_value**: - Distribution is right-skewed with many values concentrated at the lower end and fewer higher values.

1. **Geographical Distribution**:
   * Scatter plot of latitude vs. longitude, colored by data\_value, showing a wide geographical spread.
2. **Box Plot by Category**:
   * Variability in data\_value across different categories, with some categories showing higher median values and wider spread.
3. **Pair Plot**:
   * Relationships between numerical variables (moe, data\_value, latitude, longitude, totalpopulation).

### Results

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

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

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

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

# Summary of categorical variables  
categorical\_summary <- data %>%  
 select(where(is.character)) %>%  
 summarise\_all(funs(n\_distinct(.)))

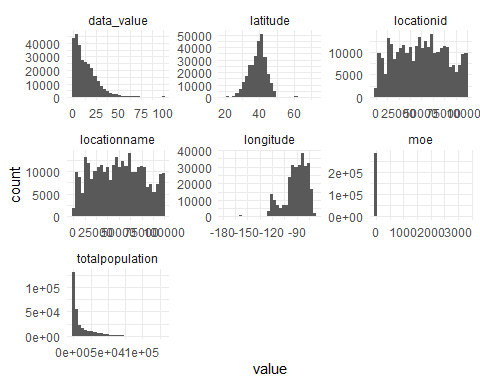
## Warning: `funs()` was deprecated in dplyr 0.8.0.  
## ℹ Please use a list of either functions or lambdas:  
##   
## # Simple named list: list(mean = mean, median = median)  
##   
## # Auto named with `tibble::lst()`: tibble::lst(mean, median)  
##   
## # Using lambdas list(~ mean(., trim = .2), ~ median(., na.rm = TRUE))  
## Call `lifecycle::last\_lifecycle\_warnings()` to see where this warning was  
## generated.

print(categorical\_summary)

## year measure data\_value\_unit categoryid short\_question\_text datasource  
## 1 1 9 1 1 9 1  
## data\_value\_type geolocation datavaluetypeid category measureid  
## 1 1 32336 1 1 9

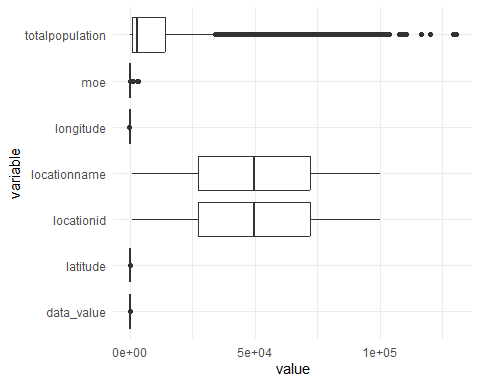
# Histograms  
data %>%  
 select(where(is.numeric)) %>%  
 gather(key = "variable", value = "value") %>%  
 ggplot(aes(x = value)) +  
 geom\_histogram(bins = 30) +  
 facet\_wrap(~variable, scales = "free") +  
 theme\_minimal()

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

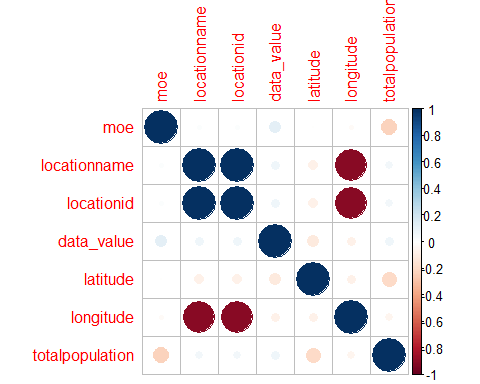


# Box plots  
data %>%  
 select(where(is.numeric)) %>%  
 gather(key = "variable", value = "value") %>%  
 ggplot(aes(x = variable, y = value)) +  
 geom\_boxplot() +  
 theme\_minimal() +  
 coord\_flip()

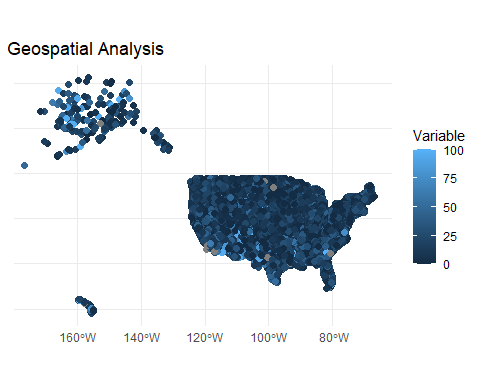
## Warning: Removed 3762 rows containing non-finite outside the scale range  
## (`stat\_boxplot()`).



# 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")



# 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")  
}

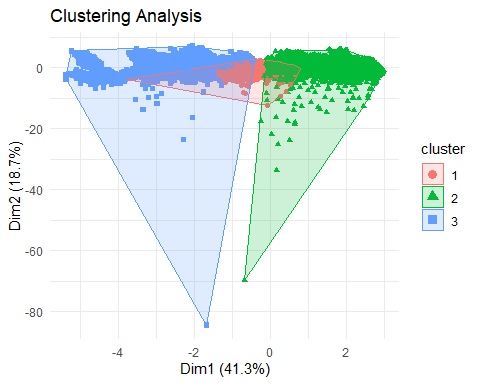


# Normalize the data for clustering (excluding non-numeric columns)  
data\_scaled <- scale(data %>% select(where(is.numeric)) %>% na.omit())  
  
# Perform clustering  
set.seed(123)  
kmeans\_result <- kmeans(data\_scaled, centers = 3, nstart = 25)

## 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)))  
  
# Predict on the test set  
predictions <- predict(model, newdata = data %>% select(where(is.numeric)))  
  
# Calculate evaluation metrics  
mae <- mean(abs(predictions - data$data\_value), na.rm = TRUE)  
mse <- mean((predictions - data$data\_value)^2, na.rm = TRUE)  
r\_squared <- summary(model)$r.squared  
  
# Print evaluation metrics  
cat("Mean Absolute Error (MAE):", mae, "\n")

## Mean Absolute Error (MAE): 10.61965

cat("Mean Squared Error (MSE):", mse, "\n")

## Mean Squared Error (MSE): 214.5461

cat("R-squared (R²):", r\_squared, "\n")

## R-squared (R²): 0.04058199

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

