Northeastern University
Mechanical and Industrial Engineering Department
IE 6600: Computation and Visualization
Professor Sivarit (Tony) Sultornsanee
Summer 1 2024

Project 3: Analysis of Health Profiles in the USA

Project Report

Group 6

Team Members

Team Members	NUID
Sanjana Rao	002473503
Atharv Nirhali	002202811
Snehal Yadav	002416545
Shreya Ale	002192493

Table of Contents

Contents

Introduction

The "500 Cities: Local Data for Better Health" dataset is an essential resource for analyzing health dynamics in urban areas across the United States. It includes extensive data on health outcomes, preventive measures, and unhealthy behaviors for 500 cities and various census tracts. The choice to utilize this dataset is driven by its ability to provide detailed insights into the health challenges and successes of urban populations at a local level. It offers critical information on conditions such as obesity, smoking rates, and access to preventive care, enabling public health officials and policymakers to identify areas needing intervention and areas of achievement.

Turning this dataset into meaningful visualizations presents an exciting opportunity. Using tools like Plotly to create interactive maps and comprehensive data graphics, the goal is to make the complex data clear and understandable. These visualizations go beyond simple charts and maps; they act as a powerful storytelling medium that can illustrate trends, compare health metrics across regions, and inform decision-making. The visual representations will enable stakeholders to engage with the data, explore various health indicators, and derive insights essential for developing effective public health strategies. This visualization effort aims to make the data not only accessible but also actionable, facilitating targeted health interventions and improved public health outcomes.

Figure 1: 500 Cities: Local Data for Better Health

Project Objectives

- The project aims to analyze the NYPD Hate Crimes dataset to uncover trends and patterns in hate crimes over recent years. By examining data such as the geographical distribution, bias motives, and offense categories, we seek to identify areas with higher incidences and the most common types of bias-related offenses. Additionally, we will study the relationship between reported hate crimes and arrests to understand law enforcement responses.
- This analysis will help in understanding the temporal patterns, comparing data across years, and potentially predicting future hotspots, thereby providing valuable insights for policymakers and law enforcement agencies.

Raw Data

Geospatial dataset sourced from data.gov (*Figure 2*).
• Contains over 810,000 records with 24 different variables.
• Includes metrics related to health outcomes, unhealthy behaviors, and preventive measures.

• Examples of measures: diagnosed diabetes, current smoking, lack of health insurance, obesity.
• Key columns: State and City Descriptors, Geolocation, Health Measures (e.g., prevalence of diseases, health behaviors), Confidence Limits, and Population Counts.
• Data is mostly complete; notable missing values include Geolocation (56 missing), City Name (56 missing), and TractFIPS (28,056 missing).

Year	StateAbbr	StateDesc	CityName	GeographicL	DataSource	Category	UniqueID	Measure	Data_Value_	DataValueTy	Data_Value_	Data_Value	Low_Confide	High_Confid	Data_Value_	Data_Value_	PopulationCo	GeoLocation	CategoryID	MeasureId	CityFIPS	TractFIPS	Short_Question_Text
2017	CA	California	Hawthorne	Census Tract	BRFSS	Health Outco	0632548-06(	Arthritis amo	%	CrdPrv	Crude prevale	14.6	13.9	15.2			4407	(33.9055479	HLTHOUT	ARTHRITIS	632548	6.038E+09	Arthritis
2017	CA	California	Hawthorne	City	BRFSS	Unhealthy Be	632548	Current smok	%	CrdPrv	Crude prevale	15.4	15	15.9			84293	(33.9146677	UNHBEH	CSMOKING	632548		Current Smoking
2017	CA	California	Hayward	City	BRFSS	Health Outco	633000	Coronary hea	%	AgeAdjPrv	Age-adjustec	4.8	4.7	4.8			144186	(37.6329591	HLTHOUT	CHD	633000		Coronary Heart Disease
2017	CA	California	Hayward	City	BRFSS	Unhealthy Be	633000	Obesity amo	%	CrdPrv	Crude prevale	24.2	24.1	24.4			144186	(37.6329591	UNHBEH	OBESITY	633000		Obesity
2017	CA	California	Hemet	City	BRFSS	Prevention	633182	Cholesterol s	%	AgeAdjPrv	Age-adjustec	78	77.6	78.3			78657	(33.7352277	PREVENT	CHOLSCREEN	633182		Cholesterol Screening
2017	CA	California	Indio	Census Tract	BRFSS	Health Outco	0636448-06(	Arthritis amo	%	CrdPrv	Crude prevale	22	21.1	22.8			5006	(33.7144617	HLTHOUT	ARTHRITIS	636448	6.065E+09	Arthritis
2017	CA	California	Indio	City	BRFSS	Unhealthy Be	636448	Binge drinkin	%	AgeAdjPrv	Age-adjustec	17.7	17.5	17.9			76036	(33.7298067	UNHBEH	BINGE	636448		Binge Drinking
2017	CA	California	Indio	City	BRFSS	Health Outco	636448	Chronic obstr	%	AgeAdjPrv	Age-adjustec	6	5.8	6.2			76036	(33.7298067	HLTHOUT	COPD	636448		COPD
2017	CA	California	Inglewood	Census Tract	BRFSS	Health Outco	0636546-06(	Diagnosed di	%	CrdPrv	Crude prevale	12.7	12	13.5			2472	(33.9439711	HLTHOUT	DIABETES	636546	6.038E+09	Diabetes
2016	CA	California	Inglewood	City	BRFSS	Prevention	636546	Mammograp	%	CrdPrv	Crude prevale	82.5	82	83			109673	(33.9565748	PREVENT	MAMMOUS	636546		Mammography
2017	CA	California	Lakewood	City	BRFSS	Unhealthy Be	639892	Obesity amo	%	CrdPrv	Crude prevale	22.1	21.9	22.2			80048	(33.8470531	UNHBEH	OBESITY	639892		Obesity
2016	CA	California	Lakewood	City	BRFSS	Health Outco	639892	All teeth lost	%	CrdPrv	Crude prevale	8.2	7.5	8.9			80048	(33.8470531	HLTHOUT	TEETHLOST	639892		Teeth Loss
2017	CA	California	Livermore	City	BRFSS	Health Outco	641992	Current asthr	%	CrdPrv	Crude prevale	8.9	8.8	9			80968	(37.6865101	HLTHOUT	CASTHMA	641992		Current Asthma
2017	US	United States		US	BRFSS	Prevention	59	Current lack c	%	AgeAdjPrv	Age-adjustec	15.2	14.9	15.5			308745538		PREVENT	ACCESS2			Health Insurance
2017	AL	Alabama	Hoover	City	BRFSS	Health Outco	135896	Chronic kidne	%	CrdPrv	Crude prevale	2.4	2.4	2.5			81619	(33.3767602	HLTHOUT	KIDNEY	135896		Chronic Kidney Disease
2016	AL	Alabama	Hoover	Census Tract	BRFSS	Prevention	0135896-01(	Mammograp	%	CrdPrv	Crude prevale	81.9	78.7	84.5			1636	(33.3923792	PREVENT	MAMMOUS	135896	1.073E+09	Mammography
2016	AL	Alabama	Hoover	City	BRFSS	Health Outco	135896	All teeth lost	%	CrdPrv	Crude prevale	8.3	7.3	9.3			81619	(33.3767602	HLTHOUT	TEETHLOST	135896		Teeth Loss
2017	AL	Alabama	Huntsville	City	BRFSS	Health Outco	137000	Coronary hea	%	CrdPrv	Crude prevale	6.7	6.6	6.8			180105	(34.6989692	HLTHOUT	CHD	137000		Coronary Heart Disease
2017	AL	Alabama	Huntsville	Census Tract	BRFSS	Health Outco	0137000-01(	Diagnosed di	%	CrdPrv	Crude prevale	9.3	8.5	10.3			4387	(34.6127555	HLTHOUT	DIABETES	137000	1.089E+09	Diabetes
2017	AL	Alabama	Huntsville	Census Tract	BRFSS	Unhealthy Be	0137000-01(	Obesity amo	%	CrdPrv	Crude prevale	30.3	29.2	31.5			2654	(34.7636374	UNHBEH	OBESITY	137000	1.089E+09	Obesity
2017	AL	Alabama	Huntsville	City	BRFSS	Health Outco	137000	Stroke amon	%	CrdPrv	Crude prevale	3.6	3.5	3.6			180105	(34.6989692	HLTHOUT	STROKE	137000		Stroke
2017	AL	Alabama	Mobile	City	BRFSS	Unhealthy Be	150000	Obesity amo	%	AgeAdjPrv	Age-adjustec	38.2	38	38.3			195111	(30.6776248	UNHBEH	OBESITY	150000		Obesity
2017	AL	Alabama	Montgomery	City	BRFSS	Prevention	151000	Cholesterol s	%	CrdPrv	Crude prevale	80.2	79.9	80.4			205764	(32.3472645	PREVENT	CHOLSCREEN	151000		Cholesterol Screening
2016	AL	Alabama	Montgomery	City	BRFSS	Prevention	151000	Visits to dent	%	CrdPrv	Crude prevale	60.6	60	61.2			205764	(32.3472645	PREVENT	DENTAL	151000		Dental Visit
2017	AL	Alabama	Tuscaloosa	City	BRFSS	Health Outco	177256	Stroke amon	%	AgeAdjPrv	Age-adjustec	4.2	4.1	4.3			90468	(33.2336083	HLTHOUT	STROKE	177256		Stroke
2017	AK	Alaska	Anchorage	Census Tract	BRFSS	Health Outco	0203000-02(	Diagnosed di	%	CrdPrv	Crude prevale	7.4	6.9	7.8			4993	(61.1593945	HLTHOUT	DIABETES	203000	2.02E+09	Diabetes
2016	AK	Alaska	Anchorage	Census Tract	BRFSS	Prevention	0203000-02(	Mammograp	%	CrdPrv	Crude prevale	74	69.4	77.9			5217	(61.0562469	PREVENT	MAMMOUS	203000	2.02E+09	Mammography
2017	AZ	Arizona	Avondale	Census Tract	BRFSS	Unhealthy Be	0404720-04(	Obesity amo	%	CrdPrv	Crude prevale	30.6	29.6	31.5			3978	(33.4505348	UNHBEH	OBESITY	404720	4.013E+09	Obesity
2017	AZ	Arizona	Chandler	City	BRFSS	Health Outco	412000	Coronary hea	%	CrdPrv	Crude prevale	3.8	3.8	3.9			236123	(33.2831898	HLTHOUT	CHD	412000		Coronary Heart Disease
2017	AZ	Arizona	Chandler	City	BRFSS	Unhealthy Be	412000	No leisure-tir	%	CrdPrv	Crude prevale	20.9	20.6	21.2			236123	(33.2831898	UNHBEH	LPA	412000		Physical Inactivity
2017	AZ	Arizona	Glendale	City	BRFSS	Unhealthy Be	427820	Current smok	%	CrdPrv	Crude prevale	18.7	18.4	19.1			226721	(33.5796123	UNHBEH	CSMOKING	427820		Current Smoking
2016	AZ	Arizona	Glendale	Census Tract	BRFSS	Prevention	0427820-04(	Mammograp	%	CrdPrv	Crude prevale	81.2	77.7	83.5			4362	(33.6610843	PREVENT	MAMMOUS	427820	4.014E+09	Mammography
2016	AZ	Arizona	Glendale	City	BRFSS	Unhealthy Be	427820	Sleeping less	%	CrdPrv	Crude prevale	37.3	37.2	37.5			226721	(33.5796123	UNHBEH	SLEEP	427820		Sleep <7 hours
2016	AZ	Arizona	Peoria	City	BRFSS	Prevention	454050	Mammograp	%	CrdPrv	Crude prevale	79.2	78.7	79.7			154065	(33.7847206	PREVENT	MAMMOUS	454050		Mammography
2016	AZ	Arizona	Phoenix	Census Tract	BRFSS	Prevention	0455000-04(	Mammograp	%	CrdPrv	Crude prevale	78.7	75.9	81.3			3849	(33.6510442	PREVENT	MAMMOUS	455000	4.014E+09	Mammography
2017	AZ	Arizona	Phoenix	Census Tract	BRFSS	Health Outco	0455000-04(	Stroke amon	%	CrdPrv	Crude prevale	3.4	3	3.8			2655	(33.4658545	HLTHOUT	STROKE	455000	4.013E+09	Stroke
2017	AZ	Arizona	Scottsdale	City	BRFSS	Prevention	465000	Cholesterol s	%	AgeAdjPrv	Age-adjustec	82.7	82.5	82.9			217385	(33.6872493	PREVENT	CHOLSCREEN	465000		Cholesterol Screening
2017	AZ	Arizona	Surprise	City	BRFSS	Health Outco	471510	Coronary hea	%	CrdPrv	Crude prevale	5.9	5.8	6.1			117517	(33.6803835	HLTHOUT	CHD	471510		Coronary Heart Disease
2017	AZ	Arizona	Tempe	Census Tract	BRFSS	Health Outco	0473000-04(	Diagnosed di	%	CrdPrv	Crude prevale	7	6.4	7.9			2225	(33.3419171	HLTHOUT	DIABETES	473000	4.013E+09	Diabetes
2017	AZ	Arizona	Tucson	Census Tract	BRFSS	Health Outco	0477000-04(	High blood pi	%	CrdPrv	Crude prevale	21.9	20.9	22.9			2138	(32.2851840	HLTHOUT	BPHIGH	477000	4.019E+09	High Blood Pressure
2017	AZ	Arizona	Yuma	City	BRFSS	Health Outco	485540	Diagnosed di	%	CrdPrv	Crude prevale	11.1	11	11.3			93064	(32.5986027	HLTHOUT	DIABETES	485540		Diabetes
2017	AR	Arkansas	Jonesboro	City	BRFSS	Health Outco	535710	High blood pi	%	CrdPrv	Crude prevale	33.3	32.9	33.6			67263	(35.8208121	HLTHOUT	BPHIGH	535710		High Blood Pressure

Figure 2: Raw Data from the dataset

Data Pre-Processing

The initial data cleaning and preparation analysis (*Figure 3*) indicates the following:
• **Missing Values**: Most columns have no missing values; however, there are significant gaps in the Data_Value, Low_Confidence_Limit, High_Confidence_Limit columns, and extensive missing data in Data_Value_Footnote and Data_Value_Footnote_Symbol. Additionally, the Geolocation, CityName, and CityFIPS columns each have 56 missing entries, while TractFIPS has 28,056 missing entries, highlighting a lack of specific location details in some records.
• **Duplicates**: The dataset contains no duplicate rows, which maintains the integrity of the data.
• **Geolocation Column**: The Geolocation column, essential for mapping, has 56 missing entries. This column must be formatted appropriately for geospatial analysis, either as separate latitude and longitude columns or as a single column in a tuple format compatible with mapping libraries.

```
Missing Values:
 Year                            0
StateAbbr                        0
StateDesc                        0
CityName                        56
GeographicLevel                  0
DataSource                       0
Category                         0
UniqueID                         0
Measure                          0
Data_Value_Unit                  0
DataValueTypeID                  0
Data_Value_Type                  0
Data_Value                   22792
Low_Confidence_Limit         22792
High_Confidence_Limit        22792
Data_Value_Footnote_Symbol   787309
Data_Value_Footnote          787309
PopulationCount                  0
GeoLocation                     56
CategoryID                       0
MeasureId                        0
CityFIPS                        56
TractFIPS                    28056
Short_Question_Text              0
dtype: int64

Duplicate Values: 0

Missing Geolocation Values: 56
```

Figure 3: Screenshot showing summary of missing values

```
# Load the dataset
file_path = '500_Cities__Local_Data.csv'
data = pd.read_csv(file_path)

# Initial data summary
final_summary = {
    'total_rows': data.shape[0],
    'total_columns': data.shape[1],
    'column_names': data.columns
}

print(final_summary)

# Extract latitude and longitude from the 'GeoLocation' column
data[['Latitude', 'Longitude']] = data['GeoLocation'].str.extract(r'\(([^,]+), ([^)]+)\)')

# Convert latitude and longitude to numeric values
data['Latitude'] = pd.to_numeric(data['Latitude'], errors='coerce')
data['Longitude'] = pd.to_numeric(data['Longitude'], errors='coerce')

# Columns to Drop
columns_to_drop = [
    'Data_Value_Footnote_Symbol', 'Data_Value_Footnote',
    'Low_Confidence_Limit', 'High_Confidence_Limit',
    'CityFIPS', 'TractFIPS', 'UniqueID'
]
data_cleaned = data.drop(columns=columns_to_drop)

# Stratified Sampling based on CityName
# Calculate the number of unique cities
num_cities = data_cleaned['CityName'].nunique()

# Determine the sample size per city to approximately reduce the dataset to 20,000 rows
sample_size_per_city = 20000 // num_cities

# Perform stratified sampling
sampled_data = data_cleaned.groupby('CityName', group_keys=False).apply(lambda x: x.sample(min(len(x), sample_size_per_city)))

# Display a sample of the cleaned data
print(sampled_data[['Short_Question_Text', 'Latitude', 'Longitude']].head())

# Save the cleaned and sampled dataset to a new CSV file
sampled_data.to_csv('Cleaned_Sampled_Cities_Data_1.csv', index=False)

# Display the final data summary
final_sample_summary = {
    'total_rows': sampled_data.shape[0],
    'total_columns': sampled_data.shape[1],
    'column_names': sampled_data.columns
}

print(final_sample_summary)
```

Figure 4: Screenshot of data preparation

Data Visualization

Further, we went on to draw some insights from the dataset to enhance our understanding of the dataset.

1. Trends in Unhealthy Behaviors:

The below interactive graph (*Figure 5*) provides a detailed visual representation of adult smoking rates across various cities and census tracts in the United States. This visualization is key to understanding geographical variations in smoking behavior and identifying areas with notably high or low prevalence rates.

Insights from Specific Locations:

- **New York, NY:** Smoking rates in New York City vary significantly across boroughs. For example, Staten Island has higher smoking rates compared to Manhattan, likely due to demographic differences and socioeconomic conditions. This visualization helps stakeholders identify areas within the city where smoking cessation programs are most needed.
- **Los Angeles, CA:** In Los Angeles, the data reveals lower smoking rates in affluent neighborhoods such as West Hollywood and Santa Monica, while higher rates are observed in some eastern parts of the city. This suggests a need for targeted public health messaging and support for smoking cessation in these areas.

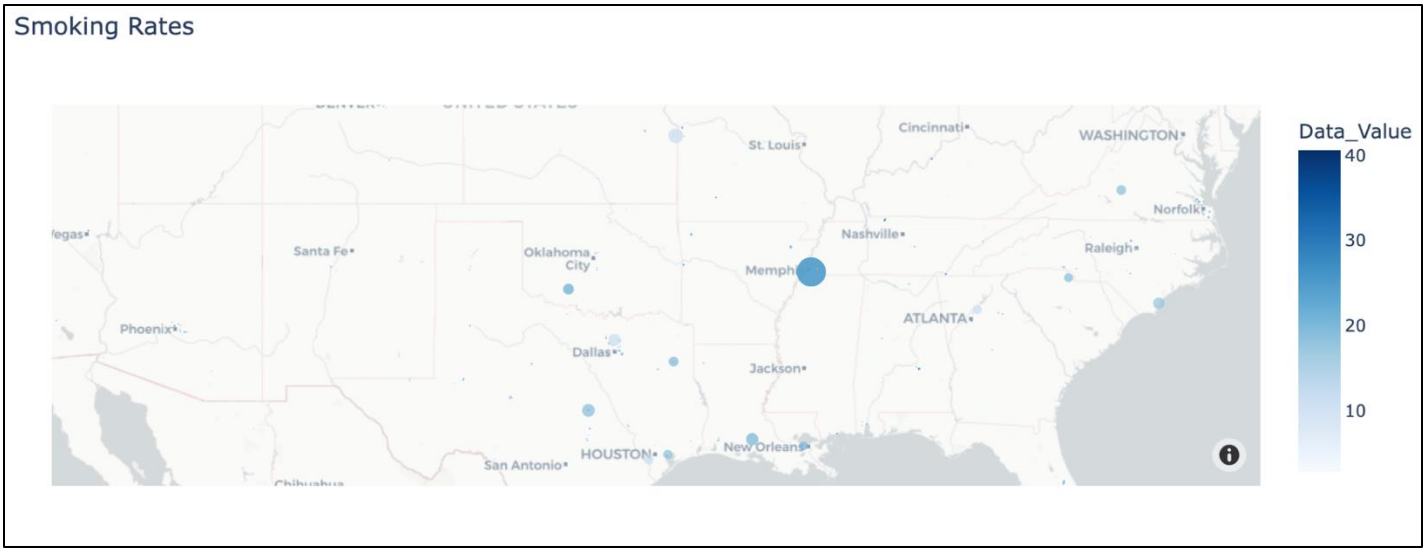

Figure 5: Figure plotting trends in Unhealthy Behaviors

- **Chicago, IL:** The map of Chicago shows higher smoking rates in the south and west sides of the city, regions known for various socioeconomic challenges. This correlation between economic hardship and smoking prevalence indicates that resources and support should be directed to these neighborhoods.

Overall Results and Insights:

- **State Comparisons:** At the state level, the visualization shows that states like Kentucky and West Virginia have higher overall smoking rates compared to states such as Utah and California. This is consistent with broader public health data, which links higher tobacco use to regions with historical ties to tobacco farming and less stringent anti-smoking regulations.
- **Urban vs. Rural:** Urban areas generally exhibit varied smoking rates, with city centers often showing lower rates due to more anti-smoking ordinances and available alternative activities. In contrast, rural areas or smaller towns typically have higher smoking rates, making them a focus for state health departments.
- **Effectiveness of Public Health Policies:** By analyzing trends over time, stakeholders can evaluate the effectiveness of public health policies aimed at reducing smoking rates. This includes assessing the impact of tobacco taxes, smoking bans in public areas, and public health campaigns.

2. Obesity Rates:

The interactive graph (*Figure 6*) provides an illuminating view into the prevalence of obesity across different cities and census tracts in the United States. This visualization helps highlight geographic disparities in obesity rates and identifies areas that may require more focused public health interventions.

Insights from New Locations:

- Houston, TX: The obesity rates in Houston are notably higher in the eastern parts of the city, which are also areas characterized by lower income levels and reduced access to healthcare services. This suggests that socioeconomic factors play a significant role in obesity prevalence and highlights the need for community-based health programs focusing on diet, exercise, and education in these areas.

Figure 6: Obesity Rates

- San Francisco, CA: Contrasting sharply with Houston, San Francisco exhibits lower overall obesity rates, particularly in the central and northern parts of the city. This could be attributed to higher socio-economic status, greater access to healthy food options, and a culture that promotes active lifestyles. However, there are pockets of higher rates in the southern districts, which could benefit from targeted health initiatives.
- Atlanta, GA: In Atlanta, higher obesity rates are observed in the suburban areas surrounding the city, compared to the central urban area. This pattern may reflect urban lifestyle factors such as walking and the availability of different food choices. It underscores the potential impact of urban planning in promoting health through recreational spaces and improved food landscapes.

Overall Results and Insights:

- State Comparisons:
  The visualization allows for comparisons between states, showing higher obesity rates in Southern states like Mississippi and Alabama and lower rates in states like Colorado and Vermont. This geographic pattern underscores the influence of both cultural diet preferences and outdoor activity opportunities available in different regions.
  a. Impact of Urbanization:
     Urban centers with high walkability scores and extensive public transit systems tend to have lower obesity rates. This correlation suggests that urban planning and infrastructure development can significantly influence public health outcomes by encouraging more active lifestyles.
  b. Public Health Policy Effectiveness:
     The data provides a basis to analyze the effectiveness of public health policies targeting obesity. For instance, areas with active public health campaigns promoting nutrition and physical activity can be compared against those without such initiatives to measure impact.
  c. Economic Disparities:
     There is a clear link between obesity rates and economic conditions within cities and states. Lower-income areas often have higher obesity rates, possibly due to less access to healthy food options and safe, accessible places for physical activity. This calls for policies that address economic barriers to healthy living.

3. State-wise data distribution on Obesity:

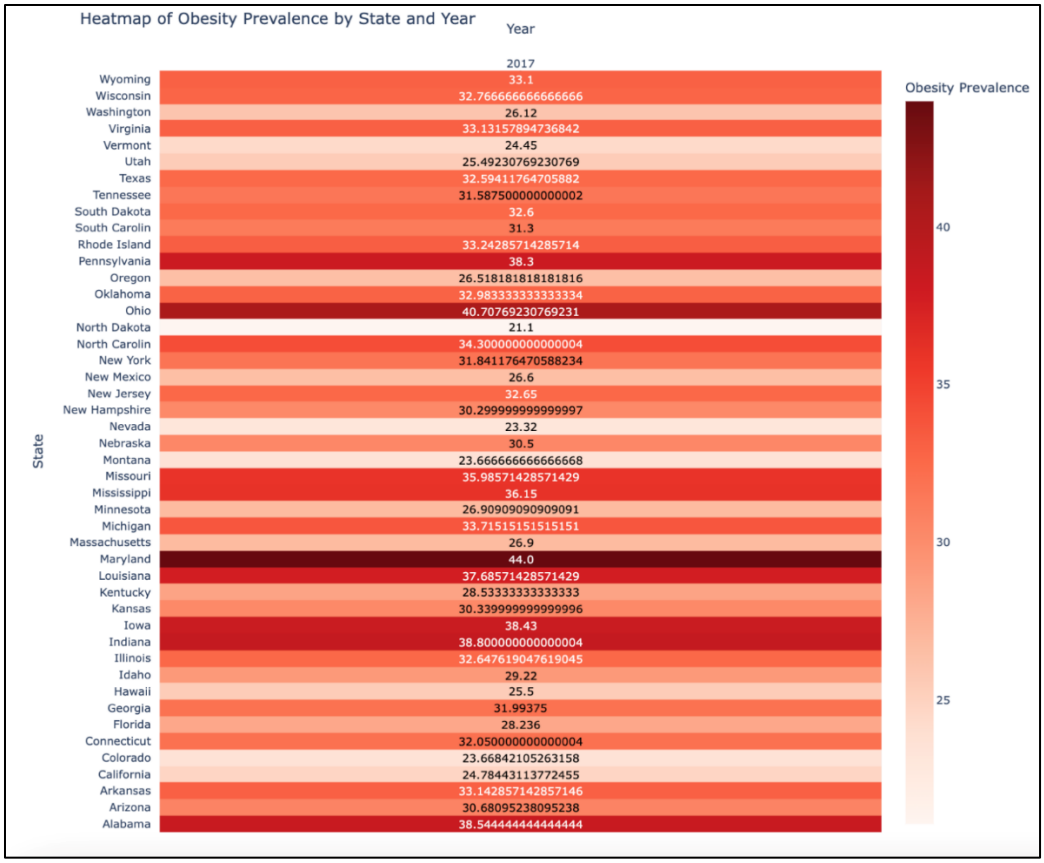

Figure 7: Heatmap of Obesity Prevalence by State and Year

The heatmap (*Figure 7*) provides a visual representation of obesity prevalence across various states in the year 2017. The color gradient indicates the intensity of obesity rates, with darker shades representing higher prevalence.

Overall Analysis:

- **General Trends:** The heatmap reveals a broad range of obesity prevalence across states, indicating significant geographic variability in obesity rates.
- **High Prevalence Areas:** States like Ohio, Alabama, Maryland, and Louisiana show darker shades, indicating higher obesity rates, exceeding 38%. Ohio stands out with a prevalence of 40.71%, while Maryland tops the chart with the highest rate at 44.0%. These states are likely grappling with considerable public health challenges related to obesity.
- **Moderate Prevalence Areas:** Many states fall in the mid-range, with obesity rates between 30% and 35%. Examples include Texas (32.59%), Wisconsin (32.77%), and Nebraska (30.5%).
- **Low Prevalence Areas:** States such as California, Washington, and Utah display lighter shades, indicating lower obesity rates. California has an obesity prevalence of 24.78%, Washington is at 26.12%, and Utah is at 25.49%. These states may benefit from more effective public health policies, higher awareness, and better access to healthy lifestyles.

Highest and Lowest Obesity Prevalence States:

- **Highest:**
  - Maryland (44.0%): Likely influenced by socioeconomic factors, limited access to healthy food options, and insufficient public health initiatives.

- Ohio (40.71%): Reflects significant public health challenges, potentially due to socioeconomic disparities and lifestyle factors.

- **Lowest:**
  - California (24.78%): Benefits from strong public health policies, a culture promoting outdoor activities, and higher awareness of healthy living.
  - Washington (26.12%): Likely due to effective health interventions and a population more inclined towards maintaining a healthy lifestyle.

4. Population vs Diabetes Rates:

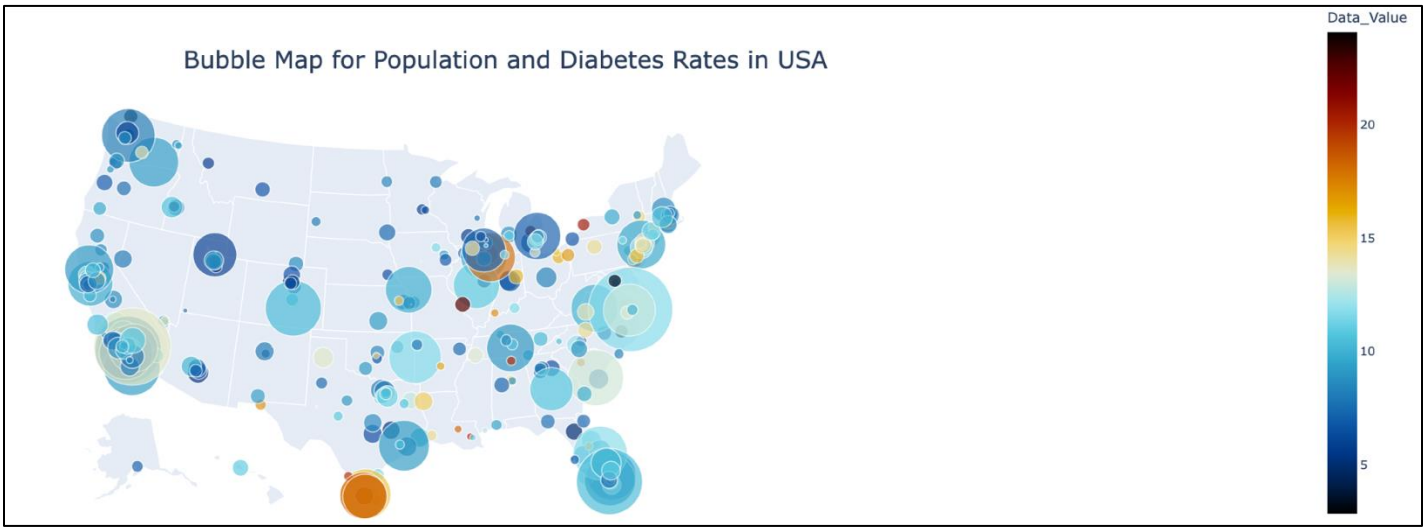

Figure 8: Bubble Map for Population and Diabetes Rates in USA

The Bubble Map (*Figure 8*) visualization vividly illustrates the prevalence of diagnosed diabetes across various geographic locations in the United States, alongside population data. This visualization is instrumental in identifying trends and disparities in diabetes rates, shedding light on how demographic factors and location influence the incidence of this chronic condition.

Insights from Specific Locations:

- Los Angeles, CA: Despite its large population, Los Angeles has relatively low diabetes rates compared to other major cities, suggesting effective public health interventions or lifestyle factors that reduce risk. However, areas like East Los Angeles exhibit higher rates, highlighting pockets of vulnerability that could benefit from targeted health programs.
- Detroit, MI: Detroit displays higher diabetes rates, particularly in economically challenged areas. The city's socioeconomic issues may contribute to health disparities, including limited access to healthcare and nutritious food, which are essential for managing and preventing diabetes.
- Miami, FL: Miami shows varied diabetes rates, with generally higher levels in suburban areas compared to the urban core. This discrepancy may reflect differences in lifestyle and healthcare access between these regions, indicating the need for region-specific strategies to combat diabetes effectively.

Overall Results and Insights:

- Population Density vs. Diabetes Rates: The bubble map indicates that diabetes rates are not directly correlated with population density. Instead, socio-economic and lifestyle factors play a more significant role in influencing diabetes prevalence.

- Regional Disparities: Southern states, known for higher diabetes prevalence, are prominently visible on the map. This trend underscores the impact of diet, exercise, and healthcare access, which vary widely across the country.
- Urban vs. Rural: Urban areas with more healthcare facilities and health education programs typically have lower diabetes rates compared to rural areas, which may face challenges such as limited healthcare access and higher poverty rates. However, urban areas with significant poverty also show increased diabetes rates, indicating that urbanization alone does not protect against diabetes.
- Impact of Healthcare Access: Locations with better healthcare infrastructure and access to preventive care services exhibit lower diabetes prevalence, highlighting the critical role of continuous healthcare management in preventing and controlling this disease.

5. Scatter Plots for Coronary Heart Disease and Cancer Prevalence in US Cities:

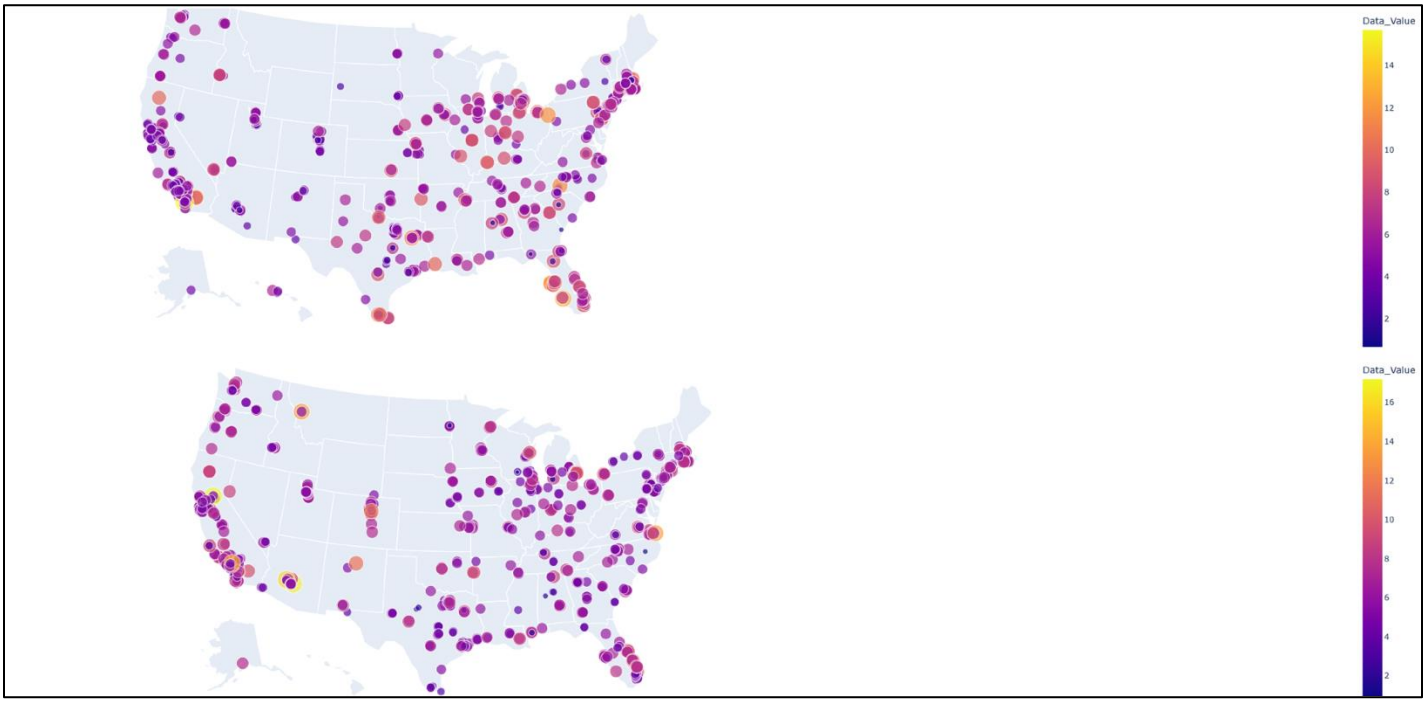

Figure 9: Scatter Plots for Coronary Hart Disease and Cancer Prevalence in US Cities

The scatter plots (*Figure 9*) provided in the image visualize the prevalence of coronary heart disease and cancer (excluding skin cancer) across various US cities. The visualizations are created using Plotly Express and display data points representing different cities, with color and size indicating the prevalence rates of these health conditions.

Insights from the Scatter Plots:

Coronary Heart Disease Prevalence:

- The scatter plot for coronary heart disease shows various US cities with different prevalence rates.
- Cities with higher prevalence rates are marked with larger and warmer-colored points.
- There might be visible clusters in certain regions indicating higher prevalence, which could correlate with socio-economic factors, healthcare access, or lifestyle patterns prevalent in those areas.

Cancer Prevalence (excluding skin cancer):

- Similarly, the scatter plot for cancer prevalence highlights cities with varying rates.
- The size and color-coding help quickly identify cities with higher cancer rates.

- Patterns or clusters might emerge, suggesting regions where cancer prevalence is notably high, possibly due to environmental factors, population demographics, or healthcare practices.

Key Insights:

- **Regional Health Disparities**: Both plots may reveal regional disparities in the prevalence of coronary heart disease and cancer. For example, certain areas might consistently show higher rates for both conditions, indicating a need for targeted public health interventions.
- **Urban vs. Rural Trends**: The visualization can help identify if there are significant differences in prevalence rates between urban and rural areas, guiding resource allocation and policymaking.
- **Health Infrastructure**: Areas with lower prevalence might indicate better healthcare infrastructure or more effective public health programs, serving as models for other regions.

6. Comparison of Coronary Heart Disease and Obesity Prevalence by state:

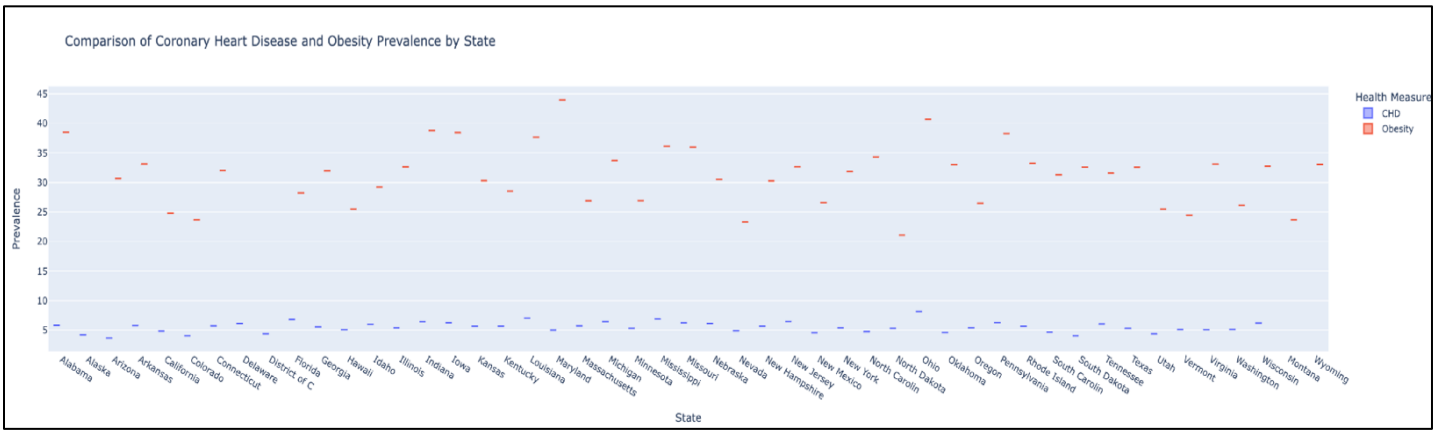

Figure 10: Comparison of Coronary Heart Disease and Obesity Prevalence by state

The graph (*Figure 10*) provides a visual comparison of the prevalence of Coronary Heart Disease (CHD) and Obesity across different states. A box plot is created using Plotly to visualize the comparison of CHD and Obesity prevalence by state. Each state has two box plots representing the prevalence of CHD (blue) and Obesity (red):

Detailed Analysis

- Prevalence of Obesity vs. CHD:
  Obesity prevalence is significantly higher than CHD prevalence in all states. The range of obesity prevalence across states is roughly between 25% and 45%. The range of CHD prevalence is much lower, roughly between 5% and 10%.
- State-wise Variations:
  States like West Virginia, Mississippi, and Alabama show high obesity prevalence. CHD prevalence is relatively consistent across states with slight variations.
- Comparative Analysis: There is no clear correlation visible between states with high obesity rates and states with high CHD rates based on this graph alone.

Implications

- Health Policies: States with higher obesity rates might need to focus more on obesity prevention and management programs.
- Public Health Interventions: States showing higher CHD prevalence should consider cardiovascular health programs and screenings.

Conclusion

The graph highlights the significant difference between the prevalence of obesity and CHD across the states. Obesity rates are alarmingly high in comparison to CHD, indicating a potential area of focus for public health initiatives. The consistent CHD rates suggest uniformity in the prevalence of heart disease, but the higher obesity rates could imply future increases in CHD if not addressed.

7. Analysis of Top 5 cities with highest prevalence of Arthritis among adults aged >=18 Years:

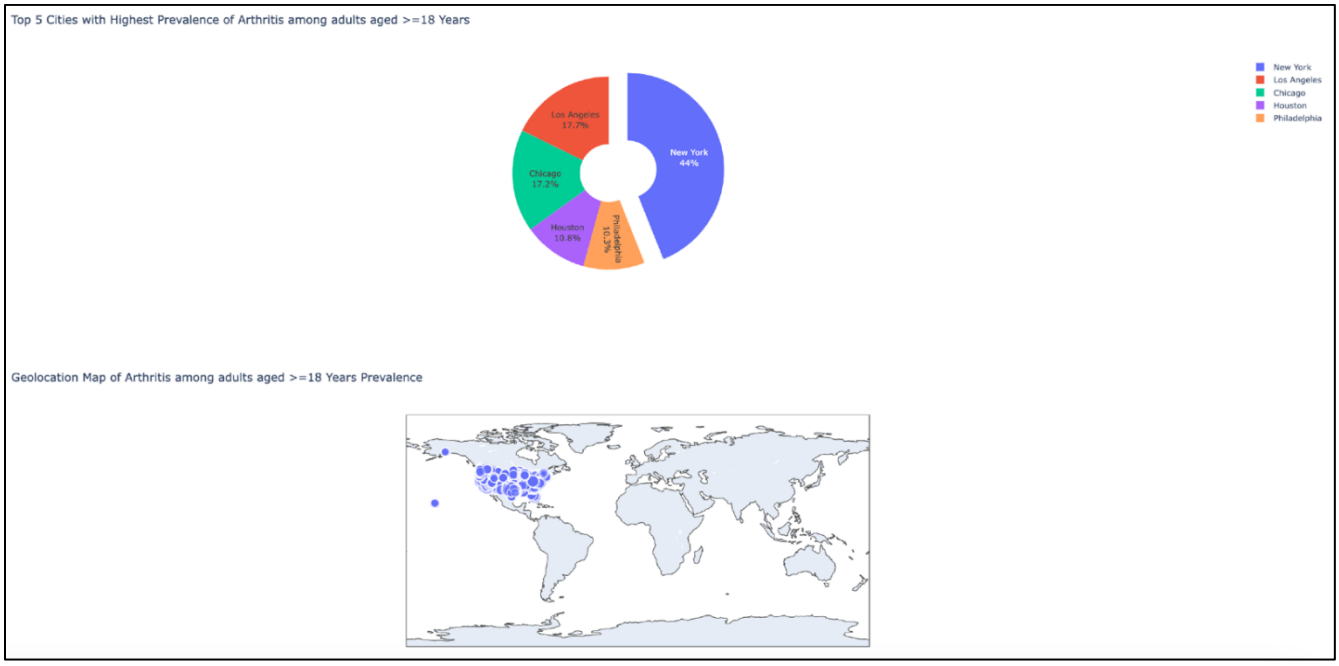

Figure 11: Analysis of Top 5 cities with highest prevalence of Arthritis among adults aged >=18 Years

Pie Chart: Top 5 Cities with Highest Prevalence of Arthritis among Adults Aged ≥18 Years

- New York:
  Prevalence: 44%
  Analysis: New York has the highest prevalence of arthritis among the top 5 cities. This indicates that nearly half of the adult population in New York reports having arthritis. Various factors could contribute to this high prevalence, including population density, lifestyle, and access to healthcare.
- Los Angeles:
  Prevalence: 17.7%
  Analysis: Los Angeles ranks second with a significantly lower prevalence compared to New York. The climate, lifestyle, and demographic factors in Los Angeles might influence the lower prevalence rate.
- Chicago:
  Prevalence: 17.2%
  Analysis: Chicago is close behind Los Angeles with a similar prevalence rate. Factors such as weather, industrial background, and healthcare infrastructure might impact the arthritis rates in Chicago.

- Philadelphia:
  Prevalence: 10.3%
  Analysis: Philadelphia has a lower prevalence rate compared to New York, Los Angeles, and Chicago. This might be due to differences in population demographics, healthcare access, and local environmental factors.
- Houston:
  Prevalence: 10.8%
  Analysis: Houston has a slightly higher prevalence rate than Philadelphia. The climate and lifestyle in Houston, along with other socio-economic factors, could play a role in the arthritis prevalence.

Geolocation Map: Arthritis Prevalence among Adults Aged ≥18 Years

- Map Overview: The map displays the geographical distribution of arthritis prevalence among adults aged 18 years and older. The size of the markers on the map indicates the relative prevalence of arthritis in each location.
- Geographical Insights: Concentration in the United States: The map highlights a concentration of data points in the United States, indicating that the dataset primarily covers US cities.
- High Prevalence Areas: The larger markers, particularly in the northeastern part of the United States (e.g., New York), indicate higher prevalence rates in those regions.
- Distribution: The distribution of arthritis prevalence varies across the US, with some regions showing higher rates than others. Factors such as climate, urbanization, healthcare access, and socio-economic conditions could influence these variations.

General Insights

- Health Infrastructure: Cities with better healthcare infrastructure might have more accurate reporting and diagnosis of arthritis cases, influencing the prevalence rates.
- Lifestyle Factors: Urban lifestyle, including diet, physical activity, and occupational factors, might contribute to the varying prevalence of arthritis in different cities.
- Environmental Factors: Climate and environmental conditions, such as humidity and temperature, can affect arthritis symptoms and prevalence.

By analyzing the graphs (*Figure 11*), we can infer that New York has the highest prevalence of arthritis among adults aged 18 years and older, followed by Los Angeles, Chicago, Philadelphia, and Houston. The geolocation map further supports these findings and provides a visual representation of the distribution of arthritis prevalence across different regions.

8. Prevention Measures Analysis: Health Insurance Coverage

- **Geographical Distribution**:
  - The map (*Figure 12*) shows health insurance coverage data spread across the United States, indicating areas where data on health insurance is available.

- **Coverage Levels**:
  - The color intensity of the data points ranges from light to dark purple. Darker shades indicate higher values of health insurance coverage. We can infer that areas with darker shades have a higher percentage of insured individuals.
  - The areas with lighter shades or smaller dots have lower coverage, indicating a potential need for increased health insurance access or enrollment efforts in these regions.

- **Population Impact**:
    - The size of the data points is proportional to the population count in each location. Larger points indicate areas with a higher population count. This helps in understanding the impact of health insurance coverage in densely populated regions.
    - For instance, large circles in darker shades signify regions with both high population and high health insurance coverage, which is positive.

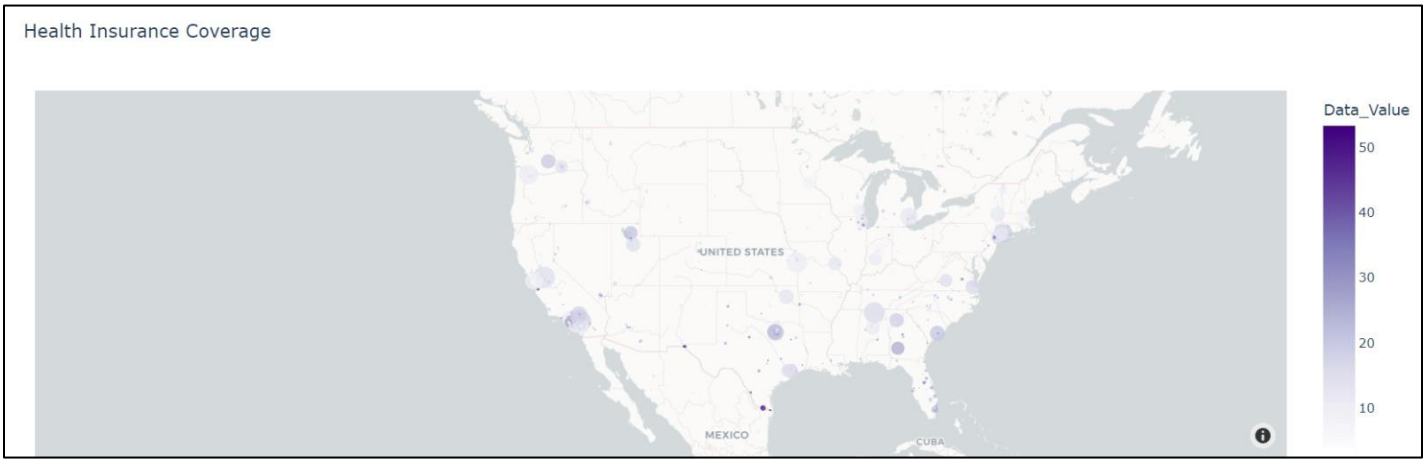

Figure 12: Health Insurance Coverage

- **Regional Variations**:
    - By examining the distribution of coverage across different states or regions, we can identify patterns. Some regions might have uniformly high coverage, while others may exhibit significant disparities.
    - For example, the map might show clusters of high coverage in urban areas, while rural areas may have lower coverage, reflecting the typical urban-rural divide in access to healthcare services.

- **High and Low Coverage Areas**:
    - The data points in regions like the East Coast and parts of California are likely to be darker and larger, indicating higher coverage in populous and economically stronger areas.
    - Conversely, smaller or lighter data points in states like those in the Midwest or certain Southern states might indicate lower coverage, which can guide policy makers and health organizations in targeting these areas for improvement.

- **Policy Implications**:
    - Insights from this map can help in directing resources and efforts to areas with low health insurance coverage.
    - It can also highlight the success of regions with high coverage, potentially serving as models for other areas.

- **Urban vs. Rural Divide**:
    - The map might show that urban areas generally have better health insurance coverage compared to rural areas. This can be used to advocate for policies aimed at improving healthcare access in less populated areas.

9. Aggregate data to get average prevalence of core preventive services for older women by state:

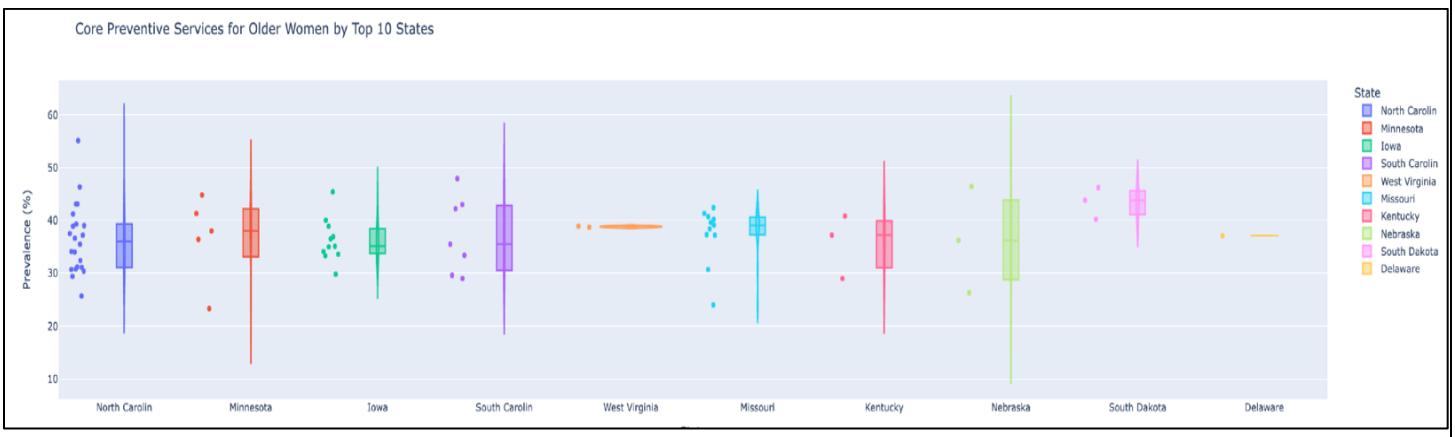

Figure 13: *Average prevalence of core preventive services for older women by state*

The graph (*Figure 13*) is a violin plot with box plots included, showing the distribution of the prevalence (%) of core preventive services for older women across the top 10 states. Each state's data is represented by a different color.

Key Insights

- North Carolina:
  Distribution: The data shows a relatively wide range of prevalence rates, with some values reaching as high as 60%.
  Box Plot: The interquartile range (IQR) is quite large, indicating variability in the data. The median is around 40%, and there are several outliers below the lower quartile.
- Minnesota:
  Distribution: Minnesota also shows a wide distribution, with prevalence rates mostly concentrated around the median of approximately 40%.
  Box Plot: The IQR is wide, suggesting variability. There are outliers below the lower quartile and above the upper quartile.
- Iowa:
  Distribution: Iowa has a relatively narrower range of prevalence rates compared to North Carolina and Minnesota.
  Box Plot: The median is slightly below 40%, with a smaller IQR indicating less variability. There are outliers on both sides of the quartiles.
- South Carolina:
  Distribution: South Carolina shows a broad distribution with prevalence rates spreading widely.
  Box Plot: The median is around 40%, with a large IQR. There are several outliers, indicating a varied distribution of data points.
- West Virginia:
  Distribution: West Virginia has a tighter clustering of prevalence rates.
  Box Plot: The median is around 35%, and the IQR is relatively small. This suggests more consistent prevalence rates. There are outliers, but not as many as in other states.
- Missouri:
  Distribution: Missouri shows a moderate distribution of prevalence rates.

Box Plot: The median is just below 40%, with a moderate IQR. There are a few outliers, indicating some variability.

- Kentucky:
  Distribution: Kentucky's data distribution is relatively wide.
  Box Plot: The median is around 40%, with a wide IQR. There are outliers on both ends, showing variability in the prevalence rates.
- Nebraska:
  Distribution: Nebraska shows a wide distribution of prevalence rates.
  Box Plot: The median is around 40%, with a large IQR. There are several outliers, indicating a varied distribution.
- South Dakota:
  Distribution: South Dakota's data is moderately spread out.
  Box Plot: The median is slightly above 35%, with a moderate IQR. There are outliers on both ends, indicating some variability.
- Delaware:
  Distribution: Delaware has a narrow range of prevalence rates.
  Box Plot: The median is just above 35%, with a small IQR indicating less variability. There are outliers above the upper quartile.

General Observations

- Variability: States like North Carolina, Minnesota, and South Carolina show high variability in the prevalence rates of core preventive services for older women, indicating inconsistency in service delivery or reporting.
- Consistency: States like West Virginia and Delaware show more consistent data with narrower IQRs and fewer outliers.
- Median Values: The median values for most states hover around 35-40%, suggesting that approximately this percentage of older women are receiving core preventive services.
- Outliers: The presence of outliers in nearly all states indicates that there are significant deviations in certain areas which could be due to various factors like access to healthcare, socio-economic conditions, and public health policies.

Conclusion

The violin plot with box plots provides a comprehensive view of the distribution of prevalence rates for core preventive services for older women across the top 10 states. It highlights the variability and consistency within each state, allowing for a detailed comparison of how these services are distributed among older women in different regions.

10. Interactive bar chart comparing health metrics:

This graph (*Figure 14*) examines an interactive bar chart visualization that compares average health metric values across different cities in the United States. The health metrics included in the analysis are:

- Cancer (excluding skin cancer) among adults aged >= 18 Years
- Arthritis among adults aged >= 18 Years
- Stroke among adults aged >= 18 Years
- All teeth lost among adults aged >= 65 Years

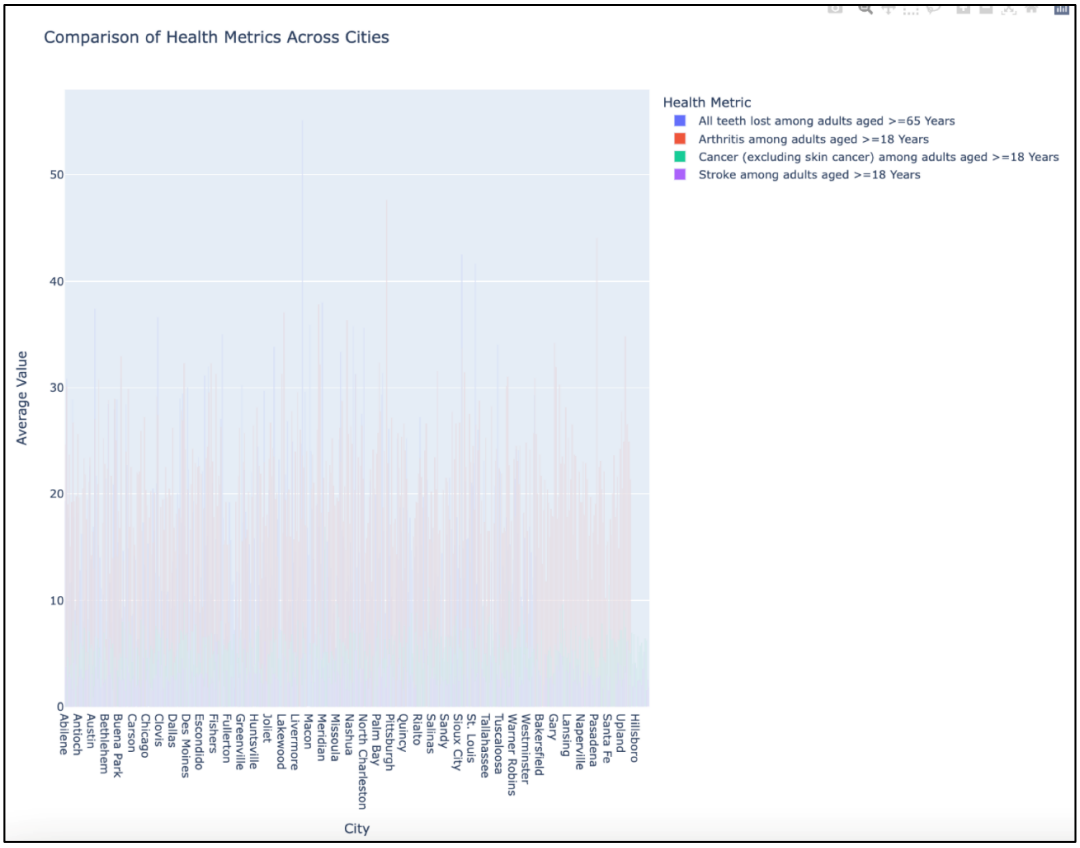

Figure 14: Interactive bar chart comparing health metrics

Data Representation: The data is presented in a bar chart format, with each city represented by a group of bars. Each bar within a city group corresponds to the average value for a specific health metric. The y-axis displays the average value, while the x-axis displays the city name.

Analysis Insights

The chart allows for comparisons of average health metric values between cities and across different health metrics within a city. Key observations include:

- City Comparisons: By comparing the heights of bars within each city group, it's possible to identify variations in average health metric values. A city might have a higher average value for cancer compared to another city, while the other city might have a higher average value for arthritis.
- Health Metric Variations: The overall height and distribution of bars across different health metrics within a city group can reveal variations. For instance, a city might have a higher average value for all teeth lost compared to cancer, suggesting a potential higher prevalence of dental problems in the older adult population.
- Cities of Interest: Cities with consistently high or low bars across all health metrics might warrant further investigation. High bars could indicate areas with risk factors contributing to these health issues, while low bars could suggest areas with successful health interventions or demographics with lower risks.

Inferences and Conclusion

The project analyzed the "500 Cities: Local Data for Better Health" dataset, focusing on health outcomes, preventive measures, and unhealthy behaviors across various cities in the USA. The data visualizations aimed to highlight trends and patterns that could inform public health policies and interventions. Here are the key inferences and conclusions drawn from the analysis:

1. Trends in Unhealthy Behaviors

- **Smoking Rates**: Geographic variations were significant, with urban areas generally exhibiting lower smoking rates due to more anti-smoking ordinances and available alternative activities. Rural areas or smaller towns typically had higher smoking rates.
- **Specific Insights**:
    - **New York, NY**: Higher smoking rates in Staten Island compared to Manhattan, suggesting targeted interventions in specific boroughs.
    - **Los Angeles, CA**: Lower smoking rates in affluent neighborhoods like West Hollywood and Santa Monica, with higher rates in eastern parts, indicating socioeconomic influences.
    - **Chicago, IL**: Higher smoking rates in the south and west sides, correlating with socioeconomic challenges.

2. Obesity Rates

- **General Trends**: Obesity rates were higher in southern states like Mississippi and Alabama, while states like Colorado and Vermont had lower rates, reflecting cultural and environmental influences.
- **Urban vs. Rural**: Urban centers with high walkability scores and extensive public transit systems had lower obesity rates, highlighting the impact of urban planning on public health.
- **Specific Insights**:
    - **Houston, TX**: Higher obesity rates in eastern parts of the city, indicating a need for community-based health programs.
    - **San Francisco, CA**: Lower overall obesity rates in central and northern parts, with higher rates in southern districts needing targeted initiatives.
    - **Atlanta, GA**: Higher obesity rates in suburban areas compared to the central urban area, suggesting urban lifestyle factors play a role.

3. Diabetes Prevalence

- **Population Density vs. Diabetes Rates**: Diabetes rates were not directly correlated with population density. Socioeconomic and lifestyle factors played a more significant role.
- **Regional Disparities**: Southern states showed higher diabetes prevalence, emphasizing the influence of diet, exercise, and healthcare access.
- **Specific Insights**:
    - **Los Angeles, CA**: Relatively low diabetes rates despite a large population, with higher rates in East Los Angeles.
    - **Detroit, MI**: Higher diabetes rates in economically challenged areas, indicating limited access to healthcare and nutritious food.
    - **Miami, FL**: Varied diabetes rates with higher levels in suburban areas compared to the urban core.

4. Arthritis Prevalence

- **City Comparisons**:
    - **New York, NY**: Highest prevalence of arthritis among the top 5 cities, possibly due to population density, lifestyle, and healthcare access.
    - **Los Angeles, CA**: Significantly lower prevalence compared to New York, influenced by climate, lifestyle, and demographics.
    - **Chicago, IL**: Similar prevalence to Los Angeles, impacted by weather, industrial background, and healthcare infrastructure.
    - **Philadelphia, PA**: Lower prevalence, potentially due to population demographics and healthcare access.
    - **Houston, TX**: Slightly higher prevalence than Philadelphia, influenced by climate and lifestyle factors.

5. Health Insurance Coverage

- **Geographical Distribution**: Areas with higher population densities generally exhibited higher health insurance coverage.
- **Regional Variations**: Urban areas typically had better coverage compared to rural areas, reflecting the urban-rural divide in healthcare access.
- **Policy Implications**: Insights can guide resource allocation and efforts to improve healthcare access in regions with lower coverage.

6. Preventive Services for Older Women

- **Variability**: States like North Carolina, Minnesota, and South Carolina showed high variability in the prevalence rates of core preventive services, indicating inconsistency in service delivery or reporting.
- **Consistency**: States like West Virginia and Delaware showed more consistent data with narrower interquartile ranges (IQRs) and fewer outliers.
- **Median Values**: Median values for most states hovered around 35-40%, suggesting that this percentage of older women received core preventive services.

Conclusion

The analysis provided critical insights into health disparities and the effectiveness of public health policies across different regions in the USA. The visualizations highlighted significant geographic and socioeconomic variations in health behaviors and outcomes, emphasizing the need for targeted public health interventions. Policymakers and health officials can use these insights to develop strategies that address specific regional needs, ultimately improving health outcomes and reducing disparities across the country.

References

[1] City Health Profiles (Dataset) [*Online*] – Available: https://catalog.data.gov/dataset/city-health-profiles

[2] Kate Conquest, "As Preemption Efforts Grow, Cities See Fewer Pathways to Pursue Healthy Policies" [*Online*] - Available: https://www.cityhealth.org/resource/preemption-efforts/