Urban Health Profiles

A Detailed Analysis of Health Behaviors and Outcomes in 500 U.S. Cities

Group 12

Shalini Dutta
Navisha Shetty
Rishab Radesh
Prateek Shetty

Table of Contents

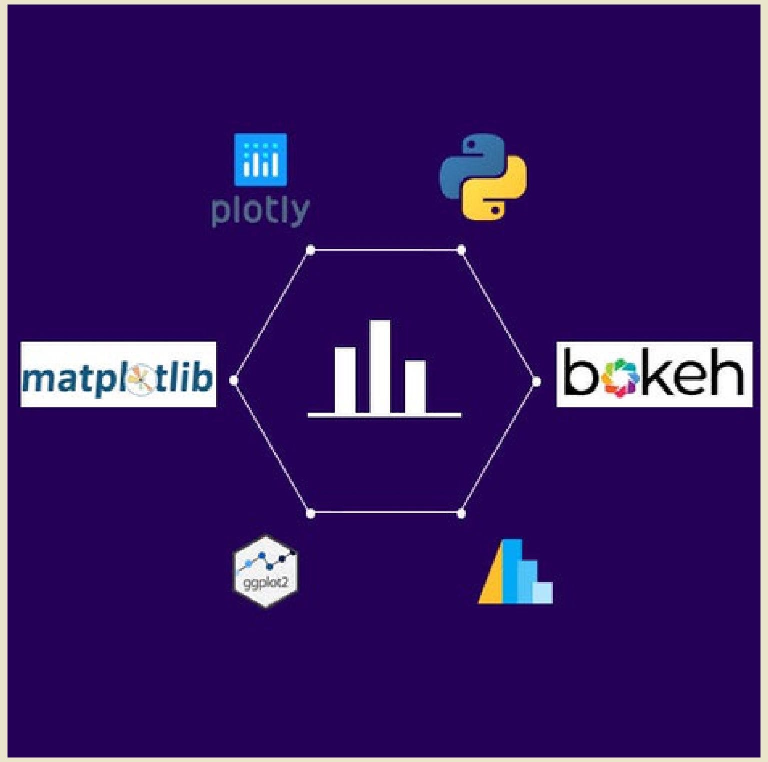
Project Overview

Objective

- Analyze and visualize geospatial data to uncover insights.
- Uncover trends and insights with interactive visualizations.
- Inform decision-making with data-driven insights.

Tools Used

- Utilize Plotly for advanced geospatial data visualization.
- Create dynamic, interactive maps and charts.
- Leverage advanced libraries for impactful data presentation.



Dataset Overview

- Geospatial dataset from data.gov.
- Over 810,000 records with 24 variables.
- Health-related metrics: Health Outcomes, Unhealthy Behaviors, and Prevention.
- Examples of measures include diagnosed diabetes, current smoking, lack of health insurance, obesity.
- Key columns include State and City Descriptors, Geolocation, Health Measures (e.g., prevalence of diseases, health behaviors), Confidence Limits, and Population Counts
- Mostly complete data; notable exceptions: GeoLocation(56 missing), CityName (56 missing), and TractFIPS (28,056 missing)

Year	StateAbbr	StateDesc	CityName	GeographicLevel	DataSource	Category	UniqueID	Measure	Data_Value	PopulationCount	GeoLocation	CityFIPS	TractFIPS	Short_Question_Text

Data Analysis

```
In [9]: # Initial data summary
        initial_summary = {
            'total_rows': data.shape[0],
            'total_columns': data.shape[1],
            'column_names': data.columns
        }

        initial_summary
Out[9]: {'total_rows': 810103,
         'total_columns': 24,
         'column_names': Index(['Year', 'StateAbbr', 'StateDesc', 'CityName', 'GeographicLevel',
                'DataSource', 'Category', 'UniqueID', 'Measure', 'Data_Value_Unit',
                'DataValueTypeID', 'Data_Value_Type', 'Data_Value',
                'Low_Confidence_Limit', 'High_Confidence_Limit',
                'Data_Value_Footnote_Symbol', 'Data_Value_Footnote', 'PopulationCount',
                'GeoLocation', 'CategoryID', 'MeasureId', 'CityFIPS', 'TractFIPS',
                'Short_Question_Text'],
               dtype='object')}
```

Findings from the initial data analysis

- Missing Values: There are no missing values for most of the columns, but there are significant gaps.
- Duplicates: There are no duplicate rows in the dataset, which is positive for the integrity of the data.
- GeoLocation Column: The GeoLocation column, crucial for mapping, has 56 missing entries. This column needs to be in a suitable format for geospatial analysis, typically as separate latitude and longitude columns.

Given these findings, the next steps are:

- Reduce dimensionality of the dataset since it is too large.
- Addressing missing GeoLocation entries: exclude these records from mapping analyses since their exact locations are unknown.
- Extracting latitude and longitude from the GeoLocation column into separate, usable columns for mapping.
- Proceeding with the cleaned dataset to create interactive maps and visualizations.

PopulationCount	GeoLocation
6502	(40.5741076886, -73.9968626717)
2802	(40.6974507413, -73.9408852337)
2658	(40.6123965321, -73.9655935668)

Data Preparation

To reduce the dataset to 20,000 rows without creating a highly skewed dataset with respect to geospatial columns and CityName, we can consider the following steps:

- Columns to Drop: Identify and drop columns that may not be critical for interactive visualization and advanced analysis, like footnotes, confidence limits, and potentially redundant identifiers.
- Rows to Sample: Since we aim for geographic and city-based diversity, stratified sampling based on CityName could be a good approach, ensuring representation from all cities.

```
{'total_cities': 474,
 'city_records_distribution':         CityName  RecordCount
 0      New York        59911
 1   Los Angeles        28119
 2       Chicago        22369
 3       Houston        16787
 4  Philadelphia        10707}
```

Short_Question_Text	Latitude	Longitude
Diabetes	40.574108	-73.996863
Current Smoking	40.697451	-73.940885
Health Insurance	40.612397	-73.965594

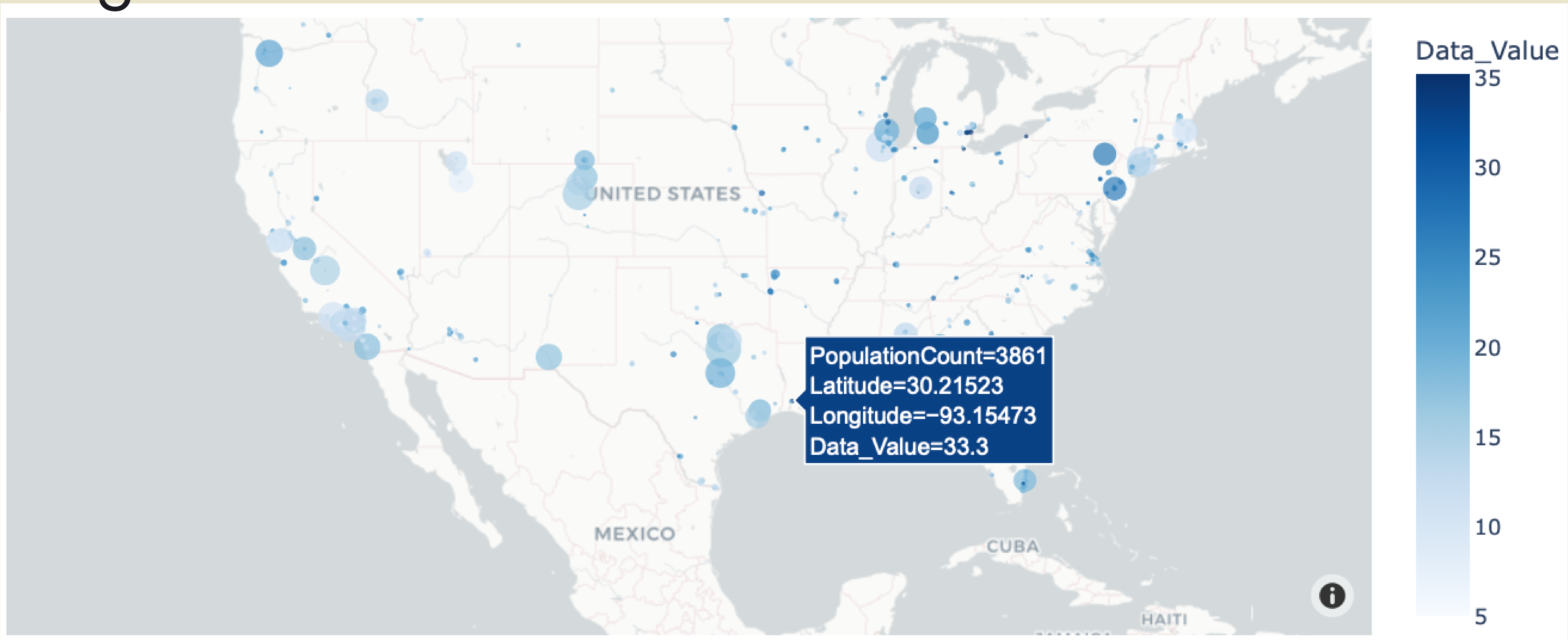
```
# Initial data summary
final_summary = {
    'total_rows': data.shape[0],
    'total_columns': data.shape[1],
    'column_names': data.columns
}

final_summary
```

```
{'total_rows': 19908,
 'total_columns': 15,
 'column_names': Index(['Year', 'StateAbbr', 'StateDesc', 'CityName', 'GeographicLevel',
        'DataSource', 'Category', 'UniqueID', 'Measure', 'Data_Value',
        'PopulationCount', 'GeoLocation', 'CityFIPS', 'TractFIPS',
        'Short_Question_Text'],
       dtype='object')}
```

Trends in Unhealthy Behaviors

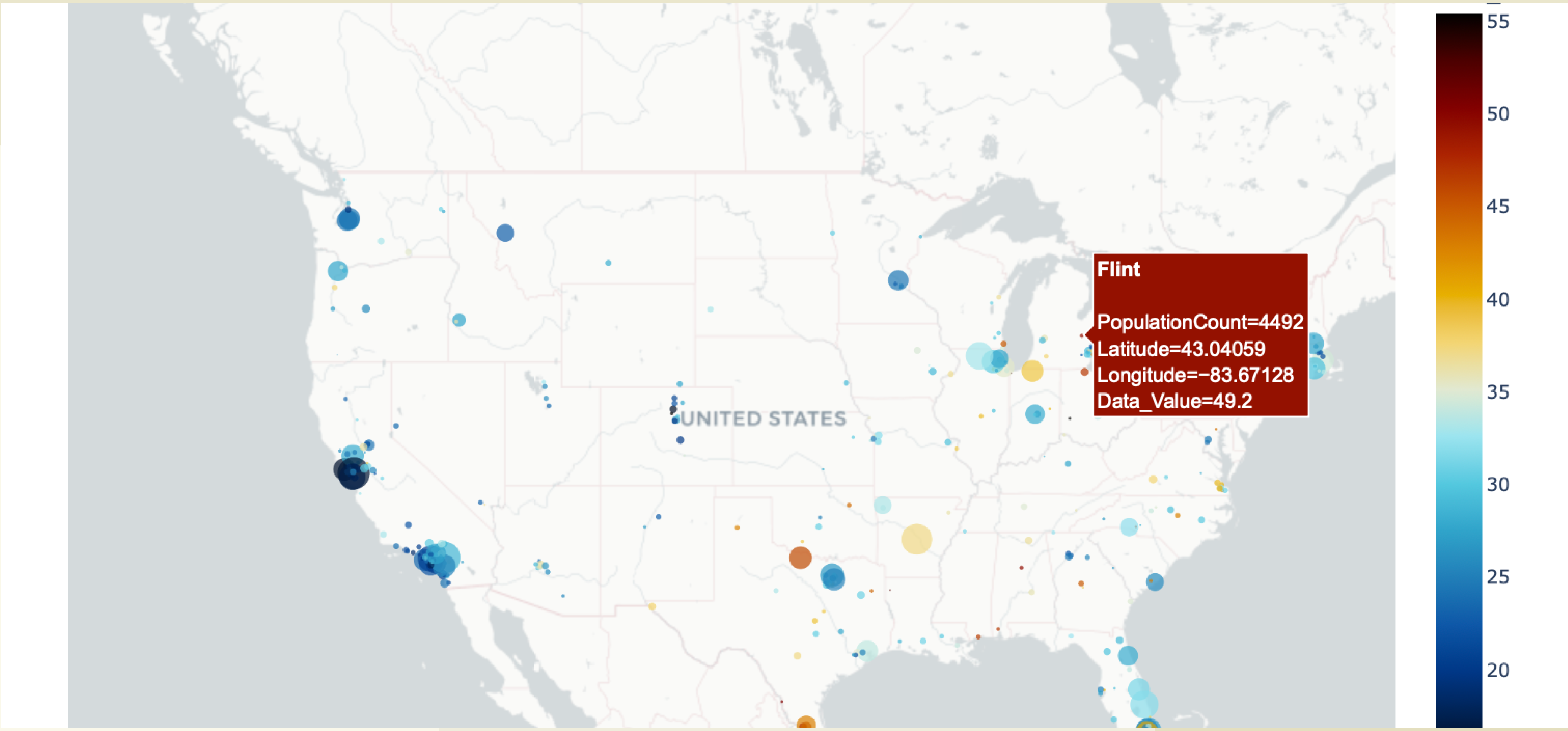
Smoking Rates



- The interactive map visualizes smoking rates by location, with point size reflecting area population and color denoting adult current smoking prevalence.
- The highest smoking prevalence is in Louisiana, New Orleans while NYC in NewYork having highest population falls in the average range of 16 for adult smokers.
- The color gradient helps identify regions with varying levels of smoking prevalence, potentially indicating areas where public health interventions could be more urgently needed.

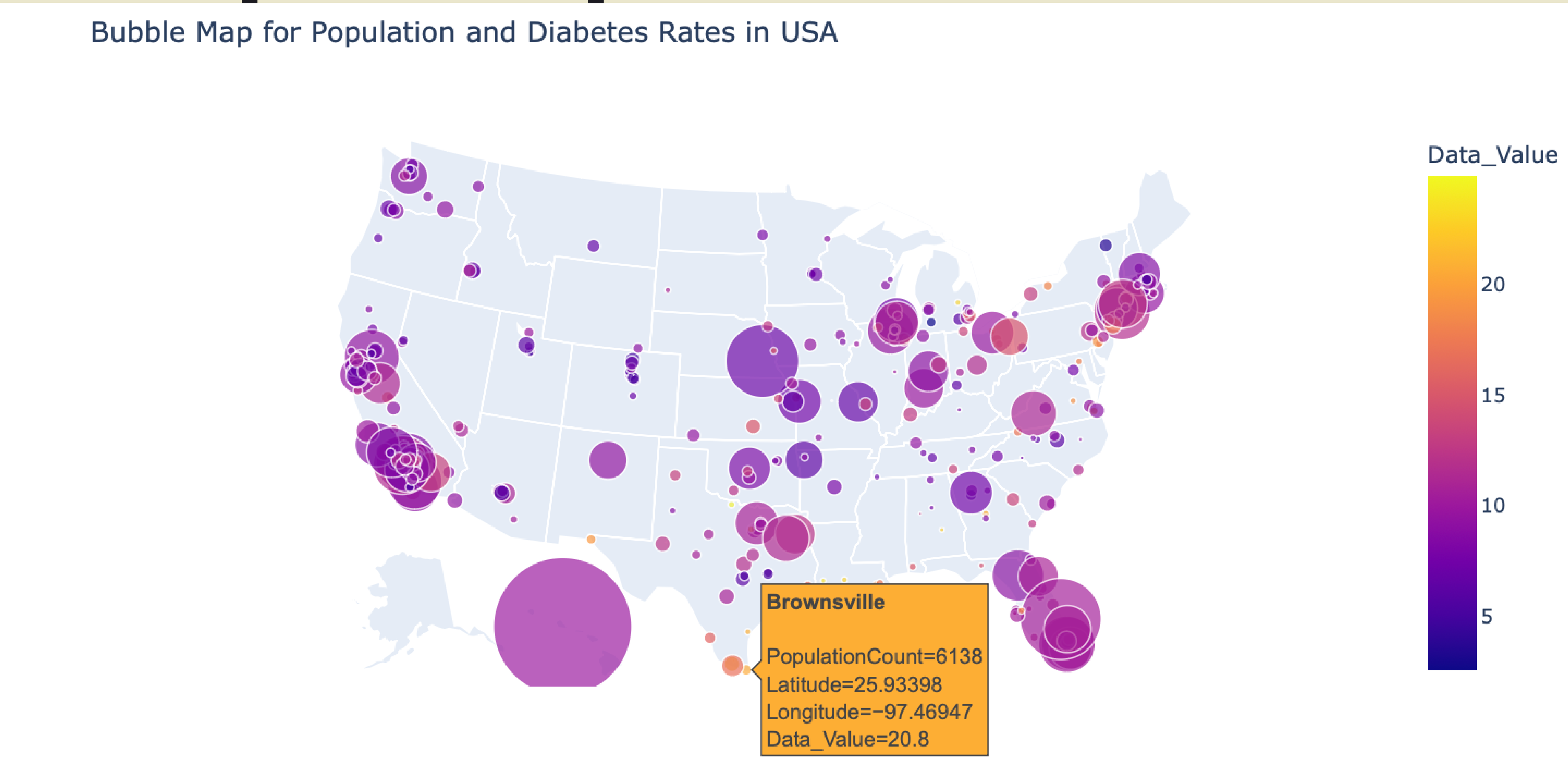
Trends in Unhealthy Behaviors

Obesity Rates



- The interactive map visualizes obesity rates by location, with point size reflecting area population and color denoting average rate of obesity.
- The highest obesity rate is in Flint, Michigan while NYC in NewYork having highest population falls just above average with a value of 27.
- Identifying cities with higher obesity rates can guide targeted health interventions, such as nutritional education and accessible physical activity programs, to combat obesity.

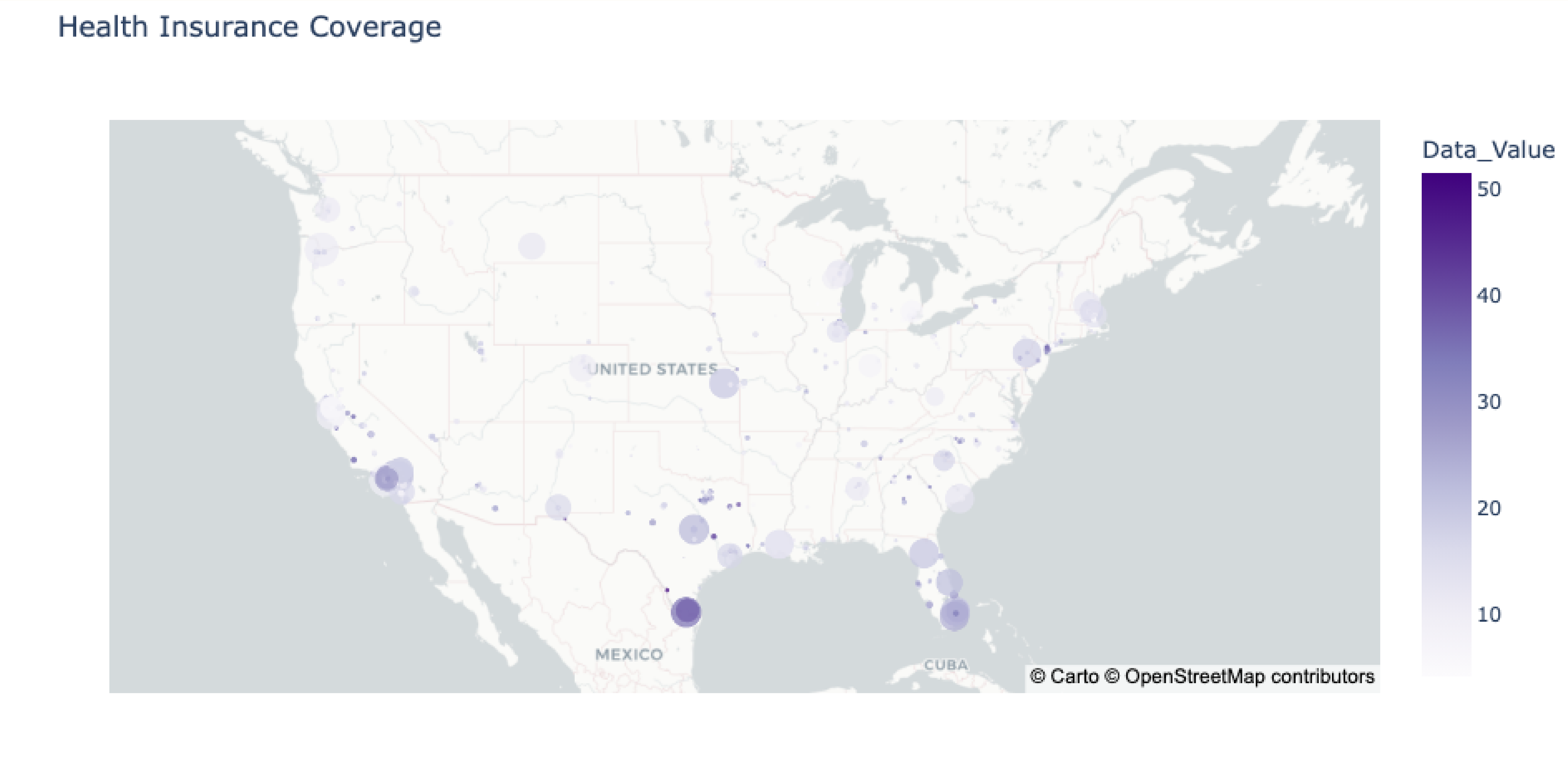
Bubble Map for Population vs Diabities Rates



- The geo bubble map combines population size and diabetes rates for cities across the USA, with bubble size representing population count and color indicating the average rate of diabetes.
- The highest diabetes rates is in Brownsville, Texas while NYC in NewYork having highest population falls just above average with a value of 17.45.
- The map can be instrumental in allocating healthcare resources, planning diabetes awareness campaigns, and implementing preventative measures, especially in densely populated areas with high diabetes rates

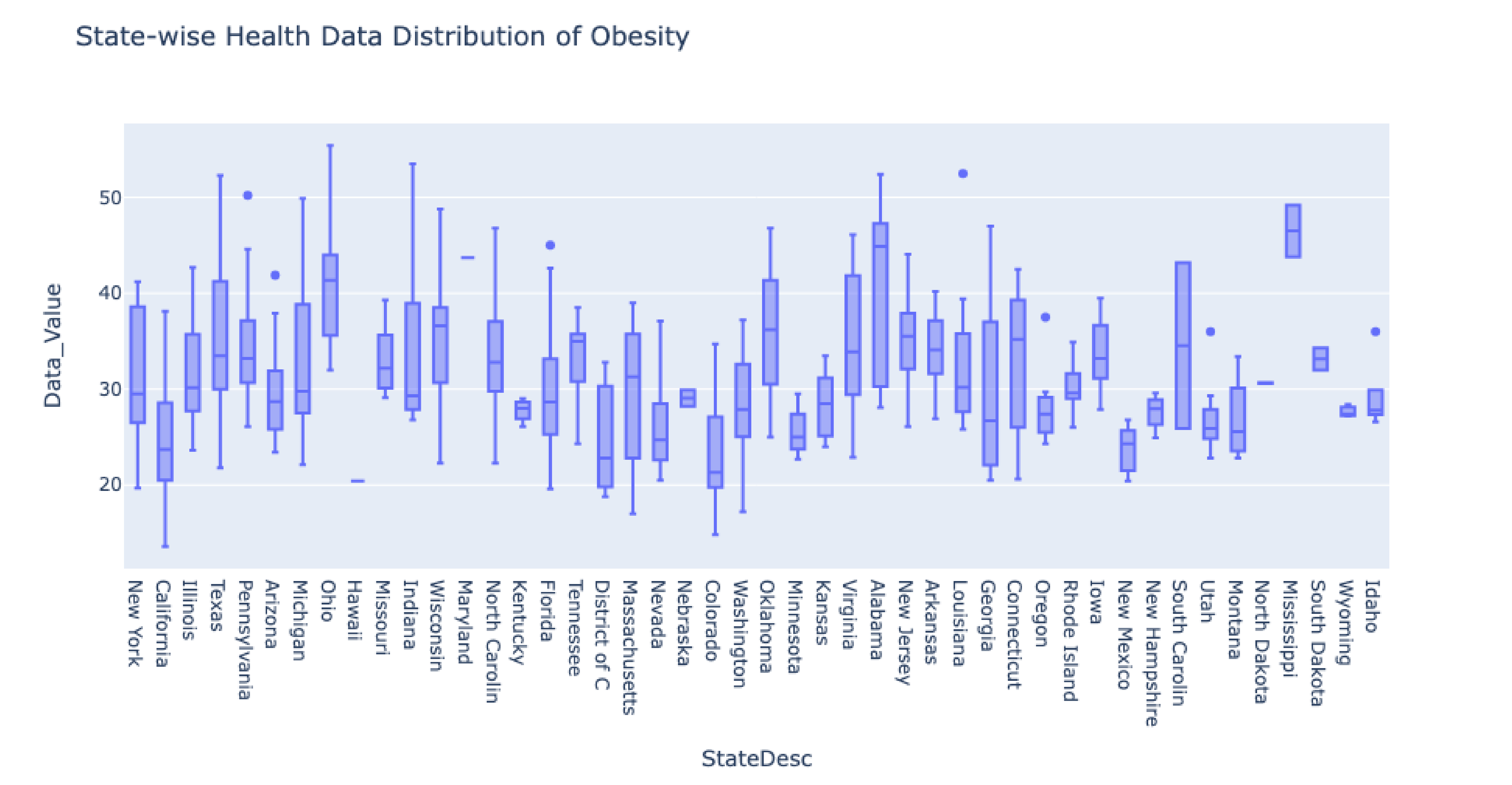
Health Insurance Coverage Map

- Points sized by population, colored by uninsured rate.
- Highlights areas with high uninsured rates, indicating potential access challenges.
- Aims to direct efforts to increase insurance enrollment and address health disparities.



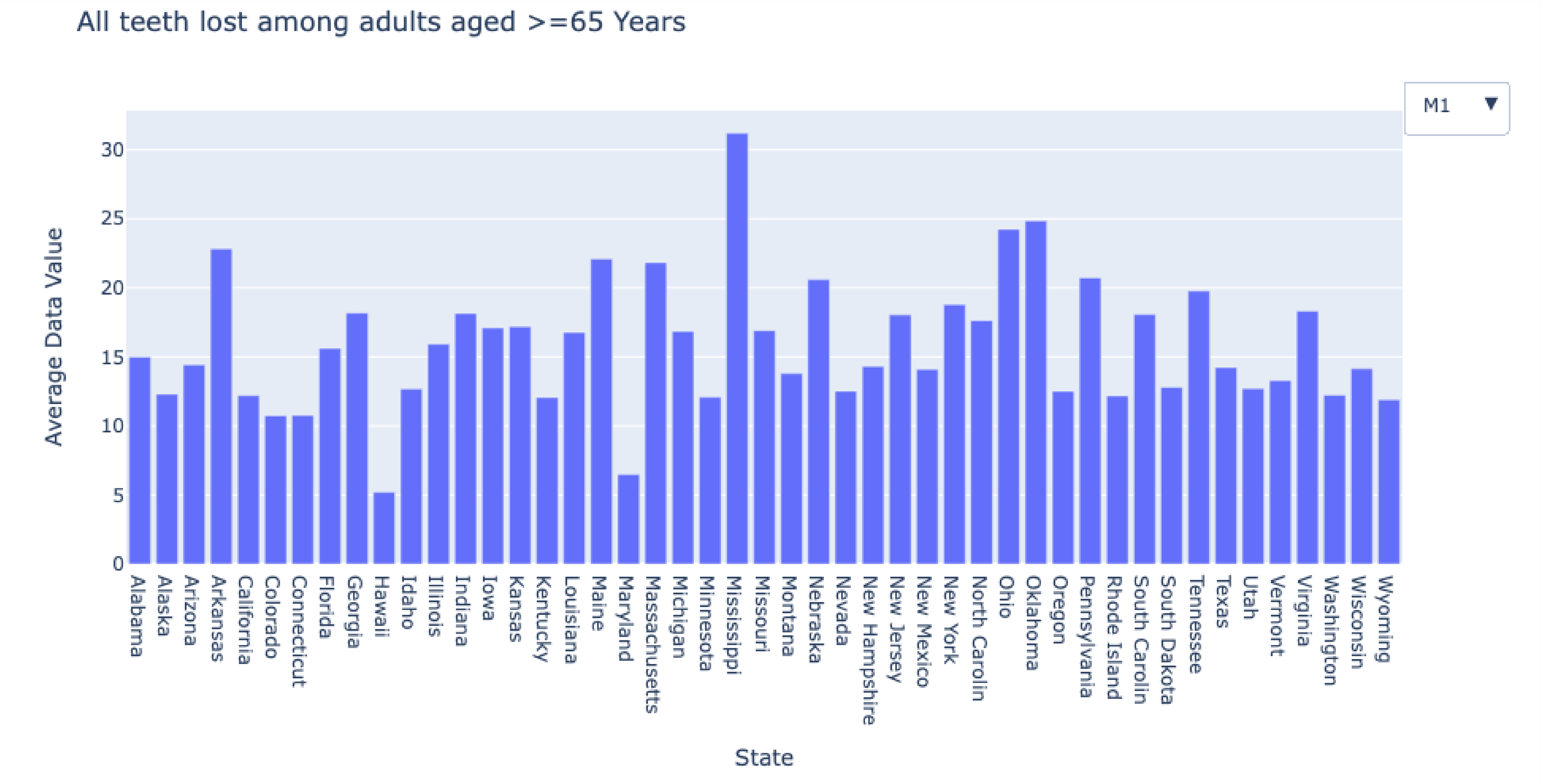
State-wise Health Data Distribution of Obesity

- Shows median, quartiles, and outliers for state obesity rates.
- Allows comparison and identification of states with significant obesity concerns.
- Supports policy and interventions to reduce obesity at the state level.



Interactive bar chart comparing health metrics

- Users can select different health measures for comparison.
- Enables exploration and comparison of health outcomes for various conditions.
- Aids in informed policy making and program development targeting health needs.



Conclusion

- Regions show significant variation in health insurance coverage, highlighting areas for targeted enrollment initiatives.
- State-to-state differences in obesity rates indicate the need for localized health policies.
- Disparities across health metrics emphasize the necessity for tailored interventions.
- Population size and health outcomes are correlated, suggesting urbanization impacts health.

Thank You

Half Coders

Group 12