UNITED STATES OBESITY RISK FACTOR DATA ANALYSIS

BRIAN DESNOYERS, ALANKRIT JOSHI, RAHUL KONDAKRINDI, PRASANNA VIKASH PEDDINTI, JUNYI WANG

AGENDA

- Introduction
- Exploratory Analysis
- Data Mining Analysis
- Discussion

AGENDA

- Introduction
- Exploratory Analysis
- Data Mining Analysis
- Discussion

BACKGROUND

- Obesity is a major challenge facing the healthcare system in the United States
- United States healthcare system is projected to pay $150 billion annually (Hurt et al., 2010)
- Environmental changes can be tied to specific regions and have a significant impact on health (National Heart, Lung, and Blood Institute, 2013; National Institute of Diabetes and Digestive and Kidney Diseases, 2012)

NUTRITION, PHYSICAL ACTIVITY, AND OBESITY DATASET

- From Centers for Disease Control and Prevention (CDC)
- Percentage of the population suffering from adult obesity, as well as associated behaviors
- Potential risk factor features

OBJECTIVE:

EXPLORE THE
DISTRIBUTION
OF SOCIOECONOMIC RISK FACTORS
ACROSS THE UNITED STATES AND
VISUALIZE
HOW THESE GROUPINGS CORRESPOND
TO OBESITY AND HEALTH

PREPROCESSING

- Reading dataset- already in CSV format
- Extract useful columns
- Filter from “question-based” format to numeric features aggregated by state
- Calculate overweight percentages to include obese adults

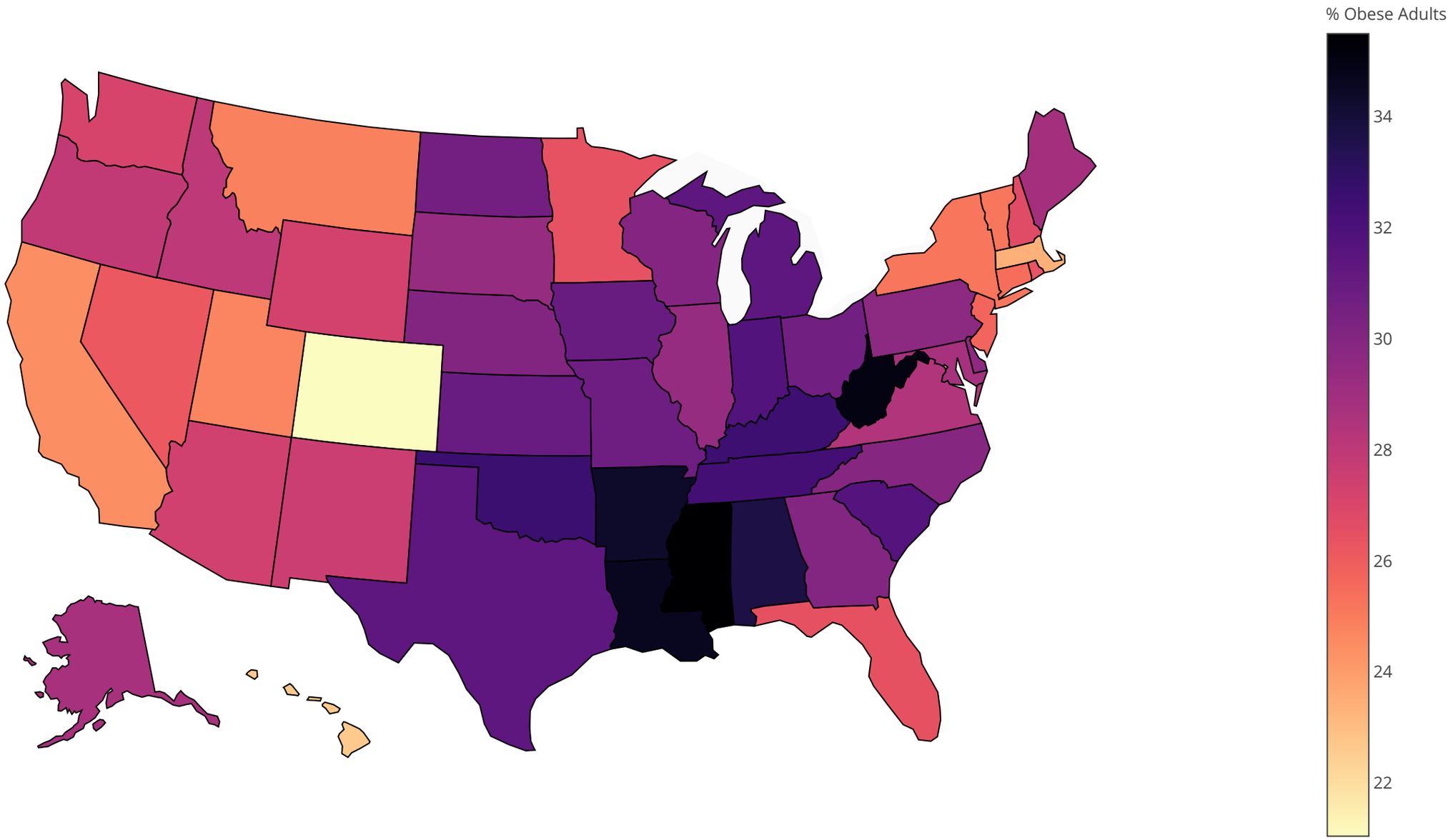
AGENDA

- Introduction
- Exploratory Analysis
- Data Mining Analysis
- Discussion

VISUALIZING FEATURES BY STATE

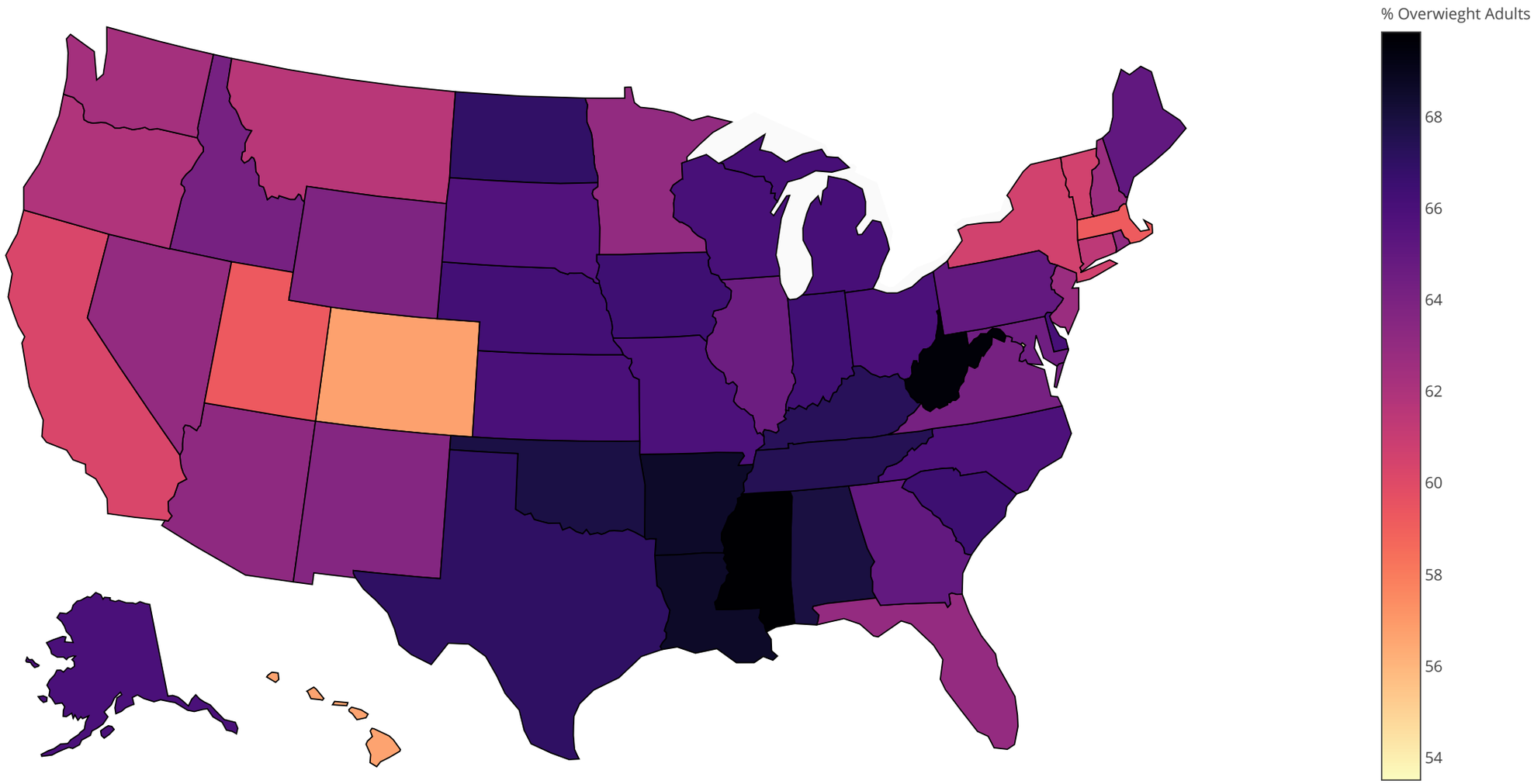
- Percentage of adults who:
  - are obese
  - are overweight
  - are inactive
  - engage in vigorous aerobic activity
  - eat less than one serving of fruit daily
  - eat less than one serving of vegetables daily

PERCENTAGE OF OBESE ADULTS BY STATE

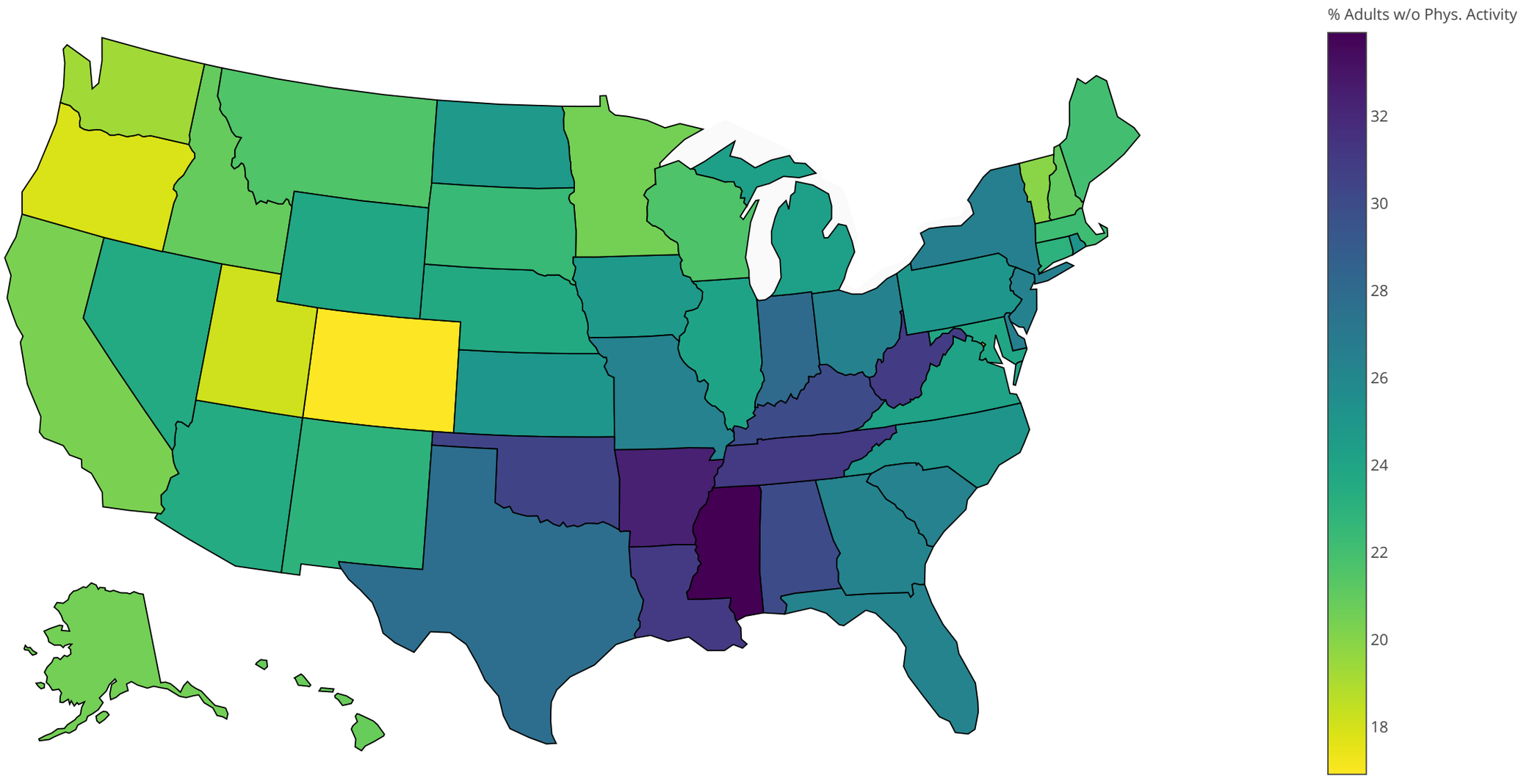


DEMO

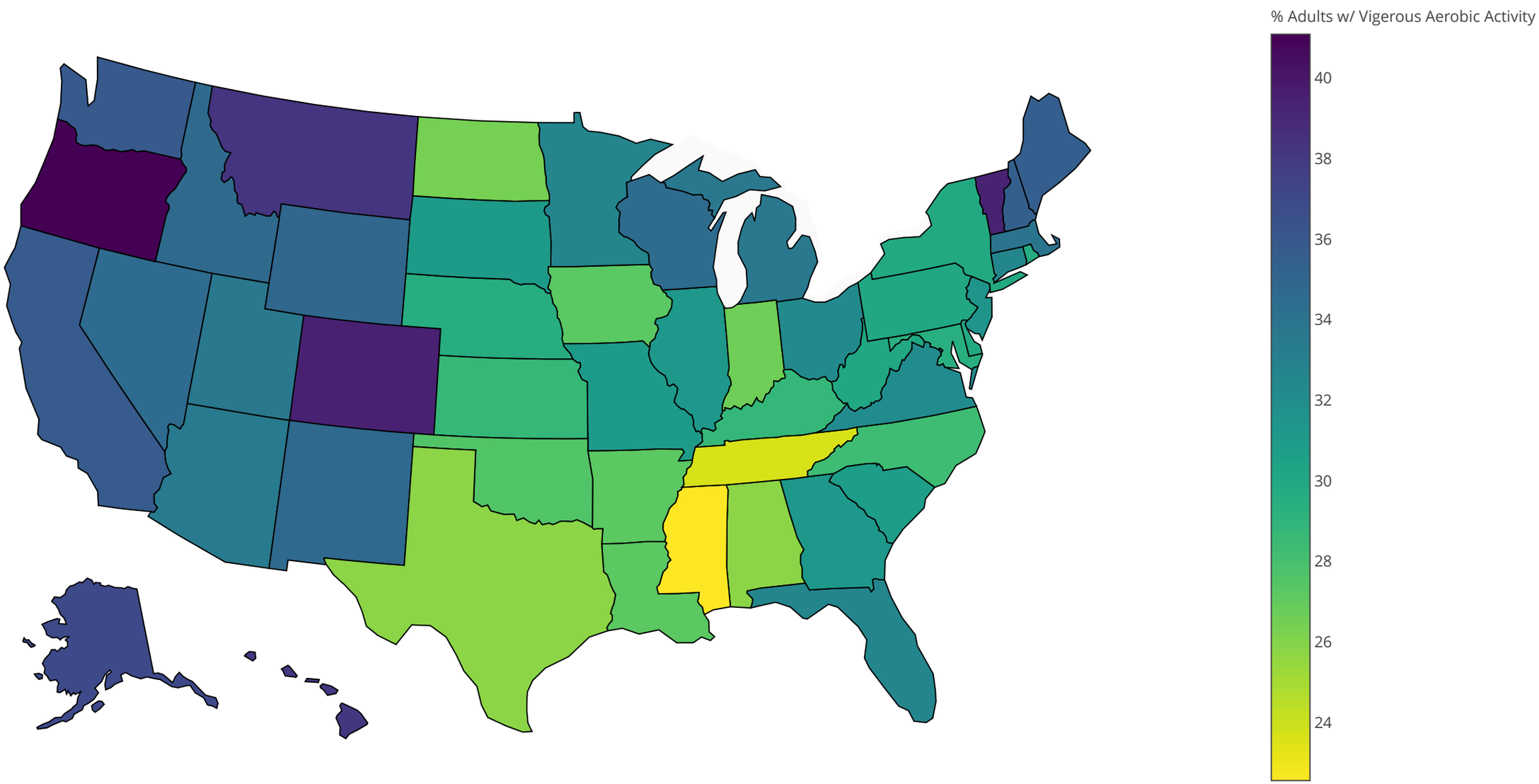
PERCENTAGE OF OVERWEIGHT ADULTS BY STATE



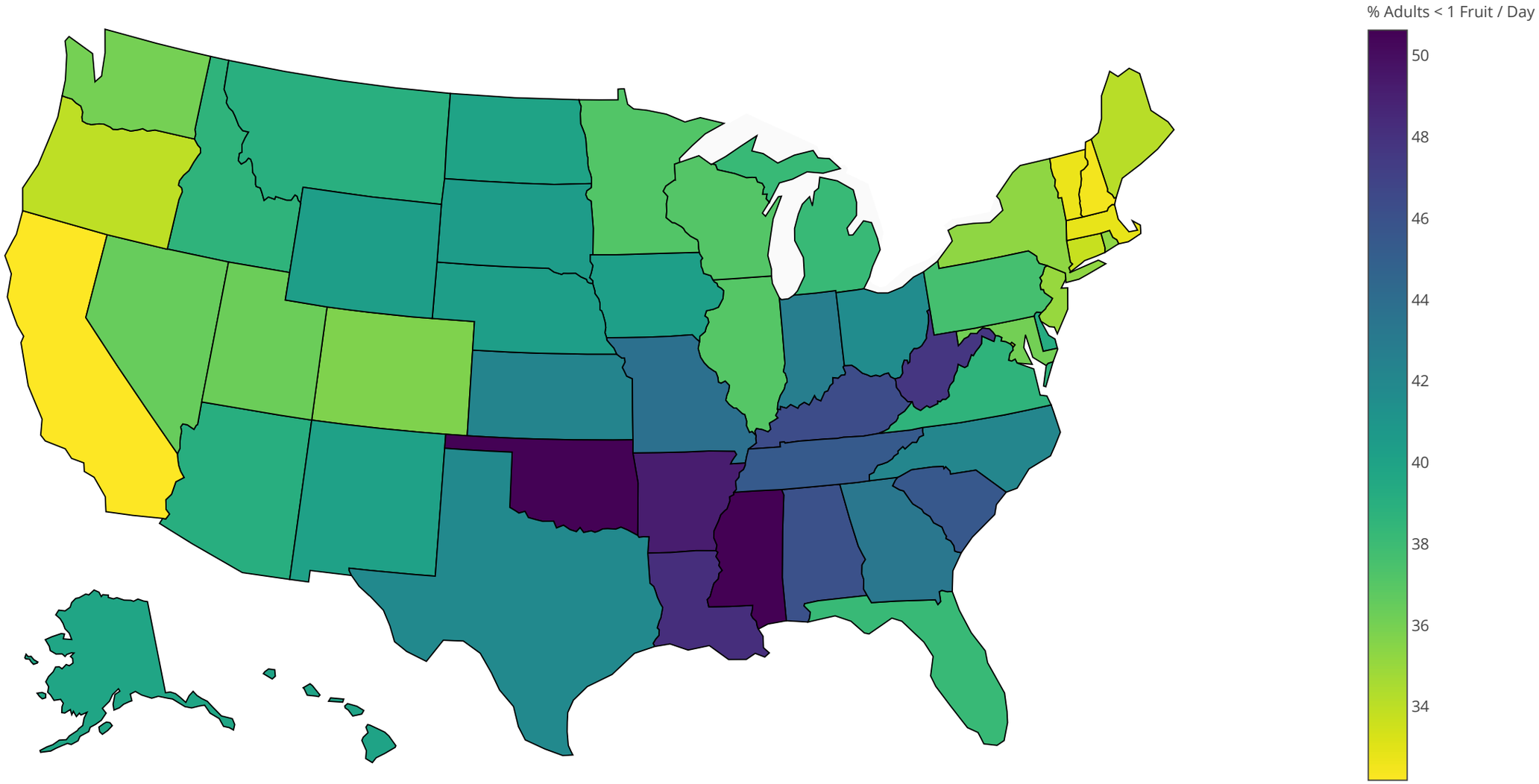
PERCENTAGE OF PHYSICALLY INACTIVE ADULTS BY STATE



PERCENTAGE OF ADULTS W/ VIGOROUS AEROBIC ACTIVITY BY STATE

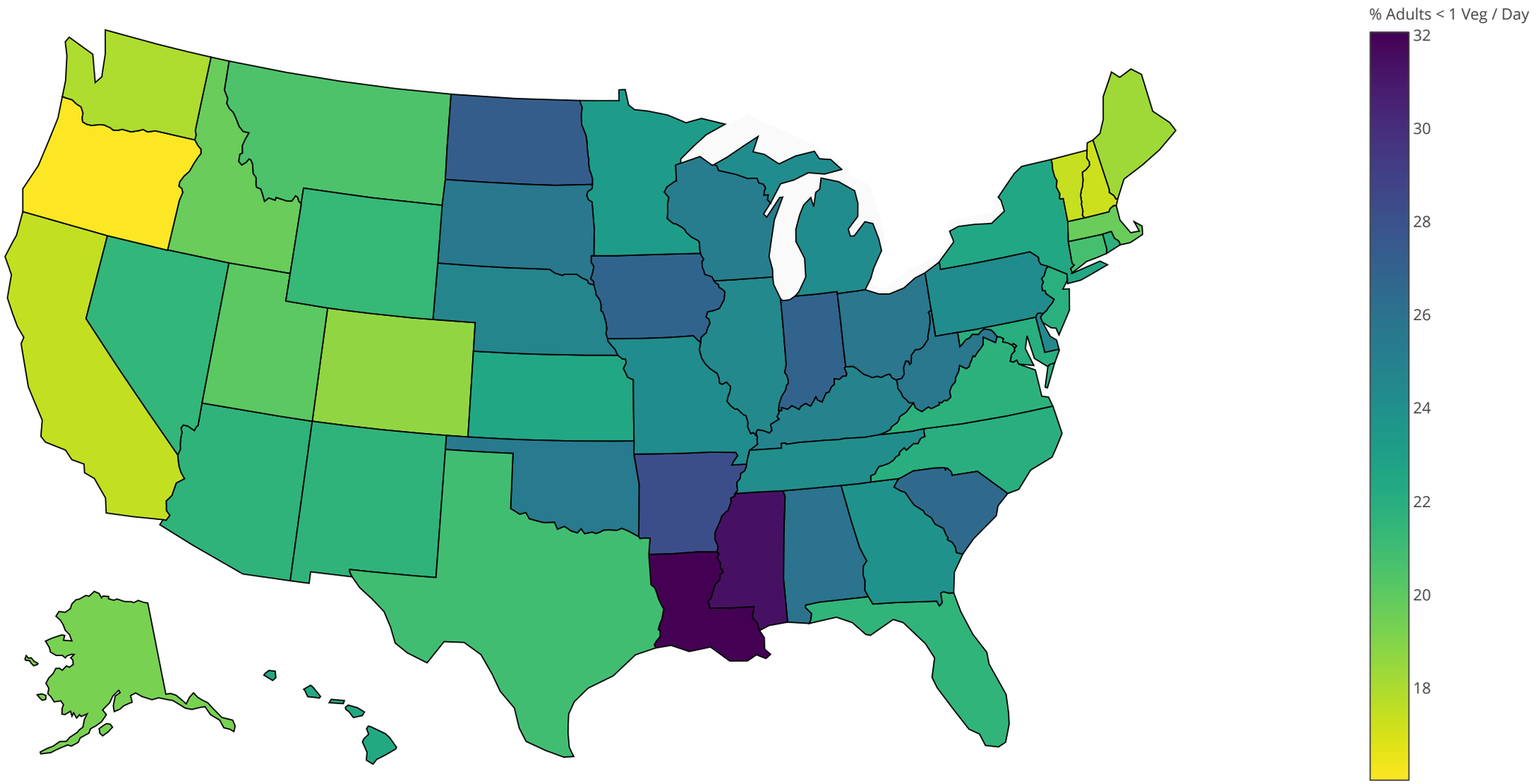


PERCENTAGE OF ADULTS WITH FRUIT MALNUTRITION* BY STATE



* defining fruit malnutrition as consuming less than one serving of fruit per day

PERCENTAGE OF ADULTS WITH VEGETABLE MALNUTRITION* BY STATE

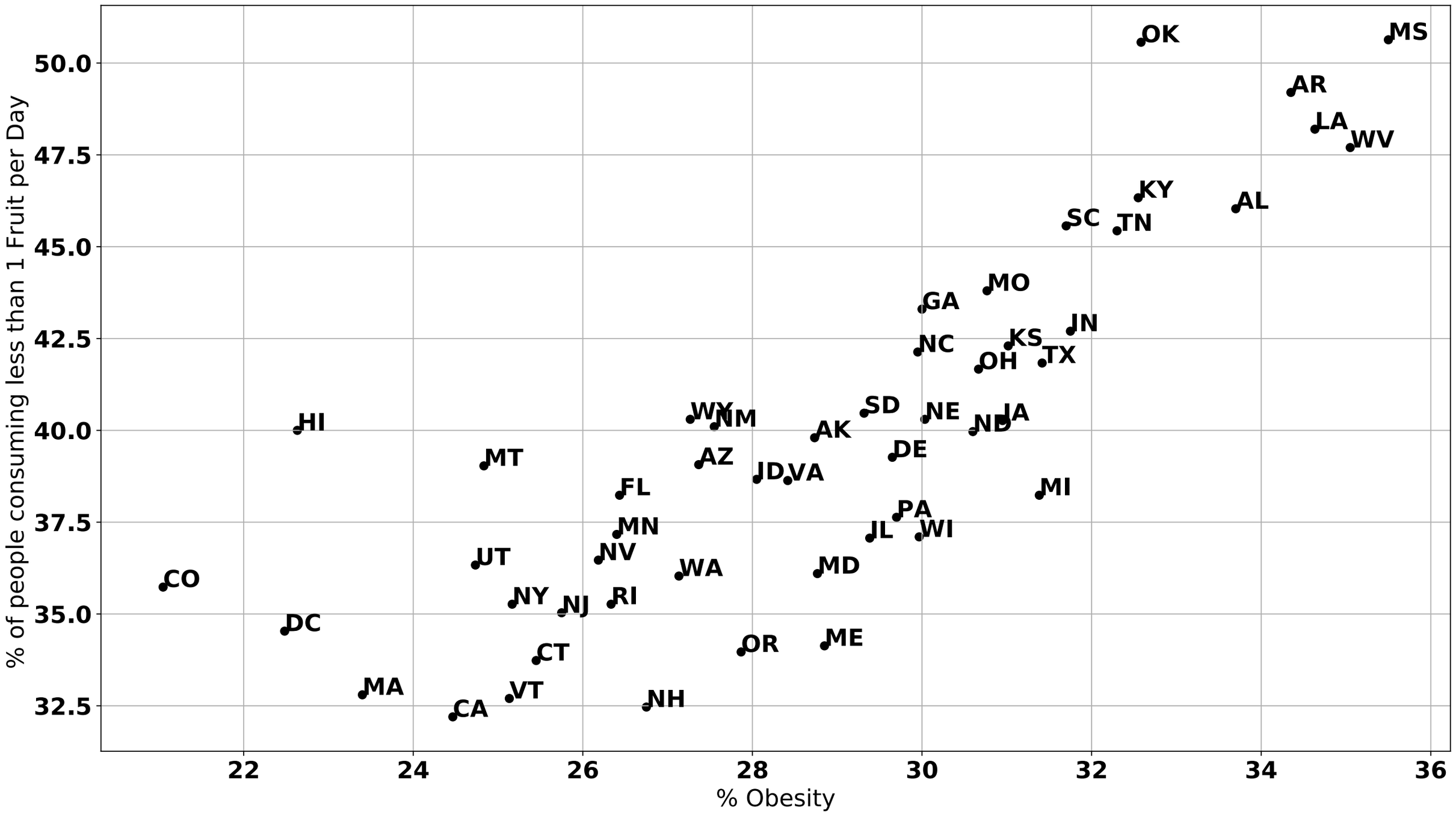


* defining fruit malnutrition as consuming less than one serving of vegetables per day

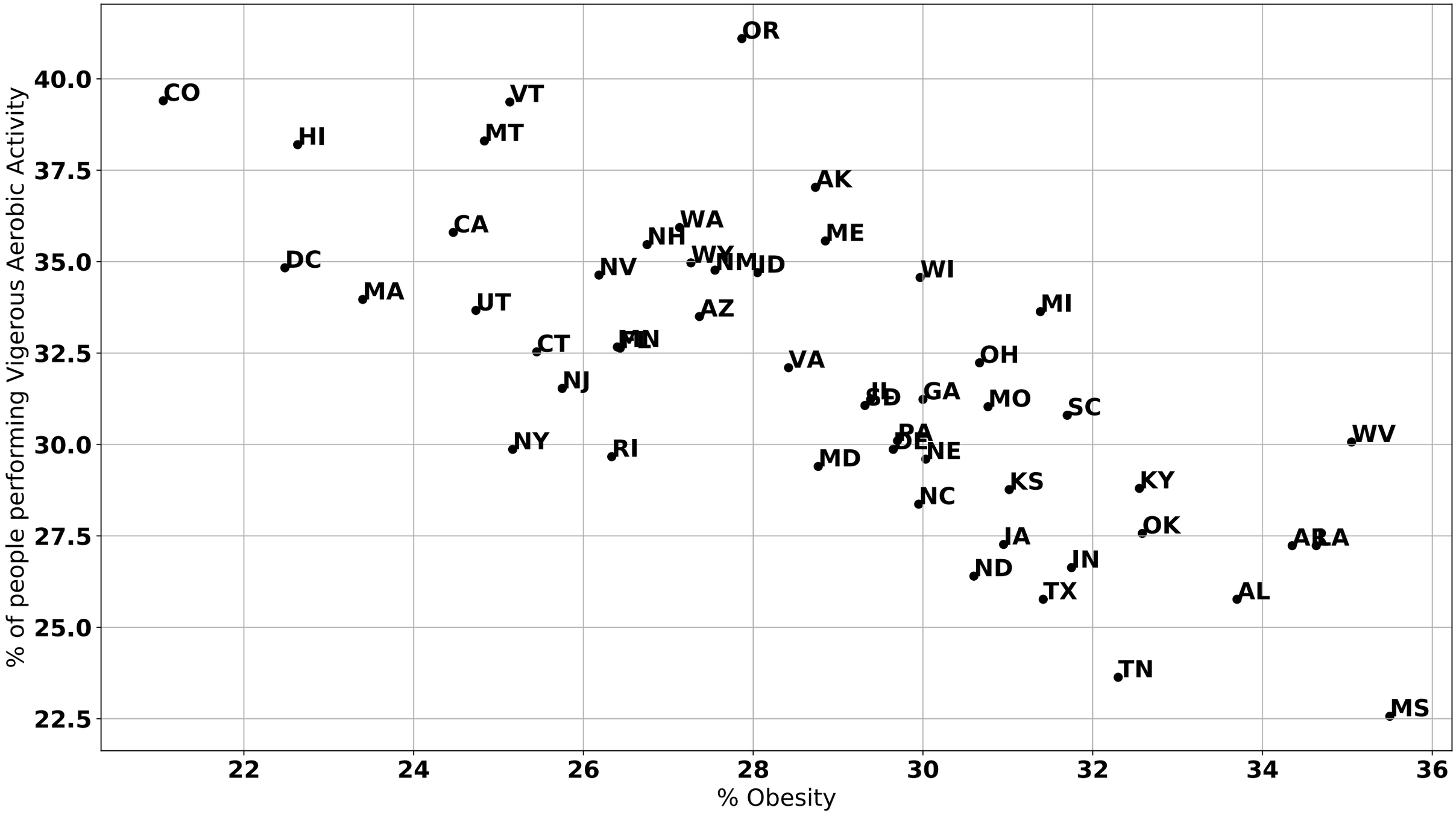
PRELIMINARY TREND VISUALIZATION

- Visualized state-wise relationships of:
  - fruit malnutrition and obesity
  - vigorous aerobic activity and obesity

STATE-WISE FRUIT MALNUTRITION AND OBESITY



STATE-WISE VIGOROUS AEROBIC ACTIVITY AND OBESITY



AGENDA

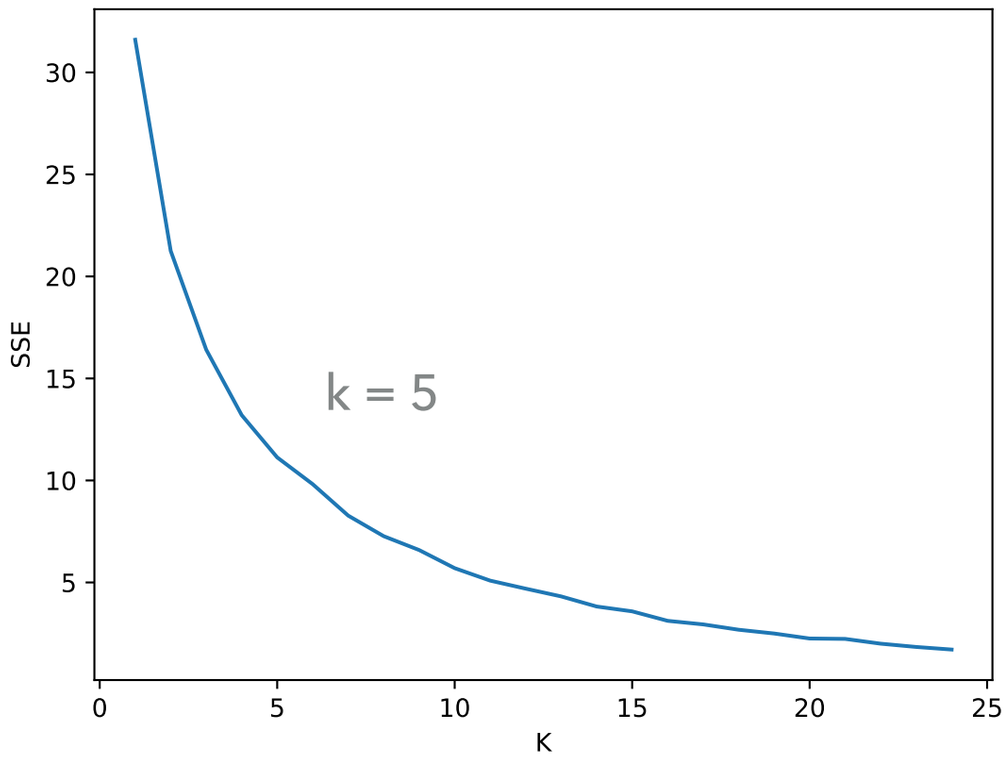
- Introduction
- Exploratory Analysis
- Data Mining Analysis
- Discussion

METHODS

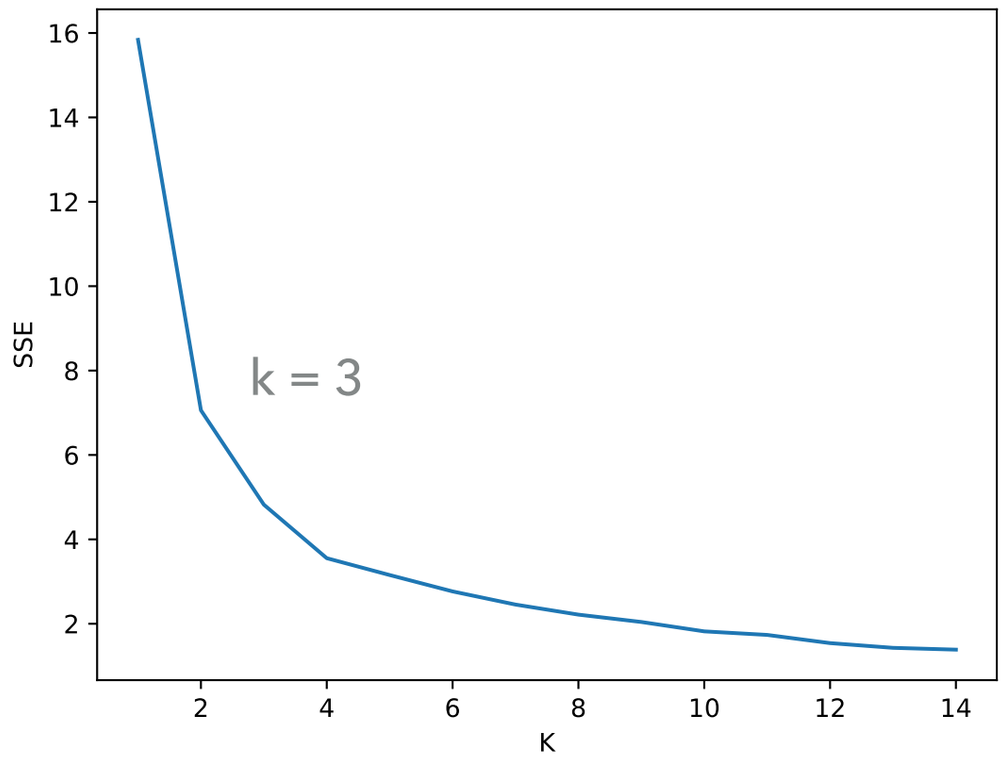
- Performed k-means, DBSCAN, and hierarchical clustering across CDC Health Behaviors and Outcomes and Socioeconomic Risk Factor datasets
- Normalized data to be scaled from 0 to 1

K-MEANS- CHOOSING A VALUE FOR K

- "Elbow" method



SOCIOECONOMIC FACTORS



CDC HEALTH BEHAVIOR DATA

DBSCAN- PARAMETER TUNING

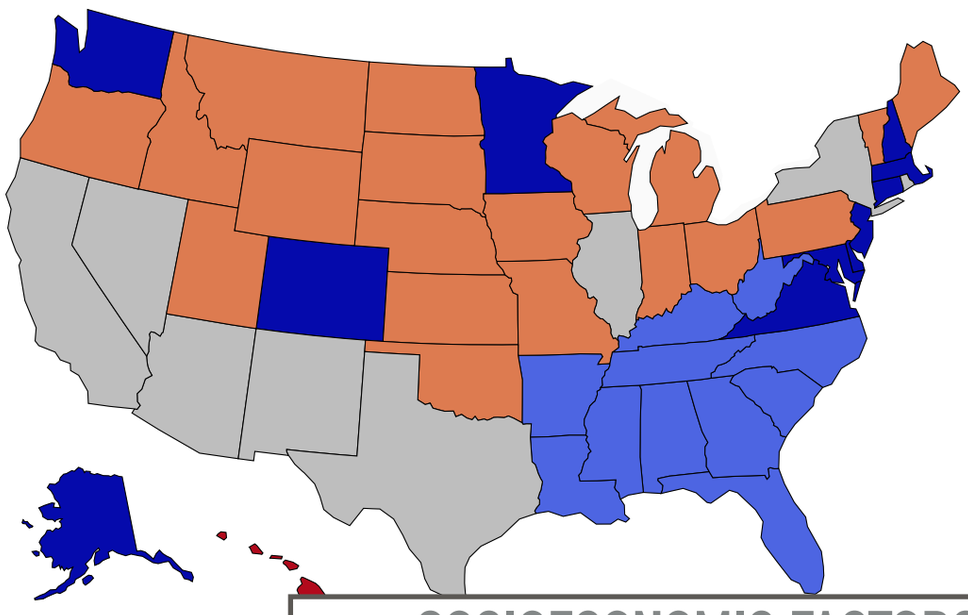
- Tuned ε and minPts to achieve less than 5% noise
- For socioeconomic factors:
  - ε = 0.6 and minPts = 6
- For CDC health behavior data:
  - ε = 0.2 and minPts = 3

HIERARCHICAL CLUSTERING– DISTANCE METRIC

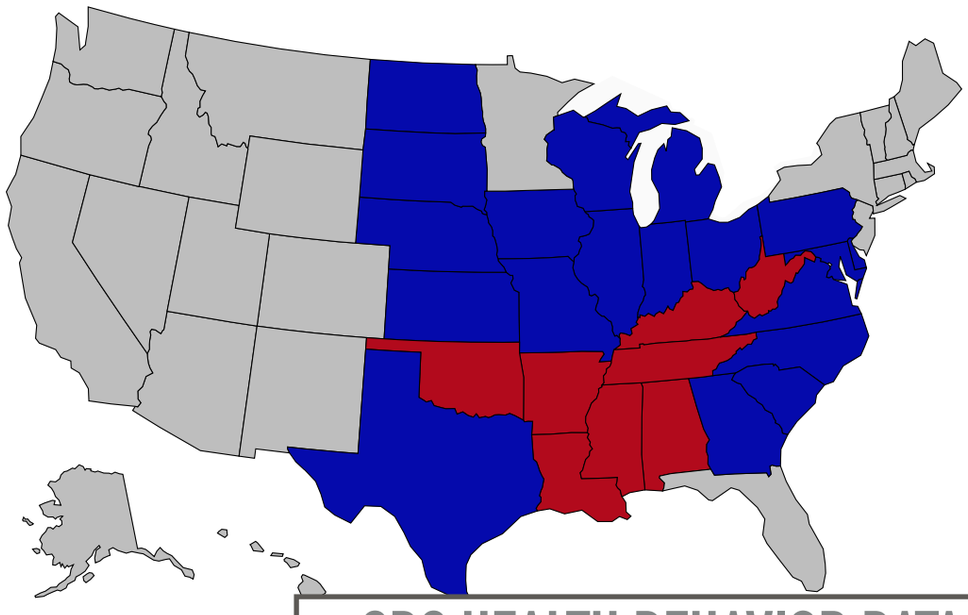
- Selected Ward distance metric after comparing results from single, complete, and average linkage

K-MEANS CLUSTERING RESULTS

- States were grouped into 5 clusters based on Socio-Economic factors
- States were grouped into 3 clusters based on CDC Health Behavior Data
- Similar trends in Socio-Economic conditions and CDC Data can be found in southern states like Texas, Alabama, Louisiana



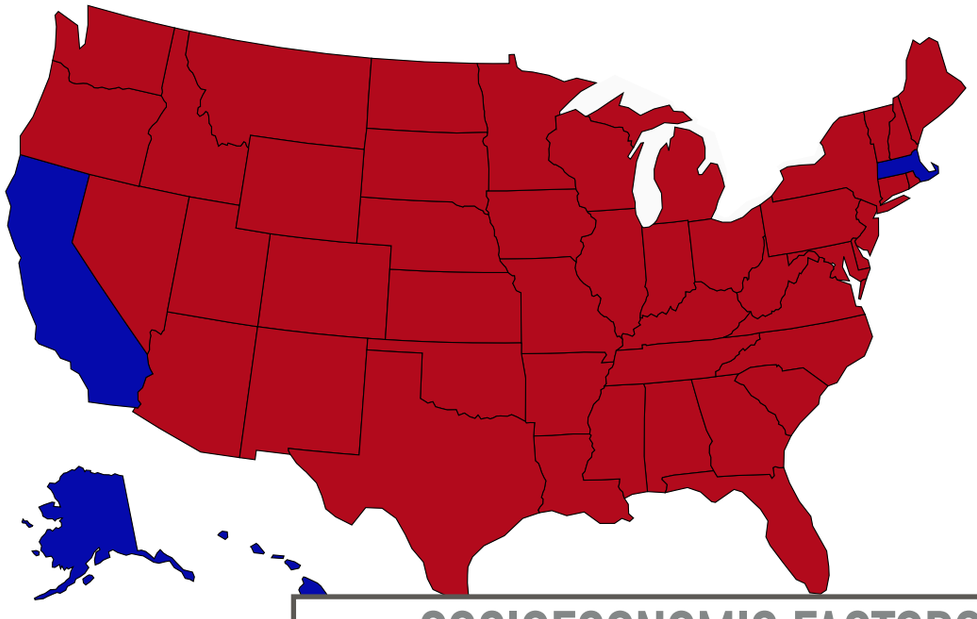
SOCIOECONOMIC FACTORS



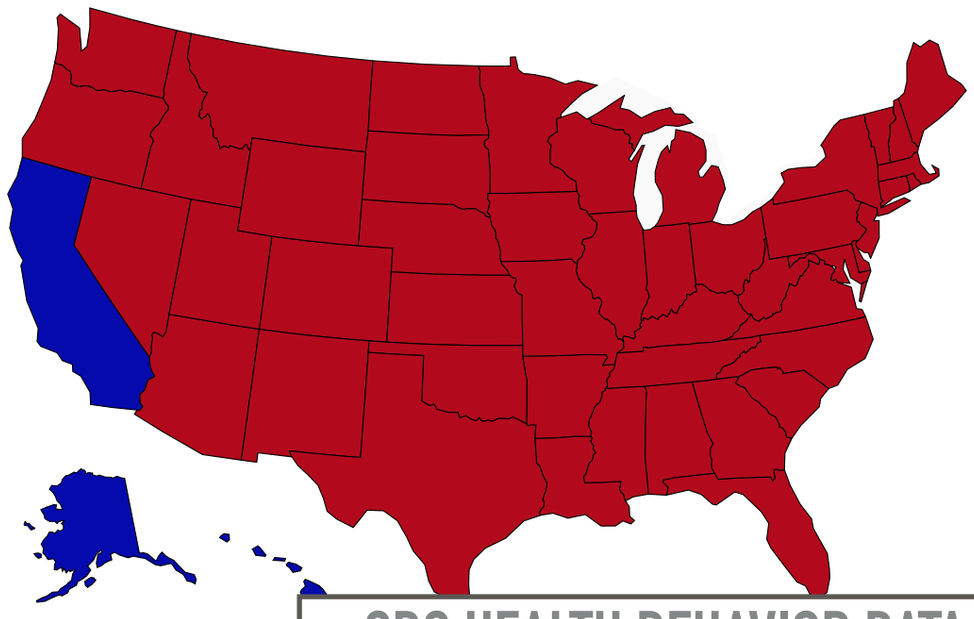
CDC HEALTH BEHAVIOR DATA

DBSCAN CLUSTERING RESULTS

- Based on Socio-Economic data, states were grouped into a single cluster with four states namely California, Massachusetts, Alaska, Hawaii as the noise
- Based on CDC Health Behavior data, states were grouped into a single cluster with three states namely California, Alaska, Hawaii as the noise points
- The consistent clustering highlights the relation between the Socio-Economic factors and Health & Obesity trends

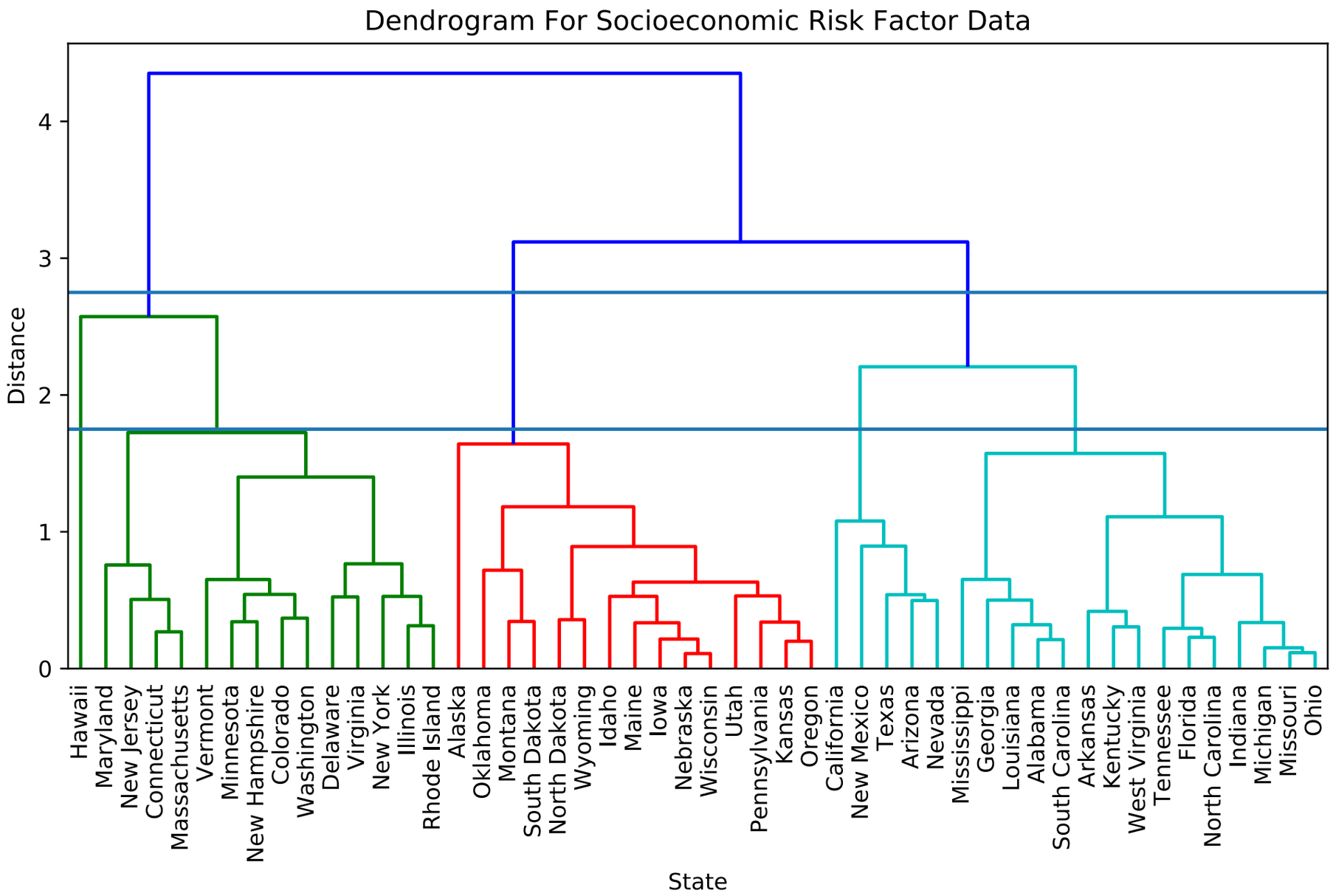


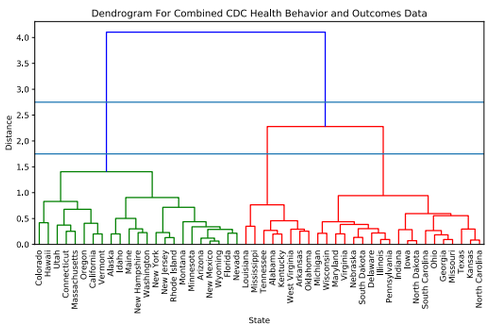
SOCIOECONOMIC FACTORS



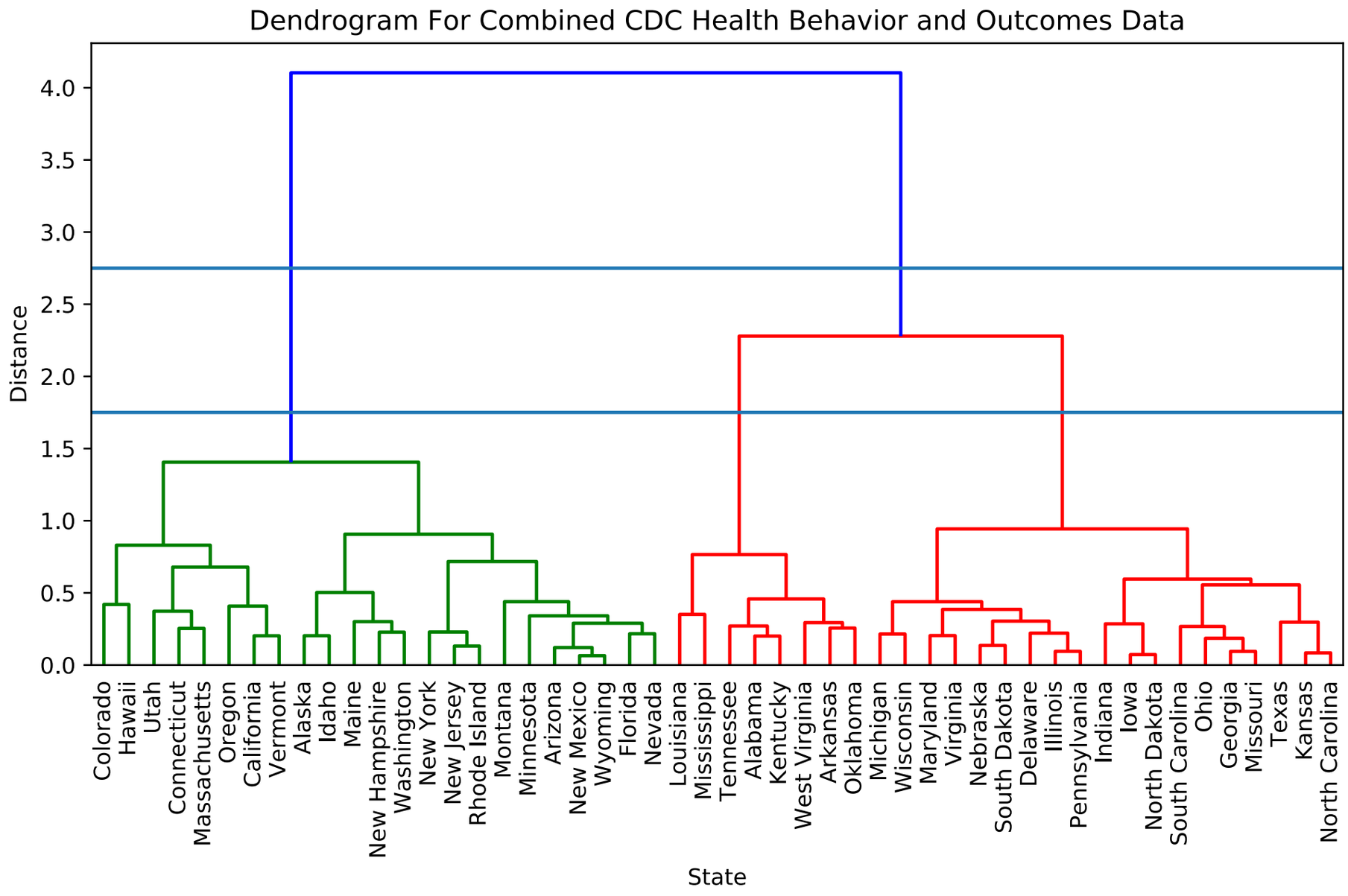
CDC HEALTH BEHAVIOR DATA

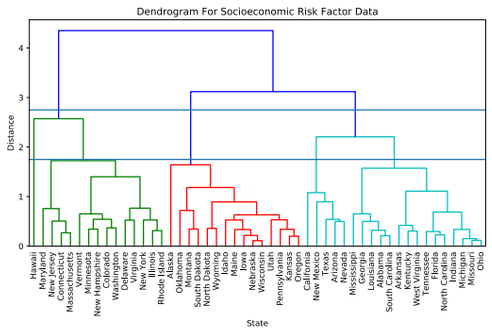
AGGLOMERATIVE HIERARCHICAL CLUSTERING RESULTS



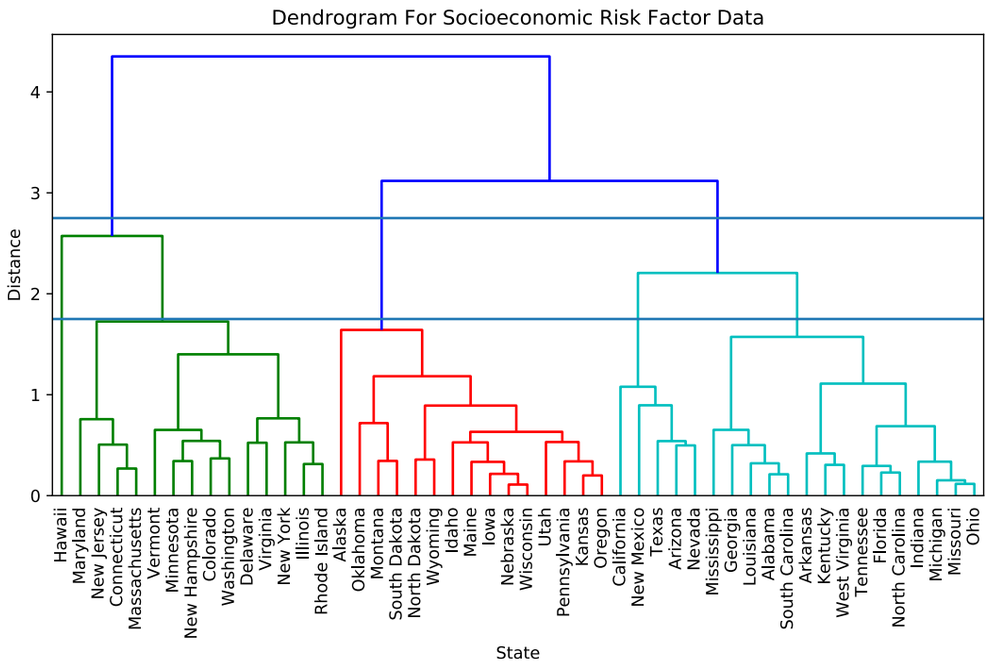


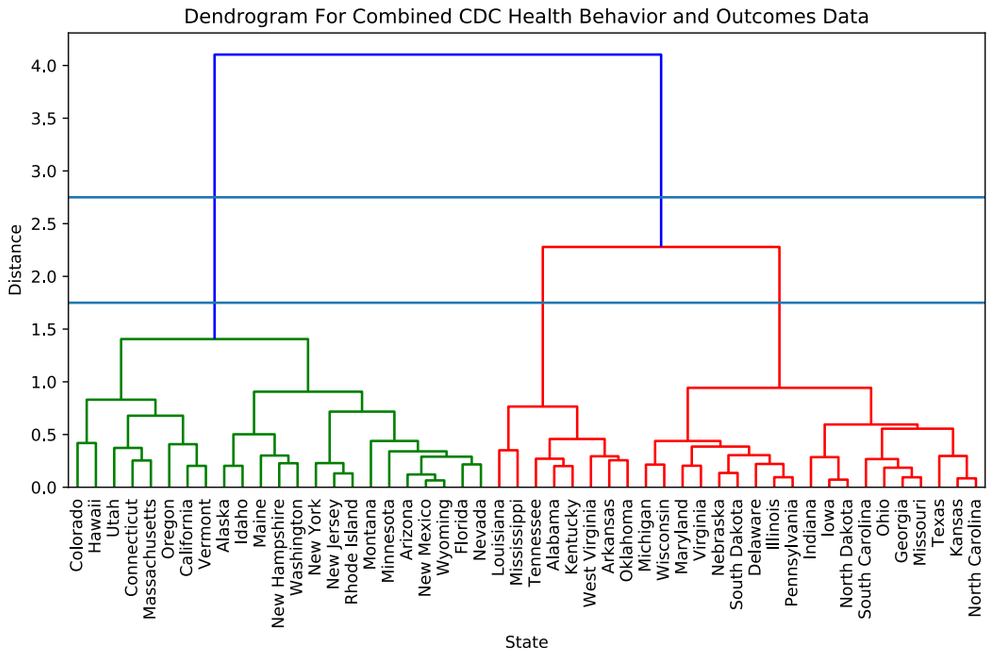
AGGLOMERATIVE HIERARCHICAL CLUSTERING RESULTS





AGGLOMERATIVE HIERARCHICAL CLUSTERING RESULTS





AGENDA

- Introduction
- Exploratory Analysis
- Data Mining Analysis
- Discussion

INTERPRETING PRINCIPAL COMPONENTS– CDC HEALTH BEHAVIOR

PC1		PC2	
% < 1 Fruit / Day	0.459181	% Vigorous Aerobic	-0.627665
% Obesity	0.418273	% Overweight	-0.410559
% Inactive	0.411431	% Obesity	-0.409127
% Overweight	0.407122	% Inactive	0.329247
% Vigorous Aerobic	-0.379775	% < 1 Fruit / Day	-0.29556
% < 1 Veg/ Day	0.367353	% < 1 Veg/ Day	0.272643

INTERPRETING PRINCIPAL COMPONENTS– SOCIOECONOMIC RISK

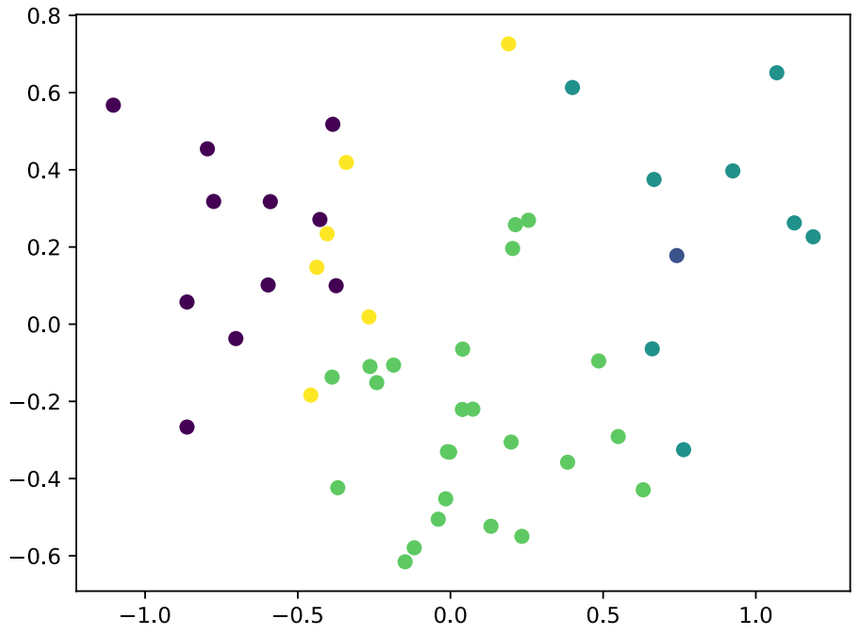
PC1

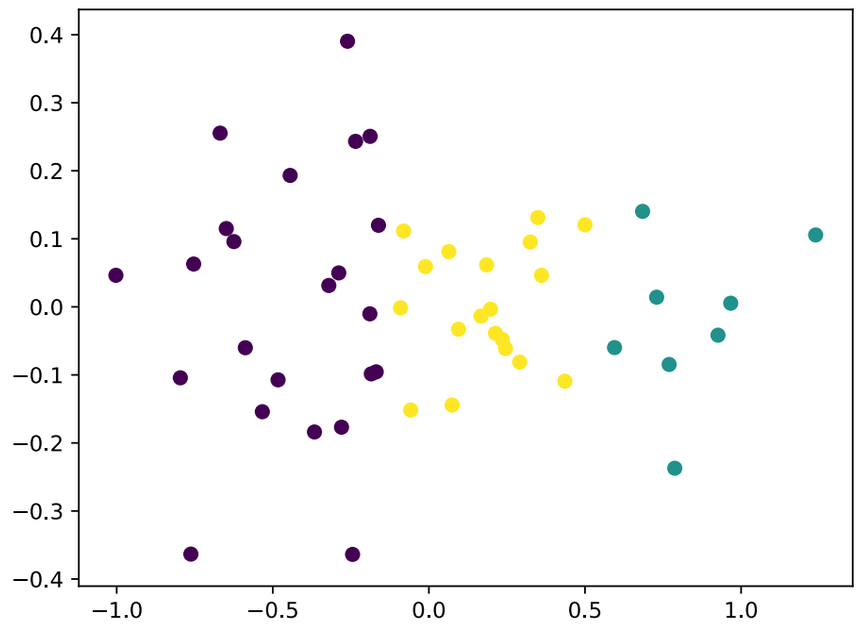
$ Median family	0.461288
$ Median household	0.452155
$ Per capita income	0.389594
% Bachelor's degree	0.382095
% Advanced degree	0.370484
% High school	0.318128
% Black or African	-0.143937
% Asian	0.109871
% Some other race	0.072519
% Two or more races	0.070956
% Native Hawaiian	0.052622
% White	-0.025696
% American Indian	-0.003568

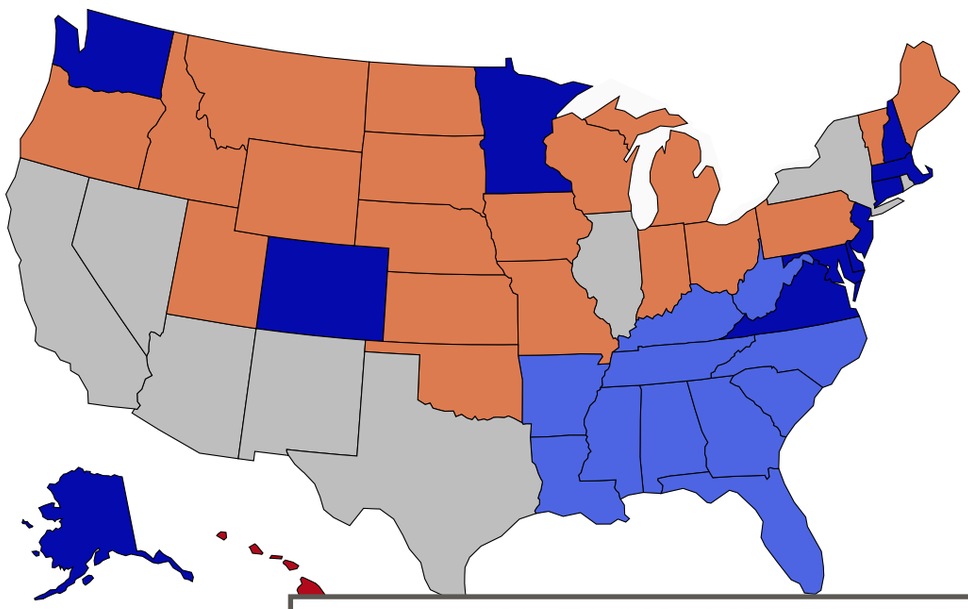
PC2

% High school	-0.575076
% Black or African	0.526907
% White	-0.370121
% Some other race	0.322323
% Advanced degree	0.254567
% American Indian	-0.218517
% Asian	0.127889
% Bachelor's degree	0.094785
$ Median household	0.077591
$ Per capita income	0.072180
$ Median family	0.033356
% Native Hawaiian	0.018870
% Two or more races	0.013513

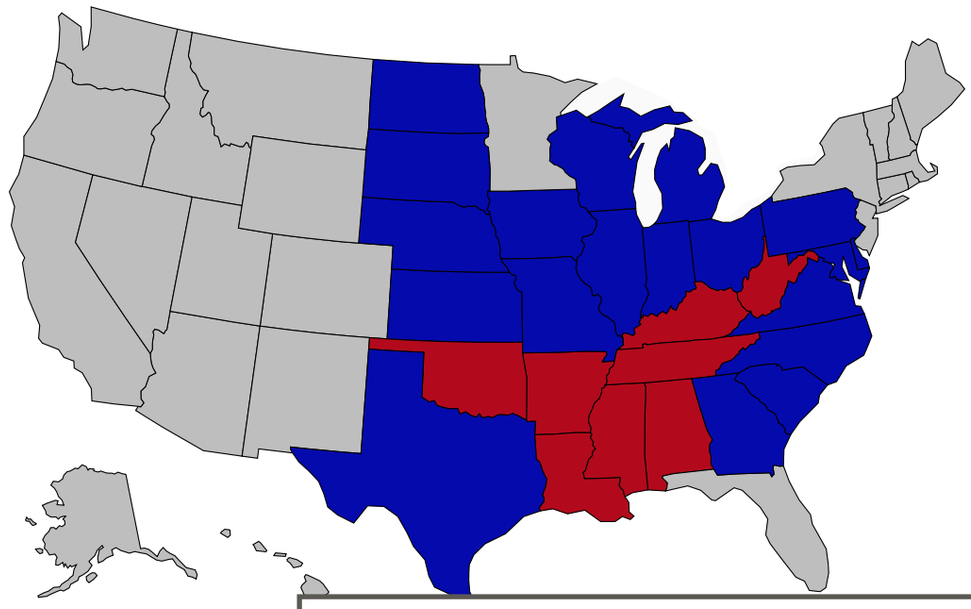
K-MEANS CLUSTERING





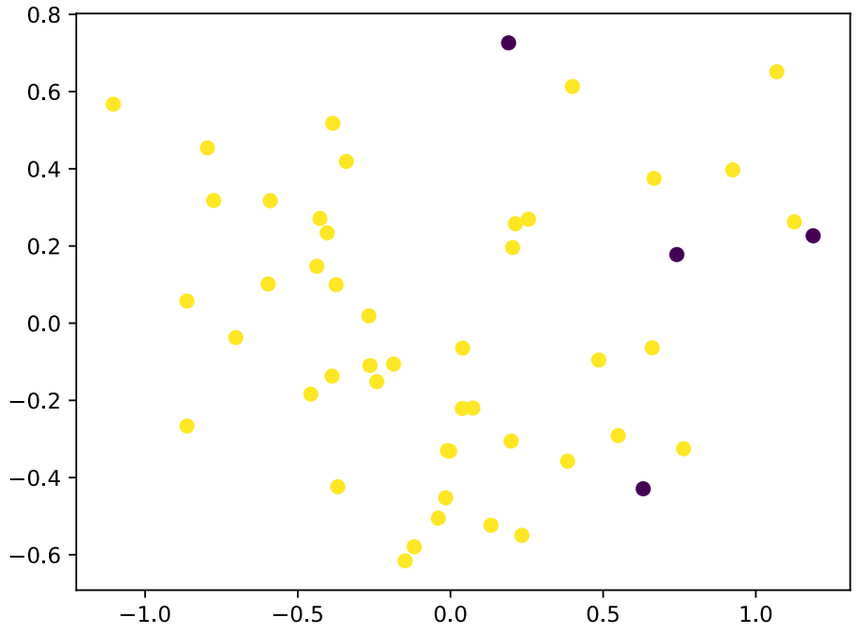


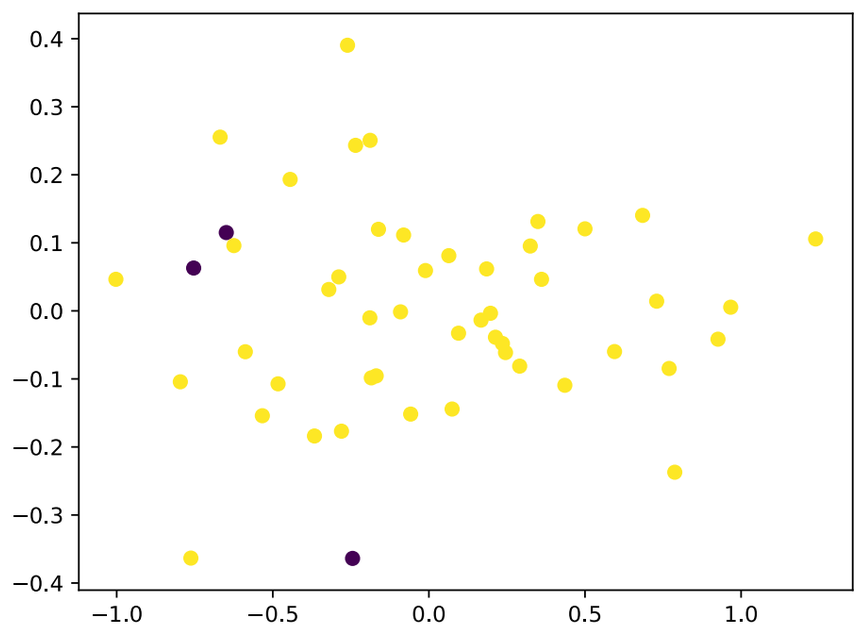
SOCIO ECONOMIC FACTORS

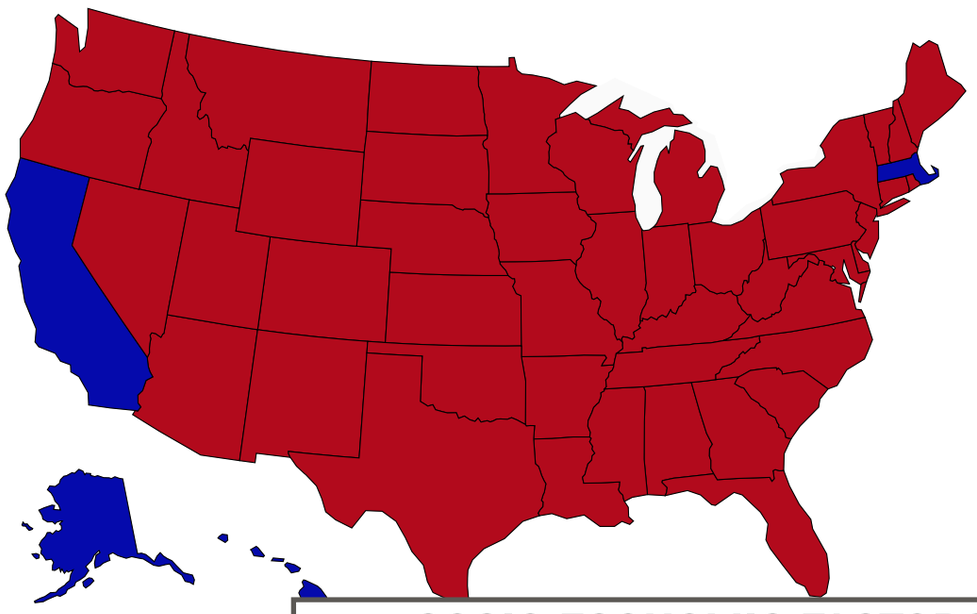


CDC HEALTH BEHAVIOR DATA

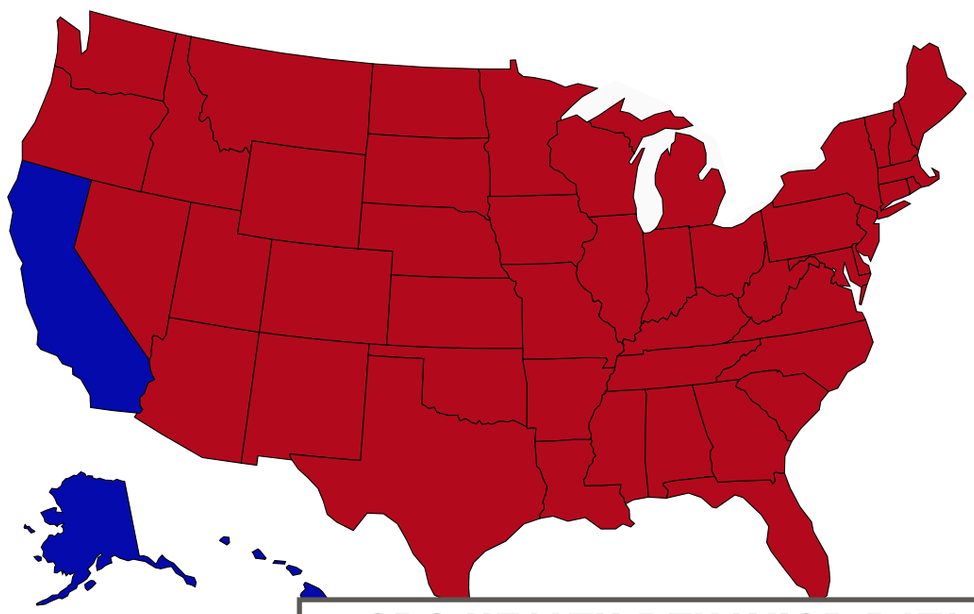
DBSCAN CLUSTERING





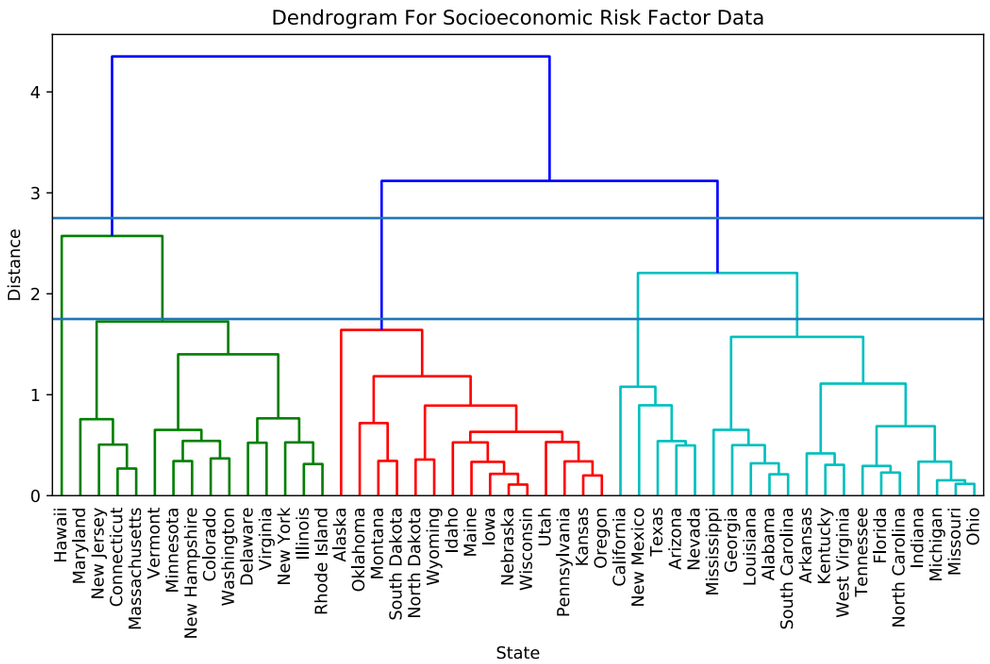


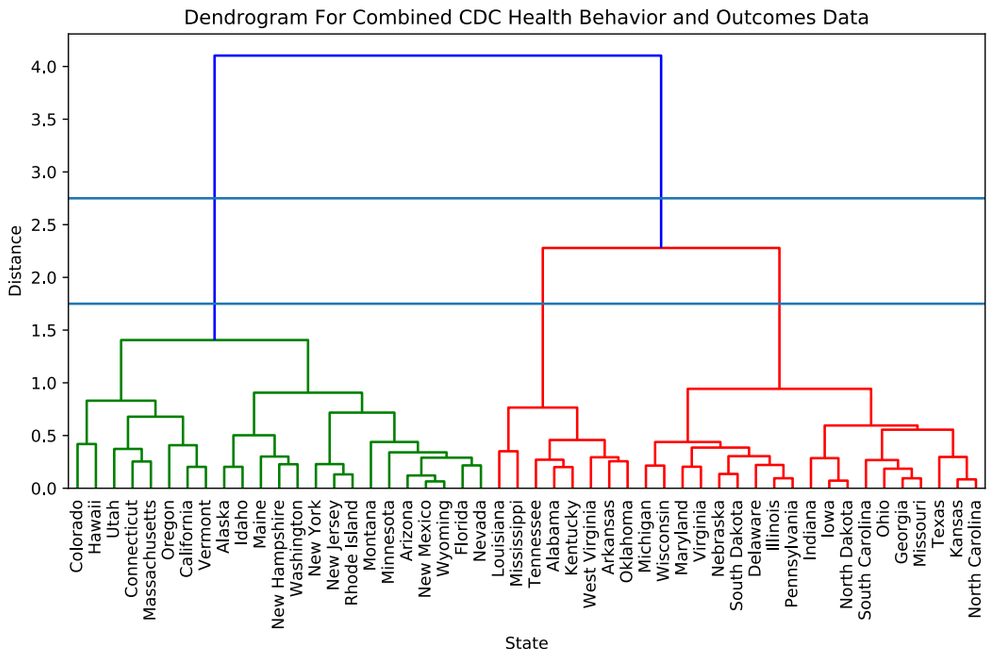
SOCIO ECONOMIC FACTORS



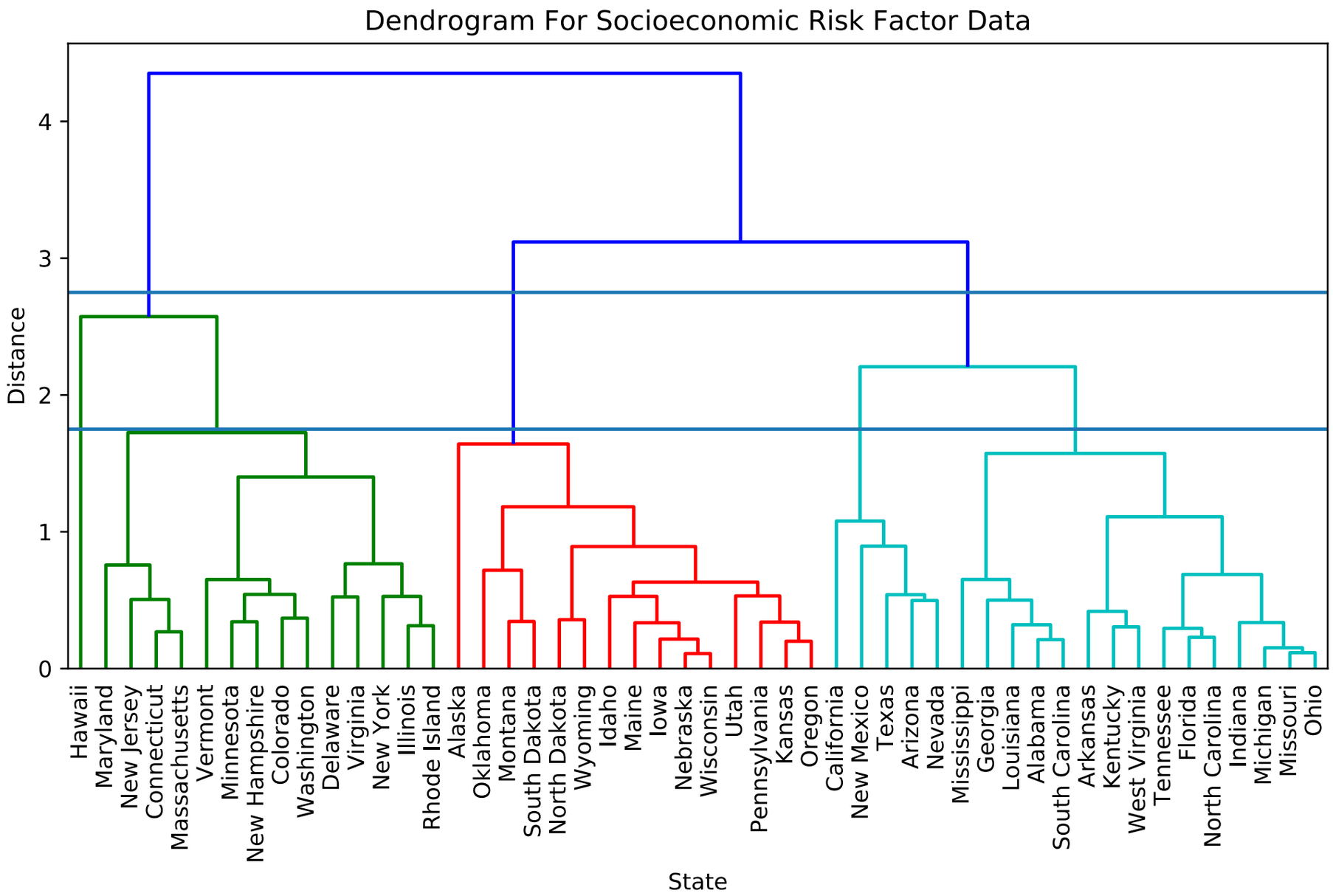
CDC HEALTH BEHAVIOR DATA

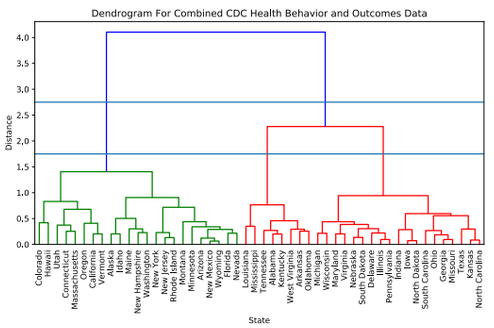
AGGLOMERATIVE HIERARCHICAL CLUSTERING





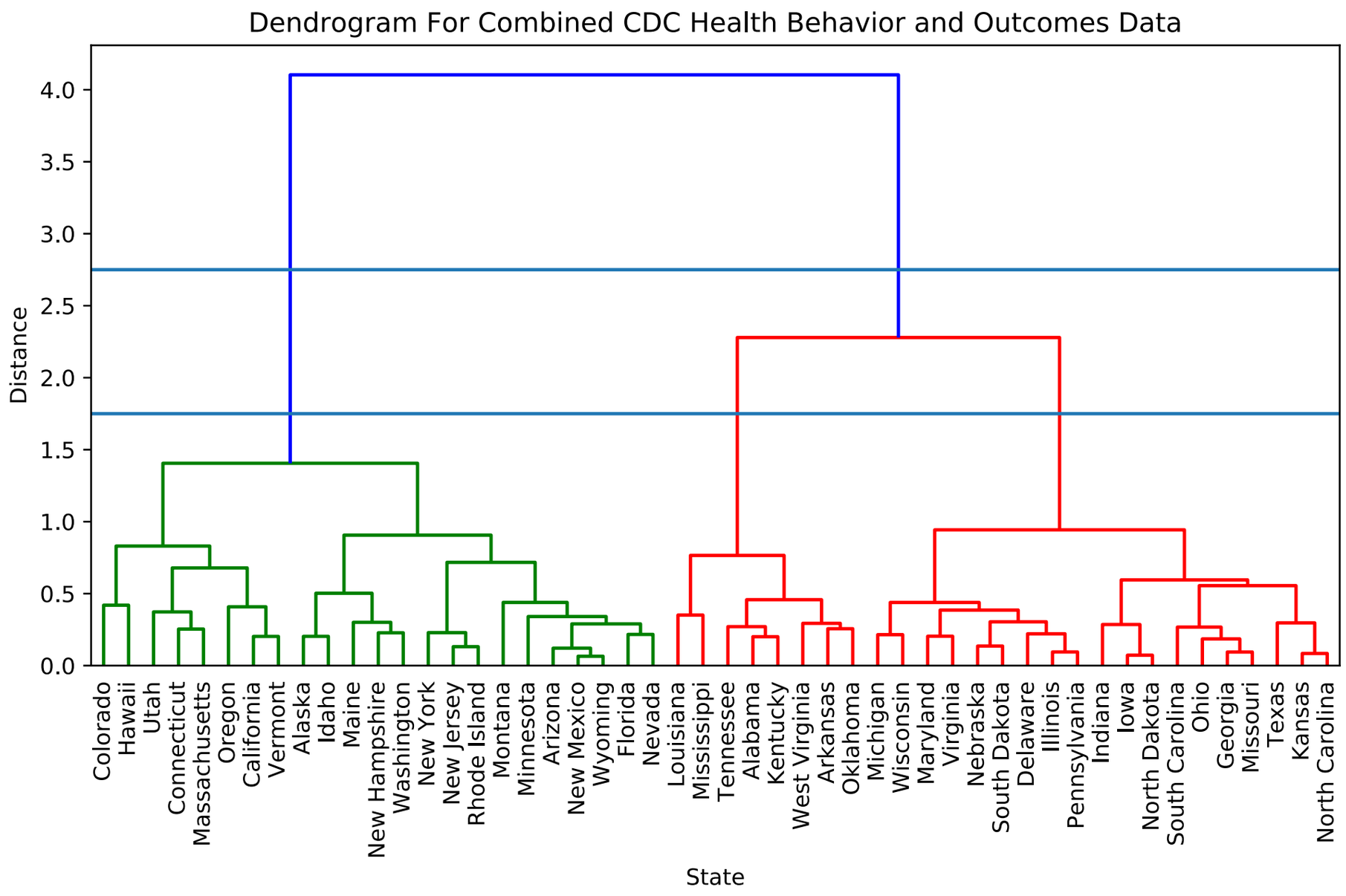
AGGLOMERATIVE HIERARCHICAL CLUSTERING





cophenetic correlation distance = 0.549033284833

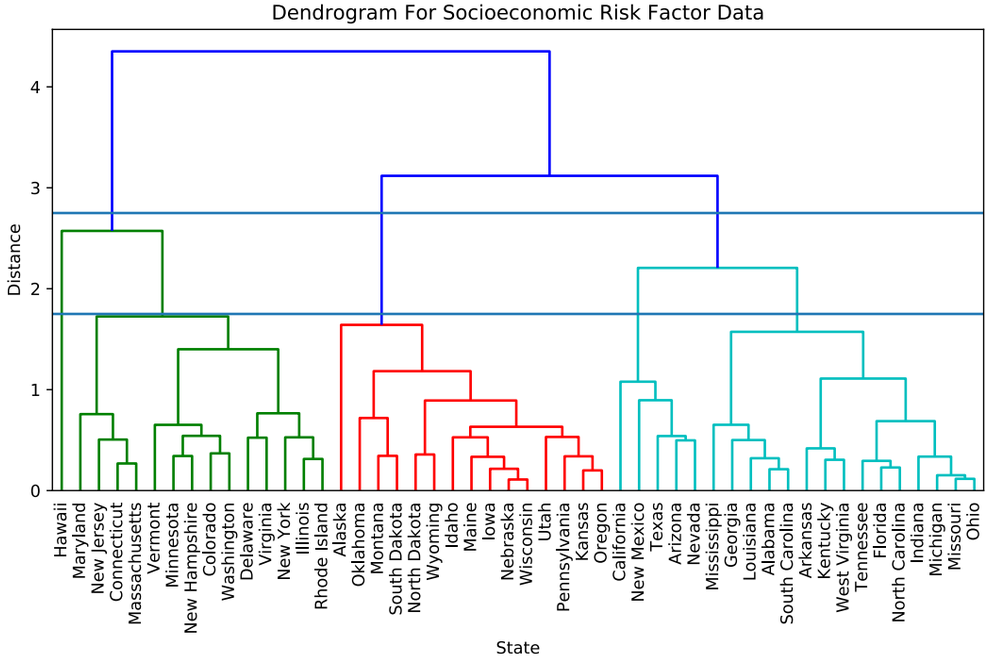
AGGLOMERATIVE HIERARCHICAL CLUSTERING

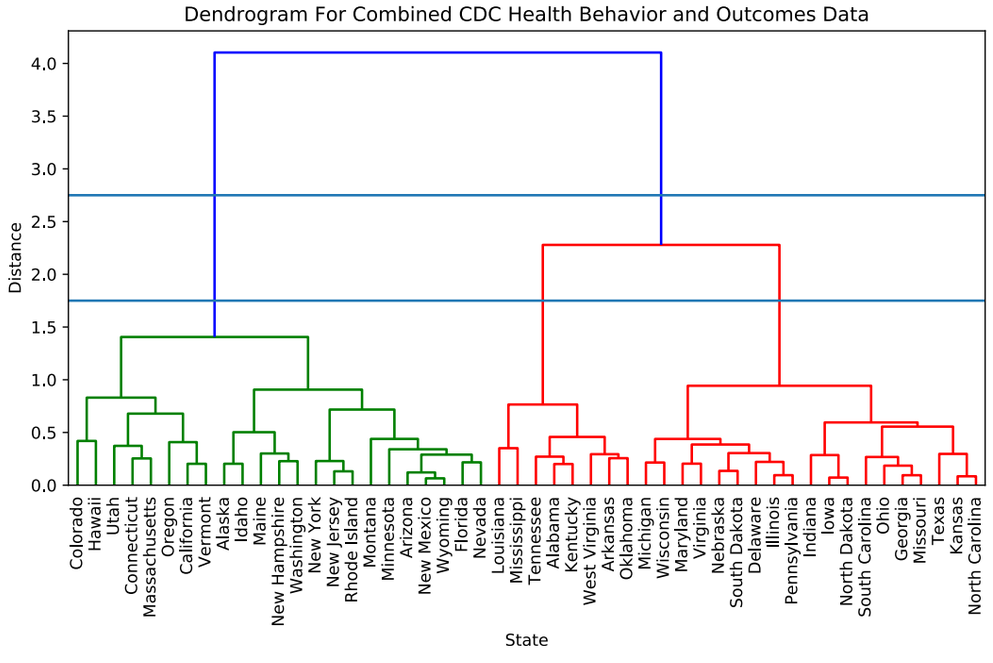


cophenetic correlation distance = 0.597008648578

SIMILARITY BETWEEN CLUSTERINGS

- Percent similarity- percentage of times that a pair of states in the same/different cluster in one clustering is also in the same/different cluster for the other





% SIMILARITY = 56.8979591837%