BRFSS - Behavioral Risk Factors Surveillance System

1. Introduction

BRFSS is one of the largest ongoing health surveys in the world, designed to track modifiable risk factors for chronic diseases and other major causes of death. This project uses survey data from the CDC's Behavioral Risk Factor Surveillance System (BRFSS), collected annually from 1984 through 2023. Because it's state-based, it captures long-term patterns in health behaviors across the U.S. population. Such a wide dataset offers exceptional opportunity to examine how health patterns differ across demographics.

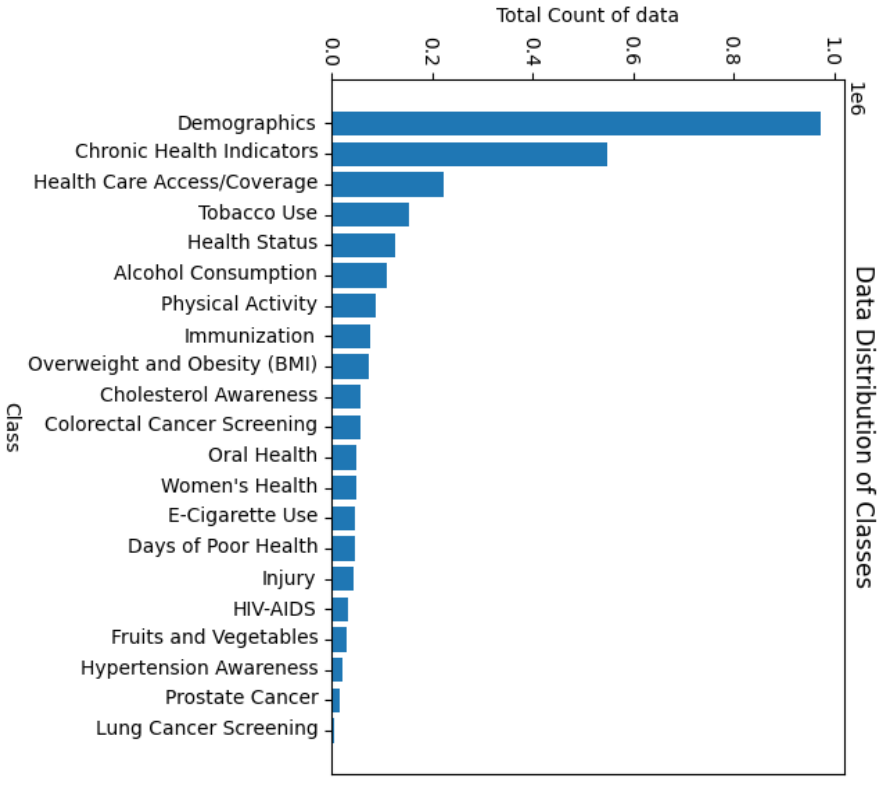
We chose this dataset because it offers a person view of health behavior like age, income, location, race/ethnicity, etc., which can be scientifically meaningful and socially impactful. This vast variety of data helped us explore the idea of how different groups experience different risks and outcomes. This can also help us understand disparities across racial/age or socio-economic groups.

Survey data like this is especially valuable for machine learning because it gives us clues about why certain behaviors or conditions show up. If a model can learn then it can provide insights to help predict similar patterns for future respondents or targeted populations.

The overall goal of this project is to visualize relationship between features in the dataset and identify meaningful patterns. By understanding this we can build predictive models, targeted interventions and do better resource planning.

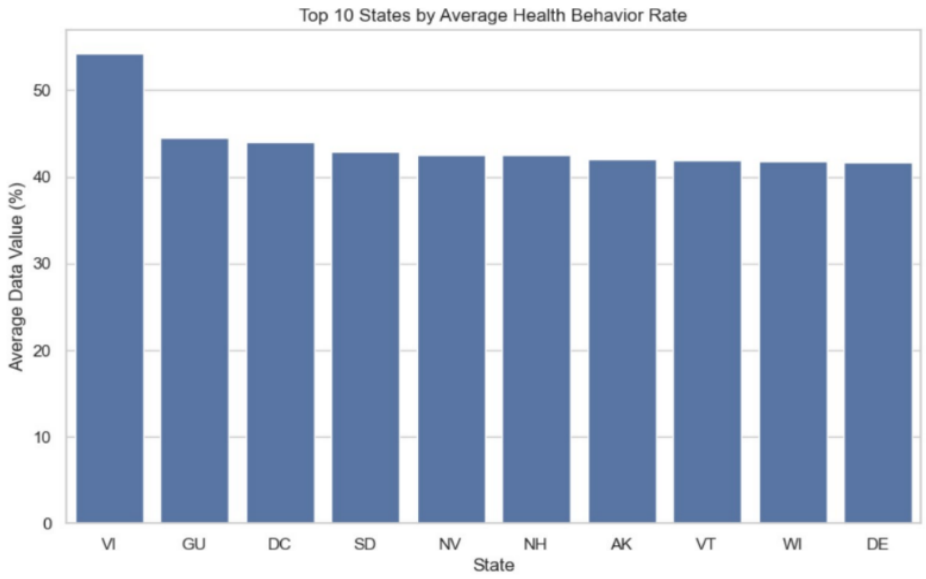
2. Figures

Figure 1: Class distribution of Survey Responses



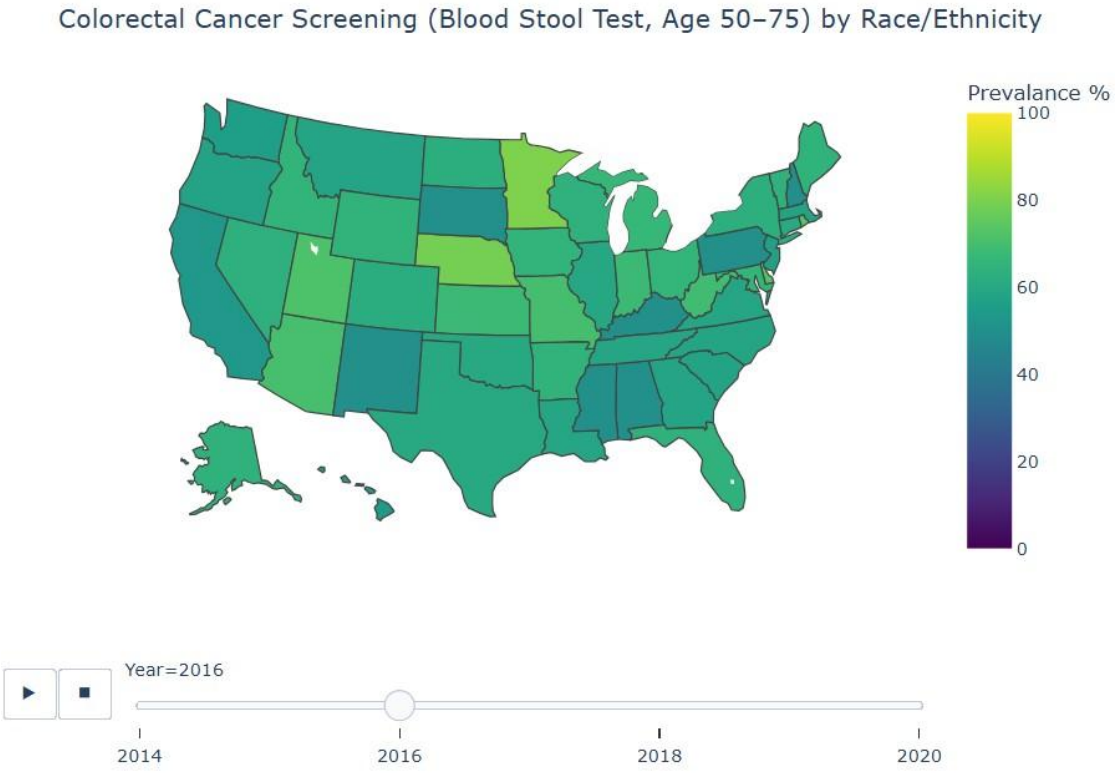
This bar chart shows the distribution of survey responses across major health classes. The demographic category like disability status, education, and employment has largest number of responses while categories like Lung Cancer Screening have significantly fewer responses.

Figure 2: Top 10 states by average health behavior



This bar plot shows mean health behavior percentages by state. Virgin Islands display the highest averages while mainland states Delaware, Wisconsin and Vermont have slightly lower but similar averages. Overall, it visually ranks these locations by their health behavior performance.

Figure 3: Choropleth map of USA for colorectal cancer screening



This choropleth map visualizes colorectal blood stool test screening records for adults aged between 50-75 years in the year 2016. Minnesota shows the highest prevalence while New Mexico shows the lowest prevalence

Figure 4: Bar chart of K-Means clustering on survey responses

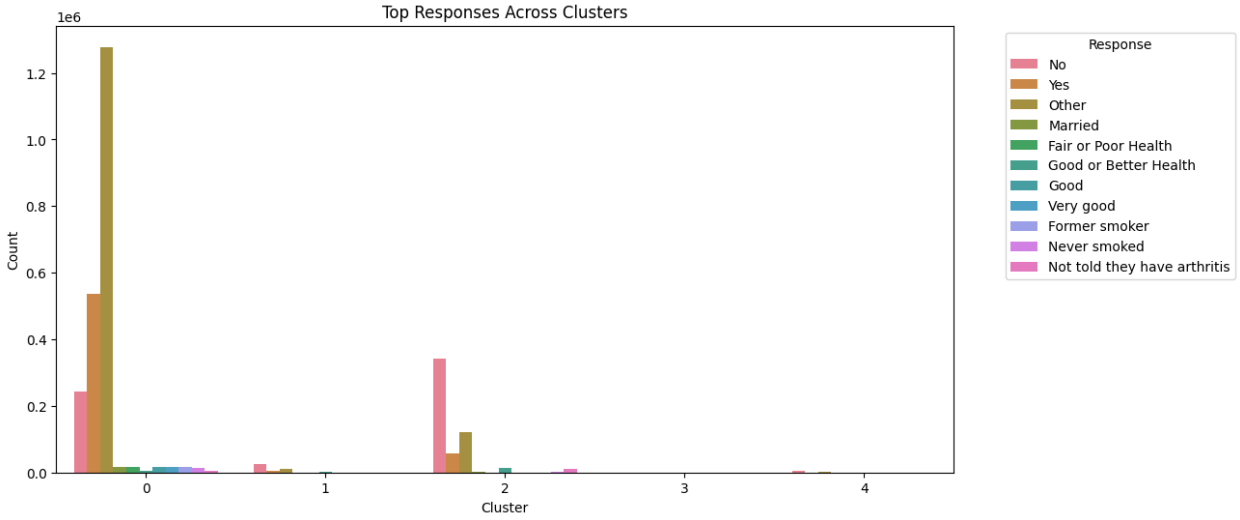


Figure 5: K-Means clustering on cancer data subset

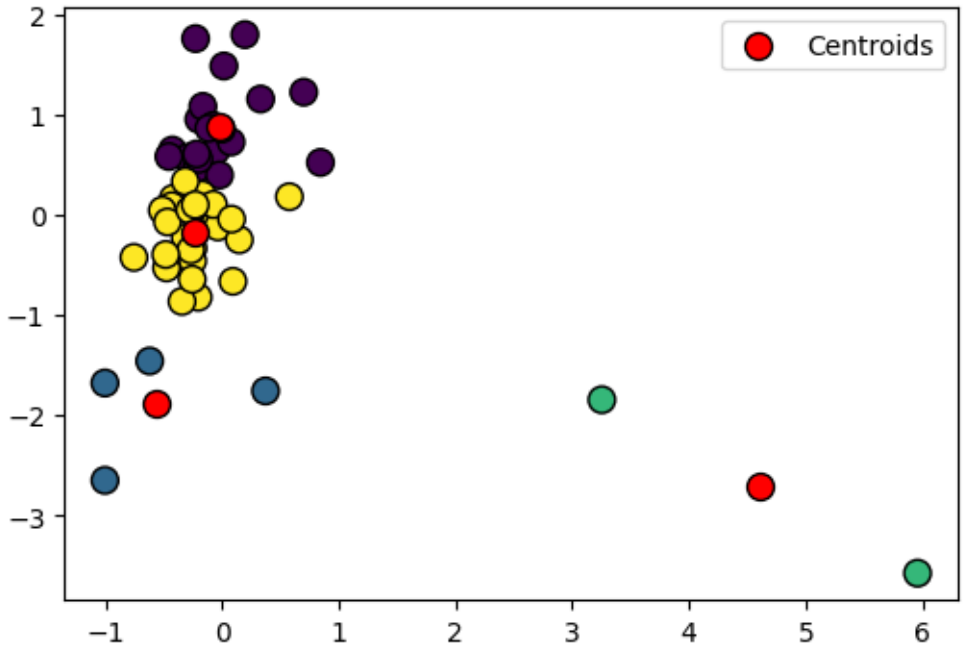
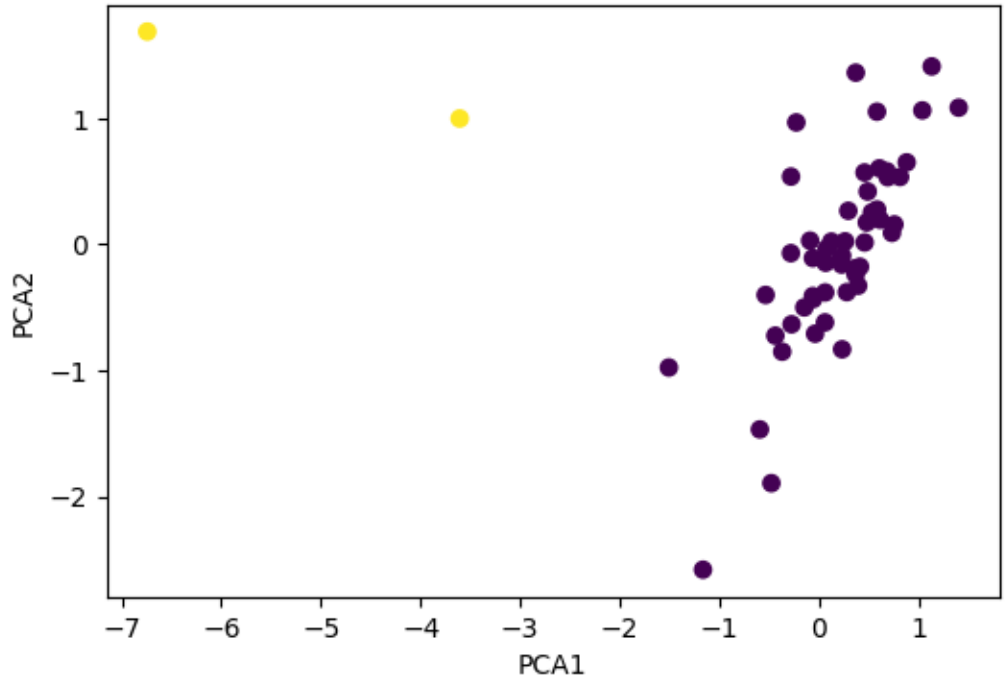


Figure 6: PCA + K-Means on cancer data subset



3. Methods

3.1 Data Exploration

We analyzed the BRFSS dataset which consisted of ~27M rows and 27 columns. We visualized different class distribution Fig.1 against the total number of survey responses. We identified states with high and low health indicator averages in Fig.2 and explored colorectal cancer screening rates using 2016 subset in Fig.3. We created a cancer specific subset by filtering class containing cancer to focus screening behavior in this area,

3.2 Data Preprocessing

We dropped two columns Data_Value_Footnote and Data_Value_Footnote_Symbol because it had more than 80% null values.

We have a combination of continuous and categorical columns. We scaled numeric features using StandardScaler and encoded for categorical features using OneHotEncoder using sklearn library.

Any missing numerical values were replaced with the median, and any missing categorical values were replaced with the most frequently present value.

Pseudo code:
num_imputer = SimpleImputer(strategy='median') df_new[num_features] = num_imputer.fit_transform(df_new[num_features]) cat_imputer = SimpleImputer(strategy='most_frequent') df_new[cat_features] = cat_imputer.fit_transform(df_new[cat_features])

For cancer subset we generated two derived features per state ‘average screening rate’ which holds mean of the data_value column per state and ‘racial disparity’ which hold standard deviation of data_value column per state

Pseudo code:
df_canLoc = df_subset.groupby(['Locationdesc','Class'])['Data_value'].mean().reset_index(name='avg_screening') df_race = df_subset[df_subset['Break_Out_Category']=='Race/Ethnicity'] df_race_disp = df_race.groupby(['Locationdesc','Class'])['Data_value'].std().reset_index(name='race_disparity') df_state_profile = pd.merge(df_canLoc, df_race_disp, on=['Locationdesc','Class'])

3.3 Model 1: K-Means Clustering

(i) Full Dataset

We applied K-Means with log-transformed and polynomial-expanded numerical features. We chose ‘Response’ as target variable for our dataset. As per observation the Response column had hundreds of unique values and plotting all of them would be messy, so we kept the 10 most frequent responses and group the rest as "Other" to simplify analysis (Fig.4).

(ii) Cancer Subset

We created this subset to isolate relevant observations, reduce noise from unrelated class types and reflect meaningful patterns in the screening rates. We applied k-Means on this subset using two derived features that partitioned the data into features and similar observations of k clusters. The elbow method suggested k=4 clusters.

3.4 Model 2: PCA + K-Means Clustering

(i) Full Dataset

For the second model, we applied PCA with 95% variance retention, reducing dimensionality from 994 encoded features to 82 components. K-Means was applied to these PCA features

(ii) Cancer Subset

For state-level colorectal screening, PCA reduced the 2 engineered features into principal components revealing two stable clusters. K-Means was run with k = 2, determined by silhouette analysis.

3.5 Links to Jupyter notebooks:

https://github.com/MousumyCSE/CDC_BRFSS_project/blob/Milestone4/ML_project.ipynb
https://github.com/MousumyCSE/CDC_BRFSS_project/blob/Milestone4/MilestoneIV_part2.ipynb

4. Results

4.1 Method 1: K-Means Clustering

(i) Full Dataset

To evaluate cluster quality, we calculated the silhouette score on a random sample of 10,000 rows due to the dataset's size. The sampled silhouette score is 0.384, indicated moderate cluster separation. Larger clusters capture most of the responses and smaller clusters represent specialized or less frequent patterns.

In terms of underfitting vs overfitting, underfitting would result in few clusters, high inertia, low silhouette score. Overfitting would result in too many clusters, very low inertia, but clusters may be meaningless (splitting natural groups unnecessarily). k=5 is a balance, representing main population segments without over-segmenting.

(ii) Cancer Subset

The results for cancer subset Fig.5 indicated that partial separation of cluster as it was difficult to interpret at the state level. Some clusters grouped individual racial groups rather than showing overall state-level screening behavior. Outliers skewed centroid.

4.2 Model 2: PCA + K-Means Clustering

(i) Full Dataset

After applying PCA with 95% variance retention, the dataset was reduced from 994 original features to 82 principal components. Running K-Means on these PCA-compressed features produced a silhouette score of 0.443 (sampled), which is noticeably higher than the score from the original K-Means model without PCA. This indicates that PCA helped remove noise and redundant information, allowing K-Means to form more compact and well-separated clusters.

(ii) Cancer Subset

We applied PCA Fig.6 to the state-level colorectal cancer screening data to transform the original features (avg_screening and race_disparity) into principal components that capture the directions of maximum variance. We got a silhouette score of 0.799 which shows strong cluster separation. We also observed cluster 1 showed high screening and moderate to low racial disparity while cluster 2 showed outlier with unusual pattern in screening or disparity. Also, the outlier state with the lowest PCA1(lowest overall screening) was 'Virgin Islands'

5. Discussion

5.1 Model 1: K-Means Clustering

K-Means on the full dataset captured broad behavioral patterns but struggled with high dimensionality and noise. Cancer subset clusters lacked interpretability due to skewed and sparse state-level observations.

5.2 Model 2: PCA + K-Means Clustering

PCA improved cluster compactness and removed correlated noise. The PCA + K-Means approach produced clearer separation, especially in the cancer subset where it highlighted meaningful state differences and clear outliers. However, the negative test silhouette score indicates overfitting in the 2-component PCA version—PCA compressed too aggressively, causing cluster instability.

5.3 Shortcomings:

- BRFSS responses are highly unbalanced and uneven across demographics.
- PCA with 2 components oversimplified some structures.
- Using only screening rate and disparity left out important predictors (income, education, access).

6. Conclusions

This project demonstrated that unsupervised learning can uncover meaningful structure in BRFSS survey data especially when combined with dimensionality reduction. PCA highly improved clustering as we could see strong cluster separation. Outliers displayed unique screening patterns. Model 2 is superior to Model 1 for interpretability and pattern recognition.

6.1 Challenges

The dataset was extremely large, making full-scale clustering computationally impractical. To address this, we sampled the data and focused on a targeted subset defined by specific classes and topics to obtain more meaningful and interpretable clustering results.

6.2 Future Work

- Adding socioeconomic variables to enrich state-level profiling could improve pattern recognition
- Trying Gaussian Mixture Models to better handle overlapping or non-spherical clusters.
- Use PCA with more components (3–5 PCs) to reduce overfitting and to capture additional structure
- Employing semi-supervised prediction of cluster membership

7. Statement of Collaboration

Name	Title	Contribution
Aditi Das	Coder/Writer	Contributed in visualization, data modeling and documentation
Mousumy Kundu	Coder/Writer	Contributed in visualization, data modeling and documentation
Hayley Baek	Coder/Writer	Contributed in visualization, data modeling and documentation