BRFSS - Behavioral Risk Factors Surveillance System

1. Introduction

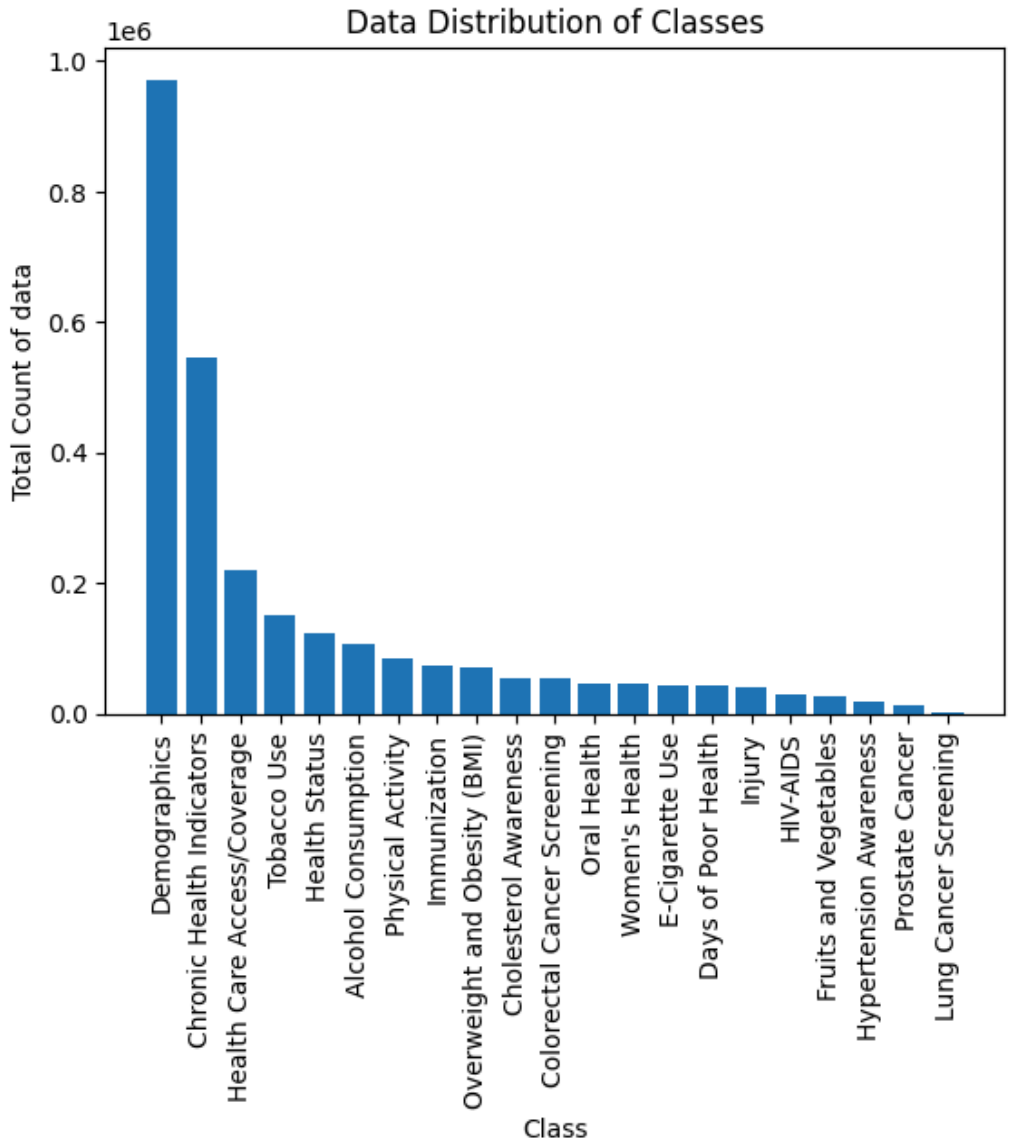
BRFSS is one of the largest ongoing health surveys in the world, designed to track modifiable risk factors for chronic diseases and other major causes of death. This project uses survey data from the CDC's Behavioral Risk Factor Surveillance System (BRFSS), collected annually from 1984 through 2023. Its scope of survey responses across the entire United States for nearly forty years serves as a valuable tool for understanding long-term health patterns across different demographic measures like age, income, and race/ethnicity. Survey question topics include depression, substances like alcohol and nicotine, and broad conditions like arthritis.

We chose this dataset because of its potential to provide valuable insights for what kinds of measurable risk factors contribute to long-term chronic diseases and increased mortality rates. In this way, we believe that the scientific and social impact of understanding public health was represented well in this dataset. Instead of having individual survey responses, this dataset maintains much of its valuable information by aggregating demographic measures; by doing this, entire populations are still represented while losing minimal survey responses which made analysis more computationally lightweight.

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 areas like race, age, or socio-economic groups. Survey data like the BRFSS is especially valuable for machine learning because it gives us clues about why certain behaviors or conditions show up in contributing to chronic diseases. A good predictive model can provide insights to predict similar patterns for future respondents and better resource planning for targeted populations. The overall goal of this project therefore is to visualize relationships between features in the dataset and identify meaningful patterns.

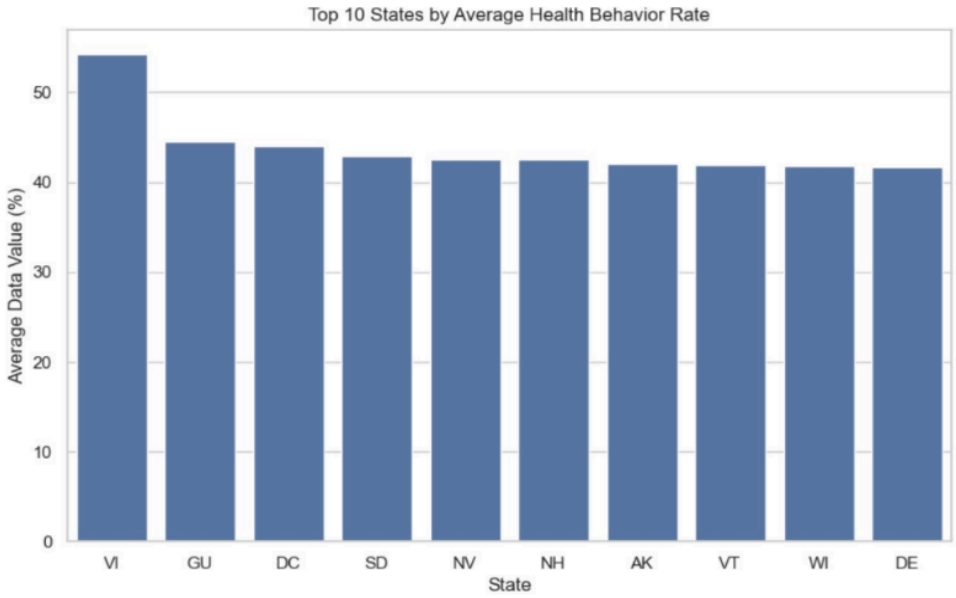
2. Figures

Figure 1: Class Distribution of Survey Responses



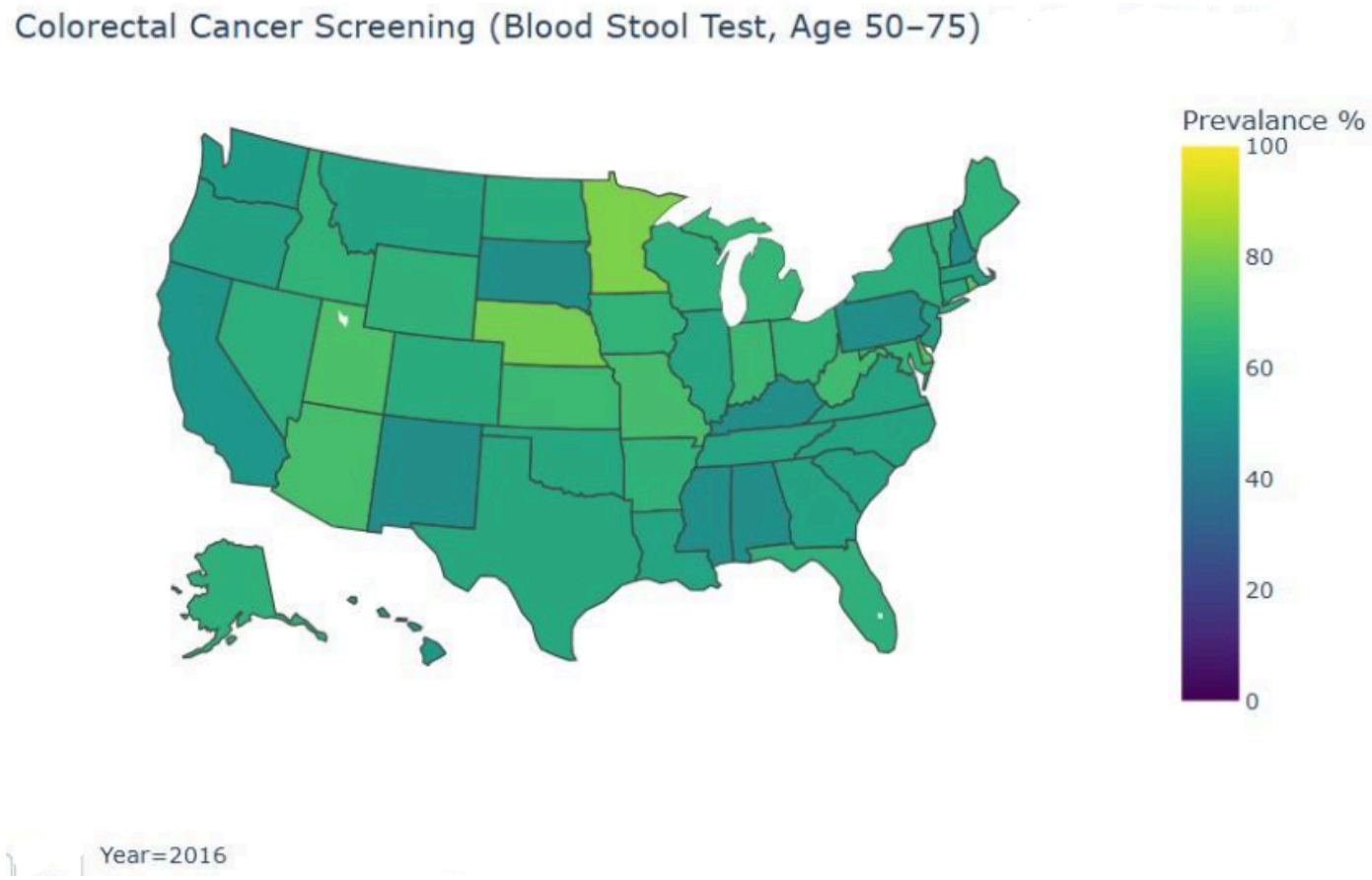
This bar chart shows the distribution of survey responses across different health classes. The demographic category being a type of response class that includes topics like disability status, education, and employment has the largest number of responses. This is in contrast to a class like Lung Cancer Screening which has significantly fewer responses. This is because every survey respondent is required to answer all of the demographic questions which have many topics, but response classes like Lung Cancer Screening have less varied responses. This gives us insight as to broadly which health classes are more epidemic compared to others across the entire United States.

Figure 2: Top 10 States by Average Health Behavior



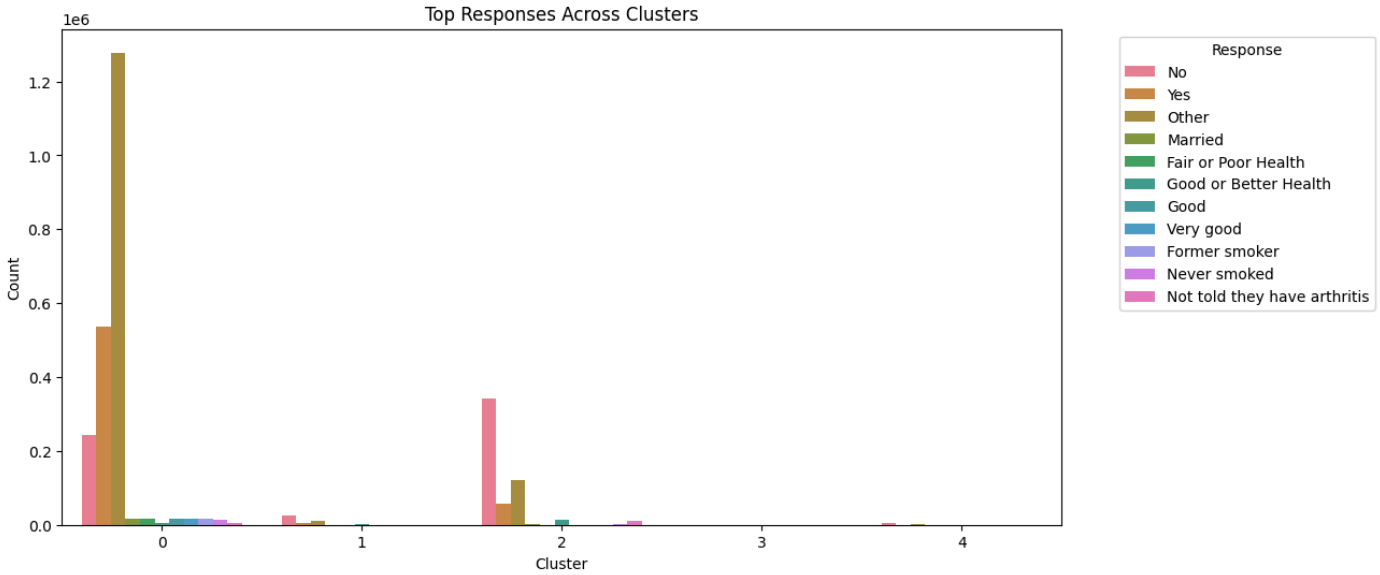
This bar chart shows the means of demographic variability. Higher values represent a lower variability of intersectional demographics. Data Value measures what percent of survey respondents represent a single demographic group compared to other demographic groups of the same responses within the same 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 states by demographic homogeneity.

Figure 3: Map of Colorectal Cancer Screening



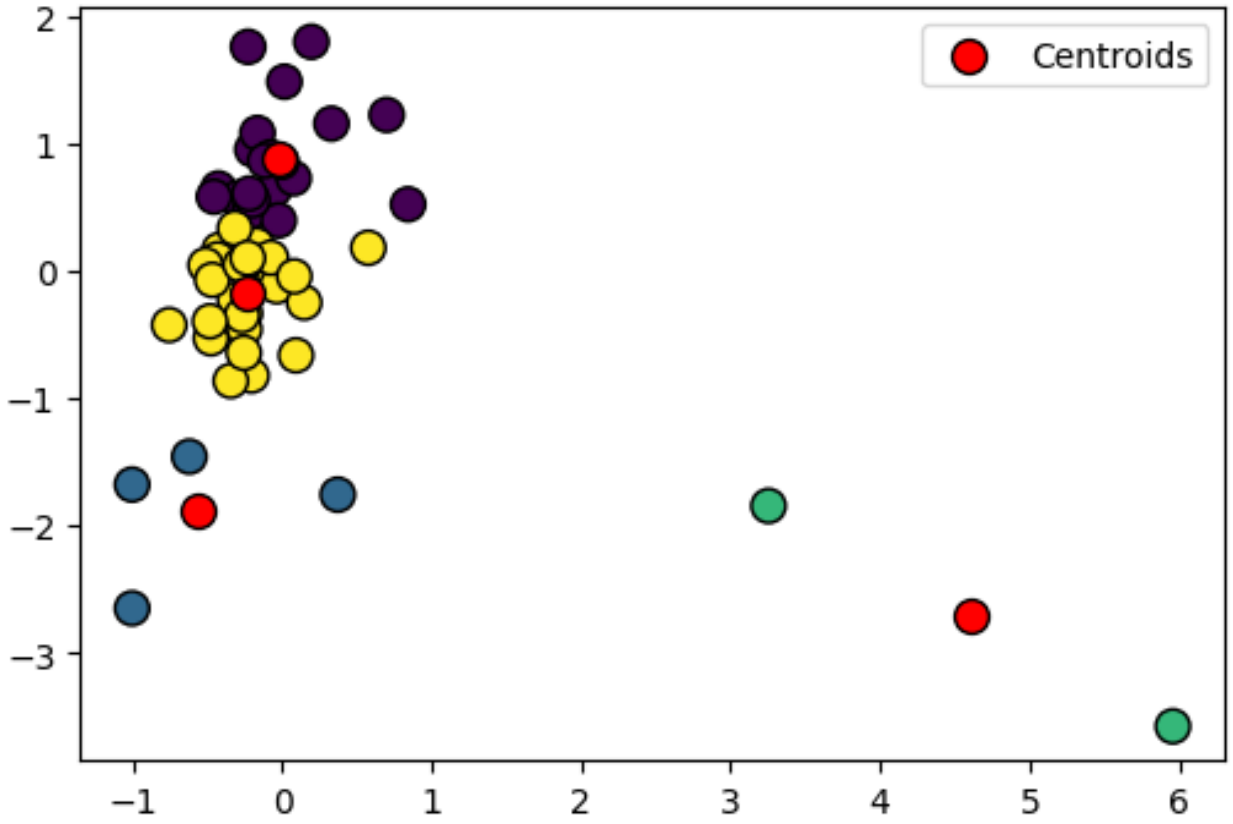
The BRFSS data carries geographic information which allows us to create choropleth maps. This map is an example of the type of demographically specific analysis we can perform for specific diseases like colorectal cancer. By choosing the 50-75 year old age group and focusing on survey responses for 2016, this graph shows that Minnesota has the highest prevalence of blood stool test screening records while New Mexico shows the lowest prevalence.

Figure 4: K-Means Clustering on Survey Responses



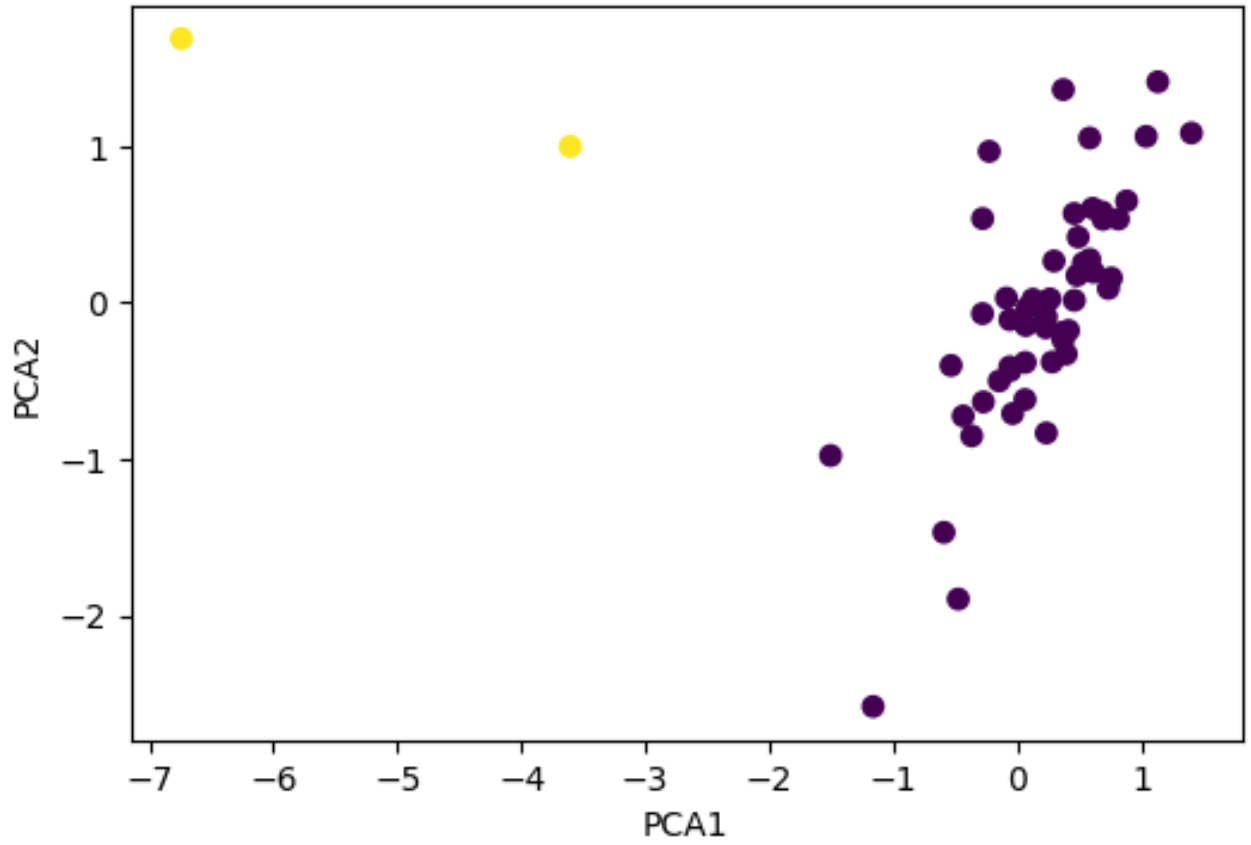
This bar chart shows the distribution of responses across five k-means clusters. Cluster 0 represents the largest segment, containing a majority of "Other", "Yes", and "No" responses, indicating a diverse population with mixed responses. Cluster 2 also captures a large group but has a higher proportion of "No" responses, suggesting a population segment less likely to respond positively. The next most prevalent response after “Other” is “Yes” meaning that for any given survey question about engaging in risky health behaviors like frequent drinking and smoking or having symptoms of health conditions like joint pain or higher than average lethargy, we can cluster demographic groups based on their responses. Clusters 1, 3, and 4 are smaller and contain niche populations or outliers, with "No" dominating but with fewer overall responses. While there are less clear response distributions in these other clusters, this can be largely attributed to the scaling of the graph from the “Other” response in the first cluster. These results indicate that K-Means successfully separates populations by general response patterns, with larger clusters representing the main population and smaller clusters highlighting specialized groups.

Figure 5: K-Means Clustering on Cancer Data Subset



This figure captures another subset of data, specifically the data that pertains to different types of cancer. It illustrates four k-means clusters with their centroids pointed out in red. The elbow method was used to find the ideal number of clusters to be 4. From all of the attributes in our data, there is a clear segmentation between the green and blue clusters from the yellow and purple clusters. Between the yellow and purple clusters, the density of the points in that area of the graph necessitates the data being separated into two clusters because of the potential of variability between more similar segments of data.

Figure 6: PCA + K-Means on Cancer Data Subset



In continuation of the cancer data subset, this figure illustrates PCA and k-means clustering. The silhouette method determined that two clusters is the ideal number of segmentations in the data. The silhouette score for k = 2 is around 0.8 whereas the silhouette score for k = 3 dropped significantly to around 0.43 and never went above 0.5 for higher values of k. The outlier states in the dataset for these clusters is the Virgin Islands. This graph demonstrates that the Virgin Islands are an outlier for survey responses specifically regarding cancer health and that there should be special consideration for cancer care between US citizens in the Virgin Islands than in the rest of America.

3. Methods

3.1 Data Exploration

The BRFSS dataset has a shape of around 2.76 million rows and 27 columns. The most important columns for this dataset include: Year and State for location and time information, Class, Topic, Question, and Response for the surveys, and Break_Out and Break_Out_Category for the demographic information. Class groups different question subjects like Cancer and Substance Use, and Topic further specifies the Class category into values like Colorectal Cancer or Alcohol Consumption. Break_Out_Category is for different demographic groups like Age, Gender, Race/Ethnicity, and Income, and Break_Out is the actual values like 18-24, female, White, and under $50,000. We were able to drop some unnecessary columns that didn't directly pertain to our data analysis like LocationID or GeoLocation, the latter containing latitude and longitude coordinates for the states that were already represented in the State column.

3.2 Data Preprocessing

The data has an approximate equal combination of continuous and categorical columns. Python's Sklearn library was used to scale numeric features using StandardScaler and to encode categorical features using OneHotEncoder. Missing numerical values were replaced with the median, and missing categorical values were replaced with the most frequently present value. Two columns Data_Value_Footnote and Data_Value_Footnote_Symbol were dropped because they had more than 80% null values.

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 the cancer data subset, two derived features Average Screening Rate calculates the mean of the demographic variability per state and Racial Disparity calculates the standard deviation of the demographic variability per state. The demographic variability appears in the data as what percentage of respondents belong to what demographic category, so aggregating these demographic differences is important for finding how different groups of people between states respond to different cancer questions.

Pseudo code:

```
df_canLoc = cancer subset dataframe
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 = race disparity dataframe
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

The first model applies K-means clustering with log-transformed and polynomial-expanded numerical features. The column 'Response' was chosen as the target variable for this model to evaluate the differences in responses between survey takers. The different clusters of responses had hundreds of unique values and plotting all of them was impractical, so the 10 most frequent responses were chosen and the rest of the other values for Response was grouped as "Other" to simplify analysis as shown in figure 4.

(ii) Cancer Subset

The subset of data specifically pertaining to cancer questions sought out to isolate relevant observations, reduce noise from other class types, and reveal patterns more specifically in the screening rates for cancer health. K-means clustering was applied on this cancer subset, and the data was further partitioned into two derived features. The elbow method was used both in the full dataset and in the cancer data subset to evaluate the ideal number of clusters for k-means clustering.

3.4 Model 2: PCA + K-Means Clustering

(i) Full Dataset

The second model builds upon a PCA (Principal Component Analysis) + K-Means pipeline. PCA was applied because the BRFSS dataset is high-dimensional, noisy, and contains highly correlated variables. By transforming data into uncorrelated components that capture maximum variance, PCA reduces dimensionality, removes noise, and preserves important patterns. This will make clustering more effective, as algorithms like K-Means perform better in lower-dimensional spaces. PCA was applied with a 95% variance retention, reducing the dimensionality from 994 encoded features to 82 components.

(ii) Cancer Subset

For state-level colorectal screening, application of PCA reduced the 2 engineered features into principal components revealing two stable clusters. The "Response" column was kept as the target variable for the full dataset and for the cancer data subset to keep consistency in evaluating the performance of a different predictive model.

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, the silhouette score was calculated on a random sample of 10,000 rows due to performance issues if the score was calculated on the full over two million rows of data. The sampled silhouette score is 0.384, indicating moderate cluster separation. Larger clusters capture the more frequent responses and smaller clusters represent specialized or less frequent patterns. Within the context of the BRFSS dataset, these clusters represent how different populations respond to the same set of behavioral risk questions. The respondents that answered questions similarly together carried the same behavioral risks, so having smaller clusters illustrated that more specific risk behaviors were exhibited in chronic illnesses. Cluster 0 had the most Yes responses out of any other cluster, and they also showed the highest responses for long-term smoking and self-reported fair or poor health. Cluster 2 had more No than Yes responses, indicated never having smoked, and self-reported good or better health.

(ii) Cancer Subset

The cancer data subset as seen in figure 5 indicates a partial separation of clusters that was difficult to interpret on a strictly geographic level. Some of the clusters grouped individual racial groups rather than showing overall state-level screening behavior. Outliers skewed the centroids.

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

PCA was additionally applied as seen in figure 6 to the state-level colorectal cancer screening data to transform the original features (Average Screening and Race Disparity) into principal components that capture the directions of maximum variance. This analysis yielded a silhouette score of 0.799 which shows strong cluster separation. 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. In terms of underfitting vs overfitting, underfitting would result in few clusters, high inertia, and a low silhouette score. Overfitting would result in too many clusters and very low inertia, but clusters may be meaningless (splitting natural groups unnecessarily). Specifying k-means for 5 clusters was found to be a good balance, representing main population segments without over-segmenting. The cancer subset clusters lacked interpretability due to skewed and sparse state-level observations.

5.2 Model 2: PCA + K-Means Clustering

Applying PCA improved cluster compactness and removed correlated noise. The PCA + k-means approach produced clearer separation, especially in the cancer subset where the algorithm 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:

In general, the BRFSS responses were shown to be highly unbalanced and unevenly distributed across the different demographic groups. For example, there were more “Yes” responses than any other response, indicating that the survey respondents were largely engaged in some risk behaviors than the survey respondents that weren’t. PCA application with two components also oversimplified some structures for certain subsets of data like with the cancer data, so there was some cluster instability. Lastly, using only screening rates and racial disparities in the cancer data subset inadvertently filtered out other considerations like income and education.

6. Conclusions

This analysis demonstrated that unsupervised learning can uncover meaningful structure in BRFSS data especially when combined with dimensionality reduction. PCA highly improved the k-means clustering algorithm to allow for more strong cluster separation and better interpretability than without PCA application. Outliers displayed unique screening patterns.

6.1 Challenges

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

6.2 Future Work

For potential future work, the addition of further socioeconomic variables could enrich the state-level profile to improve pattern recognition. Additionally, trying Gaussian mixture models could better handle overlapping or non-spherical clusters. Using PCA with more components, potentially between three and five components as opposed to simply two, could reduce overfitting and capture any additional structures. Lastly, employing semi-supervised prediction of cluster memberships could be a potential route of analysis.

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