Clustering Algorithms with 2018 BRFSS Data

By Heather Knudson

Behavioral Risk Factor Surveillance System

- Yearly telephone surveys in all 50 states, D.C., & 3 territories conducted by the CDC
- More than 400,000 interviews each year
- Questions on health, behavior, demographics

BRFSS™

Data Cleaning

Data Cleaning

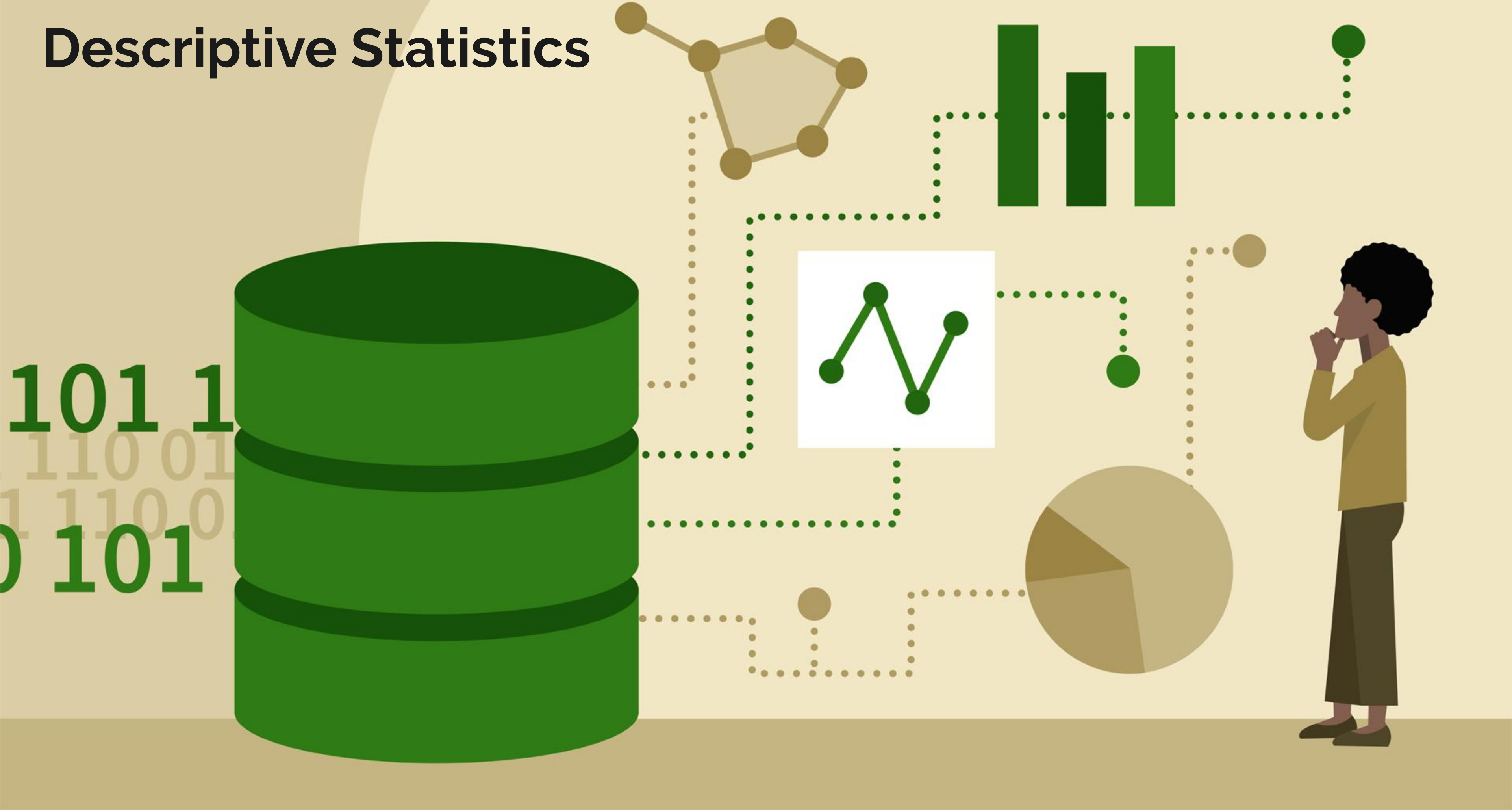
1. Removed columns with > 20% of data missing
2. For columns with < 20% missing, imputed code ‘1000 to stand in for missing values
3. For continuous variables, imputed with the mean
4. Removed columns that were copies of other columns, favoring those with the most info
5. Re-coded all columns so the first code was 0, not 1
6. Re-coded columns that were originally in ‘backwards’ order

Total: 437,436 observations & 68 features

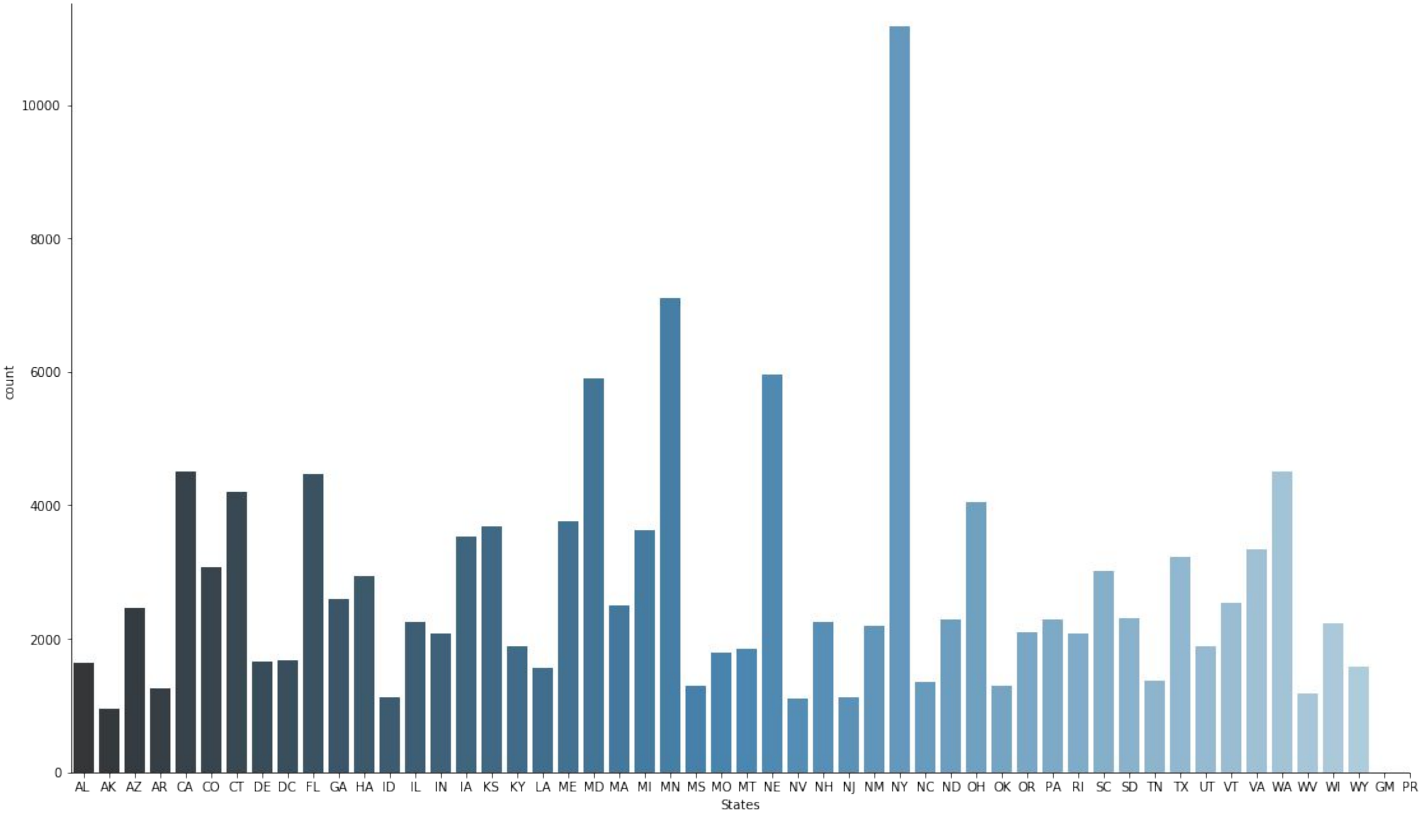
Research Question

How can 2018 BRFSS respondents be segmented based on their demographics, behaviors, and health outcomes?

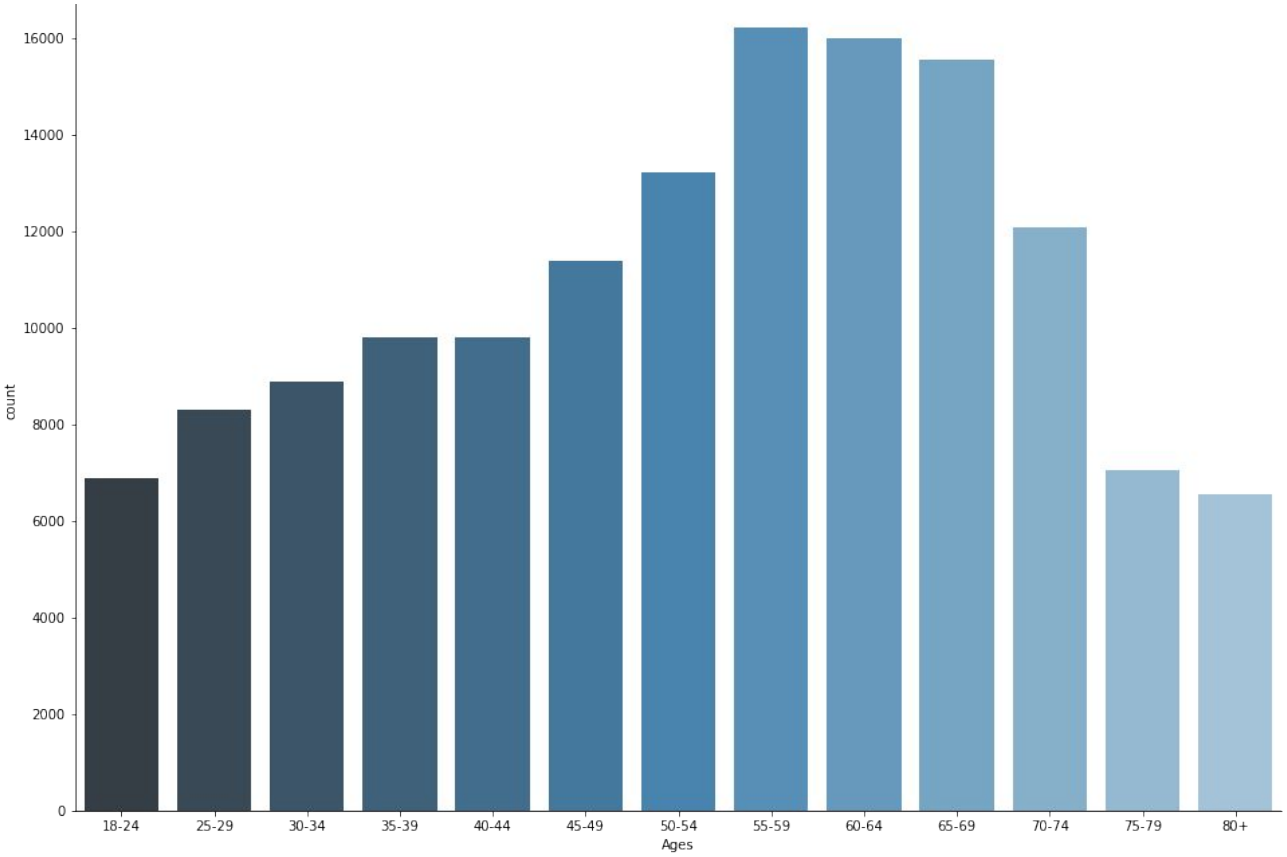
Descriptive Statistics



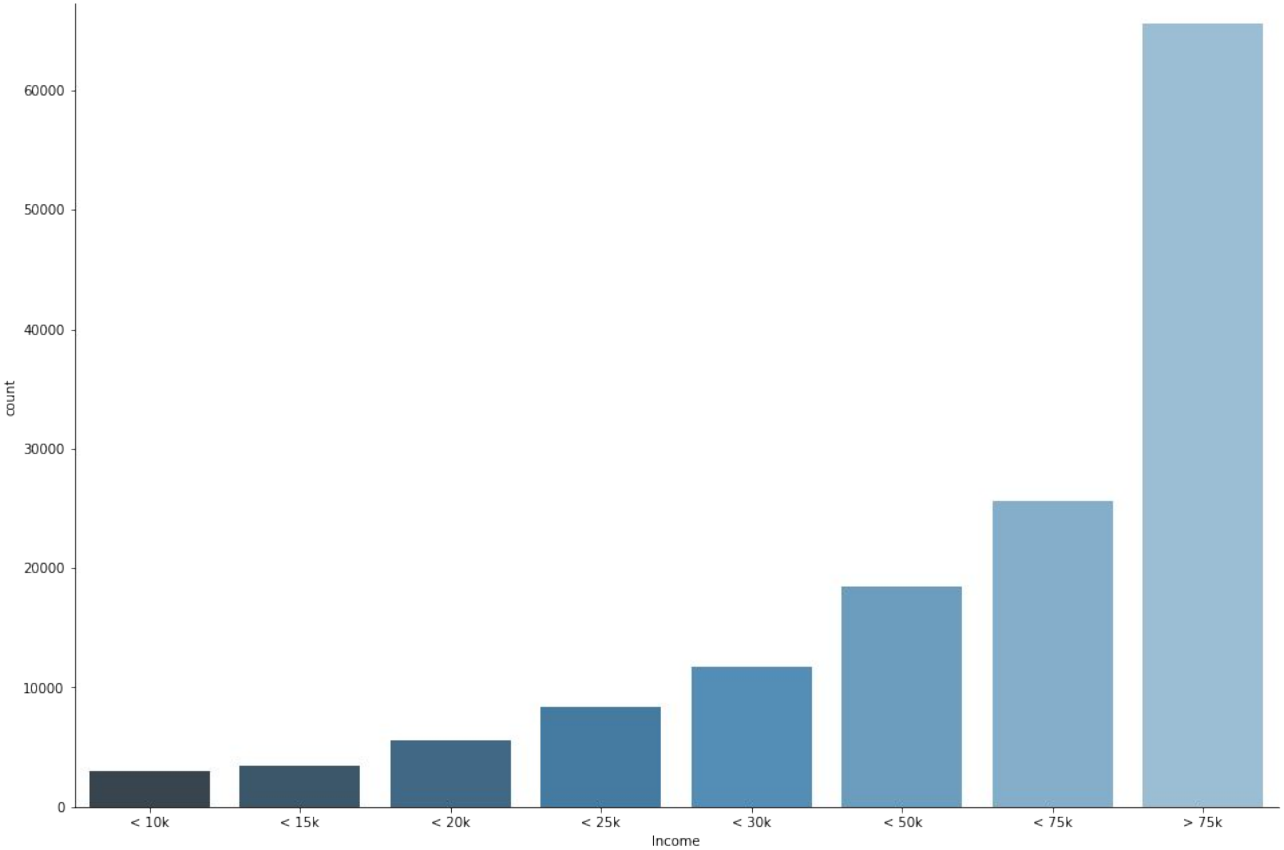
Most respondents lived in NY, followed by MN



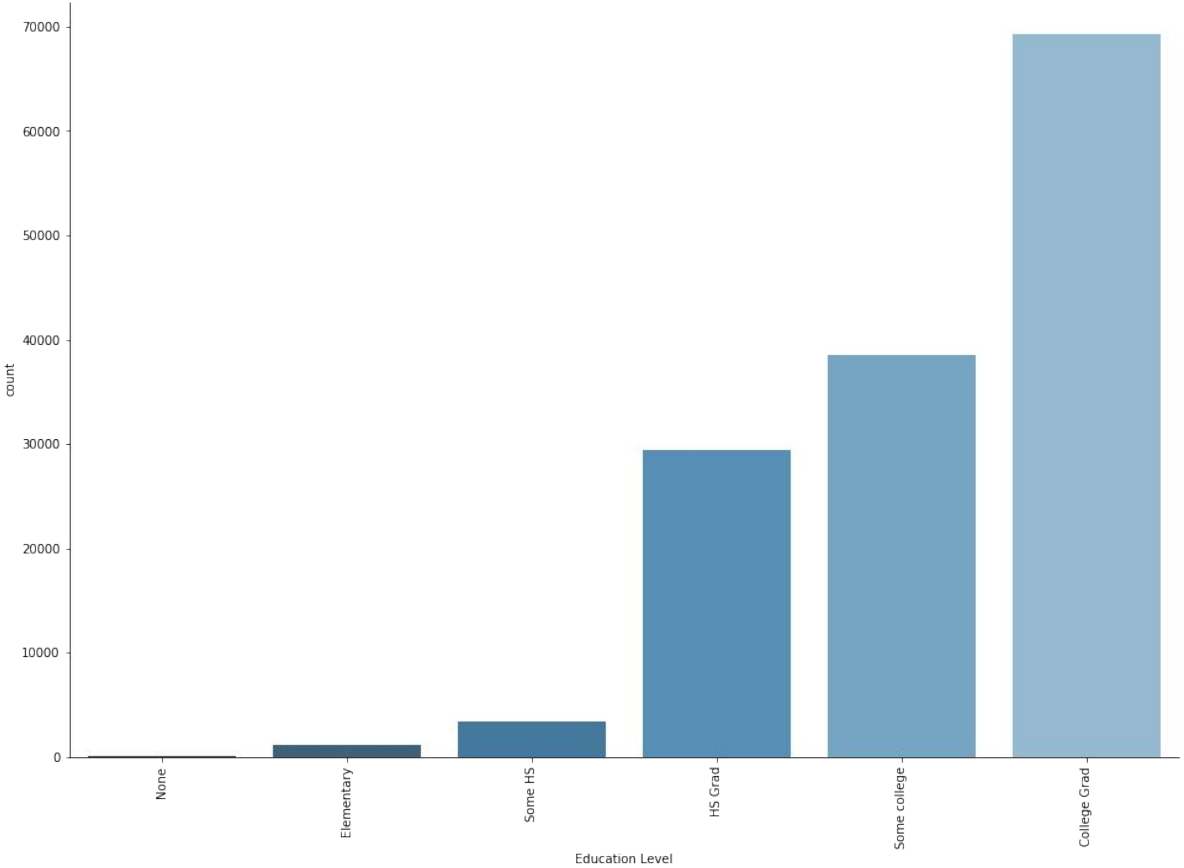
Most respondents were in their 50s & 60s



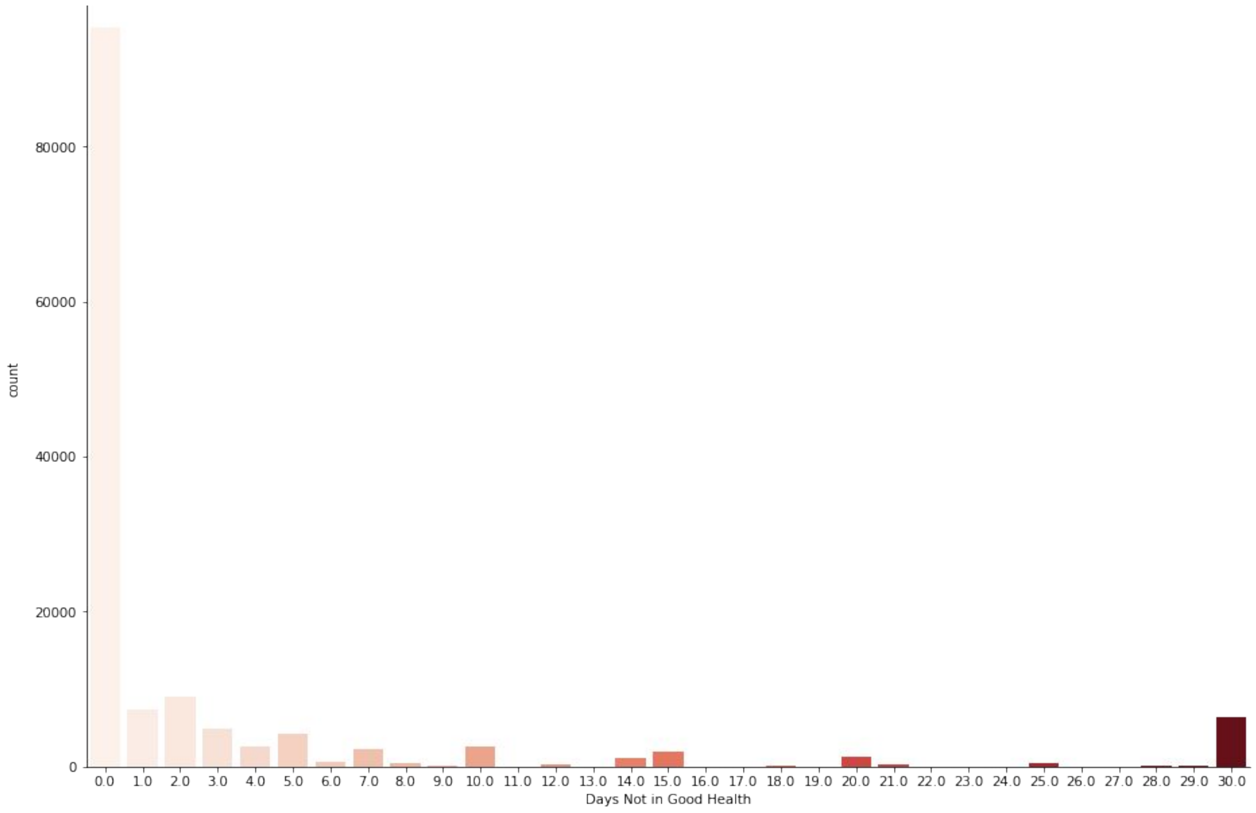
Most respondents made more than $75,000



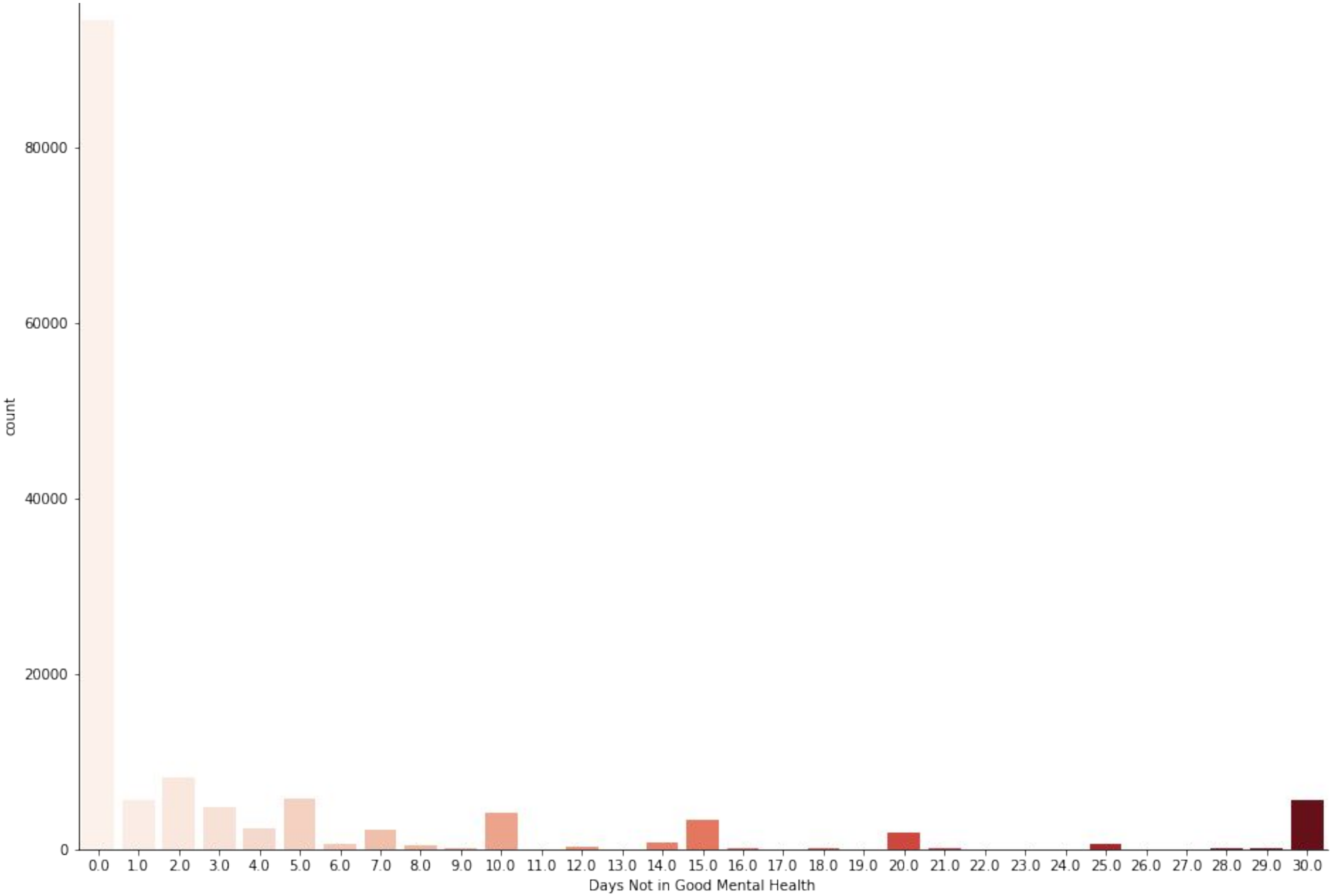
Most respondents had a college degree



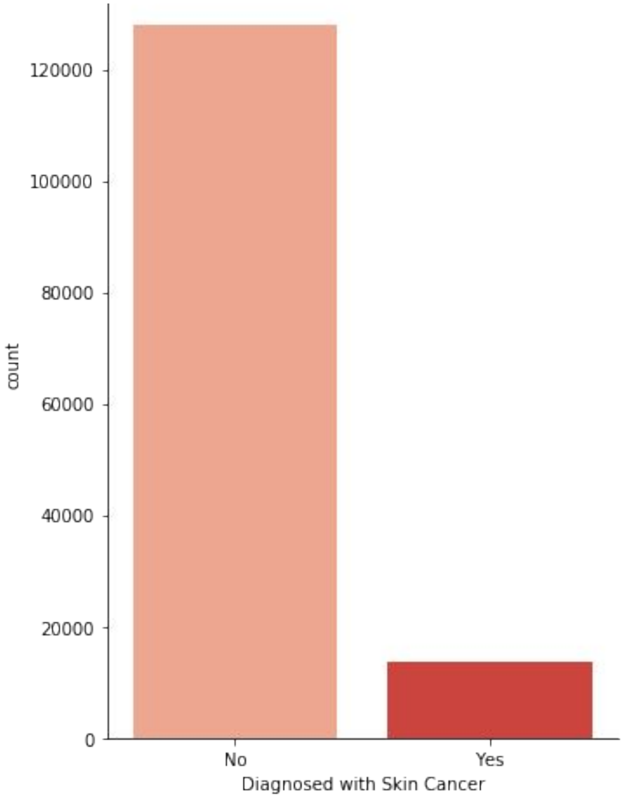
Most respondents were in good physical health

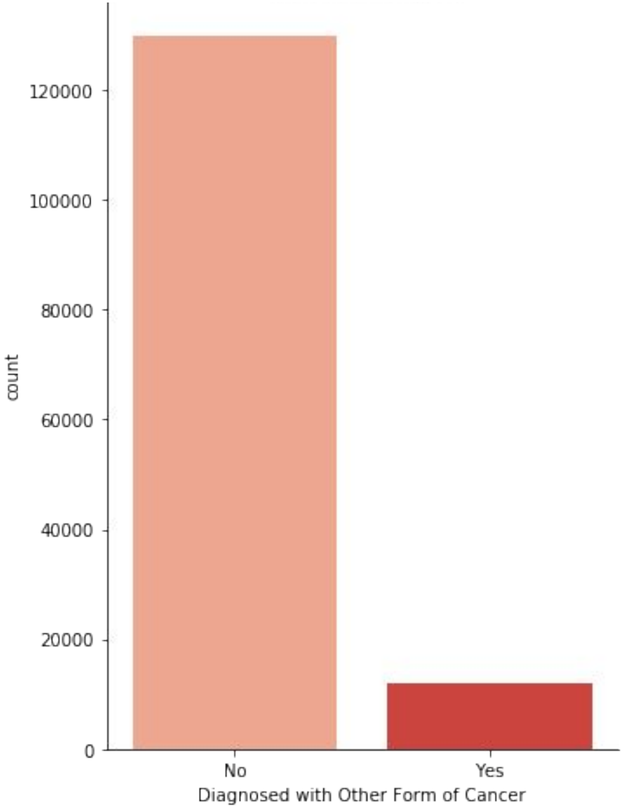


Most respondents were in good mental health

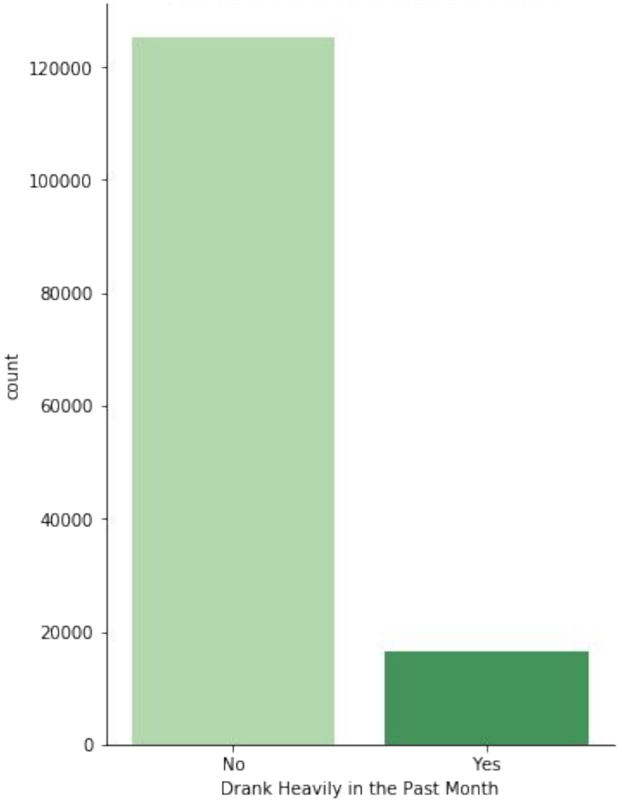


An equal number of respondents were diagnosed with skin cancer as all other forms of cancer

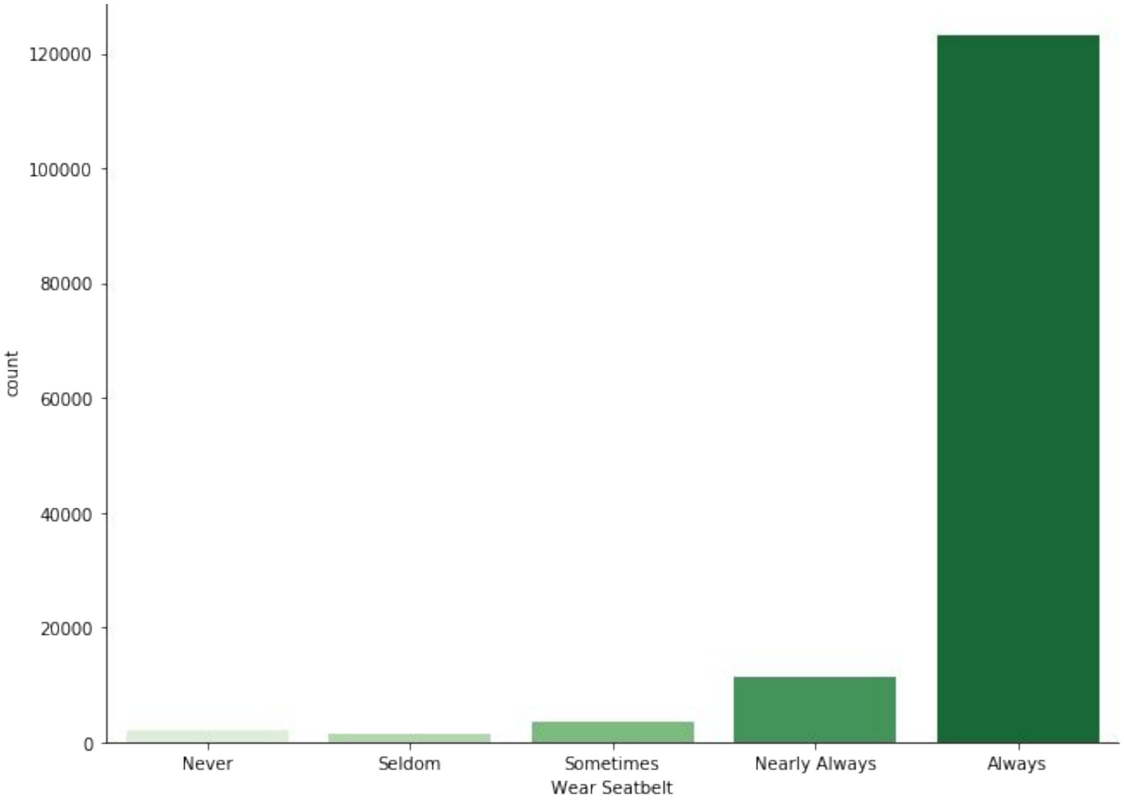




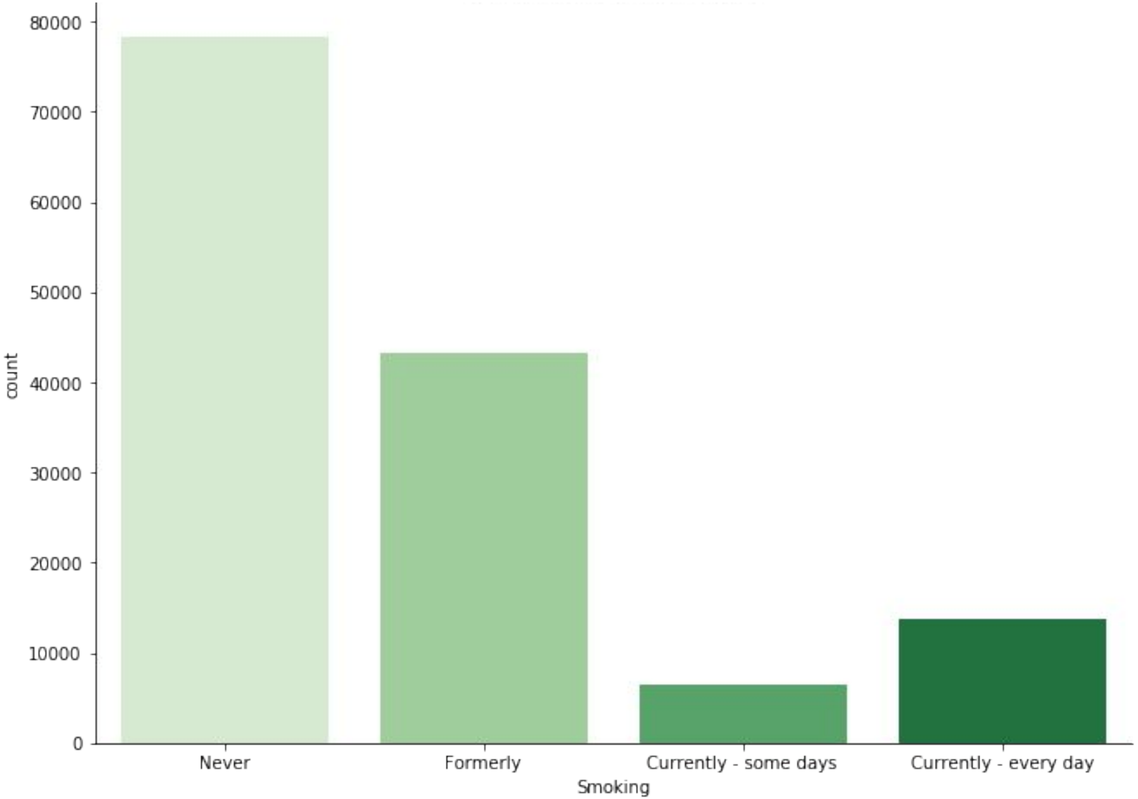
Just 6% of respondents were heavy drinkers



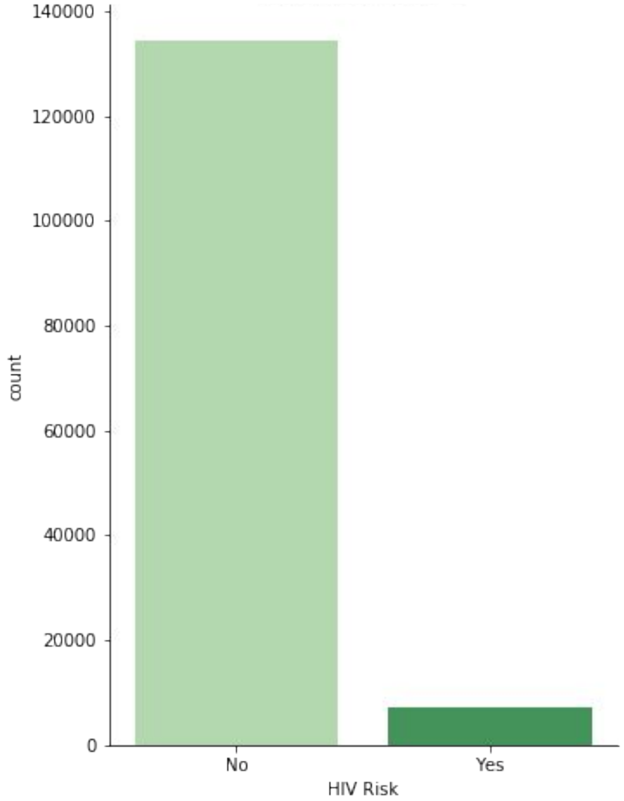
Most respondents always wore their seatbelt



10% of respondents were former smokers, while 4% smoked every day

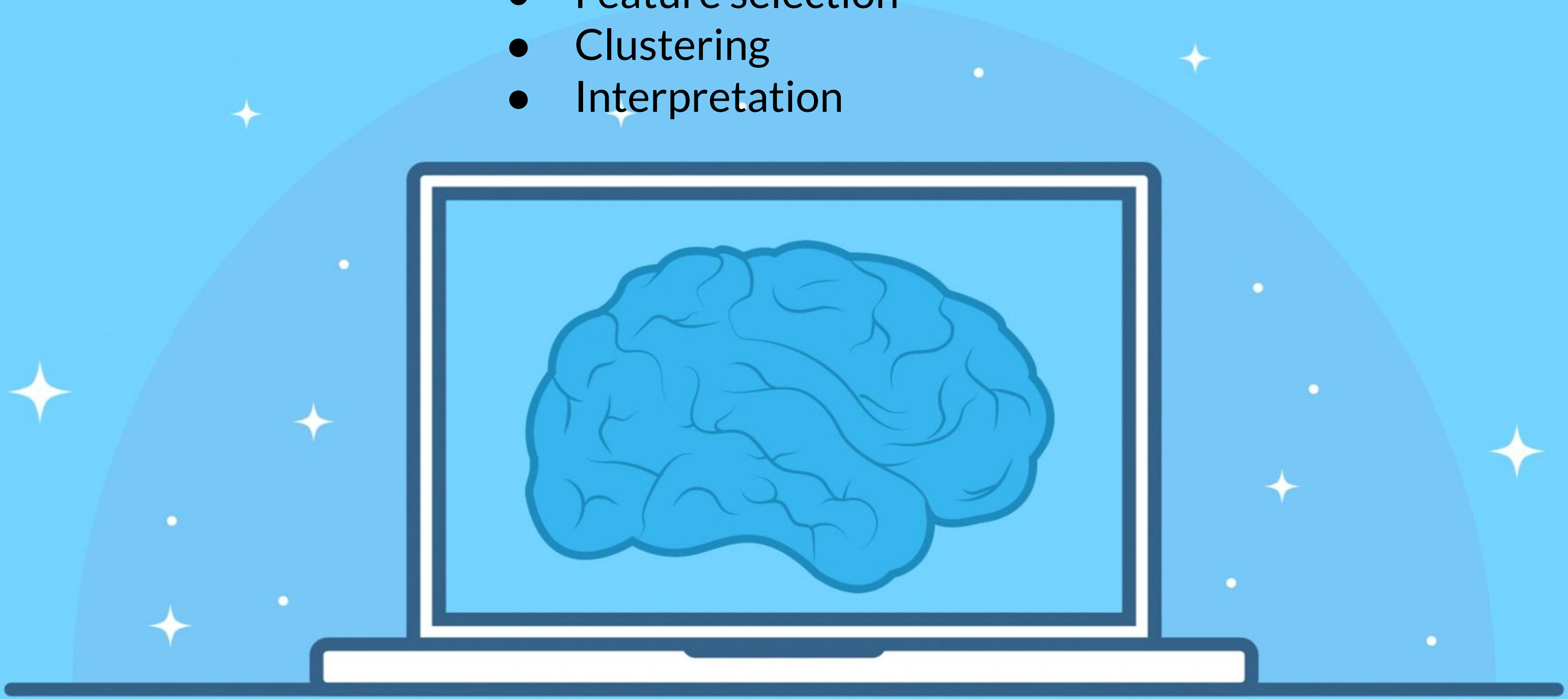


Less than 1% of respondents were considered at risk for HIV

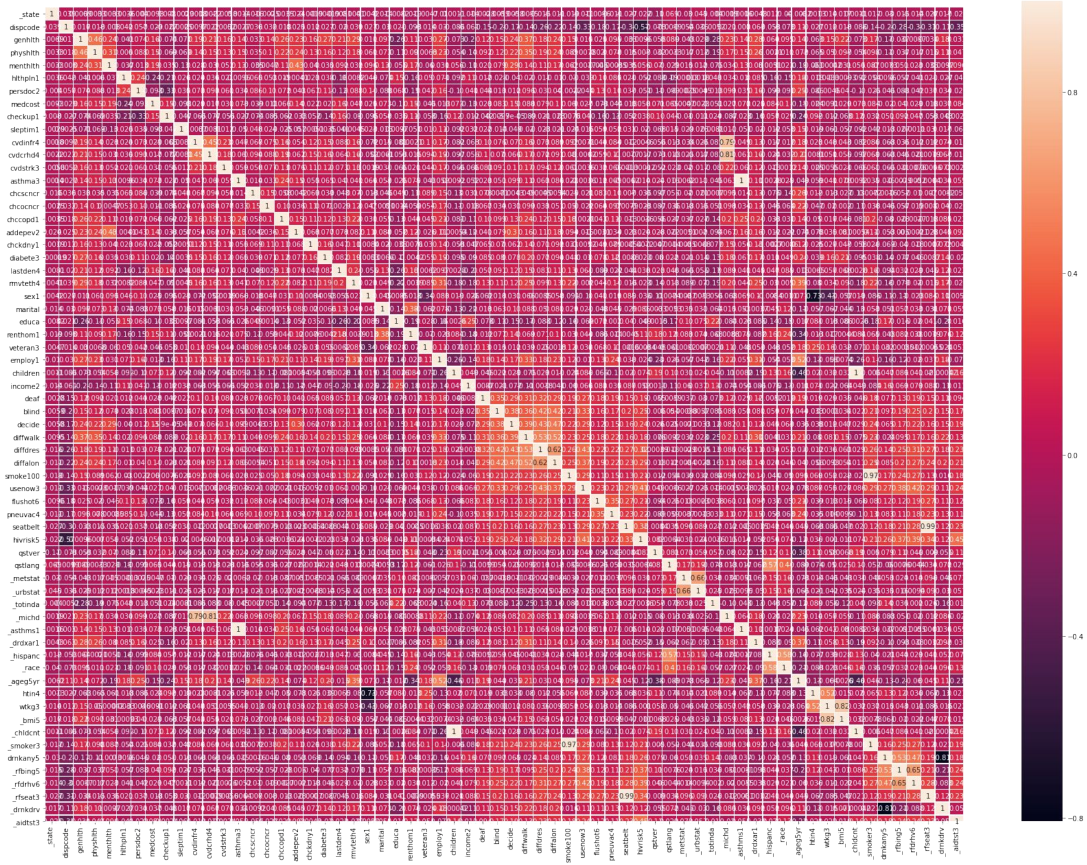


Methods

- Feature selection
- Clustering
- Interpretation



Feature Selection - Multicollinearity Heatmap



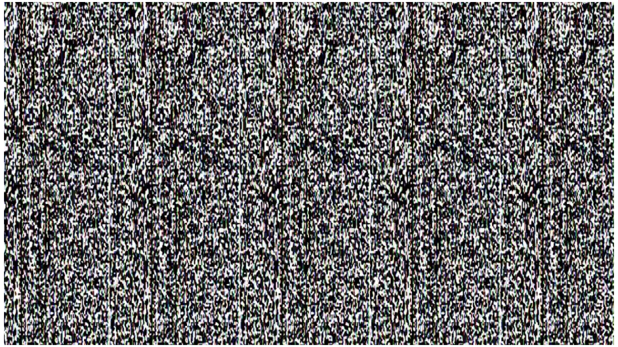
Feature Selection - Variance Inflation Factor

Result: Removed total of 13 variables

Total: 55 features & 437,436 observations

DBSCAN

Result: Too much noise - 99.99%



Best DBSCAN: .18 eps, 300 min_samples

→ 7 clusters

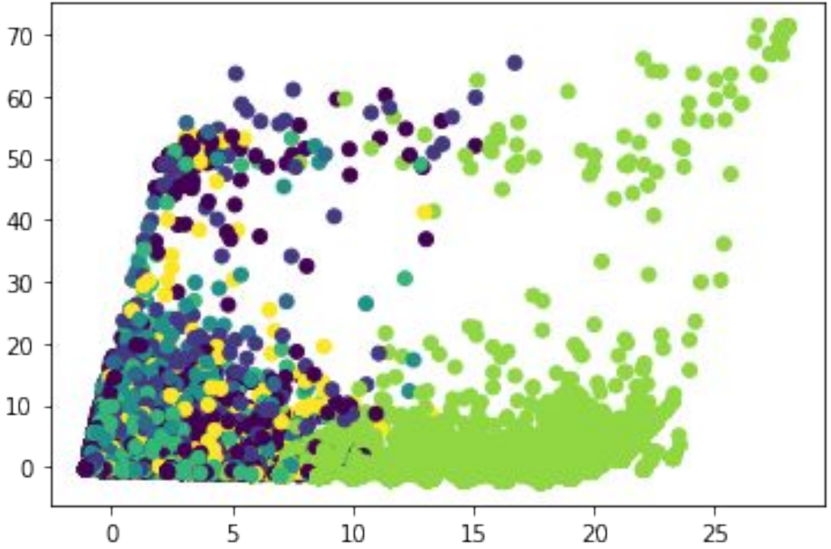
Silhouette Score: -.45

👎

K-Modes

K=7

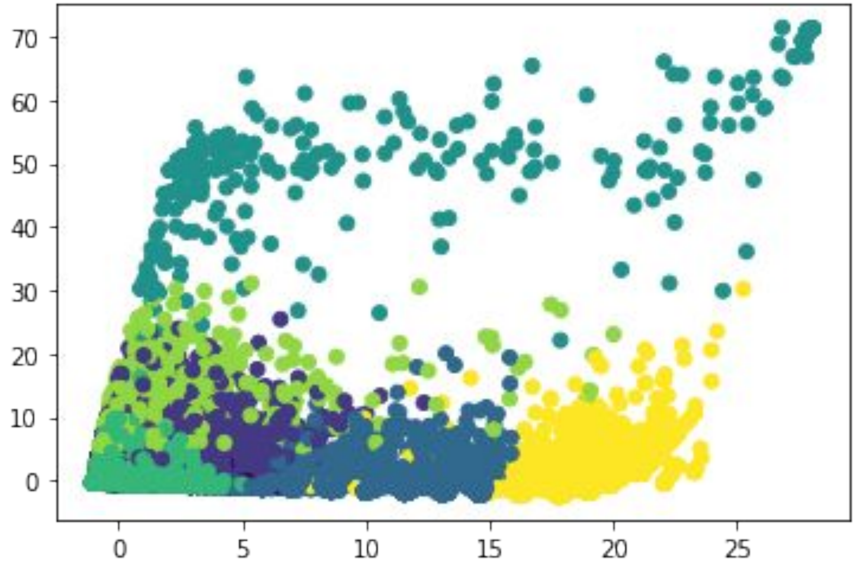
Score: .04



K-Means

K=7

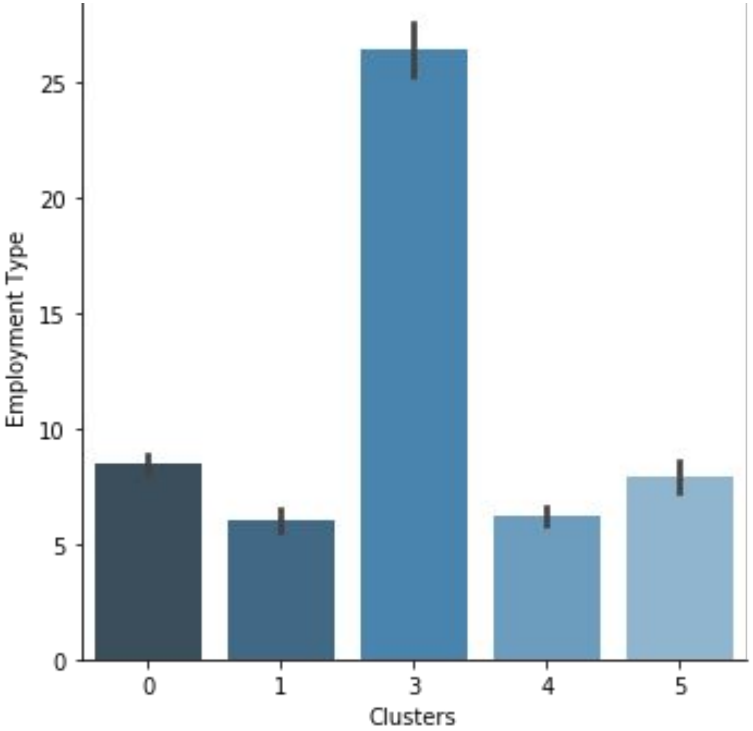
Score: .25



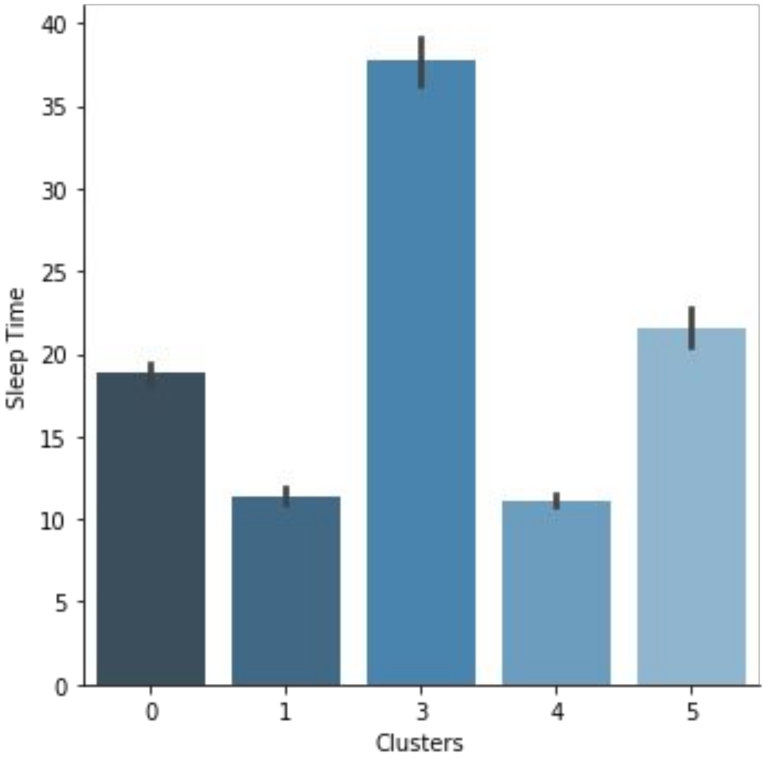
Cluster Breakdown

Cluster Number	Count
0	131,595
4	121,006
1	63,451
3	60,640
5	40,294
2	11,720
6	8,730

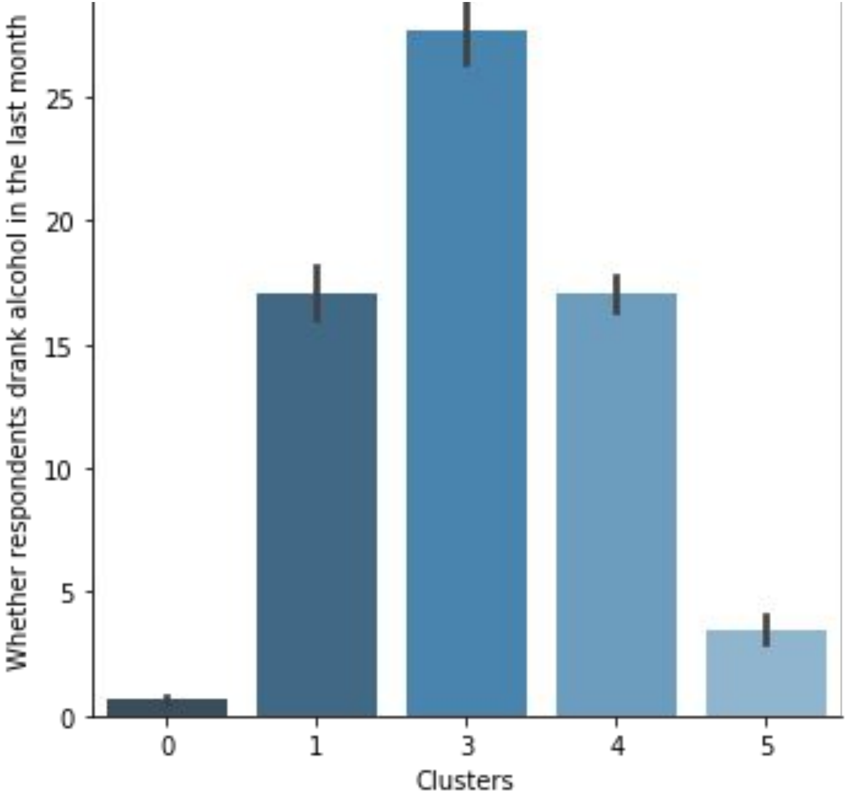
Clusters 3 & 6 were more likely to be retired or unable to work



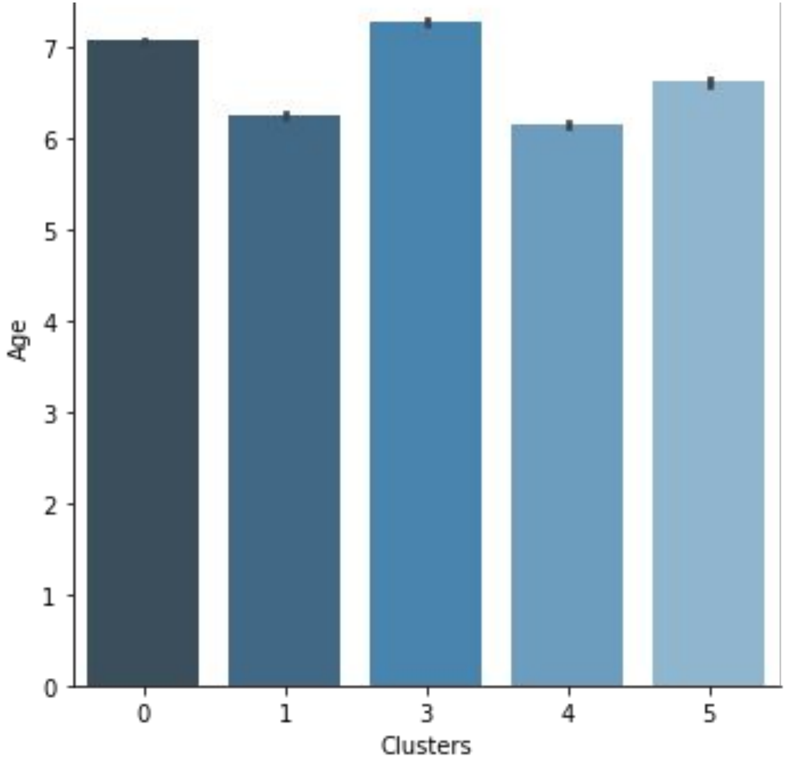
Cluster 3’s average sleep time was highest



Clusters 3, 1, & 4 were more likely to have drunk alcohol recently



Clusters 3 & 0 were older on average



Limitations

- Slowness of running the algorithms
- Could have tried a few difference values for eps, if there was more time
- Categorical data not the best to cluster with

Conclusions & Future Research Possibilities

- Possibility that the dataset was clustered based on life stages of respondents
- Potential other aspects at play too, such as affluence and lifestyle
- Would be nice to learn more about the clusters and plot them alongside more features
- Would have been good idea to try HDBSCAN