# Mental Illness, Aging and Self Reported Health in Baby Boomers

## Identifying Mental Health Trends in Aging Adults

Methodology

My personal background includes a large stint in Medicare insurance sales. To give a little context these insurance products work with Medicare in the United States for individuals over the age of 65 years old (some exceptions).

The baby boomer generation (individuals 55-75 years old) is often under served as our society has become entangled with technology. Research illustrates an increase in diagnosis of depression and other mental illness in aging adults. Additional research on these trends is available via The Family Institute at [Northwestern.](https://counseling.northwestern.edu/blog/boom-in-aging-adults-could-overwhelm-mental-health-care-field/)

Due to COVID-19's effect on mental health around the world, it is especially important to understand what trends effect our elders in an attempt to work toward favorable living conditions for all. The Centers for Disease Control and Prevention released the following [Alzheimer's Disease and Healthy Aging Dataset.](https://catalog.data.gov/dataset/alzheimers-disease-and-healthy-aging-data)

Our analysis will look for commonalities across individuals based on ethnicity, age, gender and geographic region. While no definitive conclusions will be drawn due to limitations of the dataset, K-Means clustering algorithms will help identify highly related features and insight into baby boomers as they age in our society.

## Scrub

During the scrubbing phase a few key decisions were made in order to make the dataset more digestible during the exploratory data analysis. The columns with duplicate values (such as abbreviations for an existing column) and redundancies were removed. Examples of this were columns **Datasource** and **Response** which represented the same values across the dataset; BRFSS (Behavioral Risk Factor Surveillance System) and None respectively.

Next, generic labels like **StratificationCategory** were renamed to more appropriately reflect the data they headed. These simplifications were done in an effort to make the data more digestible as many of the values represented large phrases that made visualization challenging. Special care was taken when reducing the **Topic** values as they represent the key factors of our analysis. Please see our jupyter notebook for additional insight into these changes. Lastly, our analysis focused on individuals who reside on the West Coast. The states that were included were Arizona, California, Colorado, Nevada, Oregon and Washington.

## Exploratory Data Analysis

Three key columns represent the stratifications of the data set. They are ‘Ethnicity\_Gender’, ‘Age\_Group’ and ‘GenderEthnLabel’. With so many factors, the analysis looked to find the most significant variation amongst divided categories (age, gender, ethnicity) before narrowing down on one category and exploring the subdivisions. For reference, each category is divided as follows:

Ethnicity: ‘White, non-Hispanic’, ’Asian/Pacific Islander’, ‘Native Am/Alaskan Native’, ‘Black, non-Hispanic’, ‘Hispanic’, ‘Overall’

Gender: ‘Male’, ‘Female’, ‘Overall’

Age: ’50-64’, ‘65+’, ‘Overall’

### Overall (By State)

# Overall Age, Ethnicity, Gender

overall = data.loc[(data['Ethnicity\_Gender'] == 'Overall') &

(data['Age\_Group'] == 'Overall') &

(data['Value'] != 0)] # removing null values

A key obstacle in the analysis was the fact that health questionnaires did not always address the same topics (per year) so the distribution of values was in question. To make up for this, the ‘overall’ analysis was conducted on questionnaires that occurred in the same year, on topics that occurred across each questionnaire.

INSERT VISUAL HERE

With respect to obesity, binge drinking (in the last month) and self-reporting of mental distress; individuals in Oregon reported the highest markers over the three categories (2018 average). Interestingly enough, Colorado reported the lowest percentage of obesity in adults and mental distress while reporting the highest percentage of adult binge drinking (past month).

With respect to sufficient sleep, lifetime diagnosis of depression and lack of leisure time (in the past month); Colorado questionnaires reported the highest sufficient sleep, lowest lifetime depression diagnosis and the lowest ‘No Leisure’ scores (translated as most individuals having sufficient leisure).

Seems like 2018 was a great year for baby boomers in Colorado, when compared to other West Coast states.

### Ethnicity

In the following exploratory analysis, all states and all ages were represented with the only stratification being ethnicity.

ethn\_df = data.loc[(data['Age\_Group'] == 'Overall') &

(data['GenderEthnLabel'] == 'Race/Ethnicity') &

(data['Ethnicity\_Gender'].isin(ethnicity)) &

(data['Value'] != 0)]

INSERT VISUAL HERE

Unfortunately, the last row of features includes several ethnicities that were not represented across each of the topics and as such, our analysis will continue by considering age and gender. It is interesting to note commonalities amongst ethnic groups who report certain qualities but ultimately no significant trends are apparent and additional factors like sample size and financial status may heavily influence these features.

## Age

The data frame was subdivided into age categories ‘50-64’ and ‘65+’

agedf = data.loc[(data['Age\_Group'] != 'Overall') &

(data['GenderEthnLabel'] == 'Overall') &

(data['Ethnicity\_Gender'] == 'Overall') &

(data['Value'] != 0)]

INSERT VISUAL HERE

Interesting trends begin to emerge with baby boomers in the 50-64 age group representing substantially higher percentages of smoking, need for assistance (related to cognitive decline), life time depression diagnosis, mental distress, binge drinking (in the last month), being a caregiver of another (in the last month), expecting to care for another (in the next two years) as well as effects to the performance of daily activities when compared to the reports of those over the age of 65 years old.

While this data is limited in scope; research has attempted to quantify the stress of getting closer to retirement with respect to readiness due to increases in the cost of living, medical expenses, dependence on technology and long-term care. Many of the significant variations may be influenced by several outside factors but feature commonalities help us understand some of the more pressing stressors apparent in aging baby boomers.

## Gender

The data frame was subdivided into gender categories ‘Male’ and ‘Female’

genderdf = data.loc[(data['Age\_Group'].isin(Overall)) &

(data['GenderEthnLabel'] == 'Gender') &

(data['Ethnicity\_Gender'] != 'Overall') &

(data['Value'] != 0)]

INSERT VISUAL HERE

Gender observations resulting from the health questionnaires reported Female

Health questionnaires report larger female values with respect to binge drinking, lifetime depression diagnosis, high blood pressure and issues with daily activities and a need for assistance as a result of self-reported decline in cognition. These factors are respective of all ages and ethnicities (in the data).

Our clustering analysis will continue using age as it showed the most variation amongst values per feature.

df = data2.loc[(data2['Age\_Group'] != 'Overall') &

(data2['GenderEthnLabel'] == 'Overall') &

(data2['Ethnicity\_Gender'] == 'Overall') &

(data2['Value'] != 0)]

df\_age = df[['Topic','Value']]

df\_age = df\_age.sort\_values('Topic').groupby('Topic')['Value'].apply(lambda df\_age: df\_age.reset\_index(drop=True)).unstack().T

Due to the frequency of the questionnaires with several topics being asked periodically versus a handful of topics being asked consistently – an average value fill (per column with respect to that columns mean) will be used to address null values initially, but these categories will be dropped as the analysis becomes more granular.

Prior to the fill, any column with more than 50 percent values is dropped.

df\_age.dropna(thresh=24,axis=1,inplace=True) # keeping columns with at least half valid vals

Null values are addressed

# filling values with respective mean

fillcols = df\_age.columns

for col in fillcols:

df\_age[col] = df\_age[col].fillna(value=df\_age[col].mean())

Lastly, the data is prepared using a standard scaler and our first round of modeling begins.

## Modeling

Modeling is completed in three stages. The first clustering analysis represents both age groups, the second clustering represents only individuals age 50-64 and the last clustering represents individuals age 65 and over. Analysis is completed using dimensionality reduction (Principal Component Analysis) followed by K-Means Clustering.

### Both Age Groups – 5 Clusters

In order to retain as much information as possible, the data frame is transformed into three components using PCA. Those components are added to our data frame for analysis.

# Reducing dimensionality | three components

pca = PCA(n\_components=3)

principal\_comp = pca.fit\_transform(df\_scaled) # passing scaled data

# Creating dataframe out of 3 component result

pca\_df = pd.DataFrame(data = principal\_comp, columns=['PCA1','PCA2','PCA3'])

The optimal number of clusters is selected using the elbow method and silhouette coefficient. Five clusters are selected.

# 5 Clusters

km = cluster.KMeans(n\_clusters=5, max\_iter=300, random\_state=101)

pca\_df['Five\_Clusters'] = km.fit\_predict(pca\_df)

INSERT VISUAL HERE

INSERT VISUAL HERE

As we tried to profile and inspect the 5 clusters, we decided the data did not draw meaningful and interpretable trends. This is in addition to any effects that may exist as a result of filling nulls with mean values and using a groupby function with a mean aggregation in order to observe trends. Ultimately, we decided to move forward with a clustering analysis per age group.

## Model | Age 50-64

The data was divided into two groups, halving the number of values per feature from 48 to 24. Null values were not filled but dropped in order to maintain the integrity of the dataset.

INSERT VISUAL HERE

Our multicollinearity plot for this age group reflects a significant positive correlation between lifetime diagnosis of depression and self-reported fair-poor health. The same plot illustrated a significant negative relationship between self-reported disability and self-reported good-excellent health.

With half as much data, PCA was used to reduce dimensionality to two components prior to selecting the optimal number of clusters for analysis.

# Reducing dimensionality | two components

pca = PCA(n\_components=2)

principal\_comp = pca.fit\_transform(df\_scaled) # passing scaled data

# Creating dataframe out of 2 component result

pca\_df = pd.DataFrame(data = principal\_comp, columns=['PCA1','PCA2'])

Three clusters were selected using the elbow method and silhouette coefficient.

INSERT VISUAL HERE

INSERT VISUAL HERE

A groupby function using a mean aggregation is used to profile and inspect the clusters

three = df50\_cluster.drop(['PCA1','PCA2'],axis=1).groupby('Three\_Clusters').mean().head()

three.style.background\_gradient(cmap='Blues')

INSERT VISUAL HERE

Group number ‘2’ is highlighted by the highest (or tied for highest) scores across multiple features: binge drinking, lifetime depression diagnosis, fair-poor health (self-reported), influenza vaccine (in the last year) and no leisure (in the last month).

## Model | Age 65+

Multicollinearity in features for individuals over 65 years old

INSERT VISUAL HERE

Our multicollinearity plot for this age group reflects fewer relationships between features. A strong negative correlation exists between self-reported disability and having received a pneumococcal vaccine (at any point in life).

PCA was used to reduce dimensionality to two components prior to selecting the optimal number of clusters for analysis.

# Reducing dimensionality | two components

pca = PCA(n\_components=2)

principal\_comp = pca.fit\_transform(df\_scaled) # passing scaled data

# Creating dataframe out of 2 component result

pca\_df = pd.DataFrame(data = principal\_comp, columns=['PCA1','PCA2'])

Three clusters were selected using the elbow method and silhouette coefficient.

INSERT VISUAL HERE

INSERT VISUAL HERE

A groupby function using a mean aggregation is used to profile and inspect the clusters

three = df50\_cluster.drop(['PCA1','PCA2'],axis=1).groupby('Three\_Clusters').mean().head()

three.style.background\_gradient(cmap='Blues')

INSERT VISUAL HERE

While there is not one group that shows significant scores across the majority of categories, the largest group is cluster ‘2’ with high scores across activity limitations, lifetime diagnosis of depression, and no leisure (in the last month).

# Interpretations and Limitations

The CDC’s Alzheimer’s Disease and Healthy Aging data set illustrated strong relationships between features in individuals whose reported age was 50-64 at the time of the health questionnaire.

The clustering analysis identified a single group with high scores across the majority of the included features. High scores were observed in categories: binge drinking (in the last month), lifetime depression diagnosis, fair-poor self-reported health and no leisure (past month).

The study is limited by missing data as a result of variation in questionaires per year as well as topics included on each questionnaire. In addition, the majority of the values were recorded with direct relation to self reporting. This causes issues as some reports may be influenced by events that happened at the center the day before or after changing mood. In addition, not every ethnicity is represented across every feature, this may or may not be representative of the population in the area and no additional information is provided to confirm this.

In the future I’d like to have more stuff. Finance. Other things like times visited by others, other stud like that

8 years of professional experience

How much experience do you have??

I have been working with machine learning for about a year now. I don’t have professional experience with BFSI but I have about 8 years of experience in the insurance industry which is highly regulated and works with sensitive personal health information. I considered getting into finance and banking several years back as an investment representative but pursued my M.B.A. instead.

What pay are you looking for??

Definitely looking for a W2 position. My hourly rate is flexible depending on the total compensation package (health benefits, 401k, etc.) I’m looking for 75,000-90,000 in annual salary.