NEP G

DECEMBER 2024

6001 - INTRODUCTION TO PYTHON PROGRAMMING

PROJECT- I

MS Data Science
University of New Haven

Presented by: (Nep G)
Stuti Bimali
Ayush Dangol
Binay Dhakal
Hrishabh Mahaju

GITHUB LINK: HTTPS://GITHUB.COM/ADONGOL123/ALZHEIMER-S_DISEASE

Alzheimer's Disease and Healthy Aging Data

Alzheimer's Disease and Healthy Aging Data contains data from BRFSS. The Behavioral Risk Factor Surveillance System (BRFSS) is the nation’s premier system of health-related telephone surveys that collect state data about U.S. residents regarding their health-related risk behaviors, chronic health conditions, and use of preventive services.

Objectives

1. Data PreProcessing
2. Analyzing Trends and Growth Rates Over Time
3. Confidence Interval Analysis
4. K-Means Clustering Analysis on Alzheimer’s Disease Dataset

Data PreProcessing And Imputation

1

Data Loading: We used pandas to read the data from the CSV we downloaded from the Alzheimer's Official Website. The dataset had 31 features/series/columns and 284142 rows/samples.

2

Transform the data: Renamed the columns by stripping the spaces and adding an underscore to them. We then checked for missing values using isna().sum() method on a data frame to see if there are any missing values.

3

Analyze Missing Values And Imputation: Found out that 9 features/series had missing values, out of which 4 of them were numerical columns and the remaining were categorical. We imputed the numerical values using the mean of the series and also imputed the missing categorical values using the mode of the column.

4

Summary of dataset: After preprocessing we again used describe() method to see the summary of both numerical and categorical series so that we don't miss any missing values.

Data PreProcessing And Imputation

DATA TYPES OF SERIES

```
RowId                          object
YearStart                       int64
YearEnd                         int64
LocationAbbr                   object
LocationDesc                   object
Datasource                     object
Class                          object
Topic                          object
Question                       object
Data_Value_Unit                object
DataValueTypeID                object
Data_Value_Type                object
Data_Value                    float64
Data_Value_Alt                float64
Data_Value_Footnote_Symbol     object
Data_Value_Footnote            object
Low_Confidence_Limit          float64
High_Confidence_Limit         float64
StratificationCategory1        object
Stratification1                object
StratificationCategory2        object
Stratification2                object
Geolocation                    object
ClassID                        object
TopicID                        object
QuestionID                     object
LocationID                      int64
StratificationCategoryID1      object
StratificationID1              object
StratificationCategoryID2      object
StratificationID2              object
dtype: object
```

% OF MISSING VALUES

```
Percentage of missing values:
 rowid                          0.000000
yearstart                       0.000000
yearend                         0.000000
locationabbr                    0.000000
locationdesc                    0.000000
datasource                      0.000000
class                           0.000000
topic                           0.000000
question                        0.000000
data_value_unit                 0.000000
datavaluetypeid                 0.000000
data_value_type                 0.000000
data_value                     32.143787
data_value_alt                 32.143787
data_value_footnote_symbol     61.295409
data_value_footnote            61.295409
low_confidence_limit           32.218046
high_confidence_limit          32.218046
stratificationcategory1         0.000000
stratification1                 0.000000
stratificationcategory2        12.976962
stratification2                12.976962
geolocation                    10.730198
classid                         0.000000
topicid                         0.000000
questionid                      0.000000
locationid                      0.000000
stratificationcategoryid1       0.000000
stratificationid1               0.000000
stratificationcategoryid2       0.000000
stratificationid2               0.000000
dtype: float64
```

SUMMARY OF NUMERICAL COLUMNS

	yearstart	yearend	data_value	data_value_alt	low_confidence_limit	high_confidence_limit	location
count	284142.000000	284142.000000	284142.000000	284142.000000	284142.000000	284142.000000	284142.00000
mean	2018.596065	2018.657735	37.676757	37.676757	33.027824	42.595333	800.3226
std	2.302815	2.360105	20.769560	20.769560	19.997904	21.534500	2511.5649
min	2015.000000	2015.000000	0.000000	0.000000	-0.700000	1.300000	1.00000
25%	2017.000000	2017.000000	24.200000	24.200000	19.400000	29.200000	19.00000
50%	2019.000000	2019.000000	37.676757	37.676757	33.027824	42.595333	34.00000
75%	2021.000000	2021.000000	42.400000	42.400000	35.700000	50.000000	49.00000
max	2022.000000	2022.000000	100.000000	100.000000	99.600000	100.000000	9004.00000

SUMMARY OF CATEGORICAL COLUMNS

	rowid	locationabbr	locationdesc	datasource	class	topic	question	da
count	284142	284142	284142	284142	284142	284142	284142	
unique	36046	59	59	1	7	39	39	
top	BRFSS~2022~2022~42~Q03~TMC01~AGE~RACE	US	United States, DC & Territories	BRFSS	Overall Health	Frequent mental distress	Percentage of older adults who are experiencin...	
freq	15	6132	6132	284142	96753	11092	11092	

Analyzing Trends and Growth Rates Over Time

1 **Yearly Growth Analysis:** We examined the annual growth rate of reported data entries, which showed significant year-over-year changes, reflecting shifts in how data is being reported.

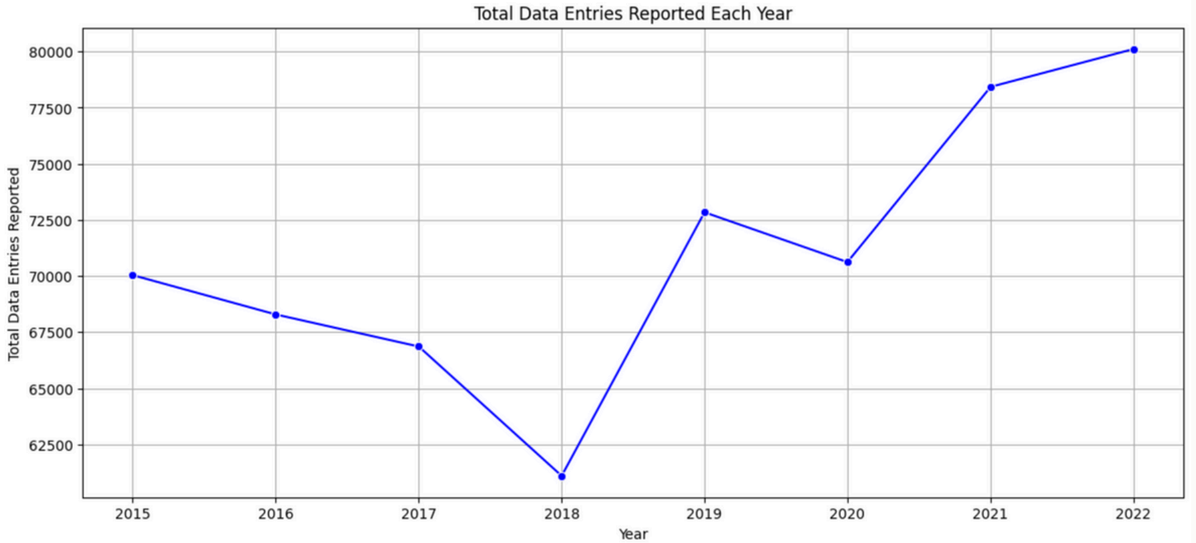
2 **Trends Over Time:** We looked at the data values over the years, identifying whether the values showed consistent patterns or fluctuations.

3 **State-wise Insights:** A regional analysis was performed to identify the top 10 states with the highest average data values, providing insights into geographical differences.

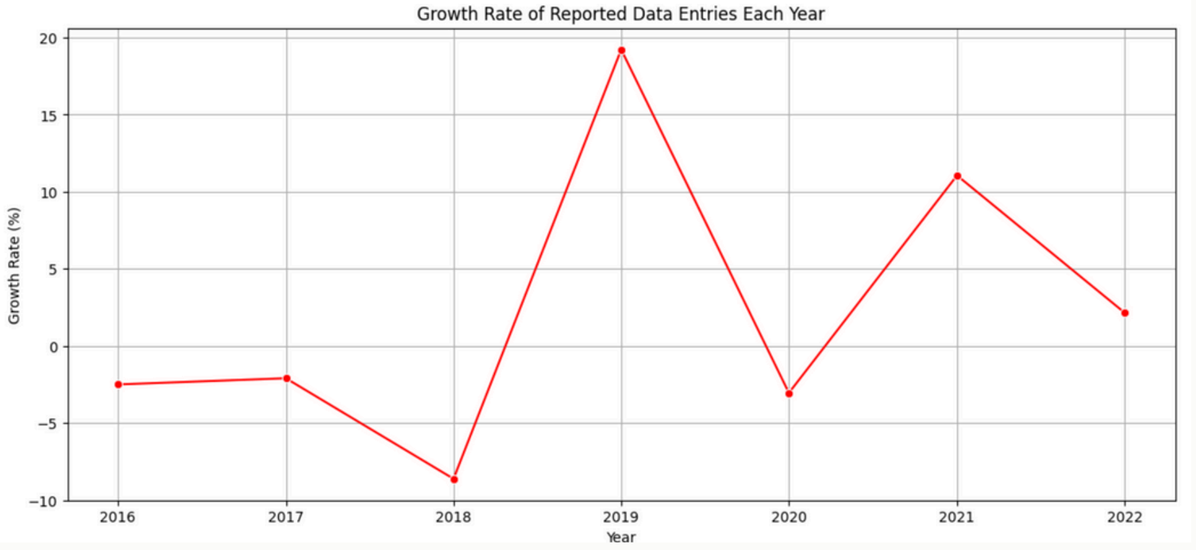
4 **Heatmap Visualization:** A heatmap was created to visualize how data values were distributed across states and years, revealing both temporal and geographical variations in the dataset.

Yearly Growth Analysis

- **Analysis Focus:** Examined the number of data entries reported each year.
    - **Approach:** Combined counts of data entries for the start and end years and aggregated them.

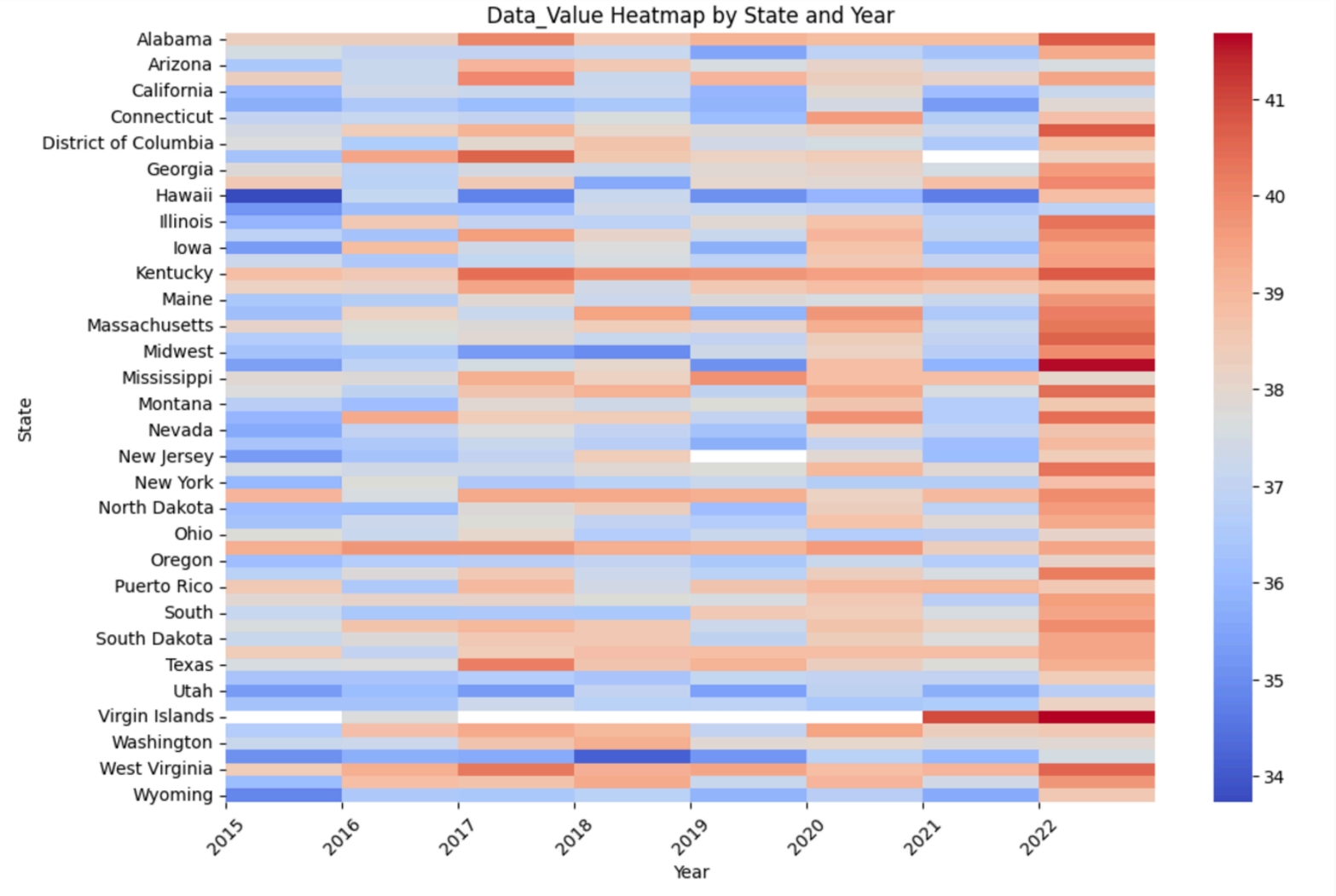


- **Analysis Focus:** Calculated the annual growth rate in reported data entries.
    - **Approach:** Used percentage change (pct_change) to measure the growth or decline in data entries from one year to the next.
    - **pct_change** => (New Value - Old Value) / Old Value * 100



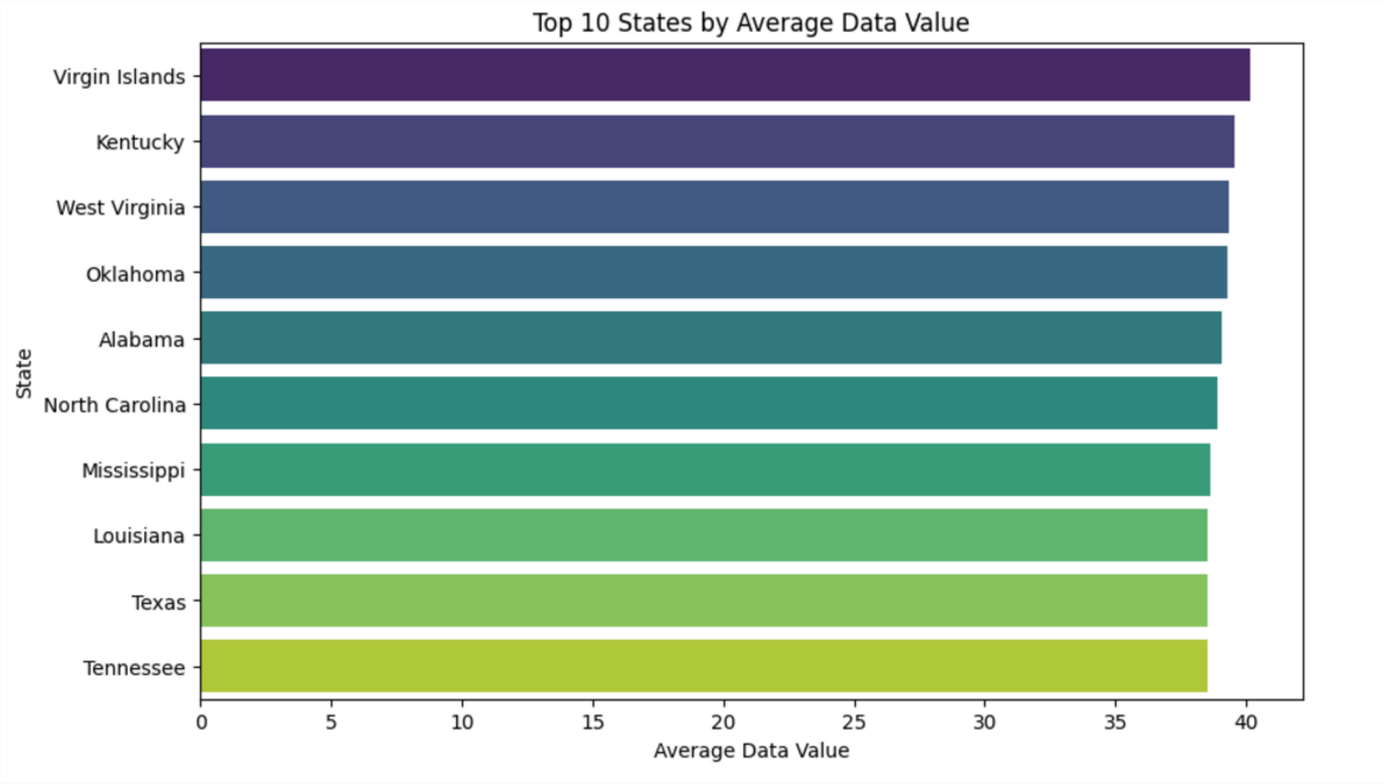
Heatmap of Data Values

- **Analysis Focus:** Examined the distribution of data values across states and years.
    - **Approach:** Created a heatmap showing data values for each state across different years.



Aggregate data by location

- **Analysis Focus:** Identified the states with the highest average data values.
    - **Approach:** Grouped the data by location and calculated the average data value for each state.



Confidence Interval Analysis

1

What is a Confidence Interval?

- A range of values used to estimate a population parameter.
- Indicates the reliability of the estimate.

2

Why is it Important?

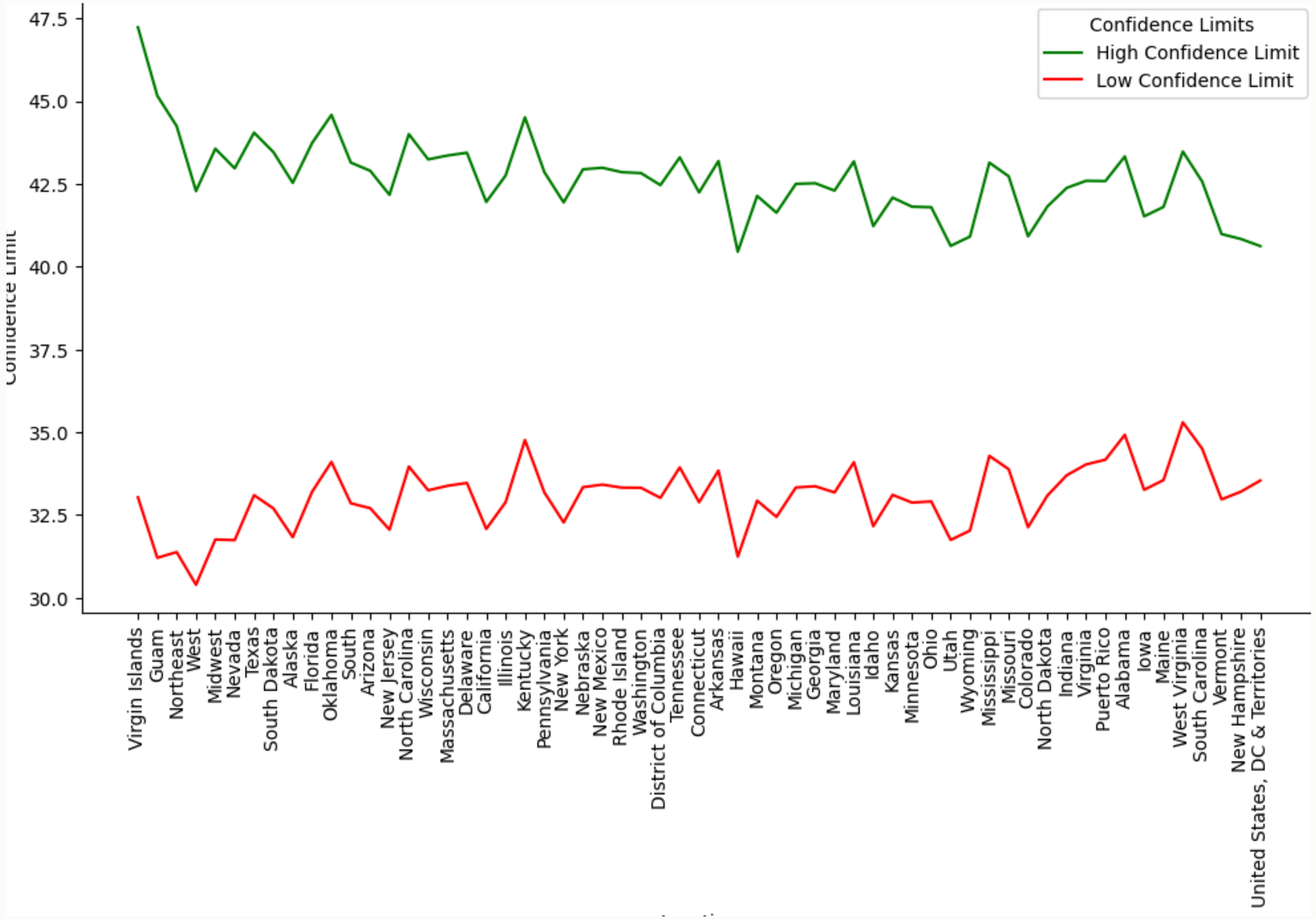
- Helps in making inferences from sample data.
- Essential in research and decision-making.

3

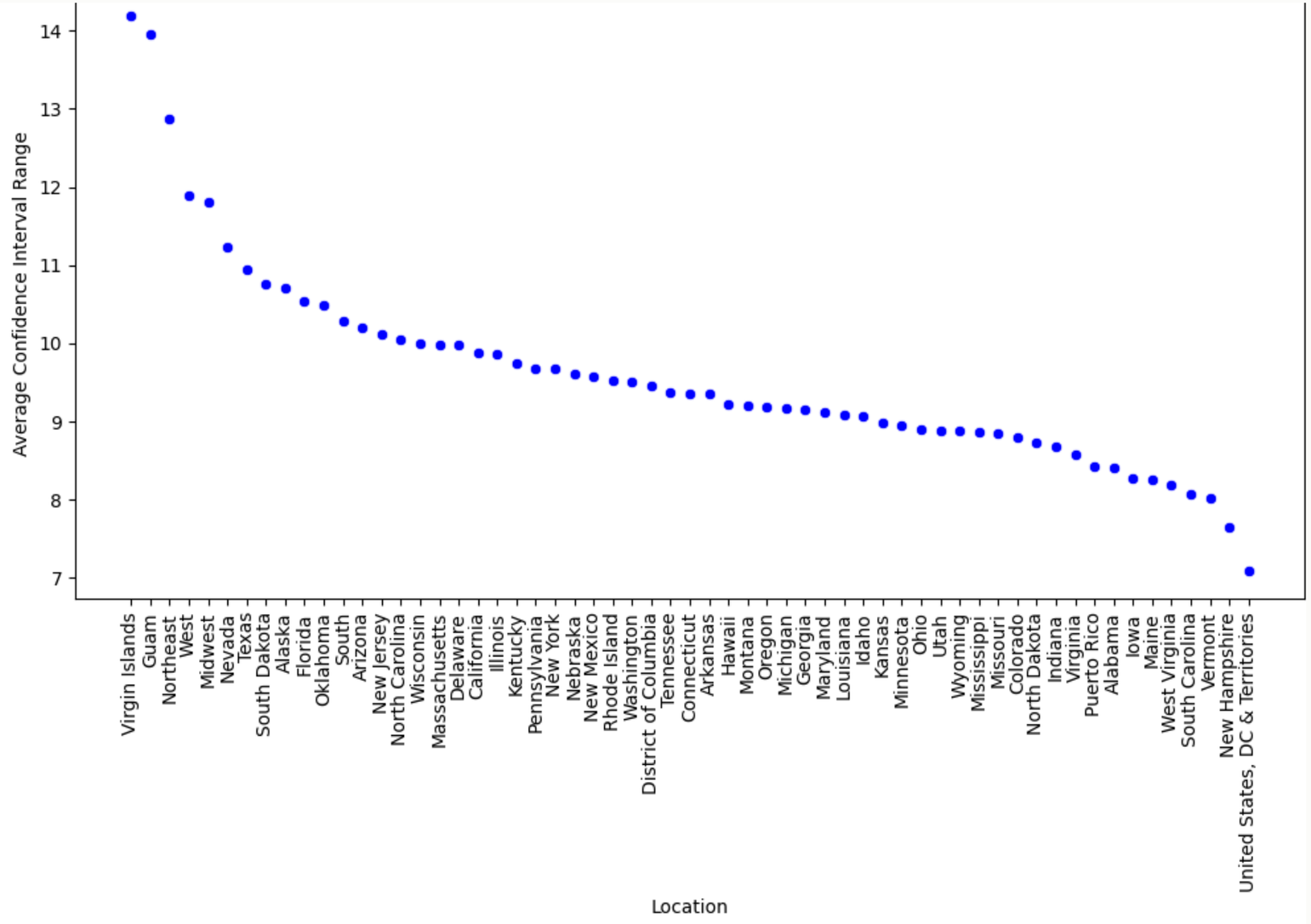
Example:

- if we get 93% confidence level, it means that out of 100, 93 times the value may be in between upper and lower limit.

Precision of Data in different location



Difference in Confidence Limit



Confidence level for Cognitive Decline

	locationdesc	topic	Average_Confidence_Level
0	Alabama	Functional difficulties associated with subjec...	43.507593
1	Alabama	Need assistance with day-to-day activities bec...	39.722299
2	Alabama	Subjective cognitive decline or memory loss am...	23.044620
3	Alabama	Talked with health care professional about sub...	42.136269
4	Alaska	Functional difficulties associated with subjec...	36.316158
...	...	...	...
223	Wisconsin	Talked with health care professional about sub...	41.924095
224	Wyoming	Functional difficulties associated with subjec...	34.289122
225	Wyoming	Need assistance with day-to-day activities bec...	31.605789

	locationdesc	topic	Average_Confidence_Level
164	Puerto Rico	Functional difficulties associated with subjec...	53.250863
165	Puerto Rico	Need assistance with day-to-day activities bec...	52.277863
118	Nevada	Subjective cognitive decline or memory loss am...	26.344122
167	Puerto Rico	Talked with health care professional about sub...	53.246863

K-Means Clustering Analysis on Alzheimer's Disease Dataset

- K-means clustering is an unsupervised learning method, meaning that the algorithm doesn't have ground truth to compare its output to. Instead, it investigates the structure of the data by grouping data points into distinct subgroups.
- K-means clustering is a popular algorithm that categorizes data points into clusters based on their proximity to a cluster center. The goal is to minimize the sum of distances between data points and their assigned clusters.
- Here's how the k-means clustering algorithm works:

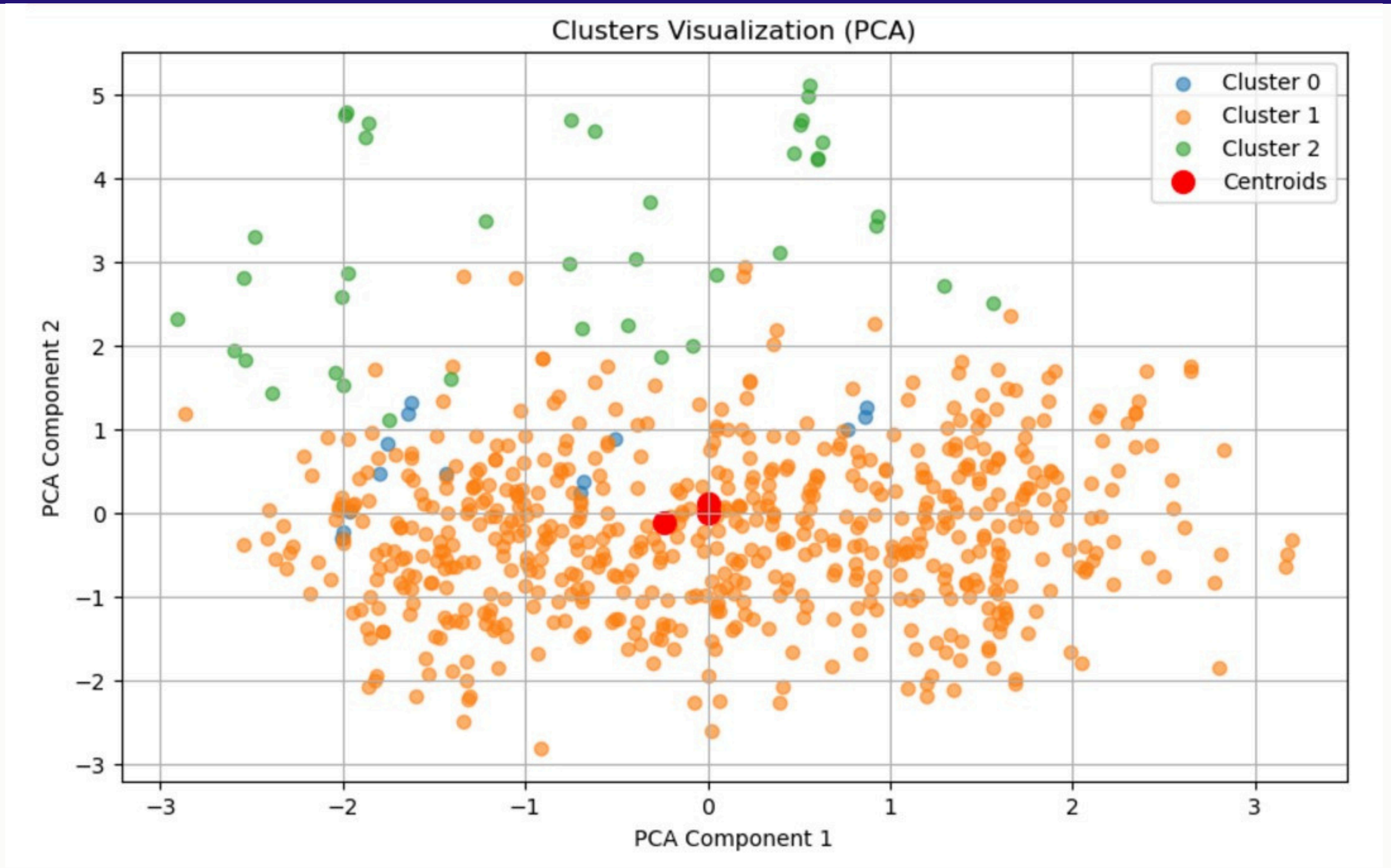
1. Randomly place centroids: Start by randomly placing a set number of cluster centers, or centroids.
2. Assign data points to centroids: Assign each data point to the centroid that is closest to it.
3. Update centroids: Recalculate the location of each centroid as the mean of all the points assigned to it.
4. Repeat: Repeat steps 2–3 until the centroids stop moving or the points stop switching clusters.

- The number of clusters, or k, determines the size of the clusters. A higher k value results in smaller clusters with more detail, while a lower k value results in larger clusters with less detail.
- K-means clustering is used in a variety of applications, including market segmentation, image segmentation and compression, and document clustering.

K-Means Clustering Analysis on Alzheimer's Disease Dataset

- Key Points:
  - Objective: Group locations and demographics based on Alzheimer's prevalence rates (Data_Value) and related demographic characteristics.
- Methodology:
  - Dimensionality Reduction: Principal Component Analysis (PCA) used to transform the data into 2 main components (PCA1 and PCA2) for visualization.
- Clustering:
  - Applied K-Means to group data points into 3 clusters based on patterns in prevalence and demographics.

K-Means Clustering Analysis on Alzheimer's Disease Dataset



K-Means Clustering Analysis on Alzheimer's Disease Dataset

- Findings:
- Cluster Overview:
  - Cluster 1 (Blue): Represents outliers or unique locations with distinct prevalence patterns.
  - Cluster 2 (Orange): Largest group, likely representing average Alzheimer's prevalence rates and demographic characteristics.
  - Cluster 3 (Green): Represents locations or demographics with significantly higher prevalence rates or unique demographic factors.
- PCA Axes Interpretation:
  - PCA1: Captures primary variation in prevalence rates.
  - PCA2: Highlights demographic differences (e.g., age or gender variability).
- Actionable Insights:
  - Investigate Cluster 1 for potential outliers or extreme cases.
  - Explore how demographic factors contribute to differences in prevalence within Clusters 2 and 3.

Any Questions ?