Final Project Report

ISM 6419: Data Visualization

Exploring Hidden Patterns for Alzheimer’s Disease and Other Conditions that Attack the Cognitive Ability of the Elderly

By Caitlin Salerno

Introduction

One in nine Americans 65 years of age or older have Alzheimer's disease, according to the Alzheimer's Association, yet so little is known about it (Alzheimer's Association, 2021). A sickeningly slow condition which causes someone to become a shell of who they once were in front of their own and loved one's eyes, Alzheimer's disease and other dementias are not only frightening, but alarmingly common. Coming from a family where we share a hereditary gene that makes us more likely to develop early-onset Alzheimer's, as my grandmother did at 52 years of age, Alzheimer's is a disease that I tend to research often in the chances that someone in my family (or I) get diagnosed. However, I am never able to easily gather large amounts of research or visualizations on the many studies being done on Alzheimer's and other forms of dementia, as the information is mostly kept out of the public domain, only providing published summaries of the data.

The main motivation for my project is to take publicly available information I can find, clean it up (which took weeks of work to do, as the data that is available is very hard to interpret) , and create visualizations that both my family and I can learn from in order to help prevent ourselves from suffering the horrible fate that Alzheimer's brings. I would also like to share this information publicly as so many others have the same fears that I do revolving around dementia.

To speak further of my ambitions toward the project, I took not only information from the BRFSS but also the U.S. Department of Agriculture 2021 population data by state to ensure that the survey represented all states correctly. This means that I linked 4 separate data sources for my project. Lastly, I moved my data throughout Excel, CSS, and Tableau Data Prep to clean the BRFSS data, along with using Tableau to visualize it, so I used multiple pieces of software we learned about in this course to solve the problem of the hard-to-interpret data.

My main research questions that drove me to choose my variables explained in the next section are as follows:

-Does the demographic data of an individual impact one's chances of developing Alzheimer's?

-Is someone more likely to hurt themselves if they develop Alzheimer's or another dementia?

-Does a person's habits or medical history affect their chances of developing Alzheimer's?

-Taking the elderly population percentages into account, does geographical region affect one's chances of developing Alzheimer's or another dementia?

The Data - Methodology

While information strictly related to Alzheimer's and other forms of dementia was hard to find publicly from reputable sources, I gathered three years of data (2019, 2020, and 2021) from the Behavioral Risk Factor Surveillance System (BRFSS), a yearly mass survey conducted by the Centers for Disease Control and Prevention (CDC) (CDC - 2019 BRFSS Survey Data and Documentation, 2019, 2020, 2021). This information is publicly available but is very hard to interpret as the data comes in the form of an ASCII file with no delimiters. The data came from three separate pages and therefore three separate files.

The BRFSS datasets contain information on hundreds of different topics, including health conditions, habits, demographic information and risk factors/behaviors. It is the largest health survey (due to the number interviewed) that has been continuously done in the world, with the first yearly survey being completed in 1984 (CDC - about BRFSS, 2020).

I also imported population data, as of 2021, from the U.S. Department of Agriculture to compare the distribution of the population surveyed to the actual population distribution in the United States (Population, 2022). This would help me prepare later visuals.

Lastly, as one of the columns in the BRFSS data was about which state the individual was from but was in numerical format, I translated the numbers to which state it corresponded to via a third table that I built myself in Excel.

The Data – Cleaning

Each year in the BRFSS data had its own codebook, which are all contained in the zip file that this report lies in. I went through each codebook, marking which variables I wished to keep from the thousands of variables to choose from. Each dataset has its variables in different locations, so each codebook needed to be looked at. Some variables were also present in certain years and not all three years. While there was no research question directly asking someone if he or she has Alzheimer's or another form of dementia, there were two variables that were important for tracking this disease: Memory Loss (specifically, confusion or memory loss that is getting worse) and Decisions (whether a physical, mental or emotional condition was making it hard to make or remember decisions). I then went through the codebooks again, marking where I should set the width of the delimiters when importing my data based on the column numbers provided in the code book. I tracked these items in an Excel document labeled "Variables project data visualization" which is also in the same zip file as this report.

Once I knew what fixed width to set my delimiters to, I imported the ASCII files by importing a folder into Excel to use the power query editor there. I set the delimiters for all three of these ASCII files, making three separate Excel files (by year that the survey was conducted), which I also added Column titles into since the ASCII files did not have titles. These titles helped me in my next steps of transforming the data. Lastly, I added an ID number to each row. The ID number included the year, so that the data could be merged later and still have unique IDs for rows.

My next step to transforming the data was to put all three Excel files that I formed into the Tableau Prep Builder application. I took out all the columns that were grouped together from the manual delimiters but not needed (and not given column names because of this). I also renamed all the fields in the three files to match one another, including the variable order. I merged all three files into one large CSV file with over 1 million rows, and the variables that I still had an interest in but were not consistently present across all three years were placed into separate data files. The Tableau Prep Builder data flow for this project can be found in the zip folder that this report lies in.

I took the merged data file and dove back into the codebooks for the three years of datasets. For every variable that was numerical in my data but actually a categorical variable, I transformed it by creating a second variable (so the original was not lost) and using Excel formulas, matched the number to the codebook. Since the State variable had so many different possible answers, I created a separate table for making the link between the numerical values and the categorical answer that was given in the survey. Some examples of the different formulas used are as follows:

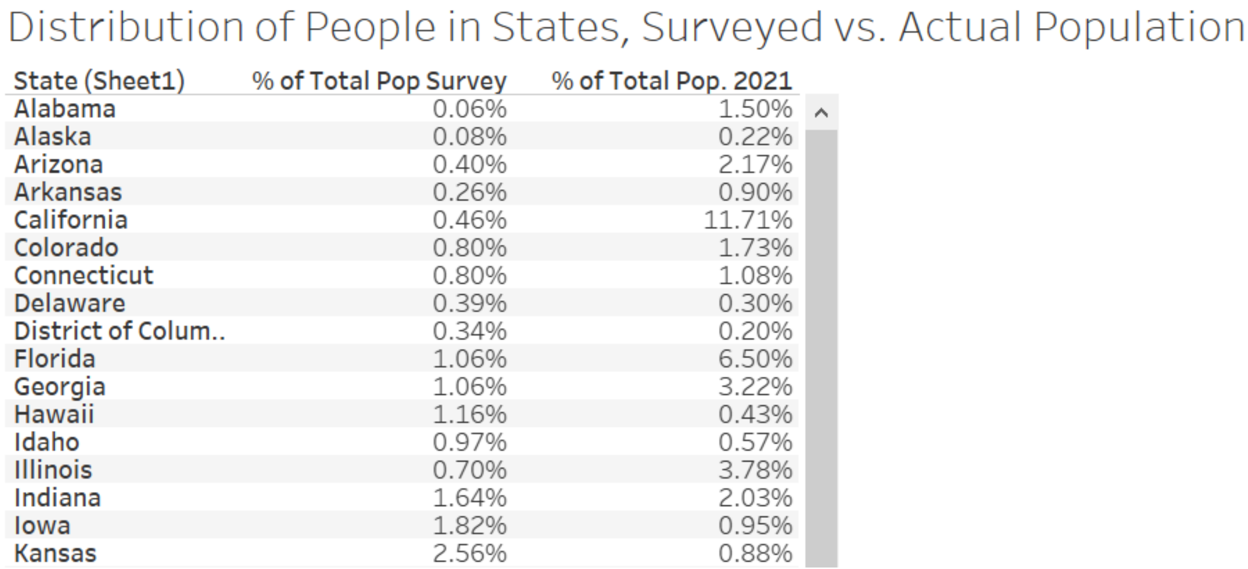
<u>SmokeFreq2 (Smoking Frequency)</u>
=IF(AU3=1, "Every Day", IF(AU3=2, "Some Days", IF(AU3=3, "Not at All", IF(AS3=2, "Not at All", AU3))))

<u>Race</u>
=IF(BE2=1,"White",IF(BE2=2,"Black",IF(BE2=3,"American Indian/Alaska Native",IF(BE2=4,"Asian",IF(BE2=5,"Native American/Pacific Islander",IF(BE2=6,"Other Race",IF(BE2=7,"Multiracial",IF(BE2=8,"Hispanic","N/A"))))))))

I followed the same process for a variable in one of the separated tables with the variables not consistent across all three years, as I only ended up using information from one and not both of the tables created in the Tableau Prep Builder data flow. Now that my BRFSS data was cleaned, I was able to start the data visualization process. The population data gathered from the U.S. Department of Agriculture was already clean for the most part, and I just needed to delete the row for the overall United States population before importing it for use.

The Data – Analysis

After importing my datasets and linking them correctly in Tableau, the first visualization I did was to compare the population distribution of the survey across all U.S. states to the actual distribution of the population as of 2021. What I would find would allow me to make decisions on how I needed to set up future visualizations, especially ones that include maps.



Distribution of People in States, Surveyed vs. Actual Population

State (Sheet1)	% of Total Pop Survey	% of Total Pop. 2021
Alabama	0.06%	1.50%
Alaska	0.08%	0.22%
Arizona	0.40%	2.17%
Arkansas	0.26%	0.90%
California	0.46%	11.71%
Colorado	0.80%	1.73%
Connecticut	0.80%	1.08%
Delaware	0.39%	0.30%
District of Colum..	0.34%	0.20%
Florida	1.06%	6.50%
Georgia	1.06%	3.22%
Hawaii	1.16%	0.43%
Idaho	0.97%	0.57%
Illinois	0.70%	3.78%
Indiana	1.64%	2.03%
Iowa	1.82%	0.95%
Kansas	2.56%	0.88%

Figure 1: Screenshot of PopDist Worksheet

According to the visualization created, the population distribution of the survey does not match the population distribution of the different states in the U.S. accurately. This means that for any visualizations that include the state, using the percentage of a factor occurring based on the state would be far better than using an average or sum function, as the incorrect distribution would make weighted normalization much harder.

My next visualizations tackled my research question, “Is someone more likely to hurt themselves if they develop Alzheimer’s or another dementia?”. I found that a great variable to answer this question would be to use the ‘Falls’ variable, which tracks the number of times that someone has fallen in the last year. I filtered my data to exclude missing data points in the survey, and only include those who are ages 65 and older. I would consider these two visualizations to be simple. I created two bar graphs with this variable, and used it to make a dashboard. One visualization focuses on comparing the average number of falls to the Memory Loss variable, as mentioned earlier to be a good link to whether someone may have Alzheimer's. The other visualization focuses on comparing the average number falls to the Decision variable, also mentioned before.

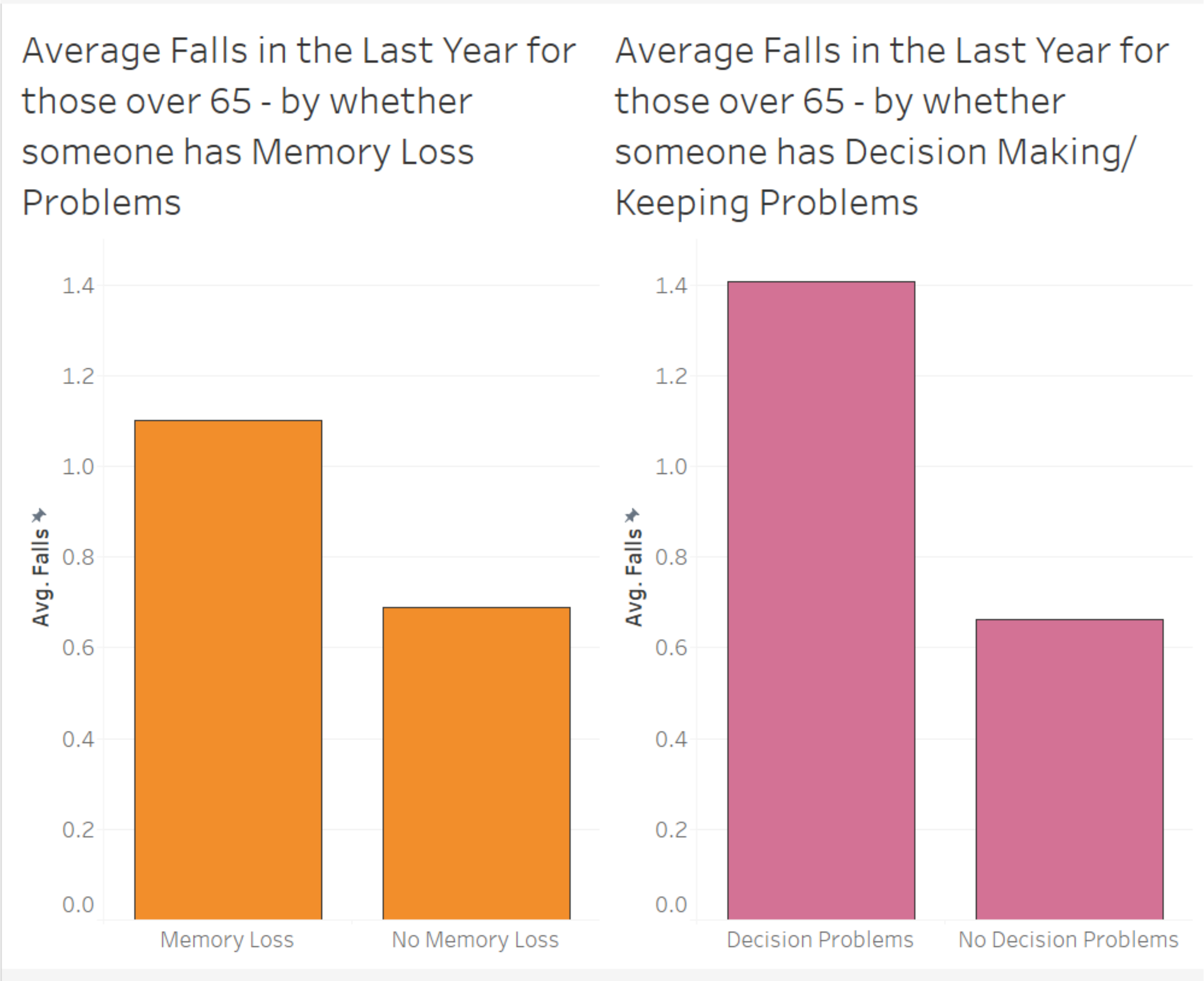


Figure 2: Screenshot of FallsDashboard in Tableau

Both visualizations tell an interesting story regarding the affect that cognitive decline can have on one's own personal safety. Having memory loss issues or decision making/keeping problems both are linked to a higher average amount of falls per year for the elderly.

My next research question, "Does the demographic data of an individual impact one's chances of developing Alzheimer's?" was addressed via my next few visualizations. I decided to look into the area of race, as I have seen many studies on gender before. I created a dashboard with three simple visualizations, all being pie charts. The first pie chart represented the distribution of race for all elderly people across the United States (for the survey), the second pie chart represented the distribution of race for all elderly people who have memory problems, and the third pie chart represented the distribution for all elderly people who have problems making and remembering decisions. Since my findings from the first visualization showed that the distribution of the survey population across different states was not accurately representing the actual population distribution for the United States, I assumed that this would also be the case for the distribution of race. Due to this assumption, I compared the distribution of the elderly with memory loss and decision making/remembering problems with the distribution of race in the entire survey rather than the population distribution of race in the United States.

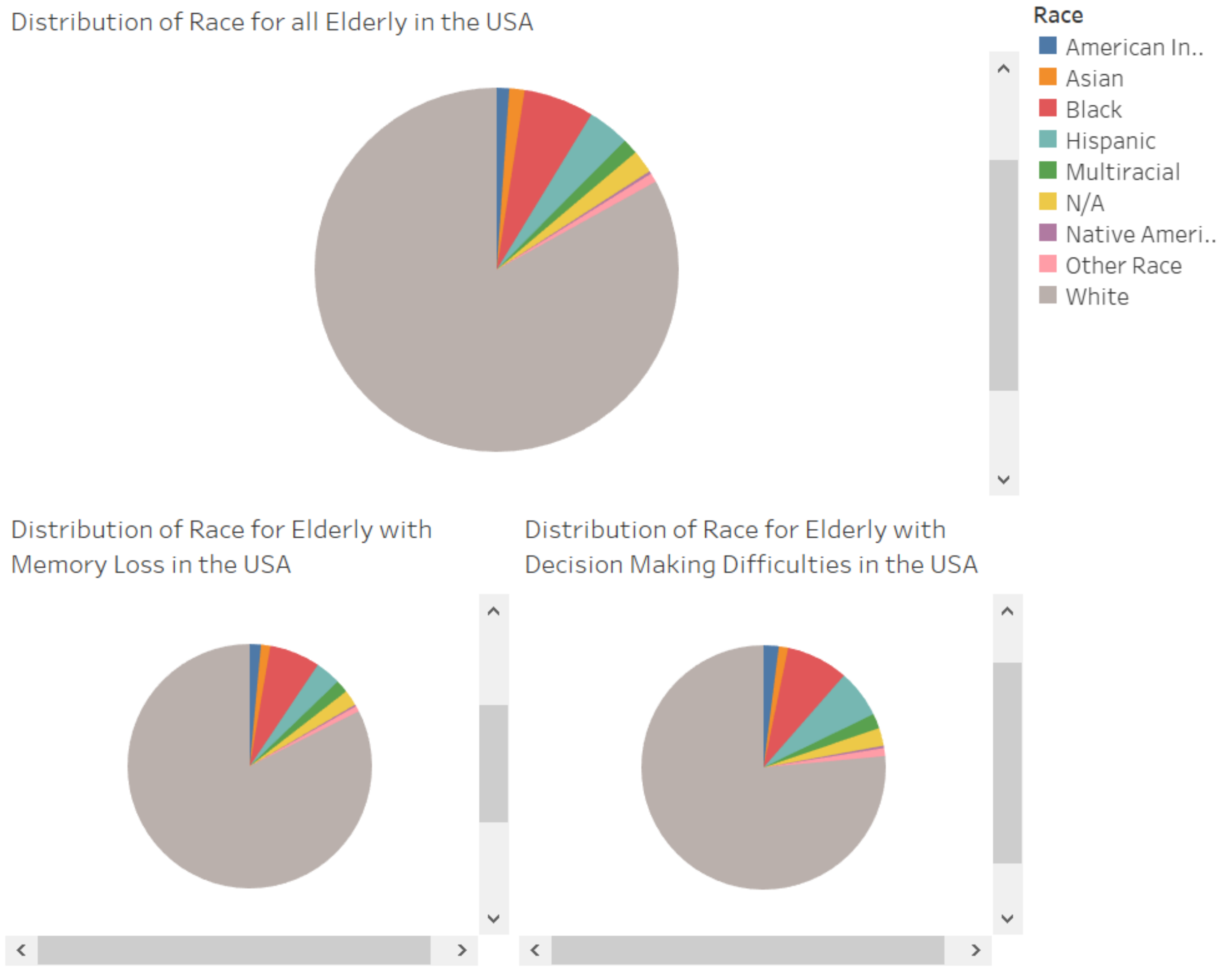


Figure 3: Screenshot of RaceDashboard in Tableau

Using the dashboard was the best way to compare these distributions to one another. To the naked eye it seems as if all distributions are relatively the same, but upon inspecting each element, it is clear that there is a higher percentage of elderly Black, Hispanic and American Indian/Alaska Native people that have difficulty making and remembering decisions than their normal population distribution percentages without this factor being singled out. However, the distribution of race for elderly people with memory loss in the USA seems to be nearly identical to the distribution of race for the elderly in the entire study.

For the research question "-Does a person's habits or medical history affect their chances of developing Alzheimer's?", I used multiple individual charts that worked best on their own, without a dashboard, due to their complexity. The first part of this question that I dove into was related to habits. I compared the percentage of elderly people who had memory loss problems to ones that did not, depending on whether they had every smoked 100 cigarettes or more in their lifetime. I filtered both the memory loss and smoking variables to only include survey answers that were “yes” or “no” to make the visualization clearer, and also ensured that only those who answered in the survey that they were age 65 and older were included in this chart.

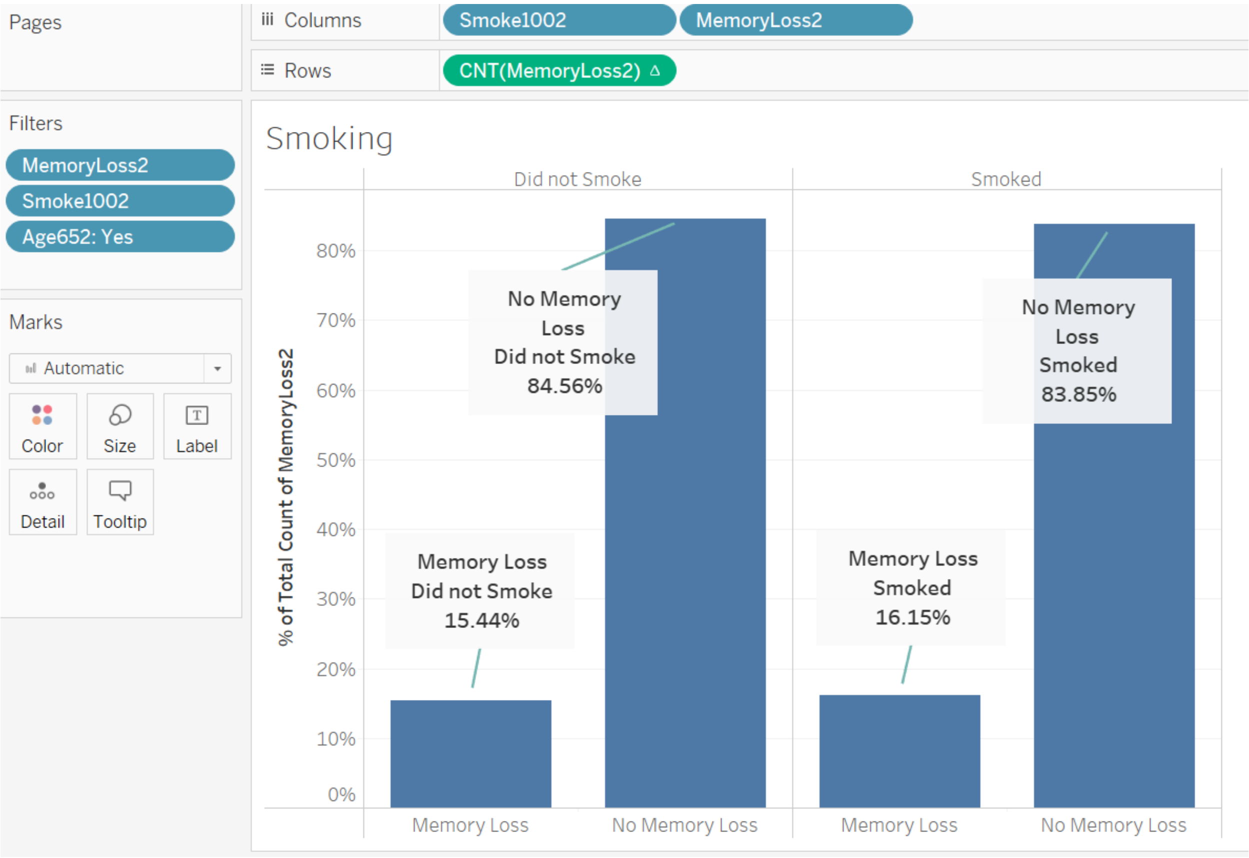
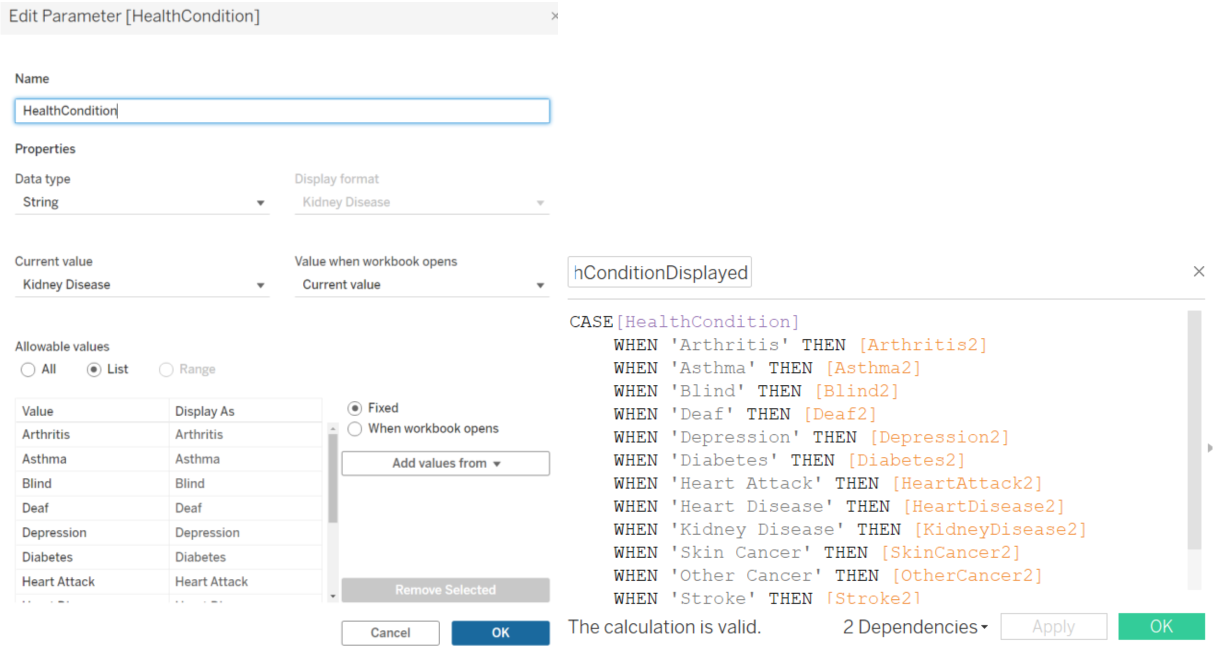


Figure 4: Screenshot of Smoking visualization in Tableau

According to the visualization, elderly people who smoked 100 cigarettes or more in their lifetime were more likely to experience memory loss issues.

Continuing to search for answers to my research question, I then focused on pre-existing health conditions to see if these had any affect on memory loss or decision making/remembering. I created two complex and interactive visualizations where I used parameters to allow the user to choose between several different pre-existing health conditions and to see whether it affects an elderly person’s chances of having memory problems (first interactive graph) or decision making/remembering problems (second interactive graph).



Edit Parameter [HealthCondition]

Name: HealthCondition

Properties

Data type: String

Display format: Kidney Disease

Current value: Kidney Disease

Value when workbook opens: Current value

Allowable values: ○ All ● List ○ Range

Value	Display As
Arthritis	Arthritis
Asthma	Asthma
Blind	Blind
Deaf	Deaf
Depression	Depression
Diabetes	Diabetes
Heart Attack	Heart Attack

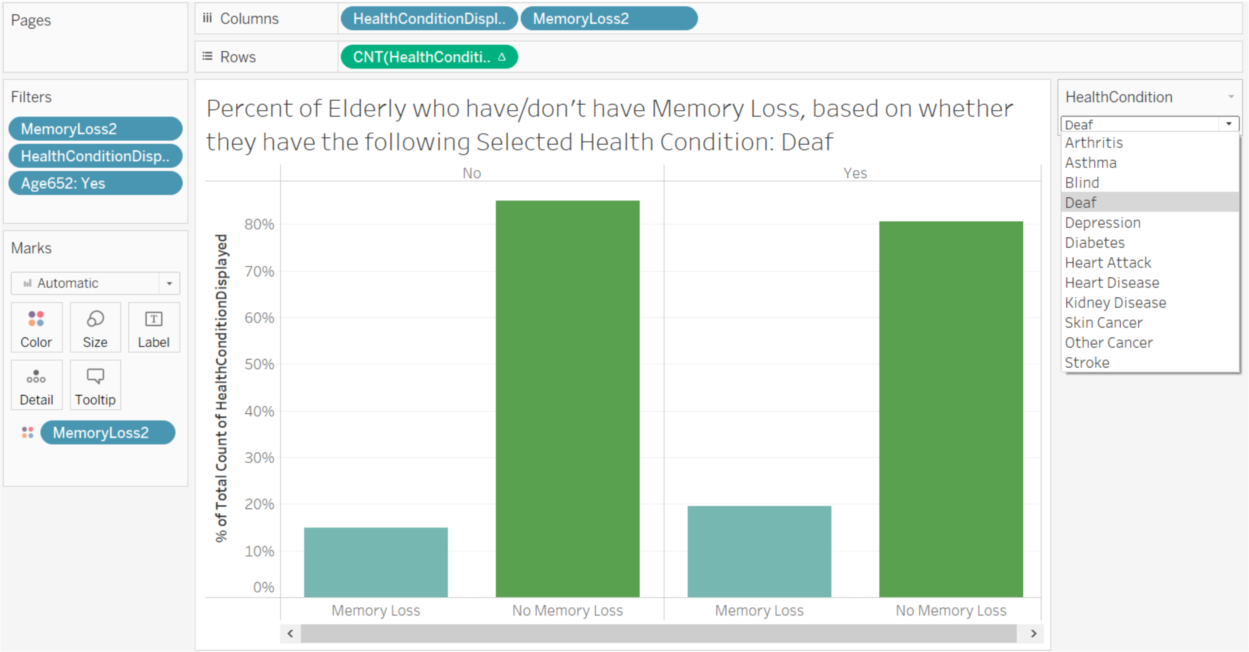
● Fixed ○ When workbook opens

hConditionDisplayed

```
CASE [HealthCondition]
    WHEN 'Arthritis' THEN [Arthritis2]
    WHEN 'Asthma' THEN [Asthma2]
    WHEN 'Blind' THEN [Blind2]
    WHEN 'Deaf' THEN [Deaf2]
    WHEN 'Depression' THEN [Depression2]
    WHEN 'Diabetes' THEN [Diabetes2]
    WHEN 'Heart Attack' THEN [HeartAttack2]
    WHEN 'Heart Disease' THEN [HeartDisease2]
    WHEN 'Kidney Disease' THEN [KidneyDisease2]
    WHEN 'Skin Cancer' THEN [SkinCancer2]
    WHEN 'Other Cancer' THEN [OtherCancer2]
    WHEN 'Stroke' THEN [Stroke2]
```

The calculation is valid.

Figure 5: HealthCondition Parameter and HealthConditionDisplayed calculated field in Tableau



Pages

Columns: HealthConditionDispl.. MemoryLoss2

Rows: CNT(HealthConditi.. Δ

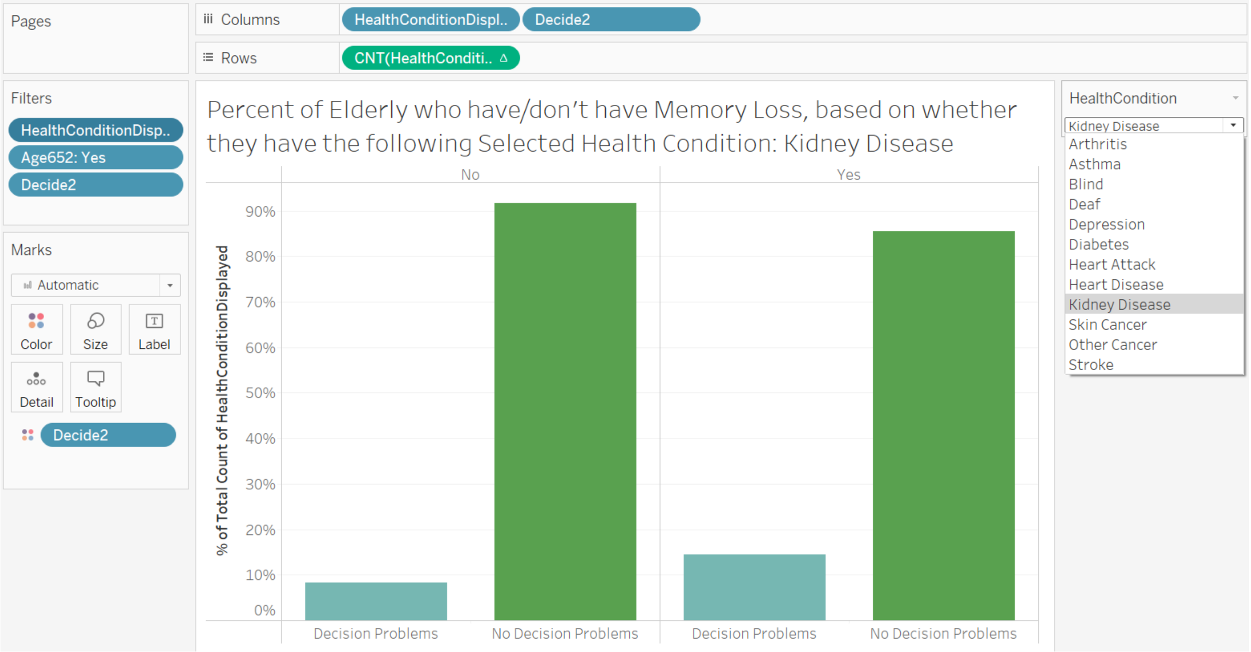
Filters: MemoryLoss2, HealthConditionDisp.., Age652: Yes

Marks: Automatic; Color: MemoryLoss2

HealthCondition: Deaf

Percent of Elderly who have/don't have Memory Loss, based on whether they have the following Selected Health Condition: Deaf

Figure 6: Screenshot of MemoryLoss Interactive visualization in Tableau



Columns: HealthConditionDispl.. Decide2

Rows: CNT(HealthConditi.. Δ

Filters: HealthConditionDisp.. Age652: Yes Decide2

Marks: Automatic

Color: Decide2

HealthCondition: Kidney Disease

Percent of Elderly who have/don't have Memory Loss, based on whether they have the following Selected Health Condition: Kidney Disease

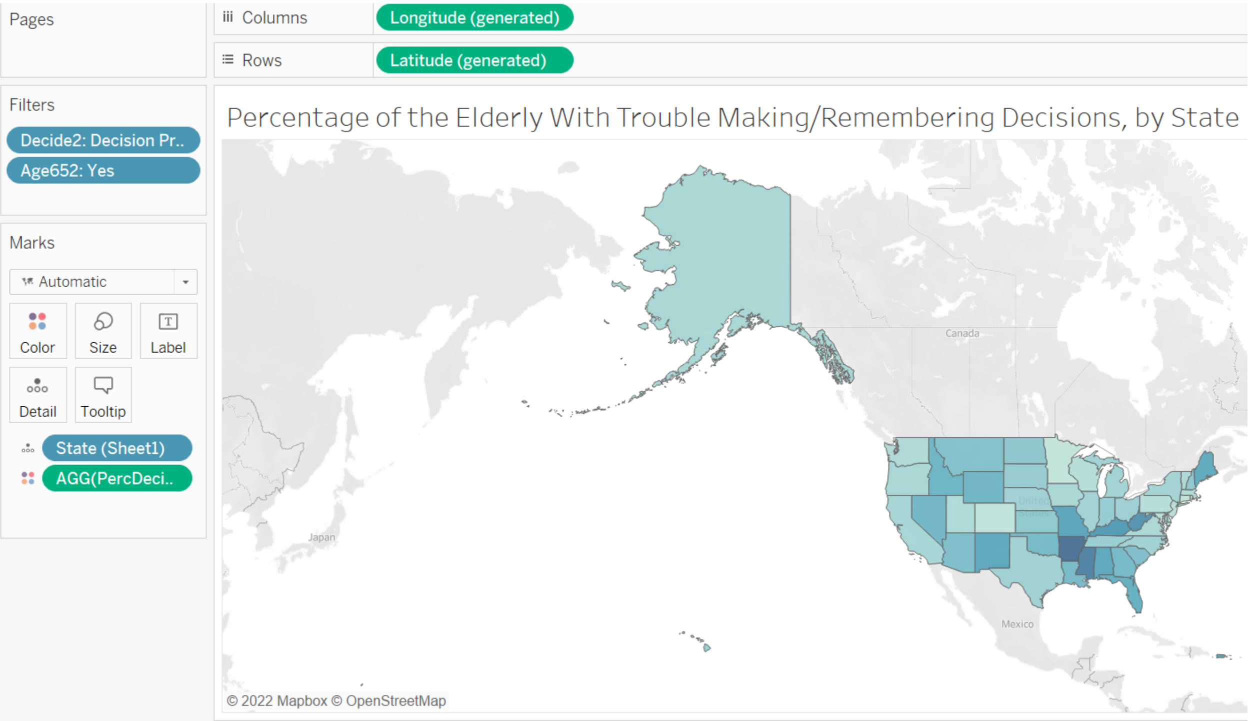
Figure 7: Screenshot of Decision Interactive visualization in Tableau

The parameter lists are expanded so that all the pre-existing conditions that the user can compare are shown. In both visualizations I filtered out the data so that only those who answered “yes” or “no” to having memory problems or decision making/remembering problems were included in the graphs (those who refused to answer or did not know were excluded). I did the same filtering for all answers for the health condition parameters, except for the condition of diabetes, where people who answered “yes”, “no”, “pre-diabetes” and “yes, only during pregnancy” were included in the graph (making four separate bar charts for this parameter selection rather than two). Lastly, the graphs were filtered to only include the elderly from this survey.

When going through the different parameter selections and comparing the side-by-side charts, it was found that having each and every pre-existing health condition that could be chosen in this parameter list led to a higher chance of having memory loss problems. For diabetes, those who had (during pregnancy) or currently have the disease also experienced a higher percentage of those who had memory loss problems than those who did not have diabetes or who were pre-diabetic. Being pre-diabetic did not present much of a difference between those who did not have diabetes.

When going through each parameter selection for the decision-related visualization, the findings were similar to the interactive memory loss visualization, for almost all health conditions did lead to a higher percentage of the elderly having problems making/remembering decisions. However, two health conditions presented different findings: skin cancer and diabetes. Those with skin cancer had virtually no difference from those without skin cancer regarding decision making and remembering, and rather than just those who had (during pregnancy) and currently have diabetes having a larger percentage of the elderly with decision problems, those with pre-diabetes also had a much higher percentage as well.

Due to the last research question I had, “Taking the elderly population percentages into account, does geographical region affect one’s chances of developing Alzheimer’s or another dementia?”, being impacted by my first visualization (where I found that the population across the states was not taken into account by the surveyors), I chose to use percentages rather than weighted population numbers to visualize the data. To do this, I made two calculated fields to find the percentage of the elderly in each state that had memory loss problems or decision making/remembering problems. I then presented these calculated fields in two map charts, where the percentages in the visualization are not that of the entire survey population but are representing the percent of elderly people in each state who answered “yes” to having memory loss problems or decision problems. This is important to note as the percentages in my data will not add up to be exactly 100%, and this was intentional.



Columns: Longitude (generated)

Rows: Latitude (generated)

Filters: Decide2: Decision Pr..; Age652: Yes

Marks: Automatic

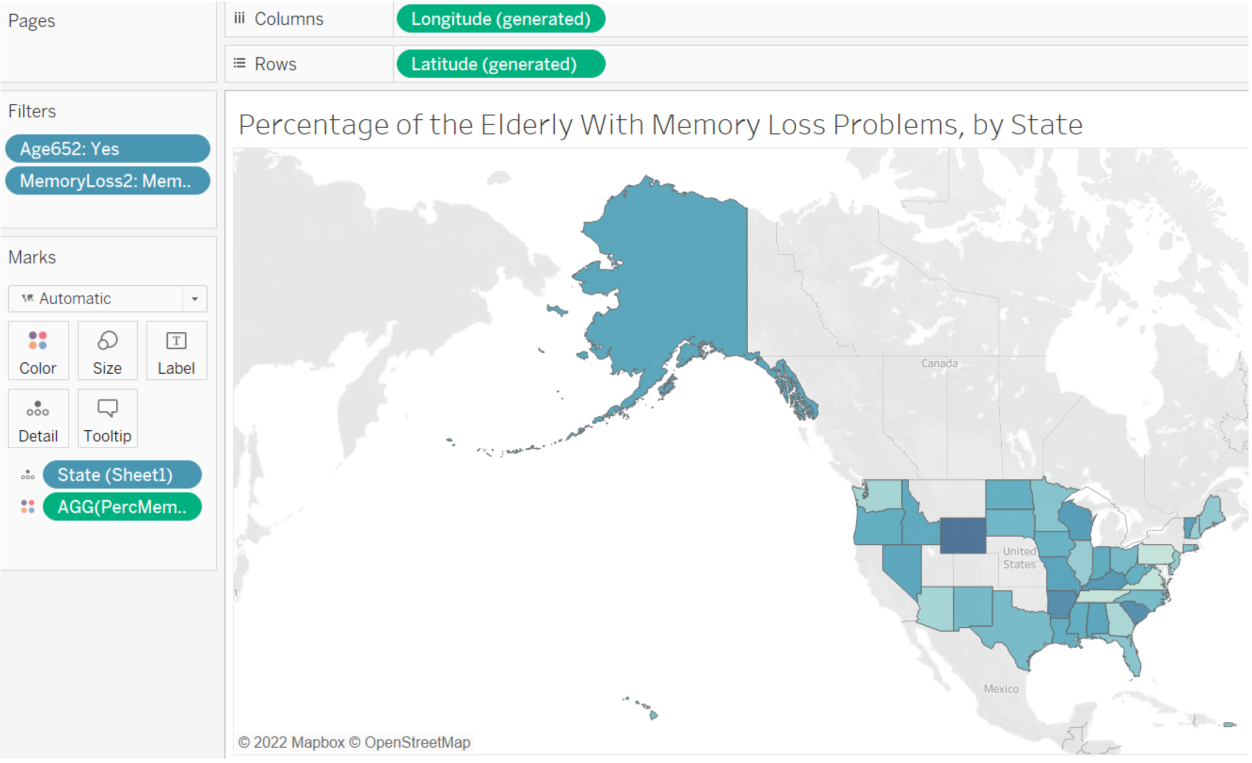
State (Sheet1)

AGG(PercDeci..

Percentage of the Elderly With Trouble Making/Remembering Decisions, by State

© 2022 Mapbox © OpenStreetMap

Figure 8: Screenshot of MapDecision visualization in Tableau



Pages

Columns: Longitude (generated)

Rows: Latitude (generated)

Filters

- Age652: Yes
- MemoryLoss2: Mem..

Marks

Automatic

- State (Sheet1)
- AGG(PercMem..

Percentage of the Elderly With Memory Loss Problems, by State

© 2022 Mapbox © OpenStreetMap

Figure 9: Screenshot of MapMemory visualization in Tableau

When analyzing the visuals, Arkansas seems to have the largest percentage of its elderly population with decision making and remembering problems, with Mississippi, Kentucky and West Virginia not too far behind. Upon the first glance at the map regarding those with memory problems, it is seen that not all states were asked this survey question, which means that this data visualization can be misleading if we use it to rank states that have the highest percentage, as these missing states could potentially have much higher percentages of the elderly with memory problems. However, it can be stated that Wyoming, Arkansas, Kentucky and South Carolina have higher percentages than the other states that actively have data.

Conclusions from Findings

All of the visualizations that I focused on creating were an effort to answer the research questions that I had listed in my introduction, and I was able to find several pieces of information to further my knowledge of the subjects.

Looking at the findings from figure 2, in regard to whether someone who has Alzheimer’s is more likely to be a danger to themselves: Since Alzheimer's is a disease that affects both memory and decision making, we can gather from this information that having the disease can be dangerous for their safety in terms of them having a higher chance of falling more often. An elderly person falling can be dangerous and even fatal, so this is an important finding.

My findings from figure 3 show that although there is a difference in the distribution of race for elderly people if they have trouble making or remembering decisions when compared to the overall elderly survey population, there were no differences found regarding memory loss. Due to Alzheimer's and other dementias affecting both decision making and memory loss, not just one of these factors, it is probably another factor that is causing decision making/remembering to be distributed differently among different races. This is a finding that can make people uneasy, as no one is safe from Alzheimer's due to it affecting all races relatively equally.

When analyzing figure 4, which is centered about one's habits affecting their chances of developing Alzheimer's, it was found in this simple visualization that those who have smoked 100 cigarettes or more, even if they no longer smoke, are more likely to have problems with memory loss when reaching an elderly age. This makes sense as smoking can lead to many other health conditions which can affect one's neurological processes, and dementia is a neurological disease.

Regarding the analysis from figures 5 and 6, which focuses on whether pre-existing conditions affects one's chances of having Alzheimer's, the fact that dementia affects both memory and decision-making means that certain health conditions (specifically, those with skin cancer) which had different results when comparing the memory loss graphs to the decision making/remembering graphs cannot be considered as potential risk factors for Alzheimer's. Those who actively have diabetes or has diabetes during pregnancy remained at consistently higher chances for having memory loss problems can also be considered as potential factors, but those with pre-diabetes cannot be considered due to not having this higher percentage for memory loss problems, and only for decision making problems.

Figures 5 and 6 show two specific states that were repeated as having a higher percentage if the elderly who are struggling with memory loss and decision making/remembering than most other states represented in the survey – Arkansas and Kentucky. Both of these states are close to each other, so this geographical region may be an area, due to several different factors not shown in this data, that people will have a higher chance of developing forms of dementia in.

Additional Questions and Closing Thoughts

As most research projects end up producing more questions than answers, I thought of several more questions stemming from my visualizations:

-What other habits, such as long-term dietary habits or more specific questions related to exercise, can be asked as survey questions on the 2022 BRFSS to dive further into the relationship between habits and memory loss and decision making/remembering?

-How do these different factors that were analyzed over my visualizations compare when only looking at those who are 75 and older (73% of those with Alzheimer's are 75 and older) (1)?

-What is the best way to gather information about those with Alzheimer’s, as most people in the advanced stages of the disease are unable to speak or answer questions?

-How does this data from the United States compare to other countries?

Something to note from this information is that even with these findings, we cannot definitively connect any of these links to Alzheimer’s or other dementias due to this specific question on the survey never being asked. These findings, however, point to researchers the importance of gathering this information, for public data on Alzheimer's is hard to come by and even harder to interpret. Diseases involving types of dementia are incredibly common yet so little is known about them due to the lack of extensive research on the subject. My hope is that my findings produce a call-to-action from data scientists, researchers and doctors around the world to find more patterns and perform more research on Alzheimer's, perhaps even looking into these newly formulated questions, before the deadly and cruel disease takes them too.

References

Alzheimer's Association. (2021). *Facts and Figures*. Alzheimer’s Disease and Dementia. https://www.alz.org/alzheimers-dementia/facts-figures

CDC - 2019 BRFSS Survey Data and Documentation. (2020, August 31). Www.cdc.gov. https://www.cdc.gov/brfss/annual_data/annual_2019.html

CDC - 2020 BRFSS Survey Data and Documentation. (2021, August 27). Www.cdc.gov. https://www.cdc.gov/brfss/annual_data/annual_2020.html

CDC - 2021 BRFSS Survey Data and Documentation. (2022, August 30). Www.cdc.gov. https://www.cdc.gov/brfss/annual_data/annual_2021.html

CDC - About BRFSS. (2020). Behavioral Risk Factor Surveillance System. https://www.cdc.gov/brfss/about/index.html

Population. (2022, June 3). Data.ers.usda.gov. https://data.ers.usda.gov/reports.aspx?ID=17827