DSC 520 Final - Alzheimer's Disease

Alex Hamedaninia

2023-08-10

Introduction

For this research paper, I am interested in exploring the complex disease Alzheimer's. As defined by the National Institute on Aging, Alzheimer's disease is "a brain disorder that slowly destroys memory and thinking skills and, eventually, the ability to carry out the simplest of tasks. In most people with the disease – those with the late-onset type symptoms first appear in their mid-60s." One of the hardest parts of Alzheimer's is those affected are not always aware of their symptoms. In this paper, we will be analyzing the data to determine what characteristics, including demographic qualities, symptoms, etc. may correlate with Alzheimer's. Do certain qualities correlate with Alzheimer's disease?

Personally, I am interested in this topic as this disease runs heavy in my family, and while genetics play a big role in this disease, I would like to see if there's any other factors that may indicate a higher risk of disease. For others who know Alzheimer's runs in their family, this may be important information for them as well.

Alzheimer's research involves gathering data from those affected by the disease and tracking all the possible factors that could have influenced the disease, such as weight, age, smoking, vaccines, overall health, etc. Data science techniques, such as linear regression and predictive modeling, can use these different factors to identify individuals at risk of developing Alzheimer's or predict disease progression in diagnosed patients. By analyzing the health data of those affected by Alzheimer's, we can use this knowledge to detect subtle clinical markers that may be associated with the disease.

Problem Statement

Can we identify any other qualities, such as demographic qualities (age, race, gender, etc.) or symptoms that correlate with the onset of Alzheimer's?

How we will address the problem

Using the data sets, I will first clean the data to ensure quality input. Then I will want to visualize the data, using different plots to determine if there are any correlations we can find between the different variables provided in the data sets. If there are any correlations present, I will want to dive into them further, and perhaps build a model and see if I can accurately predict certain factors contributing to the onset of Alzheimer's. More than anything, I want to find any possible correlations between certain symptom and demographic factors that may correlate to the onset of Alzheimer's.

One of the scarier aspects of Alzheimer's is that the person affected will have a harder time recognizing their symptoms, and it is usually the loved ones surrounding the individual who pushes them to get tested and eventually diagnosed.If we can identify specific factors that exhibit a correlation with the onset of Alzheimer's, similar to the hereditary nature of the disease, we can be proactive in addressing individuals possessing such factors, enabling them to monitor their condition attentively and potentially mitigate the onset of Alzheimer's.

Datasets

1. Alzheimer's Disease and Healthy Aging Indicators: Cognitive Decline from data.gov. The original purpose of this data also provides information pertaining to individual's diagnosed with Alzheimer's, including their age, geographical location, race, year diagnosed, and gender, with 37 columns and 21,015 rows/cases. This dataset contains data from the BRFSS 2015-2020 dataset. This dataset contains similar issues as the first dataset provided, with a unexplained 'Q##' columns with empty response columns following, but the race variables are much more clear, as well as the location variables. There are a few cases of the data being imputed, such as a '~' marking an empty value, but otherwise fields are left blank.
2. Diagnosis and Symptoms Checklist [ADNI1,GO,2] provided by the Alzheimer's Disease Neuroimaging Initiative (ADNI)[https://adni.loni.usc.edu/data-samples/adni-data-inventory/). The original purpose of this study is to discover and to document the symptoms of Alzheimer's Disease to allow for advancement in treatment to be made. From the ADNI study, we will be working with their Diagnonsis and Symptoms Checklist [ADNI1,GO,2] data set, which contains data from 2006 and variables indicating if patients had various symptoms associated with Alzheimer's. There are 39 variables with 4,884 rows. There are 1's and 2's indicating if symptoms were present, 2 being true and 1 being false. There are long questions/responses in some of the variables that will need to be categorized as well. Any missing values are left as blank or filled with NA.
3. Alzheimer Disease and Healthy Aging Data In US from (Kaggle)[https://www.kaggle.com/datasets/ananthu19/alzheimer-disease-and-healthy-aging-data-in-us]. The original purpose is to provide information about Alzheimer's disease, including prevelance, incidence, risk factors, and outcomes. The data is to be used to explore patterns and explore potential factors and interventions to potentially delay the onset of Alzheimer's. It contains data from 2015-2020 from the Center for Disease Control (CDC), BRFSS, National Health and Nutrition Examination Survey (NHANES), and National Health Interview Survey (NHIS). There are 29 variables. There are no peculiarities in the data set that I've noticed just yet, and it seems the data was imputed by replacing with the mean. I have not seen any missing values.

During this study, we will be utilizing the R packacges ggplot2, dplyr, zoo, and lubridate.

In this study, we will be working with many binary variables, and for these we will use many bar plots to determine the most significant cases. We will also utilize scatter plots to determine any correlations.

Cleaning the data

To import the data, visit the webpages listed above (data sets from the ADNI study will require permission), and obtain the csv files for the relevant files. We will import them below.

```
adni_age_df <- read.csv('dsc520_final/ADNI_Participant_Age_Distribution_08-04-2023.csv')
symptoms_df <- read.csv('dsc520_final/ADSXLIST_04Aug2023.csv')
gov_alz_decline_df <- read.csv('dsc520_final/Alzheimer_s_Disease_and_Healthy_Aging_Indicators__Cognitive
```

Beginning with our symptoms_df, this dataframe gives us all the symptoms that patients were experiencing, marked with 2 for True and 1 for False. I plan to create a new, updated dataframe for symptoms, keeping the variables that are most useful for this analysis. This includes keeping ID, EXAMDATE, and all the symptoms, converting these all to true or false. For missing values, we will examine the data to see how many values are missing, but we will most likely omit them from our analysis, depending on how frequent they are.

Let's begin cleaning up the symptoms dataframe, utilizing various functions from the dplyr package.

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
```

```
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
symptoms_df.2 <- symptoms_df %>% select(-c('Phase', 'RID', 'SITEID','VISCODE','VISCODE2', 'USERDATE', 'U

# let's convert the examdate column from character values to dates using the lubridate package
library(lubridate)
```

```
##
## Attaching package: 'lubridate'
```

```
## The following objects are masked from 'package:base':
##
##     date, intersect, setdiff, union
```

```
symptoms_df.2$EXAMDATE <- ymd(symptoms_df.2$EXAMDATE)

# now to convert the symptomatic variables to true/false. 1 = False, 2 = True
symptoms_df.3 <- symptoms_df.2 %>% mutate(across(starts_with('AX'), ~ as.logical(.x - 1)))

# checking for missing values
colSums(is.na(symptoms_df.3))
```

```
##       ID EXAMDATE AXNAUSEA  AXVOMIT AXDIARRH AXCONSTP AXABDOMN AXSWEATN
##        0      683        0        0        0        0        1        0
##  AXDIZZY AXENERGY AXDROWSY AXVISION AXHDACHE AXDRYMTH AXBREATH  AXCOUGH
##        0        0        0        0        0        0        0        0
## AXPALPIT  AXCHEST AXURNDIS AXURNFRQ  AXANKLE AXMUSCLE   AXRASH AXINSOMN
##        0        0        0        0        0        0        0        0
## AXDPMOOD AXCRYING AXELMOOD AXWANDER   AXFALL
##        0        0        0        0        0
```

```
# missing 683 exam date entries.
# We can remedy this by performing linear interpolation to fill in the missing dates based on neighbori
library(zoo)
```

```
##
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
symptoms_df.3 <- symptoms_df.3 %>% arrange(EXAMDATE) %>% mutate(imputed_date = ifelse(is.na(EXAMDATE), i

# convert imputed_date back to actual dates
symptoms_df.3$imputed_date <- as.Date(symptoms_df.3$imputed_date, origin = "1970-01-01")

#now to check for missing values again
colSums(is.na(symptoms_df.3))
```

```
##           ID     EXAMDATE     AXNAUSEA      AXVOMIT     AXDIARRH     AXCONSTP
##            0          683            0            0            0            0
##     AXABDOMN     AXSWEATN      AXDIZZY     AXENERGY     AXDROWSY     AXVISION
```

```
##           1           0           0           0           0           0
##  AXHDACHE  AXDRYMTH  AXBREATH   AXCOUGH  AXPALPIT   AXCHEST
##           0           0           0           0           0           0
##  AXURNDIS  AXURNFRQ   AXANKLE  AXMUSCLE   AXRASH  AXINSOMN
##           0           0           0           0           0           0
##  AXDPMOOD  AXCRYING  AXELMOOD  AXWANDER   AXFALL imputed_date
##           0           0           0           0           0           0
```

```
# shows no more missing values in our imputed_date column. There is 1 missing value in AXABDOMN, howeve

# now let's move imputed_date up to the front for visualizing purposes, and this dataframe will be read
symptoms_df.3 <- symptoms_df.3 %>% select(ID, imputed_date, everything())
```

Now that our symptoms dataframe is cleaned up, we can begin to clean our Alzheimer's Disease and Healthy Aging Indicators: Cognitive Decline dataframe. Let's take a look at it.

```
head(gov_alz_decline_df)
```

```
##   YearStart YearEnd LocationAbbr         LocationDesc Datasource
## 1      2016    2021           ND         North Dakota      BRFSS
## 2      2019    2019           DC District of Columbia      BRFSS
## 3      2016    2021           FL              Florida      BRFSS
## 4      2021    2021           WI            Wisconsin      BRFSS
## 5      2020    2020           VT              Vermont      BRFSS
## 6      2020    2020           PR          Puerto Rico      BRFSS
##              Class
## 1 Cognitive Decline
## 2 Cognitive Decline
## 3 Cognitive Decline
## 4 Cognitive Decline
## 5 Cognitive Decline
## 6 Cognitive Decline
##                                                                                                  T
## 1 Functional difficulties associated with subjective cognitive decline or memory loss among older adu
## 2                                        Subjective cognitive decline or memory loss among older adu
## 3                Talked with health care professional about subjective cognitive decline or memory l
## 4 Functional difficulties associated with subjective cognitive decline or memory loss among older adu
## 5     Need assistance with day-to-day activities because of subjective cognitive decline or memory l
## 6                Talked with health care professional about subjective cognitive decline or memory l
##
## 1 Percentage of older adults who reported subjective cognitive decline or memory loss that interferes
## 2          Percentage of older adults who reported subjective cognitive decline or memory loss that
## 3                                Percentage of older adults with subjective cognitive decline or mem
## 4 Percentage of older adults who reported subjective cognitive decline or memory loss that interferes
## 5             Percentage of older adults who reported that as a result of subjective cognitive decli
## 6                                Percentage of older adults with subjective cognitive decline or mem
##   Response Data_Value_Unit DataValueTypeID Data_Value_Type Data_Value
## 1       NA               %           PRCTG      Percentage       17.1
## 2       NA               %           PRCTG      Percentage       12.4
## 3       NA               %           PRCTG      Percentage         NA
## 4       NA               %           PRCTG      Percentage         NA
## 5       NA               %           PRCTG      Percentage       24.6
## 6       NA               %           PRCTG      Percentage         NA
##   Data_Value_Alt Data_Value_Footnote_Symbol
## 1           17.1
```

```
## 2           12.4
## 3             NA                       ****
## 4             NA                       ****
## 5           24.6
## 6             NA                       ****
##
## 1
## 2
## 3 Sample size of denominator and/or age group for age-standardization is less than 50 or relative sta
## 4 Sample size of denominator and/or age group for age-standardization is less than 50 or relative sta
## 5
## 6 Sample size of denominator and/or age group for age-standardization is less than 50 or relative sta
##   Low_Confidence_Limit High_Confidence_Limit Sample_Size
## 1                 11.7                  24.2          NA
## 2                  9.2                  16.5          NA
## 3                                               NA
## 4                                               NA
## 5                 17.7                  33.0          NA
## 6                                               NA
##   StratificationCategory1   Stratification1 StratificationCategory2
## 1              Age Group 65 years or older          Race/Ethnicity
## 2              Age Group 65 years or older                  Gender
## 3              Age Group           Overall          Race/Ethnicity
## 4              Age Group           Overall          Race/Ethnicity
## 5              Age Group           Overall          Race/Ethnicity
## 6              Age Group       50-64 years                  Gender
##            Stratification2 StratificationCategory3 Stratification3 ClassID
## 1     White, non-Hispanic                      NA              NA     C06
## 2                  Female                      NA              NA     C06
## 3 Native Am/Alaskan Native                      NA              NA     C06
## 4     Black, non-Hispanic                      NA              NA     C06
## 5     White, non-Hispanic                      NA              NA     C06
## 6                    Male                      NA              NA     C06
##   TopicID QuestionID ResponseID LocationID StratificationCategoryID1
## 1   TCC02        Q31         NA         38                       AGE
## 2   TCC01        Q30         NA         11                       AGE
## 3   TCC04        Q42         NA         12                       AGE
## 4   TCC02        Q31         NA         55                       AGE
## 5   TCC03        Q41         NA         50                       AGE
## 6   TCC04        Q42         NA         72                       AGE
##   StratificationID1 StratificationCategoryID2 StratificationID2
## 1            65PLUS                      RACE              WHT
## 2            65PLUS                    GENDER           FEMALE
## 3       AGE_OVERALL                      RACE              NAA
## 4       AGE_OVERALL                      RACE              BLK
## 5       AGE_OVERALL                      RACE              WHT
## 6              5064                    GENDER             MALE
##   StratificationCategoryID3 StratificationID3 Report
## 1                        NA               NA     NA
## 2                        NA               NA     NA
## 3                        NA               NA     NA
## 4                        NA               NA     NA
## 5                        NA               NA     NA
## 6                        NA               NA     NA
```

We can clean up this dataframe in quite a few different ways. 1. First, I will filter the columns to only have variables we need. This will be any variables that tell us information relevant to the patient, and excludes variables that are either repetitive, or only relevant to describing the data (e.g. Datasource, Class) 2. I will remove any empty values in the Data_Value column as this is the only response value we have in this dataset. In this case, removing NA values will not affect any other aspect of our data analysis. 3. I will separate the StratificationCategory2 variable into Race/Ethnicity and Gender, and combine the category of age group and the corresponding age group it belongs with.

```
# first let's condense this dataframe to only variables that we need.
gov_alz_decline_df.2 <- gov_alz_decline_df %>% select(-c('Sample_Size', 'StratificationCategory3', 'Stra

# since we only have one question and response in this dataset, we will remove any cases with NA as it
gov_alz_decline_df.2 <- gov_alz_decline_df.2 %>% filter(!is.na(Data_Value))

gov_alz_decline_df.3 <- gov_alz_decline_df.2

# let's separate StratificationCategory2 into Race/Ethnicity and Gender. I will also rename some of the
gov_alz_decline_df.3 <- gov_alz_decline_df.3 %>% mutate(gender = ifelse(StratificationCategory2 == 'Gen

# now let's repeat the same for race/ethnicity
gov_alz_decline_df.3 <- gov_alz_decline_df.3 %>% mutate(Race.Ethnicity = ifelse(StratificationCategory2

# checking to see if there's distinct values in DataValueTypeID
gov_alz_decline_df.3 %>% distinct(DataValueTypeID)
```

```
##   DataValueTypeID
## 1           PRCTG
```

```
# it is only PRCTG, so we can remove this variable since it doesn't tell us any useful information for

# now we can remove StratificationCategory2 and its response
gov_alz_decline_df.3 <- gov_alz_decline_df.3 %>% select(-c('StratificationCategory2', 'Stratification2'

# now let's rename Stratifcation1 to Age Group and remove the category for it, and our dataset will be
gov_alz_decline_df.3 <- gov_alz_decline_df.3 %>% rename('Age_Group' = 'Stratification1')
gov_alz_decline_df.3 <- gov_alz_decline_df.3 %>% select(-c('StratificationCategory1'))
# ta da!
gov_alz_decline_df.3 <- gov_alz_decline_df.3 %>% select(-c('Class', 'Datasource', 'Response', 'Data_Valu
```

Finally, our adni_age_df dataset only has two columns, and doesn't need much cleanup. Thus, we can leave it as is.

The final data sets

Our two final data sets we will work with are symptoms_df.3 that show us the various symptoms people with Alzheimer's experience, gov_alz_decline_df.3 shows us the demographics as well as a response to different questions, and adni_age_df which shows us the various age groups afflicted with Alzheimer's. Here they are below:

```
head(symptoms_df.3)
```

```
##     ID imputed_date   EXAMDATE AXNAUSEA AXVOMIT AXDIARRH AXCONSTP AXABDOMN
## 1 5844   2006-02-06 2006-02-06    FALSE   FALSE    FALSE    FALSE    FALSE
## 2    2   2006-03-06 2006-03-06    FALSE   FALSE    FALSE    FALSE    FALSE
## 3    4   2006-03-09 2006-03-09    FALSE   FALSE    FALSE    FALSE    FALSE
## 4    6   2006-03-13 2006-03-13    FALSE   FALSE    FALSE    FALSE    FALSE
```

```
## 5    8   2006-03-20 2006-03-20    FALSE   FALSE    FALSE     TRUE    FALSE
## 6   14   2006-04-12 2006-04-12    FALSE   FALSE    FALSE    FALSE    FALSE
##   AXSWEATN AXDIZZY AXENERGY AXDROWSY AXVISION AXHDACHE AXDRYMTH AXBREATH
## 1    FALSE   FALSE    FALSE    FALSE    FALSE    FALSE    FALSE    FALSE
## 2    FALSE    TRUE    FALSE    FALSE    FALSE    FALSE    FALSE    FALSE
## 3    FALSE   FALSE    FALSE    FALSE    FALSE     TRUE    FALSE    FALSE
## 4    FALSE   FALSE    FALSE    FALSE    FALSE    FALSE    FALSE    FALSE
## 5    FALSE   FALSE    FALSE    FALSE    FALSE    FALSE    FALSE    FALSE
## 6    FALSE   FALSE    FALSE    FALSE     TRUE    FALSE     TRUE     TRUE
##   AXCOUGH AXPALPIT AXCHEST AXURNDIS AXURNFRQ AXANKLE AXMUSCLE AXRASH AXINSOMN
## 1   FALSE    FALSE   FALSE    FALSE    FALSE   FALSE    FALSE  FALSE    FALSE
## 2   FALSE    FALSE   FALSE    FALSE    FALSE   FALSE    FALSE  FALSE    FALSE
## 3    TRUE    FALSE   FALSE    FALSE    FALSE   FALSE    FALSE   TRUE    FALSE
## 4   FALSE     TRUE   FALSE    FALSE    FALSE   FALSE     TRUE  FALSE     TRUE
## 5   FALSE    FALSE   FALSE    FALSE    FALSE   FALSE    FALSE  FALSE    FALSE
## 6    TRUE    FALSE   FALSE    FALSE    FALSE   FALSE    FALSE  FALSE    FALSE
##   AXDPMOOD AXCRYING AXELMOOD AXWANDER AXFALL
## 1    FALSE    FALSE    FALSE    FALSE  FALSE
## 2    FALSE    FALSE    FALSE    FALSE  FALSE
## 3    FALSE    FALSE    FALSE    FALSE  FALSE
## 4    FALSE    FALSE    FALSE    FALSE  FALSE
## 5    FALSE    FALSE    FALSE    FALSE   TRUE
## 6    FALSE    FALSE     TRUE    FALSE  FALSE
```

```r
head(gov_alz_decline_df.3)
```

```
##   YearStart YearEnd LocationAbbr         LocationDesc
## 1      2016    2021           ND         North Dakota
## 2      2019    2019           DC District of Columbia
## 3      2020    2020           VT              Vermont
## 4      2020    2020         WEST                 West
## 5      2015    2015           GA              Georgia
## 6      2016    2016           CT          Connecticut
##                                                                                                          To
## 1 Functional difficulties associated with subjective cognitive decline or memory loss among older adu
## 2                                            Subjective cognitive decline or memory loss among older adu
## 3     Need assistance with day-to-day activities because of subjective cognitive decline or memory l
## 4                                            Subjective cognitive decline or memory loss among older adu
## 5 Functional difficulties associated with subjective cognitive decline or memory loss among older adu
## 6 Functional difficulties associated with subjective cognitive decline or memory loss among older adu
##
## 1 Percentage of older adults who reported subjective cognitive decline or memory loss that interferes
## 2           Percentage of older adults who reported subjective cognitive decline or memory loss that
## 3              Percentage of older adults who reported that as a result of subjective cognitive decli
## 4           Percentage of older adults who reported subjective cognitive decline or memory loss that
## 5 Percentage of older adults who reported subjective cognitive decline or memory loss that interferes
## 6 Percentage of older adults who reported subjective cognitive decline or memory loss that interferes
##   Data_Value Low_Confidence_Limit High_Confidence_Limit         Age_Group
## 1       17.1                 11.7                  24.2 65 years or older
## 2       12.4                  9.2                  16.5 65 years or older
## 3       24.6                 17.7                  33.0           Overall
## 4        1.2                  0.6                   2.1       50-64 years
## 5       41.2                 34.9                  47.8           Overall
## 6       30.6                 22.7                  39.8           Overall
##   gender         Race.Ethnicity
```

```
## 1     NA    White, non-Hispanic
## 2 Female                     NA
## 3     NA    White, non-Hispanic
## 4     NA Asian/Pacific Islander
## 5     NA                     NA
## 6     NA    White, non-Hispanic
```

```
adni_age_df
```

```
##   Age.Group Number.of.Subjects
## 1     40-49                 2
## 2     50-59               131
## 3     60-69               816
## 4     70-79              1295
## 5     80-89               472
## 6  Above 89                29
## 7   Unknown                 5
```

What are different ways to look at the data?

There are quite a few different ways to look at the data. From the symptoms_df.3 dataset, there are several variables that could be grouped together, such as AXDROWSY and AXENERGY, as well as AXNAUSEA and AXVOMIT, since most of these symptoms coincide with each other. Some of the symptoms may be from other illnesses that are unrelated, but our data analysis should tell us if a symptom is not significant. As for the gov_alz_decline dataset, we have a few different ways to look at the dataset as well. As we only have one data value percentage as a response for the question provided, which are different questions, we can dissect this data in different ways, as well as interpret the percentages in different ways.

Plan to slice and dice data

We have sliced and diced the data to a certain point now in gov_alz_decline_df, creating new variables by removing the StratificationCategory1 and renaming Stratification1 label to 'Age Group'. Similarly, we used StratificationCategory2 and Stratification2 to create two new variables, Race.Ethnicity and Gender, allowing us to more easily identify these two variables. We also removed StratificationCategory3 and Stratification3 as these values were mostly NA and of no use to our analysis. We could further separate our variables, such as Topic and Question into the different questions that are listed, to allow us to more clearly differentiate between what is being asked and the response given. We could also remove either to LocationAbbr or LocationDesc, as these are just repeated values and we only need one. I'd mostly likely retain LocationDesc and it is more self-evident. There are quite a few unknown variables after splitting the data like this, and it is unapparent of how to deal with these NAs.

As for symptoms_df, we removed all unnecessary variables. We also imputed the data column to present as a date value and not as a character value. As for slicing and dicing the data, we could rename the variable names to be more intuitive. However, there does not seem to be much further slicing and dicing necessary for this data set. However, there are considerations for trying to combine the two datasets into one, though that may be a bit difficult because of the difficult numbers of rows.

Summarizing data to answer key questions

```
symptoms_df.3 %>% summarize_all(mean)
```

```
##        ID imputed_date EXAMDATE   AXNAUSEA    AXVOMIT  AXDIARRH  AXCONSTP
## 1 3713.16   2008-03-28     <NA> 0.03009828 0.01556102 0.1009419 0.1119984
##   AXABDOMN  AXSWEATN   AXDIZZY  AXENERGY  AXDROWSY   AXVISION   AXHDACHE
## 1       NA 0.0542588 0.1269451 0.2223587 0.1470106 0.06101556 0.06572482
```

```
##    AXDRYMTH   AXBREATH   AXCOUGH   AXPALPIT    AXCHEST   AXURNDIS AXURNFRQ
## 1 0.1064701 0.08660934 0.1191646 0.03153153 0.03091728 0.02313677 0.245905
##      AXANKLE  AXMUSCLE     AXRASH  AXINSOMN  AXDPMOOD   AXCRYING   AXELMOOD
## 1 0.09602785 0.3710074 0.07575758 0.1287879 0.1470106 0.04013104 0.01494676
##     AXWANDER     AXFALL
## 1 0.01781327 0.07903358
```

Considering that 0 is False and 1 is true, we want to look for values that are closer to 1 to indicate that there a certain symptom is more present than others. Energy at 0.22, Urinating Frequency at 0.25, muscle pain at 0.37 indicates these are variables to pay extra attention to. Patients are experiencing these symptoms more often than other ones. We can take this a step further and see if they have any correlation to the dates and see if they show up earlier than others. If so, we'd be on our way of answering our question of what indicates the onset of Alzheimer's.

We can also summarize our gov_alz_decline dataframe as well.

```
gov_alz_decline_df.3 %>% summarize(mean(Data_Value))
```

```
##   mean(Data_Value)
## 1         31.40467
```

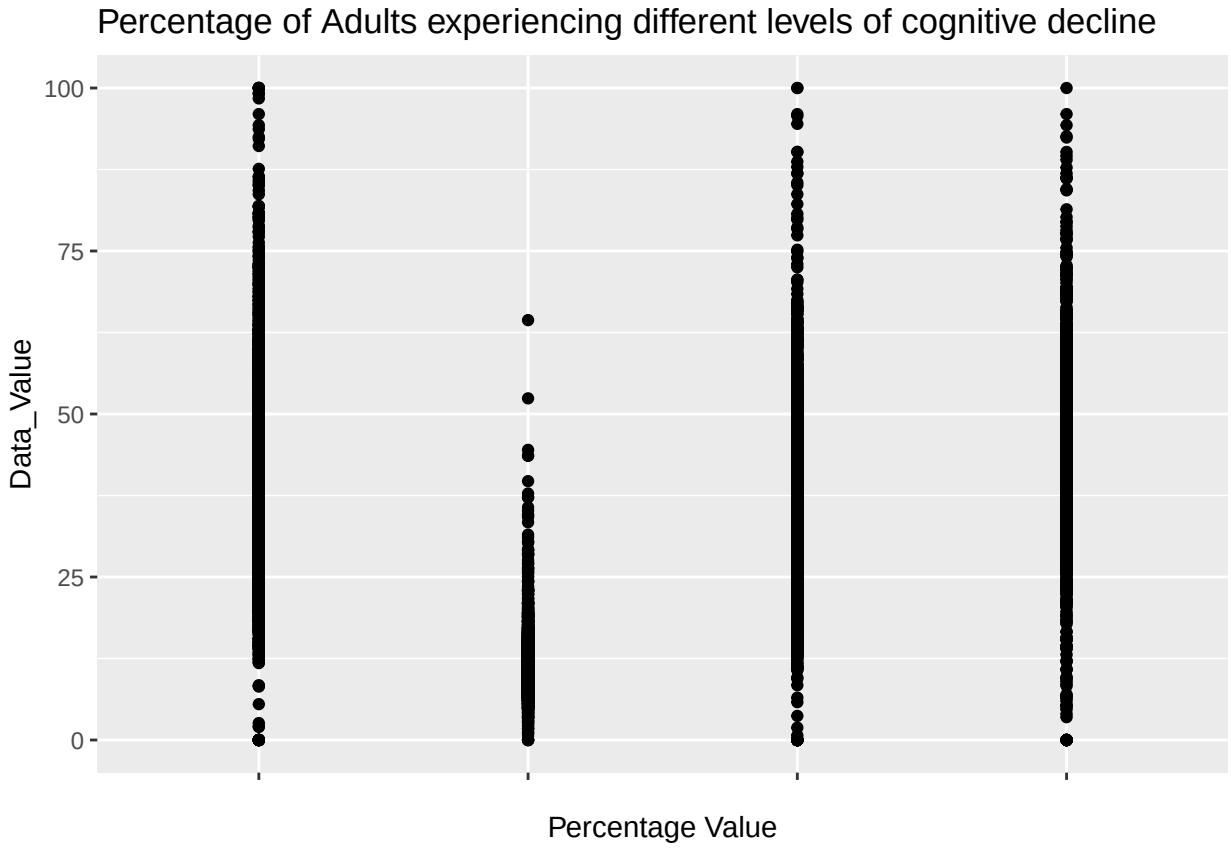
This value is the mean of the different responses to the questions under the Question variable, which show that slicing the data into the two different questions will be useful in learning more about the data. However, these findings show that 31% of the cases experienced some decline in cognitive function.

As for the adni_age_df, when we take a look at the values, we can see an strong upwards trend as the age groups increase. This tells us that Alzheimer's affects higher age groups.

Analysis

Let's utilize some bar plots for the symptoms dataframe since most of our variables are binary.

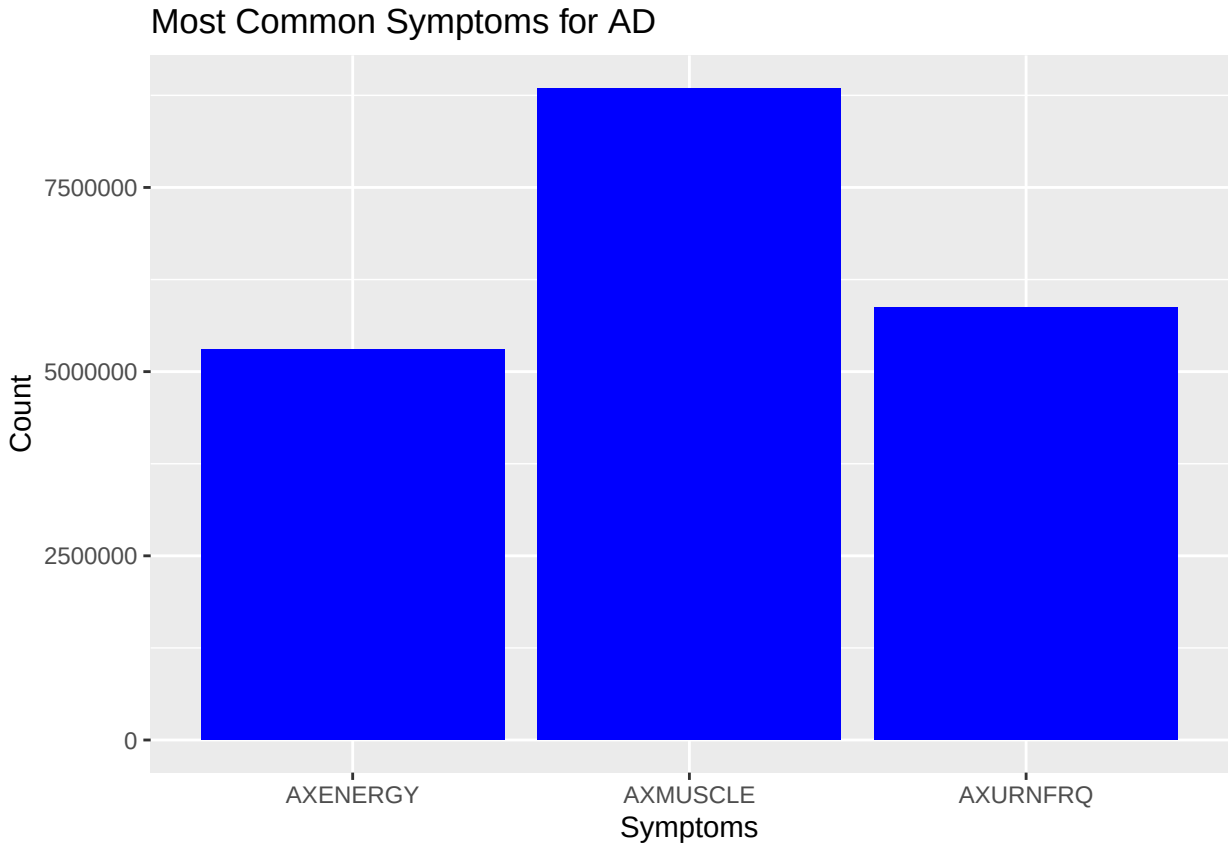
```
library(ggplot2)
ggplot(data=gov_alz_decline_df.3, aes(x=Question, y=Data_Value)) +  geom_point(position= position_dodge
```



1. "Percentage of older adults who reported subjective cognitive decline or memory loss that interferes with their ability to engage in social activities or household chores" 2. Percentage of older adults who reported subjective cognitive decline or memory loss that is happening more often or is getting worse in the preceding 12 months" 3. "Percentage of older adults who reported that as a result of subjective cognitive decline or memory loss that they need assistance with day-to-day activities" 4. "Percentage of older adults with subjective cognitive decline or memory loss who reported talking with a health care professional about it"

We can see that for most subjects in this study, there is a very high percentage of adults who find that Alzheimer's affects their day-to-day activities such as being social or household chores, and find that they need assistance and talk with their health care professional about it. For the lack of response to question 2, this may be because when one is affected by Alzheimer's they are not always aware. It is very difficult to be aware of cognitive decline, and those affected by this disease usually rely on others help them through this time.

```
# let's create a bar graph comparing the different symptoms we found to be more prevalent: Energy, Urin
# first let's convert the logical factors (TRUE/FALSE) to numeric (1/0)
binary_vars <- c('AXENERGY','AXURNFRQ','AXMUSCLE')
binary_df <- symptoms_df.3[, c(binary_vars)]
binary_df$AXENERGY <- sum(as.integer(as.logical(binary_df$AXENERGY)))
binary_df$AXURNFRQ <- sum(as.integer(as.logical(binary_df$AXURNFRQ)))
binary_df$AXMUSCLE <- sum(as.integer(as.logical(binary_df$AXMUSCLE)))

data_long <- tidyr::gather(binary_df, variable, value)

ggplot(data_long, aes(x = variable, y = value)) + geom_bar(stat = 'identity', fill = 'blue') + labs(x =
```

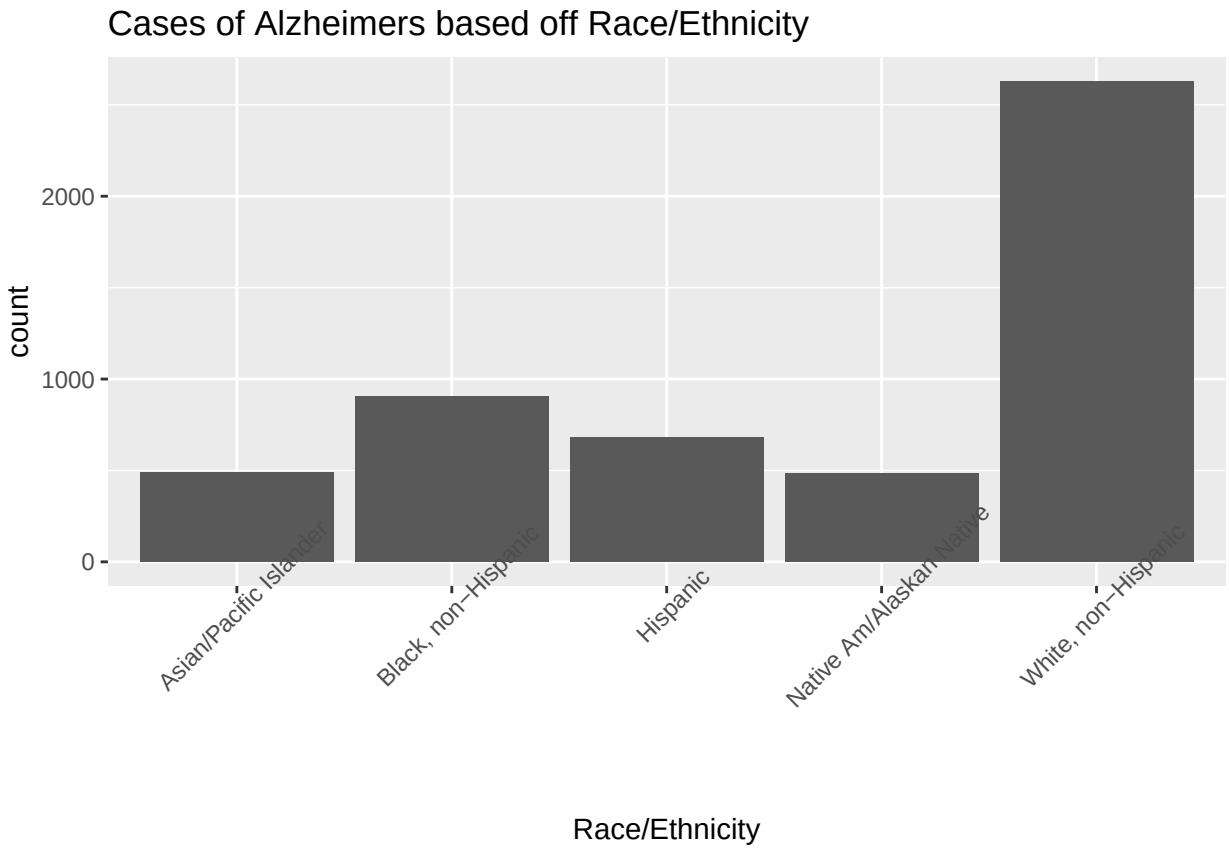


We can see this confirms that muscle pain is the number one symptom experienced by those with Alzheimer's, followed by frequent urination closely followed by a lack of energy. This is an interesting observation that requires further investigation – is muscle pain correlated with those experiencing Alzheimer's? Or is it due to the fact that those with Alzheimer's are older, and just likely to experience muscle pain with or without the disease. To further investigate this, we could compare folks who do have Alzheimer's compared to those who do not, and measure how much muscle pain each one experiences. If there is a positive correlation, we could say that muscle pain may be an indicator for Alzheimer's. However, we may find that the result is also statistically insignificant if muscle pain is just a symptom of old age.

Let's also create a scatterplot comparing the race/ethnicity and gender of those affected with Alzheimer's.

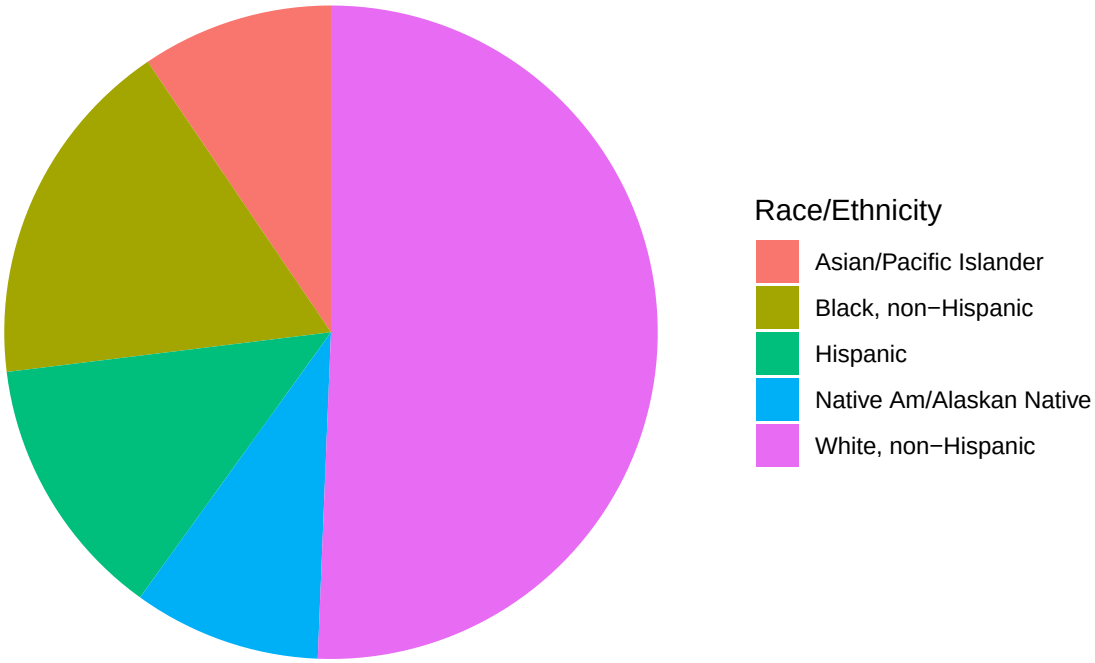
```
# filter out the NA column for races for plotting
filtered_df <- gov_alz_decline_df.3 %>% filter(!Race.Ethnicity == 'NA')

ggplot(filtered_df, aes(x=Race.Ethnicity)) + geom_histogram(stat='count') + theme(axis.text.x = element_
```

```
## Warning in geom_histogram(stat = "count"): Ignoring unknown parameters:
## `binwidth`, `bins`, and `pad`
```



Interesting to see that the high majority of those affected by Alzheimer's in this study are of white, non-hispanic descent. Is this due to a sample that doesn't represent the true population, or is there a high correlation between Alzheimer's and being white? We can see how much of the population the white demographic takes up by looking at a pie chart.

```
ggplot(filtered_df, aes(x = factor(1), fill = Race.Ethnicity)) + geom_bar(stat='count', width = 1) +
  coord_polar(theta = 'y') +
  labs(fill = "Race/Ethnicity", title = "Cases of Alzheimer's based on Race/Ethnicity") + theme_void() +
```

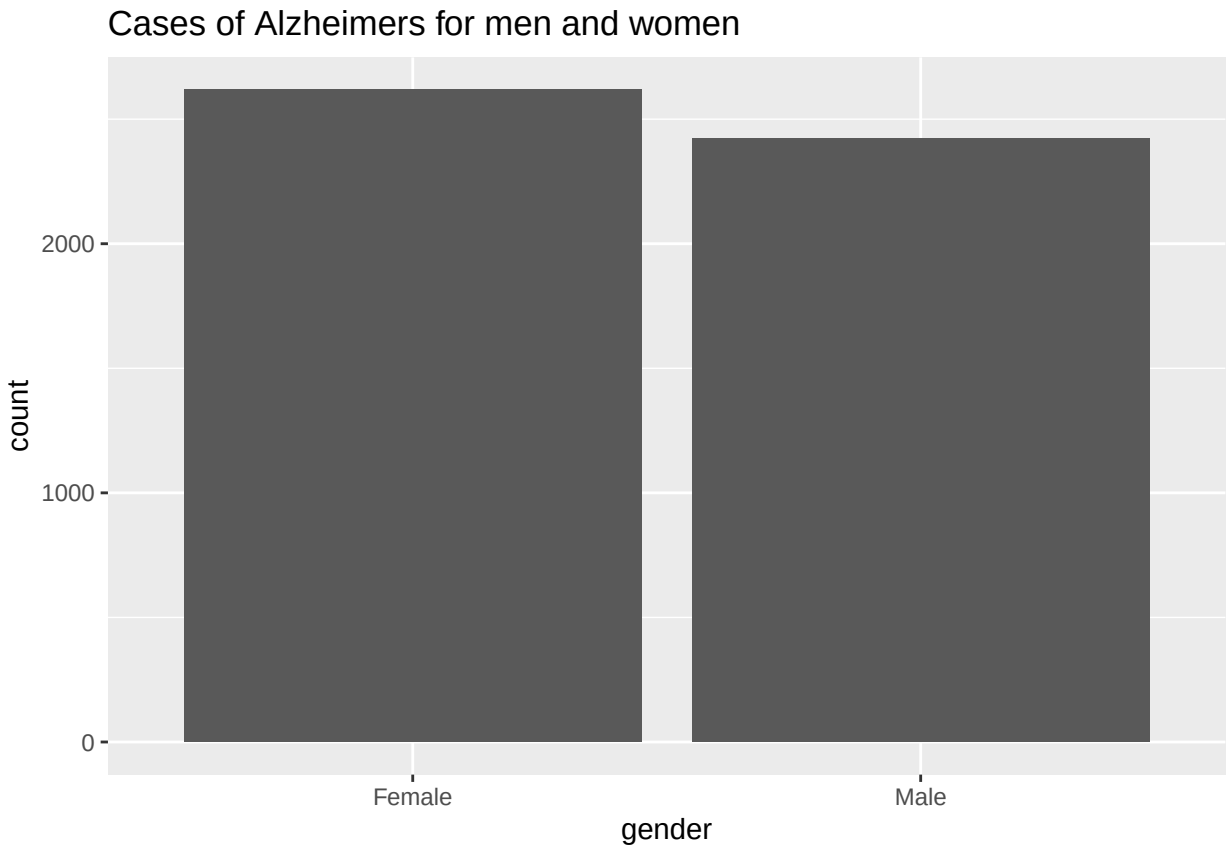


We can see that the white population makes up half of the total population of those afflicted by Alzheimer's. Is this due to the nature of the sample taken, or are white folks much more prone to this disease? To further investigate this, we could compare to another dataset to determine if this pattern continues.

Let's also compare the rate of Alzheimer's between males and females

```
# filter out the NA column for races for plotting
filtered_df.2 <- gov_alz_decline_df.3 %>% filter(!gender == 'NA')

ggplot(filtered_df.2, aes(x=gender)) + geom_histogram(stat='count') + labs(title = 'Cases of Alzheimers
```

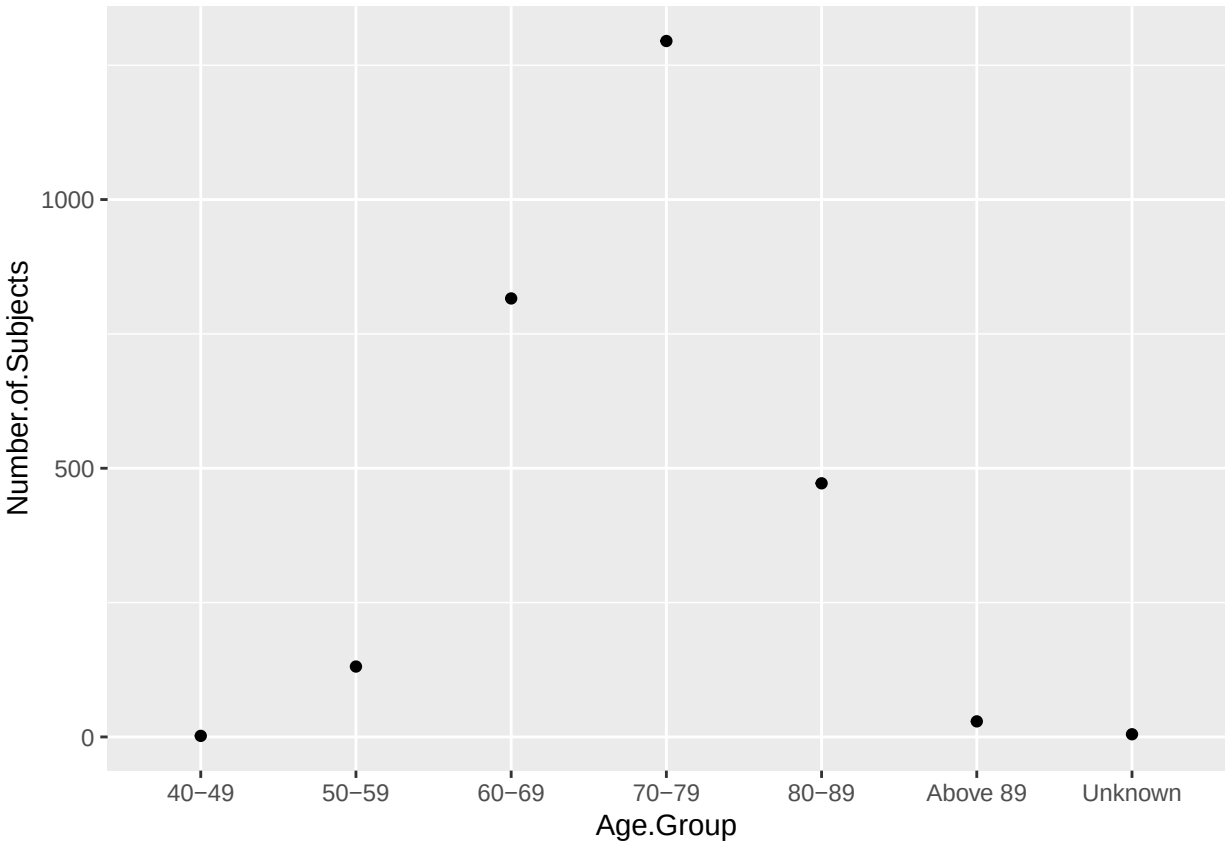
```
## Warning in geom_histogram(stat = "count"): Ignoring unknown parameters:
## `binwidth`, `bins`, and `pad`
```



These two are nearly equal, but personally I am surprised to see that women are slightly more affected by men – I thought it would be the other way around.

Now let's plot to see the different age groups affected by Alzheimer's.

```
ggplot(adni_age_df, aes(x = Age.Group, y = Number.of.Subjects)) + geom_point()
```



There is a distinct trend that as the age group peaks between 70-79, there are the most cases of Alzheimer's. Sadly, we can only assume it decreases from there as the patients pass away.

Implications

We can see there are a few different implications we have learned while analyzing our data. Our original problem statement was if we can identify any other qualities, such as demographic qualities (age, race, gender, etc.) or symptoms that correlate with the onset of Alzheimer's?

In an attempt to answer this question, we obtained data from the Alzheimer's Disease Neuroimaging Initiative (ADNI) and the government database of Alzheimer's Disease and Healthy Aging Indicators: Cognitive Decline. These data sets contained important demographic and symptomatic information that we used to find patterns and insights into Alzheimer's disease. We did this by making some plots to see if any variables stood out, indicating further research to be done.

For the symptoms dataframe we created, a logistic regression model would work very well to predict any symptoms that future patients may experience, and may be able to predict which symptoms are co-morbid, or are experienced together. As for the government dataframe that contained many different demographic features, a random forest model may prove effective as it will be able to handle the different features and non-linear relationships between the variables.

Let's take a look at the different insights we gained from our analysis of the plots we made.

Muscle Pain

We have seen that muscle pain seems to be the most reported symptom of those afflicted with Alzheimer's. This can indicate that there may be a correlation with Alzheimer's and muscle pain, however further investigation must be conducted to determine if this observation is significant to Alzheimer's, or may be

due to old age. We could investigate this further if we had a data set that measured the amount of back pain experienced by those afflicted by Alzheimer's, and those not afflicted, with age as a variable. So while we cannot say that muscle pain is an indicator of Alzheimer's, we can say many patients with Alzheimer's experience back pain and we can investigate this further.

Demographic information

From our graphs, we saw that the white demographic had the strongest majority of those afflicted by Alzheimer's. In the pie chart, we saw the white population takes up just more than half of the entire sample population. Before we can say that this indicates the white demographic is much more likely to have Alzheimer's, I would compare this data set to another one to determine if this pattern continues, because it is a very large disproportionate amount of white folks in this study.

We also saw that between males and females, there were slightly more females, indicating there may be a stronger trend for women to be afflicted by Alzheimer's. Again, further investigation is required before we can claim this statement.

Age

From our last plot, we saw that the rise in Alzheimer's starts between 50-59, with the peak being between 60-69. This can imply that in most cases, the official diagnosis may come around 60-69 years of age. However, these symptoms start appearing years earlier, which can be indicated by the 50-59 age group.

Limitations

There are quite a few limitations in this study. We could improve on this study by adding in one more data set that compares symptoms experienced by those with Alzheimer's and those without the disease. This would allow us to create a predictive model that could perhaps predict a reasonable probability of someone with certain features/characteristics becoming afflicted with Alzheimer's. There are also different graphs that could be made by someone with higher expertise skill.

This study could also go further by creating a logistic regression model for the symptoms_df data set. This model could estimate the relationship between different variables, providing insight as to whether certain symptoms come together and perhaps how this could impact the quality of life of those afflicted by Alzheimer's.

A random forest model could also be created for the cognitive decline dataframe (gov_alz_decline_df) for all the different features and non-linear relationships between the variables. This could use predictive modeling to determine if a person comes from a certain demographic, what is the likelihood they may have Alzheimer's? Note: This kind of study would require more information, and even then may not be fully accurate, causing either a lot of unnecessary distress or harmful ignorancy.

Concluding Remarks

The overall implications for someone who may be interested at looking at this study for insight into what factors contribute to the onset of Alzheimer's may come to a disheartening conclusion – what we have discovered here is that it seems the factors that affect Alzheimer's are ones out of our control: race, gender, etc. Just as this disease is hereditary and we cannot control our genes, we cannot control the aspects of what the insights this study has shown us.

Sources

U.S. Department of Health and Human Services. What is alzheimer's disease? National Institute on Aging. ht
tps://www.nia.nih.gov/health/what-alzheimers-disease#:~:text=Alzheimer's%20disease%20is%20a%20brain,appear%20in%20