Alzheimer’s Mortality and Socioeconomic Factors

Doha Sidahmed, Halle Hwang, Alex Cao

I. Background and Motivation

Alzheimer’s disease is the most common cause of dementia. It is a neurodegenerative disorder that leads to memory loss, cognitive decline, and eventually takes away people’s ability to carry out simple tasks. Unfortunately with limitations of current medication and technology, this disease is progressive and irreversible, and ultimately it will lead to death.

Over 6.9 million of American age 65 and older live with Alzheimer’s, while globally over 55 million people’s lives are affected by dementia, with Alzheimer’s being the primary cause. Currently there is no definitive test for Alzheimer’s disease, which makes it hard to diagnose the older adults who might have been experiencing symptoms such as memory loss, confusion and disorientation, difficulty with language and communication, change in mood and behavior, etc. The exact cause of Alzheimer’s disease still remains unknown, however, through observation, many factors could potentially contribute to the complex condition, such as genetics, age, life style, living environment, brain abnormalities, or other health conditions.

Through research, we found articles that explored the role of many social determinants of health in shaping the risk, incidence, and prevalence of Alzheimer’s disease and other related dementia conditions. They highlighted how socioeconomic factors such as income, education, employment, and neighborhood environment influence Alzheimer’s risk. On the social aspect, this motivated us to investigate the relationship between income and education and Alzheimer’s mortality rate. We believe that the results could help policymakers prioritize funding for education and healthcare in high risk populations in the future. In addition, the articles mentioned how living environments and quality of life might also have an impact on the chance of getting Alzheimer’s disease. We consider stress as a potential factor and thus want to investigate whether people have a higher chance of getting Alzheimer’s while under the influence of depression or anxiety.

All the above leads into our research question: **How do socioeconomic factors affect mortality rates of Alzheimer’s disease in the United States, and is there a relationship between Alzheimer’s and Depression?** We will be analyzing data and trying to figure out if there is a correlation between Alzheimer’s rate and the following three factors: **education level, income, and depression/anxiety rate**.

II. **Datasets**

CDC Alzheimer’s Data (https://www.cdc.gov/nchs/pressroom/sosmap/alzheimers_mortality/alzheimers_disease.htm?) : Includes data on Alzheimer’s mortality rates by state for 2022.

FRED State Income Data (https://fred.stlouisfed.org/release/tables?rid=249&eid=259515&od=2022-01-01#) : Median household income by state for 2022.

FRED Bachelor’s by State (https://fred.stlouisfed.org/release/tables?rid=330&eid=391444&od=2022-01-01#) : Bachelor’s degree attainment rates by state for 2022.

CDC Indicators of Anxiety or Depression Data (https://data.cdc.gov/NCHS/Indicators-of-Anxiety-or-Depression-Based-on-Repor/8pt5-q6wp/about_data) : Indicators of anxiety or depression based on reported frequency of symptoms during last 7 days.

The Federal Reserve Economic Data is a database that holds various types of economic data in the United States, such as household income, employment, etc. It is widely used by researchers and policymakers. For our project, we used a dataset they provided that contains the median household income by state in 2022. The contents of this dataset include the name of the state and its median household income, making 50 samples. This data was necessary because it provides us the necessary information to analyze how income levels vary across states and assess their potential relationship with Alzheimer's mortality rates. This database also provided us with a 50-sample dataset that includes the percentage of adults with a bachelor's degree or higher by state in 2022. Education level is an important factor in our analysis to examine whether there is a correlation between education level and Alzheimer's mortality rates

The Centers for Disease Control and Prevention (CDC) provides a dataset containing Alzheimer's disease mortality by state in 2022. The contents of the dataset include the states, death rate, and number of deaths. This dataset provides the necessary information for understanding the geographic distribution of Alzheimer's-related deaths across the United States. By using the death rate and the total number of deaths, we can assess which states experience higher or lower mortality burdens due to the disease.

The CDC also provides a dataset that includes indicators of anxiety or depression based on the reported frequency of symptoms in the last seven days by state. The data was collected through the Household Pulse Survey. The relevant contents of the dataset include the indicator (symptom of anxiety disorder or symptom of depression disorder), the state where it was reported, and the reported percentage/rate of the mental health indicator. This data provides insight into the mental health status of state populations, which can be compared with Alzheimer's mortality rates. By examining the prevalence of mental health disorders, potential correlations between them and Alzheimer's disease mortality can be explored.

III. **Methods**

General data cleaning and processing included merging all the different datasets together by state. To ensure consistency, columns had to be renamed, unnecessary columns were omitted, and small details such as wrong value types were fixed. Additionally, a new column was created for education level. Each state was classified as "low" or "high" for education based on whether the percentage of adults with a bachelor's degree exceeded the median value for the entire dataset. States with a bachelor's degree percentage above the mean were classified as "high," while those below the median were classified as "low." This allowed for a clearer analysis of the relationship between education and mortality rates, making it easier to identify patterns and trends.

To analyze the relationships between education and mortality, the appropriate statistical tests would be a linear regression model and a Pearson's correlation test. The linear regression model will determine whether the percentage of adults with a bachelor's degree significantly predicts Alzheimer's mortality rates. This analysis will help assess whether higher education levels are associated with lower or higher mortality rates from Alzheimer's disease. The Pearson's correlation test will measure the strength and direction of the linear relationship between the percentage of adults with a bachelor's degree and Alzheimer's mortality rates. A negative correlation would suggest that higher educational attainment is linked to lower mortality rates, while a positive correlation would show the opposite. For our research question examining the relationship between income and mortality, we will be conducting a Pearson correlation test and looking at a linear regression model. The Pearson correlation tests whether there is a statistically significant relationship between two variables, and in this case, it measures the strength and direction between median household income and Alzheimer's mortality rate. The linear regression model estimates how the mortality rate changes based on income, helping us understand if income is a significant predictor of mortality.

For our research question examining the relationship between depression rate and mortality, we conducted a correlation test and linear regression model on the data on depression rate and mortality rate. The correlation test will indicate whether there is a statistically significant relationship between depression rate and mortality rate. The correlation coefficient will reflect the strength and direction of this relationship. We will also conduct a linear

regression model to help estimate how the mortality rate changes based on depression rates. The p-value will indicate whether the relationship is significant, and the r-squared value will explain how much variability is explained by the variable of depression rate when it comes to predicting the mortality rate of Alzheimer's.

```
library(tidyverse)
```

```
## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
## ✓ dplyr     1.1.4     ✓ readr     2.1.5
## ✓ forcats   1.0.0     ✓ stringr   1.5.1
## ✓ ggplot2   3.5.1     ✓ tibble    3.2.1
## ✓ lubridate 1.9.3     ✓ tidyr     1.3.1
## ✓ purrr     1.0.2
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to becom
e errors
```

```
library(usdata)
```

```
## Warning: package 'usdata' was built under R version 4.4.3
```

```
library(ggplot2)
library(maps)
```

```
##
## Attaching package: 'maps'
##
## The following object is masked from 'package:purrr':
##
##     map
```

```
library(dplyr)
library(scales)
```

```
##
## Attaching package: 'scales'
##
## The following object is masked from 'package:purrr':
##
##     discard
##
## The following object is masked from 'package:readr':
##
##     col_factor
```

```
alzheimers_mortality_state <- read.csv("alzheimers.csv")
income_2022 <- read.csv("Median Household Income data - Sheet1.csv")
education_2022 <- read.csv("Bachelor's Attainment by Percentage - Sheet1.csv")
depression_2022 <- read.csv("Indicators_of_Anxiety_or_Depression_Based_on_Reported_Frequency_of_
Symptoms_During_Last_7_Days_20250316.csv")

depression_2022 <- depression_2022 %>%
  group_by(State) %>%
    summarize(value = mean(Value))

alzheimers_mortality_state$STATE <- abbr2state(alzheimers_mortality_state$STATE)

income_2022 <- income_2022 %>%
  rename(STATE = Name)

education_2022 <- education_2022 %>%
  rename(STATE = Name)

df <- merge(alzheimers_mortality_state, income_2022, by="STATE")
df <- merge(df, education_2022, by="STATE")
df <- df %>%
  filter(YEAR == 2022)

df <- df %>%
  rename(`Median Household Income` = X2022.x)
df <- df %>%
  rename(`Bachelor's degree percentage` = X2022.y)
```

```
depression_2022 <- depression_2022 %>%
  rename(STATE = State)

depression_2022 <- depression_2022 %>%
  rename(depression_rate = value)

df <- merge(df, depression_2022, by="STATE")
df <- subset(df, select = -c(`URL`))
df$DEATHS= as.numeric(gsub("[,]", "", df$DEATHS))
df$DEATHS <- as.integer(df$DEATHS)
df$`Median Household Income` = as.numeric(gsub("[,]", "", df$`Median Household Income`))
df$`Median Household Income` <- as.integer(df$`Median Household Income`)
df$education_level <- cut(df$`Bachelor's degree percentage`,
                          breaks = 2,
                          labels = c("Low", "High"))

tapply(df$`Bachelor's degree percentage`, df$education_level, range)
```

```
## $Low
## [1] 24.8 35.6
##
## $High
## [1] 35.9 46.6
```

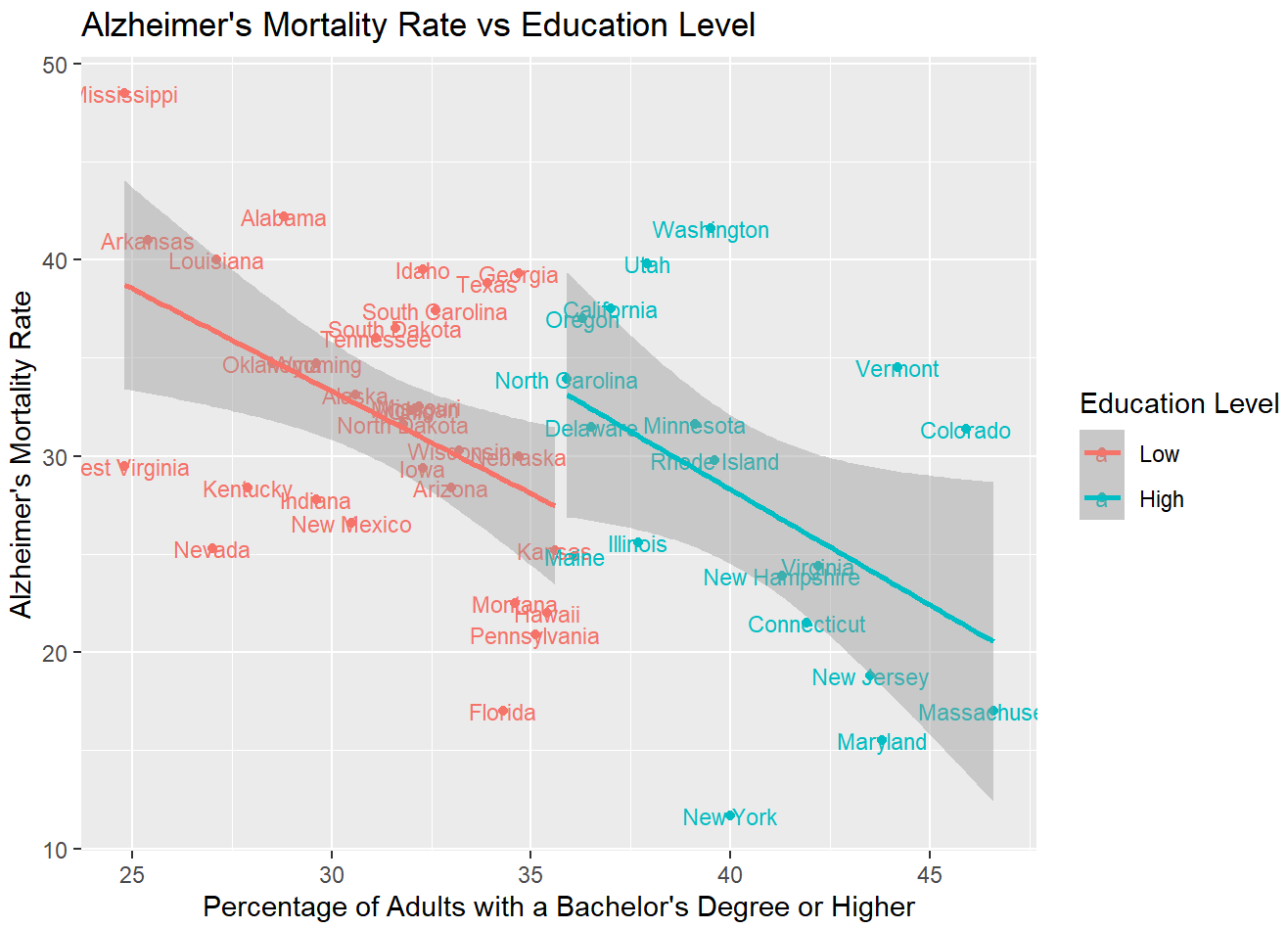
IV. Visualizations and Findings

Doha Education Analysis

```
df %>%
  ggplot(aes(x = `Bachelor's degree percentage`, y=RATE, col = education_level, label = STATE))+
  geom_point() +
  geom_text(size = 3) +
  geom_smooth(method = "lm") +
  labs(title = "Alzheimer's Mortality Rate vs Education Level",
       x = "Percentage of Adults with a Bachelor's Degree or Higher",
       y = "Alzheimer's Mortality Rate",
       color = "Education Level")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```

```
## Warning: The following aesthetics were dropped during statistical transformation: label.
## ℹ This can happen when ggplot fails to infer the correct grouping structure in
##   the data.
## ℹ Did you forget to specify a `group` aesthetic or to convert a numerical
##   variable into a factor?
```




```
model <- lm(RATE ~ `Bachelor's degree percentage`, data = df)
summary(model)
```

```
##
## Call:
## lm(formula = RATE ~ `Bachelor's degree percentage`, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15.2118  -5.2735  -0.0608   4.3798  14.3520
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     53.8036     6.3633   8.455 4.56e-11 ***
## `Bachelor's degree percentage`  -0.6723     0.1815  -3.705 0.000547 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.975 on 48 degrees of freedom
## Multiple R-squared:  0.2224, Adjusted R-squared:  0.2062
## F-statistic: 13.73 on 1 and 48 DF,  p-value: 0.0005467
```

```
correlation <- cor.test(df$`Bachelor's degree percentage`, df$RATE, method = "pearson")
print(correlation)
```

```
##
##  Pearson's product-moment correlation
##
## data:  df$`Bachelor's degree percentage` and df$RATE
## t = -3.7048, df = 48, p-value = 0.0005467
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.6628918 -0.2223952
## sample estimates:
##       cor
## -0.4715523
```

The result of the linear regression model indicates a statistically significant negative relationship between the percentage of adults with bachelor's degrees and Alzheimer's mortality rates in the US. This is because the bachelor's degree percentage coefficient was -0.6723, suggesting that for every 1% increase in the percentage of adults with a bachelor's degree, the Alzheimer's mortality rate may decrease by 0.6723, highlighting that higher education is associated with lower mortality rates. The p-value was 0.000547, showing that the regression model is significant. Pearson's correlation test resulted in a correlation coefficient of -0.472 and a p-value of 0.000547. This suggests a moderate, negative correlation between Alzheimer's mortality rate and bachelor's degree percentage. To visually represent this relationship, a scatterplot was created with the percentage of adults with a bachelor's degree on the x-axis and the Alzheimer's mortality rate on the y-axis, with each state labeled. The scatterplot color-codes the states based on their education level. The regression line on the plot shows a downward trend, meaning that as the percentage of people with a bachelor's degree increases, the Alzheimer's mortality rate decreases. The visualization allows for an easy comparison of states with different education levels and highlights the negative trend. The regression analysis is supported through the scatterplot.

Halle Income Analysis

```
df$STATE <- tolower(df$STATE)
lm_model <- lm(RATE ~ `Median Household Income`, data = df)
summary(lm_model)
```

```
##
## Call:
## lm(formula = RATE ~ `Median Household Income`, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -18.5779  -5.4622   0.5593   4.7632  14.6070
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 4.872e+01  6.231e+00   7.818 4.15e-10 ***
## `Median Household Income` -2.336e-04  7.888e-05  -2.962  0.00475 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.273 on 48 degrees of freedom
## Multiple R-squared:  0.1545, Adjusted R-squared:  0.1369
## F-statistic: 8.772 on 1 and 48 DF,  p-value: 0.004745
```

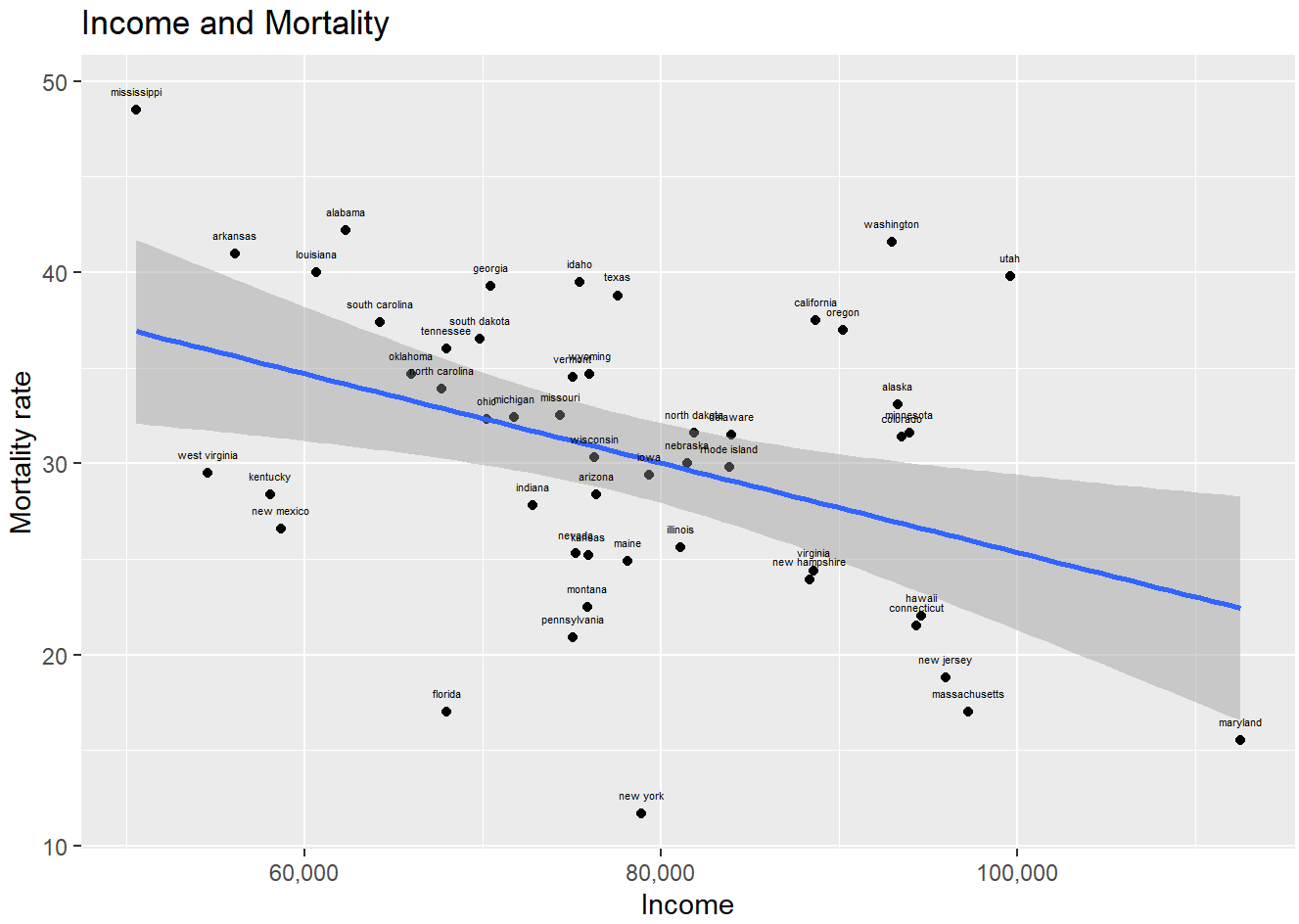
```
cor.test(df$`Median Household Income`, df$RATE, method = "pearson")
```

```
##
##  Pearson's product-moment correlation
##
## data:  df$`Median Household Income` and df$RATE
## t = -2.9617, df = 48, p-value = 0.004745
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.6052099 -0.1288276
## sample estimates:
##        cor
## -0.3930799
```

```
ggplot(df, aes(x= `Median Household Income`, y= RATE, label=STATE)) +
  geom_point() +
  geom_smooth(method='lm') +
  labs(title="Income and Mortality",
       x = "Income",
       y = "Mortality rate") +
  scale_x_continuous(labels = comma) + geom_text(size= 1.5, nudge_y=1)
```

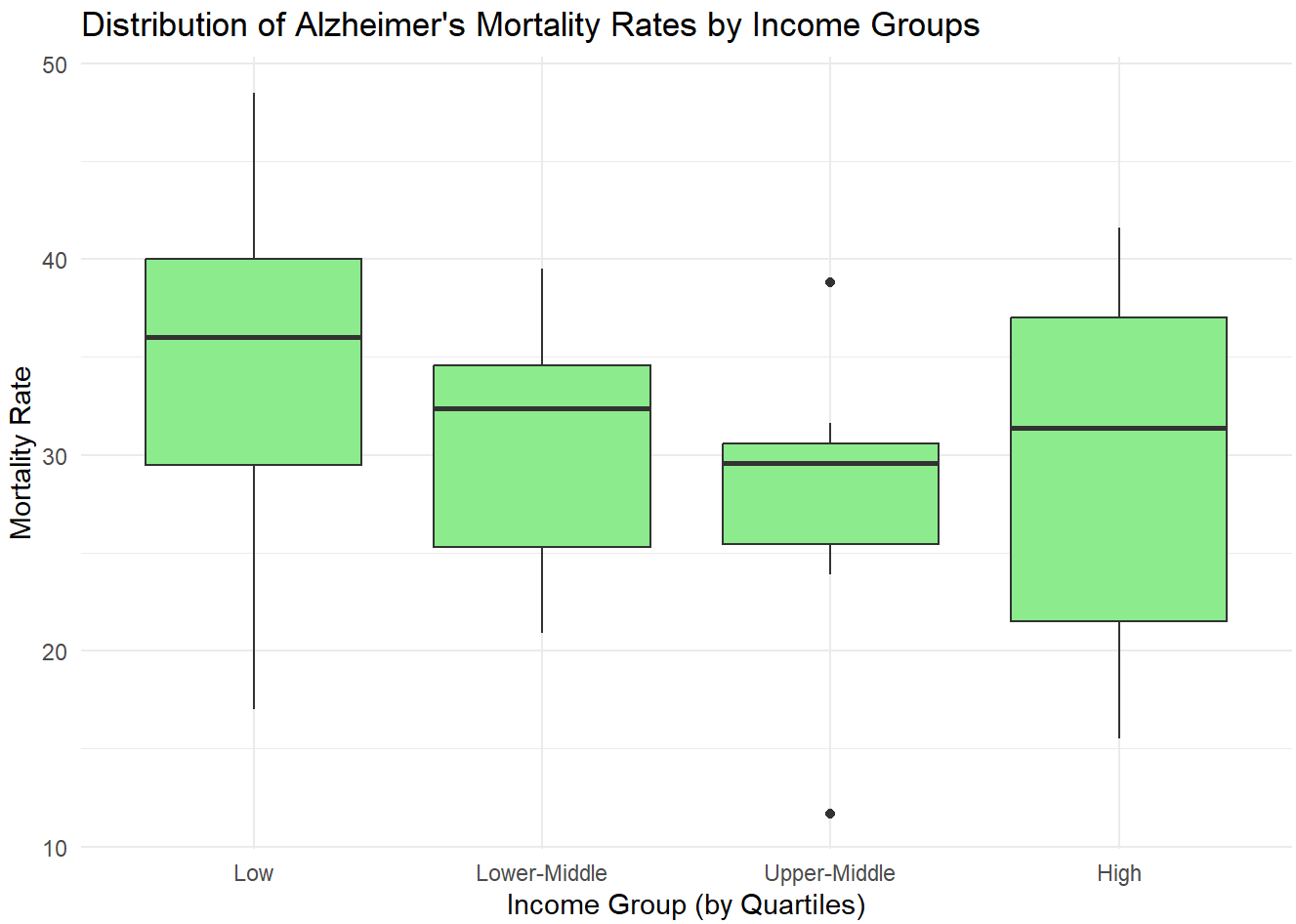
```
## `geom_smooth()` using formula = 'y ~ x'
```

```
## Warning: The following aesthetics were dropped during statistical transformation: label.
## ℹ This can happen when ggplot fails to infer the correct grouping structure in
##   the data.
## ℹ Did you forget to specify a `group` aesthetic or to convert a numerical
##   variable into a factor?
```




```
df$IncomeGroup <- cut(df$`Median Household Income`,
                      breaks = quantile(df$`Median Household Income`,
                                        probs = seq(0, 1, 0.25),
                                        na.rm = TRUE),
                      include.lowest = TRUE,
                      labels = c("Low", "Lower-Middle", "Upper-Middle", "High"))

ggplot(df, aes(x = IncomeGroup, y = RATE)) +
  geom_boxplot(fill = "lightgreen") +
  labs(title = "Distribution of Alzheimer's Mortality Rates by Income Groups",
       x = "Income Group (by Quartiles)",
       y = "Mortality Rate") +
  theme_minimal()
```



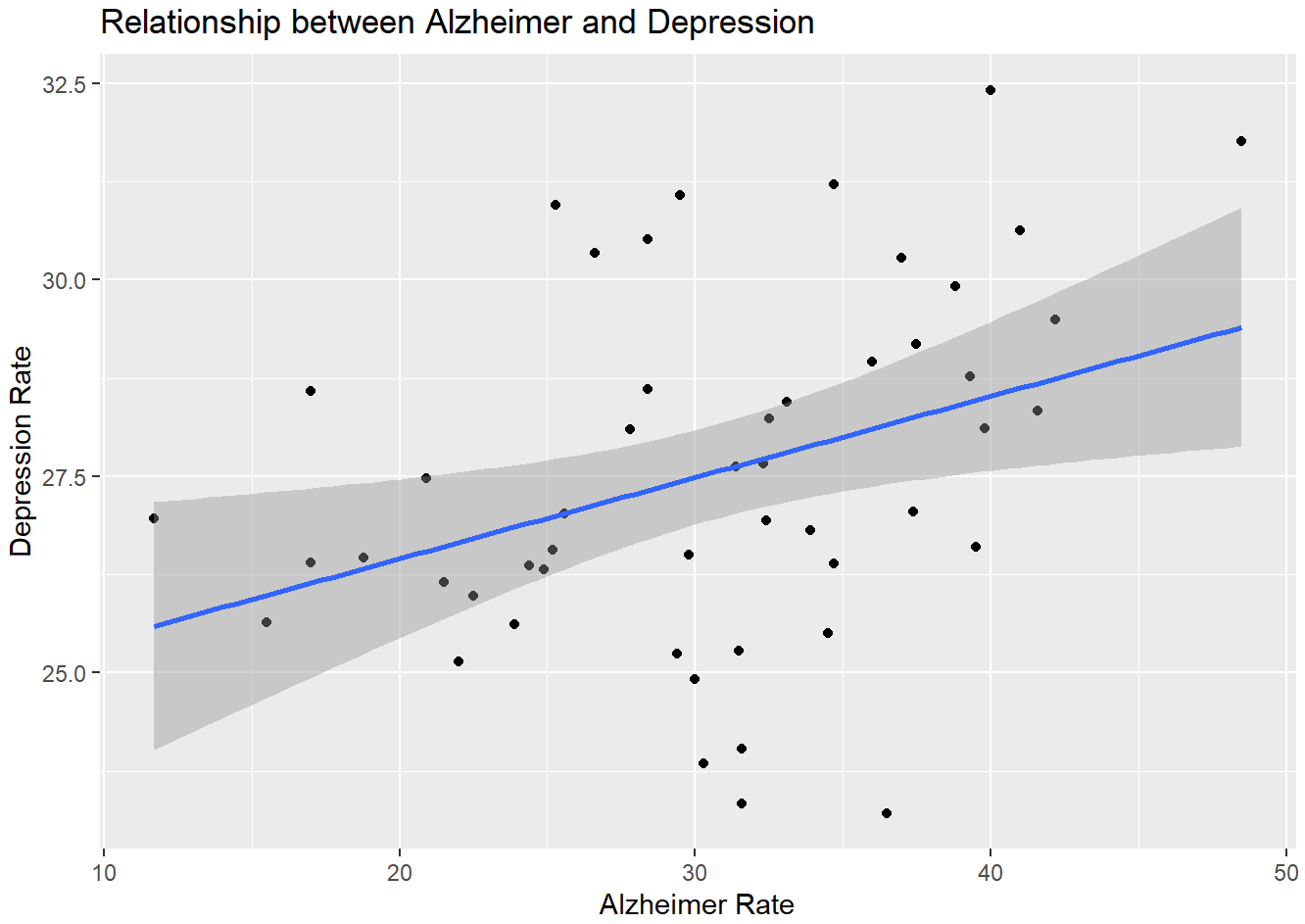
Mississippi had the lowest median income at $50,540 while Maryland had the highest at $112,500. Mississippi also had the highest mortality rate at 48.5% and New York had the lowest at 11.7%. Using Pearson's Correlation test, we got a p-value of 0.004745, indicating that there is a statistically significant relationship between income and Alzheimer's mortality rate. A simple linear regression model estimated an income coefficient of -2.336e-04. This indicates that as income increases by $1, Alzheimer's mortality rate decreases by 0.0002336.

The scatter plot displays the income vs mortality rate, and the linear regression line follows a negative trend. Higher-income states tend to experience lower Alzheimer's mortality rates. For the boxplot, I separated income into quartiles and visualized the distribution of Alzheimer's mortality rates across these groups. The boxplot shows that lower-income states tend to have higher median mortality rates than higher-income states. The low-income states have a higher median mortality rate, and larger spread, which suggests greater variability. The upper-middle states have the smallest spread and lowest median, and the high-income states have a wider range but the median is still lower than the low and low-middle states. This suggests that higher income levels generally have lower mortality rates than lower income levels.

Alex Depression Analysis

```
df %>%
  ggplot(aes(x = RATE, y = depression_rate)) +
    geom_point() +
      geom_smooth(method = 'lm') +
        labs(title = "Relationship between Alzheimer and Depression",
             x = "Alzheimer Rate",
             y = "Depression Rate")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```




```
result <- cor.test(df$RATE, df$depression_rate)
result
```

```
##
##  Pearson's product-moment correlation
##
## data:  df$RATE and df$depression_rate
## t = 2.6829, df = 48, p-value = 0.009983
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.09200903 0.58105253
## sample estimates:
##       cor
## 0.3611086
```

```
model <- lm(RATE ~ depression_rate, data = df)
summary(model)
```

```
##
## Call:
## lm(formula = RATE ~ depression_rate, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -18.092  -5.261   1.535   5.454  12.654
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -4.2353    12.9943  -0.326  0.74589
## depression_rate   1.2619     0.4703   2.683  0.00998 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.376 on 48 degrees of freedom
## Multiple R-squared:  0.1304, Adjusted R-squared:  0.1123
## F-statistic: 7.198 on 1 and 48 DF,  p-value: 0.009983
```

We created a scatter plot with Alzheimer's mortality rate on x-axis and depression rate on y-axis. The regression line on the scatter plot shows an upward trend, indicating that there is a positive correlation between the two variables. By conducting correlation test and linear regression model, we have gotten the result of 0.3611086, indicating that the correlation is positive, but the strength is a bit under moderate level. The linear regression model we applied later results in the correlation coefficient of 1.2619, meaning that for every 1% increase in Alzheimer's rate, we expect a 1.2619 increase in log-odds of depression rate. It is statistically significant because we got a p-value of 0.00998.

V. Discussion Our analysis supports our hypothesis that higher income and education levels are associated with lower Alzheimer's mortality rates. In our background research, we found that "education may help Alzheimer's patients cope better with the effects of brain atrophy by increasing cognitive reserve." Additionally, lower-income states are generally associated with fewer resources for healthcare, and our background research also stated that a lower socioeconomic status results in an increased risk for Alzheimer's. However, one unexpected finding was the variability within the high-income states, as we originally assumed that the wealthiest states would have the lowest rates overall. This spread shows that other factors may be more involved.

While Alzheimer's mortality is associated with higher depression rates, the correlation coefficient was lower than expected, suggesting that other factors influence this relationship. Initially, we hypothesized that people with depression may be at a higher risk of dying from Alzheimer's due to potential common biological factors, such as stress and cognitive decline, which could increase vulnerability to Alzheimer's. However, the correlation coefficient suggests that additional factors play a more significant role in determining Alzheimer's mortality.

For the next steps, we can investigate other social determinants such as lifestyle habits, healthcare accessibility, and environmental factors (i.e. air pollution) to provide more insight into what affects Alzheimer's mortality. It would also be helpful to explore these relationships with a longitudinal cohort study, and maybe look into urban vs rural differences as well.

Given these findings, we propose two policy recommendations.

1. The first is to **strengthen educational programs on Alzheimer's risks and prevention**. Invest in community-based initiatives that promote early detection, cognitive development, mental health awareness, and healthy lifestyle choices. Increasing awareness and promoting healthy habits through social engagement and mental stimulation can help delay the onset and progression of Alzheimer's. These

programs should prioritize low-income and less-educated communities, where gaps in education and access to healthcare may contribute to higher mortality rates. The p-values from our analyses shows a statistically significant relationship between educational attainment or income and Alzheimer's mortality rates, which suggests that targeted educational efforts could help individuals practice healthier behaviors and reduce risk.

2. Our second policy is to **increase federal funding for research**. Recent government budget cuts have led to a $1.5 billion freeze in NIH medical research funding, with Alzheimer's research centers falling a $65 million delay. These financial constraints slow progress and limit critical research on prevention and treatment. We recommend a focus on socioeconomic disparities and how these impact diagnosis, progression, and mortality rates. Since our analysis shows statistically significant relationships between education, income, and Alzheimer's mortality, addressing social determinants of health will allow for better treatment strategies and help develop targeted healthcare interventions for at-risk populations.