STA 032 Spring Quarter 2024

The Grim Reapers Report - Analysis of Death Rates Around the World

Submitted by: Dongmin Wu, Alberto Ramirez, Taran Wariyar, Alexis Coppinger

Instructor: Amy T. Kim

University of California Davis

Introduction:

Our report focuses on worldwide mortality rates. We aimed to analyze correlations and trends between the death rates spanning multiple and individual countries. Our dataset provides mortality data from 204 countries and territories, including 31 different causes of death from 1990 to 2019.

We had four questions in mind while analyzing our data:

1. What are the most common causes of death worldwide and in the United States in 2019?
2. Over time, what does the mortality rate of HIV look like?
3. On average, proportionally, does America have more deaths than the rest of the world due to Alzheimer's/Dementia?
4. How related are deaths caused by cardiovascular disease to deaths caused by neoplasms?

These questions are interesting to us because of how they compare and contrast the difference in human experience temporally and spatially. The first question shows us how the distribution of deaths around the world can be different depending on where you live. The second question, looking at trends across time and seeing increases and decreases in deaths related to HIV really shows how our medical system has improved over the past 20 years. The third question shows the difference between the country we live in and the rest of the world in relation to Alzheimer's, one of our biggest causes of mortality; And the fourth question compares the two largest causes of death in the United States and in the world. Trying to understand death rates gives us a lot of insight when trying to figure out how the world is doing in terms of medical research, but also in terms of preventable factors that could lead to unnecessary deaths.

Data Preparation:

Our data was gathered through the website kaggle.com, which is a large data-bank used by many people in order to understand trends in the world. Our specific dataset was called ‘Cause of Deaths around the World (Historical Data)’ and was gathered through ‘Our World in Data,’ a trustworthy organization bent on producing factual and accurate open source data. Our data contained 204 countries and territories and 31 different causes of death. It represents the amount of people who died for each specific cause of death by country and year.

The way we prepared our data was partially through manipulating it to fit what we wanted to show, and partly removing outliers that would pose an issue within our statistics.
In terms of manipulation, we created a new column within the data to gather the total deaths which helped us gauge proportions needed for our statistical analysis. Later in the process, we removed outliers through R and made sure our data had decent distributions so that we could have clean distributions to work with for our statistics.

Moreover, our statistical analysis was mainly correlational rather than causational. Our only explanatory variables were time for death rates lowering for HIV. Other than this we are simply comparing data to each other and offering potential explanations for why these phenomena may be happening.

Descriptive Statistics:

Looking at our findings, we found a few interesting and surprising findings that we weren't expecting, but stumbled upon and included in this report.

For example, one fact that surprised us a lot was that cardiovascular disease is the leading cause of death around the world by a large margin as of 2019. In fact, cardiovascular disease vastly outcompetes its next highest killer neoplasms by a high margin. We are 95% confident that the true mean of deaths caused by cardiovascular disease is larger than those of neoplasms by about 9432 to 10796 deaths over the past 2 decades. Given the westernization of the world, it makes sense, but it's interesting to note that cardiovascular diseases are a product of a lack of exercise and a sedentary lifestyle/bad diet. Knowing that cardiovascular disease outperforms neoplasms is surprising given that cancer is much more talked about in the general public.

Moreover, another interesting finding to us was that the United States has a much higher death toll than average when it comes to Alzheimer's. One of the main culprits of this issue may be the lack of exercise and sleep common to most Americans given our fetishization of high working hours, as well as a poor diet. We have sufficient evidence that the United States' rate of Alzheimer's deaths is higher than that of the rest of the world on average at the 5% significant level. It is interesting to speculate about why this is the case in the United States.

Question 1: What are the most common causes of death worldwide and in the United States in 2019?

—Global: Cardiovascular disease is the most common cause of death, with a much higher number of deaths than other diseases, followed by cancer. The number of deaths from these two diseases is significantly higher than other diseases, which reflects the high mortality rate of non communicable diseases worldwide.

—United States: Similar to global data, cardiovascular disease and neoplasms are also the main causes of death in the United States. In addition, the number of deaths from Alzheimer's disease and other forms of dementia is relatively high, which implies the impact of these diseases on the elderly population.

Question 2: Over time, what does the mortality rate of HIV look like?

The death by HIV/AIDS was continuous increasing until start to decrease significantly during 2004 and 2005, hypothetically by factors including:
—The popularization of antiretroviral therapy: In 1996, the introduction of an efficient antiretroviral therapy (HAART) made a big breakthrough in HIV treatment. This treatment significantly extends the lifespan of HIV individuals and reduces the mortality rate of HIV/AIDS. Since the early 2000s, this treatment method has gradually become popular worldwide, especially in developed countries and some developing countries.
—Global Health Initiative and Financial Support: After 2000, global investment in againsting HIV/AIDS has significantly increased. For example, the establishment of The Global Fund and the US President's Emergency AIDS Relief Plan (PEPFAR) has provided significant financial support for HIV treatment and prevention in low-income countries. This helps to expand the coverage of antiretroviral therapy and improve the survival rates of many HIV individuals.

Question 3: On average, proportionally, does America have more deaths than the rest of the world due to Alzheimer's/Dementia?

μ_1 = America's average proportion of deaths by Alzheimer's

μ_2= World's average proportion of deaths by Alzheimer's

$H_0 : \mu_1 < \mu_2$ vs. $H_A : \mu_1 >= \mu_2$

Null Hypothesis: People do not die at proportionally higher rates in the United States than around the rest of the world.
People around the world die at higher rates from Alzheimer's than in the United States
Alternative Hypothesis: People die at proportionally higher rates in the United States than around the rest of the world.
We will give ourselves a significance level of 5%, or $\alpha = 0.05$.
Our Z score is: $Z_s = 1.6476$
Our P-Value = 0.0497
Given that the p-value is below 0.05, we can reject the null.

In other words, we have sufficient evidence that the United States' rate of Alzheimer's deaths is higher than that of the rest of the world on average at the 5% significant level.

It is interesting to think about why this may be the case in the United States knowing that large causes of Alzheimer's are having a sedentary lifestyle, a bad diet, mental attrition, and lack of sleep. This could, however, also be due to the fact that older people in the United States live longer than non-developed countries, and Alzheimer's is also mainly linked to old age.

Question 4: How related are deaths caused by cardiovascular disease to deaths caused by neoplasms?

μ_1 = Deaths caused by cardiovascular disease

μ_2= Deaths caused by neoplasms

$H_0 : \mu_1 = \mu_2$ vs. $H_A : \mu_1 \neq \mu_2$

Null: The true difference in means between deaths caused by cardiovascular disease and deaths caused by neoplasms is 0.
Alternative: The true difference in means between deaths caused by cardiovascular disease and deaths caused by neoplasms is not equal to 0.
Our test statistic: $t_s = 29.087$

The estimated SD that the sample difference is away from 0.

P-value < 0.0001 (2.2^-16)

If the death rates are equal, the chances of that happening are basically 0.

Using a significance level of 5%, or $\alpha = 0.05$, we reject the null hypothesis since our P-Value is smaller than α.

We have sufficient evidence that the true difference in means between deaths caused by cardiovascular disease and neoplasms is not equal to 0.

We are 95% confident that the true mean of deaths caused by cardiovascular disease is larger than those of neoplasms by about 9432 to 10796 deaths. This is a troubling finding, knowing that a large cause for cardiovascular disease is lifestyle. It is unfortunate that we live in a society that requires hours of sedentary work, cheaper bad diets than good ones, and forced commute times through driving rather than having walkable cities.

Conclusion:

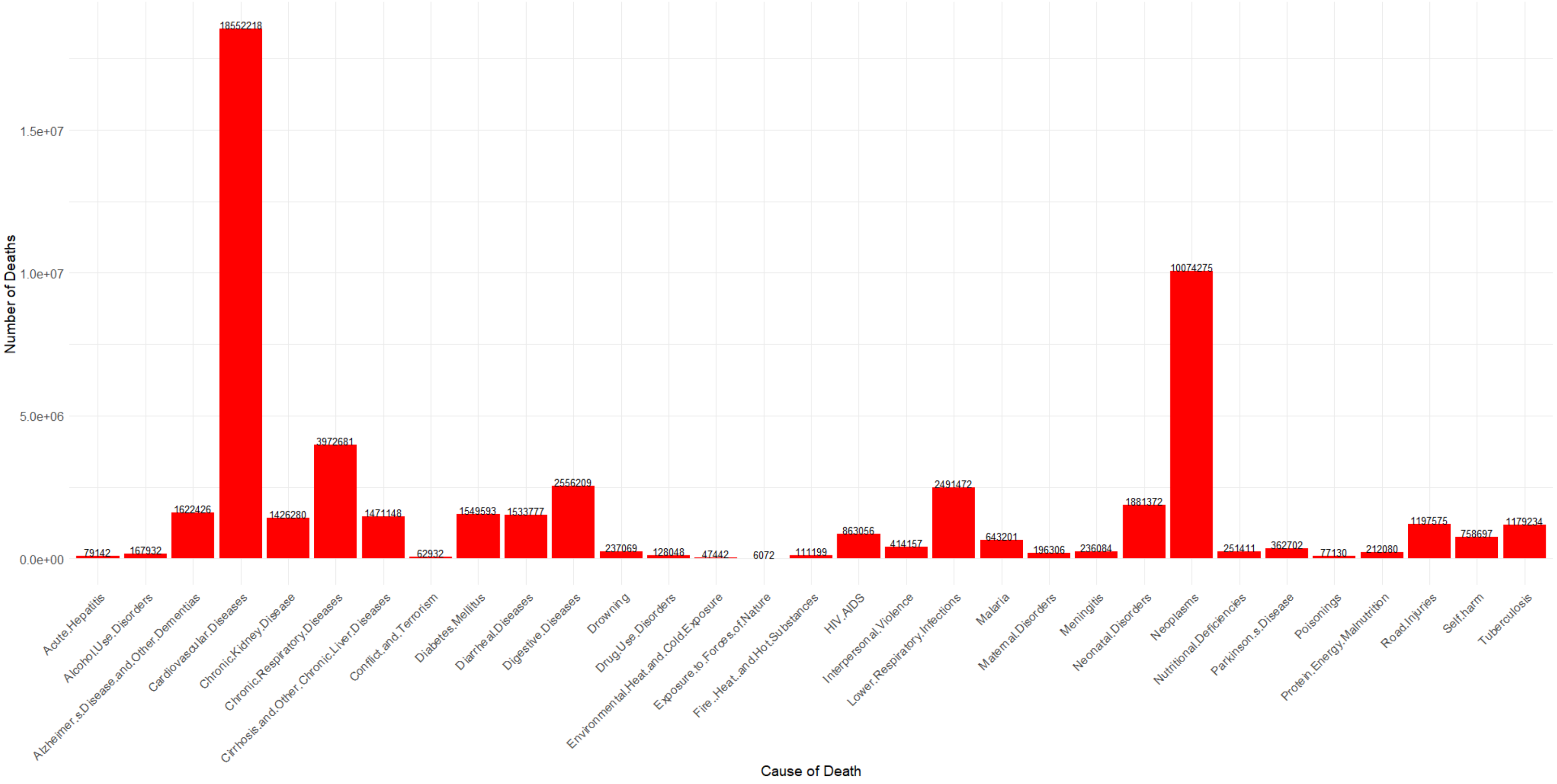
In our report we analyzed data from 204 countries and territories, spanning 31 causes of death from 1990 to 2019. We focused on four main questions and ideas: comparing the causes of death between the U.S. and the world, the trend in HIV mortality over time, the proportional death rates from Alzheimer's/Dementia in the U.S. versus the rest of the world, and the relatedness between deaths caused by malnutrition and deaths caused by cardiovascular disease and neoplasms.

Some next steps that could enhance our findings could be looking at factors between not just countries, but social income levels, healthcare access, education, etc. to better understand the relationship between social factors and causes of death. Another could be to obtain data that includes every type of death. Unfortunately, from our sample, we were only able to obtain causes of death that are relatively common. Having an 'other kinds of death' category with all the uncommon deaths possible would be much better to fully assess the proportions of rates of death in relation to all deaths possible within a country.

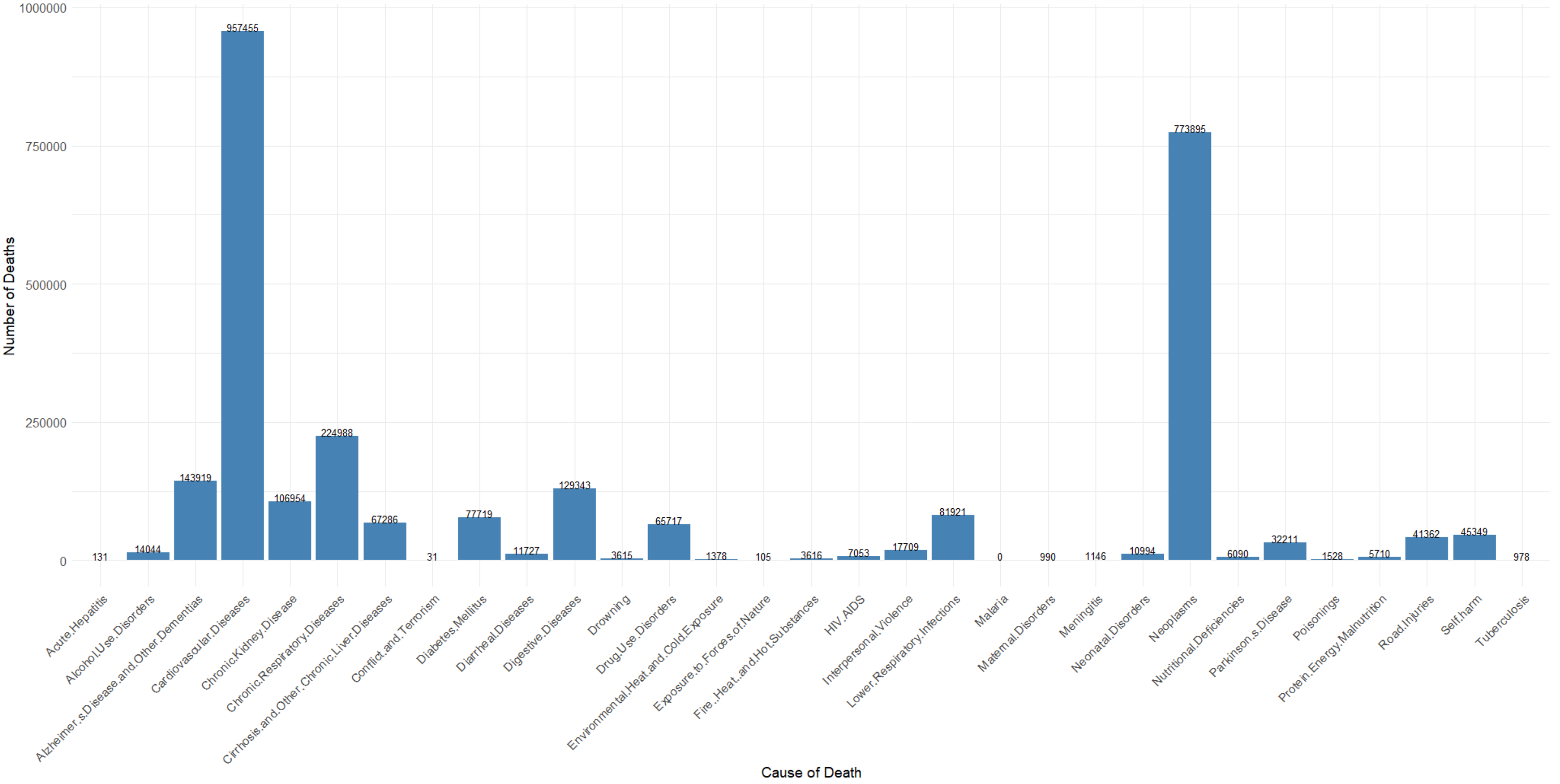
Visualization:

Q1:

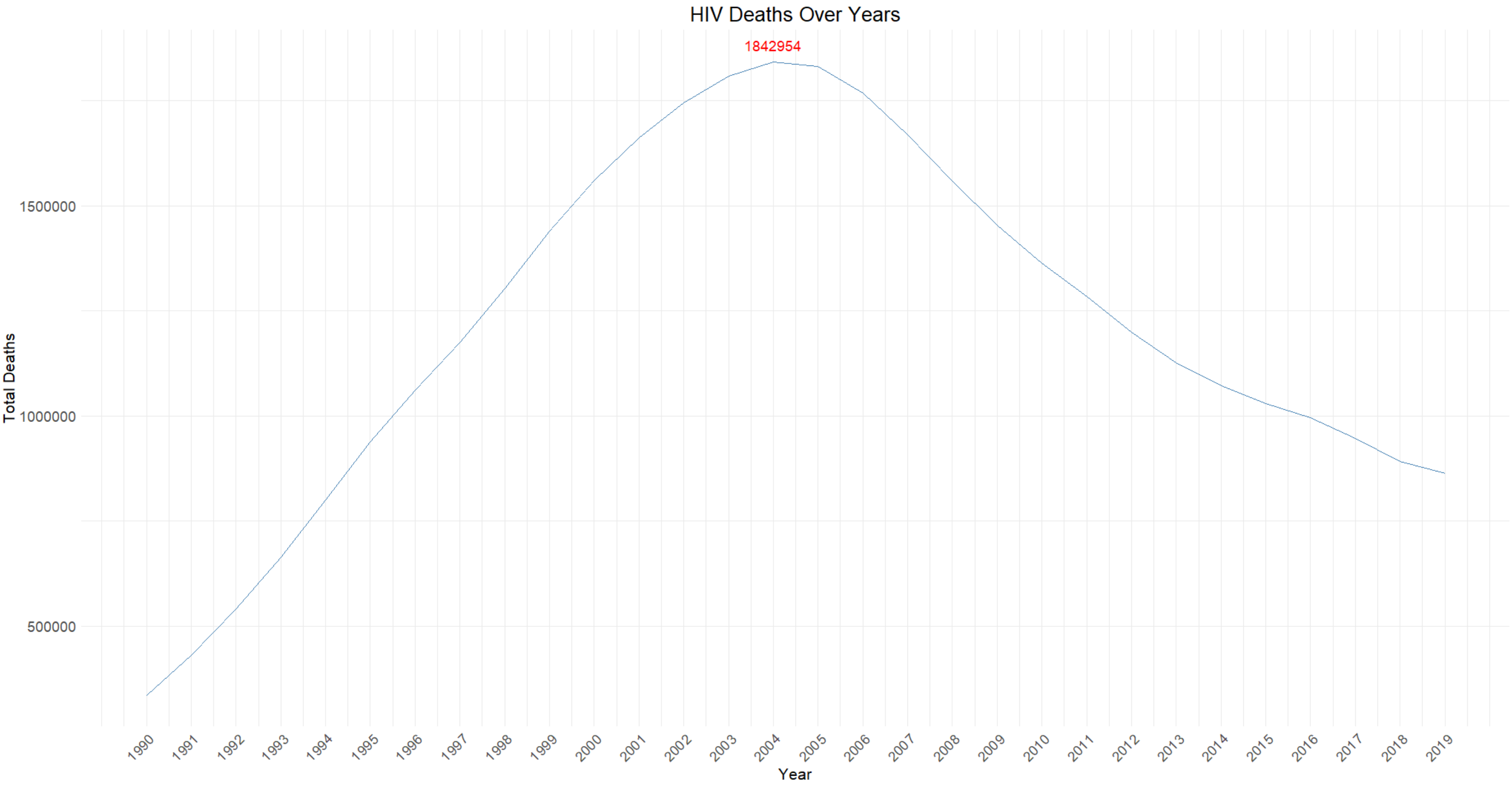
Global Number of Deaths by Cause, 2019



Number of Deaths by Cause in the United States, 2019



Q2:


HIV Deaths Over Years
1842954
Total Deaths
1500000
1000000
500000
Year
1990
1991
1992
1993
1994
1995
1996
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019

Appendix:

Question 1:

```
library(ggplot2)
library(dplyr)
library(tidyr)

global_2019_data <- subset(data, Year == 2019)

global_deaths <- global_2019_data %>%
  select(4:ncol(global_2019_data)) %>% # cause col from 4 to end
  summarise_all(sum) # sum every col

#convert data to long
long <- pivot_longer(global_deaths, cols = everything(), names_to = "Cause", values_to = "Deaths")


windows()

ggplot(long, aes(x = Cause, y = Deaths)) +
  geom_bar(stat = "identity", fill = "red") +
  geom_text(aes(label = Deaths), position = position_nudge(y = 5), angle = 0, color = "black", size = 3.5, vjust = 0) +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1, size = 12),
        axis.text.y = element_text(size = 12),
        axis.title = element_text(size = 14),
        plot.title = element_text(size = 16), ) +
  labs(x = "Cause of Death", y = "Number of Deaths", title = "Global Number of Deaths by Cause, 2019")



usa_2019_data <- subset(data, Year == 2019 & Country.Territory == "United States")


deaths <- usa_2019_data[, 4:ncol(usa_2019_data)] #cause start from 4 to end


long_data <- pivot_longer(deaths, cols = everything(), names_to = "Cause", values_to = "Deaths")#convert to long


p <- ggplot(long_data, aes(x = Cause, y = Deaths)) +
  geom_bar(stat = "identity", fill = "steelblue") +
  geom_text(aes(label = Deaths), position = position_nudge(y = 5), angle = 0, color = "black", size = 3.5, vjust = 0) +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1, size = 12),
       axis.text.y = element_text(size = 12),
       axis.title = element_text(size = 14),
       plot.title = element_text(size = 16),)+
  labs(x = "Cause of Death", y = "Number of Deaths", title = "Number of Deaths by Cause in the United States, 2019")

windows()
print(p)
```

Question 2:

```
data <- read.csv("E:/Study/STA32/Project/cause_of_deaths.csv")

# hiv deaths summary
hivSummary <- data %>%
  group_by(Year) %>%
  summarise(HIV_AIDS = sum(HIV.AIDS, na.rm = TRUE))

#highest point for hyphothesis for death drops
max_point <- hivSummary[which.max(hivSummary$HIV_AIDS),]


p <- ggplot(hivSummary, aes(x = Year, y = HIV_AIDS)) +
  geom_line(color = "steelblue") +
  geom_text(data = max_point, aes(label = HIV_AIDS, vjust = -1), size = 5, color = "red") +  # Label highest point
  scale_x_continuous(breaks = seq(min(hivSummary$Year), max(hivSummary$Year), by = 1)) +  # More years on x-axis
  labs(title = "HIV Deaths Over Years",
       x = "Year",
       y = "Total Deaths") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1, size = 14),  # x-axis text size
        axis.text.y = element_text(size = 14),  # y-axis text size
        axis.title = element_text(size = 16),  # axis title size
        plot.title = element_text(size = 20, hjust = 0.5),  #title size
        legend.title = element_text(size = 14),  #legend title size
        legend.text = element_text(size = 12))  # legend text size
windows()

print(p)
```

Question 3 and 4 (made in the same document):

```
deaths = read.csv("C:/Users/AlexC/Desktop/School Stuff/Davis/Q3/STA 032/cause_of_deaths.csv")
summary(deaths)

#US Data
```

```r
usa_data = subset(deaths, Code == "USA")
USA_prop = usa_data$Alzheimer.s.Disease.and.Other.Dementias / usa_data$Total.Deaths

#clean US data
Q1_USA = quantile(USA_prop, 0.25)
Q3_USA = quantile(USA_prop, 0.75)
IQR_USA = Q3_USA - Q1_USA

USA_lower_bound = Q1_USA - 1.5*IQR_USA
USA_upper_bound = Q3_USA + 1.5*IQR_USA

fixed_USA_data = USA_prop[USA_prop >= USA_lower_bound & USA_prop <= USA_upper_bound]

#USA mean/SD
mean_USA = mean(fixed_USA_data)
sd_USA = sd(fixed_USA_data)
print(mean_USA * 100)
print(sd_USA * 100)


#World Data
world_data = subset(deaths, Code != "USA")
world_data_prop = world_data$Alzheimer.s.Disease.and.Other.Dementias / world_data$Total.Deaths

#Clean World Data
Q1_w = quantile(world_data_prop, 0.25)
Q3_w = quantile(world_data_prop, 0.75)
IQR_w = Q3_w - Q1_w

lower_bound_w = Q1_w - 1.5*IQR_w
upper_bound_w = Q3_w + 1.5*IQR_w

fixed_world_data = world_data_prop[world_data_prop >= lower_bound_w & world_data_prop <= upper_bound_w]

#World mean/SD
mean_world = mean(fixed_world_data)
sd_world = sd(fixed_world_data)
print(mean_world * 100)
print(sd_world * 100)

#This is using the z-test formula
z_score = (mean_USA - mean_world) / sqrt(sd_USA^2 + sd_world^2)

p_value = 1-pnorm(z_score)

print(z_score)
print(p_value)

### This is now for question 4

#Cleaning the data for cardiovascular disease and neoplasms
Q1_card = quantile(deaths$Cardiovascular.Diseases, 0.25)
Q3_card = quantile(deaths$Cardiovascular.Diseases, 0.75)
IQR_card = Q3_card - Q1_card

card_lower_bound = Q1_card - 1.5*IQR_card
card_upper_bound = Q3_card + 1.5*IQR_card

fixed_card_data = deaths$Cardiovascular.Diseases[deaths$Cardiovascular.Diseases >= card_lower_bound &
deaths$Cardiovascular.Diseases <= card_upper_bound]


Q1_neo = quantile(deaths$Neoplasms, 0.25)
Q3_neo = quantile(deaths$Neoplasms, 0.75)
IQR_neo = Q3_neo - Q1_neo

neo_lower_bound = Q1_neo - 1.5*IQR_neo
neo_upper_bound = Q3_neo + 1.5*IQR_neo

fixed_neo_data = deaths$Neoplasms[deaths$Neoplasms >= neo_lower_bound & deaths$Neoplasms <= neo_upper_bound]

#Calculating the t-test for the confidence intervals, t-score and p-value
t_test = t.test(fixed_card_data, fixed_neo_data, alternative = "two.sided", conf.level = 0.95)
print(t_test)
```

Contributions:

Alexis Coppinger:

Organized the layout of the paper, coordinated the communication. Came up with and did the calculations for questions 3 and 4.

Alberto Ramirez:

Helped analyze data to come up with the conclusion and observations.

Taran Wariyar:

Helped organize and clean data for question 2 and helped organize the format and appendix of the project report.

Daniel Wu:

Generated questions 1 and 2 and visualization for question 1 and 2. Relevant R appendix was pasted.