DSCI 610 Project 1

Collin Hoskins

2021-03-12

Configuration and Library References

```
library(knitr)
library(rmdformats)
library(ggplot2)
library(reticulate)
library(reshape2)
library(ggpubr)
library(kableExtra)
library(tidyverse)
library(foreign)
library(lubridate)
use_python("C:/Users/Collin/anaconda3/python.exe")
opts_chunk$set(echo=FALSE,
                 cache=TRUE,
               prompt=FALSE,
               tidy=TRUE,
               comment=NA,
               message=FALSE,
               warning=FALSE)
opts_knit$set(width=75)
options(max.print="75")
py_run_string('import os')
py_run_string("os.environ['QT_QPA_PLATFORM_PLUGIN_PATH'] =
         'C:/Users/Collin/anaconda3/Library/plugins/platforms'")
setwd("C:/Users/Collin/Desktop/Academic/Courses/DSCI 610/Project 1")
```

Question 1

(20 pts) Summarize the contents of each dataset briefly for the users. Illustrate the survey design(s) used for data collection.

Summary of Datasets

According to the technical notes regarding these datasets, the provisional death counts are ultimately processed by the National Center for Health Statistics (NCHS). Furthermore, the data that are received by the NCHS come from many different sources, as death certificates are required for each death, and these certificates take time to process in some cases. Also, some states utilize laboratory and autopsy results to confirm the cause of death, which increases the amount of time for the state the send the data to the NCHS. These steps are essentially the design for data collection.

1

Provisional COVID-19 Deaths by Sex, Age, and State

The Provisional COVID-19 Deaths by Sex, Age, and State dataset contains categorical variables (Sex, Month, Group, State), as well as text and numberic variables. The key takeaway from this dataset are the 16 different variables that provide a thorough dataset for many different types of analyses. This dataset contains data for Pneumonia as well as Influenza deaths.

Provisional COVID-19 Deaths by Sex, Age, and Week

The Provisional COVID-19 Deaths by Sex, Age, and Week dataset is similar to the preceeding dataset mentioned. This dataset contains information regarding COVID-19 deaths, as well as the total deaths from all causes within a given week. The week variable is numberic, which is helpful for time-series analyses. Like the previous dataset, this dataset has multiple categorical variables.

Provisional COVID-19 Deaths by Conditions

The Provisional COVID-19 Deaths by Conditions dataset contains many of the same variables as the first two datasets, as well as data regarding varying conditions and the condition that contributed to a COVID-19 death.

2

Question 2

Loading the datasets

```
deathsByState <- read.csv("COVID_DEATHS_SEX_AGE_STATE.CSV")
deathsByWeek <- read.csv("COVID_DEATHS_SEX_AGE_WEEK.csv")
deathsByCondition <- read.csv("COVID_DEATHS_CONDITIONS.csv")
```

Cleaning the desired dataset

```
q21df <- filter(deathsByState, Group == "By Total" & State == "United States" & Sex != "All
        Sexes" & Age.Group != "All Ages")


adjDf <- melt(q21df, measure.vars = c("COVID.19.Deaths","Pneumonia.Deaths",
        "Influenza.Deaths"))
```

Comparing COVID-19, Pneumonia, and Influenza deaths across sex and age group.

```
ggplot(adjDf) + geom_col(aes(variable, value, fill = variable)) + facet_grid(Sex~Age.Group)
        + xlab(" ") + ylab("Number of Deaths") + theme(text = element_text(size = 16),
        axis.title.x = element_blank(), axis.text.x = element_blank(), axis.ticks.x =
        element_blank())
```

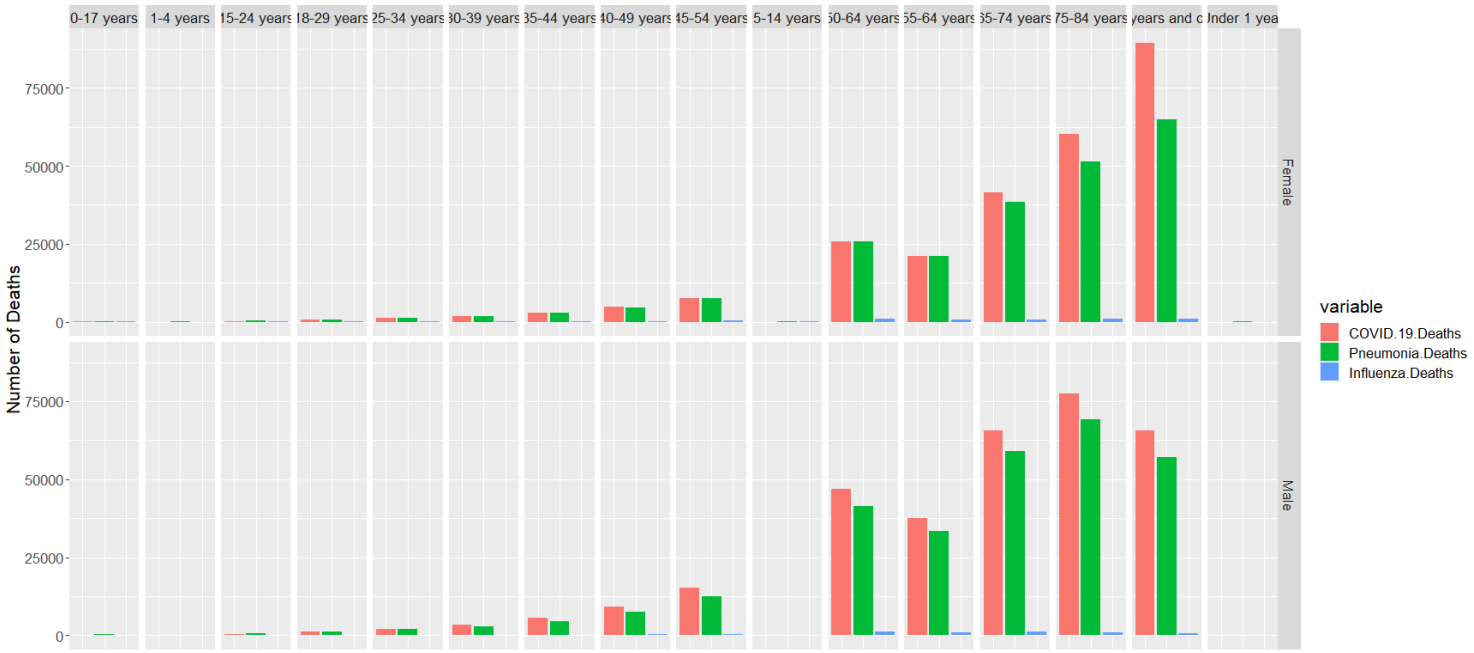
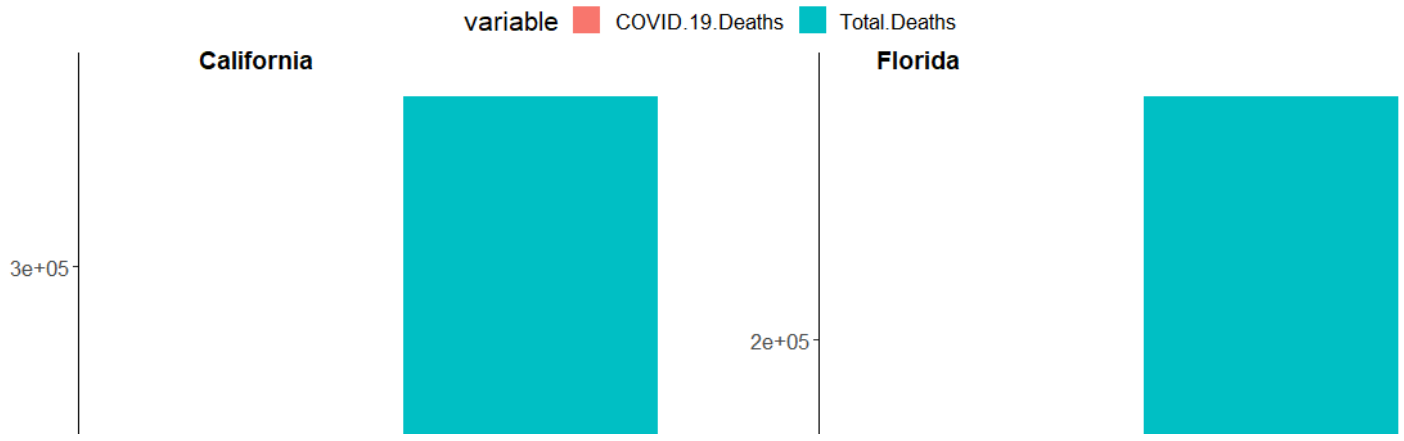


Figure 1. Comparison of COVID-19, Pneumonia, and Influenza deaths across sex and age groups.

Across all age groups, COVID-19 and Pneumonia deaths are far greater compared to Influenza deaths. For females, however, COVID-19 deaths are significantly more prevalent than Pneumonia deaths compared to the same difference in males for ages 65 and older. (Figure 1.).

Comparing COVID-19 deaths with all deaths in California, Florida, New York, and Texas.

```
calif <- filter(deathsByState, Group == "By Total" & State == "California" & Sex == "All
        Sexes" & Age.Group == "All Ages")

florida <- filter(deathsByState, Group == "By Total" &  State == "Florida"& Sex == "All
        Sexes" & Age.Group == "All Ages")

NY <- filter(deathsByState, Group == "By Total" &  State == "New York"& Sex == "All Sexes"
        & Age.Group == "All Ages")

texas <- filter(deathsByState, Group == "By Total" &  State == "Texas"& Sex == "All Sexes"
        & Age.Group == "All Ages")

CA_melt <- melt(calif, measure.vars = c("COVID.19.Deaths", "Total.Deaths"))

FL_melt <- melt(florida, measure.vars = c("COVID.19.Deaths", "Total.Deaths"))

NY_melt <- melt(NY, measure.vars = c("COVID.19.Deaths", "Total.Deaths"))

TX_melt <- melt(texas, measure.vars = c("COVID.19.Deaths", "Total.Deaths"))

A <- ggplot(CA_melt) + geom_col(aes(variable, value, fill = variable)) + theme_classic() +
        theme(text = element_text(size = 16), axis.title.x = element_blank(), axis.text.x
        = element_blank(), axis.ticks.x = element_blank())

B <- ggplot(FL_melt) + geom_col(aes(variable, value, fill = variable)) + theme_classic() +
        theme(text = element_text(size = 16), axis.title.x = element_blank(), axis.text.x
        = element_blank(), axis.ticks.x = element_blank())

C <- ggplot(NY_melt) + geom_col(aes(variable, value, fill = variable)) + theme_classic() +
        theme(text = element_text(size = 16), axis.title.x = element_blank(), axis.text.x
        = element_blank(), axis.ticks.x = element_blank())

D <- ggplot(TX_melt) + geom_col(aes(variable, value, fill = variable)) + theme_classic() +
        theme(text = element_text(size = 16), axis.title.x = element_blank(), axis.text.x
        = element_blank(), axis.ticks.x = element_blank())

ggarrange(A, B, C, D, labels = c("California", "Florida", "New York", "Texas"), hjust =
        -2.0, vjust = 1.25, common.legend = T)
```



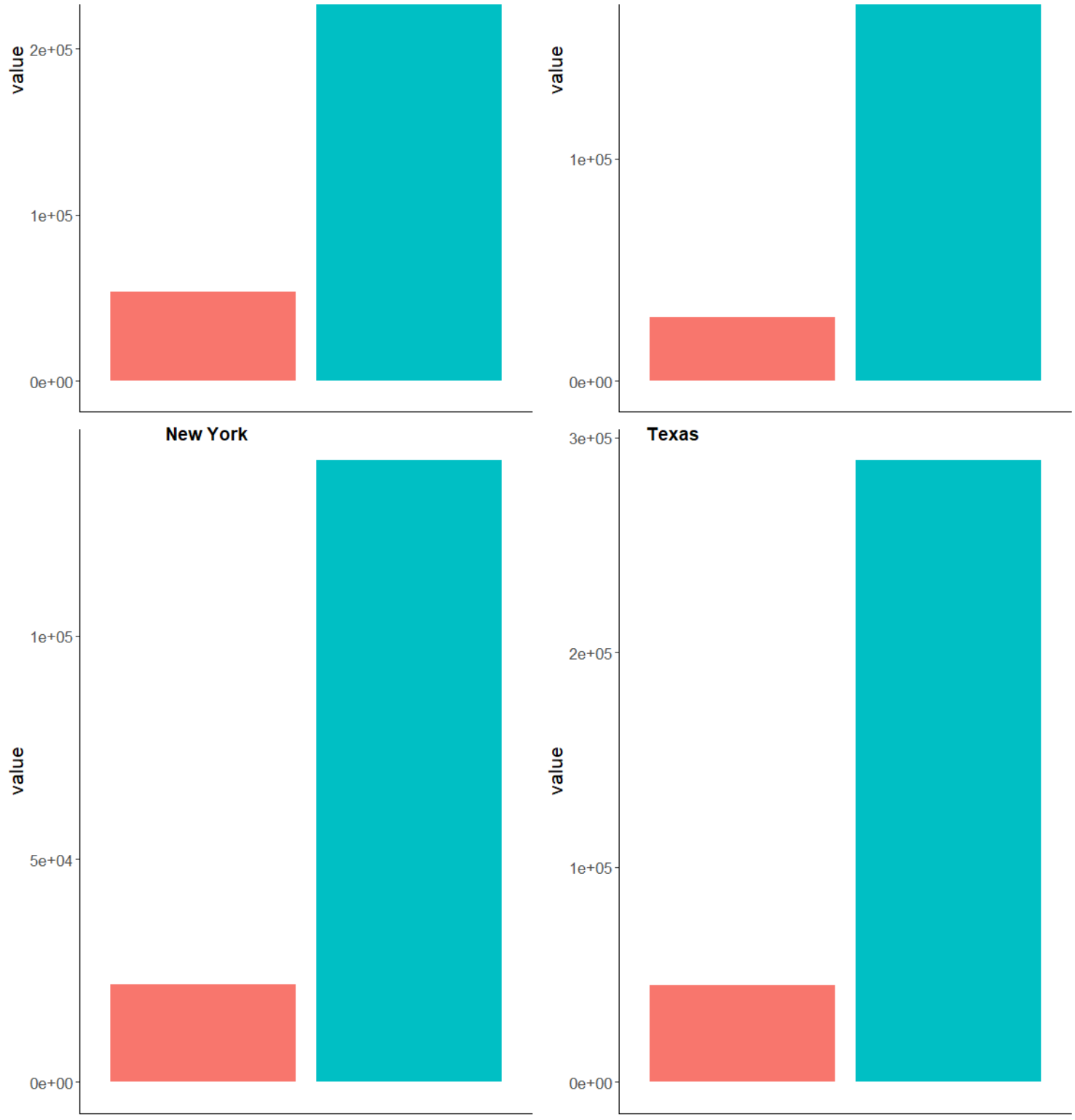


Figure 2. Comparison of COVID-19 deaths with total deaths in California, Florida, New York, and Texas.

```
altStates <- filter(deathsByState,  State == "California" | State == "Florida" | State ==
        "New York" | State == "Texas" )

altStates <- filter(altStates, Group == "By Total" & Sex == "All Sexes" & Age.Group == "All
        Ages")

altStatesMelt <- melt(altStates, measure.vars = c("COVID.19.Deaths","Total.Deaths"))
```

```
ggplot(altStatesMelt) + geom_col(aes(variable, value, fill = variable)) +
        facet_wrap(~State, nrow = 1) + xlab(" ") + ylab("Number of Deaths") + theme(text =
        element_text(size = 16), axis.title.x = element_blank(), axis.text.x =
        element_blank(), axis.ticks.x = element_blank())
```

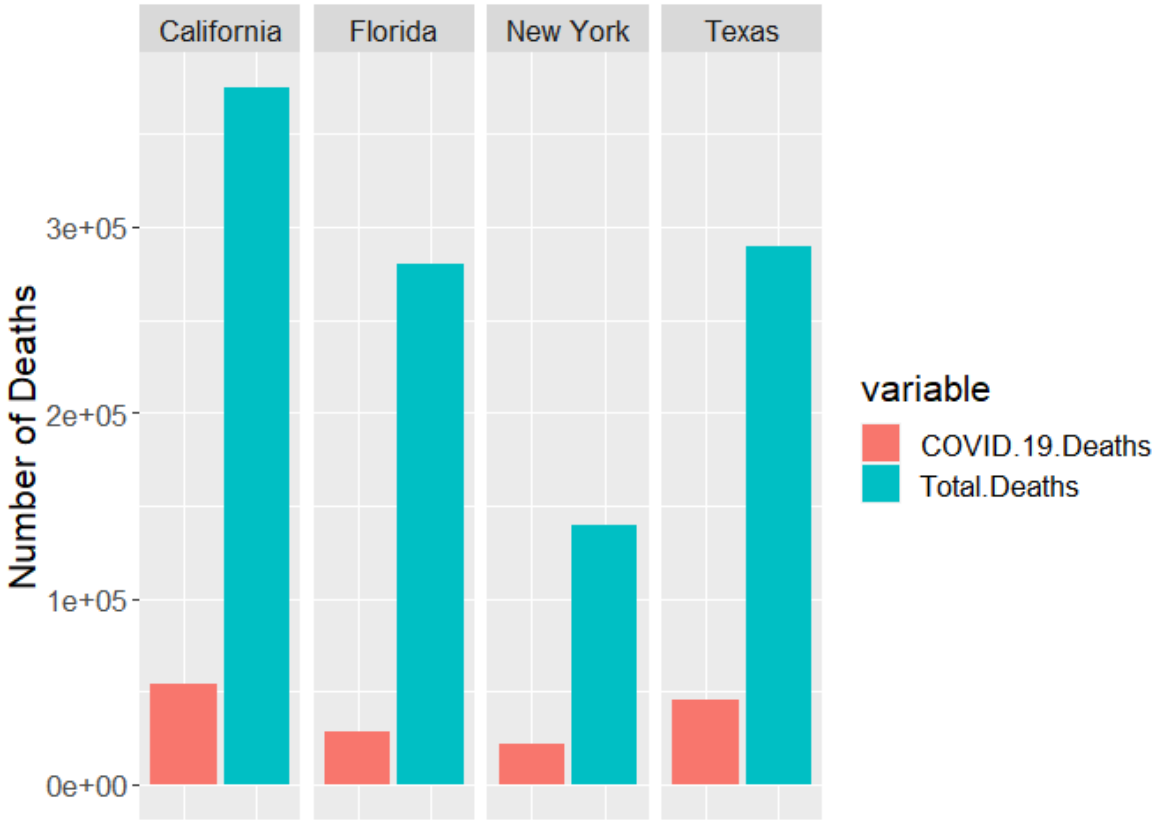


Figure 3. Alternative visualisation to compare COVID-19 deaths with total deaths in California, Florida, New York, and Texas.

The proportions of COVID-19 deaths that contributed to overall deaths in California and Texas are much greater than the proportion of COVID-19 deaths to overall deaths in Florida and New York (Figure 3.).

```
q31df <- filter(deathsByWeek, Sex != "All Sex" & Age.Group == "All Ages")

ggplot(q31df, aes(x = MMWR.Week, y = COVID.19.Deaths, color = Sex)) + geom_point() +
        ylab("Number of Deaths") + xlab("Week") + theme_classic()
```

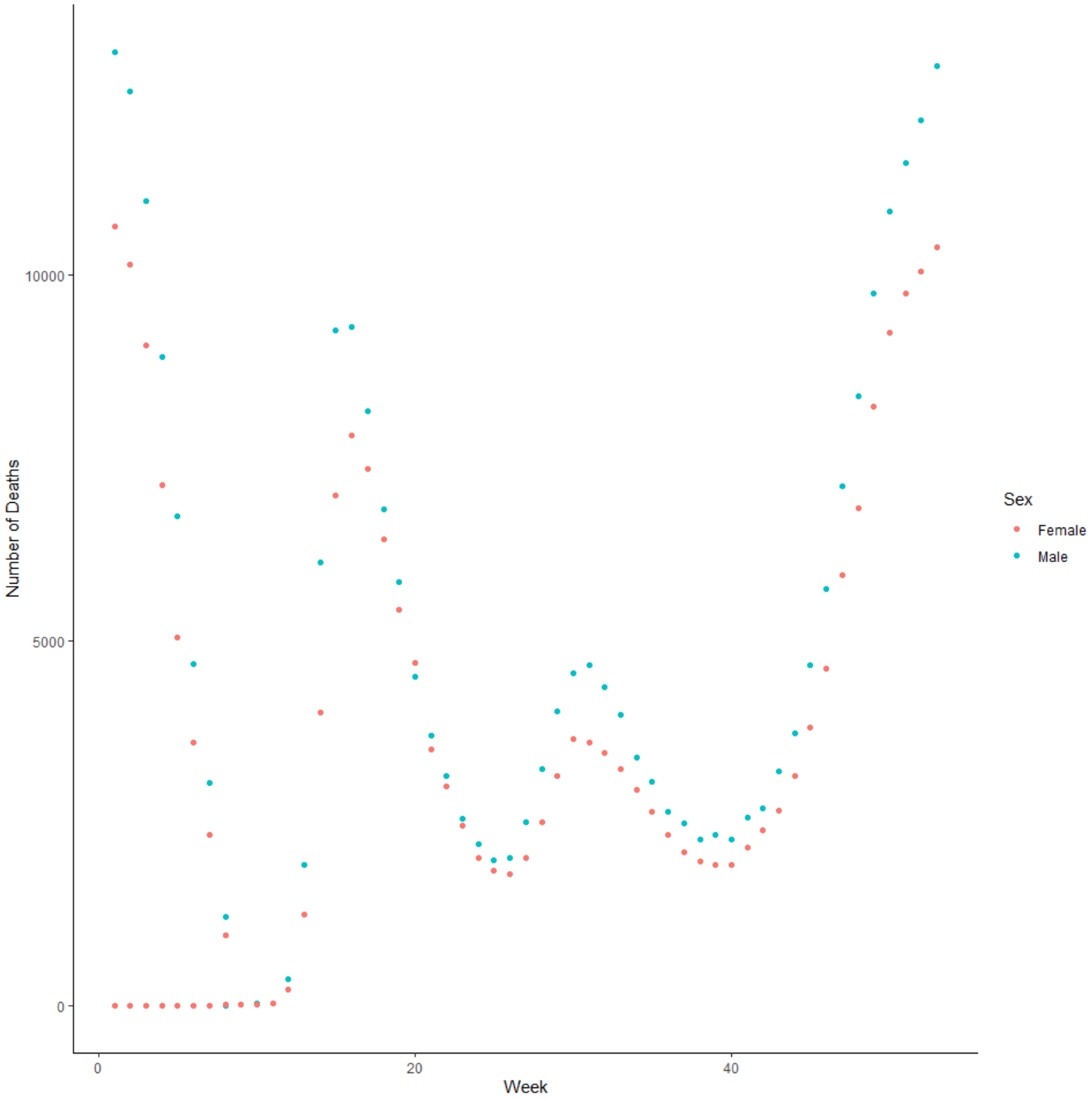


Figure 4. COVID-19 deaths by week for males and females in the United States.

Further Analysis on the time-series

```
testDataM <- filter(q31df, Sex == "Male")
testDataF <- filter(q31df, Sex == "Female")
```

```
ggplot(q31df, aes(x = MMWR.Week, y = COVID.19.Deaths, color = Sex)) + geom_point() +
        ylab("Number of Deaths") + xlab("Week") + geom_hline(aes(yintercept =
        mean(testDataM$COVID.19.Deaths), color = "Male Average Weekly Deaths")) +
        geom_hline(aes(yintercept = mean(testDataF$COVID.19.Deaths), color = "Female
        Average Weekly Deaths")) + theme_classic()
```

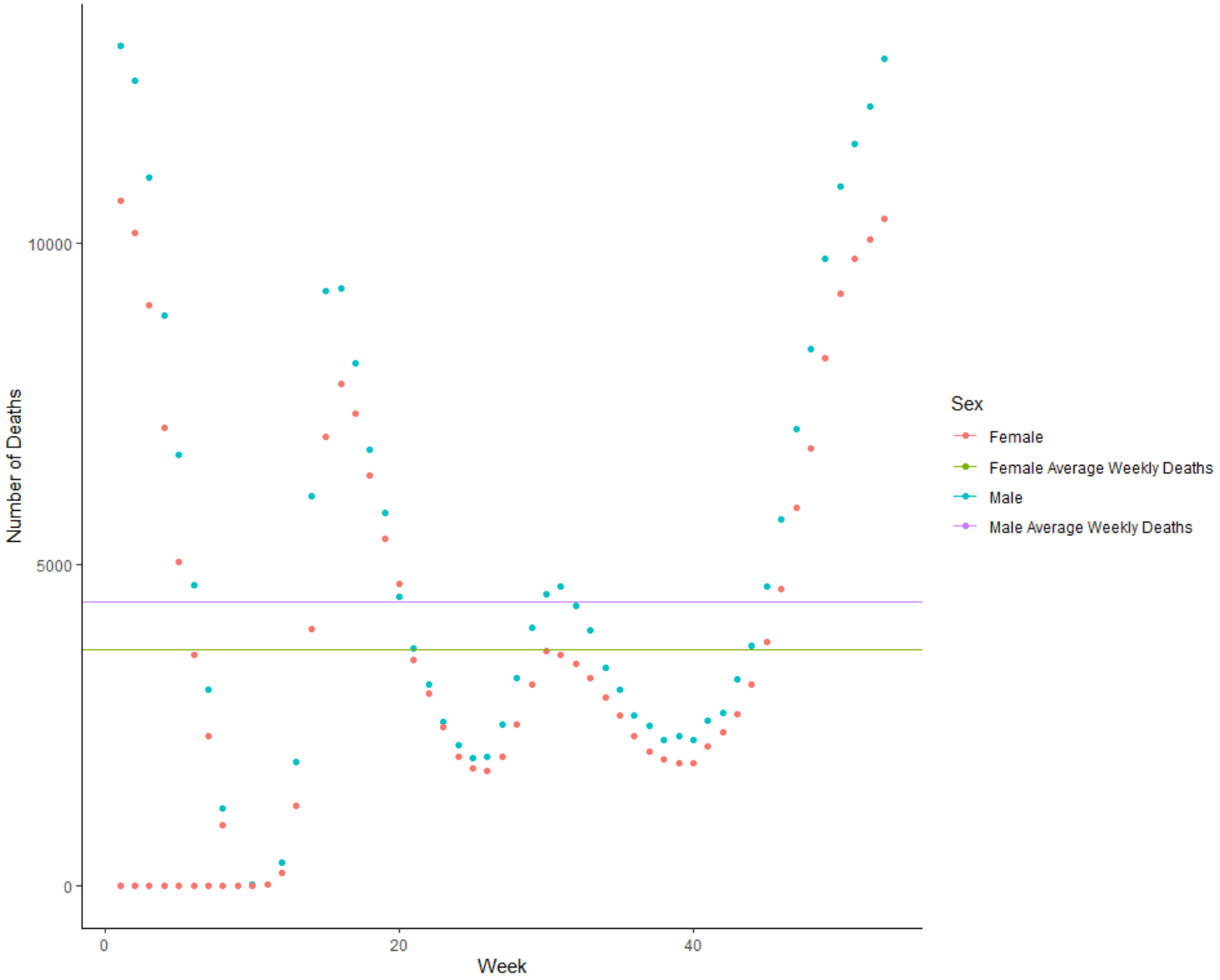


Figure 5. COVID-19 deaths by week for males and females in the United States, with average weekly death as a y-intercept included.

```
ttest <- t.test(testDataM$COVID.19.Deaths, testDataF$COVID.19.Deaths)
print(paste("The t.test in the difference in means revealed a p-value of", ttest$p.value))
```

```
[1] "The t.test in the difference in means revealed a p-value of 0.237876911800083"
```

Since the time-series for males and females followed the same momemtum, I decided to use a t.test to investigate if the mean weekly deaths are significantly different between the sexes. As mentioned above, the p-value reveals no significant difference between the means. However, if these data were reproduced for 1000 times more counts, the results could reveal a different result.

Question 4

```
q41df <- filter(deathsByCondition, Group == "By Total" & State == "United States" &
         Age.Group != "All Ages" & Age.Group != "Not stated")

sums <- as.data.frame(aggregate(COVID.19.Deaths ~ Age.Group, q41df, sum))
props <- format(prop.table(sums$COVID.19.Deaths), nsmall =4)

label_facet <- function(original_var, custom_name){
  lev <- levels(as.factor(original_var))
  lab <- paste0("(",custom_name, "%)", ": ", lev)
  names(lab) <- lev
  return(lab)
}




ggplot(q41df) + geom_col(aes(Condition.Group, COVID.19.Deaths, fill = Condition.Group)) +
         facet_wrap(~Age.Group, labeller = labeller(Age.Group =
         label_facet(q41df$Age.Group, props)), nrow = 4) + xlab(" ") + ylab("Number of
         Deaths") + theme(text = element_text(size = 16), axis.title.x = element_blank(),
         axis.text.x = element_blank(), axis.ticks.x = element_blank())
```

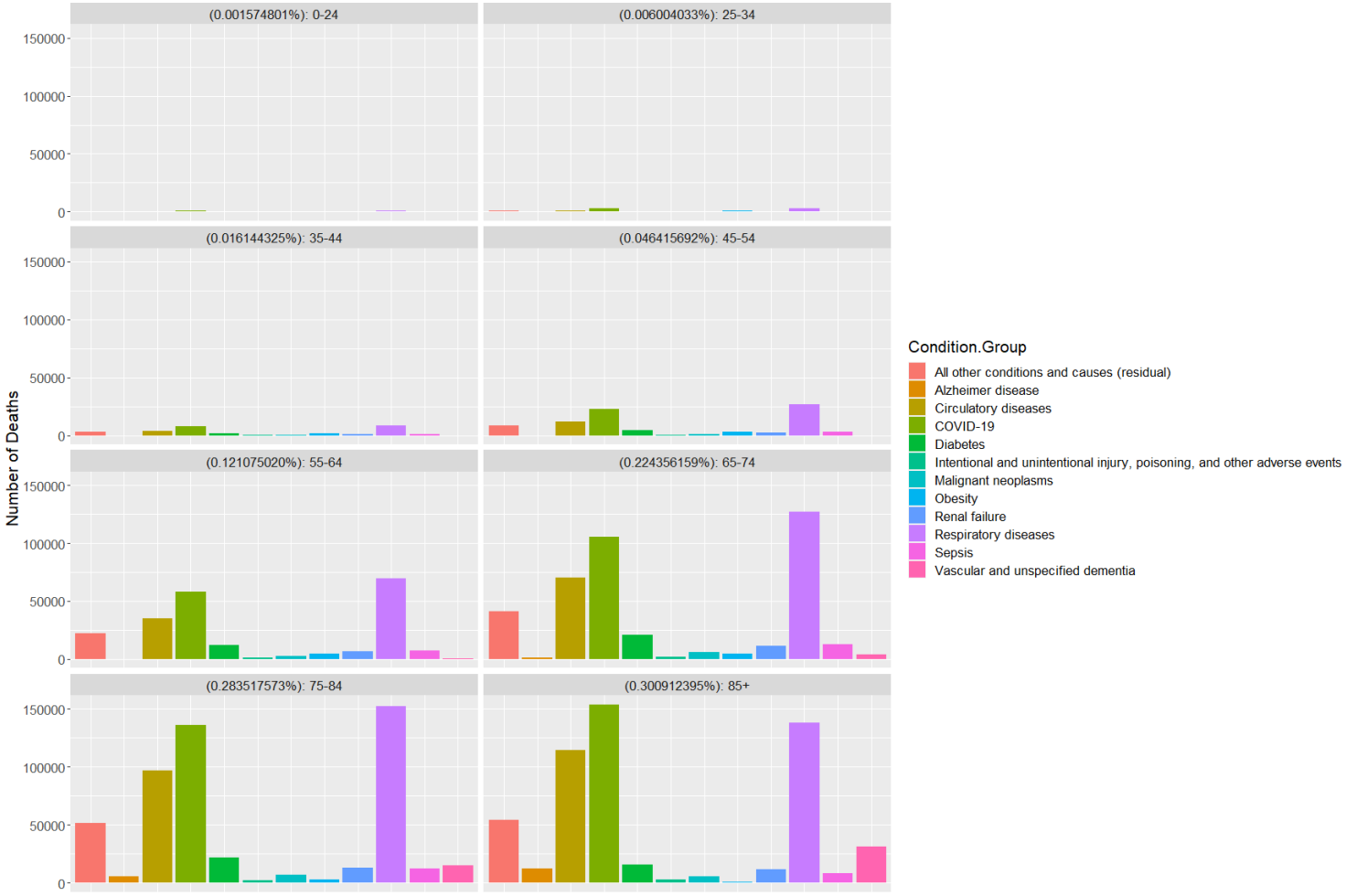


Figure 6. Comparison of COVID-19 deaths with different associated health conditions by age group.

The health conditions most commonly associated with deaths from COVID-19 are Alzheimer disease, circulatory diseases, respiratory diseases, and diabetes (Figure 6.). The most COVID-19 deaths for these condtions occur in all the age groups of individuals 55 years and older. The group most susceptible to die from COVID-19 are those 85 and older. As seen in the facet labels, about 30% of all COVID-19 deaths are those 85 years or older.

Question 5

```
condCalif <- select(filter(deathsByCondition, State == "California" & Group == "By Total" &
        COVID.19.Deaths != "NA" & Age.Group == "All Ages"), State, Condition.Group,
        Age.Group, COVID.19.Deaths)

condNY <- select(filter(deathsByCondition, State == "New York" & Group == "By Total" &
        COVID.19.Deaths != "NA" & Age.Group == "All Ages"), State, Condition.Group,
        Age.Group, COVID.19.Deaths)

names(condCalif)[names(condCalif) == "COVID.19.Deaths"] <- "COVID-19_Death_CA"
names(condNY)[names(condNY) == "COVID.19.Deaths"] <- "COVID-19_Death_NY"

comb <- inner_join(condCalif, condNY, by = "Condition.Group")
```

```
comb1 <- distinct(filter(comb, Condition.Group == "Respiratory diseases" | Condition.Group
        == "Circulatory diseases"), Condition.Group, .keep_all = T)

meltedcomb1 <- melt(comb1, measure.vars = c("COVID-19_Death_NY", "COVID-19_Death_CA"))

ggplot(meltedcomb1) + geom_col(aes(variable, value, fill = variable)) +
        facet_wrap(~Condition.Group) + xlab(" ") + ylab("Number of Deaths") + theme(text =
        element_text(size = 16), axis.title.x = element_blank(), axis.text.x =
        element_blank(), axis.ticks.x = element_blank())
```

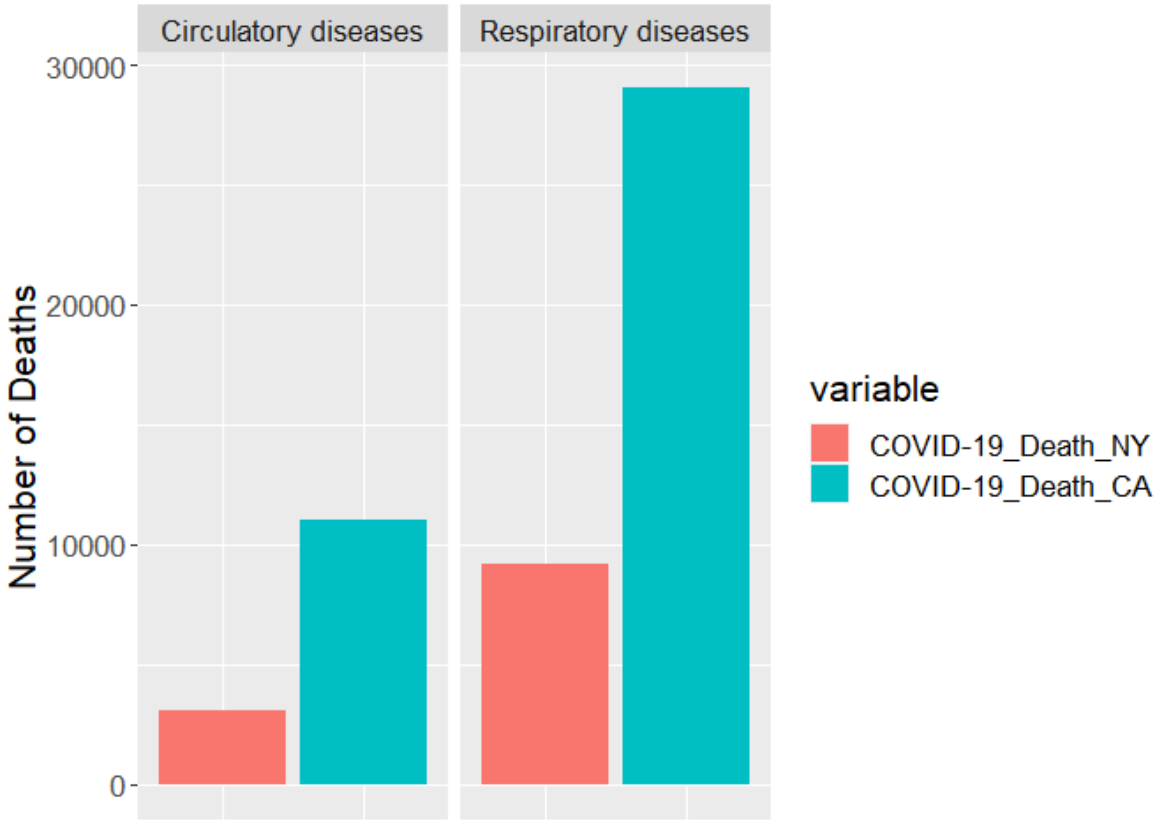


Figure 7. Comparing COVID-19 deaths in New York (red) and California (blue) by condition.

It is clear that Caliornia experienced a greater magnitude of COVID-19 deaths given these two conditions. It would be interesting to compare the number of COVID-19 deaths to the population age groups in each state. I suspect since California has many more retirees that immigrate, the population of individuals 65 or older is much greater than the population of individuals in New York who are 65 or older.

References

1. Technical Notes↩

2. Index of COVID-19 Suveillance and Ad-Hoc Data Files↩