COVID Mortality Analysis Project

Dane Pearson, Dhriti Avala

2025-12-02

Statistical Analysis of United States COVID-19 Data

Install packages:

```
library(dplyr)
library(lubridate)
library(ggplot2)
library(tidyr)
library(maps)
```

Load data:

```
original_df <- read.csv(
  "COVID-19_Deaths_by_Sex_Age.csv",
  stringsAsFactors = TRUE
)

state_populations_data <- read.csv(
  "state_populations.csv",
  stringsAsFactors = TRUE
)
```

EDA:

First convert Covid-19 Death column from factor to numeric, and remove commas in numbers:

```r
original_df$COVID.19.Deaths <- as.numeric(gsub(
  ",",
  "",
  original_df$COVID.19.Deaths
))

# Ensure it was converted correctly
# class(original_df$COVID.19.Deaths)
```

Check for missing values:

```r
# names(original_df)

# Count num of missing values in the columns we care about
# (Some columns we will use have missing values intentionally based on schema)
colSums(is.na(original_df[, c(
  "Group",
  "State",
  "Sex",
  "Age.Group",
  "COVID.19.Deaths"
)]))
```

```
          Group           State             Sex       Age.Group COVID.19.Deaths
              0               0               0               0           39430
```

I have chosen not to remove the missing values (NAs) in "COVID.19.Deaths" because the CDC intentionally "suppresses" (replaces with NA) populations that are too small for confidentiality purposes. I do not want to create bias in the data by removing data from smaller populations.

Verify no "Month" values fall outside 1-12:

```r
value_range <- range(original_df$Month, na.rm = TRUE)
print(paste(
  "Range of 'Month' column values:",
  value_range[1],
  "to",
```

```
  value_range[2]
))
```

```
[1] "Range of 'Month' column values: 1 to 12"
```

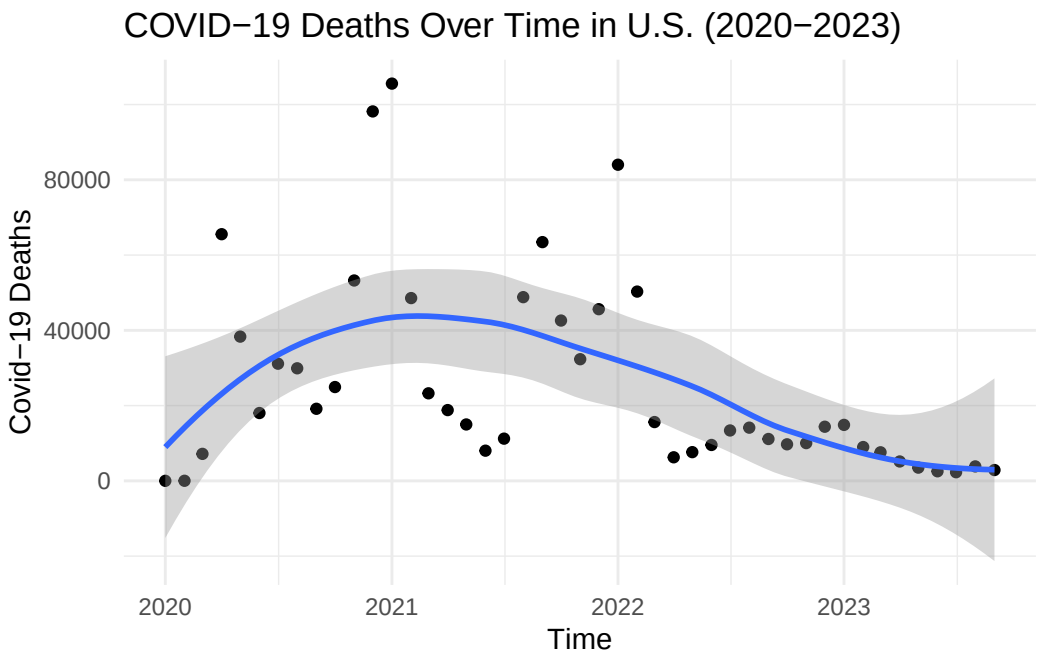
Ensure no negative values for Covid-19 Deaths:

```
min_value <- min(original_df$COVID.19.Deaths)
print(paste("Minimum value:", min_value))
```

```
[1] "Minimum value: NA"
```

Plot total COVID-19 deaths over time in the U.S.

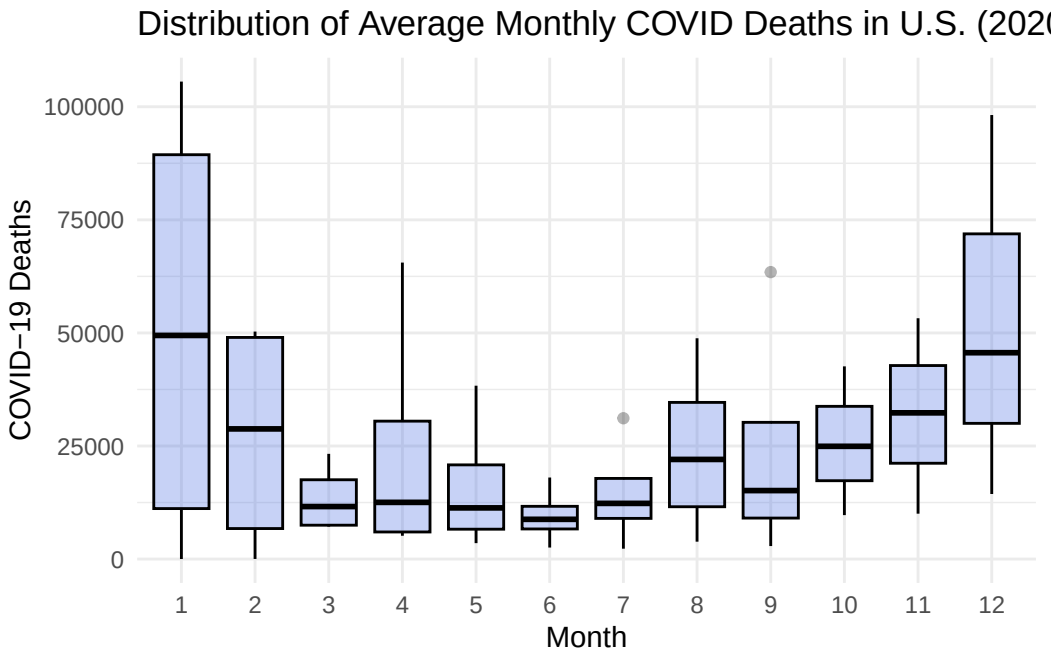
```
deaths_per_month_df <- original_df %>%
  filter(
    Group == "By Month" &
      Sex == "All Sexes" &
      Age.Group == "All Ages" &
      State == "United States"
  ) %>%
  select(Year, Month, State, Sex, Age.Group, COVID.19.Deaths) %>%
  mutate(
    MonthDate = ymd(paste(Year, Month, "01", sep = "-"))
  )

# deaths_per_month_df

ggplot(data = deaths_per_month_df, aes(x = MonthDate, y = COVID.19.Deaths)) +
  geom_point() +
  geom_smooth() +
  labs(
    title = "COVID-19 Deaths Over Time in U.S. (2020-2023)",
    x = "Time",
    y = "Covid-19 Deaths"
  ) +
  theme_minimal()
```



Deaths from COVID-19 spiked drastically in late 2020 and early 2021, with the death rate lowering in mid 2021 and leveling off by early 2023. By mid-2021, the US began administering booster shots, over 70% of the US had received at least one dose of the vaccine, and 60% were fully vaccinated. This could help to explain how the national death rate began declining at about the same time, since the vaccine has been an effective way to prevent covid-related hospitalizations and transmission of the virus.

Did certain months have higher average deaths across all years?

```
# Boxplot for each month across all years in the dataset
ggplot(
  data = deaths_per_month_df,
  aes(x = factor(Month), y = COVID.19.Deaths)
) +
  geom_boxplot(color = "black", fill = "royalblue", alpha = 0.3) +
  labs(
    title = "Distribution of Average Monthly COVID Deaths in U.S. (2020-2023)",
    x = "Month",
    y = "COVID-19 Deaths"
  ) +
  theme_minimal()
```



Based on the boxplot, we can see that the average deaths per month across all years (2020-2023) peaked in January and December. This indicates a clear pattern of there being higher death rates in colder months, suggesting that the cold weather increases likelihood and spread of the virus. This may be due to the cold and dry air helping the virus survive longer and compromising people's immune systems, in addition to people spending more time indoors in close contact, allowing it to spread more easily.

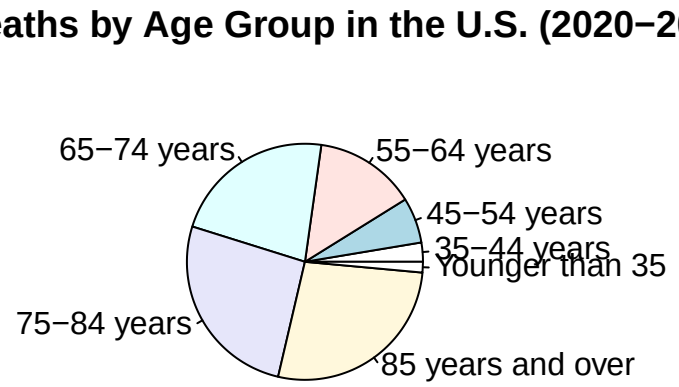
How did deaths vary across age groups in the US?

```
deaths_by_ages <- original_df %>%
  filter(
    Group == "By Total" &
      Sex == "All Sexes" &
      State == "United States" &
      Age.Group != "All Ages"
  ) %>%
  select(Age.Group, COVID.19.Deaths)

# I realized there are overlapping age groups so I have to remove the redundant ones
deaths_by_ages_clean <- deaths_by_ages %>%
  filter(
    !Age.Group %in%
```

```
      c(
        "0-17 years",
        "18-29 years",
        "30-39 years",
        "40-49 years",
        "50-64 years"
      )
  )

deaths_by_ages_clean
```

```
           Age.Group COVID.19.Deaths
1       Under 1 year             519
2          1-4 years             285
3          5-14 years            509
4         15-24 years           3021
5         25-34 years          12401
6         35-44 years          30108
7         45-54 years          71388
8         55-64 years         159712
9         65-74 years         256806
10        75-84 years         300162
11 85 years and over          311863
```

```
# Add threshold to group together smaller slices
threshold <- 0.02 * sum(deaths_by_ages_clean$COVID.19.Deaths)

# Group the smaller slices
deaths_by_ages_clean$group <- ifelse(
  deaths_by_ages_clean$COVID.19.Deaths < threshold,
  "Younger than 35",
  as.character(deaths_by_ages_clean$Age.Group)
)

agg <- aggregate(COVID.19.Deaths ~ group, deaths_by_ages_clean, sum)

# Plot as pie chart
pie(
  agg$COVID.19.Deaths,
  labels = agg$group,
  main = "Total Deaths by Age Group in the U.S. (2020-2023)"
)
```



Statistical Analysis:

Multiple Linear Regression:

The dataset used for the Multiple Linear Regression has one row for each unique combination of Year × Month × State × Sex × Age Group.

```
monthly_age_sex_deaths <- original_df %>%
  filter(Group == "By Month") %>%
  distinct(Year, Month, State, Sex, Age.Group, .keep_all = TRUE) %>%
  select(Group, Year, Month, State, Sex, Age.Group, COVID.19.Deaths)

# str(monthly_age_sex_deaths)

mlr_model <- lm(
  COVID.19.Deaths ~ Age.Group + Sex + State + Month + Year,
  data = monthly_age_sex_deaths
)

summary(mlr_model)
```

```
Call:
lm(formula = COVID.19.Deaths ~ Age.Group + Sex + State + Month +
    Year, data = monthly_age_sex_deaths)

Residuals:
   Min     1Q Median     3Q    Max
 -2928    -70     21     85 102667

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                71381.3381  6832.0143  10.448  < 2e-16 ***
Age.Group1-4 years             4.0277    19.7679   0.204   0.8385
Age.Group15-24 years           1.4024    21.1037   0.066   0.9470
Age.Group18-29 years          -6.3890    22.0894  -0.289   0.7724
Age.Group25-34 years         -10.8493    22.5928  -0.480   0.6311
Age.Group30-39 years         -12.6140    22.9492  -0.550   0.5826
Age.Group35-44 years         -10.5877    23.1118  -0.458   0.6469
Age.Group40-49 years           1.7864    22.9118   0.078   0.9379
Age.Group45-54 years          22.6607    22.5647   1.004   0.3153
Age.Group5-14 years            4.1916    19.9835   0.210   0.8339
Age.Group50-64 years         126.6602    21.3776   5.925 3.14e-09 ***
Age.Group55-64 years          93.0336    21.6358   4.300 1.71e-05 ***
Age.Group65-74 years         162.7918    21.0012   7.752 9.18e-15 ***
Age.Group75-84 years         188.6531    20.6370   9.141  < 2e-16 ***
Age.Group85 years and over   194.6390    20.5031   9.493  < 2e-16 ***
Age.GroupAll Ages            676.3222    19.8157  34.131  < 2e-16 ***
Age.GroupUnder 1 year          3.5548    19.9633   0.178   0.8587
SexFemale                    -86.2035     8.9825  -9.597  < 2e-16 ***
SexMale                      -70.4190     8.9800  -7.842 4.49e-15 ***
StateAlaska                   -4.0366    38.5321  -0.105   0.9166
StateArizona                  19.2110    37.8281   0.508   0.6116
StateArkansas                -22.1971    38.0787  -0.583   0.5599
StateCalifornia              224.0761    37.3770   5.995 2.04e-09 ***
StateColorado                -22.0569    38.2132  -0.577   0.5638
StateConnecticut             -17.2656    38.1154  -0.453   0.6506
StateDelaware                -20.0495    38.2321  -0.524   0.6000
StateDistrict of Columbia    -17.0758    38.9760  -0.438   0.6613
StateFlorida                 148.7674    37.6189   3.955 7.67e-05 ***
StateGeorgia                  33.5657    37.9583   0.884   0.3765
StateHawaii                  -17.4650    38.4890  -0.454   0.6500
StateIdaho                   -19.7119    37.8406  -0.521   0.6024
StateIllinois                 36.5472    37.7533   0.968   0.3330
StateIndiana                   7.5665    37.9899   0.199   0.8421
```

```
StateIowa                   -22.6590    37.9456  -0.597   0.5504
StateKansas                 -22.0909    38.2286  -0.578   0.5634
StateKentucky                -7.6749    37.4587  -0.205   0.8377
StateLouisiana              -14.5021    38.5898  -0.376   0.7071
StateMaine                  -19.1378    37.9594  -0.504   0.6141
StateMaryland               -14.3763    38.0722  -0.378   0.7057
StateMassachusetts           -4.4530    37.9784  -0.117   0.9067
StateMichigan                34.5846    37.9498   0.911   0.3621
StateMinnesota              -20.5118    38.1414  -0.538   0.5907
StateMississippi            -19.0817    37.9883  -0.502   0.6155
StateMissouri                -0.4660    38.1211  -0.012   0.9902
StateMontana                -16.3791    37.7202  -0.434   0.6641
StateNebraska               -23.4494    38.1549  -0.615   0.5388
StateNevada                 -22.2456    38.1882  -0.583   0.5602
StateNew Hampshire          -17.1524    37.7489  -0.454   0.6496
StateNew Jersey              31.3584    37.6610   0.833   0.4050
StateNew Mexico             -31.3601    38.7457  -0.809   0.4183
StateNew York                47.4535    38.2159   1.242   0.2143
StateNew York City           40.6236    38.1456   1.065   0.2869
StateNorth Carolina          28.3945    37.6628   0.754   0.4509
StateNorth Dakota           -11.4572    38.3934  -0.298   0.7654
StateOhio                    64.5683    37.8984   1.704   0.0884 .
StateOklahoma               -10.8262    38.1649  -0.284   0.7767
StateOregon                 -31.4349    38.1264  -0.824   0.4097
StatePennsylvania            72.4253    37.8460   1.914   0.0557 .
StatePuerto Rico            -38.2134    38.8736  -0.983   0.3256
StateRhode Island           -15.7779    37.8870  -0.416   0.6771
StateSouth Carolina          -7.2531    38.0550  -0.191   0.8488
StateSouth Dakota           -17.5510    38.5195  -0.456   0.6487
StateTennessee               19.8542    37.5737   0.528   0.5972
StateTexas                  220.4549    37.6380   5.857 4.72e-09 ***
StateUnited States         2207.9685    35.0522  62.991  < 2e-16 ***
StateUtah                   -30.9477    38.7014  -0.800   0.4239
StateVermont                 -7.9599    37.3059  -0.213   0.8310
StateVirginia                -3.4589    38.0916  -0.091   0.9276
StateWashington             -19.9133    37.8666  -0.526   0.5990
StateWest Virginia          -20.9380    37.3821  -0.560   0.5754
StateWisconsin              -14.3144    38.2570  -0.374   0.7083
StateWyoming                 -1.9018    37.8510  -0.050   0.9599
Month                         0.1065     1.0821   0.098   0.9216
Year                        -35.3129     3.3796 -10.449  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 1085 on 87644 degrees of freedom
  (36212 observations deleted due to missingness)
Multiple R-squared:  0.1207,    Adjusted R-squared:   0.12
F-statistic: 164.9 on 73 and 87644 DF,  p-value: < 2.2e-16
```

Two-way ANOVA:

How do age, sex, and state influence mortality by each month?

```r
anova_model <- aov(
  COVID.19.Deaths ~ Age.Group * Sex,
  data = monthly_age_sex_deaths
)
summary(anova_model)
```

```
                 Df    Sum Sq   Mean Sq F value Pr(>F)
Age.Group        16 2.890e+09 180623701 138.812 <2e-16 ***
Sex               2 1.220e+08  60988974  46.871 <2e-16 ***
Age.Group:Sex    32 3.124e+08   9761157   7.502 <2e-16 ***
Residuals     87667 1.141e+11   1301207
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
36212 observations deleted due to missingness
```

Choropleth heatmap:

First, we need to get the populations of each state by doing a left join of US census data on the State column of the original dataset.

Below you can see the first few rows of each dataset, with the last one being the left join of the two of them with an added death rate column:

```r
# Right dataset
state_pops_clean <- state_populations_data %>%
  select(
    NAME,
```

```
    POPESTIMATE2020,
    POPESTIMATE2021,
    POPESTIMATE2022,
    POPESTIMATE2023
  ) %>%
  rename(State = NAME)

head(state_pops_clean, 3)
```

```
             State POPESTIMATE2020 POPESTIMATE2021 POPESTIMATE2022
1    United States       331526933       332048977       333271411
2 Northeast Region        57430477        57243423        57026847
3      New England        15057898        15106108        15120739
  POPESTIMATE2023
1       334914895
2        56983517
3        15159777
```

```
# Left dataset where each row is a state and its population during a year
left_df <- original_df %>%
  filter(
    Group == "By Year",
    Sex == "All Sexes",
    Age.Group == "All Ages"
  ) %>%
  select(State, Year, COVID.19.Deaths)

head(left_df, 3)
```

```
     State Year COVID.19.Deaths
1  Alabama 2020            6706
2 Colorado 2020            5073
3   Kansas 2023             428
```

```
# Make the right dataset long so it has one row for each state and year pair for compatible j
state_pops_long <- state_pops_clean %>%
  pivot_longer(
    cols = starts_with("POPESTIMATE"),
    names_to = "Year",
    values_to = "Population"
  ) %>%
```

```
  mutate(
    Year = as.integer(gsub("POPESTIMATE", "", Year)) # Removes "POPESTIMATE"
  )

# Join the two datasets using a left join
joined_df <- left_df %>%
  left_join(state_pops_long, by = c("State", "Year")) %>%
  mutate(
    deaths_per_100000 = (COVID.19.Deaths / Population) * 100000,
    State = tolower(State) # state names must be lowercase for choropleth map
  )

head(joined_df, 3)
```

```
     State Year COVID.19.Deaths Population deaths_per_100000
1  alabama 2020            6706    5031864         133.27069
2 colorado 2020            5073    5785219          87.68899
3   kansas 2023             428    2940546          14.55512
```

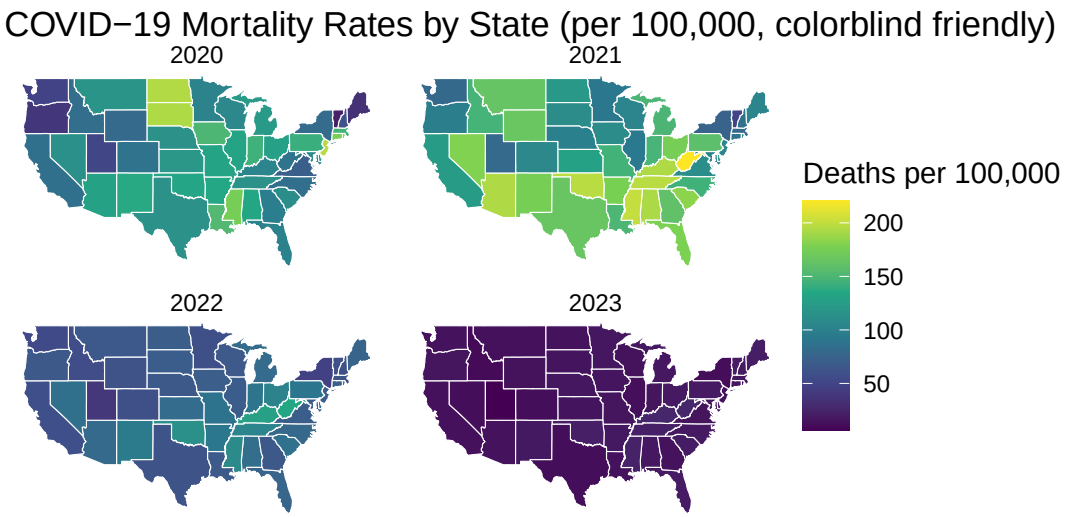
Now prepare subsets of the joined data by each year:

```
joined_2020 <- joined_df %>%
  filter(Year == "2020")

joined_2021 <- joined_df %>%
  filter(Year == "2021")

joined_2022 <- joined_df %>%
  filter(Year == "2022")

joined_2023 <- joined_df %>%
  filter(Year == "2023")
```

Finally, plot the choropleth maps across 2020-2023:

```
# Load US map geometry
us_map <- map_data("state")

# Merge the polygon map with death rate dataframe
plot_map_df <- us_map %>%
```

```r
  left_join(joined_df, by = c("region" = "State"))

# Plot the heatmaps on a 2x2 grid by year
ggplot(plot_map_df, aes(long, lat, group = group, fill = deaths_per_100000)) +
  geom_polygon(color = "white", linewidth = 0.2) +
  coord_fixed(1.3) +
  scale_fill_viridis_c(option = "viridis") +
  facet_wrap(~Year, ncol = 2) +
  labs(
    title = "COVID-19 Mortality Rates by State (per 100,000, colorblind friendly)",
    fill = "Deaths per 100,000"
  ) +
  theme_minimal() +
  theme_void() + # Remove background and axes
  theme(
    axis.title = element_blank(),
    axis.text = element_blank(),
    axis.ticks = element_blank()
  )
```



Panel data plot: