DATA 606: Data Project Proposal

Souleymane Doumbia & Fomba Kassoh

2023-10-29

Data Preparation

```
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
```

```
library(readr)
library(stringr)
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v forcats   1.0.0     v purrr     1.0.2
## v ggplot2   3.4.4     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.0

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
# load data
covid_data <- read_csv("https://raw.githubusercontent.com/hawa1983/DATA606-Project/main/covid_data1.csv
```

```
## Rows: 203224 Columns: 10
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (10): cdc_case_earliest_dt, onset_dt, current_status, sex, age_group, ra...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```r
glimpse(covid_data)
```

```
## Rows: 203,224
## Columns: 10
## $ cdc_case_earliest_dt    <chr> "10/2/2020", "10/2/2020", "10/2/2020", "10/2/2~
## $ onset_dt                <chr> "10/2/2020", NA, NA, "10/2/2020", "10/2/2020",~
## $ current_status          <chr> "Laboratory-confirmed case", "Laboratory-confi~
## $ sex                     <chr> "Male", "Male", "Male", "Male", "Male", "Male"~
## $ age_group               <chr> "10 - 19 Years", "10 - 19 Years", "10 - 19 Yea~
## $ race_ethnicity_combined <chr> "Black, Non-Hispanic", "Black, Non-Hispanic", ~
## $ hosp_yn                 <chr> "No", "No", "No", "No", "No", "No", "No", "No"~
## $ icu_yn                  <chr> "No", "Missing", "Missing", "Missing", "Missin~
## $ death_yn                <chr> "No", "No", "No", "No", "No", "No", "No", "No"~
## $ medcond_yn              <chr> "No", "Missing", "Missing", "Missing", "Missin~
```

Research question

You should phrase your research question in a way that matches up with the scope of inference your dataset allows for.

Among individuals tested for a specific condition in October 2020, how does age group, sex, and race influence the likelihood of requiring hospitalization? Specifically, are certain age groups, genders, or racial/ethnic groups at a higher risk of severe outcomes as indicated by hospitalization?

Cases

What are the cases, and how many are there? Each case represents a Covid-19 patient in the united states. There 203,224 observations in the given data set.

Data collection

Describe the method of data collection. Data is collected by the Center for Disease Control and Prevention (CDC). Data is submitted by hospitals.

Type of study

What type of study is this (observational/experiment)? This is an observational study.

Data Source

If you collected the data, state self-collected. If not, provide a citation/link. Data is collected by CDC and is available online here: https://data.cdc.gov/Case-Surveillance/COVID-19-Case-Surveillance-Public-Use-Data/vbim-akqf. For this project, data was extracted by downloading the csv format table.

Dependent Variable

What is the response variable? Is it quantitative or qualitative?

The response variable is "Death Status (Did the patient die as a result of covid?)," which is a categorical variable. It is recorded by hospitals that admit covid patients.

Independent Variable(s)

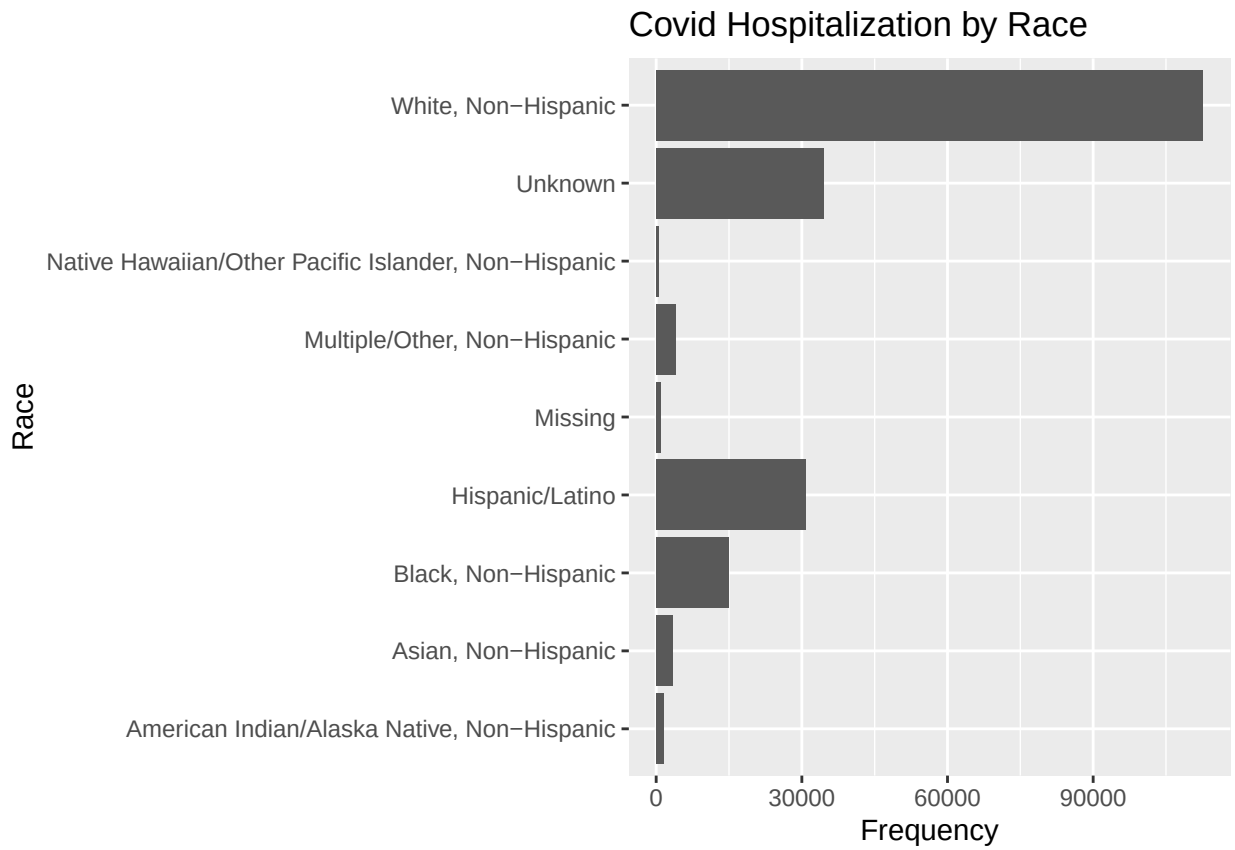
The independent variables are race and gender

Relevant summary statistics

Provide summary statistics for each the variables. Also include appropriate visualizations related to your research question (e.g. scatter plot, boxplots, etc). This step requires the use of R, hence a code chunk is provided below. Insert more code chunks as needed.

Covid hospitalization by race

```
ggplot(covid_data, aes(race_ethnicity_combined)) +
  geom_bar() +
  labs(title = "Covid Hospitalization by Race",
       x = "Race",
       y = "Frequency") +
  coord_flip()
```



Proportion of Covid infection by gender

```
covid_data |>
  count(sex)|>
  mutate(proportion = round(n/sum(n),6))
```

```
## # A tibble: 5 x 3
##   sex          n proportion
##   <chr>    <int>      <dbl>
## 1 Female  106350   0.523
## 2 Male     96203   0.473
## 3 Missing    115   0.000566
## 4 Other        2   0.00001
## 5 Unknown    554   0.00273
```

Proportion pf Icu admission

```
covid_data |>
  count(icu_yn)|>
  mutate(proportion = round(n/sum(n),6))
```

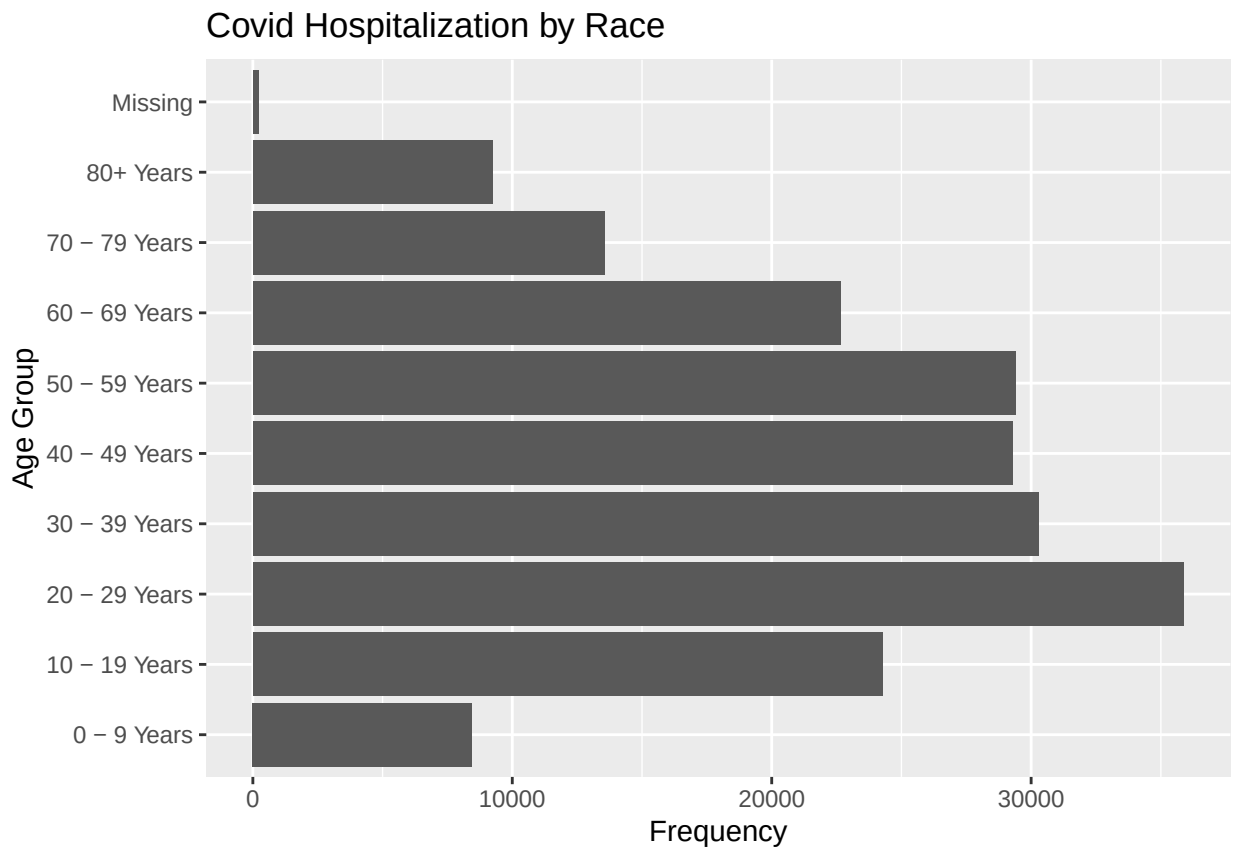
```
## # A tibble: 4 x 3
##   icu_yn       n proportion
##   <chr>    <int>      <dbl>
## 1 Missing 155239    0.764
## 2 No       44204    0.218
## 3 Unknown   2203    0.0108
## 4 Yes       1578    0.00776
```

Distribution of Covid infection by age group

```
covid_data |>
  count(age_group)|>
  mutate(proportion = round(n/sum(n),6))
```

```
## # A tibble: 10 x 3
##    age_group         n proportion
##    <chr>         <int>      <dbl>
##  1 0 - 9 Years    8442    0.0415
##  2 10 - 19 Years 24279    0.119
##  3 20 - 29 Years 35886    0.177
##  4 30 - 39 Years 30289    0.149
##  5 40 - 49 Years 29279    0.144
##  6 50 - 59 Years 29393    0.145
##  7 60 - 69 Years 22644    0.111
##  8 70 - 79 Years 13558    0.0667
##  9 80+ Years      9230    0.0454
## 10 Missing         224    0.00110
```

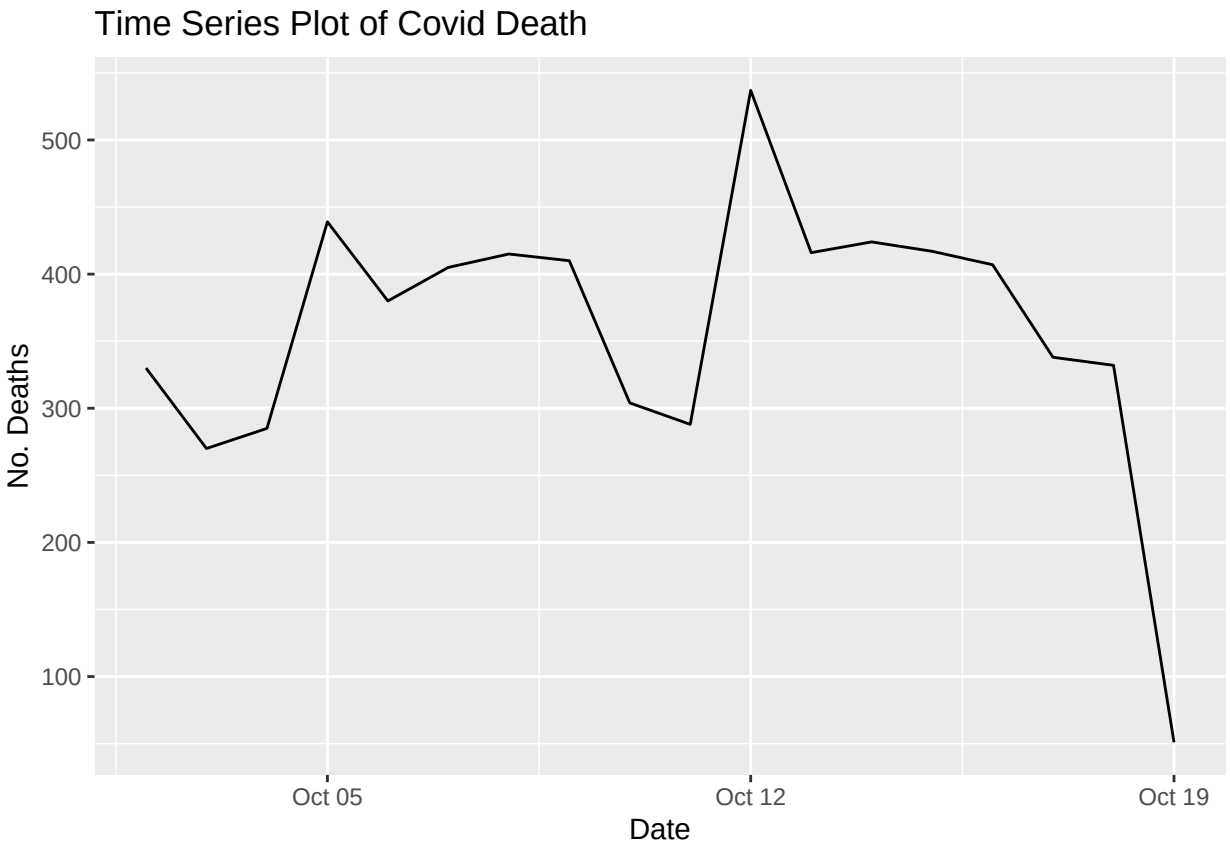
```
ggplot(covid_data, aes(age_group)) +
  geom_bar() +
  labs(title = "Covid Hospitalization by Race",
       x = "Age Group",
       y = "Frequency") +
  coord_flip()
```



Time series of covid deaths.

```
death_yes <- covid_data %>%
  filter(death_yn == "Yes") %>%
  mutate(cdc_case_earliest_dt = as.Date(cdc_case_earliest_dt, format = "%m/%d/%Y")) %>%
  count(cdc_case_earliest_dt) %>%
  arrange(cdc_case_earliest_dt)

# Assuming you have a data frame named 'mydata'
ggplot(death_yes, aes(x = cdc_case_earliest_dt, y = n)) +
  geom_line() +           # Line plot
  labs(x = "Date", y = "No. Deaths") +  # Labels for axes
  ggtitle("Time Series Plot of Covid Death")   # Title for the plot
```


Time Series Plot of Covid Death
No. Deaths
500
400
300
200
100
Oct 05
Oct 12
Oct 19
Date