STAT 597 - COVID DATA: Preliminary Exploratory Analysis

Rumil

9/29/2021

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.2     v dplyr   1.0.7
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
covid <- read_csv("C:/Users/RUMIL/Desktop/APU/STAT 597 - Soeun Kim (Statistical Consulting)/COVID19 597
```

```
##
## -- Column specification --------------------------------------------------------
## cols(
##   cdc_report_dt = col_character(),
##   pos_spec_dt = col_character(),
##   current_status = col_character(),
##   sex = col_character(),
##   age_group = col_character(),
##   `Race and ethnicity (combined)` = col_character(),
##   hosp_yn = col_character(),
##   icu_yn = col_character(),
##   death_yn = col_character(),
##   medcond_yn = col_character()
## )
```

```
glimpse(covid)
```

```
## Rows: 3,522
## Columns: 10
## $ cdc_report_dt                  <chr> "8/16/20", "8/16/20", "8/16/20", "8/16~
## $ pos_spec_dt                    <chr> "8/5/20", "8/10/20", "8/10/20", "8/10/~
## $ current_status                 <chr> "Laboratory-confirmed case", "Laborato~
## $ sex                            <chr> "Female", "Male", "Female", "Male", "F~
## $ age_group                      <chr> "0 - 9 Years", "0 - 9 Years", "0 - 9 Y~
```

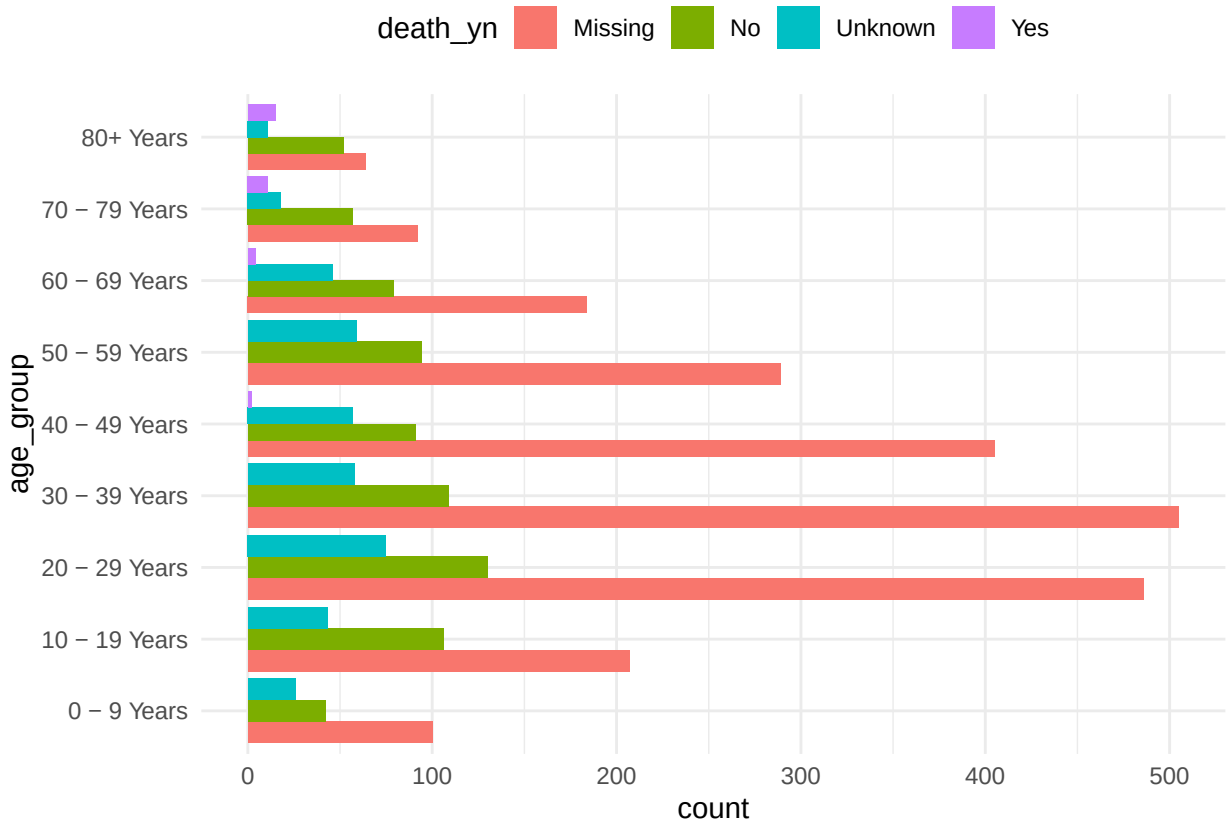
```
## $ `Race and ethnicity (combined)` <chr> "White, Non-Hispanic", "Asian, Non-His~
## $ hosp_yn                         <chr> "Missing", "Missing", "Missing", "Unkn~
## $ icu_yn                          <chr> "Missing", "Missing", "Missing", "Unkn~
## $ death_yn                        <chr> "Missing", "Missing", "No", "Unknown",~
## $ medcond_yn                      <chr> "Missing", "Missing", "No", "Unknown",~
```

```
covid %>% select(
  current_status,
  sex,
  age_group,
  `Race and ethnicity (combined)`,
  hosp_yn,
  icu_yn,
  death_yn,
  medcond_yn
) %>% as_factor()
```

```
## # A tibble: 3,522 x 8
##    current_status   sex    age_group `Race and ethnicit~ hosp_yn icu_yn death_yn
##    <chr>            <chr>  <chr>     <chr>               <chr>   <chr>  <chr>
##  1 Laboratory-conf~ Female 0 - 9 Ye~ White, Non-Hispanic Missing Missi~ Missing
##  2 Laboratory-conf~ Male   0 - 9 Ye~ Asian, Non-Hispanic Missing Missi~ Missing
##  3 Laboratory-conf~ Female 0 - 9 Ye~ Black, Non-Hispanic Missing Missi~ No
##  4 Laboratory-conf~ Male   0 - 9 Ye~ Hispanic/Latino     Unknown Unkno~ Unknown
##  5 Laboratory-conf~ Female 0 - 9 Ye~ White, Non-Hispanic Missing Missi~ Missing
##  6 Laboratory-conf~ Female 0 - 9 Ye~ Hispanic/Latino     No      Unkno~ No
##  7 Laboratory-conf~ Male   0 - 9 Ye~ Hispanic/Latino     Unknown Unkno~ Unknown
##  8 Laboratory-conf~ Female 0 - 9 Ye~ American Indian/Al~ Missing Missi~ Missing
##  9 Laboratory-conf~ Male   0 - 9 Ye~ White, Non-Hispanic Missing Missi~ No
## 10 Laboratory-conf~ Male   0 - 9 Ye~ Multiple/Other, No~ Unknown Unkno~ Unknown
## # ... with 3,512 more rows, and 1 more variable: medcond_yn <chr>
```

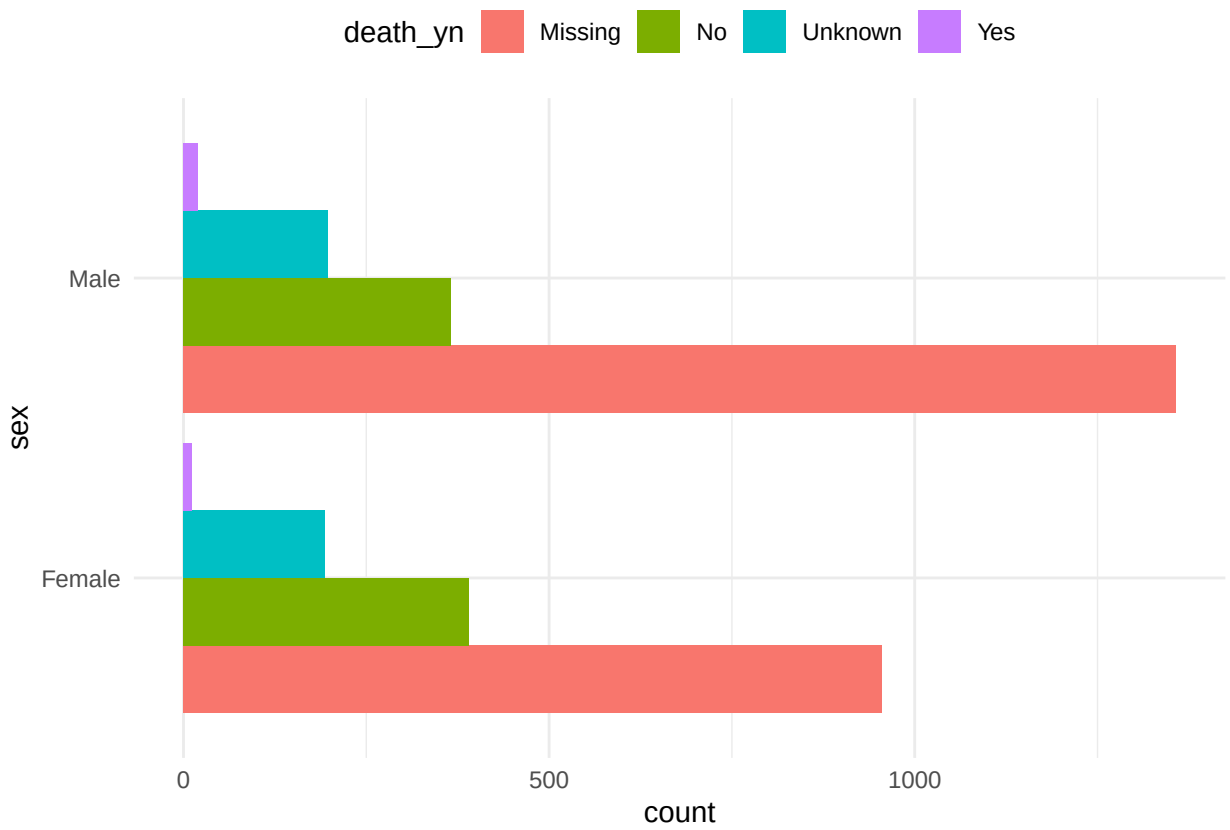
Histograms

```
#Deaths by Age group
covid %>%  filter(age_group != "Unknown") %>%
ggplot() +
  aes(x = age_group, fill = death_yn) +
  geom_bar(position = "dodge")+
  scale_fill_hue(direction = 1) +
  theme_minimal() +
  theme(legend.position = "top") + coord_flip()
```



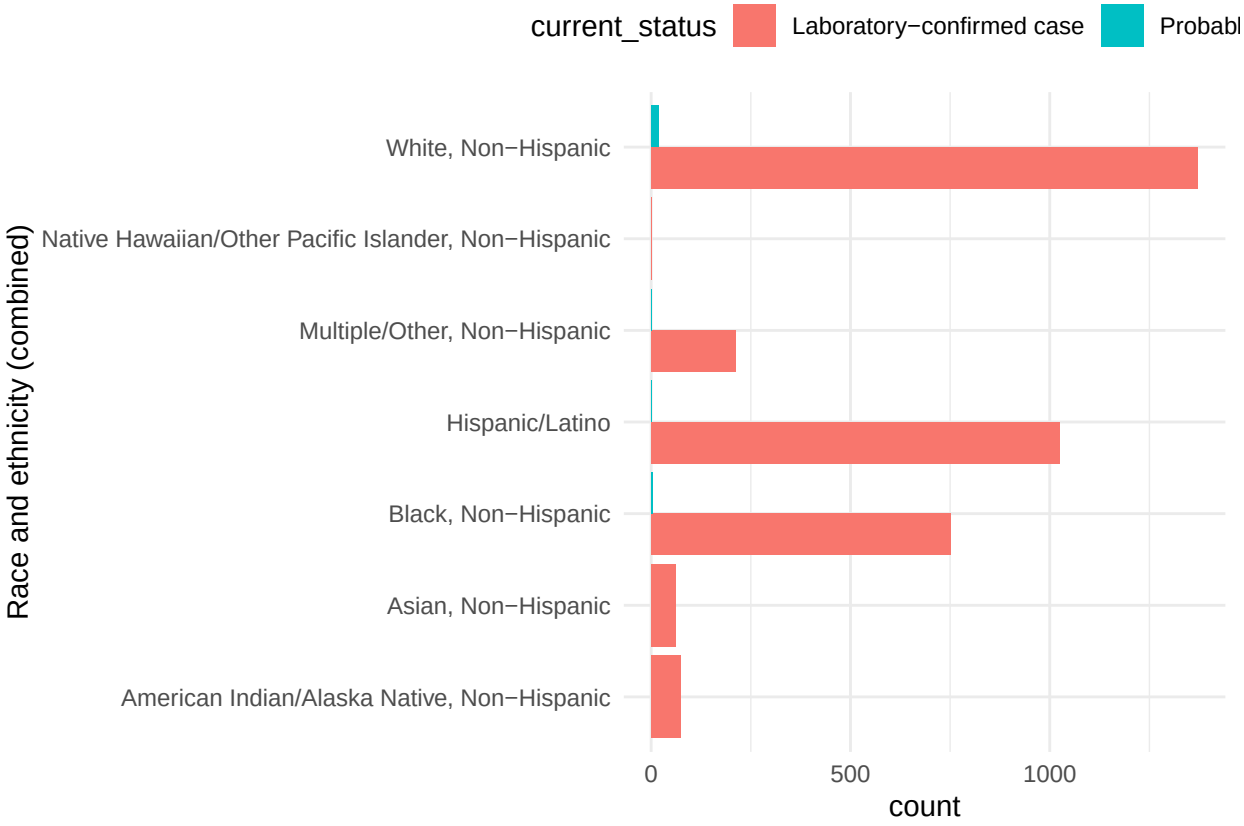
Notice the amount of "Missing" there are when filling out these forms. I'd be curious where most of these "Missing" observations are coming from

```
#Deaths by Age group
 covid  %>% filter(sex != "Unknown", sex != "Missing") %>%
ggplot() +
  aes(x = sex, fill = death_yn) +
  geom_bar(position = "dodge")+
  scale_fill_hue(direction = 1) +
  theme_minimal() +
  theme(legend.position = "top") + coord_flip()
```

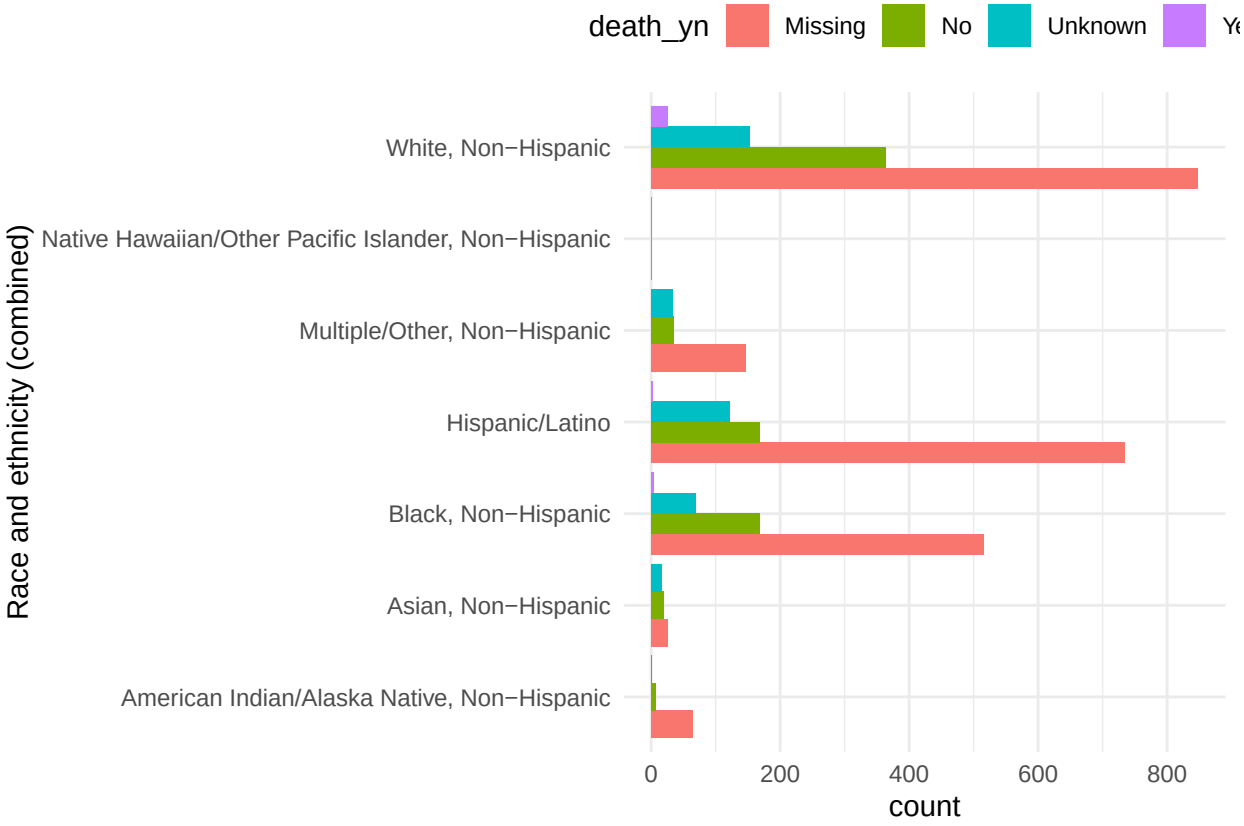


The amount of missing cases is curious.

```
#current status by race/ethnicity
covid  %>%
ggplot() +
  aes(x = `Race and ethnicity (combined)`, fill = current_status) +
  geom_bar(position = "dodge")+
  scale_fill_hue(direction = 1) +
  theme_minimal() +
  theme(legend.position = "top") + coord_flip()
```




```
#death status by race/ethnicity
covid  %>%
ggplot() +
  aes(x = `Race and ethnicity (combined)`, fill = death_yn) +
  geom_bar(position = "dodge")+
  scale_fill_hue(direction = 1) +
  theme_minimal() +
  theme(legend.position = "top") + coord_flip()
```


death_yn
Missing
No
Unknown
Ye
Race and ethnicity (combined)
White, Non−Hispanic
Native Hawaiian/Other Pacific Islander, Non−Hispanic
Multiple/Other, Non−Hispanic
Hispanic/Latino
Black, Non−Hispanic
Asian, Non−Hispanic
American Indian/Alaska Native, Non−Hispanic
0
200
400
600
800
count