illnesses.qmd

```
library(tidyverse)
```

```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.1.3     v readr     2.1.4
v forcats   1.0.0     v stringr   1.5.0
v ggplot2   3.4.3     v tibble    3.2.1
v lubridate 1.9.2     v tidyr     1.3.0
v purrr     1.0.2
-- 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 becom
```

```
library(tidymodels)
```

```
-- Attaching packages -------------------------------------- tidymodels 1.1.1 --
v broom        1.0.5     v rsample      1.2.0
v dials        1.2.0     v tune         1.1.2
v infer        1.0.5     v workflows    1.1.3
v modeldata    1.2.0     v workflowsets 1.0.1
v parsnip      1.1.1     v yardstick    1.2.0
v recipes      1.0.8
-- Conflicts ----------------------------------------- tidymodels_conflicts() --
x scales::discard() masks purrr::discard()
x dplyr::filter()   masks stats::filter()
x recipes::fixed()  masks stringr::fixed()
x dplyr::lag()      masks stats::lag()
x yardstick::spec() masks readr::spec()
x recipes::step()   masks stats::step()
* Dig deeper into tidy modeling with R at https://www.tmwr.org
```

```r
covid <- read_csv("Covid Data.csv")
```

```
Rows: 1048575 Columns: 21
-- Column specification --------------------------------------------------------
Delimiter: ","
chr  (1): DATE_DIED
dbl (20): USMER, MEDICAL_UNIT, SEX, PATIENT_TYPE, INTUBED, PNEUMONIA, AGE, P...

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.
```

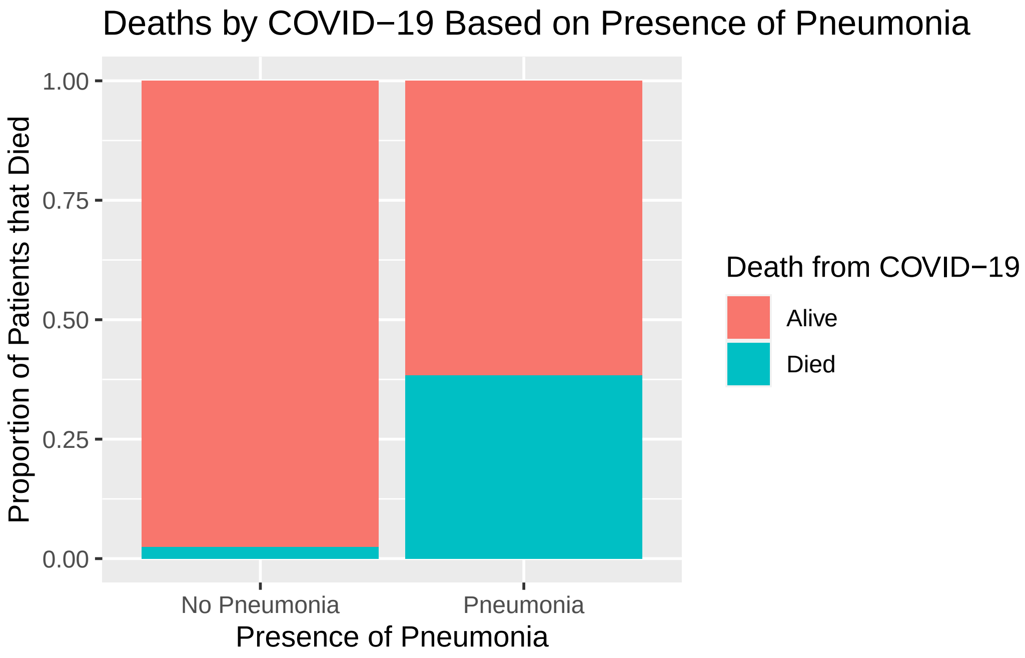
```r
# 0 = did not die
# 1 = died
covid_health <- covid |>
  mutate(
    died = if_else(DATE_DIED == "9999-99-99", 0, 1)
  ) |>
  filter(PNEUMONIA != 97 & PNEUMONIA != 99 & PNEUMONIA != 98) |>
  filter(DIABETES != 97 & DIABETES != 99 & DIABETES != 98) |>
  filter(INMSUPR != 97 & INMSUPR != 99 & INMSUPR != 98) |>
  filter(HIPERTENSION != 97 & HIPERTENSION != 99 & HIPERTENSION != 98) |>
  filter(OTHER_DISEASE != 97 & OTHER_DISEASE != 99 & OTHER_DISEASE != 98) |>
  filter(CARDIOVASCULAR != 97 & CARDIOVASCULAR != 99 & CARDIOVASCULAR != 98) |>
  filter(RENAL_CHRONIC != 97 & RENAL_CHRONIC != 99 & RENAL_CHRONIC != 98) |>
  filter(ASTHMA != 97 & ASTHMA != 99 & ASTHMA != 98) |>
  filter(AGE != 97 & AGE != 99 & AGE != 98) |>
  filter(SEX != 97 & SEX != 99 & SEX != 98)

covid_health
```

```
# A tibble: 1,025,722 x 22
   USMER MEDICAL_UNIT   SEX PATIENT_TYPE DATE_DIED  INTUBED PNEUMONIA   AGE
   <dbl>        <dbl> <dbl>        <dbl> <chr>        <dbl>     <dbl> <dbl>
 1     2            1     1            1 03/05/2020      97         1    65
 2     2            1     2            1 03/06/2020      97         1    72
 3     2            1     2            2 09/06/2020       1         2    55
 4     2            1     1            1 12/06/2020      97         2    53
 5     2            1     2            1 21/06/2020      97         2    68
 6     2            1     1            2 9999-99-99       2         1    40
 7     2            1     1            1 9999-99-99      97         2    64
 8     2            1     1            1 9999-99-99     97          1    64
 9     2            1     1            2 9999-99-99      2          2    37
10     2            1     1            2 9999-99-99      2          2    25
# i 1,025,712 more rows
# i 14 more variables: PREGNANT <dbl>, DIABETES <dbl>, COPD <dbl>,
#   ASTHMA <dbl>, INMSUPR <dbl>, HIPERTENSION <dbl>, OTHER_DISEASE <dbl>,
#   CARDIOVASCULAR <dbl>, OBESITY <dbl>, RENAL_CHRONIC <dbl>, TOBACCO <dbl>,
#   CLASIFFICATION_FINAL <dbl>, ICU <dbl>, died <dbl>
```

```r
# pneumonia specifically influential bc it's a lung disease

covid_health |>
  mutate(PNEUMONIA = if_else(PNEUMONIA == 1, "Pneumonia", "No Pneumonia")) |>
  mutate(died = if_else(died == 1, "Died", "Alive")) |>
  ggplot(aes(x = PNEUMONIA, fill = died)) +
  geom_bar(position = "fill")  +
  labs(title = "Deaths by COVID-19 Based on Presence of Pneumonia", y = "Proportion of Pat
```



```r
covid_long <- covid_health |>
  select(PNEUMONIA, DIABETES, ASTHMA, INMSUPR, HIPERTENSION, OTHER_DISEASE, CARDIOVASCULAR
  pivot_longer(
    cols = c('PNEUMONIA', 'DIABETES', 'ASTHMA', 'INMSUPR', 'HIPERTENSION', 'OTHER_DISEASE'
    names_to = "Health Condition",
    values_to = "Presence"
  ) |>
  mutate(Presence = if_else(Presence == 1, "Yes", "No"))

covid_long
```

```
# A tibble: 8,205,776 x 3
    died `Health Condition` Presence
   <dbl> <chr>              <chr>
 1     1 PNEUMONIA          Yes
 2     1 DIABETES           No
 3     1 ASTHMA             No
 4     1 INMSUPR            No
 5     1 HIPERTENSION       Yes
 6     1 OTHER_DISEASE      No
 7     1 CARDIOVASCULAR     No
 8     1 RENAL_CHRONIC      No
 9     1 PNEUMONIA          Yes
10     1 DIABETES           No
# i 8,205,766 more rows
```

```
logit_mod_health <- glm(died ~ as.factor(PNEUMONIA) + as.factor(DIABETES) + as.factor(ASTH
                        data = covid_health,
                        family = "binomial")

tidy(logit_mod_health)
```

```
# A tibble: 8 x 5
  term                       estimate std.error statistic   p.value
  <chr>                         <dbl>     <dbl>     <dbl>     <dbl>
1 (Intercept)                   1.02    0.0464       22.0 2.13e-107
2 as.factor(PNEUMONIA)2        -2.94    0.00907    -324.  0
3 as.factor(DIABETES)2         -0.717   0.0107      -67.2 0
4 as.factor(ASTHMA)2            0.514   0.0308       16.7 1.33e- 62
5 as.factor(INMSUPR)2          -0.285   0.0277      -10.3 8.55e- 25
6 as.factor(HIPERTENSION)2     -0.749   0.0103      -72.7 0
7 as.factor(CARDIOVASCULAR)2   -0.230   0.0224      -10.3 8.32e- 25
8 as.factor(RENAL_CHRONIC)2    -0.555   0.0216      -25.7 1.22e-145
```

```r
logit_health_aug <- augment(logit_mod_health)

logit_health_aug
```

```
# A tibble: 1,025,722 x 14
    died `as.factor(PNEUMONIA)` `as.factor(DIABETES)` `as.factor(ASTHMA)`
   <dbl> <fct>                  <fct>                 <fct>
 1     1 1                      2                     2
 2     1 1                      2                     2
 3     1 2                      1                     2
 4     1 2                      2                     2
 5     1 2                      1                     2
 6     0 1                      2                     2
 7     0 2                      2                     2
 8     0 1                      1                     2
 9     0 2                      1                     2
10     0 2                      2                     2
# i 1,025,712 more rows
# i 10 more variables: `as.factor(INMSUPR)` <fct>,
#   `as.factor(HIPERTENSION)` <fct>, `as.factor(CARDIOVASCULAR)` <fct>,
#   `as.factor(RENAL_CHRONIC)` <fct>, .fitted <dbl>, .resid <dbl>, .hat <dbl>,
#   .sigma <dbl>, .cooksd <dbl>, .std.resid <dbl>
```

```r
logit_health_aug <- logit_health_aug |>
  mutate(prob = exp(.fitted)/(1 + exp(.fitted)),
         pred_died = ifelse(prob > 0.5, "Died", "Did Not Die")) %>%
  select(.fitted, prob, pred_died, died)

logit_health_aug
```

```
# A tibble: 1,025,722 x 4
   .fitted   prob pred_died    died
     <dbl>  <dbl> <chr>       <dbl>
 1  -0.252 0.437  Did Not Die     1
 2   0.303 0.575  Died            1
 3  -3.22  0.0383 Did Not Die     1
 4  -3.94  0.0191 Did Not Die     1
 5  -2.47  0.0777 Did Not Die     1
 6  -1.00  0.269  Did Not Die     0
 7  -3.94  0.0191 Did Not Die     0
 8   1.31  0.787  Died            0
 9  -2.47  0.0777 Did Not Die     0
10  -3.94  0.0191 Did Not Die     0
# i 1,025,712 more rows
```

```
table(logit_health_aug$pred_died, logit_health_aug$died)
```

```
                  0      1
  Did Not Die 937687  61126
  Died         13203  13706
```

##DEMOGRAPHICS:

```
covid_health$PREGNANT <- NULL
covid_health
```

```
# A tibble: 1,025,722 x 21
   USMER MEDICAL_UNIT   SEX PATIENT_TYPE DATE_DIED  INTUBED PNEUMONIA   AGE
   <dbl>        <dbl> <dbl>        <dbl> <chr>        <dbl>     <dbl> <dbl>
 1     2            1     1            1 03/05/2020      97         1    65
 2     2            1     2            1 03/06/2020      97         1    72
 3     2            1     2            2 09/06/2020       1         2    55
 4     2            1     1            1 12/06/2020      97         2    53
 5     2            1     2            1 21/06/2020      97         2    68
 6     2            1     1            2 9999-99-99       2         1    40
 7     2            1     1            1 9999-99-99      97         2    64
 8     2            1     1            1 9999-99-99      97         1    64
 9     2            1     1            2 9999-99-99       2         2    37
10     2            1     1            2 9999-99-99       2         2    25
# i 1,025,712 more rows
# i 13 more variables: DIABETES <dbl>, COPD <dbl>, ASTHMA <dbl>, INMSUPR <dbl>,
#   HIPERTENSION <dbl>, OTHER_DISEASE <dbl>, CARDIOVASCULAR <dbl>,
#   OBESITY <dbl>, RENAL_CHRONIC <dbl>, TOBACCO <dbl>,
#   CLASIFFICATION_FINAL <dbl>, ICU <dbl>, died <dbl>
```

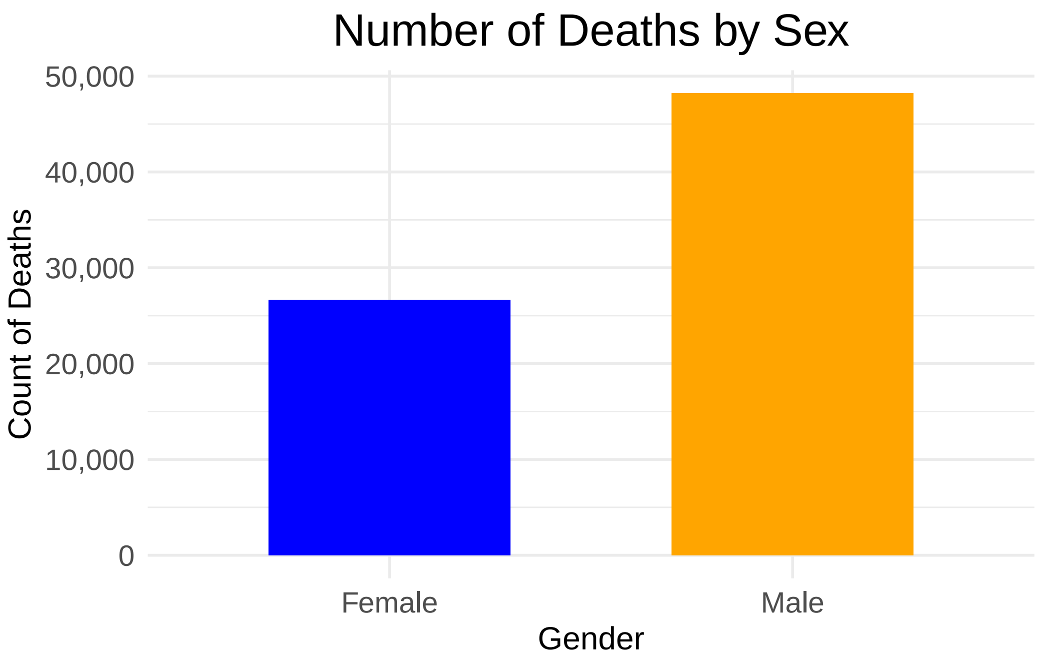
```
covid_health<- covid_health |>
mutate(DIED = ifelse(DATE_DIED == '9999-99-99', 0, 1)) |>
mutate(SEX = as.factor(SEX)) |>
mutate(DIED = as.factor(DIED))
covid_health
```

```
# A tibble: 1,025,722 x 22
   USMER MEDICAL_UNIT SEX   PATIENT_TYPE DATE_DIED  INTUBED PNEUMONIA   AGE
   <dbl>        <dbl> <fct>        <dbl> <chr>        <dbl>     <dbl> <dbl>
 1     2            1 1                1 03/05/2020      97         1    65
 2     2            1 2                1 03/06/2020      97         1    72
 3     2            1 2                2 09/06/2020       1         2    55
 4     2            1 1                1 12/06/2020      97         2    53
 5     2            1 2                1 21/06/2020      97         2    68
 6     2            1 1                2 9999-99-99       2         1    40
 7     2            1 1                1 9999-99-99      97         2    64
 8     2            1 1                1 9999-99-99      97         1    64
 9     2            1 1                2 9999-99-99       2         2    37
10     2            1 1                2 9999-99-99       2         2    25
# i 1,025,712 more rows
# i 14 more variables: DIABETES <dbl>, COPD <dbl>, ASTHMA <dbl>, INMSUPR <dbl>,
#   HIPERTENSION <dbl>, OTHER_DISEASE <dbl>, CARDIOVASCULAR <dbl>,
#   OBESITY <dbl>, RENAL_CHRONIC <dbl>, TOBACCO <dbl>,
#   CLASIFFICATION_FINAL <dbl>, ICU <dbl>, died <dbl>, DIED <fct>
```

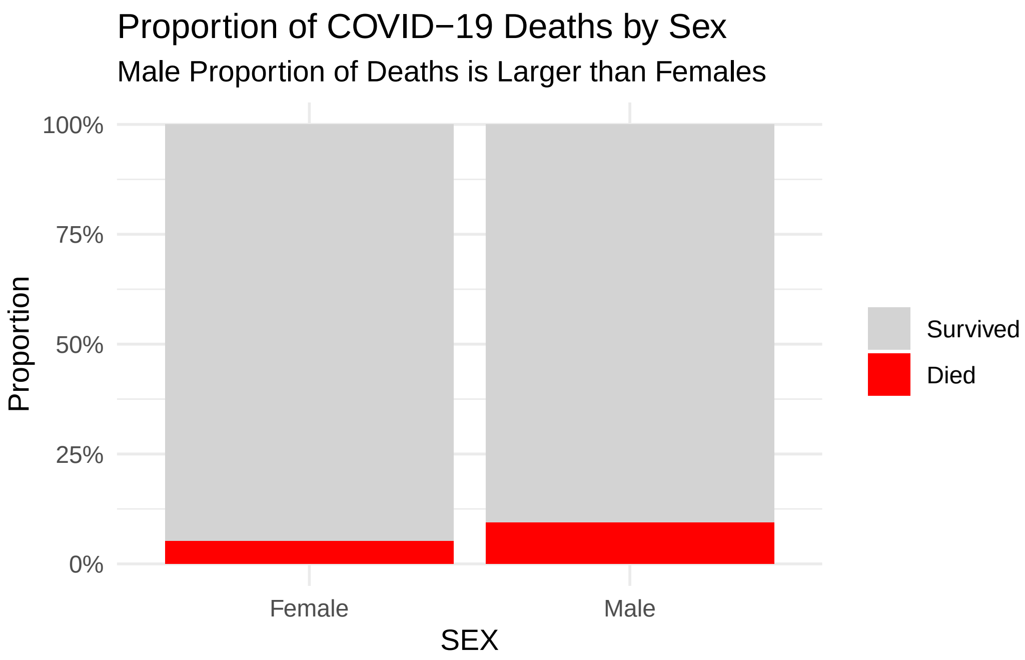
```
covid_health$SEX <- factor(covid_health$SEX, levels = c(1, 2), labels = c("Female", "Male"
covid_health
```

```
# A tibble: 1,025,722 x 22
   USMER MEDICAL_UNIT SEX    PATIENT_TYPE DATE_DIED  INTUBED PNEUMONIA   AGE
   <dbl>        <dbl> <fct>         <dbl> <chr>        <dbl>     <dbl> <dbl>
 1     2            1 Female            1 03/05/2020      97         1    65
 2     2            1 Male              1 03/06/2020      97         1    72
 3     2            1 Male              2 09/06/2020       1         2    55
 4     2            1 Female            1 12/06/2020      97         2    53
 5     2            1 Male              1 21/06/2020      97         2    68
 6     2            1 Female            2 9999-99-99       2         1    40
 7     2            1 Female            1 9999-99-99      97         2    64
 8     2            1 Female            1 9999-99-99      97         1    64
 9     2            1 Female            2 9999-99-99       2         2    37
10     2            1 Female            2 9999-99-99       2         2    25
# i 1,025,712 more rows
# i 14 more variables: DIABETES <dbl>, COPD <dbl>, ASTHMA <dbl>, INMSUPR <dbl>,
#   HIPERTENSION <dbl>, OTHER_DISEASE <dbl>, CARDIOVASCULAR <dbl>,
#   OBESITY <dbl>, RENAL_CHRONIC <dbl>, TOBACCO <dbl>,
#   CLASIFFICATION_FINAL <dbl>, ICU <dbl>, died <dbl>, DIED <fct>
```

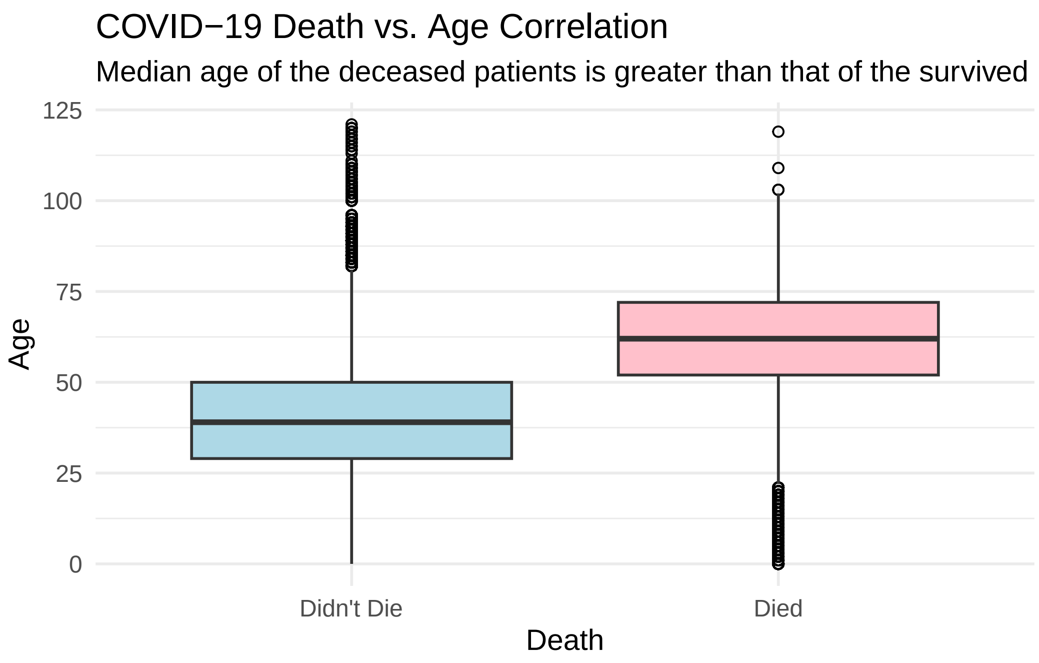
```r
library(scales)
covid_health |>
filter(DIED == 1) |>
ggplot(aes(x = SEX, fill = SEX)) +
geom_bar(stat = "count", width = 0.6) + # Adjust bar width for aesthetics
scale_fill_manual(values = c("blue", "orange")) + # Change colors for clarity
labs(
title = "Number of Deaths by Sex",
x = "Gender",
y = "Count of Deaths"
) +
theme_minimal() +
theme(
text = element_text(size = 14), # Adjust text size for better readability
plot.title = element_text(hjust = 0.5), # Center the plot title
axis.title = element_text(size = 12), # Specify axis title size
legend.position = "none" # Remove legend if redundant
) +
scale_y_continuous(labels = comma)
```



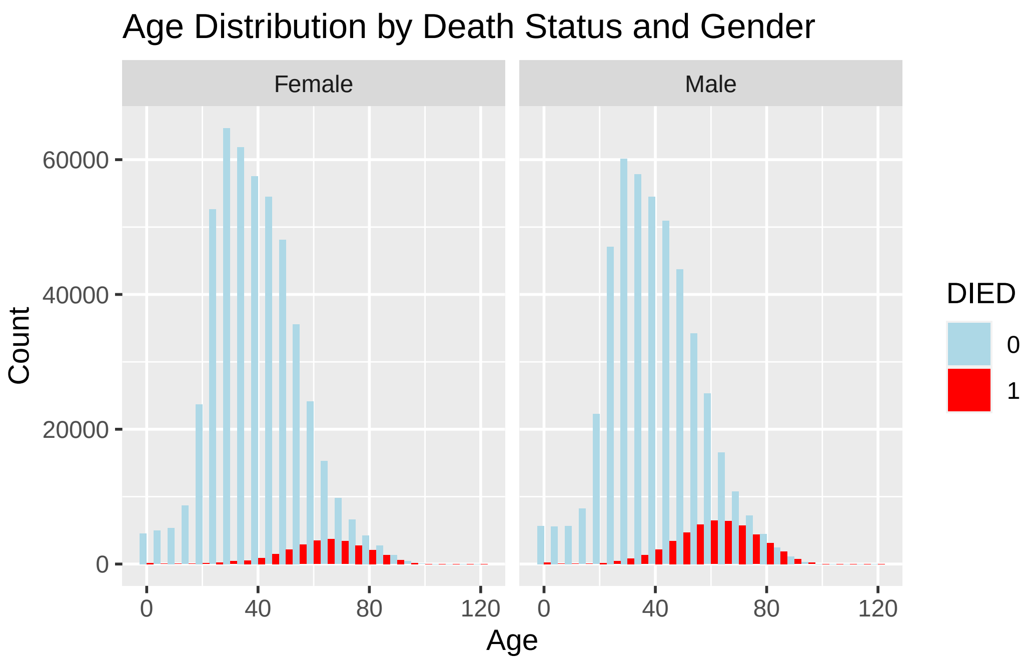
```
covid_summary <- covid_health |>
count(SEX, DIED) |>
group_by(SEX) |>
mutate(Proportion = n / sum(n))
# Plot
ggplot(covid_summary, aes(x = SEX, y = Proportion, fill = DIED)) +
geom_col() +
scale_y_continuous(labels = scales::percent_format()) +
labs(
title = "Proportion of COVID-19 Deaths by Sex",
subtitle = "Male Proportion of Deaths is Larger than Females",
x = "SEX",
y = "Proportion"
) +
scale_fill_manual(values = c("0" = "lightgrey", "1" = "red"),
labels = c("Survived", "Died")) +
theme_minimal() +
theme(legend.title = element_blank())
```



```r
covid_health |>
mutate(DIED = factor(DIED, levels = c(0, 1), labels = c("Didn't Die", "Died"))) |>
ggplot(aes(x = DIED, y = AGE, fill = DIED)) +
geom_boxplot(outlier.colour = "black", outlier.shape = 1) +
labs(title = "COVID-19 Death vs. Age Correlation",
subtitle = "Median age of the deceased patients is greater than that of the survived", x =
y = "Age") +
scale_fill_manual(values = c("Didn't Die" = "lightblue", "Died" = "pink")) +
theme_minimal() +
theme(legend.position = "none")
```



```
ggplot(covid_health, aes(x = AGE, fill = DIED)) +
geom_histogram(binwidth = 5, position = "dodge") +
facet_wrap(~SEX) +
labs(title = "Age Distribution by Death Status and Gender",
x = "Age",
y = "Count") +
scale_fill_manual(values = c("0" = "lightblue", "1" = "red"))
```



```
logit_mod_demo <- glm(DIED ~ as.factor(SEX) + AGE, data = covid_health, family = "binomial
tidy(logit_mod_demo)
```

```
# A tibble: 3 x 5
  term               estimate std.error statistic p.value
  <chr>                 <dbl>     <dbl>     <dbl>   <dbl>
1 (Intercept)          -6.75    0.0168      -401.        0
2 as.factor(SEX)Male    0.638   0.00852       74.8       0
3 AGE                   0.0761  0.000263     289.        0
```

SEX male: Holding age constant, we predict the odds of a male patient in the COVID-19 dataset passing away to be around e^0.63753001(1.8918) times that of a female patient. AGE: Holding sex constant, we predict that for each additional year in age of the patient, the odds of passing away are multiplied by e^0.07605202(1.0790).

```
logit_demo_aug <- augment(logit_mod_demo)
logit_demo_aug
```

```
# A tibble: 1,025,722 x 9
   DIED  `as.factor(SEX)`   AGE .fitted .resid    .hat .sigma .cooksd .std.resid
   <fct> <fct>            <dbl>   <dbl>  <dbl>   <dbl>  <dbl>   <dbl>      <dbl>
 1 1     Female              65  -1.81   1.98  5.80e-6  0.644 1.18e-5      1.98
 2 1     Male                72  -0.637  1.46  9.62e-6  0.644 6.07e-6      1.46
 3 1     Male                55  -1.93   2.03  3.06e-6  0.644 7.04e-6      2.03
 4 1     Female              53  -2.72   2.36  2.74e-6  0.644 1.39e-5      2.36
 5 1     Male                68  -0.942  1.59  7.15e-6  0.644 6.11e-6      1.59
 6 0     Female              40  -3.71  -0.220 1.62e-6  0.644 1.32e-8     -0.220
 7 0     Female              64  -1.88  -0.532 5.40e-6  0.644 2.74e-7     -0.532
 8 0     Female              64  -1.88  -0.532 5.40e-6  0.644 2.74e-7     -0.532
 9 0     Female              37  -3.94  -0.197 1.46e-6  0.644 9.47e-9     -0.197
10 0     Female              25  -4.85  -0.125 9.52e-7  0.644 2.49e-9     -0.125
# i 1,025,712 more rows
```

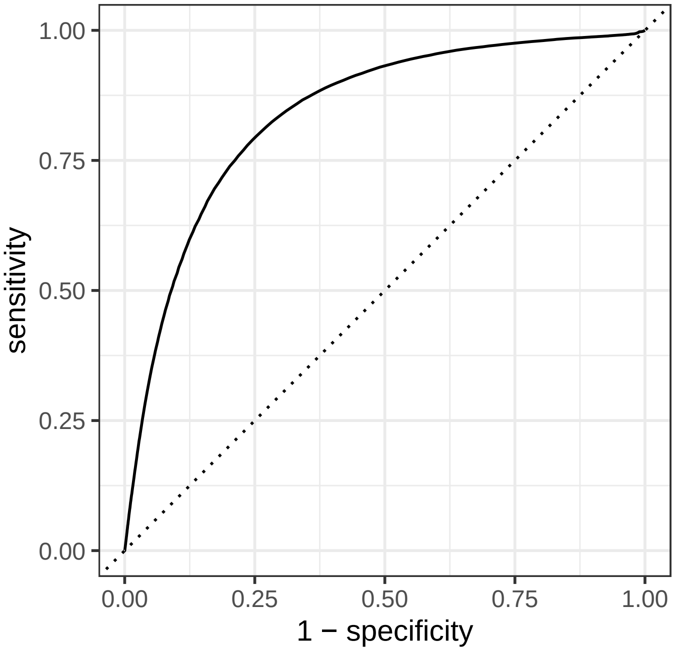
```
logit_demo_aug <- logit_demo_aug |>
mutate(prob = exp(.fitted)/(1 + exp(.fitted)),
pred_died = ifelse(prob > 0.5, "Died", "Did Not Die")) |>
select(.fitted, prob, pred_died, DIED)
logit_demo_aug
```

```
# A tibble: 1,025,722 x 4
   .fitted    prob pred_died   DIED
     <dbl>   <dbl> <chr>       <fct>
 1  -1.81  0.141   Did Not Die 1
 2  -0.637 0.346   Did Not Die 1
 3  -1.93  0.127   Did Not Die 1
 4  -2.72  0.0618  Did Not Die 1
 5  -0.942 0.281   Did Not Die 1
 6  -3.71  0.0239  Did Not Die 0
 7  -1.88  0.132   Did Not Die 0
 8  -1.88  0.132   Did Not Die 0
 9  -3.94  0.0191  Did Not Die 0
10  -4.85  0.00777 Did Not Die 0
# i 1,025,712 more rows
```

```
logit_demo_aug |>
roc_auc(
truth = DIED,
prob, event_level = "second"
)
```

```
# A tibble: 1 x 3
  .metric .estimator .estimate
  <chr>   <chr>          <dbl>
1 roc_auc binary         0.840
```

```
logit_demo_aug |>
roc_curve(
truth = DIED,
.fitted,
event_level = "second"
) |>
autoplot()
```



The AUC I achieved is 0.8395. It means that there is approximately an 83.95% chance that the model will be able to distinguish between a patient who died and one who did not die from COVID-19.

```
library(Stat2Data)
library(pROC)
```

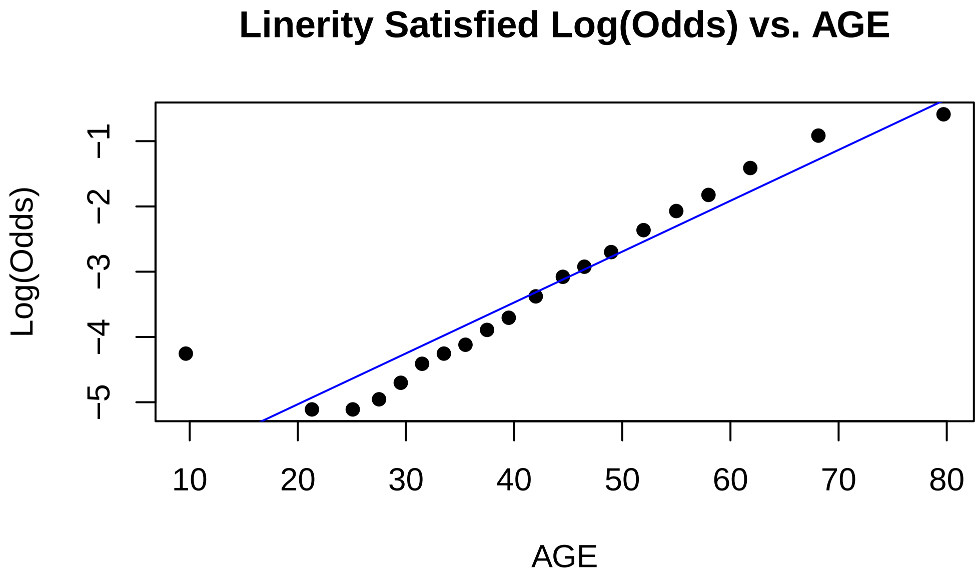
```
Type 'citation("pROC")' for a citation.
```

```
Attaching package: 'pROC'


The following objects are masked from 'package:stats':

    cov, smooth, var
```

```
emplogitplot1(DIED ~ AGE, data = covid_health, ngroups = 20,
main = "Linerity Satisfied Log(Odds) vs. AGE")
```



Since age is the only continous variable in the model, we must check our linearity assumption. From the plot graphed above, we see that it displays a linear relationship between the log odds of dying and age. The graph displays points that are relatively linearly uniform and follow the trend of the line. From the graph, we do not see that it contain any distinct trend that would challenge our assumption of linearity