Analysis of COVID-19 Dataset

Ayorinde Ayomide David

2025-07-27

Importing the dependencies

```
if(!require(pacman)) install.packages("pacman")  # Installing the package manager

pacman::p_load(tidyverse, # Metapackage
               here, # R library for coercing Rmarkdown into reading dataset from a seperate folder
               visdat, # R library for graphical inspection of dataset
               inspectdf, # R library for the distribution of variables
               gtsummary,
               ggplot2
               )
```

Loading the COVID-19 dataset into R

```
covid <- read_csv(here("Data/COVID19_line_list_data.csv"))
```

Exploring and inspecting the dataset

```
# Exploring the dataset

dim(covid)
```

```
## [1] 1085   27
```

```
head(covid, n = 10)
```

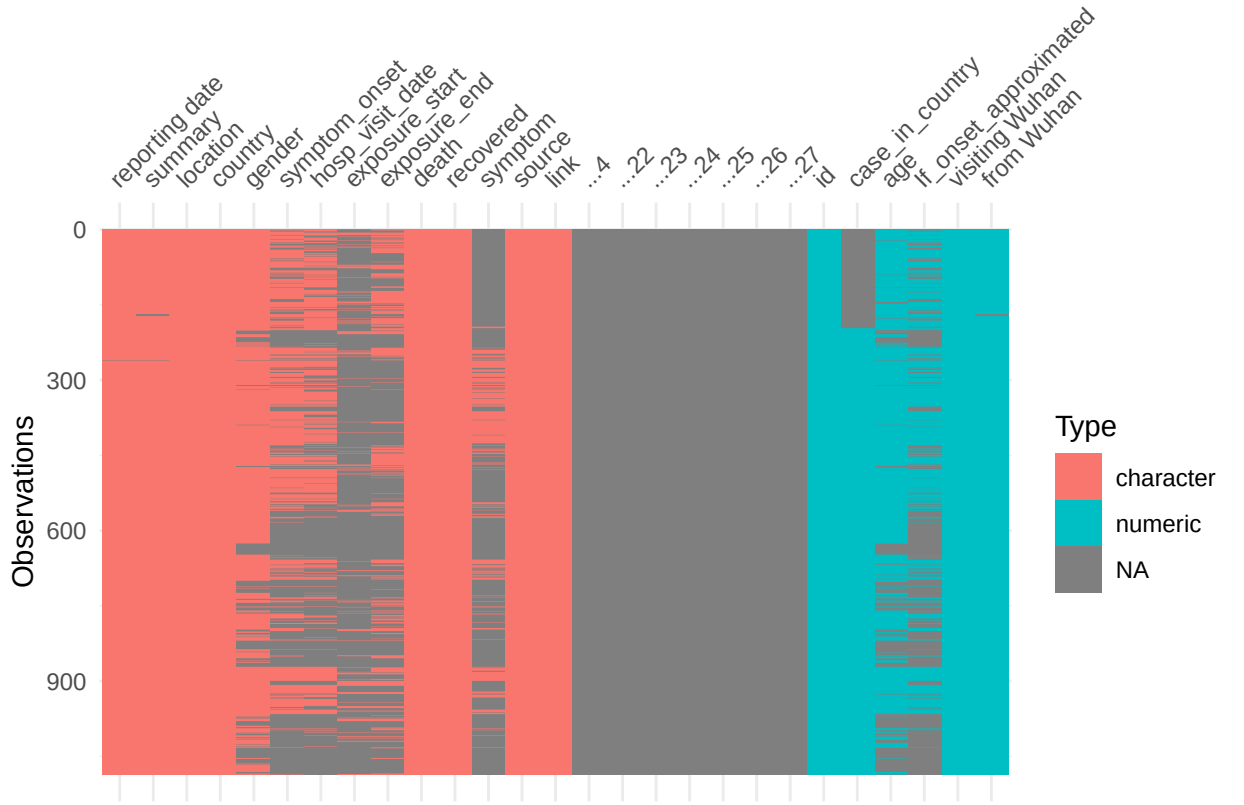
```
## # A tibble: 10 x 27
##       id case_in_country `reporting date` ...4  summary  location country gender
##    <dbl>           <dbl> <chr>            <lgl> <chr>    <chr>    <chr>   <chr>
##  1     1              NA 1/20/2020        NA    First c~ Shenzhe~ China   male
##  2     2              NA 1/20/2020        NA    First c~ Shanghai China   female
##  3     3              NA 1/21/2020        NA    First c~ Zhejiang China   male
##  4     4              NA 1/21/2020        NA    new con~ Tianjin  China   female
##  5     5              NA 1/21/2020        NA    new con~ Tianjin  China   male
##  6     6              NA 1/21/2020        NA    First c~ Chongqi~ China   female
##  7     7              NA 1/21/2020        NA    First c~ Sichuan  China   male
```

```
##  8     8              NA 1/21/2020        NA    new con~ Beijing  China   male
##  9     9              NA 1/21/2020        NA    new con~ Beijing  China   male
## 10    10              NA 1/21/2020        NA    new con~ Beijing  China   male
## # i 19 more variables: age <dbl>, symptom_onset <chr>,
## #   If_onset_approximated <dbl>, hosp_visit_date <chr>, exposure_start <chr>,
## #   exposure_end <chr>, `visiting Wuhan` <dbl>, `from Wuhan` <dbl>,
## #   death <chr>, recovered <chr>, symptom <chr>, source <chr>, link <chr>,
## #   ...22 <lgl>, ...23 <lgl>, ...24 <lgl>, ...25 <lgl>, ...26 <lgl>,
## #   ...27 <lgl>
```

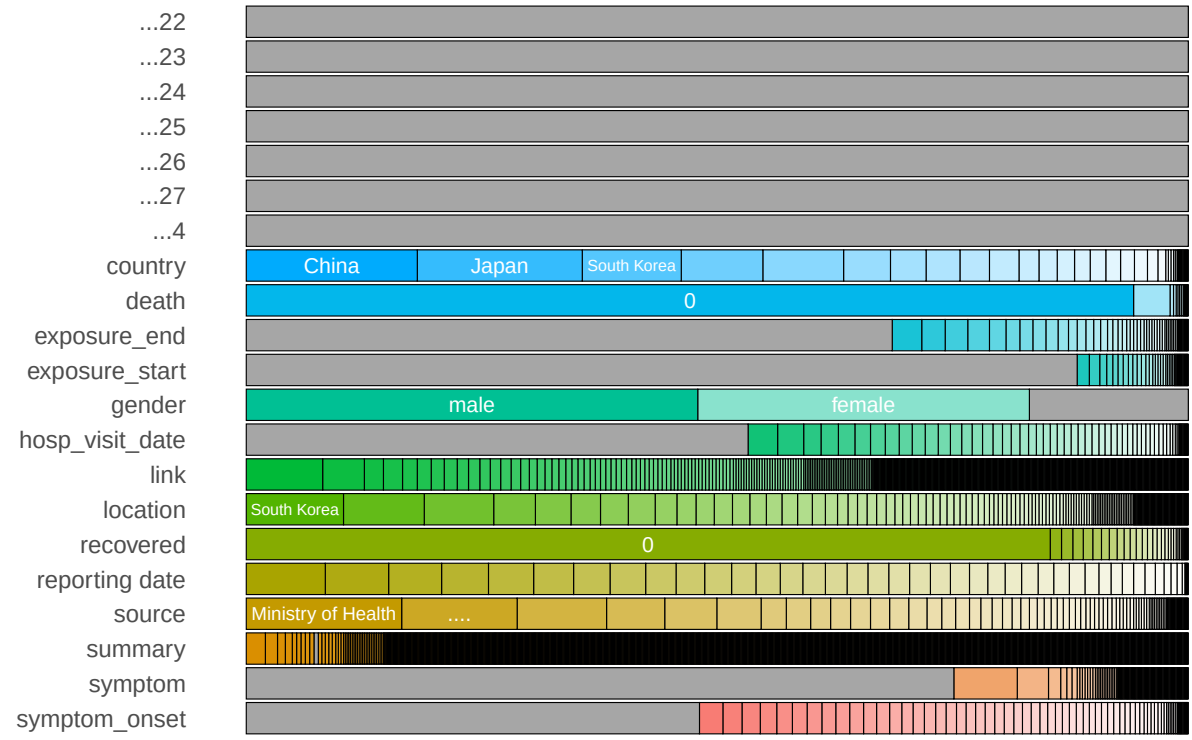
```
tail(covid, n = 10)
```

```
## # A tibble: 10 x 27
##       id case_in_country `reporting date` ...4  summary  location country gender
##    <dbl>           <dbl> <chr>            <lgl> <chr>    <chr>    <chr>   <chr>
##  1  1076              14 2/25/2020        NA    new COV~ Bahrain  Bahrain male
##  2  1077              15 2/25/2020        NA    new COV~ Bahrain  Bahrain male
##  3  1078              16 2/25/2020        NA    new COV~ Bahrain  Bahrain female
##  4  1079              17 2/25/2020        NA    new COV~ Bahrain  Bahrain female
##  5  1080               1 2/25/2020        NA    new COV~ Innsbru~ Austria <NA>
##  6  1081               2 2/25/2020        NA    new COV~ Innsbru~ Austria <NA>
##  7  1082               1 2/24/2020        NA    new COV~ Afghani~ Afghan~ <NA>
##  8  1083               1 2/26/2020        NA    new COV~ Algeria  Algeria male
##  9  1084               1 2/25/2020        NA    new COV~ Croatia  Croatia male
## 10  1085               1 2/25/2020        NA    new COV~ Bern     Switze~ male
## # i 19 more variables: age <dbl>, symptom_onset <chr>,
## #   If_onset_approximated <dbl>, hosp_visit_date <chr>, exposure_start <chr>,
## #   exposure_end <chr>, `visiting Wuhan` <dbl>, `from Wuhan` <dbl>,
## #   death <chr>, recovered <chr>, symptom <chr>, source <chr>, link <chr>,
## #   ...22 <lgl>, ...23 <lgl>, ...24 <lgl>, ...25 <lgl>, ...26 <lgl>,
## #   ...27 <lgl>
```

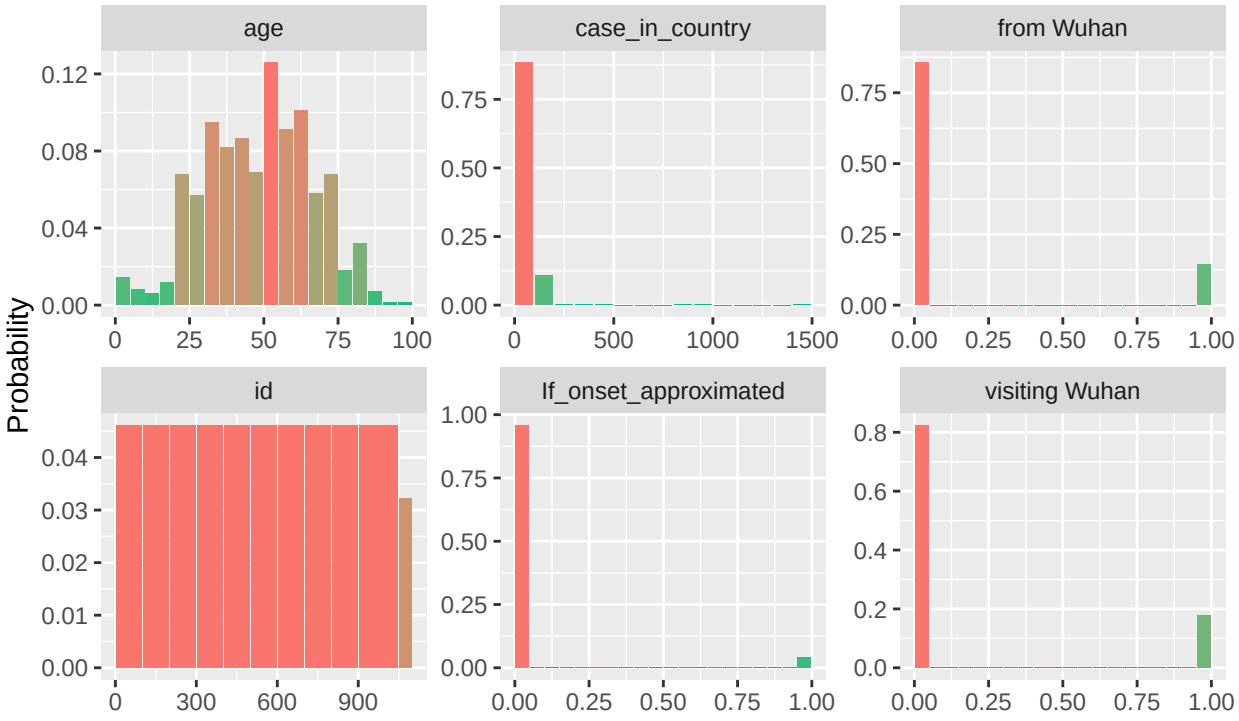
```
# Inspecting the dataset

vis_dat(covid)
```




```
inspect_cat(covid) %>%
  show_plot()
```




```
inspect_num(covid) %>%
  show_plot()
```



Selecting, cleaning, transformation and manipulation of the variables of interest

```
covid_selected <- covid %>%
  select(id,
         reporting_date = 'reporting date', # The initial variable name has to go into quote because it
         gender,
         death,
         age,
         country)
```

```
covid_selected %>%
  select(death) %>%
  unique()
```

```
## # A tibble: 14 x 1
##    death
##    <chr>
##  1 0
##  2 1
##  3 2/14/2020
##  4 2/26/2020
##  5 2/13/2020
##  6 2/28/2020
```

```
##  7 2/27/2020
##  8 2/25/2020
##  9 2/23/2020
## 10 2/24/2020
## 11 2/22/2020
## 12 02/01/20
## 13 2/19/2020
## 14 2/21/2020
```

```
covid_selected <- covid_selected %>%
  mutate(death = as.integer(covid$death != 0)) # This overwrite the initial death column by leaving ent

covid_selected %>% # checking to confirm if the changes has been effected
  select(death) %>%
  unique()
```

```
## # A tibble: 2 x 1
##   death
##   <int>
## 1     0
## 2     1
```

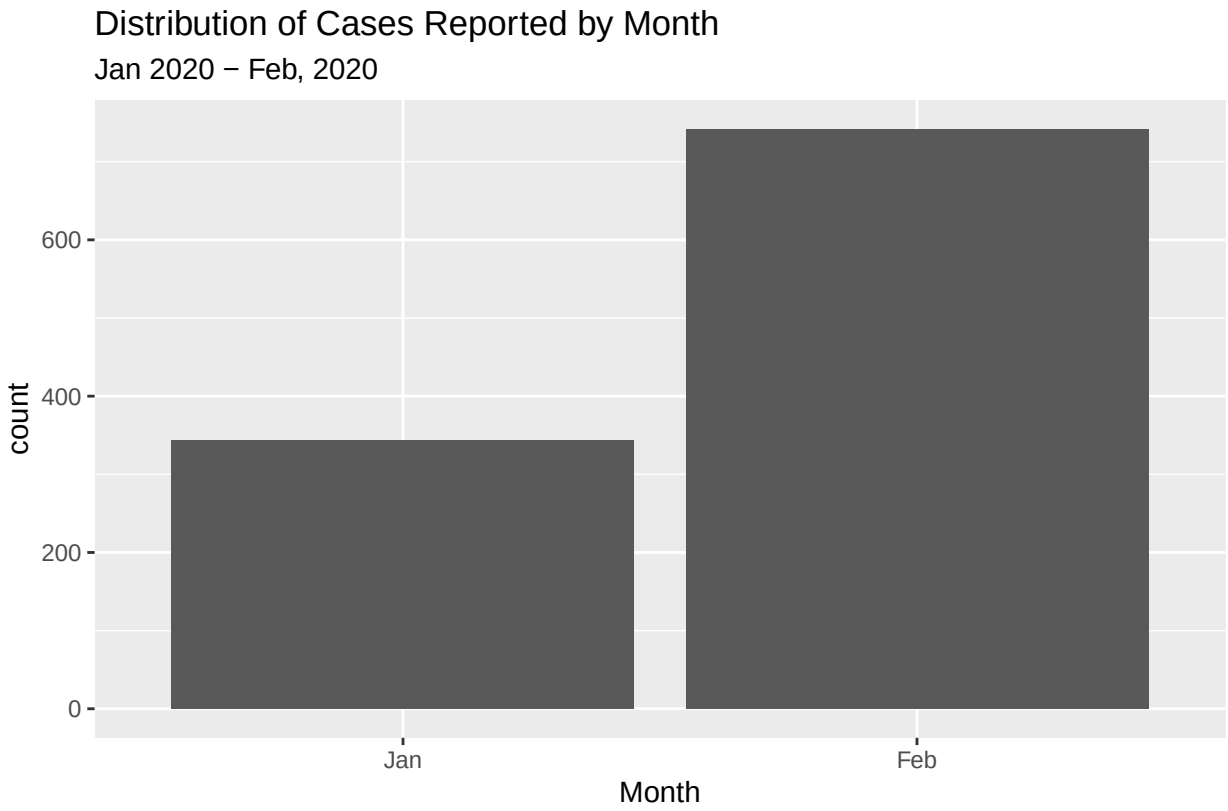
```
covid_selected <- covid_selected %>%
  mutate(reporting_date = mdy(reporting_date)) # This overwrite the initial reporting date by convertin
```

```
covid_selected <- covid_selected %>%
  mutate(month = month(reporting_date, label = T),
         month = replace_na(month,"Feb")) # The first mutate chunk create a new column for month and th
```

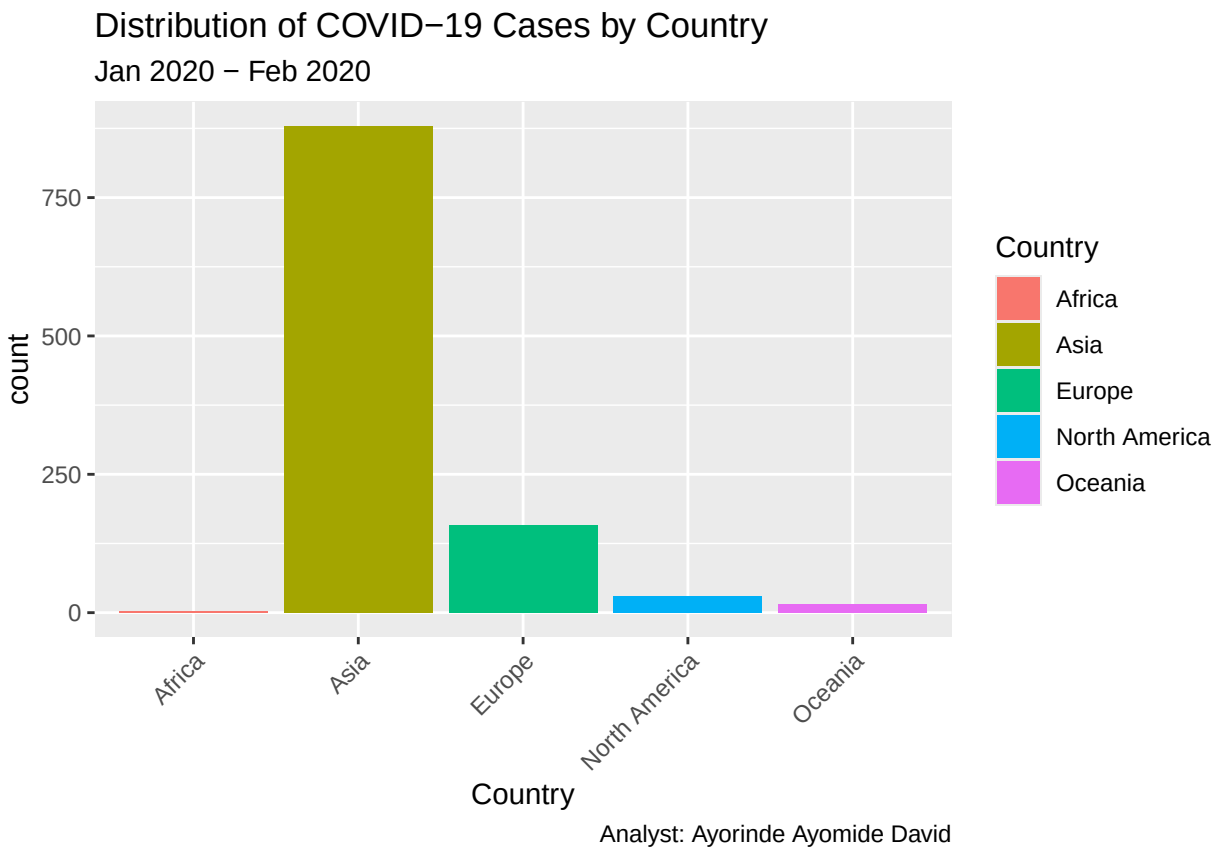
```
covid_selected <- covid_selected %>% # Creating a column for continent
  mutate(continent = case_when(
    country %in% c("USA", "Canada") ~ "North America",
    country %in% c("France", "Germany", "Italy", "Russia", "UK", "Finland", "Spain",
                   "Sweden", "Belgium", "Austria", "Croatia", "Switzerland") ~ "Europe",
    country %in% c("China", "Japan", "Malaysia", "Nepal", "Singapore", "South Korea",
                   "Taiwan", "Thailand", "Vietnam", "Cambodia", "Sri Lanka", "UAE",
                   "Hong Kong", "India", "Phillipines", "Iran", "Israel", "Lebanon",
                   "Kuwait", "Bahrain", "Afghanistan") ~ "Asia",
    country %in% c("Australia") ~ "Oceania",
    country %in% c("Egypt", "Algeria") ~ "Africa",
    TRUE ~ "Other"))
```

Visualizing some of the variables of interest

```
ggplot(covid_selected, mapping = aes(x = month)) +
  geom_bar() +
  labs(title = "Distribution of Cases Reported by Month",
       subtitle = "Jan 2020 - Feb, 2020",
       x = "Month",
       caption = "Analyst: Ayorinde Ayomide David")
```




```
ggplot(covid_selected,mapping = aes(x = continent, fill = continent)) +
   geom_bar() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  labs(title = "Distribution of COVID-19 Cases by Country",
       subtitle = "Jan 2020 - Feb 2020",
       x = "Country",
       caption = "Analyst: Ayorinde Ayomide David",
       fill = "Country")
```



Analyzing fatality by country

```
covid_selected %>%
  group_by(country) %>%
  summarise(number_of_death = sum(death == 1))
```

```
## # A tibble: 38 x 2
##    country     number_of_death
##    <chr>                 <int>
##  1 Afghanistan               0
##  2 Algeria                   0
##  3 Australia                 0
##  4 Austria                   0
##  5 Bahrain                   0
##  6 Belgium                   0
##  7 Cambodia                  0
##  8 Canada                    0
##  9 China                    39
## 10 Croatia                   0
## # i 28 more rows
```

Statistical Analysis

Two-sample t-test

H_0: There is no significant difference between the age of those alive and dead ($\mu_1 = \mu_2$)

H_1: There is a significant difference between the age of those alive and dead ($\mu_1 \neq \mu_2$)

```
covid_selected %>%
  group_by(death) %>%
  summarise(gender_death_mean = mean(age, na.rm = T))
```

```
## # A tibble: 2 x 2
##   death gender_death_mean
##   <int>             <dbl>
## 1     0              48.1
## 2     1              68.6
```

We can see that there is a difference of about 20(in years) between the ages of those that are dead and alive. Now, the question:

Is this really significant?

Let's confirm using **t.test**

```
dead <- covid_selected %>%
  filter(death == 1)
alive <- covid_selected %>%
  filter(death == 0)

t.test(alive$age,
       dead$age,
       conf.level = 0.95,
       alternative = "two.sided")
```

```
##
##  Welch Two Sample t-test
##
## data:  alive$age and dead$age
## t = -10.839, df = 72.234, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -24.28669 -16.74114
## sample estimates:
## mean of x mean of y
##  48.07229  68.58621
```

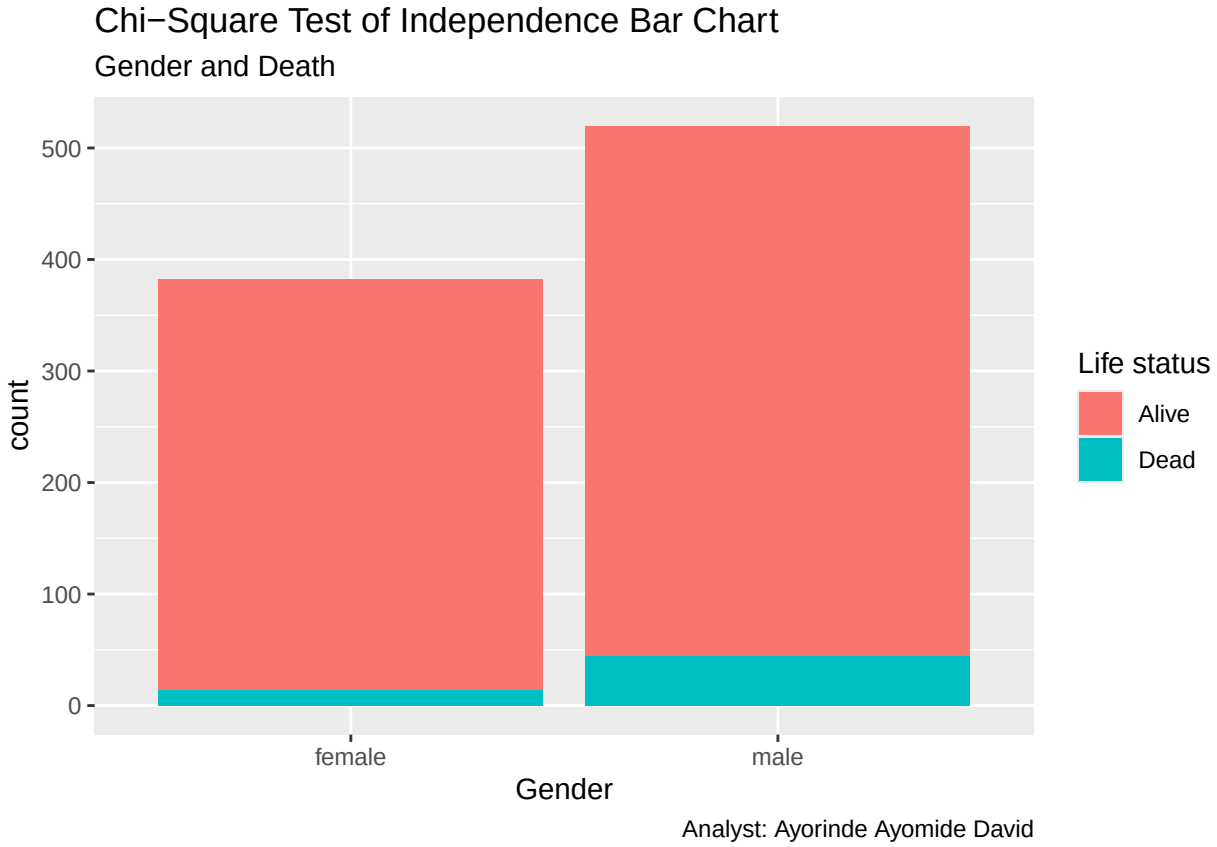
Decision rule: If p-value is < 0.05, we reject null hypothesis, otherwise, we fail to reject null hypothesis

Conclusion: Since the p-value is < 0.05, we reject null hypothesis and conclude that there is a significant difference between the age of those that are dead and those that are alive. In order words, older people are more likely/prone to death if tested positive for COVID-19

Test of Independence

```
covid_selected <- covid_selected %>%
  drop_na(gender) %>%
  mutate(gender_cov = factor(gender),
         death_cov = factor(case_when(death == 1 ~ "Dead",
```

```
                                death == 0 ~ "Alive")))

ggplot(covid_selected, mapping = aes(x = gender_cov,
                               fill = death_cov)) +
  geom_bar() +
  labs(title = "Chi-Square Test of Independence Bar Chart",
       subtitle = "Gender and Death",
       x = "Gender",
       caption = "Analyst: Ayorinde Ayomide David",
       fill = "Life status")
```



Are the proportions of gender independent of life status?

The question above leads us to the hypothesis below

H_0: The variables are independent i.e There is no relationship between the variables

H_1: The variables are not independent i.e There is a relationship between the variables

```
covid_selected %>%
  select(gender_cov,
         death_cov) %>%
  table() %>%
  chisq.test()
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
```

```
##
## data:  .
## X-squared = 7.6428, df = 1, p-value = 0.0057
```

Decision rule: If p-value is < 0.05, we reject null hypothesis, otherwise, we fail to reject null hypothesis

Conclusion: Since the p-value is < 0.05, we reject null hypothesis and conclude that there is a relationship between the death and gender variable