COVID-19 Mortality Analysis Report

Rahma Touaibi

2025-12-23

1. Introduction

The primary goal of this analysis is to identify key risk factors associated with COVID-19 mortality. By examining patient demographics—specifically age, gender, and location—we can pinpoint which populations experienced the highest fatality rates. This report uses custom R code to clean data, perform statistical t-tests, and generate visualizations to communicate trends clearly.

---

2. Data Import and Cleaning

In this stage, we import the raw dataset, handle missing values for age by imputing the mean, and clean the death column to ensure it is in a binary format.

```
#import data
COVID19 <- read.csv("C:\\Users\\admin\\Desktop\\Data Analytics\\Projects For Portfolio\\Unfinished Proj
```

#make a copy

```
data<-COVID19
```

#call libraries

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

```
# check unique value in death col
unique(data$death)
```

```
##  [1] "0"         "1"         "2/14/2020" "2/26/2020" "2/13/2020" "2/28/2020"
##  [7] "2/27/2020" "2/25/2020" "2/23/2020" "2/24/2020" "2/22/2020" "02/01/20"
## [13] "2/19/2020" "2/21/2020"
```

```r
#clean death
data$death<-as.integer(data$death!=0)

#death rate
death_rate<-sum(data$death)/nrow(data)

#clean age
data$age[is.na(data$age)]<-mean(data$age,na.rm = TRUE)
data$age<-round(data$age,0)

# clean data
data$X<-NULL
data$symptom<-NULL
data$X.1<-NULL
data$X.2<-NULL
data$X.3<-NULL
data$X.4<-NULL
data$X.5<-NULL
data$X.6<-NULL
data$link<-NULL
data$summary<-NULL
data$source<-NULL
```

→ The data was successfully cleaned and metadata columns were removed. The calculated overall death rate provides a baseline for the following demographic comparisons.

3. Statistical Analysis: Age vs. Mortality

- We subset the data into two groups dead and alive to compare their mean ages and determine statistical significance.

```r
#compare death as AGE
#death and alive total
dead = subset(data,death==1)
alive= subset(data,death==0)
mean(dead$age)
```

```
## [1] 67.03175
```

```r
mean(alive$age)
```

```
## [1] 48.28669
```

```r
#test
t.test(dead$age,alive$age,alternative = "two.sided",conf.level = 0.99)
```

```
##
##  Welch Two Sample t-test
##
## data:  dead$age and alive$age
## t = 10.194, df = 71.687, p-value = 1.367e-15
## alternative hypothesis: true difference in means is not equal to 0
## 99 percent confidence interval:
##  13.87932 23.61079
## sample estimates:
## mean of x mean of y
##  67.03175  48.28669
```

→ The Welch Two Sample t-test shows that the mean age of deceased COVID-19 patients (67.0 years) is significantly higher than that of alive patients (48.3 years), t = 10.19, p < 0.001.

4. Statistical Analysis: Gender vs. Mortality

- We examine the death rates between men and women to investigate gender disparity.

```r
#compare death as GENDER
men=subset(data,gender=="male")
woman=subset(data,gender=="female")
mean(men$death) #8.5%
```

```
## [1] 0.08461538
```

```r
mean(woman$death)#3.7%
```

```
## [1] 0.03664921
```

#test

```r
t.test(men$death,woman$death,alternative="two.sided",conf.level = 0.99)
```

```
##
##  Welch Two Sample t-test
##
## data:  men$death and woman$death
## t = 3.084, df = 894.06, p-value = 0.002105
## alternative hypothesis: true difference in means is not equal to 0
## 99 percent confidence interval:
##  0.007817675 0.088114665
## sample estimates:
##  mean of x  mean of y
## 0.08461538 0.03664921
```

→ The t-test shows a mean death rate of (8.46%) for men, which is significantly higher than for women (3.7%) .The 99% confidence interval confirms the gender disparity in mortality is clearly above zero.

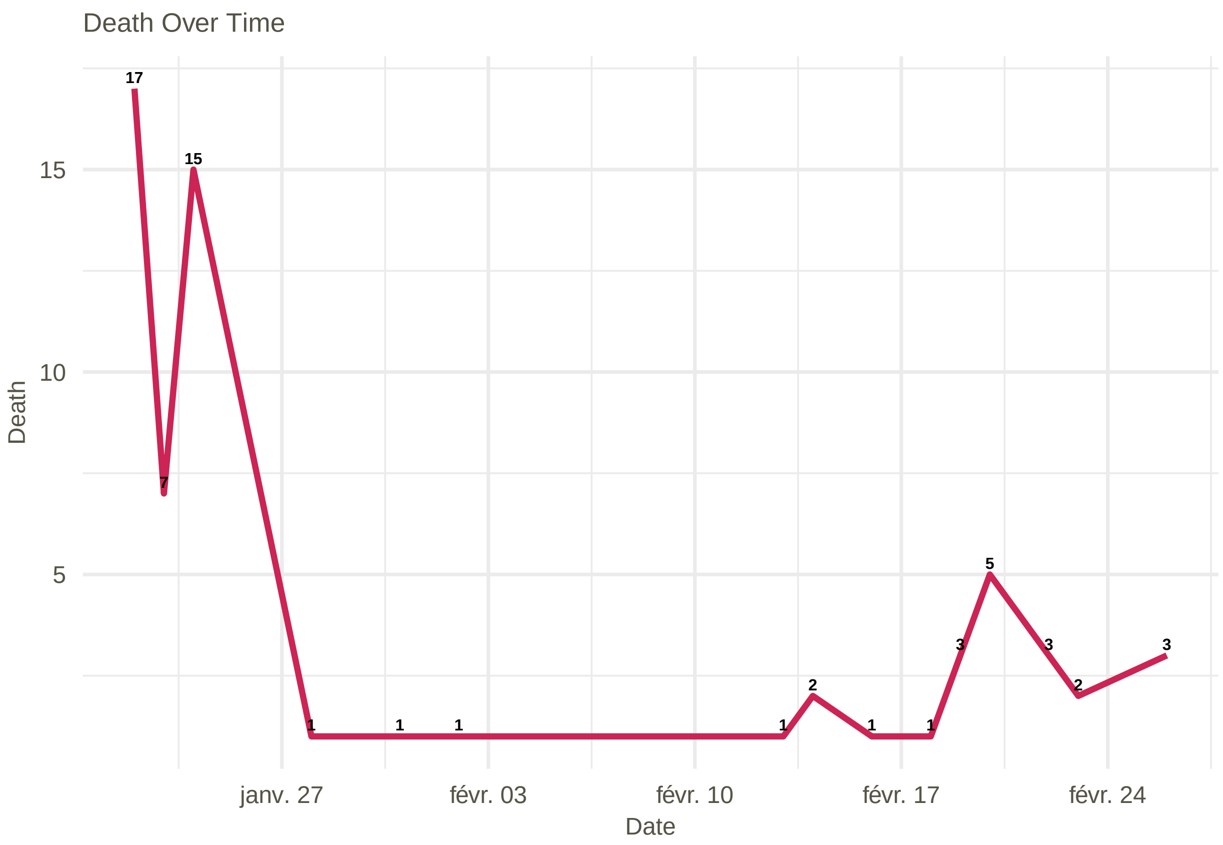
5. Visualizing Death Trends Over Time

- This section tracks the progression of fatalities using a time-series line plot.

```
#fix date
data<- data %>%
  mutate(reporting.date = mdy(reporting.date))  # convert to date format

#death over time
daily_death<- data%>%
  filter(death==1)%>%
  group_by(reporting.date)%>%
  summarise(death_count=n())

#size figure
ggplot(daily_death, aes(x = reporting.date, y = death_count)) +
  geom_line(color = "#CC2454",linewidth=1.1) +
  geom_text(aes(label = death_count),
            vjust = -0.5,
            size = 2.2,
            color = "black",
            fontface = "bold") +
  theme_minimal(base_size = 14) +
  ggtitle("Death Over Time") +
  xlab("Date") +
  ylab("Death") +
  theme(
    plot.title = element_text(size = 10, hjust = 0,color = "#545245"),
    axis.text.x = element_text(size = 9, angle = 0,color = "#545245"),
    axis.text.y = element_text(size = 9,color = "#545245"),
    axis.title.x = element_text(size = 9,color = "#545245"),
    axis.title.y = element_text(size = 9,color = "#545245")
  )
```



→ COVID-19 deaths were highest at the beginning, with peaks of 17 and 15 deaths, then sharply dropped to very low numbers for several days.

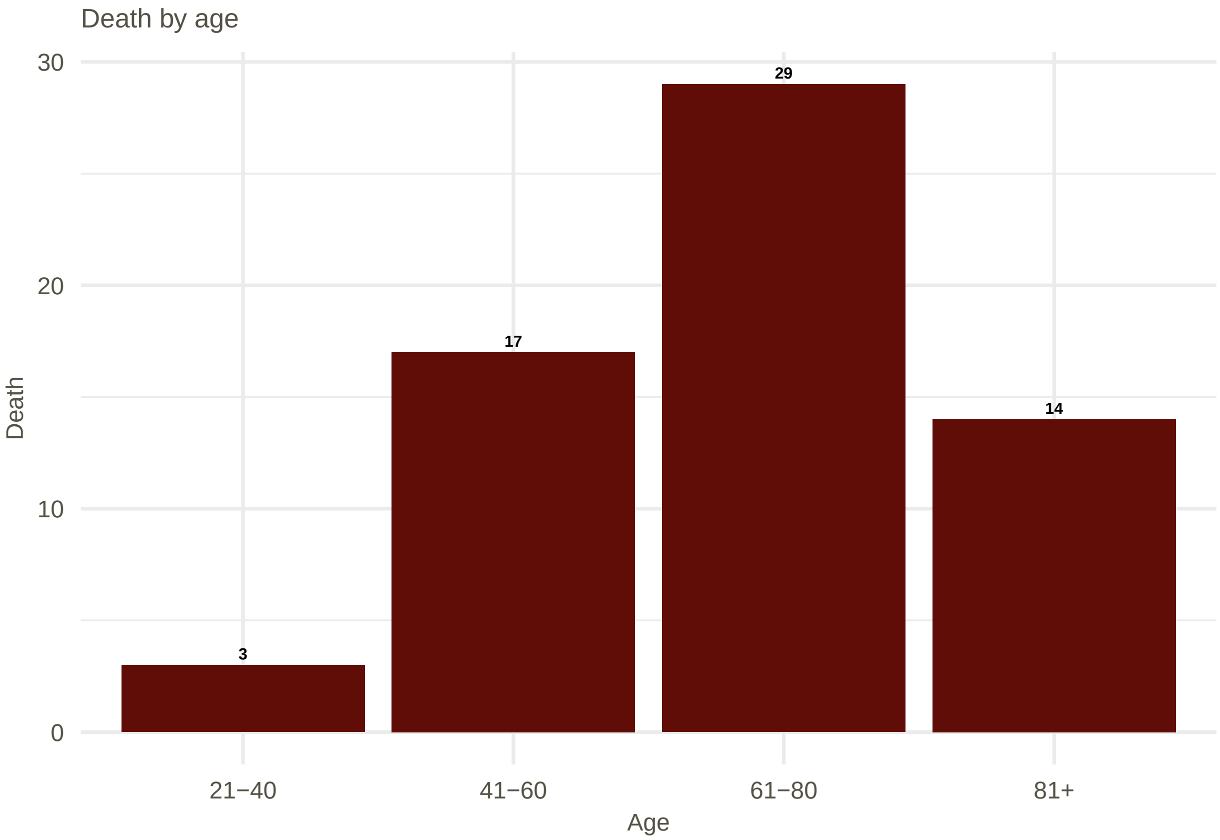
6. Mortality by Age Group

- We categorized ages into groups to identify which life stage was most vulnerable to the virus.

```
#Death by age
data<-data%>%
  mutate(age_group=cut(age,breaks=c(0,20,40,60,80,120),labels=c("0-20","21-40","41-60","61-80","81+")))

age_death<- data%>%
  filter(death==1)%>%
  group_by(age_group)%>%
  summarise(death_count=n())

#viz
ggplot(age_death,aes(age_group,death_count))+
geom_bar(stat ="identity",fill="#600D07")+
geom_text(aes(label = death_count),
          vjust =-0.5,
          size = 2.2,
          color = "black",
          fontface = "bold") +
  theme_minimal(base_size = 14) +
  ggtitle("Death by age") +
  xlab("Age") +
  ylab("Death")+
    theme(
    plot.title = element_text(size = 10, hjust = 0,color = "#545245"),
    axis.text.x = element_text(size = 9, angle = 0,color = "#545245"),
    axis.text.y = element_text(size = 9,color = "#545245"),
    axis.title.x = element_text(size = 9,color = "#545245"),
    axis.title.y = element_text(size = 9,color = "#545245")
  )
```



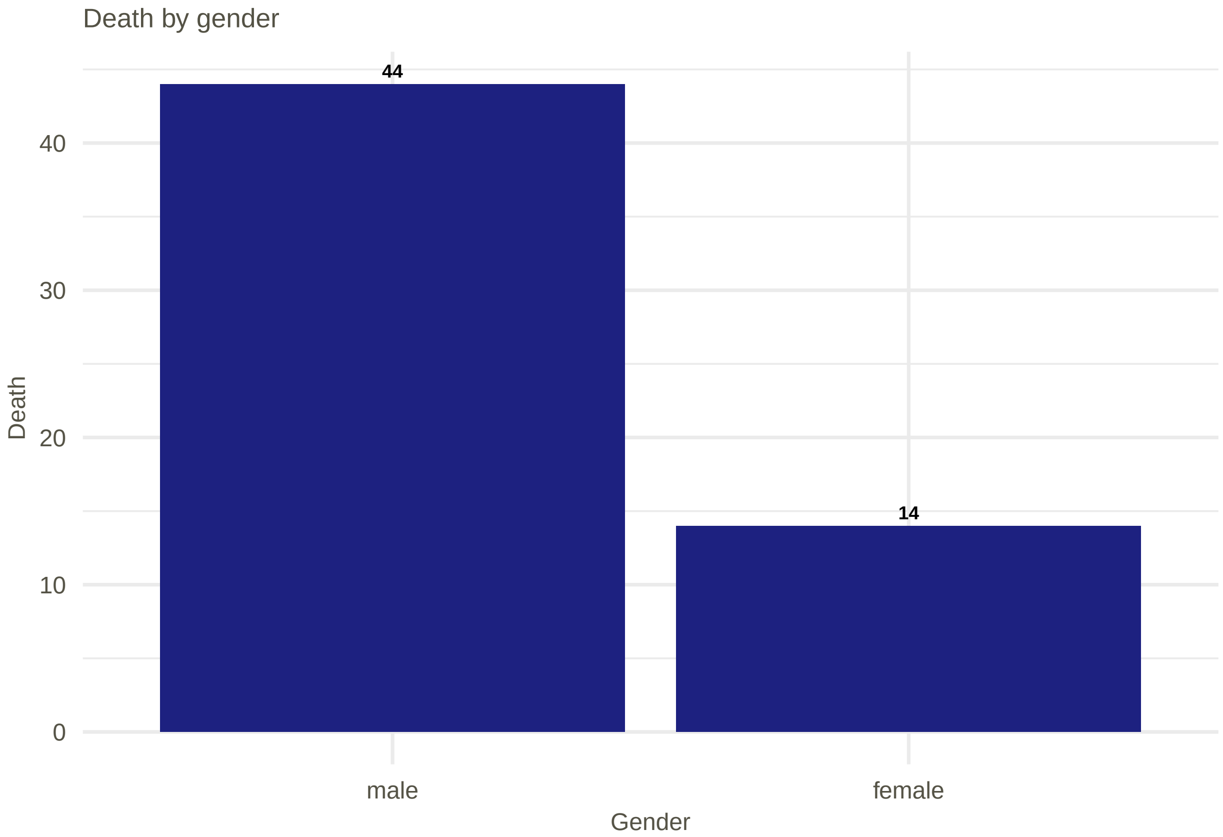
→ Most deaths occur in the 61-80 age group. Death rates are significantly lower in the younger (21-60) and oldest (81+) age groups.

7. Mortality by Gender

- The following bar chart visualizes the distribution of deaths by gender.

```
# Gender vs Death Rate
gender_death<- data%>%
  filter(death==1)%>%
  group_by(gender)%>%
  summarise(count_death=n())
gender_death<-gender_death[!is.na(gender_death$gender), ]
```

```r
ggplot(gender_death,aes(reorder(gender,-count_death),count_death))+
geom_bar(stat="identity", fill="#1D2180")+
  geom_text(aes(label =count_death),
            vjust =-0.5,
            size = 2.5,
            color = "black",
            fontface = "bold") +
  theme_minimal(base_size = 14) +
  ggtitle("Death by gender") +
  xlab("Gender") +
  ylab("Death") +
  theme(
    plot.title = element_text(size = 10, hjust = 0,color = "#545245"),
    axis.text.x = element_text(size = 9, angle = 0,color = "#545245"),
    axis.text.y = element_text(size = 9,color = "#545245"),
    axis.title.x = element_text(size = 9,color = "#545245"),
    axis.title.y = element_text(size = 9,color = "#545245")
  )
```



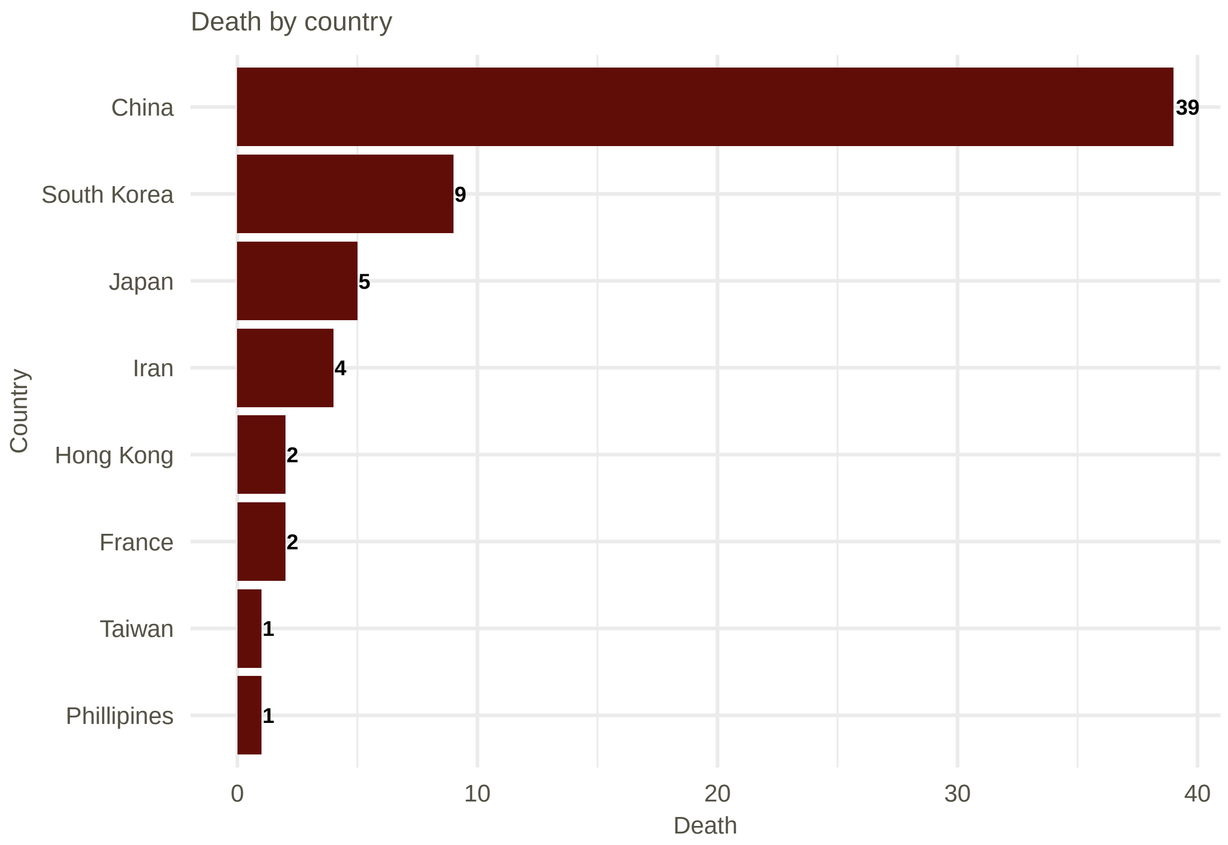
→ A stark gender disparity in deaths, with males accounting for 44 deaths compared to 14 for females. The mortality rate for males is approximately three times higher than that of females in the represented data.

8. Mortality by Country

- The following bar chart visualizes the distribution of deaths by country.

```r
#death by location
country_death<-data%>%
  filter(death==1)%>%
  group_by(country)%>%
  summarise(count_death=n())

ggplot(country_death, aes(reorder(country, count_death), count_death)) +
  geom_bar(stat="identity", fill="#600D07") +
  coord_flip() +
  geom_text(aes(label = count_death),
            hjust = -0.1,
            size = 2.8,
            color = "black",
            fontface = "bold") +
  theme_minimal(base_size = 14) +
  ggtitle("Death by country") +
  xlab("Country") +
  ylab("Death") +
  theme(
    plot.title = element_text(size = 10, hjust = 0, color = "#545245"),
    axis.text.x = element_text(size = 9, angle = 0, color = "#545245"),
    axis.text.y = element_text(size = 9, color = "#545245"),
    axis.title.x = element_text(size = 9, color = "#545245"),
    axis.title.y = element_text(size = 9, color = "#545245")
  )
```



–> a significant disparity in death counts by country, with China reporting 39 deaths, which is more than four times higher than the next highest country, South Korea at 9. The remaining countries listed, including Japan, Iran, and several others, each report 5 or fewer deaths in the dataset.

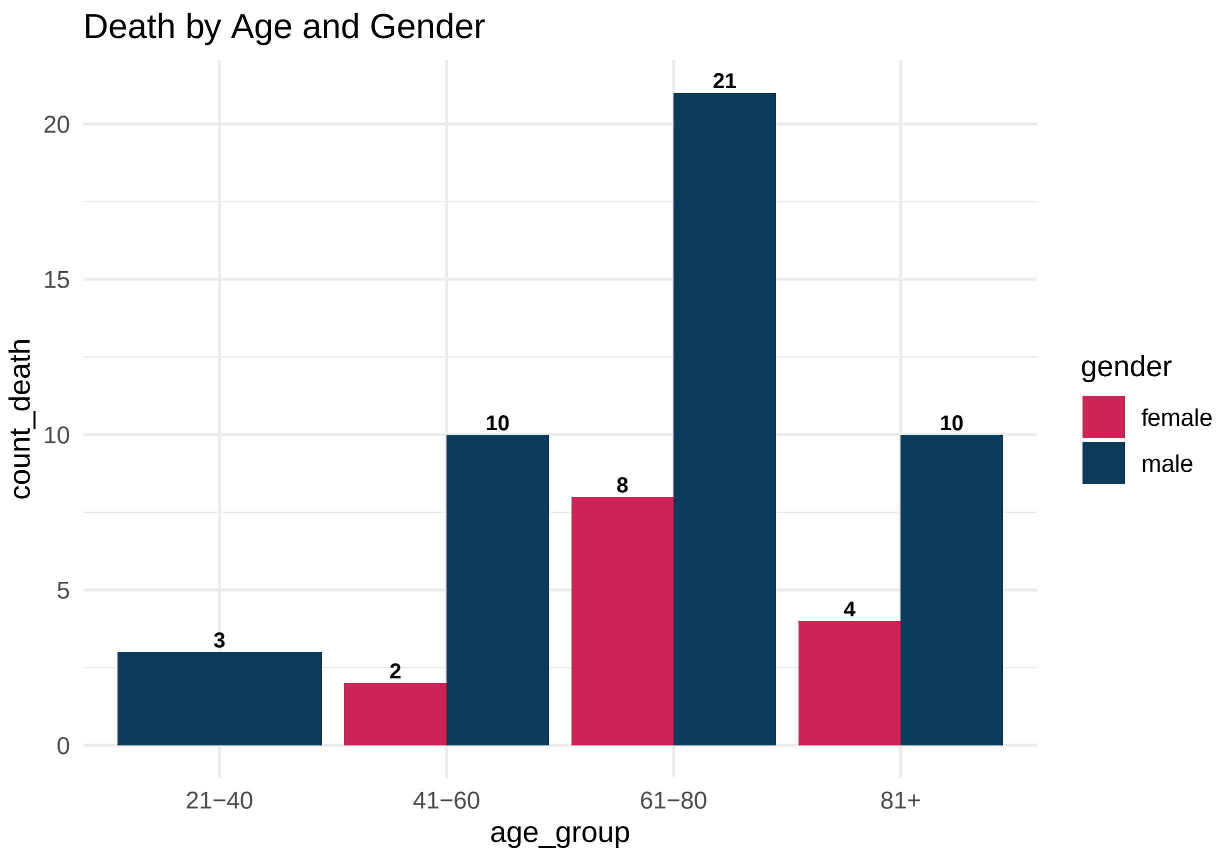
9. Age and Gender Intersection

- This grouped visualization shows the interaction between age and gender in relation to mortality.

```
#death by age and gender
age_gender_death <- data %>%
  filter(death == 1) %>%
  group_by(age_group, gender) %>%
  summarise(count_death = n()) %>%
  drop_na()
```

```
## ‘summarise()‘ has grouped output by ’age_group’. You can override using the
## ‘.groups‘ argument.
```

```
ggplot(age_gender_death, aes(age_group, count_death, fill = gender)) +
  geom_bar(stat = "identity", position = "dodge") +
  geom_text(aes(label = count_death),
            position = position_dodge(width = 0.9),
            vjust = -0.3,
            size=2.8,
            color = "black",
            fontface = "bold") +
  scale_fill_manual(values = c("male" = "#0B3C5D", "female" = "#CC2454"))+
  theme_minimal() +
  ggtitle("Death by Age and Gender")
```



→ Mortality increases significantly with age, peaking in the 61–80 age group for both genders. Across all age groups, males consistently record higher numbers of deaths than females.

10. Conclusion

Based on the analysis, we have reached the following conclusions:

Age Impact: There is a direct correlation between advanced age and mortality risk, specifically peaking in the 61-80 bracket.

Gender Disparity: Males face a significantly higher risk of death compared to females (8.5% vs 3.7%).

Outbreak Timing: The dataset captures an initial peak in fatalities followed by a decline to sporadic levels.

11. Recommendations

Based on these findings, we recommend the following actions:

Targeted Protection: Public health policies should prioritize protection and early intervention for males over the age of 60.

Resource Allocation: Medical resources should be scaled to handle higher demand within older demographic segments during waves.

Research Focus: Further investigate lifestyle or biological factors that contribute to the observed gender gap in mortality.