

Using Patient Data to Predict COVID-19 Fatality Risk: A Logistic Regression Approach

Bin Han

2025-07-08

Table of contents

1	Project description	1
2	Loading data and R library	2
2.1	Exploration data analysis	3
2.2	Data Overview	3
2.3	Correlation plot	6
2.4	Bar charts to understand the variables	7
3	Logistic Regression	22
3.1	Initial model	22
3.2	Model selection	23
4	Model validation	29
4.1	Prediction plot	29
4.2	Model accuracy	32

1 Project description

During the COVID-19 pandemic, healthcare systems worldwide faced immense pressure due to limited medical resources and overwhelming patient volumes. Early identification of high-risk patients was essential for effective resource allocation and clinical decision-making. This project aims to develop an explainable logistic regression model to predict the probability of death in COVID-19 patients based on their demographics, symptoms, and pre-existing health conditions.

The dataset includes variables such as age, sex, presence of pneumonia, comorbidities (e.g., diabetes, hypertension, COPD), lifestyle factors, and care pathway (hospitalization vs. sent home). Missing values were coded as 999 and appropriately handled. **Exploratory analysis revealed that females and patients sent home had a lower risk of death,**

while pneumonia, advanced age, and several chronic health conditions significantly increased mortality risk. Surprisingly, pregnancy, asthma, and tobacco use showed minimal association with death.

After cleaning the data and applying stepwise feature selection, the final model demonstrated strong performance with an accuracy of **89.62%**. The model provides clear, interpretable insights into how individual risk factors contribute to mortality, supporting its potential use in clinical environments. Visualization of death probabilities across age and gender groups further enhances its utility for early risk assessment and proactive care planning.

2 Loading data and R library

Load the data

```
library(tidyverse)
```

```
-- Attaching core tidyverse packages ----- tidyverse 2.0.0 --
v dplyr      1.1.4      v readr      2.1.5
v forcats    1.0.0      v stringr    1.5.1
v ggplot2    3.5.2      v tibble     3.2.1
v lubridate  1.9.4      v tidyr      1.3.1
v purrr      1.0.4
-- 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 be
```

```
library(car)
```

Loading required package: carData

Attaching package: 'car'

The following object is masked from 'package:dplyr':

recode

The following object is masked from 'package:purrr':

some

```
library(corrplot)
```

```
corrplot 0.95 loaded
```

```
data<-read.csv("covid19.csv")
getwd()
```

```
[1] "/home/binhan/project1/R-kurs/CovidA3"
```

2.1 Exploration data analysis

2.2 Data Overview

```
glimpse(data)
```

```
Rows: 10,000
```

```
Columns: 16
```

```
$ sex           <int> 1, 0, 1, 1, 0, 1, 0, 0, 1, 1, 1, 1, 0, 0, 0, 0, 1, 1, 1~
$ patient_type  <int> 0, 1, 0, 0, 0, 1, 1, 0, 1, 1, 0, 0, 0, 0, 1, 1, 1, 1, 1~
$ pneumonia     <int> 1, 0, 0, 0, 1, 0, 0, 1, 0, 0, 1, 1, 0, 1, 1, 0, 0, 0, 0~
$ age           <int> 65, 26, 73, 64, 84, 21, 34, 78, 26, 22, 58, 55, 75, 36, ~
$ pregnant      <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
$ diabetes      <int> 1, 0, 0, 0, 0, 0, 0, 1, 0, 0, 1, 1, 0, 0, 0, 0, 0, 0, 0~
$ copd          <int> 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
$ asthma        <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0~
$ inmsupr       <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
$ hypertension  <int> 1, 0, 1, 1, 1, 0, 1, 0, 0, 0, 1, 0, 1, 0, 0, 0, 0, 1, 0~
$ other_disease <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0~
$ cardiovascular <int> 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
$ obesity       <int> 0, 0, 1, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0~
$ renal_chronic <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
$ tobacco       <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 0, 0, 0, 0~
$ dead          <int> 0, 0, 1, 1, 1, 0, 0, 0, 0, 0, 1, 1, 1, 0, 0, 0, 0, 0, 0~
```

```
summary(data)
```

sex	patient_type	pneumonia	age
Min. :0.0000	Min. :0.0000	Min. :0.000	Min. : 0.00
1st Qu.:0.0000	1st Qu.:0.0000	1st Qu.:0.000	1st Qu.: 33.00
Median :0.0000	Median :1.0000	Median :0.000	Median : 44.00

Mean :0.4636	Mean :0.7196	Mean :0.217	Mean : 45.16
3rd Qu.:1.0000	3rd Qu.:1.0000	3rd Qu.:0.000	3rd Qu.: 56.00
Max. :1.0000	Max. :1.0000	Max. :1.000	Max. :100.00
pregnant	diabetes	copd	asthma
Min. : 0.000	Min. : 0.000	Min. : 0.000	Min. : 0.000
1st Qu.: 0.000	1st Qu.: 0.000	1st Qu.: 0.000	1st Qu.: 0.000
Median : 0.000	Median : 0.000	Median : 0.000	Median : 0.000
Mean : 2.803	Mean : 4.557	Mean : 4.012	Mean : 4.023
3rd Qu.: 0.000	3rd Qu.: 0.000	3rd Qu.: 0.000	3rd Qu.: 0.000
Max. :999.000	Max. :999.000	Max. :999.000	Max. :999.000
inmsupr	hypertension	other_disease	cardiovascular
Min. : 0.000	Min. : 0.000	Min. : 0.000	Min. : 0.000
1st Qu.: 0.000	1st Qu.: 0.000	1st Qu.: 0.000	1st Qu.: 0.000
Median : 0.000	Median : 0.000	Median : 0.000	Median : 0.000
Mean : 4.908	Mean : 4.191	Mean : 6.621	Mean : 4.016
3rd Qu.: 0.000	3rd Qu.: 0.000	3rd Qu.: 0.000	3rd Qu.: 0.000
Max. :999.000	Max. :999.000	Max. :999.000	Max. :999.000
obesity	renal_chronic	tobacco	dead
Min. : 0.00	Min. : 0.000	Min. : 0.00	Min. :0.0000
1st Qu.: 0.00	1st Qu.: 0.000	1st Qu.: 0.00	1st Qu.:0.0000
Median : 0.00	Median : 0.000	Median : 0.00	Median :0.0000
Mean : 4.18	Mean : 4.215	Mean : 4.57	Mean :0.1344
3rd Qu.: 0.00	3rd Qu.: 0.000	3rd Qu.: 0.00	3rd Qu.:0.0000
Max. :999.00	Max. :999.000	Max. :999.00	Max. :1.0000

There are many 999 which represents missing value in the data frame. in order to use it for analysis, we convert the missing number 999 to NA.

```
data[data==999]<-NA
summary(data)
```

sex	patient_type	pneumonia	age
Min. :0.0000	Min. :0.0000	Min. :0.000	Min. : 0.00
1st Qu.:0.0000	1st Qu.:0.0000	1st Qu.:0.000	1st Qu.: 33.00
Median :0.0000	Median :1.0000	Median :0.000	Median : 44.00
Mean :0.4636	Mean :0.7196	Mean :0.217	Mean : 45.16
3rd Qu.:1.0000	3rd Qu.:1.0000	3rd Qu.:0.000	3rd Qu.: 56.00
Max. :1.0000	Max. :1.0000	Max. :1.000	Max. :100.00
pregnant	diabetes	copd	asthma
Min. :0.000000	Min. :0.0000	Min. :0.00000	Min. :0.00000
1st Qu.:0.000000	1st Qu.:0.0000	1st Qu.:0.00000	1st Qu.:0.00000
Median :0.000000	Median :0.0000	Median :0.00000	Median :0.00000
Mean :0.005616	Mean :0.1622	Mean :0.01586	Mean :0.02711
3rd Qu.:0.000000	3rd Qu.:0.0000	3rd Qu.:0.00000	3rd Qu.:0.00000

Max. :1.000000	Max. :1.0000	Max. :1.00000	Max. :1.00000
NA's :28	NA's :44	NA's :40	NA's :40
inmsupr	hypertension	other_disease	cardiovascular
Min. :0.00000	Min. :0.000	Min. :0.00000	Min. :0.00000
1st Qu.:0.00000	1st Qu.:0.000	1st Qu.:0.00000	1st Qu.:0.00000
Median :0.00000	Median :0.000	Median :0.00000	Median :0.00000
Mean :0.01316	Mean :0.196	Mean :0.02788	Mean :0.01978
3rd Qu.:0.00000	3rd Qu.:0.000	3rd Qu.:0.00000	3rd Qu.:0.00000
Max. :1.00000	Max. :1.000	Max. :1.00000	Max. :1.00000
NA's :49	NA's :40	NA's :66	NA's :40
obesity	renal_chronic	tobacco	dead
Min. :0.0000	Min. :0.00000	Min. :0.00000	Min. :0.0000
1st Qu.:0.0000	1st Qu.:0.00000	1st Qu.:0.00000	1st Qu.:0.0000
Median :0.0000	Median :0.00000	Median :0.00000	Median :0.0000
Mean :0.1849	Mean :0.01958	Mean :0.07474	Mean :0.1344
3rd Qu.:0.0000	3rd Qu.:0.00000	3rd Qu.:0.00000	3rd Qu.:0.0000
Max. :1.0000	Max. :1.00000	Max. :1.00000	Max. :1.0000
NA's :40	NA's :42	NA's :45	

To know more about the categorical variables, we can use the count function.

```
data|>
  count(sex, name="count")
```

```
sex count
1  0  5364
2  1  4636
```

Sex of patient, 1 for female and 0 for male. Therefore the numbers of observations from man and woman are close.

```
data|>
  count(pregnant, name="count")
```

```
pregnant count
1         0  9916
2         1    56
3        NA    28
```

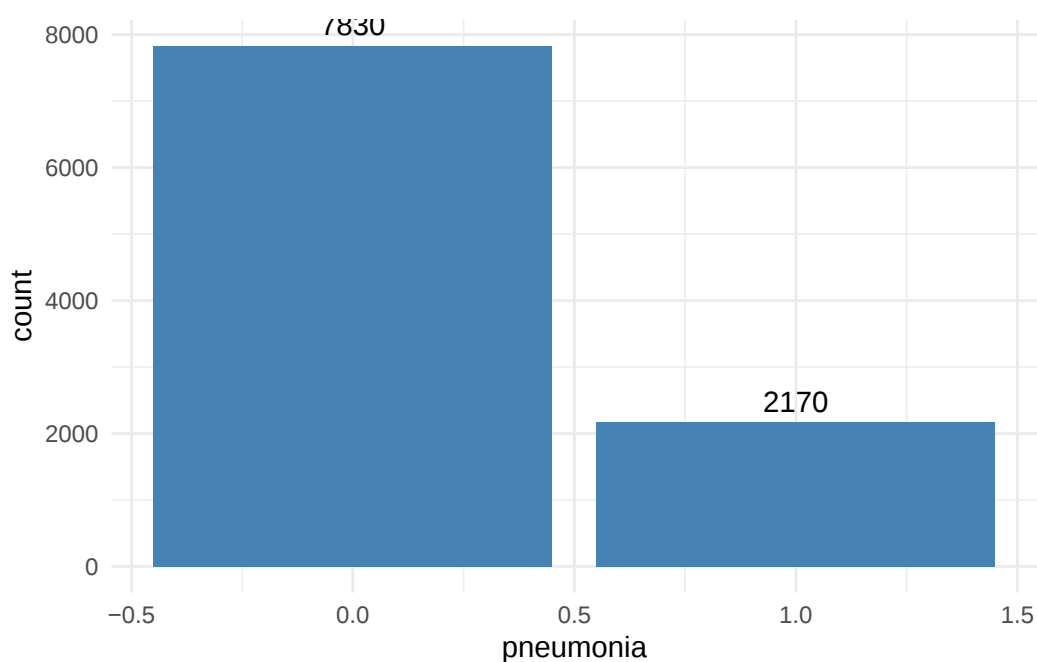
Among the data set, the pregnant samples are 58 of 10 000.

```
data|>
  count(tobacco, name="count")
```

	tobacco	count
1	0	9211
2	1	744
3	NA	45

Among the data set, the smoking samples are 744 of 10 000. Other than count function, we can also use the bar plot to know the variable.

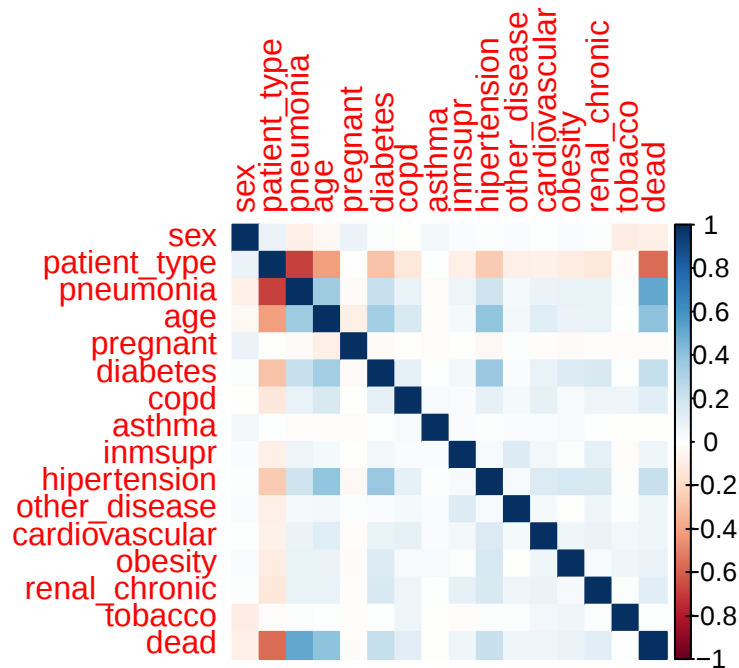
```
ggplot(data, aes(x=pneumonia))+
  geom_bar(fill="steelblue")+
  geom_text(stat="count", aes(label=after_stat(count)), vjust=-0.5)+
  theme_minimal()
```



Among the data set, we have 2170 observations who have pneumonia among the 10 000 cases.

2.3 Correlation plot

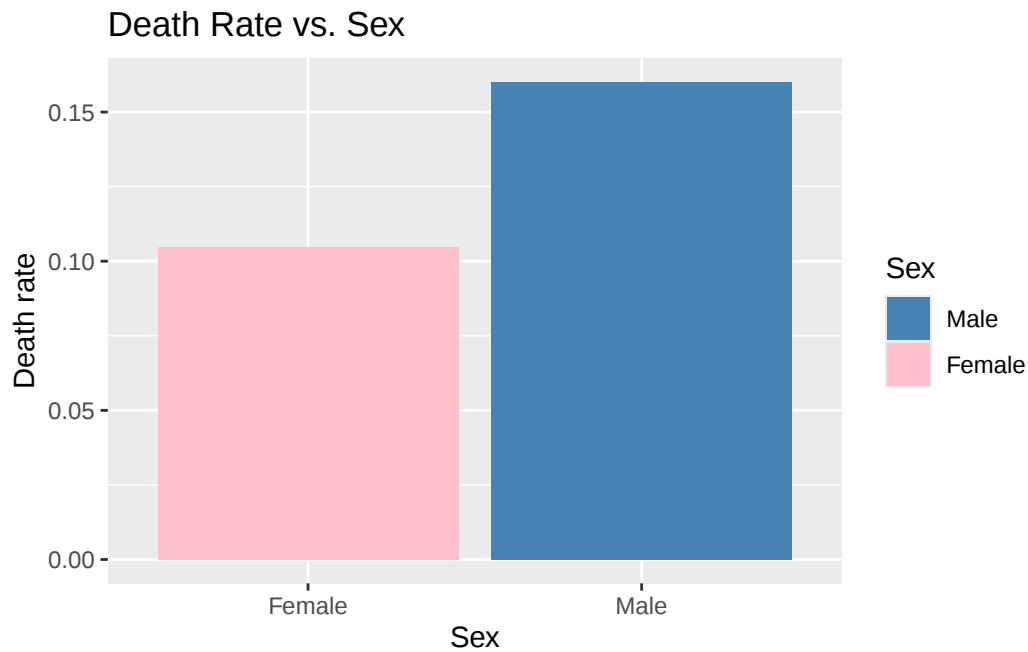
```
data%>%
  select(where(is.numeric))%>%
  cor(,use="pairwise.complete.obs")%>%
  corrplot(method="color")
```



From the correlation plot, we can see that the response variable `dead` has invert proportion to `sex` and `patient_type`, which means female gender and the patient returned home could decrease the risk of death. On the contrary, the pneumonia, age, diabetes, chronic obstructive, pulmonary disease hypertension, other disease, cardiovascular, obesity, chronic renal disease could increase the death probability. Surprisingly, the pregnancy, asthma and tobacco have no strong effect of death.

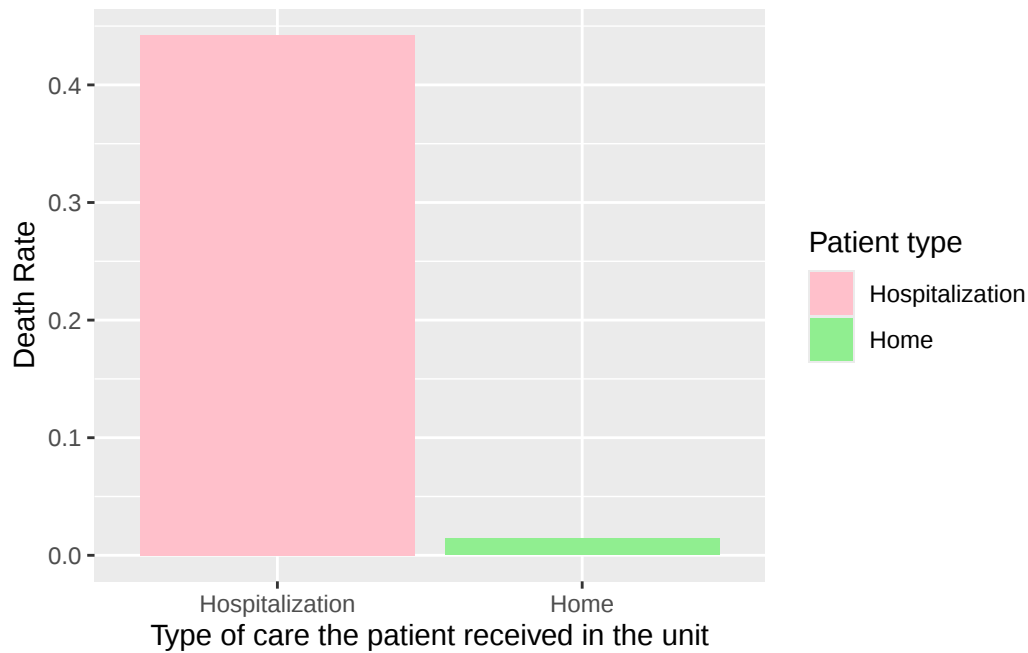
2.4 Bar charts to understand the variables

```
data%>%
  mutate(sex=factor(sex, levels=c(0,1), labels=c("Male", "Female")))%>%
  group_by(sex)%>%
  summarise(rate=sum(dead==1)/n())%>%
  ggplot(aes(x=as.character(sex), y=rate, fill=sex))+
  geom_col()+
  scale_fill_manual(
    name="Sex",
    values=c("Male"="steelblue", "Female"="pink"))+
  labs(x="Sex",
       y="Death rate",
       title="Death Rate vs. Sex")
```



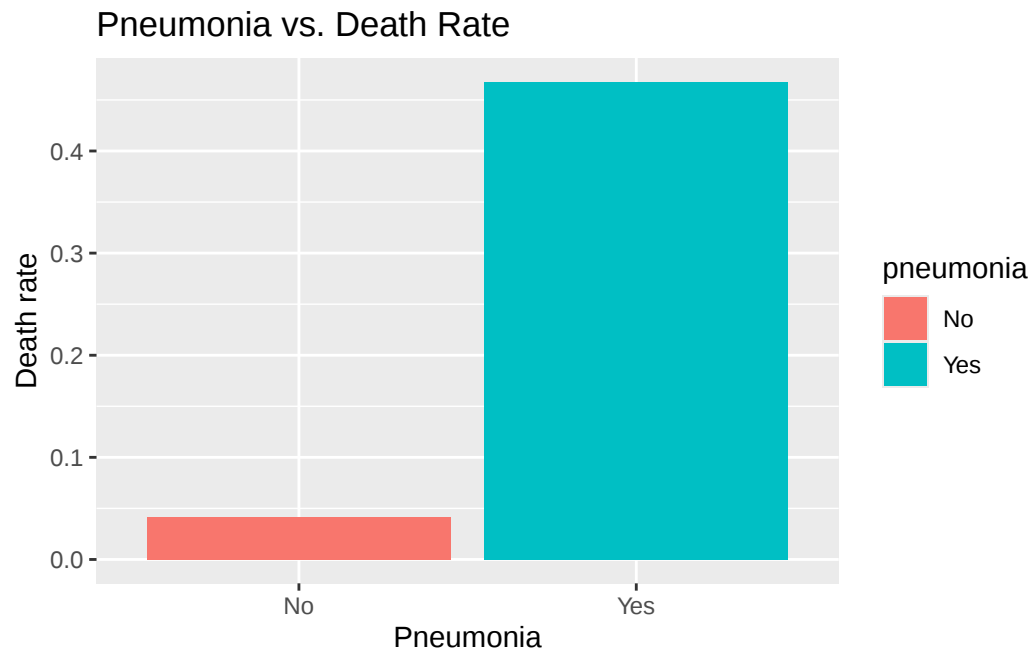
According to the bar charts, the COVID-19 death rate for males is 16%, compared to 10.5% for females.

```
data%>%
  mutate(patient_type=factor(patient_type, levels=c(0,1), labels=c("Hospitalization", "Home")))
  group_by(patient_type)%>%
  summarise(rate=sum(dead==1)/n())%>%
  ggplot(aes(patient_type, y=rate, fill=patient_type))+
  geom_col()+
  scale_fill_manual(name="Patient type",
                    values=c("Hospitalization"="pink",
                              "Home"="lightgreen"))+
  labs(y="Death Rate",
       x="Type of care the patient received in the unit")
```



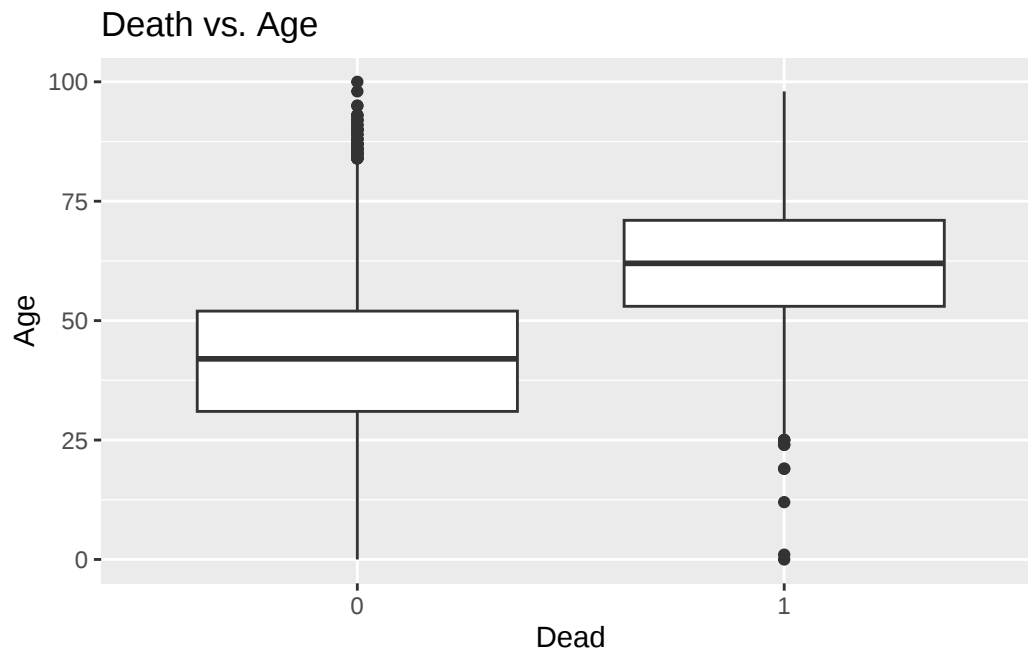
According to the bar charts, the COVID-19 death rate for Hospitalization is 44.3%, compared to 1.4% for home stay. Indeed the severe cases of COVID patient will stay and hospital even in ICU and have much high rate of death, on the contrary, the non-severe cases will be sent to home, and not likely to die.

```
data%>%
  mutate(pneumonia=factor(pneumonia, levels=c(0,1), labels=c("No","Yes")))%>%
  group_by(pneumonia)%>%
  summarise(rate=sum(dead==1)/n())%>%
  ggplot(aes(x=pneumonia, y=rate, fill=pneumonia))+
  geom_col()+
  labs(y="Death rate",
       x="Pneumonia",
       title="Pneumonia vs. Death Rate")
```



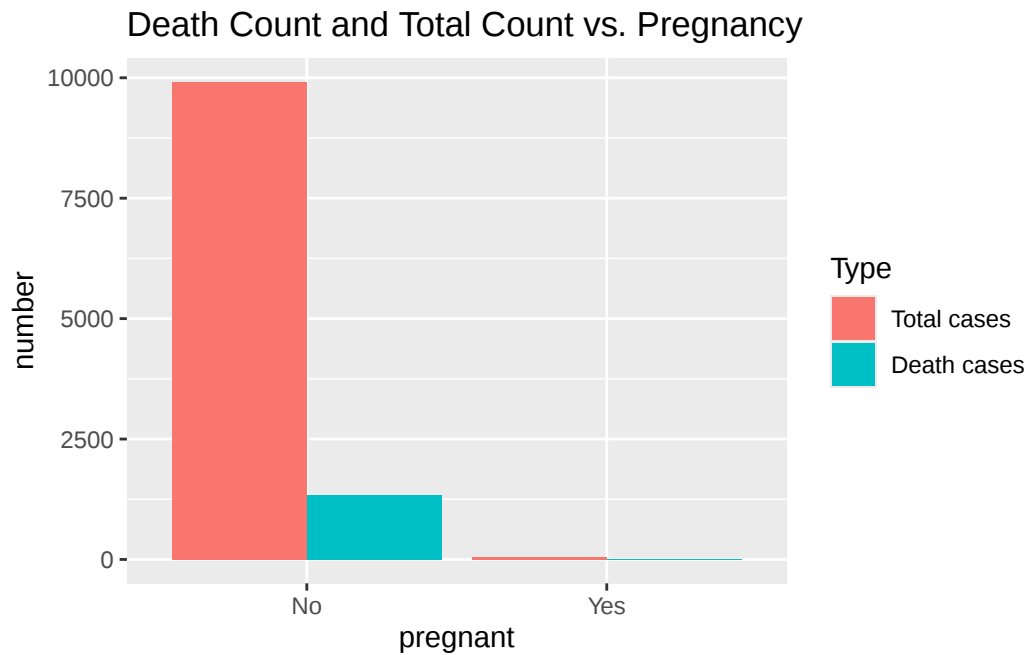
According to bar charts, the COVID death rate for pneumonia patients is 46.8%, compared to 4.2% for non-pneumonia samples.

```
data%>%  
  ggplot(aes(x=as.character(dead), y=age))+  
  geom_boxplot()+  
  labs(x="Dead",  
       y="Age",  
       title="Death vs. Age")
```



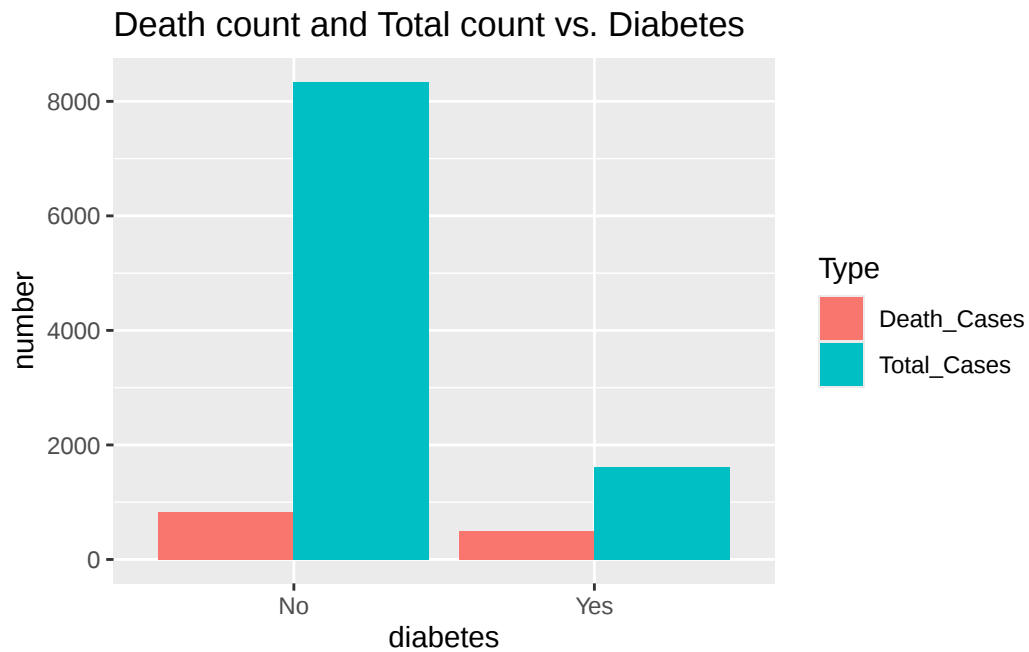
According to box plot, the COVID death group have the mean age 62 years old, older than the non-dead group which is 42 years old.

```
data%>%
  filter(!is.na(pregnant))%>%
  mutate(pregnant=factor(pregnant, levels=c(0,1), labels=c("No", "Yes")))%>%
  group_by(pregnant)%>%
  summarise(death=sum(dead==1),count=n(), rate=death/count)%>%
  pivot_longer(cols=c(death, count), names_to="type", values_to="number")%>%
  ggplot(aes(x=pregnant, y=number, fill=type))+
  geom_col(position="dodge")+
  labs(title="Death Count and Total Count vs. Pregnancy")+
  scale_fill_discrete(name="Type",
    labels=c("Total cases", "Death cases"))
```



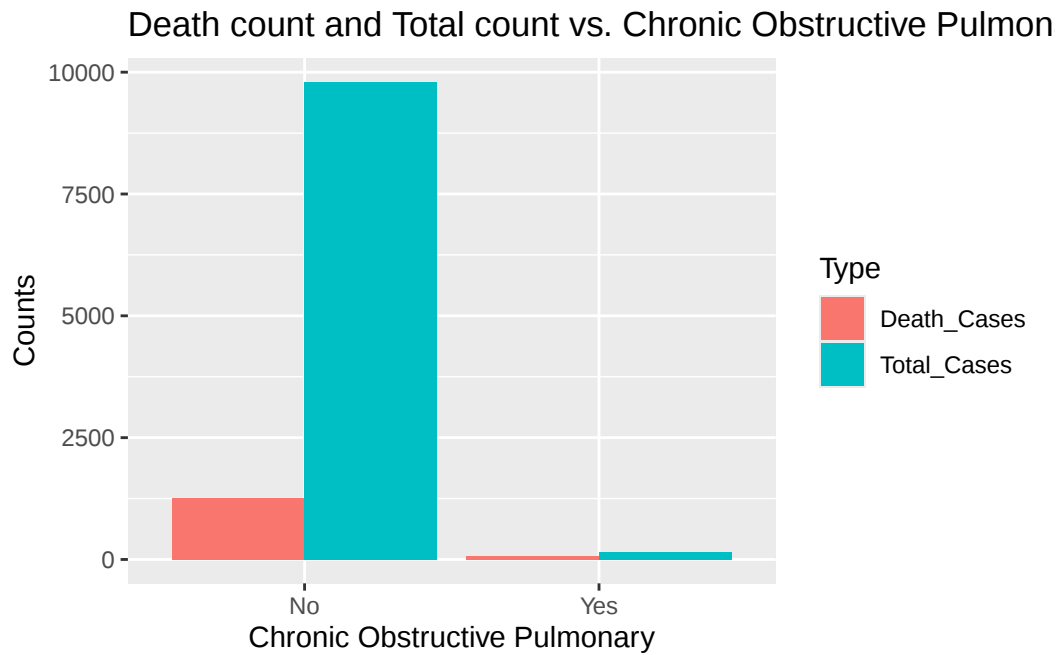
According to the bar chart, the COVID death rate for pregnancy sample are only 5.4%, compared to the 13.5% of non-pregnancy samples. It seems pregnancy could reduce risk of COVID death.

```
data%>%
  filter(!is.na(diabetes))%>%
  mutate(diabetes=factor(diabetes, levels=c("0", "1"), labels=c("No", "Yes")))%>%
  group_by(diabetes)%>%
  summarise(Death_Cases=sum(dead==1), Total_Cases=n(), Rate=Death_Cases/Total_Cases)%>%
  pivot_longer(cols=c(Death_Cases, Total_Cases), names_to="Type", values_to="number")%>%
  ggplot(aes(x=diabetes, y=number, fill=Type))+
  geom_col(position="dodge")+
  labs(title="Death count and Total count vs. Diabetes")
```



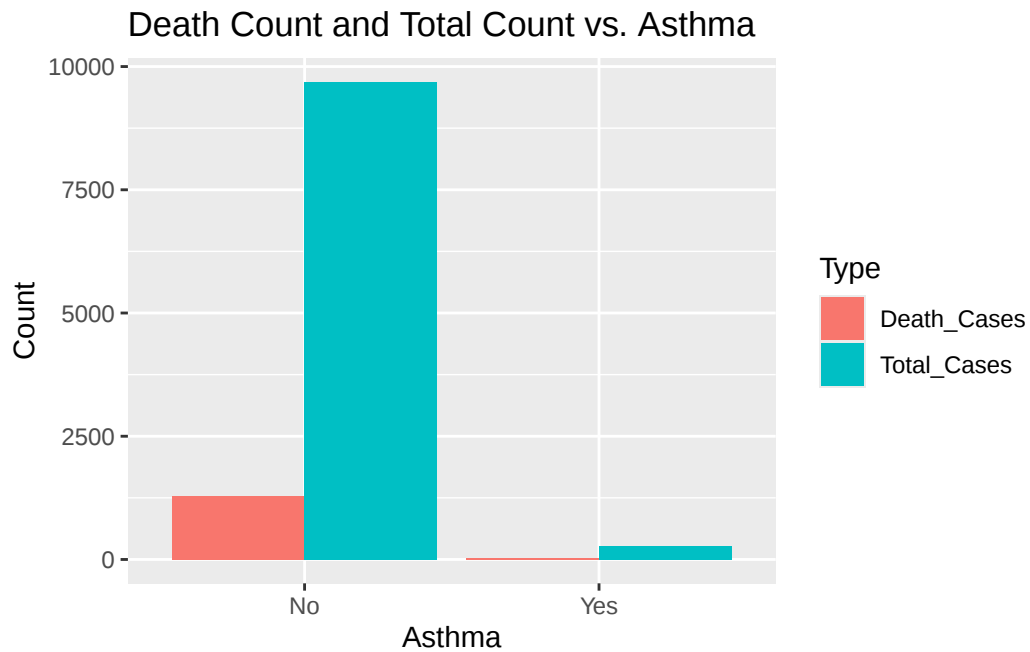
According to the bar chart, the COVID death rate of diabetes group is 31.2%, compared to the 9.9% of non-diabetes group.

```
data%>%
  filter(!is.na(copd))%>%
  mutate(copd=factor(copd, levels=c(0,1), labels=c("No", "Yes")))%>%
  group_by(copd)%>%
  summarise(Death_Cases=sum(dead==1), Total_Cases=n(), rate=Death_Cases/Total_Cases)%>%
  pivot_longer(cols=c(Death_Cases, Total_Cases), names_to = "Type", values_to="number")%>%
  ggplot(aes(x=copd,y=number,fill=Type))+
  geom_col(position="dodge")+
  labs(title="Death count and Total count vs. Chronic Obstructive Pulmonary",
       x="Chronic Obstructive Pulmonary",
       y="Counts")
```



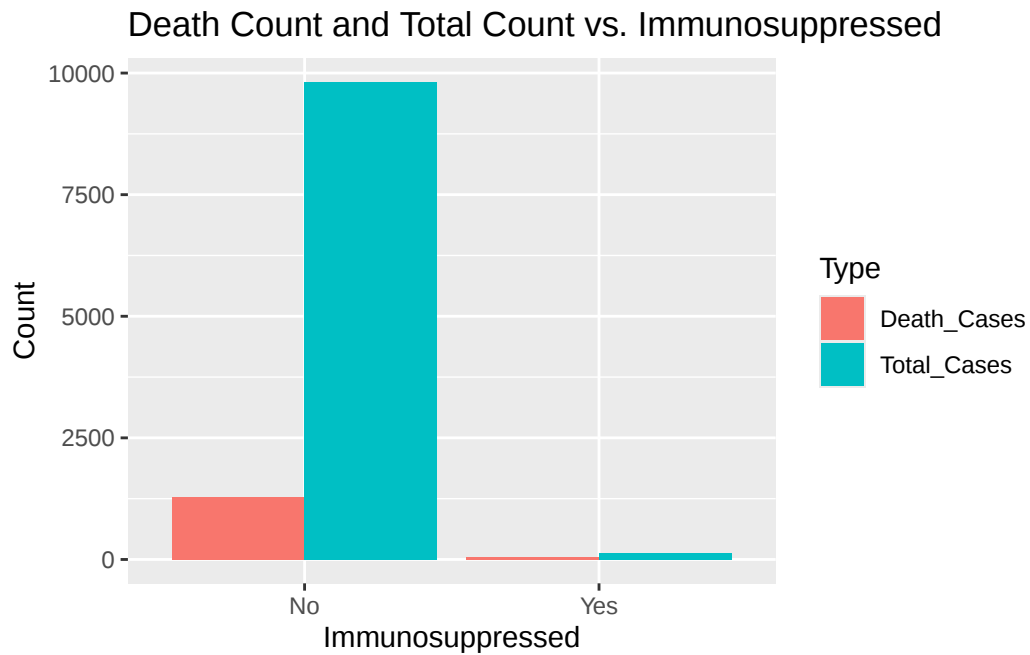
According to the bar chart, the COVID death rate for Chronic obstructive pulmonary disease group is 47.5%, compared to the 12.8% of non-Chronic obstructive pulmonary group.

```
data%>%
  filter(!is.na(asthma))%>%
  mutate(asthma=factor(asthma, levels=c(0, 1), labels=c("No", "Yes")))%>%
  group_by(asthma)%>%
  summarise(Death_Cases=sum(dead==1), Total_Cases=n(), Rate=Death_Cases/Total_Cases)%>%
  pivot_longer(cols=c(Death_Cases, Total_Cases), names_to="Type", values_to="Number")%>%
  ggplot(aes(x=asthma, y=Number, fill=Type))+
  geom_col(position="dodge")+
  labs(title="Death Count and Total Count vs. Asthma",
       x="Asthma",
       y="Count")
```



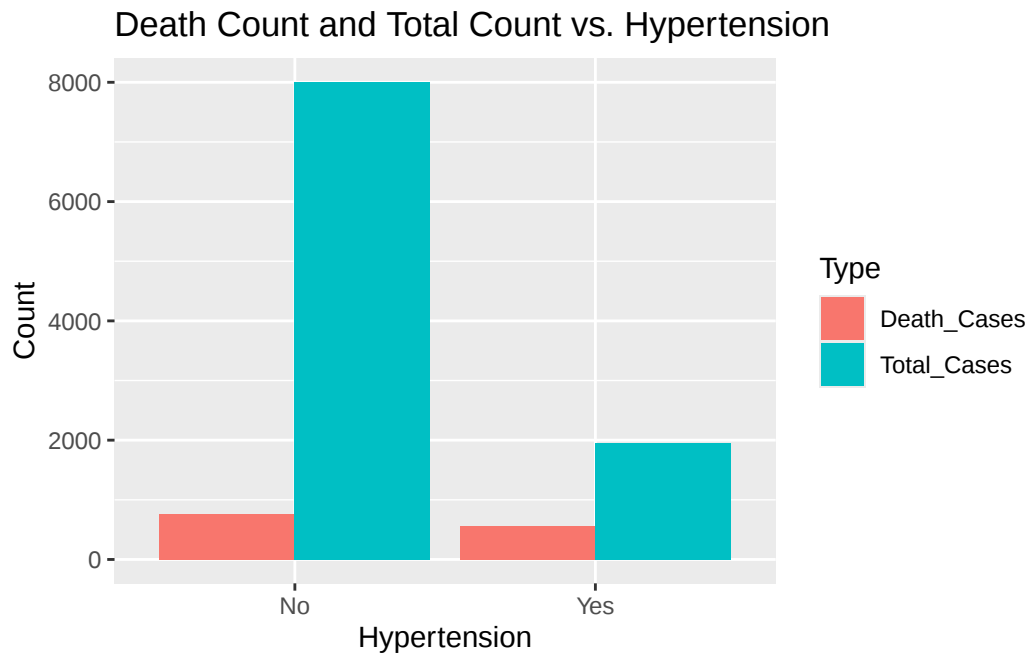
According to the bar chart, the COVID death rate for asthma group is 13.4%, compared to the 12.6% of non-asthma group. It aligns with the correlation plot that asthma has no strong influence of COVID death.

```
data%>%
  filter(!is.na(inmsupr))%>%
  mutate(inmsupr=factor(inmsupr, levels=c(0, 1), labels=c("No", "Yes")))%>%
  group_by(inmsupr)%>%
  summarise(Death_Cases=sum(dead==1), Total_Cases=n(), Rate=Death_Cases/Total_Cases)%>%
  pivot_longer(cols=c(Death_Cases, Total_Cases), names_to = "Type", values_to="Count")%>%
  ggplot(aes(x=inmsupr, y=Count, fill=Type))+
  geom_col(position="dodge")+
  labs(x="Immunosuppressed",
       title="Death Count and Total Count vs. Immunosuppressed")
```



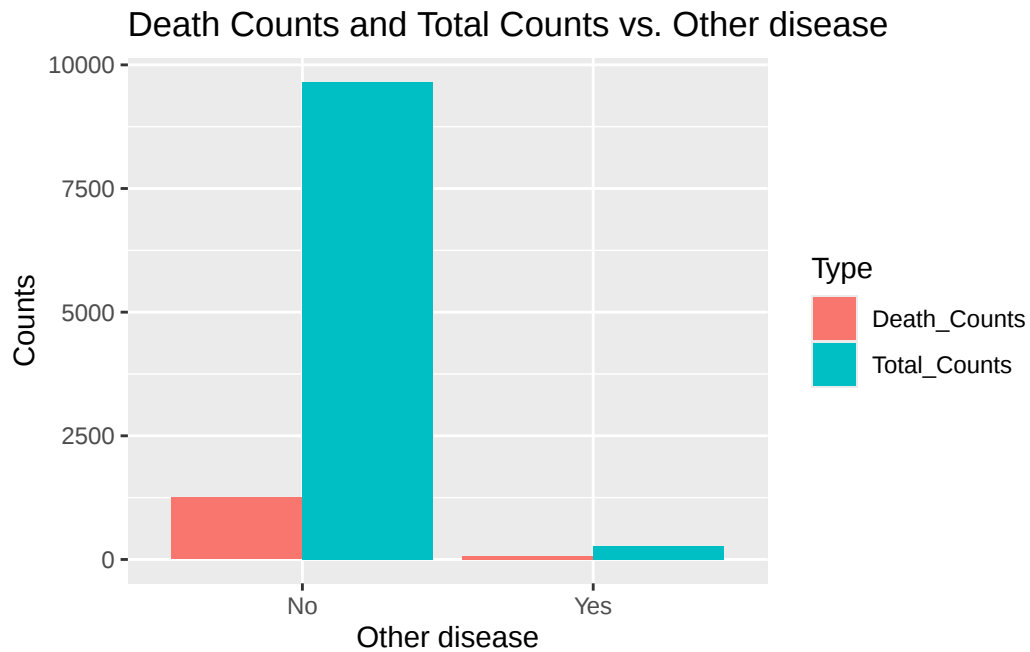
According to the bar chart, the COVID death rate for immunosuppressed group is 33.6%, compared to the 13.1% of non-immunosuppressed group.

```
data%>%
  filter(!is.na(hipertension))%>%
  mutate(hipertension=factor(hipertension, levels=c(0,1), labels=c("No", "Yes")))%>%
  group_by(hipertension)%>%
  summarise(Death_Cases=sum(dead==1), Total_Cases=n(), Rate=Death_Cases/Total_Cases) %>%
  pivot_longer(cols=c(Death_Cases, Total_Cases), names_to="Type", values_to="Count")%>%
  ggplot(aes(x=hipertension, y=Count, fill=Type))+
  geom_col(position="dodge")+
  labs(x="Hypertension",
       title="Death Count and Total Count vs. Hypertension")
```



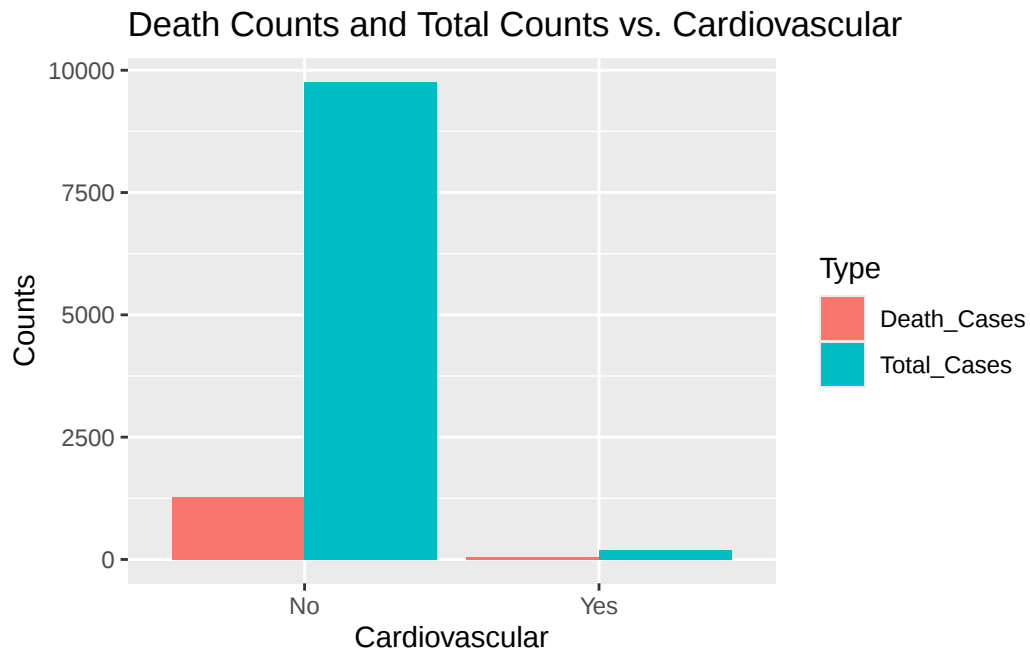
According to the bar plot, the COVID death rate for hypertension group is 28.9%, compare to the 9.6% of non-hypertension group.

```
data%>%
  filter(!is.na(other_disease))%>%
  mutate(other_disease=factor(other_disease, levels=c(0,1), labels=c("No", "Yes")))%>%
  group_by(other_disease)%>%
  summarise(Death_Counts=sum(dead==1), Total_Counts=n(), Rate=Death_Counts/Total_Counts)%>%
  pivot_longer(cols=c(Death_Counts, Total_Counts), names_to="Type", values_to="Counts")%>%
  ggplot(aes(x=other_disease, y=Counts, fill=Type))+
  geom_col(position="dodge")+
  labs(title="Death Counts and Total Counts vs. Other disease",
       x="Other disease")
```



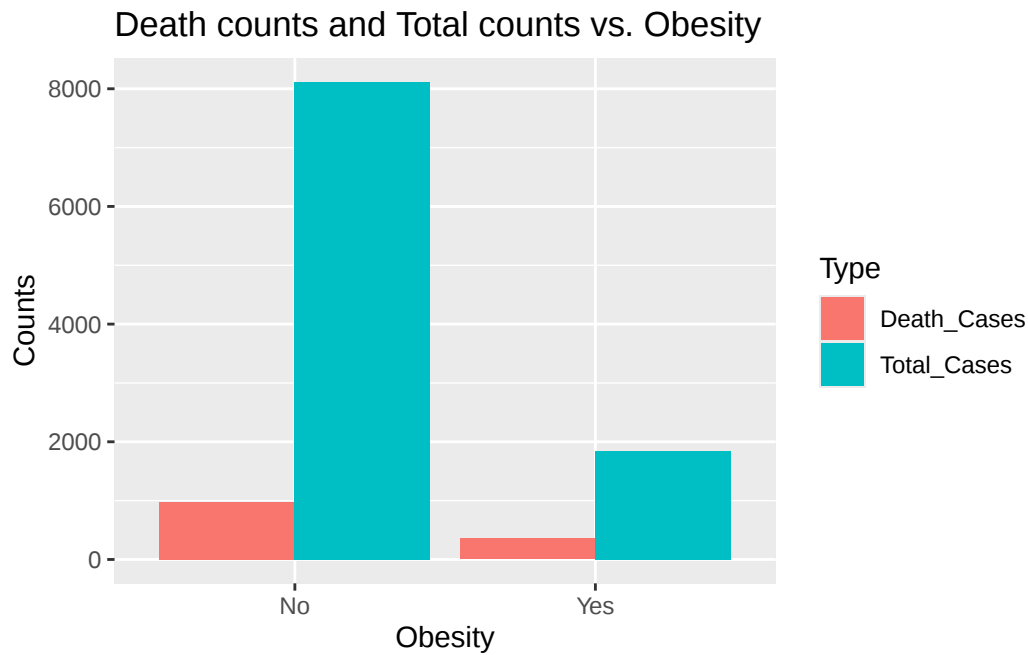
According to the bar plot, the COVID death rate for other disease group is 26%, compared to 13.0% of non-other disease group. which means the risk of death could be doubled for the other disease than the non-other disease.

```
data%>%
  filter(!is.na(cardiovascular))%>%
  mutate(cardiovascular=factor(cardiovascular, levels=c(0,1), label=c("No", "Yes")))%>%
  group_by(cardiovascular)%>%
  summarise(Death_Cases=sum(dead==1), Total_Cases=n(), Rate=Death_Cases/Total_Cases)%>%
  pivot_longer(cols=c(Death_Cases, Total_Cases), names_to="Type", values_to="Counts")%>%
  ggplot(aes(x=cardiovascular, y=Counts, fill=Type))+
  geom_col(position="dodge")+
  labs(x="Cardiovascular",
       title="Death Counts and Total Counts vs. Cardiovascular")
```



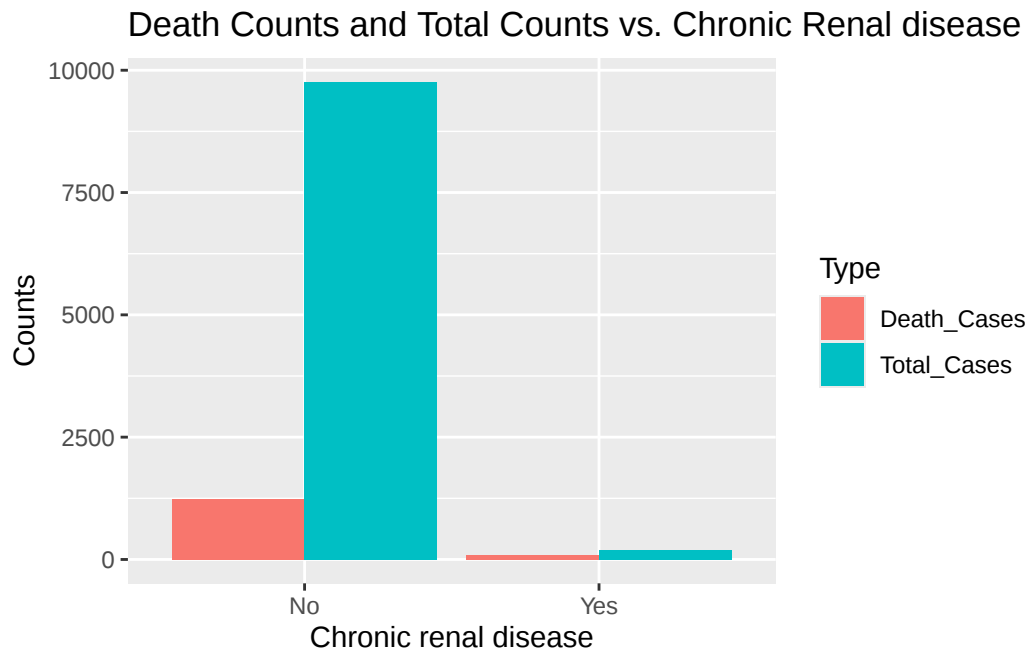
According to the plot, the COVID death rate for the cardiovascular group is 29.4%, compare to the 13% for non-cardiovascular group.

```
data%>%
  filter(!is.na(obesity))%>%
  mutate(obesity=factor(obesity, levels=c(0,1), labels=c("No", "Yes")))%>%
  group_by(obesity)%>%
  summarise(Death_Cases=sum(dead==1), Total_Cases=n(), Rate=Death_Cases/Total_Cases)%>%
  pivot_longer(cols=c(Death_Cases, Total_Cases), names_to="Type", values_to="Counts") %>%
  ggplot(aes(x=obesity, y=Counts, fill=Type))+
  geom_col(position="dodge")+
  labs(title="Death counts and Total counts vs. Obesity",
        x="Obesity")
```



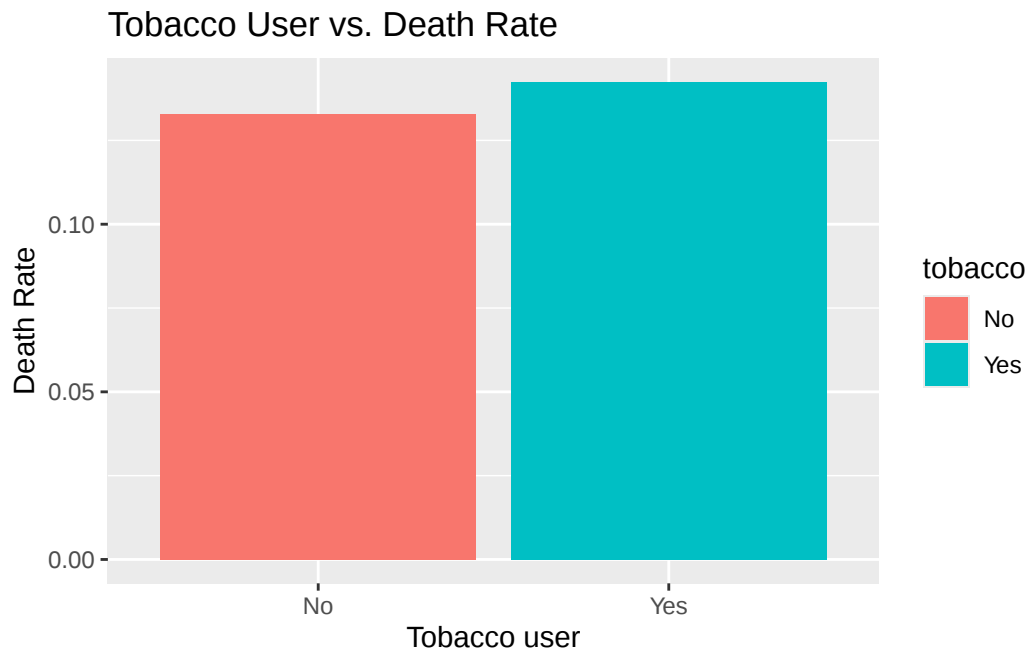
According to the bar plot, the COVID death rate for the obesity group is 19.3%, compared to 12% for non-obesity group.

```
data%>%
  filter(!is.na(renal_chronic))%>%
  mutate(renal_chronic=factor(renal_chronic, levels=c(0,1), labels=c("No", "Yes")))%>%
  group_by(renal_chronic)%>%
  summarise(Death_Cases=sum(dead==1), Total_Cases=n(), Rate=Death_Cases/Total_Cases)%>%
  pivot_longer(cols=c(Death_Cases, Total_Cases), names_to="Type", values_to="Counts")%>%
  ggplot(aes(x=renal_chronic, y=Counts, fill=Type))+
  geom_col(position="dodge")+
  labs(x="Chronic renal disease",
       title="Death Counts and Total Counts vs. Chronic Renal disease")
```



According to the bar plot, the COVID death rate for the chronic renal disease group is 44.6%, compared to 12.7% for non-chronic renal disease group.

```
data%>%
  filter(!is.na(tobacco))%>%
  mutate(tobacco=factor(tobacco, levels=c(0,1), labels=c("No", "Yes")))%>%
  group_by(tobacco)%>%
  summarise(rate=sum(dead==1)/n())%>%
  ggplot(aes(x=tobacco, y=rate, fill=tobacco))+
  geom_col()+
  labs(x="Tobacco user",
       y="Death Rate",
       title="Tobacco User vs. Death Rate")
```



According to the bar plot, the COVID death rate for tobacco group is 14.2, compared to 13.3% for non-tobacco group. it seems the tobacco has no strong influence to the COVID death, which aligns with the correlation plot.

3 Logistic Regression

3.1 Initial model

Convert numerical to factor for the categorical variable.

```
data<-na.omit(data)
data$sex<-factor(data$sex, levels=c(0,1), labels=c("Male", "Female"))
data$patient_type<-factor(data$patient_type, levels=c(0,1), labels=c("Hospitalization", "H
data<-data%>%
  mutate(across(
    .cols=-c(sex, patient_type, age),
    .fns=~ factor(.x, levels=c(0,1), labels=c("No", "Yes"))
  ))
model1<-glm(dead~., data=data, family=binomial,na.action=na.omit)
summary(model1)#4.77.7
```

Call:

```
glm(formula = dead ~ ., family = binomial, data = data, na.action = na.omit)
```

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)	
(Intercept)	-4.215678	0.194356	-21.690	< 2e-16	***
sexFemale	-0.372940	0.083564	-4.463	8.08e-06	***
patient_typeHome	-2.645381	0.121841	-21.712	< 2e-16	***
pneumoniaYes	1.166860	0.090749	12.858	< 2e-16	***
age	0.054172	0.003015	17.969	< 2e-16	***
pregnantYes	0.943970	0.655453	1.440	0.149816	
diabetesYes	0.142597	0.089949	1.585	0.112896	
copdYes	0.236625	0.214927	1.101	0.270916	
asthmaYes	0.145458	0.255453	0.569	0.569077	
inmsuprYes	0.289245	0.254108	1.138	0.255005	
hipertensionYes	0.111080	0.090806	1.223	0.221231	
other_diseaseYes	0.383521	0.192124	1.996	0.045910	*
cardiovascularYes	-0.430550	0.218019	-1.975	0.048287	*
obesityYes	0.336188	0.093131	3.610	0.000306	***
renal_chronicYes	0.736563	0.197762	3.724	0.000196	***
tobaccoYes	-0.191194	0.152418	-1.254	0.209694	

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 7763.7 on 9886 degrees of freedom
Residual deviance: 4145.7 on 9871 degrees of freedom
AIC: 4177.7

Number of Fisher Scoring iterations: 7

3.2 Model selection

```
step_model=step(model1, direction="both", trace=1)
```

Start: AIC=4177.7

dead ~ sex + patient_type + pneumonia + age + pregnant + diabetes +
copd + asthma + inmsupr + hipertension + other_disease +
cardiovascular + obesity + renal_chronic + tobacco

	Df	Deviance	AIC
- asthma	1	4146.0	4176.0
- copd	1	4146.9	4176.9
- inmsupr	1	4147.0	4177.0

- hypertension	1	4147.2	4177.2
- tobacco	1	4147.3	4177.3
- pregnant	1	4147.4	4177.4
<none>		4145.7	4177.7
- diabetes	1	4148.2	4178.2
- cardiovascular	1	4149.6	4179.6
- other_disease	1	4149.6	4179.6
- obesity	1	4158.6	4188.6
- renal_chronic	1	4159.7	4189.7
- sex	1	4165.8	4195.8
- pneumonia	1	4317.2	4347.2
- age	1	4512.1	4542.1
- patient_type	1	4739.5	4769.5

Step: AIC=4176.02

dead ~ sex + patient_type + pneumonia + age + pregnant + diabetes +
 copd + inmsupr + hypertension + other_disease + cardiovascular +
 obesity + renal_chronic + tobacco

	Df	Deviance	AIC
- copd	1	4147.3	4175.3
- inmsupr	1	4147.4	4175.4
- hypertension	1	4147.5	4175.5
- tobacco	1	4147.6	4175.6
- pregnant	1	4147.7	4175.7
<none>		4146.0	4176.0
- diabetes	1	4148.5	4176.5
+ asthma	1	4145.7	4177.7
- cardiovascular	1	4149.9	4177.9
- other_disease	1	4150.0	4178.0
- obesity	1	4159.1	4187.1
- renal_chronic	1	4160.1	4188.1
- sex	1	4165.9	4193.9
- pneumonia	1	4317.3	4345.3
- age	1	4512.2	4540.2
- patient_type	1	4739.9	4767.9

Step: AIC=4175.34

dead ~ sex + patient_type + pneumonia + age + pregnant + diabetes +
 inmsupr + hypertension + other_disease + cardiovascular +
 obesity + renal_chronic + tobacco

	Df	Deviance	AIC
- tobacco	1	4148.6	4174.6
- inmsupr	1	4148.7	4174.7
- hypertension	1	4148.9	4174.9

- pregnant	1	4149.1	4175.1
<none>		4147.3	4175.3
- diabetes	1	4149.9	4175.9
+ copd	1	4146.0	4176.0
- cardiovascular	1	4150.9	4176.9
+ asthma	1	4146.9	4176.9
- other_disease	1	4151.4	4177.4
- obesity	1	4160.5	4186.5
- renal_chronic	1	4161.8	4187.8
- sex	1	4167.0	4193.0
- pneumonia	1	4318.1	4344.1
- age	1	4529.4	4555.4
- patient_type	1	4743.4	4769.4

Step: AIC=4174.63

```
dead ~ sex + patient_type + pneumonia + age + pregnant + diabetes +
      inmsupr + hipertension + other_disease + cardiovascular +
      obesity + renal_chronic
```

	Df	Deviance	AIC
- inmsupr	1	4150.1	4174.1
- hipertension	1	4150.1	4174.1
- pregnant	1	4150.4	4174.4
<none>		4148.6	4174.6
- diabetes	1	4151.1	4175.1
+ tobacco	1	4147.3	4175.3
+ copd	1	4147.6	4175.6
+ asthma	1	4148.2	4176.2
- other_disease	1	4152.5	4176.5
- cardiovascular	1	4152.6	4176.6
- obesity	1	4161.4	4185.4
- renal_chronic	1	4163.0	4187.0
- sex	1	4167.4	4191.4
- pneumonia	1	4319.3	4343.3
- age	1	4530.9	4554.9
- patient_type	1	4744.4	4768.4

Step: AIC=4174.07

```
dead ~ sex + patient_type + pneumonia + age + pregnant + diabetes +
      hipertension + other_disease + cardiovascular + obesity +
      renal_chronic
```

	Df	Deviance	AIC
- hipertension	1	4151.5	4173.5
- pregnant	1	4151.7	4173.7
<none>		4150.1	4174.1

+ inmsupr	1	4148.6	4174.6
- diabetes	1	4152.7	4174.7
+ tobacco	1	4148.7	4174.7
+ copd	1	4149.0	4175.0
+ asthma	1	4149.6	4175.6
- cardiovascular	1	4153.9	4175.9
- other_disease	1	4154.7	4176.7
- obesity	1	4162.7	4184.7
- renal_chronic	1	4165.4	4187.4
- sex	1	4168.3	4190.3
- pneumonia	1	4322.1	4344.1
- age	1	4531.5	4553.5
- patient_type	1	4747.4	4769.4

Step: AIC=4173.52

dead ~ sex + patient_type + pneumonia + age + pregnant + diabetes +
other_disease + cardiovascular + obesity + renal_chronic

	Df	Deviance	AIC
- pregnant	1	4153.2	4173.2
<none>		4151.5	4173.5
+ hypertension	1	4150.1	4174.1
+ inmsupr	1	4150.1	4174.1
+ tobacco	1	4150.2	4174.2
+ copd	1	4150.5	4174.5
+ asthma	1	4151.0	4175.0
- cardiovascular	1	4155.1	4175.1
- diabetes	1	4155.3	4175.3
- other_disease	1	4156.2	4176.2
- obesity	1	4165.4	4185.4
- renal_chronic	1	4168.1	4188.1
- sex	1	4169.0	4189.0
- pneumonia	1	4323.6	4343.6
- age	1	4572.2	4592.2
- patient_type	1	4750.2	4770.2

Step: AIC=4173.18

dead ~ sex + patient_type + pneumonia + age + diabetes + other_disease +
cardiovascular + obesity + renal_chronic

	Df	Deviance	AIC
<none>		4153.2	4173.2
+ pregnant	1	4151.5	4173.5
+ hypertension	1	4151.7	4173.7
+ inmsupr	1	4151.8	4173.8
+ tobacco	1	4151.8	4173.8

```

+ copd          1  4152.1 4174.1
- cardiovascular 1  4156.7 4174.7
+ asthma        1  4152.7 4174.7
- diabetes      1  4156.9 4174.9
- other_disease 1  4158.0 4176.0
- obesity       1  4166.7 4184.7
- renal_chronic 1  4169.7 4187.7
- sex           1  4170.0 4188.0
- pneumonia     1  4324.5 4342.5
- age           1  4573.3 4591.3
- patient_type  1  4754.0 4772.0

```

```
summary(step_model)#4173.2
```

Call:

```

glm(formula = dead ~ sex + patient_type + pneumonia + age + diabetes +
     other_disease + cardiovascular + obesity + renal_chronic,
     family = binomial, data = data, na.action = na.omit)

```

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)	
(Intercept)	-4.243348	0.190645	-22.258	< 2e-16	***
sexFemale	-0.333959	0.081797	-4.083	4.45e-05	***
patient_typeHome	-2.656950	0.121795	-21.815	< 2e-16	***
pneumoniaYes	1.163655	0.090571	12.848	< 2e-16	***
age	0.055035	0.002886	19.068	< 2e-16	***
diabetesYes	0.169102	0.087173	1.940	0.052400	.
other_diseaseYes	0.421382	0.190050	2.217	0.026608	*
cardiovascularYes	-0.402562	0.215013	-1.872	0.061169	.
obesityYes	0.339451	0.092045	3.688	0.000226	***
renal_chronicYes	0.788654	0.195366	4.037	5.42e-05	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

```

Null deviance: 7763.7 on 9886 degrees of freedom
Residual deviance: 4153.2 on 9877 degrees of freedom
AIC: 4173.2

```

Number of Fisher Scoring iterations: 7

```
vif(step_model)
```

	sex	patient_type	pneumonia	age	diabetes
	1.021460	1.237874	1.224106	1.070543	1.079871
other_disease		cardiovascular	obesity	renal_chronic	
	1.011292	1.036287	1.035004	1.040190	

```
model2 <- glm(dead ~ sex + patient_type + pneumonia + age + diabetes +  
              other_disease + obesity + renal_chronic,  
              data = data, family = binomial)  
summary(model2)#4174.7
```

Call:

```
glm(formula = dead ~ sex + patient_type + pneumonia + age + diabetes +  
     other_disease + obesity + renal_chronic, family = binomial,  
     data = data)
```

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)	
(Intercept)	-4.208451	0.189427	-22.217	< 2e-16	***
sexFemale	-0.336155	0.081769	-4.111	3.94e-05	***
patient_typeHome	-2.659968	0.121770	-21.844	< 2e-16	***
pneumoniaYes	1.157785	0.090452	12.800	< 2e-16	***
age	0.054388	0.002861	19.012	< 2e-16	***
diabetesYes	0.161562	0.087076	1.855	0.063539	.
other_diseaseYes	0.407101	0.190022	2.142	0.032162	*
obesityYes	0.327149	0.091811	3.563	0.000366	***
renal_chronicYes	0.763846	0.195204	3.913	9.11e-05	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 7763.7 on 9886 degrees of freedom
Residual deviance: 4156.7 on 9878 degrees of freedom
AIC: 4174.7

Number of Fisher Scoring iterations: 7

```
vif(model2)
```

	sex	patient_type	pneumonia	age	diabetes
--	-----	--------------	-----------	-----	----------

1.020813	1.237986	1.222191	1.053810	1.078448
other_disease	obesity	renal_chronic		
1.009622	1.029843	1.034781		

After the step_model, the negative coefficient means the cardiovascular could decrease the risk of Covid death and also the statistical unsignificance, we remove this variable in our final model. We take model2 as our final model.

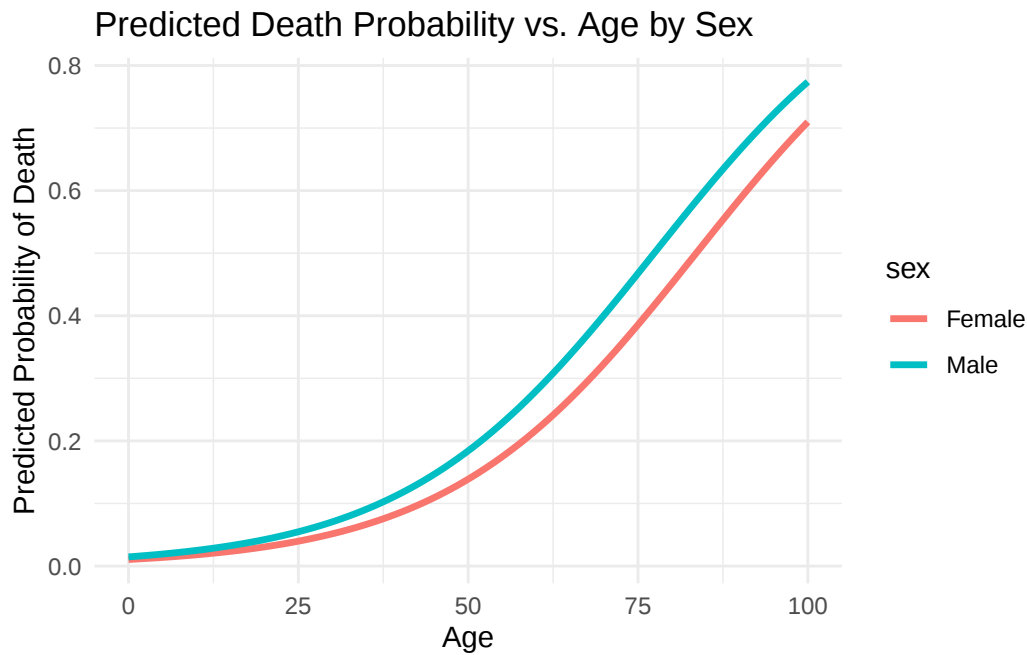
4 Model validation

4.1 Prediction plot

```
# Vary sex
new_data <- expand.grid(
  age = seq(min(data$age), max(data$age), length.out = 100),
  sex = factor(c("Male", "Female")),
  patient_type = factor("Hospitalization", levels = c("Home", "Hospitalization")),
  pneumonia = factor("No", levels = c("No", "Yes")),
  diabetes = factor("No", levels = c("No", "Yes")),
  other_disease = factor("No", levels = c("No", "Yes")),
  obesity = factor("No", levels = c("No", "Yes")),
  renal_chronic = factor("No", levels = c("No", "Yes"))
)

# Predict
new_data$predicted_prob <- predict(model2, newdata = new_data, type = "response")

# Plot by sex
ggplot(new_data, aes(x = age, y = predicted_prob, color = sex)) +
  geom_line(linewidth = 1.2) +
  labs(
    title = "Predicted Death Probability vs. Age by Sex",
    x = "Age",
    y = "Predicted Probability of Death"
  ) +
  theme_minimal()
```



From the line plot, we see the percentage of death due to COVID increases by the age of the patient. Also the female risk is lower than the male for COVID death.

```
library(purrr)
# 1. Find top 12 profiles (excluding sex)
top_profiles <- data %>%
  group_by(patient_type, pneumonia, diabetes, other_disease, obesity, renal_chronic) %>%
  summarise(count = n(), .groups = "drop") %>%
  arrange(desc(count)) %>%
  slice_head(n = 12)

# 2. Generate prediction data for each profile + both sexes
age_seq <- seq(min(data$age), max(data$age), length.out = 100)

# Expand to both Male and Female
new_data <- top_profiles %>%
  mutate(profile_id = row_number()) %>%
  group_split(profile_id) %>%
  map_dfr(function(profile) {
    profile_info <- profile %>% select(-count, -profile_id)

    # Add both sexes
    sexes <- c("Male", "Female")

    map_dfr(sexes, function(sex_val) {
      profile_replicated <- profile_info[rep(1, length(age_seq)), ]
```

```

    profile_replicated$sex <- factor(sex_val, levels = c("Male", "Female"))
    profile_replicated$age <- age_seq
    profile_replicated
  })
})

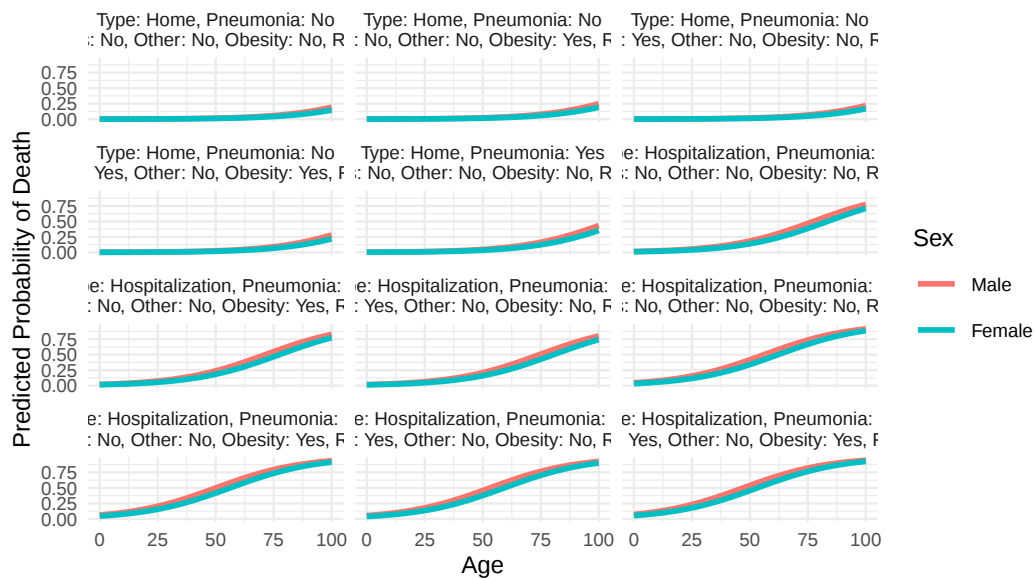
# 3. Predict
new_data$predicted_prob <- predict(model2, newdata = new_data, type = "response")

# 4. Create facet label
new_data$profile_label <- with(new_data, paste0(
  "Type: ", patient_type, ", Pneumonia: ", pneumonia, "\n",
  "Diabetes: ", diabetes, ", Other: ", other_disease,
  ", Obesity: ", obesity, ", Renal: ", renal_chronic
))

# 5. Plot with one curve per sex in each profile
ggplot(new_data, aes(x = age, y = predicted_prob, color = sex)) +
  geom_line(linewidth = 1) +
  facet_wrap(~ profile_label, ncol = 3) +
  labs(
    title = "Predicted Death Probability vs. Age\n(Top 12 Most Common Patient Profiles)",
    x = "Age",
    y = "Predicted Probability of Death",
    color = "Sex"
  ) +
  theme_minimal(base_size = 9) +
  theme(strip.text = element_text(size = 7))

```

Predicted Death Probability vs. Age (Top 12 Most Common Patient Profiles)



For the top most patient profiles in our data set, we use the final model to predict the COVID risk by age for different gender groups. We wish this could help the clinician decision makers.

4.2 Model accuracy

```
model_data<-model.frame(model2)

actual<-ifelse(model_data$dead=="Yes", 1,0)
predict_probs<-predict(model2, type="response")
predict_class<-ifelse(predict_probs>0.5, 1, 0)

Confusion_matrix=table(Predicted=predict_class, Actual=actual)
Confusion_matrix
```

```
      Actual
Predicted  0    1
      0 8152  609
      1  417  709
```

```
accuracy<-mean(predict_class==actual)
print(paste("Accuracy is", round(accuracy*100, 2),"%"))#89.62%
```

```
[1] "Accuracy is 89.62 %"
```

Our final model has the accuracy of 89.62%, which works quite well.