subject6

```
data = read.csv(file = 'Subject6_smoking.csv')
head(data,5)
```

```
##   Smoker Status  Age
## 1    Yes  Alive 21.0
## 2    Yes  Alive 19.3
## 3     No   Dead 57.5
## 4     No  Alive 47.1
## 5    Yes  Alive 81.4
```

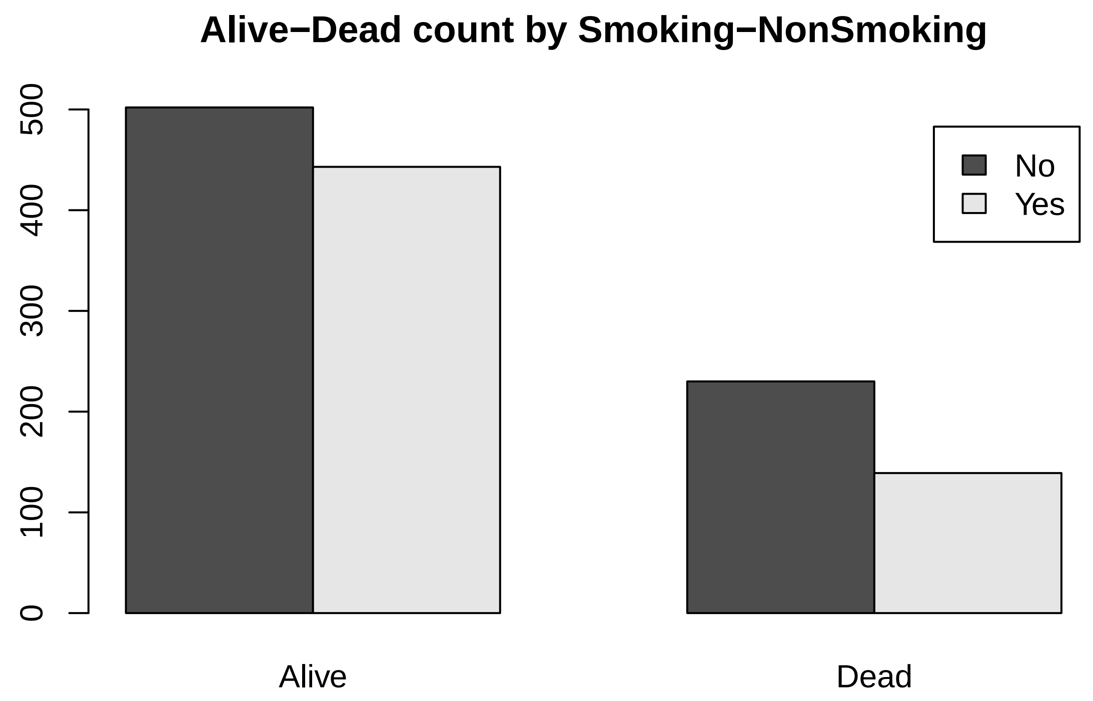
1

Two way table betwen if they are alive or dead and if yes or no they smoke

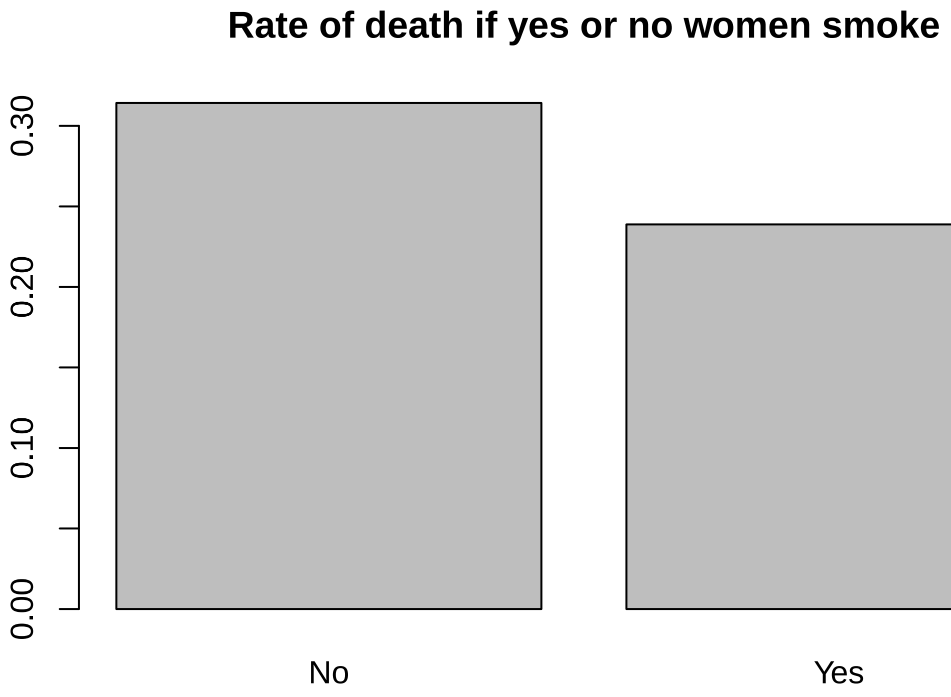
```
datatable = table(data$Smoker, data$Status)
addmargins(datatable)
```

```
##
##        Alive Dead  Sum
##   No     502  230  732
##   Yes    443  139  582
##   Sum    945  369 1314
```

```
barplot(datatable, legend=TRUE, beside=TRUE, main='Alive-Dead count by Smoking-NonSmoking')
```



```
#rate of death in the smoking and non smoking categories
rate_vector = (addmargins(datatable)[, 'Dead'] / addmargins(datatable)[, 'Sum'])[c('No', 'Yes')]
barplot(rate_vector, main="Rate of death if yes or no women smoke")
```



Explanation

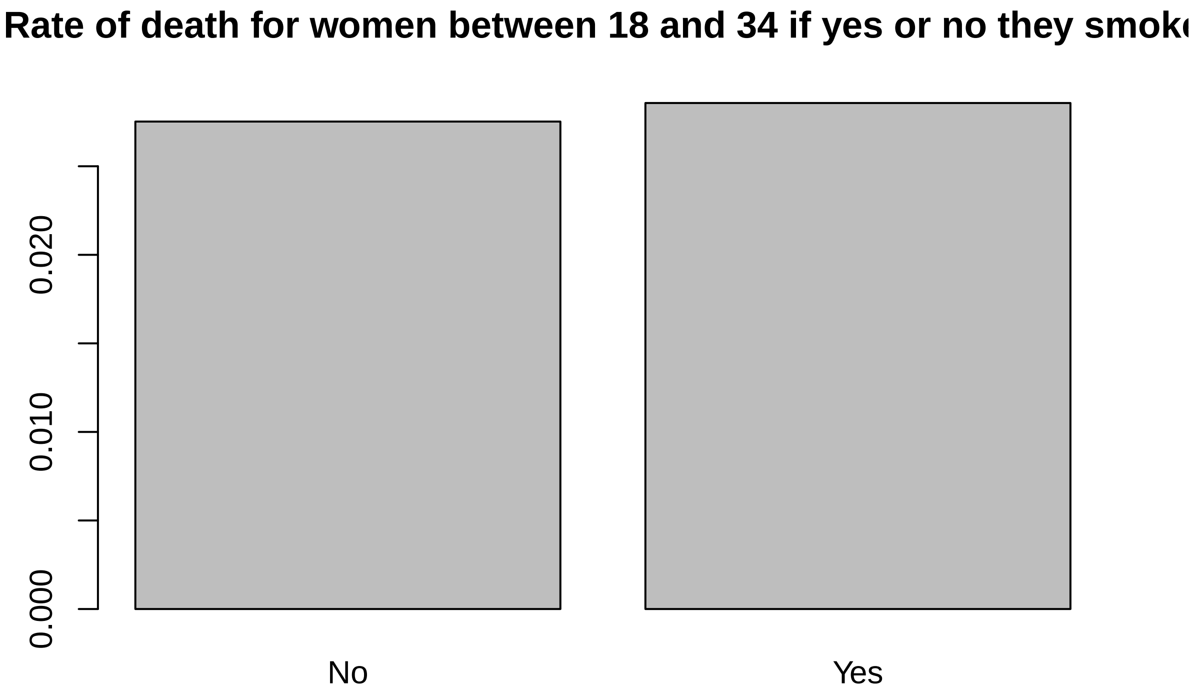
We can observe that the rate of mortality is higher in the women that didn't smoke and that is suprising because we would expect the contrary.

2

```
data_18_34 = data[data$Age > 18 & data$Age < 34, ]
datatable_18_34 = addmargins(table(data_18_34$Smoker, data_18_34$Status))
datatable_18_34
```

```
##
##       Alive Dead Sum
##   No    212    6 218
##   Yes   170    5 175
##   Sum   382   11 393
```

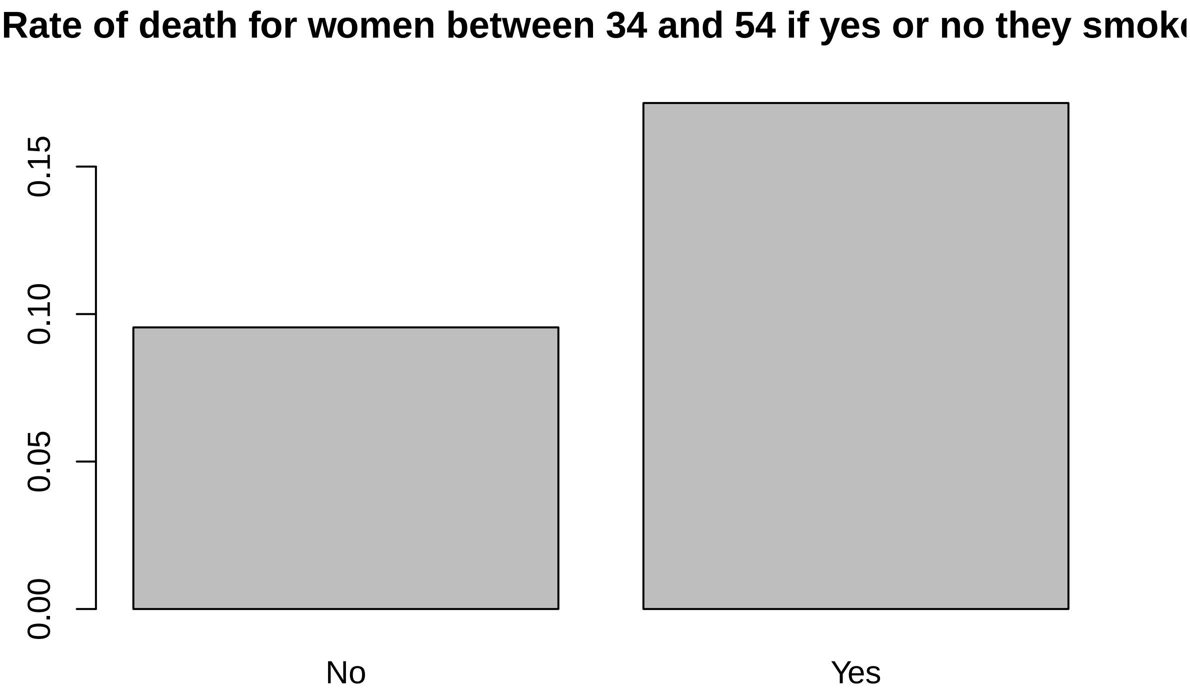
```
rate_vector_18_34 = (datatable_18_34[, 'Dead'] / datatable_18_34[, 'Sum'])[c('No', 'Yes')]
barplot(rate_vector_18_34, main="Rate of death for women between 18 and 34 if yes or no they smoke")
```



```
data_34_54 = data[data$Age >= 34 & data$Age < 54, ]
datatable_34_54 = addmargins(table(data_34_54$Smoker, data_34_54$Status))
datatable_34_54
```

```
##
##        Alive Dead Sum
##   No     180   19 199
##   Yes    198   41 239
##   Sum    378   60 438
```

```
rate_vector_34_54 = (datatable_34_54[, 'Dead'] / datatable_34_54[, 'Sum'])[c('No', 'Yes')]
barplot(rate_vector_34_54, main="Rate of death for women between 34 and 54 if yes or no they smoke")
```

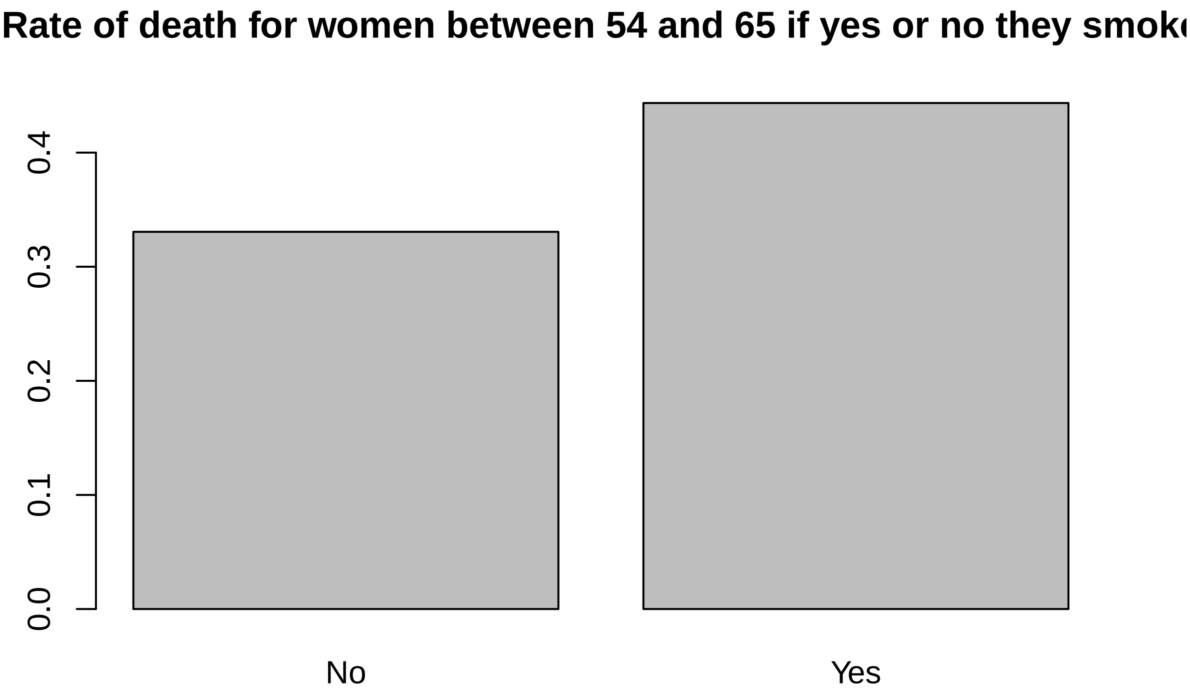


Women wo smoke die more in this category.

```
data_54_65 = data[data$Age >= 54 & data$Age < 65, ]
datatable_54_65 = addmargins(table(data_54_65$Smoker, data_54_65$Status))
datatable_54_65
```

```
##
##        Alive Dead Sum
##   No      81   40 121
##   Yes     64   51 115
##   Sum    145   91 236
```

```
rate_vector_54_65 = (datatable_54_65[, 'Dead'] / datatable_54_65[, 'Sum'])[c('No', 'Yes')]
barplot(rate_vector_54_65, main="Rate of death for women between 54 and 65 if yes or no they smoke")
```



Women wo smoke die more in this category.

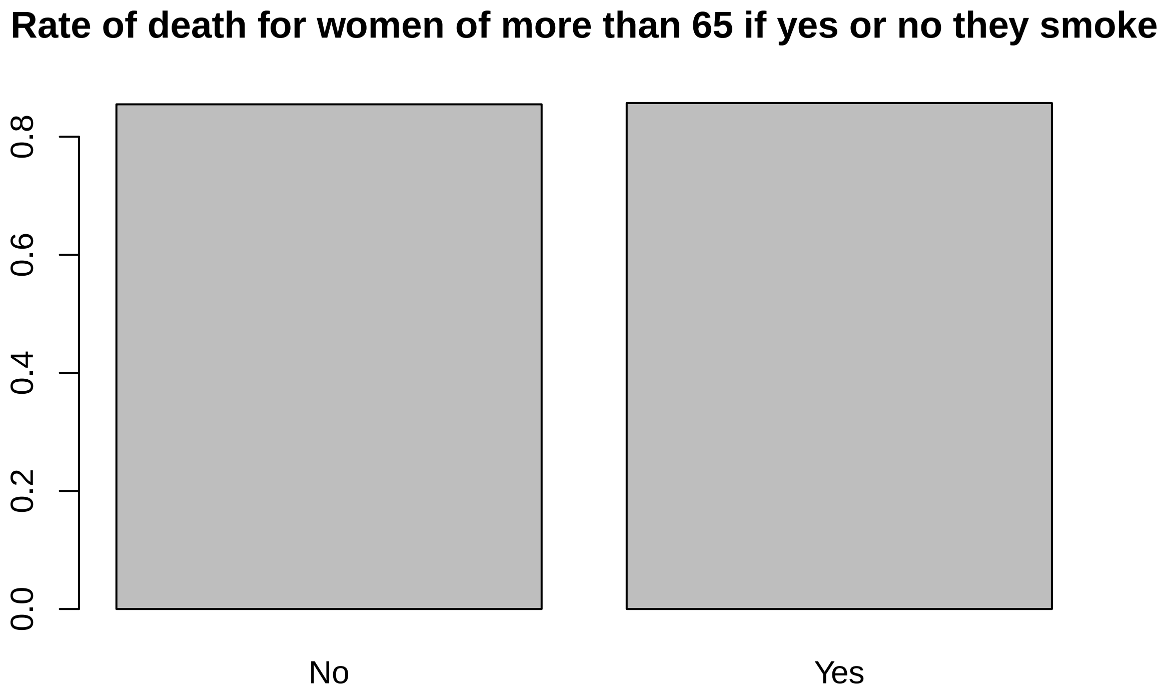
```r
data_65 = data[data$Age >= 65, ]
datatable_65 = addmargins(table(data_65$Smoker, data_65$Status))
datatable_65
```

```
##
##       Alive Dead Sum
##   No     28  165 193
##   Yes     7   42  49
##   Sum    35  207 242
```

Explanations

We can observe that for womens older than 65, there is much more deaths in the category that don't smoke. This make the data unbalance. This high number of death brings the total rate as observe previously to a high number.

```r
rate_vector_65 = (datatable_65[, 'Dead'] / datatable_65[, 'Sum'])[c('No', 'Yes')]
barplot(rate_vector_65, main="Rate of death for women of more than 65 if yes or no they smoke")
```

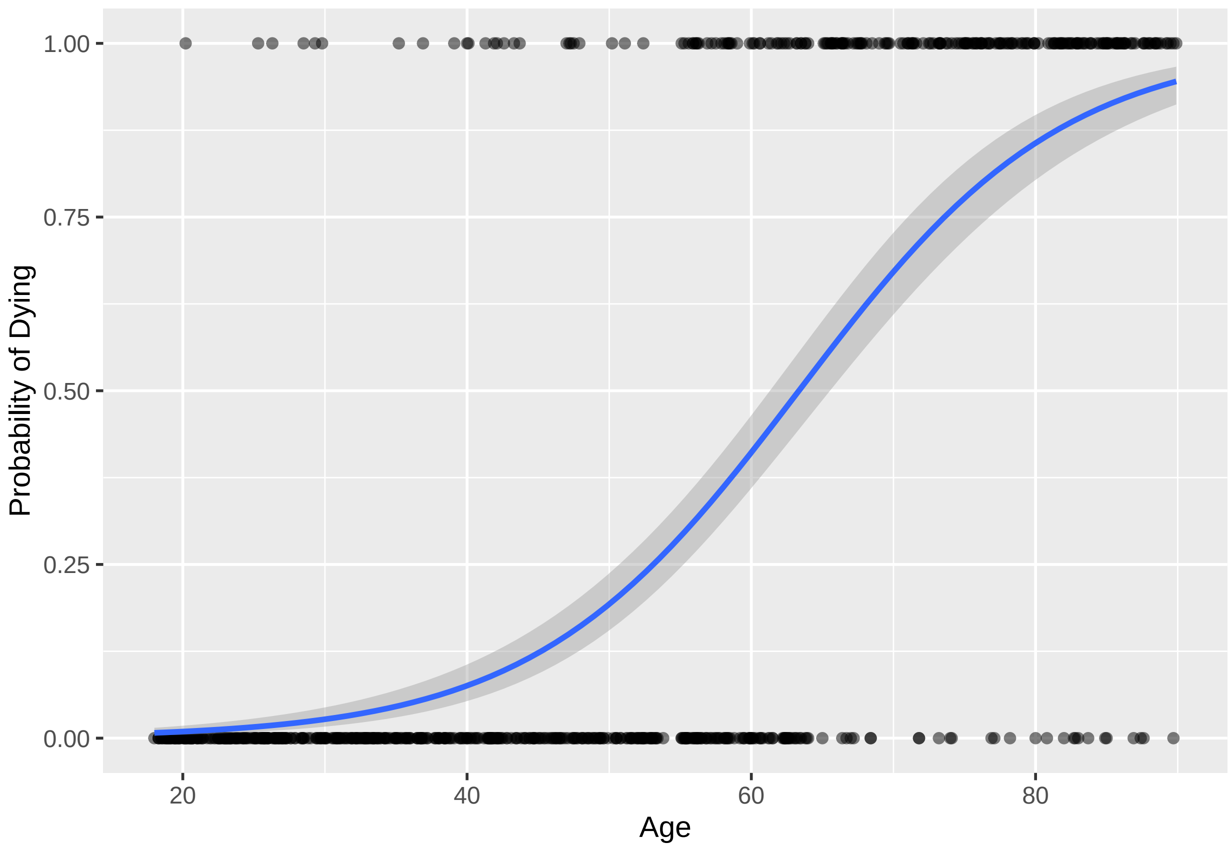


The relative rate in this category is similar thought.

3

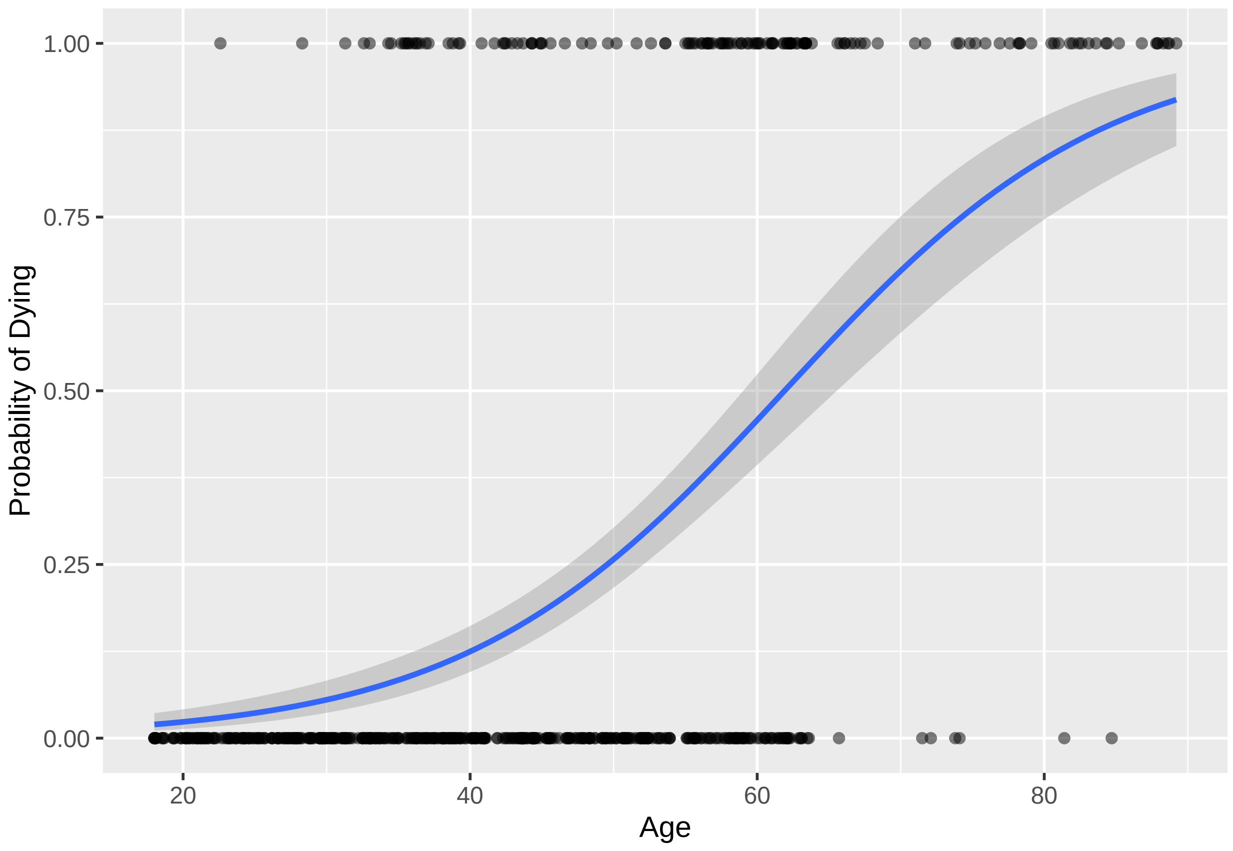
```
#change Alive and Dead in 0 and 1 in the Status column
data1 <- data %>%
      mutate(Status = ifelse(Status == "Alive",0,1))

data_no = data1[data1$Smoker == 'No',]
#plot logistic regression curve
ggplot(data_no, aes(x=Age, y=Status)) +
  geom_point(alpha=.5) +
  stat_smooth(method="glm", method.args = list(family=binomial))+
  labs(
    y = "Probability of Dying"
    )
```

```
## ‘geom_smooth()‘ using formula ’y ~ x’
```



```
data_yes = data1[data1$Smoker == 'Yes',]
#plot logistic regression curve
ggplot(data_yes, aes(x=Age, y=Status)) +
  geom_point(alpha=.5) +
  stat_smooth(method="glm", method.args = list(family=binomial)) +
  labs(
    y = "Probability of Dying"
    )
```

```
## `geom_smooth()` using formula 'y ~ x'
```



Explanation

We can observe that the probability of death increase more quickly for the womens that smoke