final_hi2021

Fenwick

2023-08-01

R Markdown

```
stroke_dt <- read.csv("stroke.csv")
colnames(stroke_dt)
```

```
 [1] "id"                "gender"            "age"
 [4] "hypertension"      "heart_disease"     "ever_married"
 [7] "work_type"         "Residence_type"    "avg_glucose_level"
[10] "bmi"               "smoking_status"    "stroke"
```

```
# PART 1: CLEANING, PROCESSING, and PREPARING THE DATA
# Exploratory Data Analysis
library(dplyr)
```

```

Attaching package: 'dplyr'

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

    filter, lag

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

    intersect, setdiff, setequal, union
```

```
stroke_dt %>% count(gender)
```

```
  gender    n
1 Female 2994
2   Male 2115
3  Other    1
```

```
stroke_dt %>% count(smoking_status)
```

```
   smoking_status    n
1         Unknown 1544
2 formerly smoked  885
3    never smoked 1892
4          smokes  789
```

```
stroke_dt %>% count(Residence_type)
```

```
  Residence_type    n
1          Rural 2514
2          Urban 2596
```

```
stroke_dt %>% count(work_type)
```

```
      work_type    n
1      Govt_job  657
2  Never_worked   22
3       Private 2925
4 Self-employed  819
5      children  687
```

```
# Handle Missing Data
library(dplyr)
stroke_dt1 <- stroke_dt %>%
  mutate_all(~replace(., . == 'N/A', NA))

# Imputing missing data
missing_dt <- apply(stroke_dt1, MARGIN = 2, function(x){sum(is.na(x))/length(x)*100})
missing_dt # shows the percentage of missing data in each column, bmi missing 3.9%
```

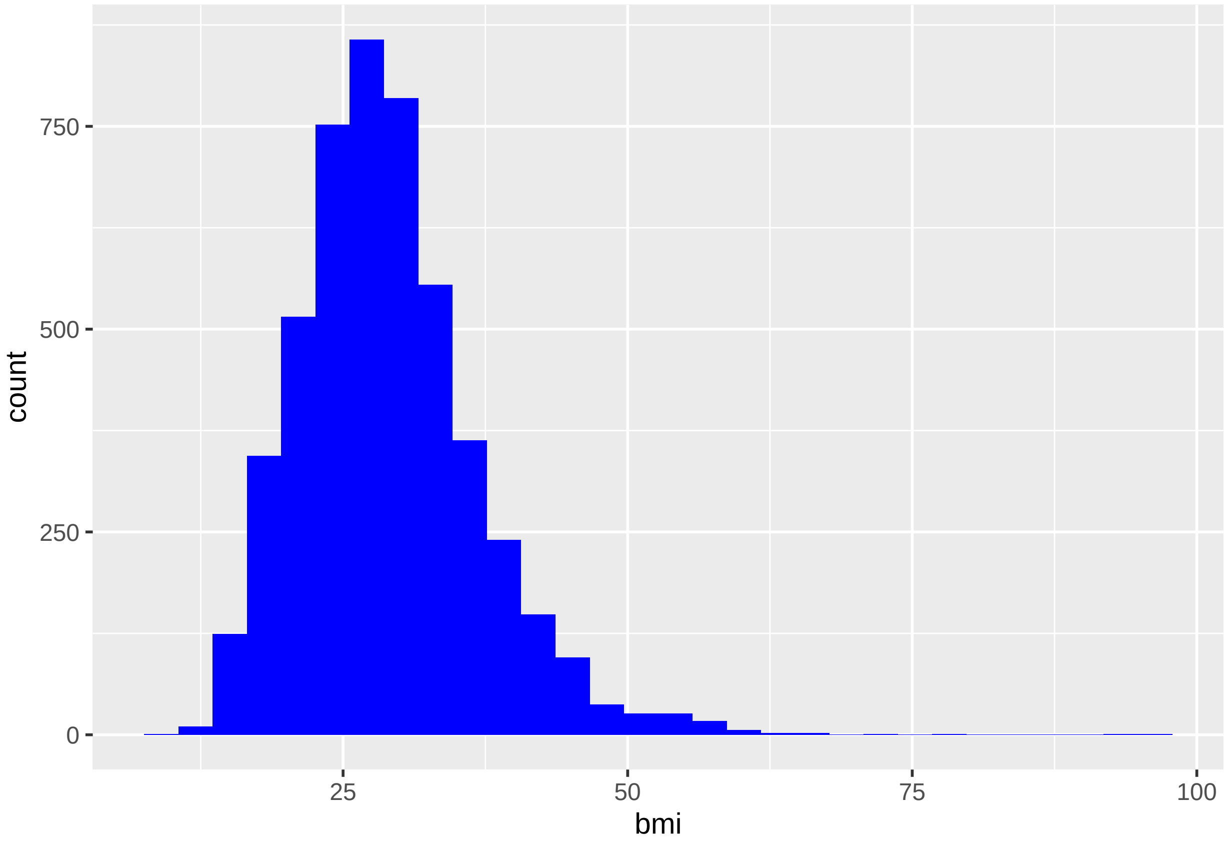
```
               id            gender               age      hypertension
         0.000000          0.000000          0.000000          0.000000
    heart_disease      ever_married         work_type    Residence_type
         0.000000          0.000000          0.000000          0.000000
avg_glucose_level               bmi    smoking_status            stroke
         0.000000          3.933464          0.000000          0.000000
```

```
new_stroke <- stroke_dt1[,missing_dt < 20] # blank before the comma because we want to keep all rows
apply(new_stroke, 2, function(x){sum(is.na(x))/length(x)*100})
```

```
               id            gender               age      hypertension
         0.000000          0.000000          0.000000          0.000000
    heart_disease      ever_married         work_type    Residence_type
         0.000000          0.000000          0.000000          0.000000
avg_glucose_level               bmi    smoking_status            stroke
         0.000000          3.933464          0.000000          0.000000
```

```
library(ggplot2) # determine if normal distribution before imputing data
stroke_dt1$bmi <- as.numeric(stroke_dt1$bmi) # convert BMI to numeric
ggplot(data=subset(stroke_dt1, !is.na(bmi)), aes(x=bmi)) +
  geom_histogram(fill = 'blue')
```

```
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



```r
# Imputing data with the mean
imputed_bmi <- data.frame(
  original = stroke_dt1$bmi,
  bmi_imputed = replace(stroke_dt1$bmi, is.na(stroke_dt1$bmi), mean(stroke_dt1$bmi, na.rm = TRUE))
)

# Combining the imputed data with the original dataframe, use cbind()
stroke_dt2 <- cbind(stroke_dt1, imputed_bmi)

stroke_dt3 <- stroke_dt2[ -c(10,13) ] # remove ID, ever_married, old BMI columns

prop.table(table(stroke_dt3$stroke)) # see ratio of stroke and non stroke patients
```

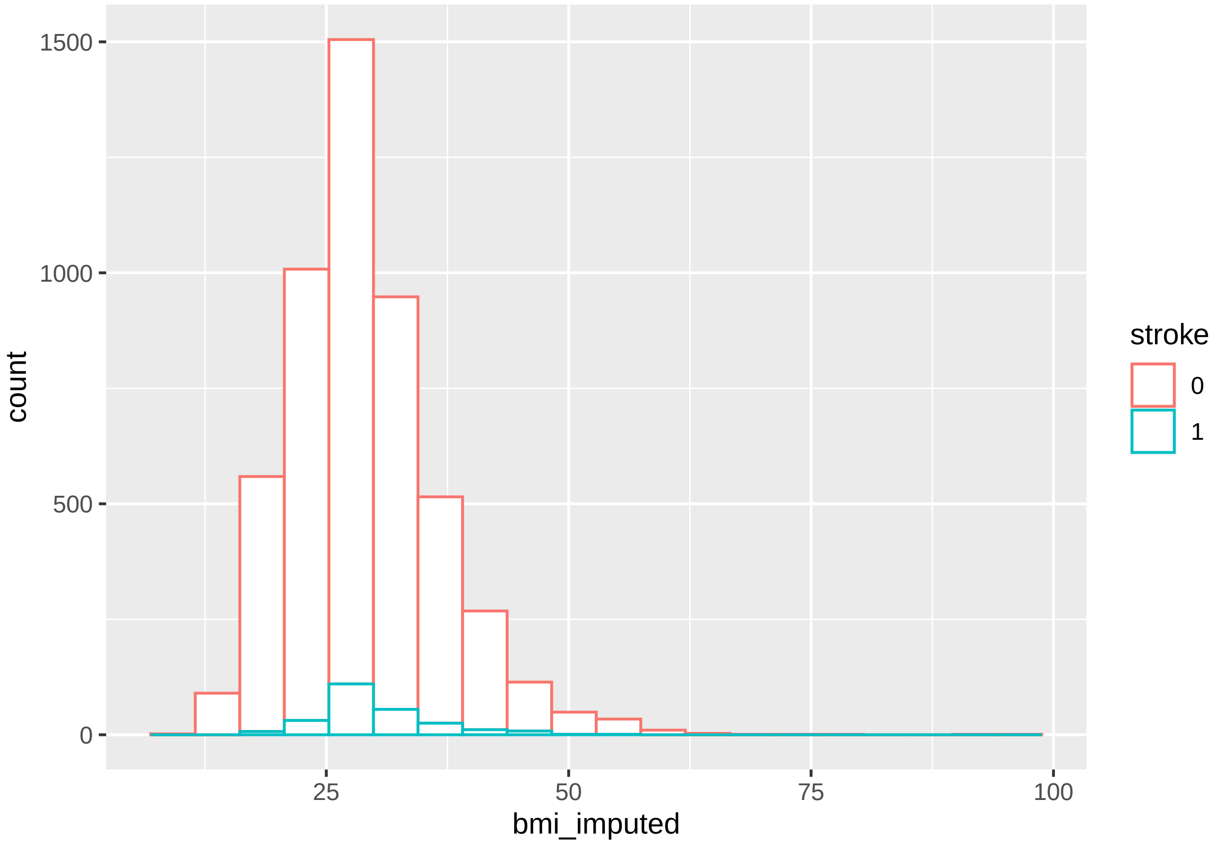
```
         0          1 
0.95127202 0.04872798 
```

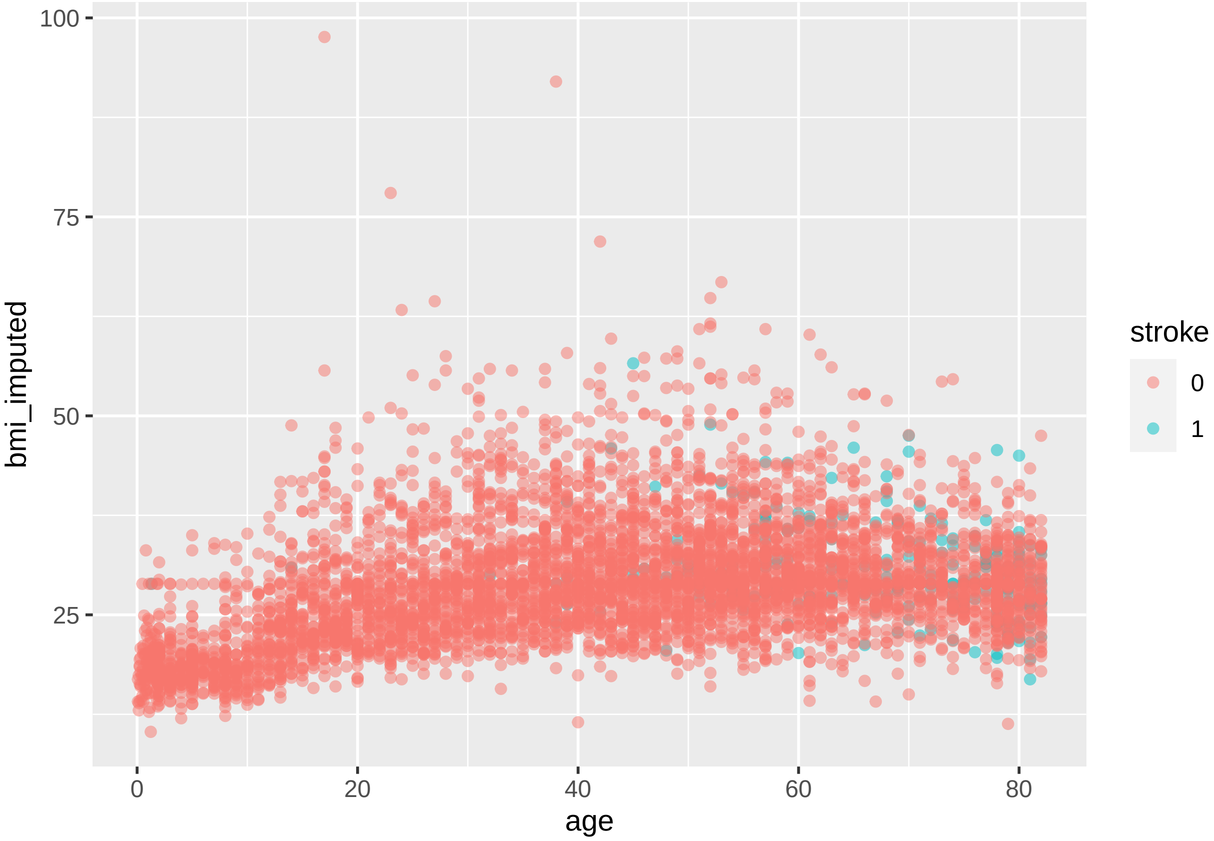
```r
stroke_pos <- subset(stroke_dt3, stroke == '1') # creating separate df for + stroke and - stroke
summary(stroke_pos)
```

```
       id           gender               age         hypertension   
 Min.   :  210   Length:249         Min.   : 1.32   Min.   :0.0000  
 1st Qu.:17013   Class :character   1st Qu.:59.00   1st Qu.:0.0000  
 Median :36706   Mode  :character   Median :71.00   Median :0.0000  
 Mean   :37115                      Mean   :67.73   Mean   :0.2651  
 3rd Qu.:56669                      3rd Qu.:78.00   3rd Qu.:1.0000  
 Max.   :72918                      Max.   :82.00   Max.   :1.0000  
 heart_disease    ever_married        work_type         Residence_type    
 Min.   :0.0000   Length:249         Length:249         Length:249        
 1st Qu.:0.0000   Class :character   Class :character   Class :character  
 Median :0.0000   Mode  :character   Mode  :character   Mode  :character  
 Mean   :0.1888                                                           
 3rd Qu.:0.0000                                                           
 Max.   :1.0000                                                           
 avg_glucose_level smoking_status         stroke   bmi_imputed   
 Min.   : 56.11    Length:249         Min.   :1   Min.   :16.90  
 1st Qu.: 79.79    Class :character   1st Qu.:1   1st Qu.:27.00  
 Median :105.22    Mode  :character   Median :1   Median :28.89  
 Mean   :132.54                       Mean   :1   Mean   :30.22  
 3rd Qu.:196.71                       3rd Qu.:1   3rd Qu.:32.50  
 Max.   :271.74                       Max.   :1   Max.   :56.60  
```

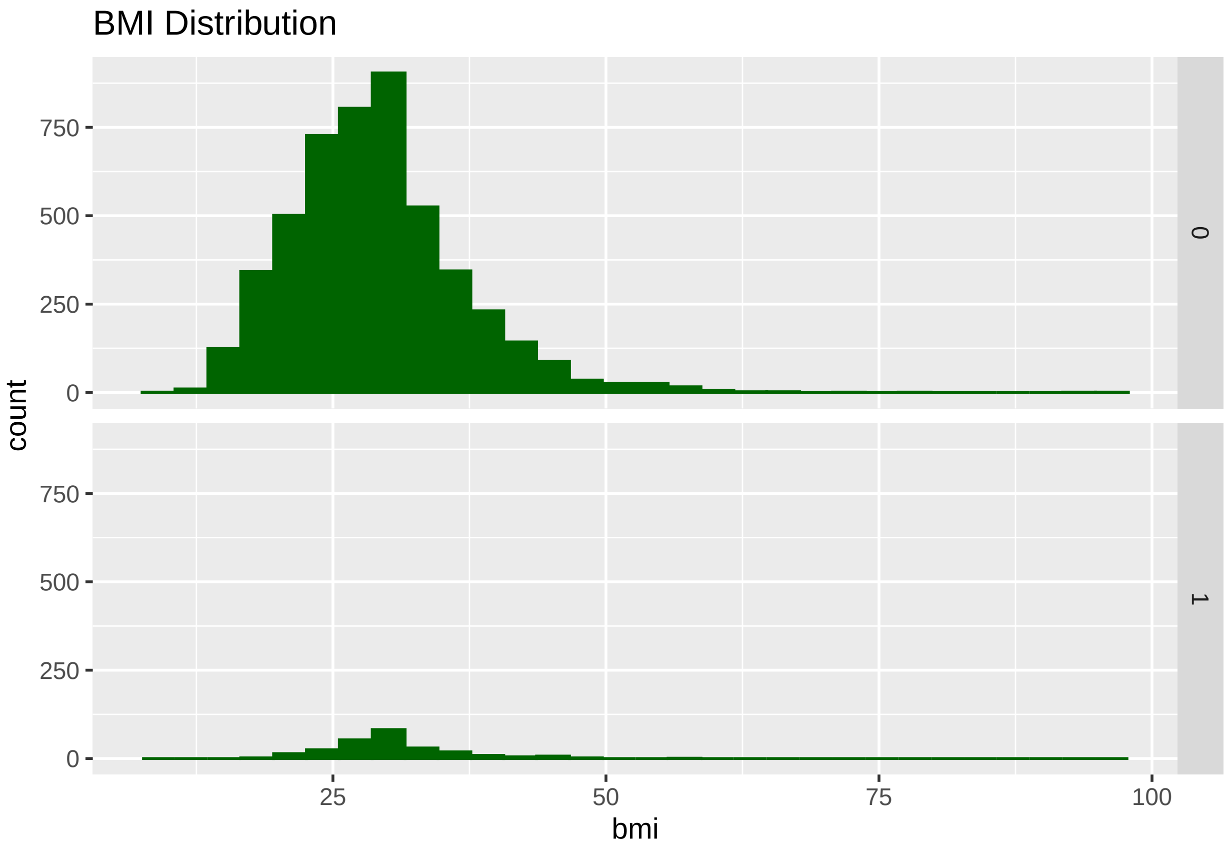
```r
stroke_neg <- subset(stroke_dt3, stroke == '0')
summary(stroke_neg)
```

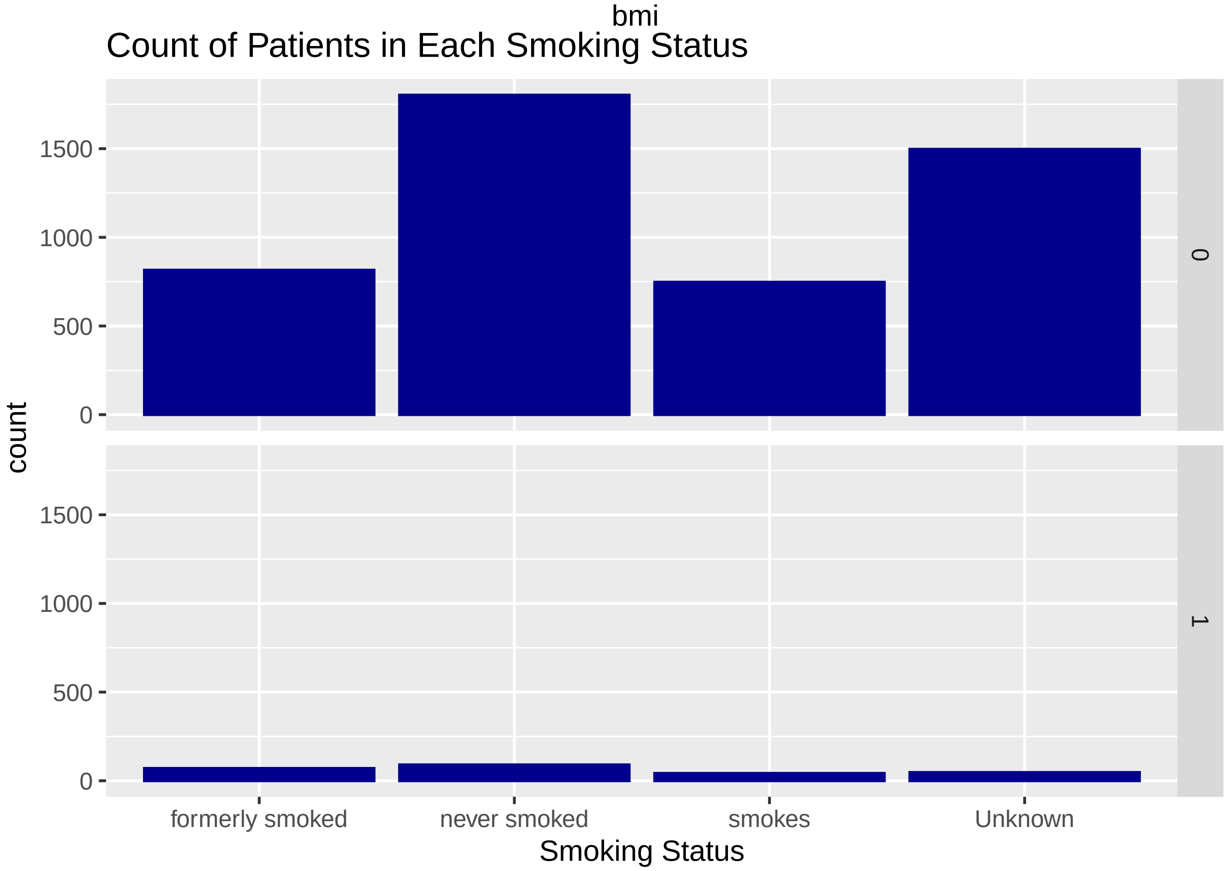
```
       id           gender               age         hypertension    
 Min.   :   67   Length:4861        Min.   : 0.08   Min.   :0.00000  
 1st Qu.:17762   Class :character   1st Qu.:24.00   1st Qu.:0.00000  
 Median :36958   Mode  :character   Median :43.00   Median :0.00000  
 Mean   :36487                      Mean   :41.97   Mean   :0.08887  
 3rd Qu.:54497                      3rd Qu.:59.00   3rd Qu.:0.00000  
 Max.   :72940                      Max.   :82.00   Max.   :1.00000  
 heart_disease     ever_married        work_type         Residence_type    
 Min.   :0.00000   Length:4861        Length:4861        Length:4861       
 1st Qu.:0.00000   Class :character   Class :character   Class :character  
 Median :0.00000   Mode  :character   Mode  :character   Mode  :character  
 Mean   :0.04711                                                           
 3rd Qu.:0.00000                                                           
 Max.   :1.00000                                                           
 avg_glucose_level smoking_status         stroke   bmi_imputed   
 Min.   : 55.12    Length:4861        Min.   :0   Min.   :10.30  
 1st Qu.: 77.12    Class :character   1st Qu.:0   1st Qu.:23.60  
 Median : 91.47    Mode  :character   Median :0   Median :28.30  
 Mean   :104.80                       Mean   :0   Mean   :28.83  
 3rd Qu.:112.83                       3rd Qu.:0   3rd Qu.:32.80  
 Max.   :267.76                       Max.   :0   Max.   :97.60  
```

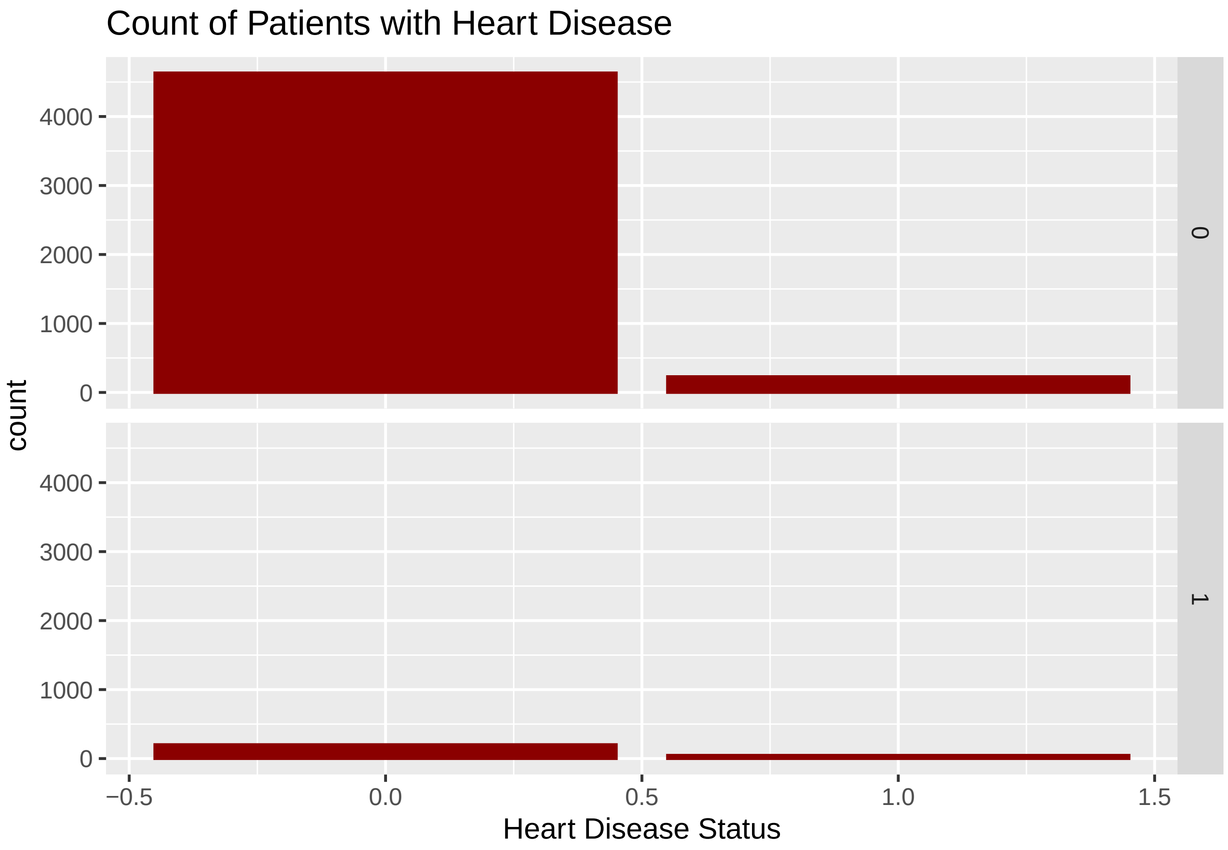




```
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```


BMI Distribution
count
750
500
250
0
0
1
25
50
75
100
bmi


Count of Patients in Each Smoking Status
count
1500
1000
500
0
0
1
formerly smoked
never smoked
smokes
Unknown
Smoking Status



```
    Welch Two Sample t-test

data:  stroke_dt3$bmi_imputed by stroke_dt3$stroke
t = -3.6104, df = 295.21, p-value = 0.0003591
alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
95 percent confidence interval:
 -2.1513954 -0.6334067
sample estimates:
mean in group 0 mean in group 1 
       28.82539        30.21779 


    Welch Two Sample t-test

data:  stroke_dt3$avg_glucose_level by stroke_dt3$stroke
t = -6.9824, df = 260.89, p-value = 2.401e-11
alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
95 percent confidence interval:
 -35.57474 -19.92371
sample estimates:
mean in group 0 mean in group 1 
       104.7955        132.5447 


    Welch Two Sample t-test

data:  stroke_dt3$age by stroke_dt3$stroke
```

```
t = -29.686, df = 331.65, p-value < 2.2e-16
alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
95 percent confidence interval:
 -27.4634 -24.0499
sample estimates:
mean in group 0 mean in group 1
       41.97154        67.72819
```

```
               Df Sum Sq Mean Sq F value  Pr(>F)
bmi_imputed     1   0.36   0.359   7.878 0.00502 **
hypertension    1   3.60   3.599  78.905 < 2e-16 ***
Residuals    5107 232.91   0.046
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
                           Df Sum Sq Mean Sq F value   Pr(>F)
bmi_imputed                 1   0.36   0.359    7.90  0.00496 **
hypertension                1   3.60   3.599   79.13  < 2e-16 ***
bmi_imputed:hypertension    1   0.70   0.697   15.33 9.15e-05 ***
Residuals                5106 232.21   0.045
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```