

Stroke dataset

Florencia Luque and Seyed Amirhossein Mosaddad

October 6, 2024

```
##
## Adjuntando el paquete: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

Introduction

This dataset is a data obtain from *kaggle* and is used to predict if a patient will probably get a stroke based on characteristic of them like gender, age, bmi, glucose levels.

The stroke variable have a 4.8% of people have had one. We want to check the distributions of the variables and possibles explanations of which variable can make an impact to get a stroke before creating a model to proved or been proved wrong about it.

The data have the follow variables:

- 1) id: unique identifier
- 2) gender: "Male", "Female" or "Other"
- 3) age: age of the patient
- 4) hypertension: 0 if the patient doesn't have hypertension, 1 if the patient has hypertension
- 5) heart_disease: 0 if the patient doesn't have any heart diseases, 1 if the patient has a heart disease
- 6) ever_married: "No" or "Yes"
- 7) work_type: "children", "Govt_jov", "Never_worked", "Private" or "Self-employed"
- 8) Residence_type: "Rural" or "Urban"
- 9) avg_glucose_level: average glucose level in blood
- 10) bmi: body mass index
- 11) smoking_status: "formerly smoked", "never smoked", "smokes" or "Unknown"*
- 12) stroke: 1 if the patient had a stroke or 0 if not

The next is a summary of the data.

```
summary(data)
```

```
##          id          gender          age          hypertension heart_disease
## Min.      : 67   Female:2994   Min.      : 0.08   0:4612        0:4834
## 1st Qu.:17741   Male  :2115   1st Qu.:25.00   1: 498        1: 276
## Median :36932   Other :    1   Median :45.00
## Mean      :36518                               Mean      :43.23
## 3rd Qu.:54682                               3rd Qu.:61.00
## Max.      :72940                               Max.      :82.00
##
## ever_married      work_type      Residence_type avg_glucose_level
## No :1757      children      : 687   Rural:2514   Min.      : 55.12
## Yes:3353      Govt_job      : 657   Urban:2596   1st Qu.: 77.25
##                               Never_worked : 22   Median : 91.89
##                               Private       :2925   Mean    :106.15
##                               Self-employed: 819   3rd Qu.:114.09
##                               Max.      :271.74
##
##          bmi          smoking_status stroke
## Min.      :10.30   formerly smoked: 885   0:4861
## 1st Qu.:23.50   never smoked    :1892   1: 249
## Median :28.10   smokes         : 789
## Mean      :28.89   Unknown        :1544
## 3rd Qu.:33.10
## Max.      :97.60
## NA's      :201
```

Dataset

We will start the analysis with the categorical variables.

Categorical Variables

Gender

This variable has 2 categories Male and Female (there's one person who is Other but it's only one so we can't make any assumptions about this data).

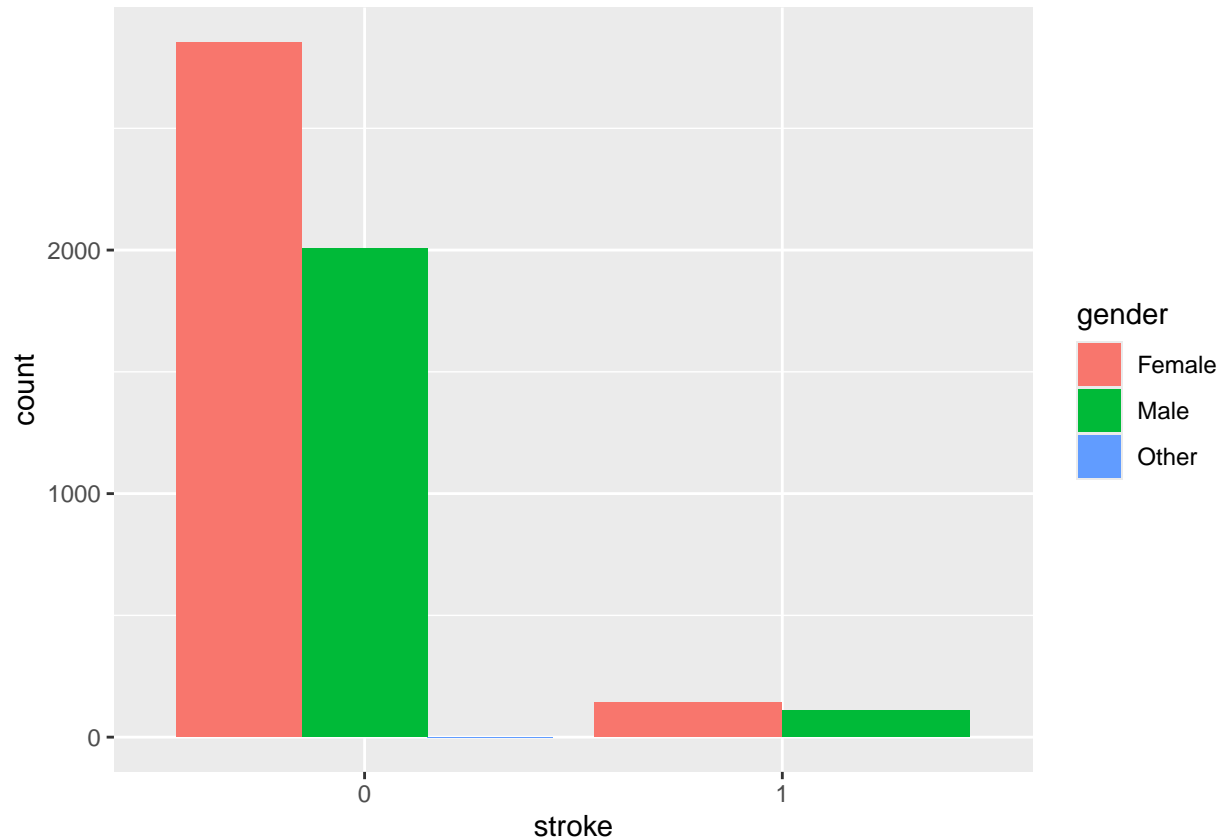
The next table shows a summary of the quantity of people getting a stroke by gender and the corresponding percentage.

```
stroke_gender = data %>% group_by(stroke,gender) %>% summarise(n = n(),.groups = "drop") %>% group_by(gender)
pander(stroke_gender)
```

stroke	gender	n	percent
0	Female	2853	95.29
0	Male	2007	94.89
0	Other	1	100
1	Female	141	4.709

stroke	gender	n	percent
1	Male	108	5.106

```
ggplot(data)+aes(x=stroke,fill=gender) + geom_bar(position=position_dodge())
```



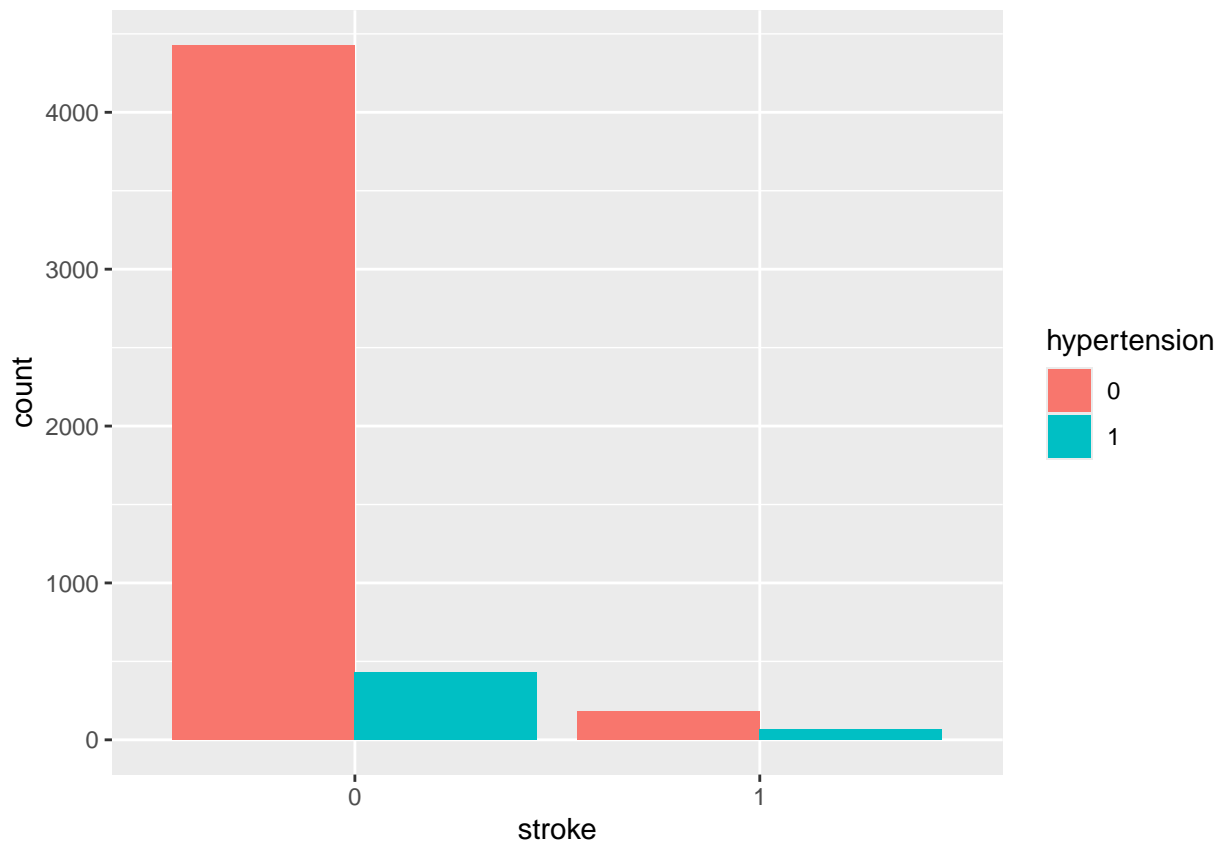
As you can see there's a higher % of male that have had a stroke in the data. This number is a little bit higher than the population so we don't think this would have a great impact in the future model.

Hypertension

```
pander(data %>% group_by(stroke,hypertension) %>% summarise(n = n(),.groups = "drop")%>%
  group_by(stroke) %>% mutate(percent = n/sum(n)*100))
```

stroke	hypertension	n	percent
0	0	4429	91.11
0	1	432	8.887
1	0	183	73.49
1	1	66	26.51

```
ggplot(data)+aes(x=stroke,fill=hypertension) + geom_bar(position=position_dodge())
```

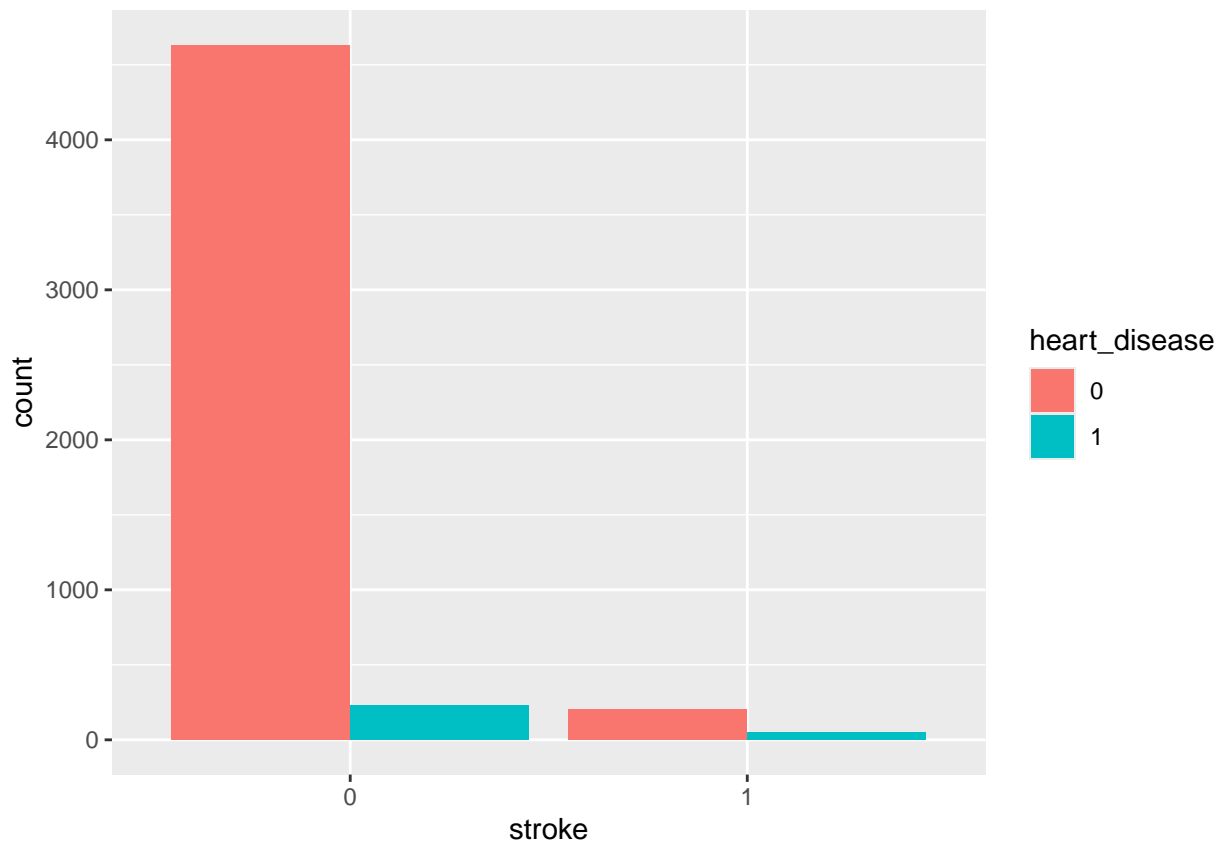


Heart disease

```
pander(data %>% group_by(stroke,heart_disease) %>% summarise(n = n(),.groups = "drop")%>%
  group_by(heart_disease) %>% mutate(percent = n/sum(n)*100))
```

stroke	heart_disease	n	percent
0	0	4632	95.82
0	1	229	82.97
1	0	202	4.179
1	1	47	17.03

```
ggplot(data)+aes(x=stroke,fill=heart_disease) + geom_bar(position=position_dodge())
```

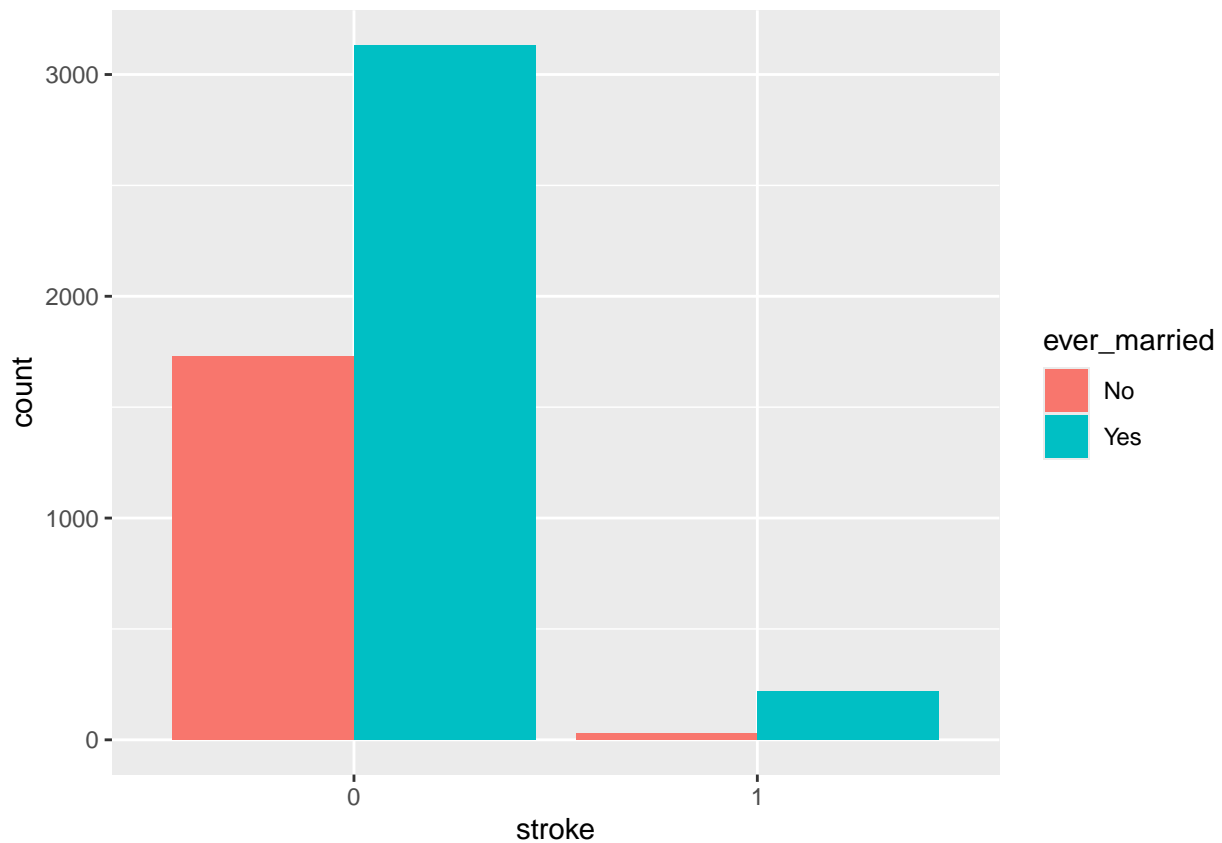


Ever Married

```
pander(data %>% group_by(stroke,ever_married) %>% summarise(n = n(),.groups = "drop")%>%
  group_by(ever_married) %>% mutate(percent = n/sum(n)*100))
```

stroke	ever_married	n	percent
0	No	1728	98.35
0	Yes	3133	93.44
1	No	29	1.651
1	Yes	220	6.561

```
ggplot(data)+aes(x=stroke,fill=ever_married) + geom_bar(position=position_dodge())
```

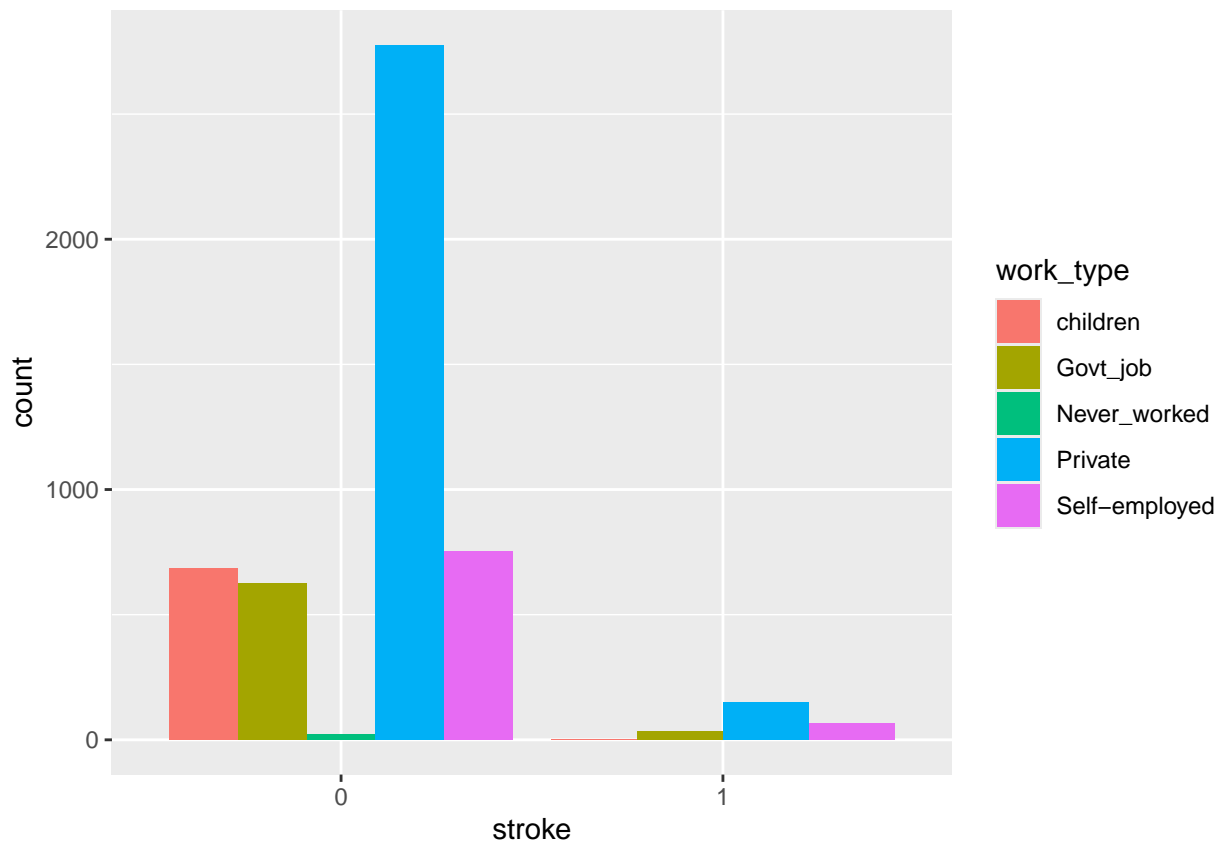


Work Type

```
pander(data %>% group_by(stroke,work_type) %>% summarise(n = n(),.groups = "drop")%>%
  group_by(work_type) %>% mutate(percent = n/sum(n)*100))
```

stroke	work_type	n	percent
0	children	685	99.71
0	Govt_job	624	94.98
0	Never_worked	22	100
0	Private	2776	94.91
0	Self-employed	754	92.06
1	children	2	0.2911
1	Govt_job	33	5.023
1	Private	149	5.094
1	Self-employed	65	7.937

```
ggplot(data)+aes(x=stroke,fill=work_type) + geom_bar(position=position_dodge())
```

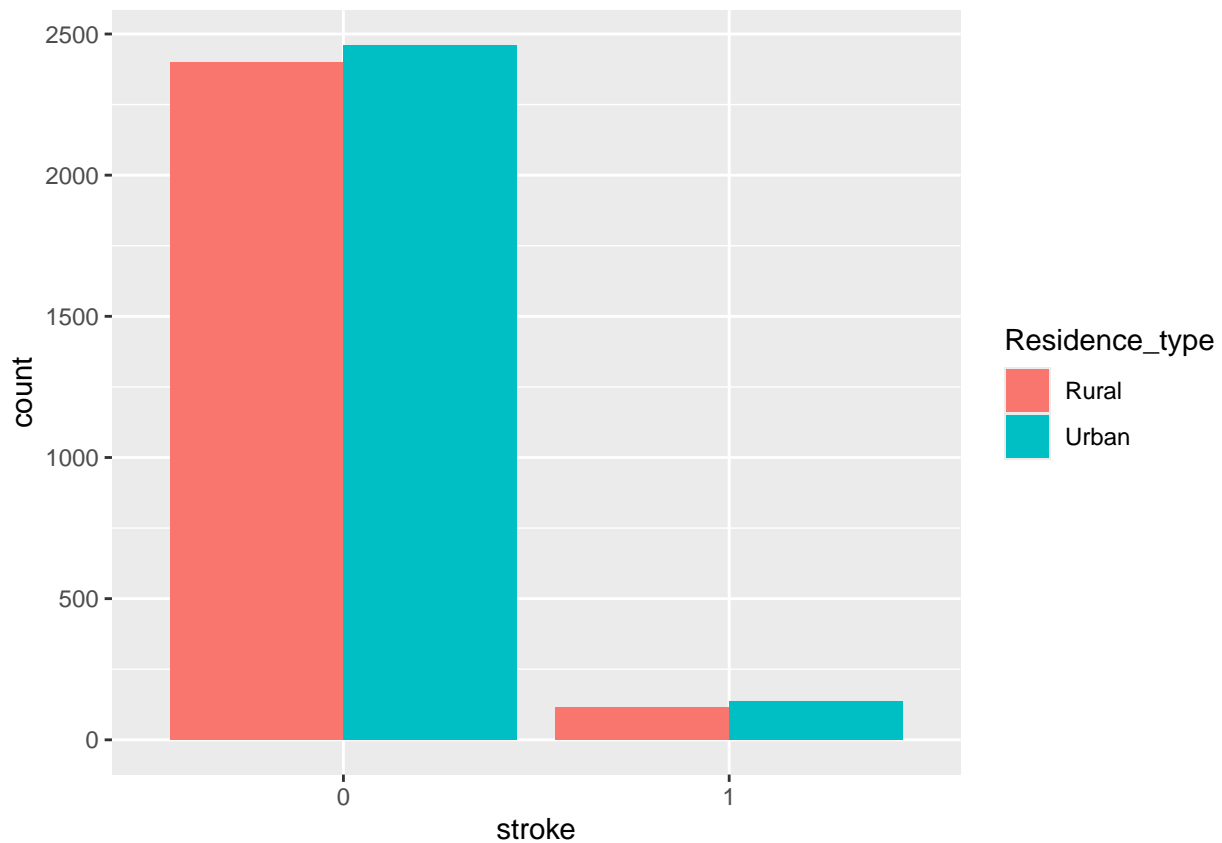


Resident Type

```
pander(data %>% group_by(stroke,Residence_type) %>% summarise(n = n(),.groups = "drop")%>%
  group_by(Residence_type) %>% mutate(percent = n/sum(n)*100))
```

stroke	Residence_type	n	percent
0	Rural	2400	95.47
0	Urban	2461	94.8
1	Rural	114	4.535
1	Urban	135	5.2

```
ggplot(data)+aes(x=stroke,fill=Residence_type) + geom_bar(position=position_dodge())
```



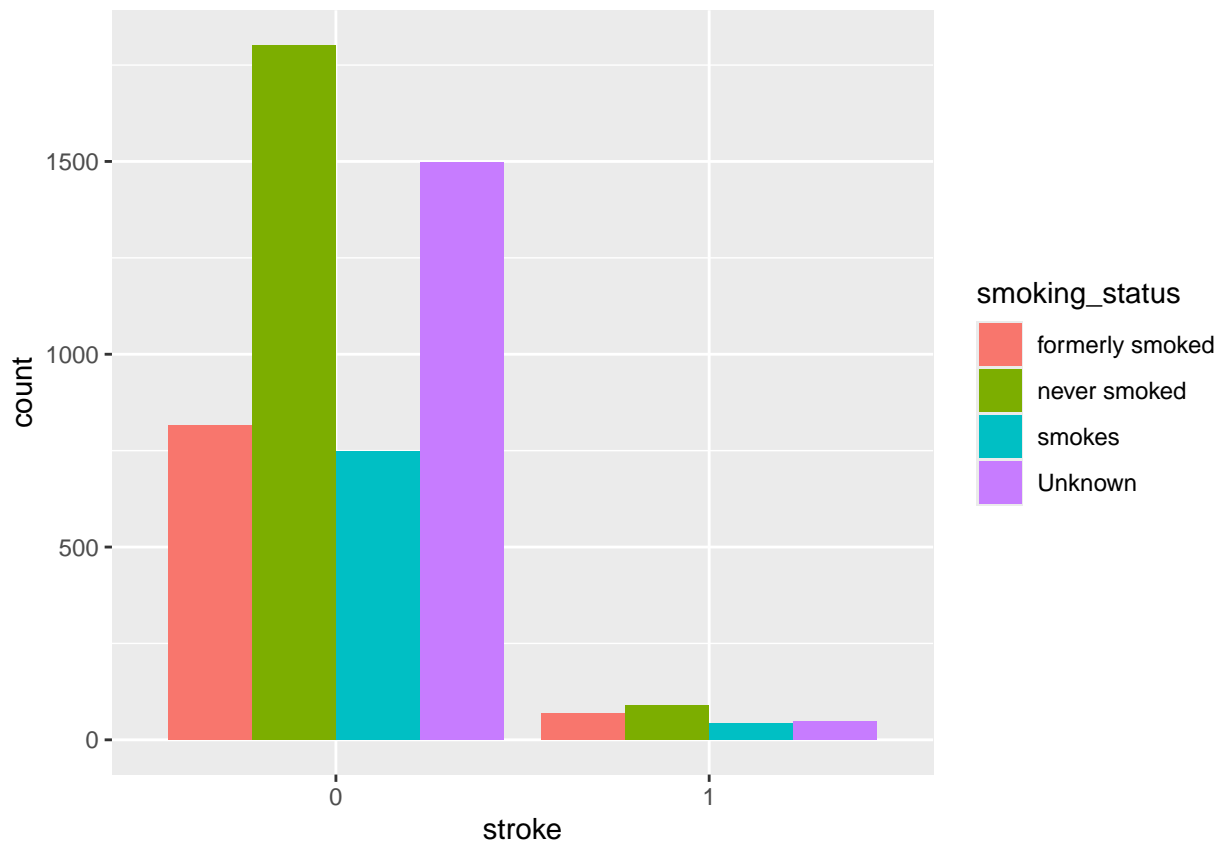
This could mean something

Smoking Status

```
pander(data %>% group_by(stroke,smoking_status) %>% summarise(n = n(),.groups = "drop")%>%
  group_by(smoking_status) %>% mutate(percent = n/sum(n)*100))
```

stroke	smoking_status	n	percent
0	formerly smoked	815	92.09
0	never smoked	1802	95.24
0	smokes	747	94.68
0	Unknown	1497	96.96
1	formerly smoked	70	7.91
1	never smoked	90	4.757
1	smokes	42	5.323
1	Unknown	47	3.044

```
ggplot(data)+aes(x=stroke,fill=smoking_status) + geom_bar(position=position_dodge())
```

It's

very clear that this variable should be important in the model ## Numeric Variables

BMI

The first variable to evaluate will be the *bmi*. This variable is a metric that represent the relation between height and weight of a person. As you can see we have 3.9334638 % percentage of NA. However this is less than 5% so to treat this variable we will deleted all the rows with the NA in *bmi*

```
summary(data$bmi)
```

```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.     NA's
##  10.30   23.50   28.10   28.89   33.10   97.60      201
```

Next, we will check the frequency of the data using the cut between (0-18.5] as underweight, (18.5-24.9] as normal, (24.9-29.9] as overweight, (29.9-34.9] as obese and over this as extremely obese. This value leaves over 50% of our population in normal and overweight and almost 40% in the largest side in obese and extremely obese.

```
data = na.omit(data)
freq_table <- cut(data$bmi, breaks = c(0, 18.5, 24.9, 29.9, 34.9, Inf),
                  labels = c("Underweight", "Normal", "Overweight", "Obese", "Extremely Obese"))
data$cat_weight = freq_table

# Create frequency table
bmi_freq_table <- table(freq_table)
print(bmi_freq_table)
```

```
## freq_table
##      Underweight      Normal      Overweight      Obese Extremely Obese
##           349           1231           1409           1000           920
```

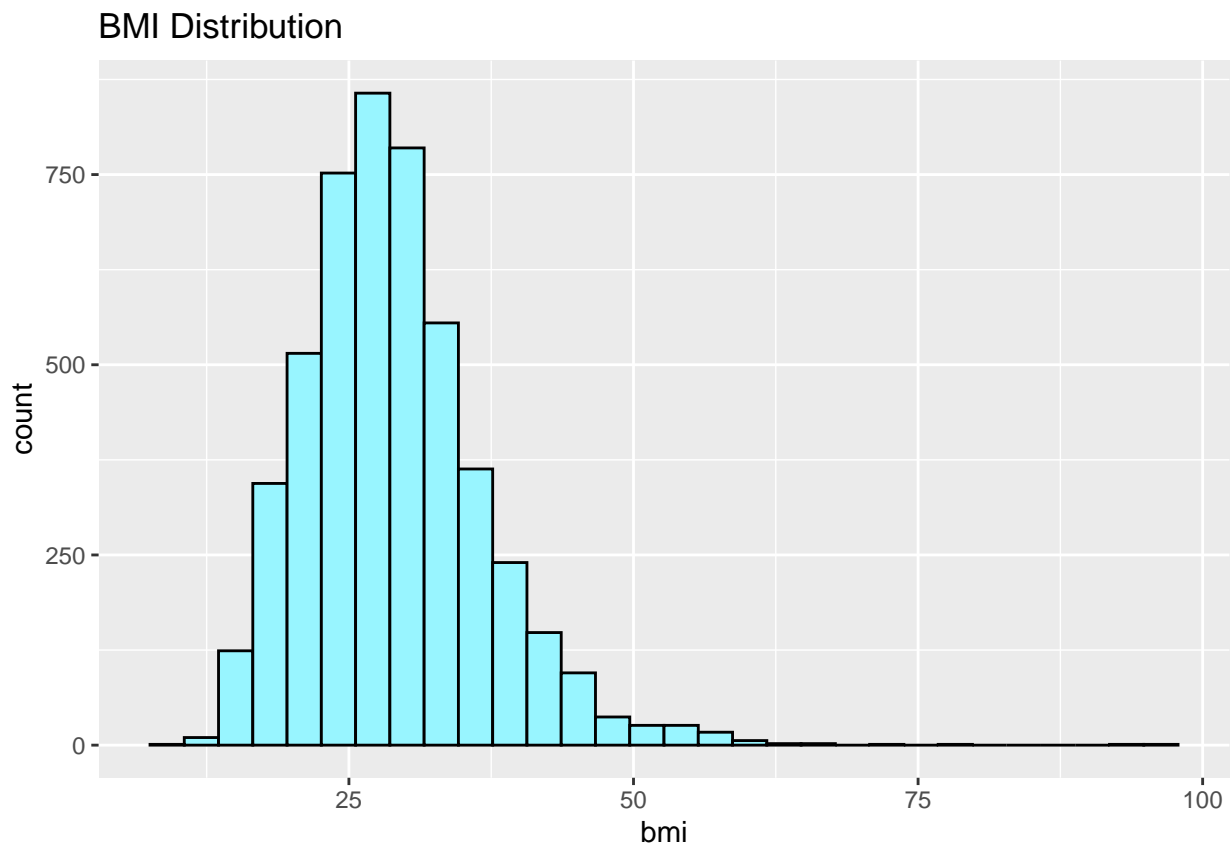
```
bmi_rel_freq <- prop.table(bmi_freq_table)*100
print(bmi_rel_freq)
```

```
## freq_table
##      Underweight      Normal      Overweight      Obese Extremely Obese
##      7.109391      25.076390      28.702383      20.370748      18.741088
```

If we check this with a graph we can see that the graph looks a bit symmetrical with an inclination to the right. This could mean that the distribution is not a normal like could it seem. We are going to check the kurtosis and skewness shape to check if there's a problem.

```
ggplot(data, aes(x=bmi)) +
  geom_histogram(color="black", fill="cadetblue1")+
  labs(title = "BMI Distribution")
```

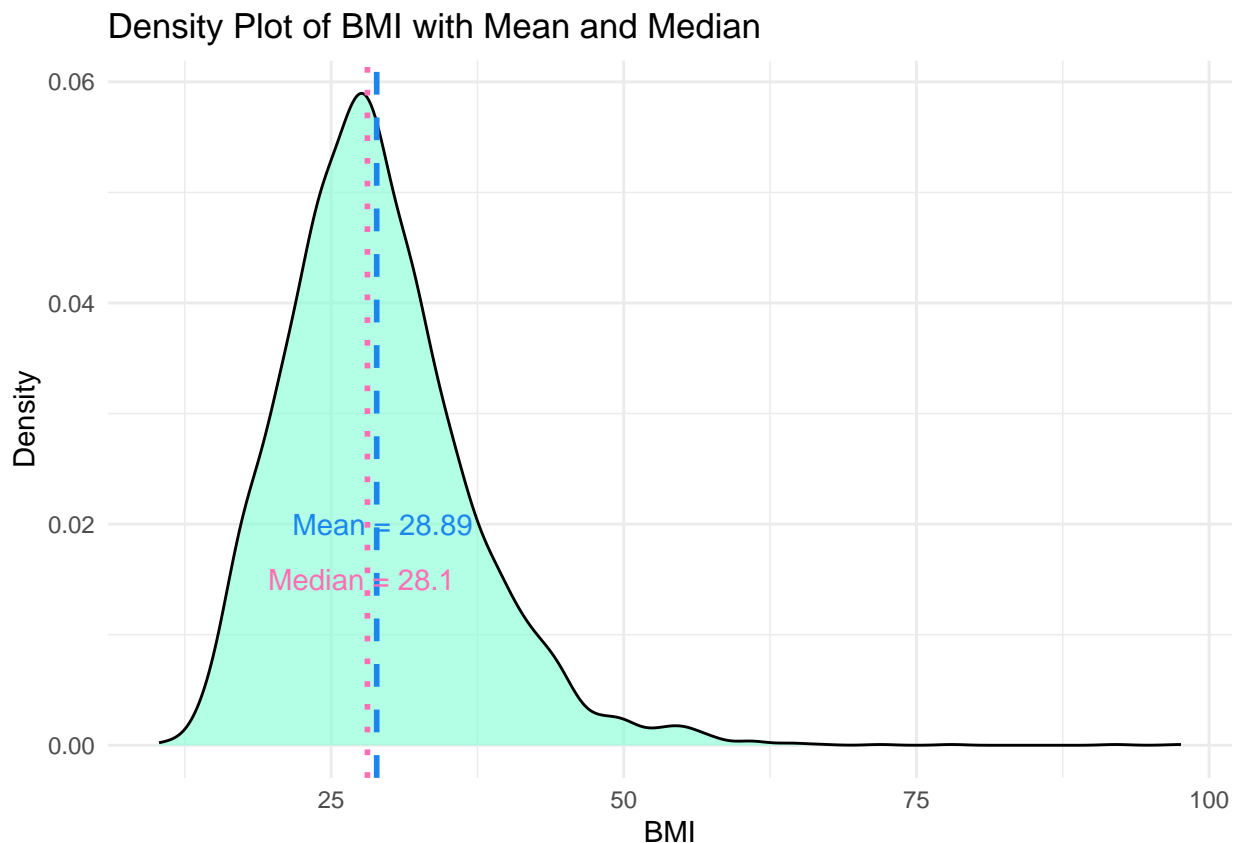
```
## 'stat_bin()' using 'bins = 30'. Pick better value with 'binwidth'.
```



```
bmi_mean <- mean(data$bmi)
bmi_median <- median(data$bmi)

# Density plot with a vertical line at the mean
```

```
ggplot(data, aes(x = data$bmi)) +
  geom_density(fill = "aquamarine1", alpha = 0.6) +
  geom_vline(aes(xintercept = bmi_mean), color = "dodgerblue2", linetype = "dashed", size = 1) +
  geom_vline(aes(xintercept = bmi_median), color = "hotpink1", linetype = "dotted", size = 1) +
  labs(title = "Density Plot of BMI with Mean and Median", x = "BMI", y = "Density") +
  annotate("text", x = bmi_mean + 0.5, y = 0.02, label = paste("Mean =", round(bmi_mean, 2)), color = "dodgerblue2") +
  annotate("text", x = bmi_median - 0.5, y = 0.015, label = paste("Median =", round(bmi_median, 2)), color = "hotpink1") +
  theme_minimal()
```



```
skewness(data$bmi)
```

```
## [1] 1.054695
```

```
kurtosis(data$bmi)
```

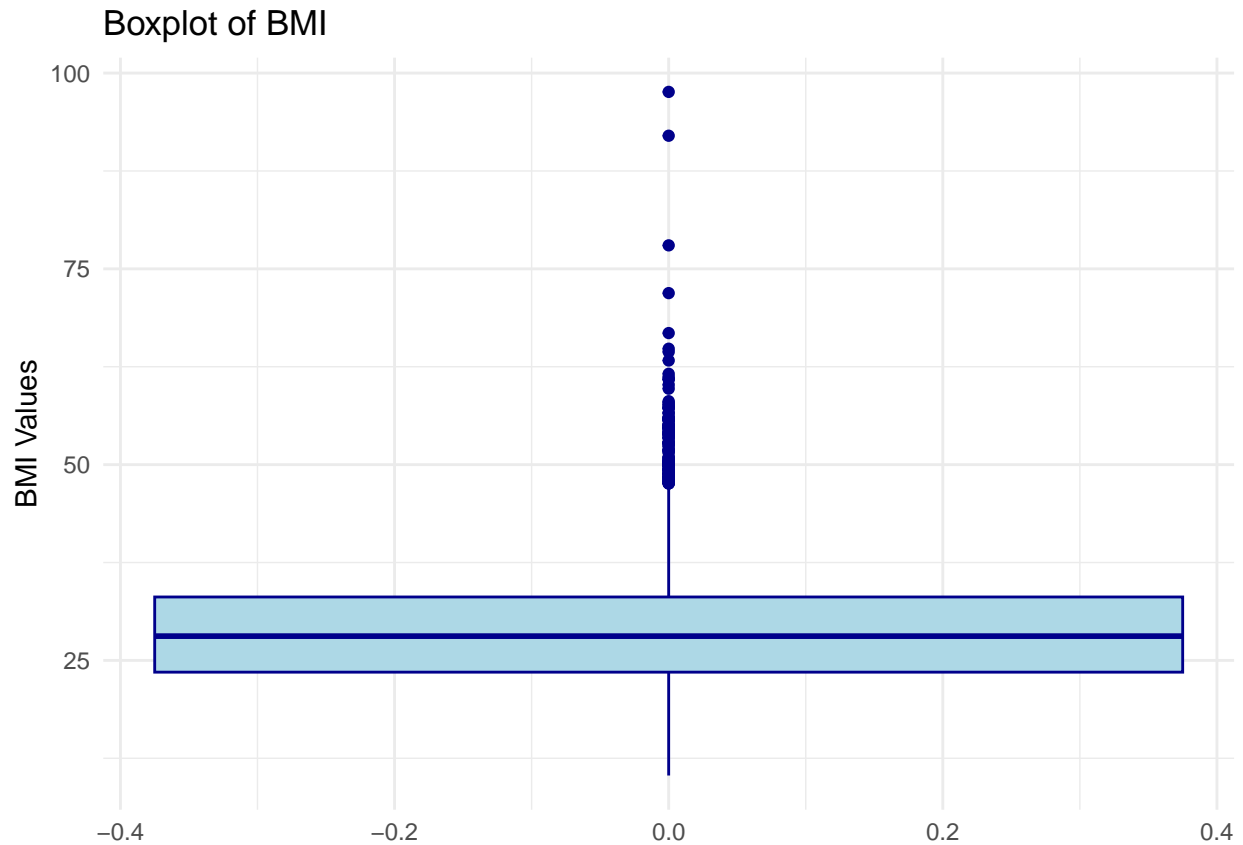
```
## [1] 3.355423
```

The values of the shape tell us that this is not a normal distribution. There's a lot of people in the center of the data but the quantity of people with an extremely high BMI that change the weight of the tail in the distribution.

Looking only at the graph we expected that the Kurtosis was close to 3 but a 3.35 shows that this data have a heavier tail than a normal distribution. This means that we have outliers in the data. In this case there's a lot of values over 30. If we combined the Skewness value of 1.05 and the Kurtosis this tells us that the distribution is skewed to the right and that the heavier tails are on this side explaining the outliers.

To see if the outliers are correctly in the heavier side we can check the boxplot of the *bmi*

```
ggplot(data, aes(y = data$bmi)) +
  geom_boxplot(fill = "lightblue", color = "darkblue") + # Change box and border color
  labs(title = "Boxplot of BMI", y = "BMI Values") + # Custom title and labels
  theme_minimal()
```



As you can see with the boxplot we can make sure that the *bmi* data have outliers within the largest values. The problem with this data is that we can not be sure if this is a mistake in the part of measure or maybe exist people with those values.

```
data_sub_bmi_gender = data %>% group_by(gender, stroke) %>% summarise(mean = mean(bmi))
```

BMI with gender

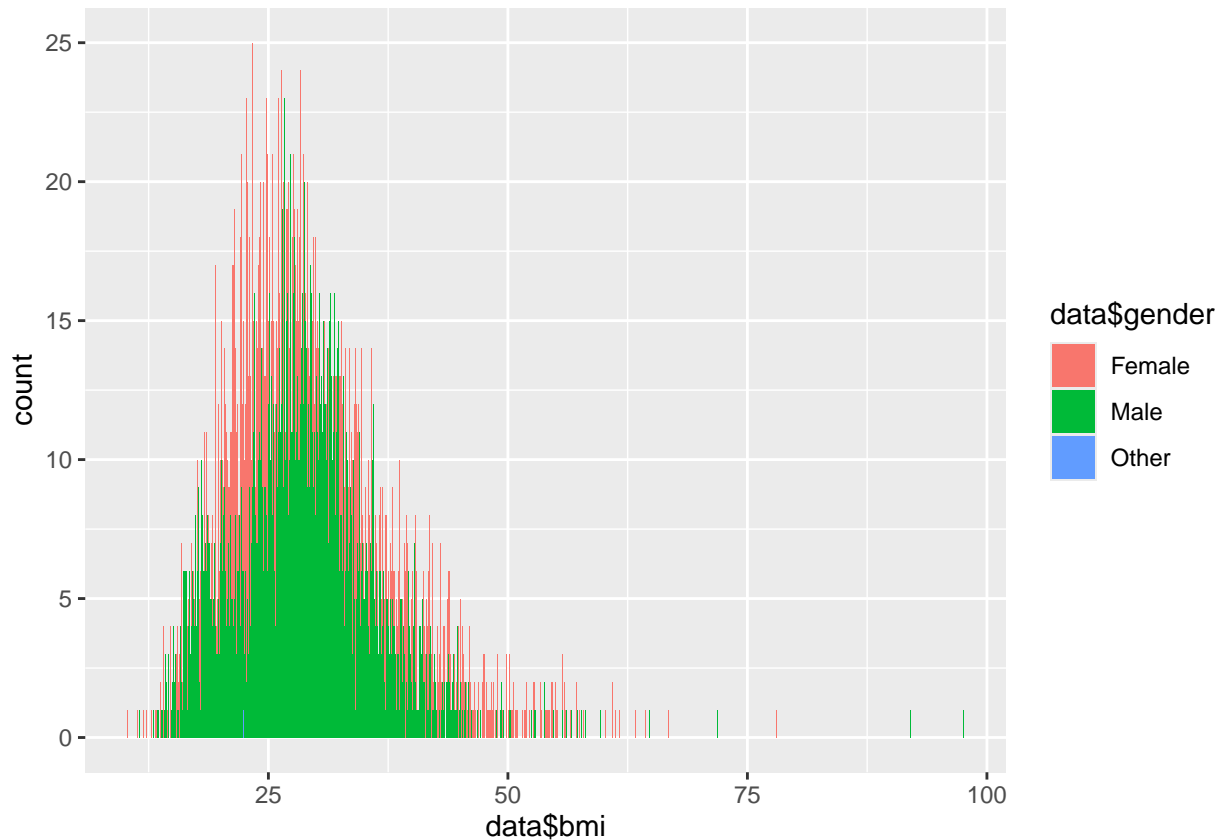
'summarise()' has grouped output by 'gender'. You can override using the
'.groups' argument.

```
print(data_sub_bmi_gender)
```

```
## # A tibble: 5 x 3
## # Groups:   gender [3]
##   gender stroke mean
##   <fct>   <fct> <dbl>
## 1 Female 0      29.0
```

```
## 2 Female 1      30.2
## 3 Male   0      28.5
## 4 Male   1      30.8
## 5 Other  0      22.4
```

```
ggplot(data, aes(x = data$bmi, fill = data$gender)) +
  geom_bar(position = position_dodge())
```



```
#geom_histogram(position = "identity", bins = 30)
```

```
data_sub_bmi_hyper= data %>% group_by(hypertension,stroke) %>% summarise(mean = mean(bmi))
```

BMI with hypertension

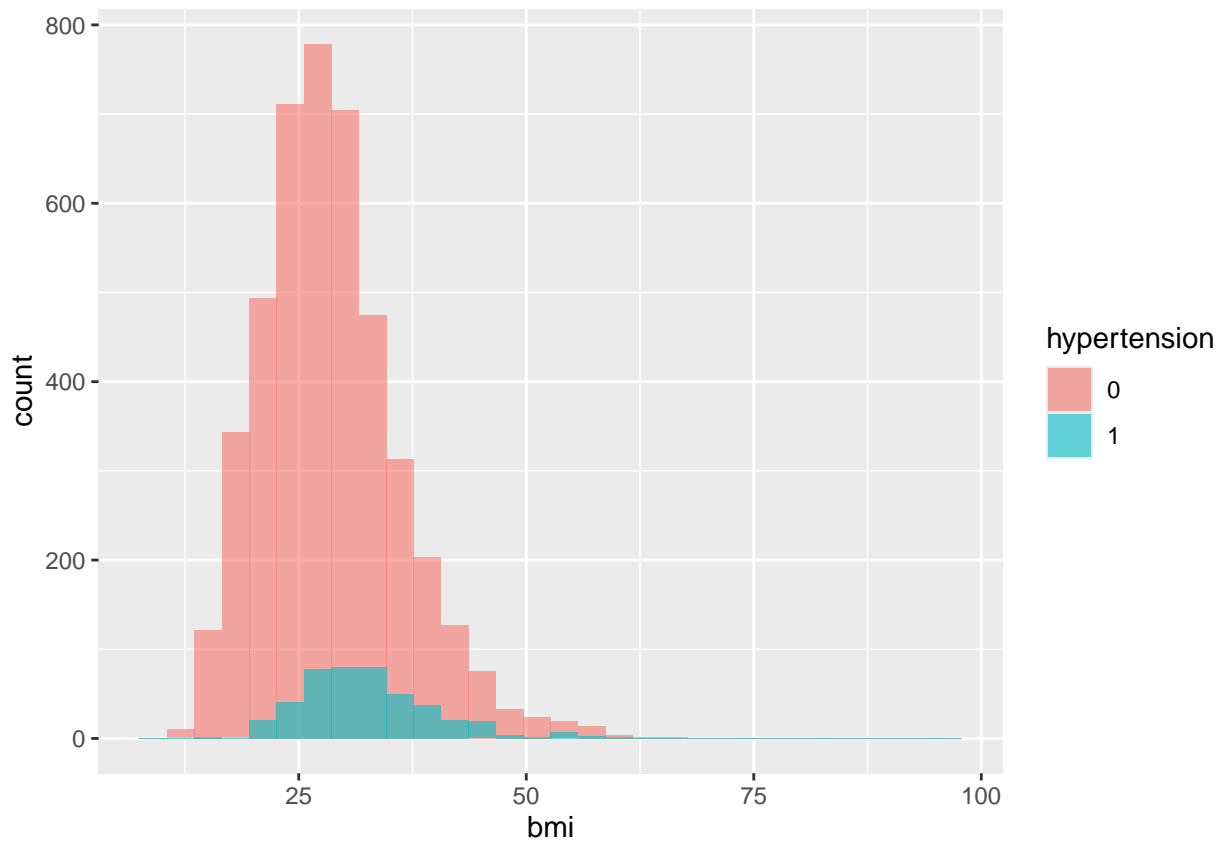
```
## 'summarise()' has grouped output by 'hypertension'. You can override using the
## '.groups' argument.
```

```
print(data_sub_bmi_hyper)
```

```
## # A tibble: 4 x 3
## # Groups:   hypertension [2]
##   hypertension stroke mean
```

```
##   <fct>      <fct> <dbl>
## 1 0         0      28.4
## 2 0         1      30.3
## 3 1         0      33.4
## 4 1         1      30.9
```

```
ggplot(data, aes(x = bmi, fill = hypertension)) +
  geom_histogram(alpha = 0.6, position = "identity", bins = 30)
```



```
data_sub_bmi_catw= data %>% group_by(cat_weight,stroke) %>% summarise(mean = mean(bmi))
```

BMI with stroke

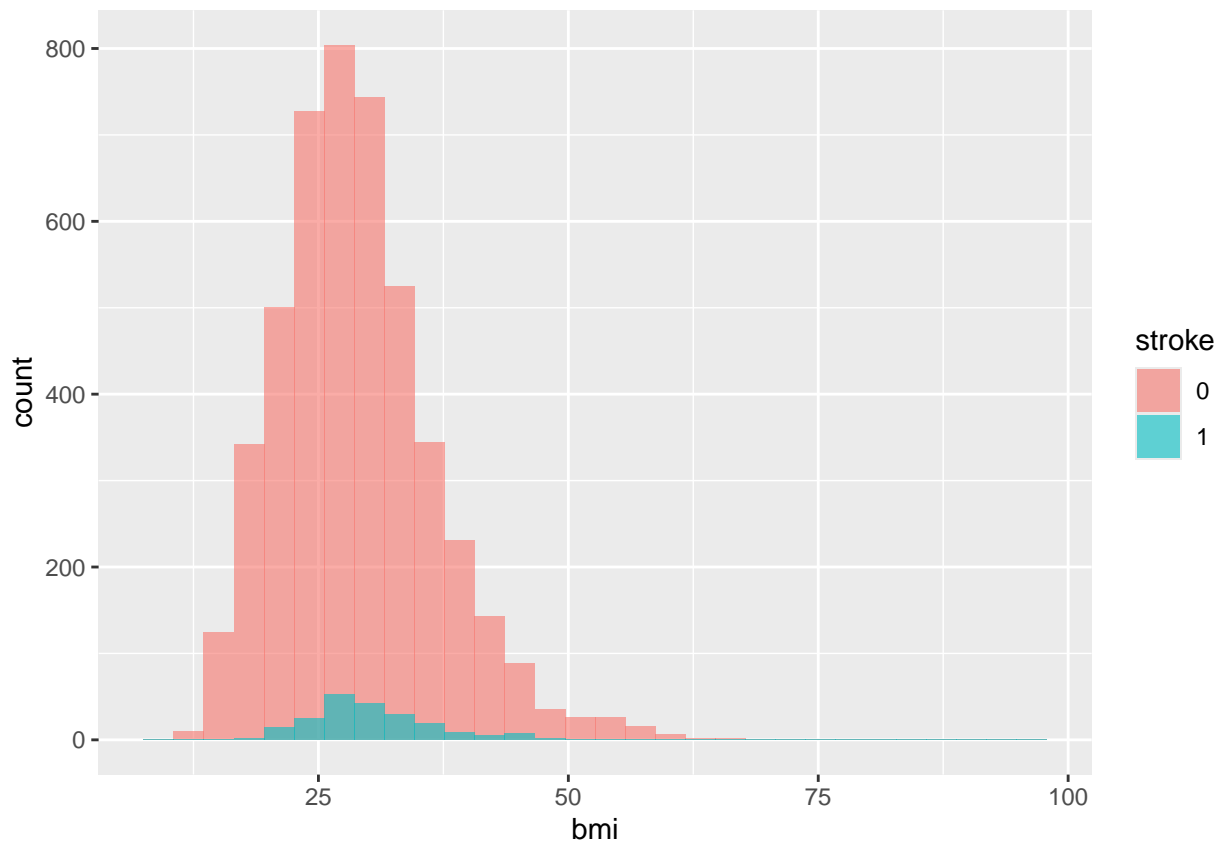
```
## 'summarise()' has grouped output by 'cat_weight'. You can override using the
## '.groups' argument.
```

```
print(data_sub_bmi_catw)
```

```
## # A tibble: 10 x 3
## # Groups:   cat_weight [5]
##   cat_weight      stroke mean
##   <fct>         <fct> <dbl>
## 1 0         0      28.4
## 2 0         1      30.3
## 3 1         0      33.4
## 4 1         1      30.9
```

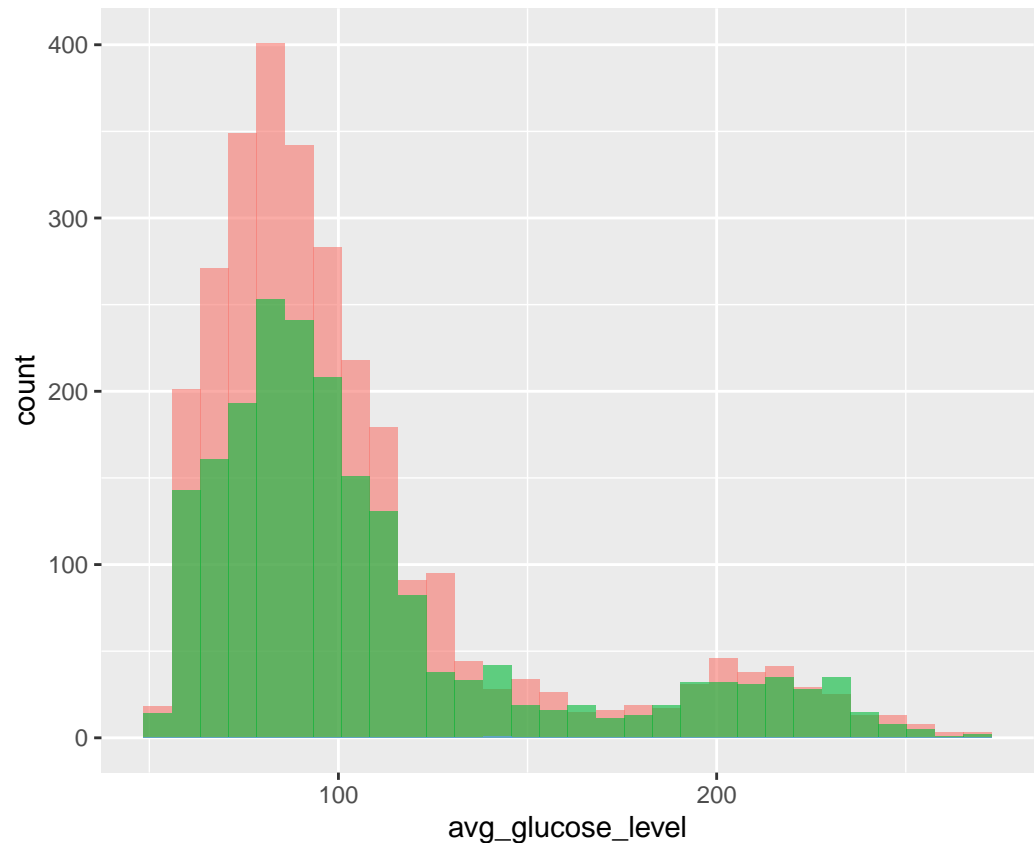
```
## 1 Underweight    0    16.7
## 2 Underweight    1    16.9
## 3 Normal         0    22.2
## 4 Normal         1    22.6
## 5 Overweight     0    27.5
## 6 Overweight     1    27.6
## 7 Obese          0    32.2
## 8 Obese          1    32.1
## 9 Extremely Obese 0    41.1
## 10 Extremely Obese 1    40.3
```

```
ggplot(data, aes(x = bmi, fill = stroke)) +
  geom_histogram(alpha = 0.6, position = "identity", bins = 30)
```



Glucose levels

```
ggplot(data, aes(x = avg_glucose_level, fill = gender)) +
  geom_histogram(alpha = 0.6, position = "identity", bins = 30)
```



Glucose levels with gender

```
data_sub_glu_hyp= data %>% group_by(hypertension,stroke) %>% summarise(mean = mean(avg_glucose_level))
```

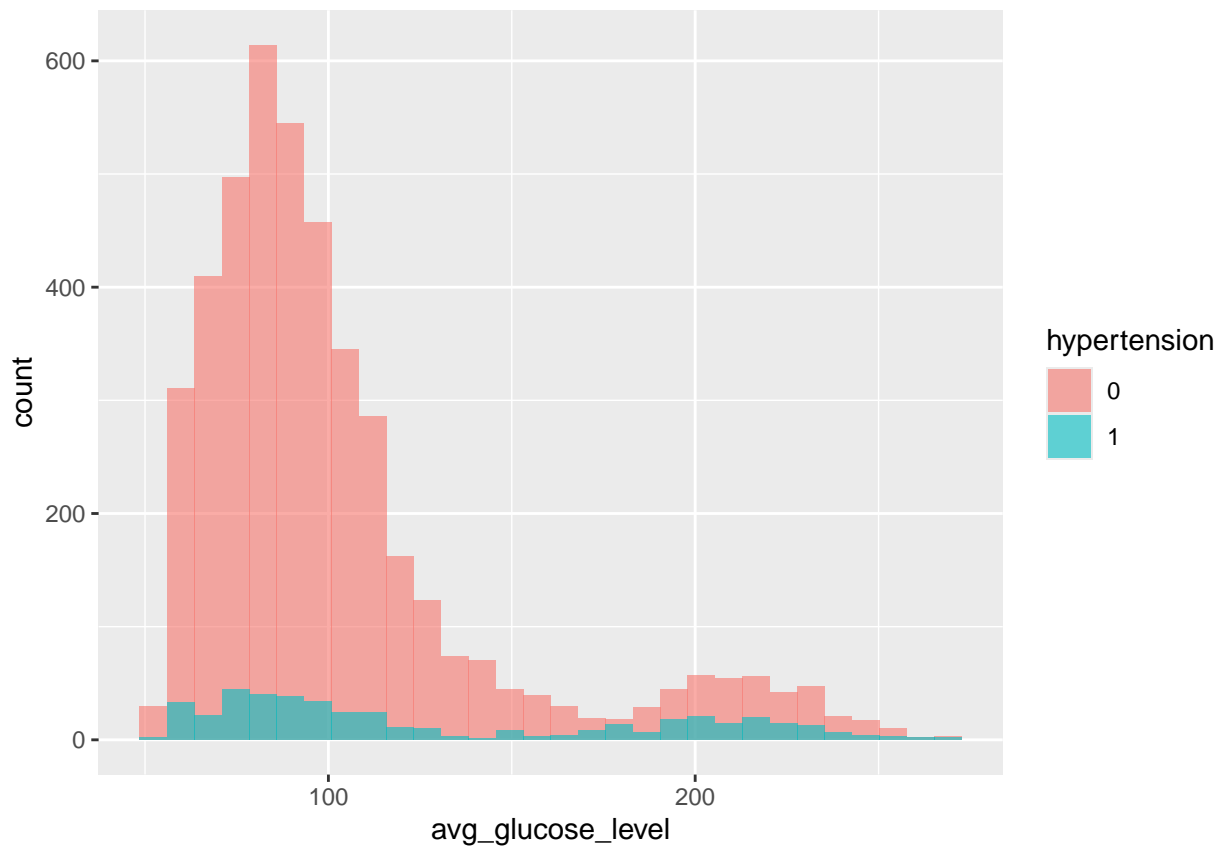
Glucose levels with hypertension

'summarise()' has grouped output by 'hypertension'. You can override using the
'.groups' argument.

```
print(data_sub_glu_hyp)
```

```
## # A tibble: 4 x 3
## # Groups:   hypertension [2]
##   hypertension stroke mean
##   <fct>         <fct> <dbl>
## 1 0             0     102.
## 2 0             1     130.
## 3 1             0     128.
## 4 1             1     146.
```

```
ggplot(data, aes(x = avg_glucose_level, fill = hypertension)) +
  geom_histogram(alpha = 0.6, position = "identity", bins = 30)
```

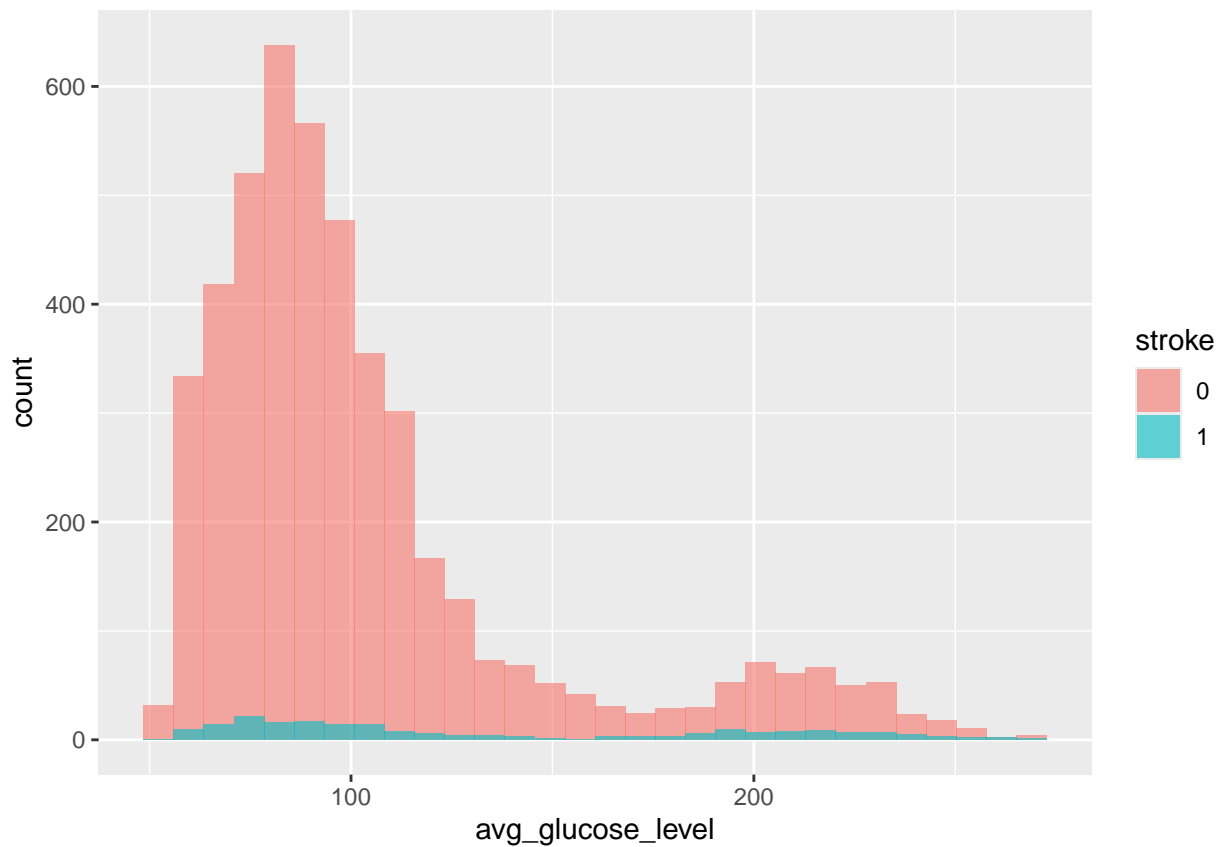



```
data_sub_glu= data %>% group_by(stroke) %>% summarise(mean = mean(avg_glucose_level))
print(data_sub_glu)
```

Glucose levels with stroke

```
## # A tibble: 2 x 2
##   stroke mean
##   <fct> <dbl>
## 1 0      104.
## 2 1      135.
```

```
ggplot(data, aes(x = avg_glucose_level, fill = stroke)) +
  geom_histogram(alpha = 0.6, position = "identity", bins = 30)
```



Age

```
data_sub_age_gender = data %>% group_by(gender, stroke) %>% summarise(mean = mean(age))
```

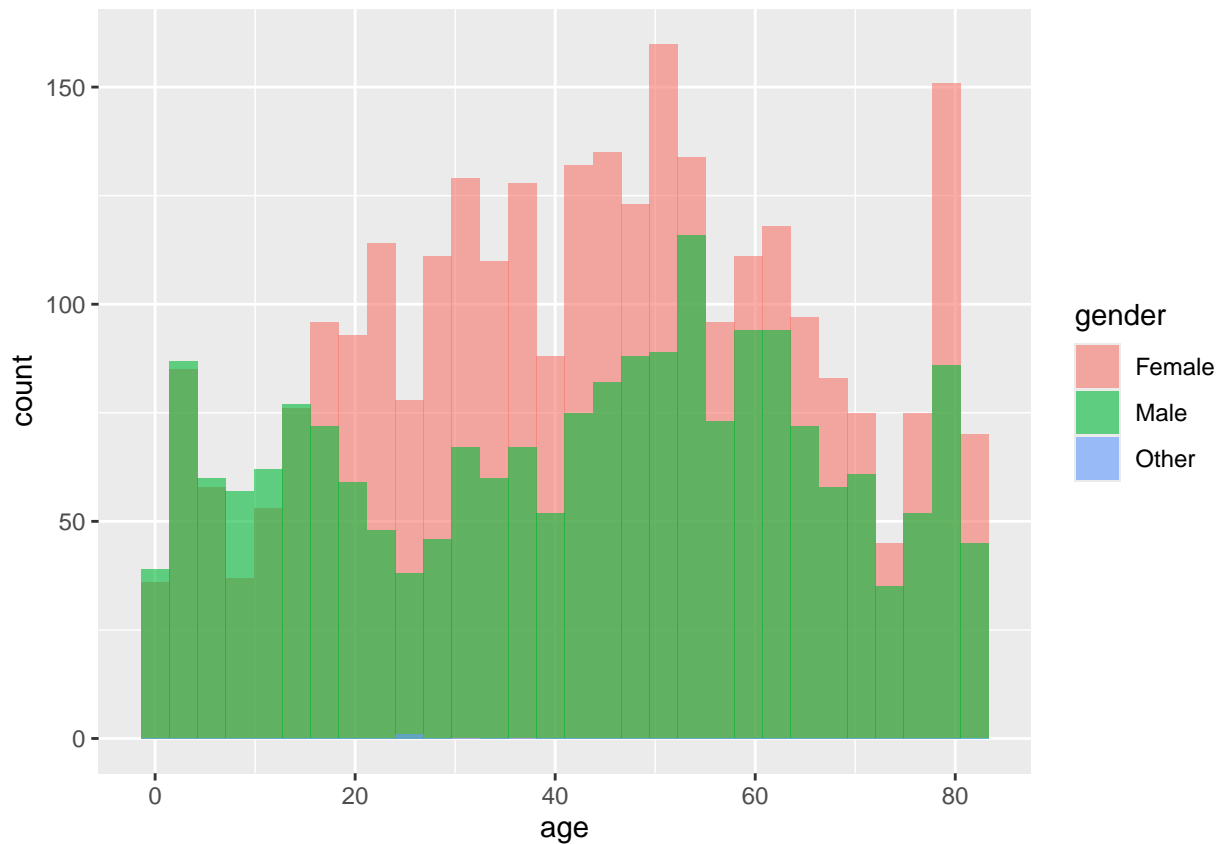
Age levels with gender

'summarise()' has grouped output by 'gender'. You can override using the
'.groups' argument.

```
print(data_sub_age_gender)
```

```
## # A tibble: 5 x 3
## # Groups:   gender [3]
##   gender stroke mean
##   <fct> <fct> <dbl>
## 1 Female 0      42.4
## 2 Female 1      67.2
## 3 Male   0      40.8
## 4 Male   1      68.3
## 5 Other  0       26
```

```
ggplot(data, aes(x = age, fill = gender)) +  
  geom_histogram(alpha = 0.6, position = "identity", bins = 30)
```



```
data_sub_age_hyp= data %>% group_by(hypertension,stroke) %>% summarise(mean = mean(age))
```

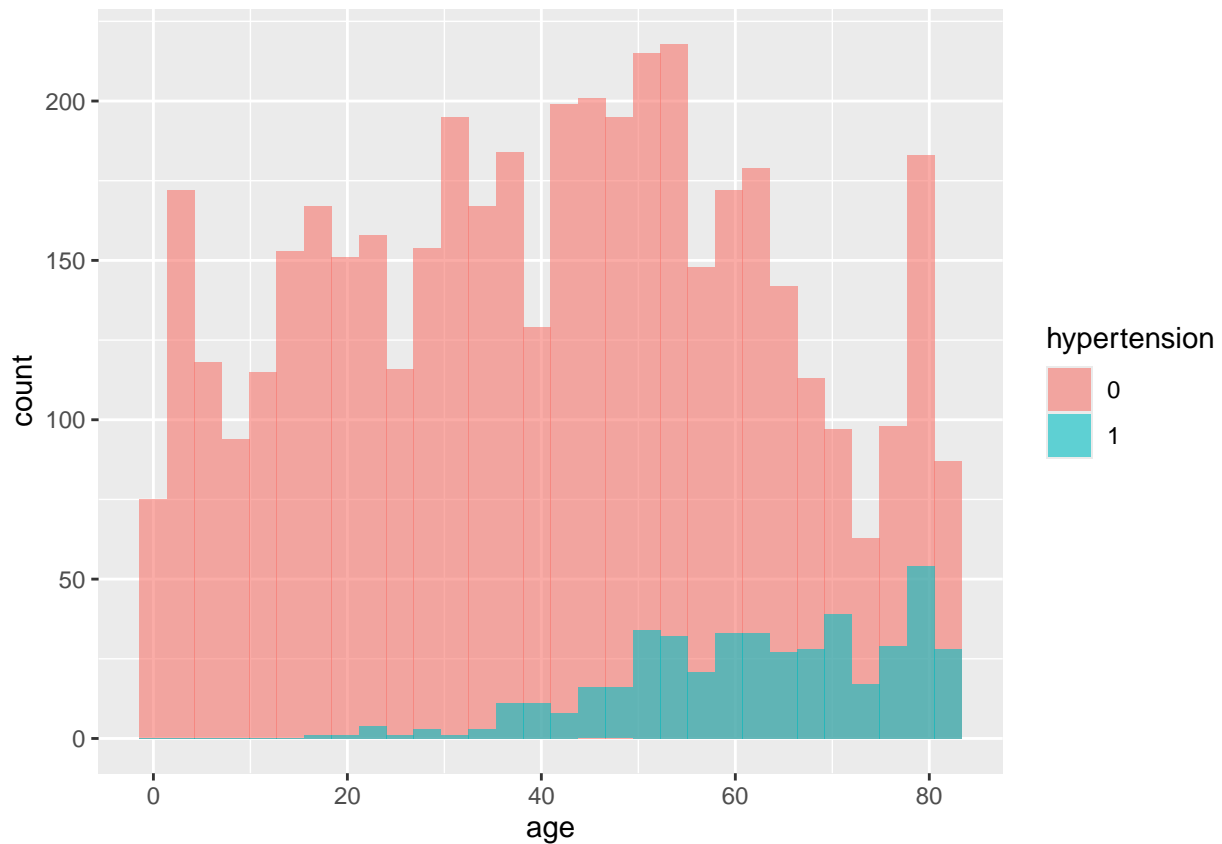
Age levels with hypertension

'summarise()' has grouped output by 'hypertension'. You can override using the
'.groups' argument.

```
print(data_sub_age_hyp)
```

```
## # A tibble: 4 x 3  
## # Groups:   hypertension [2]  
##   hypertension stroke mean  
##   <fct>         <fct> <dbl>  
## 1 0             0      40.0  
## 2 0             1      67.0  
## 3 1             0      61.2  
## 4 1             1      69.6
```

```
ggplot(data, aes(x = age, fill = hypertension)) +
  geom_histogram(alpha = 0.6, position = "identity", bins = 30)
```



jljh ##### Age levels with stroke

```
data_sub_age= data %>% group_by(stroke) %>% summarise(mean = mean(age))
print(data_sub_age)
```

```
## # A tibble: 2 x 2
##   stroke mean
##   <fct> <dbl>
## 1 0      41.8
## 2 1      67.7
```

```
ggplot(data, aes(x = age, fill = stroke)) +
  geom_histogram(alpha = 0.6, position = "identity", bins = 30)
```

