

# 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

##
## Adjuntando el paquete: 'agricolae'

## The following objects are masked from 'package:e1071':
##
##   kurtosis, skewness
```

## 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\_joy", "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   0:4611         0:4833
##  1st Qu.:17740   Male  :2115   1st Qu.:25.00   1: 498         1: 276
##  Median :36922
##  Mean    :36514
##  3rd Qu.:54643
##  Max.    :72940
##          Max.    :82.00
##
##  ever_married      work_type      Residence_type avg_glucose_level
##  No :1756      children      : 687   Rural:2513   Min.    : 55.12
##  Yes:3353      Govt_job      : 657   Urban:2596   1st Qu.: 77.24
##                      Never_worked : 22   Median   : 91.88
##                      Private       :2924   Mean     :106.14
##                      Self-employed: 819   3rd Qu.:114.09
##                      Max.         :271.74
##
##          bmi          smoking_status stroke
##  Min.      :10.30   formerly smoked: 884   0:4860
##  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

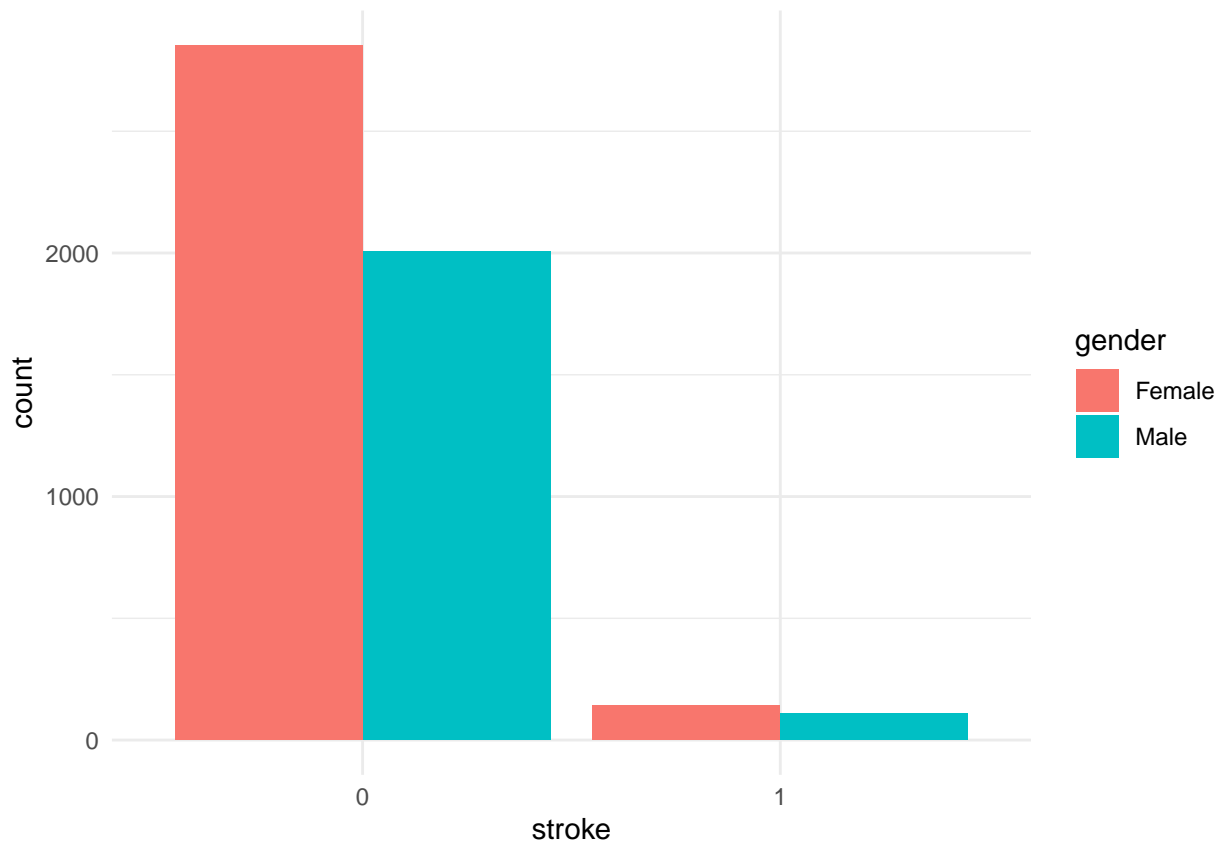
This variable has 2 categories Male and Female (there's one person who is Other but it's only one so we can't make any assumptions about this data).

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
1	Female	141	4.709
1	Male	108	5.106

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



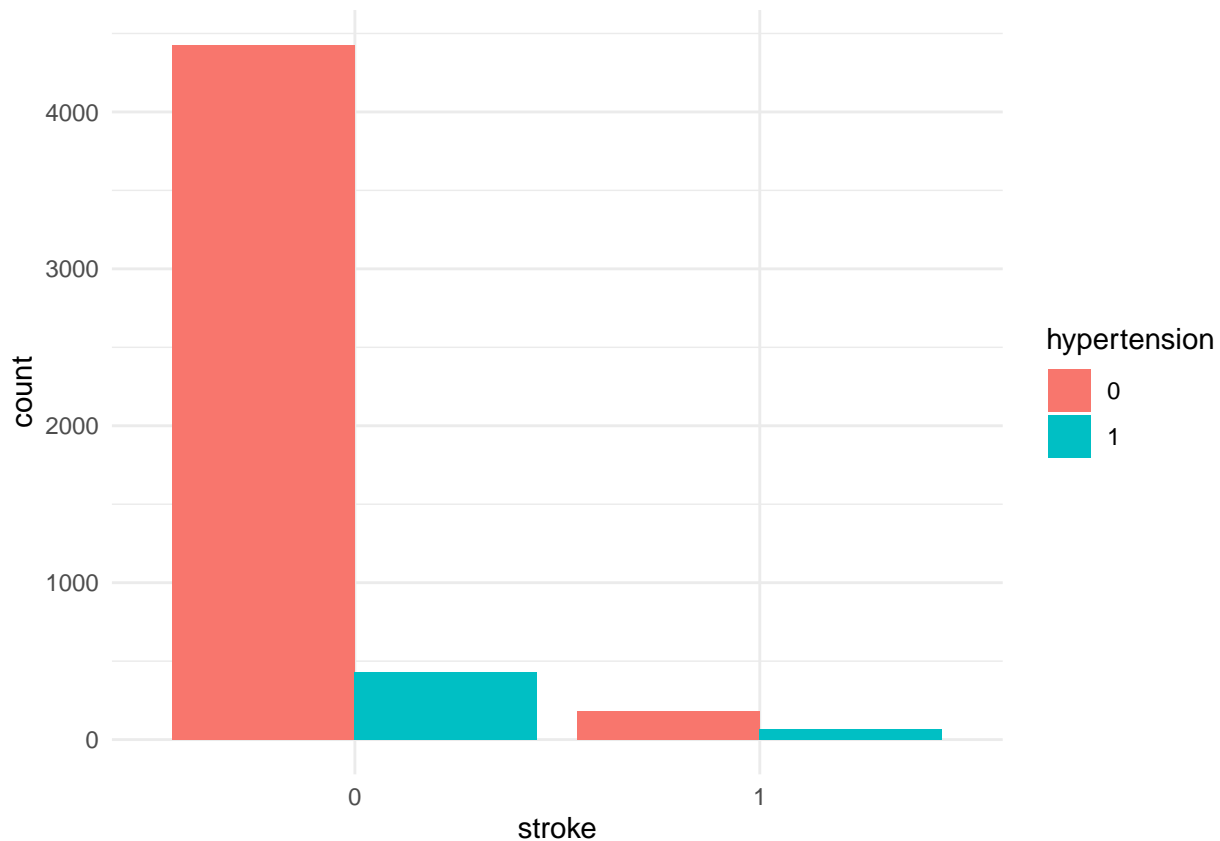
As you can see there's a higher % of male that have had a stroke in the data. This number is a little bit higher than the population so we don't think this would have a great impact in the future model.

## Hypertension

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

stroke	hypertension	n	percent
0	0	4428	96.03
0	1	432	86.75
1	0	183	3.969
1	1	66	13.25

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

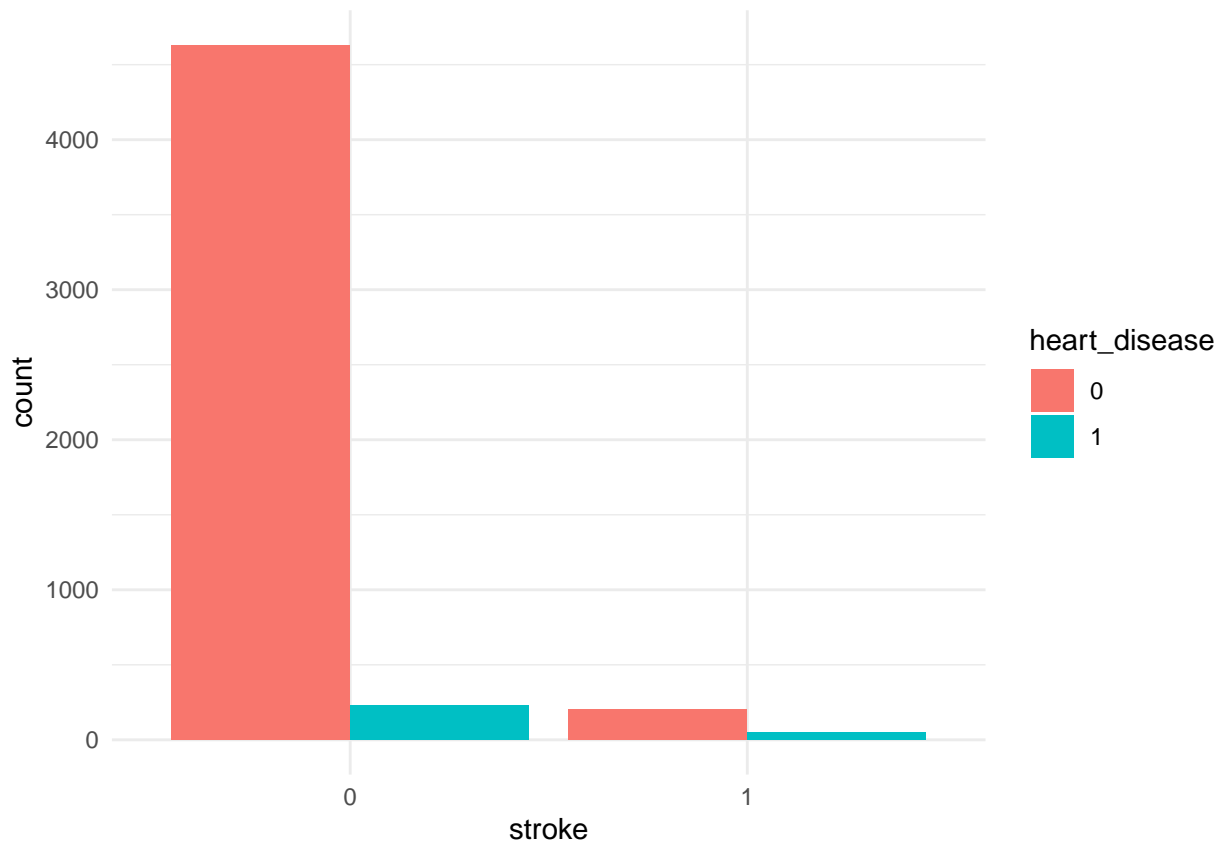


The % of people who have had a stroke and have hypertension are significantly higher than the population. There's a difference approx. 9%. This could mean that if you have hypertension you could be more likely to get a stroke. ### Heart disease

```
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	4631	95.82
0	1	229	82.97
1	0	202	4.18
1	1	47	17.03

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



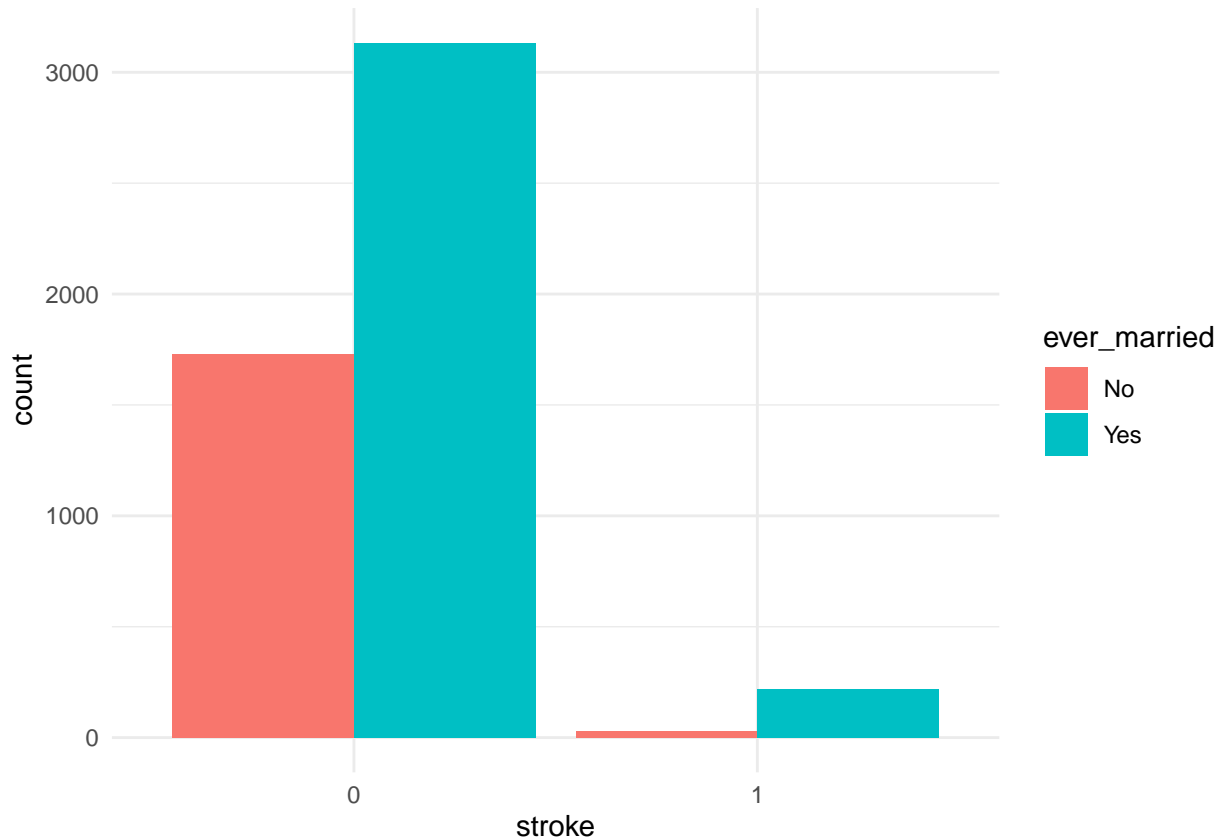
The data of the people who have a heart diseased look even more likely to get a stroke than the ones who have hypertension. Maybe there's a relation between hypertension and having a heart diseased. Heart diseased is a variable to check for more details.

## 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	1727	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())+theme_minimal()
```

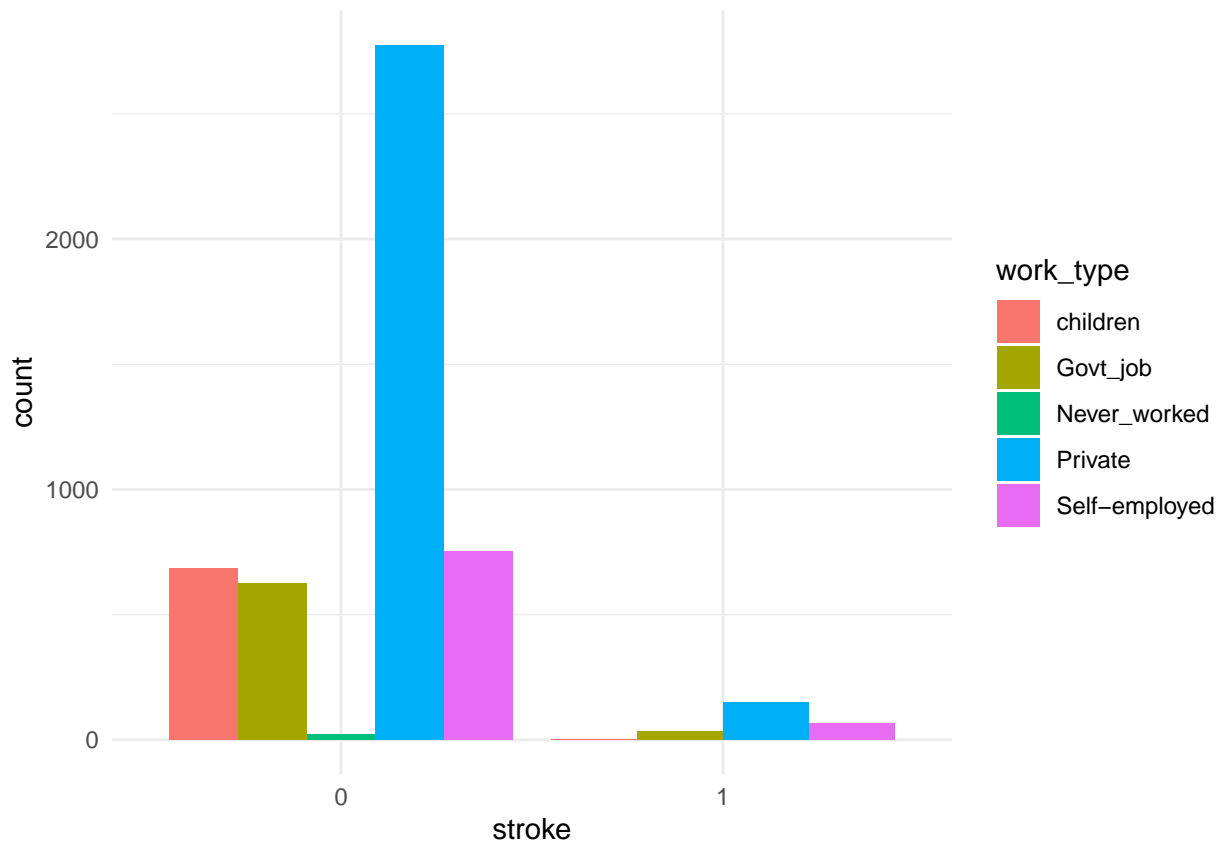


there's more people that have had a stroke you have been married than not been married. The difference with the population is big but with the ones that never married maybe not get married lower your changes or getting a stroke. ### Work Type

```
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	2775	94.9
0	Self-employed	754	92.06
1	children	2	0.2911
1	Govt_job	33	5.023
1	Private	149	5.096
1	Self-employed	65	7.937

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

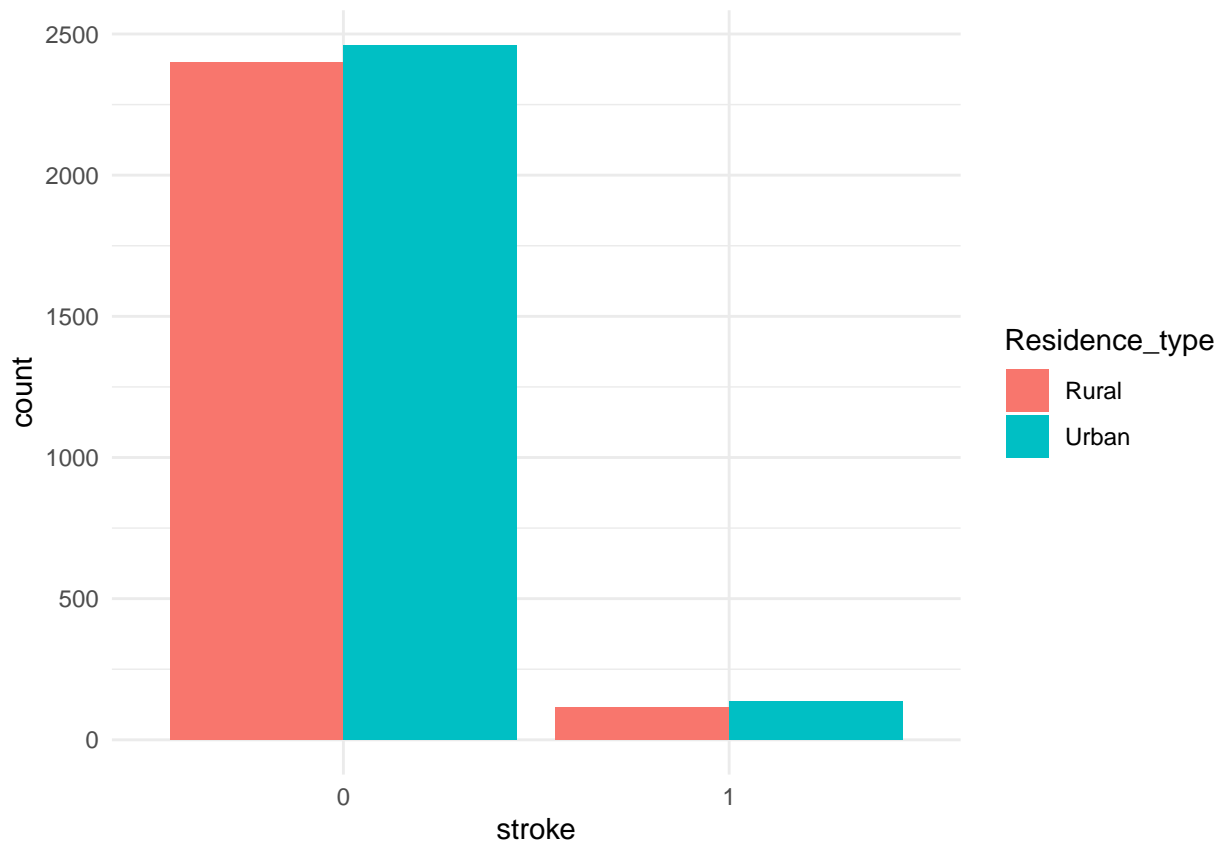


Clearly the self employed have the highest % of people who have had a stroke. This could be because of stress and the people who work with children have the lowest. Hope that working with children reduces your chances. ### Resident Type

```
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	2399	95.46
0	Urban	2461	94.8
1	Rural	114	4.536
1	Urban	135	5.2

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



The part of the city that you live looks like it doesn't matter because the % of stroke is very similar to the population and almost the same between the types of resident. This could mean something

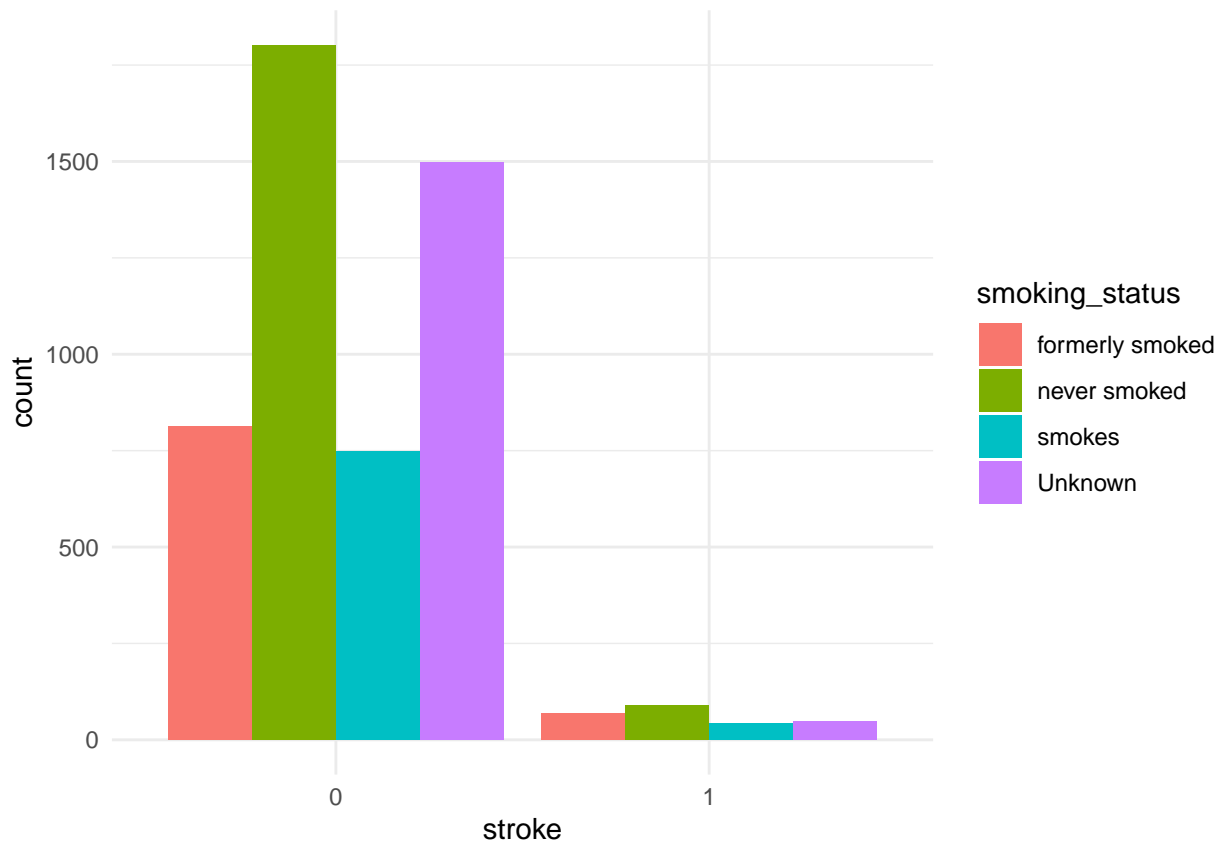
### Smoking Status

```
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	814	92.08
0	never smoked	1802	95.24
0	smokes	747	94.68
0	Unknown	1497	96.96
1	formerly smoked	70	7.919
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())+theme_minimal()
```





This results call the attention because the people you formerly smoked have 3% higher than the population but obe who smokes or never have similar numbers. This could be a multivariate problem because this type of data probably mix more than one variable at a time. 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)
frec_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 = frec_table
```

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

```
## frec_table
##      Underweight      Normal      Overweight      Obese Extremely Obese
##           349           1230           1409           1000           920
```

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

```
## frec_table
##      Underweight      Normal      Overweight      Obese Extremely Obese
##      7.110839      25.061125      28.708231      20.374898      18.744906
```

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.

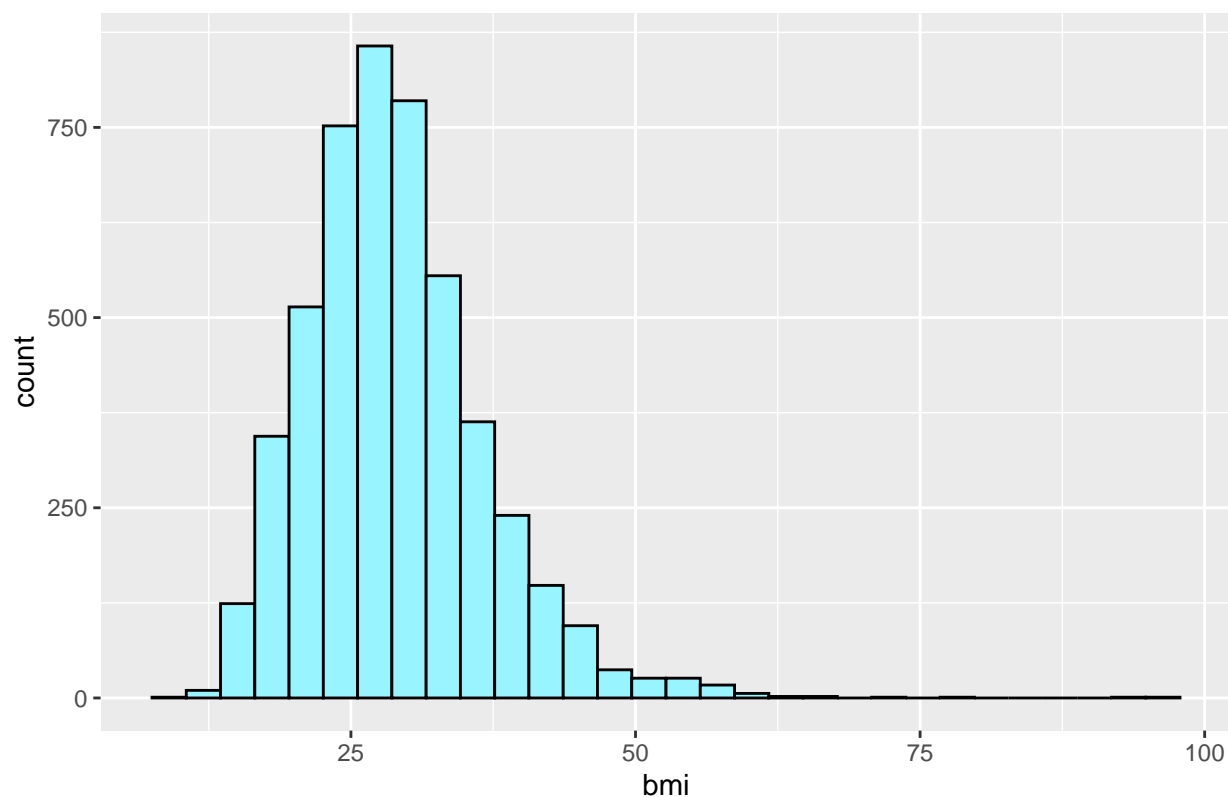
```
stat.freq(aux_bmi)
```

```
## $variance
## [1] 63.33528
##
## $mean
## [1] 28.84576
##
## $median
## [1] 28.00923
##
## $mode
##      [- -]      mode
## [1,] 25 30 27.22513
```

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

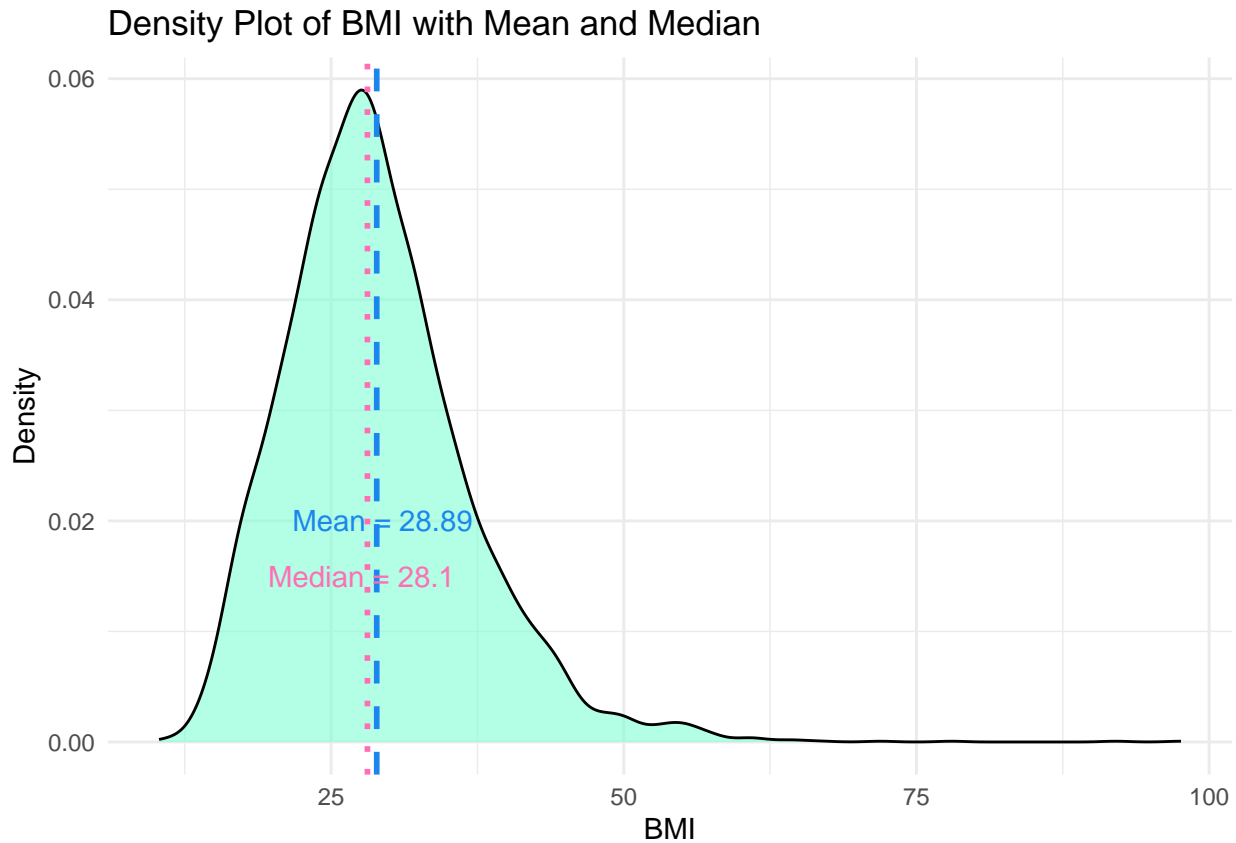
```
## 'stat_bin()' using 'bins = 30'. Pick better value with 'binwidth'.
```

# BMI Distribution



```
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] 1.055063
```

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

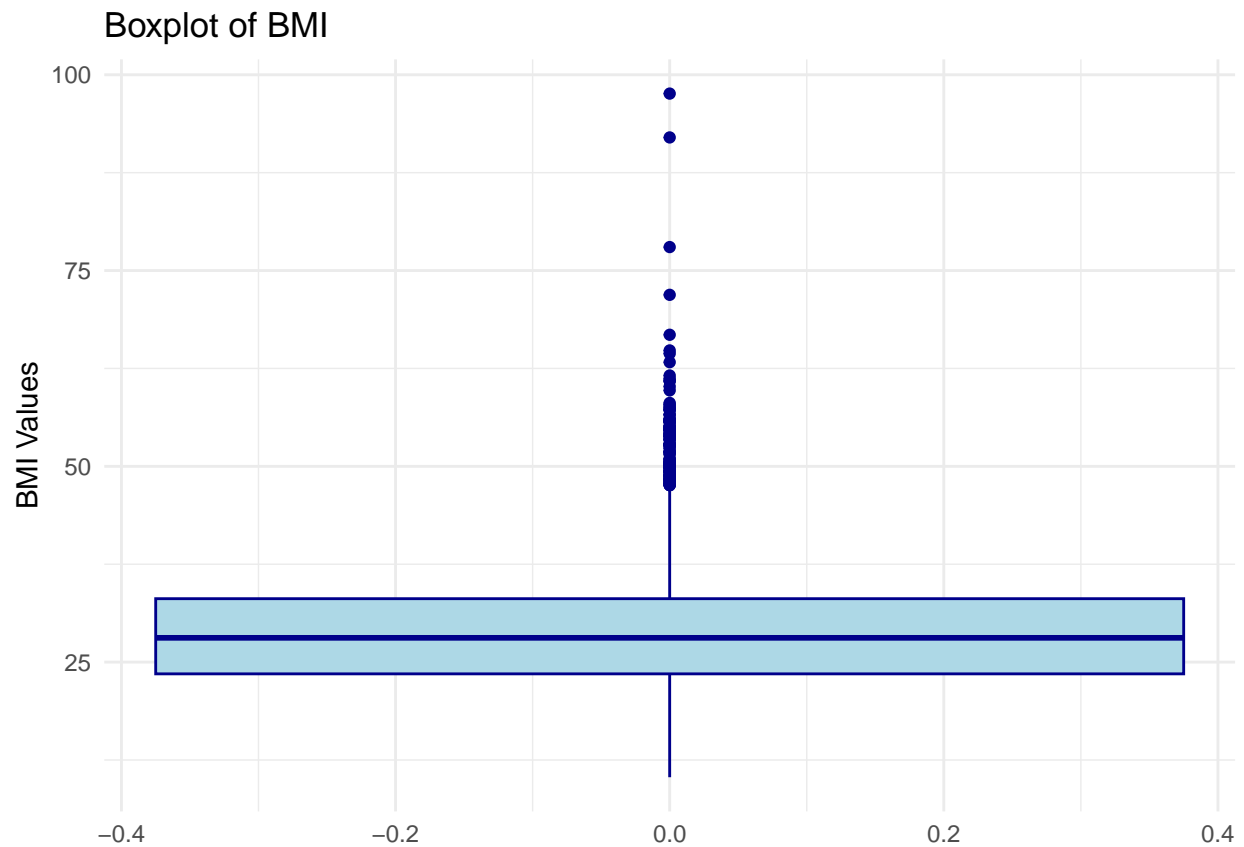
```
## [1] 3.36233
```

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 tail is to this side explaining the outliers.

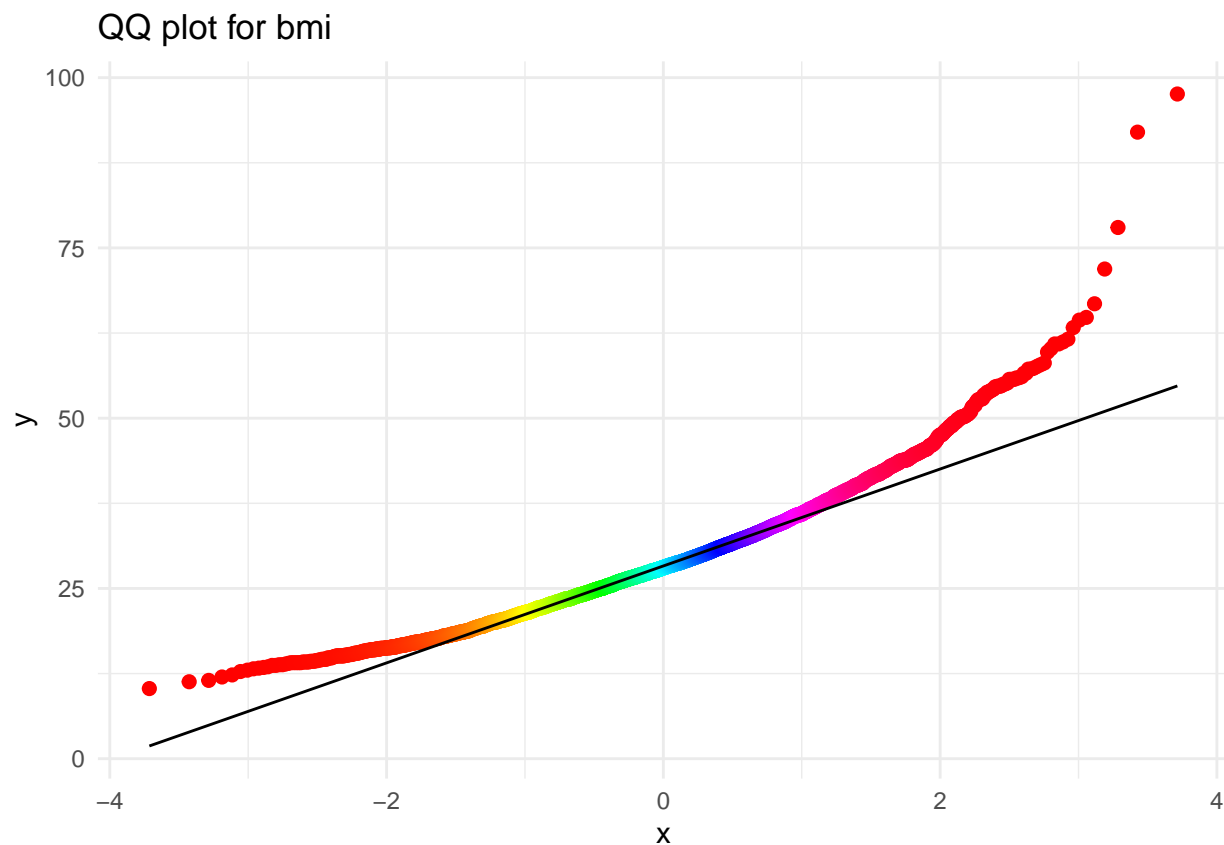
To see if the outliers are correctly in the heavier side we can check the boxplot of the *bmi*

```
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. YOu can see this in the qqplot next.

```
ggplot(data, aes(sample=bmi)) + stat_qq(size=2,color=rainbow(4908))+stat_qq_line()+theme_minimal()+  
  labs(title = "QQ plot for bmi")
```

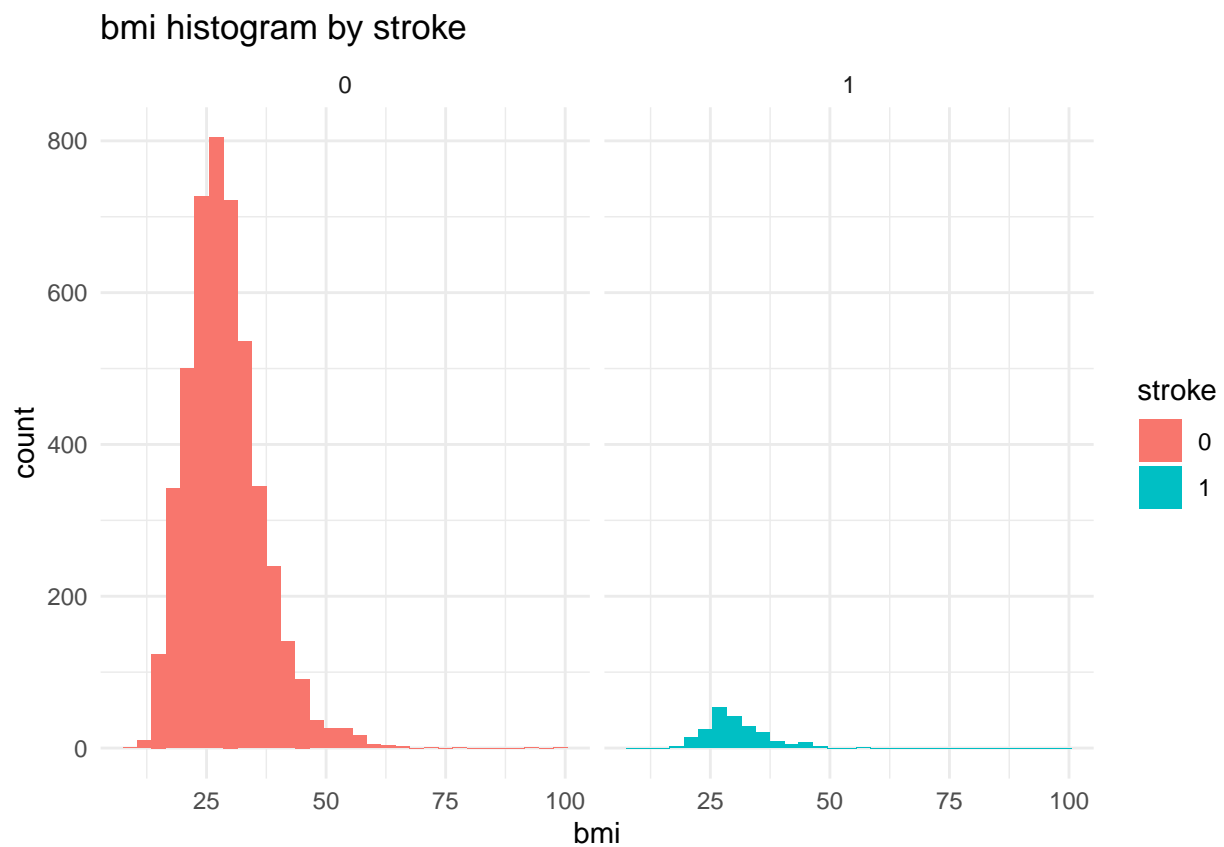


```
data_sub_bmi_catw= data %>% group_by(cat_weight,stroke) %>% summarise(mean = mean(bmi),.groups = "drop")
print(data_sub_bmi_catw)
```

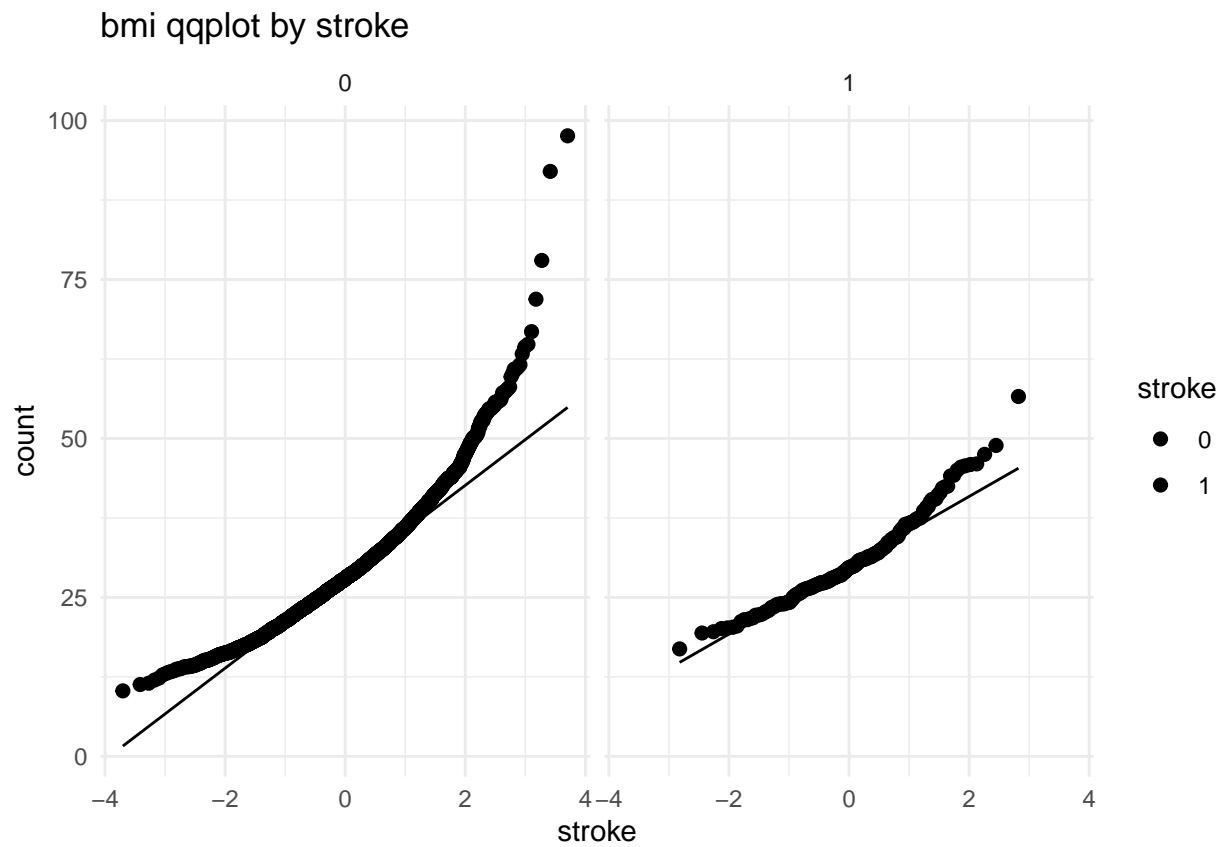
### BMI with stroke

```
## # A tibble: 10 x 3
##   cat_weight   stroke mean
##   <fct>       <fct> <dbl>
## 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(binwidth = 3)+
  facet_wrap(~stroke)+theme_minimal()+labs(title = "bmi histogram by stroke",x="bmi",y="count")
```

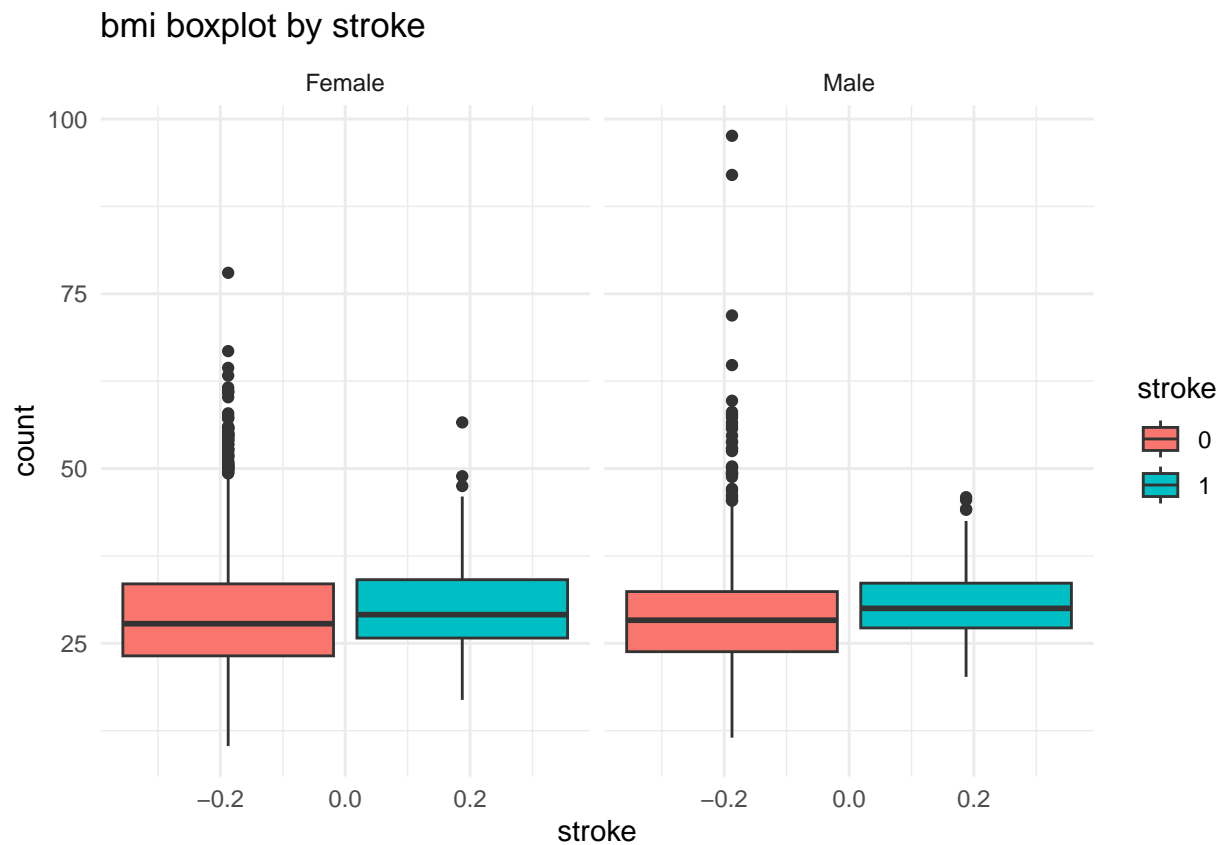


```
ggplot(data, aes(sample = bmi, fill = stroke))+  
  stat_qq(size=2)+ stat_qq_line() +  
  facet_wrap(~stroke)+theme_minimal()+labs(title = "bmi qqplot by stroke",x="stroke",y="count")
```



```
ggplot(data, aes(y = bmi, fill =stroke))+  
  geom_boxplot()+  
  facet_wrap(~gender)+theme_minimal()+labs(title = "bmi boxplot by stroke",x="stroke",y="count")
```





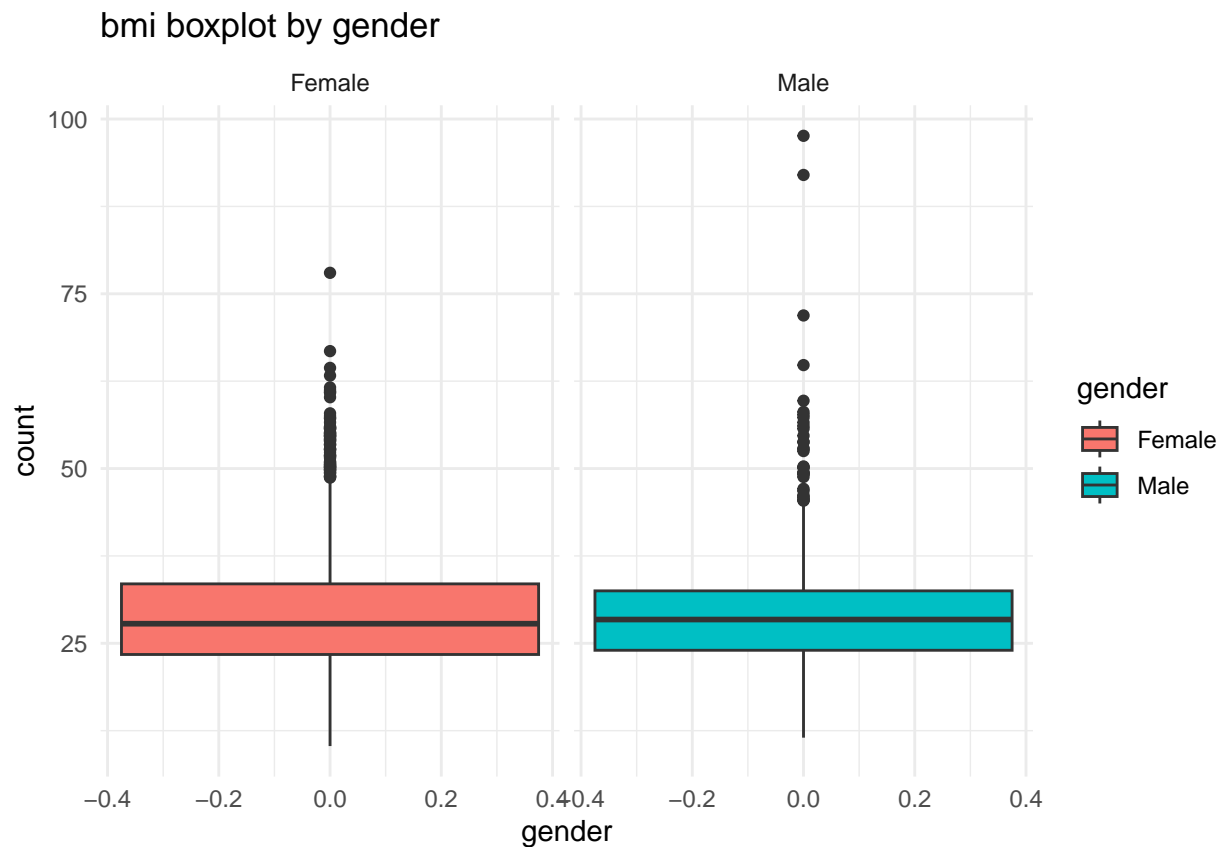
*#gender*

```
data_sub_bmi_gender= data %>% group_by(gender,stroke) %>% summarise(mean = mean(bmi),.groups = "drop")
print(data_sub_bmi_gender)
```

BMI with gender

```
## # A tibble: 4 x 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
```

```
ggplot(data, aes(y = bmi, fill = gender))+
  geom_boxplot()+
  facet_wrap(~gender)+theme_minimal()+labs(title = "bmi boxplot by gender",x="gender",y="count")
```



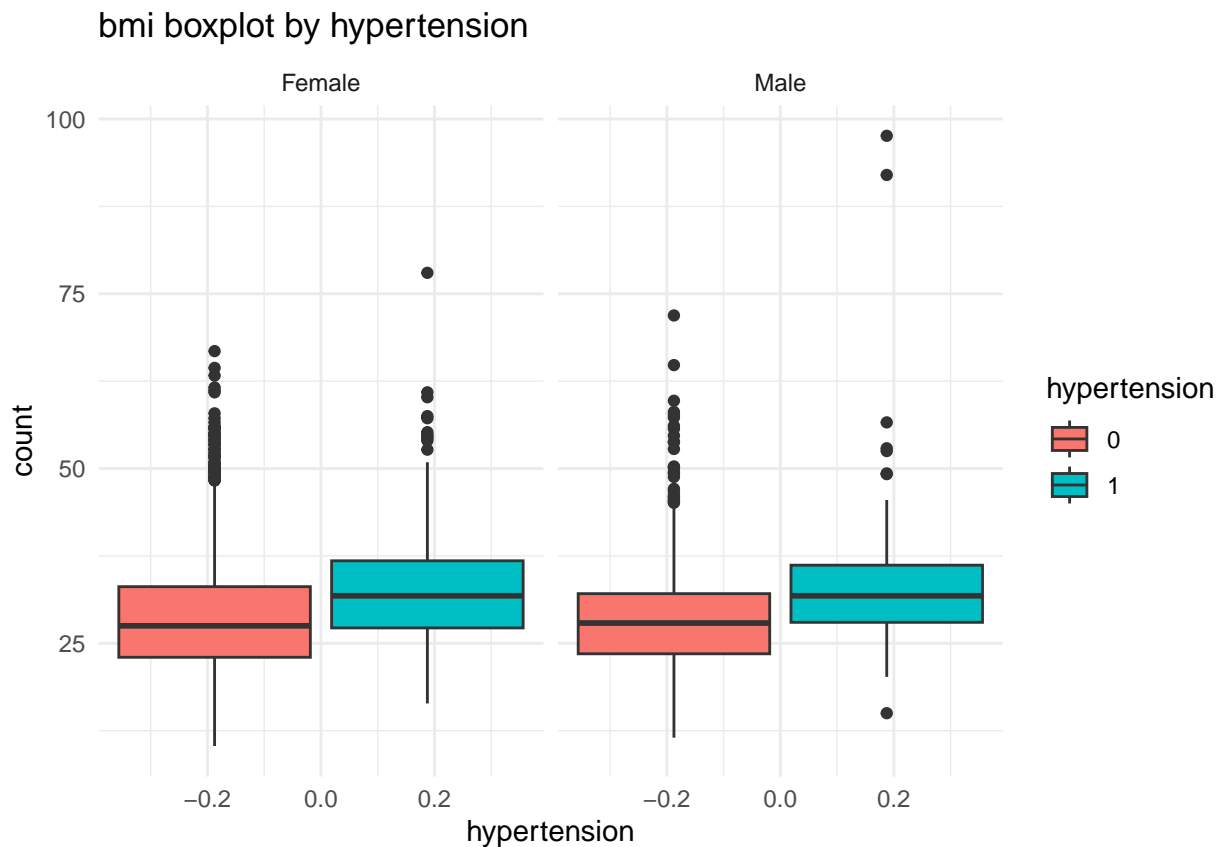
*#hypertension*

```
data_sub_bmi_hyper= data %>% group_by(hypertension,stroke) %>% summarise(mean = mean(bmi),.groups = "dr
print(data_sub_bmi_hyper)
```

### BMI with hypertension

```
## # A tibble: 4 x 3
##   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(y = bmi, fill =hypertension))+
  geom_boxplot()+
  facet_wrap(~gender)+theme_minimal()+labs(title = "bmi boxplot by hypertension",x="hypertension",y="co
```

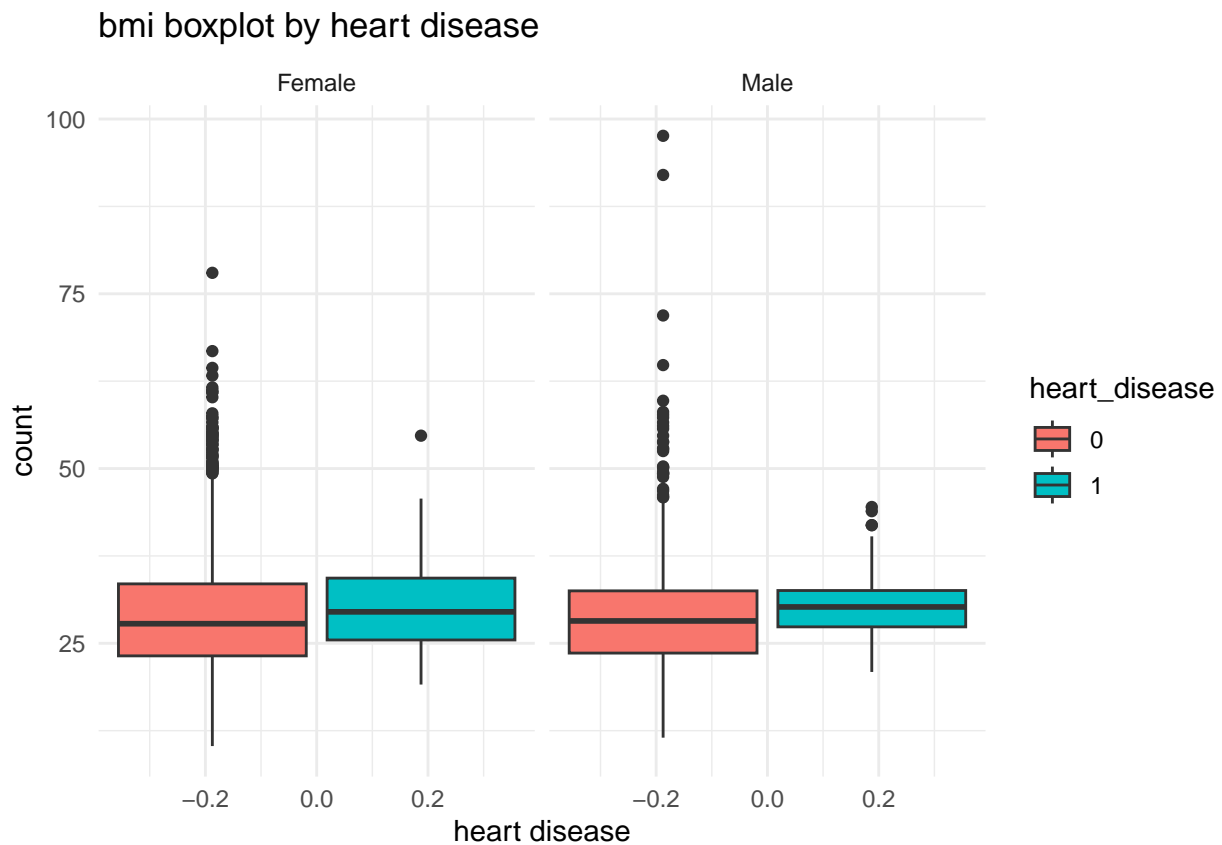


```
data_sub_bmi_hd= data %>% group_by(heart_disease,stroke) %>% summarise(mean = mean(bmi),.groups = "drop")
print(data_sub_bmi_hd)
```

### BMI by heart diseased

```
## # A tibble: 4 x 3
##   heart_disease stroke mean
##   <fct>         <fct> <dbl>
## 1 0             0     28.8
## 2 0             1     30.4
## 3 1             0     30.2
## 4 1             1     30.8
```

```
ggplot(data, aes(y = bmi, fill =heart_disease))+
  geom_boxplot()+
  facet_wrap(~gender)+theme_minimal()+labs(title = "bmi boxplot by heart disease",x="heart disease",y="bmi")
```

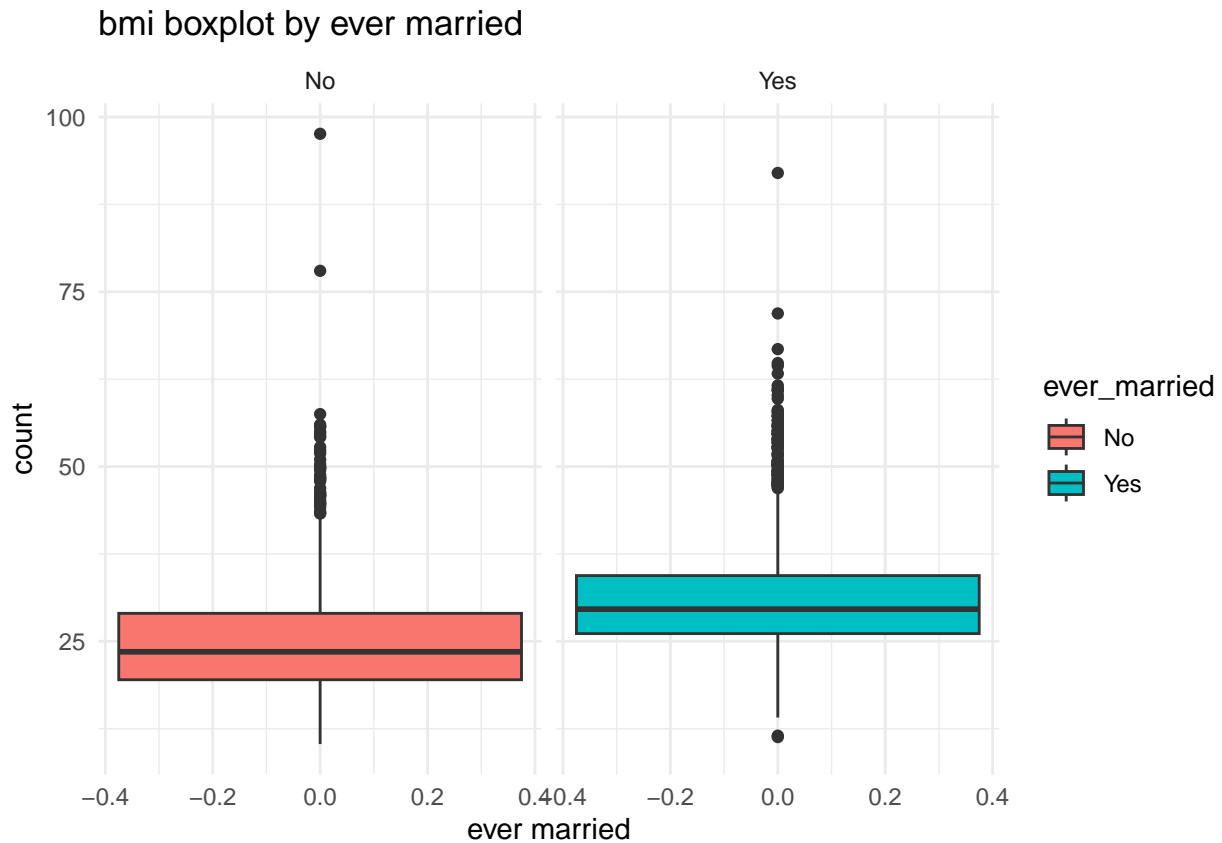


```
data_sub_bmi_em= data %>% group_by(ever_married,stroke) %>% summarise(mean = mean(bmi),.groups = "drop")
print(data_sub_bmi_em)
```

### BMI by ever married

```
## # A tibble: 4 x 3
##   ever_married stroke  mean
##   <fct>         <fct> <dbl>
## 1 No           0      25.2
## 2 No           1      29.9
## 3 Yes          0      30.9
## 4 Yes          1      30.5
```

```
ggplot(data, aes(y = bmi, fill =ever_married))+
  geom_boxplot()+
  facet_wrap(~ever_married)+theme_minimal()+labs(title = "bmi boxplot by ever married",x="ever married")
```

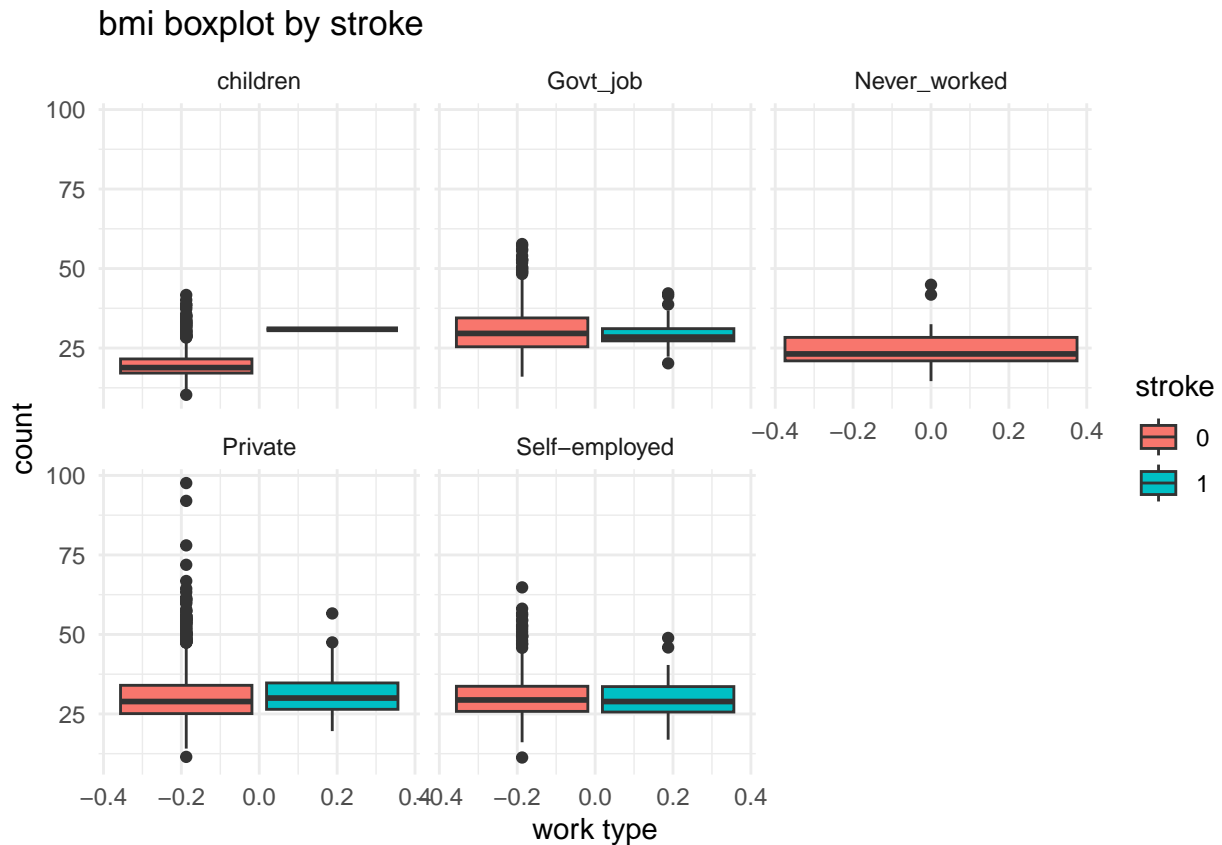


```
data_sub_bmi_wt= data %>% group_by(work_type,stroke) %>% summarise(mean = mean(bmi),.groups = "drop")
print(data_sub_bmi_wt)
```

### BMI by work type

```
## # A tibble: 9 x 3
##   work_type   stroke mean
##   <fct>       <fct> <dbl>
## 1 children    0      20.0
## 2 children    1      30.9
## 3 Govt_job    0      30.6
## 4 Govt_job    1      29.4
## 5 Never_worked 0      25.5
## 6 Private     0      30.3
## 7 Private     1      31.1
## 8 Self-employed 0      30.3
## 9 Self-employed 1      29.6
```

```
ggplot(data, aes(y = bmi, fill =stroke))+
  geom_boxplot()+
  facet_wrap(~work_type)+theme_minimal()+labs(title = "bmi boxplot by stroke",x="work type",y="count")
```

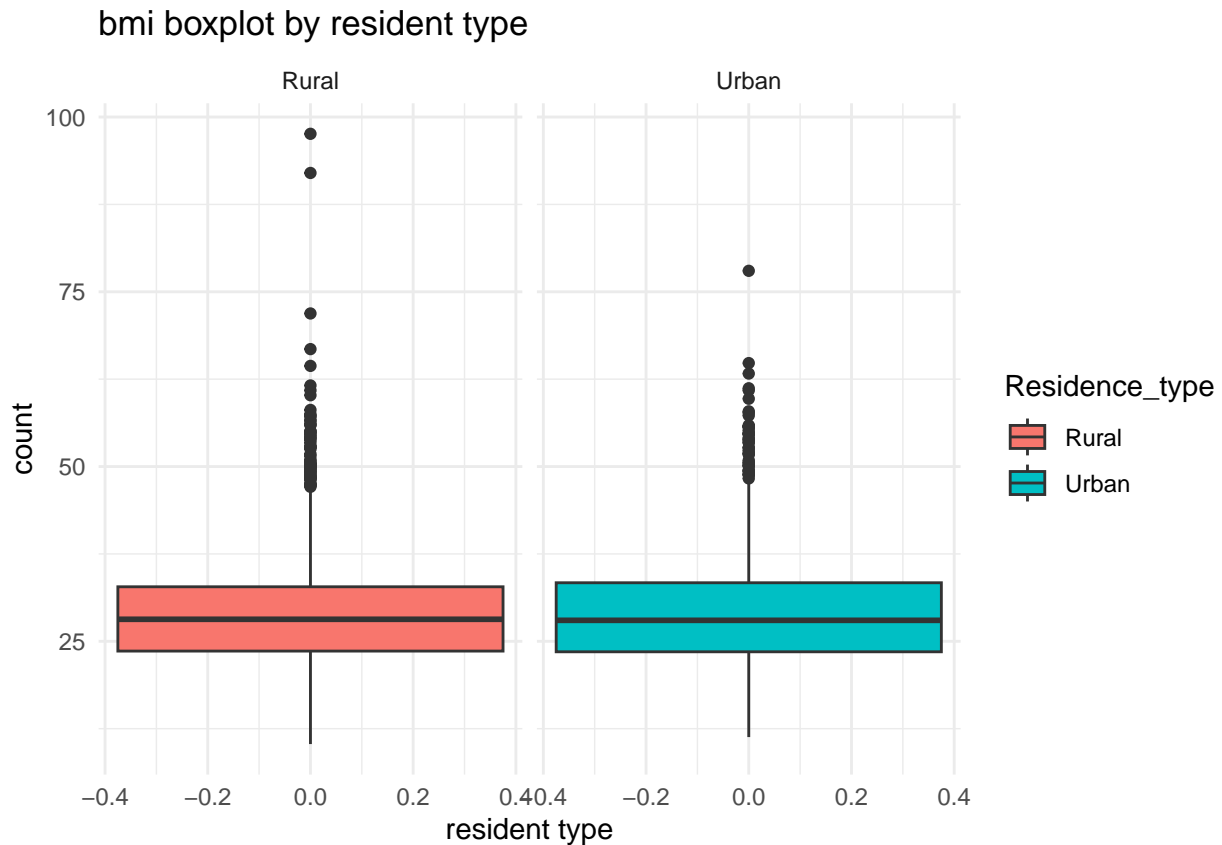


```
data_sub_bmi_tr= data %>% group_by(Residence_type,stroke) %>% summarise(mean = mean(bmi),.groups = "drop")
print(data_sub_bmi_tr)
```

### BMI by resident type

```
## # A tibble: 4 x 3
##   Residence_type stroke  mean
##   <fct>          <fct> <dbl>
## 1 Rural          0      28.8
## 2 Rural          1      30.1
## 3 Urban          0      28.8
## 4 Urban          1      30.8
```

```
ggplot(data, aes(y = bmi, fill =Residence_type))+
  geom_boxplot()+
  facet_wrap(~Residence_type)+theme_minimal()+labs(title = "bmi boxplot by resident type",x="resident type")
```



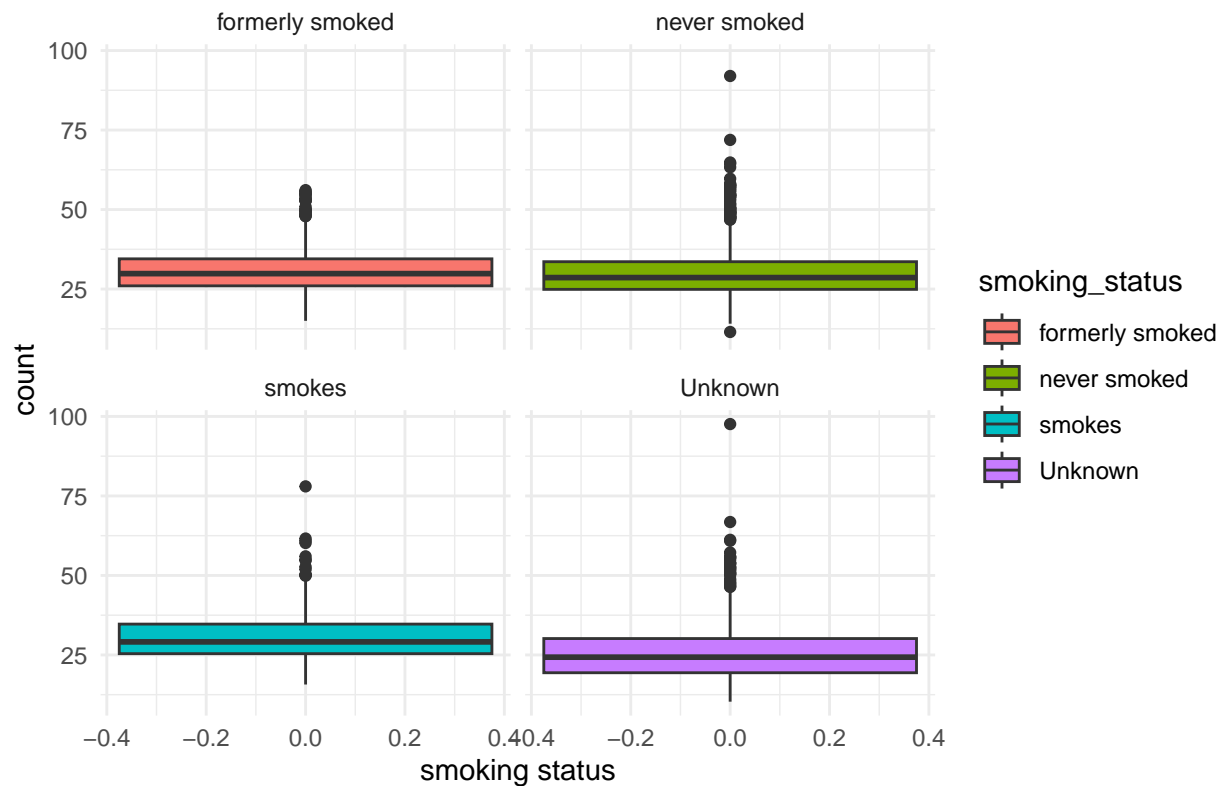
```
data_sub_bmi_catw= data %>% group_by(smoking_status,stroke) %>% summarise(mean = mean(bmi),.groups = "drop")
print(data_sub_bmi_catw)
```

### BMI by smoking status

```
## # A tibble: 8 x 3
##   smoking_status stroke mean
##   <fct>          <fct> <dbl>
## 1 formerly smoked 0      30.7
## 2 formerly smoked 1      31.0
## 3 never smoked   0      30.0
## 4 never smoked   1      30.4
## 5 smokes         0      30.5
## 6 smokes         1      30.6
## 7 Unknown        0      25.6
## 8 Unknown        1      29.4
```

```
ggplot(data, aes(y = bmi, fill =smoking_status))+
  geom_boxplot()+
  facet_wrap(~smoking_status)+theme_minimal()+labs(title = "bmi boxplot by smoking status",x="smoking status")
```

## bmi boxplot by smoking status



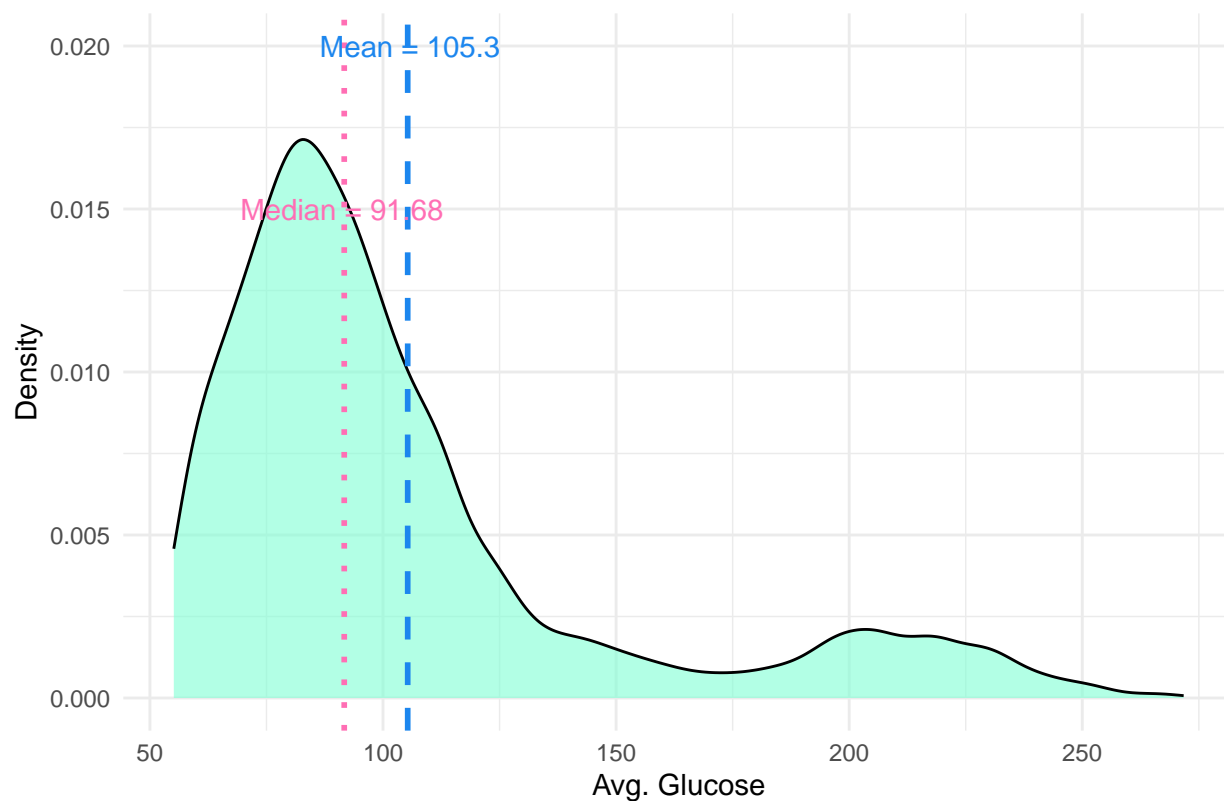
## Glucose levels

```
glc_mean <- mean(data$avg_glucose_level)
glc_median <- median(data$avg_glucose_level)

# Density plot with a vertical line at the mean
ggplot(data, aes(x = avg_glucose_level)) +
  geom_density(fill = "aquamarine1", alpha = 0.6) +
  geom_vline(aes(xintercept = glc_mean), color = "dodgerblue2", linetype = "dashed", size = 1) +
  geom_vline(aes(xintercept = glc_median), color = "hotpink1", linetype = "dotted", size = 1) +
  labs(title = "Density Plot of avg glucose with Mean and Median", x = "Avg. Glucose", y = "Density") +
  annotate("text", x = glc_mean + 0.5, y = 0.02, label = paste("Mean =", round(glc_mean, 2)), color = "dodgerblue2") +
  annotate("text", x = glc_median - 0.5, y = 0.015, label = paste("Median =", round(glc_median, 2)), color = "hotpink1") +
  theme_minimal()
```



Density Plot of avg glucose with Mean and Median



ffd

```
skewness(data$avg_glucose_level)
```

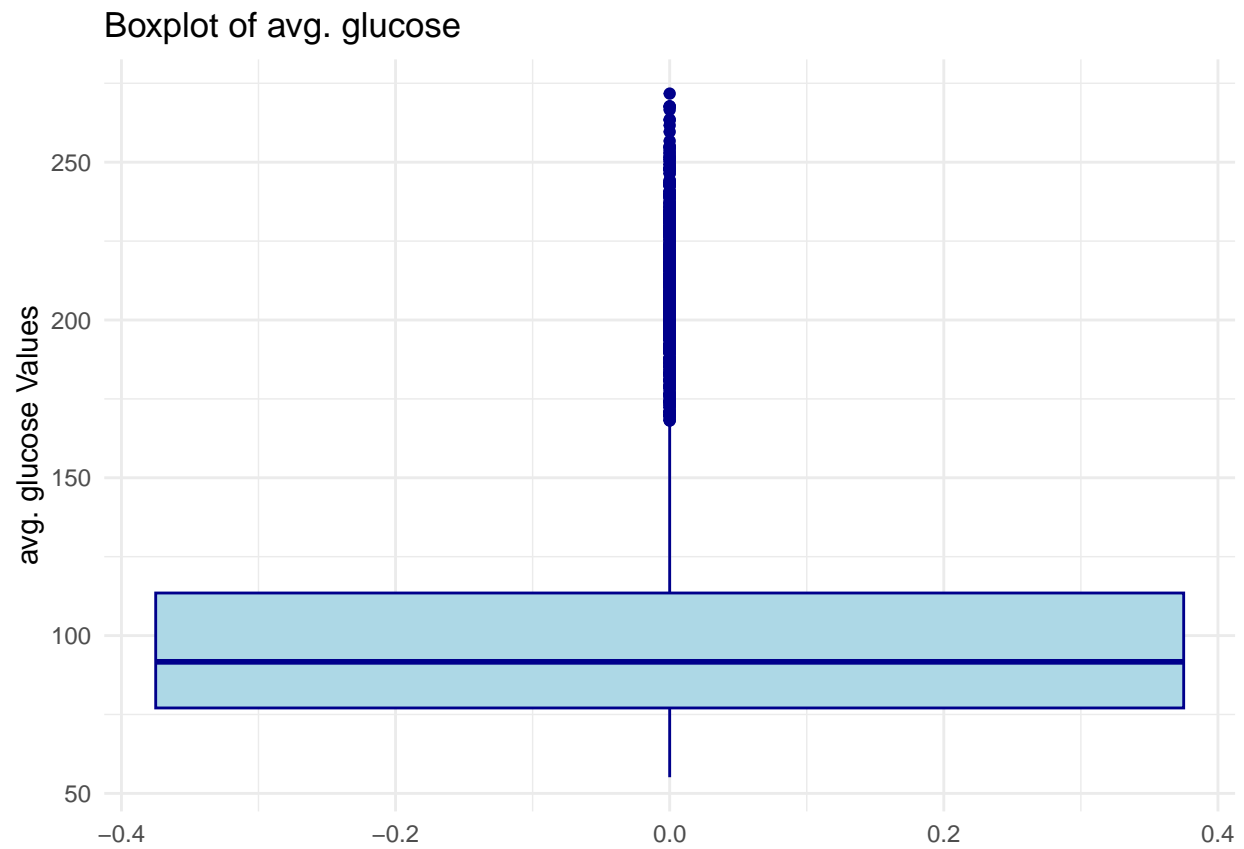
```
## [1] 1.614619
```

```
kurtosis(data$avg_glucose_level)
```

```
## [1] 1.907085
```

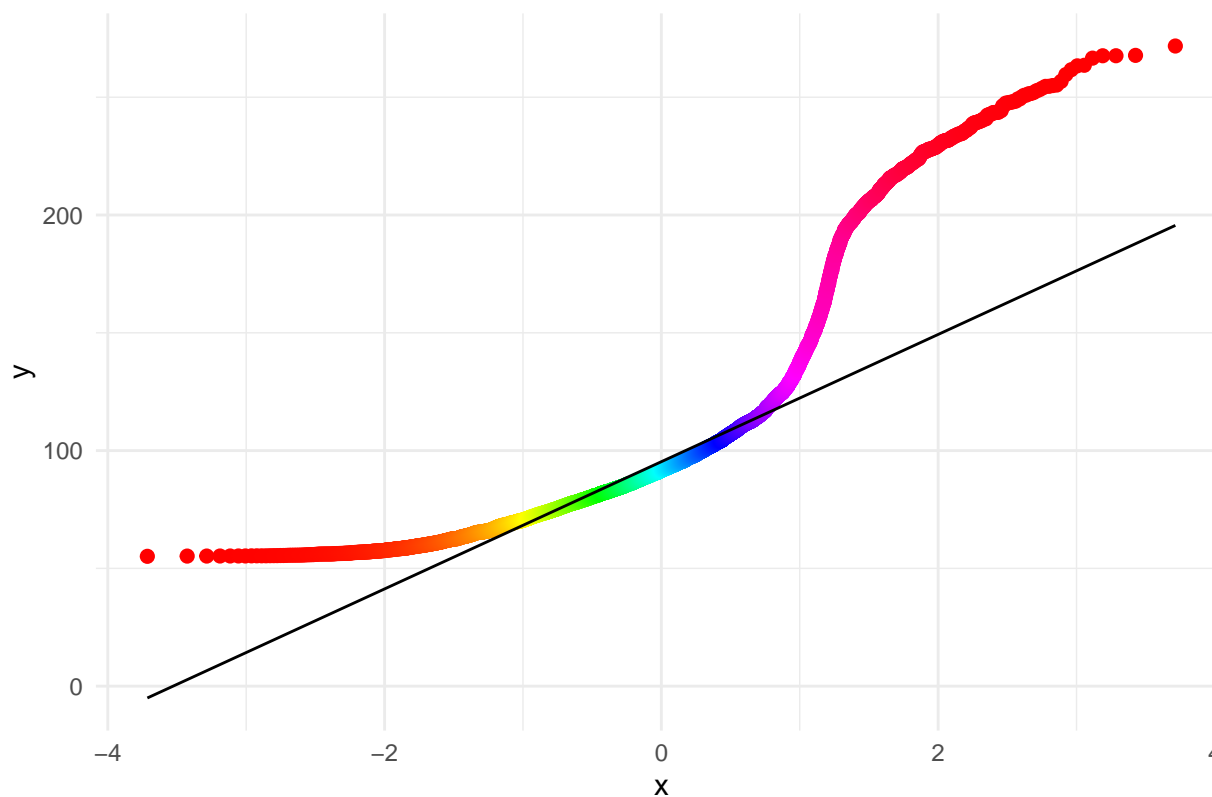
djfi

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



```
ggplot(data, aes(sample=avg_glucose_level)) + stat_qq(size=2,color=rainbow(4908))+stat_qq_line()+theme_minimal()
labs(title = "QQ plot for avg. glucose")
```

QQ plot for avg. glucose



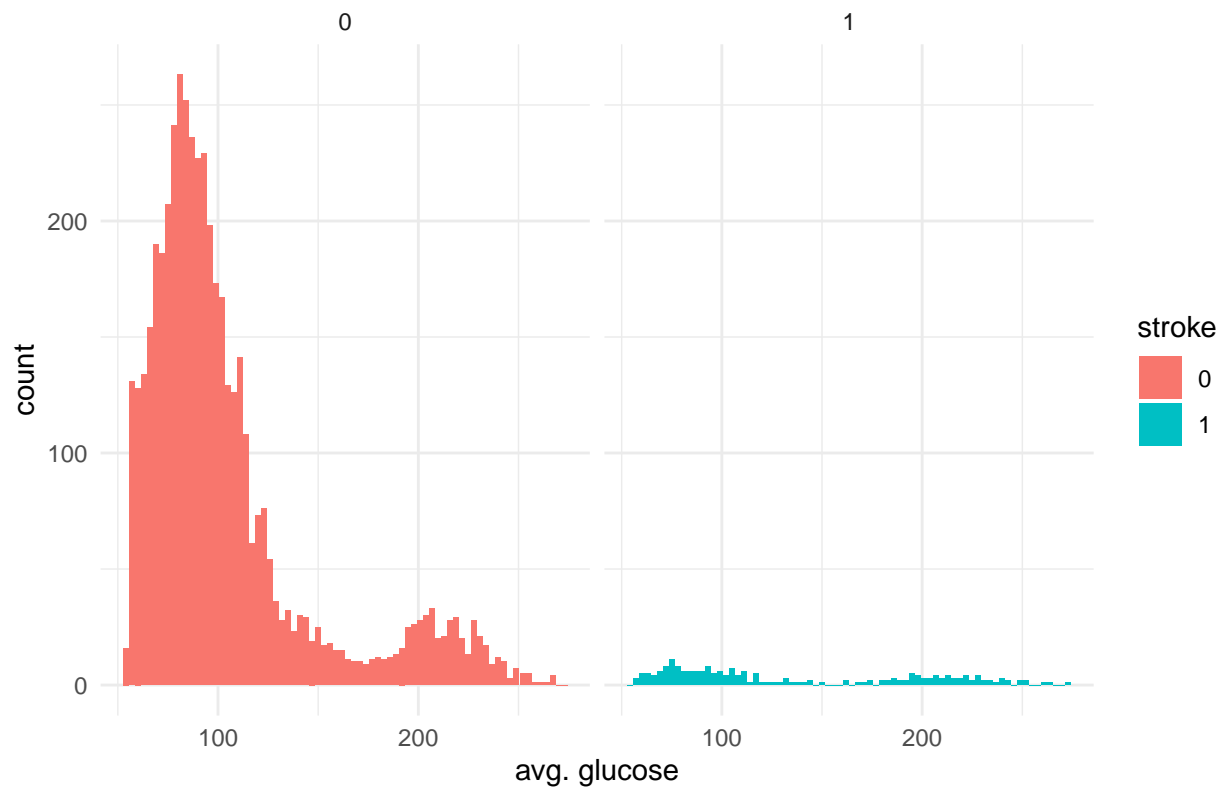
```
data_sub_glu_st= data %>% group_by(stroke) %>% summarise(mean = mean(avg_glucose_level),.groups = "drop")
print(data_sub_glu_st)
```

Avg. Glucose levels with stroke

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

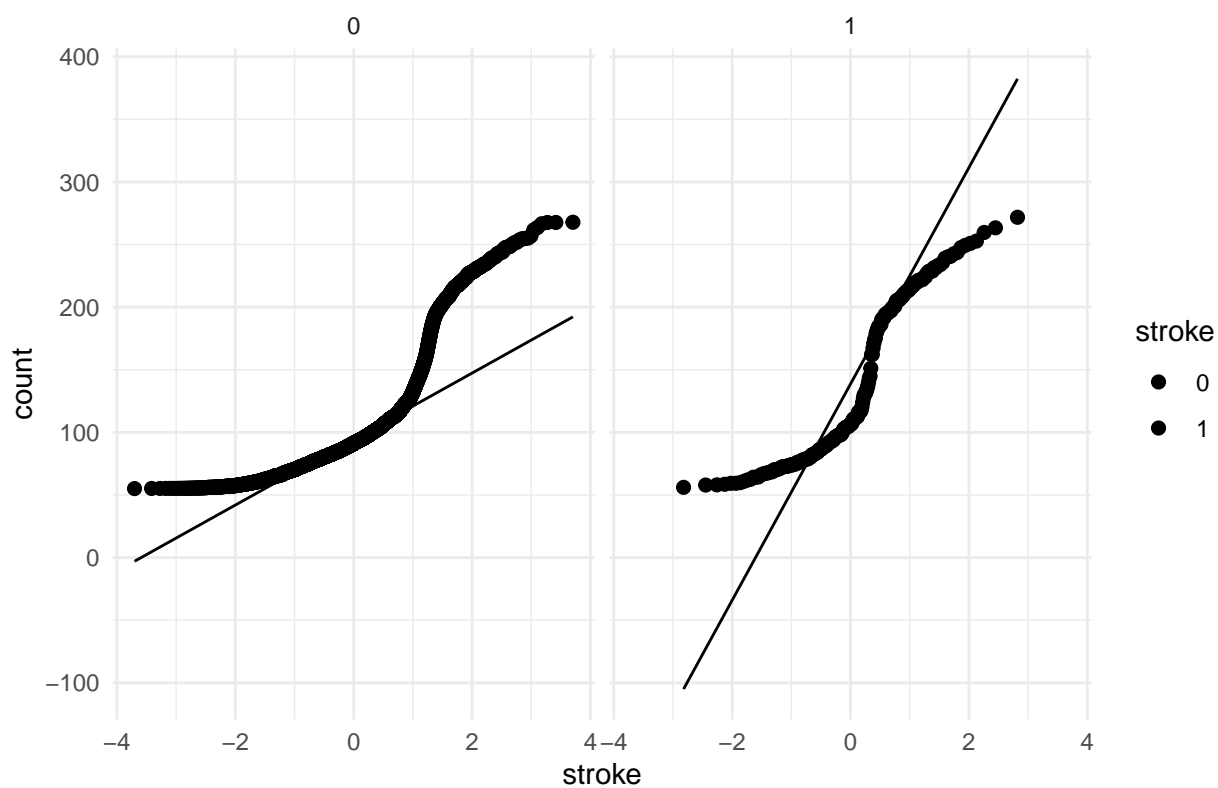
```
ggplot(data, aes(x = avg_glucose_level, fill = stroke)) +
  geom_histogram(binwidth = 3)+
  facet_wrap(~stroke)+theme_minimal()+labs(title = "avg. glucose histogram by stroke",x="avg. glucose",y="count")
```

avg. glucose histogram by stroke

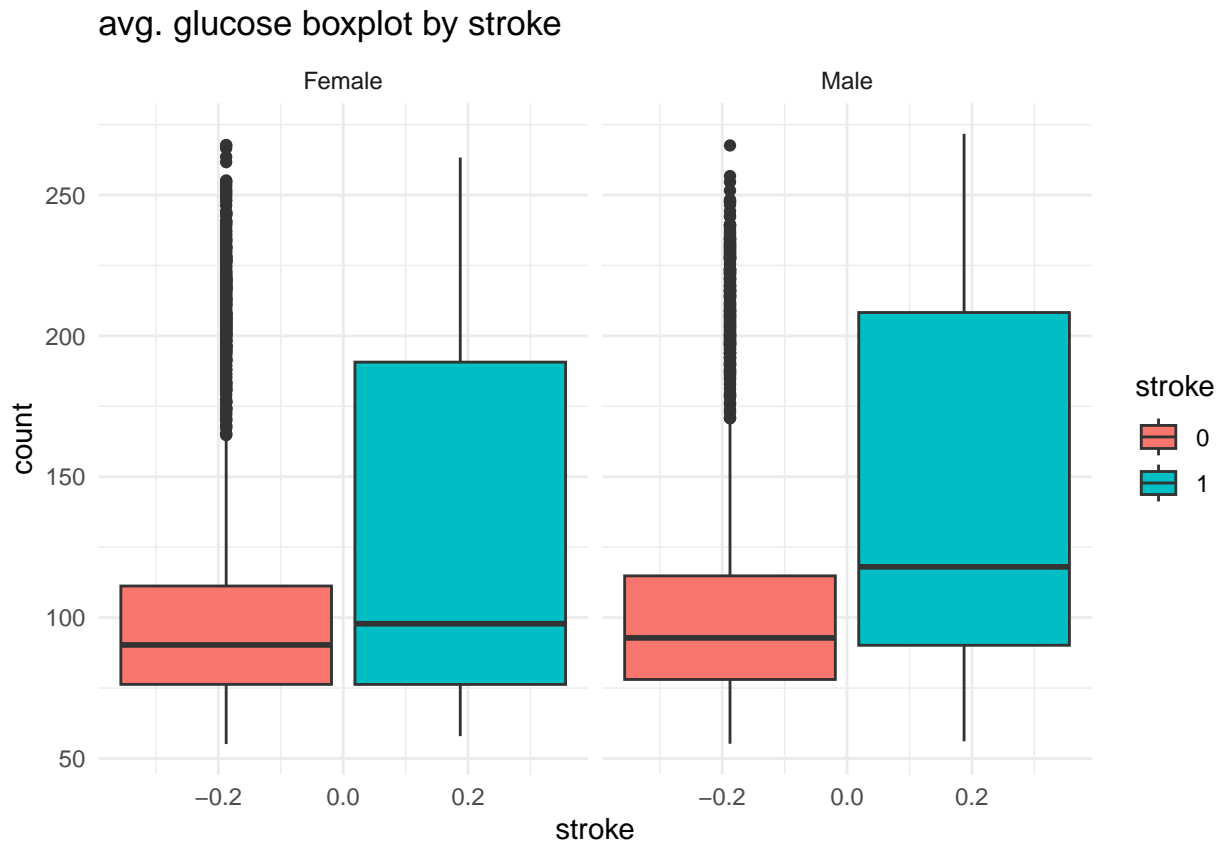


```
ggplot(data, aes(sample = avg_glucose_level, fill = stroke))+  
  stat_qq(size=2)+ stat_qq_line() +  
  facet_wrap(~stroke)+theme_minimal()+labs(title = "avg. glucose qqplot by stroke",x="stroke",y="count")
```

avg. glucose qqplot by stroke



```
ggplot(data, aes(y = avg_glucose_level, fill =stroke))+
  geom_boxplot()+
  facet_wrap(~gender)+theme_minimal()+labs(title = "avg. glucose boxplot by stroke",x="stroke",y="count")
```

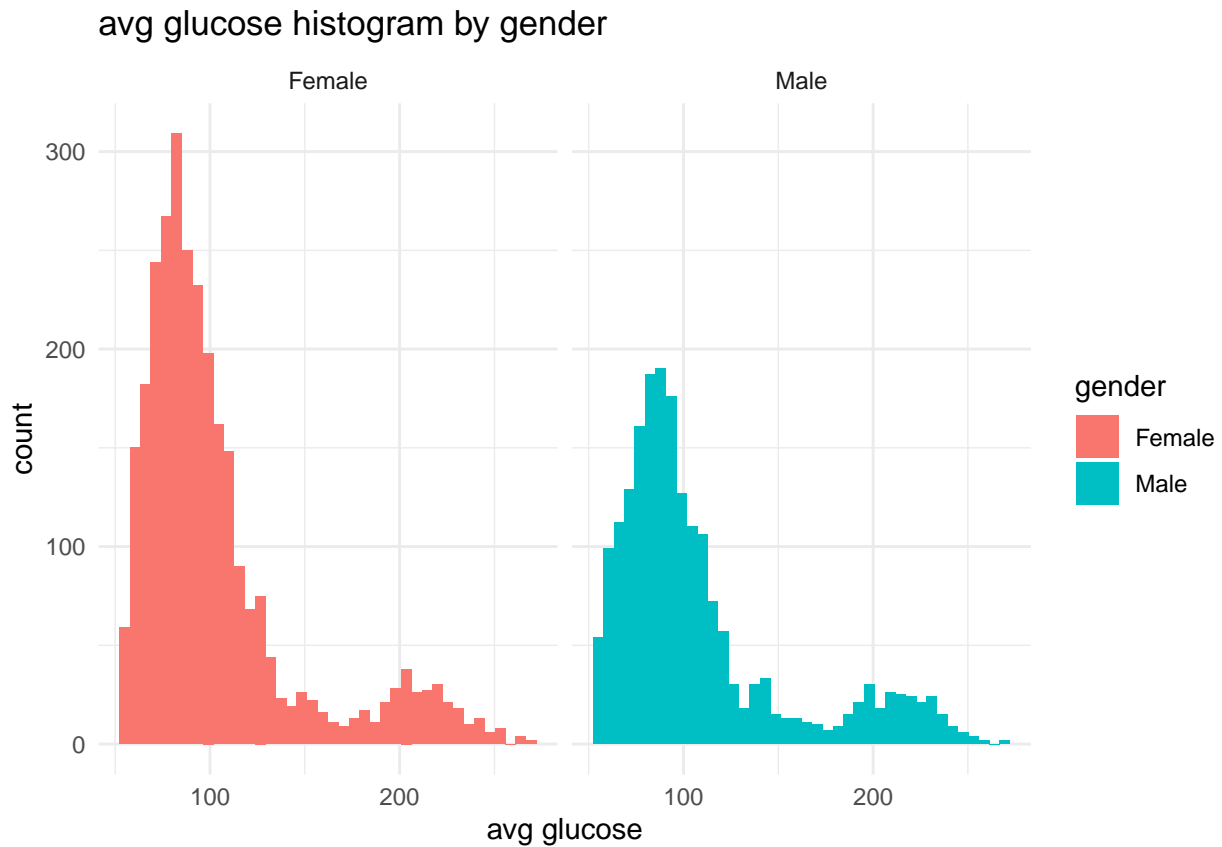


```
data_sub_glu_gd= data %>% group_by(gender,stroke) %>% summarise(mean = mean(avg_glucose_level),.groups = "drop")
print(data_sub_glu_gd)
```

### Avg. Glucose levels with gender

```
## # A tibble: 4 x 3
##   gender stroke mean
##   <fct> <fct> <dbl>
## 1 Female 0      102.
## 2 Female 1      126.
## 3 Male  0      106.
## 4 Male  1      146.
```

```
ggplot(data, aes(x = avg_glucose_level, fill = gender)) +
  geom_histogram(binwidth = 5.5)+
  facet_wrap(~gender)+theme_minimal()+labs(title = "avg glucose histogram by gender",x="avg glucose",y="count")
```

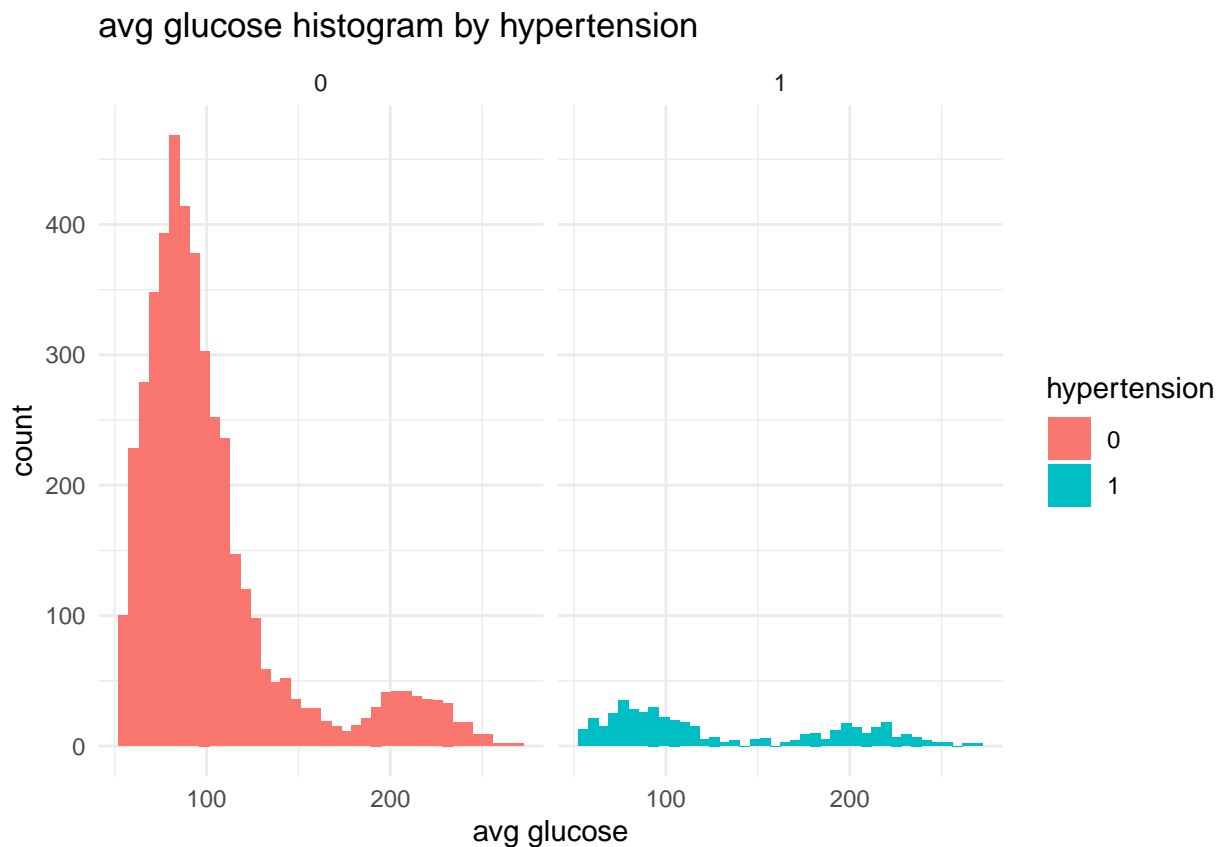


```
data_sub_glu_hyp= data %>% group_by(hypertension,stroke) %>% summarise(mean = mean(avg_glucose_level),...)
print(data_sub_glu_hyp)
```

### Glucose levels with hypertension

```
## # A tibble: 4 x 3
##   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(binwidth = 5.5)+
  facet_wrap(~hypertension)+theme_minimal()+labs(title = "avg glucose histogram by hypertension",x="avg
```



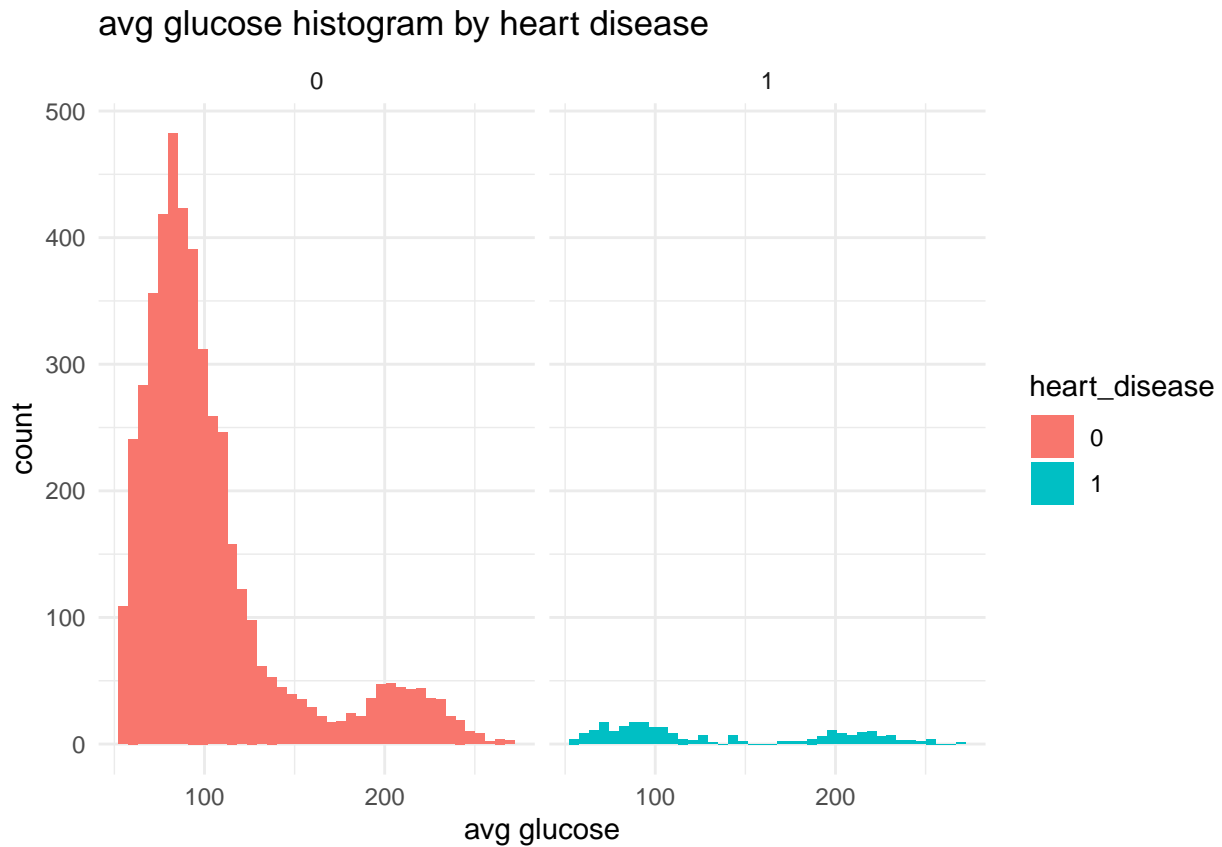
```
data_sub_glu_hd= data %>% group_by(heart_disease,stroke) %>% summarise(mean = mean(avg_glucose_level),...)
print(data_sub_glu_hd)
```

Avg. Glucose levels with heart disease

```
## # A tibble: 4 x 3
##   heart_disease stroke  mean
##   <fct>         <fct> <dbl>
## 1 0             0     103.
## 2 0             1     127.
## 3 1             0     129.
## 4 1             1     165.
```

```
ggplot(data, aes(x = avg_glucose_level, fill = heart_disease)) +
  geom_histogram(binwidth = 5.5)+
  facet_wrap(~heart_disease)+theme_minimal()+labs(title = "avg glucose histogram by heart disease",x="a
```



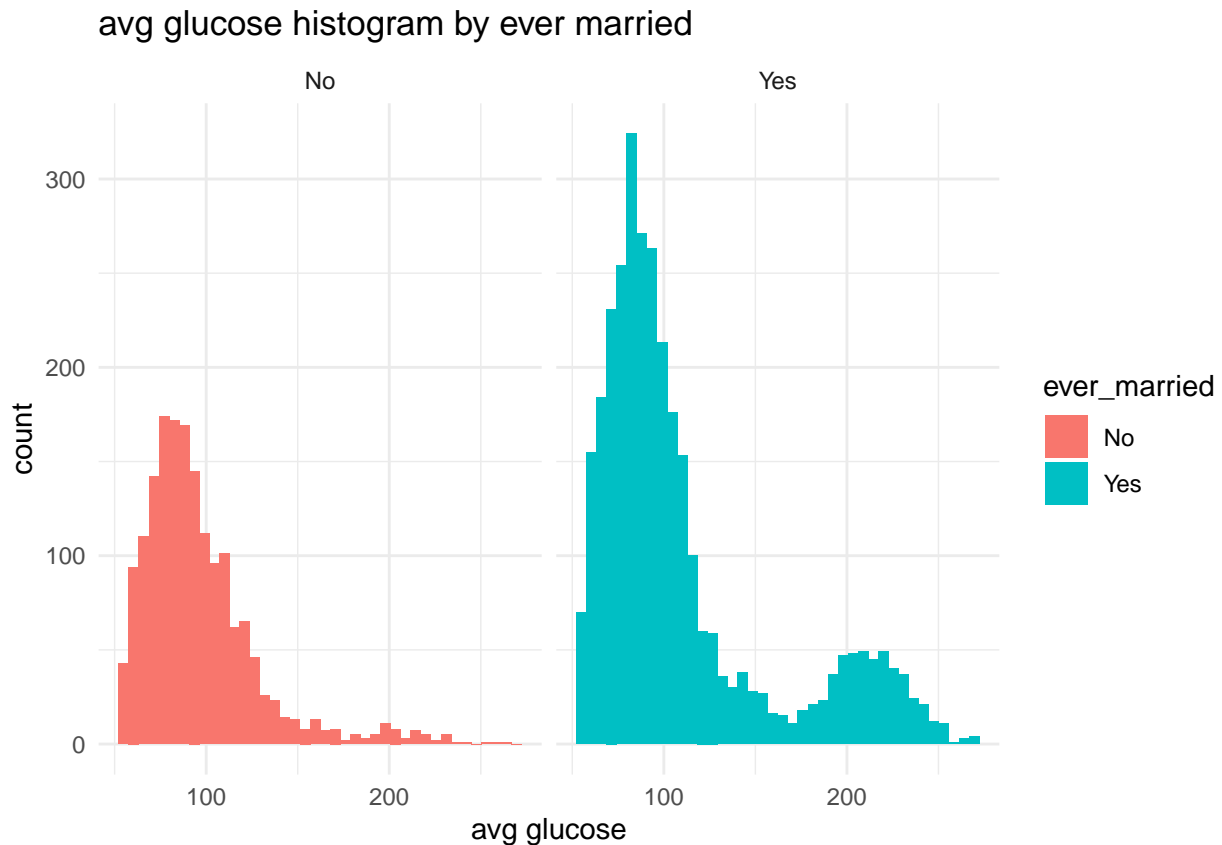


```
data_sub_glu_em= data %>% group_by(ever_married,stroke) %>% summarise(mean = mean(avg_glucose_level),.g
print(data_sub_glu_em)
```

Avg. Glucose levels with ever married

```
## # A tibble: 4 x 3
##   ever_married stroke   mean
##   <fct>         <fct> <dbl>
## 1 No          0      95.9
## 2 No          1     107.
## 3 Yes         0     108.
## 4 Yes         1     138.
```

```
ggplot(data, aes(x = avg_glucose_level, fill = ever_married)) +
  geom_histogram(binwidth = 5.5)+
  facet_wrap(~ever_married)+theme_minimal()+labs(title = "avg glucose histogram by ever married",x="avg
```

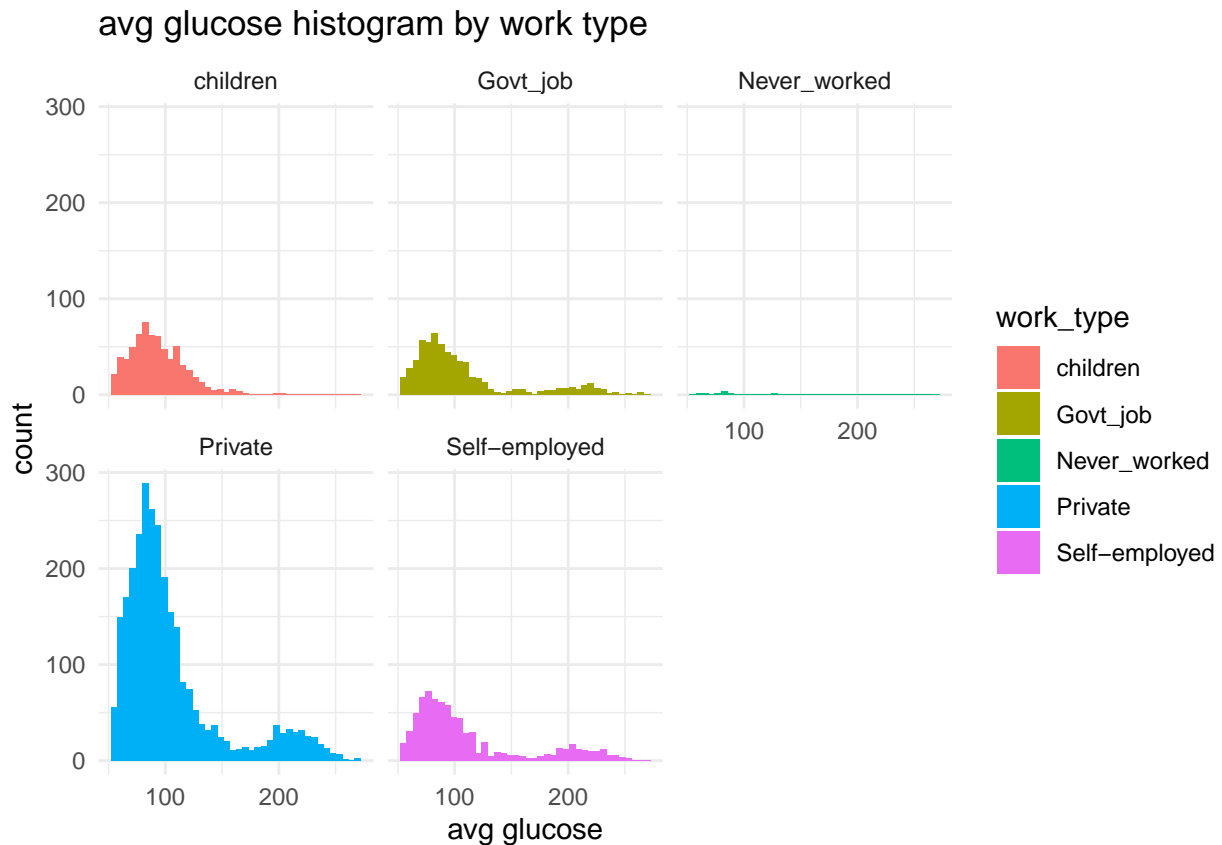


```
data_sub_glu_wt= data %>% group_by(work_type,stroke) %>% summarise(mean = mean(avg_glucose_level),.group
print(data_sub_glu_wt)
```

#### Avg. Glucose levels with work type

```
## # A tibble: 9 x 3
##   work_type   stroke mean
##   <fct>       <fct> <dbl>
## 1 children    0      94.1
## 2 children    1      57.9
## 3 Govt_job    0     106.
## 4 Govt_job    1     137.
## 5 Never_worked 0      96.0
## 6 Private     0     104.
## 7 Private     1     139.
## 8 Self-employed 0     112.
## 9 Self-employed 1     123.
```

```
ggplot(data, aes(x = avg_glucose_level, fill = work_type)) +
  geom_histogram(binwidth = 5.5)+
  facet_wrap(~work_type)+theme_minimal()+labs(title = "avg glucose histogram by work type",x="avg glucose")
```

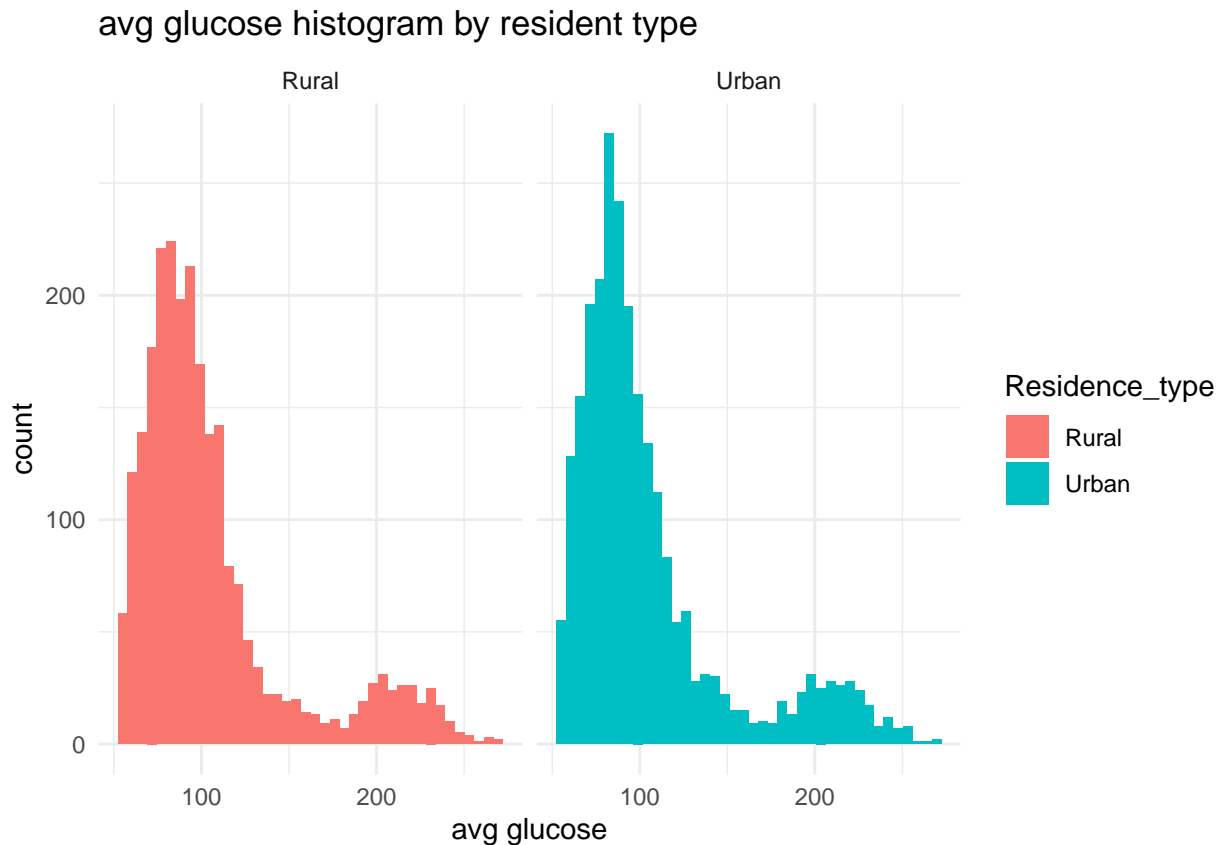


```
data_sub_glu_rt= data %>% group_by(Residence_type,stroke) %>% summarise(mean = mean(avg_glucose_level),
print(data_sub_glu_rt)
```

#### Avg. Glucose levels with resident type

```
## # A tibble: 4 x 3
##   Residence_type stroke mean
##   <fct>          <fct> <dbl>
## 1 Rural          0    104.
## 2 Rural          1    133.
## 3 Urban          0    104.
## 4 Urban          1    136.
```

```
ggplot(data, aes(x = avg_glucose_level, fill = Residence_type)) +
  geom_histogram(binwidth = 5.5)+
  facet_wrap(~Residence_type)+theme_minimal()+labs(title = "avg glucose histogram by resident type",x="")
```

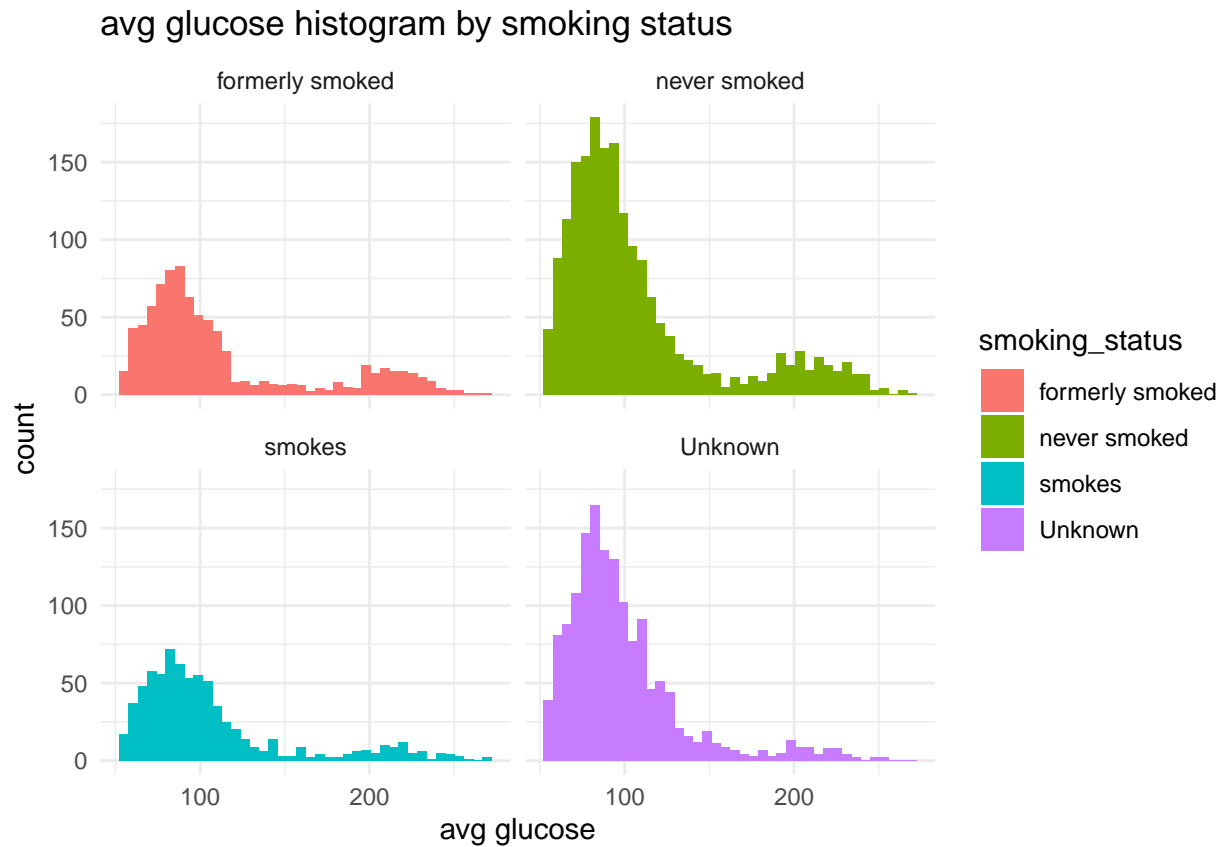


```
data_sub_glu_st= data %>% group_by(smoking_status,stroke) %>% summarise(mean = mean(avg_glucose_level),
print(data_sub_glu_st)
```

### Avg. Glucose levels with smoking status

```
## # A tibble: 8 x 3
##   smoking_status stroke mean
##   <fct>          <fct> <dbl>
## 1 formerly smoked 0     111.
## 2 formerly smoked 1     139.
## 3 never smoked   0     106.
## 4 never smoked   1     133.
## 5 smokes         0     105.
## 6 smokes         1     141.
## 7 Unknown        0      97.9
## 8 Unknown        1     121.
```

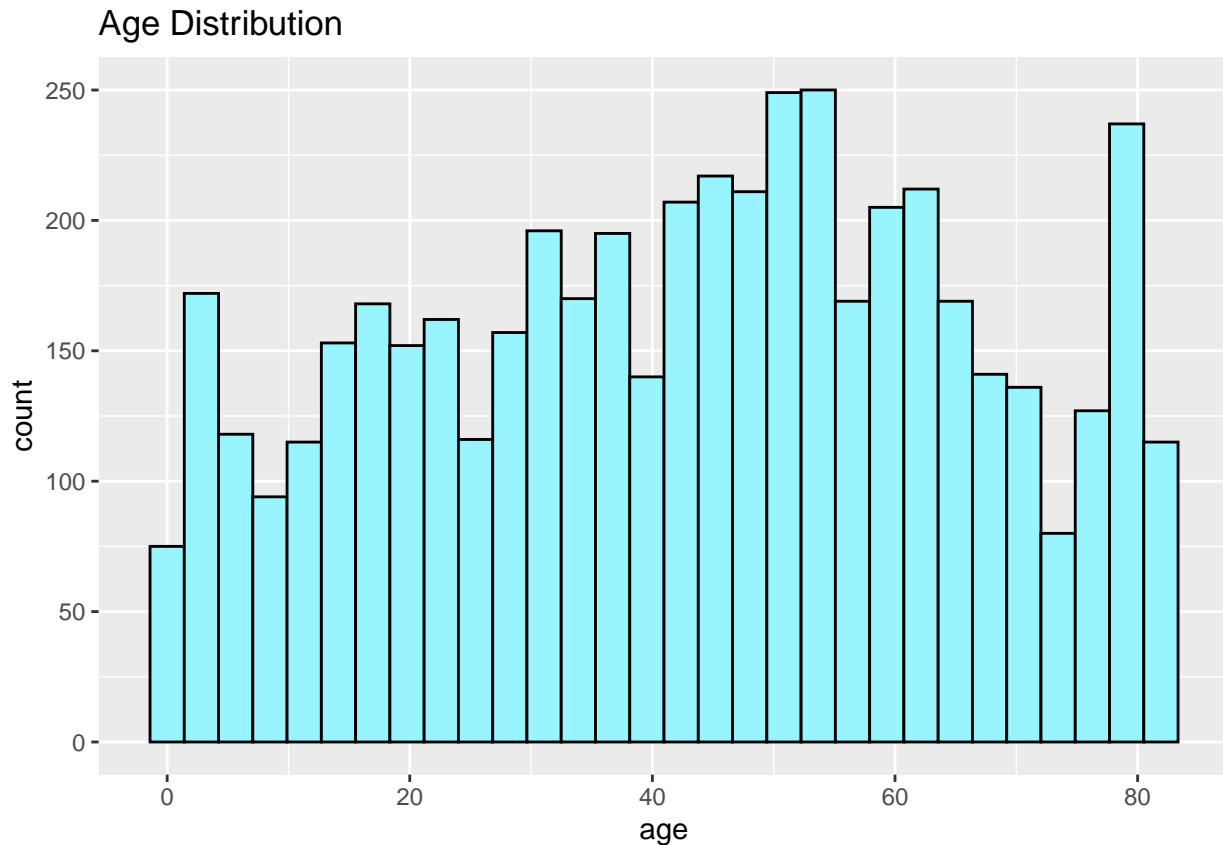
```
ggplot(data, aes(x = avg_glucose_level, fill = smoking_status)) +
  geom_histogram(binwidth = 5.5)+
  facet_wrap(~smoking_status)+theme_minimal()+labs(title = "avg glucose histogram by smoking status",x=
```



Age

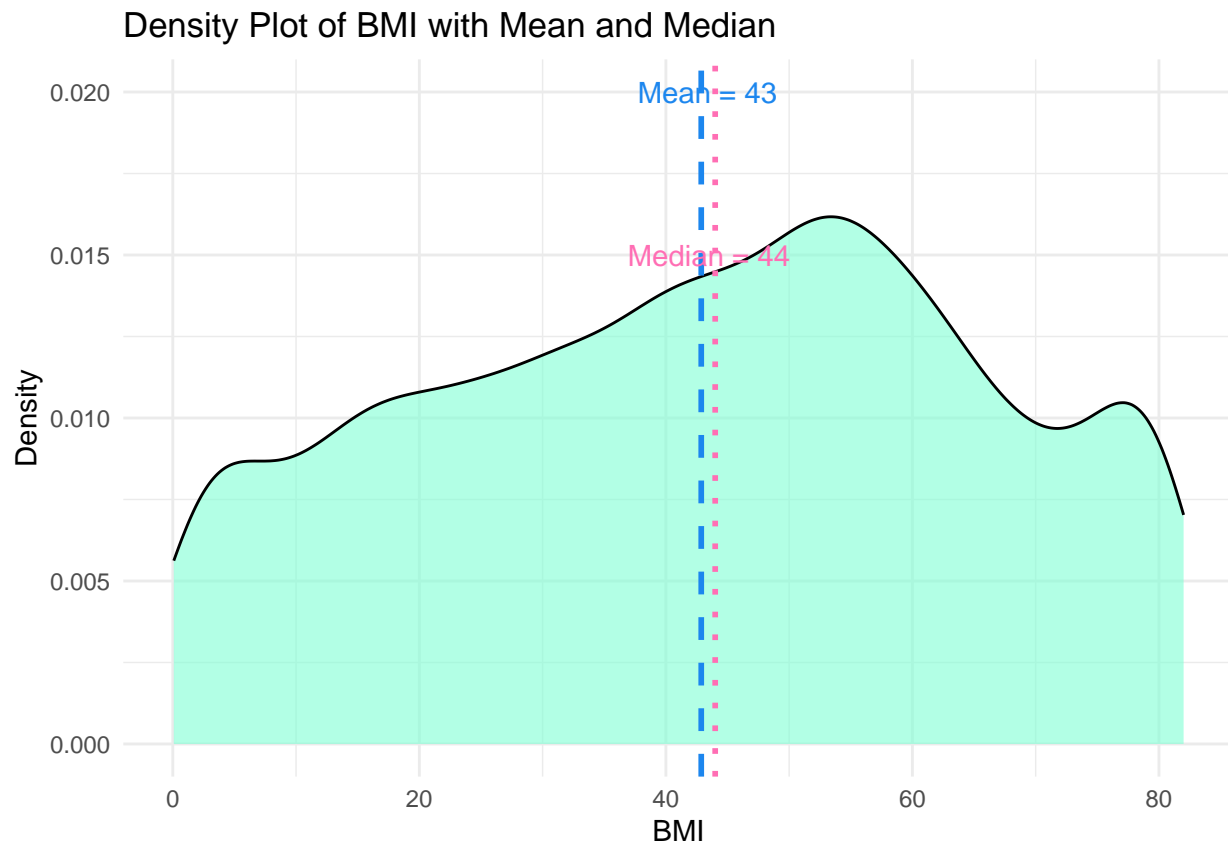
```
ggplot(data, aes(x=age)) +  
  geom_histogram(color="black", fill="cadetblue1")+  
  labs(title = "Age Distribution")
```

## 'stat\_bin()' using 'bins = 30'. Pick better value with 'binwidth'.



```
age_mean <- mean(data$age)
age_median <- median(data$age)

# Density plot with a vertical line at the mean
ggplot(data, aes(x = data$age)) +
  geom_density(fill = "aquamarine1", alpha = 0.6) +
  geom_vline(aes(xintercept = age_mean), color = "dodgerblue2", linetype = "dashed", size = 1) +
  geom_vline(aes(xintercept = age_median), color = "hotpink1", linetype = "dotted", size = 1) +
  labs(title = "Density Plot of BMI with Mean and Median", x = "BMI", y = "Density") +
  annotate("text", x = age_mean + 0.5, y = 0.02, label = paste("Mean =", round(age_mean, 0)), color = "dodgerblue2") +
  annotate("text", x = age_median - 0.5, y = 0.015, label = paste("Median =", round(age_median, 0)), color = "hotpink1") +
  theme_minimal()
```

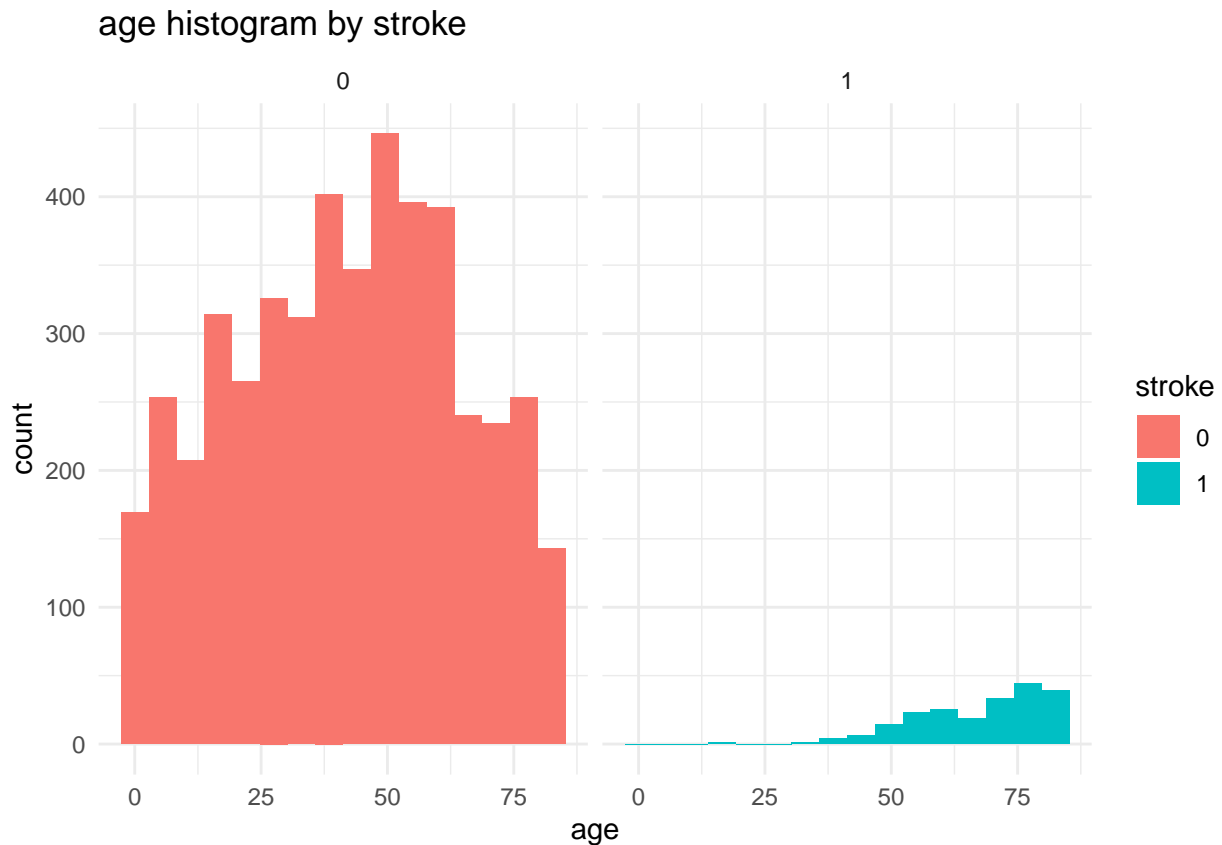


#### Age levels with stroke

```
data_sub_age= data %>% group_by(stroke) %>% summarise(mean = mean(age),.groups = "drop")
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(binwidth = 5.5)+
  facet_wrap(~stroke)+theme_minimal()+labs(title = "age histogram by stroke",x="age",y="count")
```



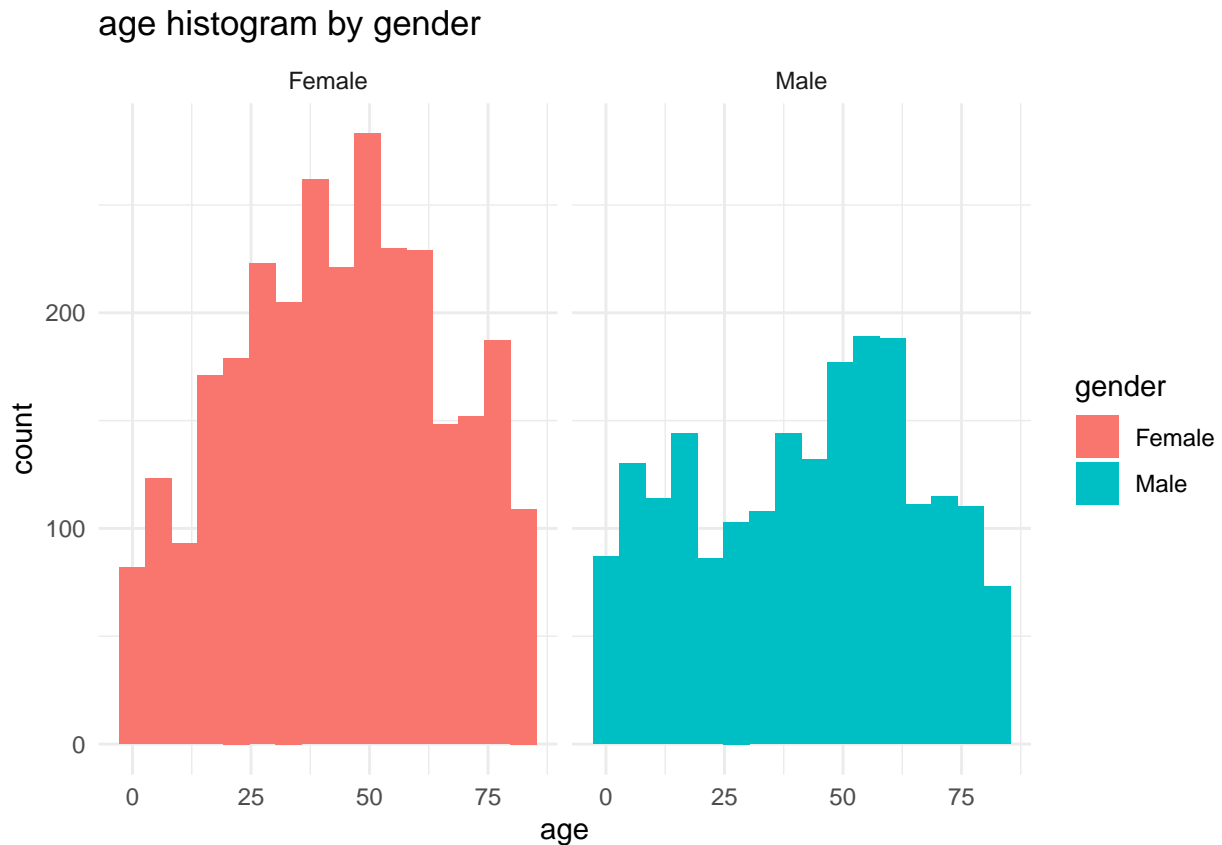
```
data_sub_age_gender = data %>% group_by(gender,stroke) %>% summarise(mean = mean(age),.groups = "drop")
print(data_sub_age_gender)
```

### Age levels with gender

```
## # A tibble: 4 x 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
```

```
ggplot(data, aes(x = age, fill = gender)) +
  geom_histogram(binwidth = 5.5)+
  facet_wrap(~gender)+theme_minimal()+labs(title = "age histogram by gender",x="age",y="count")
```





```
data_sub_age_hyp= data %>% group_by(hypertension,stroke) %>% summarise(mean = mean(age),.groups = "drop")
print(data_sub_age_hyp)
```

#### Age levels with hypertension

```
## # A tibble: 4 x 3
##   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(binwidth = 5.5)+
  facet_wrap(~hypertension)+theme_minimal()+labs(title = "age histogram by hypertension",x="age",y="count")
```

