

Stroke dataset

Florencia Luque and Seyed Amirhossein Mosaddad

October 6, 2024

```
##  
## 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  
  
##  
## Attaching package: 'agridcolae'  
  
## 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_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(stroke_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
##
## 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
##           Private     :2924
##           Self-employed: 819
##
##      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).

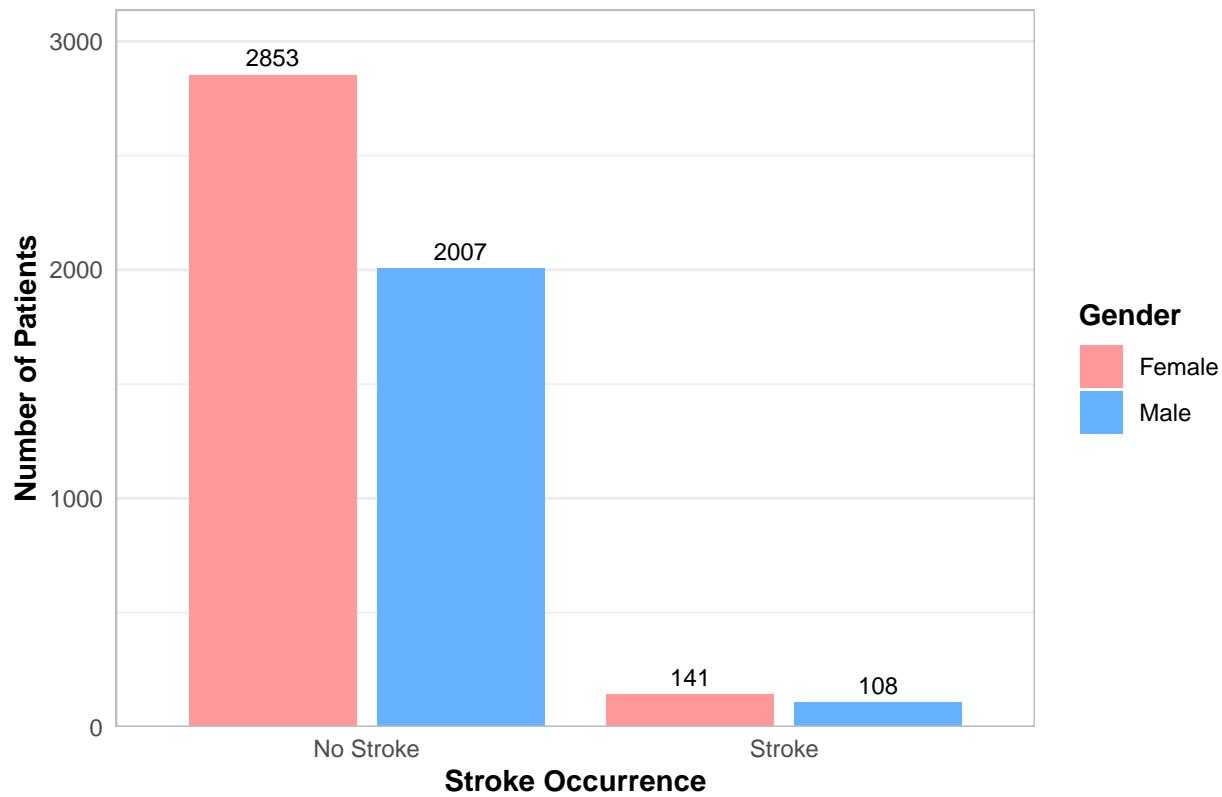
The next table shows a summary of the quantity of people getting a stroke by gender and the corresponding percentage.

```
stroke_gender = stroke_data %>% group_by(stroke,gender) %>% summarise(n = n(),.groups = "drop") %>% gro
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(stroke_data, aes(x = stroke, fill = gender)) +
  geom_bar(position = position_dodge(width = 0.9), width = 0.8) +
  geom_text(stat = 'count', aes(label = after_stat(count)),
            position = position_dodge(width = 0.9), vjust = -0.5, size = 3) +
  scale_x_discrete(labels = c("No Stroke", "Stroke")) +
  scale_fill_manual(values = c("Female" = "#FF9999", "Male" = "#66B2FF")) +
  labs(title = "Distribution of Stroke Cases by Gender",
       x = "Stroke Occurrence",
       y = "Number of Patients",
       fill = "Gender") +
  theme_minimal() +
  theme(
    plot.title = element_text(hjust = 0.5, face = "bold"),
    axis.title = element_text(face = "bold"),
    legend.title = element_text(face = "bold"),
    panel.grid.major.x = element_blank(),
    panel.border = element_rect(color = "grey", fill = NA, linewidth = 0.5) # Changed 'size' to 'linewidth'
  ) +
  scale_y_continuous(expand = expansion(mult = c(0, 0.1)))
```

Distribution of Stroke Cases by Gender



The majority of both female and male patients did not have a stroke, with 95.29% of females and 94.89% of males in the “no stroke” category. However, 4.71% of females and 5.11% of males did experience a stroke. As we can see there is a higher percentage of males that have had a stroke in the data. This number is slightly higher than the stroke rate for population.

Hypertension

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

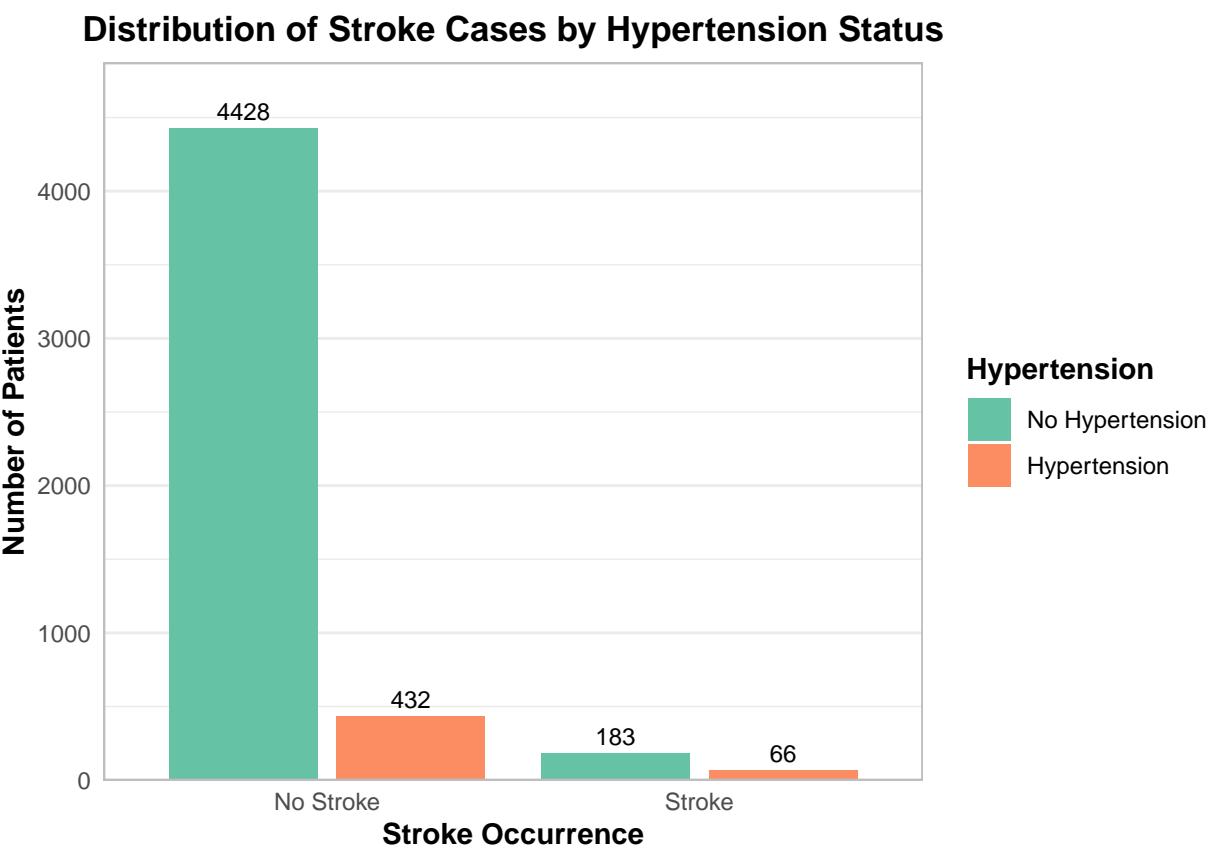
stroke	hypertension	n	percent
0	0	4428	96.03
0	1	432	86.75
1	0	183	3.969
1	1	66	13.25

```
ggplot(stroke_data, aes(x = factor(stroke), fill = factor(hypertension))) +
  geom_bar(position = position_dodge(width = 0.9), width = 0.8) +
  geom_text(stat = 'count', aes(label = after_stat(count)),
            position = position_dodge(width = 0.9), vjust = -0.5, size = 3) +
  scale_x_discrete(labels = c("No Stroke", "Stroke")) +
  scale_fill_manual(values = c("0" = "#66C2A5", "1" = "#FC8D62"),
```

```

    labels = c("No Hypertension", "Hypertension")) +
  labs(title = "Distribution of Stroke Cases by Hypertension Status",
       x = "Stroke Occurrence",
       y = "Number of Patients",
       fill = "Hypertension") +
  theme_minimal() +
  theme(
    plot.title = element_text(hjust = 0.5, face = "bold"),
    axis.title = element_text(face = "bold"),
    legend.title = element_text(face = "bold"),
    panel.grid.major.x = element_blank(),
    panel.border = element_rect(color = "grey", fill = NA, linewidth = 0.5)
  ) +
  scale_y_continuous(expand = expansion(mult = c(0, 0.1)))

```



The stroke rate among hypertensive individuals (13.25%) is more than three times higher than among non-hypertensive individuals (3.969%). This difference suggests a strong correlation between hypertension and increased stroke risk. In addition, The hypertensive group's stroke rate (13.25%) is above the overall population rate (4.8%). The non-hypertensive group's rate (3.969%) is slightly below the overall rate. This suggests that Having hypertension is associated with an increase in stroke risk compared to not having hypertension.

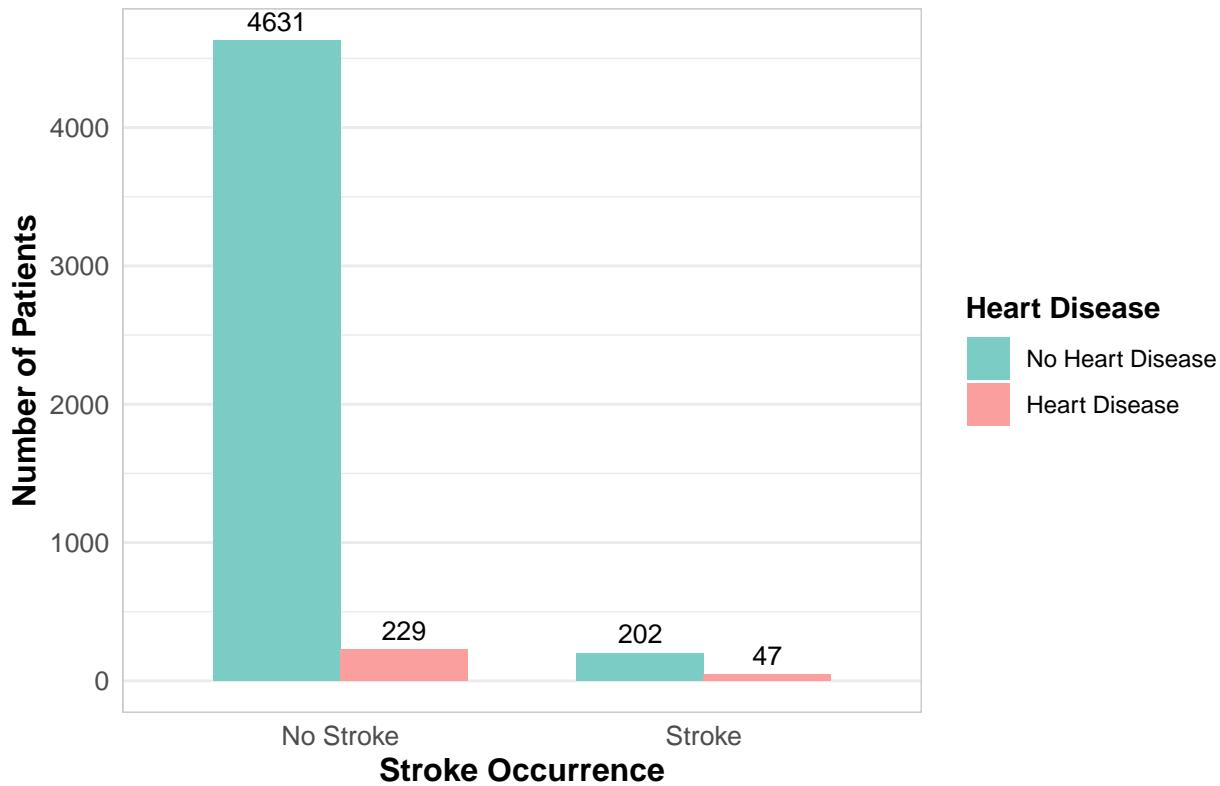
Heart disease

```
pander(stroke_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(stroke_data, aes(x = stroke, y = ..count.., fill = heart_disease)) +  
  geom_bar(position = "dodge", width = 0.7) +  
  scale_fill_manual(values = c("0" = "#7BCCC4", "1" = "#FA9E9E"), labels = c("No Heart Disease", "Heart  
  Disease"))  
  labs(  
    title = "Stroke Occurrence by Heart Disease Status",  
    x = "Stroke Occurrence",  
    y = "Number of Patients",  
    fill = "Heart Disease")  
  ) +  
  theme_minimal() +  
  theme(  
    plot.title = element_text(hjust = 0.5, face = "bold", size = 16),  
    axis.title = element_text(face = "bold", size = 12),  
    axis.text = element_text(size = 10),  
    legend.title = element_text(face = "bold"),  
    legend.position = "right",  
    panel.grid.major.x = element_blank(),  
    panel.border = element_rect(fill = NA, color = "gray80")  
  ) +  
  geom_text(stat = "count",  
    aes(label = ..count..),  
    position = position_dodge(width = 0.7),  
    vjust = -0.5,  
    size = 3.5)  
  
## Warning: The dot-dot notation ('..count..') was deprecated in ggplot2 3.4.0.  
## i Please use 'after_stat(count)' instead.  
## This warning is displayed once every 8 hours.  
## Call 'lifecycle::last_lifecycle_warnings()' to see where this warning was  
## generated.
```

Stroke Occurrence by Heart Disease Status



The data of the people who have a heart disease look even more likely to get a stroke than the ones who have hypertension. There can be a correlation between hypertension and having a heart disease. The majority of patients without heart disease (95.82%) did not experience a stroke, whereas those with heart disease have a lower percentage (82.97%) of avoiding a stroke. Among the patients who had a stroke, those with heart disease were more likely to experience one (17.03%) compared to those without heart disease (4.18%).

Ever Married

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

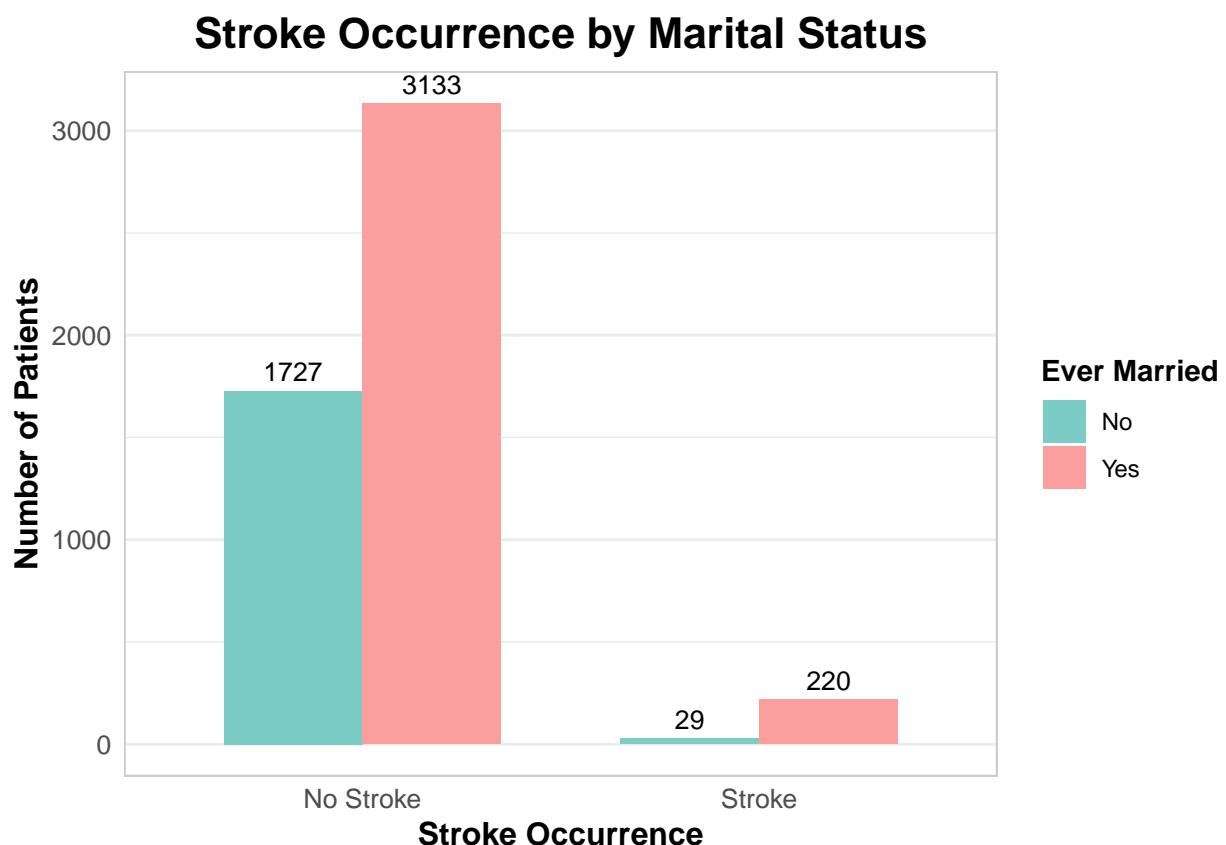
stroke	ever_married	n	percent
0	No	1727	98.35
0	Yes	3133	93.44
1	No	29	1.651
1	Yes	220	6.561

```
ggplot(stroke_data, aes(x = factor(stroke), y = ..count.., fill = ever_married)) +  
geom_bar(position = "dodge", width = 0.7) +  
scale_fill_manual(values = c("No" = "#7BCCC4", "Yes" = "#FA9E9E")) +  
scale_x_discrete(labels = c("0" = "No Stroke", "1" = "Stroke")) +  
labs(
```

```

title = "Stroke Occurrence by Marital Status",
x = "Stroke Occurrence",
y = "Number of Patients",
fill = "Ever Married"
) +
theme_minimal() +
theme(
  plot.title = element_text(hjust = 0.5, face = "bold", size = 16),
  axis.title = element_text(face = "bold", size = 12),
  axis.text = element_text(size = 10),
  legend.title = element_text(face = "bold"),
  legend.position = "right",
  panel.grid.major.x = element_blank(),
  panel.border = element_rect(fill = NA, color = "gray80")
) +
geom_text(stat = "count",
  aes(label = ..count..),
  position = position_dodge(width = 0.7),
  vjust = -0.5,
  size = 3.5)

```



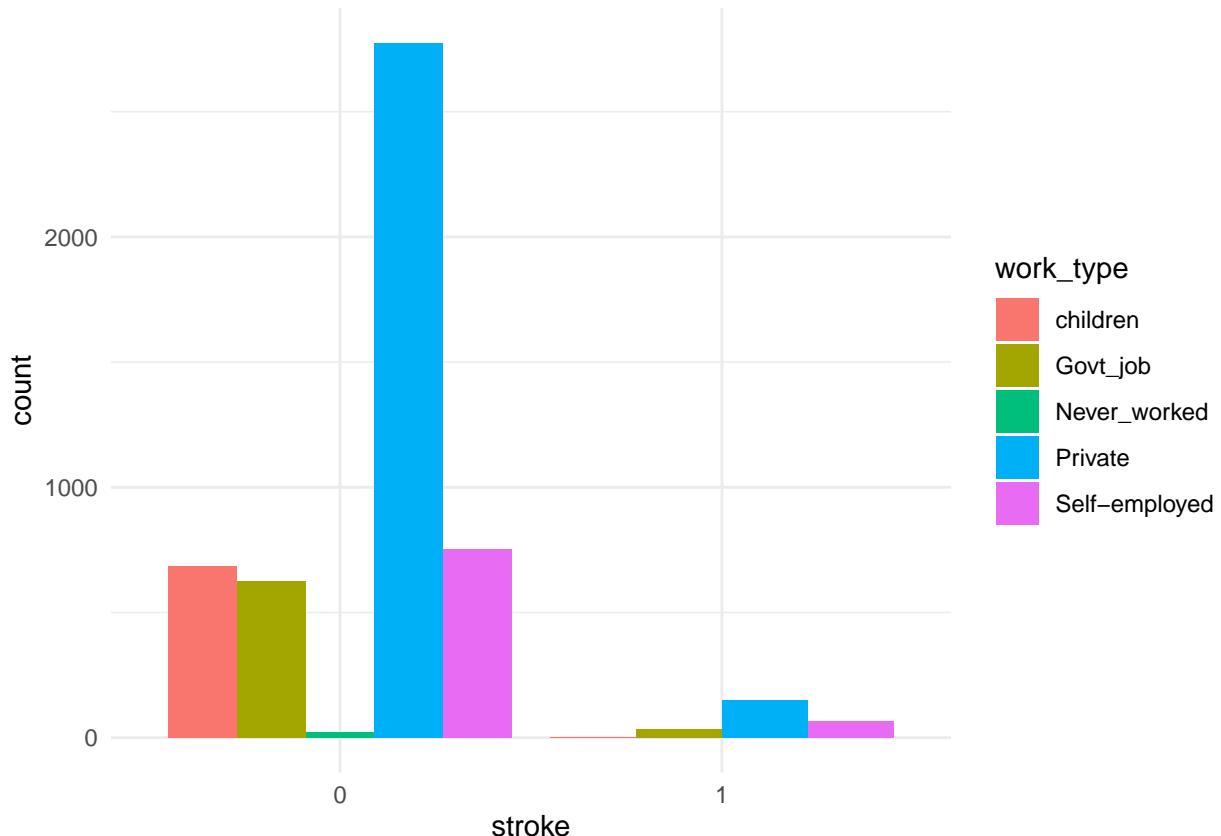
There are more people who are ever married and had a stroke than not ever married. Among individuals who have never been married, 98.35% did not experience a stroke, while 1.65% did. In contrast, among those who have been married, 93.44% did not have a stroke, and 6.56% did. This suggests that those who have ever been married are more likely to experience a stroke compared to those who have never been married.

Work Type

```
pander(stroke_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(stroke_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 han 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.

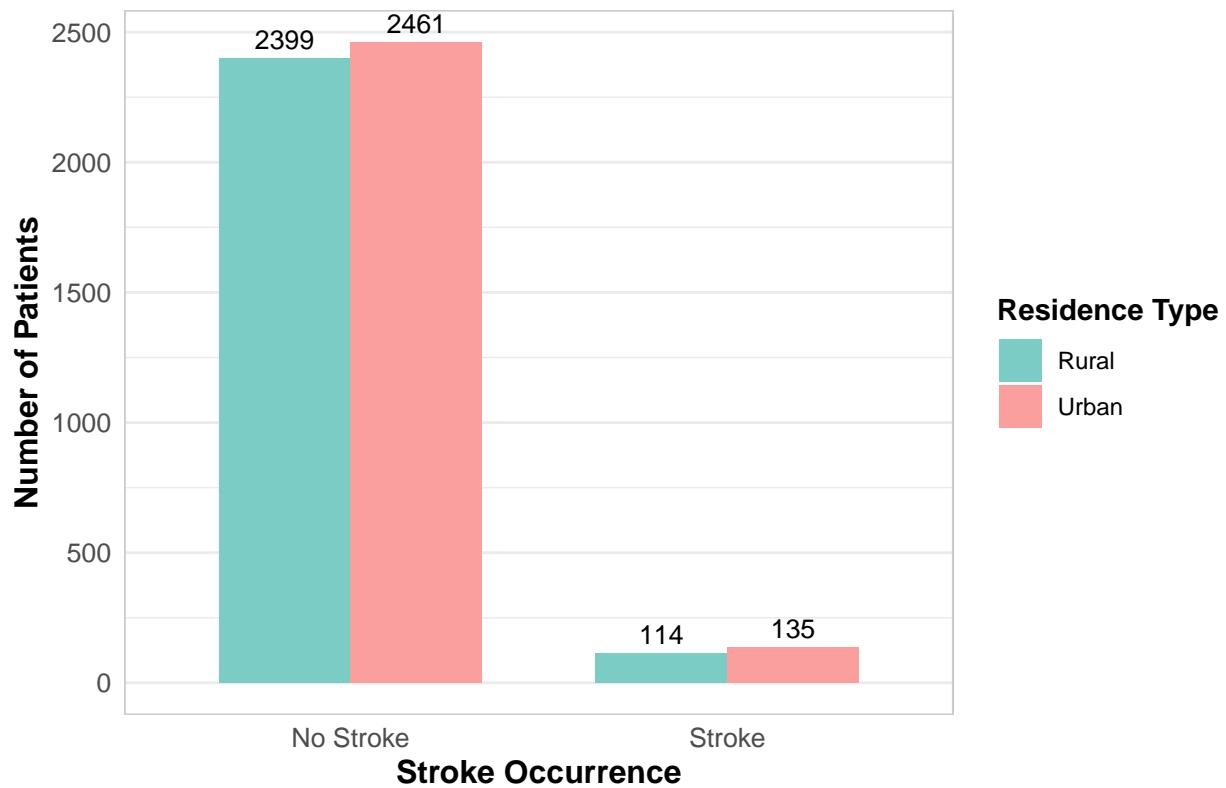
Resident Type

```
pander(stroke_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(stroke_data, aes(x = factor(stroke), y = ..count.., fill = Residence_type)) +  
  geom_bar(position = "dodge", width = 0.7) +  
  scale_fill_manual(values = c("Rural" = "#7BCCC4", "Urban" = "#FA9E9E")) +  
  scale_x_discrete(labels = c("0" = "No Stroke", "1" = "Stroke")) +  
  labs(  
    title = "Stroke Occurrence by Residence Type",  
    x = "Stroke Occurrence",  
    y = "Number of Patients",  
    fill = "Residence Type"  
) +  
  theme_minimal() +  
  theme(  
    plot.title = element_text(hjust = 0.5, face = "bold", size = 16),  
    axis.title = element_text(face = "bold", size = 12),  
    axis.text = element_text(size = 10),  
    legend.title = element_text(face = "bold"),  
    legend.position = "right",  
    panel.grid.major.x = element_blank(),  
    panel.border = element_rect(fill = NA, color = "gray80")  
) +  
  geom_text(stat = "count",  
    aes(label = ..count..),  
    position = position_dodge(width = 0.7),  
    vjust = -0.5,  
    size = 3.5)
```

Stroke Occurrence by Residence Type



The part of the city that one lives looks like it might not have much impact because the percentage of stroke is very similar to the population and almost the same between the types of resident. A slightly higher percentage of urban residents (5.2%) experienced a stroke compared to rural residents (4.54%). In both groups, the majority of individuals did not experience a stroke, with 95.46% of rural residents and 94.8% of urban residents falling into the “no stroke” category. The data suggest that while stroke occurrence is relatively similar across residence types.

Smoking Status

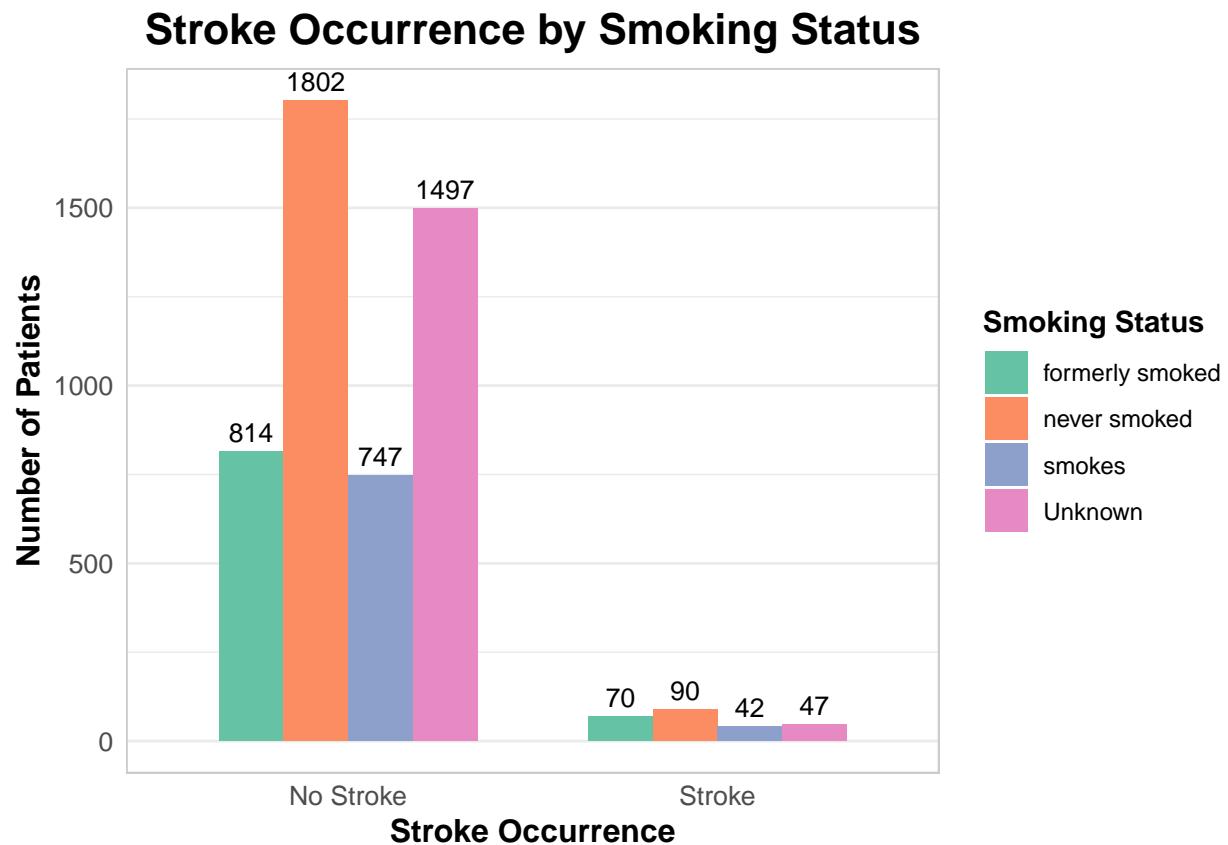
```
pander(stroke_data %>% group_by(stroke,smoking_status) %>% summarise(n = n(),.groups = "drop")%>%  
group_by(smoking_status) %>% mutate(percent = n/sum(n)*100))
```

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(stroke_data, aes(x = factor(stroke), y = ..count.., fill = smoking_status)) +
  geom_bar(position = "dodge", width = 0.7) +
  scale_fill_brewer(palette = "Set2") +
  scale_x_discrete(labels = c("0" = "No Stroke", "1" = "Stroke")) +
  labs(
    title = "Stroke Occurrence by Smoking Status",
    x = "Stroke Occurrence",
    y = "Number of Patients",
    fill = "Smoking Status"
  ) +
  theme_minimal() +
  theme(
    plot.title = element_text(hjust = 0.5, face = "bold", size = 16),
    axis.title = element_text(face = "bold", size = 12),
    axis.text = element_text(size = 10),
    legend.title = element_text(face = "bold"),
    legend.position = "right",
    panel.grid.major.x = element_blank(),
    panel.border = element_rect(fill = NA, color = "gray80")
  ) +
  geom_text(stat = "count",
            aes(label = ..count..),
            position = position_dodge(width = 0.7),
            vjust = -0.5,
            size = 3.5)

```



The people who formerly smoked have 3% higher than the population. Those who formerly smoked have the highest stroke rate (7.92%), while current smokers have a stroke rate of 5.32%. People who have never smoked have a lower stroke rate (4.76%), and those with unknown smoking status have the lowest stroke rate (3.04%). The data suggest that individuals who have smoked at some point, either formerly or currently, are at a higher risk for stroke compared to those who have never smoked or whose smoking status is unknown. However, the highest number of patients is in the “never smoked” group.

Numeric Variables

BMI

The first continuous 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(stroke_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.

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

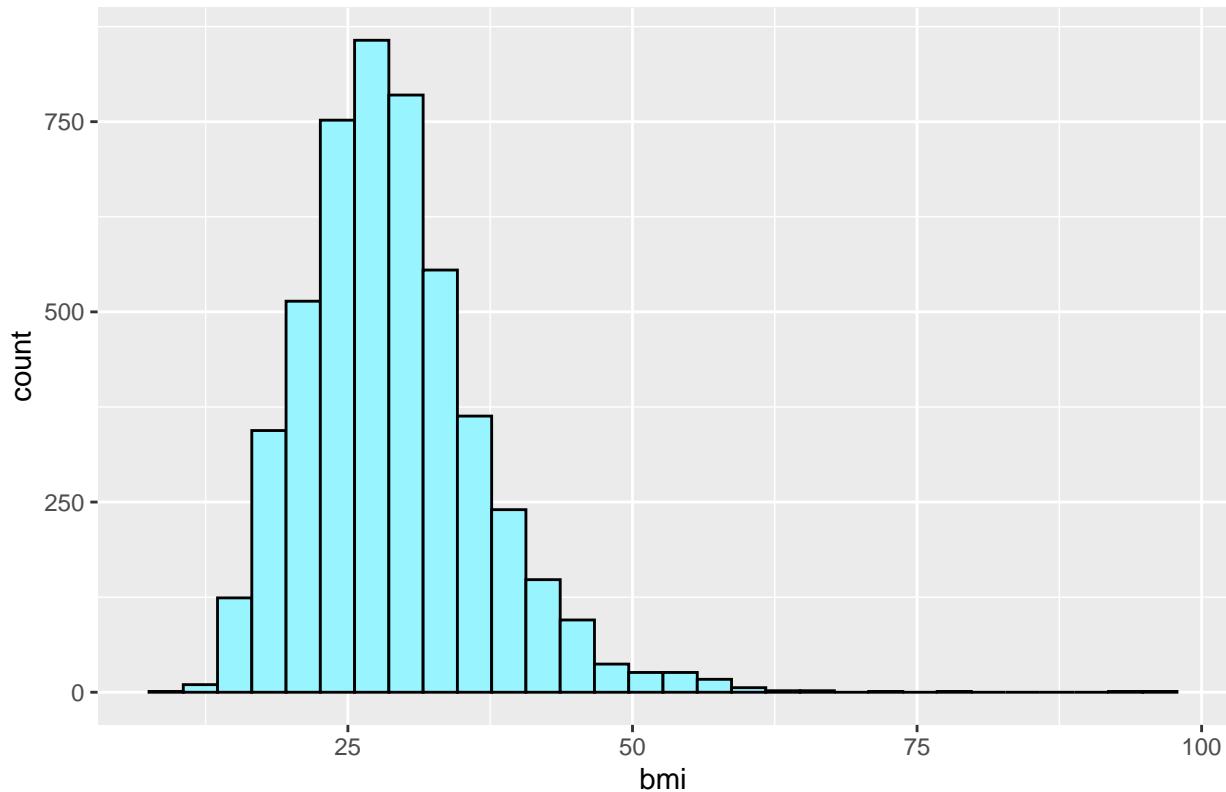
## $variance
## [1] 63.33528
##
## $mean
## [1] 28.84576
##
## $median
## [1] 28.00923
##
## $mode
##      [- -] mode
## [1,] 25 30 27.22513

ggplot(stroke_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

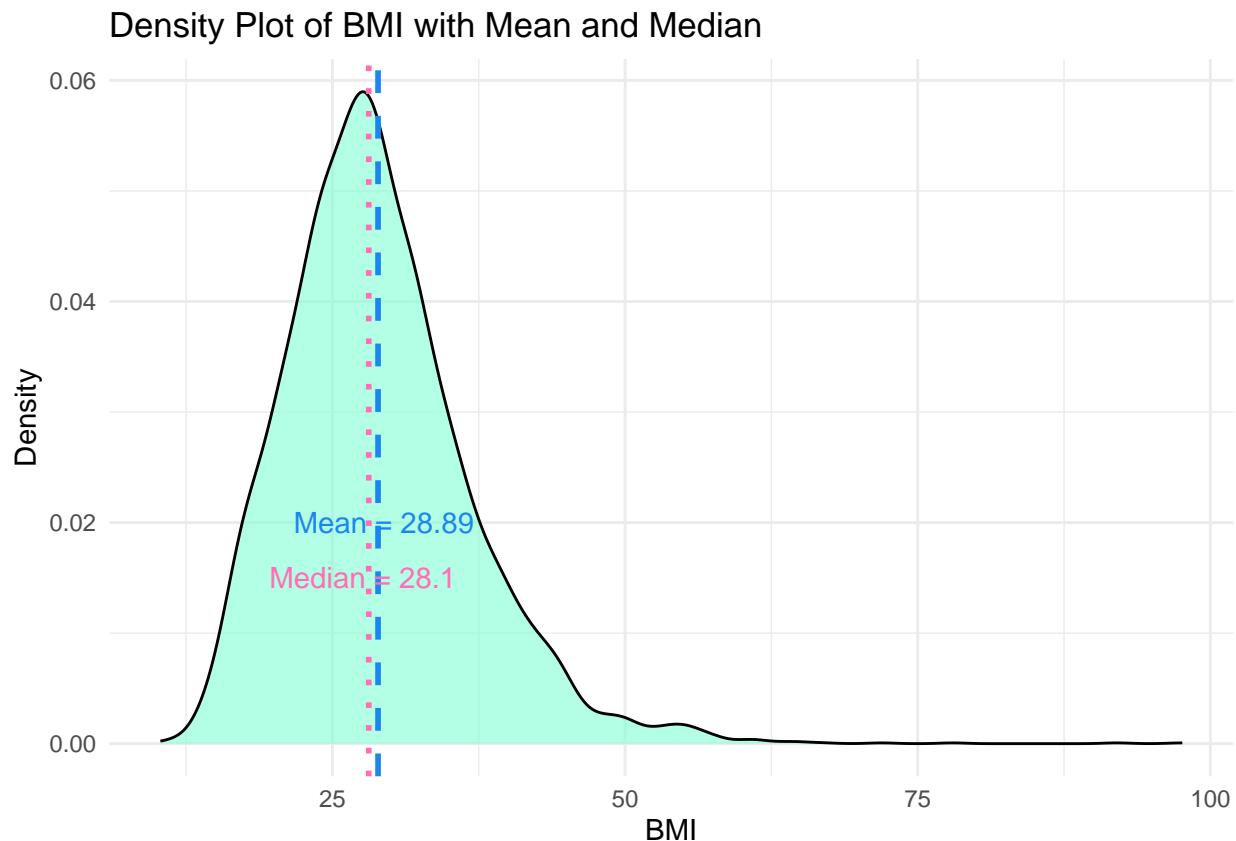


```

bmi_mean <- mean(stroke_data$bmi)
bmi_median <- median(stroke_data$bmi)

```

```
# Density plot with a vertical line at the mean
ggplot(stroke_data, aes(x = stroke_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(stroke_data$bmi)
```

```
## [1] 1.055063
```

```
kurtosis(stroke_data$bmi)
```

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

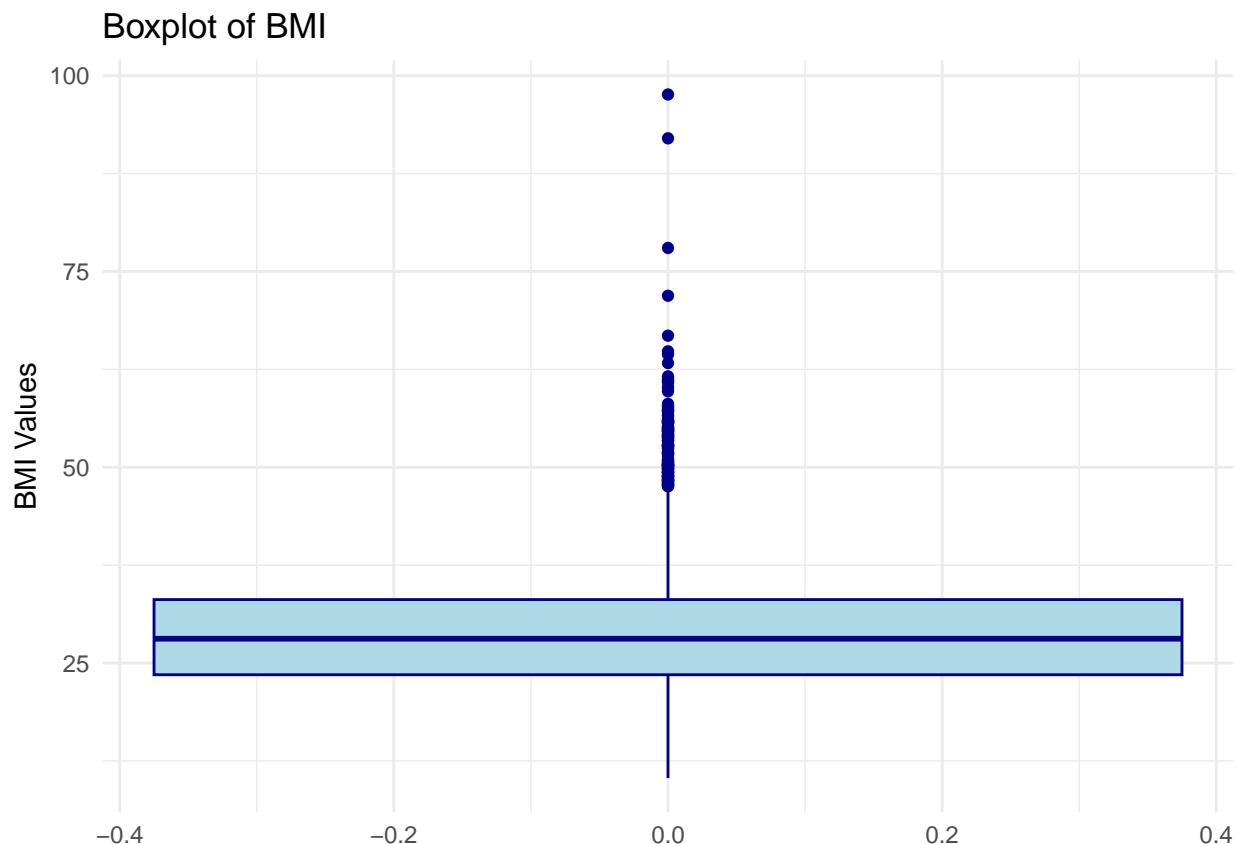
The values of the shape tell us that this is not a normal distribution. There's a lot of people in the center of the data but the quantity of people with extremely high BMI that change the weight of the tail in the distribution.

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 tell us that the the distribution is skewed to the right and that the heavier tails 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*

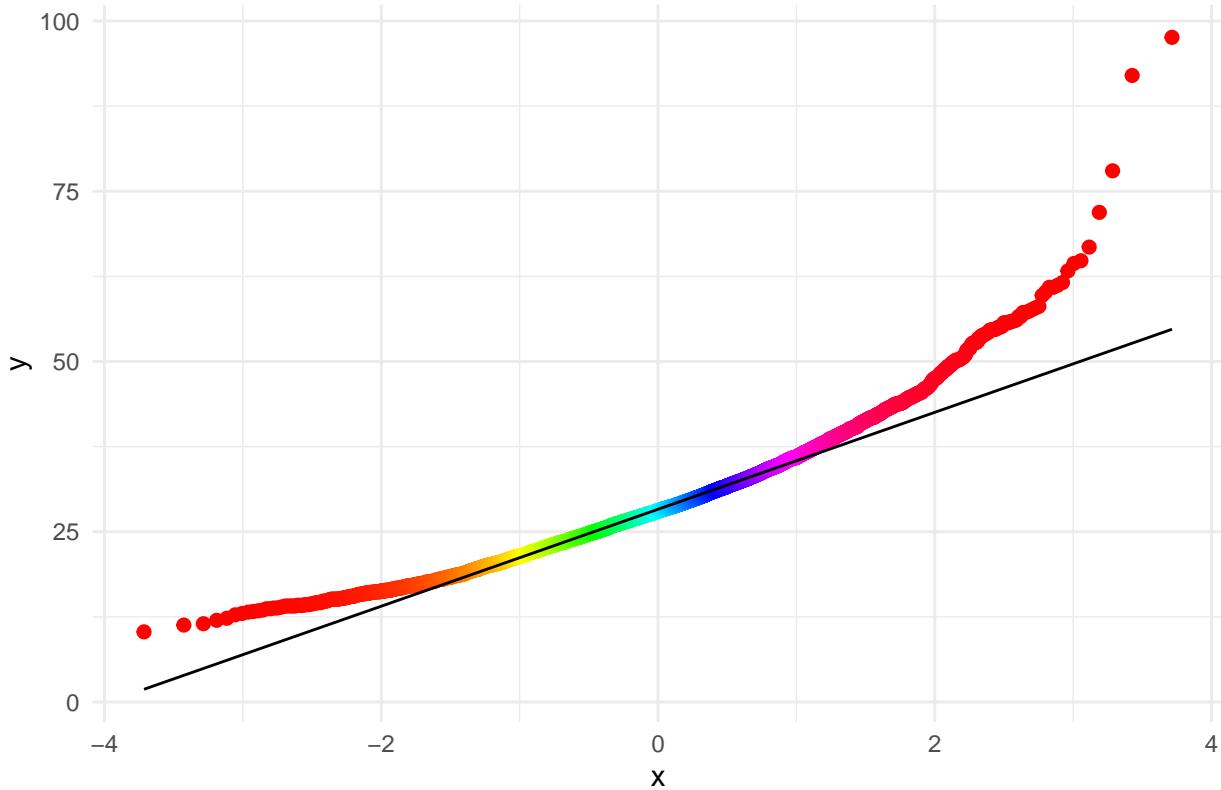
```
ggplot(stroke_data, aes(y = stroke_data$bmi)) +  
  geom_boxplot(fill = "lightblue", color = "darkblue") + # Change box and border color  
  labs(title = "Boxplot of BMI", y = "BMI Values") + # Custom title and labels  
  theme_minimal()
```



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

QQ plot for bmi



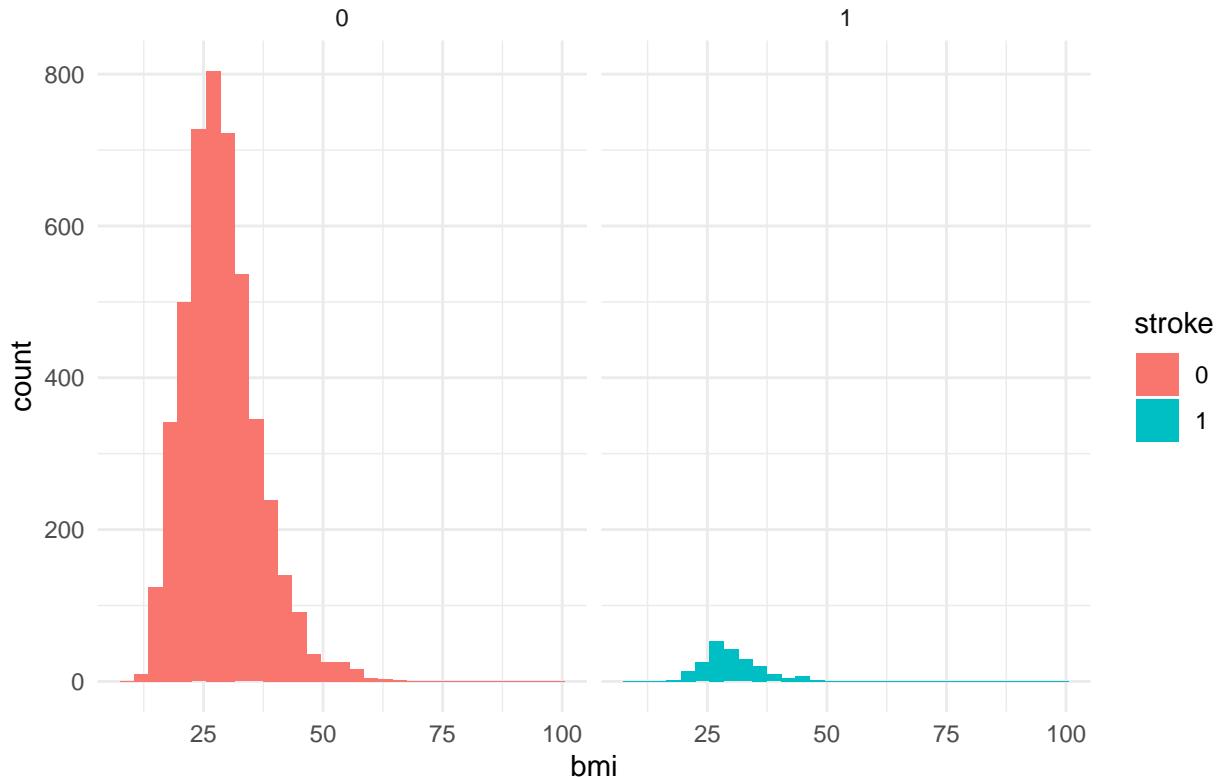
```
data_sub_bmi_catw= stroke_data %>% group_by(cat_weight,stroke) %>% summarise(mean = mean(bmi),.groups = print(data_sub_bmi_catw)
```

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
```

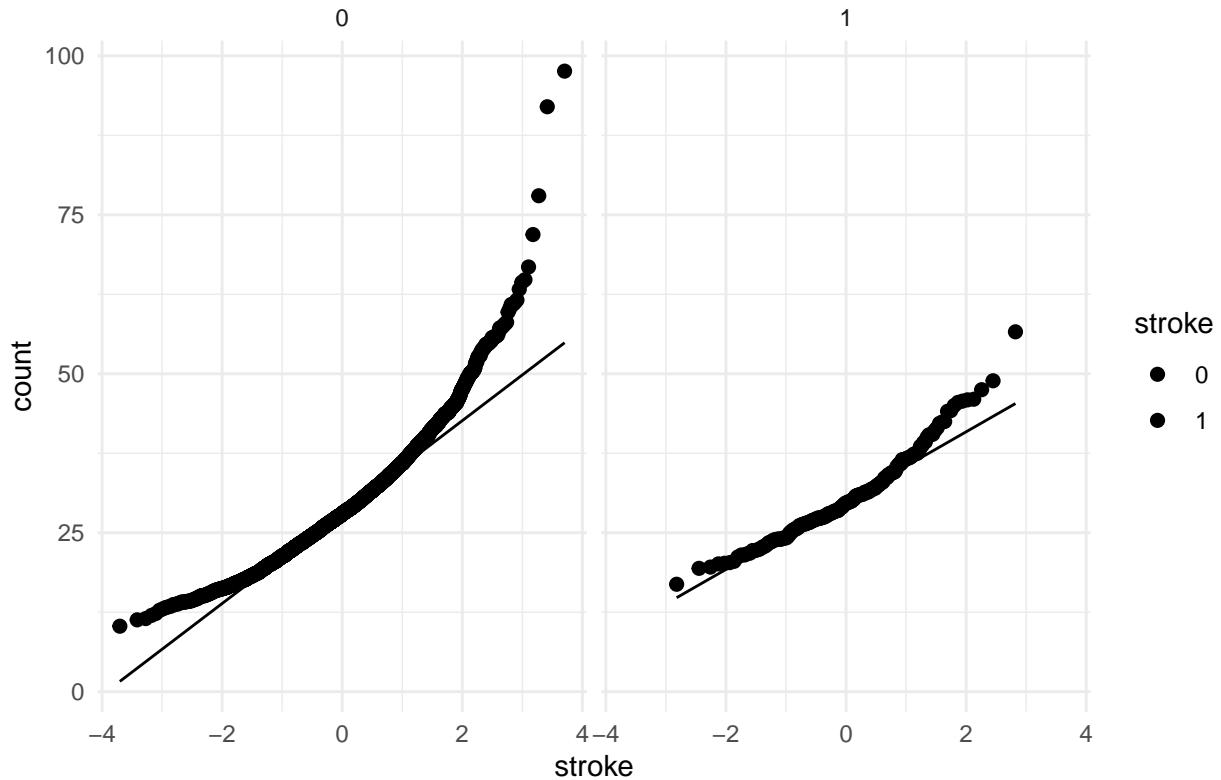
```
ggplot(stroke_data, aes(x = bmi, fill = stroke)) +
  geom_histogram(binwidth = 3) +
  facet_wrap(~stroke)+theme_minimal()+labs(title = "bmi histogram by stroke",x="bmi",y="count")
```

bmi histogram by stroke



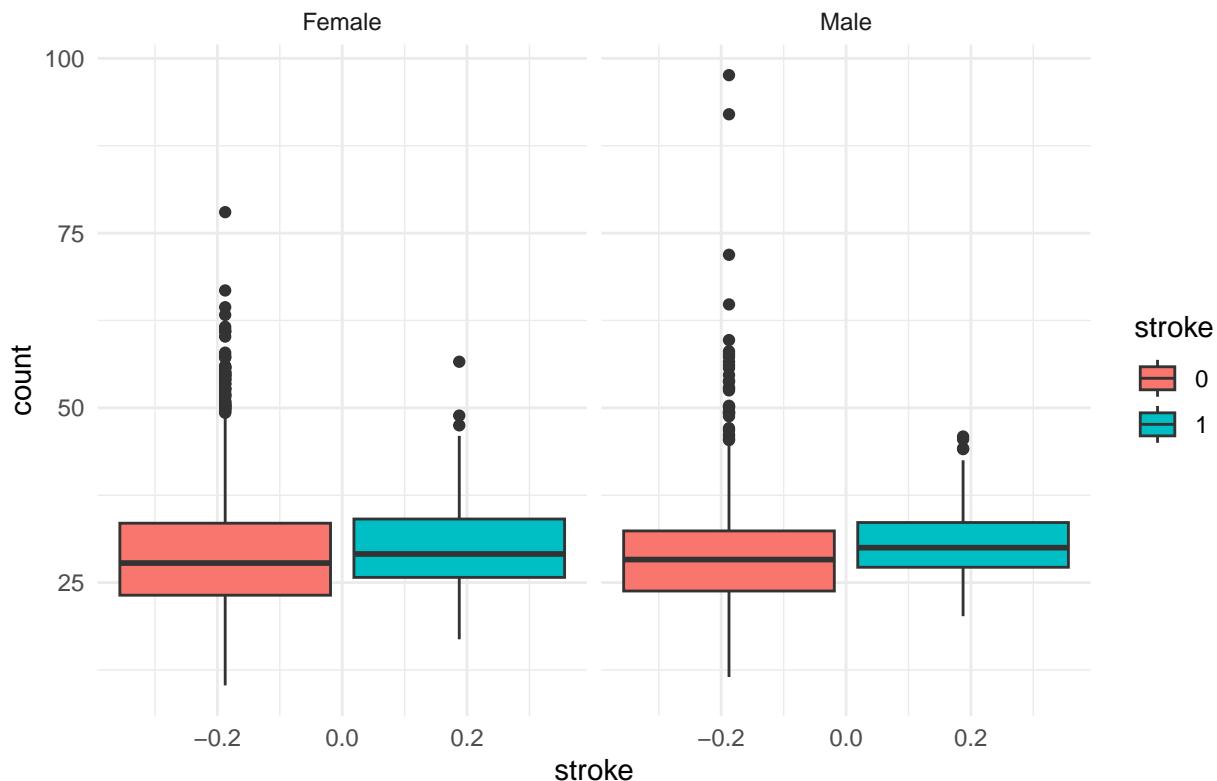
```
ggplot(stroke_data, aes(sample = bmi, fill = stroke))+
  stat_qq(size=2)+stat_qq_line() +
  facet_wrap(~stroke)+theme_minimal()+labs(title = "bmi qqplot by stroke",x="stroke",y="count")
```

bmi qqplot by stroke



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

bmi boxplot by stroke



```
#gender
```

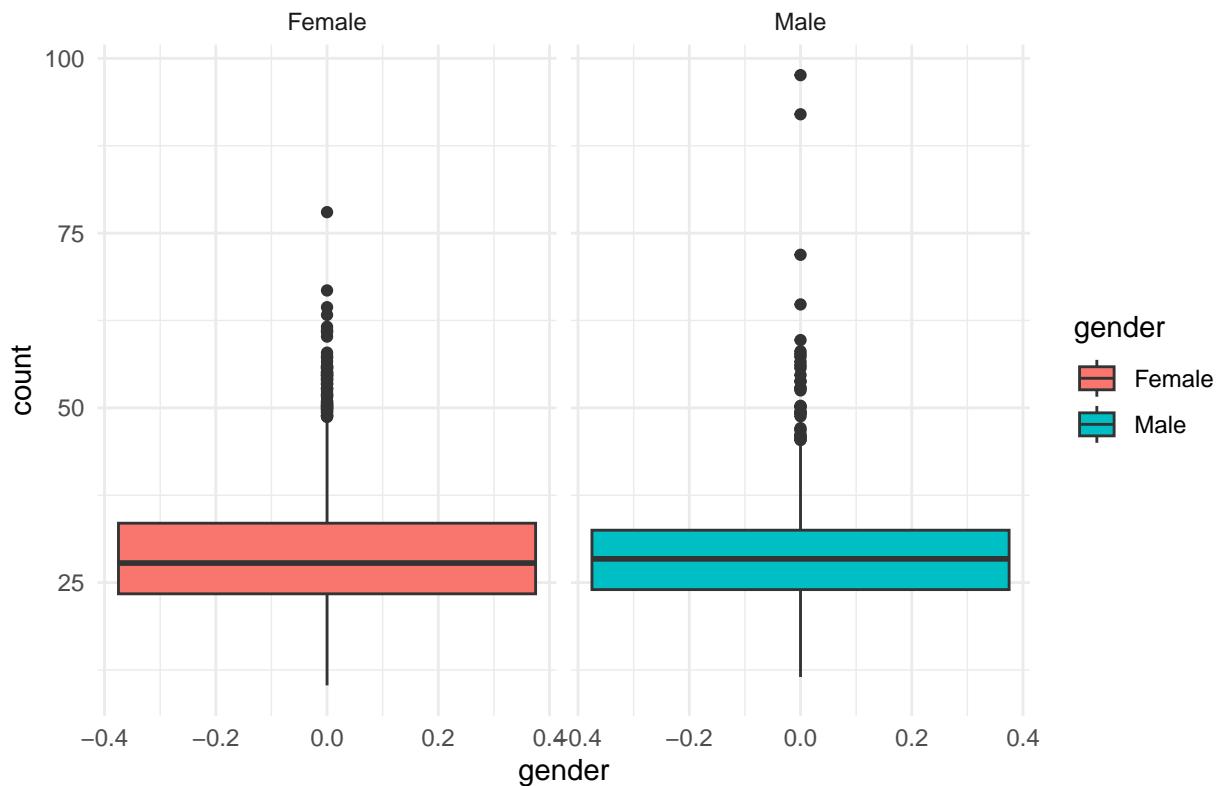
```
data_sub_bmi_gender<- stroke_data %>% group_by(gender,stroke) %>% summarise(mean = mean(bmi),.groups = "drop")  
print(data_sub_bmi_gender)
```

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

bmi boxplot by gender



```
#hypertension
```

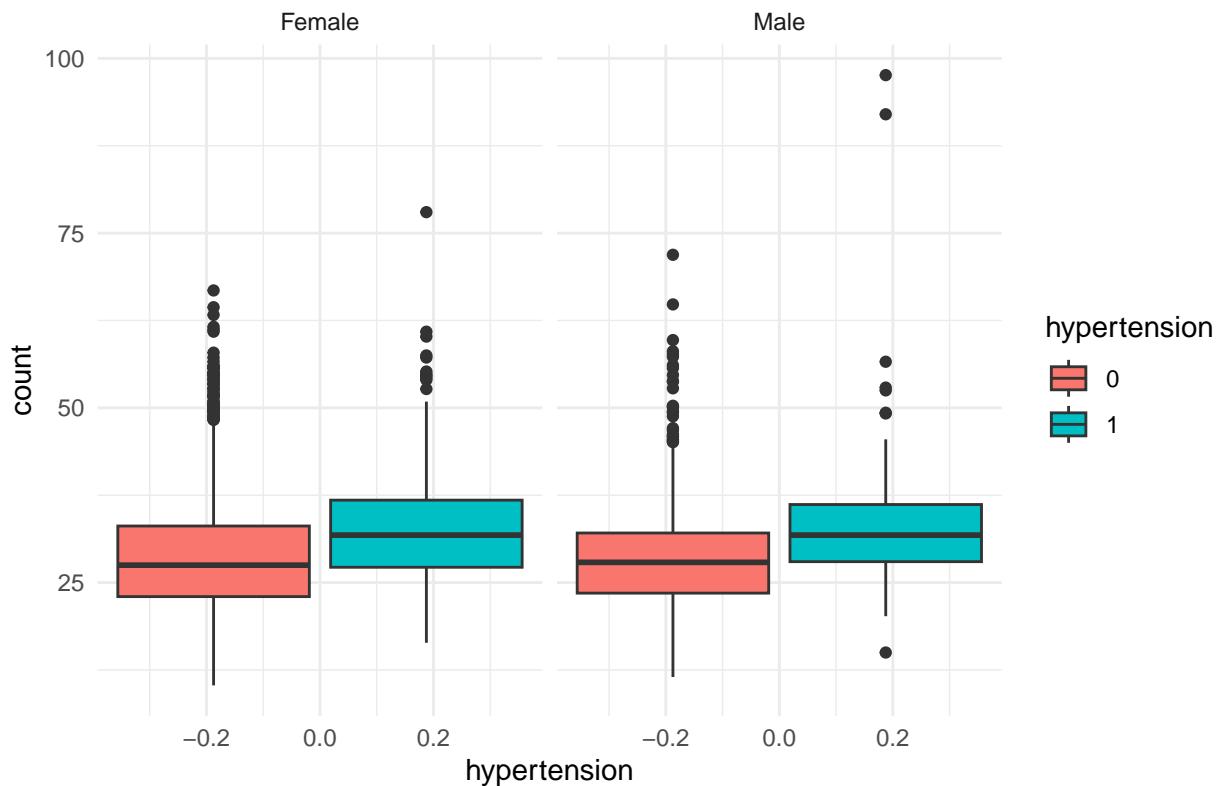
```
data_sub_bmi_hyper= stroke_data %>% group_by(hypertension,stroke) %>% summarise(mean = mean(bmi),.group=TRUE)
print(data_sub_bmi_hyper)
```

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

bmi boxplot by hypertension



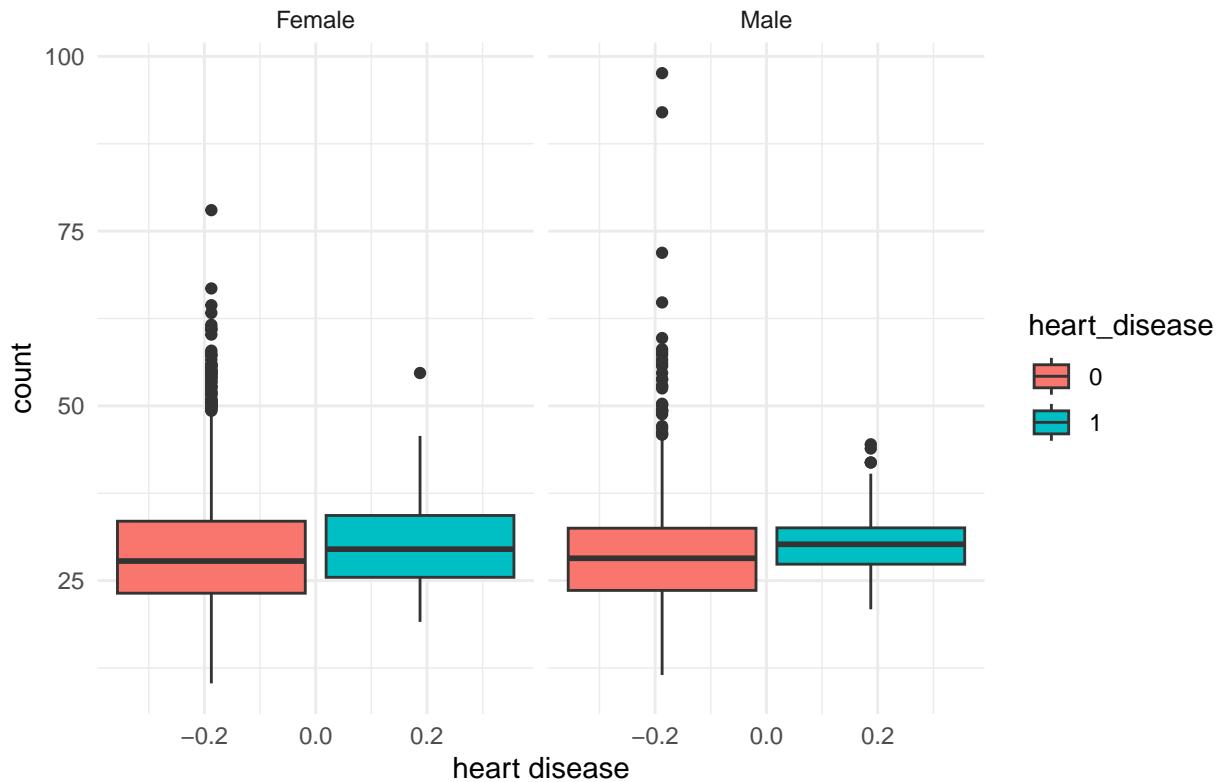
```
data_sub_bmi_hd = stroke_data %>% group_by(heart_disease, stroke) %>% summarise(mean = mean(bmi), .groups = "drop")
print(data_sub_bmi_hd)
```

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(stroke_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")
```

bmi boxplot by heart disease



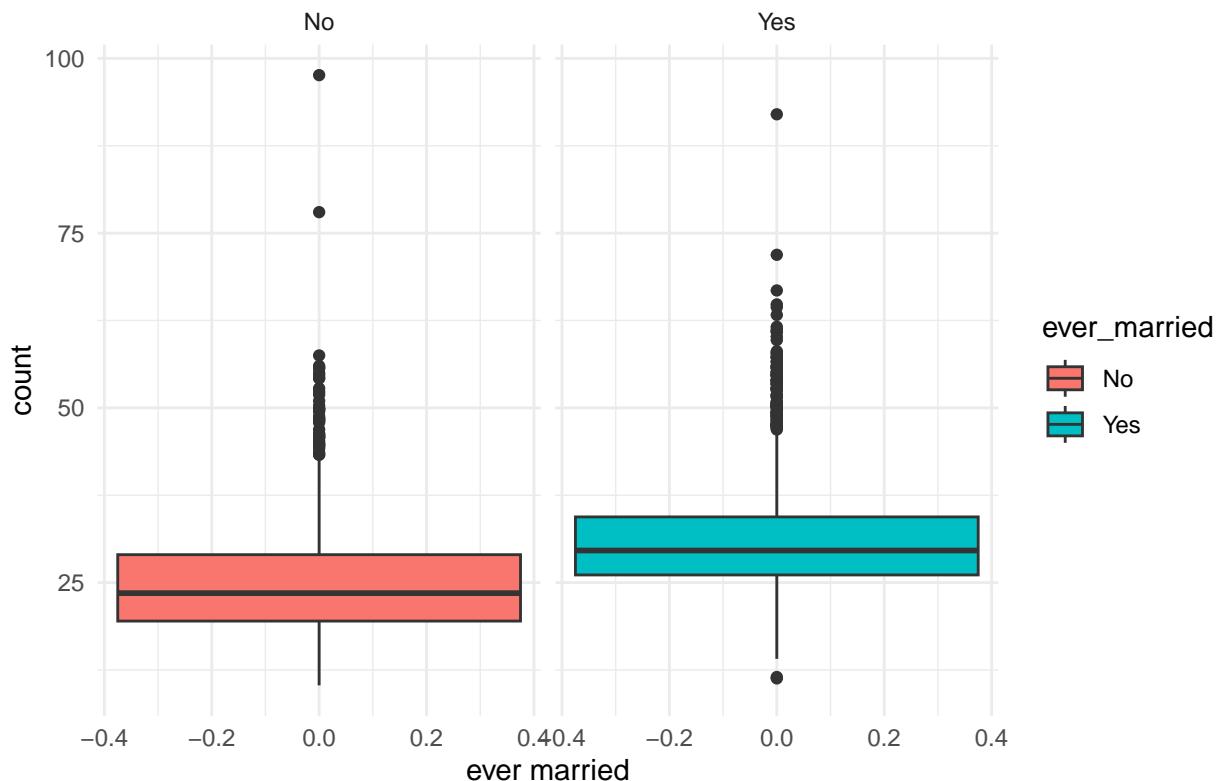
```
s#### BMI by ever married
```

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

```
## # 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(stroke_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")
```

bmi boxplot by ever married



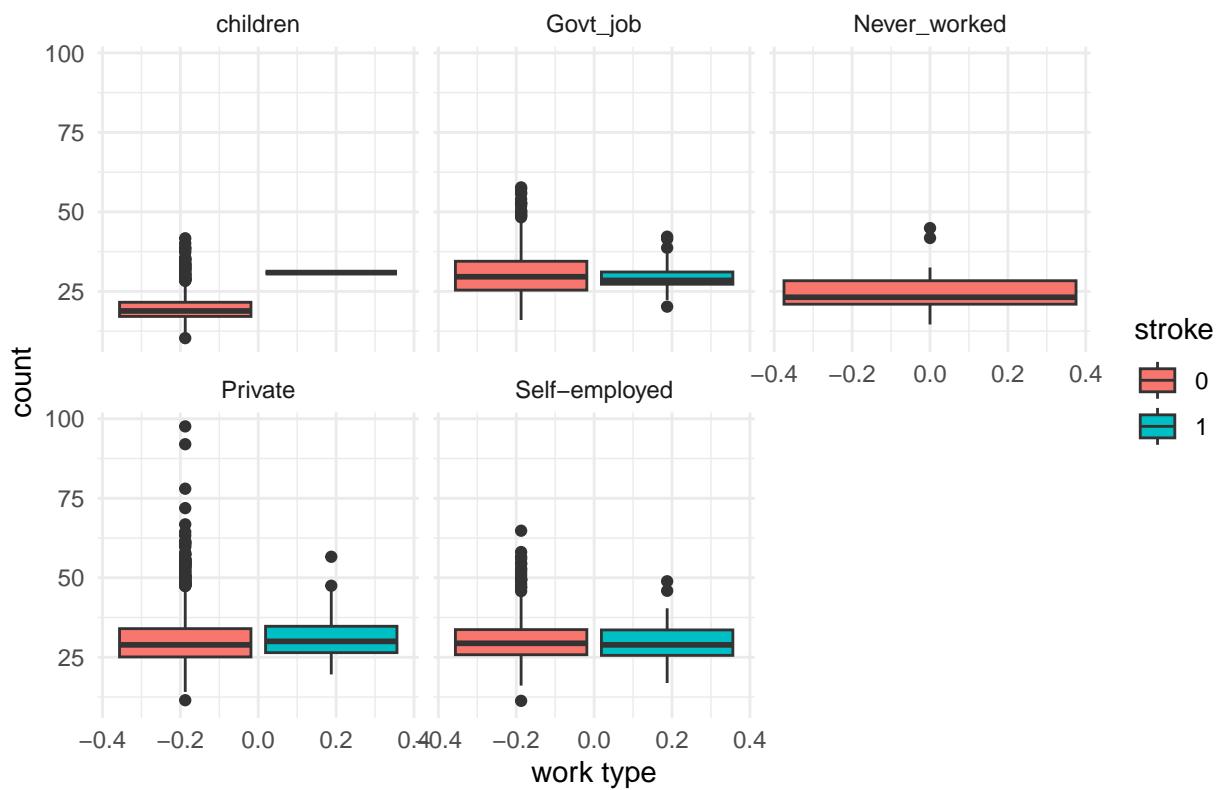
```
data_sub_bmi_wt = stroke_data %>% group_by(work_type, stroke) %>% summarise(mean = mean(bmi), .groups = "d")
print(data_sub_bmi_wt)
```

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(stroke_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")
```

bmi boxplot by stroke



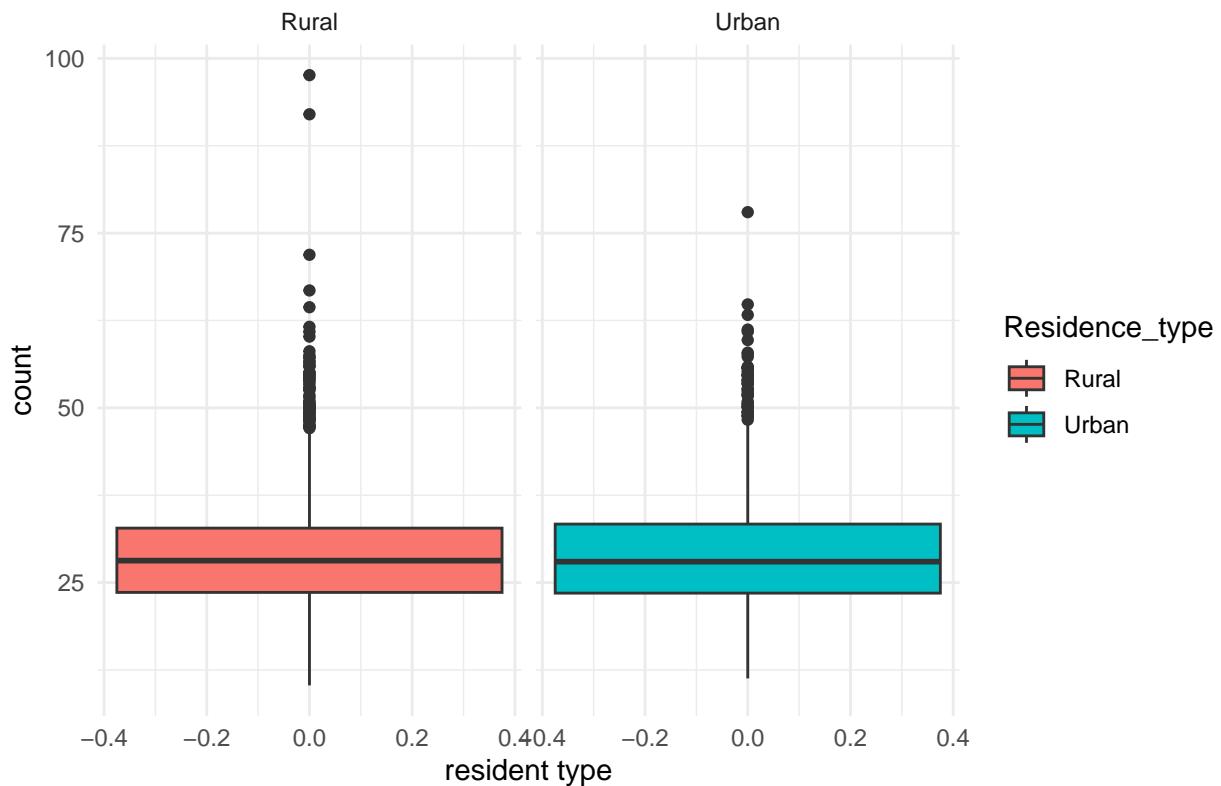
```
data_sub_bmi_tr = stroke_data %>% group_by(Residence_type,stroke) %>% summarise(mean = mean(bmi),.groups = "drop")
print(data_sub_bmi_tr)
```

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(stroke_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")
```

bmi boxplot by resident type



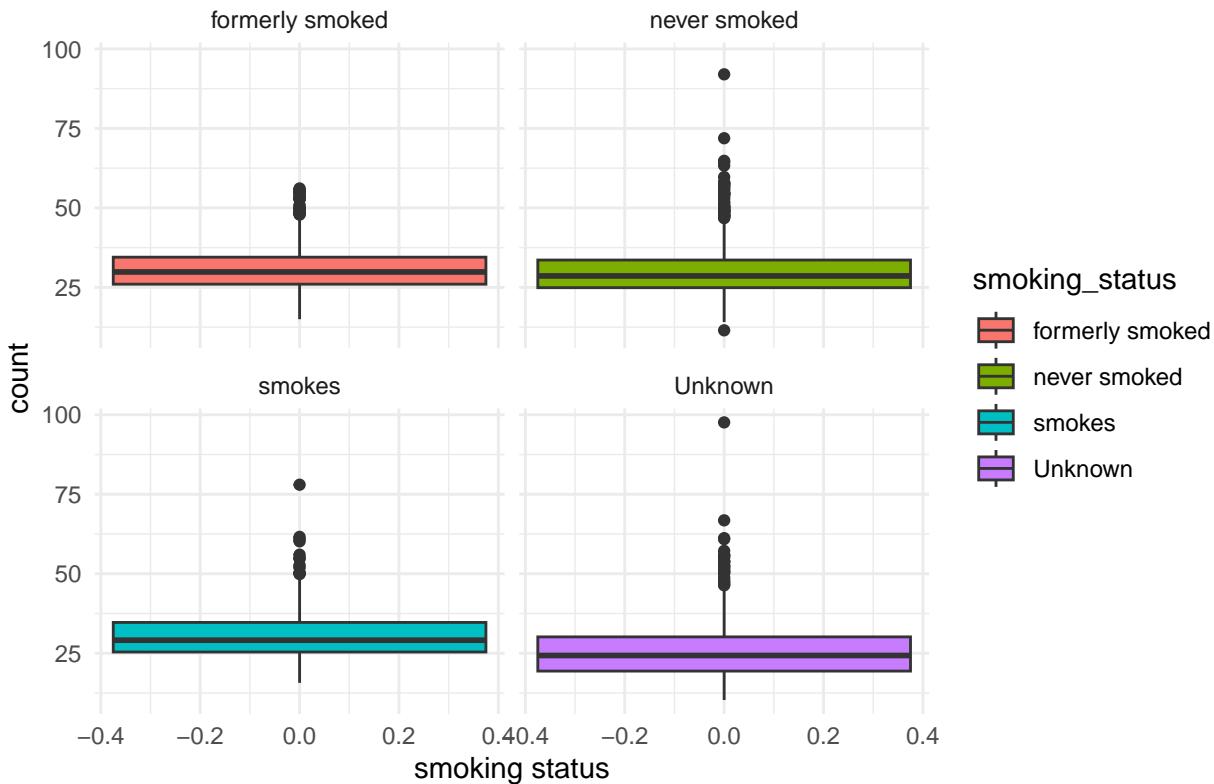
```
data_sub_bmi_catw= stroke_data %>% group_by(smoking_status,stroke) %>% summarise(mean = mean(bmi),.group=TRUE)
print(data_sub_bmi_catw)
```

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(stroke_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

The second continuous variable that we examine is average glucose levels. This variable measures on average, the amount of glucose found in blood.

```
summary(stroke_data$avg_glucose_level)
```

```
##      Min. 1st Qu. Median     Mean 3rd Qu.    Max.
##    55.12   77.07  91.68 105.30 113.50 271.74
```

Next, we examined the frequency distribution of average glucose levels using categories based on common medical standards. We used the cut function to create the following categories:

1. Normal: (0-99.9] mg/dL
2. Prediabetes: (99.9-125.9] mg/dL
3. Diabetes: (125.9-Inf] mg/dL

These categories correspond to the standard clinical definitions for fasting glucose levels:

- Normal: Less than 100 mg/dL
- Prediabetes: 100 to 125 mg/dL

- Diabetes: 126 mg/dL or higher

```

glc_mean <- mean(stroke_data$avg_glucose_level)
glc_median <- median(stroke_data$avg_glucose_level)

# Remove any NA values if present
stroke_data = na.omit(stroke_data)

# Create categories for glucose levels
# Categories based on common medical standards:
# Normal: <100 mg/dL
# Prediabetes: 100-125 mg/dL
# Diabetes: >=126 mg/dL
glucose_categories <- cut(stroke_data$avg_glucose_level,
                           breaks = c(0, 99.9, 125.9, Inf),
                           labels = c("Normal", "Prediabetes", "Diabetes"))

# Reorder the levels to put Normal in the middle
glucose_categories <- factor(glucose_categories, levels = c("Prediabetes", "Normal", "Diabetes"))

# Add the categories to the data frame
stroke_data$glucose_category = glucose_categories

# Create frequency table
glucose_freq_table <- table(glucose_categories)
print(glucose_freq_table)

## glucose_categories
## Prediabetes      Normal     Diabetes
##        969         3024        915

# Calculate relative frequencies
glucose_rel_freq <- prop.table(glucose_freq_table) * 100
print(glucose_rel_freq)

## glucose_categories
## Prediabetes      Normal     Diabetes
##    19.74328     61.61369    18.64303

# Custom function to calculate frequency statistics
custom_stat_freq <- function(x) {
  list(
    variance = var(x, na.rm = TRUE),
    mean = mean(x, na.rm = TRUE),
    median = median(x, na.rm = TRUE),
    mode = as.numeric(names(which.max(table(round(x, 1))))))
  )
}

# Calculate frequency statistics for glucose levels
glucose_freq_stats <- custom_stat_freq(stroke_data$avg_glucose_level)

```

```

# Print the results
print(glucose_freq_stats)

## $variance
## [1] 1973.63
##
## $mean
## [1] 105.2974
##
## $median
## [1] 91.68
##
## $mode
## [1] 82.4

# If you want a more formatted output:
cat("Variance:", glucose_freq_stats$variance, "\n")

## Variance: 1973.63

cat("Mean:", glucose_freq_stats$mean, "\n")

## Mean: 105.2974

cat("Median:", glucose_freq_stats$median, "\n")

## Median: 91.68

cat("Mode:", glucose_freq_stats$mode, "\n")

## Mode: 82.4

```

As we can see:

- The mean glucose level (105.30 mg/dL) falls within the prediabetes range, suggesting that on average, the population has slightly elevated glucose levels.
- However, the median (91.68 mg/dL) is in the normal range, indicating that more than half of the population has normal glucose levels.
- The mode (82.40 mg/dL) being well within the normal range suggests that normal glucose levels are the most common in the dataset.

In addition, the high variance (1973.63) indicates considerable spread in the glucose levels, which is consistent with the presence of both normal and elevated (prediabetic and diabetic) levels in the population.

The mean being higher than the median suggests a right-skewed distribution. This is consistent with having a majority in the normal range but a significant number of higher values pulling the mean up. While the majority (61.61%) have normal glucose levels, the combined 38.38% with prediabetes or diabetes represents a significant portion of the population at increased risk for various health issues, including potentially higher stroke risk.

The mean (105.30 mg/dL) being in the prediabetic range while the median (91.68 mg/dL) is in the normal range highlights the impact of the higher values on the overall distribution.

```

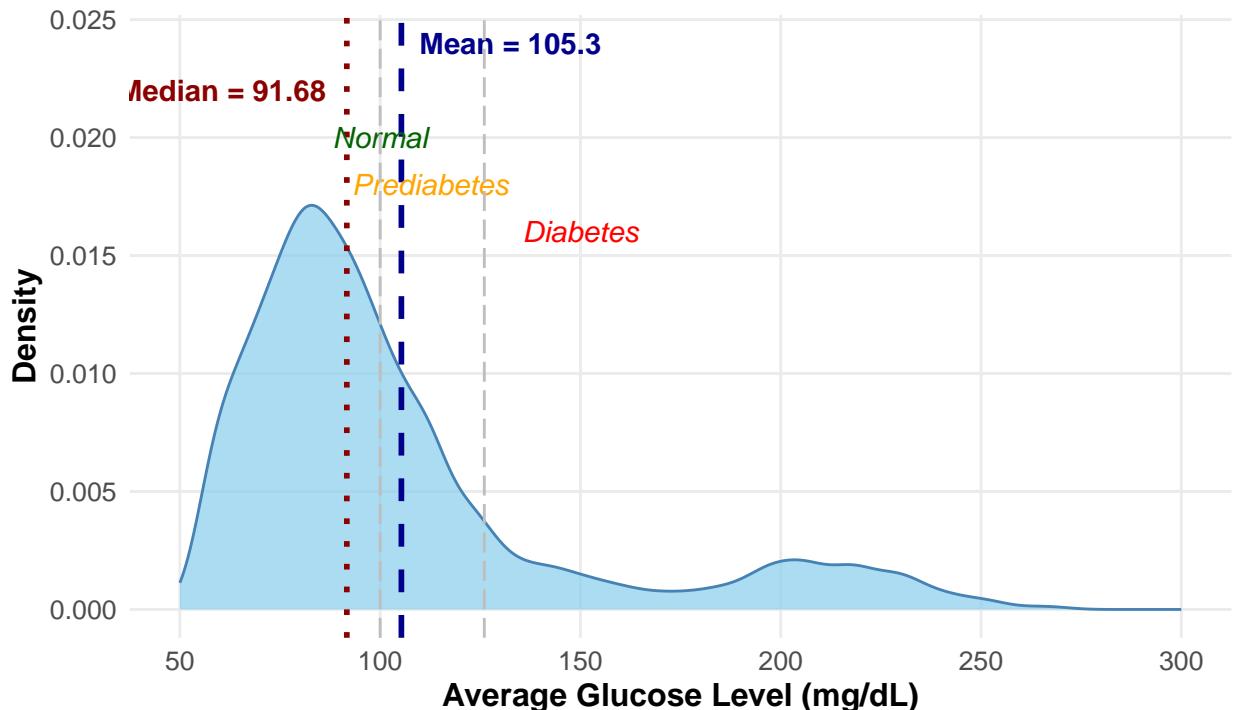
# Calculate mean and median
glc_mean <- mean(stroke_data$avg_glucose_level, na.rm = TRUE)
glc_median <- median(stroke_data$avg_glucose_level, na.rm = TRUE)

# Create the plot
ggplot(stroke_data, aes(x = avg_glucose_level)) +
  geom_density(fill = "skyblue", color = "steelblue", alpha = 0.7) +
  geom_vline(aes(xintercept = glc_mean), color = "darkblue", linetype = "dashed", size = 1) +
  geom_vline(aes(xintercept = glc_median), color = "darkred", linetype = "dotted", size = 1) +
  labs(
    title = "Distribution of Average Glucose Levels",
    subtitle = "With Mean and Median",
    x = "Average Glucose Level (mg/dL)",
    y = "Density",
    caption = "Graph x"
  ) +
  annotate("text", x = glc_mean, y = 0.024,
           label = paste("Mean =", round(glc_mean, 2)),
           color = "darkblue", hjust = -0.1, size = 4, fontface = "bold") +
  annotate("text", x = glc_median, y = 0.022,
           label = paste("Median =", round(glc_median, 2)),
           color = "darkred", hjust = 1.1, size = 4, fontface = "bold") +
  annotate("text", x = 100, y = 0.02, label = "Normal", color = "darkgreen", size = 4, fontface = 3) +
  annotate("text", x = 112.5, y = 0.018, label = "Prediabetes", color = "orange", size = 4, fontface = 3) +
  annotate("text", x = 150, y = 0.016, label = "Diabetes", color = "red", size = 4, fontface = 3) +
  geom_vline(xintercept = c(100, 126), color = "gray", linetype = "longdash", size = 0.5) +
  scale_x_continuous(breaks = seq(50, 300, by = 50), limits = c(50, 300)) +
  theme_minimal() +
  theme(
    plot.title = element_text(size = 16, face = "bold"),
    plot.subtitle = element_text(size = 12, face = "italic"),
    axis.title = element_text(size = 12, face = "bold"),
    axis.text = element_text(size = 10),
    legend.position = "none",
    panel.grid.minor = element_blank()
  )

```

Distribution of Average Glucose Levels

With Mean and Median



Graph x

As we can see, the graph confirms our findings from the frequency tables we had above. The distribution is not normal and heavily skewed to the right due to extreme values of Glucose level.

```
skewness(stroke_data$avg_glucose_level)
```

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

```
kurtosis(stroke_data$avg_glucose_level)
```

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

The value of Skewness also confirms this substantial right skew. This suggests a long tail on the right side of the distribution, indicating the presence of some very high glucose values that are pulling the mean higher than the median.

The positive value of Kurtosis also confirms that the distribution has more extreme values (in the tails) than a normal distribution would. Our value of 1.907085 suggests the distribution has heavier tails and a higher, sharper peak compared to a normal distribution.

```
# Calculate summary statistics
q1 <- quantile(stroke_data$avg_glucose_level, 0.25)
q3 <- quantile(stroke_data$avg_glucose_level, 0.75)
iqr <- q3 - q1
lower_whisker <- max(min(stroke_data$avg_glucose_level), q1 - 1.5 * iqr)
upper_whisker <- min(max(stroke_data$avg_glucose_level), q3 + 1.5 * iqr)
```

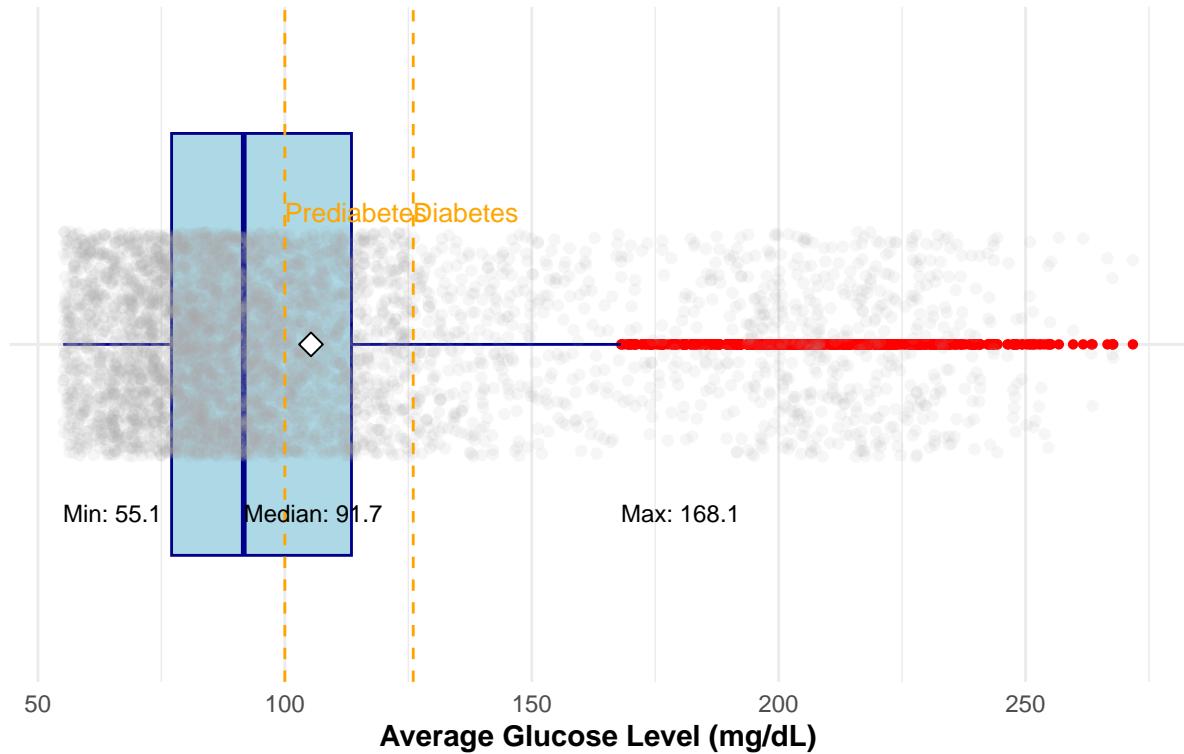
```

ggplot(stroke_data, aes(y = avg_glucose_level, x = "")) +
  geom_boxplot(fill = "lightblue", color = "darkblue", outlier.colour = "red", outlier.shape = 16) +
  geom_jitter(width = 0.2, alpha = 0.1, color = "darkgrey") + # Add jittered points
  stat_summary(fun = mean, geom = "point", shape = 23, size = 3, fill = "white") + # Add mean point
  geom_hline(yintercept = c(100, 126), linetype = "dashed", color = "orange", size = 0.5) + # Add reference lines
  scale_y_continuous(breaks = seq(0, 300, by = 50)) +
  coord_flip() + # Flip coordinates for horizontal boxplot
  labs(
    title = "Distribution of Average Glucose Levels",
    subtitle = "With reference lines for prediabetes (100 mg/dL) and diabetes (126 mg/dL)",
    x = "",
    y = "Average Glucose Level (mg/dL)"
  ) +
  annotate("text", x = 1.2, y = c(100, 126), label = c("Prediabetes", "Diabetes"),
           hjust = 0, vjust = -0.5, color = "orange", size = 3.5) +
  annotate("text", x = 0.7, y = lower_whisker, label = paste("Min:", round(lower_whisker, 1)),
           hjust = 0, size = 3) +
  annotate("text", x = 0.7, y = upper_whisker, label = paste("Max:", round(upper_whisker, 1)),
           hjust = 0, size = 3) +
  annotate("text", x = 0.7, y = median(stroke_data$avg_glucose_level),
           label = paste("Median:", round(median(stroke_data$avg_glucose_level), 1)),
           hjust = 0, size = 3) +
  theme_minimal() +
  theme(
    plot.title = element_text(face = "bold", size = 14),
    plot.subtitle = element_text(face = "italic", size = 10),
    axis.title = element_text(face = "bold"),
    axis.text.y = element_blank(),
    axis.ticks.y = element_blank()
  )

```

Distribution of Average Glucose Levels

With reference lines for prediabetes (100 mg/dL) and diabetes (126 mg/dL)



The box plot also confirms our findings that there is a concentration of values around the peak (likely in the normal glucose range) with a notable number of extreme values (likely in the diabetic range).

```
ggplot(stroke_data, aes(sample = avg_glucose_level)) +  
  stat_qq(size = 2, alpha = 0.6) +  
  stat_qq_line(color = "red", size = 1) +  
  geom_abline(intercept = 0, slope = 1, color = "blue", linetype = "dashed", size = 1) +  
  labs(  
    title = "Q-Q Plot for Average Glucose Levels",  
    subtitle = "Comparing sample quantiles to theoretical quantiles of a normal distribution",  
    x = "Theoretical Quantiles",  
    y = "Sample Quantiles (Glucose Levels mg/dL)"  
) +  
  scale_y_continuous(breaks = seq(0, 300, by = 50)) +  
  annotate("text", x = -2, y = max(stroke_data$avg_glucose_level),  
          label = "Reference Line (y=x)", color = "blue", hjust = 0, vjust = 1) +  
  annotate("text", x = 2, y = min(stroke_data$avg_glucose_level),  
          label = "QQ Line", color = "red", hjust = 1, vjust = 0) +  
  # Add reference lines for important glucose levels  
  geom_hline(yintercept = 100, color = "darkgreen", linetype = "dotted") +  
  geom_hline(yintercept = 126, color = "darkgreen", linetype = "dotted") +  
  annotate("text", x = -3, y = 100, label = "Prediabetes threshold (100 mg/dL)",  
          color = "darkgreen", hjust = 0, vjust = -0.5) +  
  annotate("text", x = -3, y = 126, label = "Diabetes threshold (126 mg/dL)",  
          color = "darkgreen", hjust = 0, vjust = -0.5) +  
  theme_minimal()
```

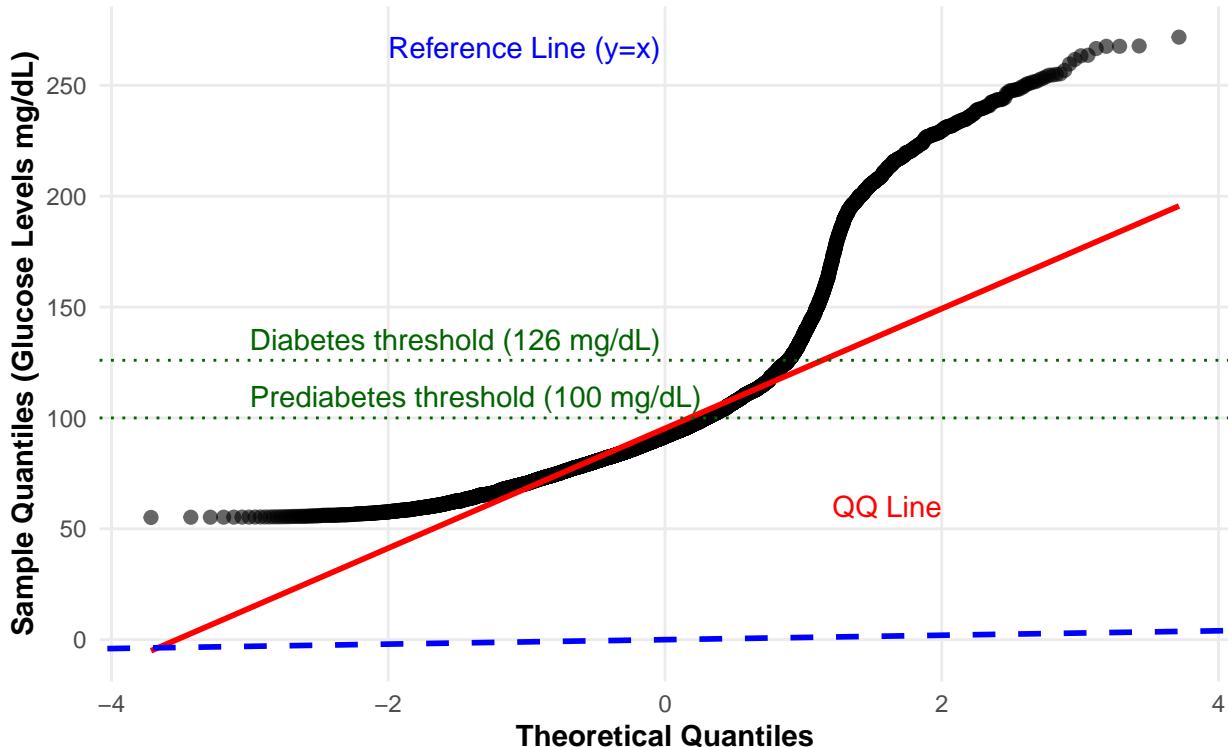
```

theme(
  plot.title = element_text(face = "bold", size = 14),
  plot.subtitle = element_text(face = "italic", size = 10),
  axis.title = element_text(face = "bold"),
  panel.grid.minor = element_blank()
)

```

Q-Q Plot for Average Glucose Levels

Comparing sample quantiles to theoretical quantiles of a normal distribution



This Q-Q plot also confirms that the average Glucose levels do not follow a normal distribution and are heavily skewed to the right.

```

# categorize glucose levels
dataG <- stroke_data %>%
  mutate(glucose_category = case_when(
    avg_glucose_level < 100 ~ "Normal",
    avg_glucose_level >= 100 & avg_glucose_level < 126 ~ "Prediabetes",
    avg_glucose_level >= 126 ~ "Diabetes"
  ))

# Now summarize mean glucose levels by category and stroke
data_sub_glucose_cat <- dataG %>%
  group_by(glucose_category, stroke) %>%
  summarise(mean = mean(avg_glucose_level), .groups = "drop")

```

```
# Print the summary
pander(data_sub_glucose_cat)
```

Avg. Glucose levels with stroke

glucose_category	stroke	mean
Diabetes	0	183.5
Diabetes	1	202
Normal	0	79.49
Normal	1	79.21
Prediabetes	0	110.6
Prediabetes	1	109.7

From the table, we can see that the most significant difference is in the Diabetes category, where those with stroke had substantially higher mean glucose levels. In the Normal and Prediabetes categories, there is minimal difference between stroke and no-stroke groups. The data suggests that very high glucose levels (Diabetes range) may be associated with increased stroke risk.

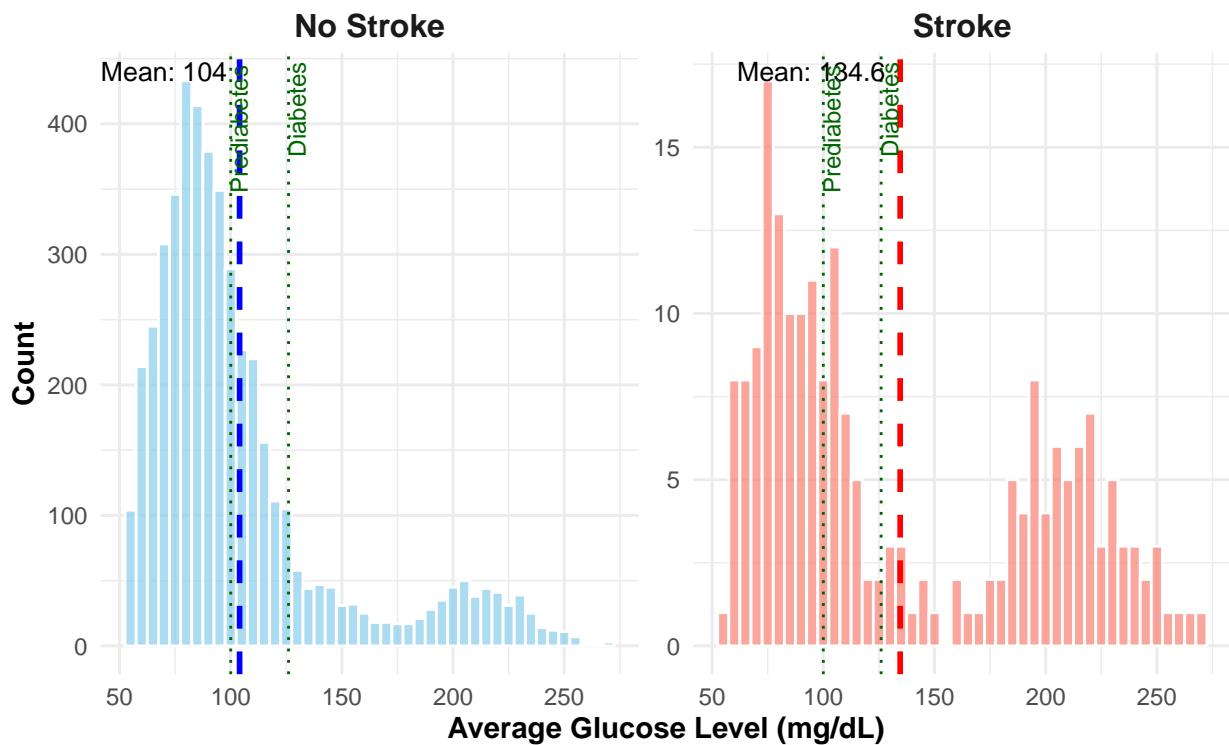
```
# Calculate means for annotation
mean_glucose <- stroke_data %>%
  group_by(stroke) %>%
  summarise(mean_glucose = mean(avg_glucose_level, na.rm = TRUE))

ggplot(stroke_data, aes(x = avg_glucose_level, fill = stroke)) +
  geom_histogram(binwidth = 5, color = "white", alpha = 0.7) +
  geom_vline(data = mean_glucose, aes(xintercept = mean_glucose, color = stroke),
             linetype = "dashed", size = 1) +
  facet_wrap(~stroke, scales = "free_y",
             labeller = labeller(stroke = c("0" = "No Stroke", "1" = "Stroke"))) +
  scale_fill_manual(values = c("0" = "skyblue", "1" = "salmon"),
                     labels = c("0" = "No Stroke", "1" = "Stroke")) +
  scale_color_manual(values = c("0" = "blue", "1" = "red")) +
  scale_x_continuous(breaks = seq(0, 300, by = 50)) +
  labs(
    title = "Distribution of Average Glucose Levels by Stroke Status",
    subtitle = "With mean glucose level and clinical thresholds indicated",
    x = "Average Glucose Level (mg/dL)",
    y = "Count",
    fill = "Stroke Status"
  ) +
  geom_vline(xintercept = c(100, 126), linetype = "dotted", color = "darkgreen", size = 0.5) +
  theme_minimal() +
  theme(
    plot.title = element_text(face = "bold", size = 14),
    plot.subtitle = element_text(face = "italic", size = 10),
    axis.title = element_text(face = "bold"),
    legend.position = "none",
    strip.text = element_text(face = "bold", size = 12)
  ) +
  geom_text(data = mean_glucose,
            aes(x = mean_glucose, y = Inf, label = paste("Mean:", round(mean_glucose, 1))),
            vjust = 1.5, hjust = 1.1, color = "black", size = 3.5) +
```

```
# Add annotations for thresholds to both facets
geom_text(data = data.frame(x = c(100, 126),
                             y = Inf,
                             label = c("Prediabetes", "Diabetes"),
                             stroke = factor(c(0, 0))), # Add to first facet
           aes(x = x, y = y, label = label),
           vjust = 1, hjust = 1.1, color = "darkgreen", angle = 90, size = 3) +
geom_text(data = data.frame(x = c(100, 126),
                             y = Inf,
                             label = c("Prediabetes", "Diabetes"),
                             stroke = factor(c(1, 1))), # Add to second facet
           aes(x = x, y = y, label = label),
           vjust = 1, hjust = 1.1, color = "darkgreen", angle = 90, size = 3)
```

Distribution of Average Glucose Levels by Stroke Status

With mean glucose level and clinical thresholds indicated



As we can see, On average, those who had a stroke have significantly higher glucose levels (mean: 134.68 mg/dL) compared to those without a stroke (mean: 104 mg/dL). A higher proportion of stroke patients fall into the prediabetes and diabetes ranges, as indicated by the clinical thresholds. This suggests a potential association between elevated glucose levels and stroke occurrence.

```
ggplot(stroke_data, aes(sample = avg_glucose_level, color = factor(stroke))) +
  stat_qq(size = 2) +
  stat_qq_line() +
  facet_wrap(~stroke, labeller = labeller(stroke = c("0" = "No Stroke", "1" = "Stroke"))) +
  scale_color_manual(values = c("0" = "skyblue", "1" = "salmon")) +
  labs(
    title = "Q-Q Plot of Average Glucose Levels by Stroke Status",
```

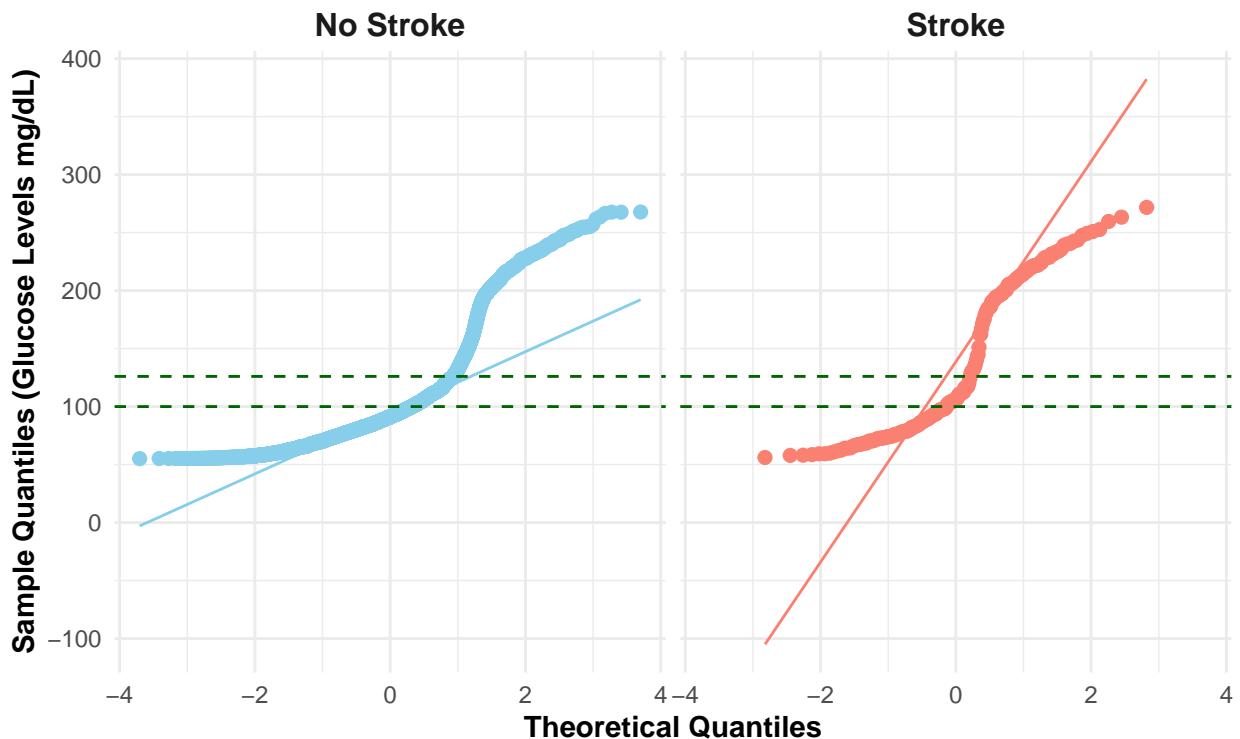
```

subtitle = "Comparing the distribution to a normal distribution",
x = "Theoretical Quantiles",
y = "Sample Quantiles (Glucose Levels mg/dL)"
) +
theme_minimal() +
theme(
  plot.title = element_text(face = "bold", size = 14),
  plot.subtitle = element_text(face = "italic", size = 10),
  axis.title = element_text(face = "bold"),
  legend.position = "none",
  strip.text = element_text(face = "bold", size = 12)
) +
geom_hline(yintercept = c(100, 126), linetype = "dashed", color = "darkgreen", size = 0.5)

```

Q-Q Plot of Average Glucose Levels by Stroke Status

Comparing the distribution to a normal distribution



From these two Q-Q plots, we can observe that for both groups, the points deviate from the diagonal line, indicating that the glucose levels do not follow a normal distribution. In addition, the deviation is more pronounced in the stroke group, where higher glucose levels show a greater departure from normality, particularly in the upper quantiles. This suggests that the stroke group has a higher concentration of extreme glucose values compared to the no-stroke group, further supporting the observed association between elevated glucose levels and stroke.

```

data_sub_glu_gd= stroke_data %>% group_by(gender,stroke) %>% summarise(mean_avg_glucose_level = mean(average))
pander(data_sub_glu_gd)

```

Avg. Glucose levels with gender/Stroke

gender	stroke	mean_avg_glucose_level
Female	0	102.3
Female	1	126.2
Male	0	106.4
Male	1	145.8

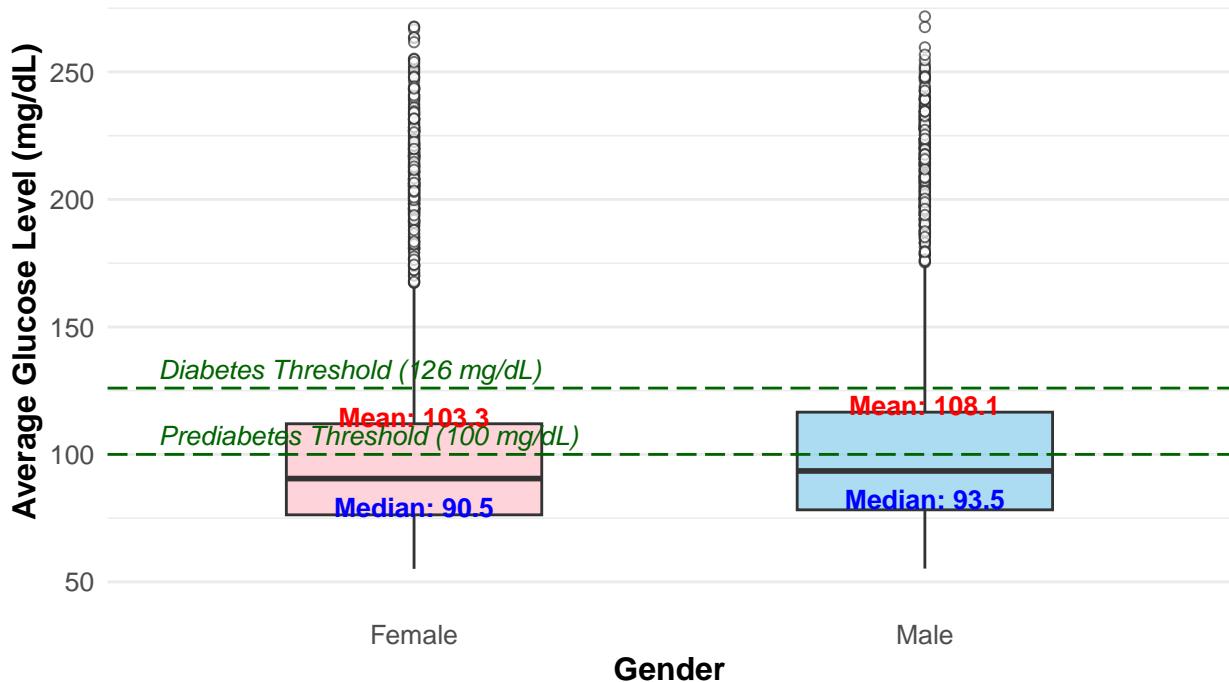
From above, we can see the mean average glucose levels for males and females with and without a history of stroke. For both genders, individuals who have had a stroke exhibit significantly higher average glucose levels compared to those who have not. In particular, males with a stroke have the highest mean glucose level (145.85 mg/dL), followed by females with a stroke (126.21 mg/dL). Males without a stroke have a slightly higher mean glucose level (106.39 mg/dL) compared to females without a stroke (102.34 mg/dL).

```
# Calculate mean and median glucose levels for each gender
gender_stats <- stroke_data %>%
  group_by(gender) %>%
  summarise(
    mean_glucose = mean(avg_glucose_level, na.rm = TRUE),
    median_glucose = median(avg_glucose_level, na.rm = TRUE)
  )

ggplot(stroke_data, aes(x = gender, y = avg_glucose_level, fill = gender)) +
  geom_boxplot(width = 0.5, alpha = 0.7, outlier.shape = 21, outlier.fill = "white") +
  geom_hline(yintercept = c(100, 126), color = "darkgreen", linetype = "longdash", size = 0.5) +
  scale_fill_manual(values = c("Male" = "skyblue", "Female" = "pink")) +
  scale_y_continuous(breaks = seq(0, 300, by = 50)) +
  labs(
    title = "Distribution of Average Glucose Levels by Gender",
    subtitle = "With mean, median, and clinical thresholds indicated",
    x = "Gender",
    y = "Average Glucose Level (mg/dL)",
    fill = "Gender"
  ) +
  theme_minimal() +
  theme(
    plot.title = element_text(face = "bold", size = 16),
    plot.subtitle = element_text(face = "italic", size = 12),
    axis.title = element_text(face = "bold", size = 12),
    axis.text = element_text(size = 10),
    legend.position = "none",
    panel.grid.major.x = element_blank()
  ) +
  geom_text(data = gender_stats, aes(x = gender, y = mean_glucose,
                                      label = paste("Mean:", round(mean_glucose, 1))),
            color = "red", vjust = -1, size = 3.5, fontface = "bold") +
  geom_text(data = gender_stats, aes(x = gender, y = median_glucose,
                                      label = paste("Median:", round(median_glucose, 1))),
            color = "blue", vjust = 2, size = 3.5, fontface = "bold") +
  annotate("text", x = 0.5, y = c(100, 126),
           label = c("Prediabetes Threshold (100 mg/dL)", "Diabetes Threshold (126 mg/dL)"),
           hjust = 0, vjust = -0.5, color = "darkgreen", size = 3.5, fontface = "italic") +
  coord_cartesian(ylim = c(50, max(stroke_data$avg_glucose_level, na.rm = TRUE) * 1.05))
```

Distribution of Average Glucose Levels by Gender

With mean, median, and clinical thresholds indicated



The plot shows the distribution of average glucose levels by gender, with both male and female groups represented. The median glucose level for females is 90.5 mg/dL, while for males, it is slightly higher at 93.5 mg/dL. The mean glucose level is also higher in males (108.1 mg/dL) compared to females (103.3 mg/dL). Clinical thresholds are indicated with the prediabetes threshold at 100 mg/dL and the diabetes threshold at 126 mg/dL. Both genders have average glucose levels that mostly fall below the diabetes threshold, with more variability observed among males.

```
# Define labels for stroke and gender
stroke_labels <- c("0" = "No Stroke", "1" = "Stroke")
gender_labels <- c("Male" = "Male", "Female" = "Female")

# Calculate summary statistics
summary_stats <- stroke_data %>%
  group_by(gender, stroke) %>%
  summarise(
    mean_glucose = mean(avg_glucose_level, na.rm = TRUE),
    median_glucose = median(avg_glucose_level, na.rm = TRUE),
    .groups = "drop"
  )

# Create the corrected violin plot
ggplot(stroke_data, aes(x = gender, y = avg_glucose_level, fill = stroke)) +
  geom_violin(trim = FALSE, alpha = 0.7) +
  geom_boxplot(width = 0.1, fill = "white", alpha = 0.5) +
  scale_fill_manual(values = c("0" = "skyblue", "1" = "salmon"), labels = stroke_labels) +
  facet_wrap(~stroke, labeller = as_labeller(stroke_labels)) +
```

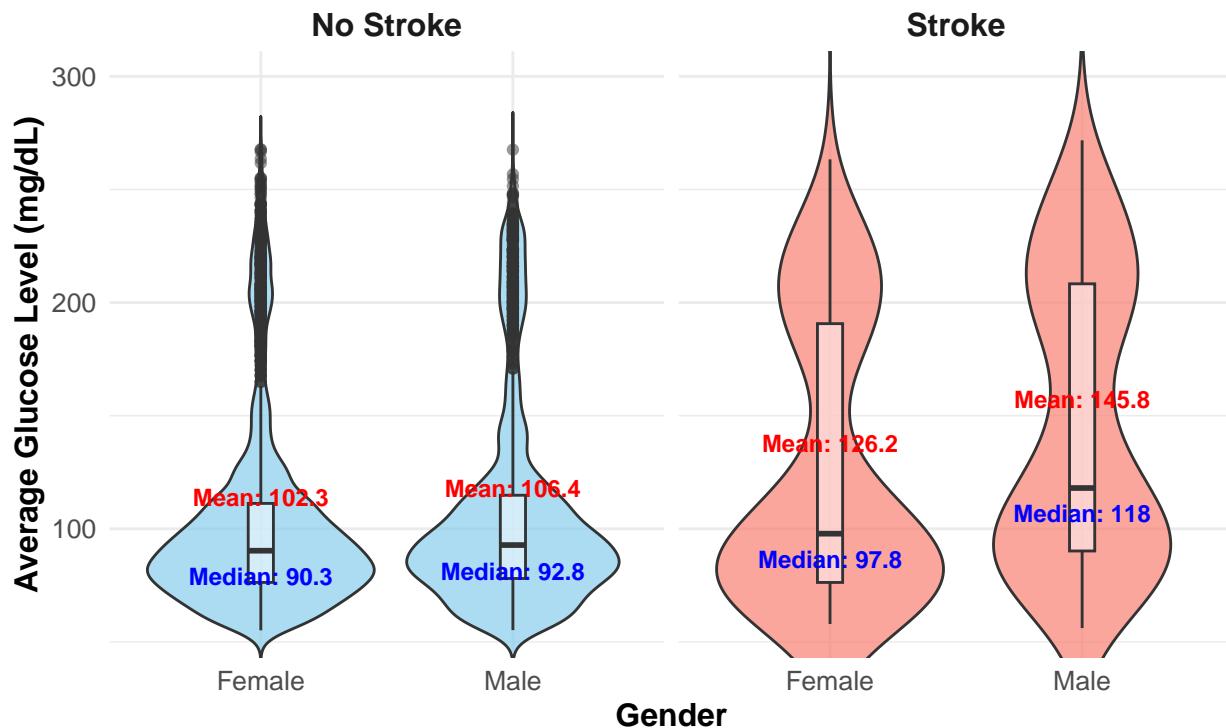
```

labs(
  title = "Distribution of Average Glucose Levels by Stroke Status and Gender",
  subtitle = "Violin plot with embedded box plot, mean, median, and jittered points",
  x = "Gender",
  y = "Average Glucose Level (mg/dL)",
  fill = "Stroke Status"
) +
theme_minimal() +
theme(
  plot.title = element_text(face = "bold", size = 16),
  plot.subtitle = element_text(face = "italic", size = 12),
  axis.title = element_text(face = "bold", size = 12),
  axis.text = element_text(size = 10),
  legend.position = "none",
  strip.text = element_text(size = 12, face = "bold")
) +
geom_text(data = summary_stats,
  aes(x = gender, y = mean_glucose,
      label = paste("Mean:", round(mean_glucose, 1))),
  color = "red", vjust = -1, size = 3, fontface = "bold") +
geom_text(data = summary_stats,
  aes(x = gender, y = median_glucose,
      label = paste("Median:", round(median_glucose, 1))),
  color = "blue", vjust = 2, size = 3, fontface = "bold") +
coord_cartesian(ylim = c(min(stroke_data$avg_glucose_level, na.rm = TRUE),
  max(stroke_data$avg_glucose_level, na.rm = TRUE) * 1.1))

```

Distribution of Average Glucose Levels by Stroke Status

Violin plot with embedded box plot, mean, median, and jittered points



The plot shows the shape of the distribution, along with the mean and median glucose levels, and clinical thresholds for prediabetes and diabetes. Males who had a stroke have the highest mean glucose level (145.8 mg/dL) and a distribution skewed towards higher glucose levels. Females who had a stroke have a mean of 126.2 mg/dL, while individuals without a stroke tend to have lower glucose levels, with males averaging 106.4 mg/dL and females 102.3 mg/dL. This highlights that stroke patients tend to have higher glucose levels, particularly among males.

```
data_sub_glu_hyp= stroke_data %>% group_by(hypertension,stroke) %>% summarise(mean = mean(avg_glucose_1))
pander(data_sub_glu_hyp)
```

Avg. Glucose levels with hypertension/Stroke

hypertension	stroke	mean
0	0	101.8
0	1	130
1	0	128.2
1	1	145.9

The data indicates that individuals with both hypertension and a history of stroke have the highest mean glucose level (145.90 mg/dL), followed by those with hypertension but no stroke (128.16 mg/dL). Individuals without hypertension who have had a stroke have a mean glucose level of 130.01 mg/dL, while those without

hypertension or stroke have the lowest mean glucose level (101.80 mg/dL). These findings suggest that both hypertension and stroke are associated with elevated glucose levels.

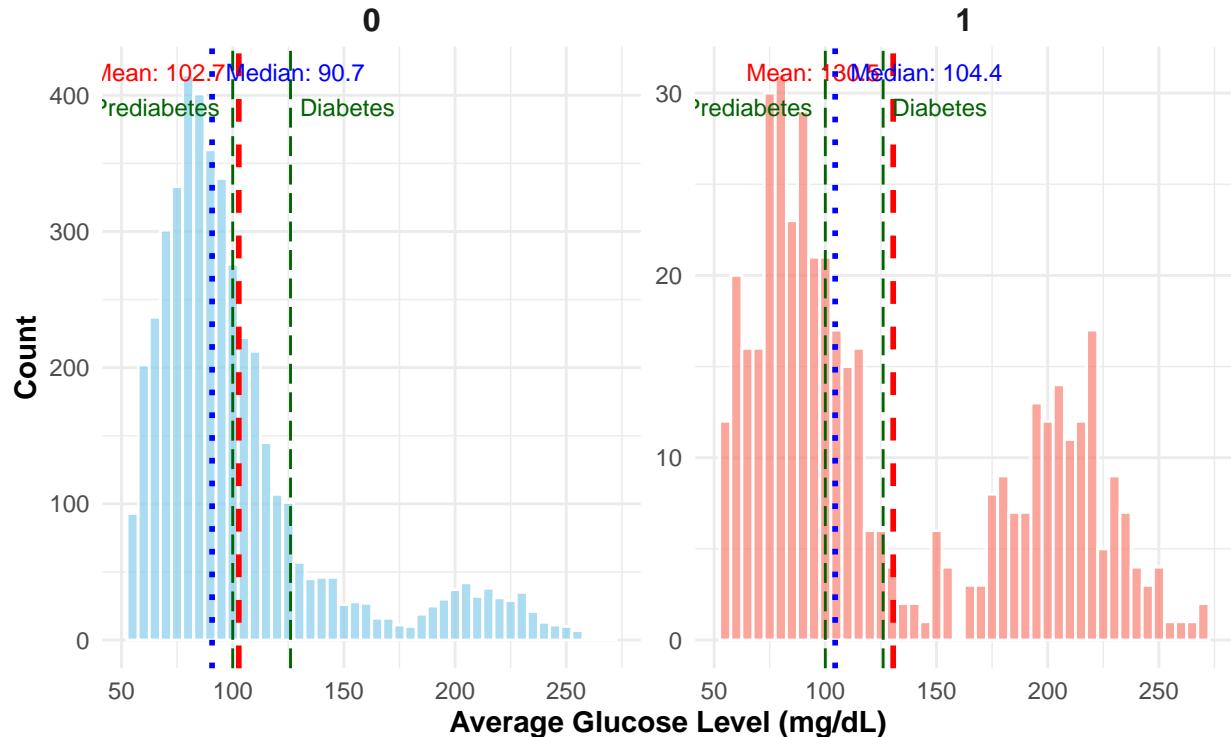
```
# Calculate summary statistics without using factor function
summary_stats <- stroke_data %>%
  group_by(hypertension) %>%
  summarise(
    mean_glucose = mean(avg_glucose_level, na.rm = TRUE),
    median_glucose = median(avg_glucose_level, na.rm = TRUE),
    .groups = "drop"
  )

# Create the plot without factor function and with correct labels
ggplot(stroke_data, aes(x = avg_glucose_level, fill = hypertension)) +
  geom_histogram(binwidth = 5, color = "white", alpha = 0.7) +
  facet_wrap(~hypertension, scales = "free_y") +
  geom_vline(data = summary_stats, aes(xintercept = mean_glucose), color = "red", linetype = "dashed",
  geom_vline(data = summary_stats, aes(xintercept = median_glucose), color = "blue", linetype = "dotted",
  geom_vline(xintercept = c(100, 126), color = "darkgreen", linetype = "longdash", size = 0.5) +
  
  # Correct label mappings without using factor
  scale_fill_manual(values = c("0" = "skyblue", "1" = "salmon"),
                    labels = c("No Hypertension", "Hypertension")) +
  
  # Adjust x-axis breaks
  scale_x_continuous(breaks = seq(50, 300, by = 50)) +
  
  # Add plot labels
  labs(
    title = "Distribution of Average Glucose Levels by Hypertension Status",
    subtitle = "With mean, median, and clinical thresholds indicated",
    x = "Average Glucose Level (mg/dL)",
    y = "Count",
    fill = "Hypertension"
  ) +
  
  # Customize theme
  theme_minimal() +
  theme(
    plot.title = element_text(face = "bold", size = 14),
    plot.subtitle = element_text(face = "italic", size = 10),
    axis.title = element_text(face = "bold"),
    legend.position = "none",
    strip.text = element_text(face = "bold", size = 12)
  ) +
  
  # Add mean and median text labels
  geom_text(data = summary_stats, aes(x = mean_glucose, y = Inf, label = paste("Mean:", round(mean_glucose,
    color = "red", hjust = 1.1, vjust = 2, size = 3) +
  geom_text(data = summary_stats, aes(x = median_glucose, y = Inf, label = paste("Median:", round(median_glucose,
    color = "blue", hjust = -0.1, vjust = 2, size = 3) +
  
  # Annotate clinical thresholds
  annotate("text", x = 100, y = Inf, label = "Prediabetes", color = "darkgreen", hjust = 1.1, vjust = 4)
```

```
annotate("text", x = 126, y = Inf, label = "Diabetes", color = "darkgreen", hjust = -0.1, vjust = 4, ...)
```

Distribution of Average Glucose Levels by Hypertension Status

With mean, median, and clinical thresholds indicated



No Hypertension Group (left panel): Individuals without hypertension have a mean glucose level of 102.7 mg/dL, with the majority falling below the prediabetes threshold (100 mg/dL). The median glucose level is slightly lower at 90.7 mg/dL, showing that most of the data is concentrated in the lower glucose ranges. This group has a healthy distribution, as very few individuals approach the diabetes threshold.

Hypertension Group (right panel): Individuals with hypertension have a higher mean glucose level of 130.5 mg/dL and a median of 104.4 mg/dL, with the majority falling above the prediabetes threshold. A substantial portion of this group reaches or exceeds the diabetes threshold, indicating that hypertension is associated with significantly higher glucose levels. This group also has a wider spread, with several individuals displaying glucose levels well beyond 200 mg/dL.

```
# Calculate mean and median glucose levels for each group
summary_stats <- stroke_data %>%
  group_by(hypertension, stroke) %>%
  summarise(
    mean_glucose = mean(avg_glucose_level, na.rm = TRUE),
    median_glucose = median(avg_glucose_level, na.rm = TRUE),
    .groups = "drop"
  )

# Create the plot
ggplot(stroke_data, aes(x = interaction(stroke, hypertension), y = avg_glucose_level, fill = interaction(...))
  geom_violin(trim = FALSE, alpha = 0.7) +
  geom_boxplot(width = 0.1, fill = "white", alpha = 0.5) +
```

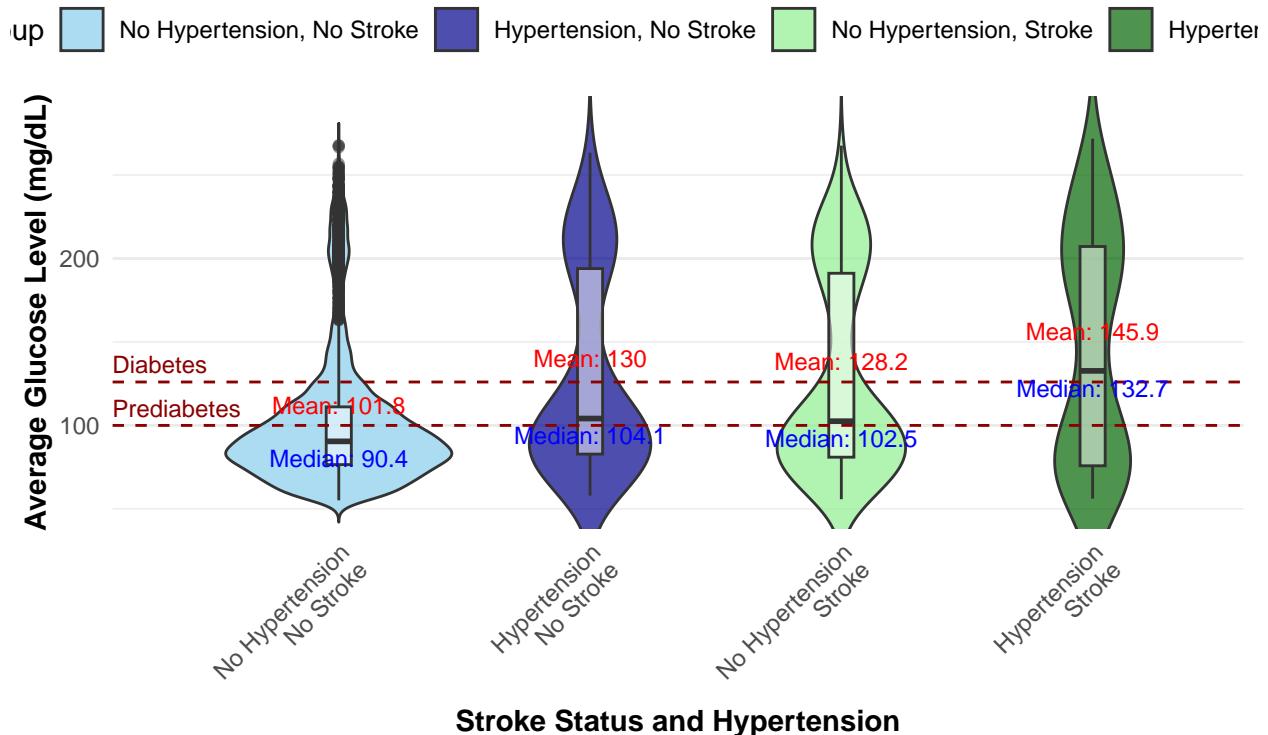
```

scale_fill_manual(values = c("0.0" = "skyblue", "0.1" = "lightgreen",
                           "1.0" = "darkblue", "1.1" = "darkgreen"),
                  labels = c("No Hypertension, No Stroke", "Hypertension, No Stroke",
                            "No Hypertension, Stroke", "Hypertension, Stroke")) +
scale_x_discrete(labels = c("No Hypertension\nNo Stroke", "Hypertension\nNo Stroke",
                            "No Hypertension\nStroke", "Hypertension\nStroke")) +
geom_hline(yintercept = c(100, 126), linetype = "dashed", color = "darkred", size = 0.5) +
geom_text(data = summary_stats, aes(x = interaction(stroke, hypertension), y = mean_glucose,
                                     label = paste("Mean:", round(mean_glucose, 1))),
           position = position_dodge(width = 0.75), vjust = -0.5, color = "red", size = 3) +
geom_text(data = summary_stats, aes(x = interaction(stroke, hypertension), y = median_glucose,
                                     label = paste("Median:", round(median_glucose, 1))),
           position = position_dodge(width = 0.75), vjust = 1.5, color = "blue", size = 3) +
labs(
  title = "Distribution of Average Glucose Levels by Stroke Status and Hypertension",
  subtitle = "Violin plot with embedded box plot, mean, median, and clinical thresholds",
  x = "Stroke Status and Hypertension",
  y = "Average Glucose Level (mg/dL)",
  fill = "Group"
) +
theme_minimal() +
theme(
  plot.title = element_text(face = "bold", size = 14),
  plot.subtitle = element_text(face = "italic", size = 10),
  axis.title = element_text(face = "bold"),
  legend.position = "top",
  axis.text.x = element_text(angle = 45, hjust = 1),
  panel.grid.major.x = element_blank()
) +
annotate("text", x = 0.1, y = c(100, 126), label = c("Prediabetes", "Diabetes"),
         hjust = 0, vjust = -0.5, color = "darkred", size = 3) +
coord_cartesian(ylim = c(50, max(stroke_data$avg_glucose_level, na.rm = TRUE) * 1.05))

```

Distribution of Average Glucose Levels by Stroke Status and Hypertension

Violin plot with embedded box plot, mean, median, and clinical thresholds



This violin plot shows the distribution of average glucose levels across four groups categorized by stroke status and hypertension. The group with neither hypertension nor stroke has the lowest median (90.4 mg/dL) and mean (101.8 mg/dL) glucose levels, both below the prediabetes threshold of 100 mg/dL. In contrast, the groups with hypertension (with or without stroke) exhibit significantly higher glucose levels. The group with both hypertension and stroke has the highest mean (145.9 mg/dL) and median (132.7 mg/dL), surpassing the diabetes threshold of 126 mg/dL. The data suggests that hypertension and stroke are associated with elevated glucose levels, with a clear increase in glucose levels as health conditions worsen.

```
data_sub_glu_hd= stroke_data %>% group_by(heart_disease,stroke) %>% summarise(mean_glucose_level = mean(pander(data_sub_glu_hd))
```

Avg. Glucose levels with heart disease/Stroke

heart_disease	stroke	mean_glucose_level
0	0	102.8
0	1	127.3
1	0	129.5
1	1	165.5

We can observe that individuals with neither heart disease nor stroke have a mean glucose level of 102.85 mg/dL, which is below the prediabetes threshold. Those without heart disease but with a stroke have a higher mean glucose level of 127.26 mg/dL, approaching the diabetes threshold. For individuals with heart

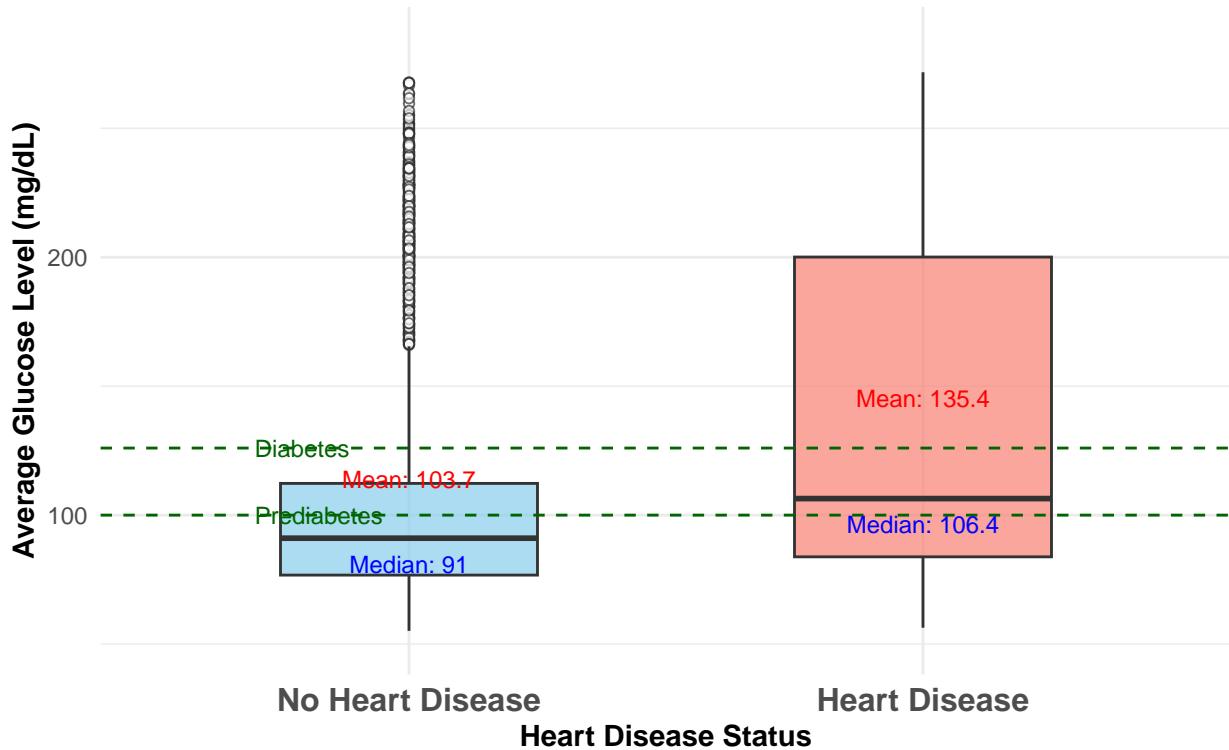
disease but no stroke, the mean glucose level is 129.46 mg/dL, exceeding the diabetes threshold. The highest glucose level, 165.46 mg/dL, is observed in individuals with both heart disease and stroke, indicating a strong association between these conditions and elevated glucose levels.

```
# Calculate summary statistics
summary_stats <- stroke_data %>%
  group_by(heart_disease) %>%
  summarise(
    mean_glucose = mean(avg_glucose_level, na.rm = TRUE),
    median_glucose = median(avg_glucose_level, na.rm = TRUE),
    .groups = "drop"
  )

# Create the plot
ggplot(stroke_data, aes(x = factor(heart_disease), y = avg_glucose_level, fill = factor(heart_disease))) +
  geom_boxplot(width = 0.5, outlier.shape = 21, outlier.fill = "white", alpha = 0.7) +
  geom_hline(yintercept = c(100, 126), linetype = "dashed", color = "darkgreen", size = 0.5) +
  scale_fill_manual(values = c("0" = "skyblue", "1" = "salmon"),
                     labels = c("0" = "No Heart Disease", "1" = "Heart Disease")) +
  scale_x_discrete(labels = c("0" = "No Heart Disease", "1" = "Heart Disease")) +
  geom_text(data = summary_stats, aes(x = factor(heart_disease), y = mean_glucose,
                                       label = paste("Mean:", round(mean_glucose, 1))),
            position = position_dodge(width = 0.75), vjust = -1, color = "red", size = 3) +
  geom_text(data = summary_stats, aes(x = factor(heart_disease), y = median_glucose,
                                       label = paste("Median:", round(median_glucose, 1))),
            position = position_dodge(width = 0.75), vjust = 2, color = "blue", size = 3) +
  labs(
    title = "Distribution of Average Glucose Levels by Heart Disease Status",
    subtitle = "With mean, median, and clinical thresholds indicated",
    x = "Heart Disease Status",
    y = "Average Glucose Level (mg/dL)",
    fill = "Heart Disease Status"
  ) +
  theme_minimal() +
  theme(
    plot.title = element_text(face = "bold", size = 14),
    plot.subtitle = element_text(face = "italic", size = 10),
    axis.title = element_text(face = "bold"),
    legend.position = "none",
    axis.text.x = element_text(face = "bold", size = 12)
  ) +
  annotate("text", x = 0.7, y = 100, label = "Prediabetes", color = "darkgreen", hjust = 0, size = 3) +
  annotate("text", x = 0.7, y = 126, label = "Diabetes", color = "darkgreen", hjust = 0, size = 3) +
  coord_cartesian(ylim = c(50, max(stroke_data$avg_glucose_level, na.rm = TRUE) * 1.05))
```

Distribution of Average Glucose Levels by Heart Disease Status

With mean, median, and clinical thresholds indicated



From the boxplot, we can see that for individuals without heart disease, the median glucose level is 91 mg/dL, which is below the prediabetes threshold, while the mean is slightly higher at 103.7 mg/dL, just above the prediabetes threshold. In contrast, individuals with heart disease have a significantly higher mean glucose level of 135.4 mg/dL, well above the diabetes threshold, with a median of 106.4 mg/dL.

```
# Calculate mean and median glucose levels for each group
summary_stats <- stroke_data %>%
  group_by(stroke, heart_disease) %>%
  summarise(
    mean_glucose = mean(avg_glucose_level, na.rm = TRUE),
    median_glucose = median(avg_glucose_level, na.rm = TRUE),
    .groups = "drop"
  )

# Create the plot with corrected label order
ggplot(stroke_data, aes(x = interaction(stroke, heart_disease), y = avg_glucose_level, fill = interaction(stroke, heart_disease)),
       geom_violin(trim = FALSE, alpha = 0.7) +
       geom_boxplot(width = 0.1, fill = "white", alpha = 0.5) +
       scale_fill_manual(values = c("0.0" = "skyblue",      # No Stroke, No Heart Disease
                                   "0.1" = "darkgreen",   # No Stroke, Heart Disease
                                   "1.0" = "lightgreen",  # Stroke, No Heart Disease
                                   "1.1" = "darkblue"),   # Stroke, Heart Disease
       labels = c("No Stroke, No Heart Disease",
                 "No Stroke, Heart Disease",
                 "Stroke, No Heart Disease",
                 "Stroke, Heart Disease")) +
```

```

# Correct x-axis label order: "No Stroke" comes first
scale_x_discrete(labels = c("No Stroke\nNo Heart Disease",
                           "No Stroke\nHeart Disease",
                           "Stroke\nNo Heart Disease",
                           "Stroke\nHeart Disease")) +


# Add clinical threshold lines
geom_hline(yintercept = c(100, 126), linetype = "dashed", color = "darkred", size = 0.5) +


# Add mean and median glucose levels
geom_text(data = summary_stats, aes(x = interaction(stroke, heart_disease), y = mean_glucose,
                                      label = paste("Mean:", round(mean_glucose, 1))),
           position = position_dodge(width = 0.75), vjust = -0.5, color = "red", size = 3) +
geom_text(data = summary_stats, aes(x = interaction(stroke, heart_disease), y = median_glucose,
                                      label = paste("Median:", round(median_glucose, 1))),
           position = position_dodge(width = 0.75), vjust = 1.5, color = "blue", size = 3) +


# Labels for the plot
labs(
  title = "Distribution of Average Glucose Levels by Stroke Status and Heart Disease",
  subtitle = "Violin plot with embedded box plot, mean, median, and clinical thresholds",
  x = "Stroke Status and Heart Disease",
  y = "Average Glucose Level (mg/dL)",
  fill = "Group"
) +


# Customize the theme
theme_minimal() +
theme(
  plot.title = element_text(face = "bold", size = 14),
  plot.subtitle = element_text(face = "italic", size = 10),
  axis.title = element_text(face = "bold"),
  legend.position = "top",
  axis.text.x = element_text(angle = 45, hjust = 1),
  panel.grid.major.x = element_blank()
) +

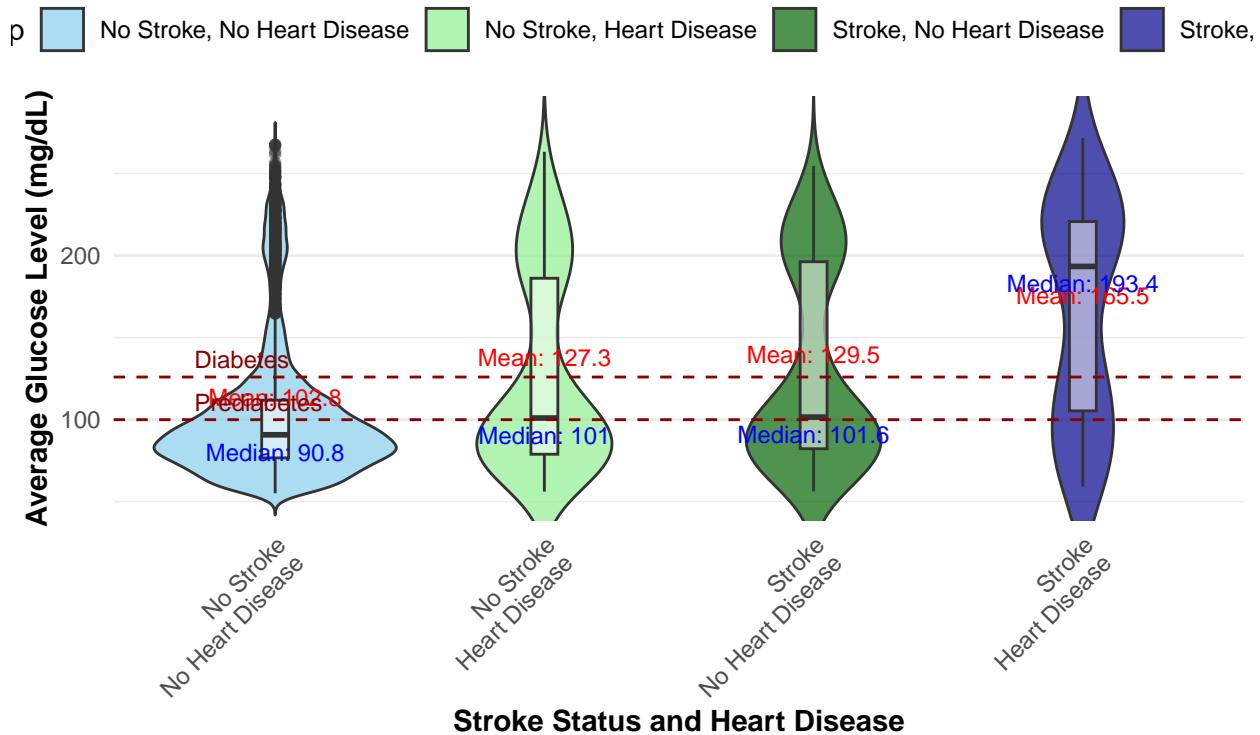

# Annotate clinical thresholds
annotate("text", x = 0.7, y = c(100, 126), label = c("Prediabetes", "Diabetes"),
         hjust = 0, vjust = -0.5, color = "darkred", size = 3) +


# Adjust y-axis limits
coord_cartesian(ylim = c(50, max(stroke_data$avg_glucose_level, na.rm = TRUE) * 1.05))

```

Distribution of Average Glucose Levels by Stroke Status and Heart Disease

Violin plot with embedded box plot, mean, median, and clinical thresholds



From this violin plot we can see that, individuals without heart disease or stroke have the lowest mean glucose level at 102.8 mg/dL, which is just above the prediabetes threshold, and a median of 90.8 mg/dL, below the prediabetes threshold. The group with heart disease but no stroke has a higher mean glucose level of 127.3 mg/dL, approaching the diabetes threshold, and a median of 101 mg/dL. Individuals without heart disease but with a stroke have a mean glucose level of 129.5 mg/dL, exceeding the diabetes threshold, with a median of 101.6 mg/dL. Finally, the group with both heart disease and stroke shows the highest mean glucose level at 165.5 mg/dL and a median of 193.4 mg/dL, far exceeding the diabetes threshold.

```
data_sub_glu_em = stroke_data %>% group_by(ever_married,stroke) %>% summarise(mean = mean(avg_glucose_level))
pander(data_sub_glu_em)
```

Avg. Glucose levels with ever married/Stroke

ever_married	stroke	mean
No	0	95.91
No	1	106.9
Yes	0	108.5
Yes	1	138

Based on the table:

1. Not married, no stroke individuals have an average glucose level of 95.91 mg/dL, which is below the prediabetes threshold.

2. Not married, stroke individuals show a higher mean glucose level of 106.86 mg/dL, which exceeds the prediabetes threshold.
3. Married, no stroke individuals have a mean glucose level of 108.50 mg/dL, also above the prediabetes threshold.
4. Married, stroke individuals exhibit the highest mean glucose level of 138 mg/dL, which is above the diabetes threshold.

This suggests that both marital status and stroke occurrence are associated with elevated glucose levels, with married individuals, especially those who have had a stroke, showing the highest average glucose levels.

```
# Calculate summary statistics
summary_stats <- stroke_data %>%
  group_by(ever_married) %>%
  summarise(
    mean_glucose = mean(avg_glucose_level, na.rm = TRUE),
    median_glucose = median(avg_glucose_level, na.rm = TRUE),
    .groups = "drop"
  )

# Create the plot
ggplot(stroke_data, aes(x = ever_married, y = avg_glucose_level, fill = ever_married)) +
  geom_boxplot(width = 0.5, alpha = 0.7, outlier.shape = 21, outlier.fill = "white") +
  geom_hline(yintercept = c(100, 126), color = "darkgreen", linetype = "longdash", size = 0.5) +
  # Set the correct fill colors
  scale_fill_manual(values = c("No" = "skyblue", "Yes" = "lightgreen"),
                    labels = c("Never Married", "Ever Married")) +
  # Adjust y-axis scale
  scale_y_continuous(breaks = seq(0, 300, by = 50)) +
  # Add labels for the plot
  labs(
    title = "Distribution of Average Glucose Levels by Marital Status",
    subtitle = "With mean, median, and clinical thresholds indicated",
    x = "Marital Status",
    y = "Average Glucose Level (mg/dL)",
    fill = "Marital Status"
  ) +
  # Customize the theme
  theme_minimal() +
  theme(
    plot.title = element_text(face = "bold", size = 16),
    plot.subtitle = element_text(face = "italic", size = 12),
    axis.title = element_text(face = "bold", size = 12),
    axis.text = element_text(size = 10),
    legend.position = "none",
    panel.grid.major.x = element_blank()
  ) +
  # Add mean and median glucose text labels
```

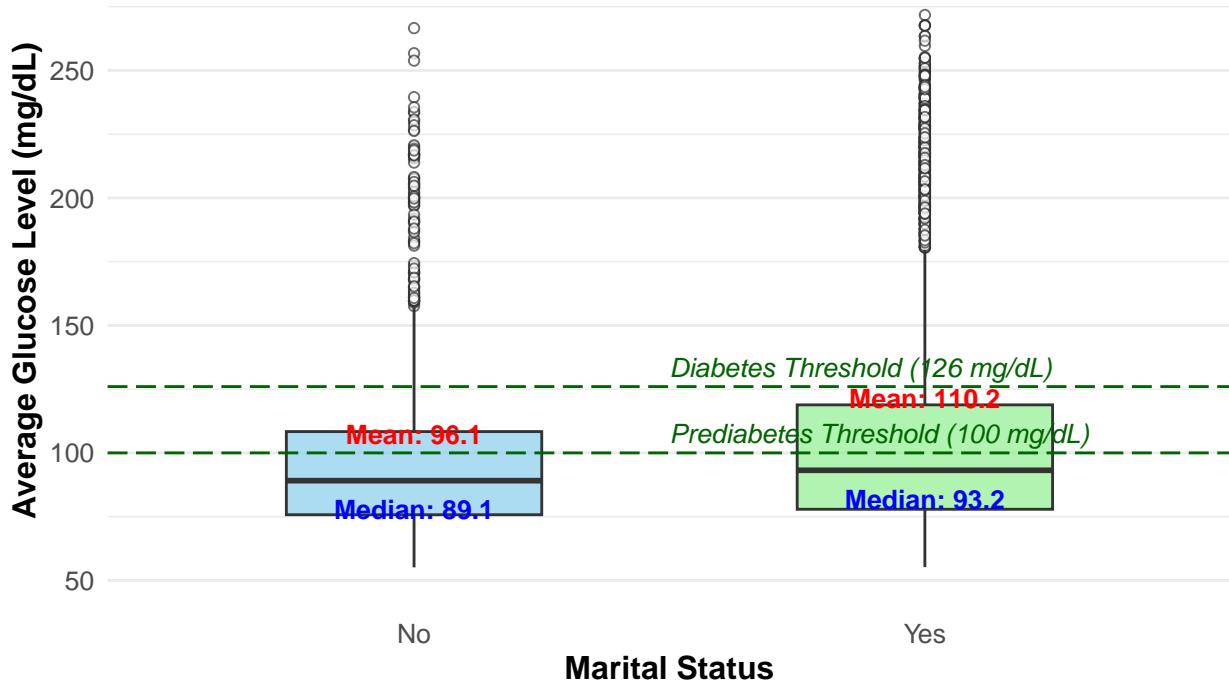
```

geom_text(data = summary_stats, aes(x = ever_married, y = mean_glucose,
                                     label = paste("Mean:", round(mean_glucose, 1))),
           color = "red", vjust = -1, size = 3.5, fontface = "bold") +
geom_text(data = summary_stats, aes(x = ever_married, y = median_glucose,
                                     label = paste("Median:", round(median_glucose, 1))),
           color = "blue", vjust = 2, size = 3.5, fontface = "bold") +
# Annotate with clinical thresholds
annotate("text", x = 1.5, y = c(100, 126),
         label = c("Prediabetes Threshold (100 mg/dL)", "Diabetes Threshold (126 mg/dL)"),
         hjust = 0, vjust = -0.5, color = "darkgreen", size = 3.5, fontface = "italic") +
# Adjust y-axis limits
coord_cartesian(ylim = c(50, max(stroke_data$avg_glucose_level, na.rm = TRUE) * 1.05))

```

Distribution of Average Glucose Levels by Marital Status

With mean, median, and clinical thresholds indicated



We can see that Individuals who are not married have a mean glucose level of 96.1 mg/dL, which is just below the prediabetes threshold of 100 mg/dL, and a median of 89.1 mg/dL, indicating that most individuals in this group have glucose levels in the normal range. For those who are ever married, the mean glucose level is 110.2 mg/dL, which exceeds the prediabetes threshold but is still below the diabetes threshold. The median is 93.2 mg/dL.

```

# Calculate mean and median glucose levels for each group
summary_stats <- stroke_data %>%
  group_by(ever_married, stroke) %>%
  summarise(
    mean_glucose = mean(avg_glucose_level, na.rm = TRUE),

```

```

median_glucose = median(avg_glucose_level, na.rm = TRUE),
  .groups = "drop"
)

# Create the plot
ggplot(stroke_data, aes(x = interaction(stroke, ever_married), y = avg_glucose_level, fill = interaction(stroke, ever_married)))
  geom_violin(trim = FALSE, alpha = 0.7) +
  geom_boxplot(width = 0.1, fill = "white", alpha = 0.5) +
  # Adjust the fill scale to match the exact interaction levels
  scale_fill_manual(values = c("0.No" = "skyblue",
                               "0.Yes" = "lightgreen",
                               "1.No" = "darkblue",
                               "1.Yes" = "darkgreen"),
                     labels = c("Never Married, No Stroke",
                               "Ever Married, No Stroke",
                               "Never Married, Stroke",
                               "Ever Married, Stroke")) +
  # Adjust the x-axis labels to match
  scale_x_discrete(labels = c("Never Married\nNo Stroke",
                               "Ever Married\nNo Stroke",
                               "Never Married\nStroke",
                               "Ever Married\nStroke")) +
  # Add horizontal lines for clinical thresholds
  geom_hline(yintercept = c(100, 126), linetype = "dashed", color = "darkred", size = 0.5) +
  # Add text for mean and median glucose levels
  geom_text(data = summary_stats, aes(x = interaction(stroke, ever_married), y = mean_glucose,
                                       label = paste("Mean:", round(mean_glucose, 1))),
            position = position_dodge(width = 0.75), vjust = -0.5, color = "red", size = 3) +
  geom_text(data = summary_stats, aes(x = interaction(stroke, ever_married), y = median_glucose,
                                       label = paste("Median:", round(median_glucose, 1))),
            position = position_dodge(width = 0.75), vjust = 1.5, color = "blue", size = 3) +
  # Add labels and titles
  labs(
    title = "Distribution of Average Glucose Levels by Stroke Status and Marital Status",
    subtitle = "Violin plot with embedded box plot, mean, median, and clinical thresholds",
    x = "Stroke Status and Marital Status",
    y = "Average Glucose Level (mg/dL)",
    fill = "Group"
  ) +
  # Customize the theme
  theme_minimal() +
  theme(
    plot.title = element_text(face = "bold", size = 14),
    plot.subtitle = element_text(face = "italic", size = 10),
    axis.title = element_text(face = "bold"),
    legend.position = "top",
    axis.text.x = element_text(angle = 45, hjust = 1),
    panel.grid.major.x = element_blank()
)

```

```

) + 

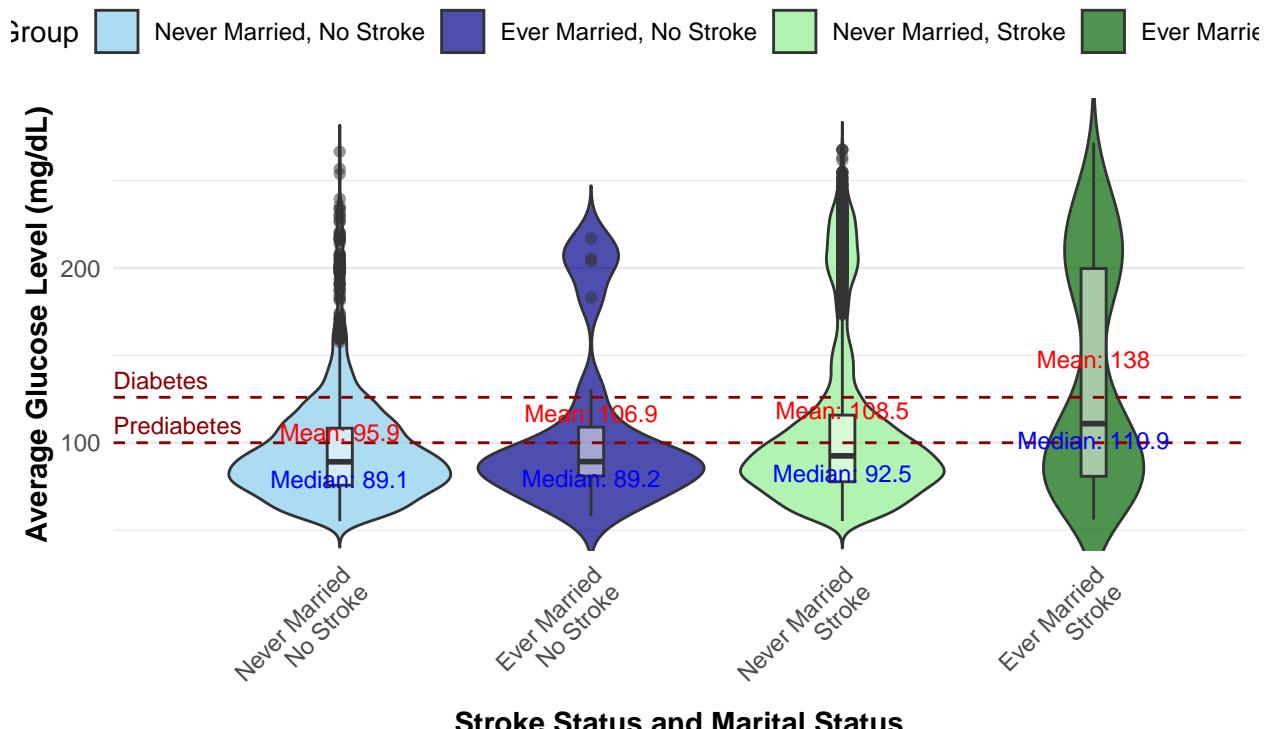
# Annotate with clinical labels
annotate("text", x = 0.1, y = c(100, 126), label = c("Prediabetes", "Diabetes"),
      hjust = 0, vjust = -0.5, color = "darkred", size = 3) + 

# Set y-axis limits
coord_cartesian(ylim = c(50, max(stroke_data$avg_glucose_level, na.rm = TRUE) * 1.05))

```

Distribution of Average Glucose Levels by Stroke Status and Marital Status

Violin plot with embedded box plot, mean, median, and clinical thresholds



1. Never married, no stroke individuals have a mean glucose level of 95.9 mg/dL, just below the prediabetes threshold, and a median of 89.10 mg/dL, suggesting relatively normal glucose levels.
2. Ever married, no stroke individuals have a higher mean glucose level of 106.9 mg/dL, slightly above the prediabetes threshold, and a median of 89.2 mg/dL.
3. Never married, stroke individuals show an elevated mean glucose level of 108.5 mg/dL, also above the prediabetes threshold, and a median of 92.5 mg/dL.
4. Ever married, stroke individuals have the highest mean glucose level of 138 mg/dL, above the diabetes threshold, and a median of 110.9 mg/dL.

Those who have experienced a stroke, regardless of marital status, tend to have higher glucose levels, with the “ever married, stroke” group showing the highest risk, as their glucose levels exceed the diabetes threshold.

```
data_sub_glu_wt = stroke_data %>% group_by(work_type,stroke) %>% summarise(mean = mean(avg_glucose_level))
pander(data_sub_glu_wt)
```

Avg. Glucose levels with work type/Stroke

work_type	stroke	mean
children	0	94.06
children	1	57.93
Govt_job	0	106
Govt_job	1	137.3
Never_worked	0	96.04
Private	0	104.1
Private	1	139.3
Self-employed	0	111.6
Self-employed	1	123.2

1. Children type without a stroke have a mean glucose level of 94.01 mg/dL, while those with a stroke have significantly lower levels, at 57.93.2 mg/dL, likely due to the small sample size in this group.
2. Government job workers without a stroke have a mean glucose level of 105.96 mg/dL, which rises to 137.3 mg/dL for those with a stroke.
3. Individuals who never worked show similar glucose levels to government job workers, with a mean of 96.04 mg/dL for those without a stroke.
4. Private sector workers without a stroke have a mean glucose level of 104.05 mg/dL, while those with a stroke have a much higher mean at 139.32 mg/dL.
5. Self-employed individuals have a mean glucose level of 111.59 mg/dL without a stroke, which increases to 123.20 mg/dL in those who have had a stroke.

This data suggests that individuals in government jobs, the private sector, and self-employed groups with a stroke tend to have the highest glucose levels.

```
# Calculate summary statistics
summary_stats <- stroke_data %>%
  group_by(work_type) %>%
  summarise(
    mean_glucose = mean(avg_glucose_level, na.rm = TRUE),
    median_glucose = median(avg_glucose_level, na.rm = TRUE),
    .groups = "drop"
  )

ggplot(stroke_data, aes(x = work_type, y = avg_glucose_level, fill = work_type)) +
  geom_boxplot(width = 0.7, alpha = 0.7, outlier.shape = 21, outlier.fill = "white") +
  geom_hline(yintercept = c(100, 126), color = "darkgreen", linetype = "longdash", size = 0.5) +
  scale_fill_brewer(palette = "Set3") +
  scale_y_continuous(breaks = seq(0, 300, by = 50)) +
  labs(
    title = "Distribution of Average Glucose Levels by Work Type",
    subtitle = "With mean, median, and clinical thresholds indicated",
```

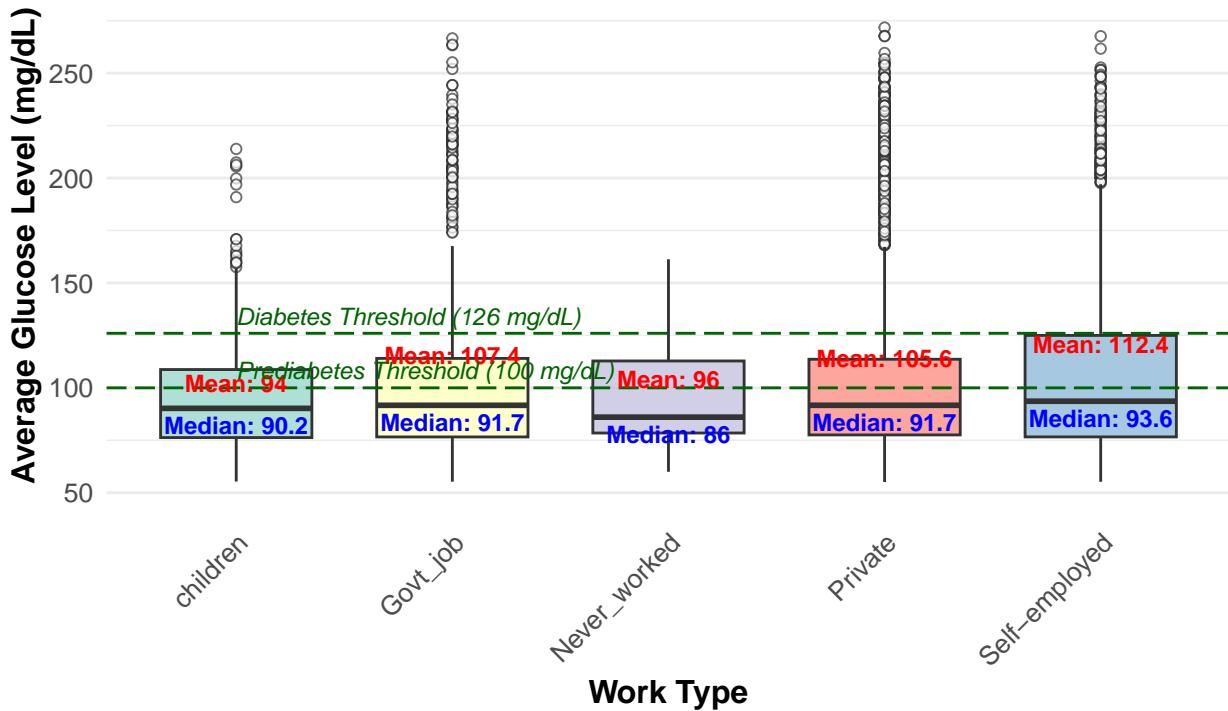
```

x = "Work Type",
y = "Average Glucose Level (mg/dL)",
fill = "Work Type"
) +
theme_minimal() +
theme(
  plot.title = element_text(face = "bold", size = 16),
  plot.subtitle = element_text(face = "italic", size = 12),
  axis.title = element_text(face = "bold", size = 12),
  axis.text = element_text(size = 10),
  axis.text.x = element_text(angle = 45, hjust = 1),
  legend.position = "none",
  panel.grid.major.x = element_blank()
) +
geom_text(data = summary_stats, aes(x = work_type, y = mean_glucose,
                                     label = paste("Mean:", round(mean_glucose, 1))),
           color = "red", vjust = -0.5, size = 3, fontface = "bold") +
geom_text(data = summary_stats, aes(x = work_type, y = median_glucose,
                                     label = paste("Median:", round(median_glucose, 1))),
           color = "blue", vjust = 1.5, size = 3, fontface = "bold") +
annotate("text", x = 1, y = c(100, 126),
         label = c("Prediabetes Threshold (100 mg/dL)", "Diabetes Threshold (126 mg/dL)"),
         hjust = 0, vjust = -0.5, color = "darkgreen", size = 3, fontface = "italic") +
coord_cartesian(ylim = c(50, max(stroke_data$avg_glucose_level, na.rm = TRUE) * 1.05))

```

Distribution of Average Glucose Levels by Work Type

With mean, median, and clinical thresholds indicated



From this plot, we can observe that:

- People whose work involve children have the lowest glucose levels, with a mean of 94 mg/dL and a median of 93 mg/dL.
- Government job workers have a mean glucose level of 107.4 mg/dL, slightly above the prediabetes threshold of 100 mg/dL.
- Those who never worked show a mean glucose level of 96 mg/dL, just below the prediabetes threshold, with a median of 95 mg/dL.
- Private sector workers have a mean glucose level of 105.6 mg/dL, slightly above the prediabetes threshold of 100 mg/dL.
- Self-employed individuals show the highest glucose levels, with a mean of 112.4 mg/dL, well above the prediabetes threshold of 100 mg/dL.

The self-employed and government job workers show elevated glucose levels compared to the other groups, suggesting a higher risk of prediabetes or diabetes.

```
data_sub_glu_rt = stroke_data %>% group_by(Residence_type, stroke) %>% summarise(mean = mean(avg_glucose_level))
pander(data_sub_glu_rt)
```

Avg. Glucose levels with resident type/Stroke

Residence_type	stroke	mean
Rural	0	104.5
Rural	1	132.5
Urban	0	103.5
Urban	1	136.5

This tells us that in rural areas, individuals without a stroke have a mean glucose level of 104.47 mg/dL, while those with a stroke show significantly higher glucose levels at 132.51 mg/dL.

For urban residents, those without a stroke have a mean glucose level of 103.53 mg/dL, similar to rural residents without a stroke. However, urban individuals with a stroke exhibit the highest mean glucose levels at 136.46 mg/dL.

```
# Calculate summary statistics
summary_stats <- stroke_data %>%
  group_by(Residence_type) %>%
  summarise(
    mean_glucose = mean(avg_glucose_level, na.rm = TRUE),
    median_glucose = median(avg_glucose_level, na.rm = TRUE),
    .groups = "drop"
  )

ggplot(stroke_data, aes(x = Residence_type, y = avg_glucose_level, fill = Residence_type)) +
  geom_boxplot(width = 0.7, alpha = 0.7, outlier.shape = 21, outlier.fill = "white") +
  geom_hline(yintercept = c(100, 126), color = "darkgreen", linetype = "longdash", size = 0.5) +
  scale_fill_manual(values = c("Rural" = "forestgreen", "Urban" = "orange")) +
  scale_y_continuous(breaks = seq(0, 300, by = 50)) +
  labs(
    title = "Distribution of Average Glucose Levels by Residence Type",
    subtitle = "With mean, median, and clinical thresholds indicated",
    x = "Residence Type",
    y = "Average Glucose Level (mg/dL)",
    fill = "Residence Type"
  ) +
  theme_minimal() +
```

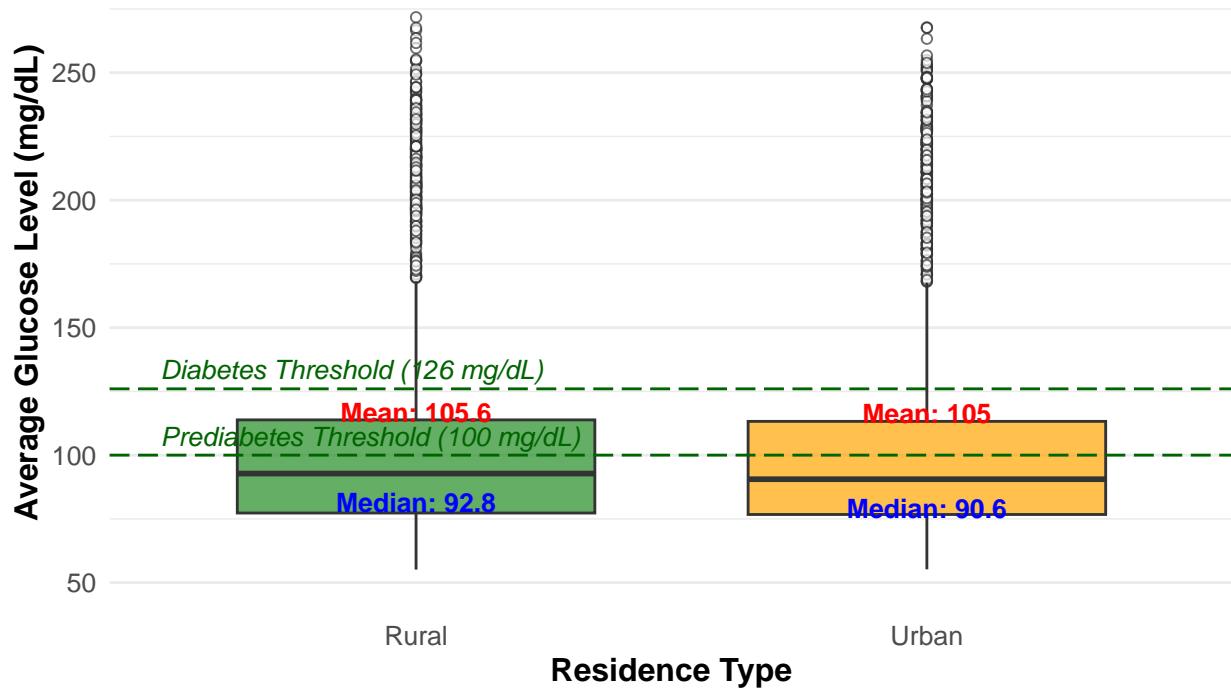
```

theme(
  plot.title = element_text(face = "bold", size = 16),
  plot.subtitle = element_text(face = "italic", size = 12),
  axis.title = element_text(face = "bold", size = 12),
  axis.text = element_text(size = 10),
  legend.position = "none",
  panel.grid.major.x = element_blank()
) +
  geom_text(data = summary_stats, aes(x = Residence_type, y = mean_glucose,
                                       label = paste("Mean:", round(mean_glucose, 1))),
            color = "red", vjust = -1, size = 3.5, fontface = "bold") +
  geom_text(data = summary_stats, aes(x = Residence_type, y = median_glucose,
                                       label = paste("Median:", round(median_glucose, 1))),
            color = "blue", vjust = 2, size = 3.5, fontface = "bold") +
  annotate("text", x = 0.5, y = c(100, 126),
           label = c("Prediabetes Threshold (100 mg/dL)", "Diabetes Threshold (126 mg/dL)"),
           hjust = 0, vjust = -0.5, color = "darkgreen", size = 3.5, fontface = "italic") +
  coord_cartesian(ylim = c(50, max(stroke_data$avg_glucose_level, na.rm = TRUE) * 1.05))

```

Distribution of Average Glucose Levels by Residence Type

With mean, median, and clinical thresholds indicated



For rural residents, the mean glucose level is 105.6 mg/dL, slightly above the prediabetes threshold, while the median is 92.8 mg/dL, suggesting that the majority of individuals in this group have normal glucose levels, but some outliers raise the mean.

For urban residents, the mean glucose level is 105 mg/dL, very similar to the rural group, and also above the prediabetes threshold. The median glucose level is 90.6 mg/dL, which is slightly lower than the rural group's median.

```

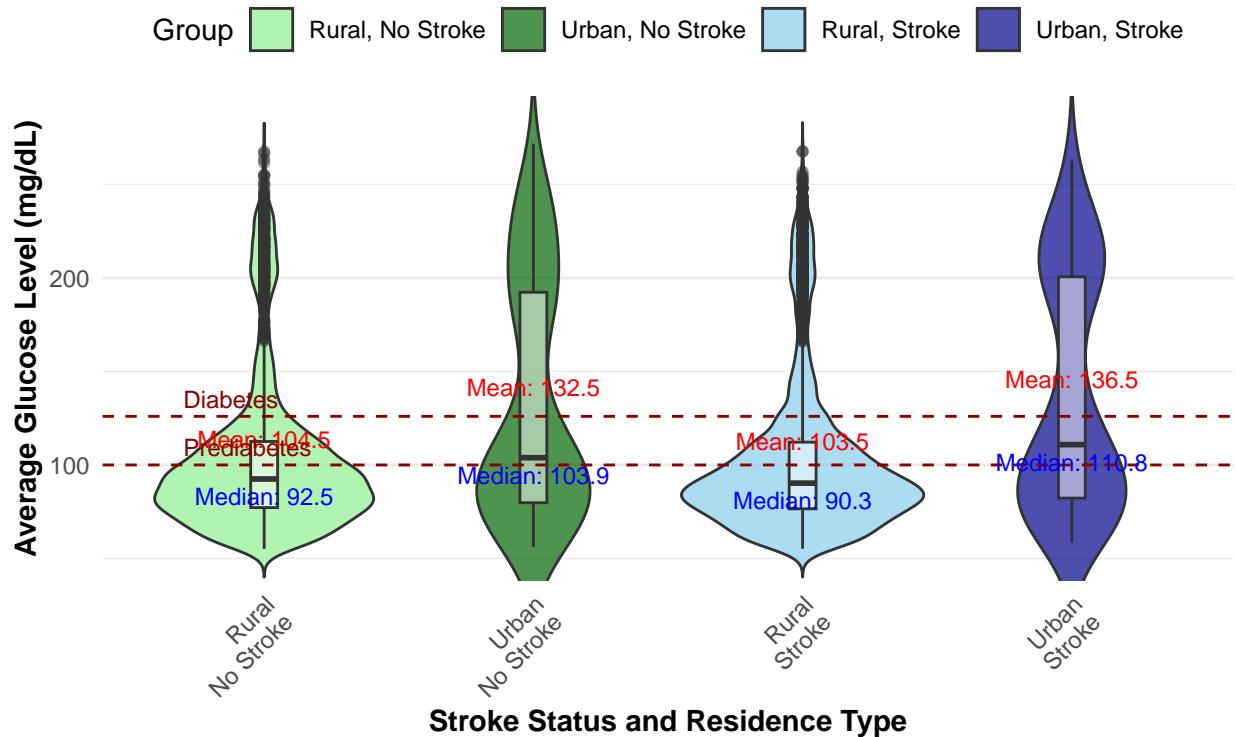
# Calculate mean and median glucose levels for each group
summary_stats <- stroke_data %>%
  group_by(Residence_type, stroke) %>%
  summarise(
    mean_glucose = mean(avg_glucose_level, na.rm = TRUE),
    median_glucose = median(avg_glucose_level, na.rm = TRUE),
    .groups = "drop"
  )

# Create the plot
# Adjust the fill scale to match the exact interaction levels
ggplot(stroke_data, aes(x = interaction(stroke, Residence_type), y = avg_glucose_level, fill = interact
  geom_violin(trim = FALSE, alpha = 0.7) +
  geom_boxplot(width = 0.1, fill = "white", alpha = 0.5) +
  scale_fill_manual(values = c("0.Rural" = "lightgreen",
                               "0.Urban" = "skyblue",
                               "1.Rural" = "darkgreen",
                               "1.Urban" = "darkblue"),
                     labels = c("Rural, No Stroke", "Urban, No Stroke",
                               "Rural, Stroke", "Urban, Stroke")) +
  scale_x_discrete(labels = c("Rural\nNo Stroke", "Urban\nNo Stroke",
                               "Rural\nStroke", "Urban\nStroke")) +
  geom_hline(yintercept = c(100, 126), linetype = "dashed", color = "darkred", size = 0.5) +
  geom_text(data = summary_stats, aes(x = interaction(stroke, Residence_type), y = mean_glucose,
                                       label = paste("Mean:", round(mean_glucose, 1))),
            position = position_dodge(width = 0.75), vjust = -0.5, color = "red", size = 3) +
  geom_text(data = summary_stats, aes(x = interaction(stroke, Residence_type), y = median_glucose,
                                       label = paste("Median:", round(median_glucose, 1))),
            position = position_dodge(width = 0.75), vjust = 1.5, color = "blue", size = 3) +
  labs(
    title = "Distribution of Average Glucose Levels by Stroke Status and Residence Type",
    subtitle = "Violin plot with embedded box plot, mean, median, and clinical thresholds",
    x = "Stroke Status and Residence Type",
    y = "Average Glucose Level (mg/dL)",
    fill = "Group"
  ) +
  theme_minimal() +
  theme(
    plot.title = element_text(face = "bold", size = 14),
    plot.subtitle = element_text(face = "italic", size = 10),
    axis.title = element_text(face = "bold"),
    legend.position = "top",
    axis.text.x = element_text(angle = 45, hjust = 1),
    panel.grid.major.x = element_blank()
  ) +
  annotate("text", x = 0.7, y = c(100, 126), label = c("Prediabetes", "Diabetes"),
           hjust = 0, vjust = -0.5, color = "darkred", size = 3) +
  coord_cartesian(ylim = c(50, max(stroke_data$avg_glucose_level, na.rm = TRUE) * 1.05))

```

Distribution of Average Glucose Levels by Stroke Status and Residence Type

Violin plot with embedded box plot, mean, median, and clinical thresholds



This plot shows us that individuals living in rural areas without a stroke have a mean glucose level of 104.5 mg/dL, just above the prediabetes threshold, while their median glucose level is lower at 92.5 mg/dL. Urban residents without a stroke show higher glucose levels, with a mean of 132.5 mg/dL, which is well above the diabetes threshold, and a median of 103.9 mg/dL.

Among those with a stroke, rural residents have a mean glucose level of 103.5 mg/dL and a median of 90.3 mg/dL, suggesting slightly lower glucose levels compared to those without a stroke. In contrast, urban residents with a stroke exhibit the highest glucose levels, with a mean of 136.5 mg/dL and a median of 110.8 mg/dL, both above the diabetes threshold.

```
data_sub_glu_st = stroke_data %>% group_by(smoking_status,stroke) %>% summarise(mean = mean(avg_glucose))
pander(data_sub_glu_st)
```

Avg. Glucose levels with smoking status/Stroke

smoking_status	stroke	mean
formerly smoked	0	110.6
formerly smoked	1	138.7
never smoked	0	105.9
never smoked	1	133.3
smokes	0	104.5
smokes	1	141.4
Unknown	0	97.89

smoking_status	stroke	mean
Unknown	1	120.9

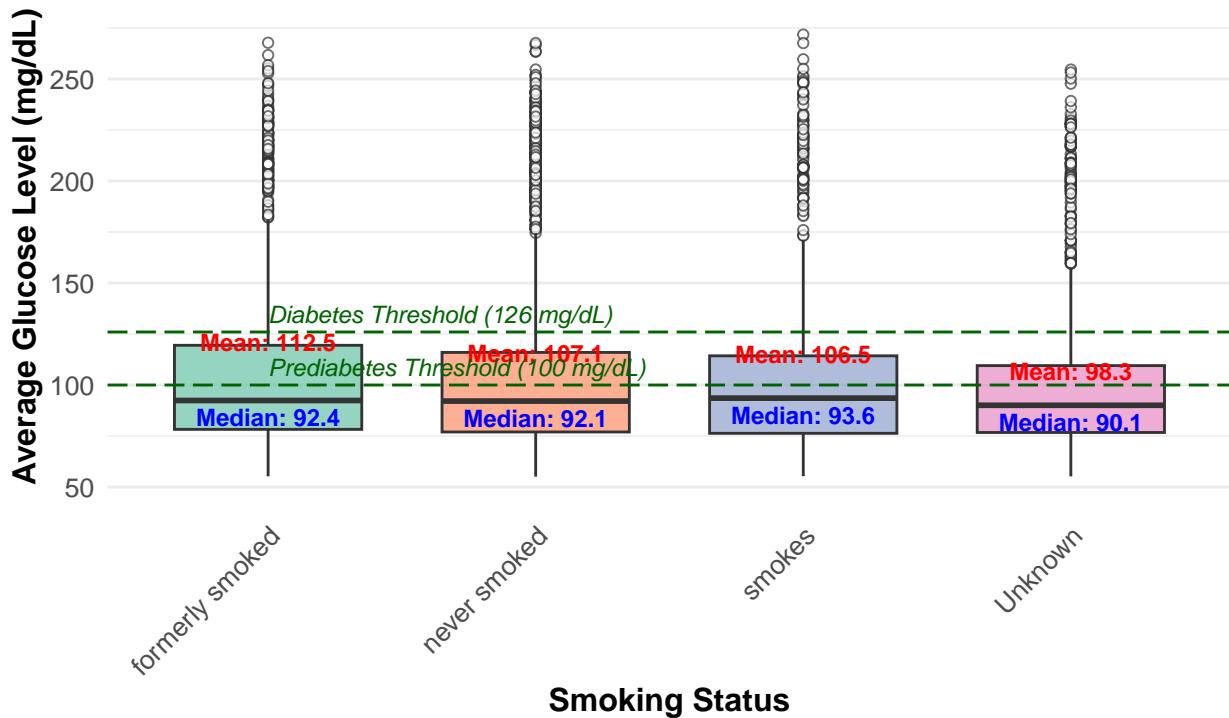
We can see that across all groups, individuals who have had a stroke tend to have higher glucose levels than those without a stroke. For example, former smokers with a stroke have higher glucose levels than those without a stroke, and a similar pattern is observed for those who never smoked or currently smoke. The highest glucose levels are seen in smokers who had a stroke, while those with an “unknown” smoking status and no stroke have the lowest glucose levels.

```
# Calculate summary statistics
summary_stats <- stroke_data %>%
  group_by(smoking_status) %>%
  summarise(
    mean_glucose = mean(avg_glucose_level, na.rm = TRUE),
    median_glucose = median(avg_glucose_level, na.rm = TRUE),
    .groups = "drop"
  )

ggplot(stroke_data, aes(x = smoking_status, y = avg_glucose_level, fill = smoking_status)) +
  geom_boxplot(width = 0.7, alpha = 0.7, outlier.shape = 21, outlier.fill = "white") +
  geom_hline(yintercept = c(100, 126), color = "darkgreen", linetype = "longdash", size = 0.5) +
  scale_fill_brewer(palette = "Set2") +
  scale_y_continuous(breaks = seq(0, 300, by = 50)) +
  labs(
    title = "Distribution of Average Glucose Levels by Smoking Status",
    subtitle = "With mean, median, and clinical thresholds indicated",
    x = "Smoking Status",
    y = "Average Glucose Level (mg/dL)",
    fill = "Smoking Status"
  ) +
  theme_minimal() +
  theme(
    plot.title = element_text(face = "bold", size = 16),
    plot.subtitle = element_text(face = "italic", size = 12),
    axis.title = element_text(face = "bold", size = 12),
    axis.text = element_text(size = 10),
    axis.text.x = element_text(angle = 45, hjust = 1),
    legend.position = "none",
    panel.grid.major.x = element_blank()
  ) +
  geom_text(data = summary_stats, aes(x = smoking_status, y = mean_glucose,
                                       label = paste("Mean:", round(mean_glucose, 1))),
            color = "red", vjust = -0.5, size = 3, fontface = "bold") +
  geom_text(data = summary_stats, aes(x = smoking_status, y = median_glucose,
                                       label = paste("Median:", round(median_glucose, 1))),
            color = "blue", vjust = 1.5, size = 3, fontface = "bold") +
  annotate("text", x = 1, y = c(100, 126),
           label = c("Prediabetes Threshold (100 mg/dL)", "Diabetes Threshold (126 mg/dL)"),
           hjust = 0, vjust = -0.5, color = "darkgreen", size = 3, fontface = "italic") +
  coord_cartesian(ylim = c(50, max(stroke_data$avg_glucose_level, na.rm = TRUE) * 1.05))
```

Distribution of Average Glucose Levels by Smoking Stat

With mean, median, and clinical thresholds indicated



For individuals who formerly smoked, the mean glucose level is 112.5 mg/dL, which is above the prediabetes threshold but below the diabetes threshold. The median is 92.4 mg/dL, indicating that many individuals in this group have glucose levels within a normal range, but some outliers raise the mean.

In the “never smoked” group, the mean glucose level is 107.1 mg/dL, still above the prediabetes threshold, with a median of 92.1 mg/dL. Similar to the “formerly smoked” group, most people have normal glucose levels, but the mean is influenced by higher values.

For current smokers, the mean glucose level is 106.5 mg/dL, with a median of 93.6 mg/dL. This group has slightly higher glucose levels compared to those who never smoked, again with outliers pushing the mean into the prediabetes range.

Individuals with an “unknown” smoking status have a mean glucose level of 98.3 mg/dL, below the prediabetes threshold, and a median of 90.1 mg/dL. This group appears to have lower glucose levels compared to the other categories, with both the mean and median closer to normal levels.

Age

The last variable that we consider is age.

```
summary(stroke_data$age)
```

```
##      Min. 1st Qu. Median      Mean 3rd Qu.      Max.
##      0.08   25.00  44.00    42.87   60.00   82.00
```

We can see that 1. The youngest individual is 0.08 years old (approximately 1 month). 2. 25% of the individuals are younger than 25 years. 3. The median age is 44 years, meaning that half of the individuals

are younger than 45 and half are older. 4. The average age is 42.87 years, slightly lower than the median, indicating a small skew towards younger individuals. 5. 75% of the individuals are younger than 60 years. 6. The oldest individual is 82 years old.

This distribution suggests that the age variable is fairly spread out, with a slight skew towards younger ages. The majority of individuals are between 25 and 61 years old.

```
# Calculate frequency statistics for age
cat("More Statistics for Age:\n")
```

```
## More Statistics for Age:
```

```
age_stats = agricolae::stat.freq(a)
age_stats
```

```
## $variance
## [1] 508.1025
##
## $mean
## [1] 42.40831
##
## $median
## [1] 43.73278
##
## $mode
##      [- -] mode
## [1,] 50 55 52.98165
```

```
cat("Skewness:", skewness(stroke_data$age), "\n")
```

```
## Skewness: -0.1194497
```

```
cat("Kurtosis:", kurtosis(stroke_data$age), "\n")
```

```
## Kurtosis: -0.9880331
```

The most frequently occurring age is around 53 years. The skewness is -0.120, which is slightly negative. This suggests that the age distribution is slightly skewed to the left, indicating that there are slightly more older individuals pulling the distribution in that direction. The kurtosis is -0.99, this means the age distribution has lighter tails and a flatter peak compared to a normal distribution, suggesting fewer extreme values.

```
# Calculate summary statistics
age_stats <- stroke_data %>%
  summarise(
    mean_age = mean(age, na.rm = TRUE),
    median_age = median(age, na.rm = TRUE),
    sd_age = sd(age, na.rm = TRUE)
  )

age_stats = agricolae::stat.freq(a)
```

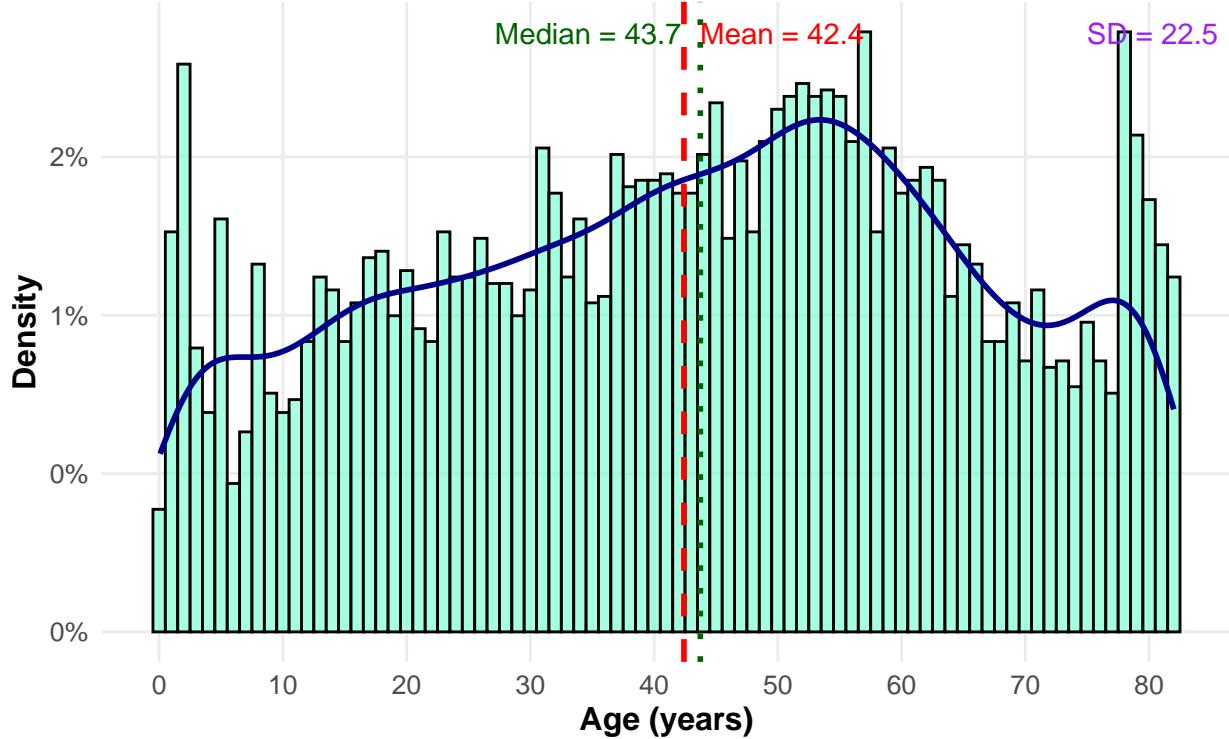
```

ggplot(stroke_data, aes(x = age)) +
  # Histogram
  geom_histogram(aes(y = ..density..), binwidth = 1, fill = "aquamarine", color = "black", alpha = 0.7)
  # Density plot
  geom_density(color = "darkblue", size = 1) +
  # Mean line
  geom_vline(aes(xintercept = age_stats$mean), color = "red", linetype = "dashed", size = 1) +
  # Median line
  geom_vline(aes(xintercept = age_stats$median), color = "darkgreen", linetype = "dotted", size = 1) +
  # Annotations
  annotate("text", x = age_stats$mean, y = Inf, label = paste("Mean =", round(age_stats$mean, 1)),
           color = "red", vjust = 2, hjust = -0.1, size = 4) +
  annotate("text", x = age_stats$median, y = Inf, label = paste("Median =", round(age_stats$median, 1)),
           color = "darkgreen", vjust = 2, hjust = 1.1, size = 4) +
  annotate("text", x = Inf, y = Inf,
           label = paste("SD =", round(sqrt(age_stats$variance), 1)),
           color = "purple", vjust = 2, hjust = 1.1, size = 4) +
  # Labels and title
  labs(
    title = "Age Distribution",
    subtitle = "Histogram with Density Overlay, Mean, and Median",
    x = "Age (years)",
    y = "Density"
  ) +
  # Theme customization
  theme_minimal() +
  theme(
    plot.title = element_text(face = "bold", size = 16),
    plot.subtitle = element_text(face = "italic", size = 12),
    axis.title = element_text(face = "bold", size = 12),
    axis.text = element_text(size = 10),
    panel.grid.minor = element_blank()
  ) +
  # Scale customization
  scale_x_continuous(breaks = seq(0, max(stroke_data$age, na.rm = TRUE), by = 10)) +
  scale_y_continuous(labels = scales::percent_format(accuracy = 1))

```

Age Distribution

Histogram with Density Overlay, Mean, and Median



This density plot confirms our findings from earlier tables. This plot shows multiple peaks, with significant frequencies at various points such as early childhood (around 0-5 years) and older age groups (around 70-80 years). This multimodal characteristic is not typical of a normal distribution, which should be unimodal and bell-shaped. In addition, it has asymmetry, particularly with a longer tail on the right and The dispersion appears broader than what we would expect in a normal distribution, with a wide range of ages and a higher frequency in the older age groups.

```
# Calculate summary statistics
age_stats <- stroke_data %>%
  summarise(
    mean_age = mean(age, na.rm = TRUE),
    median_age = median(age, na.rm = TRUE),
    q1_age = quantile(age, 0.25, na.rm = TRUE),
    q3_age = quantile(age, 0.75, na.rm = TRUE)
  )

# Create the boxplot
ggplot(stroke_data, aes(x = "Age", y = age)) +
  geom_boxplot(fill = "lightblue", color = "darkblue", alpha = 0.7, width = 0.5) +
  geom_jitter(width = 0.2, alpha = 0.1, color = "darkgrey") +
  stat_summary(fun = mean, geom = "point", shape = 23, size = 3, fill = "white") +
  geom_text(data = age_stats, aes(x = "Age", y = mean_age, label = paste("Mean:", round(mean_age, 1))), color = "red", vjust = -1, hjust = -0.1, size = 3.5) +
  geom_text(data = age_stats, aes(x = "Age", y = median_age, label = paste("Median:", median_age)), color = "blue", vjust = 2, hjust = -0.1, size = 3.5) +
  geom_text(data = age_stats, aes(x = "Age", y = q1_age, label = paste("Q1:", q1_age)),
```

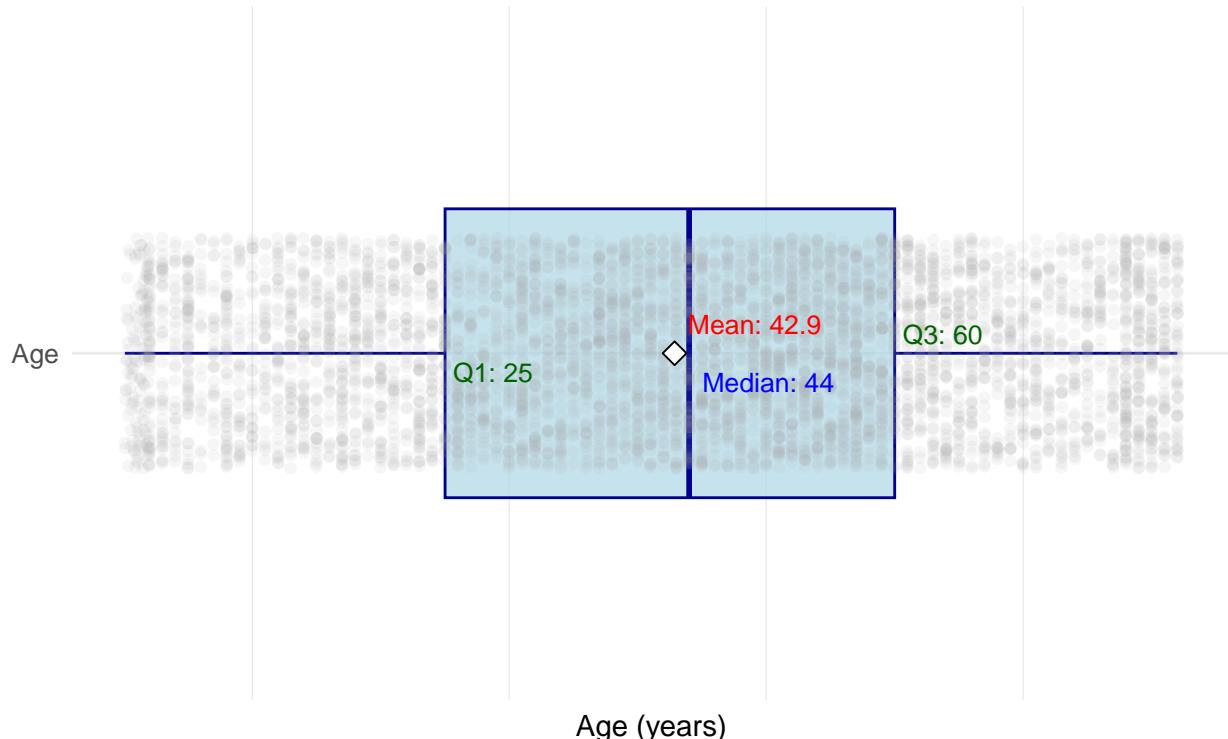
```

        color = "darkgreen", vjust = 1.5, hjust = -0.1, size = 3.5) +
  geom_text(data = age_stats, aes(x = "Age", y = q3_age, label = paste("Q3:", q3_age)),
            color = "darkgreen", vjust = -0.5, hjust = -0.1, size = 3.5) +
  labs(
    title = "Distribution of Age",
    subtitle = "Boxplot with jittered points and summary statistics",
    y = "Age (years)",
    x = NULL
  ) +
  theme_minimal() +
  theme(
    plot.title = element_text(face = "bold", size = 16),
    plot.subtitle = element_text(face = "italic", size = 12),
    axis.title.y = element_text(face = "bold", size = 12),
    axis.text = element_text(size = 10),
    panel.grid.major.x = element_blank(),
    axis.text.x = element_blank(),
    axis.ticks.x = element_blank()
  ) +
  coord_flip()

```

Distribution of Age

Boxplot with jittered points and summary statistics



As expected, The boxplot shows moderate spread in the data, with a few potential outliers on both sides of the whiskers. This confirms a wide age distribution, with a slight skew to the left.

```
# Calculate summary statistics
age_stats <- stroke_data %>%
```

```

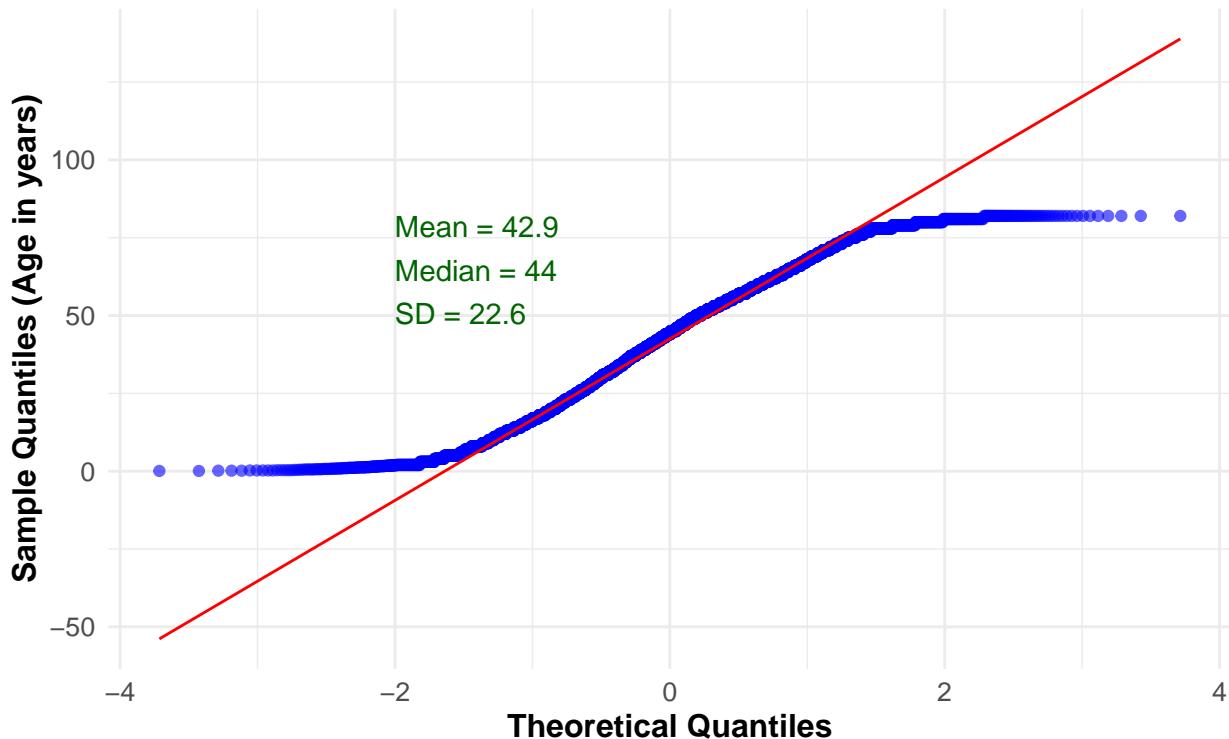
summarise(
  mean_age = mean(age, na.rm = TRUE),
  median_age = median(age, na.rm = TRUE),
  sd_age = sd(age, na.rm = TRUE)
)

# Create the Q-Q plot
ggplot(stroke_data, aes(sample = age)) +
  stat_qq(color = "blue", alpha = 0.6) +
  stat_qq_line(color = "red") +
  labs(
    title = "Q-Q Plot of Age Distribution",
    subtitle = "Assessing normality of age distribution",
    x = "Theoretical Quantiles",
    y = "Sample Quantiles (Age in years)"
  ) +
  theme_minimal() +
  theme(
    plot.title = element_text(face = "bold", size = 16),
    plot.subtitle = element_text(face = "italic", size = 12),
    axis.title = element_text(face = "bold", size = 12),
    axis.text = element_text(size = 10)
  ) +
  annotate("text", x = -2, y = max(stroke_data$age, na.rm = TRUE),
           label = paste("Mean =", round(age_stats$mean_age, 1),
                         "\nMedian =", round(age_stats$median_age, 1),
                         "\nSD =", round(age_stats$sd_age, 1)),
           hjust = 0, vjust = 1, size = 4, color = "darkgreen")

```

Q-Q Plot of Age Distribution

Assessing normality of age distribution



From the Q-Q plot we can see that there are deviations at the tails with standard deviation of 22.6 years. Which suggests a spread with some outliers, particularly in the higher age range.

```
data_sub_age = stroke_data %>% group_by(stroke) %>% summarise(Age_mean = mean(age), .groups = "drop")
pander(data_sub_age)
```

Age levels with Stroke

stroke	Age_mean
0	41.76
1	67.71

For those who have not had a stroke, the mean age is 41.8 years, indicating that the population without stroke tends to be younger. In contrast, individuals who have experienced a stroke are significantly older, with a mean age of 67.7 years. This data suggests that stroke occurrence is more common among older individuals, while younger people are less likely to have had a stroke.

```
# Calculate summary statistics
summary_stats <- stroke_data %>%
  group_by(stroke) %>%
  summarise(
    mean_age = mean(age, na.rm = TRUE),
```

```

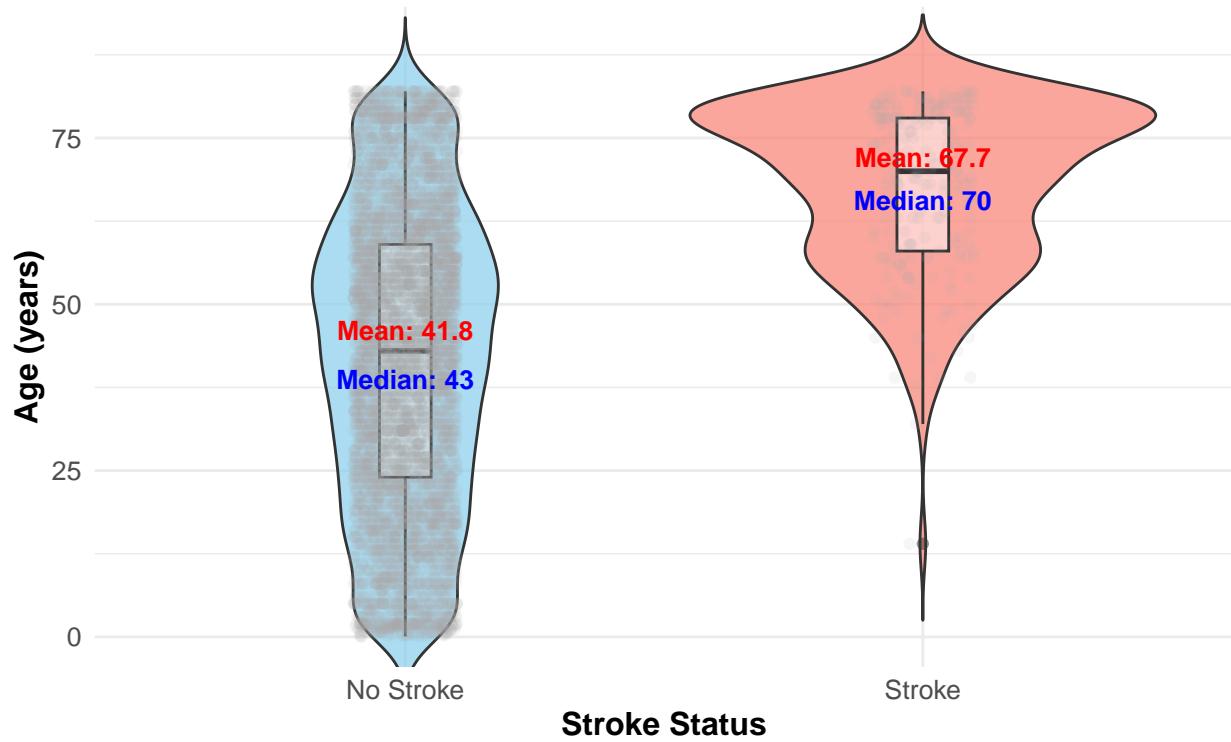
median_age = median(age, na.rm = TRUE),
  .groups = "drop"
)

# Create the violin plot
ggplot(stroke_data, aes(x = as.factor(stroke), y = age, fill = as.factor(stroke))) +
  geom_violin(trim = FALSE, alpha = 0.7) +
  geom_boxplot(width = 0.1, fill = "white", alpha = 0.5) +
  geom_jitter(width = 0.1, alpha = 0.1, color = "darkgrey") +
  scale_fill_manual(values = c("0" = "skyblue", "1" = "salmon")) +
  labs(
    title = "Age Distribution by Stroke Occurrence",
    subtitle = "Violin plot with embedded box plot, mean, median, and jittered points",
    x = "Stroke Status",
    y = "Age (years)",
    fill = "Stroke Status"
  ) +
  theme_minimal() +
  theme(
    plot.title = element_text(face = "bold", size = 16),
    plot.subtitle = element_text(face = "italic", size = 12),
    axis.title = element_text(face = "bold", size = 12),
    axis.text = element_text(size = 10),
    legend.position = "none"
  ) +
  geom_text(data = summary_stats,
            aes(x = as.factor(stroke), y = mean_age,
                 label = paste("Mean:", round(mean_age, 1))),
            color = "red", vjust = -1, size = 3.5, fontface = "bold") +
  geom_text(data = summary_stats,
            aes(x = as.factor(stroke), y = median_age,
                 label = paste("Median:", round(median_age, 1))),
            color = "blue", vjust = 2, size = 3.5, fontface = "bold") +
  coord_cartesian(ylim = c(0, max(stroke_data$age, na.rm = TRUE) * 1.1)) +
  scale_x_discrete(labels = c("0" = "No Stroke", "1" = "Stroke"))

```

Age Distribution by Stroke Occurrence

Violin plot with embedded box plot, mean, median, and jittered points



The violin plot confirms that for individuals who have not had a stroke, the mean age is 41.8 years and the median is 43 years, with a relatively wide distribution, suggesting that people of various ages may not experience strokes. In contrast, the group that has experienced a stroke is significantly older, with a mean age of 67.7 years and a median of 70 years. The distribution is more concentrated in the older age range, indicating that strokes are much more common among older individuals.

```
data_sub_age = stroke_data %>% group_by(heart_disease, stroke) %>% summarise(Age_mean = mean(age), .group = TRUE)
pander(data_sub_age)
```

Age levels with Heart Disease/Stroke

heart_disease	stroke	Age_mean
0	0	40.6
0	1	66.71
1	0	67.55
1	1	71.95

For those without heart disease and no stroke, the mean age is 40.6 years, whereas for those with no heart disease but with a stroke, the mean age is significantly higher at 66.7 years. Among individuals with heart disease, those without a stroke have a mean age of 67.6 years, while those with both heart disease and stroke have the highest mean age of 72 years. This suggests that both heart disease and stroke are more prevalent among older individuals.

```

# Define labels for heart disease
heart_disease_labels <- c("0" = "No Heart Disease", "1" = "Heart Disease")

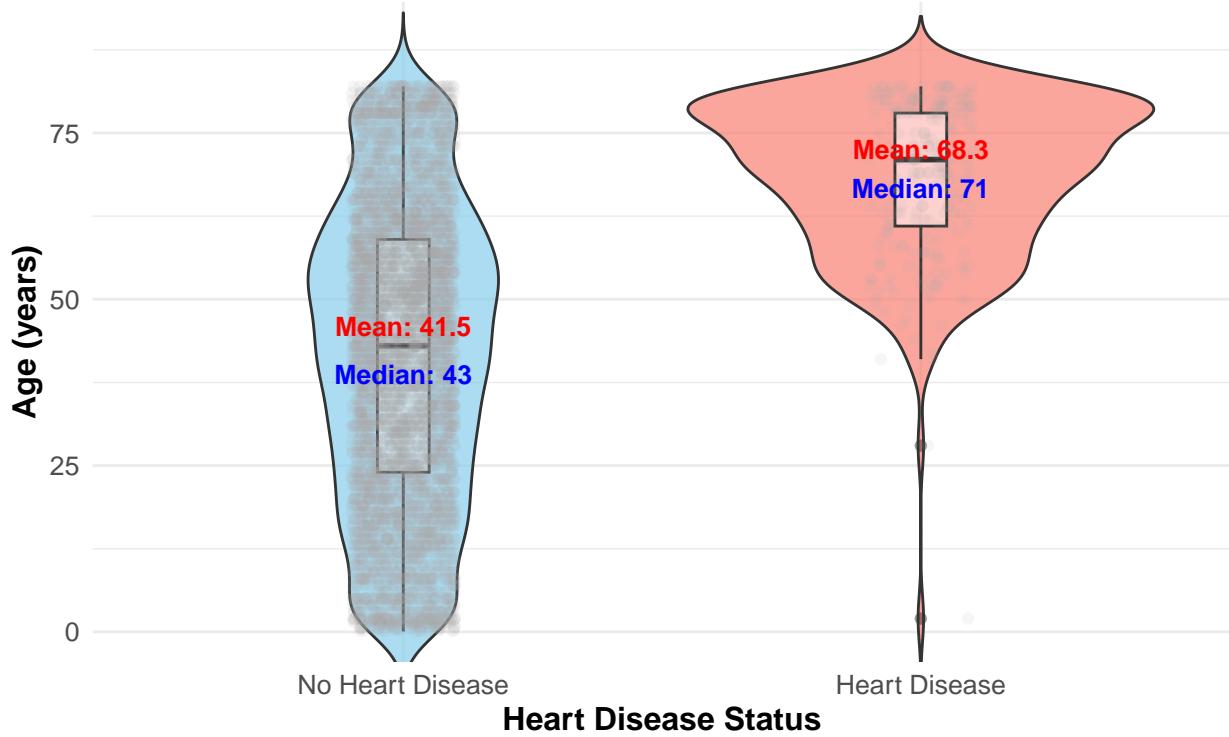
# Calculate summary statistics
summary_stats <- stroke_data %>%
  group_by(heart_disease) %>%
  summarise(
    mean_age = mean(age, na.rm = TRUE),
    median_age = median(age, na.rm = TRUE),
    .groups = "drop"
  )

# Create the violin plot
ggplot(stroke_data, aes(x = heart_disease, y = age, fill = heart_disease)) +
  geom_violin(trim = FALSE, alpha = 0.7) +
  geom_boxplot(width = 0.1, fill = "white", alpha = 0.5) +
  geom_jitter(width = 0.1, alpha = 0.1, color = "darkgrey") +
  scale_fill_manual(values = c("0" = "skyblue", "1" = "salmon"),
                     labels = heart_disease_labels) +
  scale_x_discrete(labels = heart_disease_labels) +
  labs(
    title = "Age Distribution by Heart Disease Status",
    subtitle = "Violin plot with embedded box plot, mean, median, and jittered points",
    x = "Heart Disease Status",
    y = "Age (years)",
    fill = "Heart Disease Status"
  ) +
  theme_minimal() +
  theme(
    plot.title = element_text(face = "bold", size = 16),
    plot.subtitle = element_text(face = "italic", size = 12),
    axis.title = element_text(face = "bold", size = 12),
    axis.text = element_text(size = 10),
    legend.position = "none"
  ) +
  geom_text(data = summary_stats,
            aes(x = heart_disease, y = mean_age,
                 label = paste("Mean:", round(mean_age, 1))),
            color = "red", vjust = -1, size = 3.5, fontface = "bold") +
  geom_text(data = summary_stats,
            aes(x = heart_disease, y = median_age,
                 label = paste("Median:", round(median_age, 1))),
            color = "blue", vjust = 2, size = 3.5, fontface = "bold") +
  coord_cartesian(ylim = c(0, max(stroke_data$age, na.rm = TRUE) * 1.1))

```

Age Distribution by Heart Disease Status

Violin plot with embedded box plot, mean, median, and jittered points



For those without heart disease, the mean age is 41.5 years, and the median is 43 years, with a broader spread of ages. In contrast, individuals with heart disease have a much higher mean age of 68.3 years and a median of 71 years, with a tighter distribution around the higher age range. The plot highlights that heart disease is more prevalent among older individuals, with a visible difference in the central tendency between the two groups.

```
# Define labels for stroke and heart disease
stroke_labels <- c("0" = "No Stroke", "1" = "Stroke")
heart_disease_labels <- c("0" = "No Heart Disease", "1" = "Heart Disease")

# Calculate summary statistics
summary_stats <- stroke_data %>%
  group_by(heart_disease, stroke) %>%
  summarise(
    mean_age = mean(age, na.rm = TRUE),
    median_age = median(age, na.rm = TRUE),
    .groups = "drop"
  )

# Create the boxplot
ggplot(stroke_data, aes(x = stroke, y = age, fill = stroke)) +
  geom_boxplot(alpha = 0.7) +
  geom_jitter(width = 0.2, alpha = 0.1, color = "darkgrey") +
  facet_wrap(~heart_disease, labeller = as_labeller(heart_disease_labels)) +
  scale_fill_manual(values = c("0" = "skyblue", "1" = "salmon"), labels = stroke_labels) +
  scale_x_discrete(labels = stroke_labels) +
```

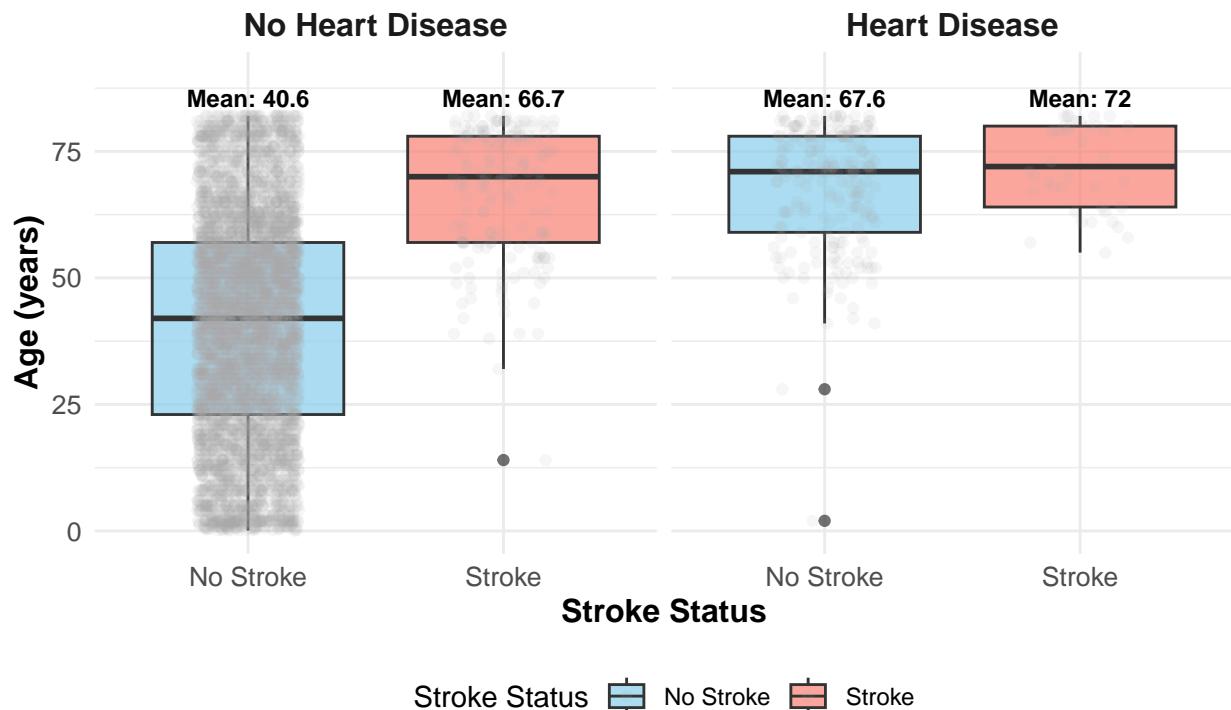
```

  labs(
    title = "Age Distribution by Stroke Status and Heart Disease",
    subtitle = "Boxplots with jittered points",
    x = "Stroke Status",
    y = "Age (years)",
    fill = "Stroke Status"
  ) +
  theme_minimal() +
  theme(
    plot.title = element_text(face = "bold", size = 16),
    plot.subtitle = element_text(face = "italic", size = 12),
    axis.title = element_text(face = "bold", size = 12),
    axis.text = element_text(size = 10),
    legend.position = "bottom",
    strip.text = element_text(face = "bold", size = 12)
  ) +
  geom_text(data = summary_stats,
            aes(x = stroke, y = max(stroke_data$age, na.rm = TRUE),
                label = paste("Mean:", round(mean_age, 1))),
            vjust = -0.5, size = 3, fontface = "bold") +
  coord_cartesian(ylim = c(0, max(stroke_data$age, na.rm = TRUE) * 1.1))

```

Age Distribution by Stroke Status and Heart Disease

Boxplots with jittered points



Individuals without heart disease and no stroke have the lowest mean age (40.6 years), while those with both stroke and heart disease have the highest mean (72 years). There is a clear trend of older age in individuals with stroke, regardless of heart disease status. Additionally, the spread of ages is more significant in the no-stroke group for those without heart disease, indicating greater variability in this population.

```
data_sub_age_gender = stroke_data %>% group_by(smoking_status, stroke) %>% summarise(Age_mean = mean(age))
pander(data_sub_age_gender)
```

Age levels with Smoking Status/Stroke

smoking_status	stroke	Age_mean
formerly smoked	0	53.98
formerly smoked	1	68.35
never smoked	0	45.33
never smoked	1	70.49
smokes	0	46.13
smokes	1	62.38
Unknown	0	28.79
Unknown	1	65.59

We can see that former smokers tend to be older on average, with those who experienced a stroke being significantly older (68.4 years) than those without a stroke (54 years). Similarly, individuals who never smoked but had a stroke are much older (mean age of 70.5 years) compared to those who never smoked and did not experience a stroke (45.3 years). Smokers, on the other hand, are generally younger compared to former smokers and non-smokers. However, smokers who had a stroke (62.3 years) are older than those who did not (46.1 years). Interestingly, individuals with an unknown smoking status show a stark difference based on stroke status: those with a stroke are much older (65.6 years) than those without (28.8 years).

```
# Calculate summary statistics
summary_stats <- stroke_data %>%
  group_by(smoking_status) %>%
  summarise(
    mean_age = mean(age, na.rm = TRUE),
    median_age = median(age, na.rm = TRUE),
    .groups = "drop"
  )

# Create the violin plot
ggplot(stroke_data, aes(x = smoking_status, y = age, fill = smoking_status)) +
  geom_violin(trim = FALSE, alpha = 0.7) +
  geom_boxplot(width = 0.1, fill = "white", alpha = 0.5) +
  geom_jitter(width = 0.1, alpha = 0.1, color = "darkgrey") +
  scale_fill_brewer(palette = "Set3") +
  labs(
    title = "Age Distribution by Smoking Status",
    subtitle = "Violin plot with embedded box plot, mean, median, and jittered points",
    x = "Smoking Status",
    y = "Age (years)",
    fill = "Smoking Status"
  ) +
  theme_minimal() +
  theme(
    plot.title = element_text(face = "bold", size = 16),
    plot.subtitle = element_text(face = "italic", size = 12),
    axis.title = element_text(face = "bold", size = 12),
```

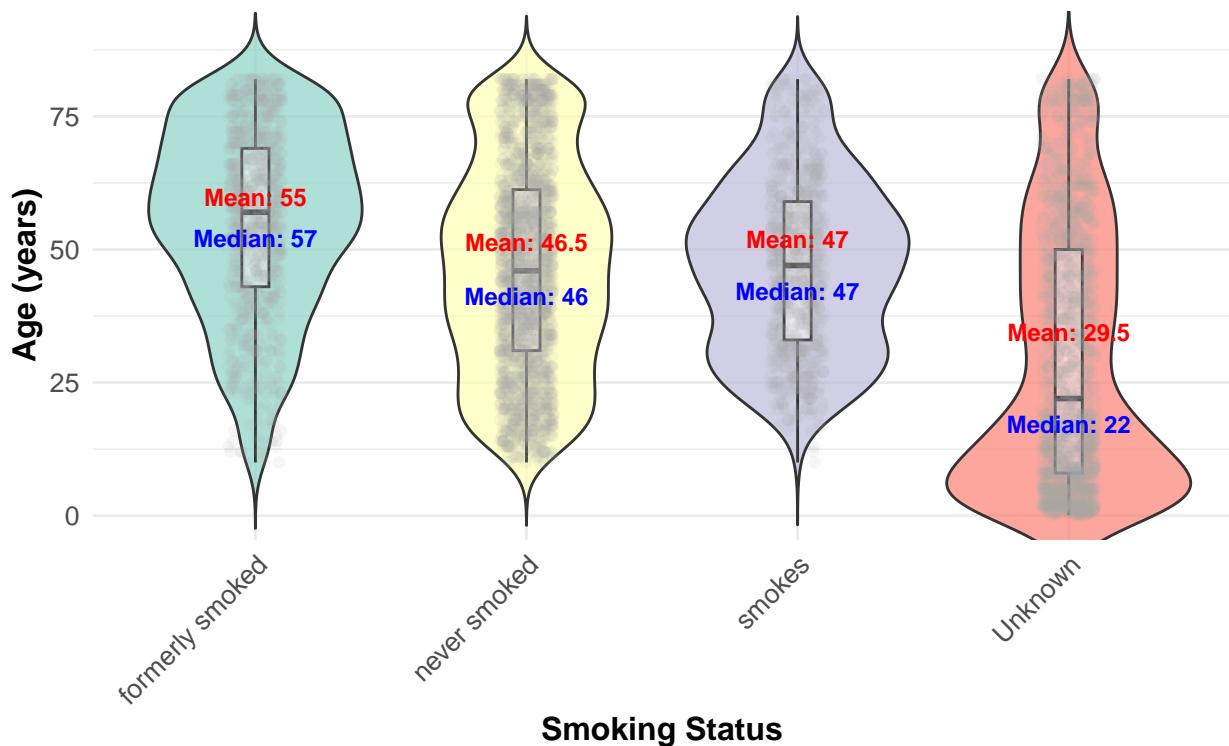
```

axis.text = element_text(size = 10),
axis.text.x = element_text(angle = 45, hjust = 1),
legend.position = "none"
) +
geom_text(data = summary_stats,
aes(x = smoking_status, y = mean_age,
label = paste("Mean:", round(mean_age, 1))),
color = "red", vjust = -1, size = 3, fontface = "bold") +
geom_text(data = summary_stats,
aes(x = smoking_status, y = median_age,
label = paste("Median:", round(median_age, 1))),
color = "blue", vjust = 2, size = 3, fontface = "bold") +
coord_cartesian(ylim = c(0, max(stroke_data$age, na.rm = TRUE) * 1.1))

```

Age Distribution by Smoking Status

Violin plot with embedded box plot, mean, median, and jittered points



Above indicates that former smokers tend to be the oldest group, with a mean age of 55 years and a median of 57 years, showing that people who have quit smoking are generally older. Those who never smoked or currently smoke have similar age profiles, with mean and median ages around 46 years, indicating that these groups are somewhat younger than former smokers. The “unknown” smoking status group stands out with a significantly lower mean age of 29.5 years and a median of 22 years, suggesting that this group consists of much younger individuals compared to the others.

```

# Define labels for stroke
stroke_labels <- c("0" = "No Stroke", "1" = "Stroke")

# Calculate summary statistics
summary_stats <- stroke_data %>%

```

```

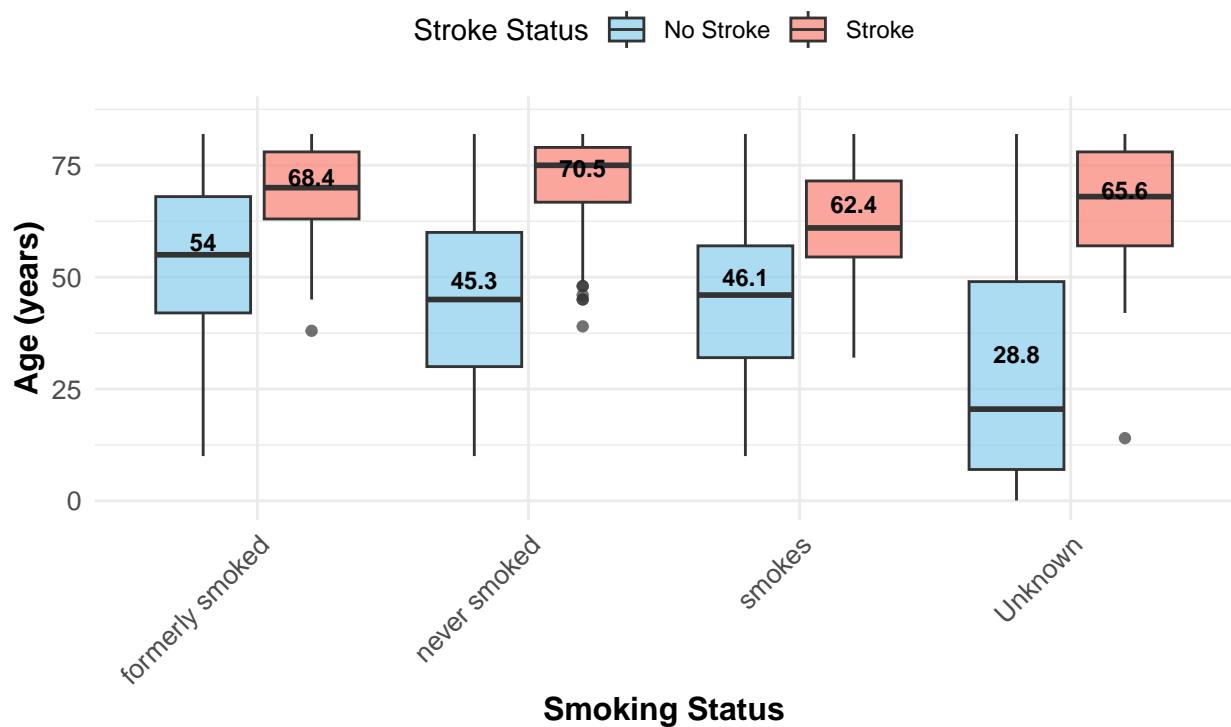
group_by(smoking_status, stroke) %>%
summarise(
  mean_age = mean(age, na.rm = TRUE),
  median_age = median(age, na.rm = TRUE),
  .groups = "drop"
)

# Create the comparative boxplot
ggplot(stroke_data, aes(x = smoking_status, y = age, fill = as.factor(stroke))) +
  geom_boxplot(position = position_dodge(width = 0.8), width = 0.7, alpha = 0.7) +
  scale_fill_manual(values = c("0" = "skyblue", "1" = "salmon"),
                     labels = stroke_labels) +
  labs(
    title = "Age Distribution by Smoking Status and Stroke Occurrence",
    subtitle = "Comparative boxplots with jittered points",
    x = "Smoking Status",
    y = "Age (years)",
    fill = "Stroke Status"
) +
  theme_minimal() +
  theme(
    plot.title = element_text(face = "bold", size = 16),
    plot.subtitle = element_text(face = "italic", size = 12),
    axis.title = element_text(face = "bold", size = 12),
    axis.text = element_text(size = 10),
    axis.text.x = element_text(angle = 45, hjust = 1),
    legend.position = "top"
) +
  geom_text(data = summary_stats,
            aes(x = smoking_status, y = mean_age, label = round(mean_age, 1), group = stroke),
            position = position_dodge(width = 0.8),
            vjust = -0.5, size = 3, fontface = "bold") +
  coord_cartesian(ylim = c(0, max(stroke_data$age, na.rm = TRUE) * 1.05))

```

Age Distribution by Smoking Status and Stroke Occurrence

Comparative boxplots with jittered points



This comparative boxplot tells us that former smokers who had a stroke tend to be older (mean age of 68.4 years) compared to those who didn't have a stroke (mean age of 54 years). A similar pattern is seen among those who never smoked: individuals who had a stroke are significantly older (70.5 years) than those who did not (45.3 years).

For current smokers, those with a stroke have a mean age of 62.4 years, while those without a stroke are younger (46.1 years). Interestingly, in the "unknown" smoking status group, those who had a stroke are much older (65.6 years) compared to those who did not (28.8 years).

```
#stroke_data <- load_stroke_data()
data_sub_age_hyp= stroke_data %>% group_by(hypertension,stroke) %>% summarise(Age_mean = mean(age), .group=TRUE)
print(data_sub_age_hyp)
```

Age levels with hypertension/Stroke

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

Among individuals without hypertension, those who have not had a stroke are younger, with a mean age of 40 years, while those who have experienced a stroke are significantly older, with a mean age of 67 years. For

individuals with hypertension, the pattern is similar: those without a stroke have a mean age of 61.2 years, and those with a stroke are older still, with a mean age of 69.6 years.

The data suggests that both stroke and hypertension are associated with older age, with those having both conditions (hypertension and stroke) being the oldest group. Conversely, individuals without either condition tend to be much younger.

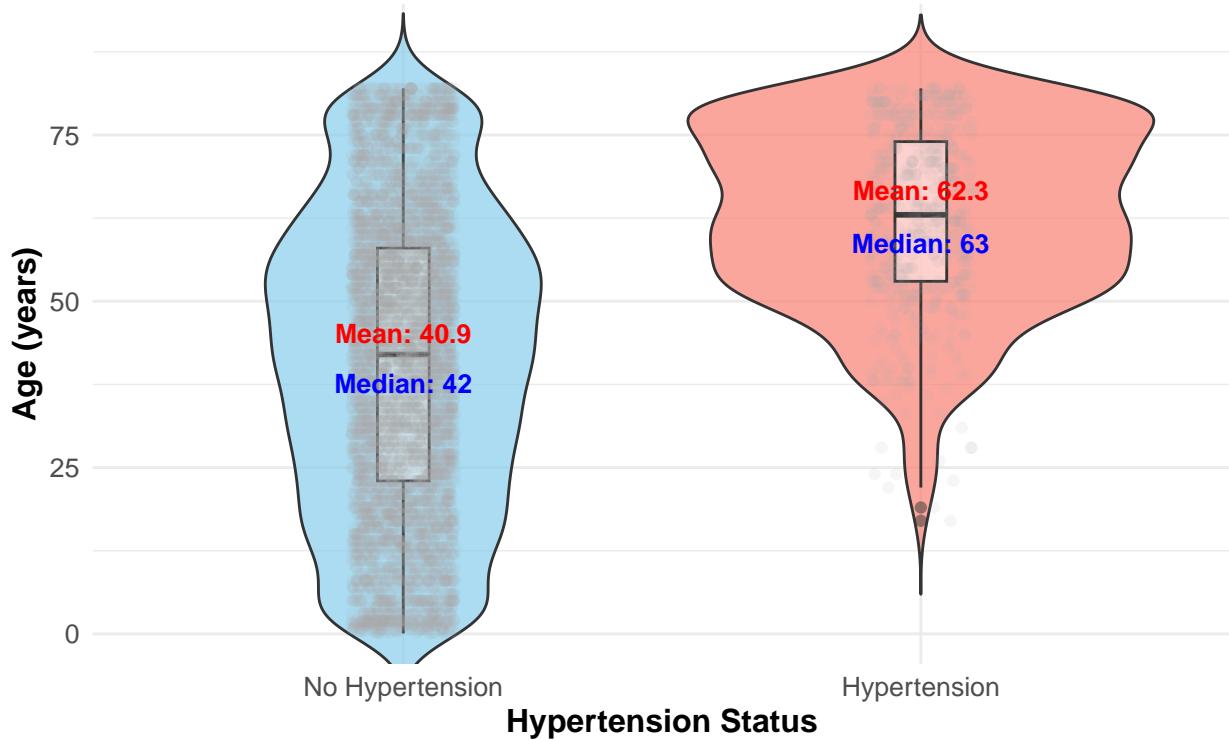
```
# Define labels for hypertension
hypertension_labels <- c("0" = "No Hypertension", "1" = "Hypertension")

# Calculate summary statistics
summary_stats <- stroke_data %>%
  group_by(hypertension) %>%
  summarise(
    mean_age = mean(age, na.rm = TRUE),
    median_age = median(age, na.rm = TRUE),
    .groups = "drop"
  )

# Create the violin plot
ggplot(stroke_data, aes(x = hypertension, y = age, fill = hypertension)) +
  geom_violin(trim = FALSE, alpha = 0.7) +
  geom_boxplot(width = 0.1, fill = "white", alpha = 0.5) +
  geom_jitter(width = 0.1, alpha = 0.1, color = "darkgrey") +
  scale_fill_manual(values = c("0" = "skyblue", "1" = "salmon"),
                     labels = hypertension_labels) +
  scale_x_discrete(labels = hypertension_labels) +
  labs(
    title = "Age Distribution by Hypertension Status",
    subtitle = "Violin plot with embedded box plot, mean, median, and jittered points",
    x = "Hypertension Status",
    y = "Age (years)",
    fill = "Hypertension Status"
  ) +
  theme_minimal() +
  theme(
    plot.title = element_text(face = "bold", size = 16),
    plot.subtitle = element_text(face = "italic", size = 12),
    axis.title = element_text(face = "bold", size = 12),
    axis.text = element_text(size = 10),
    legend.position = "none"
  ) +
  geom_text(data = summary_stats,
            aes(x = hypertension, y = mean_age,
                label = paste("Mean:", round(mean_age, 1))),
            color = "red", vjust = -1, size = 3.5, fontface = "bold") +
  geom_text(data = summary_stats,
            aes(x = hypertension, y = median_age,
                label = paste("Median:", round(median_age, 1))),
            color = "blue", vjust = 2, size = 3.5, fontface = "bold") +
  coord_cartesian(ylim = c(0, max(stroke_data$age, na.rm = TRUE) * 1.1))
```

Age Distribution by Hypertension Status

Violin plot with embedded box plot, mean, median, and jittered points



The plots indicate that individuals without hypertension tend to be younger, with a mean age of 40.9 years and a median of 42 years. The distribution for this group shows a wide spread with a relatively high concentration of younger individuals.

In contrast, individuals with hypertension are significantly older, with a mean age of 62.3 years and a median of 63 years. The distribution for this group is more tightly concentrated around the older age range. The stark difference between the two groups suggests that hypertension is more prevalent among older individuals, while younger individuals tend to have no hypertension.

```
# Define labels for hypertension and stroke
hypertension_labels <- c("0" = "No Hypertension", "1" = "Hypertension")
stroke_labels <- c("0" = "No Stroke", "1" = "Stroke")

# Calculate summary statistics
summary_stats <- stroke_data %>%
  group_by(hypertension, stroke) %>%
  summarise(
    mean_age = mean(age, na.rm = TRUE),
    median_age = median(age, na.rm = TRUE),
    .groups = "drop"
  )

# Create the comparative boxplot
ggplot(stroke_data, aes(x = hypertension, y = age, fill = stroke)) +
  geom_boxplot(position = position_dodge(width = 0.8), width = 0.7, alpha = 0.7) +
  geom_jitter(position = position_jitterdodge(dodge.width = 0.8), alpha = 0.1, color = "darkgrey") +
  scale_fill_manual(values = c("0" = "skyblue", "1" = "salmon"))
```

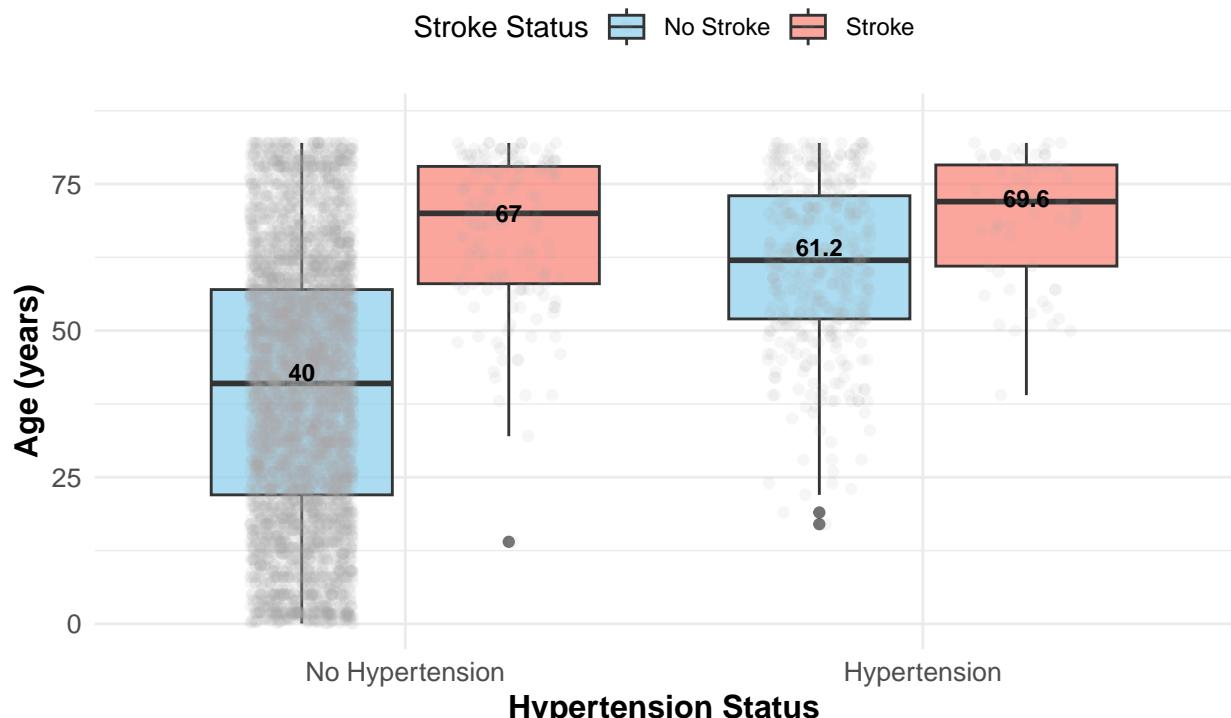
```

        labels = stroke_labels) +
scale_x_discrete(labels = hypertension_labels) +
labs(
  title = "Age Distribution by Hypertension Status and Stroke Occurrence",
  subtitle = "Comparative boxplots with jittered points",
  x = "Hypertension Status",
  y = "Age (years)",
  fill = "Stroke Status"
) +
theme_minimal() +
theme(
  plot.title = element_text(face = "bold", size = 16),
  plot.subtitle = element_text(face = "italic", size = 12),
  axis.title = element_text(face = "bold", size = 12),
  axis.text = element_text(size = 10),
  legend.position = "top"
) +
geom_text(data = summary_stats,
  aes(x = hypertension, y = mean_age, label = round(mean_age, 1), group = stroke),
  position = position_dodge(width = 0.8),
  vjust = -0.5, size = 3, fontface = "bold") +
coord_cartesian(ylim = c(0, max(stroke_data$age, na.rm = TRUE) * 1.05))

```

Age Distribution by Hypertension Status and Stroke Occurrence

Comparative boxplots with jittered points



Among individuals without hypertension, those who have experienced a stroke are significantly older, with a mean age of 67 years, compared to those without a stroke who have a mean age of 40 years. This shows that stroke is much more common in older individuals within the no-hypertension group.

For those with hypertension, the trend persists: individuals with a stroke are older, with a mean age of 69.6 years, compared to those without a stroke, who have a mean age of 61.2 years.

```
data_sub_age_hyp = stroke_data %>% group_by(ever_married, stroke) %>% summarise(Age_mean = mean(age), .groups = "drop")
pander(data_sub_age_hyp)
```

Age levels with Marital Status/Stroke

ever_married	stroke	Age_mean
No	0	21.2
No	1	66.52
Yes	0	53.22
Yes	1	67.86

For individuals who have never been married, those without a stroke are much younger, with a mean age of 21.2 years, while those who have had a stroke are considerably older, with a mean age of 66.5 years. This suggests that stroke is rare among younger, unmarried individuals but more prevalent in older unmarried individuals.

Among those who have been married, the mean age without a stroke is 53.2 years, and it increases to 67.9 years for those who have experienced a stroke. This indicates that married individuals tend to be older on average, and those who suffer from strokes are generally older as well.

```
# Calculate summary statistics
summary_stats <- stroke_data %>%
  group_by(ever_married) %>%
  summarise(
    mean_age = mean(age, na.rm = TRUE),
    median_age = median(age, na.rm = TRUE),
    .groups = "drop"
  )

# Create the violin plot
ggplot(stroke_data, aes(x = ever_married, y = age, fill = ever_married)) +
  geom_violin(trim = FALSE, alpha = 0.7) +
  geom_boxplot(width = 0.1, fill = "white", alpha = 0.5) +
  geom_jitter(width = 0.1, alpha = 0.1, color = "darkgrey") +
  scale_fill_manual(values = c("No" = "skyblue", "Yes" = "salmon"),
                     labels = c("No" = "Never Married", "Yes" = "Ever Married")) +
  scale_x_discrete(labels = c("No" = "Never Married", "Yes" = "Ever Married")) +
  labs(
    title = "Age Distribution by Marital Status",
    subtitle = "Violin plot with embedded box plot, mean, median, and jittered points",
    x = "Marital Status",
    y = "Age (years)",
    fill = "Marital Status"
  ) +
  theme_minimal() +
  theme(
    plot.title = element_text(face = "bold", size = 16),
```

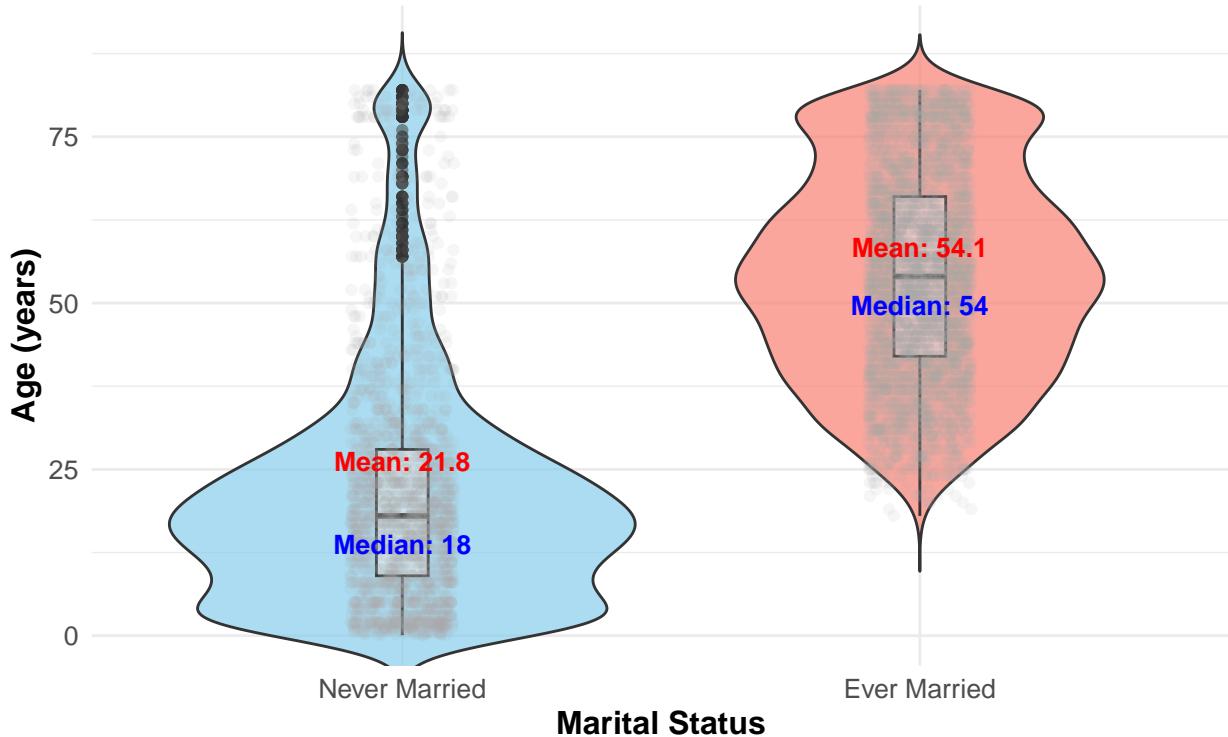
```

plot.subtitle = element_text(face = "italic", size = 12),
axis.title = element_text(face = "bold", size = 12),
axis.text = element_text(size = 10),
legend.position = "none"
) +
geom_text(data = summary_stats,
aes(x = ever_married, y = mean_age,
label = paste("Mean:", round(mean_age, 1))),
color = "red", vjust = -1, size = 3.5, fontface = "bold") +
geom_text(data = summary_stats,
aes(x = ever_married, y = median_age,
label = paste("Median:", round(median_age, 1))),
color = "blue", vjust = 2, size = 3.5, fontface = "bold") +
coord_cartesian(ylim = c(0, max(stroke_data$age, na.rm = TRUE) * 1.1))

```

Age Distribution by Marital Status

Violin plot with embedded box plot, mean, median, and jittered points



The “never married” group is notably younger, with a mean age of 21.8 years and a median age of 18 years, indicating a concentration of young individuals in this group. The distribution is narrow and primarily concentrated in the lower age range, suggesting that most of the individuals who have never married are quite young.

On the other hand, the “ever married” group is significantly older, with a mean age of 54.1 years and a median age of 54 years. The distribution for this group is much wider and more evenly spread across the higher age ranges, indicating a broader age spectrum among those who have been married.

```

# Calculate summary statistics
summary_stats <- stroke_data %>%
  group_by(ever_married, stroke) %>%

```

```

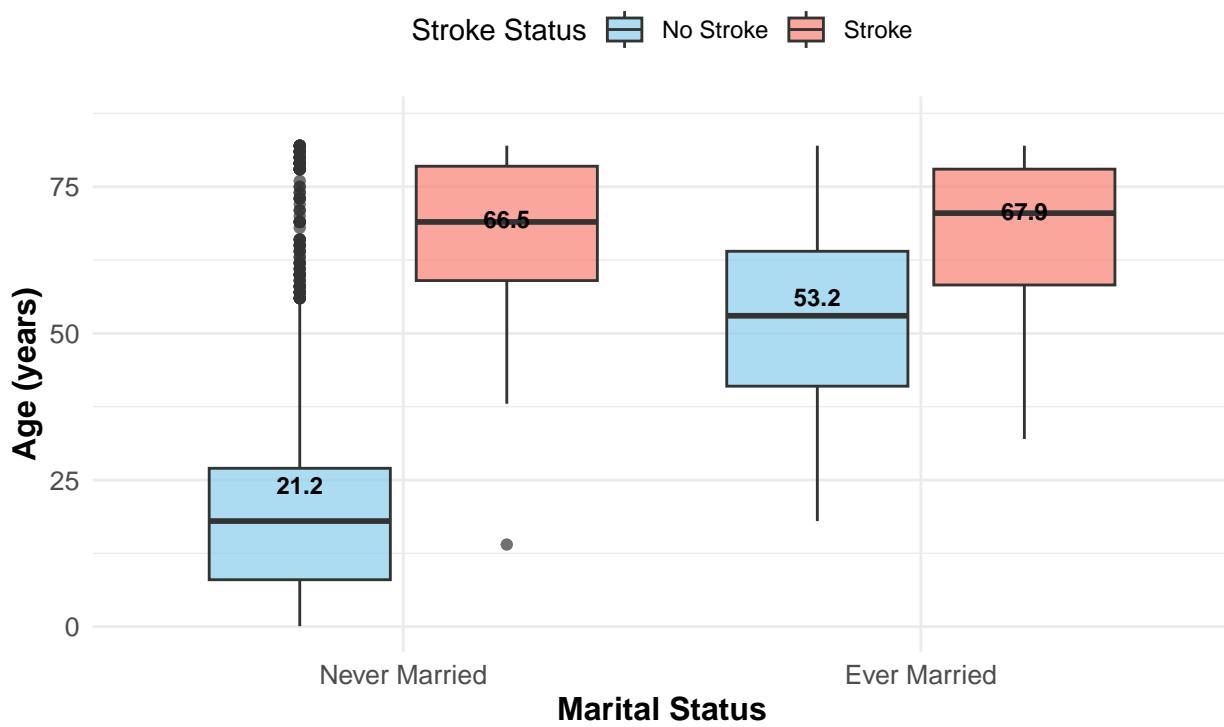
summarise(
  mean_age = mean(age, na.rm = TRUE),
  median_age = median(age, na.rm = TRUE),
  .groups = "drop"
)

# Create the comparative boxplot
ggplot(stroke_data, aes(x = ever_married, y = age, fill = stroke)) +
  geom_boxplot(position = position_dodge(width = 0.8), width = 0.7, alpha = 0.7) +
  # Correct fill scale mapping with actual factor levels
  scale_fill_manual(values = c("0" = "skyblue", "1" = "salmon"),
                     labels = c("No Stroke", "Stroke")) +
  # Correct x-axis labels
  scale_x_discrete(labels = c("No" = "Never Married", "Yes" = "Ever Married")) +
  # Add plot labels
  labs(
    title = "Age Distribution by Marital Status and Stroke Occurrence",
    subtitle = "Comparative boxplots with jittered points",
    x = "Marital Status",
    y = "Age (years)",
    fill = "Stroke Status"
  ) +
  # Customize the theme
  theme_minimal() +
  theme(
    plot.title = element_text(face = "bold", size = 16),
    plot.subtitle = element_text(face = "italic", size = 12),
    axis.title = element_text(face = "bold", size = 12),
    axis.text = element_text(size = 10),
    legend.position = "top"
  ) +
  # Add text for mean ages
  geom_text(data = summary_stats,
            aes(x = ever_married, y = mean_age, label = round(mean_age, 1), group = stroke),
            position = position_dodge(width = 0.8),
            vjust = -0.5, size = 3, fontface = "bold") +
  # Adjust y-axis limits
  coord_cartesian(ylim = c(0, max(stroke_data$age, na.rm = TRUE) * 1.05))

```

Age Distribution by Marital Status and Stroke Occurrence

Comparative boxplots with jittered points



For individuals who have never been married, those without a stroke tend to be much younger, with a mean age of 21.2 years. However, those who have had a stroke are significantly older, with a mean age of 66.5 years, indicating that stroke is rare among younger, unmarried individuals but more common in older unmarried people.

Among those who have been married, the pattern is similar but at older ages. Individuals without a stroke have a mean age of 53.2 years, while those who have experienced a stroke are even older, with a mean age of 67.9 years. This suggests that stroke is more common in older married individuals compared to their counterparts without a stroke.

```
data_sub_age_hyp = stroke_data %>% group_by(gender,stroke) %>% summarise(Age_mean = mean(age), .groups = pander(data_sub_age_hyp)
```

Age levels with Gender/Stroke

gender	stroke	Age_mean
Female	0	42.41
Female	1	67.24
Male	0	40.83
Male	1	68.35

For females, the mean age of those who have not experienced a stroke is 42.4 years, while females who have had a stroke are significantly older, with a mean age of 67.2 years.

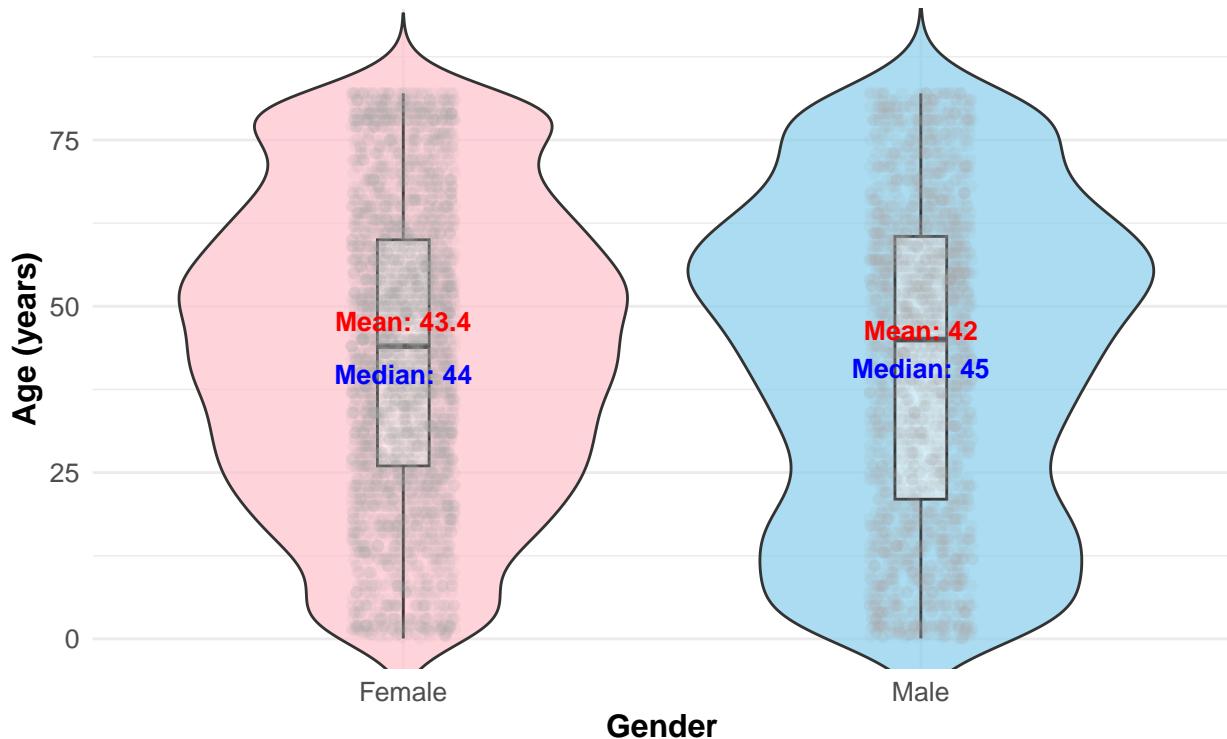
For males, a similar pattern is observed. Males without a stroke have a mean age of 40.8 years, whereas males who have experienced a stroke are older, with a mean age of 68.3 years.

```
# Calculate summary statistics
summary_stats <- stroke_data %>%
  group_by(gender) %>%
  summarise(
    mean_age = mean(age, na.rm = TRUE),
    median_age = median(age, na.rm = TRUE),
    .groups = "drop"
  )

# Create the violin plot
ggplot(stroke_data, aes(x = gender, y = age, fill = gender)) +
  geom_violin(trim = FALSE, alpha = 0.7) +
  geom_boxplot(width = 0.1, fill = "white", alpha = 0.5) +
  geom_jitter(width = 0.1, alpha = 0.1, color = "darkgrey") +
  scale_fill_manual(values = c("Male" = "skyblue", "Female" = "pink", "Other" = "purple")) +
  labs(
    title = "Age Distribution by Gender",
    subtitle = "Violin plot with embedded box plot, mean, median, and jittered points",
    x = "Gender",
    y = "Age (years)",
    fill = "Gender"
  ) +
  theme_minimal() +
  theme(
    plot.title = element_text(face = "bold", size = 16),
    plot.subtitle = element_text(face = "italic", size = 12),
    axis.title = element_text(face = "bold", size = 12),
    axis.text = element_text(size = 10),
    legend.position = "none"
  ) +
  geom_text(data = summary_stats,
            aes(x = gender, y = mean_age,
                 label = paste("Mean:", round(mean_age, 1))),
            color = "red", vjust = -1, size = 3.5, fontface = "bold") +
  geom_text(data = summary_stats,
            aes(x = gender, y = median_age,
                 label = paste("Median:", round(median_age, 1))),
            color = "blue", vjust = 2, size = 3.5, fontface = "bold") +
  coord_cartesian(ylim = c(0, max(stroke_data$age, na.rm = TRUE) * 1.1))
```

Age Distribution by Gender

Violin plot with embedded box plot, mean, median, and jittered points



For females, the mean age is 43.4 years, and the median is 42 years. The distribution for females shows a relatively even spread across ages, though it is slightly more concentrated around the central values.

For males, the mean age is 42 years, and the median is 45 years, showing a similar spread to females but with a slightly different central tendency. The distribution for males also shows a broad spread, with a similar concentration around middle ages.

```
# Ensure gender and stroke are factors with descriptive labels
levels(stroke_data$stroke) <- c("No Stroke", "Stroke")

# Calculate summary statistics
summary_stats <- stroke_data %>%
  group_by(gender, stroke) %>%
  summarise(
    mean_age = mean(age, na.rm = TRUE),
    median_age = median(age, na.rm = TRUE),
    .groups = "drop"
  )

# Create the comparative boxplot
ggplot(stroke_data, aes(x = gender, y = age, fill = stroke)) +
  geom_boxplot(position = position_dodge(width = 0.8), width = 0.7, alpha = 0.7) +
  geom_jitter(position = position_jitterdodge(dodge.width = 0.8), alpha = 0.1, color = "darkgrey") +
  scale_fill_manual(values = c("No Stroke" = "skyblue", "Stroke" = "salmon")) +
  labs(
    title = "Age Distribution by Gender and Stroke Occurrence",
    subtitle = "Comparative boxplots with jittered points",
```

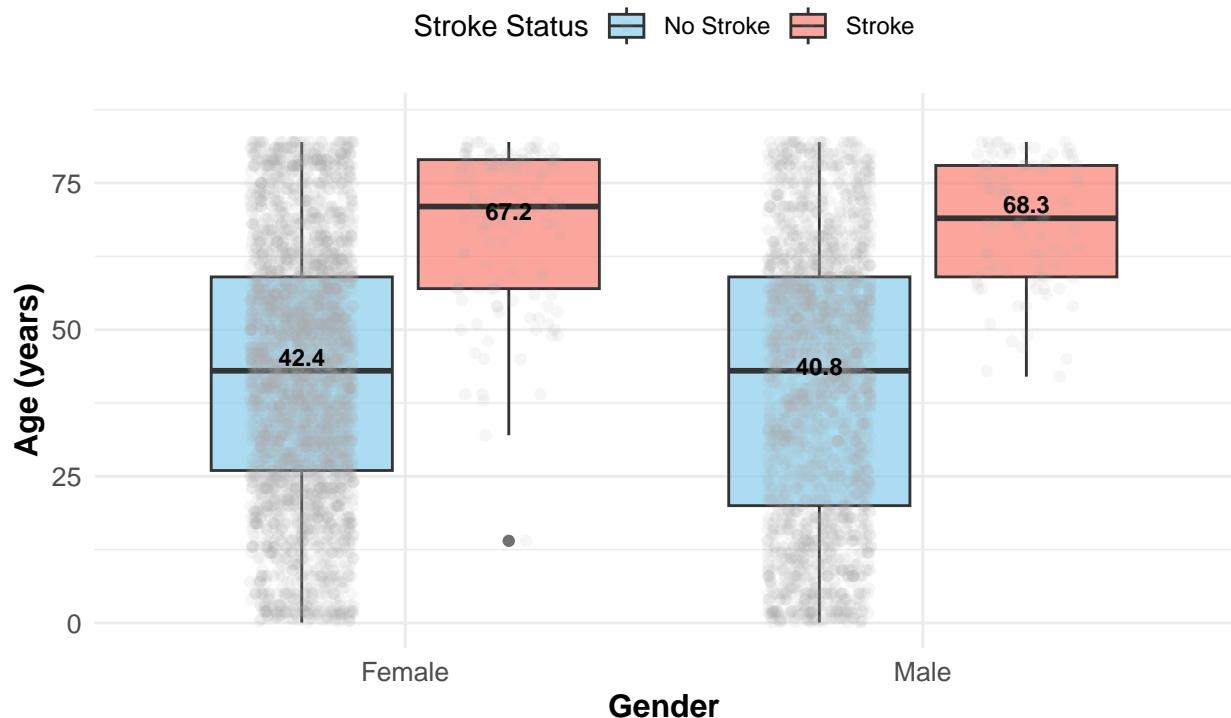
```

x = "Gender",
y = "Age (years)",
fill = "Stroke Status"
) +
theme_minimal() +
theme(
  plot.title = element_text(face = "bold", size = 16),
  plot.subtitle = element_text(face = "italic", size = 12),
  axis.title = element_text(face = "bold", size = 12),
  axis.text = element_text(size = 10),
  legend.position = "top"
) +
geom_text(data = summary_stats,
  aes(x = gender, y = mean_age, label = round(mean_age, 1), group = stroke),
  position = position_dodge(width = 0.8),
  vjust = -0.5, size = 3, fontface = "bold") +
coord_cartesian(ylim = c(0, max(stroke_data$age, na.rm = TRUE) * 1.05))

```

Age Distribution by Gender and Stroke Occurrence

Comparative boxplots with jittered points



Among females, the mean age of those without a stroke is 42.4 years, while females who have experienced a stroke are significantly older, with a mean age of 67.2 years. This demonstrates a clear distinction between younger females who haven't had a stroke and older females who have.

For males, the pattern is similar: males without a stroke have a mean age of 40.8 years, whereas males who have had a stroke are much older, with a mean age of 68.3 years. The trend indicates that stroke occurrence is linked to older age in both genders.