Stroke

Bethelhem Kassa

4/6/2023

```
library(tidyverse)

library(ggplot2)
stroke_data<-read.csv("C:/Users/nebar/Downloads/stroke data.csv")
#To see the column and their data types
str(stroke_data)

## 'data.frame':    43400 obs. of  15 variables:
##  $ id             : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ gender         : chr  "Male" "Male" "Female" "Female" ...
##  $ age            : num  3 58 8 70 14 47 52 75 32 74 ...
##  $ married        : chr  "No" "Yes" "No" "Yes" ...
##  $ hypertension   : int  0 1 0 0 0 0 0 0 0 1 ...
##  $ heart_disease  : int  0 0 0 0 0 0 0 1 0 0 ...
##  $ occupation     : chr  "A" "B" "B" "B" ...
##  $ residence      : chr  "Rural" "Urban" "Urban" "Rural" ...
##  $ metric_1       : num  95.1 88 110.9 69 161.3 ...
##  $ metric_2       : num  18 39.2 17.6 35.9 19.1 50.1 17.7 27 32.3 54.6 ...
##  $ metric_3       : int  1 1 0 0 1 0 1 1 0 0 ...
##  $ metric_4       : num  99.3 99.7 96.3 95.5 95.1 ...
##  $ metric_5       : num  95.1 88 110.9 69 161.3 ...
##  $ smoking_status: chr  "" "never smoked" "" "formerly smoked" ...
##  $ stroke         : int  0 0 0 0 0 0 0 0 0 0 ...

#To list the summary statistics of the variables
summary(stroke_data)

##        id           gender               age            married
##  Min.   :    1   Length:43400       Min.   : -10.00   Length:43400
##  1st Qu.:10851   Class :character   1st Qu.:  24.00   Class :character
##  Median :21701   Mode  :character   Median :  44.00   Mode  :character
##  Mean   :21701                      Mean   :  42.26
##  3rd Qu.:32550                      3rd Qu.:  60.00
##  Max.   :43400                      Max.   :1000.00
##
##   hypertension     heart_disease      occupation         residence
##  Min.   :0.00000   Min.   :0.00000   Length:43400       Length:43400
##  1st Qu.:0.00000   1st Qu.:0.00000   Class :character   Class :character
##  Median :0.00000   Median :0.00000   Mode  :character   Mode  :character
##  Mean   :0.09357   Mean   :0.04751
##  3rd Qu.:0.00000   3rd Qu.:0.00000
##  Max.   :1.00000   Max.   :1.00000
```

```
##
##     metric_1         metric_2        metric_3         metric_4
##  Min.   : 55.00   Min.   :10.10   Min.   :0.0000   Min.   : 87.42
##  1st Qu.: 77.54   1st Qu.:23.20   1st Qu.:0.0000   1st Qu.: 96.59
##  Median : 91.58   Median :27.70   Median :0.0000   Median : 97.61
##  Mean   :104.48   Mean   :28.61   Mean   :0.2899   Mean   : 97.53
##  3rd Qu.:112.07   3rd Qu.:32.90   3rd Qu.:1.0000   3rd Qu.: 98.70
##  Max.   :291.05   Max.   :97.60   Max.   :1.0000   Max.   :100.00
##                   NA's   :1462
##     metric_5      smoking_status         stroke
##  Min.   : 55.00   Length:43400       Min.   :0.00000
##  1st Qu.: 77.54   Class :character   1st Qu.:0.00000
##  Median : 91.58   Mode  :character   Median :0.00000
##  Mean   :104.48                      Mean   :0.01804
##  3rd Qu.:112.07                      3rd Qu.:0.00000
##  Max.   :291.05                      Max.   :1.00000
##

#since id has no statistical value, we need to remove it
stroke_data<-stroke_data[,-1]

#To find which columns have missing values
sapply(stroke_data,function(x)sum(is.na(x)))

##         gender            age        married   hypertension  heart_disease
##              0              0              0              0              0
##     occupation      residence       metric_1       metric_2       metric_3
##              0              0              0           1462              0
##       metric_4       metric_5 smoking_status         stroke
##              0              0              0              0

#Replacing the missing values of metric_2 with mean value since it has
numerical values
stroke_data1<-stroke_data
stroke_data1$metric_2[is.na(stroke_data1$metric_2)]<-
mean(stroke_data1$metric_2,na.rm=TRUE)

# List the distinct / unique values
calc_mode <- function(smoking_status)
  {
distinct_values <- unique(stroke_data1$smoking_status)

# Count the occurrence of each distinct value
distinct_tabulate <- tabulate(match(stroke_data1$smoking_status,
distinct_values))

 # Return the value with the highest occurrence
distinct_values[which.max(distinct_tabulate)]
  }
calc_mode(stoke_data1$smoking_status)
```

```
## [1] "never smoked"
```

```
#Replacing the smoking status missing value with the mode
strokedf<-stroke_data1%>%
  group_by(smoking_status)%>%
mutate(smoking_status=if_else(is.na(smoking_status),calc_mode(smoking_status)
,smoking_status))
```

```
#Check again if there are any missing values left
sapply(strokedf,function(x)sum(is.na(x)))
```

```
##         gender            age        married   hypertension  heart_disease
##              0              0              0              0              0
##     occupation      residence       metric_1       metric_2       metric_3
##              0              0              0              0              0
##       metric_4       metric_5 smoking_status         stroke
##              0              0              0              0
```

```
#Change the categorical variables in to factor
strokedf$hypertension<-as.factor(strokedf$hypertension)
strokedf$heart_disease<-as.factor(strokedf$heart_disease)
strokedf$metric_3<-as.factor(strokedf$metric_3)
strokedf$stroke<-as.factor(strokedf$stroke)
strokedf$gender<-as.factor(strokedf$gender)
strokedf$occupation<-as.factor(strokedf$occupation)
strokedf$smoking_status<-as.factor(strokedf$smoking_status)
strokedf$married<-as.factor(strokedf$married)
strokedf$residence<-as.factor(strokedf$residence)
```

```
#since there is invalid age which is negative, we have to remove it
age1<-strokedf[which(strokedf$age<0),]
dim(age1)
```

```
## [1]  1 14
```

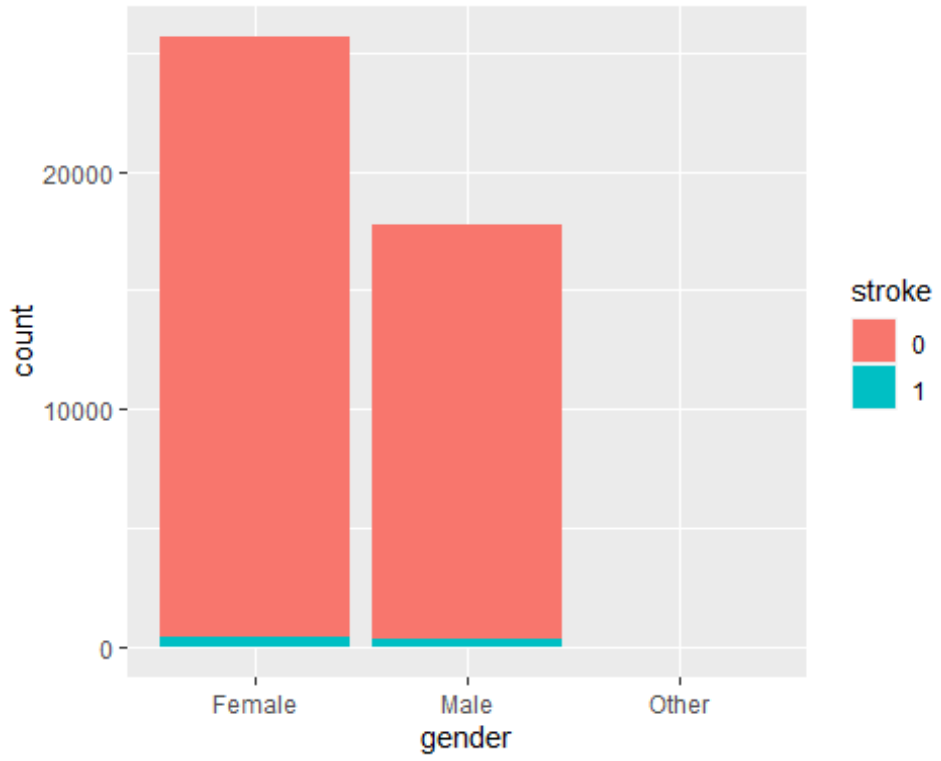
```
strokedf<-strokedf[-which(strokedf$age<0),]
dim(strokedf)
```

```
## [1] 43399    14
```

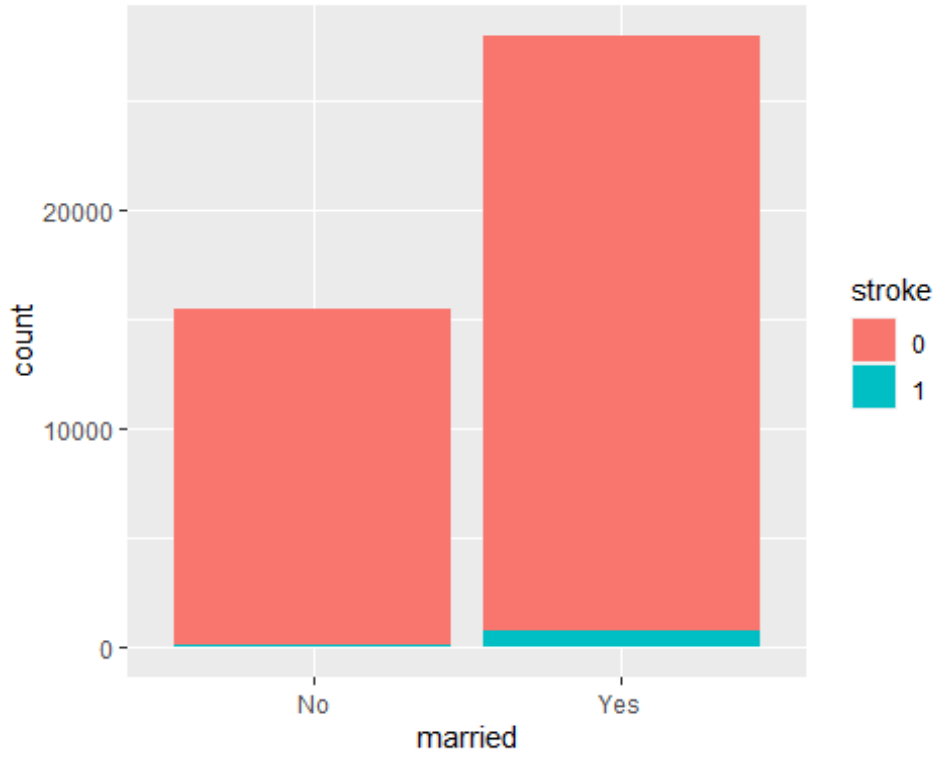
```
summary(strokedf$age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.08   24.00   44.00   42.26   60.00 1000.00
```

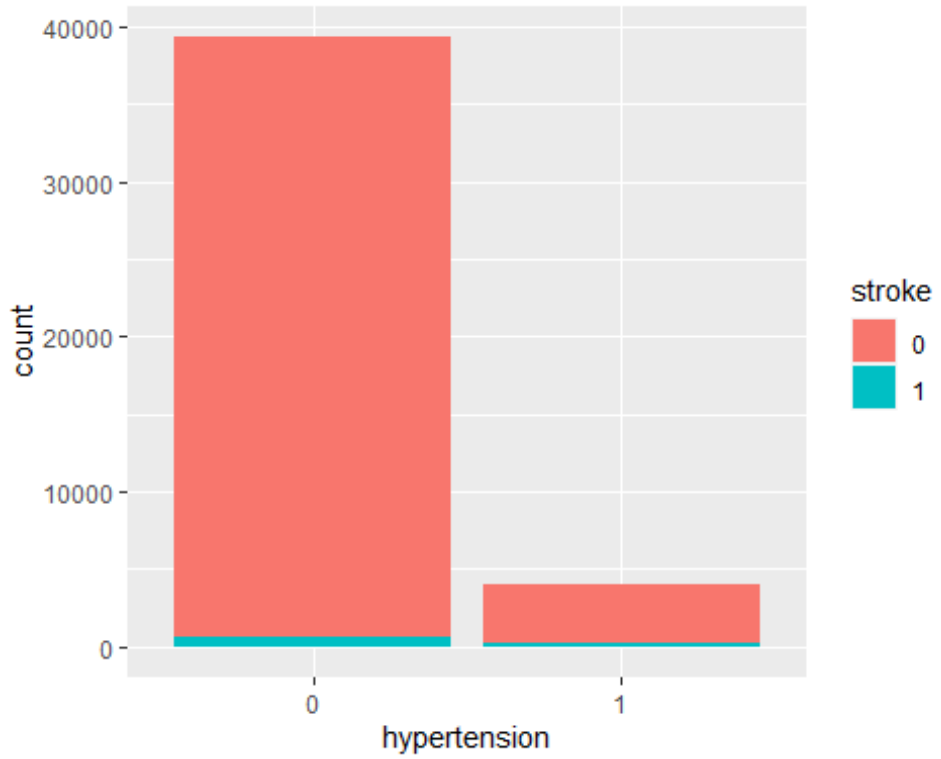
```
#PLot the distribution and relationship of each predictor variable with the
response variable(stroke)
ggplot(data= strokedf,aes(x=gender,fill=stroke))+geom_bar()
```



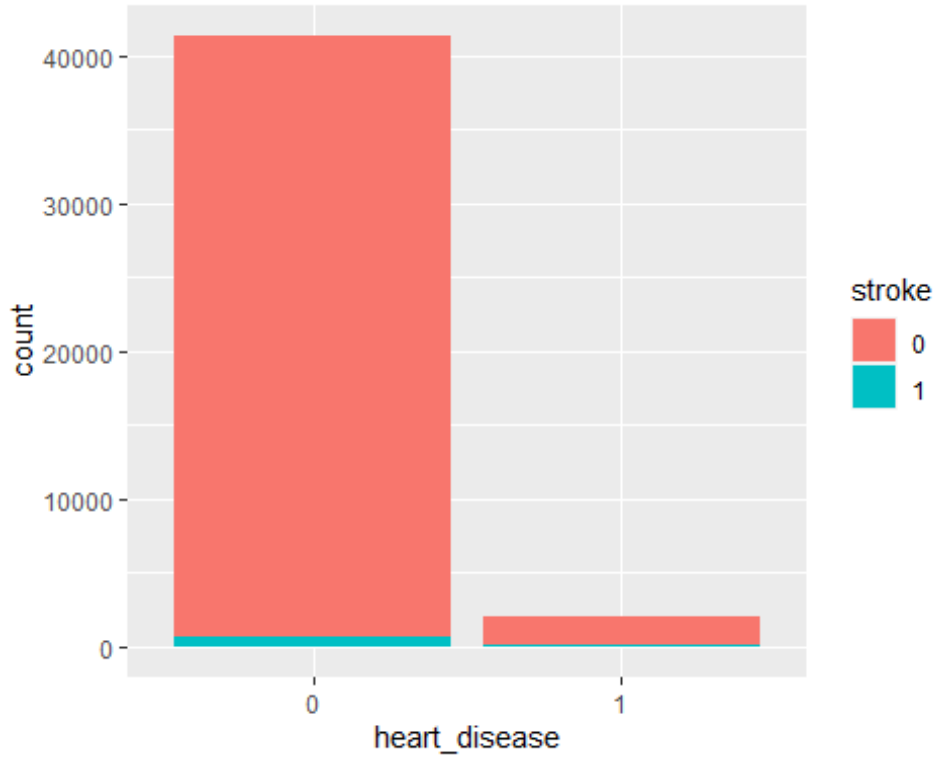

```
ggplot(data= strokedf,aes(x=married,fill=stroke))+geom_bar()
```



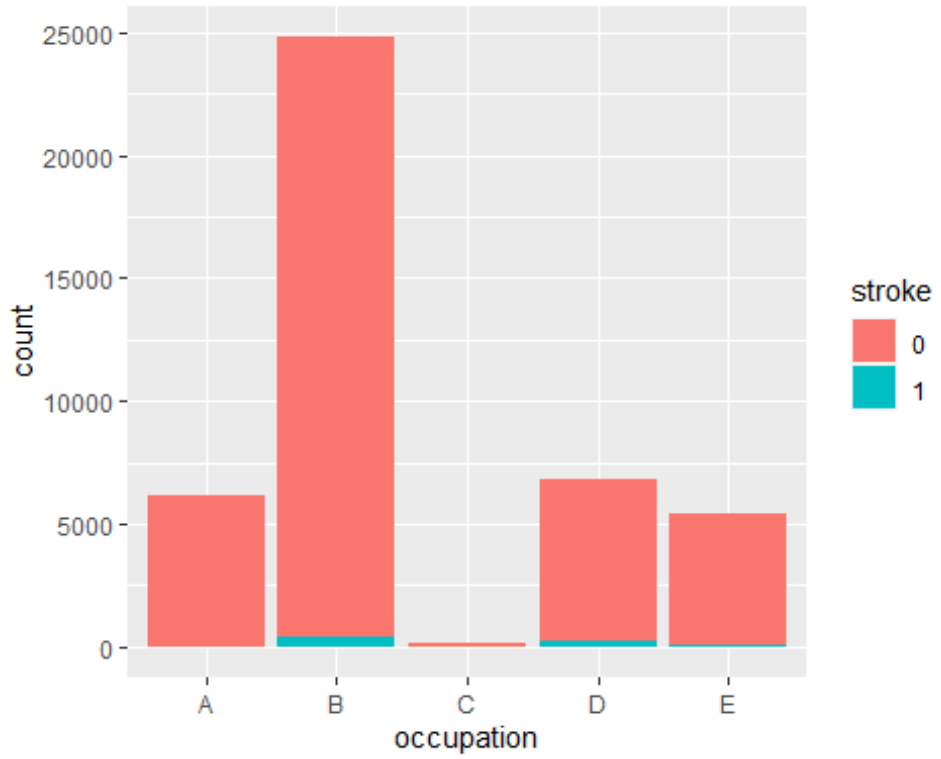

```
ggplot(data= strokedf,aes(x=hypertension,fill=stroke))+geom_bar()
```



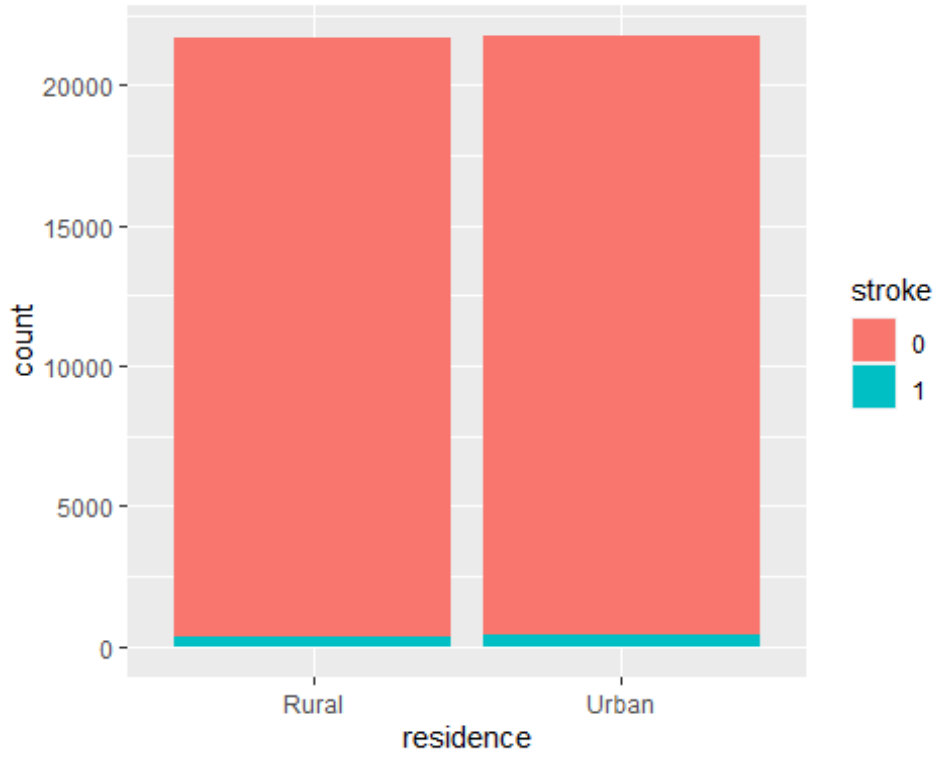

```
ggplot(data= strokedf,aes(x=heart_disease,fill=stroke))+geom_bar()
```



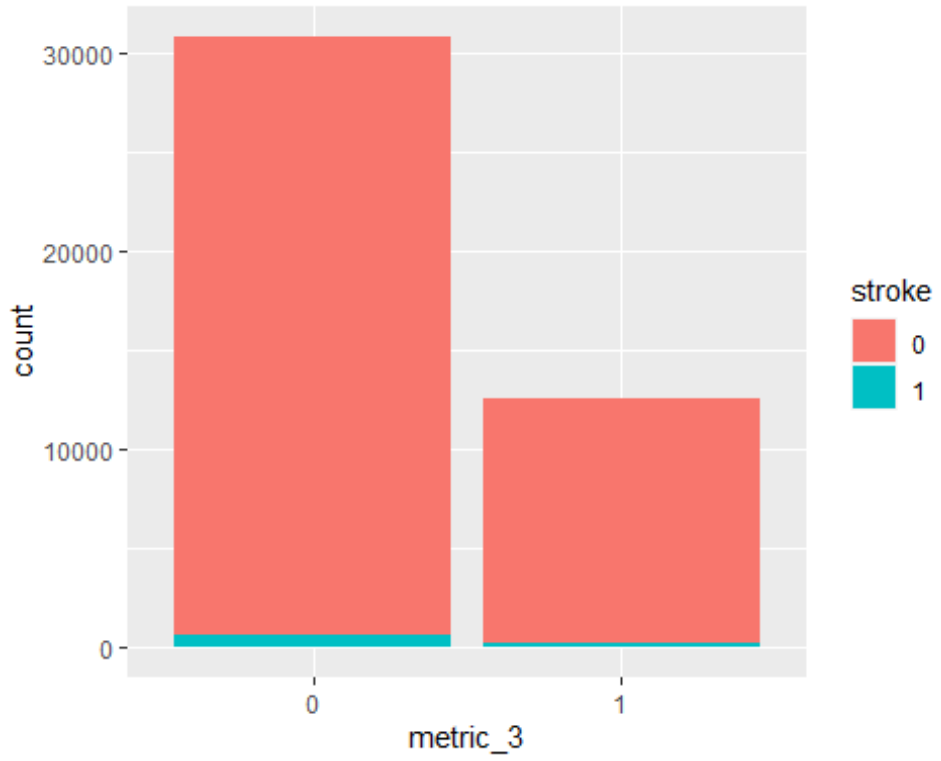

```
ggplot(data= strokedf,aes(x=occupation,fill=stroke))+geom_bar()
```



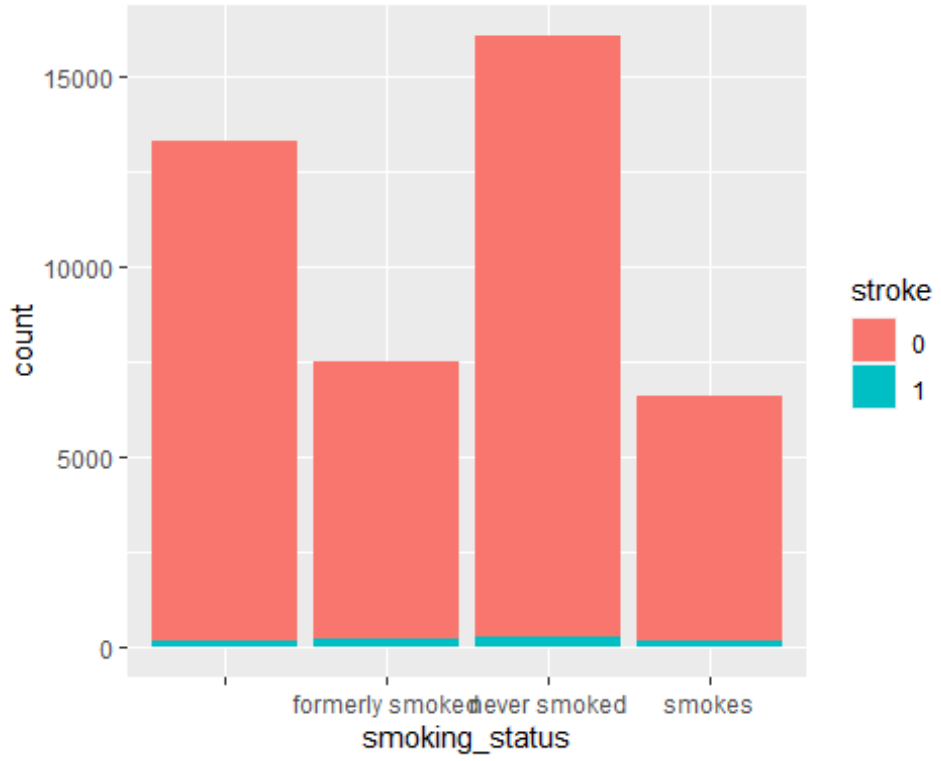

```
ggplot(data= strokedf,aes(x=residence,fill=stroke))+geom_bar()
```




```
ggplot(data= strokedf,aes(x=metric_3,fill=stroke))+geom_bar()
```

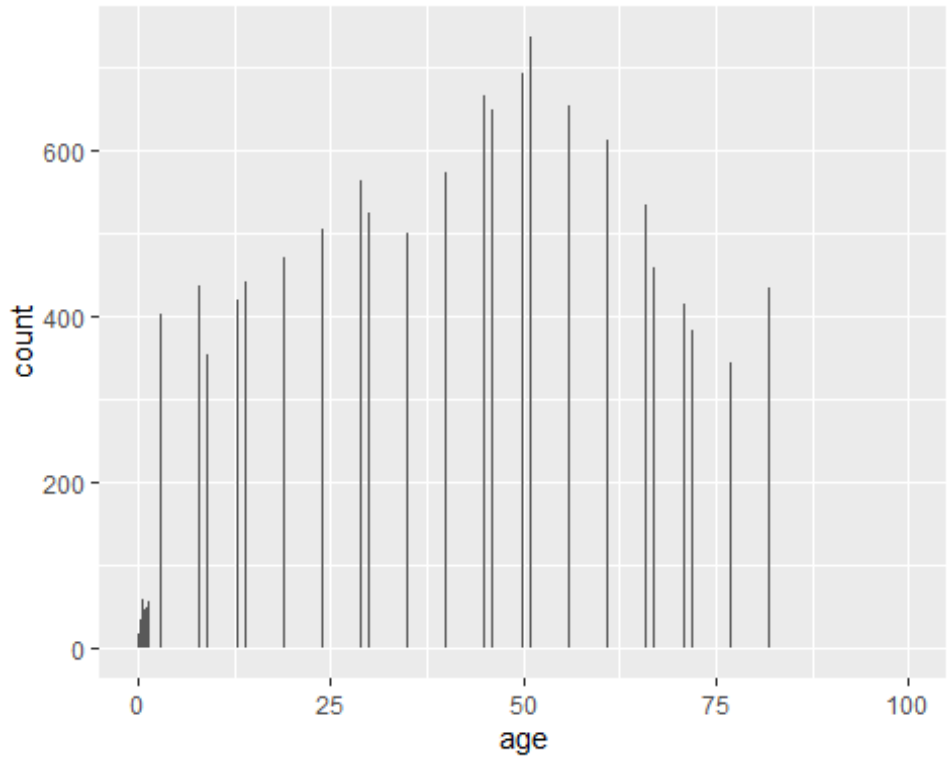



```
ggplot(data= strokedf,aes(x=smoking_status,fill=stroke))+geom_bar()
```

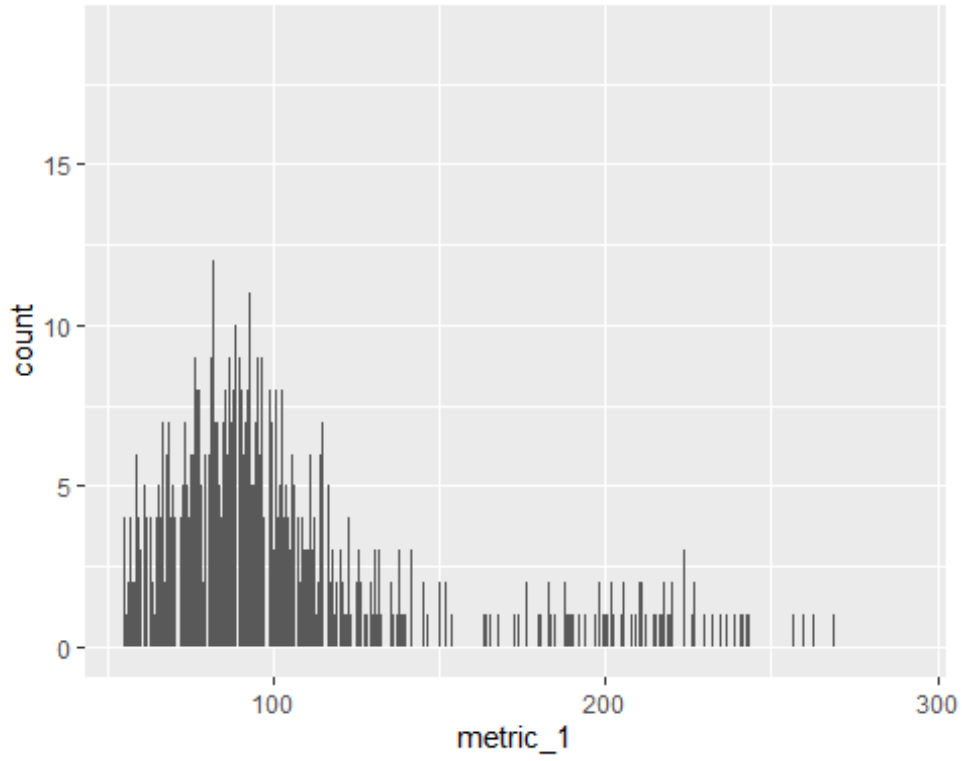



```
ggplot(data= strokedf,aes(x=age))+geom_bar()+xlim(0,100)
```

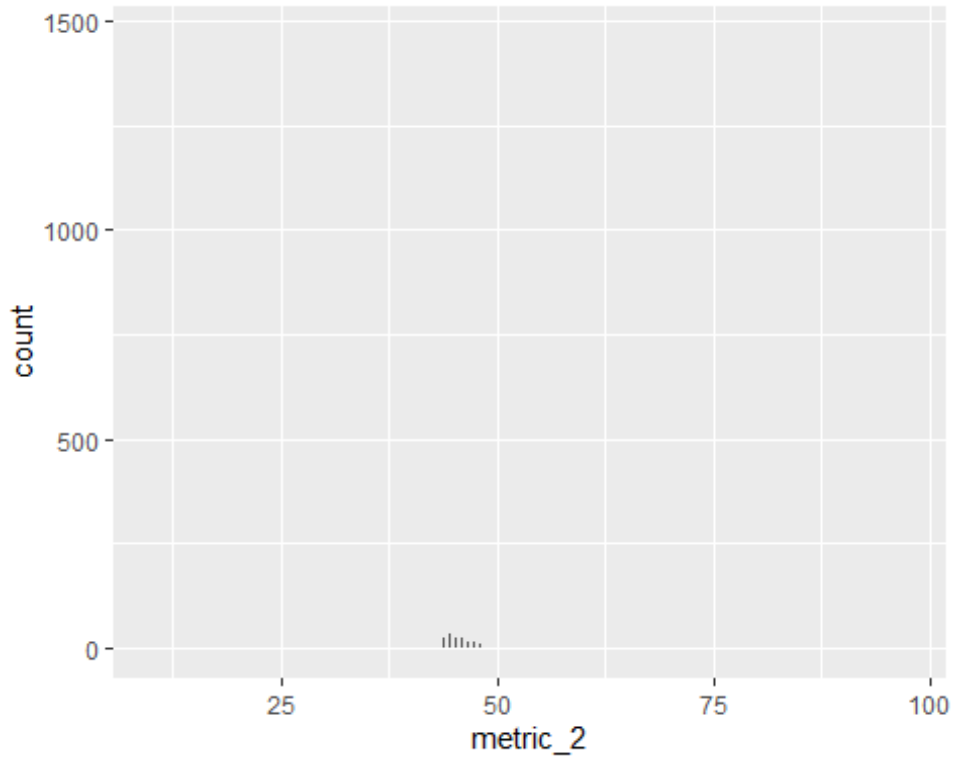
```
## Warning: Removed 2 rows containing non-finite values (`stat_count()`).
```



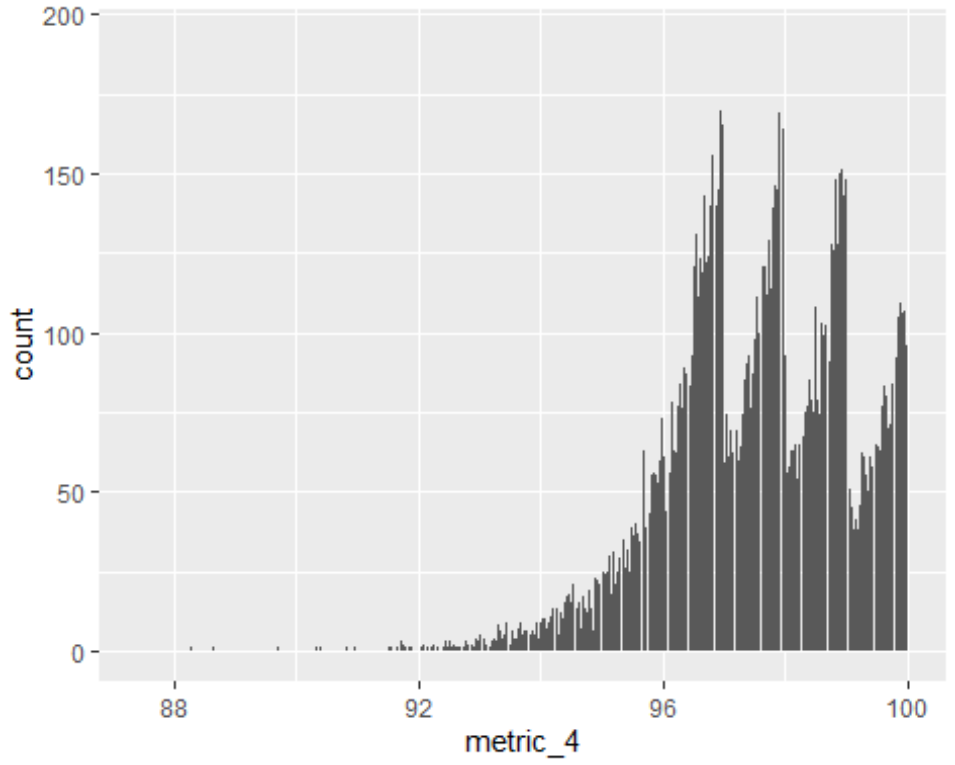

```
ggplot(data= strokedf,aes(x=metric_1))+geom_bar()
```



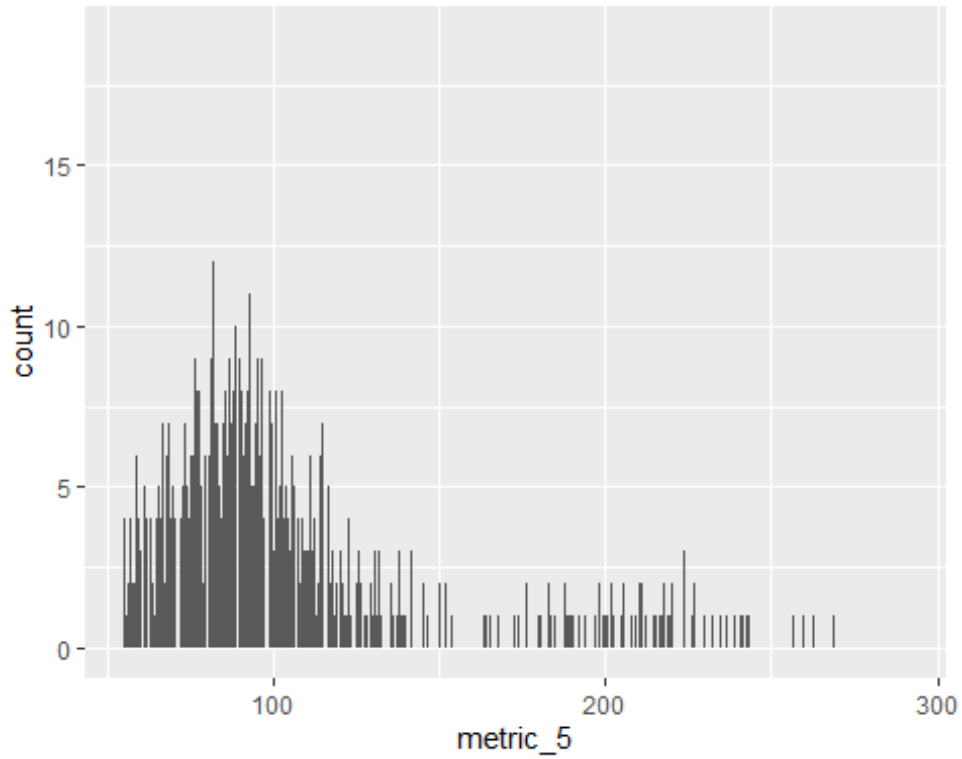
```
ggplot(data= strokedf,aes(x=metric_2))+geom_bar()
```




```
ggplot(data= strokedf,aes(x=metric_4))+geom_bar()
```

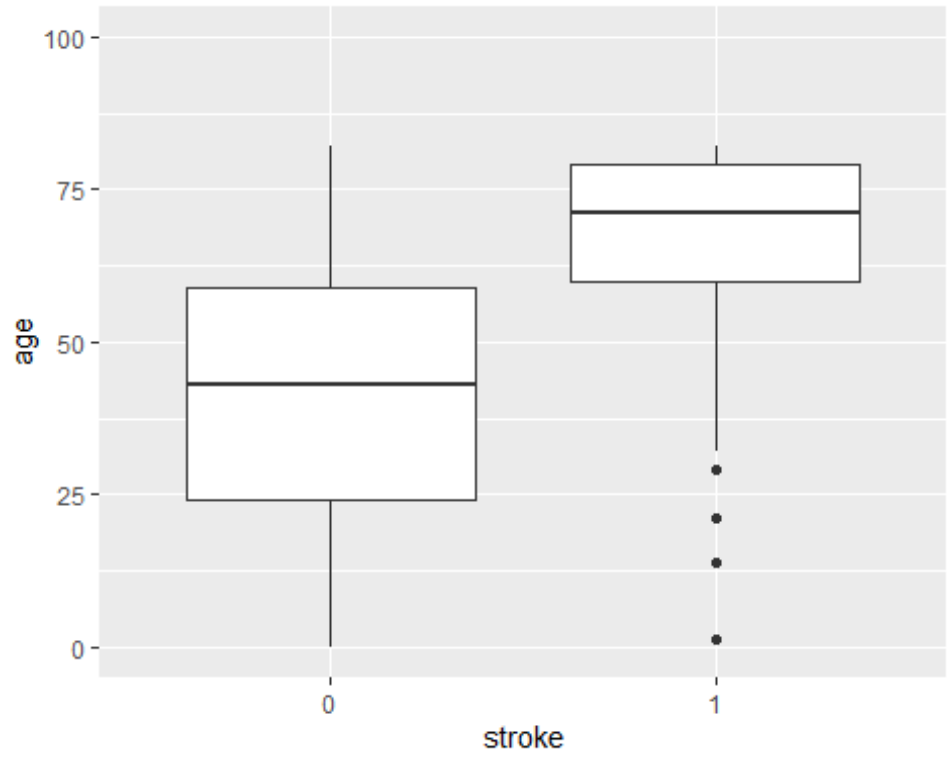


```
ggplot(data= strokedf,aes(x=metric_5))+geom_bar()
```

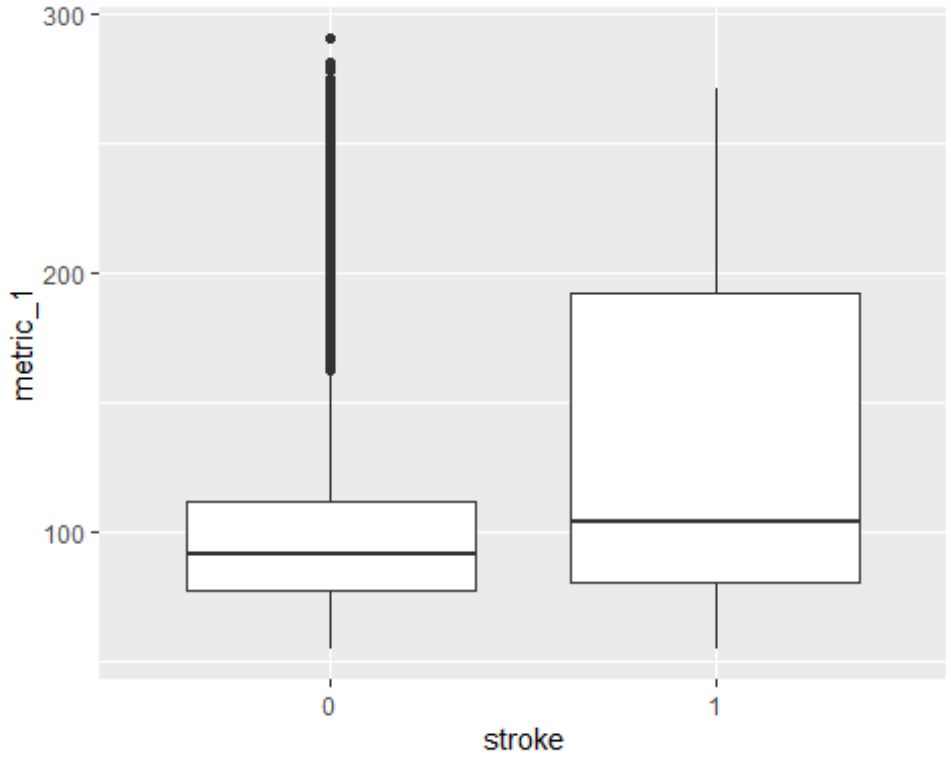



```
ggplot(data= strokedf,aes(x=stroke,y=age))+geom_boxplot()+ylim(0,100)
```

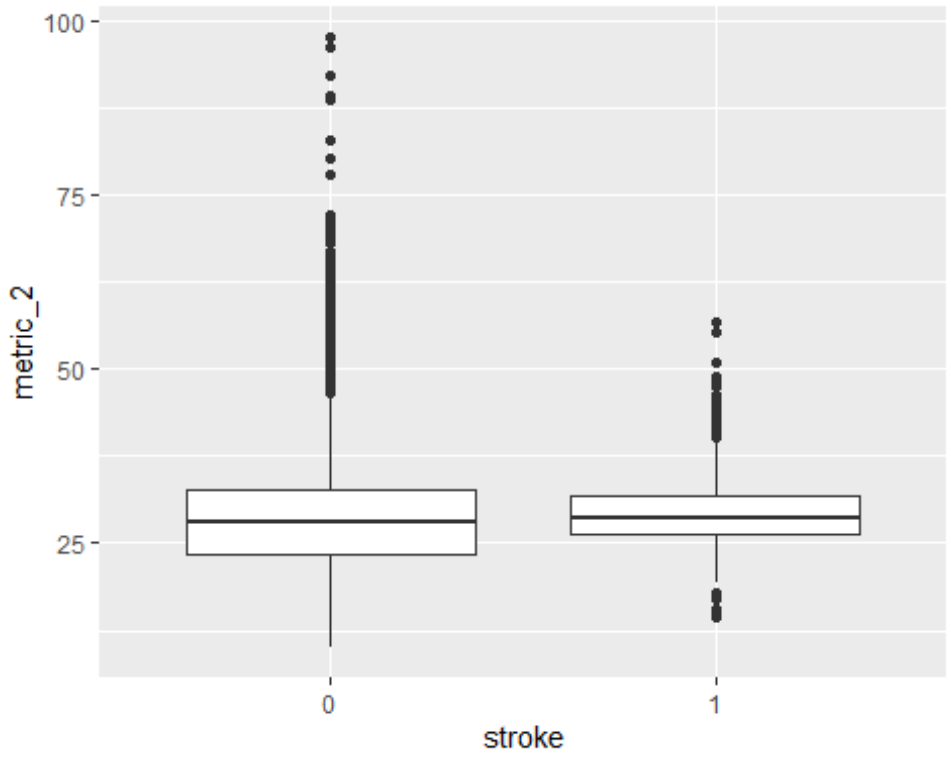
```
## Warning: Removed 2 rows containing non-finite values (`stat_boxplot()`).
```



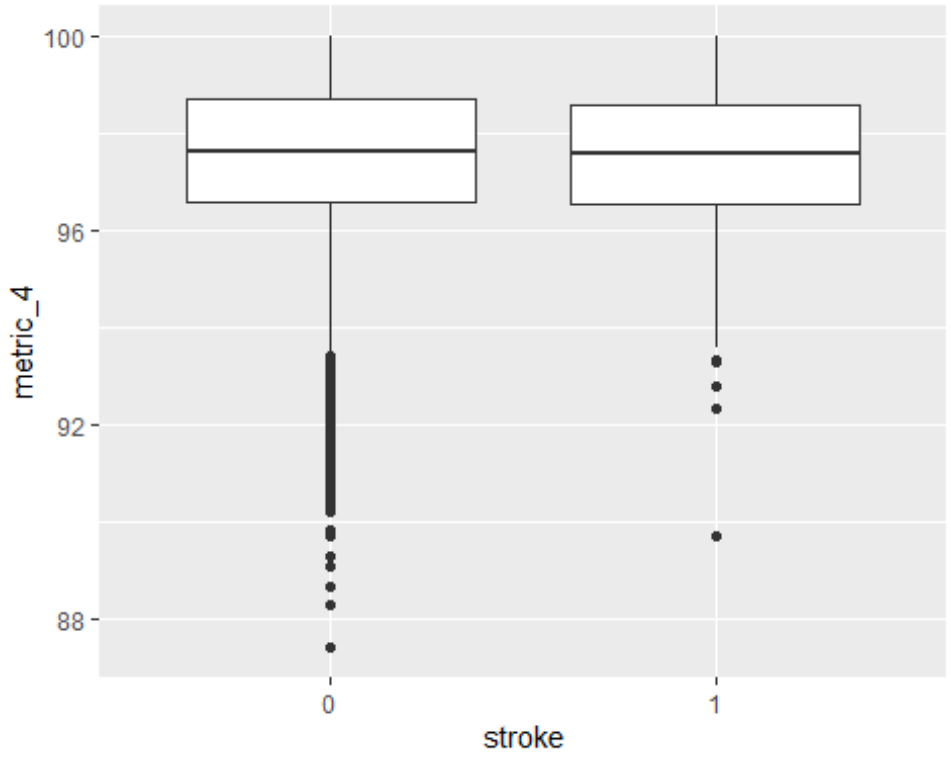

```
ggplot(data= strokedf,aes(x=stroke,y=metric_1))+geom_boxplot()
```



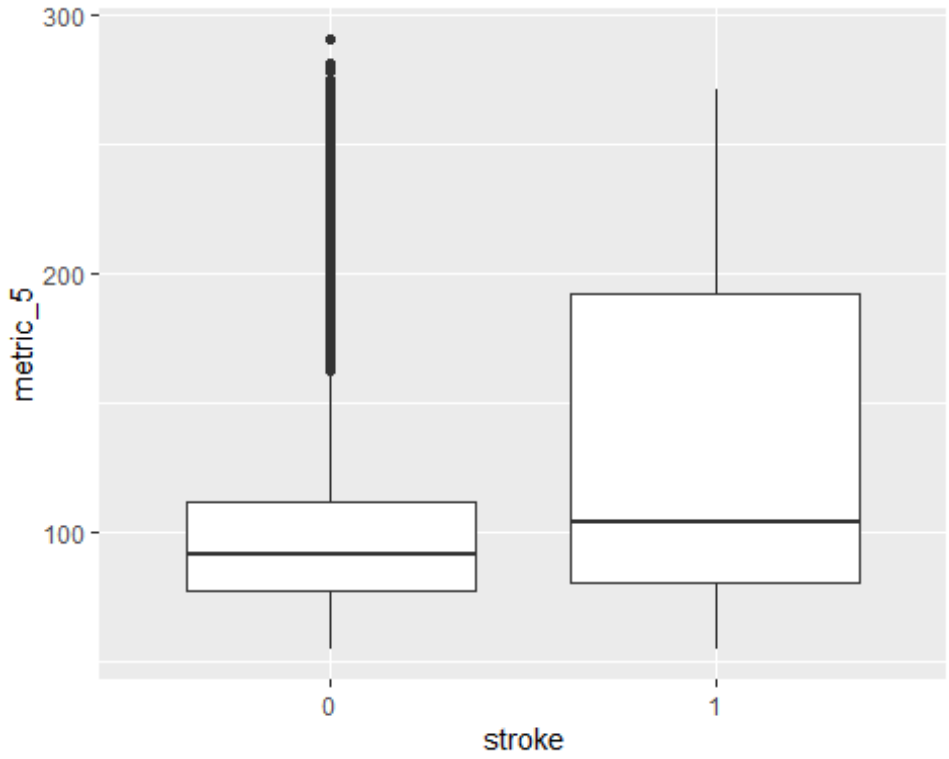

```
ggplot(data= strokedf,aes(x=stroke,y=metric_2))+geom_boxplot()
```




```
ggplot(data= strokedf,aes(x=stroke,y=metric_4))+geom_boxplot()
```




```
ggplot(data= strokedf,aes(x=stroke,y=metric_5))+geom_boxplot()
```




```
#model including all the variables(logistic regression)
strmodel<-glm(stroke~.,data=strokedf,family=binomial)

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

summary(strmodel)

##
## Call:
## glm(formula = stroke ~ ., family = binomial, data = strokedf)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -8.4904  -0.2039  -0.1227  -0.0714   4.2982
##
## Coefficients: (1 not defined because of singularities)
##                                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)                   -4.804e+00  2.641e+00  -1.819   0.0689 .
## genderMale                     7.328e-02  7.577e-02   0.967   0.3335
## genderOther                   -1.139e+01  6.926e+02  -0.016   0.9869
## age                            5.288e-02  2.888e-03  18.313  < 2e-16 ***
## marriedYes                     3.367e-02  1.252e-01   0.269   0.7880
## hypertension1                  3.796e-01  8.802e-02   4.313 1.61e-05 ***
## heart_disease1                 7.216e-01  9.499e-02   7.597 3.04e-14 ***
## occupationB                    2.072e+00  1.020e+00   2.031   0.0422 *
## occupationC                   -8.456e+00  1.754e+02  -0.048   0.9616
## occupationD                    2.137e+00  1.024e+00   2.087   0.0369 *
```

```
## occupationE                     1.939e+00  1.025e+00   1.893   0.0584 .
## residenceUrban                  3.380e-02  7.370e-02   0.459   0.6465
## metric_1                        4.239e-03  6.598e-04   6.424 1.33e-10 ***
## metric_2                       -1.491e-02  6.097e-03  -2.445   0.0145 *
## metric_31                      -4.400e-02  8.220e-02  -0.535   0.5925
## metric_4                       -4.554e-02  2.504e-02  -1.819   0.0690 .
## metric_5                               NA         NA      NA       NA
## smoking_statusformerly smoked   4.483e-02  1.115e-01   0.402   0.6876
## smoking_statusnever smoked     -3.588e-02  1.060e-01  -0.339   0.7350
## smoking_statussmokes            1.334e-01  1.247e-01   1.070   0.2846
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 7839.4  on 43398  degrees of freedom
## Residual deviance: 6618.2  on 43380  degrees of freedom
## AIC: 6656.2
##
## Number of Fisher Scoring iterations: 15
```

```
#reduced the model with the one which are significant using the significant
level of 0.05( significant variables are those with p-value less than 0.05)
reduced_stroke<-glm(stroke ~ age + hypertension + heart_disease + metric_1 +
metric_2, data= strokedf,family= binomial())
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
summary(reduced_stroke)
```

```
##
## Call:
## glm(formula = stroke ~ age + hypertension + heart_disease + metric_1 +
##     metric_2, family = binomial(), data = strokedf)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -8.4904  -0.2023  -0.1200  -0.0684   3.8351
##
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -7.3667876  0.2458120 -29.969  < 2e-16 ***
## age             0.0560437  0.0025610  21.883  < 2e-16 ***
## hypertension1   0.3725149  0.0877426   4.246 2.18e-05 ***
## heart_disease1  0.7342933  0.0939762   7.814 5.56e-15 ***
## metric_1        0.0042405  0.0006582   6.442 1.18e-10 ***
## metric_2       -0.0125436  0.0059433  -2.111   0.0348 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
```

```
##
##     Null deviance: 7839.4  on 43398  degrees of freedom
## Residual deviance: 6622.0  on 43393  degrees of freedom
## AIC: 6634
##
## Number of Fisher Scoring iterations: 9
```

```
#from the above output we can see from the coefficient estimate that metric_2 has a negative effect while the others(age,hypertension,heart disease and metric_1) have a positive effect

#From above reduced model output we can see that
# y = -7.367 + 0.056 * age + 0.373 * hypertension1 + 0.734 * heart_disease1 + 0.004 * metric_1 - 0.013 * metric_2
```

```
library(pscl)

#calculate the R2 for the reduced_stroke model
pscl::pR2(reduced_stroke)["McFadden"]
```

```
## fitting null model for pseudo-r2
```

```
##  McFadden
## 0.1552868
```

```
#higher values indicates more importance and the result matches up the p-value we have seen in the reduced model. Age is the most predictor variable and then heart disease the second predictor variable
caret::varImp(reduced_stroke)
```

```
##                  Overall
## age            21.883104
## hypertension1   4.245544
## heart_disease1  7.813606
## metric_1        6.442165
## metric_2        2.110533
```

```
#Check for multicollinearity
car::vif(reduced_stroke)
```

```
##           age  hypertension heart_disease      metric_1      metric_2
##      1.117587      1.053961      1.084451      1.095118      1.053295
```

```
#since none of the VIF values have above 5, there is no issue of mutlicollinearity
```

```
#prepare the training and test data
sample<-sample(c(TRUE,FALSE),nrow(strokedf),replace=TRUE, prob=c(0.7,0.3))
traindf<-strokedf[sample,]
testdf<-strokedf[!sample,]
```

```r
#building the model on the train dataset
stroke_model<-glm(stroke~ age+hypertension+heart_disease+metric_1+metric_2,
                  family = "binomial",data=traindf)

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

summary(stroke_model)

##
## Call:
## glm(formula = stroke ~ age + hypertension + heart_disease + metric_1 +
##     metric_2, family = "binomial", data = traindf)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -8.4904  -0.2050  -0.1263  -0.0759   3.7573
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -6.9643542  0.2798559 -24.886  < 2e-16 ***
## age             0.0511874  0.0029052  17.620  < 2e-16 ***
## hypertension1   0.3966024  0.1047670   3.786 0.000153 ***
## heart_disease1  0.7066877  0.1127293   6.269 3.64e-10 ***
## metric_1        0.0045861  0.0007832   5.855 4.76e-09 ***
## metric_2       -0.0169053  0.0070893  -2.385 0.017096 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 5545.7  on 30463  degrees of freedom
## Residual deviance: 4752.9  on 30458  degrees of freedom
## AIC: 4764.9
##
## Number of Fisher Scoring iterations: 9

#building the model on test data
stroke_model1<-glm(stroke~ age+hypertension+heart_disease+metric_1+metric_2,
                  family = "binomial",data=testdf)
summary(stroke_model1)

##
## Call:
## glm(formula = stroke ~ age + hypertension + heart_disease + metric_1 +
##     metric_2, family = "binomial", data = testdf)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.8262  -0.1920  -0.1022  -0.0493   3.4259
##
## Coefficients:
```

```
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -8.5892220  0.5152730 -16.669  < 2e-16 ***
## age             0.0705373  0.0054289  12.993  < 2e-16 ***
## hypertension1   0.2960892  0.1612270   1.836  0.06629 .
## heart_disease1  0.8065238  0.1705344   4.729 2.25e-06 ***
## metric_1        0.0033448  0.0012151   2.753  0.00591 **
## metric_2       -0.0001435  0.0109156  -0.013  0.98951
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2293.4  on 12934  degrees of freedom
## Residual deviance: 1874.6  on 12929  degrees of freedom
## AIC: 1886.6
##
## Number of Fisher Scoring iterations: 8
```

```
#prediction on the test dataset
pred<-predict(stroke_model1, newdata = testdf,type ="response")


#taking the probability cutoff as 0.5,if stroke prediction is greater than
0.5, it is stroke else no stroke
stroke_pred_num<-ifelse(pred > 0.5,1,0)
stroke_pred<-factor(stroke_pred_num,levels =c(0,1))
stroke_act<-testdf$stroke

#calculating the accuracy
mean(stroke_pred == stroke_act)
```

```
## [1] 0.9823734
```

```
#In general,from the GLM developed, we can see that age,,metric_1, having
hypertension & heart disease predict the chance of having stroke with the
proportion of accuracy 0.98
```