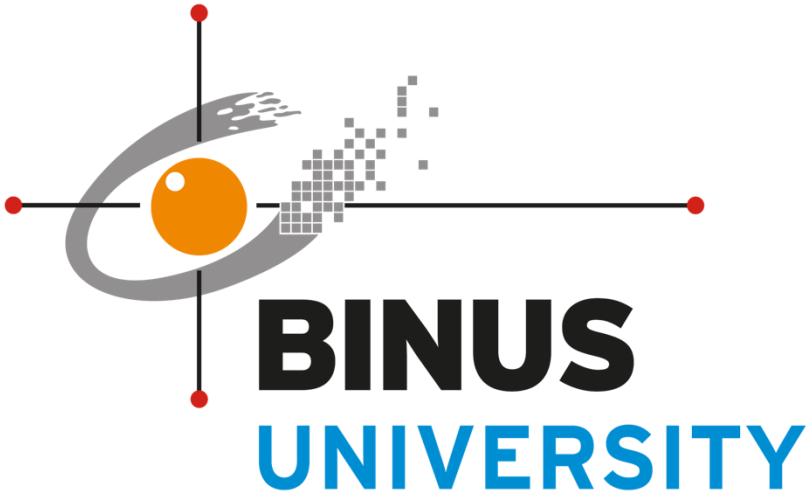
Analisa Data

“Stroke Prediction Dataset”



Disusun oleh:

Patrick Jonathan

2440064791

JURUSAN TEKNIK INFORMATIKA DAN STATISTIKA

PROGRAM STUDI DATA MINING AND VISUALIZATION

UNIVERSITAS BINA NUSANTARA

2021

INTRODUCTION

Pendahuluan

Dataset yang saya pakai untuk dianalisis diambil dari kaggle.com yang berjudul Stroke Prediction Dataset (https://www.kaggle.com/fedesoriano/stroke-prediction-dataset). Alasan saya memilih dataset ini karena dataset ini memiliki variabel yang cukup bervariasi dan saya rasa cukup penting dan menarik untuk dianalisa. Karena menurut Organisasi Kesehatan Dunia (WHO), stroke adalah penyebab kematian ke-2 secara global, bertanggung jawab atas sekitar 11% dari total kematian. Sehingga berdasarkan dataset tersebut, saya akan menganalisa dan memprediksi kemungkinan seorang pasien terkena stroke berdasarkan beberapa ciri-ciri serta faktor kesehatan dari pasien tersebut.

Tujuan

1. Untuk memenuhi Ujian Tengah Semester Mata Kuliah Data Mining and Visualization
2. Melakukan analisis dan memvisualisasikan hasil analisis dengan menggunakan program Rstudio
3. Menganalisa dan memprediksi kemungkinan seorang pasien terkena stroke berdasarkan beberapa ciri-ciri serta faktor kesehatan dari pasien tersebut

DATA DESCRIPTION

Dataset ini digunakan untuk memprediksi kemungkinan seorang pasien terkena stroke berdasarkan ciri-ciri serta faktor kesehatan seperti jenis kelamin, usia, berbagai penyakit, status merokok, dan masih banyak lagi. Dataset ini memiliki 5110 data dengan 12 variabel yang berupa informasi yang relevan tentang pasien.

Berikut adalah beberapa variabel berupa informasi yang relevan tentang pasien :

- id : id untuk setiap pasien
- gender : jenis kelamin setiap pasien ("Male" atau "Female")
- age : umur pasien
- hypertension : mengidentifikasi apakah pasien memiliki penyakit hipertensi (0 jika pasien tidak memiliki penyakit hipertensi dan 1 jika pasien memiliki penyakit hipertensi)
- heart_disease : mengidentifikasi apakah pasien memiliki penyakit jantung (0 jika pasien tidak memiliki penyakit jantung dan 1 jika pasien memiliki penyakit jantung)
- ever_married : mengidentifikasi apakah pasien sudah menikah atau belum (“Yes”, “No”)
- work_type : jenis pekerjaan pasien ("children", "Govt_jov", "Never_worked", "Private", atau "Self-employed")
- Residence_type : tipe pemukiman pasien ("Rural" atau "Urban")
- avg_glucose_level : Rata-rata kadar glukosa dalam darah pasien
- bmi : Indeks massa tubuh pasien
- smoking_status : Status merokok pasien ("formerly smoked", "never smoked", "smokes" atau "Unknown")
- stroke : mengidentifikasi apakah pasien memiliki penyakit stroke (0 jika pasien tidak memiliki penyakit stroke dan 1 jika pasien memiliki penyakit stroke)

DATA EXPLORATION AND VISUALIZATION

1. RStudio Code :

```
# Nama : Patrick Jonathan
# NIM : 2440064791

# Import library yang akan digunakan
library(dplyr)
library(skimr)
library(ggplot2)
library(ggpubr)

# Import csv
strokeDataSet <- read.csv("C:\\Users\\Patrick Jonathan\\Documents\\Patrick
Jonathan\\Sem 3\\UTS\\Data Mining\\healthcare-dataset-stroke-data.csv")
View(strokeDataSet)

# Melihat banyaknya data keseluruhan dan total variabel nya
dim(strokeDataSet)

# Menampilkan 5 data teratas dan 5 data terbawah
head(strokeDataSet,5)
tail(strokeDataSet,5)

# summary
summary(strokeDataSet)
glimpse(strokeDataSet)
skim(strokeDataSet)

# Check duplicate data
sum(duplicated(strokeDataSet))

# Check missing value dalam data
```

```
sum(is.na(strokeDataSet))

# Check adanya outlier dalam data numerik
boxplot(strokeDataSet$age, strokeDataSet$avg_glucose_level, names =
c("Age","Average Glucose Level"), main = "Check Outliers")
# Terdapat outlier dalam data avg_glucose_level

# Menampilkan Density Plot
ggarrange(ggdensity(strokeDataSet$age, fill = 'red'),
          ggdensity(strokeDataSet$avg_glucose_level, fill = 'green'),
          labels= c("Age", "Average Glucose Level")
)

# Menampilkan Scatter Plot
ggarrange(ggqqplot(strokeDataSet$age),
          ggqqplot(strokeDataSet$avg_glucose_level),
          labels= c("Age", "Average Glucose Level")
)

# Visualisasi Histogram
hist(strokeDataSet$age, main = "Age")
hist(strokeDataSet$hypertension, main = "Hypertension")
hist(strokeDataSet$heart_disease, main = "Heart Disease")
hist(strokeDataSet$avg_glucose_level, main = "Average Glucose Level")
hist(strokeDataSet$stroke, main = "Stroke")

# Korelasi antara seluruh variabel numerik
cor(strokeDataSet[,unlist(lapply(strokeDataSet, is.numeric))])

# Split data berdasarkan Ever Married
everMarriedTest = split(strokeDataSet, strokeDataSet$ever_married)

# Check Average Glucose Level dan Age berdasarkan Ever Married
plot(strokeDataSet$age, strokeDataSet$avg_glucose_level,
```

```
    xlab = "Age", ylab = "Average Glucose Level",
    main = "Corelation between Average Glucose Level and Age")
plot(everMarriedTest$Yes$age, everMarriedTest$Yes$avg_glucose_level,
    xlab = "Age", ylab = "Average Glucose Level",
    main = "Corelation between Average Glucose Level and Age (Married)")
plot(everMarriedTest$No$age, everMarriedTest$No$avg_glucose_level,
    xlab = "Age", ylab = "Average Glucose Level",
    main = "Corelation between Average Glucose Level and Age (Not Married)")

cor(strokeDataSet$age, strokeDataSet$avg_glucose_level)
cor(everMarriedTest$Yes$age, everMarriedTest$Yes$avg_glucose_level)
cor(everMarriedTest$No$age, everMarriedTest$No$avg_glucose_level)

# Check Stroke berdasarkan EverMarried
par(mfrow=c(1,2))
hist(everMarriedTest$Yes$stroke, xlab = "Stroke", main = "Ever Married")
hist(everMarriedTest$No$stroke, xlab = "Stroke", main = "Not Married")

# Split data berdasarkan Gender
gender = split(strokeDataSet, strokeDataSet$gender)

# Check Stroke berdasarkan Gender
par(mfrow=c(1,2))
hist(gender$Male$stroke, xlab = "Stroke", main = "Male")
hist(gender$Female$stroke, xlab = "Stroke", main = "Female")

# Split data berdasarkan Hypertension
hyper = split(strokeDataSet, strokeDataSet$hypertension)

# Check Stroke berdasarkan Hypertension
par(mfrow=c(1,2))
hist(hyper$`1`$stroke, xlab = "Stroke", main = "Hypertension")
hist(hyper$`0`$stroke, xlab = "Stroke", main = "Not Hypertension")
```

```
# Split data berdasarkan Heart Disease
heartdis = split(strokeDataSet, strokeDataSet$heart_disease)

# Check Stroke berdasarkan Heart Disease
par(mfrow=c(1,2))
hist(heartdis$`1`$stroke, xlab = "Stroke", main = "Heart Disease")
hist(heartdis$`0`$stroke, xlab = "Stroke", main = "Have no Heart Disease")

# Split data berdasarkan Residence
residence = split(strokeDataSet, strokeDataSet$Residence_type)

# Check Stroke berdasarkan Residence
par(mfrow=c(1,2))
hist(residence$Rural$stroke, xlab = "Stroke", main = "Rural")
hist(residence$Urban$stroke, xlab = "Stroke", main = "Urban")

# Split data berdasarkan Smoking Status
smoke = split(strokeDataSet, strokeDataSet$smoking_status)

# Check Stroke berdasarkan Smoking Status
par(mfrow=c(2,2))
hist(smoke$smokes$stroke, xlab = "Stroke", main = "Smokes")
hist(smoke$`never smoked`$stroke, xlab = "Stroke", main = "Never Smoked")
hist(smoke$`formerly smoked`$stroke, xlab = "Stroke", main = "Formerly Smoked")
hist(smoke$Unknown$stroke, xlab = "Stroke", main = "Unknown")

# Check korelasi antara stroke dengan variabel numerik lainnya
cor(strokeDataSet$stroke, strokeDataSet$age)
cor(strokeDataSet$stroke, strokeDataSet$hypertension)
cor(strokeDataSet$stroke, strokeDataSet$heart_disease)
cor(strokeDataSet$stroke, strokeDataSet$avg_glucose_level)
```

2. Visualization :

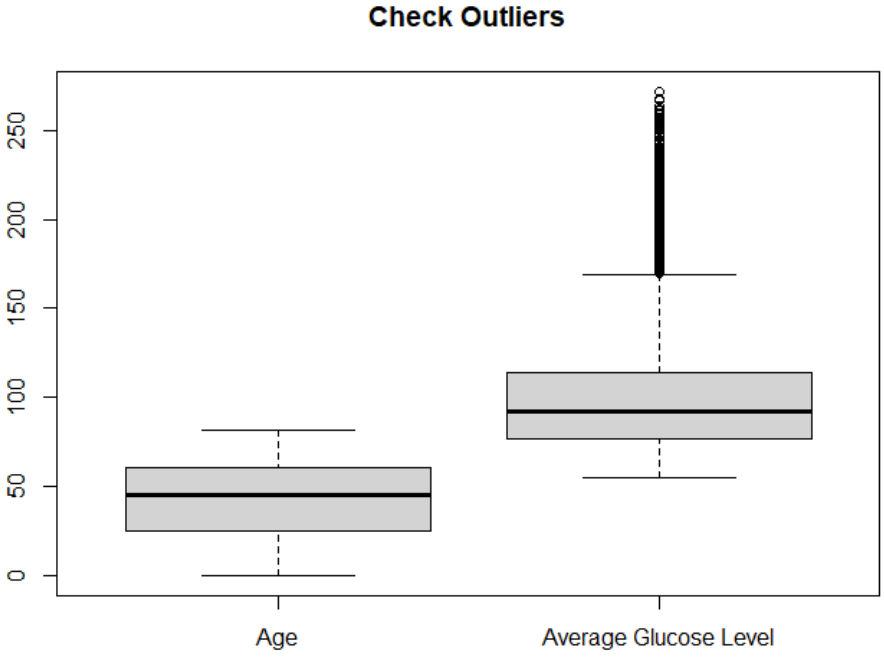
```
> # Import library yang akan digunakan
> library(dplyr)

Attaching package: ‘dplyr’

The following objects are masked from ‘package:stats’:

    filter, lag

The following objects are masked from ‘package:base’:

    intersect, setdiff, setequal, union

> library(skimr)
> library(ggplot2)
> library(ggpubr)
> # Import csv
> strokeDataSet <- read.csv("C:\\Users\\Patrick Jonathan\\Documents\\Patrick Jonathan\\Sem 3\\UTS\\Data Mining\\healthcare-dataset-stroke-data.csv")
> View(strokeDataSet)
> |
```

```
> # Melihat banyaknya data keseluruhan dan total variabel nya
> dim(strokeDataSet)
[1] 5110   12
> # Menampilkan 5 data teratas dan 5 data terbawah
> head(strokeDataSet,5)
     id gender age hypertension heart_disease ever_married     work_type Residence_type avg_glucose_level  bmi  smoking_status stroke
1  9046   Male  67            0             1          Yes       Private          Urban            228.69 36.6 formerly smoked      1
2 51676 Female  61            0             0          Yes Self-employed          Rural            202.21  N/A    never smoked      1
3 31112   Male  80            0             1          Yes       Private          Rural            105.92 32.5    never smoked      1
4 60182 Female  49            0             0          Yes       Private          Urban            171.23 34.4          smokes      1
5  1665 Female  79            1             0          Yes Self-employed          Rural            174.12   24    never smoked      1
> tail(strokeDataSet,5)
       id gender age hypertension heart_disease ever_married     work_type Residence_type avg_glucose_level  bmi  smoking_status stroke
5106 18234 Female  80            1             0          Yes       Private          Urban             83.75  N/A    never smoked      0
5107 44873 Female  81            0             0          Yes Self-employed          Urban            125.20   40    never smoked      0
5108 19723 Female  35            0             0          Yes Self-employed          Rural             82.99 30.6    never smoked      0
5109 37544   Male  51            0             0          Yes       Private          Rural            166.29 25.6 formerly smoked      0
5110 44679 Female  44            0             0          Yes      Govt_job          Urban             85.28 26.2         Unknown      0
> # summary
> summary(strokeDataSet)
       id            gender               age         hypertension     heart_disease     ever_married        work_type
 Min.   :   67   Length:5110        Min.   : 0.08   Min.   :0.00000   Min.   :0.00000   Length:5110        Length:5110
 1st Qu.:17741   Class :character   1st Qu.:25.00   1st Qu.:0.00000   1st Qu.:0.00000   Class :character   Class :character
 Median :36932   Mode  :character   Median :45.00   Median :0.00000   Median :0.00000   Mode  :character   Mode  :character
 Mean   :36518                      Mean   :43.23   Mean   :0.09746   Mean   :0.05401
 3rd Qu.:54682                      3rd Qu.:61.00   3rd Qu.:0.00000   3rd Qu.:0.00000
 Max.   :72940                      Max.   :82.00   Max.   :1.00000   Max.   :1.00000
 Residence_type     avg_glucose_level     bmi            smoking_status         stroke
 Length:5110        Min.   : 55.12    Length:5110        Length:5110        Min.   :0.00000
 Class :character   1st Qu.: 77.25    Class :character   Class :character   1st Qu.:0.00000
 Mode  :character   Median : 91.89    Mode  :character   Mode  :character   Median :0.00000
                    Mean   :106.15                                          Mean   :0.04873
                    3rd Qu.:114.09                                          3rd Qu.:0.00000
                    Max.   :271.74                                          Max.   :1.00000
> |
```

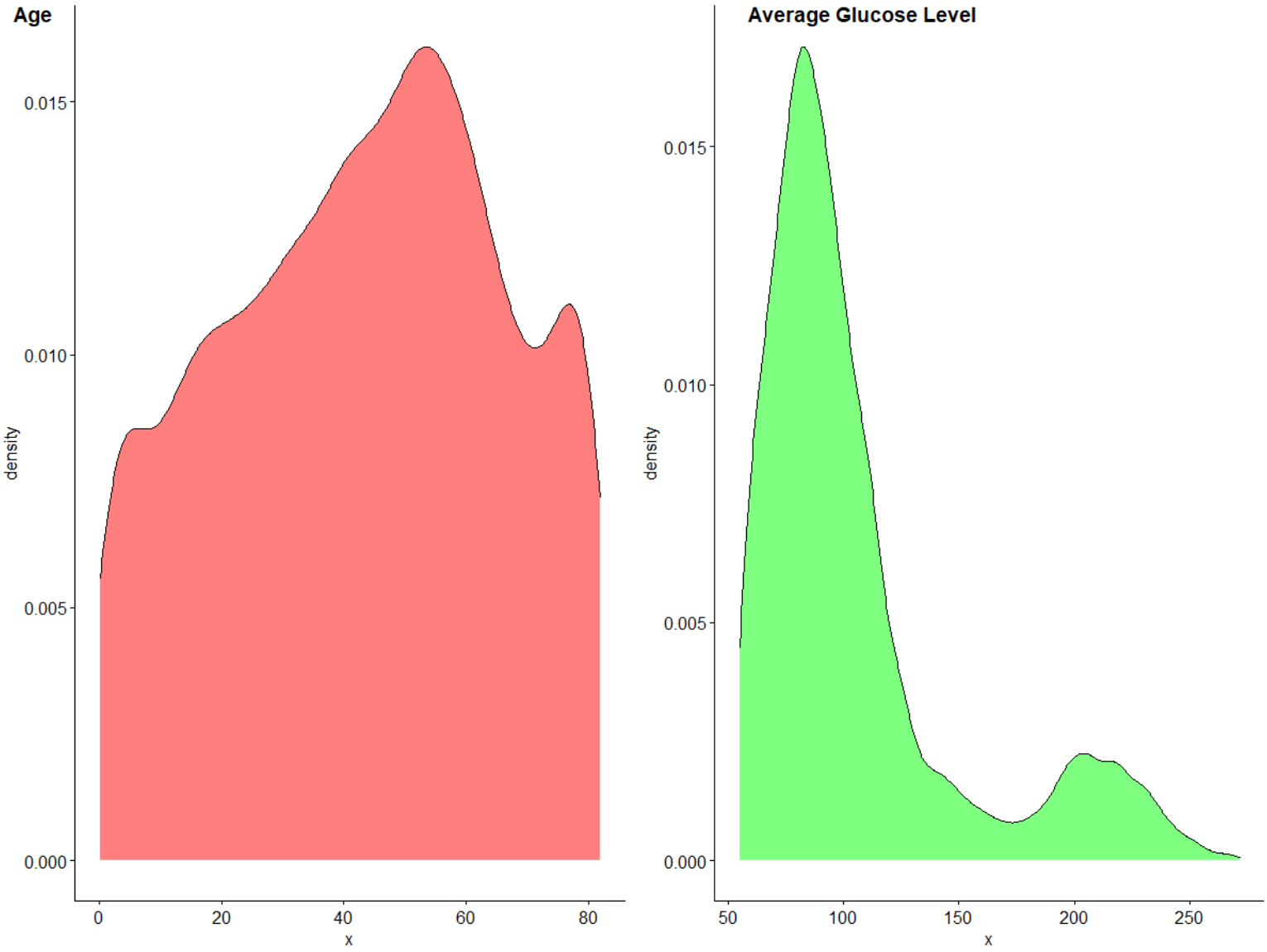
```
> glimpse(strokeDataSet)
Rows: 5,110
Columns: 12
$ id               <int> 9046, 51676, 31112, 60182, 1665, 56669, 53882, 10434, 27419, 60491, 12109, 12095, 12175, 8213, 5317, 58202, 56112,~
$ gender           <chr> "Male", "Female", "Male", "Female", "Female", "Male", "Male", "Female", "Female", "Female", "Female", "Female", "F~
$ age              <dbl> 67, 61, 80, 49, 79, 81, 74, 69, 59, 78, 81, 61, 54, 78, 79, 50, 64, 75, 60, 57, 71, 52, 79, 82, 71, 80, 65, 58, 69~
$ hypertension     <int> 0, 0, 0, 0, 1, 0, 1, 0, 0, 0, 1, 0, 0, 0, 0, 1, 0, 1, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 1, 0, 0, 1, 0, 1, ~
$ heart_disease    <int> 1, 0, 1, 0, 0, 0, 1, 0, 0, 0, 1, 0, 1, 1, 0, 1, 0, 0, 1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 1, 0, 1, 0, 0, ~
$ ever_married     <chr> "Yes", "Yes", "Yes", "Yes", "Yes", "Yes", "Yes", "No", "Yes", "Yes", "Yes", "Yes", "Yes", "Yes", "Yes", "Yes", "Ye~
$ work_type        <chr> "Private", "Self-employed", "Private", "Private", "Self-employed", "Private", "Private", "Private", "Private", "Pr~
$ Residence_type   <chr> "Urban", "Rural", "Rural", "Urban", "Rural", "Urban", "Rural", "Urban", "Rural", "Urban", "Rural", "Rural", "Urban~
$ avg_glucose_level <dbl> 228.69, 202.21, 105.92, 171.23, 174.12, 186.21, 70.09, 94.39, 76.15, 58.57, 80.43, 120.46, 104.51, 219.84, 214.09,~
$ bmi              <chr> "36.6", "N/A", "32.5", "34.4", "24", "29", "27.4", "22.8", "N/A", "24.2", "29.7", "36.8", "27.3", "N/A", "28.2", "~
$ smoking_status   <chr> "formerly smoked", "never smoked", "never smoked", "smokes", "never smoked", "formerly smoked", "never smoked", "n~
$ stroke           <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ~
```

```
> skim(strokeDataSet)
-- Data Summary ------------------------
                           Values
Name                       strokeDataSet
Number of rows             5110
Number of columns          12
_____________________
Column type frequency:
  character                6
  numeric                  6
_____________________
Group variables            None

-- Variable type: character ------------------------------------------------------------------------------------------
# A tibble: 6 x 8
  skim_variable  n_missing complete_rate   min   max empty n_unique whitespace
* <chr>              <int>         <dbl> <int> <int> <int>    <int>      <int>
1 gender                 0             1     4     6     0        3          0
2 ever_married           0             1     2     3     0        2          0
3 work_type              0             1     7    13     0        5          0
4 Residence_type         0             1     5     5     0        2          0
5 bmi                    0             1     2     4     0      419          0
6 smoking_status         0             1     6    15     0        4          0

-- Variable type: numeric --------------------------------------------------------------------------------------------
# A tibble: 6 x 11
  skim_variable     n_missing complete_rate       mean        sd    p0     p25     p50    p75   p100 hist
* <chr>                 <int>         <dbl>      <dbl>     <dbl> <dbl>   <dbl>   <dbl>  <dbl>  <dbl> <chr>
1 id                        0             1 36518.     21162.    67    17741.  36932   54682  72940  ▇▇▇▇▇
2 age                       0             1    43.2       22.6    0.08    25      45      61     82  ▅▆▆▆▇
3 hypertension              0             1     0.0975     0.297  0        0       0       0      1  ▇▁▁▁▁
4 heart_disease             0             1     0.0540     0.226  0        0       0       0      1  ▇▁▁▁▁
5 avg_glucose_level         0             1   106.        45.3   55.1     77.2    91.9   114.   272. ▇▂▁▁▁
6 stroke                    0             1     0.0487     0.215  0        0       0       0      1  ▇▁▁▁▁
> |
```

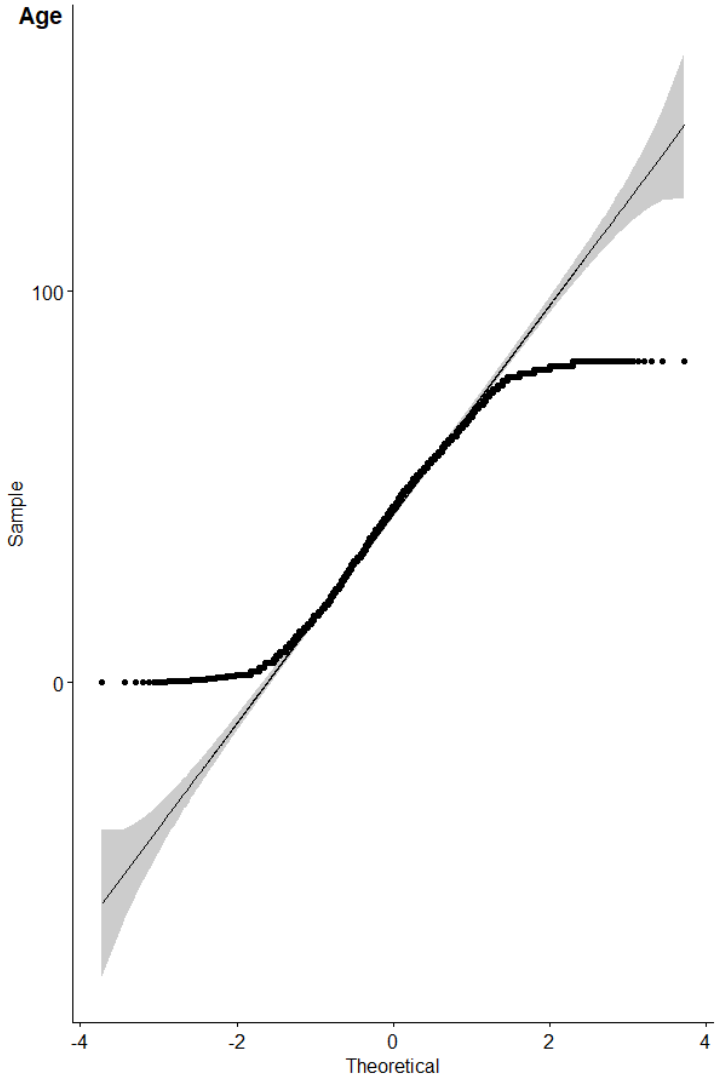
```
> # Check duplicate data
> sum(duplicated(strokeDataSet))
[1] 0
> # Check missing value dalam data
> sum(is.na(strokeDataSet))
[1] 0
> # Check adanya outlier dalam data numerik
> boxplot(strokeDataSet$age, strokeDataSet$avg_glucose_level, names = c("Age","Average Glucose Level"), main = "Check Outliers")
> # Terdapat outlier dalam data avg_glucose_level
> |
```

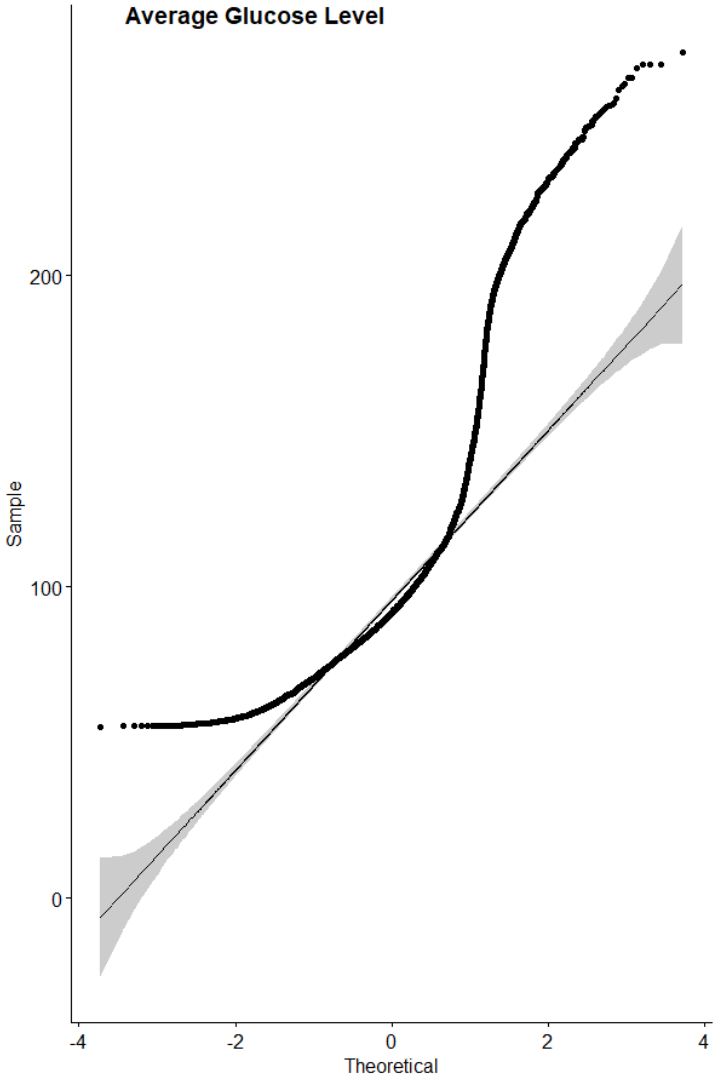


```
> # Menampilkan Density Plot
> ggarrange(ggdensity(strokeDataSet$age, fill = 'red'),
+           ggdensity(strokeDataSet$avg_glucose_level, fill = 'green'),
+           labels= c("Age", "Average Glucose Level")
+ )
>
```

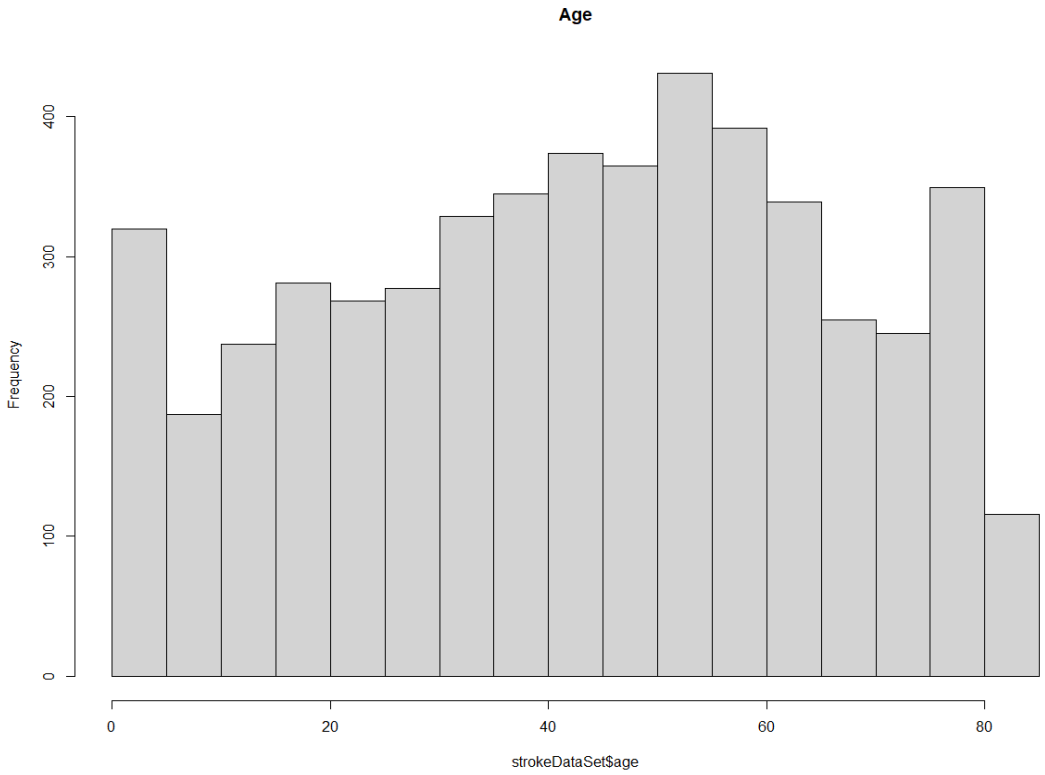


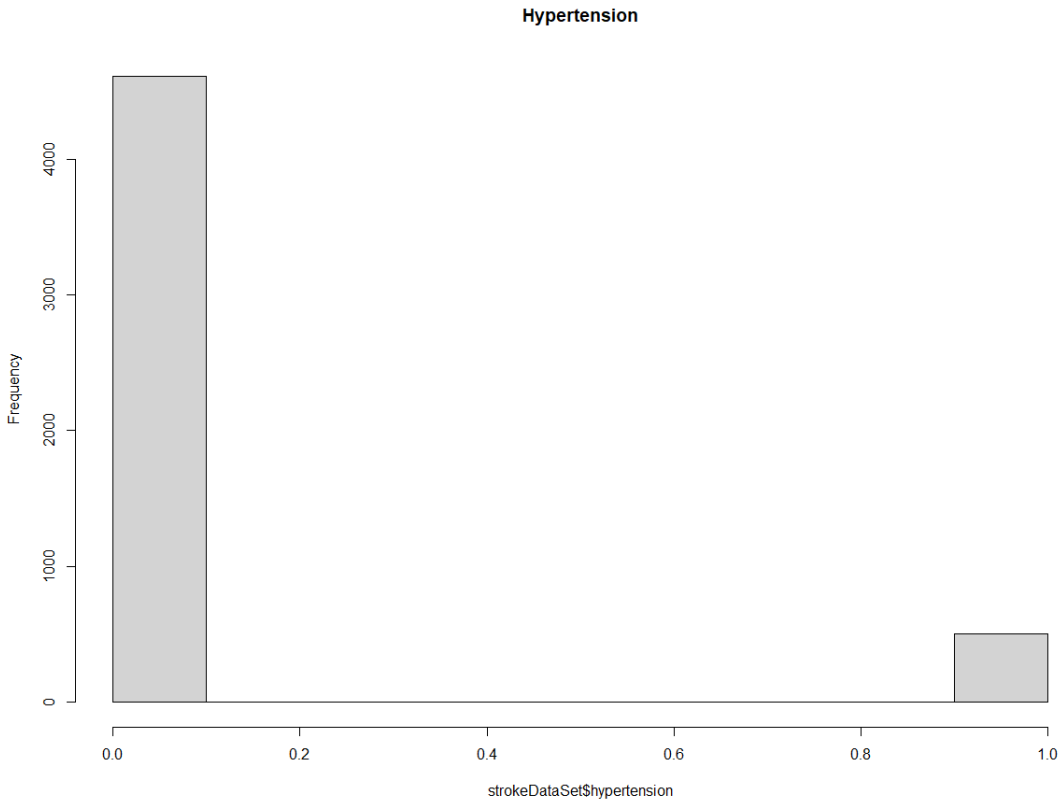
```
> # Menampilkan Scatter Plot
> ggarrange(ggqqplot(strokeDataSet$age),
+           ggqqplot(strokeDataSet$avg_glucose_level),
+           labels= c("Age", "Average Glucose Level")
+ )
>
```

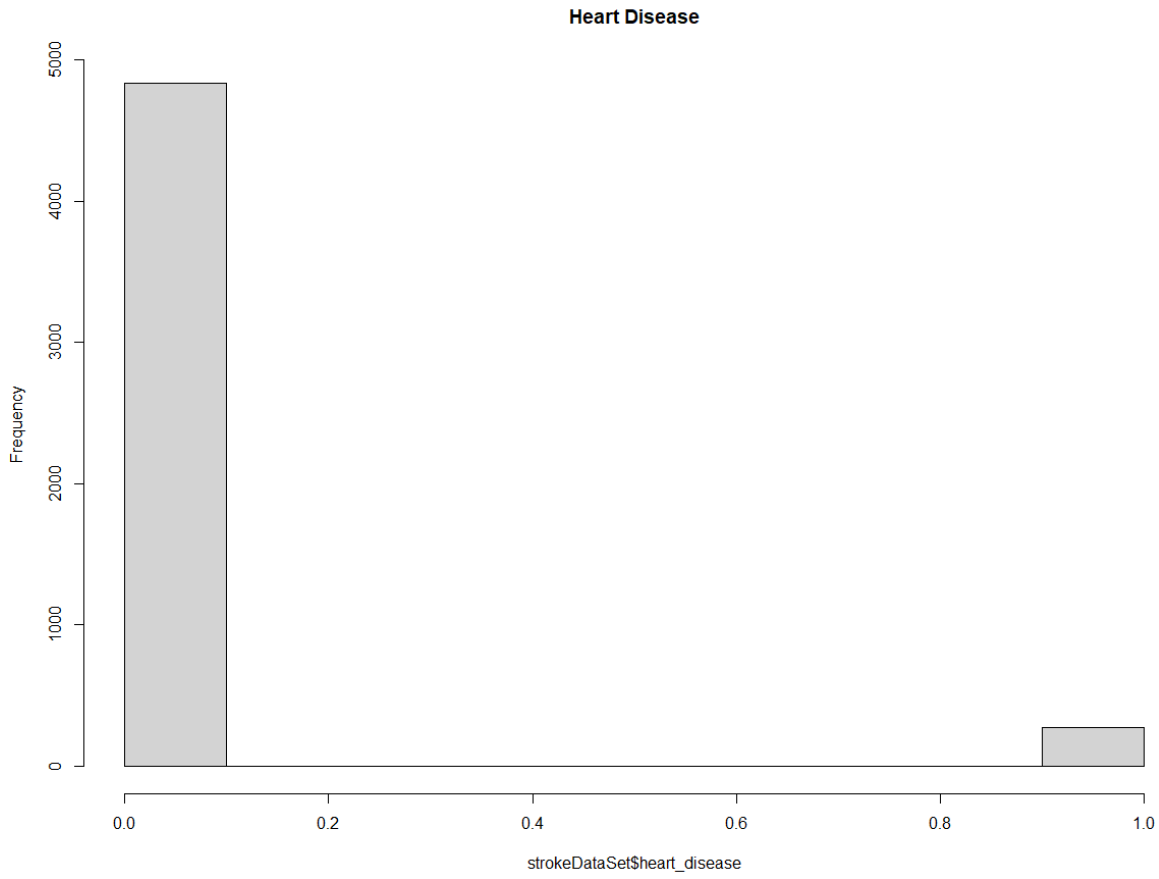


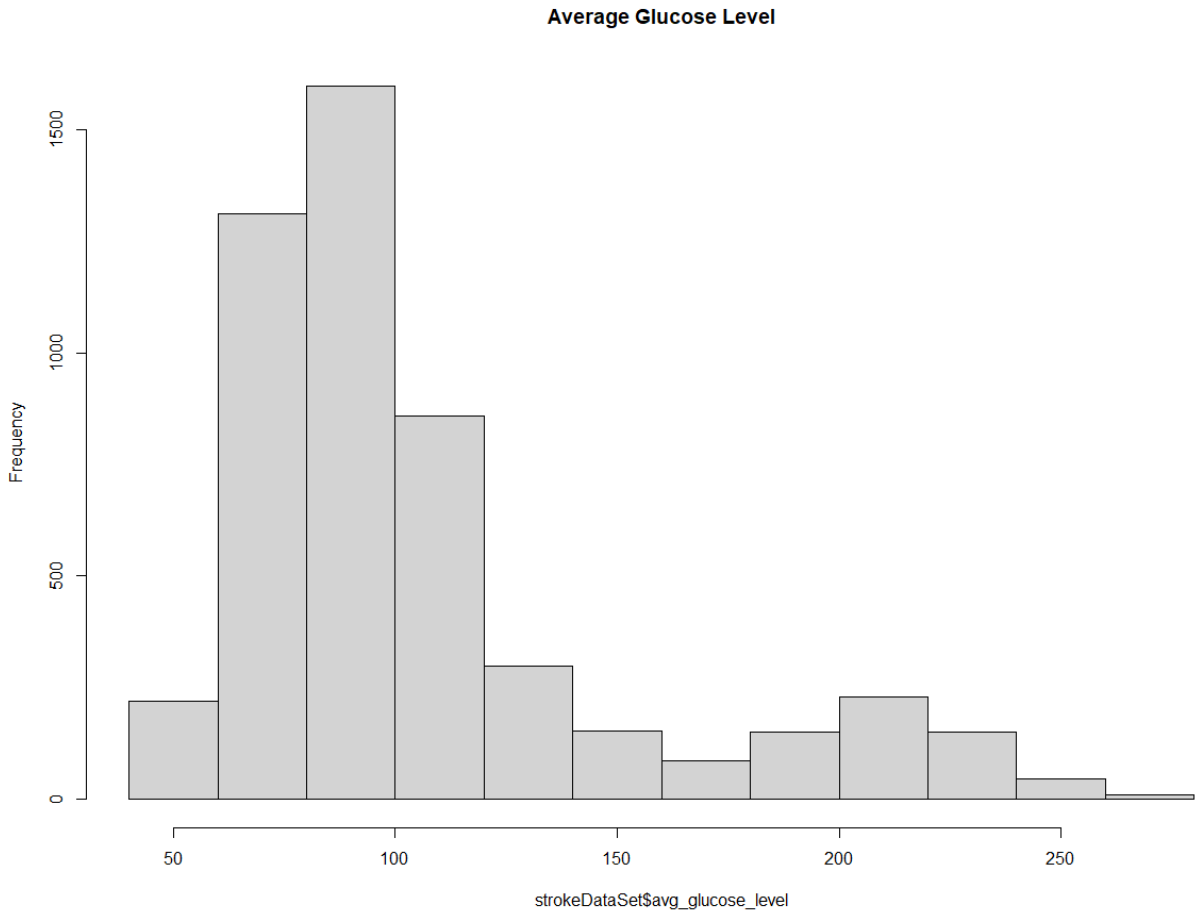


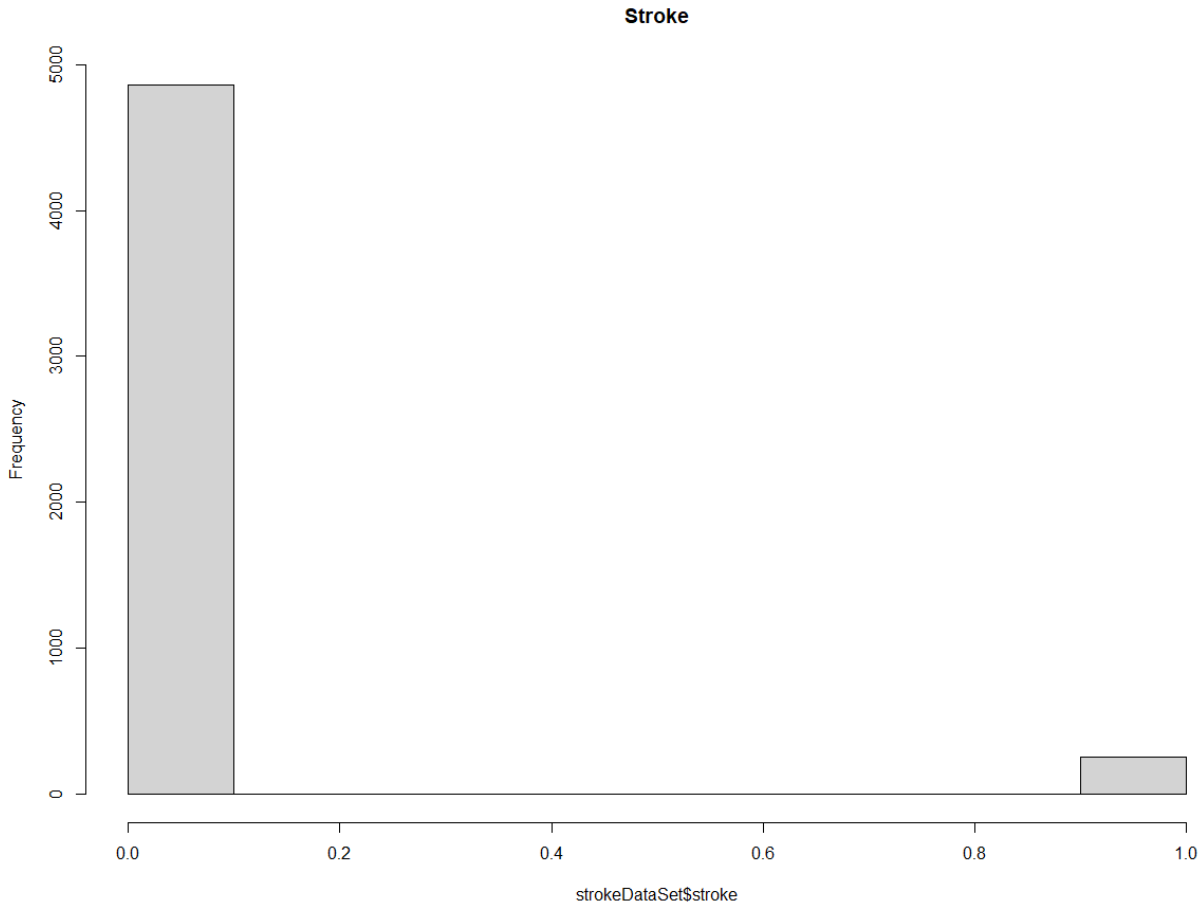
```
> # Visualisasi Histogram
> hist(strokeDataSet$age, main = "Age")
> hist(strokeDataSet$hypertension, main = "Hypertension")
> hist(strokeDataSet$heart_disease, main = "Heart Disease")
> hist(strokeDataSet$avg_glucose_level, main = "Average Glucose Level")
> hist(strokeDataSet$stroke, main = "Stroke")
>
```





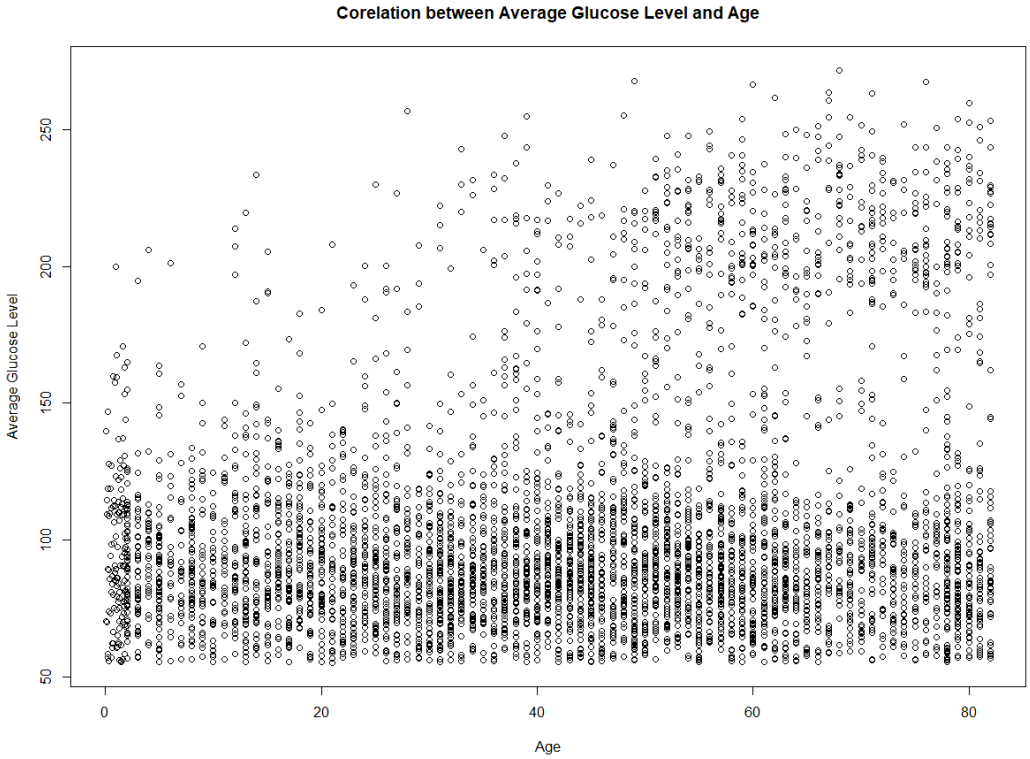

Heart Disease
Frequency
0
1000
2000
3000
4000
5000
0.0
0.2
0.4
0.6
0.8
1.0
strokeDataSet$heart_disease


Average Glucose Level
Frequency
0
500
1000
1500
50
100
150
200
250
strokeDataSet$avg_glucose_level

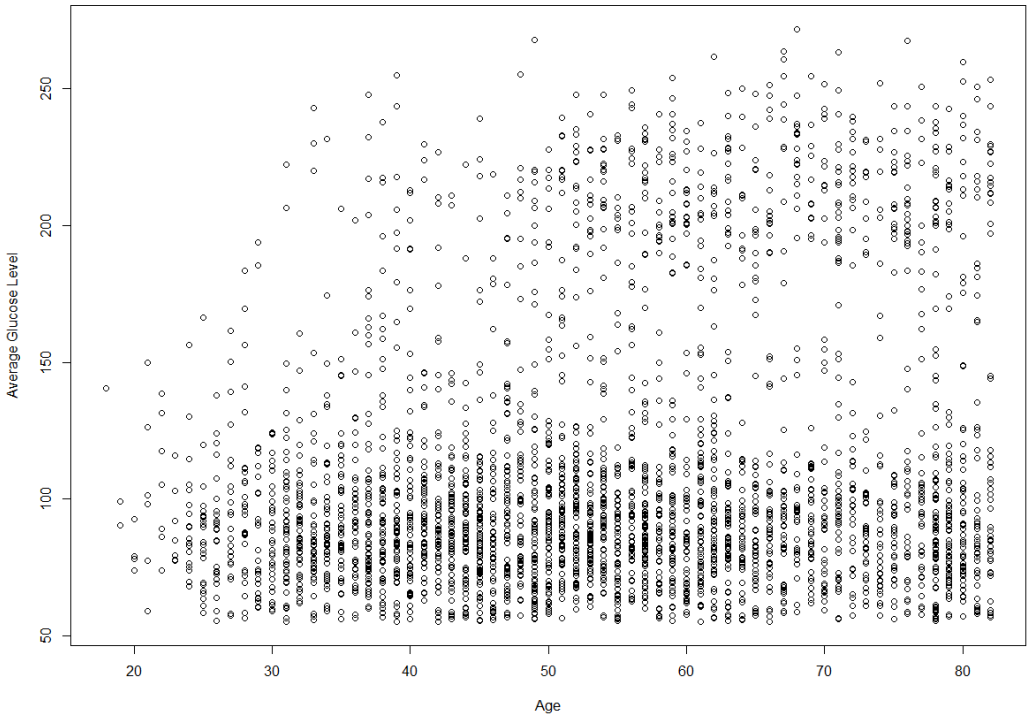



```
> cor(strokeDataSet[,unlist(lapply(strokeDataSet, is.numeric))])
                        id         age hypertension heart_disease avg_glucose_level     stroke
id                1.000000000 0.003538065  0.003549615  -0.001295941       0.001092355 0.00638817
age               0.003538065 1.000000000  0.276397628   0.263795916       0.238171114 0.24525735
hypertension      0.003549615 0.276397628  1.000000000   0.108306076       0.174473811 0.12790382
heart_disease    -0.001295941 0.263795916  0.108306076   1.000000000       0.161857332 0.13491400
avg_glucose_level 0.001092355 0.238171114  0.174473811   0.161857332       1.000000000 0.13194544
stroke            0.006388170 0.245257346  0.127903823   0.134913997       0.131945441 1.00000000
```

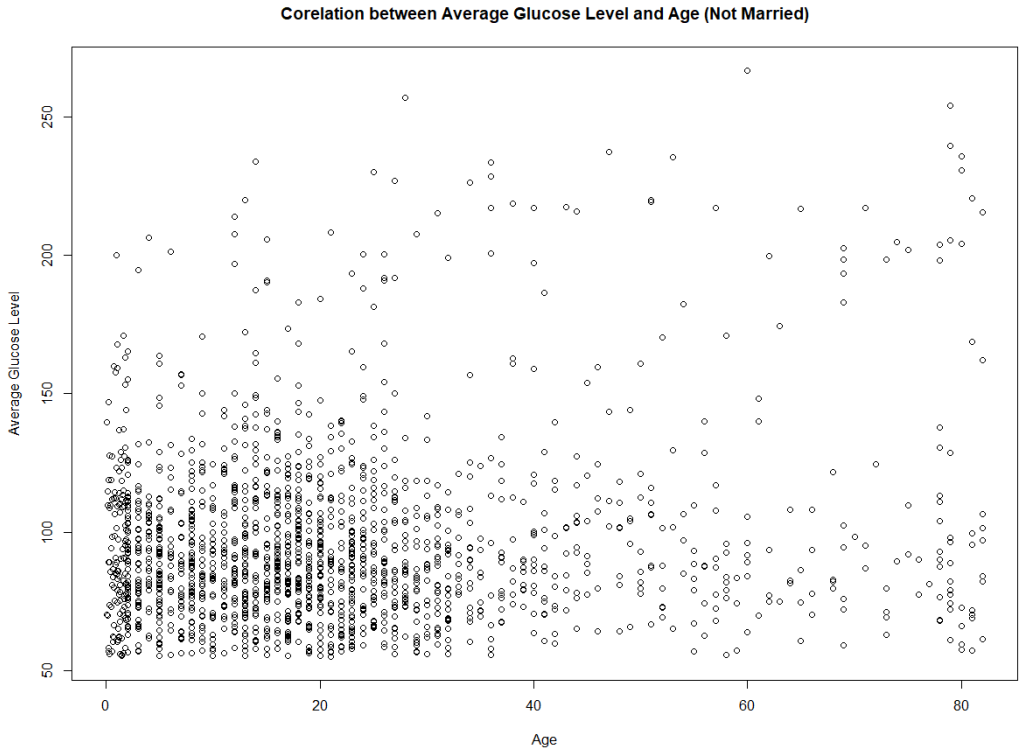
```
> # Split data berdasarkan Ever Married
> everMarriedTest = split(strokeDataSet, strokeDataSet$ever_married)
> # Check Average Glucose Level dan Age berdasarkan Ever Married
> plot(strokeDataSet$age, strokeDataSet$avg_glucose_level,
+      xlab = "Age", ylab = "Average Glucose Level",
+      main = "Corelation between Average Glucose Level and Age")
>
```



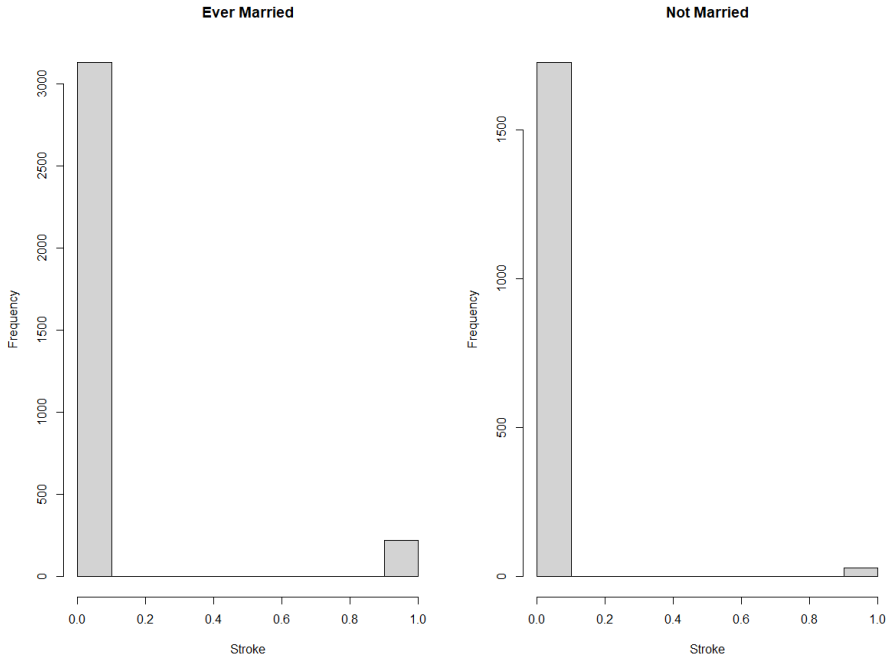

```
> plot(everMarriedTest$Yes$age, everMarriedTest$Yes$avg_glucose_level,
+      xlab = "Age", ylab = "Average Glucose Level",
+      main = "Corelation between Average Glucose Level and Age (Married)")
>
```



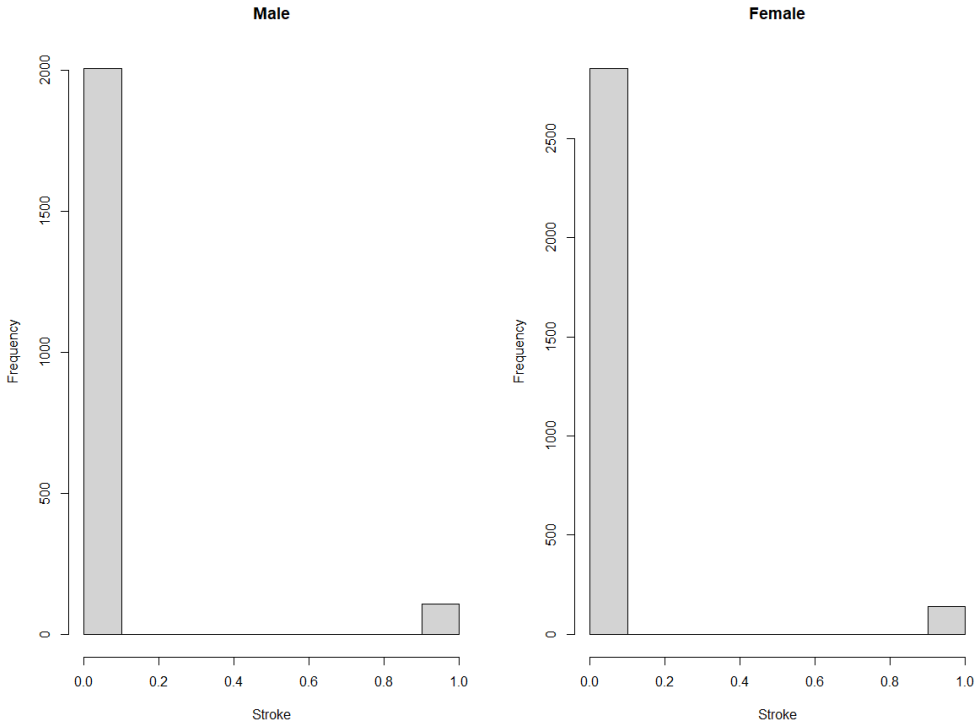
```
> plot(everMarriedTest$No$age, everMarriedTest$No$avg_glucose_level,
+      xlab = "Age", ylab = "Average Glucose Level",
+      main = "Corelation between Average Glucose Level and Age (Not Married)")
> |
```



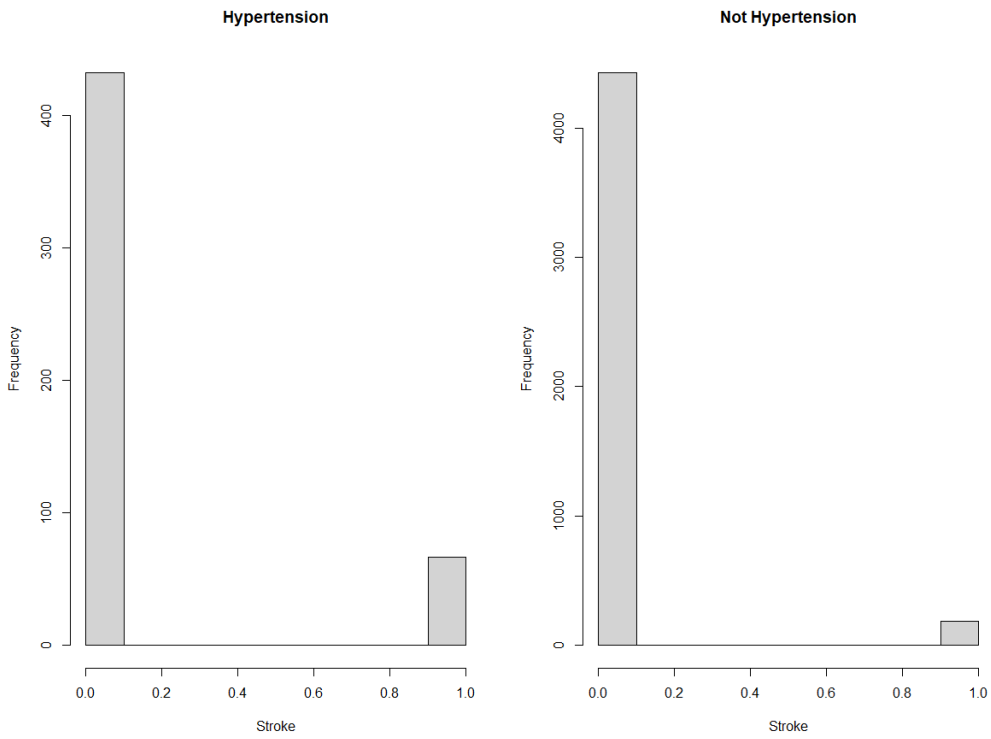

```
> cor(strokeDataSet$age, strokeDataSet$avg_glucose_level)
[1] 0.2381711
> cor(everMarriedTest$Yes$age, everMarriedTest$Yes$avg_glucose_level)
[1] 0.2094176
> cor(everMarriedTest$No$age, everMarriedTest$No$avg_glucose_level)
[1] 0.142774
> |
> # Check Stroke berdasarkan EverMarried
> par(mfrow=c(1,2))
> hist(everMarriedTest$Yes$stroke, xlab = "Stroke", main = "Ever Married")
> hist(everMarriedTest$No$stroke, xlab = "Stroke", main = "Not Married")
> |
```



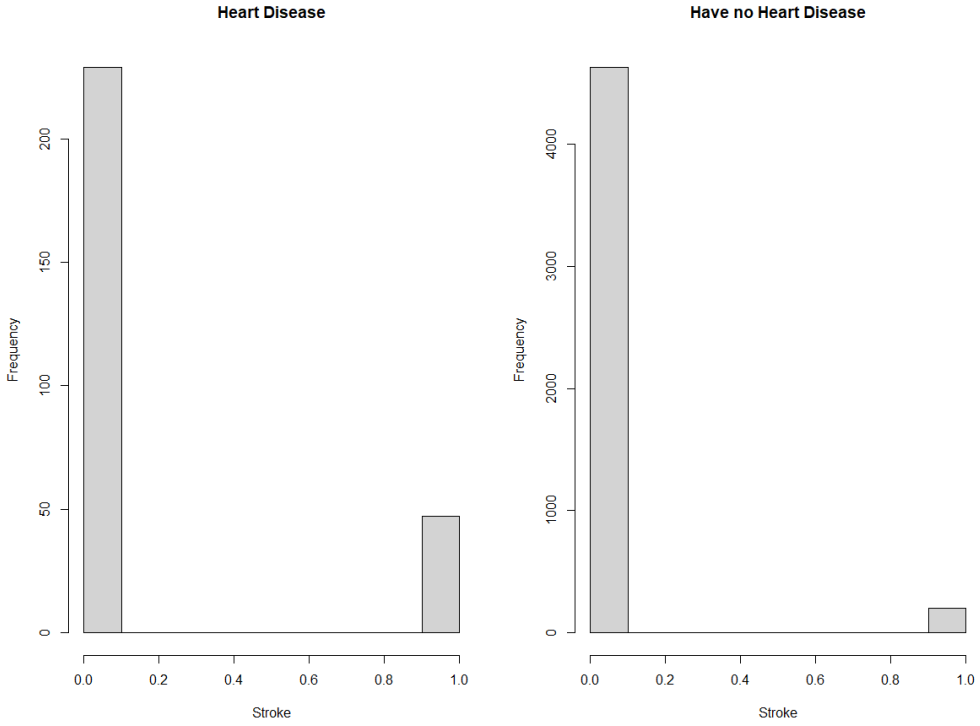
```
> # Split data berdasarkan Gender
> gender = split(strokeDataSet, strokeDataSet$gender)
> # Check Stroke berdasarkan Gender
> par(mfrow=c(1,2))
> hist(gender$Male$stroke, xlab = "Stroke", main = "Male")
> hist(gender$Female$stroke, xlab = "Stroke", main = "Female")
>
```



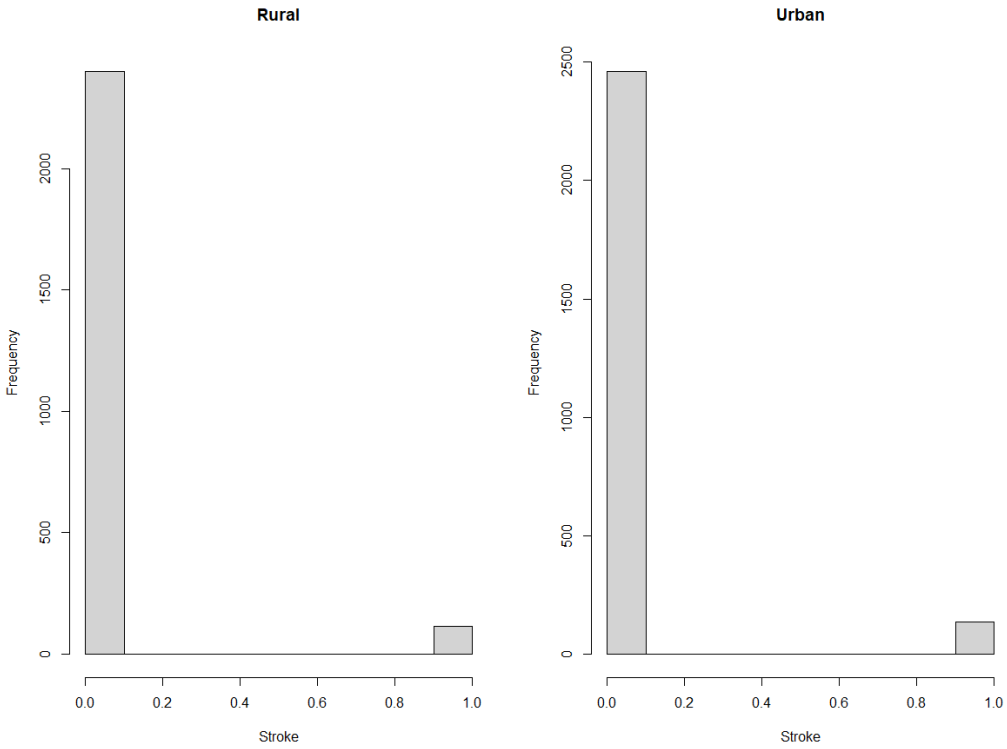

```
> # Split data berdasarkan Hypertension
> hyper = split(strokeDataSet, strokeDataSet$hypertension)
> # Check Stroke berdasarkan Hypertension
> par(mfrow=c(1,2))
> hist(hyper$`1`$stroke, xlab = "Stroke", main = "Hypertension")
> hist(hyper$`0`$stroke, xlab = "Stroke", main = "Not Hypertension")
>
```



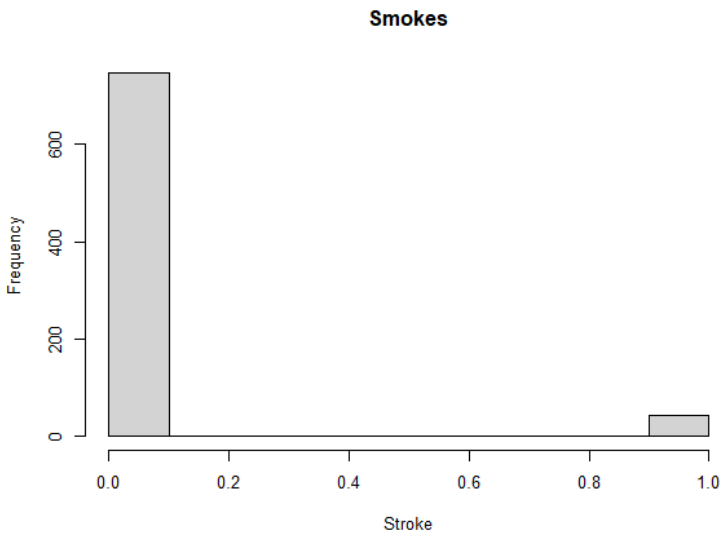
```
> # Split data berdasarkan Heart Disease
> heartdis = split(strokeDataSet, strokeDataSet$heart_disease)
> # Check Stroke berdasarkan Heart Disease
> par(mfrow=c(1,2))
> hist(heartdis$`1`$stroke, xlab = "Stroke", main = "Heart Disease")
> hist(heartdis$`0`$stroke, xlab = "Stroke", main = "Have no Heart Disease")
>
```

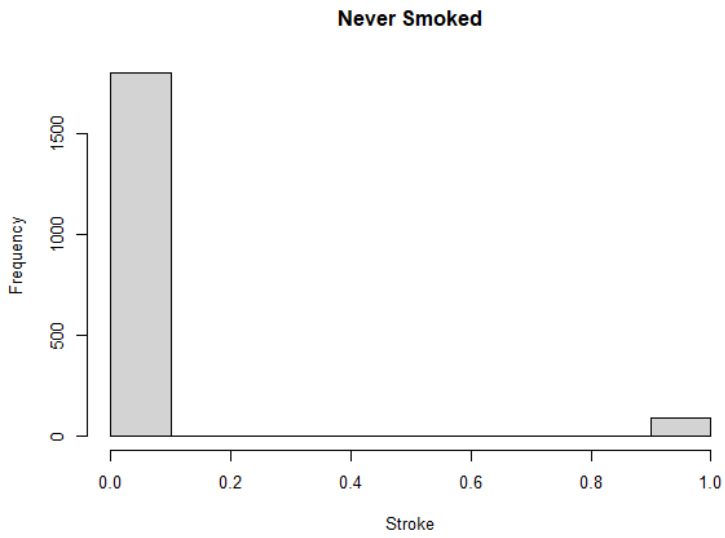


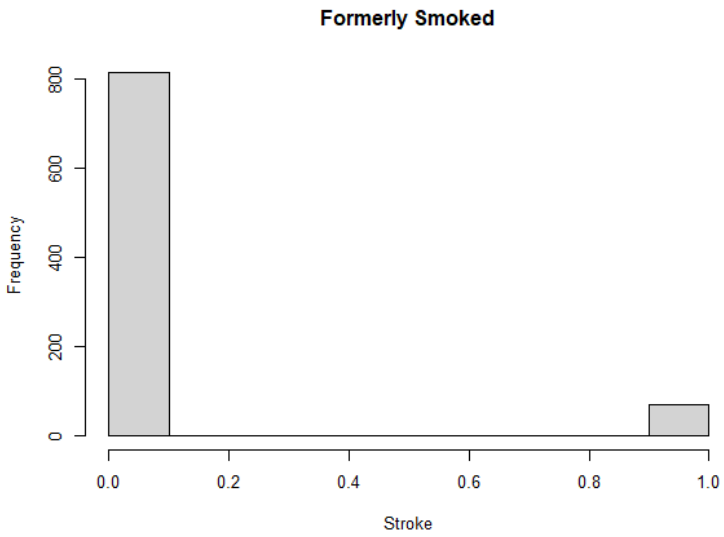

```
> # Split data berdasarkan Residence
> residence = split(strokeDataSet, strokeDataSet$Residence_type)
> # Check Stroke berdasarkan Residence
> par(mfrow=c(1,2))
> hist(residence$Rural$stroke, xlab = "Stroke", main = "Rural")
> hist(residence$Urban$stroke, xlab = "Stroke", main = "Urban")
>
```

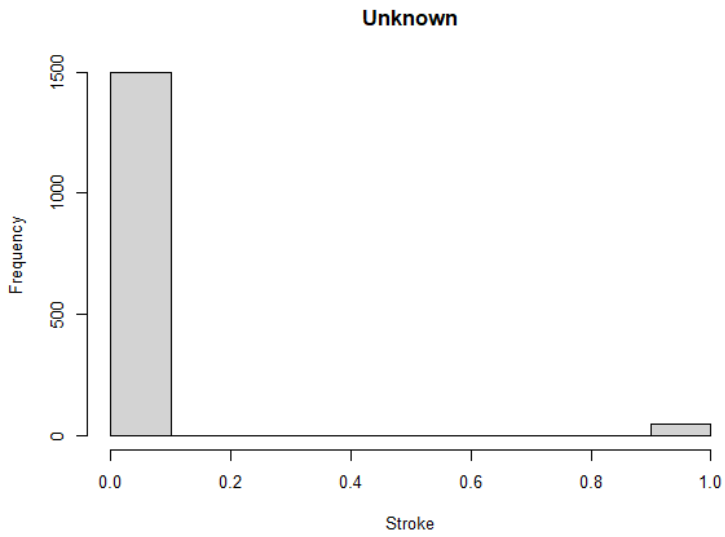


```
> # Split data berdasarkan Smoking Status
> smoke = split(strokeDataSet, strokeDataSet$smoking_status)
> # Check Stroke berdasarkan Smoking Status
> par(mfrow=c(2,2))
> hist(smoke$smokes$stroke, xlab = "Stroke", main = "Smokes")
> hist(smoke$`never smoked`$stroke, xlab = "Stroke", main = "Never Smoked")
> hist(smoke$`formerly smoked`$stroke, xlab = "Stroke", main = "Formerly Smoked")
> hist(smoke$Unknown$stroke, xlab = "Stroke", main = "Unknown")
>
```










```
> # Check korelasi antara stroke dengan variabel numerik lainnya
> cor(strokeDataSet$stroke, strokeDataSet$age)
[1] 0.2452573
> cor(strokeDataSet$stroke, strokeDataSet$hypertension)
[1] 0.1279038
> cor(strokeDataSet$stroke, strokeDataSet$heart_disease)
[1] 0.134914
> cor(strokeDataSet$stroke, strokeDataSet$avg_glucose_level)
[1] 0.1319454
>
```

Discussion / Analysis

Data set Stroke Prediction memiliki total 5110 data dengan 12 variabel berupa 6 variabel numerik dan 6 variabel kategorik. Data set ini tidak memiliki dublicate data maupun missing data. Namun data set ini memiliki outlier pada variabel numerik avg_glucose_level. Bentuk distribusi nya ditunjukkan dengan kurva density, scatter plot, dan histogram untuk variabel numerik. Pada kedua variabel, baik variabel age maupun variabel avg_glucose_level menunjukkan bentuk distribusi yang tidak normal. Karena pada kurva density, bentuk pada variabel age condong ke arah kiri (left-skewed), sedangkan bentuk pada variabel avg_glucose_level condong ke arah kanan (right-skewed). Pada scatter plot juga dapat dilihat bahwa kedua variabel tersebut tidak berdistribusi normal.

Nilai korelasi berkisar diantara -1 (korelasi lemah) sampai 1 (korelasi kuat). Semakin besar angka korelasi nya maka korelasi antar kedua variabel semakin kuat. Korelasi yang ditunjukkan antar variabel numerik menunjukkan angka yang cukup kecil. Meskipun begitu korelasi antara variabel age dan stroke menunjukkan angka korelasi yang paling besar, yang artinya semakin tua umur dari pasien maka semakin tinggi peluang pasien tersebut terkena penyakit stroke. Korelasi yang paling kecil ditunjukkan oleh korelasi antara variabel heart disease dan variabel id, yang artinya id dari seorang pasien tidak ada hubungannya dengan peluang pasien tersebut terkena penyakit jantung.

Selanjutnya pada histogram menampilkan bahwa, pasien yang telah menikah memiliki peluang terkena penyakit stroke yang lebih besar dibandingkan pasien yang belum menikah, pasien perempuan memiliki peluang terkena penyakit stroke yang lebih besar dibandingkan pasien laki-laki, pasien yang tidak memiliki penyakit hipertensi memiliki peluang terkena penyakit stroke yang lebih besar dibandingkan pasien yang memiliki penyakit hipertensi, pasien yang tidak memiliki penyakit jantung memiliki peluang terkena penyakit stroke yang lebih besar dibandingkan pasien yang memiliki penyakit jantung, pasien yang bermukim di “urban” memiliki peluang terkena penyakit stroke yang lebih besar dibandingkan pasien yang bermukim di “rural”, pasien yang tidak merokok memiliki peluang terkena penyakit stroke yang lebih besar dibandingkan dengan pasien yang merokok.

Kesimpulan :

Semua variabel memiliki hubungan atau mempengaruhi variabel stroke, walaupun memiliki nilai korelasi yang kecil. Beberapa variabel numerik seperti age, hypertension, heart_disease, avg_glucose_level, sedikit mempengaruhi variabel stroke. Namun variabel yang paling mempengaruhi adalah variabel age, dimana semakin tua umur dari pasien maka semakin tinggi peluang pasien tersebut terkena penyakit stroke. Menurut histogram, beberapa variabel kategorik lainnya juga cukup mempengaruhi seperti pasien yang sudah menikah, pasien perempuan, pasien yang bermukim di “urban”, serta pasien yang tidak merokok memiliki peluang terkena penyakit stroke yang lebih tinggi.