portfolio\_R.R

2024-01-02

#### Case Study - Data analysis of stroke prediction dataset ####  
  
### Source: https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset ###  
  
  
### setting the project directory and loading packages ###  
setwd("C:/Users/ewwci/OneDrive/Desktop/r studies/case\_study")  
  
# packages needed   
  
library(tidyverse)

## Warning: package 'ggplot2' was built under R version 4.3.2

## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──  
## ✔ dplyr 1.1.3 ✔ readr 2.1.4  
## ✔ forcats 1.0.0 ✔ stringr 1.5.0  
## ✔ ggplot2 3.4.4 ✔ tibble 3.2.1  
## ✔ lubridate 1.9.2 ✔ tidyr 1.3.0  
## ✔ purrr 1.0.2   
## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ dplyr::lag() masks stats::lag()  
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors

library(janitor)

##   
## Attaching package: 'janitor'  
##   
## The following objects are masked from 'package:stats':  
##   
## chisq.test, fisher.test

library(corrplot)

## Warning: package 'corrplot' was built under R version 4.3.2

## corrplot 0.92 loaded

options(repos = c(CRAN = "https://cloud.r-project.org"))  
install.packages("caTools")

## Installing package into 'C:/Users/ewwci/AppData/Local/R/win-library/4.3'  
## (as 'lib' is unspecified)

## package 'caTools' successfully unpacked and MD5 sums checked

## Warning: cannot remove prior installation of package 'caTools'

## Warning in file.copy(savedcopy, lib, recursive = TRUE): problem copying  
## C:\Users\ewwci\AppData\Local\R\win-library\4.3\00LOCK\caTools\libs\x64\caTools.dll  
## to C:\Users\ewwci\AppData\Local\R\win-library\4.3\caTools\libs\x64\caTools.dll:  
## Permission denied

## Warning: restored 'caTools'

##   
## The downloaded binary packages are in  
## C:\Users\ewwci\AppData\Local\Temp\RtmpMb4TYS\downloaded\_packages

### importing data and checking the structure ###  
  
stroke <- read.csv("healthcare-dataset-stroke-data.csv")  
str(stroke)

## 'data.frame': 5110 obs. of 12 variables:  
## $ id : int 9046 51676 31112 60182 1665 56669 53882 10434 27419 60491 ...  
## $ gender : chr "Male" "Female" "Male" "Female" ...  
## $ age : num 67 61 80 49 79 81 74 69 59 78 ...  
## $ hypertension : int 0 0 0 0 1 0 1 0 0 0 ...  
## $ heart\_disease : int 1 0 1 0 0 0 1 0 0 0 ...  
## $ ever\_married : chr "Yes" "Yes" "Yes" "Yes" ...  
## $ work\_type : chr "Private" "Self-employed" "Private" "Private" ...  
## $ Residence\_type : chr "Urban" "Rural" "Rural" "Urban" ...  
## $ avg\_glucose\_level: num 229 202 106 171 174 ...  
## $ bmi : chr "36.6" "N/A" "32.5" "34.4" ...  
## $ smoking\_status : chr "formerly smoked" "never smoked" "never smoked" "smokes" ...  
## $ stroke : int 1 1 1 1 1 1 1 1 1 1 ...

head(stroke)

## id gender age hypertension heart\_disease ever\_married work\_type  
## 1 9046 Male 67 0 1 Yes Private  
## 2 51676 Female 61 0 0 Yes Self-employed  
## 3 31112 Male 80 0 1 Yes Private  
## 4 60182 Female 49 0 0 Yes Private  
## 5 1665 Female 79 1 0 Yes Self-employed  
## 6 56669 Male 81 0 0 Yes Private  
## Residence\_type avg\_glucose\_level bmi smoking\_status stroke  
## 1 Urban 228.69 36.6 formerly smoked 1  
## 2 Rural 202.21 N/A never smoked 1  
## 3 Rural 105.92 32.5 never smoked 1  
## 4 Urban 171.23 34.4 smokes 1  
## 5 Rural 174.12 24 never smoked 1  
## 6 Urban 186.21 29 formerly smoked 1

#### Data transformation - Data cleaning ####  
  
# unifying column names   
stroke <- select\_all(stroke, tolower)  
  
# checking new format  
head(stroke)

## id gender age hypertension heart\_disease ever\_married work\_type  
## 1 9046 Male 67 0 1 Yes Private  
## 2 51676 Female 61 0 0 Yes Self-employed  
## 3 31112 Male 80 0 1 Yes Private  
## 4 60182 Female 49 0 0 Yes Private  
## 5 1665 Female 79 1 0 Yes Self-employed  
## 6 56669 Male 81 0 0 Yes Private  
## residence\_type avg\_glucose\_level bmi smoking\_status stroke  
## 1 Urban 228.69 36.6 formerly smoked 1  
## 2 Rural 202.21 N/A never smoked 1  
## 3 Rural 105.92 32.5 never smoked 1  
## 4 Urban 171.23 34.4 smokes 1  
## 5 Rural 174.12 24 never smoked 1  
## 6 Urban 186.21 29 formerly smoked 1

# deleting id column as it is not going to be needed in the data analysis process   
head(stroke[,-1])

## gender age hypertension heart\_disease ever\_married work\_type  
## 1 Male 67 0 1 Yes Private  
## 2 Female 61 0 0 Yes Self-employed  
## 3 Male 80 0 1 Yes Private  
## 4 Female 49 0 0 Yes Private  
## 5 Female 79 1 0 Yes Self-employed  
## 6 Male 81 0 0 Yes Private  
## residence\_type avg\_glucose\_level bmi smoking\_status stroke  
## 1 Urban 228.69 36.6 formerly smoked 1  
## 2 Rural 202.21 N/A never smoked 1  
## 3 Rural 105.92 32.5 never smoked 1  
## 4 Urban 171.23 34.4 smokes 1  
## 5 Rural 174.12 24 never smoked 1  
## 6 Urban 186.21 29 formerly smoked 1

stroke <- stroke[,-1]  
  
# checking the structure again  
  
str(stroke)

## 'data.frame': 5110 obs. of 11 variables:  
## $ gender : chr "Male" "Female" "Male" "Female" ...  
## $ age : num 67 61 80 49 79 81 74 69 59 78 ...  
## $ hypertension : int 0 0 0 0 1 0 1 0 0 0 ...  
## $ heart\_disease : int 1 0 1 0 0 0 1 0 0 0 ...  
## $ ever\_married : chr "Yes" "Yes" "Yes" "Yes" ...  
## $ work\_type : chr "Private" "Self-employed" "Private" "Private" ...  
## $ residence\_type : chr "Urban" "Rural" "Rural" "Urban" ...  
## $ avg\_glucose\_level: num 229 202 106 171 174 ...  
## $ bmi : chr "36.6" "N/A" "32.5" "34.4" ...  
## $ smoking\_status : chr "formerly smoked" "never smoked" "never smoked" "smokes" ...  
## $ stroke : int 1 1 1 1 1 1 1 1 1 1 ...

summary(stroke)

## gender age hypertension heart\_disease   
## Length:5110 Min. : 0.08 Min. :0.00000 Min. :0.00000   
## Class :character 1st Qu.:25.00 1st Qu.:0.00000 1st Qu.:0.00000   
## Mode :character Median :45.00 Median :0.00000 Median :0.00000   
## Mean :43.23 Mean :0.09746 Mean :0.05401   
## 3rd Qu.:61.00 3rd Qu.:0.00000 3rd Qu.:0.00000   
## Max. :82.00 Max. :1.00000 Max. :1.00000   
## ever\_married work\_type residence\_type avg\_glucose\_level  
## Length:5110 Length:5110 Length:5110 Min. : 55.12   
## Class :character Class :character Class :character 1st Qu.: 77.25   
## Mode :character Mode :character Mode :character Median : 91.89   
## Mean :106.15   
## 3rd Qu.:114.09   
## Max. :271.74   
## bmi smoking\_status stroke   
## Length:5110 Length:5110 Min. :0.00000   
## Class :character Class :character 1st Qu.:0.00000   
## Mode :character Mode :character Median :0.00000   
## Mean :0.04873   
## 3rd Qu.:0.00000   
## Max. :1.00000

# hypertension, heart\_disease, stroke are integer type but can be transformed to factor  
# ever\_married, work\_type, Residence\_type, smoking\_status are character but can be transformed to factor  
# bmi is character but should be changed to number so that we can check for missing data  
  
# converting bmi from character to numeric  
suppressWarnings(stroke$bmi <- as.numeric(as.character(stroke$bmi)))  
  
#checking new structure  
str(stroke)

## 'data.frame': 5110 obs. of 11 variables:  
## $ gender : chr "Male" "Female" "Male" "Female" ...  
## $ age : num 67 61 80 49 79 81 74 69 59 78 ...  
## $ hypertension : int 0 0 0 0 1 0 1 0 0 0 ...  
## $ heart\_disease : int 1 0 1 0 0 0 1 0 0 0 ...  
## $ ever\_married : chr "Yes" "Yes" "Yes" "Yes" ...  
## $ work\_type : chr "Private" "Self-employed" "Private" "Private" ...  
## $ residence\_type : chr "Urban" "Rural" "Rural" "Urban" ...  
## $ avg\_glucose\_level: num 229 202 106 171 174 ...  
## $ bmi : num 36.6 NA 32.5 34.4 24 29 27.4 22.8 NA 24.2 ...  
## $ smoking\_status : chr "formerly smoked" "never smoked" "never smoked" "smokes" ...  
## $ stroke : int 1 1 1 1 1 1 1 1 1 1 ...

#Are there any missing values?  
stroke[!complete.cases(stroke),]

## gender age hypertension heart\_disease ever\_married work\_type  
## 2 Female 61.00 0 0 Yes Self-employed  
## 9 Female 59.00 0 0 Yes Private  
## 14 Male 78.00 0 1 Yes Private  
## 20 Male 57.00 0 1 No Govt\_job  
## 28 Male 58.00 0 0 Yes Private  
## 30 Male 59.00 0 0 Yes Private  
## 44 Female 63.00 0 0 Yes Private  
## 47 Female 75.00 0 1 No Self-employed  
## 51 Female 76.00 0 0 No Private  
## 52 Male 78.00 1 0 Yes Private  
## 55 Female 63.00 0 0 Yes Govt\_job  
## 58 Male 78.00 0 0 Yes Private  
## 65 Male 75.00 0 0 Yes Private  
## 71 Female 76.00 0 0 Yes Govt\_job  
## 79 Female 51.00 0 0 Yes Private  
## 82 Female 66.00 0 0 Yes Self-employed  
## 85 Male 58.00 0 0 Yes Private  
## 106 Male 58.00 0 0 Yes Private  
## 113 Female 76.00 0 0 Yes Self-employed  
## 125 Female 72.00 0 0 Yes Private  
## 127 Male 78.00 1 0 Yes Self-employed  
## 130 Female 75.00 0 0 Yes Govt\_job  
## 134 Female 38.00 0 0 Yes Private  
## 147 Male 65.00 0 0 Yes Self-employed  
## 151 Female 79.00 0 0 Yes Private  
## 161 Female 76.00 0 0 Yes Private  
## 162 Male 71.00 0 1 Yes Private  
## 163 Female 1.32 0 0 No children  
## 168 Male 79.00 1 0 Yes Private  
## 171 Male 64.00 0 0 Yes Self-employed  
## 172 Female 79.00 1 1 No Self-employed  
## 175 Female 78.00 0 0 Yes Self-employed  
## 179 Female 80.00 0 0 Yes Govt\_job  
## 184 Female 77.00 0 0 No Private  
## 190 Male 61.00 0 1 Yes Private  
## 199 Male 79.00 0 0 Yes Private  
## 201 Male 74.00 0 0 Yes Private  
## 219 Female 76.00 1 1 Yes Self-employed  
## 228 Male 74.00 0 0 Yes Self-employed  
## 248 Male 71.00 1 0 Yes Self-employed  
## 343 Male 34.00 0 1 Yes Private  
## 361 Female 76.00 1 0 Yes Self-employed  
## 433 Female 63.00 0 0 Yes Govt\_job  
## 478 Male 61.00 0 0 Yes Govt\_job  
## 480 Male 54.00 1 0 Yes Private  
## 523 Male 40.00 0 0 Yes Private  
## 669 Female 48.00 1 0 No Private  
## 672 Male 61.00 0 1 Yes Private  
## 681 Male 31.00 1 0 Yes Govt\_job  
## 730 Female 43.00 0 0 Yes Govt\_job  
## 743 Female 9.00 0 0 No children  
## 866 Male 52.00 0 0 Yes Private  
## 868 Female 77.00 0 1 Yes Private  
## 873 Female 17.00 0 0 No Private  
## 880 Female 71.00 0 0 Yes Self-employed  
## 904 Female 35.00 0 0 No Govt\_job  
## 937 Female 23.00 0 0 No Private  
## 966 Male 40.00 0 0 No Private  
## 1103 Female 23.00 0 0 No Private  
## 1107 Female 71.00 1 0 Yes Self-employed  
## 1116 Male 13.00 0 0 No children  
## 1184 Male 73.00 1 0 Yes Self-employed  
## 1195 Female 3.00 0 0 No children  
## 1215 Male 51.00 0 0 Yes Private  
## 1236 Male 35.00 0 0 Yes Private  
## 1242 Female 73.00 0 0 Yes Self-employed  
## 1278 Female 6.00 0 0 No children  
## 1294 Male 46.00 1 0 Yes Private  
## 1301 Female 71.00 0 0 Yes Private  
## 1307 Female 54.00 1 0 Yes Private  
## 1325 Female 80.00 0 0 Yes Govt\_job  
## 1343 Female 49.00 0 0 Yes Private  
## 1353 Male 72.00 0 0 Yes Self-employed  
## 1428 Male 25.00 0 0 Yes Private  
## 1458 Male 27.00 0 0 No Private  
## 1467 Male 51.00 1 0 Yes Private  
## 1469 Male 48.00 0 0 Yes Self-employed  
## 1472 Male 7.00 0 0 No children  
## 1504 Female 61.00 1 1 Yes Private  
## 1529 Female 25.00 0 0 Yes Govt\_job  
## 1540 Male 30.00 0 0 Yes Private  
## 1547 Male 71.00 1 0 Yes Self-employed  
## 1597 Male 47.00 0 0 No Private  
## 1641 Male 76.00 0 1 Yes Private  
## 1645 Female 29.00 0 0 No Private  
## 1647 Female 48.00 0 1 Yes Self-employed  
## 1651 Female 57.00 1 0 Yes Private  
## 1670 Male 58.00 0 0 Yes Private  
## 1671 Male 45.00 0 1 Yes Private  
## 1682 Male 66.00 0 0 Yes Private  
## 1719 Male 59.00 0 1 Yes Govt\_job  
## 1720 Male 34.00 0 0 Yes Private  
## 1731 Male 69.00 1 0 Yes Private  
## 1754 Male 66.00 0 0 Yes Self-employed  
## 1757 Female 48.00 0 0 Yes Self-employed  
## 1780 Male 32.00 0 0 No Private  
## 1817 Male 60.00 0 0 Yes Self-employed  
## 1837 Female 30.00 0 0 No Govt\_job  
## 1838 Female 60.00 0 0 Yes Self-employed  
## 1867 Male 10.00 0 0 No children  
## 1895 Male 20.00 0 0 No Private  
## 1907 Male 77.00 0 0 Yes Private  
## 1913 Male 67.00 0 0 Yes Private  
## 1982 Female 42.00 0 0 Yes Private  
## 1994 Female 60.00 1 0 Yes Private  
## 2031 Male 0.48 0 0 No children  
## 2104 Male 35.00 0 0 Yes Private  
## 2106 Male 50.00 1 0 No Private  
## 2110 Female 19.00 0 0 No Private  
## 2193 Female 77.00 1 0 Yes Self-employed  
## 2216 Male 67.00 0 1 Yes Private  
## 2264 Female 20.00 0 0 No Private  
## 2286 Male 49.00 0 0 Yes Private  
## 2322 Male 77.00 0 1 Yes Govt\_job  
## 2323 Female 52.00 1 0 Yes Self-employed  
## 2340 Male 43.00 0 0 Yes Govt\_job  
## 2344 Female 69.00 1 0 Yes Govt\_job  
## 2478 Female 34.00 1 0 Yes Self-employed  
## 2495 Male 78.00 0 1 Yes Self-employed  
## 2503 Male 76.00 0 1 Yes Self-employed  
## 2516 Male 62.00 1 1 Yes Private  
## 2530 Female 71.00 1 0 Yes Private  
## 2533 Male 79.00 0 1 Yes Private  
## 2542 Male 79.00 0 1 Yes Private  
## 2583 Male 54.00 0 0 Yes Private  
## 2698 Female 73.00 1 0 Yes Private  
## 2740 Female 5.00 0 0 No children  
## 2753 Female 38.00 0 0 Yes Private  
## 2769 Male 72.00 1 0 Yes Private  
## 2789 Male 14.00 0 0 No Private  
## 2817 Male 50.00 0 0 Yes Private  
## 2829 Male 29.00 1 0 Yes Private  
## 2856 Male 75.00 1 0 Yes Private  
## 2868 Female 68.00 1 1 Yes Private  
## 2880 Female 33.00 1 0 No Private  
## 2898 Male 63.00 0 1 Yes Self-employed  
## 2915 Female 56.00 0 0 Yes Private  
## 2961 Male 70.00 0 0 Yes Govt\_job  
## 2998 Male 71.00 0 1 Yes Private  
## 3008 Female 73.00 0 0 No Self-employed  
## 3029 Female 67.00 1 0 Yes Private  
## 3049 Female 62.00 1 0 Yes Self-employed  
## 3060 Female 38.00 0 0 Yes Private  
## 3075 Female 47.00 0 0 Yes Self-employed  
## 3105 Female 42.00 0 0 Yes Private  
## 3112 Male 58.00 0 0 Yes Govt\_job  
## 3136 Male 44.00 1 0 Yes Private  
## 3162 Male 42.00 0 0 Yes Private  
## 3163 Male 78.00 1 0 Yes Self-employed  
## 3165 Female 68.00 0 0 No Private  
## 3177 Male 39.00 0 0 Yes Private  
## 3198 Male 60.00 0 0 Yes Self-employed  
## 3215 Female 31.00 0 0 Yes Self-employed  
## 3216 Male 67.00 0 0 Yes Private  
## 3217 Female 52.00 1 0 Yes Self-employed  
## 3376 Female 53.00 0 0 Yes Private  
## 3383 Female 33.00 0 0 No Private  
## 3426 Female 53.00 0 0 No Private  
## 3432 Female 49.00 0 0 Yes Private  
## 3504 Male 52.00 0 1 No Self-employed  
## 3563 Female 41.00 0 0 Yes Self-employed  
## 3606 Male 1.88 0 0 No children  
## 3630 Male 34.00 0 0 Yes Private  
## 3682 Female 16.00 0 0 No Private  
## 3700 Female 45.00 0 0 Yes Private  
## 3706 Male 1.08 0 0 No children  
## 3727 Male 1.80 0 0 No children  
## 3735 Female 13.00 0 0 No children  
## 3803 Female 61.00 0 0 Yes Private  
## 3809 Female 37.00 0 0 No Govt\_job  
## 3873 Male 32.00 1 0 No Govt\_job  
## 3914 Female 79.00 0 0 Yes Private  
## 3941 Male 8.00 0 0 No children  
## 3946 Female 75.00 0 1 Yes Self-employed  
## 3952 Female 79.00 1 0 Yes Self-employed  
## 4047 Female 69.00 0 1 Yes Private  
## 4070 Male 31.00 0 0 Yes Private  
## 4165 Female 82.00 1 0 Yes Private  
## 4203 Male 32.00 1 0 No Private  
## 4231 Female 17.00 0 0 No Private  
## 4256 Female 18.00 0 0 No Private  
## 4284 Male 59.00 1 0 Yes Govt\_job  
## 4287 Male 3.00 0 0 No children  
## 4423 Female 20.00 0 0 No Govt\_job  
## 4452 Female 78.00 0 0 Yes Govt\_job  
## 4523 Male 52.00 1 0 Yes Govt\_job  
## 4562 Female 65.00 0 1 Yes Private  
## 4617 Male 59.00 0 0 Yes Private  
## 4685 Female 78.00 1 1 Yes Private  
## 4714 Female 70.00 0 1 Yes Self-employed  
## 4751 Female 70.00 0 1 Yes Self-employed  
## 4791 Male 37.00 0 0 Yes Private  
## 4922 Male 72.00 0 1 Yes Private  
## 4935 Male 1.32 0 0 No children  
## 4950 Male 58.00 0 0 Yes Govt\_job  
## 4985 Male 31.00 0 0 No Private  
## 5040 Male 41.00 0 0 No Private  
## 5049 Male 40.00 0 0 Yes Private  
## 5094 Female 45.00 1 0 Yes Govt\_job  
## 5100 Male 40.00 0 0 Yes Private  
## 5106 Female 80.00 1 0 Yes Private  
## residence\_type avg\_glucose\_level bmi smoking\_status stroke  
## 2 Rural 202.21 NA never smoked 1  
## 9 Rural 76.15 NA Unknown 1  
## 14 Urban 219.84 NA Unknown 1  
## 20 Urban 217.08 NA Unknown 1  
## 28 Rural 189.84 NA Unknown 1  
## 30 Rural 211.78 NA formerly smoked 1  
## 44 Urban 90.90 NA formerly smoked 1  
## 47 Urban 109.78 NA Unknown 1  
## 51 Urban 89.96 NA Unknown 1  
## 52 Urban 75.32 NA formerly smoked 1  
## 55 Urban 197.54 NA never smoked 1  
## 58 Urban 237.75 NA formerly smoked 1  
## 65 Urban 104.72 NA Unknown 1  
## 71 Rural 62.57 NA formerly smoked 1  
## 79 Urban 165.31 NA never smoked 1  
## 82 Urban 101.45 NA Unknown 1  
## 85 Urban 71.20 NA Unknown 1  
## 106 Urban 82.30 NA smokes 1  
## 113 Urban 106.41 NA formerly smoked 1  
## 125 Urban 219.91 NA Unknown 1  
## 127 Urban 93.13 NA formerly smoked 1  
## 130 Urban 62.48 NA Unknown 1  
## 134 Rural 101.45 NA formerly smoked 1  
## 147 Urban 68.43 NA formerly smoked 1  
## 151 Rural 169.67 NA Unknown 1  
## 161 Urban 57.92 NA formerly smoked 1  
## 162 Urban 81.76 NA smokes 1  
## 163 Urban 70.37 NA Unknown 1  
## 168 Rural 75.02 NA never smoked 1  
## 171 Rural 111.98 NA formerly smoked 1  
## 172 Rural 60.94 NA never smoked 1  
## 175 Rural 60.67 NA formerly smoked 1  
## 179 Urban 110.66 NA Unknown 1  
## 184 Urban 81.32 NA Unknown 1  
## 190 Urban 209.86 NA Unknown 1  
## 199 Rural 114.77 NA formerly smoked 1  
## 201 Urban 167.13 NA Unknown 1  
## 219 Urban 199.86 NA smokes 1  
## 228 Rural 60.98 NA never smoked 1  
## 248 Rural 87.80 NA Unknown 1  
## 343 Urban 106.23 NA formerly smoked 0  
## 361 Urban 209.58 NA never smoked 0  
## 433 Rural 79.92 NA smokes 0  
## 478 Urban 184.15 NA Unknown 0  
## 480 Rural 198.69 NA smokes 0  
## 523 Rural 89.77 NA smokes 0  
## 669 Rural 118.14 NA formerly smoked 0  
## 672 Urban 88.27 NA never smoked 0  
## 681 Urban 92.11 NA never smoked 0  
## 730 Rural 107.42 NA never smoked 0  
## 743 Urban 95.81 NA Unknown 0  
## 866 Urban 226.70 NA smokes 0  
## 868 Rural 183.10 NA never smoked 0  
## 873 Rural 83.23 NA never smoked 0  
## 880 Urban 91.35 NA formerly smoked 0  
## 904 Urban 83.76 NA smokes 0  
## 937 Urban 110.16 NA never smoked 0  
## 966 Urban 88.27 NA formerly smoked 0  
## 1103 Rural 193.22 NA smokes 0  
## 1107 Rural 66.12 NA never smoked 0  
## 1116 Urban 71.73 NA Unknown 0  
## 1184 Rural 102.06 NA Unknown 0  
## 1195 Urban 79.63 NA Unknown 0  
## 1215 Rural 217.71 NA formerly smoked 0  
## 1236 Rural 115.92 NA formerly smoked 0  
## 1242 Rural 79.69 NA formerly smoked 0  
## 1278 Urban 201.25 NA Unknown 0  
## 1294 Rural 73.72 NA smokes 0  
## 1301 Urban 214.77 NA Unknown 0  
## 1307 Rural 98.74 NA never smoked 0  
## 1325 Urban 84.86 NA Unknown 0  
## 1343 Rural 67.27 NA formerly smoked 0  
## 1353 Rural 72.09 NA smokes 0  
## 1428 Rural 78.29 NA smokes 0  
## 1458 Rural 191.79 NA smokes 0  
## 1467 Rural 163.56 NA formerly smoked 0  
## 1469 Rural 216.88 NA smokes 0  
## 1472 Urban 87.94 NA Unknown 0  
## 1504 Urban 237.58 NA formerly smoked 0  
## 1529 Urban 93.23 NA smokes 0  
## 1540 Urban 91.23 NA smokes 0  
## 1547 Rural 93.60 NA never smoked 0  
## 1597 Rural 237.17 NA Unknown 0  
## 1641 Urban 79.05 NA Unknown 0  
## 1645 Urban 81.43 NA formerly smoked 0  
## 1647 Urban 101.22 NA formerly smoked 0  
## 1651 Urban 210.00 NA never smoked 0  
## 1670 Urban 94.00 NA Unknown 0  
## 1671 Rural 93.77 NA Unknown 0  
## 1682 Urban 190.40 NA formerly smoked 0  
## 1719 Urban 188.69 NA formerly smoked 0  
## 1720 Rural 86.51 NA formerly smoked 0  
## 1731 Rural 107.11 NA smokes 0  
## 1754 Urban 71.38 NA formerly smoked 0  
## 1757 Rural 209.90 NA smokes 0  
## 1780 Rural 95.58 NA smokes 0  
## 1817 Urban 185.71 NA Unknown 0  
## 1837 Urban 88.20 NA smokes 0  
## 1838 Urban 203.04 NA smokes 0  
## 1867 Rural 99.87 NA formerly smoked 0  
## 1895 Rural 70.96 NA Unknown 0  
## 1907 Urban 74.26 NA formerly smoked 0  
## 1913 Urban 92.73 NA never smoked 0  
## 1982 Urban 208.06 NA smokes 0  
## 1994 Urban 109.00 NA Unknown 0  
## 2031 Urban 73.02 NA Unknown 0  
## 2104 Rural 77.48 NA formerly smoked 0  
## 2106 Urban 81.96 NA formerly smoked 0  
## 2110 Rural 72.39 NA smokes 0  
## 2193 Urban 109.51 NA never smoked 0  
## 2216 Rural 97.24 NA Unknown 0  
## 2264 Urban 89.03 NA smokes 0  
## 2286 Rural 79.64 NA smokes 0  
## 2322 Rural 106.03 NA Unknown 0  
## 2323 Rural 111.38 NA smokes 0  
## 2340 Rural 80.07 NA never smoked 0  
## 2344 Urban 112.20 NA never smoked 0  
## 2478 Urban 100.61 NA Unknown 0  
## 2495 Urban 243.73 NA smokes 0  
## 2503 Urban 67.03 NA never smoked 0  
## 2516 Rural 176.25 NA never smoked 0  
## 2530 Urban 105.55 NA smokes 0  
## 2533 Urban 213.38 NA Unknown 0  
## 2542 Rural 82.27 NA never smoked 0  
## 2583 Rural 74.06 NA never smoked 0  
## 2698 Rural 217.84 NA never smoked 0  
## 2740 Rural 105.18 NA Unknown 0  
## 2753 Rural 217.55 NA smokes 0  
## 2769 Rural 231.71 NA Unknown 0  
## 2789 Rural 110.72 NA never smoked 0  
## 2817 Urban 67.02 NA formerly smoked 0  
## 2829 Urban 77.55 NA formerly smoked 0  
## 2856 Rural 198.79 NA smokes 0  
## 2868 Rural 233.30 NA Unknown 0  
## 2880 Rural 97.87 NA smokes 0  
## 2898 Urban 82.72 NA never smoked 0  
## 2915 Urban 102.97 NA smokes 0  
## 2961 Urban 202.55 NA formerly smoked 0  
## 2998 Urban 204.98 NA formerly smoked 0  
## 3008 Rural 69.35 NA never smoked 0  
## 3029 Rural 85.48 NA smokes 0  
## 3049 Urban 75.78 NA smokes 0  
## 3060 Urban 91.44 NA Unknown 0  
## 3075 Rural 195.61 NA never smoked 0  
## 3105 Urban 73.37 NA smokes 0  
## 3112 Urban 160.87 NA formerly smoked 0  
## 3136 Rural 84.10 NA Unknown 0  
## 3162 Urban 177.91 NA Unknown 0  
## 3163 Urban 90.19 NA Unknown 0  
## 3165 Urban 82.85 NA smokes 0  
## 3177 Rural 84.18 NA smokes 0  
## 3198 Rural 212.02 NA Unknown 0  
## 3215 Urban 87.23 NA formerly smoked 0  
## 3216 Urban 260.85 NA Unknown 0  
## 3217 Rural 104.45 NA never smoked 0  
## 3376 Urban 227.68 NA never smoked 0  
## 3383 Urban 84.40 NA smokes 0  
## 3426 Rural 235.45 NA formerly smoked 0  
## 3432 Rural 107.55 NA Unknown 0  
## 3504 Rural 79.81 NA formerly smoked 0  
## 3563 Rural 76.66 NA Unknown 0  
## 3606 Rural 143.97 NA Unknown 0  
## 3630 Urban 99.23 NA smokes 0  
## 3682 Urban 89.45 NA Unknown 0  
## 3700 Urban 202.66 NA never smoked 0  
## 3706 Rural 74.50 NA Unknown 0  
## 3727 Urban 68.80 NA Unknown 0  
## 3735 Rural 219.81 NA Unknown 0  
## 3803 Rural 219.38 NA never smoked 0  
## 3809 Rural 72.08 NA formerly smoked 0  
## 3873 Urban 58.24 NA formerly smoked 0  
## 3914 Rural 208.05 NA smokes 0  
## 3941 Urban 78.76 NA Unknown 0  
## 3946 Urban 83.88 NA smokes 0  
## 3952 Rural 92.43 NA never smoked 0  
## 4047 Urban 207.60 NA never smoked 0  
## 4070 Urban 108.62 NA smokes 0  
## 4165 Urban 222.52 NA formerly smoked 0  
## 4203 Rural 74.43 NA Unknown 0  
## 4231 Urban 92.97 NA formerly smoked 0  
## 4256 Rural 101.12 NA smokes 0  
## 4284 Rural 253.93 NA formerly smoked 0  
## 4287 Rural 194.75 NA Unknown 0  
## 4423 Rural 79.53 NA never smoked 0  
## 4452 Urban 101.76 NA smokes 0  
## 4523 Rural 116.62 NA smokes 0  
## 4562 Rural 57.52 NA formerly smoked 0  
## 4617 Urban 223.16 NA Unknown 0  
## 4685 Rural 206.53 NA never smoked 0  
## 4714 Rural 65.68 NA Unknown 0  
## 4751 Urban 240.69 NA smokes 0  
## 4791 Rural 107.06 NA smokes 0  
## 4922 Rural 238.27 NA smokes 0  
## 4935 Rural 107.02 NA Unknown 0  
## 4950 Urban 84.94 NA never smoked 0  
## 4985 Urban 215.07 NA smokes 0  
## 5040 Rural 70.15 NA formerly smoked 0  
## 5049 Urban 191.15 NA smokes 0  
## 5094 Rural 95.02 NA smokes 0  
## 5100 Rural 83.94 NA smokes 0  
## 5106 Urban 83.75 NA never smoked 0

# bmi  
  
summary(stroke$bmi)

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

median(stroke$bmi)

## [1] NA

median(stroke$bmi,na.rm = TRUE)

## [1] 28.1

# use median to fill missing values in bmi   
stroke$bmi[is.na(stroke$bmi)] <- median(stroke$bmi,na.rm = TRUE)  
  
#check again for missing values  
stroke[!complete.cases(stroke),] #none

## [1] gender age hypertension heart\_disease   
## [5] ever\_married work\_type residence\_type avg\_glucose\_level  
## [9] bmi smoking\_status stroke   
## <0 rows> (or 0-length row.names)

## Check and transform variables to factors ##  
  
  
# check if there are no errors  
# Gender  
summary(factor(stroke$gender))

## Female Male Other   
## 2994 2115 1

# since there is only one row containing "other" we treat it as an outlier and remove it   
stroke = stroke[!stroke$gender == 'Other',]  
# transform to factor  
stroke$gender <- as.factor(stroke$gender)  
  
  
# Work\_type  
summary(factor(stroke$work\_type))

## children Govt\_job Never\_worked Private Self-employed   
## 687 657 22 2924 819

stroke$work\_type <- factor(stroke$work\_type, levels = c("children", "Govt\_job", "Never\_worked", "Private", "Self-employed"))  
  
  
# Smoking\_status  
summary(factor(stroke$smoking\_status))

## formerly smoked never smoked smokes Unknown   
## 884 1892 789 1544

stroke$smoking\_status <- as.factor(stroke$smoking\_status)  
  
  
# Residence\_type  
summary(factor(stroke$residence\_type))

## Rural Urban   
## 2513 2596

stroke$residence\_type <- as.factor(stroke$residence\_type)  
  
# Hypertension  
summary(factor(stroke$hypertension))

## 0 1   
## 4611 498

stroke$hypertension <- factor(stroke$hypertension, levels = c("0", "1"),  
 labels = c("No", "Yes"))  
  
# Heart\_disease  
summary(factor(stroke$heart\_disease))

## 0 1   
## 4833 276

stroke$heart\_disease <- factor(stroke$heart\_disease, levels = c("0", "1"),  
 labels = c("No", "Yes"))  
  
# Stroke   
summary(factor(stroke$stroke))

## 0 1   
## 4860 249

stroke$stroke <- factor(stroke$stroke, levels = c("0", "1"),  
 labels = c("No", "Yes"))  
  
# Ever\_married  
summary(factor(stroke$ever\_married))

## No Yes   
## 1756 3353

stroke$ever\_married <- as.factor(stroke$ever\_married)  
  
str(stroke)

## 'data.frame': 5109 obs. of 11 variables:  
## $ gender : Factor w/ 2 levels "Female","Male": 2 1 2 1 1 2 2 1 1 1 ...  
## $ age : num 67 61 80 49 79 81 74 69 59 78 ...  
## $ hypertension : Factor w/ 2 levels "No","Yes": 1 1 1 1 2 1 2 1 1 1 ...  
## $ heart\_disease : Factor w/ 2 levels "No","Yes": 2 1 2 1 1 1 2 1 1 1 ...  
## $ ever\_married : Factor w/ 2 levels "No","Yes": 2 2 2 2 2 2 2 1 2 2 ...  
## $ work\_type : Factor w/ 5 levels "children","Govt\_job",..: 4 5 4 4 5 4 4 4 4 4 ...  
## $ residence\_type : Factor w/ 2 levels "Rural","Urban": 2 1 1 2 1 2 1 2 1 2 ...  
## $ avg\_glucose\_level: num 229 202 106 171 174 ...  
## $ bmi : num 36.6 28.1 32.5 34.4 24 29 27.4 22.8 28.1 24.2 ...  
## $ smoking\_status : Factor w/ 4 levels "formerly smoked",..: 1 2 2 3 2 1 2 2 4 4 ...  
## $ stroke : Factor w/ 2 levels "No","Yes": 2 2 2 2 2 2 2 2 2 2 ...

#### Data Analysis ####  
  
summary(stroke) # statistics

## gender age hypertension heart\_disease ever\_married  
## Female:2994 Min. : 0.08 No :4611 No :4833 No :1756   
## Male :2115 1st Qu.:25.00 Yes: 498 Yes: 276 Yes:3353   
## Median :45.00   
## Mean :43.23   
## 3rd Qu.:61.00   
## Max. :82.00   
## work\_type residence\_type avg\_glucose\_level bmi   
## children : 687 Rural:2513 Min. : 55.12 Min. :10.30   
## Govt\_job : 657 Urban:2596 1st Qu.: 77.24 1st Qu.:23.80   
## Never\_worked : 22 Median : 91.88 Median :28.10   
## Private :2924 Mean :106.14 Mean :28.86   
## Self-employed: 819 3rd Qu.:114.09 3rd Qu.:32.80   
## Max. :271.74 Max. :97.60   
## smoking\_status stroke   
## formerly smoked: 884 No :4860   
## never smoked :1892 Yes: 249   
## smokes : 789   
## Unknown :1544   
##   
##

# Age statistics  
summary(stroke$age)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.08 25.00 45.00 43.23 61.00 82.00

# Age analysis   
# Based on the age column statistics we can observe that there is a wide range of values, Min 0.08/ Max 82 indicating that the   
# population sample was prepared correctly assuring the data wasn't biased. Moreover we can see that both Median and Mean are similar,  
# the median (45.00) and mean (43.23) which suggests that the distribution of ages in the data-set is not heavily skewed   
# and the data is symmetrically distributed. This similarity between the median and mean ages is a positive aspect.   
# It implies that the central tendency of the age distribution is robust and not heavily influenced by extremely young or old ages.   
# The one thing that can be concerning is Min value 0.08 and should be investigated, as it may be an error or a special case that needs attention.  
# It can however be treated as an outlier during the analysis process.  
  
# Average Glucose level statistics  
summary(stroke$avg\_glucose\_level)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 55.12 77.24 91.88 106.14 114.09 271.74

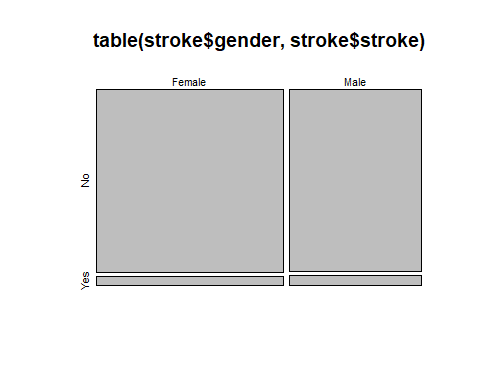
# The he glucose levels in this data-set range from a minimum of 55.12 to a maximum of 271.74. The majority of levels (50%) fall between  
# 77.24 and 114.09. The mean glucose level is higher than the median, indicating a potential right skewness influenced by higher glucose values.  
# The right skewness suggests that there are some individuals with elevated glucose levels, potentially outliers, that contribute to the higher average.  
# Understanding the distribution of glucose levels is crucial, especially in the context of health data where extreme values may have clinical   
# implications. Further exploration and consideration of clinical significance are warranted, particularly for the higher glucose values.  
  
# Bmi statistics   
summary(stroke$bmi)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 10.30 23.80 28.10 28.86 32.80 97.60

# The BMI distribution in this data-set ranges from 10.30 to 97.60. The median and mean are relatively close, suggesting a moderately   
# symmetric distribution, but the presence of a higher maximum BMI (97.60) indicates the potential presence of outliers.   
# Extremely low or high BMI values may have clinical significance. Very low BMI values (e.g., 10.30) might be unusual and could indicate  
# potential data quality issues or represent individuals with severe underweight conditions. Similarly, very high BMI values (e.g., 97.60)  
# may warrant closer examination due to their potential impact on the overall analysis and their relevance in the context of stroke prediction.  
  
# Analyzing if there is gender and stroke correlation  
table(stroke$gender, stroke$stroke)

##   
## No Yes  
## Female 2853 141  
## Male 2007 108

mosaicplot(table(stroke$gender, stroke$stroke))



# The numbers suggest a correlation between gender and stroke, according to the mosaicplot we can see more men had strokes than women   
# but it's important to note that correlation does not imply causation. There appears to be some association between gender and the   
# occurrence of stroke. However, additional statistical analysis and consideration of other factors would be necessary to draw more   
# definitive conclusions about the nature and strength of this association.  
  
# Other factors/stroke correlations   
  
table(stroke$hypertension, stroke$stroke)

##   
## No Yes  
## No 4428 183  
## Yes 432 66

table(stroke$heart\_disease, stroke$stroke)

##   
## No Yes  
## No 4631 202  
## Yes 229 47

table(stroke$ever\_married, stroke$stroke)

##   
## No Yes  
## No 1727 29  
## Yes 3133 220

table(stroke$work\_type, stroke$stroke)

##   
## No Yes  
## children 685 2  
## Govt\_job 624 33  
## Never\_worked 22 0  
## Private 2775 149  
## Self-employed 754 65

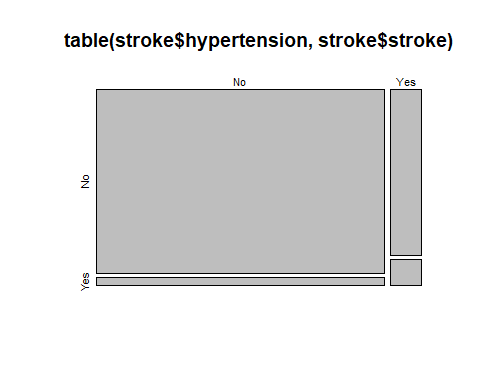
table(stroke$residence\_type, stroke$stroke)

##   
## No Yes  
## Rural 2399 114  
## Urban 2461 135

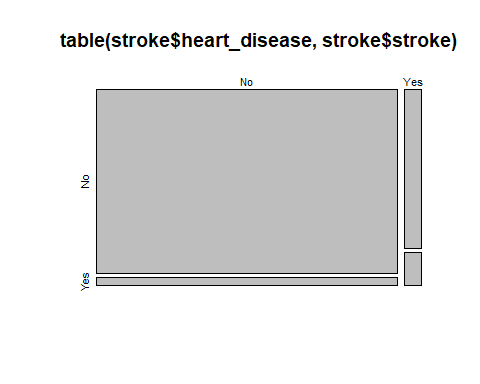
table(stroke$smoking\_status, stroke$stroke)

##   
## No Yes  
## formerly smoked 814 70  
## never smoked 1802 90  
## smokes 747 42  
## Unknown 1497 47

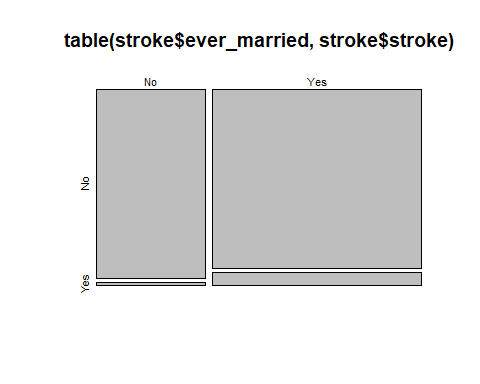
#### Data Visualization ####  
  
  
### mosaic plots to help visualize the correlations between factor type values ###  
  
mosaicplot(table(stroke$hypertension, stroke$stroke))



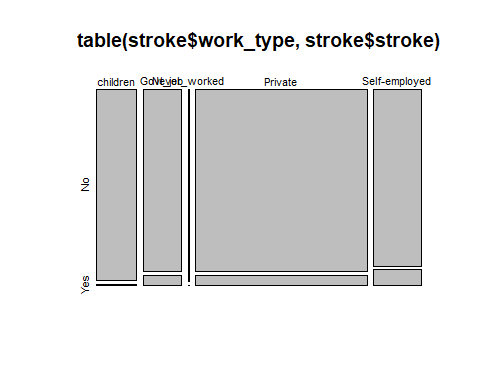
# According to the plot, people with hypertension had more cases of stroke than those who have no hypertension.  
  
mosaicplot(table(stroke$heart\_disease, stroke$stroke))



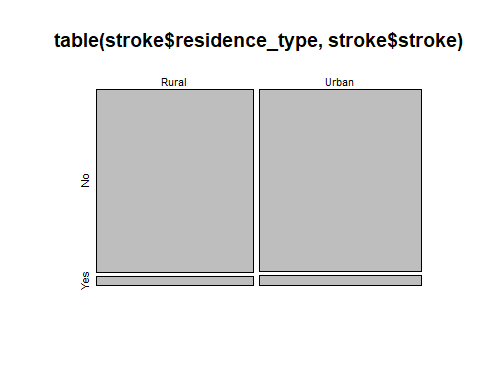
# Same as with hypertension, there are more cases of stroke in the group with heart diseases present.  
  
mosaicplot(table(stroke$ever\_married, stroke$stroke))



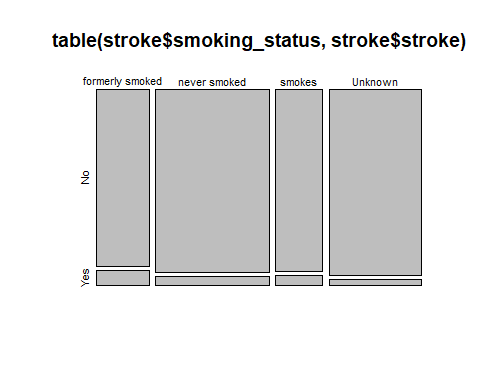
# Married people suffered more cases of stroke.  
  
mosaicplot(table(stroke$work\_type, stroke$stroke))



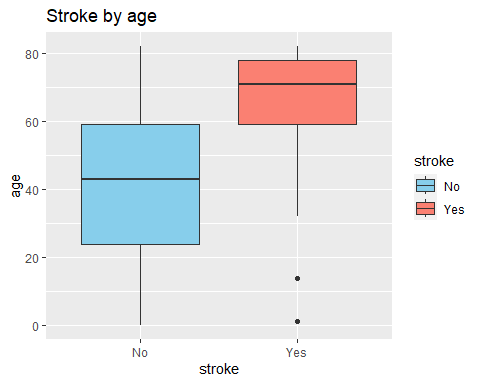
# Stroke seems to have occurred mostly in the self-employed, private, government job groups. In comparison  
# children and never worked groups show significantly less cases of stroke. This might suggest there is an important  
# correlation between actively working and stroke.  
  
mosaicplot(table(stroke$residence\_type, stroke$stroke))



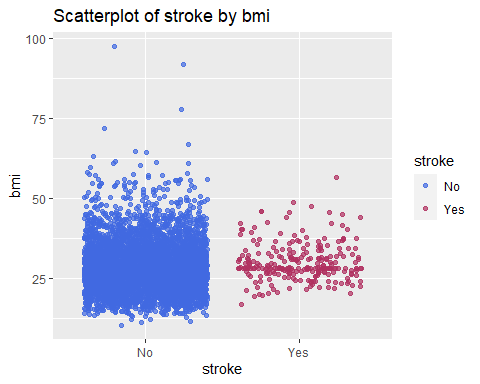
# There doesn't seem to be any difference in stroke cases depending on type of residency.  
  
mosaicplot(table(stroke$smoking\_status, stroke$stroke))



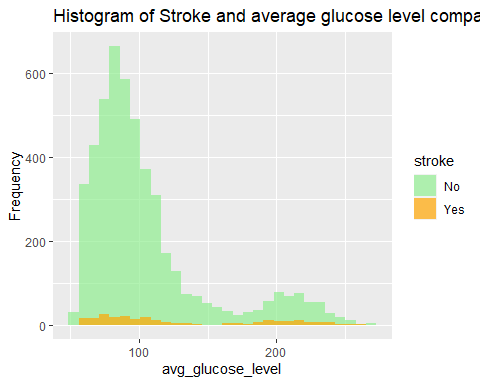
# According to the plot, people who formerly smoked had more recorded cases of stroke compared to other groups.  
  
### Data Visualization of numerical values ###  
  
## Stroke by age (Boxplot) ##  
  
ggplot(data = stroke, aes(x = stroke, y = age, fill = stroke)) +  
 geom\_boxplot() +  
 labs(title = 'Stroke by age') +  
 scale\_fill\_manual(values = c("skyblue", "salmon"))



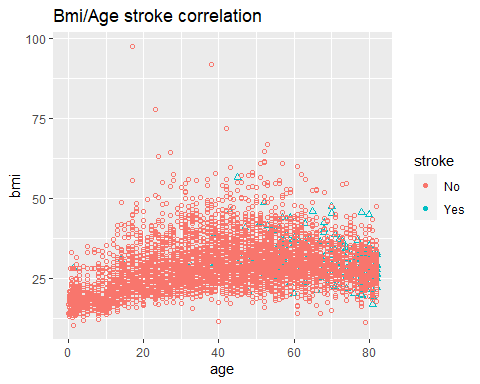
# As previously stated in age analysis this boxplot shows that people in the 60/80 years old group, with few   
# exceptions (outliers), are the ones that suffered stroke.  
  
## Stroke by bmi (Scatterplot) ##  
  
ggplot(data = stroke, aes(x = stroke, y = bmi, color = stroke)) +  
 geom\_point(position = "jitter", alpha = 0.7) +   
 labs(title = 'Scatterplot of stroke by bmi', x = 'stroke', y = 'bmi') +  
 scale\_color\_manual(values = c("royalblue", "maroon"))



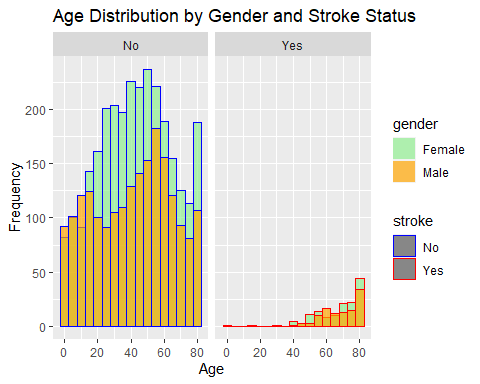
# Based on the chart we can assume that bmi alone doesn't seem to have any correlation to suffering stroke.  
  
## Stroke/ avg\_glucose\_level correlation (Histogram) ##  
  
ggplot(data = stroke, aes(x = avg\_glucose\_level , fill = stroke)) +  
 geom\_histogram(position = "identity", alpha = 0.7, bins = 30) +  
 labs(title = 'Histogram of Stroke and average glucose level comparison', x = 'avg\_glucose\_level', y = 'Frequency') +  
 scale\_fill\_manual(values = c("lightgreen", "orange"))



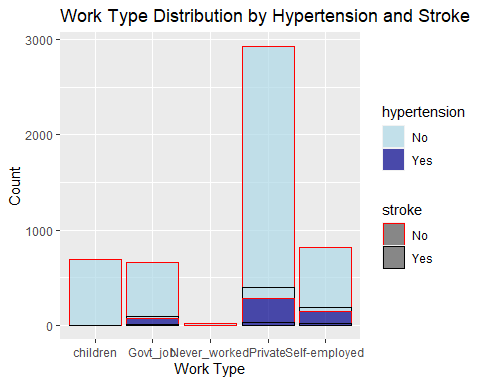
# Just like with bmi, according to the data presented we can see that average glucose level doesn't correlate with stroke.   
  
## Bmi/Age stroke cases ##  
  
ggplot(data = stroke, aes(x= age, y = bmi, color = stroke))+  
 geom\_point(shape = stroke$stroke)+  
 labs(title = 'Bmi/Age stroke correlation')



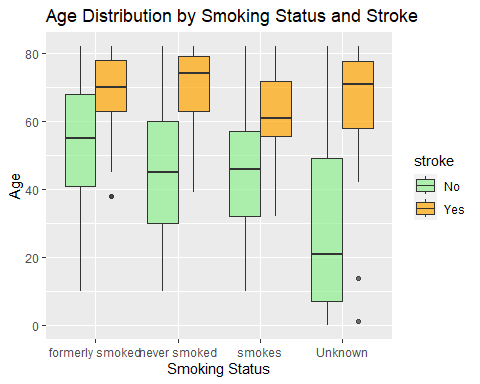
# As we can see, stroke patients are more likely to appear among people above 50 years old. However as mentioned before   
# bmi doesn't seem to correlate with age when it comes to suffering stroke.  
  
### Data Visualization of multiple factors ###  
  
## Age/Gender/Stroke correlation ##  
  
ggplot(stroke, aes(x = age, fill = gender, color = stroke)) +  
 geom\_histogram(binwidth = 5, position = "identity", alpha = 0.7) +  
 facet\_grid(. ~ stroke) +  
 labs(title = 'Age Distribution by Gender and Stroke Status', x = 'Age', y = 'Frequency') +  
 scale\_fill\_manual(values = c("lightgreen", "orange")) +  
 scale\_color\_manual(values = c("blue", "red"))



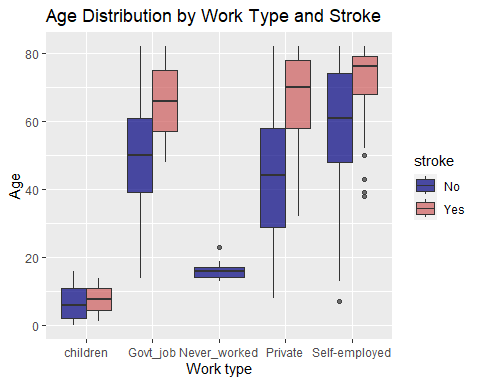
# Based on the chart we can see that men are more prone to suffering stroke rather than women, especially those who are 50+ years old.  
  
## Work Type/Hypertension/Stroke ##  
  
ggplot(stroke, aes(x = work\_type, fill = hypertension, color = stroke)) +  
 geom\_bar(position = "stack", alpha = 0.7) +  
 labs(title = 'Work Type Distribution by Hypertension and Stroke', x = 'Work Type', y = 'Count') +  
 scale\_fill\_manual(values = c("lightblue", "darkblue")) +  
 scale\_color\_manual(values = c("red", "black"))



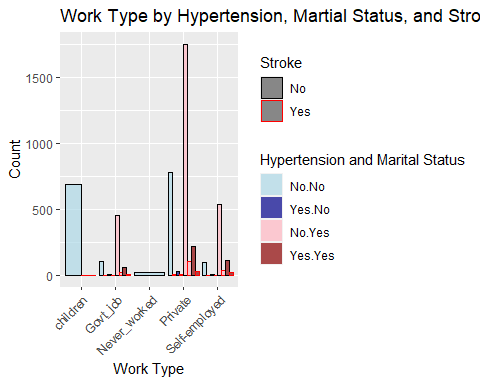
# According to the chart work type combined with hypertension seems to have big impact on probability of having stroke.  
# Those working government jobs, self-employed and private and additionally suffering from hypertension have more cases of stroke.  
  
  
## Age/Smoking/Stroke ##  
  
ggplot(stroke, aes(x = smoking\_status, y = age, fill = stroke)) +  
 geom\_boxplot(position = "dodge", alpha = 0.7) +  
 labs(title = 'Age Distribution by Smoking Status and Stroke', x = 'Smoking Status', y = 'Age') +  
 scale\_fill\_manual(values = c("lightgreen", "orange"))



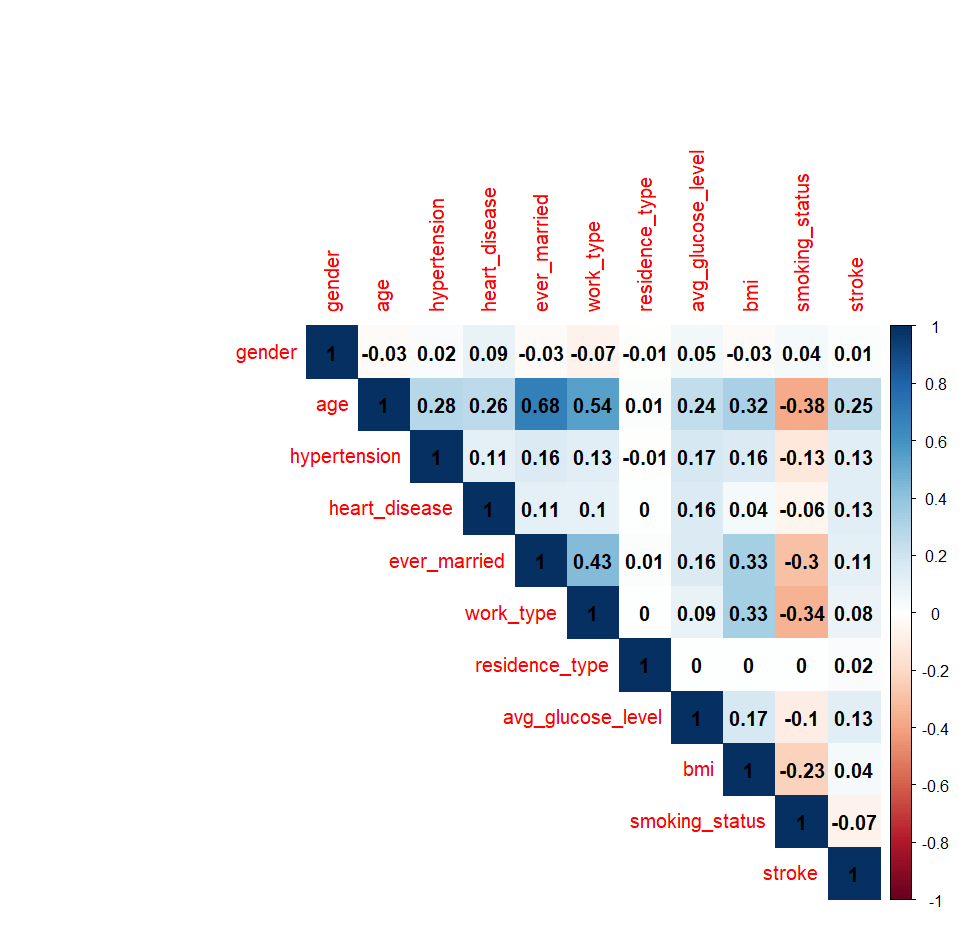
# Based on this chart smoking doesn't seem to be a factor in suffering stroke, however we can notice again that the   
# age seems to be one of the main factors that have to be examined more closely.  
  
## Age/Work Type/Stroke ##  
  
ggplot(stroke, aes(x = work\_type, y = age, fill = stroke)) +  
 geom\_boxplot(position = "dodge", alpha = 0.7) +  
 labs(title = 'Age Distribution by Work Type and Stroke', x = 'Work type', y = 'Age') +  
 scale\_fill\_manual(values = c("navy", "indianred"))



# This boxplot shows more clearly the correlation between age and working status when it comes to suffering stroke.  
  
  
## Work type/Hypertension/Martial Status/Stroke ##  
  
ggplot(stroke, aes(x = work\_type, fill = interaction(hypertension, ever\_married), color = stroke)) +  
 geom\_bar(position = "dodge", alpha = 0.7, stat = "count") +  
 labs(title = 'Work Type by Hypertension, Martial Status, and Stroke',  
 x = 'Work Type', y = 'Count') +  
 scale\_fill\_manual(values = c("lightblue", "darkblue", "lightpink", "darkred"),  
 name = "Hypertension and Marital Status") +  
 scale\_color\_manual(values = c("black", "red"), name = "Stroke") +  
 theme(axis.text.x = element\_text(angle = 45, hjust = 1))



# Based on this chart we can see that as stated before hypertension, work type and martial status are main factors.  
  
#### Logistic Regression ####  
  
### Heatmap - correlation of risk factors ###  
  
# Checking for missing values  
if (any(is.na(stroke))) {  
 cat("Warning: There are missing values in the data. Please handle them before creating the heatmap.\n")  
}  
  
# Checking for infinite values  
if (any(sapply(stroke, function(x) any(is.infinite(x))))) {  
 cat("Warning: There are infinite values in the data. Please handle them before creating the heatmap.\n")  
}  
  
# Convert all columns to numeric  
stroke <- as.data.frame(sapply(stroke, as.numeric))  
  
# Calculate correlation  
stroke.cor <- round(cor(stroke), 2)  
  
# Create heatmap using corrplot  
  
corrplot(stroke.cor, method = "color", type = "upper", addCoef.col = "black")



## Heatmap conclusion ##  
  
# Upon examination of the heatmap, it is discernible that a substantial proportion of features exhibits a lack   
# of noteworthy correlation with one another, rendering them conducive to regression analysis. Noteworthy   
# correlation is primarily confined to the relationship between age and ever\_married, albeit the underlying   
# rationale for this association is straightforward. Among the array of features, age manifests the most   
# elevated correlation coefficients, particularly in relation to the occurrence of stroke.  
  
  
### Creating a Logistic Regression Model ###  
  
str(stroke)

## 'data.frame': 5109 obs. of 11 variables:  
## $ gender : num 2 1 2 1 1 2 2 1 1 1 ...  
## $ age : num 67 61 80 49 79 81 74 69 59 78 ...  
## $ hypertension : num 1 1 1 1 2 1 2 1 1 1 ...  
## $ heart\_disease : num 2 1 2 1 1 1 2 1 1 1 ...  
## $ ever\_married : num 2 2 2 2 2 2 2 1 2 2 ...  
## $ work\_type : num 4 5 4 4 5 4 4 4 4 4 ...  
## $ residence\_type : num 2 1 1 2 1 2 1 2 1 2 ...  
## $ avg\_glucose\_level: num 229 202 106 171 174 ...  
## $ bmi : num 36.6 28.1 32.5 34.4 24 29 27.4 22.8 28.1 24.2 ...  
## $ smoking\_status : num 1 2 2 3 2 1 2 2 4 4 ...  
## $ stroke : num 2 2 2 2 2 2 2 2 2 2 ...

# splitting data into training and testing sets before training the model  
  
install.packages("caTools")

## Installing package into 'C:/Users/ewwci/AppData/Local/R/win-library/4.3'  
## (as 'lib' is unspecified)

## package 'caTools' successfully unpacked and MD5 sums checked

## Warning: cannot remove prior installation of package 'caTools'

## Warning in file.copy(savedcopy, lib, recursive = TRUE): problem copying  
## C:\Users\ewwci\AppData\Local\R\win-library\4.3\00LOCK\caTools\libs\x64\caTools.dll  
## to C:\Users\ewwci\AppData\Local\R\win-library\4.3\caTools\libs\x64\caTools.dll:  
## Permission denied

## Warning: restored 'caTools'

##   
## The downloaded binary packages are in  
## C:\Users\ewwci\AppData\Local\Temp\RtmpMb4TYS\downloaded\_packages

library(caTools)

## Warning: package 'caTools' was built under R version 4.3.2

set.seed(123)  
split <- sample.split(stroke$stroke, SplitRatio = 0.7)  
train\_data <- subset(stroke, split == TRUE)  
test\_data <- subset(stroke, split == FALSE)  
  
# Training the logistic regression model  
  
# Checking unique values in the 'stroke' variable in train\_data  
unique(train\_data$stroke)

## [1] 2 1

# Converting variable to 0 and 1  
train\_data$stroke <- as.factor(ifelse(train\_data$stroke == "1", 1, 0))  
  
# Training the logistic regression model  
model <- glm(stroke ~ ., data = train\_data, family = "binomial")  
  
summary(model)

##   
## Call:  
## glm(formula = stroke ~ ., family = "binomial", data = train\_data)  
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 8.537425 0.931092 9.169 < 2e-16 \*\*\*  
## gender -0.239770 0.167734 -1.429 0.15287   
## age -0.074484 0.006656 -11.191 < 2e-16 \*\*\*  
## hypertension -0.436022 0.194657 -2.240 0.02509 \*   
## heart\_disease -0.240324 0.230359 -1.043 0.29683   
## ever\_married 0.194655 0.264630 0.736 0.46199   
## work\_type 0.038102 0.089047 0.428 0.66873   
## residence\_type -0.184290 0.166290 -1.108 0.26776   
## avg\_glucose\_level -0.004018 0.001424 -2.821 0.00479 \*\*   
## bmi -0.001275 0.013751 -0.093 0.92615   
## smoking\_status 0.052924 0.080317 0.659 0.50993   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 1391.4 on 3575 degrees of freedom  
## Residual deviance: 1093.6 on 3565 degrees of freedom  
## AIC: 1115.6  
##   
## Number of Fisher Scoring iterations: 7

# The logistic regression model was fitted to predict the occurrence of stroke based on various predictor variables:  
  
# Significant Predictors:  
  
# Age is a highly significant predictor (p < 2e-16), with a negative coefficient of -0.074484, suggesting a decrease in the log-odds of stroke as age increases.  
# Hypertension is a significant predictor (p = 0.02509), with a negative coefficient of -0.436022, indicating a decrease in the log-odds of stroke for individuals with hypertension.  
# Average glucose level is a significant predictor (p = 0.00479), with a negative coefficient of -0.004018, implying a decrease in the log-odds of stroke for higher average glucose levels.  
  
# Non-Significant Predictors:  
  
# Gender, heart disease, ever\_married, work\_type, residence\_type, BMI, and smoking\_status are not found to be significant predictors of stroke based on their p-values.  
  
# Model Fit:  
  
# The model's null deviance is 1391.4, and the residual deviance is 1093.6. A lower residual deviance suggests a better fit of the model to the data.  
# The AIC is 1115.6, providing a measure of the model's goodness of fit while penalizing for the number of parameters.  
  
# Additional information:  
  
# The intercept is significantly different from zero, indicating the presence of stroke cases even when other predictors are zero.  
# The number of Fisher Scoring iterations during model estimation is 7.  
  
# Overall, the logistic regression model suggests that age, hypertension, and average glucose level are important predictors   
# in determining the likelihood of stroke in the dataset.  
# It provides insights into the relationships between these factors and the occurrence of stroke.  
  
## The deviance of the regression model ( Anova ) ##   
  
anova(model, test="Chisq")

## Analysis of Deviance Table  
##   
## Model: binomial, link: logit  
##   
## Response: stroke  
##   
## Terms added sequentially (first to last)  
##   
##   
## Df Deviance Resid. Df Resid. Dev Pr(>Chi)   
## NULL 3575 1391.4   
## gender 1 1.816 3574 1389.6 0.177823   
## age 1 276.434 3573 1113.1 < 2.2e-16 \*\*\*  
## hypertension 1 6.956 3572 1106.2 0.008356 \*\*   
## heart\_disease 1 1.708 3571 1104.5 0.191302   
## ever\_married 1 0.280 3570 1104.2 0.596684   
## work\_type 1 0.359 3569 1103.8 0.548903   
## residence\_type 1 1.445 3568 1102.4 0.229259   
## avg\_glucose\_level 1 8.293 3567 1094.1 0.003981 \*\*   
## bmi 1 0.011 3566 1094.1 0.915222   
## smoking\_status 1 0.437 3565 1093.6 0.508715   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

# In summary, age, hypertension, and avg\_glucose\_level are significant predictors of stroke in this model, while gender,  
# heart\_disease, ever\_married, work\_type, residence\_type, bmi, and smoking\_status do not significantly contribute to explaining  
# the variation in the response variable.  
  
# Obtaining predicted probabilities  
predicted\_probabilities <- predict(model, newdata = test\_data, type = "response")  
  
## Confusion Matrix ##  
conf\_matrix <- table(true\_labels = test\_data$stroke, predicted\_labels = ifelse(predicted\_probabilities > 0.5, 1, 0))  
  
# Display Confusion Matrix  
conf\_matrix

## predicted\_labels  
## true\_labels 1  
## 1 1458  
## 2 75

#Based on the provided confusion matrix:  
   
# True Positive (TP): 1458  
# True Negative (TN): 0  
# False Positive (FP): 0  
# False Negative (FN): 75  
  
# This means that the model correctly predicted 1458 instances of stroke (True Positives), and it incorrectly predicted   
# 75 instances as non-stroke that were actually strokes (False Negatives). There are no instances predicted as strokes that  
# are actually non-strokes (False Positives), and no instances correctly predicted as non-strokes (True Negatives).   
# The model seems to have good sensitivity but may need further tuning for specificity.  
  
## Creating ROC curve ##  
install.packages("pROC")

## Installing package into 'C:/Users/ewwci/AppData/Local/R/win-library/4.3'  
## (as 'lib' is unspecified)

## package 'pROC' successfully unpacked and MD5 sums checked

## Warning: cannot remove prior installation of package 'pROC'

## Warning in file.copy(savedcopy, lib, recursive = TRUE): problem copying  
## C:\Users\ewwci\AppData\Local\R\win-library\4.3\00LOCK\pROC\libs\x64\pROC.dll to  
## C:\Users\ewwci\AppData\Local\R\win-library\4.3\pROC\libs\x64\pROC.dll:  
## Permission denied

## Warning: restored 'pROC'

##   
## The downloaded binary packages are in  
## C:\Users\ewwci\AppData\Local\Temp\RtmpMb4TYS\downloaded\_packages

library(pROC)

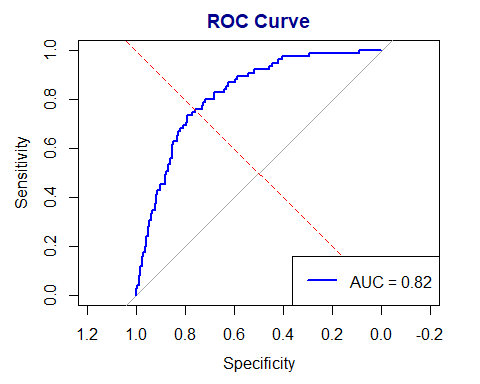
## Warning: package 'pROC' was built under R version 4.3.2

## Type 'citation("pROC")' for a citation.  
##   
## Attaching package: 'pROC'  
##   
## The following objects are masked from 'package:stats':  
##   
## cov, smooth, var

roc\_curve <- roc(test\_data$stroke, predicted\_probabilities)

## Setting levels: control = 1, case = 2  
## Setting direction: controls > cases

# Plotting the ROC curve  
plot(roc\_curve, col = "blue", main = "ROC Curve", col.main = "darkblue", lwd = 2)  
  
# Add labels and legend  
abline(a = 0, b = 1, lty = 2, col = "red") # Diagonal line for reference  
legend("bottomright", legend = paste("AUC =", round(auc(roc\_curve), 2)), col = "blue", lwd = 2)



# The ROC analysis for the logistic regression model yielded an Area Under the Curve (AUC) of 0.82.  
# This AUC value signifies excellent discriminatory power, indicating the model's strong ability to distinguish between   
# positive and negative cases.  
# Overall, with an AUC of 0.82, the logistic regression model demonstrates high accuracy in making predictions for   
# the binary classification task in the dataset.  
  
#### Summary ####  
  
# Conclusion and Recommendations  
  
## Data Transformation and Cleaning:  
  
# The dataset underwent necessary transformations, including unifying column names, removing unnecessary ID columns, and converting data types to facilitate analysis.  
  
# Missing values in the BMI column were handled by replacing them with the median BMI value, ensuring data completeness.  
  
## Data Analysis:  
  
# Statistical summaries and visualizations were conducted on numerical variables such as age, average glucose level, and BMI.  
  
# Age emerged as a crucial factor, with a wide range of values. The distribution was relatively symmetric, with a notable concentration of stroke cases in the 60-80 age group.  
  
# Average glucose level and BMI were also explored, revealing potential outliers and the need for further investigation.  
  
# Mosaic plots and correlation analyses were performed to identify potential correlations between various factors and the occurrence of strokes.  
  
## Logistic Regression:  
  
# A logistic regression model was trained to predict the likelihood of stroke based on several predictor variables.  
  
# Significant predictors included age, hypertension, and average glucose level, while gender, heart disease, marital status, work type, residence type, BMI, and smoking status did not significantly contribute to the model.  
  
# The model demonstrated good fit, with the ROC analysis yielding an Area Under the Curve (AUC) of 0.82, indicating excellent discriminatory power.  
  
## Recommendations:  
  
# Age Awareness:  
  
# - Pay close attention to older individuals, as age emerged as a decisive factor in the risk of stroke. Regular health check-ups and targeted interventions for this demographic could be beneficial.  
  
# Hypertension Management:  
  
# - Given that hypertension was a significant predictor, serious attention to hypertension management is crucial. Regular monitoring and appropriate medical interventions for individuals with hypertension may help mitigate the risk of stroke.  
  
# Occupational Considerations:  
  
# - Individuals in self-employed roles were found to have a reduced risk of stroke. Exploring the lifestyle and occupational factors associated with self-employment could provide insights into mitigating stroke risk.  
  
# Further Investigation:  
  
# - Explore anomalies and outliers in variables such as average glucose level and BMI, as they may have clinical implications and could be further investigated to enhance stroke prediction accuracy.  
  
# Model Refinement:  
  
# - While the logistic regression model demonstrated good overall performance, continued refinement and exploration of additional features or machine learning algorithms could enhance predictive accuracy, especially in addressing the imbalances in the dataset.  
  
# In conclusion, this analysis provides valuable insights into factors influencing stroke occurrence. Implementing the recommendations could contribute to proactive health management and the development of more accurate stroke prediction models.

In this stroke prediction data analysis project, comprehensive transformations and cleaning processes were applied to the dataset, ensuring uniformity and completeness. Notable findings emerged during the exploration of numerical variables, with age identified as a critical factor, particularly in the 60-80 age group. The logistic regression model, trained to predict stroke likelihood, highlighted significant predictors such as age, hypertension, and average glucose level. Recommendations include heightened awareness for age-related risks, diligent hypertension management, and further investigation into occupational influences, notably the reduced stroke risk associated with self-employment. The model's good fit, as evidenced by an Area Under the Curve (AUC) of 0.82, underscores its accuracy in distinguishing positive and negative cases. Practical suggestions encompass ongoing refinement of the model, addressing imbalances in the dataset, and exploring additional features for enhanced predictive capabilities. Overall, the analysis offers valuable insights for proactive health management and the continual improvement of stroke prediction models.