ASSIGNMENT (STROKE PREDICTION)

Deepasri

2023-02-19

ASSUMPTION:

This dataset is about “Stroke prediction”. And I assume many things using this dataset. Some of my attributes are categorical and some of them are numerical.

In this dataset I assume that the stroke is caused because of hypertension . Even though we can see , the gender wise analysis there is also a random increase of glucose level for men .And most of the people who are affected come under the category of married status and age is also the main reason . We can also assume that some of the people are affected because of smoking. My assumption may be wrong . But I created it only on my assumption .

R programming Code:

df=read.csv("healthcare-dataset-stroke-data.csv")  
head(dfs)

## 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

summary(df)

## id gender age hypertension   
## Min. : 67 Length:5110 Min. : 0.08 Min. :0.00000   
## 1st Qu.:17741 Class :character 1st Qu.:25.00 1st Qu.:0.00000   
## Median :36932 Mode :character Median :45.00 Median :0.00000   
## Mean :36518 Mean :43.23 Mean :0.09746   
## 3rd Qu.:54682 3rd Qu.:61.00 3rd Qu.:0.00000   
## Max. :72940 Max. :82.00 Max. :1.00000   
## heart\_disease ever\_married work\_type Residence\_type   
## Min. :0.00000 Length:5110 Length:5110 Length:5110   
## 1st Qu.:0.00000 Class :character Class :character Class :character   
## Median :0.00000 Mode :character Mode :character Mode :character   
## Mean :0.05401   
## 3rd Qu.:0.00000   
## Max. :1.00000   
## avg\_glucose\_level bmi smoking\_status stroke   
## Min. : 55.12 Length:5110 Length:5110 Min. :0.00000   
## 1st Qu.: 77.25 Class :character Class :character 1st Qu.:0.00000   
## 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

str(df)

## '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 ...

*#installing libraries*  
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(lattice)  
library(rmarkdown)  
library(ggplot2)  
  
glimpse(df)

## Rows: 5,110  
## Columns: 12  
## $ id <int> 9046, 51676, 31112, 60182, 1665, 56669, 53882, 10434…  
## $ gender <chr> "Male", "Female", "Male", "Female", "Female", "Male"…  
## $ age <dbl> 67, 61, 80, 49, 79, 81, 74, 69, 59, 78, 81, 61, 54, …  
## $ hypertension <int> 0, 0, 0, 0, 1, 0, 1, 0, 0, 0, 1, 0, 0, 0, 0, 1, 0, 1…  
## $ heart\_disease <int> 1, 0, 1, 0, 0, 0, 1, 0, 0, 0, 0, 1, 0, 1, 1, 0, 1, 0…  
## $ ever\_married <chr> "Yes", "Yes", "Yes", "Yes", "Yes", "Yes", "Yes", "No…  
## $ work\_type <chr> "Private", "Self-employed", "Private", "Private", "S…  
## $ Residence\_type <chr> "Urban", "Rural", "Rural", "Urban", "Rural", "Urban"…  
## $ avg\_glucose\_level <dbl> 228.69, 202.21, 105.92, 171.23, 174.12, 186.21, 70.0…  
## $ bmi <chr> "36.6", "N/A", "32.5", "34.4", "24", "29", "27.4", "…  
## $ smoking\_status <chr> "formerly smoked", "never smoked", "never smoked", "…  
## $ stroke <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1…

*#dimension*  
dim(df)

## [1] 5110 12

*#remove null values*  
*#df\_no\_na= na.omit(df)*  
*#f\_no\_na*  
names(df)

## [1] "id" "gender" "age"   
## [4] "hypertension" "heart\_disease" "ever\_married"   
## [7] "work\_type" "Residence\_type" "avg\_glucose\_level"  
## [10] "bmi" "smoking\_status" "stroke"

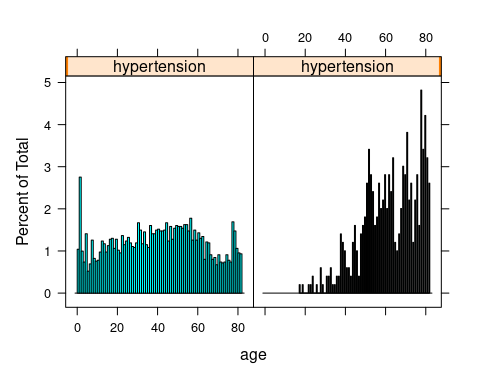
*#histogram*  
histogram(~age|hypertension,data=df,breaks=100,Col=c('skyblue','purple'),mains="hypertension wise analysis")

INSIGHT:

The mean and median is slightly close to each other and it shows about 68% of standard deviation .Most of them are caused due to high hypertension.

INFERENCE:

hypertension is more common in older age groups . This shows that it is skewed right. And the median shows that it is equally in the normal aged people who are mostly affected by stroke.



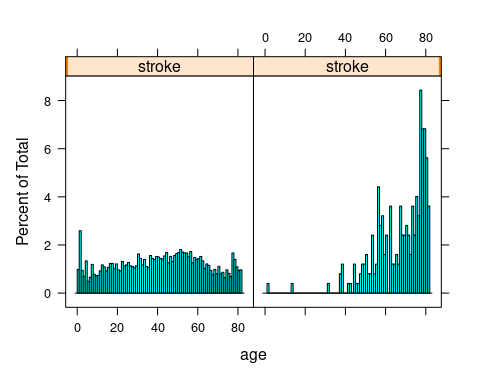
histogram(~age|stroke,data=df,breaks=80,Col=c('yellow','green'),mains="stroke wise analysis")

INSIGHT:

The dataset is having insight about the age, that aged people are highly affected by stroke because of various reasons.

INFERENCE:

The dataset is right- skewed ,with a majority of individuals being between the ages of 40-70. The shape of the histogram tells that older people are affected a lot.The age is the significant risk for the stroke.



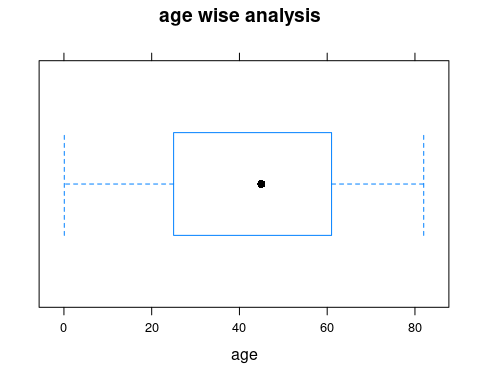
*#boxplot*  
bwplot(~age,data=df,breaks=100,Col=c('pink','gray'),main="age wise analysis")

INSIGHT:

The boxplot shows that the median age value is around 50 years old. It has a slightly interquartile range.

INFERENCE:

My assumption is that the people who are affected by the stroke are around the age of 50 years exactly. And they lie between inter quartile range and they are about to 50% of the age and some of them are under the outliers in the plot.



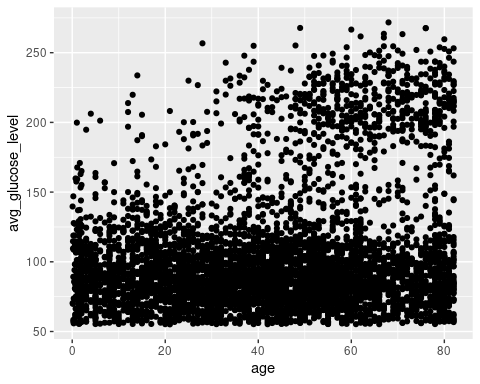
*#scatterplot*  
ggplot(df,aes(x=age,y=avg\_glucose\_level))+geom\_point()

INSIGHT:

The age increases the same way glucose levels also increase. There is a positive relationship. There are few outliers in the following graph.

INFERENCE:

The plot shows that there is a positive correlation between age and average glucose level. There is an age increase and glucose level also increases. The outliers represent in the data collection or other anomalies. The statistical test for correlation analysis.



## Including Plots

You can also embed plots, for example:

```

Note that the echo = FALSE parameter was added to the code chunk to prevent printing of the R code that generated the plot.