AMSA-HW1

ID : 111024517 Name : 鄭家豪

due on 02/23

About Data

Accroding to WHO,stroke is the 2nd leading cause of death globally.
This dataset is used to predict whether a patient is likely to get stoke based on the 11 clinical features.

The resourse : https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset

Attribute Information

- (discrete)id: unique identifier
- (discrete)gender: “Male”, “Female” or “Other”
- (continuous)age: age of the patient
- (discrete)hypertension: 0 if the patient doesn’t have hypertension, 1 if the patient has hypertension
- (discrete)heart_disease: 0 if the patient doesn’t have any heart diseases, 1 if the patient has a heart disease
- (discrete)ever_married: “No” or “Yes”
- (discrete)work_type: “children”, “Govt_jov”, “Never_worked”, “Private” or “Self-employed”
- (discrete)Residence_type: “Rural” or “Urban”
- (continuous)avg_glucose_level: average glucose level in blood
- (continuous)bmi: body mass index
- (discrete)smoking_status: “formerly smoked”, “never smoked”, “smokes” or “Unknown”
- (discrete)stroke: 1 if the patient had a stroke or 0 if not

Simple EDA

Check data

```
## 'data.frame':    5110 obs. of  12 variables:
```

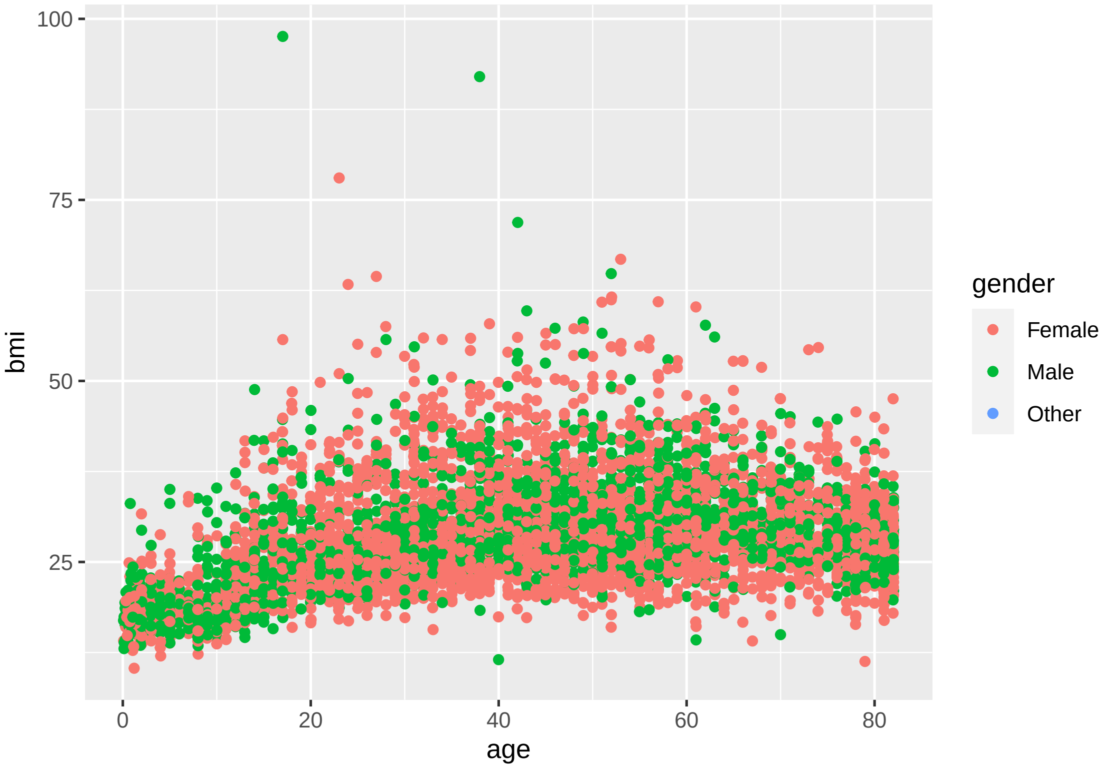
```
##  $ id               : int  9046 51676 31112 60182 1665 56669 53882 10434 27419 60491 ...
##  $ gender           : chr  "Male" "Female" "Male" "Female" ...
##  $ age              : num  67 61 80 49 79 81 74 69 59 78 ...
##  $ hypertension     : int  0 0 0 0 1 0 1 0 0 0 ...
##  $ heart_disease    : 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 ...
```

From the above , we find the “NA” value in bmi variable.We chose to remove all NA’s , assuming that they are not important for the analysis . Then , obtained :

```
## 'data.frame':    4909 obs. of  12 variables:
##  $ id               : int  9046 31112 60182 1665 56669 53882 10434 60491 12109 12095 ...
##  $ gender           : chr  "Male" "Male" "Female" "Female" ...
##  $ age              : num  67 80 49 79 81 74 69 78 81 61 ...
##  $ hypertension     : int  0 0 0 1 0 1 0 0 1 0 ...
##  $ heart_disease    : int  1 1 0 0 0 1 0 0 0 1 ...
##  $ ever_married     : chr  "Yes" "Yes" "Yes" "Yes" ...
##  $ work_type        : chr  "Private" "Private" "Private" "Self-employed" ...
##  $ Residence_type   : chr  "Urban" "Rural" "Urban" "Rural" ...
##  $ avg_glucose_level: num  229 106 171 174 186 ...
##  $ bmi              : num  36.6 32.5 34.4 24 29 27.4 22.8 24.2 29.7 36.8 ...
##  $ smoking_status   : chr  "formerly smoked" "never smoked" "smokes" "never smoked" ...
##  $ stroke           : int  1 1 1 1 1 1 1 1 1 1 ...
```

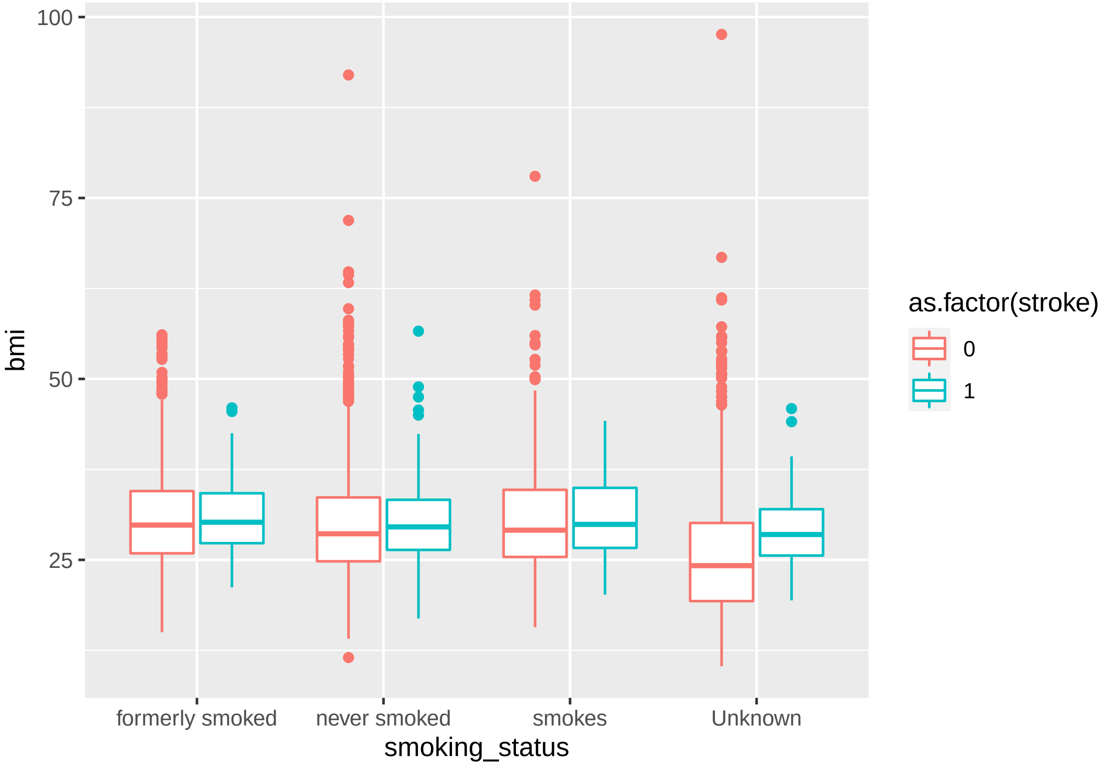
Here we have 4909 observations and 12 variables after removing the row data containing the NA values.

Scatter plot(Analysis of the sampling distribution)



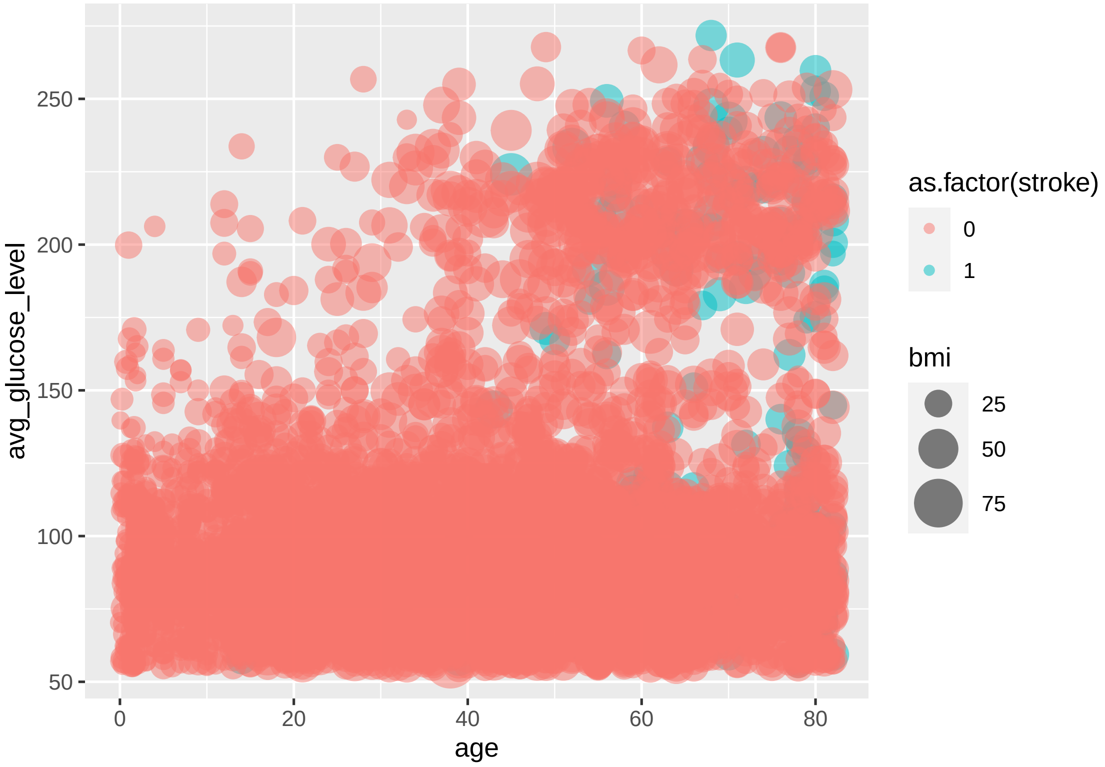
這裡展示出 age-bmi 的 scatter plot 且以 gender 來區別每個點的散佈狀況，可以大致上觀察出，年齡對於男女比例是差不多的且女生的 BMI 相對男生沒有明顯的落差，可能意味著收集資料時，針對不同年齡的男女 bmi 落差不會很顯著。不過有幾個 outlier 很明顯有所區別，其 BMI 值接近 100，屬於男性。

Boxplot(Analysis of information on smoking habits)



這裡展示出 smoking_status-BMI 的 boxplot 且以中風與否來做區別，主要是想了解，對於 4 種不同抽菸習慣，基於中風情況不同，bmi 分布是否相似？從上面可以看出，除了未知抽菸習慣有明顯的一點不同外，其他三種抽菸習慣的分布都蠻相似的。

Bubble plot(Analysis of information on glucose)



這裡展示出以 bmi 大小表示 bubble 形狀，age-血糖的 Bubble plot ，另外以中風與否來做區別。

這裡可以看出一些資訊，年齡大且平均血糖較高的人，較容易得到中風。另外，Bmi 並沒有對於中風與否提供太明顯的區別，因為沒有中風的人當中，具有高 Bmi 的人是佔多數，以及中風的人當中，bmi 值似乎都差不多。