Student ID: 20125106

Full name: Le Van Hoang Phi

1 Giới thiệu về bộ dữ liệu

1.1 Mô tả chung

- Khung dữ liệu gồm có 5110 dòng và 12 cột bao gồm: id, gender, age, hypertension, heart_disease, ever_married, work_type, Residence_type, avg_glucose_level, bmi, smoking_status, stroke.
- Dữ liệu được lấy từ địa chỉ này.
- Vì lí do dữ liệu có một vài phần không cần thiết như là các hàng có giá trị gender là "Other" hoặc là các hàng chứa cột không có giá trị nào, nên trong bài báo cáo này em đã xoá những thành phần đấy đi và còn lại 4908 dòng.

1.2 Định dạng, cấu trúc của bộ dữ liệu

Bộ dữ liệu có các cột sau:

STT	Thuộc tính	Ý nghĩa
1	id	Dùng để định danh người thực hiện quan sát
2	gender	Giới tính của người thực hiện quan sát ("Male" hoặc "Female")
3	age	Tuổi của người thực hiện quan sát
4	hypertension	Xác định người thực hiện quan sát có cao huyết áp hay không (1 = Có, 0 = Không có)
5	heart_disease	Xác định người thực hiện quan sát có bệnh tim hay không (1 = Có, 0 = Không có)
6	ever_married	Xác định người thực hiện quan sát đã từng kết hôn hay chưa (1 = Có, 0 = Không có)
7	work_type	Xác định loại công việc mà người thực hiện quan sát đang làm ("children", "Govt_jov", "Never_worked", "Private" hoặc "Self-employed")
8	Residence_type	Xác định nơi cư trú ("Rural" hoặc "Urban")
9	avg_glucose_level	Lượng đường trung bình trong máu
10	bmi	Chỉ số khối cơ thể
11	smoking_status	Tình trạng hút thuốc ("formerly smoked", "never smoked", "smokes" hoặc "Unknown")
12	stroke	Xác định người quan sát có bị đột quỵ hay không (1 = Có, 0 = Không có)

2 Thống kê mô tả dựa trên bộ dữ liệu

2.1 Tạo môi trường làm việc và load dữ liệu

- Code tạo bộ dữ liệu:

```
## Thiet lap moi truong lam viec
setwd("E:/20125106/Sem3_Year2/STAT452/Final/")
## Doc file du lieu csv tren bien data
data <- read.csv("healthcare-dataset-stroke-data.csv", na = "N/A")
## Xoa cac hang co chua cot khong co gia tri nao
data <- na.omit(data)
## Xoa cac hang co chua gia tri cua cot gender la "Other"
data <- data[!(data$gender == "Other"),]

attach(data)
```

- Code cung cấp thống kê mô tả căn bản và tần số của R:

```
data $ hypertension <- factor(data $ hypertension)
levels(data $ hypertension) <- c("No", "Yes")

data $ heart_disease <- factor(data $ heart_disease)
levels(data $ heart_disease) <- c("No", "Yes")

data $ stroke <- factor(data $ stroke)
levels(data $ stroke) <- c("No", "Yes")

data $ gender <- factor(data $ gender)

data $ work_type <- factor(data $ work_type)

data $ Residence_type <- factor(data $ Residence_type)

data $ smoking_status <- factor(data $ smoking_status)

summary(data)
```

- Kết quả nhận được sau khi xử lí bộ dữ liệu:

```
       id                gender             age          hypertension
 Min.   :   77         Female:2897      Min.   : 0.08      No :4457
 1st Qu.:18603         Male  :2011      1st Qu.:25.00      Yes: 451
 Median :37581                          Median :44.00
 Mean   :37060                          Mean   :42.87
 3rd Qu.:55182                          3rd Qu.:60.00
 Max.   :72940                          Max.   :82.00

 heart_disease       ever_married            work_type
 No :4665            Length:4908          children    : 671
 Yes: 243            Class :character     Govt_job    : 630
```

```
                    Mode  :character      Never_worked : 22
                                          Private      :2810
                                          Self-employed: 775

Residence_type      avg_glucose_level        bmi
Rural:2418          Min.   : 55.12       Min.   :10.30
Urban:2490          1st Qu.: 77.07       1st Qu.:23.50
                    Median : 91.68       Median :28.10
                    Mean   :105.30       Mean   :28.89
                    3rd Qu.:113.50       3rd Qu.:33.10
                    Max.   :271.74       Max.   :97.60
   smoking_status              stroke
formerly smoked: 836          No :4699
never smoked  :1852           Yes: 209
smokes        : 737
Unknown       :1483
```

2.2 Các thống kê cơ bản về các cột thuộc tính dữ liệu

2.2.1 Thuộc tính "stroke":

- Vẽ biểu đồ Pie thể hiện tỉ lệ người bị đột quỵ.
  - Code:

```
df.data$stroke <- data$stroke

df.stroke <- df.data %>% group_by(stroke) %>% summarise(count = n()) %>%
             mutate(count = round(count * 100 / sum(count), digits = 2))

df2 <- df.stroke %>%
  mutate(csum = rev(cumsum(rev(count))),
         pos = count/2 + lead(csum, 1),
         pos = if_else(is.na(pos), count/2, pos))

ggplot(df.stroke, aes(x = "" , y = count, fill = fct_inorder(stroke))) +
  geom_col(width = 1, color = 1) +
  coord_polar(theta = "y") +
  scale_fill_brewer(palette = "Pastel1") +
  geom_label_repel(data = df2,
                   aes(y = pos, label = paste0(count, "%")),
                   size = 4.5, nudge_x = 1, show.legend = FALSE) +
  guides(fill = guide_legend(title = "Stroke")) +
  theme_void()

ggsave(filename = "stroke.png", device = "png",width = 19,height = 10,
units="cm")
```

– Kết quả:

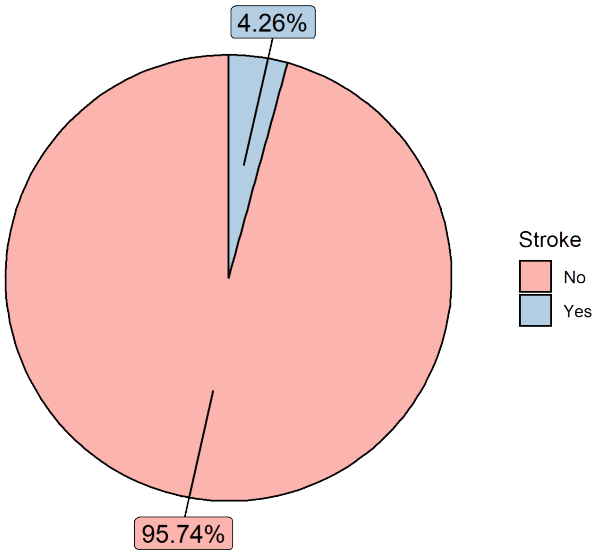


Figure 1: Biểu đồ thể hiện tỉ lệ đột quỵ.

- Nhận xét: Tỉ lệ người khảo sát bị đột quỵ chiếm số lượng thấp.

2.2.2 Thuộc tính "gender":

- Lập bảng thống kê tỉ lệ đối tượng ở từng giới tính:

  – Code:

```
df.data $ gender <- data$gender

df.gender <- df.data %>% group_by(gender) %>% summarise(count = n()) %>%
           mutate(count = round(count * 100 / sum(count), digits = 2))

df2 <- df.gender %>% mutate(csum = rev(cumsum(rev(df.gender$count))),
        pos = df.gender$count / 2 + lead(csum, 1),
        pos = if_else(is.na(pos), df.gender$count/2, pos))

ggplot(df.gender, aes(x = "" , y = count, fill = fct_inorder(gender))) +
  geom_col(width = 1, color = 1) +
  coord_polar(theta = "y", start = 0) +
  scale_fill_brewer(palette = "Pastel1") +
  geom_label_repel(data = df2,
                   aes(y = pos, label = paste0(count, "%")),
                   size = 4.5, nudge_x = 1, show.legend = FALSE) +
  guides(fill = guide_legend(title = "Gender")) +
  theme_void()
```

– Kết quả:

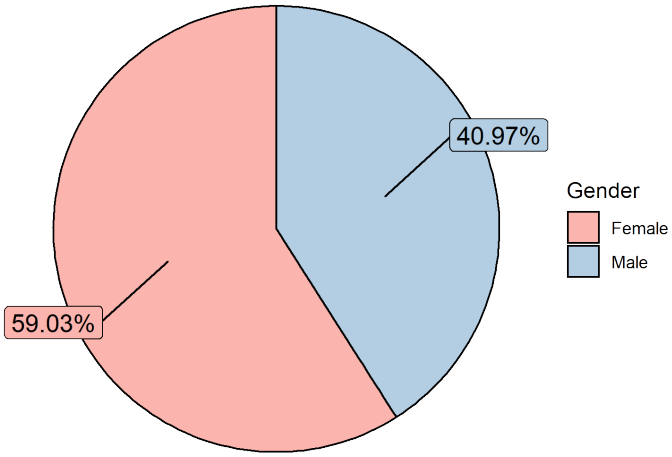


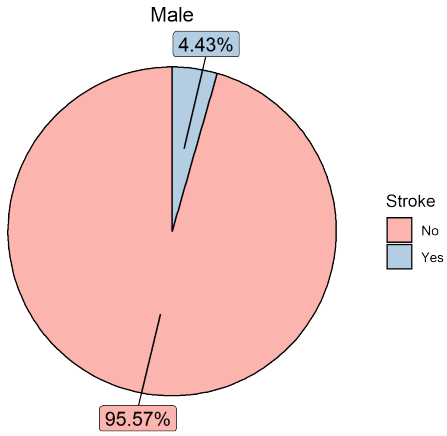
Figure 2: Biểu đồ thể hiện tỉ lệ của hai giới tính.

- Lập biểu đồ thống kê tỉ lệ đột quỵ ở từng giới tính:
  – Code:

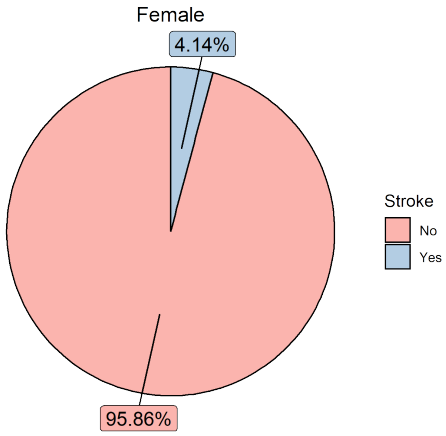
```
## ------------------------------ Male

df.GenderVsStroke <- df.data %>% group_by(gender, stroke) %>% summarise(count
= n()) %>% mutate(count = round(count * 100 / sum(count), digits = 2))

df.MaleVsStroke <- df.GenderVsStroke[3:4,]

df2 <- df.MaleVsStroke %>%
  mutate(csum = rev(cumsum(rev(df.MaleVsStroke$count))),
         pos = df.MaleVsStroke$count / 2 + lead(csum, 1),
         pos = if_else(is.na(pos), df.MaleVsStroke$count/2, pos))

ggplot(df.MaleVsStroke, aes(x = "" , y = count, fill = fct_inorder(stroke)))
  + geom_col(width = 1, color = 1) +
  coord_polar(theta = "y") + facet_wrap(~gender, ncol = 2, scale = "fixed") +
  scale_fill_brewer(palette = "Pastel1") +
  geom_label_repel(data = df2,
                   aes(y = pos, label = paste0(count, "%")),
                   size = 4.5, nudge_x = 1, show.legend = FALSE) +
  guides(fill = guide_legend(title = "Stroke")) +
  ggtitle("Male") +
  theme_void() +
  theme(plot.title = element_text(hjust = 0.5))

ggsave(filename = "male.png", device = "png",width = 19,height = 10,
       units="cm")
```

```
## ------------------------------ Female

df.FemaleVsStroke <- df.GenderVsStroke[1:2,]

df2 <- df.FemaleVsStroke %>%
  mutate(csum = rev(cumsum(rev(df.FemaleVsStroke$count))),
         pos = df.FemaleVsStroke$count / 2 + lead(csum, 1),
         pos = if_else(is.na(pos), df.FemaleVsStroke$count/2, pos))

ggplot(df.FemaleVsStroke, aes(x = "" , y = count,
  fill = fct_inorder(stroke))) + geom_col(width = 1, color = 1) +
  coord_polar(theta = "y") +
  scale_fill_brewer(palette = "Pastel1") +
  geom_label_repel(data = df2,
                   aes(y = pos, label = paste0(count, "%")),
                   size = 4.5, nudge_x = 1, show.legend = FALSE) +
  guides(fill = guide_legend(title = "Stroke")) +
  ggtitle("Female") +
  theme_void() +
  theme(plot.title = element_text(hjust = 0.5))

ggsave(filename = "female.png", device = "png",width = 19,height = 10,
       units="cm")
```

– Kết quả:



(a) Biểu đồ tỉ lệ đột quỵ ở nam giới



(b) Biểu đồ tỉ lệ đột quỵ ở nữ giới

- Nhận xét:
  - Tỉ lệ người được khảo sát có giới tính là nữ nhiều hơn so với nam (59.03% so với 40.97%).
  - Tuy nhiên tỉ lệ đột quỵ của hai giới gần bằng nhau.

2.2.3 Thuộc tính "age":

- Chia thành 3 nhóm tuổi: $\begin{cases} \text{Trẻ: dưới 40 tuổi,} \\ \text{Trung niên: trong khoảng từ 40 đến 60 tuổi} \\ \text{Già: lớn hơn 60 tuổi} \end{cases}$
- Định tính hoá thuộc tính "age" rồi lập bảng thống kê số lượng người được khảo sát theo 3 nhóm tuổi:
  - Code:

```
df.age = 1:length(data$age)

for (i in 1:length(data$age)) {
  if (age[i] < 40) {
    df.age[i] = "Tre"
  }
  else if (age[i] >= 40 && age[i] <= 60) {
    df.age[i] = "Trung nien"
  }
  else {
    df.age[i] = "Gia"
  }
}

table(df.age)
```

  - Kết quả:

```
> table(df.age)
df.age
      Gia       Tre      Trung nien
     1217      2113            1578
```

- Vẽ biểu đồ tần số dựa vào nhóm tuổi và bệnh đột quỵ của người được khảo sát:
  - Code:

```
df.data $ age <- df.age

df.AgeVsStroke <- df.data %>% group_by(age, stroke) %>%
                  summarise(count = n())

ggplot(df.AgeVsStroke, aes(x = age, y = count, fill = stroke)) +
  geom_bar(stat = "identity", color = 1) +
  xlab("Age") + ylab("Count") +
  scale_fill_discrete(name = "Stroke", labels = c("No", "Yes")) +
  theme_minimal()

ggsave(filename = "age_bar.png", device = "png",width = 19,height = 10,
       units="cm")
```

– Kết quả:

```
> df.AgeVsStroke
# A tibble: 6 × 3
# Groups:  age [3]
  age        stroke count
  <chr>      <fct>  <int>
1 Gia        No      1071
2 Gia        Yes      146
3 Tre        No      2107
4 Tre        Yes        6
5 Trung nien No      1521
6 Trung nien Yes       57
```

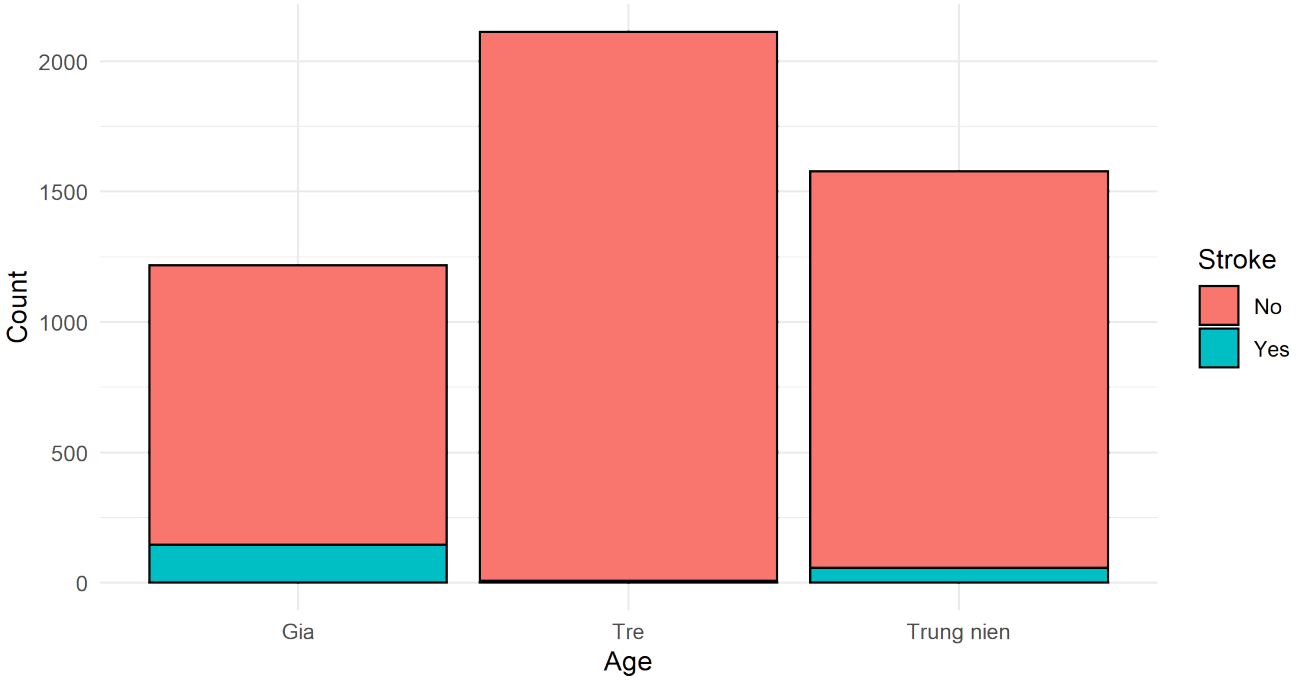


Figure 4: Biểu đồ thể hiện tần số đột quỵ của từng nhóm tuổi.

- Vẽ biểu đồ Box Plot tuổi theo hai nhóm đối tượng đột quỵ và không bị đột quỵ:
  – Code:

```
df.data$oAge <- data$age

ggplot(df.data, aes(x = oAge, y = stroke, fill = stroke)) +
  geom_boxplot() +
  xlab("Age") +
  ylab("Target") +
  theme(legend.position = "none")

ggsave(filename = "age_boxplot.png", device = "png",width = 19,height = 10,
       units="cm")
```

– Kết quả:

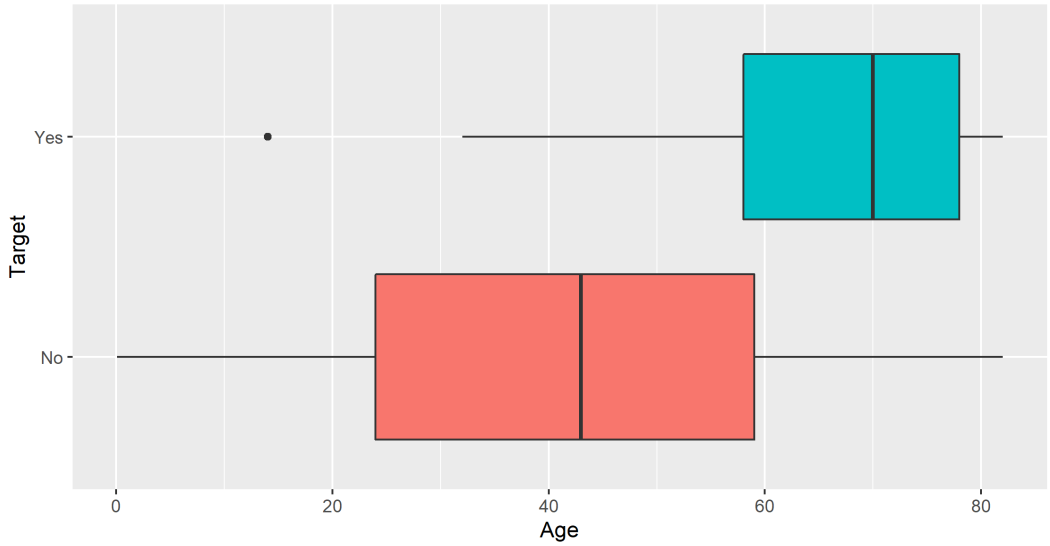


Figure 5: Box plot tuổi theo nhóm đối tượng.

- Nhận xét:
  - Nhóm tuổi già có tỉ lệ bị đột quỵ là cao nhất (12%) mặc dù có số lượng được khảo sát là ít nhất (1217).
  - Nhóm tuổi trẻ vẫn có khả năng bị đột quỵ.
  - Trung bình tuổi của nhóm bị đột quỵ cao hơn so với trung bình tuổi của nhóm không bị, hơn nữa tuổi của đối tượng bị đột quỵ tập trung nhiều trong khoảng từ 60 tuổi trở lên.

2.2.4 Thuộc tính "hypertension":

- Lập bảng thống kê tỉ lệ người mắc bệnh cao huyết áp:
  - Code:

```
df.data $ hypertension <- data$hypertension

df.hypertension <- df.data %>% group_by(hypertension) %>%
  summarise(count = n()) %>%
  mutate(count = round(count * 100 / sum(count), digits = 2))

df2 <- df.hypertension %>%
  mutate(csum = rev(cumsum(rev(df.hypertension$count))),
         pos = df.hypertension$count / 2 + lead(csum, 1),
         pos = if_else(is.na(pos), df.hypertension$count/2, pos))


ggplot(df.hypertension, aes(x = "" , y = count,
  fill = fct_inorder(hypertension))) +
  geom_col(width = 1, color = 1) +
```

```
  coord_polar(theta = "y") +
  scale_fill_brewer(palette = "Pastel1") +
  geom_label_repel(data = df2,
                   aes(y = pos, label = paste0(count, "%")),
                   size = 4.5, nudge_x = 1, show.legend = FALSE) +
  guides(fill = guide_legend(title = "hypertension")) +
  theme_void()

ggsave(filename = "hypertension_pie.png", device = "png",width = 19,
       height = 10, units="cm")
```

– Kết quả:

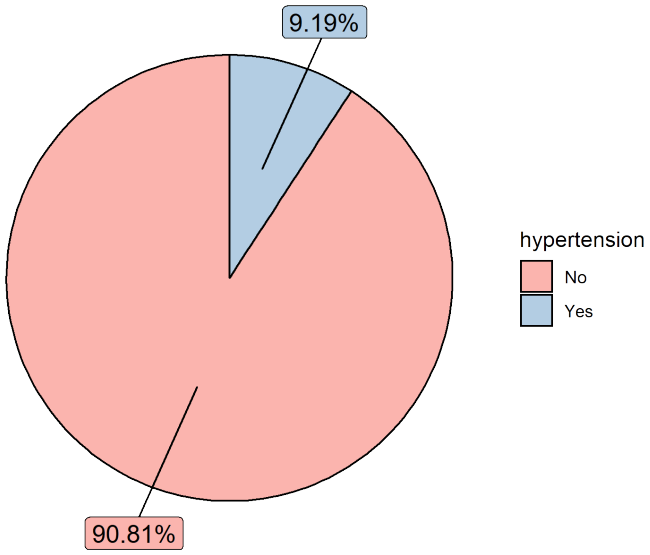


Figure 6: Biểu đồ thể hiện tỉ lệ của người mắc bệnh cao huyết áp.

- Lập bảng thống kê tần số của 2 nhóm mắc bệnh và không mắc bệnh cao huyết áp đối với đột quỵ:

  – Code:

```
df.HypertensionVsStroke <- df.data %>% group_by(hypertension, stroke) %>%
  summarise(count = n())

df.HypertensionVsStroke

ggplot(df.HypertensionVsStroke, aes(x = hypertension, y = count,
  fill = stroke)) +
  geom_col(position = "dodge", color = 1) +
  coord_flip() + xlab("Hypertension") + ylab("Frequency") +
  scale_fill_discrete(name = "Stroke", labels = c("No","Yes")) +
  theme_minimal()

ggsave(filename = "hypertension_bar.png", device = "png",width = 19,
       height = 10,units="cm")
```

– Kết quả:

```
> df.HypertensionVsStroke
# A tibble: 4 × 3
# Groups:  hypertension [2]
  hypertension stroke count
  <fct>        <fct>  <int>
1 No           No      4308
2 No           Yes      149
3 Yes          No       391
4 Yes          Yes       60
```

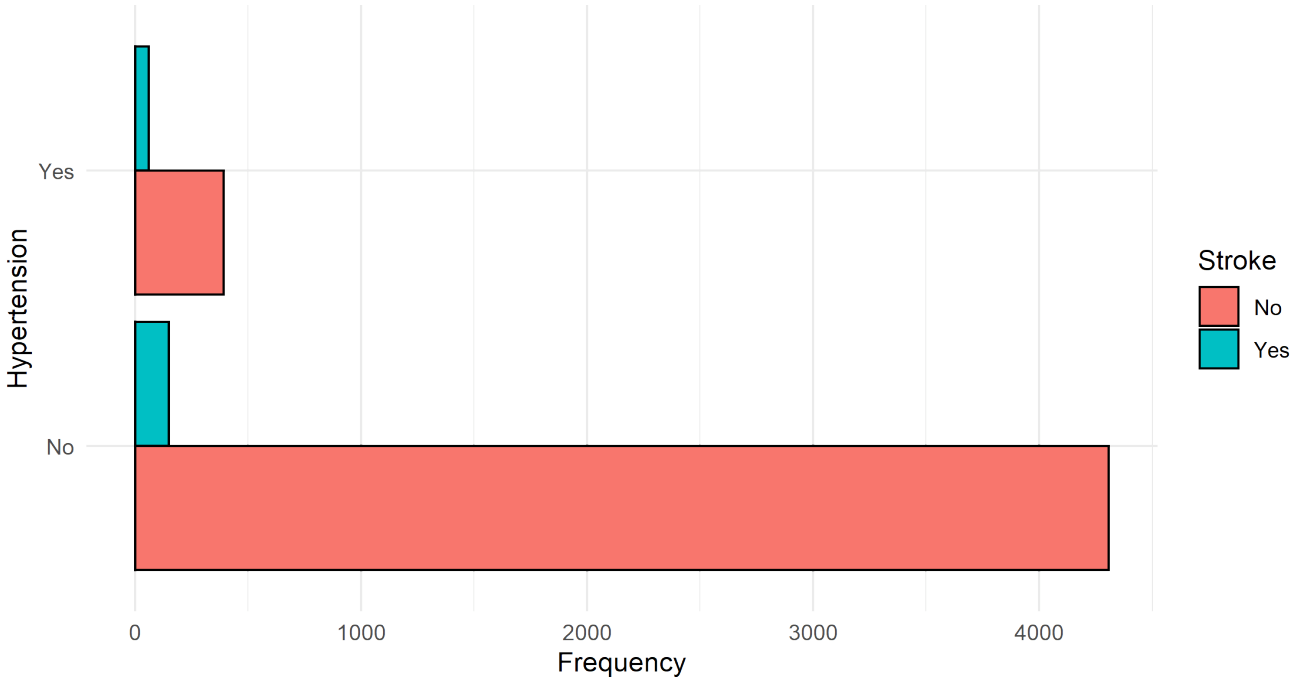


Figure 7: Biểu đồ thể hiện tần số đột quỵ của hai nhóm mắc và không mắc bệnh cao huyết áp.

- Nhận xét:
  - Số người khảo sát mắc bệnh cao huyết áp chiếm số lượng không lớn trong tập dữ liệu (khoảng 10%).
  - Trong số những người mắc bệnh cao huyết áp thì khoảng 29% trong số đấy bị đột quỵ, và cao hơn so với tỉ lệ người bị đột quỵ trong nhóm không mắc bệnh cao huyết áp (khoảng 3%).

2.2.5 Thuộc tính "heart_disease":

- Lập bảng thống kê tỉ lệ người mắc bệnh tim:
  - Code:

```
df.data $ heart_disease <- data$heart_disease
```

```
df.heart_disease <- df.data %>% group_by(heart_disease) %>%
  summarise(count = n()) %>%
  mutate(count = round(count * 100 / sum(count), digits = 2))

df2 <- df.heart_disease %>%
  mutate(csum = rev(cumsum(rev(df.heart_disease$count))),
         pos = df.heart_disease$count / 2 + lead(csum, 1),
         pos = if_else(is.na(pos), df.heart_disease$count/2, pos))


ggplot(df.heart_disease, aes(x = "" , y = count,
       fill = fct_inorder(heart_disease))) +
  geom_col(width = 1, color = 1) +
  coord_polar(theta = "y") +
  scale_fill_brewer(palette = "Pastel1") +
  geom_label_repel(data = df2,
                   aes(y = pos, label = paste0(count, "%")),
                   size = 4.5, nudge_x = 1, show.legend = FALSE) +
  guides(fill = guide_legend(title = "HeartDisease")) +
  theme_void()

ggsave(filename = "heart_disease_pie.png", device = "png",width = 19,
       height = 10,units="cm")
```

– Kết quả:

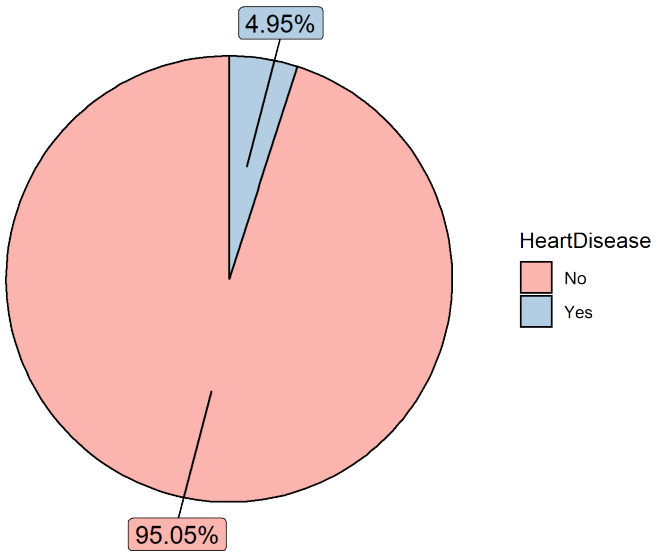


Figure 8: Biểu đồ thể hiện tỉ lệ của người mắc bệnh tim.

- Lập bảng thống kê tần số của 2 nhóm mắc bệnh và không mắc bệnh tim đối với đột quỵ:

  – Code:

```
df.HeartDiseaseVsStroke <- df.data %>% group_by(heart_disease, stroke) %>%
  summarise(count = n())
```

```
df.HeartDiseaseVsStroke

ggplot(df.HeartDiseaseVsStroke, aes(x = heart_disease, y = count,
       fill = stroke)) +
  geom_col(position = "dodge", color = 1) +
  coord_flip() + xlab("HeartDisease") + ylab("Frequency") +
  scale_fill_discrete(name = "Stroke", labels = c("No","Yes")) +
  theme_minimal()

ggsave(filename = "heart_disease_bar.png", device = "png",width = 19,
height = 10,units="cm")
```

– Kết quả:

```
> df.HeartDiseaseVsStroke
# A tibble: 4 × 3
# Groups:  heart_disease [2]
  heart_disease stroke count
  <fct>         <fct>  <int>
1 No            No      4496
2 No            Yes      169
3 Yes           No       203
4 Yes           Yes       40
```

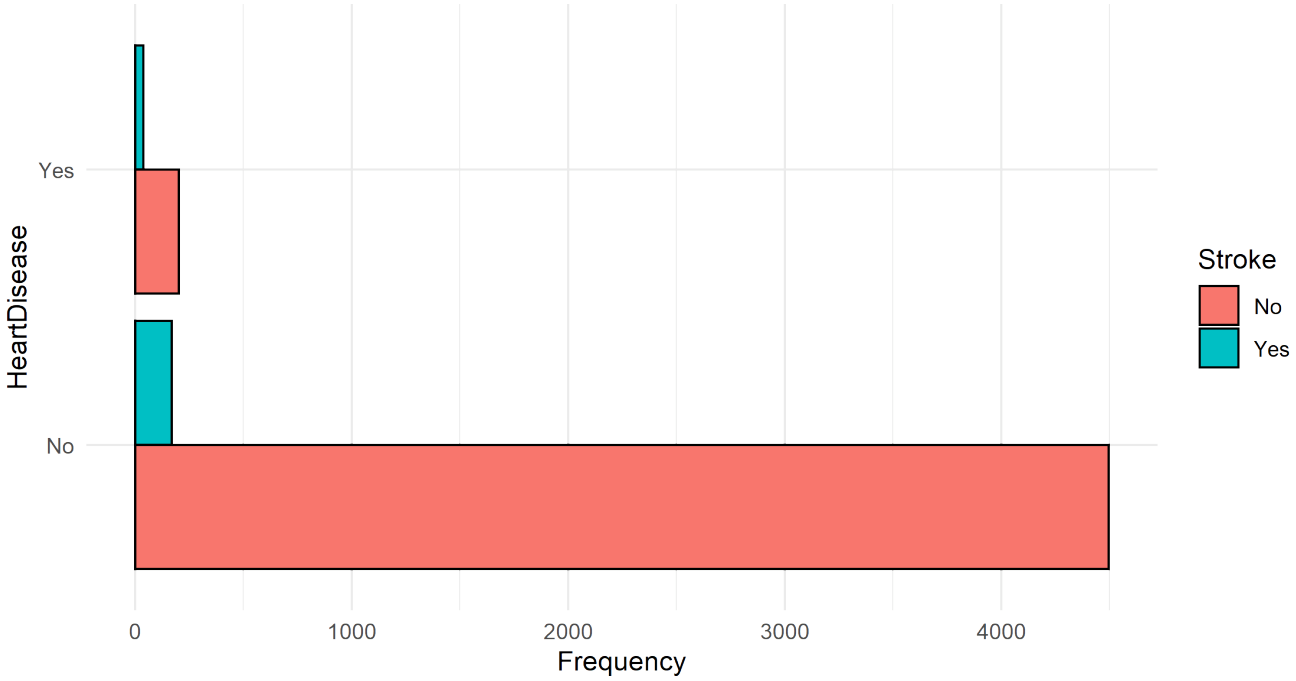


Figure 9: Biểu đồ thể hiện tần số đột quỵ của hai nhóm mắc và không mắc bệnh tim.

- Nhận xét:
  - Số người khảo sát mắc bệnh tim chiếm số lượng không lớn trong tập dữ liệu (khoảng 5%).

– Trong số những người mắc bệnh tim thì khoảng 16% trong số đấy bị đột quỵ, và cao hơn so với tỉ lệ người bị đột quỵ trong nhóm không mắc bệnh tim (khoảng 4%).

2.2.6 Thuộc tính "ever_married":

- Lập bảng thống kê tỉ lệ người đã từng kết hôn:
  - Code:

```
df.data $ married <- data$ever_married
df.married <- df.data %>% group_by(married) %>%
  summarise(count = n()) %>%
  mutate(count = round(count * 100 / sum(count), digits = 2))

df2 <- df.married %>%
  mutate(csum = rev(cumsum(rev(df.married$count))),
         pos = df.married$count / 2 + lead(csum, 1),
         pos = if_else(is.na(pos), df.married$count/2, pos))


ggplot(df.married, aes(x = "" , y = count, fill = fct_inorder(married))) +
  geom_col(width = 1, color = 1) +
  coord_polar(theta = "y") +
  scale_fill_brewer(palette = "Pastel1") +
  geom_label_repel(data = df2,
                   aes(y = pos, label = paste0(count, "%")),
                   size = 4.5, nudge_x = 1, show.legend = FALSE) +
  guides(fill = guide_legend(title = "married")) +
  theme_void()

ggsave(filename = "ever_married_pie.png", device = "png",width = 19,
       height = 10,units="cm")
```

  - Kết quả:

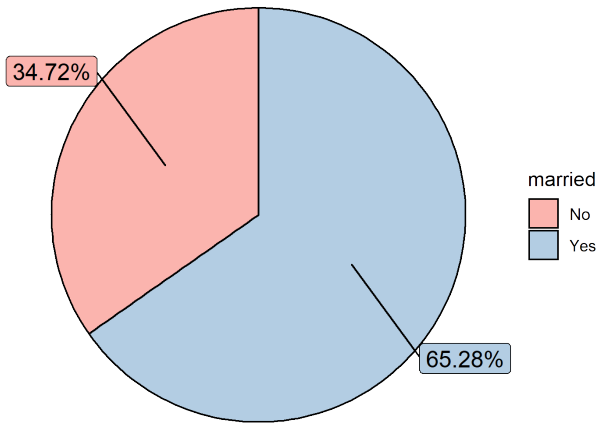


Figure 10: Biểu đồ thể hiện tỉ lệ của người mắc bệnh tim.

- Lập bảng thống kê tần số của 2 nhóm đã từng và chưa từng kết hôn đối với đột quỵ:
  - Code:

```
df.marriedVsStroke <- df.data %>% group_by(married, stroke) %>%
  summarise(count = n())

df.marriedVsStroke

ggplot(df.marriedVsStroke, aes(x = married, y = count, fill = stroke)) +
  geom_col(position = "dodge", color = 1) +
  coord_flip() + xlab("Ever married") + ylab("Frequency") +
  scale_fill_discrete(name = "Stroke", labels = c("No","Yes")) +
  theme_minimal()

ggsave(filename = "ever_married_bar.png", device = "png",width = 19,
    height = 10,units="cm")
```

  - Kết quả:

```
> df.marriedVsStroke
# A tibble: 4 × 3
# Groups:  married [2]
  married stroke count
  <chr>   <fct>  <int>
1 No      No      1681
2 No      Yes       23
3 Yes     No      3018
4 Yes     Yes      186
```

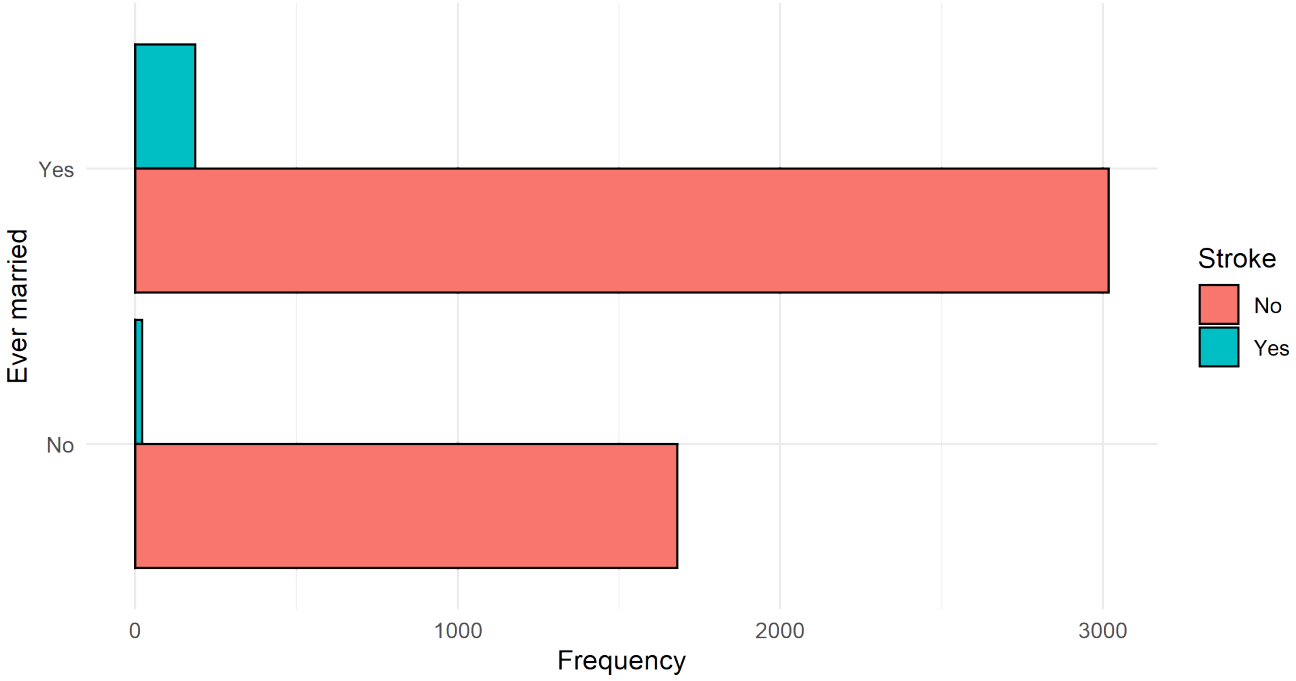


Figure 11: Biểu đồ thể hiện tần số đột quỵ của hai nhóm đã từng và chưa từng kết hôn

- Nhận xét:
  - Số người khảo sát đã từng kết hôn chiếm số lượng lớn trong tập dữ liệu (khoảng 65%).
  - Tỉ lệ những người đã từng kết hôn có xu hướng bị đột quỵ cao hơn so với trung bình (6% so với 4%), và đặc biệt cao hơn so khi so với nhóm người chưa từng kết hôn (6% so với 1%).

2.2.7 Thuộc tính "work_type":

- Lập bảng thống kê tỉ lệ việc làm:
  - Code:

```
df.data$work_type <- data$work_type

df.work_type <- df.data %>% group_by(work_type) %>%
  summarise(count = n()) %>%
mutate(count = round(count * 100 / sum(count), digits = 2))

df2 <- df.work_type %>%
  mutate(csum = rev(cumsum(rev(df.work_type$count))),
         pos = df.work_type$count / 2 + lead(csum, 1),
         pos = if_else(is.na(pos), df.work_type$count/2, pos))


ggplot(df.work_type, aes(x = "" , y = count,
       fill = fct_inorder(work_type))) +
  geom_col(width = 1, color = 1) +
  coord_polar(theta = "y") +
  scale_fill_brewer(palette = "Pastel1") +
  geom_label_repel(data = df2,
                   aes(y = pos, label = paste0(count, "%")),
                   size = 4.5, nudge_x = 1, show.legend = FALSE) +
  guides(fill = guide_legend(title = "Work type")) +
  theme_void()

ggsave(filename = "work_type_pie.png", device = "png",width = 19,
       height = 10,units="cm")
```

– Kết quả:

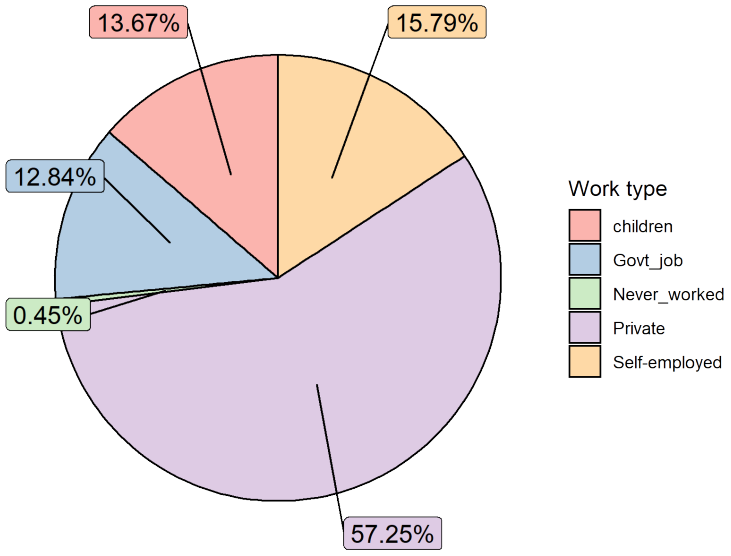


Figure 12: Biểu đồ thể hiện tỉ lệ việc làm.

- Lập bảng thống kê tần số của 5 nhóm việc làm về đột quỵ:
  – Code:

```
df.WorktypeVsStroke <- df.data %>% group_by(work_type, stroke) %>% summarise(count = n(

df.WorktypeVsStroke

ggplot(df.WorktypeVsStroke, aes(x = work_type, y = count,
      fill = stroke)) +
  geom_bar(stat = "identity", color = 1) +
  xlab("Work type") + ylab("Count") +
  scale_fill_discrete(name = "Stroke", labels = c("No", "Yes")) +
  theme_minimal()

ggsave(filename = "work_type_bar.png", device = "png",width = 19,
      height = 10,units="cm")
```

  – Kết quả:

```
> df.WorktypeVsStroke
# A tibble: 9 × 3
# Groups:  work_type [5]
  work_type    stroke count
  <fct>        <fct>  <int>
1 children     No       670
2 children     Yes        1
3 Govt_job     No       602
4 Govt_job     Yes       28
5 Never_worked No        22
6 Private      No      2683
```

```
7 Private      Yes     127
8 Self-employed No     722
9 Self-employed Yes     53
```

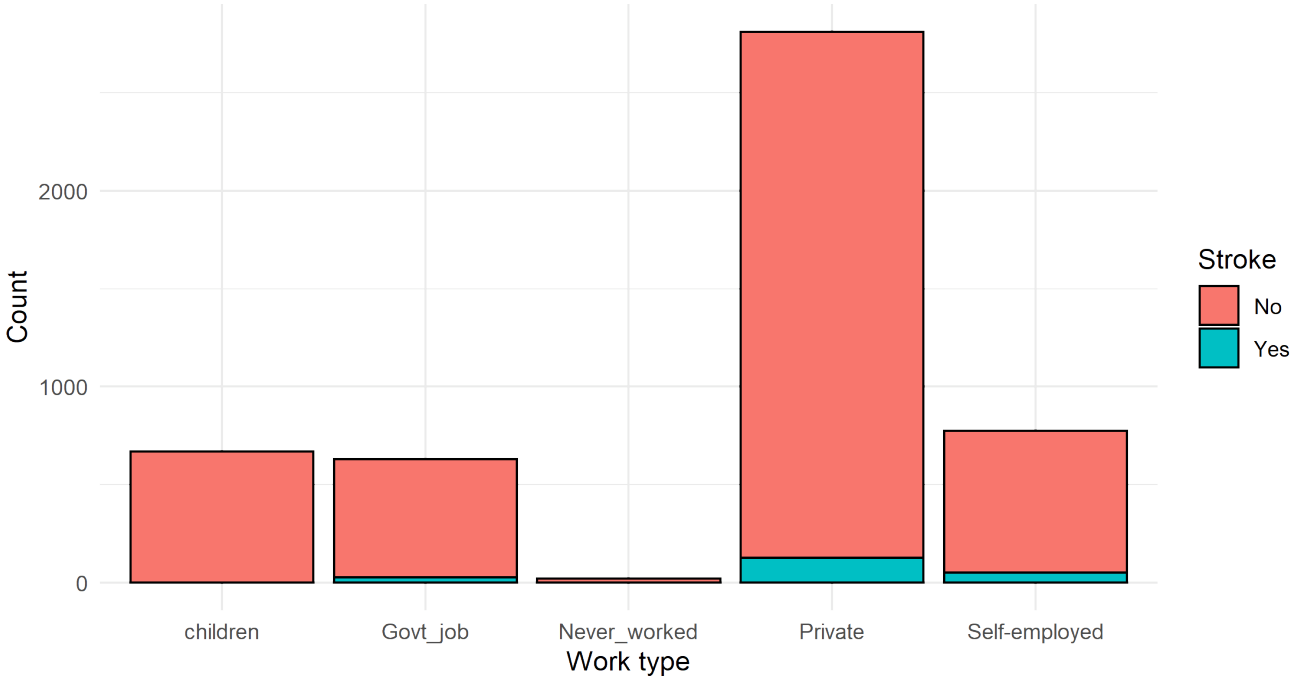


Figure 13: Biểu đồ thể hiện tần số đột quỵ của các nhóm việc làm.

- Nhận xét:
  - Số người khảo sát có việc làm là "Private" chiếm số lượng lớn trong tập dữ liệu (khoảng 57%).
  - Tỉ lệ những người có việc làm là "Self-employed" có xu hướng bị đột quỵ cao hơn so với trung bình (6.75% so với 4%).

2.2.8 Thuộc tính "Residence_type":

- Lập bảng thống kê tỉ lệ nơi cư trú:
  - Code:

```
df.data $ Residence_type <- data$Residence_type

df.residence <- df.data %>% group_by(Residence_type) %>%
  summarise(count = n()) %>%
  mutate(count = round(count * 100 / sum(count), digits = 2))

df2 <- df.residence %>%
  mutate(csum = rev(cumsum(rev(df.residence$count))),
         pos = df.residence$count / 2 + lead(csum, 1),
         pos = if_else(is.na(pos), df.residence$count/2, pos))
```

```
ggplot(df.residence, aes(x = "" , y = count,
       fill = fct_inorder(Residence_type))) +
  geom_col(width = 1, color = 1) +
  coord_polar(theta = "y") +
  scale_fill_brewer(palette = "Pastel1") +
  geom_label_repel(data = df2,
                   aes(y = pos, label = paste0(count, "%")),
                   size = 4.5, nudge_x = 1, show.legend = FALSE) +
  guides(fill = guide_legend(title = "Residence type")) +
  theme_void()

ggsave(filename = "residence_type_pie.png", device = "png",width = 19,
       height = 10,units="cm")
```

– Kết quả:

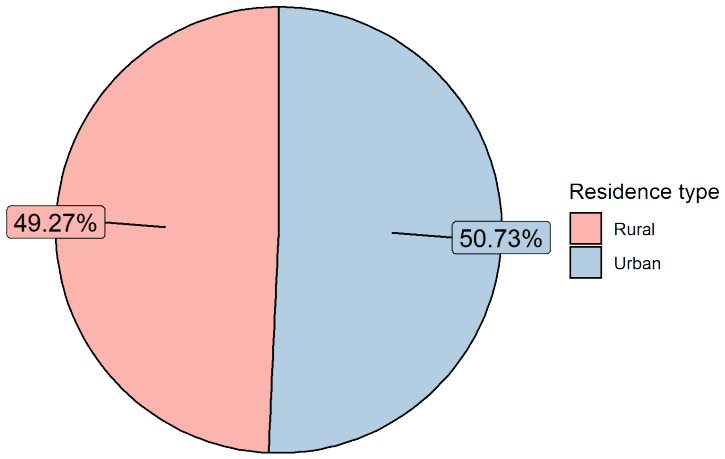


Figure 14: Biểu đồ thể hiện tỉ lệ nơi cư trú.

- Lập bảng thống kê tần số của 2 nhóm nơi cư trú về đột quỵ:

  – Code:

```
df.ResidenceVsStroke <- df.data %>% group_by(Residence_type, stroke) %>%
  summarise(count = n())

df.ResidenceVsStroke

ggplot(df.ResidenceVsStroke, aes(x = Residence_type, y = count,
       fill = stroke)) +
  geom_col(position = "dodge", color = 1) +
  coord_flip() + xlab("Residence type") + ylab("Frequency") +
  scale_fill_discrete(name = "Stroke", labels = c("No","Yes")) +
  theme_minimal()
```

```
ggsave(filename = "residence_type_bar.png", device = "png",width = 19,
       height = 10,units="cm")
```

– Kết quả:

```
> df.ResidenceVsStroke
# A tibble: 4 × 3
# Groups:  Residence_type [2]
  Residence_type stroke count
  <fct>          <fct>  <int>
1 Rural          No      2318
2 Rural          Yes      100
3 Urban          No      2381
4 Urban          Yes      109
```

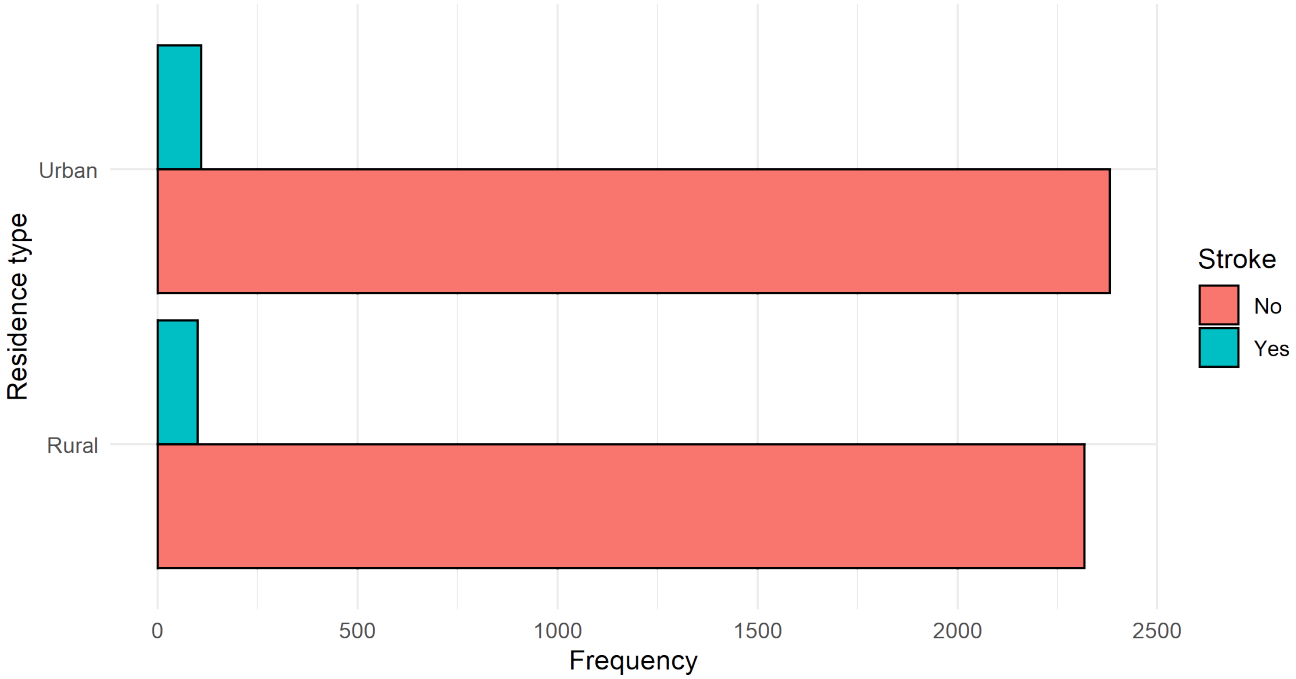


Figure 15: Biểu đồ thể hiện tần số đột quỵ của hai nhóm nơi cư trú

- Nhận xét:
  – Số người khảo sát đến từ hai nơi cư trú gần như là tương đương với nhau. Đồng thời tỉ lệ đột quỵ của cả 2 nơi đều như nhau.

2.2.9 Thuộc tính "avg_glucose_level":

- Vẽ biểu đồ boxplot thể hiện mức độ đường huyết của hai nhóm đối tượng bị đột quỵ và không bị đột quỵ:
  – Code:

```
df.data$avg_glucose_level <- data$avg_glucose_level
```

```
ggplot(df.data, aes(x = avg_glucose_level, y = stroke, fill = stroke)) +
  geom_boxplot() +
  xlab("Average glucose level") +
  ylab("Target") +
  theme(legend.position = "none")

ggsave(filename = "glucose_boxplot.png", device = "png",width = 19,
       height = 10,units="cm")
```

– Kết quả:

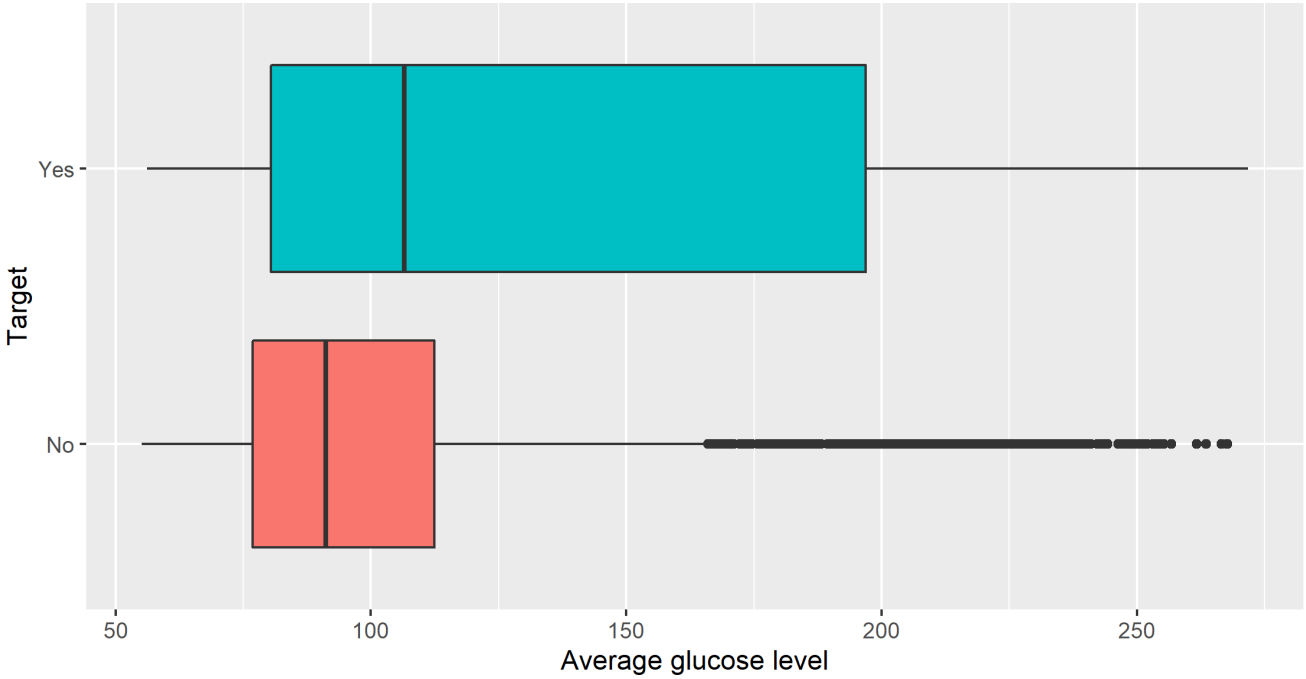


Figure 16: Boxplot mức độ đường huyết của hai nhóm đối tượng bị đột quỵ và không bị đột quỵ.

- Vẽ đồ thị histogram mức độ đường huyết của hai nhóm đối tượng:
  – Code:

```
ggplot(df.data, aes(x = avg_glucose_level, fill = stroke)) +
  geom_histogram(aes(y = ..density..)) +
  geom_density(alpha = .2) +
  facet_grid(stroke ~ .) +
  xlab("Average glucose level") +
  ylab("Density") +
  scale_fill_discrete(name = "Stroke", labels = c("No","Yes")) +
  theme(legend.position = "none")

ggsave(filename = "glucose_histogram.png", device = "png",width = 19,
       height = 10,units="cm")
```

  – Kết quả:

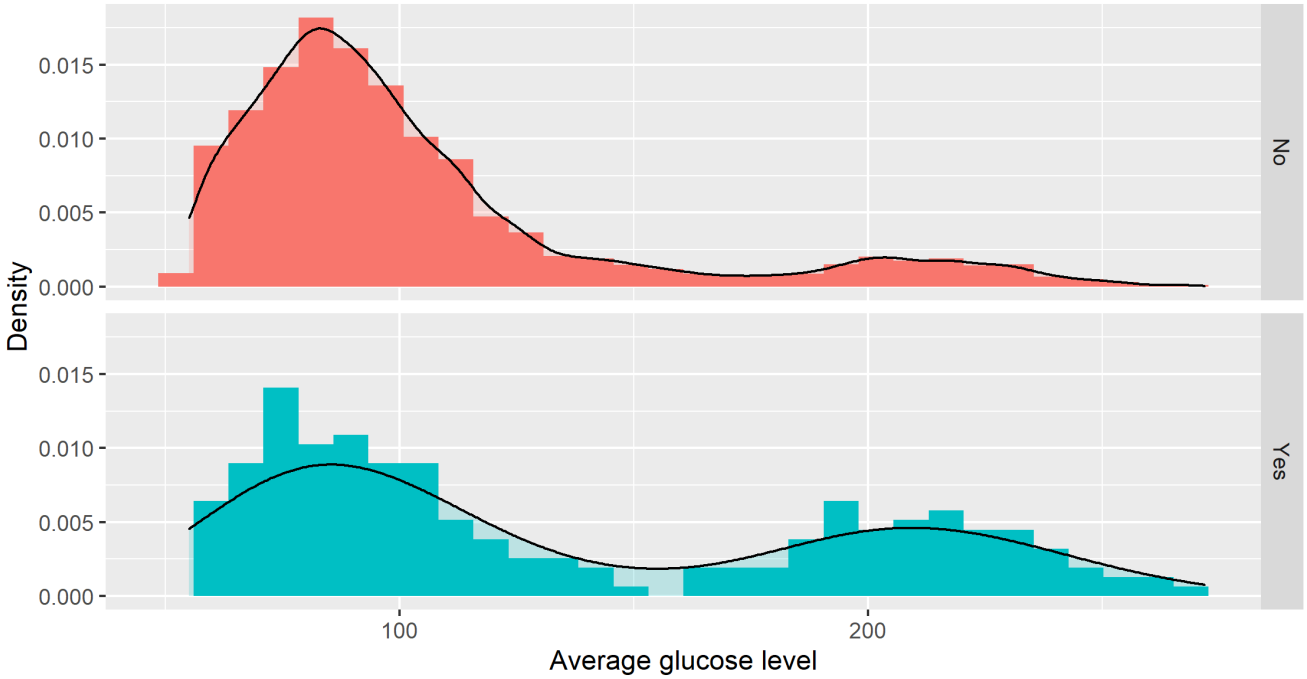


Figure 17: Histogram mức độ đường huyết của hai nhóm đối tượng.

- Định tính hoá rồi lập bảng thống kê số lượng đối tượng theo mức độ đường huyết:
  - Chia thành 3 nhóm mức độ: $\begin{cases} \text{Bình thường: dưới 140 mg/dl,} \\ \text{Cao: trong khoảng 140 đến 200 mg/dl,} \\ \text{Quá cao: trên 200 mg/dl.} \end{cases}$
  - Code:

```
## Dinh tinh hoa

df.glucose <- 1:length(data$avg_glucose_level)

for (i in 1:length(data$avg_glucose_level)) {
  if (avg_glucose_level[i] < 140) {
    df.glucose[i] = "Binh thuong"
  }
  else if (avg_glucose_level[i] >= 140 && avg_glucose_level[i] <= 200) {
    df.glucose[i] = "Cao"
  }
  else df.glucose[i] = "Qua cao"
}

## Ve bar chart

df.data$glucose <- df.glucose

df.GluVsStroke <- df.data %>% group_by(glucose, stroke) %>%
  summarise(count = n())
```

```
df.GluVsStroke

ggplot(df.GluVsStroke, aes(x = glucose, y = count, fill = stroke)) +
  geom_col(position = "dodge", color = 1) +
  xlab("Glucose") + ylab("Frequency") +
  scale_fill_discrete(name = "Stroke", labels = c("No","Yes")) +
  theme_minimal()

ggsave(filename = "glucose_bar.png", device = "png",width = 19,
       height = 10,units="cm")
```

– Kết quả:

```
> df.GluVsStroke
# A tibble: 6 × 3
# Groups:  glucose [3]
  glucose    stroke count
  <chr>      <fct>  <int>
1 Binh thuong No    4025
2 Binh thuong Yes    129
3 Cao         No     332
4 Cao         Yes     31
5 Qua cao     No     342
6 Qua cao     Yes     49
```

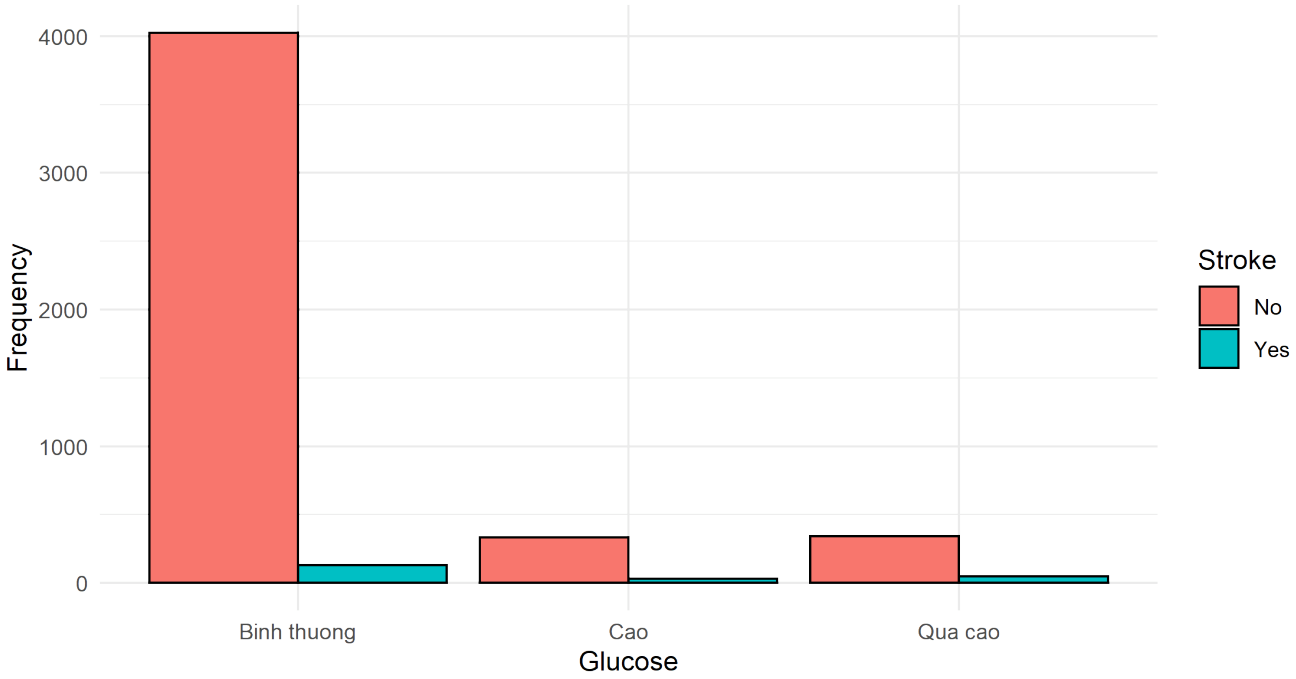


Figure 18: Biểu đồ thống kê số lượng đối tượng theo mức độ đường huyết.

- Nhận xét:
  – Từ biểu đồ boxplot, ta có thể thấy mức đường huyết trung bình của nhóm đối tượng đột quỵ cao hơn so với nhóm không bị đột quỵ.

– Những người có mức đường huyết từ 140 mg/dl trở lên có xu hướng bị đột quỵ cao hơn so với những người có mức đường huyết $<$ 140 mg/dl.

2.2.10 Thuộc tính "bmi":

- Vẽ biểu đồ boxplot thể hiện chỉ số khối cơ thể của hai nhóm đối tượng bị đột quỵ và không bị đột quỵ:
  - Code:

```
df.data$bmi <- data$bmi

ggplot(df.data, aes(x = bmi, y = stroke, fill = stroke)) +
  geom_boxplot() +
  xlab("BMI") +
  ylab("Target") +
  theme(legend.position = "none")

ggsave(filename = "bmi_boxplot.png", device = "png",width = 19,
       height = 10,units="cm")
```

  - Kết quả:

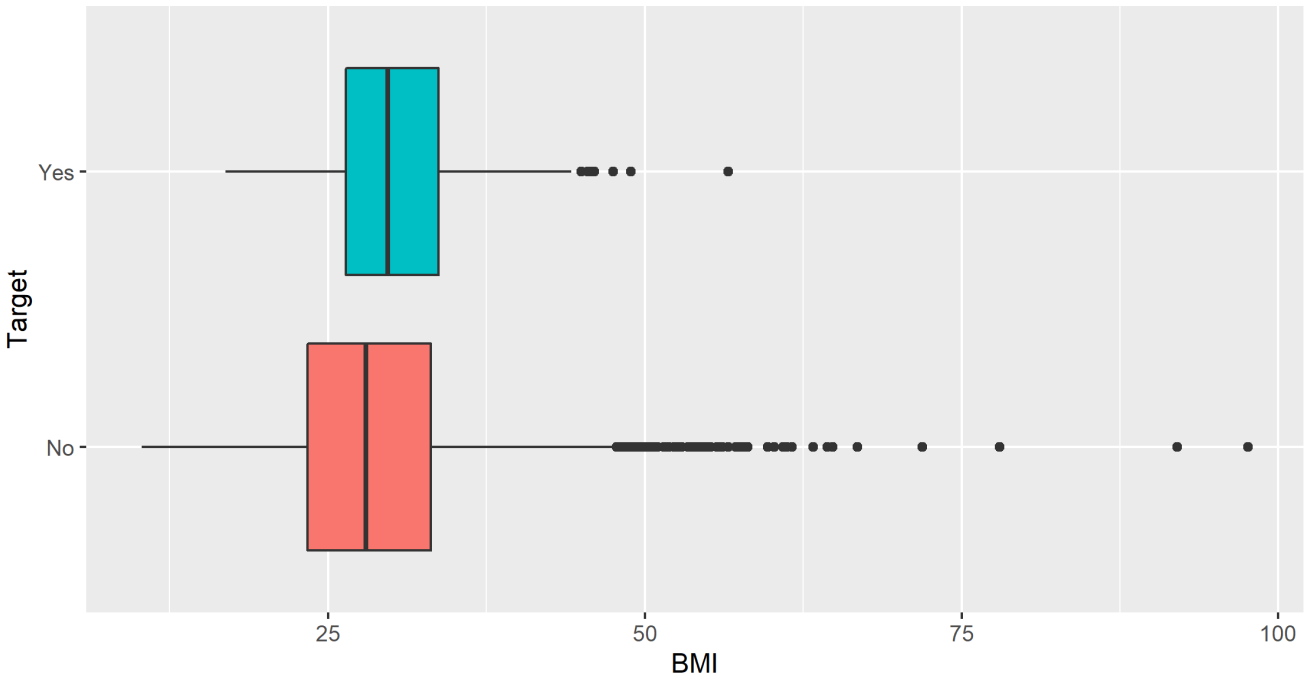


Figure 19: Boxplot chỉ số khối cơ thể của hai nhóm đối tượng bị đột quỵ và không bị đột quỵ.

- Vẽ đồ thị histogram chỉ số khối cơ thể của hai nhóm đối tượng:
  - Code:

```
ggplot(df.data, aes(x = bmi, fill = stroke)) +
  geom_histogram(aes(y = ..density..)) +
  geom_density(alpha = .2) +
```

```
  facet_grid(stroke ~ .) +
  xlab("BMI") +
  ylab("Density") +
  scale_fill_discrete(name = "Stroke", labels = c("No","Yes")) +
  theme(legend.position = "none")

ggsave(filename = "bmi_histogram.png", device = "png",width = 19,
       height = 10,units="cm")
```

– Kết quả:

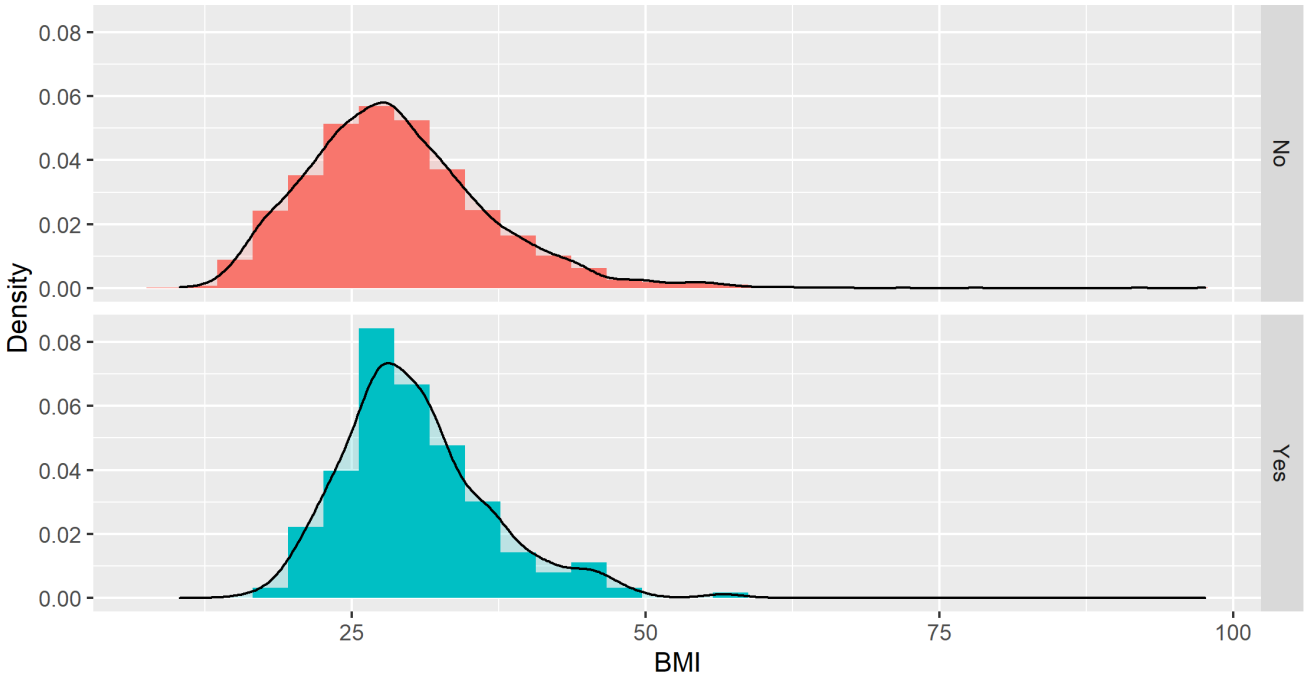


Figure 20: Histogram chỉ số khối cơ thể của hai nhóm đối tượng.

- Định tính hoá rồi lập bảng thống kê số lượng đối tượng theo chỉ số khối cơ thể:

– Chia thành 4 nhóm: $\begin{cases} \text{Gầy: dưới } 18.50 \text{ kg/m}^2 \\ \text{Bình thường: từ } 18.50 \text{ đến } 24.99 \text{ kg/m}^2, \\ \text{Thừa cân: từ } 25.00 \text{ đến } 30.00 \text{ kg/m}^2, \\ \text{Béo phì: trên } 30.00 \text{ kg/m}^2. \end{cases}$

– Code:

```
## Dinh tinh hoa

df.bmi <- 1:length(data$bmi)

for (i in 1:length(data$bmi)) {
  if (bmi[i] < 18.5) {
    df.bmi[i] = "Gay"
  }
  else if (bmi[i] >= 18.5 && bmi[i] <= 24.99) {
```

```
    df.bmi[i] = "Binh thuong"
  }
  else if (bmi[i] >= 25 && bmi[i] <= 30) {
    df.bmi[i] = "Thua can"
  }
  else df.bmi[i] = "Beo phi"
}

## Bar chart

df.data$df.bmi <- df.bmi

df.BmiVsStroke <- df.data %>% group_by(df.bmi, stroke) %>%
  summarise(count = n())

df.BmiVsStroke

ggplot(df.BmiVsStroke, aes(x = df.bmi, y = count, fill = stroke)) +
  geom_col(position = "dodge", color = 1) +
  xlab("BMI") + ylab("Frequency") +
  scale_fill_discrete(name = "Stroke", labels = c("No","Yes")) +
  theme_minimal()

ggsave(filename = "bmi_bar.png", device = "png",width = 19,
       height = 10,units="cm")
```

– Kết quả:

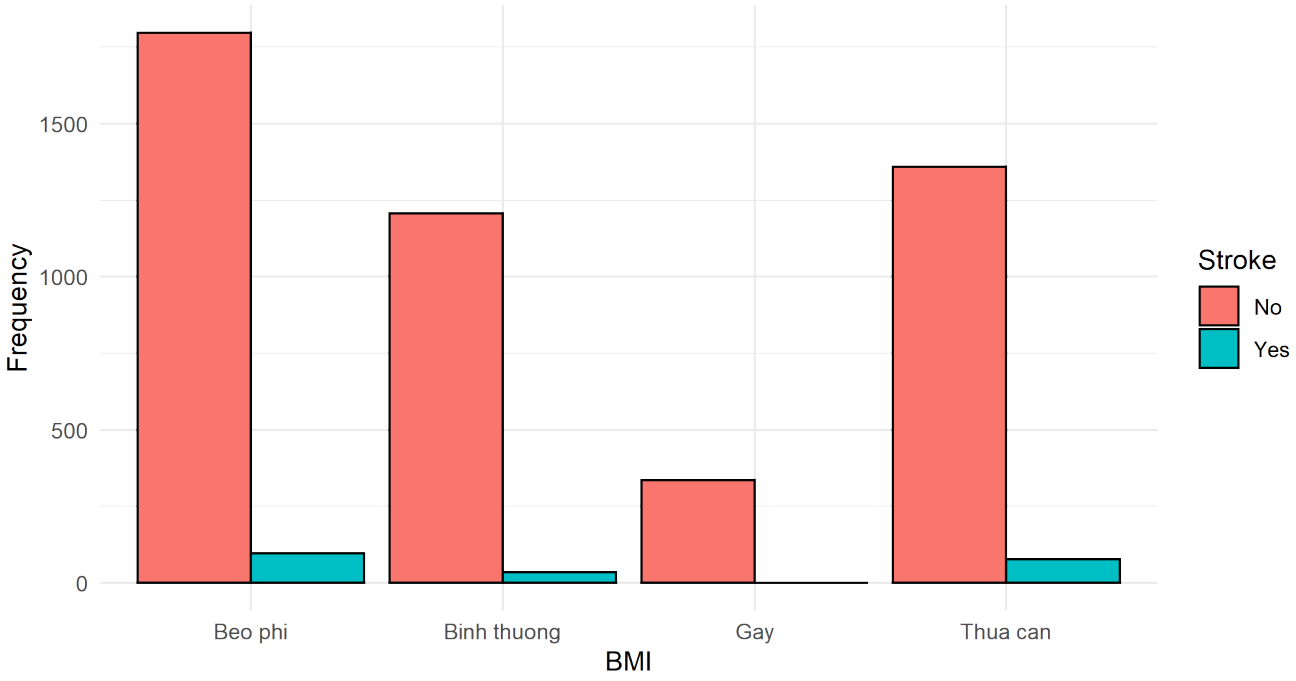


Figure 21: Biểu đồ thống kê số lượng đối tượng theo chỉ số khối lượng cơ thể.

```
> df.BmiVsStroke
# A tibble: 8 × 3
# Groups:  df.bmi [4]
  df.bmi     stroke count
  <chr>      <fct>  <int>
1 Beo phi    No      1797
2 Beo phi    Yes       96
3 Binh thuong No     1207
4 Binh thuong Yes      35
5 Gay        No       336
6 Gay        Yes        1
7 Thua can   No      1359
8 Thua can   Yes       77
```

- Nhận xét:
  - Từ biểu đồ boxplot, ta có thể thấy bmi trung bình của nhóm đối tượng đột quỵ cao hơn so với nhóm không bị đột quỵ.
  - Những người có bmi từ 25 kg/m^2 trở lên có xu hướng bị đột quỵ cao hơn so với những người có mức đường huyết $<$ 25 kg/m^2.

2.2.11 Thuộc tính "smoking_status":

- Lập bảng thống kê tỉ lệ tình trạng hút thuốc:
  - Code:

```
df.data$smoking_status <- data$smoking_status

df.smoking_status <- df.data %>% group_by(smoking_status) %>%
  summarise(count = n()) %>%
  mutate(count = round(count * 100 / sum(count), digits = 2))

df2 <- df.smoking_status %>%
  mutate(csum = rev(cumsum(rev(df.smoking_status$count))),
         pos = df.smoking_status$count / 2 + lead(csum, 1),
         pos = if_else(is.na(pos), df.smoking_status$count/2, pos))


ggplot(df.smoking_status, aes(x = "" , y = count,
       fill = fct_inorder(smoking_status))) +
  geom_col(width = 1, color = 1) +
  coord_polar(theta = "y") +
  scale_fill_brewer(palette = "Pastel1") +
  geom_label_repel(data = df2,
                   aes(y = pos, label = paste0(count, "%")),
                   size = 4.5, nudge_x = 1, show.legend = FALSE) +
  guides(fill = guide_legend(title = "Smoking status")) +
  theme_void()

ggsave(filename = "smoking_status_pie.png", device = "png",width = 19,
       height = 10,units="cm")
```

– Kết quả:

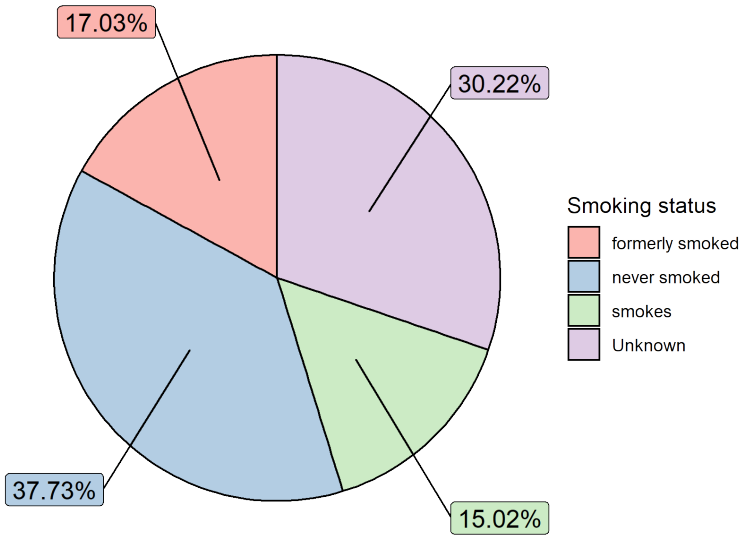


Figure 22: Biểu đồ thể hiện tỉ lệ của tình trạng hút thuốc.

- Lập bảng thống kê tần số của 4 nhóm tình trạng hút thuốc đối với đột quỵ:

  – Code:

```
df.SmokingVsStroke <- df.data %>% group_by(smoking_status, stroke) %>% summarise(count

df.SmokingVsStroke

ggplot(df.SmokingVsStroke, aes(x = smoking_status, y = count,
       fill = stroke)) +
  geom_bar(stat = "identity", color = 1) +
  xlab("Smoking status") + ylab("Count") +
  scale_fill_discrete(name = "Stroke", labels = c("No", "Yes")) +
  theme_minimal()

ggsave(filename = "smoking_status_bar.png", device = "png",width = 19,
       height = 10,units="cm")
```

  – Kết quả:

```
    > df.SmokingVsStroke
# A tibble: 8 × 3
# Groups:  smoking_status [4]
  smoking_status stroke count
  <fct>          <fct>  <int>
1 formerly smoked No       779
2 formerly smoked Yes       57
3 never smoked   No       1768
4 never smoked   Yes        84
5 smokes         No        698
6 smokes         Yes        39
```

```
7 Unknown        No      1454
8 Unknown        Yes       29
```

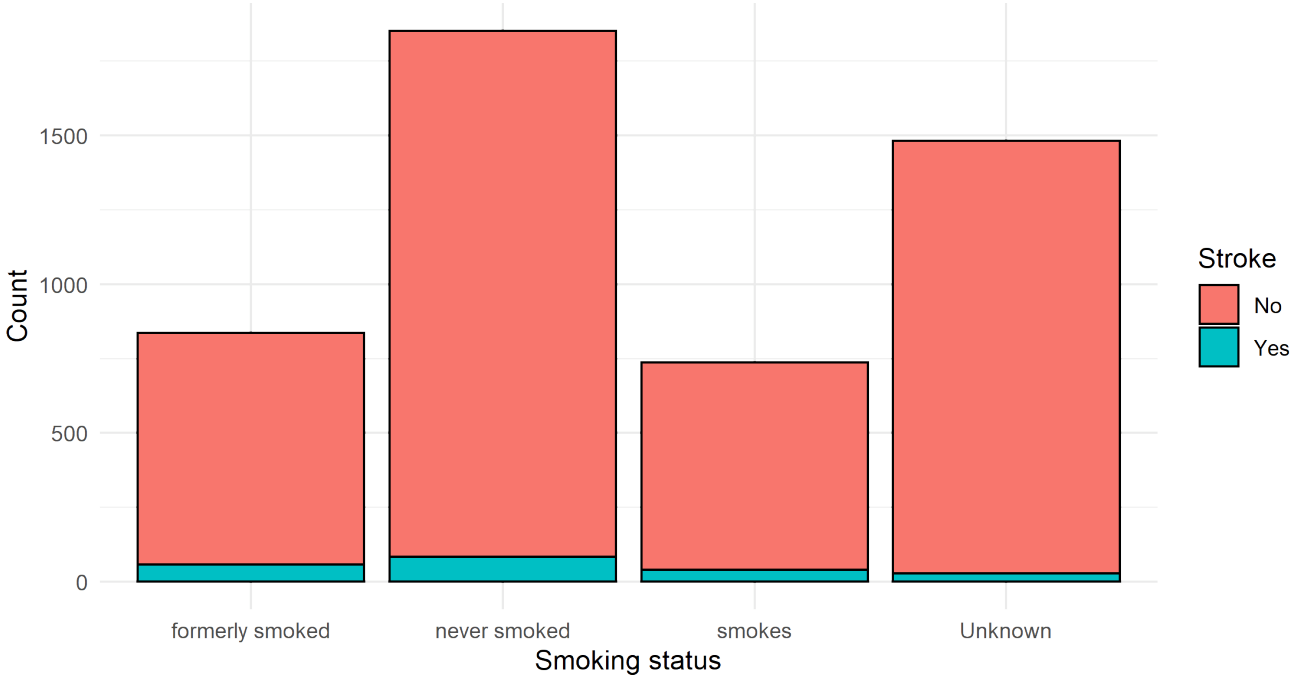


Figure 23: Biểu đồ thể hiện tần số đột quỵ theo bốn nhóm tình trạng hút thuốc.

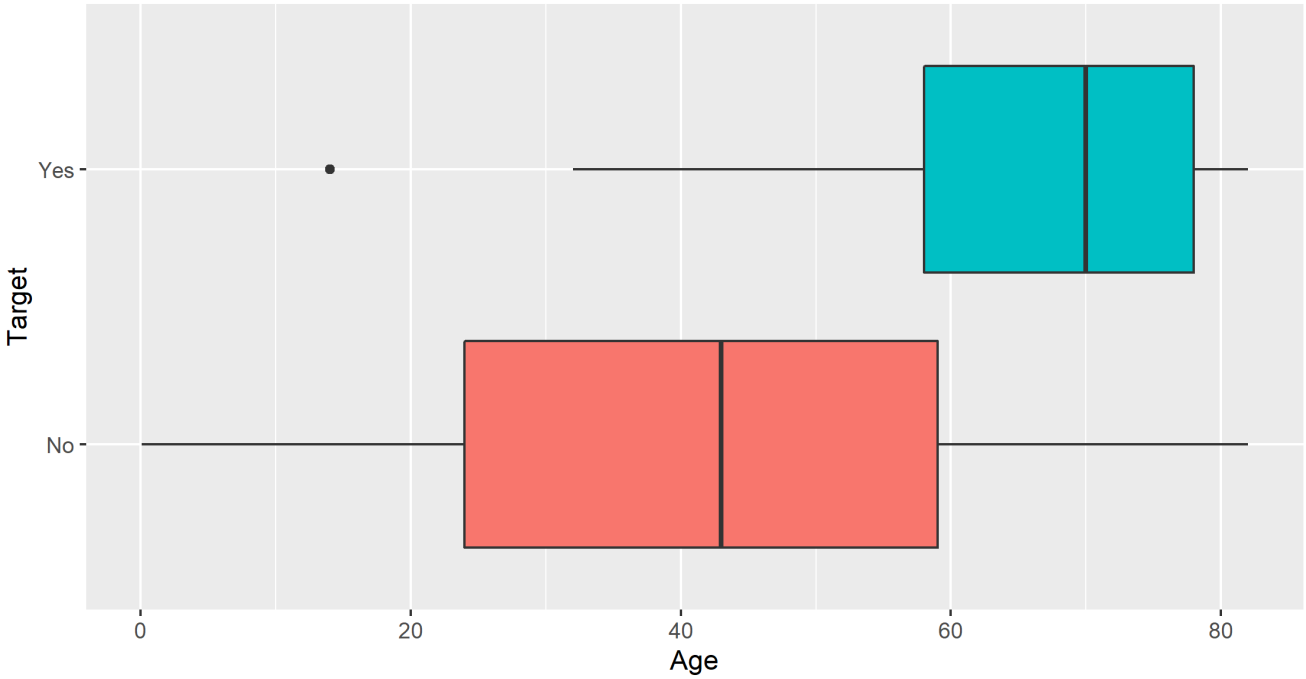
- Nhận xét:
  - Số người khảo sát không hút thuốc hoặc không có thông tin nào về tình trạng hút thuốc chiếm số lượng chủ yếu trong tập dữ liệu.
  - Tỉ lệ đột quỵ ở hai nhóm "formerly smoked" và "smokes" chiếm tỉ lệ cao hơn so với hai nhóm còn lại, mặc dù số lượng được khảo sát thấp gần như gấp đôi so với hai nhóm đấy.

3 Những suy diễn trên bộ dữ liệu

Trong phần này, dựa vào những thống kê và nhận xét trong phần mô tả, ta đặt ra một số bài toán thống kê thể hiện mối quan hệ giữa các thuộc tính và việc đột quỵ.

3.1 Đối với thuộc tính "age":

Bài toán 1: Tuổi trung bình của những người bị đột quỵ cao hơn so với những trường hợp khác.



- Code:

```
t.test(age~df.data$stroke, alternative = "less")
```

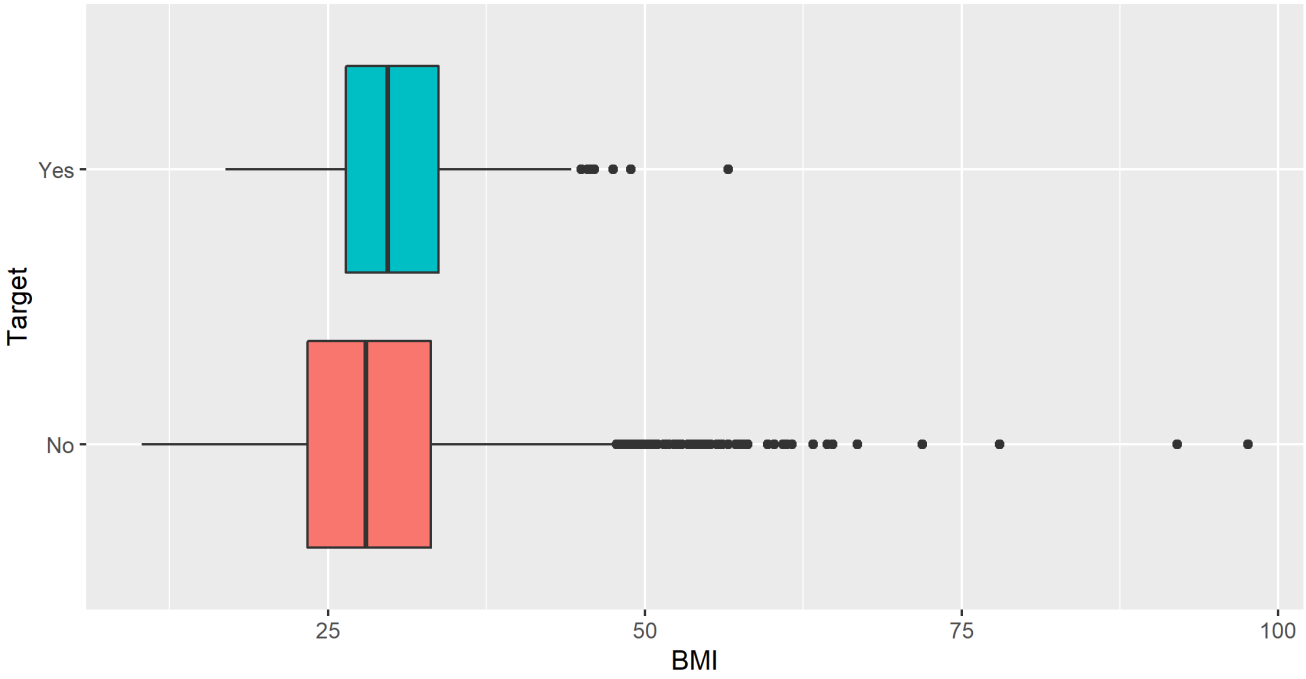
- Kết quả:

```
> t.test(age~df.data$stroke, alternative = "less")

   Welch Two Sample t-test

data:  age by df.data$stroke
t = -28.286, df = 271.68, p-value < 2.2e-16
alternative hypothesis: true difference in means between group No and group Yes
is less than 0
95 percent confidence interval:
     -Inf -24.43501
sample estimates:
 mean in group No mean in group Yes
        41.76381         67.71292
```

- **Nhận xét:** Với giá trị p-value rất nhỏ như kết quả ở trên đã đưa ra, chúng ta sẽ chấp nhận giả thuyết là "Tuổi trung bình của những người bị đột quỵ cao hơn so với những trường hợp khác".

3.2 Đối với thuộc tính "bmi":

Bài toán 2: Chỉ số khối lượng trung bình của những người bị đột quỵ cao hơn so với những trường hợp khác.



- Code:

```
t.test(bmi~df.data$stroke, alternative = "less")
```

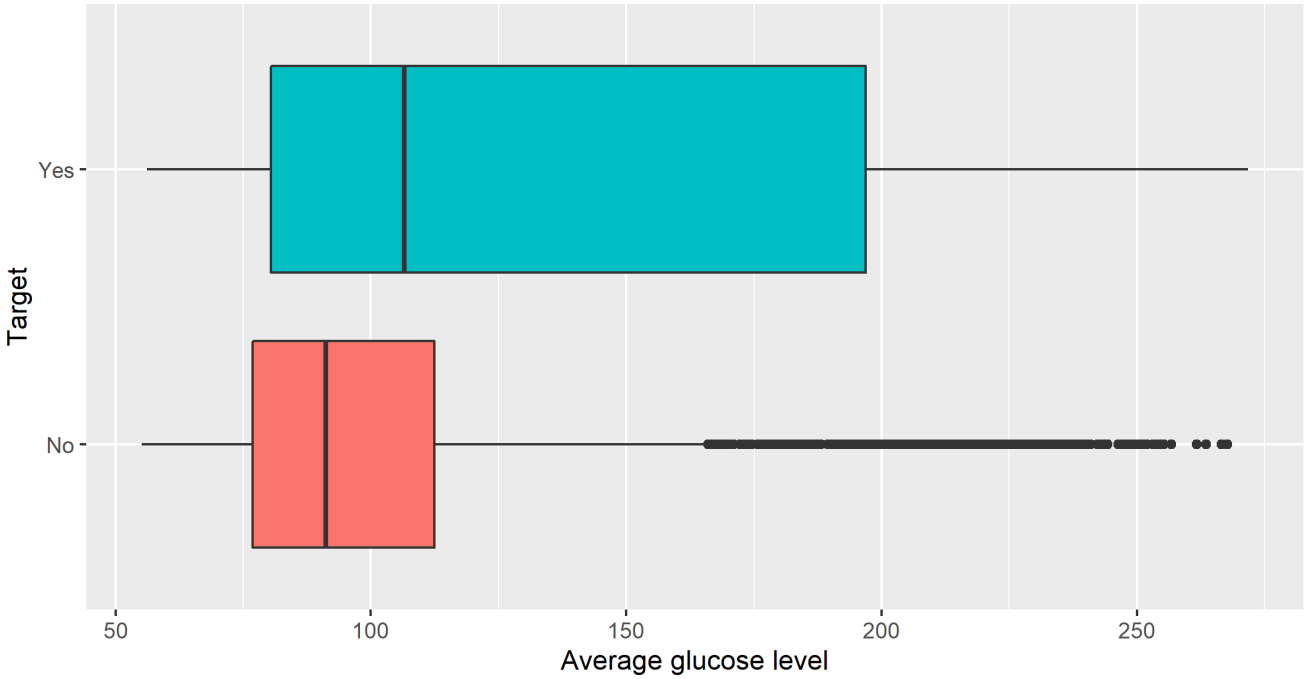
- Kết quả:

```
> t.test(bmi~df.data$stroke, alternative = "less")

   Welch Two Sample t-test

data:  bmi by df.data$stroke
t = -3.6374, df = 237.84, p-value = 0.0001689
alternative hypothesis: true difference in means between group No and group Yes
is less than 0
95 percent confidence interval:
     -Inf -0.89922
sample estimates:
 mean in group No mean in group Yes
         28.82443          30.47129
```

- **Nhận xét:** Với giá trị p-value rất nhỏ như kết quả ở trên đã đưa ra, chúng ta sẽ chấp nhận giả thuyết là "Chỉ số khối lượng trung bình của những người bị đột quỵ cao hơn so với những trường hợp khác".

3.3 Đối với thuộc tính "avg_glucose_level":

Bài toán 3: Đường huyết trung bình của những người bị đột quỵ cao hơn so với những trường hợp khác.



- Code:

```
t.test(avg_glucose_level~df.data$stroke, alternative = "less")
```

- Kết quả:

```
> t.test(avg_glucose_level~df.data$stroke, alternative = "less")

   Welch Two Sample t-test

data:  avg_glucose_level by df.data$stroke
t = -7.0034, df = 216.86, p-value = 1.547e-11
alternative hypothesis: true difference in means between group No and group Yes
is less than 0
95 percent confidence interval:
     -Inf -23.36397
sample estimates:
 mean in group No mean in group Yes
        103.9954         134.5714
```

- **Nhận xét:** Với giá trị p-value rất nhỏ như kết quả ở trên đã đưa ra, chúng ta sẽ chấp nhận giả thuyết là "Đường huyết trung bình của những người bị đột quỵ cao hơn so với những trường hợp khác".

3.4 Đối với thuộc tính "heart_disease":

Bài toán 4: Tỉ lệ đột quỵ ở những người có bệnh tim cao hơn so với những người không mắc bệnh tim.

- Lập bảng thống kê tần số:

- Code:

```
heart_diseaseVSstroke <- table(df.data$heart_disease, df.data$stroke)
heart_diseaseVSstroke
```

- Kết quả:

```
> heart_diseaseVSstroke
              stroke
heart_disease  No  Yes
       No    4496  169
       Yes    203   40
```

- Kiểm định giả thuyết:
  - Code:

```
prop.test(heart_diseaseVSstroke, correct = FALSE, alternative = "greater")
```

  - Kết quả:

```
> prop.test(heart_diseaseVSstroke, correct = FALSE, alternative = "greater")

   2-sample test for equality of proportions without continuity correction

data: heart_diseaseVSstroke
X-squared = 93.372, df = 1, p-value < 2.2e-16
alternative hypothesis: greater
95 percent confidence interval:
 0.08899519 1.00000000
sample estimates:
   prop 1    prop 2
0.9637728 0.8353909
```

- Nhận xét:
  - Có thể thấy p-value rất nhỏ từ kết quả nhận được, chúng ta sẽ chấp nhận giả thuyết từ kiểm định trên là "Tỉ lệ không bị đột quỵ của những người không mắc bệnh tim cao hơn người mắc bệnh tim", hoặc có thể hiểu rằng "Tỉ lệ đột quỵ ở những người có bệnh tim cao hơn so với những người không mắc bệnh tim".

3.5 Đối với thuộc tính "hypertension":

Bài toán 5: Tỉ lệ đột quỵ ở những người bị cao huyết áp cao hơn so với những trường hợp còn lại.

- Lập bảng thống kê tần số:
  - Code:

```
hypertensionVSstroke <- table(df.data$hypertension, df.data$stroke)
hypertensionVSstroke
```

- Kết quả:

```
> hypertensionVSstroke
            stroke
hypertension No  Yes
       No  4308  149
       Yes  391   60
```

- Kiểm định giả thuyết:
  - Code:

```
prop.test(hypertensionVSstroke, correct = FALSE, alternative = "greater")
```

  - Kết quả:

```
> prop.test(hypertensionVSstroke, correct = FALSE, alternative = "greater")

   2-sample test for equality of proportions without continuity correction

data: hypertensionVSstroke
X-squared = 99.667, df = 1, p-value < 2.2e-16
alternative hypothesis: greater
95 percent confidence interval:
 0.07293261 1.00000000
sample estimates:
   prop 1    prop 2
0.9665694 0.8669623
```

- Nhận xét:
  - Có thể thấy p-value rất nhỏ từ kết quả nhận được, chúng ta sẽ chấp nhận giả thuyết từ kiểm định trên là "Tỉ lệ không bị đột quỵ của những người không bị cao huyết áp cao hơn người bị cao huyết áp", hoặc có thể hiểu rằng "Tỉ lệ đột quỵ ở những người bị cao huyết áp cao hơn so với những trường hợp còn lại".

3.6 Mối quan hệ giữa hai thuộc tính "ever_married" và "heart_disease"

Bài toán 6: Khảo sát mối liên hệ giữa hai thuộc tính "ever_married" và "heart_disease"

- Code:

```
chisq.test(df.data $ married, df.data $ heart_disease, correct = FALSE)
```

- Kết quả:

```
> chisq.test(df.data $ married, df.data $ heart_disease, correct = FALSE)

 Pearson's Chi-squared test
```

```
data:  df.data$married and df.data$heart_disease
X-squared = 60.693, df = 1, p-value = 6.67e-15
```

- **Nhận xét:** Có thể thấy rằng p-value rất nhỏ, điều này chỉ ra hai thuộc tính "ever_married" và "heart_disease" có mối quan hệ với nhau.

4 Hồi quy

4.1 Các mô hình hồi quy đơn

4.1.1 Đồ thị phân tán

- Đường huyết trung bình (avg_glucose_level) và tuổi (age):
  - Code:

```
ggplot(data, aes(age, avg_glucose_level, color = age, size = stroke)) +
  geom_point(alpha = 0.7) +
  xlab("Tuoi") +
  ylab("Duong huyet trung binh")

ggsave(filename = "agevsglu.png", device = "png",width = 19,height = 10,
units="cm")
```

  - Kết quả:

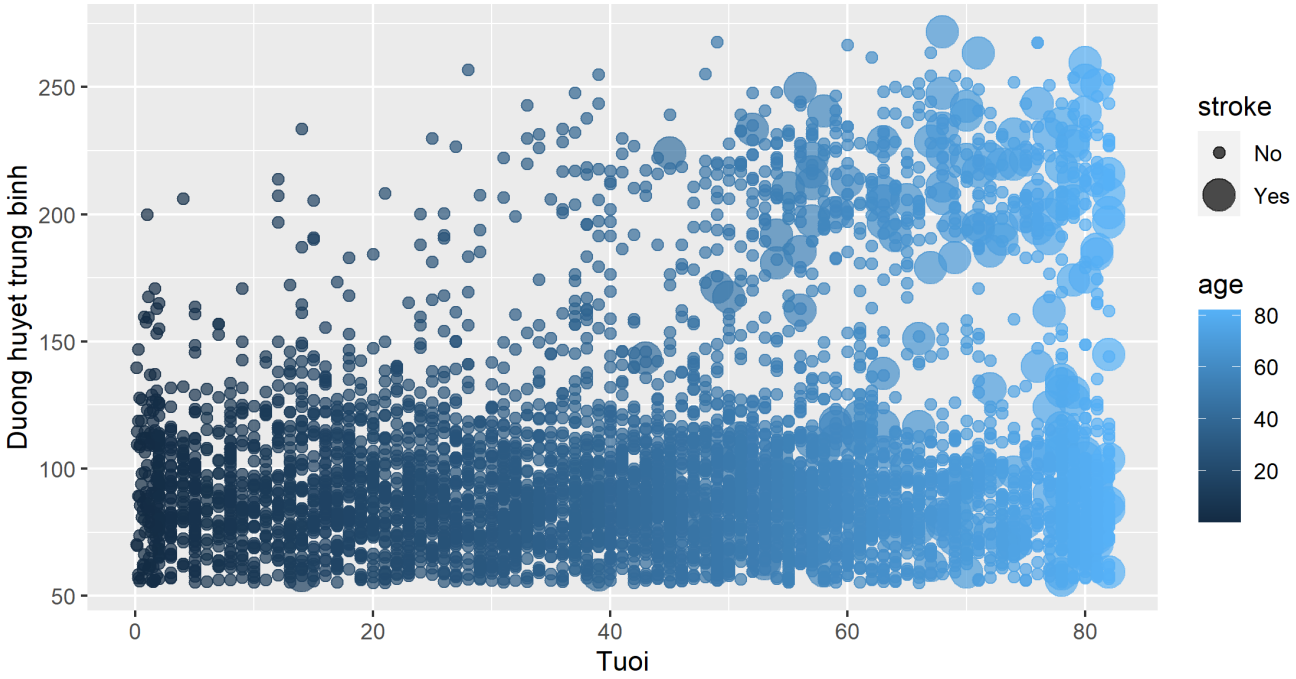


Figure 24: Biểu đồ phân tán của đường huyết trung bình và tuổi.

- Đường huyết trung bình và bmi:
  - Code:

```
ggplot(data, aes(bmi, avg_glucose_level, color = avg_glucose_level,
    size = stroke)) +
  geom_point(alpha = 0.7) +
  xlab("BMI") +
  ylab("Duong huyet trung binh")

ggsave(filename = "bmivsglu.png", device = "png",width = 19,height = 10,
units="cm")
```

– Kết quả:

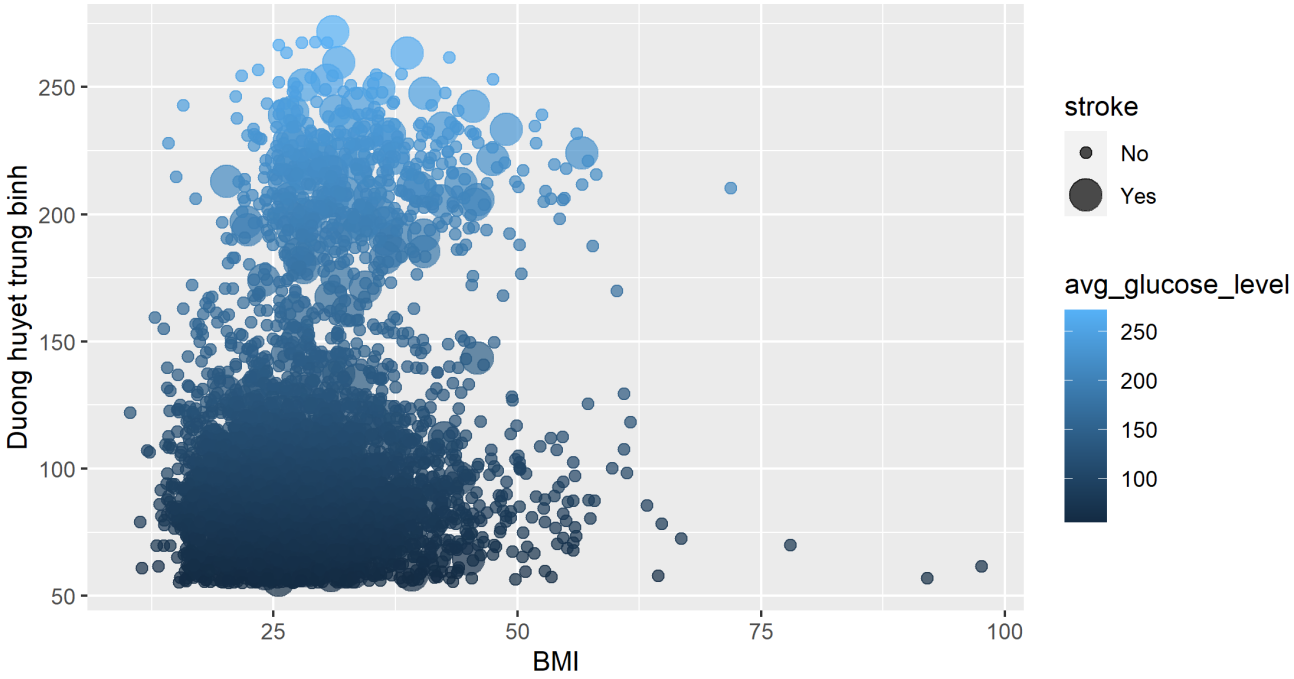


Figure 25: Biểu đồ phân tán của đường huyết trung bình và bmi.

- **Nhận xét:**
  - Mật độ những người đột quỵ ở mức đường huyết từ 150 mg/dl trở lên khá cao.
  - Từ biểu đồ phân tán của độ tuổi và đường huyết trung bình ở trên, ta có thể thấy rằng mật độ phân bố của mức đường huyết trung bình cao (150 - 250 mg/dl) khá cao ở độ tuổi > 50.
  - Từ biểu đồ phân tán của đường huyết trung bình và bmi, ta có thể thấy rằng dường như không có mối quan hệ nào giữa hai đại lượng này.

Sau đây ta sẽ lập mô hình hồi quy để tính đường huyết trung bình dựa trên các thuộc tính "age" và "bmi".

4.1.2 Mô hình hồi quy đơn tính đường huyết trung bình từ tuổi:

- Phương trình mô hình hồi quy:

$$avg\_glucose\_level = \beta_1 + \beta_2 age + \varepsilon$$

- Code:

```
model1 <- lm(avg_glucose_level~age)
model1
```

- Kết quả:

```
> model1

Call:
lm(formula = avg_glucose_level ~ age)

Coefficients:
(Intercept)          age
    85.3714       0.4648
```

- Từ kết quả ở trên, ta nhận được phương trình mô hình hồi quy:

$$avg\_glucose\_level = 85.3714 + 0.4648 \times age + \varepsilon$$

Với $\begin{cases} \beta_1 = 85.3714, \text{ nghĩa là khi tuổi bằng 0 thì đường huyết trung bình là 85.3714,} \\ \beta_2 = 0.4648, \text{ nghĩa là khi đối tượng già thêm 1 tuổi thì đường huyết tăng 0.4648.} \end{cases}$

- Dùng hàm "confint" để ước lượng khoảng tin cậy cho hệ số hồi quy:
  - Code:

```
confint(model1)
```

  - Kết quả:

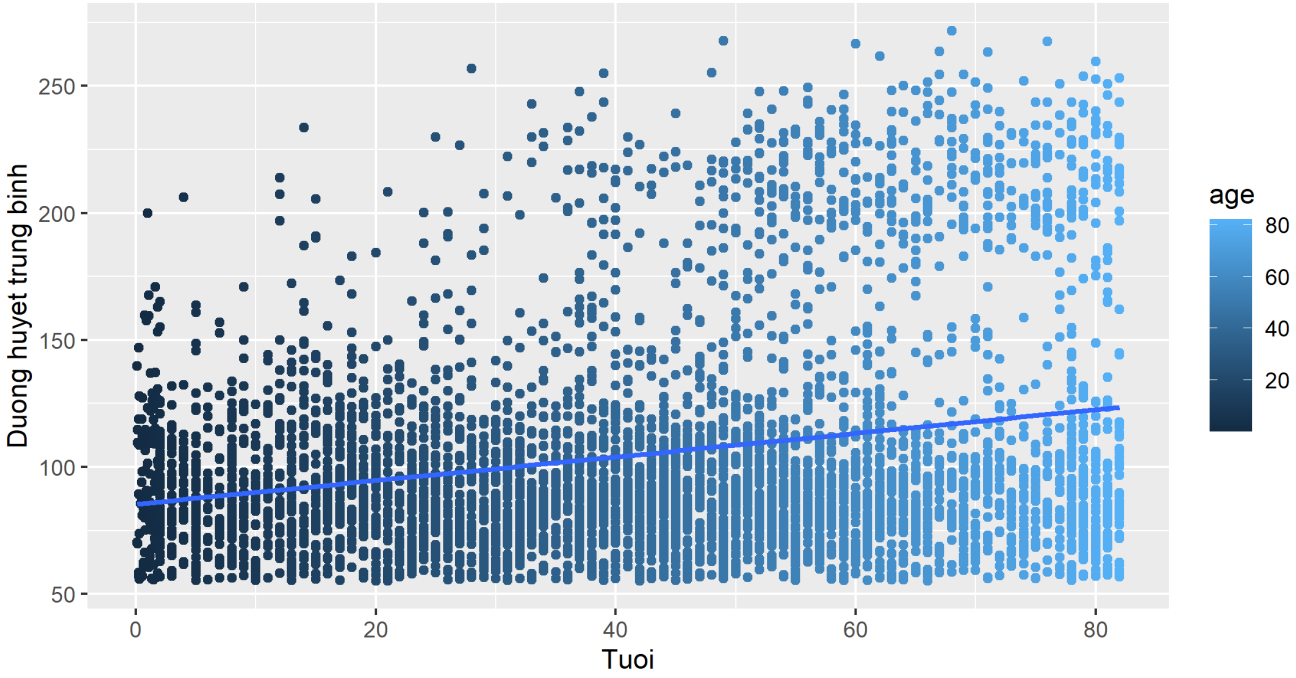
```
> confint(model1)
                2.5 %    97.5 %
(Intercept) 82.7764774 87.9662419
age         0.4112452 0.5183837
```

  - Ta có khoảng tin cậy với mức ý nghĩa 95% là: $\begin{cases} \beta_1 : (82.7764774,\ 87.9662419), \\ \beta_2 : (0.4112452,\ 0.5183837). \end{cases}$

- Biểu diễn phương trình hồi quy lên đồ thị phân tán của "age" và "avg_glucose_level":
  - Code:

```
ggplot(data, aes(age, avg_glucose_level, color = age)) +
  geom_point(alpha = 1) +
  geom_smooth(method = "lm", se = FALSE) +
  xlab("Tuoi") +
  ylab("Duong huyet trung binh")

ggsave(filename = "model1.png", device = "png",width = 19,height = 10,
units="cm")
```

– Kết quả:



- Kiểm tra những thông số khác:

```
summary(model1)
```

Kết quả:

```
> summary(model1)

Call:
lm(formula = avg_glucose_level ~ age)

Residuals:
   Min     1Q Median     3Q    Max
-66.74 -28.64 -11.97  12.93 159.61

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 85.37136  1.32362  64.50  <2e-16 ***
age          0.46481  0.02733  17.01  <2e-16 ***
---
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 43.18 on 4906 degrees of freedom
Multiple R-squared: 0.0557, Adjusted R-squared: 0.0555
F-statistic: 289.4 on 1 and 4906 DF, p-value: < 2.2e-16
```

- **Nhận xét:** Với p-value rất nhỏ thì chúng ta không thể nào phủ định mối liên hệ giữa đường huyết trung bình và tuổi, tuy nhiên mô hình này chưa hoàn toàn tốt bởi vì $R^2 = 0.0555$.

4.1.3 Mô hình hồi quy đơn tính đường huyết trung bình từ bmi:

- Phương trình mô hình hồi quy:

$$avg\_glucose\_level = \beta_1 + \beta_2 bmi + \varepsilon$$

- Code:

```
model2 <- lm(avg_glucose_level~bmi)
model2
```

- Kết quả:

```
> model2

Call:
lm(formula = avg_glucose_level ~ bmi)

Coefficients:
(Intercept)          bmi
    76.5868       0.9936
```

- Từ kết quả ở trên, ta nhận được phương trình mô hình hồi quy:

$$avg\_glucose\_level = 76.5868 + 0.9936 \times bmi + \varepsilon$$

Với $\begin{cases} \beta_1 = 76.5868, \text{ nghĩa là khi bmi bằng 0 thì đường huyết trung bình là 76.5868,} \\ \beta_2 = 0.9936, \text{ nghĩa là khi bmi tăng thêm 1 thì đường huyết tăng 0.9936.} \end{cases}$

- Dùng hàm "confint" để ước lượng khoảng tin cậy cho hệ số hồi quy:
  - Code:

```
confint(model2)
```

  - Kết quả:

```
> confint(model2)
                2.5 %    97.5 %
(Intercept) 71.9201937 81.253419
bmi          0.8377825 1.149484
```

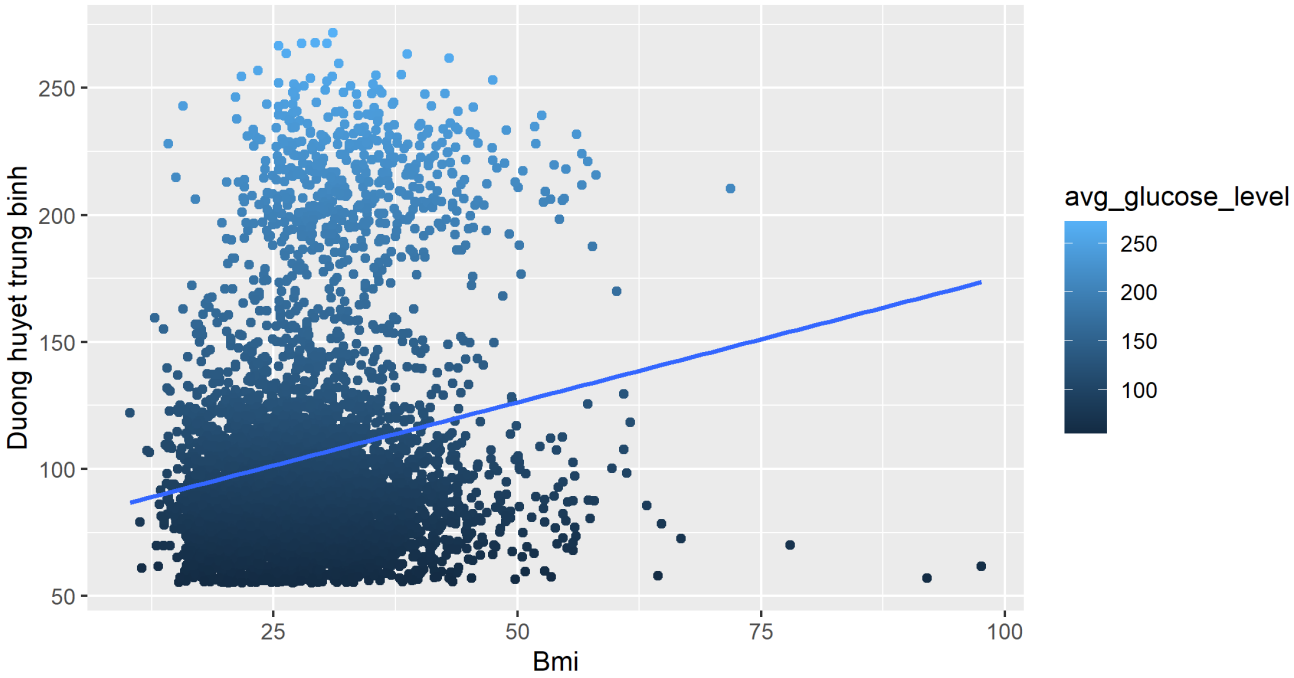
  - Ta có khoảng tin cậy với mức ý nghĩa 95% là: $\begin{cases} \beta_1 : (71.9201937,\ 81.253419), \\ \beta_2 : (0.8377825,\ 1.149484). \end{cases}$

- Biểu diễn phương trình hồi quy lên đồ thị phân tán của "bmi" và "avg_glucose_level":
  - Code:

```
ggplot(data, aes(bmi, avg_glucose_level, color = avg_glucose_level)) +
  geom_point(alpha = 1) +
  geom_smooth(method = "lm", se = FALSE) +
```

```
  xlab("Bmi") +
  ylab("Duong huyet trung binh")

ggsave(filename = "model2.png", device = "png",width = 19,height = 10,
units="cm")
```

– Kết quả:



- Kiểm tra những thông số khác:

```
summary(model2)
```

Kết quả:

```
> summary(model2)

Call:
lm(formula = avg_glucose_level ~ bmi)

Residuals:
    Min      1Q  Median      3Q     Max
-111.89 -28.62 -12.42   10.96  164.67

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 76.5868    2.3804   32.17   <2e-16 ***
bmi          0.9936    0.0795   12.50   <2e-16 ***
---
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1    ‘ ’ 1

Residual standard error: 43.74 on 4906 degrees of freedom
```

```
Multiple R-squared: 0.03086, Adjusted R-squared: 0.03066
F-statistic: 156.2 on 1 and 4906 DF, p-value: < 2.2e-16
```

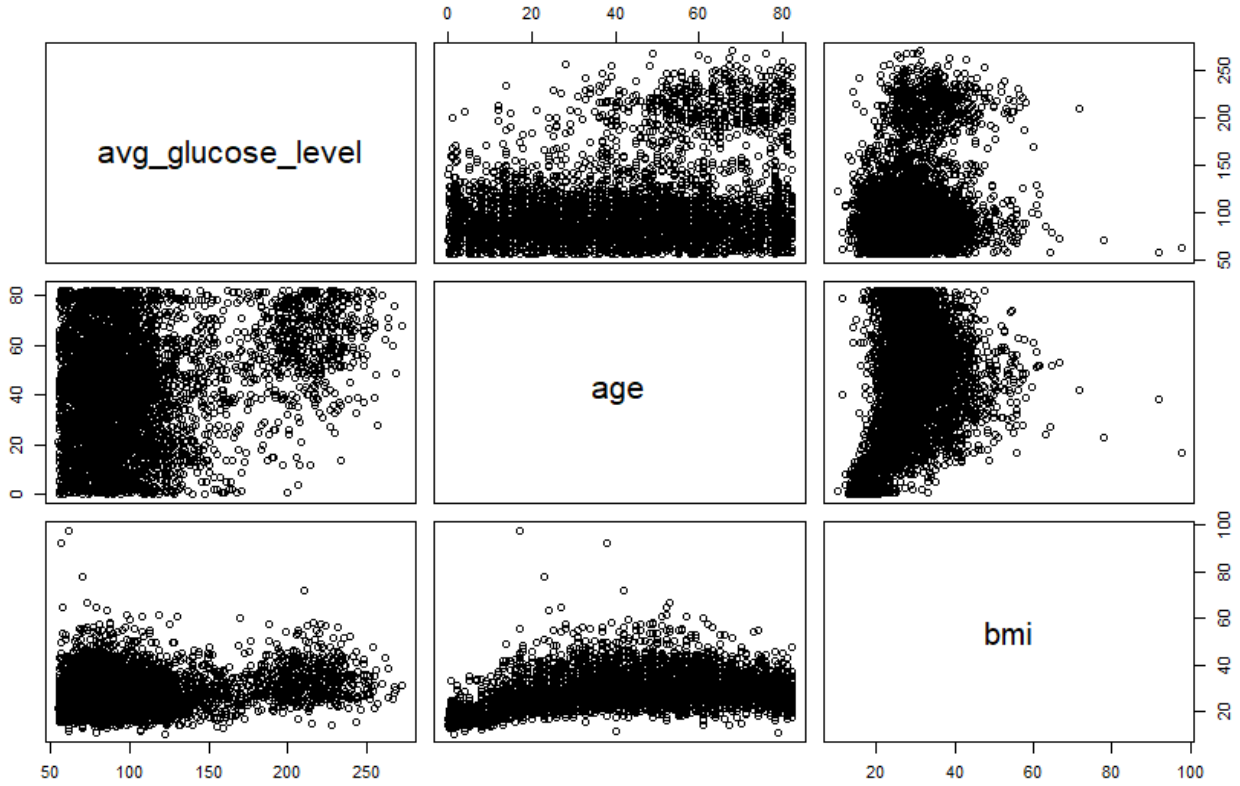
- **Nhận xét:** Với p-value rất nhỏ thì chúng ta không thể nào phủ định mối liên hệ giữa đường huyết trung bình và bmi, tuy nhiên mô hình này chưa hoàn toàn tốt bởi vì $R^2 = 0.03066$.

4.2 Mô hình hồi quy đa

- Kiểm tra sự phụ thuộc:
  - Code:

```
pairs(avg_glucose_level ~ age + bmi)
```

  - Kết quả:



  - **Nhận xét:** Mối quan hệ giữa đường huyết trung bình với các biến còn lại phần nào đã được thể hiện qua trong phần mô hình hồi quy đơn. Khi tuổi tăng hoặc bmi tăng thì đường huyết trung bình đều tăng nhẹ.

- Phương trình mô hình hồi quy:

$$avg\_glucose\_level = \beta_1 + \beta_2 age + \beta_3 bmi + \varepsilon$$

- Code:

```
model3 <- lm(avg_glucose_level ~ age + bmi)
model3
```

- Kết quả:

```
> model3

Call:
lm(formula = avg_glucose_level ~ age + bmi)

Coefficients:
(Intercept)         age         bmi
    70.6065      0.3932      0.6173
```

- Từ kết quả ở trên, ta nhận được phương trình mô hình hồi quy:

$$avg\_glucose\_level = 70.6065 + 0.3932 \times age + 0.6173 \times bmi + \varepsilon$$

Với $\begin{cases} \beta_1 = 70.6065, \text{ nghĩa là khi tuổi và bmi bằng 0 thì đường huyết trung bình là 70.6065,} \\ \beta_2 = 0.3932, \text{ nghĩa là khi tuổi tăng thêm 1 thì đường huyết tăng 0.3932,} \\ \beta_3 = 0.6173, \text{ nghĩa là khi bmi tăng thêm 1 thì đường huyết tăng 0.6173.} \end{cases}$

- Dùng hàm "confint" để ước lượng khoảng tin cậy cho hệ số hồi quy:
  - Code:

```
confint(model3)
```

  - Kết quả:

```
> confint(model3)
                2.5 %    97.5 %
(Intercept) 65.9455730 75.2674324
age         0.3366651 0.4496746
bmi         0.4550140 0.7795559
```

  - Ta có khoảng tin cậy với mức ý nghĩa 95% là: $\begin{cases} \beta_1 : (65.9455730,\ 75.2674324), \\ \beta_2 : (0.3366651,\ 0.4496746), \\ \beta_3 : (0.4550140,\ 0.7795559). \end{cases}$

- Kiểm tra những thông số khác:

```
summary(model3)
```

Kết quả:

```
> summary(model3)

Call:
lm(formula = avg_glucose_level ~ age + bmi)
```

```
Residuals:
   Min     1Q Median     3Q    Max
-85.44 -28.60 -11.84 13.46 160.68

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 70.60650  2.37748 29.698 < 2e-16 ***
age          0.39317  0.02882 13.641 < 2e-16 ***
bmi          0.61728  0.08277  7.458 1.04e-13 ***
---
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1    ‘ ’ 1

Residual standard error: 42.94 on 4905 degrees of freedom
Multiple R-squared: 0.06628, Adjusted R-squared: 0.0659
F-statistic: 174.1 on 2 and 4905 DF, p-value: < 2.2e-16
```

- **Nhận xét:** Như những mô hình đơn đã thực hiện, cả hai bmi và độ tuổi đều không thể bằng 0 được. Về phần R^2, tuy đã cải thiện hơn so với hai mô hình hồi quy đơn, nhưng nó vẫn còn rất nhỏ. Điều này cho thấy model này hiện tại chưa tốt.