Projekt końcowy

Nikola Jędrzejczyk

Zbiór danych:

Dane zostały pobrane ze strony Kaggle: https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset

```
dane1 <- read.csv("full_data.csv")
```

Zbiór ten zawiera informacje zebrane od osób, z różnych grup wiekowych. Pacjenci odpowiadali na pytania związane ze zdrowiem i prowadzonym trybem życia.

Celem badania było porównanie danych o osobach, które miały udar mózgu oraz tych, które go nie przeszły, aby skutecznie przewidywać możliwość zachorowania na wspomniany przed chwilą udar u osób, których wyniki są bardzo podobne do tych, którzy go przeszli.

Poza podstawowymi danymi takimi jak wiek i płeć, zebrano informacje między innymi na temat wskaźnika BMI, poziomu glukozy, rodzaju pracy, miejsca zamieszkania oraz czy pacjent choruje na nadciśnienie.

Opis zmiennych:

zmienna:	opis:
`gender:`	płeć (kobieta lub mężczyzna)
`age:`	wiek
`hypertension:`	nadciśnienie (tak lub nie)
`heart_disease:`	choroby serca (tak lub nie)
`ever_married:`	czy pacjent jest żonaty? (tak lub nie)
`work_type:`	rodzaj pracy (prywatna, własna działalność gospodarcza, praca rządowa, nie pracuje, bo to dziecko)
`residence_type:`	miejsce zamieszkania (miasto lub wieś)
`glucose_level:`	poziom glukozy
`bmi:`	wskaźnik masy ciała
`smoking_status:`	status palacza (nigdy nie palił, palił kiedyś, pali teraz, nieznany)
`stroke:`	czy pacjent miał udar mózgu? (0 - niemiał lub 1 - miał)

Hipotezy badawcze:

1) Osoby, które mają niepoprawne BMI lub palą papierosy częściej przechodzą udar mózgu.
2) Nieodpowiednie wyniki BMI i zły poziom glukozy we krwi zwiększają prawdopodobieństwo wystąpienia udaru mózgu.
3) Osoby mieszkające w mieście i będące w związku małżeńskim mają wyższe BMI.

Jako cel obrano sprawdzenie czy złe wyniki, wskaźniki lub warunki życia mogą ostrzegać przed zwiększonym ryzykiem groźnych chorób.

Wstępne czyszczenie i przygotowanie danych do pracy:

Pracę rozpoczynamy od zamiany dbl na int w przypadku wieku oraz angielskich nazw i oznaczeń: 0, 1 z oryginalnego pliku na polskie nazwy, między innymi dla wygody. Zdecydowano się również, aby zmienić status palenia z nieznanego na nigdy w przypadku dzieci poniżej 10 roku życia.

```
head(dane1, 1)
```

```
##   gender age hypertension heart_disease ever_married work_type Residence_type
## 1   Male  67            0             1          Yes   Private          Urban
##   glucose_level  bmi  smoking_status stroke
## 1        228.69 36.6 formerly smoked      1
```

```
dane1$gender <- factor(dane1$gender , levels = c(kobieta = "Female",
                mezczyzna = "Male"), labels = c("kobieta", "mezczyzna"))

dane1$hypertension <- factor(dane1$hypertension , levels = c(nie = 0,
                      tak = 1), labels = c("nie", "tak"))

dane1$heart_disease <- factor(dane1$heart_disease , levels = c(nie = 0,
                       tak = 1), labels = c("nie", "tak"))

dane1$ever_married <- factor(dane1$ever_married , levels = c(nie = "No",
                      tak = "Yes"), labels = c("nie", "tak"))

dane1$Residence_type <- factor(dane1$Residence_type , levels = c(wies = "Rural",
                        miasto = "Urban"), labels = c("wies", "miasto"))

dane1$work_type <- factor(dane1$work_type , levels = c(prywatna = "Private",
                  dzialalnosc = "Self-employed", rzadowa = "Govt_job",
                  nie_pracuje_to_dziecko = "children"),
                  labels = c("prywatna", "dzialalnosc", "rzadowa", "nie_pracuje_to_dziecko"))

dane1$smoking_status <- factor(dane1$smoking_status , levels =
                        c(nigdy = "never smoked", kiedys = "formerly smoked",
                        pali = "smokes", nieznany = "Unknown"),
                        labels = c("nigdy", "kiedys", "pali", "nieznany"))

dane1$stroke <- factor(dane1$stroke , levels = c(nie_mial = 0,
                mial = 1), labels = c("nie_mial", "mial"))

dane1$age <- as.integer(dane1$age) #zmieniamy na pełne, ukończone lata

dane1$smoking_status[which(dane1$age < 11)] <- factor(dane1$smoking_status[which(dane1$age < 11)], leve

head(dane1, 2) #wyświetlamy dla sprawdzenia czy zaszły zmiany:
```

```
##      gender age hypertension heart_disease ever_married work_type
## 1 mezczyzna  67          nie           tak          tak  prywatna
## 2 mezczyzna  80          nie           tak          tak  prywatna
##   Residence_type glucose_level  bmi smoking_status stroke
## 1         miasto        228.69 36.6         kiedys   mial
## 2           wies        105.92 32.5          nigdy   mial
```

Podstawowe statystyki:

```
lapply(dane1, summary)
```

```
## $gender
##   kobieta mezczyzna
##      2907      2074
##
## $age
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   25.00   45.00   43.41   61.00   82.00
##
## $hypertension
##  nie  tak
## 4502  479
##
## $heart_disease
##  nie  tak
## 4706  275
##
## $ever_married
##  nie  tak
## 1701 3280
##
## $work_type
##               prywatna            dzialalnosc                rzadowa
##                   2860                    804                    644
## nie_pracuje_to_dziecko
##                    673
##
## $Residence_type
##   wies miasto
##   2449   2532
##
## $glucose_level
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   55.12   77.23   91.85  105.94  113.86  271.74
##
## $bmi
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    14.0    23.7    28.1    28.5    32.6    48.9
##
## $smoking_status
##    nigdy   kiedys     pali nieznany     NA's
##     2316      863      775     1017       10
##
## $stroke
## nie_mial     mial
##     4733      248
```

Wniosek:

Na początku możemy przytoczyć pare faktów:

1) W badaniu brały udział również dzieci i niemowlęta.

2) Większość badanych choruje na nadciśnienie, ale też w przypadku większości nie stwierdzono chorób serca.

Zdrowy dorosły człowiek o prawidłowym stosunku wagi do wzrostu ma wskaźnik w granicach 18,5–24,9. BMI poniżej 18,5 informuje o niedowadze. Osoby ze wskaźnikiem BMI 25,0-29,9 cierpią na nadwagę. Współczynnik BMI powyżej 30,0 to już otyłość.

3) Wnioskujemy z tego, że średnie BMI badanych jest niepokojąco wysokie. Maksymalna wartość i mediana są niemalże dwukrotnie większe niż zdrowe BMI. A Najmniejsza wartość BMI świadczy o niedowadze.

Norma poziomu cukru we krwi dla osób dorosłych wynosi: na czczo 80–140. Natomiast norma glukozy na czczo u dzieci to 70–100 mg/dl.

4) W przypadku poziomu glukozy: średnia i mediana wypadają bardzo dobrze, jednak jak możemy zauważyć znów pojawiają się osoby ze zbyt niskim i ze zbyt, bo aż ponad dwukrotnie za wysokim wynikiem.

5) W badaniu bierze udział o 1/3 więcej kobiet niż mężczyzn, ale grupy zawierające osoby mieszkające na wsi i w mieście są zbliżonej wielkości.Na szczęście zdecydowanie przeważa ilość osób dorosłych nad dziećmi.

6) Zdecydowana większość badanych nie miała udaru mózgu.

Sprawdzamy braki:

```
lapply(lapply(dane1, is.na), sum)
```

```
## $gender
## [1] 0
##
## $age
## [1] 0
##
## $hypertension
## [1] 0
##
## $heart_disease
## [1] 0
##
## $ever_married
## [1] 0
##
## $work_type
## [1] 0
##
## $Residence_type
## [1] 0
##
## $glucose_level
## [1] 0
##
## $bmi
## [1] 0
```

```
##
## $smoking_status
## [1] 10
##
## $stroke
## [1] 0
```

Wniosek: Na szczęście nie mamy żdanych braków.

Analiza zależności:

Początkowo staramy się zidentyfikować zmienne zależne, niezależne oraz towarzyszące. W dalszej części na podstawie wykresów pokażemy, że wiek jest zmienną towarzyszącą.

Sprawdzamy korelacje między zmiennymi numerycznymi:

```
cor(dane1[sapply(dane1, is.numeric)])
```

```
##                     age glucose_level       bmi
## age           1.0000000     0.2366505 0.3740642
## glucose_level 0.2366505     1.0000000 0.1863482
## bmi           0.3740642     0.1863482 1.0000000
```

Wniosek:

1) Występuje tylko korelacja dodatnia, zważając na to, że są to realne dane to mimo, iż korelacje nie wydają się wysokie - uważamy, że jest w porządku.
2) Wykorzystamy interpretacje wykresów do wyciągnięcia wiecej wniosków.

Dla hipotezy pierwszej:

```
library(GGally)
```

```
## Loading required package: ggplot2
```

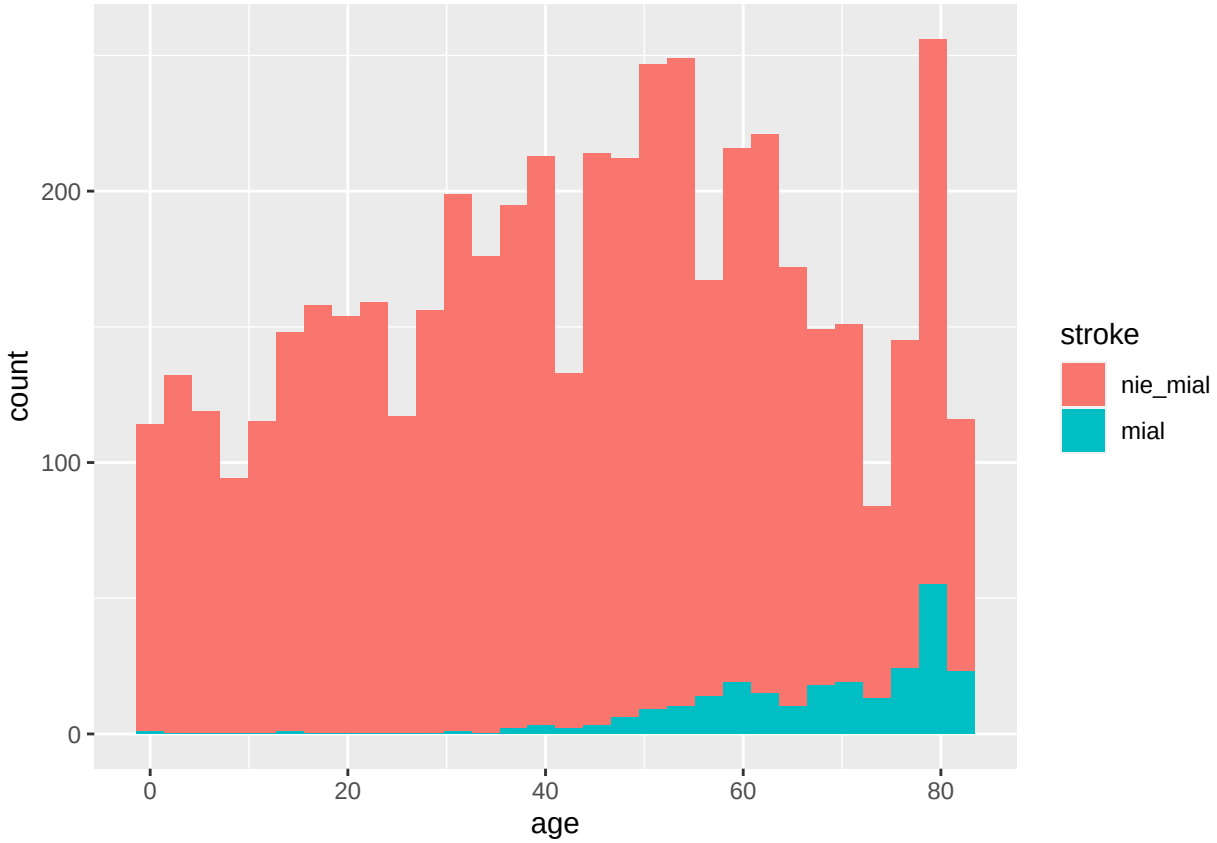
```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

```
library(ggplot2)
```

Zaczniemy od porównania wieku oraz występowania udaru:

```
ggplot(dane1, aes(x = age, fill = stroke)) + geom_histogram()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

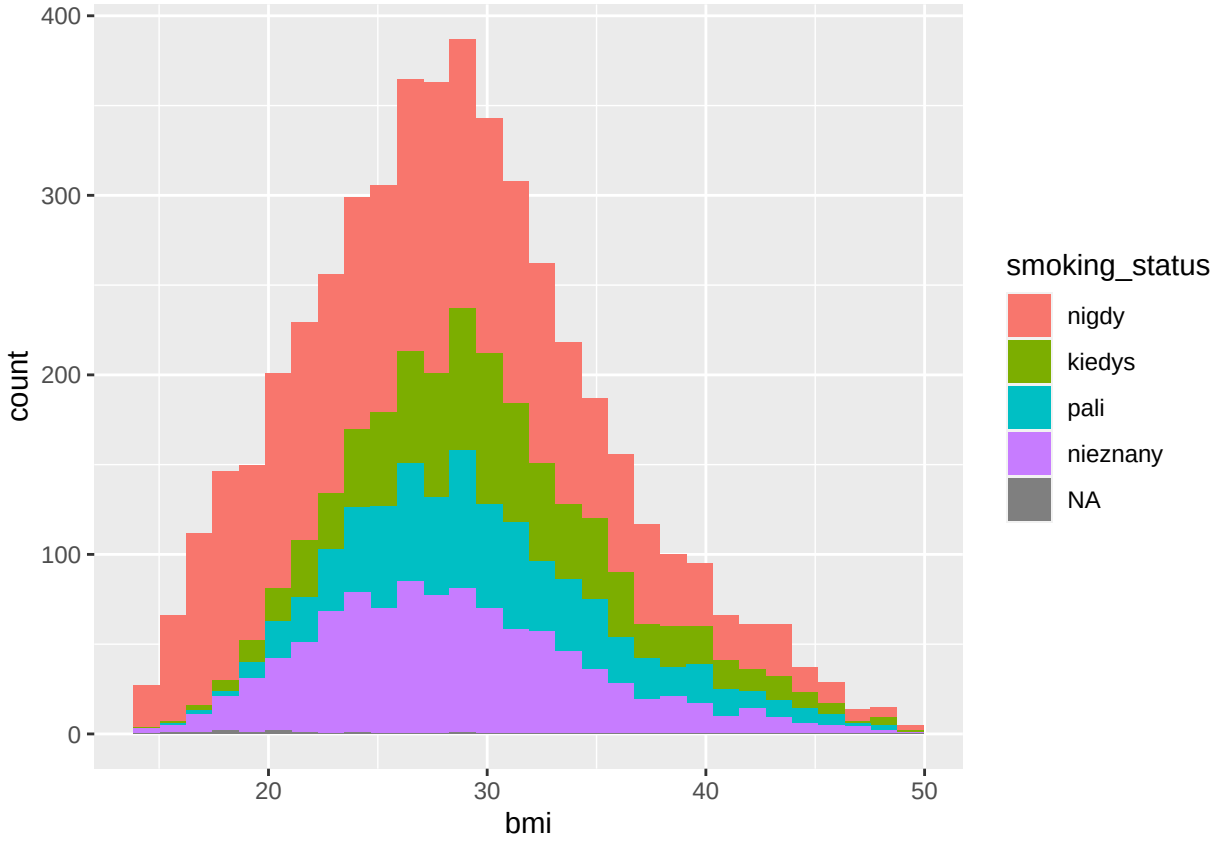


Wniosek: Widzimy wyraźnie, że z wiekiem rośnie ilość osób, które miały udar mózgu, zatem im człowiek straszy tym większe prawdopodobieństwo, że będzie miał udar. Wnioskujemy z tego, że wiek jest zmienną towarzyszącą.

Teraz spójrzmy jak wygląda status palenia w zależności od BMI:

```
ggplot(dane1, aes(x = bmi, fill = smoking_status)) + geom_histogram()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Wniosek: Na wykresie widać, że w każdej grupie "statusu palenia", BMI rozkłada się w podobny sposób. Co wypada pozytywnie - zdecydowana większość nie pali, bądź przestała, bo paliła kiedyś. Niestety status nieznany może trochę negatywnie wpływać na wyniki.

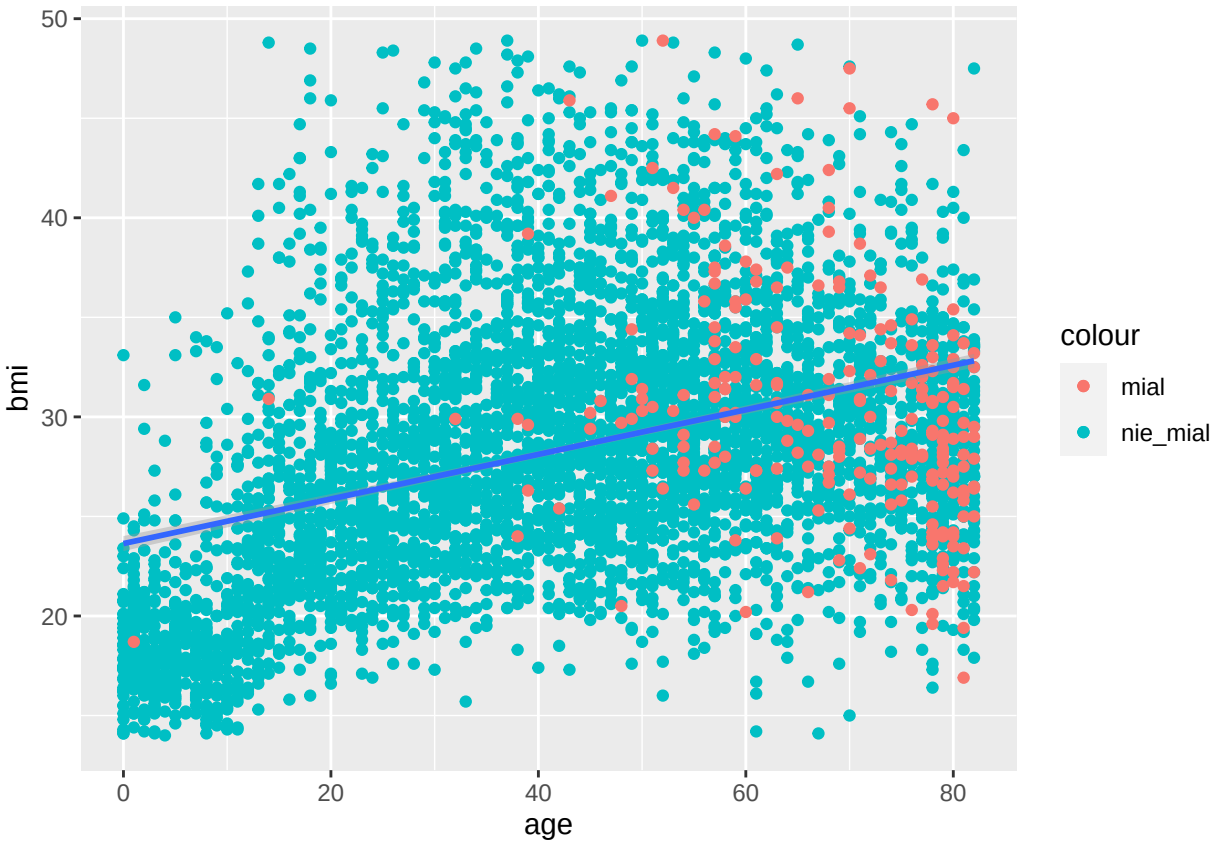
Spójrzmy jeszcze czy BMI rośnie z wiekiem i jak mają się te dwie zmienne do udaru:

W tym celu musimy przygotować sobie pomocnicze dane:

```
dane1_nie_mial <- dane1[dane1$stroke == "nie_mial",]
dane1_mial <- dane1[dane1$stroke == "mial",]

ggplot(dane1, aes(x = age, y = bmi)) +
    geom_point(
        data = dane1_nie_mial,
        aes(colour = "nie_mial")) +
    geom_point(
        data = dane1_mial,
        aes(colour = "mial")) +
    geom_smooth(method = "lm")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



Wniosek: Mamy trend rosnący.

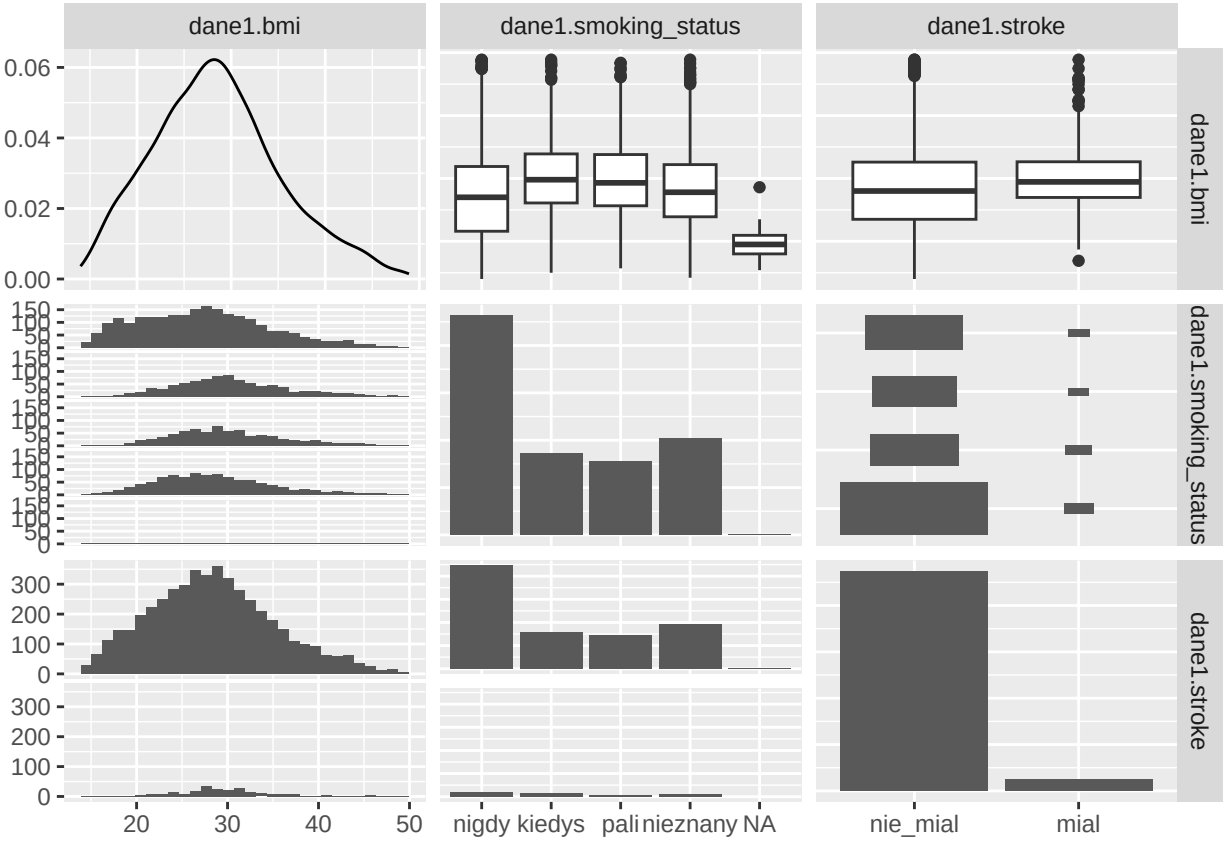
Na koniec porównajmy zmienne wystepujące w hipotezie pierwszej:

```
df1 = data.frame(dane1$bmi, dane1$smoking_status, dane1$stroke)
ggpairs(df1)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## Warning: Removed 10 rows containing non-finite values (`stat_g_gally_count()`).
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Wniosek:

1) Rzuca się w oczy, że przeważa ilość osób, które nie doświadczyły udaru.
2) Jeżeli spróbujemy porównać osoby które miały udar i te, które nie miały w zależności od tego czy palą - wyraźnie widać, że w przypadku osób, które nie miały udaru, znacznie więcej pacjentów wybrało opcję "nigdy nie paliłem" niż w przypadku osób, które miały udar.
3) Porównując BMI i ststus palenia - definitywnie odznacza się, że osoby, które nigdy nie paliły częściej mają niższe i lepsze BMI niż w pozostałych sytuacjach.

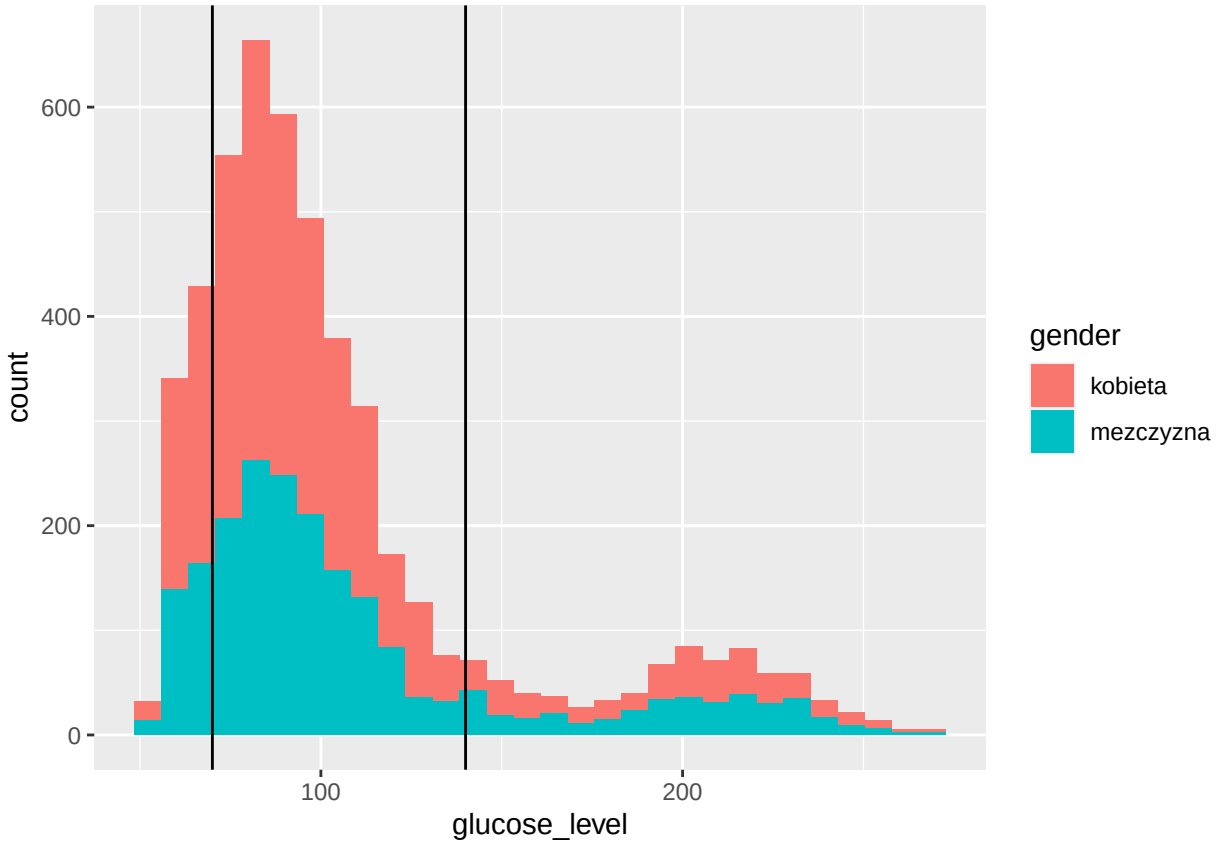
Dla hipotezy drugiej:

Z ciekawości spójrzmy na poziom glukozy i płeć:

Zaznaczę również przedział dobrzego poziomu glukozy

```
ggplot(dane1, aes(x = glucose_level, fill = gender)) + geom_histogram() + geom_vline(xintercept = 70) +
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

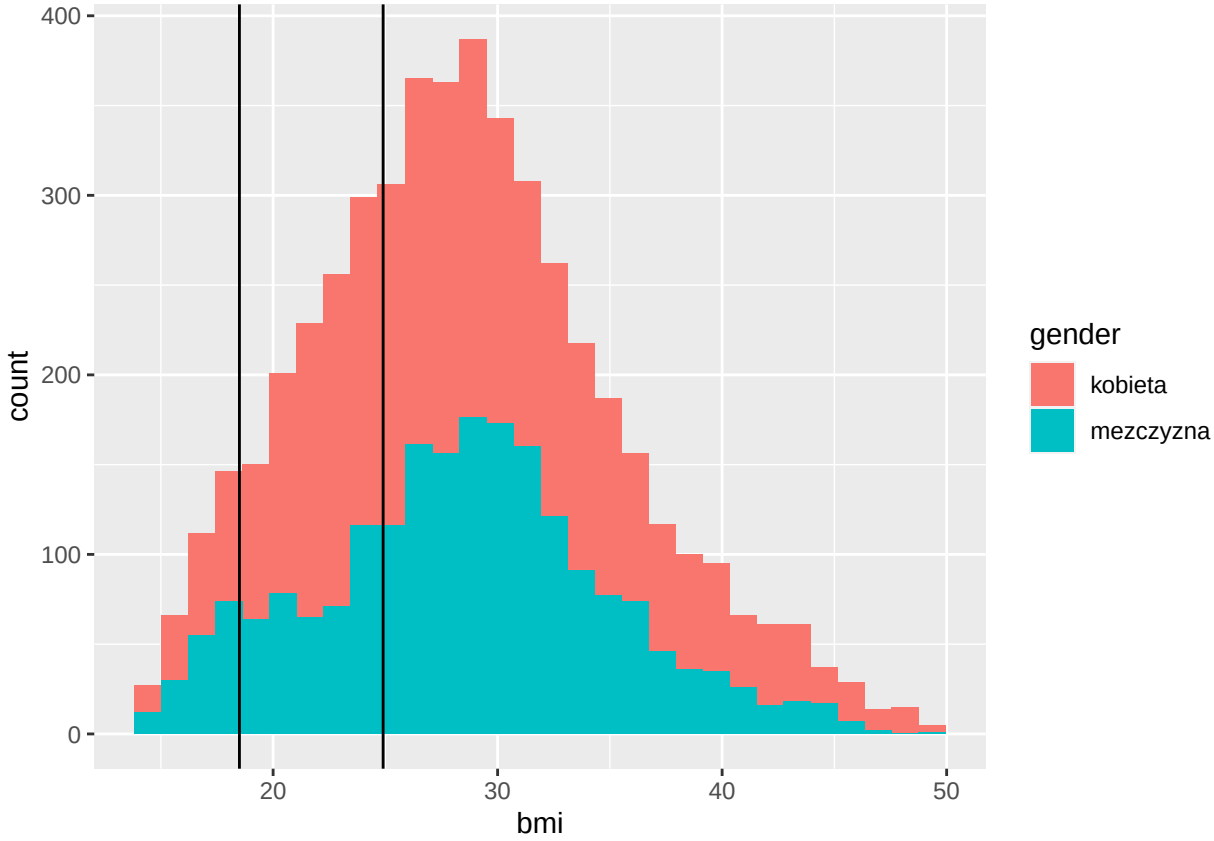


Wniosek: Mimo, iż przeważa ilośc kobiet w tym badaniu widzimy, że płeć nie ma znaczenia. Rozkłady ilościowe wyglądają bardzo podobnie. Wnioskujemy z tego, że średnio tyle samo kobiet co mężczyzn ma za niski lub za wysoki poziom glukozy.

A czy jest różnica w przypadku BMI? Dodaję również przedziały dla poprawnego BMI.

```
ggplot(dane1, aes(x = bmi, fill = gender)) + geom_histogram() + geom_vline(xintercept = 18.5) + geom_vl
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Wniosek: Jest dokładnie tak samo jak w przypadku glukozy.

Teraz chcemy zobaczyć jaka jest zależność między poziomem BMI i poziomem glukozy w przełożeniu na wystąpienie udaru:

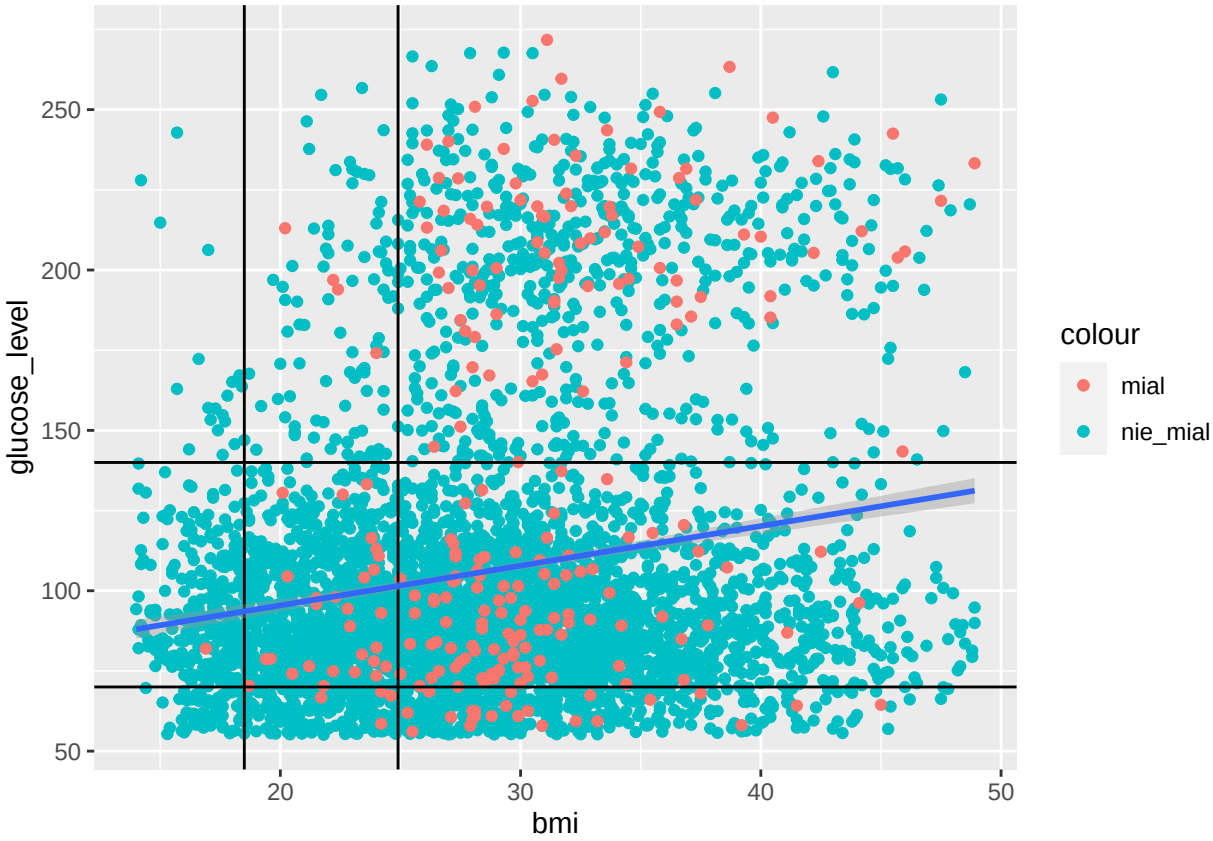
W tym celu musimy przygotować sobie pomocnicze dane:

```
dane1_nie_mial <- dane1[dane1$stroke == "nie_mial",]
dane1_mial <- dane1[dane1$stroke == "mial",]

ggplot(dane1, aes(x = bmi, y = glucose_level )) +
    geom_point(
        data = dane1_nie_mial,
        aes(colour = "nie_mial")) +
    geom_point(
        data = dane1_mial,
        aes(colour = "mial")) +

    geom_hline(yintercept = 70) +
    geom_hline(yintercept = 140) +
    geom_vline(xintercept = 18.5) +
    geom_vline(xintercept = 24.9) +
    geom_smooth(method = "lm")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



Wniosek: Z wykresu można wyraźnie zauważyć, że przy nieodpowiednim poziomie glukozy i BMI dużo częściej występuje udar. Zapewne ma to też związek z tym, że niewiele osób ma "idealne" wyniki.

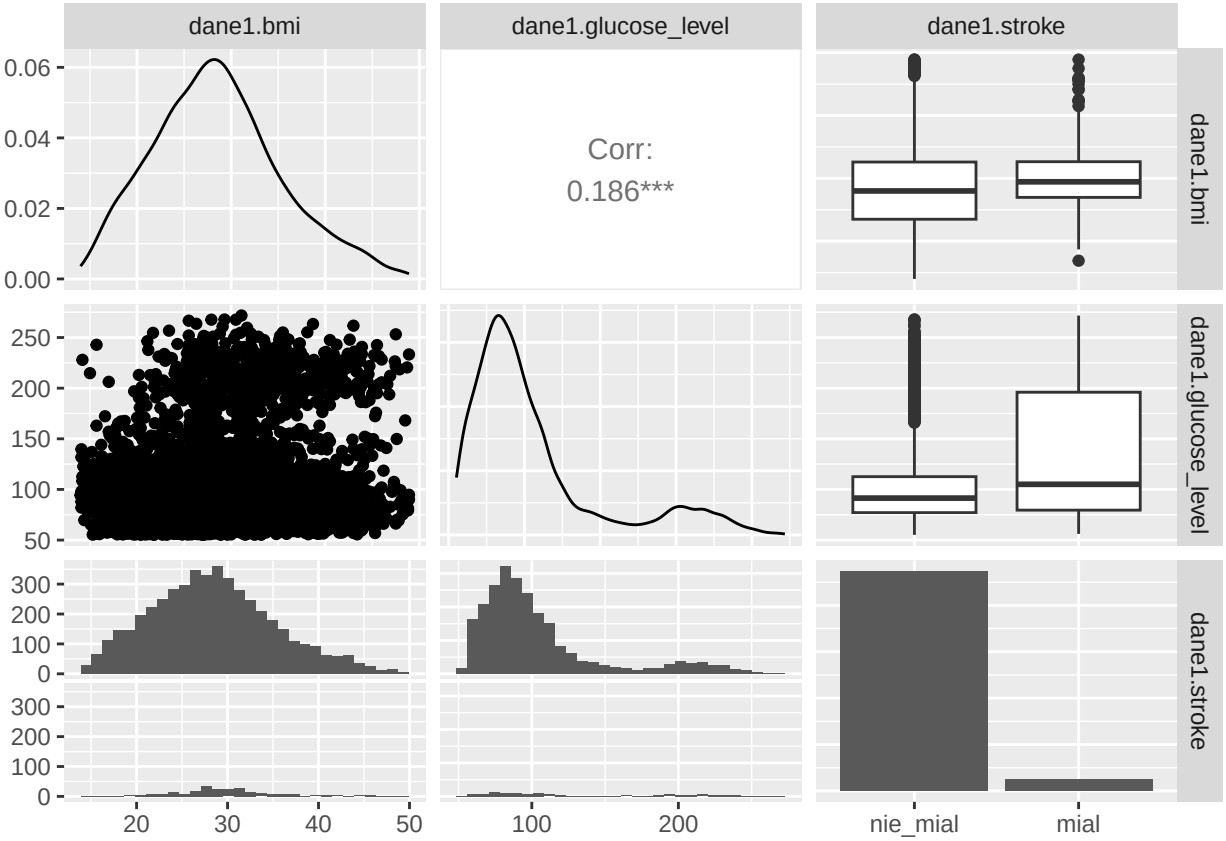
Na wykresie zaznaczono przedział z odpowiednim poziomem glukozy i BMI. Osoby, które mają dobre wyniki należą do powstałego na wykresie prostokąta.

Czerwone kropki przeważają u góry , na prawo, co znaczy, że wyższe BMI towarzyszy przy występowaniu udaru częściej niż niższe. Największe skupisko znajduje się jednak w granicy dobrego poziomu glukozy i trochę zbyt wysokiego BMI.

Na koniec porównajmy zmienne wystepujące w hipotezie drugiej:

```
df2 = data.frame(dane1$bmi, dane1$glucose_level, dane1$stroke)
ggpairs(df2)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Wniosek:

1) Korelacja między BMI, a poziomem glukozy jest dodatnia. Wydaje się słaba, ale zważając na to, że są to realne dane, a nie laboratoryjne - korelacja ta jest w porządku. Poza tym - trzy gwiazdki sugerują, że są to istotne statystycznie dane.

2) Niepokojąco dużo osób z wysokim BMI ma strasznie wysoki poziom glukozy. Widzimy to wyraźnie na środkowym wykresie w pierwszej kolumnie. Lewy dolny róg jest pełen punktów, u góry jednak dominuje środkowa część. Przy wysokim BMI jednak tych punktów wydaje się być więcej.

3) Mediana BMI widoczna na wykresie pudełkowym jest wyższa w przypadku osób, które miały udar.

Tak samo w przypadku poziomu glukozy.

4) Jest spora różnica między kwantylami 1 i 3 w przypadku poziomu glukozy i wystąpienia udaru.

Dla hipotezy trzeciej:

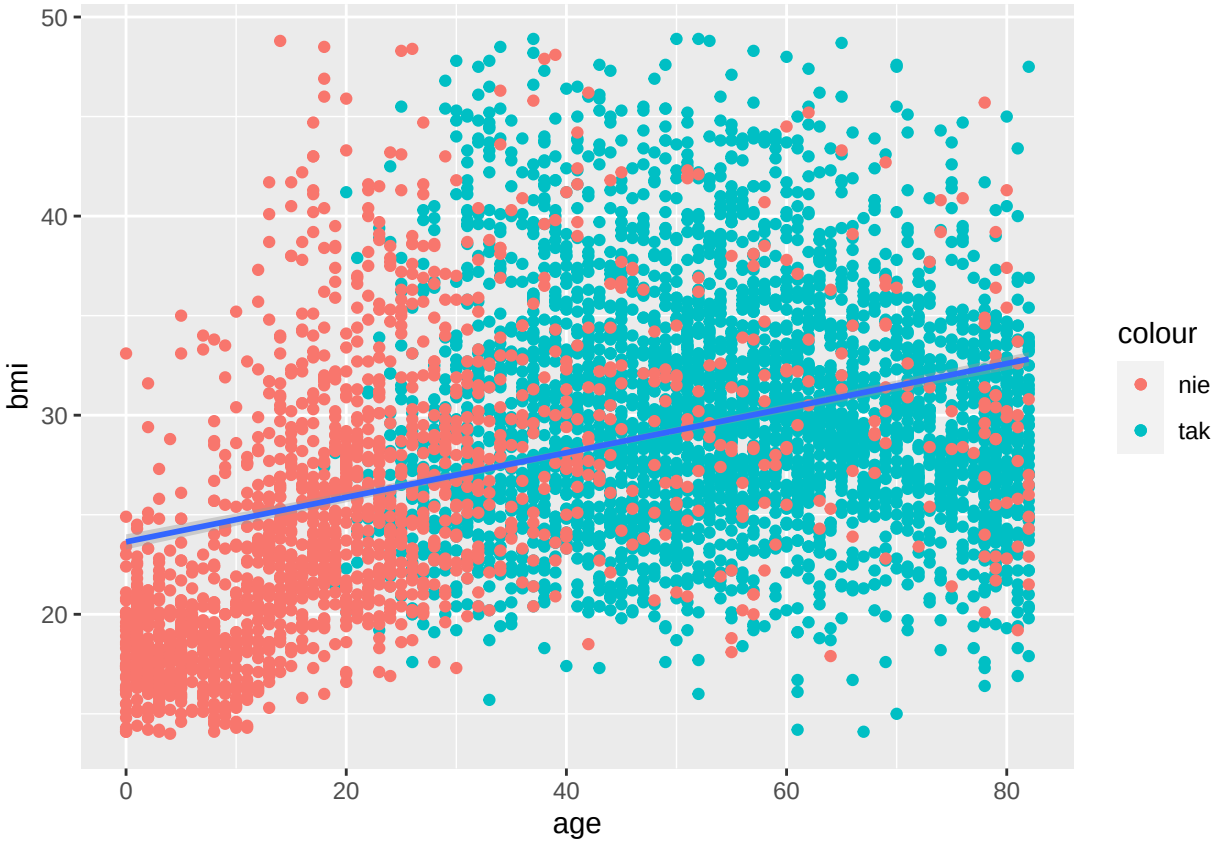
Na początek porównajmy zmienne wiek, bmi i status małżeński:

Musimy zaznaczyć fakt, że dopiero od pewnego wieku można wziąć ślub, więc dzieci tutaj odpadają.

```
dane1_nie <- dane1[dane1$ever_married == "nie",]
dane1_tak <- dane1[dane1$ever_married == "tak",]

ggplot(dane1, aes(x = age, y = bmi)) +
    geom_point(
        data = dane1_tak,
        aes(colour = "tak")) +
    geom_point(
        data = dane1_nie,
        aes(colour = "nie")) + geom_smooth(method = "lm")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```

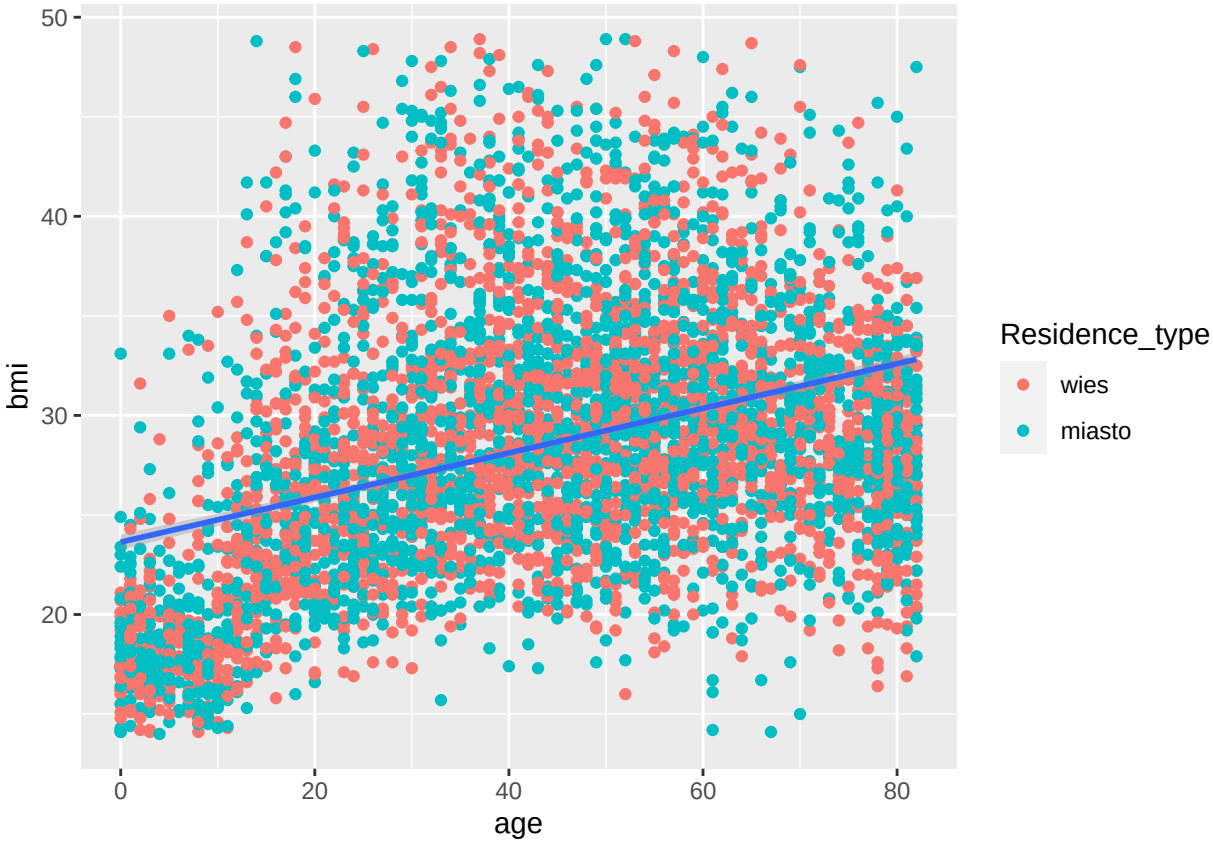


Wniosek: Zdecydowanie widać, że średnio ludzie w związku małżeńskim mają dużo wyższe BMI. Z wiekiem występuje trend rosnący.

Spójrzmy jeszcze na dane o miejscu zamieszkania:

```
ggplot(dane1, aes(x = age, y = bmi)) + geom_point(aes(color = Residence_type)) + geom_smooth(method = "
```

```
## `geom_smooth()` using formula = 'y ~ x'
```

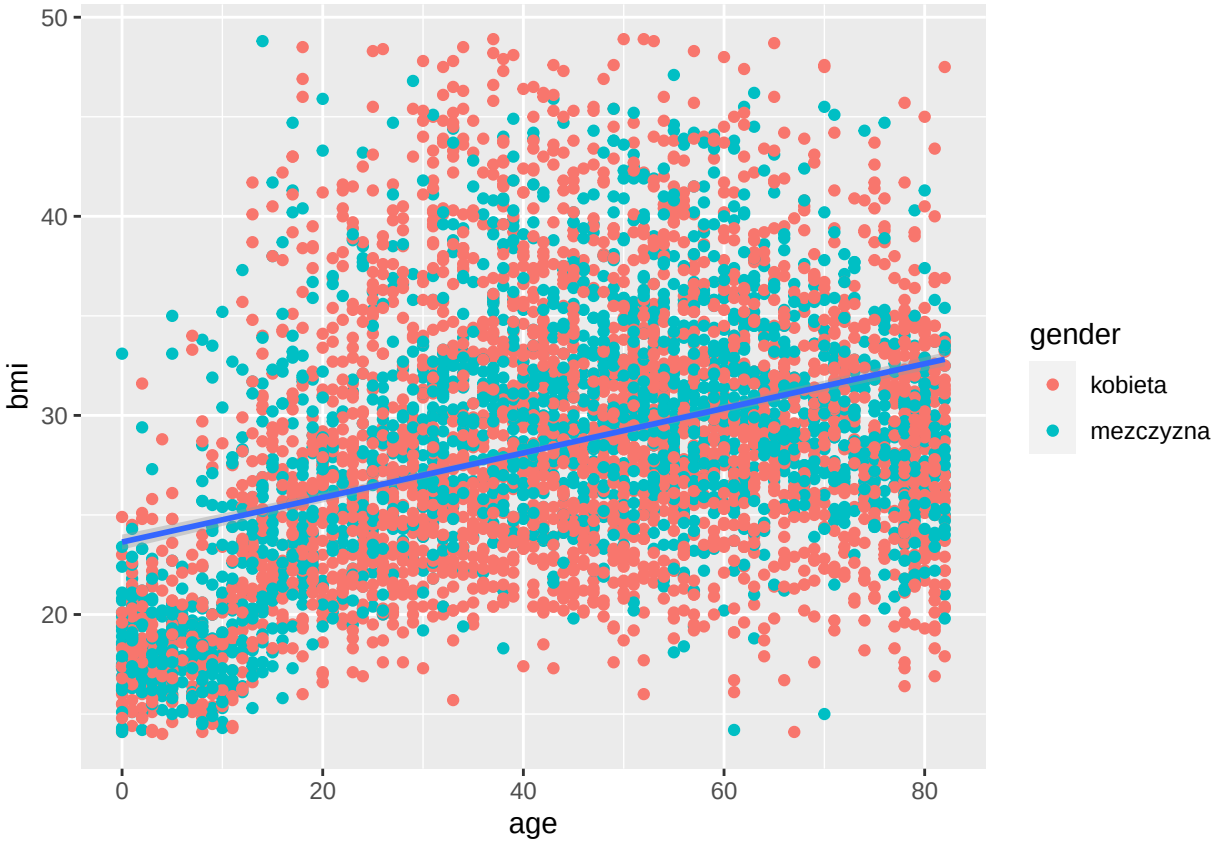


Wniosek: Z racji, że w badaniu wzięła udział podobna ilość osób mieszkających na wsi i w mieście ciężko wyciągnąć jendoznaczny wniosek z wykresu.

Dla upewnienia spójrzmy ostatni raz na zależności między wiekiem, BMI, a płcią:

```
ggplot(dane1, aes(x = age, y = bmi)) + geom_point(aes(color = gender)) + geom_smooth(method = "lm")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```

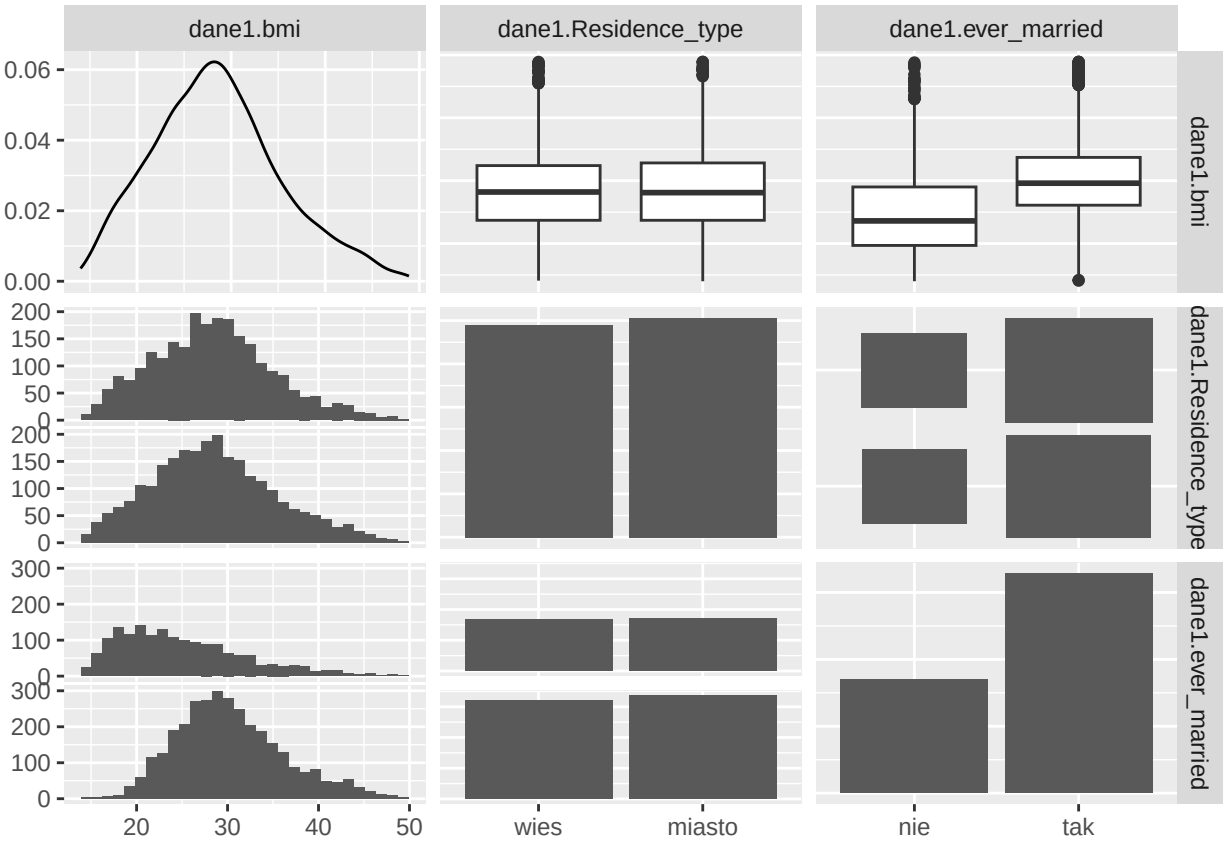


Wniosek: Może się delikatnie zdawać, że mężczyźni mają ciut wyższe BMI, jednak tak jak wcześniej zauważono: niezależnie od płci, BMI rośnie z wiekiem.

Na koniec porównajmy zmienne wystepujące w hipotezie trzeciej:

```
df3 = data.frame(dane1$bmi, dane1$Residence_type, dane1$ever_married)
ggpairs(df3)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Wniosek:

1) W przypadku miejsca zamieszkania nie widac prawie żadnej różnicy.

2) Za to status małżeński ma ogromny wpływ na wysokość BMI. W lewym dolnym rogu mamy wykres, który bardzo dobrze pokazuje, iż osoby, które wyszły za mąż mają dużo wyższe BMI i mają problemy z nadwagą czy otyłością. W przypadku osób, które nie mają ślubu pojawia się częściej inny problem - niedowaga. Na wykresie pudełkowym w prawym górnym rogu widać, że średnie BMI jest dużo większe dla osób po ślubie.

Model pierwszy:

Po pierwsze musimy zamienić w zmiennej stroke oznaczenia na 0 - nie miał i 1 - miał, ponieważ inaczej nie będziemy mogli zbudowac modelu.

```
dane1_od_nowa <- read.csv("full_data.csv")
dane1$stroke <- dane1_od_nowa$stroke
```

Wykresy diagnostyczne:

```
model1 <- lm(stroke ~ bmi + smoking_status, dane1)
summary(model1)
```

```
##
## Call:
## lm(formula = stroke ~ bmi + smoking_status, data = dane1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.10888 -0.05779 -0.04480 -0.03479  0.97672
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)             -0.001984   0.013405  -0.148  0.88233
## bmi                      0.001495   0.000462   3.236  0.00122 **
## smoking_statuskiedys     0.037767   0.008767   4.308 1.68e-05 ***
## smoking_statuspali       0.011177   0.009099   1.228  0.21934
## smoking_statusnieznany   0.004628   0.008183   0.566  0.57169
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2171 on 4966 degrees of freedom
##   (10 observations deleted due to missingness)
## Multiple R-squared:  0.007014,   Adjusted R-squared:  0.006214
## F-statistic:  8.77 on 4 and 4966 DF,  p-value: 4.728e-07
```

Wniosek:

1) Sprawdzono i nie ma interakcji. Zatem korzystamy z "+".
2) Wyraźnie widać, że wysokość BMI oraz status palenia mają wpływ na to czy osoba przeszła udar. (Patrzymy na ilość gwiazdek)

Sprawdzimy jeszcze kryterium i błąd średniokwadratowy:

```
extractAIC(model1)
```

```
## [1]      5.00 -15182.19
```

```
sqrt(mean(model1$residuals^2))
```

```
## [1] 0.2169514
```

Wniosek:

RMSE wypada całkiem nieźle, tak samo jak kryterium AIC.

Wniosek do hipotezy pierwszej:

Model potwierdza słuszność hipotezy. BMI, tak jak status palenia wywierają znaczący wpływ na wystąpienie udaru.

Model drugi:

```
model2 <- lm(stroke ~ bmi + glucose_level, dane1)
summary(model2)
```

```
##
## Call:
## lm(formula = stroke ~ bmi + glucose_level, data = dane1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.16070 -0.05397 -0.03994 -0.02854  0.98395
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -4.552e-02  1.405e-02  -3.240   0.0012 **
## bmi            1.065e-03  4.578e-04   2.327   0.0200 *
## glucose_level  6.130e-04  6.896e-05   8.890   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2155 on 4978 degrees of freedom
## Multiple R-squared:  0.01882,    Adjusted R-squared:  0.01842
## F-statistic: 47.73 on 2 and 4978 DF,  p-value: < 2.2e-16
```

```
cor.test(dane1$bmi, dane1$glucose_level)
```

```
##
##  Pearson's product-moment correlation
##
## data:  dane1$bmi and dane1$glucose_level
## t = 13.384, df = 4979, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.1594010 0.2130179
## sample estimates:
##       cor
## 0.1863482
```

```
t.test(dane1$bmi, dane1$glucose_level)
```

```
##
##  Welch Two Sample t-test
##
## data:  dane1$bmi and dane1$glucose_level
## t = -119.91, df = 5205.9, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -78.71159 -76.17919
## sample estimates:
## mean of x mean of y
##  28.49817 105.94356
```

Wniosek:

1) Poziom glukozy jest bardziej istotny statystycznie niż BMI, ale gdyby przyjęto poziom istotności 0.05 to również można by powiedzieć, że jest on istotny statystycznie.
2) Nie wystepuje interakcja między poziomem glukozy i BMI.

Sprawdzimy jeszcze kryterium i błąd średniokwadratowy:

```
extractAIC(model2)
```

```
## [1]      3.00 -15285.79
```

```
sqrt(mean(model2$residuals^2))
```

```
## [1] 0.215453
```

Wniosek: Tak jak poprzednio jest całkiem.

Wniosek do hipotezy drugiej:

Niestety hipoteza nie jest całkowicie prawdziwa, ponieważ to poziom glukozy ma dużo większy wpływ na występowanie udaru.

Nie mozemy jednak powiedzieć, że obie te zmienne w połączeniu częściej powodują udar.

Model trzeci:

```
model3 <- lm(bmi ~ ever_married * Residence_type, dane1)
summary(model3)
```

```
##
## Call:
## lm(formula = bmi ~ ever_married * Residence_type, data = dane1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -16.2789  -4.6701  -0.8701   3.6299  23.7299
##
## Coefficients:
##                                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)                         24.91655    0.21679 114.935   <2e-16 ***
## ever_marriedtak                      5.33280    0.26795  19.902   <2e-16 ***
## Residence_typemiasto                 0.15351    0.30578   0.502    0.616
## ever_marriedtak:Residence_typemiasto -0.02396    0.37684  -0.064    0.949
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.305 on 4977 degrees of freedom
## Multiple R-squared:  0.1383, Adjusted R-squared:  0.1377
## F-statistic: 266.2 on 3 and 4977 DF,  p-value: < 2.2e-16
```

Wniosek:

1) R-squared wynosi 0.1383
2) Nie występuje interakcja między statusem cywilnym, a miejscem zamieszkania.
3) Możemy tez wywnioskować, że życie w małżeństwie wywiera ogromny wpływ na wysokość BMI.
4) Miejsce zamieszkania nie ma żadnego istotnego statystycznie wpływu na BMI.

Wniosek do hipotezy trzeciej:

Ponownie nie możemy potwierdzić słuszności hipotezy. Owszem życie w związku małżeńskim wywiera ogromny wpływ na BMI, ale życie w mieście nie ma żadnego istotnego statystycznie znaczenia.