Bank Marketing data (with social/economic context)

Maciej Maecki

24 października 2019

Streszczenie

W pliku Bank Marketing data.csv znajduj si dane charakteryzujce klientw pewnego banku oraz kampanie marketingowe skierowane do tych klientw. Doczone s ponadto wskaniki spoeczne i ekonomiczne. Na podstawie tych danych naley zbudowa model prognozujcy szans, e klient w wyniku prowadzonej kampanii zaoy lokat terminow.

Spis treści

1 Wprowadzenie

1.1 Opis problemu

W ramach kampani marketingowej organizowanej przez pewien bank w latach midzy majem 2008 rok, a listopadem 2010 roku, byy zbierane informacje na temat klientw tego banku. Na podstawie tych danych planowane jest przewidzenie, czy i jakie rodzaj klientw kupi lokat terminow w tym banku.

1.2 Opis danych

Nasze dane zawieraj 21 column danych. Kolumny moemy podzieli na 3 grupy:

I: Zmienne zwizane z danymi klienta bankowego:

1. Wiek (age): wiek klienta.
2. Praca (job): rodzaj pracy klienta.
3. Stan cywilny (marital): stan cywilny klienta.
4. Edukacja (education): edukacja klienta.
5. Domylnie (default): Klient wczeniej domylnie mia kredyt.
6. Mieszkanie (housing): Klient ma kredyt mieszkaniowy.
7. Poyczka (loan): Klient ma osobist poyczk.

II: Zmienne zwizane z ostatnim kontaktem biecej kampanii marketingowej:

8. Kontakt (contact): Typ komunikacji kontaktowej (telefonicznej lub komrkowej).
9. Miesic (month): Ostatni kontakt miesica roku.
10. Dzie tygodnia (day of week): dzie ostatniego kontaktu tygodnia.
11. Czas trwania (duration): czas trwania ostatniego kontaktu w sekundach. Jeli czas trwania wynosi 0, nigdy nie skontaktowalimy si z klientem, aby zaoy konto lokaty terminowej.
12. Kampania (campaign): liczba kontaktw wykonanych podczas tej kampanii i dla tego klienta
13. Liczba dni (pdays): liczba dni, ktre upyny od ostatniego kontaktu klienta z poprzedniej kampanii (warto liczbowa; 999 oznacza, e klient wczeniej si nie skontaktowa)
14. Poprzedni (previous): liczba kontaktw wykonanych przed t kampani i dla tego klienta (numerycznie)
15. Poutcome: wynik poprzedniej kampanii marketingowej (kategorycznie: poraka, nieistniejca, sukces)

III: Atrybuty kontekstu spoecznego i gospodarczego:

16. Emp.var.rate: wskanik zmiennoci zatrudnienia - wskanik kwartalny
17. Cons.price.idx: wskanik cen konsumpcyjnych - wskanik miesiczny
18. Cons.conf.idx: wskanik zaufania konsumentw - wskanik miesiczny
19. Euribor3m: stawka 3-miesiczna euribor - wskanik dzienny
20. Liczba zatrudnionych (nr employed): liczba pracownikw - wskanik kwartalny

Zmienna wyjciowa (podany cel):

21. y - czy klient subskrybowa lokat? (dwjkowy: tak, nie)

1.3 Wstpna eksploracja danych

```
## Error in contrib.url(repos, "source"): trying to use CRAN without setting a mirror
## Error in contrib.url(repos, "source"): trying to use CRAN without setting a mirror
## Error in contrib.url(repos, "source"): trying to use CRAN without setting a mirror
## Error in contrib.url(repos, "source"): trying to use CRAN without setting a mirror
## Error in contrib.url(repos, "source"): trying to use CRAN without setting a mirror
## Error in contrib.url(repos, "source"): trying to use CRAN without setting a mirror
## Error in contrib.url(repos, "source"): trying to use CRAN without setting a mirror
## Error in contrib.url(repos, "source"): trying to use CRAN without setting a mirror
## Error in contrib.url(repos, "source"): trying to use CRAN without setting a mirror
## Error in contrib.url(repos, "source"): trying to use CRAN without setting a mirror
## Error in contrib.url(repos, "source"): trying to use CRAN without setting a mirror
## Error in contrib.url(repos, "source"): trying to use CRAN without setting a mirror
```

Badane dane zawieraj 4119 wierszy oraz 21 kolumn o nastjcych nazwach:

```
##  [1] "age"            "job"            "marital"        "education"
##  [5] "default"        "housing"        "loan"           "contact"
##  [9] "month"          "day_of_week"    "duration"       "campaign"
## [13] "pdays"          "previous"       "poutcome"       "emp.var.rate"
## [17] "cons.price.idx" "cons.conf.idx"  "euribor3m"      "nr.employed"
## [21] "y"
```

Struktura danych:

```
str(df_bank)

## 'data.frame': 4119 obs. of  21 variables:
##  $ age           : int  30 39 25 38 47 32 32 41 31 35 ...
##  $ job           : chr  "blue-collar" "services" "services" "services" ...
##  $ marital       : chr  "married" "single" "married" "married" ...
##  $ education     : chr  "basic.9y" "high.school" "high.school" "basic.9y" ...
##  $ default       : chr  "no" "no" "no" "no" ...
##  $ housing       : chr  "yes" "no" "yes" "unknown" ...
##  $ loan          : chr  "no" "no" "no" "unknown" ...
##  $ contact       : chr  "cellular" "telephone" "telephone" "telephone" ...
##  $ month         : chr  "may" "may" "jun" "jun" ...
##  $ day_of_week   : chr  "fri" "fri" "wed" "fri" ...
##  $ duration      : int  487 346 227 17 58 128 290 44 68 170 ...
##  $ campaign      : int  2 4 1 3 1 3 4 2 1 1 ...
##  $ pdays         : int  999 999 999 999 999 999 999 999 999 999 ...
##  $ previous      : int  0 0 0 0 0 2 0 0 1 0 ...
##  $ poutcome      : chr  "nonexistent" "nonexistent" "nonexistent" "nonexistent" ...
##  $ emp.var.rate  : num  -1.8 1.1 1.4 1.4 -0.1 -1.1 -1.1 -0.1 -0.1 1.1 ...
##  $ cons.price.idx: num  92.9 94 94.5 94.5 93.2 ...
##  $ cons.conf.idx : num  -46.2 -36.4 -41.8 -41.8 -42 -37.5 -37.5 -42 -42 -36.4 ...
##  $ euribor3m     : num  1.31 4.86 4.96 4.96 4.19 ...
##  $ nr.employed   : num  5099 5191 5228 5228 5196 ...
##  $ y             : chr  "no" "no" "no" "no" ...
```

Czy w danych znajduj sirtoci typu `NaN` lub `Na` ?

```
## [1] FALSE
```

Jednake wiemy, e w danych wystj wartoci brakujce i s one opisa "unknown". W danych znajduje si30 rekordartoci "unknownżozmieszczonych w 1029 rch wierszach. To stanowi 24.98% wszystkich wierszy w naszej badzie danych, wiie moemy pozwolibie na usunie tych wszystkich informacji. W tabeli 1 znajduj siformacje na temat liczny nieznanych wartoci w kadej z kolumn z osobna.

```
## Error: nie znaleziono obiektu 'Number_of_unknown'
## Error in kable(table_unknown, "latex", booktabs = F, caption = "Liczba nieznanych
warto<U+009C>ci w poszczegulnych kolumnach."): nie znaleziono obiektu 'table_unknown'
```

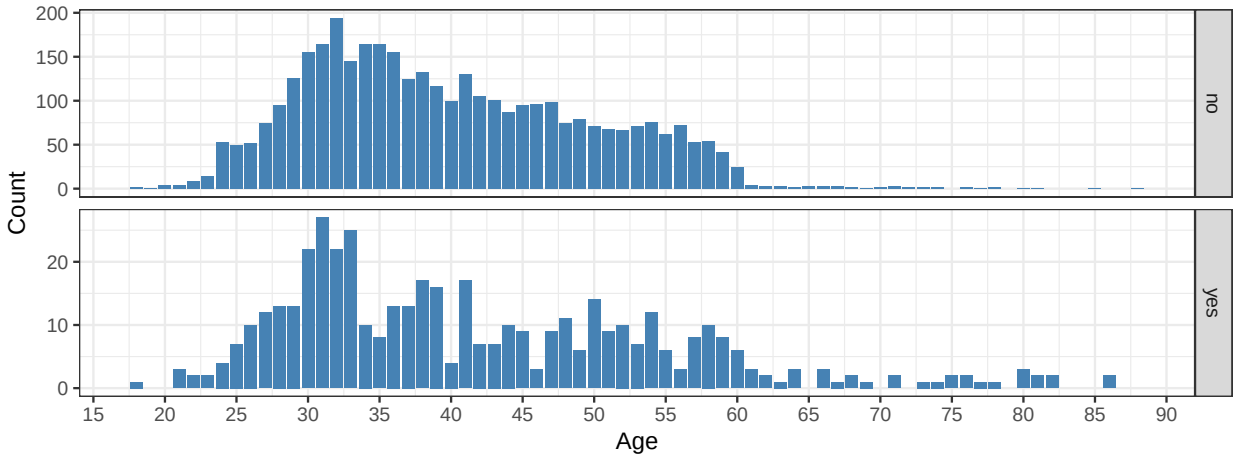
2 Analiza eksploracyjna

W tej sekcji zostanie omny kady parametr z osobna. Nastie dane zostan odpowiednio przygotowane do wykorzystania ich w modelach predykcyjnych.

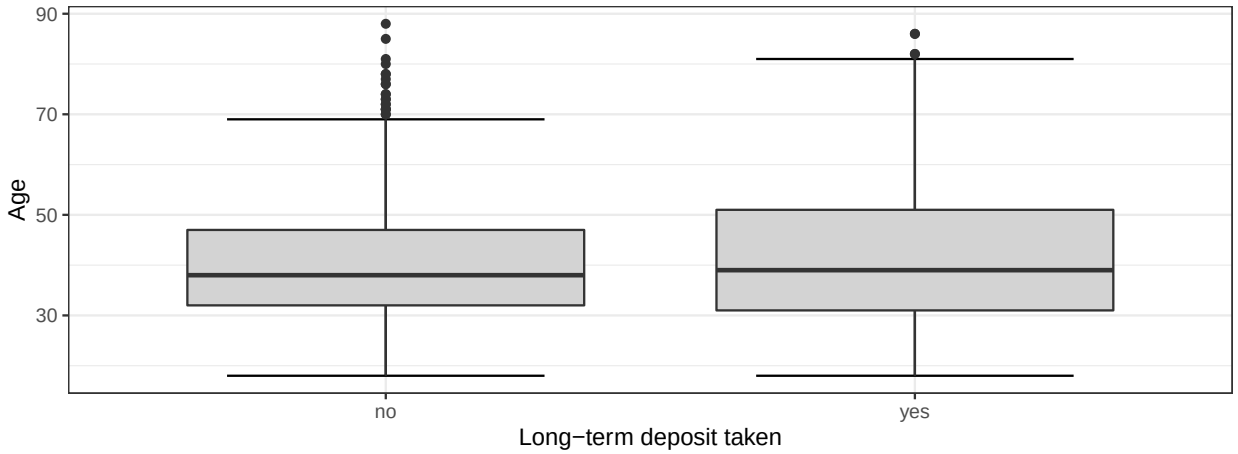
2.1 Age

W jakim wieku byy osoby, z kti skontaktowano sidczas tej kampani?

```
## Error in dimnames(x) <- dnx: 'dimnames()' zastosowane do nie-tablicy
```

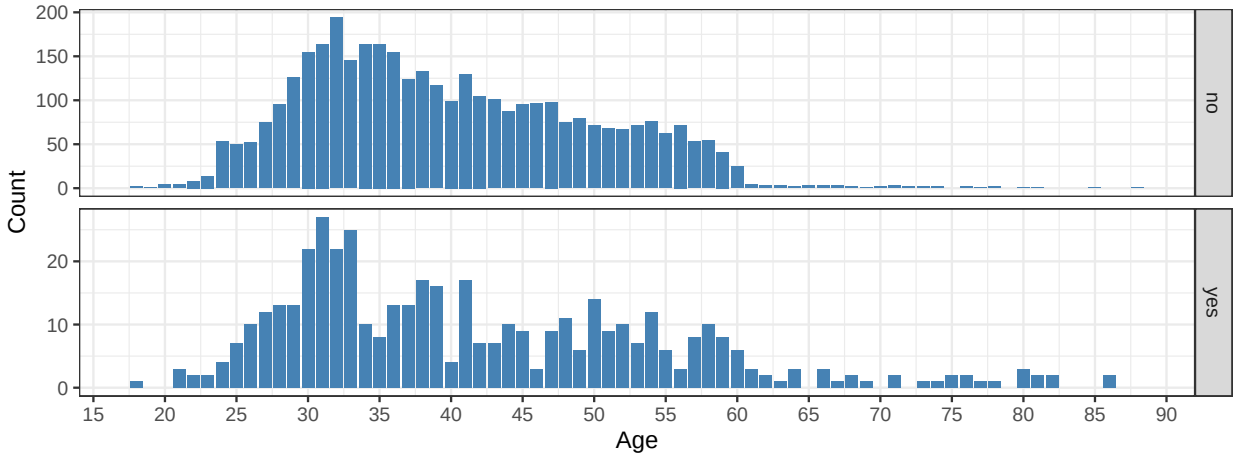


Rysunek 1: Histogram wieku klient<U+663C><U+3E33>w w zale<U+623C><U+3E66>no<U+393C><U+3E63>ci od wzi<U+653C><U+3E61>cia lokaty d<U+623C><U+3E33>ugoterminowej.

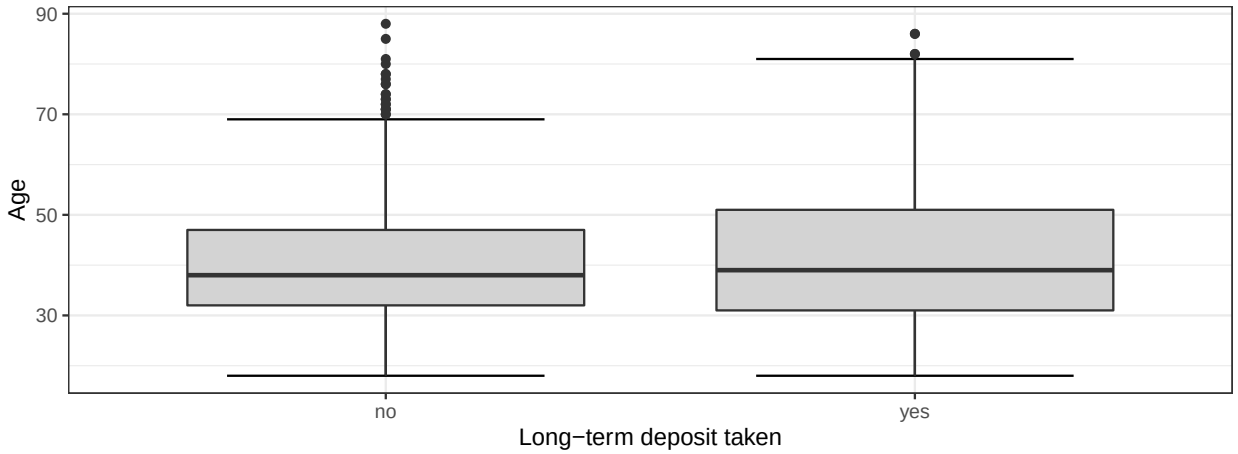


Rysunek 2: Boxplot wieku klient<U+663C><U+3E33>w w zale<U+623C><U+3E66>no<U+393C><U+3E63>ci od wzi<U+653C><U+3E61>cia lokaty d<U+623C><U+3E33>ugoterminowej.

2.2



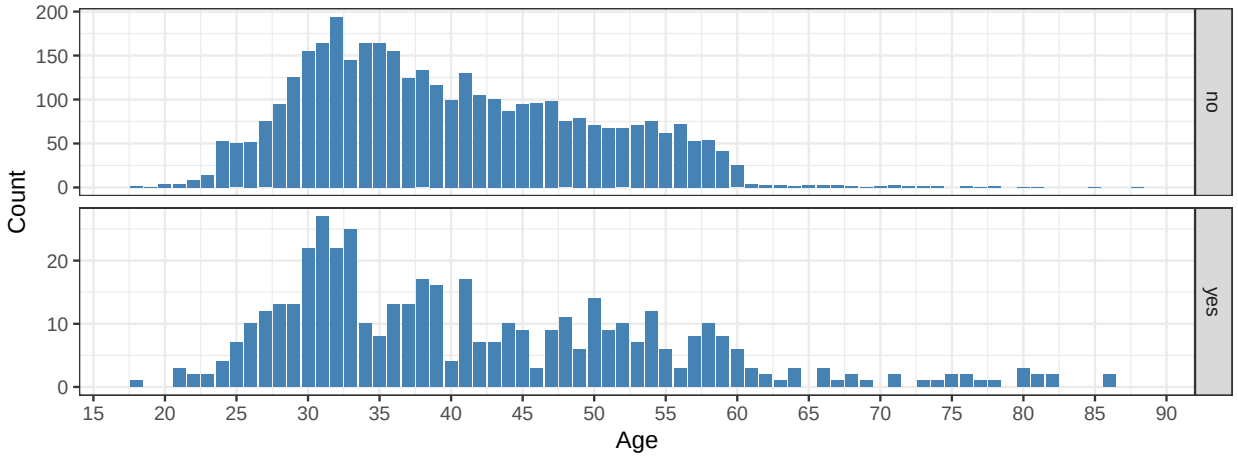
Rysunek 3: New Figure



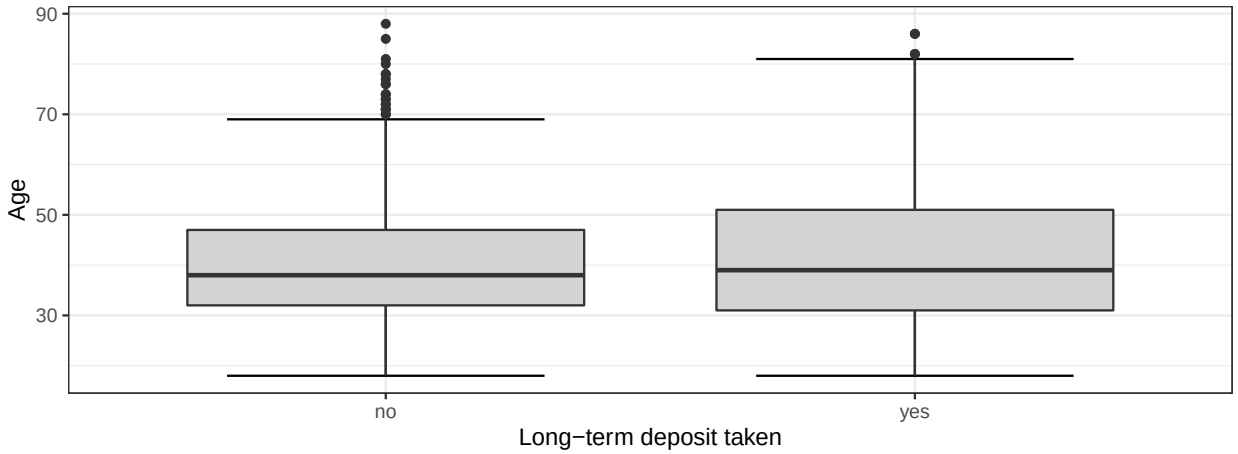
Rysunek 4: New Figure

```
## Error in dimnames(x) <- dnx: 'dimnames()' zastosowane do nie-tablicy
```

2.3



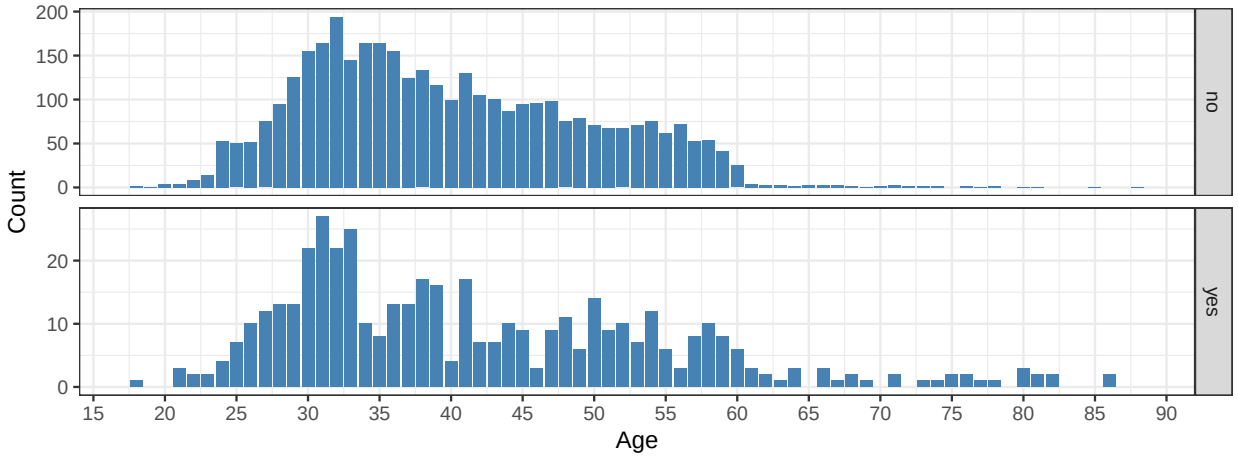
Rysunek 5: New Figure



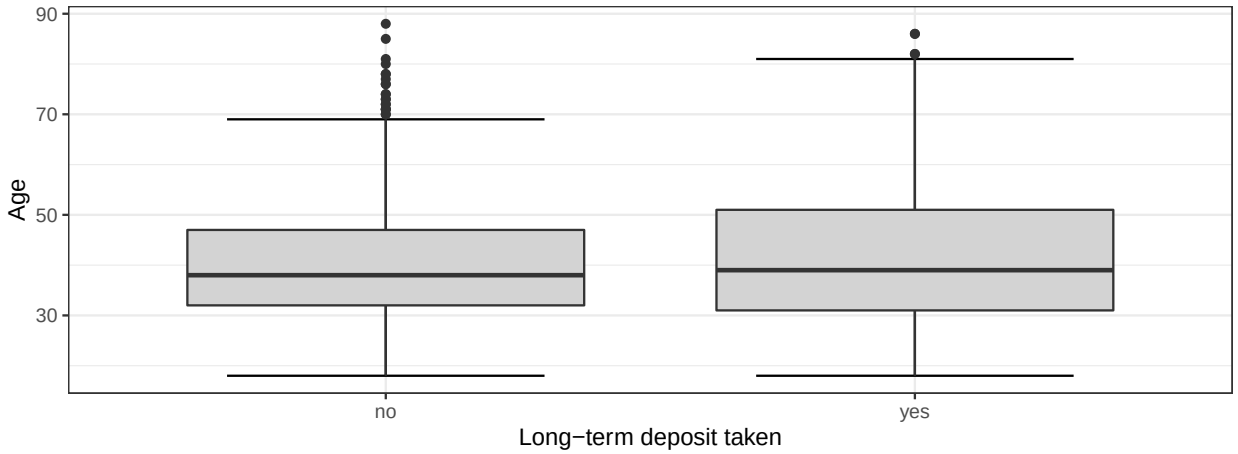
Rysunek 6: New Figure

```
## Error in dimnames(x) <- dnx: 'dimnames()' zastosowane do nie-tablicy
```

2.4



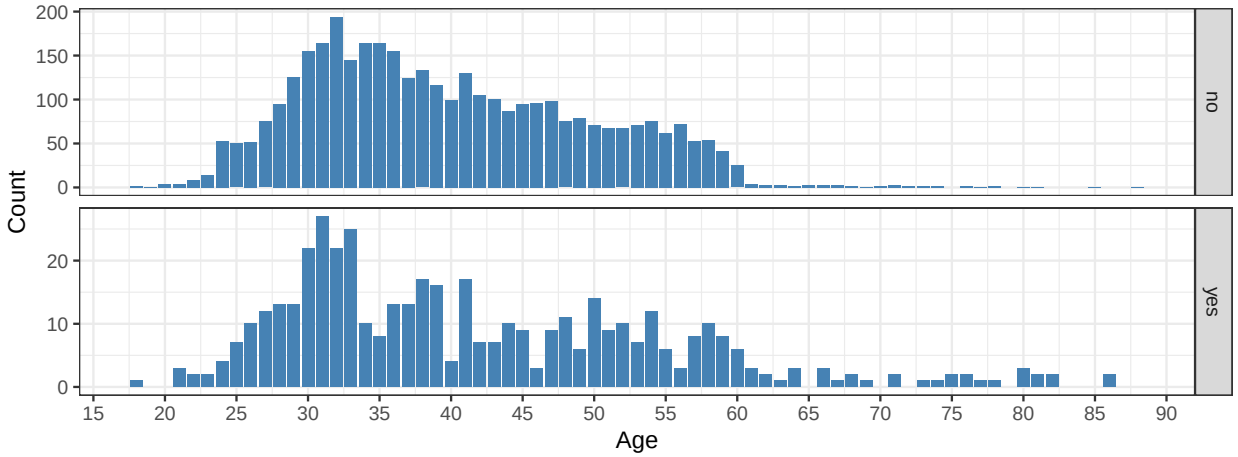
Rysunek 7: New Figure



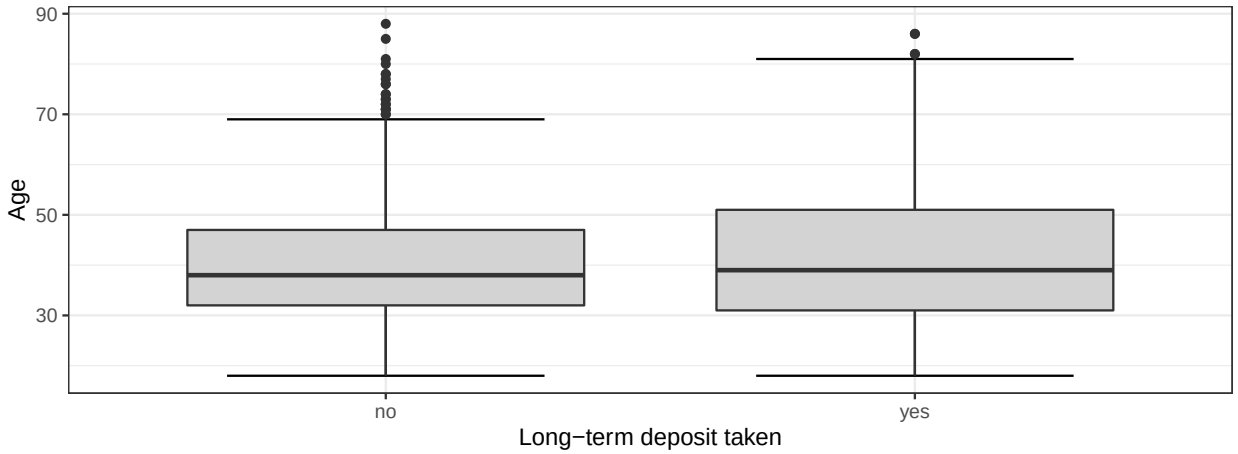
Rysunek 8: New Figure

```
## Error in dimnames(x) <- dnx: 'dimnames()' zastosowane do nie-tablicy
```

2.5



Rysunek 9: New Figure

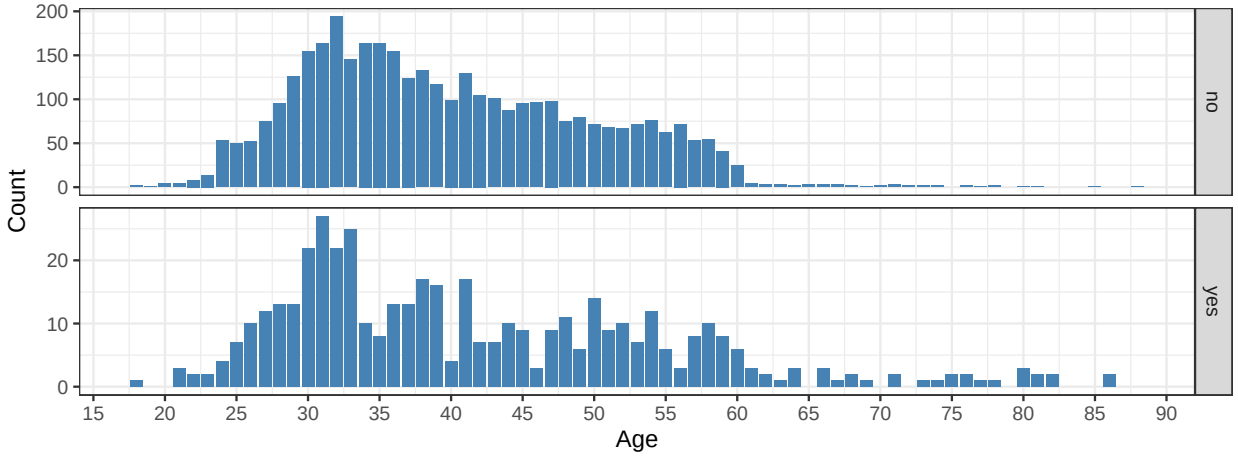


Rysunek 10: New Figure

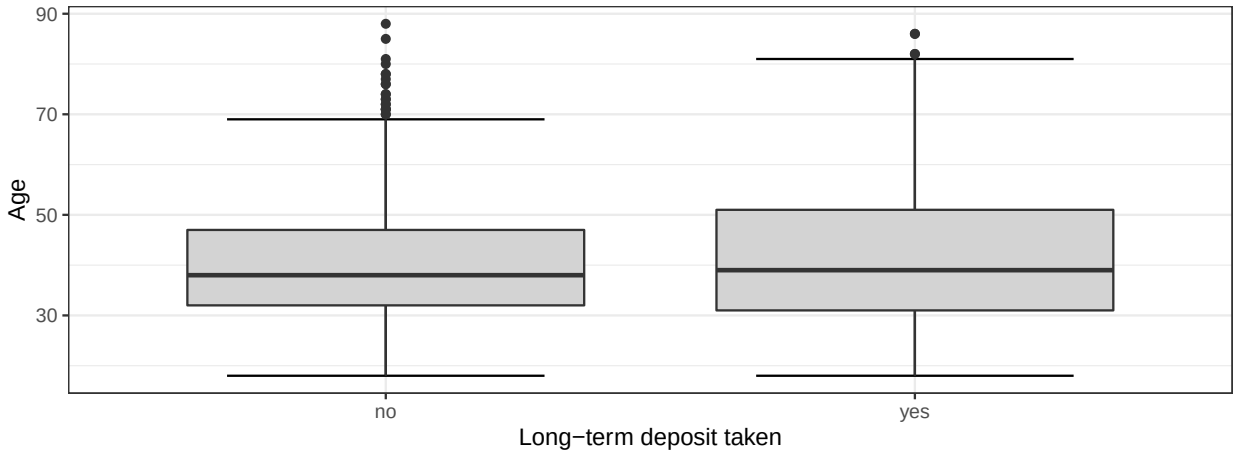
```
## Error in dimnames(x) <- dnx: 'dimnames()' zastosowane do nie-tablicy
```

2.6

```
## Error in dimnames(x) <- dnx: 'dimnames()' zastosowane do nie-tablicy
```



Rysunek 11: New Figure



Rysunek 12: New Figure