Raport projektu z analizy i wizualizacji danych

Julia Murawska, Weronika Piechaczyk, Mateusz Nowak

Uniwersytet im. Adama Mickiewicza w Poznaniu

Spis treści:

1. Źródło danych

Pobraliśmy dane ze strony kaggle.com, których właścicielem jest fedesoriano. Można je tam znaleźć pod nazwą „Stroke Prediction Dataset" lub po nazwie właściciela zbioru. Pochodzą one z 26.01.2021 roku i nigdy nie były aktualizowane. Są to dane publiczne.

2. Zawartość danych

Dane składają się z 12 kolumn (z czego pierwsza oznacza id osoby badanej) oraz 5111 wierszy, które odpowiadają odpowiednio: w pierwszym - nazwy kolumn (zmiennych) oraz pozostałym - odpowiedzi osób badanych. Zmienne, które postanowiliśmy wykorzystać to:

- gender - płeć
- age - wiek
- hypertension - nadciśnienie
- heart disease - choroby serca
- avg glucose level - średni poziom glukozy
- bmi - wskaźnik masy ciała (BMI)
- stroke - udar

Pominęliśmy zmienne: id – indeks, ever_married – czy badany był kiedykolwiek w związku małżeńskim, work_type – rodzaj wykonywanej pracy, residence_type – miejsce zamieszkania (miejskie/wiejskie), smoking status - odpowiedź na pytanie o częstość palenia papierosów.

3. Proces czyszczenia danych

Niektóre dane dotyczące wieku nie były podawane jako liczby całkowite, więc zostały one zaokrąglone. Usunięte zostały wiersze z brakiem danych w zmiennej BMI. Po czyszczeniu zostało 4909 danych do analizy.

4. Cel i kierunek przeprowadzonych analiz

Opis danych dotyczących udaru oraz analiza poszczególnych zmiennych, dzięki której dowiemy się między innymi:

- w jakim wieku są badani,
- jaka jest średnia wieku,
- jakie jest ich najczęstsze BMI,
- jaka jest minimalna i maksymalna wartość poziomu glukozy wśród badanych,
- jaki procent badanych miało udar, a jaki posiadają choroby serca,
- jaki procent respondentów to kobiety a jaki mężczyźni.

Próbujemy odpowiedź na pytania dotyczące korelacji niektórych zmiennych:

1. Czy jest związek problemu nadciśnienia z chorobami serca?
2. Czy choroby serca mają wpływ na udar?

3. Czy wraz z wiekiem średni poziom glukozy jest coraz wyższy?

Chcemy opisać zależności pomiędzy nadciśnieniem a chorobami serca, chorobami serca a udarem oraz wiekiem i średnim poziomem glukozy.

5. Charakterystyka zmiennych

a) Wiek:

Tabela 1: Dane statystyczne dla zmiennej „wiek”

Statystyka:	Wartość [w latach]:
Średnia	43
Dominanta	78
Odchylenie standardowe	26
Odchylenie przeciętne	19
Pierwszy kwartyl	25
Mediana/drugi kwartyl	44
Trzeci kwartyl	60
Wartość minimalna	0
Wartość maksymalna	82
Wariancja	510

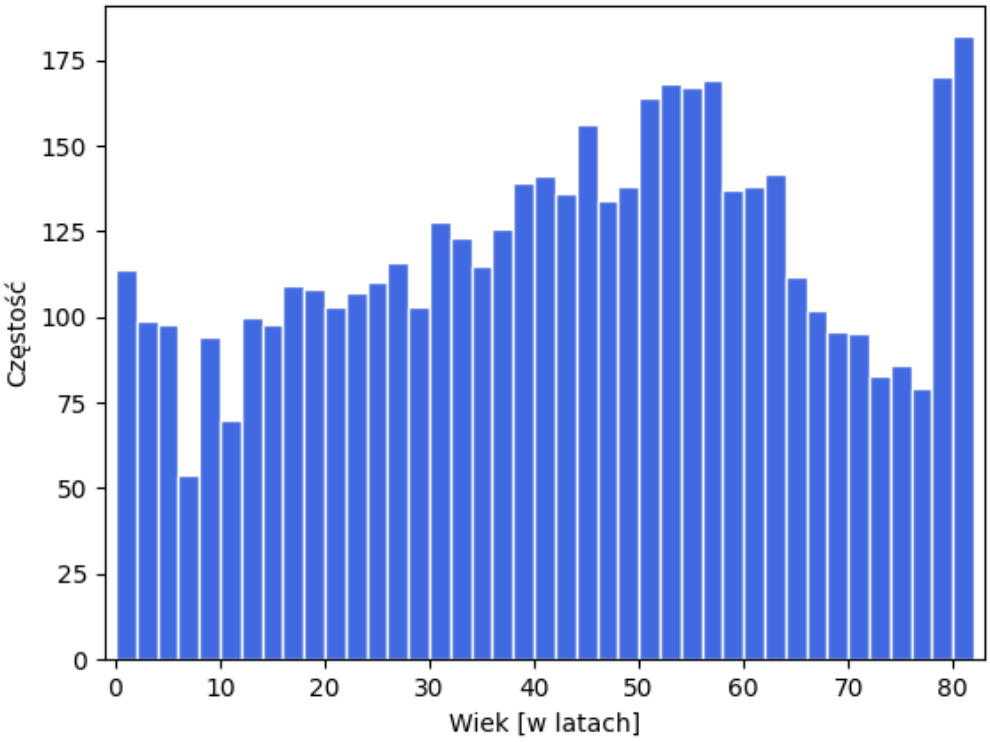
Z tabeli 1, znajdującej się powyżej możemy odczytać, że średnia tej zmiennej w zaokrągleniu to 43 lata, dominanta (a więc najczęściej występująca wartość wieku) wynosi 78. Mediana to 44, czyli połowa badanych miała mniej niż lub dokładnie 44 lata, a druga połowa również tyle samo lub więcej. Odchylenie standardowe wskazuje na duże zróżnicowanie wieku badanych, ponieważ wynosi ono 26 lat. Odchylenie przeciętne wskazuje na to, że wiek zmienia się średnio o 19 lat od średniej

dla przeciętnego badanego. Pierwszy kwartyl równy 25 mówi nam, że 25% badanych ma 25 lub mniej lat, natomiast trzeci, że 25% ma 60 lat lub więcej. Wartość minimalna wynosi 1 rok, a najstarszy badany miał 82 lata. Wariancja, która wynosi 510, mówi nam o dużym zróżnicowaniu wieku wśród osób badanych.

Tabela 2: Rozkład zmiennej „wiek”

Miara:	Wartość:
Skośność	-0.12
Test skośności	-3.49, alfa ≈ 0.00
Kurtoza	-0.98
Test kurtozy	-31.89, alfa ≈ 0.00

Skośność bliska 0 wskazuje na brak asymetrii wyników. Ze względu na ujemną kurtozę stwierdzamy, że rozkład jest platykurtyczny, co też możemy zaobserwować na poniższym wykresie.



Wykres 1: Histogram zmiennej „wiek”

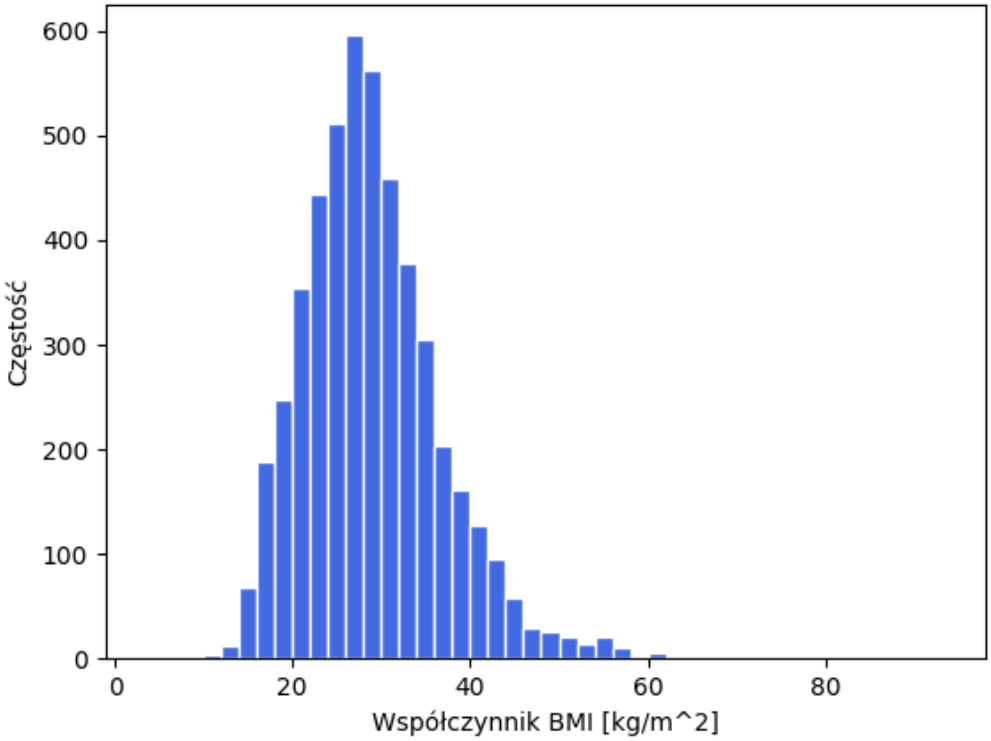
b) BMI:

Tabela 3: Dane statystyczne dla zmiennej „BMI”

Statystyka:	Wartość [w kg/m²]
Średnia	29
Dominanta	29
Odchylenie standardowe	8
Odchylenie przeciętne	6
Pierwszy kwartyl	24
Mediana/drugi kwartyl	28
Trzeci kwartyl	33
Wartość minimalna	10

Wartość maksymalna	98
Wariancja	62

W tabeli 3 można zauważyć, że średnia w zaokrągleniu to 29 kg/m², dominanta, czyli najczęstsza wartość wynosi również 29. Mediana to 28, czyli połowa badanych miała mniej niż lub dokładnie tyle, a druga połowa również tyle samo lub więcej. Odchylenie standardowe wskazuje na małe zróżnicowanie BMI badanych, ponieważ wynosi ono 8 jednostek. Odchylenie przeciętne wskazuje na to, że BMI różni się średnio o 6 jednostek od średniej dla przeciętnego badanego. Pierwszy kwartyl równy 24 mówi nam, że 25% badanych ma wynik BMI 24 lub mniejszy, natomiast trzeci, że 25% ma 33 lub więcej. Wartość minimalna wynosi 10 jednostek, a wartość maksymalna 98. Wariancja, która wynosi 62, mówi nam o małym zróżnicowaniu BMI wśród osób badanych.



Wykres 2: Histogram zmiennej „BMI”

Tabela 4: Rozkład zmiennej ,,BMI

Miara:	Wartość:
Skośność	1.05
Test skośności	25.13, alfa ≈ 0.00
Kurtoza	3.36
Test kurtozy	19.74, alfa ≈ 0.00

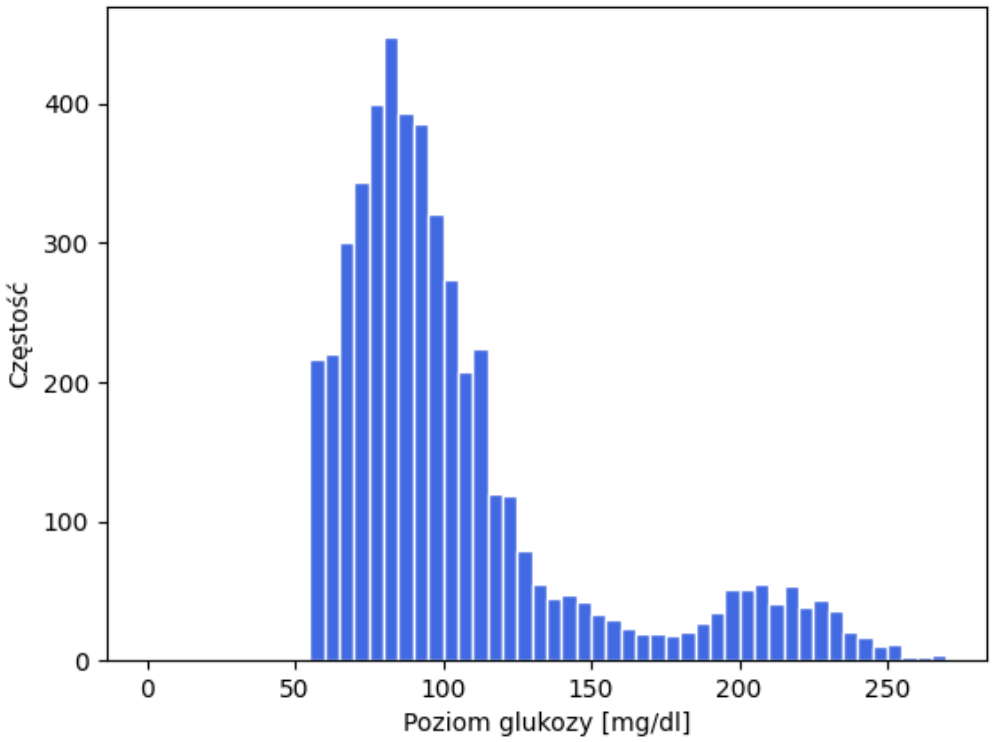
Skośność większa od 0 wskazuje na to, że wykres jest lewoskośny. Ze względu na dodatnią kurtozę stwierdzamy, że rozkład jest leptokurtyczny.

c) Poziom glukozy:

Tabela 5: Dane statystyczne dla zmiennej ,,średni poziom glukozy”

Statystyka:	Wartość [w mg/dl]
Średnia	105
Dominanta	94
Odchylenie standardowe	44
Odchylenie przeciętne	32
Pierwszy kwartyl	77
Mediana/drugi kwartyl	92
Trzeci kwartyl	114
Wartość minimalna	55
Wartość maksymalna	272
Wariancja	1974

Tabela 5 pokazuje, że średnia w zaokrągleniu to 105 mg/dl, dominanta, czyli najczęstsza wartość wynosi 94. Mediana to 92, czyli połowa badanych miała mniej niż lub dokładnie tyle, a druga połowa również tyle samo lub więcej. Odchylenie standardowe równe 44 wskazuje na duże zróżnicowanie badanych, ponieważ wynosi ono jednostek. Odchylenie przeciętne, wskazuje na to, że poziom glukozy różni się średnio o 32 jednostek od średniej dla przeciętnego badanego. Pierwszy kwartyl mówi nam, że 25% badanych ma wynik 77 lub mniejszy, natomiast trzeci, że 25% ma 114 lub więcej. Wartość minimalna wynosi 55 jednostek, a wartość maksymalna 272. Wariancja, która wynosi, 1974 mówi nam o dużym zróżnicowaniu wśród osób badanych.



Wykres 3: Histogram zmiennej „średni poziom glukozy”

Tabela 6: Rozkład zmiennej „średni poziom glukozy”

Miara:	Wartość:
Skośność	1.61

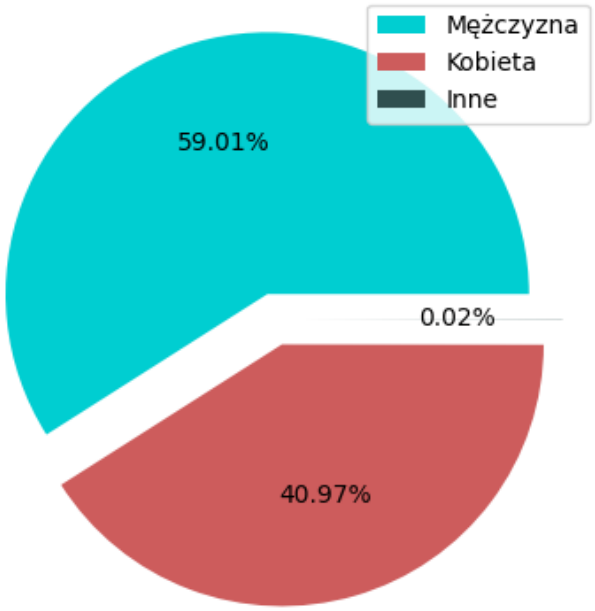
Test skośności	33.55, alfa ≈ 0.00
Kurtoza	1.91
Test kurtozy	14.75, alfa ≈ 0.00

Skośność większa od 0 wskazuje na to, że wykres jest lewoskośny. Ze względu na dodatnią kurtozę stwierdzamy, że rozkład jest leptokurtyczny.

d) Płeć:

Tabela 7: Dane zmiennej “płeć

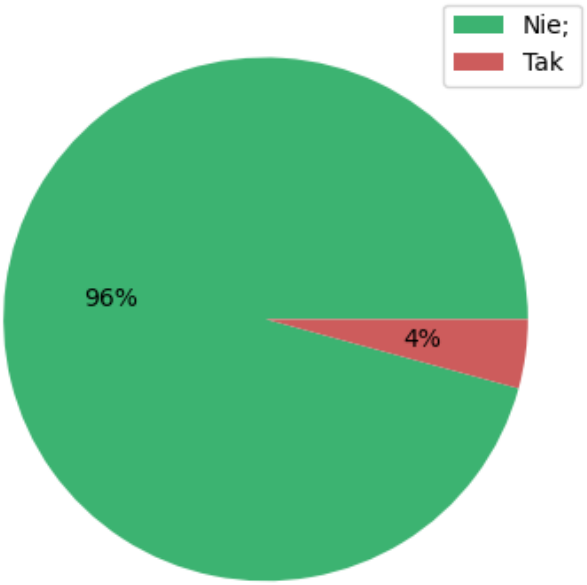
Liczba kobiet	Liczba mężczyzn	Inne
2897	2011	1



Wykres 4: Wykres kołowy zmiennej „płeć”

Liczba kobiet, które udzieliły odpowiedzi to 2897, co stanowi 59.01% wszystkich respondentów, a mężczyźni, których było 2011, to 40.97%. Jedna osoba udzieliła odpowiedzi “inne” i stanowi 0.02%.

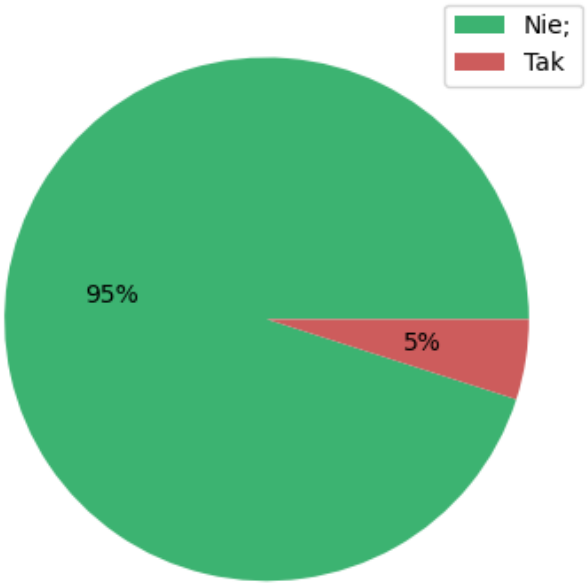
e) Udar:



Wykres 5: Wykres kołowy zmiennej „udar”

Wiemy, że badani na pytanie o udar udzielali jednej z dwóch odpowiedzi „tak” - jeśli kiedykolwiek go mieli, „nie” - jeśli nigdy go nie doświadczyli. Dominantą w tym przypadku jest odpowiedź „nie”. Większość odpowiadających, czyli 96% nie miało udaru, natomiast 4% miało.

f) Choroby serca:

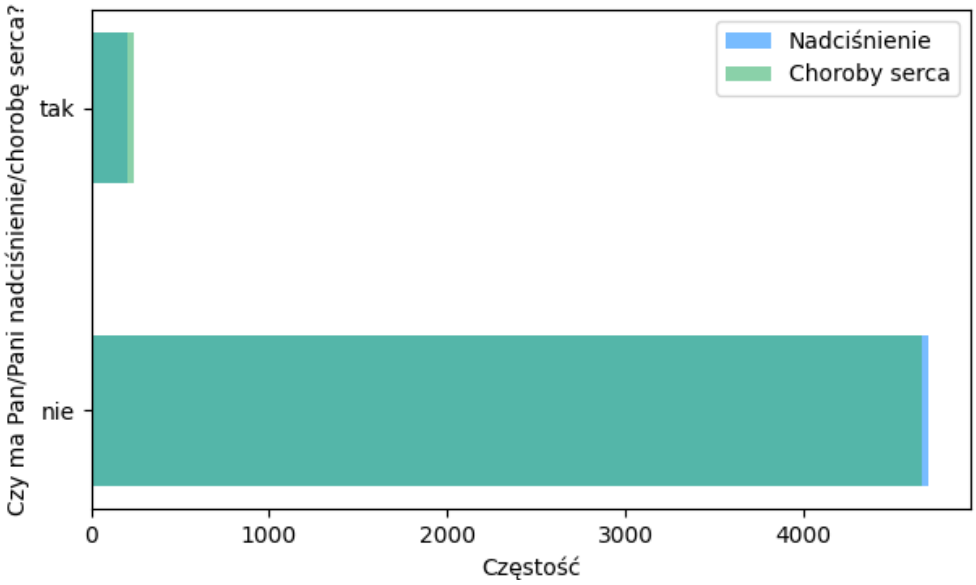


Wykres 6: Wykres kołowy zmiennej „choroby serca”

Wiemy, że badani również na pytanie o choroby serca udzielali jednej z dwóch odpowiedzi „tak” - jeśli jakieś mają, „nie” - w przeciwnym przypadku. Dominantą jest odpowiedź „nie”. Większość respondentów – 95% nie miało udaru, natomiast 5% je ma.

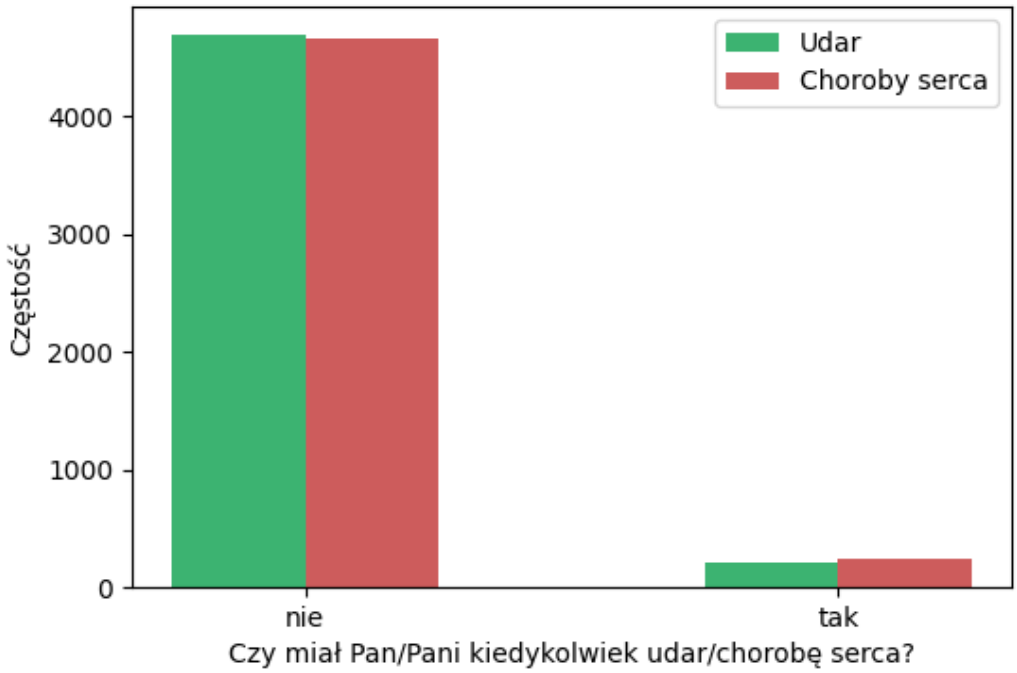
6. Charakterystyka między zmiennymi.

a) Zmienne: nadciśnienie i choroby serca



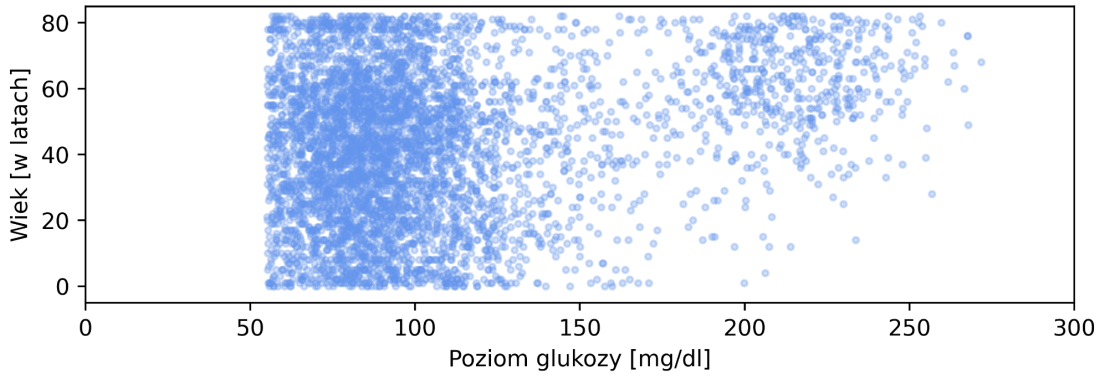
Wykres 7: Podwójny wykres słupkowy zmiennych „nadciśnienie” i “choroby serca”

b) Zmienne: udar i choroby serca



Wykres 8: Podwójny wykres słupkowy zmiennych „udar” i “choroby serca”

c) Zmienne: wiek i średni poziom glukozy



Wykres 9: Wykres rozrzutu zmiennych „wiek” i “średni poziom glukozy”

Tabela 8: Macierz korelacji Pearsona między zmiennymi

	wiek	nadciśnienie	choroby serca	poziom glukozy	bmi	udar
wiek	1.00	0.27	0.26	0.24	0.33	0.23
nadciśnienie	0.27	1.00	0.16	0.18	0.17	0.14
choroby serca	0.26	0.16	1.00	0.15	0.04	0.14
poziom glukozy	0.24	0.18	0.15	1.00	0.18	0.14
bmi	0.33	0.17	0.04	0.18	1.00	0.04
udar	0.23	0.14	0.14	0.14	0.04	1.00

8. Wnioski

Z wykresu 7 wynika, że większość respondentów nie posiada żadnej z tych chorób, ale więcej osób ma choroby serca niż nadciśnienie.

Jeśli chodzi o porównanie danych dotyczących udaru i chorób serca, w wykresie 8 również większość udzieliła odpowiedzi negująco na oba pytania, ale można zauważyć, że także więcej osób posiada choroby serca, niż miało kiedykolwiek udar.

Natomiast wykres opisujący zależność pomiędzy wiekiem, a średnim poziomem glukozy z wykresu 9 pozwala odpowiedzieć twierdząco na pytanie „Czy wraz z wiekiem średni poziom glukozy jest coraz wyższy?". Można wywnioskować z tego, że częściej na cukrzycę chorują osoby starsze.

Z tabeli 8 możemy zauważyć, że wszystkie korelacje pomiędzy naszymi zmiennymi są słabe (mimo swojej istotności).