Projekt Zaliczeniowy – Inżynieria Oprogramowania

Weronika Kozłowska, Michał Zielonka

Informatyka i Ekonometria, N52-12

Projekt zaliczeniowy opracowany został wspólnie przez wyżej wymienione osoby. Źródło wykorzystanych tabel, wykresów oraz zrzutów ekranu stanowiło opracowanie własne w programach R oraz MS Excel na podstawie wykorzystywanych danych.

Charakterystyka problemu

Choroby sercowo-naczyniowe (CVD) są główną przyczyną zgonów na całym świecie, pochłaniając około 17,9 miliona istnień ludzkich każdego roku, co stanowi 31% wszystkich zgonów na świecie. Ten zbiór danych zawiera 6 cech, które można wykorzystać do przewidywania możliwej choroby serca.

Osoby z chorobami układu krążenia lub osoby z wysokim ryzykiem sercowo-naczyniowym potrzebują wczesnego wykrycia choroby, ale również określenia ryzyka zachorowania. W tym celu można wspomagać się zbudowanymi modelami statystycznymi objaśniającymi dane zjawisko. Jest to bardzo pomocne w określaniu **istotnych cech** mających wpływ na zachorowanie, a dodatkowo w określeniu **prawdopodobieństwa** tego, że dany pacjent zachoruje. **Zbudowanie takiego modelu, który będzie to umożliwiać jest głównym celem tego badania.**

Źródło danych

Zbiór danych wykorzystywany w badaniu został utworzony poprzez połączenie różnych zbiorów danych już dostępnych niezależnie. W rozważanym zbiorze danych połączono 5 zbiorów dotyczących osób z obecną (lub nie) **chorobą wieńcową serca** ostatecznie otrzymując **918 przebadanych**. Znajduje się w nim 6 głównych cech, które są podstawowymi czynnikami mogącymi świadczyć o ryzyku wystąpienia choroby wieńcowej. Badanie zostanie opracowane na podstawie poniższych źródeł.

Źródło 1: https://pubmed.ncbi.nlm.nih.gov/2756873/
Źródło 2: https://www.kaggle.com/datasets/fedesoriano/heart-failure-prediction?fbclid=IwAR3WrdfwNXhCAtgbijKFhhyx3T8ue57v33WDZ0eiuYQl4dUjWDoQz6JvKj0

Twórcy zbiorów danych:

- Hungarian Institute of Cardiology. Budapest: Andras Janosi, M.D.
- University Hospital, Zurich, Switzerland: William Steinbrunn, M.D.
- University Hospital, Basel, Switzerland: Matthias Pfisterer, M.D.
- V.A. Medical Center, Long Beach and Cleveland Clinic Foundation: Robert Detrano, M.D., Ph.D.

Rozważane cechy w badaniu

Nasz zestaw danych zawiera 6 cech, które można wykorzystać do przewidywania (objaśniania) możliwej choroby wieńcowej serca (HeartDisease). **Skupimy się na analizie zmiennych: ExerciseAngina, MaxHR, Cholesterol, RestingBP, Sex i Age w kontekście zmiennej HeartDisease.**

Tabela 1.

Zmienna	Opis
ExerciseAngina	Dławica wywołana wysiłkiem (Y = tak, N = nie)
MaxHR	Osiągalne maksymalne tętno (uderzenia na minutę)
Cholesterol	Poziom cholesterolu [mg/dl] (Miligramm Per Deciliter)
RestingBP	Spoczynkowe ciśnienie krwi [mm Hg]
Sex	Płeć
Age	Wiek
HeartDisease	Występowanie choroby serca (0 = nie, 1 = tak)

Uwaga: Po głębszej analizie zbioru danych i jego źródłach, zauważono i zweryfikowano, że zmienna "Cholesterol" na stronie Kaggle wpisaną ma błędną jednostkę w opisie zmiennych. **Zamiast mm/dl powinno być mg/dl** - jednostka powszechnie stosowana do mierzenia cholesterolu.

Potrzebne biblioteki

W celu przeprowadzenia badania, należy wgrać odpowiednie biblioteki do programu R:

- ggplot2
- corrplot
- readr
- car

Import danych

Przed rozpoczęciem wgrano zbiór danych 'heart.csv'.

Przygotowanie typu zmiennych

Należy zmienić **typ zmiennych na kategorialny** dla Sex, ExerciseAngina oraz HeartDisease, ponieważ są to **zmienne dychotomiczne**, czyli takie, które przyjmują tylko dwie wartości.
Dla reszty zmiennych zadbano o to, aby przyjmowały typ zmiennych jako **numeryczny**.

Wstępna weryfikacja poprawności danych

Zgodnie ze wcześniejszą informacją, która została umieszczona w sekcji „Rozważane zmienne w badaniu", skupimy się na analizie zmiennych: ExerciseAngina, MaxHR, Cholesterol, RestingBP, Sex i Age w kontekście zmiennej HeartDisease.

Na początek spójrzmy na podstawowe statystyki naszego zbioru danych.

Tabela 2.

Statystyki	MaxHR	Cholesterol	RestingBP	Age
Minimum	60	0	0	28
1 kwartyl	120	173,2	120	47
Mediana	138	223	130	54
Średnia	136,8	198,8	132,4	53,51
3 kwartyl	156	267	140	60
Maximum	202	603	200	77

Tabela 3.

Kobieta/Mężczyzna	Sex
F	193
M	725

Nie występuje/Występuje	ExerciseAngina
N	547
Y	371

Nie występuje/Występuje	HeartDisease
0	410
1	508

Po weryfikacji danych, można zauważyć, że **występują jednostki, które przyjmują wartości nieosiągalne dla żywego człowieka**. Należy zwrócić uwagę na to, że w wartościach minimalnych w powyższej tabeli dla zmiennych **RestingBP i Cholesterol występują zera**.

Ponieważ są to zmienne odpowiadające za poziom cholesterolu [mg/dl] oraz za spoczynkowe ciśnienie krwi [mm Hg], to **wartość 0 jest niepoprawna**. W związku z tym, należy je **zidentyfikować** i następnie **usunąć** z naszego zbioru danych.

W programie **zidentyfikowaliśmy** wszystkie przypadki, w których dla zmiennej Cholesterol widnieje **wartość 0**. Sprawdzona została również liczba takich przypadków i wynosi ona **172**.

Analogicznie zidentyfikowaliśmy jednostki, w których zmienna **RestingBP** przyjmuje **wartość 0**.

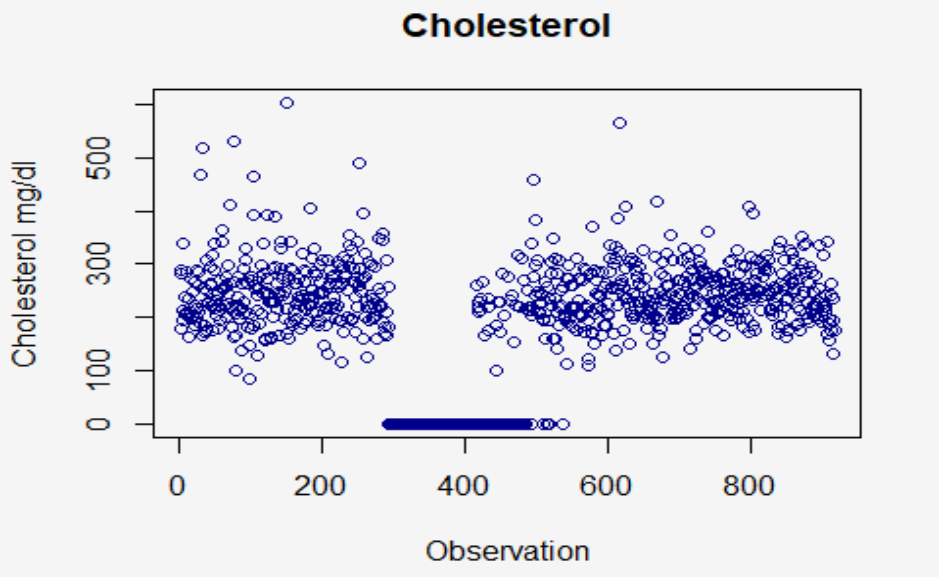
Widać, że jest to jeden przypadek, który jednocześnie ma przypisaną wartość 0 również dla zmiennej Cholesterol. **W takim razie łączna liczba jednostek, które usuniemy wynosi 172**.

Wielkość próby po usunięciu danych wynosić będzie: **746**.

W rzeczywistym badaniu należałoby przyjrzeć się tym nieprawdopodobnym obserwacjom i zastosować inny schemat postępowania, aby nie stracić części ważnych informacji przy usuwaniu obserwacji. Na potrzeby projektu możemy usunąć te obserwacje.

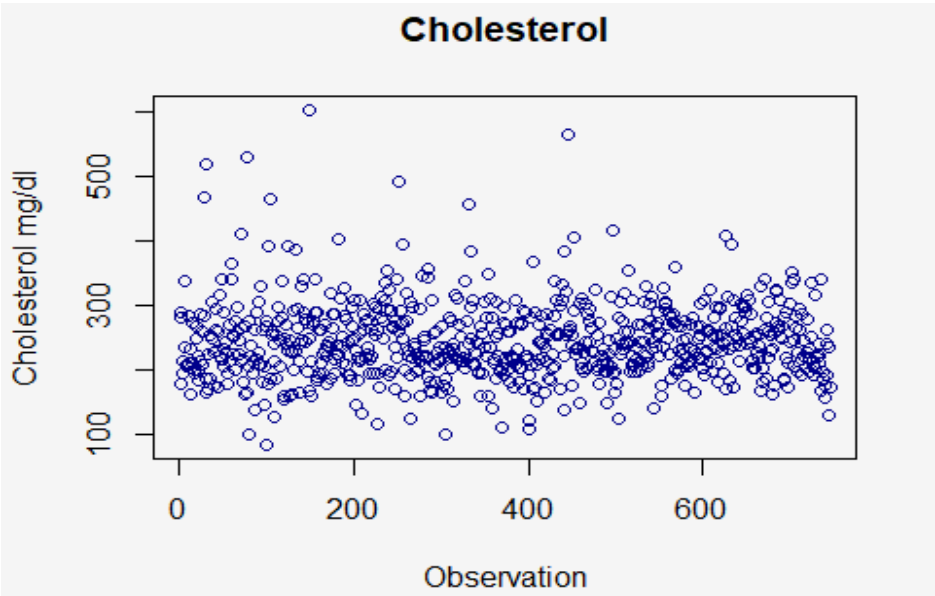
Sprawdźmy, jak wygląda rozrzut zmiennej Cholesterol przed usunięciem danych.

Wykres 1.



Na powyższym wykresie wyraźnie widać jednostki przyjmujące wartość 0, które zidentyfikowaliśmy do usunięcia. Po usunięciu wyżej wspomnianych jednostek spójrzmy na wykres raz jeszcze.

Wykres 2.



Obserwacje zostały usunięte poprawnie. Zauważyć można również “pas” pomiędzy wartościami 180mg/dl - 300mg/dl, w którym skupia się stosunkowo najwięcej obserwacji.

Analiza wstępna

Spójrzmy, jak wyglądają podstawowe statystyki naszego zbioru danych po usunięciu obserwacji i czym charakteryzują się poszczególne zmienne.

Tabela 4.

Statystyki	MaxHR	Cholesterol	RestingBP	Age
Minimum	69	85	92	28
1 kwartyl	122	207,2	120	46
Mediana	140	237	130	54
Średnia	140,2	244,6	133	52,88
3 kwartyl	160	275	140	59
Maximum	202	603	200	77

Tabela 5.

Kobieta/Mężczyzna	Sex
F	182
M	564

Nie występuje/ Występuje	ExerciseAngina
N	459
Y	287

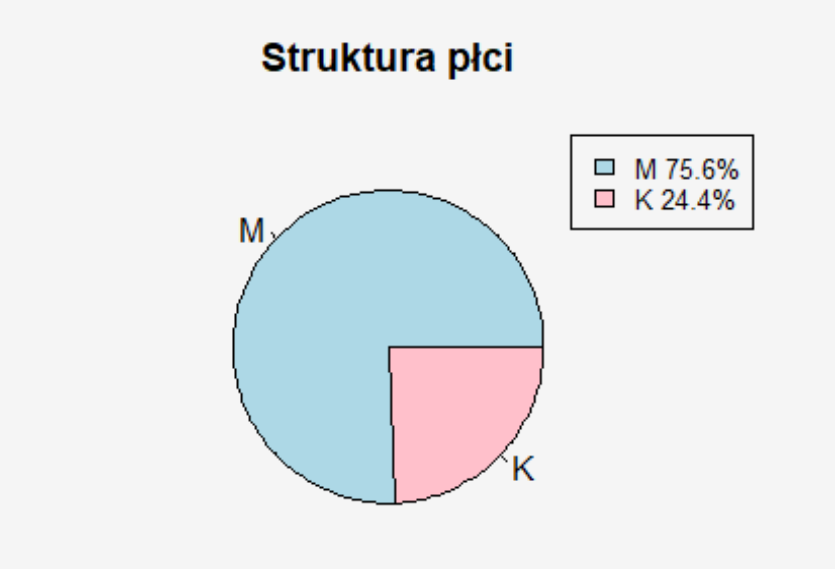
Nie występuje/ Występuje	HeartDisease
0	390
1	356

Analizę powyższych wyników przedstawiliśmy poniżej zajmując się każda zmienną w osobnej sekcji.

Sex - Płeć

Warto zwrócić uwagę na strukturę płci

Wykres 3.



Udział mężczyzn w całej próbie to 75.6% (kolor niebieski), natomiast kobiet 24.4% (kolor różowy).

Age - Wiek

W celu zwizualizowania liczebności jednostek w zależności od płci, przydzieliliśmy każdą obserwację do poniższych kategorii wiekowych:

- 28 – 41 lat (<42)
- 42 – 53 lat (42 - 53)
- 54 -60 lat (54 – 60)
- Powyżej 60 lat (>60)

Uwaga: W nawiasach powyżej zapisano notację, z której skorzystaliśmy do odpowiedniego oznaczenia grup wiekowych na wykresie.

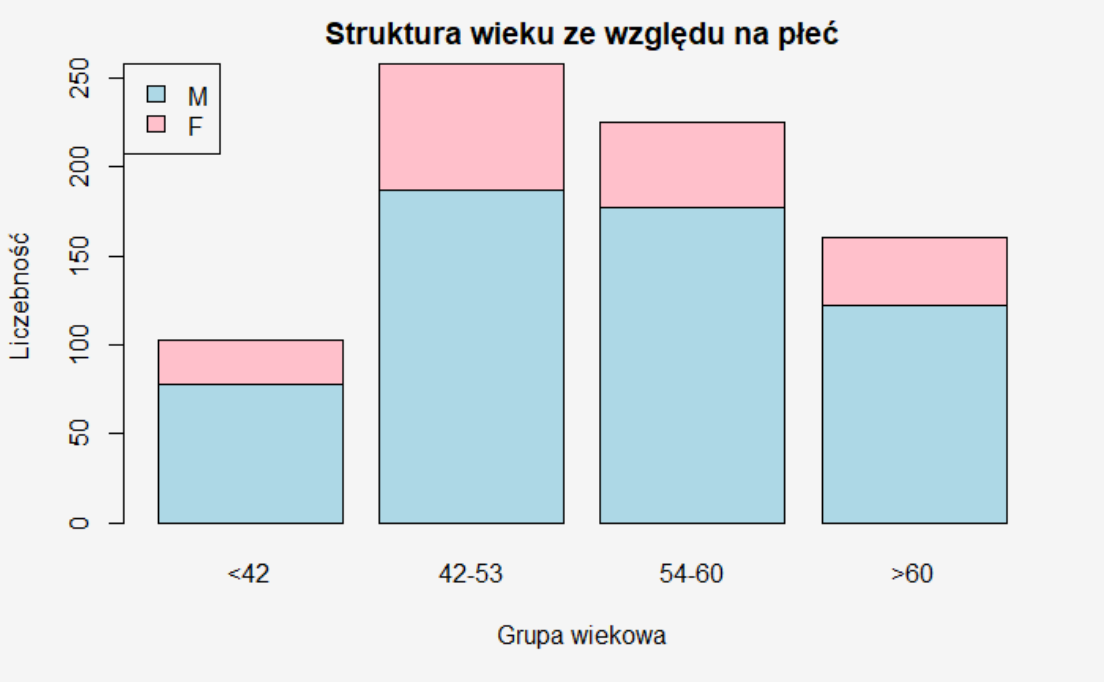
Stworzyliśmy macierz danych, obrazującą liczebności jednostek w poszczególnej grupie wiekowej w zależności od płci.

Tabela 6.

Mężczyzna/Kobieta	<42	42-53	54-60	>60
M	78	187	177	122
F	25	71	48	38
Razem	103	258	225	160

Wizualizujemy uzyskane dane.

Wykres 4.



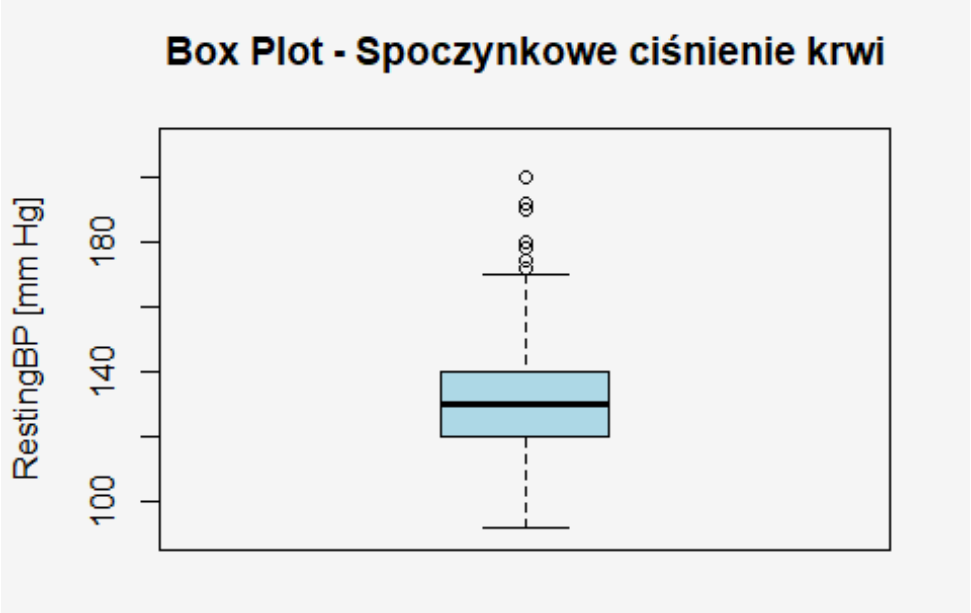
Można zaobserwować, że najwięcej badanych jest w wieku 42-53 lat, co ma miejsce również, gdy podzielimy liczebność ze względu na płeć. Można także zauważyć, że najmniej liczna grupa to "<42" lata.

Ważna uwaga: Najmłodsza badana osoba jest w wieku 28 lat, co wydaje się być odległe od 42-latka będącego w tej samej grupie, natomiast **podział wiekowy został zaproponowany przez autorów** projektu i jest stworzony tylko w celach zaprezentowania analizy wstępnej - nie będzie miało to znaczenia w późniejszej analizie właściwej.

RestingBP – Spoczynkowe ciśnienie krwi [mm Hg]

Zobaczmy, jak prezentuje się wykres ramka-wąsy dla ciśnienia krwi.

Wykres 5.

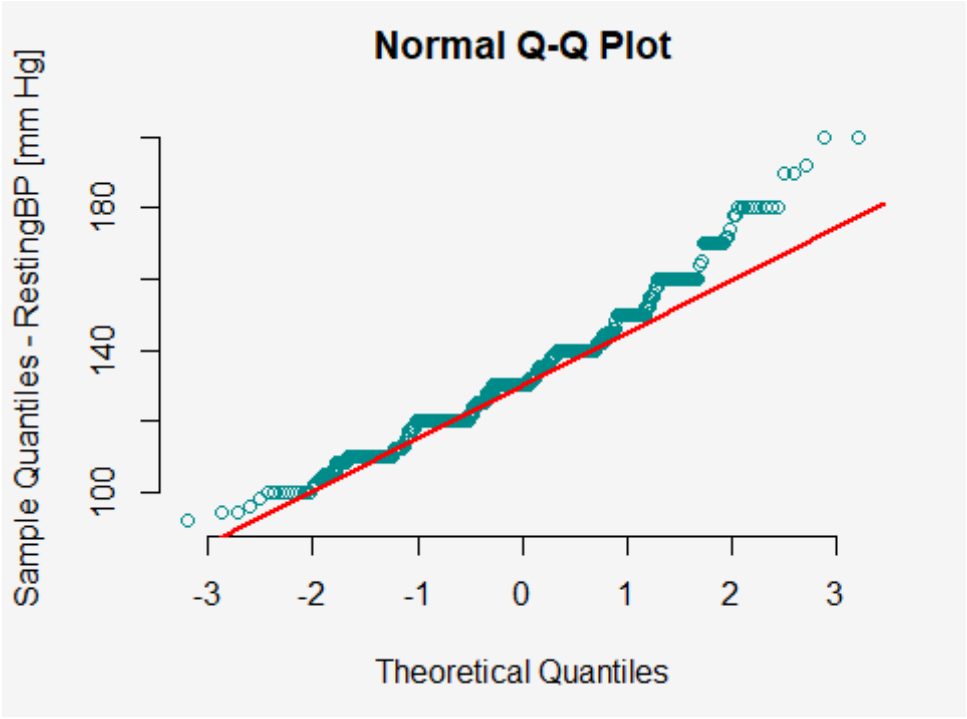


Możemy zaobserwować jednostki, które znajdują się **nad górnym "wąsem"**, czyli poza wartością rozstępu międzykwartylowego przemnożonego przez 1.5 i po dodaniu do tego wartości trzeciego kwartyla. Wyróżnia je wyjątkowo wysokie ciśnienie krwi. Jednostki te zidentyfikowane zostały w programie R.

Warto zwrócić uwagę, że **75%** (15 obserwacji) z 20 pokazanych, to osoby będące **w dwóch najstarszych przedziałach wiekowych**.

Spójrzmy jak wygląda **wykres kwantylowy**, który służy do sprawdzenia dopasowania teoretycznego rozkładu (normalnego) do zaobserwowanych danych. Typ wykresu kwantyl-kwantyl to jeden z najczęściej wykorzystywanych wykresów do **graficznego testowania normalności rozkładu**.

Wykres 6.

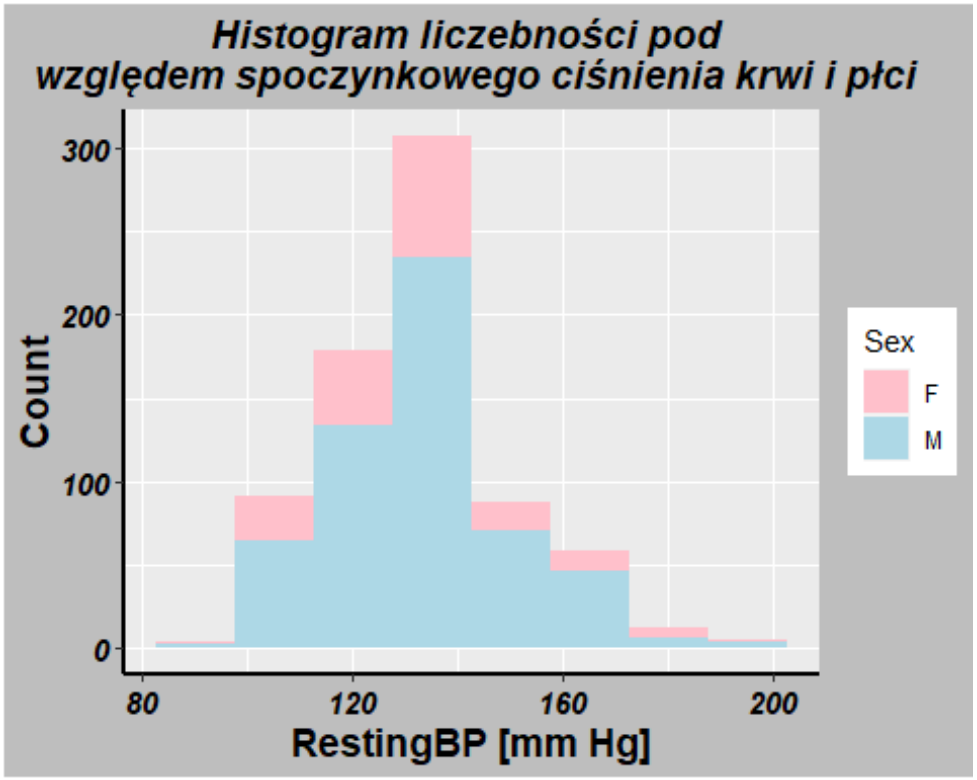


Z powyższego wykresu możemy wywnioskować, że rozkład zmiennej może charakteryzować się trochę **"dłuższym prawym ogonem"**, co oznacza, że będą występować jednostki osiągające wyższe wartości **- zauważyć można niedopasowanie do czerwonej linii szczególnie po prawej stronie wykresu**.

Potwierdza to również wykres Box Plot, na którym widać, że jednostki nietypowe występują tylko dla wysokich wartości. **Gdyby punkty danych znajdowały się idealnie na pokazanej linii, oznaczałoby to, że rozkład zmiennej jest dobrze dopasowany do rozkładu teoretycznego - normalnego.**

W celu weryfikacji tych obserwacji spójrzmy jak przedstawia się histogram zmiennej RestingBP.

Wykres 7.



Każdy wydzielony słupek bierze pod uwagę zakres 15 kolejnych wartości spoczynkowego ciśnienia krwi. Należy pamiętać, że wykres przedstawia zgrupowaną liczność badanych w każdym przedziale wyróżniając kobiety (kolor różowy) i mężczyzn (kolor niebieski).

Przykładowo w słupku obrazującym grupę, w której spoczynkowe ciśnienie krwi wynosi 128 mm Hg - 142 mm Hg widzimy, że badanych jest najwięcej w obydwóch kategoriach - mężczyźni i kobiety.

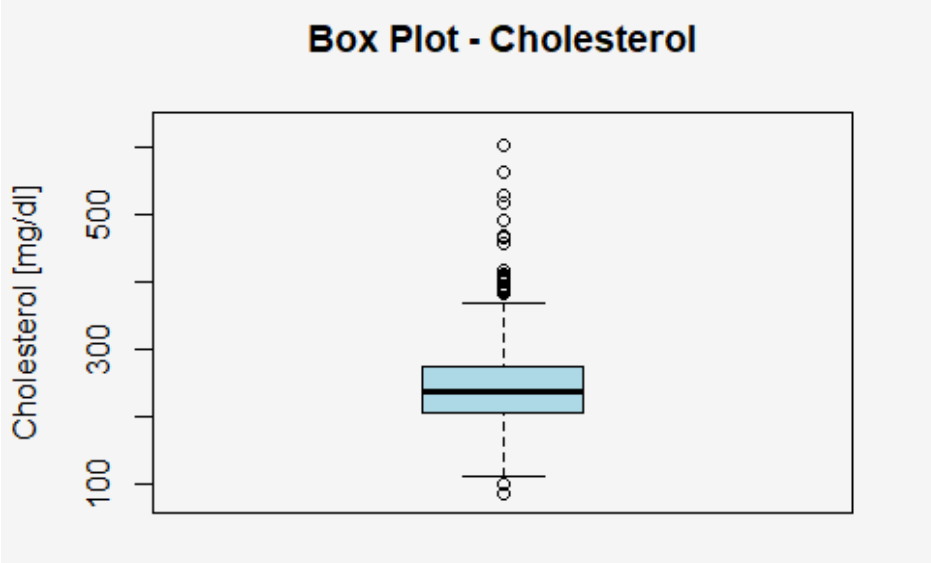
Widzimy też **małą liczbę osób o skrajnych wartościach ciśnienia w prawym "ogonie"** rozkładu.
Są ta najpewniej jednostki pokazane poprzednio na wykresie Box Plot jako wartości skrajne.

Wykonanie histogramu utwierdziło nas również w przypuszczeniu, które mieliśmy po zwizualizowaniu wykresu kwantylowego.

Cholesterol

Spójrzmy, jak przedstawia się wykres Box Plot zmiennej Cholesterol.

Wykres 8.



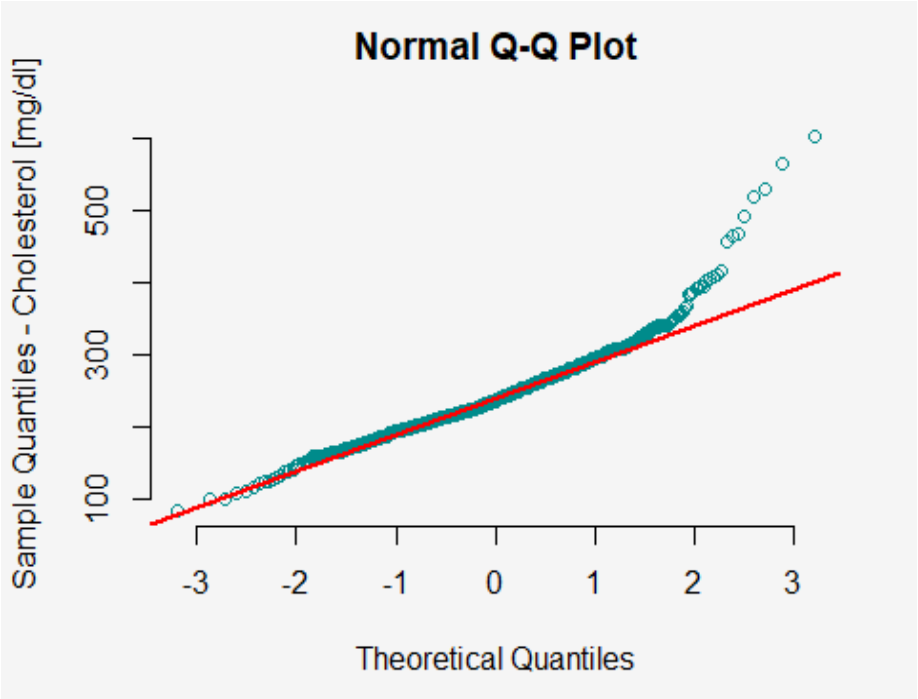
W tym przypadku możemy zaobserwować jednostki, które znajdują się **nad górnym jak i pod dolnym "wąsem"**, czyli poza wartością rozstępu międzykwartylowego przemnożonego przez 1.5 i po dodaniu/odjęciu do tego wartości trzeciego/pierwszego kwartyla.

Wyróżnia je wyjątkowo wysoki/niski poziom cholesterolu odpowiednio dla danej grupy jednostek.

Jednostki skrajnie wysokie i skrajnie niskie zostały zidentyfikowane w programie R.

Spójrzmy, jak wygląda wykres kwantylowy, aby sprawdzić graficzne dopasowanie normalności rozkładu.

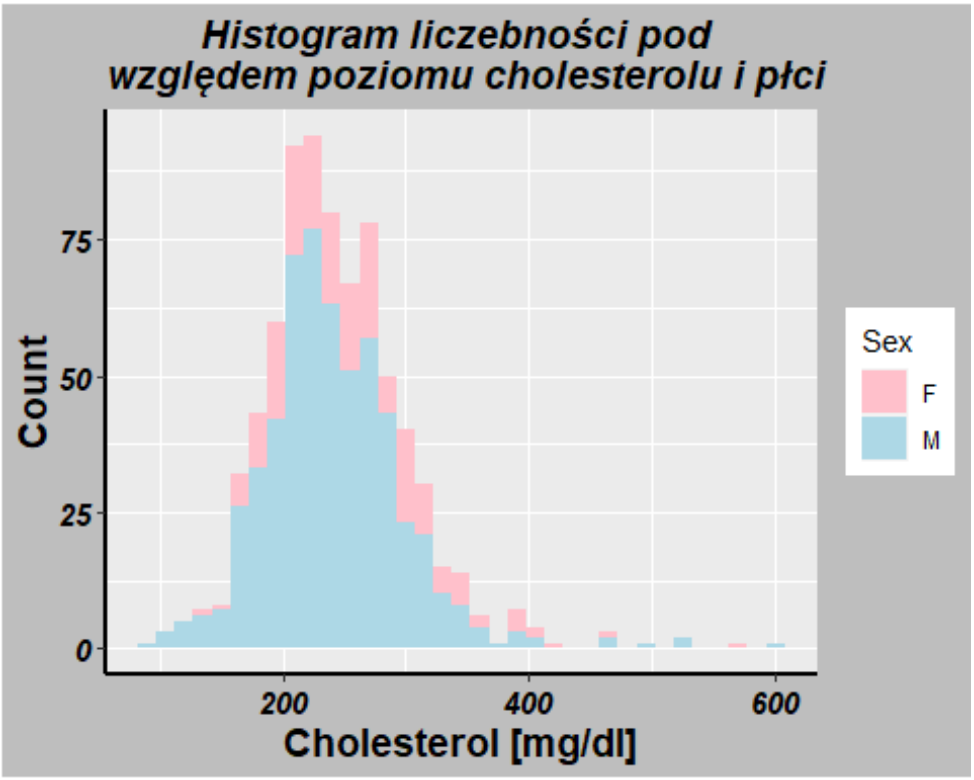
Wykres 9.



Dla zmiennej Cholesterol widzimy znacznie **lepsze dopasowanie**, niż było to w przypadku zmiennej RestingBP. **Większość punktów danych pokrywa się z czerwoną linią**, co oznacza dopasowanie do rozkładu teoretycznego. Widać jednak wyraźne **odstępstwo po prawej stronie wykresu** - punkty są zaznaczone nad linią, co oznacza, że rozkład tej zmiennej będzie **mocno prawoskośny**.

W celu weryfikacji tych obserwacji spójrzmy, jak przedstawia się histogram zmiennej Cholesterol.

Wykres 10.



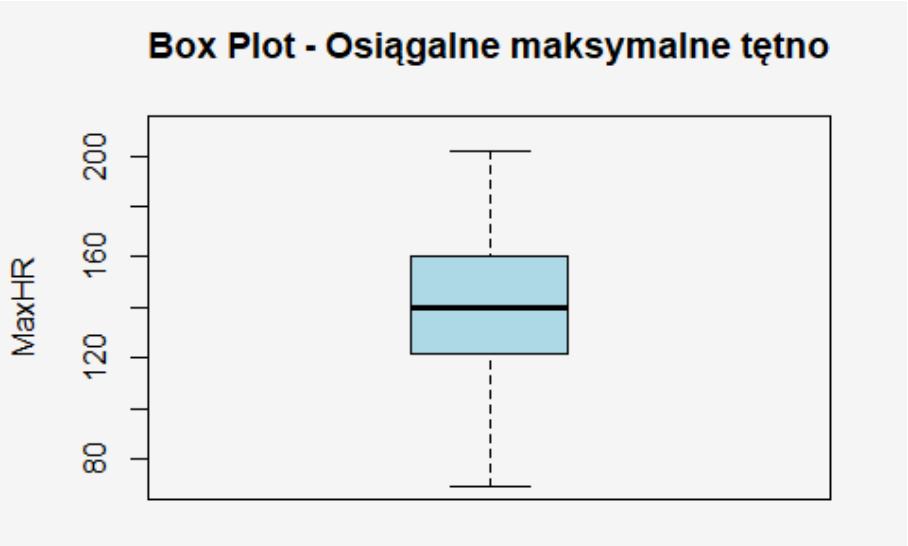
Każdy wydzielony słupek bierze pod uwagę zakres 15 kolejnych wartości poziomu cholesterolu. Można zauważyć, że w przedziale od 200 mg/dl do 250 mg/dl skupionych jest najwięcej obserwacji. Widzimy też, że **osoby o skrajnie niskim poziomie cholesterolu to mężczyźni** (pierwsze 3 niebieskie słupki od lewej).

Potwierdza to również, poprzednia identyfikacja jednostek nietypowych na podstawie wykresu pudełkowego. **Widoczny jest tu również "ogon"** z wartościami sięgającymi aż do 600 mg/dl, co wydedukowaliśmy po wykresie kwantyl-kwantyl.

MaxHR – Osiągalne maksymalne tętno (uderzenia na minutę)

Spójrzmy jak przedstawia się wykres Box Plot zmiennej MaxHR.

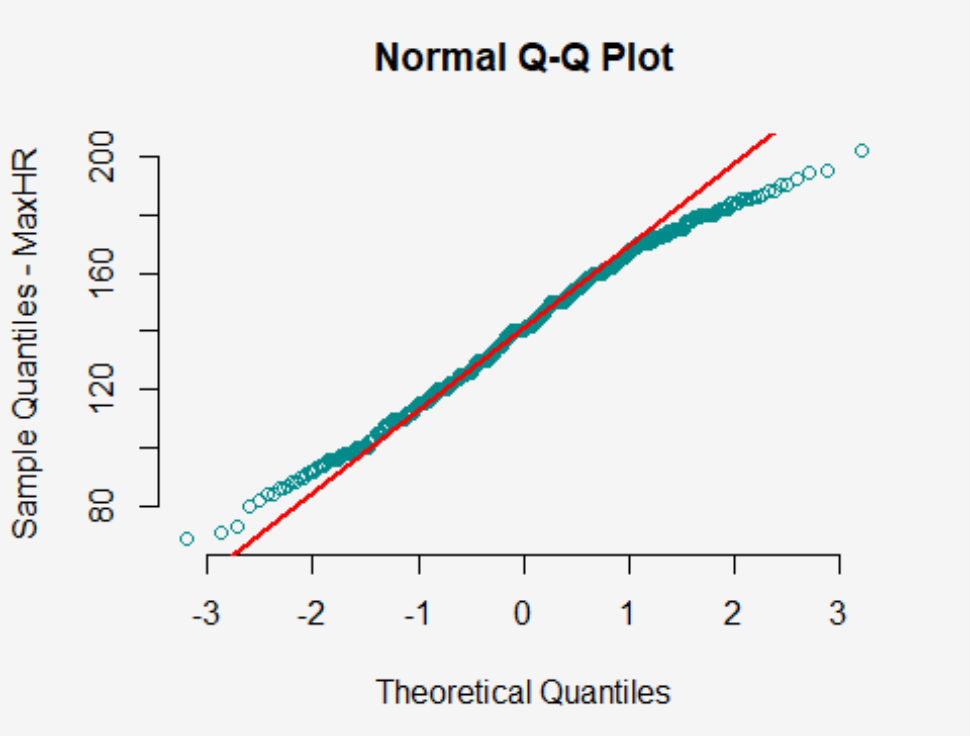
Wykres 11.



Powyższy wykres wskazuje na **symetryczny rozkład tej zmiennej**. Średnia (140,23) i mediana (140) są bardzo zbliżone. **Wizualizacja nie wykazuje jednostek nietypowych.**

Zobaczmy jak wygląda dopasowanie rozkładu zmiennej MaxHR do rozkładu normalnego.

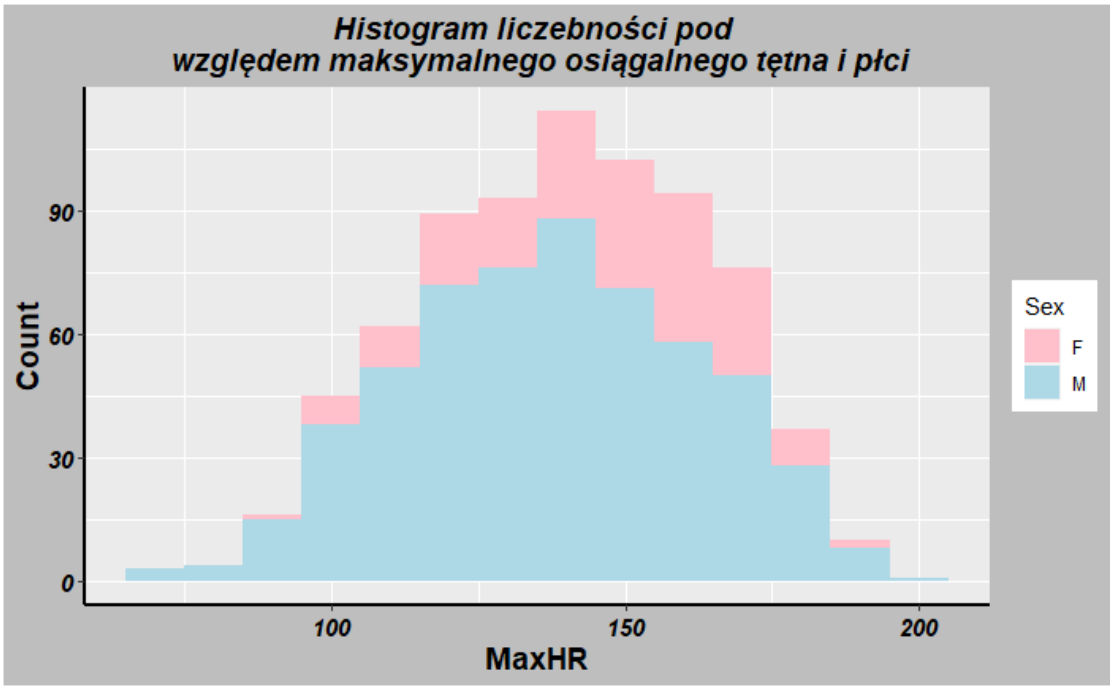
Wykres 12.



W **środkowej części** tego wykresu **dopasowanie** do rozkładu teoretycznego jest **bardzo dobre**, natomiast **na obu końcach** widać wyraźne **odstępstwa od czerwonej linii**.

Po lewej stronie punkty danych zaznaczone są ponad linią, natomiast po prawej stronie wykresu punkty są poniżej prostej. **Oznacza to, że obserwacje dla zmiennej MaxHR będą znajdywały się znacznie bliżej średniej po obu końcach rozkładu, niż wyglądałoby to w przypadku rozkładu teoretycznego**.

Wykres 13.



Tak jak wskazywał wykres pudełkowy, rozkład tej zmiennej wydaje się być **zbliżony do rozkładu symetrycznego**. **Nie widać "długich ogonów**" o małej liczebności, które wskazywałyby na występowanie jednostek skrajnych i obecność rozkładu skośnego.

Potwierdziliśmy również przypuszczenia dokonane po zwizualizowaniu wykresu kwantylowego. **Obserwacje są mocniej "ściśnięte"** bliżej średniej rozkładu, niż jest to w przypadku rozkładu normalnego.

ExerciseAngina – Dławica wywołana wysiłkiem

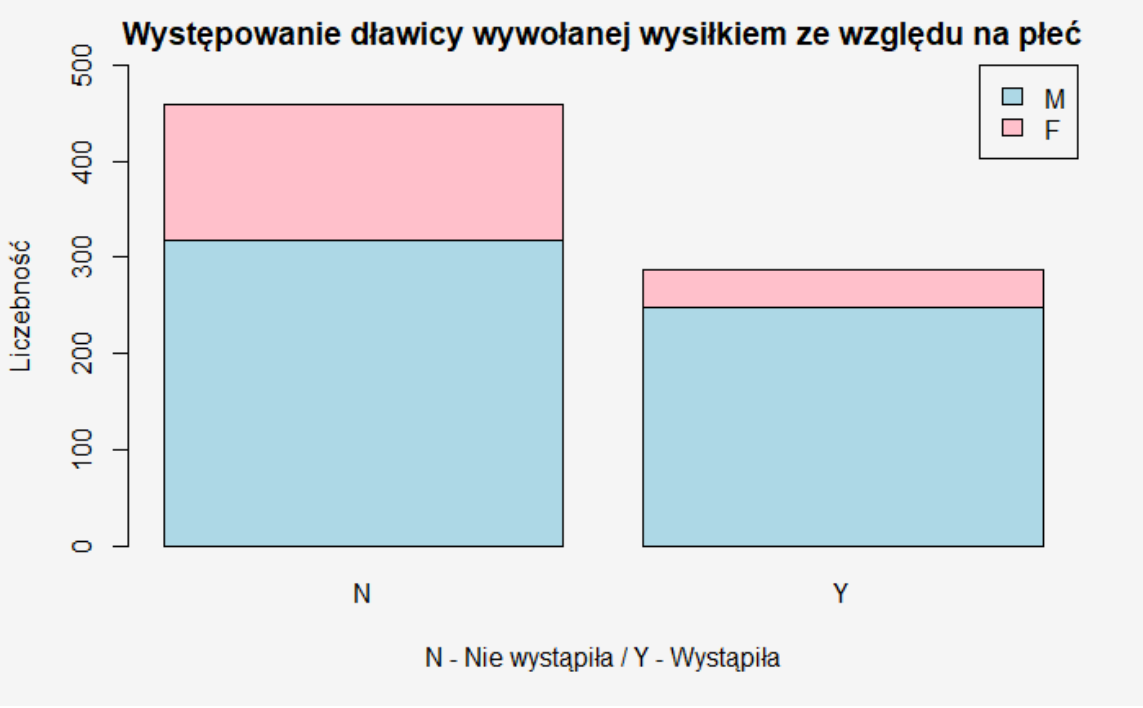
W celu zwizualizowania podziału jednostek ze względu na płeć i występowanie dławicy stworzyliśmy macierz danych, obrazującą liczebności jednostek w poszczególnej grupie (Y - dławica wystąpiła, N - dławica nie wystąpiła) w zależności od zmiennej Sex.

Tabela 7.

Mężczyzna/Kobieta	Nie wystąpiła	Wystąpiła
M	317	247
F	142	40
Razem	459	287

Wizualizujemy uzyskane dane.

Wykres 14.



Można zauważyć, że udział grupy kobiet, u której nie wystąpiła dławica wywołana wysiłkiem (ok. 31%) w stosunku do wszystkich osób, które nie chorowały na dławicę jest większy od udziału kobiet, które chorowały na dławicę (ok. 14%) w stosunku do wszystkich takich osób o ok. 17 punktów procentowych.

HeartDisease – Występowanie wieńcowej choroby serca

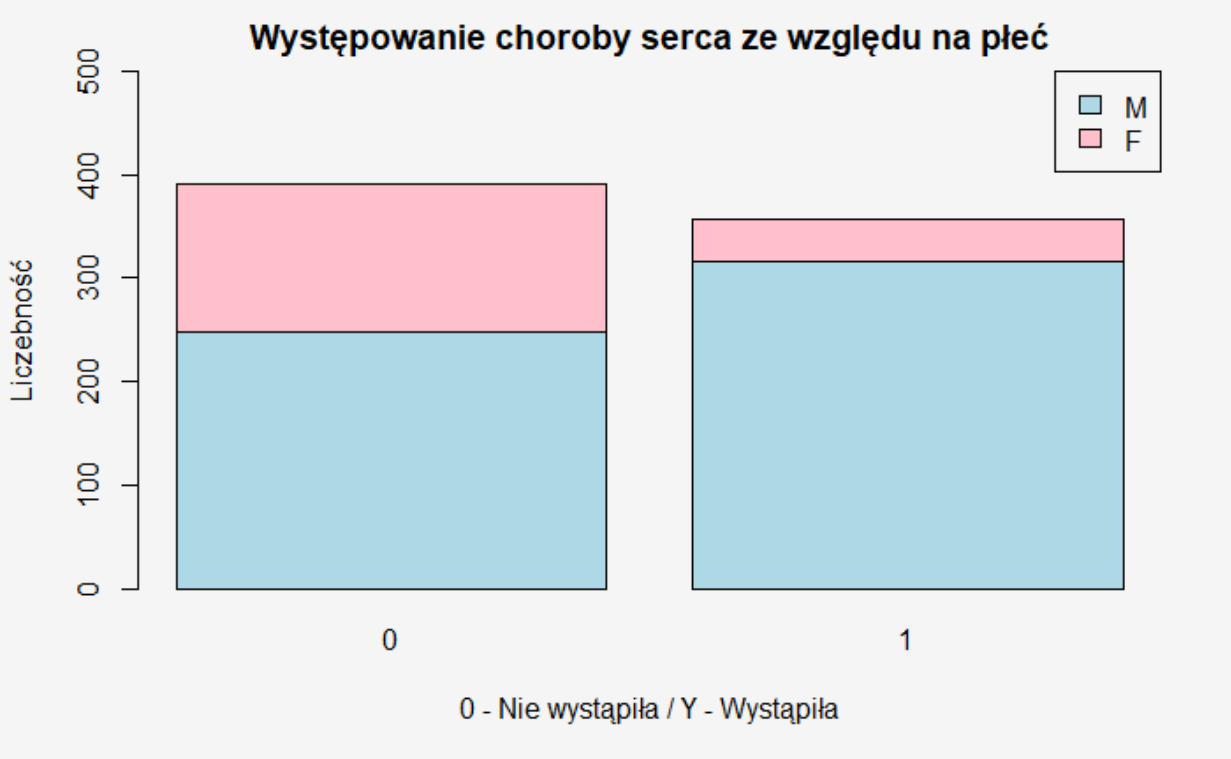
Analogicznie jak poprzednio w celu zwizualizowania podziału jednostek ze względu na płeć i występowanie choroby serca tworzymy macierz danych, obrazującą liczebności jednostek w poszczególnej grupie (0 = nie wystąpiła wieńcowa choroba serca, 1 = wystąpiła wieńcowa choroba serca) w zależności od zmiennej "Sex".

Tabela 8.

Mężczyzna/Kobieta	Nie wystąpiła	Wystąpiła
M	248	316
F	142	40
Razem	390	356

Wizualizujemy uzyskane dane.

Wykres 11.



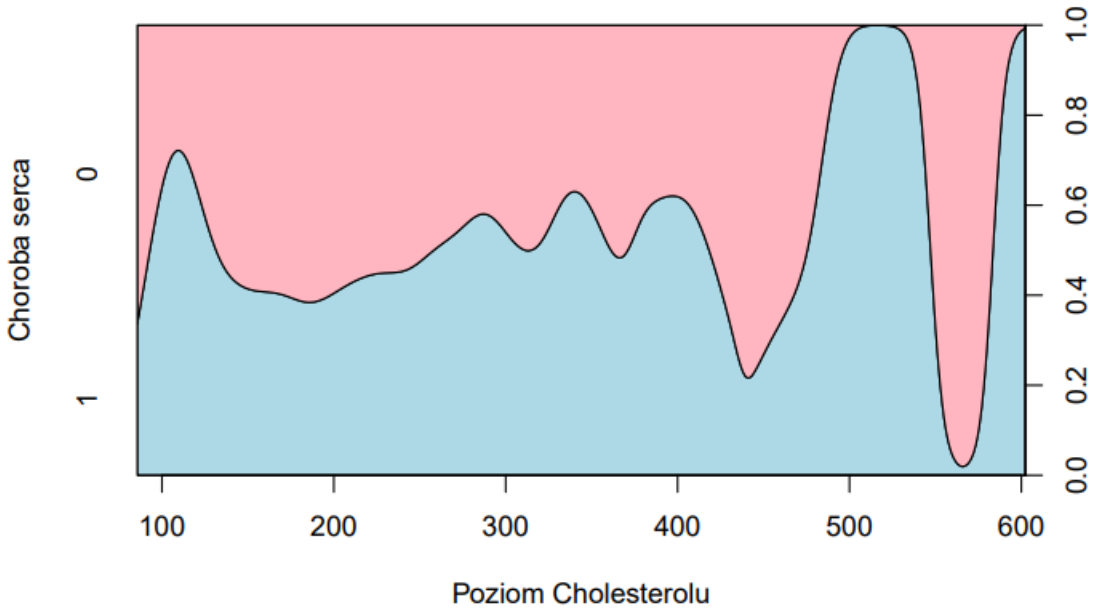
Można zauważyć, że udział grupy mężczyzn, u której wystąpiła choroba wieńcowa serca (ok. 89%) w stosunku do wszystkich osób, które chorowały jest większy od udziału mężczyzn, którzy nie doświadczyli choroby wieńcowej serca (ok. 64%) w stosunku do wszystkich takich osób o ok. 25 punktów procentowych.

Model logitowy

Modele logitowe są wykorzystywane do objaśniania zmiennej jakościowej dychotomicznej, która reprezentowana jest przez zmienną zero-jedynkową. Przeprowadzona analiza ma na celu zbudowanie modelu opisującego występowanie chorób serca (1- tak, 0-nie), stąd zastosowanie regresji logistycznej.

Analizę rozpoczynamy od wizualizacji kilku zmiennych w kontekście zmiennej „HeartDisease", czyli choroby wieńcowej serca.

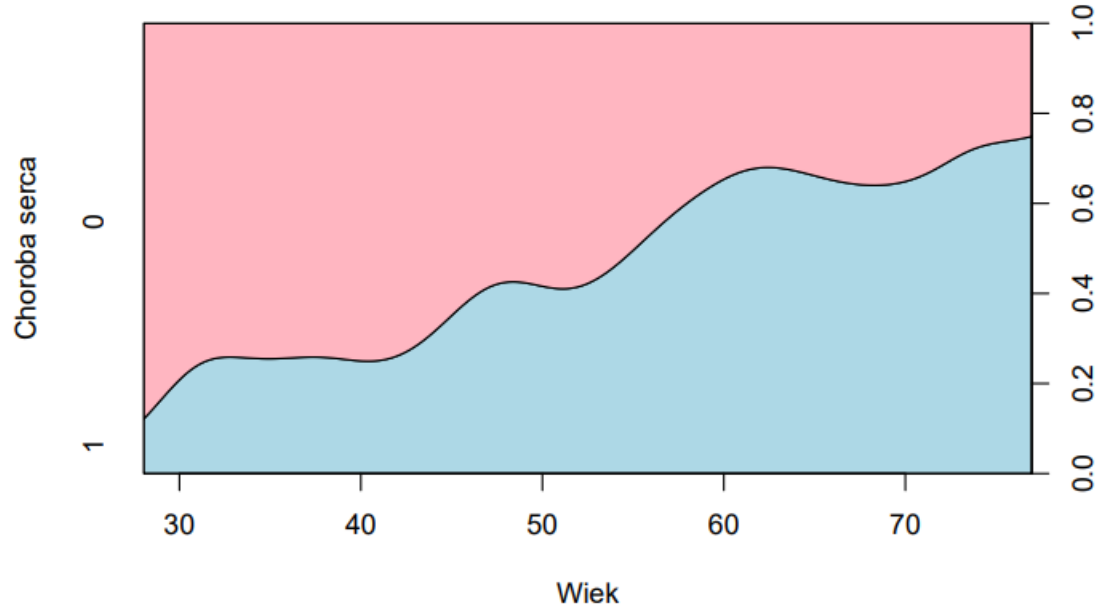
Wykres 12.



Na powyższym wykresie przedstawiona została zależność pomiędzy poziomem cholesterolu a chorobą wieńcową serca. Kolor niebieski opisuje występowanie choroby wieńcowej serca, natomiast kolor różowy- jej brak.
Jak można zauważyć, rozkład jest nierównomierny i trudno na tym etapie o wniosek dotyczący wpływu poziomu cholesterolu na chorobę wieńcową serca.

Kolejny wykres przedstawia zależność między wiekiem a chorobą wieńcową serca.

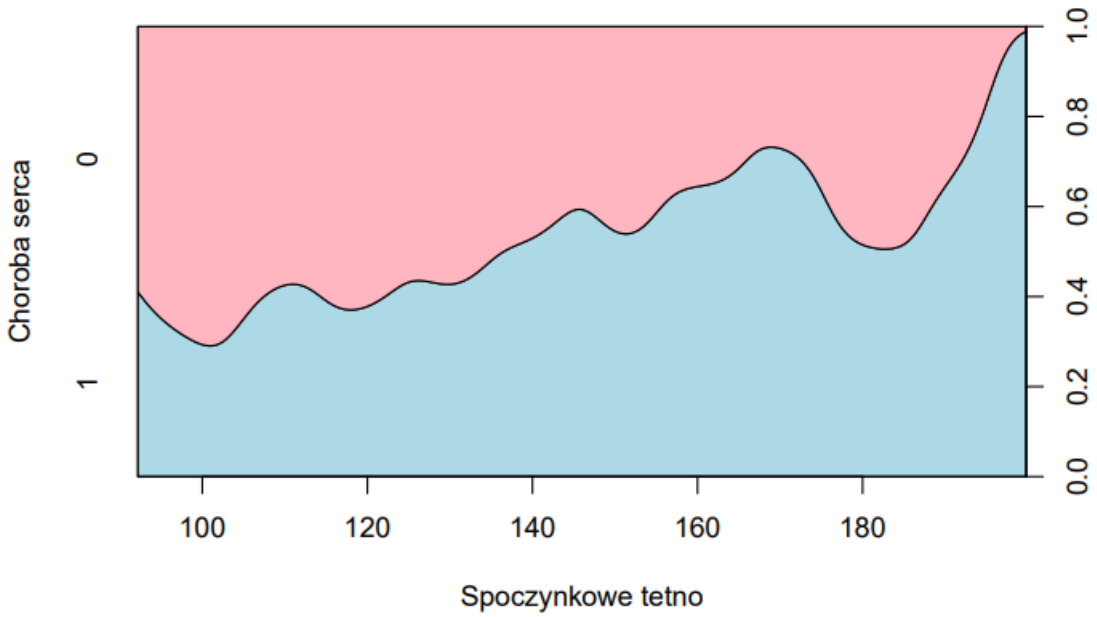
Wykres 13.



W tym przypadku zależność jest zauważalna - wraz ze wzrostem wieku, zwiększa się liczba osób z chorobą wieńcową serca.

Następna wizualizacja przedstawia zależność pomiędzy chorobą wieńcową serca a spoczynkowym tętnem.

Wykres 14.



Podobnie jak w poprzednim przypadku - wraz ze wzrostem spoczynkowego tętna wzrasta prawdopodobieństwo występowania choroby wieńcowej serca.

W dalszej analizie zbiór zostanie podzielony na dwie części - na zbiór **uczący i testowy.**
Zbiór uczący/treningowy służy do budowy modelu, a zbiór testowy służy do oceny modelu.
Dokonamy losowego podziału w proporcji: 70% i 30% odpowiednio.

Rozpoczynamy od porównania udziału występowania choroby wieńcowej serca w ogólnej liczbie jednostek, na zbiorze uczącym oraz testowym.

Tabela 9.

	Brak występowania	Występowanie
Ogólna liczba jednostek	0.5227882	0.4772118
Zbiór uczący	0.5229885	0.4770115
Zbiór testowy	0.5223214	0.4776786

Po podziale na zbiór uczący i testowy, udział osób u których występowały choroby wieńcowe serca- lub nie, nie zmienił się znacząco. Oznacza to poprawne wyznaczenie liczebności zbiorów.

Sprawdzenie korelacji parami zmiennych objaśniających

Należy sprawdzić, czy do budowy modelu można wykorzystać wszystkie wybrane zmienne ze zbioru danych. W tym celu utworzona została tabela z korelacjami między zmiennymi.

Tabela 10.

	Age	RestingBP	Cholesterol	MaxHR
Age	1.00000000	0.27959543	0.06491249	-0.38739260
RestingBP	0.27959543	1.00000000	0.08506986	-0.08943196
Cholesterol	0.06491249	0.08506986	1.00000000	-0.03753796
MaxHR	-0.38739260	-0.08943196	-0.03753796	1.00000000

Zaleca się unikania w modelu zmiennych nadmiernie skorelowanych, tzn. |r|>=0.7.
Nadmierna korelacja w tym przypadku nie występuje. Można wykorzystać do budowy modelu wszystkie zmienne dostępne w tym zbiorze danych.

Dodatkowo można zbudować modele liniowe, które pokażą, czy występują znaczące zależności między zmiennymi objaśniającymi. Pozwoli to na stwierdzenie, czy któreś zmienne wpływają na siebie, co w dalszej analizie powinno wykluczyć stosowanie ich razem.

Weźmy pod uwagę zmienną MaxHR oraz Sex w kontekście zmiennej objaśnianej Cholesterol.

```
Call:
lm(formula = Cholesterol ~ MaxHR, data = heart_uczacy)

Residuals:
    Min      1Q  Median      3Q     Max
-160.93  -37.97   -8.07   31.15  355.69

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 258.83552   15.40734  16.799   <2e-16 ***
MaxHR        -0.09219    0.10762  -0.857    0.392
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 60.11 on 520 degrees of freedom
Multiple R-squared:  0.001409,  Adjusted R-squared:  -0.0005113
F-statistic: 0.7338 on 1 and 520 DF,  p-value: 0.3921
```

```
Call:
lm(formula = Cholesterol ~ Sex, data = heart_uczacy)

Residuals:
    Min      1Q  Median      3Q     Max
-157.82  -38.82   -7.82   32.41  360.18

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  255.508      5.380   47.49   <2e-16 ***
SexM         -12.691      6.161   -2.06   0.0399 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 59.91 on 520 degrees of freedom
Multiple R-squared:  0.008094,  Adjusted R-squared:  0.006186
F-statistic: 4.243 on 1 and 520 DF,  p-value: 0.03991
```

Jak widać, osiągalne maksymalne tętno nie wpływa istotnie statystycznie na objaśnianie cholesterolu. Natomiast parametr dla zmiennej płeć na poziomie istotności równym 5% jest istotny statystycznie, co może wskazywać na zależność między tymi zmiennymi.
Jednak na tym etapie nie można założyć czy zmienna Cholesterol ma być wykluczona z kolejnych etapów analizy.

Można również sprawdzić, czy dławica wywołana wysiłkiem wpływa na spoczynkowe tętno.

```
Call:
lm(formula = RestingBP ~ ExerciseAngina, data = heart_uczacy)

Residuals:
    Min      1Q  Median      3Q     Max
-45.262 -10.625  -0.625   9.375  62.738

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     130.6250     0.9721 134.373  < 2e-16 ***
ExerciseAnginaY   6.6374     1.5627   4.247 2.56e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 17.39 on 520 degrees of freedom
Multiple R-squared:  0.03353,   Adjusted R-squared:  0.03167
F-statistic: 18.04 on 1 and 520 DF,  p-value: 2.562e-05
```

Dławica wywołana wysiłkiem jest zmienną istotną statystycznie przy objaśnianiu spoczynkowego tętna. Prawdopodobnie obie te zmienne nie będą brane razem pod uwagę przy objaśnianiu zmiennej HeartDisease w dalszym etapie analizy, jednak teraz nie można tego jednoznacznie stwierdzić.

Estymacja modeli dwumianowych logitowych jednoczynnikowych

Estymujemy model dla zmiennej dychotomicznej Y (family = binomial) z domyślną funkcją wiążącą link = logit.

Rozpoczniemy od zbudowania modeli z jedną zmienną objaśniającą, aby sprawdzić istotność parametrów.

Tabela 11.

Model	Współczynniki modelu
HeartDisease ~ Age	Estimate Std. Error z value Pr(>|z|) (Intercept) -3.49501823 0.54524051 -6.410049 1.454731e-10 Age 0.06434826 0.01013052 6.351919 2.126456e-10
HeartDisease ~ Cholesterol	Estimate Std. Error z value Pr(>|z|) (Intercept) -0.742454259 0.376412606 -1.972448 0.04855850 Cholesterol 0.002645377 0.001489285 1.776273 0.07568798
HeartDisease ~ Sex	Estimate Std. Error z value Pr(>|z|) (Intercept) -1.533930 0.2350712 -6.525386 6.782644e-11 SexM 1.817217 0.2559525 7.099819 1.249201e-12
HeartDisease ~ MaxHR	Estimate Std. Error z value Pr(>|z|) (Intercept) 4.75773256 0.613136143 7.759667 8.515239e-15 MaxHR -0.03442997 0.004300829 -8.005427 1.190533e-15
HeartDisease ~ RestingBP	Estimate Std. Error z value Pr(>|z|) (Intercept) -2.54712652 0.695594134 -3.661800 0.0002504495 RestingBP 0.01842735 0.005180037 3.557377 0.0003745760
HeartDisease ~ ExerciseAngina	Estimate Std. Error z value Pr(>|z|) (Intercept) -1.082014 0.1285681 -8.415887 3.899338e-17 ExerciseAnginaY 2.679618 0.2278023 11.762909 6.060972e-32

Na podstawie modelu *HeartDisease ~ Cholesterol* zmienna Cholesterol nie wpływa istotnie statystycznie na zmienną HeartDisease - występowanie choroby wieńcowej serca (p-value > przyjętego poziomu istotności na poziomie 0.05). Parametr okazał się nieistotny statystycznie.

Dla pozostałych modeli wykorzystane zmienne okazały się statystycznie istotne, co oznacza, że na występowanie chorób wieńcowych serca wpływają: wiek, płeć, osiągalne maksymalne tętno, spoczynkowe ciśnienie krwi oraz występowanie dławicy wywołanej wysiłkiem.

Porównanie dobroci dopasowania modeli logitowych

W kolejnym kroku porównane zostaną modele logitowe.
Pod uwagę weźmiemy miary kryterium Akaike, McFaddena oraz Cragga Uhlera.

Tabela 12.

Model	Kryterium AIC	McFadden	Cragg Uhler
HeartDisease ~ Age	681.4121	0.062459644	0.110509339
HeartDisease ~ Cholesterol	723.3180	0.004461809	0.008215015
HeartDisease ~ Sex	663.7691	0.086877685	0.151180151
HeartDisease ~ MaxHR	649.9041	0.106066836	0.182190939
HeartDisease ~ RestingBP	713.3221	0.018296176	0.033366482
HeartDisease ~ ExerciseAngina	549.1708	0.245481965	0.384379248

Na podstawie kryterium informacyjnego i miar dopasowania pseudo-R^2 można znaleźć potwierdzenie o najgorszym dopasowaniu modelu *HeartDisease ~ Cholesterol*, w którym zmienna Cholesterol okazała się nieistotna statystycznie.

Najlepszym modelem jest model ze zmienną objaśniającą ExerciseAngina (dławica wywołana wysiłkiem), ponieważ kryterium AIC ma wartość najniższą, a miary pseudo-R^2 osiągnęły wartość największą dla tego modelu.

Interpretacja parametrów modelu HeartDisease ~ ExerciseAngina

```
Call:
glm(formula = HeartDisease ~ ExerciseAngina, family = binomial,
    data = heart_uczacy)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.8878  -0.7640  -0.7640   0.6071   1.6576

Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)     -1.0820     0.1286  -8.416   <2e-16 ***
ExerciseAnginaY  2.6796     0.2278  11.763   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 722.54  on 521  degrees of freedom
Residual deviance: 545.17  on 520  degrees of freedom
AIC: 549.17

Number of Fisher Scoring iterations: 4
```

Szansa wystąpienia choroby wieńcowej serca w grupie referencyjnej, czyli dla osób bez dławicy wywołanej wysiłkiem (ExerciseAngina=0) wynosi 0,34.

Szansa wystąpienia choroby wieńcowej serca u osoby z dławicą wywołaną wysiłkiem jest ok. 14,58 razy większa niż dla osoby bez dławicy.

W kolejnym kroku zbudujemy model wykorzystując cały zestaw zmiennych, które okazały się istotne statystycznie: wiek, płeć, osiągalne maksymalne tętno, spoczynkowe ciśnienie krwi oraz występowanie dławicy wywołanej wysiłkiem.

```
Call:
glm(formula = HeartDisease ~ Age + MaxHR + RestingBP + Sex +
    ExerciseAngina, family = binomial, data = heart_uczacy)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.5304  -0.7354  -0.2767   0.5544   2.4440

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)    -3.064176   1.395029  -2.196  0.02806 *
Age             0.043919   0.013552   3.241  0.00119 **
MaxHR          -0.013519   0.005226  -2.587  0.00968 **
RestingBP       0.002693   0.006633   0.406  0.68473
SexM            1.711542   0.303184   5.645 1.65e-08 ***
ExerciseAnginaY 2.301751   0.253420   9.083  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 722.54  on 521  degrees of freedom
Residual deviance: 478.44  on 516  degrees of freedom
AIC: 490.44

Number of Fisher Scoring iterations: 5
```

W modelu, w którym uwzględnione zostały wiek, maksymalne osiągalne tętno, spoczynkowe ciśnienie krwi, płeć oraz dławica wywołana wysiłkiem zmienna RestingBP jest nieistotna statystycznie. Pozostałe zmienne są istotne statystycznie.

Zbudujemy model nieuwzględniający zmiennej RestingBP.

```
Call:
glm(formula = HeartDisease ~ Age + MaxHR + Sex + ExerciseAngina,
    family = binomial, data = heart_uczacy)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.5172  -0.7377  -0.2823   0.5582   2.4488

Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)     -2.798867   1.231696  -2.272 0.023064 *
Age              0.045315   0.013127   3.452 0.000556 ***
MaxHR           -0.013401   0.005219  -2.568 0.010233 *
SexM             1.711356   0.302265   5.662  1.5e-08 ***
ExerciseAnginaY  2.314910   0.251602   9.201  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 722.54  on 521  degrees of freedom
Residual deviance: 478.61  on 517  degrees of freedom
AIC: 488.61

Number of Fisher Scoring iterations: 5
```

Po usunięciu tej zmiennej na poziomie istotności 5% już wszystkie zmienne są istotne statystycznie.

Tabela 13.

Test ilorazu wiarygodności	Test Walda
Likelihood ratio test Model 1: HeartDisease ~ Age + MaxHR + Sex + ExerciseAngina Model 2: HeartDisease ~ 1 #Df LogLik Df Chisq Pr(>Chisq) 1 5 -239.30 2 1 -361.27 -4 243.93 < 2.2e-16 *** --- Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1	Wald test Model 1: HeartDisease ~ Age + MaxHR + Sex + ExerciseAngina Model 2: HeartDisease ~ 1 Res.Df Df F Pr(>F) 1 517 2 521 -4 36.334 < 2.2e-16 *** --- Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

W modelu, w którym zmiennymi objaśniającymi są wiek, płeć, dławica wywołana wysiłkiem oraz maksymalne osiągalne tętno (Age, Sex, ExerciseAngina i MaxHR) wszystkie zmienne są istotne statystycznie.

Dodatkowo oba testy wykazały, że w modelu istnieje parametr przy zmiennej objaśniającej, który różni się statystycznie istotnie od 0.

Można postawić hipotezę, że jest to najlepszy model objaśniający występowanie choroby serca.

Aby jednak to sprawdzić dokonamy oceny modelu porównując go z modelem, który okazał się być najlepszy dotychczas - z jedną zmienną objaśniającą - ExerciseAngina.

Tabela 14.

Model	Kryterium AIC	McFadden	Cragg Uhler
HeartDisease ~ ExerciseAngina	549.1708	0.2454820	0.3843792
HeartDisease ~ Age + Sex + ExerciseAngina + MaxHR	488.6070	0.3376065	0.4981021

Model *HeartDisease ~ Age + Sex + ExerciseAngina + MaxHR*, ma mniejszą wartość kryterium AIC oraz wyższe wartości miar pseudo-R^2. Oznacza to, że jest lepszym niż ten, w którym zmienną egzogeniczną jest ExerciseAngina.

Postać oraz interpretacja modelu logitowego

*Logit(p)= -2.798867 + 0.045315 *Age - 0.013401*MaxHR + 1.711356 *SexM + 2.314910*ExerciseAnginaY*

- ***Szansa w grupie referencyjnej***

  Interpretacja nie ma tu sensu, ponieważ najmłodsza badana osoba jest w wieku 28 lat.

- ***Iloraz szans $e^{0,045315}$=1, 046357***

Przy wzroście wieku o rok szansa wystąpienia choroby wieńcowej serca wzrasta średnio o 4,63% przy pozostałych zmiennych ustalonych (ceteris paribus).

- ***Iloraz szans $e^{(-0.013401)}$ -> (1 - $e^{(-0.013401)}$)*100 = 1,33116***

Przy wzroście maksymalnego tętna o 1 jednostkę szansa wystąpienia choroby wieńcowej serca maleje średnio o 1,33116% przy pozostałych zmiennych ustalonych (ceteris paribus).

- ***Iloraz szans $e^{1,711356}$=5.536464***

Szansa wystąpienia choroby wieńcowej serca dla mężczyzn jest prawie 6 razy większa niż dla kobiet przy pozostałych zmiennych ustalonych (ceteris paribus).

- ***Iloraz szans $e^{2.314910}$=10.12401***

Szansa wystąpienia choroby wieńcowej serca dla osób z dławicą jest 10 razy większa niż dla osób bez dławicy przy pozostałych zmiennych ustalonych (ceteris paribus).

Model probitowy

Estymujemy model dla zmiennej dychotomicznej Y family = binomial z funkcją wiążącą probit (link = probit). Na podstawie modelu logitowego, który okazał się najlepszym modelem, zbudujemy model probitowy, aby porównać go z modelem logitowym.

Wyniki oszacowania modelu probitowego:

```
Call:
glm(formula = HeartDisease ~ Age + Sex + ExerciseAngina + MaxHR,
    family = binomial(link = probit), data = heart_uczacy)

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)     -1.678246   0.703713  -2.385 0.017086 *
Age              0.026521   0.007468   3.551 0.000384 ***
SexM             0.976780   0.167123   5.845 5.08e-09 ***
ExerciseAnginaY  1.370497   0.142505   9.617  < 2e-16 ***
MaxHR           -0.007498   0.003012  -2.490 0.012790 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 722.54  on 521  degrees of freedom
Residual deviance: 479.00  on 517  degrees of freedom
AIC: 489

Number of Fisher Scoring iterations: 5
```

Testy istotności parametrów modelu

W celu sprawdzenia istotności parametrów modelu przeprowadzono testy: Walda oraz ilorazu wiarygodności.

```
> lrtest(probit1)
Likelihood ratio test

Model 1: HeartDisease ~ Age + Sex + ExerciseAngina + MaxHR
Model 2: HeartDisease ~ 1
  #Df  LogLik Df  Chisq Pr(>Chisq)
1   5 -239.50
2   1 -361.27 -4 243.55  < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> waldtest(probit1)
Wald test

Model 1: HeartDisease ~ Age + Sex + ExerciseAngina + MaxHR
Model 2: HeartDisease ~ 1
  Res.Df Df      F    Pr(>F)
1    517
2    521 -4 45.093 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Parametry okazały się istotne statystycznie. Oba testy wykazały, że w modelu istnieje parametr przy zmiennej objaśniającej, który różni się statystycznie istotnie od 0. Odrzucamy H0 na korzyść H1.

Wartości p-value przy każdym oszacowanym parametrze są mniejsze od poziomu istotności 0,05. Oznacza to, że wszystkie wykorzystane zmienne wpływają istotnie na występowanie choroby wieńcowej serca.

Interpretacja parametrów modelu probitowego

- Jeśli wzrasta wiek to wzrasta prawdopodobieństwo wystąpienia choroby wieńcowej serca – zmienna jest stymulantą modelu.
- Dla mężczyzny prawdopodobieństwo wystąpienia choroby wieńcowej serca też jest większe – zmienna również jest stymulantą.
- Występowanie dławicy serca również jest stymulantą modelu w kontrze do MaxHR (destymulanta).

Porównanie modelu logitowego i probitowego

Tabela 15.

Model	Kryterium AIC	MCFadden	Cragg Uhler
logitowy	488.6070	0.3376065	0.4981021
probitowy	488.9959	0.3370682	0.4974789

Najlepszym modelem jest model logitowy, ponieważ kryterium AIC ma wartość najniższą, a miary pseudo-R2 osiągnęły wartość największą dla tego modelu. Jednak są to różnice niewielkie (na poziomie kilku miejsc po przecinku).

Przykładowa predykcja prawdopodobieństwa dla różnych przypadków dla modelu logitowego i probitowego

Tabela 16.

Przypadek	Prawdopodobieństwo	
	Model logitowy	Model probitowy
Kobieta, 65 lat, z dławicą, tętno 130	0,672	0,671
Mężczyzna, 65 lat, z dławicą, tętno 130	**0,919**	**0,922**
Kobieta, 35 lat, bez dławicy, tętno 130	0,049	0,042
Kobieta, 35 lat, z dławicą, tętno 130	0,345	0,361
Kobieta, 65 lat, bez dławicy, tętno 100	0,233	0,241
Mężczyzna, 65 lat, bez dławicy, tętno 100	0,627	0,607
Mężczyzna, 35 lat, z dławicą	0,814	0,801
Mężczyzna, 35 lat, bez dławicy	0,301	0,300

Największe prawdopodobieństwo wystąpienia choroby wieńcowej serca jest dla mężczyzny, 65 lat, z dławicą serca., tętno 130. Najmniejsze dla kobiety, 35 lat, bez dławicy serca, tętno 130.
Model logitowy i probitowy daje zbliżone wartości prawdopodobieństwa.

Porównanie jakości predykcji modeli logitowego i probitowego

Tworzymy tablice trafności dla punktu odcięcia p-proporcja z próby uczącej*

Tabela 17.

Tablica trafności dla modelu logitowego (próba ucząca)			
	Przewidywane		
Obserwowane		0	1
	0	231	42
	1	62	187

Tablica trafności dla modelu probitowego (próba ucząca)			
	Przewidywane		
Obserwowane		0	1
	0	231	42
	1	63	186

Tablica trafności dla modelu logitowego (próba testowa)			
	Przewidywane		
Obserwowane		0	1
	0	94	23
	1	30	77

Tablica trafności dla modelu probitowego (próba testowa)			
	Przewidywane		
Obserwowane		0	1
	0	95	22
	1	30	77

W utworzonych tablicach trafności nie ma znaczących różnic między wartościami w modelach logitowym i probitowym zarówno na zbiorze uczącym i testowym. Może to oznaczać, że oba modele podobnie przewidują choroby serca.

Miary trafności prognoz

Tabela 18.

Model	*ACC*	*ER*	*SENS*	*SPEC*	*PPV*	*NPV*
model_logit_uczacy	*0.8007663*	*0.1992337*	*0.7510040*	*0.8461538*	*0.8165939*	*0.7883959*
model_probit_uczacy	*0.7988506*	*0.2011494*	*0.7469880*	*0.8461538*	*0.8157895*	*0.7857143*
model_logit_testowy	*0.7633929*	*0.2366071*	*0.7196262*	*0.8034188*	*0.7700000*	*0.7580645*
model_probit_testowy	*0.7678571*	*0.2321429*	*0.7196262*	*0.8119658*	*0.7777778*	*0.7600000*

Interpretacja miar dla próby uczącej:

- ACC(logit)=0,8007663
- ACC(probit)=0,7988506

Udział liczby trafnie sklasyfikowanych jednostek w ogólnej liczbie jednostek (dla modelu logitowego) wynosi 80% - to umiarkowany wynik.

- ER(logit)=0,1992337
- ER(probit)=0,2011494

Udział liczby źle sklasyfikowanych jednostek w ogólnej liczbie jednostek (dla modelu logitowego) wynosi 19.92%.

- SENS(logit)=0,7510040
- SENS(probit)=0,7469880

Dla obu modeli 75% jedynek (osób z chorobą serca) zostało trafnie oszacowanych w liczbie wszystkich obserwowanych jedynek (osób z chorobą serca).

- SPEC(logit)=0.*8461538*
- SPEC(probit)= *0.8461538*

Dla modelu logitowego 84,61% zer (osób bez choroby serca) zostało trafnie oszacowanych w liczbie wszystkich obserwowanych zer (osób bez choroby serca). Dla modelu probitowego było to 84,61%.

- PPV(logit)=*0.8165939*
- PPV(probit)= *0.8157895*

Wśród sklasyfikowanych jedynek 81% było w rzeczywistości jedynkami. Dla modelu logitowego udział liczby trafnie oszacowanych jedynek (osób z chorobą serca) w liczbie wszystkich prognozowanych jedynek (osób z chorobą wieńcową serca) wyniósł 81%. Dla modelu probitowego było to 81,57%.

- NPV(logit)= *0.7883959*
- NPV(probit)= *0.7857143*

Wśród sklasyfikowanych zer 78,89% było w rzeczywistości zerami. Dla modelu logitowego udział liczby trafnie oszacowanych zer (osób bez choroby wieńcowej serca) w liczbie wszystkich prognozowanych zer (osób bez choroby serca) wyniósł 78,89%. Dla modelu probitowego było to 78,57%.

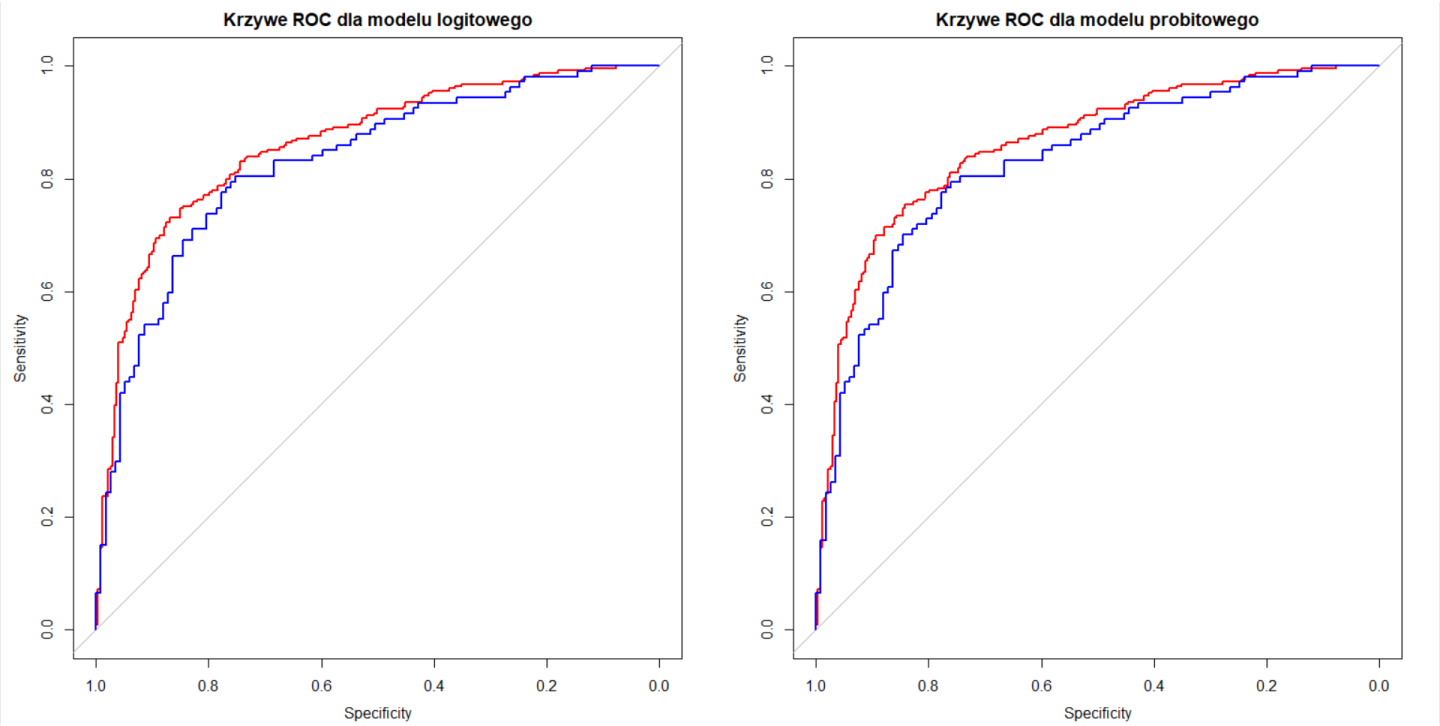
Dodatkowo, na zbiorze testowym miary nie pogorszyły się, co świadczy na korzyść modelu.

Krzywa ROC

Krzywa ROC prezentuje jakość predykcji modelu dla wszystkich możliwych punktów odcięcia p* (jest niezależna od wyboru p*). Dla modeli oszacowanych na zbiorze uczącym porównana została poniżej jakość predykcji na zbiorze uczącym i testowym:

- krzywa czerwona - ROC wyznaczona na zbiorze uczącym.
- krzywa niebieska - ROC wyznaczona na zbiorze testowym.

Wykres 19.



AUC - pole powierzchni pod krzywą ROC

Tabela 19.

	Pole powieszchni pod krzywą ROC	
	Model logitowy	Model probitowy
Zbiór uczący	0.8623	0.8624
Zbiór testowy	0.83	0.8301

Pole powierzchni pod krzywą ROC jest ważną statystyką odzwierciedlającą prawdopodobieństwo, że kolejność predykcji na podstawie obserwacji zmiennej testowanej będzie prawidłowa. W przypadku zbioru uczącego, wartość obliczona na podstawie modelu logitowego oraz probitowego jest taka sama i wynosi 0,862. Jest to wynik mówiący o bardzo dobrej klasyfikacji.

Na zbiorze testowym obie wartości wynoszą 0,83, co jest wynikiem mniejszym niż na poprzednim zbiorze, ale w dalszym ciągu jest to bardzo dobra klasyfikacja.

Porównanie miar jakości predykcji dla dwóch punktów odcięcia

Punkt odcięcia jako proporcja z próby uczącej p*= 0.4770115

Punkt odcięcia według indeksu Youdena dla próby uczącej p*= 0.5052504

Tabela 20.

Punkt odcięcia	***accuracy***	***sensitivity***	***specificity***	***ppv***	***npv***	***youden***
Proporcja z próby uczącej	*0.8007663*	*0.7510040*	*0.8461538*	*0.8165939*	*0.7883959*	*1.5971579*
Indeks Youdena dla próby uczącej	*0.8026820*	*0.7309237*	*0.8681319*	*0.8348624*	*0.7796053*	*1.5990556*

Dla punktu odcięcia według indeksu Youdena dla próby uczącej p*= 0.5052504 miara SENS wskazuje, że 73,09% jedynek (osób z chorobą serca) zostało trafnie oszacowanych w liczbie wszystkich obserwowanych jedynek (osób z chorobą serca). To mniej niż w przypadku punktu odcięcia jako proporcja z próby uczącej p*= 0.4770115, gdzie odsetek ten wynosił 75,1%. W tym przypadku indeks Youdena okazałby się gorszy, jeśli chodzi o wykrywanie choroby wieńcowej serca.

W przypadku tego problemu bardziej zależy nam na trafnym wykrywaniu chorób niż trafnym wykrywaniu zer, czyli braków tej choroby.

Najlepszym okazał się być model logitowy, w którym choroba wieńcowa serca objaśniana jest przez płeć, wiek, dławicę wywołaną wysiłkiem i maksymalne osiągalne tętno.