Kamil Marciniak

Raport 1

18 marca 2021

Spis treści

1. Wstepne infromacje

Raport składa się z dwóch części w pierwszej z nich będę analizował dane grades o uczniach pewnej szkoły, a w drugiej będę analizował dane income o zarobkach Amerykanów. Celem raportu jest analiza tego jak płeć wpływa na wyniki uczniów w szkole oraz w testach psychologicznych oraz jak wpływa na zarobki wśród osób dorosłych. W pierwszej części raportu będę analizował dane grades zawierające informacje o 78 uczniach pewnej szkoły z USA. Spodziewałem się danych o 89 uczniach jednak z nieznanych powodów jest o 11 mniej obserwacji.Dla każdego ucznia mamy dane 4 wartości. Są to odpowiednio:

— średnia ocen której odpowiadają wartości od 0 do 11 przy czym 0 oznacza ocenę F(najgorszą) a 11 oznacza ocenę A(najlepszą)

— wynik standardowego testu IQ, korzystając z `https://www.britannica.com/science/intelligence-test` widzimy, że testami najbardziej rozpowszechnionymi są skala inteligencji Stanforda-Bineta oraz skala Wechslera, Jednak dla moich rozważań kluczowe jest, że niezależnie od dobranego testu wyniki IQ dla większości ludzi leżą w przedziale 70-130 przy czym im wyższy wynik tym lepsze dana osoba ma zdolności

— płeć której odpowiadają dwie wartości-kobieta/mężczyzna

— punktacja na teście psychologicznym Piersa-Harrisa która waha się wśród badanych uczniów od 20 do 80. Wartości powyżej 60 świadczą o wysokiej samoocenie a wartości poniżej 40 świadczą o niskiej samoocenie, a wartości od 40 do 59 są uznawane przez psychologów za typowe. Wiedzę jak interpretować wyniki tego testu zaczerpnąłem z `https://www.montgomeryschoolsmd.org/uploadedFiles/community-engagement/linkages-to-learning/Piers-Harris-2-Interpretation-Overview.pdf`

2. Tabele statystyk dla danych grades

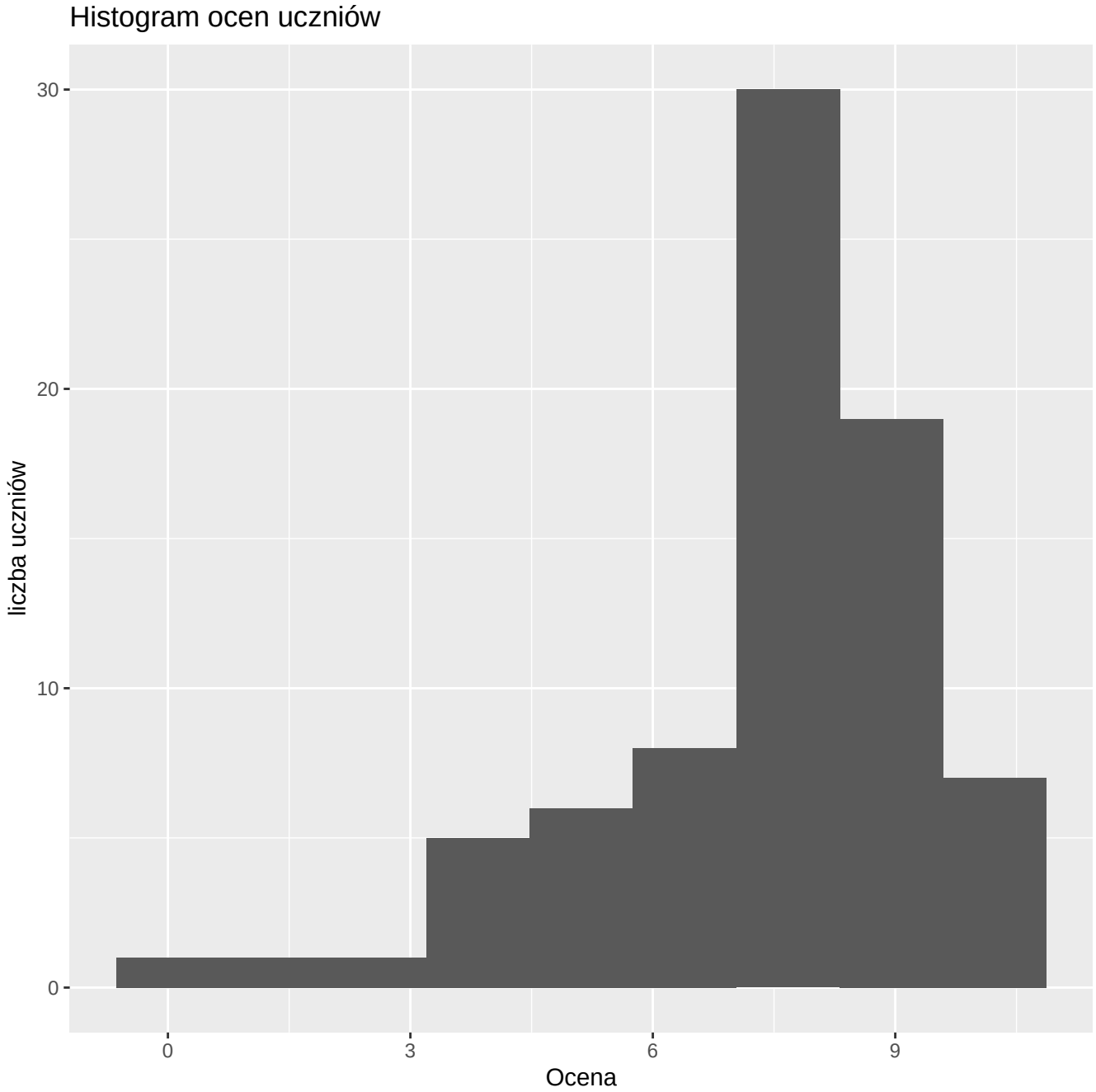
Na początek warto pokazać statystyki dla naszych danych. W pierwszej tabelce znajdują się wyniki zbiorcze. Odnotujmy że wartości ocen mieszczą się w przedziale od 0.53 do 10.76. Wartości IQ należą do przedziału od 72 do 136, a wyniki testu Piersa-Harrisa przyjmują wartości od 20 do 80. Warto również zauważyć, że najwyższy współczynnik zmienności ma zmienna Ocena, najniższy zmienna IQ.

	min	max	mediana	1Q	3Q	średnia	odchylenie	wariancja	wz
ocena	0.53	10.76	7.83	6.28	8.98	7.45	2.10	4.41	0.28
iq	72.00	136.00	110.00	103.00	117.50	108.92	13.17	173.47	0.12
ph	20.00	80.00	59.50	51.00	66.00	56.96	12.41	154.06	0.22

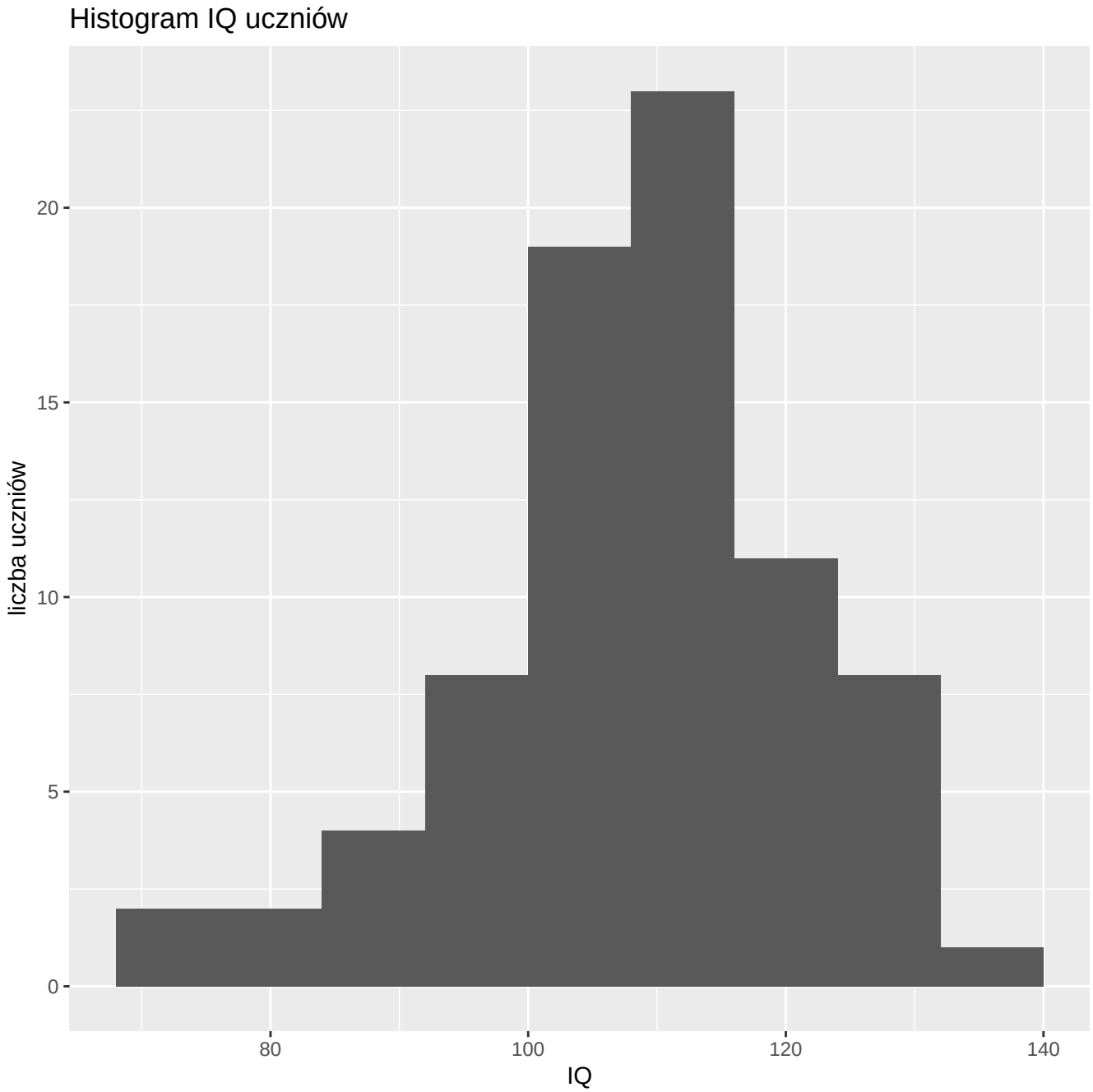
W poniższej tabelce znajdują się wyniki uczniów z podziałem na płeć. Zauważmy że wyniki testu Piersa-Harrisa maja większy współczynnik zmienności niż wyniki IQ. Dokładniejszą analizę danych przedstawię razem z analizą histogramów.

	min	max	mediana	1Q	3Q	średnia	odchylenie	wariancja	wz
ocena_ch	0.53	10.76	7.88	6.36	9.08	7.28	2.34	5.50	0.32
ocena_dz	3.41	10.70	7.83	6.37	8.95	7.68	1.69	2.87	0.22
iq_ch	77.00	136.00	111.00	106.00	119.00	111.35	11.95	142.85	0.11
iq_dz	72.00	132.00	106.00	97.50	114.00	105.44	14.22	202.25	0.13
ph_ch	20.00	80.00	59.00	51.00	67.00	57.85	12.39	153.55	0.21
ph_dz	21.00	72.00	60.00	52.75	64.00	55.69	12.53	156.93	0.22

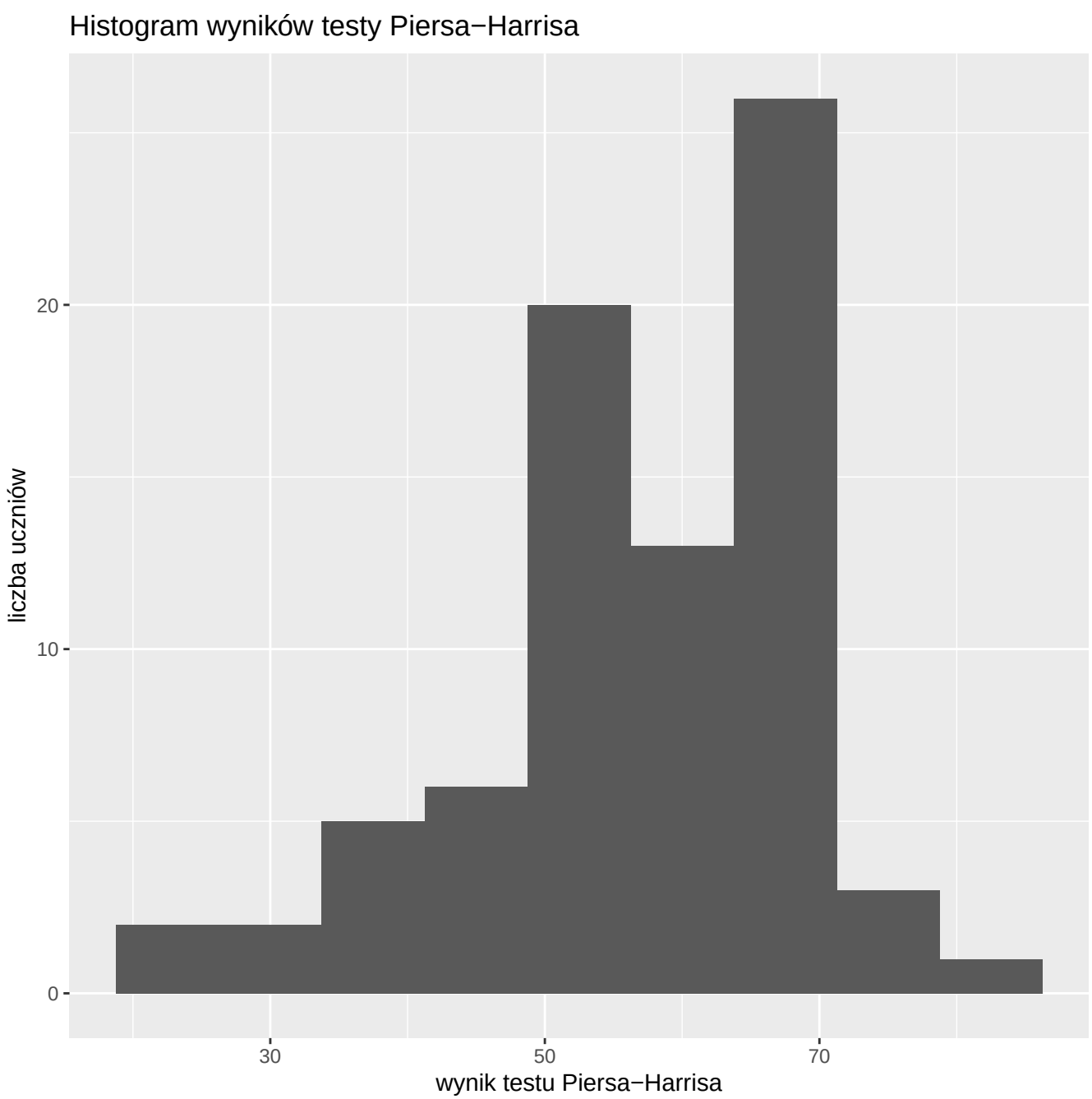
3. Histogramy dla zbioru grades



Histogram dotyczący zmiennej Ocena jest asymetryczny i skośny w lewo. Wykres jest jednomodalny. Mediana wynosi 7.83 a średnia 7.45. Rozstęp wynosi 10.23. Odchylenie standardowe jest dość niskie bo wynosi 2.1 więc spory odsetek danych jest blisko średniej. Warto również odnotować, że rozstęp międzykwartylowy jest dość mały bo wynosi niecałe 3 punkty. Przy czym pierwszy kwartyl wynosi około 6 a trzeci kwartyl około 9.

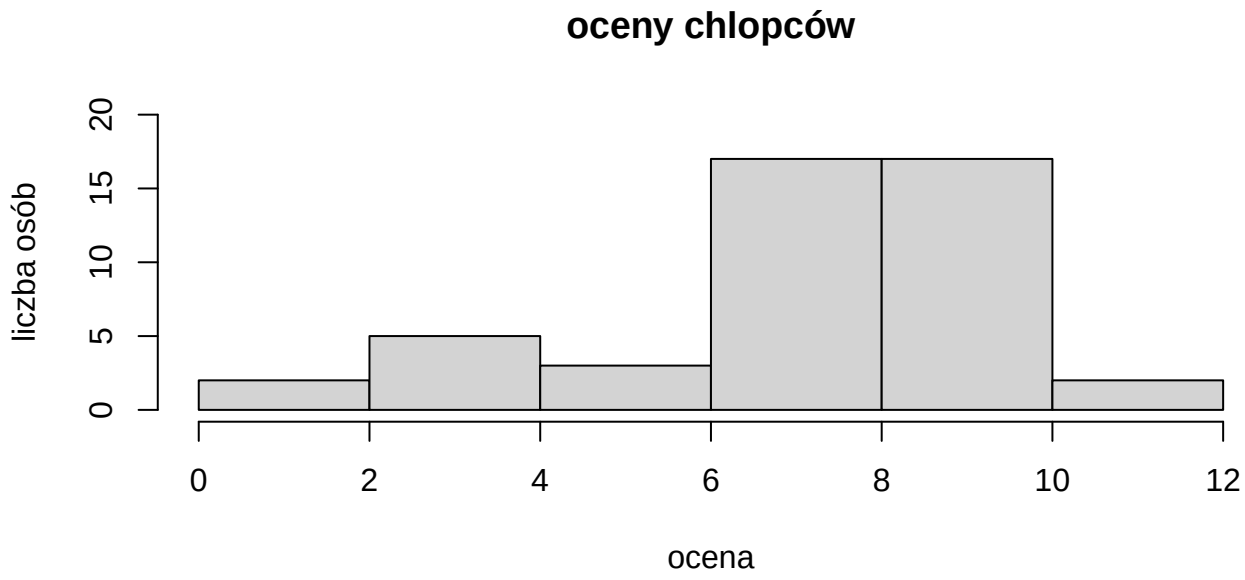


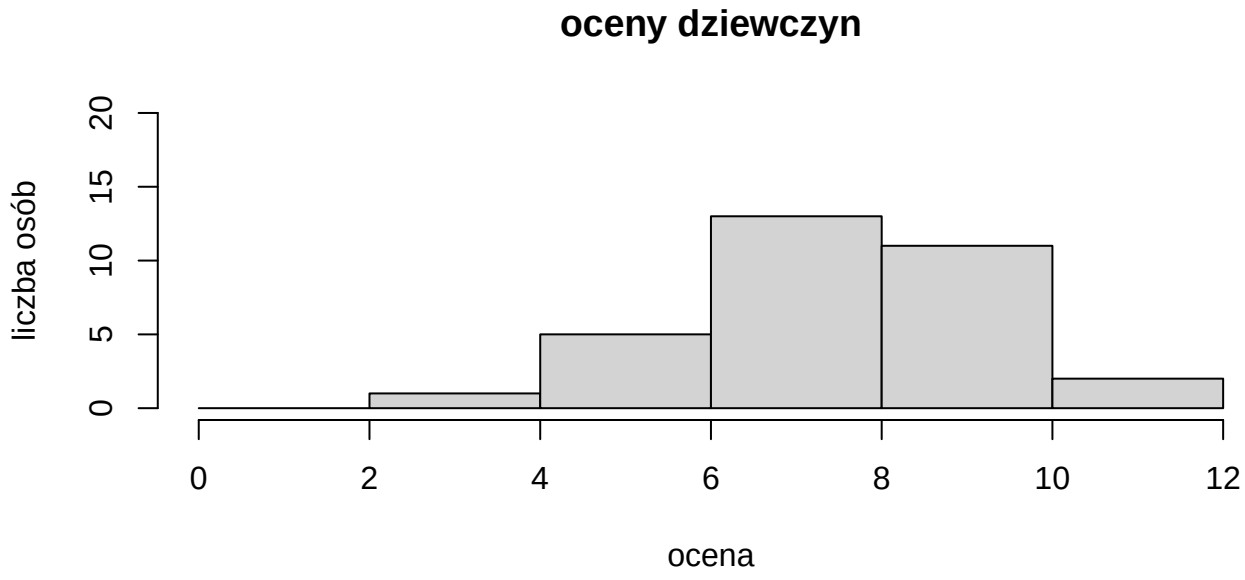
Aby upewnić się co do skośności policzyłem odpowiedni współczynnik który dla danych IQ wynosi -0.58. Histogram dotyczący IQ jest asymetryczny, skośny w lewo i jednomodalny. Rozstęp wynosi 64 a odchylenie standardowe 13.17 więc spora częśc danych znajduje się blisko średniej. Średnia (108.92) jest niewiele niższa od mediany (110). Niski rozstęp międzykwartylowy (14.5) również świadczy o silnej koncentracji danych wokół środka rozkładu.



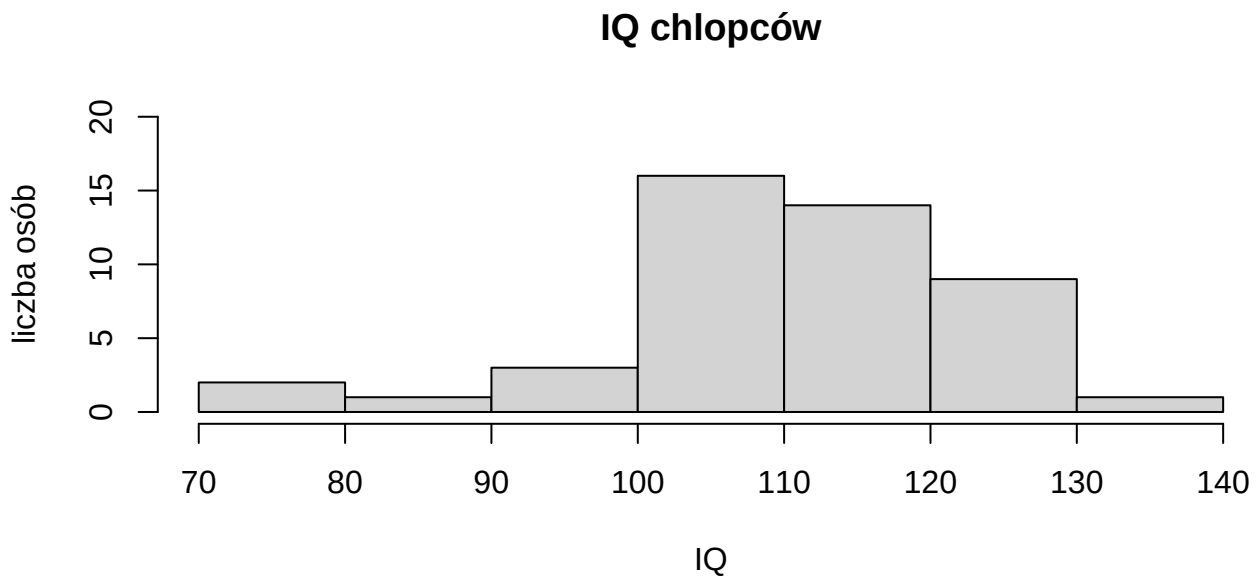
Histogram dotyczący wyników testu Piersa-Harrisa jest asymetryczny, skośny w lewo i dwumodalny.Rozstęp wynosi 60 a odchylenie standardowe 12.41. Co oznacza, że spory odsetek danych znajduje się blisko średniej.Mediana (59.5) jest wyższa od średniej (56.96). A pierwszyb kwartyl(51) różni się od trzeciego kwartyla o 15 punktów. Nie powinniśmy również przeoczyć faktu, że badani uczniowie uzyskali w tym teście wyniki istotnie lepsze od oczekiwań dla losowej próby.

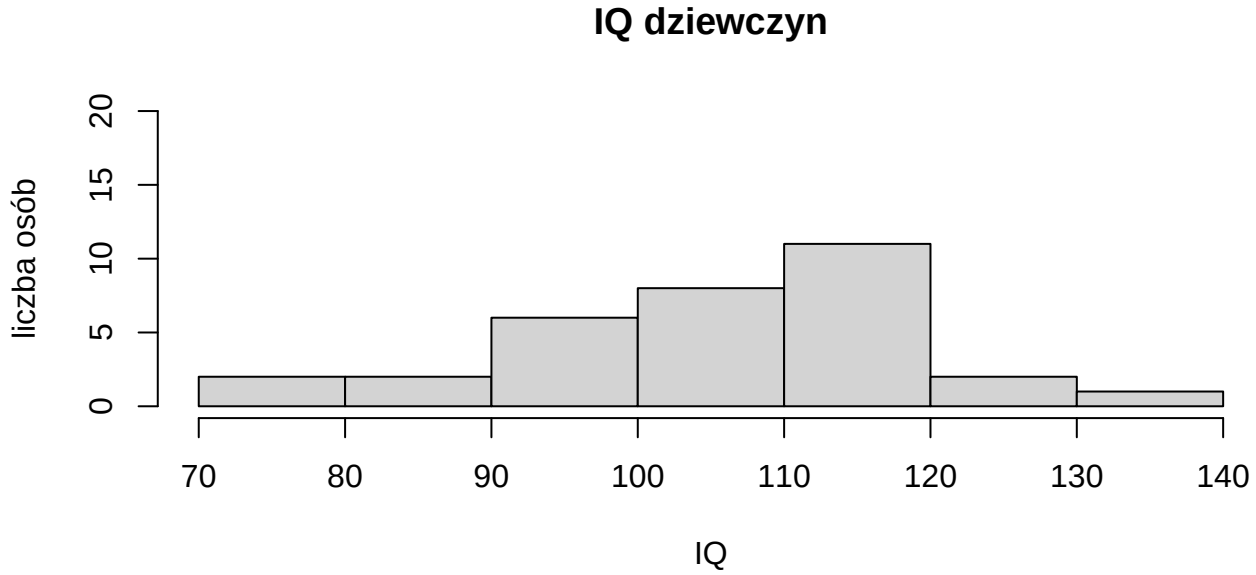
Teraz czas przeanalizować wpływ płci na oceny uczniów oraz na wyniki testu IQ oraz testu Piersa-Harrisa. Najpierw odnotujmy, że mamy informacje o 46 chłopcach oraz o 32 dziewczynach. Do analizy posłużą nam histogramy odpowiednich zmiennych z podziałem na płeć. Przy czym dla czytelności, mimo pewnej różnicy w liczebności grup chłopców i dziewczyn zdecydowałem aby skale jak i szerokość słupków na histogramach były takie same dla obu płci.



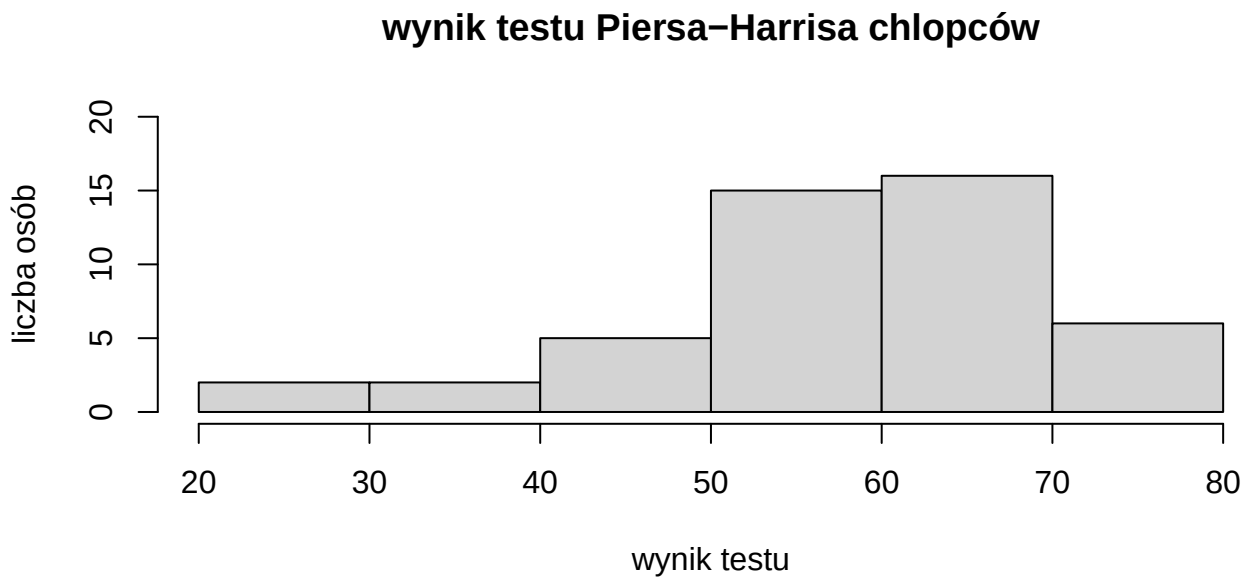


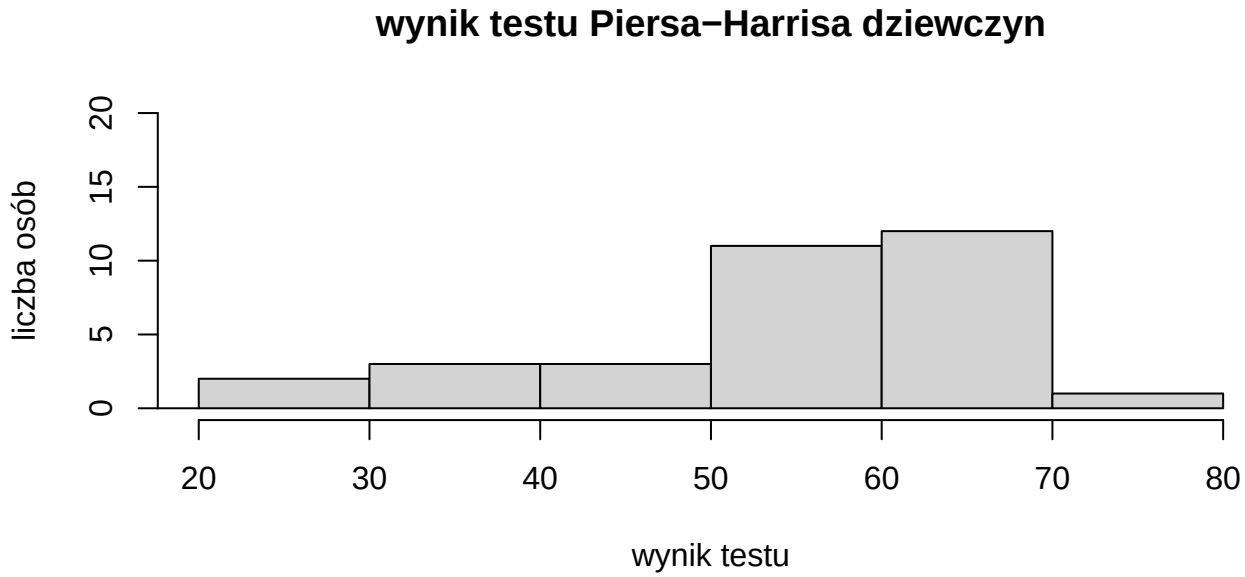
Jak widzimy histogramy są dość podobne. Wiemy również że kwartyle i mediana również są bardzo podobne dla obu płci. Jedyną różnicą którą da się zauważyć w naszych statystykach jest trochę niższa średnia ocen dla chłopców, ale jak widzimy na histogramie u chłopców mamy więcej obserwacji o bardzo niskiej ocenie. A wartości skrajne bardzo wpływają na średnią. Poza tym drobnym szczegółem rozkład ocen u obu płci jest podobny





Przechodząc jednak do analizy IQ ze względu na płeć na histogramach widzimy wyraźną różnicę między chłopcami i dziewczynkami. U chłopców widzimy dość duży odsetek wyników powyżej 120 punktów a u dziewczynek takie wyniki są już wyjątkami. Potwierdzają to również obliczone statystyki. Kwartyle mediana i średnia dla chłopców są wyższe od wyników dziewczyn.

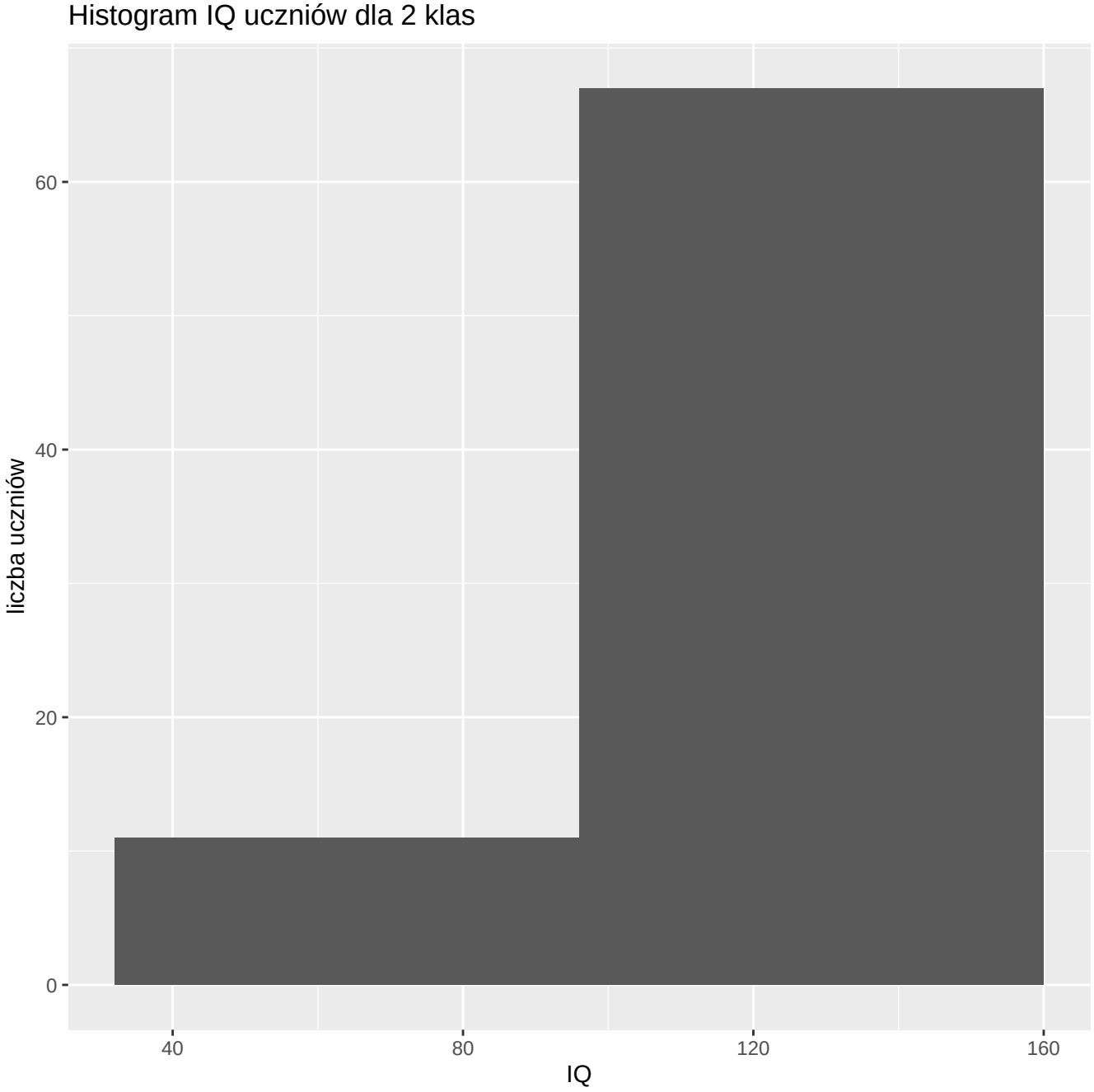


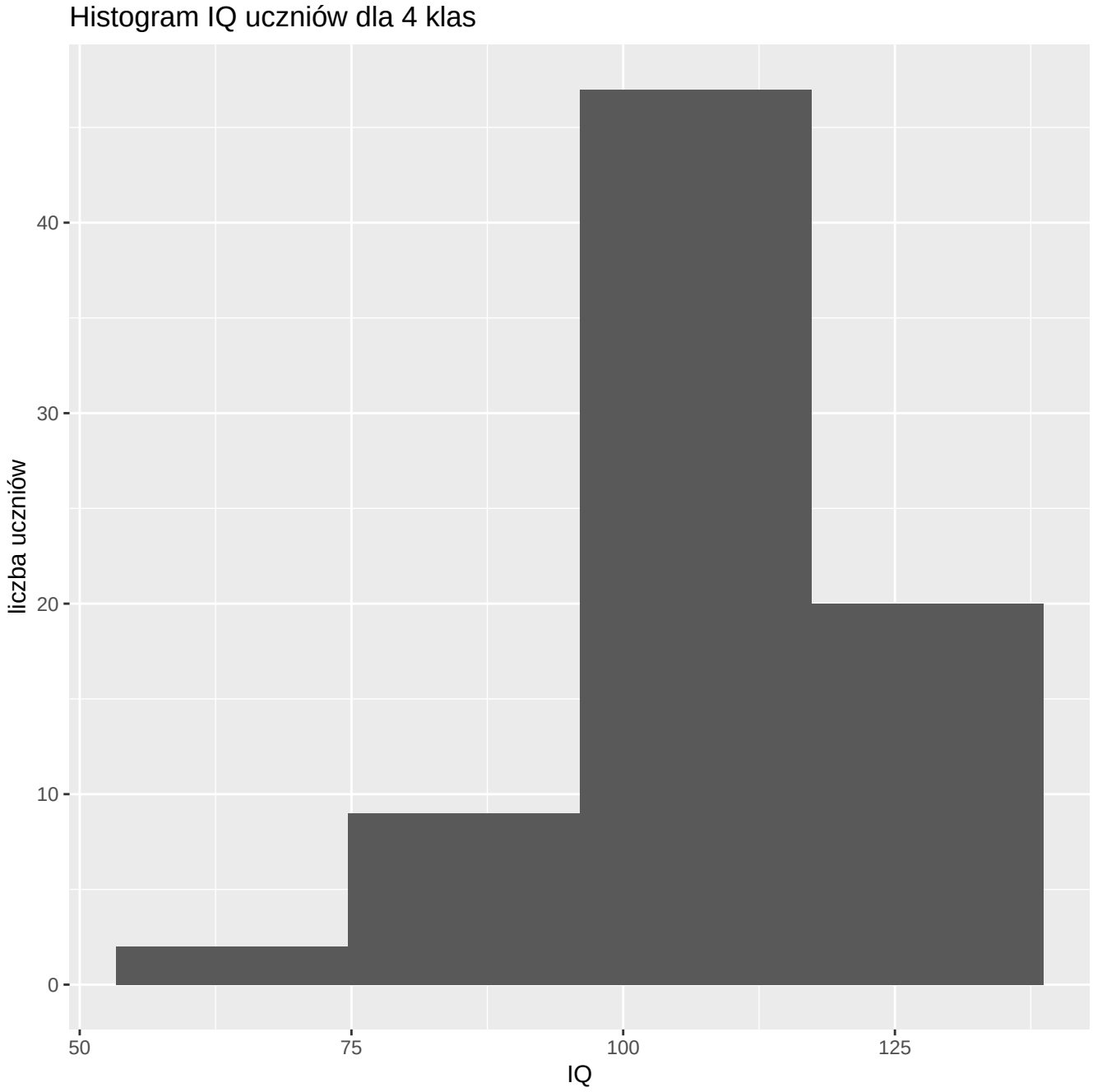


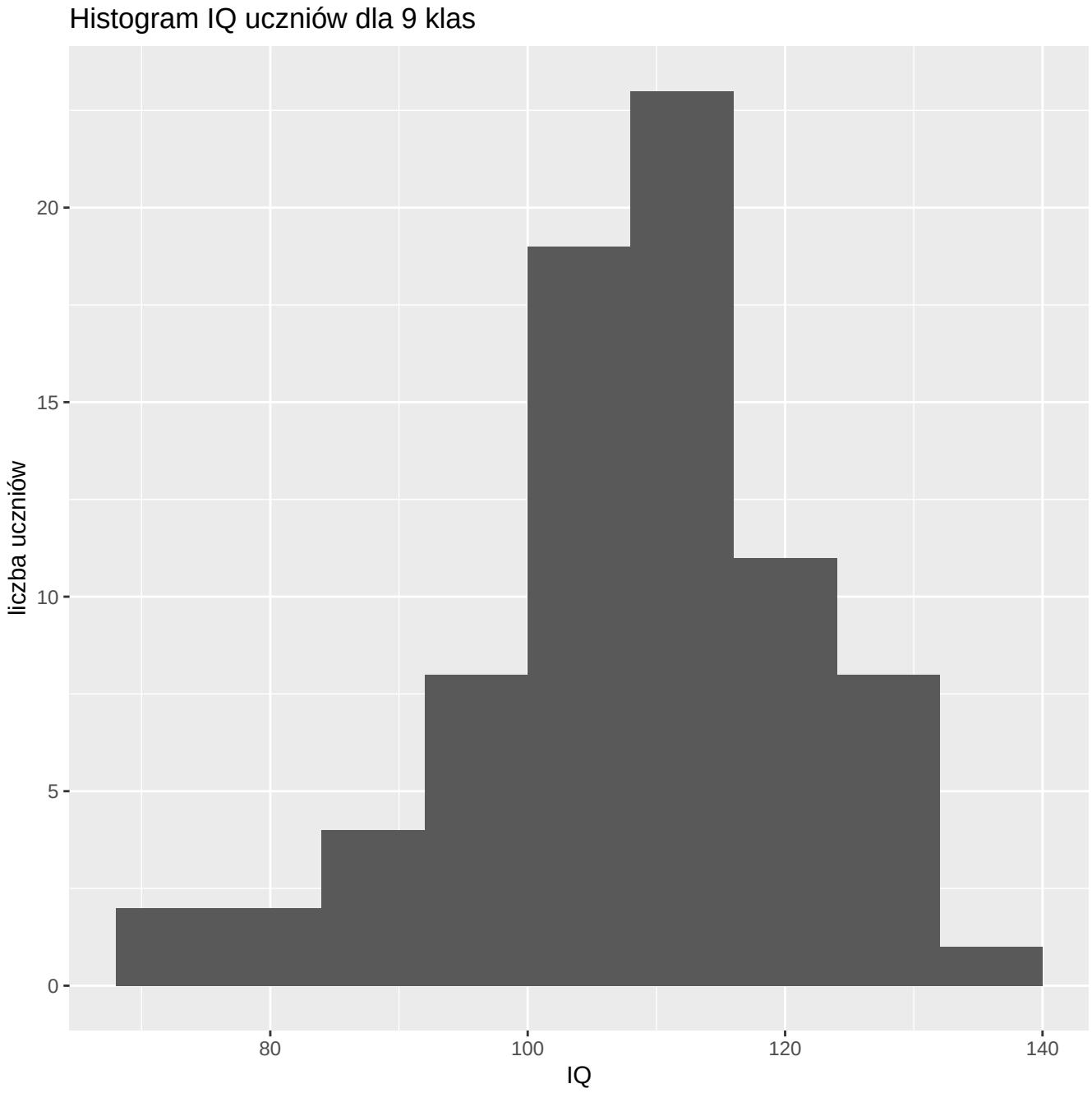
Jeśli chodzi o histogramy obejmujące wyniki testu Piersa-Harrisa to rozkład dla obu płci jest dość podobny z wyjątkiem tego, że wśród najlepszych wyników(70 i więcej) jest wyraźnie większy odsetek chłopców(nawet biorąc poprawkę na to, że dziewczyn jest mniej).Ma to zresztą swoje odwzorowanie w kwartylach. Pierwszy i drugi kwartyl są nieznacznie wyższe dla dziewczyn (odpowiednio 1,75 i 1 punkt więcej), jednakże trzeci kwartyl jest o 3 punkty wyższy dla chłopców. Wiele obserwacji z wysokim wynikiem u chłopców wpłynęło również na to że średnia jest o około 2 punkty wyższa dla chłopców.

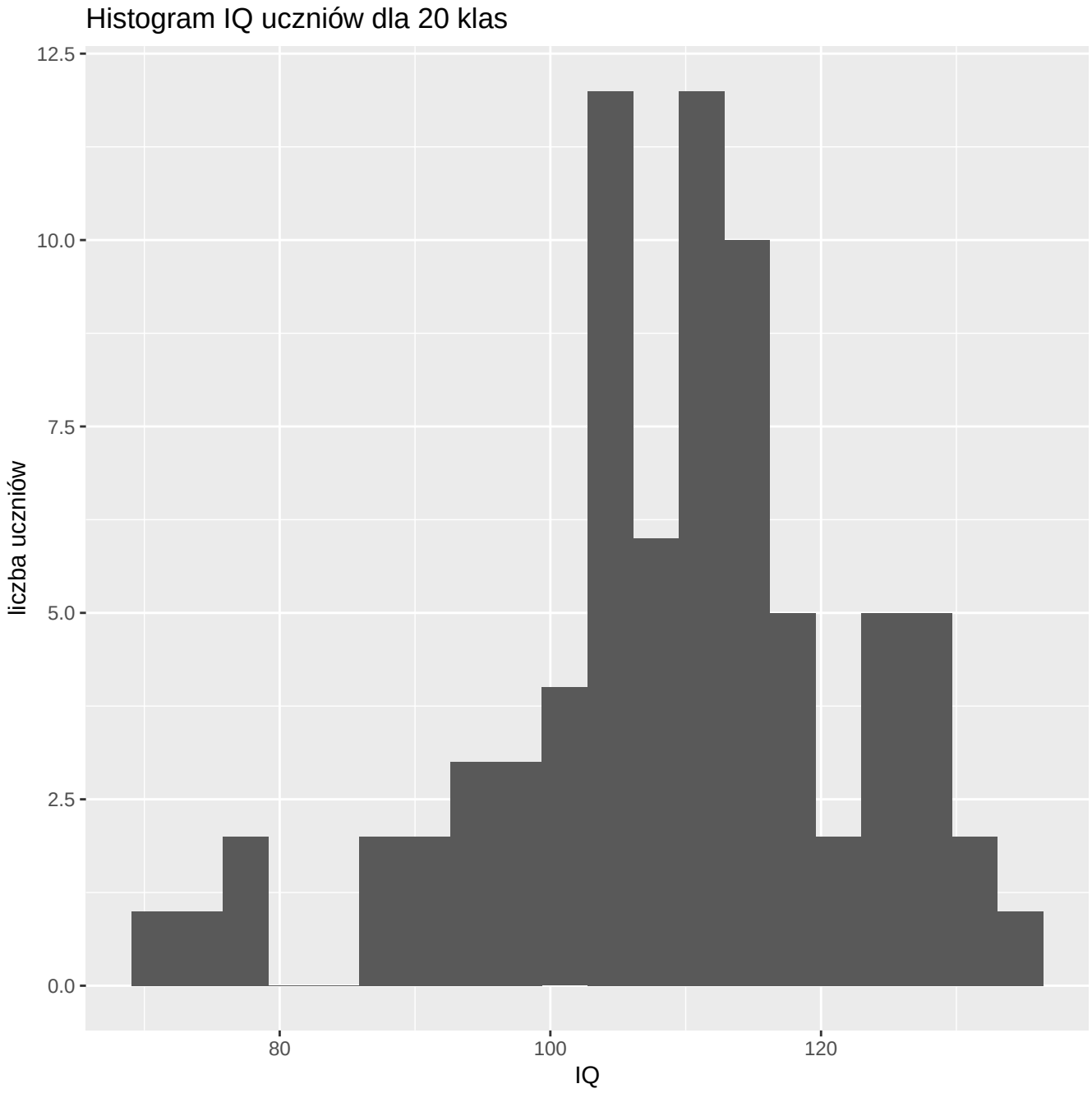
4. Histogramy dla różnej liczby klas

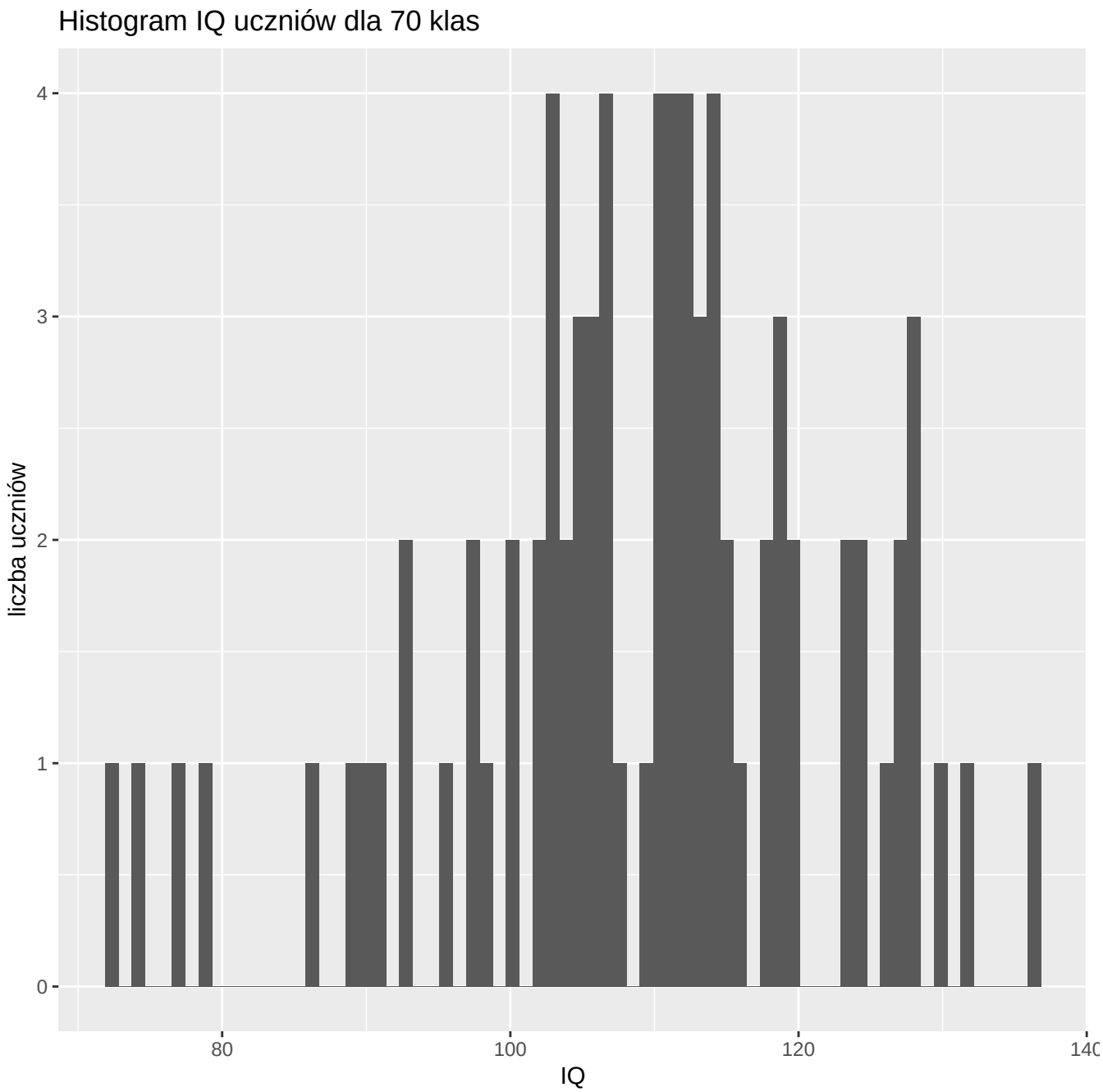
Narysujmy teraz histogramy zmiennej IQ ze zbioru grades dla różnej ilości klas. Zdecydowałem się narysować 5 histogramów które mają odpowiednio 2,4,9,20,70 koszyków.




Histogram IQ uczniów dla 4 klas
liczba uczniów
0
10
20
30
40
50
75
100
125
IQ


Histogram IQ uczniów dla 9 klas
liczba uczniów
0
5
10
15
20
80
100
120
140
IQ


Histogram IQ uczniów dla 20 klas
liczba uczniów
12.5
10.0
7.5
5.0
2.5
0.0
80
100
120
IQ



Prawidłową liczbę koszyków policzyłem ze wzoru dla którego dla rozmiaru danych n liczymy pierwiastek z tej liczby i zaokrąglamy w górę do najblizszej liczby całkowitej. Jak widzimy, dla bardzo małej liczby klas kształt rozkładu nie jest dobrze rozpoznawalny. Gdy liczba klas rośnie do optymalnej, która dla naszych danych wynosi 9 koszyków, to wtedy mamy najlepszą możliwość aby rozpoznać jaki jest rozkład danych oraz histogram wygląda estetycznie. Gdy liczba klas jeszcze bardziej rośnie to w histogramie robią się "dziury" i wykres przestaję być czytelny.

5. Analiza danych incame.dat

Zbiór income.dat zawiera dane zebrane przez Bureau of Labor Statistics obejmujące informacje o 55899 osobach i dla każdej obserwacji mamy 6 zmiennych.Dane obejmują

— numer identyfikacji osoby
— wiek w latach
— wykształcenie w skali od 1 do 6 (1 = podstwawowe, 2 = niepełne średnie, 3 = średnie, 4 = niepełne wyższe, 5 = wyższe (licencjat), 6 = wyższe (magisterium))
— płeć gdzie 1 oznacza mężczyznę a 2 kobietę,
— roczne zarobki ankietowanego w dolarach(przy czym z racji tego, że mamy sporo danych o ujemnych zarobkach rozsądnym będzie przypuścić, że chodzi tutaj o przychód)
— sektor zatrudnienia (5 = sektor prywatny, 6 = sektor publiczny, 7 = samozatrudnienie).

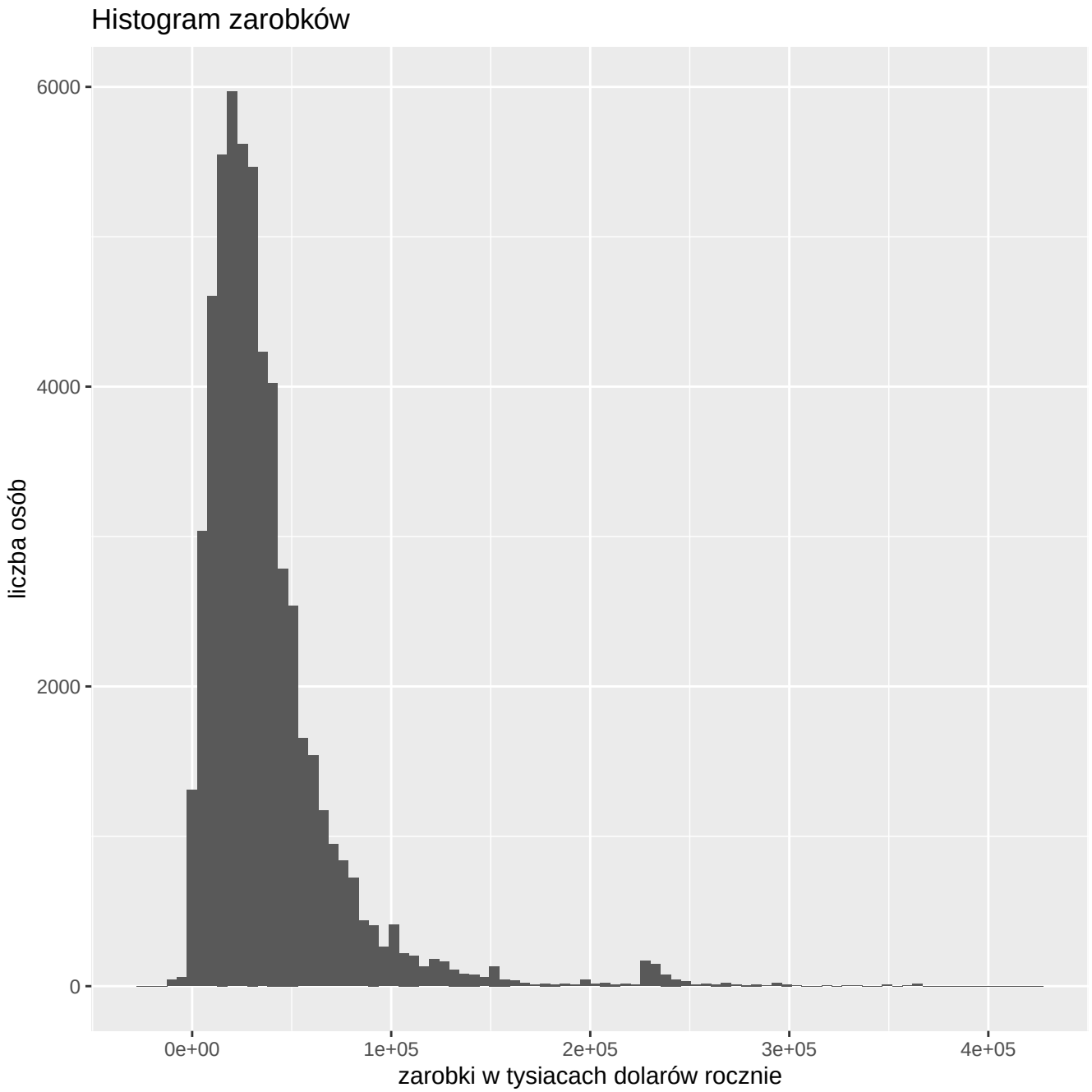
6. Tabela statystyk dla danych income

Rozpocznijmy od statystyk dla naszych danych. W jednej tabelce przedstawiono dane zbiorcze jak i te z podziałem na płcie.

	min	max	mediana	1Q	3Q	średnia	odch	wariancja	wz
ogółem	-24998	425510	29717	17000	46504	37865	36158	1307402975	0.95
mężczyzni	-24998	425510	36000	22146	55852	46489	41978	1762121911	0.90
kobiety	-9999	385068	23012	13004	36200	28422	25279	639035926	0.89

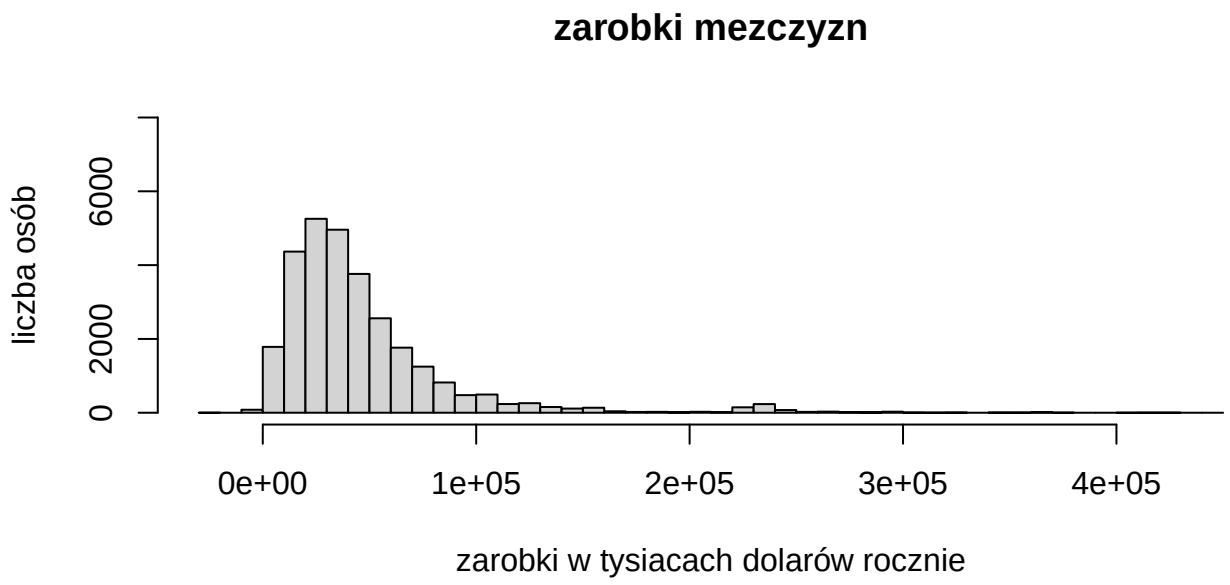
Z powyższej tabelki widzimy, że dane mają bardzo duży rozstęp rzędu około 450 tysięcy dolarów. Zauważmy też, że współczynnik zmienności dla zarobków jest bardzo duży. Powodem tego jest wysoka wariancja i odchylenie standardowe.Odnotujmy również, że dla statystyk uwzględnionych w tabelce wyraźnie widzimy, że zarówno kwartyle jak i średnia są o wiele wyższe dla mężczyzn. Dla pierwszego kwartyla jest to 10 tysięcy dolarów więcej a dla trzeciego kwartyla aż 20 tysięcy więcej.

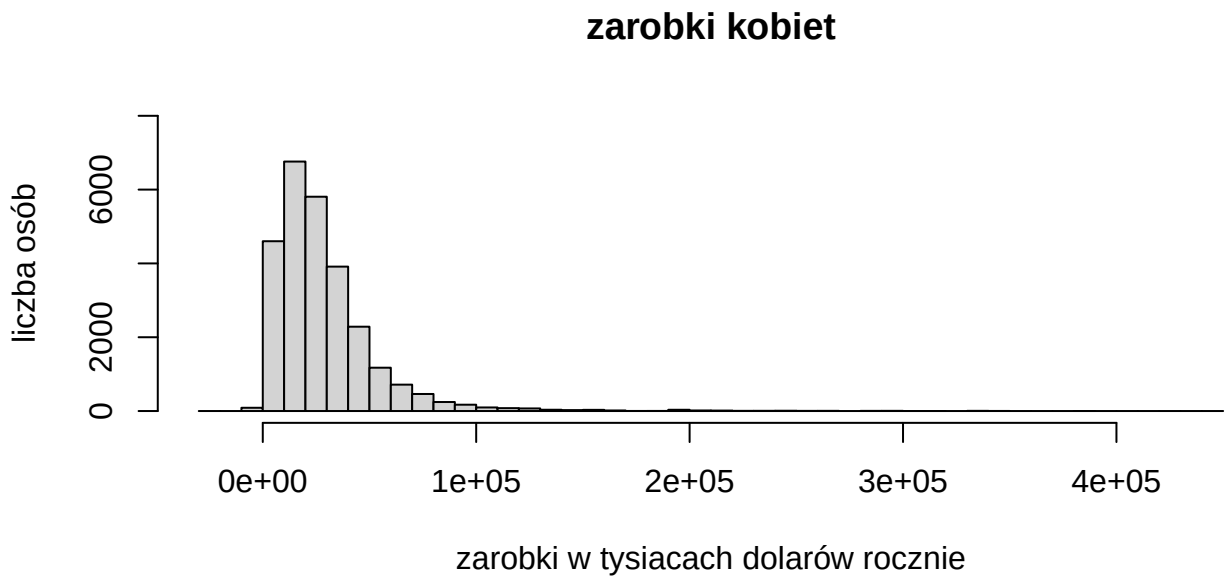
Zbadajmy teraz jak wygląda histogram zarobków dla naszych danych



Histogram zarobków jest asymetryczny i skośny w prawo.Mediana(29717 dolarów) jest istotnie niższa od średniej(37865 dolarów) oraz wykres jest jednomodalny. Bardzo duży odsetek danych skupiony jest na lewym krańcu wykresu. Świadczą o tym niskie wartości kwartyli. Pierwszy kwartyl to tylko 17 tysięcy dolarów, a trzeci kwartyl to 46.5 tysięcy. Dodatkowo mamy również stosunkowo niski rozstęp międzykwartylowy(około 30 tysięcy dolarów). Jednocześnie w danych mamy dużo obserwacji odstających w górę i są to osoby zarabiające kwoty stu tysięcy dolarów rocznie i więcej.

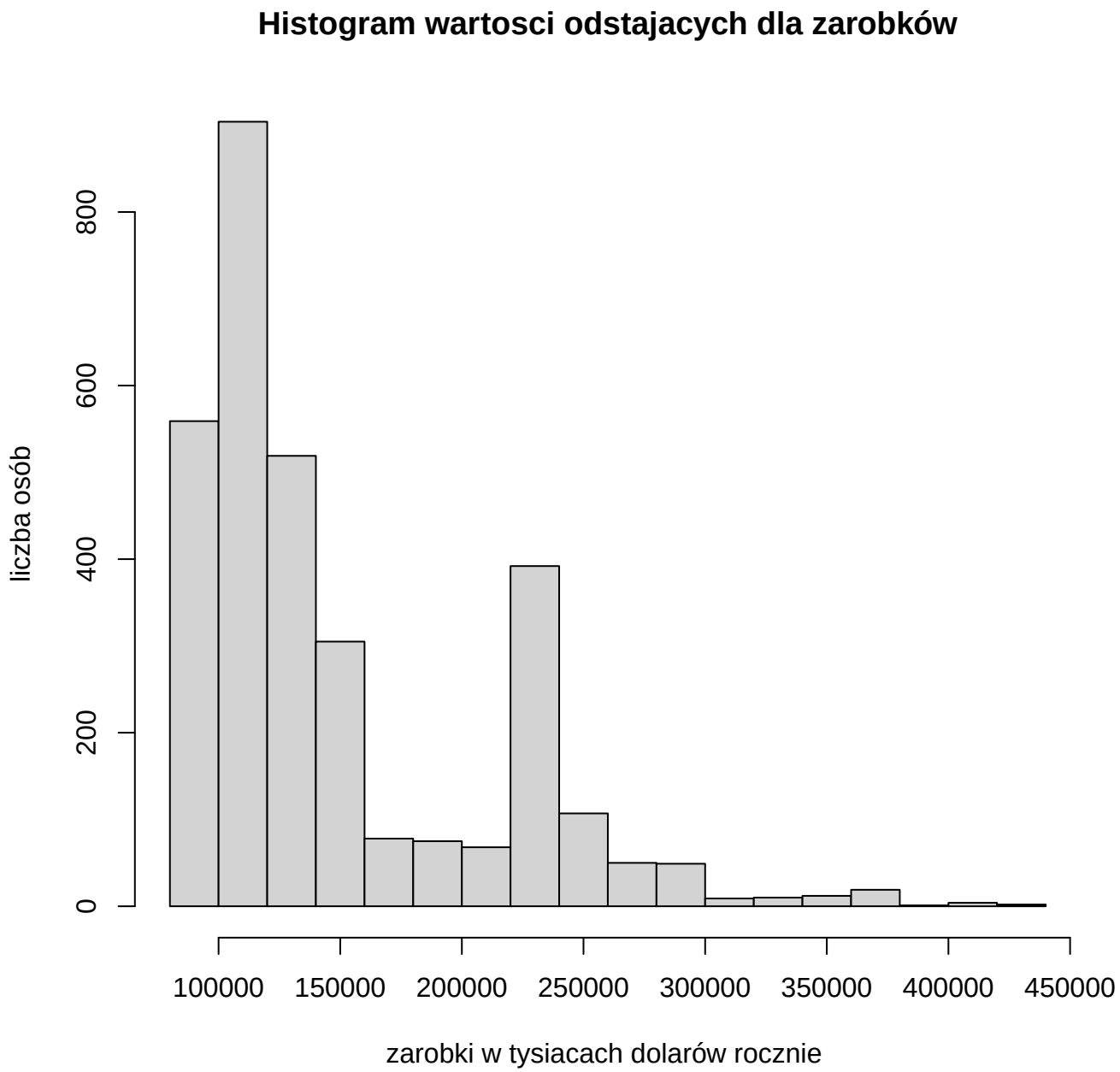
Popatrzmy teraz na histogramy zarobków z podziałem na płcie.





Najpierw odnotujmy, że dla czytelności histogramy mają ustawione takie same skale na obu osiach oraz taką samą szerokość słupków z racji tego że liczba kobiet jest porównywalna do liczby mężczyzn(około 27000 kobiet i około 29000 mężczyzn). Dla mężczyzn dominanta histogramu przypada na przedział od 20 do 30 tysięcy dolarów a dla kobiet na przedział od 10 do 20 tysięcy. Na powyższych histogramach widzimy również, że dla mężczyzn istnieje spory odsetek obserwacji gdzie mężczyźni zarabiają około 100 tysięcy dolarów, a dla kobiet wraz ze wzrostem zarobków odsetek kobiet mogących się pochwalić dobrymi zarobkami drastycznie spada.

Na poniższym histogramie uwzględniono wszystkie obserwacje odstające jeśli chodzi o zarobki



Jak widzimy obserwacji odstających jest bardzo dużo a dokładnie 3163. Nie jest to jednak zaskoczeniem ponieważ Stany Zjednoczone są krajem dużych nierówności społecznych i o ile ogromna część społeczeństwa zarabia niewiele to jest również pewna grupa Amerykanów których zarobki istotnie odstają w górę od reszty.

7. Podsumowanie

Jak widzimy z powyższej analizy płeć w pewien sposób różnicuje badanych uczniów i dorosłych pod względem badanych cech. Jeśli chodzi o uczniów amerykańskiej szkoły to o ile oceny i wynik testu Piersa-Harrisa mają dość podobny rozkład dla obu płci to już wynik testu IQ ma średnio wyższe wyniki dla chłopców. Badając zarobki dorosłych Amerykanów ujawnia się zjawisko które socjologowie i ekonomiści określają mianem gender pay gap `https://www.glassdoor.com/research/gender-pay-gap-2019/#` W zalinkowanym źródle możemy potwierdzić, że zaobserwowana w naszych danych różnica zarobków kobiet i mężczyzn na korzyść mężczyzn jest ogólną prawidłowością w wielu krajach.