Karolina Bakalarz i Amelia Bieda

Analiza wybranych danych rzeczywistych z wykorzystaniem metod statystyki opisowej

1. Wstęp

1.1. Cel

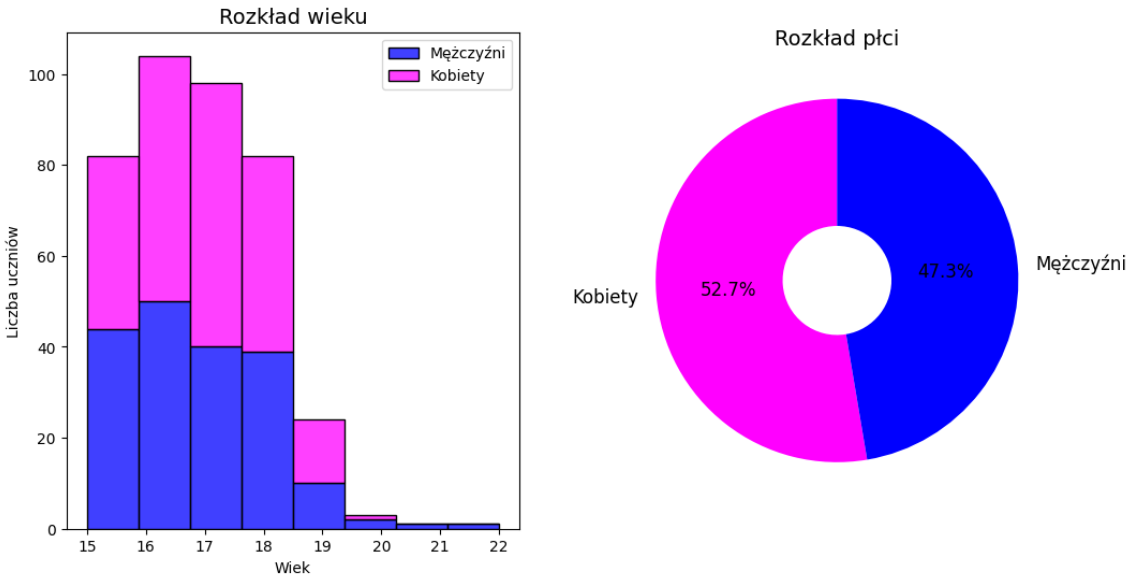
Celem naszej pracy jest porównanie spożycia alkoholu między kobietami a mężczyznami - w trakcie tygodnia oraz przez weekend.

1.2. Opis danych

Zbiór danych został zaczerpnięty ze strony *Kaggle*.

Analizowana przez nas baza danych dotyczy spożycia alkoholu wśród 395 uczniów. 88% osób jest ze szkoły Gabriel Pereira, reszta z Mousinho da Silveira. 52,7% (208 osób) badanych to kobiety, natomiast 47,3% (187 osób) to mężczyźni. Większość uczniów jest w przedziale wiekowym 15-18, najstarsza osoba ma 22 lata.

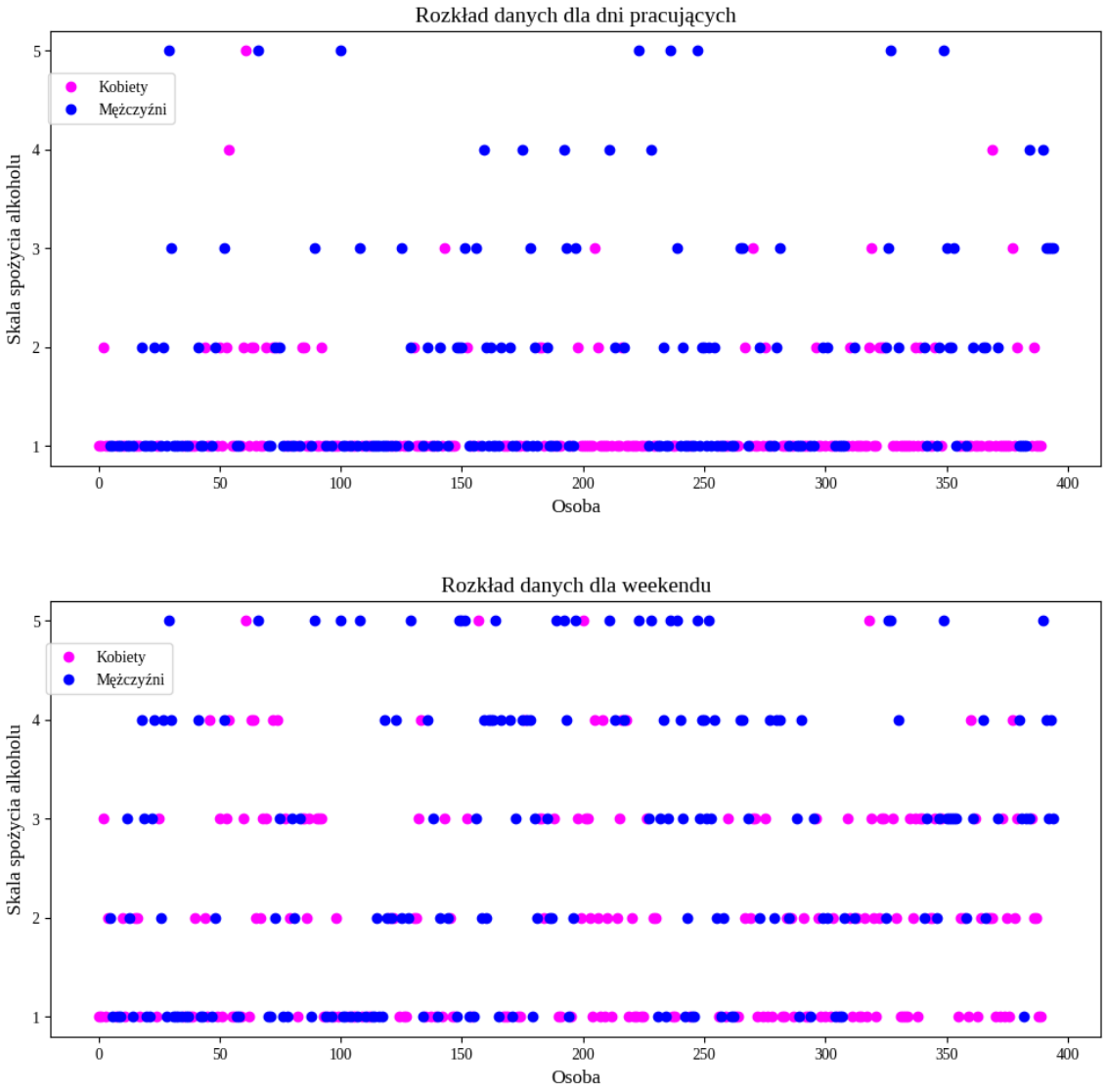
Rysunek 1. Histogramy rozkładu wieku oraz płci



Zajmiemy się interpretacją odpowiedzi uczniów na ankietę dotyczącą ilości spożywanego alkoholu w ciągu tygodnia. Możliwe odpowiedzi są w skali

od 1 do 5, gdzie 1 to bardzo niskie (będziemy uznawać za brak konsumpcji alkoholu), 5 to bardzo wysokie. Ankieta ta została podzielona na dwie części
— DALC (*workday alcohol consumption*) przedstawia średnie spożycie alkoholu od poniedziałku do piątku,
— WALC (*weekend alcohol consumption*) dotyczy średniego spożycia alkoholu w ciągu weekendu.

Rysunek 2. Rozkłady punktów danych dla badanych statystyk



Ponad 88% zbadanych osób nie pije, bądź spożywa bardzo mało alkoholu w ciągu tygodnia. Niespełna 4,5% spożywa wtedy dużo alkoholu (tabela 1). Konsumpcja alkoholu wzrasta na koniec tygodnia, gdzie niecałe 60% pije bardzo mało, ale 20% ma alkohol pod dostatkiem (tabela 2). Są również uczniowie, którzy nie piją alkoholu w tygodniu, ale piją w weekendy (tabela 3).

Tabela 1. Odpowiedzi - spożycie alkoholu w dni pracujące

Skala odpowiedzi	Liczebność	Procent
1	276	69,87 %
2	75	18,99 %
3	26	6,58 %
4	9	2,28 %
5	9	2,28 %

Tabela 2. Odpowiedzi - spożycie alkoholu przez weekend

Skala odpowiedzi	Liczebność	Procent
1	151	38,23 %
2	85	21,52 %
3	80	20,25 %
4	51	12,91 %
5	28	7,09 %

Tabela 3. Odpowiedzi dla WALC udzielone przez osoby niepijące w dni powszednie

Skala odpowiedzi	Liczebność	Procent
1	150	54,35 %
2	65	23,55 %
3	42	15,22 %
4	15	5,43 %
5	4	1,45 %

2. Podstawowe statystyki

2.1. Miary położenia

Średnia arytmetyczna to klasyczna miara położenia, tendencji centralnej, będąca ilorazem sumy zaobserwowanych wartości zmiennej mierzalnej przez liczbę obserwacji.

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

Tabela 4. Wyniki dla średniej arytmetycznej

	DALC	WALC
Kobiety	1.2548	1.9567
Mężczyźni	1.7326	2.6631

Średnia harmoniczna to klasyczna miara położenia, równa odwrotności średniej arytmetycznej z odwrotności poszczególnych wartości badanej cechy.

$$x_h = n \div \sum_{i=1}^{n} \frac{1}{x_i}.$$

Tabela 5. Wyniki dla średniej harmonicznej

	DALC	WALC
Kobiety	1.1191	1.4986
Mężczyźni	1.3324	1.8753

Średnia geometryczna to klasyczna miara położenia, będąca pierwiastkiem stopnia n z iloczynu n wartości zmiennej.

$$x_g = \sqrt[n]{\prod_{i=1}^{n} x_i}.$$

Tabela 6. Wyniki dla średniej geometrycznej

	DALC	WALC
Kobiety	1.1705	1.7022
Mężczyźni	1.4915	2.2535

Średnia winsorowska jest statystyczną miarą tendencji centralnej zbliżoną do zwykłej średniej arytmetycznej lub mediany, a najbardziej podobną do średniej ucinanej.

$$x_w = \frac{1}{n}[(k+1)x_{k+1} + \sum_{i=k+2}^{n-k-1} x_i + (k+1)x_{n-k}].$$

Tabela 7. Wyniki dla średniej winsorowskiej

	DALC	WALC
Kobiety	1.2548	1.9567
Mężczyźni	1.7326	2.6631

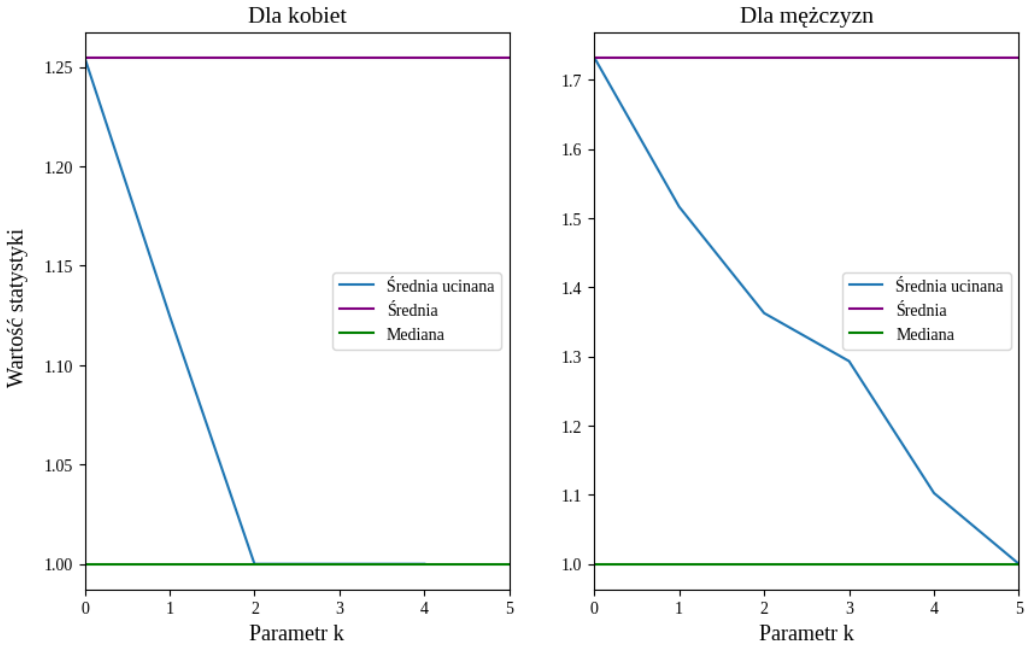
Średnia ucinana - jej obserwacje porządkuje się od najmniejszej do największej, następnie odrzuca mały procent najbardziej ekstremalnych obserwacji na obu krańcach (wartości najmniejsze oraz największe w próbce), a następnie oblicza się średnią z pozostałych obserwacji.

$$x_u = \frac{1}{n-2k} \sum_{i=k+1}^{n-k} x_i.$$

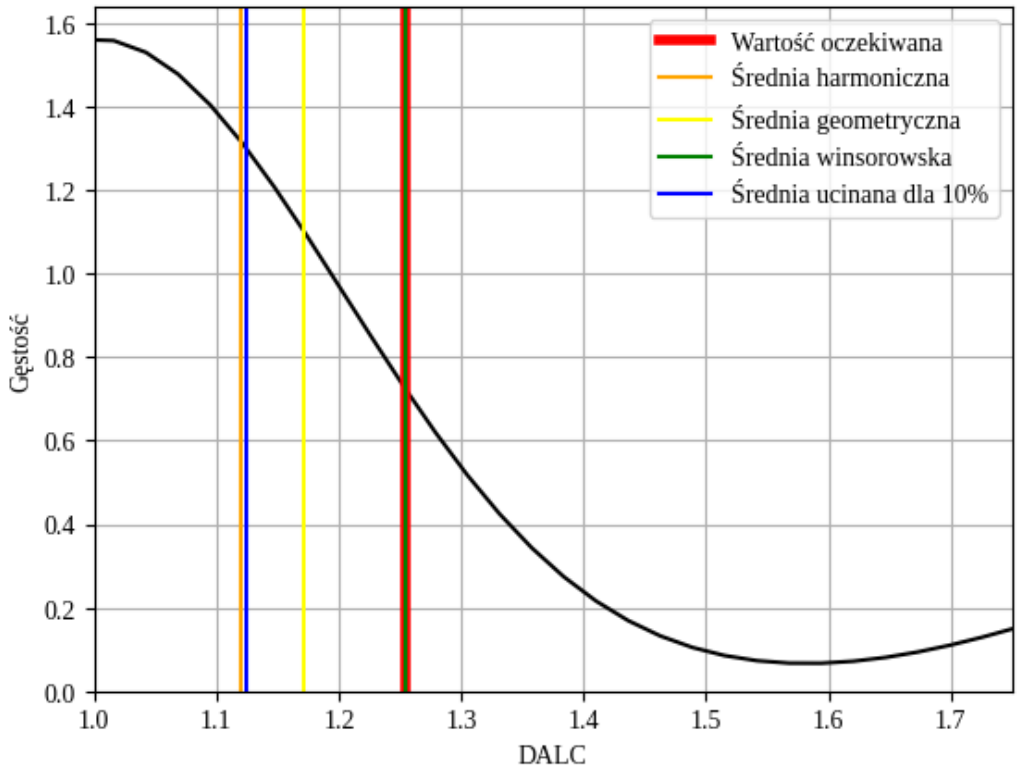
Tabela 8. Wyniki dla średniej ucinanej
z odrzuceniem 10% największych i najmniejszych wartości

	DALC	WALC
Kobiety	1.1250	1.8155
Mężczyźni	1.5166	2.5828

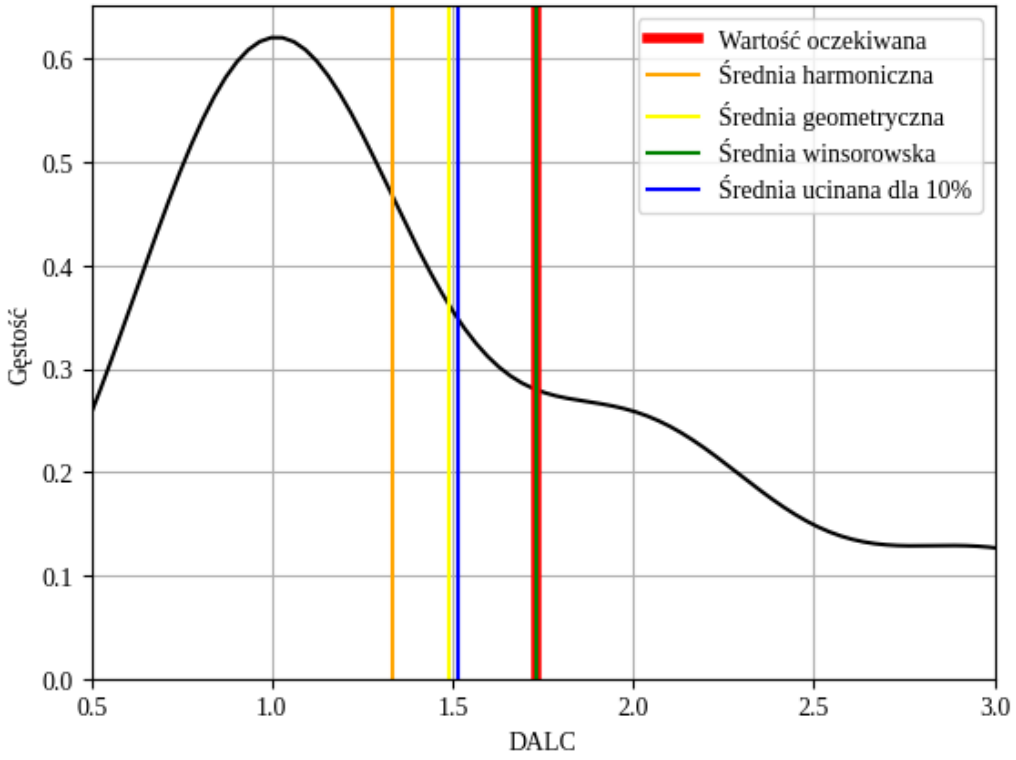
Rysunek 3. Wykresy średniej ucinanej z parametrem k dla dni pracujących



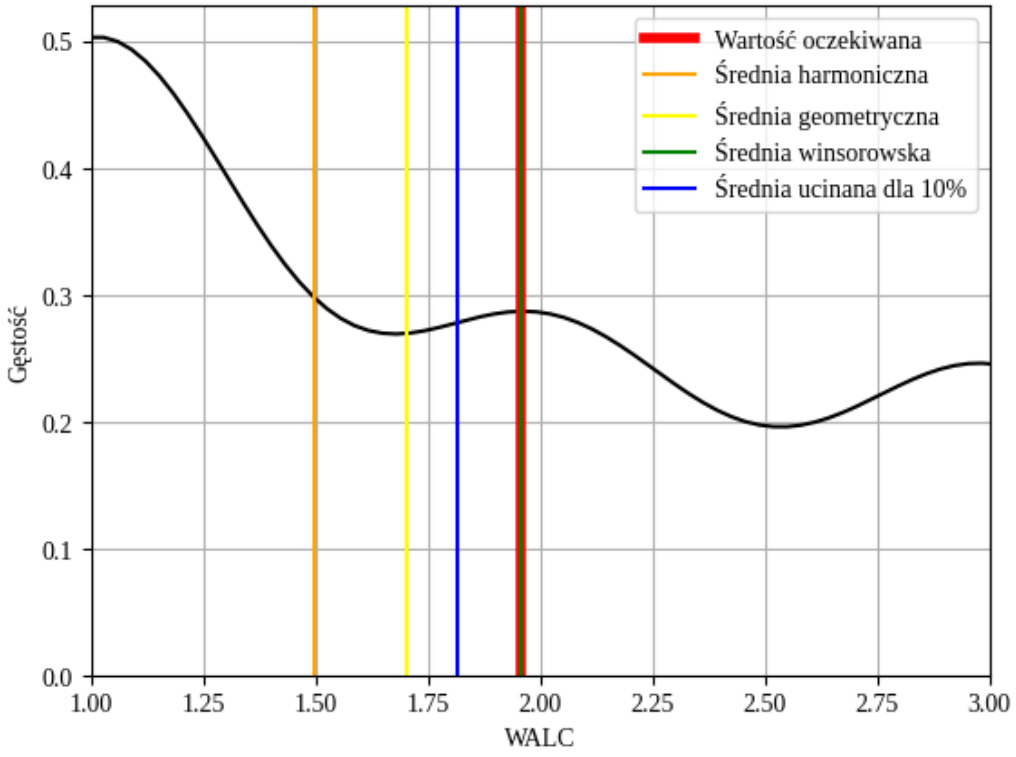
Rysunek 4. Porównanie średnich dla DALC u kobiet



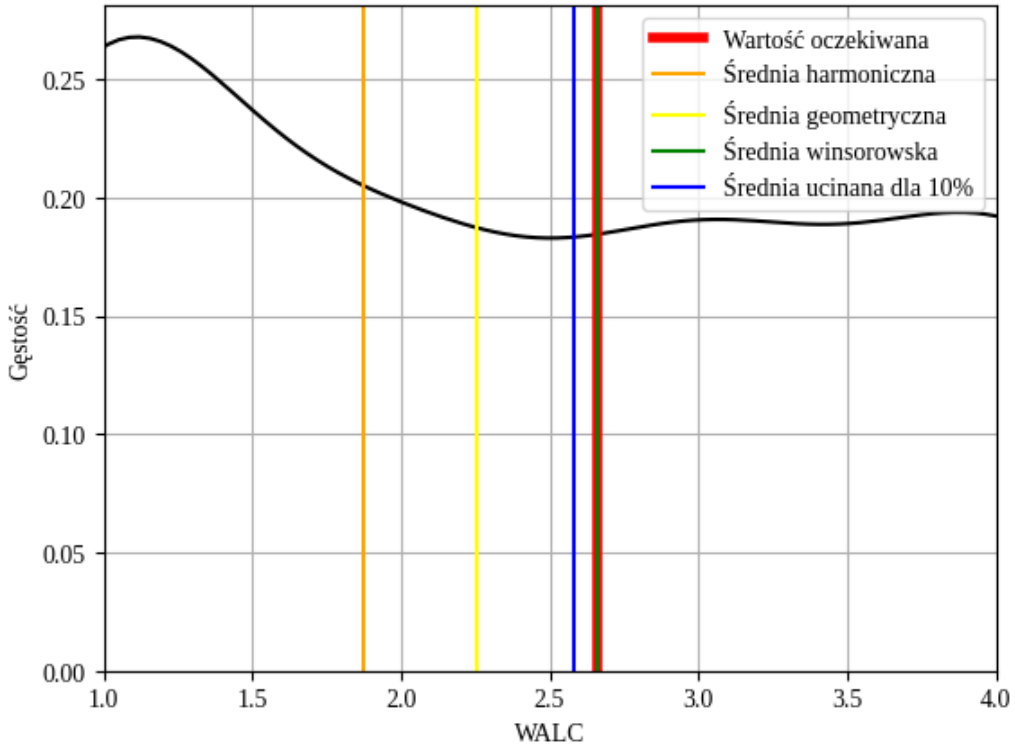
Rysunek 5. Porównanie średnich dla DALC u mężczyzn



Rysunek 6. Porównanie średnich dla WALC u kobiet



Rysunek 7. Porównanie średnich dla WALC u mężczyzn



Pierwszy kwartyl Q1 to mediana obserwacji mniejszych od Q2 (wartość, poniżej której znajduje się 25% danych).

Tabela 9. Wyniki dla pierwszego kwartylu

	DALC	WALC
Kobiety	1.0	1.0
Mężczyźni	1.0	1.0

Mediana to drugi kwartyl (Q2).

$$x_{med} = \begin{cases} x_{(n+1)/2}, & \text{gdy n nieparzyste,} \\ 0.5(x_{n/2} + x_{(n/2)+1}) & \text{gdy n parzyste.} \end{cases}$$

Tabela 10. Wyniki dla mediany

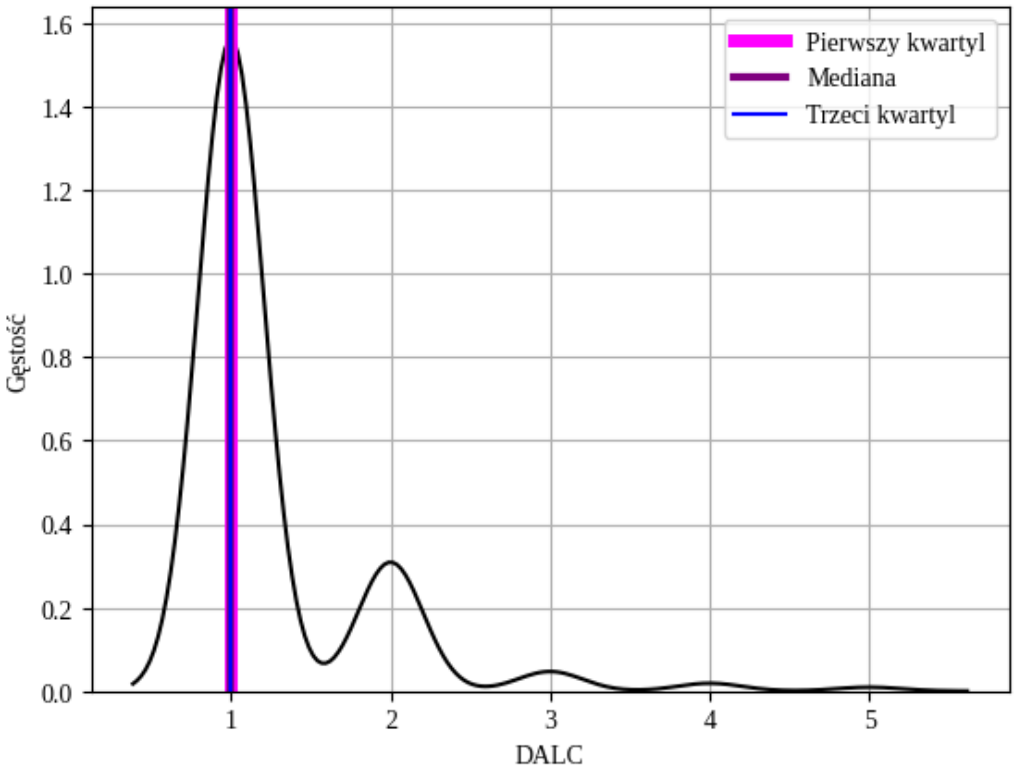
-	DALC	WALC
Kobiety	1.0	2.0
Mężczyźni	1.0	3.0

Trzeci kwartyl Q3 to mediana obserwacji większych od Q2 (75% danych jest mniejszych niż Q3).

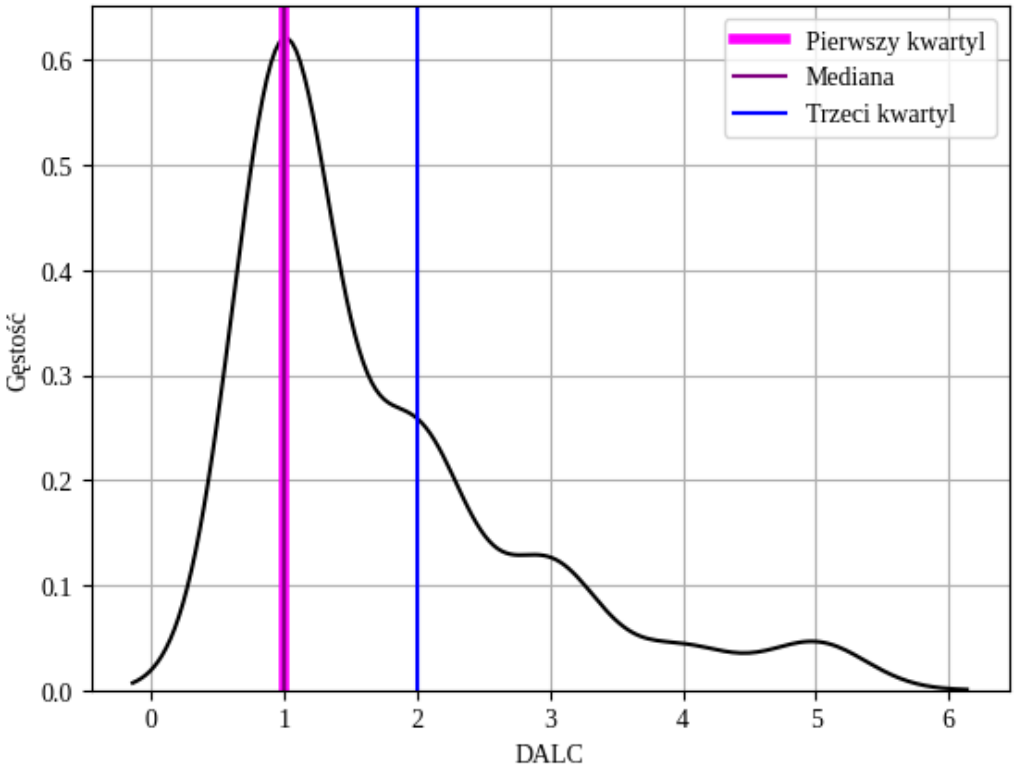
Tabela 11. Wyniki dla trzeciego kwartylu

	DALC	WALC
Kobiety	1.0	3.0
Mężczyźni	2.0	4.0

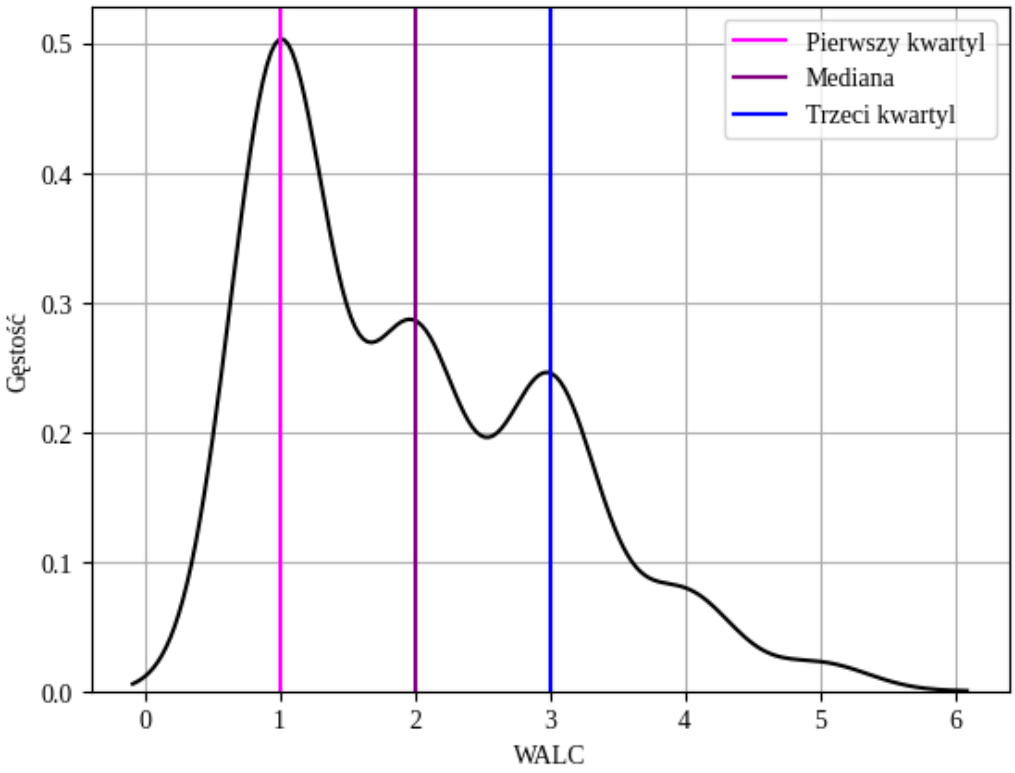
Rysunek 8. Porównanie kwartyli dla DALC u kobiet



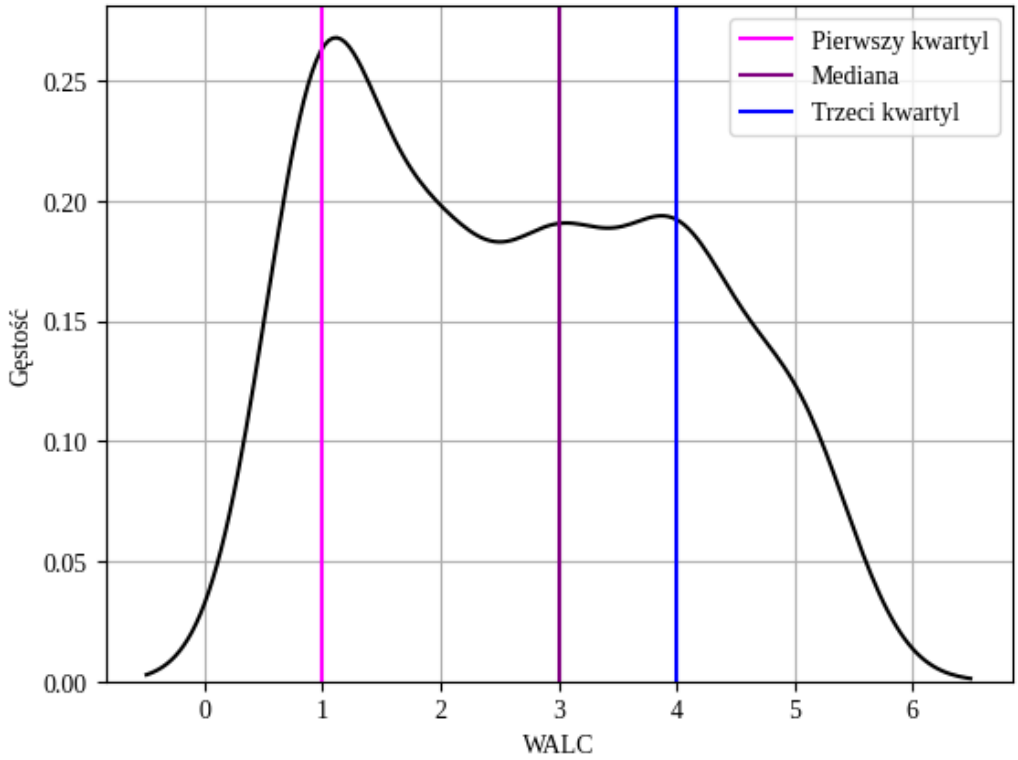
Rysunek 9. Porównanie kwartyli dla DALC u mężczyzn



Rysunek 10. Porównanie kwartyli dla WALC u kobiet



Rysunek 11. Porównanie kwartyli dla WALC u mężczyzn



2.2. Miary rozproszenia

Rozstęp międzykwartylowy to różnica między wartością trzeciego i pierwszego kwartyla, która obejmuje 50% obserwacji.

$$IQR = Q3 - Q1.$$

Tabela 12. Wyniki dla rozstępu międzykwartylowego

-	DALC	WALC
Kobiety	0.0	2.0
Mężczyźni	1.0	3.0

Rozstęp to pozycyjna miara dyspersji liczona jako różnica między największą a najmniejszą wartością badanej cechy mierzalnej zaobserwowaną w próbie.

$$R = x_{max} - x_1.$$

Tabela 13. Wyniki dla rozstępu

-	DALC	WALC
Kobiety	4.0	4.0
Mężczyźni	4.0	4.0

Odchylenie standardowe jest pierwiastkiem kwadratowym z wariancji, informuje o tym, jak szeroko wartości danej wielkości są rozrzucone wokół jej średniej.

$$s = \sqrt{Var(X)},$$

gdzie $Var(X)$ to wariancja.

Tabela 14. Wyniki dla odchylenia standardowego

-	DALC	WALC
S Kobiety	0.5944	1.0530
Mężczyźni	1.0763	1.4138

Odchylenie przeciętne od wartości średniej to średnia arytmetyczna z odchyleń bezwzględnych dla wszystkich elementów zbioru danych.

$$d_i = \frac{1}{n}\sum_{i=1}^{n} |x_i - \overline{x}|$$

Tabela 15. Wyniki dla odchylenia przeciętnego od wartości średniej

-	DALC	WALC
Kobiety	0.5944	1.0530
Mężczyźni	1.0763	1.4138

Współczynnik zmienności to iloraz odchylenia standardowego i odpowiadającej jej miary średniej.

$$\nu = (s \div \overline{x}) \cdot (100\%)$$

Tabela 16. Wyniki dla współczynnika zmienności

-	DALC	WALC
Kobiety	47.32%	53.81%
Mężczyźni	62.12%	53.09%

2.3. Miary asymetrii

Współczynnik skośności służy do badania kierunku i siły asymetrii rozkładu cechy.

$$\alpha = \frac{n}{(n-1)(n-2)}\sum_{i=1}^{n}\left(\frac{x_i - \overline{x}}{s}\right)^3$$

Tabela 17. Wyniki dla współczynnika skośności

-	DALC	WALC
Kobiety	3.0072	0.8276
Mężczyźni	1.5497	0.2345

2.4. Miary spłaszczenia

Kurtoza określa intensywność występowania wartości skrajnych, mierzy więc ona, co się dzieje w "ogonach" rozkładu.

$$K = \frac{n-1}{(n-2)(n-3)}\Big((n+1)K_1 - 3(n-1)\Big) + 3, \text{gdzie } K_1 = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_1 - \overline{x})^4}{(\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2)^2}$$

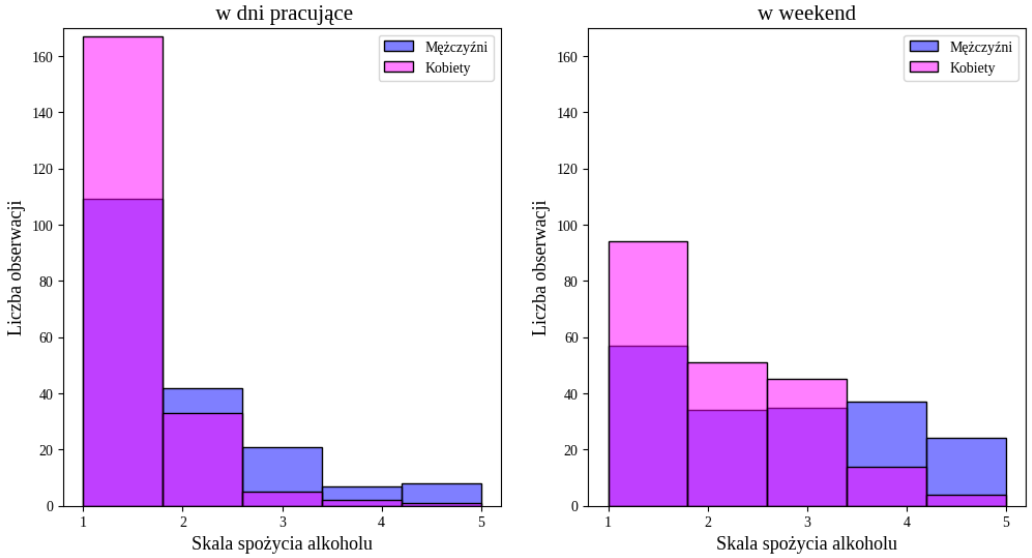
Tabela 18. Wyniki dla kurtozy

-	DALC	WALC
Kobiety	11.1549	-0.1871
Mężczyźni	1.7123	-1.2907

3. Wizualizacja danych

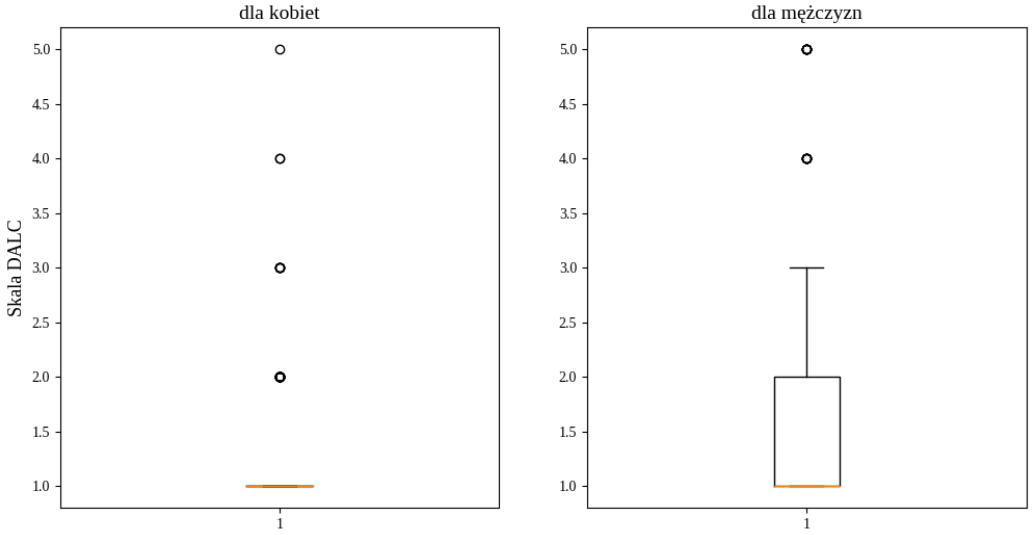
Histogramy

Rysunek 12. Histogramy liczebności dla spożycia alkoholu

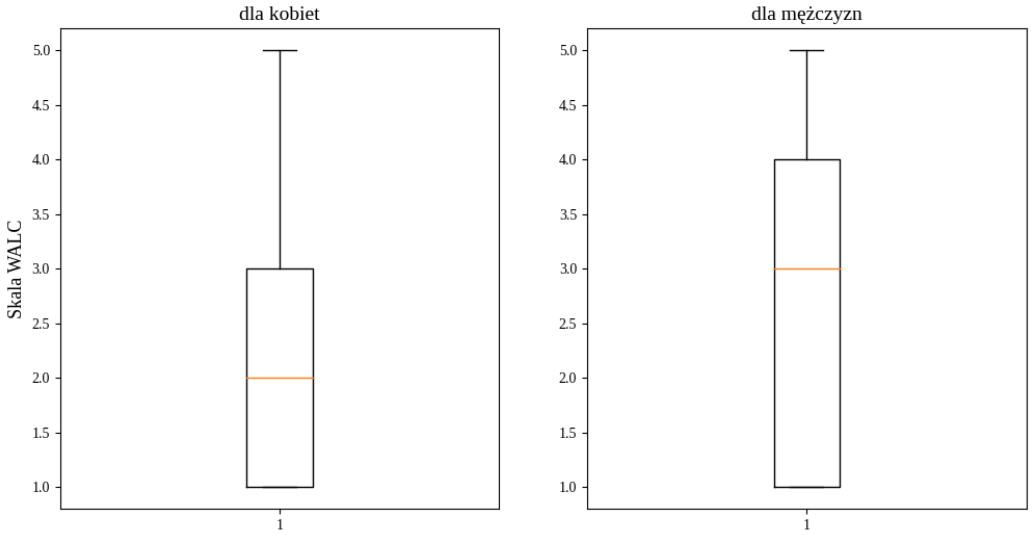


Wykresy pudełkowe

Rysunek 13. Wykresy pudełkowe spożycia alkoholu w dni pracujące

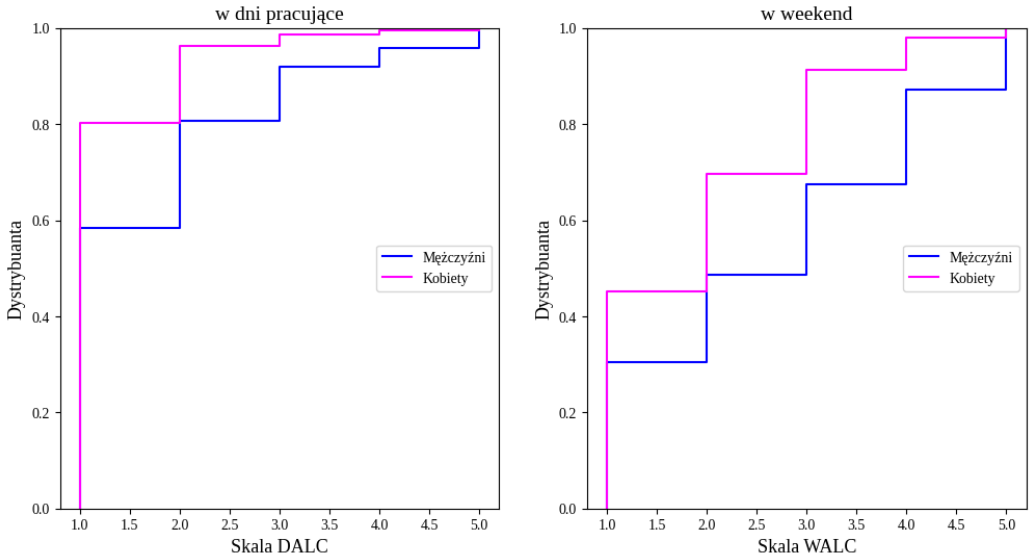


Rysunek 14. Wykresy pudełkowe spożycia alkoholu w weekend



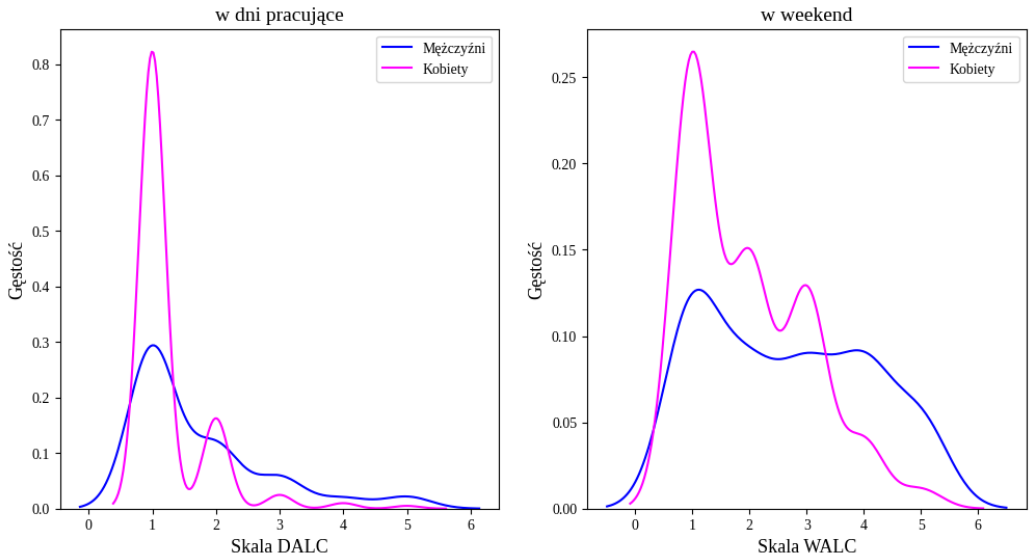
Dystrybuanty empiryczne

Rysunek 15. Dystrybuanty empiryczne dla spożycia alkoholu



Gęstości empiryczne

Rysunek 16. Gęstości empiryczne dla spożycia alkoholu



4. Interpretacja wyników

Z przeprowadzonej analizy danych wynika, że średnio badani uczniowie nie piją w ogóle lub piją bardzo mało alkoholu. Jego konsumpcja zwiększa się wraz z nadejściem weekendu, co przedstawiają wszystkie średnie. Średnia harmoniczna jest zawsze najmniejszą z reszty miar.

Rozstęp międzykwartylowy jest większy dla statystyki WALC niż dla DALC, co oznacza większą zmienność w skali odpowiedzi. To samo obserwujemy, porównując dane pod względem płci. Wyniki IQR pokazują, że dla dni pracujących odpowiedzi kobiet w większości są takie same, a dla mężczyzn mają małe rozproszenie. Patrząc na IQR widzimy również, dla statystyki weekendowej kobiety wciąż mają mniejszą zmienność odpowiedzi niż mężczyźni.

Współczynnik zmienności to sposób normalizacji odchylenia standardowego przez wartość średnią, co umożliwia porównanie zmienności między różnymi zmiennymi, niezależnie od ich jednostek miary. Im większy współczynnik zmienności, tym większa zmienność danych w stosunku do ich średniej. Współczynnik zmienności dla rozkładu spożycia alkoholu w dni powszednie przez mężczyzn wynosi aż 62.12%. Oznacza to, że odchylenie standardowe stanowi około 62.12% średniej wartości. Jest to dość wysoka zmienność danych. Dla pozostałych rozkładów zmienność danych jest umiarkowana

Współczynnik skośności w każdym porównywanym przez nas przypadku był dodatni. Skośność charakteryzuje stopień asymetrii rozkładu wokół jego średniej. Skośność dodatnia określa rozkład z asymetrią rozciągającą się w kierunku wartości dodatnich. W przypadku rozkładu normalnego skośność wynosi 0, co oznacza, że rozkład jest idealnie symetryczny. Analizowane przez nas dane, co możemy zaobserwować na pokazanych wyżej wykresach, posiadają rozkłady prawoskośne, prawe ramię rozkładu jest wydłużone. Najsilniejszą asymetrię prawostronną ma rozkład spożycia alkoholu w dni powszednie przez kobiety.

Kurtoza dla spożycia alkoholu w dni pracujące ma wartości dodatnie zarówno dla kobiet jak i mężczyzn oznacza to, że rozkład danych jest bardziej spiętrzony wokół średniej niż w rozkładzie normalnym. Może to sugerować, że istnieją pewne obszary danych, gdzie wartości są znacznie bardziej skoncentrowane niż przeciętnie.

Ujemna kurtoza dla spożycia alkoholu przez mężczyzn i kobiety w weekend wskazuje na bardziej spłaszczony rozkład danych w porównaniu z rozkładem normalnym. To oznacza, że dane są rozproszone bardziej równomiernie wokół średniej.

Próbkowa kurtoza dla spożycia alkoholu w dni pracujące dla kobiet jest jedyna większa od trzech, co wskazuje na rozkład ciężko ogonowy. W pozostałych przypadkach mamy do czynienia z rozkładem lekko ogonowym. Histogramy liczebności oraz gęstości empiryczne są jednomodalne, prawostronnie skośne. Zarówno dla kobiet jak i mężczyzn modą histogramu oraz modą gęstości jest jedynka, co oznacza, że badani uczniowie w większości nie spożywają alkoholu. Widać to również na wykresach dystrybuant empirycznych, gdzie w każdym przypadku skok jest największy przy jedynce. Jednak dla kobiet największe wartości gęstości są dużo większe niż dla mężczyzn. Wynika to z tego, że mężczyźni konsumują większe ilości alkoholu niż kobiety, co bardzo dobrze widzimy dla gęstości i histogramu spożycia alkoholu przez weekend, gdzie wykresy odpowiedzi mężczyzn są o wiele bardziej wypłaszczone.

Na wykresach pudełkowych współrzędna y dolnej podstawy ramki (pudełka) jest równa pierwszemu kwartylowi, a współrzędna y górnej podstawy ramki jest równa trzeciemu kwartylowi. Zatem długość boku odpowiada

rozstępowi międzykwartylowemu. Poziomy pomarańczowy odcinek wyznacza medianę cechy w próbie. Wszystkie wartości obserwowane na wykresach pudełkowych zgadzają się z wartościami wyznaczonymi przez obliczenia w sestatystyk podstawowych statystyk. Na wykresie pudełkowym spożycia alkoholu w dni pracujące dla mężczyzn (13) wyraźnie widać prawostronną (dodatnią) skośność rozkładu, ponieważ górny wąs jest dłuższy od dolnego wąsa, oraz mediana jest bliżej pierwszego kwartyla. Zarówno dla dni pracujących jak i dla weekendu rozstęp międzykwartylowy u płci męskiej jest znacznie większy niż u płci żeńskiej. To również potwierdza, że mężczyźni spożywają większe ilości trunków od kobiet.

5. Podsumowanie

Po powyższej analizie danych zdecydowanie można uznać, iż uczniowie płci męskiej spożywają więcej alkoholu niż ucznnice płci żeńskiej. Większa zmienności w konsumpcji alkoholu występuje u mężczyzn. Niezależnie od płci spożycie alkoholu wzrasta na weekend. Te wnioski potwierdzają wszystkie statystyki.