Sprawozdanie 1

Aleksandra Szczur (276047), Agnieszka Staszkiewicz (268791)

Spis treści

Wstęp

Cel pracy

Na każdym etapie edukacji wyniki egzaminów stanowią istotną miarę zdobytej wiedzy. Aby osiągnąć lepsze rezultaty, uczniowie poświęcają znaczną ilość czasu na naukę. Czy jednak jedynie czas spędzony na nauce decyduje o wynikach? W niniejszym sprawozdaniu przeanalizujemy różne czynniki i zbadamy ich wpływ na wyniki egzaminów końcowych.

Opis danych

Zbiór danych "Student Perfomance Factors", dostępny na platformie Kaggle , przedstawia analizę wyników egzaminów końcowych uczniów oraz czynników wpływających na ich osiągnięcia.

Dane obejmują 6607 rekordów, z których każdy odnosi się do indywidualnego ucznia. Każdy rekord składa się z 20 kolumn, reprezentujących różnorodne czynniki potencjalnie oddziałujące na ostateczny wynik ucznia:

- **Hours_Studied** (Godziny nauki) - ilość godzin w tygodniu spędzonych na nauce
- **Attendace** (Frekwencja) - frekwencja ucznia wyrażona w procentach
- **Parental_Involvement** (Zaangażowanie rodziców) - poziom zaangażowania rodziców w edukację dziecka (*Low* - niski poziom , *Medium* - średni poziom , *High* - wysoki poziom)
- **Access_to_Resourses** (Dostęp do dodatkowych materiałów) - dostęp do zasobów edukacyjnych (*Low* - niski dostęp do zasobów edukacyjnych, *Medium* - średni dostęp do zasobów edukacyjnych, *High* - wysoki dostęp do zasobów edukacyjcych)
- **Extracurricular_Activities** (Zajęcia pozalekcyjne) - uczestnictwo w zajęciach pozalekcyjnych (*Yes* - uczeń uczestniczący w zajęciach pozalekcyjnych, *No* - uczeń nieuczestniczący w zajęciach pozalekcyjnych)
- **Sleep_Hours** (Godziny snu) - średnia ilość godzin spędzona na spaniu w ciągu jednej nocy
- **Previous_Scores** (Poprzedni wynik) - wynik z poprzednich egzaminów
- **Motivation_Level** (Motywacja) - poziom motywacji ucznia (*Low* - niska motywacja, *Medium* - średnia motywacja, *High* -wysoka motywacja)
- **Internet_Access** (Dostęp do internetu) - dostęp do internetu (*Yes* - uczeń z dostępem do internetu, *No* - uczeń bez dostępu do internetu)
- **Tutoring_Sessions** (Godziny korepetycji) - ilość godzin korepetycji ucznia w ciągu miesiąca
- **Family_Income** (Dochód rodziny) - dochód rodziny ucznia (*Low* - niski dochód, *Medium* - średni dochód, *High* - wysoki dochód)
- **Teacher_Quality** (Kompetencje nauczyciela) - kompetencje nauczyciela (*Low* - niskie kompetencje, *Medium* - średnie kompetencje, *High*- wysokie kompetencje)
- **School_Type** (Typ szkoły) - szkoła, do której uczęszcza uczeń (*Public* - szkoła publiczna, *Private* - szkoła prywatna)
- **Peer_Influence** (Wpływ otoczenia) - wpływ otoczenia na ucznia (*Positive* - wpływ pozytywny, *Neutral* - wpływ neutralny, *Negative* - wpływ negatywny)
- **Psychical_Activity** (Godziny aktywności fizycznej) - średnia ilość godzin w tygodniu poświęcona aktywności fizycznej
- **Learning_Disabilities** (Dysfunckja)- dysfunkcja do uczenia się (*Yes* - uczeń z dysfunkcją, *No* - uczeń bez dysfunkcji)

- **Parental_Education_Level** (Wykształcenie rodziców)- najwyższy poziom wykształcenia rodziców (*High School* - średnie, *College* - wyższe, *Postgraduate* - studia podyplomowe)
- **Distance_From_Home** (Odległość od szkoły) - odległość szkoły od domu (*Near* - mała odległość, *Moderate* - umierkowana odległość, *Far* - duża odległość)
- **Gender** (Płeć)- płeć ucznia (*Male* - mężczyzna, *Female*- kobieta)
- **Exam_Score** (Wynik egzaminu) - wynik egzaminu końcowego

Wczytanie danych

Wczytujemy dane. W celu zlikwidowania potencjalnych nieprawidłowości wynikających z braków danych, zdecydowałyśmy się na usunięcie wierszy zawierających puste komórki. Po usunięciu wierszy, w których znajdują się braki danych w naszym zbiorze pozostaje 6378 wierszy. Dodatkowo, zmieniłyśmy nazwy zmiennych z angielskich na polskie, podane w nawiasach w spisie powyżej.

Poniżej przedstawiamy tabele opisującą zmienne z bazy danych, ich typy oraz możliwe wartości.

Tabela 1: Tabela opisująca zmienne.

Zmienne	Typ danych	Możliwe wartości
Godziny nauki	numeric	[1,44]
Frekwencja	numeric	[60,100]
Zaangażowanie rodziców	character	High, Low, Medium
Dostęp do dodatkowych materiałów	character	High, Low, Medium
Zajęcia pozalekcyjne	character	No, Yes
Godziny snu	numeric	[4,10]
Poprzedni wynik	numeric	[50,100]
Motywacja	character	High, Low, Medium
Dostęp do internetu	character	No, Yes
Godziny korepetycji	numeric	[0,8]
Dochód rodziny	character	High, Low, Medium
Kompetencje nauczyciela	character	High, Low, Medium
Typ szkoły	character	Private, Public
Wpływ otoczenia	character	Negative, Neutral, Positive
Godziny aktywności fizycznej	numeric	[0,6]
Dysfunkcja	character	No, Yes
Wykształcenie rodziców	character	College, High School, Postgraduate
Odległość od szkoły	character	Far, Moderate, Near
Płeć	character	Female, Male
Wynik egzaminu	numeric	[55,101]

W naszej analizie będziemy korzystać tylko z kolumn: *Płeć, Godziny nauki, Frekwencja, Godziny korepetycji, Dostęp do internetu, Dostęp do dodatkowych materiałów, Wynik egzaminu.*

Statystyka opisowa

Statystyka opisowa jest fundamentalnym zagadnieniem w statystyce, mającym na celu opisanie i zrozumienie danych za pomocą różnorodnych technik i wskaźników statystycznych. Jest to podstawowy proces, który umożliwia poznanie charakterystyki zbioru danych oraz identyfikację podstawowych cech, które są istotne dla dalszej analizy.

Tabela statystyki opisowej

W poniższej tabeli przedstawione zostały statystyki opisowe zmiennych numerycznych z naszego zbioru danych:

Tabela 2: Tabela przedstawiająca statystki opisowe zmiennych .

Zmienna	Min	Max	Średnia	Mediana	Q1 - kwartyl 1	Q3 - kwartyl 3	Wariancja	Skośność	Kurtoza	Współczynnik zmienności
Godziny nauki	1	44	19.977	20	16	24	35.826	0.016	3.006	0.300
Frekwencja	60	100	80.021	80	70	90	133.419	0.006	1.804	0.144
Godziny snu	4	10	7.035	7	6	8	2.155	-0.027	2.496	0.209
Poprzedni wynik	50	100	75.066	75	63	88	207.371	-0.001	1.808	0.192
Godziny korepetycji	0	8	1.495	1	1	2	1.523	0.825	3.662	0.825
Godziny aktywności fizycznej	0	6	2.973	3	2	4	1.059	-0.037	2.946	0.346
Wynik egzaminu	55	101	67.252	67	65	69	15.321	1.665	13.647	0.058

Z tabeli możemy wyczytać:

Min - minimalna wartość zmiennej.

Max - maksymalna wartość zmiennej.

Średnia - średnia arytmetyczna wszystkich wartości zmiennej. Średnia arytmetyczna $\bar{x}$ jest miarą centralnej tendencji, która oblicza się jako suma wszystkich wartości zmiennej x_i, dla badanej zbiorowości, podzielona przez liczbę jednostek tej zbiorowości (n).

Mediana - służy do określania wartości centralnej w uporządkowanym zbiorze danych.

Kwartyl 1- punkt, który dzieli najmniejszą część zbioru danych na 25% najniższych wartości. Kwartyl pierwszy jest medianą dolnej połowy próby. Jest to wartość, poniżej której znajduje sie 25% obserwacji.

Kwartyl 3 - punkt, który dzieli największą część zbioru danych na 25% najwyższych wartości. Kwartyl trzeci jest medianą górnej połowy próby. Jest to wartość powyżej której znajduje się 75% obserwacji.

Wariancja - miara rozproszenia wartości w zbiorze danych, która określa, jak bardzo te wartości różnią się od średniej arytmetycznej. Im większa wariancja, tym bardziej zróżnicowane są wartości.

Skośność - miara asymetrii obserwacji, informująca o tym jak wyniki dla danej zmiennej kształtują się wokół średniej ($\alpha > \mathbf{0}$ - asymetria dodatnia i rozkład jest prowostronnie skośny, $\alpha < \mathbf{0}$ - asymetria ujemna i rozkład jest lewostronnie skośny, $\alpha = \mathbf{0}$ - rozkład symetryczny)

Kurtoza - miara spłaszczenia, która informuje o koncentracji rozkładu ($\mathbf{K} > \mathbf{3}$ - rozkład ciężkoogonowy, $\mathbf{K} < \mathbf{3}$ - rozkład lekkoogonowy, $\mathbf{K} = \mathbf{3}$ - kurtoza rozkładu normalnego).

Współczynnik zmienności - miara rozproszenia danych, która pozwala porównywać zmienność między różnymi zbiorami danych o różnych średnich wartościach. Określa stosunek odchylenia standardowego do średniej arytmetycznej danych ($V \in [0, 0.2)$ - zmienność mała, $V \in [0.2, 0.4)$ - zmienność przeciętna, $V \in [0.4, 1)$ - zmienność duża, $V \in [1, 1.5)$ - zmienność bardzo duża, $V \in [1.5, \infty)$ - zmienność skrajnie duża).

Na podstawie tabeli możemy wyciągnąć wnioski dla zmiennych używanych przez nas w sprawozdaniu

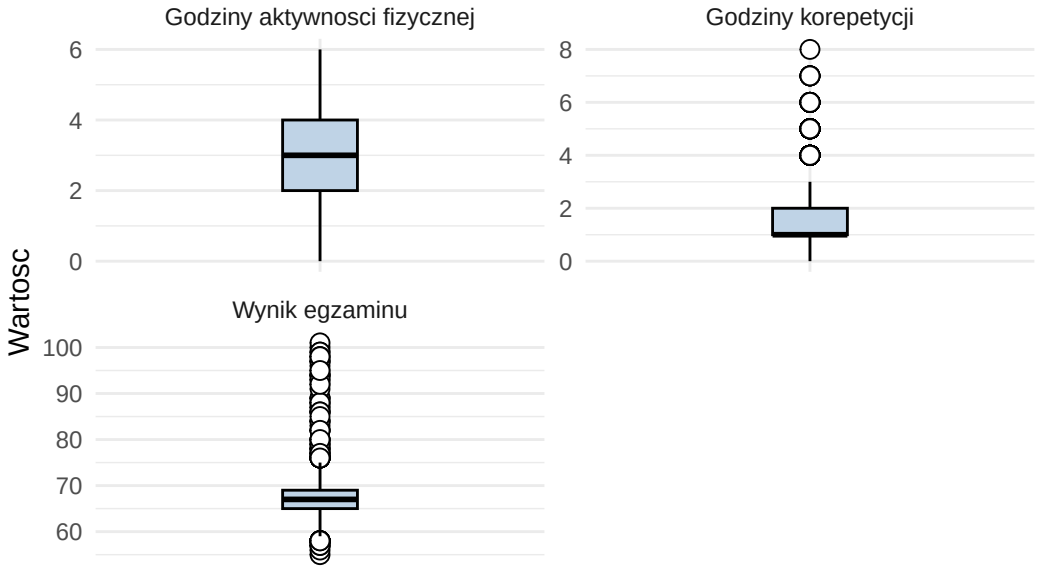
- Średnia arytmetyczna i mediana liczby godzin nauki są do siebie zbliżone, co oznacza, że rozkład zmiennej jest w miarę symetryczny. Wysoka wartość wariancji świadczy o dużej rozpiętości danych, a współczynnik zmienności 0.3 oznacza zmienność przeciętną w liczbie godzin nauki. Dodatkowo, skośność i kurtoza wskazują, że rozkład jest zbliżony do normalnego.
- Podobnie, zbliżone wartości średniej i mediany frekwencji świadczą o równomiernym rozkładzie wokół średniej. Mimo wysokiej wariancji, która sugeruje dużą zmienność danych, współczynnik zmienności wskazuje na niewielkie odchylenie w proporcji do średniej. Skośność i kurtoza natomiast pokazują, że rozkład jest bardziej spłaszczony niż normalny.
- W przypadku godzin korepetycji zarówno średnia jak i mediana sugerują, że wartości są silniej skupione wokół niższych przedziałów. Ujemna skośność wskazuje na asymetrię ujemną, a kurtoza większa od 3 na rozkład ciężkoogonowy. Dane charekteryzują się przy tym duża zmiennością.
- Wyniki egzaminów wykazują symetryczny rozkład, co wynika z podobieństwa wartości średniej i mediany. Wariancja wskazuje na stosunkowo jednorodne wyniki, a niski współczynnik zmienności potwierdza niewielkie różnice między nimi. Dodatkowo, skośność sugeruje, że rozkład jest lekko przesunięty w stronę wyższych wyników, a bardzo wysoka kurtoza na wyraźne skupienie wartości blisko mediany.

Wykresy pudełkowe

Wykres pudełkowy pozwala szybko zobaczyć rozkład danych, wykryć wartości odstające i porównać rozkłady między grupami. Składa się z pudełka oraz dwóch, tak zwanych, "wąsów". Z wykresu pudełkowego możemy odczytać:

- medianę - linia wewnątrz pudełka,
- kwartyl pierwszy ($Q1$) - dolna granica pudełka,
- kwartyl trzeci ($Q3$) - górna granica pudełka,
- rozstęp międzykwartylowy (IQR) - wysokość pudełka (różnica między $Q3$ a $Q1$),
- zakres wartości - wąsy wyChodzące z pudełka,
- wartości odstające - punkty nad i pod pudełkiem. Wartości odstające to wartości większe od $Q3 + 1.5IQR$ lub mniejsze od $Q1 - 1.5IQR$

Poniżej przedstawiono wykresy pudełkowe dla zmiennych ilościowych z naszych danych:



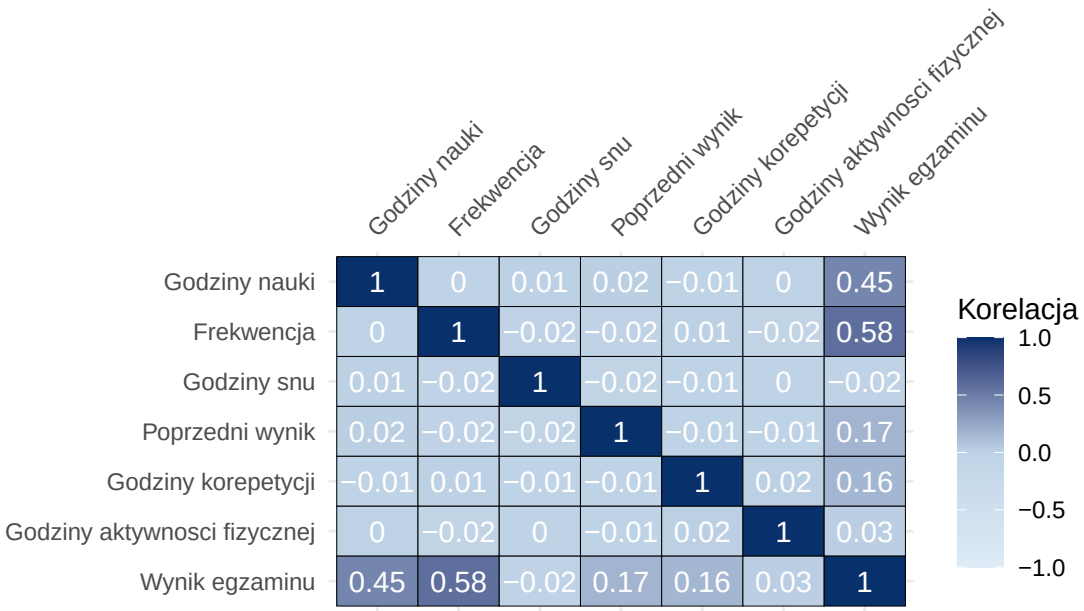
Rysunek 1: Wykresy pudełkowe dla zmiennych ilościowych.

Wykresy pudełkowe w czytelny sposób ilustrują wartości obliczone w tabeli statystyk opisowych, potwierdzając ich poprawność. Dodatkowo umożliwiają łatwe zidentyfikowanie danych, w których występują wartości odstające.

Macierz korelacji

Korelacja Pearsona to miara związku pomiędzy dwiema zmiennymi. Korelacja to inaczej unormowana kowariancja. Wartość kowariancji to liczba z przedziału $[-1, 1]$ ($\rho = -1$ - silna korelacja ujemna, $\rho = 1$ - silna korelacja dodatnia, $\rho = 0$ - brak korelacji).

Poniżej przedstawiona jest macierz korelacji zmiennych ilościowych z naszych danych. Patrząc na wyliczone wartości będziemy w stanie odczytać korelację między każdym czynnikiem ilościowym wpływającym na wynik egzaminu końcowego.



Rysunek 2: Macierz korelacji dla zmiennych ilościowych.

Z wykresu wynika, że największe wartości korelacji mają zależności wyniku końcowego egzaminu do godzin poświeconym nauce oraz do obecności na zajęciach i wynoszą kolejno: 0.45, 0.58. Korelacje te są dodatnie, co oznacza, że im wiecej się uczymy lub uczęszczamy na zajęcia - tym większy wynik egzaminu możemy osiągnąć. Co zaskakujące, korelacja w przypadku zależności *Wynik egzaminu* do *Frekwencja* jest wyższa od korelacji *Wynik egzaminu* do *Godziny nauki*. Oznacza to, że większy wpływ na lepszy wynik egzaminu ma obecność na zajęciach, a nie ilość poświęconych na naukę godzin.

Pytania badawcze

W naszym sprawozdaniu skupimy się na trzech pytaniach badawczych:

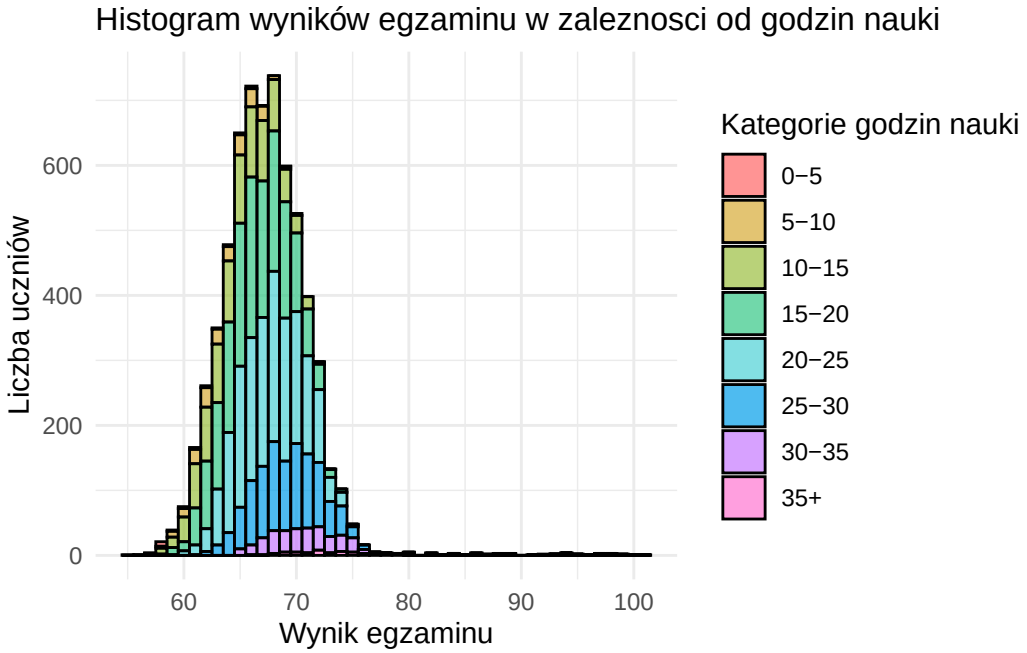
1. Jak ilość godzin spędzonych na nauce wpływa na wynik egzaminu końcowego?
2. Jak obecność na zajęciach wpływa na wynik egzaminu końcowego?
3. Jak dodatkowe wsparcie edukacyjne wpływa na wynik egzaminu?

Analiza danych

Jak ilość godzin spędzonych na nauce wpływa na wynik egzaminu końcowego?

Z analizy macierzy koreleacji wiemy już, że spośród wszystkich zmiennych ilościowych, liczba godzin poświęconych na naukę ma jedno z największych oddziaływań na wynik egzaminu końcowego. Przyjrzyjmy się temu, jak znaczące są te różnice.

W pierwszej kolejności przyjrzyjmy się liczbie uczniów osiągających dany wynik na egzaminie w zależności od liczby godzin poświęconych na naukę. Godziny nauki zostały podzielone na osiem kategorii: $[0, 5), [5, 10), [10, 15), [15, 20), [20, 25), [25, 30), [30, 35)$ oraz $[35, \infty)$.



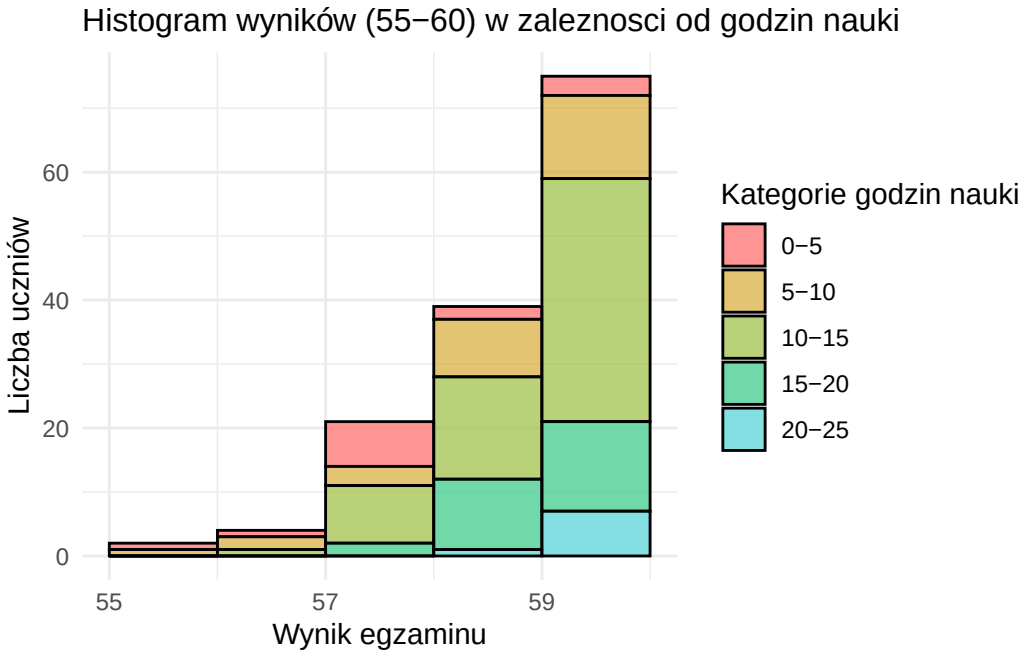
Rysunek 3: Histogram wyników egzaminu w zależności od godzin nauki.

Największe skupienie w przedziale $[60, 70]$ świadczy o tym, że większość uczniów uzyskuje właśnie takie wyniki. Zgadza się to z wyliczoną w tabeli statystyki opisowej średnią zmiennej *Wynik egzaminu* wynoszącą 67.252 i medianą wynoszącą 67. Rozkład jest prawostronnie skośny i wyraźnie skupiony wokół mediany, co potwierdza skośność i kurtozę.

Słupki podzielone są według kategorii godzin spędzonych na nauce. Widać, że te kategorie faktycznie przyczyniają się do rozkładu wyniku z egzaminu. Uczniowie z niższych kategorii osiągają głównie niższe wyniki, co widać po większym ich skupieniu po lewej stronie. Uczniowie, którzy natomiast spędzili więcej czasu na nauce uzyskują wyniki wyższe, co potwierdza pojawienie się kategorii $[30, \infty)$ dopiero po prawej stronie rozkładu.

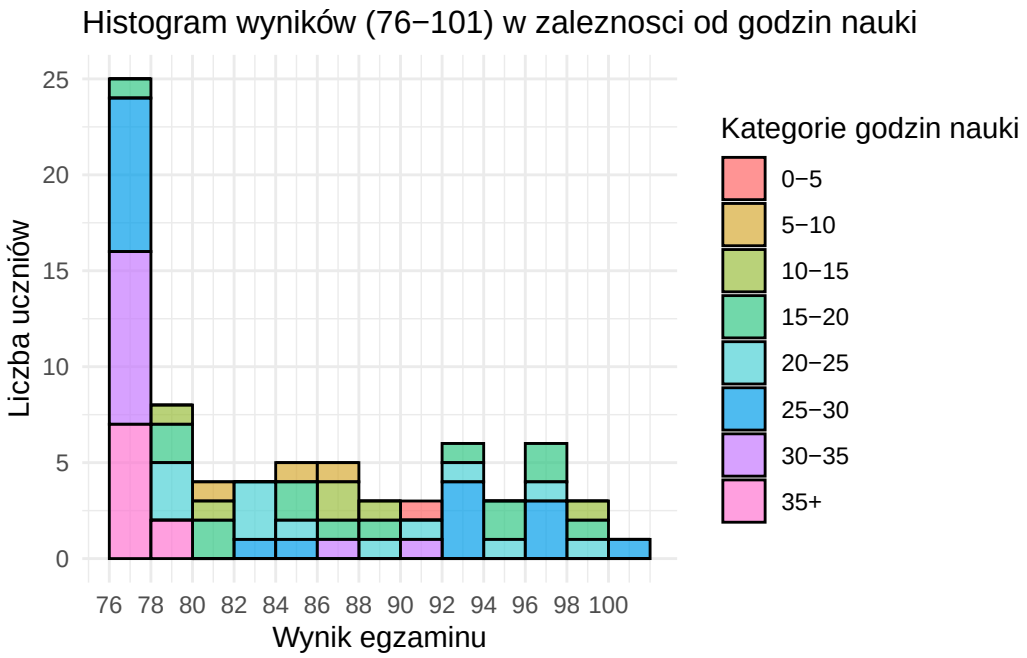
Możemy stwierdzić, że kategorie $[15, 25)$ najlepiej reprezentują uczniów uzyskujących wyniki zbliżone do średniej.

Dla lepszego zauważenia tych tendencji skupmy się na histogramach wartości odstających. Wartości odstające przyjęłyśmy na podstawie wykresu pudełkowego zmiennej *Wynik egzaminu.*



Rysunek 4: Histogram wyników (55-60) w zależności od godzin nauki.

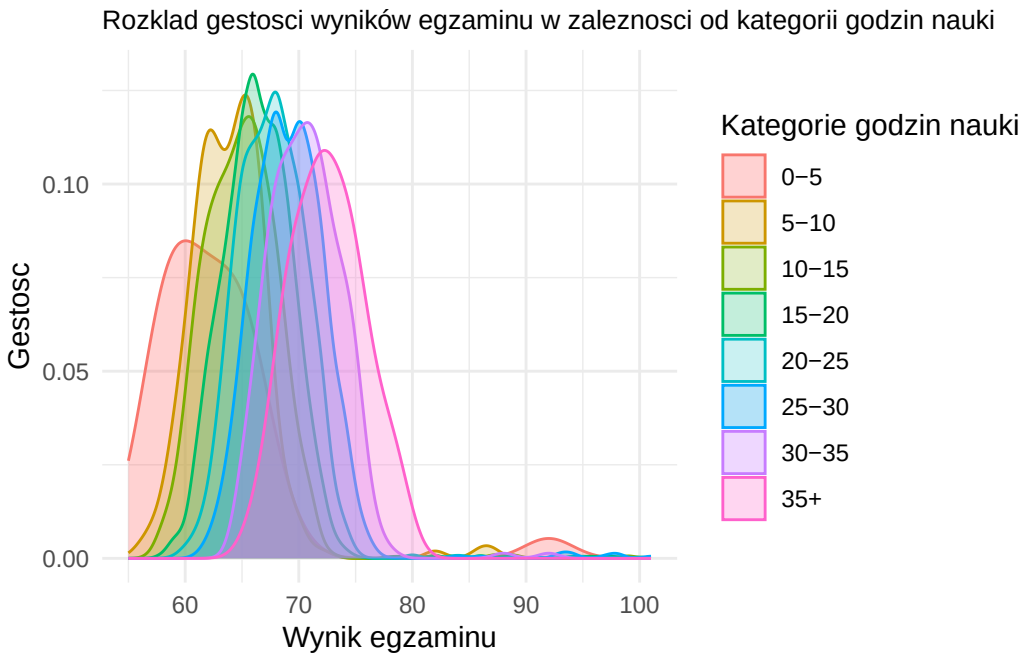
Z wykresu wynika, że kategoria godzin nauki która dominuje w przypadku niższych wyników egzaminu to $[10, 15)$. Widzimy też, że w miarę zwiększania wyniku, zwiększa się również ilość uczniów przeznaczających więcej czasu na naukę. Uczniów z kategorii $[20, 25)$ jest zauważalnie mniej. Nie pojawiają się również uczniowie spędzający więcej niż 25 godzin na nauce.



Rysunek 5: Histogram wyników (76-101) w zależności od godzin nauki.

Z wykresu ilości uczniów uzyskujących wyższe wyniki zauważamy znacznie większy stosunek uczniów uczących się więcej niż 15 godzin do reszty. Można stwierdzić, że kategorie [15.30) są dominujące. W gronie uczniów uzyskujących wysokie wyniki, przeciwnie do tych o niskich wynikach, znajdują sie osoby uczące się więcej niż 25 godzin, a liczba osób uczących się mniej niż 10 godzin jest praktycznie niezauważalna.

Aby lepiej zrozumieć, jak różne kategorie godzin poświęconych nauce wpływają na wyniki egzaminu, skupny się na rozkładach gęstości wyników. Wykresy gęstości pozwolą nam lepiej uchwycić kształt i różnice w zależności od godzin poświęconych na naukę do egzaminu.



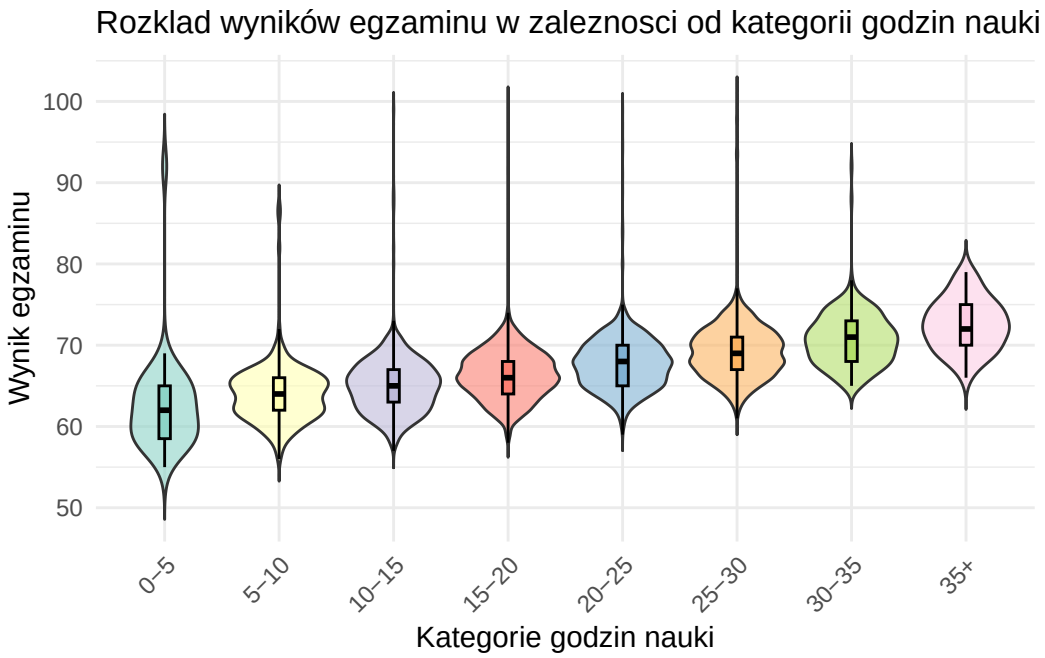
Rysunek 6: Rozkład gęstości wyników egzaminu w zależności od kategorii godzin nauki.

Widzimy, że osoby uczące się mniej uzyskują niższe wyniki egzaminu. Wraz ze wzrostem czasu nauki, szczyt krzywych przesuwa się w stronę wyższych wyników egzaminu.

Kategorie z większą liczbą godzin nauki mają tendencję do wyższych wyników, ale wyniki te mogą być bardziej rozproszone.

W kategoriach z bardzo wysoką liczbą godzin nauki, gęstość jest niska, ale rozciąga się na wyższe wyniki, co świadczy o rzadszych, ale lepszych wynikach.

Aby uzupełnić analizę i zachować informację o rozkładzie danych, zastosujemy wykresy skrzypcowe. Ten typ wizualizacji łączy w sobie zalety wykresów gęstościowych i pudełkowych.



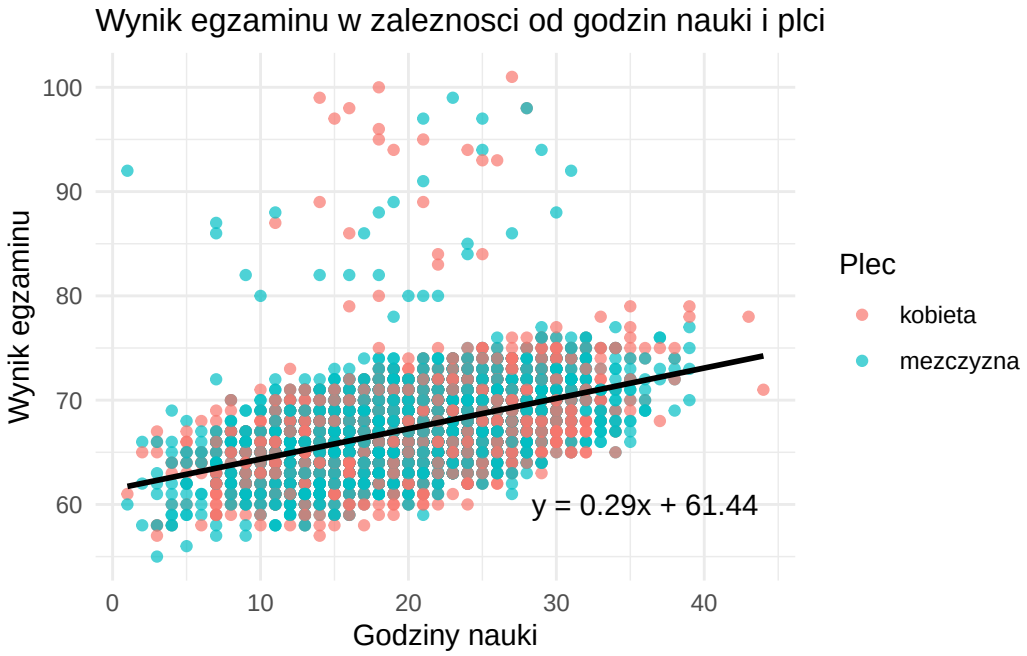
Rysunek 7: Rozkład wyników egzaminu w zależności od kategorii godzin nauki.

W najniższej kategorii czasu poświęconego na naukę wyniki są niskie, gęstość skoncentrowana w niższym przedziale, a mediana najniższa.

Wraz ze wzrostem czasu nauki wyniki rosną. Mediana przesuwa się w górę, ale wyniki pozostają rozproszone.

Przy najwyższej kategorii czasu poświęconego na naukę wyniki wciąż są wysokie, ale dłuższe skrzypce wskazuje na znaczącą różnorodność wyników. Możliwe, że efektywność nauki spada przy tak długim czasie.

Wykresy skrzypcowe wyraźnie sugerują istnienie zależności - osoby poświęcające więcej czasu na naukę osiągają lepsze wyniki egzaminu. Aby dokładnie ocenić siłę tej relacji i pogłębić naszą analizę zastosujemy model regresji liniowej.



Rysunek 8: Wynik egzaminu w zależności od godzin nauki i płci.

Z równania regresji wynika, że z każdą dodatkowo przepracowaną godziną nauki, średni wynik wzrasta o około 0.29 punktu. Wartość 61.44 to punkt przecięcia linii regresji z osią Y, czyli przewidywany wynik egzaminu przy zerowych godzinach nauki.

Zauważamy dodatnią zależność między liczbą godzin nauki a wynikiem egzaminu - im więcej czasu poświęcono na naukę, tym wyższy średni wynik. Zależność nie jest jednak bardzo silna (niewielkie nachylenie linii), co oznacza, że poza nauką istnieją inne czynniki wpływające na wynik egzaminu.

Wyniki egzaminu są podobnie rozproszone dla obu płci. Nie widać wyraźnych różnic między mężczyznami a kobietami w poziomie wyników egzaminu przy danej liczbie godzin nauki.

Mimo dodatniej zależności, zauważamy zmienność wyników. Nie wszystkie osoby osiągają wysokie wyniki nawet przy dużej liczbie godzin nauki.

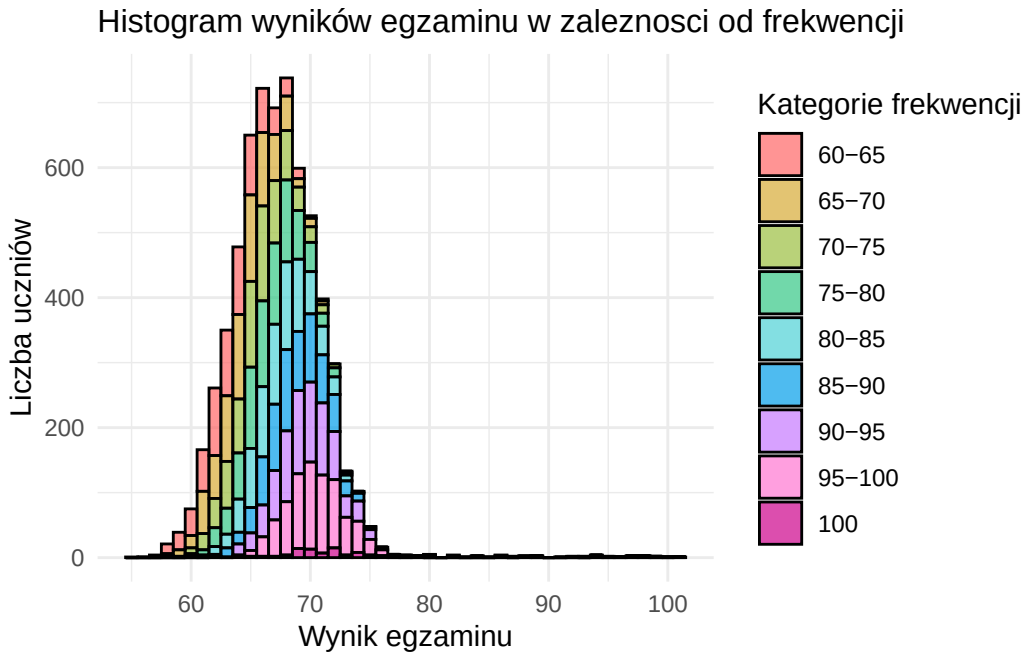
Wyniki egzaminu przy mniejszej liczbie godzin nauki są widocznie bardziej skoncentrowane w dolnej części wykresu.

Analiza wykresu sugeruje, że nauka ma pozytywny wpływ na wynik egzaminu, jednak efektywność nauki jest stosunkowo niska.

Podsumowując wszystkie powyższe wykresy, dłuższy czas nauki pozytywnie wpływa na wyniki egzaminu końcowego. Studenci poświęcający więcej czasu na naukę osiągają lepsze wyniki, niż ci uczący się mniej.

Jak obecność na zajeciach wpływa na wynik egzaminu końcowego?

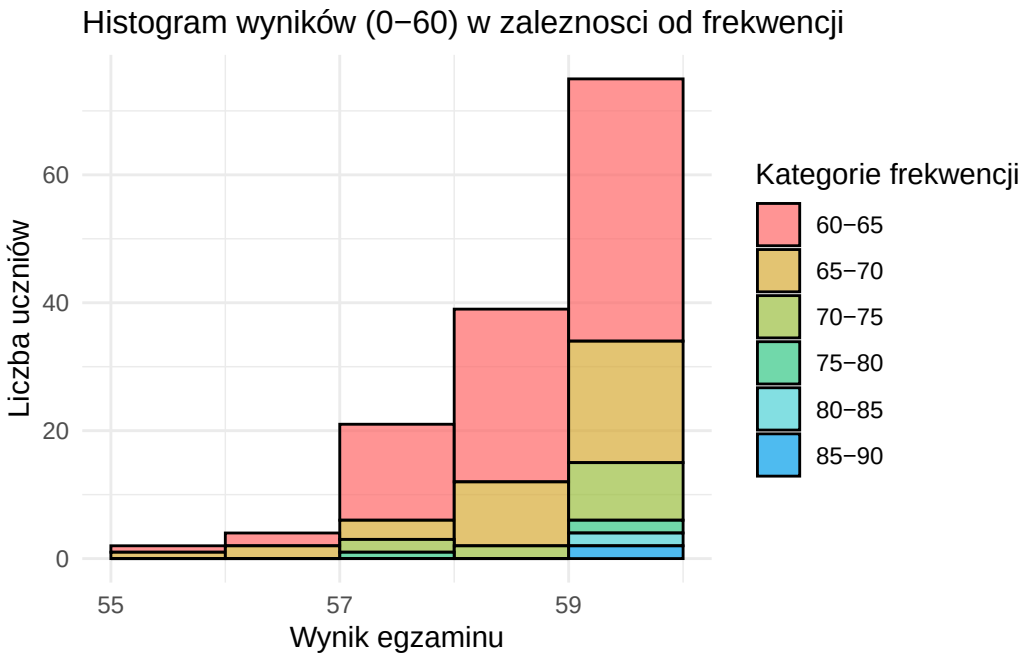
Przyjrzyjmy się następnie zależności o najwyższej wartości obliczonej korelacji Pearsona, czyli relacji między frekwencją, a wynikiem egzaminu.



Rysunek 9: Histogram wyników egzaminu w zależności od frekwencji.

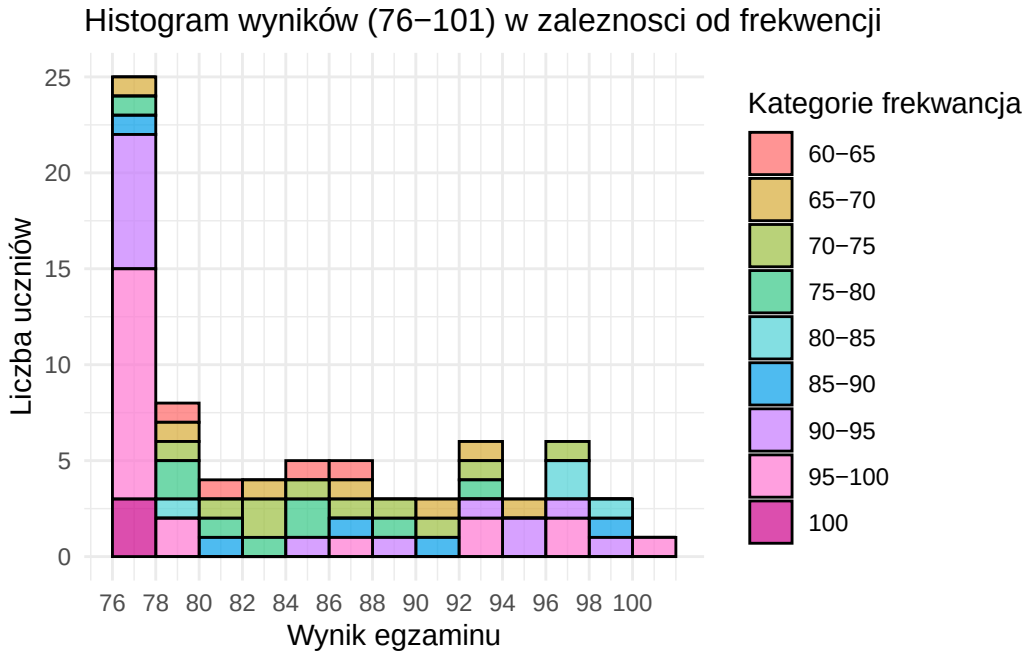
Słupki tego samego histogramu, co w poprzednim pytaniu badawczym, tym razem podzielone zostały według kategorii frekwencji na zajęciach. Anlizując kolory odpowiadające poszczególnym kategoriom, możemy zauważyć, że frekwencja rzeczywiście wpływa na wyniki egzaminu. Uczniowie z niższą frekwencją dominują po lewej stronie rozkładu, co wskazuje na osiąganie przez nich niższych wyników. Z kolei osoby z wysoką frekwencją, w szczególności tej powyżej 90% częściej pojawiają się po prawej stronie rozkładu, co świadczy o uzyskiwaniu wyższych wyników.

Przyjrzyjmy się następnie histogramom wartości odstających.



Rysunek 10: Histogram wyników (0-60) w zależności od frekwencji.

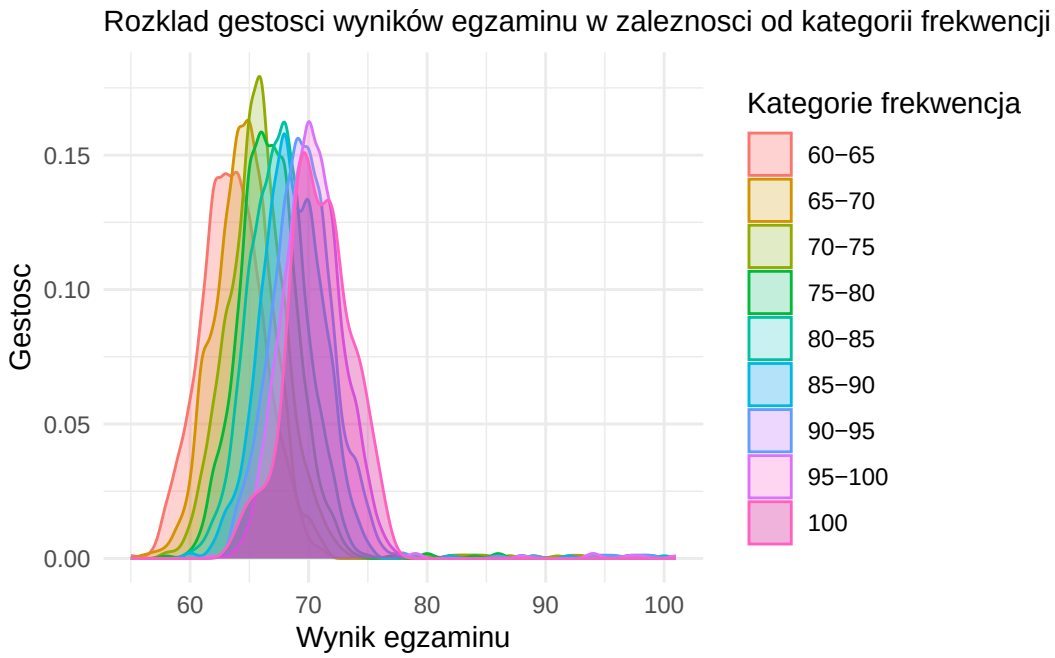
Większość uczniów, którzy uzyskali najniższe wyniki, ma bardzo niską frekwencję (głównie w kategoriach $[60, 70)$. Uczniowie z wyższą frekwencją praktycznie nie pojawiają się w tym przedziale wyników. Można założyć, że niskie wyniki są prawdopodobnie związane z niską obecnością na zajeciach.



Rysunek 11: Histogram wyników (76-101) w zależności od frekwencji.

Wysokie wyniki są osiągane przez uczniów z różnymi poziomami frekwencji, natomiast znacznie przeważają osoby z frekwencją na poziomie 70+, przeciwnie do wartości najniższych.

Przyjrzyjmy się teraz wykresom gęstości wyników egzaminu dla poszczególnych kategorii frekwencji, aby zobaczyć, które grupy osiągają lepsze wyniki i jak różnią się między sobą.



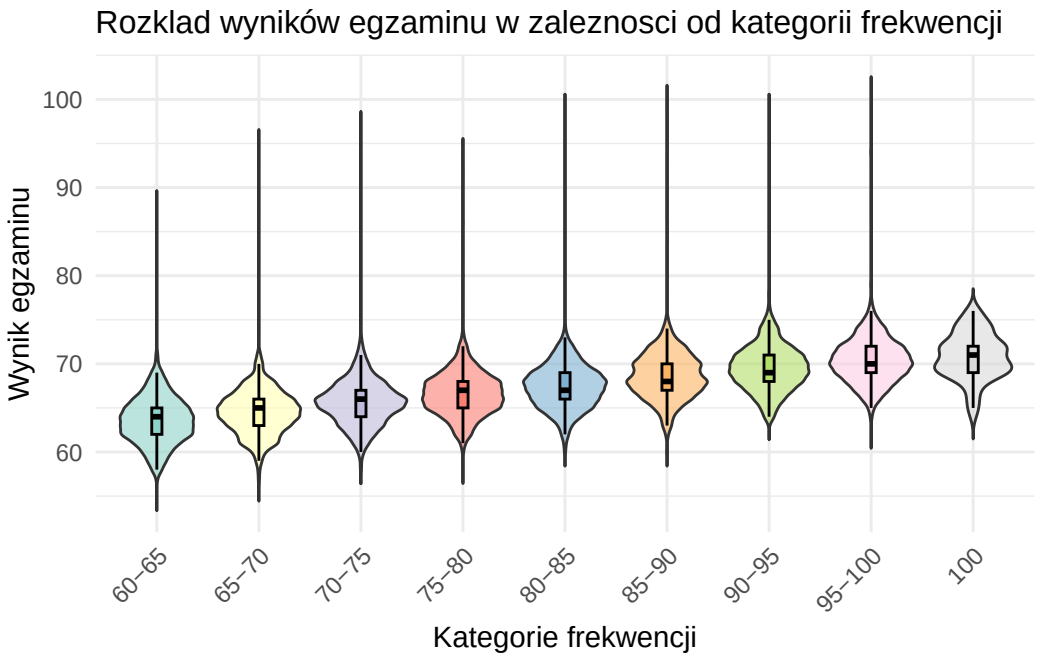
Rysunek 12: Rozkład gęstości wyników egzaminu w zależności od kategorii frekwencji.

Na wykresie widzimy, że grupy o wyższej frekwencji mają wyższe wyniki egzaminów, co widać po przesunięciu ich rozkładów w prawo.

Grupy o niższej frekwencji mają wyniki skoncentrowane bliżej lewej strony, co świadczy o gorzej napisanym egzaminie.

Potwierdza to pozytywną korelację między frekwencją a wynikami egzaminu.

Przeanalizujmy teraz wykresy skrzypcowe zależności wyniku egzaminu od kategorii frekwencji.

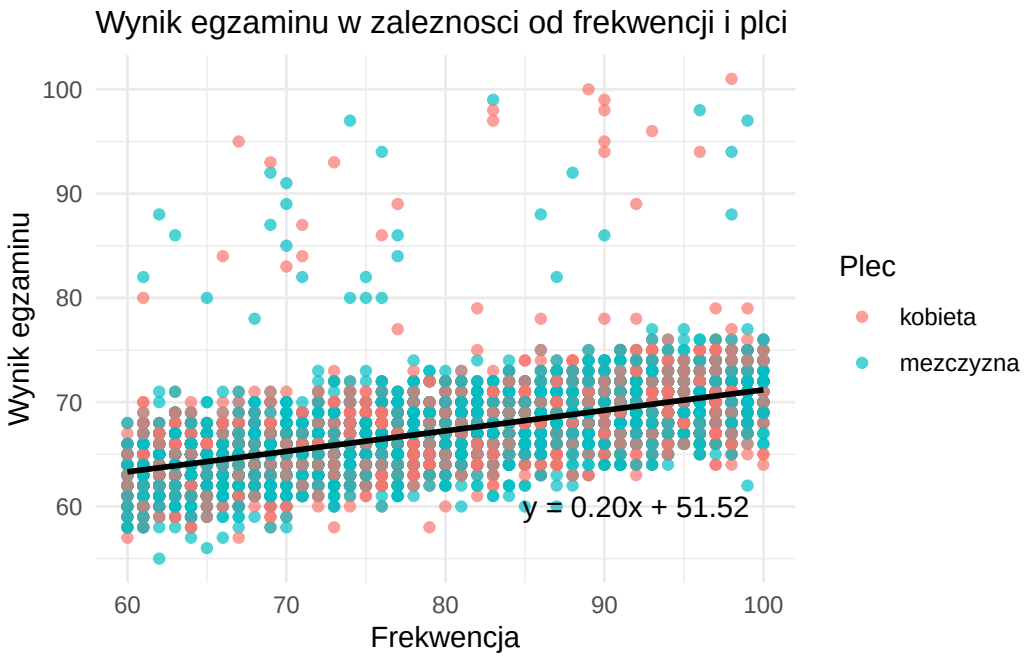


Rysunek 13: Rozkład wyników egzaminu w zależności od kategorii frekwencji.

Widzimy, że im wyższe kategorie frekwencji, tym wyższe mediany wyników.

W każdej kategorii wysokości pudełek pozostają podobne, co świadczy o podobnym rozrzucie wartości.

W grupach o niższej frekwencji pojawiają się zarówno bardzo niskie, jak i sporadycznie wysokie wyniki, co świadczy o mniejszej przewidywalności.



Rysunek 14: Wynik egzaminu w zależności od frekwencji i płci.

Z równania regresji wynika, że każdemu wzrostowi frekwencji o 1% odpowiada średni wzrost wyniku egzaminu o 0.2 punktu. Punkt przecięcia regresji z osią Y równy 51.52 oznacza przewidywany wynik egzaminu, gdy frekwencja wynosi 0%.

Płeć nie ma silnego wpływu na wynik egzaminu w analizie z użyciem frekwencji.

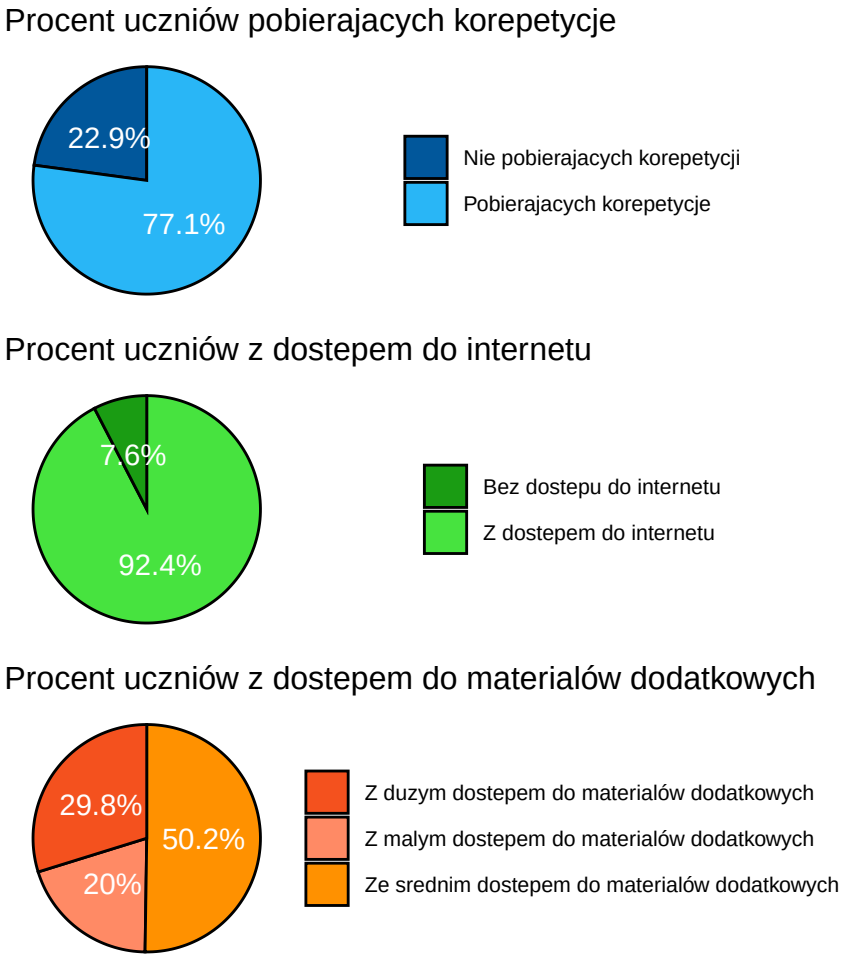
Linia trendu wskazuje dodatnią zależność między frekwencją a wynikiem egzaminu, jest to jednak słaba zależność.

Podsumowując wszystkie powyższe wykresy, frekwencja na zajęciach ma pozytywny wpływ na wynik egzaminu końcowego. Uczniowie spędzający więcej czasu na zajęciach uzyskują lepsze wyniki od tych pojawiających się na nich rzadko.

Jak dodatkowe wsparcie edukacyjne wpływa na wynik egzaminu?

W nauce i przyswajaniu wiedzy, oprócz zajęć w szkole, pomocne jest również dodatkowe wsparcie edukacyjne. W tym pytaniu badawczym sprawdzimy czy uczęszczanie na korepetycje, dostęp do internetu i dodatkowych źródeł wpływa na końcowy wynik egzaminu.

W pierwszej kolejności zbadajmy procentowy udział osób z dodatkowym wsparciem edukacyjnym:



Rysunek 15: Procentowy udział osób z dodatkowym wsparciem edukacyjnym.

W wykresów wynika, że ponad 77% uczniów pobiera minimum jedną godzinę korepetycji w miesiącu, aż 92.4% uczniów posiada dostęp do internetu, a w sumie 80% uczniów ma minimum średni dostęp do materiałów dodatkowych. Te wysokie wartości świadczą o tym, że większość uczniów ma szeroki dostęp do wsparcia edukacyjnego i technologii, co może pozytywnie wpływać na ich proces nauki i wyniki końcowe. Sprawdzimy czy faktycznie tak jest omawiając każdy przypadek osobno.

Korepetycje

Na początku zbadamy ilość osób uczęszczających na korepetycje. Naszym warunkiem zakwalifikowania do tej grupy jest odbycie minimum jednej godziny takich zajęć w ciągu miesiąca.

Rysunek 16: Rozkład wyników egzaminu dla uczniów z korepetycjami.

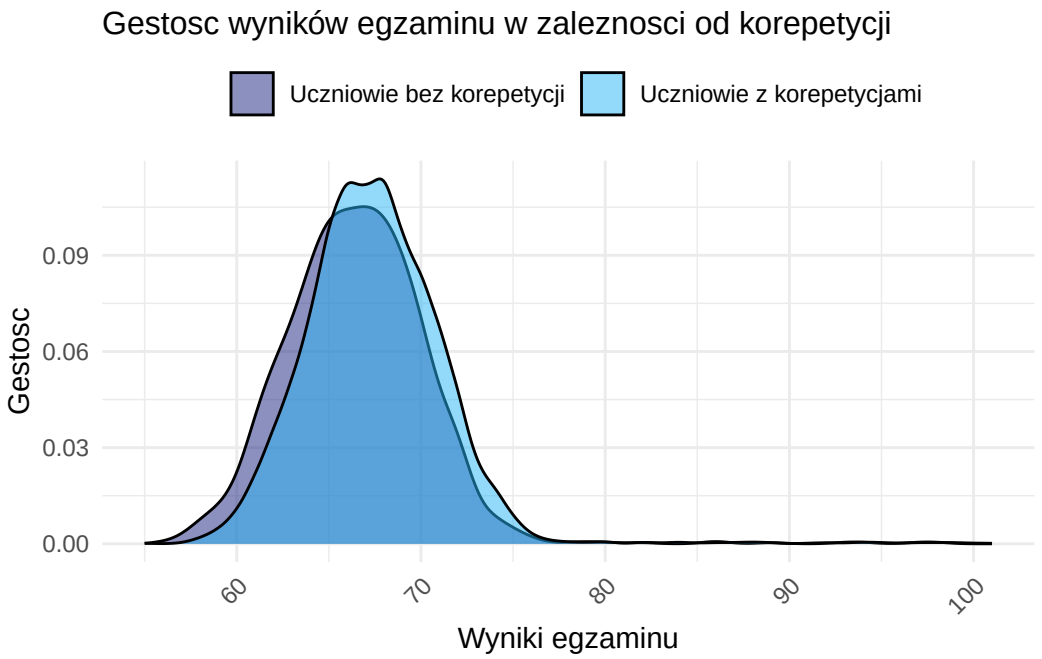
Z wykresu wynika, że suma osób pobierających korepetycje to 4919 i, że większość tych osób uzyskało wynik $(65, 70]$. Zbadajmy teraz wyniki egzaminu osób niepobierających korepetycji.

Rysunek 17: Rozkład wyników egzaminu dla uczniów bez korepetycji.

W wykresu wynika, że suma osób niepobierających korepetycji to 1456 i, że większość tych osób uzyskało wynik $(61, 66]$, czyli mniej od osób uczęszczających na korepetycje. Jednakże,

podobnie dużą grupą są również osoby, które uzyskały wynik z przedziału $(66, 71]$, a są to wartości bardzo podobne do wykresu poprzedniego.

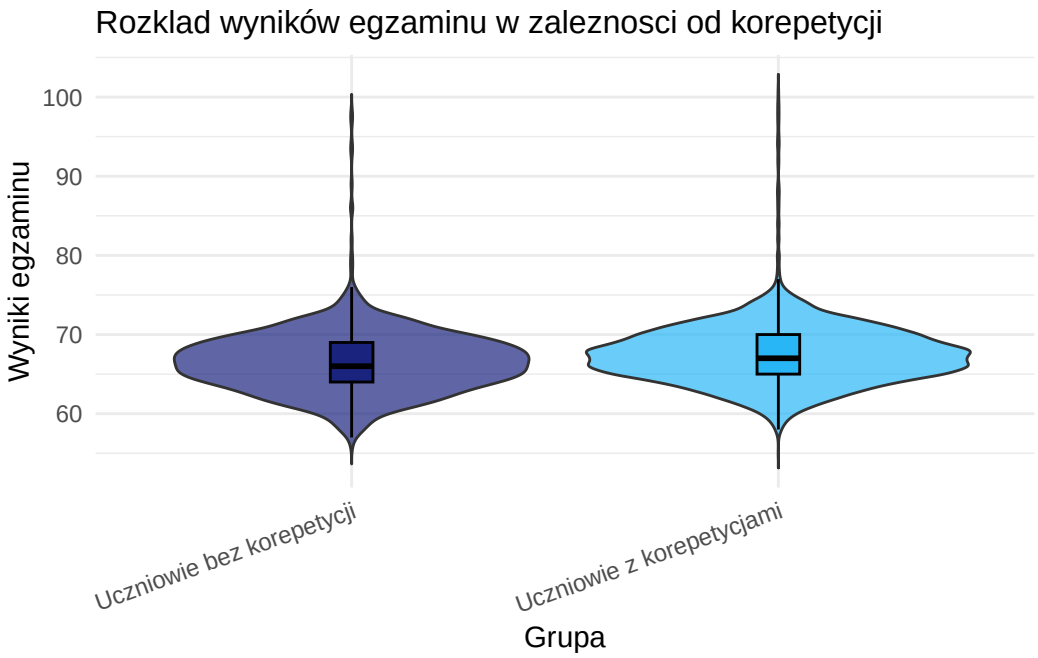
W celu lepszego porównania danych przedstawmy wykres gęstości wyniku z egzaminu dla obu tych grup:



Rysunek 18: Gęstość wyników egzaminu w zależności od korepetycji.

Z wykresu wynika, że wyniki uczniów bez korepetycji są skoncentrowane bardziej po lewej stronie niż wyniki uczniów pobierających korepetycje. Sugeruje to niższe wyniki tych pierwszych i większą szansę na uzyskanie lepszego wyniku gdy pobieramy korepetycje. Mimo tego, że nie są to wartości bardzo się różniące, wykres wskazuje, że korepetycje mają pozytywny wpływ na wyniki egzaminu.

Przeanalizujmy teraz wykres skrzypcowy porównujący rozkład wyników egzaminu tych dwóch grup.



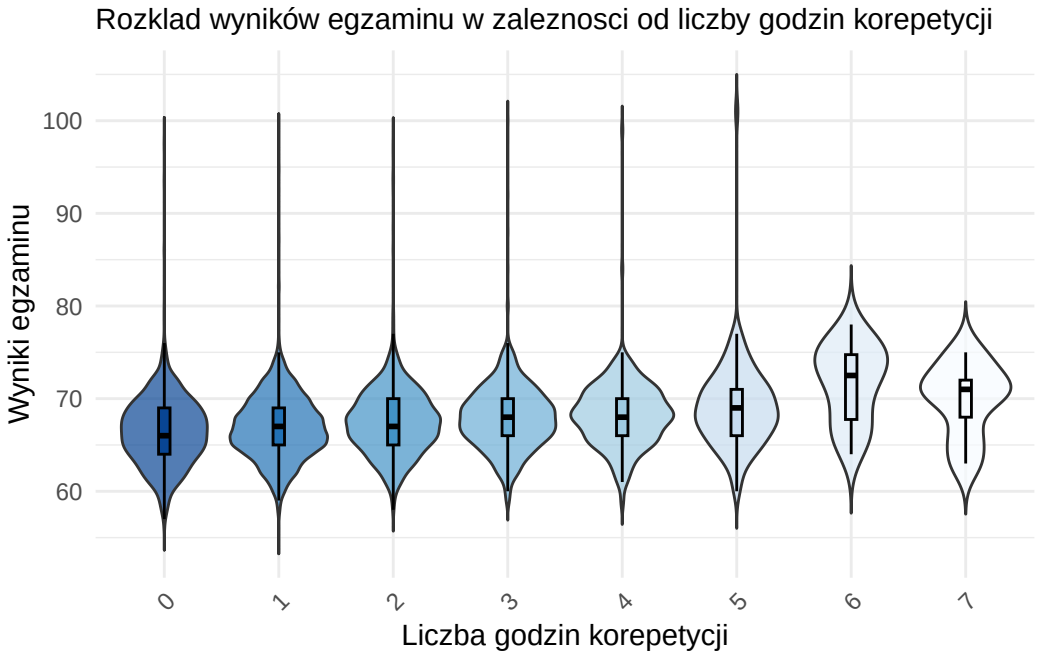
Rysunek 19: Rozkład wyników egzaminu w zależności od korepetycji.

Wykresy w obu przypadkach mają podobne szerokości, co wkazuje na podobne zróżnicowanie wyników zarówno w pierwszej, jak i w drugiej grupie.

Wysokości pudełka, reprezentujące kwartyle (50% wyników) są podobnej wielkości co oznacza, że rozkład wyników dla 50% uczniów w każdej grupie jest podobny.

Kreska w środku pudełek, reprezentuje medianę wyników. W przypadku uczniów pobierających korepetycje jest ona wyższa, co wskazuje na wyższe wyniki w tej grupie.

Sprawdźmy teraz jak wygląda rozkład wyników egzaminu przy konkretnych ilościach godzin korepetycji:



Rysunek 20: Rozkład wyników egzaminu w zależności od liczby godzin korepetycji.

W wykresie pominęliśmy liczbę godzin korepetycji równą 8, ponieważ w zbiorze danych mamy tylko jeden taki przypadek.

Z wykresu wynika, że dla 0 godzin korepetycji wyniki skupiają się w dolnej części wykresu, a mediana jest najniższa, co oznacza najniższe wyniki egzaminu.

Dla większej liczby godzin korepetycji (1, 2, 3, 4) mediana przesuwa się lekko w górę, a wyniki są podobnie zróżnicowane. Oznacza to pozytywny wpływ większej liczby korepetycji na wynik egzaminu.

Bardziej zauważalne różnice w wynikach obserwujemy przy większej liczbie godzin korepetycji. Chociaż dla 5 godzin i więcej rozkłady wyników egzaminów wykazują większe zróżnicowanie, to jednocześnie mediany tych wyników systematycznie rosną, co wskazuje na pozytywny wpływ dodatkowych godzin nauki.

Ogólnie, im więcej godzin przeznaczmy na korepetycje, tym większą szansę mamy na uzyskanie większego wyniku.

Podsumowując wszystkie powyższe wykresy, uczęszczanie na korepetycje ma pozytywny wpływ na wynik egzaminu. Dodatkowo, im więcej godzin przeznaczymy na korepetycje tym większa szansa na wyższy wynik.

Dostęp do internetu

Zbadajmy teraz wpływ dostępu do internetu, czyli, bazując na wykresie kołowym, najbardziej popularnej pomocy naukowej.

Przeanalizujmy ilość osób z dostępem do internetu i ich wyniki z egzaminu:

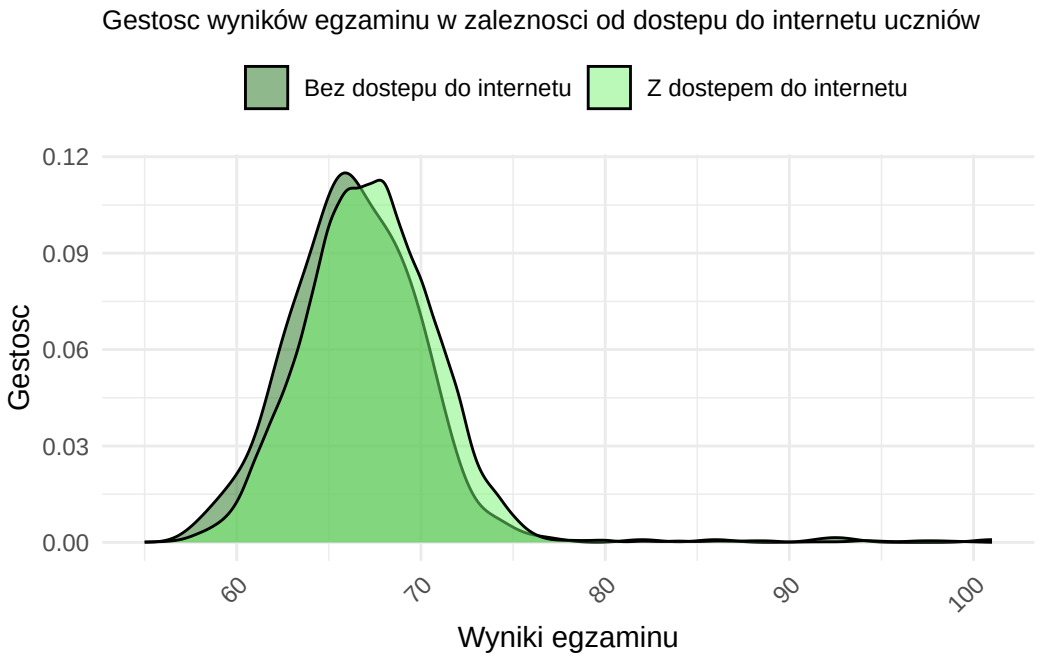
Rysunek 21: Rozkład wyników egzaminu dla uczniów z dostępem do internetu.

Z wykresu wynika, że suma osób z dostępem do internetu to 5893 i, że większość tych osób uzyskało wynik $(65, 70]$. Zbadajmy teraz wyniki egzaminu osób bez dostępu do internetu.

Rysunek 22: Rozkład wyników egzaminu dla uczniów bez dostępu do internetu.

Z wykresu wynika, że suma osób bez dostępu do internetu to 485 i, że większość tych osób uzyskało wynik $(63, 68]$, czyli mniej od osób z dostępem do internetu.

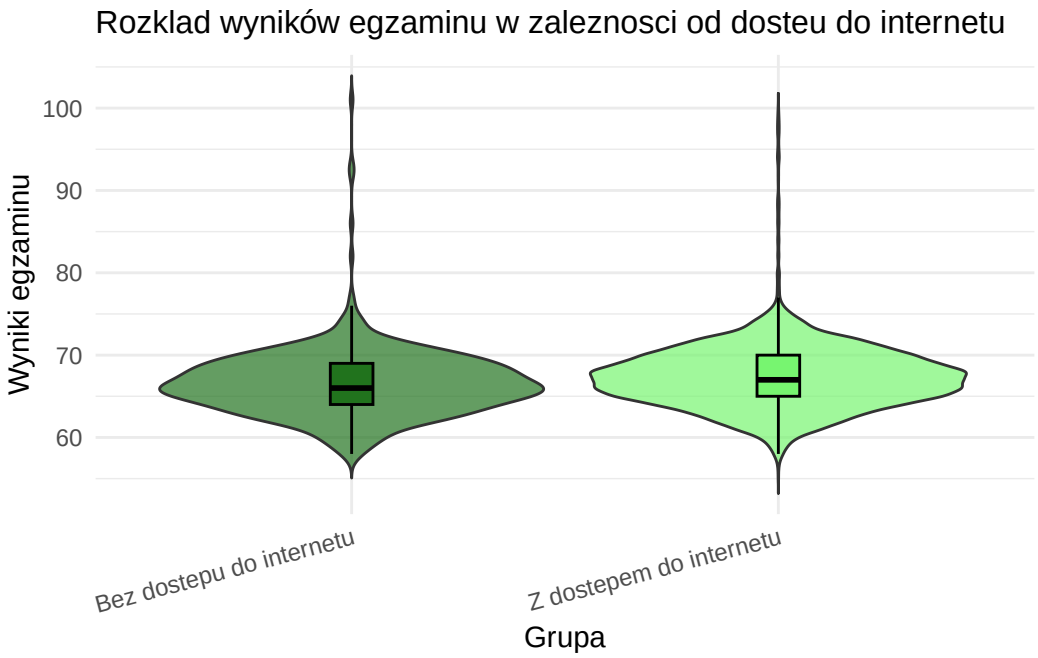
W celu lepszego porównania danych przedstawmy wykres gęstości wyniku z egzaminu dla obu tych grup:



Rysunek 23: Gęstość wyników egzaminu w zależności od dostępu do internetu uczniów.

Z wykresu wynika, że wyniki uczniów bez dostępu do internetu są skoncentrowane bardziej po lewej stronie niż wyniki uczniów z dostępem do internetu. Oznacza to, że wyniki osób z dostępem do internetu są wyższe. Dostęp do internetu ma pozytywny wpływ na wynik egzaminu.

Przeanalizujmy teraz wykres skrzypcowe dla obu grup:



Rysunek 24: Rozkład wyników egzaminu w zależności od dostępu do internetu.

Wykresy w obu przypadkach mają podobne szerokości, co wskazuje na podobne zróżnicowanie wyników zarówno w pierwszej, jak i w drugiej grupie.

Wysokości pudełka, czyli 50% wyników, są podobne w obu grupach wielkości co oznacza, że rozkład wyników dla 50% uczniów w każdej grupie jest podobny.

Mediana w przypadku uczniów z dostępem do internetu jest większa, co wskazuje na wyższe wyniki w tej grupie.

Podsumowując, dostęp do internetu ma pozytywny wpływ na wynik egzaminu.

Dostęp do dodatkowych zasobów

Zajmijmy się teraz już ostatnią dodatkową pomocną naukową, czyli dostępem do dodatkowych materiałów.

Rysunek 25: Rozkład wyników egzaminu dla uczniów z dużym dostępem do zasobów.

Z wykresu wynika, że suma osób z wysokim dostępem do dodatkowych materiałów to 1898 i, że większość tych osób uzyskało wynik $(66, 71]$. Zbadajmy teraz wyniki egzaminu osób ze średnim dostępem do dodatkowych materiałów.

Rysunek 26: Rozkład wyników egzaminu dla uczniów ze średnim dostępem do zasobów.

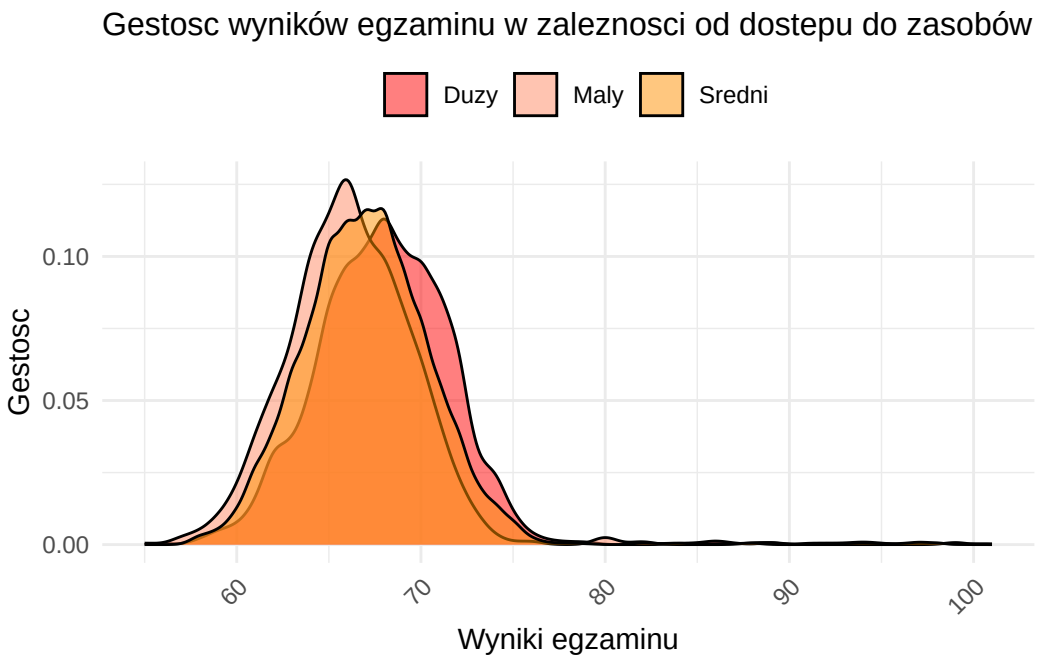
Z wykresu wynika, że suma osób ze średnim dostępem do dodatkowych materiałów to 3202 i, że większość tych osób uzyskało wynik $(63, 68]$, czyli mniej niż w przypadku dużego dostępu.

Zbadajmy teraz wyniki egzaminu osób z niskim dostępem do dodatkowych materiałów.

Rysunek 27: Rozkład wyników egzaminu dla uczniów z niskim dostępem do zasobów.

Z wykresu wynika, że suma osób z niskim dostępem do dodatkowych materiałów to 1273 i, że większość tych osób uzyskało wynik $(65, 70]$, czyli podobnie jak dla uczniów z wysokim dostępem do materiałów dodatkowych.

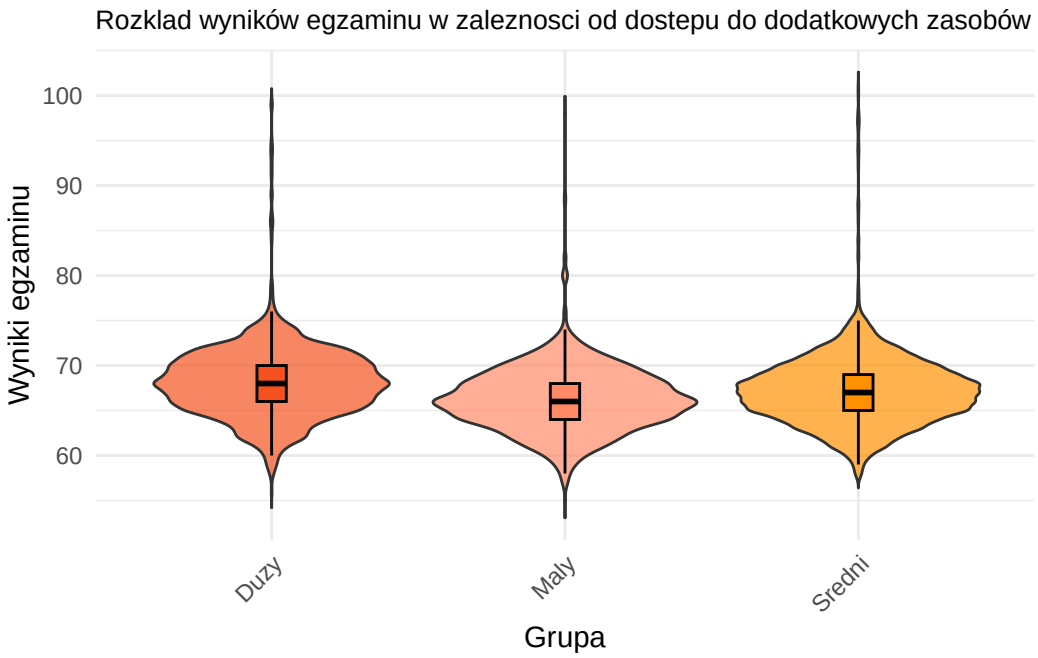
Zbadajmy teraz gęstości wszystkich tych przypadków:



Rysunek 28: Gęstość wyników egzaminu w zależności od dostępu do zasobów.

Z wykresu wynika, że im większy dostęp do dodatkowych materiałów, tym gęstość koncentruje się bardziej po prawej stronie. Oznacza to, że im większą ilością materiałów dysponujemy, tym większa szansa na otrzymanie wyższego wyniku, a im większa dostępność, tym bardziej pozytywny wpływ na wynik egzaminu.

Przeanalizujmy teraz wykres skrzypcowy dla wszystkich trzech grup:



Rysunek 29: Rozkład wyników egzaminu w zależności od dostępu do dodatkowych zasobów.

Wysokości pudełek w każdej grupie są podobne co oznacza, że rozkład wyników dla 50% uczniów w każdym wypadku jest podobny.

Mediana w miarę zwiększania dostępu do dodatkowych źródeł się zwiększa, co wskazuje na wyższe wyniki dla uczniów ze średnim i wysokim dostępem do dodatkowych materiałów.

Podsumowując, im większy dostęp do dodatkowych materiałów, tym lepszy wpływ na wynik końcowy.

Z naszych wykresów wynika, że każda dodatkowa pomoc naukowa ma pozytywny wpływ na wynik końcowy egzaminu.

Im więcej godzin korepetycji weźmiemy i im większy dostęp do materiałów dodatkowych, tym większa szansa na uzyskanie wyższego wyniku. Dostęp do internetu również korzystnie wpływa na wynik końcowy.

Kilka kategorii na raz

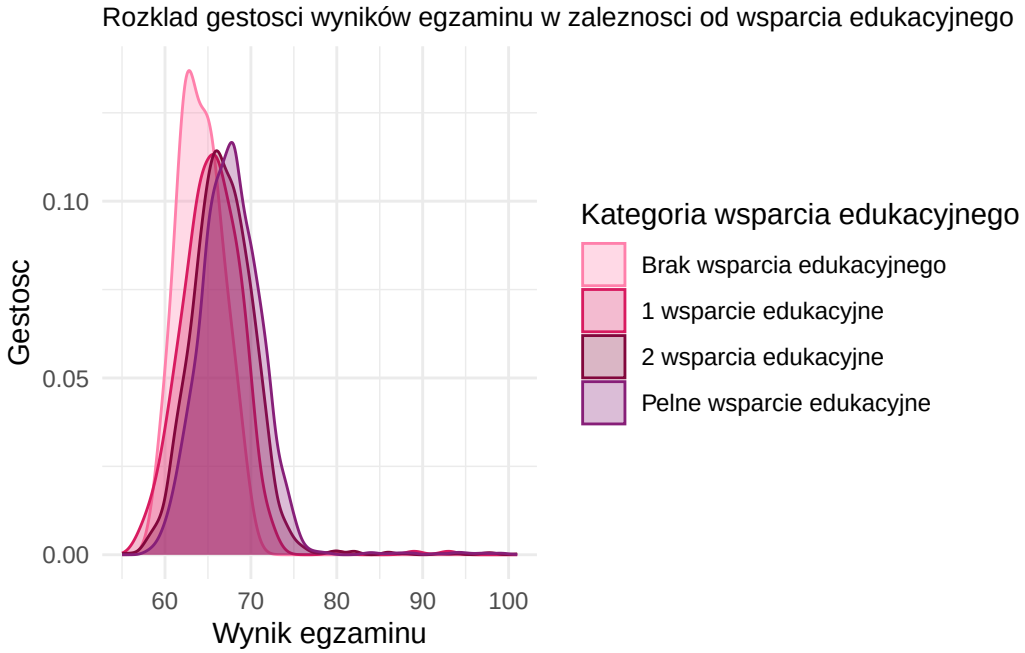
Wiemy już, że pojedyńcze wsparcie edukacyjne wpływa pozytywnie na wynik egzaminu. Sprawdźmy jednak jeszcze jak kilka jego kategorii na raz wpływa na efektywność uczniów na egzaminie.

Za dodatkowe wsparcie edukacyjne przyjmijmy:

- przynajmniej jedną godzinę korepetycji

- dostęp do internetu
- przynajmniej średni dostęp do materiałów dodatkowych

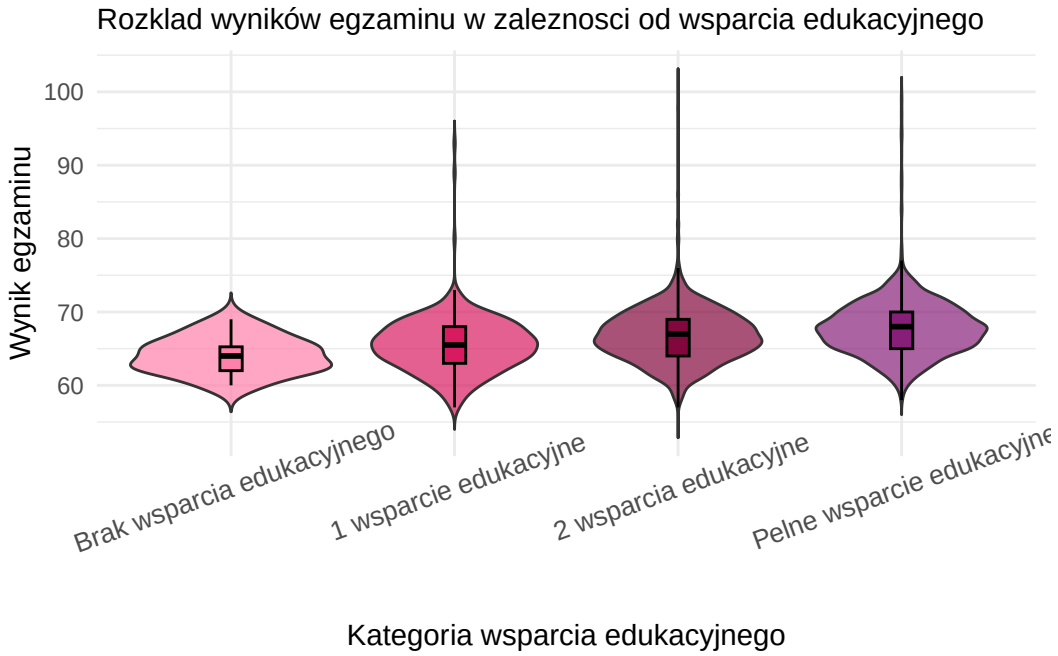
Wykres gęstości wyników egzaminu w zależności od ilości dodatkowych materiałów przedstawia się następująco:



Rysunek 30: Rozkład gęstości wyników egzaminu w zależności od wsparcia edukacyjnego.

Z wykresu wynika, że im większe wsparcie edukacyjne, tym gęstość koncentruje się wokół wyższych wyników. Oznacza to, że im więcej wsparcia edukacyjnego dostajemy, tym większa szansa na otrzymanie wyższego wyniku.

Przeanalizujmy teraz wykresy skrzypcowe każdego przypadku:



Rysunek 31: Rozkład wyników egzaminu w zależności od wsparcia edukacyjnego.

Z wykresu wynika, że im większe wsparcie edukacyjne, tym większa mediana, a co za tym idzie - wyższe wyniki egzaminu. W przypadku braku wsparcia edukacyjnego wartości są najmniej rozproszone i nie ma wartości odstających.

Podsumowując, większe wsparcie edukacyjne ma zauważalnie bardziej pozytywny wpływ na wynik egzaminu końcowego.

Podsumowanie

Na podstawie przedstawionych statystyk i wykresów wybranego przez nas zbioru danych wynika, że zarówno czas poświęcony na naukę, frekwencja, jak i dostęp do dodatkowych materiałów edukacyjnych wpływa korzystnie na wynik egzaminu końcowego.

Z naszych analiz wynika, że liczba godzin pozytywnie koreluje z wynikami egzaminów. Im więcej czasu uczniowie poświęcają na naukę, tym wyższe wyniki osiągają. Zależność ta została potwierdzona za pomocą modelu regresji liniowej, który wskazał wzrost wyniku o 0.29 punktu za każdą dodatkową godzinę nauki.

Obecność na zajęciach również ma silny wpływ na wyniki egzaminów. Uczniowie z wyższą frekwencją osiągają lepsze wyniki w porównaniu do uczniów mniej uczęszczających na zajęcia. Ta relacja również została potwierdzona za pomocą modelu regresji liniowej, który wskazał wzrost wyniku o 0.2 punktu za każdy dodatkowy punk procentowy obecności.

Uczniowie korzystający z internetu, korepetycji i ci z większym dostępem do materiałów edukacyjnych również osiągali wyższe wyniki w porównaniu do uczniów, którzy nie mieli takowego wsparcia. Dodatkowo, kombinacja kilku rodzajów wsparcia edukacyjnego prowadziła do najwyższych wyników egzaminacyjnych.

Wyniki sugerują, że zadbanie o wysoką frekwencję, dobrą organizację czasu nauki oraz zapewnienie dostępu do różnorodnych materiałów edukacyjnych może prowadzić do lepszych wyników uczniów. Warto promować te czynniki jako elementy wspierające sukces w szkole. Większa świadomość rodziców, nauczycieli i uczniów na ten temat może skutkować znaczną poprawą wyników egzaminacyjnych.