



**Politechnika
Śląska**

Dokumentacja projektowa

Analiza Statystyczna wyników studentów

*Analiza statystyczna wyników studentów i badanie czynników,
wpływających na te wyniki w nauce*

Kierunek: Informatyka

Członkowie zespołu:

Michał Tarnawa

Michał Żarłok

Wojciech Grzywocz

Gliwice, 2025/2026

Spis treści

1	Opis danych	3
1.1	Geneza zbioru i jego struktura	3
1.2	Analiza danych i ich wpływ	4
1.2.1	Analiza struktury płci	4
1.2.2	Analiza struktury wieku	5
1.2.3	Analiza wyników i ocen	6
1.2.4	Analiza wyników w podziale na wydziały	7
1.2.5	Wpływ aktywności na wynik końcowy	8
1.2.6	Wpływ obecności na ocenę końcową	9
1.2.7	Analiza czasu poświęconego na naukę w tygodniu	11
1.2.8	Zależność oceny końcowej od czasu nauki	12
1.2.9	Analiza długości snu	13
1.2.10	Wpływ długości snu na wyniki	14
1.2.11	Analiza poziomu stresu	14
1.2.12	Wpływ stresu na wynik końcowy	15
2	Testy parametryczne (t-studenta)	17
2.1	Założenia	17
2.2	Wyniki w nauce w zależności od płci	17
2.2.1	Hipotezy	17
2.2.2	Zbadanie jednorodności wariancji	17
2.2.3	Sprawdzenie czy zmienna Total_Score(ostateczny wynik) ma rozkład normalny	18
2.2.4	Liczebność próby	18
2.2.5	Badanie hipotezy	18
2.3	Wynik końcowy, a dostęp do internetu w domu.	20
2.3.1	Hipotezy	20
2.3.2	Zbadanie jednorodności wariancji w rozkładzie	20
2.3.3	Sprawdzenie czy zmienna Total_Score ma rozkład normalny w podziale na grupy	20
2.3.4	Przeprowadzenie testu zgodności	21
3	Testy nieparametryczne	23
3.1	Porównanie poziomu wyników z projektu i z zadań	23
3.1.1	Hipotezy badawcze	23
3.1.2	Sprawdzenie czy obie zmienne mają rozkład normalny	23
3.1.3	Badanie hipotezy	24
3.2	Porównanie poziomu stresu z względem płci	25
3.2.1	Hipotezy badawcze	25

3.2.2	Sprawdzenie czy zmienna Stress_Level__(1-10) ma rozkład normalny w podziale na grupy	25
3.2.3	Badanie hipotezy	26
3.3	Porównanie wyników z połowy semestru (Mid_Term) z wynikiem z finalnego egzaminu (Final_Score)	28
3.3.1	Hipotezy badawcze	28
3.3.2	Sprawdzenie czy zmienne Mid_term oraz Final_Score mają rozkład normalny	28
3.3.3	Badanie hipotezy i przeprowadzenie testu	29
4	Nieparametryczna ANOVA	30
4.1	Sprawdzenie zależności między poziomem dochodów w rodzinie a oceną końcową	30
4.1.1	Hipotezy badawcze	30
4.1.2	Sprawdzenie czy finalna ocena ma rozkład normalny . .	30
4.1.3	Badanie hipotezy	32
5	Parametryczna ANOVA	33
5.1	Wpływ godzin snu na wynik końcowy	33
5.1.1	Hipotezy	33
5.1.2	Sprawdzenie czy Total_Score ma rozkład normalny względem grup	33
5.1.3	Sprawdzenie jednorodności wariancji względem grup . .	34
5.1.4	Badanie hipotezy	35
6	Regresja liniowa	36
6.1	Sprawdzanie w jaki sposób wpływają poszczególne czynniki na ocenę końcową	36
6.1.1	Poprawiona Regresja	37
6.1.2	Analiza reszt	38
7	Wnioski	39
8	Bibliografia	40

1 Opis danych

1.1 Geneza zbioru i jego struktura

Analiza zbioru danych z serwisu Kaggle: Students Grading Dataset.

Zbiór ten zawiera informacje o wynikach studentów oraz czynnikach, które mogą mieć na nie wpływ. Są to:

- **Student_ID**: ID Studenta - Unikalny identyfikator każdego studenta.
- **First_Name**: Imię - Imię studenta.
- **Last_Name**: Nazwisko - Nazwisko studenta.
- **Email**: Email - Adres e-mail kontaktowy.
- **Gender**: Płeć - Płeć: Mężczyzna (Male), Kobieta (Female), Inna (Other).
- **Age**: Wiek - Wiek studenta.
- **Department**: Wydział/Kierunek - Kierunek studiów studenta.
- **Attendance (%)**: Frekwencja (%) - Procentowa frekwencja na zajęciach (0–100%).
- **Midterm_Score**: Wynik z egzaminu śródsesemestralnego (0–100).
- **Final_Score**: Wynik z egzaminu końcowego (0–100).
- **Assignments_Avg**: Średnia z zadań domowych -(0–100).
- **Quizzes_Avg**: Średnia z kartkówek/kolowiów (0–100).
- **Participation_Score**: Ocena z aktywności (0–10).
- **Projects_Score**: Ocena z projektów - (0–100).
- **Total_Score**: Łączny wynik -
- **Grade**: Ocena końcowa - Ocena literowa (A, B, C, D, F).
- **Study_Hours_per_Week**: Średnia liczba godzin nauki w tygodniu.
- **Extracurricular_Activities**: Udział w zajęciach pozalekcyjnych (Tak/Nie).
- **Internet_Access_at_Home**: Dostęp do internetu w domu (Tak/Nie).

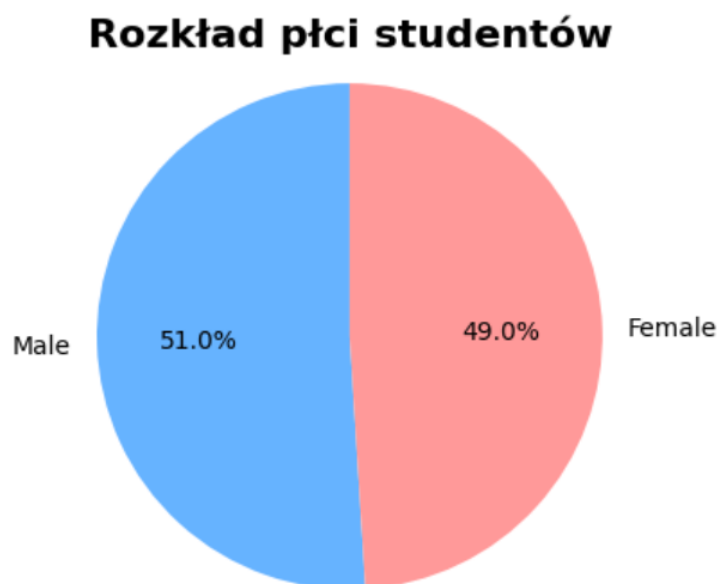
- **Parent_Education_Level**: Najwyższy poziom wykształcenia rodziców (Brak, Szkoła średnia, Licencjat, Magister, Doktorat/PhD).
- **Family_Income_Level**: Poziom dochodów rodziny (Niski, Średni, Wysoki).
- **Stress_Level** (1–10): Samoocena poziomu stresu (1: niski – 10: wysoki).
- **Sleep_Hours_per_Night**: Średnia liczba godzin snu na dobę.

Baza danych posiada 5000 rekordów i nie zawiera brakujących wartości (null).

W analizie naszego zbioru usunięto następujące kolumny: **Student_ID**, **First_Name**, **Last_Name**, **Email**, gdyż zawierają one wartości tekstowe, których nie można sensownie przekształcić do dalszej analizy (nie wnoszą wkładu w modele matematyczne).

1.2 Analiza danych i ich wpływ

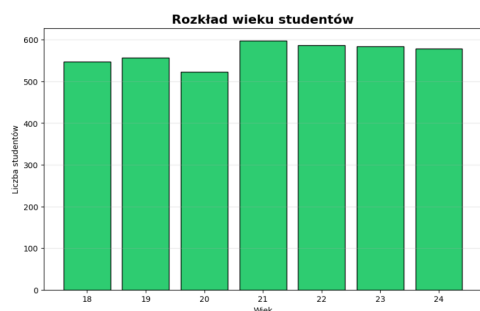
1.2.1 Analiza struktury płci



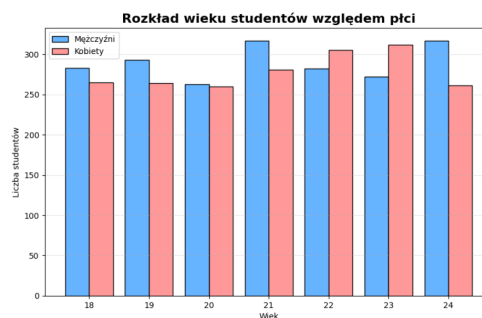
Rysunek 1: Wykres kołowy ukazując rozkład płci w analizowanym zbiorze

Analizując rysunek 1 widzimy, że mamy prawie równomierny podział, gdzie mężczyźni stanowią 51% badanej populacji, a kobiety 49%, co jest wartością podobną, do proporcji przy urodzeniu.

1.2.2 Analiza struktury wieku



Rysunek 2: Rozkład wieku studentów

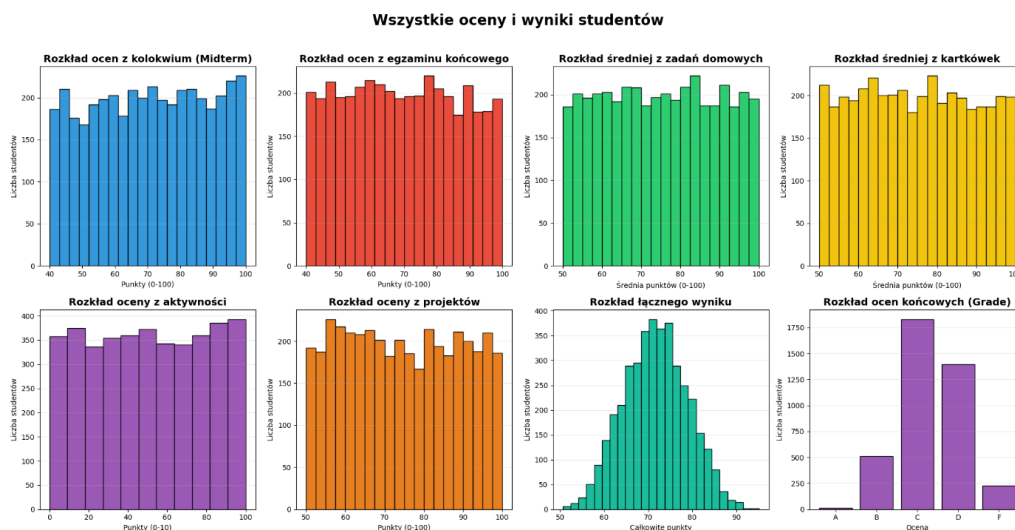


Rysunek 3: Rozkład wieku względem płci

Analizując Rysunek 2, widzimy, że dane obejmują osoby w wieku od 18 do 24 lat. Rozkład jest wyglądem na równomierny, a liczebność w każdej grupie wiekowej jest podobna i zamyka się w granicach 500–600 osób. Rozkład przypomina rozkład jednostajny (nie jest to teza, tylko hipoteza).

Łącząc to z analizą Rysunku 3 widzimy, mężczyźni dominują w ilościach pod względem wieku. Wyjątkie są osoby w wieku 22 i 23, gdzie jest więcej kobiet. Największa przepaść jest wśród osób w wieku 24 lat.

1.2.3 Analiza wyników i ocen



Rysunek 4: Zbiorcze zestawienie rozkładów ocen i wyników studentów

Jak widzimy na Rysunku 4, rozkłady punktów z kategorii: **Midterm_Score**, **Final_Score**, **Assignments_Avg** oraz **Quizzes_Avg** posiadają bardzo podobny kształt. Są one stosunkowo równomierne, z niewielkim odchyleniem w stronę wyższych ocen (60–100 pkt). Oznacza to, że studenci są dobrze przygotowani.

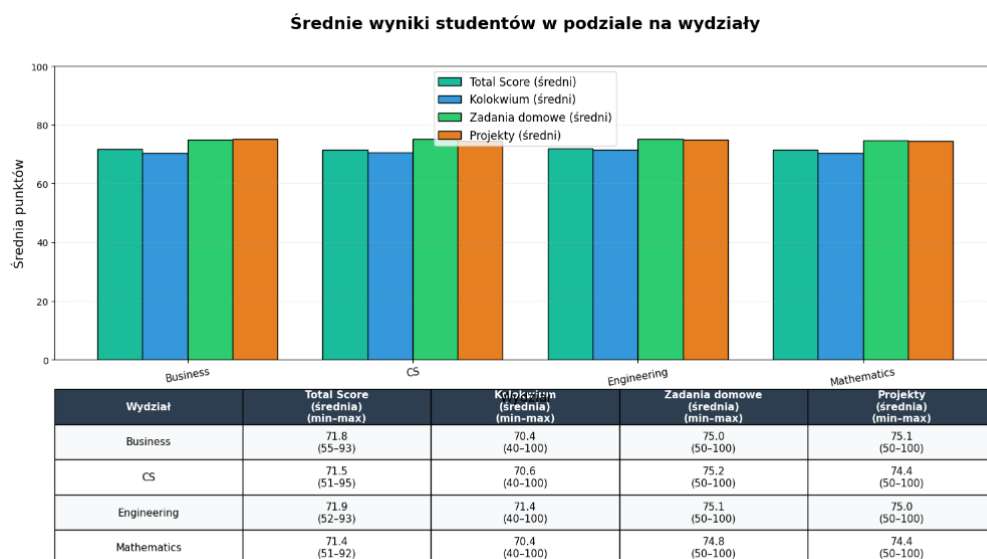
Natomiast wykres oceny z projektów wyróżnia się na tle pozostałych – posiada wyraźny szczyt w przedziale 80–100 punktów, co sugeruje, że z tą formą zaliczenia studenci radzili sobie najlepiej.

Wykres oceny z aktywności charakteryzuje się dość równomiernym rozkładem w zakresie 0–10, z lekkim nagromadzeniem przy wyższych wartościach. Wskazuje to na fakt, że większość studentów wykazywała umiarkowaną lub wysoką aktywność na zajęciach.

Wykres łącznego wyniku przypomina rozkład normalny. Szczyt wartości przypada na okolice 70–75 punktów, co zgadza się z wcześniej wyliczoną średnią wynoszącą 71,6. Potwierdza to teorię, że wiele zmiennych losowych w przyrodzie bazuje na tym rozkładzie. Wykres ocen końcowych jest odwzorowaniem wykresu łącznego wyniku, zrzuconym na 5 kategorii (ocen literowych).

Podsumowując, można przypuszczać, że osoby osiągające bardzo dobre wyniki (np. z egzaminów) wcale nie musiały otrzymać równie wysokich not z innych kategorii, takich jak aktywność czy projekty.

1.2.4 Analiza wyników w podziale na wydziały

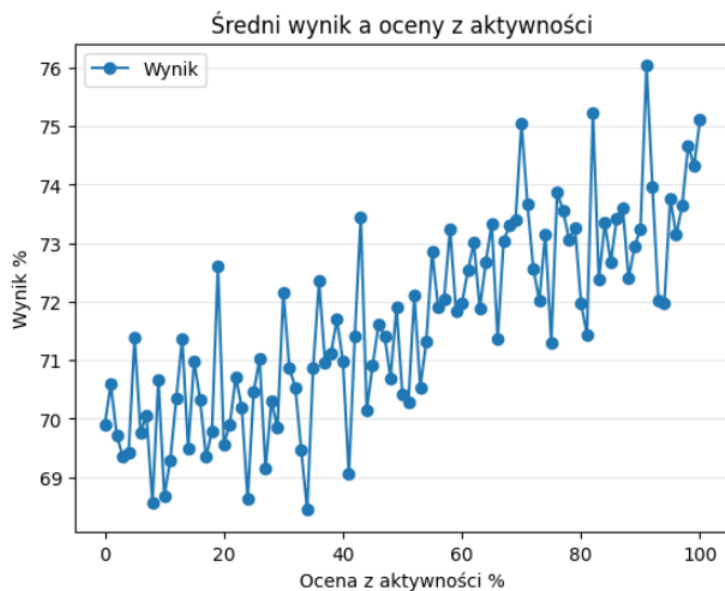


Rysunek 5: Średnie wyniki studentów na poszczególnych wydziałach

Jak widać na Rysunku 5, niezależnie od wydziału wyniki są podobne. Studenci najlepiej radzą sobie z zadaniami domowymi oraz projektami. Najślabiej wypada kolokwium, natomiast wynik łączny jest nieco wyższy od niego, co może świadczyć o potencjalnej mobilizacji w okresie egzaminów końcowych.

Analizując tabelę, nie zaobserwowane znaczących dysproporcji między wydziałami - różnice oscylują w granicach jednego punktu procentowego (p.p.). Należy jednak zauważyć, że najlepiej prezentują się wyniki studentów z wydziału Inżynierii, którzy osiągnęli najwyższe średnie w 3 z 4 analizowanych przez nas kategorii.

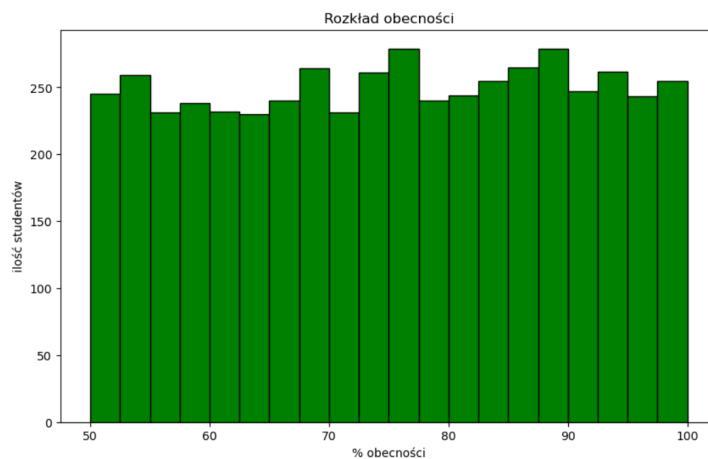
1.2.5 Wpływ aktywności na wynik końcowy



Rysunek 6: Średni wynik końcowy a ocena z aktywności

Wykres przedstawiony na Rysunku 6 nie jest regularny. Wartości rosną i maleją. Można jednak zauważyć zależność, iż im wyższa ocena z aktywności, tym lepsza ocena finalna. Gdyby poddać te dane regresji liniowej (pomijając fakt, wielkiego rozrzutu danych), współczynnik kierunkowy prostej byłby dodatni ($a > 0$), co świadczy o trendzie wzrostowym.

1.2.6 Wpływ obecności na ocenę końcową



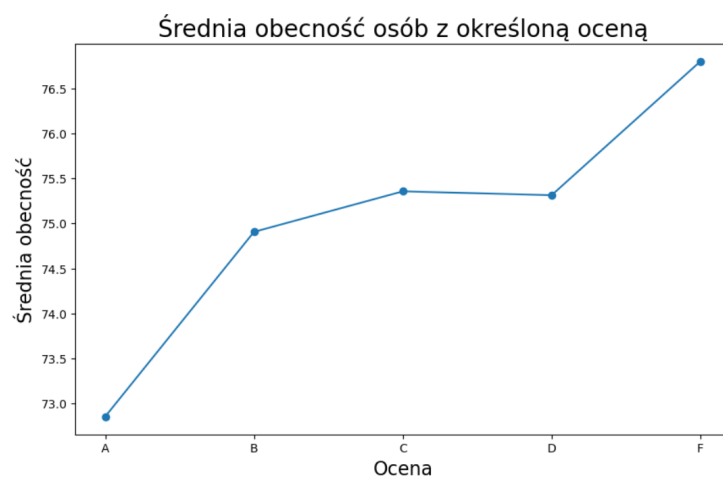
Rysunek 7: Histogram obecności

Na rysunku 7 widzimy, że obecność studentów ma rozkład dosyć równomierny. Średnia obecność wynosi: 75,36%, natomiast mediana wynosi: 75,67%. Wpływ obecności na wynik na koniec semestru prezentują się następująco:



Rysunek 8: Średni wynik końcowy osób z określonym poziomem obecności

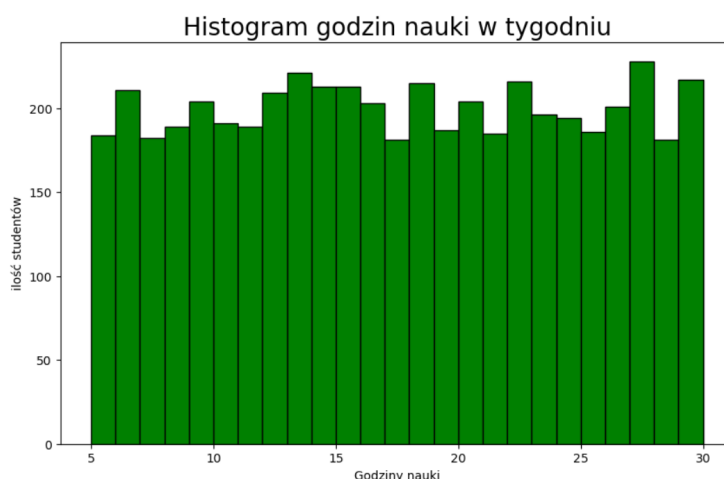
Na Wykresie 8 możemy zauważyć delikatny trend spadkowy, aby jednak dokładniej zbadać wpływ obecności na ocenę możemy odwrócić podejście i sprawdzić jak prezentują się średnia obecność dla osób osiągających daną ocenę końcową.



Rysunek 9: Średni obecność dla osób z daną oceną końcową

Na Wykresie 9 możemy zauważyć, że osoby, które uzyskały najlepsze wyniki najmniej uczęszczały na zajęcia. Może to świadczyć o tym, że osoby zdolniejsze wolały uczyć się indywidualnie, a mając więcej czasu na naukę indywidualną, przełożyło się to na lepsze wyniki końcowe.

1.2.7 Analiza czasu poświęconego na naukę w tygodniu



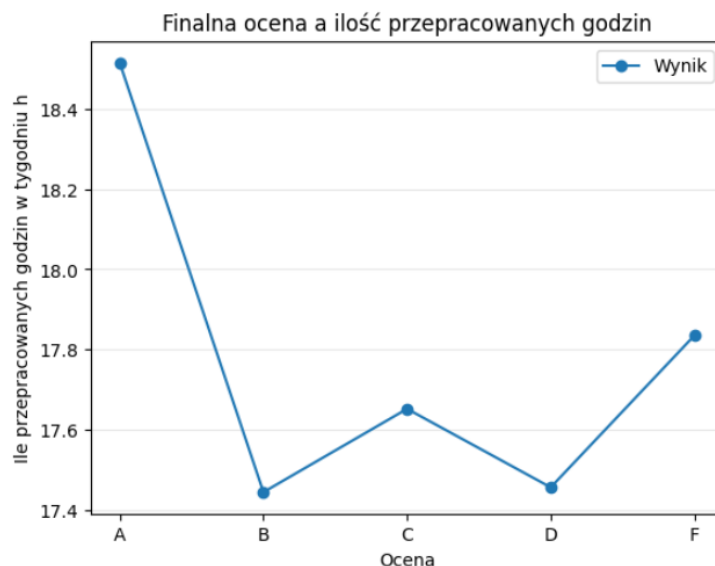
Rysunek 10: Histogram średniego czasu poświęconego na naukę w tygodniu

Analizując Rysunek 10, możemy zauważyć, że średni czas poświęcony na naukę w tygodniu ma rozkład dosyć równomierny. Możemy wyróżnić jednak sporą grupę studentów, którzy na naukę poświęcają między 12 a 17 godzin. Najwyższy słupek na histogramie to z kolei osoby, które poświęcały między 27 a 28 godzin na naukę w tygodniu.

Dodatkowo, obliczony średni czas nauki w tygodniu (z uwzględnieniem wszystkich 5000 studentów) wyniósł: 17.52 godziny w tygodniu.

Mediana to natomiast: 17.4 godziny w tygodniu.

1.2.8 Zależność oceny końcowej od czasu nauki



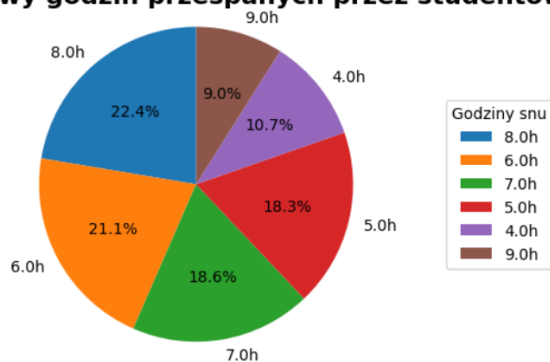
Rysunek 11: Finalna ocena a średnia ilość godzin nauki w tygodniu

Analizując Rysunek 11, można zauważyć, że osoby z najwyższą oceną (A) poświęcały na naukę najwięcej czasu. Co ciekawe, studenci z najgorszą oceną (F) uplasowali się na drugim miejscu pod względem liczby przepracowanych godzin.

Taka anomalia może sugerować dwie hipotezy: albo ankietowani w tej grupie podali nieprawdziwe dane, albo poświęcili dużo czasu na przyswojenie materiału, którego nie byli w stanie zrozumieć (nieefektywna nauka). Z kolei najmniej czasu na naukę poświęciły osoby, które uzyskały ocenę B.

1.2.9 Analiza długości snu

Wykres kołowy godzin przespanych przez studentów



Rysunek 12: Wykres kołowy godzin przespanych przez studentów

Analizując Rysunek 12, widzimy, że aż 68,6% studentów śpi mniej niż wymagane 8 godzin. Co więcej, 29% badanych deklaruje, że przesypia jedynie 4 lub 5 godzin na dobę.

1.2.10 Wpływ długości snu na wyniki



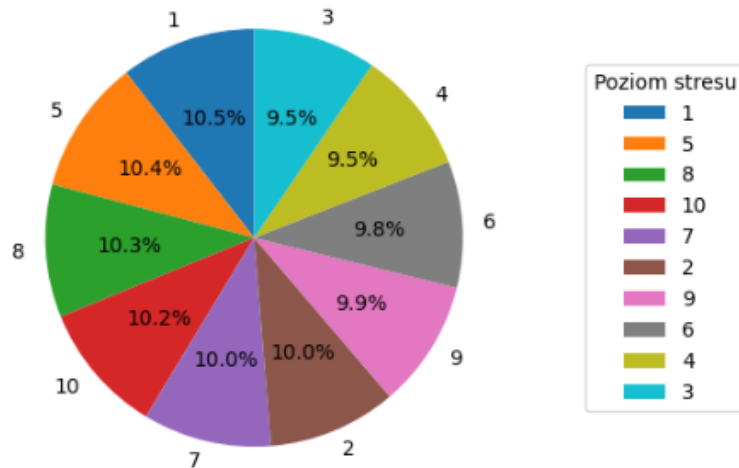
Rysunek 13: Zależność między średnim wynikiem a liczbą godzin snu

Jak widać na Rysunku 13, najlepsze wyniki osiągają osoby, które śpią 8 godzin dziennie, co jest zgodne z ogólnymi zaleceniami. Najgorzej wypadają studenci, którzy śpią jedynie 4 godziny. Co istotne, dłuższy sen (powyżej 8 godzin) wcale nie przekłada się na polepszenie wyników.

1.2.11 Analiza poziomu stresu

Analizując Rysunek 14 (poniżej), widzimy, że każdy poziom stresu rozkłada się równomiernie. Udział każdej kategorii wynosi w przybliżeniu około 10%, co wskazuje na brak znaczących odstępstw w rozkładzie tej zmiennej. Należy podkreślić fakt, iż przedstawione dane, to samoocena poziomu stresu.

Rozkład stresu studentów



Rysunek 14: Rozkład poziomu stresu wśród studentów

1.2.12 Wpływ stresu na wynik końcowy



Rysunek 15: Średni wynik końcowy w zależności od poziomu stresu

Jak widzimy na Rysunku 15, stres nie zawsze wpływa korzystnie na wyniki. Osoby najbardziej zestresowane osiągają najgorsze rezultaty, które są jednak tylko nieznacznie niższe od wyników osób nieodczuwających stresu.

Co ciekawe, najlepsze wyniki uzyskują studenci, których poziom stresu mieści się w przedziale 8-9 - w tej grupie średni wynik przekracza 72%. Wniosek może być taki, że lekki stres pobudza do działania.

2 Testy parametryczne (t-studenta)

2.1 Założenia

Przed przeprowadzeniem testów zakładamy następujące postulaty:

- współczynnik istotności $\alpha = 0.05$,
- ze względu na dużą liczebność próby pomijamy test Shapiro-Wilka podczas testów normalności - przy takiej liczebności jest on nieskuteczny,
- liczebność próby wynosi 5000, a więc jest ona spełniona w każdym przypadku.

2.2 Wyniki w nauce w zależności od płci

2.2.1 Hipotezy

Niech posiadamy następujące hipotezy dotyczące średnich wyników semestralnych:

$$\begin{array}{l} H_0 : \mu_m = \mu_f \\ H_1 : \mu_m \neq \mu_f \end{array}$$

gdzie:

- μ_m – średni wynik całkowity (Total_Score) dla mężczyzn
- μ_f – średni wynik całkowity (Total_Score) dla kobiet

2.2.2 Zbadanie jednorodności wariancji

Po wykonaniu testu:

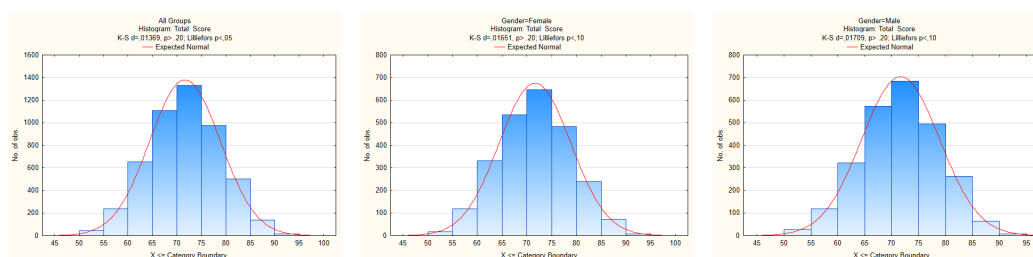
Tabela 1: Wyniki testu Levene’a jednorodności wariancji dla zmiennej Total_Score (Kobiety vs Mężczyźni)

Porównanie	F	p
Kobiety vs Mężczyźni	1,002	0,951

Wniosek: Wartość $p = 0,9514 \gg 0,05$ (a także znacznie większa niż 0,10 czy 0,20). **Nie ma podstaw do odrzucenia hipotezy zerowej** o równości wariancji między grupą kobiet i mężczyzn.

2.2.3 Sprawdzenie czy zmienna Total_Score(ostateczny wynik) ma rozkład normalny

Przed rozpoczęciem weryfikacji hipotezy H_0 należy najpierw sprawdzić, czy zmienna Total_Score ma rozkład normalny. W tym celu wykorzystano test zgodności rozkładu w programie Statistica. Wyniki wskazują, że nie ma podstaw do odrzucenia hipotezy o normalności rozkładu ($p > 0,2 > 0,05$). Zatem można przystąpić do wykonania testu parametrycznego.



(a) Wykres dla wszystkich kategorii

(b) Wykres dla kobiet

(c) Wykres dla mężczyzn

Rysunek 16: Wykresy dopasowania do normalności:

Analizując otrzymane wyniki:

Tabela 2: Wyniki testu normalności Kołmogorowa-Smirnowa dla zmiennej Total_Score

Grupa	Statystyka d	p
Wszyscy badani	0,01369	$> 0,20$
Kobiety	0,01651	$> 0,20$
Mężczyźni	0,01709	$> 0,20$

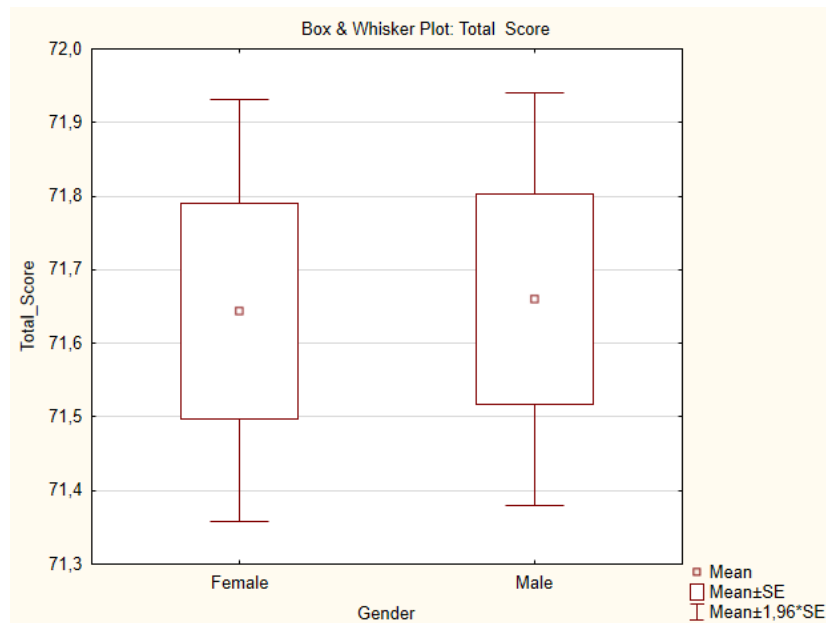
Na podstawie uzyskanych wyników stwierdzono, że nie ma podstaw do odrzucenia hipotezy, że rozkład jest normalny. $p > 0.2 > 0.05$. A więc możemy wykonać test.

2.2.4 Liczebność próby

Z wcześniejszych analiz wynika, że obie płcie są stosunkowo równoliczne, a liczebność każdej grupy wyraźnie przekracza 30 obserwacji.

2.2.5 Badanie hipotezy

Dokonamy teraz testu t-Studenta dla prób niezależnych, grupowanych, ale najpierw przeanalizujemy wykres wąsowy dla obu płci.



Rysunek 17: Wykresy pudełkowy dla średniej obu płci

Jak możemy zauważyć na wykresie 17, zarówno średnie, jak i kwantyle dla obu płci są niemal identyczne. W związku z tym przeprowadzono test t-Studenta dla prób niezależnych.

Po jego przeprowadzeniu możemy zauważyć, iż średnie wyniki `Total_Score` nie różnią się między kobietami a mężczyznami.

Tabela 3: Statystyki opisowe oraz wyniki testu t-Studenta dla zmiennej `Total_Score` w podziale na płeć

Płeć	<i>N</i>	<i>M</i>	<i>SD</i>	<i>t</i> (4998)	<i>p</i>
Kobiety	2449	71,64	7,24	-0,075	0,941
Mężczyźni	2551	71,66	7,23		

Uwaga. *M* – średnia, *SD* – odchylenie standardowe, *N* – liczebność próby.

Test równości wariancji wykazał brak istotnych różnic ($F = 1,002$, $p = 0,951$). Wyniki testu t-Studenta dla prób niezależnych: $t(4998) = -0,075$, $p = 0,941$.

Oznacza to, że nie stwierdzono istotnej statystycznie różnicy między średnimi wynikami kobiet i mężczyzn. Średnie wyniki w obu grupach są statystycznie porównywalne.

Wniosek: Nie ma podstaw do odrzucenia hipotezy zerowej $H_0 : \mu_m = \mu_f$

2.3 Wynik końcowy, a dostęp do internetu w domu.

2.3.1 Hipotezy

Niech posiadamy następujące hipotezy dotyczące średnich wyników semestralnych:

$$\begin{array}{l} H_0 : \mu_t = \mu_n \\ H_1 : \mu_t \neq \mu_n \end{array}$$

gdzie:

- μ_t – średni finalny wynik semestralny dla osób z dostępem do internetu
- μ_n – średni finalny wynik semestralny dla osób bez dostępu do internetu

2.3.2 Zbadanie jednorodności wariancji w rozkładzie

Przed rozpoczęciem badania hipotezy sprawdzimy, czy zmienna Total_Score podzielona na grupy osób z dostępem do internetu i z brakiem dostępu do niego spełnia założenie jednorodności wariancji.

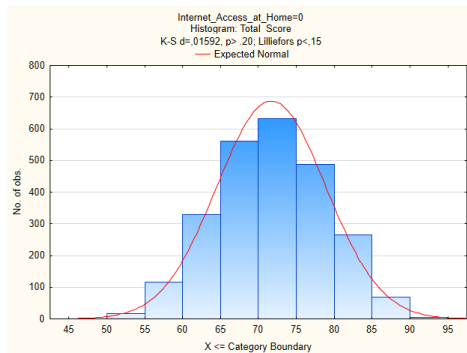
Tabela 4: Wyniki testów jednorodności wariancji dla zmiennej Total_Score

Test	Statystyka F	p
Test Levene'a	0,073	0,788
Test Browna-Forsythe'a	0,075	0,784

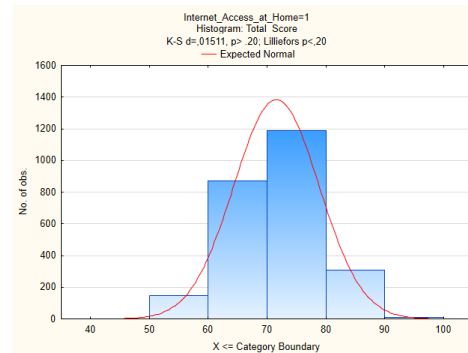
Analizując wyniki testów jednorodności wariancji (Tabela 4), widzimy, że wartość p dla obu testów jest znacznie większa od założonego poziomu istotności $\alpha = 0,05$. Oznacza to, że nie ma podstaw do odrzucenia hipotezy o równości wariancji w obu grupach. Założenie o jednorodności wariancji niezbędne do wykonania testu t-Studenta zostało spełnione.

2.3.3 Sprawdzenie czy zmienna Total_Score ma rozkład normalny w podziale na grupy

Wiemy już z wcześniejszego testu, że cała zmienna Total_Score ma rozkład normalny. Musimy teraz sprawdzić, czy ten rozkład jest zachowany po podzieleniu jej na grupy w zależności od dostępu do internetu.



(a) Wykres dla osób bez dostępu do internetu



(b) Wykres dla osób z dostępem do internetu

Rysunek 18: Wykresy dopasowania do normalności zmiennej `Total_Score` względem dostępu do internetu:

Analizując rysunek 18 możemy powiedzieć, że w obu grupach rozkład przypomina rozkład normalny, ale w celu potwierdzenia tej tezy potrzebny będzie test Kołmogorowa–Smirnowa.

Po wykonaniu tego testu:

Tabela 5: Testy normalności rozkładu zmiennej `Total_Score` (Projekt) w podziałach na dostęp do internetu w domu

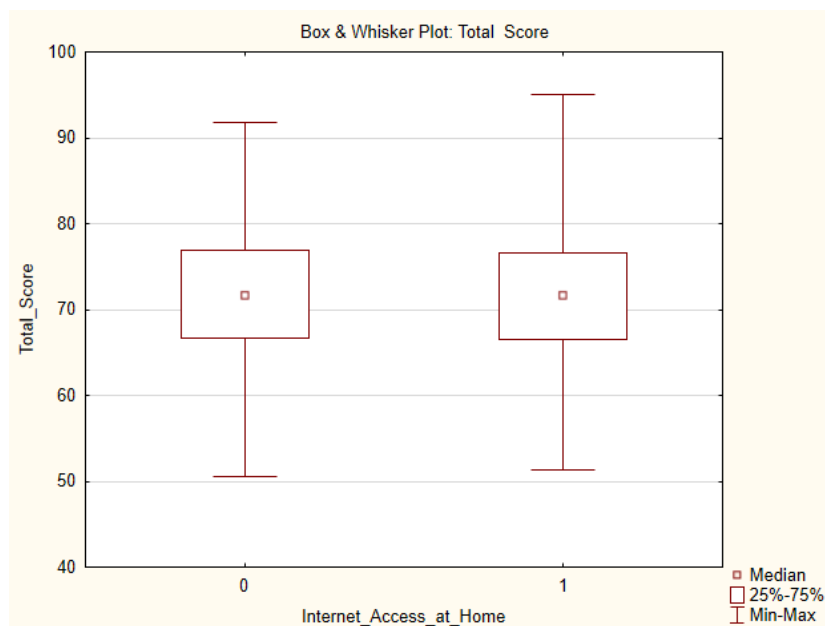
Grupa	Statystyka	p
Internet_Access_at_Home = 0	$D = 0,01592$ (K-S), Lilliefors	$p > 0,20$; $p < 0,15$
Internet_Access_at_Home = 1	$D = 0,01511$ (K-S), Lilliefors	$p > 0,20$; $p < 0,20$

Przyglądając się tabelce 5 jasno możemy zauważyć, iż wyniki testów Kołmogorowa–Smirnowa oraz Lillieforsa przekraczają poziom istotności wynoszący 0,05, a samo p jest $> 0,2$ – z wyjątkiem testu Lillieforsa dla grupy `Internet_Access_at_Home = 0`, gdzie wartość ta jest $> 0,15$.

Nie ma podstaw do odrzucenia hipotezy o normalności rozkładu.

2.3.4 Przeprowadzenie testu zgodności

Teraz dokonamy testu parametrycznego t-Studenta dla grup niezależnych. Najpierw porównamy średnie na podstawie wykresu pudełkowego z wąsami.



Rysunek 19: Wykresy pudełkowy dla średniej obu płci

Przyglądając się rysunkowi 19 możemy wstępnie założyć potwierdzenie hipotezy zerowej, że grupy mają taki sam rozkład, jednakże do ostatecznego podjęcia decyzji potrzebne jest wykonanie testu t-Studenta.

Po jego wykonaniu wyniki prezentują się następująco:

Tabela 6: Statystyki opisowe oraz wyniki testu t-Studenta dla zmiennej Total_Score w podziale na dostęp do internetu

Grupa (Dostęp do internetu)	<i>N</i>	<i>M</i>	<i>SD</i>	<i>t</i> (4998)	<i>p</i>
Brak dostępu (0)	2480	71,71	7,19	0,551	0,582
Dostęp w domu (1)	2520	71,60	7,27		

Uwaga. *M* – średnia, *SD* – odchylenie standardowe, *N* – liczebność próby.

Analizując wyniki testu z tabeli 6 jasno widzimy, iż $p = 0,582$, co jest większe od poziomu istotności $\alpha = 0,05$. Same średnie wyników Total_Score dla osób bez dostępu do internetu jak i z dostępem do niego są zbliżone (odpowiednio 71,71 oraz 71,60). To samo tyczy się odchylenia standardowego (7,19 i 7,27).

Wniosek:

Nie ma podstaw do odrzucenia hipotezy H_0 .

Wychodząc poza statystykę możemy się domyślać, iż osoby, które nie mają dostępu do internetu w domu, mają do niego dostęp na uczelni, co

sprawia, że nie ma to takiego wpływu. Oprócz tego mogą nie mieć dostępu do internetu w domu, ale mają dostęp do internetu mobilnego.

3 Testy nieparametryczne

3.1 Porównanie poziomów wyników z projektu i z zadań

3.1.1 Hipotezy badawcze

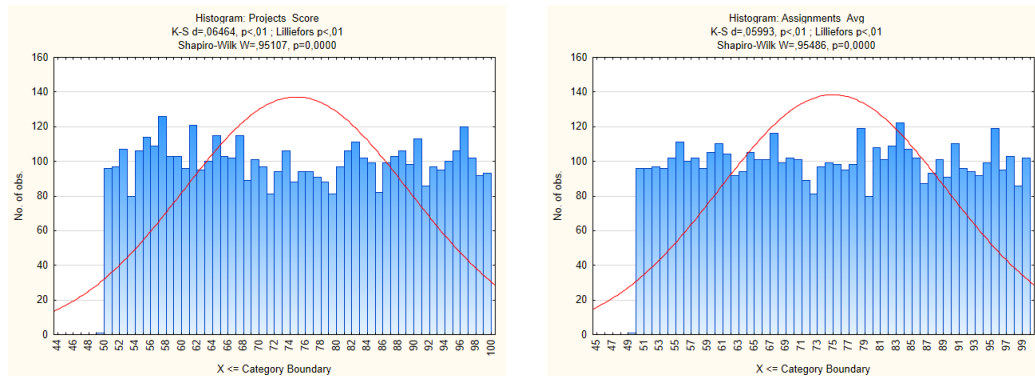
Celem analizy jest ustalenie, czy wyniki uzyskiwane przez studentów z projektów (*Projects_Score*) oraz zadań (*Assignments_Avg*) mają ten sam rozkład.

W tym celu sformułowano następujące hipotezy statystyczne :

$$\begin{aligned} H_0 : F_{\text{Projects}}(x) &= F_{\text{Assignments}}(x) \quad (\text{rozkłady zmiennych są identyczne}) \\ H_1 : F_{\text{Projects}}(x) &\neq F_{\text{Assignments}}(x) \quad (\text{rozkłady zmiennych różnią się}) \end{aligned}$$

3.1.2 Sprawdzenie czy obie zmienne mają rozkład normalny

Sprawdźmy, czy zmienne mają rozkład normalny z współczynnikiem istotności $\alpha = 0,05$.



(a) Rozkład zmiennej *Projects_Score* (ocena z projektu)

(b) Rozkład zmiennej *Assignments_Avg* (średnia z zadań)

Rysunek 20: Porównanie rozkładów empirycznych z rozkładem normalnym dla dwóch zmiennych częściowych

Wyniki testów normalności dla obu zmiennych jasno dowodzą, że zmienne nie mają rozkładu normalnego. Zarówno ocena z samego projektu (*Projects_Score*), jak i średnia z zadań/ćwiczeń (*Assignments_Avg*) mają rozkłady różne od rozkładu normalnego.

Tabela 7: Testy normalności dla zmiennych częściowych oceny w projekcie ($N \approx 5000$)

Zmienna	Kolmogorov-Smirnov	Lilliefors (p)	Shapiro-Wilk (W, p)
<i>Projects_Score</i>	$d = 0,0646$	$p < 0,01$	$W = 0,9511, p < 0,001$
<i>Assignments_Avg</i>	$d = 0,0599$	$p < 0,01$	$W = 0,9549, p < 0,001$

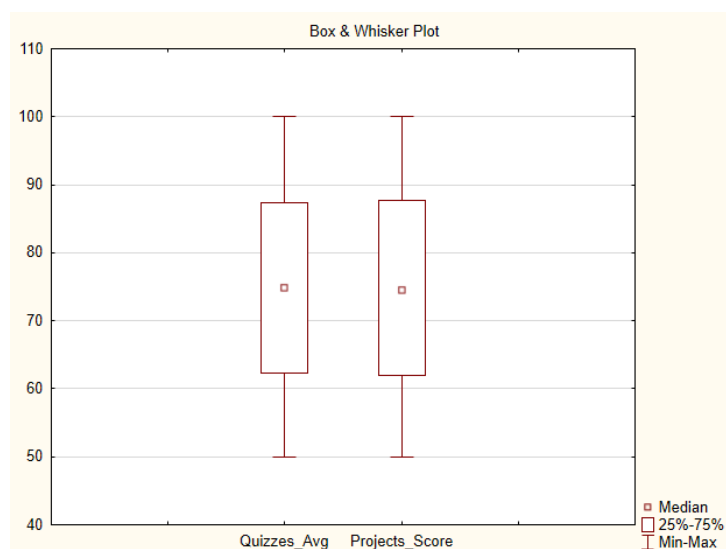
W obu przypadkach hipoteza o normalności rozkładu zostaje odrzucona. Rozkłady obu zmiennych są istotnie nienormalne, co potwierdza również test Kołmogorowa–Smirnowa z poprawką Lillieforsa ($p < 0,01$).

Podsumowanie

Odrzucamy hipotezę o rozkładzie normalnym tych zmiennych dokonujemy testu nieparametrycznego.

3.1.3 Badanie hipotezy

Dokonamy teraz zbadania hipotezy za pomocą testu: Comparing 2 Dependent Samples (variables).



Rysunek 21: Porównanie rozkładów wyników z quizów (*Assignments_Avg*) i projektu (*Projects_Score*) – diagram pudełkowy

Z wykresu 21 wynika, iż możemy przypuszczać, że hipoteza zerowa jest prawdziwa.

Wyniki i interpretacja (przy $\alpha = 0,05$): Oba nieparametryczne testy porównujące wyniki z quizów/ćwiczeń (*Assignments_Avg*) z wynikiem

Tabela 8: Porównanie wyników z quizów i projektu przy użyciu testów nieparametrycznych ($N = 4997$)

Test	Statystyka	p
Test znaków (Sign Test)	$Z = 1,103$	$p = 0,2698$
Test Wilcoxona dla par (Matched Pairs)	$T = 6\,175\,279, Z = 0,671$	$p = 0,5020$

samego projektu (*Projects_Score*) wskazują na brak istotnych różnic między tymi dwiema zmiennymi: - test znaków: $Z = 1,103$, $p = 0,2698$, - test Wilcoxona dla par: $Z = 0,671$, $p = 0,5020$.

W obu przypadkach $p > 0,05$, dlatego nie ma podstaw do odrzucenia hipotezy zerowej.

Wniosek: Nie ma podstaw do odrzucenia hipotezy H_0 .

3.2 Porównanie poziomu stresu z względem płci

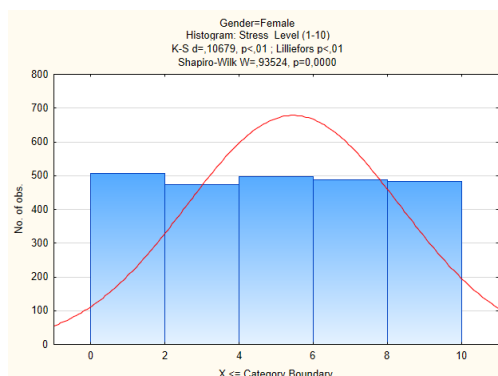
3.2.1 Hipotezy badawcze

Celem analizy jest ustalenie, czy poziom stresu deklarowany przez studentów różni się w zależności od płci. Ze względu na porządkowy charakter zmiennej *Stress_Level* oraz planowane zastosowanie testów nieparametrycznych, sformułowano hipotezy dotyczące identyczności rozkładów:

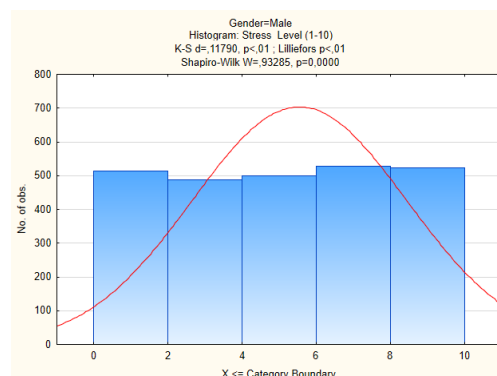
$H_0 : F_{\text{male}}(x) = F_{\text{female}}(x)$ (rozkłady poziomu stresu u obu płci są identyczne) $H_1 : F_{\text{male}}(x) \neq F_{\text{female}}(x)$ (rozkłady poziomu stresu u obu płci różnią się)
--

3.2.2 Sprawdzenie czy zmienna *Stress_Level* (1-10) ma rozkład normalny w podziale na grupy

Sprawdźmy, czy zmienne mają rozkład normalny przy współczynniku istotności $\alpha = 0,05$.



(a) Rozkład zmiennej *Stress_Level* (1-10) dla mężczyzn



(b) Rozkład zmiennej *Stress_Level* (1-10) dla kobiet

Rysunek 22: Porównanie rozkładów zmiennej *Stress_Level* (1-10) z rozkładem normalnym dla dwóch grup

Analizując rysunek 22 widzimy, że niezależnie od płci wykres poziomu stresu nie przypomina rozkładu normalnego. Rozkład wygląda bardziej na jednostajny.

Po dokonaniu testów:

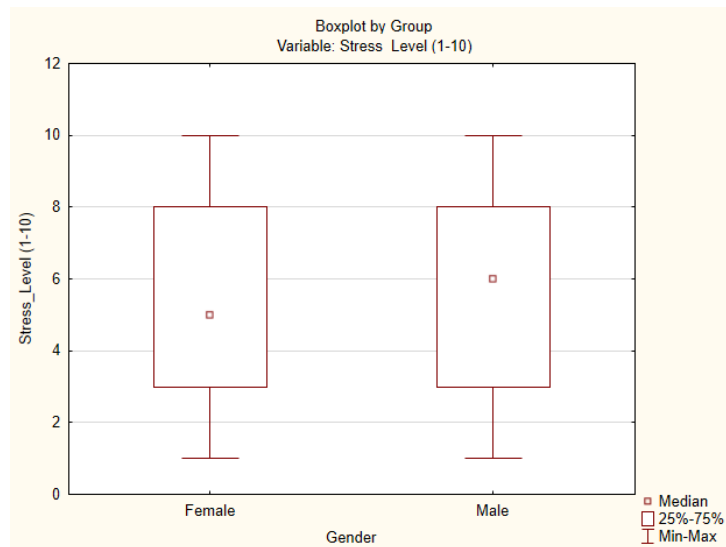
Tabela 9: Wyniki testów normalności rozkładu dla zmiennej *Stress_Level*

Grupa	Kolmogorov-Smirnov	Lilliefors (d, p)	Shapiro-Wilk (W, p)
Mężczyźni	$d = 0,1179, p < 0,01$		$W = 0,9329, p < 0,001$
Kobiety	$d = 0,1068, p < 0,01$		$W = 0,9352, p < 0,001$

Z tabeli 9 widzimy, że niezależnie od testu – Kołmogorowa–Smirnowa z poprawką Lillieforsa ani testu Shapiro–Wilka – zmienna nie ma rozkładu normalnego, gdyż $p < 0,05$. Zatem odrzucamy hipotezę o rozkładzie normalnym w tych grupach. Zatem możemy dokonać testu nieparametrycznego..

3.2.3 Badanie hipotezy

Teraz rozpoczniemy próby potwierdzenia naszej hipotezy H_0 . Wykonamy test Wald–Wolfowitza, Kołmogorowa–Smirnowa oraz Manna–Whitneya. Ale najpierw przeanalizujemy wykres pudełkowy:



Rysunek 23: Wykresu pudełkowy dla zmiennej Stress_Level_(1-10) względem płci

Analizując rysunek ?? możemy wstępnie potwierdzić naszą hipotezę zerową o braku różnic między wartościami dla obu płci.

Wykonanie testów:

Tabela 10: Wyniki testów nieparametrycznych dla porównania grup według zmiennej Płeć

Test	Wynik statystyczny ($Z/D, p$)
Mann-Whitney U	$Z = -1,054, p = 0,292$
Wald-Wolfowitz (Serie)	$Z = 0,157, p = 0,875$
Kolmogorov-Smirnov	$D = 0,025, p > 0,10$

Patrząc na wyniki testów w tabeli 10 widzimy, że dla każdego z trzech rodzajów testów $p > 0,05$. Oznacza to, że nie ma podstaw do odrzucenia hipotezy H_0 .

Wniosek: Nie ma podstaw do odrzucenia hipotezy H_0 .

3.3 Porównanie wyników z połowy semestru (Mid_Term) z wynikiem z finalnego egzaminu (Final_Score)

3.3.1 Hipotezy badawcze

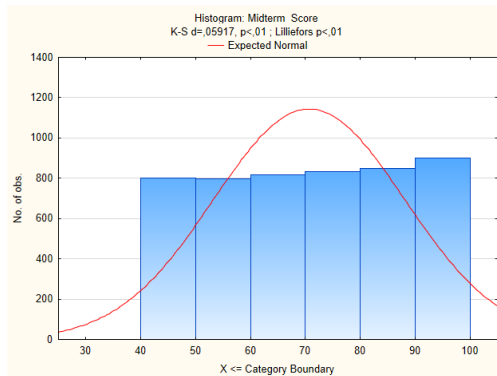
Celem analizy jest ustalenie, czy wyniki uzyskiwane przez studentów w połowie semestru (*Mid_Term*) oraz na egzaminie końcowym (*Final_Score*) mają ten sam rozkład.

Sformułowano następujące hipotezy:

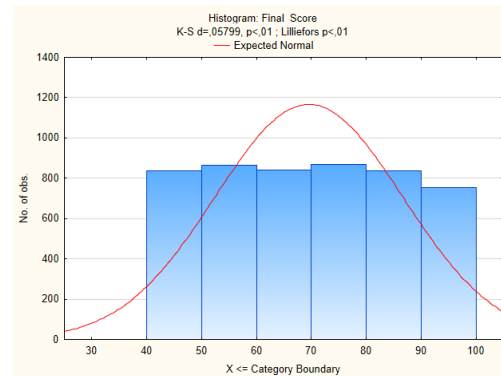
$$\begin{aligned} H_0 : F_{\text{Mid_Term}}(x) &= F_{\text{Final_Score}}(x) \quad (\text{rozkłady wyników są identyczne}) \\ H_1 : F_{\text{Mid_Term}}(x) &\neq F_{\text{Final_Score}}(x) \quad (\text{rozkłady wyników różnią się}) \end{aligned}$$

3.3.2 Sprawdzenie czy zmienne Mid_term oraz Final_Score mają rozkład normalny

Przed rozpoczęciem testu sprawdzimy, czy nasze zmienne mają rozkład normalny. Najpierw sprawdzimy to na wykresie z wizualizacją rozkładu zmiennych.



(a) Rozkład zmiennej *Midterm_Score* (wynik połówkowy)



(b) Rozkład zmiennej *Final_Score* (finalny egzamin)

Rysunek 24: Porównanie rozkładów zmiennych *Midterm_Score* oraz *Final_Score*

Uważnie przyglądając się wykresom umieszczonym na rysunku 24 można stwierdzić, iż zmienne nie posiadają rozkładu normalnego, a przeprowadzenie testu jest tylko i wyłącznie formalnością. Obydwie zmienne posiadają rozkład podobny do jednostajnego.

Przeprowadzenie wyników:

Tabela 11: Wyniki testów normalności dla zmiennych Mid_Term oraz Final_Score

Zmienna	Kołmogorow-Smirnow (poprawka Lillieforsa)		Shapiro-Wilk	
	D	Istotność p	W	p
Mid_Term	0,05917	$< 0,01$	0,95429	0,000
Final_Score	0,05799	$< 0,01$	0,95745	0,000

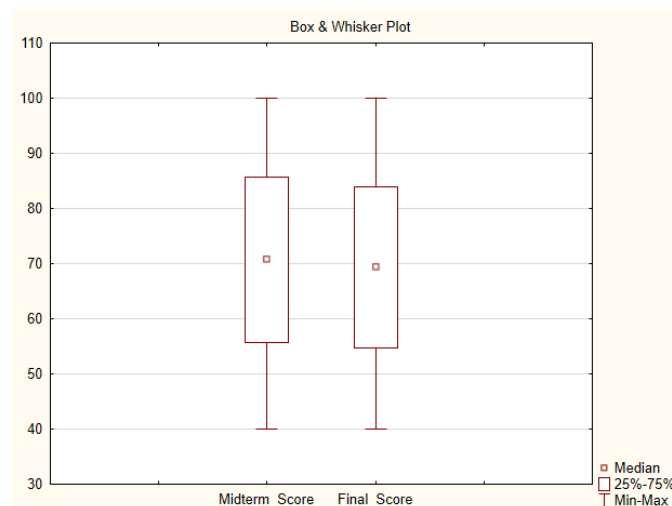
Analizując tabelę 11 obydwie testy – Kołmogorowa–Smirnowa (z poprawką Lillieforsa) oraz Shapiro–Wilka – otrzymały wyniki parametru p poniżej poziomu istotności $\alpha = 0,05$. W przypadku testu Kołmogorowa–Smirnowa p jest mniejsze od 0,01, a przy teście Shapiro–Wilka prawie równe 0.

Wniosek: Odrzucamy hipotezę o rozkładzie normalnym tych dwóch zmiennych. Zostanie przeprowadzony test nieparametryczny.

3.3.3 Badanie hipotezy i przeprowadzenie testu

Teraz wykonamy analizę nieparametryczną porównania dwóch zależnych próbek (jedna osoba pisała te dwa sprawdziany).

Przed tym przeanalizujemy wykres pudełkowy:



Rysunek 25: Wykresu pudełkowy dla zmiennych MidTerm_Score oraz Final_Score

Analizując wykres 25 możemy, zauważyć odchylenie 3 kwartyła w zmiennej Final_Score. Sama mediana testu wynosi 70.86 dla Midterm_Score oraz 69.485 dla Final_Score. Jest to różnica równa dwóch punktą, co jest znacz-

nie większą różnicą niż w poprzednich testach. Aby potwierdzić(lub odrzucić) hipotezę H_0 , potrzebne będzie przeprowadzenie testu.

Po przeprowadzeniu testów znaków oraz kolejności par Wilcoxon'a:

Tabela 12: Wyniki testów nieparametrycznych dla porównania zmiennych Midterm_Score i Final_Score (próby zależne)

Test	N ważnych	Statystyka Z	p	Decyzja
Test kolejności par Wilcoxona	4999	3,422	0,0006	Odrzucenie H_0
Test znaków	4999	3,281	0,0010	Odrzucenie H_0

Uwaga. Wyniki istotne statystycznie ($p < 0,05$) oznaczono kolorem czerwonym.

Analizując tabelkę 12 można zauważyć, iż wyniki parametru p dla testu znaków oraz testu kolejności par Wilcoxona są mniejsze od poziomu istotności $\alpha = 0,05$. Jest to odpowiednio 0,001 oraz 0,0006.

Wniosek: Odrzucamy hipotezę H_0 na rzecz hipotezy H_1 , która zakłada, że rozkład dwóch zmiennych jest różny.

4 Nieparametryczna ANOVA

4.1 Sprawdzenie zależności między poziomem dochodów w rodzinie a oceną końcową

4.1.1 Hipotezy badawcze

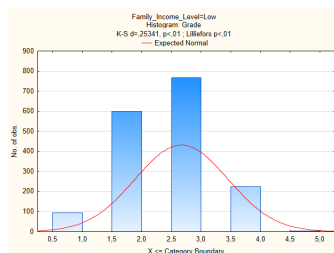
Celem analizy jest zweryfikowanie, czy poziom dochodów w rodzinie studenta różnicuje ocenę końcową (**Grade**). Ze względu na planowane użycie testów nieparametrycznych (ANOVA rang Kruskala-Wallisa), hipotezy dotyczą porównania rozkładów i median w trzech grupach dochodowych.

Sformułowano następujące hipotezy statystyczne:

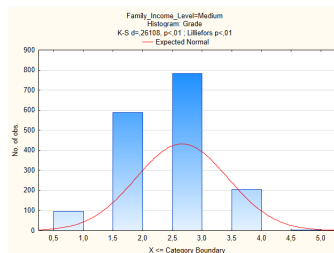
$$\begin{aligned}
 H_0 : F_{\text{Low}}(x) &= F_{\text{Medium}}(x) = F_{\text{High}}(x) \\
 &\text{(rozkłady ocen są identyczne we wszystkich grupach)} \\
 H_1 : &\text{istnieje co najmniej jedna para grup, dla której} \\
 &F_i(x) \neq F_j(x)
 \end{aligned}$$

4.1.2 Sprawdzenie czy finalna ocena ma rozkład normalny

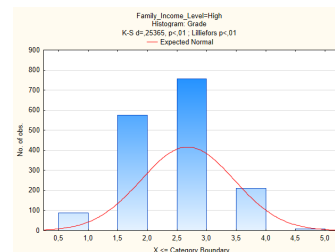
Z rysunku 26 jasno wynika, że żadna ze zmiennych nie ma rozkładu normalnego. Aby się upewnić, dokonajmy testów normalności.



(a) Niski dochód rodziny



(b) Średni dochód rodziny



(c) Wysoki dochód rodziny

Rysunek 26: Rozkłady zmiennej **Grade** (ocena końcowa z przedmiotu) w trzech grupach dochodu rodziny z nałożonym rozkładem normalnym oraz wykresami Q-Q

Z rysunku 26 jasno wynika, że żadna ze zmiennych nie ma rozkładu normalnego. Aby się upewnić, dokonajmy testów normalności.

Tabela 13: Testy normalności rozkładu oceny końcowej (**Grade**) w podziale na poziom dochodów rodziny

Grupa dochodów rodziny	Kolmogorov-Smirnov	Lilliefors (p)	Shapiro-Wilk (W , p)
Wszystkie grupy łącznie	$d = 0,2561$	$p < 0,01$	$W = 0,8610$, $p < 0,001$
Niski dochód	$d = 0,2534$	$p < 0,01$	$W = 0,8600$, $p < 0,001$
Średni dochód	$d = 0,2611$	$p < 0,01$	$W = 0,8584$, $p < 0,001$
Wysoki dochód	$d = 0,2537$	$p < 0,01$	$W = 0,8634$, $p < 0,001$

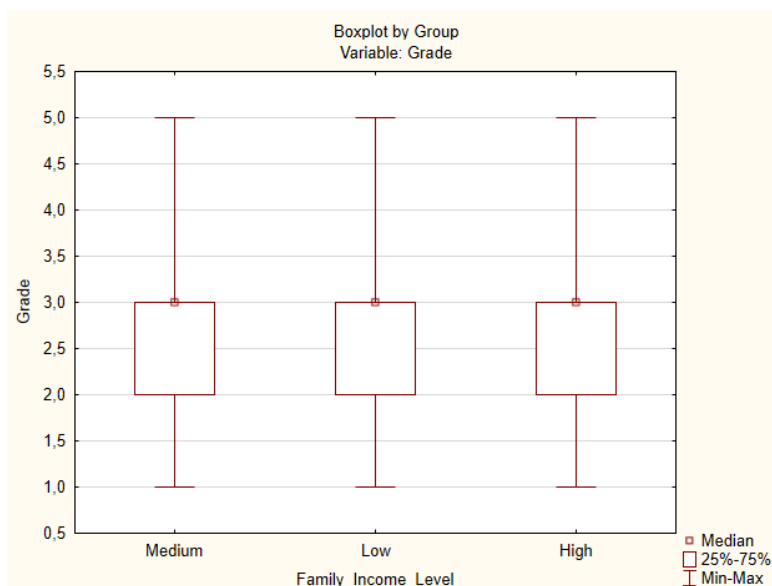
Interpretacja (przy $\alpha = 0,05$):

We wszystkich trzech grupach dochodów rodziny rozkład oceny końcowej (**Grade**) nie ma rozkładu normalnego. Wszystkie testy (Shapiro–Wilka oraz Kołmogorowa–Smirnowa z poprawką Lillieforsa) odrzucają hipotezę zerową o normalności z wartościami parametru $p < 0,001$. Wartości statystyki Shapiro–Wilka wahają się zaledwie od 0,858 do 0,863 – są to wyniki, które wskazują na odstępstwa od rozkładu normalnego.

Wniosek Rozkład nie jest normalny, należy wykonać test nieparametryczny.

4.1.3 Badanie hipotezy

Badając hipotezy dokonamy testu (Comparing multiple independent samples (groups))



Rysunek 27: Diagram pudełkowy oceny końcowej (**Grade**) w podziale na poziom dochodów rodziny (Low / Medium / High)

Analizując wykres pudełkowy 27 możemy wnioskować, iż istnieje duże prawdopodobieństwo potwierdzenia hipotezy H_0 .

Tabela 14: Nieparametryczne porównanie oceny końcowej (**Grade**) między trzema poziomami dochodów rodziny ($N = 5000$)

Test	Statystyka	df	p
Kruskal-Wallis (ANOVA rangowa)	$H = 0,319$	2	0,8524
Test median (Chi-kwadrat)	$\chi^2 = 0,955$	2	0,6203

Wyniki i interpretacja (przy $\alpha = 0,05$): - Test Kruskala-Wallisa: $H(2, N = 5000) = 0,319$, $p = 0,8524$ - Test median: $\chi^2(2) = 0,955$, $p = 0,6203$

W obu testach wartości p są wyraźnie wyższe od poziomu istotności 0,05.

Wniosek końcowy: Nie ma podstaw do odrzucenia hipotezy H_0 .

5 Parametryczna ANOVA

5.1 Wpływ godzin snu na wynik końcowy

5.1.1 Hipotezy

Celem analizy jest ustalenie, czy liczba godzin snu wpływa na ocenę końcową (*Total_Score*).

W związku z tym sformułowano dwie hipotezy:

H_0 : rozkłady i średnie zmiennej *Total_Score* są identyczne we wszystkich grupach liczby godzin snu (4–9 h)

H_1 : istnieje co najmniej jedna para grup liczby godzin snu, w której rozkłady i zmienne zmiennej *Total_Score* różnią się

$H_0 : \mu_4 = \mu_5 = \mu_6 = \mu_7 = \mu_8 = \mu_9$
 $H_1 : \exists_{i,j} : \mu_i \neq \mu_j$ (nie wszystkie średnie są równe)

gdzie:

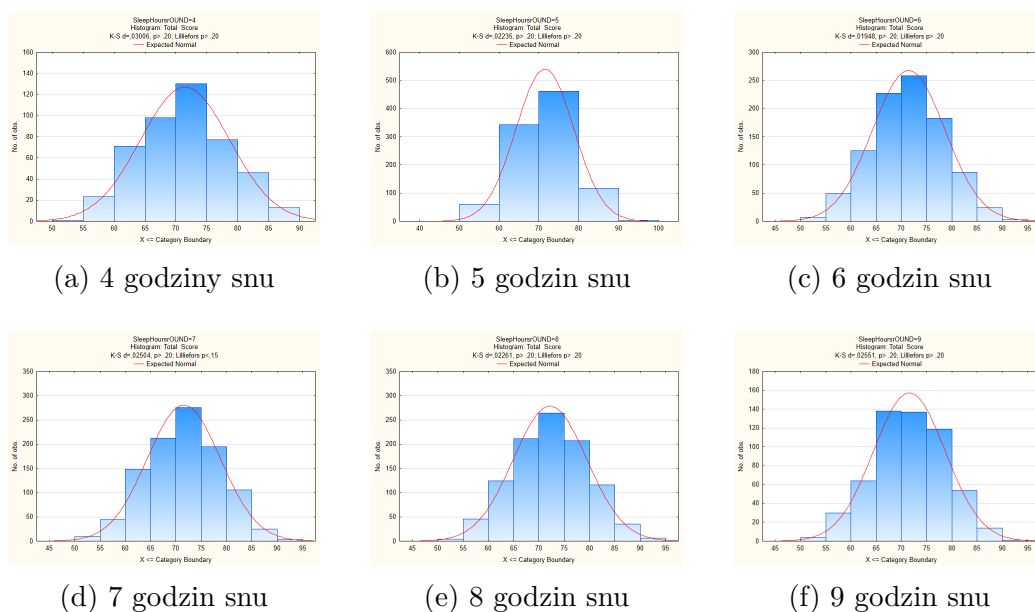
- μ_i – średnia wartość *Total_Score* w grupie osób śpiących dokładnie i godzin na dobę ($i = 4, 5, 6, 7, 8, 9$).

5.1.2 Sprawdzenie czy *Total_Score* ma rozkład normalny względem grup

Tabela 15: Testy normalności rozkładu oceny końcowej projektu (*Total_Score*) w zależności od liczby godzin snu

H snu	n	Kolmogorov-Smirnov (d, p)	Lilliefors (p)
4 h	120–150	d = 0,03006, $p > 0,20$	$p > 0,20$
5 h	300–350	d = 0,02235, $p > 0,20$	$p > 0,20$
6 h	450–500	d = 0,01948, $p > 0,20$	$p > 0,20$
7 h	400–450	d = 0,02504, $p > 0,20$	$p < 0,15$
8 h	250–300	d = 0,02261, $p > 0,20$	$p > 0,20$
9 h	80–100	d = 0,02551, $p > 0,20$	$p > 0,20$

Wyniki wszystkich testów normalności jasno wskazują na $p > 0,05$. Oznacza to, że nie ma podstaw do odrzucenia hipotezy o rozkładzie normalnym wewnątrz grup.



Rysunek 28: Wyniki dla różnych długości snu (4–9 godzin)

Wniosek: Rozkład zmiennej Total_Score **względem grup** jest normalny. Zatem wykonamy test parametryczny.

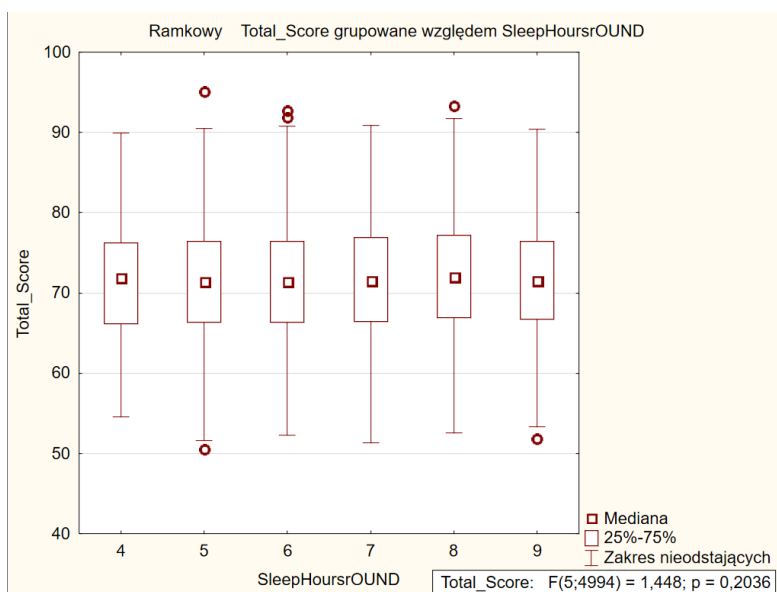
5.1.3 Sprawdzenie jednorodności wariancji względem grup

Na rysunku 29 możemy zauważyć, że wyniki osiągane przez studentów z różnym poziomem snu są zbliżone do siebie. Co ciekawe, najwyższa mediana była w grupie studentów śpiących po 8 godzin na dobę, a większa ilość przespanych godzin niekoniecznie przekładała się na lepsze wyniki w nauce. Teraz dokonamy analizy jednorodności wariancji w grupach:

Tabela 16: Wyniki testów jednorodności wariancji dla zmiennej Total_Score w grupach w zależności od poziomu snu:

Test	Statystyka F	p
Test Levene'a	0,081	0,995
Test Browna-Forsythe'a	0,078	0,996

Jak widzimy w tabeli nie mamy podstaw do odrzucenia hipotezy o równości wariancji.



Rysunek 29: Diagram pudełkowy wyniku końcowego (**Total Score**) w podziale na grupy o różnym poziomie snu

5.1.4 Badanie hipotezy

Aby zbadać dokładnie różnicę między tymi sześcioma grupami, przeprowadzimy test ANOVA parametryczny, ponieważ wcześniej zbadaliśmy, że nie ma podstaw do odrzucenia hipotezy o rozkładzie normalnym w żadnej z grup oraz analiza jednorodności wariancji nie wykazała istotnie różnej wariancji w grupach.

Tabela 17: ANOVA - porównanie wyniku końcowego (**Total Score**) między grupami z różnym poziomem snu ($N = 5000$)

Test	Statystyka	df	p
ANOVA	$F = 1,448$	5	0,204

Wyniki i interpretacja (przy $\alpha = 0,05$): - ANOVA: $F = 1,448$, $p = 0,204$

Uzyskane p jest wyższe od poziomu istotności 0,05.

Wniosek końcowy: Nie ma podstaw do odrzucenia hipotezy H_0 o równości średnich w grupach. Czyli różnice w wynikach końcowych osiąganych przez studentów, którzy spali różną ilość godzin, nie są statystycznie istotne, co może rodzić hipotezę, że sen nie ma aż tak dużego wpływu na wynik końcowy, przynajmniej statystycznie.

6 Regresja liniowa

6.1 Sprawdzanie w jaki sposób wpływają poszczególne czynniki na ocenę końcową

		Regression Summary for Dependent Variable: Grade (Projekt)					
		R= ,92775670 R2= ,86073250 Adjusted R2= ,86050927					
		F(8,4991)=3855,8 p<0,0000 Std.Error of estimate: ,29100					
N=5000		b*	Std.Err. of b*	b	Std.Err. of b	t(4991)	p-value
Intercept				-4,51782	0,049218	-91,7913	0,000000
Midterm_Score		0,336398	0,005288	0,01503	0,000236	63,6190	0,000000
Final_Score		0,546872	0,005284	0,02490	0,000241	103,4979	0,000000
Assignments_Avg		0,285253	0,005288	0,01543	0,000286	53,9413	0,000000
Quizzes_Avg		0,185856	0,005286	0,01004	0,000286	35,1601	0,000000
Participation_Score		0,193673	0,005286	0,00521	0,000142	36,6393	0,000000
Projects_Score		0,554327	0,005286	0,02970	0,000283	104,8633	0,000000
Study_Hours_per_Week		0,004079	0,005285	0,00044	0,000572	0,7718	0,440265
Sleep_Hours_per_Night		-0,000971	0,005285	-0,00052	0,002847	-0,1838	0,854167

Rysunek 30: Wyniki pierwszej regresji liniowej uwzględniającej wszystkie zmienne wejściowe.

Z tej regresji (rysunek: 30) wynika, że ilość godzin przepracowanych oraz ilość godzin przespanych nie są istotne statystycznie.

	Variables currently in the Equation; DV: Grade (Projekt)						
Variable	b* in	Partial Cor.	Semipart Cor.	Tolerance	R-square	t(4991)	p-value
Midterm_Score	0.336398	0.669178	0.336061	0.997996	0.002004	63.6190	0.000000
Final_Score	0.546872	0.825929	0.546717	0.999431	0.000569	103.4979	0.000000
Assignments_Avg	0.285253	0.606862	0.284939	0.997801	0.002199	53.9413	0.000000
Quizzes_Avg	0.185856	0.445556	0.185730	0.998640	0.001360	35.1601	0.000000
Participation_Score	0.193673	0.460391	0.193543	0.998661	0.001339	36.6393	0.000000
Projects_Score	0.554327	0.829346	0.553930	0.998568	0.001432	104.8633	0.000000
Study_Hours_per_Week	0.004079	0.010924	0.004077	0.998976	0.001024	0.7718	0.440265
Sleep_Hours_per_Night	-0.000971	-0.002602	-0.000971	0.998976	0.001024	-0.1838	0.854167

Rysunek 31: Macierz korelacji dla wszystkich pierwotnych zmiennych.

A ponadto, jak wskazuje macierz korelacji (rysunek: 31), korelacja tych czynników z oceną końcową jest bardzo niska, dlatego należy usunąć te czynniki i przeprowadzić kolejną regresję.

6.1.1 Poprawiona Regresja

N=5000	Regression Summary for Dependent Variable: Grade (PurgedData)					
	R= ,92774725 R2= ,86071495 Adjusted R2= ,86054758 F(6,4993)=5142,4 p<0,0000 Std.Error of estimate: ,29096					
	b*	Std.Err. of b*	b	Std.Err. of b	t(4993)	p-value
Intercept			-4,51275	0,044777	-100,782	0,00
Midterm_Score	0,336417	0,005286	0,01503	0,000236	63,641	0,00
Final_Score	0,546881	0,005283	0,02491	0,000241	103,521	0,00
Assignments_Avg	0,285190	0,005286	0,01543	0,000286	53,948	0,00
Quizzes_Avg	0,185737	0,005283	0,01003	0,000285	35,159	0,00
Participation_Score	0,193628	0,005285	0,00520	0,000142	36,639	0,00
Projects_Score	0,554320	0,005285	0,02970	0,000283	104,877	0,00

Rysunek 32: Statystyki poprawionego modelu regresji po usunięciu nieistotnych zmiennych.

Jak możemy zobaczyć na wykresie 32, wszystkie składowe są istotne statystycznie, ponieważ współczynnik p jest mniejszy od 0,05. Ponadto z tej regresji wynika, że wzór na ocenę końcową to:

$$\begin{aligned}
 \text{Grade} = & -4,51275 \\
 & + 0,015 \times \text{Midterm Score} \\
 & + 0,025 \times \text{Final Score} \\
 & + 0,015 \times \text{Assignments Score} \\
 & + 0,01 \times \text{Quizzes Score} \\
 & + 0,005 \times \text{Participation Score} \\
 & + 0,03 \times \text{Projects Score}
 \end{aligned}$$

Z tego wzoru oraz ze zmiennych b^* wynika, że najbardziej wpływającymi na ocenę elementami są: oceny z egzaminu końcowego oraz z projektów, następnie zadania, kolokwium z połowy materiału, a na końcu quizy i obecność. R^2 , czyli współczynnik determinacji, wynosi 0,86, co oznacza, że ten wzór wyjaśnia 86% zmienności oceny końcowej.

Variable	Variables currently in the Equation: DV: Grade (Projekt)						
	b* in	Partial Cor.	Semipart Cor.	Tolerance	R-square	t(4993)	p-value
Midterm_Score	0,336417	0,669233	0,336132	0,998307	0,001693	63,6412	0,00
Final_Score	0,546881	0,825936	0,546765	0,999574	0,000426	103,5212	0,00
Assignments_Avg	0,285190	0,606835	0,284938	0,998229	0,001771	53,9484	0,00
Quizzes_Avg	0,185737	0,445477	0,185700	0,999606	0,000394	35,1593	0,00
Participation_Score	0,193628	0,460313	0,193513	0,998822	0,001178	36,6387	0,00
Projects_Score	0,554320	0,829328	0,553926	0,998577	0,001423	104,8769	0,00

Rysunek 33: Korelacje parametrów wchodzących w skład poprawionego modelu regresji.

I tak samo w przypadku korelacji (rysunek: 33) – najniższa wartość wynosi około 0,45, więc można uznać, że wszystkie parametry są skorelowane z oceną końcową.

6.1.2 Analiza reszt

Cook's Distances: Grade (Projekt)									
Sorted									
Case	Observed Value	Predicted Value	Residual	Standard Pred. v.	Standard Residual	Std. Err.	Mahalanobis Distance	Deleted Residual	Cook's Distance
4117	4,000000	3,527461	0,472539	1,18565	1,62405	0,015578	13,33013	0,473897	0,001086
1482	2,000000	1,509526	0,490474	-1,60595	1,68569	0,014468	11,36105	0,491690	0,001009
541	3,000000	2,534366	0,465634	-0,18819	1,60032	0,015122	12,50366	0,466895	0,000994
3890	1,000000	1,485565	-0,485565	-1,63909	-1,66882	0,014405	11,25349	-0,486758	0,000980
4958	3,000000	2,537774	0,462226	-0,18347	1,58861	0,015056	12,38510	0,463467	0,000970
848	2,000000	2,497672	-0,497672	-0,23895	-1,71043	0,013958	10,50511	-0,498820	0,000966
4436	1,000000	1,509530	-0,509530	-1,60667	-1,74190	0,013969	10,03345	-0,507951	0,000861
3151	2,000000	1,516278	0,483722	-1,59660	1,66249	0,014282	11,04535	0,484890	0,000956
2268	3,000000	3,477513	-0,477513	1,11655	-1,64115	0,014456	11,34027	-0,478994	0,000955
1677	2,000000	2,550593	-0,550593	-0,22868	-1,73594	0,013526	9,80332	-0,558167	0,000634
3391	4,000000	4,509141	-0,509141	2,54370	-1,74985	0,013394	9,59374	-0,510222	0,000531
244	2,000000	1,516796	0,483204	-1,59589	1,66071	0,014065	10,68214	0,484336	0,000925
1156	1,000000	0,572038	0,427962	-2,90286	1,47085	0,015864	13,86037	0,429238	0,000924
4636	3,000000	2,538856	0,461144	-0,18198	1,58489	0,014641	11,62753	0,462215	0,000912
4633	2,000000	1,510389	0,489631	-1,60478	1,68279	0,013594	9,91231	0,490702	0,000887
2120	1,000000	1,474017	-0,474017	-1,65507	-1,62913	0,013992	10,56019	-0,475115	0,000881
2301	2,000000	2,454573	-0,454573	-0,28857	-1,56231	0,014580	11,55228	-0,455717	0,000880

Rysunek 34: Wykres odległości Cooka służący do identyfikacji obserwacji wpływowych.

Jak można zauważyć na rysunku 34, w naszym zbiorze danych najwyższy współczynnik Cooka wynosi 0,001, czyli ten przypadek ma bardzo mały wpływ na regresję liniową.

Mahalanobis distances: Grade (Projekt)									
Sorted									
Case	Observed Value	Predicted Value	Residual	Standard Pred. v.	Standard Residual	Std. Err.	Mahalanobis Distance	Deleted Residual	Cook's Distance
2519	1,037353	-0,037353	-2,25915	-0,12838	0,016019	14,15185	-0,037466	0,000007	
4450	2,646986	0,353094	-0,03250	1,21353	0,015945	14,07251	0,354517	0,000636	
1156	0,572038	0,427962	-2,90286	1,47085	0,015864	13,86037	0,429238	0,000924	
4117	3,527461	0,472539	1,18565	1,62405	0,015578	13,33013	0,473897	0,001086	
462	2,696632	0,393368	-0,08822	1,35195	0,015501	13,18845	0,394467	0,000745	
845	3,667112	0,332888	1,37884	1,14409	0,015408	13,01918	0,333824	0,000527	
708	1,847141	0,152859	-1,13889	0,52535	0,015396	12,99707	0,153288	0,000111	
4119	4,305646	-0,305646	2,26219	-1,05047	0,015389	12,98404	-0,306505	0,000443	
1105	5,021163	-0,021163	3,25203	-0,07274	0,015251	12,73363	-0,021222	0,000002	
541	2,534366	0,465634	-0,18819	1,60032	0,015122	12,50366	0,466895	0,000994	
2109	4,852264	0,147736	3,01838	0,50775	0,015079	12,42556	0,148134	0,000099	
4958	2,537774	0,462226	-0,18347	1,58861	0,015056	12,38510	0,463467	0,000970	
1779	3,215466	-0,215467	0,75404	-0,74053	0,014986	12,26214	-0,216040	0,000209	
3128	1,989255	0,010745	-0,94229	0,03693	0,014961	12,21627	0,010773	0,000001	
3550	3,212737	-0,212737	0,75026	-0,73115	0,014922	12,14854	-0,213298	0,000202	
3042	1,661571	0,338429	-1,39561	1,16313	0,014903	12,11487	0,339319	0,000510	
1679	1,796710	0,203290	-1,20866	0,69868	0,014893	12,09779	0,203824	0,000184	

Rysunek 35: Wykres odległości Mahalanobisa dla analizowanych przypadków.

W przypadku odległości Mahalanobisa (rysunek: 35) największa wartość wynosi 14,15, co oznacza odległość danego studenta od centrum rozkładu

prawdopodobieństwa.

										Standard Residual: Grade (Projekt)									
Standard Residuals										Sorted									
Case	-1,8	1,75	
										Observed Value	Predicted Value	Residual	Standard Pred. v	Standard Residual	Std Err. Pred Val	Mahalanobis Distance	Deleted Residual	Cook's Distance	
866 .*	3,000000	3,518801	-0,518802	1,17367	-1,78305	0,011910	7,37636	-0,519672	0,000764	
3187 .*	3,000000	3,517465	-0,517465	1,17182	-1,77846	0,010705	5,76743	-0,518167	0,000613	
987 .*	2,000000	2,515866	-0,515866	-0,21378	-1,77296	0,010279	5,23960	-0,516511	0,000562	
1924 .*	1,000000	1,514548	-0,514548	-1,59900	-1,76843	0,011544	6,86867	-0,515359	0,000705	
3232 .*	2,000000	2,514261	-0,514261	-0,21600	-1,76744	0,011777	7,19065	-0,515105	0,000734	
2283 .*	2,000000	2,512752	-0,512752	-0,21809	-1,76226	0,011319	6,56498	-0,513529	0,000673	
3914 .*	1,000000	1,512005	-0,512005	-1,60252	-1,75969	0,010267	5,22414	-0,512643	0,000552	
3204 .*	3,000000	3,510508	-0,510508	1,16220	-1,75455	0,011864	7,31169	-0,511359	0,000734	
3117 .*	2,000000	2,510424	-0,510424	-0,22131	-1,75426	0,010203	5,14665	-0,511053	0,000542	
130 .*	3,000000	3,509423	-0,509423	1,16070	-1,75082	0,011529	6,84940	-0,510224	0,000690	
4444 .*	3,000000	3,509357	-0,509357	1,16061	-1,75059	0,012645	8,44146	-0,510321	0,000830	
3391 .*	4,000000	4,509141	-0,509141	2,54370	-1,74985	0,013394	9,59374	-0,510222	0,000931	
2632 .*	2,000000	2,508734	-0,508734	-0,22365	-1,74845	0,011894	7,35423	-0,509586	0,000732	
1974 .*	2,000000	2,508602	-0,508602	-0,22383	-1,74800	0,011328	6,57785	-0,509374	0,000664	
2598	*	3,000000	2,491622	0,508378	-0,24732	1,74723	0,011472	6,77179	0,509170	0,000680	
3801	*	3,000000	2,491819	0,508181	-0,24705	1,74655	0,011593	6,93596	0,508989	0,000694	
4846 .*	2,000000	2,507616	-0,507616	-0,22519	-1,74461	0,010395	5,38077	-0,508264	0,000556	

Rysunek 36: Rozkład reszt standaryzowanych w celu weryfikacji poprawności modelu.

Dla reszt standaryzowanych przedstawionych na rysunku 36 można założyć, że wyniki nie przekraczają wartości 2 oraz -2 , co oznacza, że nie ma żadnych przypadków odstających.

7 Wnioski

Zauważyliśmy między innymi, że płeć, poziom zamożności rodziców, ilość snu i dostęp do internetu nie wpływają statystycznie istotnie na wyniki w nauce. Ocena końcowa natomiast rosła wraz ze wzrostem aktywności studenta, natomiast w kontekście obecności studenci z najlepszymi ocenami notowali najniższą obecność. Testy statystyczne pozwoliły nam również dowiedzieć, że różnica pomiędzy oceną z egzaminu śródsesemestralnego jest istotnie różna od oceny z egzaminu końcowego. Jeśli chodzi o poziom stresu między płciami, to nie ma statystycznie istotnej różnicy między płciami. Wyniki z projektów i z zadań również nie wykazywały statystycznie istotnej różnicy. Najlepsze wyniki w nauce notowali studenci przesypiający zalecane 8 godzin snu na dobę, natomiast umiarkowany stres motywował studentów do działania. Jeśli chodzi o czas poświęcony na naukę, to najwięcej go poświęcili studenci z najwyższą oceną, a następnie studenci z najniższą oceną, chociaż w tym przypadku mogło to świadczyć o nieefektywności ich nauki.

Udało się nam przeanalizować wszystkie najważniejsze cechy zbioru danych. Analiza dotyczyła zarówno wpływu aktywności typowo akademickiej i naukowej, jak i wpływu czynników środowiskowych i zewnętrznych na wyniki osiągnięte przez studentów na koniec semestru. Wykorzystaliśmy również wszystkie znane nam metody statystyczne z odpowiednim ich dopasowaniem do danego przypadku i charakteru badanych zmiennych.

8 Bibliografia

1. Wikipedia, <https://pl.wikipedia.org>
2. StatSoft Polska, Internetowy Podręcznik Statystyki, <https://www.statsoft.pl/textbook/stathome.html>
3. M. Bratijczuk, A. Chydzinski, *Statystyka matematyczna*, Wydawnictwo Politechniki Śląskiej, Gliwice 2012.
4. Grokopedia, <https://grokopedia.pl>
5. Overleaf, *Writefull and Grammar Checker integration*, https://www.overleaf.com/learn/how-to/Writefull_on_Overleaf