Wnioskowanie Statystyczne - Projekt zaliczeniowy

Alicja Kalwat, Łukasz Bielawski
Modelowanie Matematyczne i Analiza Danych,
Uniwersytet Gdański

30.01.2024r.

Wstęp

Do projektu użyjemy zestawu danych zdrowotnych zaczerpniętych z Kaggle'a (Link do strony), który obejmuje różnorodne aspekty życia codziennego osób. Poniżej przedstawiamy krótki opis poszczególnych zmiennych, które zostały zebrane, a które posłużą do dalszej analizy.

Zmienna	Opis
`Person.ID`	Numer identyfikacyjny dla każdej osoby.
`Gender`	Płeć osoby (Male/Female).
`Age`	Wiek osoby w latach.
`Occupation`	Zawód lub profesja osoby.
`Sleep.Duration`	Ilość godzin snu danej osoby dziennie.
`Quality.of.Sleep`	Subiektywna ocena jakości snu, skala 1-10.
`Physical.Activity.Level`	Ilość minut aktywności fizycznej dziennie.
`Stress.Level`	Subiektywna ocena poziomu stresu, skala 1-10.
`BMI.Category`	Kategoria BMI osoby (Normal, Overweight, Obese).
`Blood.Pressure`	Ciśnienie krwi osoby (skurczowe/rozkurczowe).
`Heart.Rate`	Puls osoby, wyrażona w uderzeniach na minutę.
`Daily.Steps`	Ilość kroków, jakie osoba wykonuje dziennie.
`Sleep.Disorder`	Obecność lub brak zaburzeń snu (None, Insomnia, Sleep Apnea).

Celem projektu jest zidentyfikowanie zależności między różnymi zmiennymi zdrowotnymi, co pozwoli na lepsze zrozumienie wpływu różnych czynników na zdrowie i styl życia. Skupimy się analizie powyższych danych uwzględniając trzy hipotezy badawcze.

W dalszej części pracy skupimy się na zależności Ciśnienia Krwi oraz Jakości Snu od innych czynników zdrowotnych. Wydaje się, że obie z tych zmiennych mogą być determinowane przez różnorodne czynniki, takie jak wiek, BMI, ilość snu, czy poziom stresu.

Analiza ma na celu ustalenie, jakie zmienne mają istotny wpływ na zmienne zależne. Zajmiemy się więc wyżej opisanymi wstępnymi pomysłami, a finalne wnioski i głębsze rozważania pojawią się w trakcie dokładniejszego zbadania relacji między poszczególnymi czynnikami zdrowotnymi.

Przygotowanie danych oraz wstępna analiza

Aby przystąpić do analizy, należy odpowiednio przygotować dane.

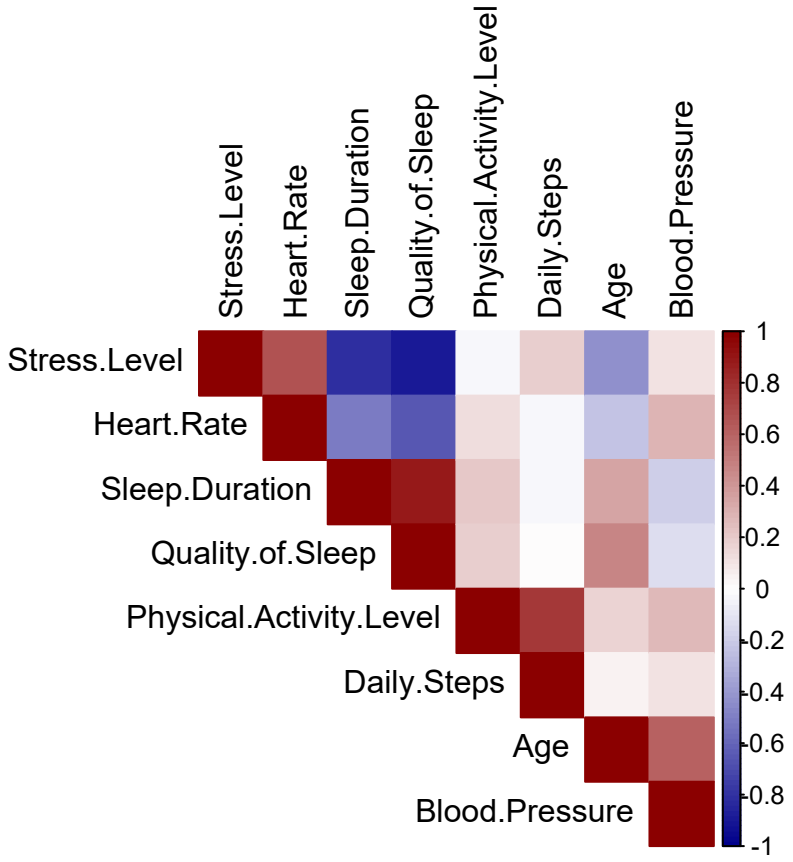
Kolumny, które są zmiennymi kategorycznymi (`Gender`,`Occupation`, `Sleep.Disorder`, a także `BMI.Category`) konwertujemy na faktory.

Kolejno zajmiemy się kolumną `Blood.Pressure`. Przydałoby się, aby miała ona wartości numeryczne, dlatego w dalszej analizie będziemy zajmować się tylko ciśnieniem skurczowym - zostawiamy tylko pierwsze wartości z dwóch wartości.

W następnym kroku zmienimy wartości BMI na 0 - Normal, 1 - Overweight, oraz 2 - Obese, aby ułatwiło to nam dalszą pracę.

- **Korelacje zmiennych numerycznych**

Jako, że część zmiennych w naszych danych jest zmiennymi numerycznymi, możemy za pomocą funkcji `cor()` wyliczyć dla nich korelacje, aby znaleźć między nimi jakieś zależności. Na poniższym wykresie przedstawiono właśnie owe wartości.

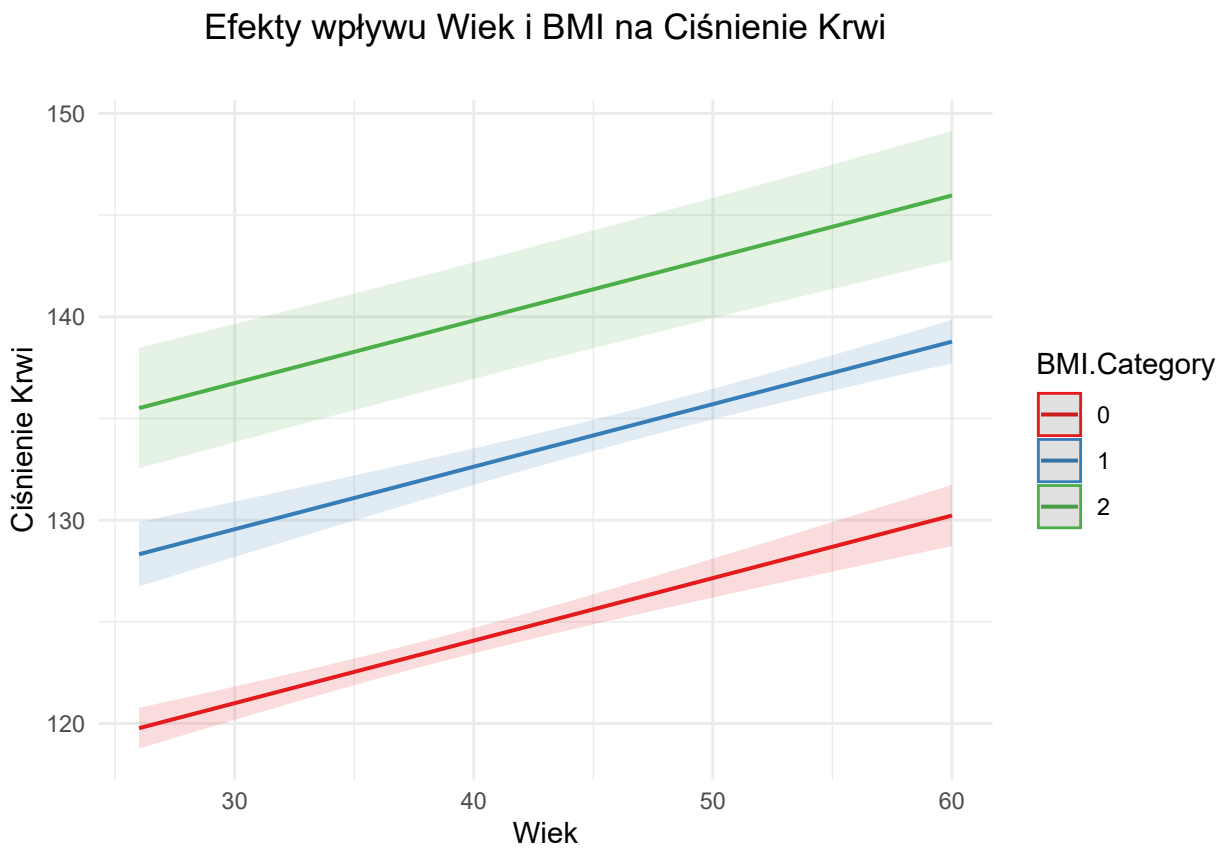


Hipoteza 1: Ciśnienie Krwi zależy od Wieku oraz BMI

Na wykresie korelacji widzimy, że istnieje znacząca korelacja między Ciśnieniem Krwi a Wiekiem. Dodatkowo, porównując średnie wartości Ciśnienia Krwi z BMI (tabela powyżej), także widać tutaj jakąś korelację. Natomiast nie widać, żeby była jakaś korelacja między BMI a Płcią. W przypadku budowy modeli lepsze rezultaty można osiągnąć, gdy zmienne niezależne są słabo lub niezbyt silnie skorelowane między sobą, więc jest to dla nas dobra wiadomość.

BMI.Category	Blood.Pressure_mean	Age_mean
0	123.6065	38.47222
1	135.0541	47.88514
2	139.2000	38.00000

Aby zbadać zależność Ciśnienia Krwi od Wieku i BMI stworzymy model regresji liniowej. W naszym modelu uwzględnimy wiek jako zmienną ciągłą oraz BMI jako zmienną kategoryczną. Poniżej przedstawiono wykres obrazujący zależność `Blood.Pressure` od `Age` oraz `BMI.Category`. Jak widać, dla każdej kategorii BMI im wyższy wiek, tym wyższe jest ciśnienie krwi. Podobną zależność możemy zaobserwować w przypadku kategorii BMI, dla 0 (Normal) mamy najniższe wartości, natomiast dla 1 (Obese) - najwyższe.



Podsumowanie modelu Blood.Pressure ~ Age + BMI.Category + Gender

term	estimate	std.error	statistic	p.value
(Intercept)	111.773848	1.2860519	86.91239	0
Age	0.307563	0.0324245	9.48552	0
BMI.Category1	8.552508	0.5776963	14.80451	0
BMI.Category2	15.738757	1.4869204	10.58480	0

Z podsumowania modelu powyżej możemy odczytać, że:

- Współczynnik dla Wieku wynosi 0.30756, co oznacza, że każdy dodatkowy rok wieku przewiduje wzrost ciśnienia krwi o 0.30756 jednostki, przy założeniu stałej wartości BMI.
- Współczynniki dla BMI.Category1 i BMI.Category2 (8.55251 i 15.73876) oznaczają, że osoby przynależące do tych kategorii mają średnio wyższe ciśnienie krwi w porównaniu do kategorii referencyjnej BMI.Category0 (przy założeniu stałej wartości wieku).

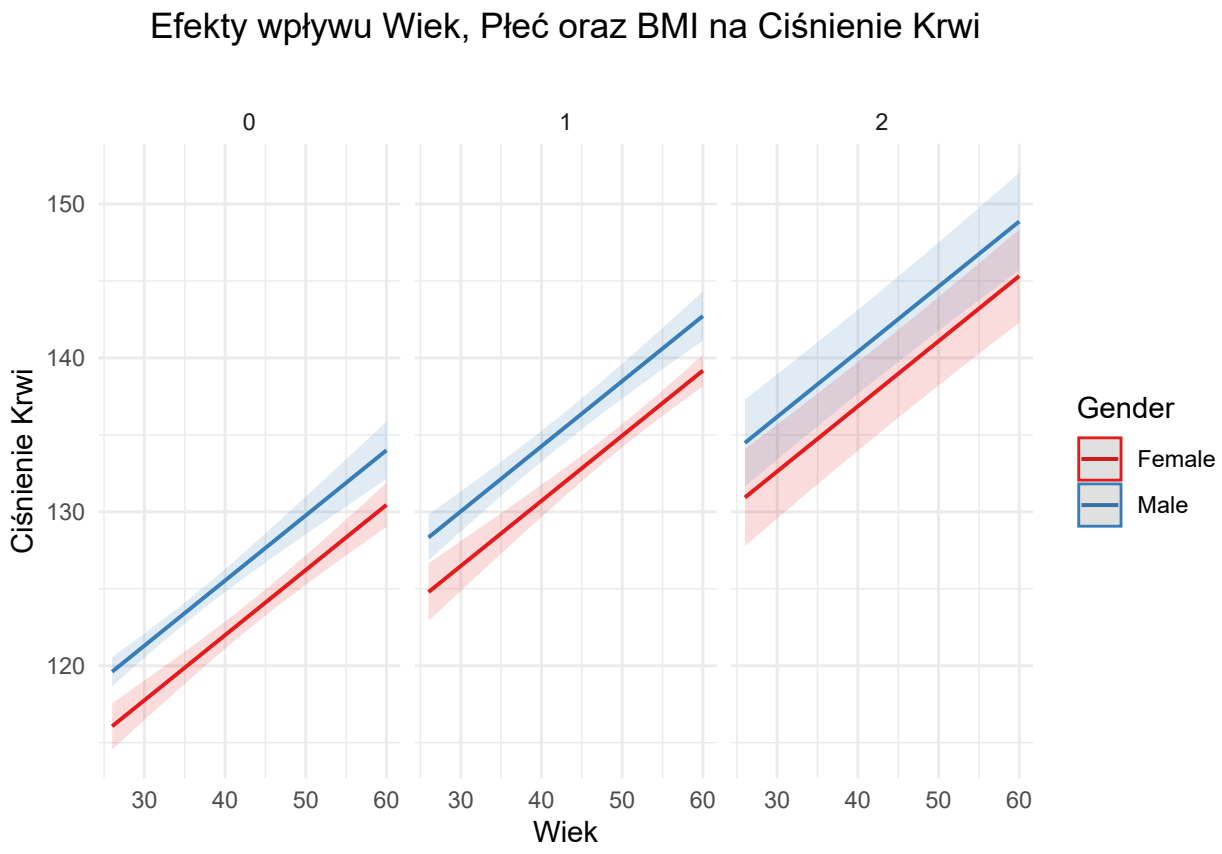
Statystyki t-testu dla współczynników (Estimate, Std. Error, t value, Pr(>|t|)) pozwalają stwierdzić, czy dany współczynnik jest statystycznie istotny. Wszystkie trzy współczynniki Age, BMI.Category1, BMI.Category2 mają bardzo małe wartości p-value, co sugeruje, że są one statystycznie istotne.

Podsumowując, model sugeruje, że wiek, kategoria BMI1 i kategoria BMI2 mają statystycznie istotny wpływ na przewidywane ciśnienie krwi, a model ogółem jest istotny statystycznie. Potwierdza to naszą hipotezę, więc możemy stwierdzić, że wiek i kategoria BMI mają wpływ na ciśnienie krwi.

Podążając za pierwszą hipotezą, która zakładała, że Ciśnienie Krwi zależy od Wieku oraz BMI, postawiliśmy dodatkową hipotezę badawczą:

Hipoteza 1b: Ciśnienie Krwi zależy od Wieku oraz BMI, ale nie zależy od płci.

Ciekawi nas, czy płeć może być zmienną towarzyszącą dla powyższego modelu. Podobnie, jak w przypadku pierwszej hipotezy, użyjemy modelu regresji liniowej, aby zbadać związki między Ciśnieniem Krwi, Wiekiem, BMI i Płcią. Wprowadzimy zmienną płci jako zmienną dodatkową do naszego modelu, aby sprawdzić, czy ma ona istotny wpływ na poziom ciśnienia krwi.



Podsumowanie modelu Blood.Pressure ~ Age + BMI.Category + Gender

term	estimate	std.error	statistic	p.value
(Intercept)	105.0595640	1.6295412	64.47187	0
Age	0.4231979	0.0360020	11.75485	0
GenderMale	3.5460850	0.5677360	6.24601	0
BMI.Category1	8.7232806	0.5508069	15.83728	0
BMI.Category2	14.8674398	1.4228177	10.44929	0

Analizując wyniki modelu regresji liniowej, w którym uwzględniono zmienne Age, Gender, BMI.Category jako predyktory dla zmiennej zależnej Blood.Pressure, możemy wyciągnąć następujące wnioski:

Wpływ Wieku (Age):

- Współczynnik dla GenderMale wynosi 3.5461. Oznacza to, że w porównaniu do płci żeńskiej, płci męskiej przypisuje się średnio wyższe ciśnienie krwi o 3.5461 jednostki, przy założeniu stałych wartości pozostałych zmiennych.

Wyniki więc sugerują, że płci męskiej przypisuje się średnio wyższe ciśnienie krwi, biorąc pod uwagę wiek i kategorię BMI. Należy jednak pamiętać, że analiza statystyczna nie zawsze pozwala na jednoznaczne wnioski przyczynowo-skutkowe. Wpływ na wyniki analizy mogą mieć na przykład różnice w strukturze wiekowej między płciami.

Na poniższym wykresie pudełkowym możemy przyjrzeć się temu, jaki jest rozkład wieku dla kobiet i dla mężczyzn. Widać, że w naszym zbiorze danych duża część kobiet jest starsza niż mężczyźni.

Jeśli wyniki wskazują, że kobiety mają niższe ciśnienie krwi, nawet pomimo przewagi starszych kobiet w badaniu, to może to potwierdzić fakt, że Płeć jest również istotną zmienną w kwestii modelowania Ciśnienia Krwi, więc nasza hipoteza 1b zostaje odrzucona.

Hipoteza 2: Najlepszym modelem predyktującym zmienną Jakość snu jest model zależny od Poziomu stresu i Długości snu.

Analizując diagram korelacji zmiennych można zauważyć, że zmienna Jakości snu posiada parę istotnych korelacji z innymi zmiennymi, lecz najbardziej wyróżniają się korelacje z Poziomem stresu (ok. -0.90) oraz Długością Snu (ok. 0.88). Sugeruje to więc, aby stworzyć model, którego celem będzie opisywanie Jakości Snu za pomocą tych cech.

Podsumowanie modelu Quality.of.Sleep ~ Sleep.Duration + Stress.Level

term	estimate	std.error	statistic	p.value
(Intercept)	4.0592843	0.5660178	7.171655	0
Stress.Level	-0.3407688	0.0273608	-12.454654	0
Sleep.Duration	0.7145435	0.0611790	11.679549	0

Hipotezą jest, że jest to najlepszy model do przewidywania Jakości Snu. Dzielimy więc nasze dane na zbiór treningowy i testowy, uczymy model na zbiorze treningowym, a na koniec sprawdzamy jakość predykcji na zbiorze testowym. Po wielokrotnym sprawdzeniu tego modelu średni wynik prawidłowych wyników to ok. 83%. Żeby sprawdzić, czy jest to najlepszy model, musimy jednak rozważyć inne opcje i dokonać głębszej analizy.

Z diagramu korelacji widzimy, że poza korelacją Jakości Snu ze zmiennymi wyjaśniającymi - Długością snu i Poziomu stresu, istnieje też silna korelacja pomiędzy zmiennymi wyjaśniającymi (ok. -0.81). Podnosi to pytanie, czy aby na pewno obie zmienne są potrzebne w tym modelu, mimo że wyniki okazały się istotne statystycznie. Tworzymy więc model Jakość snu ~ Długość snu * Poziom stresu, aby uwzględnić też zależność pomiędzy zmiennymi wyjaśniającymi.

Podsumowanie modelu Quality.of.Sleep ~ Sleep.Duration * Stress.Level

term	estimate	std.error	statistic	p.value
(Intercept)	1.7133329	0.7105860	2.411155	0.0163887
Sleep.Duration	1.0604210	0.0931431	11.384860	0.0000000
Stress.Level	0.1631108	0.1130279	1.443102	0.1498378
Sleep.Duration:Stress.Level	-0.0762622	0.0162204	-4.701619	0.0000036

Wyniki są intresujące, gdyż okazuję się, że zmienna z Poziomem stresu nie jest istotna dla tego modelu, ale zależność między zmiennymi wyjaśniającymi jest istotna. Czyli sam Poziom stresu nie pomaga nam znacząco w poprawieniu modelu, ale jego wartość w stosunku do długości snu już tak. Tworzymy więc model uwzględniający tylko te istotne zmienne i zależności.

Wyniki predykcji dla tego modelu są lepsze (ok. 86.6%). Upewniamy się też t.testem, że średni wynik poprawnych predykcji jest znacząco wyższy niż dla pierwszego modelu. Obala to naszą wstępną hipotezę, jednak spróbujmy zrobić jeszcze lepszy model, dodając zmienną towarzyszącą.

Patrząc na diagram korelacji, pozostałe zmienne albo nie mają istotnej korelacji z Jakością snu, albo mają ją również ze zmienną Długości snu. W naszych danych mamy też jednak zmienne kategoryczne, które mogą mieć istotny wpływ. Popatrzmy na średnią jakość snu w zależności od kategorii BMI.

BMI.Category	Średnia_Jakość_snu	Średni_poziom_stresu	Śrdnia_długość_snu
0	7.638889	5.134259	7.387963
1	6.898649	5.729730	6.770270
2	6.400000	5.700000	6.960000

Widzimy że jakość snu dosyć znacząco różni się w zależności od grupy BMI. W tym samym czasie, Poziom stresu i Długość snu wydaje się niezbyt zależna od kategorii. Wydaje się więc to dobry kandydat na zmienną towarzyszącą w naszym modelu.

Podsumowanie modelu Quality.of.Sleep ~ Sleep.Duration + Sleep.Duration:Stress.Level + BMI.Category

term	estimate	std.error	statistic	p.value
(Intercept)	4.1557086	0.4405866	9.432218	0.00e+00
Sleep.Duration	0.8069299	0.0432068	18.675982	0.00e+00
BMI.Category	-0.2099379	0.0488480	-4.297776	2.45e-05
Sleep.Duration:Stress.Level	-0.0613334	0.0034211	-17.927761	0.00e+00

Ponownie wyniki predykcji znacząco się poprawiły (ok. 91%), a t.test potwierdza istotność tej różnicy w stosunku do poprzedniego modelu. Wszystkie współczynniki są istotne, więc wydaje się to być najlepszy z dotychczasowych modeli, co ostatecznie obala hipotezę - sytuacja była bardziej skomplikowana niż się wydawało i był potrzebny bardziej skomplikowany model.

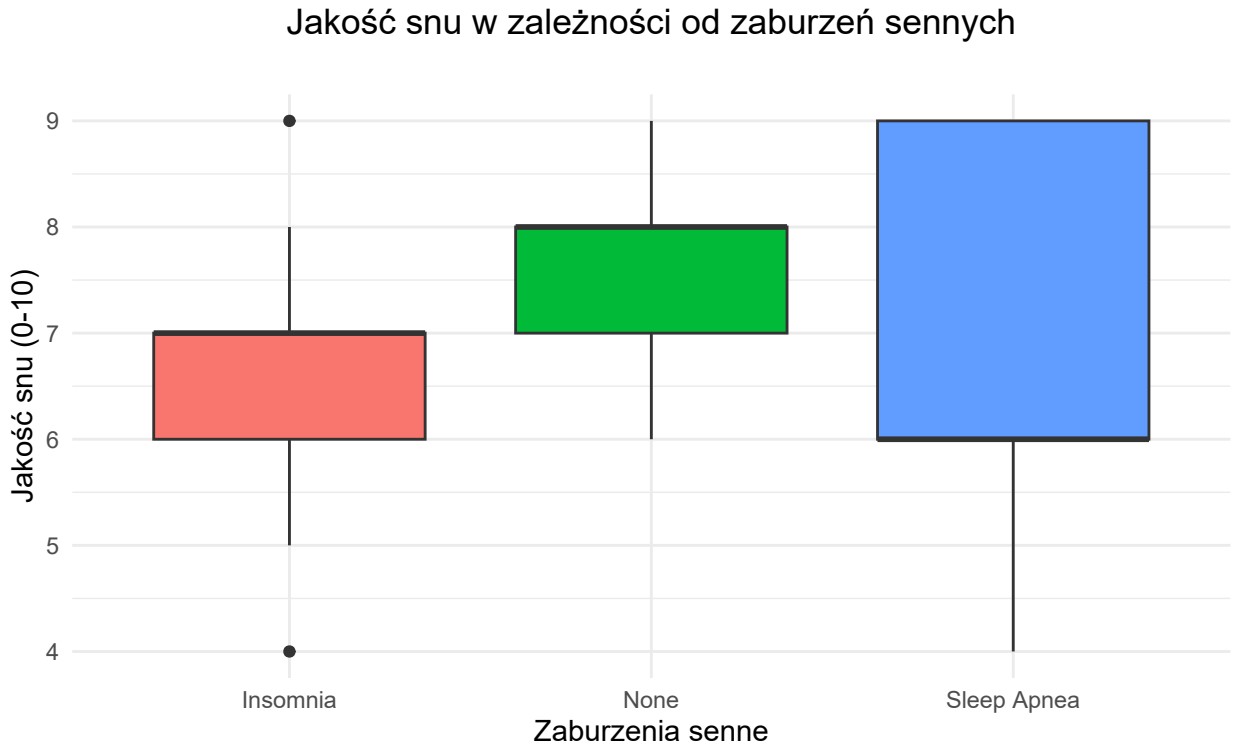
Porównanie dokładności wszystkich modeli

Model	Dokładność
Jakość ~ Długość snu + Poziom Stresu	0.8331672
Jakość ~ Długość snu + Długość snu:Poziom Stresu	0.8710072
Jakość ~ Długość snu + Długość snu:Poziom Stresu + Kategoria BMI	0.9097675

Należy jednak pamiętać, że nie świadczy to jeszcze o bezpośrednim wpływie zmiennej BMI na Jakość snu. Aby to sprawdzić, trezba dokonać dogłębniejszej analizy .

Hipoteza 3: Zmienna Kategoria BMI bezpośrednio wpływa na jakość snu.

Zauważyliśmy już, że zachodzi korelacja między kategorią BMI a jakością snu. Istnieje jednak przypuszczenie, że nie zależy to bezpośrednio od tej zmiennej, a od innej zmiennej, korelującej z nią. Na przykład wydaje się, że na Jakość snu powinny wpływać zaburzenia senne takie jak insomnia i bezdech senny. Możliwe, że BMI ma istotny wpływ na występowanie tych zaburzeń, przez co pośrednio wpływa na Jakość snu, ale nie bezpośrednio. Na początku sprawdźmy, czy rzeczywiście posiadanie zaburzenia wpływa istotnie Jakość snu, do czego wykorzystamy model ANOVA.



Podsumowanie modelu ANOVA Quality.of.Sleep ~ Sleep.Disorder

term	df	sumsq	meansq	statistic	p.value
Sleep.Disorder	2	69.21481	34.607406	27.6006	0
Residuals	371	465.18358	1.253864	NA	NA

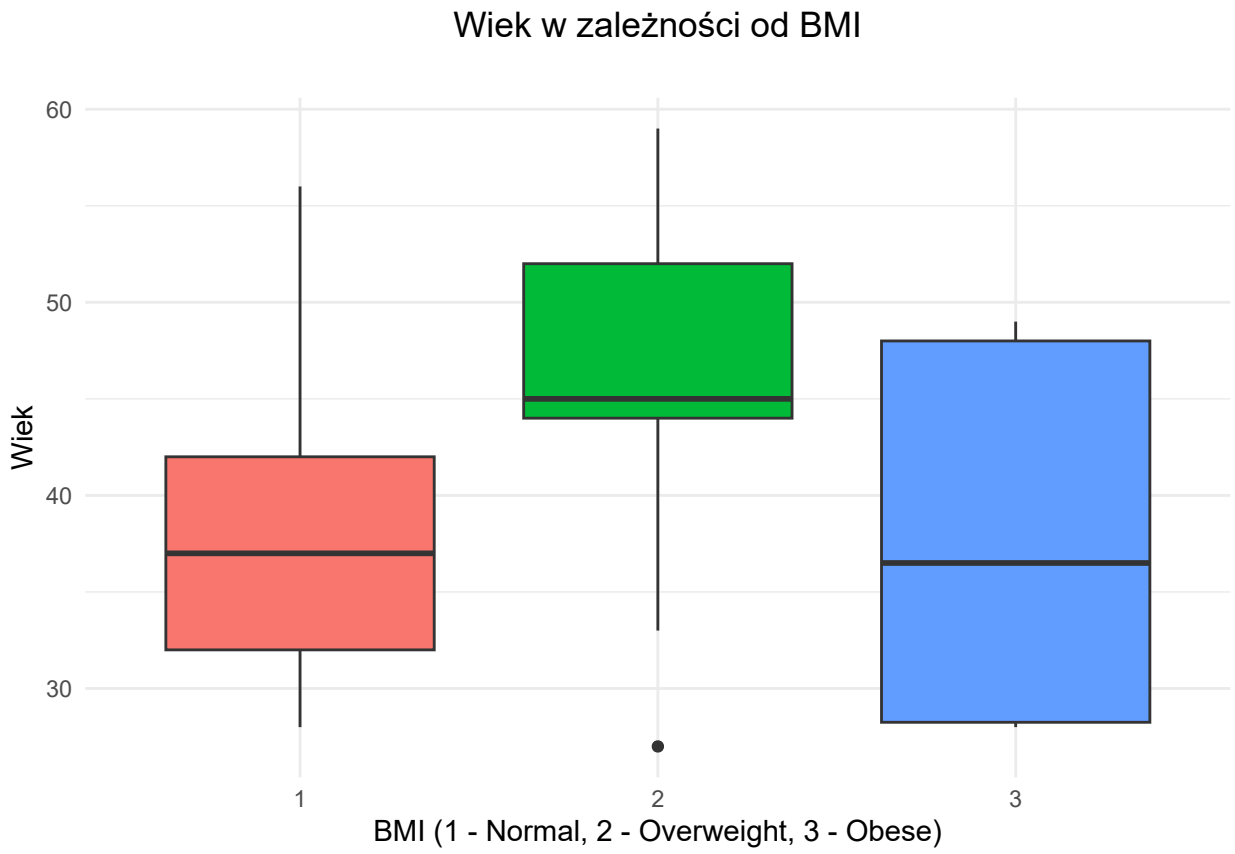
Widzimy istotny wpływ zaburzeń sennych na jakość snu, co nie jest zaskoczeniem. Chcemy teraz sprawdzić, czy BMI wpływa na jakość snu poprzez korelacje z zaburzeniami, czy jest istotne niezależnie od nich. Stworzymy więc model ANOVA, w którym głównym efektem będzie zaburzenie, a BMI będzie efektem interakcji.

Podsumowanie modelu ANOVA Quality.of.Sleep ~ Sleep.Disorder*BMI.Category

term	df	sumsq	meansq	statistic	p.value
Sleep.Disorder	2	69.214812	34.607406	28.410726	0.0000000
BMI.Category	2	10.284417	5.142209	4.221463	0.0153981
Sleep.Disorder:BMI.Category	3	9.070683	3.023561	2.482173	0.0606740
Residuals	366	445.828483	1.218111	NA	NA

Widzimy, że p-value dla kategorii BMI jest istotne statystycznie w tym modelu. Oznacza to, że niezależnie czy dana osoba ma zaburzenie senne czy nie, BMI jest istotne i wpływa na jakość snu. Biorąc na przykład grupę zdrowych osób, wśród nich można zauważyć pogorszenie jakości snu przy większym BMI.

Istnieje też jednak druga wątpliwość, że kluczowy dla jakości snu jest wiek. Im osoba starsza tym gorszej jakości może mieć sen, a przy okazji ma też większe BMI, co nieprawidłowo sugeruje, że BMI przyczynia się gorszej jakości snu. Sprawdźmy, czy wiek wpływa na jakość snu i czy istnieje korelacja między wiekiem i BMI.



Podsumowanie modelu ANOVA Age ~ BMI.Category

term	df	sumsq	meansq	statistic	p.value
BMI.Category	2	7961.389	3980.69471	73.48592	0
Residuals	371	20096.881	54.16949	NA	NA

Podsumowanie modelu Quality.of.Sleep ~ Age

term	estimate	std.error	statistic	p.value
(Intercept)	4.5548651	0.2713705	16.78468	0
Age	0.0653787	0.0063015	10.37513	0

Można wyciągnąć wnioski, że istnieje zarówno korelacja między wiekiem a kategorią BMI, jak i między wiekiem i jakością snu, więc nasze podejrzenia się potwierdzają. Trzeba więc ponownie sprawdzić, czy wpływ BMI na Jakość snu jest niezależny od drugiej zmiennej. Wykonujemy podobny model ANOVA.

Podsumowanie modelu ANOVA Quality.of.Sleep ~ Age*BMI.Category

term	df	sumsq	meansq	statistic	p.value
Age	1	119.931711	119.9317109	242.845435	0.0000000
BMI.Category	2	225.066042	112.5330208	227.864092	0.0000000
Age:BMI.Category	2	7.660066	3.8300328	7.755297	0.0005023
Residuals	368	181.740578	0.4938603	NA	NA

Również tutaj p-value dla kategorii BMI jest nieduże, więc BMI ma istotny wpływ na jakość snu niezależnie od wieku. Mimo wątpliwości nie byliśmy więc w stanie obalić naszej hipotezy, BMI jest istotnym czynnikiem i niezależnie od innych czynników wpływa na jakość snu. Na podstawie tych danych więc potwierdzamy tę hipotezę.