## Projekt indywidualny

## Data Mining

## Wykonał:

## Ivan Kaliankovich

Spis Treści

[Projekt indywidualny 1](#_Toc165731220)

[Data Mining 1](#_Toc165731221)

[Wykonał: 1](#_Toc165731222)

[Ivan Kaliankovich 1](#_Toc165731223)

[Zad 1 3](#_Toc165731224)

[a) Dodanie bazy 3](#_Toc165731225)

[b) Kwantyl 5](#_Toc165731226)

[e) Wykresy 5](#_Toc165731227)

[f) Amerykańskie vs Nieamyrykańskie 7](#_Toc165731228)

[g) Wykresy 7](#_Toc165731229)

[h) Histogram 8](#_Toc165731230)

[Zad2 Śmiertelność a zanieczyszczenie powietrza 9](#_Toc165731231)

[a) Wczytanie oraz analiza statystyczna 9](#_Toc165731232)

[b) Współczynnik nachylenia oraz błąd standardowy 10](#_Toc165731233)

[c) Log(N0x) 10](#_Toc165731234)

[d) Residua 11](#_Toc165731235)

[Zad3 Oszczędności 11](#_Toc165731236)

[a) Wczytać dane savings 11](#_Toc165731237)

[b) Wykres reszt 12](#_Toc165731238)

[c) Wartość dźwigni 13](#_Toc165731239)

[d) DFFITS, DFBETAS, odległości Cooke’a 14](#_Toc165731240)

[e) Model bez największej odległości Cooke’a 14](#_Toc165731241)

[f) 15](#_Toc165731242)

[zad4 Domy w Chicago 16](#_Toc165731243)

[a) Analiza wpływu liczby sypialni w modelu 16](#_Toc165731244)

[b) Szacowanie ceny własnego domu 16](#_Toc165731245)

[zad5 30 Wysp 16](#_Toc165731246)

[zad6 Irys 17](#_Toc165731247)

[a) Drzewo klasyfikacyjne 17](#_Toc165731248)

[b) Macierz błędów 18](#_Toc165731249)

Zadania wykonywałem używając specjalnego środowiska do programów w języku R: RStudio. Program dodatkowo instaluje język R oraz można wykonywać polecenia jedno po drugim co wspomaga w śledzeniu każdego z poleceń.

# Zad 1

## Dodanie bazy

Kod:  
Obraz zawierający tekst, Czcionka, zrzut ekranu, linia

Opis wygenerowany automatycznie

Za pomocą poleceń library oraz data uzyskuję dostęp do zmiennych oraz jestem w stanie otworzyć plik car93 z podglądem każdej wartości  
Obraz zawierający tekst, numer, zrzut ekranu, Czcionka

Opis wygenerowany automatycznie

Za pomocą zmiennej cols\_to\_summary oraz head z parametrem 20 wyświetlam wybrane parametry takie jak Min.Price, MPG.city,MPG.highway, Weight, Origin, Type.  
Output (pierwsze 20 wierszy z bazy zgodnie z moim założeniem):  
Obraz zawierający tekst, zrzut ekranu, numer

Opis wygenerowany automatycznie  
Dodane wartości do bazy zgodne z poleceniem:   
Obraz zawierający tekst, zrzut ekranu, numer, Czcionka

Opis wygenerowany automatycznie

## b) Kwantyl

Obraz zawierający tekst, zrzut ekranu, Czcionka, numer

Opis wygenerowany automatycznie  
Powyższe polecenia wraz z wyjściem określają, samochodów droższych od ceny kwantyla rzędu 0.95 jest 5 w bazie.

## e) Wykresy

Kod   
Obraz zawierający tekst, Czcionka, linia, zrzut ekranu

Opis wygenerowany automatycznie  
Wykresy są wystarczająco czytelne w podstawowych kolorach, dlatego postanowiłem je nie zmieniać. Sportowych samochodów wykonany program naliczył Obraz zawierający tekst, zrzut ekranu, diagram, design

Opis wygenerowany automatycznie  
Obraz zawierający diagram, tekst, krąg

Opis wygenerowany automatycznie

## f) Amerykańskie vs Nieamyrykańskie

Kod: Obraz zawierający tekst, Czcionka, zrzut ekranu

Opis wygenerowany automatycznie  
Wykresy:   
Obraz zawierający tekst, diagram, zrzut ekranu, Prostokąt

Opis wygenerowany automatycznie  
Wyniki są spodziewane, mediana dystansu przypadająca na jeden galon samochodów nieamerykańskich jest większa od samochodów amerykańskich. Niektóre przypadki samochodów nieamerykańskich mają drastycznie dużą efektywność.

## g) Wykresy

Obraz zawierający tekst, Czcionka, zrzut ekranu

Opis wygenerowany automatycznie

Obraz zawierający tekst, diagram, zrzut ekranu, numer

Opis wygenerowany automatycznie

## h) Histogram

Narysowany za pomocą polecenia   
*hist(Cars93$Weight, main="Histogram częstości dla wagi samochodu", xlab="Waga samochodu", ylab="Częstość")*

Obraz zawierający tekst, zrzut ekranu, diagram, linia

Opis wygenerowany automatycznie

# Zad2 Śmiertelność a zanieczyszczenie powietrza

# Wczytanie oraz analiza statystyczna

Udało się wczytać dane za pomocą funkcji „data”. Zacząłem używać biblioteki ggplot2 do wykonywania wykresów.

Obraz zawierający tekst, zrzut ekranu, linia, Wykres

Opis wygenerowany automatycznie  
Statystyki opisowe dopasowania modelu.

Obraz zawierający tekst, zrzut ekranu, Czcionka, numer

Opis wygenerowany automatycznie

## Współczynnik nachylenia oraz błąd standardowy

Obraz zawierający tekst, zrzut ekranu, Czcionka, linia

Opis wygenerowany automatycznie

Na podstawie powyższych wyników można stwierdzić, że dopasowany model nie dobrze opisuje dane. Współczynnik regresji dla zmiennej NOx nie jest istotny statystycznie oraz współczynnik determinacji jest bardzo niski, co oznacza, że model nie tłumaczy znaczącej części zmienności w zmiennej objaśnianej.

## Log(N0x)

Obraz zawierający tekst, zrzut ekranu, linia, diagram

Opis wygenerowany automatycznie

Obraz zawierający tekst, Czcionka, zrzut ekranu, linia

Opis wygenerowany automatycznie

## Residua

Obserwacje o dużych residuach   
Obraz zawierający tekst, Czcionka, zrzut ekranu, linia

Opis wygenerowany automatycznie  
Po utworzeniu modelu pomijającym obserwację uzyskałem następujące wyniki:   
"R-squared dla oryginalnego modelu: 0.0870955860032189"

"R-squared dla nowego modelu: 0.32201142269628"

Wartość R-squared dla nowego modelu wzrosła do 0.322, co oznacza, że nowy model tłumaczy około 32.2% zmienności w Mortality, co jest znacznie lepszym wynikiem. Te wyniki sugerują, że pominięcie obserwacji z dużymi residuami studentyzowanymi pozwoliło uzyskać lepszy model, który lepiej opisuje dane.

# Zad3 Oszczędności

## a) Wczytać dane savings

Do pliku savings dodałem ręcznie kolumnę „Savings”, gdyż była nieobecna w programie.  
Obraz zawierający tekst, zrzut ekranu, Czcionka, numer

Opis wygenerowany automatycznie

## Wykres reszt

Obraz zawierający tekst, linia, diagram, Wykres

Opis wygenerowany automatycznie

Obraz zawierający tekst, zrzut ekranu, Czcionka, Jaskrawoniebieski

Opis wygenerowany automatycznie

## Wartość dźwigni



Obraz zawierający tekst, zrzut ekranu, Czcionka

Opis wygenerowany automatycznie

## DFFITS, DFBETAS, odległości Cooke’a

Obraz zawierający tekst, zrzut ekranu, Czcionka

Opis wygenerowany automatycznie

W celu ustalenia obserwacji wpływowej ustawiłem wartość progową na 0.1 oraz wyznaczyłem wszystkie wartości powyższe od progu.   
Obraz zawierający tekst, zrzut ekranu, Czcionka, linia

Opis wygenerowany automatycznie

## Model bez największej odległości Cooke’a

Obraz zawierający tekst, zrzut ekranu, Czcionka, numer

Opis wygenerowany automatycznie  
Zmienna dpi jest istotna w modelu dopasowanym.

## f)

# zad4 Domy w Chicago

Po dopasowaniu modelu liniowego uzyskałem następujące wyniki:

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 13.712572 9.514111 1.441 0.1677

Bedroom -7.756208 3.109374 -2.494 0.0232 \*

Space 0.011626 0.008981 1.295 0.2128

Room 5.097706 2.764303 1.844 0.0827.

Lot 0.228063 0.195434 1.167 0.2593

Tax 0.003374 0.006859 0.492 0.6291

Bathroom 5.718372 4.276867 1.337 0.1988

Garage 3.613603 2.064997 1.750 0.0982.

Condition -2.162027 4.137400 -0.523 0.6080

Istnieje istotny statystycznie ujemny związek między liczbą sypialni a ceną domu. Co oznacza, że zwiększenie liczby sypialni o 1 skutkuje średnio zmniejszeniem ceny domu o 7.76 jednostek.  
Liczba łazienek oraz liczba pokoi również wpływa na cenę mieszkania, natomiast są zaprzeczające przypadki. Szerokość frontu działki w stopach, powierzchnia oraz roczny podatek nieznacząco wpływają na cenę mieszkania.

## Analiza wpływu liczby sypialni w modelu

Przy zwiększeniu liczby sypialni o 1 dla wszystkich mieszkań wyniki modelu są identyczne:

**Bedroom -7.756208 3.109374 -2.494 0.0232 \***

Po dodaniu jednej sypialni do wszystkich mieszkań wszystkie inne czynniki pozostają takie same (tj. powierzchnia, liczba pokoi, itp.), model się nie zmieni. Z tego powodu żadna zależność pomiędzy zmienną zależną a niezależną się nie zmieniła, czyli wyniki numeryczne pozostają identyczne.

W przypadku stworzenie modelu liniowego, w którym do parametru Bedroom dodałem 1 w każdym przypadku oraz Price zależy jedynie od liczby sypialni uzyskałem następujący wynik

**Bedroom 3.921 1.797 2.182 0.0391 \***

W tym przypadku wyraźnie widać, że cena teraz jest mniej zależna od liczby sypialni, niż w poprzednich przypadkach.

## b) Szacowanie ceny własnego domu

Użyłem funkcji predict z 95 procentowym udziałem ufności  
Uzyskany wynik dla mojego domu to oczekiwana cena = 54.60 jednostek, z dolnym ograniczeniem ufności około 45.82 jednostek i górnym ograniczeniem ufności około 63.39 jednostek.

# zad5 30 Wysp

a)

# zad6 Irys

## Drzewo klasyfikacyjne

Dane zostały wczytane ze strony internetowej oraz podzielone na zbiory treningowe i testowy w stosunku 6 do 4 z góry w dół uzyskanej tabeli.

Drzewo klasyfikacyjne, które zostało wygenerowane prezentuje się następująco w postaci tekstowej w konsoli programu:

N = 120

node), split, n, loss, yval, (yprob)

\* denotes terminal node

1) root 120 75 Iris-virginica (0.33333333 0.29166667 0.37500000)

2) Petal.Length< 2.45 40 0 Iris-setosa (1.00000000 0.00000000 0.00000000) \*

3) Petal.Length>=2.45 80 35 Iris-virginica (0.00000000 0.43750000 0.56250000)

6) Petal.Length< 4.75 32 1 Iris-versicolor (0.00000000 0.96875000 0.03125000) \*

7) Petal.Length>=4.75 48 4 Iris-virginica (0.00000000 0.08333333 0.91666667) \*

**Root** : Całkowita liczba obserwacji w tym węźle wynosi 120. Na podstawie atrybutu Petal.Length model podjął decyzję, że najbardziej odpowiednią klasą dla obserwacji w tym węźle jest Iris-virginica. Rozkład prawdopodobieństwa dla klas: Iris-setosa (33.33%), Iris-versicolor (29.17%), Iris-virginica (37.5%).  
**Węzeł 2**: Jeśli długość płatka (Petal.Length) jest mniejsza niż 2.45, to drzewo klasyfikuje kwiat jako Iris-setosa. W tym węźle jest 40 obserwacji należących do tej klasy.  
**Węzeł 3**: Jeśli długość płatka jest większa lub równa 2.45, drzewo przechodzi do kolejnego węzła. W tym węźle, jeśli długość płatka jest mniejsza niż 4.75, to drzewo klasyfikuje kwiat jako Iris-versicolor, a jeśli jest większa lub równa 4.75, jako Iris-virginica.  
**Węzeł 6**: W tym węźle są 32 obserwacje, dla których długość płatka jest mniejsza niż 4.75. Drzewo klasyfikuje je jako Iris-versicolor z bardzo wysokim prawdopodobieństwem (96.88%).  
**Węzeł 7**: W tym węźle są 48 obserwacji, dla których długość płatka jest większa lub równa 4.75. Drzewo klasyfikuje je jako Iris-virginica z wysokim prawdopodobieństwem (91.67%).

Postać graficzna doskonale ilustruję wszystkie zależności

Obraz zawierający tekst, zrzut ekranu, diagram

Opis wygenerowany automatycznie

## Macierz błędów

Algorytm został użyty na zbiorze testowym i uzyskał dokładność na poziomie 91.(6) %. Z 60 poddanych analizie gatunków prawidłowo określone one zostały w 55 przypadkach. Wszystkie przypadki Iris-setosa zostały określone prawidłowo. Najwięcej błędów, czyli 4 algorytm popełnił przy błędnej klasyfikacji Iris-virginica jako Iris-versicolor. Jeden przypadek był w przeciwną stronę. Parameter LodygaWid jest do poprawienia do lepszej precyzji modelu.