

Palmer Penguins: Statystyczna analiza danych

408453, Łukasz Wala, poniedziałek 14⁴⁰

*AGH, Wydział Informatyki, Elektroniki i Telekomunikacji
Rachunek prawdopodobieństwa i statystyka 2021/2022*

Kraków, 23 stycznia 2022

Ja, niżej podpisany własnoręcznym podpisem deklaruje, że przygotowałem przedstawiony do oceny projekt samodzielnie i żadna jego część nie jest kopią pracy innej osoby.

.....

1 Streszczenie raportu

Raport powstał w oparciu o analizę cech 344 pingwinów zamieszkujących archipelag Palmera zebranych w latach 2007-2009.

2 Opis danych

Dane do projektu pochodzą z strony projektu **palmerpenguins**. Jest to projekt open-source mający dostarczyć dataset do wizualizacji i eksploracji danych będący alternatywą dla *Iris Dataset*, klasyka wśród materiałów do nauki statystycznej analizy danych i uczenia maszynowego.

W skład projektu wchodzi dwa zestawy danych: *penguins_raw* oraz jego uproszczona wersja *penguins*. W tym projekcie użyta zostanie wersja *penguins*. Dane zawierają 344 rekordy, gdzie każdy odpowiada innemu pingwinowi opisanemu ośmioma atrybutami:

- **species** - gatunek pingwina (zmienna jakościowa spośród *Adelie*, *Chinstrap*, *Gentoo*),
- **island** - wyspa w archipelagu Palmera (zmienna jakościowa spośród *Biscoe*, *Dream*, *Torgersen*),
- **bill_length_mm** - długość dzioba (liczba zmiennoprzecinkowa, w milimetrach),
- **bill_depth_mm** - głębokość dzioba (liczba zmiennoprzecinkowa, w milimetrach),

- **flipper_length_mm** - długość płetwy (liczba całkowita, w milimetrach),
- **body_mass_g** - masa ciała (liczba całkowita, w gramach),
- **sex** - płeć (zmienna jakościowa spośród *male*, *female*),
- **year** - rok obserwacji (liczba całkowita, z zamiarem zmiany na zmienną jakościową)

Do załadowania ich do środowiska R można podejść na dwa sposoby. Z racji tego, że jest to zestaw danych przygotowanych z myślą o edukacji, został spakowany do pakietu R, który można zainstalować poprzez CRAN, a następnie załadować:

```
> install.packages("palmerpenguins", repos = "https://cloud.r-project.org/")
> library(palmerpenguins)
> data(package = 'palmerpenguins')
> penguins_data <- penguins
```

Alternatywnym podejściem będzie pobranie pliku *.csv* z repozytorium projektu i załadowanie go:

```
> penguins_data <- read.csv(file = "penguins.csv", header = TRUE, stringsAsFactors = TRUE)
```

Czyszczenie danych sprowadzi się tylko do usunięcia rekordów z wartościami nieznanymi, które są numeryczne (są tylko dwa takie wiersze i nie zawierają one żadnego z pomiarów, więc nie powoduje to dużej utraty informacji). W przypadku nieznanymi zmiennymi jakościowymi (tylko płci pingwinów w tym przypadku), jedną z metod byłoby zastąpienie ich losowymi wartościami, jednak ze względu na potencjalnie dużą korelację płci z pozostałymi cechami, nie jest to najlepsze rozwiązanie. Najlepszą metodą byłoby zbadanie zależności cech i płci, i na tej podstawie uzupełnienie danych, jednak, dla uproszczenia oraz ze względu na niewielką ilość takich wierszy, tutaj zostaną usunięte. Inną niewielką zmianą będzie konwersja typu roku obserwacji z liczby na zmienną jakościową, w tym przypadku ma to więcej sensu, bo bardziej interesującymi informacjami są cechy pingwinów w danym roku niż średnia lat wykonywania pomiarów:

```
> penguins_data <- na.omit(penguins_data)
> penguins_data$year <- as.factor(penguins_data$year)
```

Wówczas dane wyglądać będą następująco:

```
> str(penguins_data)

'data.frame':      333 obs. of  8 variables:
 $ species      : Factor w/ 3 levels "Adelie","Chinstrap",...: 1 1 1 1 1 1 1 1 1 1 ...
 $ island       : Factor w/ 3 levels "Biscoe","Dream",...: 3 3 3 3 3 3 3 3 3 3 ...
 $ bill_length_mm: num   39.1 39.5 40.3 36.7 39.3 38.9 39.2 41.1 38.6 34.6 ...
 $ bill_depth_mm: num   18.7 17.4 18 19.3 20.6 17.8 19.6 17.6 21.2 21.1 ...
 $ flipper_length_mm: int   181 186 195 193 190 181 195 182 191 198 ...
```

```

$ body_mass_g      : int   3750 3800 3250 3450 3650 3625 4675 3200 3800 4400 ...
$ sex              : Factor w/ 2 levels "female","male": 2 1 1 1 2 1 2 1 2 2 ...
$ year            : Factor w/ 3 levels "2007","2008",...: 1 1 1 1 1 1 1 1 1 1 ...
- attr(*, "na.action")= 'omit' Named int [1:11] 4 9 10 11 12 48 179 219 257 269 ...
..- attr(*, "names")= chr [1:11] "4" "9" "10" "11" ...

```

3 Analiza danych

3.1 Wydobywanie podstawowych informacji z danych

Na początku zostaną obliczone estymatory punktowe podstawowych wielkości statystyki opisowej:

```
> summary(penguins_data)
```

species	island	bill_length_mm	bill_depth_mm
Adelie :146	Biscoe :163	Min. :32.10	Min. :13.10
Chinstrap: 68	Dream :123	1st Qu.:39.50	1st Qu.:15.60
Gentoo :119	Torgersen: 47	Median :44.50	Median :17.30
		Mean :43.99	Mean :17.16
		3rd Qu.:48.60	3rd Qu.:18.70
		Max. :59.60	Max. :21.50

flipper_length_mm	body_mass_g	sex	year
Min. :172	Min. :2700	female:165	2007:103
1st Qu.:190	1st Qu.:3550	male :168	2008:113
Median :197	Median :4050		2009:117
Mean :201	Mean :4207		
3rd Qu.:213	3rd Qu.:4775		
Max. :231	Max. :6300		

```

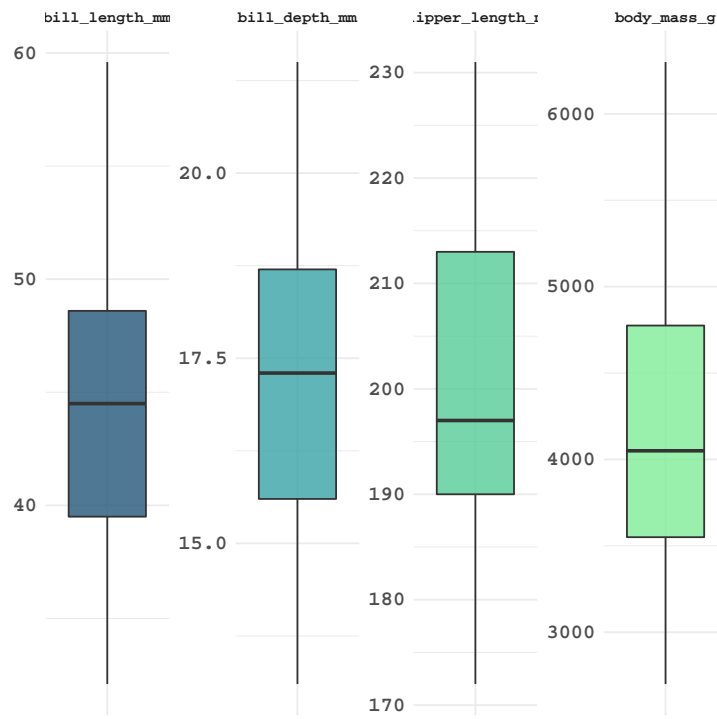
> # zmienne jakościowe ignorowane
> describeBy(penguins_data[c(3:6)])

```

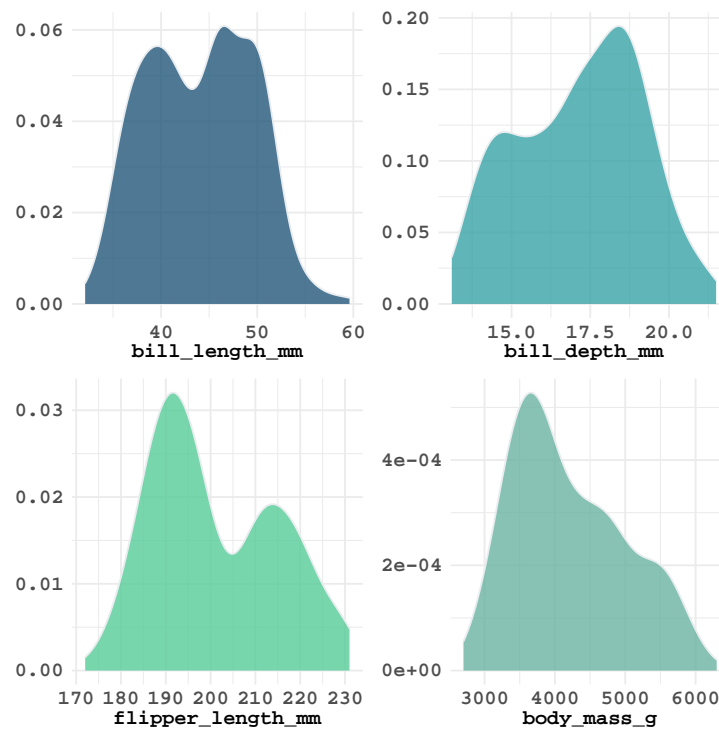
	vars	n	mean	sd	median	trimmed	mad	min	max
bill_length_mm	1	333	43.99	5.47	44.5	43.98	6.97	32.1	59.6
bill_depth_mm	2	333	17.16	1.97	17.3	17.19	2.22	13.1	21.5
flipper_length_mm	3	333	200.97	14.02	197.0	200.36	16.31	172.0	231.0
body_mass_g	4	333	4207.06	805.22	4050.0	4159.46	889.56	2700.0	6300.0

	range	skew	kurtosis	se
bill_length_mm	27.5	0.04	-0.90	0.30
bill_depth_mm	8.4	-0.15	-0.91	0.11
flipper_length_mm	59.0	0.36	-0.98	0.77
body_mass_g	3600.0	0.47	-0.75	44.13

Poniżej kilka wykresów przedstawiających cechy numeryczne:



Jak widać, dla wszystkich cech, z wyjątkiem *bill_depth_mm*, *skew*, czyli współczynnik skośności przyjmuje wartości dodatnie, oznacza to, że rozkłady są (w niewielkim stopniu, bo wartości *skew* nadal bliskie zera) prawostronnie asymetryczne (mają wydłużone prawe ramię rozkładu). Rozkład *bill_depth_mm* jest w podobnym stopniu lewostronnie asymetryczny. Na podstawie wartości kurtozy, która dla każdej cechy jest ujemna, można stwierdzić, że intensywność wartości skrajnych jest mniejsza niż w przypadku rozkładu normalnego. Dobrze obrazują to poniższe wykresy gęstości prawdopodobieństwa:



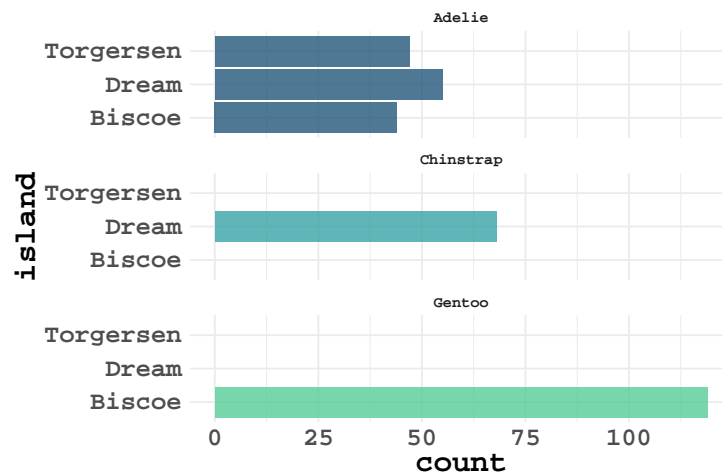
Badanie, czy dane pochodzą z jakiegoś rozkładu, zostanie przeprowadzone w dalszej części dokumentu, ale już teraz, na podstawie wykresów gęstości i ich podobieństwa do wykresu gęstości rozkładu normalnego, można snuć pewne przypuszczenia na ten temat.

3.2 Badanie zależności pomiędzy danymi

Na początku pod lupę zostaną wzięte zmienne jakościowe i zależności pomiędzy nimi.

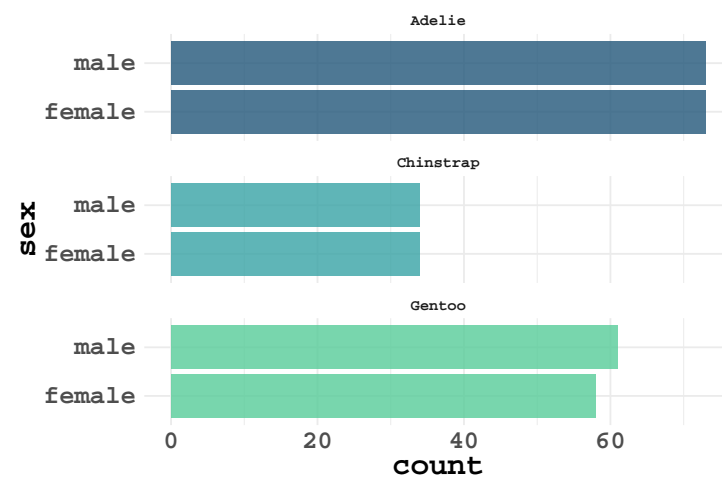
```
> # gatunek, a wyspa, na której został zaobserwowany
> penguins_data %>% count(species, island)
```

	species	island	n
1	Adelie	Biscoe	44
2	Adelie	Dream	55
3	Adelie	Torgersen	47
4	Chinstrap	Dream	68
5	Gentoo	Biscoe	119



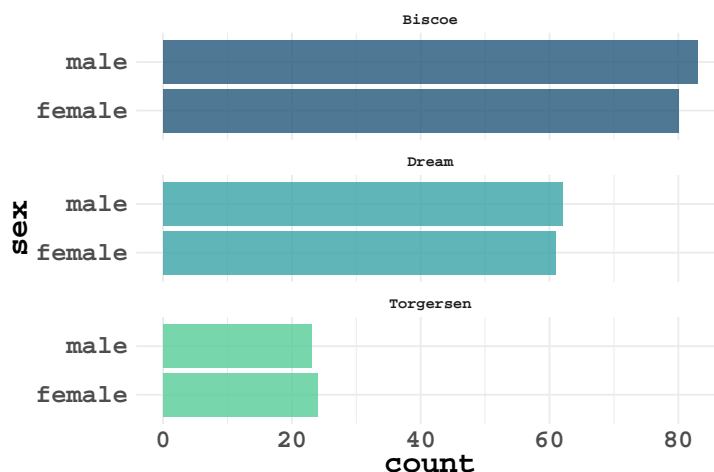
```
> # gatunek, a pleć
> penguins_data %>% count(sex, species)
```

```
  sex  species  n
1 female  Adelie 73
2 female Chinstrap 34
3 female  Gentoo 58
4  male  Adelie 73
5  male Chinstrap 34
6  male  Gentoo 61
```



```
> # wyspa, na której dokonano pomiaru, a pleć
> penguins_data %>% count(sex, island)
```

	sex	island	n
1	female	Biscoe	80
2	female	Dream	61
3	female	Torgersen	24
4	male	Biscoe	83
5	male	Dream	62
6	male	Torgersen	23

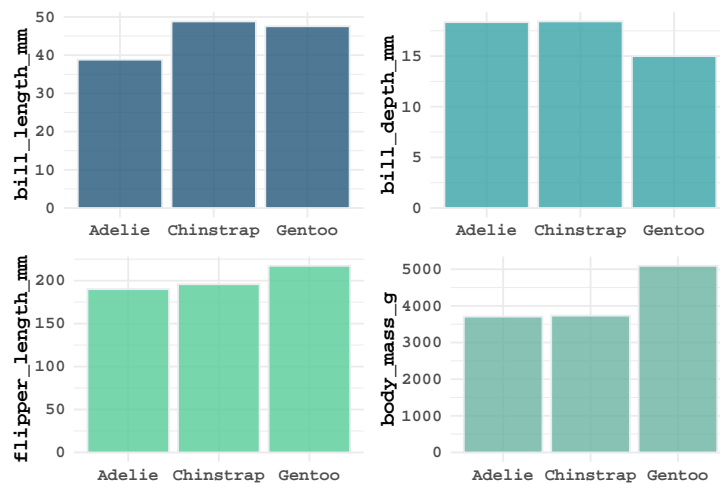


Ciekawą cechą, którą można tutaj zauważyć, jest to, że tylko gatunek *Adelie* żyje wszystkie na wszystkich trzech wyspach, reszta gatunków zamieszkuje jedną wyspę. Rozkład płci względem gatunku oraz wyspy występowanie jest praktycznie równomierny, trudno zaobserwować inne zależności pomiędzy cechami jakościowymi.

Teraz zależności pomiędzy cechami numerycznymi i jakościowymi. Dla uproszczenia pod uwagę wzięte zostaną tylko średnie wartości grupowane po cechach jakościowych.

```
> # zależność pomiędzy gatunkiem a cechami numerycznymi
> penguins_data %>% group_by(species) %>% summarise(across(-c(island, sex, year), mean))

# A tibble: 3 x 5
  species    bill_length_mm bill_depth_mm flipper_length_mm body_mass_g
  <fct>          <dbl>          <dbl>          <dbl>          <dbl>
1 Adelie        38.8            18.3            190.        3706.
2 Chinstrap    48.8            18.4            196.        3733.
3 Gentoo       47.6            15.0            217.        5092.
```



Jak widać na podstawie średnich, cechy numeryczne pingwinów z gatunków *Adelie* oraz *Chinstrap* nie różnią się bardzo z wyjątkiem długości dzioba (średnio nieznacznie dłuższe u *Chinstrap*). Pingwiny gatunku *Gentoo* wyróżniają się natomiast średnio krótszymi głębokościami dzioba, ale większymi masami ciała.

```
> # zależność pomiędzy wyspą występowania a cechami numerycznymi
> penguins_data %>% group_by(island) %>% summarise(across(-c(species, sex, year), mean))
```

```
# A tibble: 3 x 5
  island   bill_length_mm bill_depth_mm flipper_length_mm body_mass_g
  <fct>         <dbl>         <dbl>         <dbl>         <dbl>
1 Biscoe       45.2           15.9           210.          4719.
2 Dream        44.2           18.3           193.          3719.
3 Torgersen    39.0           18.5           192.          3709.
```

Należy pamiętać, że gatunek jest mocno skorelowany z wyspą występowania, więc grupowanie danych po wyspach nie wprowadza wiele nowych danych.

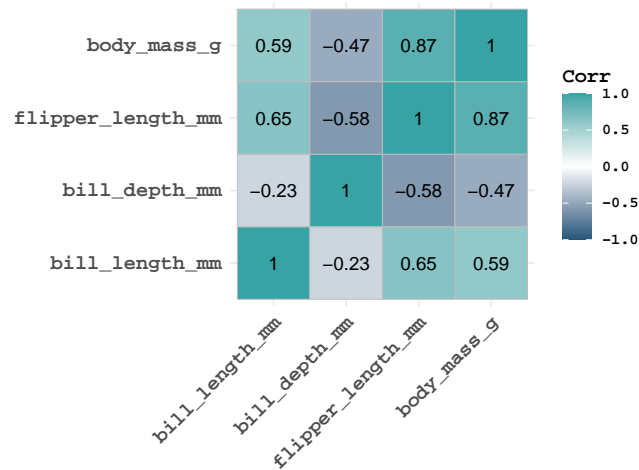
```
> # zależność pomiędzy płcią a cechami numerycznymi
> penguins_data %>% group_by(species, sex) %>% summarise(across(-c(island, year), mean))
```

```
# A tibble: 6 x 6
# Groups:   species [3]
  species sex   bill_length_mm bill_depth_mm flipper_length_mm body_mass_g
  <fct>   <fct>         <dbl>         <dbl>         <dbl>         <dbl>
1 Adelie female      37.3           17.6           188.          3369.
2 Adelie male       40.4           19.1           192.          4043.
3 Chinstrap female    46.6           17.6           192.          3527.
4 Chinstrap male     51.1           19.3           200.          3939.
5 Gentoo female     45.6           14.2           213.          4680.
6 Gentoo male      49.5           15.7           222.          5485.
```

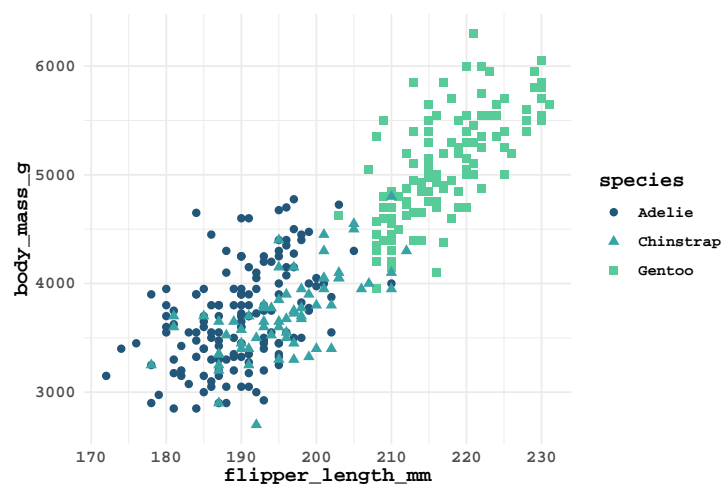

Samce pingwinów, niezależnie od gatunku, przeważają pod względem każdej cechy, szczególnie masy ciała.

Teraz pora na zbadanie zależności pomiędzy cechami numerycznymi. Również można by to rozbić, rozbijając dane na grupy względem cech jakościowych, ale tutaj, dla uproszczenia, użyte zostaną dane z całej próby. Najpierw obliczona zostanie macierz korelacji dla cech numerycznych:

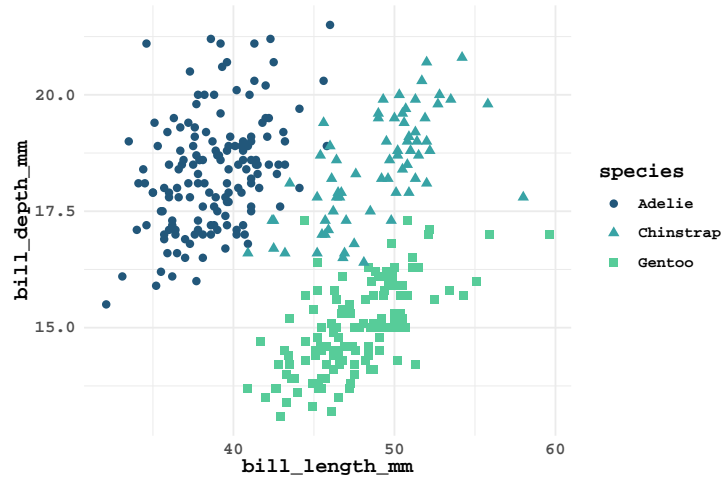
```
> corr <- round(cor(penguins_data[c(3:6)]), 2)
```



Można zaobserwować m.in. wysoki współczynnik korelacji pomiędzy masą ciała pingwina, a długością płetwy (skrzydła). Łatwo to zaobserwować na poniższym wykresie:



Natomiast np. długość i głębokość dzioba wykazują niższy poziom korelacji:



3.3 Estymatory przedziałowe

3.4 Testowanie hipotez

Testować będziemy hipotezy zerowe... **H0**: ... wobec hipotezy alternatywnej **H1**: ... Korzystam ze statystyki t -Studenta postaci

$$t = \sum_{i=1}^n \frac{\text{licznik} X_i}{\text{mianownik}^2}$$

3.5 Regresja

W ten sposób możemy zapisać równania w \LaTeX , znakiem AND wyróżniamy je, a dwa slashy służą do przejścia do kolejnej linii.

$$\begin{aligned} y &= a \cdot x + b + \varepsilon, \\ z &= 3 \cdot y. \end{aligned}$$

4 Wnioski

Wnioski płynące z przeprowadzonej analizy, są następujące:

- wniosek pierwszy,
- wniosek drugi,
- i kolejne.