BRNO UNIVERSITY OF TECHNOLOGY
FACULTY OF INFORMATION TECHNOLOGY

Ukládání a příprava dat

Projekt 2

22. listopadu 2022

Team xziska03

Žiška Marek xziska03
Osvald Martin xosval03
Daša Nosková xnosko05

Obsah

Kapitola 1

Exploratívna analýza

1.1 Atribúty dátovej sady

Dátová sada obsahuje 2 súbory avšak, súbor penguins_size je podsetom súboru penguins_iter. V datasete sa nachádza 17 stĺpcov, ktoré sa dajú rozdeliť na numerické a kategorické typy. Jeden stĺpec Date egg je typu `object`, ale dá sa pokladať za dátumový údaj. Za kategorické sú pokladané hodnoty typu `object`.

Z tabuľky 1.1 je vidieť, že niektorí tučniaci sa vyskytujú v datasete viackrát, ale nakoľko ide o tučniaka v rôznych štúdiách a tučniak časom rastie, tento fakt sa dá ignorovať.

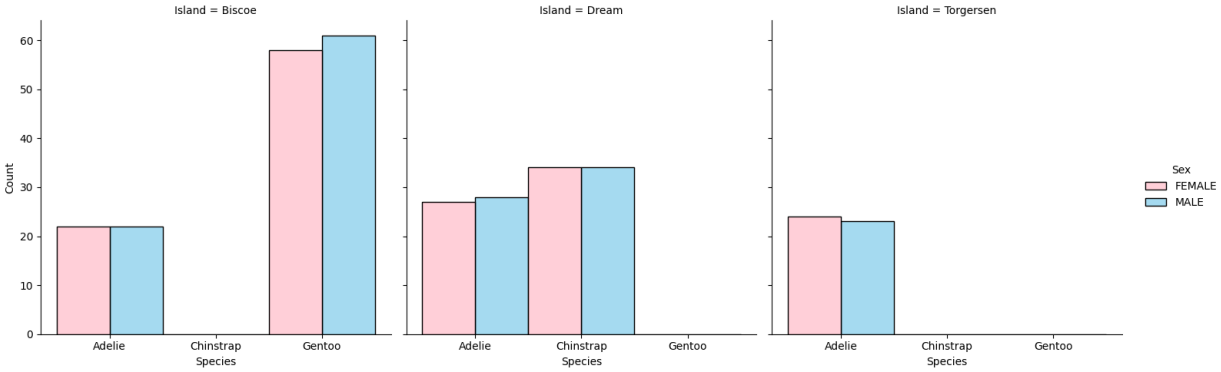
	typ	počet	priemer	std	min	25%	50%	75%	max
Sample Number	int64	344.0	63.15	40.43	1.00	29.00	58.00	95.25	152.00
Culmen Length (mm)	float64	342.0	43.92	5.46	32.10	39.22	44.45	48.50	59.60
Culmen Depth (mm)	float64	342.0	17.15	1.97	13.10	15.60	17.30	18.70	21.50
Flipper Length (mm)	float64	342.0	200.92	14.06	172.00	190.00	197.00	213.00	231.00
Body Mass (g)	float64	342.0	4201.75	801.95	2700.00	3550.00	4050.00	4750.00	6300.00
Delta 15 N (o/oo)	float64	330.0	8.73	0.55	7.63	8.30	8.65	9.17	10.03
Delta 13 C (o/oo)	float64	331.0	-25.69	0.79	-27.02	-26.32	-25.83	-25.06	-23.79

Tabulka 1.1: Numerické atribúty a ich štatistické miery

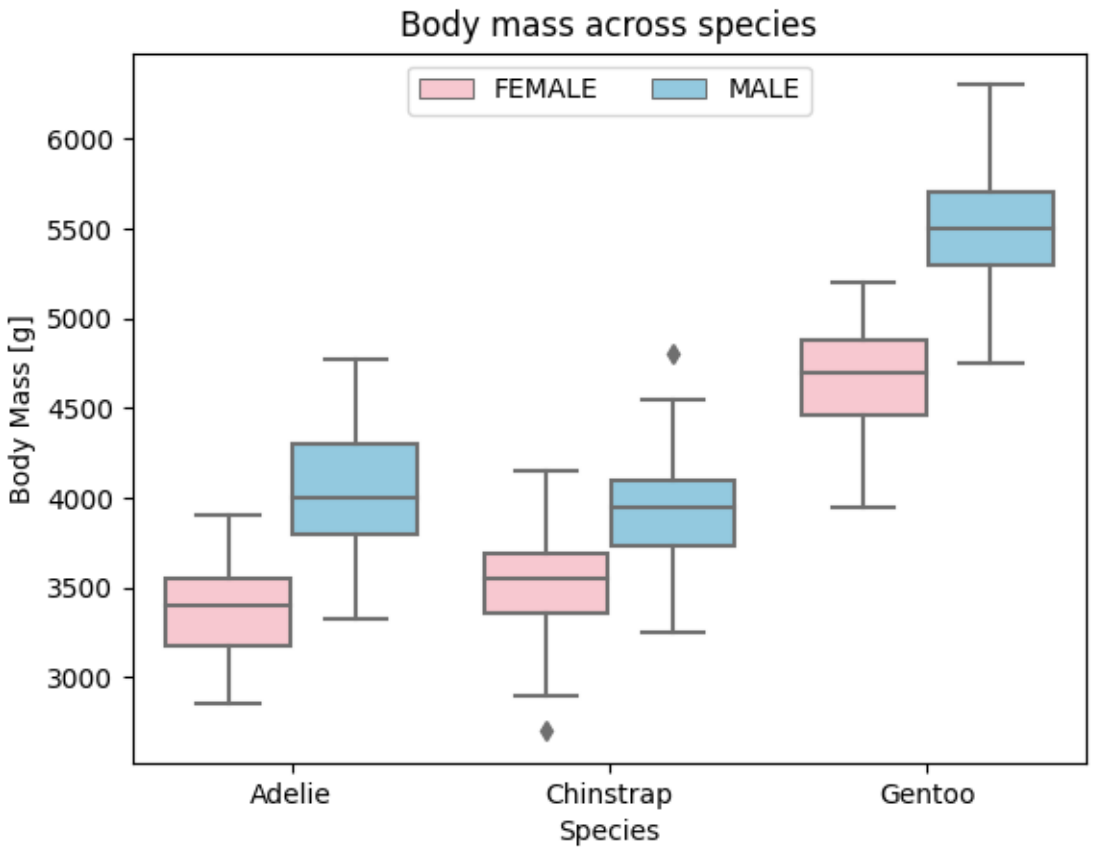
	count	unique	top	freq	kategórie
studyName	344	3	PAL0910	120	PAL0708, PAL0809, PAL0910
Species	344	3	Adelie Penguin (Pygoscelis adeliae)	152	Adelie Penguin (Pygoscelis adeliae), Chinstrap penguin (Pygoscelis antarctica), Gentoo penguin (Pygoscelis papua)
Region	344	1	Anvers	344	Anvers
Island	344	3	Biscoe	168	Torgersen, Biscoe, Dream
Stage	344	1	Adult, 1 Egg Stage	344	Adult, 1 Egg Stage
Individual ID	344	190	N21A1	3	
Clutch Completion	344	2	Yes	308	Yes, No
Sex	334	3	MALE	168	Male, Female
Comments	26	7	Nest never observed with full clutch.	13	

Tabulka 1.2: Kategorické atribúty s počtom unikátnych hodnôt, najfrekvenovanejšou hodnotou a so všetkými hodnotami jednotlivých kategórií

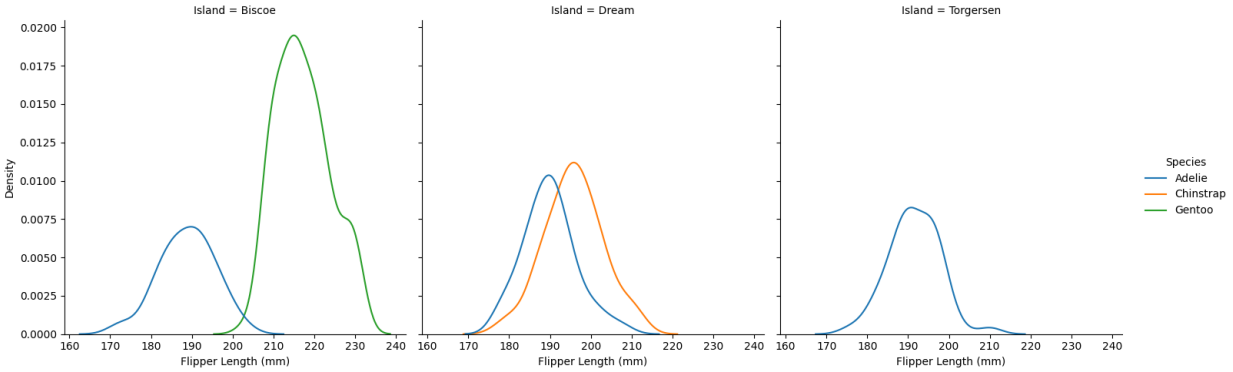
1.2 Rozloženie hodnôt



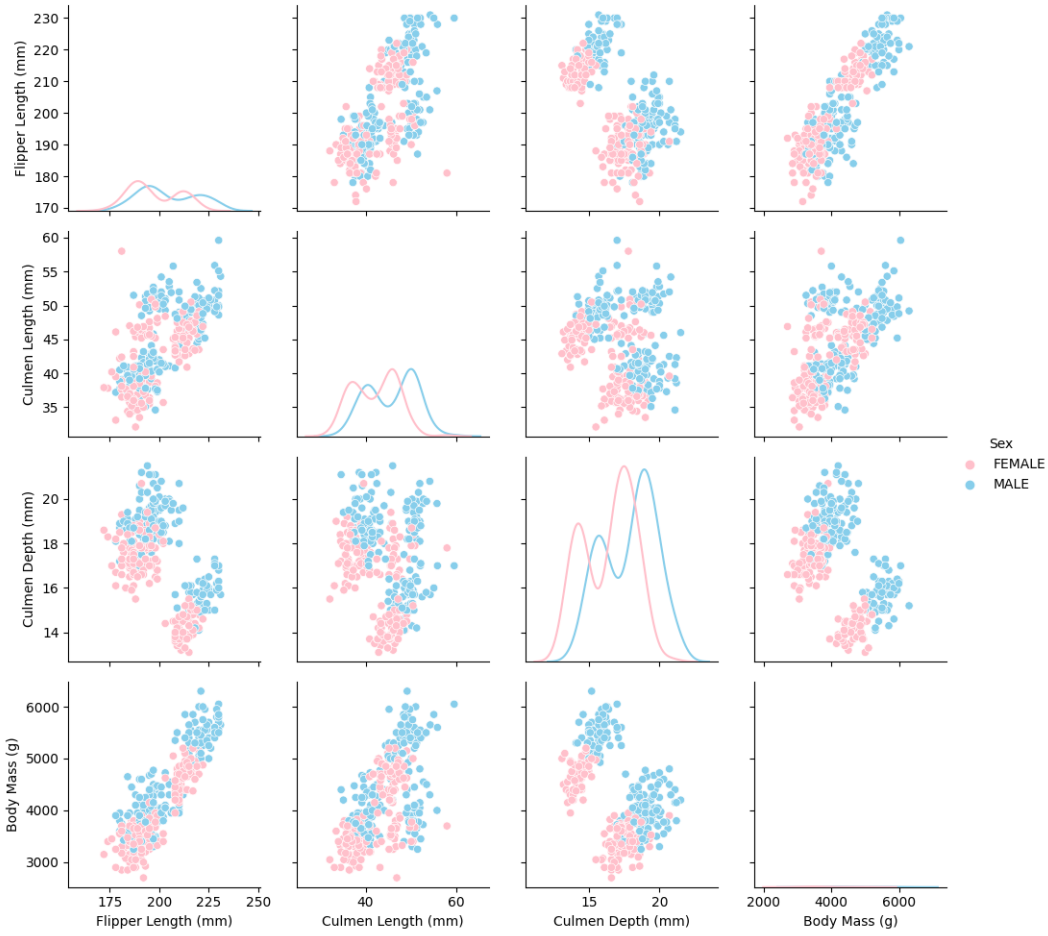
Obrázek 1.1: Histogram sledujúci početnosť tučniakov na ostrovoch. Histogram bol zvolený, pretože ide o kategorické atribúty a porovnáva sa početnosť tučniakov. Z grafu je vidieť, že pohlavie tučniakov je vyrovnané na všetkých ostrovoch. Z grafu je vidieť aj aké druhy tučniakov žijú na jednotlivých ostrovoch.



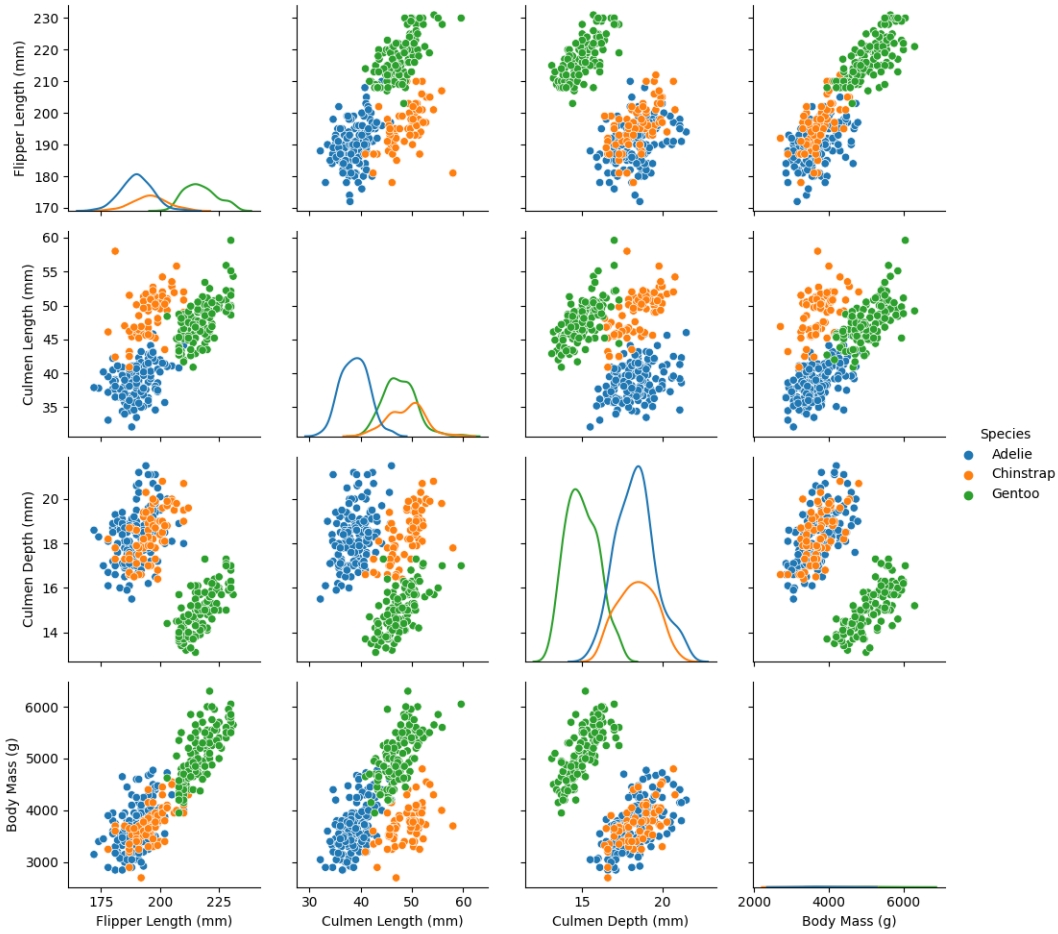
Obrázek 1.2: Krabicový graf zobrazujúci váhu. Tučniaci druhu Gentoo sú najväčšie, čo sa týka hmotnosti, a medzi pohlaviami tučniakov Chinstrap je najmenší rozdiel váhy. Najmenšiu váhu dosahujú samičky druhu Adelie. Najemnšiu váhu u samčekov dosahuje druh Chinstrap.



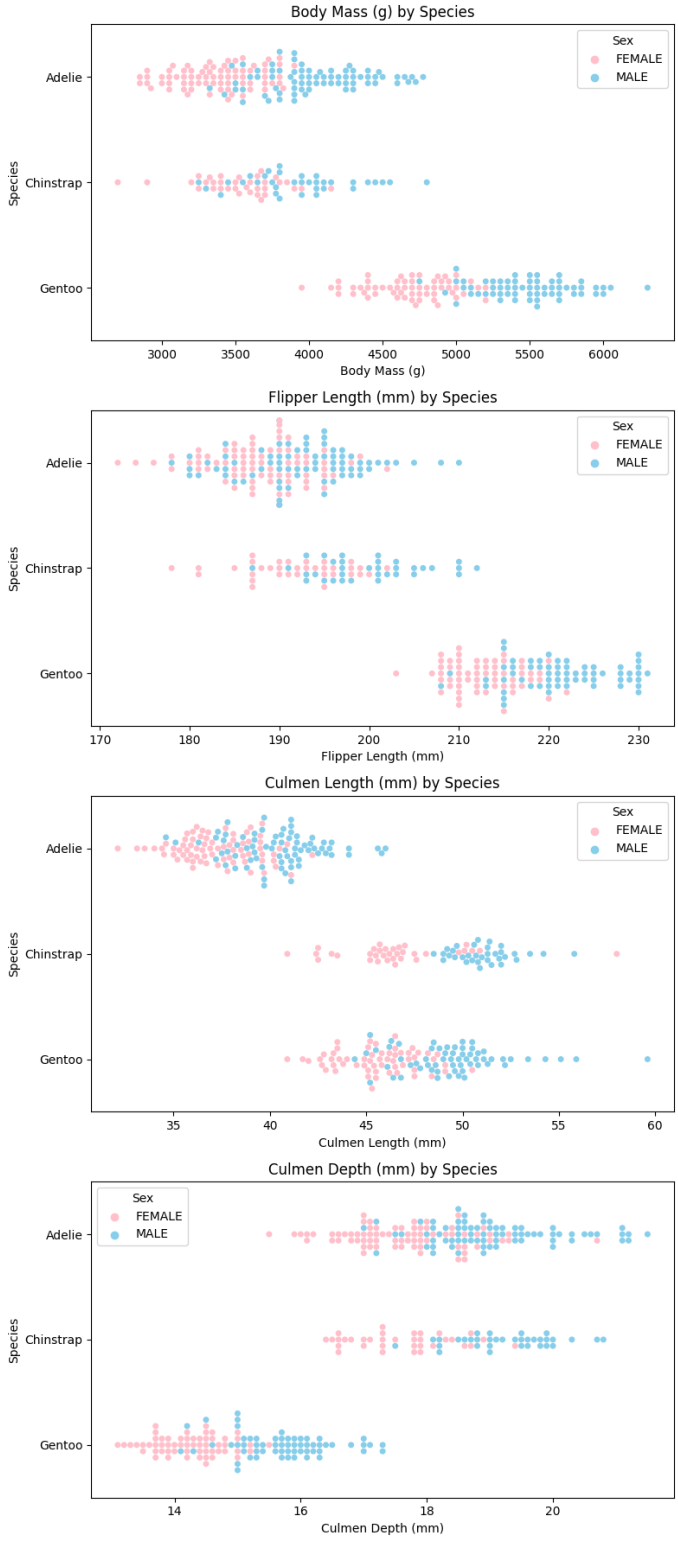
Obrázek 1.3: Veľkosť plutví tučniakov. Druh Adelie sa nachádza na všetkých 3 ostrovoch, avšak na ostrove Biscoe veľkosť ich plutví dosahuje najmenšiu veľkosť. Druh Chinstrap má o pár milimetrov väčšiu veľkosť plutví ako Chinstrap a najväčšie plutvy má druh Gentoo. Tento graf a graf 1.2, nám dáva prvotnú predstavu o tom, ktorý druh je najväčší.



Obrázek 1.4: Zložený graf zobrazujúci vzťahy medzi jednotlivými vlasnostami tučniakov a ich pohlavím. Z grafu môžme vidieť určitú koreláciu medzi jednotlivými atribútmi, a vytvorené zhluky u niektorých atribútov, nám značia rozdielnie rozloženie veľkostí podľa druhu. Vidíme, že samčekovia sú prirodzene väčší ako samičky, ale taktiež vidíme zopár bodov, kde majú niektoré samičky väčšie, popr. menšie hodnoty.



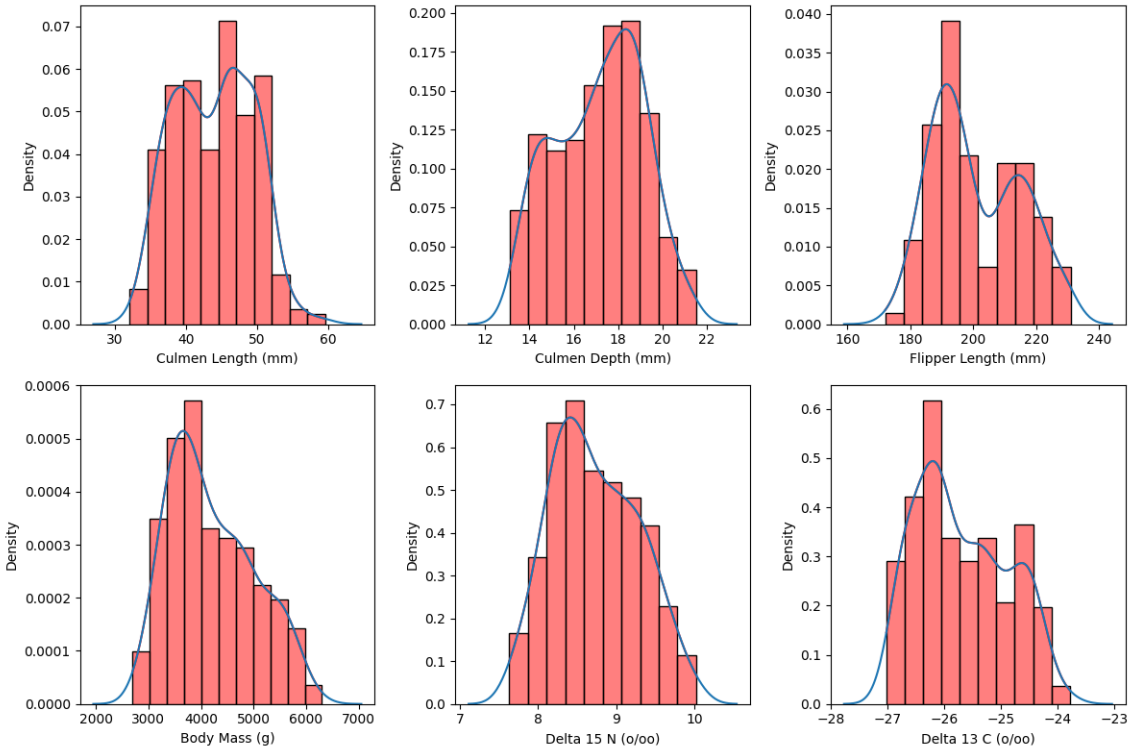
Obrázek 1.5: Zložený graf zobrazujúci vzťahy medzi jednotlivými vlasnostami tučniakov a ich druhom. Po zistení z predchádzajúceho grafu 1.4, bol vytvorený rovnaký graf, ktorý farebne odlišuje tučniakov podľa druhu. Z grafu je vidieť, že tučniaci druhu Gentoo sú najväčšie, ale majú najtenší zobák. Najmenšiu veľkosť zobáku majú tučniaky Adelie, tučniaci Chinstrap a Gentoo majú dĺžku zobáku podobnú, avšak Chinstrap má zobák hrubší.



Obrázek 1.6: Pre lepšie zobrazenie všetkých hodnôt tučniakov bol použitý swarmplot zobrazujúci jednotlivé vlastnosti tučniakov podľa pohlavia a ich druhu. Alternatívou by bol bodový graf, napr. v prípade väčšieho počtu vzorkov. Z grafu je vidieť, aké hodnoty nadobúdajú jednotlivé druhy podľa pohlavia pre numerické atribúty. V grafe sa nachádza pár osamotených bodov, ktoré by mohli byť potencionálne odľahlé hodnoty. Z grafu je možné vyčítať, podobnosti a rozlišnosti pre atribúty naprieč druhmi.

1.3 Odlahlé hodnoty

Dataset na základe prvého pohľadu podľa štatistických mierok nevyzerá, že obsahuje odľahlé hodnoty, ktoré by ovplyvňovali extrémnym spôsobom priemer, ale po analýze grafov, hlavne grafu 1.6, sa tu teoreticky nachádzá zopár odľahlých hodnôt. Pre zistenie, či tieto hodnoty sú naozaj odlahlé bolo použité pravidlo 3σ v kombinácii s normalizáciou hodnôt pomocou z-skóre. Bolo zistené, že dataset neobsahuje žiadne odľahlé hodnoty, čo dokazuje aj graf 1.7. Dátová sada neobsahuje žiadne odľahlé hodnoty ani pre jednotlivé kategórie tučniakov (males, females a druhy tučniakov). Pre overenie správnosti bola použitá aj metóda IQR, ale ani táto metóda nenašla žiadne extrémne odľahlé hodnoty.



Obrázek 1.7: Rozloženie numerických atribútov.

1.4 Chýbajúce hodnoty

Stĺpec	Počet chyb. hodnôt
Culmen Length (mm)	2
Culmen Depth (mm)	2
Flipper Length (mm)	2
Body Mass (g)	2
Sex	11
Delta 15 N (o/oo)	14
Delta 13 C (o/oo)	13

Tabulka 1.3: Celkový počet chýbajúcich hodnôt. Kategória Sex obsahuje 1 záznam chybného typu '.', ktorý je započítaný do počtu v tabuľke.

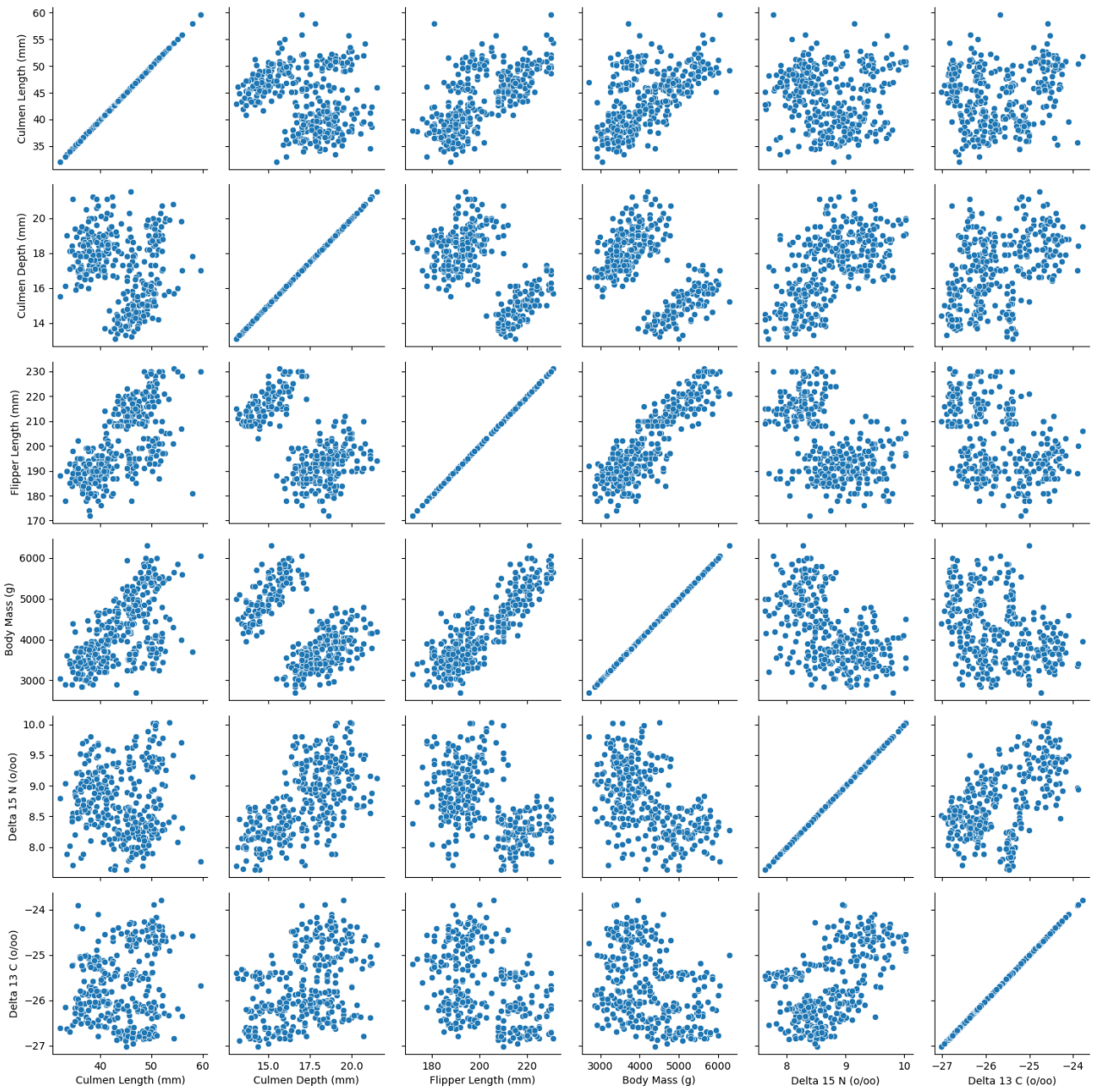
Stĺpec/Index	0	3	8	11	12	13	15	39	41	46	47	250	339
Culmen Length (mm)	39.1	NaN	34.1	37.8	41.1	38.6	36.6	39.8	40.8	41.1	37.5	47.3	NaN
Culmen Depth (mm)	18.7	NaN	18.1	17.3	17.6	21.2	17.8	19.1	18.4	19.0	18.9	15.3	NaN
Flipper Length (mm)	181.0	NaN	193.0	180.0	182.0	191.0	185.0	184.0	195.0	182.0	179.0	222.0	NaN
Body Mass (g)	3750.0	NaN	3475.0	3700.0	3200.0	3800.0	3700.0	4650.0	3900.0	3425.0	2975.0	5250.0	NaN
Sex	MALE	NaN	NaN	NaN	FEMALE	MALE	FEMALE	MALE	MALE	MALE	NaN	MALE	NaN
Delta 15 N (o/oo)	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Delta 13 C (o/oo)	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN

Tabulka 1.4: Riadky s viacerými chýbajúcimi hodnotami.

1.5 Korelačná analýza

Najvyššiu koreláciu má telesná hmostnosť s dĺžkou plutvy, ale korelácia sa nachádza aj pri iných parametroch, kde ale ako je vidieť z grafu 1.5 záleží na druhu tučniaka. Okrem grafu 1.8 a 1.5 zobrazuje koreláciu aj graf 1.4.

	Culmen Length	Culmen Depth	Flipper Length	Body Mass	Delta 15 N	Delta 13 C
Culmen Length (mm)	1.000	-0.235	0.656	0.595	-0.060	0.189
Culmen Depth (mm)	-0.235	1.000	-0.584	-0.472	0.606	0.430
Flipper Length (mm)	0.656	-0.584	1.000	0.871	-0.508	-0.376
Body Mass (g)	0.595	-0.472	0.871	1.000	-0.538	-0.375
Delta 15 N (o/oo)	-0.060	0.606	-0.508	-0.538	1.000	0.571
Delta 13 C (o/oo)	0.189	0.430	-0.376	-0.375	0.571	1.000



Obrázek 1.8: Bodový graf zobrazujúci koreláciu jednotlivých numerických atribútov

Kapitola 2

Úprava dátovej sady pre dolovacie algoritmy

Připravte 2 varianty datové sady vhodné pro dolovací algoritmy. Můžete uvažovat dolovací úlohu uvedenou u datové sady nebo navrhnout vlastní dolovací úlohy. V případě vlastní dolovací úlohy ji specifikujte v dokumentaci. V rámci přípravy datové sady proveďte následující kroky: Odstraňte z datové sady atributy, které jsou pro danou dolovací úlohu irelevantní. Vypořádejte se s chybějícími hodnotami. Pro odstranění těchto hodnot využijte alespoň dvě různé metody pro odstranění chybějících hodnot. Vypořádejte se s odlehlými hodnotami, jsou-li v datové sadě přítomny. Pro jednu variantu datové sady proveďte diskretizaci numerických atributů tak, aby výsledná datová sada byla vhodná pro algoritmy, které vyžadují na vstupu kategorické atributy. Pro druhou variantu datové sady proveďte vhodnou transformaci kategorických atributů na numerické atributy. Dále pak proveďte normalizaci numerických atributů, které má smysl normalizovat. Výsledná datová sada by měla být vhodná pro metody vyžadující numerické vstupy.

2.1 Očistenie datasetov

V rámci tejto časti bolo úlohou vytvoriť dva datasety pre ktoré bolo potrebné odstrániť atribúty, ktoré sú irelevantné pre našu dolovaciu úlohu "Klasifikace druhů tučňáků na základě ostatních atributů nebo shluková analýza.". Bližšie informácie o jednotlivých atribútoch sme si priblížili v tabuľke1.1 z predošlej časti. Z hľadiska našej dolovacej úlohy, sme ako relevantné atribúty zvolili: Voľba atribútu "Species" je z hľadiska dolovacej úlohy zrejmý. Voľba atribútu "Island"

Stĺpec	**Počet chyb. hodnôt**	**Typ**
Species	0	category
Island	0	category
Culmen Depth (mm)	2	float64
Flipper Length (mm)	2	float64
Body Mass (g)	2	int64
Culmen Depth (mm)	2	int64

Tabulka 2.1: Vybrané atribúty pre naše datové sady, ich celkový počet chýbajúcich hodnôt a typ hodnôt v jednotlivých záznamoch.

je vhodná z dôvodu, že by sme dokázali klasifikovať jednotlivé druhy tučniakov na základe výskytu tučniakov na danom ostrove(viď obrázok 1.3). Ako príklad druh "Adelie" je jediný druh tučniaka, ktorý sa vyskytuje na ostrove "Torgersen". Ostatné atribúty, boli volené na základe existencie pomerne jasne rozlíšitelných zhlukov z obrázku1.5, ktoré by boli vhodné na klasifikáciu, ale taktiež aj pre zhlukovú analýzu.

2.2 Chýbajúce hodnoty

V rámci zvolených atribútov existovali len dva záznamy, ktoré obsahovali niekoľko chýbajúcich hodnôt v stĺpcoch Body Mass, Flipper Length, Culmen Length a Culmen Depth. Na eliminovanie týchto hodnôt boli použité dva spôsoby, každý na jeden z dvoch výsledných datasetov. Nakoľko sa v zázname súčasne nevyskytovalo viacero hodnôt bolo viac možností, ako sa s tým vysporiadať. Prvým spôsobom bolo vymazanie záznamu s chybajúcimi hodnotami. Tohoto sme docielili jednoducho pomocou `dropna()`. Druhý spôsob pozostával z metódy binning, tj. nahradenia chýbajúcich hodnôt s priemernou hodnotou daného atribútu pre jednotlivé druhy tučniakov.

2.3 Odľahlé hodnoty

Z kapitoly 1.3 vyplýva, že v datasete sa nenachádzajú významne odľahlé hodnoty.

2.4 Transformácia a diskretizácia

Jeden z vytváraných datasetov mal byť vhodný pro metody vyžadujúce numerické vstupy. Preto bolo potrebné vykonať transformáciu dvoch kategorických atribútov v jednom z našich výstupných datasetov. Konkrétna transformácia hodnôt je popísaná v nasledujúcich dvoch tabuľkách. Ostatné atribúty zostali rovnaké, pretože obsahovali numerické hodnoty.

Island	**Transformovaná numerická hodnota**
Torgersen	0
Dream	1
Biscoe	2

Tabulka 2.2: Transformácia kategorického atribútu špecifikujeho ostrov na ktorom sa jednotlivý tučniaci vyskytli.

Species	**Transformovaná numerická hodnota**
Gentoo	10
Chinstrap	11
Adelie	12

Tabulka 2.3: Transformácia kategorického atribútu špecifikujeho druh tučniaka.

Pre druhý dataset bolo potrebné vytvoriť dataset, ktorý by mal byť vhodný pre algoritmy, ktoré vyžadujú na vstupu kategorické atribúty. Preto bolo potrebné vykonať diskretizáciu štyroch numerických atribútov v druhom z našich výstupných datasetov. Konkrétna diskretizácia hodnôt je prebiehala pomocou pridelovania hodnôt do istého intervalu hodnôt. Interval do ktorého sa hodnoty rozdelovali sa určil pomocou maximálnej a minimálnej hodnoty daného stĺpca. Tento interval bol následne rozdelený do 12 košov/interval, do ktorých sa priradzovala príslušnosť na základe hodnoty daného záznamu. Ostatné kategorické atribúty zostali rovnaké.

Stĺpec	**Kôš 1**	**Kôš 2**	$\cdots$	**Kôš 12**
Flipper Length (mm)	$(171.9, 177.3]$	$(177.364, 182.7]$	$\cdots$	$(225.6, 231.0]$
Body Mass (g)	$(2699.9, 3027.2]$	$(3027.2, 3354.5]$	$\cdots$	$(5972.7, 6300.0]$
Culmen Length (mm)	$(32.0, 34.6]$	$(34.6, 37.1]$	$\cdots$	$(57.1, 59.6]$
Culmen Depth (mm)	$(13.0, 13.8]$	$(13.8, 14.6]$	$\cdots$	$(20.7, 21.5]$

Tabulka 2.4: Kategorizácia numerických atribútov.