2. projekt do předmětu UPA
Příprava dat a jejich popisná charakteristika

Bc. Sebastián Krajňák - vedúci (xkrajn05)
Bc. Dušan Morbitzer (xmorbi00)
Bc. Richard Gajda (xgajda06)

Zima 2022

Obsah

Časť I

Explorativní analýza zvolené datové sady

Kapitola 1

Průzkum jednotlivých atributů datové sady

Celý projekt bol vypracovaný v Python Jupyter Notebooku, pre prácu s dátovou sadou bola použitá knižnica Pandas, pre účely exploratívnej analýzy a zobrazovania dát v grafoch boli použité knižnice Matplotlib a Seaborn.

Dátovú sadu sme si zvolili tučniakov. Sada obsahuje 17 atribútov, pričom cieľový výstup tvoria dve varianty dátovej sady, upravenej a pripravenej pre klasifikáciu druhu tučniakov. Pomocou Pandas metódy `info()` získavame stručné informácie o dátovej sade, viď. Obr. 1.1

```
RangeIndex: 344 entries, 0 to 343
Data columns (total 17 columns):
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
 0   studyName           344 non-null    string
 1   Sample Number       344 non-null    int64
 2   Species             344 non-null    string
 3   Region              344 non-null    string
 4   Island              344 non-null    string
 5   Stage               344 non-null    string
 6   Individual ID       344 non-null    string
 7   Clutch Completion   344 non-null    string
 8   Date Egg            344 non-null    string
 9   Culmen Length (mm)  342 non-null    float64
 10  Culmen Depth (mm)   342 non-null    float64
 11  Flipper Length (mm) 342 non-null    float64
 12  Body Mass (g)       342 non-null    float64
 13  Sex                 334 non-null    string
 14  Delta 15 N (o/oo)   330 non-null    float64
 15  Delta 13 C (o/oo)   331 non-null    float64
 16  Comments            26 non-null     string
dtypes: float64(6), int64(1), string(10)
memory usage: 45.8 KB
None
```

Obr. 1.1: Stručné informácie o dátovej sade

Celkovo sada obsahuje 344 záznamov, pričom je viditeľné, že posledných atribútov obsahuje chýbajúce alebo zašumené hodnoty, ktoré budú opravené v neskoršej kapitole. Prevažná väčšina atribútov je dátového typu `string` (Pandas automaticky inferuje dáta typu `string` ako vlastný dátový typ `object`, pre klasifikačný algoritmus sme avšak hneď po načítaní dát previedli atribúty do požadovaného dátového typu), druhé najväčšie zastúpenie má dátový typ `float64`.

Pomocou Pandas metódy `describe()` získame popisné štatistiky dát, pričom hodnoty `NaN` sú vypustené. Všetky popisné štatistiky sú zobrazené v Tabuľke 1.1 a Tabuľke 1.2.

	Sample Number	Culmen Length	Culmen Depth	Flipper Length	Body Mass
count	344.000000	342.000000	342.000000	342.000000	342.000000
mean	63.151163	43.921930	17.151170	200.915205	4201.754386
std	40.430199	5.459584	1.974793	14.061714	801.954536
min	1.000000	32.100000	13.100000	172.000000	2700.000000
25%	29.000000	39.225000	15.600000	190.000000	3550.000000
50%	58.000000	44.450000	17.300000	197.000000	4050.000000
75%	95.250000	48.500000	18.700000	213.000000	4750.000000
max	152.000000	59.600000	21.500000	231.000000	6300.000000

Tabuľka 1.1: Popisné štatistiky dát 1/2

	Delta 15 N	Delta 13 C
count	330.000000	331.000000
mean	8.733382	-25.686292
std	0.551770	0.793961
min	7.632200	-27.018540
25%	8.299890	-26.320305
50%	8.652405	-25.833520
75%	9.172123	-25.062050
max	10.025440	-23.787670

Tabuľka 1.2: Popisné štatistiky dát 2/2

Kapitola 2

Průzkum rozložení hodnot atributů

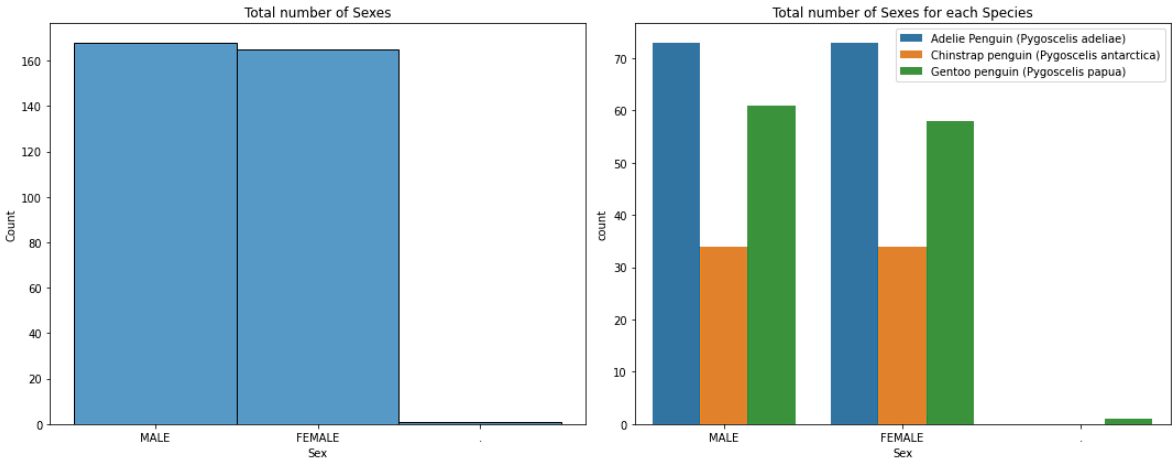
K průzkumu rozložení dat využijeme různé grafy pro jejich vykreslení za využití knihoven Matplotlib a Seaborn. Nejprve jsme se zaměřili na rozložení pohlaví v datech (viz. Obr. 2.1). Lze vidět, že zastoupení obou pohlaví je téměř stejné, avšak si lze povšimnout, že zastoupení jednotlivých druhů tučňáků je ve značném nepoměru. Také si lze povšimnout chyby v datech a to pohlaví označené ".". Z grafu 2.5 lze odhadnout, že se jedná o samici rodu Gentoo, protože se v datasetu nachází téměř stejný poměr samců a samic Gentoo (viz Obr. 2.1) a i když je samců Gentoo nepatrně více, tak chybný záznam by ještě spadal do rozmezí 1. a 3. kvartilu samic Gentoo. Naopak pokud by se jednalo o samce Gentoo, pak by se chybný záznam nacházel na hranici minima a jednalo by se téměř o odlehlou hodnotu.

Dále jsme se zaměřili na poměr tučňáků z jednotlivých ostrovů. V grafu 2.2 vidíme, že jejich poměry jsou značně nevyvážené.

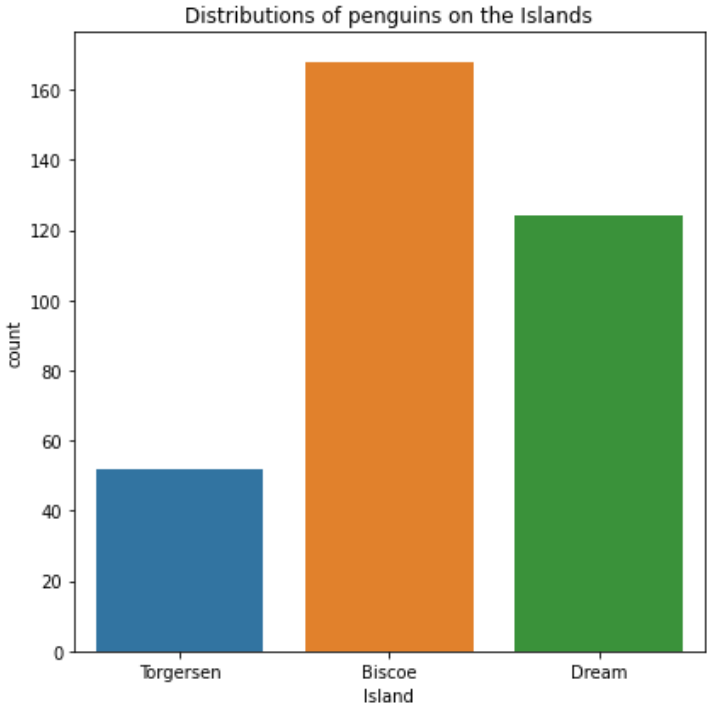
V dalším grafu 2.3 jsme se zaměřili na závislost mezi velikostí ploutve a tělesné hmotnosti. Lze si povšimnout, že samci i samice rodu Gentoo jsou výrazně těžší a mají větší ploutve, než tučňáci zbylých dvou druhů.

Graf 2.4 závislosti délky a výšky zobáku odhalil, že velikost zobáku významně nerozlišuje samce od samic s průměrnými velikostmi zobáku, ale lze usoudit, že pokud má tučňák podprůměrně malý zobák, jedná se pravděpodobněji o samici a naopak nadprůměrnými velikostmi se vyznačují spíše samci.

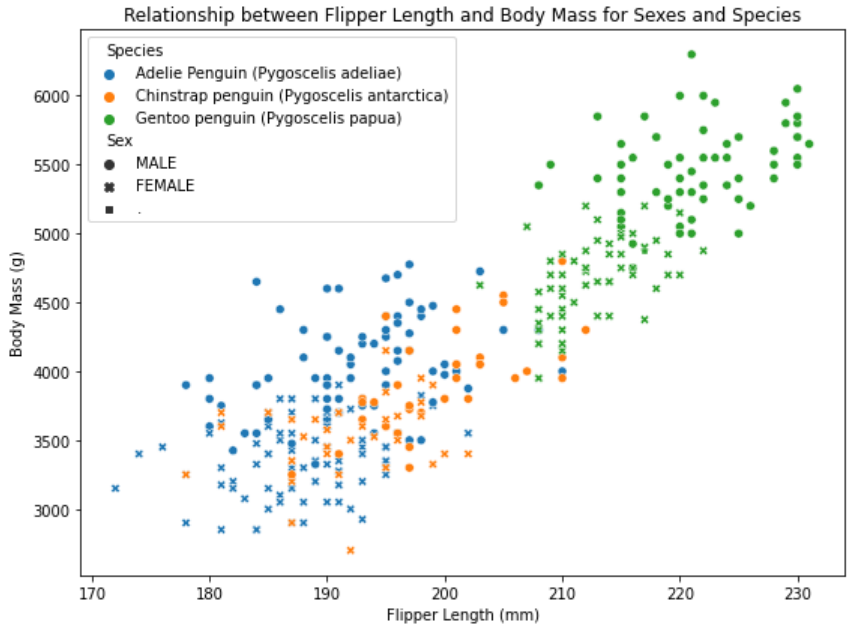
V poslední řadě jsme pozorovali rozložení tělesné hmotnosti vůči pohlaví a druhu (viz. Obr. 2.5). Lze si povšimnout, že podobně jako při porovnvání velikostí ploutví se jedinci druhu Gentoo vyznačují větší tělesnou hmotností, než ostatní dva druhy. Zároveň nám tento graf spolu s grafem 2.1 napomohl k odhadu pohlaví chybného záznamu, jak bylo již popsáno výše v kapitole.



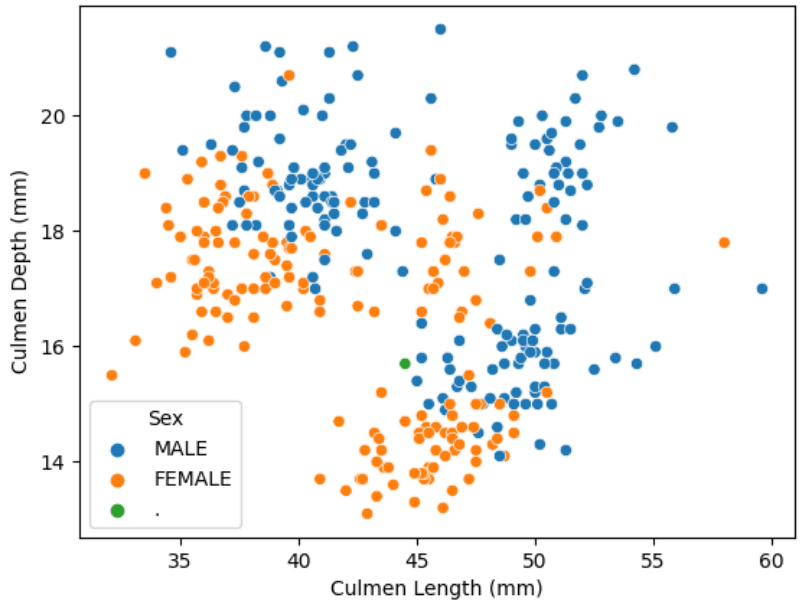
Obr. 2.1: Grafy rozložení pohlaví



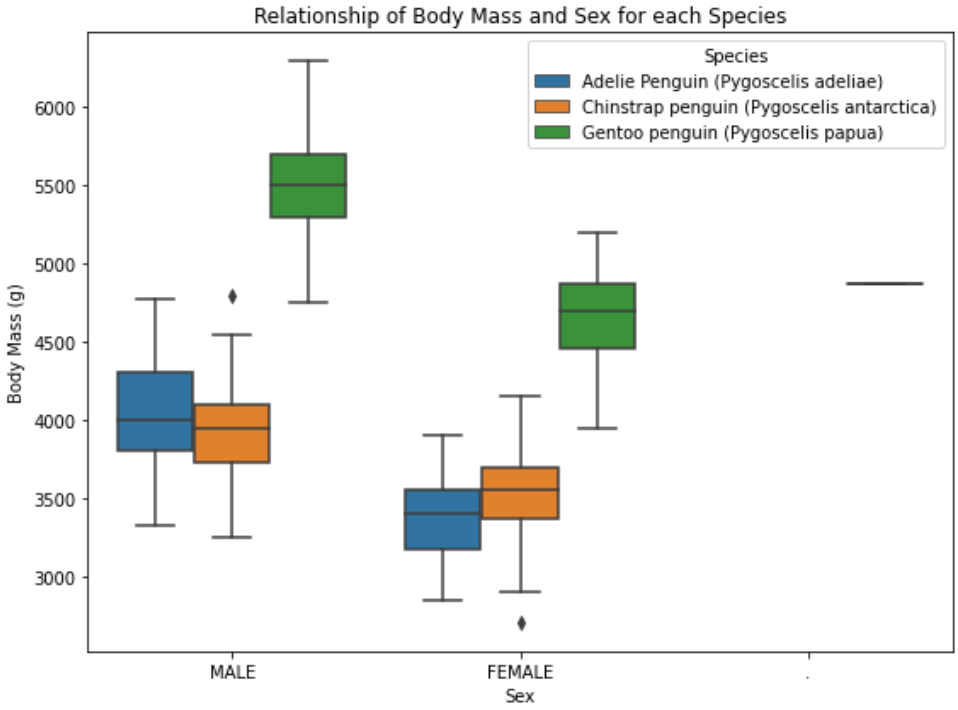
Obr. 2.2: Graf rozložení tučňáků z jednotlivých ostrovů



Obr. 2.3: Graf závislosti velikosti ploutve na celkové hmotnosti podle pohlaví i druhu



Obr. 2.4: Graf závislosti délky zobáku na jeho výšce mezi pohlavími

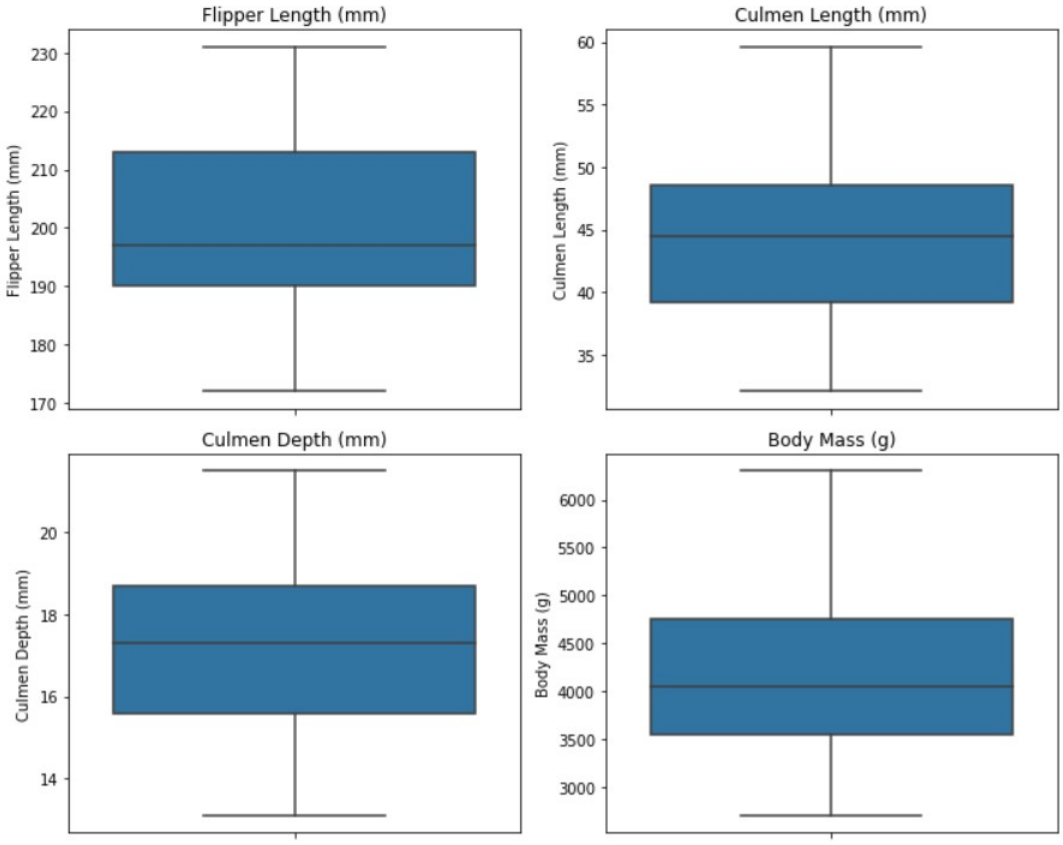


Obr. 2.5: Graf závislosti celkové hmotnosti na pohlaví a druhu

Kapitola 3

Nalezení odlehlých hodnot

Keďže pri klasifikácii budeme potrebovať hlavne 4 numerické atribúty, tj. Flipper Length (mm), Culmen Length (mm), Culmen Depth (mm) a Body Mass (g), identifkáciu odľahlých hodnôt môžeme vykonať jednoducho a to použitím krabicových grafov, zobrazených na Obr. 3.1. Z grafov je viditeľné, že atribúty neobsahujú žiadne odľahlé hodnoty.



Obr. 3.1: Krabicové grafy numerických atribútov

Kapitola 4

Analýza chybějících hodnot

Ako bolo viditeľné na výpise metódy `info()` v Kapitole 1 v dátovej sade sa nachádza značné množstvo chýbajúcich hodnôt, najmä v posledných 7 atribútoch, pričom najviac chýbajúcich hodnôt sa nachádza v atribúte Comments, obsahujúci komentáre, jednotlivým meraniam. Pomocou Pandas metódy `isnull().sum()` sme schopný získať sumy chýbajúcich hodnôt jednotlivých atribútov, konkrétne

- `Culmen Length (mm)` 2
- `Culmen Depth (mm)` 2
- `Flipper Length (mm)` 2
- `Body Mass (g)` 2
- `Sex` 10
- `Delta 15 N (o/oo)` 14
- `Delta 13 C (o/oo)` 13
- `Comments` 318

zo zoznamu boli odstránené, narozdiel od výpisu v priloženom Jupyter Notebooku, všetky atribúty, ktoré obsahovali 0 chýbajúcich hodnôt. Z výnimkou na atribúty Sex a Comments, ktoré sú typu `string`, všetky atribúty, kde chýbajú hodnoty majú typ `float64`. Metódam opravenia chýbajúcich hodnôt, sa venujeme v neskoršej kapitole, druhej časti dokumentácie.

Kapitola 5

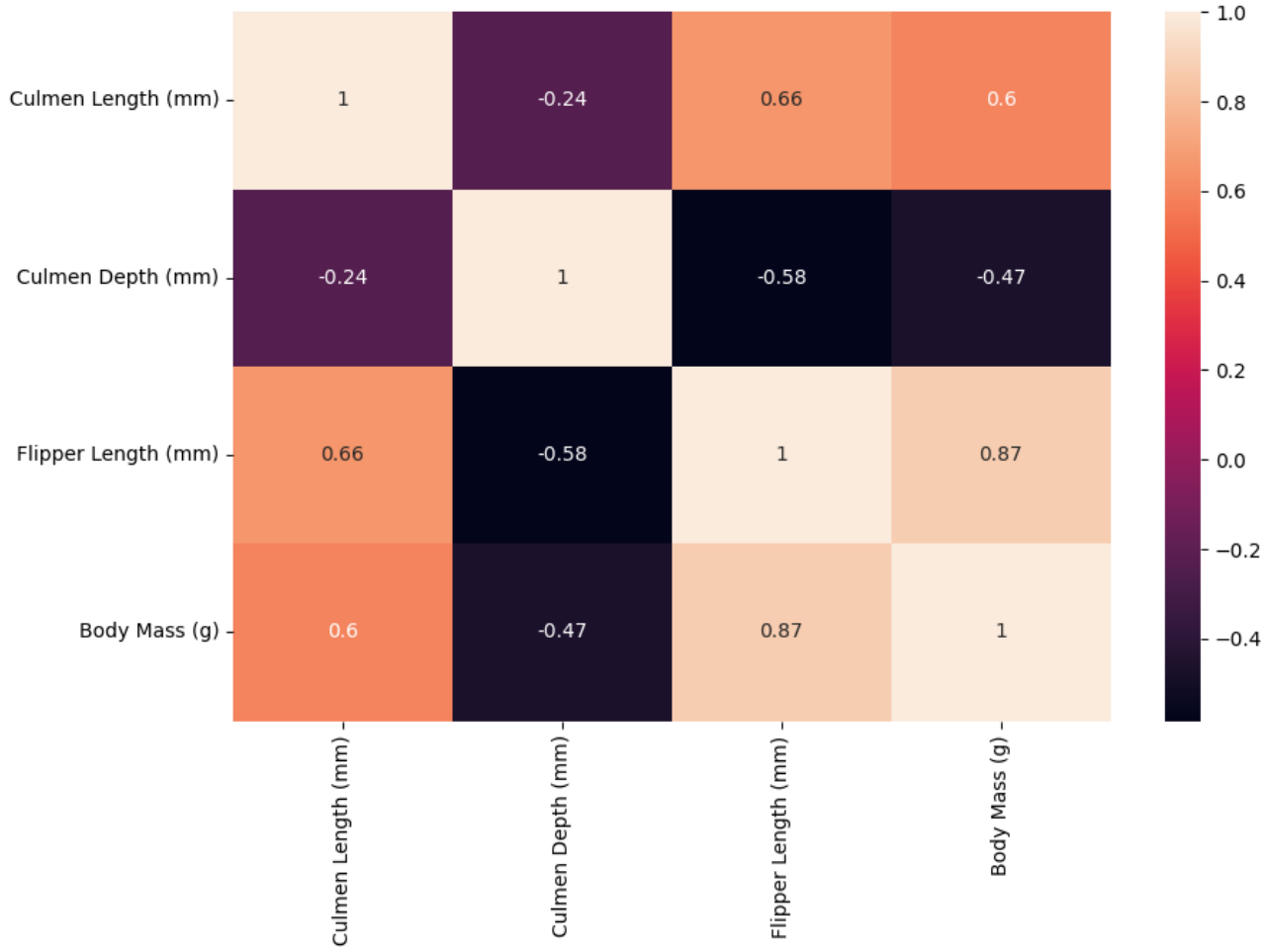
Korelační analýza numerických atributů

Pred analýzou korelácie numerických atribútov boli najprv odstránené nepotrebné atribúty popísané v Kapitole 6. Numerické atribúty Delta 15 N (o/oo) a Delta 13 C (o/oo), nebudú pri klasifikácii použité a ich prítomnosť značne a zbytočne zväčšuje grafy použité pri korelačnej analýze.

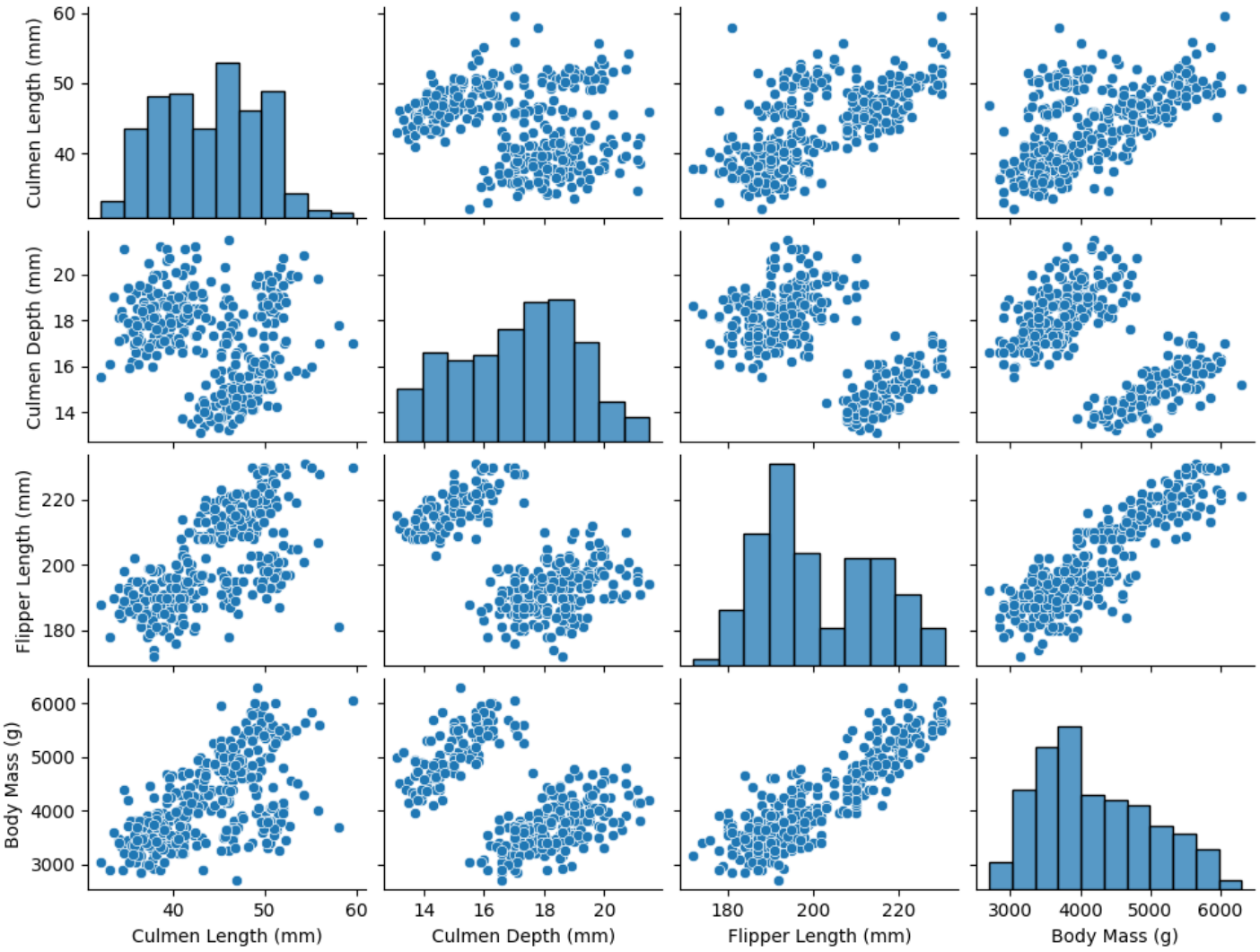
Pri korelačnej analýze boli použité dva typy grafov, pomocou metód poskytnutých knižnicou Seaborn, matica bodových grafov a korelačná mapa, zobrazené na nasledujúcich obrázkoch.

Z korelačnej mapy (viď. Obr. 5.1) je zrejmé, že v jednotlivé atribúty majú nielen pozitívnu napr. Culmen Length (mm) a Flipper Length (mm) ale aj negatívnu koreláciu napr. Culmen Depth (mm) a Flipper Length (mm). Žiadne 2 atribúty nekorelujú úplne avšak značne vysoká miera korelácie sa nachádza medzi atribútmi Body Mass (g) a Fillper Length (mm), čo môže byť vhodné pri neskoršej klasifikácii. Tomu že na seba tieto dva atribúty korelujú zodpovedá aj biologická stránka, keďže plutvy (flippers) tučniaka sú skoro rovnako dlhé ako jeho výška samotná.

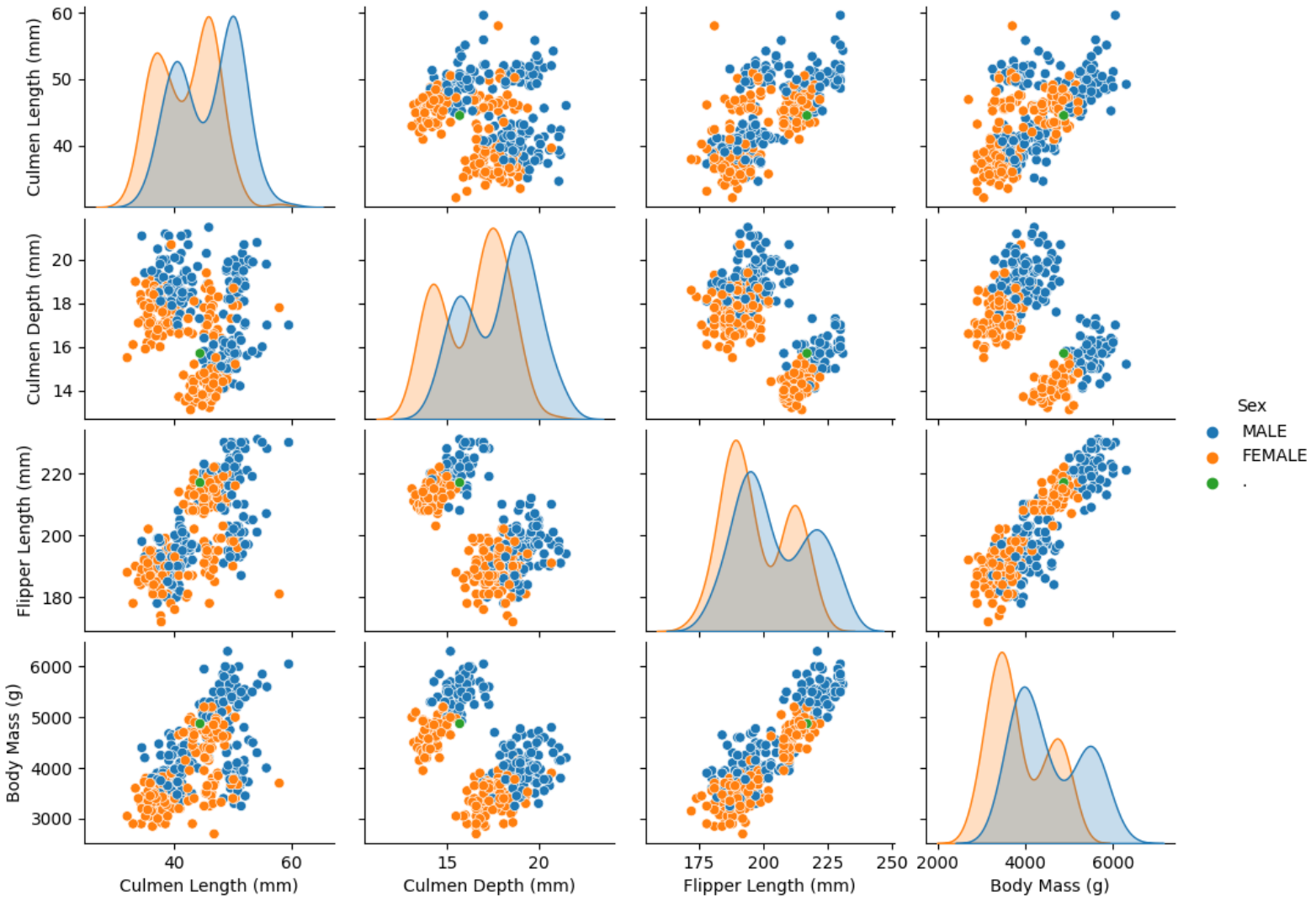
Na obrázkoch Obr. 5.2 až Obr. 5.4 sú zobrazené matice bodových grafov pre jednotlivé atribúty, podľa korelačnej mapy sa najväčšia pozitívna korelácia nachádza medzi Body Mass (g) a Flipper Length (mm), čo je viditeľné na Obr. 5.2. Ak sa však zameriame na túto koreláciu a pozrieme sa na Obr. 5.3 a Obr. 5.4 vidíme, že sú to práve samci a rodu Gentoo (Pygoscelis Papua), ktorý dosahujú väčším váham a dlhším plutvám.



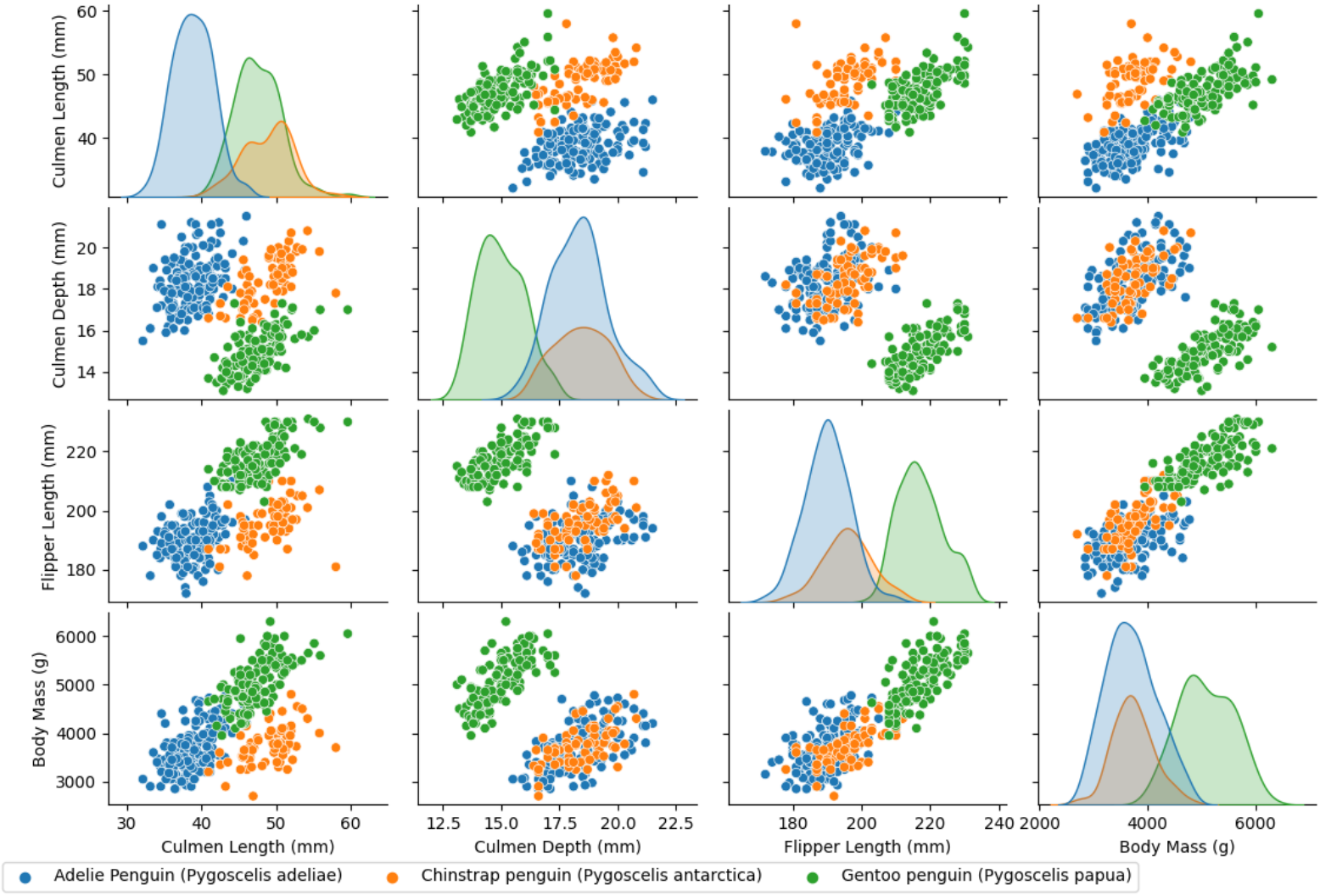
Obr. 5.1: Korelačná mapa



Obr. 5.2: Matica bodových grafov pre čisto numerické atribúty



Obr. 5.3: Matica bodových grafov doplnená kategorickým atribútom Sex



Obr. 5.4: Matica bodových grafov doplnená kategorickým atribútom Species

Časť II

Příprava variant datové sady pro dolovací algoritmus

Kapitola 6

Odstranění nepotřebných atributů

Vzhledem k tomu, že nás zajímají pouze některé atributy, můžeme zbytek atributů v této fázi odstranit. Atributy Stage a Region jsou odstraněny implicitně, protože obsahují pouze jednu unikátní hodnotu.

Odstraněny byly tyto atributy:

- Delta 15 N (o/oo)
- Delta 13 C (o/oo)
- Comments
- Sample Number
- studyName
- Individual ID
- Date Egg
- Clutch Completion

V datech nám tedy zůstávají atributy:

- Body Mass
- Flipper Length
- Culmen Depth
- Culmen Length
- Sex
- Species
- Island

Kapitola 7

Oprava chybějících a zašumených hodnot

Vzhledem k tomu, že se v datech nenacházejí odlehlé hodnoty a počet chybějících hodnot je poměrně malý, můžeme chybějící hodnoty jednoduše doplnit, což bude při klasifikaci prospěšné.

V předchozí kapitole jsme provedli zahození nepotřebných atributů, které mnohdy měly chybějící hodnoty (např. Comments). Tímto způsobem jsme se taky zbavili většího množství chybějících hodnot, kde by bylo zbytečné pokoušet se tyto hodnoty doplnit jakýmkoli způsobem.

Ve zbylých 5 atributech jsme chybějící hodnoty doplnili způsobem odpovídajícím jejich typu. Numerické hodnoty byly doplněny vypočítanou střední hodnotou. U kategorických atributů bylo doplnění provedeno na základě výpočtu modusu.

V datech se nachází jen jedna zašuměná hodnota a to pohlaví `"."` u atributu Sex, na základě přezkoumání grafů v Kapitole 2 jsme tuto zašuměnou hodnotu opravili na hodnotu FEMALE.

Kapitola 8

Vypořádání se s odlehlými hodnotami

V datové sadě se nevyskytovaly žádné odlehlé hodnoty (viz. Obr. 3.1).

Kapitola 9

Diskretizace numerických atributů

Pro použití v algoritmech vyžadujících kategorické atributy byla provedena diskretizace numerických atributů. Jedná se o atributy Culmen Length (mm), Culmen Depth (mm), Flipper Length (mm) a Body Mass (g). U jednotlivých atributů je kategorizace prováděna na základě příslušnosti do numerického intervalu pomocí funkce `cut()` knihovny Pandas. Pro Flipper Length a Culmen Depth atributy byl počet intervalů stanoven na 6, u Culmen Length na 7 a u Body Mass je počet intervalů stanoven na 8, přičemž jednotlivé intervaly jsou rovnoměrně rozloženy na oboru hodnot. Počty intervalů byly stanoveny vzhledem k rozpětí hodnot jednotlivých atributů.

Výsledná popisná tabulka atributů vypadá následovně:

	Species	Island	Culmen Length (mm)	Culmen Depth (mm)	Flipper Length (mm)	Body Mass (g)	Sex
count	344	344	344	344	344	344	344
unique	3	3	7	6	6	8	2
top	Adelie Penguin (Pygoscelis adeliae)	Biscoe	(44, 48]	(17.5, 19.0]	(181, 191]	(3600.0, 4050.0]	MALE
freq	152	168	80	104	92	79	178

Obr. 9.1: Popisná tabulka kategorických atributů

Kapitola 10

Transformace kategorických atributů na numerické

Datová sada obsahuje 3 kategorické atributy Species, Island a Sex, které je potřeba transformovat na numerické. Transformace je provedena pomocí metody `factorise()` knihovny Pandas. Tyto hodnoty nedává smysl normalizovat, protože jsou to diskrétní hodnoty vyjadřující náležitost k daným třídám.

Zbylé numerické atributy však normalizovat dává smysl pro usnadnění data miningu z výsledného datasetu. Vzhledem k tomu, že žádný z atributů neobsahuje odlehlé hodnoty, využili jsme min-max normalizaci.

Výsledná popisná tabulka normalizovaných numerických atributů pak vypadá takto:

	Species	Island	Culmen Length	Culmen Depth	Flipper Length	Body Mass	Sex
count	344.000000	344.000000	344.000000	344.000000	344.000000	344.000000	344.000000
mean	1.918605	2.209302	0.429888	0.482282	0.490088	0.417154	1.479651
std	0.893320	0.684970	0.197951	0.234408	0.237638	0.222115	0.500313
min	1.000000	1.000000	0.000000	0.000000	0.000000	0.000000	1.000000
25%	1.000000	2.000000	0.260909	0.297619	0.305085	0.236111	1.000000
50%	2.000000	2.000000	0.441818	0.500000	0.423729	0.375000	1.000000
75%	3.000000	3.000000	0.596364	0.666667	0.694915	0.569444	2.000000
max	3.000000	3.000000	1.000000	1.000000	1.000000	1.000000	2.000000

Obr. 10.1: Popisná tabulka normalizovaných numerických atributů