

Príprava dát a ich popisná charakteristika

Bc. Michal Šedý

Bc. Martina Chrípková

Bc. Martin Novotný Mlinárčsik

19. novembra 2022

1 Úvod

Cieľom druhého projektu z predmetu Ukladanie a príprava dát je príprava dát a ich popisná charakteristika. Vstupom bude dátová sada¹ vytvorená pozorovaním troch druhov tučniakov na Antarktíde. Výstup bude tvorený analýzou tohto datasetu a úpravou datasetu do podoby, ktorá bude vhodná pre zadanú dolovaciu úlohu.

2 Exploratívna analýza

2.1 Skúmanie atribútov

Numerické dáta

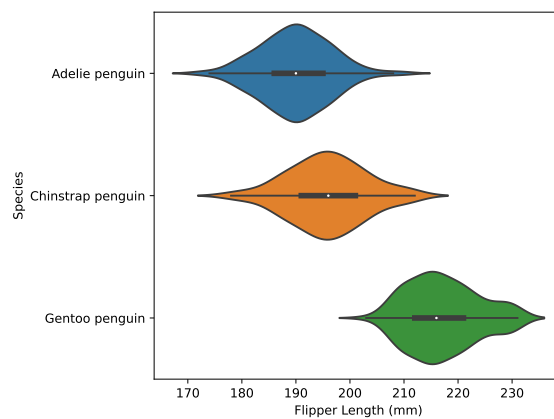
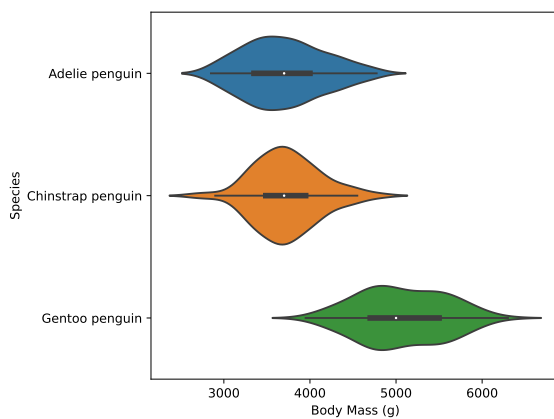
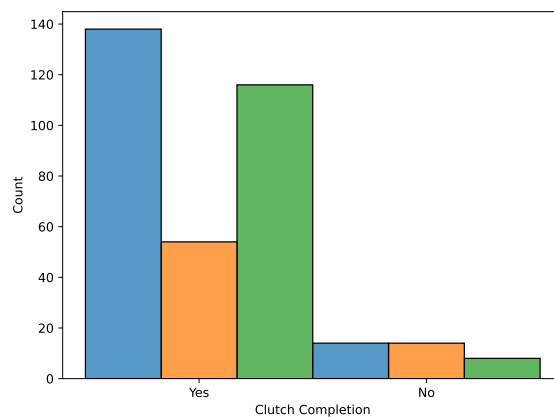
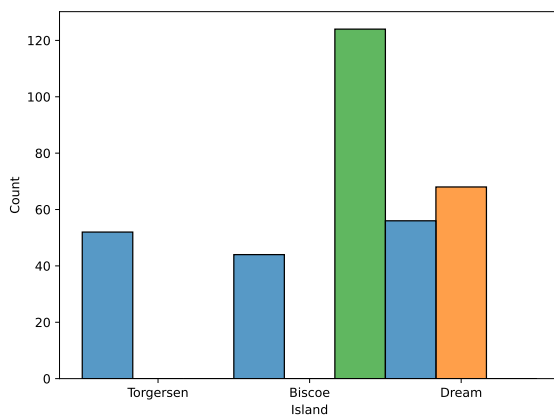
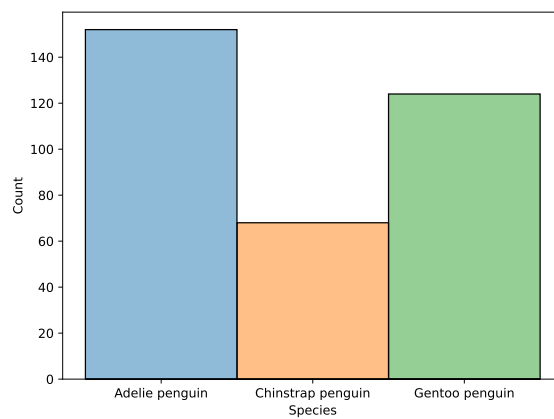
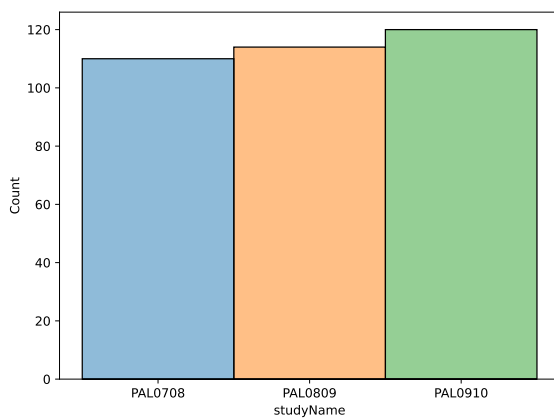
Stĺpec	Typ	Počet hodnôt	Priemerná hodnota	Rozsah hodnôt
Sample Number	int64	344	63.151	1 - 152
Culmen Length (mm)	float64	342	43.921	32.1 - 59.6
Culmen depth (mm)	float64	342	17.151	13.1 - 21.5
Flipper Length (mm)	float64	342	200.915	172 - 231
Body Mass (g)	float64	342	4201.754	2700 - 6300
Delta 15 N (ooo)	float64	330	8.733	7.632 - 10.025
Delta 13 C (ooo)	float64	331	-25.686	-27.018 - -23.787

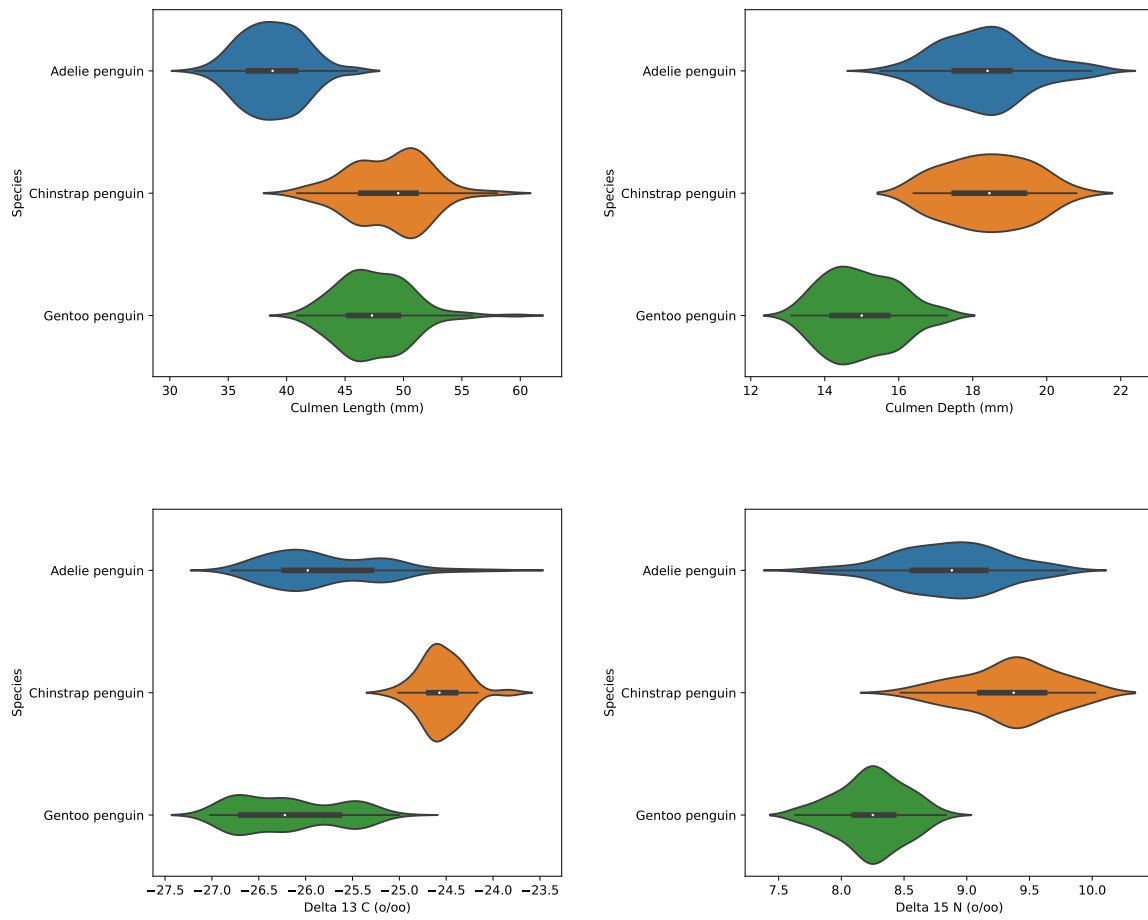
Kategorické dáta

Stĺpec	Typ	Počet hodnôt	Najčastejšia hodnota
studyName	object	3	PAL0910
Species	object	3	Adelie Penguin
Region	object	1	Anvers
Island	object	3	Biscoe
Stage	object	1	Adult, 1 Egg Stage
Individual ID	object	DOPLNIT	N61A2
Clutch Completion	object	2	Yes
Date Egg	object	DOPLNIT	11/27/07
Sex	object	2	Male

¹<https://www.kaggle.com/datasets/parulpandey/palmer-archipelago-antarctica-penguin-data>

2.1.1 Grafy atribútov





Grafy Stage, Individual ID a Date Egg sú vynechané, keďže neposkytujú dôležité informácie z pohľadu dolovania.

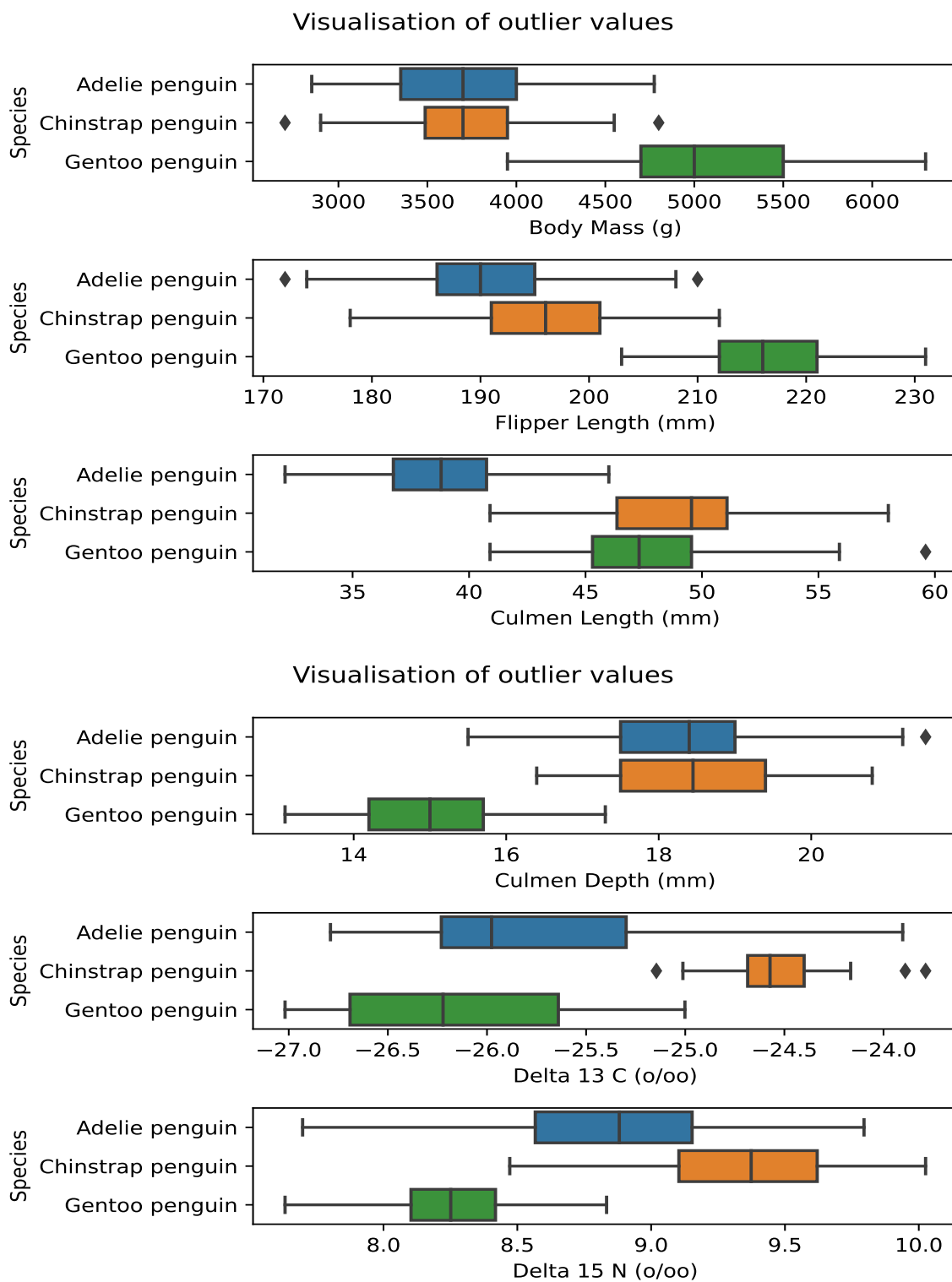
2.2 Rozloženie hodnôt atribútov

Na vizualizáciu rozloženia hodnôt atribútov bola použitá matica grafov.



2.3 Odľahlé hodnoty

Na skúmanie odľahlých hodnôt môžeme využiť krabicový graf.



Z grafov si môžeme všimnúť, že každá populácia tučniakov obsahuje odľahlé hodnoty pri nejakom atribúte. Atribút **Delta 15 N** neobsahuje odľahlé hodnoty.

2.4 Chýbajúce hodnoty

Chýbajúce atribúty pri položkách datasetu

ID	Počet chýbajúcich hodnôt	Chýbajúce hodnoty
0	2	Delta 15 N, Delta 13 C
3	7	Všetky numerické hodnoty, Sex
8	3	Sex, Delta 15 N, Delta 13 C
9	1	Sex
10	1	Sex
11	3	Sex, Delta 15 N, Delta 13 C
12	2	Delta 15 N, Delta 13 C
13	2	Delta 15 N, Delta 13 C
15	2	Delta 15 N, Delta 13 C
39	2	Delta 15 N, Delta 13 C
41	2	Delta 15 N, Delta 13 C
46	2	Delta 15 N, Delta 13 C
47	3	Sex, Delta 15 N, Delta 13 C
212	1	Delta 15 N
246	1	Sex

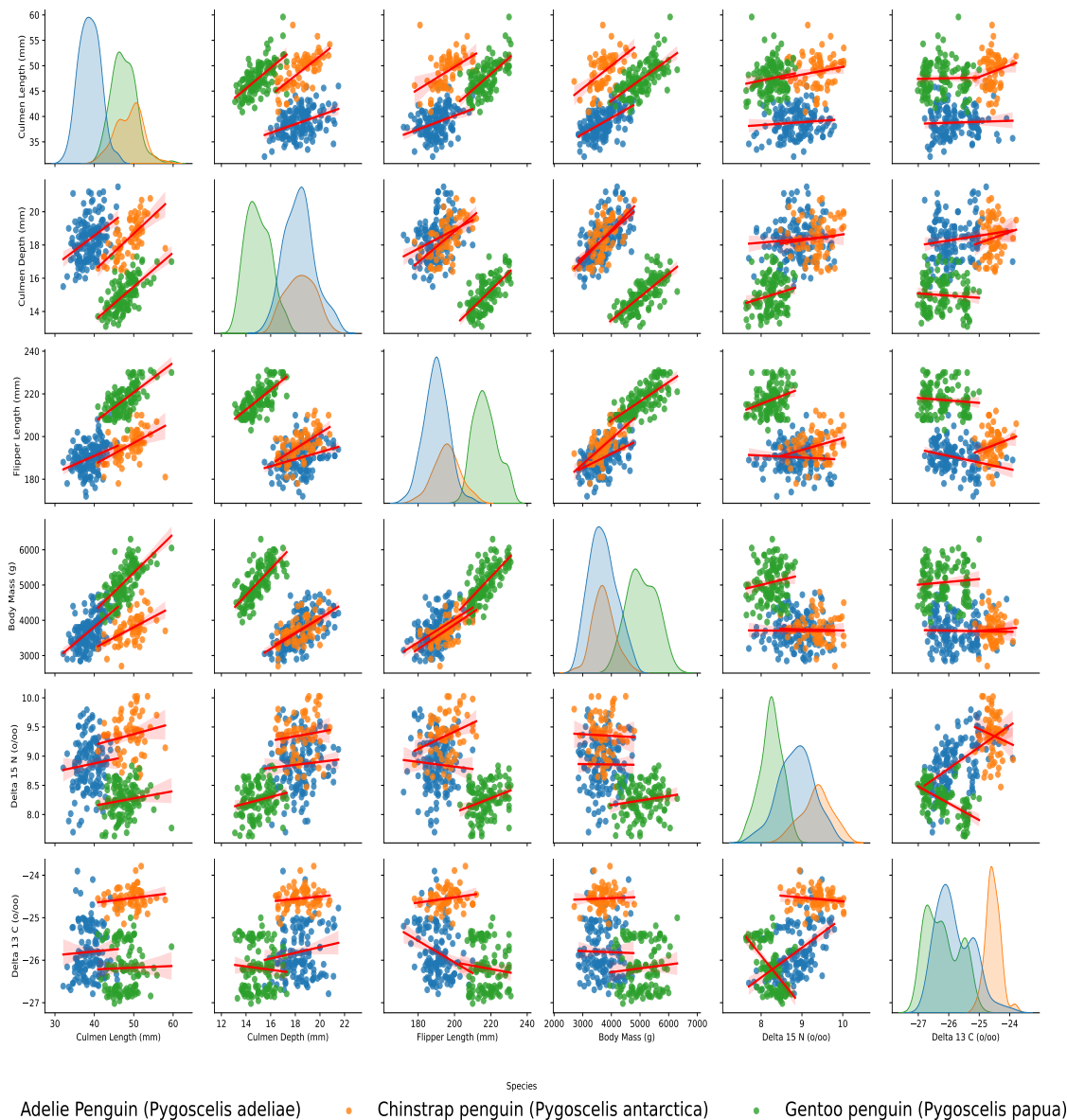
Počet chýbajúcich hodnôt jednotlivých sérií

Séria	Počet chýbajúcich hodnôt
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

2.5 Korelačná analýza

Pre potreby korelačnej analýzy môžeme využiť graf zo sekcie 3 s pridaním korelačných priamok.

Môžeme si všimnúť, že fyzické atribúty tučniakov (**Body mass**, **Flipper Length**, **Culmen Depth** a **Culmen Length**) majú medzi sebou pozitívnu koreláciu. Z grafov vyplýva, že atribúty **Delta 15 N** a **Delta 13 C** nie sú závislé od fyzických atribútov spomenutých vyššie.



3 Úprava dátovej sady

Ako dolovacia úloha bola zvolená zhluková analýza. Pri bližšom skúmaní grafov regresie si môžeme všimnúť, že niektoré grafy ukazujú náznaky zhlukovania bodov v grafe. Keďže tieto grafy vizualizujú závislosti medzi fyzickými atribútmi tučniakov, môžeme jednotlivým tučniakom priradiť druh podľa ich fyzických atribútov a podľa vzájomných závislostí medzi týmito atribútmi.

3.1 Odstránenie nepotrebných hodnôt

Prvá varianta dátovej sady bude pracovať s kategorickými dátami. Numerické hodnoty v dátovej sade sú diskretizované za použitia ”**binning**” metódy, pričom jednotlivé intervaly sú dané hodnotami kvartilov pre daný stĺpec. Výsledkom algoritmu je súbor **discrete_dataset.csv**. Ukážka výslednej dátovej sady:

Species	Island	Clutch Completion	Culmen Length (mm)	Culmen Depth (mm)
Adelie penguin	Torgersen	Yes	B	D
Adelie penguin	Torgersen	Yes	B	C
Adelie penguin	Torgersen	Yes	B	D
Adelie penguin	Torgersen	Yes	B	D
Adelie penguin	Torgersen	Yes	A	E
Adelie penguin	Torgersen	Yes	B	E

Druhá varianta dátovej sady bude pracovať s numerickými dátami. Kategorické dáta sú za pomoci kódovania prevedené na číselné hodnoty v intervale od 0 po 1. Následné sú všetky numerické dáta normalizované za použitia **MinMax normalizácie**, ktorá taktiež normalizuje hodnoty na hodnoty v intervale 0 až 1. Výsledkom algoritmu je súbor **numeric_normalized_dataset.csv**. Ukážka výslednej dátovej sady:

Species	Island	Clutch Completion	Culmen Length (mm)	Culmen Depth (mm)
0.0	0.0	0	0.2702702702702703	0.691358024691358
0.0	0.0	0	0.2857142857142857	0.5308641975308641
0.0	0.0	0	0.31660231660231647	0.6049382716049383
0.0	0.0	0	0.25835485438134437	0.64769847109803
0.0	0.0	0	0.17760617760617767	0.7654320987654323
0.0	0.0	0	0.2779922779922778	0.9259259259259262