Příprava dat a jejich popisná charakteristika

František Koleček, xkolec08@stud.fit.vut.cz
Tomáš Moravčík, xmorav41@stud.fit.vut.cz
David Sladký, xsladk07@stud.fit.vut.cz

zima 2022

Obsah

Kapitola 1

Explorativní analýza

1.1 Zadání

`penguins_size.csv` je datová sada obsahující údaje o populaci tučnáků na ostrovech u pobřeží Antarktidy. Cílem této části je provést explorativní analýzu, neboli prozkoumaní atributů datové sady a jejich rozložení, odhalit odlehlé a chybějící hodnoty a provést korelační analýzu.

1.2 Nástroje

Pro prozkoumání datové sady byl zvolen programovací jazyk Python. Konrétně byly využity knihovny pandas, matplotlib a seaborn. Tyhle knihovny je možné doinstalovat příkazem `pip install pandas matplotlib seaborn`. Script provádějící explorativní analýzu je uloženv souboru `src/exploration.py`, který na svém vyžaduje jeden parametr a to cestu k datasetu. Tento skript vydává textový výstup, tak i řadu grafů.

Příklad spuštění:
`python3 ./src/exploration.py ./dataset/penguins_size.csv`
Nebo pomocí Makefile:
`make exp`

1.3 Prozkoumání atributů datové sady

Základní popis kategorických atributů.

	Druh	Ostrov	Pohlaví
Počet	344	344	334
Unikátních	3	3	3
Nejčastější	Adelie	Biscoe	MALE
Frekvence	152	168	168

Počet jedinců jednotlivých druhů.

Jméno	Počet	Zastoupení
Adelie	152	44.2%
Gentoo	124	36.0%
Chinstrap	68	19.8%

Počet tučňáků na jednotlivých ostrovech

Jméno	Počet	Zastoupení
Biscoe	168	48.8%
Dream	124	36.0%
Torgersen	52	15.1%

Rozložení pohlaví tůčňáků

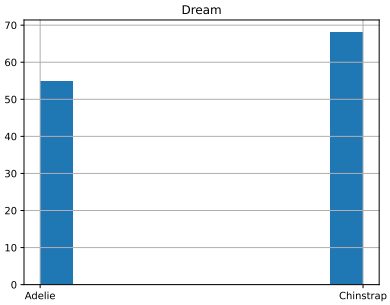
Pohlaví	Počet	Zastoupení
MALE	168	50.3%
FEMALE	165	49.4%
.	1	0.3%

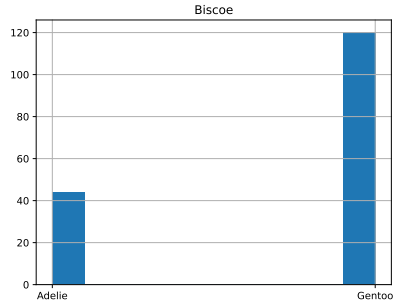
Vlastnosti číselných atributů

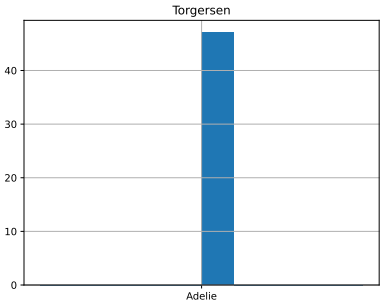
	culmen length mm	culmen depth mm	flipper length mm	body mass g
count	342.00	342.00	342.00	342.00
mean	43.92	17.15	200.91	4201.75
std	5.45	1.97	14.06	801.95
min	32.10	13.10	172.00	2700.00
25%	39.22	15.60	190.00	3550.00
50%	44.45	17.30	197.00	4050.00
75%	48.50	18.70	213.00	4750.00
max	59.60	21.50	231.00	6300.00

1.4 Rozložení atributů

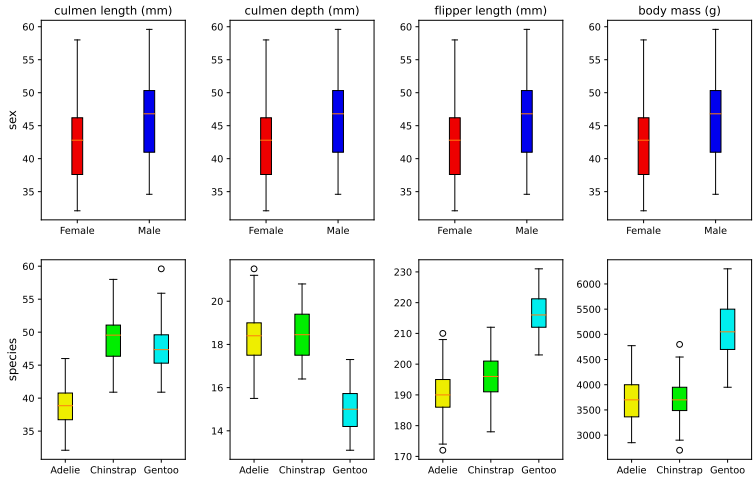
Zastoupení jednotlivých druhů na ostrovech



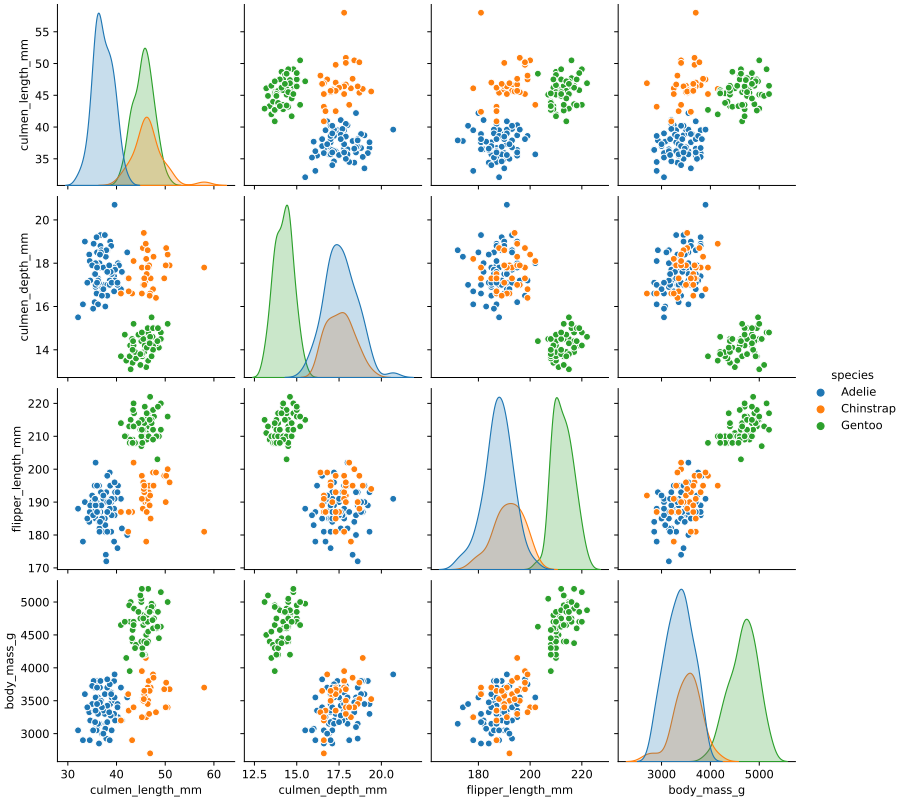




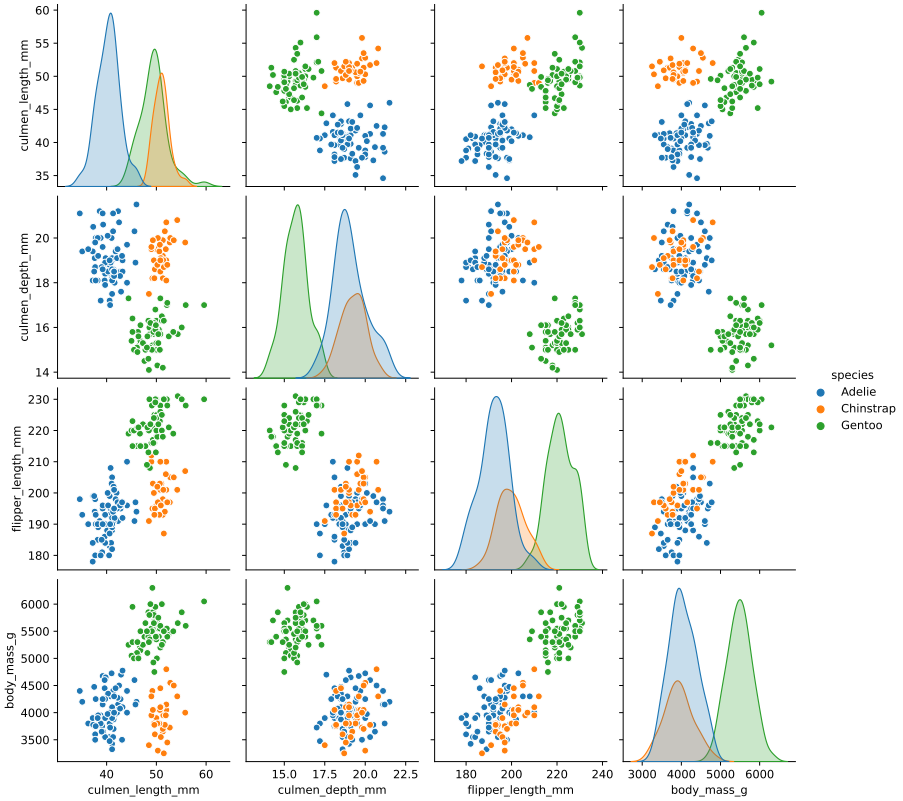
Rozložení numerických atributů v závisloti na pohlaví a druhu



Podrobnější rozložení numerických atributů v závisloti na pohlaví - Female



Podrobnější rozložení numerických atributů v závisloti na pohlaví - Male



1.5 Odlehlé hodnoty

Několik odlehlých hodnot jde vyčíst z krabicového grafu v předešlé kapitoly.

- 1 hodnota pro culmen length druhu Gentoo
- 1 hodnota pro culmen depth druhu Adelie
- 2 hodnoty pro flipper length druhu Adelie
- 2 hodnoty pro body mass druhu Chinstrap

Další odlehlé hodnoty lze identifikovat v maticových grafech a to především pro Female v řádku culmen length.

1.6 Chybějící hodnoty

Atribut	Počet
species	0
island	0
culmen length mm	2
culmen depth mm	2
flipper length mm	2
body mass g	2
sex	10

Jednotlivé řádky s chybjejícími hodnotami

Řádek 3	
Atribut	Počet
species	Adelie
island	Torgersen
culmen length mm	NaN
culmen depth mm	NaN
flipper length mm	NaN
body mass g	NaN
sex	NaN

Řádek 8	
Atribut	Počet
species	Adelie
island	Torgersen
culmen length mm	34.1
culmen depth mm	18.1
flipper length mm	193.0
body mass g	3475.0
sex	NaN

Řádek 9	
Atribut	Počet
species	Adelie
island	Torgersen
culmen length mm	42.0
culmen depth mm	20.2
flipper length mm	190.0
body mass g	4250.0
sex	NaN

Řádek 10	
Atribut	Počet
species	Adelie
island	Torgersen
culmen length mm	37.8
culmen depth mm	17.1
flipper length mm	186.0
body mass g	3300.0
sex	NaN

Řádek 11	
Atribut	Počet
species	Adelie
island	Torgersen
culmen length mm	37.8
culmen depth mm	17.3
flipper length mm	180.0
body mass g	3700.0
sex	NaN

Řádek 47	
Atribut	Počet
species	Adelie
island	Dream
culmen length mm	37.5
culmen depth mm	18.9
flipper length mm	179.0
body mass g	2975.0
sex	NaN

Řádek 246	
Atribut	Počet
species	Gentoo
island	Biscoe
culmen length mm	44.5
culmen depth mm	14.3
flipper length mm	216.0
body mass g	4100.0
sex	NaN

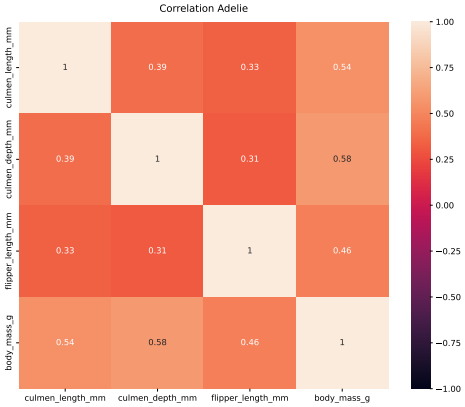
Řádek 286	
Atribut	Počet
species	Gentoo
island	Biscoe
culmen length mm	46.2
culmen depth mm	14.4
flipper length mm	214.0
body mass g	4650.0
sex	NaN

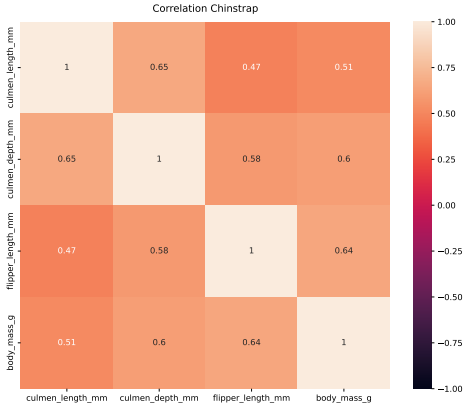
Řádek 324	
Atribut	Počet
species	Gentoo
island	Biscoe
culmen length mm	47.3
culmen depth mm	13.8
flipper length mm	216.0
body mass g	4725.0
sex	NaN

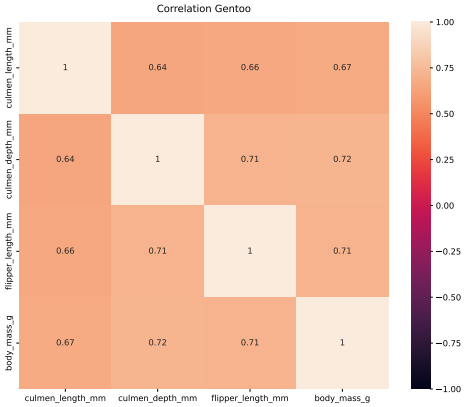
Řádek 339	
Atribut	Počet
species	Gentoo
island	Biscoe
culmen length mm	NaN
culmen depth mm	NaN
flipper length mm	NaN
body mass g	NaN
sex	NaN

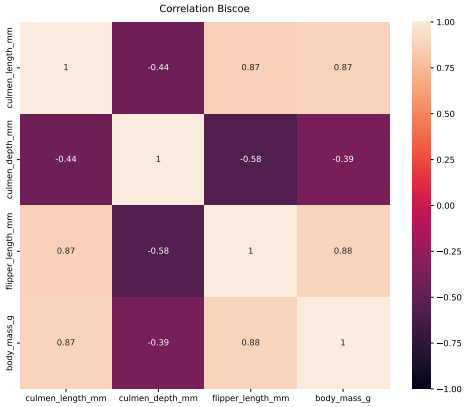
1.7 Korelační analýza

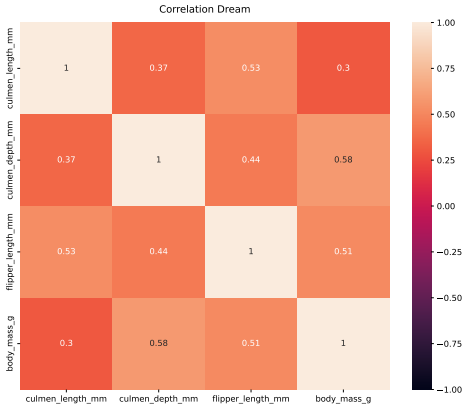
Korelace je znázorně kombinací indexů a grafu typu heatmap. Tyto grafy jsou rozděleny podle druhu, ostrova a pohlaví.

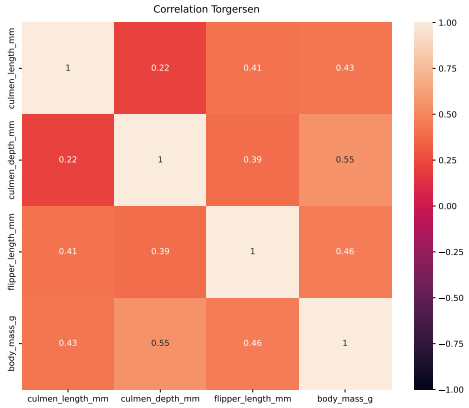


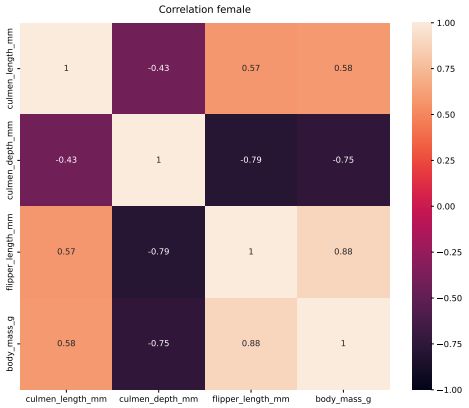


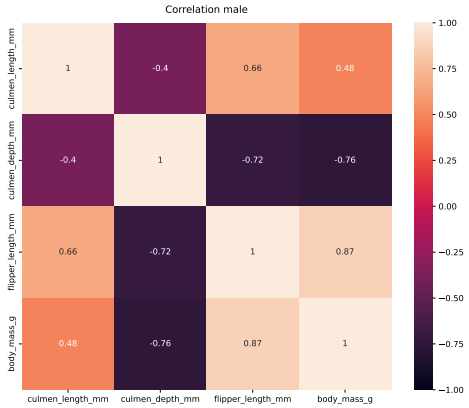





Correlation Dream
culmen_length_mm
culmen_depth_mm
flipper_length_mm
body_mass_g
1
0.37
0.53
0.3
0.37
1
0.44
0.58
0.53
0.44
1
0.51
0.3
0.58
0.51
1
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00


Correlation Torgersen
culmen_length_mm
culmen_depth_mm
flipper_length_mm
body_mass_g
1
0.22
0.41
0.43
0.22
1
0.39
0.55
0.41
0.39
1
0.46
0.43
0.55
0.46
1
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00


Correlation female
culmen_length_mm
culmen_depth_mm
flipper_length_mm
body_mass_g
1
-0.43
0.57
0.58
-0.43
1
-0.79
-0.75
0.57
-0.79
1
0.88
0.58
-0.75
0.88
1
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00


Correlation male
culmen_length_mm
culmen_depth_mm
flipper_length_mm
body_mass_g
1
-0.4
0.66
0.48
-0.4
1
-0.72
-0.76
0.66
-0.72
1
0.87
0.48
-0.76
0.87
1
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00

Kapitola 2

Příprava datové sady

2.1 Zadání

Datovou sadu obsahující informace o tučňácích (`penguins_lter.csv`) je třeba transformovat do podoby vhodné pro dolovací úlohu – klasifikace druhů tučňáků na základě ostatních atributů. Výstupem jsou dva nové datové soubory – `A.csv` je vhodný pro metody vyžadující kategorické atributy a `B.csv` je vhodný pro metody vyžadující numerické atributy.

2.2 Nástroje

Jako nástroj pro úpravu datové sady byl zvolen programovací jazyk Python s využitím knihovny Pandas. Tato knihovna obsahuje spoustu užitečných nástrojů pro zpracování souborů formátu csv a pro následnou práci s daty. Pandas není součástí základní instalace Pythonu, je třeba ji doinstalovat příkazem `pip install pandas`. Skript implementující přípravu datové sady je uložen v souboru `modify_data.py`. Skript vyžaduje tři vstupní argumenty – cestu k původní datové sadě, cestu k výstupní sadě s kategorickými atributy a cestu k výstupní sadě s numerickými atributy.

Příklad spuštění:

```
py ./src/modify_data.py ./dataset/penguins_lter.csv A.csv B.csv
```

Případně pomocí Makefile:

```
make modify
```

2.3 Postup

Zpracování dat je rozděleno na několik dílčích kroků, některé jsou společně pro obě varianty výstupů.

2.3.1 Odstranění irelevantních atributů

Jedná se o odstranění sloupců, které nejsou relevantní pro klasifikaci druhu tučňáka. V tomto případě se jedná například o různé identifikátory – například identifikátor studia, číslo sběru a individuální ID. Dále byly odstraněny informace o snůškách vajec, region, ostrov (byl ponechán u kategorické verze) a komentář. Byly ponechány informace o fyzikálních vlastnostech tučňáků, jejich pohlaví a druhu.

2.3.2 Řešení chybějících hodnot

V záznamech nebylo mnoho případů chybějících hodnot, proto byly ve většině případů odstraněny, především ty záznamy, ve kterých chyběly informace o fyzikálních vlastnostech tučňáků, nebo jejich pohlaví. Ve zbytku dat bylo několik záznamů, kde chyběly informace o složení krve. Tyto atributy byly doplněny mediánovou hodnotou příslušného sloupce.

2.3.3 Řešení odlehlých hodnot

V jednom případě byla ve sloupci Sex chyba – byla zde byla tečka namísto obvyklých `MALE` nebo `FEMALE`. Tento záznam byl proto odstraněn. Krom tohoto případu nebyly v datech žádné další významně odlehlé hodnoty.

2.3.4 Kategorizace numerických hodnot

Pro datovou sadu A byly sloupce obsahující numerické hodnoty převedeny na kategorické hodnoty – konkrétně na určité intervaly hodnot. Jedná se o rozdělení dat do „košů“ (binning), kdy interval mezi minimální a maximální hodnotou sloupce je rozdělen na několik stejně velkých intervalů. V tomto případě bylo použito dvacet košů. Touto metodou byly upraveny všechny sloupce obsahující numerické hodnoty. Výsledná datová sada je v souboru `A.csv`.

2.3.5 Převod kategorických hodnot na numerické

V případě datové sady B bylo zapotřebí provést převod kategorických dat na numerická data a normalizovat vhodné sloupce. Sloupec specifikující druh tučňáka obsahuje tři různé druhy, ty byly nahrazeny čísly 1 2 a 3. Obdobně byly hodnoty `MALE` a `FEMALE` ve sloupci `Sex` nahrazeny za hodnoty 0 a 1. Na závěr byla provedena min-max normalizace u dat popisující složení krve, konkrétně byly hodnoty převedeny na rozsah mezi 0 a 1. Díky této úpravě jsou ve výsledné datové sadě všechny hodnoty nezáporné. Výsledná datová sada je v souboru `B.csv`.