

FAKULTA INFORMAČNÍCH TECHNOLOGIÍ VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ

Ukládání a příprava dat – 2.projekt
Dokumentácia k projektu

Obsah

1	Úvod	2
2	Exploratívna analýza	2
2.1	Preskúmanie atribútov	2
2.2	Čistenie dát	3
2.3	Rozloženie hodnôt atribútov	3
2.4	Odfahľé hodnoty	6
2.5	Zisťovanie korelácie	7
3	Úprava dátovej sady	9
3.1	Kategorické dáta na numerické	9
3.2	Numerické na kategorické	9
3.3	Klasifikácia druhu tučniakov na základe dvoch atribútov	9

1 Úvod

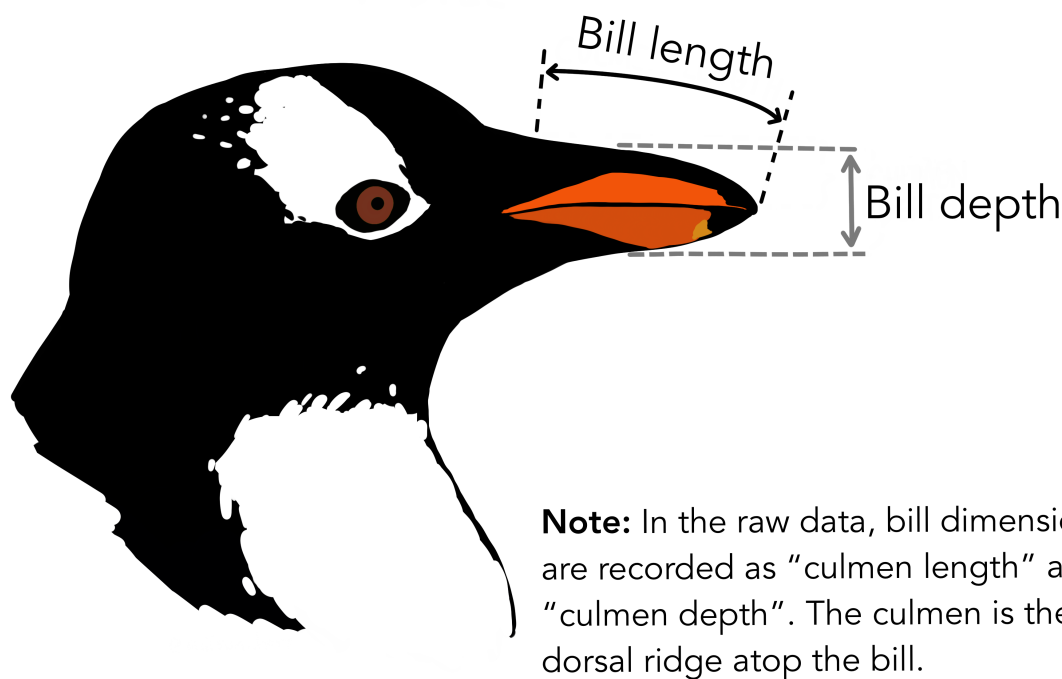
Ako dátovú sadu pre náš projekt sme si zvolili sadu tučňakov¹ z dôvodu zaujímavosti a čistoty sady v porovnaní s ostatnými sadami. Projekt sme sa rozhodli vypracovať v jazyku Python v Jupyter notebooku.

2 Exploratívna analýza

2.1 Preskúmanie atribútov

Dátová sada tučňakov obsahuje 344 vzoriek a 17 atribútov. Nie všetky atribúty sú zaujímavé na analýzu, nižšie sú uvedené najzaujímavejšie atribúty. O nepotrebných atribútoch a vysporiadaní sa s nimi sa píše neskôr.

- Druh – meranie bolo spravené na troch druhoch tučňakov (Adelie, Chinstrap, Gentoo).
- Ostrov – ostrov v Antarktíde, kde sa meranie uskutočnilo (Biscoe, Dream, Torgersen).
- Dĺžka zobáka v mm znázornená na Obr. 1, hodnoty v rozmedzí 32,1 - 59,6.
- Hĺbka zobáka v mm znázornená na Obr. 1, hodnoty v rozmedzí 13,1 - 21,5.
- Dĺžka plutvy v mm, hodnoty v rozmedzí 172 - 231.
- Váha tučňaka v g, hodnoty v rozmedzí 2700 - 6300.
- Pohlavie tučňaka - samec/samica.
- $\delta^{15}\text{N}$, $\delta^{13}\text{C}$ – merania obsahu stabilných izotopov dusíka a uhlíka z krvi tučňaka, hodnoty $\delta^{15}\text{N}$ v rozmedzí 7,63 - 10, hodnoty $\delta^{13}\text{C}$ v rozmedzí -27 - -23,8.



Obr. 1: Zobrazenie dĺžky a hĺbky zobáku².

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

²<https://github.com/allisonhorst/palmerpenguins/blob/main/README.md>

2.2 Čistenie dát

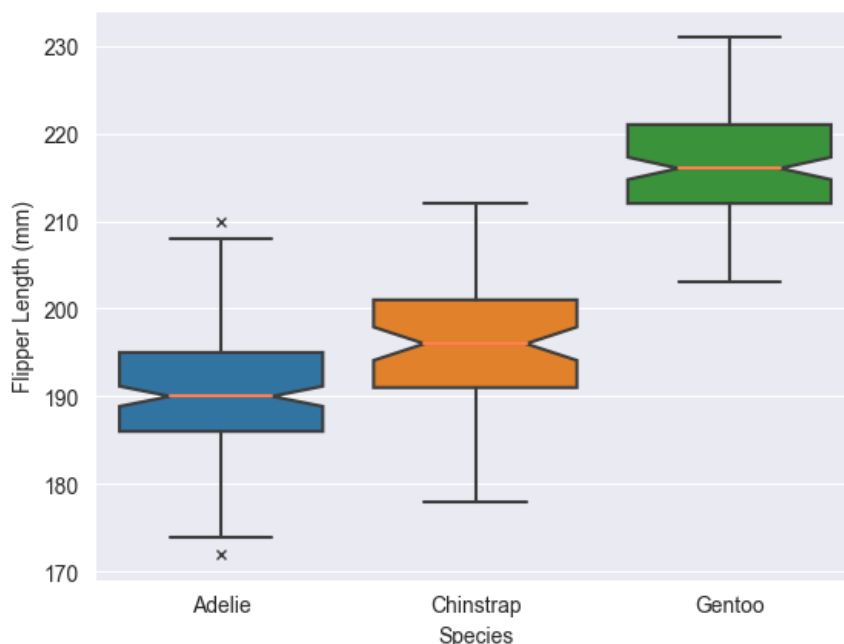
Vstupná dátová sada obsahuje niekoľko nepotrebných atribútov vzhľadom na analýzu dát, ktoré ideme robiť. Týmito atribútmi sú číslo vzorky (Sample Number), identifikačné číslo (Individual ID) a komentáre k vzorku (Comments). Komentár slúži najmä k zdôvodneniu, prečo niektoré dáta chýbajú, identifikačné číslo a číslo vzorky slúžia len na priamu identifikáciu jedinca. Sada taktiež obsahuje atribúty, ktoré majú jednu unikátnu hodnotu – región (Region), štádium (Stage). Atribúty obsahujú jedinú hodnotu preto, lebo všetky vzorky boli odobrané v regióne Anvers v Antarktíde od dospelých jedincov. Všetky tieto atribúty sme sa z vyššie uvedených dôvodov rozhodli zo sady zahodiť.

Chýbajúce hodnoty

Dátová sada obsahuje niekoľko chýbajúcich hodnôt. Z výpisu dátovej sady pre chýbajúce hodnoty sme zistili, že dve vzorky tučňakov (indexy 3 a 339) neobsahujú väčšinu parametrov. Jedna z poznámok napovedá, že dospelý jedinec nebol odmeraný, tak sme tieto chybné vzorky odstránili. Chýbajúce hodnoty zo vzoriek krvi jedincov zaberajú približne 3.5 % sady. Obe atribúty sme sa rozhodli doplniť strednou hodnotou daného atribútu. Zvyšok chýbajúcich hodnôt je chýbajúce pohlavie. Chýbajúce hodnoty tvoria približne 2 % dátovej sady. Hodnoty sme doplnili najčastejšou hodnotou z atribútu, t.j. mužským pohlavím. Následne sme zistili, že vzorka s indexom 336 obsahuje chybnú hodnotu (.) pohlavia. Nahradili sme ju modusom tiež.

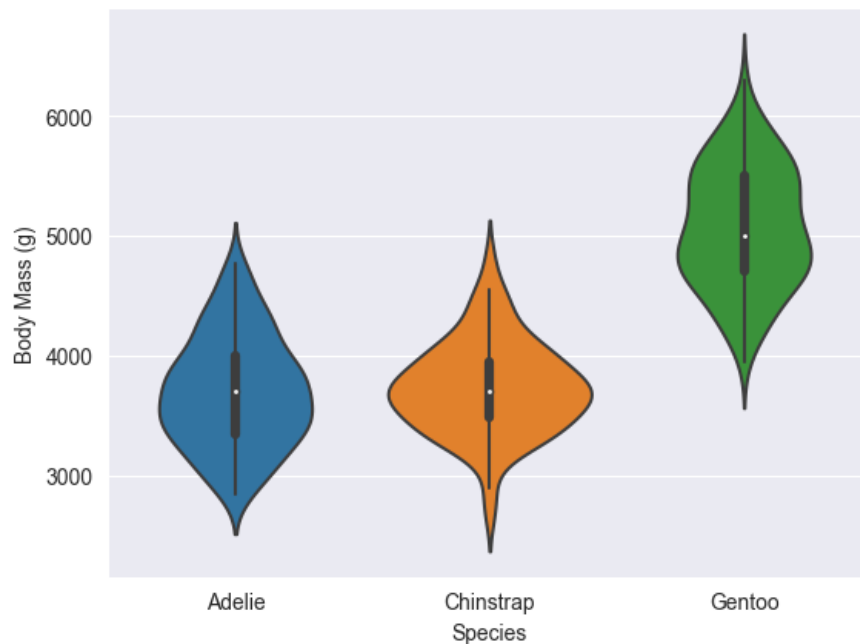
2.3 Rozloženie hodnôt atribútov

Nasledujúca sekcia obsahuje rôzne typy grafov, ktoré zobrazujú rôzne parametre a porovnania, hlavne medzi druhmi tučňakov. Prvý krabicový graf na Obr. 2 znázorňuje rozloženie dĺžky plutiev jednotlivých druhov tučňakov.



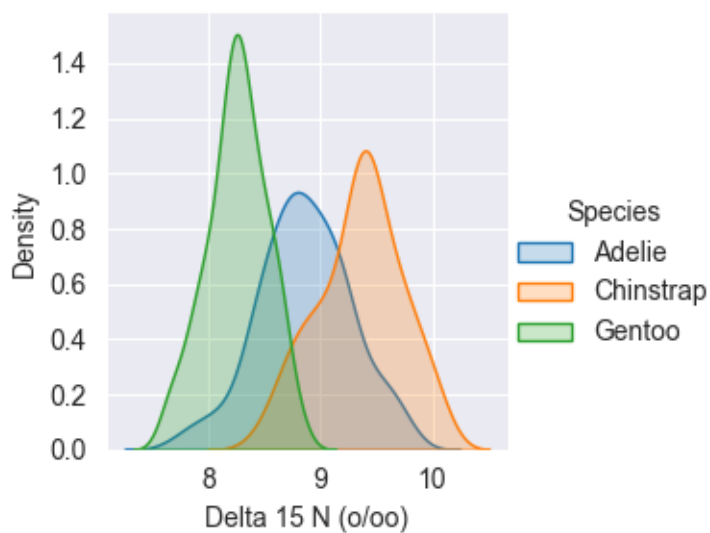
Obr. 2: Krabicový graf znázorňujúci rozloženie dĺžky plutvy medzi jednotlivými druhmi tučňakov.

Z grafu je vidieť, že merania druhu tučňaka Adelie obsahujú dve odľahlé hodnoty a že dĺžka plutvy je v rozmedzí približne 175 - 235 mm. Husľový graf na Obr. 3 znázorňuje rozloženie váhy druhov tučňakov.



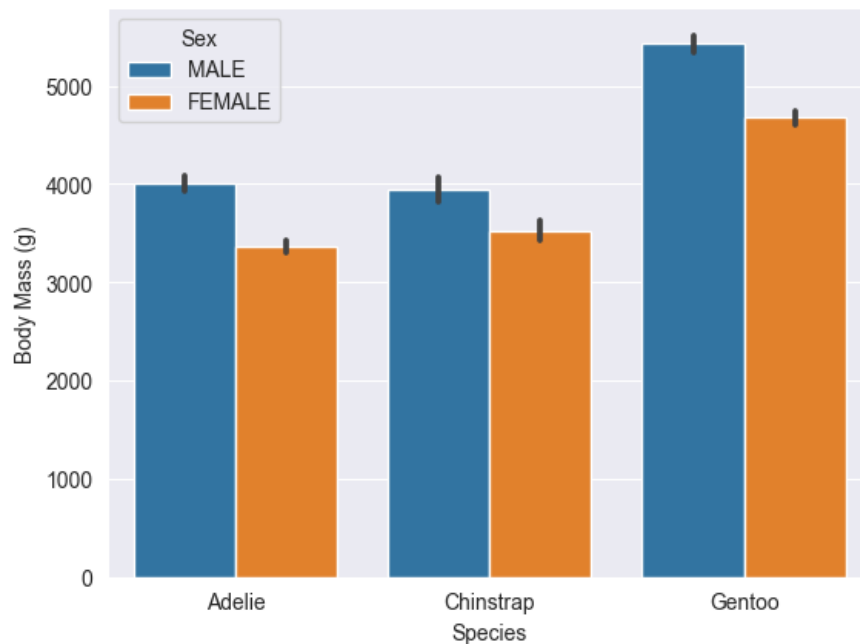
Obr. 3: Husľový graf znázorňujúci rozloženie váhy jednotlivých druhov tučniakov.

Graf rozloženia hustoty pravdepodobnosti na Obr. 4 zobrazuje rozloženie atribútu $\delta^{15}\text{N}$.



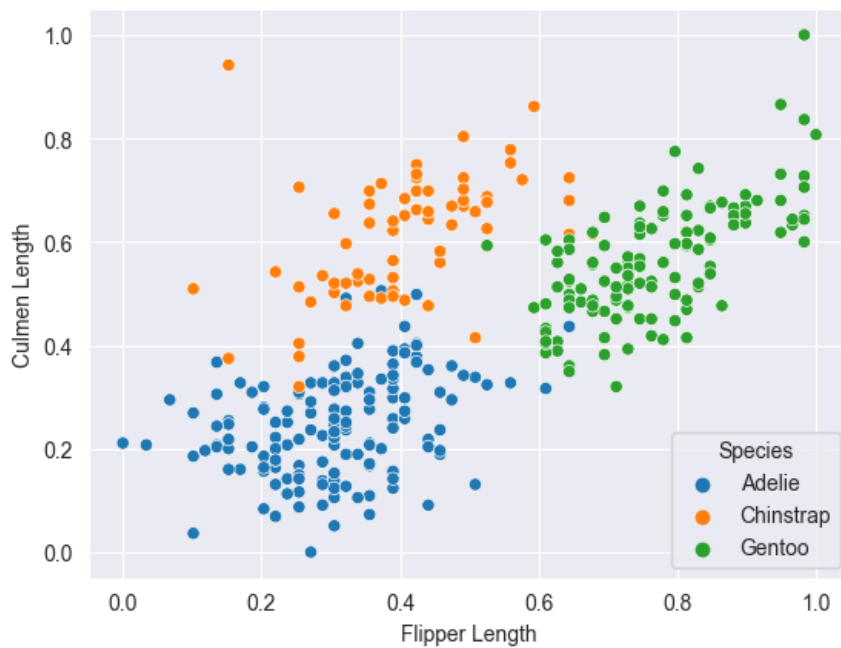
Obr. 4: Graf rozloženia hustoty pre atribút $\delta^{15}\text{N}$.

Z grafu je možné vyčítať, že všetky tri druhy majú inú strednú hodnotu pre tento parameter. Histogram na Obr. 5 zobrazuje porovnanie váhy medzi jednotlivými druhmi podľa pohlavia.



Obr. 5: Histogram znázorňujúci váhu jednotlivých druhov podľa pohlavia.

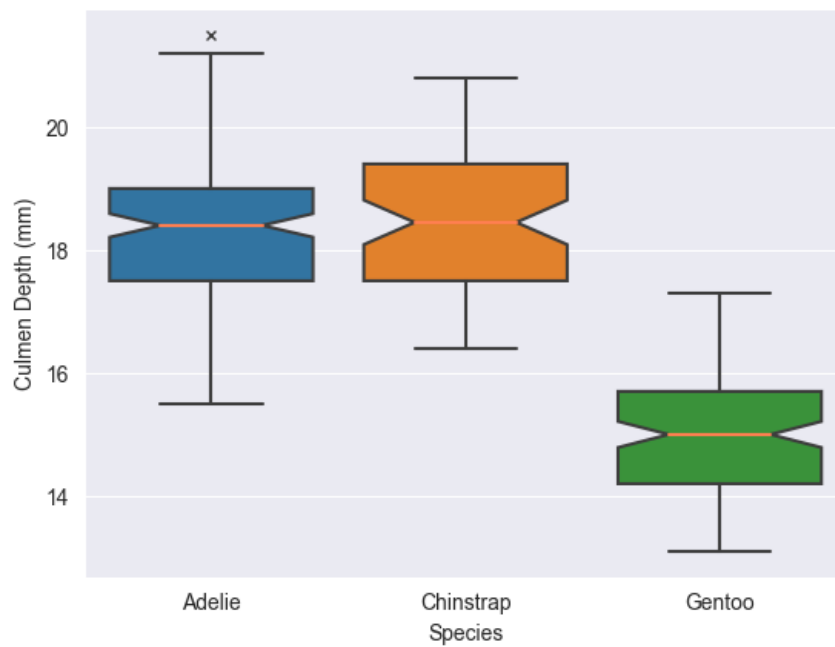
Z histogramu je prehľadne zrejmé, že tučniaky mužského pohlavia vážia približne o 0,5 až 1 kg viac ako tučniaky ženského pohlavia. Posledný graf na Obr. 6 zobrazuje normalizovanú závislosť hodnôt dĺžky plutvy a dĺžky zobáku.



Obr. 6: Normalizovaná závislosť dĺžky plutvy na dĺžke zobáku.

2.4 Odľahlé hodnoty

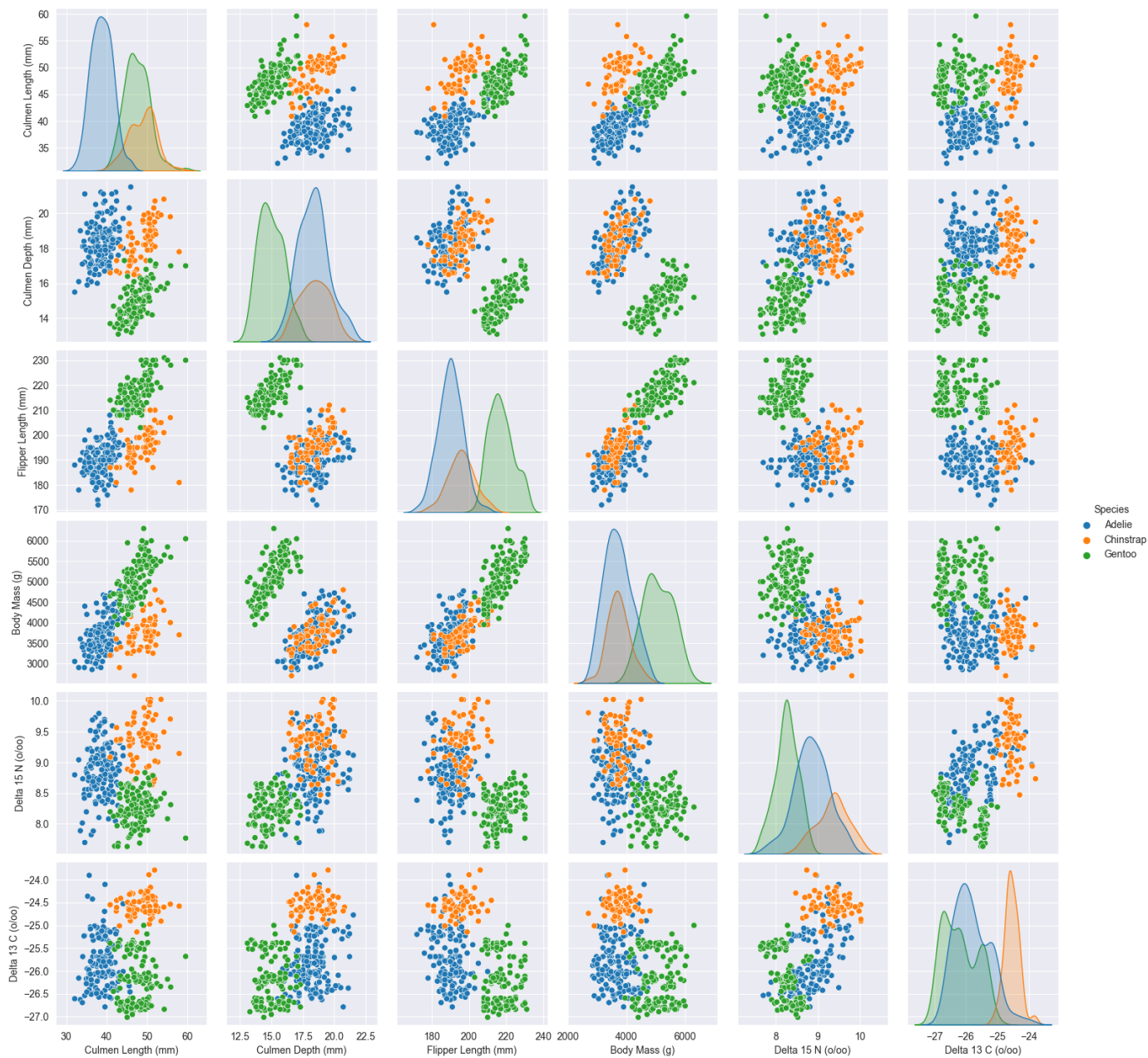
Dátová sada neobsahuje odľahlé hodnoty ako také. Ak však sadu rozdelíme podľa druhu tučniakov, tak na Obr. 2 a Obr. 7 sú v krabicových grafoch vizualizované odľahlé hodnoty mimo mezikvantilového rozpätia.



Obr. 7: Rozloženie hĺbky zobáku podľa druhu tučniaka.

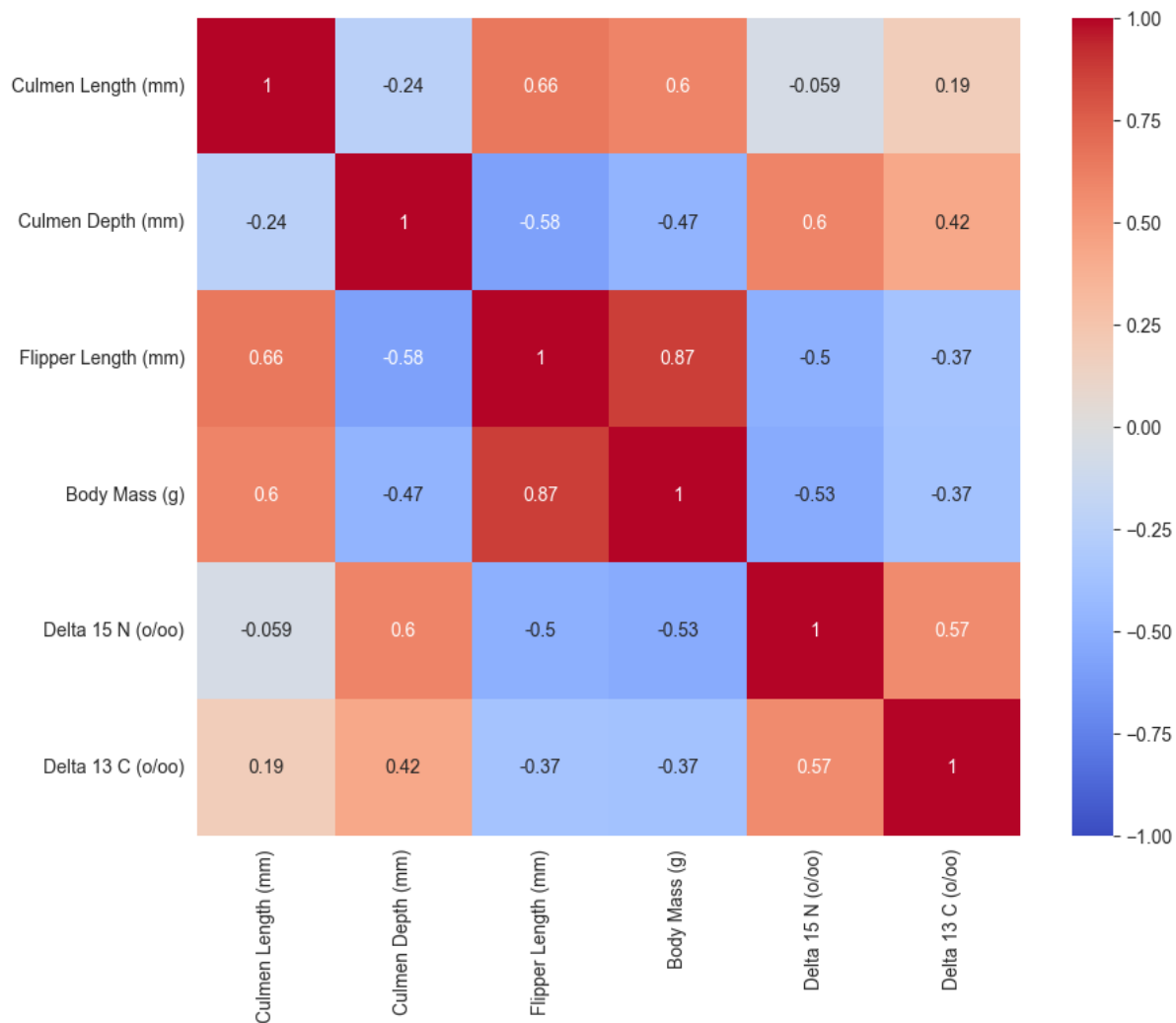
2.5 Zisťovanie korelácie

Zisťovanie korelácie sme začali zobrazením matice grafov pre všetky numerické atribúty. Matica grafov na Obr. 8 zobrazuje porovnanie jednotlivých atribútov medzi druhmi tučňakov.



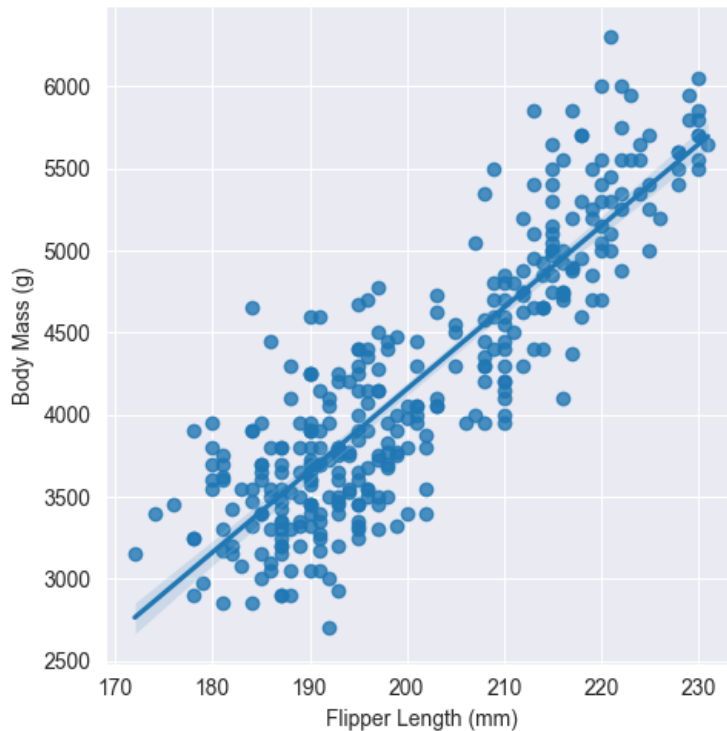
Obr. 8: Porovnanie atribútov medzi druhmi tučňakov.

Z matice je na prvý pohľad vidno, že pravdepodobne existuje u tučňakov vysoká korelácia medzi váhou a dĺžkou plutvy. Intuitívne toto tvrdenie zmysel dáva, poďme ho skúsiť overiť. Zobrazme ešte korelačnú maticu na Obr. 9, ktorá zobrazuje Pearsonove koeficienty korelácie pre každú dvojicu numerických atribútov.



Obr. 9: Korelačná matica.

Z korelačnej matice vyčítame, že koeficient korelácie je 0,87 pre túto dvojicu atribútov. Graf na Obr. 10 zobrazuje rozloženie hodnôt pre tieto atribúty preložených priamkou lineárnej regresie.



Obr. 10: Rozloženie atribútov dĺžky plutvy a váhy tučniaka preložených priamkou lineárnej regresie.

3 Úprava dátovej sady

S chýbajúcimi a nesprávnymi hodnotami sme sa vysporiadali pred začiatkom exploratívnej analýzy. Detailný postup je popísaný v Podsekcii 2.2.

3.1 Kategoricé dáta na numerické

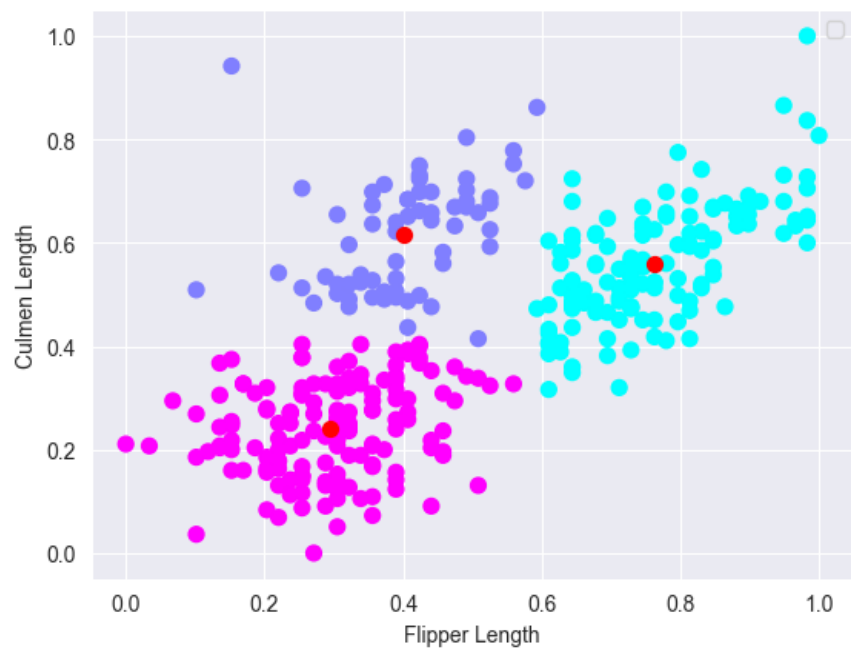
Ako prevod kategorických dát na numerické sme konvertovali ostrov na číslo. Ostrov Dream na hodnotu 0, Torgersen na hodnotu 1 a Biscoe na hodnotu 2. Do dátovej sady sme priložili normalizované hodnoty váhy, dĺžky, hĺbky zobáka, a dĺžky plutvy metódou min-max. Dátová sada má názov `numeric_normalized.csv`. Druh tučniakov sme v sade ponechali ako validačný atribút. Daný algoritmus si potom môže zvoliť zo sady hodnoty, s ktorými chce pracovať.

3.2 Numerické na kategorické

Atribúty do druhej dátovej sady sme zvolili rovnaké ako do prvej s výnimkou ostrova. Taktiež sme ponechali druh ako validačný atribút. Numerické dáta sme previedli na kategorické metódou plnenia, kde sme jednotlivé rozloženie rozdelili pomocou decilov, resp. 10-kvantilov.

3.3 Klasifikácia druhu tučniakov na základe dvoch atribútov

Ako rozšírenie k projektu sme si chceli vyskúšať klasifikáciu pomocou metódy K-means na tejto dátovej sade. Spočiatku sme sa rozhodli klasifikovať na základe dvoch numerických atribútov. Pred zvolením najvhodnejších kandidátov na atribúty sme si všetky dáta normalizovali inou metódou (min-max, z-score, max scale) a zobrazili matice grafov pre každý normalizovaný atribút inou metódou. Z daných matic sme sa snažili pohľadom nájsť dvojicu normalizovaných atribútov, ktoré majú jednotlivé zhluky druhov čo najviac oddelené. Nakoniec sme vybrali dvojicu dĺžka plutvy, dĺžka zobáka. Výsledok metódy K-means je zobrazený na Obr. 11. Výsledok metódy môžeme porovnať s Obr. 6 na ktorom sú vyobrazené pôvodné dáta.



Obr. 11: Výsledok metódy K-means.