Fakulta informačních technologií
Vysoké učení technické v Brně

Ukládání a příprava dat – 2.projekt
Dokumentácia k projektu

2022 Zbyněk Lamačka, Luděk Burda, Matej Kunda

Obsah

1 Úvod

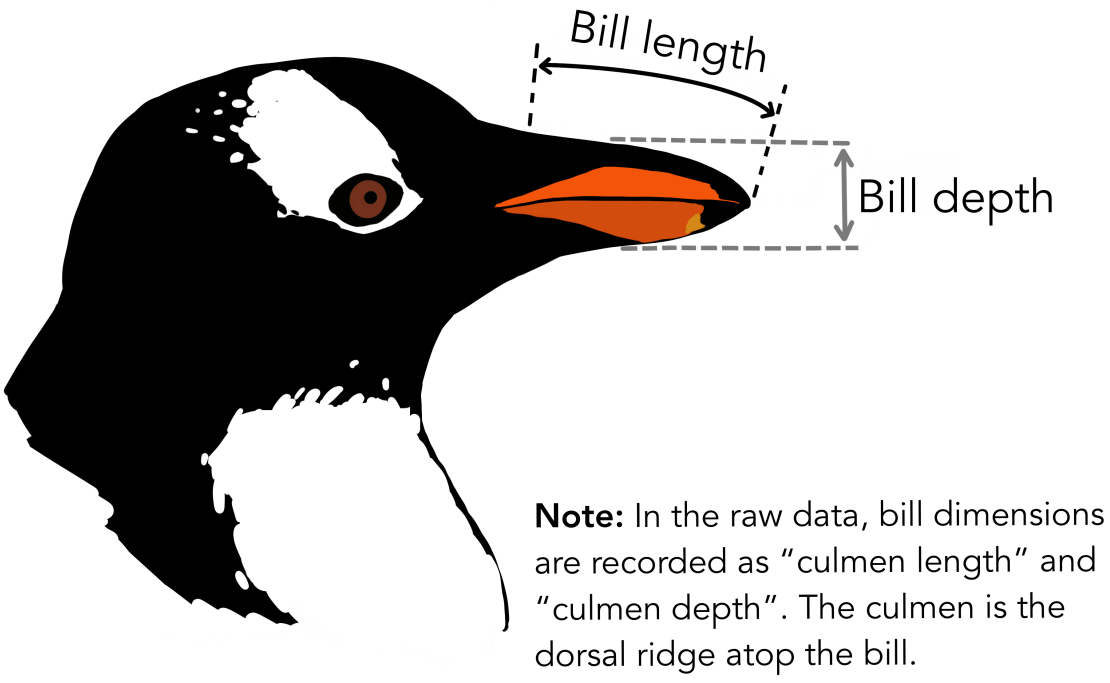
Ako dátovú sadu pre náš projekt sme si zvolili sadu tučniakov[1] z dôvodu zaujímavosti a čistoty sady v porovnaní s ostatnými sadami. Projekt sme sa rozhodli vypracovať v jazyku Python v Jupyter notebooku.

2 Exploratívna analýza

2.1 Preskúmanie atribútov

Dátová sada tučniakov 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 tučniakov (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 - 21m5.
- Dĺžka plutvy v mm, hodnoty v rozmedzí 172 - 231.
- Váha tučniaka v g, hodnoty v rozmedzí 2700 - 6300.
- Pohlavie tučniaka - samec/samica.
- δ^{15}N, δ^{13}C – merania obsahu stabilných izotopov dusíka a uhlíka z krvi tučniaka, hodnoty δ^{15}N v rozmedzí 7,63 - 10, hodnoty δ^{13}C v rozmedzí -27 - -23,8.



Obr. 1: Zobrazenie dĺžky a hĺbky zobáku[2].

[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

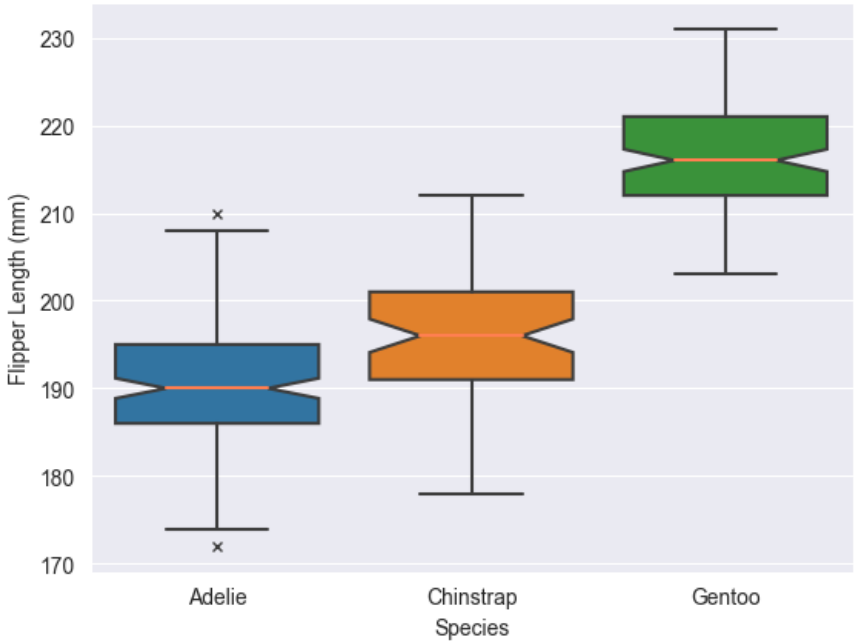
Vstupná dátová sada obsahuje niekoľko nepotrebných atribútov vzhľadom na analýzu dát, ktoré ideme robiť. Týmito atribútmi sú číslo vzorky (Sample Number), identifikačné číslo (Individual ID) a komentáre k vzorku (Comments). Komentár slúži najmä k zdôvodneniu, prečo niektoré dáta chýbaju, identifikačné 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

Dátova sada obsahuje niekoľko chýbajúcich hodnôt. Z výpisu dátovej sady pre chýbajúce hodnoty sme zistili, že dve vzorky tučniakov (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. samcom. Následne sme zistili, že vzorka s indexom 336 obsahuje chybnú hodnotu (`.`) pohlavia. Nahradili sme ju modusom tiež.

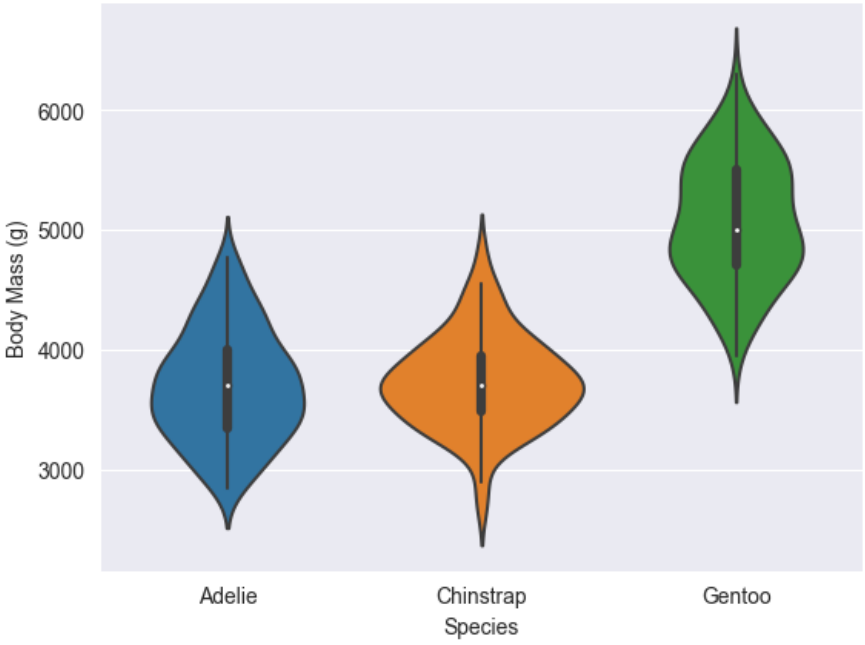
2.3 Rozloženie hodnôt atribútov

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



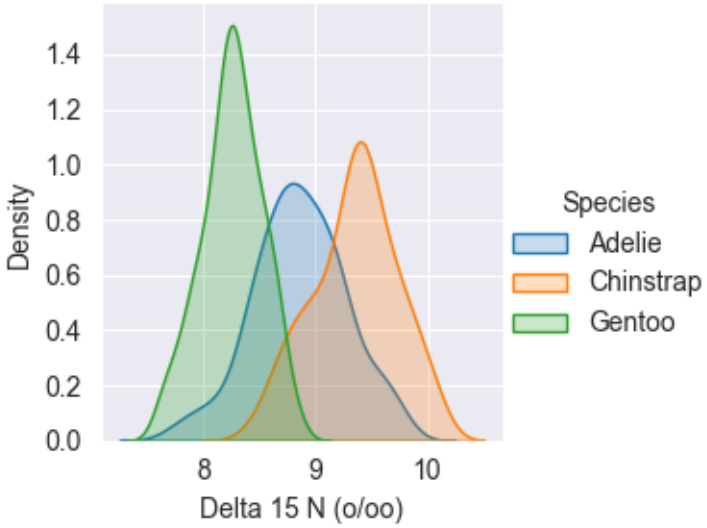
Obr. 2: Krabicový graf znázorňujúci rozloženie dĺžky plutvy medzi jednotlivými druhmi tučniakov.

Z grafu je vidieť, že merania druhu tučniaka 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 tučniakov.



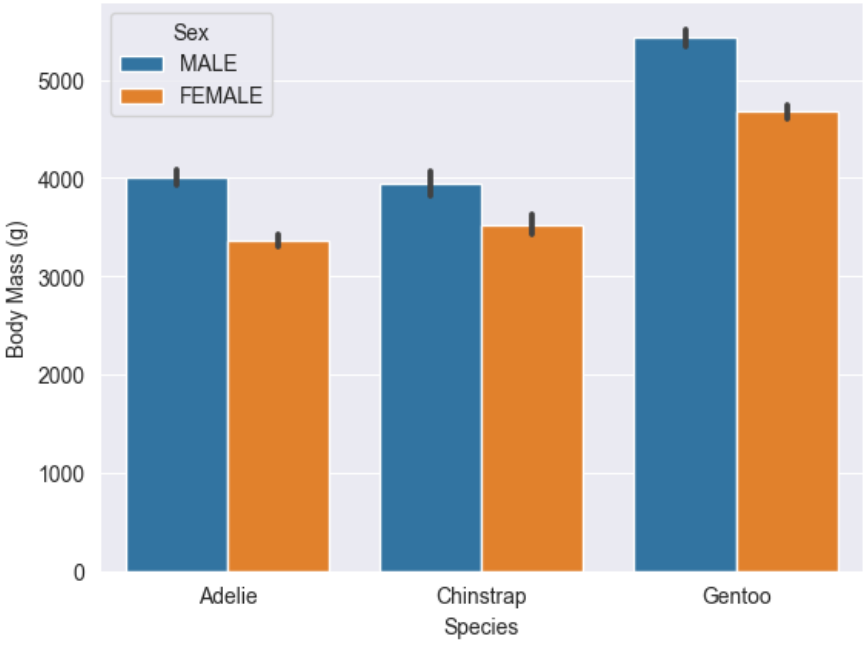
Obr. 3: Husľový graf znázorňujúci rozloženie váhy jednotlivých druhov tučniakov.

Graf rozloženia hustoty pravdepodobnosti na Obr. 4 zobrazuje rozloženie atribútu δ^{15}N.



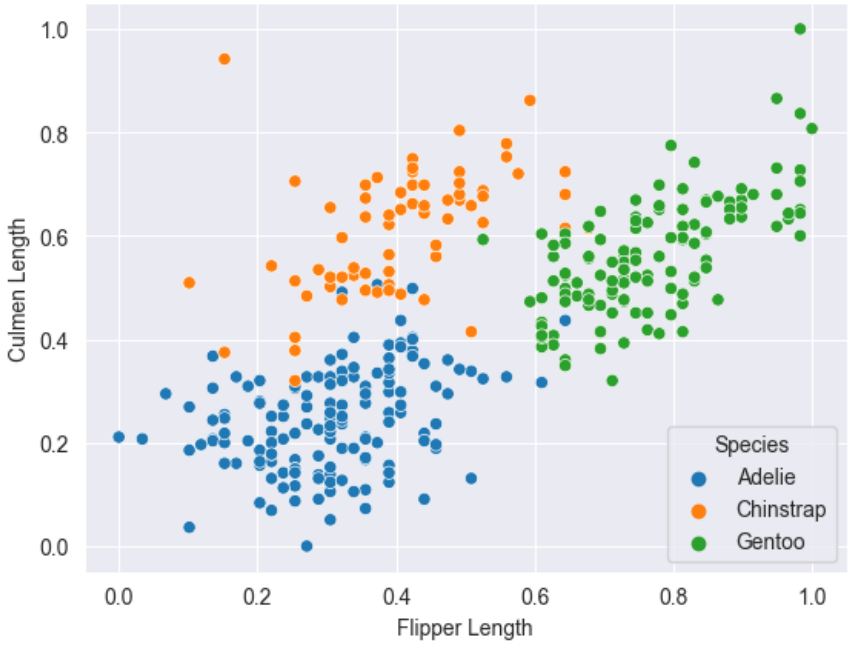
Obr. 4: Graf rozloženia hustoty pre atribút δ^{15}N.

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.

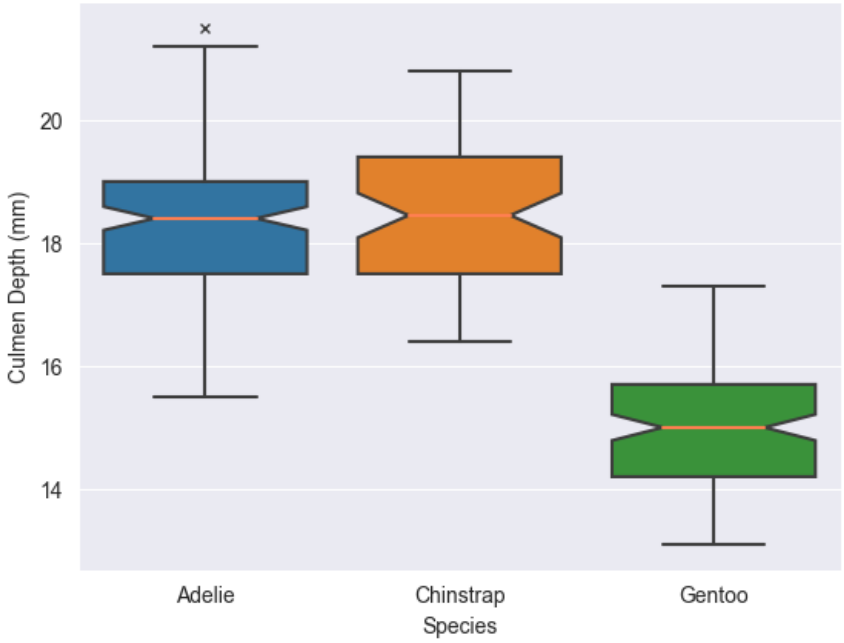
Z histogramu je prehľadne zrejmé, že samci vážia približne o 0,5 až 1 kg viac ako samičky. Posledný graf na Obr. 6 zobrazuje normalizovanú závislosť hodnôt dĺžky plutvy a dĺžky zobáku.



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

2.4 Odľahlé hodnoty

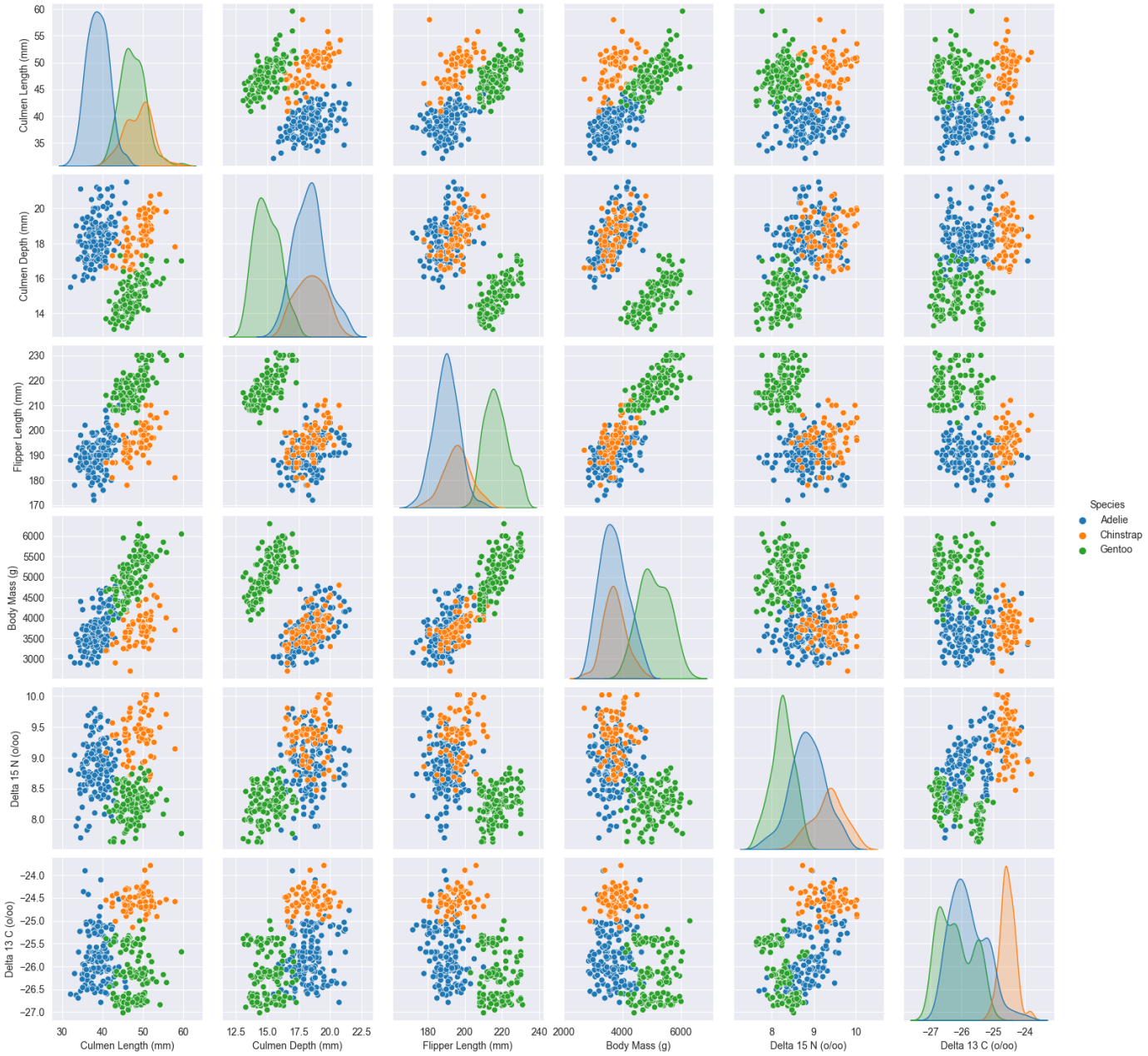
Dátová sada neobsahuje odľahlé hodnoty ako také. Ak však sadu rozdelíme podľa druhu tučniakov, tak na Obr. 2 a Obr. 7 sú v krabicových grafoch vizualizované odľahlé hodnoty mimo mezikvantilového rozpätia.



Obr. 7: Rozloženie hĺbky zobáku podľa druhu tučniaka.

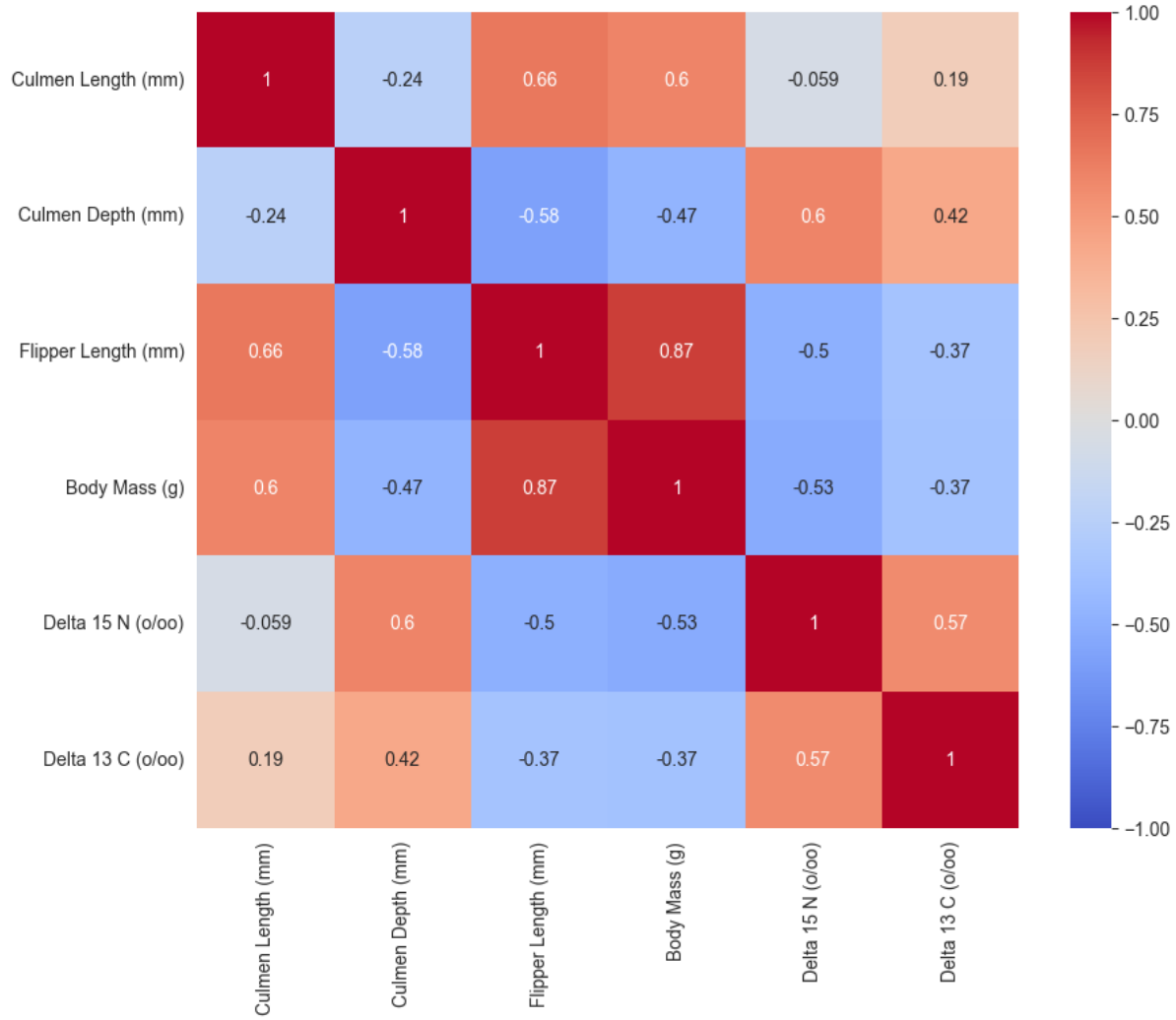
2.5 Zisťovanie korelácie

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



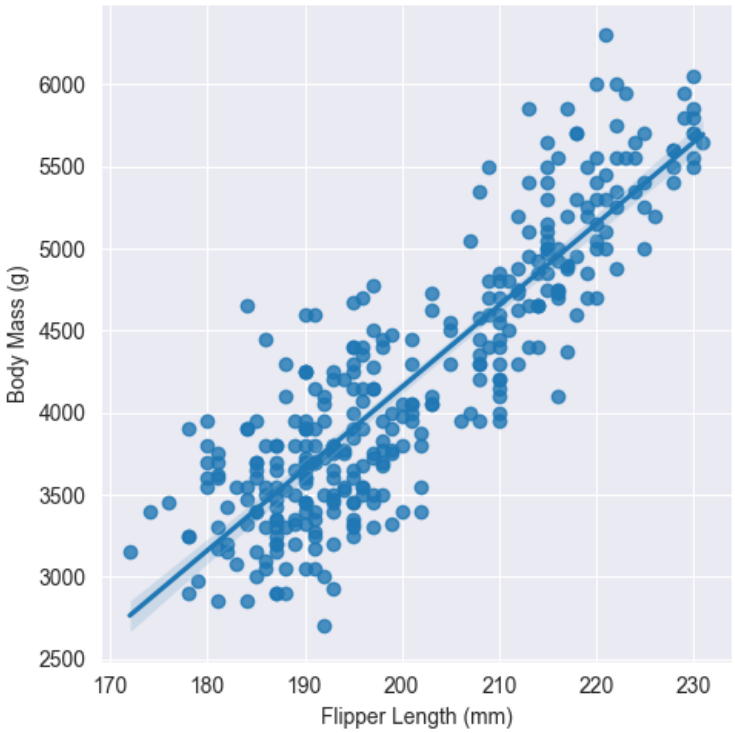
Obr. 8: Porovnanie atribútov medzi druhmi tučniakov.

Z matice je na prvý pohľad vidno, že pravdepodobne existuje u tučniakov 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 Kategorické dáta na numerické

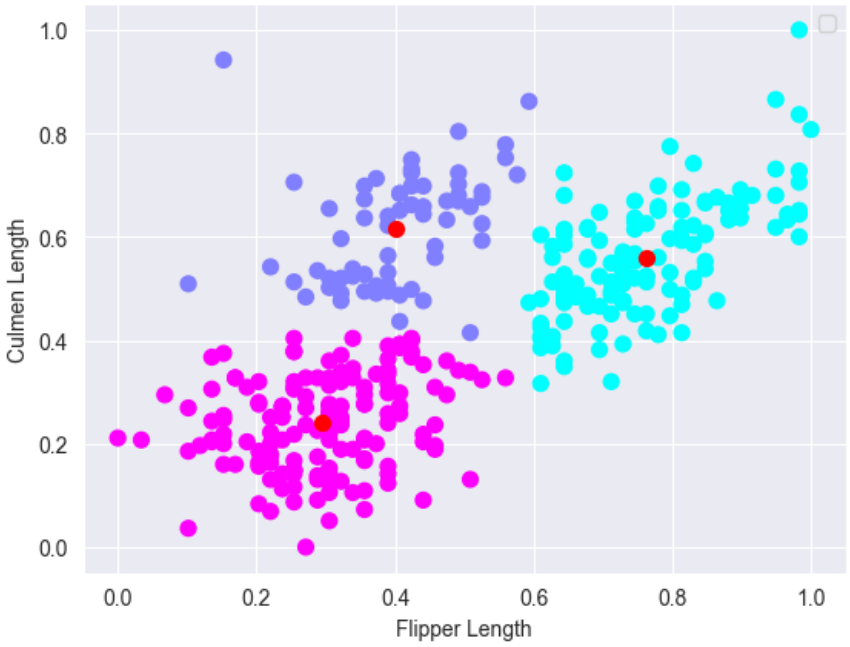
Ako prevod kategorických dát na numerické sme konvertovali ostrov na číslo. Ostrov Dream na hodnotu 0, Torgersen na hodnotu 1 a Biscoe na hodnotu 2. Do dátovej sady sme priložili normalizované hodnoty váhy, dĺžky, hĺbky zobáka, a dĺžky plutvy metódou min-max. Dátová sada má názov `numeric_normalized.csv`. 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 ostrovu. Taktiež sme ponechali druh ako validačný atribút. Numerické dáta zme previedli na kategorické metódou plnenia, kde sme jednotlivé rozloženie rozdelili na 10 rovnakých intervalov, aby sme zachovali pôvodné rozloženie. Výsledná dátová sada nesie názov `categoric.csv`.

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 matíc 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.