Príprava dát a ich popisná charakteristika

Marek Mudroň
xmudro04

Samuel Repka
xrepka07

Barbora Šmahlíková
xsmahl00

zima 2022

Obsah

Kapitola 1

Dátová sada

V tomto projekte sa zaoberáme exploráciou, čistením a tvorbou popisnej charakteristiky pre open-source dátovú sadu *Palmer Archipelago (Antarctica) penguin data*. Ide o dátovú sadu, ktorej obsahom sú informácie o fyziologických a demografických vlastnostiach troch druhov tučniakov. Dataset obsahuje súbory `penguins_lter.csv` a `penguins_size.csv`. Oba obsahujú informácie k rovnakým jedincom, no `penguins_lter.csv` poskytuje okrem informácií obsiahnutých v `penguins_size.csv` aj dáta k hodnotám stabilných izotopov uhlíka (δ^{13}C) a dusíka (δ^{15}N) v tele. Tieto údaje sa používajú už od roku 1980 poskytujú hodnotné informácie k životospráve a migračným vzorcom tučniakov. Preto pre naše účely využívame informácie obsiahnuté v `penguins_lter.csv`.

Kapitola 2

Exploratívna analýza

Táto kapitola sa venuje

2.1 Atribúty

studyName

Tento atribút obsahuje identifikátor štúdie, z ktorej dáta pochádzajú. Celkovo sú identifikátory tri, a to `PAL0708`(s počtom 110), `PAL0809`(114 prvkov) a `PAL0910`(počet prvkov 120).

Sample Number

Poradové číslo záznamu vzťahujúce sa na konkrétnu štúdiu.

Species

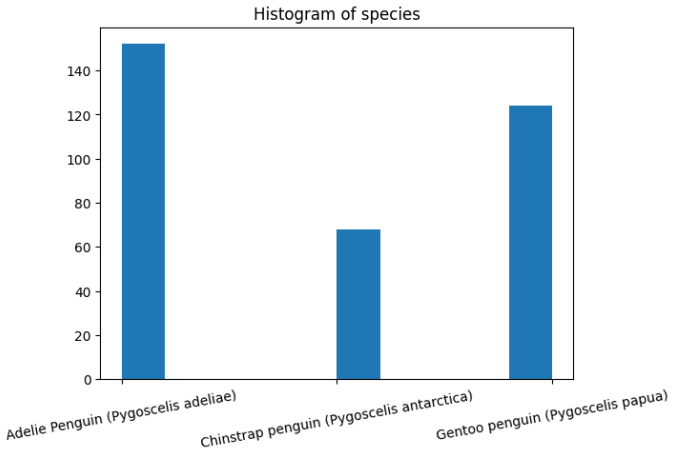
Dataset obsahuje v atribúte `Species` názov jedného z troch druhov tučniakov

- Adelie Penguin (Pygoscelis adeliae)
- Chinstrap penguin (Pygoscelis antarctica)
- Gentoo penguin (Pygoscelis papua)

V dátach dominuje Adelie penguin, nasledovaný tučniakom Gentoo a Chinstrap, ako je vidieť na grafe 2.1.

Region

V atribúte `Region` je uvedená oblasť výskytu daného jedinca. Táto je pre každého z nich rovnaká a obsahuje hodnotu *Anvers*.

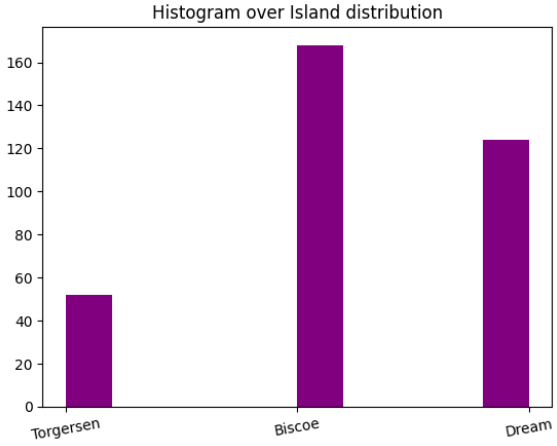


Obr. 2.1: Histogram počtu tučniakov podľa druhu

Island

Informácia obsiahnutá v tomto atribúte označuje ostrov, ktorý obýva jedinec na zázname. Demografické rozloženie sa medzi druhmi líši, ako je vidno na grafe 2.2. Atribút obsahuje tieto tri hodnoty.

- Torgersen
- Biscoe
- Dream



Obr. 2.2: Počty tučniakov na ostrovoch, ako sú uvedené v datasete.

Stage

Atribútom `Stage` uvádza pre každého jedinca rovnakú hodnotu *Adult, 1 Egg Stage*.

Individual ID

Identifikátor každého z jedincov. Neexistuje záznam bez tejto hodnoty. V datasete sa ale nachádza len 190 unikátnych identifikátorov, pričom celkový počet záznamov je 344.

Clutch completion

`Clutch Completion` obsahuje informáciu o tom, či bol jedince pozorovaný s hniezdom obsahujúcim 2 vajíčka. Obsahuje hodnoty, "Yesä "No". Skoro 90% jedincov má v záznamoch uvedené "Yes".

Date egg

Dátum, kedy bolo hniezdo prvý krát pozorované s 1 vajíčkom.

Culmen length (mm) a Culmen depth (mm)

Atribúty `Culmen Length (mm)` a `Culmen Depth (mm)` poskytujú rozmery zobáku. Culmen length označuje jeho dĺžku v milimetroch, zatiaľ čo Culmen depth označuje jeho výšku. Hodnoty pre Culmen length sú v intervale od 32.1 do 59.6 mm, so strednou hodnotou 43.92mm a štandardnou odchýlkou 5.46. Culmen depth má hodnoty od 13.1 do 21.5 mm, stredná hodnota 17.15 a odchýlka 1.97.

Flipper length (mm)

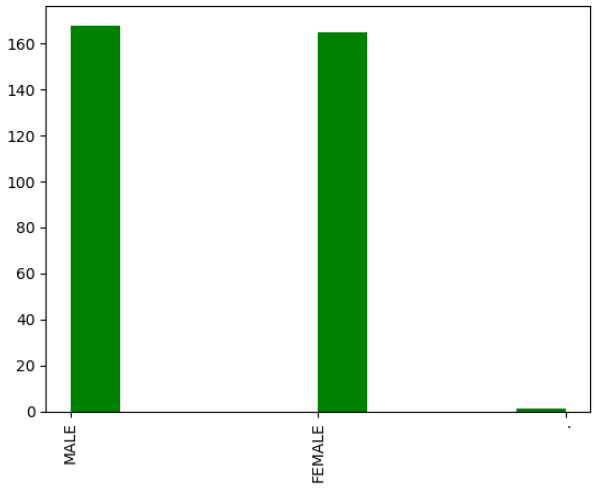
Dĺžka krídel tučniaka je obsiahnutá v atribúte `Flipper Length (mm)` a taktiež je v milimetroch. Stredná hodnota je 200.9 mm, so štandardnou odchýlkou 14 mm. Namerané maximum je 231 mm a minimum 172 mm.

Body mass (g)

Telesnú hmotnosť v gramoch poskytuje atribút `Body Mass (g)`. Najtučnejší tučniak váži 6.3 kg. Na druhom konci spektra sa nachádza tučniak s hmotnosťou 2.7 kg. Stredná hodnota je 4.2 kg a štandardná odchýlka 0.8 kg.

Sex

Pohlavie jedinca. Táto hodnota nie je vždy uvedená, a v jednom prípade jedna z položiek tohto atribútu obsahuje nevalidnú hodnotu '.' 2.3.



Obr. 2.3: Histogram pohlaví tučniakov s nevyfiltrovanou nevalidnou hodnotou.

Delta 15 N (o/oo)

Stabilný izotop dusíka. Meranie tohoto atribútu prebehlo 330 krát. Stredná hodnota je 8.73 so štandardnou odchýlkou 0.551. Namerané maximum je 10.03 a minimum 7.63.

Delta 13 C (o/oo)

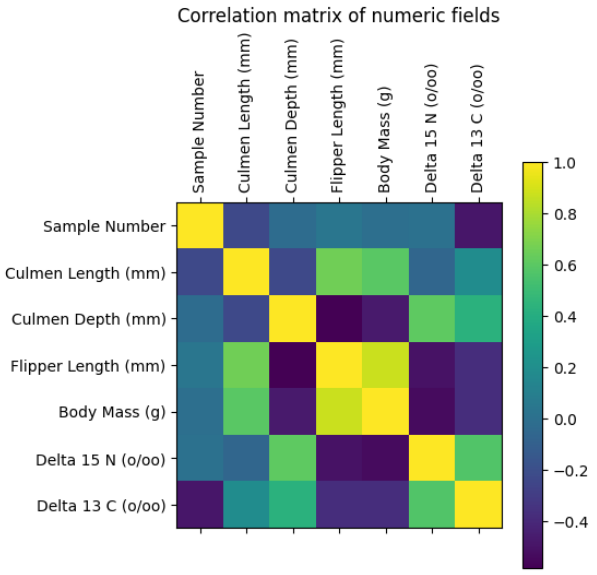
Atribút s meraniami stabilných izotopov uhlíka. Podobne ako pri Delta 15 N, dataset obsahuje 330 záznamov s týmito hodnotami. Stredná hodnota je -25.69 a štandardná odchýlka 0.79. Maximum je -23.787 a minimum je -27.02.

Comments

Tento atribút je pomerne zriedka vyplnený a obsahuje slovné poznámky k niektorým meraniam. Z hľadiska strojového spracovania dát je irelevantný.

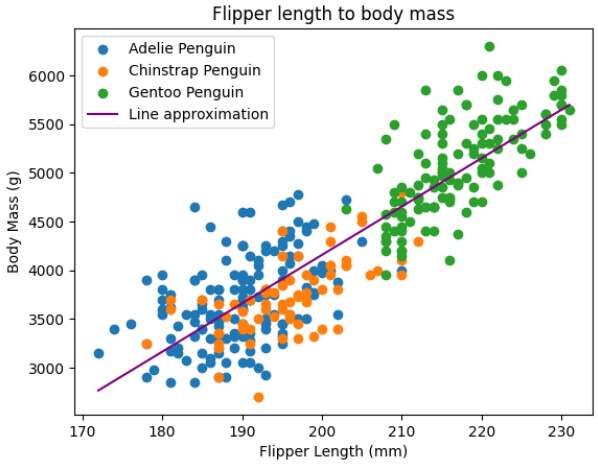
2.2 Analýza vzťahov

Na získanie prvotného pohľadu na vzťahy sme vytvorili korelačnú maticu 2.4. Môžeme z nej vyčítať rôzne informácie, ako napríklad že hmotnosť je silno korelovaná s dĺžkou krídel, čo sa dá intuitívne čakať. Nie až také intuitívne je, že výška zobáka je trochu inverzne korelovaná s dĺžkou krídel.



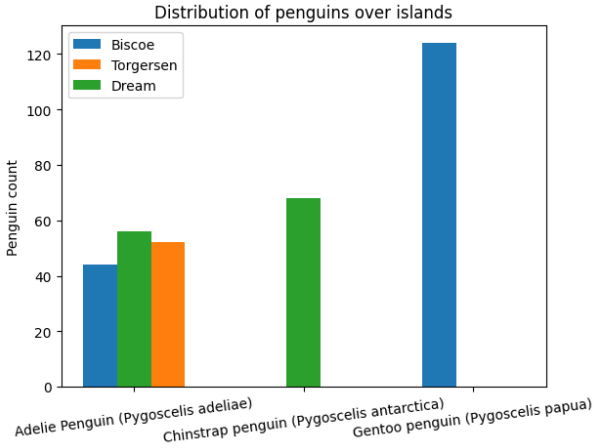
Obr. 2.4: Korelačná matica číselných atribútov

Bližší náhľad na koreláciu hmotnosti a dĺžky krídlel nám poskytol graf 2.5. Vidno, že na atribútoch existuje pomerne silná závislosť, ktorá sa dá aproximovať priamkov (aj keď nie veľmi presne.)



Obr. 2.5: Dĺžka krídel vzhľadom k hmotnosti

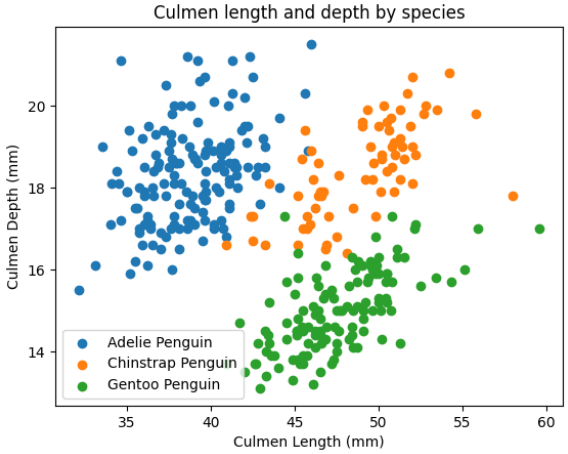
Zaujímavý je aj pohľad na dáta z hľadiska druhov, teda ktorý druh obýva ktorý ostrov. Bol predpoklad, že jeden druh obýva len jeden ostrov. Graf 2.6 to do istej miery aj potvrdzuje, pretože tučniaky druhu Chinstrap a Gentoo sa naozaj nachádzajú len na jednom ostrove. Druh Adelie je ale na všetkých.



Obr. 2.6: Distribúcia druhov po ostrovoch

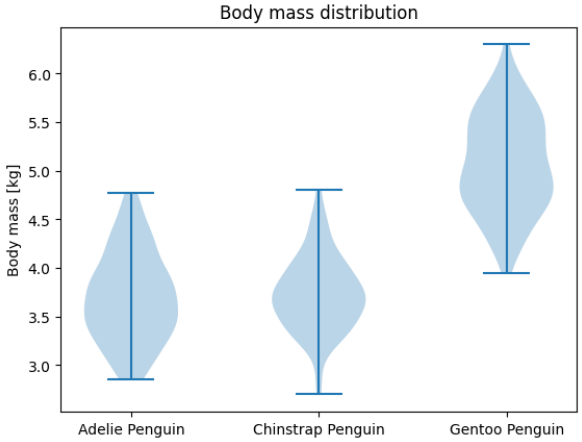
Vzťah medzi dĺžkou a výškou zobáka je niečo, čo by človek prirodzene čakal. Napriek tomu korelácia je takmer nulová. Po vygrafovaní týchto parametrov vznikol pomerne chaotický graf, ktorý ale okamžite začal dávať zmysel pri dodatočnom rozdelení podľa druhu (Fig 2.7). Na obrázku je jasne vidieť zhluky bodov, zodpovedajúce jednotlivým druhom. Rozmery zobáka by mohli byť

potenciálne klasifikačné kritérium druhu.



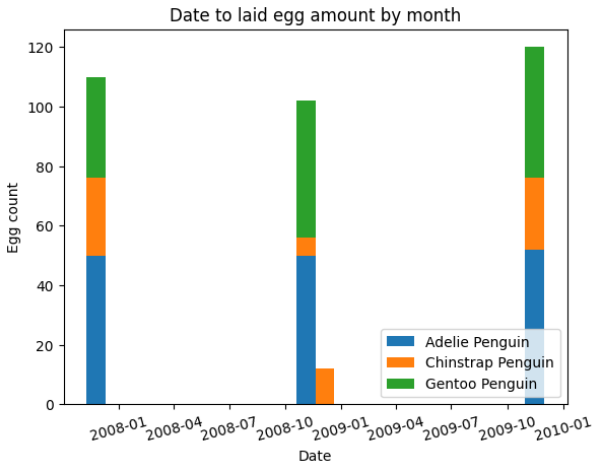
Obr. 2.7: Porovnanie výšky a dĺžky zobáka vzhľadom ku druhu.

Pre porovnanie distribúcií hmotností podľa druhu sa hodí husľový graf, ako vidno na 2.8. Prekvapivo to vyzerá, že jednotlivé druhy majú pomerne odlišné distribúcie.



Obr. 2.8: Husľový graf znázorňujúci distribúcie hmotností tučniakov rozdelených podľa druhu.

Posledný vzťah, na ktorý sme sa pozreli ako potenciálneho kandidáta na klasifikáciu druhu, bol čas kladenia vajec. Na tento účel sme spravili histogram 2.9, na ktorom ale vidno, že všeky druhy kladú vajcia v priibližne rovnakom čase.



Obr. 2.9: Histogram kladenia vajec vzhľadom na čas

2.3 Odlehlé hodnoty

Odlehlé hodnoty v datové sadě jsme hledali pomocí mezikvartilového rozpětí (IQR). Pro odlehlou hodnotu musí platit, že je menší než $q_1 - 1.5 \cdot \text{IQR}$ nebo větší než $q_3 + 1.5 \cdot \text{IQR}$, kde q_1 a q_3 jsou hodnoty prvního a třetího kvartilu. Takové hodnoty má smysl hledat pouze u numerických atributů.

Pro žádný z atributů jsme nenašli odlehlé hodnoty, když jsme uvažovali celou datovou sadu. Jelikož se ale hodnoty atributů jako délka křídel a zobáku určitě liší v závislosti na druhu tučňáka, rozhodli jsme se hledat odlehlé hodnoty atributů pro jednotlivé druhy tučňáků. Tam už jsme odlehlé hodnoty našli.

2.4 Chybějící hodnoty

Dále jsme zkoumali počet chybějících hodnot u jednotlivých atributů. Výsledky analýzy jsou uvedeny v tabulce 2.1. Celkově chybí 35 numerických hodnot, 10 hodnot u pohlaví tučňáka a 318 komentářů. Chybějící komentáře ovšem nejsou důležité, obsahují pouze nějakou přidanou informaci a neočekává se, že by tato hodnota měla být přítomna u všech tučňáků.

Pokud pomineme atribut `Comments`, je zcela vyplněných řádků v tabulce 325, těch kde tedy chybí alespoň jedna hodnota je 19. (Řádků, které jsou zaplněny zcela, včetně komentářů, je 13.) Z celkových 19 řádků je 13 řádků, ve kterých chybí hodnota více než jednoho atributu. Nejvíce hodnot chybí na řádku 3 a 339 (což odpovídá řádkům 5 a 341 po otevření v Excelu) – na každém z těchto řádků je 7 chybějících hodnot.

Název atributu	**Počet chybějících hodnot**
Name	0 / 344
Sample number	0 / 344
Species	0 / 344
Island	0 / 344
Stage	0 / 344
Individual ID	0 / 344
Clutch completion	0 / 344
Date egg	0 / 344
Sex	10 / 344
Comments	318 / 344
Culmen length	2 / 344
Culmen depth	2 / 344
Flipper length	2 / 344
Body mass	2 / 344
Delta 13	13 / 344
Delta 15	14 / 344

Tabuľka 2.1: Počty chybějících hodnot

Kapitola 3

Príprava dátovej sady pre dolovacie algoritmy

Dolovací úloha a odstranění irelevantních atributů

Jako dolovací úloha byla zvolena klasifikace druhů tučňáků na základě ostatních atributů. Jako atributy, ze kterých se dá druh tučňáka odvodit, jsme zvolili `Island`, `Culmen length`, `Culmen Depth`, `Flipper Length` a `Body Mass`. Vliv hodnot těchto atributů na druh tučňáka můžeme pozorovat v grafech 2.5, 2.6 a 2.7. Ostatní atributy považujeme v naší dolovací úloze za irelevantní a proto jsme je z datové sady odstranili.

Odlehlé hodnoty

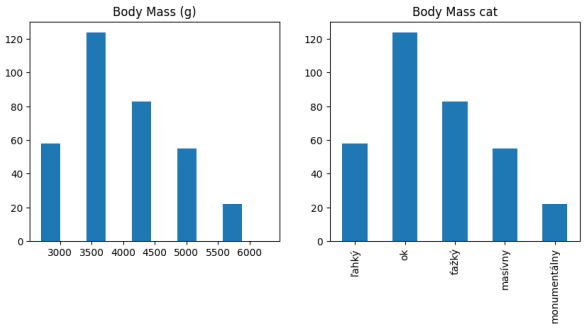
Jak již bylo diskutováno v kapitole 2.3, odlehlé hodnoty jsou v datové sadě přítomné pouze v případě, že zkoumáme hodnoty atributů pro každý druh tučňáka zvlášť. Jelikož ale v naší upravené datové sadě pro dolování nemáme atribut `Species`, toto rozdělení neprovádíme a datová sada žádné odlehlé hodnoty neobsahuje.

Chybějící hodnoty

V upravené datové sadě se vyskytují pouze dva záznamy, které obsahují chybějící hodnoty. V první variantě datové sady se s chybějícími hodnotami vypořádáme tak, že celé tyto dva záznamy odstraníme. Ve druhé variantě datové sady nahradíme hodnoty chybějících atributů průměrem ze všech přítomných hodnot pro daný atribut.

Diskretizácia kategorických atribútov

`Body Mass (g):` Pri zobrazení histogramu hodnôt sme zhodnotili, že ich distribúcia dostatočne zodpovedá normálnemu rozdeleniu. Rozhodli sme sa použiť techniku equi-width binning. Vytvoreným kategorickým atribútom je `Body Mass cat`.



Obr. 3.1: Histogram distribúcie hodnôt pôvodného atribútu `Body Mass (g)` a nového atribútu `Body Mass cat`.

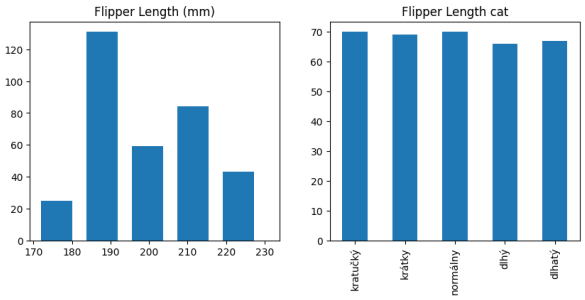
ľahký
ok
ťažký
masívny
monumentálny

Tabuľka 3.1: Ordinálne zoradenie hodnôt atribútu `Body Mass cat`

`Flipper Length (mm):` Distribúcia hodôt nezodpovedá normálnemu rozdeleniu a tak je možné, že nie je správne zastúpený pomer jednotlivých počtov tučniakov s dĺžkou krídel príslušnej dĺžky (časť intervalu mohla byť podhodnotená). Pre tento atribút používame equi-depth binning, ktorý by mal toto podhodnotenie minimalizovať. Tieto transformované hodnoty sa nachádzajú v atribúte `Flipper Length cat`.

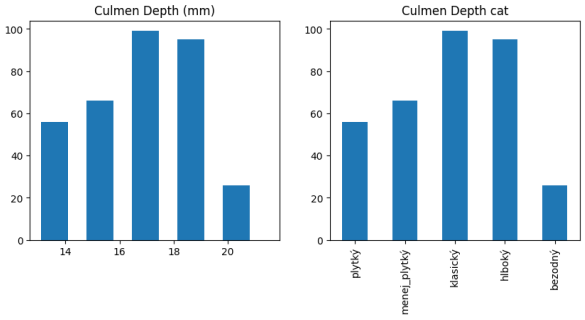
kratučký
krátky
normálny
dlhý
dlhatý

Tabuľka 3.2: Ordinálne zoradenie hodnôt atribútu `Flipper Length cat`



Obr. 3.2: Histogram distribúcie hodnôt pôvodného atribútu `Flipper Length (mm)` a nového atribútu `Flipper Length cat`.

`Culmen Depth (mm):` Po vykreslení histogramu sme opäť posúdili, že ide o normálne rozdelenie a teda používame kategórie reprezentujúce skupiny pevných dĺžok (equi-width binning). Kategorickým atribútom reprezentujúcim túto transformáciu je `Culmen Depth cat`.

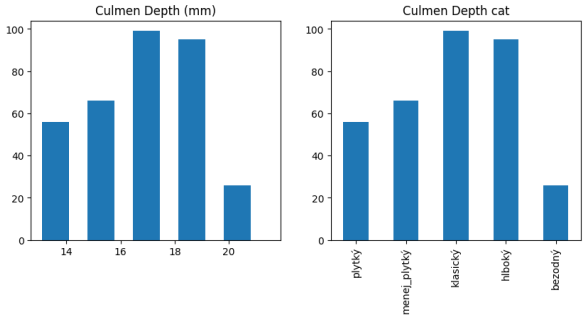


Obr. 3.3: Histogram distribúcie hodnôt pôvodného atribútu `Culmen Depth (mm)` a nového atribútu `Culmen Depth cat`.

plytký
menej_plytký
klasický
hlboký
bezodný

Tabuľka 3.3: Ordinálne zoradenie hodnôt atribútu `Culmen Depth cat`

`Culmen Length (mm):` Na rovnakom závere sme sa zhodli aj pri tomto atribúte, ktorého transformácia je obsiahnutá v atribúte `Culmen Length cat`.



Obr. 3.4: Histogram distribúcie hodnôt pôvodného atribútu `Culmen Length (mm)` a nového atribútu `Culmen Length cat`.

kratučký
krátky
normálny
dlhý
dlhatý

Tabuľka 3.4: Ordinálne zoradenie hodnôt atribútu `Culmen Length cat`

Takto vytvorenú dátovú sadu s kategorickými atribútmi sme vyexportovali do súboru `penguins_categorical.csv`.

Normalizácia numerických atribútov

Pre druhú dátovú sadu, ktorá má byť určená pre modely pracujúce s numerickými atribútmi sme normalizovali hodnoty týchto numerických atribútov. Použili sme techniku normalizácie *min-max*. Transformovanými numerickými atribútmi sú `Body Mass (g)`, `Flipper Length (mm)`, `Culmen Depth (mm)` a `Culmen Length (mm)`.

Transformácia kategorických atribútov na numerické

V datasete sa vyskytuje jediný kategorický atribút `Island`. Pre jeho prevod do numerického atribútu sme použili technik One-Hot Encoding. V našom prípade je vhodné použiť, nakoľko tento atribút nadobúda len tri rozličné hodnoty, čo znamená, že pôvodný atribút bude nahradený tromi novými atribútmi, z ktorých každý bude reprezentovať jednu z možných hodnôt atribútu `Island`. Každý z týchto atribútov bude obsahovať hodnotu 1 ak daný sledovaný objekt mal v pôvodnom atribúte jeho hodnotu. Inak bude obsahovať hodnotu 0.

Takto vytvorenú dátovú sadu s numerickými atribútmi sme vyexportovali do súboru `penguins_numerical.csv`.