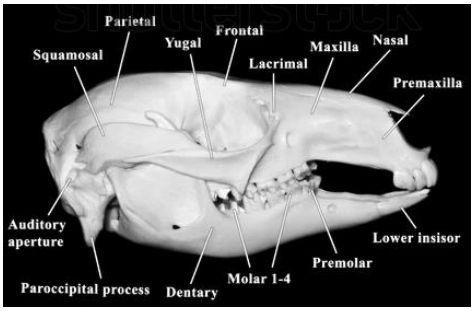
PCA kraniometrie klokanů

Aplikovaná statistika II - Domácí úkol

Tomáš Spurný

11. května 2023

Nahrání a představení dat



Dataset obsahuje údaje z **měření lebek klokanů** (jde o upravenou verzi datasetu kanga z knihovny faraway z knihy Data z roku **1985**).

U každého klokana z 148 byl určen **druh, pohlaví a 12** lebečních rozměrů. Přičemž lebka klokana je zachycena na následujícím obrázku a přejmenování[1] rozměrů v prvním bloku kódu v proměnné `VARIABLES`.

Jsou zde zachyceni klokani tří druhů – všechny se liší velikostí, vzhledem a rozšířením:

- **Klokan velký** *Macropus fuliginosus*: anglicky western grey kangaroo, je třetí největší druh klokana po klokanovi obrovském (*Macropus giganteus*) a klokanovi rudém (*Macropus rufus*). Vyskytuje se v jižní až jihozápadní Austrálii. Má šedou krátkou srst. Dle přesnějšího výskytu se rouzlišuje poddruh *Macropus fuliginosus fuliginosus* (endemit na Klokaním ostrově) a poddruh *Macropus fuliginosus melanops* (zbytek zmíněného území)
- **Klokan obrovský** *Macropus giganteus*: anglicky eastern grey kangaroo je druhý největší druh klokana, žije ve východní Austrálii a Tasmánii. Má šedou dlouhou srst a širší ušní boltce.
- *Macropus melanops*: v datech jde nejspíš o poddruh *Macropus fuliginosus melanops*.

Spolu s nahráním, byla zjištěna struktura dat a nahrány konstantní vektory pro překódování názvů proměnných a barvy pro druhy klokanů. Na začátku práce byly také nahrány knihovny pomocí kódu č. 0 v příloze.

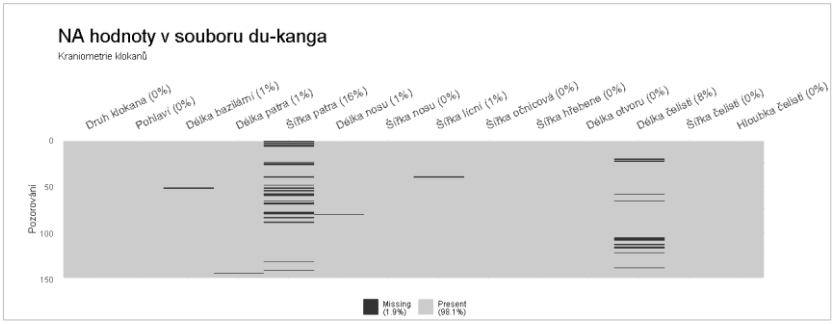
```
df <- read.delim('./du-kanga.txt', sep=" ", header=T,
                 stringsAsFactors = T, na.strings="NA")

VARIABLES <- c("Druh klokana" = "species",
               "Pohlaví" = "sex",
               "Délka bazilární" = "basilar.length",
               "Délka patra" = "palate.length",
               "Šířka patra" = "palate.width",
               "Délka nosu" = "nasal.length",
               "Šířka nosu" = "nasal.width",
               "Šířka lícní" = "zygomatic.width",
               "Šířka očnicová" = "orbital.width",
               "Šířka hřebene" = "crest.width",
               "Délka otvoru" = "foramina.length",
               "Délka čelisti" ="mandible.length",
               "Šířka čelisti" ="mandible.width",
               "Hloubka čelisti" ="mandible.depth")

KLOKANI <- c("fuliginosus" = "#90353b", "giganteus" = "#1a476f", "melanops" = "#dfac56")
REV_VARIABLES <- setNames(names(VARIABLES), VARIABLES)
setnames(df, old = VARIABLES, new=names(VARIABLES))
vis_miss(df)
```

[1] Nestuduji antropologii, takže moje překlady proměnných do češtiny nejsou nejspíš správné.

Data obsahují 1,9 % NA hodnot. Nejvíc neúplné jsou veličiny: šířka patra a délka čelisti. Chybějící pozorování byly tedy odstraněny a byly tak vytvořeny dvě tabulky: celistvá **`df0`** a **`dfNum`** s pouze numerickými veličinami.




```
df0 <- df |> #data pro analýzu
  drop_na()

dfNum <- df |> #spojité proměnné
  drop_na() |>
  select_if(is.numeric)
nrow(df0)/nrow(df) #% úplných pozorování
## [1] 0.7567568
```

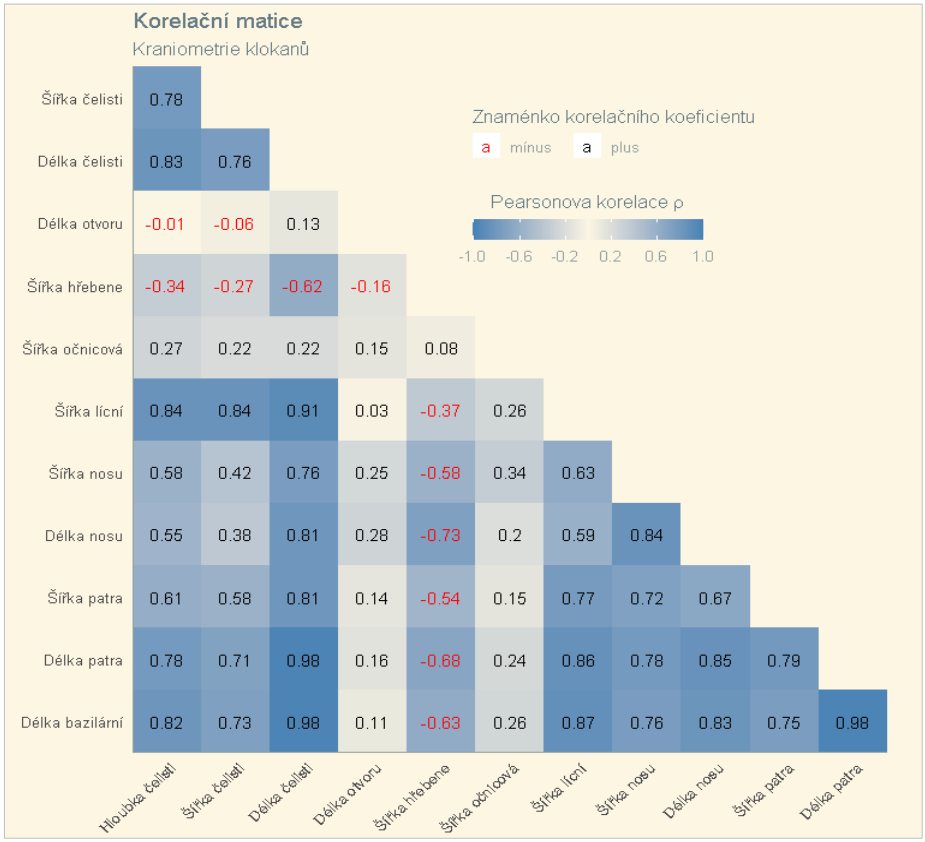
Pozůstavši data tvoří 76% originálu.

Korelační analýza

Na numerických datech byly zhodnoceny vazby mezi proměnnými pomocí korelační matice.

```
CorMat <- cor(dfNum, method = "pearson")
```

Ta byla vynesena do heatmapy níže pomocí kódu č. 1 v příloze.

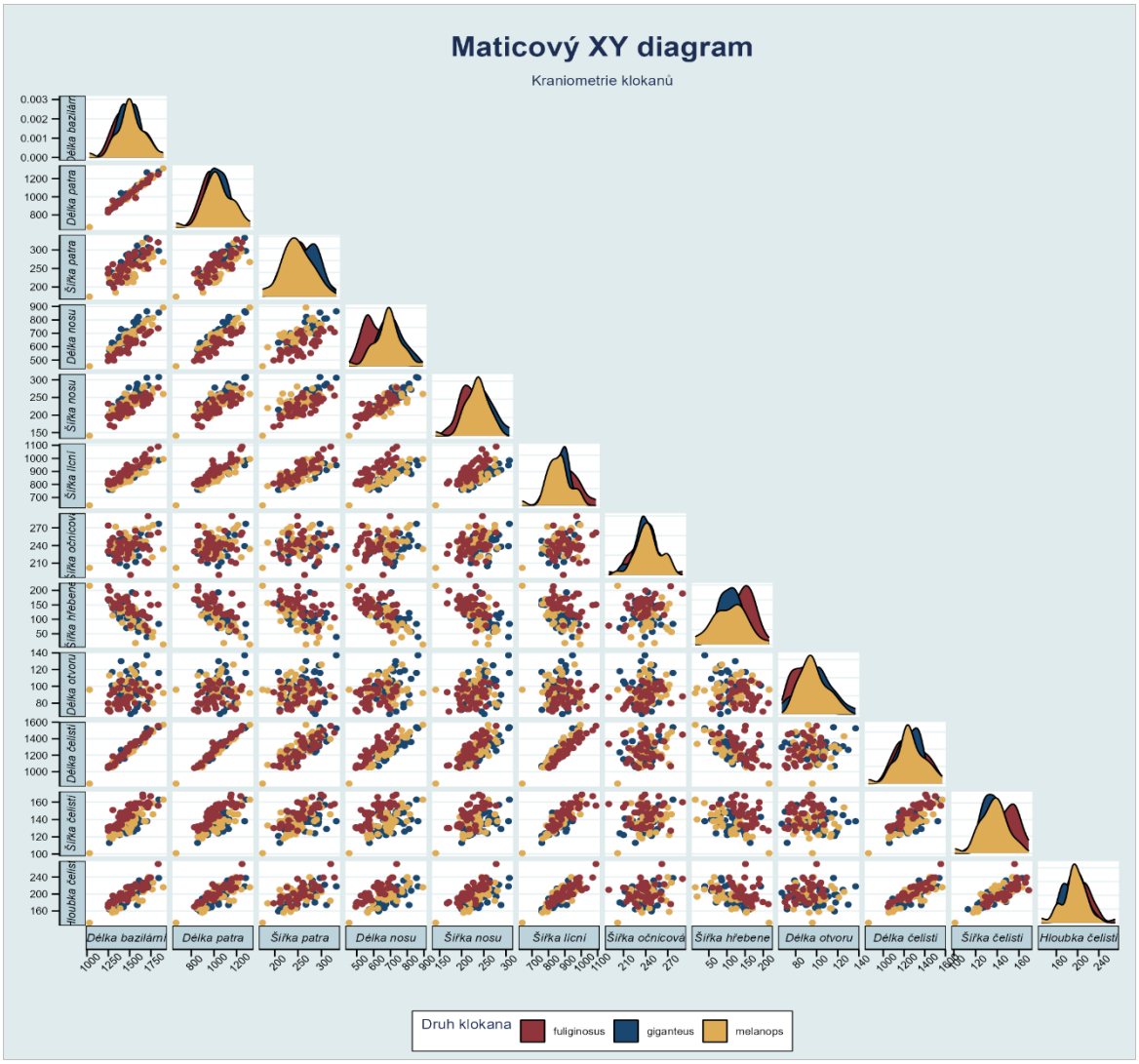


Negativní korelaci vykazuje se všemi rozměry **šířka hřebene**. Ostatní nezanedbatelné korelace jsou všechny kladné. Malé korelace vykazuje se všemi **délka otvoru a očnicová šířka**. Silné korelace se obvykle týkají rozměrů na **jedné lebeční části**.

Příklady:

- Silné pozitivní korelace (1-0,7): šířka a délka nosu (0,84)
- Negativní korelace: šířka hřebene a délka nosu (-0,73)
- Slabé korelace (0,1-0,3): šířka patra a očnicová šířka

Korelace byla zachycena i znázorněním v XY grafech pomocí kódu č. 2 v příloze. Lze pozorovat zápornou korelaci u šířky hřebene i silné korelace na jedné lebeční části (např. nos). V některých grafech se zdá, že se jednotlivé druhy systematicky liší – např. druh *Macropus fuliginosus* má menší délku čelisti při stejné lícní šířce než zbylé dva druhy a opačně např. u dvojice délka čelisti – šířka nosu.

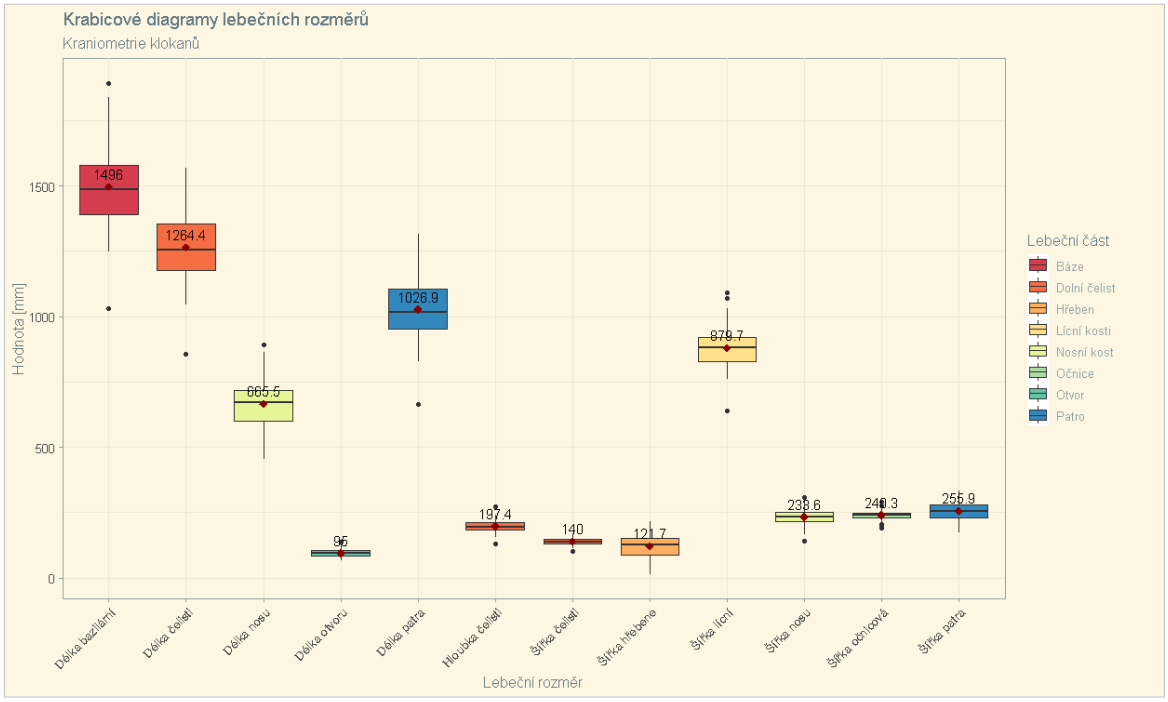


Rozpětí naměřených hodnot a škálování

Byly vypočteny průměry veličin. Nejmenším rozměrem je délka otvoru a největším bazilární délka.

```
colMeans(dfNum)
##  Délka bazilární      Délka patra      Šířka patra        Délka nosu 
##        1496.0179        1026.9107         255.9107         665.4911 
##        Šířka nosu       Šířka lícní  Šířka očnicová  Šířka hřebene 
##         233.6250         879.7143         240.2857         121.7054 
##     Délka otvoru     Délka čelisti  Šířka čelisti Hloubka čelisti 
##          95.0000        1264.4196         140.0179         197.4375
```

Proměnné byly znázorněny graficky v krabicových diagramech (kód č. 3 v příloze), barva diagramu určuje **jednu část lebky**, na které je měřen jeden nebo více rozměrů. Z polohy statistik boxplotu lze říci, že proměnné jsou přibližně **normálního** rozdělení, ale obsahují odlehlé hodnoty. Je vidět, že celkové rozpětí hodnot je veliké, proto se při PCA **provede škálování**. Lingvisticky: většina šířek je kratších než většina délek.



Vlastní analýza hlavních komponent

Barlettův test sféricity

Slouží k určení, jestli jsou veličiny dostatečně korelované, aby mělo význam provádět PCA. Nulová hypotéza tvrdí, že výběrová korelační matice je jednotková, tzn. ostatní výběrové korelační koeficienty jsou nulové.

```
cortest.bartlett(cor(dfNum), nrow(dfNum), diag = TRUE)
## $chisq
## [1] 1880.141
## $p.value
## [1] 0
## $df
## [1] 66
p <- ncol(dfNum)
(kvantil <- qchisq(0.95, df=p*(p-1)/2))
## [1] 85.96491
```

Testová statistika (1880,1) se realizuje hluboko v kritickém oboru <85,96; ∞), p-hodnota je nulová. Zamítáme tedy nulovou hypotézu a můžeme přistoupit k PCA. **Data jsou dostatečně korelovaná**.

Vlastní PCA

Pomocí příkazu `prcomp()`, byla provedena analýza hlavních komponent pro spojité proměnné souboru du-kanga.csv.

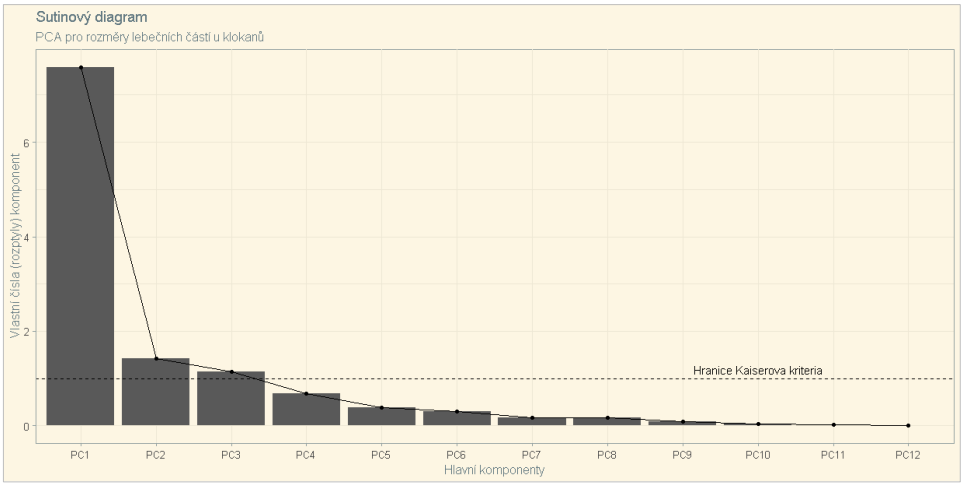
```
dataPCA <- prcomp(dfNum, center = TRUE, scale. = TRUE)
summary(dataPCA)
## Importance of components:
##                           PC1    PC2    PC3    PC4    PC5    PC6    PC7    PC8
## Standard deviation     2.7526 1.1924 1.06557 0.82094 0.62345 0.54574 0.41546 0.41124
## Proportion of Variance 0.6314 0.1185 0.09462 0.05616 0.03239 0.02482 0.01438 0.01409
## Cumulative Proportion  0.6314 0.7499 0.84450 0.90066 0.93305 0.95787 0.97225 0.98634
##                            PC9    PC10    PC11   PC12
## Standard deviation     0.30310 0.21304 0.12107 0.1094
## Proportion of Variance 0.00766 0.00378 0.00122 0.0010
## Cumulative Proportion  0.99400 0.99778 0.99900 1.0000
varTable <- summary(dataPCA)$importance[c(2,3),]
rownames(varTable) <- c("Relativní variabilita", "Kumulativní relativní variabilita")
```

Byla vytvořena tabulka podílů variabilit. Prvními třemi hlavními komponentami je vysvětleno 84,5 % variability. První komponenta vysvětluje 63,1 % variability. Další dvě, každá kolem 10 %.

Tabulka podílů variabilit												
Hlavní komponenta	**PC1**	**PC2**	**PC3**	**PC4**	**PC5**	**PC6**	**PC7**	**PC8**	**PC9**	**PC10**	**PC11**	**PC12**
Relativní variabilita	0,631	0,118	0,095	0,056	0,032	0,025	0,014	0,014	0,008	0,004	0,001	0,001
Kumul. rel. variabilita	0,631	0,75	0,845	0,901	0,933	0,958	0,972	0,986	0,994	0,998	0,999	1

Volba počtu hlavních komponent

Pomocí kódu č. 4 v příloze byl vykreslen sutinový diagram:



Podle něj lze popsat původní datový soubor bez podstatné ztráty informace prvními 2 hlavními komponentami. Podle Kaiserova kritéria jsou potřeba první 3 komponenty. To potvrzuje i požadavek na vysvětlení alespoň 80 % variability, jelikož první dvě komponenty tak činí pouze z 75 %. Omezíme se tedy na **první 3 hlavní komponenty**.

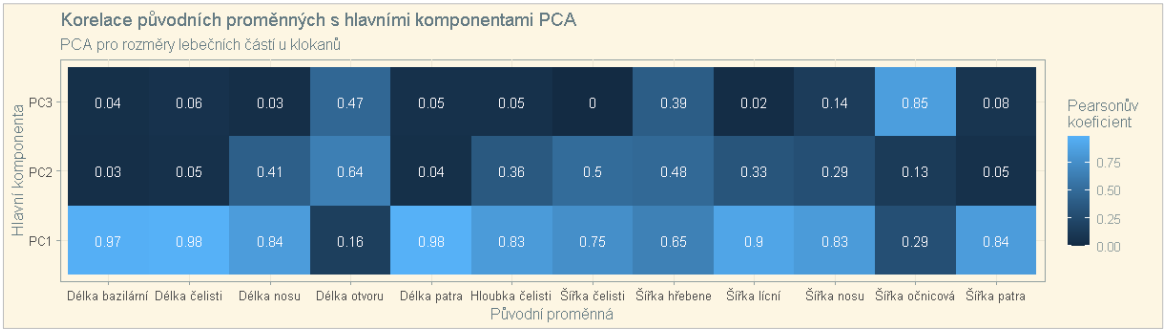
Zhodnocení PCA

Korelace: původní proměnné vs. hlavní komponenty

Byla vypočtena korelace původních proměnných a prvních třech hlavních komponent a vynesena do heatmapy.

```
df.in.pc <- dataPCA$x
CorORIGxPCA <- cor(dfNum, df.in.pc[,1:3])

CorORIGxPCA |>
  round(2) |>   data.frame() |>
  rownames_to_column("Původní proměnná") |>
  pivot_longer(cols = PC1:PC3, names_to = "Hlavní komponenta",
               values_to = "Pearsonův\nkoeficient") |>
  ggplot(aes(`Původní proměnná`, `Hlavní komponenta`, fill= `Pearsonův\nkoeficient`,
             label= `Pearsonův\nkoeficient`)) +
  geom_tile() +
  geom_text(color = "#A0ACB8") +
  scale_fill_viridis(discrete=FALSE, option = "inferno") +
  labs(title = "Korelace původních proměnných s hlavními komponentami PCA",
       subtitle = "PCA pro rozměry lebečních částí u klokanů")
```



První hlavní komponenta je vysoce korelovaná se všemi veličinami krom délky otvoru a očnicové šířky. Nejvíce s délkou patra, délkou čelisti a bazilární délkou.

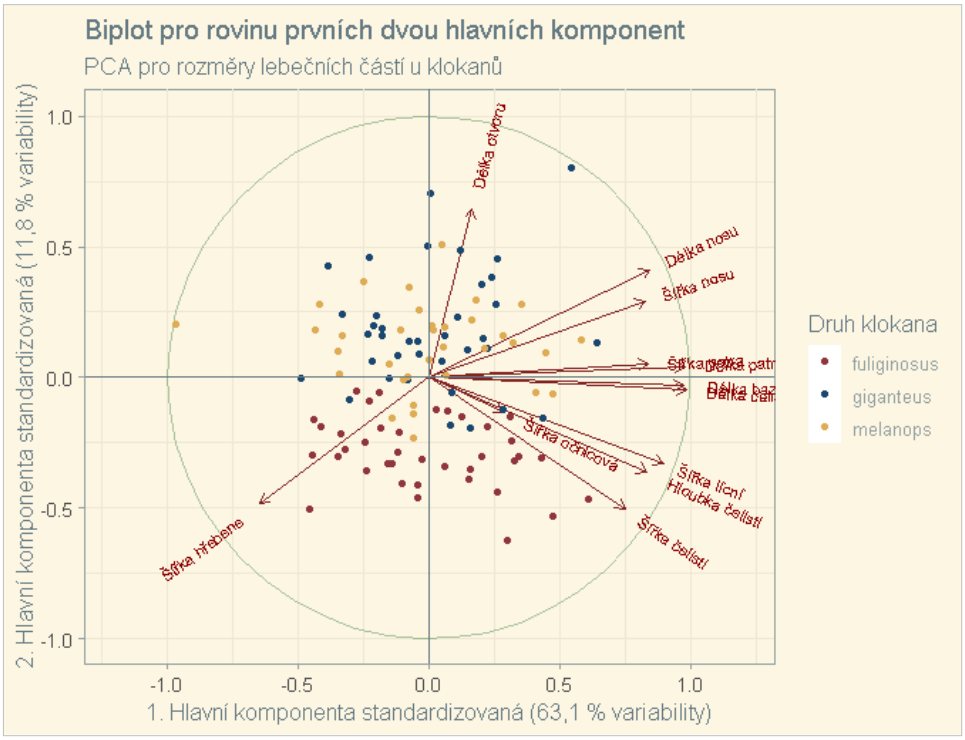
Druhá hlavní komponenta nejvíce koreluje s **délkou otvoru**. Dále i s šířkou čelisti, šířkou hřebene a délkou nosu. Má spoustu korelací, které nejsou zanedbatelné.

Třetí komponenta má vysokou korelaci s **očnicovou šířkou** a rozlišuje tedy klokany na základě ní.

Biplot

Pro vytvoření biplotu byla použita knihovna ggbiplot, resp. její upravená verze, která dokáže vykreslit i jednotkový korelační kruh a data přeškáluje.

```
source("ggbiplot-update.R")
ggbiplot2(dataPCA, circle = T, groups = df0$`Druh klokana`) +
  geom_hline(yintercept = 0, color = "#657b83") +
  geom_vline(xintercept = 0, color = "#657b83") +
  scale_x_continuous(limits = c(-1.2, 1.2)) +
  scale_y_continuous(limits = c(-1, 1)) +
  xlab("1. Hlavní komponenta standardizovaná (63,1 % variability)") +
  ylab("2. Hlavní komponenta standardizovaná (11,8 % variability)") +
  labs(title = "Biplot", subtitle = "PCA pro rozměry lebečních částí u klokanů",
       colour = "Druh klokana") +
  coord_equal()
```

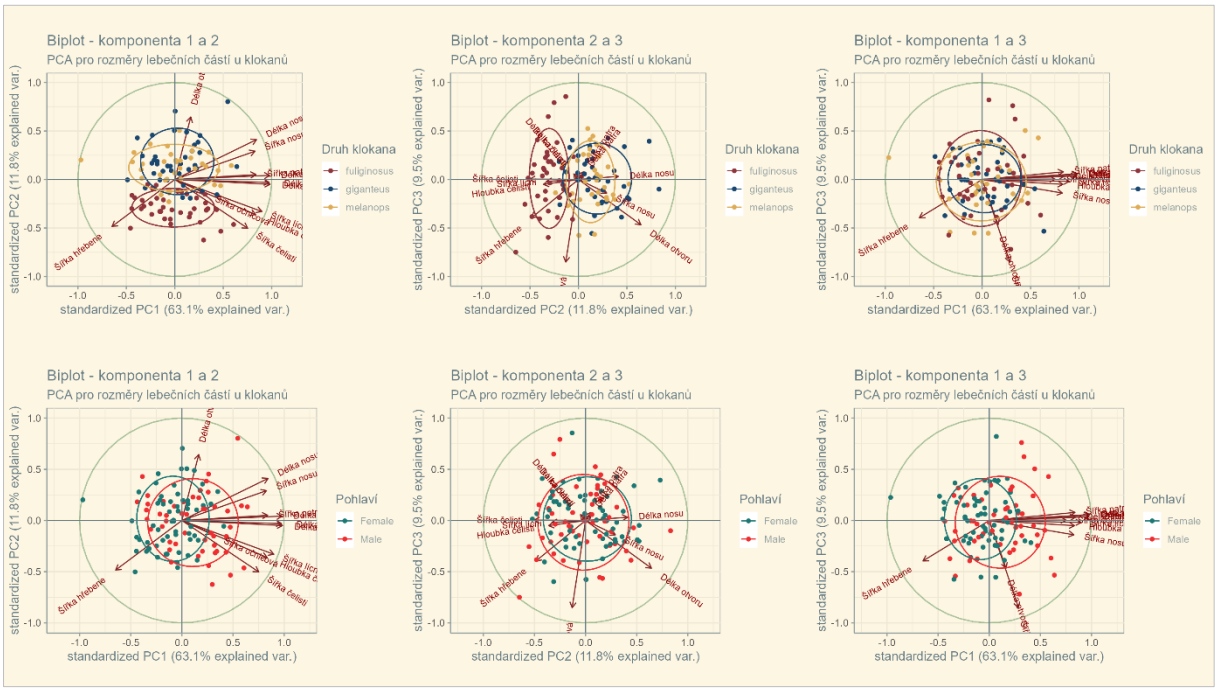


Je znát, že rozměry, které byly v původních datech silně korelované **jsou pospolu** (skupinka pro nos, čelist-líční kosti-očnice, patro-bázi-čelist). Skupina veličin **patro-báze-čelist** nejvíce souvisí s první hlavní komponentou. Šířka hřebene, která vykazovala záporné korelace stojí **oproti ní**.

Rozměry, které měly zanedbatelné korelace v původním souboru mají krátkou šipku anebo jsou orientované kolměji, resp. odpovídají druhé hlavní komponentě.

V rovině 1. a 2. komponenty lze rozlišit **klokany *Macropus fuliginosus* od zbylých dvou druhů** - ty zbylé se překrývají. Stejná situace je i v rovině 2. a 3. komponenty. V rovině 3. a 1. komponenty nejdou klokani rozeznat vůbec. Roviny nejsou dostačující k rozlišení všech tří druhů klokanů.

Všechny kombinace rovin hlavních komponent následují a byly vygenerovány pomocí kódu č. 5. Horní tři se týkají dělení na druhy, dolní tři zase dělení na pohlaví.



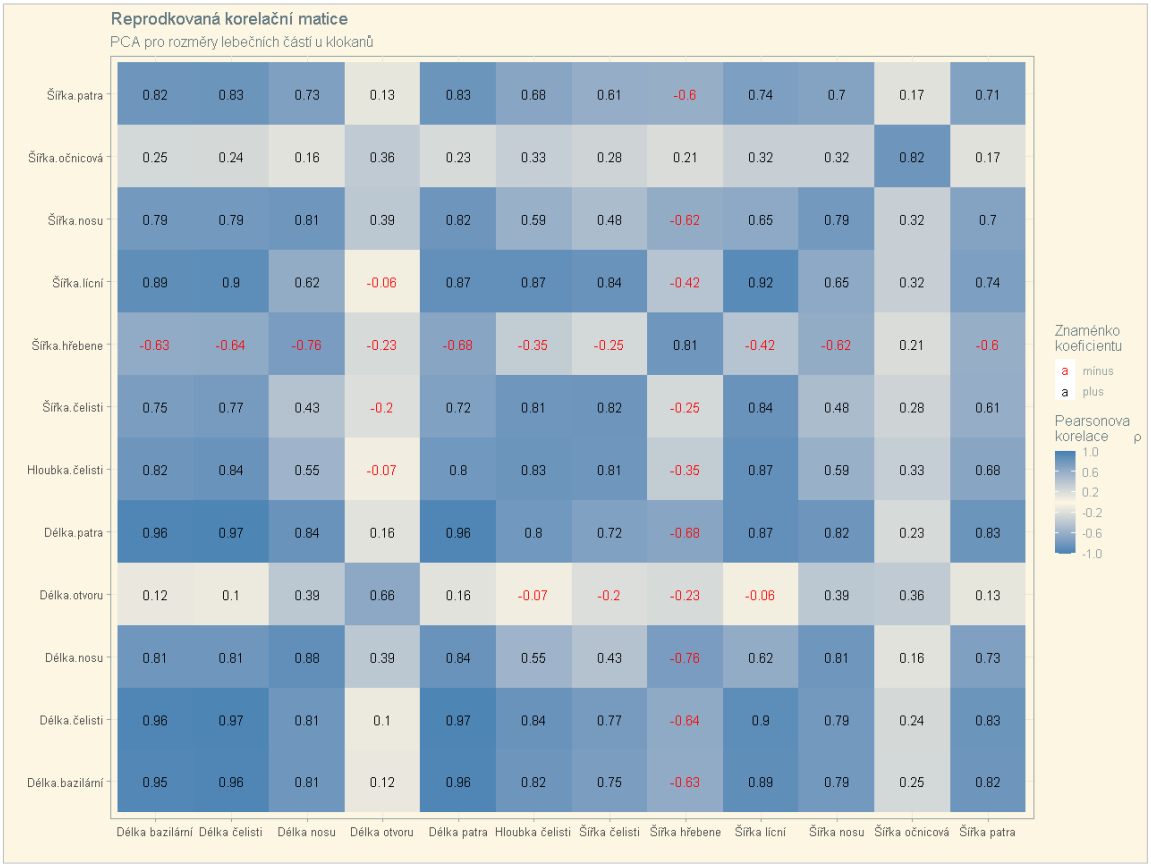
Reprodukovaná a reziduální korelační matice

Pomocí maticového násobení byla získána reprodukovaná matice výběrových korelací a residuální korelační matice.

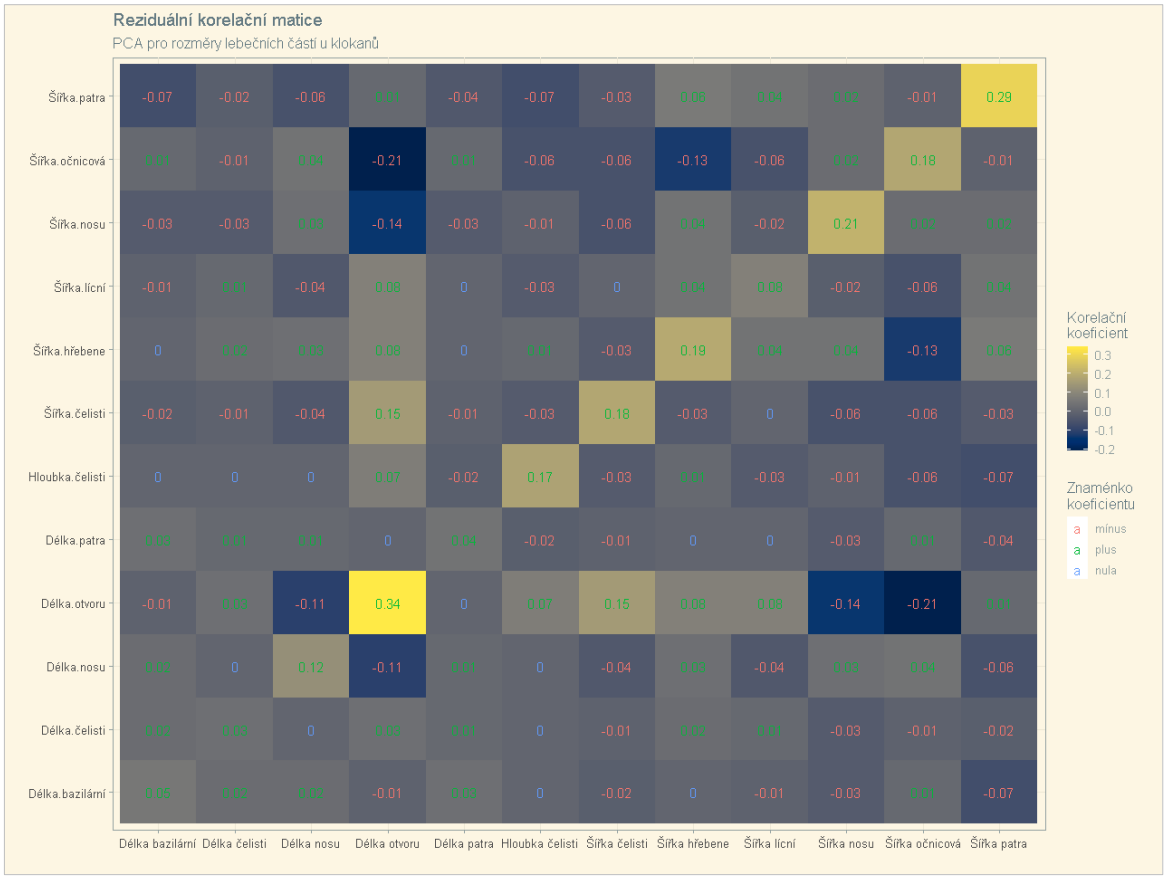
```
vl.cisla.pca <- dataPCA$sdev^2
pc <- dataPCA$rotation

corel.reproduced <- pc[,1:3] %*% diag(vl.cisla.pca[1:3]) %*% t(pc[,1:3])
corel.residual <- CorMat - corel.reproduced

corel.reproduced |>
  round(2) |>
  data.frame() |>
  rownames_to_column("V1") |>
  pivot_longer(cols = -V1, names_to = "V2", values_to = "Korelační\nkoeficient") |>
ggplot(aes(V1, V2, fill = `Korelační\nkoeficient`, label = `Korelační\nkoeficient`)) +
  geom_tile() +
  scale_fill_viridis(option = "plasma")+
  labs(x = "", y = "", title = "Reprodkovaná korelační matice",
       subtitle = "Kraniometrie klokanů")
```




```
corel.residual |>
  round(2) |>
  data.frame() |>
  rownames_to_column("V1") |>
  pivot_longer(cols = -V1, names_to = "V2", values_to = "Korelační\nkoeficient") |>
 ggplot(aes(V1, V2, fill = `Korelační\nkoeficient`, label = `Korelační\nkoeficient`)) +
  geom_tile() +
  scale_fill_viridis(option = "cividis")+
  labs(x = "", y = "", title = "Reprodkovaná korelační matice", subtitle = "Kraniometrie
klokanů")
```



V reziduální korelační matici je 17 rozdílů v intervalu (0,10; 0,35), 7 z nich leží na hlavní diagonále, která je v reprodukované matici podhodnocována. Výběrové korelace jsou spíše nadhodnocovány.

Jinak se rezidua pohybují blízko nule, takže by stávající volba hlavních komponent by mohla být dostačující k vystihnutí původního datového souboru.

Příloha – kódy

Kód č. 1 — Heatmapa korelační matice

```
CorVal <- CorMat # hodnoty pro graf
CorVal[lower.tri(CorMat, diag = T)] <- NA

CorVal |>
  round(2) |>
  as.data.frame() |>
  rownames_to_column("V1") |>
  pivot_longer(-V1, names_to = "V2") |>
  drop_na() |>
  mutate(V1_ = as.factor(V1), V2_ = as.factor(V2)) |>
  mutate(V1 = V1_ |> recode(!!! REV_VARIABLES) |> factor(levels=REV_VARIABLES),
         V2 = V2_ |> recode(!!! REV_VARIABLES) |> factor(levels=rev(REV_VARIABLES))) |>
  data.frame() |>
ggplot(aes(V2, V1, fill = value)) +
  geom_tile(aes(V2, V1, fill = value)) +
  geom_text(aes(V2, V1, label = value, color = factor(value/abs(value))), size = 4) +
  scale_fill_gradient2(name=expression("Pearsonova korelace" * ~ rho),
                       mid = "#fdf6e3", low = "steelblue", high = "steelblue",
                       breaks=seq(-1, 1, by = 0.4), limits = c(-1, 1)) +
  scale_color_manual(labels = c("mínus", "plus"), values = c("red", "black")) +
  scale_x_discrete(expand = c(0, 0)) +
  scale_y_discrete(expand = c(0, 0)) +
  labs(x = "", y = "", title = "Korelační matice", subtitle = "Kraniometrie klokanů",
       colour = "Znaménko korelačního koeficientu") +
  guides(fill = guide_colorbar(barwidth = 10, barheight = 1, title.position = "top",
                               title.hjust = 0.5)) + # guide_colorsteps
  guides(colour = guide_legend(title.position = "top")) +
  theme(axis.text.x = element_text(angle = 45, vjust = 1, hjust = 1),
        panel.grid.major = element_blank(),
        panel.border = element_blank(),
        panel.background = element_blank(),
        plot.margin = margin(1, 1, 0, 0, "cm"),
        axis.ticks = element_blank(),
        legend.justification = c(1, 0),
        legend.position = c(0.9, 0.7),
        legend.background = element_blank(),
        legend.direction = "horizontal") +
  coord_equal()
```

Kód č. 2 — Matice XY grafů

```
ggpairs(data=df0, columns = 3:14, switch="both",
mapping = aes(color=`Druh klokana`),     upper="blank",legend = 1)+
  theme_stata(base_size = 8) +
  scale_colour_manual(values = KLOKANI)  +
  scale_fill_manual(values = KLOKANI)  +
  labs(x = "", y = "", title = "Maticový XY diagram", subtitle = "Kraniometrie klokanů")+
  theme(plot.background = element_rect(fill = "#dfebec"),
        strip.background = element_rect(colour = "black", fill = "#bcd3dc"),
        strip.text = element_text(face = "italic", size = 8),
        plot.title = element_text(size = 20, face = "bold", vjust = 2),
        legend.position = "bottom",
        axis.text.y = element_text(angle = 0),
        axis.text.x = element_text(angle = 45),
        plot.subtitle = element_text(size = 10))
```

Kód č. 3 — Boxploty v jednom diagramu

```
fun_mean <- function(x){
  #funkce ke značení průměru v boxplotech
  return(data.frame(y=mean(x),label=round(mean(x,na.rm=T),1)))
}

CastiLebky <- c("bazilární" = "Báze",  "čelisti" = "Dolní čelist",
                "hřebene" = "Hřeben",  "lícní" = "Lícní kosti",
                "nosu" = "Nosní kost", "očnicová" = "Očnice",
                "otvoru" = "Otvor",    "patra" = "Patro")
dfNum |>
  pivot_longer(cols = `Délka bazilární`:`Hloubka čelisti`, names_to = "SkullDim",
               values_to = "Values") |>
  separate(SkullDim, into = c("Dimension", "SkullParts"), remove = F) |>
ggplot(aes(x = SkullDim, y =Values)) +
  geom_boxplot(aes(fill = SkullParts)) +
  scale_fill_brewer(palette="Spectral", labels = CastiLebky) +
  theme(axis.text.x = element_text(angle = 45, vjust = 1, hjust = 1)) +
  stat_summary(fun.y = mean, geom="point",colour="darkred", size=3, shape = 18) +
  stat_summary(fun.data = fun_mean, geom="text", vjust=-0.7) +
  labs(y = "Hodnota [mm]", x = "Lebeční rozměr", title = "Krabicové diagramy lebečních
       rozměrů", subtitle = "Kraniometrie klokanů", fill = "Lebeční část")
```

Kód č. 4 — Sutinový test

```
PCs <- c("PC1", "PC2", "PC3", "PC4", "PC5", "PC6", "PC7", "PC8", "PC9", "PC10", "PC11",
         "PC12")
data.frame(PC = colnames(dataPCA$x), Var = dataPCA$sdev^2) |>
  rownames_to_column("noPC") |>
  mutate(noPC = as.numeric(noPC)) |>
ggplot(aes(x = factor(PC, PCs), y = Var, group=1)) +
  geom_col() +
  geom_line() +
  geom_point() +
  geom_hline(yintercept=1, linetype="dashed") +
  annotate("text", x = "PC10", y = 1, label = "Hranice Kaiserova kriteria", vjust = -0.5) +
  labs(title = "Sutinový diagram", subtitle = "PCA pro rozměry lebečních částí u klokanů",
       x = "Hlavní komponenty", y = "Vlastní čísla (rozptyly) komponent")
```

Kód č. 5 — Biploty všech rovin

```
POHLA <- c("#1d7874", "#ee2e31") # barvy pro pohlaví
tab <- list(PC = list(rep(c(1,2,1), 2)), # list pro iteraci for loopu
            HK = list(rep(c(2,3,3), 2)),
            gr = list(rep(c("Druh klokana", "Pohlaví"), each = 3)),
            cl = list(KLOKANI, KLOKANI,KLOKANI, POHLA, POHLA,POHLA))
p <- list() # kontejner

for (each in 1:6){
  i = tab$PC[[1]][each]
  j = tab$HK[[1]][each]
  c = tab$gr[[1]][each]
  b = tab$cl[[each]]
  p[[each]] <- ggbiplot2(dataPCA, circle = T, ellipse = T,
                         choices = c(i, j), groups = df0[[c]]) +
    geom_hline(yintercept = 0, color = "#657b83") +
    geom_vline(xintercept = 0, color = "#657b83") +
    scale_x_continuous(limits = c(-1.2, 1.2)) +
    scale_y_continuous(limits = c(-1, 1)) +
    scale_colour_manual(values = b) +
    labs(title = paste0("Biplot - komponenta ", i, " a ", j),
         subtitle = "PCA pro rozměry lebečních částí u klokanů",colour = c) +
    coord_equal()
}
do.call(gridExtra::grid.arrange,c(p, nrow=2, ncol=3))
g <- do.call(gridExtra::arrangeGrob, c(p, nrow=2, ncol=3))
ggsave(filename = "gridBiplotu.png", plot = g ,width = 16, height = 9, dpi = 200,
       units = "in", device='png')
```