Project 4

This is the dataset you will be working with:

```
lemurs <- readr::read_csv('https://raw.githubusercontent.com/rfordatascience/tidytuesday/master/data/2021/2021-08-24/lemur_data.csv')

lemurs
```

```
## # A tibble: 82,609 × 54
##    taxon dlc_id hybrid sex   name     curre…¹ stud_…² dob        birth…³ estim…⁴
##    <chr> <chr>  <chr>  <chr> <chr>    <chr>   <chr>   <date>       <dbl> <chr>
##  1 OGG   0005   N      M     KANGA    N       <NA>    1961-08-25       8 <NA>
##  2 OGG   0005   N      M     KANGA    N       <NA>    1961-08-25       8 <NA>
##  3 OGG   0006   N      F     ROO      N       <NA>    1961-03-17       3 <NA>
##  4 OGG   0006   N      F     ROO      N       <NA>    1961-03-17       3 <NA>
##  5 OGG   0009   N      M     POOH BE… N       <NA>    1963-09-30       9 <NA>
##  6 OGG   0009   N      M     POOH BE… N       <NA>    1963-09-30       9 <NA>
##  7 OGG   0009   N      M     POOH BE… N       <NA>    1963-09-30       9 <NA>
##  8 OGG   0010   N      M     EEYORE   N       <NA>    1964-05-20       5 <NA>
##  9 OGG   0010   N      M     EEYORE   N       <NA>    1964-05-20       5 <NA>
## 10 OGG   0014   N      F     ROOLETTE N       <NA>    1964-10-27      10 <NA>
## # … with 82,599 more rows, 44 more variables: birth_type <chr>,
## #   birth_institution <chr>, litter_size <dbl>, expected_gestation <dbl>,
## #   estimated_concep <date>, concep_month <dbl>, dam_id <chr>, dam_name <chr>,
## #   dam_taxon <chr>, dam_dob <date>, dam_age_at_concep_y <dbl>, sire_id <chr>,
## #   sire_name <chr>, sire_taxon <chr>, sire_dob <date>,
## #   sire_age_at_concep_y <dbl>, dod <date>, age_at_death_y <dbl>,
## #   age_of_living_y <dbl>, age_last_verified_y <dbl>, …
```

```
length(unique(lemurs[["dlc_id"]]))
```

```
## [1] 2270
```

More information about the dataset can be found here: https://github.com/rfordatascience/tidytuesday/tree/master/data/2021/2021-08-24 (https://github.com/rfordatascience/tidytuesday/tree/master/data/2021/2021-08-24) and https://www.nature.com/articles/sdata201419 (https://www.nature.com/articles/sdata201419).

Question: *What are the differences between taxons when looking at their expected gestation length, litter size, age of conception of the mother and father, and weight?*

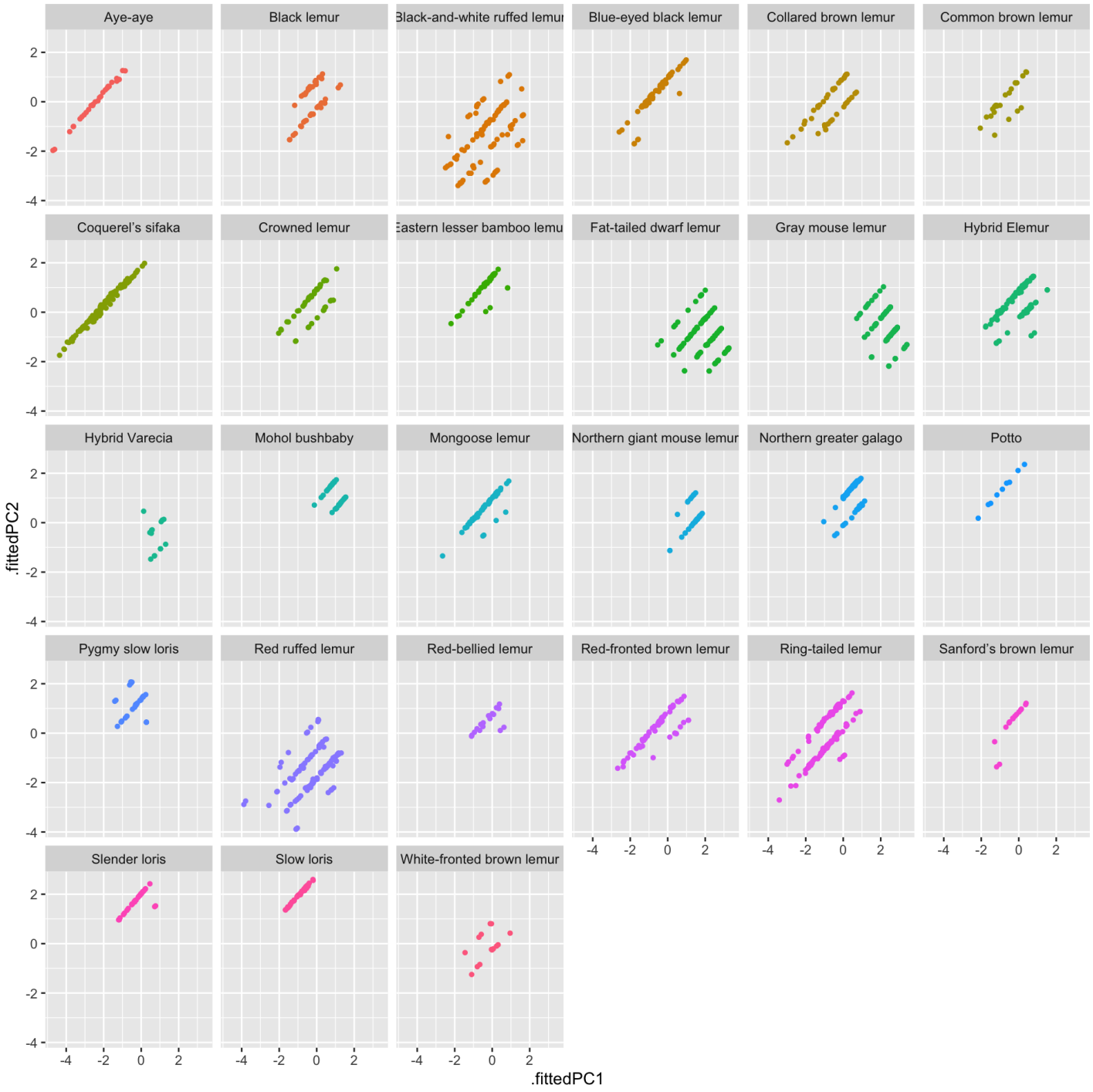
Introduction: *The Duke Lemur Center houses over 200 lemurs across 14 species – the most diverse population of lemurs on Earth, outside their native Madagascar. With information on 2270 different lemurs acquired by the DLC, this study will investigate differences between the taxons of lemurs with respect to features of their birth. We will investigate their expected gestation length (expected_gestation_d), litter size (litter_size), age of conception for the mother and father (sire_age_at_concep_y and dam_age_at_concep_y), and their weight (weight_g).*

Approach: *We only want the latest lemur information, so we'll keep only the rows which have the latest weight date per each DLC_id. Then we'll do PCA stuff: scale our numeric columns, perform PCA, calculate the components, and plot the PCs, rotation matrix, and variance explained.*

Analysis:

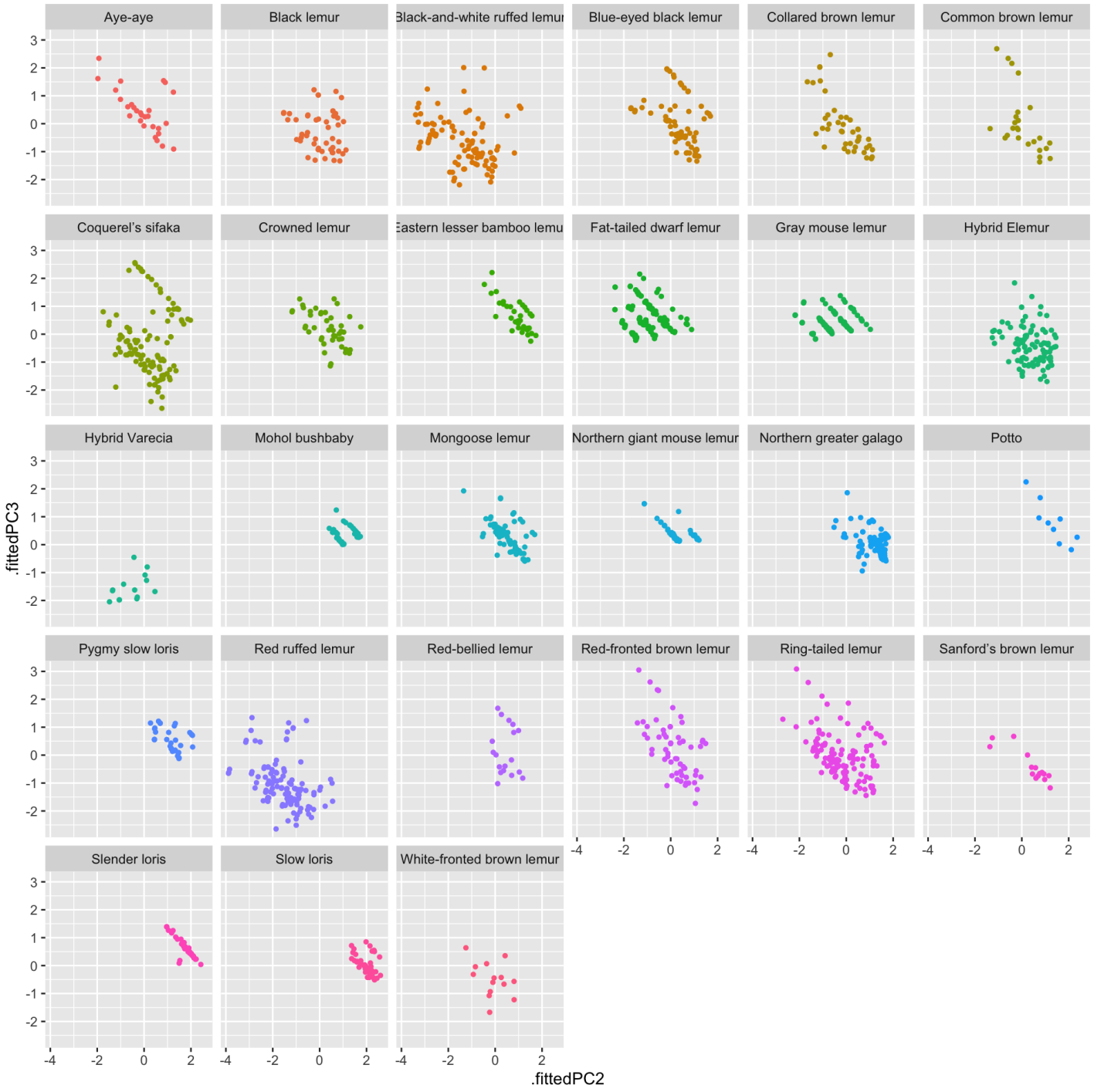
```
# getting the latest weight date rows per dlc id
lemurs_latest <- lemurs %>%
  group_by(dlc_id) %>%
  mutate(taxon = recode(taxon, CMED = "Fat-tailed dwarf lemur", DMAD = "Aye-aye", EALB =
"White-fronted brown lemur", ECOL = "Collared brown lemur", ECOR = "Crowned lemur", EFLA
= "Blue-eyed black lemur", EFUL = "Common brown lemur", EMAC = "Black lemur", EMON = "Mo
ngoose lemur", ERUB = "Red-bellied lemur", ERUF = "Red-fronted brown lemur", ESAN = "San
ford's brown lemur", EUL = "Hybrid Elemur", GMOH = "Mohol bushbaby", HGG = "Eastern less
er bamboo lemur", LCAT = "Ring-tailed lemur", LTAR = "Slender loris", MMUR = "Gray mouse
lemur", MZAZ = "Northern giant mouse lemur", NCOU = "Slow loris", NPYG = "Pygmy slow lor
is", OGG = "Northern greater galago", PCOQ = "Coquerel's sifaka", PPOT = "Potto", VAR =
"Hybrid Varecia", VRUB = "Red ruffed lemur", VVV = "Black-and-white ruffed lemur")) %>%
#more useful names from the github page
  slice(which.max(as.Date(weight_date, '%m/%d/%Y')))

# ungroup acting weird, had to make a temp variable.
temp <- lemurs_latest %>%
  ungroup() %>%
  select(dlc_id, taxon, expected_gestation, litter_size, sire_age_at_concep_y, dam_age_a
t_concep_y, weight_g) %>%
  na.omit #can't have missing values in PCA, only lost 680 lemurs out of 2270 or 30%.


pca_fit <- temp %>%
  select(expected_gestation, litter_size, sire_age_at_concep_y, dam_age_at_concep_y, wei
ght_g) %>%
  scale() %>%
  prcomp()


pca_fit %>%
  # add PCs to the original dataset
  augment(temp) %>%
  ggplot(aes(.fittedPC1, .fittedPC2)) +
  geom_point(size=1, aes(color=taxon)) +
  facet_wrap(~taxon) +
  guides(colour=FALSE)
```

```
## Warning: `guides(<scale> = FALSE)` is deprecated. Please use `guides(<scale> =
## "none")` instead.
```

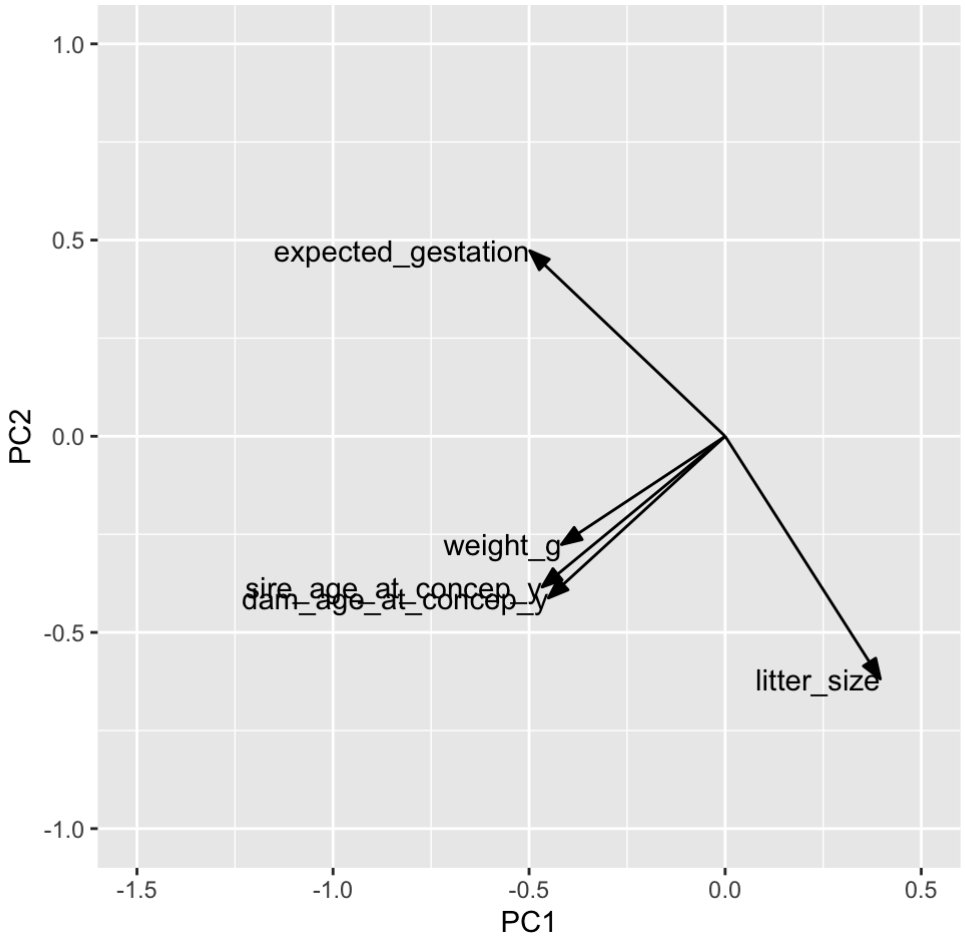



```
pca_fit %>%
  # add PCs to the original dataset
  augment(temp) %>%
  ggplot(aes(.fittedPC2, .fittedPC3)) +
  geom_point(size=1, aes(color=taxon)) +
  facet_wrap(~taxon) +
  guides(colour=FALSE)
```

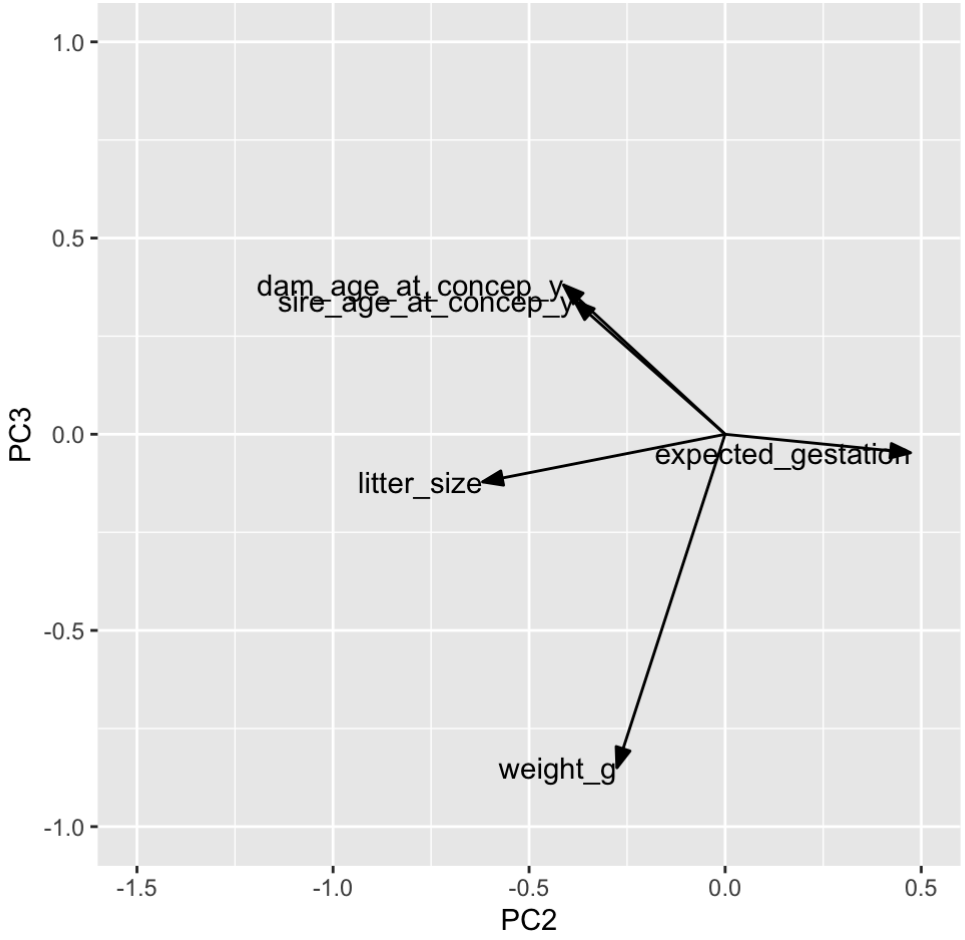
```
## Warning: `guides(<scale> = FALSE)` is deprecated. Please use `guides(<scale> =
## "none")` instead.
```


Aye-aye
Black lemur
Black-and-white ruffed lemur
Blue-eyed black lemur
Collared brown lemur
Common brown lemur
Coquerel's sifaka
Crowned lemur
Eastern lesser bamboo lemur
Fat-tailed dwarf lemur
Gray mouse lemur
Hybrid Elemur
Hybrid Varecia
Mohol bushbaby
Mongoose lemur
Northern giant mouse lemur
Northern greater galago
Potto
Pygmy slow loris
Red ruffed lemur
Red-bellied lemur
Red-fronted brown lemur
Ring-tailed lemur
Sanford's brown lemur
Slender loris
Slow loris
White-fronted brown lemur
.fittedPC3
.fittedPC2

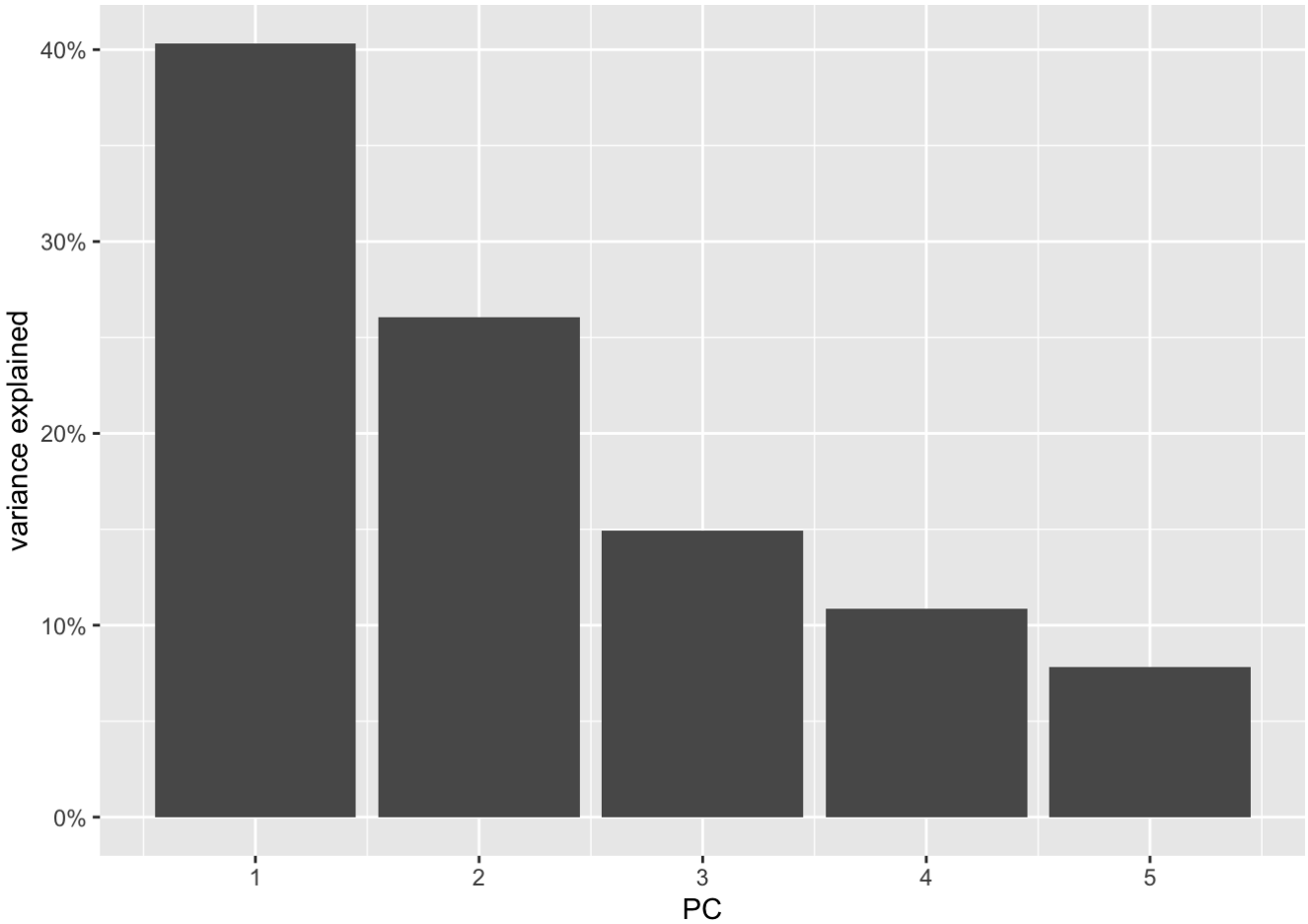
```
arrow_style <- arrow(
  angle = 20, length = grid::unit(8, "pt"),
  ends = "first", type = "closed"
)
pca_fit %>%
  # extract rotation matrix
  tidy(matrix = "rotation") %>%
  pivot_wider(
    names_from = "PC", values_from = "value",
    names_prefix = "PC"
  ) %>%
  ggplot(aes(PC1, PC2)) +
  geom_segment(
    xend = 0, yend = 0,
    arrow = arrow_style
  ) +
  geom_text(aes(label = column), hjust = 1) +
  xlim(-1.5, 0.5) + ylim(-1, 1) +
  coord_fixed()
```



```
pca_fit %>%
  # extract rotation matrix
  tidy(matrix = "rotation") %>%
  pivot_wider(
    names_from = "PC", values_from = "value",
    names_prefix = "PC"
  ) %>%
  ggplot(aes(PC2, PC3)) +
  geom_segment(
    xend = 0, yend = 0,
    arrow = arrow_style
  ) +
  geom_text(aes(label = column), hjust = 1) +
  xlim(-1.5, 0.5) + ylim(-1, 1) +
  coord_fixed()
```



```
pca_fit %>%
  # extract eigenvalues
  tidy(matrix = "eigenvalues") %>%
  ggplot(aes(PC, percent)) +
  geom_col() +
  scale_x_continuous(
    # create one axis tick per PC
    breaks = 1:6
  ) +
  scale_y_continuous(
    name = "variance explained",
    # format y axis ticks as percent values
    label = scales::label_percent(accuracy = 1)
  )
```



Discussion: *Discussion of the plots in reverse order will be helpful. Let's start with the explained variance plot. Principal components 1, 2, and 3 cover around 80% of the explained variance in the dataset. So it's best to leave the dicussion to these two PCs. All variables except for litter size contribute negatively to PC1, showing that higher litter size is an indicator of a lower gestation period, weight, and age of parents. PC2 shows how litter size has an effect on gestation periods, indicating the same relationship found in PC1. PC3 shows another relationship between the age of parents and the weight. Typically, the older the parents, the leaner the lemur.*

Let's move onto answering the question, given our interpretation of the principal components. When looking at PC1 plotted against PC2, you immediately notice "streaks" of observations that lie in the direction of the parents' ages and weight of the lemur. The groups (or "streaks") represent lemurs that have similarly sized litters (and

therefore gestation periods), spread along their weights and ages of their parents. Coquerel's sifaka lemurs tend to stick to one litter size, while Black-and-white ruffed lemurs have varying litter sizes. Weight will pull the points towards the third quadrant, so Northern giant mouse lemurs tend to be heavier than the Red ruffed lemur, for example. (It's slightly humorous that the "giant mouse" lemur is actually quite small compared to the average lemur.

Moving onto PC2 vs PC3, agian we see the "streaking" litter sizes effect, but we can observe some differences in parental age (age of fertility) for different lemurs. Older parents will pull the points towards the second quadrant, so fat tailed dwarf lemurs tend to have babies at an older age compared to the rest. We can also see which lemurs are fat by looking at points pulled towards the third quadrant, Coquerel's sifaka tend to be heavier compared to the other lemurs.