Homework 7

Weifan Liu

2026-01-28

Instructions

Answer the following questions and/or complete the exercises in RMarkdown. Please embed all of your code and push the final work to your repository. Your report should be organized, clean, and run free from errors. Remember, you must remove the `#` for any included code chunks to run.

Load the tidyverse

```
library("tidyverse")
library("janitor")
```

Data

For this assignment, we will use data from a study on elephants and the effects of poaching on tusk size.

Reference: Chiyo, Patrick I., Vincent Obanda, and David K. Korir. "Illegal tusk harvest and the decline of tusk size in the African elephant." Ecology and Evolution 5, 22: 5216–5229 (2015) (https://doi.org/10.1002/ece3.1769). Data deposited at Dryad Digital Repository (https://doi.org/10.5061/dryad.h6t7j).

1. Before starting data analysis, read the abstract of the paper to get an idea of the questions being asked. In 2-3 sentences, describe what the study is testing and the variables involved.

```
## The article mainly discussed the topic that whether human's harviesting toward ele
phant because of tusk play an important role in reducton of both the length and circu
mference of the tusk
```

2. Load `elephants.csv` and store it as a new object called `elephants`.

```
elephants <- read_csv("data/elephants.csv")
```

```
## Rows: 777 Columns: 7
## ── Column specification ────────────────────────────────────────────────────────
## Delimiter: ","
## chr (3): Years of sample collection, Elephant ID, Sex
## dbl (4): Estimated Age (years), shoulder Height in  cm, Tusk Length in cm, T...
##
## ℹ Use `spec()` to retrieve the full column specification for this data.
## ℹ Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

3. Clean the data by converting variable names to lowercase with no spaces or special characters.

```
str(elephants)
```

```
## spc_tbl_ [777 × 7] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ Years of sample collection: chr [1:777] "1966-68" "1966-68" "1966-68" "1966-68"
...
##  $ Elephant ID              : chr [1:777] "12" "34" "162" "292" ...
##  $ Sex                      : chr [1:777] "f" "f" "f" "f" ...
##  $ Estimated Age (years)    : num [1:777] 0.08 0.08 0.083 0.083 0.25 0.25 0.25 0.
5 0.5 1 ...
##  $ shoulder Height in  cm   : num [1:777] 102 89 89 92 133 100 93 108 108 124 ...
##  $ Tusk Length in cm        : num [1:777] NA NA NA NA NA NA NA NA NA NA ...
##  $ Tusk Circumference   in cm: num [1:777] NA NA NA NA NA NA NA NA NA NA ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   `Years of sample collection` = col_character(),
##   ..   `Elephant ID` = col_character(),
##   ..   Sex = col_character(),
##   ..   `Estimated Age (years)` = col_double(),
##   ..   `shoulder Height in  cm` = col_double(),
##   ..   `Tusk Length in cm` = col_double(),
##   ..   `Tusk Circumference   in cm` = col_double()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

```
elephants <- elephants %>%
  clean_names()
```

4. Use one or more of the summary functions you have learned to get an idea of the structure of the data.

```
summary(elephants)
```

```
##  years_of_sample_collection elephant_id           sex
##  Length:777                 Length:777         Length:777
##  Class :character           Class :character   Class :character
##  Mode  :character           Mode  :character   Mode  :character
##
##
##
##
##  estimated_age_years shoulder_height_in_cm tusk_length_in_cm
##  Min.   : 0.08       Min.   : 89.0         Min.   : 22.50
##  1st Qu.: 4.50       1st Qu.:177.0         1st Qu.: 60.95
##  Median :12.00       Median :220.0         Median : 88.00
##  Mean   :15.05       Mean   :210.2         Mean   : 91.59
##  3rd Qu.:23.00       3rd Qu.:244.0         3rd Qu.:116.75
##  Max.   :55.00       Max.   :340.0         Max.   :234.00
##  NA's   :3           NA's   :1             NA's   :180
##  tusk_circumference_in_cm
##  Min.   : 8.00
##  1st Qu.:16.00
##  Median :20.00
##  Mean   :21.01
##  3rd Qu.:25.00
##  Max.   :48.00
##  NA's   :163
```

```
str(elephants)
```

```
## spc_tbl_ [777 × 7] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ years_of_sample_collection: chr [1:777] "1966-68" "1966-68" "1966-68" "1966-68"
...
##  $ elephant_id               : chr [1:777] "12" "34" "162" "292" ...
##  $ sex                       : chr [1:777] "f" "f" "f" "f" ...
##  $ estimated_age_years       : num [1:777] 0.08 0.08 0.083 0.083 0.25 0.25 0.25 0.
5 0.5 1 ...
##  $ shoulder_height_in_cm     : num [1:777] 102 89 89 92 133 100 93 108 108 124 ...
##  $ tusk_length_in_cm         : num [1:777] NA NA NA NA NA NA NA NA NA NA ...
##  $ tusk_circumference_in_cm  : num [1:777] NA NA NA NA NA NA NA NA NA NA ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   `Years of sample collection` = col_character(),
##   ..   `Elephant ID` = col_character(),
##   ..   Sex = col_character(),
##   ..   `Estimated Age (years)` = col_double(),
##   ..   `shoulder Height in  cm` = col_double(),
##   ..   `Tusk Length in cm` = col_double(),
##   ..   `Tusk Circumference   in cm` = col_double()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

```
str(elephants)
```

```
## spc_tbl_ [777 × 7] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ years_of_sample_collection: chr [1:777] "1966-68" "1966-68" "1966-68" "1966-68"
...
##  $ elephant_id              : chr [1:777] "12" "34" "162" "292" ...
##  $ sex                      : chr [1:777] "f" "f" "f" "f" ...
##  $ estimated_age_years      : num [1:777] 0.08 0.08 0.083 0.083 0.25 0.25 0.25 0.
5 0.5 1 ...
##  $ shoulder_height_in_cm    : num [1:777] 102 89 89 92 133 100 93 108 108 124 ...
##  $ tusk_length_in_cm        : num [1:777] NA NA NA NA NA NA NA NA NA NA ...
##  $ tusk_circumference_in_cm : num [1:777] NA NA NA NA NA NA NA NA NA NA ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   `Years of sample collection` = col_character(),
##   ..   `Elephant ID` = col_character(),
##   ..   Sex = col_character(),
##   ..   `Estimated Age (years)` = col_double(),
##   ..   `shoulder Height in  cm` = col_double(),
##   ..   `Tusk Length in cm` = col_double(),
##   ..   `Tusk Circumference   in cm` = col_double()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

5. Use `mutate()` Change the variables `years_of_sample_collection` , `elephant_id` , and `sex` to factors. Be sure to store the output as a new dataframe and use it for the remaining questions.

```
elephants_factor <-
  elephants %>%
  mutate(across(c(years_of_sample_collection,elephant_id,sex), as.factor))
```

6. From which years were data collected? Show the sample periods below.

```
names(elephants_factor)
```

```
## [1] "years_of_sample_collection" "elephant_id"
## [3] "sex"                        "estimated_age_years"
## [5] "shoulder_height_in_cm"      "tusk_length_in_cm"
## [7] "tusk_circumference_in_cm"
```

```
elephants_factor %>%
  distinct(years_of_sample_collection)
```

```
## # A tibble: 2 × 1
##   years_of_sample_collection
##   <fct>
## 1 1966-68
## 2 2005-13
```

7. How many males and females were sampled in this study?

```
elephants_factor %>%
  filter(sex=="f") %>%
  count()
```

```
## # A tibble: 1 × 1
##       n
##   <int>
## 1   416
```

```
elephants_factor %>%
  filter(sex=="m") %>%
  count()
```

```
## # A tibble: 1 × 1
##       n
##   <int>
## 1   361
```

8. What is the mean, median, and standard deviation for age of males and females included in the study? Separate the results by year of sample collection. Does the sampling look even between years and sexes?

```
names(elephants_factor)
```

```
## [1] "years_of_sample_collection" "elephant_id"
## [3] "sex"                        "estimated_age_years"
## [5] "shoulder_height_in_cm"      "tusk_length_in_cm"
## [7] "tusk_circumference_in_cm"
```

```
elephants_factor %>%
  filter(sex=="f") %>%
  summarize(mean_estimated_age_years=mean(estimated_age_years,na.rm=T),
            median_estimated_age_years=median(estimated_age_years,na.rm=T),
            sd_estimated_age_years=sd(estimated_age_years,na.rm=T))
```

```
## # A tibble: 1 × 3
##   mean_estimated_age_years median_estimated_age_years sd_estimated_age_years
##                      <dbl>                       <dbl>                  <dbl>
## 1                     17.6                          16                   13.0
```

```
elephants_factor %>%
  filter(sex=="m") %>%
  summarize(mean_estimated_age_years=mean(estimated_age_years,na.rm=T),
            median_estimated_age_years=median(estimated_age_years,na.rm=T),
            sd_estimated_age_years=sd(estimated_age_years,na.rm=T))
```

```
## # A tibble: 1 × 3
##   mean_estimated_age_years median_estimated_age_years sd_estimated_age_years
##                      <dbl>                       <dbl>                  <dbl>
## 1                     12.1                         8.5                   10.6
```

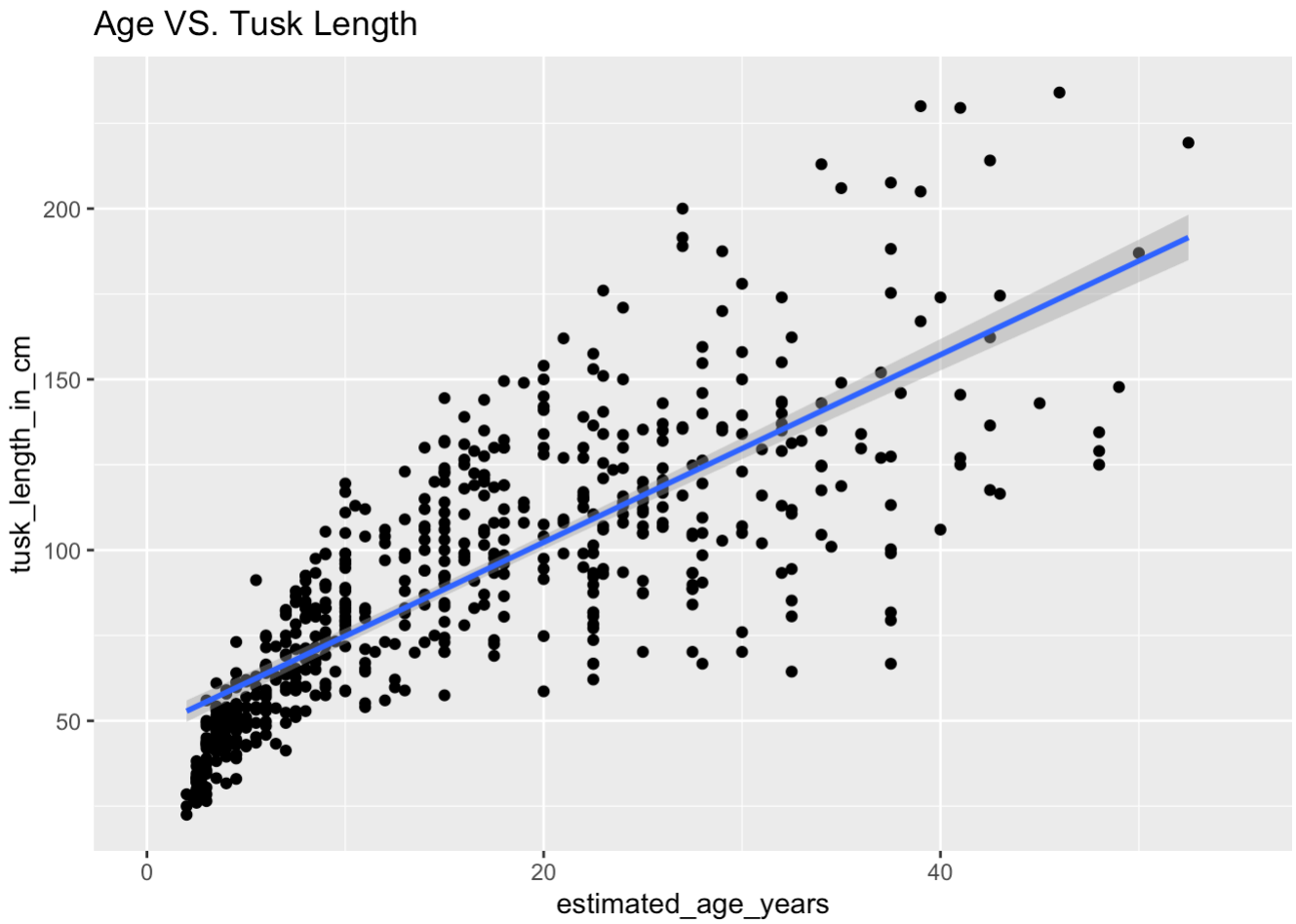
9. Is age (independent variable) a positive predictor of tusk length (dependent variable)? Create a plot that shows the relationship between these variables and add a linear model fit line.

```
names(elephants_factor)
```

```
## [1] "years_of_sample_collection" "elephant_id"
## [3] "sex"                        "estimated_age_years"
## [5] "shoulder_height_in_cm"      "tusk_length_in_cm"
## [7] "tusk_circumference_in_cm"
```

```
elephants_factor %>%
  ggplot(mapping=aes(x=estimated_age_years,y=tusk_length_in_cm))+
  geom_point(na.rm=T)+
  geom_smooth(method="lm",na.rm=T)+
  labs(title="Age VS. Tusk Length")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



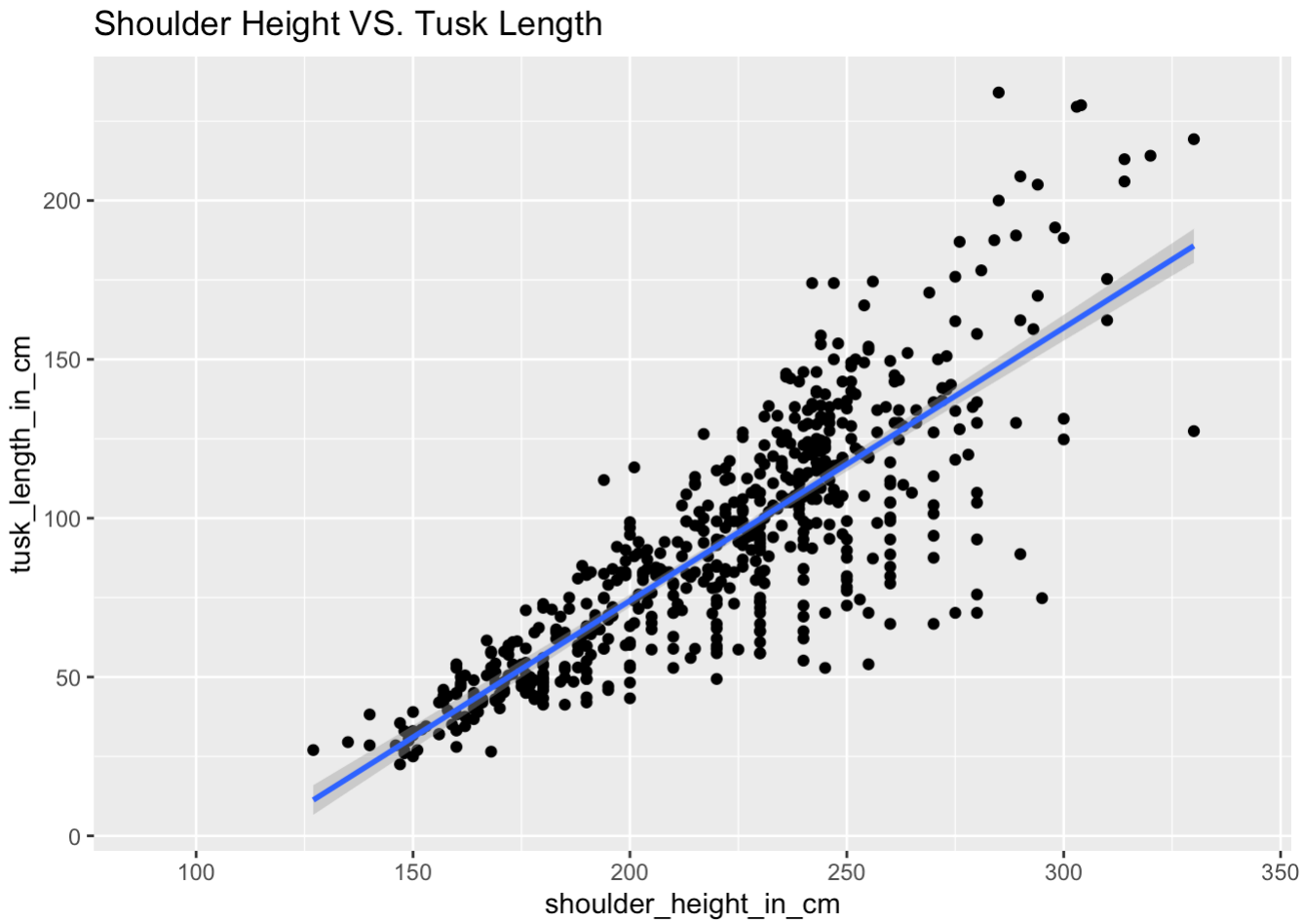
10. Is shoulder height (independent variable) a positive predictor of tusk length (dependent variable)? Create a plot that shows the relationship between these variables and add a linear model fit line.

```
names(elephants_factor)
```

```
## [1] "years_of_sample_collection" "elephant_id"
## [3] "sex"                        "estimated_age_years"
## [5] "shoulder_height_in_cm"      "tusk_length_in_cm"
## [7] "tusk_circumference_in_cm"
```

```
elephants_factor %>%
  ggplot(mapping=aes(x=shoulder_height_in_cm,y=tusk_length_in_cm))+
  geom_point(na.rm=T)+
  geom_smooth(method="lm",na.rm=T)+
  labs(title="Shoulder Height VS. Tusk Length")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



11. The authors argue that because poachers preferentially target elephants with large tusks, this has resulted in a decrease in average tusk length. Is this supported by the data? Show your code and calculations below.

```
names(elephants_factor)
```

```
## [1] "years_of_sample_collection" "elephant_id"
## [3] "sex"                        "estimated_age_years"
## [5] "shoulder_height_in_cm"      "tusk_length_in_cm"
## [7] "tusk_circumference_in_cm"
```

```
elephants_factor %>%
  distinct(years_of_sample_collection)
```

```
## # A tibble: 2 × 1
##   years_of_sample_collection
##   <fct>
## 1 1966-68
## 2 2005-13
```

```
elephants_factor %>%
  filter(years_of_sample_collection=="1966-68") %>%
  summarize(mean_tusk_length_in_cm=mean(tusk_length_in_cm,na.rm=T))
```

```
## # A tibble: 1 × 1
##   mean_tusk_length_in_cm
##                    <dbl>
## 1                   96.9
```

```
elephants_factor %>%
  filter(years_of_sample_collection=="2005-13") %>%
  summarize(mean_tusk_length_in_cm=mean(tusk_length_in_cm,na.rm=T))
```

```
## # A tibble: 1 × 1
##   mean_tusk_length_in_cm
##                    <dbl>
## 1                   77.2
```

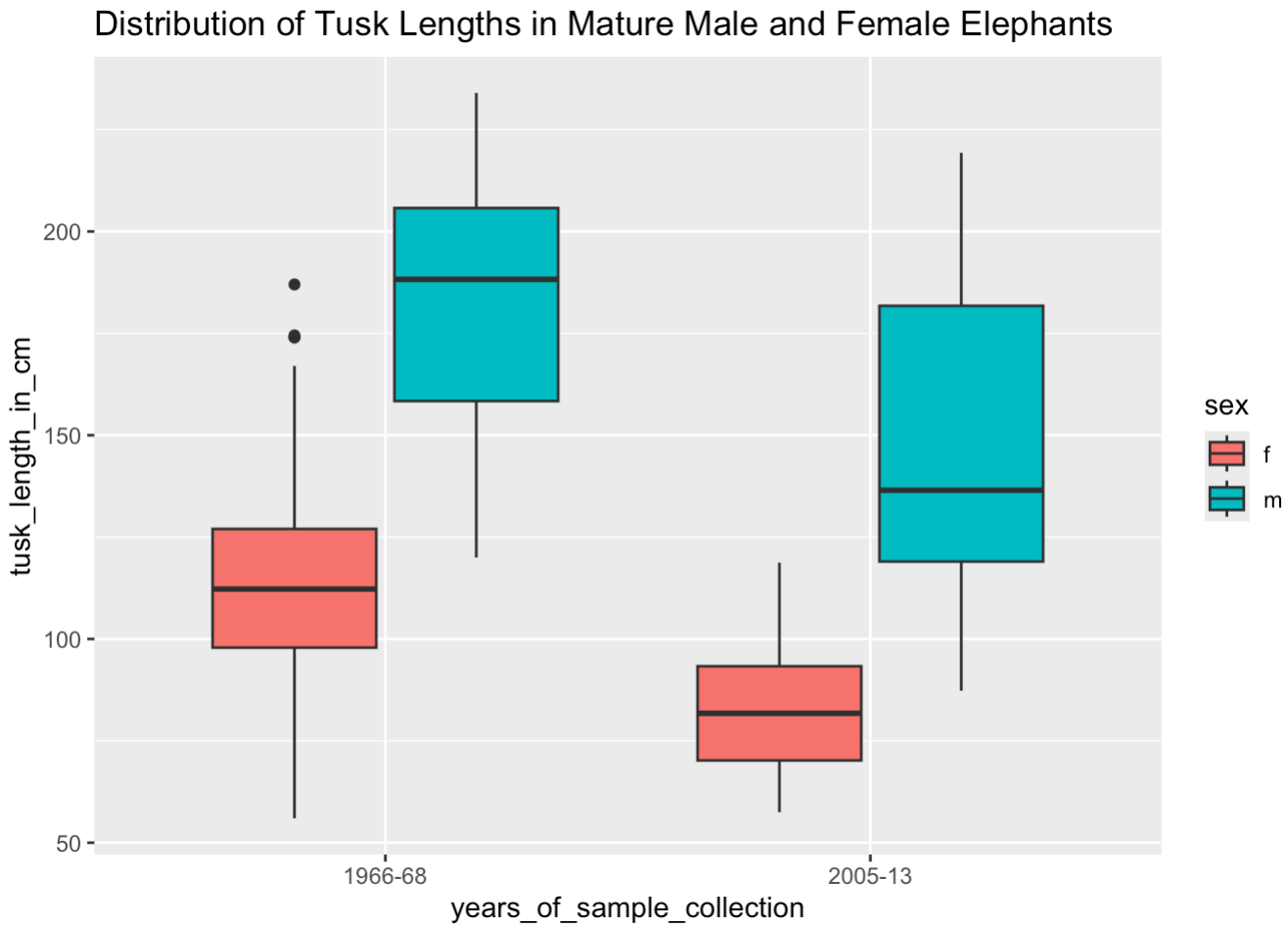
12. Male elephants reach effective sexual maturity at 25 years while females are sexually mature at 12 years. Make a new dataframe that extracts only the males and females at sexual maturity. Then, make a plot that shows the range of tusk length between the two sample periods for these mature elephants.

```
names(elephants_factor)
```

```
## [1] "years_of_sample_collection" "elephant_id"
## [3] "sex"                        "estimated_age_years"
## [5] "shoulder_height_in_cm"      "tusk_length_in_cm"
## [7] "tusk_circumference_in_cm"
```

```
elephants_factor_new <-
  elephants_factor %>%
  filter((sex=="f"&estimated_age_years>=12)|(sex=="m"&estimated_age_years>=25))
```

```
elephants_factor_new %>%
  select(sex,tusk_length_in_cm,years_of_sample_collection) %>%
ggplot(mapping=aes(x=years_of_sample_collection,y=tusk_length_in_cm,fill=sex))+
  geom_boxplot(na.rm=T)+
  labs(title="Distribution of Tusk Lengths in Mature Male and Female Elephants")
```



Submit the Homework

1. Save your work and knit the .rmd file.
2. Open the .html file and “print” it to a .pdf file in Google Chrome (not Safari).
3. Go to the class Canvas page and open Gradescope.
4. Submit your .pdf file to the homework assignment- be sure to assign the pages to the correct questions.
5. Commit and push your work to your repository.