Homework 7

Chengyi Wang

2026-01-29

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 study is about wild elephants population in Africa. The researchers measured elephants' tusk data to test if the illegal ivory harvesting is causing decline in tusk size among African elephants.

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.

```
elephants <- clean_names(elephants)
```

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

```
glimpse(elephants)
```

```
## Rows: 777
## Columns: 7
## $ years_of_sample_collection <chr> "1966-68", "1966-68", "1966-68", "1966-68",…
## $ elephant_id                <chr> "12", "34", "162", "292", "11", "152", "264…
## $ sex                        <chr> "f", "f", "f", "f", "f", "f", "f", "f", "f"…
## $ estimated_age_years        <dbl> 0.080, 0.080, 0.083, 0.083, 0.250, 0.250, 0…
## $ shoulder_height_in_cm      <dbl> 102, 89, 89, 92, 133, 100, 93, 108, 108, 12…
## $ tusk_length_in_cm          <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,…
## $ tusk_circumference_in_cm   <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,…
```

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 %>%
  mutate(across(c("years_of_sample_collection","elephant_id","sex"),as.factor))
```

```
## # A tibble: 777 × 7
##    years_of_sample_collection elephant_id sex   estimated_age_years
##    <fct>                      <fct>       <fct>               <dbl>
##  1 1966-68                    12          f                   0.08
##  2 1966-68                    34          f                   0.08
##  3 1966-68                    162         f                   0.083
##  4 1966-68                    292         f                   0.083
##  5 1966-68                    11          f                   0.25
##  6 1966-68                    152         f                   0.25
##  7 1966-68                    264         f                   0.25
##  8 1966-68                    263         f                   0.5
##  9 1966-68                    266         f                   0.5
## 10 1966-68                    217         f                   1
## # ℹ 767 more rows
## # ℹ 3 more variables: shoulder_height_in_cm <dbl>, tusk_length_in_cm <dbl>,
## #   tusk_circumference_in_cm <dbl>
```

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

```
elephants %>%
  select(years_of_sample_collection) %>%
  distinct()
```

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

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

```
elephants %>%
  select(sex) %>%
  count(sex)
```

```
## # A tibble: 2 × 2
##   sex       n
##   <chr> <int>
## 1 f       416
## 2 m       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?

```
elephants %>%
  group_by(sex,years_of_sample_collection) %>%
  summarise(mean_age=mean(estimated_age_years,na.rm=T),min_age=min(estimated_age_years,na.rm=T),ma
x_age=max(estimated_age_years,na.rm=T))
```

```
## `summarise()` has grouped output by 'sex'. You can override using the `.groups`
## argument.
```

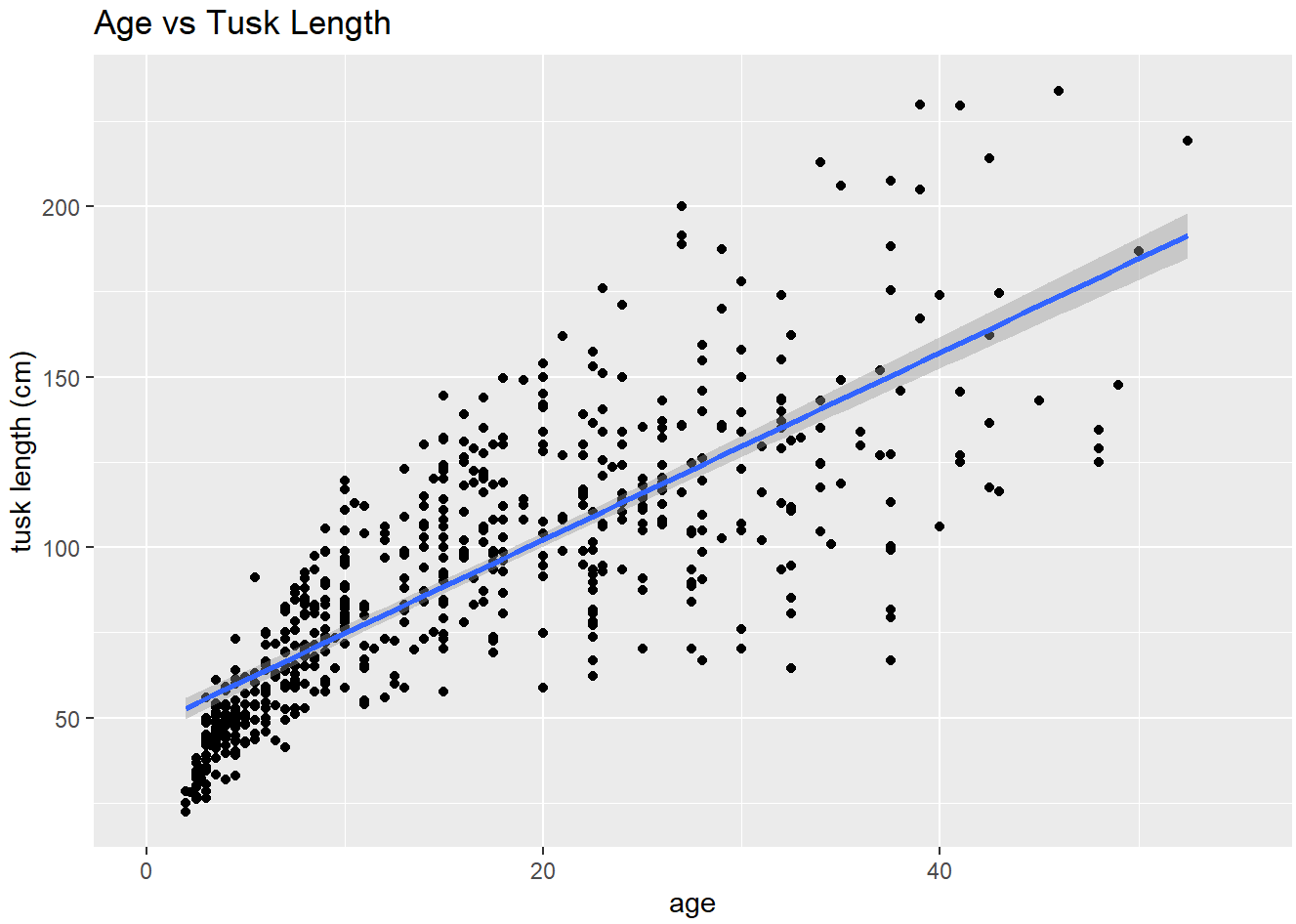
```
## # A tibble: 4 × 5
## # Groups:   sex [2]
##   sex   years_of_sample_collection mean_age min_age max_age
##   <chr> <chr>                         <dbl>   <dbl>   <dbl>
## 1 f     1966-68                        17.6    0.08    55
## 2 f     2005-13                        17.9    2.5     42.5
## 3 m     1966-68                        10.8    0.08    46
## 4 m     2005-13                        16.7    2.5     52.5
```

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.

```
elephants %>%
  ggplot(mapping=aes(x=estimated_age_years,y=tusk_length_in_cm))+
  geom_point()+
  geom_smooth(method = lm, na.rm = T)+
  labs(x="age",y="tusk length (cm)",title = "Age vs Tusk Length")
```

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

```
## Warning: Removed 182 rows containing missing values or values outside the scale range
## (`geom_point()`).
```

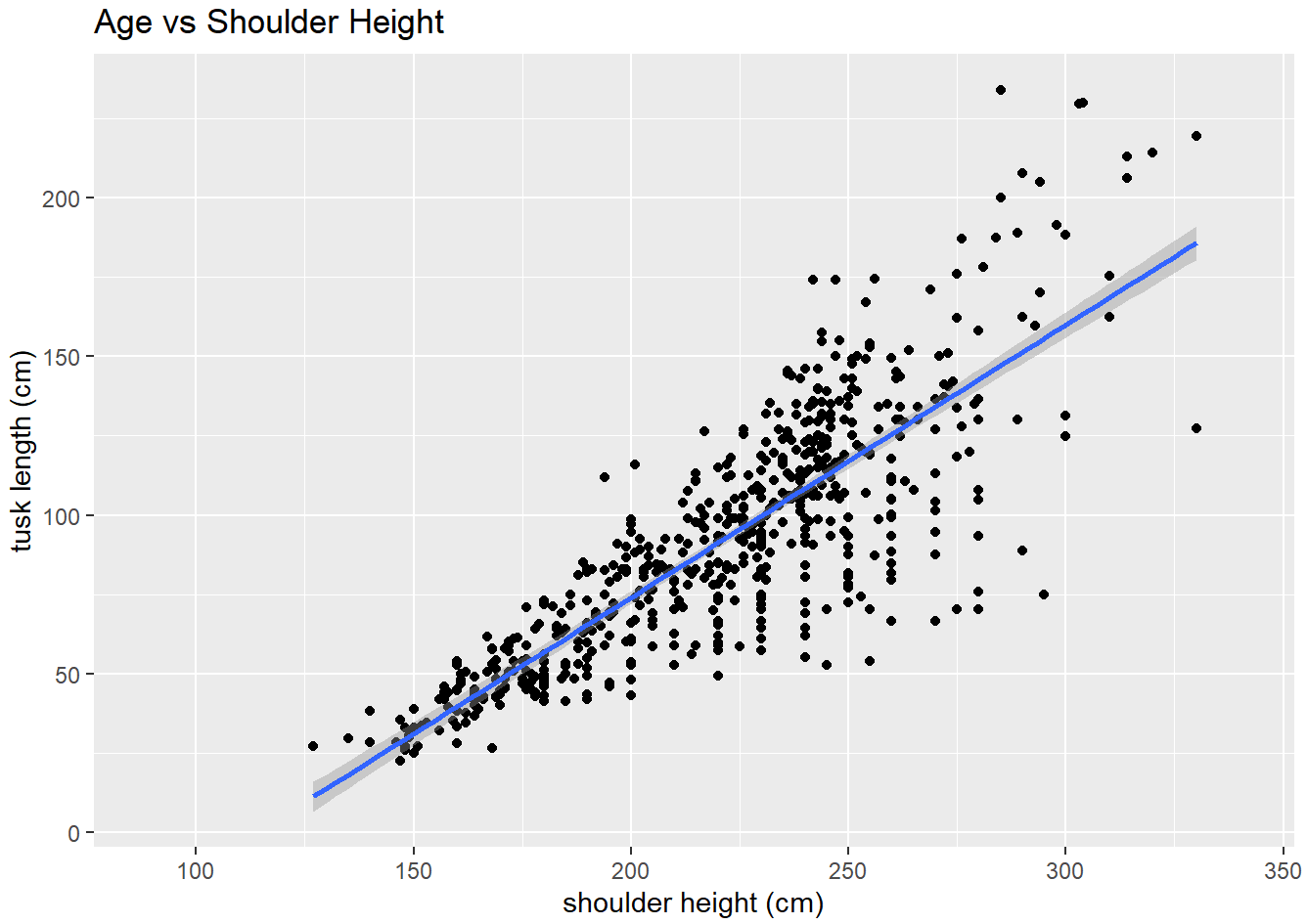


10. Is shoulder height (independent variable) a positive predictor of tusk length (dependent variable)? Create a plot that shows the relationship between these variables and add a linear model fit line.

```
elephants %>%
  ggplot(mapping=aes(x=shoulder_height_in_cm,y=tusk_length_in_cm))+
  geom_point()+
  geom_smooth(method = lm, na.rm = T)+
  labs(x="shoulder height (cm)",y="tusk length (cm)",title = "Age vs Shoulder Height")
```

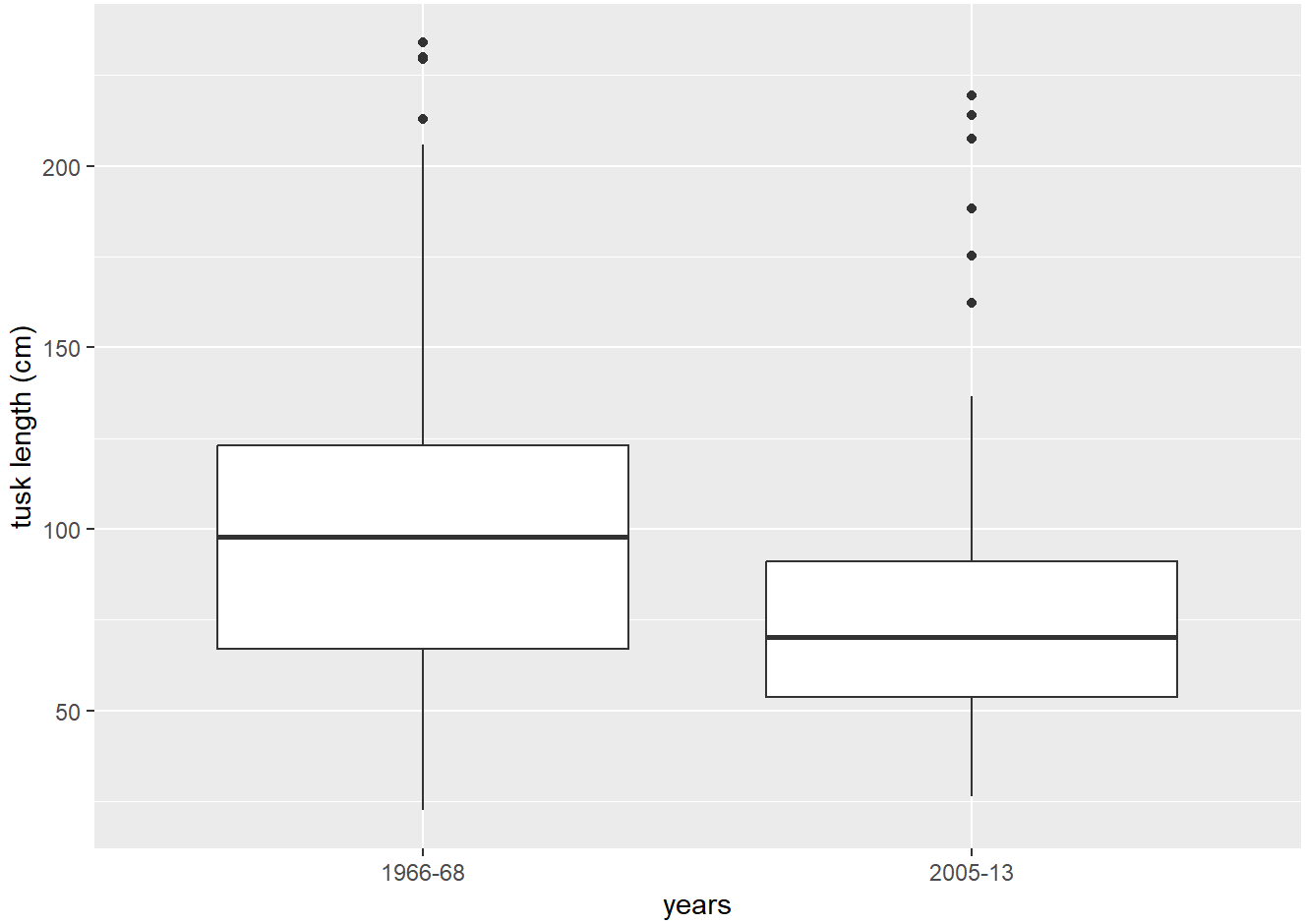
```
## `geom_smooth()` using formula = 'y ~ x'
```

```
## Warning: Removed 181 rows containing missing values or values outside the scale range
## (`geom_point()`).
```



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.

```
elephants %>%
  ggplot(mapping = aes(x=years_of_sample_collection,y=tusk_length_in_cm))+
  geom_boxplot(na.rm = T)+
  labs(x="years",y="tusk length (cm)")
```




```
elephants %>%
  group_by(years_of_sample_collection) %>%
  summarise(mean_tusk=mean(tusk_length_in_cm,na.rm=T),sd_tusk=sd(tusk_length_in_cm,na.rm=T))
```

```
## # A tibble: 2 × 3
##   years_of_sample_collection mean_tusk sd_tusk
##   <chr>                          <dbl>   <dbl>
## 1 1966-68                         96.9    39.1
## 2 2005-13                         77.2    33.9
```

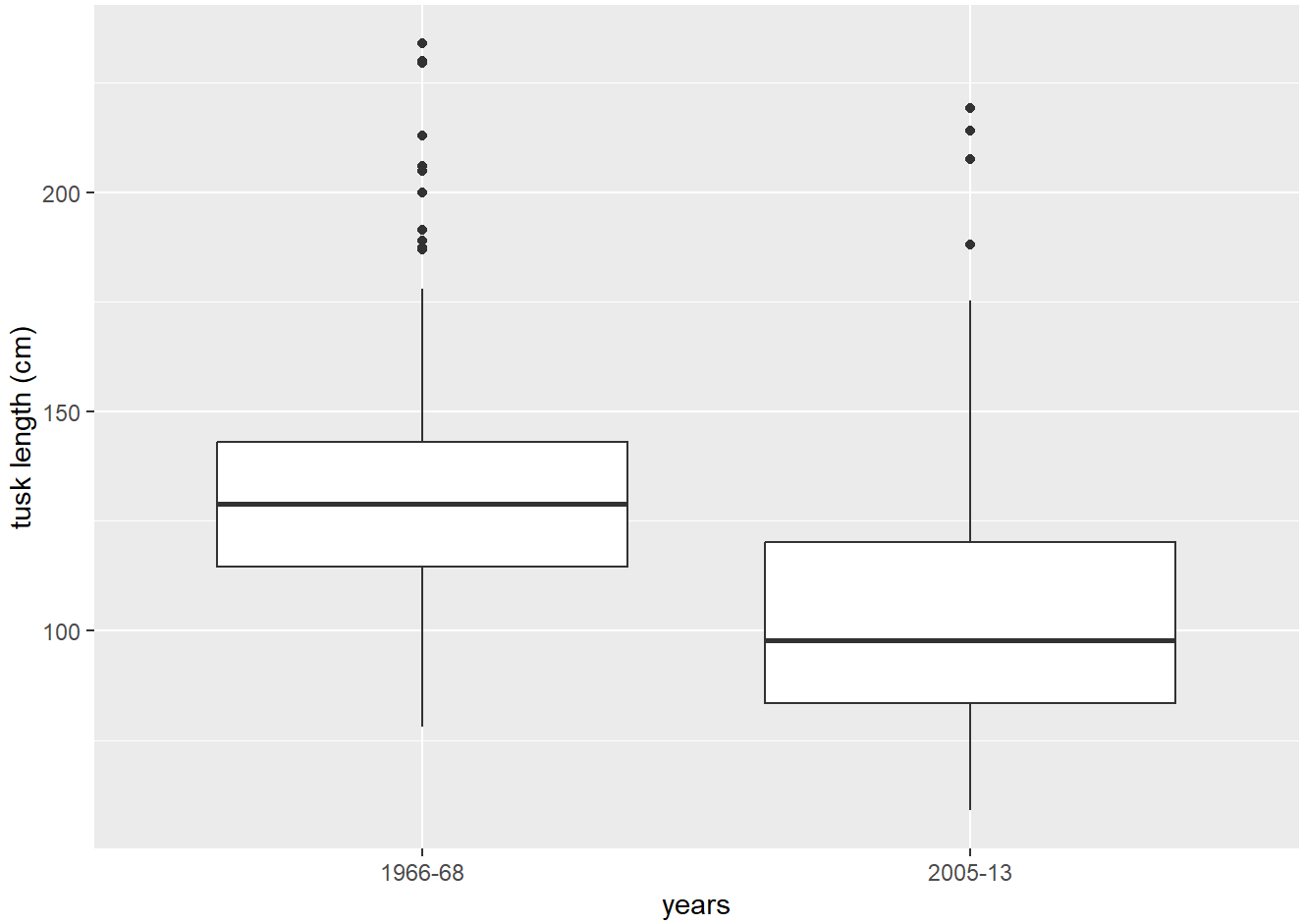
```
test_result <- t.test(tusk_length_in_cm ~ years_of_sample_collection, data = elephants)
test_result$p.value
```

```
## [1] 4.40279e-09
```

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.

```
elephants_new <- elephants %>%
  filter(sex=="m"|estimated_age_years>=25,sex=="f"|estimated_age_years>=12)
```

```
elephants_new %>%
  ggplot(mapping = aes(x=years_of_sample_collection,y=tusk_length_in_cm))+
  geom_boxplot(na.rm = T)+
  labs(x="years",y="tusk length (cm)")
```



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.