Homework 1: Basic Practice with R

Ruonan Zhao

August 29, 2024

Instructions:

The purpose of this homework assignment is to practice using R in RStudio. Make sure you read each question carefully. In each question, I will give you a task to do, and I will tell you want I want you to output. You can write as much code as you want in each code chunk, but make sure you complete the task and only print the output I asked you to print. Don't sort the data unless you are told to sort the data. You should remove the "#" sign in each code chunk before writing your code. **You should knit your RMD file to a PDF after you answer every question.**

For this assignment, we will be using data from the Box Turtle Connection. The Box Turtle Connection is a long-term study anticipating at least 100 years of data collection on box turtles. Their purpose is to learn more about the status and trends in box turtle populations, identify threats, and develop strategies for long-term conservation of the species. Eastern Box Turtle populations are in decline in North Carolina and while they are recognized as a threatened species by the International Union for Conservation of Nature, the turtles have no protection in North Carolina. There are currently more than 30 active research study sites across the state of North Carolina. Turtles are weighed, measured, photographed, and permanently marked. These data, along with voucher photos (photos that document sightings), are then entered into centralized database managed by the NC Wildlife Resources Commission. The *Turtles.csv* dataset contains data collected at The Piedmont Wildlife Center in Durham.

Before answering question 1, remember to run the *setup* code chunk to load all libraries. All of the libraries you load will need to be first installed on your computer.

After you are done, knit the RMarkdown file to PDF and submit the PDF to Gradescope under Homework 1.

Questions

Q1 (2 Points)

Write code that imports *Turtles.csv* into R and save that data into a data frame named **Turtles**. Print the first 5 rows of **Turtles**.

```
Turtles = read.csv("Turtles.csv")
head(Turtles, 5)
```

```
##   LifeStage    Sex Annuli Mass StraightlineCL  MaxCW PL_AnteriortoHinge
## 1     Adult   Male     13  410         127.00 102.00              48.00
## 2     Adult   Male     19  340         113.62  93.96              44.87
## 3  Juvenile Female      7  160          89.49  73.51              39.60
```

```
## 4     Adult   Male     16  175         127.70 101.16                 54.76
## 5  Juvenile Female      7  100          81.00  69.00                 35.00
##   PL_HingetoPosterior ShellHeightatHinge
## 1               68.00              61.00
## 2               67.61              55.88
## 3               53.65              43.48
## 4               84.72              61.97
## 5               44.00              39.00
```

Q2 (2 Points)

Create a data frame named **Male_Turtles** that contains only the male turtles. Print the first 5 rows of **Male_Turtles**.

```
Male_Turtles = Turtles[Turtles$Sex  == "Male",]
head(Male_Turtles, 5)
```

```
##    LifeStage  Sex Annuli Mass StraightlineCL  MaxCW PL_AnteriortoHinge
## 1      Adult Male     13  410         127.00 102.00              48.00
## 2      Adult Male     19  340         113.62  93.96              44.87
## 4      Adult Male     16  175         127.70 101.16              54.76
## 9      Adult Male     18  325         115.00  94.00              45.00
## 10     Adult Male     40  475         137.00 105.00              52.00
##    PL_HingetoPosterior ShellHeightatHinge
## 1                68.00              61.00
## 2                67.61              55.88
## 4                84.72              61.97
## 9                68.00              55.00
## 10               79.00              63.00
```

Q3 (3 Points)

For the male turtles sample, find the three quartiles of the variable **ShellHeightatHinge**. A turtle's shell height at hinge is the measurement of the shell's height between the hinge of the plastron and the top of the carapace.

I don't care how much code you right, but I only should see three numbers in your output.

```
quantile(Male_Turtles$ShellHeightatHinge, probs = c(0.25, 0.5, 0.75) )
```

```
##     25%     50%     75% 
## 53.5975 56.0000 59.0000
```

Q4 (2 Points)

For the male turtles, calculate the ratio of the sample average **ShellHeightatHinge** for adults to the sample average **ShellHeightatHinge** for juveniles.

I don't care how much code you right, but I only should see one number in your output.

```r
adults_Male_Turtles = Male_Turtles[Male_Turtles$LifeStage == "Adult",]

juveniles_Male_Turtles = Male_Turtles[Male_Turtles$LifeStage == "Juvenile",]

ratio = mean(adults_Male_Turtles$ShellHeightatHinge, na.rm = TRUE) / mean(juveniles_Male_Turtles$ShellHe

ratio
```

```
## [1] 1.295792
```

Q5 (2 Points)

The **Annuli** rings on a turtle represent growth on the scutes of the carapace and plastron. In the past, it was thought that annuli corresponded to age, but findings suggest that this is not the case. However, the annuli are still counted since it may yield important life history information. Calculate the sample correlation between **Annuli** and **ShellHeightatHinge** for male turtles.

I don't care how much code you right, but I only should see one number in your output.

```r
cor(Male_Turtles$Annuli, Male_Turtles$ShellHeightatHinge,)
```

```
## [1] 0.2173155
```

Q6 (4 Points)

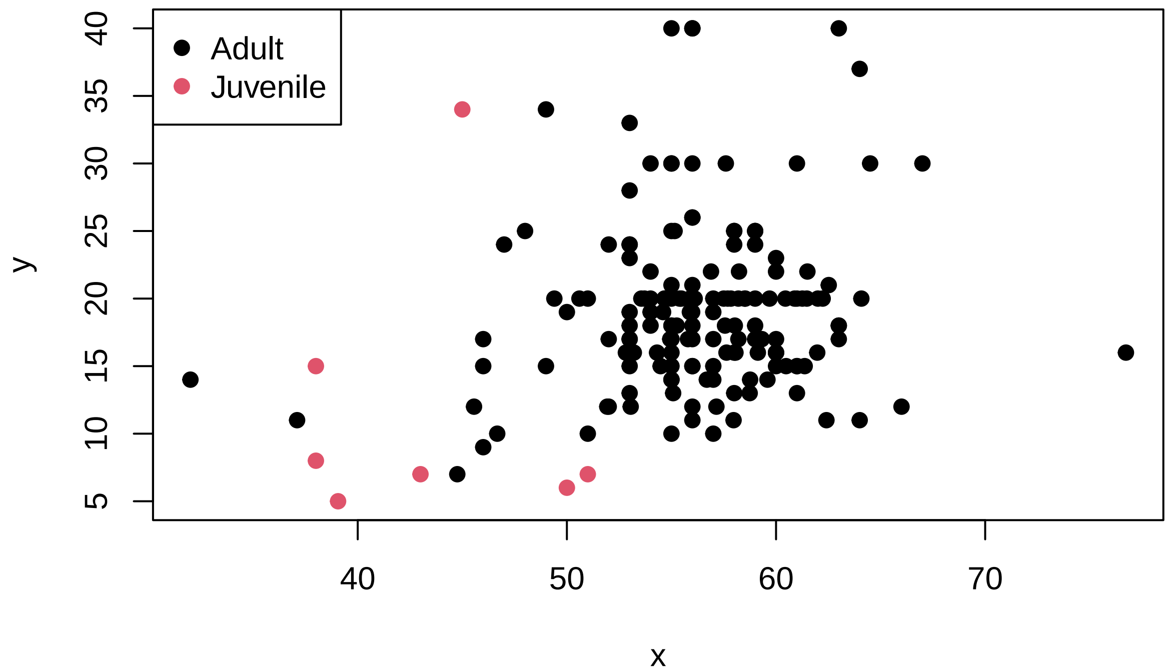
Go to the website https://r-charts.com/correlation/scatter-plot-group/ and observe the first visual that shows a scatter plot with different colors for the three groups. Modify that code in the first visual for the data in **Male_Turtles**. I want to see a scatter plot where **Annuli** is the response variable plotted on the y-axis and **ShellHeightatHinge** is the predictor variable plotted on the x-axis. I want to see a legend with two different colors so the audience can clearly see which dots are adult turtles and which dots are juvenile turtles. The legend should be placed in the top left of the plot.

You may need to run all the code on the website so you understand how the code works and what you need to do to get this to work for the turtles.

```r
# Scatter plot
x = Male_Turtles$ShellHeightatHinge
y = Male_Turtles$Annuli

plot(x, y,
     pch = 19,
     col = factor(Male_Turtles$LifeStage)
)

# Legend
legend("topleft",
       legend = levels(factor(Male_Turtles$LifeStage)),
       pch = 19,
       col = factor(levels(factor(Male_Turtles$LifeStage))))
```



Q7 (5 Points)

Perform a two-sample t-test (Welch's t-Test) for all the numeric variables in the **Male_Turtles** dataset. For each t-test, you are testing to see if there is evidence that the population mean for Adults males is different than the population mean for Juvenile males.

In your output, I want to see the output of all your t-tests. Then, I want you to write a response in complete sentences where you explain to your audience what you learned about the differences between adult and juvenile male turtles in the context of the data. Your audience should know where you found statistical significance, but more importantly, the audience should know what this means for the turtles. Be accurate in your discussion of the results. I would recommend reviewing how to interpret the results of a t-test. Since you are discussing 7 variables, you want to be concise so I am not expecting a sentence for every variable. Just make sure your explanation is precise and accurate.

Usually, I would require you to not use the variable names in your explanation since the audience wouldn't know what "PL_AnteriortoHinge" means; however, I didn't quickly fine one place where all these variables are explained so it is okay to reference the variables by their names in the dataset. I do ask you to format your variable names in your paragraph using a pair of single asterisks which *italicize* whatever is written in the middle. This makes the output in the PDF look more professional.

```
numeric_cols = sapply(Male_Turtles, is.numeric)
numeric_cols
```

```
##          LifeStage                 Sex             Annuli                Mass
##              FALSE               FALSE               TRUE                TRUE
##      StraightlineCL               MaxCW  PL_AnteriortoHinge PL_HingetoPosterior
```

```
##                  TRUE                 TRUE                 TRUE                 TRUE
##  ShellHeightatHinge
##                  TRUE
```

```r
t_annuli = t.test(Annuli~ Male_Turtles$LifeStage, data = Male_Turtles)
t_annuli
```

```
##
##  Welch Two Sample t-test
##
## data:  Annuli by Male_Turtles$LifeStage
## t = 1.8972, df = 6.1839, p-value = 0.1051
## alternative hypothesis: true difference in means between group Adult and group Juvenile is not equal
## 95 percent confidence interval:
##  -2.098724 17.062791
## sample estimates:
##    mean in group Adult mean in group Juvenile
##               19.19632               11.71429
```

```r
t_mass = t.test(Mass~ Male_Turtles$LifeStage, data = Male_Turtles)
t_mass
```

```
##
##  Welch Two Sample t-test
##
## data:  Mass by Male_Turtles$LifeStage
## t = 10.082, df = 7.4712, p-value = 1.296e-05
## alternative hypothesis: true difference in means between group Adult and group Juvenile is not equal
## 95 percent confidence interval:
##  149.1647 239.0641
## sample estimates:
##    mean in group Adult mean in group Juvenile
##               344.5429               150.4286
```

```r
t_StraightlineCL = t.test(StraightlineCL~ Male_Turtles$LifeStage, data = Male_Turtles)
t_StraightlineCL
```

```
##
##  Welch Two Sample t-test
##
## data:  StraightlineCL by Male_Turtles$LifeStage
## t = 7.4131, df = 6.5926, p-value = 0.0001981
## alternative hypothesis: true difference in means between group Adult and group Juvenile is not equal
## 95 percent confidence interval:
##  19.69580 38.49128
## sample estimates:
##    mean in group Adult mean in group Juvenile
##              120.71497               91.62143
```

```r
t_MaxCW = t.test(MaxCW~ Male_Turtles$LifeStage, data = Male_Turtles)
t_MaxCW
```

```
##
##  Welch Two Sample t-test
##
## data:  MaxCW by Male_Turtles$LifeStage
## t = 3.6589, df = 6.0944, p-value = 0.01031
## alternative hypothesis: true difference in means between group Adult and group Juvenile is not equal
## 95 percent confidence interval:
##   9.858259 49.216763
## sample estimates:
##    mean in group Adult mean in group Juvenile
##               96.02037               66.48286
```

```
t_PL_AnteriortoHinge = t.test(PL_AnteriortoHinge~ Male_Turtles$LifeStage, data = Male_Turtles)
t_PL_AnteriortoHinge
```

```
##
##  Welch Two Sample t-test
##
## data:  PL_AnteriortoHinge by Male_Turtles$LifeStage
## t = 5.0659, df = 6.4597, p-value = 0.001848
## alternative hypothesis: true difference in means between group Adult and group Juvenile is not equal
## 95 percent confidence interval:
##   5.751776 16.151590
## sample estimates:
##    mean in group Adult mean in group Juvenile
##               48.92454               37.97286
```

```
t_PL_HingetoPosterior = t.test(PL_HingetoPosterior~ Male_Turtles$LifeStage, data = Male_Turtles)
t_PL_HingetoPosterior
```

```
##
##  Welch Two Sample t-test
##
## data:  PL_HingetoPosterior by Male_Turtles$LifeStage
## t = 6.0345, df = 6.5329, p-value = 0.0006815
## alternative hypothesis: true difference in means between group Adult and group Juvenile is not equal
## 95 percent confidence interval:
##   9.918988 23.012580
## sample estimates:
##    mean in group Adult mean in group Juvenile
##               70.97436               54.50857
```

```
t_ShellHeightatHinge = t.test(ShellHeightatHinge~ Male_Turtles$LifeStage, data = Male_Turtles)
t_ShellHeightatHinge
```

```
##
##  Welch Two Sample t-test
##
## data:  ShellHeightatHinge by Male_Turtles$LifeStage
## t = 6.0769, df = 6.4425, p-value = 0.0006906
## alternative hypothesis: true difference in means between group Adult and group Juvenile is not equal
## 95 percent confidence interval:
```

```
##   7.759776 17.936981
## sample estimates:
##    mean in group Adult mean in group Juvenile
##               56.28552               43.43714
```

Response in Complete Sentences: In comparing the number of *Annuli* between adult and juvenile male turtles, the Welch's t-test produced a p-value of 0.1051, greater than 0.05, indicating no statistically significant difference between the two groups. This means we can not reject null hypothesis, which suggests the number of *Annuli* between the two groups is the same.

For the rest of the variables, the p-value are all less than 0.05, indicating significant difference between the two groups. This means we can reject null hypothesis, suggesting that the means for *Mass*, *StraightlineCL*, *MaxCW*, *PL_AnteriortoHinge*, *PL_HingetoPosterior* and *ShellHeightatHinge* differ between the two groups.

From the result, most of the variables (*Mass*, *StraightlineCL*, *MaxCW*, *PL_AnteriortoHinge*, *PL_HingetoPosterior*, *ShellHeightatHinge*) show significant differences between adult and juvenile male turtles, with adults generally having larger measurements. The exception is *Annuli*, where the difference is not statistically significant. This suggests that while many physical characteristics differ between adult and juvenile turtles, the number of *Annuli* does not vary significantly with age.

This result indicates that adult turtles tend to have larger physical dimensions than juvenile turtles, which is consistent with expectations given that adults have had more time to grow. The results suggest that different stages of life (adult vs. juvenile) are associated with significant differences in several key physical traits.