# ADA2: Class 01, R, Review

Tim Farkas

January 23, 2022

Write R code to answer the quiz questions on Learn using the dataset below.

## Rubric for grading

For these questions below:

- 3. (2 p) plot and interpretation.
- 5. (2 p) plot and interpretation.
- 7. (2 p) plot and interpretation.
- 10. (4 p) code and output appear correct, no errors.

Note that because the **Quiz 1 questions** also use this data, those questions are also in this document typeset in preformatted text, like this:

```
Quiz 1. What was the lowest recorded punting distance among the 13 participants?
```

---

## American Football Punters

### Description

Investigators studied physical characteristics and ability in 13 football punters. Each volunteer punted a football ten times. The investigators recorded the average distance for the ten punts, in feet. They also recorded the average hang time (time the ball is in the air before the receiver catches it) for the ten punts, in seconds. In addition, the investigators recorded five measures of strength and flexibility for each punter: right leg strength (pounds), left leg strength (pounds), right hamstring muscle flexibility (degrees), left hamstring muscle flexibility (degrees), and overall leg strength (foot-pounds). From the study "The relationship between selected physical performance variables and football punting ability" by the Department of Health, Physical Education and Recreation at the Virginia Polytechnic Institute and State University, 1983.

```
Variable        Description
-------------   --------------------------------
Distance        Distance travelled in feet
Hang            Time in air in seconds
R_Strength      Right leg strength in pounds
L_Strength      Left leg strength in pounds
R_Flexibility   Right leg flexibility in degrees
L_Flexibility   Left leg flexibility in degrees
O_Strength      Overall leg strength in pounds
```

Data File: `ADA2_CL_01_punting.csv`

### Source

The Relationship Between Selected Physical Performance Variables and Football Punting Ability. Department of Health, Physical Education and Recreation, Virginia Polytechnic Institute and State University, 1983.

---

# Rubric

1. Read the data set into R.

```
library(tidyverse)

# First, download the data to your computer,
#   save in the same folder as this Rmd file.

# read the data
dat_punt <- readr::read_csv("ADA2_CL_01_punting.csv", skip = 1)
str(dat_punt)
```

```
spec_tbl_df [13 x 7] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
 $ Distance     : num [1:13] 162 144 148 164 192 ...
 $ Hang         : num [1:13] 4.75 4.07 4.04 4.18 4.35 4.16 4.43 3.2 3.02 3.64 ...
 $ R_Strength   : num [1:13] 170 140 180 160 170 150 170 110 120 130 ...
 $ L_Strength   : num [1:13] 170 130 170 160 150 150 180 110 110 120 ...
 $ R_Flexibility: num [1:13] 106 92 93 103 104 101 108 86 90 85 ...
 $ L_Flexibility: num [1:13] 106 93 78 93 93 87 106 92 86 80 ...
 $ O_Strength   : num [1:13] 241 195 153 197 267 ...
 - attr(*, "spec")=
  .. cols(
  ..   Distance = col_double(),
  ..   Hang = col_double(),
  ..   R_Strength = col_double(),
  ..   L_Strength = col_double(),
  ..   R_Flexibility = col_double(),
  ..   L_Flexibility = col_double(),
  ..   O_Strength = col_double()
  .. )
 - attr(*, "problems")=<externalptr>
```

```
#dat_punt
```

2. Generate summaries `summary()` and frequency tables `table()` for each variable. Answer questions 1–7.

```
# I'll get you started with the code, the rest is up to you.
summary(dat_punt)
```

```
    Distance          Hang         R_Strength      L_Strength
 Min.   :104.9   Min.   :3.020   Min.   :110.0   Min.   :110.0
 1st Qu.:140.2   1st Qu.:3.640   1st Qu.:130.0   1st Qu.:130.0
 Median :150.2   Median :4.040   Median :150.0   Median :150.0
 Mean   :148.2   Mean   :3.921   Mean   :147.7   Mean   :143.8
```

```
 3rd Qu.:163.5   3rd Qu.:4.180   3rd Qu.:170.0   3rd Qu.:160.0
 Max.   :192.0   Max.   :4.750   Max.   :180.0   Max.   :180.0
 R_Flexibility    L_Flexibility      O_Strength
 Min.   : 85.00   Min.   : 78.00   Min.   :130.2
 1st Qu.: 90.00   1st Qu.: 86.00   1st Qu.:153.9
 Median : 93.00   Median : 93.00   Median :197.1
 Mean   : 95.69   Mean   : 91.23   Mean   :196.2
 3rd Qu.:103.00   3rd Qu.: 94.00   3rd Qu.:240.6
 Max.   :108.00   Max.   :106.00   Max.   :266.6
```

```
apply(dat_punt, 2, table)
```

```
$Distance

104.93 105.67 117.59 140.25    144  147.5 150.17    162  162.5  163.5 165.17
     1      1      1      1      1      1      1      1      1      1      1
171.75    192
     1      1

$Hang

3.02  3.2  3.6 3.64 3.68 3.85 4.04 4.07 4.16 4.18 4.35 4.43 4.75
   1    1    1    1    1    1    1    1    1    1    1    1    1

$R_Strength

110 120 130 140 150 160 170 180
  1   2   1   2   1   2   3   1

$L_Strength

110 120 130 140 150 160 170 180
  2   1   2   1   3   1   2   1

$R_Flexibility

 85  86  89  90  92  93  95 101 103 104 106 108
  1   1   1   1   2   1   1   1   1   1   1   1

$L_Flexibility

 78  80  83  86  87  92  93  94  95 106
  1   1   1   1   1   1   3   1   1   2

$O_Strength

130.24 132.68 152.99 153.92 154.64 195.49 197.09 205.88 219.25 240.57 260.56
     1      1      1      1      1      1      1      1      1      2      1
266.56
     1
```

Note that you can do even better than reading the numbers from above to answer the specific **quiz questions**. Instead, you can (not required) write code that returns the specific values you want. For example:

- 1. The minimum distance is 104.93 ft.

```
Quiz 1. What was the lowest recorded punting distance among the 13 participants?
Quiz 2. What was the highest recorded hang time among the 13 participants?
Quiz 3. Is the range of values for R_Strength the same or different than the range of values fo
Quiz 4. What percentage of the sample has a L_Strength of 110 pounds?
Quiz 5. Is the range of values for R_Flexibility the same or different than the range of values
Quiz 6. What percentage of the sample has a L_Flexibility of 106 degrees?
Quiz 7. What is the most common value for O_Strength (i.e., what is the modal value)?
```

Q1: Min of distance is 104.93.
Q2: Max of hang time is 4.75.
Q3: Range of right strength is 110, 180, and range of left strength is 110, 180.

```
pLS <- ecdf(dat_punt$L_Strength)
pLS_110 <- pLS(110) - pLS(100)
```

Q4: Percentage of sample with L_Strength of 110 lbs is 0.1538462.
Q5: Range of right strength is 85, 108, and range of left strength is 78, 106.

```
pLF <- ecdf(dat_punt$L_Flexibility)
pLF_106 <- pLF(106) - pLF(105)
```

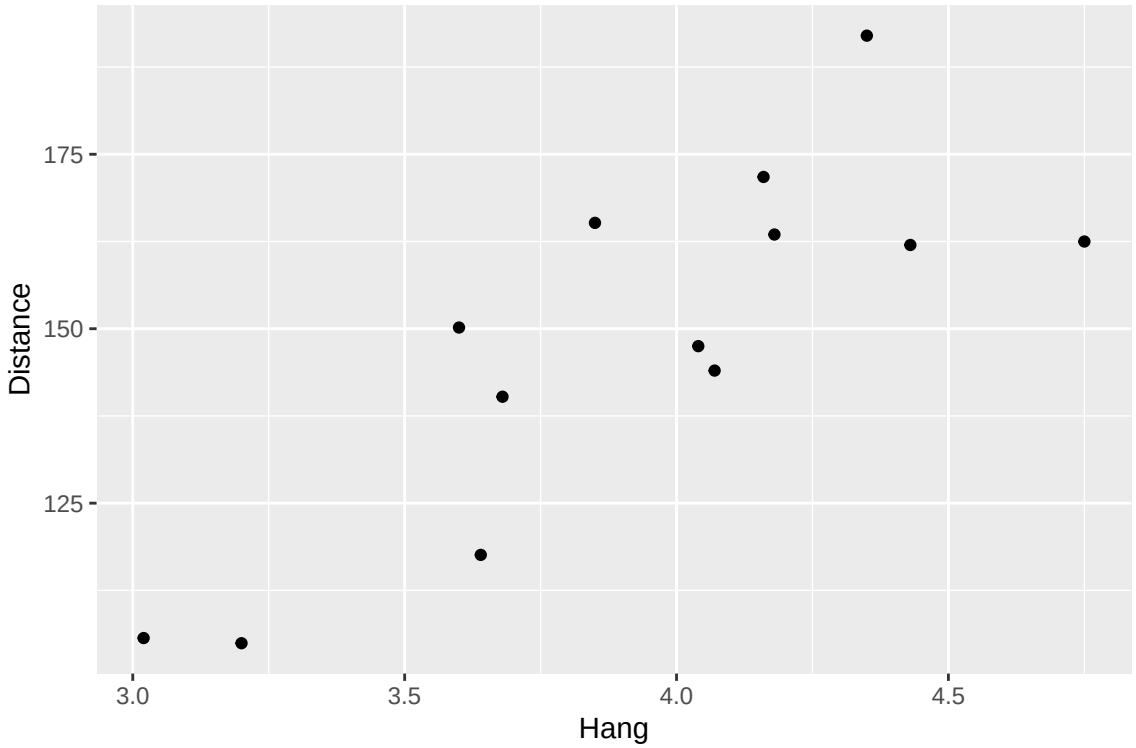
Q6: Percentage of sample with L_Flexibility of 106lbs is 0.1538462.

```
OS_mode <- dat_punt %>%
  group_by(O_Strength) %>%
  summarize(count = n()) %>%
  arrange(desc(count)) %>%
  slice(1) %>%
  pull(O_Strength)
```

Q7: The modal value for O_Strength is 240.57.

3. (2 p) Plot $y$ =`Distance` and $x$ =`Hang` and interpret the plot in terms of linearity and strength of correlation.

```
# plot distance by hang
library(ggplot2)
p <- ggplot(dat_punt, aes(x = Hang , y = Distance)) +
  geom_point()
# ...
print(p)
```




The relationship appears to be both strong (high correlation) and rather linear.

4. Calculate the Pearson correlation between `Distance` and `Hang` (read the help for performing the hypothesis test). Answer questions 8–9.

```
dhcor <- cor(dat_punt$Distance, dat_punt$Hang)
dhct <- cor.test(x = dat_punt$Distance, dat_punt$Hang)
```

```
  Quiz 8. What is the correlation between Distance and Hang?
  Quiz 9. The corresponding p-value for the correlation between Distance and Hang is ____.
```

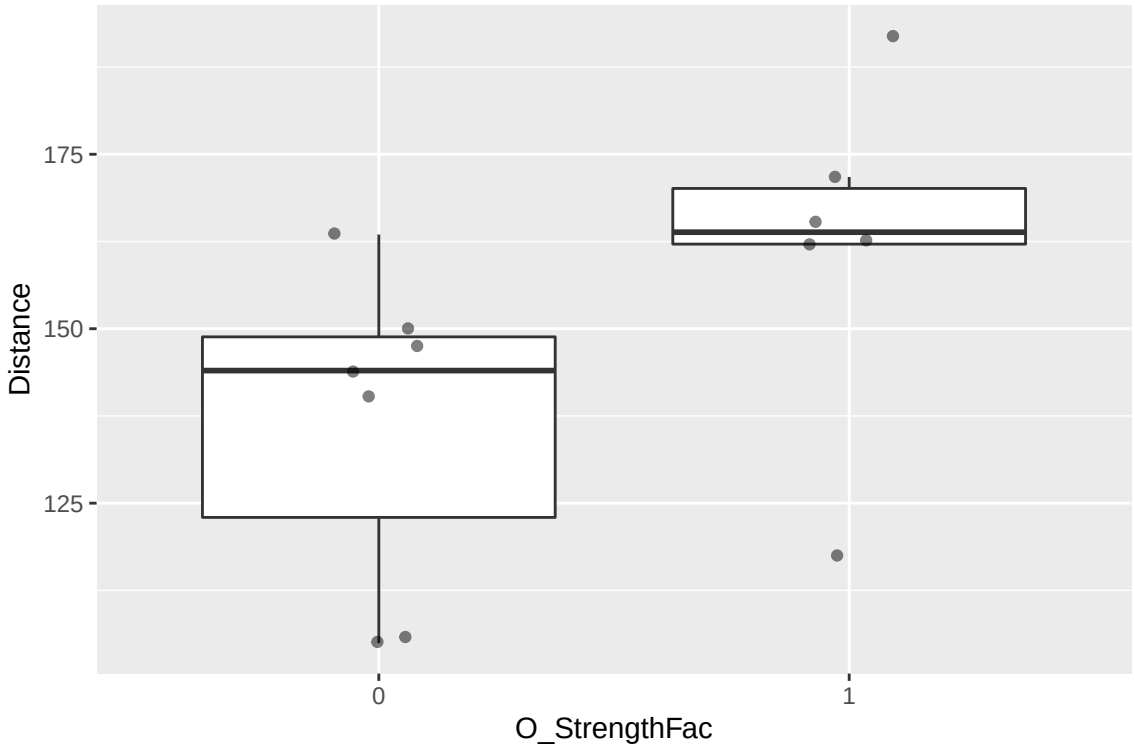
Q8: The Pearson correlation between Distance and Hang is 0.819.
Q9: The p-value for the correlation is rdhct$p.value‘.

5. (2 p) Create a new categorical (factor) variable, `O_StrengthFac`, from the quantitative variable overall leg strength (`O_Strength`) to indicate high leg strength: code less than 200 as 0 (low leg strength) and at least 200 as 1 (high leg strength).

```
library(magrittr)
# create categorical variable
dat_punt %<>%
  mutate(O_StrengthFac = as.factor(ifelse(O_Strength < 200, 0, 1)))
```

Plot $y$ =`Distance` and $x$ =`O_StrengthFac` and interpret the comparison of distance by strength group.

```
# plot distance by strength group
library(ggplot2)
 p <- ggplot(dat_punt, aes(y = Distance, x = O_StrengthFac)) +
   geom_boxplot(outlier.shape = NA) +
   geom_point(position=position_jitter(width = 0.1), alpha = 0.5)
print(p)
```




6. Use a two-sample $t$-test (assume equal variance) to test whether $H_0 : \mu_{\text{low}} = \mu_{\text{high}}$, that the population means for distance are equal for the two overall leg strength groups you created. Answer questions 10–11.

```
tt <- t.test(Distance ~ O_StrengthFac, data = dat_punt, var.equal = TRUE )
tt
```

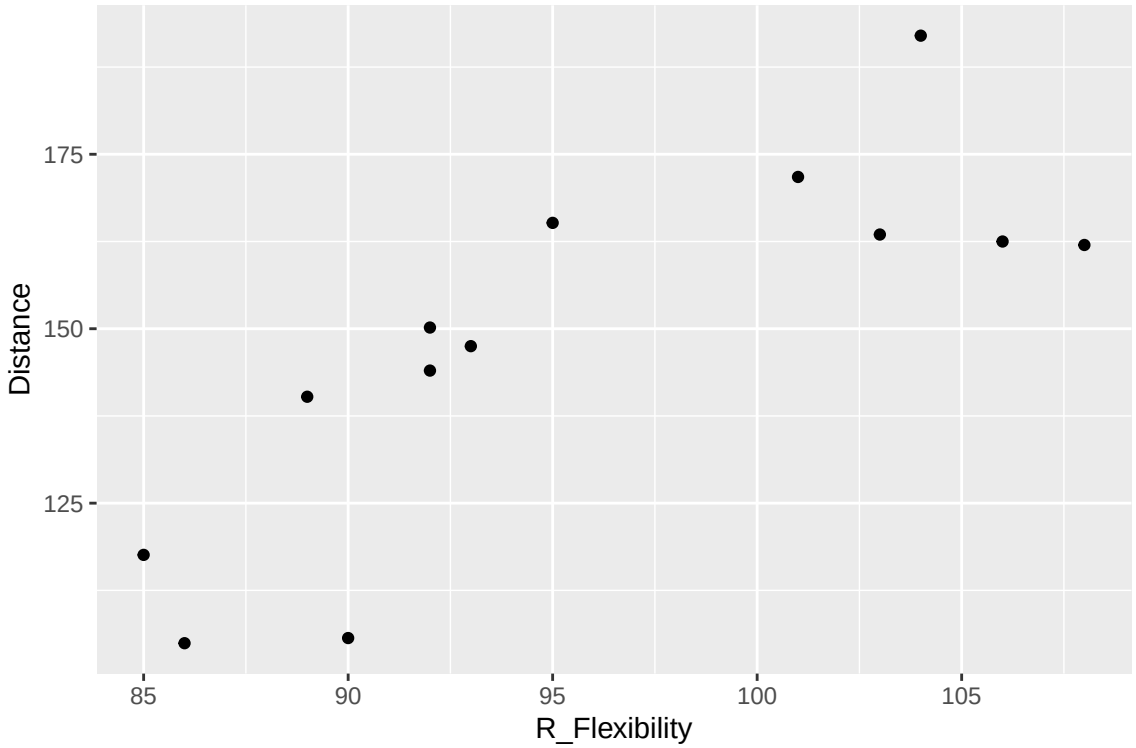
```
    Two Sample t-test

data:  Distance by O_StrengthFac
t = -1.939, df = 11, p-value = 0.07858
alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
95 percent confidence interval:
 -53.934775   3.413347
sample estimates:
mean in group 0 mean in group 1
       136.5743        161.8350
```

```
  Quiz 10. Is distance significantly associated with overall strength (categorical) at an alpha =
  Quiz 11. What is the mean distance in feet for the low and high strength groups, respectively?
```

Q10: The p-value for association between strength and distance is 0.0785776, so distance is not significant related at $\alpha = 0.05$.
Q11: The mean distance in feet for low strength is 136.5742857, and the mean for high strength is 161.835.

7. (2 p) Plot $y =$`Distance` and $x =$`R_Flexibility` and interpret the relationship.

```
library(ggplot2)
 p <- ggplot(dat_punt, aes(x = R_Flexibility , y = Distance )) +
  geom_point()
 print(p)
```




There appears to be a positive, more-or-less linear relationship between flexibility in the right leg and distance. We might investigate whether this relationship saturates, flattens out at high flexibilities.

8. Regress $y =$`Distance` on $x =$`R_Flexibility`. Answer questions 12–13.

```
an1 <- lm(Distance ~ R_Flexibility, data = dat_punt)
sum1 <- summary(an1)
sum1
```

```
Call:
lm(formula = Distance ~ R_Flexibility, data = dat_punt)

Residuals:
    Min      1Q  Median      3Q     Max
-27.267 -13.431   5.689  10.000  21.443

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   -108.9013    57.0426  -1.909  0.08266 .
R_Flexibility    2.6871     0.5943   4.522  0.00087 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 16.04 on 11 degrees of freedom
Multiple R-squared:  0.6502,    Adjusted R-squared:  0.6184
F-statistic: 20.44 on 1 and 11 DF,  p-value: 0.0008698
```

```
Quiz 12. What is the expected increase in distance for each degree increase in flexibility?
Quiz 13. Is distance significantly associated with flexibility at an alpha = 0.05 level?
```

Q12: The expected increase in distance is 2.687 per degree increase in flexibility.
Q13: The p-value associated with flexibility is 0.001, so there is a significant relationships at $\alpha = 0.05$.

9. Create a new variable which is the mean of the right leg and left leg flexibility variables, `O_Flexibility`. Generate a frequency distribution for this new variable. Answer questions 14–15.

```
dat_punt %<>%
  rowwise %>%
  dplyr::mutate(O_Flexibility = mean(c(L_Flexibility, R_Flexibility)))
```

```
  Quiz 14. What is the median value for your new variable that is the mean of the right and left
  Quiz 15. What percentage of the sample has a mean flexibility no more than 86 degrees?
```

Q14: The median for the new flexibility variable is 93.
Q15: 23.1% of the sample has a mean flexibility no more than 86 degrees.

10. (4 p) Upload your error-free program (html output as PDF file) showing your work and your plots for additional points.