

Metadata

Course: DS 5100
Module: 11 R Programming 2
Topic: HW on Tidyverse
Author: R.C. Alvarado (adapted)
Date: 7 July 2023 (revised)

Student Info

Name: Bernard Gonzales
Net ID: qja8kc
File GitHub URL: <https://github.com/bernard-gonzales/play-space/blob/main/M11-HW-2.Rmd>

Instructions

In your **private course repo** use this notebook to write code that performs the tasks below.

Save your notebook in the M11 directory.

Remember to add and commit these files to your repo.

Then push your commits to your repo on GitHub.

Be sure to fill out the **Student Info** block above.

To submit your homework, save your results as a PDF and upload it to GradeScope.

TOTAL POINTS: 7

Overview

In this homework, you will work with the Abalone dataset from the UCI Machine Learning Repository.

To get started, download and import the `abalone.data` dataset from this URL:

- <https://archive.ics.uci.edu/ml/machine-learning-databases/abalone/abalone.data>

You can pass the URL directly to `read.csv()` and that there is no header row.

Note: The instruction to print in the questions below can be accomplished either through the `print()` function or by displaying a value directly.

TOTAL POINTS: 7

Tasks

Task 0

(0 points)

Get the dataset.

```
data <- read.csv('https://archive.ics.uci.edu/ml/machine-learning-databases/abalone/abalone.data')
```

Task 1

(1 point)

Print the number of rows in the dataset.

```
nrow(data)
```

```
## [1] 4176
```

Task 2

(1 point)

The rightmost column is the number of rings. Print the maximum number of rings

```
max(data$X15)
```

```
## [1] 29
```

Task 3

(1 point)

The leftmost column is the gender with these values: M: male, F: female, I: infant.

Apply the `filter()` function from tidyverse to select only rows where gender is infant, and print the number of records.

```
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 4.3.3
```

```
## Warning: package 'ggplot2' was built under R version 4.3.3
```

```
## Warning: package 'tibble' was built under R version 4.3.3
```

```
## Warning: package 'tidyr' was built under R version 4.3.3
```

```
## Warning: package 'readr' was built under R version 4.3.3
```

```
## Warning: package 'purrr' was built under R version 4.3.3
```

```
## Warning: package 'dplyr' was built under R version 4.3.3
```

```
## Warning: package 'stringr' was built under R version 4.3.3
```

```
## Warning: package 'forcats' was built under R version 4.3.3

## Warning: package 'lubridate' was built under R version 4.3.3

## -- Attaching core tidyverse packages ----- tidyverse 2.0.0 --
## v dplyr      1.1.4      v readr      2.1.5
## v forcats    1.0.0      v stringr    1.5.1
## v ggplot2    3.5.1      v tibble     3.2.1
## v lubridate  1.9.3      v tidyr      1.3.1
## v purrr      1.0.2
## -- Conflicts ----- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()     masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors

abalone <- tibble(data)
abalone %>%
  filter(M == "I")
```

```
## # A tibble: 1,342 x 9
##   M      X0.455 X0.365 X0.095 X0.514 X0.2245 X0.101 X0.15   X15
##   <chr>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl> <dbl> <int>
## 1 I      0.33    0.255   0.08    0.205   0.0895  0.0395 0.055    7
## 2 I      0.425   0.3     0.095   0.352   0.141   0.0775 0.12     8
## 3 I      0.355   0.28    0.085   0.290   0.095   0.0395 0.115    7
## 4 I      0.38    0.275   0.1     0.226   0.08    0.049   0.085   10
## 5 I      0.24    0.175   0.045   0.07    0.0315  0.0235 0.02     5
## 6 I      0.205   0.15    0.055   0.042   0.0255  0.015   0.012    5
## 7 I      0.21    0.15    0.05    0.042   0.0175  0.0125 0.015    4
## 8 I      0.39    0.295   0.095   0.203   0.0875  0.045   0.075    7
## 9 I      0.325   0.245   0.07    0.161   0.0755  0.0255 0.045    6
## 10 I     0.52    0.41    0.12    0.595   0.238   0.111   0.19     8
## # i 1,332 more rows
```

Task 4

(1 point)

Apply the `filter()` function from `tidyverse` to select only rows where gender is infant or male, and print the number of records.

```
s <- abalone %>%
  filter(M == "I" | M == "M")
dim(s)
```

```
## [1] 2869    9
```

Task 5

(1 point)

Call the `table()` function on the abalone genders to find out how many of each gender are present. Print the result.

```
table(abalone$M)
```

```
##  
##      F      I      M  
## 1307 1342 1527
```

Task 6

(1 point)

Compute the mean value of column 2 (V2) grouped by gender.

V2 is the longest shell measurement.

Requirements: use the `%>%` operator to chain commands, and the `group_by()` and `summarize()` functions.

```
abalone %>%  
  group_by(M) %>%  
  summarize(mean(X0.455))
```

```
## # A tibble: 3 x 2  
##   M      'mean(X0.455)'  
##   <chr>          <dbl>  
## 1 F              0.579  
## 2 I              0.428  
## 3 M              0.561
```

Task 7

(1 point)

Compute the MEDIAN value of longest shell measurement for only the males.

Requirements: use the `%>%` operator to chain commands.

```
abalone %>%  
  filter(M == "M") %>%  
  summarize(median(X0.455))
```

```
## # A tibble: 1 x 1  
##   'median(X0.455)'  
##           <dbl>  
## 1           0.58
```