Metadata

```
Course:   DS 5100
Module:   11 R Programming 2
Topic:    HW on Tidyverse
Author:   R.C. Alvarado (adapted)
Date:     07 October 2022 (revised)
```

Student Info

```
Name: Ballard
Net ID: bkq5nt
File GitHub URL:https://github.com/ballard11/DS5100-2022-08-0/tree/main/lessons/M11_RDplyr
```

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 GitHib.

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.

```r
# Read Dataset

df<-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.

```r
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 4.1.3
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
df %>% count()
```

```
##      n
## 1 4176
```

```r
#Or
dim(df)
```

```
## [1] 4176    9
```

Task 2

(1 point)

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

```r
# max rings is 29
df %>%
  summarise(max = max(X15, na.rm=TRUE))
```

```
##   max
## 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.

```
# filter to rows with Infant

task3<-df %>%
  filter(M == 'I')

head(task3)
```

```
##   M X0.455 X0.365 X0.095 X0.514 X0.2245 X0.101 X0.15 X15
## 1 I  0.330  0.255  0.080 0.2050  0.0895 0.0395 0.055   7
## 2 I  0.425  0.300  0.095 0.3515  0.1410 0.0775 0.120   8
## 3 I  0.355  0.280  0.085 0.2905  0.0950 0.0395 0.115   7
## 4 I  0.380  0.275  0.100 0.2255  0.0800 0.0490 0.085  10
## 5 I  0.240  0.175  0.045 0.0700  0.0315 0.0235 0.020   5
## 6 I  0.205  0.150  0.055 0.0420  0.0255 0.0150 0.012   5
```

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.

```
# filter to rows with Infant or Male

task4 <- df %>%
  filter(M == 'I' | M == 'M')

head(task4)
```

```
##   M X0.455 X0.365 X0.095 X0.514 X0.2245 X0.101 X0.15 X15
## 1 M  0.350  0.265  0.090 0.2255  0.0995 0.0485 0.070   7
## 2 M  0.440  0.365  0.125 0.5160  0.2155 0.1140 0.155  10
## 3 I  0.330  0.255  0.080 0.2050  0.0895 0.0395 0.055   7
## 4 I  0.425  0.300  0.095 0.3515  0.1410 0.0775 0.120   8
## 5 M  0.475  0.370  0.125 0.5095  0.2165 0.1125 0.165   9
## 6 M  0.430  0.350  0.110 0.4060  0.1675 0.0810 0.135  10
```

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.

```r
# call table function

table(df$M)
```

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

```r
#Base R
```

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.

```r
# group by and summarize
df2 <- df %>%
    group_by(M) %>%
    summarize(X0.455_Mean = mean(X0.455, na.rm = TRUE))


print(df2)
```

```
## # A tibble: 3 x 2
##   M     X0.455_Mean
##   <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.

```r
# group by and summarize

df3 <- df %>%
    group_by(M) %>%
    filter(M=='M') %>%
    summarize(X0.455_Max = max(X0.455, na.rm = TRUE))

#Max shell length for males
print(df3)
```

```
## # A tibble: 1 x 2
##   M     X0.455_Max
##   <chr>      <dbl>
## 1 M           0.78
```