DS5100_HW11

Eashan Kaw

4/6/2023

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:  Eashan Kaw
Net ID: elk7ed
File GitHub URL: https://github.com/elkaw/DS5100-2023-01-O/blob/main/lessons/M10_RBasics/M11-HW.ipynb
```

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 (https://archive.ics.uci.edu/ml/datasets/Abalone) 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 (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.

```
# CODE HERE
#install.packages("tinytex")
library(tinytex)
library(dplyr)
```

```
## Registered S3 methods overwritten by 'tibble':
##   method     from
##   format.tbl pillar
##   print.tbl  pillar
```

```
##
## 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
```

```
abalone_df <- read.csv('/Users/eashankaw/Documents/Continuing Education/DS5100_programming_FOR_DATA_SCIENCE/abalo
ne.data', header = F)
head(abalone_df)
```

```
##   V1    V2    V3    V4     V5     V6     V7    V8 V9
## 1  M 0.455 0.365 0.095 0.5140 0.2245 0.1010 0.150 15
## 2  M 0.350 0.265 0.090 0.2255 0.0995 0.0485 0.070  7
## 3  F 0.530 0.420 0.135 0.6770 0.2565 0.1415 0.210  9
## 4  M 0.440 0.365 0.125 0.5160 0.2155 0.1140 0.155 10
## 5  I 0.330 0.255 0.080 0.2050 0.0895 0.0395 0.055  7
## 6  I 0.425 0.300 0.095 0.3515 0.1410 0.0775 0.120  8
```

Task 1

(1 point)

Print the number of rows in the dataset.

```
# CODE HERE
print(nrow(abalone_df))
```

```
## [1] 4177
```

Task 2

(1 point)

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

```
# CODE HERE
print(abalone_df %>% select(V9)  %>% 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.

```
# CODE HERE
#print(abalone_df$M %>% unique)
print(nrow(abalone_df %>% filter(abalone_df$V1 == "I")))
```

```
## [1] 1342
```

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.

```
# CODE HERE
print(nrow(abalone_df %>% filter(abalone_df$V1 == "I" | abalone_df$V1 == "M")))
```

```
## [1] 2870
```

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.

```
# CODE HERE
print(abalone_df$V1 %>% table)
```

```
## .
##    F    I    M
## 1307 1342 1528
```

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.

```
# CODE HERE
abalone_df %>% group_by(abalone_df$V1)  %>% summarize(mean(abalone_df$V2))
```

```
## Warning in pillar::colonnade(df, has_row_id = if (star) "*" else TRUE,
## needs_dots = needs_dots): `...` must be empty.
```

```
## # A tibble: 3 x 2
##   `abalone_df$V1` `mean(abalone_df$V2)`
##   <chr>                           <dbl>
## 1 F                               0.524
## 2 I                               0.524
## 3 M                               0.524
```

Task 7

(1 point)

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

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

```
# CODE HERE

print(abalone_df %>% filter(abalone_df$V1 == 'M') %>% select(V2)  %>% summarize(median(abalone_df$V2)))
```

```
##   median(abalone_df$V2)
## 1                 0.545
```