STAT 33A/B Lab Workbook Wk 13

Won Shil Park (3033452021)

Apr 16, 2021

This workbook is due **Apr 22, 2020** by 9:00am or by midnight Apr 23 if you attend lab.

- Knit and submit the generated PDF file on Gradescope.

For each of the following exercises, use the dogs data.

```
load(url("http://www.stat.berkeley.edu/users/nolan/data/dogs.rda"))
```

```
library(dplyr)
```

```
##
## 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
```

Exercise 1

For each of the following tasks, show the code to compute the result:

- Using base R
- Using dplyr

1. The mean and median of the longevity column (ignoring missing values).
2. The subset that contains rows 10-20 of the height, weight, and longevity columns.
3. The number of dog breeds with weight greater than 42.
4. The subset of large dogs that require daily grooming.

Part 1

```r
# Base R
mean(dogs$longevity, na.rm = TRUE)
```

```
## [1] 10.95674
```

```r
median(dogs$longevity, na.rm = TRUE)
```

```
## [1] 11.29
```

```r
# dplyr
summarize(dogs, mean(longevity, na.rm = TRUE))
```

```
##   mean(longevity, na.rm = TRUE)
## 1                      10.95674
```

```r
summarize(dogs, median(longevity, na.rm = TRUE))
```

```
##   median(longevity, na.rm = TRUE)
## 1                           11.29
```

Part 2

```r
# Base R
dogs[10:20, c("height", "weight", "longevity")]
```

```
##    height weight longevity
## 10  14.50   22.0     12.53
## 11  21.75   47.5     12.58
## 12  10.50   15.0     13.92
## 13  10.25     NA     11.42
## 14     NA   24.0     12.63
## 15  13.00   15.5     11.81
## 16   5.00    5.5     16.50
## 17  10.50     NA     11.05
## 18  20.00     NA     12.87
## 19  19.50   45.0     12.54
## 20  10.50     NA     12.80
```

```r
# dplyr
select(slice(dogs, 10:20), height, weight, longevity)
```

```
##    height weight longevity
## 1   14.50   22.0     12.53
## 2   21.75   47.5     12.58
## 3   10.50   15.0     13.92
## 4   10.25     NA     11.42
## 5      NA   24.0     12.63
```

```
## 6   13.00   15.5     11.81
## 7    5.00    5.5     16.50
## 8   10.50     NA     11.05
## 9   20.00     NA     12.87
## 10  19.50   45.0     12.54
## 11  10.50     NA     12.80
```

Part 3

```
# Base R
sum(dogs$weight > 42, na.rm = TRUE)
```

```
## [1] 37
```

```
# dplyr
summarise(dogs, sum(weight > 42, na.rm = TRUE))
```

```
##   sum(weight > 42, na.rm = TRUE)
## 1                             37
```

```
count(dogs, weight > 42)
```

```
##   weight > 42  n
## 1       FALSE 49
## 2        TRUE 37
## 3          NA 86
```

Part 4

```
# Base R
dogs[dogs$size == "large" & dogs$grooming == "daily" & !is.na(dogs$grooming),]
```

```
##               breed   group datadog popularity_all popularity lifetime_cost
## 44           Briard herding    2.71            125         79         19673
## 62  Giant Schnauzer working    2.38             95         70         26686
## 67     Afghan Hound   hound    2.08             88         66         24077
## 75           Borzoi   hound    1.89            102         71         16176
## 79 Alaskan Malamute working    1.82             58         47         21986
## 86    Saint Bernard working    1.42             49         43         20022
##    intelligence_rank longevity ailments price food_cost grooming   kids
## 44                30     11.17        1   650       466    daily   high
## 62                28     10.00        1   810      1349    daily medium
## 67                80     11.92        0   890       710    daily   high
## 75                76      9.08        0   675       466    daily medium
## 79                50     10.67        2  1210       710    daily medium
## 86                65      7.78        3   875      1217    daily   high
##    megarank_kids megarank  size weight height
## 44            44       33 large     NA   24.5
```

```
## 62            62       67 large   77.5   25.5
## 67            67       60 large   55.0   26.0
## 75            75       82 large   82.5   28.0
## 79            79       83 large   80.0   24.0
## 86            86       81 large  155.0   26.5
```

```
subset(dogs, size == "large" & grooming == "daily")
```

```
##               breed   group datadog popularity_all popularity lifetime_cost
## 44           Briard herding    2.71            125         79         19673
## 62  Giant Schnauzer working    2.38             95         70         26686
## 67     Afghan Hound   hound    2.08             88         66         24077
## 75           Borzoi   hound    1.89            102         71         16176
## 79 Alaskan Malamute working    1.82             58         47         21986
## 86    Saint Bernard working    1.42             49         43         20022
##    intelligence_rank longevity ailments price food_cost grooming   kids
## 44                30     11.17        1   650       466    daily   high
## 62                28     10.00        1   810      1349    daily medium
## 67                80     11.92        0   890       710    daily   high
## 75                76      9.08        0   675       466    daily medium
## 79                50     10.67        2  1210       710    daily medium
## 86                65      7.78        3   875      1217    daily   high
##    megarank_kids megarank  size weight height
## 44            44       33 large     NA   24.5
## 62            62       67 large   77.5   25.5
## 67            67       60 large   55.0   26.0
## 75            75       82 large   82.5   28.0
## 79            79       83 large   80.0   24.0
## 86            86       81 large  155.0   26.5
```

```
# dplyr
filter(dogs, size == "large" & grooming == "daily")
```

```
##               breed   group datadog popularity_all popularity lifetime_cost
## 1            Briard herding    2.71            125         79         19673
## 2  Giant Schnauzer working    2.38             95         70         26686
## 3     Afghan Hound   hound    2.08             88         66         24077
## 4            Borzoi   hound    1.89            102         71         16176
## 5 Alaskan Malamute working    1.82             58         47         21986
## 6    Saint Bernard working    1.42             49         43         20022
##   intelligence_rank longevity ailments price food_cost grooming   kids
## 1                30     11.17        1   650       466    daily   high
## 2                28     10.00        1   810      1349    daily medium
## 3                80     11.92        0   890       710    daily   high
## 4                76      9.08        0   675       466    daily medium
## 5                50     10.67        2  1210       710    daily medium
## 6                65      7.78        3   875      1217    daily   high
##   megarank_kids megarank  size weight height
## 1            44       33 large     NA   24.5
## 2            62       67 large   77.5   25.5
## 3            67       60 large   55.0   26.0
## 4            75       82 large   82.5   28.0
## 5            79       83 large   80.0   24.0
## 6            86       81 large  155.0   26.5
```

Exercise 2

Use dplyr and the dogs data to determine which 3 dogs cost the most.

Your answer to this exercise should be a data frame with 3 rows.

```
sorted = arrange(dogs, desc(price))
slice(sorted, 1:3)
```

```
##                   breed        group datadog popularity_all popularity
## 1       Tibetan Mastiff      working      NA            122         NA
## 2 Black Russian Terrier      working      NA            128         NA
## 3               Bulldog non-sporting    0.99              6          6
##   lifetime_cost intelligence_rank longevity ailments price food_cost grooming
## 1         23747                NA     11.92       NA  3460        NA   weekly
## 2            NA                NA        NA        0  2833        NA     <NA>
## 3         13479                78      6.29        5  2680       466   weekly
##     kids megarank_kids megarank   size weight height
## 1   high            NA       NA  large     NA     25
## 2   <NA>            NA       NA  large     NA     28
## 3 medium            87       87 medium     45     NA
```

Exercise 3

Use dplyr to answer each of the following:

1. On average, which `group` of dog has the highest lifetime cost? Which has the lowest?
2. How many dogs are there for each possible combination of `size` and `grooming`?

Part 1

```
groups = group_by(dogs, group)
costs = summarize(groups, mean_cost = mean(lifetime_cost, na.rm = TRUE))
arrange(costs, mean_cost)
```

```
## # A tibble: 7 x 2
##   group        mean_cost
##   <fct>            <dbl>
## 1 working         19165.
## 2 non-sporting    19316.
## 3 hound           19366.
## 4 toy             19506.
## 5 sporting        20299.
## 6 terrier         20504.
## 7 herding         20692.
```

You should find that herding dogs have the highest average lifetime cost. Working dogs have the lowest.

Part 2

```
count(dogs, size, grooming)
```

```
##      size grooming  n
## 1   large    daily  6
## 2   large   weekly 30
## 3   large     <NA> 18
## 4  medium    daily  8
## 5  medium   weekly 29
## 6  medium  monthly  1
## 7  medium     <NA> 22
## 8   small    daily  9
## 9   small   weekly 29
## 10  small     <NA> 20
```

```
summarize(group_by(dogs, size, grooming), n = n())
```

```
## `summarise()` has grouped output by 'size'. You can override using the `.groups` argument.
```

```
## # A tibble: 10 x 3
## # Groups:   size [3]
##    size   grooming     n
##    <fct>  <fct>    <int>
##  1 large  daily        6
##  2 large  weekly      30
##  3 large  <NA>        18
##  4 medium daily        8
##  5 medium weekly      29
##  6 medium monthly      1
##  7 medium <NA>        22
##  8 small  daily        9
##  9 small  weekly      29
## 10 small  <NA>        20
```