STAT 33A/B Lec Workbook Wk 13

Won Shil Park (3033452021)

Apr 17, 2021

This workbook is due **Apr 19, 2021** by 11:59pm PT for STAT 33A and **Apr 21, 2021** by 11:59pm PT for STAT 33B.

- Knit and submit the generated PDF file on Gradescope.

Exercise 1

How many packages are in the Tidyverse? Explore the website to find out. You can count the tidymodels packages as a single package.

YOUR ANSWER GOES HERE: There are 8 core packages in the tidyverse. In addition, there are another 12 or so packages for a total of about 20 packages.

Exercise 2

1. Read the documentation for the tibble package on the website. What's the name of the function that creates a new tibble from column vectors?
2. Create a tibble with 4 rows and 3 columns. You can make up the data in the columns, but use a different data type for each one.
3. Show how to convert the tibble from step 2 into an ordinary data frame.

Part 1

The tibble() function is used to create a tibble from vectors. Each vector should be the same length and will be used to create a column in the tibble.

Part 2

```
library(tibble)

tb = tibble(x = 1:4, y  = c(TRUE, TRUE, FALSE, TRUE), z = c("a", "bb", "ccc", "dddd"))

class(tb)
```

```
## [1] "tbl_df"     "tbl"        "data.frame"
```

```
df = data.frame(x = 1:4, y  = c(TRUE, TRUE, FALSE, TRUE), z = c("a", "bb", "ccc", "dddd"))

class(df)
```

```
## [1] "data.frame"
```

```
df
```

```
##   x     y    z
## 1 1  TRUE    a
## 2 2  TRUE   bb
## 3 3 FALSE  ccc
## 4 4  TRUE dddd
```

```
tb
```

```
## # A tibble: 4 x 3
##       x y     z
##   <int> <lgl> <chr>
## 1     1 TRUE  a
## 2     2 TRUE  bb
## 3     3 FALSE ccc
## 4     4 TRUE  dddd
```

Part 3

```
tb2 = as.data.frame(tb)

class(tb2)
```

```
## [1] "data.frame"
```

Exercise 3

Use dplyr and the dogs data to compute each of the following subsets:

1. Rows 10-30 only
2. All rows except row 51
3. All columns except `popularity_all` and `popularity`
4. Rows 1-10 with only the `breed`, `weight`, and `height` columns

You do not need to print out these subsets, just show us the code to compute them.

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

Part 1

Here the results are hidden so that the solutions PDF doesn't have pages and pages of raw data printouts.

```
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
```

```
tb1 = slice(dogs, 10:30)
tb1

df1 = dogs[10:30, ]
df1
```

Part 2

```
tb2 = slice(dogs, -51)

df2 = dogs[-51, ]

dim(tb2)

tb2[50:52, 1:4]
dogs[50:52, 1:4]
df2[50:52, 1:4]
```

Part 3

```
tb3 = select(dogs, -popularity_all, -popularity)

tb3a = select(dogs, -"popularity_all", -"popularity")

tb3b = select(dogs, -c(4,5))

df3 = dogs[ , !(names(dogs) %in% c("popularity_all", "popularity"))]

df3a = dogs[ , -c(4, 5)]

names(df3a)
```

Part 4

```
tb4 = slice(select(dogs, breed, height, weight), 1:10)
tb4a = select(slice(dogs, 1:10), breed, height, weight)
df4 = dogs[1:10, c("breed", "height", "weight")]

df4
```

```
##                      breed height weight
## 1            Border Collie   20.0     NA
## 2           Border Terrier     NA   13.5
## 3                 Brittany   19.0   35.0
## 4            Cairn Terrier   10.0   14.0
## 5   Welsh Springer Spaniel   18.0     NA
## 6   English Cocker Spaniel   16.0   30.0
## 7           Cocker Spaniel   14.5   25.0
## 8                 Papillon    9.5     NA
## 9    Australian Cattle Dog   18.5     NA
## 10       Shetland Sheepdog   14.5   22.0
```

Exercise 4

Use dplyr to show that there are no duplicated rows in the dogs data.

Explain your reasoning.

```
dim(distinct(dogs))
```

```
## [1] 172  18
```

```
dim(dogs)
```

```
## [1] 172  18
```

```
dim(dogs) == dim(distinct(dogs))
```

```
## [1] TRUE TRUE
```

Exercise 5

Use dplyr to determine for each `group` of dog, what's the shortest lifespan? You should have one result per group here.

Additionally, for each `group` of dog, what's the longest lifespan?

Here are the shortest lifespans for each group of dog:

```
groups = group_by(dogs, group)
summarize(groups, short.life = min(longevity, na.rm = TRUE))
```

```
## # A tibble: 7 x 2
##   group        short.life
##   <fct>             <dbl>
## 1 herding            7.33
## 2 hound              6.75
## 3 non-sporting       6.29
## 4 sporting           6.5
## 5 terrier            6.6
## 6 toy                9.25
## 7 working            6.5
```

Here are the longest lifespans for each group of dog:

```
summarize(groups, long.life = max(longevity, na.rm = TRUE),
          avg.wt = mean(weight, na.rm = TRUE), ct = n(), first.dog = first(breed))
```

```
## # A tibble: 7 x 5
##   group        long.life avg.wt    ct first.dog
##   <fct>            <dbl>  <dbl> <int> <chr>
## 1 herding           14.7  36.7     25 Border Collie
## 2 hound             13.6  63.8     26 Dachshund
## 3 non-sporting      14.4  27.9     19 Lhasa Apso
## 4 sporting          12.9  52.0     28 Brittany
## 5 terrier           14    23.4     28 Border Terrier
## 6 toy               16.5   9.82    19 Papillon
## 7 working           12.6 105       27 Siberian Husky
```

An alternative approach to subsetting

```
tb5 = filter(dogs, group == "toy")

df4 = dogs[dogs$group == "toy", ]

dim(tb5)
```

```
## [1] 19 18
```

```
dim(df4)
```

```
## [1] 19 18
```