

Data manipulation practice

Experiential Data science for Undergraduate Cross-disciplinary Education

Dr. Kim Dill-McFarland, U. of British Columbia

Contents

1 Data manipulation practice	1
1.1 Learning objectives	1
1.2 Setup	1
1.2.1 Install and load <code>tidyverse</code>	1
1.2.2 Load and clean data	2
1.3 Practice	2
1.3.1 <code>select</code> and <code>filter</code>	2
1.3.2 <code>rename</code> and <code>mutate</code>	3
1.3.3 <code>summarize</code> and pipes	3
1.3.4 <code>*_join</code>	3

1 Data manipulation practice

These problems are designed to help you practice concepts and functions covered in the ‘Data manipulation tutorial’.

1.1 Learning objectives

- Load tabular data using the `tidyverse`
- Subset and clean data in `dplyr` (`filter`, `select`, `rename`, `arrange`, `mutate`)
- Summarize data in `dplyr` (`group_by`, `summarize`)
- Transform data frames using `tidyr` (`gather`, `spread`) and `dplyr` (`*_join`)
- Link multiple tidyverse functions using pipes `%>%`

1.2 Setup

Open a new RStudio session, create a Project, and start a new R script to save your responses.

1.2.1 Install and load `tidyverse`

If you have not done so already, install the tidyverse packages. (More information available in our ‘RStudio tutorial’.)

```
install.packages("tidyverse")
```

*Please note that if you have **R v3.3 or older**, you may not be able to install `tidyverse`. In this case, you need to separately install each package within the tidyverse.*

For this practice, you will need to load:

R v3.4 or newer

```

library(tidyverse)

R v3.3 or older

library(readr)
library(dplyr)

## Warning: package 'dplyr' was built under R version 3.5.2
library(tidyr)

## Warning: package 'tidyr' was built under R version 3.5.2

```

1.2.2 Load and clean data

Load the pre-cleaned data from our GitHub.

```

dat <- read_csv(
  "https://raw.githubusercontent.com/EDUCE-UBC/workshop_data/master/Saanich_Data_clean.csv")

## Parsed with column specification:
## cols(
##   Cruise = col_double(),
##   Date = col_date(format = ""),
##   Depth = col_double(),
##   O2_uM = col_double(),
##   NO3_uM = col_double(),
##   H2S_uM = col_double(),
##   Depth_m = col_double()
## )

```

If you would like to learn more about Saanich Inlet and these data, checkout [our description](#).

1.3 Practice

In order to prevent conflicts, please copy the data to pdat (practice data) prior to attempting **each exercise**.

```
pdat <- dat
```

NOTE: Variable names containing O2 are O as in Oxygen, not 0 as in zero.

1.3.1 select and filter

1. Subset data based on the following criteria
 - select the Cruise, Date, Depth_m, and O2_uM variables in pdat
 - filter these data to retain observations from Cruise 72 where Depth is greater than or equal to 50 m
 - *Your resulting pdat object should be a [13, 4] data frame.*
2. Subset data based on the following criteria
 - select Cruise and Depth_m variables in pdat
 - filter these data retain observations where oxygen OR nitrate is greater than zero.
 - *Hint:* Can you filter based on a variable that you previously removed by not selecting it?
 - *Your resulting pdat object should be a [1138, 2] data frame.*

1.3.2 rename and mutate

3. Calculate the means of Depth_cm and Oxygen
 - mutate the Depth variable (currently in km) to centimeters in pdat
 - rename the O2_uM variable to ‘Oxygen’
 - calculate the means of Depth_cm and Oxygen
 - *Your means should be 102837.5 and 68.78413, respectively.*

1.3.3 summarize and pipes

4. Calculate median nitrate at 10, 100, and 200 meters
 - Filter the data to depths equal to 10, 100, or 200 meters in pdat
 - Calculate median nitrate (NO3_uM) by depth
 - Hint: Use the help ? function to find out how to deal with NAs when calculating a medians.
 - *Your medians should be 15.6, 25.5, and 0, respectively.*

1.3.4 *_join

5. Using the separated oxygen and nitrate data created here,

```
dat_O2 <- pdat %>%
  select(Cruise, Date, Depth_m, O2_uM) %>%
  arrange(O2_uM) %>%
  filter(O2_uM != 0)

dat_NO3 <- pdat %>%
  select(Cruise, Date, Depth_m, NO3_uM) %>%
  arrange(NO3_uM) %>%
  filter(NO3_uM != 0)
```

- Combine all the data, keeping only rows found in dat_NO3
 - There are 2 possible joining functions to complete this task. Try to write code for both!
 - *Your resulting pdat object should be a [989, 5] data frame.*
-