Week 5 - Class Worksheet

Tidy and Manipulate: Part II - Manipulate

Dr. Anil Dolgun

Last updated: 18 June, 2020

Required Packages

The following packages will be required or may come in handy.

```
library(tidyr)
library(dplyr)
library(readr)
library(readxl)
library(knitr)
```

Exercises

Family Incidents data set

The following exercise is based on family incidents family_incidents.csv (../data/family_incidents.csv) data set. This data set is taken from Crime Statistics Agency https://www.crimestatistics.vic.gov.au/family-violence-data-portal/download-data-tables (https://www.crimestatistics.vic.gov.au/family-violence-data-portal/download-data-tables) containing 84 observation from Local government areas about family incident rate per 100,000 population by police region from 2012 to 2017.

The variables for family_incidents.csv (../data/family_incidents.csv) data set consists of:

`Police Region` [Character]: Police regions (levels: North West Metro, Eastern, Southern Metro, Western and where Total2 is for unknown geographical location).

`Local Government Area` [Character]: 79 Local government areas of Victoria there are totals per region.

`2012-2013, ..., 2016-2017` : The year that data was collected.

---

1 From `family_incidents` data set, select only 2015-2016 and 2016-2017. Then group by police region and show a summary of mean and standard deviation of 2015-2016 and 2016-2017.

Influenza data set

The following exercises (exercise 2-4) are based on Influenza data set Influenza.xlsx (../data/Influenza.xlsx) taken from Department of Health, located at http://www9.health.gov.au/cda/source/pub_influ.cfm (http://www9.health.gov.au/cda/source/pub_influ.cfm).

This data set has 335,544 observations of infection with influenza viruses around Australia between 2008 to 2016 containing variables:

`Week Ending (Friday)` [POSIXct]: Represents the date the data of the diagnosis with time zone.

`State` [Character]: The state or territory of residence of the notified case.

`Age Group` [Character]: Age category with 19 levels.

`Sex` [Character]: Male, Female, X, Unknown

`Indigenous Status` [Character]: Indigenous, non-Indigenous, not available, unknown

`Type/Subtype` [Character]: Type/Subtype of the influenza virus.

For more information please check the Data Caveats sheet of the data set.

2

Using influenza data set, use `filter()`, `group_by()` and `summarize()` to find the counts of Indigenous people, group them with Age group, Sex and State then use `arrange()` in ascending order. Don't forget to use `factor()` with `levels()`, `labels` and `ordered` arguments where appropriate for Age Group, State variables.

3

The table down below is created from influenza data which is filtered out to show only Indigenous people. Create the same table using `filter()`, `mutate()`, `group_by()`, `summarize()` functions from `dplyr` package and `spread()` function from `tidyr` package, name it df1. The challenge is creating a new variable year using `mutate()`: You need to use `substr()` function from base package to select the first 4 digits of the `Week Ending (Friday)` column (for year). You can always go creative and find another way to create the same table as an exercise! (Hint: To select the first 4 digits of a string you can use `substr(x, 1,4)`).

year	NT	QLD	WA
2008	105	72	58
2009	1209	2365	632
2010	258	266	71
2011	439	395	175
2012	199	586	422
2013	271	201	123
2014	496	824	464
2015	274	885	311
2016	249	987	415
2017	1	1	NA

4 Repeat exercise 3, this time don't use any filter, name it `df2`, then join `df1` and `df2` using `suffix` argument to differentiate the columns of `df1`.

SA & Victorian pet ownership data sets

The following exercises (exercise 5-9) are based on subsets of the pet ownership data which are Registrations_Master_Vic.csv (../data/Registrations_Master_Vic.csv), VIC_pet.csv (../data/VIC_pet.csv), SA_pet.csv (../data/SA_pet.csv), pet1.csv (../data/pet1.csv), pet2.csv (../data/pet2.csv) and pet3.csv (../data/pet3.csv). These data sets are taken from Kaggle, located at https://www.kaggle.com/puppygogo/sa-dog-ownership-sample/data (https://www.kaggle.com/puppygogo/sa-dog-ownership-sample/data). Variables are self explanatory.

5 Use `bind_rows()` and `union()` to bind `vic_pet` and `sa_pet` data sets. Compare these two data sets you binded with `intersect()` and use `setdiff()` to prove that two data sets has no difference.

6 Read in `pet1`, `pet2` and `pet3` data sets. First apply a `left_join()` to `pet1` and `pet2`, then join this new data set with `pet3` using `left_join`. Repeat the same action, this time use `right_join()`, name it `pet_join`. Explain shortly why the results are different. Then use `setdiff()` and/or `anti_join()` to find out the different records in the data sets.

7 Use a suitable join function to join `pet2` and `pet3` data sets, only keep the rows that exists in the both data sets.

8 Use a suitable join function to join `pet2` and `pet3` data sets, only keep the rows that exists in `pet2` data set.

9 Bonus exercise: Use Registrations_Master_Vic.csv (../data/Registrations_Master_Vic.csv) data set to create a meaningful data set to tell a story using at least 4 of `select()`, `mutate()`, `filter()`, `arrange()`, `summarize()` and `group_by()`. Share your own story with your code on the discussion board. Best solution(s) will be immortalised as example solutions in this worksheet.

Finished?

If you have finished the above tasks, work through the weekly list of tasks posted on the Canvas announcement page.

Return to Course Website (../index.html)