# Week 4 - Class Worksheet

## Tidy and Manipulate: Part I - Tidy

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 exercises (exercise 1-4) are 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.

**First six observations of the family_incidents data set:**

| Police region | Local government area | 2012-13 | 2013-14 | 2014-15 | 2015-16 | 2016-17 |
|---|---|---|---|---|---|---|
| North West Metro | Banyule | 940.9 | 974.1 | 1060.8 | 1030.6 | 1014.3 |
| North West Metro | Brimbank | 986.8 | 1084.3 | 1240.5 | 1343.3 | 1105.0 |
| North West Metro | Darebin | 1020.8 | 1131.3 | 1063.6 | 1071.9 | 1052.4 |
| North West Metro | Hobsons Bay | 899.5 | 1065.7 | 1137.0 | 1205.7 | 1075.9 |
| North West Metro | Hume | 1390.7 | 1540.3 | 1554.9 | 1538.0 | 1478.6 |

| Police region | Local government area | 2012-13 | 2013-14 | 2014-15 | 2015-16 | 2016-17 |
|---|---|---|---|---|---|---|
| North West Metro | Maribyrnong | 956.8 | 1042.4 | 997.6 | 1089.4 | 891.8 |

**1** Read in the family_incidents.csv (../data/family_incidents.csv) data set using a suitable function. Check out the classes of each variable and convert Police Region column into factor, use `ordered` argument then check its' levels.

**2** Choose the suitable action for `family_incident` data set.

a) We need to combine multiple columns into a single column using `unite()` function

b) We need to gather those columns into a new pair of variables using `gather()` function

c) We need to split the variables since multiple variables are stored in one column using `separate()` function

d) We need to transform data from long format to wide format using `spread()` function

**3** Tidy the `family_incidents` data set into the form below.

| Police region | Local government area | year | cases |
|---|---|---|---|
| North West Metro | Banyule | 2012-13 | 940.9 |
| North West Metro | Brimbank | 2012-13 | 986.8 |
| North West Metro | Darebin | 2012-13 | 1020.8 |
| North West Metro | Hobsons Bay | 2012-13 | 899.5 |

**4** Using `family_incidents` data set, change the separator of the `year` column from `"-"` to `"/"` using `separate()` and `unite()` functions.

# Primary Health Network data set

The following exercises (exercise 5-8) are based on the PHN data set (PHN.xlsx (../data/PHN2017.xlsx)). This data set has the percentage of children who are immunised against Polio and Hepatitis B in Australia by age group and location in October 2017 - September 2017. `PHN` data set is taken from https://beta.health.gov.au/resources/publications/2017-phn-childhood-immunisation-coverage-data (https://beta.health.gov.au/resources/publications/2017-phn-childhood-immunisation-coverage-data) with 93 observations, containing variables:

`PHN Number` [Character]: ID of the Primary Health Network Area

`PHN Name` [Character]: Name of the Primary Health Network Area

`Age Group` [Character]: Factor with Levels 12-<15 Months, 24-<27 Months, 60-<63 Months

`%Polio` [Numeric]: Percentage of children immunised against Polio

`%HEP` [Numeric]: Percentage of children immunised against Hepatitis B

**First six observations of the PHN data set:**

| PHN Number | PHN Name | Age Group | %Polio | %HEP |
|---|---|---|---:|---:|
| PHN101 | Central and Eastern Sydney | 12-<15 Months | 94.20931 | 94.09876 |
| PHN102 | Northern Sydney | 12-<15 Months | 94.75332 | 94.58254 |
| PHN103 | Western Sydney | 12-<15 Months | 94.07182 | 93.99051 |
| PHN104 | Nepean Blue Mountains | 12-<15 Months | 94.94118 | 94.94118 |
| PHN105 | South Western Sydney | 12-<15 Months | 93.85654 | 93.92611 |
| PHN106 | South Eastern NSW | 12-<15 Months | 95.44135 | 95.54297 |

**5**

Read in the PHN.xlsx (../data/PHN2017.xlsx) data set using a suitable function. Check out the classes of each variable and convert Age Group column into factor, use `ordered` argument then check its levels.

**6**

Assume that we want to calculate the mean percentage as a column of all age groups of children immunised against Polio and Hepatitis B. Before calculating the mean, select the tidy functions that we need in order to form the data set as:

| PHN Number | PHN Name | Vaccination Type | 12-<15 Months | 24-<27 Months | 60-<63 Months |
|---|---|---|---:|---:|---:|
| PHN101 | Central and Eastern Sydney | %HEP | 94.09876 | 95.37775 | 0.00000 |
| PHN101 | Central and Eastern Sydney | %Polio | 94.20931 | 95.54502 | 92.50768 |
| PHN102 | Northern Sydney | %HEP | 94.58254 | 95.11925 | 0.00000 |
| PHN102 | Northern Sydney | %Polio | 94.75332 | 95.56295 | 92.16040 |
| PHN103 | Western Sydney | %HEP | 93.99051 | 95.89433 | 0.00000 |
| PHN103 | Western Sydney | %Polio | 94.07182 | 95.99414 | 94.36424 |

a) `unite()` , `gather()`

b) `split()` , `unite()`

c) `gather()` , `spread()`

d) `separate()` , `split()`

**7** Tidy the `PHN` data set into the form in exercise 6. Once you tidy, use `rowMeans` function from baseR package to calculate the mean percentage of children immunised and save this average as a vector. (Hint: Remember how to select columns as a sequence i.e., `df[,3:5]` .) Repeat this calculation by combining your tidy code and calculating the mean in one line whilst saving it as a vector.

**8** Bonus exercise: Use `round` function from base R package to round the values to 2 decimals places in PHN data set.

# 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)**