**P8105\_hw2\_yc384**

Ying Chen (UNI: yc384) 9/24/2019

setwd(“/Users/macbook/Documents/Statistics/P8105/HW/P8105 HW2”)

knitr::opts\_chunk$set(echo = TRUE)

library (tidyverse)

## -- Attaching packages ------------------------------------------------------------------ tidyverse 1.2.1 --

## v ggplot2 3.2.1 v purrr 0.3.2

## v tibble 2.1.3 v dplyr 0.8.3

## v tidyr 1.0.0 v stringr 1.4.0

## v readr 1.3.1 v forcats 0.4.0

## -- Conflicts --------------------------------------------------------------------- tidyverse\_conflicts() --

## x dplyr::filter() masks stats::filter()

## x dplyr::lag() masks stats::lag()

library (dplyr)

library(readxl)

rm(list = ls())

options(tibble.print\_min = 3)

**P8105 DSI Homework 1**

**This assignment reinforces ideas in Data Wrangling I**

**1. Problem 0**

* Github repo: <https://github.com/YingCarolineChen/p8105_hw2_yc384.git>
* RMarkdown file name: P8105\_hw2\_yc384
* Create a subdirectory to store local data files and the path is: “~/Documents/Statistics/P8105/HW/P8105 HW2”

**2. Problem 1**

* 1-1 Read and tidy Mr. Trash Wheel data
* Dataset: Mr. Trash Wheel / MS Excel file / Contains 8 sheets

# read in Mr. Trash Wheel sheet

# skip first rows with notes / figures

# drop last column that containing notes

MrTrashWheel =

read\_excel("./data/HealthyHarborWaterWheelTotals2018-7-28.xlsx", sheet = "Mr. Trash Wheel", skip = 1, range = "A2:N338", col\_names = TRUE) %>%

janitor::clean\_names()

# omit rows has NA for dumpster

dumpster =

drop\_na(MrTrashWheel,dumpster) %>%

# Round sports balls to nearest integer

mutate(

sports\_balls\_round = round(sports\_balls, digits = 0),

sports\_balls\_int = as.integer(sports\_balls\_round)

)

# rename to reasonable variable names

dumpster\_rename = rename(dumpster, bags\_grocery = grocery\_bags, bags\_chip = chip\_bags, bottles\_plastic = plastic\_bottles, bottles\_glass = glass\_bottles)

# Check data

head(MrTrashWheel,5)

## # A tibble: 5 x 14

## dumpster month year date weight\_tons volume\_cubic\_ya~

## <dbl> <chr> <dbl> <dttm> <dbl> <dbl>

## 1 1 May 2014 2014-05-16 00:00:00 4.31 18

## 2 2 May 2014 2014-05-16 00:00:00 2.74 13

## 3 3 May 2014 2014-05-16 00:00:00 3.45 15

## 4 4 May 2014 2014-05-17 00:00:00 3.1 15

## 5 5 May 2014 2014-05-17 00:00:00 4.06 18

## # ... with 8 more variables: plastic\_bottles <dbl>, polystyrene <dbl>,

## # cigarette\_butts <dbl>, glass\_bottles <dbl>, grocery\_bags <dbl>,

## # chip\_bags <dbl>, sports\_balls <dbl>, homes\_powered <dbl>

head(dumpster\_rename, 5)

## # A tibble: 5 x 16

## dumpster month year date weight\_tons volume\_cubic\_ya~

## <dbl> <chr> <dbl> <dttm> <dbl> <dbl>

## 1 1 May 2014 2014-05-16 00:00:00 4.31 18

## 2 2 May 2014 2014-05-16 00:00:00 2.74 13

## 3 3 May 2014 2014-05-16 00:00:00 3.45 15

## 4 4 May 2014 2014-05-17 00:00:00 3.1 15

## 5 5 May 2014 2014-05-17 00:00:00 4.06 18

## # ... with 10 more variables: bottles\_plastic <dbl>, polystyrene <dbl>,

## # cigarette\_butts <dbl>, bottles\_glass <dbl>, bags\_grocery <dbl>,

## # bags\_chip <dbl>, sports\_balls <dbl>, homes\_powered <dbl>,

## # sports\_balls\_round <dbl>, sports\_balls\_int <int>

**2. Problem 1**

* 1-2 Read and tidy precipitation data
* join datasets and some data manipulations

# read in precipitation data for 2017 and 2018

# Omit rows without precipitation data and add a variable year

precipitation\_2017 =

read\_excel("./data/HealthyHarborWaterWheelTotals2018-7-28.xlsx", sheet = "2017 Precipitation", skip = 1, range = "A2:B14", col\_names = TRUE) %>%

janitor::clean\_names() %>%

mutate(

year = 2017) %>%

drop\_na (total)

precipitation\_2018 =

read\_excel("./data/HealthyHarborWaterWheelTotals2018-7-28.xlsx", sheet = "2018 Precipitation", skip = 1, range = "A2:B14", col\_names = TRUE) %>%

janitor::clean\_names() %>%

mutate(

year = 2018) %>%

drop\_na (total)

# Combine precipitation datasets

precipitation\_join =

full\_join(precipitation\_2017, precipitation\_2018, by = NULL) %>%

# convert month to a character variable

mutate(

month = month.name[month]

)

## Joining, by = c("month", "total", "year")

head(precipitation\_join,5)

## # A tibble: 5 x 3

## month total year

## <chr> <dbl> <dbl>

## 1 January 2.34 2017

## 2 February 1.46 2017

## 3 March 3.57 2017

## 4 April 3.99 2017

## 5 May 5.64 2017

**2. Problem 1**

* 1-3 Datasets description using inline R code

dump2017 = filter(dumpster, year == 2017)

Dataset Mr. Trash Wheel contains 336 obersvations and 14 variables. The key variable for this dataset is called “dumpster”. Variable “dumpster” has 51 rows with “NA” and will be excluded for future analyses. The median number of sports balls in a dumpster in 2017 is: 8.

Dataset precipitation\_2017 contains 12 obersvations and 3 variables. The key variables for this dataset is year and month. Dataset precipitation\_2018 contains 7 obersvations and 3 variables. The key variables for this dataset are year and month. Dataset precipitation\_join is the combination of datasets precipitation\_2017 and precipitation\_2018. The total precipitation is 2018 is 23.5in.

**3. Problem 2**

* Read pols-month.csv, unemployment.csv, and snp.csv and merge them using year and month as keys
* 2-1 This code cunk will read and tidy pols-month.csv.
* break up the variable mon into integer variables year, month, and day;
* replace month number with month name;
* create a president variable taking values gop and dem, remove prez\_dem, prez\_gop & the day variable.

**2\_1 Pols**

# read in pols-month.csv

pols\_tidy =

read\_csv("./data/pols-month.csv", col\_names = TRUE) %>%

mutate (

# change date format

Date = as.Date(mon, format='%m/%d/%Y')

) %>%

janitor::clean\_names() %>%

# break date to three vars

separate(., "date", c("year", "month", "day")) %>%

mutate (

month = as.numeric(month),

#replace month number with month name

month = month.name[c(month)],

president = prez\_gop + prez\_dem

) %>%

# reorder varaibles and remove three variables

select(year, month, everything(), -prez\_dem, -prez\_gop, -day)

## Parsed with column specification:

## cols(

## mon = col\_date(format = ""),

## prez\_gop = col\_double(),

## gov\_gop = col\_double(),

## sen\_gop = col\_double(),

## rep\_gop = col\_double(),

## prez\_dem = col\_double(),

## gov\_dem = col\_double(),

## sen\_dem = col\_double(),

## rep\_dem = col\_double()

## )

**2\_2 snp**

# read in snp.csv

snp\_tidy =

read\_csv("./data/snp.csv", col\_names = TRUE) %>%

mutate (

# change date format

Date1 = as.Date(date, format='%m/%d/%Y')

) %>%

janitor::clean\_names() %>%

# break date to three vars

separate(., "date1", c("year", "month", "day")) %>%

mutate (

month = as.numeric(month),

#replace month number with month name

month = month.name[c(month)],

) %>%

# reorder varaibles and remove three variables

select(year, month, everything(), -day)

## Parsed with column specification:

## cols(

## date = col\_character(),

## close = col\_double()

## )

**2\_3 unemployment**

# read in unemployment.csv

unemployment =

read\_csv("./data/unemployment.csv", col\_names = TRUE) %>%

janitor::clean\_names()

## Parsed with column specification:

## cols(

## Year = col\_double(),

## Jan = col\_double(),

## Feb = col\_double(),

## Mar = col\_double(),

## Apr = col\_double(),

## May = col\_double(),

## Jun = col\_double(),

## Jul = col\_double(),

## Aug = col\_double(),

## Sep = col\_double(),

## Oct = col\_double(),

## Nov = col\_double(),

## Dec = col\_double()

## )

unemployment

## # A tibble: 68 x 13

## year jan feb mar apr may jun jul aug sep oct nov

## <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>

## 1 1948 3.4 3.8 4 3.9 3.5 3.6 3.6 3.9 3.8 3.7 3.8

## 2 1949 4.3 4.7 5 5.3 6.1 6.2 6.7 6.8 6.6 7.9 6.4

## 3 1950 6.5 6.4 6.3 5.8 5.5 5.4 5 4.5 4.4 4.2 4.2

## # ... with 65 more rows, and 1 more variable: dec <dbl>

unemployment\_tidy =

pivot\_longer(

unemployment,

jan:dec,

names\_to = "month",

values\_to = "unemploy\_rate"

)

unemployment\_tidy

## # A tibble: 816 x 3

## year month unemploy\_rate

## <dbl> <chr> <dbl>

## 1 1948 jan 3.4

## 2 1948 feb 3.8

## 3 1948 mar 4

## # ... with 813 more rows