Lab Session 1: Introduction to R

Isabel Fulcher

3/2/2021

Set working directory

We will first set a working directory. Your working directory should be where your plan to save your R code and also where the example datasets have been stored. In the "Files" tab, navigate to this folder. Once you are there, select the "More" dropdown and select "Set As Working Directory".

Install and load R packages

You should have already installed the tidyverse package. Now, you will need to load the package into R. This will allow you to use the functionality of the package.

```
library(tidyverse)
library(lubridate)
```

Load and view data

Load the outpatient visits ".rds" file and save it as a data frame called "outpatient". An .rds file is an R object and is the best way to store data that will be used in R.

```
outpatient <- readRDS("session1_data/example_outpatient.rds")
```

View the first six observations in the outaptient datasets.

```
head(outpatient)
```

```
## # A tibble: 6 x 2
##   date       outpatient_visits
##   <date>                 <dbl>
## 1 2016-01-01              4983
## 2 2016-02-01              5331
## 3 2016-03-01              6267
## 4 2016-04-01              6063
## 5 2016-05-01              5775
## 6 2016-06-01              4397
```

View the variable types in the dataset.

```
str(outpatient)
```

```
## Classes 'tbl_df', 'tbl' and 'data.frame':    60 obs. of  2 variables:
##  $ date             : Date, format: "2016-01-01" "2016-02-01" ...
##  $ outpatient_visits: num  4983 5331 6267 6063 5775 ...
```

Load a .csv file in R and view the first six observations. Does anything look different about this data frame?

```
data_csv <- read_csv("session1_data/example_outpatient.csv")
head(data_csv)
```

```
## # A tibble: 6 x 2
##   date   outpatient_visits
##   <chr>              <dbl>
## 1 1/1/16              4983
```

```
## 2 2/1/16              5331
## 3 3/1/16              6267
## 4 4/1/16              6063
## 5 5/1/16              5775
## 6 6/1/16              4397
```

Working with data

Filtering data

Filter the data to months that had greater than or equal to 8000 outpatient visits.

```
outpatient %>%
  filter(outpatient_visits >= 8000)
```

```
## # A tibble: 2 x 2
##   date       outpatient_visits
##   <date>                 <dbl>
## 1 2017-02-01              8091
## 2 2017-03-01              8326
```

Filter the data to view outpatient visits in 2020.

```
outpatient %>%
  filter(date >= "2020-01-01")
```

```
## # A tibble: 12 x 2
##    date       outpatient_visits
##    <date>                 <dbl>
##  1 2020-01-01              5030
##  2 2020-02-01              5908
##  3 2020-03-01              6388
##  4 2020-04-01              4389
##  5 2020-05-01              2559
##  6 2020-06-01              2865
##  7 2020-07-01              2970
##  8 2020-08-01              3144
##  9 2020-09-01              3588
## 10 2020-10-01              3385
## 11 2020-11-01              3314
## 12 2020-12-01              2582
```

Data summaries

What is the total number of outpatient visits in the dataset?

```
outpatient %>%
  summarize(sum(outpatient_visits))
```

```
## # A tibble: 1 x 1
##   `sum(outpatient_visits)`
##                      <dbl>
## 1                   280861
```

What is the mean (average) number of monthly outpatient visits?

```
outpatient %>%
  summarize(mean(outpatient_visits))
```

```
## # A tibble: 1 x 1
##   `mean(outpatient_visits)`
##                       <dbl>
## 1                     4681.
```

What is the minimum and maximum number of monthly outpatient visits?

```
outpatient %>%
  summarize(min(outpatient_visits),
            max(outpatient_visits))
```

```
## # A tibble: 1 x 2
##   `min(outpatient_visits)` `max(outpatient_visits)`
##                      <dbl>                    <dbl>
## 1                     2559                     8326
```

How many missing months are there?

```
outpatient %>%
  summarize(sum(is.na(outpatient_visits)))
```

```
## # A tibble: 1 x 1
##   `sum(is.na(outpatient_visits))`
##                             <int>
## 1                               0
```

Creating a new variable

Create a new variable that indicates the month.

```
outpatient %>%
  mutate(month = month(date))
```

```
## # A tibble: 60 x 3
##    date       outpatient_visits month
##    <date>                 <dbl> <dbl>
##  1 2016-01-01              4983     1
##  2 2016-02-01              5331     2
##  3 2016-03-01              6267     3
##  4 2016-04-01              6063     4
##  5 2016-05-01              5775     5
##  6 2016-06-01              4397     6
##  7 2016-07-01              4176     7
##  8 2016-08-01              5303     8
##  9 2016-09-01              5862     9
## 10 2016-10-01              5836    10
## # ... with 50 more rows
```

Now, save the above as a new data frame (also known as a "tibble" in R)

```
outpatient %>%
  group_by(month = month(date)) -> outpatient2

head(outpatient2)
```

```
## # A tibble: 6 x 3
## # Groups:   month [6]
##   date       outpatient_visits month
##   <date>                 <dbl> <dbl>
```

```
## 1 2016-01-01              4983     1
## 2 2016-02-01              5331     2
## 3 2016-03-01              6267     3
## 4 2016-04-01              6063     4
## 5 2016-05-01              5775     5
## 6 2016-06-01              4397     6
```

Summarize data within groupings

Over the time period, what is the average number of visits by month?

```
outpatient2 %>%
  group_by(month) %>%
  summarize(mean(outpatient_visits))
```

```
## # A tibble: 12 x 2
##    month `mean(outpatient_visits)`
##    <dbl>                     <dbl>
##  1     1                     5531.
##  2     2                     5926
##  3     3                     6443.
##  4     4                     5337.
##  5     5                     5183.
##  6     6                     4023.
##  7     7                     3746.
##  8     8                     4111.
##  9     9                     4482.
## 10    10                     4453
## 11    11                     3528
## 12    12                     3410.
```

Lab activity

Please take 15 minutes to complete the following questions.

1. How many months have less than 3000 visits?

```
outpatient %>%
  filter(outpatient_visits > 3000)
```

```
## # A tibble: 54 x 2
##    date       outpatient_visits
##    <date>                 <dbl>
##  1 2016-01-01              4983
##  2 2016-02-01              5331
##  3 2016-03-01              6267
##  4 2016-04-01              6063
##  5 2016-05-01              5775
##  6 2016-06-01              4397
##  7 2016-07-01              4176
##  8 2016-08-01              5303
##  9 2016-09-01              5862
## 10 2016-10-01              5836
## # ... with 44 more rows
```

2. How many outpatient visits occurred in March 2019?

```
outpatient %>%
  filter(date == "2019-03-01")
```

```
## # A tibble: 1 x 2
##   date       outpatient_visits
##   <date>                 <dbl>
## 1 2019-03-01              5392
```

3. How many total outpatient visits occurred in 2016?

```
outpatient %>%
  filter(date >= "2016-01-01" & date <= "2016-12-01") %>%
  summarize(sum(outpatient_visits))
```

```
## # A tibble: 1 x 1
##   `sum(outpatient_visits)`
##                      <dbl>
## 1                    63057
```

4. Load in the example_kmc.rds dataset

```
kmc <- readRDS("session1_data/example_kmc.rds")
```

5. View the KMC dataset.

```
head(kmc)
```

```
## # A tibble: 6 x 2
##   date       indicator_count_kangaroo
##   <date>                        <dbl>
## 1 2016-01-01                       22
## 2 2016-02-01                       11
## 3 2016-03-01                        9
## 4 2016-04-01                        8
## 5 2016-05-01                       NA
## 6 2016-06-01                        6
```

6. How many missing months are in this dataset?

```
kmc %>%
  summarize(sum(is.na(indicator_count_kangaroo)))
```

```
## # A tibble: 1 x 1
##   `sum(is.na(indicator_count_kangaroo))`
##                                    <int>
## 1                                     23
```

7. What is the mean (average) number of monthly Kangaroo Mother Care?

```
kmc %>%
  summarize(mean(indicator_count_kangaroo,na.rm=TRUE))
```

```
## # A tibble: 1 x 1
##   `mean(indicator_count_kangaroo, na.rm = TRUE)`
##                                            <dbl>
## 1                                           8.62
```