Session #3 - Exercises

Adrian C Lo

2/5/2020

INSTRUCTIONS

There are 3 segments with around 10 questions each that increase in difficulty. Fill in the answer within the code chunk. When you wish to test the code chunk, press the *green play button* on the right side of the code chunk to see your output.

The hints will guide you what functions are required. You can access further information with *?x* where x is the function name, e.g. ?print().

Also check this page, specifically sections 8.3 and 10.2 for additional help.

basic operations {base}

- bla
- bla

data import {readr}

There are several datasets available that we will import.

```
suppressPackageStartupMessages( library(readr) )
```

1.

data manipulations {dplyr}

In this section, 2 different datasets will be explored: coronavirus and babynames

```
suppressPackageStartupMessages( library(dplyr) )
suppressPackageStartupMessages( library(tidyr) )
```

In the following questions, you will perform exploratory analysis on the coronavirus. This dataset is contained in the like-named {coronavirus} library.

```
suppressPackageStartupMessages( library(coronavirus) )
```

1. Worldwide, how many confirmed cases of coronavirus have been found?

```
coronavirus %>%
    filter(type == "confirmed") %>%
    summarise(worldwide_confirmed = sum(cases))
```

```
## # A tibble: 1 x 1
##   worldwide_confirmed
##                 <int>
## 1               88371
```

2. Worldwide, how many people died from coronavirus?

```
coronavirus %>%
    filter(type == "death") %>%
    summarise(worldwide_confirmed = sum(cases))
```

```
## # A tibble: 1 x 1
##   worldwide_confirmed
##                 <int>
## 1                2996
```

3. Which are the top 5 countries with the most cases of confirmed coronavirus?

```
coronavirus %>%
    filter(type == "confirmed") %>%
    group_by(Country.Region) %>%
    summarise(confirmed = sum(cases)) %>%
    arrange(desc(confirmed)) %>%
    head(5)
```

```
## # A tibble: 5 x 2
##   Country.Region confirmed
##   <chr>              <int>
## 1 Mainland China     79826
## 2 South Korea         3736
## 3 Italy               1694
## 4 Iran                 978
## 5 Others               705
```

4. From which country is the last confirmed case?

```
coronavirus %>%
    filter(type == "confirmed") %>%
    arrange(desc(date)) %>%
    head(1)
```

```
## # A tibble: 1 x 7
##   Province.State Country.Region   Lat  Long date       cases type
##   <chr>          <chr>          <dbl> <dbl> <date>     <int> <chr>
## 1 ""             Armenia         40.1  45.0 2020-03-01     1 confirmed
```

5. From which country were the latest recovered cases?

```
coronavirus %>%
    filter(type == "recovered") %>%
    arrange(desc(date)) %>%
    head(1)
```

```
## # A tibble: 1 x 7
##   Province.State Country.Region   Lat  Long date       cases type
##   <chr>          <chr>          <dbl> <dbl> <date>     <int> <chr>
## 1 ""             Iran              32    53 2020-03-01    52 recovered
```

6. When and where were the most confirmed cases detected on a single day?

```
coronavirus %>%
    filter(type == "confirmed") %>%
    arrange(desc(cases)) %>%
    head(1)
```

```
## # A tibble: 1 x 7
```

```
##   Province.State Country.Region   Lat  Long date       cases type
##   <chr>          <chr>          <dbl> <dbl> <date>     <int> <chr>
## 1 Hubei          Mainland China  31.0  112. 2020-02-13 14840 confirmed
```

7. Were there any false positive confirmed cases?

```
coronavirus %>%
    filter(type == "confirmed") %>%
    filter(cases < 0)
```

```
## # A tibble: 8 x 7
##   Province.State        Country.Region   Lat   Long date       cases type
##   <chr>                 <chr>          <dbl>  <dbl> <date>     <int> <chr>
## 1 ""                    Japan           36    138   2020-01-23    -1 confi~
## 2 Queensland            Australia      -28.0  153.  2020-01-31    -1 confi~
## 3 Queensland            Australia      -28.0  153.  2020-02-02    -1 confi~
## 4 ""                    Japan           36    138   2020-02-07   -20 confi~
## 5 Lackland, TX (From D~ US              29.4  -98.6 2020-02-24    -2 confi~
## 6 Omaha, NE (From Diam~ US              41.3  -96.0 2020-02-24   -11 confi~
## 7 Travis, CA (From Dia~ US              38.3 -122.  2020-02-24    -5 confi~
## 8 From Diamond Princess Australia       35.4  140.  2020-02-29    -8 confi~
```

8. Which are the top 3 countries that have more than 20 deaths?

```
coronavirus %>%
    filter(type == "death") %>%
    group_by(Country.Region) %>%
    summarise(death = sum(cases)) %>%
    filter(death > 20) %>%
    arrange(desc(death))
```

```
## # A tibble: 3 x 2
##   Country.Region death
##   <chr>          <int>
## 1 Mainland China  2870
## 2 Iran              54
## 3 Italy             34
```

9. How many countries have a recovered-confirmed ratio of more than 0.60?

```
coronavirus %>%
    filter(type %in% c("confirmed", "recovered")) %>%
    group_by(Country.Region, type) %>%
    summarise(cases = sum(cases)) %>%

    # from {tidyr}: to have values put in separate columns
    spread(key = "type", value = "cases") %>%
    mutate(recovered = ifelse(is.na(recovered), 0, recovered)) %>%
    mutate(proportion = recovered / confirmed) %>%
    filter(proportion > 0.60) %>%
    arrange(desc(proportion))
```

```
## # A tibble: 10 x 4
## # Groups:   Country.Region [10]
##    Country.Region confirmed recovered proportion
##    <chr>              <int>     <dbl>      <dbl>
##  1 Cambodia               1         1      1
##  2 India                  3         3      1
```

```
##  3 Nepal                  1         1      1
##  4 Russia                 2         2      1
##  5 Sri Lanka              1         1      1
##  6 Vietnam               16        16      1
##  7 Macau                 10         8      0.8
##  8 Singapore            106        72      0.679
##  9 Thailand              42        28      0.667
## 10 Malaysia              29        18      0.621
```

10. What is the recovery-confirmed ratio for Italy?

```
coronavirus %>%
    filter(Country.Region == "Italy") %>%
    filter(type %in% c("confirmed", "recovered")) %>%
    group_by(type) %>%
    summarise(cases = sum(cases)) %>%

    # from {tidyr}: to have values put in separate columns
    spread(key = "type", value = "cases") %>%
    mutate(proportion = recovered / confirmed)
```

```
## # A tibble: 1 x 3
##   confirmed recovered proportion
##       <int>     <int>      <dbl>
## 1      1694        83     0.0490
```

In the following questions, you will explore the popularity of certain babynames. This dataset can be found in the like-named {babynames} library.

```
suppressPackageStartupMessages( library(babynames) )
```

12. What is the proportion of female babies that are called "Anna" in 1880 and 2017?

```
babynames %>%
    filter(sex == "F" & name == "Anna") %>%
    filter(year %in% c(1880,2017))
```

```
## # A tibble: 2 x 5
##    year sex   name      n    prop
##   <dbl> <chr> <chr> <int>   <dbl>
## 1  1880 F     Anna   2604 0.0267
## 2  2017 F     Anna   4520 0.00241
```

13. From 1880-1900, which was the most popular name for boys and girls?

```
babynames %>%
    filter(between(year, 1880, 1900)) %>%
    group_by(name, sex) %>%
    summarise(n = sum(n)) %>%
    arrange(desc(n)) %>%
    group_by(sex) %>%
    slice(1)
```

```
## # A tibble: 2 x 3
## # Groups:   sex [2]
##   name  sex        n
##   <chr> <chr>  <int>
## 1 Mary  F     239510
```

```
## 2 John  M     180444
```

14. For girls, what was the most popular name in 1880, 1917, 1943 and 2017?

```
babynames %>%
    filter(sex == "F")
```

```
## # A tibble: 1,138,293 x 5
##     year sex   name          n   prop
##    <dbl> <chr> <chr>     <int>  <dbl>
##  1  1880 F     Mary       7065 0.0724
##  2  1880 F     Anna       2604 0.0267
##  3  1880 F     Emma       2003 0.0205
##  4  1880 F     Elizabeth  1939 0.0199
##  5  1880 F     Minnie     1746 0.0179
##  6  1880 F     Margaret   1578 0.0162
##  7  1880 F     Ida        1472 0.0151
##  8  1880 F     Alice      1414 0.0145
##  9  1880 F     Bertha     1320 0.0135
## 10  1880 F     Sarah      1288 0.0132
## # ... with 1,138,283 more rows
```

data visualizations {ggplot2}

```
suppressPackageStartupMessages( library(ggplot2) )
```

data mining {rvest}

```
suppressPackageStartupMessages( library(rvest) )
```