U6614: Assignment 2: COVID-19 Country Case Data

SAMPLE SOLUTION

2020-09-22

Please submit your knitted .pdf file along with the corresponding R markdown (.rmd) via Courseworks by 11:59pm on Monday, September 21st.

Introduction.

Load packages:

```
library(tidyverse)
```

1 Load and prep the data.

Load the coronavirus.rda data from class and only keep confirmed cases. Data source: https://github.com/RamiKrispin/coronavirus/tree/master/data

```
#load raw data
load("coronavirus.rda")

#inspect raw data
str(coronavirus)
```

```
## ’data.frame':    182120 obs. of  7 variables:
##  $ date    : Date, format: "2020-01-22" "2020-01-23" ...
##  $ province: chr  "" "" "" "" ...
##  $ country : chr  "Afghanistan" "Afghanistan" "Afghanistan" "Afghanistan" ...
##  $ lat     : num  33.9 33.9 33.9 33.9 33.9 ...
##  $ long    : num  67.7 67.7 67.7 67.7 67.7 ...
##  $ type    : chr  "confirmed" "confirmed" "confirmed" "confirmed" ...
##  $ cases   : int  0 0 0 0 0 0 0 0 0 0 ...
```

```
#clean raw data as needed and assign to new data frame
covid_confirmed.df <- coronavirus %>%
  mutate(coronavirus, type.fac = as.factor(type)) %>% #convert type to factor
  filter(type.fac == "confirmed") %>%                 #keep confirmed cases
  mutate(country = if_else(province == "",
                           country,
                           paste(country, "-", province))) %>% #use province
  select(-type, -province) #drop unused type and province columns

head(covid_confirmed.df)
```

```
##         date     country      lat      long cases  type.fac
## 1 2020-01-22 Afghanistan 33.93911 67.70995     0 confirmed
## 2 2020-01-23 Afghanistan 33.93911 67.70995     0 confirmed
## 3 2020-01-24 Afghanistan 33.93911 67.70995     0 confirmed
## 4 2020-01-25 Afghanistan 33.93911 67.70995     0 confirmed
## 5 2020-01-26 Afghanistan 33.93911 67.70995     0 confirmed
## 6 2020-01-27 Afghanistan 33.93911 67.70995     0 confirmed
```

2 Describe the data.

Provide the following, along with any other information you think might be useful for the reader to know about the data.

- *unit of observation*
- *date range observed in the data*
- *number of countries (or administrative entities reporting data)*

The unit of observation is day-country; more accurately, observations are for countries except in select cases when the reporting entity is at the sub-country level such as a province (e.g. Xinjiang) or autonomous territory (e.g. Aruba).

```
covid_dates <- covid_confirmed.df %>%
  summarise(num_of_days = n_distinct(date), #reporting entities
            firstday = min(date), #first date
            lastday = max(date))  #last date
covid_dates
```

```
##   num_of_days   firstday    lastday
## 1         232 2020-01-22 2020-09-09
```

The data spans 232 days, from 2020-01-22 through 2020-09-09.

```
num_of_countries <- covid_confirmed.df %>%
  summarise(num_of_countries = n_distinct(country))
num_of_countries
```

```
##   num_of_countries
## 1              266
```

The data includes case counts for 266 different reporting entities.

3 Latest global case counts.

a. Create a new data frame that only includes observations for the most recent day only. *Note: don't hard-code a date to filter on, find the last day, store as a data object, and then refer back to (the element in) that object (see Lecture2-inclass.r for guidance)*

```
lastday <- covid_confirmed.df %>% summarise(max(date)) #find last date
covid_confirmed_last.df <- covid_confirmed.df %>%      #filter on last date
  filter(date == lastday[,1])
```

```
lastday_max <- covid_confirmed_last.df %>% summarise(max(cases))
lastday_max
```

b. What was max case count for the most recent day observed in the data?

```
##   max(cases)
## 1      95735
```

```r
#short way
max(covid_confirmed_last.df$cases)
```

```
## [1] 95735
```

The largest reported case count for September 9th was 95735 for India.

```r
covid_confirmed_last.df %>%
  select(country, cases) %>%
  arrange(desc(cases)) %>%
  head(n = 5)
```

c. List the top 5 countries (or administrative entities) by case count for the most recent day observed in the data?

```
##     country cases
## 1     India 95735
## 2    Brazil 35816
## 3        US 34256
## 4  Colombia 15318
## 5 Argentina 12259
```

```r
zerocases <- covid_confirmed_last.df %>%
  select(country, cases) %>%
  arrange(country) %>%
  filter(cases == 0)
zerocases
```

d. How many countries (or administrative entities) had zero confirmed cases for the most recent day?

```
##                                            country cases
## 1                              Antigua and Barbuda     0
## 2        Australia - Australian Capital Territory     0
## 3                  Australia - Northern Territory     0
## 4                          Australia - Queensland     0
## 5                     Australia - South Australia     0
## 6                            Australia - Tasmania     0
## 7                                        Barbados     0
## 8                                          Bhutan     0
## 9                                        Botswana     0
## 10                                         Brunei     0
## 11                                        Burundi     0
## 12                                       Cambodia     0
## 13                                       Cameroon     0
## 14                               Canada - Alberta     0
## 15                      Canada - British Columbia     0
## 16                      Canada - Diamond Princess     0
## 17                        Canada - Grand Princess     0
## 18                         Canada - New Brunswick     0
## 19             Canada - Newfoundland and Labrador     0
## 20                 Canada - Northwest Territories     0
## 21                           Canada - Nova Scotia     0
## 22                                 Canada - Yukon     0
## 23                                  China - Anhui     0
## 24                                China - Beijing     0
## 25                              China - Chongqing     0
## 26                                 China - Fujian     0
## 27                                  China - Gansu     0
## 28                                China - Guangxi     0
## 29                                China - Guizhou     0
## 30                                 China - Hainan     0
## 31                                  China - Hebei     0
## 32                           China - Heilongjiang     0
## 33                                  China - Henan     0
## 34                                  China - Hubei     0
## 35                                  China - Hunan     0
## 36                         China - Inner Mongolia     0
## 37                                China - Jiangsu     0
## 38                                China - Jiangxi     0
## 39                                  China - Jilin     0
## 40                               China - Liaoning     0
## 41                                  China - Macau     0
## 42                                China - Ningxia     0
## 43                                China - Qinghai     0
## 44                                China - Shaanxi     0
## 45                               China - Shandong     0
## 46                                 China - Shanxi     0
## 47                                China - Sichuan     0
## 48                                China - Tianjin     0
## 49                                  China - Tibet     0
## 50                               China - Xinjiang     0
## 51                                 China - Yunnan     0
## 52                               China - Zhejiang     0
## 53                                        Comoros     0
## 54                            Congo (Brazzaville)     0
## 55                            Denmark - Greenland     0
## 56                               Diamond Princess     0
## 57                                       Dominica     0
## 58                                        Eritrea     0
## 59                            France - Martinique     0
## 60                               France - Mayotte     0
## 61                         France - New Caledonia     0
## 62             France - Saint Pierre and Miquelon     0
## 63                                        Grenada     0
## 64                                  Guinea-Bissau     0
## 65                                       Holy See     0
## 66                                         Kosovo     0
## 67                                           Laos     0
## 68                                      Mauritius     0
## 69                                       Mongolia     0
## 70                                     MS Zaandam     0
## 71 Netherlands - Bonaire, Sint Eustatius and Saba     0
## 72                                      Nicaragua     0
## 73                                          Niger     0
## 74                          Saint Kitts and Nevis     0
## 75               Saint Vincent and the Grenadines     0
## 76                          Sao Tome and Principe     0
## 77                                     Seychelles     0
## 78                                        Somalia     0
## 79                                          Sudan     0
## 80                                        Taiwan*     0
## 81                                       Tanzania     0
## 82                                    Timor-Leste     0
## 83                                        Tunisia     0
## 84                      United Kingdom - Anguilla     0
## 85        United Kingdom - British Virgin Islands     0
## 86   United Kingdom - Falkland Islands (Malvinas)     0
## 87                   United Kingdom - Isle of Man     0
## 88                    United Kingdom - Montserrat     0
## 89                                 Western Sahara     0
```

```
nrow(zerocases)
```

```
## [1] 89
```

There were 89 countries reporting 0 COVID-19 cases on September 9th.

4 Oman case counts.

```
covid_confirmed_oman.df <- covid_confirmed.df %>%
  filter(country == "Oman") %>%
  arrange(desc(date))
head(covid_confirmed_oman.df)
```

a. Create a new data frame for daily confirmed case counts for Oman only. Sort in descending data order.

```
##         date country      lat     long cases  type.fac
## 1 2020-09-09    Oman 21.51258 55.92326   349 confirmed
## 2 2020-09-08    Oman 21.51258 55.92326   262 confirmed
## 3 2020-09-07    Oman 21.51258 55.92326   256 confirmed
## 4 2020-09-06    Oman 21.51258 55.92326   692 confirmed
## 5 2020-09-05    Oman 21.51258 55.92326     0 confirmed
## 6 2020-09-04    Oman 21.51258 55.92326     0 confirmed
```

```
oman <- covid_confirmed_oman.df %>%
  summarise(oman_mean = mean(cases),
            oman_min = min(cases),
            oman_max = max(cases))
oman
```

b. Find the daily mean, min, and max case counts for Oman over the duration of the pandemic and name each column appropriately.

```
##   oman_mean oman_min oman_max
## 1  379.0474        0     2164
```

The average daily case count for Oman over the period covered in this data is 379.0474138, with daily counts ranging from 0 to 2164.

c. What was the average daily case count in Oman over *last* 30 days of reported data? *[HINT: See Lecture2.1 -> Section 4.2 for examples of subsetting syntax that can help you refer to the first 30 rows of sorted data. If you're having trouble, you can also try using the `filter()` function and use the `row_number()` function as a part of the filtering condition]*

```
oman_avg_last30 <- covid_confirmed_oman.df[1:30,] %>%
  summarise(oman_last30days_mean = mean(cases))
oman_avg_last30
```

```
##   oman_last30days_mean
## 1             205.0667
```

The average daily case count in Oman over the last 30 days of reported data was 205.0666667.

```
oman_avg_first30 <- covid_confirmed_oman.df %>%
  arrange(date) %>%
  filter(row_number() < 30) %>%
  summarise(oman_first30days_mean = mean(cases))
oman_avg_first30
```

d. What was the average daily case count in Oman over the *first* 30 days of reported data?

```
##   oman_first30days_mean
## 1                     0
```

The average daily case count in Oman over the *first* 30 days of reported data is 0.