COVID19 assignment for STAT 412/612

J. Wall

2020-04-06

We will use this assignment to look at COVID-19 data

# First we will look at data in the US by county

The first portion of this work is thanks to Max Kuhn of RStudio.

theme\_set(theme\_bw())  
  
dat <-  
 read\_csv("https://raw.githubusercontent.com/nytimes/covid-19-data/master/us-counties.csv")  
  
deaths\_by\_state <- dat %>%  
 group\_by(state,date) %>%  
 summarize(total\_deaths = sum(deaths)) %>%  
 ungroup() %>%  
 filter(date == max(date)) %>%  
 arrange(desc(total\_deaths))  
  
# or  
  
deaths\_by\_state <- dat %>%  
 count(state,date, wt = deaths) %>%  
 filter(date == max(date)) %>%  
 rename(total\_deaths = n) %>%  
 arrange(desc(total\_deaths))  
  
deaths\_by\_state

## # A tibble: 55 x 3  
## state date total\_deaths  
## <chr> <date> <dbl>  
## 1 New York 2020-04-05 4161  
## 2 New Jersey 2020-04-05 917  
## 3 Michigan 2020-04-05 617  
## 4 Louisiana 2020-04-05 477  
## 5 California 2020-04-05 349  
## 6 Washington 2020-04-05 344  
## 7 Illinois 2020-04-05 283  
## 8 Massachusetts 2020-04-05 231  
## 9 Florida 2020-04-05 220  
## 10 Georgia 2020-04-05 219  
## # ... with 45 more rows

# total deaths   
sum(deaths\_by\_state$total\_deaths)

## [1] 9659

# Latest data from:  
max(dat$date)

## [1] "2020-04-05"

* **Exercise 1:** produce a tibble that has both deaths and total cases per state, arranged by the total number of cases in descending order

## # A tibble: 1,884 x 4  
## state date total\_deaths total\_cases  
## <chr> <date> <dbl> <dbl>  
## 1 New York 2020-04-05 4161 122911  
## 2 New York 2020-04-04 3568 114996  
## 3 New York 2020-04-03 2935 102945  
## 4 New York 2020-04-02 1958 92770  
## 5 New York 2020-04-01 1652 83890  
## 6 New York 2020-03-31 1282 75832  
## 7 New York 2020-03-30 1062 67216  
## 8 New York 2020-03-29 897 59568  
## 9 New York 2020-03-28 782 53364  
## 10 New York 2020-03-27 535 44636  
## # ... with 1,874 more rows

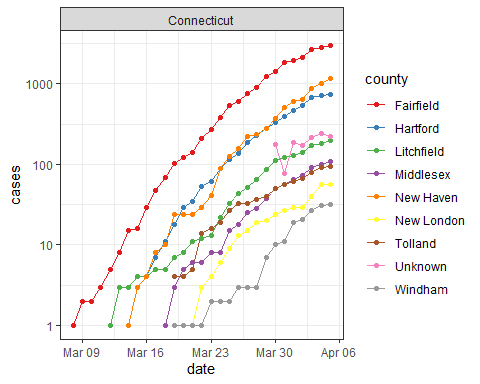
## [1] "total deaths = 9659"

## [1] "total cases = 336260"

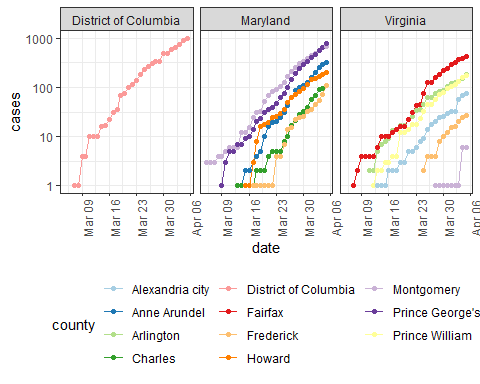
## [1] "latest data from: 2020-04-05"

Now we will choose just a few states to look at

dat\_small <-  
 dat %>%  
 filter(state %in% c("South Carolina", "Connecticut")) %>%  
 mutate(county = factor(county))  
  
dat\_small %>%  
 filter(state == "Connecticut") %>%  
 ggplot(aes(x = date, y = cases, group = county, col = county)) +  
 geom\_line() +  
 geom\_point() +  
 facet\_wrap(~ state) +  
 scale\_y\_log10() +  
 scale\_color\_brewer(palette = "Set1")



* **Exercise2:** (1 1/2 points) Do the analysis like above for the DMV area (DC, MD and VA) for both deaths and total cases. Use <http://www.theusgenweb.org/dcgenweb/geography/counties.shtml> for the counties in the DMV area. Be sure you get them all (hint: look at spelling in the dataset versus spelling you are entering)

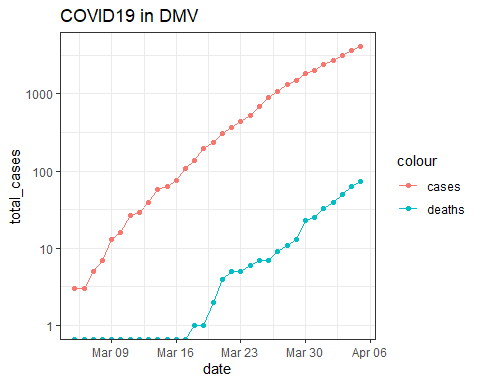
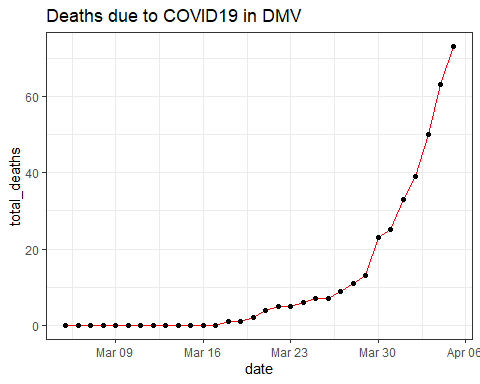
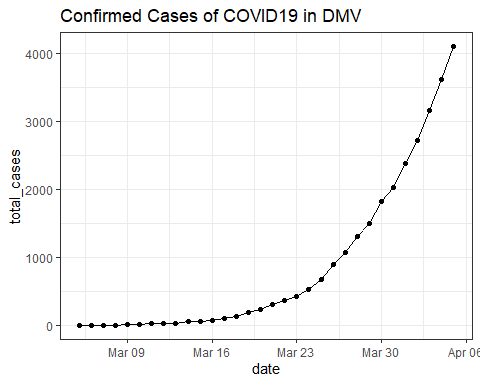


* **Exercise 3:** (1 1/2 points) Graph the total number of cases by date in the DMV area and the total deaths by date.

## [1] "total deaths = 73"

## [1] "total cases = 4106"

## [1] "latest data from: 2020-04-05"



# Get and clean worldwide Data

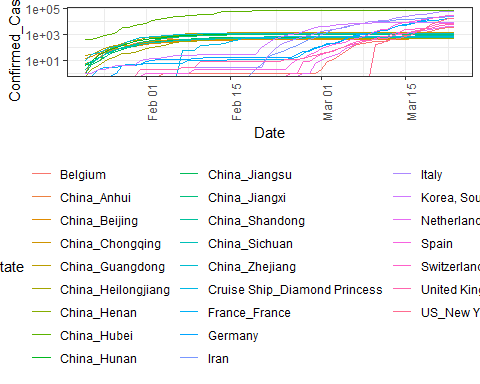
* **Exercise 4:** Get data from (<https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/archived_data/archived_time_series/time_series_19-covid-Confirmed_archived_0325.csv>). After you have read in the dataset, create a new variable called Country/State that combines the country with the province/state and tidy the dataset.

## # A tibble: 6 x 7  
## `Country/State` `Province/State` `Country/Region` Lat Long Date   
## <chr> <chr> <chr> <dbl> <dbl> <date>   
## 1 Thailand <NA> Thailand 15 101 2020-01-22  
## 2 Thailand <NA> Thailand 15 101 2020-01-23  
## 3 Thailand <NA> Thailand 15 101 2020-01-24  
## 4 Thailand <NA> Thailand 15 101 2020-01-25  
## 5 Thailand <NA> Thailand 15 101 2020-01-26  
## 6 Thailand <NA> Thailand 15 101 2020-01-27  
## # ... with 1 more variable: Confirmed\_Cases <dbl>

* **Exercise 5:** Find the 25 Country/State’s with the most Confirmed Cases (Hint: use group\_by, summarize and ungroup?)

## # A tibble: 25 x 3  
## `Country/State` `Country/Region` ttl  
## <chr> <chr> <dbl>  
## 1 China\_Hubei China 2894885  
## 2 Italy Italy 497959  
## 3 Iran Iran 252770  
## 4 Spain Spain 186200  
## 5 Korea, South Korea, South 181699  
## 6 Germany Germany 160974  
## 7 France\_France France 117724  
## 8 China\_Guangdong China 67015  
## 9 US\_New York US 64538  
## 10 China\_Henan China 61811  
## 11 China\_Zhejiang China 61207  
## 12 China\_Hunan China 50256  
## 13 China\_Anhui China 47647  
## 14 Switzerland Switzerland 46413  
## 15 China\_Jiangxi China 45181  
## 16 United Kingdom\_United Kingdom United Kingdom 35081  
## 17 China\_Shandong China 33603  
## 18 China\_Jiangsu China 30272  
## 19 China\_Chongqing China 28659  
## 20 Netherlands\_Netherlands Netherlands 28030  
## 21 Cruise Ship\_Diamond Princess Cruise Ship 26228  
## 22 China\_Sichuan China 25719  
## 23 China\_Heilongjiang China 21991  
## 24 Belgium Belgium 21324  
## 25 China\_Beijing China 20967

* **Exercise 6:** (2 points) Plot the number of cases over time for the top 25 country/states. Use appropriate scales for the axes.



* **Exercise 7:** (2 points) Get the data for the state of NY from this csse dataset and compare the growth of cases from this dataset to the growth shown in the nytimes dataset used for exercises 1-3. Does the data match? If not, how is it different?

## # A tibble: 36 x 3  
## Date Confirmed\_Cases n  
## <date> <dbl> <dbl>  
## 1 2020-03-01 NA 1  
## 2 2020-03-02 NA 1  
## 3 2020-03-03 NA 2  
## 4 2020-03-04 NA 11  
## 5 2020-03-05 NA 22  
## 6 2020-03-06 NA 44  
## 7 2020-03-07 NA 89  
## 8 2020-03-08 NA 106  
## 9 2020-03-09 NA 142  
## 10 2020-03-10 173 173  
## 11 2020-03-11 220 217  
## 12 2020-03-12 328 326  
## 13 2020-03-13 421 421  
## 14 2020-03-14 525 610  
## 15 2020-03-15 732 732  
## 16 2020-03-16 967 950  
## 17 2020-03-17 1706 1374  
## 18 2020-03-18 2495 2382  
## 19 2020-03-19 5365 4152  
## 20 2020-03-20 8310 7102  
## 21 2020-03-21 11710 10356  
## 22 2020-03-22 15793 15168  
## 23 2020-03-23 15793 20875  
## 24 2020-03-24 NA 25666  
## 25 2020-03-25 NA 33067  
## 26 2020-03-26 NA 38988  
## 27 2020-03-27 NA 44636  
## 28 2020-03-28 NA 53364  
## 29 2020-03-29 NA 59568  
## 30 2020-03-30 NA 67216  
## 31 2020-03-31 NA 75832  
## 32 2020-04-01 NA 83890  
## 33 2020-04-02 NA 92770  
## 34 2020-04-03 NA 102945  
## 35 2020-04-04 NA 114996  
## 36 2020-04-05 NA 122911

