Covid_19

lwarne

9/23/2021

COVID-19 and the Data

We assume that larger populations would have greater numbers of cases. Is this true for the State of Maine?

Library in packages

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.4     v dplyr   1.0.7
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   2.0.1     v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(lubridate)
```

```
##
## Attaching package: 'lubridate'

## The following objects are masked from 'package:base':
##
##     date, intersect, setdiff, union
```

Identify and read in Covid Data

Expected to use open source COVID-19 data on github curated by Johns Hopkins. "Use brackets <> to create document hyperlink" https://github.com/CSSEGISandData/COVID-19/tree/master/csse_covid_19_data/csse_covid_19_time_series/

Alternative Data Source Selected data from the Maine CDC data to prepare a local report.Error messages beyond my R abilities limited my ability to use the JH data files from github. https://gateway.maine.gov/dhhs-apps/mecdc_covid/cases_by_county_history.csv

```
maine_covid <- read.csv("https://gateway.maine.gov/dhhs-apps/mecdc_covid/cases_by_county_history.csv")
summary(maine_covid)
```

```
##      Date           Patient_County     Case_Status         Total_Cases
##  Length:19652       Length:19652       Length:19652       Min.   :  0.000
##  Class :character   Class :character   Class :character   1st Qu.:  0.000
##  Mode  :character   Mode  :character   Mode  :character   Median :  0.000
##                                                           Mean   :  4.831
##                                                           3rd Qu.:  4.000
##                                                           Max.   :334.000
##
##  Completed_Isolations     Deaths       Hospitalizations Population_2018
##  Min.   : 0.0000      Min.   :0.0000   Min.   :0.0000   Min.   : 16800
##  1st Qu.: 0.0000      1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.: 35311
##  Median : 0.0000      Median :0.0000   Median :0.0000   Median : 52702
##  Mean   : 0.6592      Mean   :0.0547   Mean   :0.1318   Mean   : 83650
##  3rd Qu.: 0.0000      3rd Qu.:0.0000   3rd Qu.:0.0000   3rd Qu.:111280
##  Max.   :76.0000      Max.   :7.0000   Max.   :9.0000   Max.   :293557
##                                                         NA's   :1156
```

Data Summary

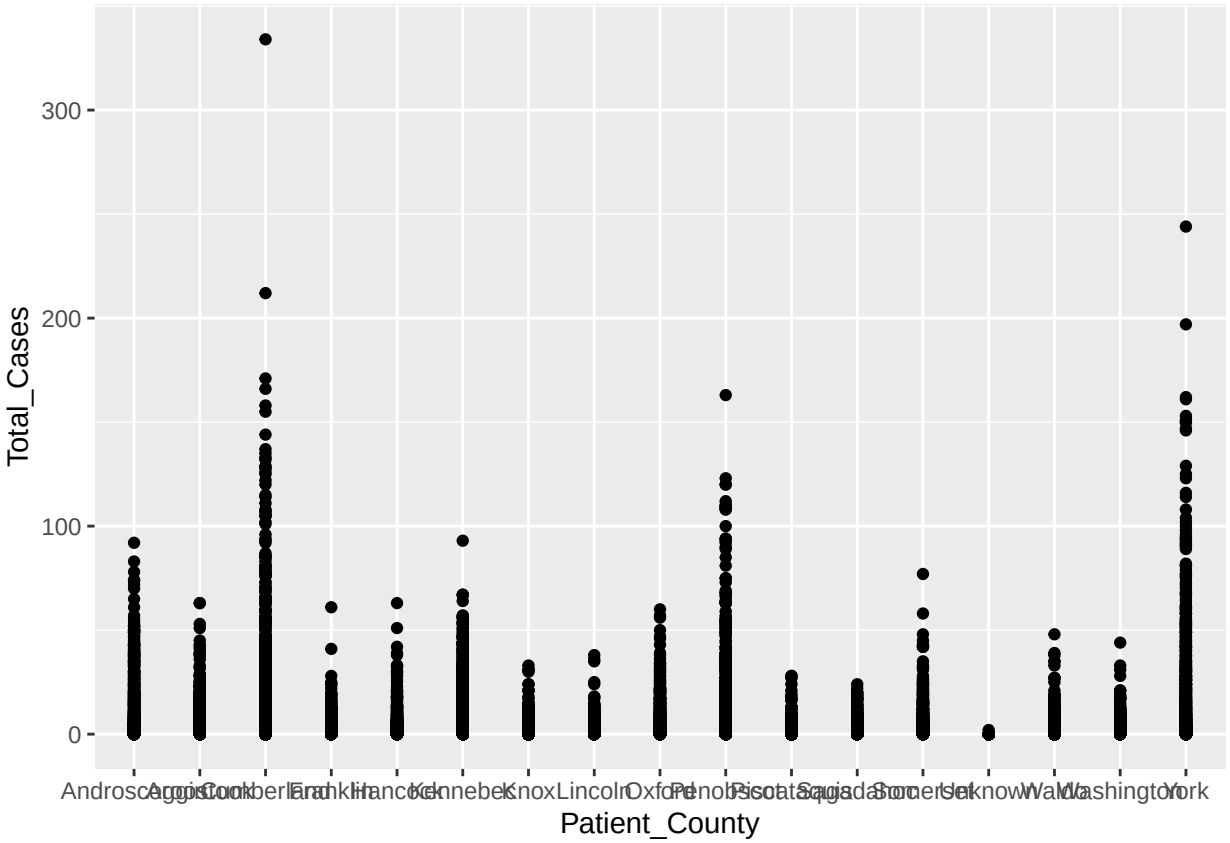
This data chunk updated to run from 3/12/2020 through 10/10/2021 provides the State of Maine Covid-19 data by County. as a data.frame With 19,652 observations and 8 columns of variables:

Date, Patient_County, Case_Status, Total_Cases, Completed Isolations, Deaths, Hospitalizations, Population_2018.

This data includes 1,156 case counts not connected to a county, meaning 5.8% of the cases patients from *unknown* locations. In addition, this data contains 2,685 observation days with counties reporting less than 10 total cases.

Daily Cases by County

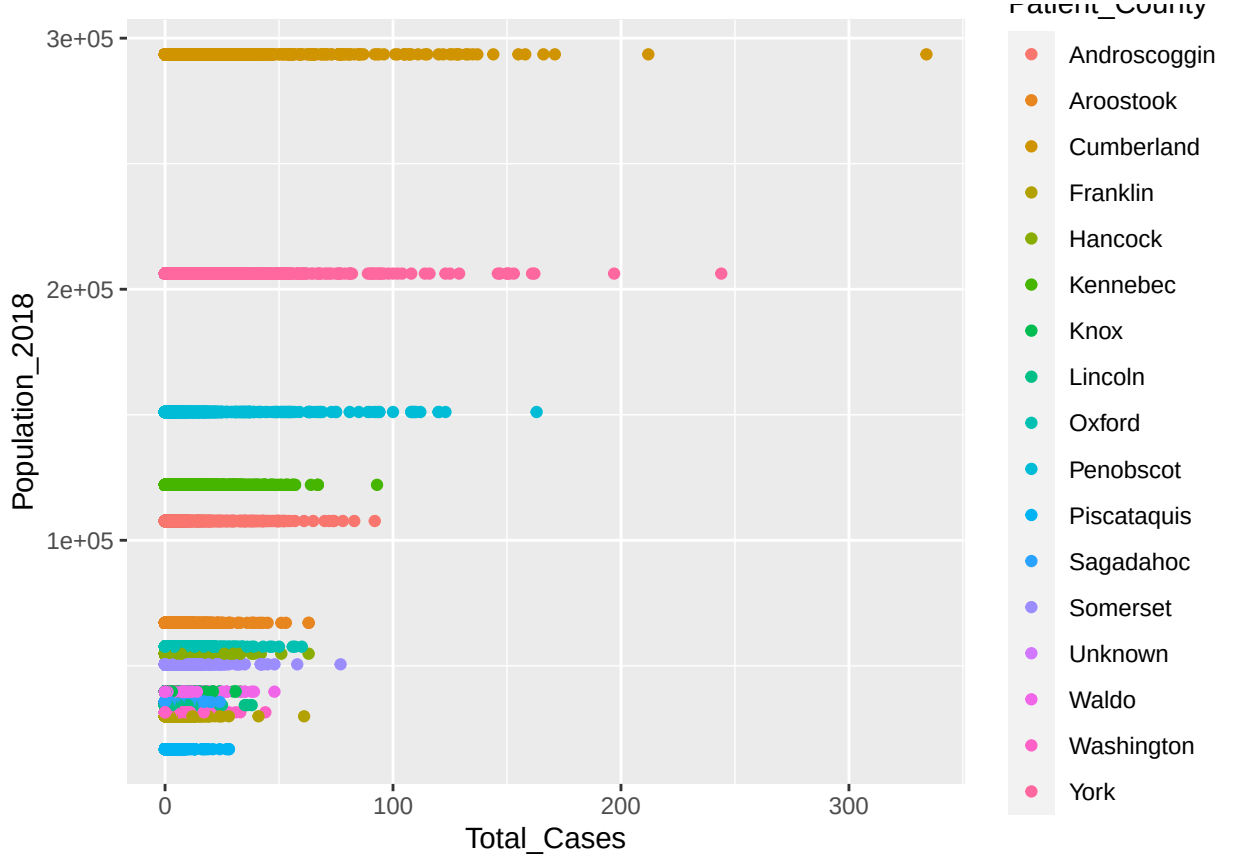
```
ggplot(data = maine_covid) +
 geom_point(mapping = aes(x = Patient_County, y = Total_Cases))
```



County Comparison Each point indicates daily total cases in the 16 Maine counties. Those counties with larger populations have signficant outliers: days with very high Covid-19 counts.

```
ggplot(data = maine_covid) +
  geom_point(mapping = aes(x = Total_Cases , y = Population_2018, color = Patient_County))
```

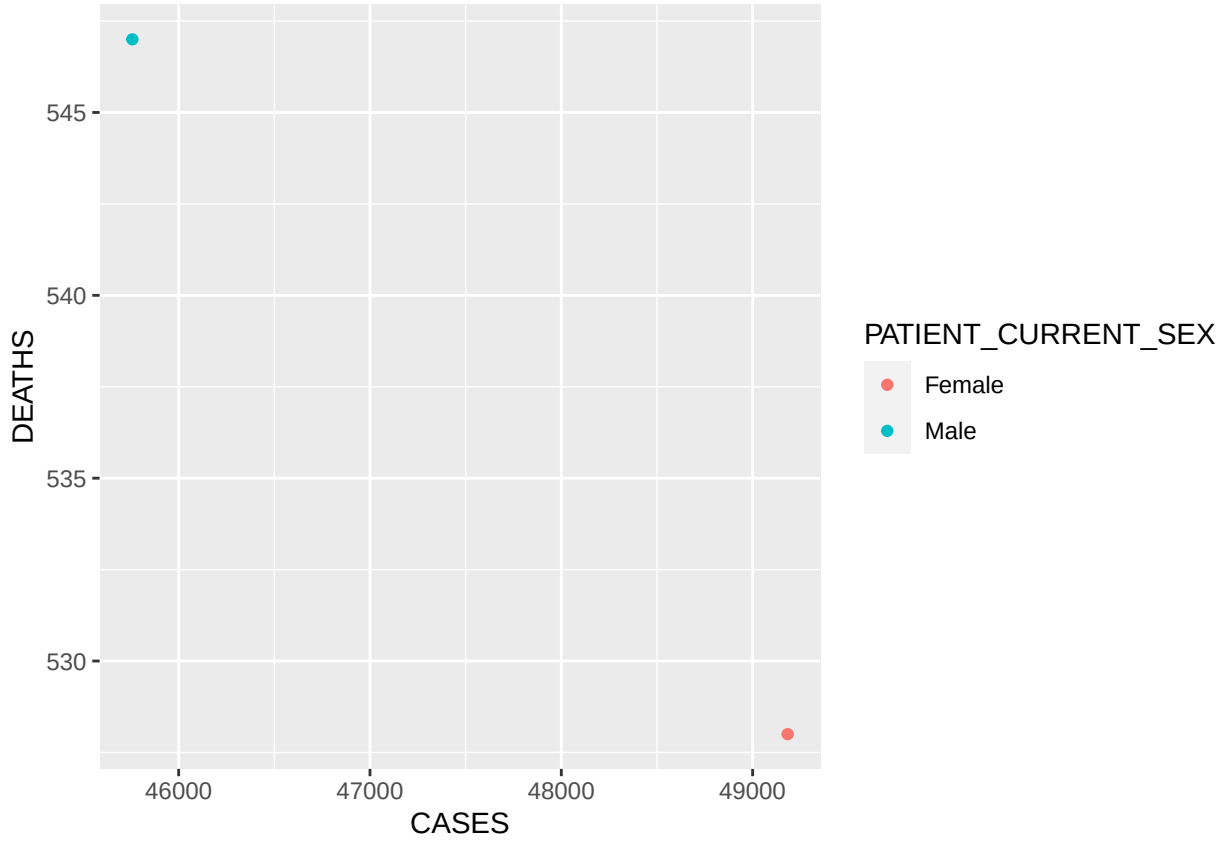
```
## Warning: Removed 1156 rows containing missing values (geom_point).
```



Compare Male and Female Stats

This data shows a difference of 19 fewer female deaths, though there were 3,425 more cases among females than males.

```
maine_sex <- read.csv("https://gateway.maine.gov/dhhs-apps/mecdc_covid/cases_by_sex.csv")
ggplot(data = maine_sex) +
  geom_point(mapping = aes(x = CASES, y = DEATHS, color = PATIENT_CURRENT_SEX))
```



Covid_ME <- maine_covid %>% group_by(Patient_County, Date) %>% select(Date, Patient_County, Total_Cases, Deaths, Hospitalizations, Population_2018) %>% ungroup() summary(Covid_ME)

```
str(maine_covid)
```

```
## 'data.frame':    19652 obs. of  8 variables:
##  $ Date                : chr  "2021-10-10" "2021-10-10" "2021-10-10" "2021-10-10" ...
##  $ Patient_County      : chr  "Androscoggin" "Androscoggin" "Aroostook" "Aroostook" ...
##  $ Case_Status         : chr  "Confirmed" "Probable" "Confirmed" "Probable" ...
##  $ Total_Cases         : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Completed_Isolations: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Deaths              : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Hospitalizations    : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Population_2018     : int  107679 107679 67111 67111 293557 293557 29897 29897 54811 54811 ...
```

Conclusions and Bias

This is raw data that requires context. While more populous areas in Maine, showed more cases per day. That needs to be a function of population. I found it interesting that the females had more total cases, but fewer deaths. I also appreciated seeing that the cases based on population showed statistical significance, even with the outliers being many more cases than typical, those outliers were within the higher populated areas.

There is no bias in the data because they are simply observed numbers. However, the signficant number of unknown counties skewed the data.