STAT 505 Fall 2022: Homework 1

David Agyemfra Atakora

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.3.6     v purrr   0.3.4
## v tibble  3.1.8     v dplyr   1.0.9
## v tidyr   1.2.0     v stringr 1.4.1
## v readr   2.1.2     v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
covid19 <- read_csv("https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_dat
```

```
## Rows: 58 Columns: 21
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr   (3): Province_State, Country_Region, ISO3
## dbl  (10): Lat, Long_, Confirmed, Deaths, FIPS, Incident_Rate, Total_Test_Re...
## lgl   (6): Recovered, Active, People_Hospitalized, Hospitalization_Rate, Peo...
## dttm  (1): Last_Update
## date  (1): Date
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
library(dplyr)
covid19
```

```
## # A tibble: 58 x 21
##    Province_St~1 Count~2 Last_Update           Lat  Long_ Confi~3 Deaths Recov~4
##    <chr>         <chr>   <dttm>              <dbl>  <dbl>   <dbl>  <dbl> <lgl>
##  1 Alabama       US      2021-08-27 04:30:55  32.3  -86.9  676795  12103 NA
##  2 Alaska        US      2021-08-27 04:30:55  61.4 -152.    86218    438 NA
##  3 American Sam~ US      2021-08-27 04:30:55 -14.3 -170.        0      0 NA
##  4 Arizona       US      2021-08-27 04:30:55  33.7 -111.   998164  18661 NA
##  5 Arkansas      US      2021-08-27 04:30:55  35.0  -92.4  443564   6806 NA
##  6 California    US      2021-08-27 04:30:55  36.1 -120.  4388404  65100 NA
##  7 Colorado      US      2021-08-27 04:30:55  39.1 -105.   618566   7352 NA
##  8 Connecticut   US      2021-08-27 04:30:55  41.6  -72.8  369920   8355 NA
##  9 Delaware      US      2021-08-27 04:30:55  39.3  -75.5  118016   1872 NA
## 10 Diamond Prin~ US      2021-08-27 04:30:55  NA     NA        49      0 NA
## # ... with 48 more rows, 13 more variables: Active <lgl>, FIPS <dbl>,
## #   Incident_Rate <dbl>, Total_Test_Results <dbl>, People_Hospitalized <lgl>,
## #   Case_Fatality_Ratio <dbl>, UID <dbl>, ISO3 <chr>, Testing_Rate <dbl>,
## #   Hospitalization_Rate <lgl>, Date <date>, People_Tested <lgl>,
## #   Mortality_Rate <lgl>, and abbreviated variable names 1: Province_State,
## #   2: Country_Region, 3: Confirmed, 4: Recovered
```

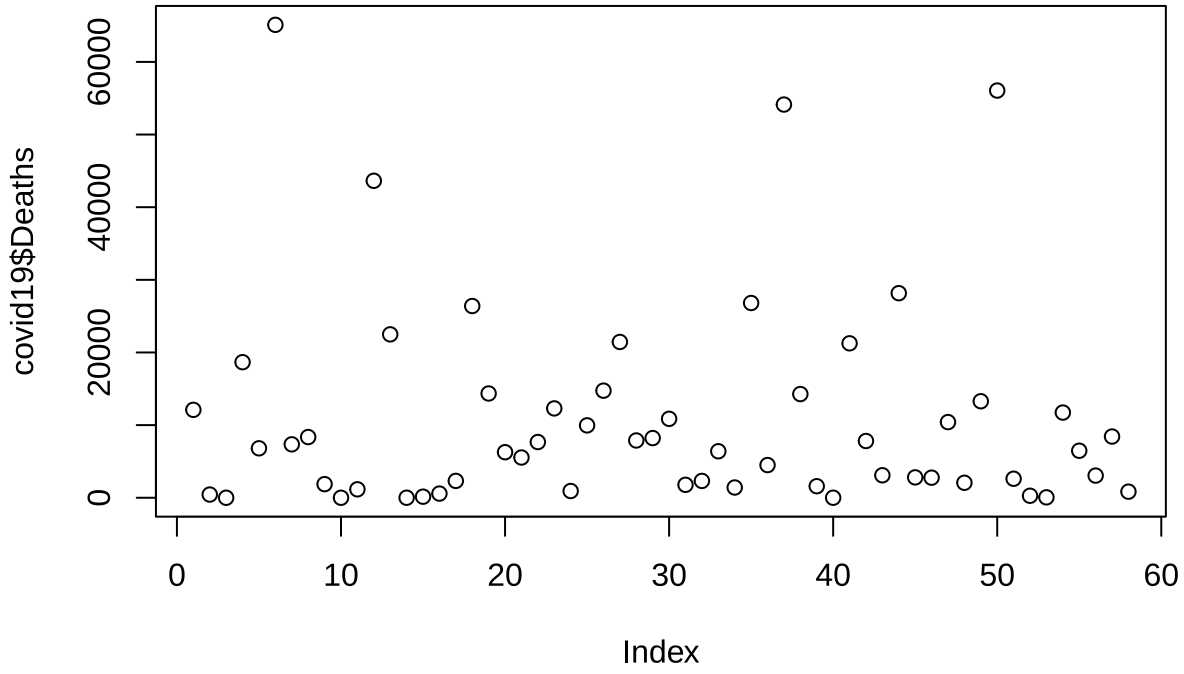
```
colnames(covid19)
```

```
##  [1] "Province_State"       "Country_Region"       "Last_Update"
##  [4] "Lat"                  "Long_"                "Confirmed"
##  [7] "Deaths"               "Recovered"            "Active"
## [10] "FIPS"                 "Incident_Rate"        "Total_Test_Results"
## [13] "People_Hospitalized"  "Case_Fatality_Ratio"  "UID"
## [16] "ISO3"                 "Testing_Rate"         "Hospitalization_Rate"
## [19] "Date"                 "People_Tested"        "Mortality_Rate"
```

```
covid19$Province_State
```

```
##  [1] "Alabama"                  "Alaska"
##  [3] "American Samoa"           "Arizona"
##  [5] "Arkansas"                 "California"
##  [7] "Colorado"                 "Connecticut"
##  [9] "Delaware"                 "Diamond Princess"
## [11] "District of Columbia"     "Florida"
## [13] "Georgia"                  "Grand Princess"
## [15] "Guam"                     "Hawaii"
## [17] "Idaho"                    "Illinois"
## [19] "Indiana"                  "Iowa"
## [21] "Kansas"                   "Kentucky"
## [23] "Louisiana"                "Maine"
## [25] "Maryland"                 "Massachusetts"
## [27] "Michigan"                 "Minnesota"
## [29] "Mississippi"              "Missouri"
## [31] "Montana"                  "Nebraska"
## [33] "Nevada"                   "New Hampshire"
## [35] "New Jersey"               "New Mexico"
## [37] "New York"                 "North Carolina"
## [39] "North Dakota"             "Northern Mariana Islands"
## [41] "Ohio"                     "Oklahoma"
## [43] "Oregon"                   "Pennsylvania"
## [45] "Puerto Rico"              "Rhode Island"
## [47] "South Carolina"           "South Dakota"
## [49] "Tennessee"                "Texas"
## [51] "Utah"                     "Vermont"
## [53] "Virgin Islands"           "Virginia"
## [55] "Washington"               "West Virginia"
## [57] "Wisconsin"                "Wyoming"
```

```
plot(covid19$Deaths)
```

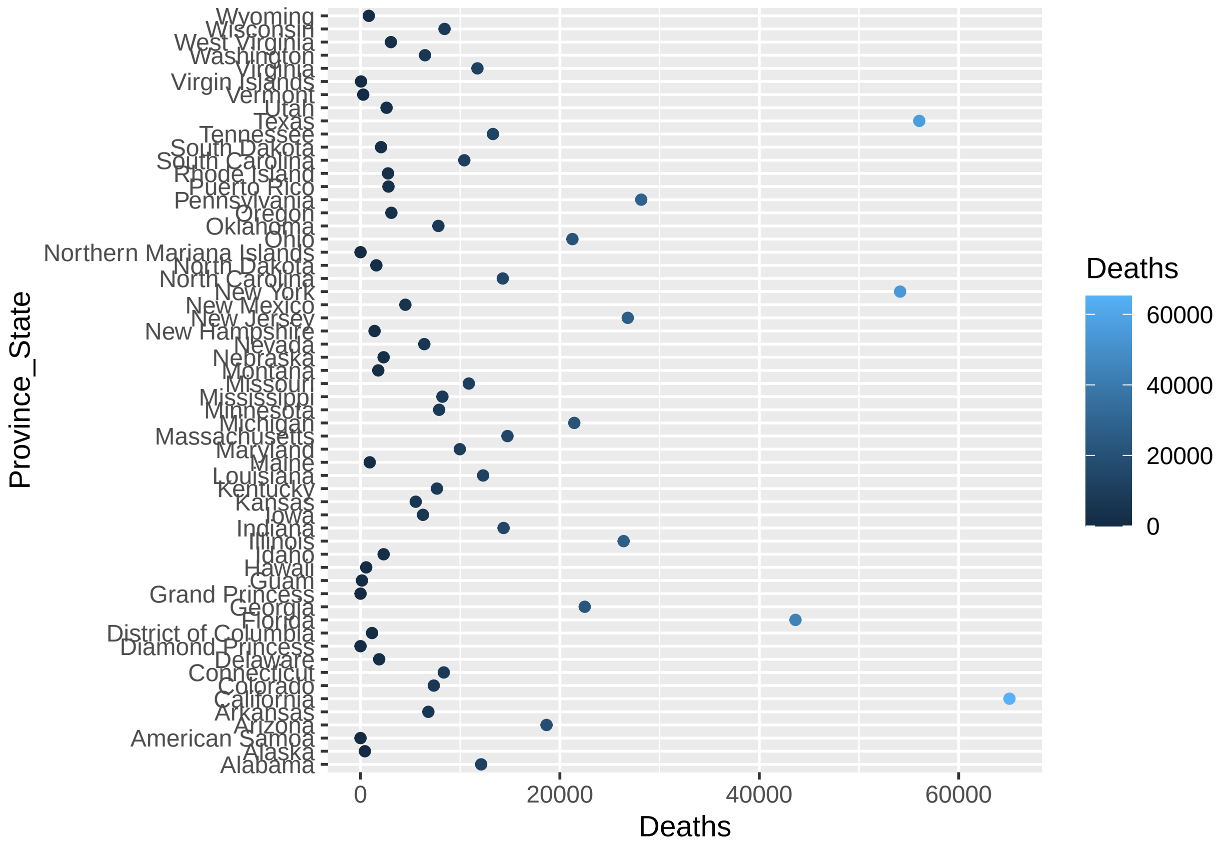


```r
covid19 %>% group_by(Province_State)
```

```
## # A tibble: 58 x 21
## # Groups:   Province_State [58]
##    Province_St~1 Count~2 Last_Update           Lat  Long_ Confi~3 Deaths Recov~4
##    <chr>         <chr>   <dttm>              <dbl>  <dbl>   <dbl>  <dbl> <lgl>
##  1 Alabama       US      2021-08-27 04:30:55  32.3  -86.9  676795  12103 NA
##  2 Alaska        US      2021-08-27 04:30:55  61.4 -152.    86218    438 NA
##  3 American Sam~ US      2021-08-27 04:30:55 -14.3 -170.        0      0 NA
##  4 Arizona       US      2021-08-27 04:30:55  33.7 -111.   998164  18661 NA
##  5 Arkansas      US      2021-08-27 04:30:55  35.0  -92.4  443564   6806 NA
##  6 California    US      2021-08-27 04:30:55  36.1 -120.  4388404  65100 NA
##  7 Colorado      US      2021-08-27 04:30:55  39.1 -105.   618566   7352 NA
##  8 Connecticut   US      2021-08-27 04:30:55  41.6  -72.8  369920   8355 NA
##  9 Delaware      US      2021-08-27 04:30:55  39.3  -75.5  118016   1872 NA
## 10 Diamond Prin~ US      2021-08-27 04:30:55  NA     NA        49      0 NA
## # ... with 48 more rows, 13 more variables: Active <lgl>, FIPS <dbl>,
## #   Incident_Rate <dbl>, Total_Test_Results <dbl>, People_Hospitalized <lgl>,
## #   Case_Fatality_Ratio <dbl>, UID <dbl>, ISO3 <chr>, Testing_Rate <dbl>,
## #   Hospitalization_Rate <lgl>, Date <date>, People_Tested <lgl>,
## #   Mortality_Rate <lgl>, and abbreviated variable names 1: Province_State,
## #   2: Country_Region, 3: Confirmed, 4: Recovered
```

```r
#ggplot(df,aes(x,y))+geom_point(aes(colour=x))
```

```
#ggplot2
library(ggplot2)
ggplot(covid19, aes(x = Deaths, y = Province_State)) +
    geom_point(aes(colour=Deaths))
```



HW1

The purpose of this homework is to make sure that you have all of the proper technology tools installed.

Answer Q1 using a .RMD file. Then upload that source file and a PDF output file to Github for your submission.

Q1.

Download a .CSV file containing up to date information about COVID-19 cases. A file from August 16th (2021) can be downloaded can be at https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse_covid_19_daily_reports_us/08-16-2021.csv

a. (4 points)

Use `dplyr` and a `group_by()` statement to summarize the data in some fashion.

b. (4 points)

Use ggplot2 to create a figure of the data

Q2. (2 points)

Make at least one post on the Microsoft Teams app. This could be creating a new channel, adding a comment, etc..

Q3. (1 points)

What are you most excited about this semester (this class or in general)?

The use of R and github

Q4. (1 points)

What are you most worried about this semester (this class or in general)?

Not being able to make out the best from this semester

Q5. (1 point)

What do you hope to learn in this class?

I want to develop a strong knowledge in Statistical Analysis and modeling. Also, how I can apply statistics to a real life dataset.

Q6. (1 point)

What degree do you hope to earn from MSU?

Master of Science in Statistics

Q7. (1 point)

What do you hope to do after graduating from MSU?

I hope to develop a career in Actuarial Science with my Statistics degree being able to make sense from any large datasets.

Q8. (1 point)

Is there anything else that you want me to know?

I am ready to learn more from this course; in and out of class.