Mandha-663-HW3

harinath

2022-10-02

```
library(IDDA)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

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

```
library(tidyr)
library(plotly)
```

```
## Warning: package 'plotly' was built under R version 4.2.2

## Loading required package: ggplot2

##
## Attaching package: 'plotly'

## The following object is masked from 'package:ggplot2':
##
##     last_plot

## The following object is masked from 'package:stats':
##
##     filter

## The following object is masked from 'package:graphics':
##
##     layout
```

```
library(htmlwidgets)
```

```
## Warning: package 'htmlwidgets' was built under R version 4.2.2
```

```
library(lubridate)
```

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

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

```r
library(ggplot2)
```

```r
setwd("C:\\Users\\harin\\OneDrive\\Documents\\Statistical_graphs_and_exploration")
```

(a) Create a time series plot for the logarithm of the cumulative infected cases for the top ten states with the largest new number of infected cases on December 31, 2020.

```r
install.packages("plotly")
```

```
## Warning: package 'plotly' is in use and will not be installed
```

```r
library(IDDA)
data("state.long")
```

```r
library(dplyr)
data_grouped_by_states <- IDDA::state.long %>%
  dplyr::filter(DATE=="2020-12-31")%>%
  group_by(State)

data_grouped_by_states
```

```
## # A tibble: 49 x 7
## # Groups:   State [49]
##     State             Region    Division            pop DATE       Infec~1 Death
##     <chr>             <fct>     <fct>             <int> <date>        <int> <int>
##  1 Alabama           South     East South Cent~ 4.89e6 2020-12-31   361226  4827
##  2 Arizona           West      Mountain         7.17e6 2020-12-31   523829  8879
##  3 Arkansas          South     West South Cent~ 3.01e6 2020-12-31   222319  3676
##  4 California        West      Pacific          3.96e7 2020-12-31  2307706 25963
##  5 Colorado          West      Mountain         5.70e6 2020-12-31   335579  4879
##  6 Connecticut       Northeast New England      3.57e6 2020-12-31   185043  5995
##  7 Delaware          South     South Atlantic   9.67e5 2020-12-31    57320   926
##  8 DistrictofColumbia South    South Atlantic   7.02e5 2020-12-31    28983   786
##  9 Florida           South     South Atlantic   2.13e7 2020-12-31  1320297 21672
## 10 Georgia           South     South Atlantic   1.05e7 2020-12-31   640442 10582
## # ... with 39 more rows, and abbreviated variable name 1: Infected
```

```r
states_sorted <- data_grouped_by_states[order(-data_grouped_by_states$Infected), ]
states_sorted
```

```
## # A tibble: 49 x 7
## # Groups:   State [49]
##     State         Region    Division                 pop DATE       Infected Death
##     <chr>         <fct>     <fct>                  <int> <date>         <int> <int>
##  1 California     West      Pacific             39557045 2020-12-31   2307706 25963
##  2 Texas         South     West South Central  28701845 2020-12-31   1772163 28155
##  3 Florida       South     South Atlantic      21299325 2020-12-31   1320297 21672
##  4 NewYork       Northeast Middle Atlantic     19542209 2020-12-31    979040 37557
##  5 Illinois      Midwest   East North Central  12741080 2020-12-31    965840 17979
##  6 Ohio          Midwest   East North Central  11689442 2020-12-31    700380  8962
##  7 Pennsylvania  Northeast Middle Atlantic     12807060 2020-12-31    646060 16020
##  8 Georgia       South     South Atlantic      10519475 2020-12-31    640442 10582
##  9 Tennessee     South     East South Central   6770010 2020-12-31    565676  6787
## 10 NorthCarolina South     South Atlantic      10383620 2020-12-31    541070  6811
## # ... with 39 more rows
```

```
library(dplyr)
plot_data <- top_n(states_sorted, 10)

## Selecting by Death
plot_data

## # A tibble: 49 x 7
## # Groups:   State [49]
##    State          Region    Division                pop DATE        Infected Death
##    <chr>          <fct>     <fct>                  <int> <date>         <int> <int>
##  1 California     West      Pacific            39557045 2020-12-31   2307706 25963
##  2 Texas          South     West South Central 28701845 2020-12-31   1772163 28155
##  3 Florida        South     South Atlantic     21299325 2020-12-31   1320297 21672
##  4 NewYork        Northeast Middle Atlantic    19542209 2020-12-31    979040 37557
##  5 Illinois       Midwest   East North Central 12741080 2020-12-31    965840 17979
##  6 Ohio           Midwest   East North Central 11689442 2020-12-31    700380  8962
##  7 Pennsylvania   Northeast Middle Atlantic    12807060 2020-12-31    646060 16020
##  8 Georgia        South     South Atlantic     10519475 2020-12-31    640442 10582
##  9 Tennessee      South     East South Central  6770010 2020-12-31    565676  6787
## 10 NorthCarolina South     South Atlantic     10383620 2020-12-31    541070  6811
## # ... with 39 more rows
```

#Time series plot for part (a)

```
library("plotly")
library("tidyr")
plot1 <- plot_ly(data = plot_data) %>%
  mutate(cum_inf_cases = cumsum(Infected))%>%
  add_trace(x = ~log(cum_inf_cases), y = ~Infected,
  name = ' State',
  type = 'scatter', mode = 'markers')
plot1
```
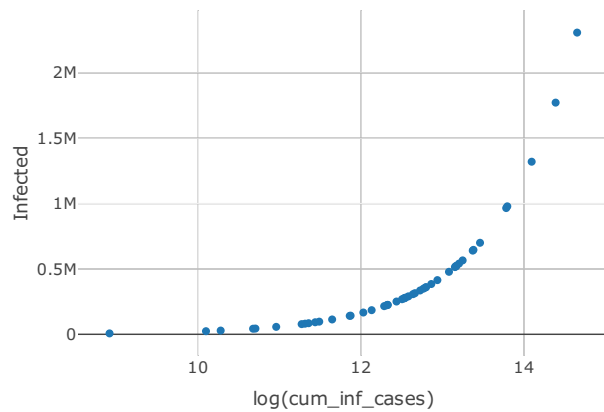
```
data_dec31<- IDDA::state.long %>%
  dplyr::filter(DATE=="2020-12-31")
data_dec31
```

```
## # A tibble: 49 x 7
##    State            Region    Division               pop DATE        Infec~1 Death
##    <chr>            <fct>     <fct>                <int> <date>        <int> <int>
##  1 Alabama          South     East South Cent~ 4.89e6 2020-12-31    361226  4827
##  2 Arizona          West      Mountain         7.17e6 2020-12-31    523829  8879
##  3 Arkansas         South     West South Cent~ 3.01e6 2020-12-31    222319  3676
##  4 California        West     Pacific          3.96e7 2020-12-31   2307706 25963
##  5 Colorado         West      Mountain         5.70e6 2020-12-31    335579  4879
##  6 Connecticut       Northeast New England     3.57e6 2020-12-31    185043  5995
##  7 Delaware         South     South Atlantic   9.67e5 2020-12-31     57320   926
##  8 DistrictofColumbia South   South Atlantic   7.02e5 2020-12-31     28983   786
##  9 Florida          South     South Atlantic   2.13e7 2020-12-31   1320297 21672
## 10 Georgia          South     South Atlantic   1.05e7 2020-12-31    640442 10582
## # ... with 39 more rows, and abbreviated variable name 1: Infected
```

```
library(dplyr)
Pie_Data<-count(data_dec31,Infected, Region)
Pie_Data
```

```
## # A tibble: 49 x 3
##    Infected Region       n
##       <int> <fct>    <int>
```

```
## 1      7403 Northeast     1
## 2     24200 Northeast     1
## 3     28983 South         1
## 4     43145 Northeast     1
## 5     44409 West          1
## 6     57320 South         1
## 7     78327 Northeast     1
## 8     81701 West          1
## 9     85334 South         1
## 10    92495 Midwest       1
## # ... with 39 more rows
```

```r
library(plotly)
plot2 <- plot_ly(Pie_Data,values=~n,labels=~factor(Region),marker=list(colors=c("blue","green")),type="
plot2
```
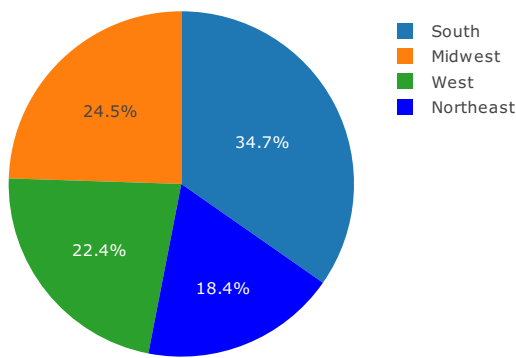

Saving plots as HTML

```r
library(htmlwidgets)
saveWidget(plot1, "hw_3_1(a)_timeseries.html", selfcontained = F, libdir = "C:\\Users\\harin\\Downloads
saveWidget(plot2, "hw_3_1(b)_piechart.html", selfcontained = F, libdir = "C:\\Users\\harin\\Downloads\\
```
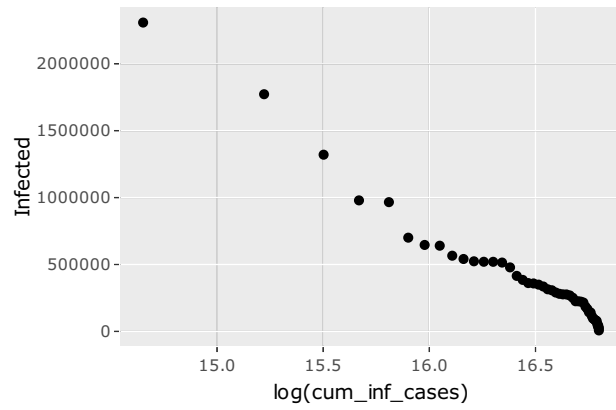
```r
cum_inf_cases = cumsum(plot_data$Infected)
```
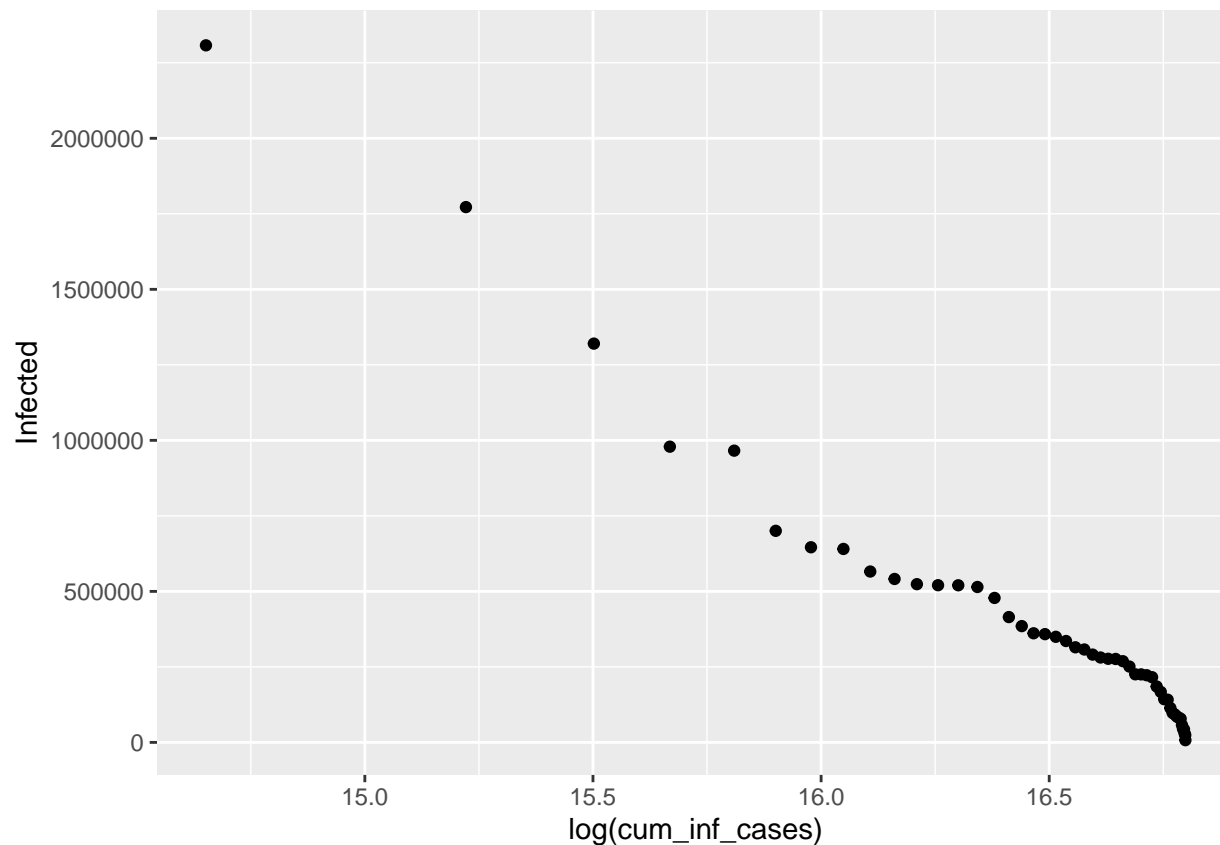
```r
# Basic Line
library(ggplot2)
library(plotly)
```

```
p <- ggplot(plot_data, aes(log(cum_inf_cases), Infected)) +
  geom_point()

ggplotly(p)
```



```
p
```

```
library(IDDA)
data(Test.state)
data(PosTest.state)
```

Transforming test.state from wide to long

```
library(reshape2)
```

```
## Warning: package 'reshape2' was built under R version 4.2.2
```

```
##
## Attaching package: 'reshape2'
```

```
## The following object is masked from 'package:tidyr':
##
##     smiths
```

```
library(tidyr)
```

```
Test.long <- Test.state %>%
  pivot_longer(X2020.12.31: X2020.11.13, names_to = "DATE",values_to = "Infected")
```

Transforming postest.state from wide to long

```
postTest.long <- PosTest.state %>%
pivot_longer (X2020.12.31: X2020.11.13, names_to = "DATE",
values_to = "Infected")
```

```
library(dplyr)
combined_data <- full_join(postTest.long, Test.long, by = "State")%>%
```

```
  group_by(State)
combined_data <- na.omit(combined_data)
#cumulative_count <- cumsum(ifelse(is.na('X2020.12.31x':'X2020.02.25x'/'X2020.12.31y':'X2020.02.25y
```

#(b)