ETC3250 Lab 6

Di Cook

SOLUTION

Purpose

This lab will be the first on exploring data using plots, using the grammar of graphics approch to constructing plots with the `ggplot2` R package.

Data examples

- Publicly available data on the zika virus was announced just this week. We will take a look at it.
- The PISA education data is one that we looked at last week, and we will revisit this data.

Zika

The zika virus has been prominent in the news for more than a year now, as the gravity of its impacts became clear. A new R package makes the incidence data available. This first exercise is to explore the data. To install the data and analysis package you would need to run these commands:

If the install does not work for you it is possible to download the data only from the web site: https://github.com/cpsievert/zikar.

There are three data sets. We want the `zika` and `latLonDat` R data frames.

What would we like to know?

1. Where are the zika incidences around the globe?
2. What is the trend of incidences?
3. Is the trend different at different locations? Are there emerging areas of incidence? Are some areas past the worst?

For each of these questions we need to work out how to make plots to address them:

1. Take the spatial coordinates and plot them on a map.
2. Aggregate counts by day, examine temporal trend by locations, and find the locations of the biggest outbreaks.
3. Find differences in the temporal trends over locations. Are some locations past the peak of the outbreak? Are some still in the main throes of zika?

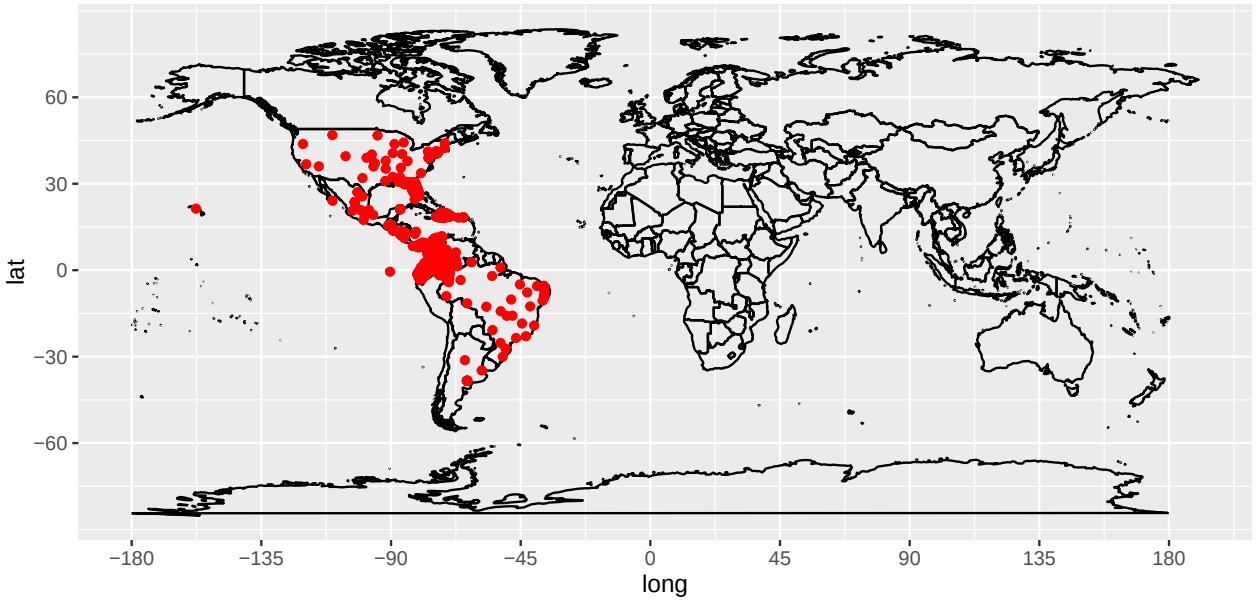
Quick plot, to show an issue with data:

An explanation from Carson Sievert, "If you go to the Colombia tab in the shiny app, it looks like Colombia reclassified basically everyone from confirmed to suspected, nationwide. Not sure why, but the folks at the CDC (which I'll be presenting this to next week) might know ;)"

We are going to handle this discrepancy by considering everything to be confirmed case, for today.

Exercise 1 Map locations

- Make a map of the world with the locations of zika incidence overlaid, using the code below.

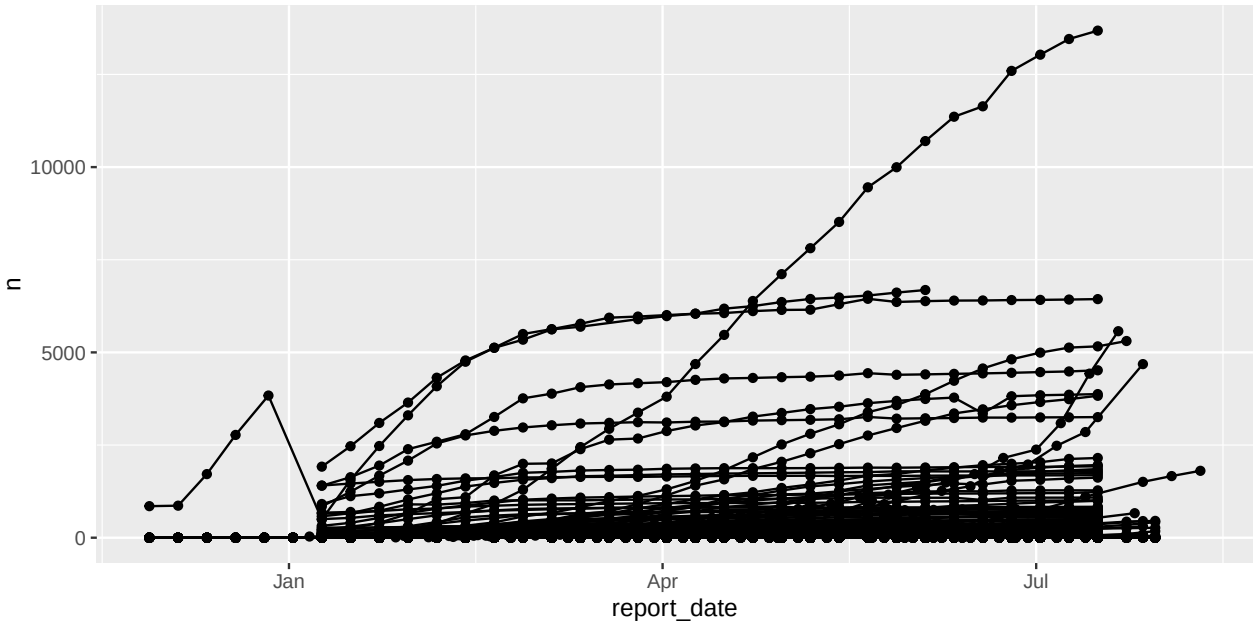


- What do you learn about the locations of zika? Does this match what is in the news? `It is only in the Americas. There are a lot more places in the USA than we would expect given the news, which only discusses Florida incidence. The history is outbreaks in Africa but there is no information in this data on outbreaks there.`

Since the incidence in this data is localised we could use google maps as the background to the locations.

Exercise 2 Examine temporal trend in counts Where are the biggest outbreaks?

We are going to use the `dplyr` package to aggregate the counts for each location by day.



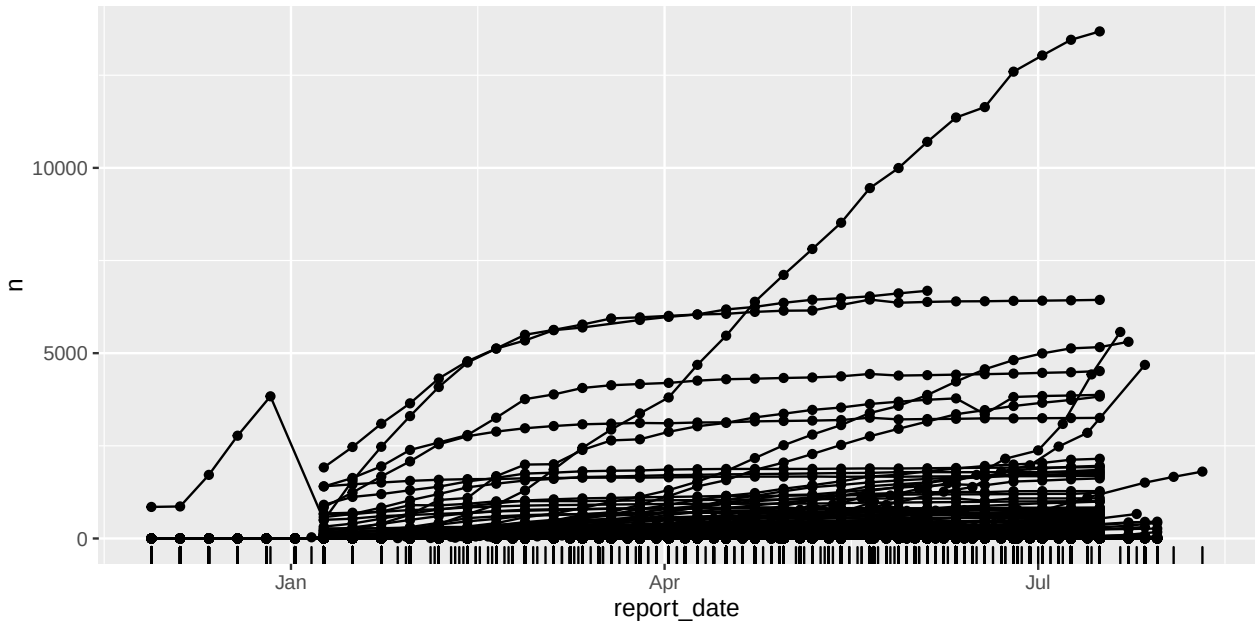
Make it interactive so that we can see the locations of the largest outbreaks.

- What do you learn about the incidence in different locations over time? `There are a handful of locations where the zika cases are really prevalent. Most locations are reporting few cases.`

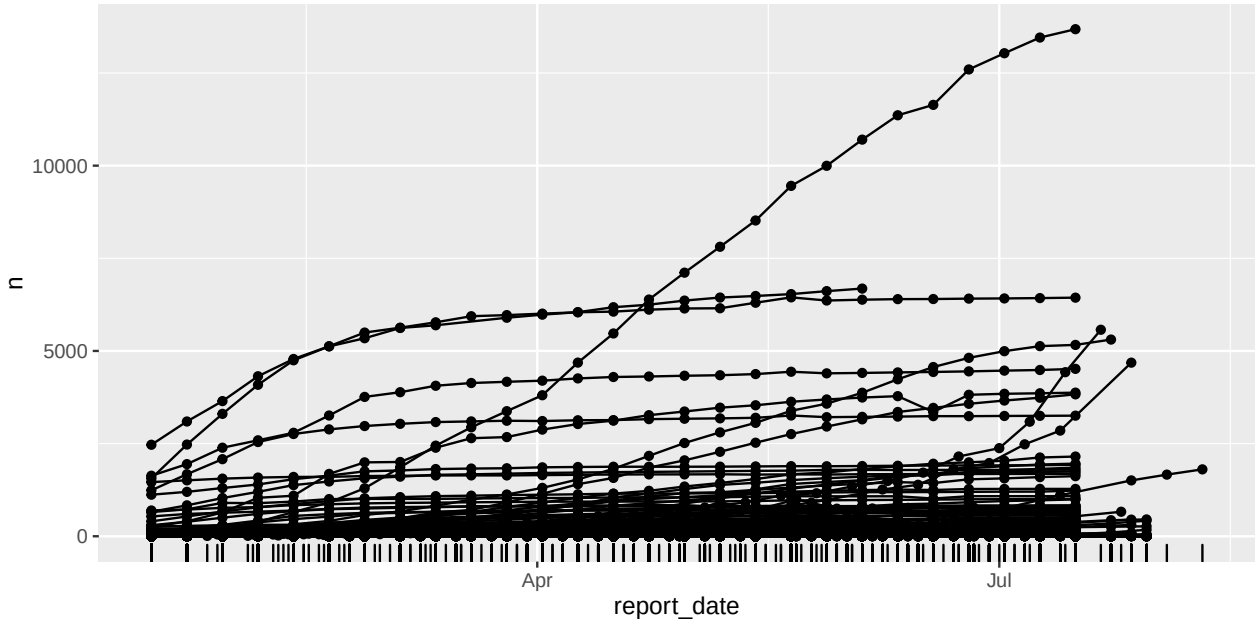
Exercise 3 What are the different patterns of outbreaks?

We are going to examine the trends in zika incidence.

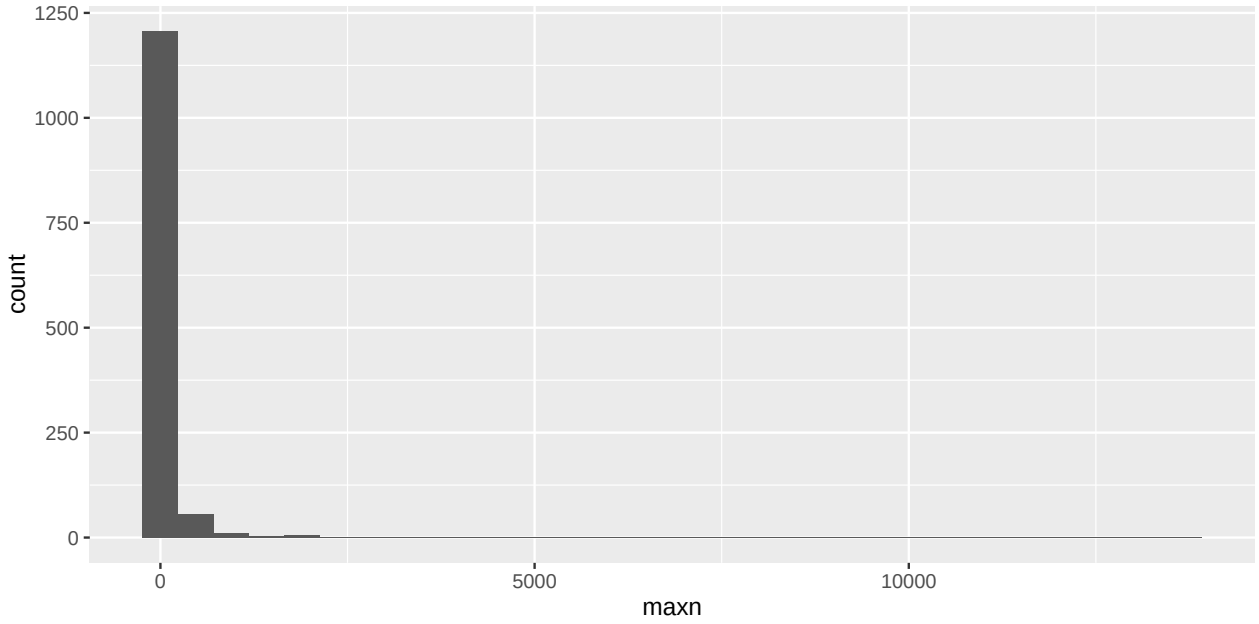
- The very first few measurements and the most recent look problematic. We will need to remove them before computing stats for each curve. First look at the temporal support, by adding a rug plot to the date axis. Are the measurements equidistant in time? `Clearly no, the days with measurements are a bit sporadic, especially early and late in the time period.`



- Trim off the early measurements. `Nothing to report on here.`

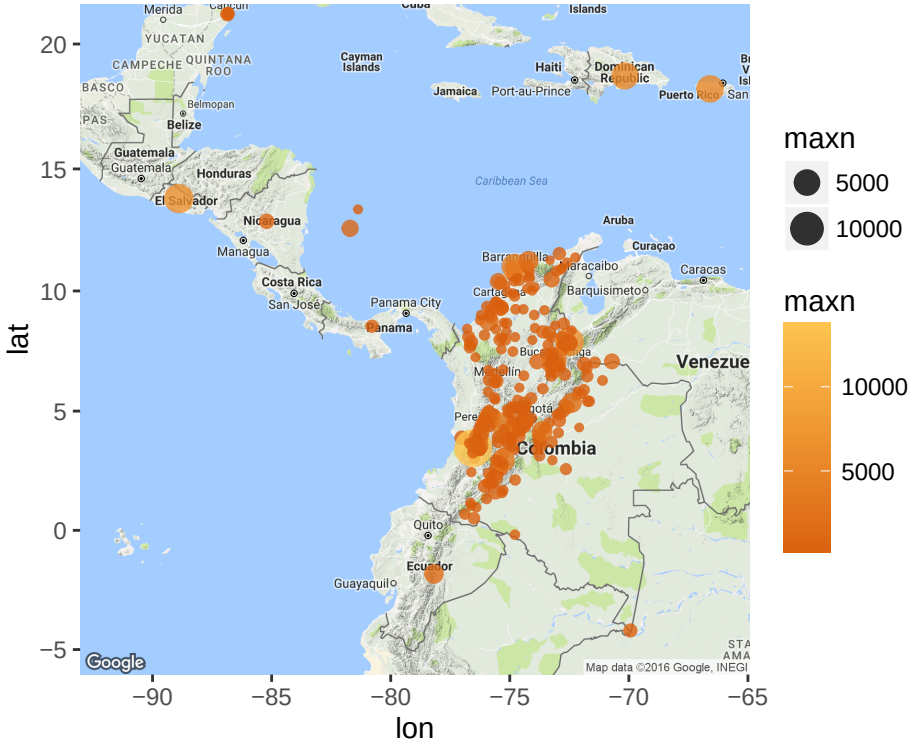


- Compute the highest daily count for every location. What do we learn about incidence for each location? `There are just a few locations with a lot of zika cases.`

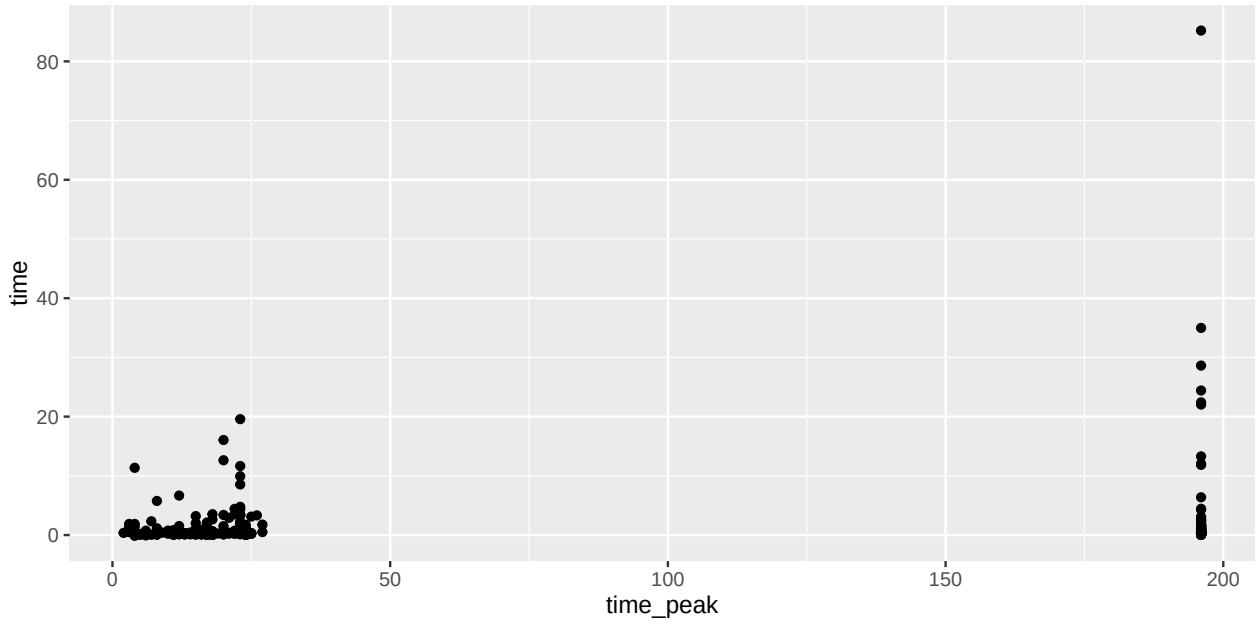



```
#    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
#     0.0     1.0     6.5   103.9    33.0 13680.0
```

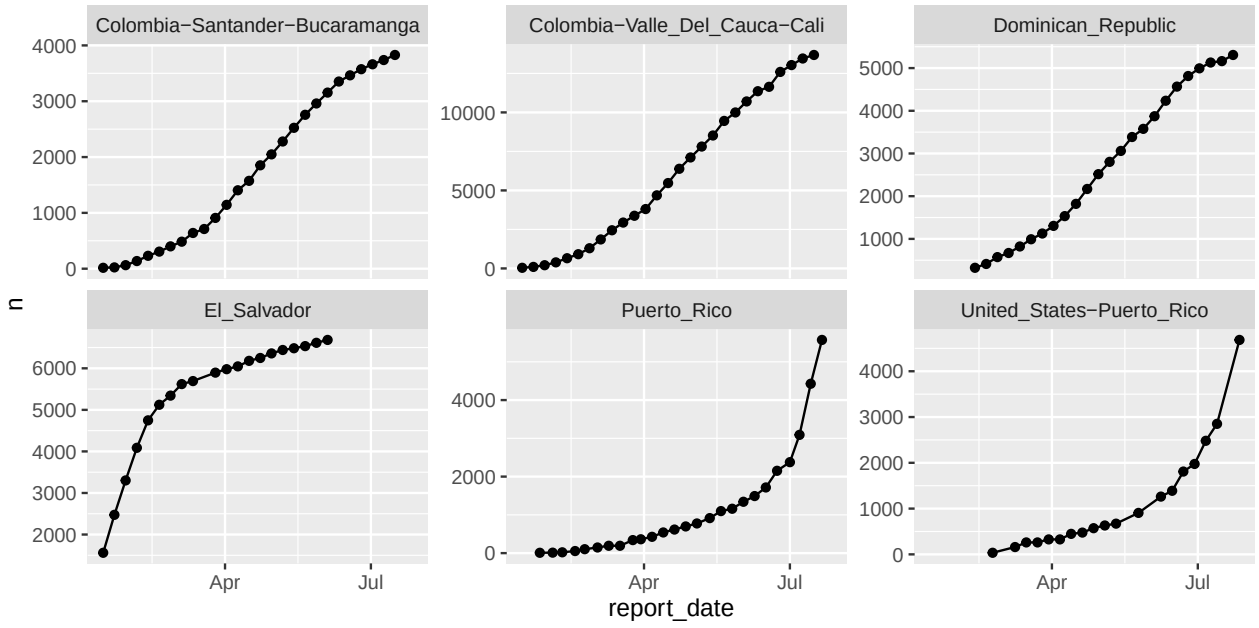
- Drop the locations with less than 33 incidences, and plot the hot spots. `Nothing to comment on here.`



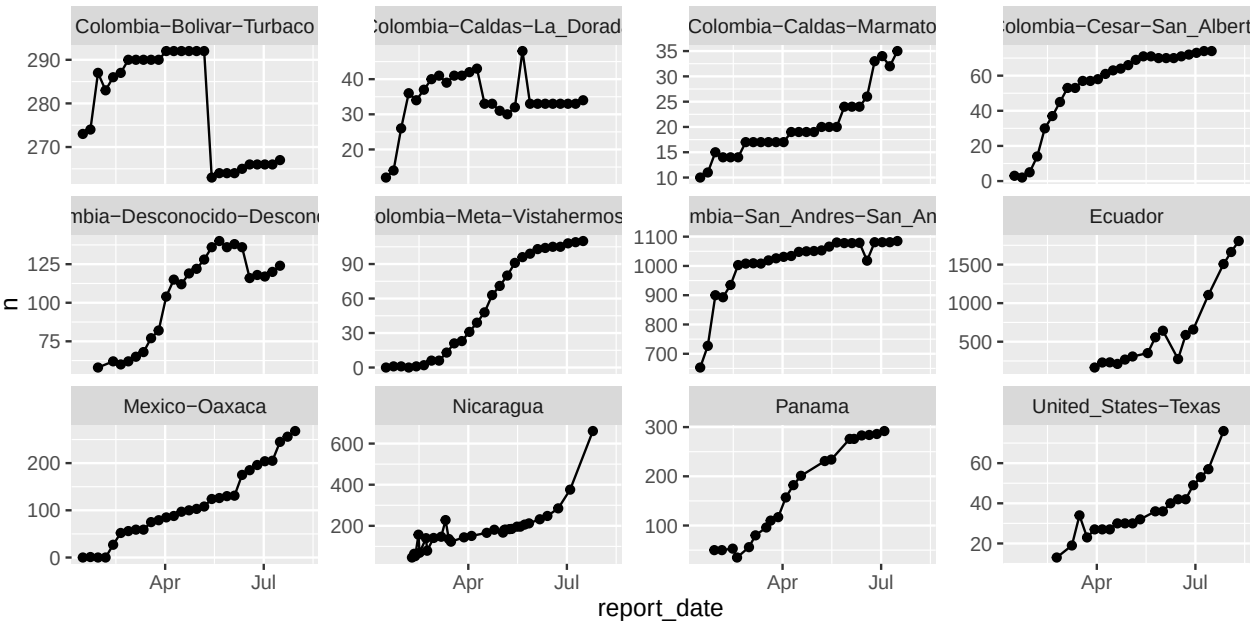
- Compute more statistics for each location. `Nothing to comment on here.`



- Look at the top locations based on trend. Which locations are still in the state of zika epidemic? `All of 6 locations are seeing an increasing number of cases. El Salvador might be tapering off.`



- Find the locations with the earliest peaks (the earliest 12). In which locations is the zika outbreak declining? Is this what you expected to find? `This is not what I expected. Most of these locations are still seeing and increase in cases. The early *peak* is basically a local anomaly.`



Exercise 4 Education

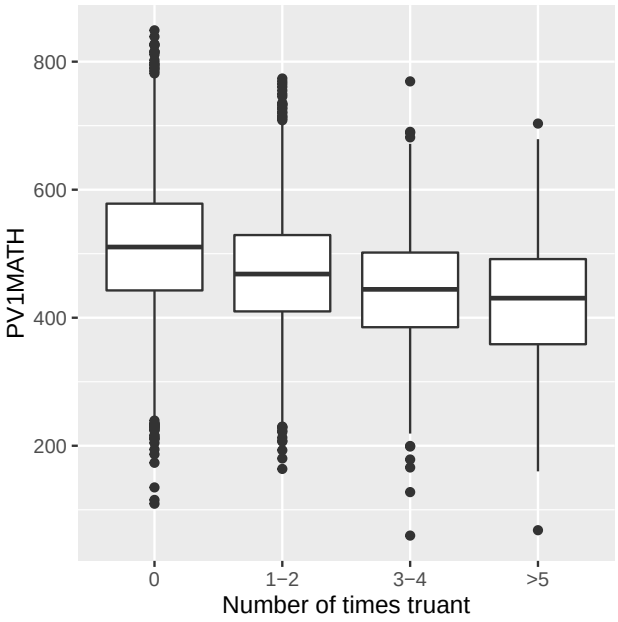
On the PISA data from last week. Make a plot to answer this question:

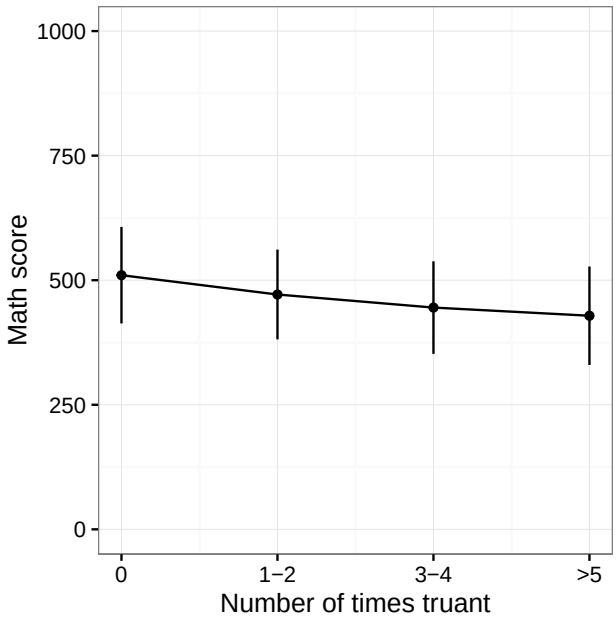
- Does truancy affect math score, on average?
- Explain your choices in the plot design. And answer the question.

Below you have the means and standard deviations of math scores against frequency of skipping school. Skipping school decreases the score on average by about 80 points once students skip school 5 or more times per month. There is a lot of variability, though, so some students who skip school 5 or more times per month still do better than some students who never skip school.

```
oz <- read.csv("../data/PISA-oz.csv") # You might need to change the directory to where the data is loc
dim(oz)
# [1] 14481    80
oz.sub <- filter(oz, !is.na(ST09Q01))
oz.sub$truancy <- factor(oz.sub$ST09Q01, levels=c("None  ", "One or two times  ", "Three or four times
p1 <- ggplot(data=oz.sub, aes(x=truancy, y=PV1MATH)) +
  geom_boxplot() +
  scale_x_discrete("Number of times truant",
                   labels=c("0", "1-2",
                            "3-4",
                            ">5"))
truancy <- summarise(group_by(oz.sub, truancy), m = mean(PV1MATH),  s = sd(PV1MATH))
truancy
# # A tibble: 4 x 3
#   truancy        m        s
#    <fctr>    <dbl>    <dbl>
# 1       0 510.0515 96.83555
# 2       1 471.2836 90.10048
```

```
# 3       2 445.0754 92.83297
# 4       3 428.6922 98.79920
truancy$truancy <- as.numeric(as.character(truancy$truancy))
p2 <- ggplot(data=truancy, aes(x=truancy, y=m)) +
  geom_point() + geom_line() +
  geom_linerange(aes(ymin=m-s, ymax=m+s)) + ylim(c(0,1000)) +
  scale_x_continuous("Number of times truant",
                     labels=c("0", "1-2",
                              "3-4",
                              ">5")) +
  ylab("Math score") + theme_bw()
library(gridExtra)
grid.arrange(p1, p2, ncol=2)
```



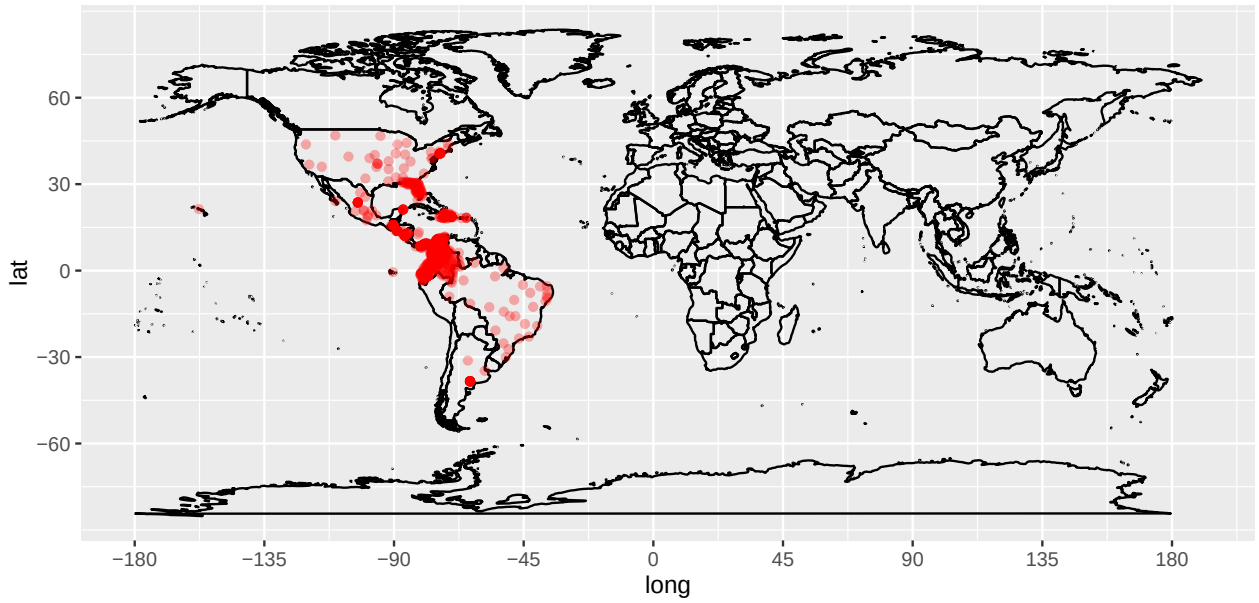


- Pick one other interesting question based on the data dictionary description information to answer using a plot. Make the plot, and summarise what you learn. `There will be variaous answers here.`

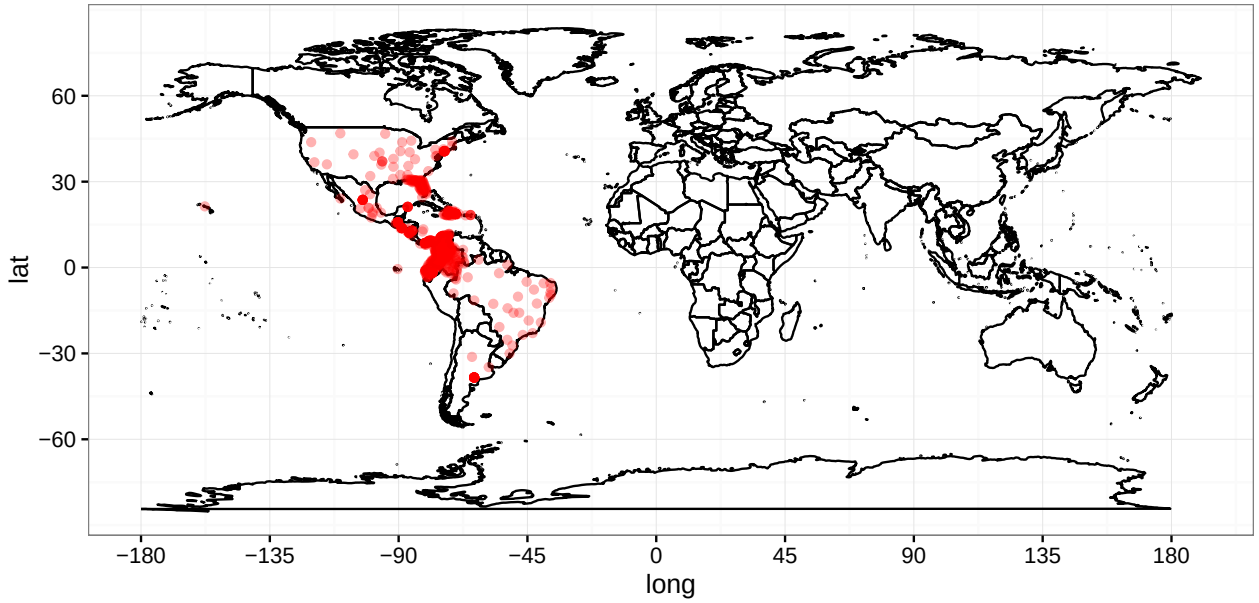
Coding exercises

- For the first map in exercise 1: Using your cheat sheet for `ggplot2` change the transparency of the points to examine the density, make the background of the map white, remove the axes and axis labels so that it looks more like a conventional map. Report your code for doing this.

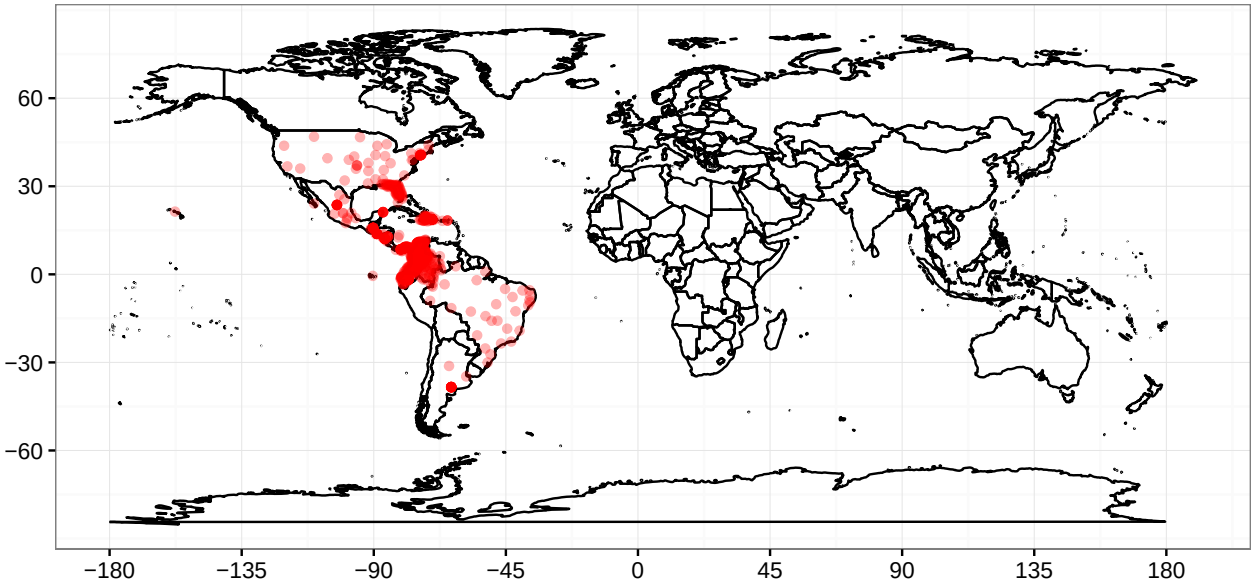
```
worldmap + geom_point(data=latLonDat, aes(x=lng, y=lat),
                      colour="red", alpha=0.3)
```



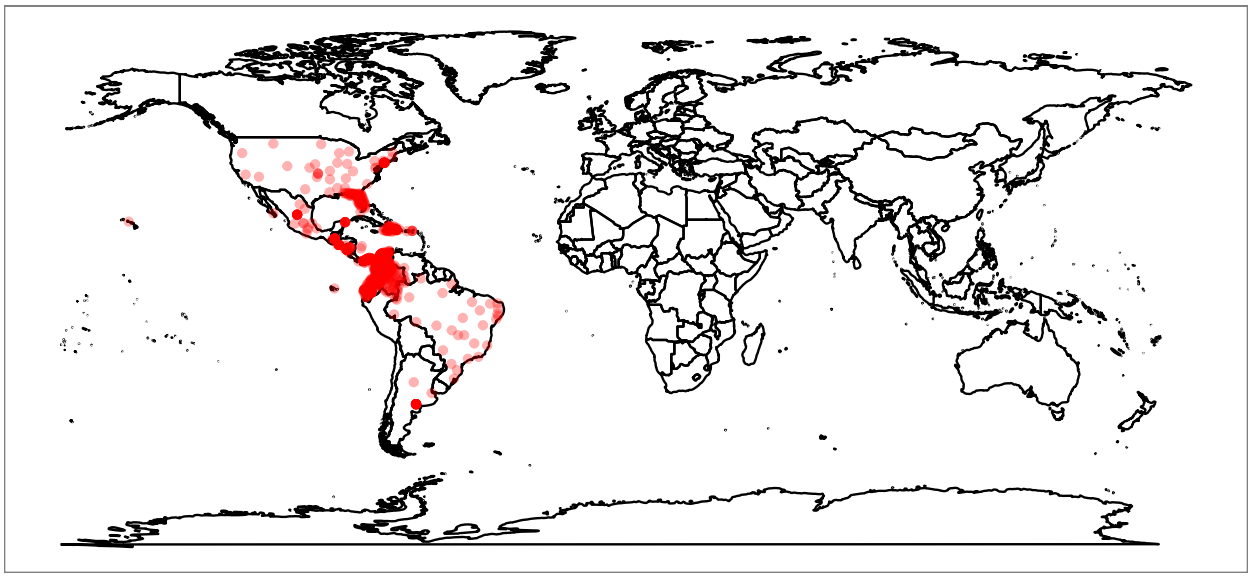

```
worldmap + geom_point(data=latLonDat, aes(x=lng, y=lat),
                      colour="red", alpha=0.3) + theme_bw()
```




```
worldmap + geom_point(data=latLonDat, aes(x=lng, y=lat),
                      colour="red", alpha=0.3) +
  xlab("") + ylab("") + theme_bw()
```

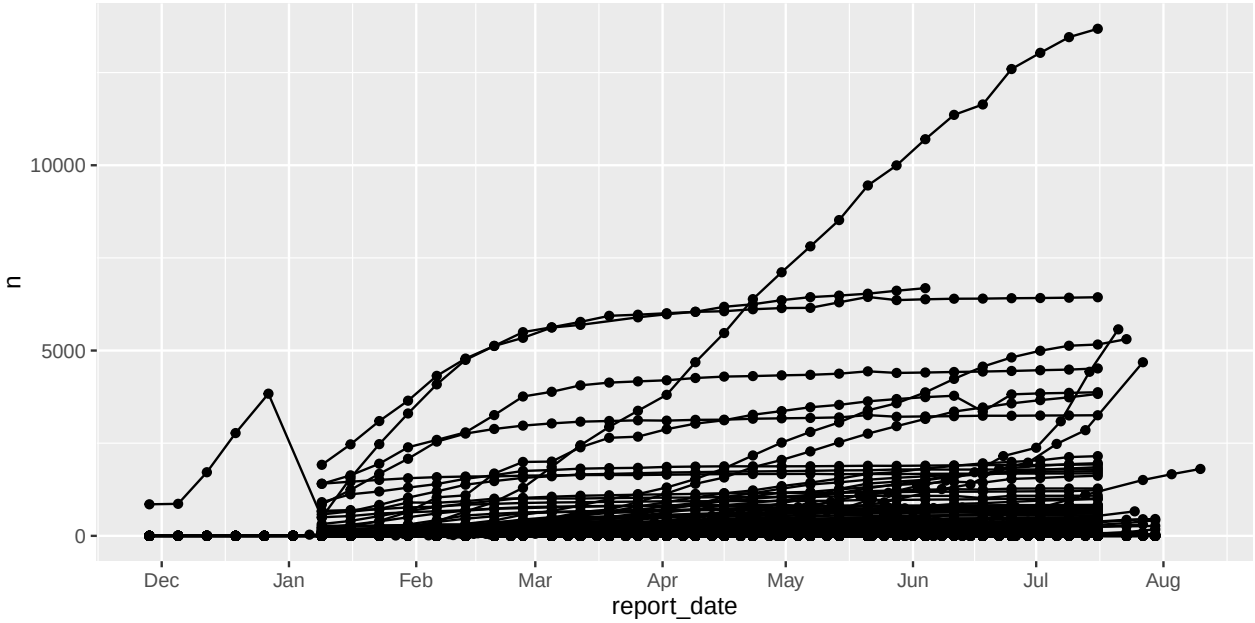



```
new_theme_empty <- theme_bw()
new_theme_empty$line <- element_blank()
new_theme_empty$strip.text <- element_blank()
new_theme_empty$axis.text <- element_blank()
new_theme_empty$plot.title <- element_blank()
new_theme_empty$axis.title <- element_blank()
worldmap + geom_point(data=latLonDat, aes(x=lng, y=lat),
                      colour="red", alpha=0.3) + new_theme_empty
```



- For the longitudinal plot in exercise 2: Change the axis labels to be more meaningful, change the date axis to show all months with the labels being a single letter, first letter of the month.

```
zika_smry <- zika %>% group_by(location, report_date) %>%
  summarise(n=sum(value, na.rm=T))
p <- ggplot(zika_smry, aes(x=report_date, y=n)) +
  geom_point() +
  geom_line(aes(group=location))
p + scale_x_date(date_breaks = "1 month", date_labels = "%b")
```

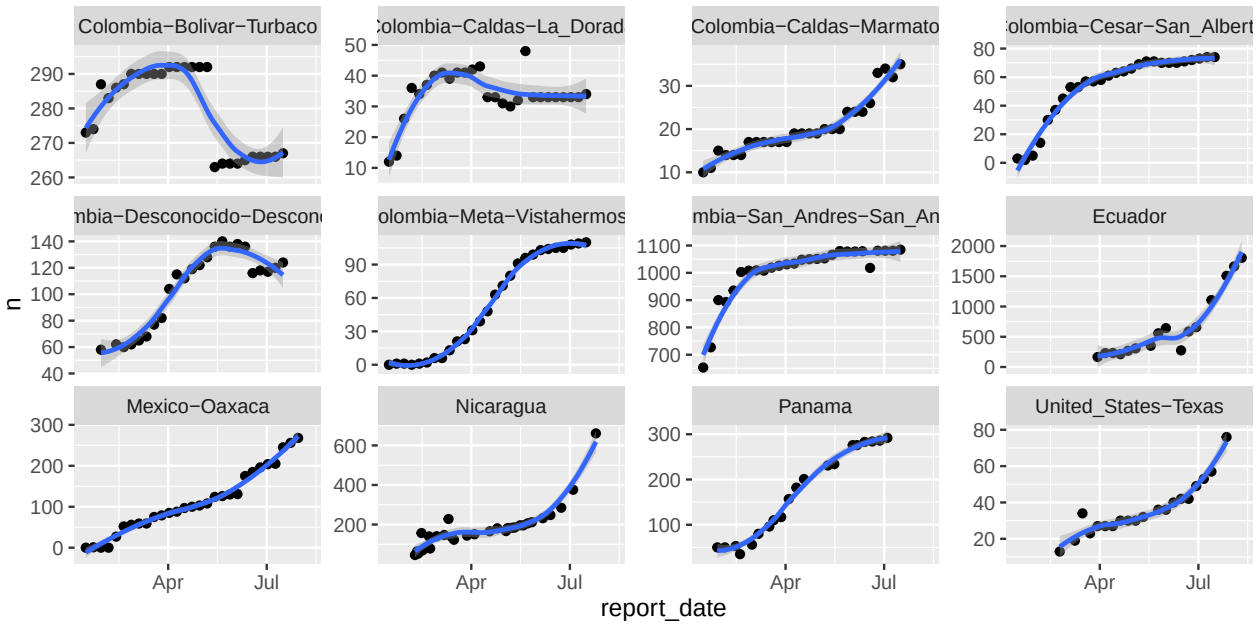


I don't know the solution to this one! I thought the code below would work, based on the help page, but it doesn't.

```
p + scale_x_date(date_breaks = "1 month",
 labels = c("D", "J", "F", "M", "A", "M", "J", "J", "A"))
```

- Change the plots in the plot of the 12 locations with the earliest peaks to have a fitted smoother to the data, and remove the lines that currenctly connect the dots.

```
p <- ggplot(zika_sub, aes(x=report_date, y=n)) +
  geom_point() +
  geom_smooth(aes(group=location)) +
  facet_wrap(~location, ncol=4, scales="free_y")
p
```



WHAT TO TURN IN

Turn in two items: a `.Rmd` document, and the output `.pdf` or `.docx` from running it. No need to include the R output and plots in your pdf, but the code should be in the Rmd file.

Resources

- RStudio cheat sheets
- ggplot2: Elegant Graphics for Data Analysis, Hadley Wickham, web site
- R Graphics Cookbook, Winston Chang