STAT 534: Homework 1 Solution

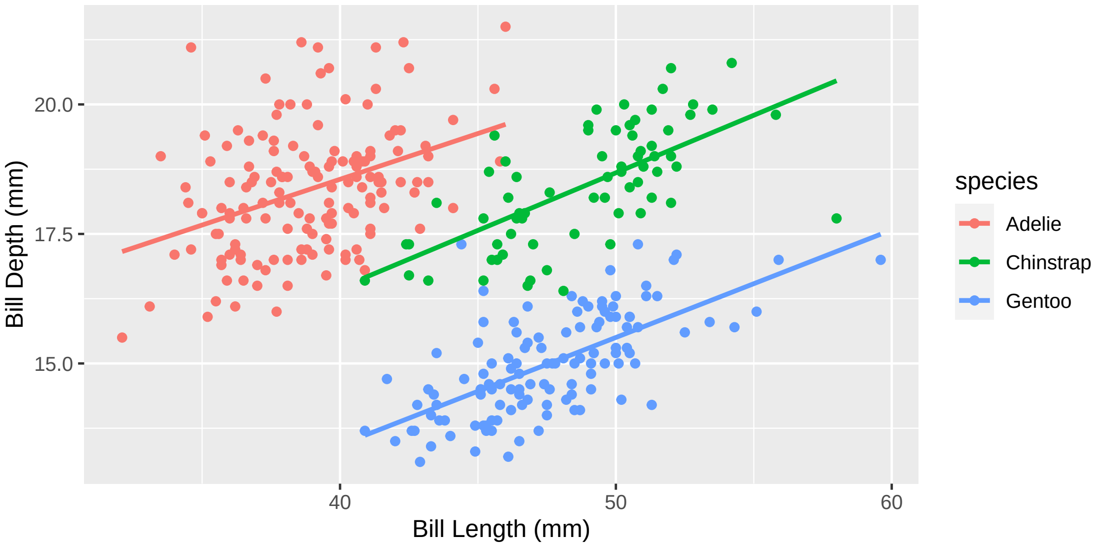
Fall 2021

Due: Monday, September 13, 12:30 pm

1. (Excerpted from Problem 3.8 in textbook) Using the *penguins* data set from the *palmerpenguins* package: (Hint: use na.omit(penguins) to remove cases with missing values.)

   (a) Create a scatterplot of bill_length_mm against bill_depth_mm where individual species are colored and a regression line is added to each species. What do you observe about the association of bill depth and bill length?

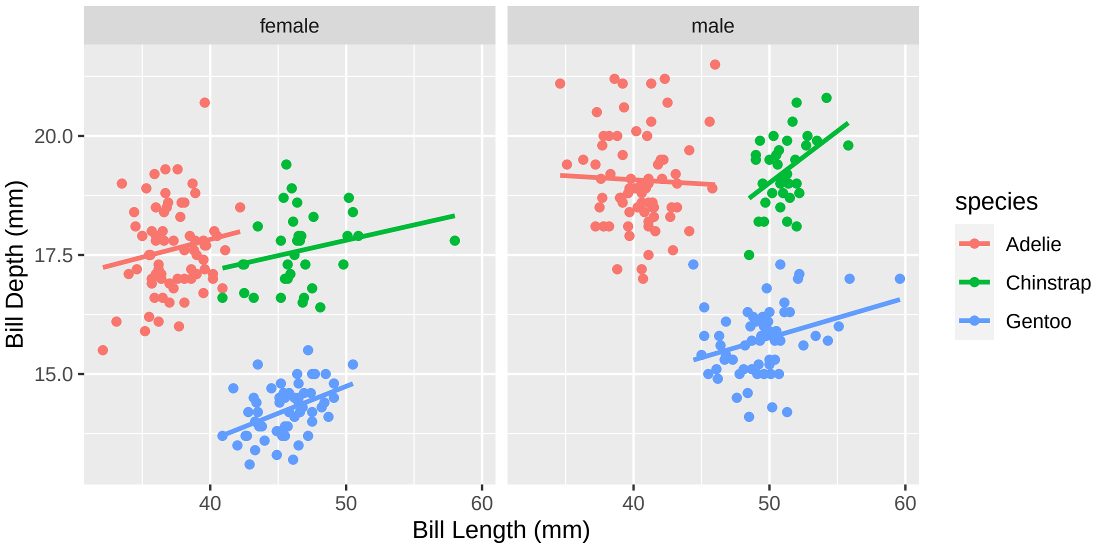
```
library("palmerpenguins")
ggplot(data = na.omit(penguins),
       mapping = aes(x = bill_length_mm,
                     y = bill_depth_mm,
                     color = species))  +
  geom_point() +
  geom_smooth(method = "lm", se = F) +
  labs(x="Bill Length (mm)", y="Bill Depth (mm)")
```



Bill depth and length are positively correlated and the relation behaves similarly across the three species since the regression lines are almost parallel.

   (b) Repeat the same scatterplot but now separate your plot into facets by sex. How would you summarize the association between bill depth and bill length?

```
library("palmerpenguins")
ggplot(data = na.omit(penguins),
       mapping = aes(x = bill_length_mm,
                     y = bill_depth_mm,
                     color = species))  +
  geom_point() +
  geom_smooth(method = "lm", se = F) +
  facet_wrap(~sex) +
  labs(x="Bill Length (mm)", y="Bill Depth (mm)")
```



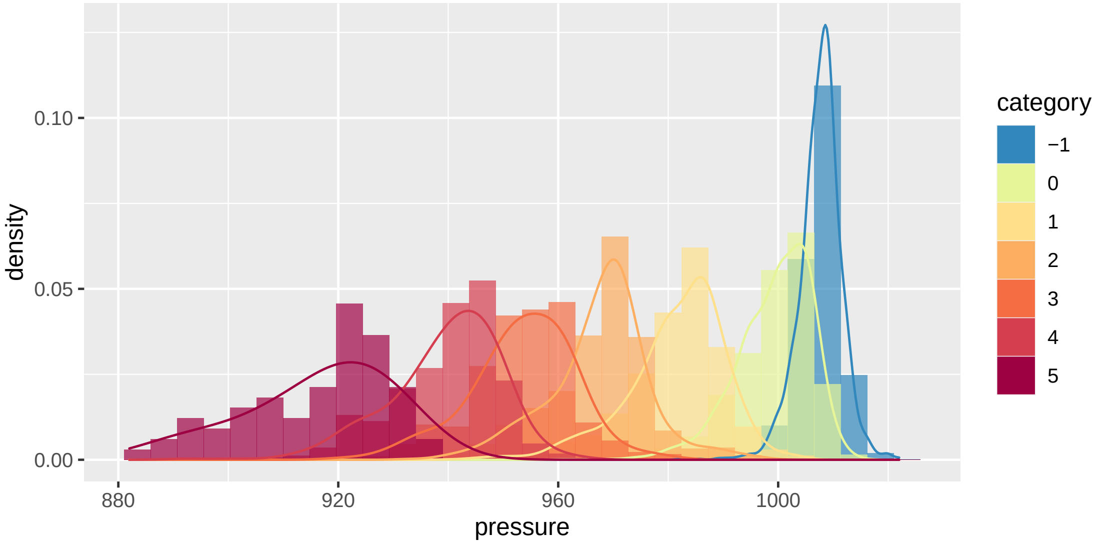
There is an interaction between species and sex embedded in the relation between bill depth and bill length. (Especially for Adelies, a deeper bill is not necessarily longer for a male bird.)

2. Using *storm* data from the *dplyr* package:

   (a) Produce a histogram of the pressure variable. Fill your bars using the category variable.

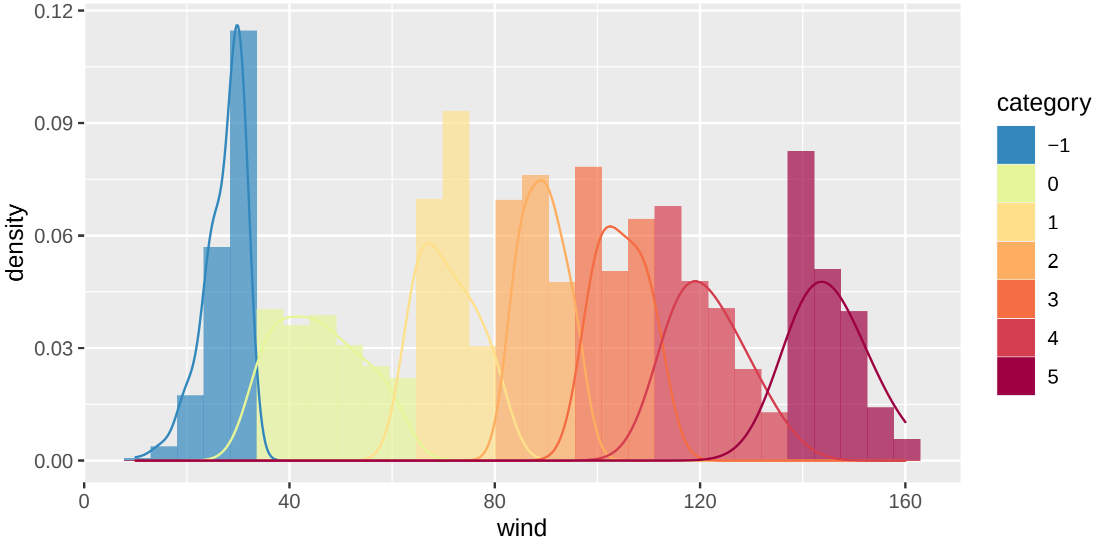
```
library(RColorBrewer)
palette_storm <- brewer.pal(11,"Spectral")[c(10,7,5:1)] # select color palette

ggplot(data = storms, mapping = aes(x = pressure)) +
  geom_histogram(mapping = aes(y=..density.., fill = category),
                 bins=30, alpha=0.7, position="identity") +
  geom_density(mapping = aes(color = category), adjust = 1.5) +
  # use position="identity" and overlay density curves to alleviate overlaps
  scale_fill_manual(values = palette_storm) +
  scale_color_manual(values = palette_storm)
```



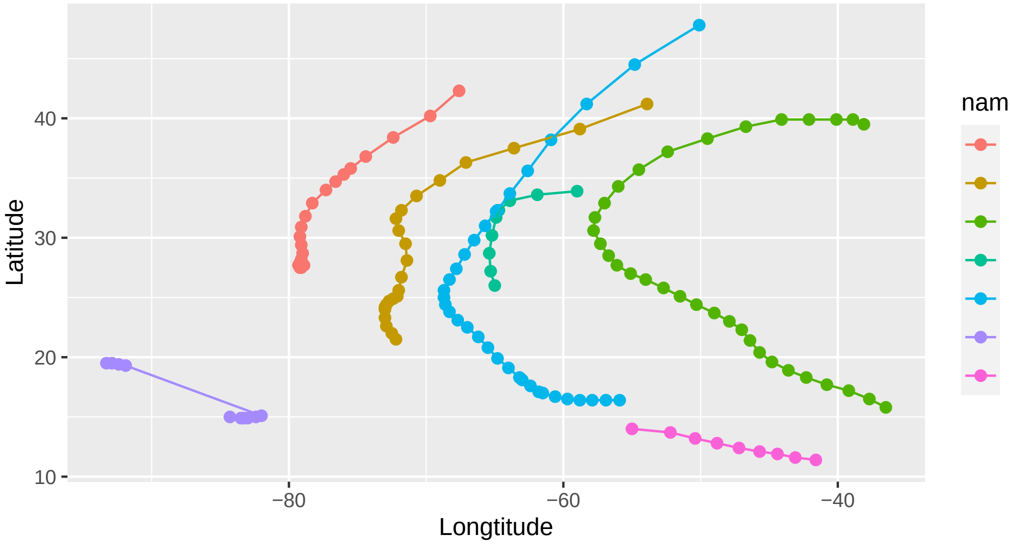
(b) Repeat part (a) with the wind speed variable.

```
ggplot(data = storms, mapping = aes(x = wind)) +
  geom_histogram(mapping = aes(y=..density.., fill = category),
                 bins=30, alpha=0.7, position="identity") +
  geom_density(mapping = aes(color = category), adjust = 3) +
  scale_fill_manual(values = palette_storm) +
  scale_color_manual(values = palette_storm)
```



(c) Use `geom_path()` to plot the path of each tropical storm in 2014. Use color to distinguish the storms from one another. Which storm in 2014 made it the furthest North?

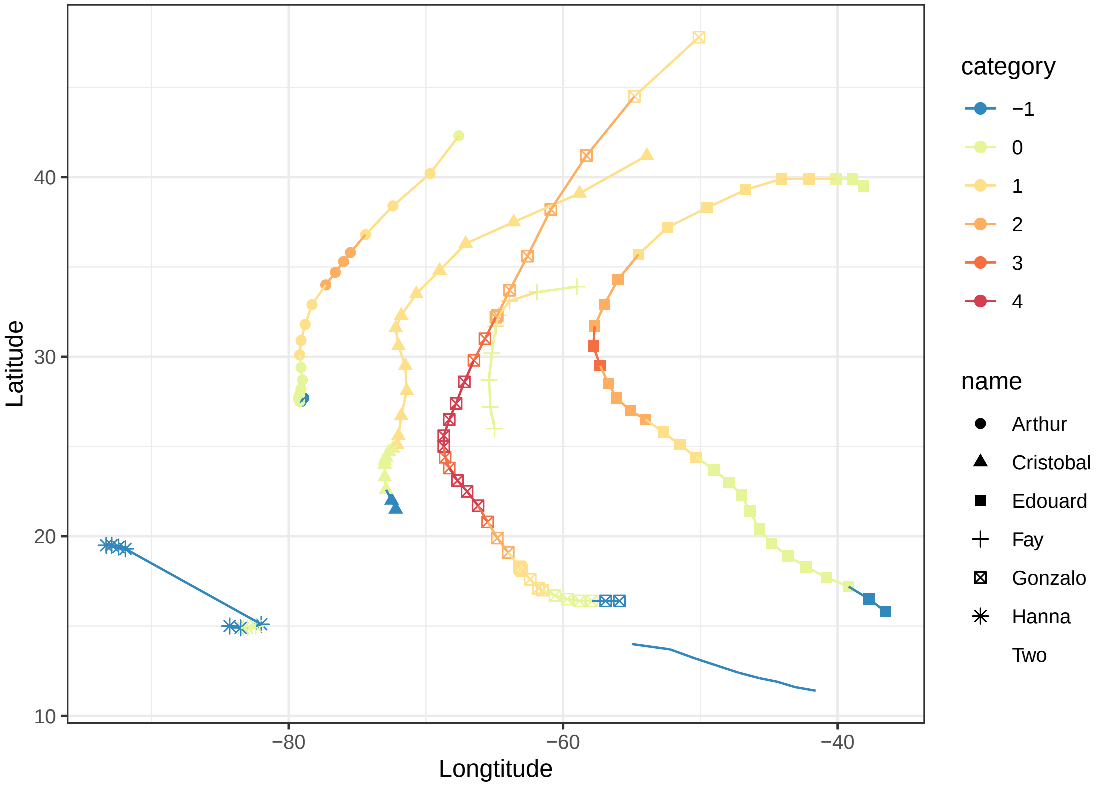
```
ggplot(data = subset(storms, year==2014),
       mapping = aes(x = long, y = lat, color = name)) +
  geom_point(size = 2) +
  geom_path() +
  labs(x="Longtitude", y="Latitude")
```



Gonzalo made it the furthest north.

(d) Challenge: show changes in the category variable along the paths in part (c). (Hint: group the data by names and add another aesthetic that maps category.)
Remarks: The group aesthetic is by default set to the interaction of all discrete variables in the plot. This choice often partitions the data correctly, but when it does not, or when no discrete variable is used in the plot, you will need to explicitly define the grouping structure by mapping group to a variable that has a different value for each group.

```
ggplot(data = subset(storms, year==2014),
       mapping = aes(x = long, y = lat,
                     color = category, # use color to show category changes
                     shape = name, # use shape to distinguish storms
                     group = name)) + # data should be grouped by storm
  geom_point(size = 2) +
  geom_path() +
  scale_color_manual(values = palette_storm) +
  labs(x="Longtitude", y="Latitude") +
  theme_bw()
```



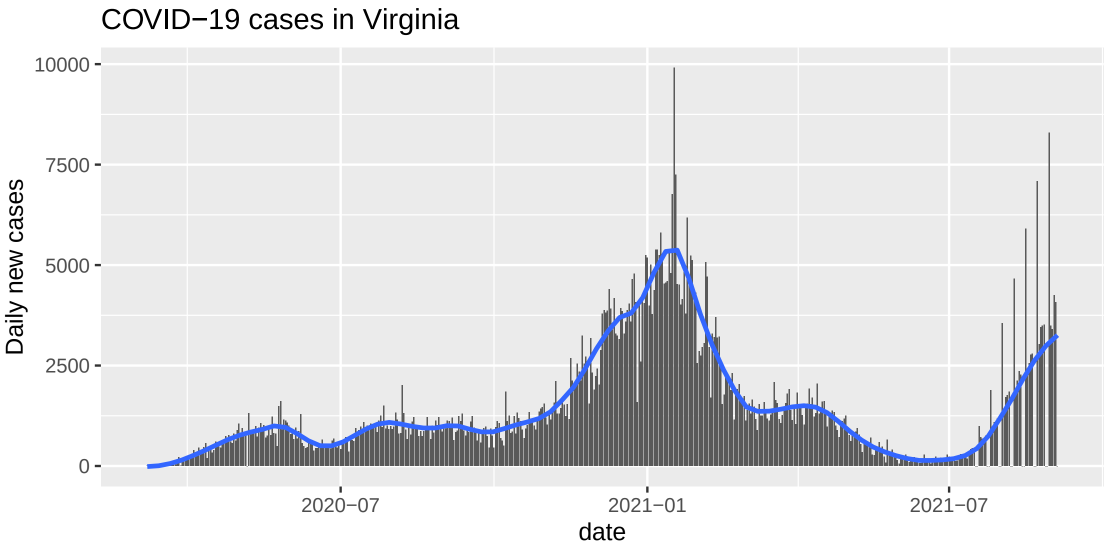
In this case, the group aesthetic is by default set to the interaction of name and category for points and paths. Although we add a new aesthetic to reflect changes in category, points and paths should still be grouped by storm.

3. COVID-19 vignette

(a) Use data in *us_cases.txt* (data source: New York Times repository of COVID-19 data) to reproduce Figure (a) with geom_col() and geom_smooth(span = 0.1, method = ”loess”, se = F).

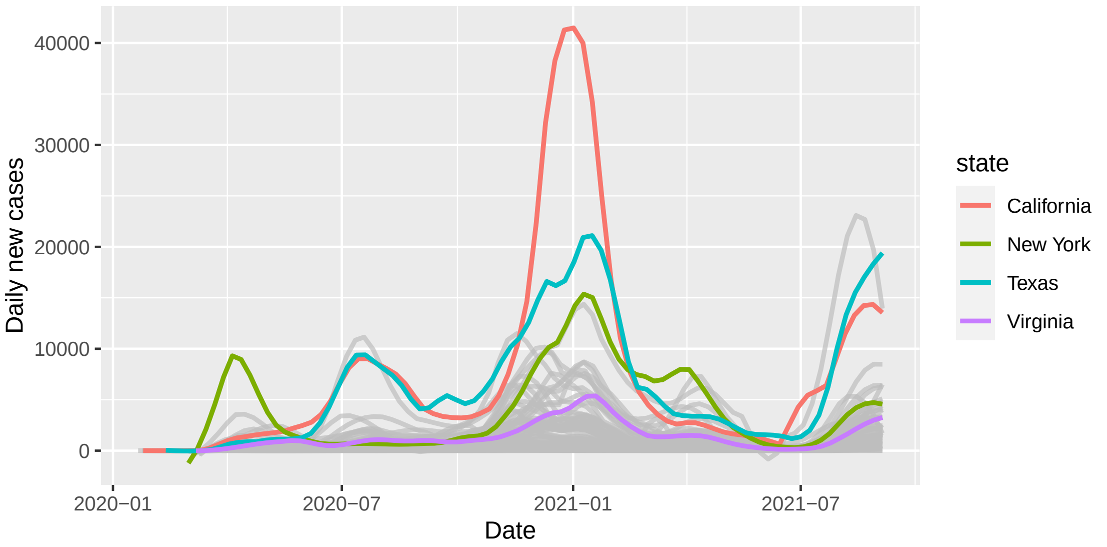
```
library(lubridate)
us_cases <- read.table("/Volumes/GoogleDrive/My Drive/STAT 534 F21/CH1.1visualization/HW1/us_ca
                       header = T, sep = " ")
```

```
us_cases <- us_cases %>%
  mutate(date = ymd(date),
         state = as.factor(state))
us_cases %>%
  filter(state == "Virginia") %>%
  ggplot(aes(x = date, y = daily_cases)) +
  geom_col() +
  geom_smooth(span = 0.1, method = "loess", se = F) +
  labs(title = "COVID-19 cases in Virginia", y = "Daily new cases")
```



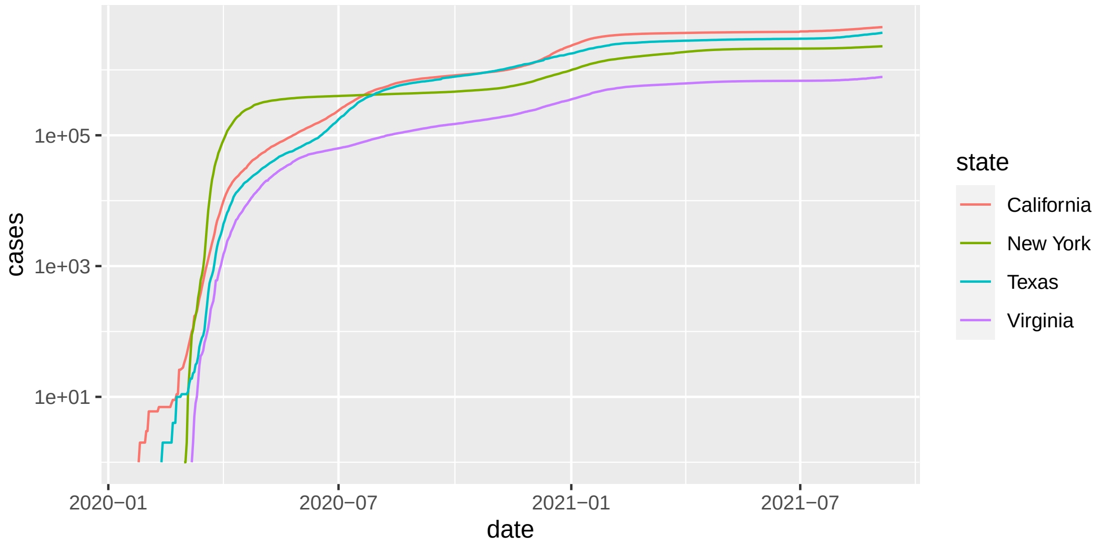
(b) Challenge: reproduce Figure (b) with `gghighlight()` function in the *gghighlight* package.

```
bigstates <- c("California", "New York", "Texas","Virginia")
us_cases %>%
  ggplot(aes(x = date, y = daily_cases, color = state)) +
  geom_smooth(span = 0.1, se = FALSE) +
  gghighlight::gghighlight(state %in% bigstates) +
  labs(
    x = "Date", y = "Daily new cases"
  )
```



(c) When tracking a disease, the rate of growth is particularly important, and is proportional to the logarithm of the case count. Reproduce the figure below with `scale_y_log10(labels = scales::comma)`.

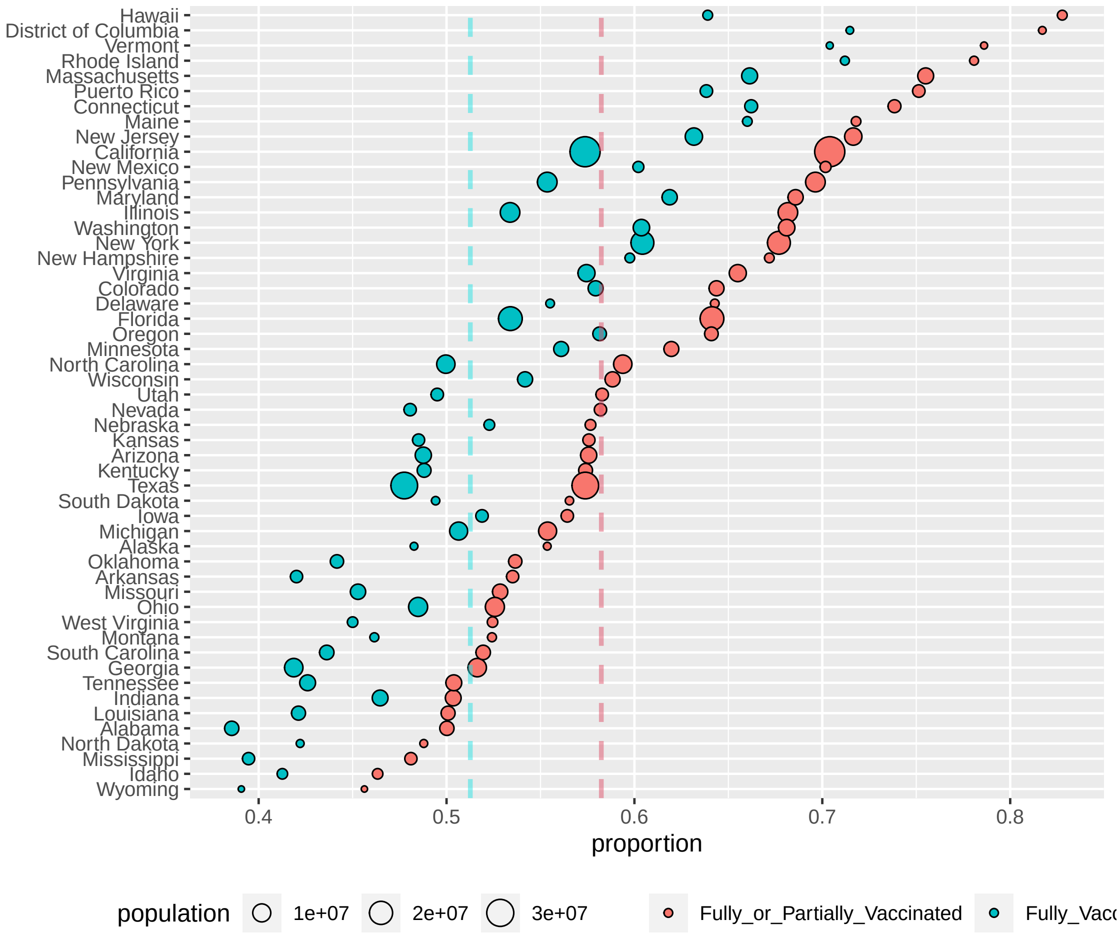
```
us_cases %>%
  filter(state %in% bigstates) %>%
  ggplot(aes(x = date, y = cases, color = state)) +
  geom_line() +
  scale_y_log10()
```



(d) Use data in *vaccine.txt* (data source: Johns Hopkins repository of COVID-19 data) to create an informative graphic that demonstrates the vaccination rollout across the US at the current stage. You may follow examples at Johns Hopkins coronavirus resource center or create your own plot.

```
vac <- read.table("/Volumes/GoogleDrive/My Drive/STAT 534 F21/CH1.1visualization/HW1/vaccine.t
                  header = T)
# calculate proportion
```

```
vac <- vac %>%
  mutate(Fully_Vaccinated = People_Fully_Vaccinated/population,
         Fully_or_Partially_Vaccinated = (People_Fully_Vaccinated+People_Partially_Vaccinated)/
# transform data for creating the bubble chart
vac1 <- vac %>%
  dplyr::select(-People_Fully_Vaccinated, -People_Partially_Vaccinated) %>%
  pivot_longer(cols = Fully_Vaccinated:Fully_or_Partially_Vaccinated,
               names_to = "ind", values_to = "prop")
# bubble chart
ggplot(data = vac1, mapping = aes(x = reorder(state, prop, max), y = prop,
                                  size = population, fill = ind)) +
  geom_point(shape = 21) +
  geom_hline(yintercept = median(vac$Fully_Vaccinated),
             linetype = "dashed", color = 5, size = 1, alpha = 0.5) +
  geom_hline(yintercept = median(vac$Fully_or_Partially_Vaccinated),
             linetype = "dashed", color = 2, size = 1, alpha = 0.5) +
  coord_flip() +
  labs(x = NULL, y = "proportion", size = "population", fill = "") +
  theme(legend.position = "bottom")
```



Note there is no standard solution to this problem.