Project 5

Read in the dataset you will be working with:

```
US_states <- readRDS(url("https://github.com/wilkelab/SDS375/blob/master/datasets/US_states.rds?raw=tru
  rename(state = name)

colony <- readr::read_csv('https://raw.githubusercontent.com/rfordatascience/tidytuesday/master/data/20

bee_colony <- colony %>%
  filter(state != "United States") %>%
  filter(state != "Other States")

bee_colony
```

```
## # A tibble: 1,170 x 10
##     year months    state colon~1 colon~2 colon~3 colon~4 colon~5 colon~6 colon~7
##    <dbl> <chr>     <chr>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>
##  1  2015 January-~ Alab~    7000    7000    1800      26    2800     250       4
##  2  2015 January-~ Ariz~   35000   35000    4600      13    3400    2100       6
##  3  2015 January-~ Arka~   13000   14000    1500      11    1200      90       1
##  4  2015 January-~ Cali~ 1440000 1690000  255000      15  250000  124000       7
##  5  2015 January-~ Colo~    3500   12500    1500      12     200     140       1
##  6  2015 January-~ Conn~    3900    3900     870      22     290      NA      NA
##  7  2015 January-~ Flor~  305000  315000   42000      13   54000   25000       8
##  8  2015 January-~ Geor~  104000  105000   14500      14   47000    9500       9
##  9  2015 January-~ Hawa~   10500   10500     380       4    3400     760       7
## 10  2015 January-~ Idaho   81000   88000    3700       4    2600    8000       9
## # ... with 1,160 more rows, and abbreviated variable names 1: colony_n,
## #   2: colony_max, 3: colony_lost, 4: colony_lost_pct, 5: colony_added,
## #   6: colony_reno, 7: colony_reno_pct
```

Provide more information about the dataset here.

Question: In this project, we attempt to answer two questions related to variation in Bee colony renovation % across different seasons, understand the trend in % of lost colonies and see if they are related to each other. They are:

- How does the colony renovation % change among the Bees in the US states for the year 2018?
- What is the relationship between colony renovation and colonies lost % across the years?

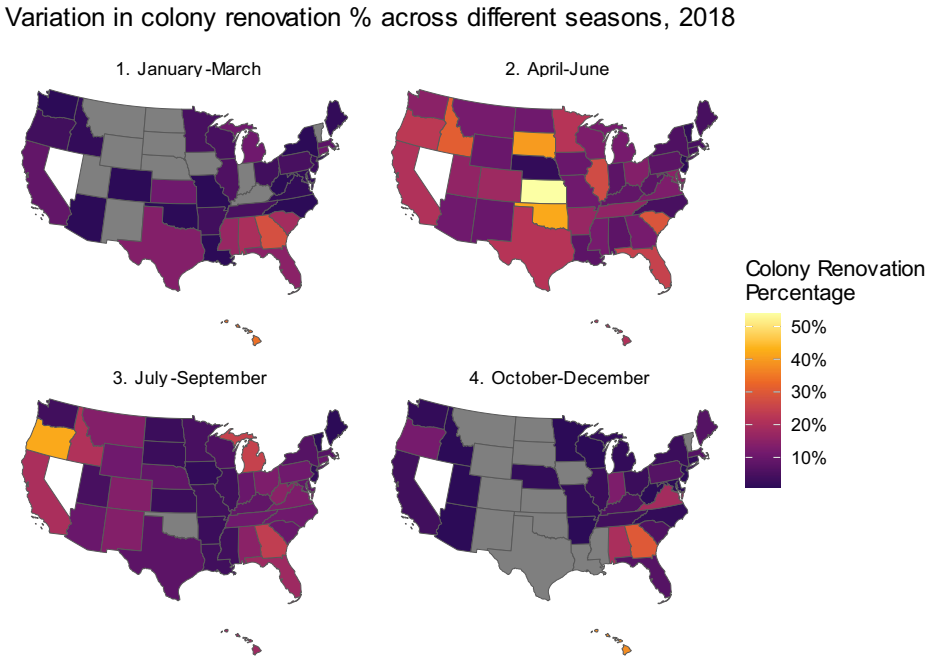
Introduction: The Honey Bee Colonies data by USDA has collected a comprehensive dataset called `bee_colony`, consisting of detailed information on honey bee colonies in terms of number of colonies, maximum, lost, percent lost, added, renovated, and percent renovated, as well as colonies lost with Colony Collapse Disorder symptoms with both over and less than five colonies. The data for operations with honey bee colonies are collected from a stratified sample of operations that responded as having honey bees on the Bee and Honey Inquiry and from the NASS list frame. There are 1170 rows, with each row representing an observation of months across each state.
We will left join the `bee_colony` with the `US_states` dataset to get the *Geospatial Analysis* and then examine the state (column `state`), the year in which observation was recorded (column `year`), and the months in which the record was observed (column `months`) for the bee colonies for the year 2018. In answering the second question, we will investigate `colony_lost_pct` for a few states and `year` across different months. It is important to include all rows that meet the relevant filters since each row represents a unique record in time.

Approach: To address the first question, we need to perform some data wrangling. Specifically, we will join the `bee_colony` data with `US_States` data to get geometry variables for plotting map and filter the data by `year`. We will then use the `case_when()` function to recode `months` into proper terms with a number to be printed in order. We will then use ggplot() with `geom_sf()` to plot the geospatial analysis and also use `girafe` function to make the plot interactive by linking each state's wiki page respectively.
To address the second inquiry, we commence by filtering the data from the year 2019 as all the data in that year for a particular season was NaN. We then choose to focus on the variables `state`, `year`, and `colony_lost_pct` Using the `transition_reveal()` function, we construct an interactive line graph to visually examine the trends. We plot this graph for four states observed through the geospatial analysis which have significant variation in colony renovation from the first question.

Analysis:

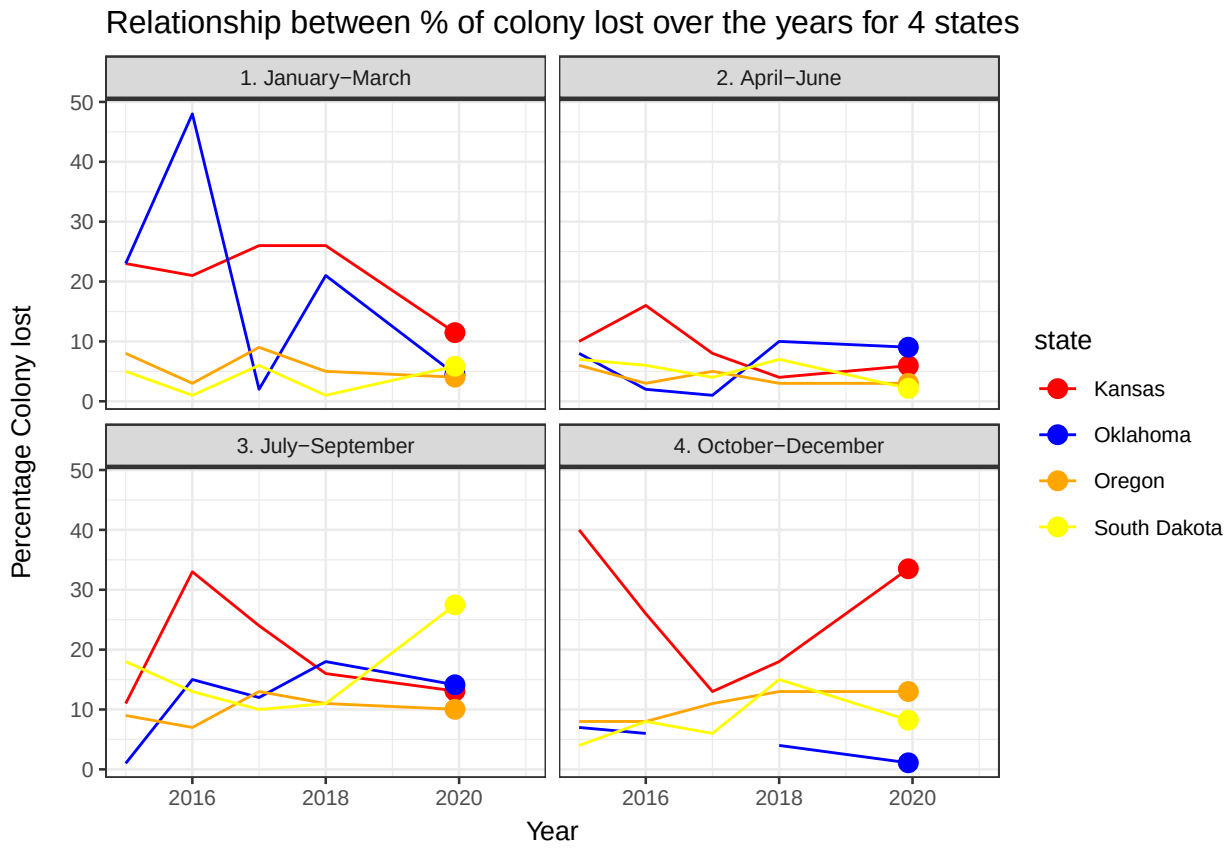
```
US_bee_map <- US_states %>%
  left_join(bee_colony, by = "state") %>%
  filter(year == 2018) %>%
  mutate(
    months = case_when(
      months == "January-March" ~ "1. January-March",
      months == "April-June" ~ "2. April-June",
      months == "July-September" ~ "3. July-September",
      months == "October-December" ~ "4. October-December",
      TRUE ~ NA_character_
    ),
    onclick = glue::glue('window.open(
      "https://en.wikipedia.org/wiki/{state}")')
  ) %>%
  ggplot(aes(fill = colony_reno_pct), size = 0.1) +
    geom_sf_interactive(
    aes(
      data_id = state, tooltip = state,
      onclick = onclick
    )
  ) +
  scale_fill_viridis_c(
    name = "Colony Renovation \nPercentage",
    option = "B",
    begin = 0.15,
    label = scales::label_percent(scale = 1)
  ) +
  facet_wrap(vars(months), nrow = 2) +
  labs(title = ("Variation in colony renovation % across different seasons, 2018\n")) +
  theme_void()

girafe(
  ggobj = US_bee_map,
  width_svg = 7.5,
  height_svg = 7.5 * 0.618,
  options = list(
    opts_tooltip(css = "background: #F5F5F5; color: #191970;")))
```




```
df <- bee_colony %>%
  filter(year != 2019) %>%
  mutate(
    months = case_when(
      months == "January-March" ~ "1. January-March",
      months == "April-June" ~ "2. April-June",
      months == "July-September" ~ "3. July-September",
      months == "October-December" ~ "4. October-December",
      TRUE ~ NA_character_
    )) %>%
  filter(state %in% c("Kansas", "Oklahoma", "South Dakota", "Oregon"))

ggplot(df, aes(x = year, y = colony_lost_pct, group = state, color = state)) +
  geom_line() +
  geom_point(size = 3) +
    scale_color_manual(values = c("red", "blue", "orange", "yellow")) +
  labs(title = "Relationship between % of colony lost over the years for 4 states",
       x = "\nYear",
       y = "Percentage Colony lost \n\n") +
  facet_wrap(vars(months)) +
  theme_bw(10) +
  transition_reveal(year)
```


Relationship between % of colony lost over the years for 4 states
1. January−March
2. April−June
3. July−September
4. October−December
Percentage Colony lost
Year
0
10
20
30
40
50
2016
2018
2020
state
Kansas
Oklahoma
Oregon
South Dakota

Discussion:
Plot1 - Variation in colony renovation % across different seasons, 2018:
When we look at the variation in colony renovation % across different seasons for the 2018, we see that there is a high colony renovation % in the months *April-June*. This can be correlated to the fact that bees collect more pollen/produce more honey during spring season. But sharp after this spring season, the renovation % falls to around 10-20% and the renovation is almost none to less in between October-March.
With this we can observe that bees renovate their colonies over the April-June months. The plot is animated/interactive and thus we can click on the states to learn more about them.
Plot2 - Relationship between % of colony lost over the years for 4 states:
Here in the 2nd plot we try to understand if there is any relation between colony lost % and colony renovated % across the years.
As per our first plot, it was understood that bees start their colony renovation in the early months of Jan-Mar and then the % is clearly visible in the graph1. Similarly, there is coincidence to colony lost % in the initial months of Jan-Mar. Even after the spring i.e, April-June months, there is a significant loss in colony and the loss substantiates over in the last 3 months of the year.