Ananlysis of Bee colony losses in the Unites States

```
# Get the Data
colony <- readr::read_csv('https://raw.githubusercontent.com/rfordatascience/tidyt
stressor <- readr::read_csv('https://raw.githubusercontent.com/rfordatascience/tid
```

The data comes from the USDA and provides information on honey bee colonies in the United States.

Question:

What are the major stressors that have led to the loss of bee colonies and which states in the United States have experienced the maximum loss of bee colonies?

Introduction:

We will use the Bee colonies datasets `colony` and `stressors` for our analysis. The dataset `colony` contains 1222 records on various aspects of honey bee colonies such as the number of colonies, maximum number of colonies, number and percentage of lost colonies, added colonies, number and percentage of renovated colonies in different states of the US from 2015 to 2021. It also includes information on colonies lost in different states over the years and months. The `stressor` dataset contains 7332 records and lists the different colony health stressors and their respective contribution percentages to colony loss in different states, and time frames.

Our focus is on identifying the primary stressors that pose the greatest threat to bee colonies, as well as determining which states in the United States have experienced the most significant percentage of bee colony losses.

We will be using the following columns from the `colony` dataset.

Variable	Class	Description
year	character	year
months	character	month

Variable	Class	Description
state	character	State Name
colony_max	integer	Maximum colonies
colony_lost	integer	Colonies lost
colony_lost_pct	integer	Percent of total colonies lost

We will be using the following columns from the `stressor` dataset.

Variable	Class	Description
year	character	Year
months	character	Month range
state	character	State Name
stressor	character	Stress type
stress_pct	double	Percent of colonies affected by stressors anytime during the quarter, colony can be affected by multiple stressors during same quarter.

Approach:

We will begin with preparing the datasets. The stressor dataset contains 843 missing values for `stress_pct` . To identify the stressors with the highest percentage contribution to colony loss, we require non-null values for `stress_pct` . Therefore, we replace these null values with the mean percentage of proportions, which takes into account the non-null `stress_pct` values of other stressors for each year, month, and state combination. This process results in the creation of the `stressor_cleaned` dataset, which contains 1222 unique records, listing the most concerning stressor for each state during different years and quarters.

Additionally, the `colony` dataset contains 72 and 47 records with missing values for `colony_max` and `colony_lost` , respectively. Since we are interested in the percentage of colony loss, we work with the remaining 1150 records in the `colony_cleaned` dataset as we have no prior information to replace the null values. We also calculate `colony_lost_pct` from `colony_max` and `colony_lost` , wherever possible.

Next, we create a grouped bar plot that shows the most damaging stressors over the years in different quarters, as we want to compare the proportions of states affected. We also plot the distribution of maximum `colony_lost_pct` in different states on a map of the United States for the entire timeframe.

Analysis:

Exploring and cleaning the datasets.

```
# Your R code here
################################# exploring stressor ################################
summary(stressor$stress_pct)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##     0.1     1.7     5.3    11.2    14.2   102.0     843
```

```
################################# exploring colony ##################################
summary(colony$colony_max)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    1700    9000   21000   79113   68750 1710000      72
```

```
summary(colony$colony_lost)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##      20     950    2200   16551    6500  502350      47
```

```
summary(colony$colony_lost_pct)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    1.00    6.00   10.00   11.38   15.00   52.00      54
```

```
################################# cleaning stressor #################################
stressor = stressor %>%
  mutate( # update max stress_pct to 100
```

```
    stress_pct = ifelse(stress_pct>100, 100, stress_pct)
  )

stressor_grp <- stressor %>%
  group_by(year, months, state) %>%
  summarize(
    null_count = sum(is.na(stress_pct)),
    # total percentage assined to different stressors
    assigned_pct = sum(stress_pct, na.rm = TRUE),
    .groups = "drop"
  ) %>%
  mutate(
    # if the total assigned_pct is >100, replace them with 100
    assigned_pct = ifelse(assigned_pct>100, 100, assigned_pct),
    # mean value of unassigned percentage
    mean_remaining_pct = ifelse(null_count>0, (100 - assigned_pct)/null_count, 0)
  )

# combining with stressor_grp to get updated stress_pct
stressor_cleaned <- left_join(stressor, stressor_grp, by = c("year", "months", "st
  mutate( # replace null stress_pct with mean_remaining_pct
    stress_pct = ifelse(is.na(stress_pct), mean_remaining_pct, stress_pct)
  ) %>%
  select(year, months, state, stressor, stress_pct)

# get records with maximum stress_pct value for each year, months, state combinati
stressor_cleaned <- stressor_cleaned %>%
  group_by(year, months, state) %>%
  mutate(rank = order(stress_pct, decreasing=TRUE)) %>%
  filter(rank==1) %>%
  ungroup() %>%
  select(-rank)


################################## cleaning colony ##################################
colony_cleaned <- colony %>%
  select(year, months, state, colony_max, colony_lost, colony_lost_pct) %>%
  filter(!is.na(colony_max) & !is.na(colony_lost)) %>%
  mutate( # replace null colony_lost_pct with mean_remaining_pct
    colony_lost_pct = ifelse(is.na(colony_lost_pct),
                             colony_lost*100/colony_max,
                             colony_lost_pct)
  )

# combining colony and stressor data
colony_cleaned <- colony_cleaned %>%
  left_join(stressor_cleaned, by = c("year", "months", "state"))
```

```
table(colony_cleaned$stressor)
```

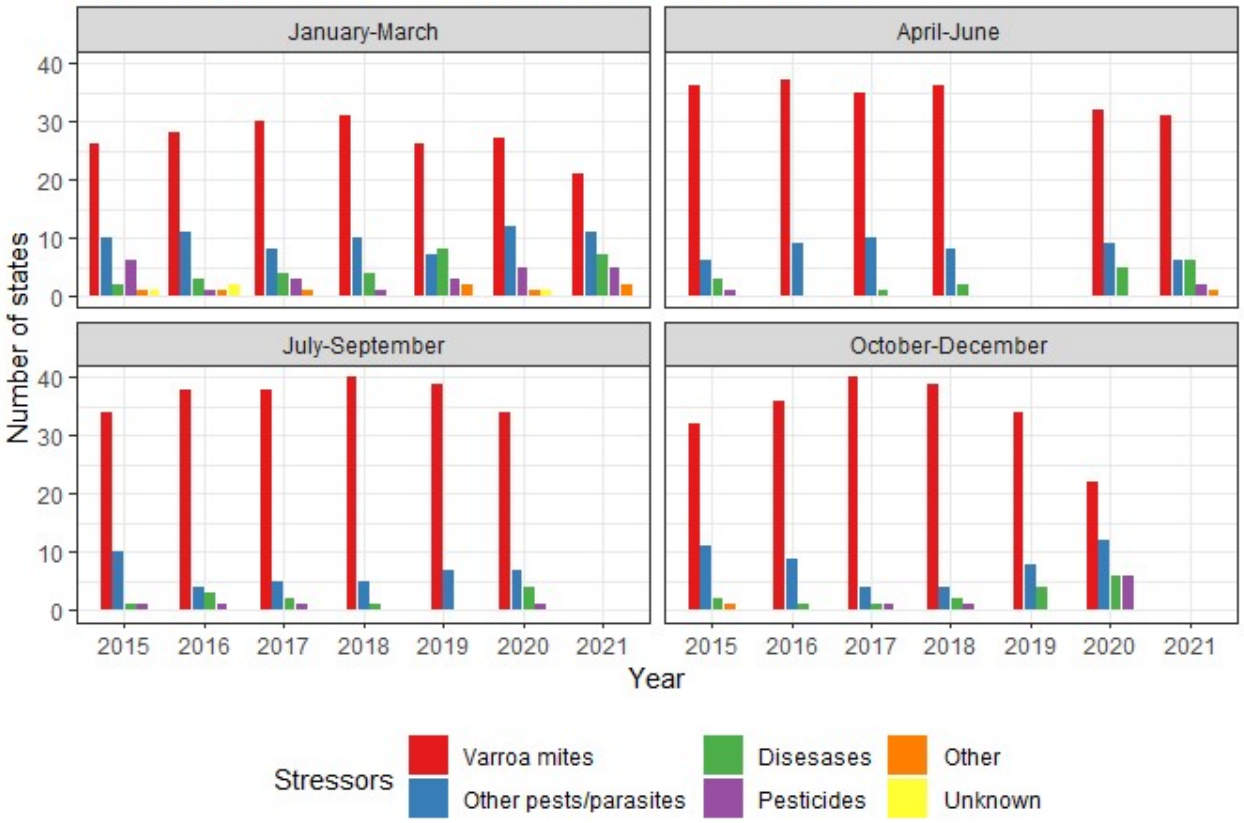
```
##
##             Disesases                 Other Other pests/parasites
##                    72                    10                   203
##            Pesticides               Unknown          Varroa mites
##                    39                     4                   822
```

Now we perform visualization on `colony_cleaned`.

```
# grouped bar plot

# setting the order according to the number of records for each stressor
stressor_order <- c("Varroa mites", "Other pests/parasites", "Disesases", "Pestici
colony_cleaned %>%
  mutate(
    months = factor(months,
                    levels = c('January-March', 'April-June', 'July-September', 'O
    stressor = factor(stressor, levels = stressor_order)
  ) %>%
  ggplot(aes(factor(year), fill = stressor)) +
  geom_bar(position=position_dodge2(preserve = "single")) +
  scale_fill_brewer(palette = "Set1") +
  labs(
    title = "Fig. 1: Major stressors that led to loss of bee colonies",
    x = "Year",
    y = "Number of states",
    fill = "Stressors"
  ) +
  theme_bw() +
  facet_wrap(~months) +
  theme(
    legend.position = "bottom"
  )
```

Fig. 1: Major stressors that led to loss of bee colonies



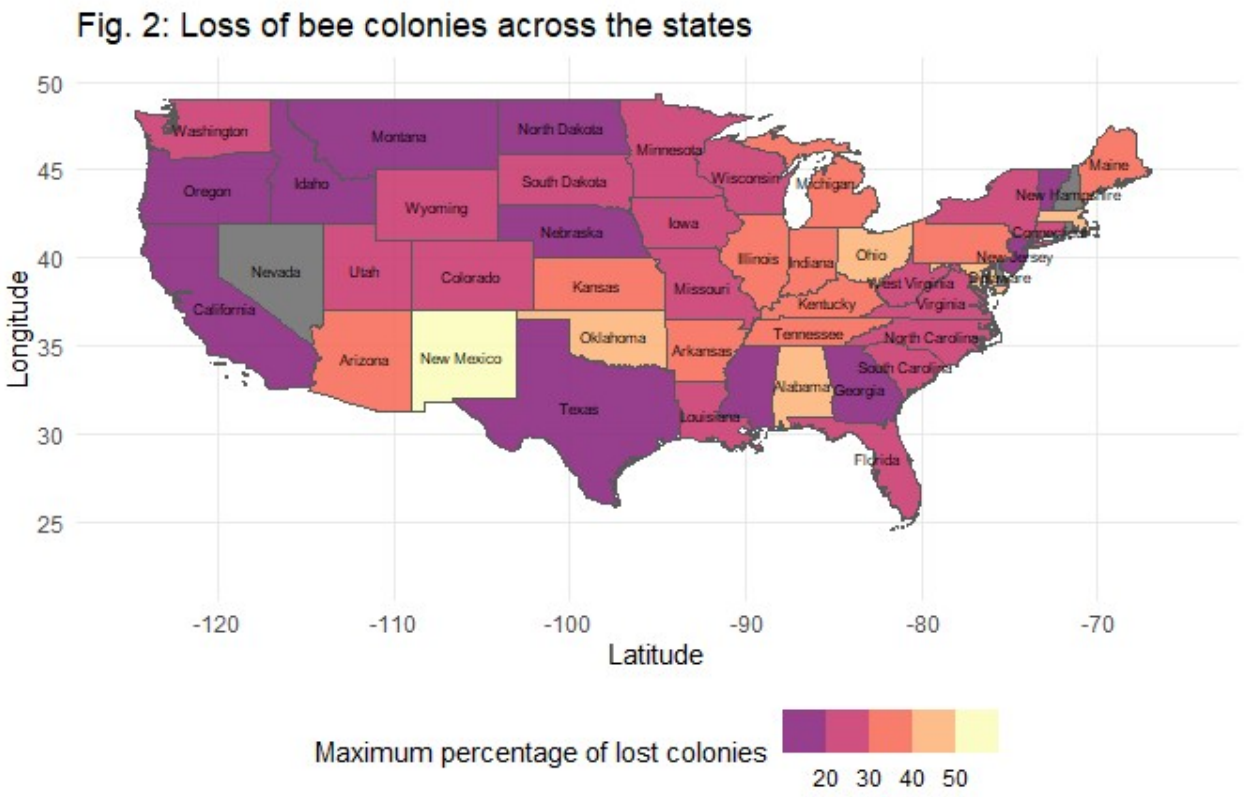

```
# geospacial analysis
# creating a simple features object to get geometry information
sf_us <- ne_states(
  country = "United States of America",
  returnclass='sf'
)

# converting to dataframe for joining with colony_cleaned
sf_us <- as.data.frame(sf_us) %>%
  select(name, geometry)

# right join because there are states with missing data
colony_sf <- colony_cleaned %>%
  right_join(sf_us, by = c("state" = "name" ))

colony_sf %>%
  group_by(state, geometry) %>%
  summarise(
    max_colony_lost = max(colony_lost_pct, rm.na =  TRUE),
    .groups = "drop"
  ) %>%
  ungroup() %>%
  mutate( # deriving latitute and longitude for labelling
```

```r
  lon = map_dbl(geometry, ~st_centroid(.x)[[1]]),
  lat = map_dbl(geometry, ~st_centroid(.x)[[2]])
  ) %>%
ggplot(aes(geometry = geometry, fill = max_colony_lost)) +
geom_sf() +
coord_sf(
  xlim = c(-125, -65),
  ylim = c(22, 50)
) +
scale_fill_viridis_b(
  option = 'A',
  alpha = 0.9,
  begin = 0.4,
  breaks = c(0, 10, 20, 30, 40, 50, 60),
  expand = c(0, 0)) +
theme_minimal() +
theme(
  legend.position = "bottom"
) +
geom_text(aes(label = state, x = lon, y = lat), size = 2, check_overlap = TRUE)
labs(
  title = "Fig. 2: Loss of bee colonies across the states",
  x = "Latitude",
  y = "Longitude",
  fill = "Maximum percentage of lost colonies"
)
```



Discussion:

Based on the information presented in Figure 1, it is evident that Varroa mites are responsible for causing the highest percentage of bee colony losses across all years and quarters. The Varroa mites caused the most damage in approximately 40 states during the last quarter of 2017. While other pests, parasites, and diseases have also contributed to colony loss, they are not nearly as widespread as Varroa mites. It is worth mentioning that the number of states impacted by Varroa mites has not shown a significant decrease over the years.

Figure 2 illustrates that New Mexico experienced the highest bee colony loss among all states, with a loss exceeding 50%. Additionally, Ohio, Oklahoma, Alabama, Massachusetts, and Maryland all lost 40%-50% of their bee colonies, which is also significant. Some states like Nevada and New Hampshire are colored in grey because there is no data available for these states.