Project 5

```
###Load Dataset
colony <- readr::read_csv('https://raw.githubusercontent.com/rfordatascience/tidytuesday/master/data/2022/2022-01-11/colony.csv')
#stressor <- readr::read_csv('https://raw.githubusercontent.com/rfordatascience/tidytuesday/master/data/2022/2022-01-11/stressor.csv')
```

- Dataset: **Bee Colonies**
  - Source: Tidy Tuesday - https://github.com/rfordatascience/tidytuesday/blob/master/data/2022/2022-01-11/readme.md#bee-colonies (https://github.com/rfordatascience/tidytuesday/blob/master/data/2022/2022-01-11/readme.md#bee-colonies)
  - Published on: January 11, 2022

Question: *Does the impact of bee colony loss vary across different seasons and geographical areas within the United States? If so, how can these regions be grouped based on their characteristic bee colony loss patterns? Emphasize the findings by employing effective visualization methods.*

Introduction:

Honey bees, the indispensable pollinators of our ecosystems, are facing a crisis of declining colony numbers. Their critical role in pollinating about one-third of the world's food supply makes their decline a grave threat to agriculture and the environment.

This project will analyze a dataset named 'colony', which contains statistical information about honey bee colonies in various states of the United States and parts of Canada between 2015 and 2021. The dataset provides details on honey bee colonies, including colony numbers, maximum colony counts, colony losses, percentage losses, colony additions, renovations, percentage renovations, and colony losses for both large (over five colonies) and small (less than five colonies) operations. Data collection occurred quarterly for operations with five or more colonies and annually for operations with less than five colonies.

Utilizing this dataset, we will conduct various analyses to address the questions posed. For ease of analysis, the questions can be categorized into three parts:Assessing Seasonal Variations in Bee Colony Loss,Geographical Clustering Based on Bee Colony Loss Patterns and Quantifying Bee Colony Loss across US States

To address the questions posted, Only the data from 2015-20 will be included for the states within United states. Following key fields from the provided dataset will be used for analysis-

- *Year*: The year corresponding to the bee colony data.
- *Months*: The three-month periods within a year for which the bee colony data was collected.
- *State*: The name of the state in which the bee colony data was collected.
- *Colony Number*: The total number of bee colonies present in the state.
- *Colony Loss*: The number of bee colonies lost in the state.
- *Colony Loss Percentage*: The proportion of bee colonies lost in the state, expressed as a percentage.

By analyzing these key fields, we can gain valuable insights into the seasonal and geographical patterns of bee colony loss in the United States.

Approach:

Prior to delving into our analysis, we must meticulously prepare the dataset to ensure its integrity and reliability. This involves selecting the relevant fields from the original dataset and eliminating any records containing missing values or “n/a” entries. To enhance the analysis, we will introduce a new column that identifies the corresponding quarter of the year based on the provided months. Additionally, we will remove data from the year 2021 and any data that does not pertain to the United States. The resulting refined dataset will be aptly named “colony_filtered” and will serve as the foundation for our subsequent analysis.To comprehensively address the questions at hand, we will adopt a multifaceted approach, employing a variety of techniques and data visualizations to thoroughly examine each aspect of the questions as below.

- Part 1: Investigating Seasonal Variations in Bee Colony Loss

  This section aims to determine whether the severity of bee colony loss varies across different seasons. To achieve this, we will create a subset named ‘colony_subset_1’ that consolidates the average colony losses for each quarter and each state. This subset will be visualized as a line graph to effectively identify any trends or patterns in colony losses over time.Line graphs are particularly useful for illustrating data variables and trends clearly.

- Part 2: Unveiling Regional Patterns in Bee Colony Loss

  This section will delve into the realm of regional patterns in bee colony loss, exploring whether there exist distinct geographical groups of states that exhibit similar loss patterns. To achieve this, we will construct a new subset named ‘colony_subset_2’ comprising data from the top two quarters (identified in Part 1) during which each state experienced the highest losses.

  To effectively identify groups of states with similar bee colony loss patterns, we will employ the powerful technique of hierarchical clustering. This method meticulously analyzes the data and constructs a hierarchical tree-like structure, known as a dendrogram, that visually represents the relationships and groupings among the states based on their colony loss patterns.While the creation of a dendrogram involves intricate steps such as calculating distance matrices and assigning clusters, the resulting visualization provides invaluable insights into the hierarchical relationships and complex structures that may remain elusive when examining raw data alone. A single glance at a well-crafted dendrogram can reveal groupings and patterns among the states, guiding our understanding of regional variations in bee colony loss.

- Part 3: Magnitude of Bee Colony Loss Across US States

  This section will explore the magnitude of variations in bee colony loss across different US states. We will quantify and compare the average bee colony loss rates among different states to identify areas with particularly high or low loss rates.To achieve this, we will utilize the subset ‘colony_subset_2’ and visualize the data on a geographical map of the United States using the powerful ‘rnaturalearth’ and ‘geom_sf’ packages in R. This approach will allow us to create a choropleth map of the United States and effectively detect any relationships between the data and geographical location.Choropleth maps are particularly useful for comparing regions with one another, making them ideal for addressing our objective of understanding the magnitude of bee colony losses across the United States.

Analysis:

Firstly,we must extract the required fields from original data set and clean the same.

```
#Clean up the data set

colony_filtered <- colony %>%
  select(year,months,state,colony_n,colony_lost,colony_lost_pct) %>%
  na.omit(colony) %>%                          #remove rows with N/A
  filter(year !=2021) %>%                      #Excluding 2021 year stats
  filter(state != "Other States") %>%          #Excluding Non US data
  mutate(Quarter_of_year = case_when(          #Adding new column to track Quarter of the year
    months == "January-March"    ~ "Q1",
    months == "April-June"       ~ "Q2",
    months == "July-September"   ~ "Q3",
    months == "October-December" ~ "Q4"))
```

To delve into a comprehensive examination of the research questions, we will systematically address each question in a multifaceted manner, as outlined in the approach section.

Analysis of PART-1 begins here-

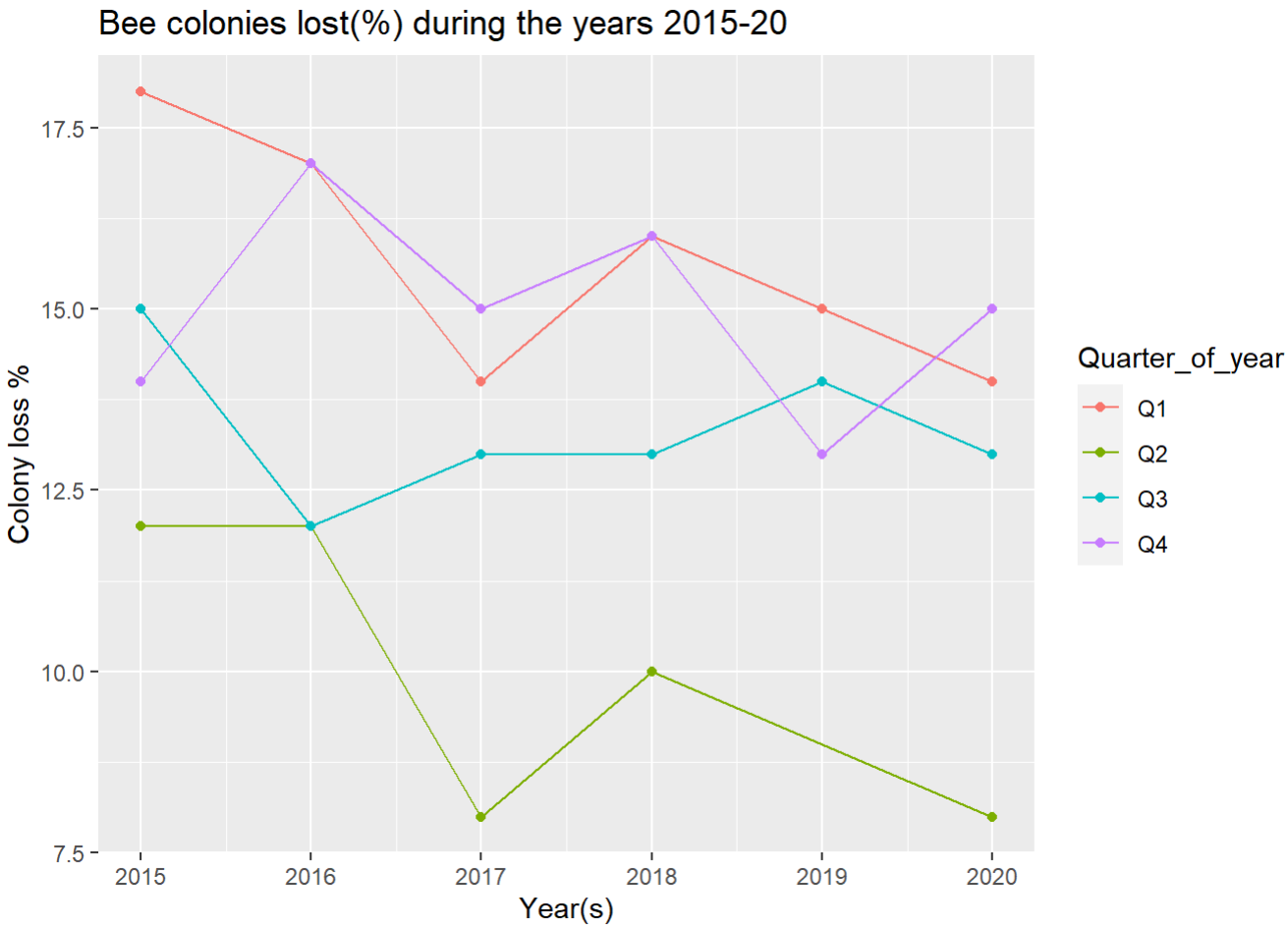
To address part 1 of the question,create a subset of the data set that focuses on the seasonal variations in bee colony loss.

```
#Create a subset-1

#Subset-1 :It has consolidated data of all the US states for each year
colony_subset_1 <-colony_filtered %>%  #Subset#1
  filter(state == "United States")

head(colony_subset_1)
```

```
## # A tibble: 6 × 7
##    year months        state colony_n colony_lost colony_lost_pct Quarter_of_year
##   <dbl> <chr>         <chr>    <dbl>       <dbl>           <dbl> <chr>
## 1  2015 January-March Unit…  2824610      500020              18 Q1
## 2  2015 April-June    Unit…  2849500      352860              12 Q2
## 3  2015 July-Septemb… Unit…  3132880      457100              15 Q3
## 4  2015 October-Dece… Unit…  2874760      412380              14 Q4
## 5  2016 January-March Unit…  2594590      428800              17 Q1
## 6  2016 April-June    Unit…  2801470      329820              12 Q2
```

Now,Plot the line graph based on above subset .This will be supportive to tackle the first part of research question.

```
#Plot using a subset
ggplot(colony_subset_1,aes(year,colony_lost_pct))+
  geom_line(aes(color=Quarter_of_year))+
  geom_point(aes(color=Quarter_of_year))+
  scale_x_continuous(name= "Year(s)",breaks=seq(2015, 2020, 1))+
  scale_y_continuous(name="Colony loss %")+
  ggtitle("Bee colonies lost(%) during the years 2015-20")
```



The seasonal pattern of bee colony loss is evident from the graph.

Bee colonies suffered higher average losses during the first (Q1) and fourth (Q4) quarters throughout the analyzed years. Interestingly, the second quarter (Q2) appears to be the safest season for bee colonies. This suggests that seasonal factors play a significant role in bee colony health and survival.

Now, let's shift our focus to parts 2 and 3 of the question, which delve into regional patterns in bee colony loss. To accomplish this, we'll require another subset of the filtered dataset, specifically one that includes bee colony losses from Q1 and Q4, periods characterized by higher average colony loss rates as indicated by the line graph.

Analysis of PART-2 begins here-

```
###Create a subset-2

#Subset-2 : It has the average colony loss % data of all the US states in Q1 & Q4 of each year i
n the period 2015-20

colony_subset_2 <- colony_filtered %>%
  filter(state != "United States") %>%
  filter(Quarter_of_year=="Q1" | Quarter_of_year=="Q4")

colony_subset_2 <- colony_subset_2  %>%
  group_by(state) %>%
  mutate(avg_colony_loss_pct = mean(colony_lost_pct)) %>%
  filter(!duplicated(state)) %>%
  select(state, avg_colony_loss_pct)

head(colony_subset_2)
```

```
## # A tibble: 6 × 2
## # Groups:   state [6]
##   state       avg_colony_loss_pct
##   <chr>                     <dbl>
## 1 Alabama                    13.3
## 2 Arizona                    19.1
## 3 Arkansas                   16.3
## 4 California                 13.3
## 5 Colorado                   14.2
## 6 Connecticut                 9.5
```

With the new dataset prepared, we're now ready to embark on the task of hierarchical clustering. This process involves assigning states to distinct clusters, a crucial step in constructing an informative dendrogram.

```
### Hierarchical clustering plot in 3 steps #####

#Step 1: Calculate the distance matrix
dist_out <- colony_subset_2 %>%
  column_to_rownames(var = "state") %>%
  scale() %>%
  dist(method = "manhattan")

sample(dist_out,100,replace=TRUE)  #Sample of Distance matrix based on method = "manhattan"
```

```
##   [1] 0.29583573 0.94297638 1.07240451 1.01693531 0.49922279 0.01848973
##   [7] 0.75807905 0.46224332 2.79194966 0.65823449 0.96146611 1.79350409
##  [13] 0.05546920 0.83203798 1.90444249 0.24036653 0.14791786 0.61016118
##  [19] 2.16329875 0.09244866 2.66252153 1.75652462 1.66407596 2.40366527
##  [25] 1.66407596 0.59167145 2.86590860 1.01693531 1.05391477 1.05391477
##  [31] 1.46068890 1.75652462 1.59011703 0.83203798 2.86590860 1.36824023
##  [37] 2.07085008 1.71954516 1.95991168 0.46224332 1.42370943 1.47917863
##  [43] 2.38517554 1.50506426 0.16640760 1.62709649 1.83048355 1.51615810
##  [49] 0.27734599 1.47917863 0.70260985 2.53309340 2.49611394 1.49766836
##  [55] 1.16485317 3.82737471 1.86746302 1.05391477 0.24036653 2.05975624
##  [61] 0.01848973 2.36668581 1.47917863 1.59011703 0.27734599 1.88595275
##  [67] 0.55469199 2.21876795 1.03542504 1.07240451 0.72109958 1.07240451
##  [73] 0.66563038 0.12942813 1.44219916 3.34664165 0.83203798 1.70105542
##  [79] 1.62709649 0.97995584 1.71954516 0.82464209 0.88750718 0.81354825
##  [85] 2.42215501 1.55313756 0.73958932 0.12203224 1.60860676 2.31121661
##  [91] 0.72109958 0.53620225 0.14791786 0.09244866 0.59167145 1.09089424
##  [97] 0.00000000 0.18489733 0.85052771 0.20338706
```

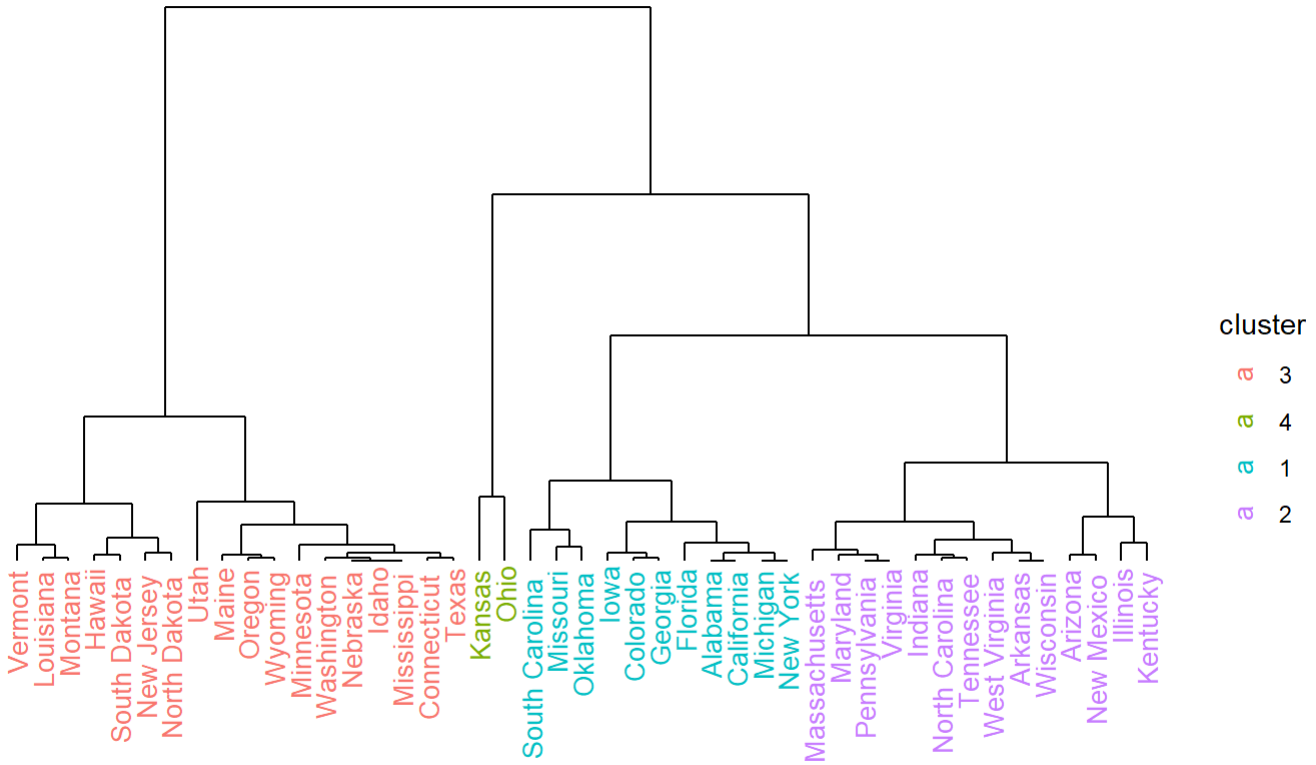
```
library(ggdendro)
```

```
## Warning: package 'ggdendro' was built under R version 4.3.2
```

```
#Step 2: Cluster
hc_out <- hclust(
  dist_out, method = "complete")
# Step 3a:prep for plotting
ddata <- dendro_data(
  hc_out,
  type = "rectangle"
)
segments <- segment(ddata)
labels <- label(ddata)
clust    <- cutree(hc_out,k=4)                    # find 4 clusters
clust.df <- data.frame(label=names(clust), cluster=factor(clust,levels = c(3,4,1,2)))

ddata[["labels"]] <- merge(ddata[["labels"]],clust.df, by="label")

clust      #View states with Cluster designations (1 - 4)
```

```
##        Alabama        Arizona       Arkansas     California       Colorado
##              1              2              2              1              1
##    Connecticut        Florida        Georgia         Hawaii          Idaho
##              3              1              1              3              3
##       Illinois        Indiana           Iowa         Kansas       Kentucky
##              2              2              1              4              2
##      Louisiana          Maine       Maryland  Massachusetts       Michigan
##              3              3              2              2              1
##      Minnesota    Mississippi       Missouri        Montana       Nebraska
##              3              3              1              3              3
##     New Jersey     New Mexico       New York North Carolina   North Dakota
##              3              2              1              2              3
##           Ohio       Oklahoma         Oregon   Pennsylvania South Carolina
##              4              1              3              2              1
##   South Dakota      Tennessee          Texas           Utah        Vermont
##              3              2              3              3              3
##       Virginia     Washington  West Virginia      Wisconsin        Wyoming
##              2              3              2              2              3
```

```
# Step 3b:All set!..Now Plot
ggplot() +
  geom_segment(data=segment(ddata), aes(x=y, y=x, xend=yend, yend=xend)) +
  geom_text(
    data=label(ddata), aes(y, x, label=label, hjust=0, color=cluster),
    hjust = 1.1,
    size = 10/.pt,
    angle = 90
       ) +
  coord_flip(ylim = c(1, 45),xlim = c(-2,4)) +
  theme(axis.line.y=element_blank(),
    axis.ticks.y=element_blank(),
    axis.text.y=element_blank(),
    axis.title.y=element_blank(),
    axis.text.x=element_blank(),
    axis.ticks.x=element_blank(),
    axis.title.x=element_blank()
  )+
  ggtitle("Hierarchical clustering of US states based on Avg loss of Bee colonies")+
  theme_void()
```



The distinct clustering patterns observed in the dendrogram clearly indicate that states can be categorized into groupings based on their bee colony loss patterns.For example ,High loss states such as Kansas and Ohio are tagged to cluster 4 This variation in bee colony loss across geographical regions provides a compelling answer to part 2 of our research question.

Analysis of PART-3 begins here-

To address the final part of our research question, we will perform geospatial analysis on the subset-2 using a choropleth map of US states as below.

```
### Prepare datasets to plot US States
library(rnaturalearth)
```

```
## Warning: package 'rnaturalearth' was built under R version 4.3.2
```

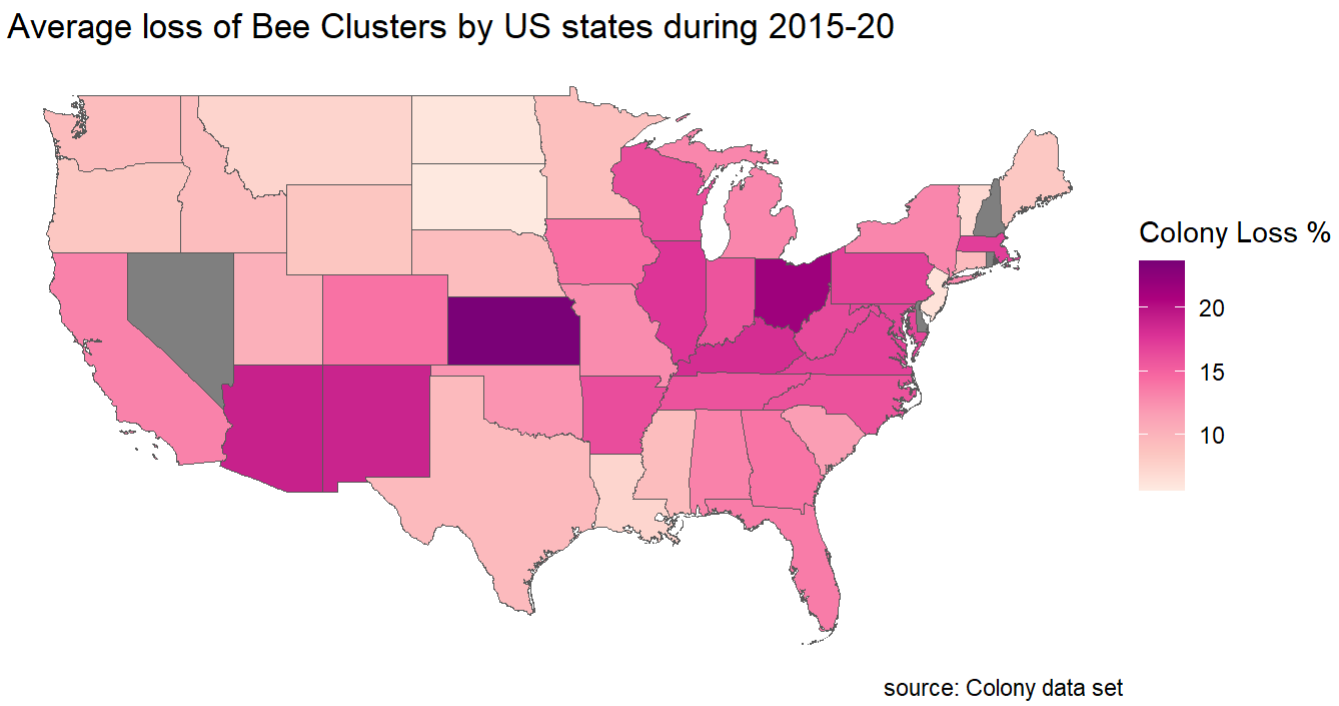
```
## The legacy packages maptools, rgdal, and rgeos, underpinning the sp package,
## which was just loaded, will retire in October 2023.
## Please refer to R-spatial evolution reports for details, especially
## https://r-spatial.org/r/2023/05/15/evolution4.html.
## It may be desirable to make the sf package available;
## package maintainers should consider adding sf to Suggests:.
## The sp package is now running under evolution status 2
##      (status 2 uses the sf package in place of rgdal)
```

```
## Support for Spatial objects (`sp`) will be deprecated in {rnaturalearth} and will be removed
in a future release of the package. Please use `sf` objects with {rnaturalearth}. For example: `
ne_download(returnclass = 'sf')`
```

```
options(max.print = .Machine$integer.max)
sf_us <- ne_states(                         #geo details of US states
  country = "United States of America",
  returnclass='sf'
)
sf_us <- sf_us %>% select(name)   %>%       # Use the dplyr package to merge the sf_us and colon
y_subset_2
  left_join(.,colony_subset_2, by = c("name"="state"))
sf_us %>%
  #filter(type == "state")%>%
  drop_na(avg_colony_loss_pct)
```

```
## Simple feature collection with 45 features and 2 fields
## Geometry type: MULTIPOLYGON
## Dimension:     XY
## Bounding box:  xmin: -178.3044 ymin: 18.90612 xmax: -66.97732 ymax: 49.36949
## Geodetic CRS:  WGS 84
## First 10 features:
##            name avg_colony_loss_pct                       geometry
## 1    Washington            9.416667 MULTIPOLYGON (((-122.753 48...
## 2         Idaho            9.333333 MULTIPOLYGON (((-117.0382 4...
## 3       Montana            7.500000 MULTIPOLYGON (((-116.0482 4...
## 4  North Dakota            6.166667 MULTIPOLYGON (((-104.0476 4...
## 5     Minnesota            9.083333 MULTIPOLYGON (((-97.22609 4...
## 6      Michigan           13.083333 MULTIPOLYGON (((-84.4913 46...
## 7          Ohio           21.583333 MULTIPOLYGON (((-80.52023 4...
## 8  Pennsylvania           17.000000 MULTIPOLYGON (((-79.76301 4...
## 9      New York           13.083333 MULTIPOLYGON (((-79.06523 4...
## 10      Vermont            7.000000 MULTIPOLYGON (((-73.35134 4...
```

```
###Plot choropleth of US states

ggplot(data=sf_us,aes(fill = avg_colony_loss_pct))+
  geom_sf()+
  scale_x_continuous(limits = c(-125, -67))+
  scale_y_continuous(limits = c(25, 50))+
  scale_fill_distiller(palette = "RdPu",
                       direction = 1) +
  labs(title = "  Average loss of Bee Clusters by US states during 2015-20 ",
       caption = "source: Colony data set",
       fill = "Colony Loss %") +
  theme_void()
```



The choropleth map vividly illustrates the substantial variations in bee colony losses across the United States. States depicted in darker shades of pink, such as those in the Northeast, experienced higher bee colony losses compared to states with lighter shades, primarily located in the West.

Having gathered substantial insights through our comprehensive analysis, we are now well-positioned to present our findings and conclusions in the discussion forum.

Discussion: *Your discussion of results here.*

This comprehensive study delves into the intricate patterns of bee colony loss across various seasons and geographical regions within the United States. Drawing upon a comprehensive dataset spanning multiple years, we have uncovered significant variations in colony loss rates, highlighting the critical role of both temporal and spatial considerations in crafting effective bee colony conservation strategies.

The line graph clearly illustrates that the impact of bee colony loss is not consistent throughout the year. Bee colonies experienced notably higher average losses during the first (Q1) and fourth (Q4) quarters, particularly during the winter and spring seasons. This seasonal pattern suggests that seasonal factors, such as weather conditions , play a crucial role in influencing bee colony health and survival. The harsh winter and early spring months pose a significant threat to bee colonies, as they are exposed to numerous unfavorable factors that elevate colony losses.

The geographical distribution of bee colony loss reveals a striking pattern, with varying rates observed across different regions of the United States. The dendrogram analysis effectively categorizes states into four distinct clusters based on their bee colony loss patterns. This clustering underscores the importance of considering regional factors when developing targeted conservation strategies. The choropleth map vividly illustrates the substantial variations in bee colony losses across the country. Notably, states in the Northeast region generally experienced higher bee colony losses compared to those in the West, highlighting the need for geographically

specific conservation approaches.The findings of this study underscore the imperative for a multifaceted approach to bee colony conservation. Targeted conservation efforts must encompass both seasonal and geographical variations in bee colony loss rates to effectively protect these essential pollinators.

By understanding the complex interplay of factors influencing bee colony health and survival, we can develop comprehensive strategies to mitigate the impact of bee colony loss and safeguard the continued pollination of our crops and ecosystems.