Project 3

Enter your name and EID here

Connor Hanna cdh3663

This is the dataset used in this project:

```
#news_orgs from Project Oasis via Data is Plural

tuesdata <- tidytuesdayR::tt_load('2022-01-11')
```

```
##
##  Downloading file 1 of 2: `colony.csv`
##  Downloading file 2 of 2: `stressor.csv`
```

```
tuesdata <- tidytuesdayR::tt_load(2022, week = 2)
```

```
##
##  Downloading file 1 of 2: `colony.csv`
##  Downloading file 2 of 2: `stressor.csv`
```

```
colony <- tuesdata$colony
stressor <- tuesdata$stressor
```

Link to the dataset: https://github.com/rfordatascience/tidytuesday/blob/master/data/2022/2022-01-11/readme.md

Part 1

Question:

What stressors are most likely to result in colony loss? Are some stressors more likely to occur together than others?

Introduction:

Bees are extremely important to the health of American agriculture and ecosystems. Farmers cultivating crops ranging from almonds to oranges frequently rent beehives to pollinate their plants before harvest. But in recent years, climate change and a wide range of other stressors have begun to have a deleterious impact on bee populations overall.

The `colony` and `stressor` data were both collected by the United States Department of Agriculture in an effort to answer crucial questions concerning which stressors are most prevalent and most destructive to beekeeping and related agricultural practices in the United States. Both datasets are formatted long, with `colony` containing variables on overall hive numbers and losses organized by state, season, and year. The smaller dataset, `stressor`, contains the same identification variables and similarly organized observations. The two unique variables, `stressor` and `stress_pct`, refer to a given pest afflicting a given percentage of

beehives in that particular state and season. The identification variables for both datasets are identical, facilitating easy merging for further analysis.

To answer the question, I will need information from the variables `year`, `months`, `state`, `stressor`, `stress_pct`, and `colony_lost_pct` variables. Only `year`, `stress_pct` and `colony_lost_pct` are numeric. The other variables `months`, `state`, and `stressor` are character variables. `year`, `months` and `state` are the observation ID variables, containing information on the year, quarter, and state from which the information was gathered. `stressor` is categorical, containing information on the pest type for the observation. `stress_pct` is a numerical variable associated with `stressor`, indicating the percentage of affected hives in the given `month`, `year`, and `state` of the observation. `colony_lost_pct` is numerical and reflects the percentage of lost colonies in the `month`, `year`, and `state` of the observation.

#revisit me

Approach:

To clean the data, I'll begin by using `spread()` and `left_join()` to merge the stressor data with the other colony information. This should result in a dataframe containing one observation for each state-year-season combination, with each observation containing information on the prevalence of all six colony collapse disorder contributors. I'll also use the `fill` option in `spread()` to convert the NA values to zeroes.

Once the data is clean and in wide format, I'll run a linear regression on the stressors and `colony_lost_pct`. Since both the stressors and `colony_lost_pct` are in percentages, this will help me to minimize any accidental influence of state-level variation in bee population. I'll then use `map()` to run the same regression across all data by state. This will help me to determine the first part of my question: Are some stressors more destructive than others?

Then I will use `prcomp()`, `geom_segment()`, `geom_text_repel()`, and `geom_col()` to generate principle components, conduct a principle component analysis on the stressor variables and `colony_lost_pct`, generate a rotation matrix, and plot the resulting eigenvalues. This will allow me to examine the stressor variables for collinearity, determining if stressors are co-occurring in separate combinations or if the colony losses are being caused by all of the stressors together. Color was added to the eigenvalue bar graph to aid in readability.

Analysis:

```
#spreading stressor
stresses <-
  stressor |>
    spread(stressor, stress_pct, fill = 0)

#merging the stresses data with the colony data
#removing country-level observations
bees <-
  left_join(colony, stresses) |>
  filter(state != "United States")
```

```
## Joining, by = c("year", "months", "state")
```

```
#looks like there's a typo in the column name for "Diseases"
bees <-
  bees |>
  rename(Diseases = Disesases)
```

```
#running a regression on colony losses
lm_ccd <-
  lm(colony_lost_pct ~ Diseases + Other + `Other pests/parasites` + Pesticides + Unknown + `Varroa mites
summary(lm_ccd)
```

```
##
## Call:
## lm(formula = colony_lost_pct ~ Diseases + Other + `Other pests/parasites` +
##     Pesticides + Unknown + `Varroa mites`, data = bees)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -21.267  -4.157  -1.061   2.713  39.538
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)               6.93295    0.38641  17.942  < 2e-16 ***
## Diseases                  0.02641    0.03139   0.841 0.400321
## Other                     0.25765    0.03225   7.989 3.31e-15 ***
## `Other pests/parasites` -0.06389    0.01744  -3.663 0.000261 ***
## Pesticides               -0.01869    0.02428  -0.770 0.441713
## Unknown                   0.34553    0.03776   9.151  < 2e-16 ***
## `Varroa mites`            0.06989    0.01293   5.406 7.85e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.586 on 1136 degrees of freedom
##   (53 observations deleted due to missingness)
## Multiple R-squared:  0.1876, Adjusted R-squared:  0.1834
## F-statistic: 43.73 on 6 and 1136 DF,  p-value: < 2.2e-16
```

```
#using map() to fit lm by state and season
#this gives us year-over-year results, controlling for seasonality and geography
models <-
  bees |>
    nest(data = -c(state)) |>
    mutate(
      fit = map(data, ~lm(colony_lost_pct ~ Diseases + Other + `Other pests/parasites` + Pesticides + U
                          `Varroa mites`, data = .x))
    )

#now we have 48 models
#fetching the model for California during January-March
summary(models$fit[[4]])
```

```
##
## Call:
## lm(formula = colony_lost_pct ~ Diseases + Other + `Other pests/parasites` +
##     Pesticides + Unknown + `Varroa mites`, data = .x)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.9241 -1.7860 -0.6182  1.9525  4.3284
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)               9.28662    3.54622   2.619   0.0174 *
## Diseases                 -0.43828    0.32044  -1.368   0.1882
## Other                     0.14368    0.30251   0.475   0.6405
```

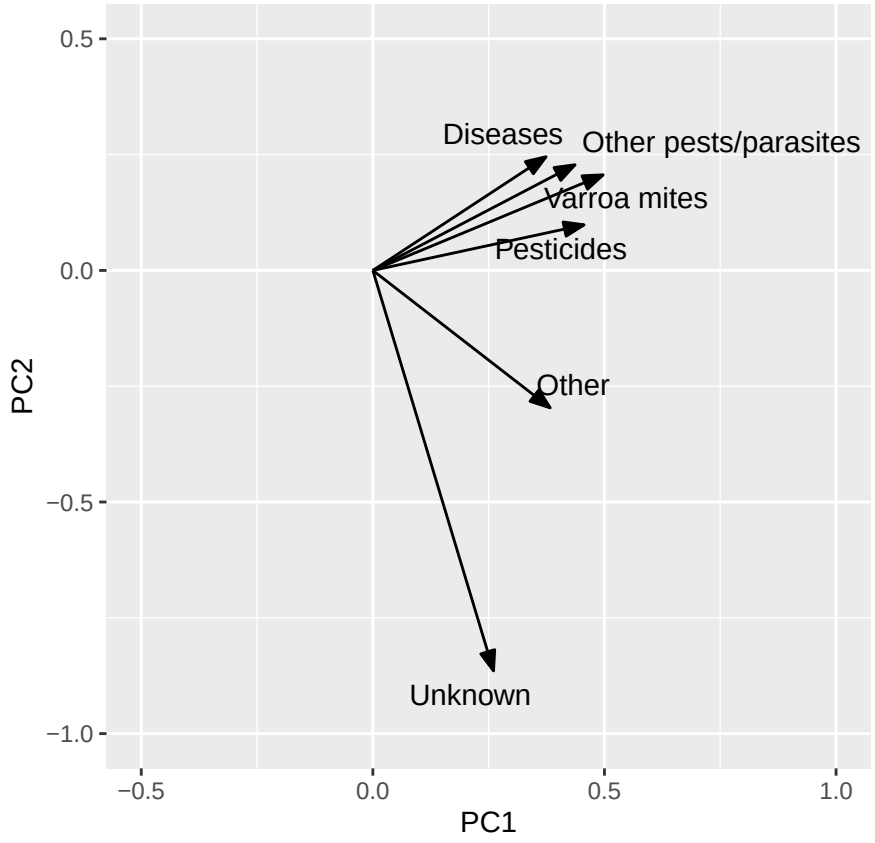
```
## `Other pests/parasites`  0.43223    0.22460   1.924   0.0703 .
## Pesticides              -0.40143    0.22960  -1.748   0.0974 .
## Unknown                  0.39425    0.43762   0.901   0.3795
## `Varroa mites`           0.03661    0.10286   0.356   0.7261
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.823 on 18 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.2802, Adjusted R-squared:  0.0403
## F-statistic: 1.168 on 6 and 18 DF,  p-value: 0.3659
```

```
#storing PCA of the bee plagues
pca_fit <-
  bees |>
  select(Diseases, Other, `Other pests/parasites`,
           Pesticides, Unknown, `Varroa mites`) |>
  na.omit() |>
  scale() |>
  prcomp()
pca_fit
```

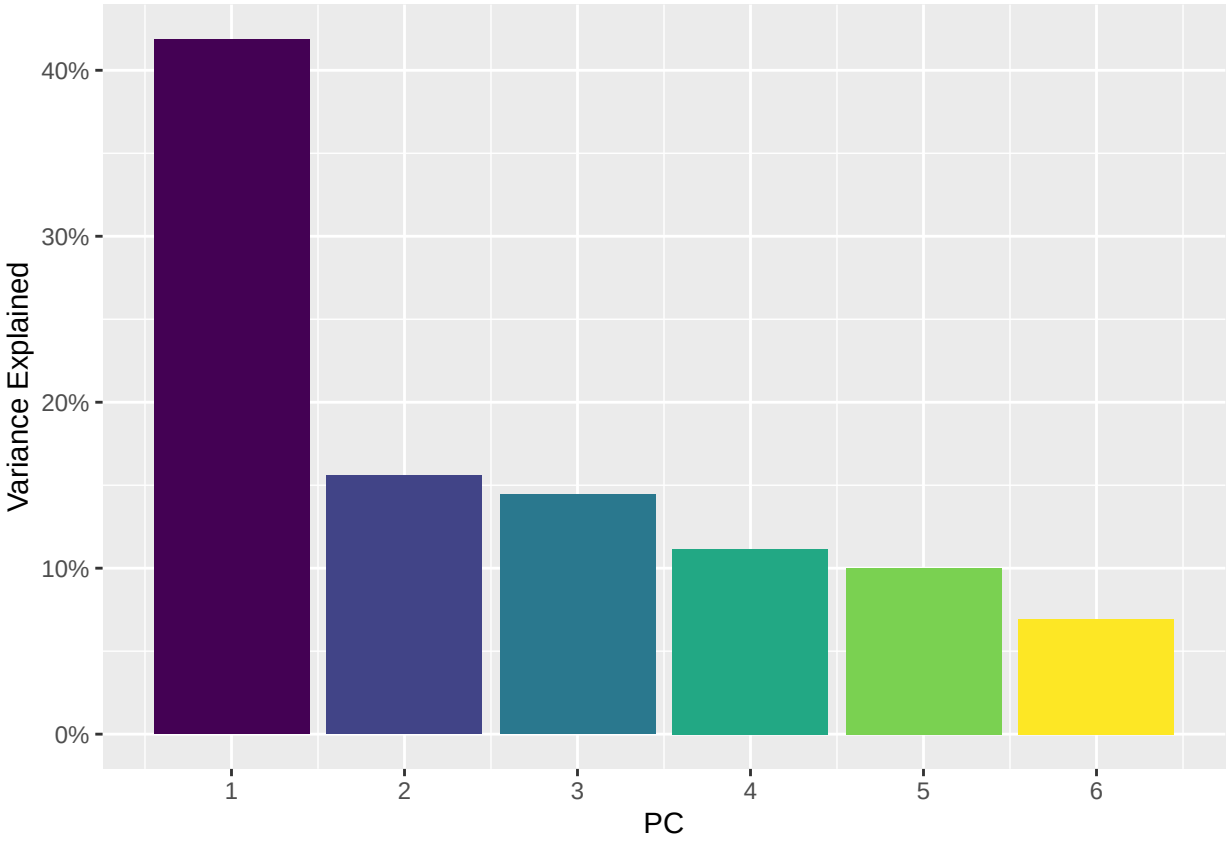
```
## Standard deviations (1, .., p=6):
## [1] 1.5848913 0.9670070 0.9308619 0.8182012 0.7750791 0.6452226
##
## Rotation (n x k) = (6 x 6):
##                              PC1         PC2         PC3         PC4         PC5
## Diseases               0.3738881  0.24538692  0.58513340 -0.55580217  0.38380311
## Other                  0.3828800 -0.29664685  0.46525803  0.71696388  0.13286173
## Other pests/parasites 0.4365653  0.22810439 -0.55666507  0.05756449  0.30323428
## Pesticides             0.4560130  0.09837163  0.09105773 -0.13317707 -0.85452151
## Unknown                0.2604559 -0.86459046 -0.19897311 -0.37212255  0.06783723
## Varroa mites           0.4972285  0.20630231 -0.28878376  0.13236975  0.09101099
##                               PC6
## Diseases               -0.03729614
## Other                  -0.13133574
## Other pests/parasites -0.59349677
## Pesticides             -0.16168020
## Unknown                 0.04448221
## Varroa mites            0.77524341
```

```
#plotting a rotation matrix
arrow_style <- arrow(
  angle = 20, length = grid::unit(8, "pt"),
  ends = "first", type = "closed"
)

pca_fit |>
    tidy(matrix = "rotation") |>
  pivot_wider(
    names_from = "PC", values_from = "value",
    names_prefix = "PC"
  ) |>
```

```
  ggplot(aes(PC1, PC2)) +
  geom_segment(
    xend = 0, yend = 0,
    arrow = arrow_style
  ) +
  geom_text_repel(aes(label = column)) +
  xlim(-.5, 1) + ylim(-1, .5) +
  coord_fixed()
```




```
#fetching the r-squared values for the principle components via eigenvalue plot
pca_fit |>
  tidy(matrix = "eigenvalues") |>
  ggplot(aes(PC, percent, fill = PC)) +
  geom_col() +
  scale_x_continuous(
    breaks = 1:6
  ) +
  scale_y_continuous(
  name = "Variance Explained",
  label = scales::label_percent(accuracy = 1)
  ) +
  scale_fill_viridis_c() +
  theme(legend.position = "none")
```



Discussion:

The results from the regression were inconclusive. Several of the coefficients were statistical zeroes - of those that weren't, the effect size was fairly modest. This suggested that the stressors contributed to colony collapse only when present simultaneously and/or in large quantities. When the same model was applied only to observations in California the effect sizes all diminished to statistical zero. With the inconsistency of effect sizes between samples and models, omitted variable bias seems the results. Furthermore, the inconsistency also suggests that there may be a high degree of collinearity between diseases. This matches the theoretical understanding of CCD, where multiple stressors combine with a change in climate to push colonies to collapse.

The results from the PCA suggested positively what the regression suggested negatively. In the rotation matrix, four of the six vectors for stressors exhibited high degrees of collinearity. This result further supports the theory that CCD is caused by co-occurrence of multiple stressors, and provided scant evidence that the stressors occurred together in distinct clusters.

Part 2

Question:

How do colony losses and pest proportions vary by season? Is seasonal variation in stressor prevalence associated with colony losses to colony collapse disorder?

Introduction:

Since the stressors exhibit a high degree of collinearity, the next logical question is *when* the stressors occur and why. In particular, seasonal variation was excluded from both the regression and the PCA. Thus, the

next target of my analysis will be examining the seasonal variation in both stressor prevalence and colony losses.

To answer the question, I will need information from the variables `year`, `months`, `state`, `stressor`, `stress_pct`, `colony_n`, and `colony_lost_pct` variables. Only `year`, `stress_pct`, `colony_n`, and `colony_lost_pct` are numeric. The other variables `months`, `state`, and `stressor` are character variables. `colony_n` represents the total number of surveyed colonies in the sample.

Approach:

To accomplish this, I start by cleaning the separate data for each figure. To construct a bar graph I join the longitudinal versions of `colony` and `stressor`, then `filter()` to remove the missing observations from 2019 and the national observations. I then `mutate()` to compute figures for the total number of bees affected by each stressor in each observation, and use `group_by()` and `summarize()` to condense that information into a table with readable assignments for the `ggplot()` aesthetics. I also use `factor()` to manually reorder columns in advance of the `ggplot()` call.

For the box plot, I use `filter()` to remove the national-level observations so they don't appear as outliers. I also use `mutate()` to convert the percentages to decimals for the `label` option in `scale_x_continuous()`.

The stacked bar graph generated by `geom_col()` was chosen to capture the two dimensions important to the first part of my question - the number of colonies subject to stress, and the types of stressors those colonies experienced. This will allow me to observe simultaneously changes in the stressor types and the quantities of affected hives, to be compared with the results in the boxplot to observe potential correlation between stressors and losses.

`geom_boxplot()` was chosen to represent variation both within and between seasons in numbers of colony that collapsed after featuring CCD symptoms. Density plots proved inadequate due to the relatively large number of outliers and the high degree of skewness in the distribution. Most seasons, relatively low percentages of colonies collapsed due to CCD. Some states and seasons, however, experienced severe losses. Boxplots were better able to visualize those outliers. I chose the `colony_lost_pct` variable again to minimize interference from variation in state size, and for easier comparison to the `stress_pct` variable used in the bar graph and regression models.

Analysis:

```
#joining the long version of stressor and colony
#dropping years with missing observations and filtering to national-level data
bees_bar <-
  left_join(colony, stressor) |>
    filter(year != 2019) |>
    filter(state == "United States")
```

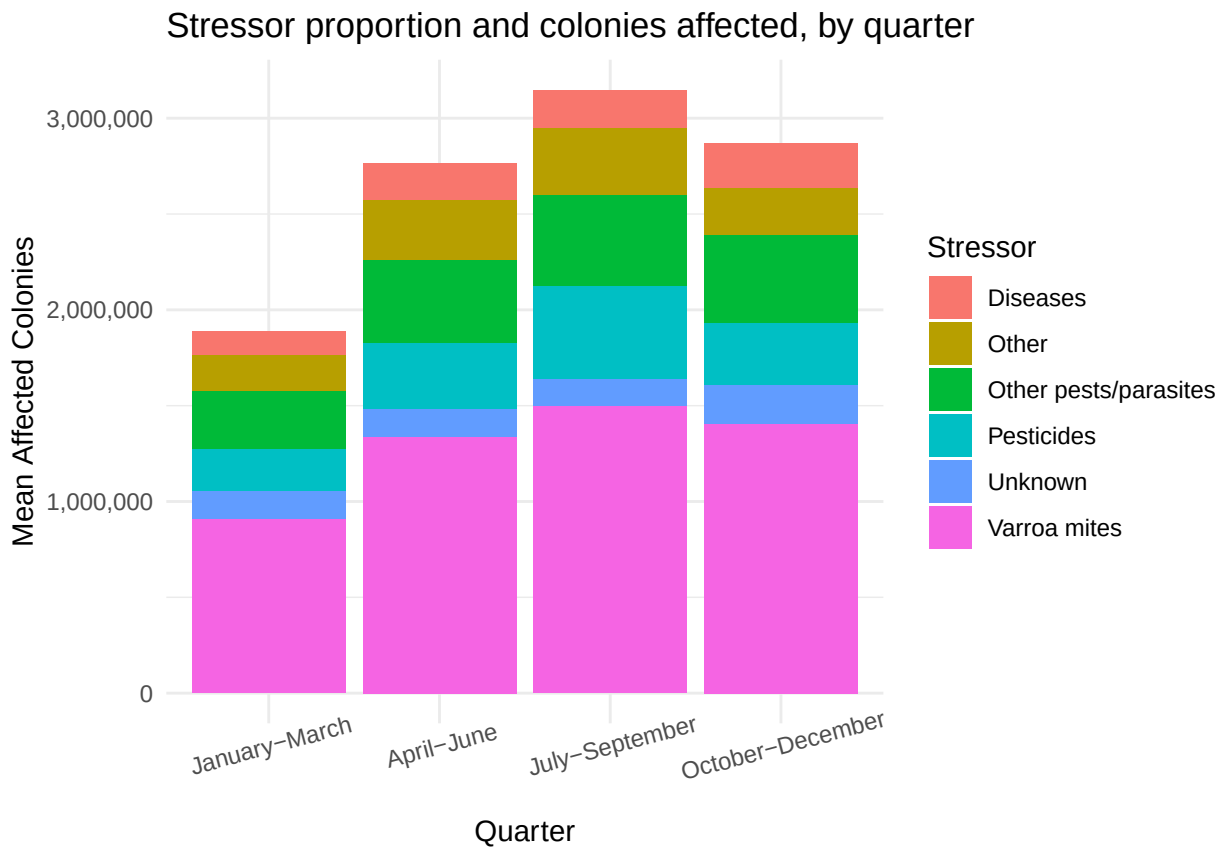
```
## Joining, by = c("year", "months", "state")
```

```
#using mutate() to generate a count of the total number of affected hives for each pest and year
#using group_by() and summarize() to generate a table with clear aesthetic assignments for ggplot
bees_bar <-
  bees_bar |>
    mutate(bees_affected = (colony_n * (stress_pct/100))) |>
    group_by(stressor, months) |>
    summarize(bees_affected = mean(bees_affected))
```

```
## `summarise()` has grouped output by 'stressor'. You can override using the `.groups` argument.
```

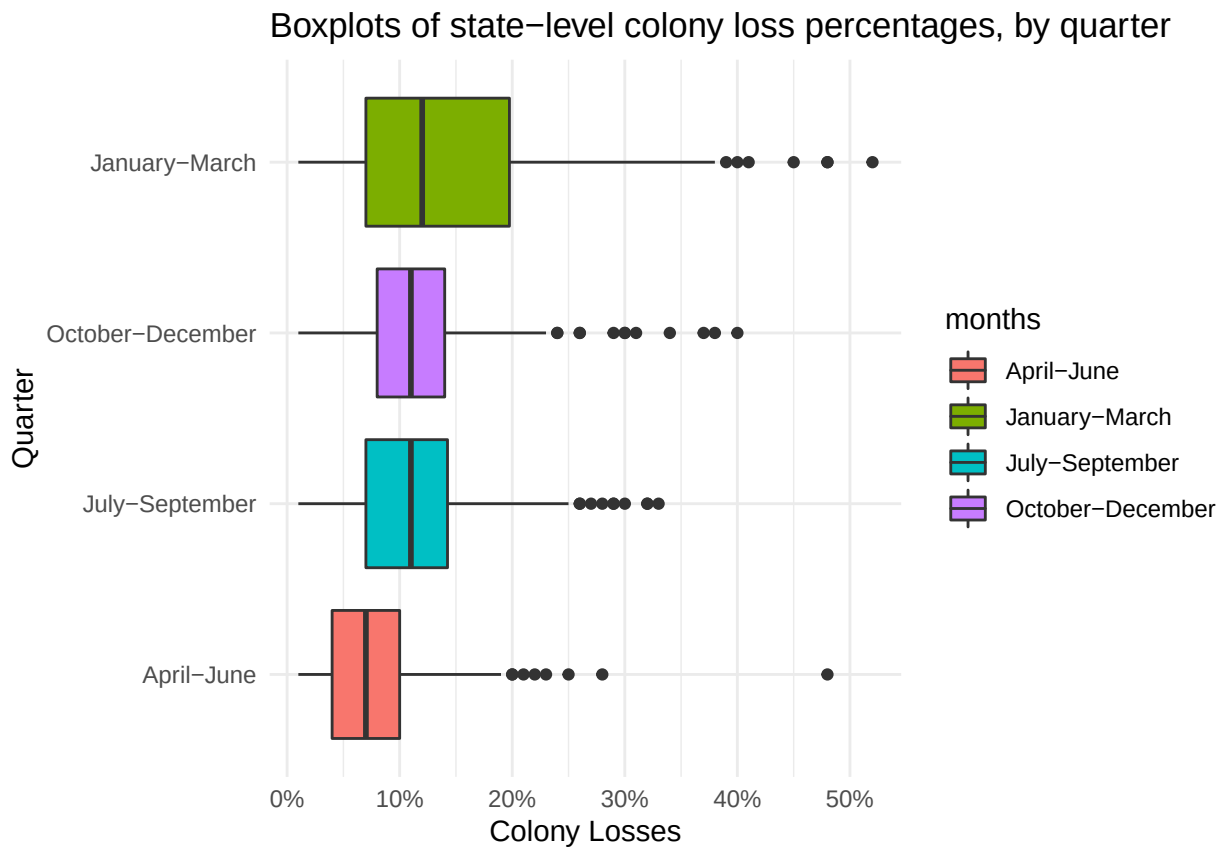
```r
#using factor() to order the variables for ggplot()
bees_bar <-
  bees_bar |>
    mutate(months = factor(months, levels = c("January-March", "April-June", "July-September", "October-

#creating a dataframe for the seasonal boxplots
bees_box <-
  colony |>
    filter(state != "United States") |>
    mutate(colony_lost_pct = (colony_lost_pct/100))
```

```r
#bar graph
  ggplot(
    ) +
  geom_col(data = bees_bar, aes(months, bees_affected, fill = stressor)
    ) +
  labs(fill = "Stressor",
       title = "Stressor proportion and colonies affected, by quarter"
       ) +
  scale_x_discrete(name = "Quarter") +
  scale_y_continuous(name = "Mean Affected Colonies",
                     labels = c("0", "1,000,000", "2,000,000", "3,000,000")) +
  scale_fill_discrete(labels = c("Diseases", "Other", "Other pests/parasites", "Pesticides", "Unknown",
  theme_minimal() +
  theme(axis.text.x = element_text(
      angle = 15))
```




```
#density plots of seasonal distributions of colony losses
bees_box |>
  ggplot(
    aes(colony_lost_pct, reorder(months, colony_lost_pct, na.rm = TRUE), fill = months)
  ) +
  geom_boxplot() +
  scale_x_continuous(name = "Colony Losses",
                     label = scales::label_percent(accuracy = 1)) +
  scale_y_discrete(name = "Quarter") +
  theme(legend.position = "none") +
  labs(title = "Boxplots of state-level colony loss percentages, by quarter") +
  theme_minimal()
```

```
## Warning: Removed 53 rows containing non-finite values (stat_boxplot).
```



Discussion:

The results again match theory, and help explain why the results from the regressions in Part 1 were inconclusive.

The bar graph shows the most colony stress occurring during the peak of summer in the July-September quarter, with slightly lower levels in April-June and October-December. January-March featured a dramatic decrease in stressors, which could be attributed to two potential causes. During the winter bees are less active and maintain lower hive populations, potentially limiting the spread of the parasites that constitute the bulk of CCD stresses. Alternatively, if there is an unobserved causal relationship between pesticide exposure and decreases in resistance to parasites, it could be the case that lower exposure to pesticides due to fallow winter fields is leading to the decreases in pest levels. My results show little proportional variation in stresses between seasons. Despite this, the boxplot results show that the most severe CCD losses occur during the January-March season, the period in which the bar graph finds by far the lowest prevalence of stressors. If colony collapse disorder is directly caused by the stressors, we would expect peak attrition to occur during the summer. So why is the deadliest season for bee colonies during the winter?

This is likely related to the annual cycle of activity in beehives. During the spring and summer, bees work actively to collect ample resources for the winter. During the fall, hives will expel workers to cut food requirements for the remaining bees, then remain largely dormant until spring. Knowing this, it is possible that either because of pesticide exposure resulting in parasite infestation, or due to higher rates of contact between colonies spreading parasites, high amounts of external stress reduce the ability of colonies to stow resources for the winter. Alternatively, colonies could be uniquely vulnerable to stressors during the winter, when lower numbers and a limited metabolism may hinder their ability to fight off pests resulting in an elevated rate of mortality. Increasingly severe winters caused by climate change may contribute in either case.

These patterns would explain why the regression model failed to consistently identify a relationship between the stressors and colony losses, through the omitted variable bias of severe weather or through the delayed

effect of stressors in contributing to CCD. They would also explain the significant collinearity from the PCA, since the stressors change together seasonally. Further, this matches theory - in which colony collapse disorder is described as a process resulting from the cumulative stresses of climate change, increased pests, and pesticide use.