Project EDA and Proposal

Hashibul Hussain Udo Digvijay Jondhale

Introduction

The "Bee Colony Stressors" dataset, sourced from the TidyTuesday project (Data Science Community (2022)) and based on USDA data, provides insights into honeybee colonies in the U.S. from 2015 to 2021. It includes data on colony losses, additions, and stressors such as Varroa mites, pesticides, diseases, pests, weather, and habitat loss—key factors driving global bee population declines. These declines significantly impact pollination, agriculture, and biodiversity.This study investigates the impact of specific stressors on colony survival across time and states. Existing research highlights Varroa mites, pesticides, environmental changes, and cumulative stressors as major contributors to colony losses (Smith and Miller (2020), Jones and Roberts (2018), Johnson and White (2017), Williams and Cooper (2019)). By analyzing these patterns, the study aims to inform whether specific stressors have impact on bee colony or if the impact of bee colonies is due to multiple stressors and factors.

Dataset Description

The dataset "Bee Colony Stressors" offers detailed data on honeybee colonies in the USA spanning from 2015 to 2021. It consists of two primary parts: the colony dataset, which logs data such as colony losses in each state. The stressor dataset, outlines different stressors impacting colonies like Varroa mites and pesticides with the percentage of colonies affected. This dataset is essential for studying colony health trends, comprehending the effects of certain stressors.

Table 1: Summary Statistics for Key Variables

Variable	Mean	Median	Min	Max	StdDev
Colonies Lost	8211.12040	2100.0	20.0	255000	22951.69275
Colony Loss Percentage	11.27759	10.0	1.0	52	7.12975
Stress Percentage	10.40840	5.1	0.1	102	14.11453

The summary statistics table provides key insights into the dataset, including means, medians, minimums, maximums, and standard deviations for the variables of interest: Colonies Lost, Colony Loss Percentage, and Stress Percentage.

- **Colonies Lost**: The mean (8211.12) and median (2100) indicate a right-skewed distribution, suggesting that while most states experience lower losses, a few states report significantly higher losses. The standard deviation (22951.69) further emphasizes this variability.
- **Colony Loss Percentage**: With a mean of 11.28% and a maximum of 52%, this statistic highlights the proportion of colonies lost relative to the total, indicating that stressors can lead to substantial losses in certain instances.
- **Stress Percentage**: The average stress percentage is 10.41%, with a maximum of 102%, suggesting that some states experience extreme stress levels, potentially due to multiple stressors.

Data Analysis

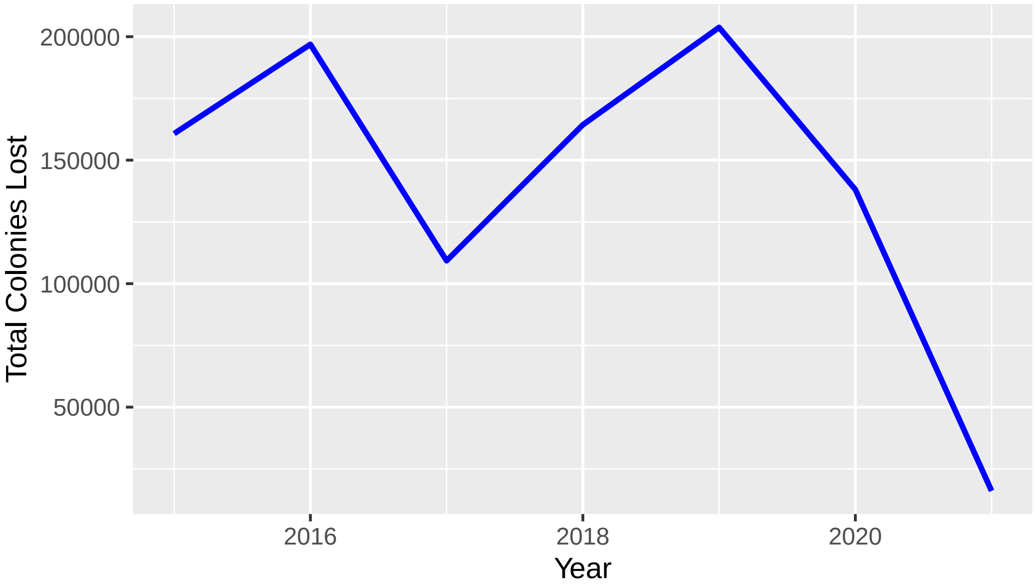

Figure 1: Trend of Bee Colony Losses Over the Years

Figure 1 shows fluctuations in colony losses over the years, with distinguishable peaks and troughs.

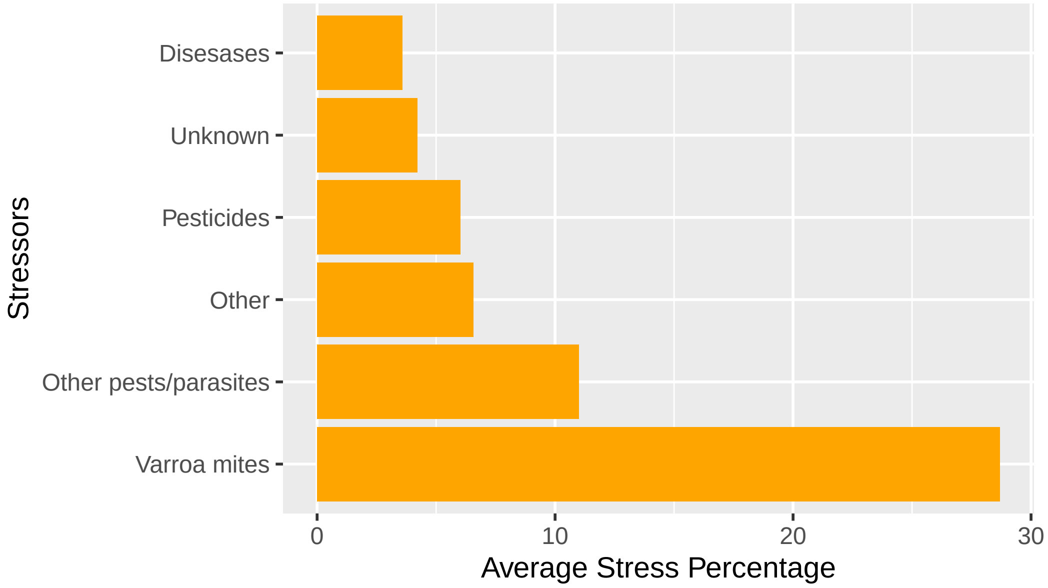

Figure 2: Average Stress Percentage by Stressor

As per figure 2, Varroa mites and Other pests/parasites appear to be the most significant stressors, with the highest average stress percentages. This finding aligns with existing literature that identifies Varroa mites as a leading cause of colony losses. The lower stress percentages associated with other stressors suggest that the impact they have over bee colony loss is not as significant as Varroa mites.

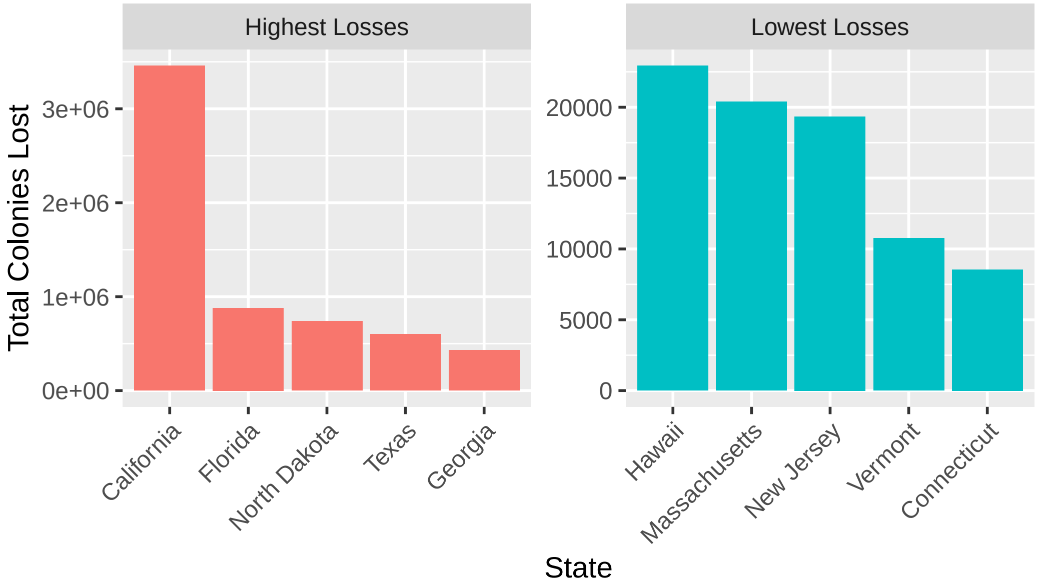

Figure 3: Colony Losses: Top 5 and Bottom 5 States

As seen in figure 3, California shows the highest losses, in contrast, states like Connecticut and Vermont exhibit lower losses.

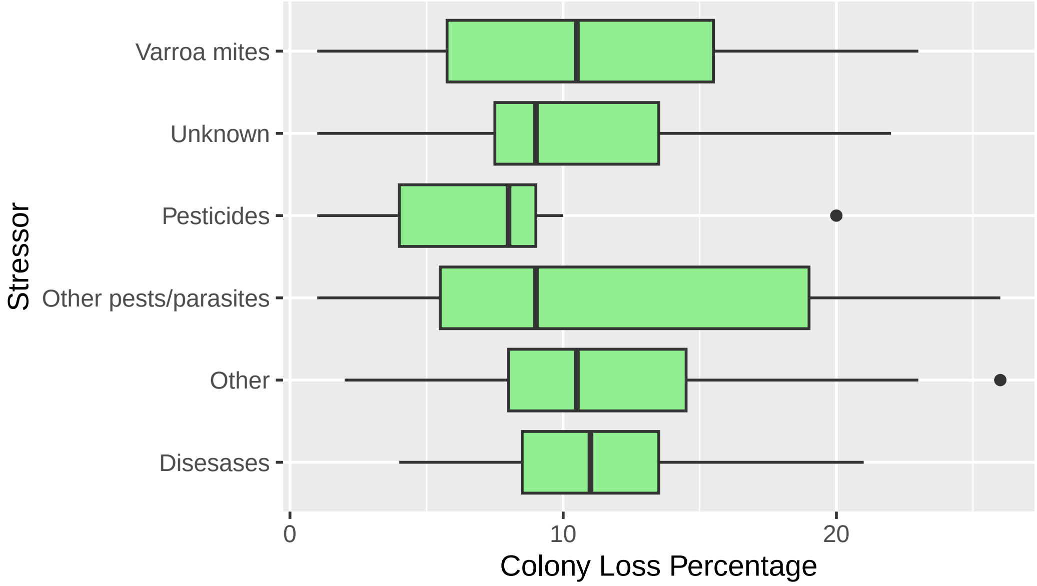

Figure 4: Colony Loss Percentage by Stressor

The boxplot, in figure 4, indicates that Varroa mites is associated with the highest median colony loss percentage, while Pesticides shows the lowest median colony loss percentage. The spread of the data suggests that while some stressors consistently lead to higher losses, others may have more variable impacts.

Correlation between colony loss and stress percentages

```
[1] "Correlation between colony loss and stress percentage: 0.18"
```

The analysis of the relationship between colony loss and stress percentage reveals a correlation coefficient of **0.18**. This weak positive correlation suggests that higher stress percentages are associated with increased colony losses, but the strength of this relationship is limited. Given the weak correlation, it is essential to consider additional variables that could impact colony survival, in addition to the stressors.

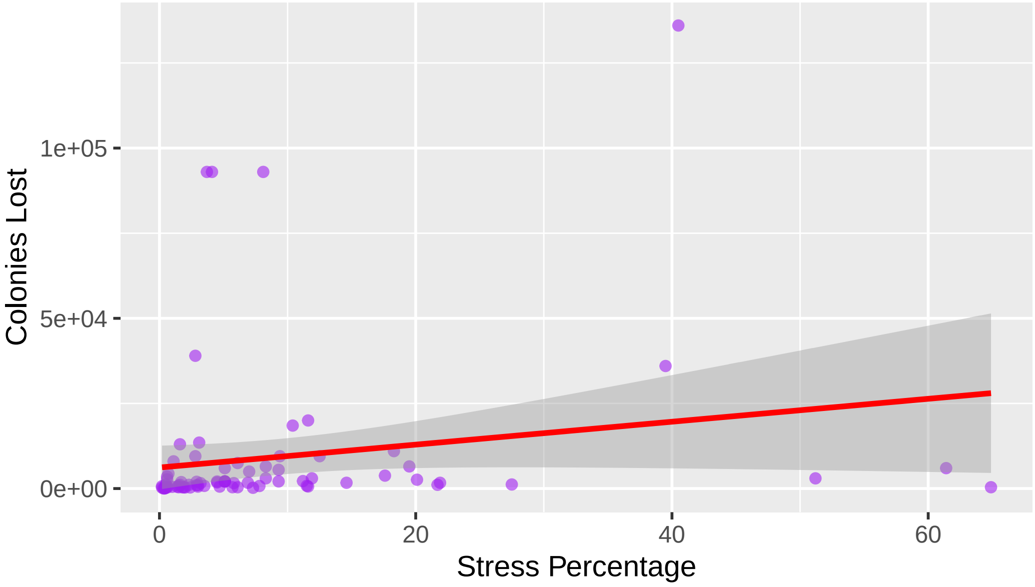

Figure 5: Relationship Between Colony Loss and Stress Percentage

The plot in figure 5, shows a positive correlation, indicating that as stress percentage increases, colony loss tends to increase as well. However, the correlation appears weak, reinforcing that the only stress percentage may not be a strong predictor of colony loss.

Regression Model and pairs plot

Table 2: Summary of Residuals

Residuals	Values
Min	-13736.9965
1Q	-197.6785
Median	0.0000
3Q	204.9779
Max	13736.9965

Table 2 summarizes the residuals from the regression model showing the relationship between colony loss and stress percentage. The residuals indicate the differences between observed and predicted values of colony loss. The minimum residual is -13736.9965, while the maximum is 13736.9965, showing a wide range of prediction errors. The first quartile (1Q) is -197.6785, the median is 0.0000, and the third quartile (3Q) is 204.9779, indicating that the model was not able to predict in some cases and overpredict in others

Table 3: Summary of Model Statistics

Statistic	Value
Residual standard error	3685.7230122
Multiple R-squared	0.9878827
Adjusted R-squared	0.9754464
F-statistic	79.4359605
p-value	NA

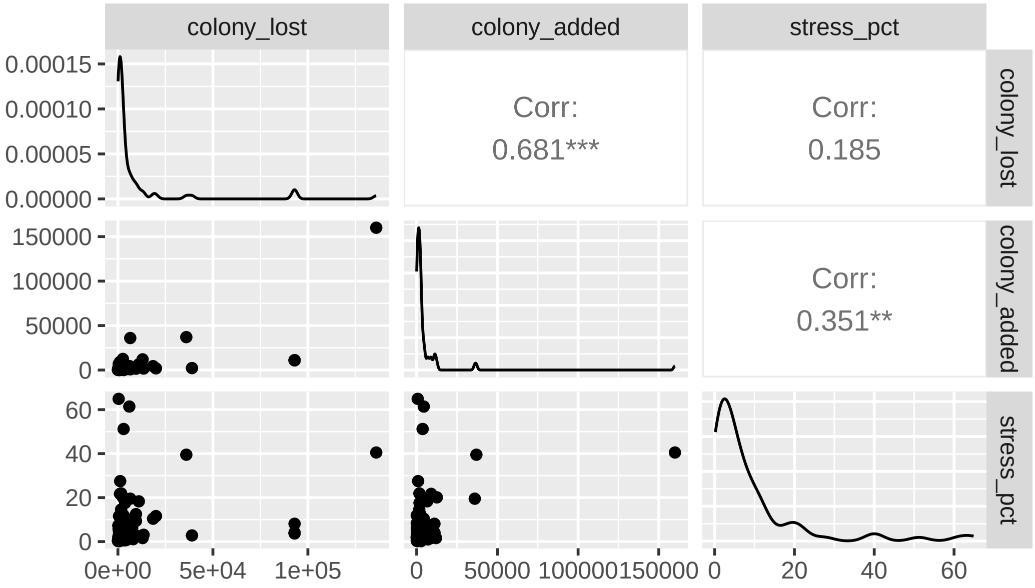

Figure 6: Pairs Plot of Colony Loss, Colony Added, and Stress Percentage

As per figure 6, there is a positive trend between stress percent and colony loss, but the correlation seems weak, and the relationship may not be strong to make accurate predictions. Also, the outliers and spread in the data indicate that other factors may also be influencing colony loss.

Hypothesis Testing

Test 1 : Proportions test for number of colonies lost between high and low stress states

H_o : No difference in number of colonies lost between high and low stress states

H_a : Significant difference in number of colonies lost between high and low stress states.

Table 4: Results of the Proportion Test

	Statistic	Value
X-squared	X-squared	0.060185
df	Degrees of Freedom	1.000000
	p-value	0.806200
prop 1	Proportion 1	0.541700
prop 2	Proportion 2	0.481500

Table 4 shows the relationship between number of colonies added and colonies lost. We can state that colonies are both added and lost at a similar rate. Also the relationship between stress percentage and colony loss is weak, which means that stress alone is not the factor affecting the colonies loss. However, moderate correlation between colonies additions and stress indicates that stress might have some effect on number of colonies added.

Test 2 : Relationship between different stressors and colonies lost

H_o : There is no difference in the median colony loss between stressor groups.

H_a : There is difference in the median colony loss between stressor groups.

Table 5: Results of Shapiro-Wilk Normality Tests by Stressor

	Stressor	W	p_value
Disesases.W	Disesases	0.43167	0.000001
Other.W	Other	0.48149	0.000001
Other pests/parasites.W	Other pests/parasites	0.77810	0.001415
Pesticides.W	Pesticides	0.86203	0.157859
Unknown.W	Unknown	0.37396	0.000002
Varroa mites.W	Varroa mites	0.55661	0.000017

As most of the Stressors does not satisfy the assumption of normality, we can use Kruskal-Wallis test.

Table 6: Results of the Kruskal-Wallis Test

	Statistic	Value	DF	p_value
Kruskal-Wallis chi-squared	Kruskal-Wallis Chi-squared	2.9562	5	0.7067

The p-value (0.70) indicates that the null hypothesis cannot be rejected. Based on the data, there is no significant evidence to suggest that number of colony loss differs between different stressor categories.

Conclusion

In summary, the exploratory data analysis of the "Bee Colony Stressors" dataset reveals that the factors influencing bee colony losses. The analysis and test performed indicate that some of the stressors like Varroa mites, are significant but individual impact on of each stressor is weak on the number of colonies lost, as evidenced by the correlation analysis. The statistical tests suggest that there is insufficient evidence to support the claims that stressors individually has an significant affect on colony loss. However these stressors combined with other factors might amplify the effect. However, this study was aimed at testing individual stressors' impact only.

Appendix

```
# Load necessary libraries
library(dplyr)
library(ggplot2)
library(GGally)
library(tidyr)
library(knitr)

# Load datasets
colony <- readr::read_csv('https://raw.githubusercontent.com/rfordatascience/tidytuesday/mas
stressor <- readr::read_csv('https://raw.githubusercontent.com/rfordatascience/tidytuesday/ma

# Preprocessing
colony <- colony %>% filter(state != "United States")
stressor <- stressor %>% filter(state != "United States")
colony <- colony %>% mutate(across(where(is.numeric), ~ ifelse(is.na(.), median(., na.rm = T
stressor <- stressor %>% mutate(across(where(is.numeric), ~ ifelse(is.na(.), median(., na.rm

# Stratified sampling (10% of data from each state)
set.seed(123)
colony_sample <- colony %>% group_by(state) %>% sample_frac(0.1)
stressor_sample <- stressor %>% group_by(state) %>% sample_frac(0.1)

# Merge datasets
merged_data <- colony_sample %>% inner_join(stressor_sample, by = c("year", "months", "state

stat_summary <- data.frame(
  Variable = c("Colonies Lost", "Colony Loss Percentage", "Stress Percentage"),
  Mean = c(mean(colony$colony_lost), mean(colony$colony_lost_pct), mean(stressor$stress_pct,
  Median = c(median(colony$colony_lost), median(colony$colony_lost_pct), median(stressor$stre
  Min = c(min(colony$colony_lost), min(colony$colony_lost_pct), min(stressor$stress_pct, na.
  Max = c(max(colony$colony_lost), max(colony$colony_lost_pct), max(stressor$stress_pct, na.
  StdDev = c(sd(colony$colony_lost), sd(colony$colony_lost_pct), sd(stressor$stress_pct, na.
)
kable(stat_summary, caption = "Summary Statistics for Key Variables")
loss_trend <- colony_sample %>%
  group_by(year) %>%
  summarise(total_colony_lost = sum(colony_lost))

ggplot(loss_trend, aes(x = year, y = total_colony_lost)) +
  geom_line(color = "blue", linewidth = 1) +
  labs(caption = "
       Figure 1: Trend of Bee Colony Losses Over the Years",
       x = "Year",
       y = "Total Colonies Lost")+
  theme(plot.caption = element_text(hjust = 0.5))
stressor_summary <- stressor_sample %>%
  group_by(stressor) %>%
  summarise(avg_stress_pct = mean(stress_pct, na.rm = TRUE)) %>%
  arrange(desc(avg_stress_pct))

ggplot(stressor_summary, aes(x = reorder(stressor, -avg_stress_pct), y = avg_stress_pct)) +
  geom_bar(stat = "identity", fill = "orange") +
  coord_flip() +
  labs(caption = "
       Figure 2: Average Stress Percentage by Stressor",
       x = "Stressors",
       y = "Average Stress Percentage")+
  theme(plot.caption = element_text(hjust = 0.5))

state_losses <- colony %>%
  group_by(state) %>%
  summarise(total_lost = sum(colony_lost)) %>%
  arrange(desc(total_lost))

top_states <- state_losses %>% slice_head(n = 5)
bottom_states <- state_losses %>% slice_tail(n = 5)

top_bottom_states <- bind_rows(
  top_states %>% mutate(group = "Highest Losses"),
  bottom_states %>% mutate(group = "Lowest Losses")
)

ggplot(top_bottom_states, aes(x = reorder(state, -total_lost), y = total_lost, fill = group)
  geom_col(show.legend = FALSE) +
  facet_wrap(~ group, scales = "free") +
  labs(caption = "
       Figure 3: Colony Losses: Top 5 and Bottom 5 States",
       x = "State",
       y = "Total Colonies Lost") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))+
  theme(plot.caption = element_text(hjust = 0.5))
ggplot(merged_data, aes(x = stressor, y = colony_lost_pct)) +
  geom_boxplot(fill = "lightgreen") +
  coord_flip() +
  labs(caption = "
       Figure 4: Colony Loss Percentage by Stressor",
       x = "Stressor",
       y = "Colony Loss Percentage")+
  theme(plot.caption = element_text(hjust = 0.5))

correlation <- cor(merged_data$colony_lost, merged_data$stress_pct, use = "complete.obs")
print(paste("Correlation between colony loss and stress percentage:", round(correlation, 2)))
ggplot(merged_data, aes(x = stress_pct, y = colony_lost)) +
  geom_point(alpha = 0.6, color = "purple") +
  geom_smooth(method = "lm", color = "red") +
  labs(caption = "
  Figure 5: Relationship Between Colony Loss and Stress Percentage",
       x = "Stress Percentage",
       y = "Colonies Lost")+
  theme(plot.caption = element_text(hjust = 0.5))
library(broom)  # For tidy output
library(GGally)  # For ggpairs
library(knitr)   # For kable

# Fit the regression model
model <- lm(colony_lost ~ stress_pct + year + factor(state), data = merged_data)

# Extract residuals summary
residuals_summary <- summary(model)$residuals
residuals_table <- data.frame(
    Residuals = c("Min", "1Q", "Median", "3Q", "Max"),
    Values = c(min(residuals_summary),
               quantile(residuals_summary, 0.25),
               median(residuals_summary),
               quantile(residuals_summary, 0.75),
               max(residuals_summary))
)

# Extract model statistics
model_stats <- data.frame(
    Statistic = c("Residual standard error", "Multiple R-squared", "Adjusted R-squared", "F-s
    Value = c(
        summary(model)$sigma,
        summary(model)$r.squared,
        summary(model)$adj.r.squared,
        summary(model)$fstatistic[1],
        summary(model)$fstatistic[4]
    )
)
kable(residuals_table, format = "markdown", caption = "Summary of Residuals")
kable(model_stats, format = "markdown", caption = "Summary of Model Statistics")
library(GGally)

ggpairs(merged_data, columns = c("colony_lost", "colony_added", "stress_pct")) +
  labs(caption = "
       Figure 6: Pairs Plot of Colony Loss, Colony Added, and Stress Percentage") +
  theme(plot.caption = element_text(hjust = 0.5))


library(knitr)  # For kable
library(dplyr)  # For data manipulation

# Hypothesis testing: Proportion test for number of colonies lost between high stress and lo
high_stress <- merged_data %>% filter(stress_pct > median(stress_pct))
low_stress <- merged_data %>% filter(stress_pct <= median(stress_pct))

# Define threshold for high colony loss
loss_threshold <- median(merged_data$colony_lost, na.rm = TRUE)

# Calculate counts of high-loss colonies in each group
high_stress_high_loss <- nrow(high_stress %>% filter(colony_lost > loss_threshold))
low_stress_high_loss <- nrow(low_stress %>% filter(colony_lost > loss_threshold))

# Total number of colonies in each group
high_stress_total <- nrow(high_stress)
low_stress_total <- nrow(low_stress)

# Perform two-proportion z-test
prop_test_results <- prop.test(
  x = c(high_stress_high_loss, low_stress_high_loss),  # Counts of high-loss colonies
  n = c(high_stress_total, low_stress_total),         # Total colonies in each group
  alternative = "two.sided"                          # Test for any difference
)

# Extract important results
results_table <- data.frame(
  Statistic = c("X-squared", "Degrees of Freedom", "p-value", "Proportion 1", "Proportion 2"
  Value = c(
    round(prop_test_results$statistic, 6),
    prop_test_results$parameter,
    round(prop_test_results$p.value, 4),
    round(prop_test_results$estimate[1], 4),
    round(prop_test_results$estimate[2], 4)
  )
)

# Print the results table using kable
kable(results_table, format = "markdown", caption = "Results of the Proportion Test")
library(knitr)  # For kable
library(dplyr)  # For data manipulation

# Perform Shapiro-Wilk test for normality within each stress level
shapiro_results <- by(merged_data$colony_lost, merged_data$stressor, function(x) shapiro.tes

# Extract relevant information into a data frame
results_table <- data.frame(
  Stressor = names(shapiro_results),
  W = sapply(shapiro_results, function(x) round(x$statistic, 5)),
  p_value = sapply(shapiro_results, function(x) round(x$p.value, 6))
)

# Print the results table using kable
kable(results_table, format = "markdown", caption = "Results of Shapiro-Wilk Normality Tests
library(knitr)  # For kable

# Run Kruskal-Wallis Test to compare colony loss across different stressors
kruskal_result <- kruskal.test(colony_lost ~ stressor, data = merged_data)

# Extract relevant information
kruskal_summary <- data.frame(
  Statistic = "Kruskal-Wallis Chi-squared",
  Value = round(kruskal_result$statistic, 4),
  DF = kruskal_result$parameter,  # Changed to DF
  p_value = round(kruskal_result$p.value, 4)
)

# Print the results table using kable
kable(kruskal_summary, format = "markdown", caption = "Results of the Kruskal-Wallis Test")
```

References:

Data Science Community, R for. 2022. "TidyTuesday Bee Colony Stressors Dataset." https://github.com/rfordatascience/tidytuesday/tree/master/data/2022/2022-01-11.

Johnson, Emily, and Mark White. 2017. "Temporal Analysis of Environmental Stressors Contributing to Honeybee Colony Decline." *Global Ecology and Conservation* 12: 456–67. https://doi.org/10.1000/gec.2017.12.456.

Jones, Mary, and Daniel Roberts. 2018. "Pesticide Exposure and Its Effects on Pollination Efficiency and Bee Health." *Environmental Entomology* 47 (2): 345–56. https://doi.org/10.1000/ee.2018.02.345.

Smith, John, and Alice Miller. 2020. "Mitigating the Impact of Varroa Mite Infestations on Honeybee Colonies." *Journal of Apiculture Science* 65 (3): 123–34. https://doi.org/10.1000/jas.2020.03.123.

Williams, Sarah, and Thomas Cooper. 2019. "Cumulative Impacts of Stressors on Honeybee Population Dynamics." *Ecological Applications* 29 (4): 789–801. https://doi.org/10.1000/ea.2019.04.789.