Project One

Mallory Tyler & Gabrielle Salvador

2024-02-29

Contents

1 Abstract

Honey bees are one of the most important pollinators in the world because of the variety of crops they pollinate. They are a crucial part of the environment and the economy because of their pollination services when they collect pollen and nectar. The analysts wished to study honey bee colony loss in the United States, and the Bee Informed Partnership (Bee Informed Partnership, n.d.a) conducts a yearly survey on the topic. They then post the data, and well as their analysis of the data, on their website. We downloaded the data and excluded the columns that did not apply to the analysis we wished to perform. The data included information about each state, the season (winter, summer, annual), and the year the data was collected. For the analysis, three states were chosen: Florida, New Jersey, and California. Plots were made for each state showing colony loss per year, and the plots each have lines for the winter, summer, and annual data. Out of the three states, Florida had the biggest changes in the percentage of colonies lost as well as a higher overall percentage loss. We originally hypothesized that winter data would show higher colony loss than summer data, but in Florida the mean colony loss for summer was higher than the mean colony loss for winter. There could be several reasons for this, including higher temperatures in winter compared to other states, hurricanes, or the hotter weather attracting pests and diseases. The correlation coefficient was calculated for each state. There was high correlation between the winter and annual data in Florida and moderate correlation between the variables in New Jersey. The correlation between summer and annual was slightly higher than winter and annual in California. Standard deviation was also calculated, which showed that the winter, summer, and annual data sets for California deviated less than the data sets from the other states. Overall, Florida had the highest loss annually. More research should be done to determine why colony loss is higher in Florida, or to determine which factors impacting honey bees cause the greatest impact on colony loss.

2 Introduction

Bees are one of the most well-known and indispensable pollinators in the world; every one in three bites of food contains products derived from pollination services. They are a crucial part of our environment and economy through their collection of pollen and nectar to meet their dietary needs. Bee pollination enhances crop quality, extends shelf life, and increases commercial value in contrast to wind pollination and self-pollination, thereby contributing positively to the economy (Klatt, B. K., Holzschuh, A., Clough, Y., Smit, I., Pawelzik, E., & Tscharntke, T., 2014). Pollination further plays a crucial role in environmental conservation by facilitating the transfer of pollen between flowers, ensuring successful plant reproduction, and preserving biodiversity. Our interest in bees stems from Mallory's long history with bees, starting from middle school. After reading a book analyzing the absence of bees from our environments, she has advocated for the importance of preserving bee colonies. As a result, we have decided on focusing on the loss of bee colonies within various states through the Bee Informed Partnership.

The Bee Informed Partnership (BIP) provides several important services to beekeepers, researchers, and the general public (Bee Informed Partnership, n.d.a). Being a national nonprofit organization, their mission revolves around enhancing honey bee health through research fueled by data collection and fostering collaboration with beekeepers. BIP employs scientifically-based, data-driven strategies aimed at enhancing the health and long-term sustainability of honey bees, scientifically referred to as Apis mellifera, as well as other vital pollinators. Providing relevant and up-to-date data, BIP assists beekeepers in making informed decisions for colony management (Bee Informed Partnership, n.d.a). Each year, BIP gathers survey data from thousands of beekeepers and collects over 10,000 field samples per season, allowing us to understand the influences of various management practices, forage availability, pests, diseases, nutrition, and environmental factors on honey bee health (Bee Informed Partnership, n.d.a). Out of 20,000 bee species globally, one in four species are at risk of extinction (The Bee Conservancy, n.d.). The objective of this study is to analyze bee colony loss in Florida, California, and New Jersey, comparing the colony loss per year in each state. We further hope to determine if there is a significant amount of loss in one of the states compared to the other two states.

We used (R Core Team, 2023) and (Posit team, 2023) for this project.

Figure 1: Honey Bee in Florida pollinating crimson bottlebrush

Figure 1 was taken by Isabella Flores, an English and Entomology student at the University of Florida.

3 Data

The data utilized in this project originates from the Bee Informed Partnership, which conducts a statewide loss survey. This survey stands as the longest-running national effort to monitor honey bee mortality rates in the U.S. In addition to estimating the level of colony mortality rates in the country, the study also determines the most prevalent practices employed by beekeepers in the U.S. in correlation with colony loss risk. For the purposes of this study, we subsetted the data to include state information, survey year, season, total colony at risk, total colony lost, and total colony alive, where

$$\text{total colonies at risk} = \text{total colonies lost} + \text{total colonies alive}. \tag{1}$$

Although the data set contained 31 columns total, we omitted the other variables due to a lack of clarity of the criteria and context of those columns. In the dataset, the total colonies lost contained some data points which had the value "[R]" or "0", thus we decided to remove those data points as well.

This data is important because it helps estimate the level of colony mortality, which can lead to people having a better understanding and addressing the challenges that honey bee colonies face. Through the Loss and Management Survey, BIP has over a decade of systematically documented loss rates covering the whole U.S. They also get data from all sides of beekeepers, including backyard beekeepers and commercial beekeeping operations. They estimate that respondents represent one in every ten managed honey bee colonies in the U.S. (Bee Informed Partnership, n.d.b).

The BIP provides information about the purpose of the study, eligibility for participants, the procedure, duration, risks, benefits, costs, and confidentiality of the data (Bee Informed Partnership, n.d.c). The purpose of the survey is to document honey bee colony losses and related beekeeping practices in the United States on an annual basis. Participants of the survey must be 18 years or older. Only U.S. beekeepers are included in the estimates. Participants are asked a series of questions about their beekeeping operation and management practices, which will take them approximately 30-45 minutes to complete. The team who created the survey does believe there are any risks associated with completing the survey, and there are no

direct benefits from taking part in the research (Bee Informed Partnership, n.d.c). However, beekeepers can learn more about colony losses and other beekeeping practices. It does not cost money to complete the survey. Finally, participation in the survey is confidential; all answers are secured in a secure, password protected database application that uses SSL encryption. The data collected is analyzed and results are posted on the BIP website and discussed during meetings, workshops, conferences and more. Personal information is never disclosed, and further information on BIP's governance and policies can be found on their website in the Governance and Data Policies section (Bee Informed Partnership, n.d.c).

4 Analysis and Results

4.1 Florida Data

```
bee = read.csv("bee.csv")

#Florida Winter
fl_wint = subset(bee, initials == "FL" & Season == "Winter")
logic.vector = logical(length(fl_wint$TotalColLost)) #create for removing unnecessary rows

for(i in seq_along(fl_wint$TotalColLost)){ #iterate by index and change values
  if(fl_wint$TotalColLost[i]=="[R]"){
    logic.vector[i] = FALSE
  }else{
    logic.vector[i] = TRUE
  }
}

fl_wint = fl_wint[logic.vector,] #remove elements with non-numerical values

florida.winter.data <- data.frame(fl_wint$SurveyYear, fl_wint$Season,
    fl_wint$TotalColAtRisk, fl_wint$TotalColLost)
#create new data frame to calculate avg loss per year

combine.florida.winter <- florida.winter.data %>%
  group_by(fl_wint.SurveyYear) %>% #group data points by year
  summarize(fl_wint.TotalColAtRisk = sum(fl_wint.TotalColAtRisk),
            fl_wint.TotalColLost = sum(strtoi(fl_wint.TotalColLost)))
#summarize total at risk and lost to determine overall avg of yr

combine.florida.winter$ratio = as.integer(combine.florida.winter$fl_wint.TotalColLost)/
  combine.florida.winter$fl_wint.TotalColAtRisk

plot(combine.florida.winter$fl_wint.SurveyYear, combine.florida.winter$ratio,
     type="l", main="Florida Colony Loss per Year", xlab = "Survey Year",
     ylab = "Percent Colony Lost", col='blue', ylim = c(0, 1))



# florida summer
fl_sum = subset(bee, initials == "FL" & Season == "Summer")
logic.vector2 = logical(length(fl_sum$TotalColLost)) #create for removing unnecessary rows
```

```r
for(i in seq_along(fl_sum$TotalColLost)){ #iterate by index and change values
  if(fl_sum$TotalColLost[i]=="[R]"){
    logic.vector2[i] = FALSE
  }else{
    logic.vector2[i] = TRUE
  }
}

fl_sum = fl_sum[logic.vector2,] #remove elements with non-numerical values

florida.summer.data <- data.frame(fl_sum$SurveyYear, fl_sum$Season,
                                  fl_sum$TotalColAtRisk, fl_sum$TotalColLost)

combine.florida.summer <- florida.summer.data %>%
  group_by(fl_sum.SurveyYear) %>%
  summarize(fl_sum.TotalColAtRisk = sum(fl_sum.TotalColAtRisk),
            fl_sum.TotalColLost = sum(strtoi(fl_sum.TotalColLost)))

combine.florida.summer$ratio = as.integer(combine.florida.summer$fl_sum.TotalColLost)/
  combine.florida.summer$fl_sum.TotalColAtRisk
combine.florida.summer = combine.florida.summer[complete.cases(combine.florida.summer),]


lines(x=combine.florida.summer$fl_sum.SurveyYear, y=combine.florida.summer$ratio, col='red')




# florida annual
fl_annual = subset(bee, initials == "FL" & Season == "Annual")
logic.vector3 = logical(length(fl_annual$TotalColLost)) #create for removing unnecessary rows

for(i in seq_along(fl_annual$TotalColLost)){ #iterate by index and change values
  if(fl_annual$TotalColLost[i]=="[R]"){
    logic.vector3[i] = FALSE
  }else{
    logic.vector3[i] = TRUE
  }
}

fl_annual = fl_annual[logic.vector3,] #remove elements with non-numerical values


florida.annual.data <- data.frame(fl_annual$SurveyYear, fl_annual$Season,
                                  fl_annual$TotalColAtRisk, fl_annual$TotalColLost)

combine.florida.annual <- florida.annual.data %>%
  group_by(fl_annual.SurveyYear) %>%
  summarize(fl_annual.TotalColAtRisk = sum(fl_annual.TotalColAtRisk),
            fl_annual.TotalColLost = sum(strtoi(fl_annual.TotalColLost)))

combine.florida.annual$ratio = as.integer(combine.florida.annual$
               fl_annual.TotalColLost)/combine.florida.annual$fl_annual.TotalColAtRisk
```

```
combine.florida.annual = combine.florida.annual[complete.cases(combine.florida.annual),]


lines(x=combine.florida.annual$fl_annual.SurveyYear,
      y=combine.florida.annual$ratio, col='green')
legend(x='topleft', legend=c('Winter', 'Summer','Annual'),
       fill=c('blue', 'red', 'green'), cex=1, title = "Season")
```

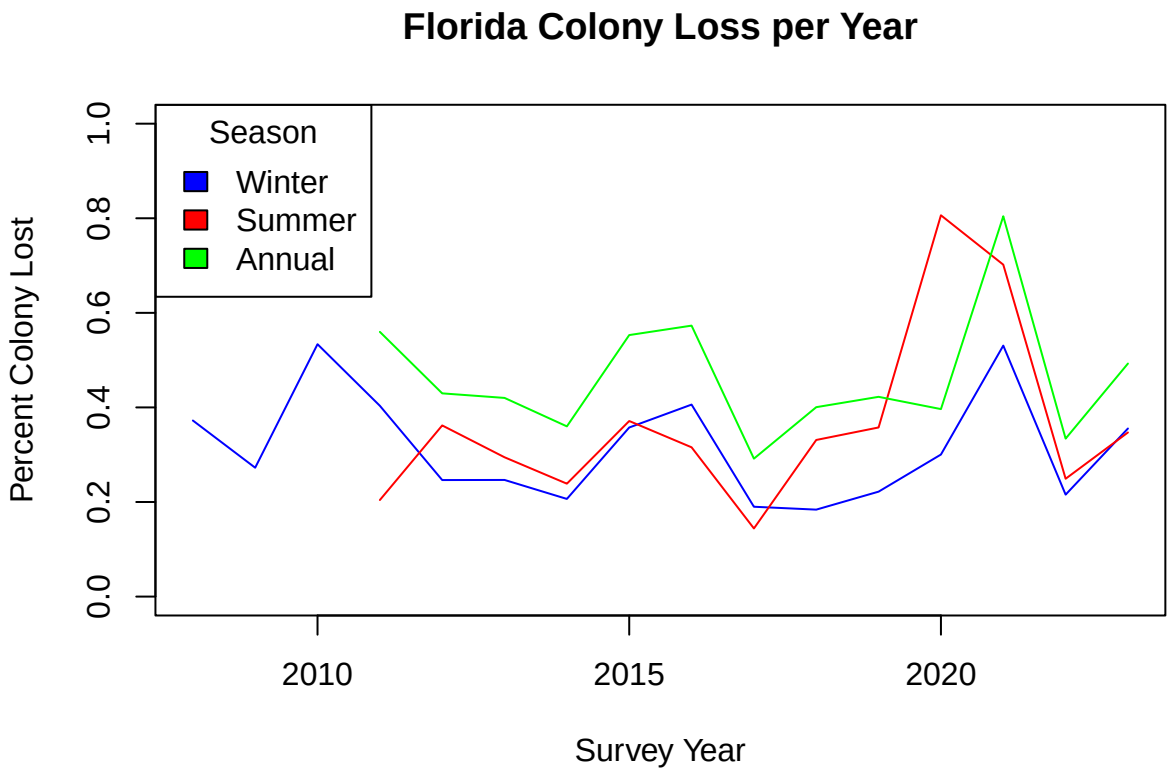

Figure 2: Florida Data

Figure 2 shows the average colony loss per year in Florida in the winter, summer, and annual seasons.

Out of the 3 graphs, Florida's data has the biggest changes in the percentage of colonies lost along with a higher overall percentage loss. The mean percentage loss is 31.5173323% in the winter data, 36.3357774% in the summer data, and 46.4438816% in the annual data. For reference, the mean percentage loss is calculated as

$$\bar{x} = \frac{\Sigma_{i=1}^{n} x_i}{n} \tag{2}$$

where x represents the percentage lost. Furthermore, the highest percent loss of the winter seasons was 53.3670779% during 2010, while the highest percentage loss during the summer seasons was 80.6337633% during 2020 and 80.4249699% annually during 2021. In our original hypothesis, we expected the highest percentage loss for the winter data set to be much higher than that of the summer data sets due to the incompatibility of the winter season and retention of colonies. However, this hypothesis was rejected as the highest colony loss of the summer data was actually much higher than that of the winter data. In addition to this, the mean of the colonies lost in the summer seasons was higher than the mean of the colonies lost in

the winter seasons, showing that the overall loss of colonies in the summer was higher than the loss during the winter. This likely may be due to the fact that Florida's climate during the winter is warmer than the climate of California and New Jersey and the weather during the summer in Florida is much rougher on the colonies, as there is often a lot of heavy thunderstorms as well as hurricanes. The warmer weather also causes a larger amount of pests and viruses to circulate, likely adding onto the loss of colonies. Varroa mites, the main pests for bees as well as their highest threat, thrives in warmer weather as well.

The correlation coefficient is calculated using the formula

$$r_{xy} = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\Sigma(x_i - \bar{x})^2 \Sigma(y_i - \bar{y})^2}} \tag{3}$$

where r_{xy} is the correlation coefficient between x and y, x_i and y_i are the actual values of x and y, and $\bar{x}$ and $\bar{y}$ are the means of the x and y values. For the case of this study, the x and y represents the different data of the summer or winter and annual data sets, respectively. To check whether the winter or summer data had a higher effect on the trend of the annual data, we conducted a correlation analysis, where the correlation coefficient between the winter and annual data was 0.943635, indicating a very strong correlation between the two data sets. On the other hand, the correlation between the summer and annual data was much lower, with a correlation coefficient of 0.4504686, indicating that the summer data had a much lower effect on the annual data. The trends of Figure 2 exemplifies this perfectly, as the winter data and annual data follow a very similar trend while the summer data follows a slightly different trend.

4.2 New Jersey Data

```
#bee = read.csv("bee.csv")

#New Jersey Winter
nj_wint = subset(bee, initials == "NJ" & Season == "Winter")
logic.vector = logical(length(nj_wint$TotalColLost))
#create for removing unnecessary rows

for(i in seq_along(nj_wint$TotalColLost)){ #iterate by index and change values
  if(nj_wint$TotalColLost[i]=="[R]"||nj_wint$TotalColLost[i]=="0"){
    logic.vector[i] = FALSE
  }else{
    logic.vector[i] = TRUE
  }
}

nj_wint = nj_wint[logic.vector,] #remove elements with non-numerical values
nj_wint = nj_wint[nj_wint$SurveyYear!="2008",]

nj.winter.data <- data.frame(nj_wint$SurveyYear, nj_wint$Season,
    nj_wint$TotalColAtRisk, nj_wint$TotalColLost)
#create new data frame to calculate avg loss per year

combine.nj.winter <- nj.winter.data %>%
  group_by(nj_wint.SurveyYear) %>% #group data points by year
  summarize(nj_wint.TotalColAtRisk = sum(nj_wint.TotalColAtRisk),
            nj_wint.TotalColLost = sum(strtoi(nj_wint.TotalColLost)))
#summarize total at risk and lost to determine overall avg of yr

combine.nj.winter$ratio = as.integer(combine.nj.winter$nj_wint.TotalColLost)/
```

```
  combine.nj.winter$nj_wint.TotalColAtRisk

plot(combine.nj.winter$nj_wint.SurveyYear, combine.nj.winter$ratio,
     type="l", main="New Jersey Colony Loss per Year", xlab = "Survey Year",
     ylab = "Percent Colony Lost", col='blue', ylim = c(0, 1))


# New Jersey summer
nj_sum = subset(bee, initials == "NJ" & Season == "Summer")
logic.vector2 = logical(length(nj_sum$TotalColLost))
#create for removing unnecessary rows

for(i in seq_along(nj_sum$TotalColLost)){ #iterate by index and change values
  if(nj_sum$TotalColLost[i]=="0"||nj_sum$TotalColLost[i]=="[R]"){
    logic.vector2[i] = FALSE
  }else{
    logic.vector2[i] = TRUE
  }
}

nj_sum = nj_sum[logic.vector2,] #remove elements with non-numerical values

nj.summer.data <- data.frame(nj_sum$SurveyYear, nj_sum$Season,
                             nj_sum$TotalColAtRisk, nj_sum$TotalColLost)

combine.nj.summer <- nj.summer.data %>%
  group_by(nj_sum.SurveyYear) %>%
  summarize(nj_sum.TotalColAtRisk = sum(nj_sum.TotalColAtRisk),
            nj_sum.TotalColLost = sum(strtoi(nj_sum.TotalColLost)))

combine.nj.summer$ratio = as.integer(combine.nj.summer$nj_sum.TotalColLost)/
  as.integer(combine.nj.summer$nj_sum.TotalColAtRisk)
combine.nj.summer = combine.nj.summer[complete.cases(combine.nj.summer),]

lines(x=combine.nj.summer$nj_sum.SurveyYear, y=combine.nj.summer$ratio, col='red')


# New Jersey annual
nj_annual = subset(bee, initials == "NJ" & Season == "Annual")
logic.vector3 = logical(length(nj_annual$TotalColLost))
#create for removing unnecessary rows

for(i in seq_along(nj_annual$TotalColLost)){ #iterate by index and change values
  if(nj_annual$TotalColLost[i]=="[R]"||nj_annual$TotalColLost[i]=="0"){
    logic.vector3[i] = FALSE
  }else{
    logic.vector3[i] = TRUE
  }
}

nj_annual = nj_annual[logic.vector3,] #remove elements with non-numerical values

nj.annual.data <- data.frame(nj_annual$SurveyYear, nj_annual$Season,
```

```
                              nj_annual$TotalColAtRisk, nj_annual$TotalColLost)

combine.nj.annual <- nj.annual.data %>%
  group_by(nj_annual.SurveyYear) %>%
  summarize(nj_annual.TotalColAtRisk = sum(nj_annual.TotalColAtRisk),
            nj_annual.TotalColLost = sum(strtoi(nj_annual.TotalColLost)))

combine.nj.annual$ratio = as.integer(combine.nj.annual$nj_annual.TotalColLost)/
  combine.nj.annual$nj_annual.TotalColAtRisk
combine.nj.annual = combine.nj.annual[complete.cases(combine.nj.annual),]

lines(x=combine.nj.annual$nj_annual.SurveyYear,
      y=combine.nj.annual$ratio, col='green')
legend(x='topleft', legend=c('Winter', 'Summer','Annual'),
       fill=c('blue', 'red', 'green'), title = "Season")
```

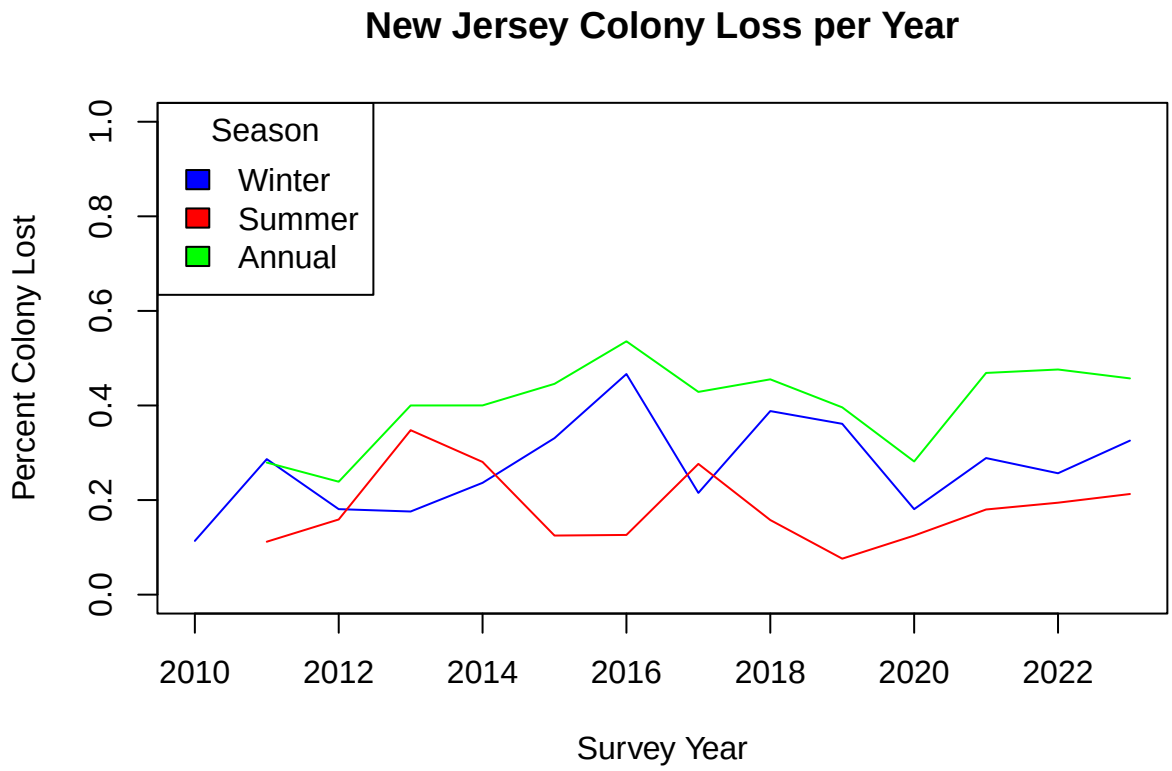

Figure 3: New Jersey Data

Figure 3 shows the average colony loss per year in New Jersey in the winter, summer, and annual seasons.

The highest percentage of colonies lost for the winter data set is 46.6480447% during 2016, 34.7766396% in 2013 for the summer data, and 53.5683297% in 2016 for the annual data. This supports our null hypothesis of a higher percentage lost during the winter season, which coincides with the fact that the New Jersey weather varies much more between the summer and winter seasons. The means of each data set are 27.1942213% for the winter data, 18.2525939% for the summer data, and 40.4883645% for the annual data, further supporting our hypothesis as the mean of the winter data is 10% higher than the summer data.

The correlation analysis for the New Jersey data further shows a strong correlation between the winter and annual data, with a correlation coefficient of 0.6358429, displaying a moderate correlation, which is much higher than the correlation between the summer and annual data, with a correlation coefficient of a mere 0.1584301, meaning that there is almost no correlation between the summer data and annual data. This is further supported by Figure 3, as the trend lines for the winter and annual data sets are moderately similar while the trend line for the summer data set is nearly opposite, such that the percent loss increases when there is a decrease in the winter and annual data and vice versa.

4.3 California Data

```
#California Winter
ca_wint = subset(bee, initials == "CA" & Season == "Winter")
logic.vector = logical(length(ca_wint$TotalColLost))
#create for removing unnecessary rows

for(i in seq_along(ca_wint$TotalColLost)){ #iterate by index and change values
  if(ca_wint$TotalColLost[i]=="[R]"||ca_wint$TotalColLost[i]=="0"){
    logic.vector[i] = FALSE
  }else{
    logic.vector[i] = TRUE
  }
}

ca_wint = ca_wint[logic.vector,] #remove elements with non-numerical values

cali.winter.data <- data.frame(ca_wint$SurveyYear, ca_wint$Season,
          ca_wint$TotalColAtRisk, ca_wint$TotalColLost)
#create new data frame to calculate avg loss per year

combine.cali.winter <- cali.winter.data %>%
  group_by(ca_wint.SurveyYear) %>% #group data points by year
  summarize(ca_wint.TotalColAtRisk = sum(ca_wint.TotalColAtRisk),
            ca_wint.TotalColLost = sum(strtoi(ca_wint.TotalColLost)))
#summarize total at risk and lost to determine overall avg of yr

combine.cali.winter$ratio = as.integer(combine.cali.winter$
              ca_wint.TotalColLost)/combine.cali.winter$ca_wint.TotalColAtRisk

plot(combine.cali.winter$ca_wint.SurveyYear, combine.cali.winter$ratio,
     type="l", main="California Colony Loss per Year", xlab = "Survey Year",
     ylab = "Percent Colony Lost", col='blue', ylim = c(0, 1))


# California summer
ca_sum = subset(bee, initials == "CA" & Season == "Summer")
logic.vector2 = logical(length(ca_sum$TotalColLost))
#create for removing unnecessary rows

for(i in seq_along(ca_sum$TotalColLost)){ #iterate by index and change values
  if(ca_sum$TotalColLost[i]=="0"||ca_sum$TotalColLost[i]=="[R]"){
    logic.vector2[i] = FALSE
  }else{
```

```
    logic.vector2[i] = TRUE
  }
}

ca_sum = ca_sum[logic.vector2,] #remove elements with non-numerical values

cali.summer.data <- data.frame(ca_sum$SurveyYear, ca_sum$Season,
                               ca_sum$TotalColAtRisk, ca_sum$TotalColLost)

combine.cali.summer <- cali.summer.data %>%
  group_by(ca_sum.SurveyYear) %>%
  summarize(ca_sum.TotalColAtRisk = sum(ca_sum.TotalColAtRisk),
            ca_sum.TotalColLost = sum(strtoi(ca_sum.TotalColLost)))

combine.cali.summer$ratio = as.integer(combine.cali.summer$
                ca_sum.TotalColLost)/combine.cali.summer$ca_sum.TotalColAtRisk
combine.cali.summer = combine.cali.summer[complete.cases(combine.cali.summer),]

lines(x=combine.cali.summer$ca_sum.SurveyYear,
      y=combine.cali.summer$ratio, col='red')


# California annual
ca_annual = subset(bee, initials == "CA" & Season == "Annual")
logic.vector3 = logical(length(ca_annual$TotalColLost))
#create for removing unnecessary rows

for(i in seq_along(ca_annual$TotalColLost)){ #iterate by index and change values
  if(ca_annual$TotalColLost[i]=="[R]"||ca_annual$TotalColLost[i]=="0"){
    logic.vector3[i] = FALSE
  }else{
    logic.vector3[i] = TRUE
  }
}

ca_annual = ca_annual[logic.vector3,] #remove elements with non-numerical values

cali.annual.data <- data.frame(ca_annual$SurveyYear, ca_annual$Season,
                               ca_annual$TotalColAtRisk, ca_annual$TotalColLost)

combine.cali.annual <- cali.annual.data %>%
  group_by(ca_annual.SurveyYear) %>%
  summarize(ca_annual.TotalColAtRisk = sum(ca_annual.TotalColAtRisk),
            ca_annual.TotalColLost = sum(strtoi(ca_annual.TotalColLost)))

combine.cali.annual$ratio = as.integer(combine.cali.annual$
          ca_annual.TotalColLost)/combine.cali.annual$ca_annual.TotalColAtRisk
combine.cali.annual = combine.cali.annual[complete.cases(combine.cali.annual),]

lines(x=combine.cali.annual$ca_annual.SurveyYear,
      y=combine.cali.annual$ratio, col='green')
legend(x='topleft', legend=c('Winter', 'Summer','Annual'),
       fill=c('blue', 'red', 'green'), title = "Season")
```

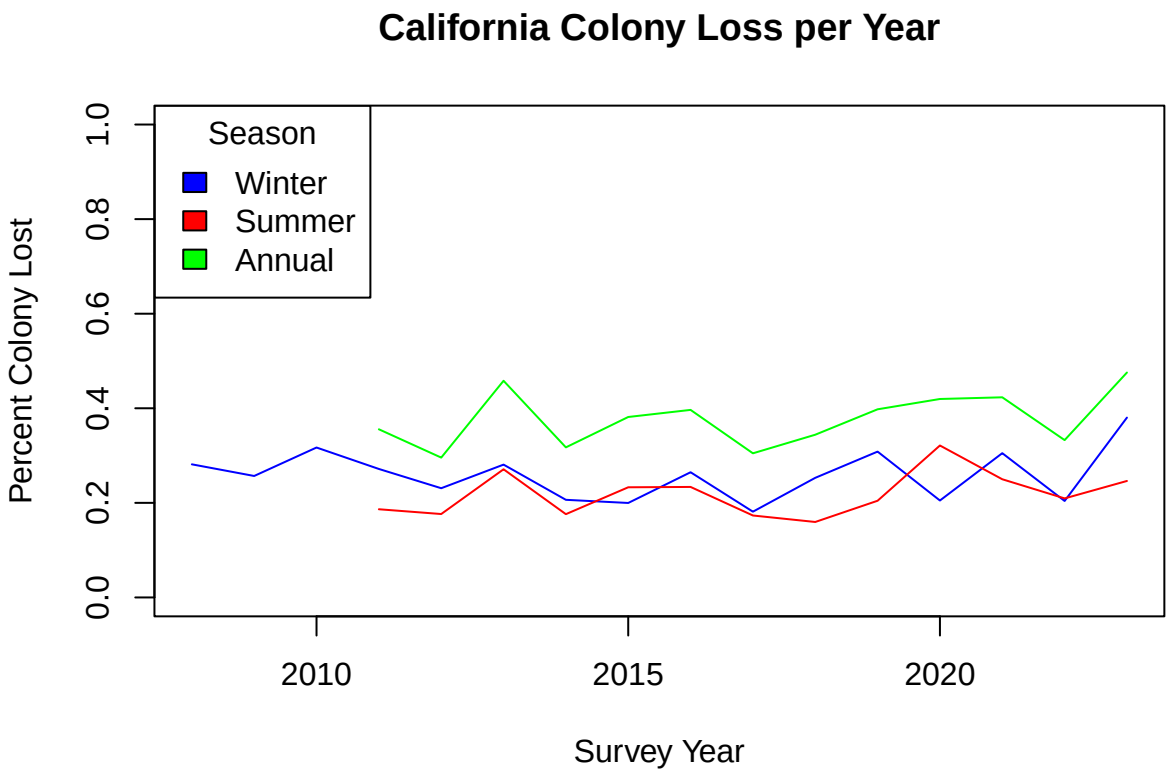

Figure 4: California Data

Figure 4 shows the average colony loss per year in California in the winter, summer, and annual seasons.

California's data, however, varies a lot less compared to the variation of the Florida and New Jersey data, as the standard deviations of the winter, summer and annual data sets are 5.3473492%, 4.6476957%, and 5.7724199%, respectively. The standard deviation is calculated using the formula

$$\sigma = \sqrt{\frac{\Sigma(x_i - \mu)^2}{n}} \tag{4}$$

where x_i represents the values of the data set, μ represents the data set's mean, and n represents the number of data points. To put this into perspective, the deviations of the Florida data are 13.4285619%, 11.3067735%, and 18.7357714%, which is nearly triple the deviation of the California data, while the deviations of the New Jersey data are 9.6807525%, 7.8586399%, and 8.7784992%, which is approximately double that of the California data. This explains why Figure 4, showing the California data, is much flatter in comparison to Figure 2, the Florida data, and Figure 3, the New Jersey data.

Additionally, the maximum percentage lost of the winter data is 38.0275008% in 2023, 32.13012% in 2020 for the summer data, and 47.5578236% in 2023 for the annual data. This supports our null hypothesis as well, as the maximum percentage lost is higher in the winter data than the summer data. This claim is further supported as the mean for the winter data is 25.9155439%, 21.8517001% for the summer data, and 37.7126641% for the annual data.

Using a correlation analysis, the correlation between the winter and annual data sets is 0.7084946, while the correlation between the summer and annual data sets is 0.7813815, showing that the correlation between the summer and annual data has a slightly higher correlation in comparison with the correlation between the winter and annual data. We can see this in Figure 4, as the summer and annual trend lines follow closer than the trend lines of the winter and annual data. Interestingly, however, the correlation between the winter and summer data is only 0.2001514, meaning that there is very little similarity between the trends of the winter and summer data sets, despite the fact that they both similarly correlate with the annual data.

It is important to note that the correlation analysis used for Florida, California, and New Jersey called for a need to omit the extra data points of the winter data sets such that the winter data sets started from 2011 rather than 2008 for California and Florida, and 2010 for New Jersey.

5 Conclusion

Initially, we predicted that winter would have a higher effect on the annual losses throughout Florida, California and New Jersey, which was rejected following our analysis of our data. By using the data from the Bee Informed Partnership, we were able to create graphs for the Florida, California, and New Jersey data. We chose Florida, California, and New Jersey specifically as we wanted to accommodate for the varying climates throughout the country. Overall, Florida has the highest loss annually, as the mean loss was 46.4438816%, while the mean loss for California was 37.7126641% and the mean loss for New Jersey was very similar to California, as it was 40.4883645% despite the fact that California deviated much less in comparison to New Jersey and Florida. Florida also had a very high peak of annual loss of 80.4249699% during 2021 while both California and New Jersey did not have a significant peak around the same time. The winter loss in Florida also had a higher correlation with the annual loss than the summer while New Jersey had a moderate correlation. The correlation in California between summer and annual data was slightly higher than winter and annual. In the future, we hope to do a more extensive study as to what may be causing the discrepancies with the Florida data as well as look into the meaning of the columns we originally omitted. We would also like to look into the loss in other parts of the country, for example a state in the Midwest. It may also be beneficial to investigate why there is a lack of variation in the California losses as opposed to the Florida and New Jersey data, as well as why the trends between the winter and summer data for New Jersey seemed nearly opposite of each other.

6 Contribution Statement

6.1 Mallory

- Abstract
- Introduction
- Data section (found data set)
- Bibliography and citation
- Check analysis and results
- Check conclusion

6.2 Gabby

- Check abstract
- Edit introduction section
- Edit data section
- Analysis and Results (chose which tests and statistics to use for the analysis)
- Conclusion

6.3 Both

- Plotting each state (each member plotted the states, then compared with other member)
- Debugging the code

Works Cited

Bee Informed Partnership. (n.d.a). *Bee informed partnership.* Bee Informed Partnership. https://beeinformed.org/

Bee Informed Partnership. (n.d.b). *Bee informed partnership data request page.* Bee Informed Partnership. https://bip2.beeinformed.org/survey/data_request/1/

Bee Informed Partnership. (n.d.c). *Bee informed partnership survey page.* Bee Informed Partnership. https://beeinformed.org/take-survey/

Klatt, B. K., Holzschuh, A., Clough, Y., Smit, I., Pawelzik, E., & Tscharntke, T. (2014). *Bee pollination improves crop quality, shelf life and commercial value.* The Royal Society Publishing. https://royalsocietypublishing.org/doi/10.1098/rspb.2013.2440

Posit team. (2023). *RStudio: Integrated development environment for r.* Posit Software, PBC. http://www.posit.co/

R Core Team. (2023). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing. https://www.R-project.org/

The Bee Conservancy. (n.d.). *Protecting bees, building habitat, and strengthening communities together.* The Bee Conservancy. https://thebeeconservancy.org/