Explore Boston and Cambridge Trees

Insert Your Name

2023-07-24

Objectives

In this worksheet we are going to explore data on trees in Boston and Cambridge. Let's be statistical sleuths and practice wrangling and visualizing data!

The Datasets

The data live in the `R` package: `bosTrees`.

```
# Run this chunk once to install the package
install.packages("devtools")
devtools::install_github("harvard-ufds/bosTrees")
```

There are two datasets in the package:

- `bosTrees`: includes data on 6,836 primary street trees located throughout Boston and was collected by the City of Boston's GIS team.
- `camTrees`: includes data on 38,050 trees and tree planting sites owned, planted or maintained by the City of Cambridge, the Massachusetts Department of Conservation and Recreation, MIT, Harvard University, and other private organizations. It is maintained by the Cambridge Public Works.

```
# Load the package with the data
library(bosTrees)

# Load the data
data(bosTrees)
data(camTrees)
```

Problem 1 Head up to the **Environment** tab and click on the two datasets. Start to explore what variables are contained in the datasets. What do you think the variables mean? Which variables might be interesting to summarize and graph?

```
#  Run this chunk to see the help files, which describe the variables
?bosTrees
?camTrees
```

Problem 2 How many trunks does a tree have?

a. Run the following code and explain what is being computed.

```
count(camTrees, Trunks)
```

```
## # A tibble: 15 x 2
##    Trunks     n
##     <dbl> <int>
##  1      0 11110
```

```
##  2      1 25950
##  3      2   399
##  4      3   339
##  5      4   126
##  6      5    64
##  7      6    26
##  8      7    13
##  9      8    11
## 10      9     3
## 11     10     2
## 12     11     3
## 13     12     1
## 14     15     2
## 15     16     1
```

Answer: For each number of trunks, R is counting the number of street trees in Cambridge with that number of trunks. So, 11,110 trees have 0 trunks, 25,950 have 1 trunk, ...

---

b. Is it really possible to have 15 or 16 trunks?? What is the `CommonName`, `Latitude`, and `Longitude` of the trees that supposedly has at least 15 trunks?

```
# Code chunk is optional
# You can also use the View window to identify these trees
```

```
# Only need the first line
filter(camTrees, Trunks >= 15) %>%
  select(CommonName, Latitude, Longitude, Trunks) %>%
  as.data.frame() # To get more digits
```

```
##     CommonName Latitude Longitude Trunks
## 1 Serviceberry 42.36623 -71.08012     16
## 2 Serviceberry 42.37478 -71.10987     15
## 3 Serviceberry 42.36626 -71.08035     15
```

Answer: They are all Serviceberries.

---

c. For the trees identified in (b) use Google Maps to try to find to a photo of these trees. Why can we only find an image of the three trees?

Hint: Determine when the Google Maps image was captured.

Answer: Two of them were planted after the Google Maps picture was taken.

---

d. From your sleuthing, does it seem possible for a tree to have 15 or 16 trunks?

Answer: Yes!

---

e. How many of the trees have no trunks? Explore some of these trees to decide whether or not they really have 0 trunks. If they actually do have trunks, then "0" should actually be replaced with what?

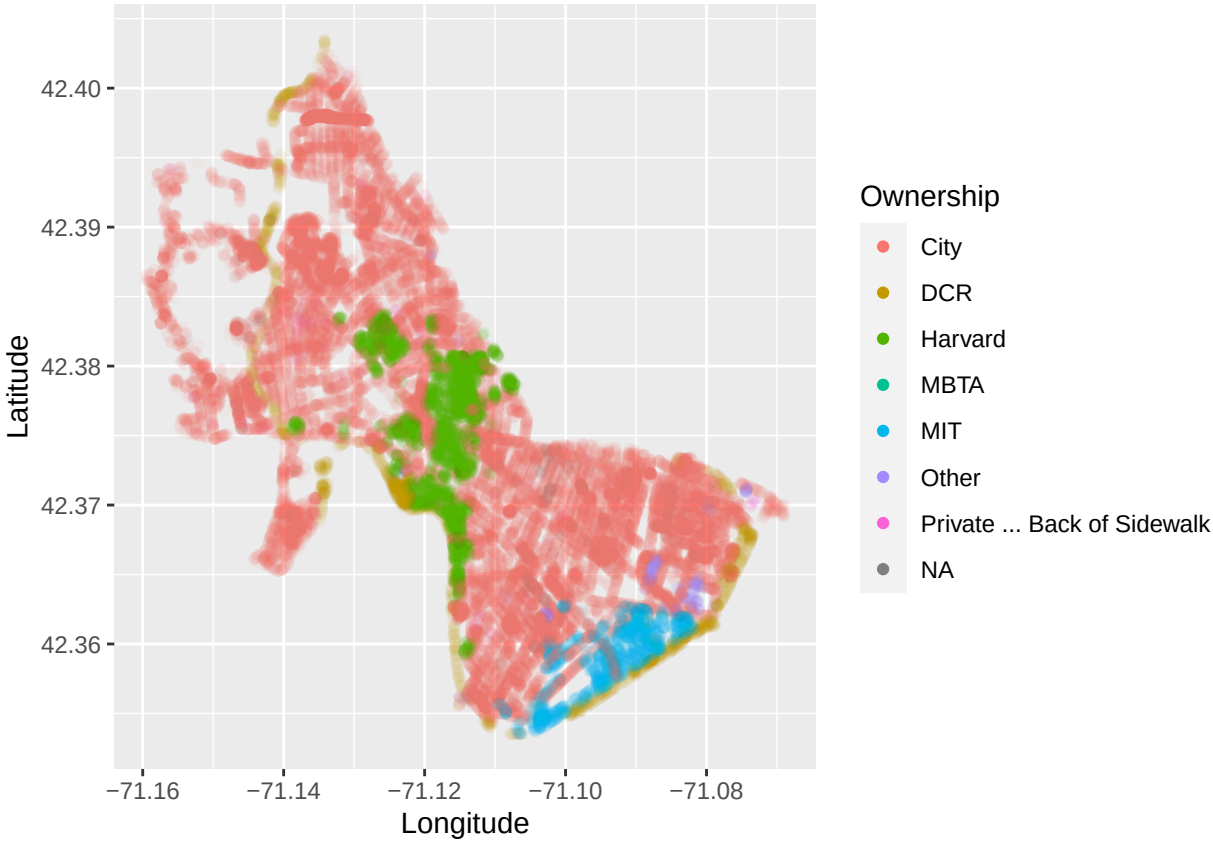
```
# Code chunk is optional
no_trunk <- filter(camTrees, Trunks == 0,
                   Diameter > 0,
                   SiteType == "Tree")
```

Answer: It looks like these trees do have trunks. It is likely that 0 should actually be coded as NA.

Problem 3 Let's explore the Cambridge trees geographically!

a. Run the following code and describe what you see in the resulting graph. Explain the shape of the data and any patterns you observe.

```
ggplot(data = camTrees, mapping = aes(x = Longitude,
                                      y = Latitude,
                                      color = Ownership)) +
  geom_point(alpha = 0.05) +
  guides(colour = guide_legend(override.aes = list(alpha = 1)))
```



Answer: Lots of potential things to observe. Here are a few potential observations:

1. The data are shaped the same as the city of Cambridge.

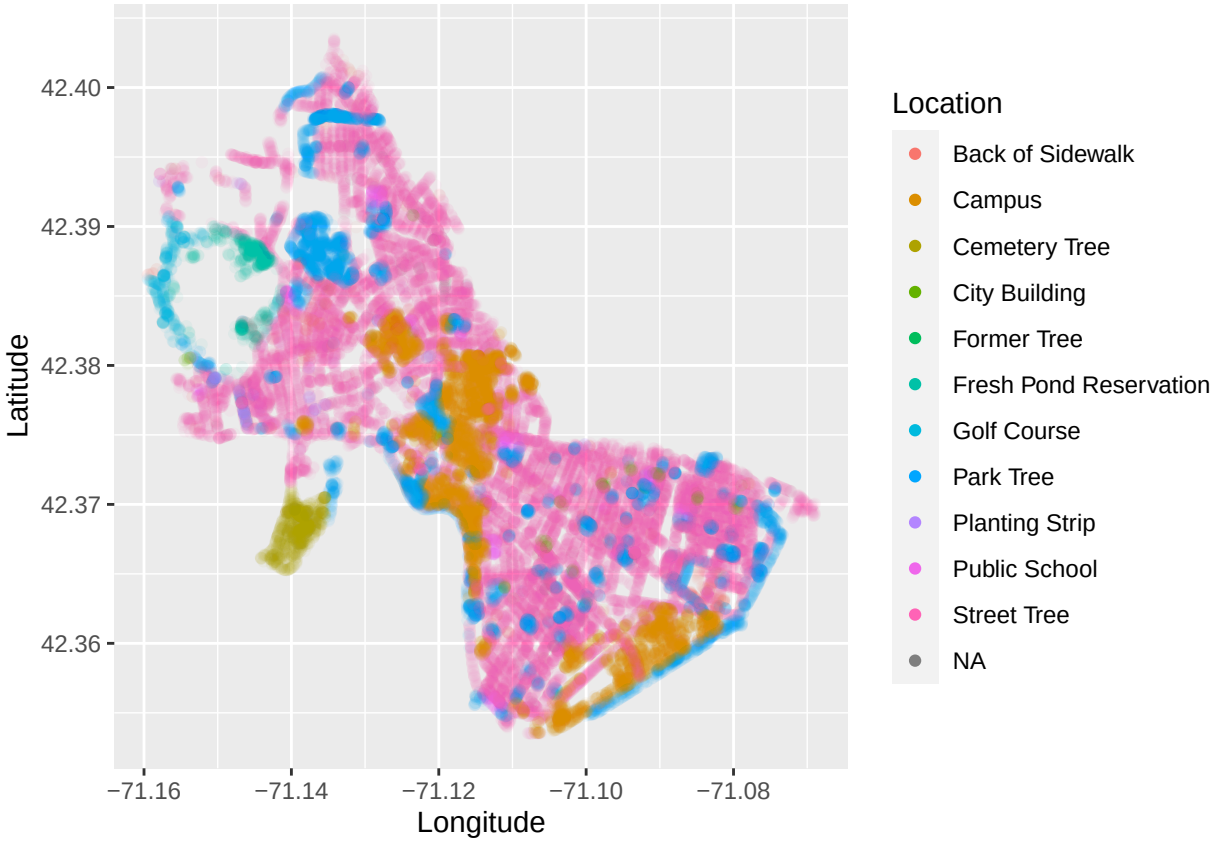
2. There are clumps where Harvard and MIT are located.

3. The DCR ownership maybe corresponds to waterways? I think this is state ownership instead of city ownership.
4. Most of the land is owned by the city.
5. There is a hole where Fresh Pond would be.

---

b. Pick one of the other variables in the `camTrees` dataset. Reproduce the scatterplot from (a) but now instead of coloring by `Ownership`, color by your new variable. Describe any interesting trends you observe.

```
ggplot(data = camTrees, mapping = aes(x = Longitude,
                                      y = Latitude,
                                      color = Location)) +
  geom_point(alpha = 0.05) +
  guides(colour = guide_legend(override.aes = list(alpha = 1)))
```



Answer: Answers will vary.

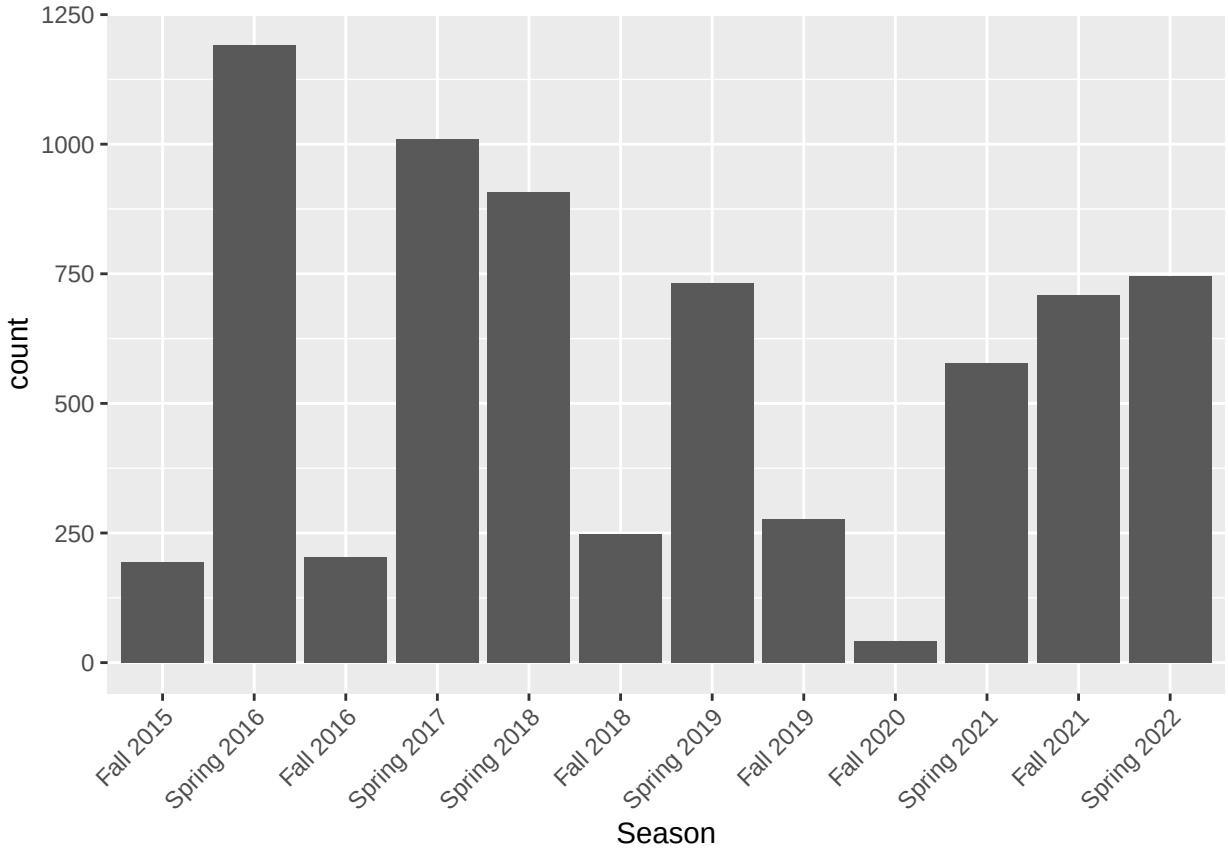
Problem 4 Let's now focus on the Boston trees and explore when the data were collected and if the data collection has changed at all over time.

a. Run the following code and answer the following questions:

- Was it more common to collect data in the fall or spring?

- Are there any missed fall or spring seasons between 2016 and 2022?

```
ggplot(data = bosTrees,
       mapping = aes(x = Season)) +
  geom_bar() +
  theme(axis.text.x = element_text(angle = 45, vjust = 1, hjust = 1))
```

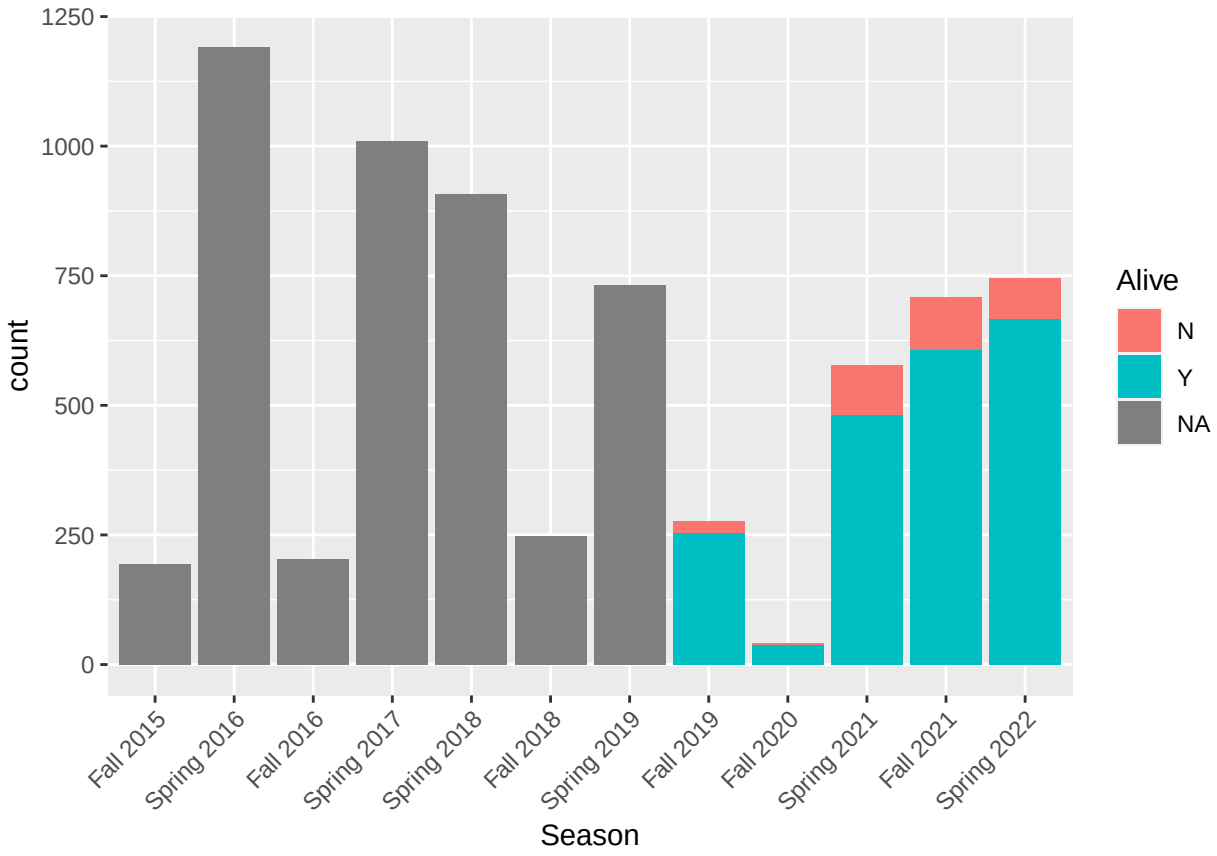


Answer: Typically data collection appears to be more common in the spring. They didn't collect data in Fall of 2017 and Spring of 2020. Also very little data were collected in Fall of 2020. It makes sense that little data were collected in 2020 because of COVID.

---

b. Recreate the plot above by this time also map `Alive` to the `fill`. What does this plot tell you about the data?

```
ggplot(data = bosTrees,
       mapping = aes(x = Season, fill = Alive)) +
  geom_bar() +
  theme(axis.text.x = element_text(angle = 45, vjust = 1, hjust = 1))
```

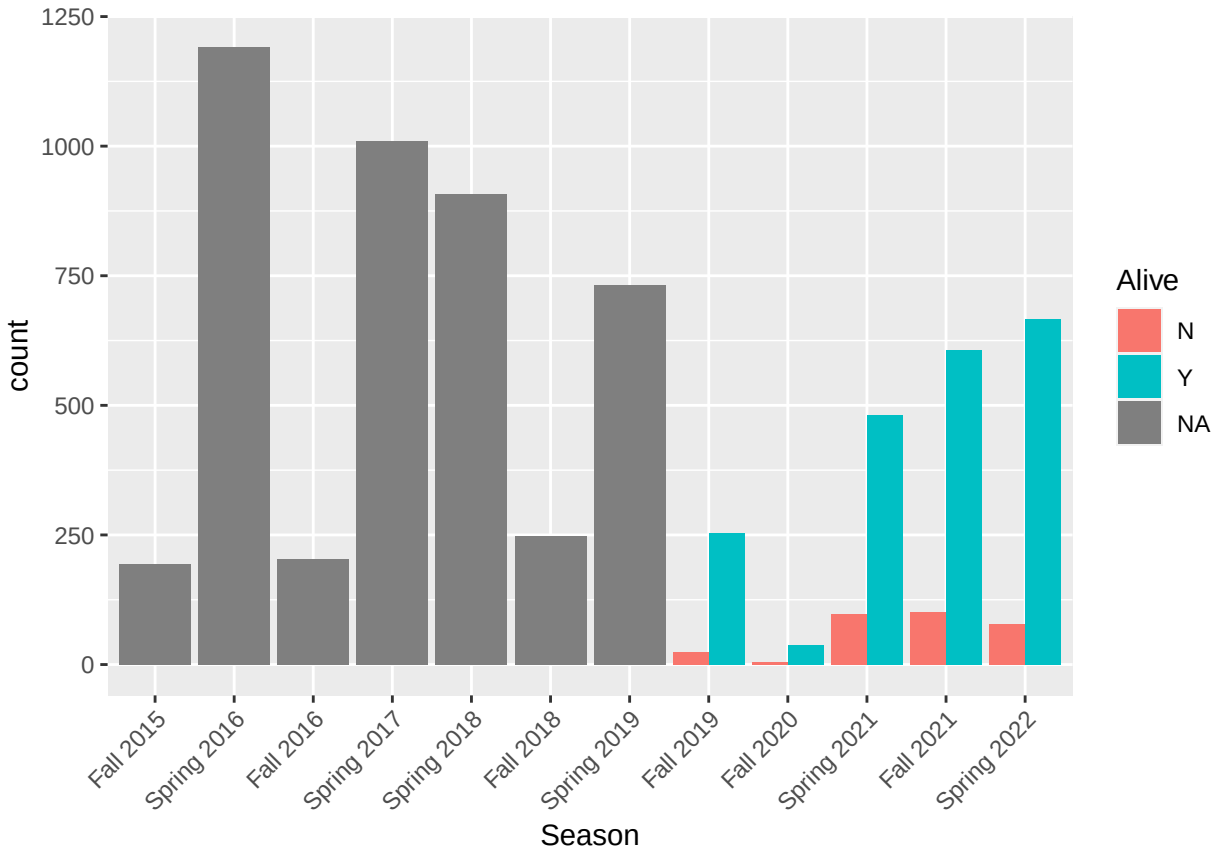


Answer: They didn't start collecting whether or not the tree was alive until Fall of 2019!

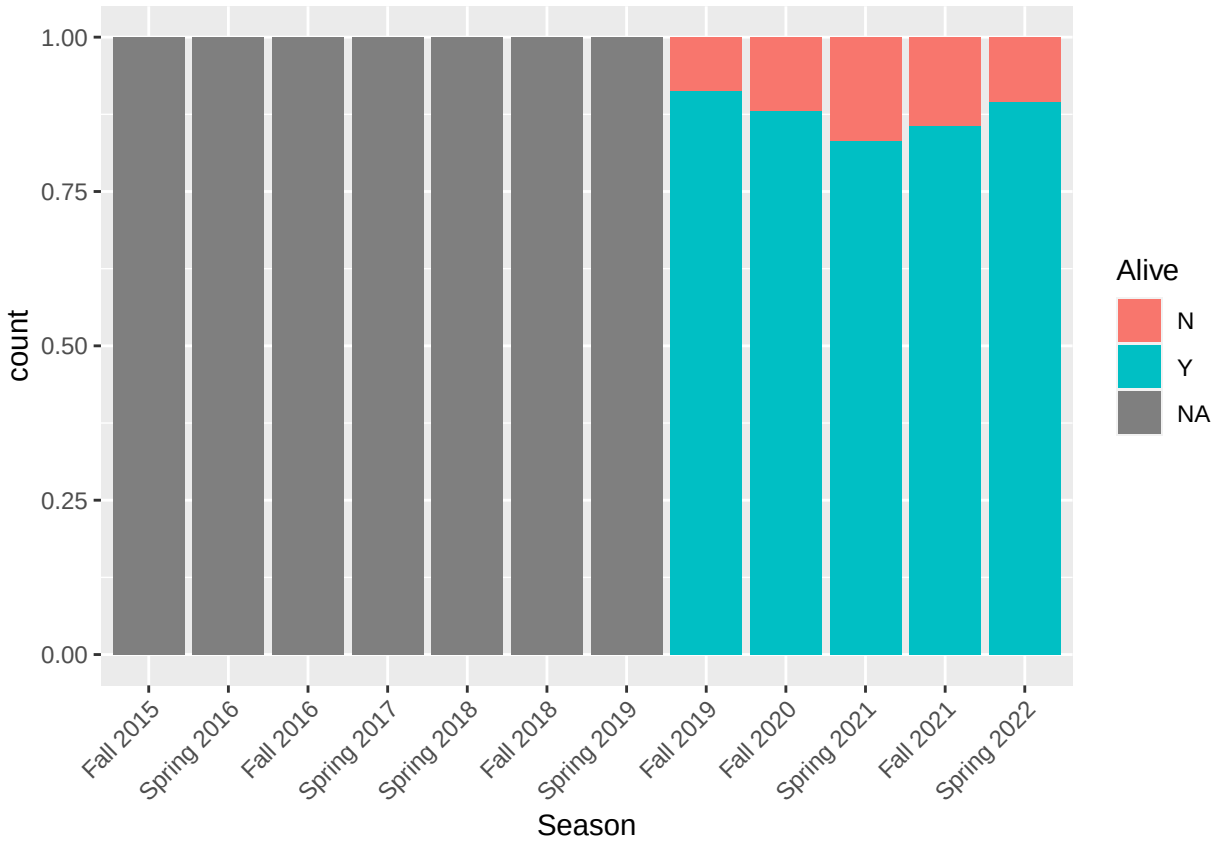
---

c. Let's enhance our plot from (b) in two different ways. First recreate the plot from (b) but add `position = "dodge"` within the `geom_bar()`. Then recreate the plot but instead add `position = "fill"`. How do the plots change? In what ways do these changes add to the story?

```
ggplot(data = bosTrees,
       mapping = aes(x = Season, fill = Alive)) +
  geom_bar(position = "dodge") +
  theme(axis.text.x = element_text(angle = 45, vjust = 1, hjust = 1))
```




```
ggplot(data = bosTrees,
       mapping = aes(x = Season, fill = Alive)) +
  geom_bar(position = "fill") +
  theme(axis.text.x = element_text(angle = 45, vjust = 1, hjust = 1))
```



Answer: The dodged bar plot makes it easier to compare the number of dead trees from year to year because the start of the bar is now at 0 for all years. The fill bar plot makes it easier to see if the **percentage** of dead trees is changing over the collect periods.

Problem 5 There is still much left to explore in these datasets but now it is your turn to pick your question(s)! Here are some potential questions to explore:

- Which tree types are most common in Boston?
- What trees are closest to the Department of Biostatistics?
- What is the distribution of the tree diameters in Cambridge? Are there any unusual values?

Feel free to ask any of us for help as you explore these data further!