

IRC_Data_Analysis_Training_2

Rachel Kenny

02/01/2021

Load packages

```
# Load packages
library(tidyverse)

## Warning: package 'tidyverse' was built under R version 4.0.2
## -- Attaching packages ----- tidyverse 1.3.0 --
## v ggplot2 3.3.2      v purrr   0.3.4
## v tibble  3.0.3      v dplyr  1.0.1
## v tidyr   1.1.1      v stringr 1.4.0
## v readr   1.3.1      v forcats 0.5.0
## Warning: package 'ggplot2' was built under R version 4.0.2
## Warning: package 'tibble' was built under R version 4.0.2
## Warning: package 'tidyr' was built under R version 4.0.2
## Warning: package 'dplyr' was built under R version 4.0.2
## -- Conflicts ----- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
library(here)

## here() starts at /Users/rachelkenny/Documents/IRC/R Code/IRC_Data_Analysis_Training
library(janitor)

##
## Attaching package: 'janitor'
## The following objects are masked from 'package:stats':
##
##   chisq.test, fisher.test
library(readxl)
```

Load data

```
# Load oak data
oak_data_raw <- read_csv(here("data", "Weir_Oak_Restoration_Data_winter19.csv"))
```

```
## Parsed with column specification:
## cols(
##   `Short ID` = col_character(),
##   Survival = col_logical(),
##   Quantity = col_double(),
##   `Height (cm)` = col_double(),
##   `Open Closed` = col_character(),
##   `Location UML` = col_character(),
##   `Water Yes No` = col_character(),
##   `Sampling Group` = col_character()
## )

oak_data <- clean_names(oak_data_raw)

# Load agua chinon veg data
ac_data_raw <- read_excel(here("data", "OCWR_AC_2019_Data.xlsx"))
ac_data <- clean_names(ac_data_raw)

hp_raw <- read_csv(here("data", "harry_potter_aggression_full.csv"))

## Parsed with column specification:
## cols(
##   Name = col_character(),
##   abb = col_character(),
##   book = col_character(),
##   aggressions = col_double(),
##   paragraphs = col_double(),
##   mentions = col_double(),
##   agg.per.mention = col_double(),
##   mentions.per.p = col_double(),
##   agg.weighted = col_double()
## )

hp_data <- clean_names(hp_raw)
```

Data wrangling

```
# dim() gives you dimensions
dim(oak_data)

## [1] 1202    8

dim(ac_data)

## [1] 1732   10

dim(hp_data)

## [1] 263    9

# head() shows first 6 lines of data
head(oak_data)

## # A tibble: 6 x 8
##   short_id survival quantity height_cm open_closed location_uml water_yes_no
##   <chr>    <lgl>      <dbl>    <dbl> <chr>      <chr>      <chr>
```

```
## 1 L_C_01_4 FALSE      0      0 Closed      Lower      No
## 2 L_C_02_1 FALSE      0      0 Closed      Lower      Yes
## 3 L_C_02_2 FALSE      0      0 Closed      Lower      Yes
## 4 L_C_02_3 FALSE      0      0 Closed      Lower      Yes
## 5 L_C_02_4 FALSE      0      0 Closed      Lower      No
## 6 L_C_03_3 FALSE      0      0 Closed      Lower      No
## # ... with 1 more variable: sampling_group <chr>
```

```
head(ac_data)
```

```
## # A tibble: 6 x 10
##   desired_habitat polygon_id transect pin_number data_type species_code
##   <chr>           <chr>       <dbl>    <dbl> <chr>      <chr>
## 1 Weedy Control  Weedy Con~      1        NA A.Belt    SOLSPP
## 2 Weedy Control  Weedy Con~      1        NA A.Belt    AMBACA
## 3 Weedy Control  Weedy Con~      1        NA A.Belt    AMBPSI
## 4 Weedy Control  Weedy Con~      1        NA A.Belt    ERIFAS
## 5 Weedy Control  Weedy Con~      1        NA A.Belt    ISOMEN
## 6 Weedy Control  Weedy Con~      1        NA A.Belt    LACSER
## # ... with 4 more variables: scientific_name <chr>, native_non_native <chr>,
## #   functional_group <chr>, layer <chr>
```

```
head(hp_data)
```

```
## # A tibble: 6 x 9
##   name  abb  book  aggressions paragraphs mentions agg_per_mention
##   <chr> <chr> <chr>      <dbl>      <dbl>    <dbl>      <dbl>
## 1 Harry harr The ~      3        3015    1074      0.00279
## 2 Herm~ herm The ~      3        3015     249      0.0120
## 3 Vold~ vold The ~      0        3015      32        0
## 4 Ron ~ ronw The ~      2        3015     513      0.00390
## 5 Seve~ snap The ~      0        3015     148        0
## 6 Albu~ albu The ~      0        3015     143        0
## # ... with 2 more variables: mentions_per_p <dbl>, agg_weighted <dbl>
```

```
# tail() shows last 6 lines of data
```

```
tail(oak_data)
```

```
## # A tibble: 6 x 8
##   short_id survival quantity height_cm open_closed location_uml water_yes_no
##   <chr>    <lgl>      <dbl>    <dbl> <chr>      <chr>      <chr>
## 1 M_0_36_3 TRUE        2      37   Open      Middle      Yes
## 2 M_0_33_3 TRUE        2      40   Open      Middle      Yes
## 3 M_0_37_3 TRUE        1      40   Open      Middle      No
## 4 M_0_30_4 TRUE        2      46   Open      Middle      Yes
## 5 M_0_09_1 TRUE        1     46.5 Open      Middle      Yes
## 6 U_0_49_3 TRUE        1      63   Open      Upper       Yes
## # ... with 1 more variable: sampling_group <chr>
```

```
tail(ac_data)
```

```
## # A tibble: 6 x 10
##   desired_habitat polygon_id transect pin_number data_type species_code
##   <chr>           <chr>       <dbl>    <dbl> <chr>      <chr>
## 1 Elderberry Scr~ OW-M2      1        NA A.Belt    SOLUMB
## 2 Elderberry Scr~ OW-M2      1        NA A.Belt    AMBACA
## 3 Elderberry Scr~ OW-M2      1        NA A.Belt    MALSAX
```

```
## 4 Elderberry Scr~ OW-M2          1          NA A.Belt    SALMEL
## 5 Elderberry Scr~ OW-M2          1          NA A.Belt    MEDPOL
## 6 Elderberry Scr~ OW-M2          1          NA A.Belt    LACSER
## # ... with 4 more variables: scientific_name <chr>, native_non_native <chr>,
## #   functional_group <chr>, layer <chr>
```

```
tail(hp_data)
```

```
## # A tibble: 6 x 9
##   name abb book aggressions paragraphs mentions agg_per_mention
##   <chr> <chr> <chr>          <dbl>         <dbl>         <dbl>         <dbl>
## 1 Seam~ seam The ~           1          6861           11          0.0909
## 2 Yaxl~ yaxl The ~           2          6861           52          0.0385
## 3 Fleu~ fleu The ~           0          6861          104           0
## 4 Greg~ goyl The ~           1          6861           26          0.0385
## 5 Marv~ marv The ~           0          6861           11           0
## 6 Mund~ mund The ~           0          6861           42           0
## # ... with 2 more variables: mentions_per_p <dbl>, agg_weighted <dbl>
```

```
# summary() shows the class and statistical summary of data
```

```
summary(oak_data)
```

```
##   short_id      survival      quantity      height_cm
## Length:1202    Mode :logical  Min.   : 0.000    Min.   : 0.000
## Class :character FALSE:525    1st Qu.: 0.000    1st Qu.: 0.000
## Mode  :character TRUE :677    Median : 1.000    Median : 9.000
##                                     Mean  : 0.891    Mean  : 8.627
##                                     3rd Qu.: 1.000    3rd Qu.:15.000
##                                     Max.   :10.000    Max.   :63.000
## open_closed    location_uhl    water_yes_no    sampling_group
## Length:1202    Length:1202    Length:1202    Length:1202
## Class :character Class :character Class :character Class :character
## Mode  :character Mode  :character Mode  :character Mode  :character
##
##
##
```

```
summary(ac_data)
```

```
##   desired_habitat    polygon_id      transect      pin_number
## Length:1732        Length:1732    Min.   :1.000    Min.   : 1.00
## Class :character    Class :character 1st Qu.:1.000    1st Qu.:13.00
## Mode  :character    Mode  :character Median :1.000    Median :25.00
##                                     Mean  :1.293    Mean  :25.42
##                                     3rd Qu.:1.000    3rd Qu.:38.00
##                                     Max.   :3.000    Max.   :50.00
##                                     NA's   :200
##   data_type      species_code      scientific_name      native_non_native
## Length:1732      Length:1732      Length:1732      Length:1732
## Class :character  Class :character  Class :character  Class :character
## Mode  :character  Mode  :character  Mode  :character  Mode  :character
##
##
##
## functional_group      layer
```

```
## Length:1732      Length:1732
## Class :character  Class :character
## Mode :character  Mode :character
##
##
##
##
```

```
summary(hp_data)
```

```
##      name          abb          book      aggressions
## Length:263      Length:263      Length:263      Min.   : 0.000
## Class :character  Class :character  Class :character  1st Qu.: 0.000
## Mode  :character  Mode  :character  Mode  :character  Median : 1.000
##                                           Mean   : 2.681
##                                           3rd Qu.: 2.500
##                                           Max.   :48.000
## paragraphs      mentions      agg_per_mention  mentions_per_p
## Min.   :3015     Min.   : 11.0     Min.   :0.000000     Min.   :0.001186
## 1st Qu.:4092     1st Qu.: 26.0     1st Qu.:0.000000     1st Qu.:0.004423
## Median :6803     Median : 52.0     Median :0.002436     Median :0.008841
## Mean   :6211     Mean   :187.7     Mean   :0.033394     Mean   :0.032342
## 3rd Qu.:6861     3rd Qu.:143.0     3rd Qu.:0.031454     3rd Qu.:0.028575
## Max.   :9275     Max.   :3509.0     Max.   :0.636364     Max.   :0.406158
## agg_weighted
## Min.   : 0.00
## 1st Qu.: 0.00
## Median : 9.45
## Mean   :223.09
## 3rd Qu.:194.60
## Max.   :5902.27
```

```
View(oak_data)
```

```
View(ac_data)
```

```
View(hp_data)
```

```
# To change the class use as.character or as.numeric
oak_data$survival <- as.character(oak_data$survival)
oak_data$survival <- as.numeric(oak_data$survival)
```

```
## Warning: NAs introduced by coercion
```

```
# To check what class of data it is, use class()
class(ac_data$transect)
```

```
## [1] "numeric"
```

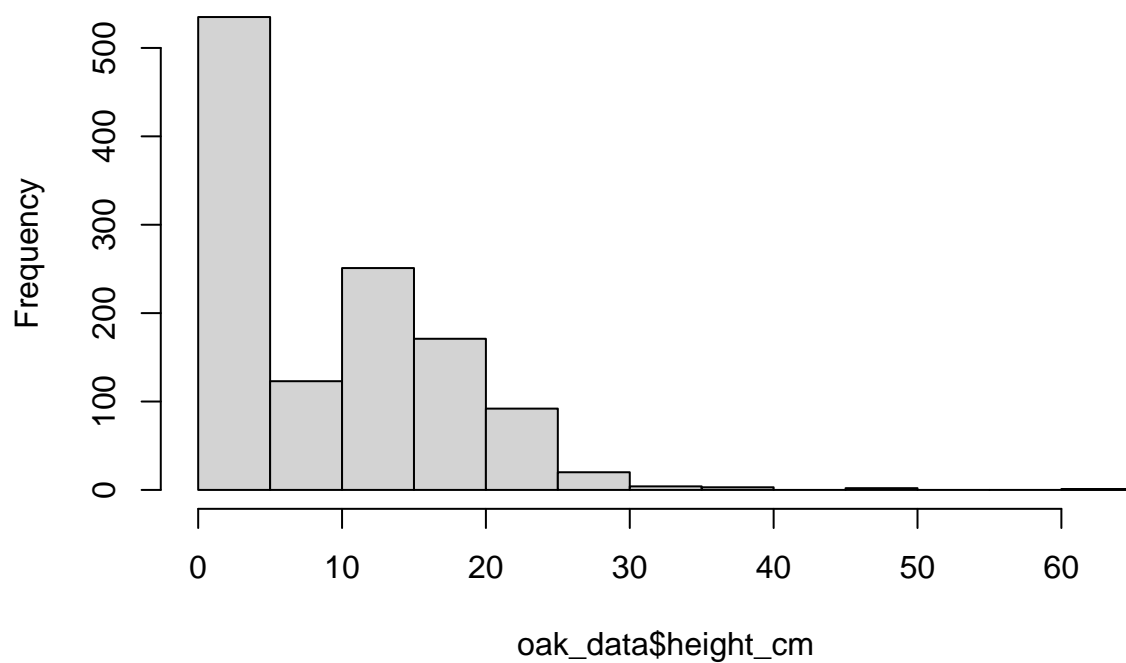
```
ac_data$transect <- as.character(ac_data$transect)
class(ac_data$transect)
```

```
## [1] "character"
```

Data visualization 1

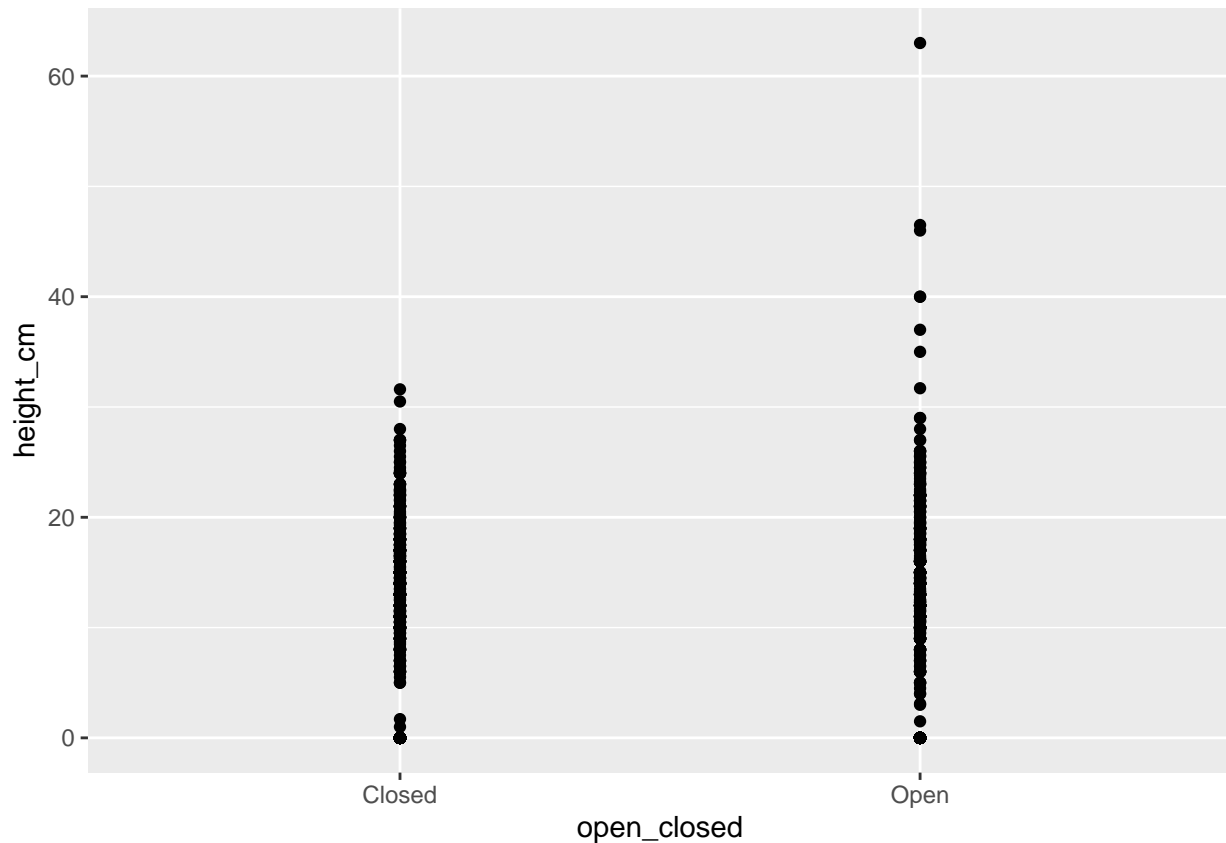
```
# histogram of oak seedling height
hist(oak_data$height_cm)
```

Histogram of oak_data\$height_cm



```
# basic ggplot scatter plot example
plot1 <- ggplot(oak_data, aes(x=open_closed, y=height_cm)) +
  geom_point()

plot1
```

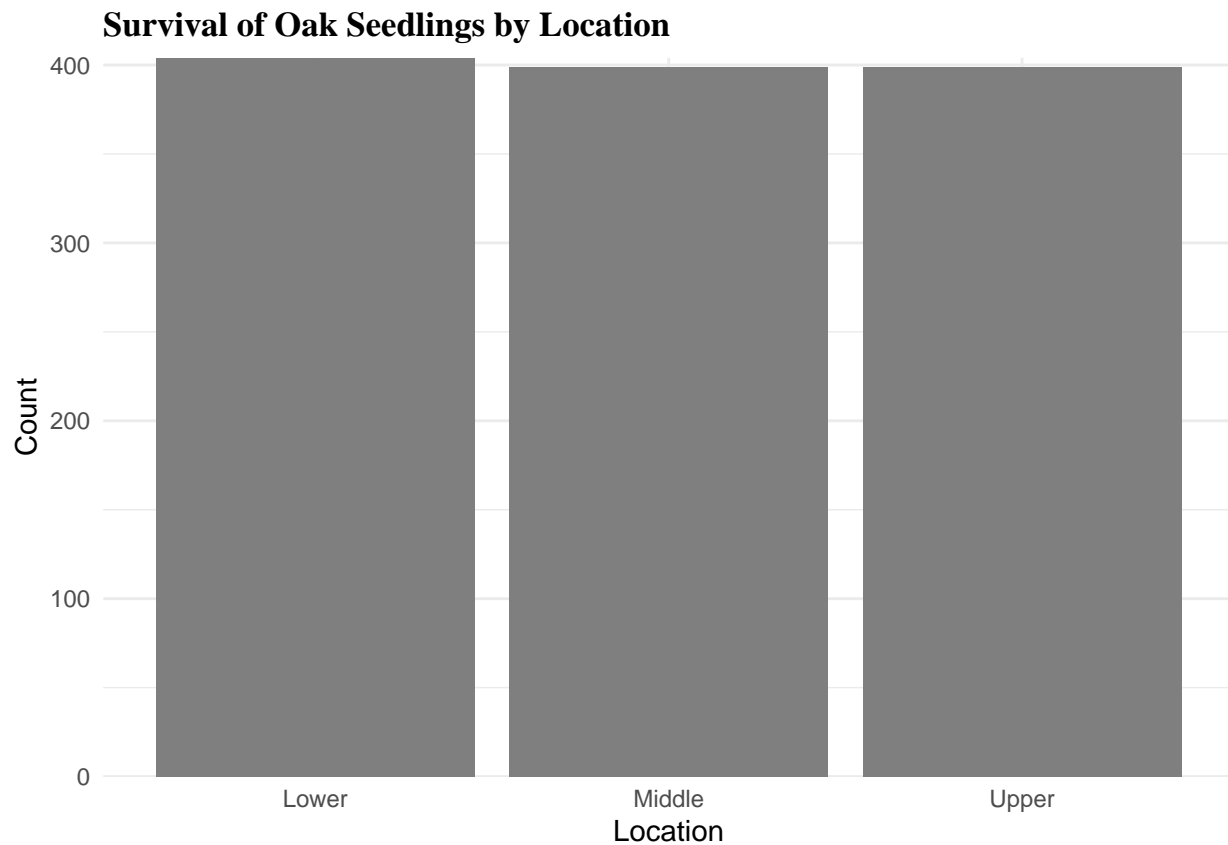


Data visualization 2

```
# Bar graphs - Example 1 using oak data

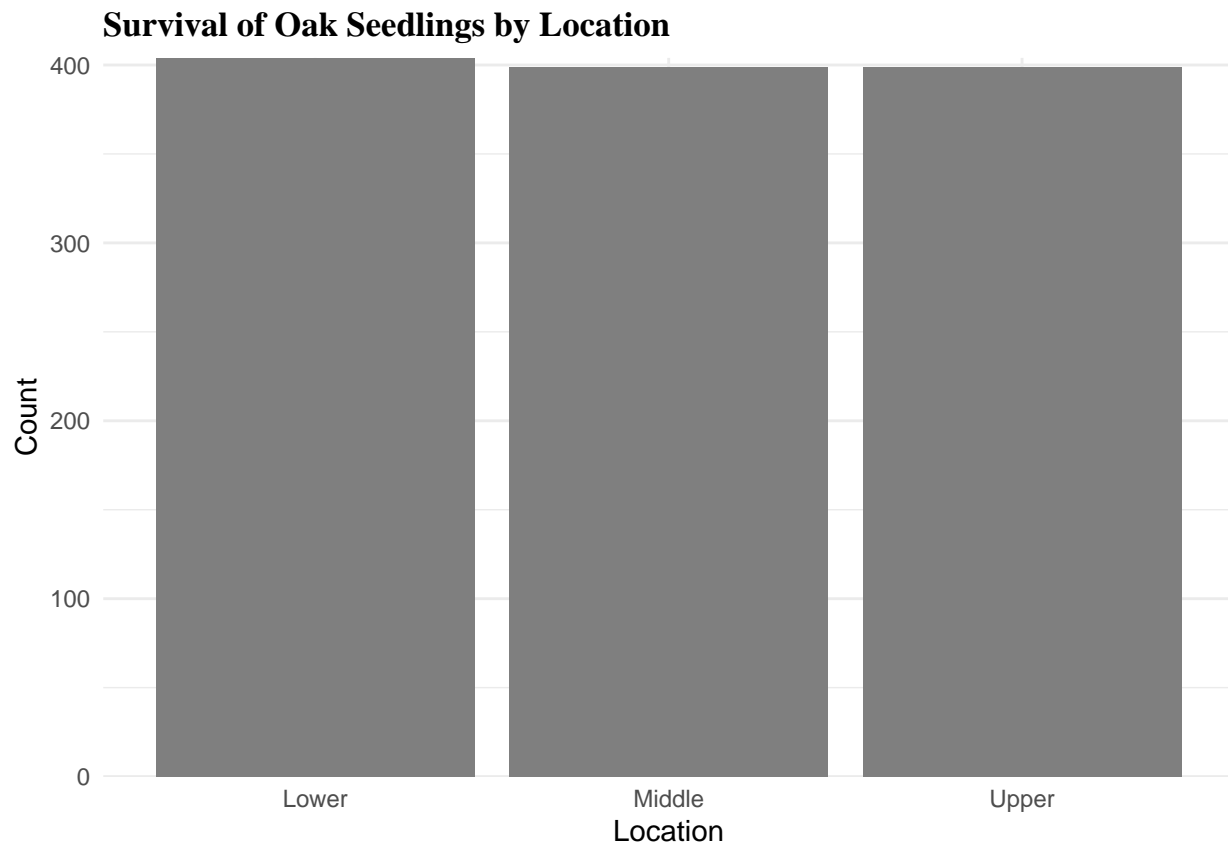
# Side by side plots
plot2 <- ggplot(oak_data, aes(location_u1, fill=survival)) +
  geom_bar(position="dodge") + #this is where the code changes each time.
  xlab("Location") +
  ylab("Count") +
  ggtitle("Survival of Oak Seedlings by Location") +
  scale_x_discrete(expand=c(0.3,0)) +
  scale_y_continuous(expand=c(0,0)) +
  theme_minimal() +
  theme(plot.title=element_text(face="bold", family = "serif"))

plot2
```



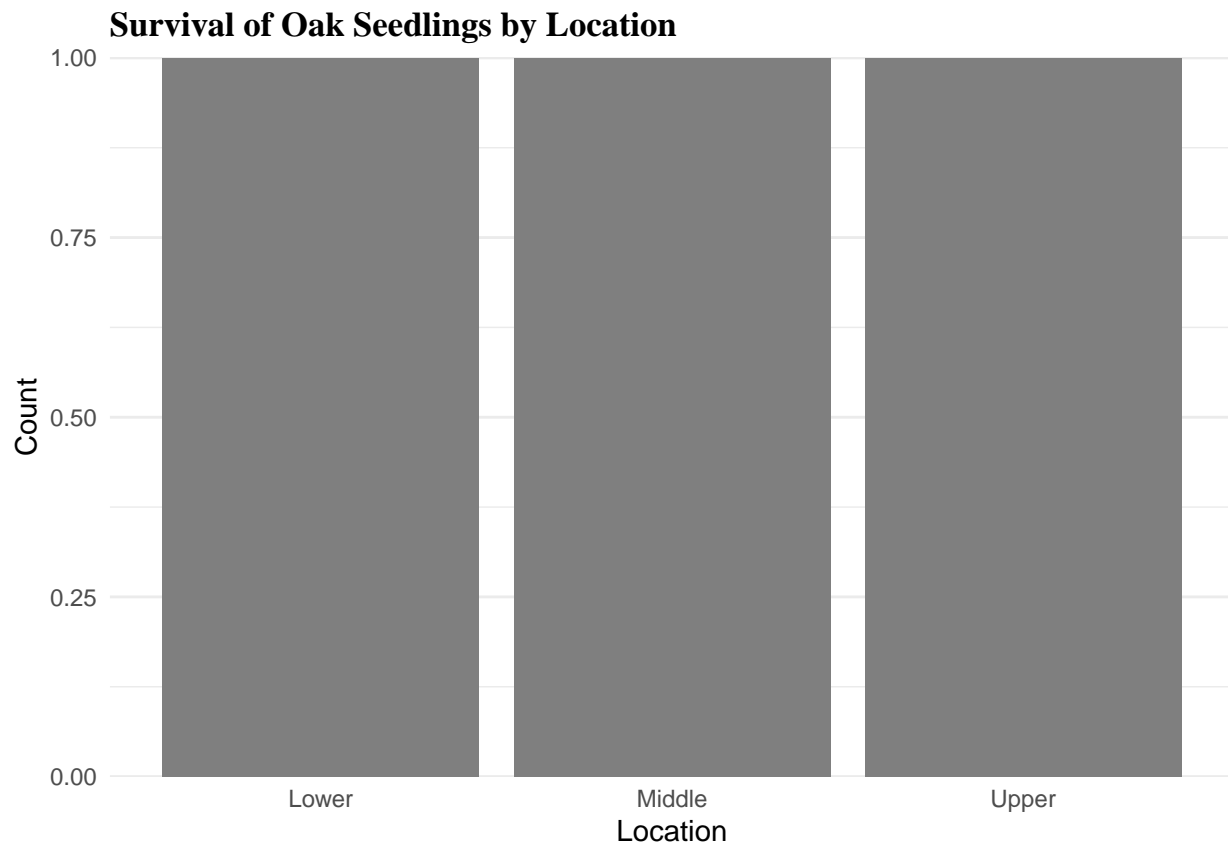
```
# Stacked plots
plot3 <- ggplot(oak_data, aes(location_u1, fill=survival)) +
  geom_bar(position="stack") + #this is where the code changes each time.
  xlab("Location") +
  ylab("Count") +
  ggtitle("Survival of Oak Seedlings by Location") +
  scale_x_discrete(expand=c(0.3,0)) +
  scale_y_continuous(expand=c(0,0)) +
  theme_minimal() +
  theme(plot.title=element_text(face="bold", family = "serif"))

plot3
```

```
# Normalized stacked plots
plot4 <- ggplot(oak_data, aes(location_u1, fill=survival)) +
  geom_bar(position="fill") + #this is where the code changes each time.
  xlab("Location") +
  ylab("Count") +
  ggtitle("Survival of Oak Seedlings by Location") +
  scale_x_discrete(expand=c(0.3,0)) +
  scale_y_continuous(expand=c(0,0)) +
  theme_minimal() +
  theme(plot.title=element_text(face="bold", family = "serif"))
```

plot4

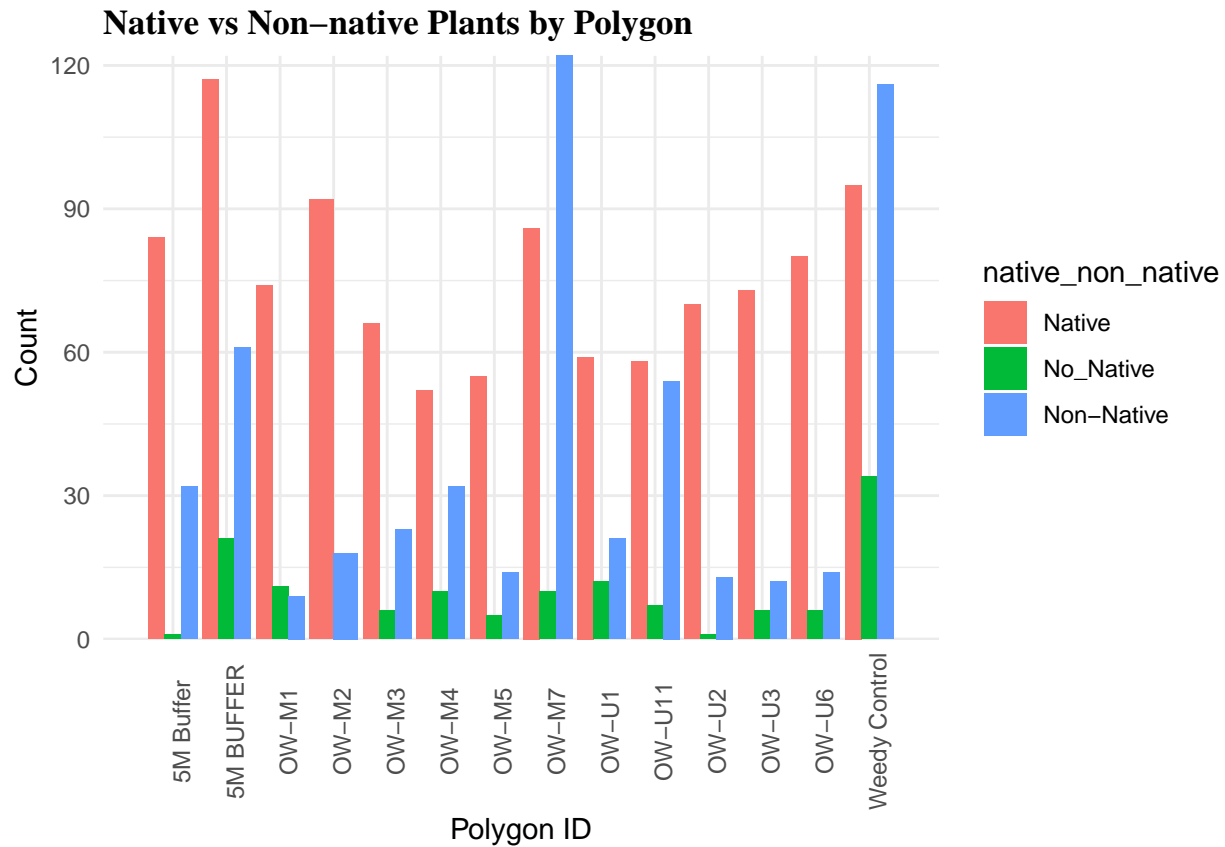


Bar graphs - Example 2 using aqua chinon veg data

Side by side plots

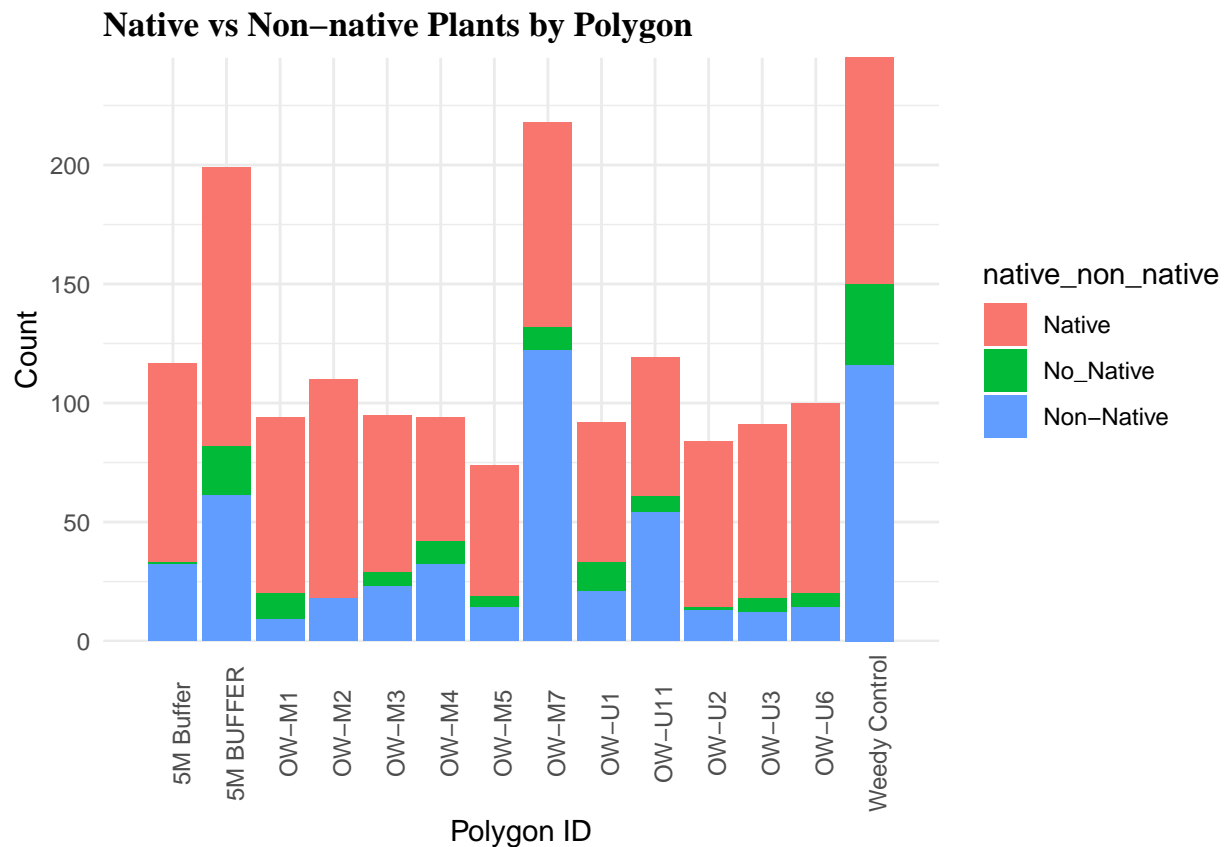
```
plot5 <- ggplot(ac_data, aes(polygon_id, fill=native_non_native)) +
  geom_bar(position="dodge") + #this is where the code changes each time.
  xlab("Polygon ID") +
  ylab("Count") +
  ggtitle("Native vs Non-native Plants by Polygon") +
  scale_x_discrete(expand=c(0.1,0)) +
  scale_y_continuous(expand=c(0,0)) +
  theme_minimal() +
  theme(plot.title=element_text(face="bold", family = "serif"), axis.text.x = element_text(angle = 90))
```

plot5



```
# Stacked plots
plot6 <- ggplot(ac_data, aes(polygon_id, fill=native_non_native)) +
  geom_bar(position="stack") + #this is where the code changes each time.
  xlab("Polygon ID") +
  ylab("Count") +
  ggtitle("Native vs Non-native Plants by Polygon") +
  scale_x_discrete(expand=c(0.1,0)) +
  scale_y_continuous(expand=c(0,0)) +
  theme_minimal() +
  theme(plot.title=element_text(face="bold", family = "serif"), axis.text.x = element_text(angle = 90))

plot6
```



```
# Normalized stacked plots
plot7 <- ggplot(ac_data, aes(polygon_id, fill=native_non_native)) +
  geom_bar(position="fill") + #this is where the code changes each time.
  xlab("Polygon ID") +
  ylab("Count") +
  ggtitle("Native vs Non-native Plants by Polygon") +
  scale_x_discrete(expand=c(0.1,0)) +
  scale_y_continuous(expand=c(0,0)) +
  theme_minimal() +
  theme(plot.title=element_text(face="bold", family = "serif"), axis.text.x = element_text(angle = 90))

plot7
```

