Project Proposal

Ava Cruzan

1 My blog link

My blog is available here

```
## here() starts at /home/alcruzan/DASC-blog

## -- Attaching packages ----------------------------------------------------------------

## v tibble  2.1.3      v dplyr   0.8.3
## v tidyr   1.0.0      v stringr 1.4.0
## v readr   1.3.1      v forcats 0.4.0
## v purrr   0.3.3

## -- Conflicts -------------------------------------------------------------------------
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()

## Observations: 500
## Variables: 24
## $ binomial_name     <fct> Abutilon pitcairnense, Acaena exigua, Acalypha dikuluwensis, Acalypha rubri
## $ country           <fct> "Pitcairn", "United States", "Congo", "Saint Helena, Ascension and Tristan
## $ continent         <fct> Oceania, North America, Africa, Africa, Oceania, Asia, North America, Nortl
## $ group             <fct> Flowering Plant, Flowering Plant, Flowering Plant, Flowering Plant, Floweri
## $ year_last_seen    <fct> 2000-2020, 1980-1999, 1940-1959, Before 1900, 1920-1939, 1920-1939, 1960-19
## $ threat_AA         <int> 0, 0, 0, 1, 1, 0, 0, 0, 0, 1, 0, 0, 1, 1, 0, 0, 0, 1, 1, 1, 1, 1, 0, 0, 0,
## $ threat_BRU        <int> 0, 0, 0, 1, 0, 0, 0, 0, 0, 1, 1, 1, 1, 0, 1, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0,
## $ threat_RCD        <int> 0, 0, 0, 0, 1, 1, 1, 0, 0, 1, 0, 0, 0, 1, 0, 0, 0, 0, 1, 0, 1, 0, 1, 0, 0,
## $ threat_ISGD       <int> 1, 1, 0, 1, 1, 0, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 1,
## $ threat_EPM        <int> 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
## $ threat_CC         <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0,
## $ threat_HID        <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
## $ threat_P          <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
## $ threat_TS         <int> 0, 0, 0, 0, 1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
## $ threat_NSM        <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 1, 0, 0,
## $ threat_GE         <int> 1, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1,
## $ threat_NA         <int> 0, 0, 0, 1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0,
## $ action_LWP        <int> 1, 0, 0, 0, 0, 0, 0, 1, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1,
## $ action_SM         <int> 1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
## $ action_LP         <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
## $ action_RM         <int> 1, 1, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1,
## $ action_EA         <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
## $ action_NA         <int> 0, 0, 1, 1, 1, 0, 1, 0, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0,
## $ red_list_category <fct> Extinct in the Wild, Extinct, Extinct, Extinct, Extinct, Extinct in the Wil
```

```
## Observations: 3,000
## Variables: 8
## $ binomial_name    <fct> Abutilon pitcairnense, Abutilon pitcairnense, Abutilon pitcairnense, Abutil
## $ country          <fct> "Pitcairn", "Pitcairn", "Pitcairn", "Pitcairn", "Pitcairn", "Pitcairn", "Ur
## $ continent         <fct> Oceania, Oceania, Oceania, Oceania, Oceania, Oceania, North America, North
## $ group            <fct> Flowering Plant, Flowering Plant, Flowering Plant, Flowering Plant, Floweri
## $ year_last_seen   <fct> 2000-2020, 2000-2020, 2000-2020, 2000-2020, 2000-2020, 2000-2020, 1980-1999
## $ red_list_category <fct> Extinct in the Wild, Extinct in the Wild, Extinct in the Wild, Extinct in t
## $ action_type      <fct> Land & Water Protection, Species Management, Law & Policy, Research & Monit
## $ action_taken     <int> 1, 1, 0, 1, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 1, 0,
```

2 Extinct Plants

For this project, I would like to examine the plants that are extinct. I am using the dataset plants.csv from the Tidy Tuesday website. The data consists of 500 observations and 24 variables. The binomial_name variable shows the species name for the plant. The country variable shows what country the plant is native to. The continent variable shows the continent the plant is from. The variable group is the plant's taxonomic group. The year_last_seen variable is the last time the plant was seen. The threat_AA, threat_BRU, threat_RCD, threat_ISGD, threat_EPM, threat_CC, threat_HID, threat_P, threat_TS, threat_NSM, threat_GE, and threat_NA variables are for telling if the plant is threatened by agriculture and aquaculture, biological resource use, commercial development, invasive species, energy production and mining, climate change, human intrusions, pollution, transportation corridor, natural system modifications, geological events, or the threat is unknown respectively. The variables action_LWP, action_SM, action_LP, action_RM, action_EA, and action_NA are for showing if there is current action in regards to land and water protection, species management, law and policy, research and monitoring, education and awareness, or if there is no current action known. The variable red_list_category shows the IUCN Red list category. Initial exploration indicates that a decent portion of the plants haven't been seen since before 1900, and it appears that many of the plants that have gone extinct are from Asia, Africa, or South America.

Additional data to understand these plants more is in actions.csv dataset also found on Tidy Tuesday. The actions.csv dataset has 3000 observations and 8 variables. These datasets contain much of the same information. The actions.csv has the same variables binomial_name, country, continent, group, year_last_seen, and red_list_category like plants.csv. However, actions.csv also has variables action_type to spell out the type of action and action_taken to tell if the action has been taken or not. With this data, it will be easier to tell how much work is getting done to combat extinction.

- Question 1: First, is there a country or continent that has a majority of the plant extinctions. If so, which country or continent. To test this, I plan to calculate how many plants have gone extinct per continent and generate data visualizations in order to easily see if there is a trend.
- Question 2: Second, is there a time period in which there were significantly more extinctions. If so, is there a trend to predict future plant extinctions. To test this, I plan to generate a line chart using ggplot to see if there is a trend. I will use other forms of data groupings to see if there are other factors such as different threat types and locations that would affect the plot.
- Question 3: Third, I would like to see how much action is being taken to combat the extinction of plants. I plan to join the two datasets in order to get more information. I would like to use data visualization to see where the most action and least action is being taken, and I would also like to see which countries and continents are taking the most action.

3 San Francisco Trees

```
## Observations: 192,987
## Variables: 12
## $ tree_id      <int> 53719, 30313, 30312, 30314, 30315, 30316, 48435, 30319, 30318, 30320, 30321, 30
```

```
## $ legal_status <fct> Permitted Site, Permitted Site, Permitted Site, DPW Maintained, Permitted Site,
## $ species      <fct> Tree(s) ::, Tree(s) ::, Tree(s) ::, Pittosporum undulatum :: Victorian Box, Aca
## $ address      <fct> 2963 Webster St, 501 Arkansas St, 501 Arkansas St, 501 Arkansas St, 1190 Sacram
## $ site_order   <int> 1, 3, 2, 1, 5, 6, 4, 2, 1, 3, 1, 3, 1, 2, 4, 1, 1, 1, 2, 4, 3, 1, 2, 1, 3, 1, 1
## $ site_info    <fct> Sidewalk: Curb side : Cutout, Sidewalk: Curb side : Cutout, Sidewalk: Curb side
## $ caretaker    <fct> Private, Private, Private, Private, Private, Private, Private, Private, Private
## $ date         <fct> 1955-09-19, 1955-10-20, 1955-10-20, 1955-10-20, 1955-10-24, 1955-10-24, 1955-10-
## $ dbh          <int> NA, NA, NA, 16, NA, NA, NA, NA, NA, NA, 2, NA, NA, NA, NA, NA, NA, NA, 4, NA, 2
## $ plot_size    <fct> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,
## $ latitude     <dbl> 37.79787, 37.75984, 37.75984, 37.75977, 37.79265, 37.79265, 37.79265, 37.77319,
## $ longitude    <dbl> -122.4341, -122.3981, -122.3981, -122.3981, -122.4124, -122.4124, -122.4124, -1
```

Using the sf_trees.csv dataset from Tidy Tuesday, I investigate what the trees are like in San Francisco. The data has 192,987 observations and 12 variables. The variables are the ID given to the tree, the legal status of the tree, the species of the tree, the street address, the order of the trees in the case that there are multiple trees at one location, where the tree is on the lot, the primary caretaker of the tree, the date it was planted, the diameter of the tree, the dimension of the plot, the latitude the plant is at, and the longitude the tree is planted. * Question 1: First, Where are most of the trees in San Francisco? Are there in locations with an especially high amount of trees? Are most of these trees public or private? To analyze this, I will create a spatial map using ggplot and compare that to a map of San Francisco to see locations of interest. * Question 2: Second, What is the most common tree in San Francisco? If so, What kind To assess this, I will use data manipulation and represent this data visually to show what kind of tree is the most popular tree. * Question 3: Third, Does site of the tree affect the diameter of the tree? To test this, I will create a graph that compares the age, diameter, and site of a tree to see if some trees are growing in less than ideal places.