Final Project Proposal

National Park Species

Moriah Ruggerio

Import the data:

```
most_visited_nps_species_data <- readr::read_csv('https://raw.githubusercontent.com/rfordatascience/tidy
```

```
## Warning: One or more parsing issues, call `problems()` on your data frame for details,
## e.g.:
##   dat <- vroom(...)
##   problems(dat)
```

```
## Rows: 61119 Columns: 28
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (21): ParkCode, ParkName, CategoryName, Order, Family, TaxonRecordStatus...
## dbl  (3): References, Observations, Vouchers
## lgl  (4): Synonyms, ParkAccepted, Sensitive, ExternalLinks
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
most_visited_nps_species_data_cleaned <- most_visited_nps_species_data %>%
  janitor::clean_names()
```

Determine what variables to remove to keep it between 10-20:

```
most_visited_nps_species_data_cleaned <- most_visited_nps_species_data_cleaned %>%
  filter(taxon_record_status == "Active", park_accepted == TRUE) %>%  #only includes data with active t
  #count(taxon_record_status)
  select(park_name:family, sci_name, common_names, occurrence:abundance, references:vouchers, te_status
  #count(taxon_record_status, park_accepted)

#no data in synonyms, external_links
#all data is not sensitive (remove variable)
#combine nativeness_tabs with native & same with occurrence and occurrence_tags
#remove park_tags
#remove inactive taxons (including data rows)
#only include park accepted/remove the row and record status
```

```
write_csv(most_visited_nps_species_data_cleaned, file = "data/most_visited_nps_species_data_cleaned.csv
```

```
most_visited_nps_species_data_cleaned %>%
  count(nativeness)
```

```
## # A tibble: 4 x 2
##   nativeness     n
```

```
##   <chr>      <int>
## 1 Native     24916
## 2 Non-native  3754
## 3 Unknown    25085
## 4 <NA>        1287
```

Introduction

This project uses the dataset `most_visited_nps_species_data_cleaned`. The National Park Service (NPS) keeps a list of every species found within each national park. The data used in this project is originally from the National Park Species List. It can be accessed here: https://irma.nps.gov/NPSpecies/Search/SpeciesList. Due to the massive size of this dataset, it was restricted to the top 15 most popular national parks by `f. hull` in the `tidytuesday` git repository on September 2, 2024. This reduced dataset (`most_visited_nps_species_data`) was accessed on March 15, 2025 from https://github.com/rfordatascience/tidytuesday/blob/main/data/2024/2024-10-08/readme.md. Since this dataset originally had too many for the confines of this assignment, I restricted the scope of the data further. Only included in this projects' dataset (`most_visited_nps_species_data_cleaned`) are species whose presence in the park has been approved by the NPS, have an active taxon (current scientific name), and a non-sensitive species status. Data on sensitive species is removed from all public access by the NPS to protect endangered or fragile species in national parks.

The dataset has 55042 rows and 19 variables. Each row is a specific species found in one of the parks. The variables are `park_name`, `category_name`, `order`, `family`, `sci_name`, `common_names`, `occurrence`, `occurrence_tags`, `nativeness`, `nativeness_tags`, `abundance`, `references`, `observations`, `vouchers`, `te_status`, `state_status`, `ozone_sensitive_status`, `g_rank`, `s_rank`. (See `README.md` for variable descriptions. The intent of this research project is to examine if there is trends between biodiversity and the popularity of a park. Some specific research questions to be examined are: (1) Do more popular parks have higher observations of birds (and other classes)? (2) Are there more native species seen in popular parks than invasives or introduced?

Data

```
glimpse(most_visited_nps_species_data_cleaned)
```

```
## Rows: 55,042
## Columns: 19
## $ park_name              <chr> "Acadia National Park", "Acadia National Park",~
## $ category_name          <chr> "Mammal", "Mammal", "Mammal", "Mammal", "Mammal~
## $ order                  <chr> "Artiodactyla", "Artiodactyla", "Carnivora", "C~
## $ family                 <chr> "Cervidae", "Cervidae", "Canidae", "Canidae", "~
## $ sci_name               <chr> "Alces alces", "Odocoileus virginianus", "Canis~
## $ common_names           <chr> "Moose", "Northern White-tailed Deer, Virginia ~
## $ occurrence             <chr> "Present", "Present", "Present", "Unconfirmed",~
## $ occurrence_tags        <chr> NA, NA, NA, NA, NA, NA, NA, "Historical", NA, N~
## $ nativeness             <chr> "Native", "Native", "Non-native", "Native", "Un~
## $ nativeness_tags        <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,~
## $ abundance              <chr> "Rare", "Abundant", "Common", NA, "Common", NA,~
## $ references             <dbl> 11, 20, 8, 2, 16, 1, 7, 10, 10, 5, 7, 6, 14, 1,~
## $ observations           <dbl> 1, 0, 1, 0, 0, 0, 0, 0, 0, 0, 1, 0, 1, 0, 0, 1,~
## $ vouchers               <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
## $ te_status              <chr> "50", "50", "SC", "E", NA, NA, "RT", NA, NA, NA~
## $ state_status           <chr> NA, NA, NA, NA, NA, "ME: SC", NA, NA, NA, NA, N~
## $ ozone_sensitive_status <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,~
```

```
## $ g_rank                 <chr> "G5", "G5", "G5", "G5", "G5", "G5", "G5", "G5",~
## $ s_rank                 <chr> "ME: S5", "ME: S5", "ME: S5", "ME: SH", "ME: S5~
```

```
skim(most_visited_nps_species_data_cleaned)
```

Table 1: Data summary

Name	most_visited_nps_species_...
Number of rows	55042
Number of columns	19
Column type frequency:	
character	16
numeric	3
Group variables	None

Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
park_name	0	1.00	18	35	0	15	0
category_name	0	1.00	4	21	0	16	0
order	556	0.99	5	22	0	501	0
family	642	0.99	6	24	0	2043	0
sci_name	0	1.00	3	69	0	38325	0
common_names	22731	0.59	3	218	0	21193	0
occurrence	122	1.00	7	16	0	4	0
occurrence_tags	53632	0.03	8	22	0	4	0
nativeness	1287	0.98	6	10	0	3	0
nativeness_tags	54921	0.00	8	17	0	4	0
abundance	13593	0.75	4	10	0	6	0
te_status	51367	0.07	1	6	0	18	0
state_status	52840	0.04	5	50	0	179	0
ozone_sensitive_status	54670	0.01	1	5	0	2	0
g_rank	19113	0.65	2	10	0	164	0
s_rank	30257	0.45	6	33	0	1271	0

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
references	0	1	1.19	2.06	0	0	1	2	88	
observations	0	1	7.15	119.04	0	0	0	0	11576	
vouchers	0	1	4.99	44.05	0	0	0	2	6931	

Data Analysis Plan

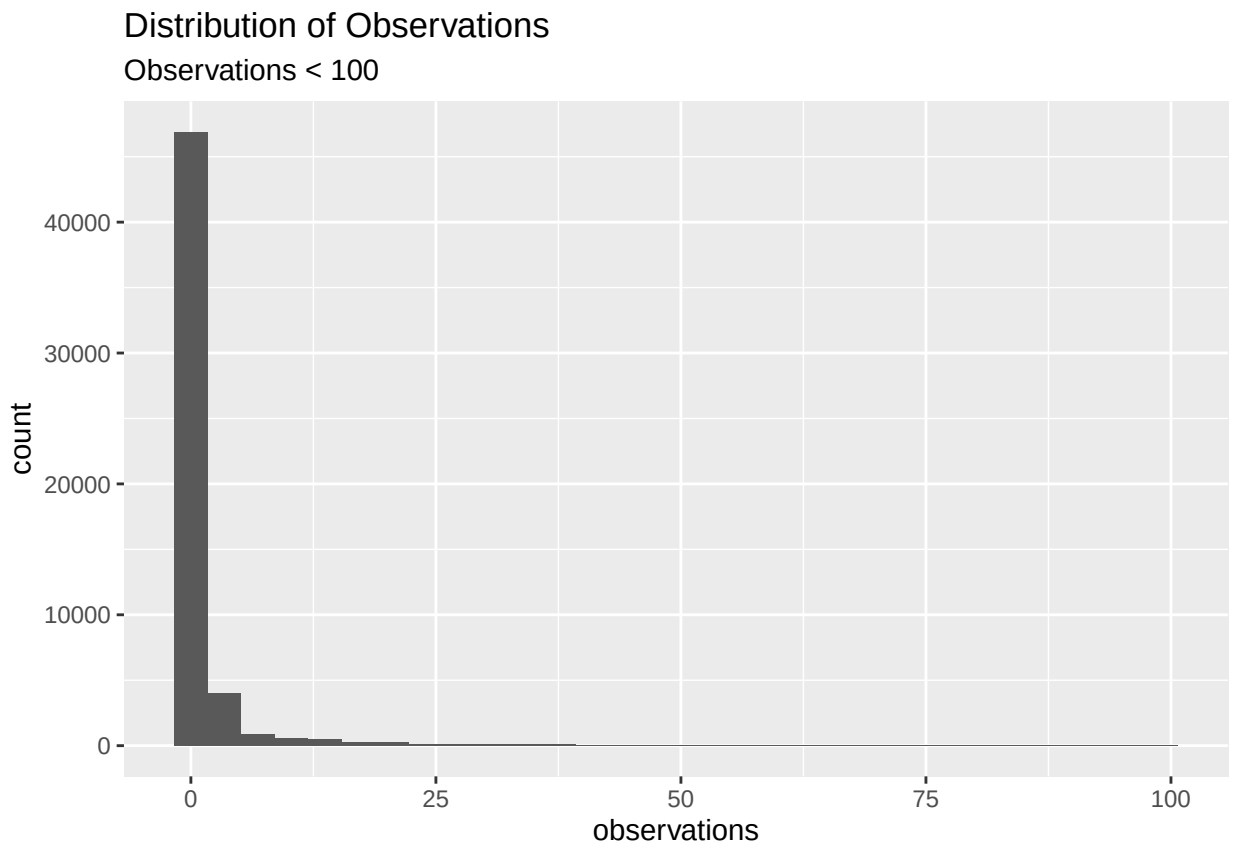
Predictor, Outcome Variables, and Comparison Groups: (1) For the first question, the predictor variable will be `evidences` (sum of `observations`, `references`, and `vouchers`) while the outcome will be `park_name` (ordered as a factor by popularity). This will be compared across each `category_name`.

(2) For the second question the predictor variable will be `nativeness` and the outcome vairable will be

`park_name` (again ordered as a factor by park popularity).

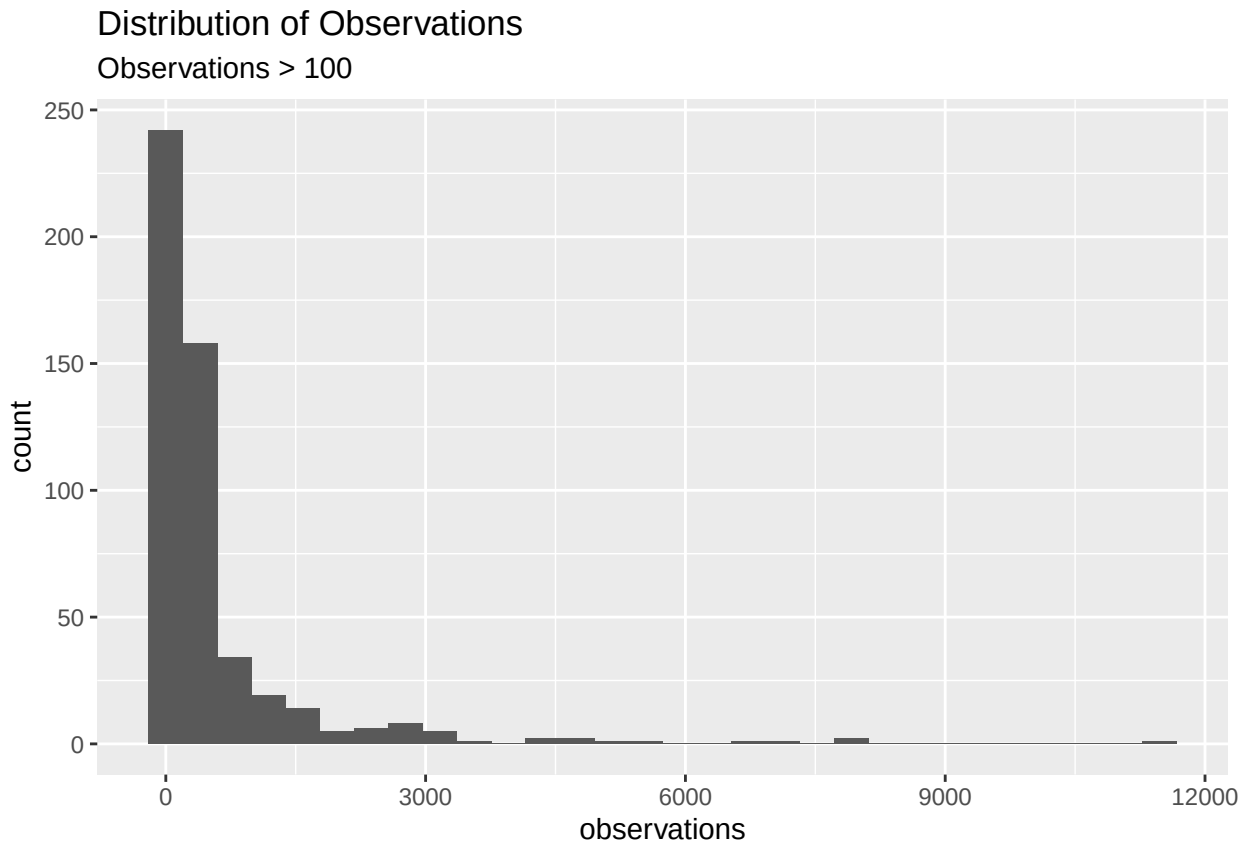
```
most_visited_nps_species_data_cleaned %>%
  filter(observations < 100) %>% # split between 2 graphs as the data was too squished
  ggplot(aes(x = observations)) +
  geom_histogram() +
  ggtitle("Distribution of Observations") + labs(subtitle = "Observations < 100")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
most_visited_nps_species_data_cleaned %>%
  filter(observations > 100) %>%
  ggplot(aes(x = observations)) +
  geom_histogram() +
  ggtitle("Distribution of Observations") + labs(subtitle = "Observations > 100")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
most_visited_nps_species_data_cleaned %>%
  group_by(park_name, category_name) %>%
  summarise(sum(observations), sum(references), sum(vouchers, mean(observations), mean(references), mean
```

```
## `summarise()` has grouped output by 'park_name'. You can override using the
## `.groups` argument.
```

```
## # A tibble: 138 x 5
## # Groups:   park_name [15]
##    park_name                 category_name `sum(observations)` `sum(references)`
##    <chr>                     <chr>                       <dbl>              <dbl>
##  1 Acadia National Park      Amphibian                      15                126
##  2 Acadia National Park      Bird                          286               4279
##  3 Acadia National Park      Fish                            1                106
##  4 Acadia National Park      Mammal                         37                476
##  5 Acadia National Park      Reptile                         5                 70
##  6 Acadia National Park      Vascular Pla~                   1               4625
##  7 Bryce Canyon National Pa~ Amphibian                       0                  8
##  8 Bryce Canyon National Pa~ Bird                            4                518
##  9 Bryce Canyon National Pa~ Fish                            0                  0
## 10 Bryce Canyon National Pa~ Mammal                          1                573
## # i 128 more rows
## # i 1 more variable:
## #   `sum(vouchers, mean(observations), mean(references), mean(vouchers))` <dbl>
```

The summary statistics help show that the amount of observations, references, and vouchers vary between parks and among the categories. The observations graphs show that there is a lot of variation in the number of observations between all the parks. It may be interesting to investigate if a particular category or park

has the massive amount of observations.

Methods

To test my questions I first would need to convert `park_name` into a factor (ordered by popularity). Then I would need to create another variable `evidences` as the sum of `observations`, `references`, and `vouchers`. It may then be helpful to create a graphic and factor by `park_name` and compare the most commonly "evidenced" categories across each park. If data follows the hyposthesized patterns, for each category `evidences` will be higher in the most popular parks. I would also want to create graphics that look at the number of different species within each park. This would be grouped by `nativeness`. If the data follows the hypothesized assotiation, parks with higher ratios of nativness will be more popular.