Day Two: What is a dataset? Worksheet

SDS 192: Introduction to Data Science

Lindsay Poirier Statistical & Data Sciences, Smith College

Spring 2022

Today, we're are going to work with a fun, light dataset!

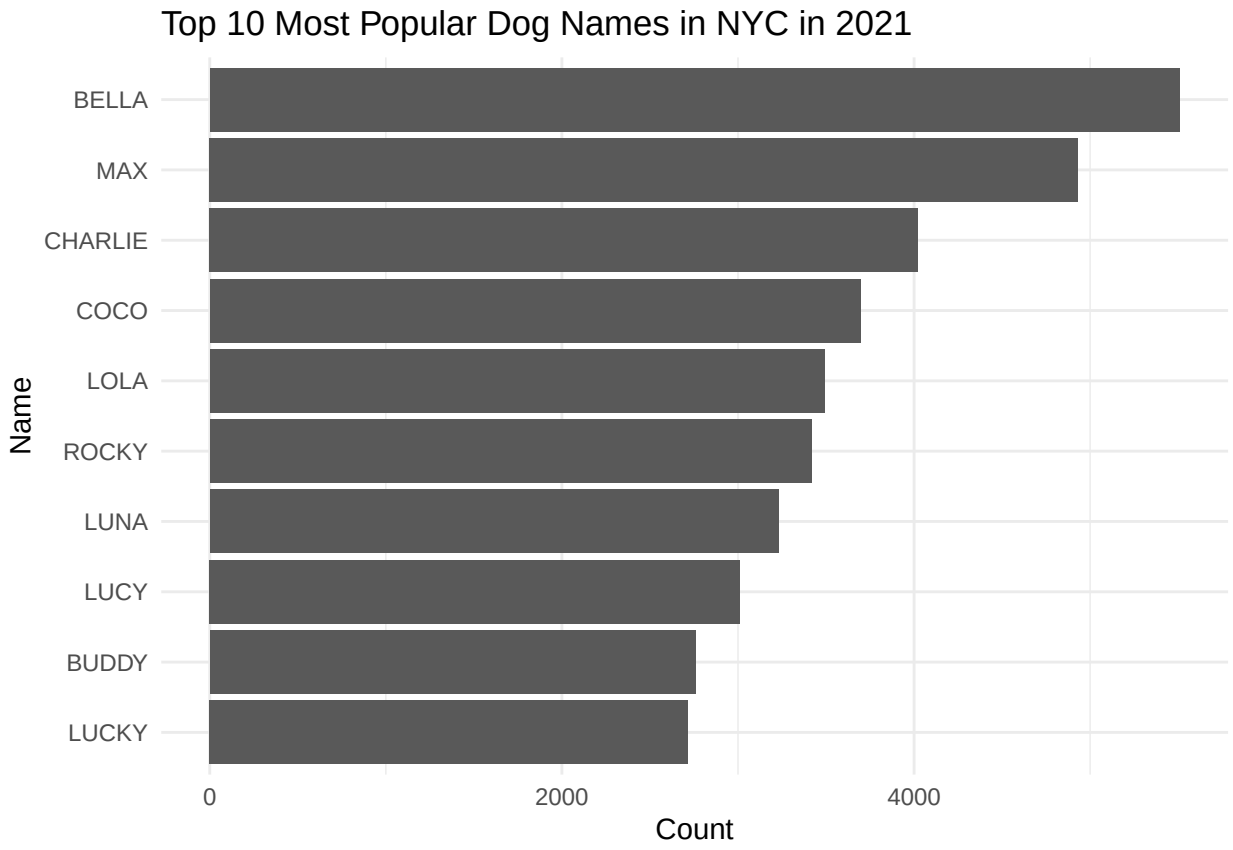
1. Navigate to the **NYC Dog Licensing Dataset**.
2. Note the information listed in the "About this Dataset" section. This is administrative metadata.
3. Note the attachment in this section with the file name: `DOHMHDataDictionary_Dog_Licenses_7.2021_.xlsx`.This contains descriptive metadata for this dataset.
4. Scroll to the 'Table Preview' at the bottom of the page. This previews this data as a rectangular dataset.
5. Answer the following questions by discussing in your groups:

- What is the unit of observation in this dataset? In other words, what does each row signify? How do you know?
- How frequently is this dataset updated? How do you know?
- What are the possible values for the AnimalGender variable in this dataset? How do you know?
- What is the value at index [4,4] in this dataset? How do you know?
- Identify the index of one missing value in this dataset.

6. As a preview of what you will be able to do in a few weeks, here is a fun visualization of the most popular dog names in 2021 in NYC!

```
library(tidyverse)
nyc_dog_names <- read.csv("https://data.cityofnewyork.us/api/views/nu7n-tubp/rows.csv")

nyc_dog_names %>%
  filter(!AnimalName %in% c("UNKNOWN", "NAME NOT PROVIDED")) %>%
  group_by(AnimalName) %>%
  summarize(count = n()) %>%
  top_n(10, count) %>%
  ggplot(aes(x = reorder(AnimalName, +count), y = count)) +
  geom_col() +
  coord_flip() +
  labs(title="Top 10 Most Popular Dog Names in NYC in 2021", x = "Name", y = "Count") +
  theme_minimal()
```


Top 10 Most Popular Dog Names in NYC in 2021
BELLA
MAX
CHARLIE
COCO
LOLA
ROCKY
LUNA
LUCY
BUDDY
LUCKY
Name
0
2000
4000
Count