Veterinary Hospital Patient List 2023

A look at popular names, breeds, and species among patients

Preston Stoltzfus

2023-08-04

## Goal

Using patient data of the active pets, find out the most popular names, species, and breeds of cats/dogs. The veterinary hospital uses a program called “cornerstone” (which is made by a company called “Idexx”) which keeps all the data that will be used for this project. Data includes all active patients in the hospital as of June 2023

## Getting Started

For this project I decided to use 3 different packages; “tidyverse”, “skimr”, “openxlsx”

library(tidyverse)

## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──  
## ✔ dplyr 1.1.2 ✔ readr 2.1.4  
## ✔ forcats 1.0.0 ✔ stringr 1.5.0  
## ✔ ggplot2 3.4.2 ✔ tibble 3.2.1  
## ✔ lubridate 1.9.2 ✔ tidyr 1.3.0  
## ✔ purrr 1.0.1   
## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ dplyr::lag() masks stats::lag()  
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors

library(skimr)  
library(openxlsx)

## Importing and getting familiar with the data set

After that I worked on importing my data and getting familiar with it. View() & head() help me get my bearings and skim() to help find errors that need cleaned.

vet\_patient\_list\_2023 <- read\_csv("C:/Users/prest/OneDrive/R/vet\_patient\_list\_2023/vet\_patient\_list\_2023.csv")

## Rows: 22566 Columns: 11  
## ── Column specification ────────────────────────────────────────────────────────  
## Delimiter: ","  
## chr (8): clientid, patientid, name, breedname, breedid, speciesname, species...  
## dbl (3): statusid, extension, active  
##   
## 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.

# getting familiar with my data  
View(vet\_patient\_list\_2023)  
head(vet\_patient\_list\_2023)

## # A tibble: 6 × 11  
## clientid patientid statusid name breedname breedid speciesname speciesid  
## <chr> <chr> <dbl> <chr> <chr> <chr> <chr> <chr>   
## 1 LTEST 18886 1 <NA> Amazon, Bl… ABF Avian AVI   
## 2 15086 23294 1 <NA> Domestic S… DSH Feline FEL   
## 3 671 16222 1 <NA> Axolotl AXO Reptile/Am… REA   
## 4 1277 2080 1 <NA> Domestic S… DSH Feline FEL   
## 5 4601 10809 1 Mask Maltese Mix MAX Canine CAN   
## 6 4601 10807 1 Mini Mom Maltese Mix MAX Canine CAN   
## # ℹ 3 more variables: extension <dbl>, active <dbl>, description <chr>

skim(vet\_patient\_list\_2023)

Data summary

|  |  |
| --- | --- |
| Name | vet\_patient\_list\_2023 |
| Number of rows | 22566 |
| Number of columns | 11 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Column type frequency: |  |
| character | 8 |
| numeric | 3 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Group variables | None |

**Variable type: character**

| skim\_variable | n\_missing | complete\_rate | min | max | empty | n\_unique | whitespace |
| --- | --- | --- | --- | --- | --- | --- | --- |
| clientid | 0 | 1 | 1 | 8 | 0 | 13128 | 0 |
| patientid | 0 | 1 | 1 | 10 | 0 | 22565 | 0 |
| name | 4 | 1 | 1 | 30 | 0 | 6799 | 0 |
| breedname | 0 | 1 | 3 | 30 | 0 | 615 | 0 |
| breedid | 0 | 1 | 1 | 3 | 0 | 564 | 0 |
| speciesname | 0 | 1 | 3 | 17 | 0 | 24 | 0 |
| speciesid | 0 | 1 | 1 | 3 | 0 | 24 | 0 |
| description | 0 | 1 | 4 | 13 | 0 | 6 | 0 |

**Variable type: numeric**

| skim\_variable | n\_missing | complete\_rate | mean | sd | p0 | p25 | p50 | p75 | p100 | hist |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| statusid | 0 | 1 | 1.0 | 0.00 | 1 | 1.00 | 1.0 | 1.00 | 1 | ▁▁▇▁▁ |
| extension | 22564 | 0 | 864.5 | 655.49 | 401 | 632.75 | 864.5 | 1096.25 | 1328 | ▇▁▁▁▇ |
| active | 0 | 1 | 1.0 | 0.00 | 1 | 1.00 | 1.0 | 1.00 | 1 | ▁▁▇▁▁ |

## I need to select out columns that I need for analysis. I use “head” and “skim” to get more familiar with this specific data set.

vet\_patient\_list\_2023\_selects <- vet\_patient\_list\_2023 %>%   
 select(name, breedname, speciesname, description)  
head(vet\_patient\_list\_2023\_selects)

## # A tibble: 6 × 4  
## name breedname speciesname description   
## <chr> <chr> <chr> <chr>   
## 1 <NA> Amazon, Blue Front Avian Female   
## 2 <NA> Domestic Shorthair Feline Female   
## 3 <NA> Axolotl Reptile/Amphibian Undetermined  
## 4 <NA> Domestic Shorthair Feline Male   
## 5 Mask Maltese Mix Canine Male   
## 6 Mini Mom Maltese Mix Canine Male

skim(vet\_patient\_list\_2023\_selects)

Data summary

|  |  |
| --- | --- |
| Name | vet\_patient\_list\_2023\_sel… |
| Number of rows | 22566 |
| Number of columns | 4 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Column type frequency: |  |
| character | 4 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Group variables | None |

**Variable type: character**

| skim\_variable | n\_missing | complete\_rate | min | max | empty | n\_unique | whitespace |
| --- | --- | --- | --- | --- | --- | --- | --- |
| name | 4 | 1 | 1 | 30 | 0 | 6799 | 0 |
| breedname | 0 | 1 | 3 | 30 | 0 | 615 | 0 |
| speciesname | 0 | 1 | 3 | 17 | 0 | 24 | 0 |
| description | 0 | 1 | 4 | 13 | 0 | 6 | 0 |

## The cleaning begins, focusing first on the “decription” (soon to be “sex”) column.

# renaming column from "description" to "sex"  
vet\_patient\_list\_2023\_selects <- vet\_patient\_list\_2023\_selects %>%  
 rename(sex = description)  
  
# changing data in data columns to lower case and cleaning any whitespace that may/may not be there  
vet\_patient\_list\_2023\_selects$breedname <- trimws(tolower(vet\_patient\_list\_2023\_selects$breedname))  
vet\_patient\_list\_2023\_selects$sex <- trimws(tolower(vet\_patient\_list\_2023\_selects$sex))  
vet\_patient\_list\_2023\_selects$name <- trimws(tolower(vet\_patient\_list\_2023\_selects$name))  
vet\_patient\_list\_2023\_selects$speciesname <- trimws(tolower(vet\_patient\_list\_2023\_selects$speciesname))  
  
# Cleaning up the "sex" column  
vet\_patient\_list\_2023\_selects$sex <- gsub("undetermined", "unknown", vet\_patient\_list\_2023\_selects$sex)  
vet\_patient\_list\_2023\_selects$sex <- gsub("neutered male", "male, neutered", vet\_patient\_list\_2023\_selects$sex)  
vet\_patient\_list\_2023\_selects$sex <- gsub("spayed female", "female, spayed", vet\_patient\_list\_2023\_selects$sex)

## “name” column

### In some cases I wanted to remove any names with variances of the word so used the function “grepl”. In other cases I wanted to remove only specific names either because it’s simplier code wise to type in and/or because sometimes “grepl” cast to wide of a net. A lot of the names that are things like “kitten”, “yellow collar”, etc are just leftover patients who came in for a litter exam and were never saw again so need to remove them. There are also several “fake” accounts used for in-house things.

rows\_to\_remove\_names <- c("general use", "kittens-general use", "zzgeneral stuff", "aaaestimates", "aaestimate",   
 "a center use", "a", "communications", "misc cat", "misc", "miscellaneous", "unknown", "cubex test", "Transfer to Downtown", "Transfer to Downtown", "zzzzzz")  
  
vet\_patient\_list\_2023\_selects <- vet\_patient\_list\_2023\_selects[!(vet\_patient\_list\_2023\_selects$name %in% rows\_to\_remove\_names), ]  
  
vet\_patient\_list\_2023\_selects <- subset(vet\_patient\_list\_2023\_selects, !grepl("client", name, ignore.case = TRUE))  
vet\_patient\_list\_2023\_selects <- subset(vet\_patient\_list\_2023\_selects, !grepl("merch", name, ignore.case = TRUE))  
vet\_patient\_list\_2023\_selects <- subset(vet\_patient\_list\_2023\_selects, !grepl("pup", name, ignore.case = TRUE))  
vet\_patient\_list\_2023\_selects <- subset(vet\_patient\_list\_2023\_selects, !grepl("kitten", name, ignore.case = TRUE))  
vet\_patient\_list\_2023\_selects <- subset(vet\_patient\_list\_2023\_selects, !grepl("\\bpink\\b", name))  
vet\_patient\_list\_2023\_selects <- subset(vet\_patient\_list\_2023\_selects, !grepl("miss blue", name, ignore.case = TRUE))  
vet\_patient\_list\_2023\_selects <- subset(vet\_patient\_list\_2023\_selects, !grepl("mr blue", name, ignore.case = TRUE))  
vet\_patient\_list\_2023\_selects <- subset(vet\_patient\_list\_2023\_selects, !grepl("yellow", name, ignore.case = TRUE))  
vet\_patient\_list\_2023\_selects <- subset(vet\_patient\_list\_2023\_selects, !grepl("mr red", name, ignore.case = TRUE))  
vet\_patient\_list\_2023\_selects <- subset(vet\_patient\_list\_2023\_selects, !grepl("miss red", name, ignore.case = TRUE))  
vet\_patient\_list\_2023\_selects <- subset(vet\_patient\_list\_2023\_selects, !grepl("red collar", name, ignore.case = TRUE))  
vet\_patient\_list\_2023\_selects <- subset(vet\_patient\_list\_2023\_selects, !grepl("\\d", name))  
vet\_patient\_list\_2023\_selects <- subset(vet\_patient\_list\_2023\_selects, !grepl("miss white", name, ignore.case = TRUE))  
vet\_patient\_list\_2023\_selects <- subset(vet\_patient\_list\_2023\_selects, !grepl("miss purple", name, ignore.case = TRUE))  
vet\_patient\_list\_2023\_selects <- subset(vet\_patient\_list\_2023\_selects, !grepl("mr black", name, ignore.case = TRUE))  
vet\_patient\_list\_2023\_selects <- subset(vet\_patient\_list\_2023\_selects, !grepl("mr gray", name, ignore.case = TRUE))  
vet\_patient\_list\_2023\_selects <- subset(vet\_patient\_list\_2023\_selects, !grepl("mr brown", name, ignore.case = TRUE))  
vet\_patient\_list\_2023\_selects <- subset(vet\_patient\_list\_2023\_selects, !grepl("mr green", name, ignore.case = TRUE))  
vet\_patient\_list\_2023\_selects <- subset(vet\_patient\_list\_2023\_selects, !grepl("mr orange", name, ignore.case = TRUE))  
vet\_patient\_list\_2023\_selects <- subset(vet\_patient\_list\_2023\_selects, !grepl("collar", name, ignore.case = TRUE))  
vet\_patient\_list\_2023\_selects <- subset(vet\_patient\_list\_2023\_selects, !grepl("test", name, ignore.case = TRUE))  
vet\_patient\_list\_2023\_selects <- subset(vet\_patient\_list\_2023\_selects, !grepl("experiment", name, ignore.case = TRUE))  
vet\_patient\_list\_2023\_selects <- subset(vet\_patient\_list\_2023\_selects, !grepl("transfer", name, ignore.case = TRUE))  
vet\_patient\_list\_2023\_selects <- subset(vet\_patient\_list\_2023\_selects, !grepl("idexx", name, ignore.case = TRUE))  
vet\_patient\_list\_2023\_selects <- subset(vet\_patient\_list\_2023\_selects, !grepl("invoice", name, ignore.case = TRUE))  
vet\_patient\_list\_2023\_selects <- subset(vet\_patient\_list\_2023\_selects, !grepl("stray", name, ignore.case = TRUE))  
vet\_patient\_list\_2023\_selects <- subset(vet\_patient\_list\_2023\_selects, !grepl("found", name, ignore.case = TRUE))  
  
vet\_patient\_list\_2023\_selects <- subset(vet\_patient\_list\_2023\_selects, !is.na(name))

## “speciesname” column

### There were severall in-house accounts used for various purposes that need to go for a proper analysis.

vet\_patient\_list\_2023\_selects <- vet\_patient\_list\_2023\_selects[!(vet\_patient\_list\_2023\_selects$speciesname %in% c("xxxxxxx", "cash small", "cash large")), ]

## “breedname” column

### I needed to change the same breeds with multiple names to a consistent name while not combining the “mixed” breeds. For example I don’t want to combine “yorkshire terrier” with “yorkshire terrier mix” but yet still need to format things properly “terrier, yorkshire”. Currently the words of the “breedname” are in different orders as well so I needed a function to find them in any way but exclude certain words.

vet\_patient\_list\_2023\_selects$breedname <- ifelse(  
 grepl("yorkshire", vet\_patient\_list\_2023\_selects$breedname) &  
 grepl("terrier", vet\_patient\_list\_2023\_selects$breedname) &  
 !grepl("mix", vet\_patient\_list\_2023\_selects$breedname),  
 "terrier, yorkshire",  
 vet\_patient\_list\_2023\_selects$breedname)  
  
vet\_patient\_list\_2023\_selects$breedname <- ifelse(  
 grepl("jack", vet\_patient\_list\_2023\_selects$breedname) &  
 grepl("russell", vet\_patient\_list\_2023\_selects$breedname) &  
 grepl("terrier", vet\_patient\_list\_2023\_selects$breedname) &  
 !grepl("mix", vet\_patient\_list\_2023\_selects$breedname),  
 "terrier, jack russell",  
 vet\_patient\_list\_2023\_selects$breedname)  
  
vet\_patient\_list\_2023\_selects$breedname <- ifelse(  
 grepl("jack", vet\_patient\_list\_2023\_selects$breedname) &  
 grepl("russell", vet\_patient\_list\_2023\_selects$breedname) &  
 grepl("terrier", vet\_patient\_list\_2023\_selects$breedname) &  
 grepl("mix", vet\_patient\_list\_2023\_selects$breedname),  
 "terrier, jack russell mix",  
 vet\_patient\_list\_2023\_selects$breedname)  
  
vet\_patient\_list\_2023\_selects$breedname <- ifelse(  
 grepl("bull", vet\_patient\_list\_2023\_selects$breedname) &  
 grepl("terrier", vet\_patient\_list\_2023\_selects$breedname) &  
 grepl("mix", vet\_patient\_list\_2023\_selects$breedname),  
 "terrier, bull mix",  
 vet\_patient\_list\_2023\_selects$breedname)  
  
vet\_patient\_list\_2023\_selects$breedname <- ifelse(  
 grepl("fox", vet\_patient\_list\_2023\_selects$breedname) &  
 grepl("(smooth)", vet\_patient\_list\_2023\_selects$breedname) &  
 grepl("terrier", vet\_patient\_list\_2023\_selects$breedname) &  
 !grepl("mix", vet\_patient\_list\_2023\_selects$breedname),  
 "terrier, fox smooth",  
 vet\_patient\_list\_2023\_selects$breedname)  
  
vet\_patient\_list\_2023\_selects$breedname <- ifelse(  
 grepl("boston", vet\_patient\_list\_2023\_selects$breedname) &  
 grepl("terrier", vet\_patient\_list\_2023\_selects$breedname) &  
 !grepl("mix", vet\_patient\_list\_2023\_selects$breedname),  
 "terrier, boston",  
 vet\_patient\_list\_2023\_selects$breedname)  
  
vet\_patient\_list\_2023\_selects$breedname <- ifelse(  
 grepl("cairn", vet\_patient\_list\_2023\_selects$breedname) &  
 grepl("terrier", vet\_patient\_list\_2023\_selects$breedname) &  
 !grepl("mix", vet\_patient\_list\_2023\_selects$breedname),  
 "terrier, cairn",  
 vet\_patient\_list\_2023\_selects$breedname)  
  
vet\_patient\_list\_2023\_selects$breedname <- ifelse(  
 grepl("tibetan", vet\_patient\_list\_2023\_selects$breedname) &  
 grepl("terrier", vet\_patient\_list\_2023\_selects$breedname) &  
 !grepl("mix", vet\_patient\_list\_2023\_selects$breedname),  
 "terrier, tibetan",  
 vet\_patient\_list\_2023\_selects$breedname)  
  
vet\_patient\_list\_2023\_selects$breedname <- ifelse(  
 grepl("american", vet\_patient\_list\_2023\_selects$breedname) &  
 grepl("staffordshire", vet\_patient\_list\_2023\_selects$breedname) &  
 !grepl("mix", vet\_patient\_list\_2023\_selects$breedname),  
 "terrier, american staffordshire",  
 vet\_patient\_list\_2023\_selects$breedname)  
  
vet\_patient\_list\_2023\_selects$breedname <- ifelse(  
 grepl("americanstaffordshire", vet\_patient\_list\_2023\_selects$breedname) &  
 grepl("terrier", vet\_patient\_list\_2023\_selects$breedname) &  
 !grepl("mix", vet\_patient\_list\_2023\_selects$breedname),  
 "terrier, american staffordshire",  
 vet\_patient\_list\_2023\_selects$breedname)  
  
vet\_patient\_list\_2023\_selects$breedname <- ifelse(  
 grepl("west", vet\_patient\_list\_2023\_selects$breedname) &  
 grepl("highland", vet\_patient\_list\_2023\_selects$breedname) &  
 grepl("terrier", vet\_patient\_list\_2023\_selects$breedname) &  
 !grepl("mix", vet\_patient\_list\_2023\_selects$breedname),  
 "terrier, west highland white",  
 vet\_patient\_list\_2023\_selects$breedname)  
  
vet\_patient\_list\_2023\_selects$breedname <- ifelse(  
 grepl("biewerterrier", vet\_patient\_list\_2023\_selects$breedname) &  
 !grepl("mix", vet\_patient\_list\_2023\_selects$breedname),  
 "terrier, biewer",  
 vet\_patient\_list\_2023\_selects$breedname)  
  
vet\_patient\_list\_2023\_selects$breedname <- ifelse(  
 grepl("aussie-doodle", vet\_patient\_list\_2023\_selects$breedname) &  
 grepl("mini", vet\_patient\_list\_2023\_selects$breedname) &  
 !grepl("mix", vet\_patient\_list\_2023\_selects$breedname),  
 "aussiedoodle, mini",  
 vet\_patient\_list\_2023\_selects$breedname)  
  
vet\_patient\_list\_2023\_selects$breedname <- ifelse(  
 grepl("labradoodle", vet\_patient\_list\_2023\_selects$breedname) &  
 grepl("mini", vet\_patient\_list\_2023\_selects$breedname) &  
 !grepl("mix", vet\_patient\_list\_2023\_selects$breedname),  
 "labradoodle, mini",  
 vet\_patient\_list\_2023\_selects$breedname)  
  
vet\_patient\_list\_2023\_selects$breedname <- ifelse(  
 grepl("cock-apoo", vet\_patient\_list\_2023\_selects$breedname) &  
 !grepl("mix", vet\_patient\_list\_2023\_selects$breedname),  
 "cockapoo",  
 vet\_patient\_list\_2023\_selects$breedname)  
  
vet\_patient\_list\_2023\_selects$breedname <- ifelse(  
 grepl("yorkie-poo", vet\_patient\_list\_2023\_selects$breedname) &  
 !grepl("mix", vet\_patient\_list\_2023\_selects$breedname),  
 "terrier, yorkie-poo",  
 vet\_patient\_list\_2023\_selects$breedname)  
  
vet\_patient\_list\_2023\_selects$breedname <- ifelse(  
 grepl("dsh", vet\_patient\_list\_2023\_selects$breedname) &  
 !grepl("mix", vet\_patient\_list\_2023\_selects$breedname),  
 "domestic shorthair",  
 vet\_patient\_list\_2023\_selects$breedname)  
  
vet\_patient\_list\_2023\_selects$breedname <- ifelse(  
 grepl("labrador", vet\_patient\_list\_2023\_selects$breedname) &  
 grepl("retriever", vet\_patient\_list\_2023\_selects$breedname) &  
 !grepl("mix", vet\_patient\_list\_2023\_selects$breedname),  
 "retriever, labrador",  
 vet\_patient\_list\_2023\_selects$breedname)  
  
vet\_patient\_list\_2023\_selects$breedname <- ifelse(  
 grepl("golden", vet\_patient\_list\_2023\_selects$breedname) &  
 grepl("retriever", vet\_patient\_list\_2023\_selects$breedname) &  
 !grepl("mix", vet\_patient\_list\_2023\_selects$breedname),  
 "retriever, golden",  
 vet\_patient\_list\_2023\_selects$breedname)  
  
vet\_patient\_list\_2023\_selects$breedname <- ifelse(  
 grepl("dakota", vet\_patient\_list\_2023\_selects$breedname) &  
 grepl("sport", vet\_patient\_list\_2023\_selects$breedname) &  
 grepl("retriever", vet\_patient\_list\_2023\_selects$breedname) &  
 !grepl("mix", vet\_patient\_list\_2023\_selects$breedname),  
 "retriever, dakota sport",  
 vet\_patient\_list\_2023\_selects$breedname)  
  
vet\_patient\_list\_2023\_selects$breedname <- ifelse(  
 grepl("retriever", vet\_patient\_list\_2023\_selects$breedname) &  
 grepl("mixed", vet\_patient\_list\_2023\_selects$breedname),  
 "retriever mix",  
 vet\_patient\_list\_2023\_selects$breedname)  
  
vet\_patient\_list\_2023\_selects$breedname <- ifelse(  
 grepl("german", vet\_patient\_list\_2023\_selects$breedname) &  
 grepl("shepherd", vet\_patient\_list\_2023\_selects$breedname) &  
 !grepl("mix", vet\_patient\_list\_2023\_selects$breedname),  
 "shepherd, german",  
 vet\_patient\_list\_2023\_selects$breedname)  
  
vet\_patient\_list\_2023\_selects$breedname <- ifelse(  
 grepl("australian", vet\_patient\_list\_2023\_selects$breedname) &  
 grepl("shepherd", vet\_patient\_list\_2023\_selects$breedname) &  
 !grepl("mix", vet\_patient\_list\_2023\_selects$breedname),  
 "shepherd, australian",  
 vet\_patient\_list\_2023\_selects$breedname)  
  
vet\_patient\_list\_2023\_selects$breedname <- ifelse(  
 grepl("anatolian", vet\_patient\_list\_2023\_selects$breedname) &  
 grepl("shepherd", vet\_patient\_list\_2023\_selects$breedname) &  
 !grepl("mix", vet\_patient\_list\_2023\_selects$breedname),  
 "shepherd, anatolian",  
 vet\_patient\_list\_2023\_selects$breedname)  
  
vet\_patient\_list\_2023\_selects$breedname <- ifelse(  
 grepl("french", vet\_patient\_list\_2023\_selects$breedname) &  
 grepl("bulldog", vet\_patient\_list\_2023\_selects$breedname) &  
 !grepl("mix", vet\_patient\_list\_2023\_selects$breedname),  
 "bulldog, french",  
 vet\_patient\_list\_2023\_selects$breedname)  
  
vet\_patient\_list\_2023\_selects$breedname <- ifelse(  
 grepl("pit", vet\_patient\_list\_2023\_selects$breedname) &  
 grepl("bull", vet\_patient\_list\_2023\_selects$breedname) &  
 grepl("terrier", vet\_patient\_list\_2023\_selects$breedname) &  
 !grepl("mix", vet\_patient\_list\_2023\_selects$breedname),  
 "pitbull",  
 vet\_patient\_list\_2023\_selects$breedname)  
  
vet\_patient\_list\_2023\_selects$breedname <- ifelse(  
 grepl("african", vet\_patient\_list\_2023\_selects$breedname) &  
 grepl("gray", vet\_patient\_list\_2023\_selects$breedname) &  
 !grepl("mix", vet\_patient\_list\_2023\_selects$breedname),  
 "parrot, african grey",  
 vet\_patient\_list\_2023\_selects$breedname)  
  
vet\_patient\_list\_2023\_selects$breedname <- ifelse(  
 grepl("belgian", vet\_patient\_list\_2023\_selects$breedname) &  
 grepl("malinois", vet\_patient\_list\_2023\_selects$breedname) &  
 !grepl("mix", vet\_patient\_list\_2023\_selects$breedname),  
 "malinois, belgian",  
 vet\_patient\_list\_2023\_selects$breedname)  
  
vet\_patient\_list\_2023\_selects$breedname <- ifelse(  
 grepl("wolfhound", vet\_patient\_list\_2023\_selects$breedname) &  
 grepl("irish", vet\_patient\_list\_2023\_selects$breedname) &  
 !grepl("mix", vet\_patient\_list\_2023\_selects$breedname),  
 "wolfhound, irish",  
 vet\_patient\_list\_2023\_selects$breedname)  
  
vet\_patient\_list\_2023\_selects$breedname <- ifelse(  
 grepl("galgo", vet\_patient\_list\_2023\_selects$breedname) &  
 grepl("spanish", vet\_patient\_list\_2023\_selects$breedname) &  
 grepl("greyhound", vet\_patient\_list\_2023\_selects$breedname) &  
 !grepl("mix", vet\_patient\_list\_2023\_selects$breedname),  
 "greyhound, spanish",  
 vet\_patient\_list\_2023\_selects$breedname)  
  
vet\_patient\_list\_2023\_selects$breedname <- ifelse(  
 grepl("corgi", vet\_patient\_list\_2023\_selects$breedname) &  
 grepl("welsh", vet\_patient\_list\_2023\_selects$breedname) &  
 grepl("pembroke", vet\_patient\_list\_2023\_selects$breedname) &  
 !grepl("mix", vet\_patient\_list\_2023\_selects$breedname),  
 "corgi, pembroke welsh",  
 vet\_patient\_list\_2023\_selects$breedname)  
  
vet\_patient\_list\_2023\_selects$breedname <- ifelse(  
 grepl("dlh", vet\_patient\_list\_2023\_selects$breedname) &  
 !grepl("mix", vet\_patient\_list\_2023\_selects$breedname),  
 "domestic longhair",  
 vet\_patient\_list\_2023\_selects$breedname)  
  
vet\_patient\_list\_2023\_selects$breedname <- ifelse(  
 grepl("dmh", vet\_patient\_list\_2023\_selects$breedname) &  
 !grepl("mix", vet\_patient\_list\_2023\_selects$breedname),  
 "domestic mediumhair",  
 vet\_patient\_list\_2023\_selects$breedname)  
  
vet\_patient\_list\_2023\_selects$breedname <- ifelse(  
 grepl("husky", vet\_patient\_list\_2023\_selects$breedname) &  
 grepl("siberian", vet\_patient\_list\_2023\_selects$breedname) &  
 !grepl("mix", vet\_patient\_list\_2023\_selects$breedname),  
 "husky, siberian",  
 vet\_patient\_list\_2023\_selects$breedname)  
  
vet\_patient\_list\_2023\_selects$breedname <- ifelse(  
 grepl("maine", vet\_patient\_list\_2023\_selects$breedname) &  
 grepl("coon", vet\_patient\_list\_2023\_selects$breedname) &  
 !grepl("mix", vet\_patient\_list\_2023\_selects$breedname),  
 "coon, maine",  
 vet\_patient\_list\_2023\_selects$breedname)  
  
vet\_patient\_list\_2023\_selects$breedname <- ifelse(  
 grepl("bullmastiff", vet\_patient\_list\_2023\_selects$breedname) &  
 !grepl("mix", vet\_patient\_list\_2023\_selects$breedname),  
 "mastiff, bull",  
 vet\_patient\_list\_2023\_selects$breedname)  
  
vet\_patient\_list\_2023\_selects$breedname <- ifelse(  
 grepl("bullmastiff", vet\_patient\_list\_2023\_selects$breedname) &  
 grepl("mix", vet\_patient\_list\_2023\_selects$breedname),  
 "mastiff, bull mix",  
 vet\_patient\_list\_2023\_selects$breedname)  
  
vet\_patient\_list\_2023\_selects$breedname <- ifelse(  
 grepl("bernese", vet\_patient\_list\_2023\_selects$breedname) &  
 grepl("mountain", vet\_patient\_list\_2023\_selects$breedname) &  
 grepl("dog", vet\_patient\_list\_2023\_selects$breedname) &  
 !grepl("mix", vet\_patient\_list\_2023\_selects$breedname),  
 "mountain dog, bernese",  
 vet\_patient\_list\_2023\_selects$breedname)  
  
vet\_patient\_list\_2023\_selects$breedname <- ifelse(  
 grepl("bernese", vet\_patient\_list\_2023\_selects$breedname) &  
 grepl("mountain", vet\_patient\_list\_2023\_selects$breedname) &  
 grepl("dog", vet\_patient\_list\_2023\_selects$breedname) &  
 grepl("mix", vet\_patient\_list\_2023\_selects$breedname),  
 "mountain dog, bernese mix",  
 vet\_patient\_list\_2023\_selects$breedname)  
  
vet\_patient\_list\_2023\_selects$breedname <- ifelse(  
 grepl("greater", vet\_patient\_list\_2023\_selects$breedname) &  
 grepl("swiss", vet\_patient\_list\_2023\_selects$breedname) &  
 grepl("mountain", vet\_patient\_list\_2023\_selects$breedname) &  
 grepl("dog", vet\_patient\_list\_2023\_selects$breedname) &  
 !grepl("mix", vet\_patient\_list\_2023\_selects$breedname),  
 "mountain dog, greater swiss",  
 vet\_patient\_list\_2023\_selects$breedname)  
  
vet\_patient\_list\_2023\_selects$breedname <- ifelse(  
 grepl("oriental", vet\_patient\_list\_2023\_selects$breedname) &  
 grepl("longhair", vet\_patient\_list\_2023\_selects$breedname) &  
 !grepl("mix", vet\_patient\_list\_2023\_selects$breedname),  
 "oriental, longhair",  
 vet\_patient\_list\_2023\_selects$breedname)  
  
vet\_patient\_list\_2023\_selects$breedname <- ifelse(  
 grepl("oriental", vet\_patient\_list\_2023\_selects$breedname) &  
 grepl("shorthair", vet\_patient\_list\_2023\_selects$breedname) &  
 !grepl("mix", vet\_patient\_list\_2023\_selects$breedname),  
 "oriental, shorthair",  
 vet\_patient\_list\_2023\_selects$breedname)  
  
vet\_patient\_list\_2023\_selects$breedname <- ifelse(  
 grepl("other", vet\_patient\_list\_2023\_selects$breedname) &  
 grepl("canin", vet\_patient\_list\_2023\_selects$breedname) &  
 !grepl("mix", vet\_patient\_list\_2023\_selects$breedname),  
 "other",  
 vet\_patient\_list\_2023\_selects$breedname)  
  
vet\_patient\_list\_2023\_selects$breedname <- ifelse(  
 grepl("other", vet\_patient\_list\_2023\_selects$breedname) &  
 grepl("reptile/amphibian", vet\_patient\_list\_2023\_selects$breedname) &  
 !grepl("mix", vet\_patient\_list\_2023\_selects$breedname),  
 "other reptile",  
 vet\_patient\_list\_2023\_selects$breedname)  
  
vet\_patient\_list\_2023\_selects$breedname <- ifelse(  
 grepl("other", vet\_patient\_list\_2023\_selects$breedname) &  
 grepl("reptile", vet\_patient\_list\_2023\_selects$breedname) &  
 !grepl("mix", vet\_patient\_list\_2023\_selects$breedname),  
 "other reptile",  
 vet\_patient\_list\_2023\_selects$breedname)  
  
vet\_patient\_list\_2023\_selects$breedname <- ifelse(  
 grepl("other", vet\_patient\_list\_2023\_selects$breedname) &  
 grepl("breed", vet\_patient\_list\_2023\_selects$breedname) &  
 !grepl("mix", vet\_patient\_list\_2023\_selects$breedname),  
 "other",  
 vet\_patient\_list\_2023\_selects$breedname)  
  
vet\_patient\_list\_2023\_selects$breedname <- ifelse(  
 grepl("miniature", vet\_patient\_list\_2023\_selects$breedname) &  
 grepl("schnauzer", vet\_patient\_list\_2023\_selects$breedname) &  
 !grepl("mix", vet\_patient\_list\_2023\_selects$breedname),  
 "schnauzer, miniature",  
 vet\_patient\_list\_2023\_selects$breedname)  
  
vet\_patient\_list\_2023\_selects$breedname <- ifelse(  
 grepl("english", vet\_patient\_list\_2023\_selects$breedname) &  
 grepl("setter", vet\_patient\_list\_2023\_selects$breedname) &  
 !grepl("mix", vet\_patient\_list\_2023\_selects$breedname),  
 "setter, english",  
 vet\_patient\_list\_2023\_selects$breedname)  
  
vet\_patient\_list\_2023\_selects$breedname <- ifelse(  
 grepl("rhodesian", vet\_patient\_list\_2023\_selects$breedname) &  
 grepl("ridgeback", vet\_patient\_list\_2023\_selects$breedname) &  
 !grepl("mix", vet\_patient\_list\_2023\_selects$breedname),  
 "ridgeback, rhodesian",  
 vet\_patient\_list\_2023\_selects$breedname)  
  
vet\_patient\_list\_2023\_selects$breedname <- ifelse(  
 grepl("shorthair", vet\_patient\_list\_2023\_selects$breedname) &  
 grepl("american", vet\_patient\_list\_2023\_selects$breedname) &  
 !grepl("mix", vet\_patient\_list\_2023\_selects$breedname),  
 "domestic shorthair",  
 vet\_patient\_list\_2023\_selects$breedname)  
  
vet\_patient\_list\_2023\_selects$breedname <- ifelse(  
 grepl("short", vet\_patient\_list\_2023\_selects$breedname) &  
 grepl("haired", vet\_patient\_list\_2023\_selects$breedname) &  
 !grepl("mix", vet\_patient\_list\_2023\_selects$breedname),  
 "shorthair, misc",  
 vet\_patient\_list\_2023\_selects$breedname)  
  
vet\_patient\_list\_2023\_selects$breedname <- ifelse(  
 grepl("short", vet\_patient\_list\_2023\_selects$breedname) &  
 grepl("haired", vet\_patient\_list\_2023\_selects$breedname) &  
 !grepl("american", vet\_patient\_list\_2023\_selects$breedname) &  
 !grepl("mix", vet\_patient\_list\_2023\_selects$breedname),  
 "shorthair, misc",  
 vet\_patient\_list\_2023\_selects$breedname)  
  
vet\_patient\_list\_2023\_selects$breedname <- ifelse(  
 grepl("shorthair", vet\_patient\_list\_2023\_selects$breedname) &  
 grepl("english", vet\_patient\_list\_2023\_selects$breedname) &  
 !grepl("mix", vet\_patient\_list\_2023\_selects$breedname),  
 "shorthair, english",  
 vet\_patient\_list\_2023\_selects$breedname)  
  
vet\_patient\_list\_2023\_selects$breedname <- ifelse(  
 grepl("short", vet\_patient\_list\_2023\_selects$breedname) &  
 grepl("haired", vet\_patient\_list\_2023\_selects$breedname) &  
 grepl("american", vet\_patient\_list\_2023\_selects$breedname) &  
 !grepl("mix", vet\_patient\_list\_2023\_selects$breedname),  
 "shorthair, english",  
 vet\_patient\_list\_2023\_selects$breedname)  
  
vet\_patient\_list\_2023\_selects$breedname <- ifelse(  
 grepl("welsh", vet\_patient\_list\_2023\_selects$breedname) &  
 grepl("springer", vet\_patient\_list\_2023\_selects$breedname) &  
 grepl("spaniel", vet\_patient\_list\_2023\_selects$breedname) &  
 !grepl("mix", vet\_patient\_list\_2023\_selects$breedname),  
 "spaniel, welsh springer",  
 vet\_patient\_list\_2023\_selects$breedname)  
  
vet\_patient\_list\_2023\_selects$breedname <- ifelse(  
 grepl("english", vet\_patient\_list\_2023\_selects$breedname) &  
 grepl("springer", vet\_patient\_list\_2023\_selects$breedname) &  
 grepl("spaniel", vet\_patient\_list\_2023\_selects$breedname) &  
 !grepl("mix", vet\_patient\_list\_2023\_selects$breedname),  
 "spaniel, english springer",  
 vet\_patient\_list\_2023\_selects$breedname)  
  
vet\_patient\_list\_2023\_selects$breedname <- ifelse(  
 grepl("cocker", vet\_patient\_list\_2023\_selects$breedname) &  
 grepl("spaniel", vet\_patient\_list\_2023\_selects$breedname) &  
 !grepl("mix", vet\_patient\_list\_2023\_selects$breedname),  
 "spaniel, cocker",  
 vet\_patient\_list\_2023\_selects$breedname)  
  
vet\_patient\_list\_2023\_selects$breedname <- ifelse(  
 grepl("clumber", vet\_patient\_list\_2023\_selects$breedname) &  
 grepl("spaniel", vet\_patient\_list\_2023\_selects$breedname) &  
 !grepl("mix", vet\_patient\_list\_2023\_selects$breedname),  
 "spaniel, clumber",  
 vet\_patient\_list\_2023\_selects$breedname)  
  
vet\_patient\_list\_2023\_selects$breedname <- ifelse(  
 grepl("cavalier", vet\_patient\_list\_2023\_selects$breedname) &  
 grepl("king", vet\_patient\_list\_2023\_selects$breedname) &  
 grepl("charles", vet\_patient\_list\_2023\_selects$breedname) &  
 grepl("spaniel", vet\_patient\_list\_2023\_selects$breedname) &  
 !grepl("mix", vet\_patient\_list\_2023\_selects$breedname),  
 "spaniel, cavalier king charles",  
 vet\_patient\_list\_2023\_selects$breedname)  
  
vet\_patient\_list\_2023\_selects$breedname <- ifelse(  
 grepl("teddy", vet\_patient\_list\_2023\_selects$breedname) &  
 !grepl("mix", vet\_patient\_list\_2023\_selects$breedname),  
 "shichon (teddy bear)",  
 vet\_patient\_list\_2023\_selects$breedname)  
  
vet\_patient\_list\_2023\_selects$breedname <- ifelse(  
 grepl("teddybear", vet\_patient\_list\_2023\_selects$breedname) &  
 !grepl("mix", vet\_patient\_list\_2023\_selects$breedname),  
 "shichon (teddy bear)",  
 vet\_patient\_list\_2023\_selects$breedname)  
  
vet\_patient\_list\_2023\_selects$breedname <- gsub("^golden$", "retriever, golden", vet\_patient\_list\_2023\_selects$breedname)

## Done with cleaning!

### Seeing as cleaning was done I needed to look at unique values, and sometimes unique values involving subcategories and create a data fram with top 10 values. I also used these during cleaning to help get a grasp of different values.

unique\_breeds <- vet\_patient\_list\_2023\_selects %>%   
 count(breedname)  
View(unique\_breeds)  
  
unique\_canine\_breeds <- vet\_patient\_list\_2023\_selects %>%  
 filter(speciesname == "canine") %>%  
 count(breedname)  
View(unique\_canine\_breeds)  
  
unique\_feline\_breeds <- vet\_patient\_list\_2023\_selects %>%  
 filter(speciesname == "feline") %>%  
 count(breedname)  
View(unique\_feline\_breeds)  
  
unique\_sex <- vet\_patient\_list\_2023\_selects %>%   
 count(sex)  
View(unique\_sex)  
  
unique\_names <- vet\_patient\_list\_2023\_selects %>%   
 count(name)  
View(unique\_names)  
  
unique\_speciesname <- vet\_patient\_list\_2023\_selects %>%   
 count(speciesname)  
View(unique\_speciesname)  
  
top10\_speciesname <- unique\_speciesname[order(-unique\_speciesname$n), ][1:10, ]  
top10\_names <- unique\_names[order(-unique\_names$n), ][1:10, ]  
top10\_feline\_breeds <- unique\_feline\_breeds[order(-unique\_feline\_breeds$n), ][1:10, ]  
top10\_canine\_breeds <- unique\_canine\_breeds[order(-unique\_canine\_breeds$n), ][1:10, ]

## Exporting data

### All finished data I put in a folder called “project\_results”. Decided to use excel because I can save 4 different data frams in separate sheets. I also made a few basic bar graphs just to have a few visuals. Lastly I wanted to export the entire cleaned data frame I was working with as well just to have that on hand.

wb <- createWorkbook()  
  
addWorksheet(wb, "Top 10 Species Name")  
writeData(wb, sheet = "Top 10 Species Name", top10\_speciesname)  
  
addWorksheet(wb, "Top 10 Names")  
writeData(wb, sheet = "Top 10 Names", top10\_names)  
  
addWorksheet(wb, "Top 10 Feline Breeds")  
writeData(wb, sheet = "Top 10 Feline Breeds", top10\_feline\_breeds)  
  
addWorksheet(wb, "Top 10 Canine Breeds")  
writeData(wb, sheet = "Top 10 Canine Breeds", top10\_canine\_breeds)  
  
saveWorkbook(wb, file.path("project\_results", "top10\_breeds\_names\_species.xlsx"))  
  
speciesname\_plot <- ggplot(top10\_speciesname, aes(x = reorder(speciesname, -n), y = n)) +  
 geom\_bar(stat = "identity", fill = "blue") +  
 labs(title = "Top 10 Species Names", x = "Species Name", y = "Count") +  
 theme(axis.text.x = element\_text(angle = 45, hjust = 1))  
  
names\_plot <- ggplot(top10\_names, aes(x = reorder(name, -n), y = n)) +  
 geom\_bar(stat = "identity", fill = "green") +  
 labs(title = "Top 10 Names", x = "Name", y = "Count") +  
 theme(axis.text.x = element\_text(angle = 45, hjust = 1))  
  
feline\_breeds\_plot <- ggplot(top10\_feline\_breeds, aes(x = reorder(breedname, -n), y = n)) +  
 geom\_bar(stat = "identity", fill = "purple") +  
 labs(title = "Top 10 Feline Breeds", x = "Feline Breed", y = "Count") +  
 theme(axis.text.x = element\_text(angle = 45, hjust = 1))  
  
canine\_breeds\_plot <- ggplot(top10\_canine\_breeds, aes(x = reorder(breedname, -n), y = n)) +  
 geom\_bar(stat = "identity", fill = "orange") +  
 labs(title = "Top 10 Canine Breeds", x = "Canine Breed", y = "Count") +  
 theme(axis.text.x = element\_text(angle = 45, hjust = 1))  
  
ggsave(filename = file.path("project\_results", "top10\_speciesname\_plot.png"), plot = speciesname\_plot, width = 10, height = 6, dpi = 300)  
ggsave(filename = file.path("project\_results", "top10\_names\_plot.png"), plot = names\_plot, width = 10, height = 6, dpi = 300)  
ggsave(filename = file.path("project\_results", "top10\_feline\_breeds\_plot.png"), plot = feline\_breeds\_plot, width = 10, height = 6, dpi = 300)  
ggsave(filename = file.path("project\_results", "top10\_canine\_breeds\_plot.png"), plot = canine\_breeds\_plot, width = 10, height = 6, dpi = 300)  
  
write.csv(vet\_patient\_list\_2023\_selects, file = file.path("project\_results", "cleaned\_vet\_patient\_list\_2023.csv"), row.names = FALSE)