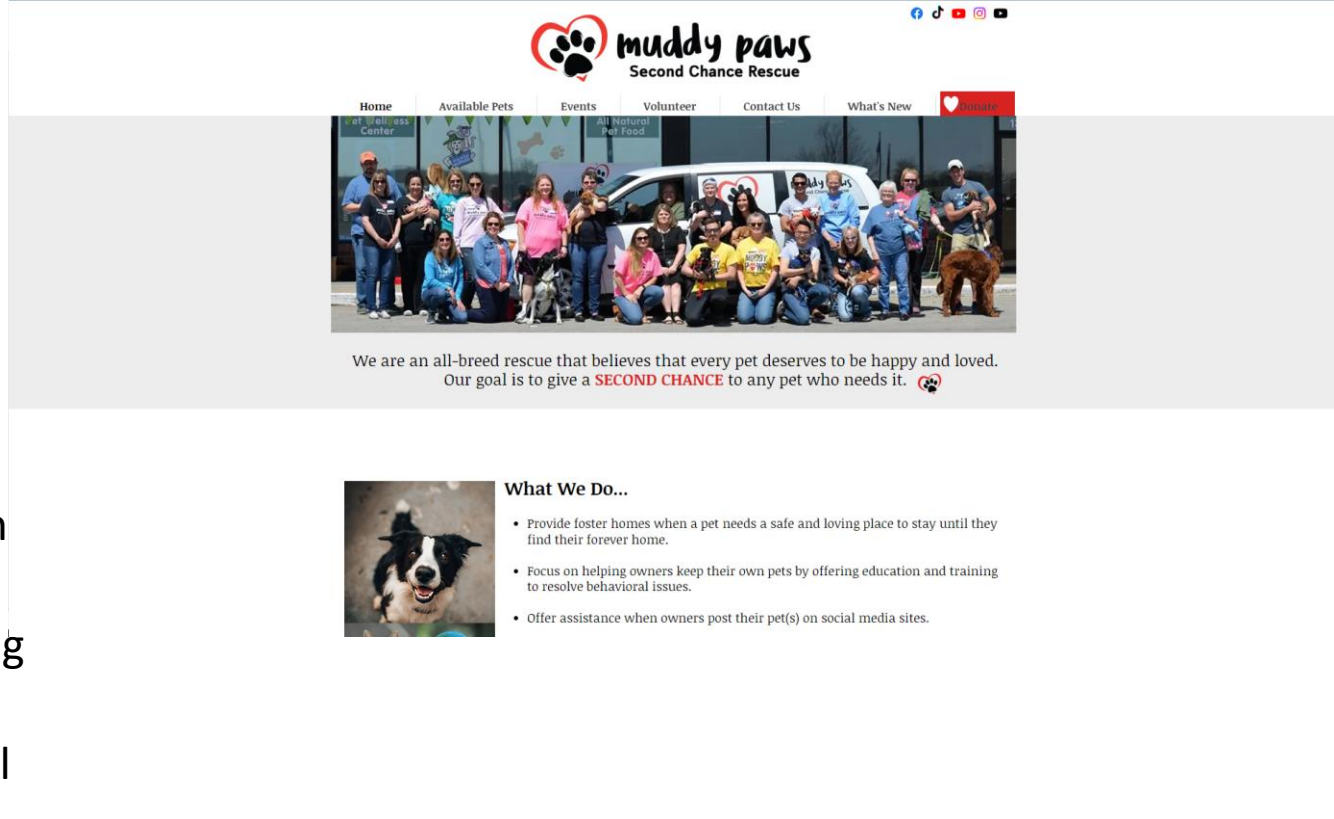
DOG LENGTH OF STAY IN RESCUE

Kaylar Fullington

BACKGROUND

- Muddy Paws Second Chance Rescue
  - Omaha, NE
  - Provide foster homes when a pet needs a safe and lovin place to stay until they find their forever home.
  - Focus on helping owners keep their own pets by offering education and training to resolve behavioral issues.
  - Offer assistance when owners post their pet(s) on social media sites.
  - Assist qualified owners with veterinary expenses needed to save the pet's life or restore its quality of life.
  - Maintain a pet food pantry for those owners that need help feeding their pets.



BUSINESS PROBLEM

- How long will animals stay at Muddy Paws?
- What characteristics most influence length of stay?

- I have chosen to focus on dogs only, to simplify analysis

THE DATA

- Muddy Paws uses the Petstablished Database software to manage their animal data
- I chose to use General Animal Records and Medical Records
- See the following slides for a description of the various data attributes

ATTRIBUTES (PG. 1)

- Petstablished ID
  - PS14507954
- Pet Name
  - Autumns Oktoberfest litter: Baum
  - Axel (Courtesy Post)
  - Bandit (St. Bernard) – Courtesy Post
- Length of Stay (Days)
  - 22
- Current Status
  - Adopted
  - Returned to Owner
  - Deceased
  - Boarding
  - Etc.

- Pet Type
  - Dog
  - (Others)
- Pet Age
  - Baby
  - Puppy
  - Young
  - Adult
  - Senior
  - (blanks)

ATTRIBUTES (PG. 2)

- Size
  - Small
  - Medium
  - Large
  - X-Large
  - (blanks)
- Shedding
  - No shedding
  - Sheds a little
  - Sheds a lot
  - None
  - Not Available
- Coat Length
  - Short
  - Medium
  - Long
  - (blanks)
- Temperament
  - Free text for the data entry person to write anything about the animal's temperament
- Breed Type
  - Mixed
  - Purebreed
  - (blanks)
- Adoption Fee
  - 500
  - 75 for both
- Gender
  - Male
  - Female

ATTRIBUTES (PG. 3)

- Age in Years
  - 1 years, 8 months
- Is Mix
  - Yes
  - No
  - N/A
- Shots up to date?
  - YES
  - NO
  - Not Sure
- Spayed/Neutered?
  - YES
  - NO
  - Not Sure
- Hypoallergenic?
  - YES
  - NO
  - Not Sure
- Housebroken?
  - YES
  - NO
  - Not Sure
- Special Need?
  - YES
  - NO
  - Not Sure
- Gets along with Cats?
  - YES
  - NO
  - Not Sure

ATTRIBUTES (PG. 4)

- Gets along with Dogs?
  - YES
  - NO
  - Not Sure
- Gets along with Kids?
  - YES
  - NO
  - Not Sure
- Internal Notes
  - Free text field for data entry person

- Behavioral Tacking Notes
  - Free text field for data entry person
- Parents
  - Fauna PS447339
  - (contains parent name and petstablished id)
- Siblings
  - Disney Litter: Zazu…
  - Contains litter name and sibling names and petstablished ids

- Pet Primary Breed
  - Chihuahua
- Pet Secondary Breed
  - Jack Russell Terrier
- Record Type
  - Diagnosis
  - Medications
  - Surgery/Procedures
- Subtype
  - Various procedure names
- Description
  - Free text describing procedure

METHODS

- Data Cleaning
- Exploratory Data Analysis
- Data Preparation

DATA CLEANING

- Iterative Process
- Format cleaning
  - Column names
    - Replace spaces with _
    - Remove special characters such as /
    - Change to lowercase
  - New Columns
    - Courtesy and Cross Post columns
    - Sentiment Score column
    - Age in Months column
    - Had Medical column
    - Litter column
- Limited Pet Type column to "dog"
- Removed outlier lengths of stay

Back

COURTESY AND CROSS POSTS

- Courtesy Post
  - The dog is still in the original owner's home, but Muddy Paws uses their platform to spread the word about the dog needing a new home
  - MP's experience with the animal is limited
- Cross Post
  - The dog is at another organization, but they want to increase the reach because the dog may be having trouble getting adopted
  - MP's experience with the animal is limited

Back

SENTIMENT SCORE

- Performed text analysis on the Temperament field and assigned a score to the dog in a new column
- There was likely not enough data here to make a functional difference, but if they are more thorough with data entry in the future this could come in handy

Back

AGE IN MONTHS

- Age in Years column contained values such as:
  - 1 year, 8 months
- This text is not useful for analysis
- Calculated the age in months (in this case, 20) and assigned it to a new column as a continuous variable

Back

HAD MEDICAL

- If the medical records contained at least one entry for a dog, this column was marked ‘yes’
- This excludes spays, neuters, and microchip implantations, which are all standard for Muddy Paws

Back

LITTER

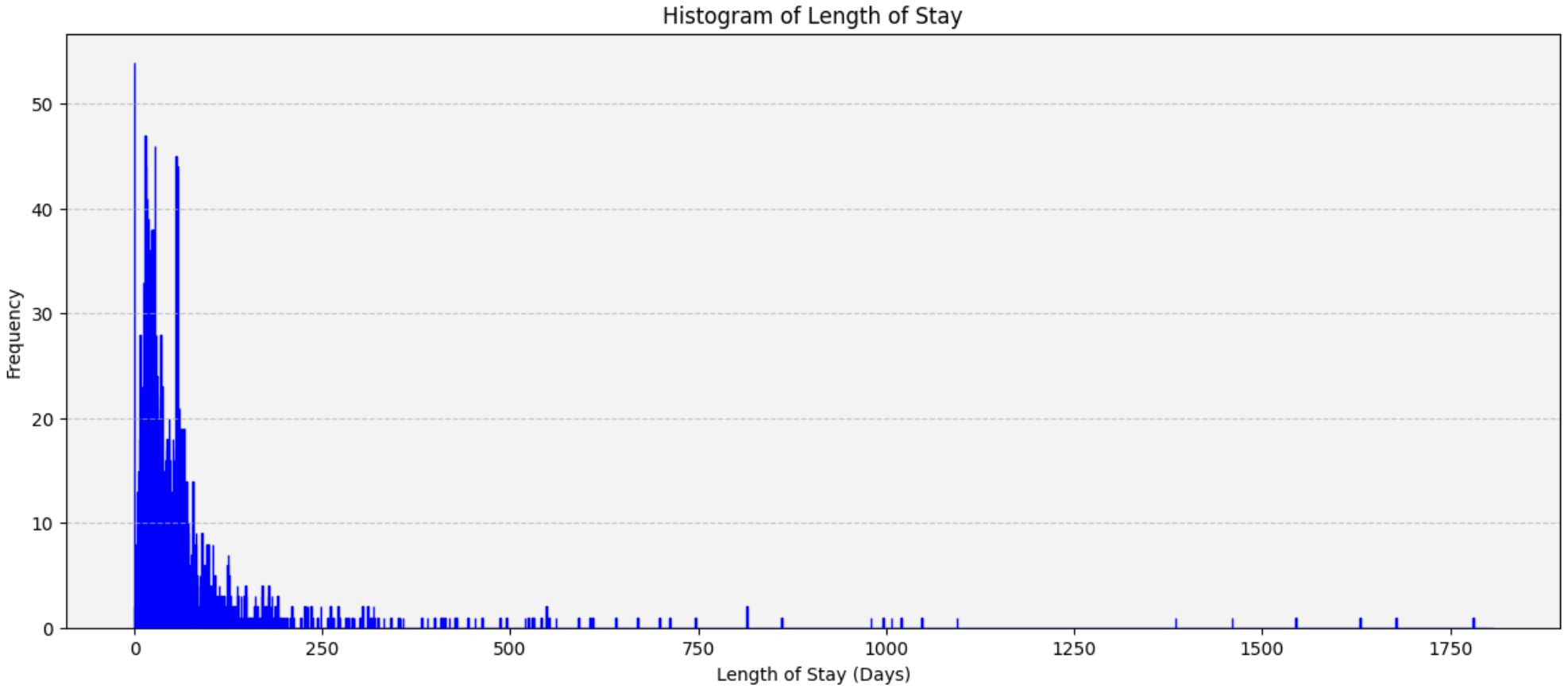
- If the Pet Name column contained references to a litter, the Litter column contained ‘yes’. Otherwise, it was ‘no’

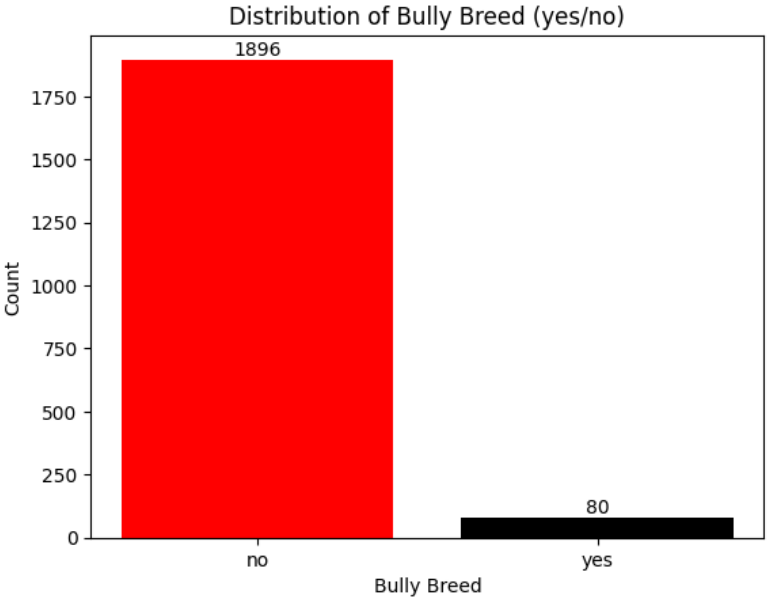
Back

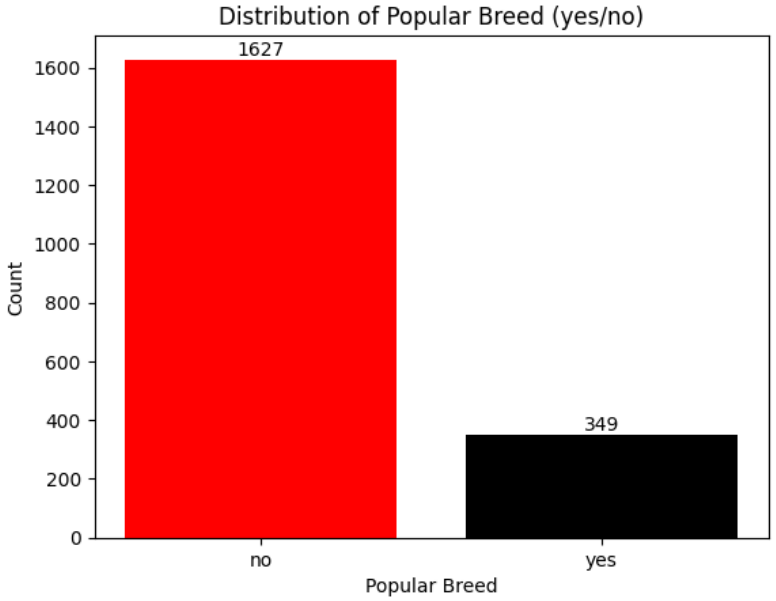
EXPLORATORY DATA ANALYSIS

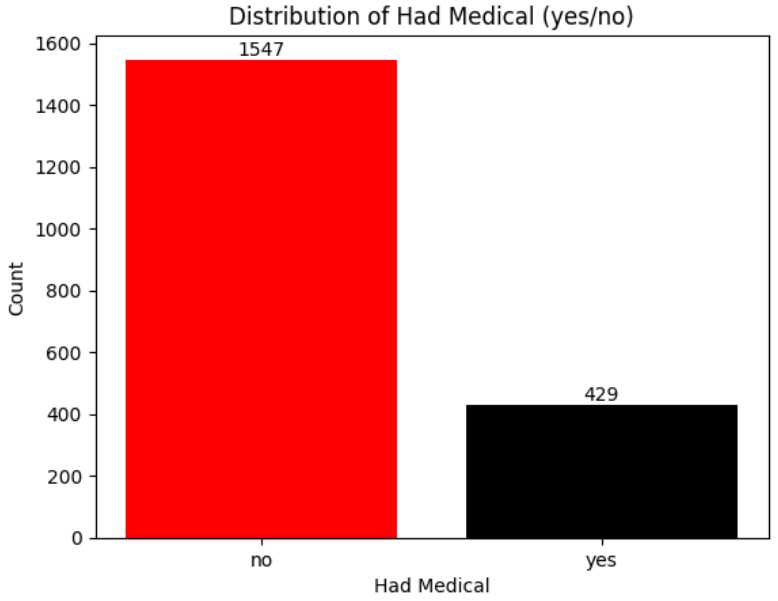
- This EDA was done on the cleaned data and not the original data
- Based on this EDA, feature engineering may be performed
- Length of Stay Distribution
  - The length of stay data was heavily right-skewed
  - This will be important in future iterations of this project
- Categorical Variable Bar Charts

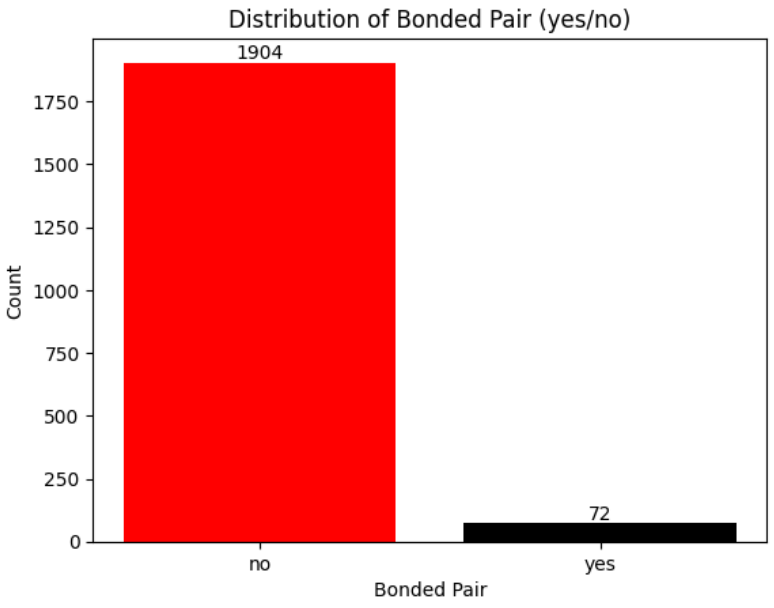
Back

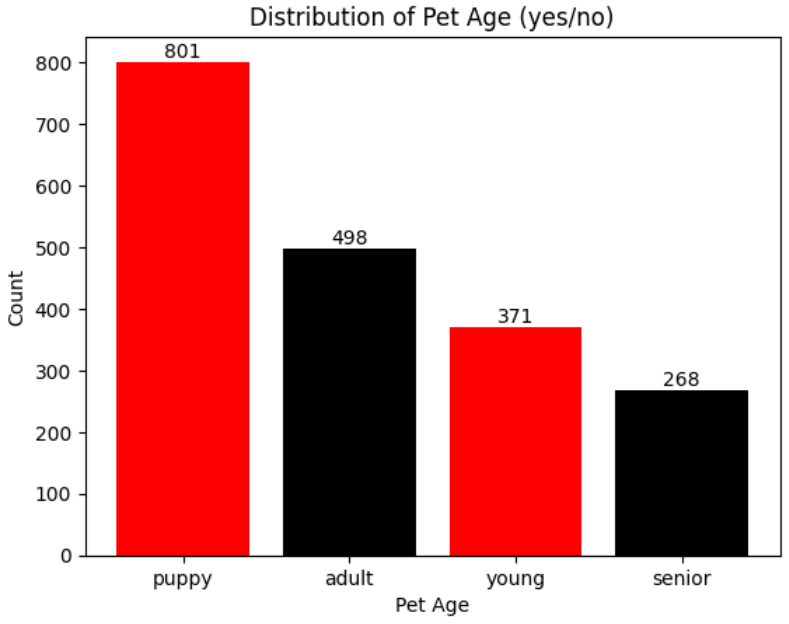

Histogram of Length of Stay
Frequency
0
10
20
30
40
50
0
250
500
750
1000
1250
1500
1750
Length of Stay (Days)

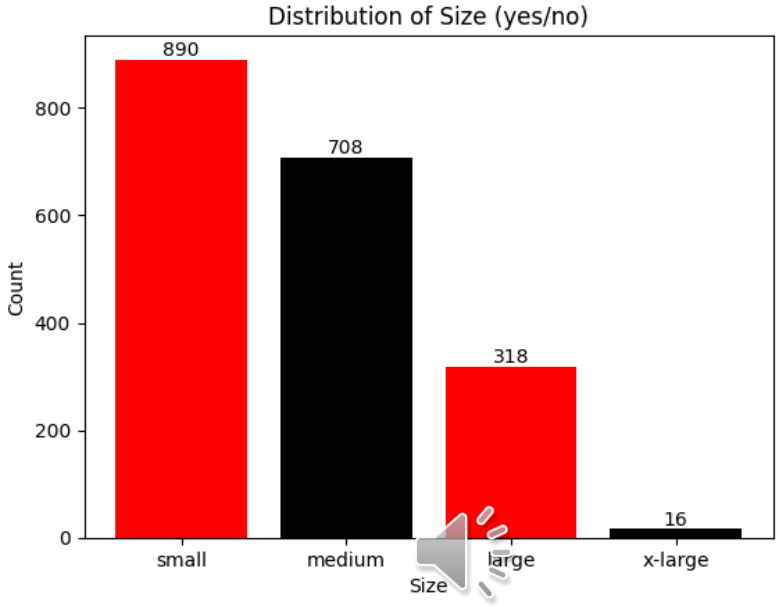

Distribution of Bully Breed (yes/no)
Count
1896
80
no
yes
Bully Breed

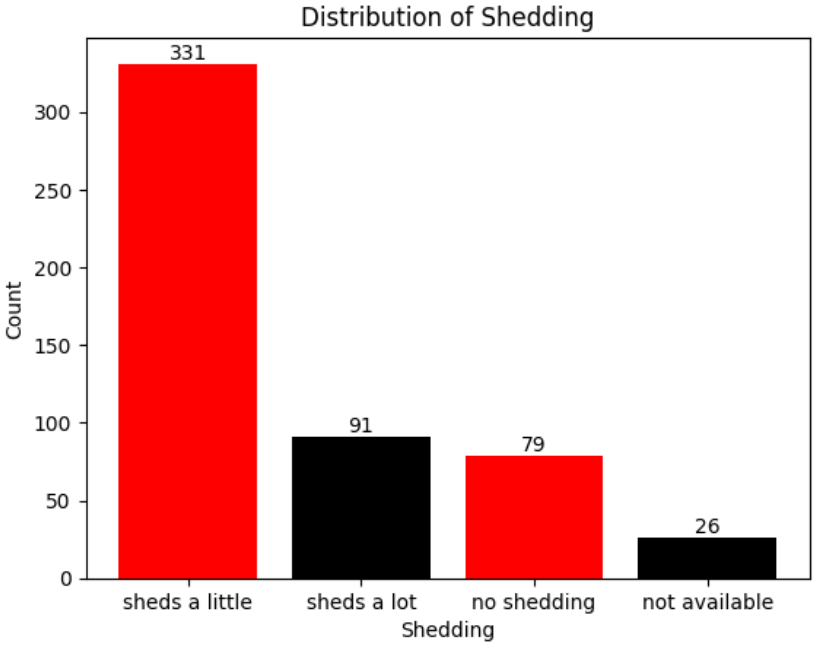

Distribution of Popular Breed (yes/no)
Count
1627
349
no
yes
Popular Breed

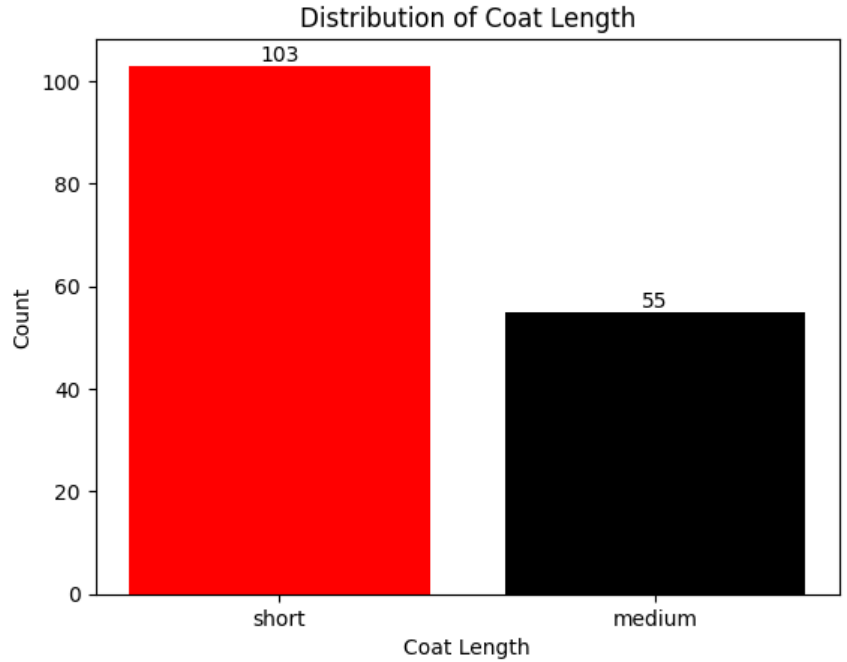

Distribution of Had Medical (yes/no)
Count
1547
429
no
yes
Had Medical

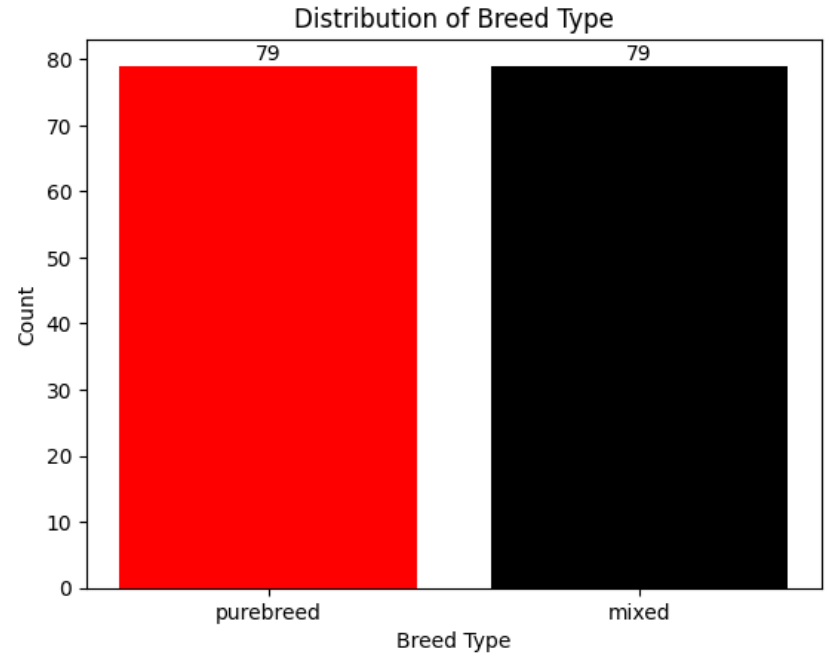

Distribution of Bonded Pair (yes/no)
Count
1904
72
no
yes
Bonded Pair

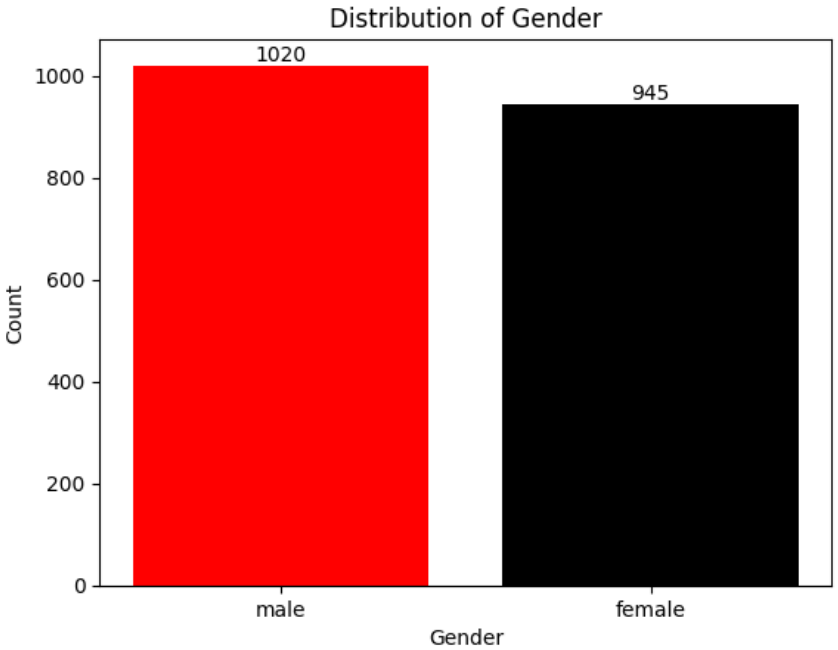

Distribution of Pet Age (yes/no)
Count
801
498
371
268
puppy
adult
young
senior
Pet Age

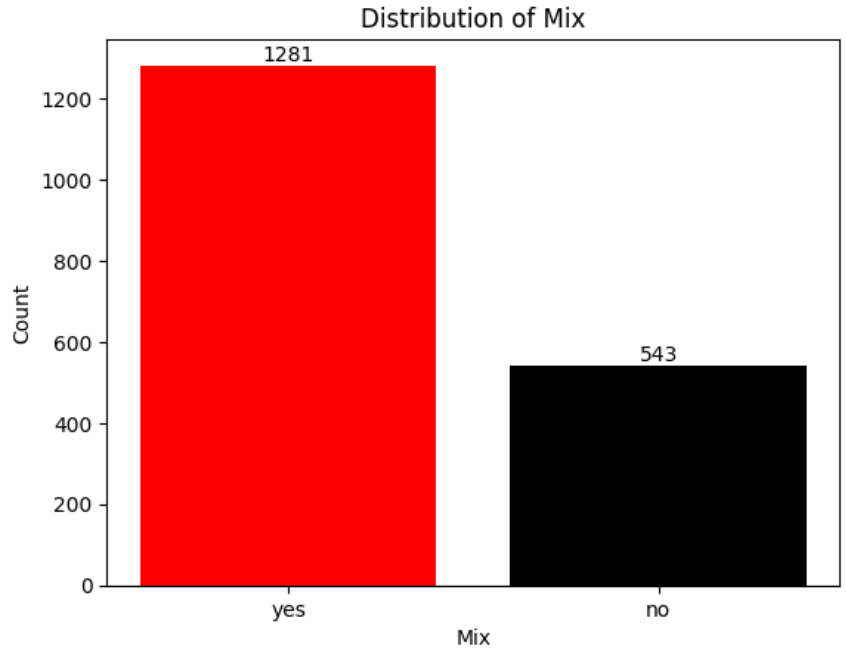

Distribution of Size (yes/no)
Count
890
708
318
16
small
medium
large
x-large
Size

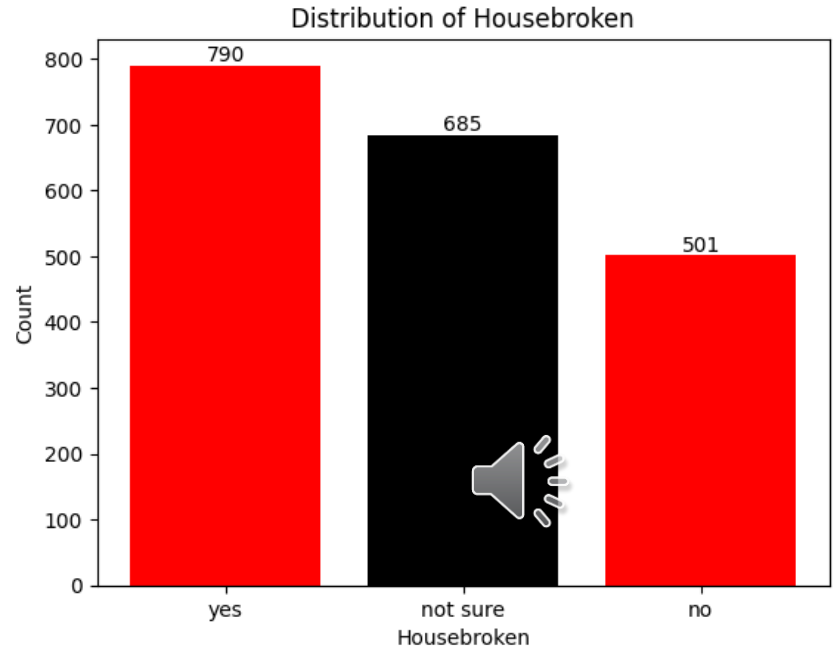

Distribution of Shedding
331
91
79
26
Count
sheds a little
sheds a lot
no shedding
not available
Shedding

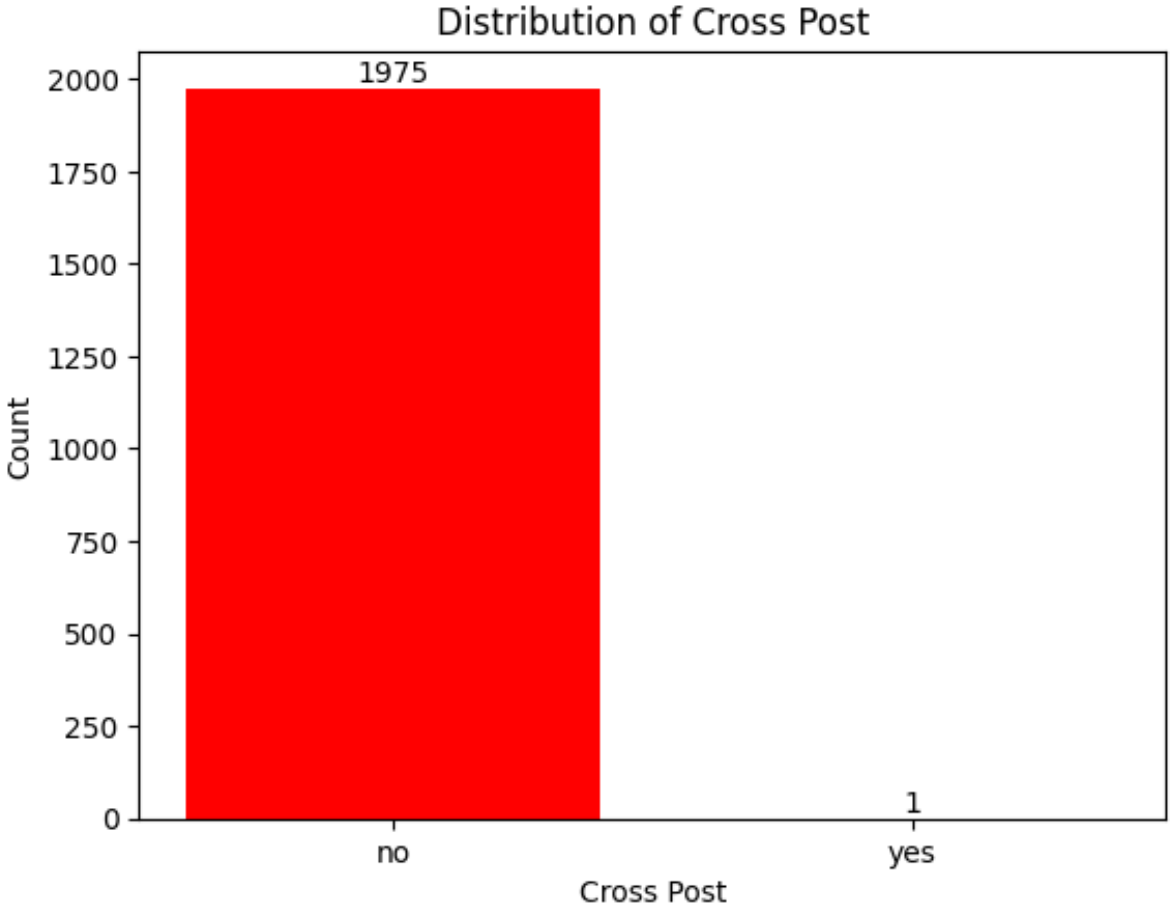

Distribution of Coat Length
103
55
Count
short
medium
Coat Length

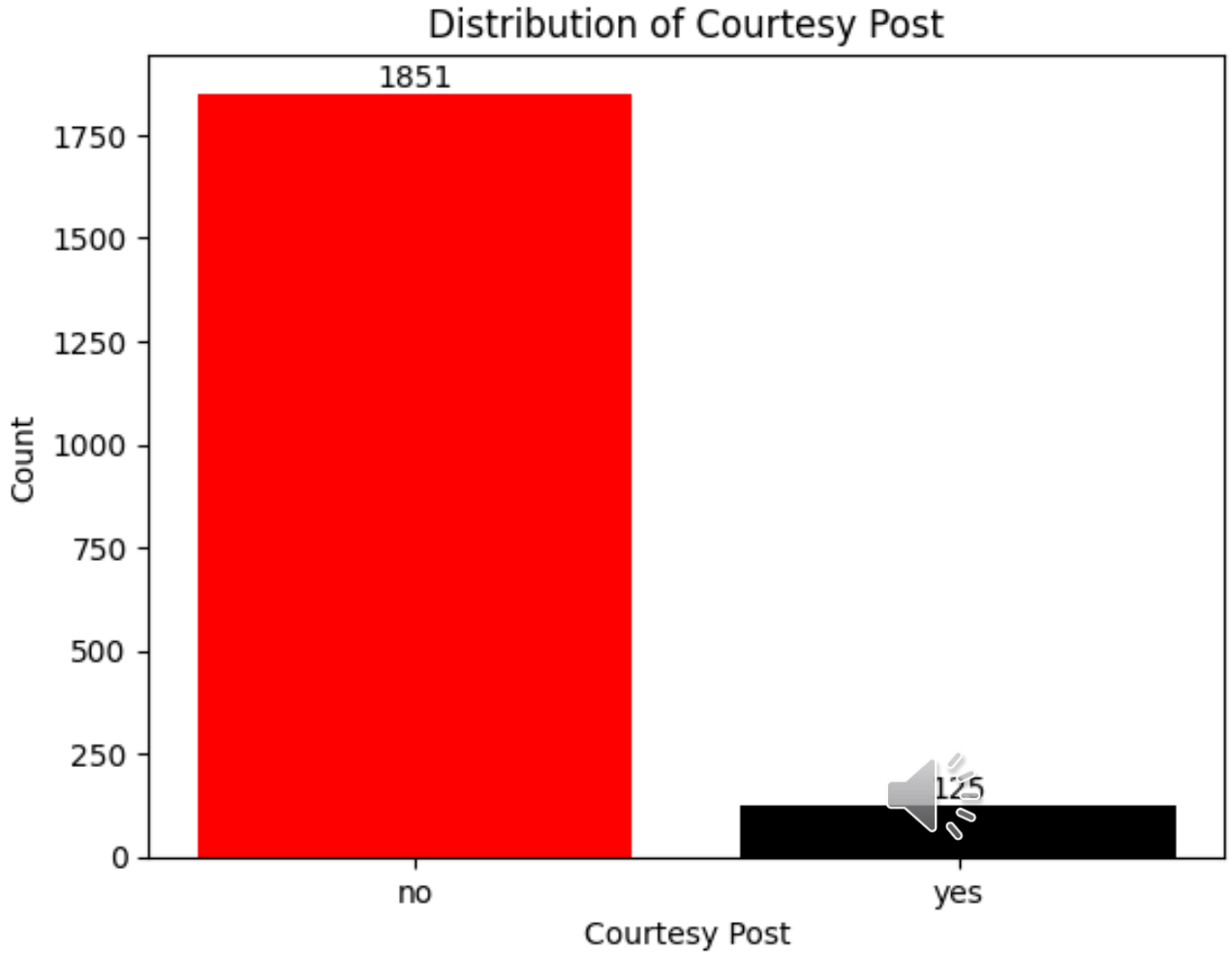

Distribution of Breed Type
79
79
Count
purebreed
mixed
Breed Type

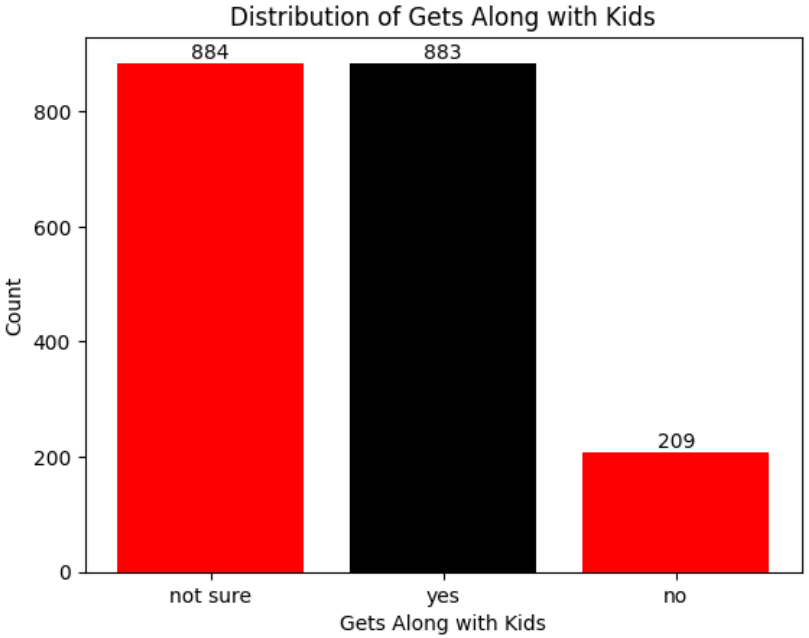

Distribution of Gender
1020
945
Count
male
female
Gender

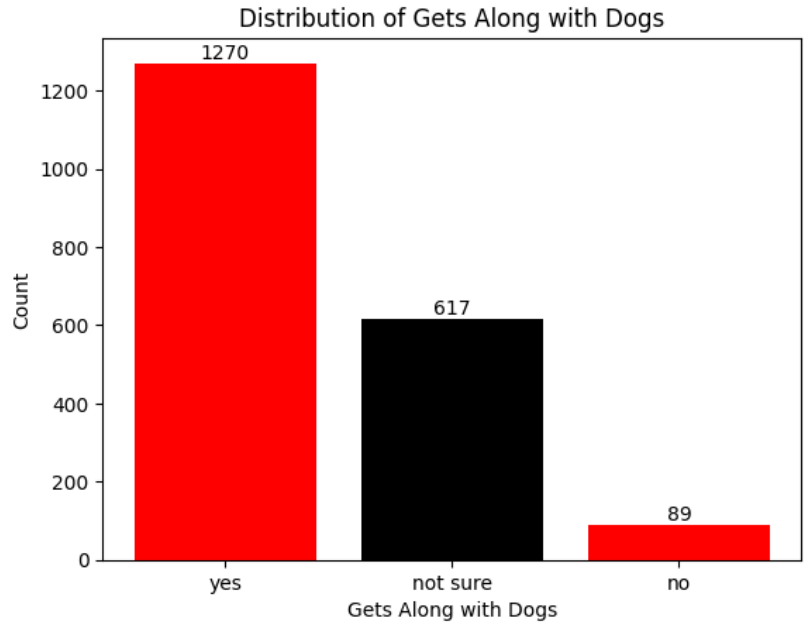

Distribution of Mix
1281
543
Count
yes
no
Mix

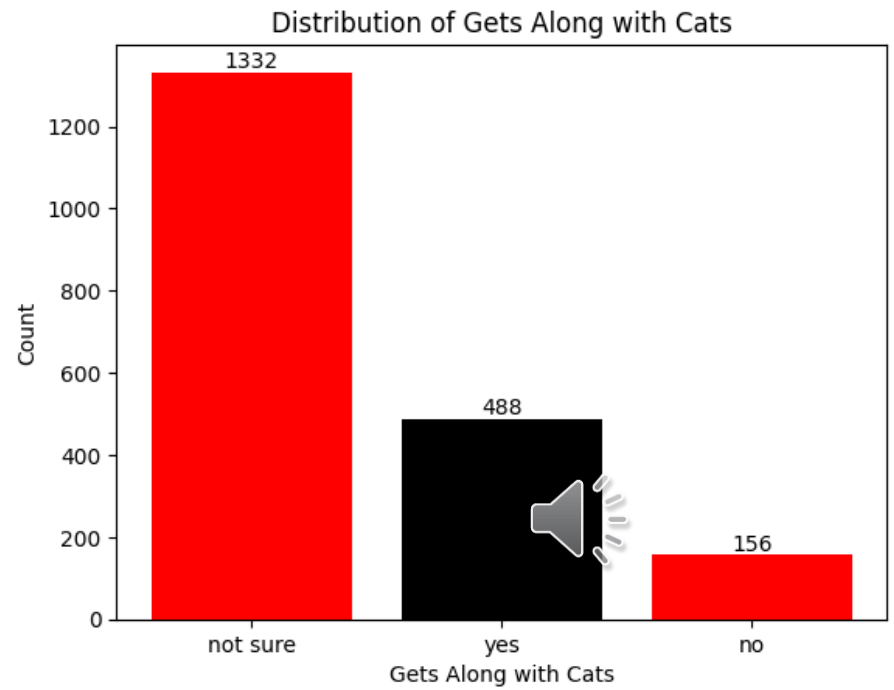

Distribution of Housebroken
790
685
501
Count
yes
not sure
no
Housebroken

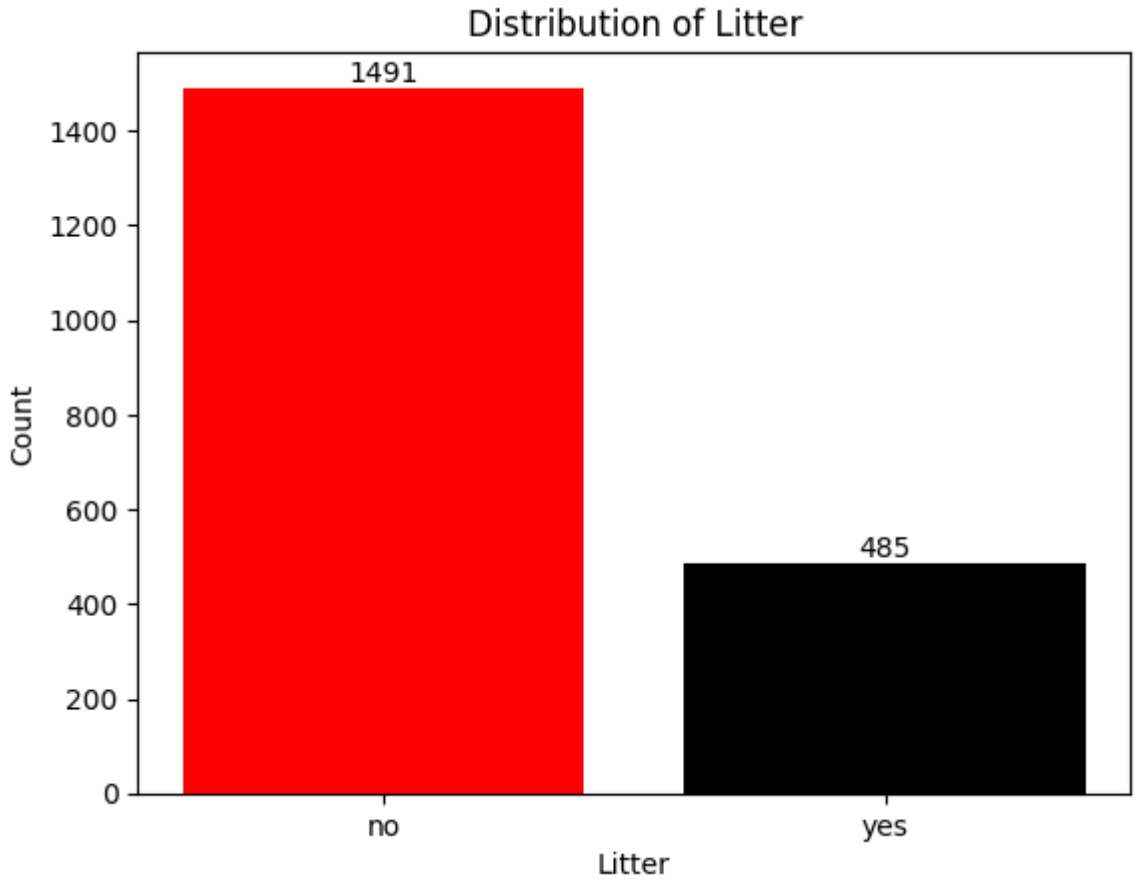

Distribution of Cross Post
2000
1750
1500
1250
1000
750
500
250
0
Count
1975
1
no
yes
Cross Post


Distribution of Courtesy Post
1750
1500
1250
1000
750
500
250
0
Count
1851
125
no
yes
Courtesy Post


Distribution of Gets Along with Kids
884
883
209
800
600
400
200
0
Count
not sure
yes
no
Gets Along with Kids


Distribution of Gets Along with Dogs
1270
617
89
1200
1000
800
600
400
200
0
Count
yes
not sure
no
Gets Along with Dogs


Distribution of Gets Along with Cats
1332
488
156
1200
1000
800
600
400
200
0
Count
not sure
yes
no
Gets Along with Cats


Distribution of Litter
1491
485
1400
1200
1000
800
600
400
200
0
Count
no
yes
Litter

DATA PREPARATION

- Determined that random forest classifier would be most effective for predicting length of stay
- This model requires categories rather than continuous variables
- Utilized Kmeans clustering to determine how many and what categories to split length of stay into
  - 10 was the best, with the highest accuracy of 64% and the smallest ranges in the categories

Back

ASSUMPTIONS

- For the purposes of analysis and modeling, I assumed that the data entered in Petstablished was accurate to the best of the data entry person's ability. I also elected to focus only on dogs to simplify the analysis. I also selected only dogs that were in the status of adopted, and assumed that they were not returned at any points.

LIMITATIONS

- The starting data set was limited to roughly 3000 rows for dogs, so the sample size I used for this analysis was relatively small. The rescue is a local organization and they have only been running for a handful of years. I selected the data from their founding all the way until November 30th , 2023. Thus, some dogs that are still in foster care were eliminated from the analysis because they have not been adopted at the cut off for the data set

CHALLENGES

- Cleaning the data for this project was challenging. The Pet Name column gave me trouble, because it contained valuable information that was not otherwise captured in an inconsistent format. For example, a Courtesy Post (within the organization's context) is an adoption listing where the dog is still in the original owner's home. There are also Cross Posts (listings where the dog is in the care of another organization). However, there are no separate columns for these crucial distinctions. Instead, they listed them in the Pet Name column. For example: "Abbey (Cross Post)" So, when cleaning the data, I had to use the Pet Name column to create new columns for Cross Posts and Courtesy Posts.
- Further challenges arose when testing different types of models on the data set. Most of the models had relatively low performance and required a lot of feature engineering and other methods to increase the accuracy without overfitting.

FUTURE USES AND ADDITIONAL APPLICATIONS

- In the future, the work done in this project could be used as the basis for work with additional local organizations and not just Muddy Paws. I also would like to expand the model to include the impact of social media promotion eventually, though that may be best with much more data than a single organization can provide.

RECOMMENDATIONS

- Improve data entry procedures
- Provide a list of key words for data entry people to put into the free text columns (such as the Temperament field), to make for easier sentiment analysis
- Add columns for Courtesy and Cross Posts
- Continue collecting more data

IMPLEMENTATION PLAN

- Once the model has an accuracy score that I am happier with, I will start feeding it more data and use it to start predicting for new intakes. I plan to keep track of the model's predictions and check back in with the same dogs periodically to determine how closely the predictions match the actual outcomes. I will run it locally for now, as the organization does not have the infrastructure to manage it the way a bigger company would.

ETHICAL ASSESSMENT

- The president wants these predictions to help her determine where to spend resources such as money and foster homes. The nature of rescue organizations means that their animals are often homeless and/or in need of expensive medical care. My concern is that the model could be used to deny care to animals who may otherwise have received it through the rescue. Inaccurate predictions could also lead to heartbreak. Animals that are identified as long stayers may be passed over when in fact they may have been quickly adopted. I do not believe the president will use the model in this way, but it is a risk to consider.

CONCLUSION

- The highest accuracy obtained during this project was 64%
- More feature engineering needs to be done, which will hopefully increase the accuracy score
  - Removing superfluous dummy variables (ex: gets_along_with_cats_yes and gets_along_with_cats_no)
  - Removing low-impact variables
- Some of the characteristics that have the highest impact include:
  - Age
  - Gender
  - Getting along with kids
  - Whether the animal is part of a litter