Pet adoption speed prediction

What opportunities exist for animal shelters in Malaysia to leverage existing pet listing data to predict the speed at which a pet will be adopted and thus plan and allocate resources more effectively?

Final Report

SIMONA TSO

MARCH 2021

1. Introduction

1.1 Problem Statement

Millions of stray animals suffer on the streets or are euthanized in shelters every day around the world. There are a limited number of shelters and foster homes available. If we are able to predict the speed at which a pet will be adopted, adoption centres and foster homes can distribute resources more efficiently and implement targeted strategies accordingly.

PetFinder.my is Malaysia's leading animal welfare platform since 2008. It has accumulated a database of over 150,000 animals. We will use this data to answer the following questions:

- What is the best predictive model to predict whether or not a pet will be adopted within 100 days?
- Which features play an important role in adoptability of a pet?
- What can adoption shelters do to increase the chance for pets in shelters to be adopted?

2. Data and pre-processing

2.2 Data Overview

4 datasets provided by https://www.kaggle.com/c/petfinder-adoption-prediction/overview was analyzed.

2.1.1 Main Data

The main data consist of 14993 rows of pet adoption data. This dataset contained adoption information for dogs and cats. It included features about name, age, breed, gender, color, size, fur length, vaccination status, deworm status, sterilization status, health, quantity, fee, location, rescuer ID, video amount, photo amount, description and the target variable, adoption speed.

2.1.2 Breed label data

This dataset consist of 307 rows that decodes the breed ID abbreviation used in the main dataset into its associated breed name.

2.1.3 Color label data

This dataset consist of 7 rows that decodes the color ID abbreviation used in the main dataset into its associated color name.

2.1.4 State label data

This dataset consist of 15 rows that decodes the state ID abbreviated used in the main dataset into its associated state name.

2.2 Data cleaning, processing and feature engineering

2.2.1 Main Data

The main data contained 14993 observations in total. 8132 of which were dog adoption data and 6861 was cat adoption. I first split the dataset into dogs and cats as these are different animals and should be treated separately.

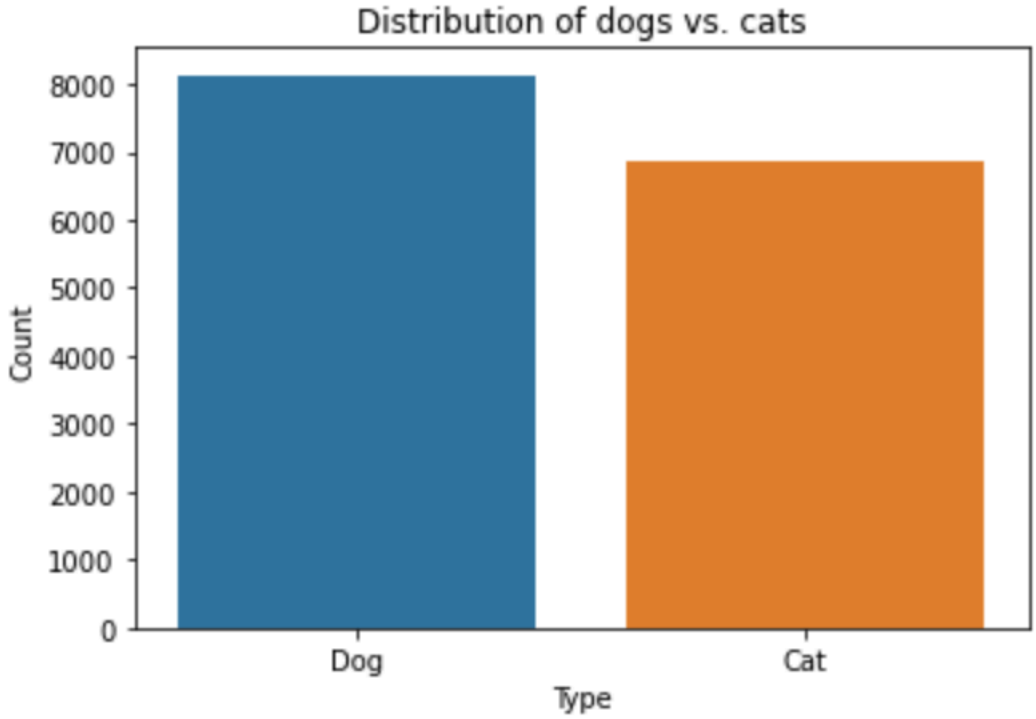


Fig 1. Distribution of the types of animals found in the pet adoption dataset.

I performed data processing and cleaning separately for both datasets, but the steps are the same. First, I removed all outliers (3 standard deviations below and above the mean) for the continuous features Age and Fee. Then, I performed one-hot-encoding for the remaining categorical features and re-binned some of the features. I also engineered several new features which will be discussed in-depth later on.

For the 'Description' feature, I used TF-IDF vectors to extract the top keywords. TF-IDF is a statistical measure that evaluates how relevant a word is to a text in a collection of texts, by multiplying the number of times a word appears in the text and the inverse document frequency of the word across a set of documents. I used TF-IDF to retrieve the top 100 keywords, then I manually combed through the list and selected 20 keywords that I thought was most relevant and created a new feature for each keyword in the dataset. For each observation, if the description contained the keyword, a 1 would be recorded into the keyword feature and 0 otherwise.

Lastly, the target variable, Adoption Speed, contained 5 classes:

- Same day adoption
- Adoption within the 1st week
- Adoption within the 1st month
- Adoption between 2-3 months
- No adoption after 100 days

I regrouped these 5 classes into two bins in order to create a suitable target variable for binary classification. The two classes I defined are:

- 0: Adopted within 100 days
- 1: No adoption after 100 days

2.2.1.1 Predictor Variables

Below is an overview of the features in the dog adoption and cat adoption dataset after data cleaning, processing and feature engineering.

Name: The name of the pet on file.

From this feature, 3 new features were created to capture information regarding pet name. The 'Name' feature was eventually dropped.

HasName: Whether or not the pet has a name on file.

This feature was engineered from the original 'Name' feature in the dataset. Each observation was classified as either 'unnamed' (0) or 'named' (1), depending on whether the 'Name' feature was filled in or left blank. Fig 2.1 shows the distribution of 'HasName' in the dog adoption dataset and fig 2.2 shows the same distribution in the cat adoption dataset. This feature was one-hot encoded afterwards, creating new features 'HasName_0' and 'HasName_1' respectively.

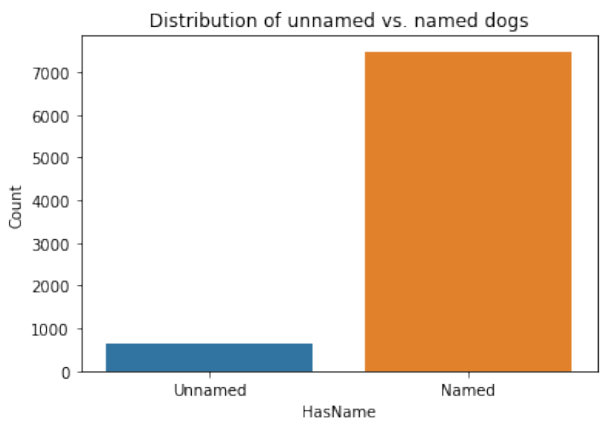


Fig 2.1 Distribution of unnamed and named dogs

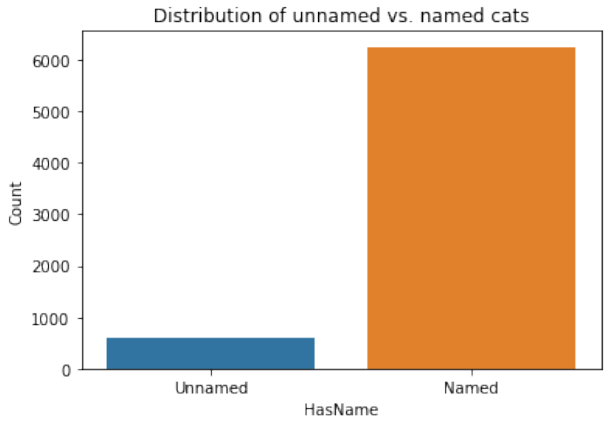


Fig 2.2 Distribution of unnamed and named cats

NameLen: The number of characters in the pet's name.

This feature was also engineered from the original 'Name' feature in the dataset. 'NameLen' depicts the total number of characters found in the 'Name' field for each observation. Fig 3.1 shows the distribution of 'NameLen' in the dog adoption dataset and fig 3.2 shows the same distribution in the cat adoption dataset.

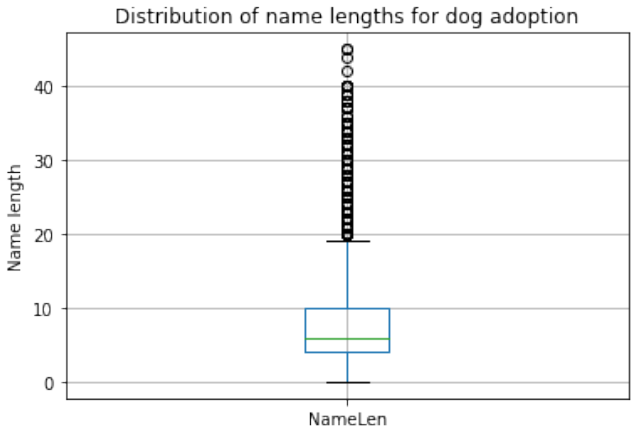


Fig 3.1 Box plot showing the distribution of dog name lengths

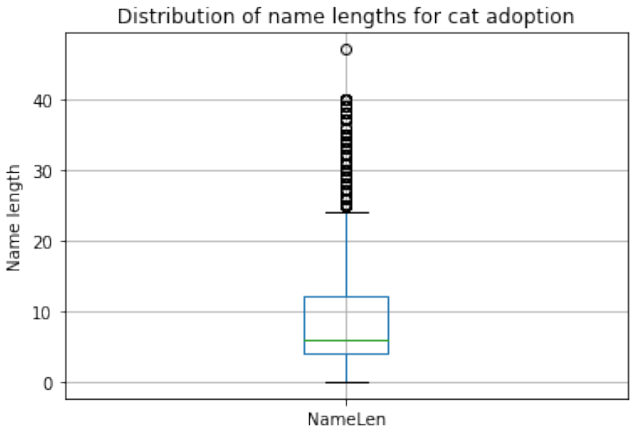


Fig 3.2 Box plot showing the distribution of cat name lengths

NameNumWords: The number of words in the pet's name.

This feature was also engineered from the original 'Name' feature in the dataset. 'NameNumWords' depicts the total number of words found in the 'Name' field for each observation. Fig 4.1 shows the distribution of 'NameNumWords' in the dog adoption dataset and fig 4.2 shows the same distribution in the cat adoption dataset.

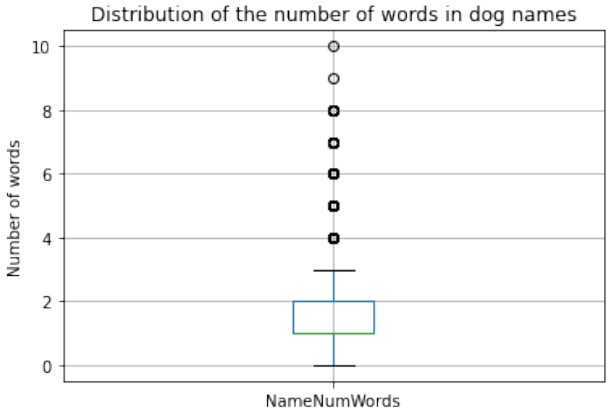


Fig 4.1 Box plot showing the distribution of dog name word count

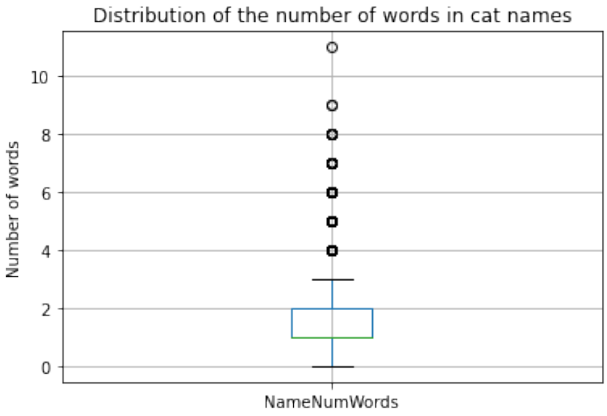


Fig 4.2 Box plot showing the distribution of cat name word count

Age: The age of the pet in months.

This feature was kept but outliers were removed. Based on this feature, a new feature 'ThreeMonths' was also engineered.

Age

For the ‘Age’ feature, outlier removal was performed by removing all records that lied 3 standard deviations above and below the mean. 210 outliers were removed in the dog adoption dataset and 168 outliers were removed in the cat adoption dataset. Fig 5.1 displays the final distribution for dog ages in months and fig 5.2 shows the final distribution for cat ages in months.

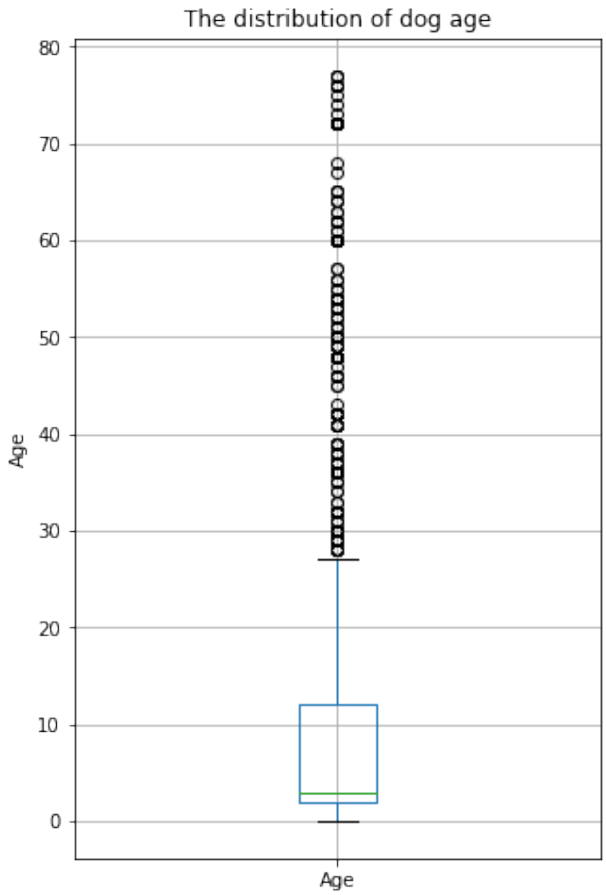


Fig 5.1 Box plot showing the distribution of dog age

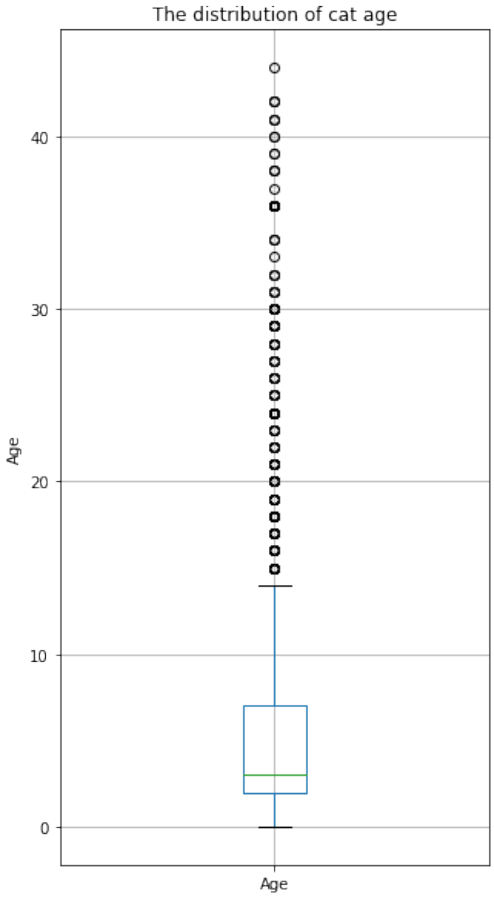


Fig 5.1 Box plot showing the distribution of cat age

ThreeMonths: Whether or not the pet is 3 months of age or younger.

From domain knowledge, I know that cats and dogs that are 3 months of age or younger are widely considered puppies and kittens and are likely to be adopted quicker. As a result, I created a new feature ‘ThreeMonths’ based on the ‘Age’ feature to store this information. Pets that are 3 months of age or younger are recorded as 1 and those that are older are recorded as 0. Fig 6.1 represents the distribution of ‘ThreeMonths’ for dogs and fig 6.2 shows the distribution for cats.

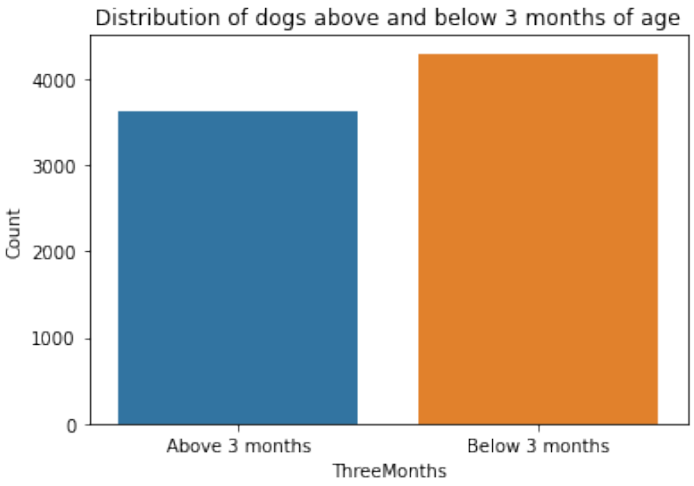


Fig 6.1 Distribution of dogs above and below 3 months old

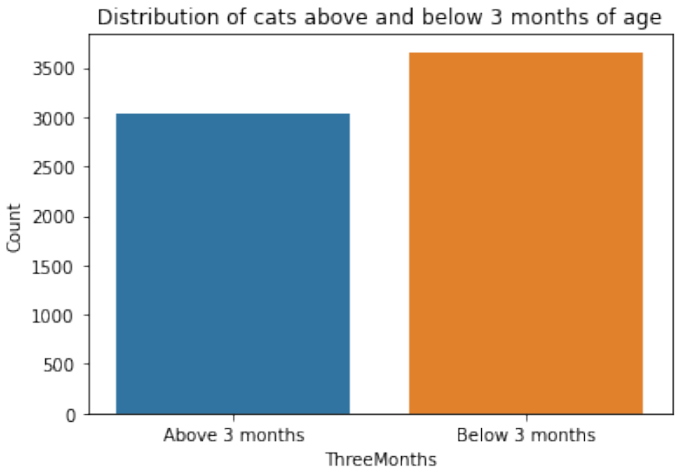


Fig 6.2 Distribution of cats above and below 3 months old

Breed1, Breed2: The primary and secondary breed for dog listings.

The original data contained two breed features: Breed1 and Breed2. Breed1 represents the pet's dominant breed and Breed2 is the pet's secondary breed. I noticed several problems with this data, below are the problematic scenarios:

- Breed1 was 0, whilst Breed2 wasn't
- Breed2 was the same as Breed1
- Breed1 was 307 (mixed breed), but Breed2 wasn't 0

The first problem appears to be an input error, I fixed this by filling in Breed1's value with Breed2s' and filling in Breed2's with 0. The second problem also appears to be an input error, I fixed this by changing Breed2's value to 0. Finally, I resolved the last issue by swapping Breed1 and Breed2's values to ensure that Breed1 always contained a known breed.

PureBreed: Whether the pet is purebred or not.

I used 'Breed1' and 'Breed2' to create a new feature 'PureBreed,' which determines whether a pet is of mixed or pure breed. For the dog adoption dataset, if 'Breed2' is equal to 0 (n/a) and 'Breed1' is not equal to 307 (mixed breed), then the observation is classified as pure breed (1) and mixed breed (0) otherwise. Fig 7.1 shows the distribution of mixed breed vs. pure breed dogs.

It was slightly more complicated for the cat adoption dataset since domestic breeds are not encoded as 307 in the dataset. From domain knowledge, I applied the below rule to determine whether or not a cat was purebred. If 'Breed2' is equal to 0 (n/a) and 'Breed1' is not equal to either 307 (mixed breed), 264 (domestic long hair), 265 (domestic medium hair) or 255 (domestic short hair), then the observation is classified

as pure breed (1) and mixed breed (0) otherwise. Fig 7.2 shows the distribution of mixed breed vs. pure breed cats.

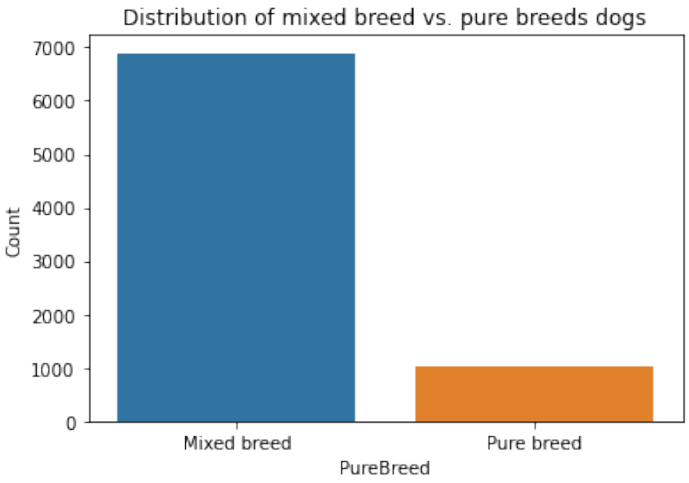


Fig 7.1 Distribution of mixed breed and pure breed dogs

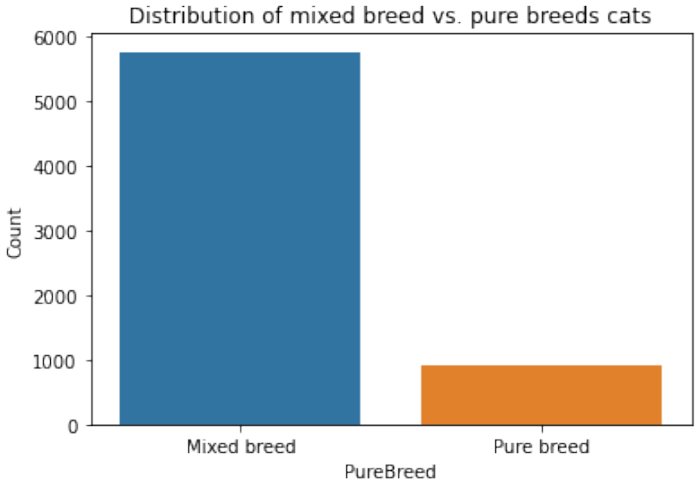


Fig 7.1 Distribution of mixed breed and pure breed cats

DomesticBreed: Whether the cat is of domestic breed or not.

I created an additional feature for cats: 'DomesticBreed'. This binary feature labels 'domestic short hair', 'domestic medium hair' and 'domestic long hair' cats as 1 and 0 otherwise. Fig 8.1 shows the distribution of domestic cats.

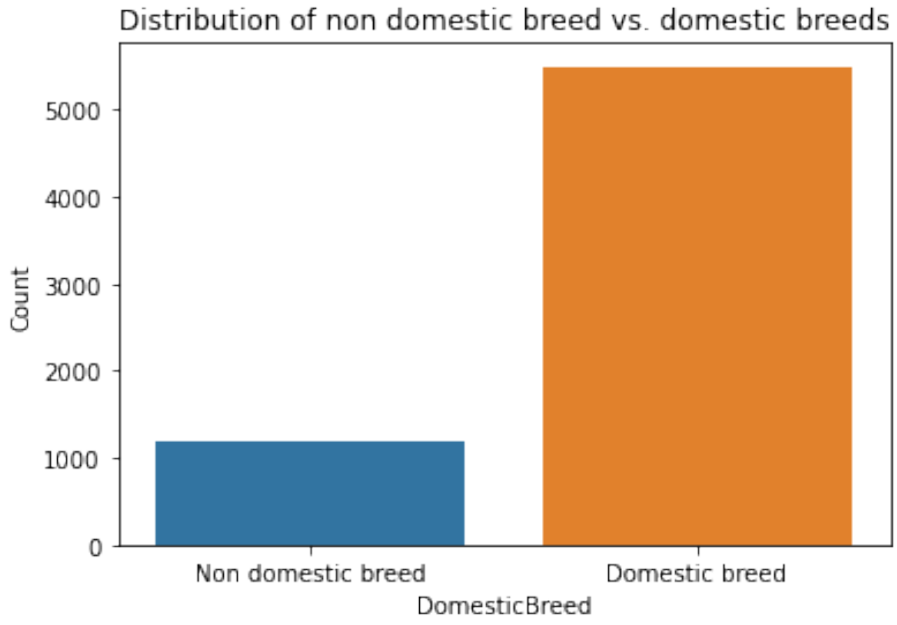


Fig 8.1 Distribution of non-domestic and domestic cats

Top 10 Breeds

Finally, I dropped the 'Breed2' column and only kept the top 10 most common breed types from 'Breed1' for both the dog and cat dataset. The remaining breeds were all grouped into 'Breed_999'. Next, I performed one hot encoding on 'Breed1' which resulted in a total of 12 one hot-encoded new features for each of the datasets. Fig 9.1 shows the new features generated for the dog adoption dataset and fig 9.2 shows the new features generate for the cat adoption dataset.

Feature	Description
Breed_20	Whether or not the dog is of breed 'Beagle'.
Breed_103	Whether or not the dog is of breed 'German Shepherd'.
Breed_109	Whether or not the dog is of breed 'Golden Retriever'.
Breed_141	Whether or not the dog is of breed 'Labrador Retriever'.
Breed_179	Whether or not the dog is of breed 'Poodle'.
Breed_189	Whether or not the dog is of breed 'Rottweiler'.
Breed_205	Whether or not the dog is of breed 'Shih Tzu'.
Breed_213	Whether or not the dog is of breed 'Spitz '.
Breed_218	Whether or not the dog is of breed 'Terrier'.
Breed_307	Whether or not the dog is of breed 'Mixed Breed'.
Breed_999	Whether or not the dog is of any of the unmentioned breeds above.

Fig 9.1 One hot encoded dog breed labels

Feature	Description
Breed_243	Whether or not the cat is of breed ‘American Shorthair’.
Breed_247	Whether or not the cat is of breed ‘Bengal’.
Breed_254	Whether or not the cat is of breed ‘Calico’.
Breed_264	Whether or not the cat is of breed ‘Domestic Long Hair’.
Breed_265	Whether or not the cat is of breed ‘Domestic Medium Hair’.
Breed_266	Whether or not the cat is of breed ‘Domestic Short Hair.
Breed_283	Whether or not the cat is of breed ‘Oriental Short Hair.
Breed_285	Whether or not the cat is of breed ‘Persian ‘.
Breed_292	Whether or not the cat is of breed ‘Siamese’.
Breed_299	Whether or not the cat is of breed ‘Tabby’.
Breed_999	Whether or not the cat is of any of the unmentioned breeds above.

Fig 9.2 One hot encoded cat breed labels

Gender: The gender of the pet (Male, Female or Mixed)

This feature depicts the pet’s gender, whether it was male, female or mixed gender. Mixed gender refers to a listing that contains multiple pets with different genders. Fig 10.1 shows the distribution of dog gender and fig 10.2 shows the distribution of cat gender.

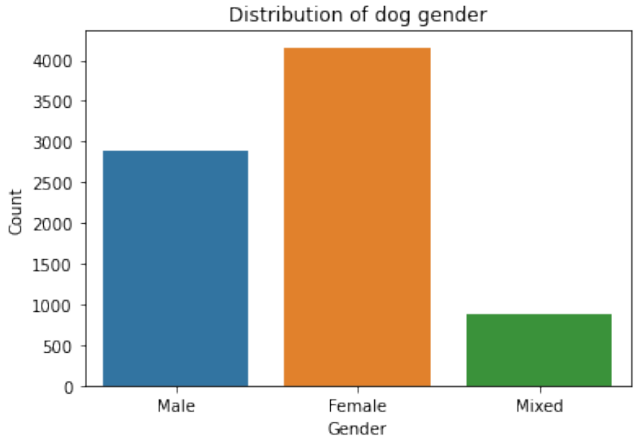


Fig 10.1 Distribution of dog genders

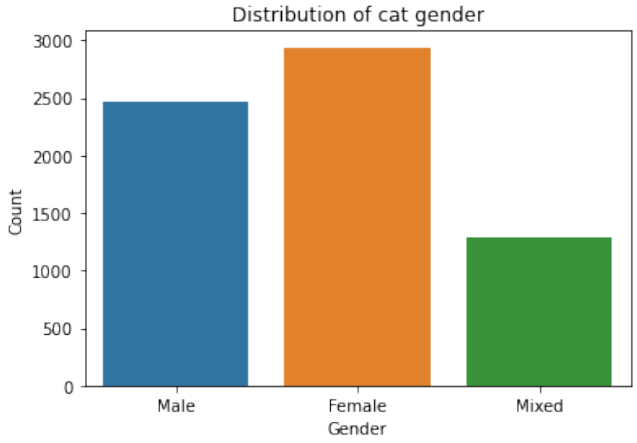


Fig 10.1 Distribution of cat genders

I performed one hot encoding on the gender feature which created 3 new columns: ‘Gender_1’, ‘Gender_2’ and ‘Gender_3’ representing male, female and mixed gender respectively.

Color1, Color2, Color3: The primary, secondary and tertiary color of the pet

The ’Color1’, ‘Color2’ and ‘Color3’ features depict the pet’s primary, secondary and tertiary fur color. Fig 11.1 shows the distribution of dog color and fig 11.2 shows the distribution of cat color.

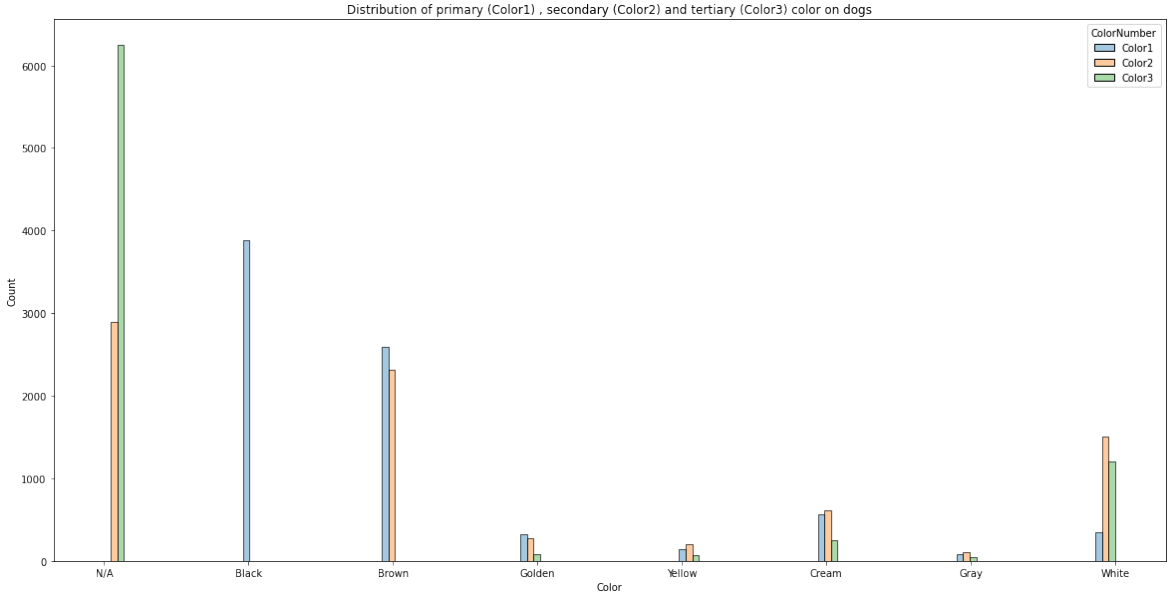


Fig 11.1 Distribution of fur colors for dogs

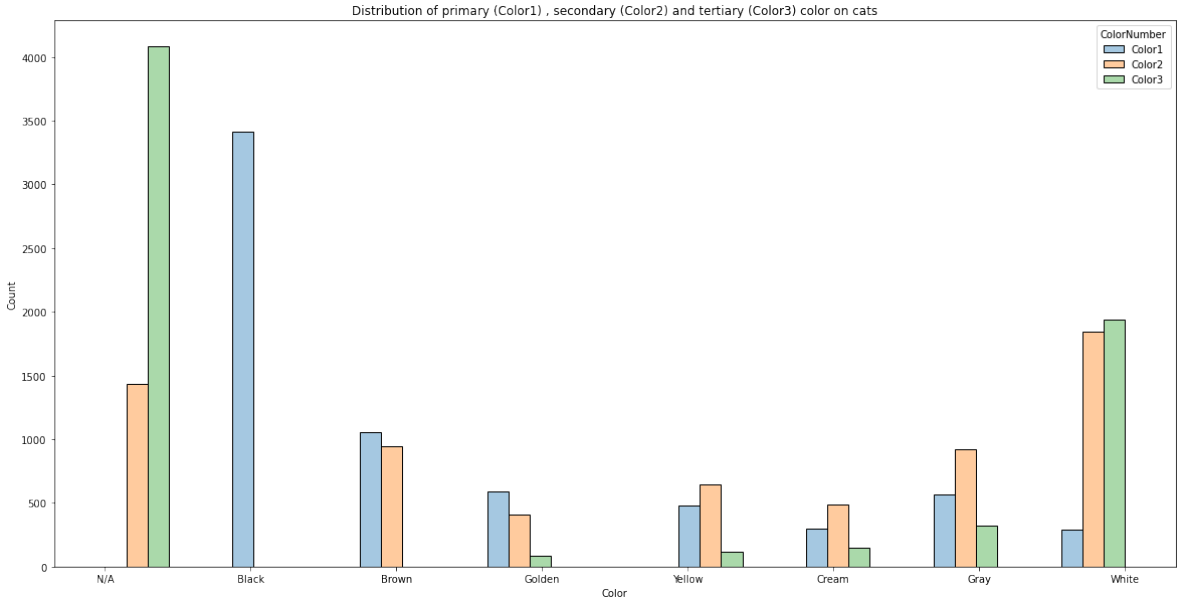


Fig 11.2 Distribution of fur colors for cats

NumColor: The number of colours apparent on a pet's fur coat (1, 2 or >= 3).

Based on 'Color1', 'Color2' and 'Color3', I engineered a new feature 'NumColor' which represents the number of colors apparent on a pet's fur coat. Fig 12.1 shows this distribution for dogs and fig 12.2 shows the distribution for cats.

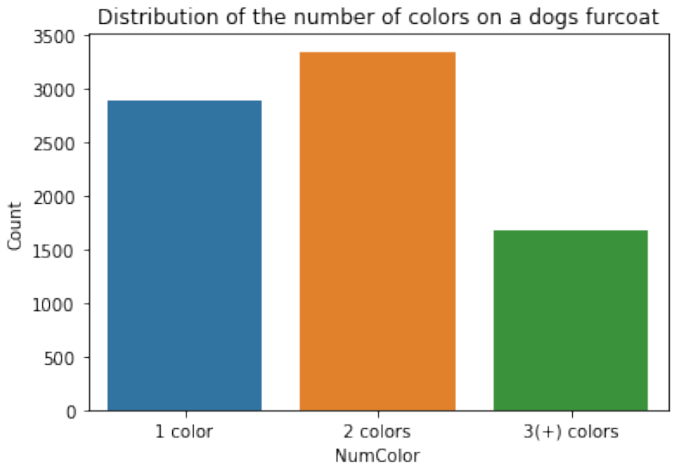


Fig 12.1 Distribution of dog color counts

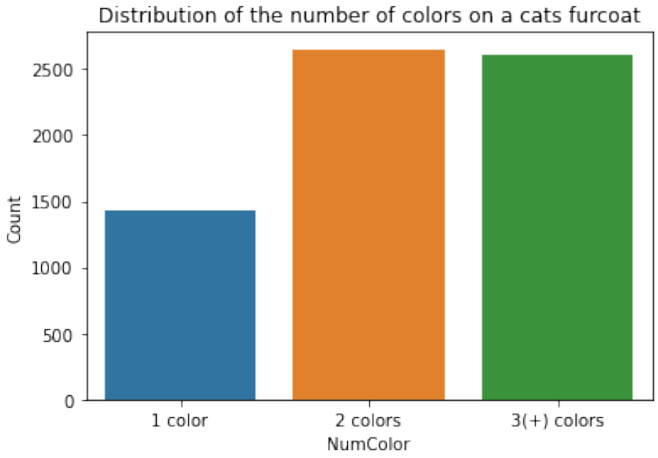


Fig 12.2 Distribution of cat color counts

One hot encoding was performed for the 'NumColor' feature to create 3 new features: 'NumColor1', 'NumColor2' and 'NumColor3'.

Color: One hot encoded dominant color.

Finally, I only kept the feature 'Color1' and dropped 'Color2' and 'Color3'. I performed one hot encoding on 'Color1' which resulted in the final 7 features: 'Color_1', 'Color_2', 'Color_3', 'Color_4', 'Color_5', 'Color_6' and 'Color_7', representing black, brown, golden, yellow, cream, gray and white respectively. Fig 13.1 shows the distribution of dominant colors for dogs and fig 13.2 shows this distribution for cats.

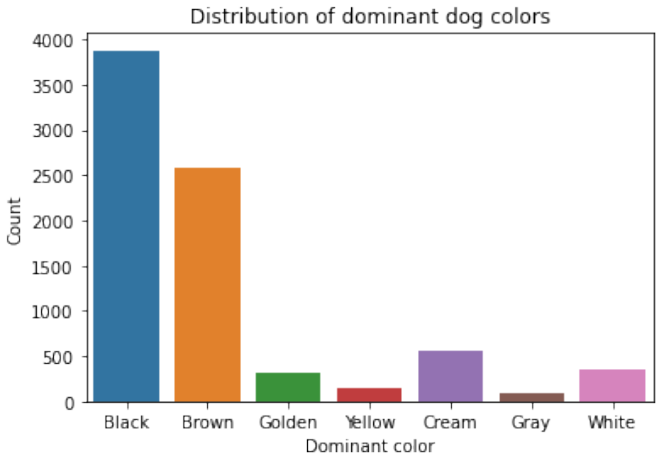


Fig 13.1 Distribution of dog dominant color

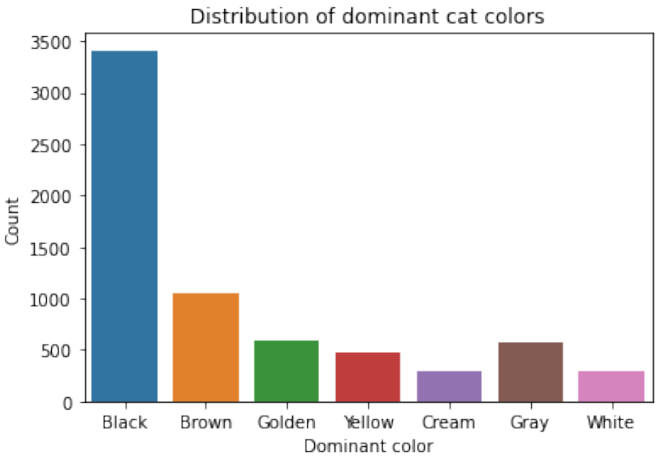


Fig 13.2 Distribution of cat dominant color

MaturitySize: The size of the pet (Small, Medium, Large, Extra Large)

Fig 14.1 shows the distribution of dog maturity sizes and fig 14.2 shows the distribution of cat maturity sizes.

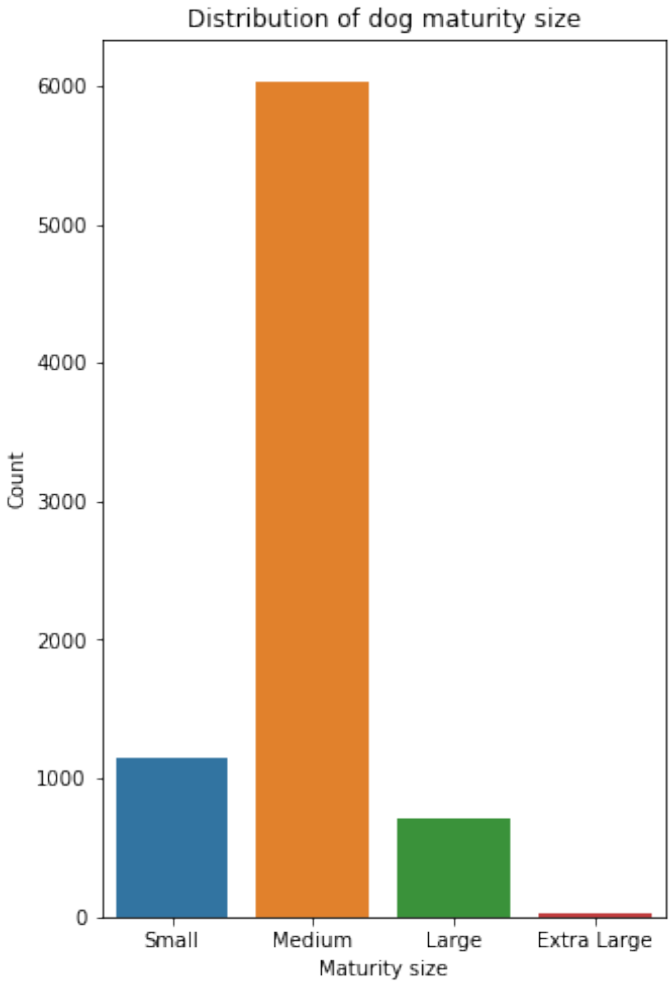


Fig 14.1 Distribution of dog maturity size

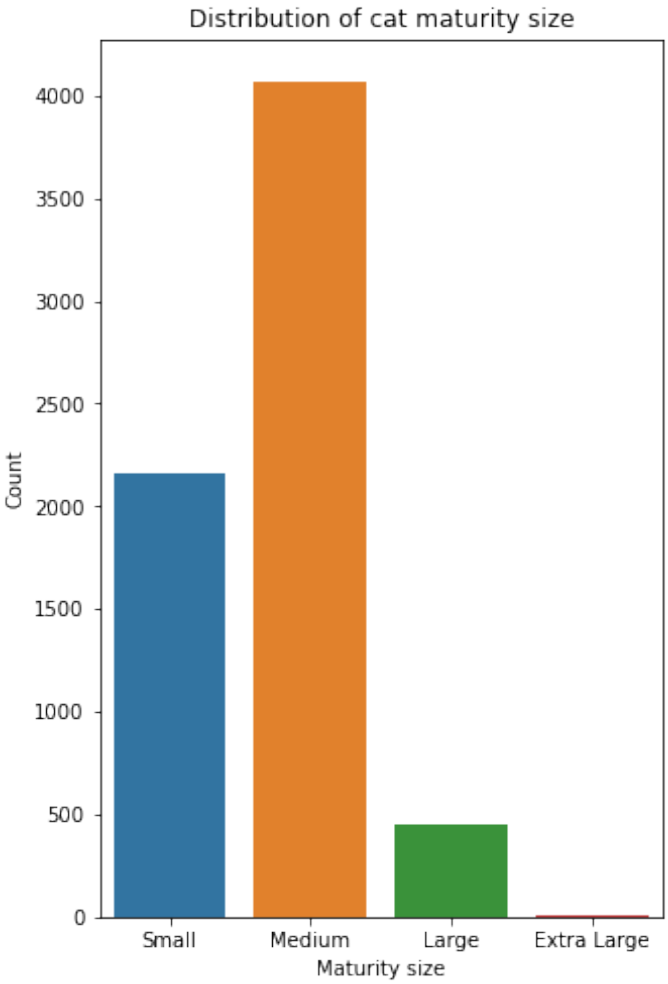


Fig 14.2 Distribution of cat maturity size

Since the number of extra-large pets are extremely small, I grouped the extra-large category together with large. I then performed one hot encoding to generate features 'Size_1', 'Size_2' and 'Size_3,' which represent small, medium and large – extra-large pets respectively.

FurLength: The fur length of the pet (Short, Medium, Long)

Fig 15.1 shows the distribution of dog fur lengths and fig 15.2 shows the distribution of cat fur lengths.

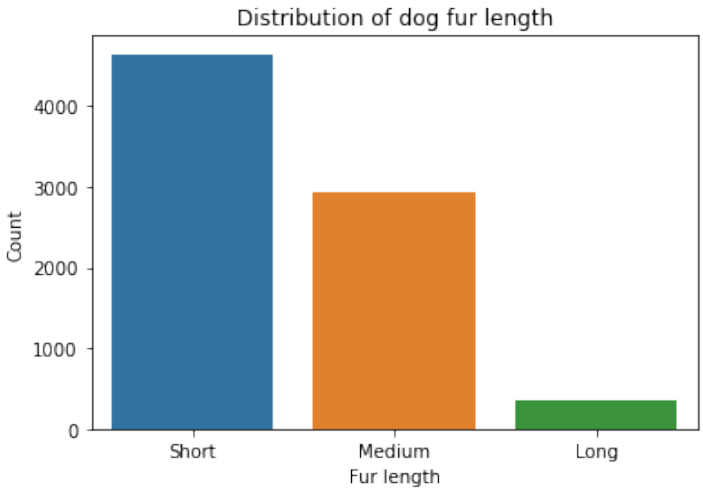


Fig 15.1 Distribution of dog fur length

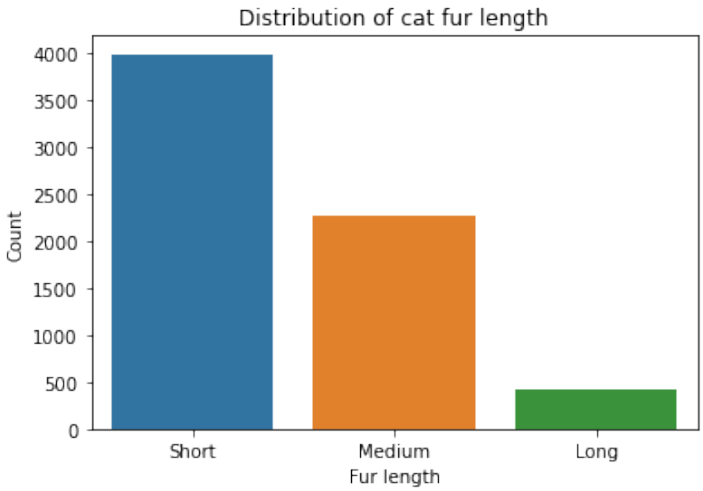


Fig 15.2 Distribution of dog fur length

Since the number of pets with long fur coats are relatively small, I grouped the long category together with medium. I then performed one hot encoding to generate features 'FurLength_1' and 'FurLength_2' which represents short and medium-long fur lengths respectively.

Vaccinated: Whether or not a pet has been vaccinated (Yes, No, Unknown)

Fig 16.1 shows the distribution of dog vaccination status and fig 16.2 shows the distribution of cat vaccination status.

I grouped together 'No' and 'Unknown' vaccination statuses since an unknown vaccination status will be treated essentially the same as a 'no' by pet adopters. I then performed one hot encoding to generate features 'Vaccinated_0' and ''Vaccinated_1' which represents not vaccinated and vaccinated respectively.

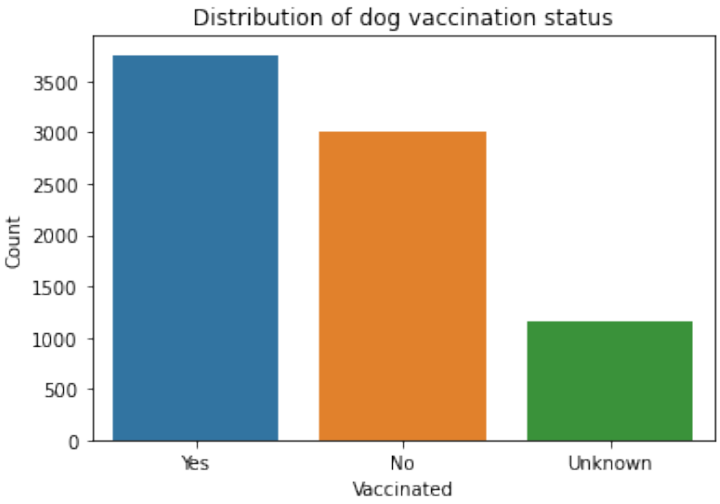


Fig 16.1 Distribution of dog vaccination status

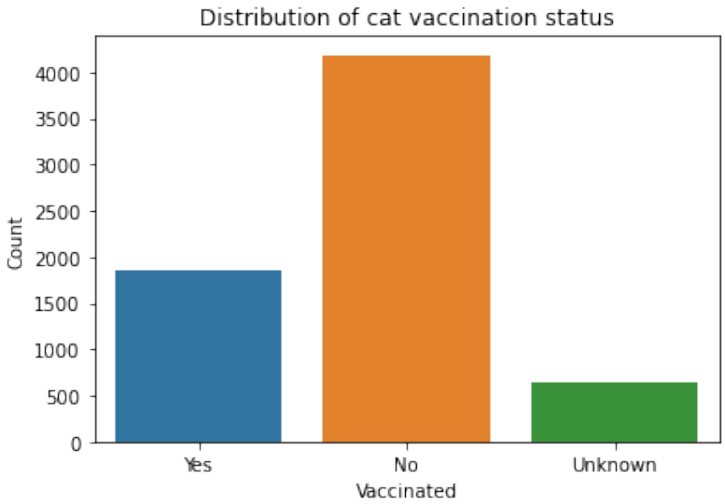


Fig 16.2 Distribution of cat vaccination status

Dewormed: Whether or not a pet has been dewormed (Yes, No, Unknown)

Fig 17.1 shows the distribution of dog deworming status and fig 17.2 shows the distribution of cat deworming status.

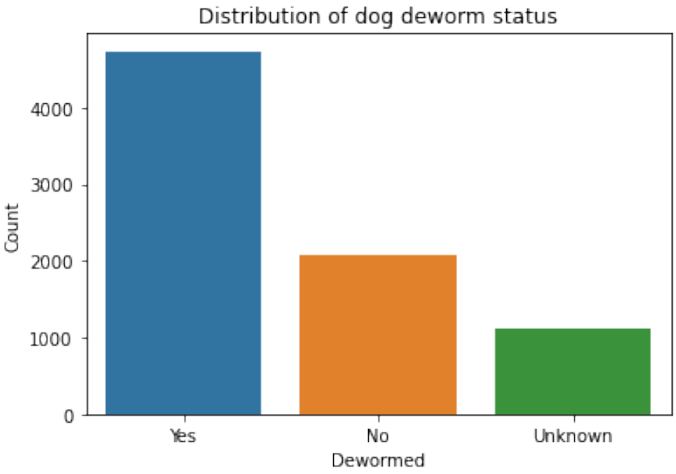


Fig 17.1 Distribution of dog deworm status

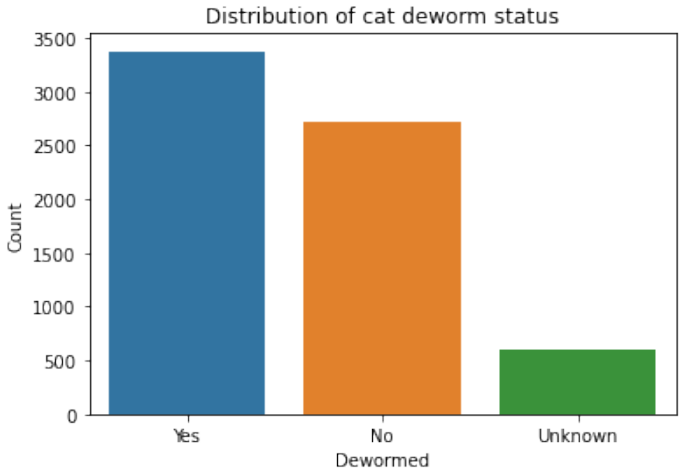


Fig 17.2 Distribution of cat deworm status

I grouped together 'No' and 'Unknown' deworming statuses since an unknown deworming status will be treated essentially the same as a 'no' by dog adopters. I then performed one hot encoding to generate features 'Dewormed_0' and ''Dewormed_1' which represents not dewormed and dewormed respectively.

Sterilized: Whether or not a dog has been sterilized (Yes, No, Unknown)

Fig 18.1 shows the distribution of dog sterilization status and fig 18.2 shows the distribution of cat sterilization status.

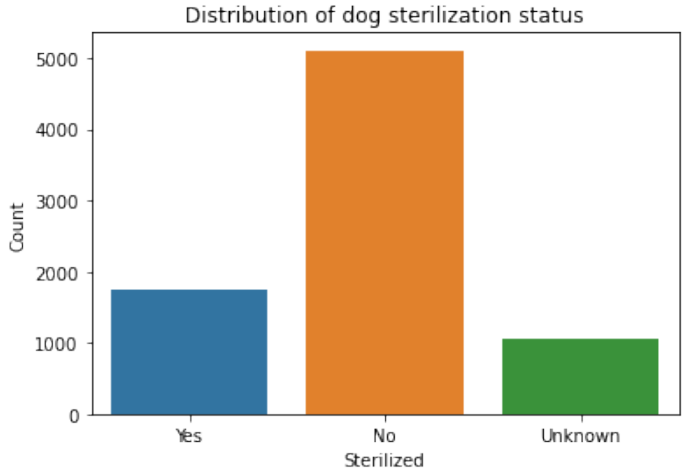


Fig 18.1 Distribution of dog sterilization

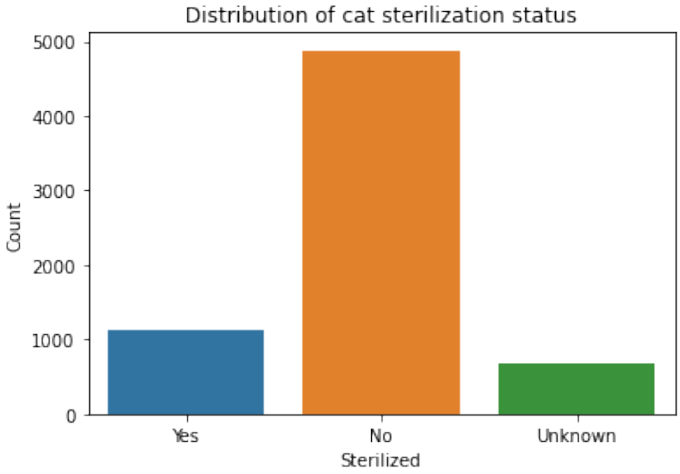


Fig 18.2 Distribution of cat sterilization

I grouped together 'No' and 'Unknown' sterilization statuses since an unknown sterilization status will be treated essentially the same as a 'no' by dog adopters. I then performed one hot encoding to generate features 'Sterilized_0' and ''Sterilized_1' which represents not sterilized and sterilized respectively.

Preparation: Whether or not a pet is fully prepared for adoption (Yes, No)

I engineered a new feature based on 'Vaccinated', 'Dewormed' and 'Sterilized' features. If a pet was sterilized, vaccinated and dewormed, they will be labelled as fully prepared (1) and 0 otherwise. Fig 19.1 shows the distribution for dog adoption preparation and fig 19.2 shows the distribution for cat adoption preparation.

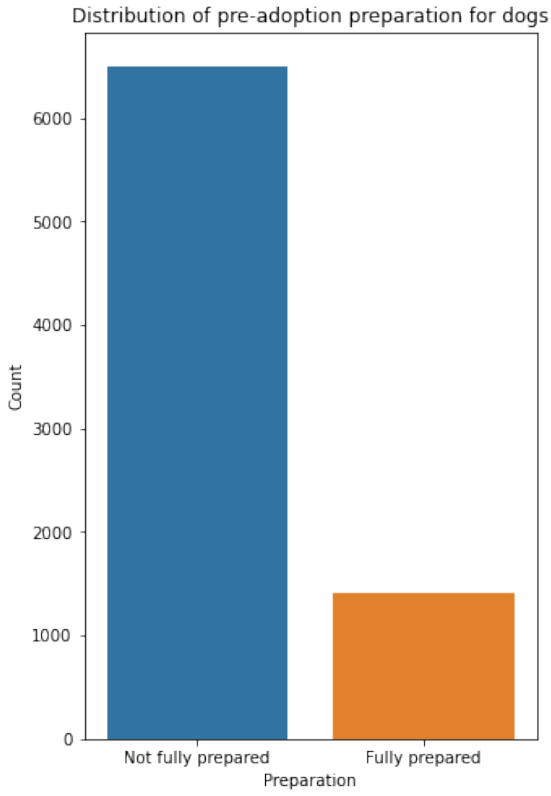


Fig 19.1 Distribution of dog preparation

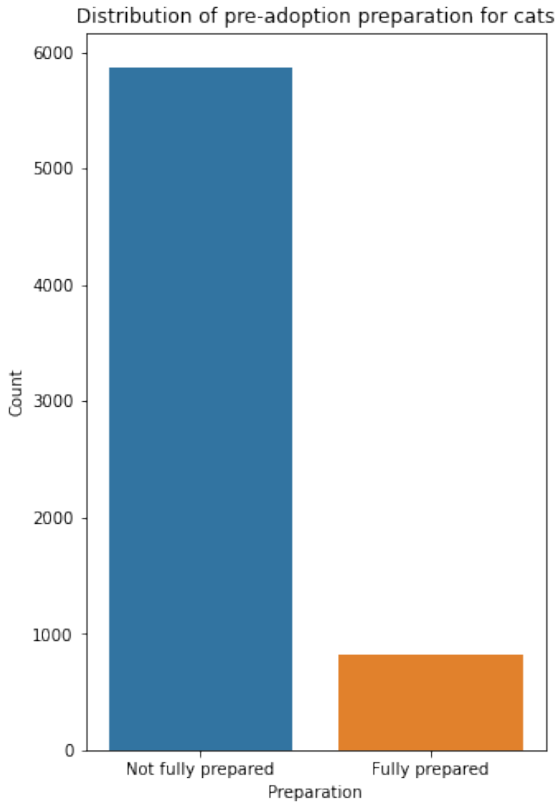


Fig 19.2 Distribution of cat preparation

Health: The health status of a pet (Healthy, Minor injury, Serious injury)

Fig 20.1 shows the distribution of dog health and fig 20.2 shows the distribution of cat health.

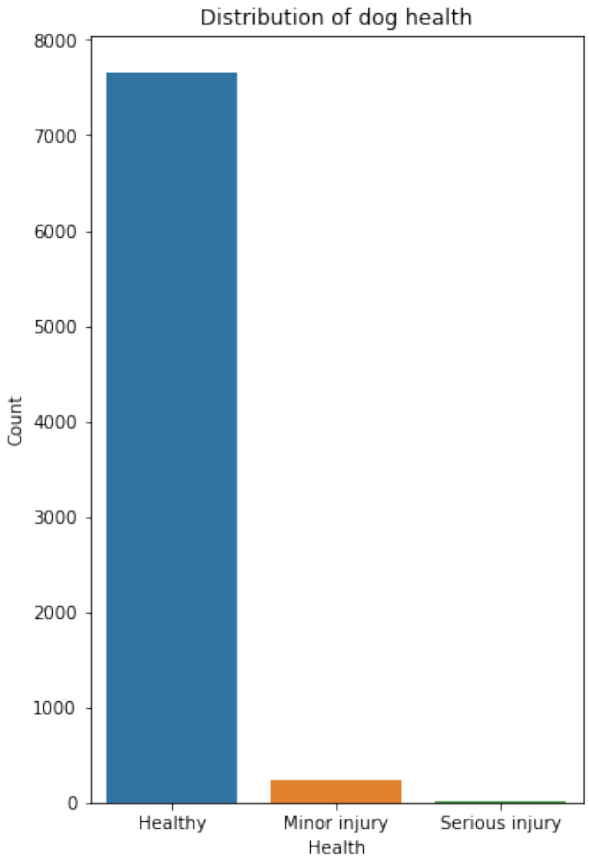


Fig 20.1 Distribution of dog health

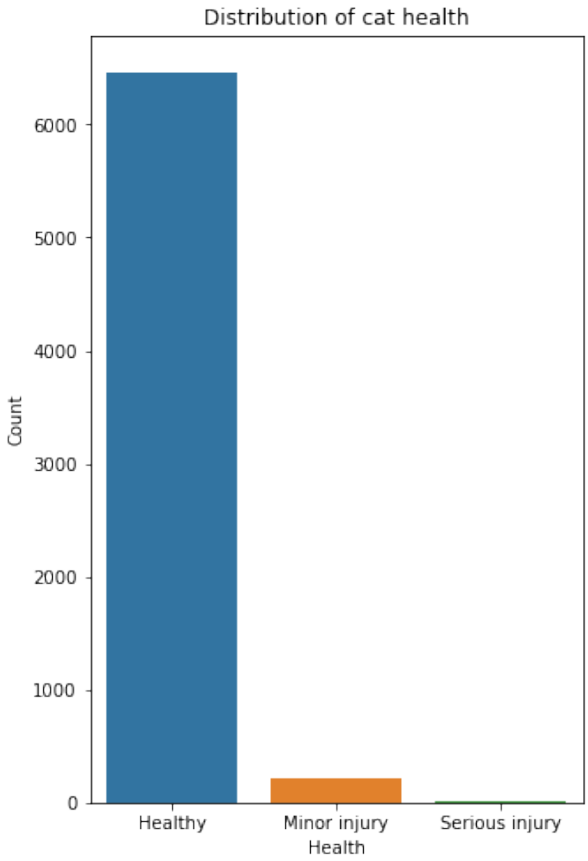


Fig 20.2 Distribution of cat health

Since a very small number of pets lie in the minor injury group and an even smaller number of pets are seriously injured, I grouped together 'Serious injury' and 'Minor injury.' I then performed one hot encoding to generate features 'Health_0' and ''Health_1' which represents injured and healthy respectively.

Cost: The foreseeable costs of the pet (0, 1, 2, 3, 4)

Based on the 'Vaccinated', 'Dewormed', 'Sterilized' and 'Health', I created a new feature 'Costly' which depicts the foreseeable cost for a cost. This ranges from 0 – 4, 0 for none and 4 representing that there are expected costs for vaccination, deworming, sterilization and health issues. Fig 21.1 shows the cost distribution for dogs and fig 21.2 shows this distribution for cats.

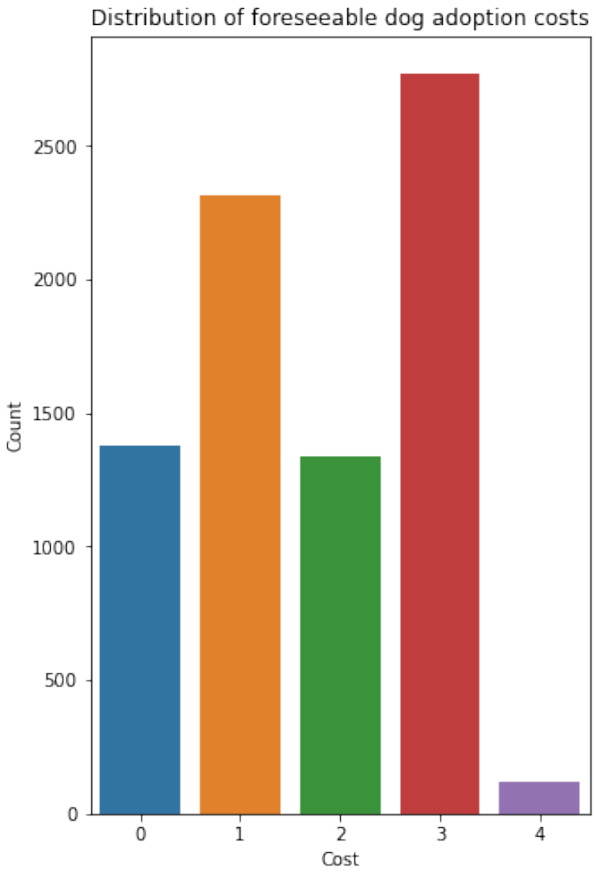


Fig 21.1 Distribution of foreseeable dog adoption cost

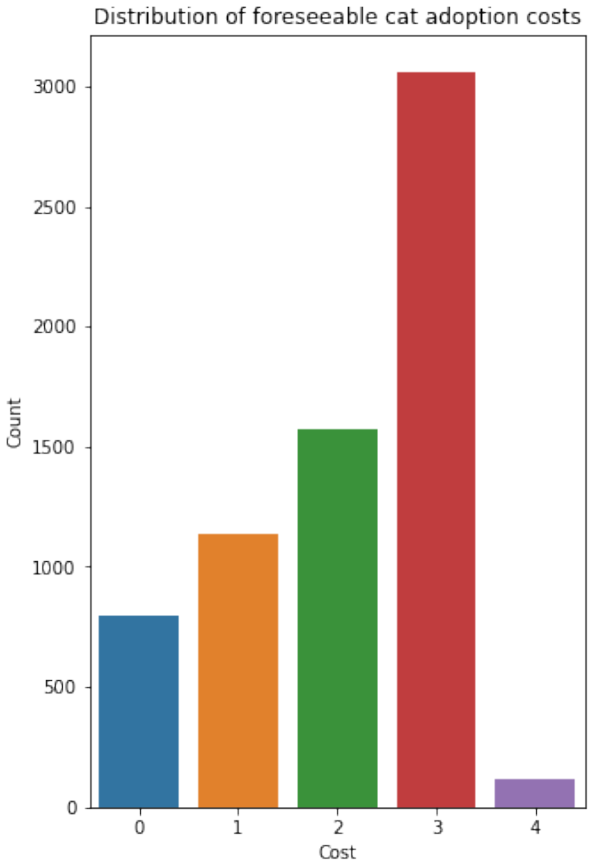


Fig 22.2 Distribution of foreseeable cat adoption cost

Quantity: The number of pets in a single listing

Fig 22.1 shows the distribution of dog quantities and fig 22.2 shows the distribution of cat quantities. The dog adoption quantities range from 1 to 20, with 75% of the listings having a quantity of 1. The cat adoption quantities also range from 1 to 20, with over 50% of the listings having a quantity of 1.

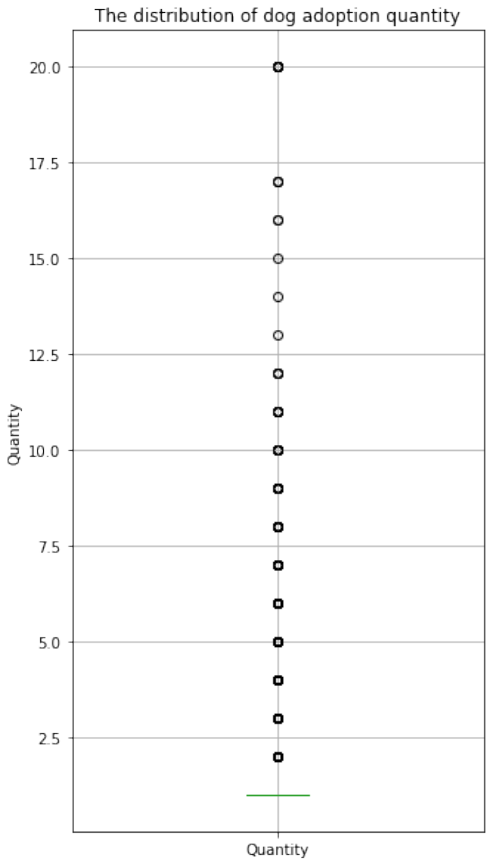


Fig 22.1 Distribution of dog quantities

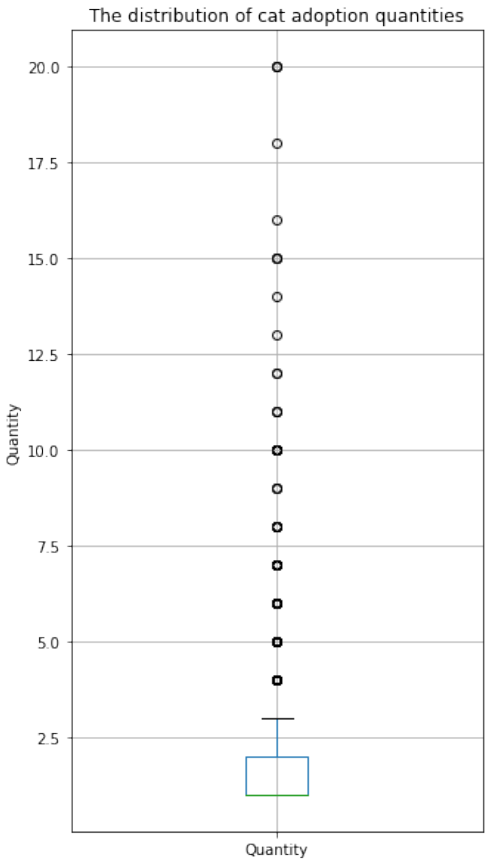


Fig 22.2 Distribution of cat quantities

Fee: The adoption fee for a particular pet listing

Fig 23.1 shows the distribution of dog adoption fees. Dog adoption fees range from 0 to $3000MYR, with 75% of the listings having no adoption fee. A total of 234 outliers were removed.

Fig 23.2 shows the distribution of cat adoption fees. Cat adoption fees range from 0 to $800MYR, with 75% of the listings having no adoption fee. A total of 109 outliers were removed.

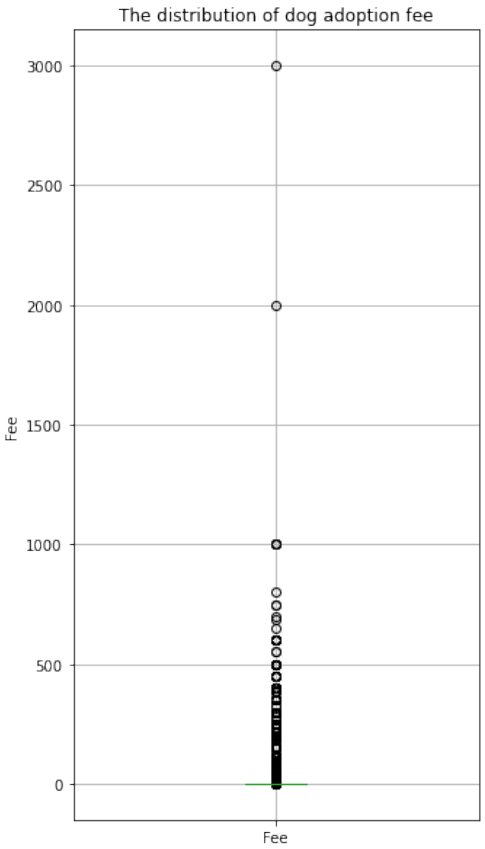


Fig 23.1 Distribution of dog adoption fee

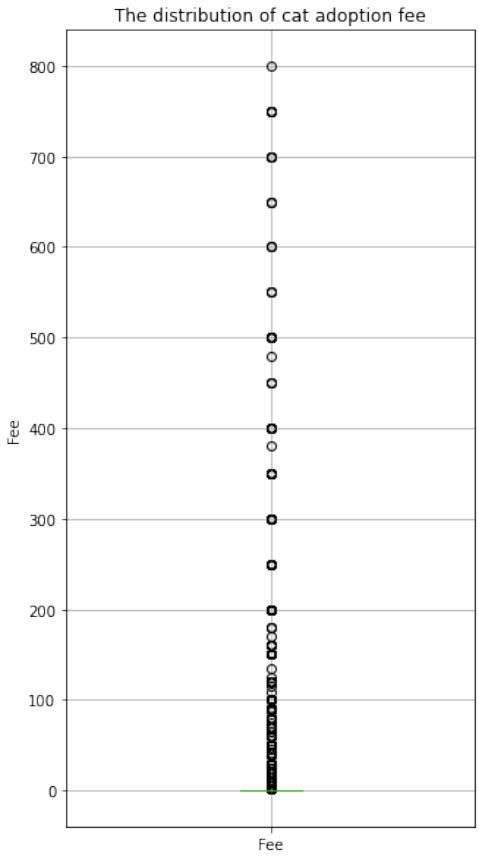


Fig 23.2 Distribution of cat adoption fee

State: The Malaysian state the adoptable pet is located in

'Density': The relative density of pets adoptable in each Malaysian state.

The 'Density' feature was derived from the 'State' feature. The normalized count values for each state was stored in a new feature 'Density' which represents the normalized number of pets found in each Malaysian state.

Top 7 states

The 'State' feature includes all 15 states of Malaysia. I kept the top 7 states with the highest number of adoptable pets for each dataset and grouped the remaining 6 states into 'others'. Fig 24.1 shows the one hot encoded state features for dog adoption. Fig 24.2 shows the one hot encoded state features for cat adoption.

Feature	Description
State_0	Whether or not the dog is located in: ‘ Kedah’, ‘Kelantan’, ‘Labuan’, ‘Pahang’, ‘Perlis’, ‘Sabah’, ‘Sarawak’, ‘Terengganu’.
State_41324	Whether or not the dog is located in ‘Melaka’
State_41326	Whether or not the dog is located in ‘Selangor’
State_41327	Whether or not the dog is located in ‘Pulau Pinang’
State_41330	Whether or not the dog is located in ‘Perak’
State_41332	Whether or not the dog is located in ‘Negeri Sembilan’
State_41336	Whether or not the dog is located in ‘Johor’
State_41401	Whether or not the dog is located in ‘Kuala Lumpur’

Fig 24.1 One-hot encoded state features for dog adoption

Feature	Description
State_0	Whether or not the cat is located in: ‘ Melaka’, ‘Kelantan’, ‘Labuan’, ‘Pahang’, ‘Perlis’, ‘Sabah’, ‘Sarawak’, ‘Terengganu’.
State_41325	Whether or not the cat is located in ‘Kedah’.
State_41326	Whether or not the cat is located in ‘Selangor’.
State_41327	Whether or not the cat is located in ‘Pulau Pinang’
State_41330	Whether or not the cat is located in ‘Perak’
State_41332	Whether or not the cat is located in ‘Negeri Sembilan’
State_41336	Whether or not the cat is located in ‘Johor’
State_41401	Whether or not the cat is located in ‘Kuala Lumpur’

Fig 24.2 One-hot encoded state features for cat adoption

RescuerID: The unique ID of the pet rescuer

'RescuerFreq': The frequency of the rescuer in the pet adoption dataset

From 'RescuerID', I created a new feature 'RescuerFreq' which represents the frequency (the total number of dog/cat listing rescued by that rescuer) of the rescuer in the pet adoption dataset.

Top 10 rescuers

There is a total of 2735 recuers in the dog adoption dataset and 2827 rescuers in the cat adoption dataset. I kept the top 10 rescuers from each dataset and grouped the remaining rescuers into a single group. I one hot encoded this feature creating 11 new features for each of the datasets.

VideoAmt: The number of videos attached in the pet adoption listing

The number of videos for each dog adoption listing ranges from 0 to 8, with over 90% of dog listings having 0 videos. The number of videos for each cat adoption listing ranges from 0 to 6, with over 95% of cat listings having 0 videos.

PhotoAmt: The number of photos attached in the pet adoption listing

The number of photos for each dog adoption listing ranges from 0 to 30, with over 75% of dog listings having 5 or less photos. Fig 25.1 shows this distribution. The number of photos for each cat adoption listing ranges from 0 to 30, with over 75% of cat listings having 5 or less photos. Fig 25.2 shows this distribution.

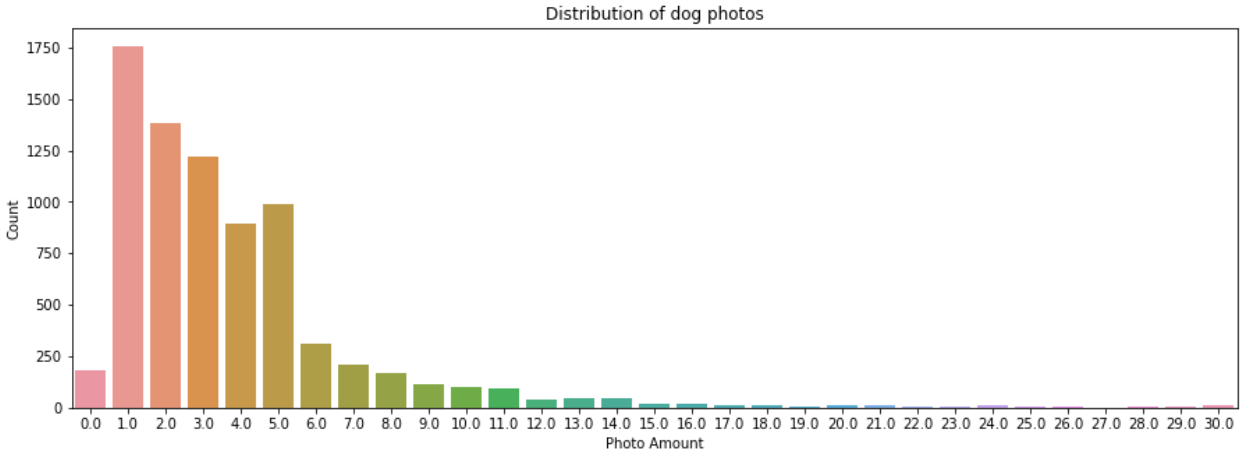


Fig 25.1 Distribution of photos per dog listing

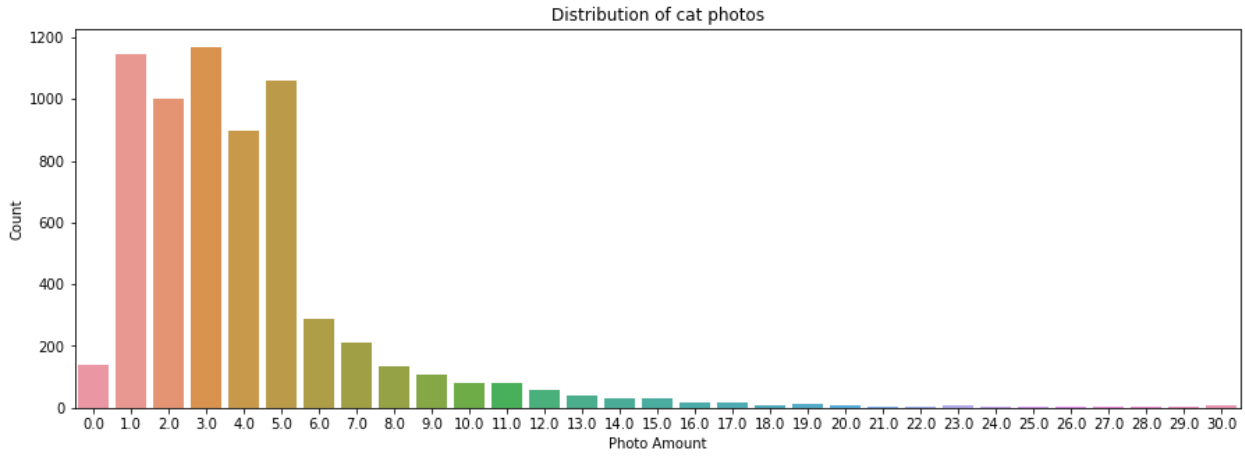


Fig 25.2 Distribution of photos per cat listing

Media: Whether or not there is at least 1 photo or video on file for the pet

This feature was derived from the 'VideoAmt' and 'PhotoAmt' features. It depicts whether or not the pet has at least 1 photo or video on file. This feature was one hot encoded to create 'Media_0' and 'Media_1' features. Fig 26.1 shows the distribution of media for dog adoption and fig 26.2 shows the distribution of media for cat adoption.

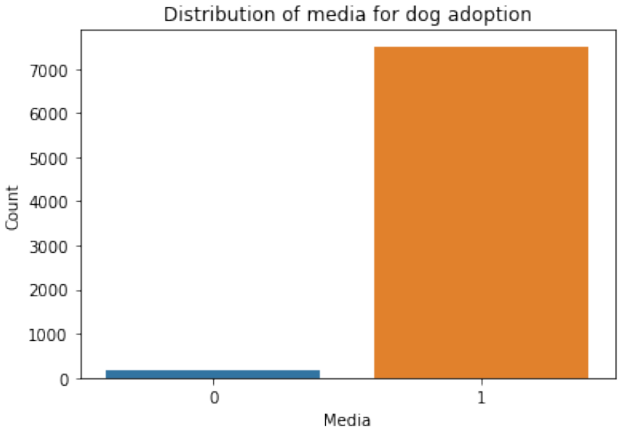


Fig 26.1 Distribution of media for dog adoption

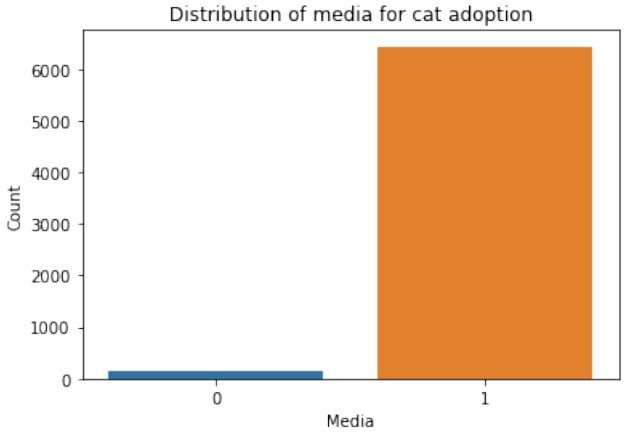


Fig 26.2 Distribution of media for cat adoption

Description: The description included in the pet adoption listing

DescLen: The length of the description included

Fig 27.1 shows the distribution of description lengths for dog adoption and fig 27.2 shows the distribution of description lengths for cat adoption.

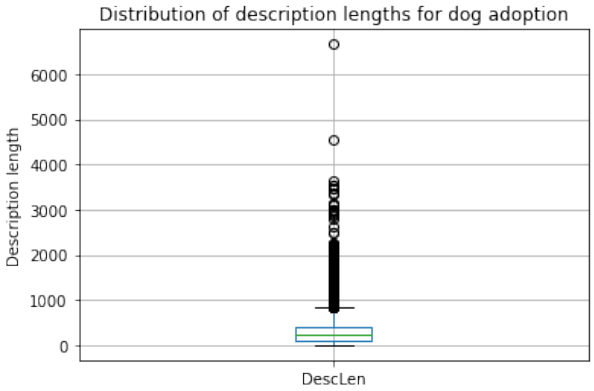


Fig 27.1 Distribution of description lengths for dog adoption

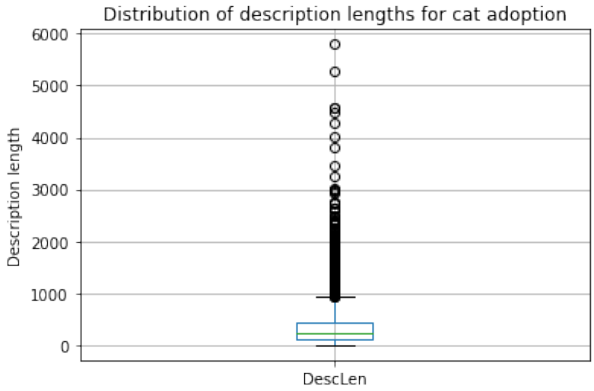


Fig 27.2 Distribution of description lengths for cat adoption

DescNumWords: The number of words in the pet description

Fig 28.1 shows the distribution of word count for dog adoption descriptions and fig 28.2 shows the distribution of word count for cat adoption descriptions.

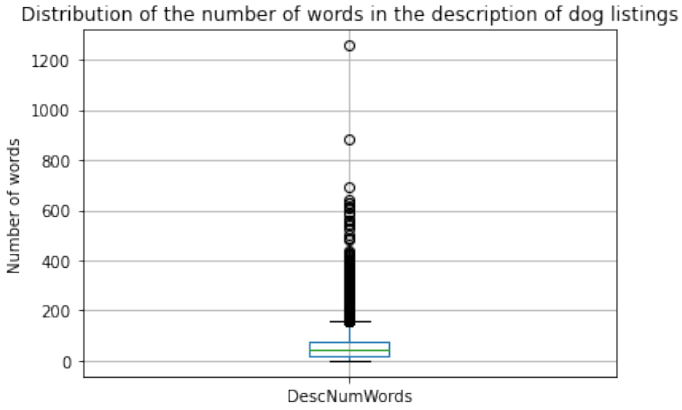


Fig 28.1 Distribution of word count in the description of dog listings

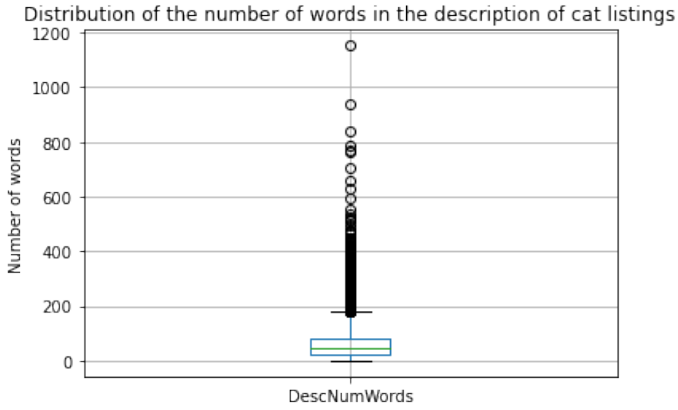


Fig 28.2 Distribution of word count in the description of cat listings

Top 20 keywords

For the dog adoption dataset, the 20 keywords I extracted from TF-IDF and manual combing are: ‘abandoned’, ‘active’, ‘adorable’, ‘attention’, ‘cute’, ‘friendly’, ‘good’, ‘guard’, ‘happy’, ‘healthy’, kind’, ‘little’, ‘lovely’, ‘playful’, ‘puppy’, ‘shelter’, ‘siblings’, ‘stray’, ‘sweet’, and ‘trained.’ Each of these keywords became a new binary feature in my dataset.

For the cat adoption dataset, the 20 keywords I extracted from TF-IDF and manual combing are: 'active', 'adorable', 'beautiful', 'birth', 'box', 'cute', 'eyes', 'family', 'friendly', 'fur', 'good', 'healthy', 'indoor', 'little', 'lovely', 'mother', 'playful', 'siblings', 'sweet' and 'trained.' Each of these keywords became a new binary feature in my dataset.

2.2.1.2 Target Variable

The target variable originally had 5 bins representing adoption made within the 'same day,' '1st week,' '1st month,' '2-3 months,' and 'no adoption after 100 days.' Fig 29.1 shows the distribution of dog adoption speed and fig 29.2 shows the distribution of cat adoption speed.

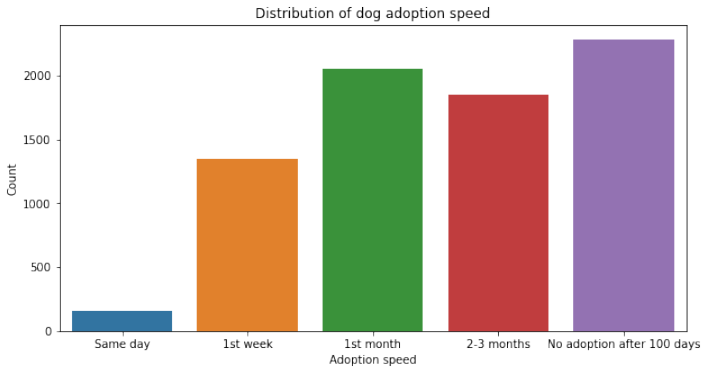


Fig 29.1 Distribution of dog adoption speed

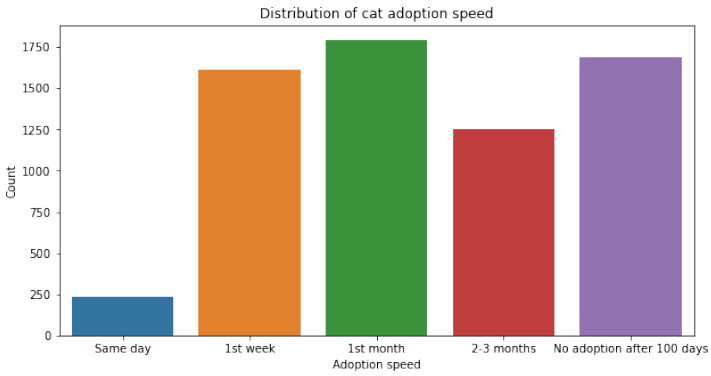


Fig 29.2 Distribution of cat adoption speed

I transformed the target variable into binary by creating a new class 'within 100 days' which groups all the original classes except for 'No adoption after 100 days' together. Fig 30.1 shows the results of this grouping for the dog adoption dataset and fig 30.2 shows the distribution for the cat adoption dataset.

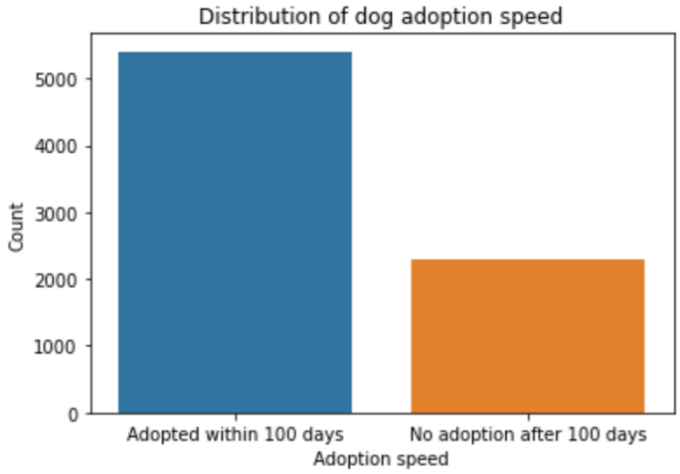


Fig 30.1 Distribution of target variable for dog adoption

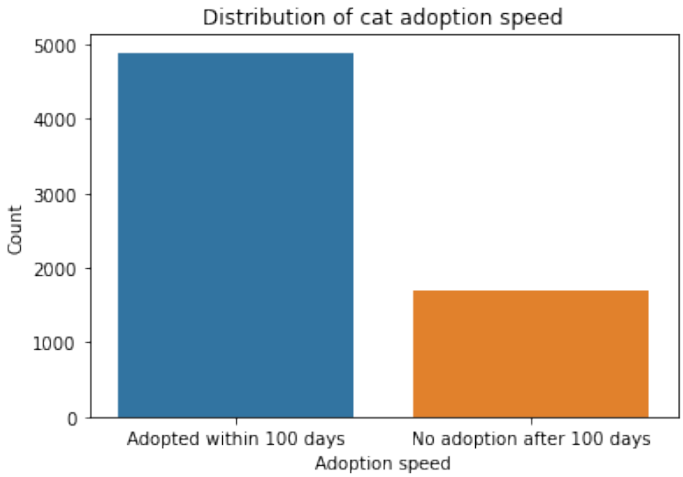


Fig 30.1 Distribution of target variable for cat adoption

3. Exploratory data analysis

3.1 Dog Adoption

Fig 31 shows the correlation matrix for all the features in the dog adoption dataset.

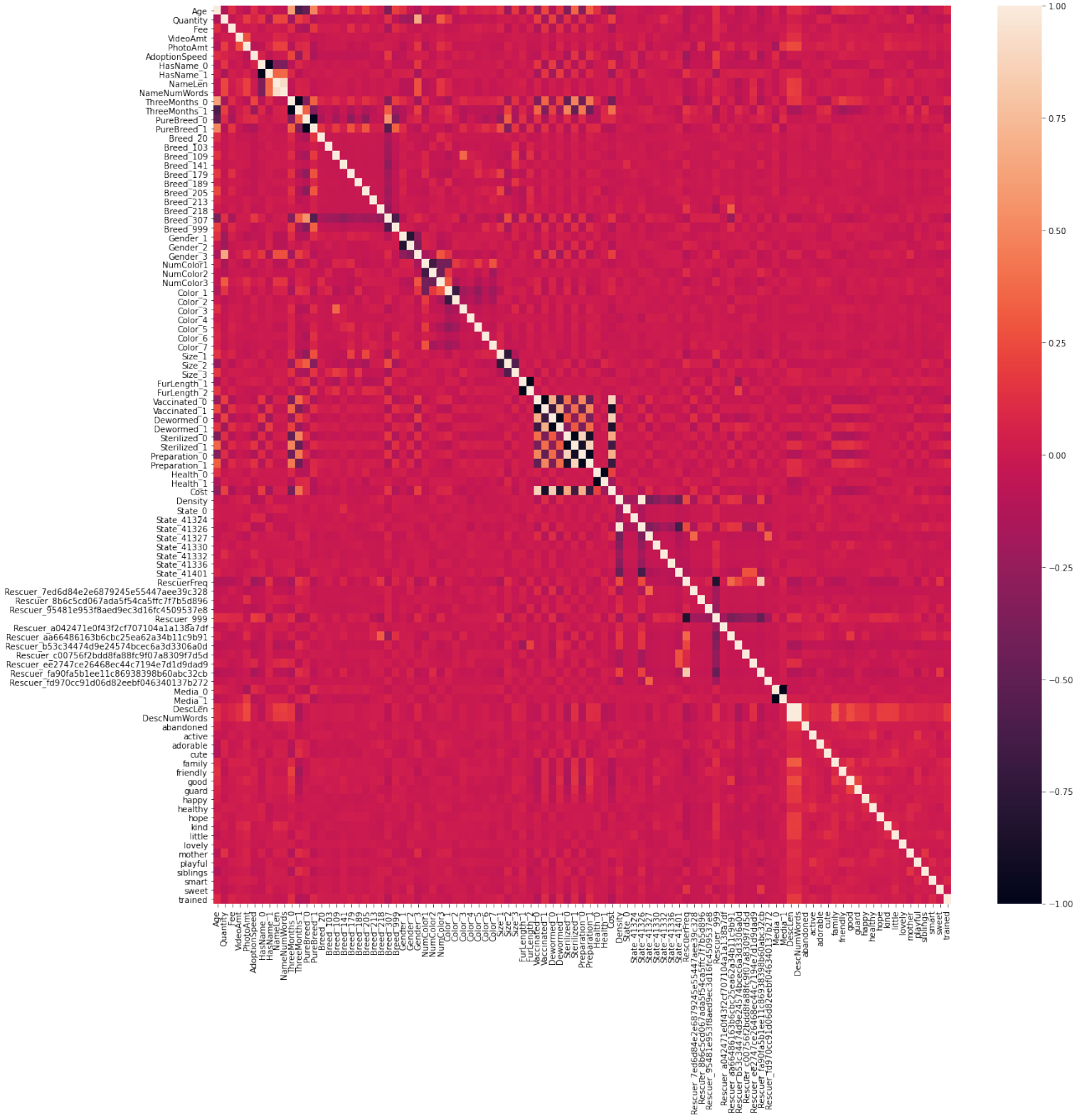


Fig 31 Heat map for the dog adoption dataset

Since there are a lot of features and the correlation matrix is extremely large and difficult to interpret, I printed out the features most positively and negatively correlated with the target variable. Fig 32 shows the correlation values with the target variable: 'AdoptionSpeed'.

```
dogCorr['AdoptionSpeed'].sort_values(ascending = False)
```

```
AdoptionSpeed                                  1.000000
ThreeMonths_0                                  0.221749
Rescuer_999                                    0.191426
Breed_307                                      0.167789
Sterilized_1                                   0.147901
                                                   ...
Preparation_0                                 -0.134943
Sterilized_0                                  -0.147901
Rescuer_fa90fa5b1ee11c86938398b60abc32cb      -0.154560
RescuerFreq                                   -0.206098
ThreeMonths_1                                 -0.221749
Name: AdoptionSpeed, Length: 99, dtype: float64
```

Fig 32 Correlation values between the most correlated features and 'AdoptionSpeed'.

Looking more closely at correlation values with the target variable 'AdoptionSpeed', we see that dog adoption speed is most positively correlated with 'ThreeMonths_0', 'Rescuer_999', 'Breed_307', 'Sterilized_1'. Dog adoption speed appears to be most negatively correlated with 'ThreeMonths_1', 'RescuerFreq', 'Rescuer_fa90fa5b1ee11c86938398b60abc32cb', 'Sterilized_0' and Preparation_0'.

Since a high value for 'AdoptionSpeed' means that the dog wasn't adopted with 100 days, we can interpret these correlation values as follows. Dogs that are adopted within 100 days are correlated with:

- Being 3 months or younger
- Being rescued from one of the top 10 rescuers
- Not having a dominant breed of 'mixed breed'
- Being rescued by a rescuer who has rescued many dogs before
- Being rescued by rescuer fa90fa5b1ee11c86938398b60abc32cb
- Not being sterilized
- Not being fully prepared for adoption

Fig 33 shows scatterplots between each of the continuous features with the target feature in the dog adoption dataset. A few plots that stand out are 'NameLen', 'NameNumWords', 'DescLen', 'DescNumWords', 'VideoAmt' and PhotoAmt'. We can see that there are more adoption listings with a longer name, a longer description and a higher number of photos/videos which have an adoption speed of 0. This suggests that

if a dog adoption listing has a longer name or a longer description or a high number of photos/videos, they are more likely to be adopted within 100 days.

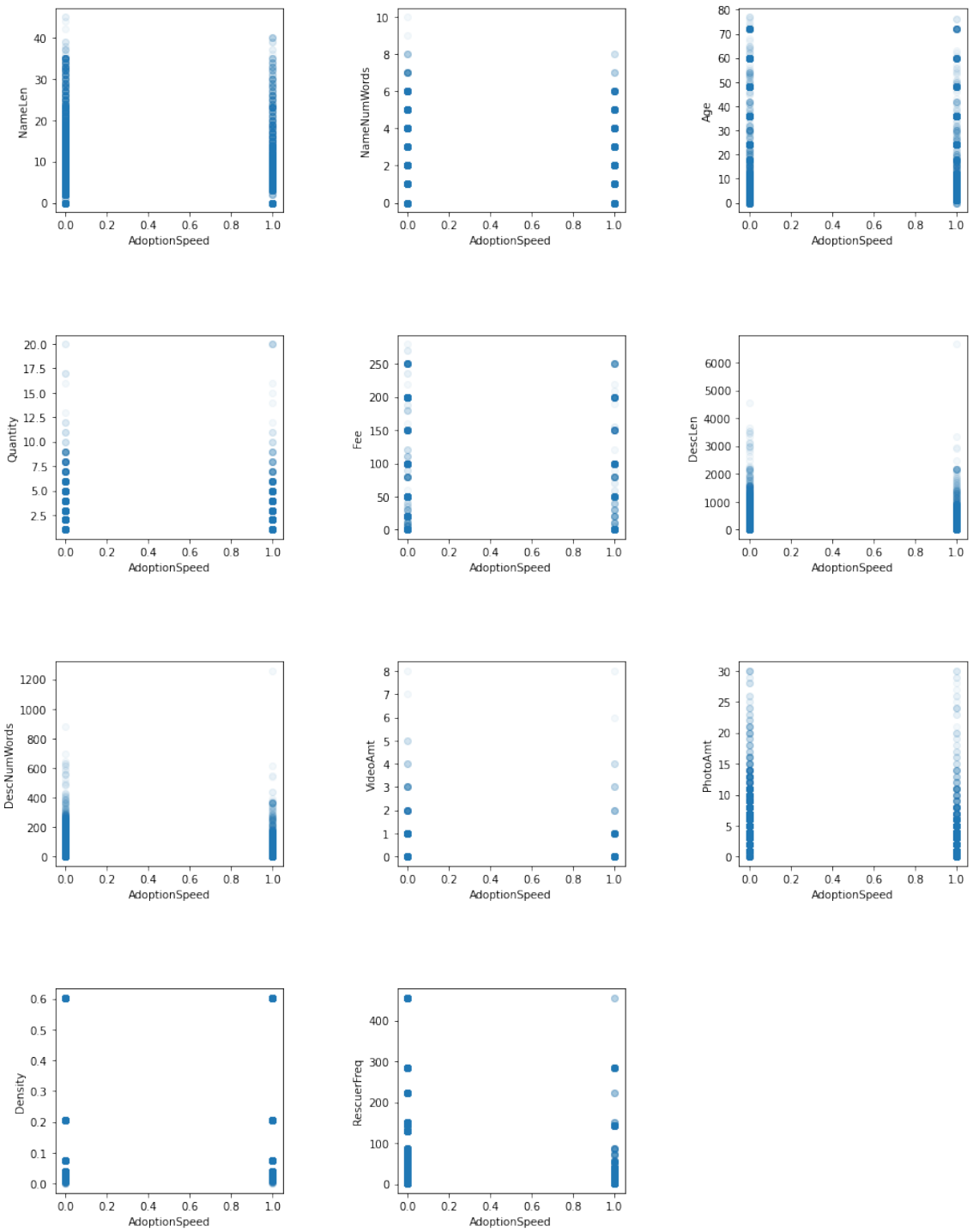


Fig 33 Scatterplots showing the distribution of all continuous features in the dog adoption dataset.

Fig 34 shows the count plots between each of the categorical features with the target feature in the dog adoption dataset. We see that almost all plots have a higher count for adoption speed 0 than 1, this means that most dogs are adopted within 100 days. The

count plots with a higher proportion of 1's than 0's are 'State_41324', 'State_41332', 'Rescuer_ee2747ce26468ec44c7194e7d1d9dad9' and 'Media_0'. This means that dogs that are located in Melaka or Negeri Sembilan, are rescued by rescuer ee2747ce26468ec44c7194e7d1d9dad9 or don't have any photos or videos on file are less likely to be adopted within 100 days. We can see that 'Breed_189' (Rottweiler's) are highly likely to be adopted with 100 days. Interestingly, we can see that dogs rescued by rescuer 95481e953f8aed9e3d16fc4509537e8 has an equal chance of being adopted within 100 days and not being adopted within 100 days.

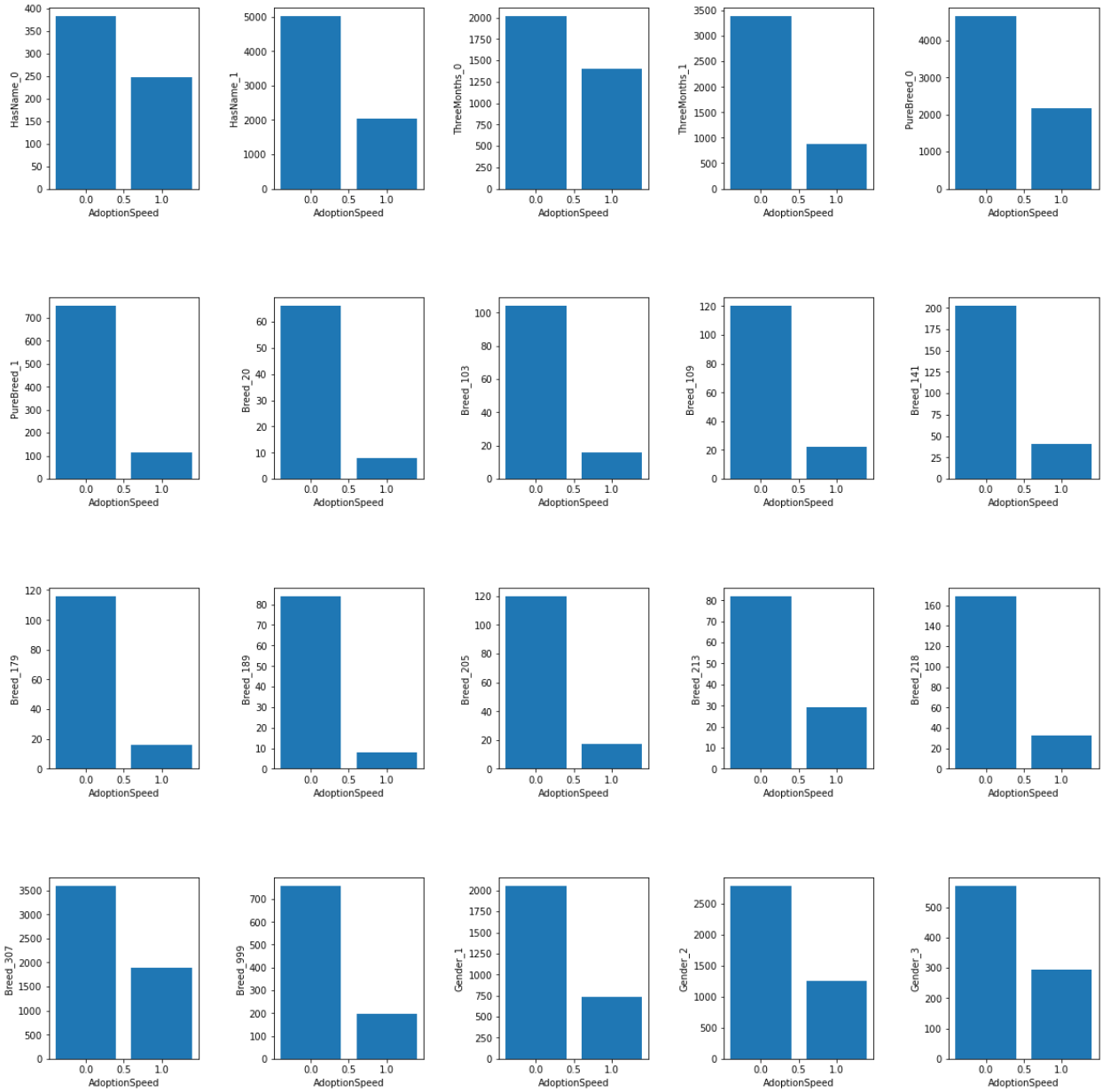


Fig 34.1 Count plots between categorical features and the target variable in the dog adoption dataset

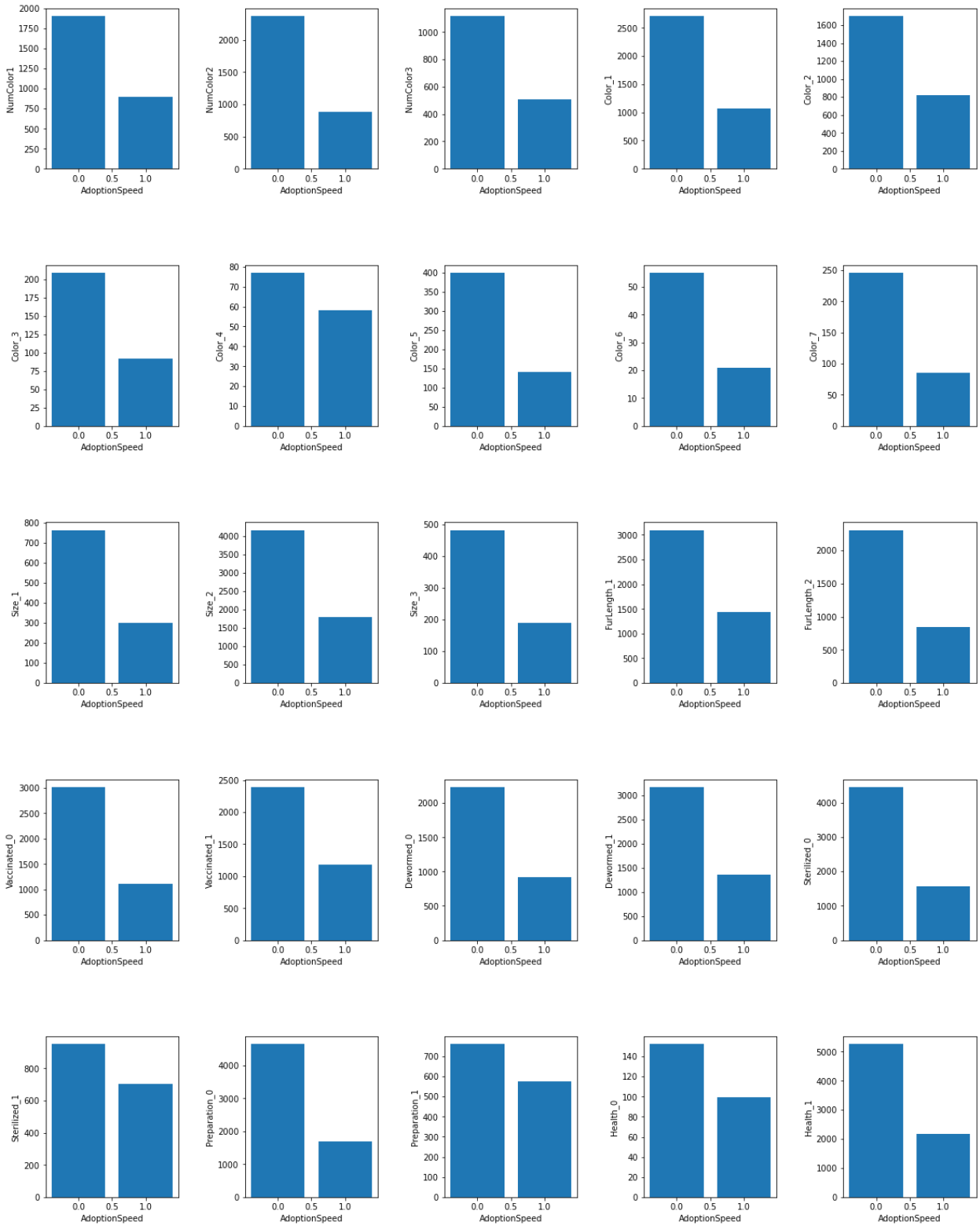


Fig 34.2 Count plots between categorical features and the target variable in the dog adoption dataset

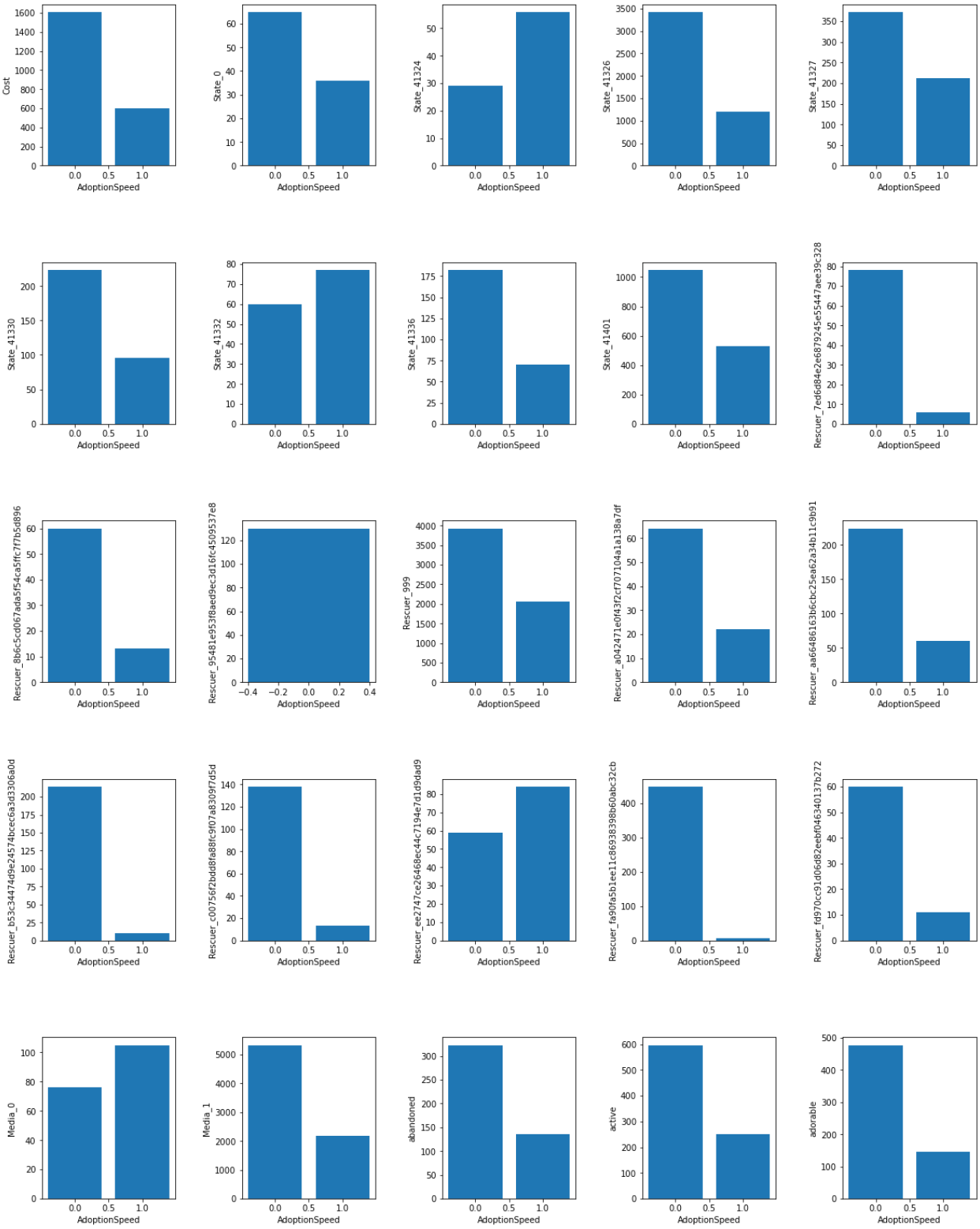


Fig 34.3 Count plots between categorical features and the target variable in the dog adoption dataset

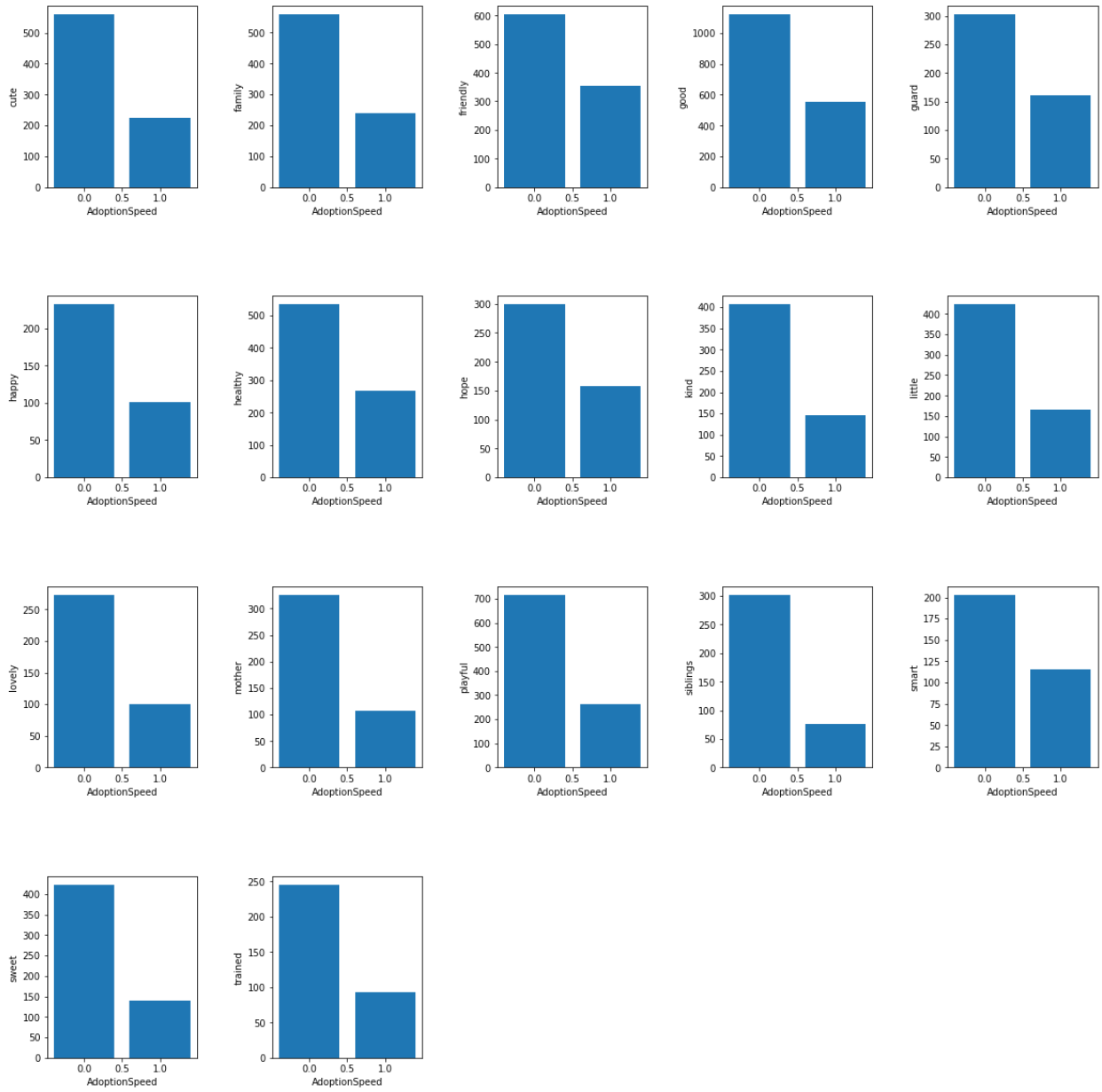


Fig 34.4 Count plots between categorical features and the target variable in the dog adoption dataset

3.2 Cat Adoption

Fig 35 shows the correlation matrix for all the features in the cat adoption dataset. Fig 36 shows the correlation values with the target variable: 'AdoptionSpeed'.

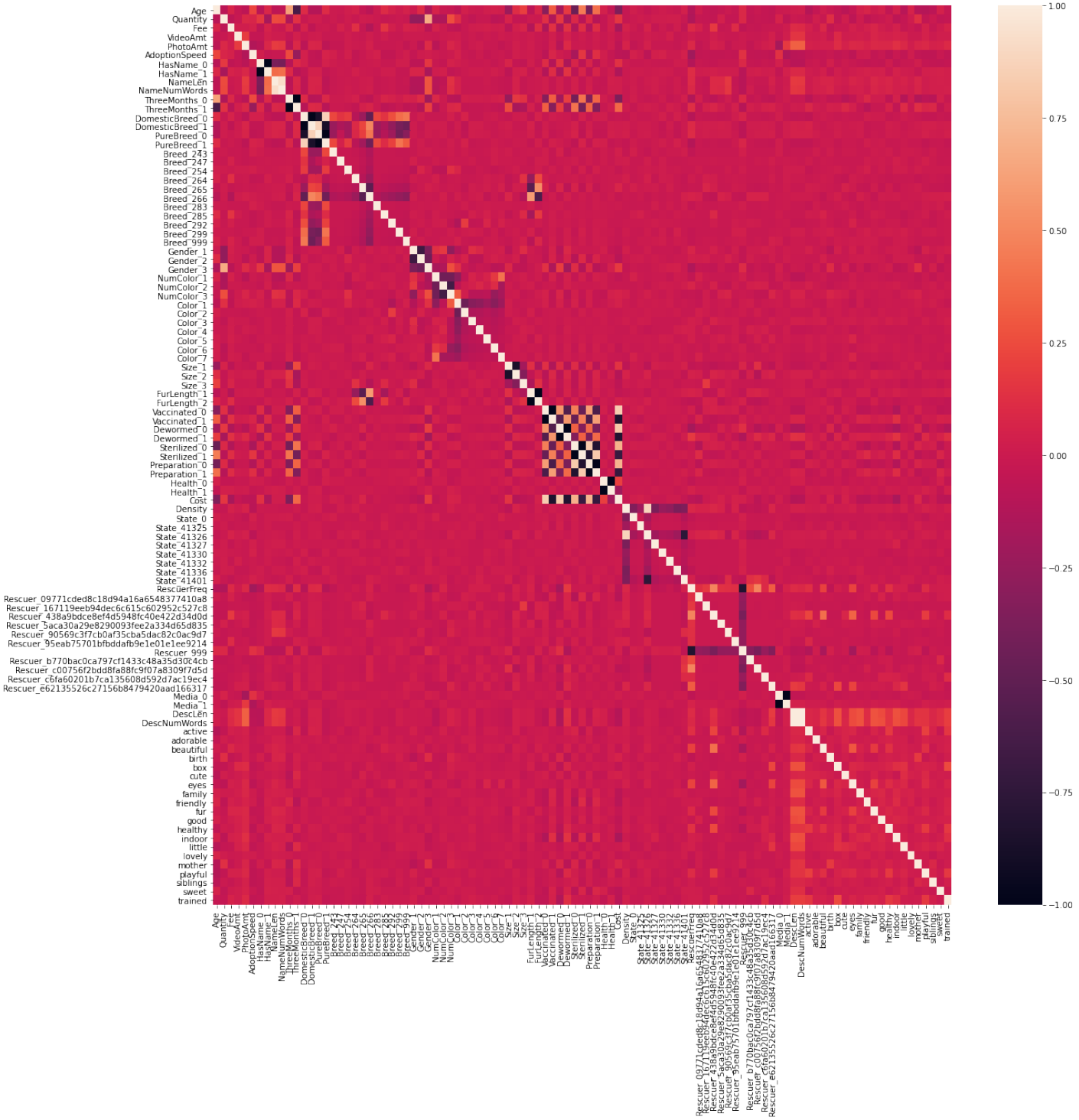


Fig 35 Heat map for the cat adoption dataset

Looking more closely at correlations with the target variable 'AdoptionSpeed', we see that cat adoption speed is most positively correlated with 'ThreeMonths_0', 'Age', 'Sterilized_1' and 'Rescuer_999'. Cat adoption speed appears to be most negatively correlated with 'ThreeMonths_1', 'RescuerFreq', 'Sterilized_0', 'Media_1' and 'Preparation_0'.

```
catCorr['AdoptionSpeed'].sort_values(ascending = False)
```

```
AdoptionSpeed    1.000000
ThreeMonths_0    0.244620
Age              0.218978
Sterilized_1     0.162493
Rescuer_999      0.140622
                    ...
Preparation_0   -0.124637
Media_1         -0.135290
Sterilized_0    -0.162493
RescuerFreq     -0.196421
ThreeMonths_1   -0.244620
Name: AdoptionSpeed, Length: 101, dtype: float64
```

Fig 36 Correlation values between the most correlated features and 'AdoptionSpeed'

Since a high value for 'AdoptionSpeed' means that the cat wasn't adopted with 100 days, we can interpret these correlation values as follows. Cats that are adopted within 100 days are correlated with:

- Being 3 months or younger
- Being of a younger age
- Not being sterilized
- Being rescued by one of the top 10 rescuers
- Being rescued by a rescuer that has rescued multiple other pets
- Having at least 1 photo or video on file
- Not being fully prepared for adoption (vaccination, deworming & sterilization procedures)

Fig 37 shows scatterplots between each of the continuous features with the target feature (Adoption speed) in the cat adoption dataset. A few plots that stand out are 'NameLen', 'NameNumWords', 'DescLen', 'DescNumWords', 'VideoAmt', PhotoAmt' and 'RescuerFreq'. We can see that there are more adoption listings with a longer name, a longer description, a higher number of photos/videos (respectively) and a higher rescuer frequency with an adoption speed of 0 than 1. This suggests that if a cat adoption listing has a long name, a longer description, a higher number of photos/videos or is rescued from a popular rescuer, they are more likely to be adopted within 100 days.

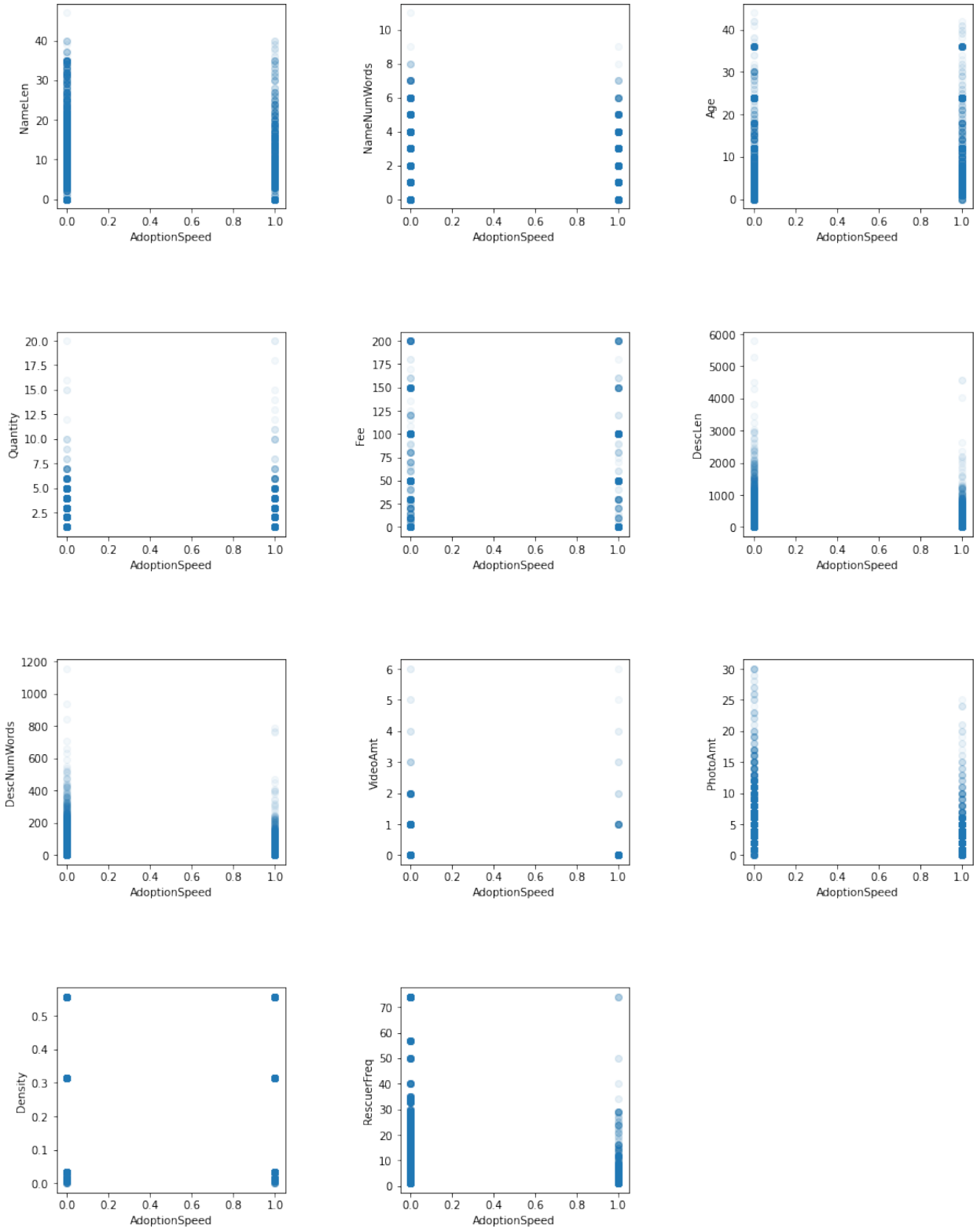


Fig 37 Scatterplots showing the distribution of all continuous variables in the cat adoption dataset

Fig 38 shows the count plots between each of the categorical features with the target feature in the cat adoption dataset. We see that almost all plots have a higher count for adoption speed 0 than 1, this means that most cats are adopted within 100 days. 'Media_0' is the only count plot with a higher proportion of 1's than 0's, this means that cats with no photos or videos on their adoption file are likely to not be adopted within 100 days. We can also see that the count plots for 'Breed_285' and 'siblings' have an exceptionally high proportion of 0's compared to 1. 'Breed_285' corresponds to the 'Persian' cat breed, this breed appear to be very adoptable based on initial exploratory data analysis. The high proportion of 0's for 'siblings' indicate that cat profiles with the word 'sibling' in their description tend to make the cat more adoptable. There are also several rescuers that have an equal number of cats between the two different adoption speeds, this shows that adoption speed is not affected by those rescuers.

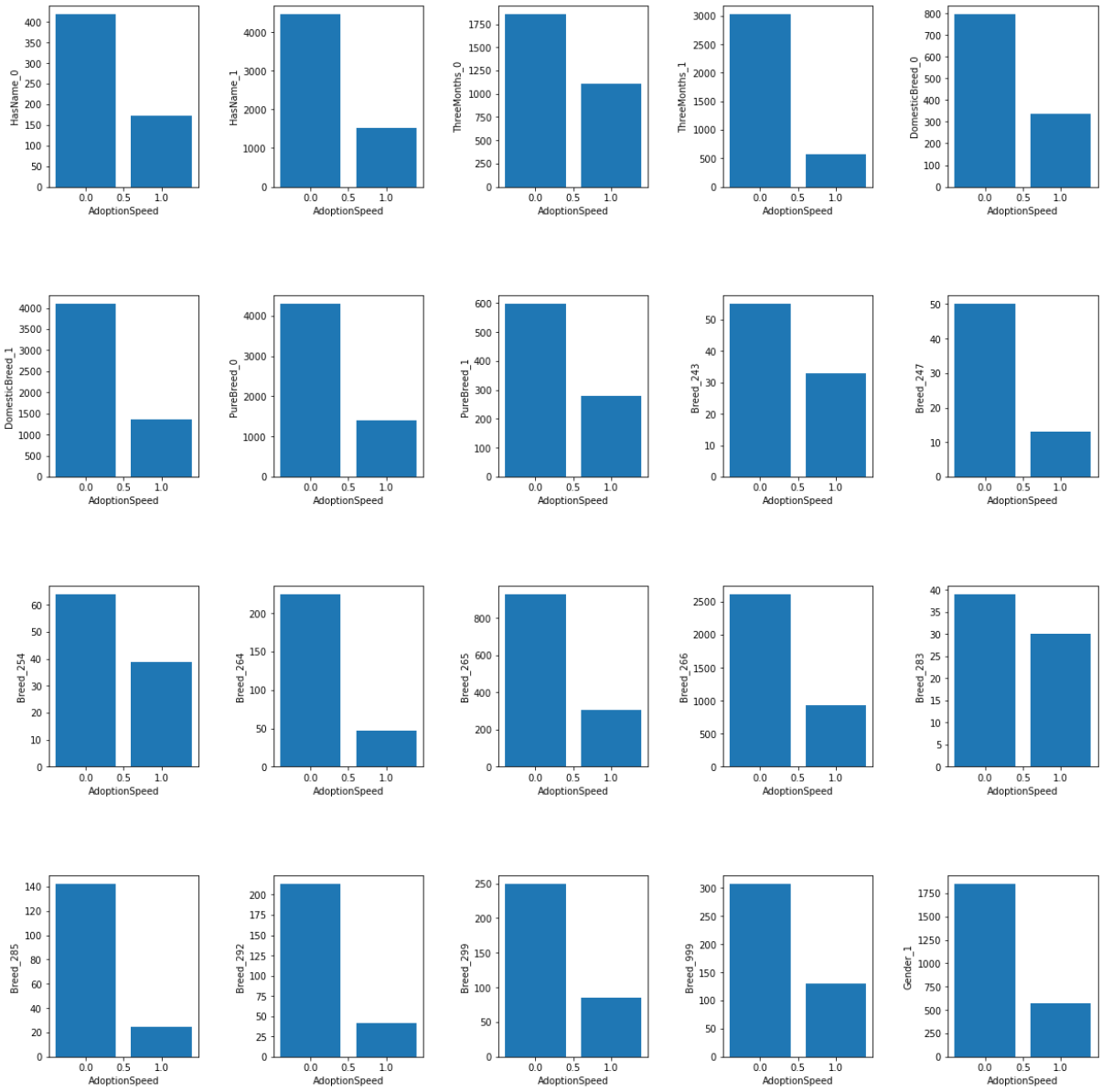


Fig 38.1 Count plots between categorical features and the target variable in the cat adoption dataset

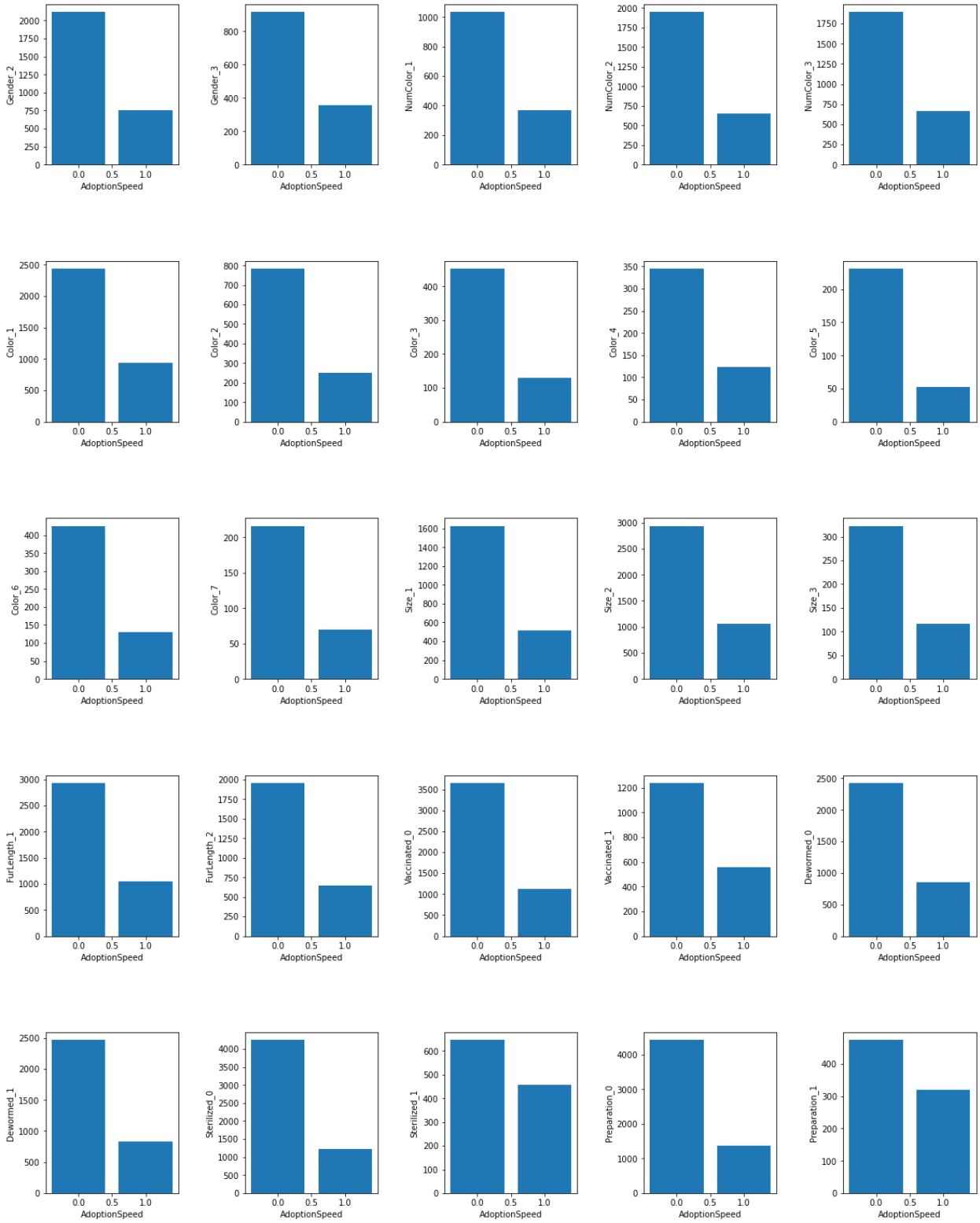


Fig 38.2 Count plots between categorical features and the target variable in the cat adoption dataset

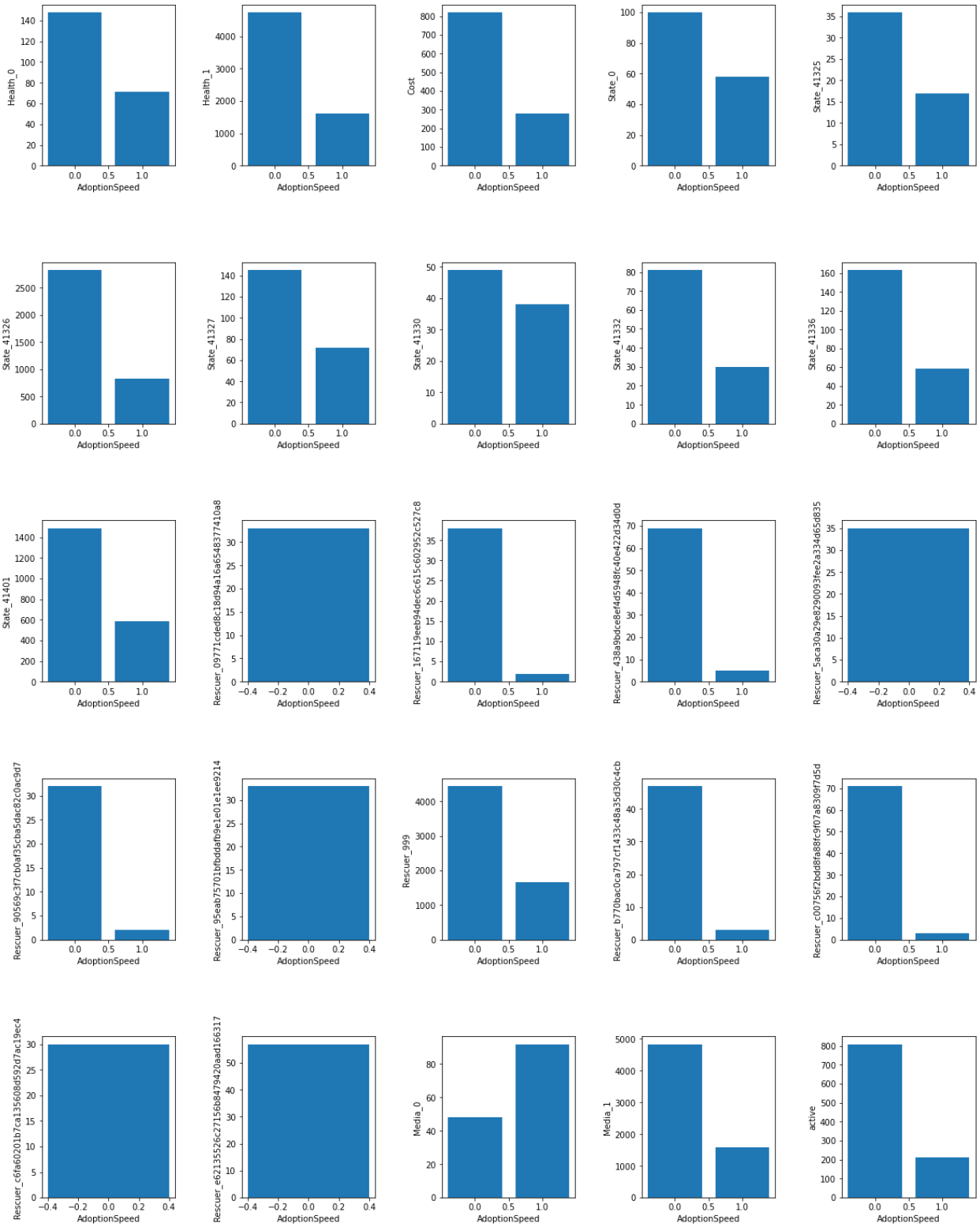


Fig 38.3 Count plots between categorical features and the target variable in the dog adoption dataset

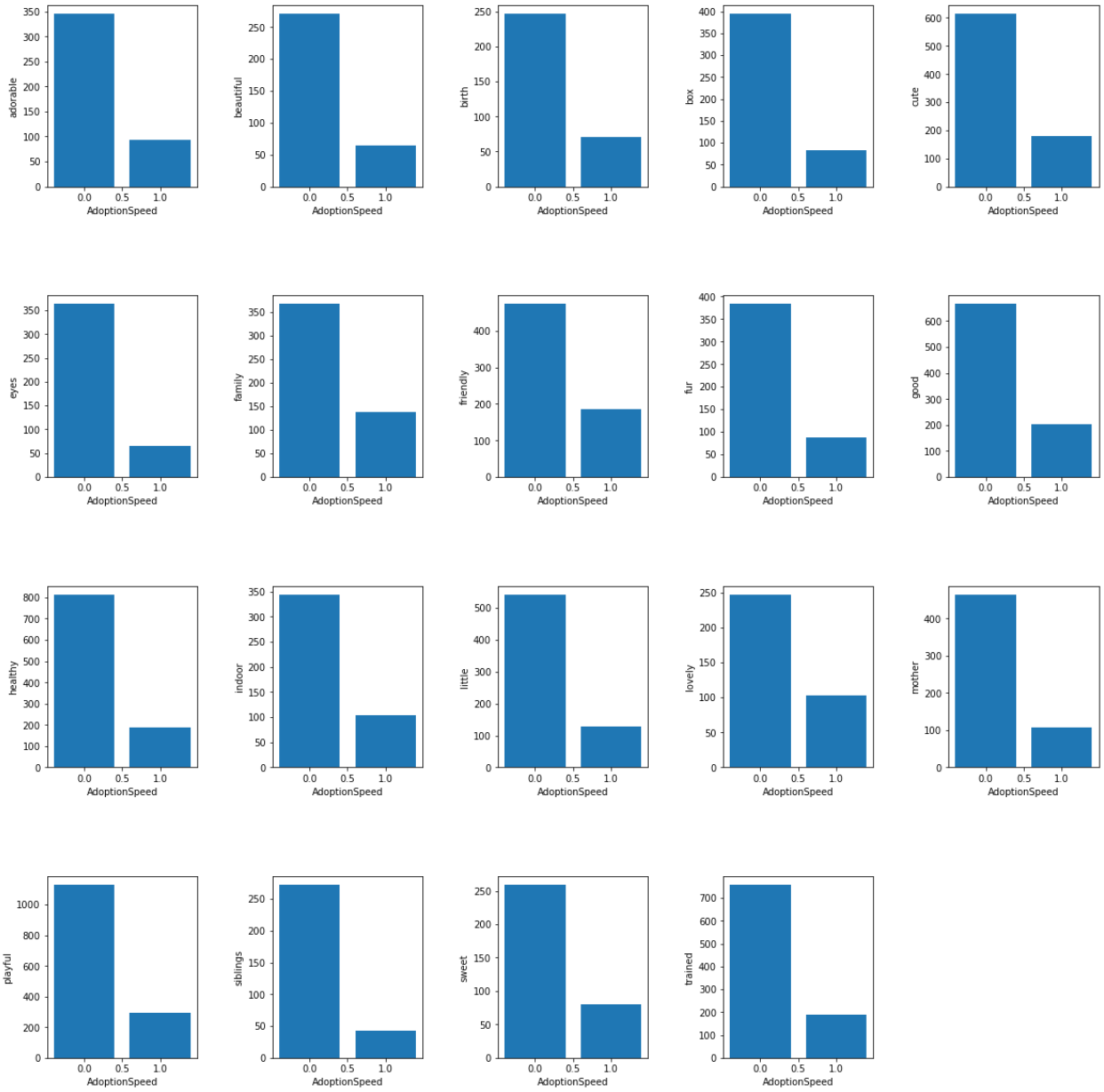


Fig 38.4 Count plots between categorical features and the target variable in the dog adoption dataset

4. Machine Learning

The goal of this report is to build and identify the best machine learning model that can accurately predict the target variable. I will build multiple classification machine learning models and select the best model - one for dog adoption and one for cat adoption.

After data cleaning, feature engineering and processing, the dog adoption dataset contains 7688 observations and the cat adoption dataset contains 6584 observations. The dog adoption dataset has 98 features and 1 target, whilst the cat adoption dataset has 100 features and 1 target.

4.1 Pre-processing

First, I split both of my datasets into train and test sets with the ratio 80:20. Next, I used weight of evidence (WOE) and information value (IV) to perform variable selection. IV uses frequency counting to measure the importance of a feature. IV ranges from 0 to 1, with values closer to 0 having little to no predictive power whilst values close to 1 represent predictors that may be too good to be true. Since there are no hard rules, I chose 0.02 and 0.8 to be the cut off for this problem. I only kept features that had an IV between 0.02 and 0.8. For the dog dataset, 73 features were removed by IV and 25 features remain. For the cat dataset, 79 features were removed by IV and 21 features remain.

Next, I used variance inflation factors (VIF) to detect multicollinearity within the datasets. VIF determines the strength of the correlation between each and every feature. This correlation value is achieved by regressing each feature against every other feature. VIF values range from 1 to infinity, where 1 means the feature is not correlated and infinity means the feature can be fully explained by another feature. Since there are no hard rules, I chose 10 to be the cut off for this problem. For the dog adoption dataset, 6 features were further removed by VIF. For the cat dataset, 7 features were further removed by VIF.

After IV and VIF pre-processing the dog adoption dataset now has 18 features and the cat adoption dataset now has 13 features. Both datasets are now ready to be fed into machine learning models.

4.2 Model Building

Since the target variable is binary, I fitted my datasets to the following classification models:

- Logistic regression
- Random forest
- XGBoost classifier

A pipeline object was created with a standard scaler step to ensure that features contribute to the model equally. I used random search with 10 folds cross validation to tune the model hyper parameters. The fit method runs 10-fold cross validation on each randomly chosen combination of hyper parameters. The combination of hyper parameters that generate the highest mean validation score are stored in the best_params_ and best_score_ attributes. The fit method also trained the full training set by using the best hyper parameters. The predict_proba method predicts the probability of the target value being 1. This is used to evaluate the best model.

Since both datasets have an unequal distribution of classes. I used a cut off based on the initial proportion of the respective datasets to predict the outcome class. For the dog adoption dataset, the proportion between the negative and positive class is 0.29, so any predicted probability above 0.29 outputted by the model is classified as 1 and 0 otherwise. The same procedure was performed for the cat adoption dataset where the cut off is set to 0.26 to reflect its class proportions.

4.2.1 Best Model

4.2.1.1 Dog adoption dataset

Fig 39 shows the summary evaluation metrics for the models trained on the dog adoption dataset. Since the target classes are slightly imbalanced and we care about the positive label most, I used F1-score as the evaluation metric. F1-score keeps a balance between precision and recall which focuses on the metrics of the positive class label. Based on F1-score, we can see that the random forest model is superior. This is further confirmed by looking at the confusion matrix which shows that the random forest model had the highest number of true positives and lowest number of false negatives. The best hyperparameters for this model are:
{'randomforestclassifier__n_estimators': 200, 'randomforestclassifier__min_samples_split': 4, 'randomforestclassifier__max_features': 'sqrt', 'randomforestclassifier__max_depth': 13, 'randomforestclassifier__criterion': 'gini'}

Model	F1-Score	Best Score	ROC AUC	Confusion Matrix
Logistic Regression	0.6176	0.7641	0.797	[[815 254] [146 323]]
Random Forest	0.6553	0.8039	0.8355	[[750 319] [85 384]]
XGBoost	0.5803	0.813	0.8433	[[968 10] [236 233]]

Fig 39 Summary metrics for models built on the dog adoption dataset

This model will be used for further analysis in subsequent sections of this report.

4.2.1.2 Cat adoption dataset

Fig 40 shows the summary evaluation metrics for the best models trained on the cat adoption dataset. We can see that random forest is the superior model when comparing F1-scores, best score and ROC AUC score. We can also see that the random forest model had the highest number of true positives and lowest number of false negatives from its confusion matrix. The best hyper-parameters for this model are:
{'randomforestclassifier__n_estimators': 1400, 'randomforestclassifier__min_samples_split': 3, 'randomforestclassifier__max_features': 'auto', 'randomforestclassifier__max_depth': 9, 'randomforestclassifier__criterion': 'gini'}

Model	F1-Score	Best Score	ROC AUC	Confusion Matrix
Logistic Regression	0.522	07438	0.7316	[[679 300] [112 225]]
Random Forest	0.5355	0.7761	0.7725	[[646 333] [92 245]]
XGBoost	0.4781	0.7687	0.7717	[[829 150] [184 153]]

Fig 40 Summary metrics for models built on the cat adoption dataset

This model will be used for further analysis in subsequent sections of this report.

5. Feature Importance and SHAP

5.1 Feature importance

Feature importance quantifies how well each input feature is at predicting the target variable. The higher the score, the more important the feature is for predicting the target variable. For a random forest classifier, feature importance describes how much a single feature contributes to the tree's total impurity reduction averaged over all the trees. The relative importance of all the input features sum up to 1.

5.1.1 Dog adoption dataset

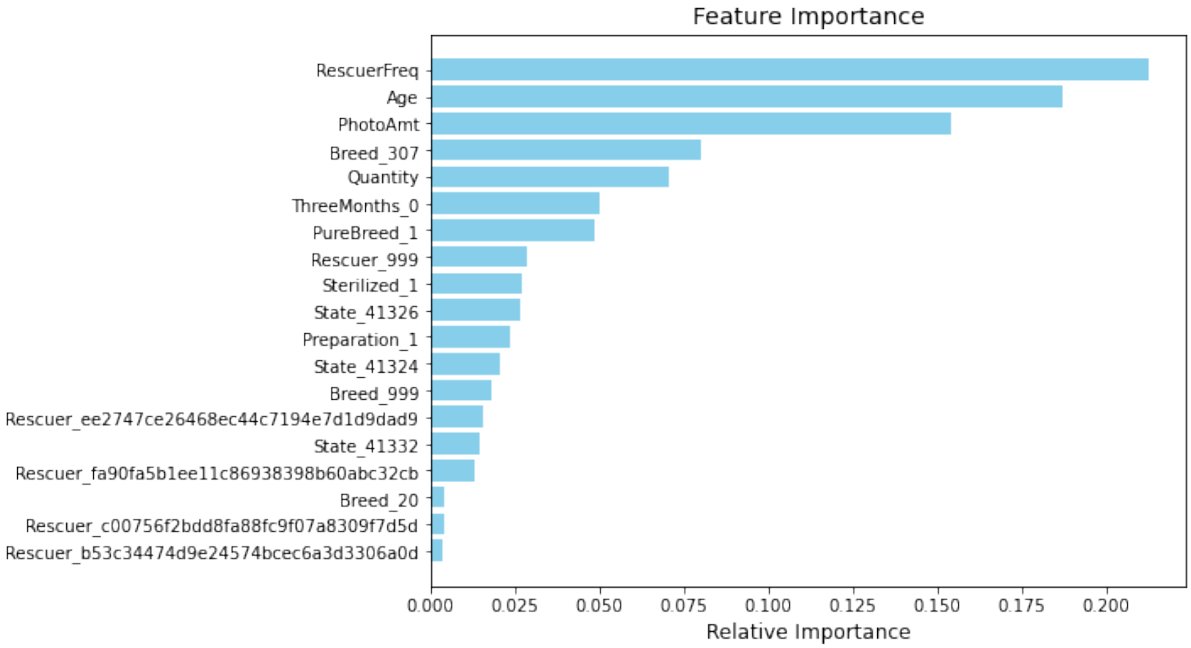


Fig 41 Feature importance plot for the dog adoption dataset

Fig 41 shows the feature importance for each input feature in the dog adoption dataset. We can see that 'RescuerFreq' is an extremely important feature. 'RescuerFreq' contributes to over 20% of the target variable prediction. On the other hand, the feature 'Rescuer_b53c34474d9e24574bcec6a3d3306a0d' only contributes 0.33% to the target variable prediction.

5.1.2 Cat adoption dataset

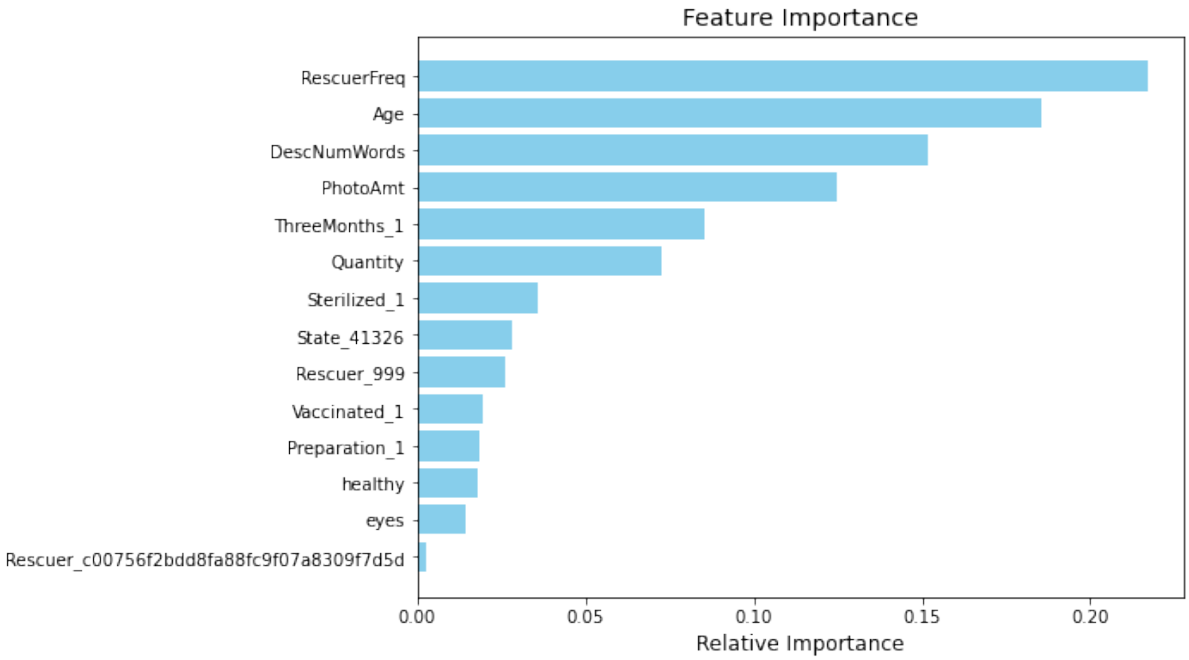


Fig 42 Feature importance plot for the cat adoption dataset

Fig 42 shows the feature importance for each input feature in the cat adoption dataset. We can see that similar to the dog adoption dataset, 'RescuerFreq' is also the most important feature, contributing to over 20% of the target variable prediction as well. Similar to the dog adoption dataset but for a different rescuer, we can see that the individual rescuer feature 'Rescuer_c00756f2bdd8fa88fc9f07a8309f7d5d' only contributes 0.26% to the target variable prediction.

5.2 SHAP

I performed SHAP (Shapley additive explanations) analysis to further understand the impact of each feature. SHAP measures the influence of a feature by comparing model predictions with and without the feature being utilized.

5.2.1 Summary plot

SHAP summary plots display the positive (red) and negative (blue) relationships between each input feature and the target variable.

5.2.1.1 Dog adoption dataset

Fig 43 shows the summary plot for the best model on the dog adoption dataset. It demonstrates the relationship of each feature in the model with the target variable (adoption speed). Since our target variable is binary (0: adoption within 100 days, 1: no adoption after 100 days), a SHAP value of below 0 means the feature is driving the target value towards 0 and a SHAP value of above 0 means the feature is driving the target value towards 1. For example, we can see that the feature 'ThreeMonths_0' has a clear distinction between low and high values. We can see that low 'ThreeMonths_0' values correspond to negative SHAP values and high 'ThreeMonths_0' values correspond to positive SHAP values. Since 'ThreeMonths_0' is also binary (0: 3 months or younger, 1: older than 3 months), we can interpret the relationship as dogs that are older than 3 months are less likely to be adopted within 100 days, whilst dogs that are 3 months or younger are more likely to be adopted within 100 days. This makes sense as puppies are generally more desirable by pet owners.

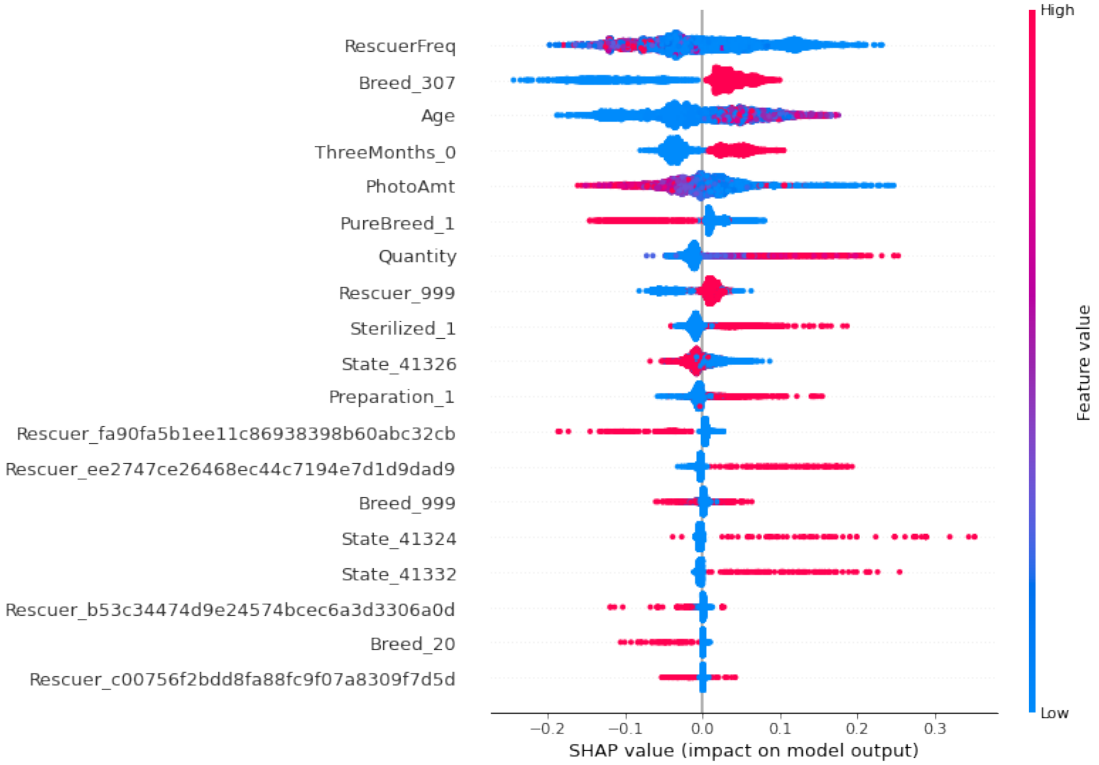


Fig 43 SHAP summary plot for the best model on the dog adoption dataset

Based on fig 43, we can identify a few trends that generally increases the probability of adoption for dogs:

- Rescues with a high rescue frequency
- Dogs not having a dominant breed of mixed breed
- Young dogs
- Dogs that are 3 months or younger
- Dogs with a lot of photos on their file
- Purebred dogs
- Listings with a low quantity of dogs
- Dogs that are rescued by the top 10 most popular rescuers
- Unsterilized dogs
- Dogs located in state 41326 (Selangor)
- Unprepared dogs (not sterilized, vaccinated or dewormed)
- Dogs rescued by rescuer fa90fa5b1ee11c86938398b60abc32cb
- Dogs not rescued by rescuer ee2747ce26468ec44c7194e7d1d9dad9
- Dogs not located in state 41324 (Melaka)
- Dogs not located in state 41332 (Negeri Sembilan)
- Dogs rescued by rescuer b53c34474d9e24574bcec6a3d3306a0d
- Dogs being of breed 20 (Rottweiler)

5.2.1.2 Cat adoption dataset

Fig 44 shows the summary plot for the best model trained on the cat adoption dataset. We can see that the general trend for many features e.g. 'RescuerFreq', 'Age' and 'Quantity' are very similar to the ones for dog adoption.

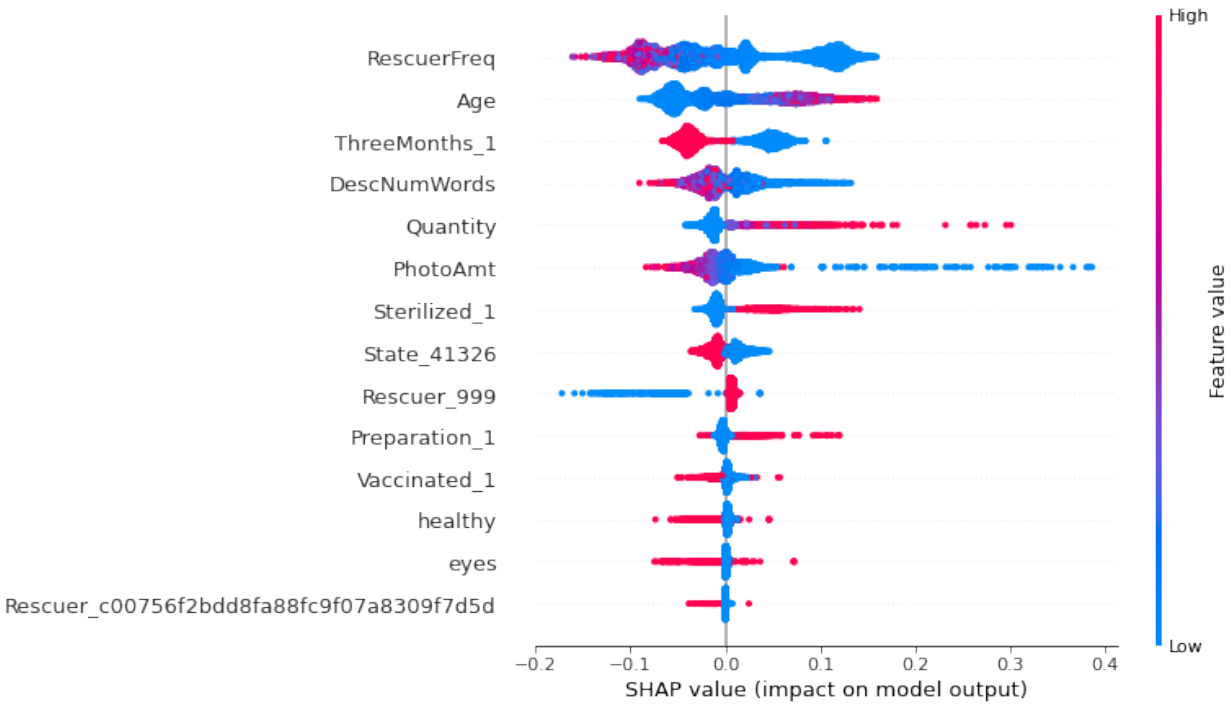


Fig 44 SHAP summary plot for the best model on the cat adoption dataset

Based on fig 44, we can identify a few trends that generally increases the probability of adoption for cats:

- Rescues with a high rescue frequency
- Young cats
- Cats that are 3 months or younger
- Profile descriptions with a high number of words
- Listings with a low quantity of cats
- Cats with a lot of photos on their file
- Unsterilized cats
- Cats located in state 41326 (Selangor)
- Cats that are rescued by the top 10 most popular rescuers
- Unprepared cats (not sterilized, vaccinated or dewormed)
- Vaccinated cats
- Profile descriptions containing the word ‘healthy’
- Profile descriptions containing the word ‘eyes’
- Cats that are rescued by rescuer c00756f2bdd8fa88fc9f07a8309f7d5d

5.2.2 Dependence plots

Dependence plots shows the marginal effect two features have on the predicted outcome of the machine learning model.

5.2.2.1 Dog adoption dataset

5.2.2.1.1 RescuerFreq

Fig 45.1 demonstrates the relationship between 'RescuerFreq' and the target variable as well as 'RescuerFreq's interaction with 'State_41326'. There is an initial downward trend between RescuerFreq and the target variable (up until ~ 80). Beyond 80, the relationship between 'RescuerFreq' and the target variable seem to be constant and have SHAP values between -0.2 to 0.1. This means that dogs are more likely to be adopted within 100 days when they are rescued by a rescuer that has rescued a high number of other dogs. Dogs are most likely to be adopted if they are adopted by a rescuer that has adopted at least 80 other dogs. This makes sense as adopters are more likely to adopt from a reputable rescue that has a lot of dogs up for adoption.

Fig 45.1 also shows the interaction between 'RescuerFreq' and 'State_41326'. We can see that most dogs are located in State_41326 (red), Selangor. This shows that the state has the most dogs up for adoption. We can also see that all the rescuers with the highest rescuer frequency are located in Selangor. The rescuers with high rescuer frequencies are likely to be adoption centres. This shows us that most adoption centres are located in Selangor as well. Being the state with the largest population and economy (in terms of GDP), it makes sense that the most developed Malaysian state have the most facilities to alleviate the stray dog problem.

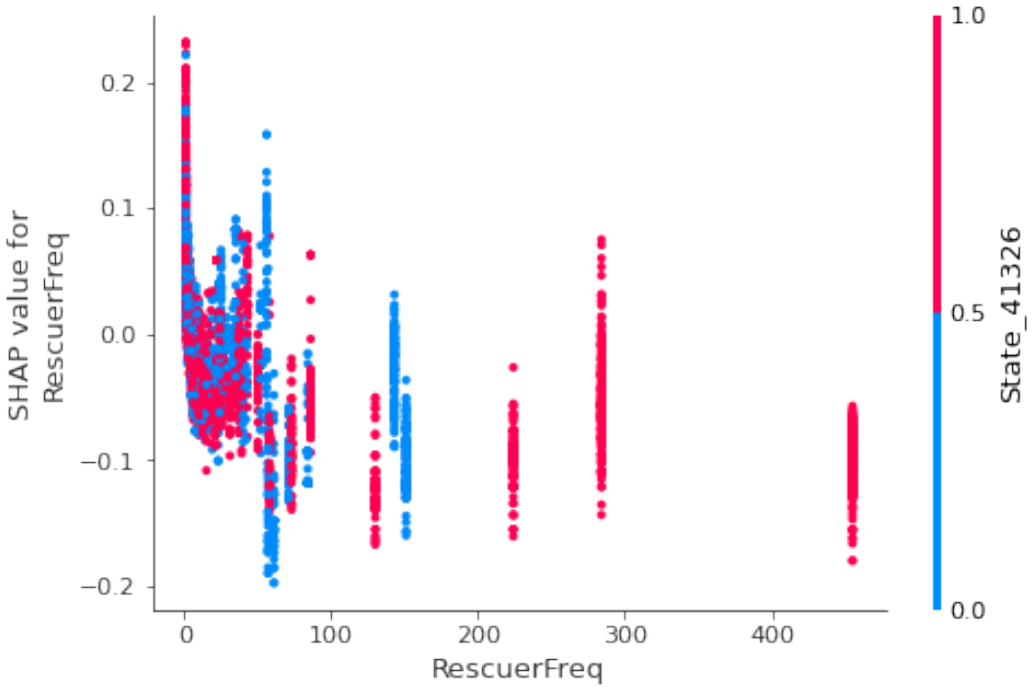


Fig 45.1 SHAP 'RescuerFreq' dependence plot with interaction to 'State_41326'

5.2.2.1.2 Breed_307

Fig 45.2 demonstrates the relationship between 'Breed_307' and the target variable as well as 'Breed_307's interaction with 'PhotoAmt'. We can see that dogs that are not of 'Breed_307' (i.e. have an identified dominant breed), have a negative shap value, whilst dogs that have 'Breed_307' as their dominant breed have a positive shap value. This means that dogs not of 'Breed_307' are more likely to be adopted within 100 days, whilst dogs that are of 'Breed_307' are less likely to be adopted within 100 days. This makes sense as dogs with a recognizable breed are rarer and more likely to be adopted.

Looking more closely at dogs that are not of 'Breed_307' (mixed breed), we can see that a lower 'PhotoAmt' (blue) generally corresponds to a lower SHAP value. This means that dogs with less photos are more likely to be adopted within 100 days. This could be because there just wasn't enough time to upload photos for dogs that were quickly adopted. On the other hand, for dogs that are of 'Breed_307' (mixed breed), we can see that those with a higher photo (red) amount have a lower SHAP value. This means that dogs of 'Breed_307' are more likely to be adopted within 100 days if they have more photos uploaded. This makes sense as people are more likely to adopt a dog if they know what it looks like, especially if they're labelled as mixed breed which has a lot of variations.

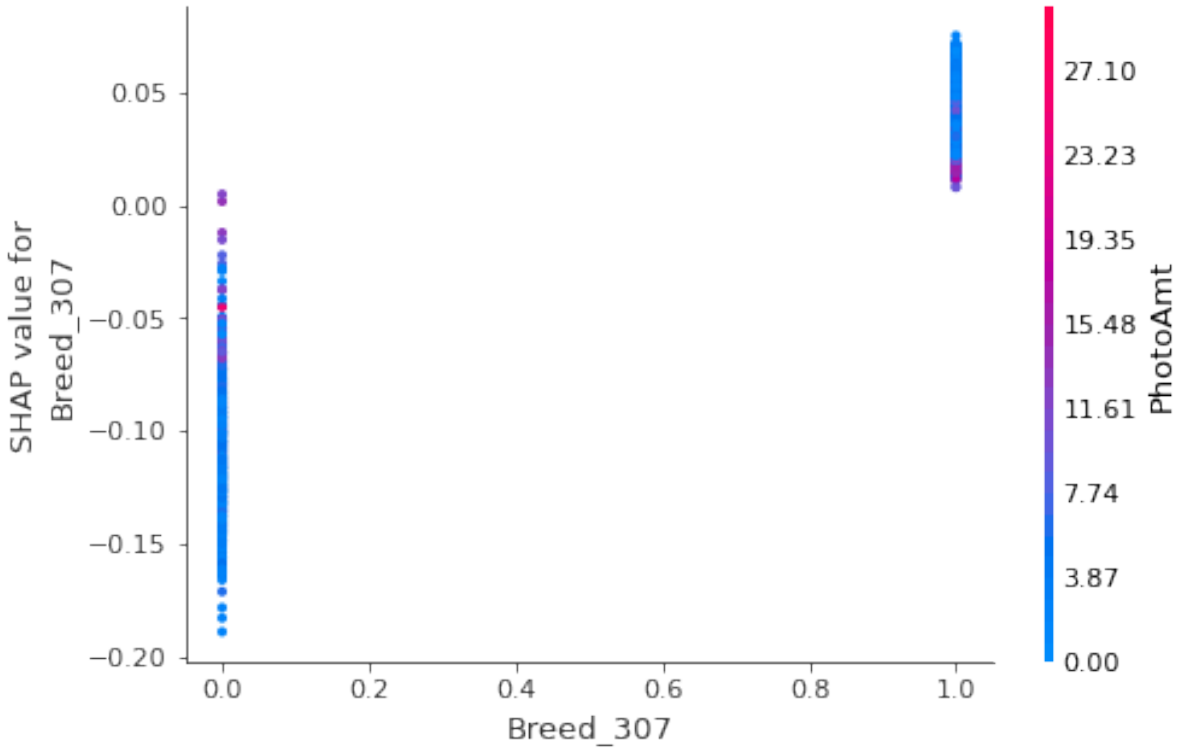


Fig 45.2 SHAP 'Breed_307' dependence plot with interaction to 'PhotoAmt'

5.2.2.1.3 Age

Fig 45.3 demonstrates the relationship between 'Age' and the target variable as well as 'Age's interaction with 'Breed_307'. From Fig 45.3 we can see that there is a general upwards trend until around 8. This tells us that the older the dog the less likely it will be adopted within 100 days. However, we do see the trend plateau after 8 months, where the SHAP values consistently lie between -0.05 to 0.2. This suggests that dogs between ages 8 to 80 months generally have a similar adoption speed. This makes sense as most adopters want to adopt dogs at a younger age so that they can grow old together.

Looking at the interaction between 'Age' and 'Breed_307', we can see that there is a flip in the trend of 'Breed_307' at 3 months. We can see that for dogs 3 months or younger, dogs that have a dominant breed of mixed breed (red) generally have a lower SHAP value than pure breed (blue) dogs of the same age. This means that mixed breed dogs are slightly more likely to be adopted within 100 days than pure breed ones that are 3 months or younger. However, we see this trend flip for dogs older than 3 months. For older dogs, not having a dominant breed of mixed breed (blue) drives the likelihood of being adopted within 100 days more than having a dominant breed of mixed breed (red).

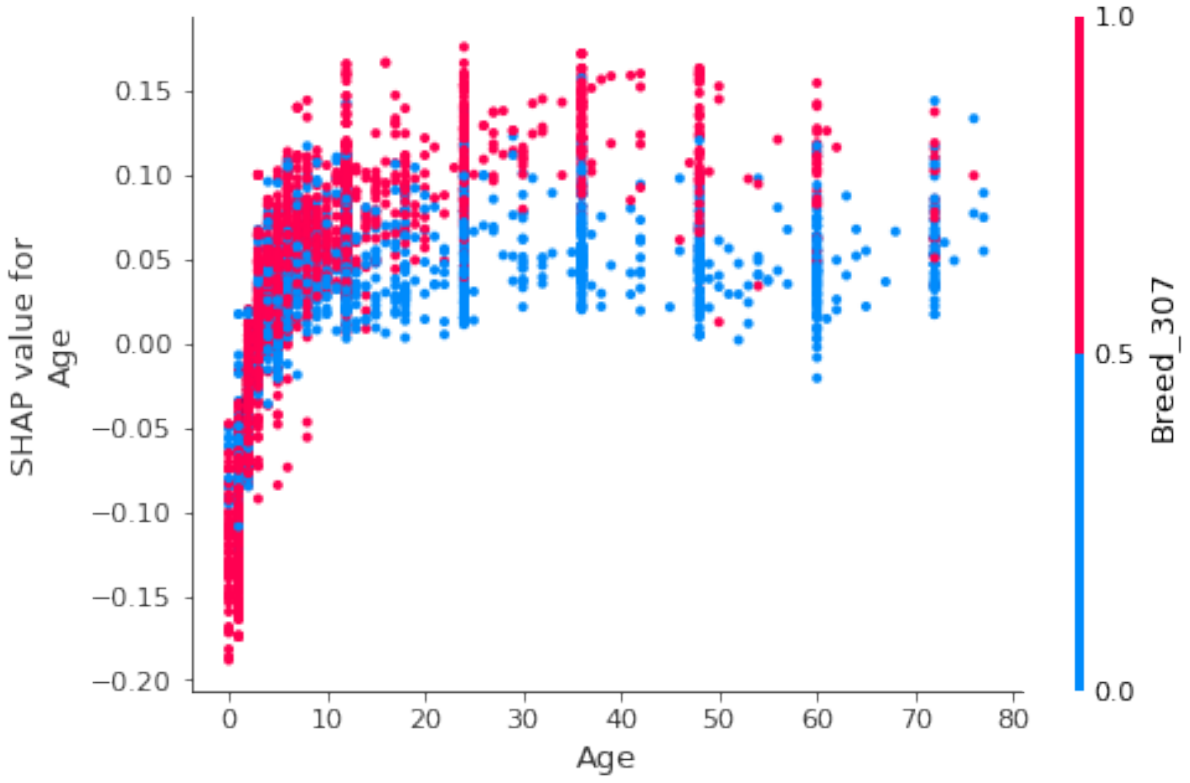


Fig 45.3 SHAP 'Age' dependence plot with interaction to 'Breed_307'

5.2.2.1.4 ThreeMonths_0

Fig 45.4 demonstrates the relationship between 'ThreeMonths_0' and the target variable as well as 'ThreeMonth_0's interaction with 'Breed_307'. It is evident that dogs that are 3 months or younger are more likely to be adopted within 100 days, whilst dogs that are older than 3 months are less likely to be adopted within 100 days. This makes sense as dogs 3 months or younger are puppies and are much more desirable.

From fig 45.4 we can see that there are a lot more dogs with the dominant breed of mixed breed (red) than other breeds. Most dogs that are 3 months or younger are of mixed dominant breed. A 3 month or younger dog with mixed dominant breed has a lower SHAP value than a 3 month or younger dog with a pure dominant breed. This means that for dogs that are 3 months or younger, they are more likely to be adopted if they're of mixed dominant breed than if they were not. On the other hand, for dogs that are older than 3 months, those with a dominant mixed breed tend to have a higher SHAP values than those with a dominant pure breed. This means that for dogs that are older than 3 months, they are more likely to be adopted if they have a dominant pure breed. This makes sense as mixed dominant breed puppies tend to exhibit similar physical traits of their pure breed counterpart, they are also less costly and are still extremely cute, thus making them more desirable than pure dominant breed puppies. However, for older dogs, it is very easy to distinguish between mixed breed and pure breed dogs, thus making pure breed dogs much more desirable.

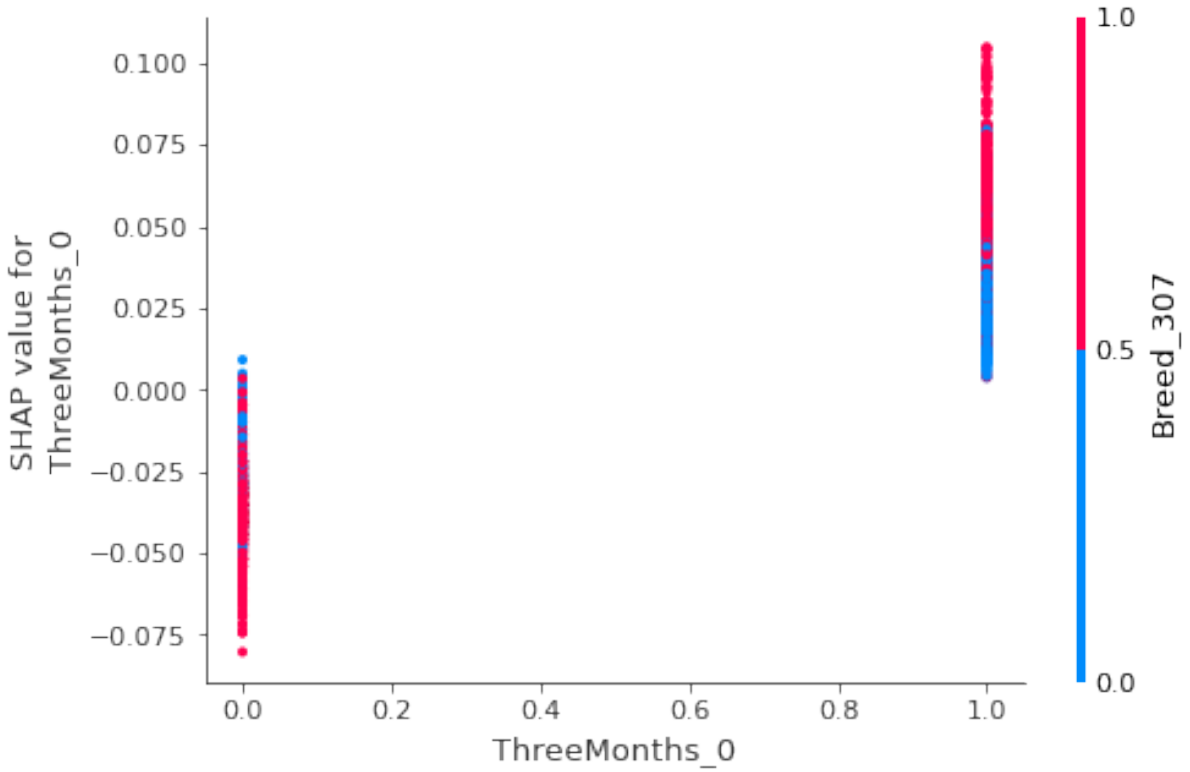


Fig 45.4 SHAP 'ThreeMonths_0' dependence plot with interaction to 'Breed_307'

5.2.2.1.5 PhotoAmt

Fig 45.5 demonstrates the relationship between 'PhotoAmt' and the target variable as well as 'PhotoAmt's interaction with 'Quantity'. We can see that there is a clear decline in SHAP values as the number of photos increases up to 10. As the number of photos increase from 11 to 14, there appears to be a slight increase in SHAP values too. After 14 photos, the SHAP values remain constant ranging from -0.1 to 0.15.

It is evident that a photo amount of 0 results in a SHAP value above 0.05, this means that when there are no photos attached on the dog listing, the dog is less likely to be adopted within 100 days. We can see that the optimal number of photos to include in a dog listing is between 7-10. Fig 45.5 suggests that adding more than 11 photos, may not necessarily increase the chances for a dog to be adopted within 100 days.

Now looking at the interaction between 'PhotoAmt' and 'Quantity,' we can see that most listings have a low quantity (i.e. most dogs are listed individually). Listings that have a high quantity tend to have a higher number of photos attached. This makes sense as a higher quantity indicates more dogs that can be photographed. We can see that listings with a high quantity are more likely to be adopted within 100 days (lower SHAP value) if they have more photos attached.

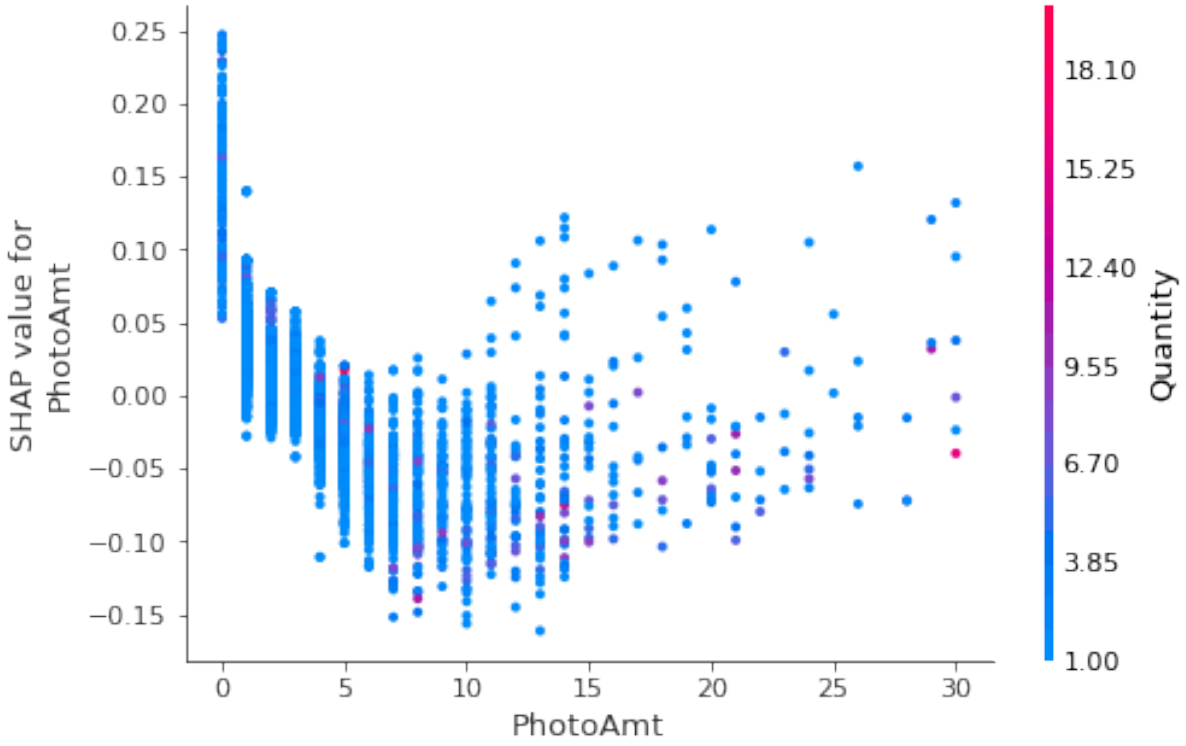


Fig 45.5 SHAP 'PhotoAmt' dependence plot with interaction to 'Quantity'

5.2.2.1.6 PureBreed_1

Fig 45.6 demonstrates the relationship between 'PureBreed_1' and the target variable as well as 'PureBreed_1's interaction with 'Breed_307'. Fig 45.6 indicates that purebred dogs are more likely to be adopted within 100 days than mixed breed dogs, as purebred dogs have negative SHAP values whilst mixed breed dogs have positive SHAP values. This makes sense are pure breed dogs are rarer and more desirable.

All purebred dogs don't have a dominant breed of 307 (mixed breed), this makes sense as this is how the feature was derived. Amongst mixed breed dogs, we can see that those with a dominant breed of 307 (red) are more likely to be adopted with 100 days since they have a lower SHAP value. This means that amongst mixed breed dogs, those that do not have an identified dominant breed are more likely to be adopted. This is likely due to their lower fee as purebred dogs tend to have a higher adoption fee. We can also see that there are outliers to this rule as there are several blue points below the red points.

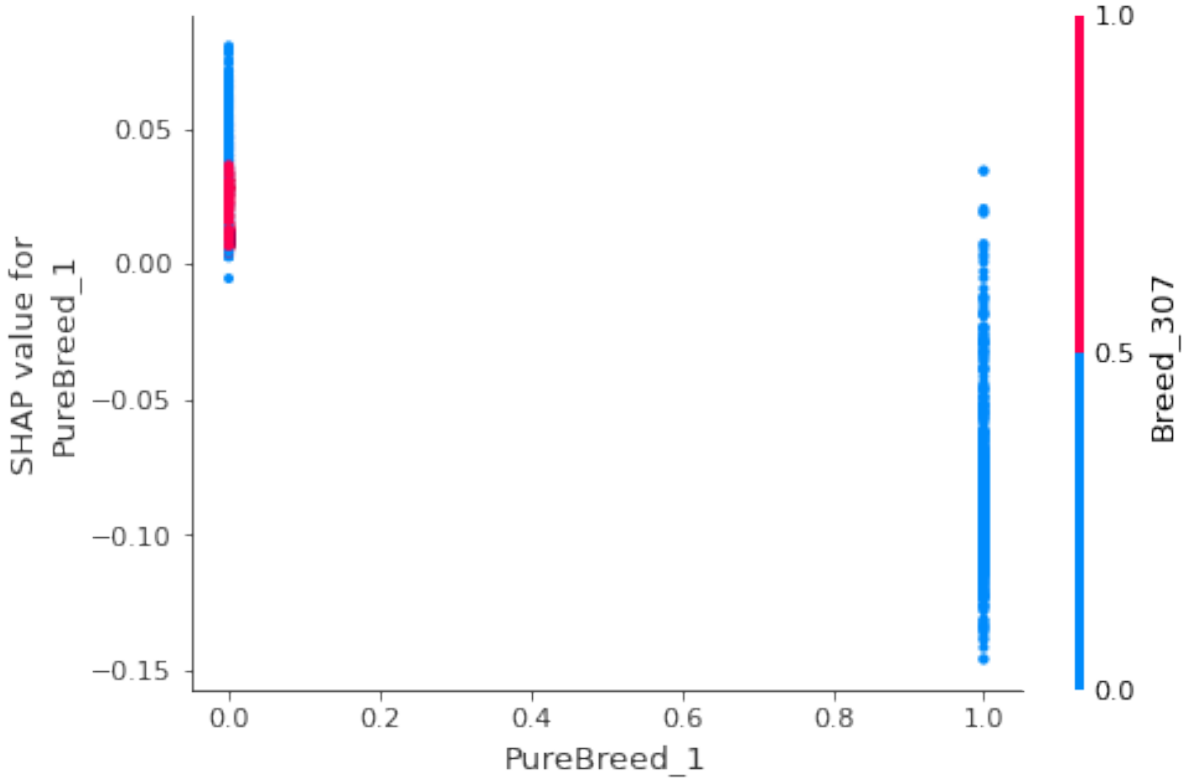


Fig 45.6 SHAP 'PureBreed_1' dependence plot with interaction to 'Breed_307'

5.2.2.1.7 Quantity

Fig 45.7 demonstrates the relationship between 'Quantity' and the target variable as well as 'Quantity's interaction with 'Rescuer_c00756f2bdd8fa88fc9f07a8309f7d5d'. In general, we can see that quantity 1 has a negative SHAP value whilst quantity 2 may have a positive or negative value. Quantities above 2 tend to have a positive SHAP value. Quantities between 4 and 20 generally have SHAP values between 0 and 0.25. This means that dog listings with smaller quantities are more likely to be adopted within 100 days than those of larger quantities. This makes sense as it takes longer for multiple dogs to be adopted.

We can see that there's a very small number of dogs rescued by 'Rescuer_c00756f2bdd8fa88fc9f07a8309f7d5d' (red). The dogs that are rescued by 'Rescuer_c00756f2bdd8fa88fc9f07a8309f7d5d' are in quantities of between 4 – 9. In general, we can see that dogs rescued by this rescuer have a higher SHAP value compared to the same quantity of dogs not rescued by this rescuer. This means that dogs rescued by 'Rescuer_c00756f2bdd8fa88fc9f07a8309f7d5d' are slightly less likely to be adopted within 100 days than the same quantity of dog rescued by other rescues.

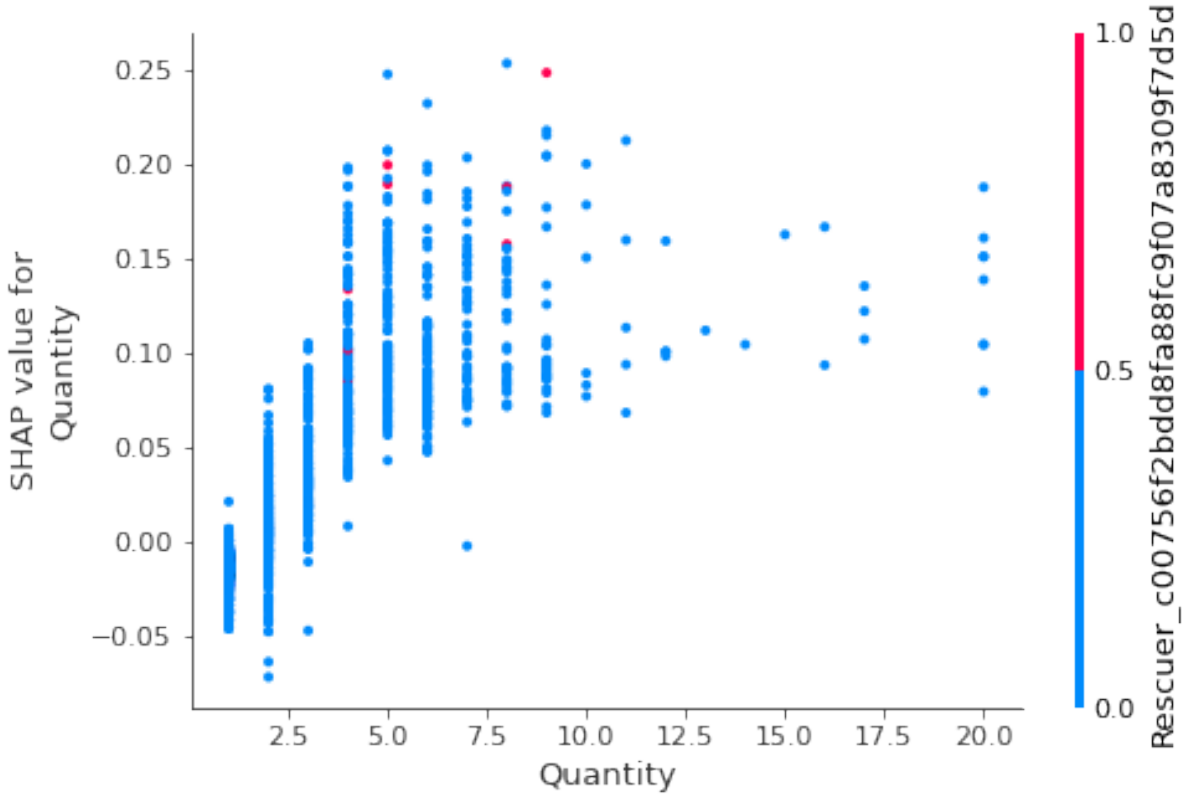


Fig 45.7 SHAP 'Quantity' dependence plot with interaction to 'Rescuer_c00756f2bdd8fa88fc9f07a8309f7d5d'

5.2.2.1.8 Preparation_1

Fig 45.8 demonstrates the relationship between 'Preparation_1' and the target variable as well as 'Preparation_1's interaction with 'PhotoAmt'. Based on Fig 45.8, we can see that there are more fully prepared dogs than not fully prepared. Interestingly, dogs that are fully prepared have a very large range of SHAP values, with a lot of them being above 0, thus suggesting that a fully prepared dog is less likely to be adopted within 100 days. For a dog that is not fully prepared, the range of SHAP values also encompass 0 but there are more values below 0, suggesting that not being fully prepared may increase the chances of a dog being adopted within 100 days.

Looking at the interaction between 'Preparation_1' and 'PhotoAmt', we can see that for dogs that aren't fully prepared, they are more likely to be adopted if there are more photos on file. On the other hand, for dogs that are fully prepared, they are actually more likely to be adopted within 100 days if there are less photos attached.

These trends could be explained by the length of stay of he dog in the rescue. Since dogs that stay at a rescue longer are more likely to be fully prepared and have their file updated with photos. On the other hand, if a dog is quickly adopted, there is less time to prepare the dog and add photos to their file.

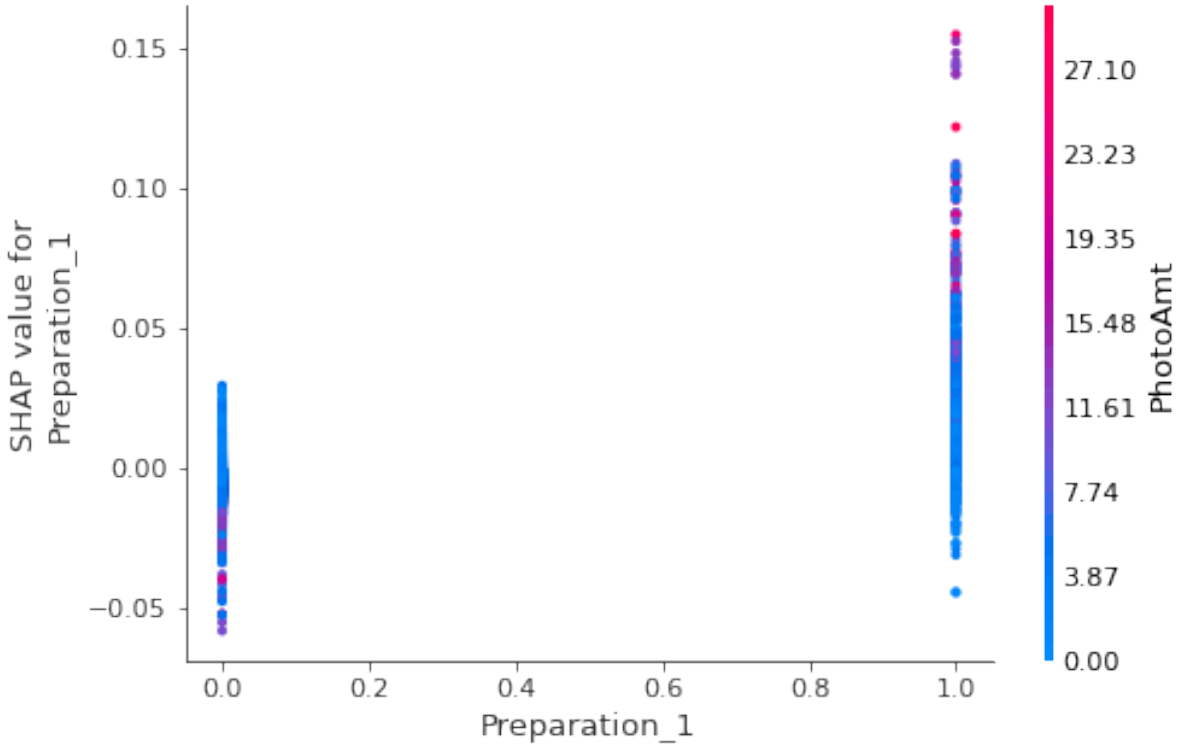


Fig 45.8 SHAP 'Preparation_1' dependence plot with interaction to 'PhotoAmt'

5.2.2.1.9 State_41326

Fig 45.9 demonstrates the relationship between 'State_41326' and the target variable as well as 'State_41326's interaction with 'ThreeMonths_0'. Based on Fig 45.9, we can see that there are more dogs located outside of state 41326 (Selangor) than within. Interestingly, dogs that are located outside of Selangor have a very large range of SHAP values, with a lot of them being above 0, thus suggesting that a dog rescued outside of Selangor is less likely to be adopted within 100 days. For a dog that is rescued in Selangor, the range of SHAP values also encompass 0 but most data points have negative SHAP values, indicating that dogs rescued in Selangor are more likely to be adopted within 100 days.

Looking at the interaction between 'State_41326' and 'ThreeMonths_0', we can see that for dogs rescued outside of Selangor, dogs that are 3 months or younger are more likely to be adopted within 100 days. On the other hand, for dogs rescued inside Selangor, dogs older than 3 months are more likely to be adopted within 100 days.

In summary. fig 45.9 tells us that dogs are usually more adoptable in Selangor than in other states. Even though dogs that are 3 months or younger are usually more desirable, older dogs appear to be more adoptable than younger ones in Selangor.

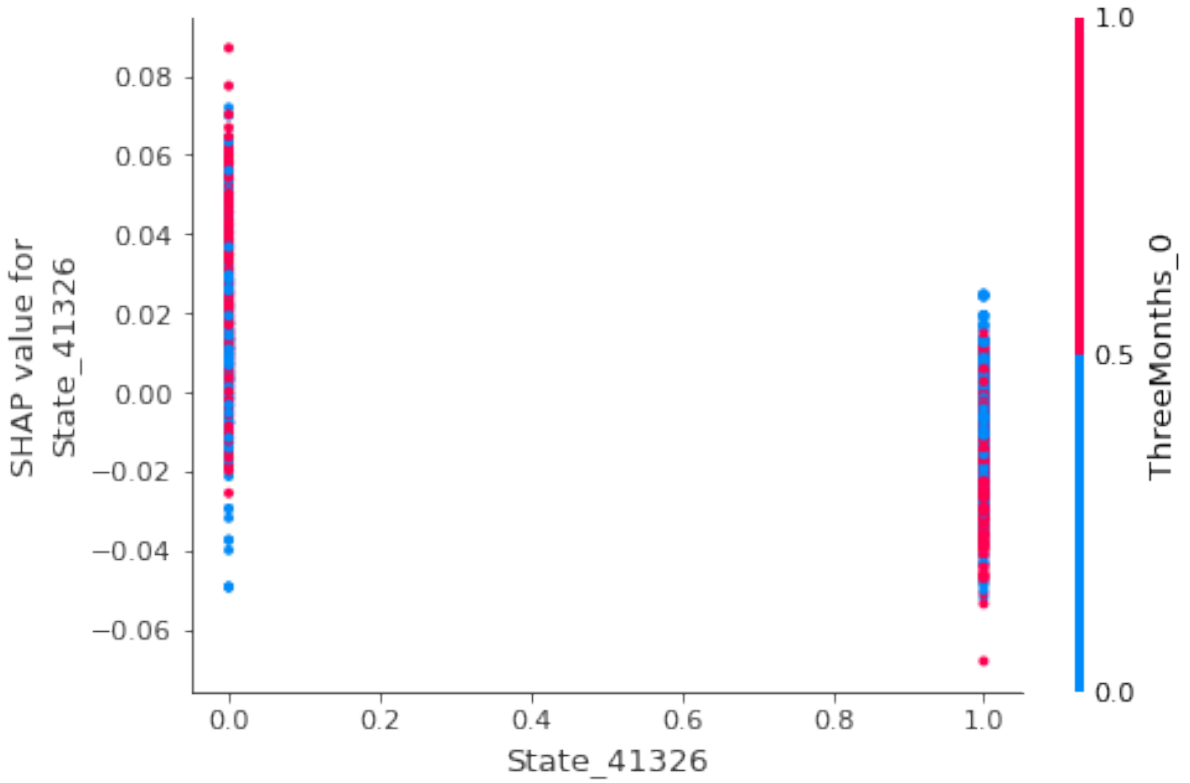


Fig 45.9 SHAP 'State_41326' dependence plot with interaction to 'ThreeMonths_0'

5.2.2.1.10 State_41324

Fig 45.10 demonstrates the relationship between 'State_41324' and the target variable as well as 'State_41324's interaction with 'ThreeMonths_0'. We can see that most dogs are located in State_41324 (Melaka). Dogs rescued in Melaka almost all have a positive SHAP value suggesting that they are less likely to be adopted within 100 days. Dogs that aren't located in State 41324 are more likely to be adopted within 100 days based on the slightly negative SHAP values.

We can see that sterilizing a dog from State 41324 will increase its chances of being adopted within 100 days, based on the lower SHAP values. For dogs that are located outside of State 41324, sterilization will actually decrease the chance of the dog being adopt.

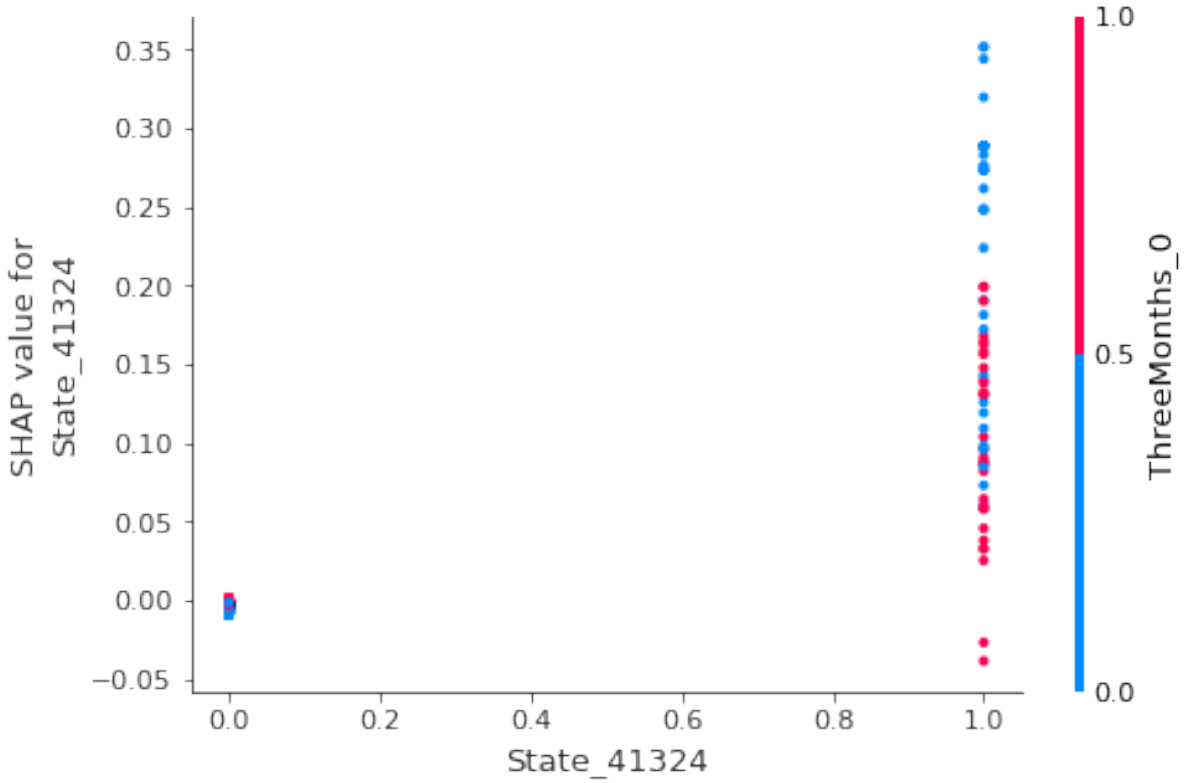


Fig 45.10 SHAP 'State_41324' dependence plot with interaction to 'ThreeMonths_0'

5.2.2.1.11 State_41332

Fig 45.11 demonstrates the relationship between 'State_41332' and the target variable as well as 'State_41332's interaction with 'Sterilized_1'. Dogs located in State_41332 (Negeri Sembilan) are less likely to be adopted within 100 days, based on the positive SHAP values. Dogs that aren't located in State 41332 are more likely to be adopted within 100 days based on the slightly negative SHAP values.

We can see that sterilizing a dog from State 41332 will increase its chances of being adopted within 100 days, based on the lower SHAP values. For dogs that are located outside of State 41332, sterilization will actually decrease the chance of the dog being adopt.

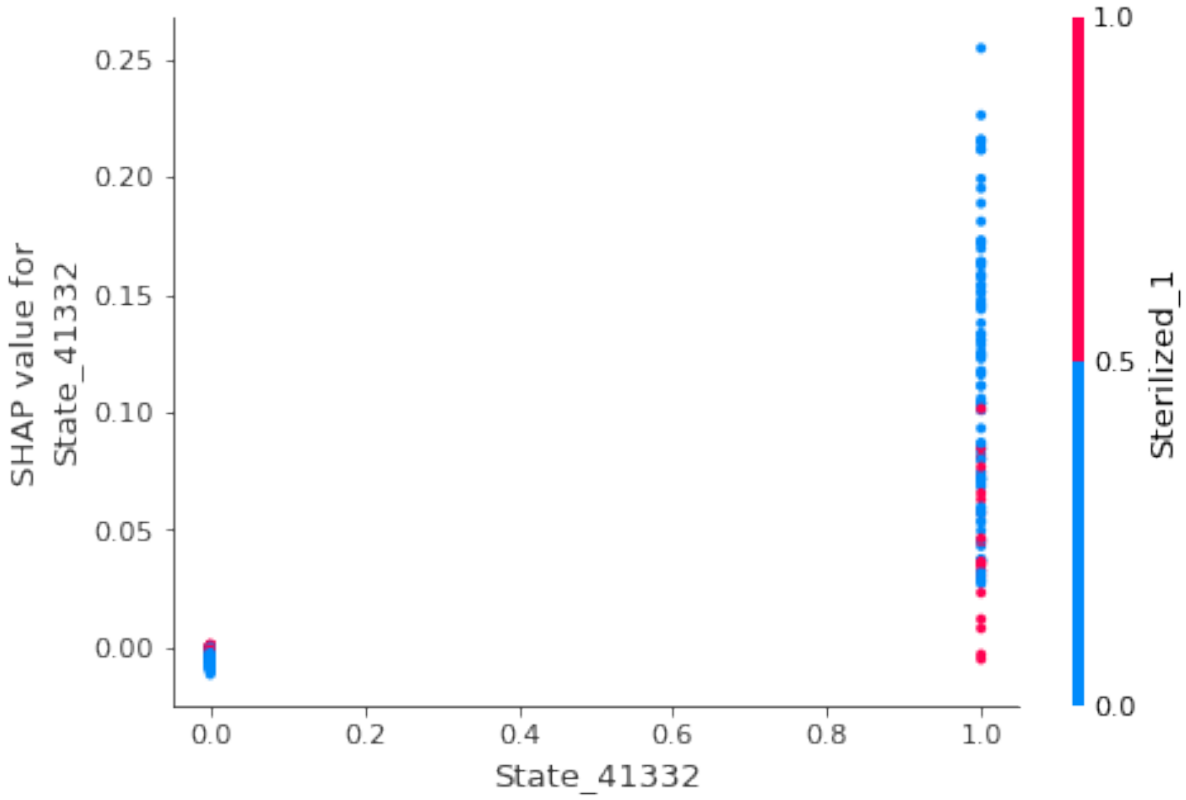


Fig 45.11 SHAP 'State_41332' dependence plot with interaction to 'Sterilized_1'

5.2.2.1.12 Rescuer_fa90fa5b1ee11c86938398b60abc32cb

Fig 45.12 demonstrates the relationship between 'Rescuer_fa90fa5b1ee11c86938398b60abc32cb' and the target variable as well as 'Rescuer_fa90fa5b1ee11c86938398b60abc32cb's interaction with 'Sterilized_1'. We can see that dogs rescued by Rescuer_fa90fa5b1ee11c86938398b60abc32cb are more likely to be adopted within 100 days, based on the negative SHAP values. Dogs rescued by other rescuers are less likely to be adopted within 100 days, based on the range of more positive SHAP values.

We can see that sterilizing a dog from 'Rescuer_fa90fa5b1ee11c86938398b60abc32cb' will increase its chances of being adopted within 100 days, based on the even lower SHAP values. For dogs that are rescued by other rescuers, sterilization will actually decrease the chance of the dog being adopt, represented by higher SHAP values.

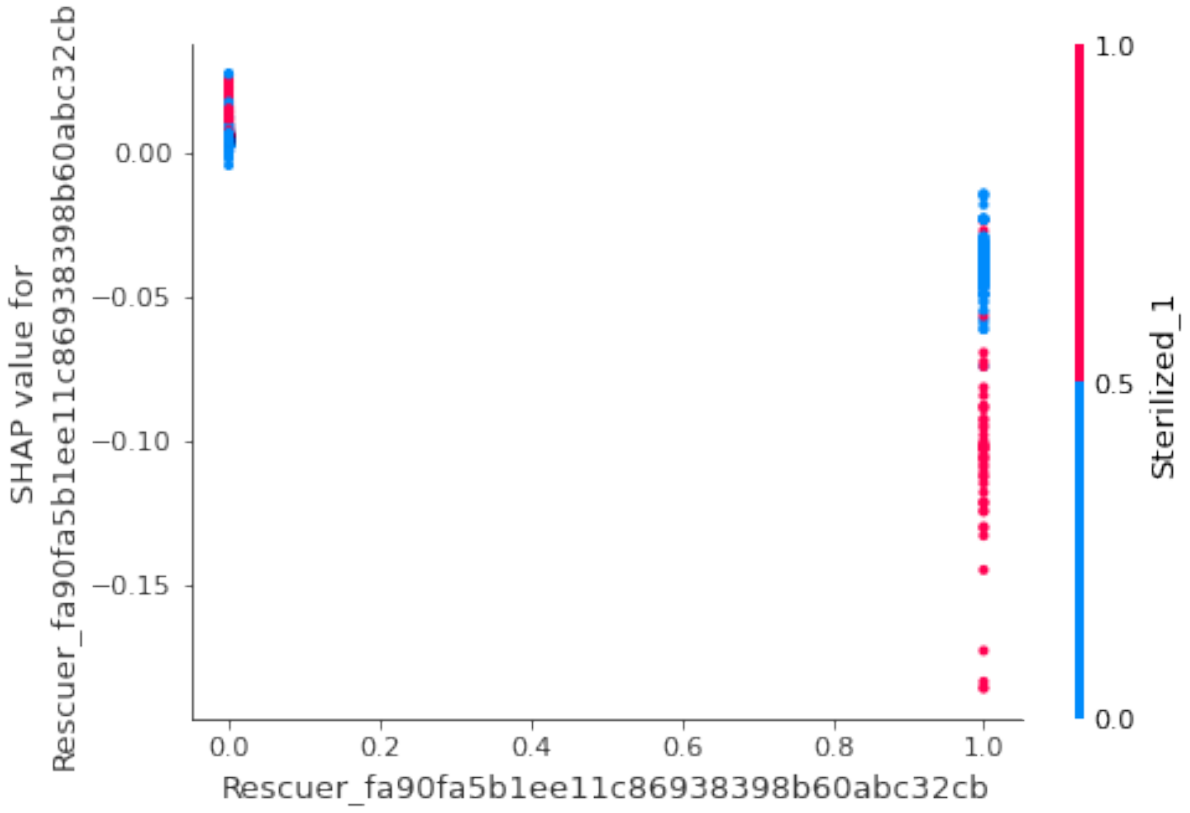


Fig 45.12 SHAP 'Rescuer_fa90fa5b1ee11c86938398b60abc32cb' dependence plot with interaction to 'Sterilized_1'

5.2.2.1.13 Rescuer_ee2747ce26468ec44c7194e7d1d9dad9

Fig 45.13 demonstrates the relationship between 'Rescuer_ee2747ce26468ec44c7194e7d1d9dad9' and the target variable as well as 'Rescuer_ee2747ce26468ec44c7194e7d1d9dad9's interaction with 'Sterilized_1'. We can see that the spread in SHAP values for dogs rescued by 'Rescuer_ee2747ce26468ec44c7194e7d1d9dad9' is much larger, ranging from 0 to 0.2. Since the range of SHAP values are above 0, we can say that a dog rescued by 'Rescuer_ee2747ce26468ec44c7194e7d1d9dad9' is not very likely to be adopted within 100 days. For dogs that are not rescued by 'Rescuer_ee2747ce26468ec44c7194e7d1d9dad9', the range of SHAP values is much smaller, ranging from 0.01 to -0.05. Since most of these SHAP values are negative, we can say that dogs not rescued by 'Rescuer_ee2747ce26468ec44c7194e7d1d9dad9' are more likely to be adopted within 100 days.

We can see that sterilizing a dog from 'Rescuer_ee2747ce26468ec44c7194e7d1d9dad9' will increase its chances of being adopted within 100 days, based on the even lower SHAP values. For dogs that are rescued by other rescuers, sterilization will actually decrease the chance of the dog being adopt, represented by higher SHAP values.

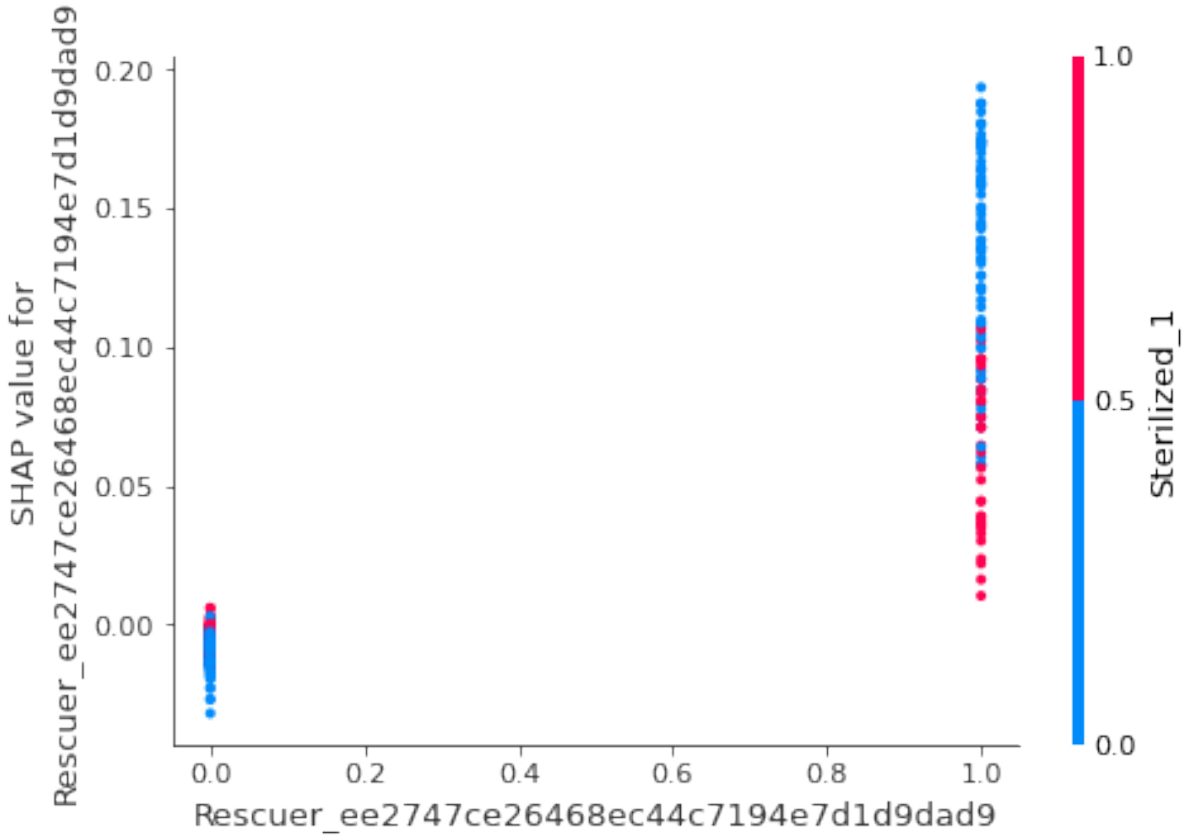


Fig 45.13 SHAP 'Rescuer_fa90fa5b1ee11c86938398b60abc32cb' dependence plot with interaction to 'Sterilized_1'

5.2.2.1.14 Rescuer_b53c34474d9e24574bcec6a3d3306a0d

Fig 45.14 demonstrates the relationship between 'Rescuer_b53c34474d9e24574bcec6a3d3306a0d' and the target variable as well as 'Rescuer_b53c34474d9e24574bcec6a3d3306a0d's interaction with 'ThreeMonths_0'. We can see that the spread in SHAP values for dogs rescued by 'Rescuer_b53c34474d9e24574bcec6a3d3306a0d' is much larger, ranging from -0.12 to 0.03. For dogs that are not rescued by 'Rescuer_b53c34474d9e24574bcec6a3d3306a0d', the range of SHAP values is much smaller, ranging from -0.02 to 0.02. Most dogs rescued by 'Rescuer_b53c34474d9e24574bcec6a3d3306a0d' have a negative SHAP value, this means that dogs rescued by this rescuer are more likely to be adopted within 100 days.

Looking at the interaction between 'Rescuer_b53c34474d9e24574bcec6a3d3306a0d' and 'ThreeMonths_0', we can see that dogs rescued by 'Rescuer_b53c34474d9e24574bcec6a3d3306a0d' are more likely to be adopted within 100 days if they are older than 3 months. For dogs that aren't rescued by 'Rescuer_b53c34474d9e24574bcec6a3d3306a0d', they are more likely to be adopted within 100 days if they are 3 months or younger. This suggests that the 'Rescuer_b53c34474d9e24574bcec6a3d3306a0d' is really good at driving adoption for older dogs.

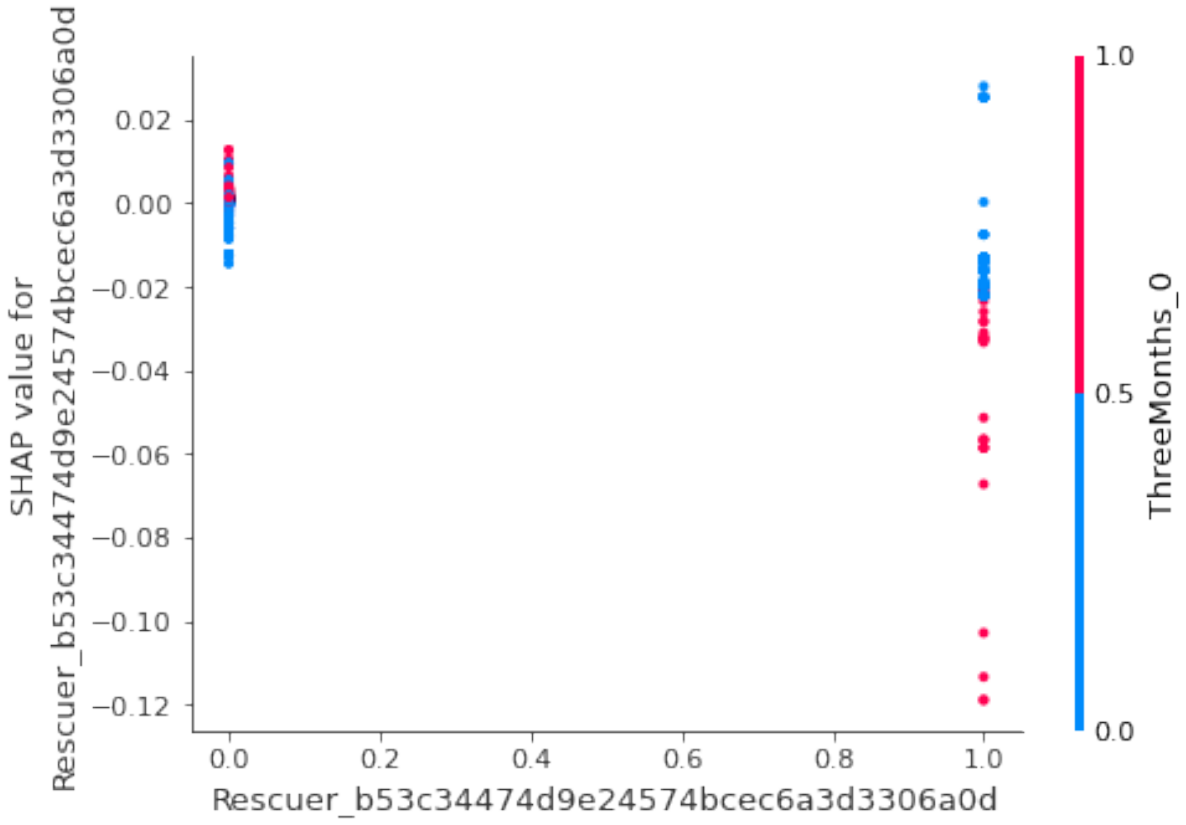


Fig 45.14 SHAP 'Rescuer_b53c34474d9e24574bcec6a3d3306a0d' dependence plot with interaction to 'ThreeMonths_0'

5.2.2.1.15 Rescuer_c00756f2bdd8fa88fc9f07a8309f7d5d

Fig 45.15 demonstrates the relationship between 'Rescuer_c00756f2bdd8fa88fc9f07a8309f7d5d' and the target variable as well as 'Rescuer_c00756f2bdd8fa88fc9f07a8309f7d5d's interaction with 'ThreeMonths_0'. We can see that the spread in SHAP values for dogs rescued by 'Rescuer_b53c34474d9e24574bcec6a3d3306a0d' is much larger, ranging from -0.06 to 0.04. For dogs that are not rescued by 'Rescuer_c00756f2bdd8fa88fc9f07a8309f7d5d', the range of SHAP values is much smaller, ranging from -0.01 to 0.015. Most dogs rescued by 'Rescuer_c00756f2bdd8fa88fc9f07a8309f7d5d' have a negative SHAP value, this means that dogs rescued by this rescuer are more likely to be adopted within 100 days.

Looking at the interaction between 'Rescuer_c00756f2bdd8fa88fc9f07a8309f7d5d' and 'ThreeMonths_0', we can see that dogs rescued by 'Rescuer_c00756f2bdd8fa88fc9f07a8309f7d5d' are more likely to be adopted within 100 days if they are older than 3 months. For dogs that aren't rescued by 'Rescuer_c00756f2bdd8fa88fc9f07a8309f7d5d', they are more likely to be adopted within 100 days if they are 3 months or younger. This suggests that the 'Rescuer_c00756f2bdd8fa88fc9f07a8309f7d5d' is really good at driving adoption for older dogs.

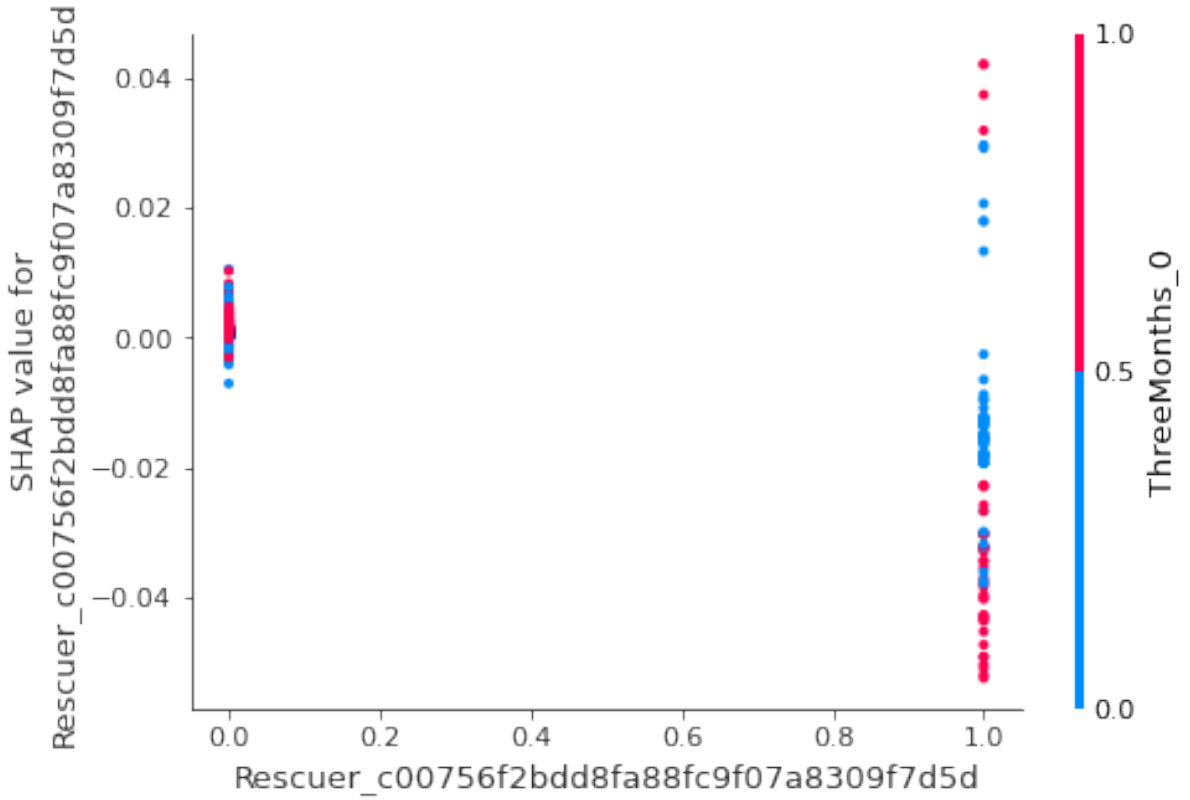


Fig 45.15 SHAP 'Rescuer_c00756f2bdd8fa88fc9f07a8309f7d5d' dependence plot with interaction to 'ThreeMonths_0'

5.2.2.1.16 Breed_20

Fig 45.16 demonstrates the relationship between 'Breed_20' and the target variable as well as 'Breed_20's interaction with 'ThreeMonths_0'. We can clearly see that dogs that are of breed 20 (Beagle) are more likely to be adopted within 100 days (based on the negative SHAP values). Dogs that do not have a dominant breed of Beagle are less likely to be adopted within 100 days represented by the positive SHAP values.

Looking at the interaction between 'Breed_20' and 'ThreeMonths_0', we can see that beagles that are older than 3 months (red) are more likely to be adopted within 100 days than beagles that are 3 months or younger (blue). For dogs that don't have a dominant breed of beagle, those that are 3 months or younger are slightly more likely to be adopted within 100 days than those that are older than 3 months.

Beagles are most known for their scent and are often used by authorities to track down scents. For this reason, older beagles may be more desirable as their senses will be fully developed and can be used to track down scents more effectively.

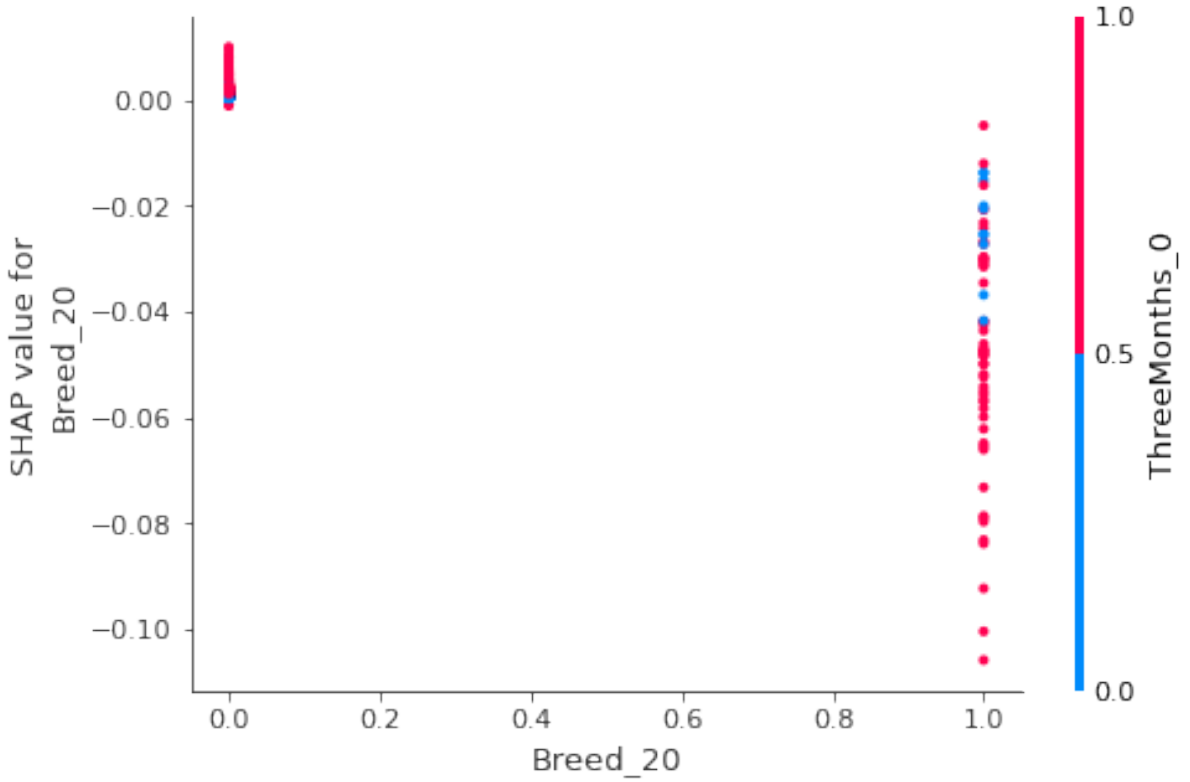


Fig 45.16 SHAP 'Breed_20' dependence plot with interaction to 'ThreeMonths_0'

5.2.2.2 Cat adoption dataset

5.2.2.2.1 RescuerFreq

Fig 46.1 demonstrates the relationship between 'RescuerFreq' and the target variable as well as 'RescuerFreq's interaction with 'Rescuer_c00756f2bdd8fa88fc9f07a8309f7d5d'. There is an initial downward trend between RescuerFreq and the target variable (up until ~ 15). Beyond 15, the relationship between 'RescuerFreq' and the target variable seem to be constant and have SHAP values between -0.2 to 0, with a single outlier for a rescuer with RescuerFreq of ~80 and having a SHAP value of ~ 0.05. This means that cats are more likely to be adopted within 100 days when they are rescued by a rescuer that has rescued at least 15 other cats.

Fig 46.1 also shows the interaction between 'RescuerFreq' and 'Rescuer_c00756f2bdd8fa88fc9f07a8309f7d5d'. We can see that 'Rescuer_c00756f2bdd8fa88fc9f07a8309f7d5d' is a rescue that has rescued almost 80 other cats, it is one of the only rescues that have such a high number of rescued cats. Cats rescued by 'Rescuer_c00756f2bdd8fa88fc9f07a8309f7d5d' (red data points) also have very small SHAP values which indicate that they are very likely to be adopted within 100 days. Other rescues with the same rescuer frequency have a higher SHAP value, indicating that cats located in other rescues are not as likely to be adopted within 100 days. This trend mades sense as cat adopters are more likely to adopt from a reputable and popular rescue.

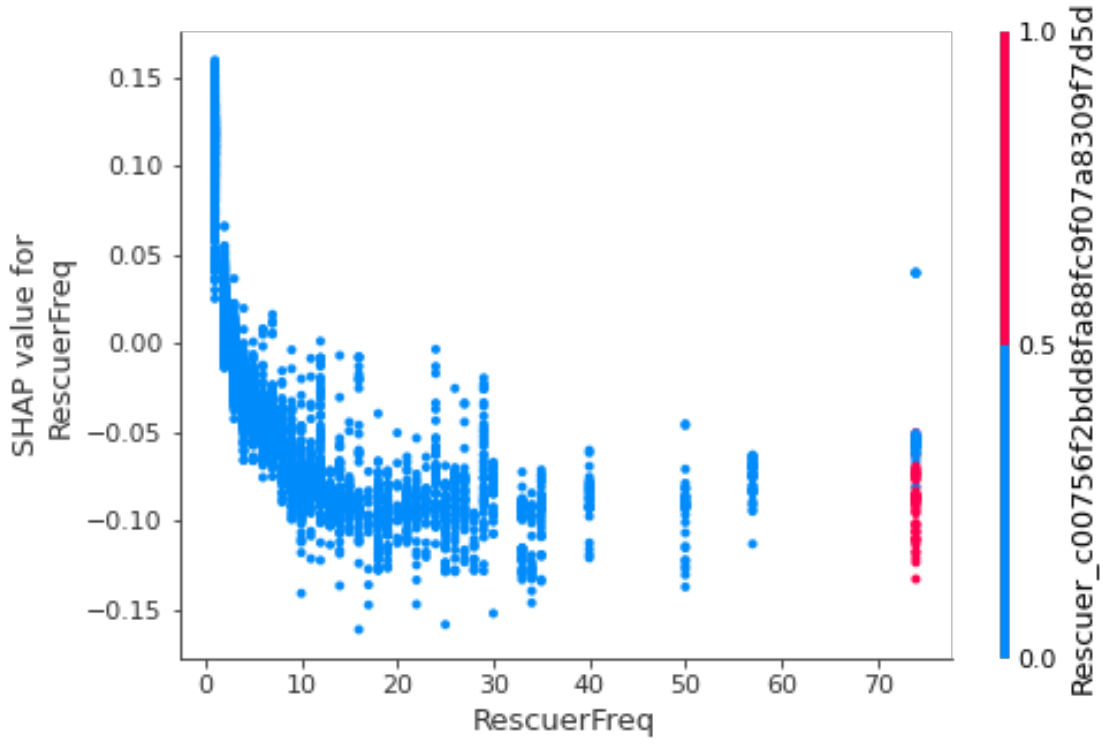


Fig 46.1 SHAP 'RescuerFreq' dependence plot with interaction to 'Rescuer_c00756f2bdd8fa88fc9f07a8309f7d5d'

5.2.2.2.2 Age

Fig 46.2 demonstrates the relationship between 'Age' and the target variable as well as 'Age's interaction with 'Sterilized_1'. From Fig 46.2 we can see that there is a general upwards trend until around 10 months (where the SHAP value consistently lies between 0.02 and 0.2).

This tells us that when cats are between 0 to 10 months old, the younger they are, the more likely they will be adopted within 100 days. Since this upwards trend plateaus at around 10 months, we can say that cats older than 10 months generally have similar likelihoods of being adopted.

Looking at the interaction between 'Age' and 'Sterilized_1', we can see more younger cats are unsterilized and more older cats are sterilized. Looking at cats of the same age, those that are sterilized (red data points) generally have lower SHAP values, thus indicating that sterilized cats are more likely to be adopted within 100 days.

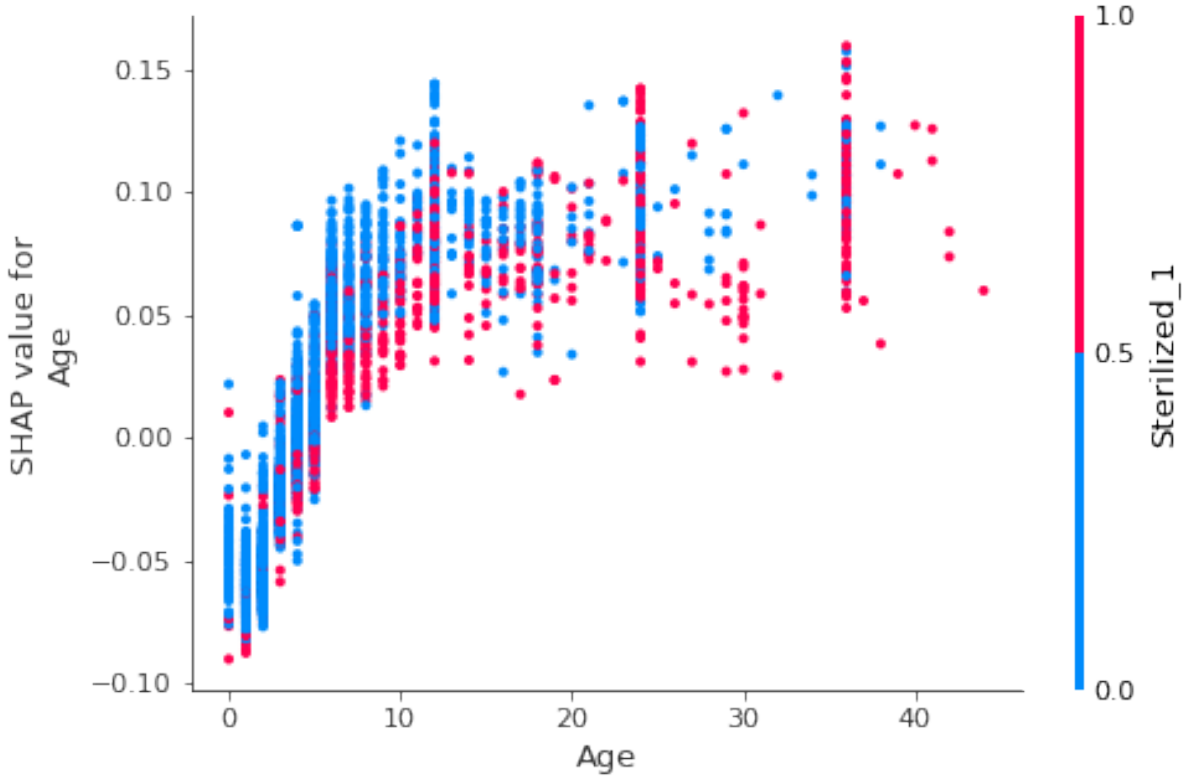


Fig 46.2 SHAP 'Age' dependence plot with interaction to 'Sterilized_1'

5.2.2.2.3 ThreeMonths_1

Fig 46.3 demonstrates the relationship between 'ThreeMonths_1' and the target variable as well as 'ThreeMonth_1's interaction with 'State_41326'. It is evident that cats that are 3 months or younger are more likely to be adopted within 100 days (lower SHAP values), whilst cats that are older than 3 months are less likely to be adopted within 100 days (higher SHAP values). This makes sense as cats 3 months or younger are kittens and are much more desirable.

Looking at the interaction between 'ThreeMonths_1' and 'State_41326', we can see that for cats that are older than 3 months, they are more likely to be adopted within 100 days if they are located in State 41326 (Selangor) than if they were not. For cats that are 3 months or younger, they are more likely to be adopted if they're not from Selangor. This indicates that Selangor is better at driving adoption for older cats than younger cats.

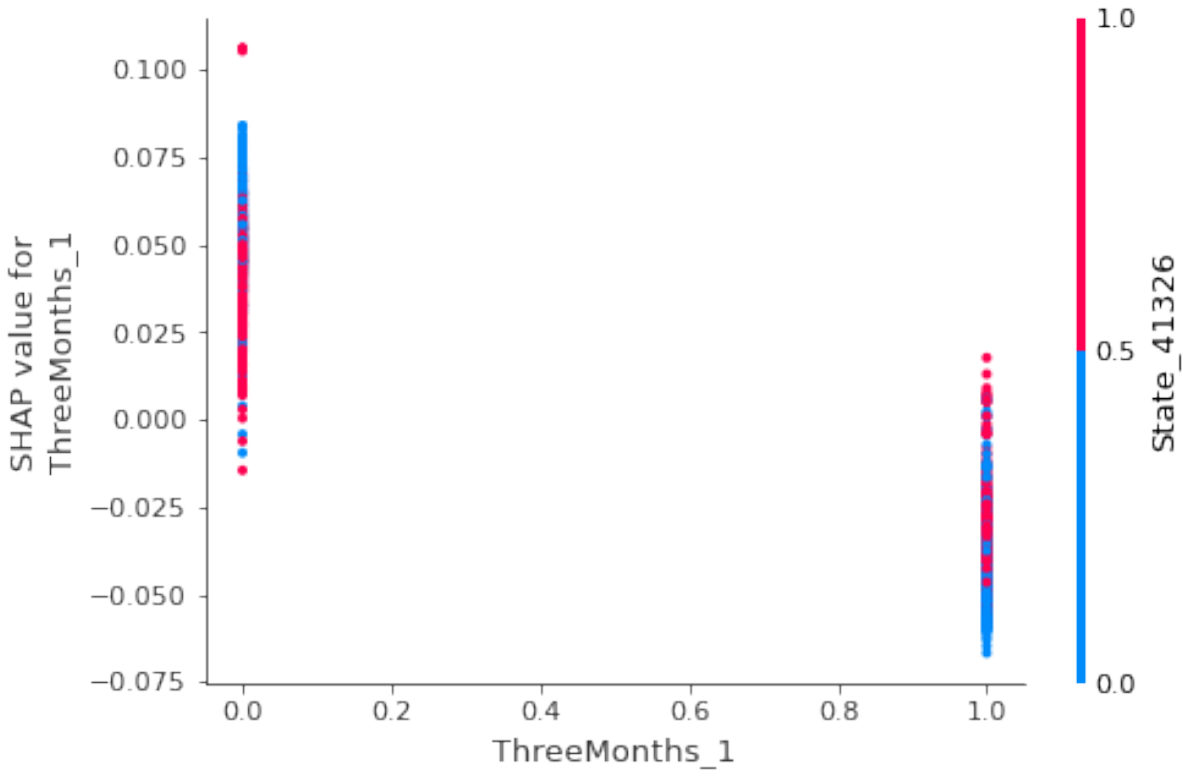


Fig 46.3 SHAP 'ThreeMonths_1' dependence plot with interaction to 'State_41326'

5.2.2.2.4 DescNumWords

Fig 46.4 demonstrates the relationship between 'DescNumWords' and the target variable as well as 'DescNumWords' interaction with 'PhotoAmt'. From Fig 46.4 we can see that there is a steep downwards trend until around 50 words, after which the trend seems to plateau and become sparser. This tells us that when cats profile descriptions are between 0 to 50 words long, the longer the description, the more likely they will be adopted within 100 days. Since this downwards trend plateaus at around 50 words, we can say that cats with profile descriptions above 50 words generally have similar likelihoods of being adopted within 100 days.

Looking at the interaction between 'DescNumWords' and 'PhotoAmt', we can that cats with a shorter description tend to have less photos attached on their file. This makes sense as rescuers that write a short description are likely to only attach a few photos (if any).

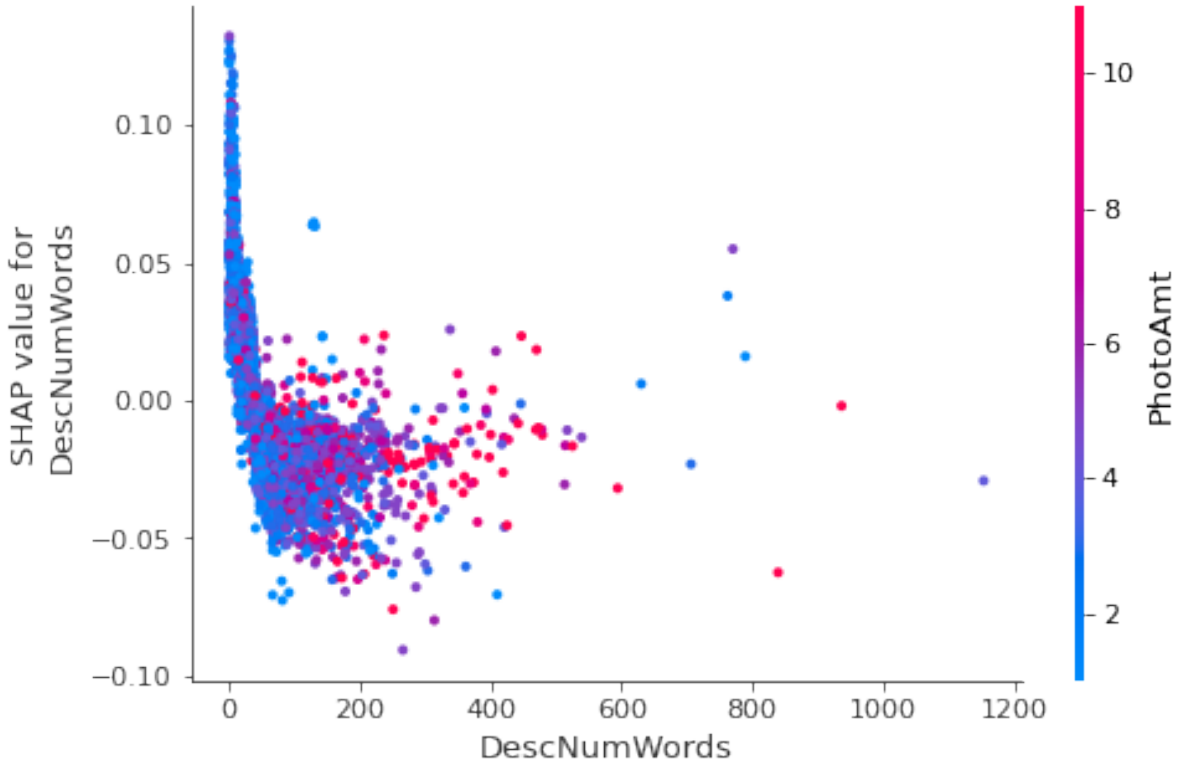


Fig 46.4 SHAP 'DescNumWords' dependence plot with interaction to 'PhotoAmt'

5.2.2.2.5 Quantity

Fig 46.5 demonstrates the relationship between 'Quantity' and the target variable as well as 'Quantity's interaction with 'Preparation_1'. In general, we can see that quantity 1 has a negative SHAP value whilst quantities above 1 generally have a positive value. Quantities between 2 and 9 generally have SHAP values between 0 and 0.2. Quantities above 9 have a larger SHAP value range of between 0.05 to 0.3. This means that cat listings with smaller quantities are more likely to be adopted within 100 days than those of larger quantities. This makes sense as it takes longer for multiple cats to be adopted.

We can see that most cats are not fully prepared (vaccinated, sterilized and dewormed), represented by blue data points. Those that are fully prepared are in quantities of between 1 – 8, with one outlier found at quantity 16. In general, we can see that fully prepared cats have a higher SHAP value compared to the same quantity of cats that are not fully prepared. This means that cats that are not fully prepared are more likely to be adopted within 100 days than cats, this could be due to the fact that cats that are quickly adopted don't have enough time to be fully prepared.

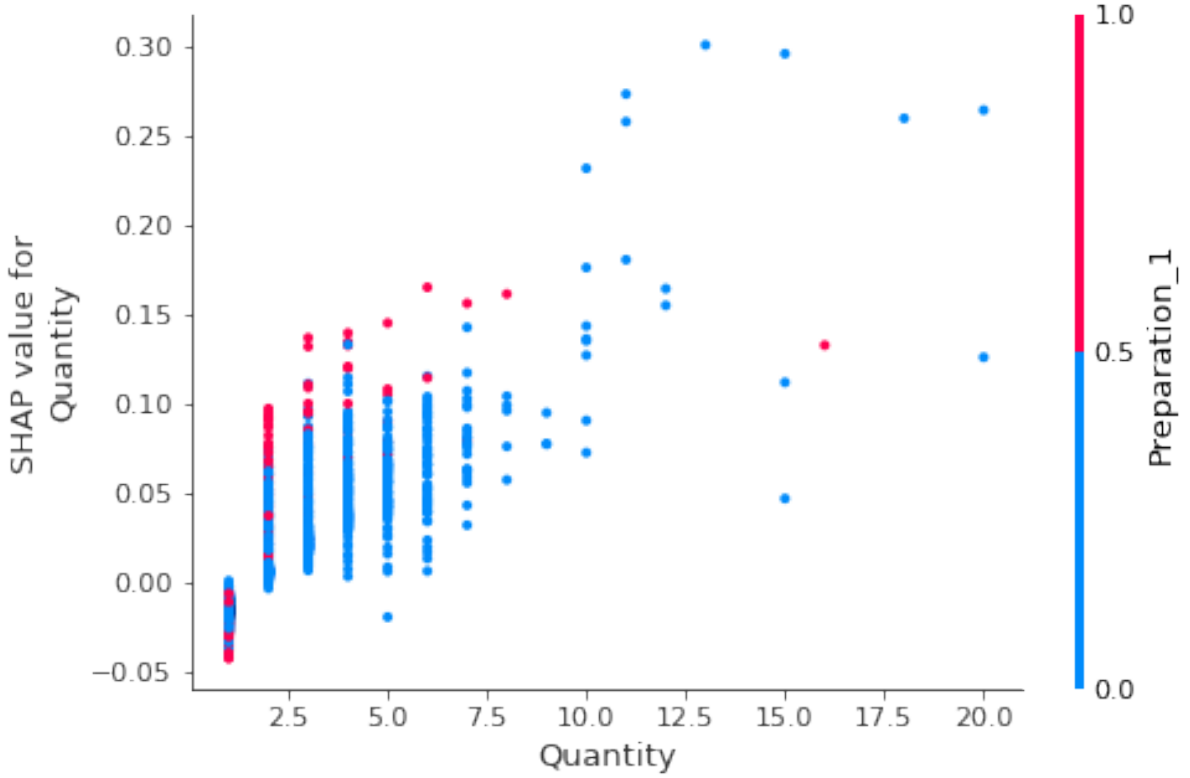


Fig 46.5 SHAP 'Quantity' dependence plot with interaction to 'Preparation_1'

5.2.2.2.6 PhotoAmt

Fig 46.6 demonstrates the relationship between 'PhotoAmt' and the target variable as well as 'PhotoAmt's interaction with 'RescuerFreq'. We can see that there is a clear decline in SHAP values as the number of photos increases up to 9. Beyond 9 photos, there doesn't appear to be any effect on photo amount and the probability of adoption within 100 days.

It is evident that a photo amount of 0 results in a SHAP value above 0.1, this means that when there are no photos attached on the cat listing, the cat is less likely to be adopted within 100 days. We can see that the optimal number of photos to include in a cat listing is between 9-12, any photos beyond that is redundant.

Now looking at the interaction between 'PhotoAmt' and 'RescuerFreq,' we can see that a low rescuer frequency generally results in lower SHAP values. This means that given the same photo amount, a cat rescued by a rescuer with a smaller frequency is more likely to be adopted within 100 days. This could be due to the urgency of smaller rescues (especially one-off rescuers), if cats aren't adopted quickly they may be transferred to larger rescues that can accommodate them. We can see that cats with 0 photos attached on their file are mostly rescued by rescues with a low rescuer frequency. The lack of photos can be explained by the rescues inexperience since they haven't rescued many cats before.

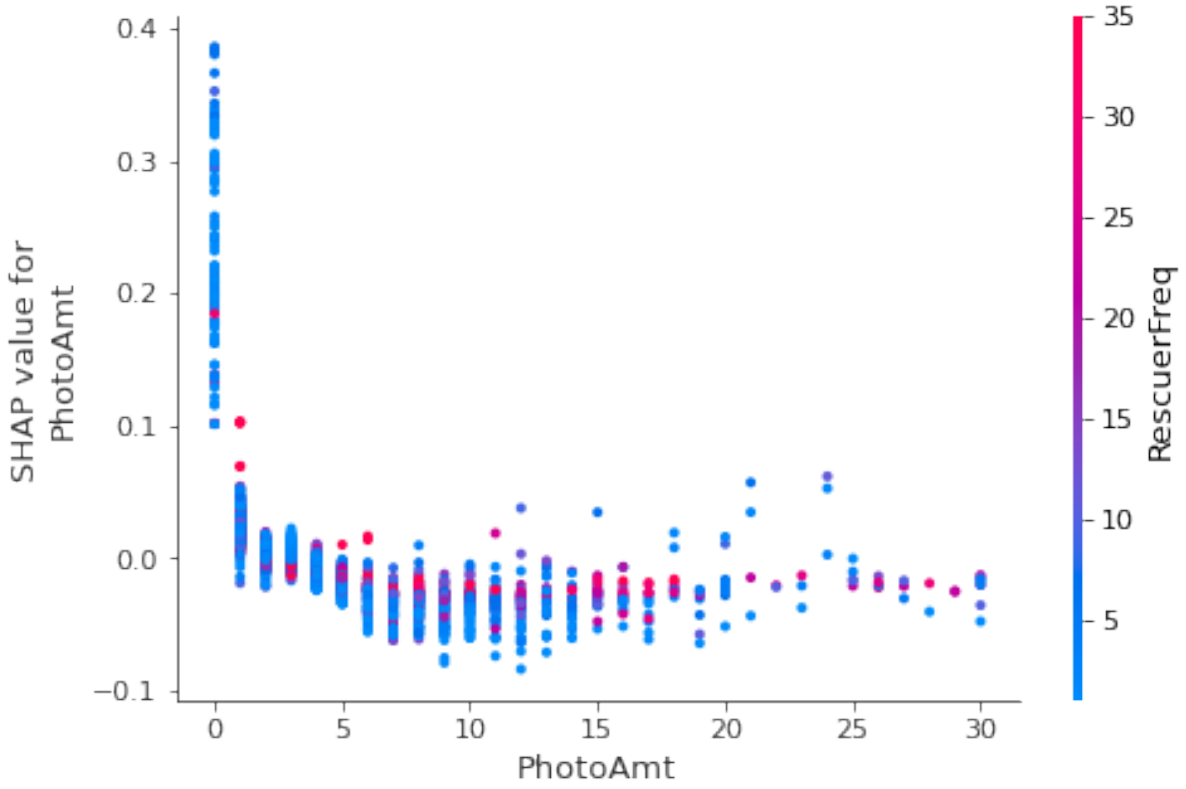


Fig 46.6 SHAP 'PhotoAmt' dependence plot with interaction to 'RescuerFreq'

5.2.2.2.7 Sterilized_1

Fig 46.7 demonstrates the relationship between 'Sterilized_1' and the target variable as well as 'Sterilized_1' interaction with 'ThreeMonths_1'. Based on Fig 46.7, we can see that there are more cats that are sterilized than not. We can see that cats that are sterilized have positive SHAP values, whilst those that are not have negative SHAP values. This means that unsterilized cats are more likely to be adopted within 100 days. Cats that are older than 3 months dominate this dataset, making it hard to unravel the relationship between sterilization and cats 3 months or younger (red data points). Unsterilized cats may be more adoptable because adopters want a cat that can have offsprings.

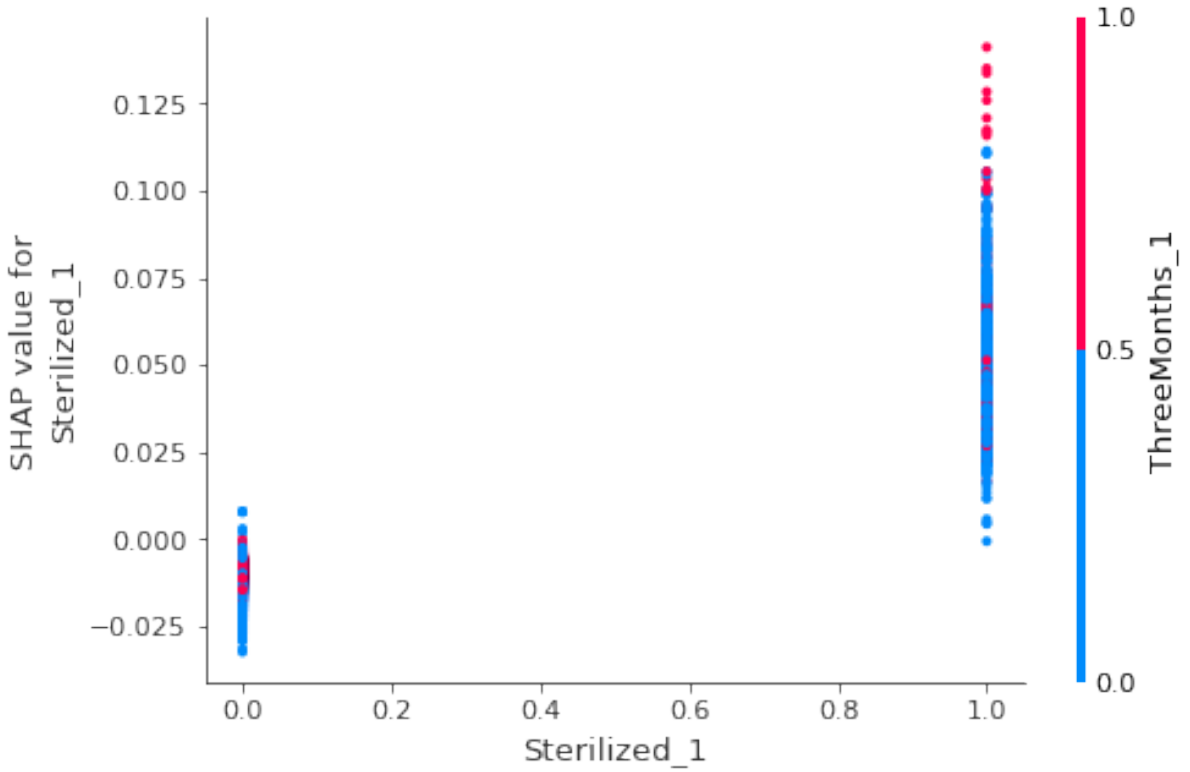


Fig 46.7 SHAP 'Sterilized_1' dependence plot with interaction to 'ThreeMonths_1'

5.2.2.2.8 State_41326

Fig 46.8 demonstrates the relationship between 'State_41326' and the target variable as well as 'State_41326's interaction with 'ThreeMonths_1'. Based on Fig 46.8, we can see that cats located in 'State_41326' (Selangor) are more likely to be adopted within 100 days based on the range of SHAP values that generally lie below 0. For cats that are located in Selangor, they are more likely to be adopted if they are older than 3 months (blue data points have more negative SHAP values). This is unusual as younger cats tend to be more adoptable.

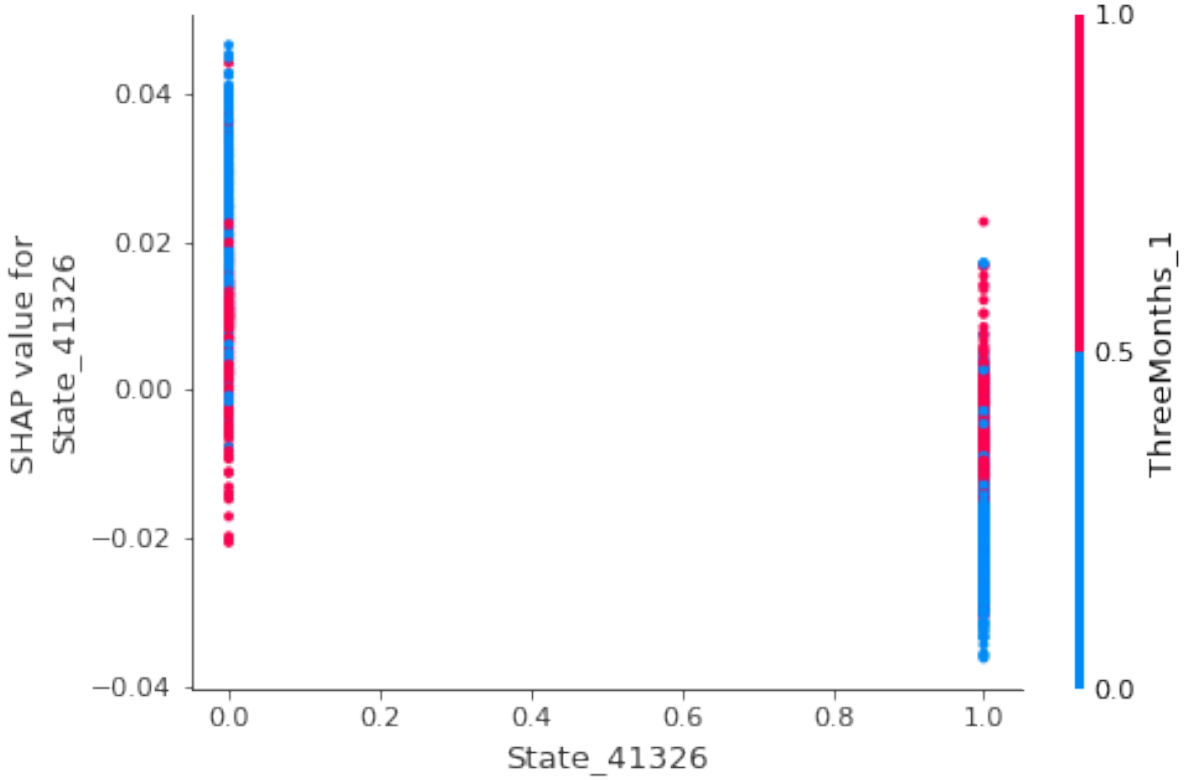


Fig 46.8 SHAP 'State_41326' dependence plot with interaction to 'ThreeMonths'

5.2.2.2.9 Rescuer_999

Fig 46.9 demonstrates the relationship between 'Rescuer_999' and the target variable as well as 'Rescuer_999's interaction with 'Sterilized_1'. Based on Fig 46.9, we can see that there are more cats that are rescued by popular rescuers (labelled as 0 for Rescuer_999) than unpopular ones. Interestingly, cats that aren't rescued by a popular rescuer have a very large range of SHAP values, with most of the data points lying below 0, thus suggesting that cat's rescued by popular rescues are more likely to be adopted within 100 days. For cats that are rescued by unpopular rescuers, the range of SHAP values also encompass 0 but there is a clear distinction where red data points generally represent positive SHAP values and blue data points represent negative SHAP values. This means that for cats rescued by unpopular rescuers, they are more likely to be adopted within 100 days if they are not sterilized and less likely to be adopted within 100 days if they are sterilized. This trend seems to be reversed for cats rescued by popular rescuers where sterilized cats have lower SHAP values than unsterilized cats. This means that cats rescued by popular rescuers are more likely to be adopted within 100 days if they are sterilized.

It makes sense that cats rescued by popular rescuers are more adoptable, this is because most adopters looking to adopt will adopt from a notable adoption centre. It also makes sense that a sterilized cat in a notable adoption centre is even more desirable because a sterilized cat signifies one less outstanding cost. For cats rescued by unpopular rescuers, these tend to be one-off rescuers who may have come across a stray cat on the street, these rescuer may not have the means to keep the cat for long so they are often adopted quickly without being sterilized.

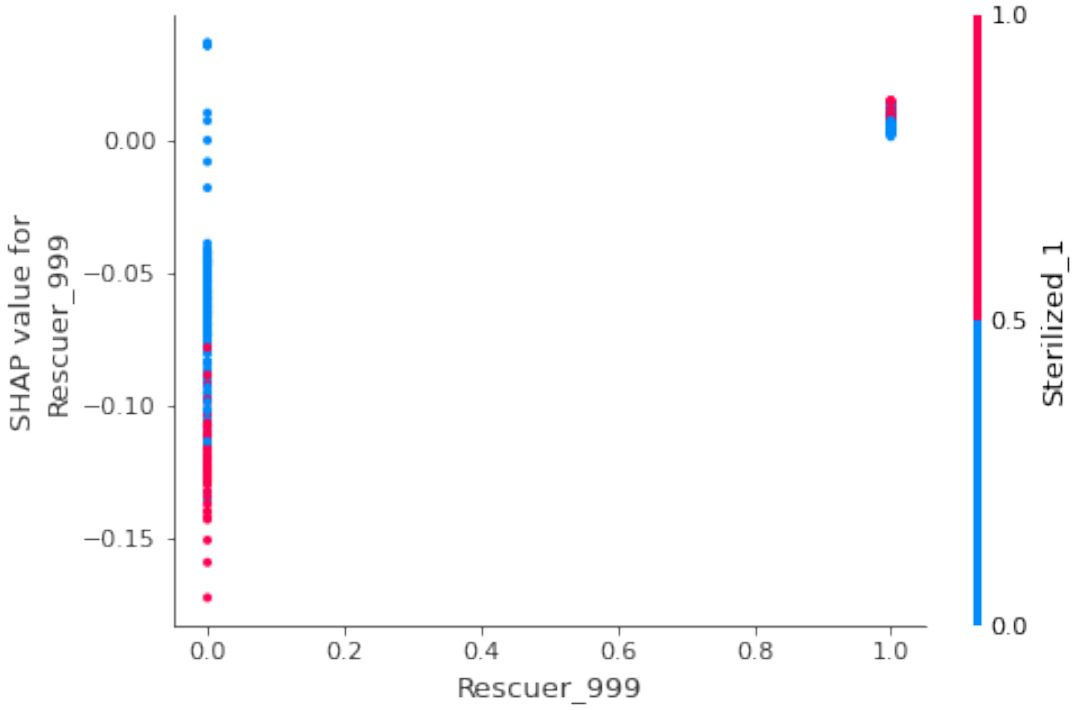


Fig 46.9 SHAP 'Rescuer_999' dependence plot with interaction to 'Sterilized_1'

5.2.2.2.10 healthy

Fig 46.10 demonstrates the relationship between 'healthy' and the target variable as well as the interaction between 'healthy' and 'RescuerFreq'. From Fig 46.10, we can see that most cat adoption profile descriptions contain the word 'healthy'. The range of SHAP values is very small for descriptions not containing the word 'healthy' and the range of SHAP values is much larger for descriptions containing the word 'healthy'. This means that in general, cats with the word 'healthy' in their description are more likely to be adopted within 100 days.

For a cat description that doesn't contain the word 'healthy', we can see that it is more likely to be adopted within 100 days (negative SHAP value) if the cat was rescued by a popular rescuer and it is more likely to not be adopted within 100 days (positive SHAP value) if the cat was rescued by an unpopular rescuer. For a cat description that contains the word 'healthy', we can see that it is more likely to be adopted within 100 days (negative SHAP values) if the cat was rescued by a lesser-known rescuer (blue) and it is more likely to not be adopted within 100 days (positive SHAP value) if the cat was rescued by a popular rescue (red).

This makes sense as the health situation of a cat may be more in question for smaller rescues as they are usually less notable. For larger, more reputable rescues, it is expected that they only release cats that are healthy or will fully disclose any health issues.

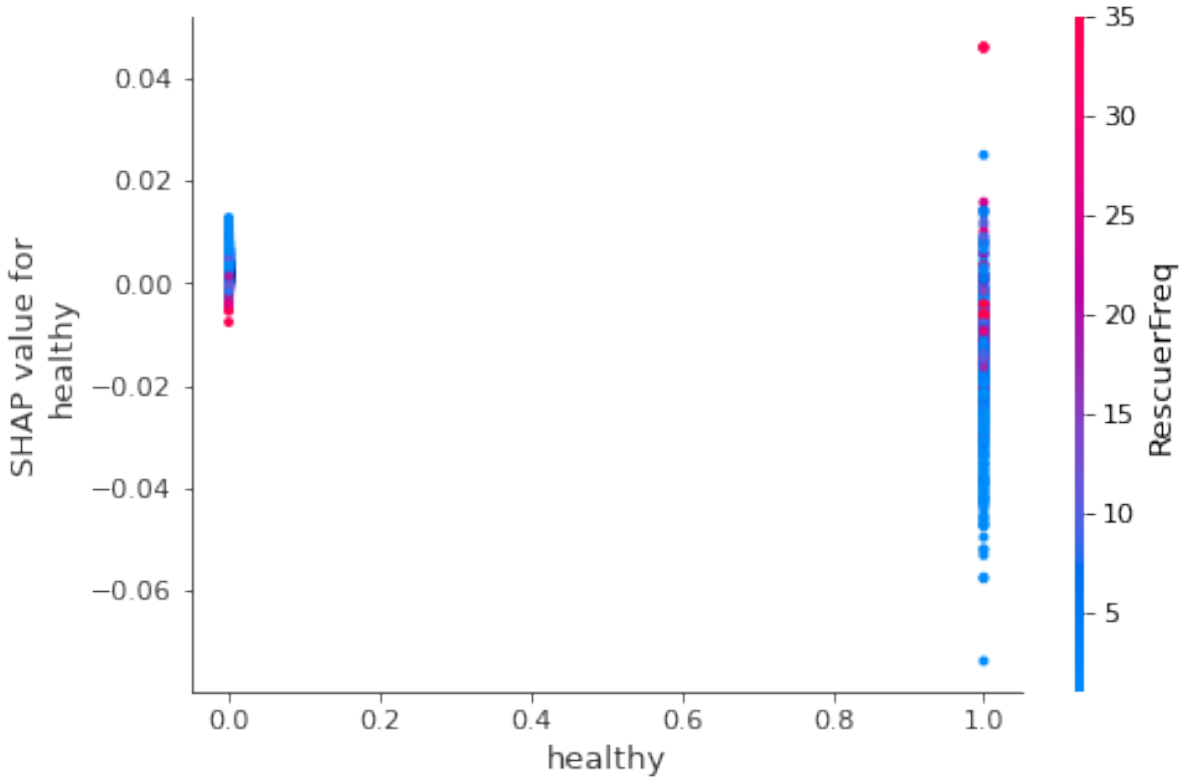


Fig 46.10 SHAP 'healthy' dependence plot with interaction to 'RescuerFreq'

5.2.2.2.11 eyes

Fig 46.11 demonstrates the relationship between ‘eyes’ and the target variable as well as ‘eyes’ interaction with ‘RescuerFreq’. From Fig 46.11, we can see that most cat adoption profile descriptions contain the word ‘eyes’. The range of SHAP values is very small for descriptions not containing the word ‘eyes’ and the range of SHAP values is much larger for descriptions containing the word ‘eyes’. In general, cats with descriptions containing the word ‘eyes’ are more likely to be adopted within 100 days.

For a cat description that doesn’t contain the word ‘eyes’, we can see that the cat is more likely to be adopted within 100 days (negative SHAP value) if they are rescued by a popular rescuer. For a cat description that contains the word ‘eyes’, we can see that the cat is more likely to be adopted within 100 days (negative SHAP values) if the cat was rescued by a lesser-known rescuer (blue).

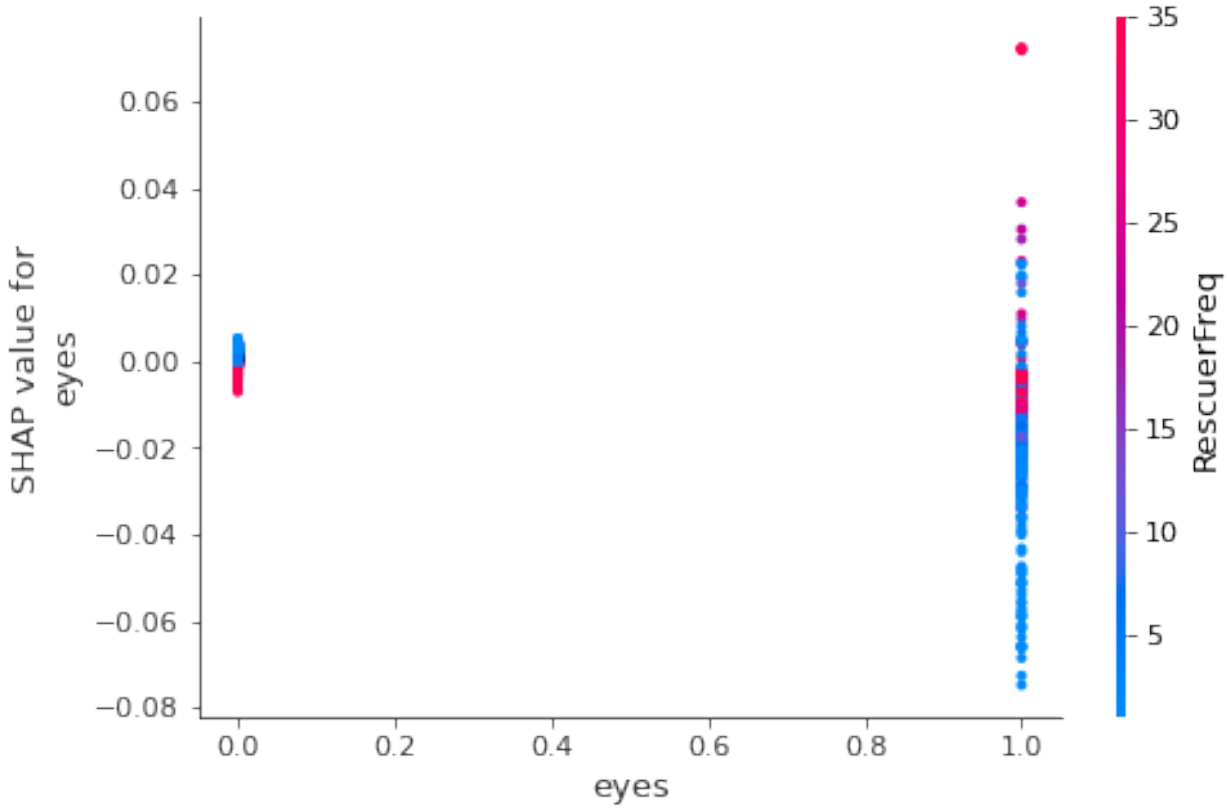


Fig 46.11 SHAP ‘eyes’ dependence plot with interaction to ‘RescuerFreq’

5.2.2.2.12 Rescuer_c00756f2bdd8fa88fc9f07a8309f7d5d

Fig 46.12 demonstrates the relationship between 'Rescuer_c00756f2bdd8fa88fc9f07a8309f7d5d' and the target variable as well as 'Rescuer_c00756f2bdd8fa88fc9f07a8309f7d5d's interaction with 'RescuerFreq'. Based on Fig 46.12, we can see that there are more cats that are rescued by 'Rescuer_c00756f2bdd8fa88fc9f07a8309f7d5d' (labelled as 1) than other rescues. Interestingly, cats that are rescued by Rescuer_c00756f2bdd8fa88fc9f07a8309f7d5d have a very large range of SHAP values, with most of the datapoints lying below 0, thus suggesting that cats rescued by Rescuer_c00756f2bdd8fa88fc9f07a8309f7d5d are more likely to be adopted within 100 days. We can also see that most rescues have a high rescuer frequency including Rescuer_c00756f2bdd8fa88fc9f07a8309f7d5d.

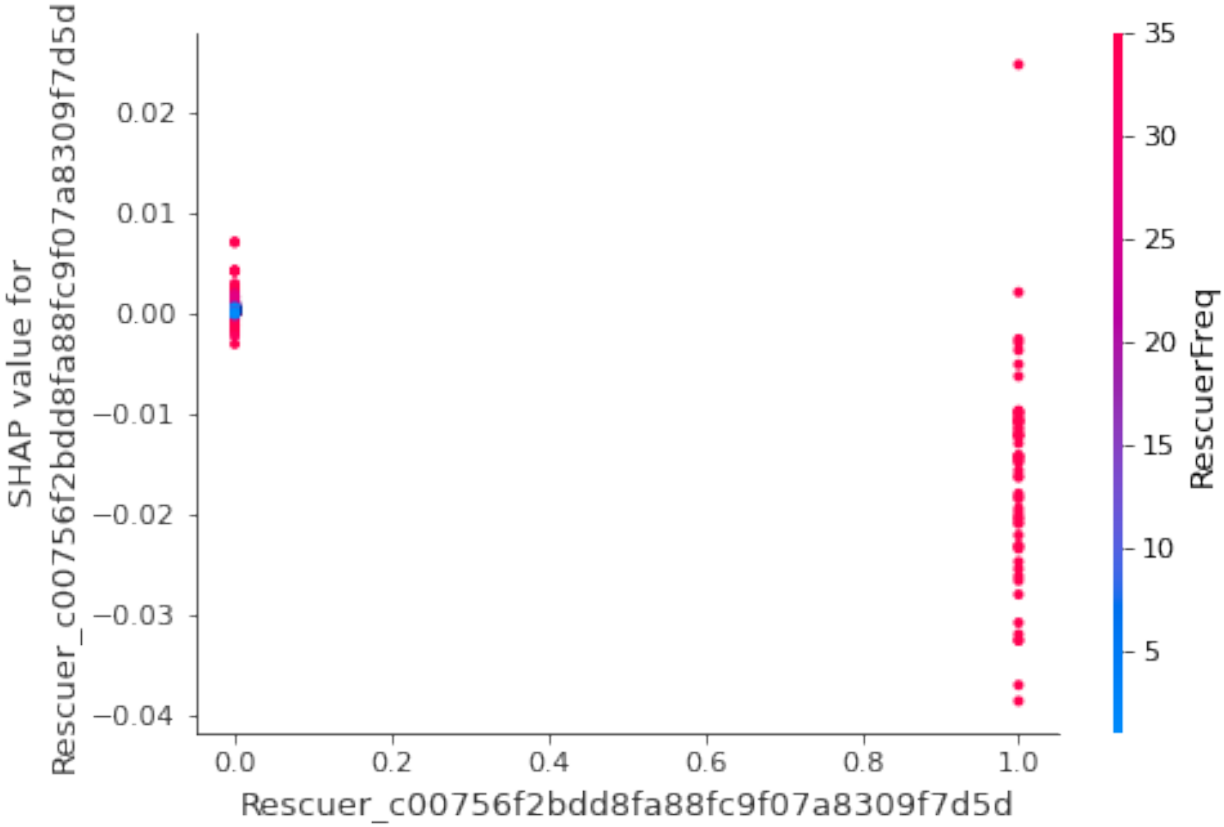


Fig 46.12 SHAP 'Rescuer_c00756f2bdd8fa88fc9f07a8309f7d5d' dependence plot with interaction to 'RescuerFreq'

5.2.3 Force plot

5.2.3.1 Dog Adoption Dataset

The base value for the force plots is 0.295. This is the mean of the target variable in the training data.

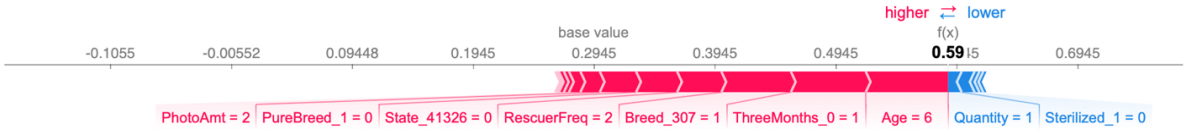
5.2.3.1.1 Example 1

Fig 47.1 shows the feature values for example 1 (PetID: e63eea2f7) from the dog test set.

Age	6
Breed_20	0
Breed_307	1
Breed_999	0
PhotoAmt	2
Preparation_1	0
PureBreed_1	0
Quantity	1
RescuerFreq	2
Rescuer_999	1
Rescuer_b53c34474d9e24574bcec6a3d3306a0d	0
Rescuer_c00756f2bdd8fa88fc9f07a8309f7d5d	0
Rescuer_ee2747ce26468ec44c7194e7d1d9dad9	0
Rescuer_fa90fa5b1ee11c86938398b60abc32cb	0
State_41324	0
State_41326	0
State_41332	0
Sterilized_1	0
ThreeMonths_0	1

Fig 47.1 Feature values for example 1 from dog test set

Fig 47.2 shows the force plot for the dog with PetID 'e63eea2f7.' We can see that the predicted probability of this dog not being adopted within 100 days is 0.59. Fig 47.2 displays the different contributions for each feature. We can see that 'PhotoAmt', 'PureBreed_1', 'State_41326', 'RescuerFreq', 'Breed_307', 'ThreeMonths_0' and 'Age' contributed positively towards the target variable, whilst 'Quantity' and 'Sterilized_1' contributed negatively towards the target variable. We can see that the 'Age' feature contributed the most towards the predicted value, this means that age is an important feature for this prediction. From previous analysis we know that the older the dog, the more likely it will not be adopted within 100 days, this is reflected by the positive contribution of the age feature here as the dog is 6 months old.



47.2 Force plot for example 1 from dog test set

5.2.3.1.2 Example 2

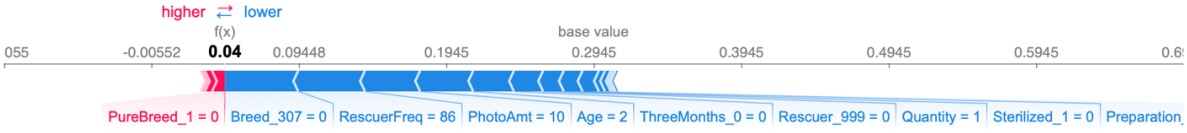
Fig 48.1 shows the feature values for example 2 (PetID: afa1e2afb) from the dog test set.

Age	2
Breed_20	0
Breed_307	0
Breed_999	0
PhotoAmt	10
Preparation_1	0
PureBreed_1	0
Quantity	1
RescuerFreq	86
Rescuer_999	0
Rescuer_b53c34474d9e24574bcec6a3d3306a0d	0
Rescuer_c00756f2bdd8fa88fc9f07a8309f7d5d	0
Rescuer_ee2747ce26468ec44c7194e7d1d9dad9	0
Rescuer_fa90fa5b1ee11c86938398b60abc32cb	0
State_41324	0
State_41326	1
State_41332	0
Sterilized_1	0
ThreeMonths_0	0

Fig 48.1 Feature values for example 2 from the dog test set

Fig 48.2 shows the force plot for example 2. We can see that the predicted probability of this dog not being adopted within 100 days is 0.04. Fig 48.2 displays the different contributions for each feature. We can see that 'PureBreed_1' contributed positively towards the target variable, whilst 'Breed_307', 'RescuerFreq', 'PhotoAmt', 'Age', 'ThreeMonths_0', 'Rescuer_999', 'Quantity', 'Sterilized_1' and 'Preparation_1' contributed negatively towards the target variable. We can see that the 'Breed_307' feature contributed the most towards the predicted value, this means that 'Breed_307'

is an important feature for this prediction. From previous analysis we know dogs of dominant breed 307 are more likely to not be adopted within 100 days, this is reflected by the negative contribution here as the dog is not of dominant breed 307.



48.2 Force plot for example 2 from the dog test set

5.2.3.2 Cat Adoption Dataset

The base value for the force plots is 0.295. This is the mean of the target variable in the training data.

5.2.3.2.1 Example 1

Fig 49.1 shows the feature values for example 1 (PetID: c4835ee5c) from the cat test set.

Age	2
DescNumWords	19
PhotoAmt	4
Preparation_1	0
Quantity	1
RescuerFreq	3
Rescuer_999	1
Rescuer_c00756f2bdd8fa88fc9f07a8309f7d5d	0
State_41326	1
Sterilized_1	0
ThreeMonths_1	1
Vaccinated_1	0
eyes	0
healthy	0

Fig 49.1 Feature values for example 1 from the cat test set

Fig 49.1 shows the force plot for example 1. We can see that the predicted probability of this cat not being adopted within 100 days is 0.12. Fig 49.1 displays the different contributions for each feature. We can see that 'Rescuer_999' and 'DescNumWords' contributed positively towards the target variable, whilst 'Age', 'ThreeMonths_1', 'PhotoAmt', 'RescuerFreq', 'Quantity', 'State_41326' and 'Sterilized_1' contributed negatively towards the target variable. We can see that the 'Age' feature contributed the most towards the predicted value, this means that age is an important feature for this prediction. From previous analysis we know that the older the cat, the more likely it will not be adopted within 100 days, this is reflected by the negative contribution of the age feature here as the dog is only 2 months old.

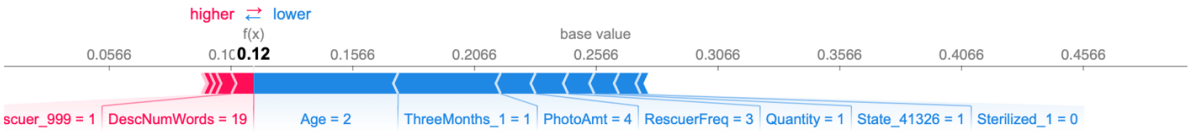


Fig 49.2 Force plot for example 1 from the cat test set

5.2.3.2.2 Example 2

Fig 50.1 shows the feature values for example 1 (PetID: 796781fbb) from the cat test set.

Age	4
DescNumWords	11
PhotoAmt	1
Preparation_1	0
Quantity	1
RescuerFreq	4
Rescuer_999	1
Rescuer_c00756f2bdd8fa88fc9f07a8309f7d5d	0
State_41326	0
Sterilized_1	0
ThreeMonths_1	0
Vaccinated_1	0
eyes	0
healthy	0

Fig 50.1 Feature values for example 2 from the cat test set

Fig 50.2 shows the force plot for example 2. We can see that the predicted probability of this cat not being adopted within 100 days is 0.34. Fig 50.2 displays the different contributions for each feature. We can see that 'Rescuer_999, 'DescNumWords', 'PhotoAmt', 'State_41326' and 'ThreeMonths_0' contributed positively towards the target variable, whilst 'RescuerFreq', 'Quantity', 'Sterilized_1' and 'Age' contributed negatively towards the target variable. We can see that the 'ThreeMonths_1' feature contributed the most towards the predicted value, this means that 'ThreeMonths_1' is an important feature for this prediction. From previous analysis we know that cats that are older than 3 months are more likely to not be adopted within 100 days, this is reflected by the positive contribution of the 'ThreeMonths_1' feature here as the cat is older than 3 months.

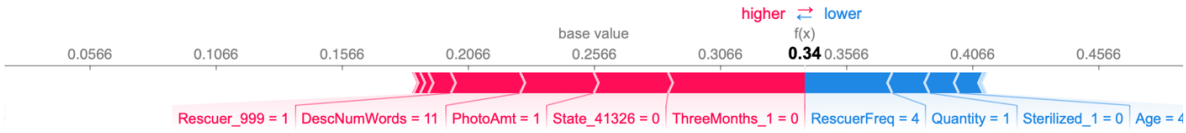


Fig 50.2 Force plot for example 2 from the cat test set

6. Conclusion

6.1 Dog adoption dataset

Based on our best machine learning model and subsequent analysis we now understand the behaviour and general trends of the dog adoption dataset with respect to different variables. The random forest classifier model gave the best predictions with an AUC score of 0.8355. Our model tells us that the average predicted probability for the training data is 0.295, which means that the average dog in the dataset has a 29.5% chance of not being adopted within 100 days.

Based on further analysis using SHAP, we were able to identify the most important features, how they impacted the model and how they interacted with other features. The most important trends to note for the dog adoption dataset are:

1. The higher the rescuer frequency, the more likely the dog will be adopted within 100 days.
2. Dogs not of dominant breed 307 (mixed breed) have a higher probability of being adopted within 100 days.
3. The younger the dog, the more likely the dog will be adopted within 100 days.
    a. Specifically, those 3 months or younger
    b. Dogs that are 3 months or younger and are of dominant breed 307 (mixed breed) have the highest probability of being adopted within 100 days
4. The more photos attached on the dog's profile, the more likely the dog will be adopted within 100 days.
    a. Optimal number of photos is between 7 - 10
5. Dogs of pure breed have a higher probability of being adopted within 100 days.
6. The smaller the quantity of dogs in one listing, the more likely the dog will be adopted within 100 days.
    a. Optimal quantity is 1
7. Unsterilized dogs are more likely to be adopted within 100 days
8. Dogs in Selangor are more likely to be adopted within 100 days
9. Dogs that are not fully prepared (sterilized, vaccinated and dewormed) are more likely to be adopted within 100 days
10. Dogs rescued by Rescuer_fa90fa5b1ee11c86938398b60abc32cb are more likely to be adopted within 100 days
    a. Dogs are even more adoptable if they're also sterilized
11. Dogs not rescued by Rescuer_ee2747ce26468ec44c7194e7d1d9dad9 are more likely to be adopted within 100 days
    a. Dogs are even more adoptable if they're also unsterilized
12. Dogs rescued by Rescuer_b53c34474d9e24574bcec6a3d3306a0d are more likely to be adopted within 100 days
    a. Dogs are even more adoptable if they're also older than 3 months
13. Beagles are more likely to be adopted within 100 days
    a. Beagles that are older than 3 months are even more adoptable

14. Dogs rescued by Rescuer_c00756f2bdd8fa88fc9f07a8309f7d5d are more likely to be adopted within 100 days
    a. Dogs are even more adoptable if they're also 3 months or younger

6.2 Cat adoption dataset

Based on our best machine learning model and subsequent analysis we now understand the behavior and general trends of the cat adoption dataset with respect to different variables. The random forest classifier model gave the best predictions with an AUC score of 0.7725. Our model tells us that the average predicted probability for the training data is 0.257, which means that the average cat in the dataset has a 25.7% chance of not being adopted within 100 days.

Based on further analysis using SHAP, we were able to identify the most important features, how they impacted the model and how they interacted with other features. The most important trends to note for the cat adoption dataset are:

1. The higher the rescuer frequency, the more likely the cat will be adopted within 100 days.
2. The younger the cat, the more likely the cat will be adopted within 100 days.
    a. Specifically, those 3 months or younger
    b. Cats are even more adoptable if they're also sterilized
3. The more words included in the description, the more likely the cat will be adopted within 100 days.
    a. Optimal number of words is between 200 – 400.
4. The smaller the quantity of cats in one listing, the more likely the cat will be adopted within 100 days.
    a. Optimal quantity is 1
5. The more photos attached on the cat's profile, the more likely the cat will be adopted within 100 days.
    a. Optimal number of photos is between 9 – 12
6. Unsterilized cats are more likely to be adopted within 100 days
7. Cats in Selangor are more likely to be adopted within 100 days
    a. Especially those that are older than 3 months
8. Cats rescued by the top 10 rescuers are more likely to be adopted within 100 days
    a. Especially those that are sterilized
9. Vaccinated cats are more likely to be adopted within 100 days
    a. Especially those that are older than 3 months
10. Cats with the word 'healthy' in their description are more likely to be adopted within 100 days
    a. Especially those rescued by unpopular rescuers
11. Cats with the word 'eyes' in their description are more likely to be adopted within 100 days
    a. Especially those rescued by unpopular rescuers
12. Cats rescued by Rescuer_c00756f2bdd8fa88fc9f07a8309f7d5d are more likely to be adopted within 100 days

7. Recommendation

It is clear that there are many factors driving pet adoption, most of which cannot be controlled, such as the pet's breed, gender or health status etc. The recommendations in this section will focus on factors rescue centres can manipulate in order to achieve a higher adoption rate.

7.1 Dog Adoption Dataset

Among all the features, rescuers are able to control photo amount, quantity, sterilization, preparation (combination of sterilization, vaccination and deworming) and state.

Based on the SHAP analysis, I would recommend rescuers to upload at least 7 photos. Adding more photos won't increase the probability of adoption much. I would not recommend adding more than 13 photos as this may actually decrease the probability of adoption.

The optimal quantity for dog adoption is 1. Adding additional dogs to the listing will decrease its adoptability. I would recommend separating groups of dogs into individual listings to increase adoptability. If that is not possible, I would recommend at most 3 dogs in one listing. If there are 4 dogs in one listing, adding additional dogs will not affect the adoptability.

Although sterilized dogs show a lower probability of adoption in our model, I would not recommend against sterilization based on domain knowledge. Sterilization is important to help dogs live a longer and healthier life by eliminating a number of health problems. When looking at interactions between different features and sterilization, we can see that dogs rescued by rescuer fa90fa5b1ee11c86938398b60abc32cb and ee2747ce26468ec44c7194e7d1d9dad9 or are located in Negeri Sembilan are much more likely to be adopted if they are sterilized. Thus, I would highly recommend sterilization in these two rescues and the all rescues in the state of Negeri Sembilan.

Similar to sterilization, preparation also shows a lower probability of adoption in our model. Vaccination and deworming are also very important to keep dogs healthy and prevent the spread of diseases, as such I would not advise against it despite the model's predictions.

It is common knowledge that young dogs are more adoptable than older dogs (especially puppies). However, our model shows that rescuer 'b53c34474d9e24574bcec6a3d3306a0d' and 'c00756f2bdd8fa88fc9f07a8309f7d5d' have high adoptability for older dogs. Rescues located in Selangor also have high adoptability for older dogs. I would recommend looking at the adoption strategies adopted in these rescues and maybe even transfer some older dogs to these rescuers if they have the capacity.

The adoption probability in Melaka and Negeri Sembilan are very low. I would recommend running some awareness programs that can educate people in those areas about the importance of adoption and seriousness of stray dogs. It may also be worth taking a look at

the strategies adopted by rescues in Selangor which have a much higher adoption probability.

Since dogs rescued by rescues with a low rescuer frequency tend to have a low probability of adoption, especially rescues with a rescuer frequency of 1. I would recommend transferring dogs that are individually rescued to larger rescues to increase their chances of adoption.

7.2 Cat Adoption Dataset

Among all the features in our model, rescuers are able to control photo amount, quantity, sterilization, preparation (combination of sterilization, vaccination and deworming), profile description and state.

First, I would recommend including at least 50 words in each cat's profile description. This allows adopters to better understand the cat, thus increases chances of adoption. For rescues with a low rescuer frequency, I would also recommend using the keywords 'healthy' and 'eyes' in the cat profile description as this seems to increase the adoptability of cats. For rescues with a high rescuer frequency, I would advice against using the keywords 'healthy' and 'eyes' as this seems to decrease the cat's adoptability.

Attaching photos will also help adopters visualize their potential cat. The more photos attached, the more likely the cat will be adopted. I would recommend attaching at least 1 photo as adding 1 photo significantly increases the chances of adoption. The optimal number of photos to attach is 9-12. Adding more than 12 photos doesn't appear to further increase the chances for adoption.

Another important feature in the model is 'Quantity'. Cat listings with a small quantity have a higher chance of being adopted. Cat listings with only 1 cat have the highest chance of being adopted. I would recommend separating groups of cats into individual listings to increase adoptability. If that is not possible, I would recommend at most 9 cats in one listing. If there are 10 cats in one listing, adding additional cats will not further diminish the adoptability.

Although sterilized cats show a lower probability of adoption in our model, I would not recommend against sterilization based on domain knowledge. Sterilization is important to help cats live a longer and healthier life by eliminating a number of health problems. For cats rescued by the top 10 rescuers, they are actually more likely to be adopted if they are sterilized. As a result, I would recommend sterilizing all cats from the top 10 rescuers. Older cats are also more adoptable if they are sterilized, I would recommend sterilizing older cats first if there are limited resources. Similar to sterilization, preparation (sterilization, vaccination and deworming) doesn't shows an improvement in cat adoptability but out of cats which are fully prepared, older cats show a higher probability of adoption than younger ones. Thus, I would recommend prioritizing the preparation of older cats before younger cats.

Rescues in Selangor have a high adoption probability, especially for older cats. Rescuer c00756f2bdd8fa88fc9f07a8309f7d5d also has a high adoption probability, especially for older cats. It may also be worth taking a look at the strategies adopted by rescues in Selangor and rescuer c00756f2bdd8fa88fc9f07a8309f7d5d to understand their success and adopt similar strategies in other rescues. Furthermore, if there is ever an overpopulation in rescues outside of Selangor, I would recommend transferring cats to rescues in Selangor to accommodate as cats in Selangor are usually adopted within 100 days.

Since cats rescued by rescues with a higher rescuer frequency have a higher chance of adoption, I would recommend transferring cats that are individually rescued to larger rescues to increase their chances of adoption. Cats rescued by rescues with a rescuer frequency of 15 or more have the highest probability of adoption.

8. Future Scope

Given more time, I think it would've been extremely useful to use image recognition to include the pet adoption images in the machine learning model. Sentiment analysis could also have been used to further analyze the description of the pet profile. Furthermore, I could incorporate more external datasets such as GDP per state to better understand the economic background of adopters in the area which may impact adoption speed in the area.