**R Statistical Programming (ISGB-799V-001)**

**Prof. John Campbell**

**Animal Shelter Project**

**Final Report**

**Group 6**

Dingsheng Kuang

Mengyuan Guo

Saverio Hans

Yezi Zhou

# Introduction

The question this project aims to answer is whether, given the data provided data set, we can predict the outcome of a certain shelter animal. The outcomes can be either positive -- adopted or returned -- or negative -- died or euthanized. We aim to use logistic regression on our data set and validate the results using the K-fold cross-validation method.

Research question:

Are shelter adoption rates influenced by more than just the animal’s physical condition?

Research purpose:

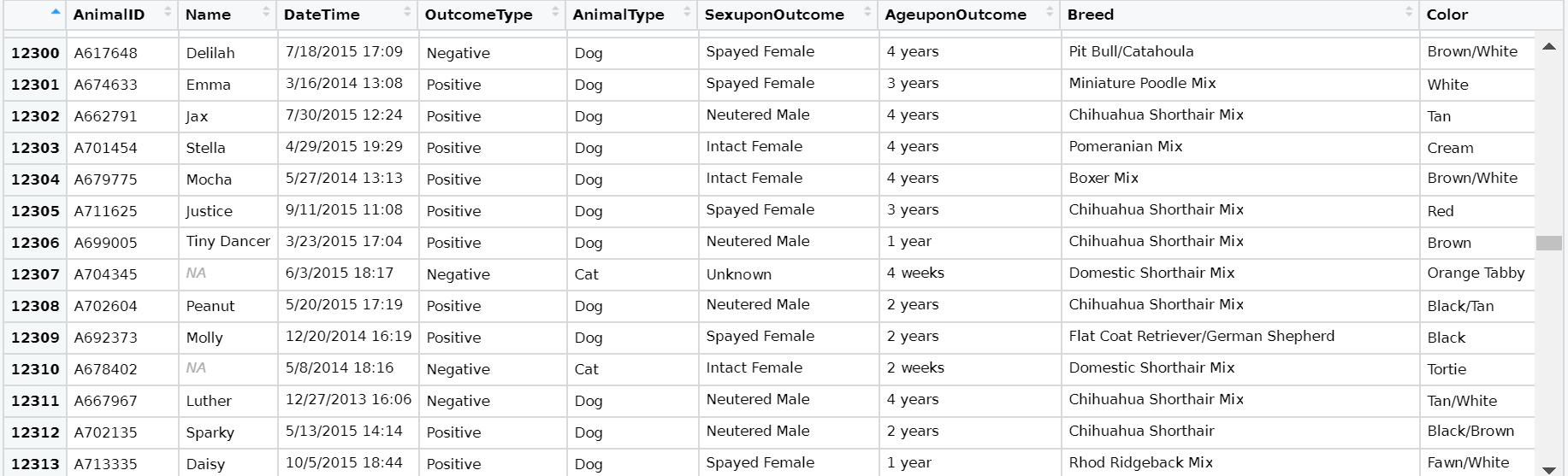
1. Build a model using the shelter animal’s’ physical condition and other external factors to predict whether the outcome of the given animal.

2. Analyze the important factors that lead to adoption and find solutions to improve the adoption rate of less-adopted animals.

# Pre-analysis

## Summary Statistics

Our initial data contained 26729 observations of 9 variables.



Data Dictionary

|  |  |
| --- | --- |
| Data | Description |
| AnimalID | Identification number of the animal |
| Name | Name of animal |
| DateTime | Date and time upon the outcome |
| OutcomeType | Either positive or negative |
| AnimalType | Either dog or cat |
| SexUponOutcome | Gender or neuter/spay info |
| AgeUponOutcome | Age of animal |
| Breed | Breed of animal |
| Color | Color of animal |

## Feature development

### Data pre-processing

1. Transfer “Name” into a dummy variable

A name may not have practical significance, but we assume that a potential owner probably expects to name his/her own pet, or feel less intimate when calling a pet named by others. So, we decided to classify animals into two kinds: with a name and without a name.

Method: with a name – 1, without a name – 0.

2. Transfer “Breed” into a dummy variable

There are 1380 different breeds in the data. The broad variability of entries in the breed column creates difficulty in processing the information. We decided to categorize the breeds into two kinds: mixed breed and purebred.

Method: mixed breed (breed name has the keyword “mix” or has a slash) – 1, purebred (other) – 0.

3. Extract days, months and seasons from “DateTime”

4. Unify the unit of “Age”

Transfer all ages into months, then categorize ages into three kinds: younger than 1 month, 1 month to 1 year, over 1-year-old.

5. Re-classify “Color”

We assume an animal’s appearance to be an important factor in its adoption probability. In our dataset, “Color” directly influences appearance. So, we cannot simply drop entries or categorize them roughly. We decided to shrink 366 colors into 114 by categorizing colors with less than 15 entries to “other”.

## Initial Plots

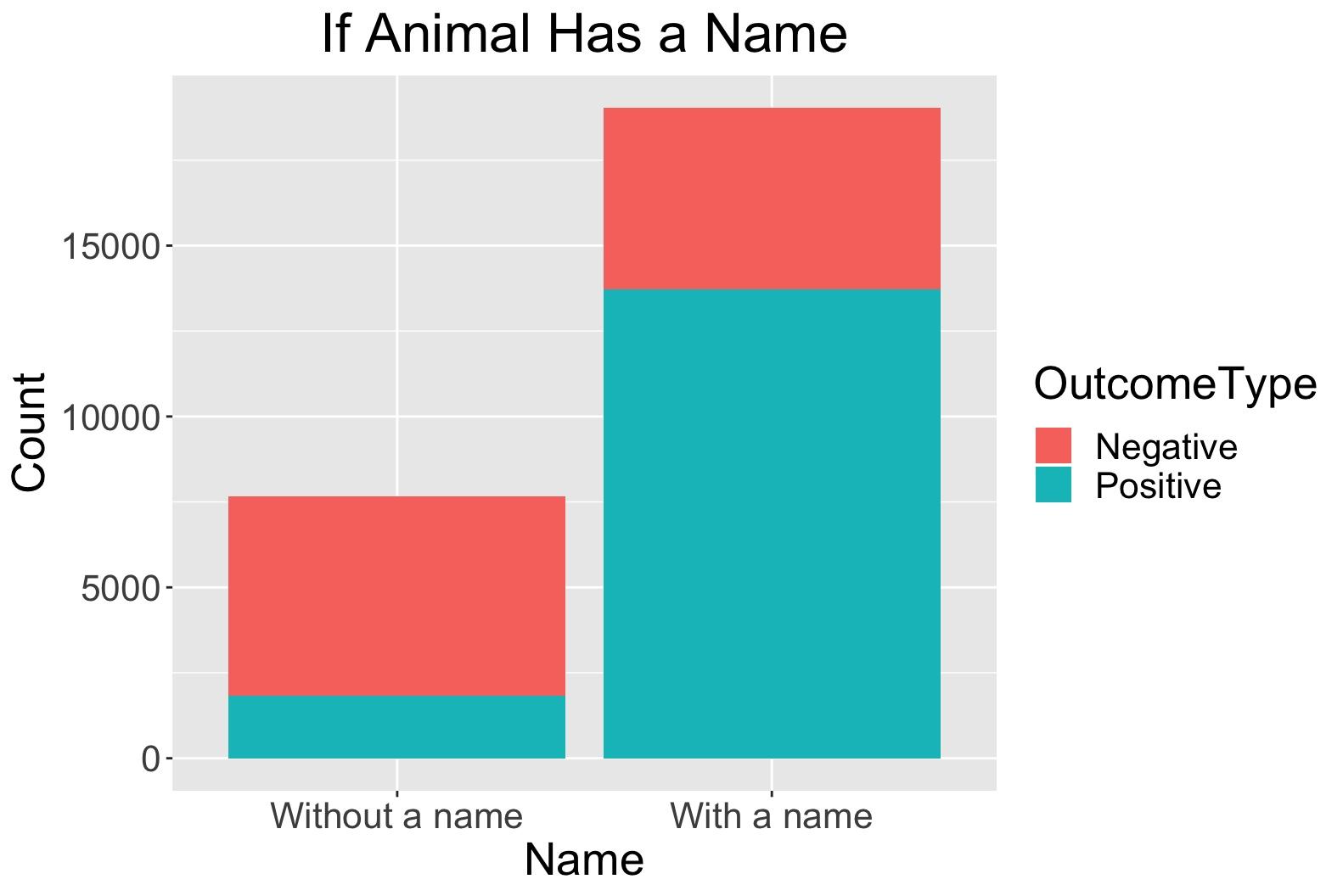


Figure.1 Animal w/ and w/o name comparison

Figure 1 shows the ratio of positive to negative for classes of animals without a name and animals with a name. The graph suggests animals without a name have a significantly lower adoption rate (<30%) compared to animals with a name (>60%). Our expectation was that the name variable will have no effect on the adoption rate of shelter animals. The results of this analysis conflicts with our initial expectations. People are more likely to adopt a shelter animal if the animal had previously been given a name.

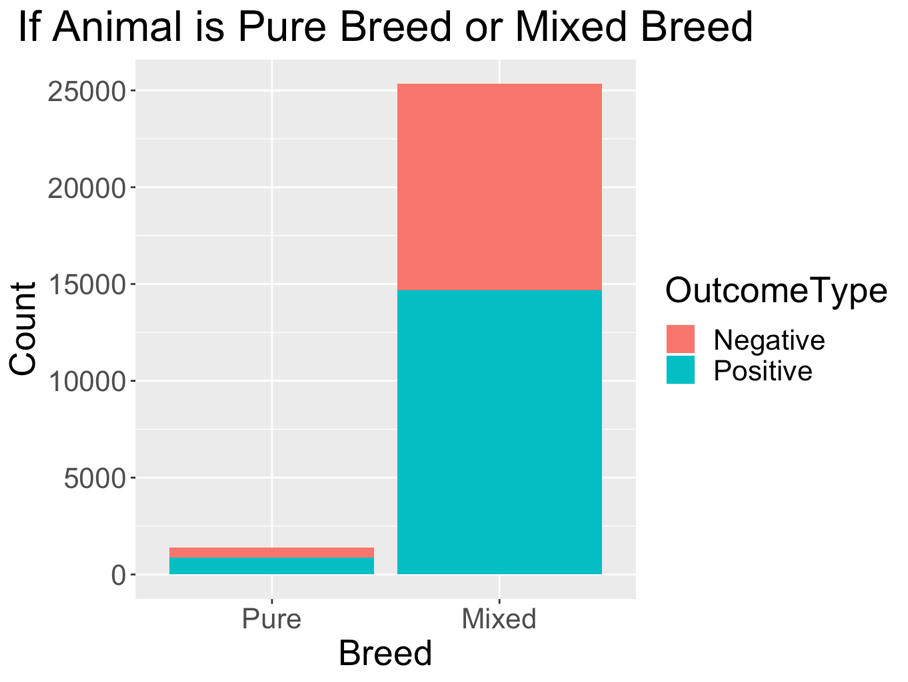


Figure.2 Effect of Animal Breed

Figure 2 shows the ratio of positive to negative for classes of animals that are pure breed and mixed breed. The ratio between the 2 classes are similar. This graph is not significant in summarizing our data as the population of the pure breed class is a low percentage of the total population of shelter animals (<10%). Another factor that we find reasonable is that pet owners who are looking for purebred dogs are more likely to visit a breeder and not a shelter. We think that people who come to shelters do not expect to adopt a purebred animal.

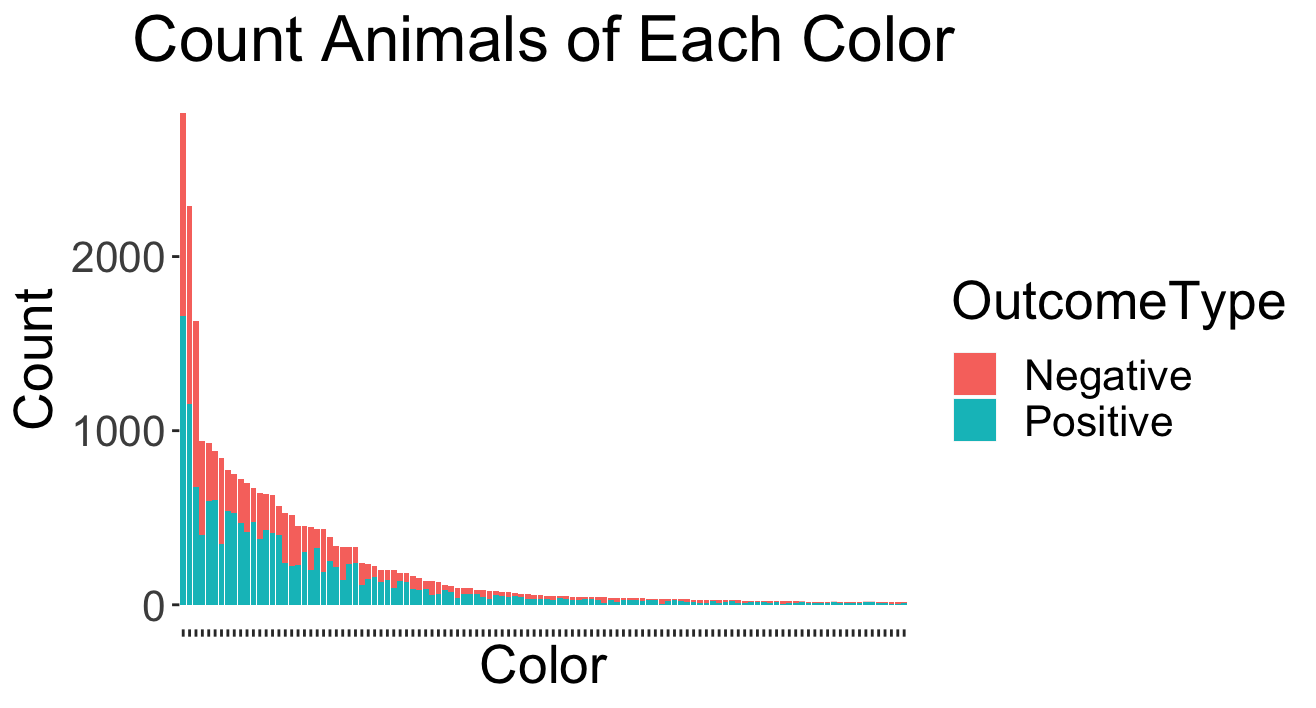


Figure.3 Outcome of Animal Color

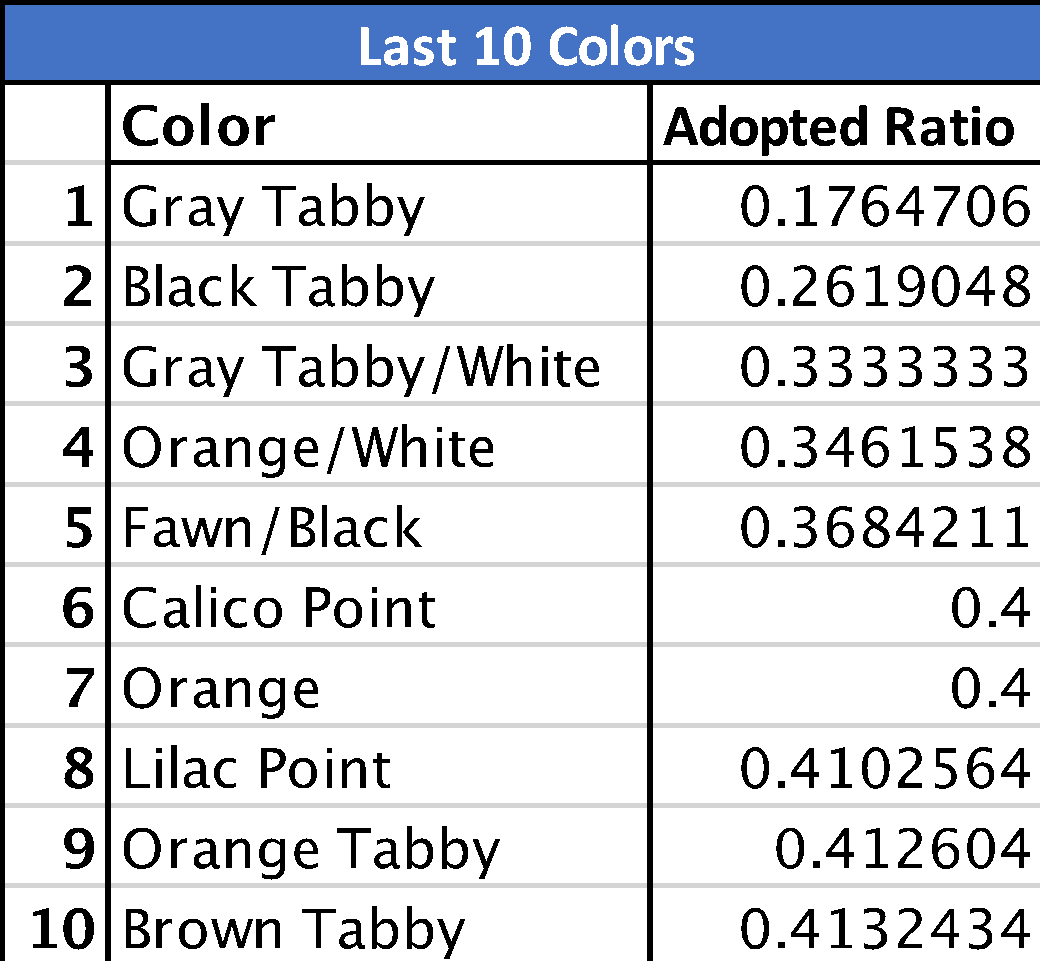
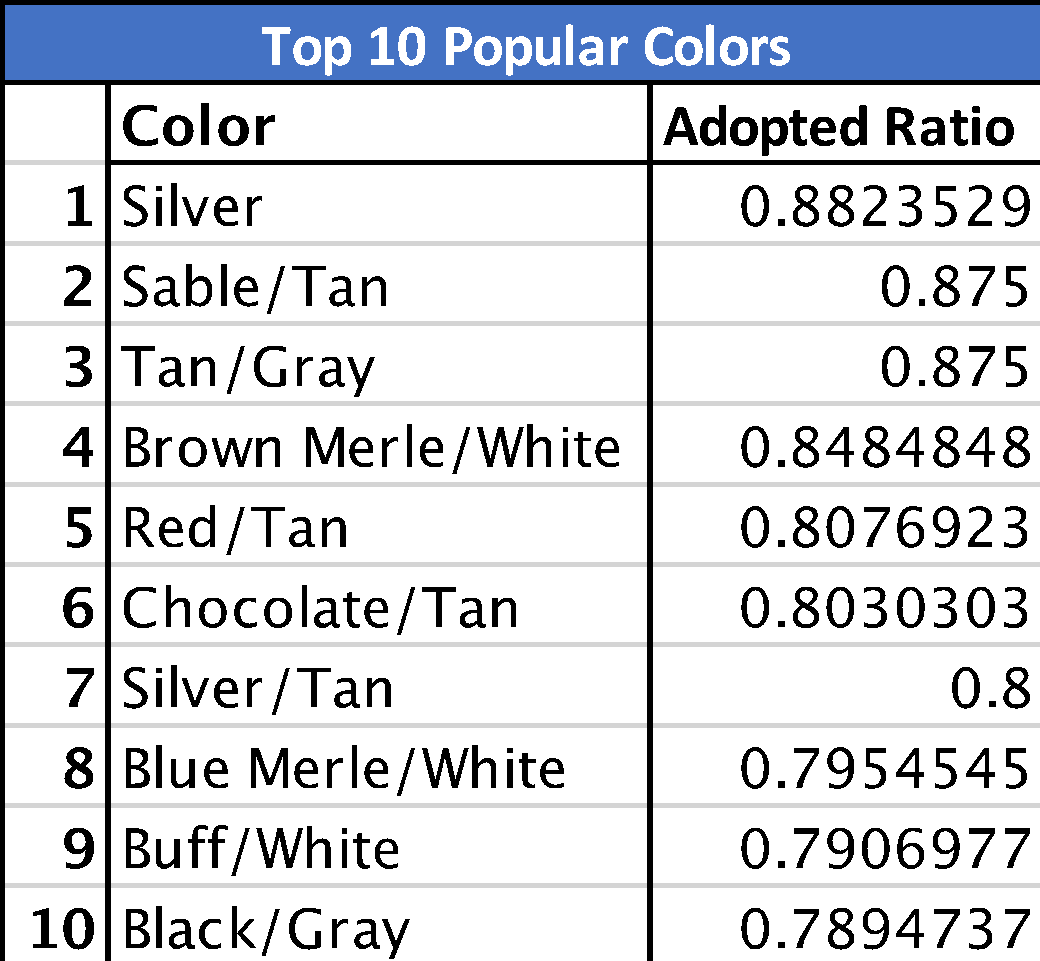


Figure.4 Top 10 and Last 10 color types

Figure.3 indicates that the adoption rate is affected by colors. Each color type has a different ratio. Figure.4 shows the most and least popular color types. Based on the two figures above, categorizing them into simplified color types like single-color, bicolor, or tricolor is not a viable method. For example, silver animals are the most popular and orange animals are in the last 10, but they are all single-color animals if being categorized.

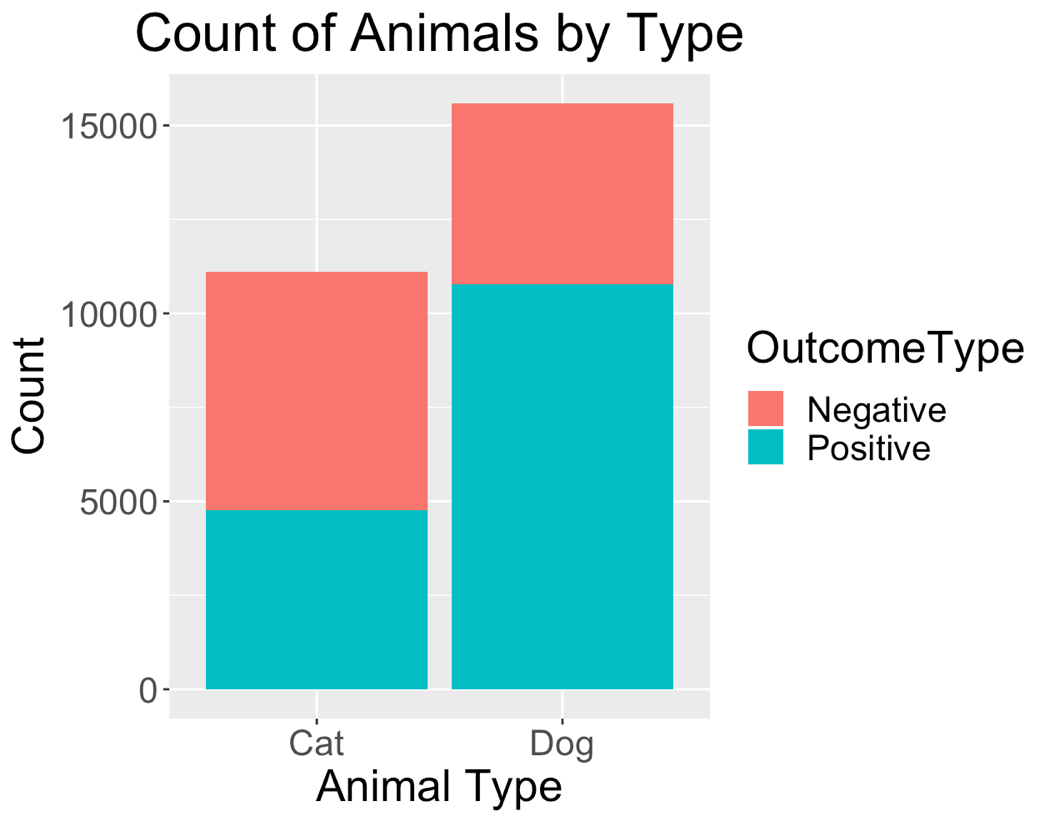


Figure.5 Outcome of Animal Type

Figure 5 shows the ratio of positive to negative outcome between the classes of animals dog and cat. Cats have a lower adoption rate (<50%) while dogs have a higher adoption rate (>60%). We conclude that dogs are more popular than cats. The animal type will be a significant feature in our model.

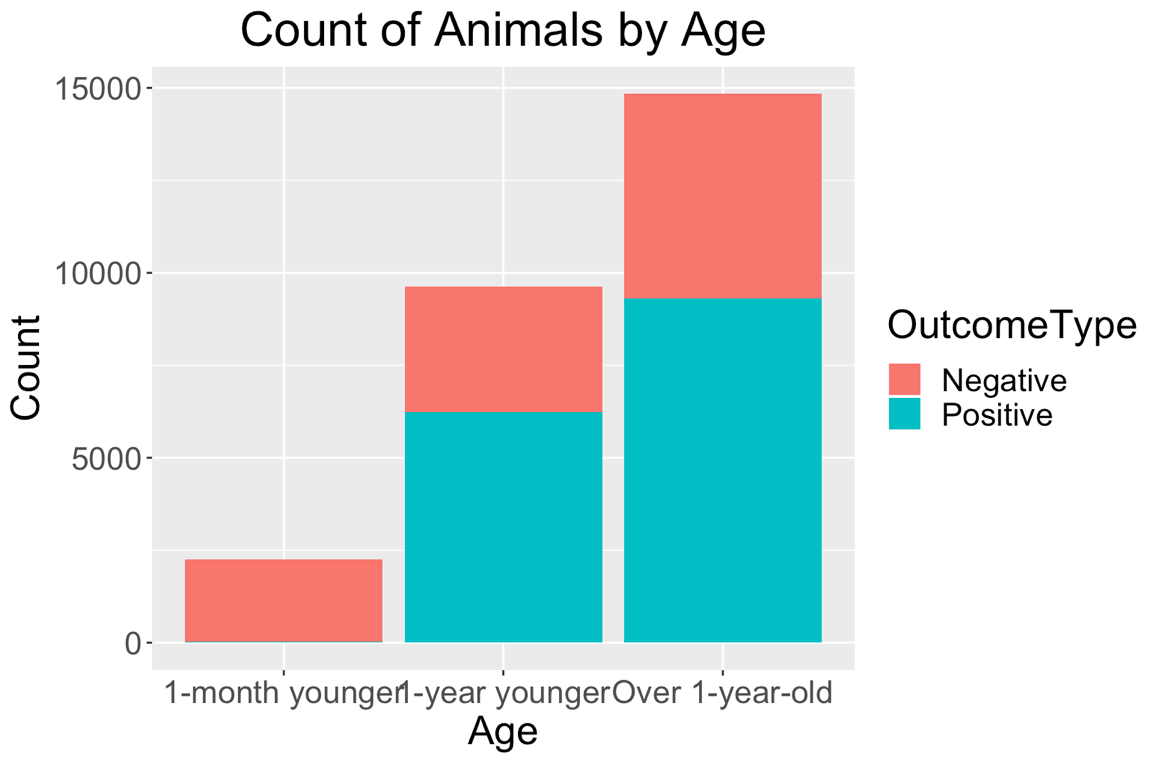
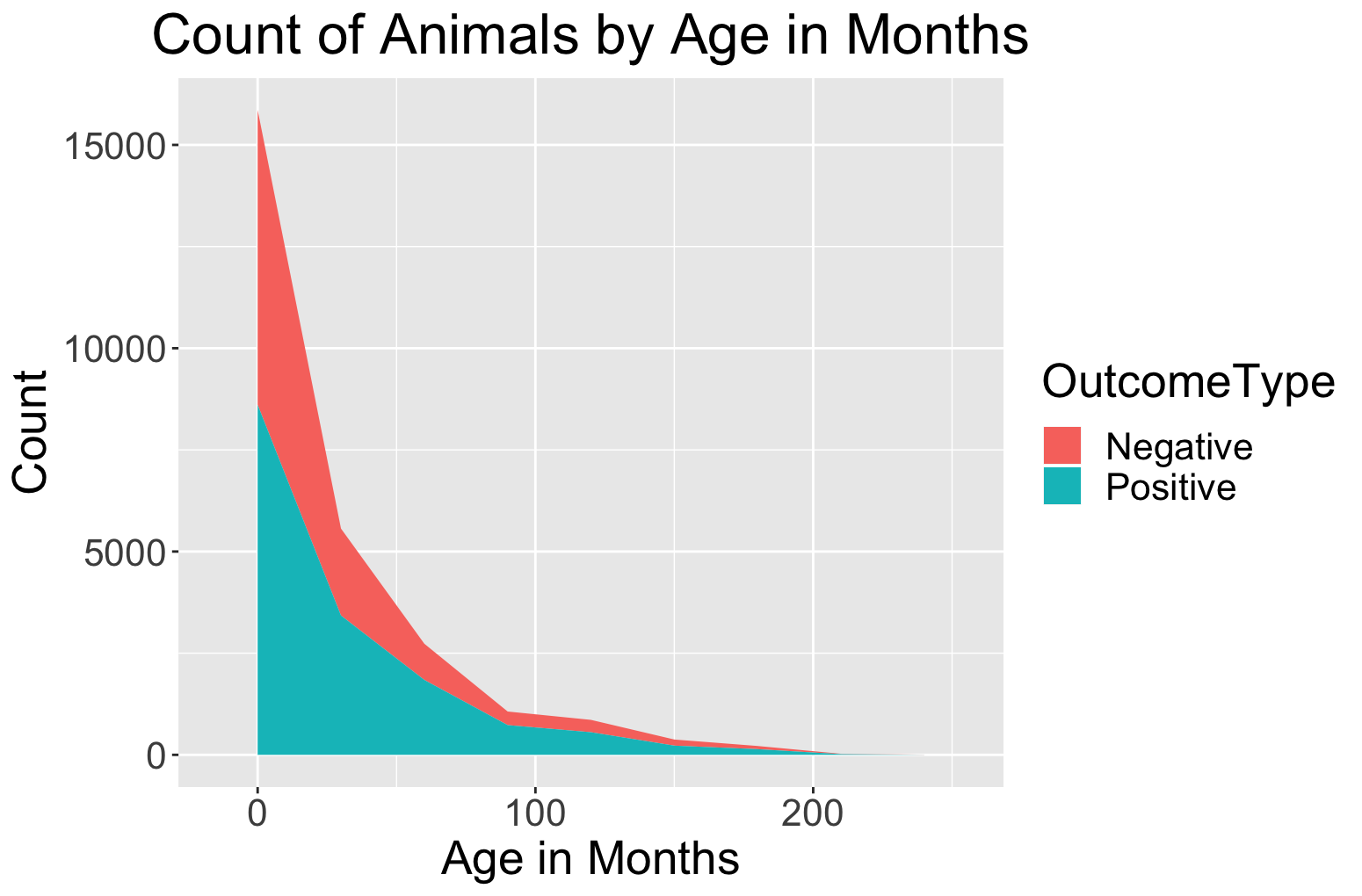


Figure.6 Outcome of Age

Figure 6 compares the adoption rate and age. The first graph shows age as a linear variable and the second graph shows age as a categorical variable that has 3 categories (1-month younger, 1-year younger and over 1-year old). The adoption rate is directly proportional to age. As age increases, adoption rate increases as well. We think younger animals are more prone to sickness and diseases, and might require more visits to the veterinarian for vaccinations. This leads to higher costs of taking care of the animal. Younger animals also need more attention and training to help it adapt to its new home and pet adopters might be deterred from this because of the hassle and lack of expertise in the subject of animal training.

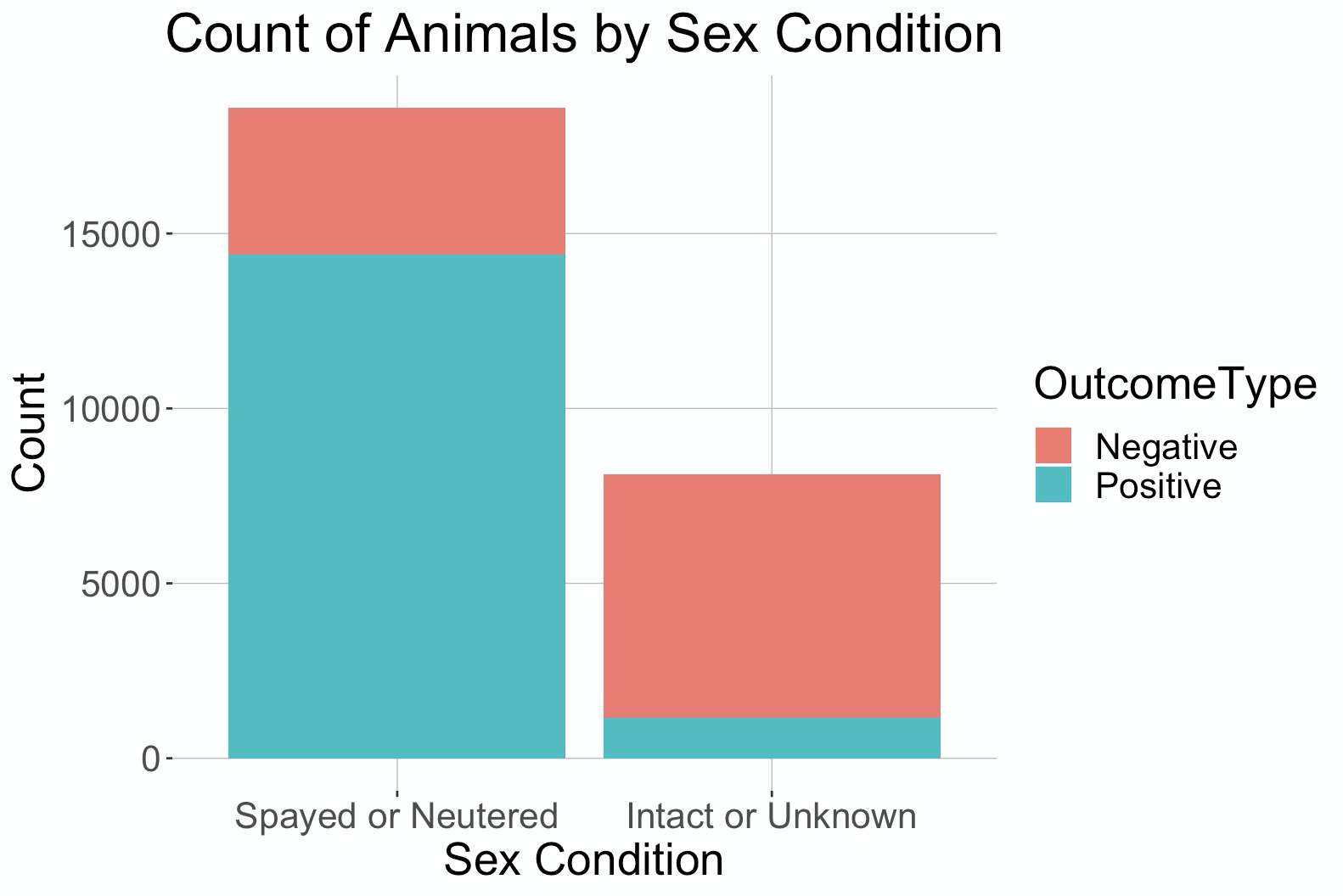
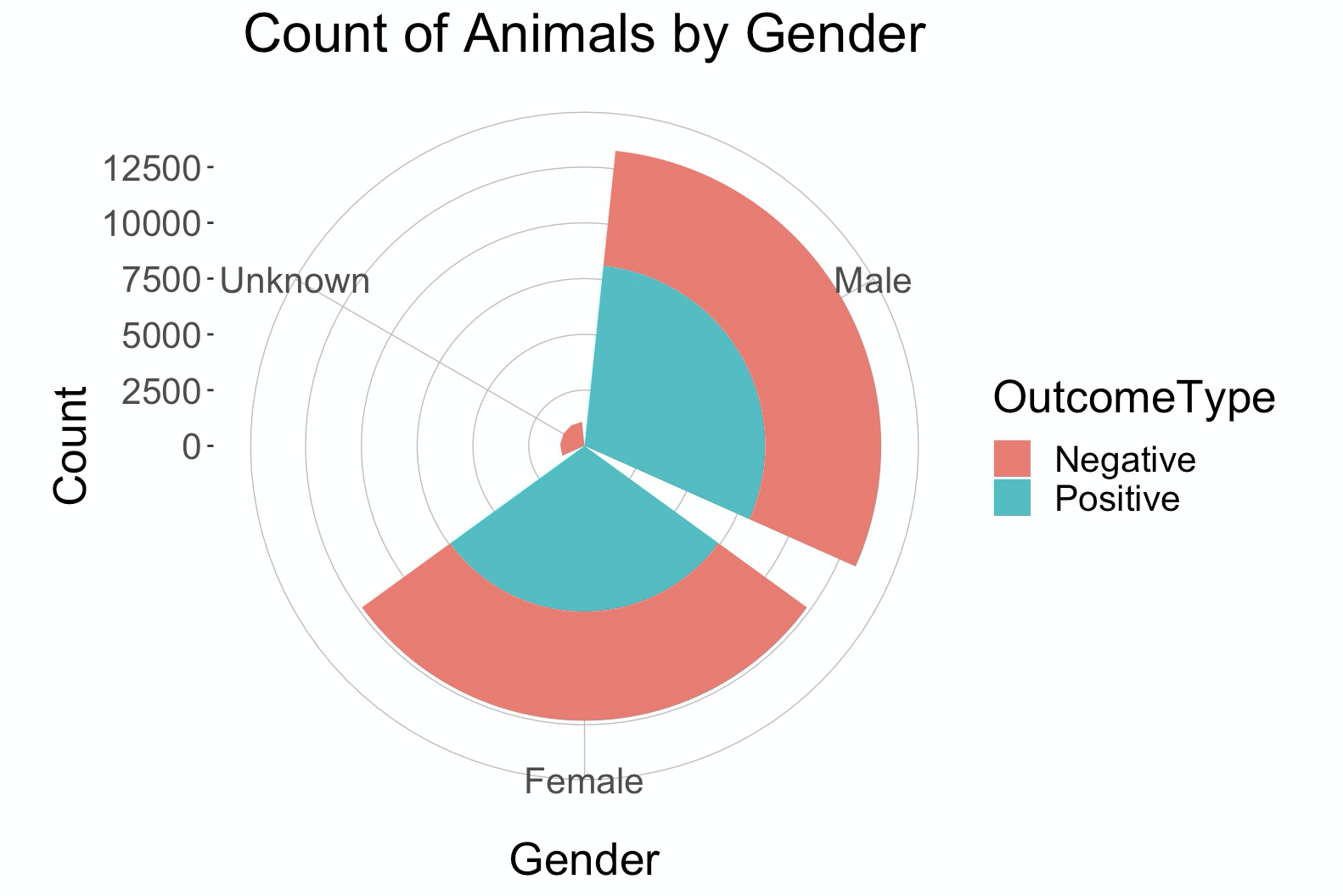


Figure.7 Outcome of Gender and Sex Condition

Figure 7 shows the adoption rate between gender and sex condition classes. The adoption rate between male and female are similar, and the gender of the animal bears no significance in determining adoption rate. The adoption rate for spayed/neutered animals is significantly higher (>75%) than intact/unknown animals (<30%). We think that pet adopters consider the spaying/neutering of the animal because of health improvements to the animal that will reduce the risk of conditions such as animal estrus. Spaying/neutering also clears the owner of risk that the animal will accidentally reproduce.

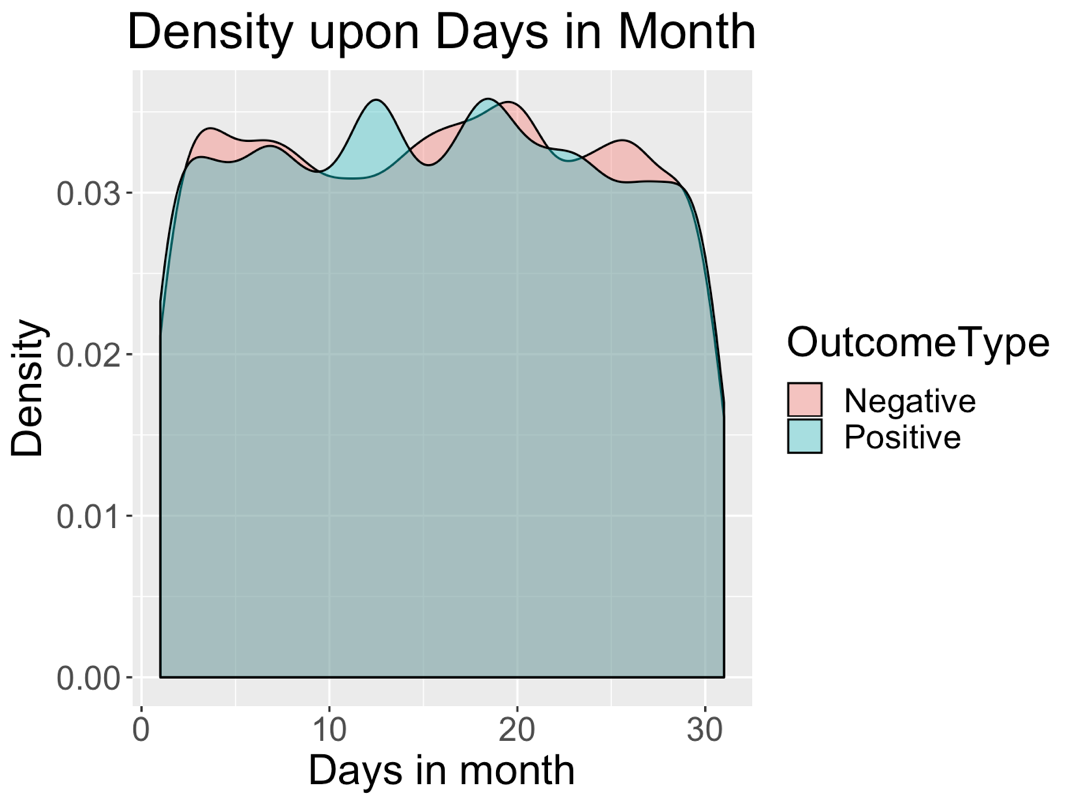


Figure.8 Outcome of different days in a month

Figure 8 shows the density of outcomes based on different days in a month. We find that during certain periods, the density of one outcome is higher than that of the other outcome. For example, more positive outcomes occur between the 10th and 14th of the month. A solution we recommend to help shelters in raising adoption rates is to show more animals and hold public viewing events in those specific periods to increase the number of adoption.

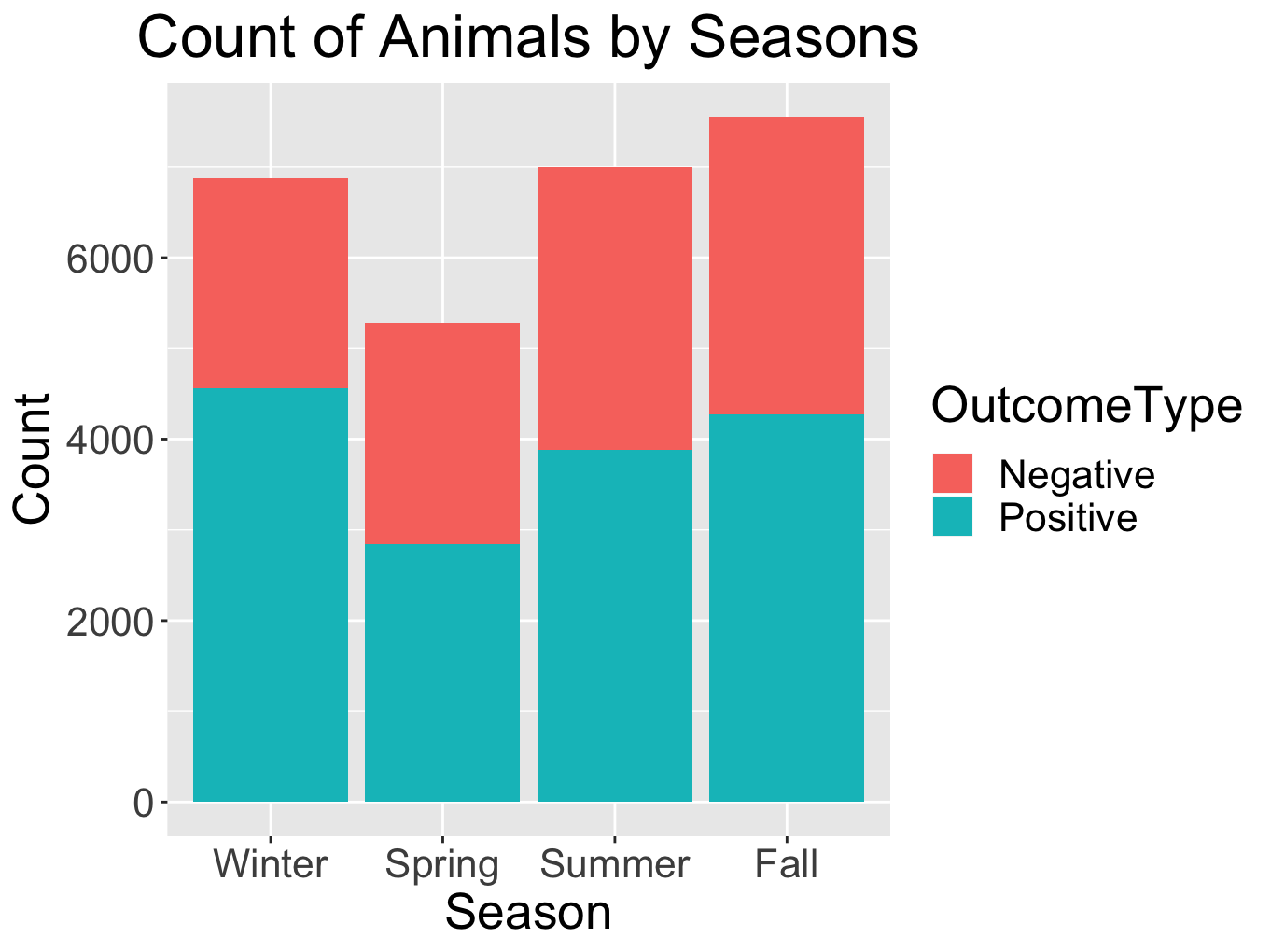


Figure.9 Outcome by Season

Figure 9 shows the adoption rate between seasons. The figure suggests that season is a factor in determining the adoption rate. The results suggest that adoption rates are higher in winter (>50%) compared to the summer or spring (both <50%). We think that social factors in the winter season (holidays, turn of the new year) might have an effect on the willingness of people to adopt an animal that leads it to have a higher adoption rate than all other seasons.

## Initial hypothesis

The adoption rate of shelter animals is influenced not only by animal’s self-condition but also by external factors – including the presence of a name and the date/season the animal is being offered for adoption.

# Modeling

## Model build-up

As our target of prediction is binominal, we built two logistic regression models and applied the K-fold cross-validation method to test our hypothesis.

According to our hypothesis, we want to test whether external factors influence the adoption result. So, we first built a model with all selected variables representing animals’ self-condition, which are: Animal Type, Sterility, Color, and Age.

A screenshot of a cell phone

Description automatically generated

The first model without external factors has an accuracy of 79.93%.

Then, we added our selected variables as external factors to build the second model, including: Name, Day, and Season.

A screenshot of a cell phone

Description automatically generated

The second model with external factors included has an accuracy of 80.71%. The average accuracy increased by .78%, compared to the first model.

## Model Output

A screenshot of a cell phone

Description automatically generated

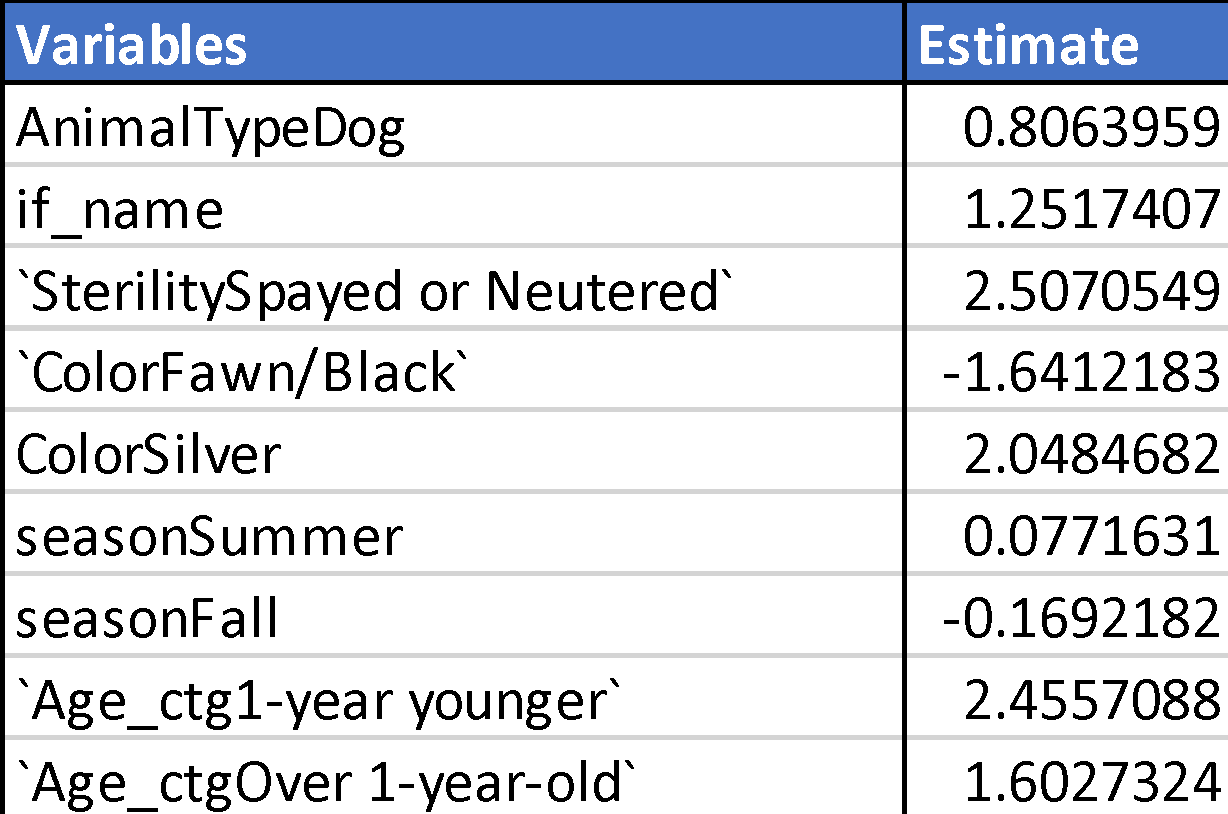
…

A screenshot of a cell phone

Description automatically generated

From the above output we see, in our second model, the addition of 122 (26709-26587) independent variables decreased the deviance from 36300 to 24051, a significant reduction in deviance. **The Residual Deviance has reduced by 12249 with a loss of 122 degrees of freedom**. By comparing to the first model (residual deviance reduced by 11264 with a loss of 117 degrees of freedom), the second model containing external factors as variables has an even better fit.

**Variables & Coefficients**



The above table shows only variables with a P-value of less than 0.1. These variables have coefficient estimates significantly different than 0.

Seen from the table, we see that **sex conditions** can be the most important influencing factor to the adoption result. Holding all other variables to a fixed value, spayed or neutered animals have odds about 1130.5% (exp(2.51)-1) higher than intact or unknown animals. Same, dogs have odds about 120% higher than cats. Some colors also play a role, including Fawn/Black and Silver. Also, the odds value for animals older than 1 month is significantly higher than animals younger than 1 month.

Among our selected external factors, **name** and **season** have been tested to be significant, meaning that changes in these external factors should influence the adoption result. For now, our hypothesis is tested to be valid.

## Test of Assumptions

As our model contains only 1 numeric variable (day), the test of linearity assumption, influential values, and multicollinearity are not needed.

# Outcomes

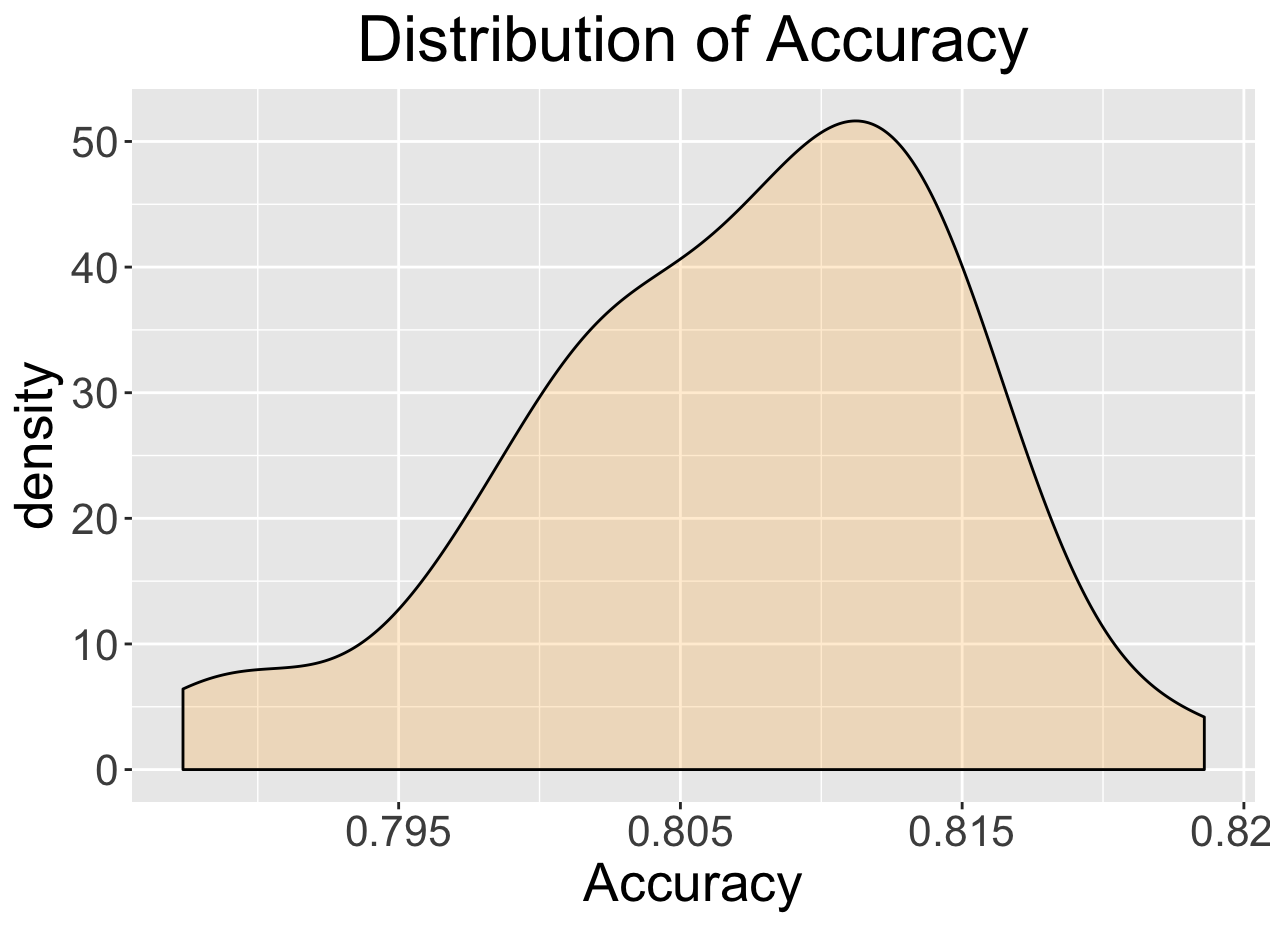


Figure.10 Distribution of Accuracy -- Model2

The average accuracy is 0.8071. According to the plot, most of the results fall between 0.80 to 0.815. In general, the model should have an accuracy between 0.80 to 0.815.

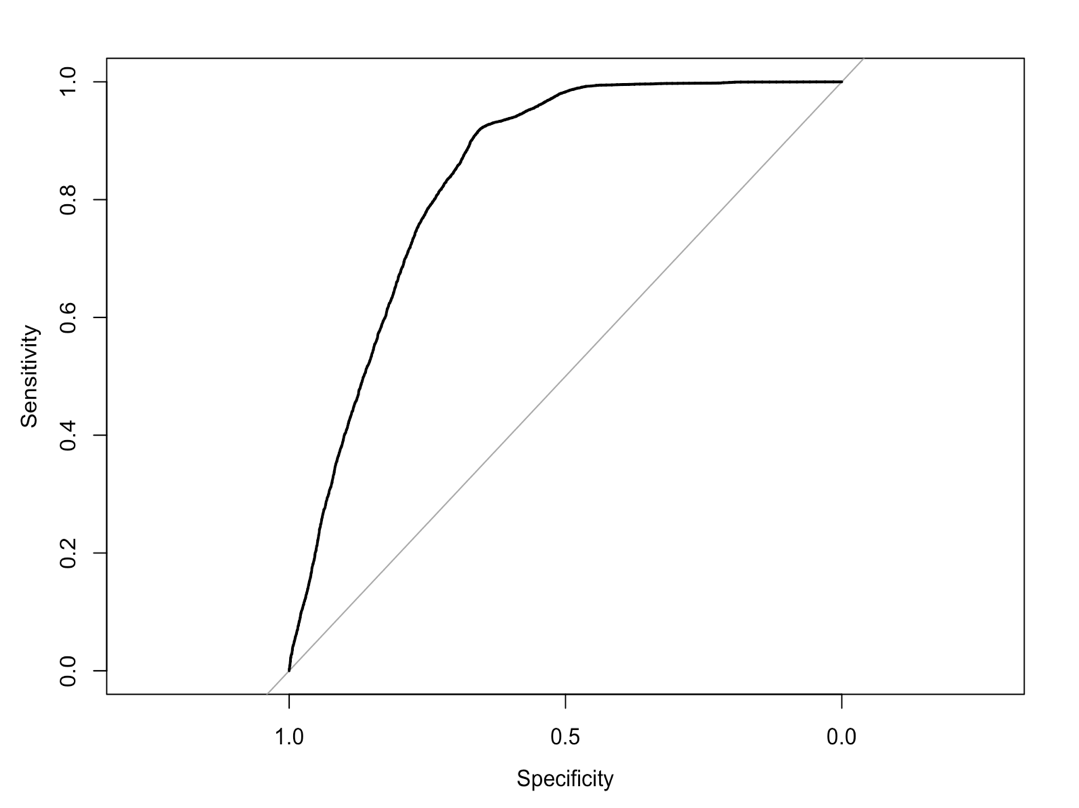


Figure.11 ROC Curve -- Model2

The model has 83.73% of the area under the ROC curve. In the ROC curve, each point on the curve represents a sensitivity/specificity pair corresponding to a particular decision threshold. A test with perfect discrimination (no overlap in the two distributions) has a ROC curve that passes through the upper left corner (100% sensitivity, 100% specificity). As the curve has more than 80% of the area under the curve, the overall accuracy of our model is relatively high.

# Conclusion

The animal’s sexual condition is the most important variable in determining whether an animal will have a positive or negative outcome. The presence of a name and the season in which the animal is being offered for adoption is also an important factor. Our model that included external variables also improved the accuracy of our original model. We conclude that our hypothesis is valid. External factors such as day, season and name do affect the adoption rate of shelter animals.

To improve adoption rates in shelters, we recommend all animals to be spayed/neutered as this will appeal positively to most adopters. Holding public viewing events at public places such as parks and shopping malls during days and seasons in which more positive outcomes happen is also a viable solution to promote the animals and give them more exposure to potential adopters. We also recommend that shelters also give their animals names instead of only a numerical identifier to give the animal more personality during private or public viewings.