Springboard Data Analytics Course

Pet Adoption Final Report for Capstone I

October 2019

John L. Parsons

Contents

[**Introduction** 3](#_Toc21399352)

[**1.1 Problem Statement** 3](#_Toc21399353)

[**1.2 Dataset Summary** 4](#_Toc21399354)

[**Data Wrangling** 5](#_Toc21399355)

[**2.1 Missing Data Points** 5](#_Toc21399356)

[**2.2 Outliers for Selected Input Variables** 5](#_Toc21399357)

[**2.3 Handling Multifactorial Categorical Variables** 6](#_Toc21399358)

[**Results** 8](#_Toc21399359)

[**3.1 Pet Adoption Results Overview** 8](#_Toc21399360)

[**3.2 Model Development** 9](#_Toc21399361)

[**3.2a Model Results** 9](#_Toc21399362)

[**3.3 Survival Analysis Model** 14](#_Toc21399363)

[**3.4 Text Analysis** 16](#_Toc21399364)

[**3.4a Text Analysis Results** 17](#_Toc21399365)

[**Discussion** 20](#_Toc21399366)

[**Conclusion** 26](#_Toc21399367)

[**References** 28](#_Toc21399368)

[**Appendix** 29](#_Toc21399369)

# **Introduction**

## **1.1 Problem Statement**

The Pet Adoption Dataset will have three business objectives. The first objective will provide the champion model to predict which animals are the most likely to be adopted from the shelters. This information will be used in the Adoption Campaign to make the necessary changes to minimize shelter length.

The second objective will build a Lifeline Survival Analysis Model. The Lifeline model can predict which shelter animals are the least likely to be adopted from the shelters. The Kaplan-Meier plots will be created to go more in-depth for the survival analysis of specific cohorts to maximize adoption rates.

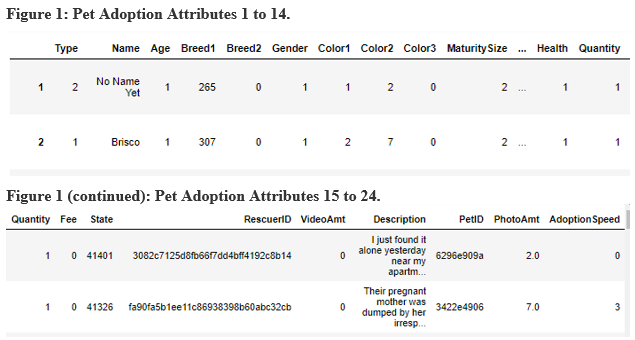
The third goal will determine if the name of a pet will increase an animal’s chance of being adopted from the shelter. The Naïve Bayes model for the text analysis will be used to determine which names are the best predictors for **AdoptionSpeed**.

## **1.2 Dataset Summary**

The Pet Adoption Dataset will be used for this project and came from the Kaggle Competition website (<https://www.kaggle.com/c/petfinder-adoption-prediction>). The goal of the competition was to predict the length of time animals stayed in the shelter and this was labeled as **AdoptionSpeed**. The **AdoptionSpeed** of the animals in the shelters was then predicted from a group of two files. These files were the training and pet images file. There were additional Excel CSV files for the description of the coat color, breed and state information and the test file for making the final prediction.

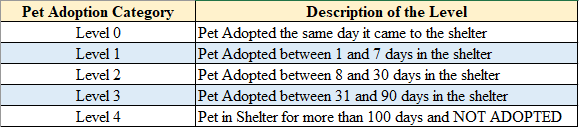
The training dataset will only be used for this project to meet the business objectives. The Pet Adoption training dataset contains a total of 24 attributes and 14,993 rows of data.The list of all 24 attributes can be seen in Figure 1 and attribute description in Figure 1A (Please note all figures with an A are in the Appendix). The Pet Adoption dataset had three types of data classifications listed in Python. The majority of these were Int64 for 19 variables, four were listed as Objects (**Name, Description, PetID** and **RescueID**) and only one was listed as a Float64 (**PhotoAmt**).

**Figure 1: List of all 24 Attributes found in the Pet Adoption Dataset.**



The **AdoptionSpeed** attribute is an ordinal variable that shows the length of time an animal stays in the shelter and is the target variable. The **AdoptionSpeed** variable has a total of five levels based on the age of the animal entering the shelter and the description of the levels can be seen in Figure 2. There were no animals that stayed in the shelter between 90 and 100 days and no animals were adopted if they were in the shelter for more than 100 days.

**Figure 2: Description of the Pet Adoption Category.**



# 

# **Data Wrangling**

## **2.1 Missing Data Points**

The training data set did not have any missing values for the numeric variables in the dataset. There were two non-numeric variables that contained missing values and this was the **Name** and the **Description** of the pet. The **Name** variable had a total of 1,257 missing values and the **Description** only had 12 missing values. The missing rows will be removed for the text analysis section of this study.

## **2.2 Outliers for Selected Input Variables**

Bivariate boxplots were set up with the **Age** and **Breed1** variables and this shows several outliers for these input attributes. A total of 9,061 out of 14,993 pets were adopted that were listed between 0 and 5 days old when entering the shelter. The **Age** attribute will have two types of transformations to reduce the number of outliers. The Age attribute will be Log10 transformed to minimize the number of outliers. The Age attribute will also be binned into a total of seven categories to evenly distribute the data. The bins will contain the following groups; ages between 0 to 1, 1 to 2, 2 to 3, 3 to 5, 5 to 10, 10 to 25 and any value more than 25 were placed in the remaining seventh bin (Figure 2A).

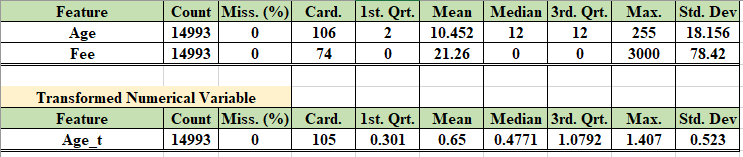
The **Breed1** Attribute had a total of 176 unique factors for this variable and the bivariate boxplot shows several outliers. There were two breeds that dominated the number of pets in these agencies and this is 307 which accounted for 5,927 of animals and 266 which accounted for 3,634 of the animals. Breed 307 was Mixed Breeds for dogs and 266 was a Domestic Short Hair animal for cats. Breeds 266 and 307 remained as one category and then the 174 unique breeds were placed into two groups. All breeds that were less than or equal to 240 were placed in the bin named 265 (remaining dog breeds) and all breeds greater than and equal to 241 (except for 266 and 307) were placed in another bin called 275 (remaining cat breeds) (Figure 3A). The goal is to isolate the two largest breeds for both types into two groups and bin the others to increase model performance.

## **2.3 Handling Multifactorial Categorical Variables**

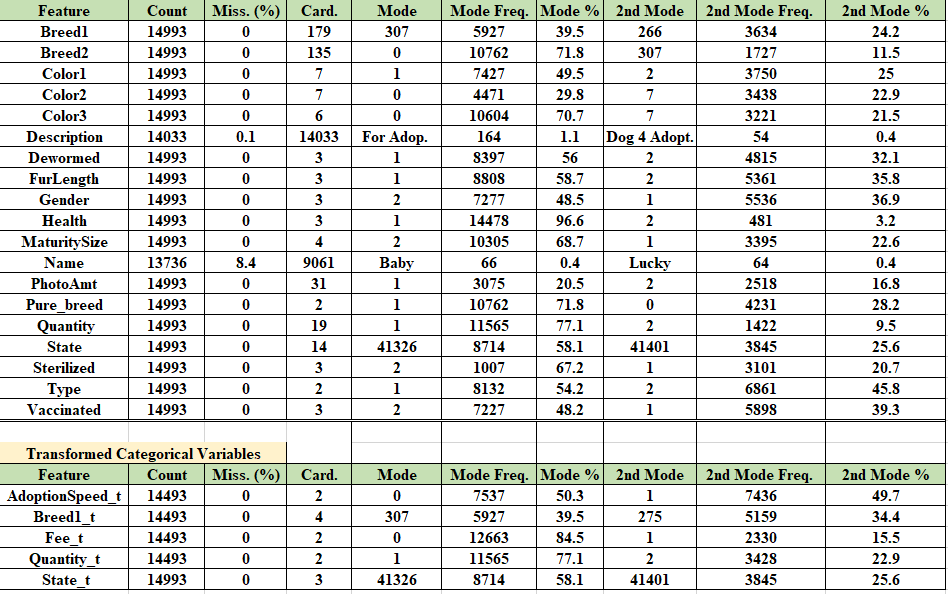
The data wrangling step for the target variable will merge the five-level ordinal variable to a binary variable and will be used for the optimized model. The target variable will have levels 0, 1, 2 and 3 binned into a 0-level and is for any animal that has been in the shelter for less than or equal to 90 days. The target variable level 4 will be binned into the 1-level for animals that has been in the shelter for more than 100 days and were not adopted. The multi-level and binned Target variable results can be seen in Figure 5A.

There are two input variables that had more than 7 levels or high cardinality for this study and are of interest for the optimized models. These two input variables are the **State** or location in Malaysia and **Quantity** or number of pets represented in the profile for each kennel. The **State** variable had a total of 14 nominal levels and 41326 accounted for a total of 8,714 of the pets and 41401 accounted for 3,845 of the pets. Number 41326 is from Selangor and 41401 is from Kuala Lumpur in Malaysia. The remaining 12 levels were merged into one group called 41000 and the binned count (Figure 4A).

**Figure 3: Data Quality Report for Continuous Features.**



**Figure 4: Data Quality Report for Categorical Features.**



The **Quantity** ordinal variable had a total of 19 levels and range from 1 to 19 for this dataset. The greatest number of adoptions occurred when only one pet was listed at a time and this accounted for 11,565 of these adoptees and only 1,422 for two pets listed at a time. These numbers continually decreased when two or more pets were included for the listing of these shelter pets. The **Quantity** variable will bin all animals that had 2 or more animals in a listing for group 2 (Figure 4A). The summary of the data quality report can be seen in Figures 3 and 4 for the continuous and categorical variables. The Fee variable had a total of 74 unique values and high cardinality. This variable was binned into two categories. All animals that were adopted for free were placed into category 0 and those that were purchased for a dollar value were binned into level 1. The Fee ranged from a value of 0 dollars and comprised 84.5% of the data set up to a value of 3,000 dollars.

# **Results**

## **3.1 Pet Adoption Results Overview**

The results will have three sections that will be discussed to meet the objectives of this report. The first section is model development using f1 scores for model optimization and selecting the champion model for predicting adoption speed. The binned target variable is unbalanced and the accuracy score does not give a good estimate of the True Positives (TP) for this dataset. The f1 score will give a more balanced value between precision and recall. The Survival Analysis Model development will be discussed next and the goal is to provide a model for prediction of AdoptionSpeed and create Kaplan-Meier plots for select cohorts to determine how they affect adoption rates. The last section will contain the text analysis section to determine the best names for the placement of these shelter pets and minimize shelter time.

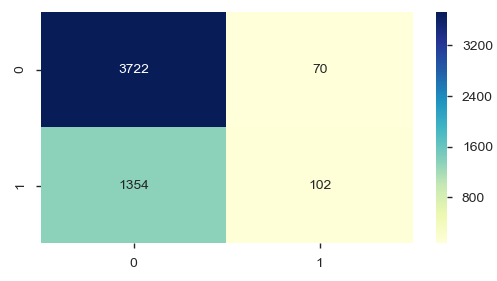
## **3.2 Model Development**

Model development is divided into three iterations. A base model was first generated from the Logistic Model using a 10-fold cross validation for the five-level target variable and the binary target variable. The scores from these two baseline models will be a reference point for model development. The first iteration contains seven models to determine the best f1 results for the binned target variable and will include the original 17 input variables. These seven models are the Random Forest, KNN Neighbors, Logistic Regression, Extra Tree Classifier, Support Vector Machine, Ada Boost and Gaussian Process. The second iteration will contain the binned target and the top nine variables selected by variable importance and logworth values. The data set was split into the training and validation test data with a 65:35 split for all models. All 14,993 entries were included in the first two runs and 9,745 samples were used for the training dataset and 5,248 was used for the test data. The final optimization step used hyperparameters from these seven models, the binary target variable and select eight input attributes. The final optimization step used only 35% of the data and this included 3,410 entries for the training set and 1,837 entries for the test dataset (65:35 split). The best model will be selected based on f1 score and run again with a balanced approach and all 14,993 entries to generate the best score for this champion model.

### **3.2a Model Results**

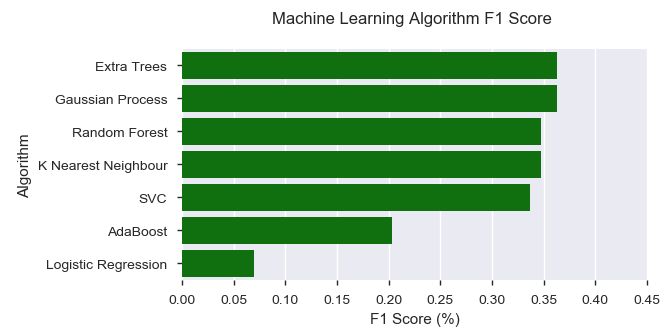
The Logistic Regression baseline model for the original multi-level target variable and all 17 original attributes with accuracy score of 33.7%. The Logistic Regression baseline model with the binary target and all 17 attributes had a f1 score of 13% and an accuracy score of 73%. These values are very low and will be used as the baseline model for this study. The confusion matrix for the binary target baseline Logistic Regression shows the model did a great job in selecting the True Negatives (TN) but did not do well in selecting the TP (Figure 5). The unbalanced dataset is affecting the model and will be adjusted in the champion model.

**Figure 5: Confusion Matrix for the Baseline Logistic Regression.**

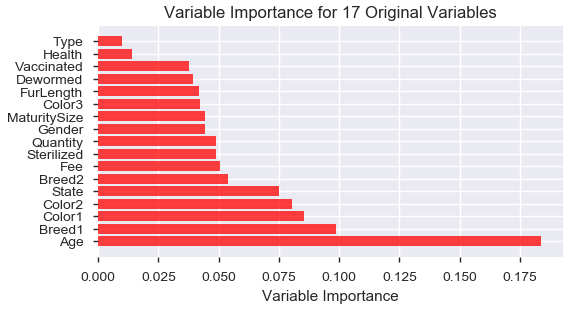


The first iteration using all seven models with the binary target and 17 original attributes can be seen in Figures 6 and 7. The average f1 scores ranged from 7% for the Logistic Regression Model up to 36.3% for the Extra Trees Classifier. The Random Forest variable importance values will be used to measure variable importance and can be divided into three groups. Age had the greatest value at 18.3% and Fee to Breed1 values ranged between 5.1% to 9.9% and the rest were less than 5%.

**Figure 6: 1st Iteration Model Accuracy Scores.**



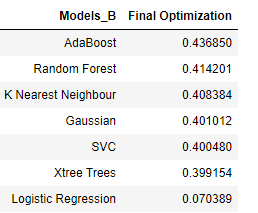
**Figure 7: 1st Iteration Model Variable Importance Values.**



Nine attributes were used for the second optimization step. These attributes were selected on variable importance, Logworth values and R squared means. These variables are **Age\_bin, Breed1, Color2, MaturitySize, State\_t, FurLength, Fee\_t, Quantity\_t and Type**. The attributes with **\_t** at the end of the name were transformed and this is explained in the Data Wrangling Section of the Report. The f1 scores ranged from 12.4% for Logistic Regression up to 37.2% for the Random Forest Model. Breed1 had the highest variable importance value of 23.1% and Age\_bin was 22.2%. The other values were similar to the previous run (Figure 8).

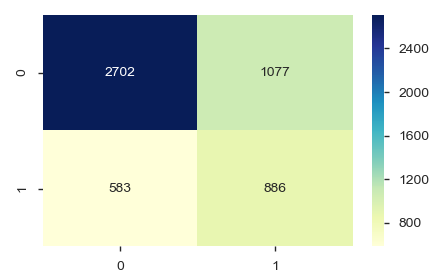
The final optimization step removed the **Type** attribute to improve model performance and contained only eight input variables. All seven models included hyperparameters to increase model performance, but used only 35% of a randomly sampled data set. These hyperparameters were; gamma (SVM), C and penalty (LOG), n restart optimization (Gaussian), n estimators and learning rates (Ada Boost), nearest neighbors’ weights and leaf size (KNN), max depth and features, minimum samples and leaf size (RF and XTREE). The results from the final optimization step is in Figure 8. The AdaBoost and Random Forest models had the best results with f1 scores of 43.7% and 41.4%. These models did a good job in selecting the

**Figure 8: Final Model Optimization Results.**



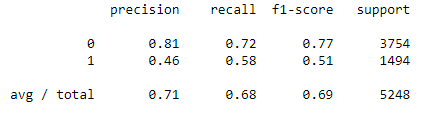
TN, but did not do a good job in selecting the TP values because of the unbalanced data. The Random Forest Model was selected to be the champion model and run again for the final optimization step because it had slightly more TP and TN values from the 2nd optimization run. The final Random Forest Model used the balanced approach with the Gini Index and had all 14,993 rows of data to generate the f1 score. The model’s f1 score increased to 52% compared to the 41% from the previous run. The Confusion Matrix, Classification Report and Variable Importance results from this run is in Figures 9, 10 and 11. The model did a better job in predicting the TN for animals that stayed in the shelter for less than 100 days compared to the TP or animals that stayed in the shelter longer than 100 days and did not leave the shelter. The top three attributes for the variable importance are **Age\_bin, Breed1** and **Color2** for this final model. The variable importance values ranged between 15.5 to 30.7% for these three variables. The transformed Breed1\_t attribute did not perform as well as the original Breed1 variable and was not included in this analysis.

**Figure 9: Champion Model Confusion Matrix.**

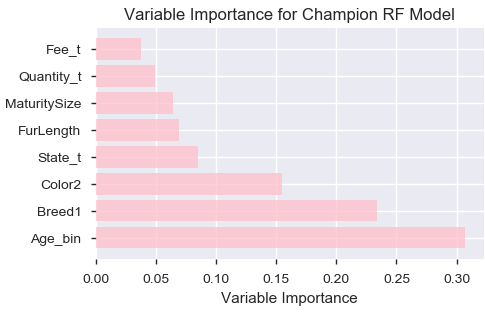


**Figure 10: Champion Model Classification Report and**

**Variable Importance Report.**



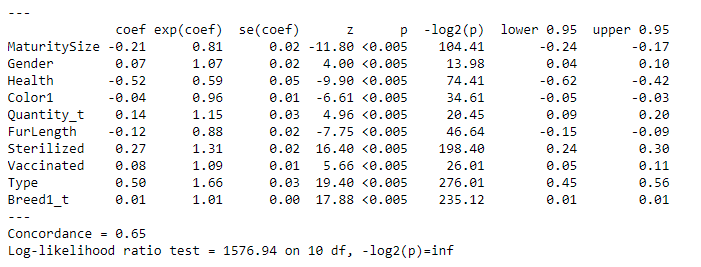
**Figure 11: Variable Importance values for Champion RF Model.**



## **3.3 Survival Analysis Model**

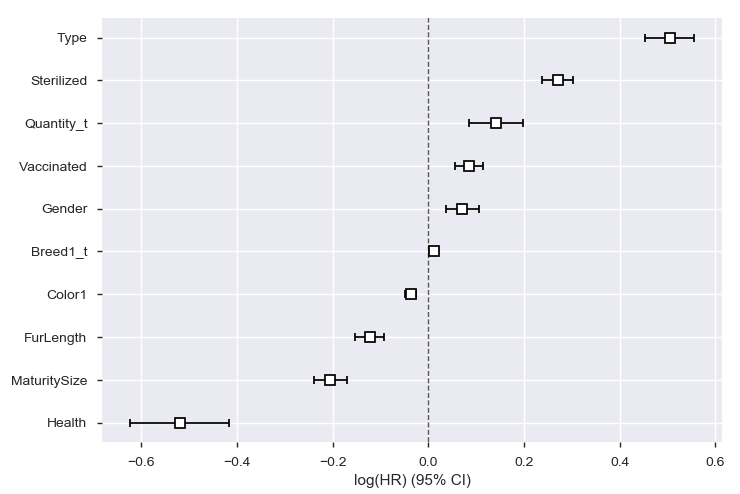
The Lifeline Survival Analysis Model in Python was used to determine which animals had the longest longevity in the shelters for the binary target variable (Figure 5A). The Cox Model was run in Python and used all 17 original attributes. The Age variable was used for the duration column and the binary target variable was used for the event column. The binary target had 0 representing the animals that were not adopted and 1 represented pet that were adopted or was the event. The input variables that had high p values on the Cox Summary report were removed and only nine variables were selected for the final model. These variables are **Gender, Color1, MaturitySize, FurLength, Vaccinated, Sterilized, Health, Quantity\_t, Breed1\_t and Type**. The Cox PH Fitter Summary report can be seen in Figure 12 and 13.

**Figure 12: Cox PH Fitter Summary Report.**



The summary report and Cox PH Fitter Plot shows a couple of attributes that had more of an impact for survival on this dataset. Health and MaturitySize had more of a negative influence compared to Quantity\_t, Sterilized and Type, which had more of a positive impact on Adoption rates. Breed1\_t had minimal influence on the adoption rates but was plotted for the interaction between 266 and the rest of the breeds. The concordance or “Goodness of Fit” value is only 65% and did a reasonable job for this model and this value ranges between 55 and 70% in other models. The Breed1\_t, Type and FurLength will be plotted using the Kaplan-Meier curves to go more in-depth for the survival analysis of these specific cohorts and the others will only be briefly discussed in this section. The Kaplan-Meier Plot (KMF Plot) used the Age as the Time and the AdoptionSpeed\_t as the event.

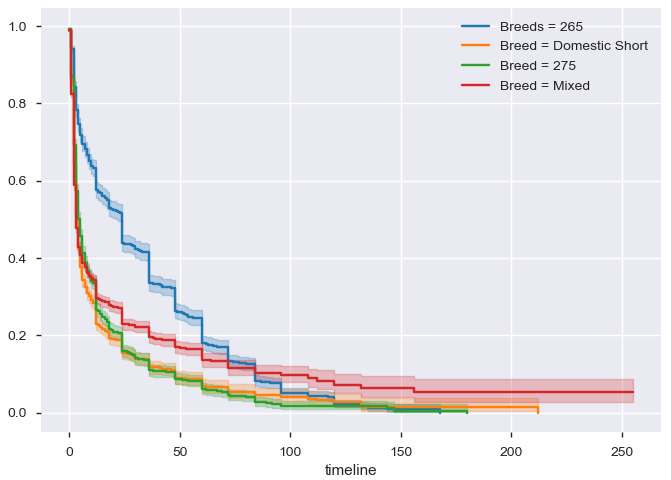
**Figure 13: Cox PH Fitter Plot.**



The Healthy individuals were much more likely to be adopted and were 97% of the data set compared to the minor and serious injured animals. The MaturitySize of 3 and 4 (large or extra-large animals) stayed in the shelters longer than the small to medium sized dogs for this dataset. The extra-large pet’s probability of being adopted dropped from 40% down to 20% around 40 months of age and then leveled off and was consistent with the large breeds. The KMF Plot for variable Type shows that cats had a higher adoption rate from Age 2 to 100 months for the data set. The Long Hair animals for FurLength was also least likely to become adopted from Age 2 until 90 months. This could be the result of the warmer climate in Malaysia or individual’s preference for specific breed types.

The Breed1\_t was had the least impact for Adoption but was plotted due to its influence on other models for the study. The interesting note from the graph is the Breed 265 which contains all dog breeds except 307 was the least likely to become adopted until age of 70 and then was more likely to become adopted compared to the 307 or mixed dog breeds. The 265 Breed had a total of 502 pets between 10 and 25 months old and 819 pets older than 25 months of age and this was 60% of the population for Breed 265 (Figure 9A). A total of 391 or 18% of the population did not get adopted compared to the 307 mixed breed which had 34% its population that was not adopted. The dogs that are in this category had higher age values and were more likely to become adopted.

**Figure 14: KMF Covariate Groups Plot on Type.**



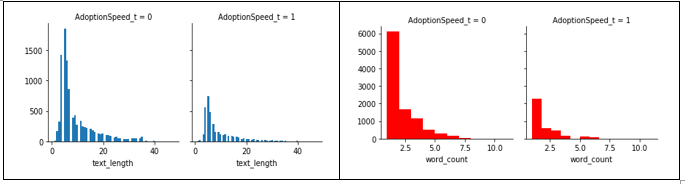
## **3.4 Text Analysis**

The text analysis was conducted for the name attribute of the training dataset. Webster’s definition of a “name” is a word or phrase that constitutes the distinctive designation of a person or thing. The goal is to determine if a word or distinctive phrase can be used to predict the adoption outcome of shelter pets. Text analysis was completed on the **Name** attribute to determine if specific names had a greater chance of staying in the shelter longer than other pets. The standard text analysis procedures were performed for the Pet Adoption Dataset. These preprocessing steps are; converting all words to lowercase and the removal of punctuation, whitespace, stop words and frequent words. The top 35 most frequent words and least 10 common words were filtered from the data set to remove the majority of names like “baby” or “lucky. The standard word frequency, TDIF, Polarity Scores were calculated and the Naïve Bayes Model was used with the binary target variable and the polarity scores, word length and word count for names.

### **3.4a Text Analysis Results**

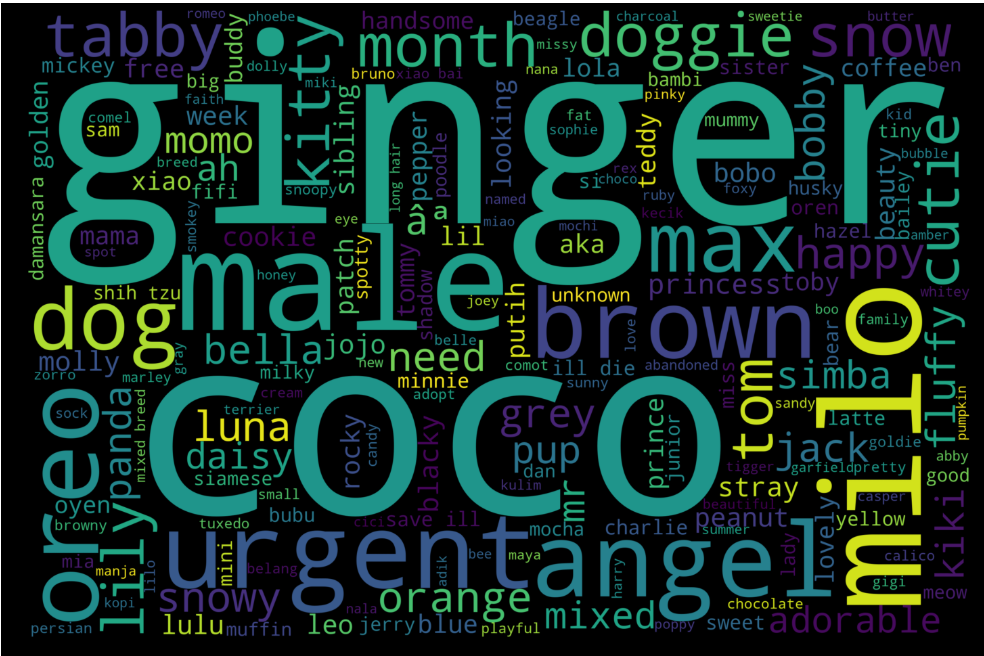
The text length and word count values were calculated and are similar in their distribution for the binned AdoptionSpeed Target variables (Figure 5A). The text length of five was most frequent with 1,842 Names for AdoptionSpeed of 0 to 743 Names for AdoptionSpeed of 1 (Figure 15). The word count of one word was also most frequent with 6,088 Names for AdoptionSpeed of 0 to 2,287 Names for AdoptionSpeed of 1 (Figure 15). The word cloud in Figure 16 shows the distribution of the most frequent names like “coco”, “oreo” or “lulu”.

**Figure 15: Text Length and Word Count for the Name Attribute.**



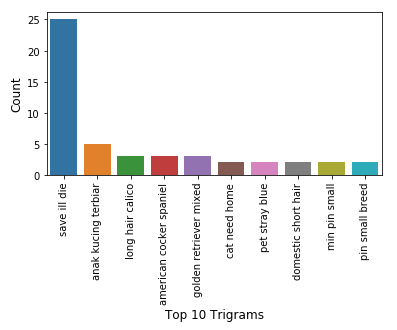
The majority of the names from the data set have a polarity score of 0 and this accounts for 94.1% of the total Names. There were only 532 names classified as a positive and 276 names as a negative sentiment. A random draw of five names for all three sentiments show the names are more of a verb describing the type or plight of the animal for these shelter pets. The most negative words describe the potential outcome of the animal who will not become adopted and they just added a number to the same phrase to separate the entries for this database. The top ten trigrams show the results and the phrase “save II die” which was used many times by adding a number at the end of this phrase to separate the animals in the shelters (Figure 17).

**Figure 16: Word Cloud for Name Attribute.**

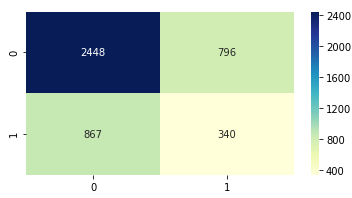


The Naïve Bayes model was used on the preprocessed Name text to give a baseline model result to determine if Names are good predictors for Adoption rate success. The baseline model had an accuracy score of 63% and a f1 score of 29% and the confusion matrix can be seen in Figure 18. The model does a good job in selecting the TN but struggles with the TP and is a consistent theme with this dataset.

**Figure 17: Trigrams for the Name Attribute.**

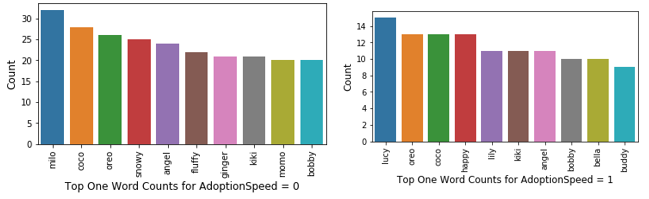


**Figure 18: Confusion Matrix for NB Base Model.**



The model optimization step looked at the number of letters or word counts a Name had to improve model performance. All Names that had 3 letters or less was binned as zero, three letters words were binned as a three and this continued until eight. All words that had eight or more letters were binned into the eight level. The words were also filtered by word count and names that only had one word for the name was tested with Naive Bayes model. The accuracy values were in the 60% range and the f1 scores ranged between 26 to 33%. The filtering by word count did a great job in selecting specific names but the accuracy score was only 64% and the f1 score was 26.5%. The top 10 words for the 1-word dataset shows the distribution of the words for the AdoptionSpeed of 0 and 1. The most common names like “puppy”, “baby”, “cat” and “girl” were filtered, but then another set of common words appeared in both AdoptionSpeed such as “coco” and become the next set of common words for both Adoption Speeds. The naming scheme for the pets seemed random and had no specific set of rules being followed from the adoption agency and made it difficult to determine if Name was a good predictor of AdoptionSpeed.

**Figure 19: Top word names for AdoptionSpeed.**

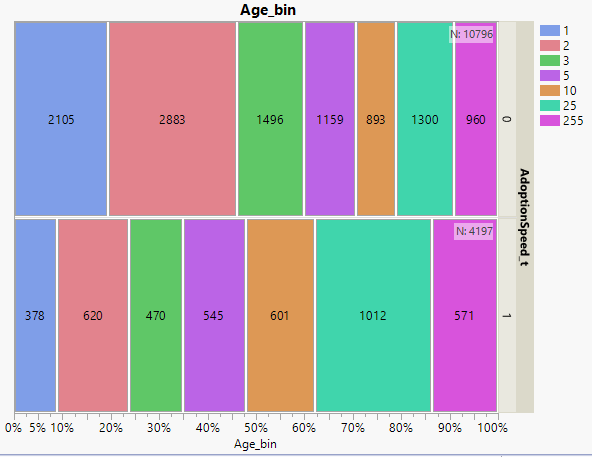


# **Discussion**

The Random Forest Model using the balanced approach containing eight attributes is the model that will be used to predict which animals are least likely to be adopted for this study. There are four recommendations that will be implemented from the findings of this study. The goal is to increase adoption rates for the pets that have been predicted to not be adopted and reduce the cost for these animals that have been in the shelter for 100 days.

The Age attribute has been the most important variable for all the models in this study. One interesting statistic in Figure 20 shows that 24% of the pets that were not adopted from the

**Figure 20: Tree Map for Age\_bin and AdoptionSpeed\_t.**

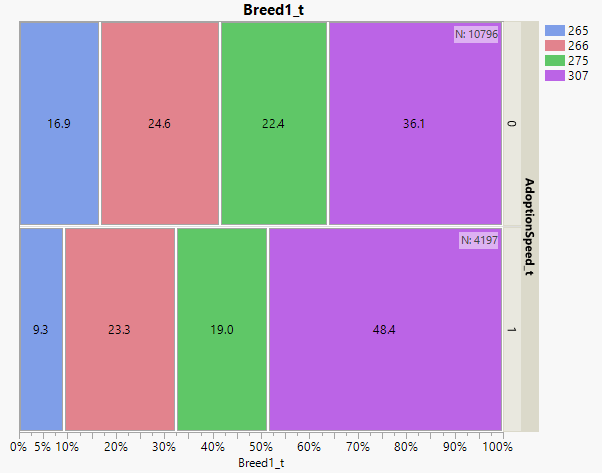


Age\_**bin** category came from animals that were older than 10 months and less than 25 months of age. This age group will be targeted to increase adoption rates from the select

shelters by implementing this new program called “Rescue Rebuild” which has shown to increase adoption rates by 25% (Huffpost, 2013). The goal is to restructure select shelters to provide animals with more open areas and reduce stress that is placed on the animals from being confined in tight kennels. This concept allows the animals personality to flourish as they play in the open area compared to the confined kennel. These animals are between 10 and 25 months of age and would be a good target group to test the Rescue Rebuild Program for the pets in Malaysia. The Rescue Rebuild program not only reduces stress, but also gives time for the potential adoptees to find the pet that they can relate too and help with the adoption process. Two groups will be randomly selected with half being the control and no access to outdoor space and the remaining half being the treatment group for the Rescue Rebuild Program. Results will be analyzed to determine if this can significantly reduce shelter time for the treatment compared to the control group pets in the Malaysia region and increase adoption rates by 25%.

The Breed1\_t attribute was also another important attribute for variable importance. The second recommendation is to focus on a select group of mixed breeds for these shelter pets not being adopted. The training data set contained a total of 176 unique breeds for Breed1 and Breed2 had a total of 135 unique counts. Breed2 was listed for animals that were not a pure breed and is listed as the secondary breed. The training dataset contained 72% Mixed Breeds compared to only 28% Pure Breeds as shown in the count plot (Figure 6A). These numbers are consistent with other agencies who state that mixed breeds make up more than 75% of the animals entering the shelters. The tree map in Figure 21 shows the distribution of all four

**Figure 21: Tree Map for Breed\_t and AdoptionSpeed\_t.**

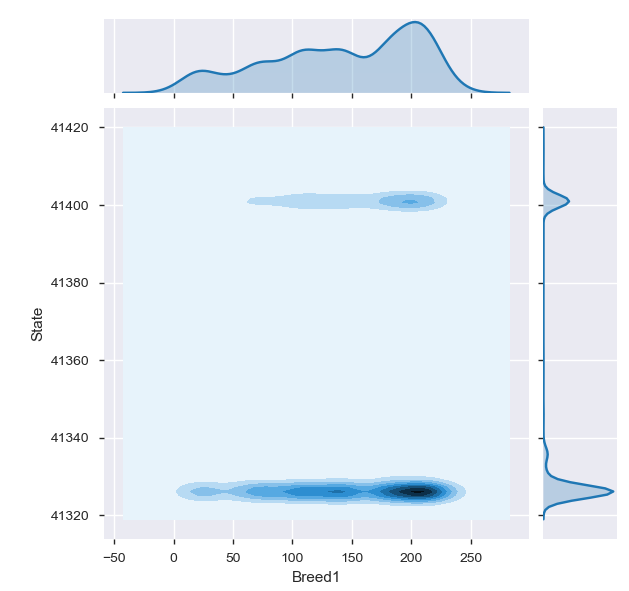


breed1\_t categories. Breed1\_t categories 265, 266 and 275 slightly decreased in the percent of the total population of animals being adopted to not getting adopted except for Breed 307. This mixed breed dog category actually increased from 36% that were adopted to 48% of the animals that were not being adopted in this study. The goal is to implement two programs

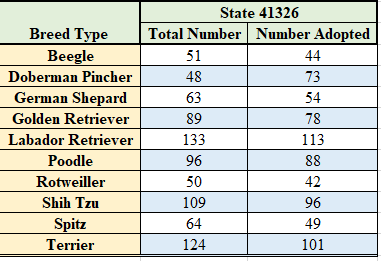
that were initiated from the Stony Brook University and is called the “Blog method” and Costa Rican Adoption organization in creating unique breed names. The Costa Rican Adoption agency had an increase in pet adoptions for uniquely created names like “Fire-Tailed Border Cocker” compared to the control group. The Huffington post actually said this increased adoption rates by 1,400% and these results need to be replicated in this study. The goal is to combine the two programs into one and create a blog for the mixed dog breeds in the shelter with a unique breed name specific to their genetic makeup and create an electronic portfolio. This study will randomly divide the mixed dog breeds into three categories which contain two treatments and the control group. Treatment 1 will create a blog for the animal with a unique name specific to their genetic makeup and with a common one-word name. Treatment 2 will contain a blog with only a common one-word pet name that is unique to that shelter and will be labeled as mixed breed. The control group will not have a blog, will name the animals as mixed breeds and will use the current naming scheme. The text analysis from this study shows there was no naming schemes being followed for these shelters. Names were created that had one words with over 15 letters which had no meaning and names with over eight words used as a slogan. The slogan was used to describe the pet’s fate or description of the event which should have been placed in the description attribute. The words were recycled and reused which made the Naive Bayes model a poor predictor in which names could be the best predictors for success. This study will test to see how effective the Blog and Naming scheme actually work with this animal population and measure its effectiveness in increasing Adoption rates.

The Breed1\_t binned attribute number 265 had very interesting results for the dogs age and adoption rates. The dogs that were older than 10 months of age contained 1,302 out of the 2218 animals and 60% of these were adopted from the shelter and this can be seen in the Kaplan Meier-Plots. The Breed1\_t column was filtered to contain all dogs except 307 and the KDE plot was created in python for these dog breeds and the State (Figure 22). The density map shows a concentration of pets around Breeds 200 and between State values 41320 and 41330. The Pandas crosstab was used to determine which breeds were the most adopted and from which state in the data set (Figure 23). This figure shows the top 10 dog breeds in this data set which came from State 41326 or Selangor, Malaysia. Selangor had a total of 1,359 of the 2218 dogs from the Breed1\_t and about 60% of these older animals were adopted out of the shelter. These are common house breeds that were adopted more frequently compared to the mixed breeds for this age. This is another example of why the Rescue Rebuild or Unique Name Identifier Programs for mixed breeds needs to be implemented to help reduce these unwanted shelter pets.

**Figure 22: KDE Density Plot for Breed1\_t 265 and State.**



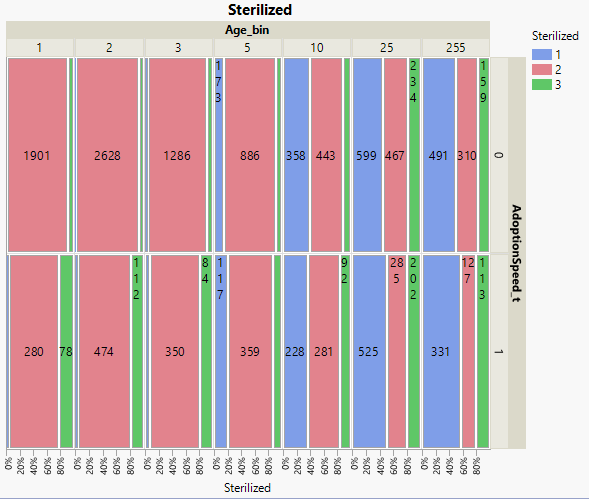
**Figure 23: Top Ten Dog Breeds for Selangor.**



This study was unique because most pets were not sterilized until 10 months of age and was a significant variable in the Tukey comparison of means but was a weak predictor for the Random Forest Model. A few animals were sterilized at five months of age and a total of 2,532 animals were sterilized after 10 months of age (Figure 24). The American Animal Hospital

**Figure 24: Tree Map comparing counts for Age\_bin, Sterilized and Adoptions**

**Rates (Sterilized values are; 1 = Sterilized, 2= Not Sterilized and 3 is Unknown).**



Association (AAHA) recommends sterilizing cats and small dogs around five to six months and medium to large dogs should be sterilized between 9 and 15 months of age. The recommendation is to sterilize all shelter pets when age appropriate to reduce the number of unwanted pregnancies and fight the battle against the pet overpopulation problem. Sterilization can reduce cancer rates, reduce unwanted behaviors such as aggression, and eliminate the heat cycle in females. There was a total of 4,527 small to medium sized animals and only 60% of this population actually was sterilized. Sterilization is a common practice that should be implemented in the shelters to help reduce unwanted pets.

# **Conclusion**

The best model to predict Pet Adoption rates with the f1 score is the Random Forest Model that uses the balanced approach and the Gini Index. The Accuracy results were 72% and the f1 Score was 52 % for the data set and the model did a great job in selecting the TN and an average job in selecting the TP from the dataset. The results from the study show 24% of the pets that were not adopted came from the Age group ranging between 10 to 25 months and would be a good target group to initiate the Rescue Rebuild Program that allows animals personality to attract potential adoptees.

The mixed dog breeds from Breed 307 composed of 48.4 % of the Pets that were not adopted from this study and is another group that could benefit from a hybrid program that created a blog for each pet and generated a uniquely created name based on their genetic make up to attract potential donors. The mixed breeds can make up 75% of the shelters population and are more difficult to adopt because people are looking for specific breeds as shown in the binned 266 Breed1\_t category. This group had a large percentage of dogs that were older than 10 months of age, but were adopted more than the mixed breed category as shown in the Kaplan-Meier plots around 70 months. This group contained the standard breeds like Beegle, Golden Retriever and Poodle.

The Text Analysis for the Name with the Naïve Bayes Model had accuracy values in the 60% range and low f1 scores ranging from 26 to 33%. The model did not do a good job in predicting Adoption Rates based on Name and the naming scheme from the Pet Adoption Agencies had no specific naming protocol. The blog and naming scheme program will be used to determine specific names can increase pet adoption rates.

# **References**

Albon, C. (2018). Machine Learning with Python Cookbook, Practical solutions from

preprocessing to deep learning. O’Reilly Media Inc in Boston, First Addition.

Animal Farm Foundation. (ND). Naming Shelter Dogs: The Framing Effect. Retrieved from:

<https://animalfarmfoundation.blog/2013/04/08/dog-names-framing/>

ASPCA, (ND). ASPCA Research: Do Cat Names Affect Adoptability

Retrieved from: <https://www.aspcapro.org/resource/saving-lives-adoption-marketing-research-data/aspca-research-do-cat-names-affect>

Davidson-Pilon, C. (ND). Lifelines. Retrieved from: <https://lifelines.readthedocs.io/en/latest/>

Crowe, E. (ND). Increasing Adoption Rates at Animal Shelters. Stony Brook University.

Retrieved fromhttps://xsrv.mm.cs.sunysb.edu/c4g/showcase/dogs%20with%20blogs.pptx

Foothills Animal Shelter. (ND). Services; Lost and Found Fees. Retrieved from:

<https://foothillsanimalshelter.org/services/lost-and-found/fees/>

Hastie, T., Tibshirani, R. and Friedman, J. (2017). The Elements of Statistical Learning: Data

Mining, Inference and Prediction. Springer Series in Statistics, 2nd Edition.

Huffpost, (2013). Mixed Breed Dog Adoption Campaign Gives Lovable Pooches New, Unique

Breed Names. Environment Section.

Retrieved from: <https://www.huffpost.com/entry/mixed-breed-dog-adoption_n_3869267?guccounter=1&guce_referrer=aHR0cHM6Ly93d3cuZ29vZ2xlLmNvbS8&guce_referrer_sig=AQAAAKEJYS7Dqs7CWUbiEhU7u7WUZn-CoFmqQMeiTJAsyyexkHCsvAsgpYtHbf3S_wepcwnsUAkvlcuxeZZTCqG6N5ntnL7h9Jguz7c6BGct5eqiYIdKuUYIBaIcVSE4buNl15mufdCKRcZNFtWoFJmN_GVRLTRBDPkdwIpi132CQm9s>

Kelly, J. (ND). Pet Adoption Fees Explained. PETMD Care. Retrieved from:

<https://www.petmd.com/dog/care/evr_dg_dog_adoption_fees_explained>

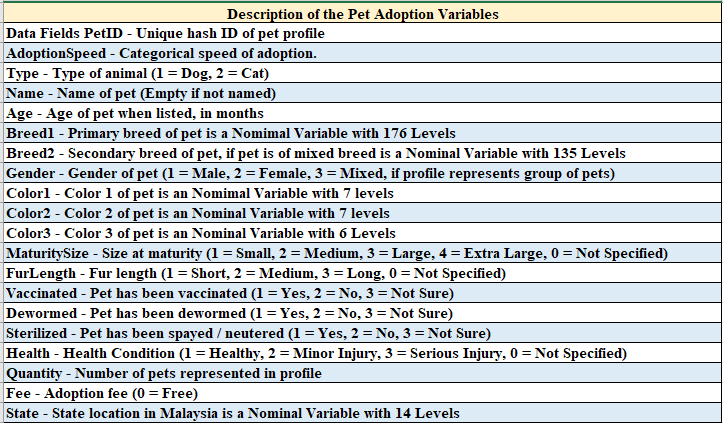
Merrian-Webster. (ND). Retrieved from: <https://www.merriam-webster.com/dictionary/name>

Purina. (2013). Nestle Purina completes acquisition of Petfinder. Retrieved from:

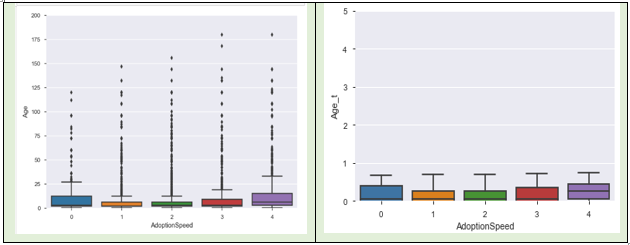
<http://newscenter.purina.com/2013-07-15-Nestle-Purina-completes-acquisition-of-Petfinder>

# **Appendix**

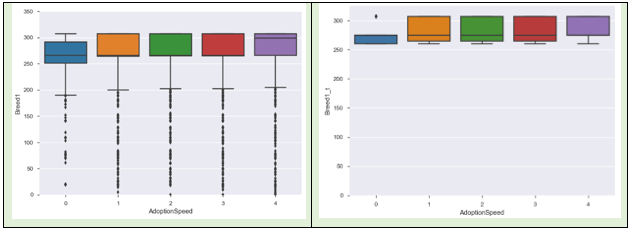
**Figure 1A: Pet Adoption Variable Description.**



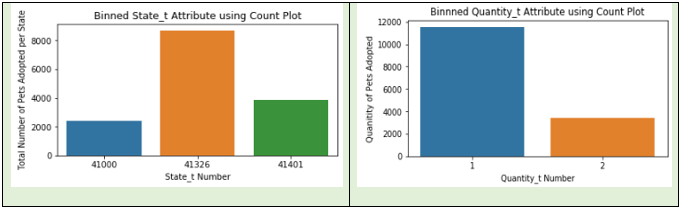
**Figure 2A: Boxplots for the Age and Log10 Transformed Age\_t variable.**



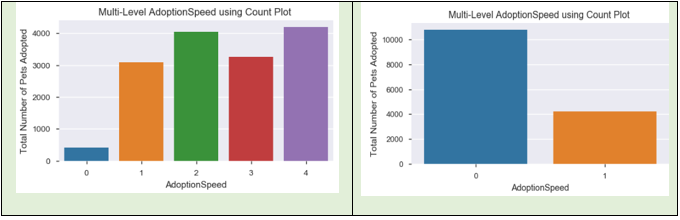
**Figure 3A: Boxplots for the Breed1 and the binned Breed1\_t variable.**



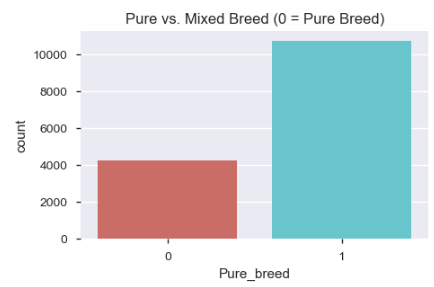
**Figure 4A: Binned Results for the State\_t and Quantity\_t Variables.**



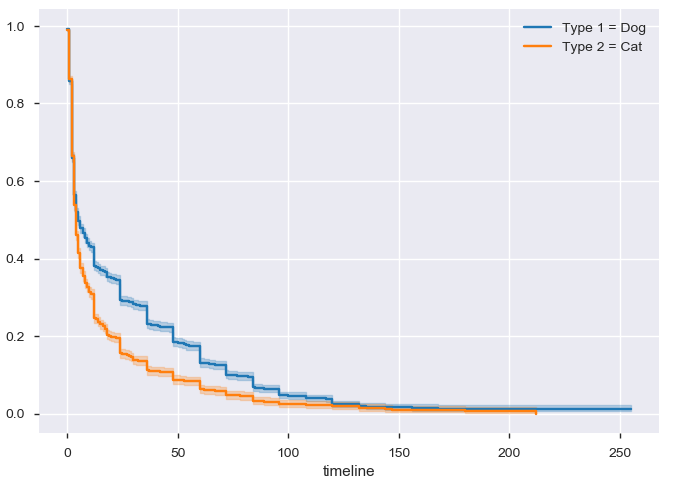
**Figure 5A: Plots Comparing the Original to binned Target Variable AdoptionSpeed.**



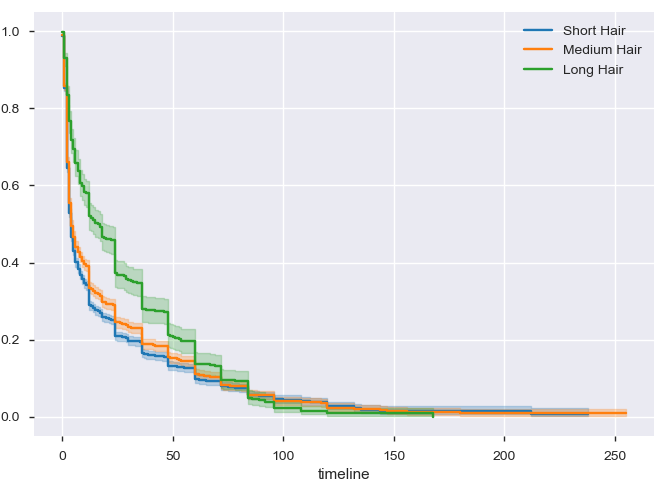
**Figure 6A: Pure verses Mixed Breed Count Plot.**



**Figure 7A: Kaplan-Meier Plot for Pet Type.**



**Figure 8A: Kaplan-Meier Plot for FurLength.**



**Figure 9A: Tree Map for the Breed1\_t and Age\_bin Category.**

