Capstone 1 Milestone Report

TITLE: A Home for All
AUTHOR: Caitlin Jansson
DATASET : https://www.kaggle.com/c/petfinder-adoption-prediction
GIT REPOSITORY: https://github.com/CJEJansson/Springboard_Projects

PROBLEM STATEMENT

There are more potential pets than pet owners, and shelters across the country are struggling to keep up. But what if there were a way to increase the likelihood that an animal gets adopted? Does the name matter? The color? Is one breed more likely to be adopted than another? Do any of these things influence the speed at which a pet is adopted?

Local area shelters are struggling. The problem is, as always, space is limited and the number of animals that need help just keeps growing! A local shelter specializing in large breed dogs has asked that some research be done. Are better pictures needed? Do they need to "rebrand" the animals by listing a different primary breed? Or is a simple name change going to help? Which animals need to go to all the adoption events and which will be adopted simply because they're listed? If they better understand the systemic trends based on the dog's identifying features, they can configure their PetFinder entries to optimize pet adoptions. The intent of this project is to investigate and find out.

GENERAL OVERVIEW OF THE DATASET:

The data is provided from the Kaggle.com PetFinder competition, targeted at increasing adoption speeds for homeless pets. It's divided into a test and train dataset and contains information on over 10,000 animals, including cats and dogs. The data is pulled from the Malaysian region, and is categorized by Malaysian state. Feature data has been converted to ordinal format (typically 0-4), and descriptive data that corresponds is separated into an independent table for each feature.

Overall the data was relatively clean, with the exception of the "Name" and "Description" columns, which will be addressed later. The first step was to remove all data on cats from the dataset, as the customer is only concerned with the dogs at this time. It was verified that there are no animals listed as being a cat and having a corresponding breed classification indicating that it is a dog.

Any values that are null in breed occur only in Breed 2 or Breed 3 column, indicating that a dog is listed as being a purebred dog. Null values were filled with zeros to balance out the data set for later use, and a row for 0 was entered into the breed table with the value "No Breed". An additional column was created to account for mixed vs. purebred listings called "BreedCount", and counts the number of breeds listed for each animal. Mixed breeds have a "BreedCount" value greater than one, as expected. The same process was applied to the animal's colors, and the creation of a column called "ColorCount."

Two rows in the Description column were found to have null values, these were originally left. However, further investigation showed that there is a lot of unclean data in the description column. As the final intent of this project will not include Natural Language Processing, the null description values were filled with "No Description". The intent going forward will be to leave the description column largely unaddressed. The logic behind this decision is that the customer has requested focused efforts on identifying which dogs will be quickly adopted via an online listing vs. which dogs will need additional efforts for adoption including adoption events and fostering.

After investigation of the "Name" column, in the attempt to address Null values, it was determined that just because an animal had a value in the name column, did not mean that the animal actually had a name. Due to the size and extent of the problem, name's were manually cleaned using excel. Some of the following are examples of name entries that were found and were replaced with a null include, but are not limited to: “Lost Dog”, “Cute Puppies”, “Boy”, “Girl”, Miscellaneous breed names (labrador, terrier, etc.), “No Name Yet”, “Please Name Me”, “Save me or I'll Die”, “Urgent home needed”, “Puppy”, and various descriptions of puppies (happy puppy, big eyes, sad puppy, bouncy puppy, etc.)

Animals that are actually unnamed were left as null values. This allowed for the creation of an additional column “Name_bool”, used to indicate whether an animal had a name or not. The value 0 was filled for no name, and 1 indicated that the animal had a name.

INITIAL FINDINGS:

The data was then explored via Exploratory Data Analysis (EDA) and analyzed using inferential statistics. Knowing that the majority of the dataset is categorical (discrete) data, Chi-squared testing was heavily used to explore the relationships between features. Additionally, Cramer's V technique was used to generate a heat map to show correlation between variables[1]. A complete summary of statistical results can be found in Appendix A.

It is known from animal rescue work in the United States that dogs with the least likely chance of adoption are senior dogs, dogs with health problems, and dogs that have black fur. Also, the breeds categorized as “aggressive,” pitbull breeds, german shepherds, dobermans, etc. have a lower chance of adoption. Overall, approximately 50% of American animals (approx. 1.2 million) go unadopted, annually. It is suspected that this is also the case in dog rescue globally.

When looking at the Kaggle dataset it quickly becomes apparent that a large number of dogs are also going unadopted, but not to the severity in the U.S. Approximately 30% of dogs go either unadopted or unreported as being adopted (Fig. 1, page 3). As expected, a very low number of dogs are adopted on the same day they are listed, at 2.08% Of these animals, the younger dogs are adopted more quickly than the senior, which is consistent with expectations. Statistical testing confirms that there is a relationship between age and life speed at a significance level of 95%.

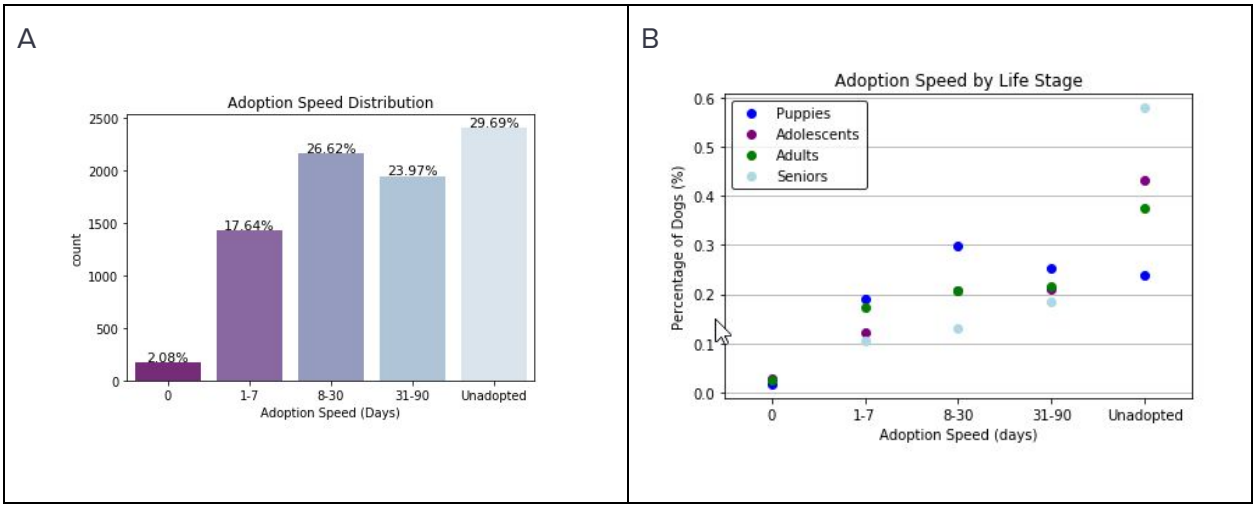
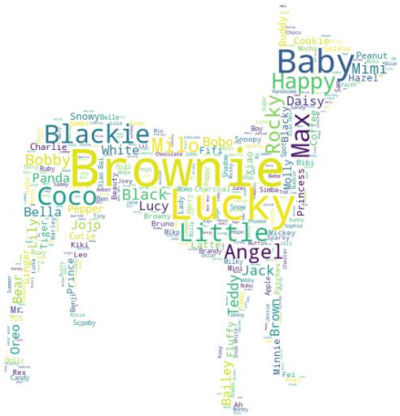


FIGURE 1: a) Distribution of adoption speeds for dogs across data set. b) Scatterplot showing adoption speed distributions by life stage

When it comes to names and appearance, dogs that have already been given a name tend to be adopted more quickly in the U.S. It's possible that the presence of a name indicates the dog's history is known, and it gives a sense of personality. First, the names listed in the dataset were explored to see what were the most popular names, and if they had any bearing on adoption speed. Most often, the color of the animal was its name, with the most popular names in this dataset being "Blackie", "Brownie", "Baby" and "Lucky". Statistical testing showed that the specific name of an animal is independent of adoption speed. However, the presence of a name did matter, and tested as being statistically related to adoption speed (Fig. 3)



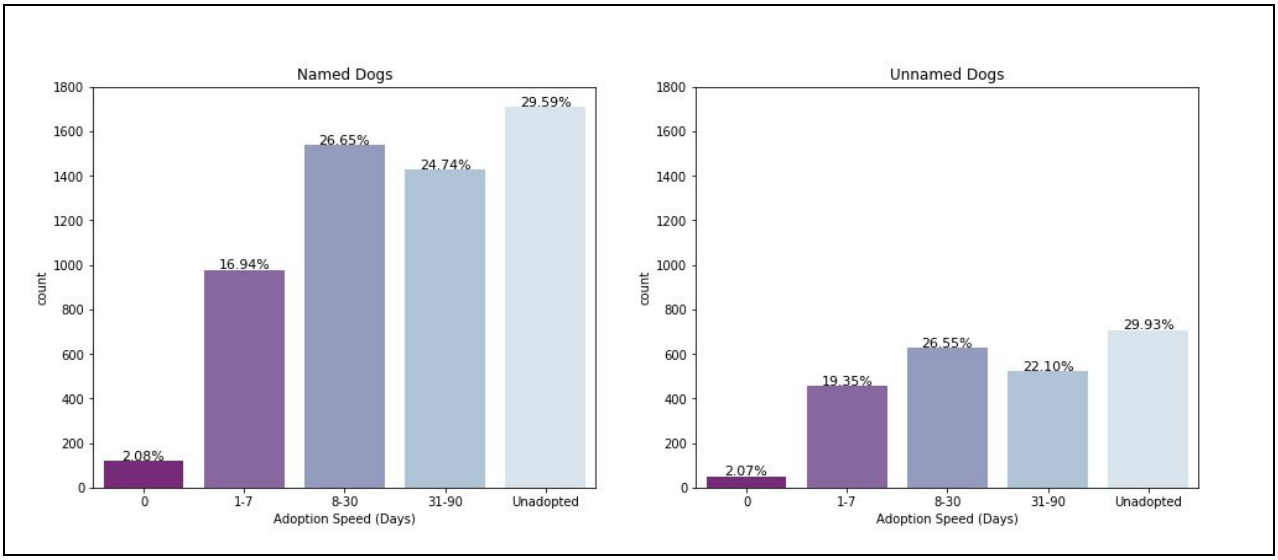
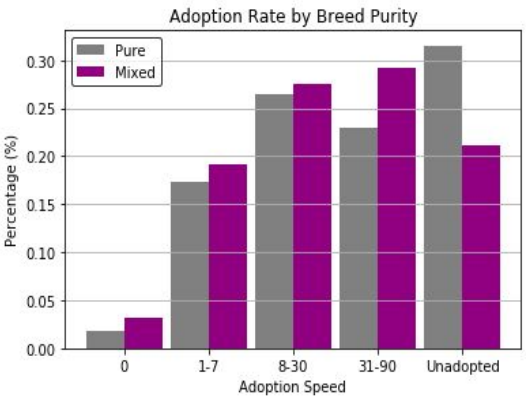


FIGURE 3: Adoption distribution of named (left) vs unnamed (right) dogs

When it comes to breed choices in the U.S., purebred dogs tend to be more popular. Particularly when it comes to labradors and other hunting dogs. The data set seems to demonstrate that in Malaysia this is not the case, with mixed breed dogs being more popular (image to right). This may be due to the availability of mixed breed dogs being higher. However, Labrador's are just as popular as they are here in the U.S. Shih Tzu's and Terriers are the second and third most popular dogs. (Fig. 4, pg.4) Statistical tests confirm that there is a relationship between breed and adoption speed, and between mixed/purebred and adoption speed. However, the actual breed has more of an impact that whether a dog is mixed.



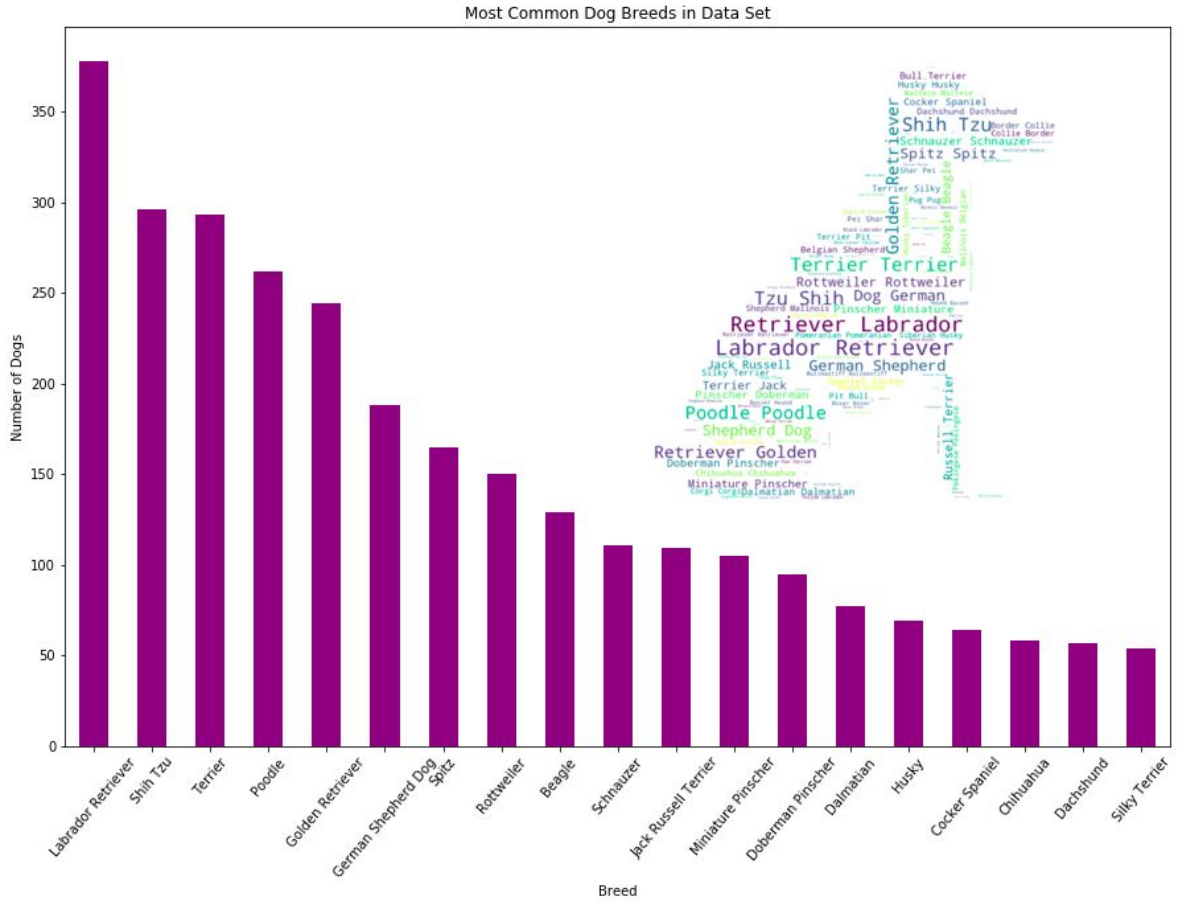


FIGURE 4: Distribution of dog breeds in data set and corresponding word map

The size of the dog also plays a role here in the U.S., with medium size breeds being more popular. In Malaysia, giant breeds are very quickly adopted, likely due to their scarcity, as they make up 0.22% of the data set. Otherwise, size has very little effect on adoption speed in Malaysia. Statistical tests confirm this relationship.

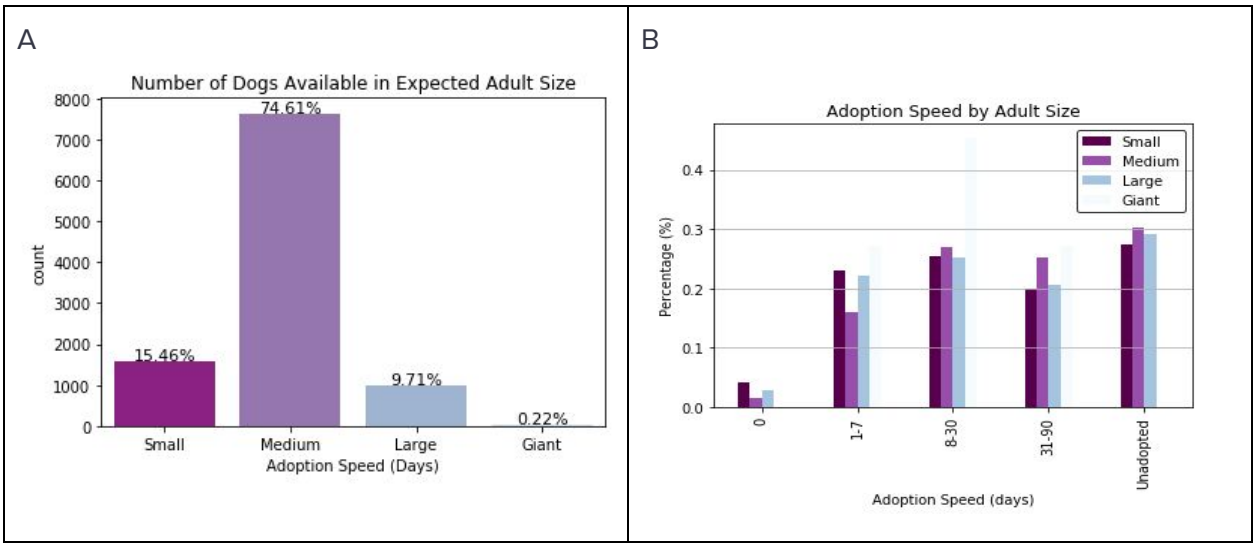


FIGURE 5: a) Distribution of dogs by adult size, b) distributions of adoption speed by adult size

When it comes to appearance, in the U.S., there's definitely an impact. Fur length is more of a factor in certain regions, but overall doesn't make much of a difference in adoption speeds. In the dataset, it appears that dogs trend towards shorter coats. While there is some difference between the distributions (Fig. 6), statistical testing shows that between fur length and adoption speed are not highly correlated but there is a relationship between the features.

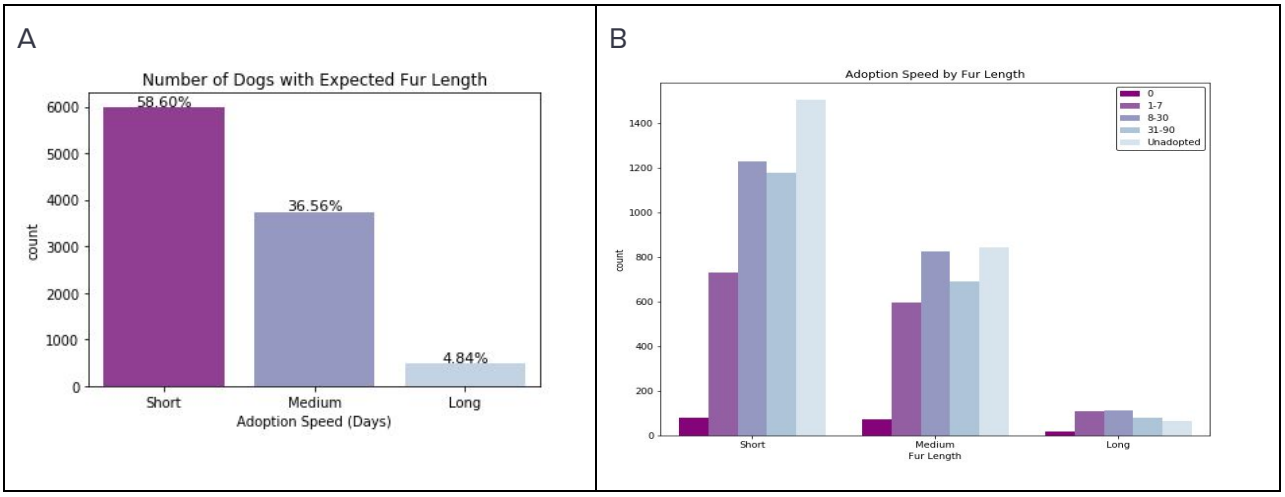


FIGURE 6: a) Distribution of dogs by fur length, b) distribution of adoption speeds by fur length

While fur length doesn't make much of a difference in the dog's adoption speed in the US, the color does. There's a well known phenomenon in the U.S. adoption groups that indicates black, or darker colored dogs go largely unadopted. This is thought to be because they appear intimidating, and they also do not photograph well. In Malaysia fur length also has very little effect on adoption speed. However, as expected, darker color dogs go unadopted more often. This may in part be due to the large number of dark colored dogs that are available. Additional investigation would be needed for conclusive evidence, but it appears the trend holds (Fig. 7). One interesting observation of note is that golden colored dogs and yellow colored dogs are separated into two groups. At first glance it appears that yellow dogs are also very unlikely to be adopted, but when combined with golden dogs this trend disappears.

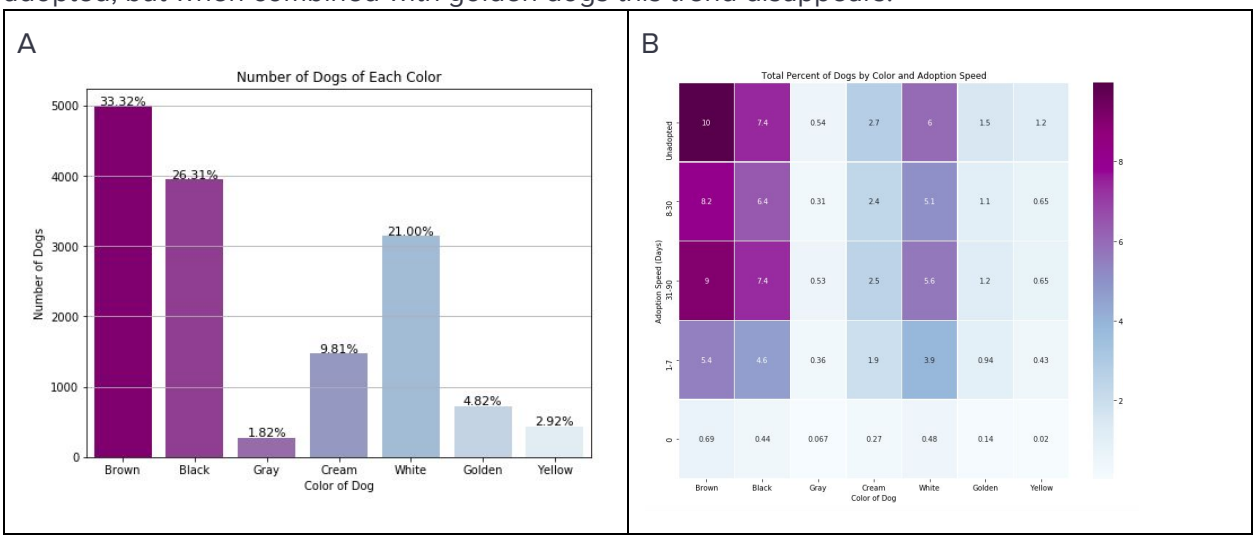


FIGURE 7: a) Distribution of dogs by color, b) heat map of percentage of dogs in each adoption speed category

Also investigated was the number of colors in a dogs coat. There were no visible trends during EDA, but there is a slight correlation between this and adoption speed.

As with dogs in the U.S., the way the listing profile is set up has an effect on the speed with which an animal will be adopted. The presence of photos and as much information as possible does make a noticeable difference here. This trend holds true in Malaysia. Also of note, the presence of a video makes little difference in adoption speed (Fig. 8).

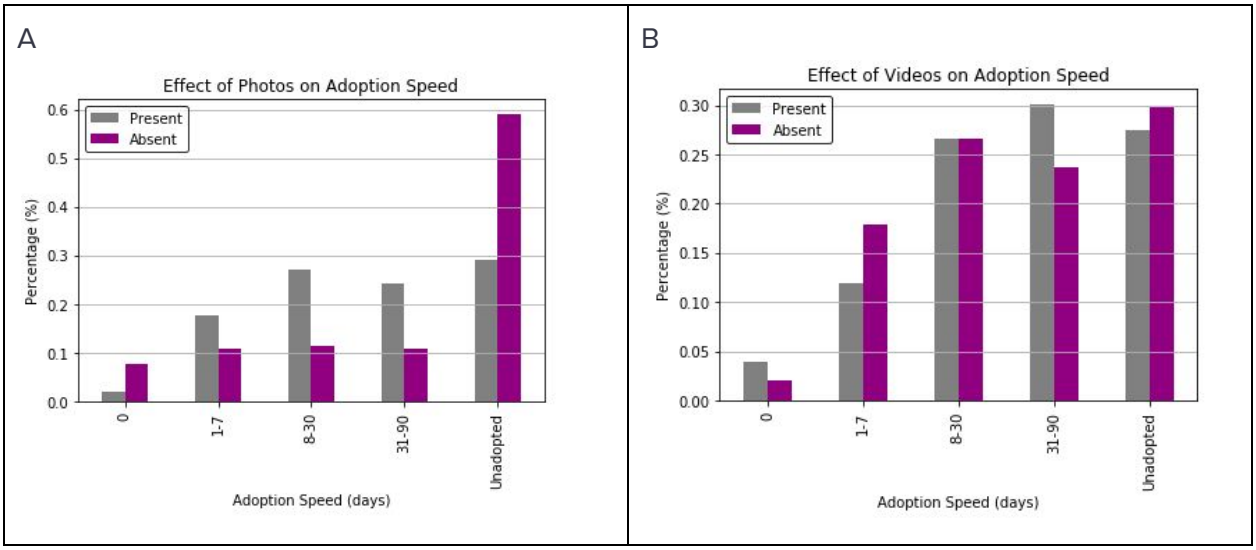
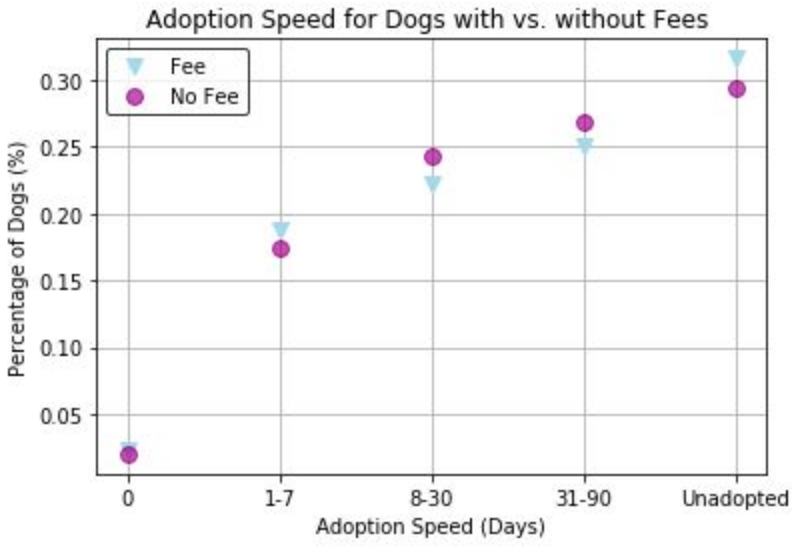


FIGURE 8: a) Distribution of adoption speed with and without photos, b) distribution of adoption speeds with and without videos

The cost of a rescue is often a contentious subject when adopting a dog. If an animal is from a particular breeder and is sold as a puppy, it's often no issue. The perception in the US is that a rescue is not as "high class", and should thus be cheaper. It was expected that the more expensive a dog is listed as, the less likely they are to be adopted. However, the opposite was found to be true when examining the data (see figure on right). This is potentially due to the fact that a specific type of dog or a purebred dog can be listed at a higher price and many people consider them to be more desirable. While cost is not highly correlated with adoption speed, statistical testing does show it had some impact.



Investigation showed that more dogs per listing the less likely you are to see an adoption. It's possible that listings with larger numbers of dogs that do not require co-adoption (for litters vs adopting siblings together) are obscured because there is no way to record adoption speed for

each animal, and the longest adoption time is all that's captured. Statistical testing was not performed on this feature due to the inability to perform correlation on a combination of continuous and discrete data.

In the U.S. the health of an animal plays a large role in adoption speed. To the point that almost all shelters and rescues guarantee the dog will be sterilized (spayed/neutered), up to date on vaccinations, and dewormed. Healthy dogs also tend to be adopted faster as opposed to those classified as "special needs", or those dogs that have a known injury or illness. In the dataset it became apparent that dogs that were unsterilized, not dewormed, and unvaccinated were fastest adopted (Fig. 9). More in-depth investigation showed that this is because these dogs were typically puppies, and puppies are adopted fastest. After removing puppies from the data set the trends more closely mirrored those in the U.S. As expected, dogs that are unhealthy or have a known injury/illness are less likely to be adopted. These relationships were confirmed during statistical testing.

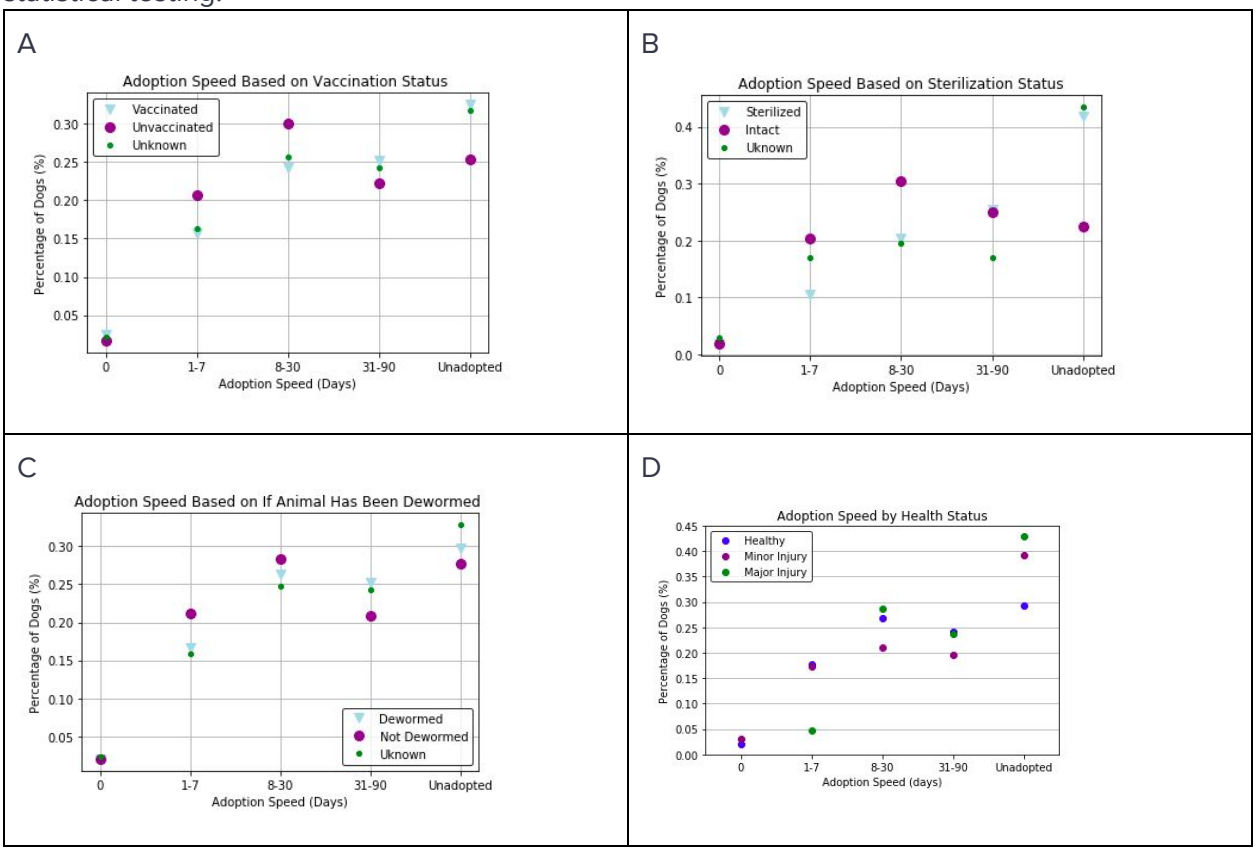


FIGURE 9: Adoption speed as compared to a) vaccination status, b) sterilization status, c) worm/unwormed status, and d) overall health (pre-existing conditions).

Overall the data showed the results expected with a few exceptions. Malaysian dog trends are slightly different than those in the U.S. As expected those dogs with the worst outlook for adoption are dark colored, seniors, or dogs with health issues. A few unexpected gems were unearthed as part of this EDA, including the insignificance of videos, the impact of age on other variables, and the preference for an unsterilized dog. Hopefully these insights will lend themselves to the development of an accurate adoption prediction engine as the project continues.

CITATIONS

[1] https://towardsdatascience.com/the-search-for-categorical-correlation-a1cf7f1888c9

[2] https://www.aspca.org/animal-homelessness/shelter-intake-and-surrender/pet-statistics

Appendix A

STATISTICAL RESULTS OF FEATURE SPECIFIC INVESTIGATION:

*Cramer's V Value reported is vs adoption speed. Green indicates

FEATURE	CONCLUSION	χ^2 VALUE	P-VALUE	CRAMER'S V VALUE*
Age/Life Stage	Age and adoption speed are related	939.16	1.31 E -44	0.06
Name	The dog's name and adoption speed are independent.	13,096.11	0.98	-
Presence of Name	Whether the dog has a name affects adoption speed.	10.46	0.03	0.02
Breed	There is a relationship between breed and adoption speed.	2738.21	2.39 E -41	-0.06
Mixed vs Purebred	There is a relationship between mixed vs purebred dogs but it's not as significant as the relationship between specific breed and adoption speed.	75.32	1.70 E -15	0.17
Size	There is a relationship between an animal's adult size and adoption speed.	102.48	1.82 E -16	0.03
Fur Length	There is a relationship between fur length and adoption speed.	106.14	2.36 E -19	-0.11
Color	There is a relationship between color and adoption speed.	74.19	4.96 E -7	-0.03
# of Colors	There is a relationship between whether a dog is a single color and adoption speed, but it has less impact than color alone.	30.85	1.4 E -4	-0.01
# of Dogs in Listing	Not tested due to the nature of data	-	-	0.07
Photos	There is a relationship between photo amount but it is not a large correlation.	401.19	5.34 E -32	-0.00
Videos	Videos and adoption speed are independent.	33.80	0.38	-0.01
Cost	There is a relationship between adoption fee and adoption speed	327.04	1.5 E -4	-0.03
Vaccination	There is a relationship between vaccination status and adoption speed.	90.03	4.59 E -16	-0.04
Sterilization	There is a relationship between sterilization status and adoption speed.	431.30	3.75 E -88	-0.07
Wormed vs Dewormed	There is a relationship between whether a dog has been treated for worms or not and adoption status.	42.40	1.14 E -6	-0.01
Pre-existing Conditions	There is a relationship between an animal's health status (healthy vs pre-existing conditions) and adoption speed.	18.51	0.02	0.03

Cramer's V Heatmap showing correlation between features: