Capstone 1: Data Storytelling

Project Proposal: A Home for All
Data Wrangling: Capstone 1 Data Wrangling

LEARNING OBJECTIVES

1. Learn to ask questions about and explore data
2. Develop skills in identifying trends, correlations, and making hypotheses
3. Practice using text and plots to communicate and present insights.

DATASET : https://www.kaggle.com/c/petfinder-adoption-prediction

APPROACH

The intent of this project is to utilize Exploratory Data Analysis (EDA) to investigate the impact of each recorded characteristic on the overall adoption speed. Specifically, the goal is to explore the relationships between each variable and adoption speed to determine those features that can be modified to increase likelihood of adoption. The table found in Appendix A summarizes the features investigated, hypotheses prior to investigation, conclusions post investigation, and a thumbnail of the resulting plots. Additional information and exploration into the data can be found on Github via this link: Capstone_1_Data_Storytelling.

GENERAL OVERVIEW OF THE DATASET:

There are listings for 10,230 dogs in the provided dataset. These dogs are located in the country of Malaysia and are available for adoption via the PetFinder website. 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 (approximately 1.2 million) go unadopted, annually[1]. 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. 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.

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. When analyzing the Kaggle data there was no apparent relationship between name and adoption speed, with unnamed and named dogs being adopted at approximately the same speed. In the data set the most common names given to dogs are related to the color of the animal, with "Blackie" and "Brownie" being the most popular.

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 demonstrates that in Malaysia

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

this is not the case, with mixed breed dogs being more popular. This may be due to the availability of mixed breed dogs, people will take what's available. 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. 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.

When it comes to appearance, in the U.S., there's definitely an impact. While fur length doesn't make much of a difference in the dog's adoption speed, 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. 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. It appears classification when setting up a pet's profile matters!

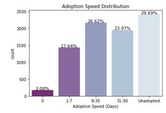
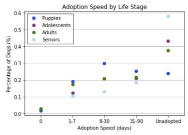
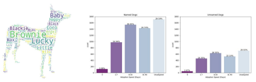
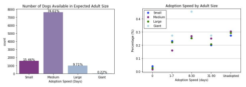
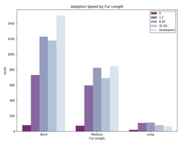
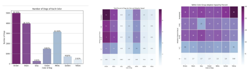
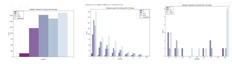
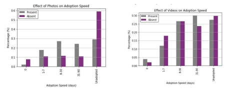
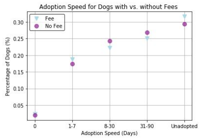
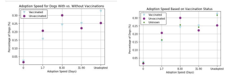
As with dogs in the U.S., the way the listing profile is set up has an effect on the speed with which an animal will be adopted. The presence of photos and as much information as possible does make a noticeable difference here. Also, the more dogs per listing the less likely you are to see an adoption. This trend holds true in Malaysia. Also of note, the presence of a video makes little difference in adoption speed, and whether a dog is free or not has little impact. 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.

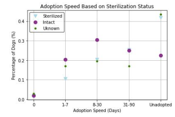
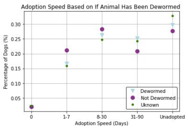
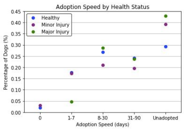
In the U.S. the health of an animal plays a large role in adoption speed. To the point that almost all shelters and rescues guarantee the dog will be sterilized (spayed/neutered), up to date on vaccinations, and dewormed. Healthy dogs also tend to be adopted faster as opposed to those classified as "special 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. 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.

Overall the data showed the results expected with a few exceptions. Malysian 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 in later projects.

Appendix A

FEATURE SPECIFIC INVESTIGATION:

FEATURE	HYPOTHESIS	CONCLUSION	GRAPH THUMBNAIL
Adoption Speed	~50% of animals will be unadopted	~30% of animals go unadopted	
Age/Life Stage	Puppies will be adopted fastest, Senior dogs will be least likely to be adopted	Puppies are adopted fastest, Senior dogs are least likely to be adopted	
Name	Named dogs will be adopted faster	Name has no effect on adoption speed.	
Breed	Purebred dogs will be adopted fastest. Labradors will be most popular.	Mixed breed dogs are adopted fastest. Labradors are the most popular	
Size	Medium dogs will be most quickly adopted.	Giant breeds are very quickly adopted. Otherwise size plays no role	
Fur Length	Fur length will have little impact on adoption speed.	Fur length has little impact on adoption speed	
Color	Black and brown dogs will be least likely to be adopted.	Brown and black dogs are least likely to be adopted.	
# of Dogs in Listing	Multiple dog listings will take longer for adoption	Multiple dog listings are slower for adoption	
Photos/Videos	Photo/video presence will increase likelihood of adoption	Photos help increase adoption speed. Videos have little impact	
Cost	HIgher cost will slow adoption	Cost has little impact.	
Vaccination	Vaccinated dogs will be adopted fastest	Vaccine status only matters in older dogs, and then it is preferred they are vaccinated.	

FEATURE	HYPOTHESIS	CONCLUSION	GRAPH THUMBNAIL
Sterilization	Sterilized dogs will be adopted faster	Unsterilized dogs are adopted faster	
Wormed vs Dewormed	Wormed animals will be adopted fastest	Worming only matters in older dogs. Then dewormed is preferred	
Pre-existing Conditions	Dogs that are healthy (have no injuries illness) will be adopted faster.	Dogs with no injury/illness are adopted faster.	 Adoption Speed by Health Status Healthy Minor Injury Major Injury Percentage of Dogs (%) 0.00 0.05 0.10 0.15 0.20 0.25 0.30 0.35 0.40 0.45 0 1-7 8-30 31-90 Unadopted Adoption Speed (days)