Capstone Project 2
Petfinder

Problem:
There are millions of stray pets around the world, some of which are fortunate enough to be adopted while many others are not. While adoption of a pet is often the definition of success, the rate at which a pet is adopted is also a key success factor - pets that take a long time to adopt contribute to over-crowded animal shelters and can prevent taking on new strays. Sadly, pets that are not adopted eventually need to be euthanized.

Client:
Petfinder.my is Malaysia's leading animal adoption agency and is sponsoring a Kaggle competition to improve the adoptability of pets. Petfinder's objectives are...

1. an improved understanding of the most important features to share about a pet up for adoption
2. model/s that can be used to predict the speed at which a pet is adopted based on the pet's listing

Petfinder intends to share the results with animal shelters around the world to help them improve their pet profiles' appeal and increase the number of (and rate of) adoptions; reducing animal suffering and euthanization.

Dataset:
Petfinder has provided the data at this link

Approach:
This is a supervised classification problem, with over 10 features, including images and a text description for each pet's profile. The planned approach is...

Data Wrangling:

1. Limited data wrangling is anticipated for this dataset and will be completed as issues arise. For example, profile descriptions not in English will probably be filtered out.

Data Storytelling:

1. Analyze the features...
    a. looking for correlation between each feature and the prediction feature
    b. looking for correlation between features.
    c. to evaluate if the data has a linear, gaussian, or other distribution type
2. Employ Lasso machine learning algorithm to identify which features are best to carry forward into machine learning.
3. Various visualizations of the data (histograms, etc...)
4. NLP: Calculate word counts, word relevance scores, metrics on most common words that lead to adoption, length of description, etc...

Machine Learning:

1. Algorithms to employ on the non-NLP features will be determined based on the Data Storytelling results.
2. For NLP, employ bag-of-words and Word-to-Vec algorithms

3. Combine most effective approaches of non-NLP and NLP methods; considering one or more of the following...
    a. Employ an ensemble between the best performing non-NLP algorithms and the NLP algorithm
    b. An algorithm that combines the non-NLP and NLP features
    c. Using the NLP output as input for non-NLP algorithms
4. *Stretch goal:* employ NLP deep learning to the profile description using the Keras library

Exclusion:

1. Not planning to work with the image data.