Data Wrangling: Maximizing Animal Shelter Funds

Import and explore dataset

For this analysis, I used the Austin Animal Shelter dataset, which includes intake and outcome data for shelter animals from October 2013 to present.

1. Import dataset using python's Pandas package
2. Explore the shape, structure, and columns of the dataset.
    a. Each row represents a shelter animal, and the dataset contains many columns with descriptive data on each animal.
    b. I determined during this step that there was already a column that displayed number of days that each animal spent in the shelter; however, it needed to be converted to a datetime value.

Adoption times by animal type

3. I used slicing to determine that there were dogs, cats, birds, and 'other' in the dataset, with dog appearing most frequently.
4. I used a boxplot to briefly analyze each group.
    a. The vast majority of animals (75% for all categories) fall into a relatively short adoption time range of less than 20 days.
    b. All groups had a substantial number of outliers beyond the whiskers of the boxplot, making it difficult to interpret this group.
5. I sliced the animals into groups by each animal type.

Adoption times by dog breed

6. I first looked at dog breeds and determined that the formatting of this column created a lot of complexity due to the presence of mixed breeds. I created a column that returns 1 or 0 if the dog is a mixed breed or pure breed, respectively.
7. I split the breed column into a list of breeds and then used Panda's get dummies function to return 1 or 0 if the list contained a specific breed.
8. I concatenated the dummies dataframe back to the dogs dataframe.
9. I then used slicing and a melt reshape the table to be long with fewer columns instead of wide, with a new column containing a breed name and a corresponding column with 1 or 0 if the dog contained that breed.
10. I used a violin plot and a boxplot to show adoption times by breed, split by purebred and mixed breed. Both plots were difficult to read due to the number of outliers, and for the boxplot, I hid outliers to get a better grasp on the majority of the data.

11. I found that due to the number of breeds, I had to create another column showing only the most common breeds and lumping the others together as “other”. I chose breeds that appear >400 times as the cutoff.

Dog adoption times by AKC Dog Group

12. Since there weren’t any clear patterns shown when I looked by breed, I pulled a table that shows breeds by AKC Dog Group in using Pandas and merged this dataframe with the dog breeds dataframe. Once again, notable patterns were difficult to determine with outliers present, so I hid them to see only the majority of the data.

Purebred v. Mixed Breed Dogs

13. I looked at all dogs (with outliers hidden) using just purebreds vs mixed breeds and saw only a slight increase overall on adoption times for mixed breeds.

Adoption time by dog age

14. I also looked at age of dogs using a scatterplot to determine if age affects adoption times, but the age variable was not continuous (rounded by year) and created odd groupings. I decided to instead use age groups as a categorical to look at age as a bar chart.

Adoption time by coat color

15. I used a very similar analysis to the get dummies for this analysis, as described above.

Repeat analyses looking only at ‘outliers’

I repeated all above analyses but sliced using adoption times > 100 days to see if we could learn more about the outlier groups hidden from the initial analyses.