Capstone Project 1: Milestone Report

For the first capstone project, I have been working with data from the AAC, which compiled shelter data on animals that were admitted/ released. This data included the specific breeds as well as the outcome, such as adopted or euthanized. This data was of interest as boosting adoptions is a goal of any animal shelter. The goal of my analysis was to see if there was a correlation between animal age and chance of being adopted. I also wanted to see if there were any breeds that resulted in higher adoptions. This is useful information to the client, AAC, as they can target specific breeds/ ages on their websites/ marketing in order to boost adoptions.

The dataset from AAC was already fairly clean with very clear column headers, but I did perform some wrangling steps to get the specific columns I needed. One of the first steps I took was creating datetime objects from the intake and outtake dates so that I could calculate the shelter length of each animal. I performed some exploration of the dataset and saw that there were numerous outcome types that did not seem sensical for AAC's marketing purposes. For example. There was Return to Owner and Transfer as outcome options. These are not true cases that I wanted to review as they are not relevant to the adoption or euthanization outcomes, so I isolated only the outcomes that resulted in adoptions or euthanizations.

I also wanted to calculate the age of each animal upon outcome, so I converted the DOB column into a datetime object and subtracted from the outcome datetime. I also wanted to block the shelter length and age of outcome into meaningful buckets, so I used the time delta to convert the shelter length into weeks and age of outcome into years. This allowed those columns to be split into more sizable chunks that could be used for comparison. Once the age of outcome and shelter length columns were cleaned and converted, I combed through them and removed any outliers (there were 5 entries that had -1 as age of outcome so were removed). I also checked for any null values in the dataset and had these removed (there were 10 null values for outcome type, so just removed due to the low number).

Once the data was cleaned and ready for analysis, I began performing the EDA, which allowed me to see interesting trends in the dataset. Initially looking at the data, I noticed that adoptions outweighed cases of euthanization, but I also noticed that there seemed to be a trend with younger animals being adopted and euthanized, which was unusual to me. I expected older animals to lead the cases of euthanization, but that was not the case. At this point in the EDA, I wanted to look at breed specific trends so see if there were any patterns of use, especially with the cases of euthanization since they differed from my expectations.

Looking at the top breeds that were euthanized, I saw that Pit Bull Mixes ages 1-2 lead the charge for dogs; however, Domestic Shorthair Mix for cats more than tripled the dog amounts. The DS mix for cats led the top three spots for euthanizations when grouped by breed and age of outcome. I found that very strange, especially since the ages ranged from 0-2. I did not expect cats to lead the euthanization stats, so my next step was to review the top breeds for adoptions. I saw that the top spot was also the Domestic Shorthair Mix aged 0-1 with over 7500 cases. The second top breed was Labrador Retriever Mix aged 0-1 with over 1200 cases. I found that to be a shocking difference between the first and second top adopted breeds and pondered why there would be such a large discrepancy. It also seemed weird to me that cats accounted for the top number of adoptions because I find dogs to be more popular. As a cat owner myself, I usually feel in minority when discussing pets with friends or colleagues.

Thinking through the data, I came to the realization that many cat owners do not specify their cats' breeds as most people do not know the distinct breeds of cats. People are familiar with the different dog breeds, so it's very common for people to throw breeds around in conversation, but the same is not true for cats. I often refer to my cats as Domestic Shorthair because it is simply easier to say that rather than list out the actual breeds. After pondering that, I felt that cat data was actually throwing my numbers off and decided to solely focus on dogs for this project.

Once the cats were removed from the dataset, I looked at the top ten adoptions and euthanizations grouped by breed and age of outcome. I plotted histograms of this data to have a clearer picture of the spread. Interestingly, I saw that the top breed euthanized was the Pit Bull Mix ages 0-3, which seemed bizarre as the ages were so young. Looking at the adoptions, I saw that the Labrador Retriever Mix aged 0-1 accounted for the top adoptions by a landslide. The next breeds were Chihuahua Shorthair Mix (aged 0-1) and Pit Bull Mix (aged 1-2). The histogram was able to clearly show labrador retrievers accounted for about double the second highest breed adopted and then trailed off from there. The EDA showed interestly trends for the data that I wanted to explore further in the inferential statistics section.

The first thing I wanted to check was the correlation between age of outcome and adoption as I expected animals with higher ages led to longer shelter times; however, there was not much of a covariance between those two variables. From there, I calculated the shelter mean and STD so that I could perform a binomial distribution against the rate of adoptions against euthanizations. I used the np.randon.binomial function to achieve this, and plotted as a histogram. I saw that on average, there was a 93% chance of an animal being adopted and the sample data spanned from 85 - 97.5%

I also was able to compute the ECDF of this dataset and could see the majority of data was between the 87.5 and 97.5% mark. I then graphed the PMF of this sample distribution to create an alternative way of viewing the data. From there I wanted to perform poisson distributions of the shelter adoptions and calculate the mean as well as the percent of data less than or equal to 3 years. I saw that the mean shelter length was 3 and 64% of adoptions occurred for shelter lengths 3 and below. I then performed a poisson distribution of the age of adoptions and saw that the mean age was 2 and 73% of adoptions occurred for animals 2 and younger.

From my EDA, I learned that Pit Bull mixes are popular breeds for adoptions as well as euthanizations, I wanted to run a simulation on this. I created a new dataframe of Pit Bull mixes that were euthanized as well as adopted and then calculated the probability of pit bulls being adopted, which was 85%. From there, I conducted a binomial distribution on this dataset and saw that from the histogram and ECDF, 75% and 90% of pit bulls were adopted out of 100, so the majority of pit bulls are being adopted. I also performed a poisson distribution on the age of pit bulls euthanized and saw that the mean age of euthanized cases was 3 and 63% percent of pit bulls are euthanized between the ages of 3 and below. That is a high percentage for such a low age and made my wonder why so many young pit bulls are being euthanized. I performed the same poisson distribution for the age of adoptions and saw that the mean age was 2 and over 75% of pit bulls were adopted between the ages of 2 and below.

Since I saw that the labrador retriever mix was the highest adopted dog breed, I wanted to also perform a poisson distribution of this breed's age of adoption for review. I saw that the mean age of adoption was 1 nad 85% of labs were adopted between the ages of 2 and below. After that, I wanted to perform a binomial distribution of labs adopted out of total population. I plotted this data as a histogram and ECDF and saw that spread was between 5 and 20%, which is the percentage of labs adopted out of the total population. I had expected this to be higher since Labs had such a large lead in the number of adoptions, but I am guessing this is due to the fact that labrador retriever mix is quite specific. Some people may just list labrador or retriever rather than actually labrador retriever mix.

I then wanted to review the amount of labrador retriever mixes euthanized verse adopted and performed a binomial distribution of the probability of labs being euthanized. From the distribution, I plotted the histogram and ECDF and saw that the data fell between 2 and 10%, which is very low and expected since this is the highest breed adopted for dogs. From the statistics performed, It seems to be in the AAC's best interest to market labrador retrievers to boost adoptions and shy away from marketing pit bull mixes as there are an alarming amount euthanized between the ages of 0-3. The amount seems too high to account for illness so there must be another reason such as overpopulation or aggression. The data for the AAC does not provide a reason for euthanizing, but that would be helpful to take a deeper look into what led to these cases.