Our data set is the intake data of a specific shelter in Austin, Texas. (<https://data.austintexas.gov/Health-and-Community-Services/Austin-Animal-Center-Outcomes/9t4d-g238>) This intake data is updated regularly, so it contains all data from October 2013 to October 2019. This data set is very large with 109,682 observations contained in it. However, since they keep track of the animal IDs, we know that there are only 98,307 unique animals represented in the set. Further analysis revealed that this set also consists of all different species of animal that entered the shelter, so a fair bit of discretion was needed to get the subset of observations we wanted to work with. In pursuit of that goal, we decided our unit of analysis was to be every unique dog that entered the shelter from the start of the data set until October 17th of this year.

     In order to get the subset we desired from the data set, a fair bit of data cleaning was needed. Since this data represents all animals coming into the shelter, we had to clean out all other species of animals as mentioned above. We then had to determine the outcomes of each dog coming into the shelter. This included things like those that were returned to their owners after being lost since we were only interested in lone animals that were adopted or not, not lost pets. To ensure a trend could actually be determined based on breed, which we felt was one of the stronger predictors of adoption outcome, we left out breeds with less than 200 dogs of that breed coming into the shelter.

Since the purpose of our study is to determine what factors affect the likelihood that a given dog will be adopted, we decided to pick a visualization technique that reflects a percentage chance that a dog would be adopted. With that in mind, we are planning on having the adoption rates be measured with a multiple line graph, each one representing a different breed. The X-axis would represent adoption rate and the Y-axis would represent age, the other predictor we predict will be the largest contributing factor. These two predictors were selected based on common knowledge of the group members.

For our second model we will use bagging and boosting to determine the importance of each predictor. Due to the nature of bagging we will present the outcome in a confusion matrix. A variable importance plot will give a more easily understandable way to understand the effects of each predictor.