Your final report should have the following components and follow the APA style as much as possible.

* **Research problem (10pts)**: Describe the task you want to achieve. What is the outcome of interest? What are you trying to predict? Why is it important? What are the potential benefits of having a predictive model for this outcome? Discuss potential applications of such a model.

As people stayed home for most of the early days of the coronavirus pandemic, the demand for goods and services increased. Some of the increased demand was not surprising and likely expected; items such as home workout equipment, trampolines, and lumber. However, some of the demand took the industry by surprise: yeast for baking, and pets. In fact, the interest in pet adoption increased so much that shelters were regularly reporting empty kennels and sifting through dozens of adoption applications for a single puppy.

Unfortunately, as vaccines were rolled out and people began returning to work and school, shelters and foster groups filled up with animals that were no longer compatible with people’s lifestyles. The decrease in demand for dogs means that people can be more selective in the kind of dog they adopt. However, the information provided by animal shelters and rescue groups are often based on a short period of time with the animal and the animal’s appearance. One shelter’s Border collie mix might be another’s spaniel or shepherd mix. This best guess breed identification can have significant impacts on a dog’s future and could be the difference between adoption and euthanasia. Objectively identifying the impact of a dog’s listed breed on the length of stay in an animal shelter could provide shelters with the information needed to shift away from listing a dog’s breed as the primary information for potential adopter and toward a more holistic evaluation of a dog’s temperament and future needs.

* **Description of the data (15pts):** Describe core features of the data, any additional features you produced from existing features and how, basic descriptive statistics about these features, and any missing data analysis you conduct. The description should be sufficiently clear that the instructor understands all the variables included in your modeling.

The data were obtained from kaggle: <https://www.kaggle.com/aaronschlegel/austin-animal-center-shelter-intakes-and-outcomes?select=aac_intakes_outcomes.csv>. It was originally provided by the Austin Animal Center in Austin, Texas. The data include information about the intake and outcome of the animal, and details on the type and condition of the animal. A brief examination of the data revealed that the animal shelter takes in animals in addition to typical domestic pets (cats and dogs). For the purposes of this study, the following types of animals were excluded: cats, birds, and animals that were classified as other, including rabbits, bats, snakes, raccoons, ferrets, reptiles, and other wild animals that live in close proximity to humans. Additionally, dog breeds with sample sizes less than 20 were excluded as this small sample made it difficult to accurately model the length of stay for the breed. The final data set included variables of the animal (breed, age on intake, sex, condition of the animal), circumstances of the animal arriving at the shelter (type of intake, month of intake), and specifics of the outcome of the animal (outcome, month of outcome, time spent in the shelter, measured in days).

Initial data visualization was performed to understand the data available. This included visualizing the number of animals taken into the shelter each month and further exploring the number of each type of animal taken in each month.

* **Description of the models (15pts):** List at least three different modeling approaches you apply to this dataset. Describe each model, why the given model was selected, which hyperparameters to be optimized and how. Also, discuss how you plan to evaluate model performance.

In order to predict the length of stay of shelter dogs, three types of modeling approaches were explored: linear regression, linear regression with ridge penalty, and bagged trees. These models were chosen for their increasing complexity to determine the extent to which the increasing complexity added value to or impacted the predictions and importance of variables used in the predictions. All models were fit with 10-fold cross validation for comparison purposes. The performances of the models will be compared using the values of R-squared, MAE, and RMSE.

* **Model fit (20pts):** Provide the results of your model evaluation. Compare and contrasts results from different fits, including a discussion of model performance. Discuss your final model selection and the evidence that led you to this selection. If it is a classification problem, how did you choose a cut-off point for binary predictions? Did you consider different cut-off points?
* **Data visualization (5pts):** Include at least two plots (or more) to help communicate your findings. The plots may be of initial data explorations, fits of individual models, and plots displaying the performance of competing models.

Scatterplot of intakes by month

* **Discussion/Conclusion (25pts):** Discuss and summarize what you learned. Which variables were the most important in predicting your outcome? Was this expected or surprising? Were different models close in performance, or were there significant gaps in performance from different modeling approaches? Are there practical/applied findings that could help the field of your interest based on your work? If yes, what are they?
* **Reproducibility (10pts):** Provide a link to the GitHub repo at the beginning of your report as a note.

The final report should be no longer than 10,000 words (approximately 15 single-spaced pages), excluding the tables, figures, and references.