I. Background

The survey data provided to us by the Dog Aging Project afforded us many avenues to analyze the health of domestic dogs across the country from both an epidemiological standpoint, such as lifestyle determinants of age-related diseases within our population, as well as a socioeconomic standpoint, such as whether the income of a given owner has a substantive effect on the health of a given canine. This analysis attempts to add a third prong to our approach, namely in the form of a psychological analysis. While the practice of neutering or spaying dogs is emphatically supported by veterinary science professionals to aid in the reduction of detrimental sexual behaviors and health risks exhibited in dogs, such as roaming, urine marking, the birthing of unplanned litters, and the development of certain types of cancers, less studies have been conducted to ascertain the relationship between neutering, spaying, and the behavioral development of dogs. This analysis explores this relationship within our population in an effort to assess if further empirical studies are warranted.

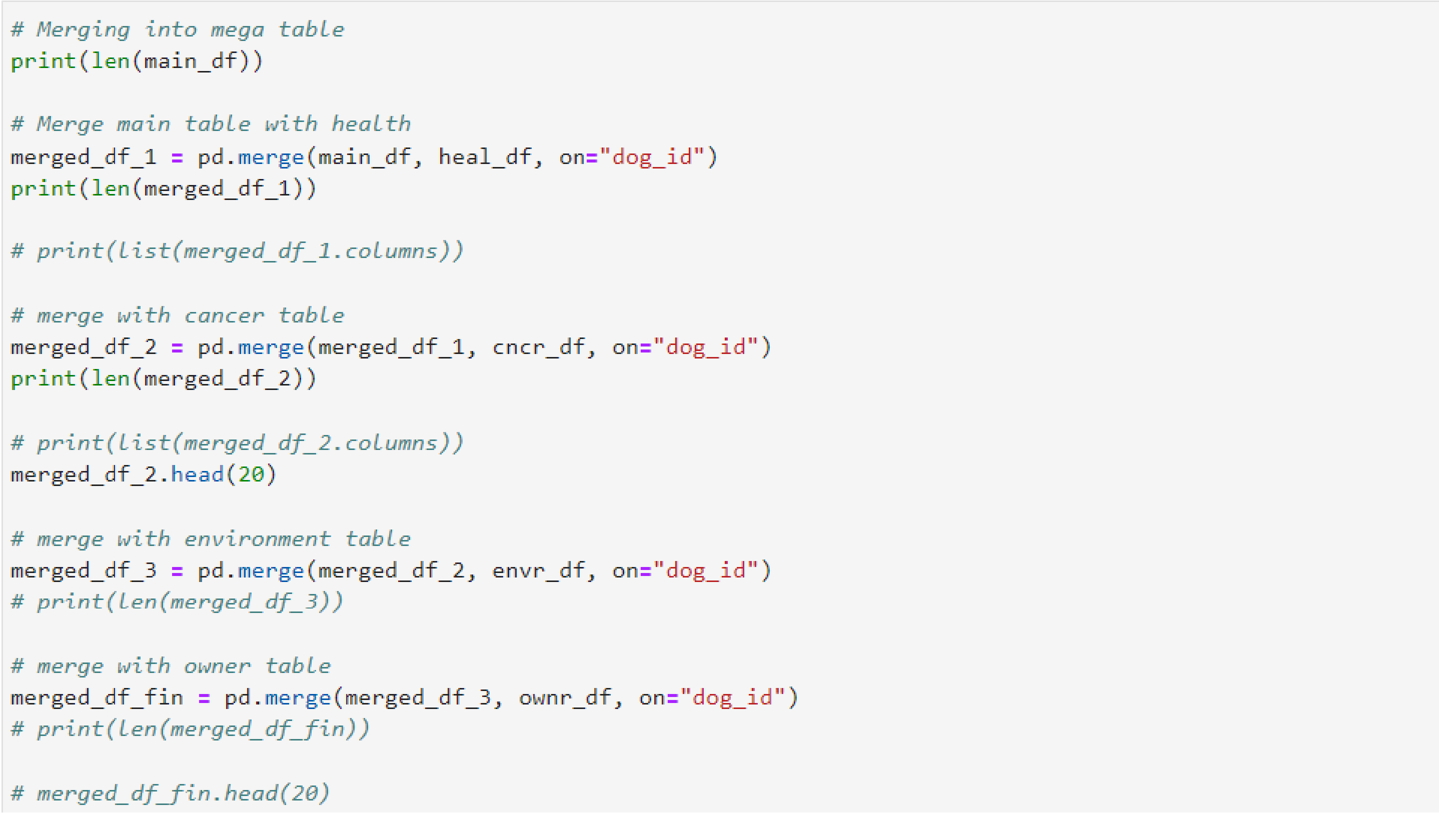
II. Data Acquisition

The data provided to the *k-means-k-9s* team came to us from the Dog Aging Project in tabular form, as csv files. The data was differentiated by the domain of the survey, as defined below:

- **CSLB:** This survey is sent to study participants annually to assess age-related cognitive and behavioral changes in dogs. The first time a participant takes the survey, a baseline cognitive score for their dog is established. Administering the survey in subsequent years allows the organization to assess the cognitive state of dogs over an extended period of time.

- **ENVIRONMENT:** The environment dataset was the only non-survey dataset provided. This dataset contains a single unique record for each dog captured and provides a snapshot of environmental factors that may impact a dog’s health over time, such as the presence of pollutants, the temperature, and walkability of the dog’s locale. This data is updated by the Dog Aging Project on a monthly basis.
- **cancer_conditions:** The cancer conditions dataset contains information related to cancer diagnoses for our canine population. Participants in the Dog Aging Project complete survey questions pertaining to cancer diagnoses upon entering the project.
- **dog_owner:** The dog_owner dataset contains owner-specific information (i.e. education level, ethnicity) as well as dog-specific information (i.e. flea and tick medication usage, inflammation medication usage). Participants in the Dog Aging Project complete survey questions pertaining to owner demographics and general dog information upon entering the project.
- **health_condition:** The health_condition dataset contains details of all non-cancer related health conditions reported in participating dogs (i.e. fungal infections, eye infections). Participants in the Dog Aging Project complete survey questions pertaining to health conditions aside from cancer upon entering the project.

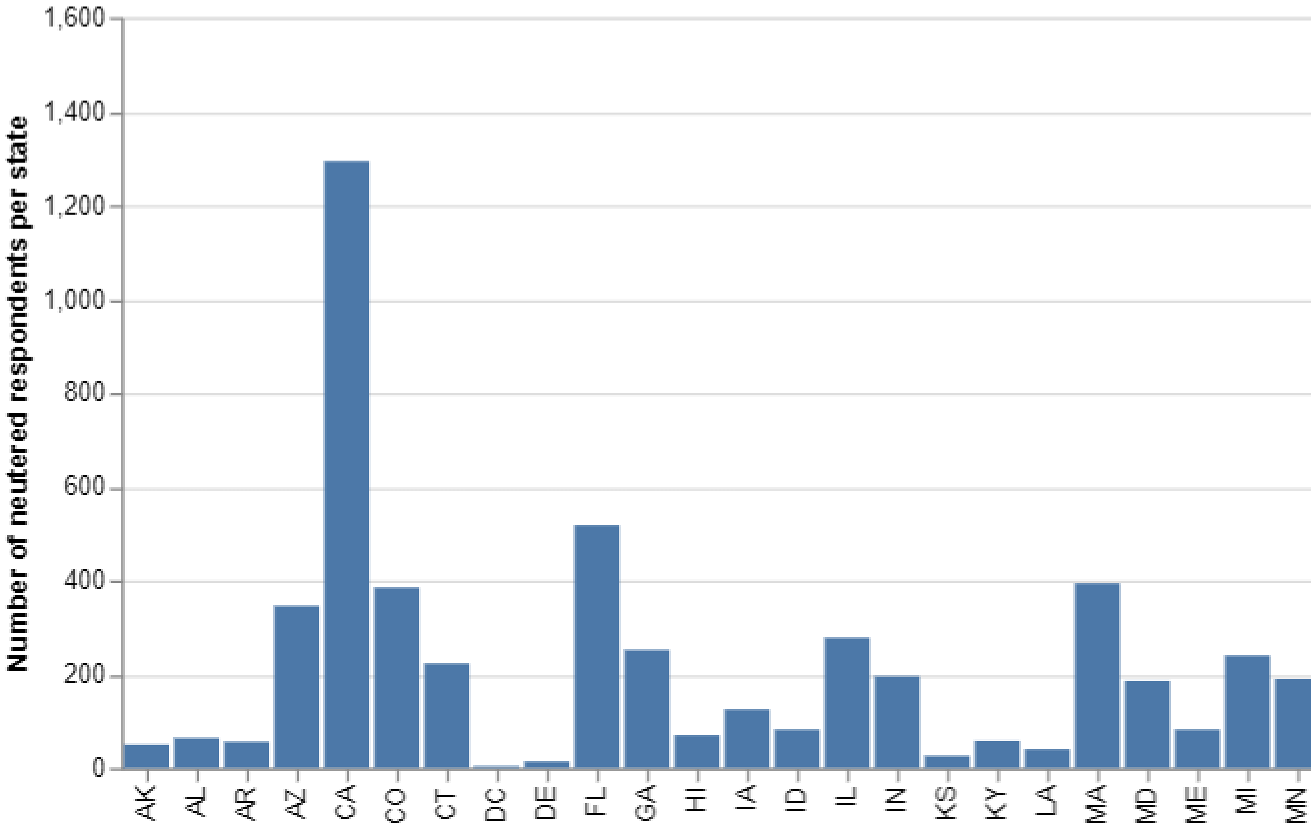
The datasets provided to the team contain records uniquely identifiable by the ‘dog_id’ field, which made merging the datasets together a very doable task.



```python
# Merging into mega table
print(len(main_df))

# Merge main table with health
merged_df_1 = pd.merge(main_df, heal_df, on="dog_id")
print(len(merged_df_1))

# print(list(merged_df_1.columns))

# merge with cancer table
merged_df_2 = pd.merge(merged_df_1, cncr_df, on="dog_id")
print(len(merged_df_2))

# print(list(merged_df_2.columns))
merged_df_2.head(20)

# merge with environment table
merged_df_3 = pd.merge(merged_df_2, envr_df, on="dog_id")
# print(len(merged_df_3))

# merge with owner table
merged_df_fin = pd.merge(merged_df_3, ownr_df, on="dog_id")
# print(len(merged_df_fin))

# merged_df_fin.head(20)
```

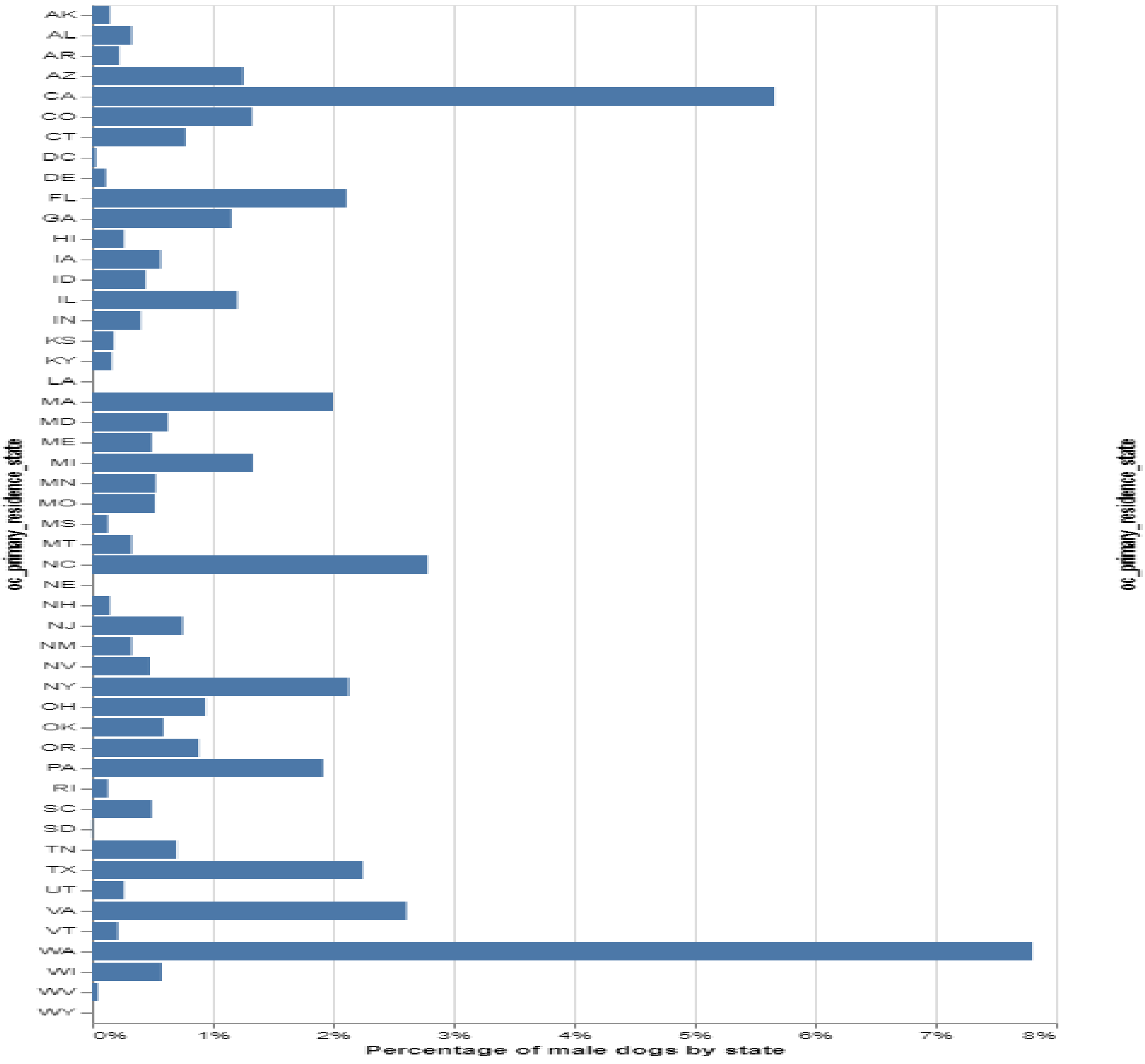
III. Exploratory Data Analysis

Prior to moving on to the model construction stage, the team conducted exploratory data analysis to allow us to make more informed assertions about our final results, while also helping us to ascertain whether any salient imbalances existed in our data. As a perfunctory step, I wanted to get a better understanding of precisely where the owners of spayed and neutered dogs in our population resided, as depicted by the raw count distribution below.



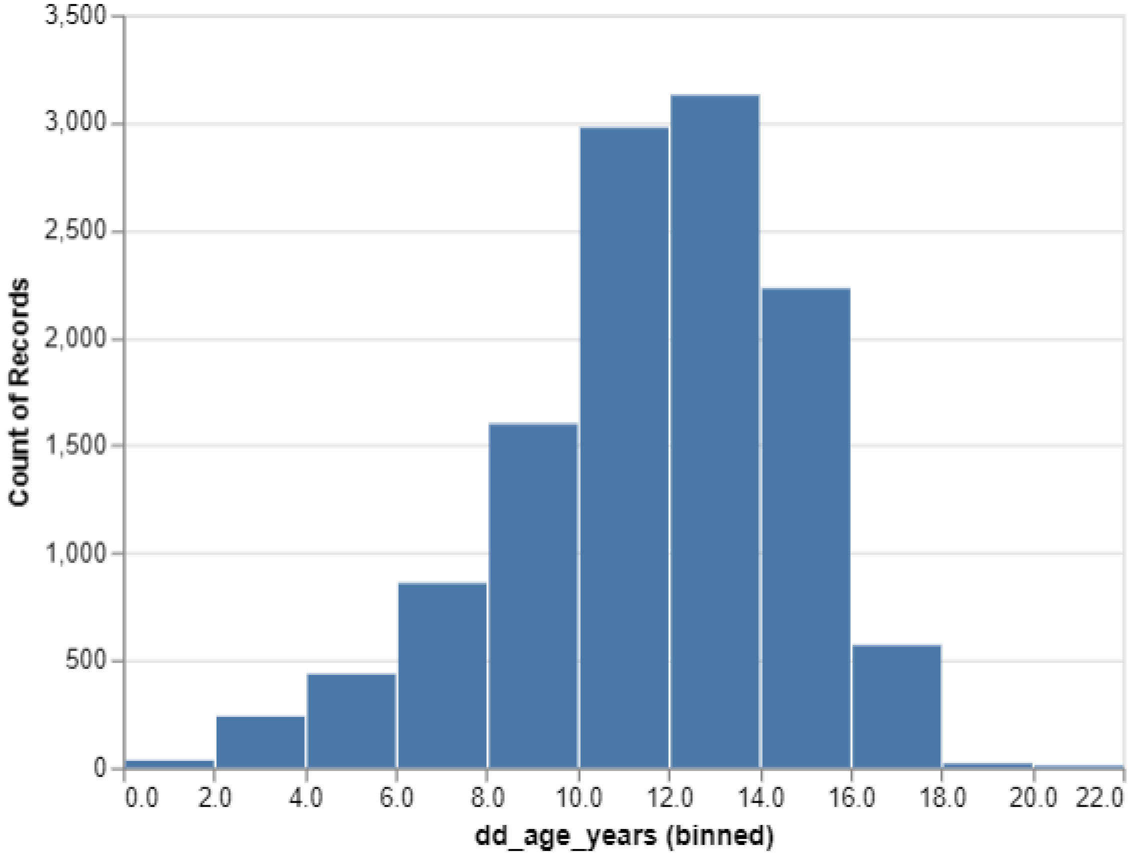
At an immediate glance, my initial intuition regarding the geographic location of the participants was confirmed. Namely, that most respondents reported a permanent address in the home state of the Dog Aging Project, Washington. California also reported a high response rate, to which an easy explanation didn’t readily come to mind.

Following this data exploration, I wanted to get a better understanding of the canine gender distribution for all participating states. Since male dogs have historically exhibited higher rates of undesirable behaviors – particularly aggression – when unneutered vs. un-spayed females, I wanted to validate whether or not male dogs made up a disproportionate percentage of our population.



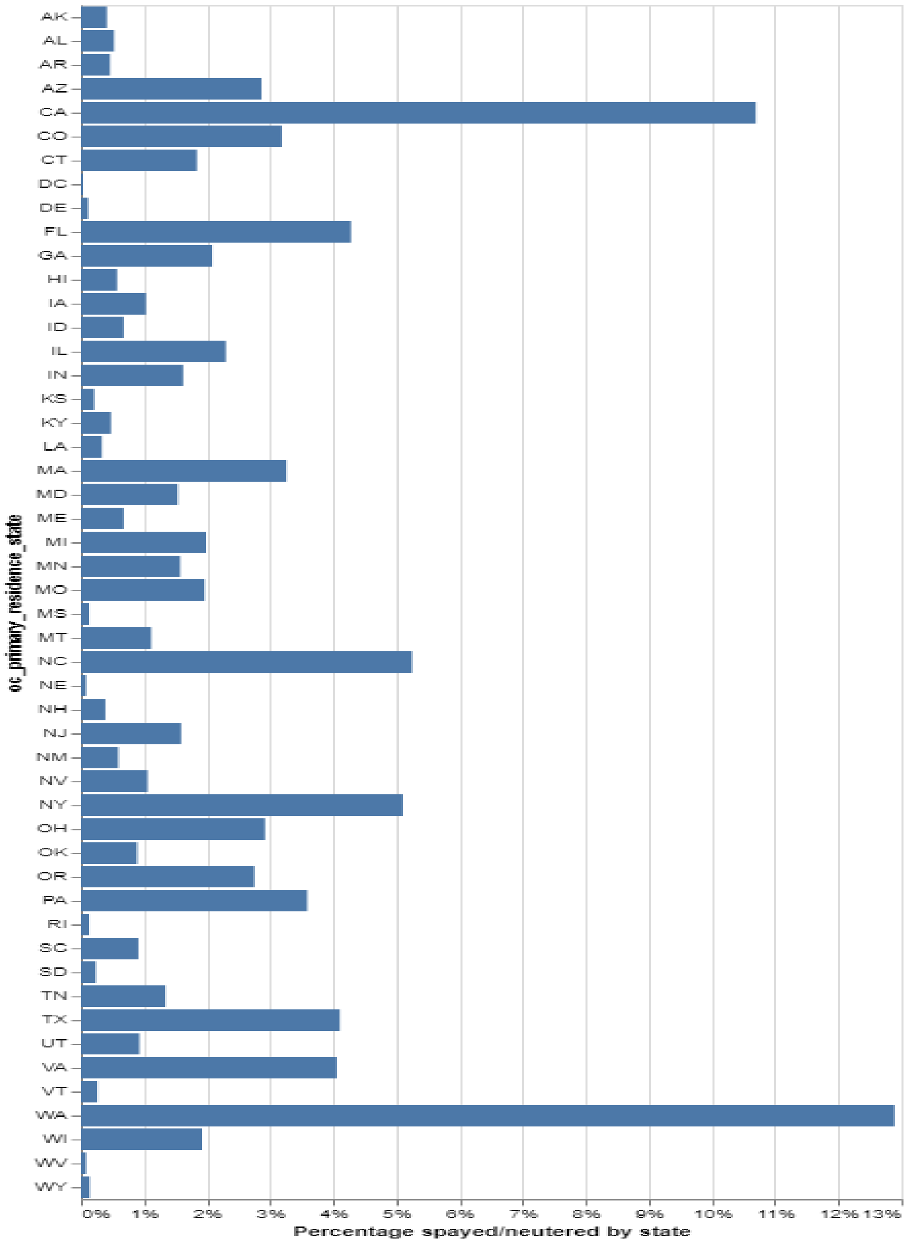
Upon constructing the concatenated bar chart visualization above displaying the gender distribution of male and female dogs by state, I was able to confirm that the distribution is indeed fairly even. However, this graph did unveil a separate issue, namely that the percentage totals for each state did not add up to 100%, indicating a general presence of non-responses for this particularly important field. Because of this, I concluded that utilizing gender as a predictor variable for my classification models would be problematic at best.

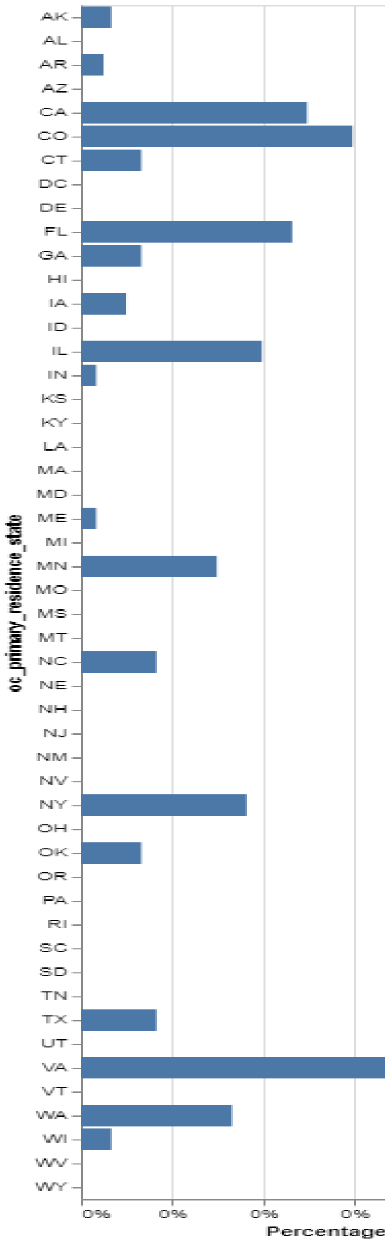
Thirdly, I constructed a histogram of our canine populations’ age distribution. Since older dogs are more prone to cognitive decline, which can affect behavior, checking for a higher presence of senior dogs within our population seemed a particularly valuable point of investigation.



The distribution was intriguing, in the sense that it was visibly skewed to the right, towards more advanced ages. Current consensus agrees that large dog breeds are considered senior as early as ages 6-7. medium dog breeds at ages 8-9, and smaller dog breeds at 10-12. The brunt of respondents appeared to fall between ages 8 and 16, suggesting that much of our population could be considered of senior age. I kept this observation in mind, in the event I discovered a large number of behavioral abnormalities in our population.

As a final explorative step, I visualized the percentage of dogs, by state, that had been spayed or neutered vs. the un-spayed and unneutered population and placed them side by side.





It was immediately apparent that, for several of the states, no respondents reported that their dog had been un-spayed or unneutered. Ultimately, I viewed this finding as interpretable in two ways. It was possible that my preconceived notion was correct, namely that the majority of dog owners have their pets spayed or neutered as a best practice. This also may have been evidence of non-response bias, in the sense that it was quite possible that owners didn't want to report that their dog was un-spayed or unneutered due to the societal stigma that may be associated with such a choice.

IV. Model Construction

Chi Square Tests

Prior to moving onto the model construction stage, I decided to conduct chi square tests in an effort to discern whether or not the behavioral variables I had identified were, at the very least, somewhat correlated with my dependent variable *dd_spayed_or_neutered*, and, in the event they were not, to discard them. To make management of my predictor variables easier, I identified several broad behavior types to bucket them into: Aggression related, fear and anxiety related, excitability related, separation related, attachment related, training related, and miscellaneous. With almost no exception, the p-values produced from the chi squared tests of 'dd_spayed_or_neutered' against the social behavior explanatory variables were very small and significant up to an alpha of 0.01, indicating that the observed data was very unlikely to have been produced by pure chance - in other words, that there was in fact a relationship between the variable 'dd_spayed_or_neutered' and the behavior variables - and that the probability that the data could deviate from the null hypothesis of 'no relationship' in future samples was nearly 100%. I was hesitant to put unwarranted stock into these results, so prior to training my models, I ensured to normalize the data first, to mitigate the chance of producing inflated results.

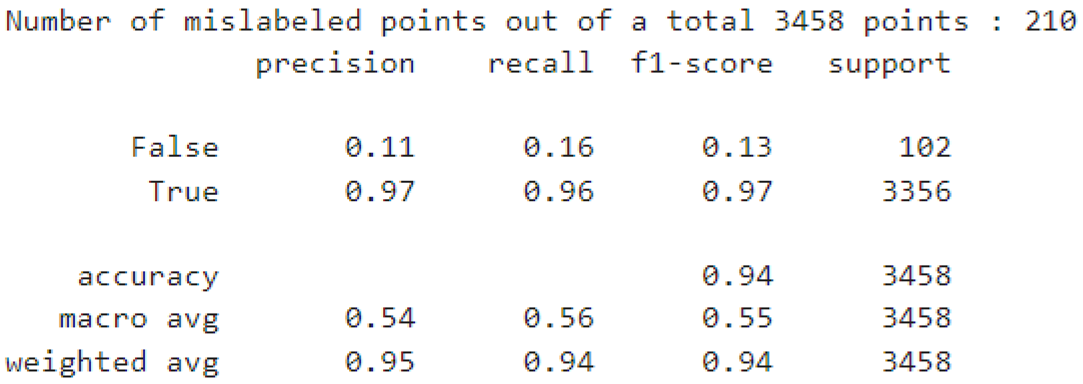
Logistic Regression

Logistic regression served as the baseline classifier for this empirical analysis. To aggregate the data, I split the records into distinct dataframes, differentiated by the aforementioned behavior types. To normalize the data, I utilized the StandardScalar class from the sklearn module. I then constructed separate logistic regression models – one per behavior type – so I could obtain a more easily interpretable picture of which behavior types contained the strongest predictors. Only six predictor variables were found to be statistically significant at any level. Three were aggression related, one was excitability related, and two were miscellaneous, with one pertaining to hyperactivity and the other pertaining to frequent urinating inside the house. On personal intuition alone, these behaviors generally seemed more likely to be exhibited by more aggressive, excitable unneutered and un-spayed canines.

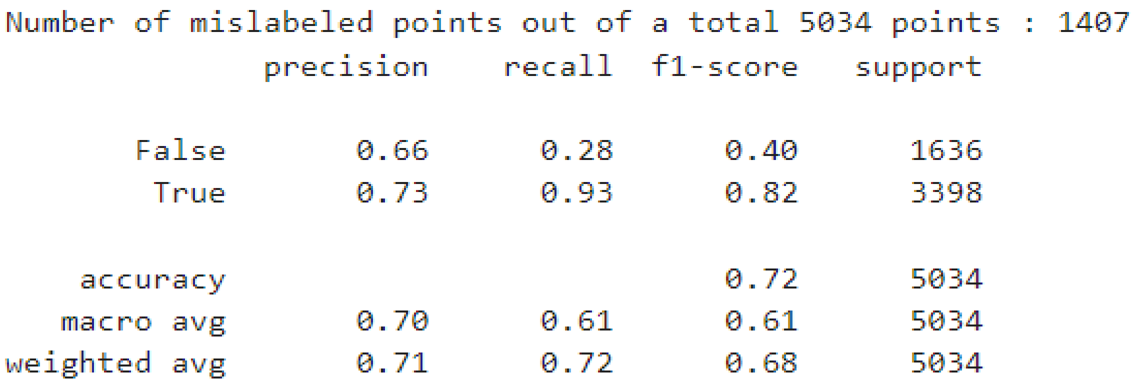
As a whole, model accuracy was mixed, with accuracy scores ranging from 39% (attachment) to 56% (miscellaneous).

Gaussian Naive Bayes

Since Naïve Bayes is generally more suitable for datasets with fewer variables and fewer records than logistic regression, I decided to utilize this classifier type as well. I followed a similar procedure to curate the data, splitting the records into distinct dataframes by behavior type and subsequently normalizing them. Upon constructing my initial set of Naïve Bayes models, I discovered that the models were highly biased towards the majority class, almost universally predicting True (spayed and neutered), a potential drawback of our data that was alluded to me by my exploratory data analysis. To combat this class imbalance, I utilized the imblearn package to oversample the minority class and reconstructed the models. Precision and recall scores for the minority class, False, generally improved, while the scores for the majority class were less inflated. The images below, displaying the classification report for the aggression model before and after normalization, exemplifies this.



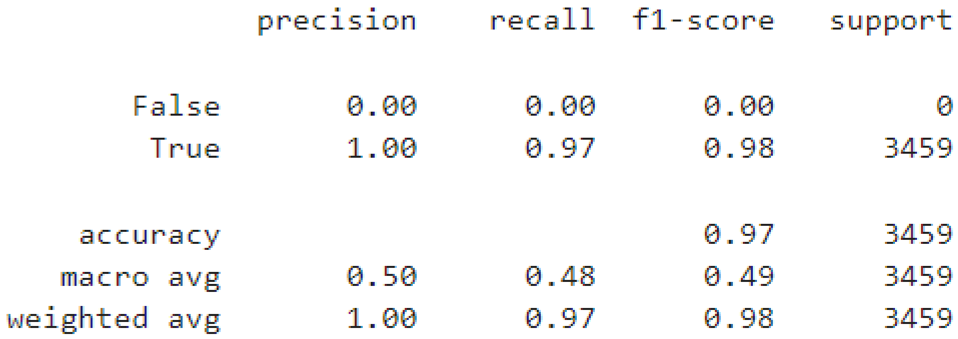
```
Number of mislabeled points out of a total 3458 points : 210
              precision    recall  f1-score   support

       False       0.11      0.16      0.13       102
        True       0.97      0.96      0.97      3356

    accuracy                           0.94      3458
   macro avg       0.54      0.56      0.55      3458
weighted avg       0.95      0.94      0.94      3458
```



```
Number of mislabeled points out of a total 5034 points : 1407
              precision    recall  f1-score   support

       False       0.66      0.28      0.40      1636
        True       0.73      0.93      0.82      3398

    accuracy                           0.72      5034
   macro avg       0.70      0.61      0.61      5034
weighted avg       0.71      0.72      0.68      5034
```

<u>Decision Trees and Random Forest</u>

As a final step, I decided to implement more robust modeling methods – namely, decision tree learning and random forests – to see if utilizing them would augment performance. While the set of decision tree models performed somewhat fairly for some behavior types after oversampling the minority class (some models still predicted the majority class one hundred percent of the time), the random forest models predicted the majority class one hundred percent of the time across the board, resulting in zero scores for precision, recall, and the f1 metric, as exemplified below.



```
              precision    recall  f1-score   support

       False       0.00      0.00      0.00         0
        True       1.00      0.97      0.98      3459

    accuracy                           0.97      3459
   macro avg       0.50      0.48      0.49      3459
weighted avg       1.00      0.97      0.98      3459
```

Due to the relatively limited number of records made available to us, it was not overly surprising that an ensemble learning method produced substandard results.

V. Concluding Thoughts

After concluding my empirical analysis, I was left with a firm affirmation that the relationship between spaying and neutering and the prevalence of certain undesirable behaviors in male and female dogs warrants further study. Although the models I constructed weren’t perfect predictors, the performance scores most produced, despite the underlying data quantity problem, were firmly average and too high to discount entirely.

The data provided to us by the Dog Aging Project was outdated, with the most recent survey responses having been recorded in the year 2020. With participation having surged to over one-hundred thousand in the year 2022, I believe that future researchers who are able to obtain a comprehensive dataset will be able to produce far more robust and productionizable results. All in all, lack of size was undoubtedly one of the biggest impediments to model performance for all team members.