Introduction:

The dataset that we are analyzing as part of this project is the 'Shelter Animal Outcome' dataset from kaggle.com. This dataset contains intake information for companion animals from the Austin animal center including breed, color, sex and age. Analyzing the dataset will help us understand the trends in animal outcomes as well as predict the outcomes.

Problem Definition:

To predict the outcomes for shelter animals based on the given dataset.

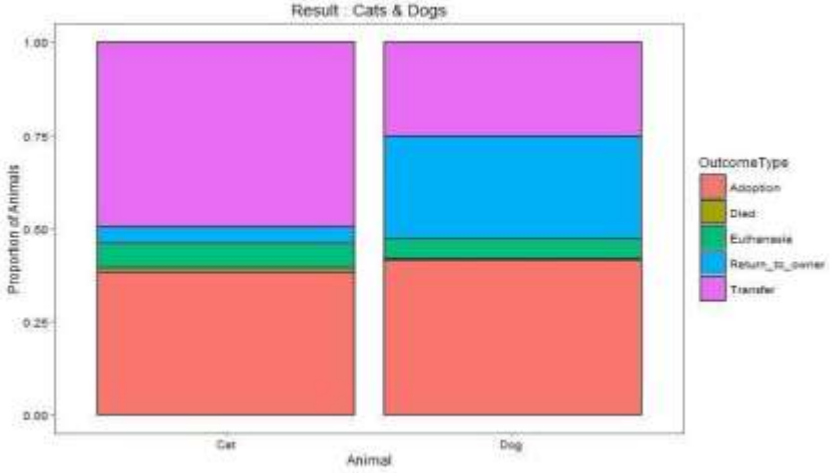
- **Dataset Details:**
  - No. of Attributes: 10
  - No. of instances: 26729
  - Type of outcome variable: Class value
- **Attribute Description:**
  - Animal ID: A unique Id for each animal
  - Name: Names of the animals.
  - DateTime: Date and time of outcome
  - OutcomeType: Outcome of each animal. Eg: Euthanasia, Adoption
  - OutcomeSubtype: Outcome subtype for each animal. Eg: Suffering, Foster
  - AnimalType: Type of animal-Dog or Cat
  - SexuponOutcome: Sex of the animal along with intactness at the time of. Ex: Neutered Male, Intact Female
  - AgeuponOutcome: Age of the animal at the time of outcome. Ex; 1 year, 3 weeks
  - Breed: Breed of the animal. Ex: Pit Bull Mix, Cairn Terrier
  - Color: Color of the animal. Ex: Tan, White, Black, Brown/White
- **Preprocessing:**
  - DateTime: This attribute is spilt into
    - Year
    - Month
    - DayTime: Morning, Afternoon, Evening, Night
  - SexuponOutcome: This attribute is split into:
    - Sex: Male or Female
    - Intact: Neutered, Spayed, Intact
  - AgeuponOutcome: This attribute is converted into AgeDays which represents age in terms of days. AgeDays is further converted to AgeType which classifies an animal as 'baby' or 'adult' depending on age.
  - Breed: This attribute contains multiple breeds separated by '/' for some instances. The first breed is taken in this case and stored in SimBreed.
  - Color: This attribute contains multiple breeds separated by '/' for some instances. The first color is taken in this case and stored in SimColor.

- **Attributes that will be considered for training**: Year, Month, DayTime, OutcomeType, AnimalType, Sex, Intact, AgeType, Breed, Color
- **Null Values:** Null values in attributes are classified as 'Unknown'

Pre-Processing Results:

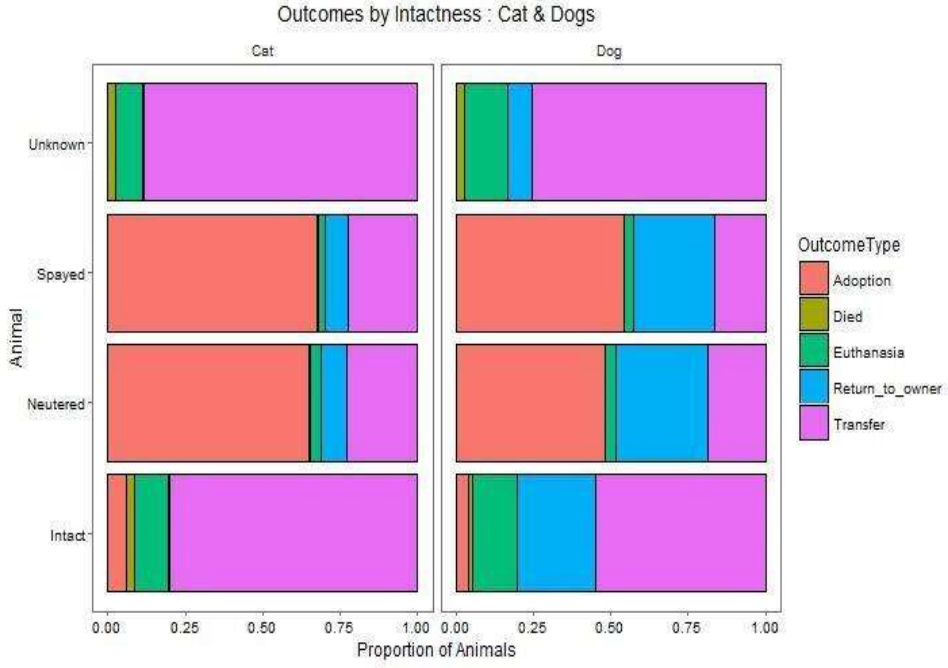
In order to understand the trend of the dataset, we have pre-processed the attributes and visualized the results as follows.

1. **Impact of Animal Type on Outcome Type**



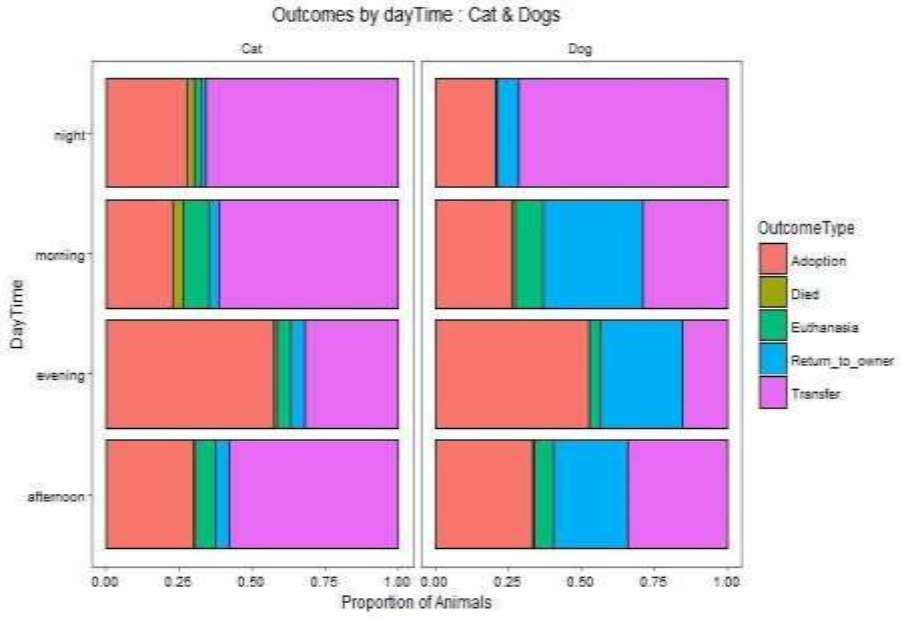
From the plot we can understand that Cats are less adopted than Dogs. Also we can see that Dogs are more likely to be returned to owners than Cats. Cats are more transferred to the shelter center.

2. **Impact of Intactness on Outcome Type**



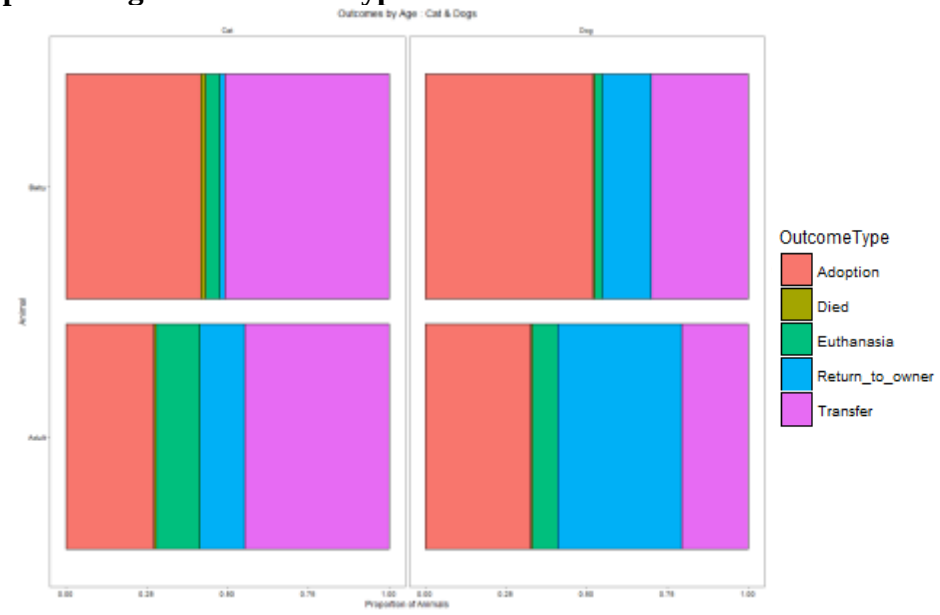
From the graph we can see that adoption rate is more for Spayed and Neutered animals in case of both Cats and Dogs. Intact animals are more likely to be transferred to the Animal Shelter.

3. Impact of Time of Outcome on Outcome Type



We can see that adoption is more during the evening time and Euthanasia rate is more during the morning time compared to the other timings.

4. Impact of Age on Outcome Type



We see that adoptions and transfers are more for a baby animal compared to an adult animal. On the other hand, euthanasia is more prominent for an adult animal compared to a baby animal.

Experimental Methodology:

- Cross validation is used on the training dataset to find the classifier with highest accuracy and also to prevent overfitting.
- The classifier with the highest accuracy is used to predict the outcomes on the test dataset.
- Language Used: R

Results:

- Results of cross validation of different classifiers.

Classifier Used	Accuracy on 3-fold cross Validation
Boosting	0.74
Bagging	0.62
Random Forest	0.671
SVM Linear	0.5672
SVM Radial	0.6547
DeepLearning	0.61
Linear Discriminant Analysis	0.568
KNN	0.823

A number of techniques were applied for all classifiers and the maximum accuracy obtained in each case is reported here. For example, for boosting we used techniques like LogitBoost, AbaBoost, GradientBoost etc.

For SVM we have tried linear kernels as well as non-linear kernels like radial.

Analysis:

We have found out that KNN is having good accuracy on the training dataset. We understood that it is because we have all the attributes as categorical and we have one attribute with 382 levels and one with 57 levels after preprocessing. In this scenario, as we are testing on large dataset containing 26729 rows and we found that KNN performs better with our dataset. Random Forest is having comparatively lesser performance because of more levels for the attributes.
We run KNN on the test dataset and predicted the outcome and wrote output to the .csv files.

Discussion:

From the results we can observe that we have predicted the outcomes as follows:

Output Type	Number instances
Transfer	2645
Adoption	6792

Return_To_owner	1980
Euthansia	40

Tabulating the results for the different Animal Type- Cat and Dog

OutputType	Number of instances for Cat(4800)	Number of instances for Dog(6656)
Transfer	34.79%	14.63%
Adoption	47.70%	67.63%
Return_To_owner	16.66%	17.72%
Euthansia	0.83%	0 %

We see that Adoption and Transfer are the most likely outcomes for cats. Adoption is the prominent outcome for dogs too followed by Return to owner. Euthanasia is least possible outcome for both categories.

Conclusion:

Adoption is the most likely outcome for the animals in the given dataset and Euthanasia is the least likely. Hence, we can see that the predictions are analogous to the real world scenario where Euthanasia is uncommon.

Challenges:

- The main challenge we faced was the running time for each classifiers in normal systems.

References:

The discussion forums for the Kaggle challenge was very much helpful in analyzing the data.
https://www.kaggle.com/c/shelter-animal-outcomes.

Contribution of Team Members:

Team Members:

Febin Steve Jose
Josina Joy
Malini Bhaskaran

- We have divided our work equally among the 3 members. Divided the stages of project as 3 and in each stage we have divided the work.
- For Pre-processing stage we split the attributes among ourselves and did the preprocessing and analysis of impact on the outcome.
- During the training and analysis stage we took 2 classifiers each and tried out a variety of methods and packages to report the accuracy.

- Prediction is done using KNN classifier as it gave the best accuracy.
- Documentation work was shared and each member completed their part.