

# SC1015

## MINI-PROJECT

### A136 | Team 2

Tai Chee Hian

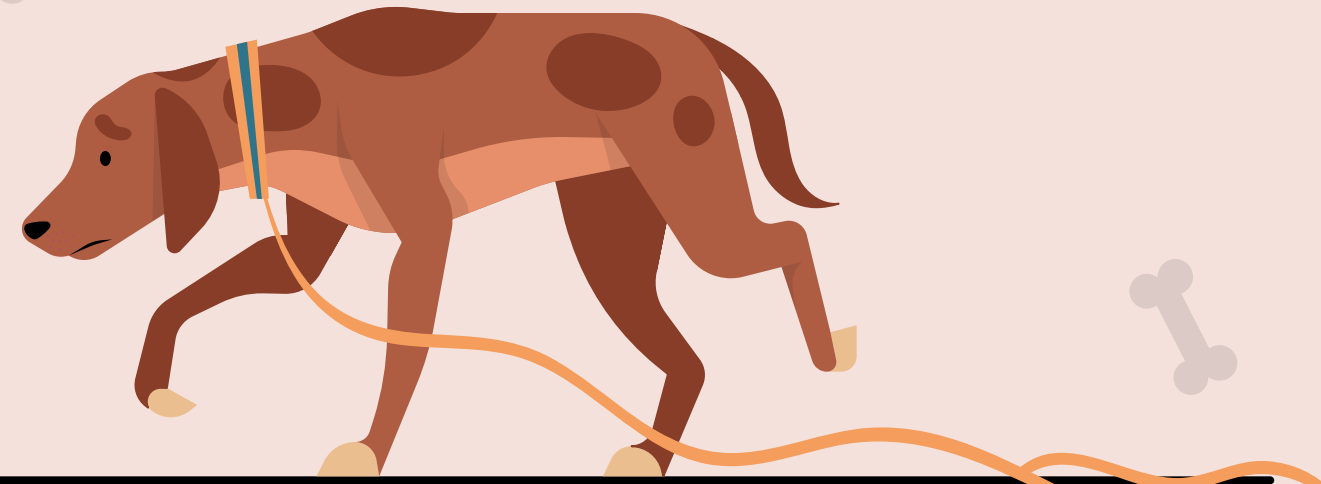
S Dhanusha

Rhea Susan George





**PROBLEM:** How can we predict the outcomes of animals in shelters based on their features?



# TABLE OF CONTENTS

**01** MOTIVATION

**02** EXPLORATORY  
DATA ANALYSIS

1

2

3

4

**03** CORE  
ANALYSIS

**04** CONCLUSION

01

# motivation

Problem definition and dataset  
used



# motivation behind the problem





# 6,300,000



Companion animals enter U.S shelters every year



dataset used:

# Austin Animal Center

- largest no-kill shelter in the U.S
- caring for and sheltering over 18,000 animals per year
- animals of all species who need shelter regardless of their age or condition



Animal ID	Name	DateTime	MonthYear	Found Location	Intake Type	Intake Condition
A786884	*Brock	01/03/2019 04:19:00 PM	01/03/2019 04:19:00 PM	2501 Magin Meadow Dr in Austin (TX)	Stray	Normal
A706918	Belle	07/05/2015 12:59:00 PM	07/05/2015 12:59:00 PM	9409 Bluegrass Dr in Austin (TX)	Stray	Normal
A724273	Runster	04/14/2016 06:43:00 PM	04/14/2016 06:43:00 PM	2818 Palomino Trail in Austin (TX)	Stray	Normal
A665644		10/21/2013 07:59:00 AM	10/21/2013 07:59:00 AM	Austin (TX)	Stray	Sick
A682524	Rio	06/29/2014 10:38:00 AM	06/29/2014 10:38:00 AM	800 Grove Blvd in Austin (TX)	Stray	Normal
A743852	Odin	02/18/2017 12:46:00 PM	02/18/2017 12:46:00 PM	Austin (TX)	Owner Surrender	Normal
A635072	Beowulf	04/16/2019 09:53:00 AM	04/16/2019 09:53:00 AM	415 East Mary Street in Austin (TX)	Public Assist	Normal
A708452	Mumble	07/30/2015 02:37:00 PM	07/30/2015 02:37:00 PM	Austin (TX)	Public Assist	Normal
A818975		06/18/2020 02:53:00 PM	06/18/2020 02:53:00 PM	Braker Lane And Metric in Travis (TX)	Stray	Normal
A774147		06/11/2018 07:45:00 AM	06/11/2018 07:45:00 AM	6600 Elm Creek in Austin (TX)	Stray	Injured
A731435	*Casey	08/08/2016 05:52:00 PM	08/08/2016 05:52:00 PM	Austin (TX)	Owner Surrender	Normal
A760053		10/11/2017 03:46:00 PM	10/11/2017 03:46:00 PM	8800 South First Street in Austin (TX)	Stray	Normal
A707375	*Candy Cane	07/11/2015 06:19:00 PM	07/11/2015 06:19:00 PM	Galilee Court And Damita Jo Dr in Manor (TX)	Stray	Normal
A696408	*Pearl	02/04/2015 12:58:00 PM	02/04/2015 12:58:00 PM	9705 Thaxton in Austin (TX)	Stray	Normal
A790209	Ziggy	03/06/2019 02:31:00 PM	03/06/2019 02:31:00 PM	4424 S Mopac Expwy in Austin (TX)	Public Assist	Normal
A743114		02/04/2017 10:10:00 AM	02/04/2017 10:10:00 AM	208 Beaver St in Austin (TX)	Stray	Injured

snippet of dataset used



# 02 exploratory data analysis

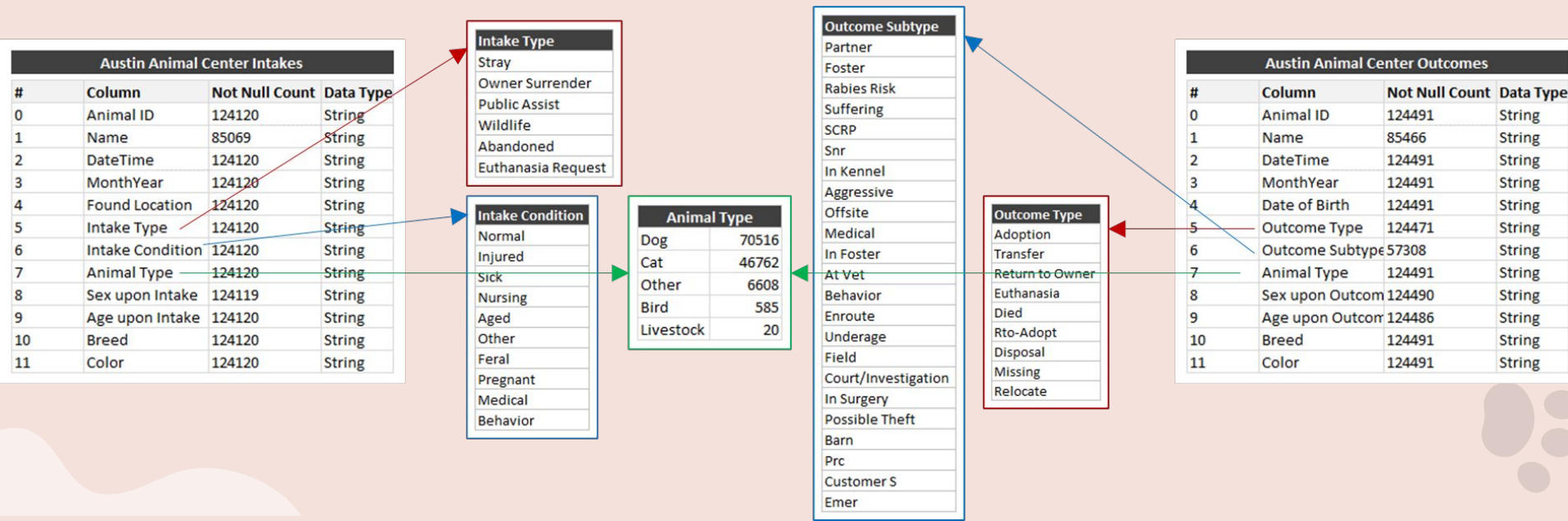
Data preparation and visualisation



# Format of Data

The data is in two separate files, **Austin\_Animal\_Center\_Intakes.csv** and **Austin\_Animal\_Center\_Outcomes.csv**

Contained in the two files is the following data:



# Data Cleaning

1. Standardise column names
  - Replace spaces with underscores; capitalize
2. Convert datetime columns from string to datetime
3. Convert AGE\_INTAKE and AGE\_OUTCOME from string to integer

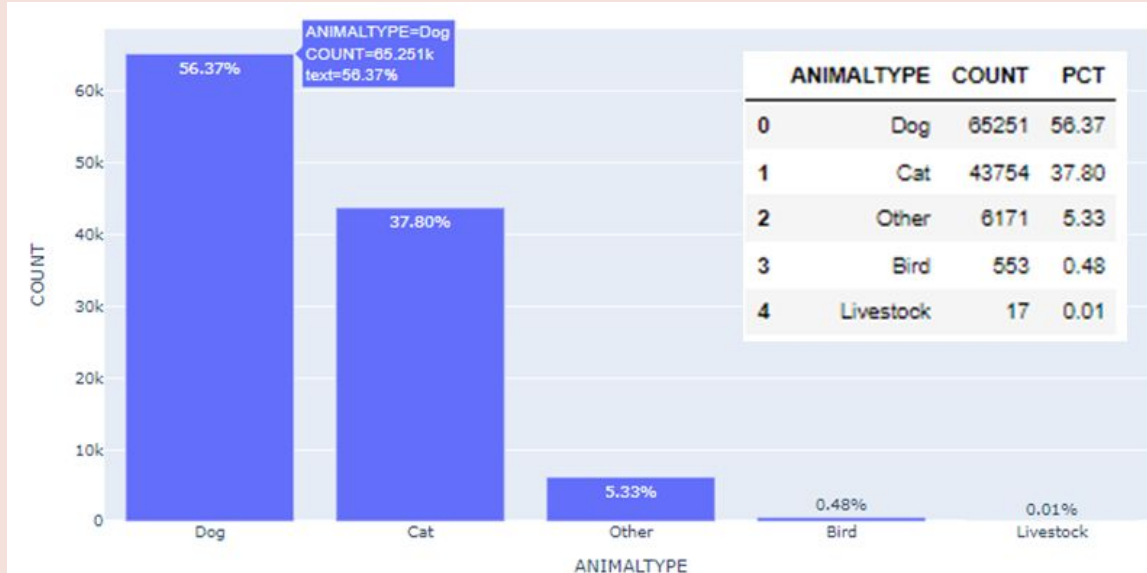
- Used months as the unit

```
def ConvertAgeAsMonths(AgeString):  
    if not pd.isnull(AgeString):  
        AgeSignatureDictionary = {"YEAR":12, "MONTH":1,"DAY":1/30,"WEEK":1/4.15}  
        AgePattern = "|".join([i for i in AgeSignatureDictionary.keys()])  
        for pattern in AgeSignatureDictionary.keys():  
            AgeValue = None  
            if len(re.findall(pattern,AgeString.upper()))>0:  
                AgeValue = (re.split(pattern,AgeString.upper())[0].strip())  
                AgeValue = np.float16(AgeValue)  
                AgeValue = AgeValue*AgeSignatureDictionary.get(pattern)  
                AgeValue = None if AgeValue<0 else np.round(AgeValue,3)  
            return AgeValue  
    return None
```

ANIMALID	DATETIME_INTAKE	LOCATION	INTAKE TYPE	INTAKE CONDITION	ANIMAL TYPE	SEX/INTAKE	AGE/INTAKE	BREED	COLOR	PREV/HIST	DATETIME_OUTCOME	DOB	OUTCOMETYPE	OUTCOMESUBTYPE	SEXOUTCOME	AGEOUTCOME	DAYS_STAY
A006100	7/3/2014 2:26	8700 Research in Austin (TX)	Public Assist	Normal	Dog	Neutered Male	72	Spinone Italiano Mix	Yellow/White	0	8/3/2014 5:10	7/9/2007	Return to Owner		Neutered Male	72	1
A006100	19/12/2014 10:21	8700 Research Blvd in Austin (TX)	Public Assist	Normal	Dog	Neutered Male	84	Spinone Italiano Mix	Yellow/White	1	20/12/2014 4:35	7/9/2007	Return to Owner		Neutered Male	84	0
A006100	7/12/2017 2:07	Colony Creek And Hunters Trace in Austin (TX)	Stray	Normal	Dog	Neutered Male	120	Spinone Italiano Mix	Yellow/White	2	7/12/2017 12:00	7/9/2007	Return to Owner		Neutered Male	120	0
A047759	2/4/2014 3:55	Austin (TX)	Owner Surrender	Normal	Dog	Neutered Male	120	Dachshund	Tricolor	0	7/4/2014 3:12	4/2/2004	Transfer	Partner	Neutered Male	120	4
A134067	16/11/2013 9:02	12034 Research Blvd in Austin (TX)	Public Assist	Injured	Dog	Neutered Male	192	Shetland Sheepdog	Brown/White	0	16/11/2013 11:54	10/16/1997	Return to Owner		Neutered Male	192	0


# Restricting Animal Type

- Impossible to make comparison on features such as breed and age between different animal types
- Dogs make up majority of the dataset (56.37%)
- **Restrict animal type to dogs**





# Feature Engineering

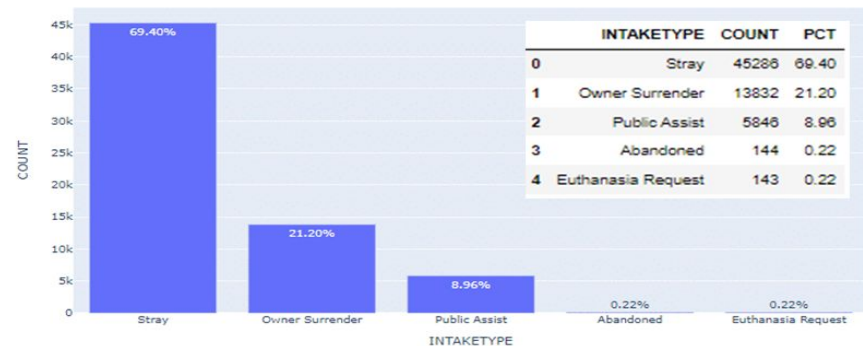
- Transform data from **qualitative to quantitative** values to make it easier to perform classification
  - The features are broken down to multiple **one-hot vectors**
  - i.e. the columns are split such that each property is indicated with a true / false (1 / 0) value
  - Performed on BREED, SEXINTAKE, INTAKETYPE, INTAKECONDITION
  - COLOR has too many values to be split into one-hot vectors
- 

# Exploring Other Features

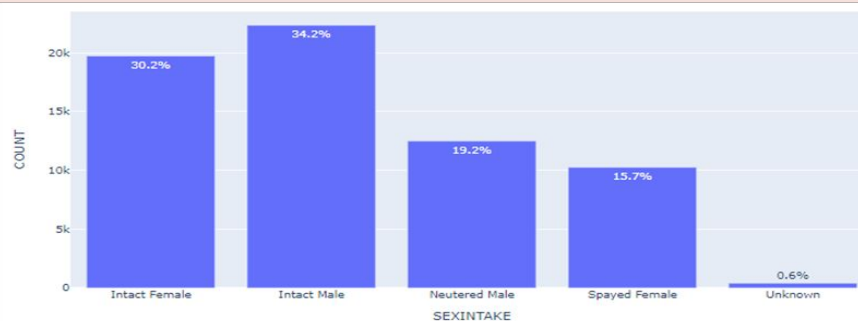
## ANIMAL TYPE of all Animals



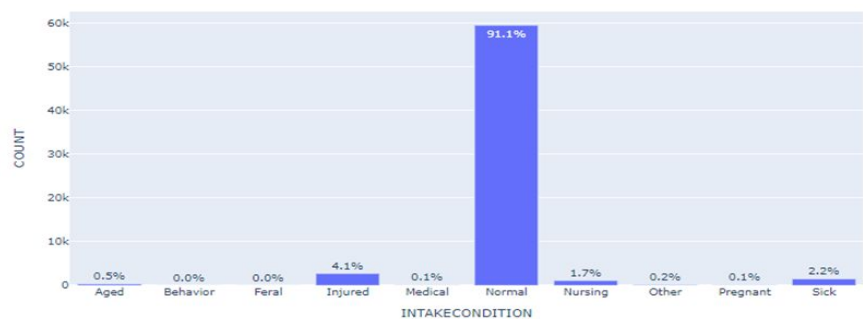
## INTAKE TYPE of Dogs



## SEX of Dogs

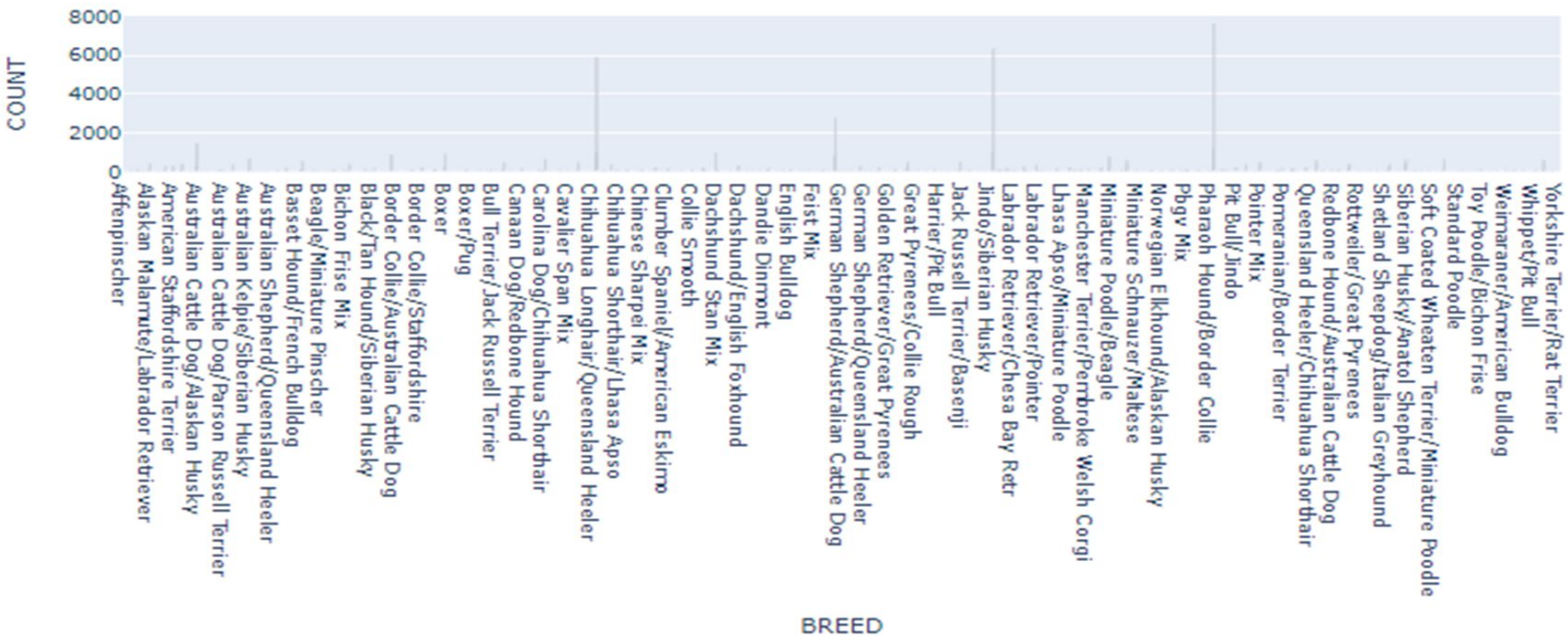


## INTAKE CONDITION of Dogs



# Exploring Other Features

BREED of Dogs





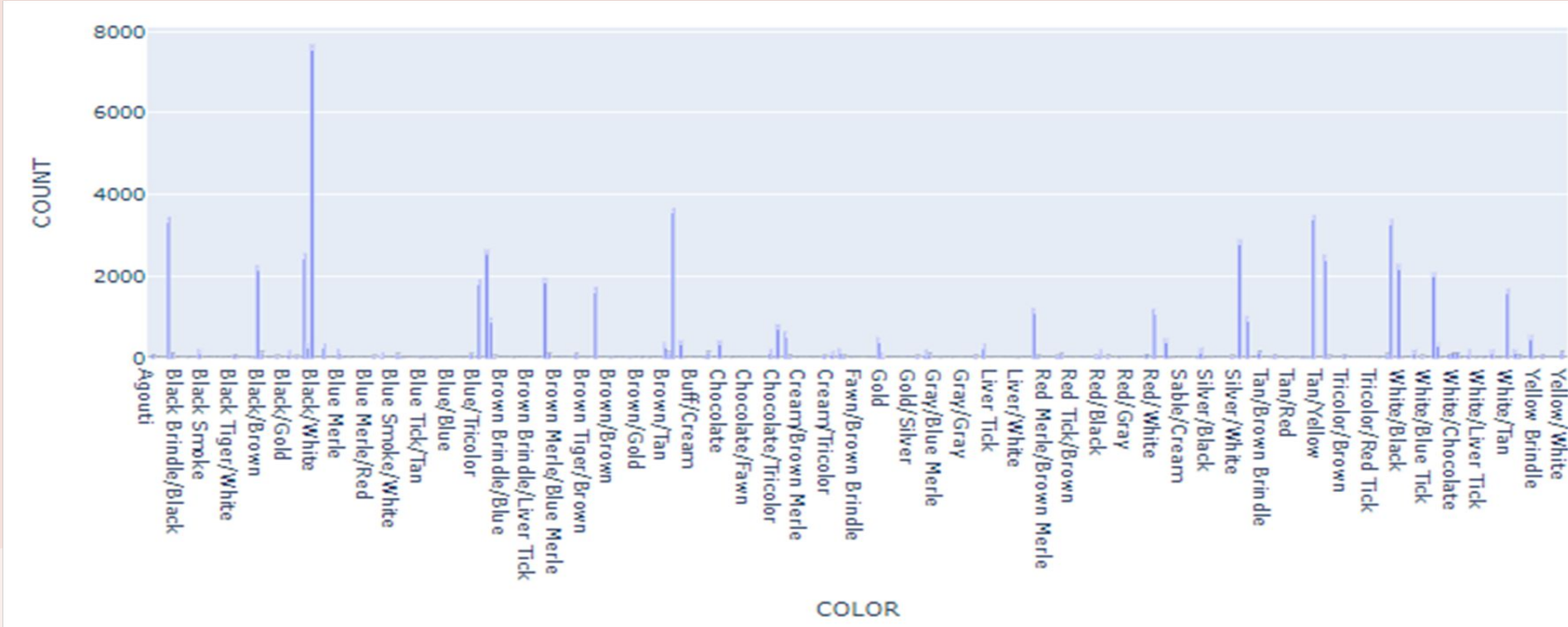


# Feature Engineering

- We can create a **new feature** to summarise the information under the BREED column
- A good way to differentiate dog breeds is to categorise them as purebred or mixed breed
  - Binary, can be represented with true / false
- Introduce new column **IS\_PUREBREED** that indicates whether or not the dog is purebred



**COLOR of Dogs**



# Feature Engineering

Breakdown of BREED, SEXINTAKE, INTAKETYPE, INTAKECONDITION

	Y	X																								
	OUTCOMETYPE	AGEOUTCOME	PREVHIST	DAYS_STAY	IS_PUREBREED	INTAKETYPE_ABANDONED	INTAKETYPE_EUTHANASIAREQUEST	INTAKETYPE_OWNERSURRENDER	INTAKETYPE_PUBLICASSIST	INTAKETYPE_STRAY	INTAKECONDITION_AGED	INTAKECONDITION_BEHAVIOR	INTAKECONDITION_FERAL	INTAKECONDITION_INJURED	INTAKECONDITION_MEDICAL	INTAKECONDITION_NORMAL	INTAKECONDITION_NURSING	INTAKECONDITION_OTHER	INTAKECONDITION_PREGNANT	INTAKECONDITION_SICK	SEXINTAKE_INTACTFEMALE	SEXINTAKE_INTACTMALE	SEXINTAKE_NEUTEREDMALE	SEXINTAKE_SPAYEDFEMALE	SEXINTAKE_UNKNOWN	
ANIMALID_TIMESTAMP																										
A006100_1394159160	RTO	72	0	1	1	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	
A006100_1418984460	RTO	84	1	0	1	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	
A006100_1512612420	RTO	120	2	0	1	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	
A134067_1384592520	RTO	192	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	1	0	0	
A141142_1384569960	RTO	180	0	1	1	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	
A163459_1415934660	RTO	180	0	0	1	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	
A178569_1395049500	RTO	180	0	5	1	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	
A189592_1442555160	RTO	216	0	0	1	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0	
A200922_1380772020	ADOPT	192	0	50	1	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	
A208755_1424135820	DIE	168	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0	
***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	

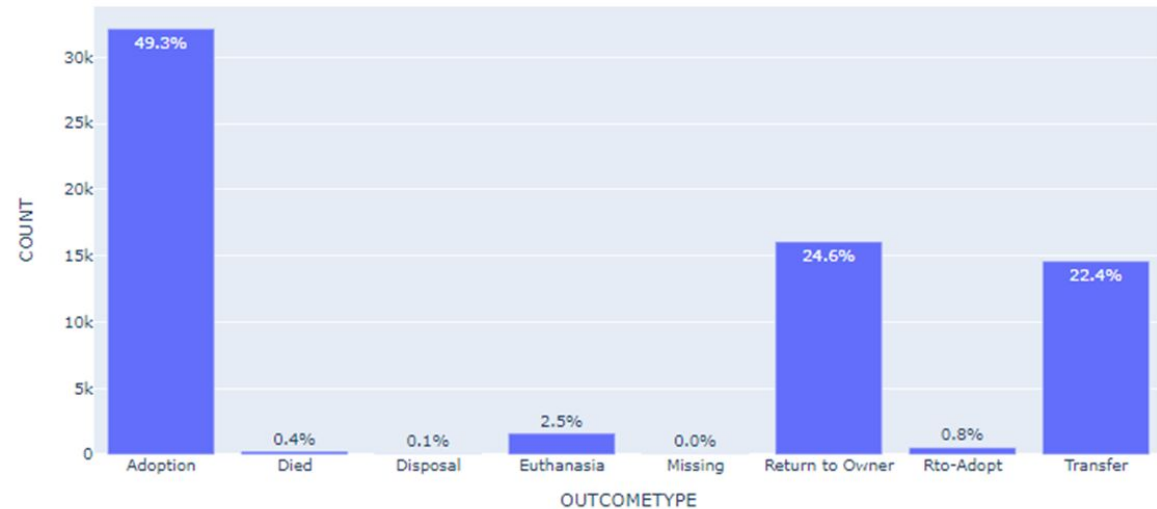
This data frame is stored as feature.csv

X - Features

Y - Target

# Redefining Outcomes

- Huge class imbalance
- Similar outcome types are grouped together
- Outcome types that are too few or ambiguous are dropped

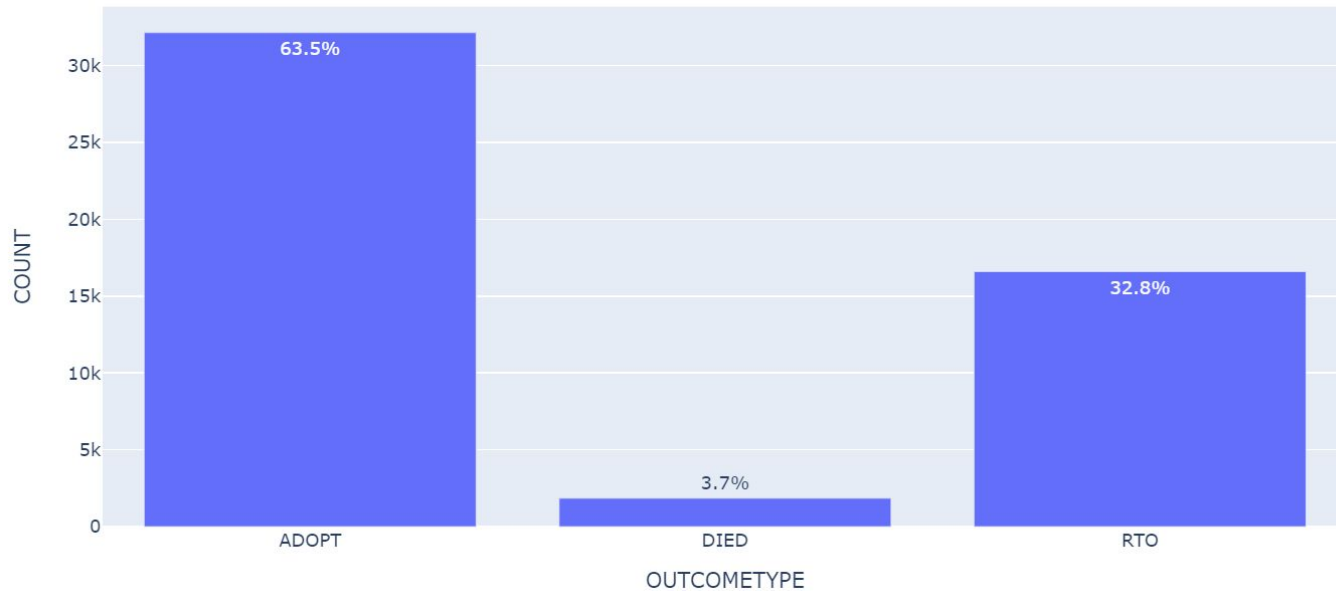


Outcome Type	PREDICTION CLASS
Adoption	ADOPT
Transfer	
Return to Owner	RTO
Euthanasia	DIE
Died	DIE
Rto-Adopt	RTO
Disposal	
Missing	
Relocate	

These are too few and ambiguous so we can safely remove these records

# Redefining Outcomes

OUTCOMETYPE Distribution [ Animal Type : Dog (#50580)]



03

# core analysis

machine learning and  
regression



# machine learning




## regression

- finding the correlation between outcomes of dogs and the intakes

## classification

- predicting the outcomes of dogs based on their features

## objective

- Shelters can predict outcomes
  - Aid resource management and allocation
- 

# machine learning tools & techniques

- a. Random forest
- b. Random forest after feature fine tuning:  
RandomForestClassifier
- c. Gradient boosting
- d. Support Vector Machines model
- e. Ensemble method





# a. RandomForestClassifier

-----  
Confusion Matrix:

-----  
[[6031 78 243]  
[ 156 115 86]  
[ 587 46 2774]]  
-----

-----  
Classification Report:

-----  
precision recall f1-score support  
  
ADOPT 0.89 0.95 0.92 6352  
DIE 0.48 0.32 0.39 357  
RTO 0.89 0.81 0.85 3407  
  
accuracy 0.88 10116  
macro avg 0.76 0.70 0.72 10116  
weighted avg 0.88 0.88 0.88 10116  
-----

#####  
FEATURE RANKING USING IMPORTANCE SCORE  
#####  
RANK IMPORTANCE FEATURE  
-----  
1. (0.51) AGEOUTCOME  
2. (0.14) PREVHIST  
3. (0.11) DAYS\_STAY  
4. (0.07) IS\_PUREBREED  
5. (0.06) INTAKETYPE\_ABANDONED  
6. (0.02) INTAKETYPE\_EUTHANSIAREQUEST  
7. (0.01) INTAKETYPE\_OWNERSURRENDER  
8. (0.01) INTAKETYPE\_PUBLICASSIST  
9. (0.01) INTAKETYPE\_STRAY  
10. (0.01) INTAKECONDITION\_AGED  
11. (0.01) INTAKECONDITION\_BEHAVIOR  
12. (0.01) INTAKECONDITION\_FERAL  
13. (0.01) INTAKECONDITION\_INJURED  
14. (0.01) INTAKECONDITION\_MEDICAL  
15. (0.01) INTAKECONDITION\_NORMAL  
16. (0.00) INTAKECONDITION\_NURSING  
17. (0.00) INTAKECONDITION\_OTHER  
18. (0.00) INTAKECONDITION\_PREGNANT  
19. (0.00) INTAKECONDITION\_SICK  
20. (0.00) SEXINTAKE\_INTACTFEMALE  
21. (0.00) SEXINTAKE\_INTACTMALE  
22. (0.00) SEXINTAKE\_NEUTEREDMALE  
23. (0.00) SEXINTAKE\_SPAYEDFEMALE  
24. (0.00) SEXINTAKE\_UNKNOWN  
#####

- Features highlighted are not necessary for the prediction models

## b. RandomForestClassifier after Feature Fine Tuning

-----  
Confusion Matrix:

-----  
[[ 6113 47 192]  
 [ 193 48 116]  
 [ 625 16 2766]]

-----  
Classification Report:

-----  
                  precision    recall  f1-score    support  
  
      ADOPT          0.88      0.96      0.92      6352  
      DIE           0.43      0.13      0.21       357  
      RTO           0.90      0.81      0.85      3407  
  
      accuracy                  0.88      10116  
      macro avg          0.74      0.64      0.66      10116  
      weighted avg      0.87      0.88      0.87      10116  
-----

```
#####  
FEATURE RANKING USING IMPORTANCE SCORE  
#####  
RANK      IMPORTANCE      FEATURE  
-----  
1.         (0.54)         AGEOUTCOME  
2.         (0.15)         PREVHIST  
3.         (0.14)         DAYS_STAY  
4.         (0.08)         IS_PUREBREED  
5.         (0.06)         INTAKETYPE_ABANDONED  
6.         (0.02)         INTAKETYPE_EUTHANASIAREQUEST  
7.         (0.01)         INTAKETYPE_OWNERSURRENDER  
8.         (0.00)         INTAKETYPE_PUBLICASSIST  
9.         (0.00)         INTAKETYPE_STRAY  
#####
```

- Unnecessary features have been removed

# c. Gradient Boosting (GBC)

-----  
Confusion Matrix:

-----  
[[6193 15 144]  
 [ 211 44 102]  
 [ 637 7 2763]]

-----  
Classification Report:

-----

	precision	recall	f1-score	support
ADOPT	0.88	0.97	0.92	6352
DIE	0.67	0.12	0.21	357
RTO	0.92	0.81	0.86	3407
accuracy			0.89	10116
macro avg	0.82	0.64	0.66	10116
weighted avg	0.89	0.89	0.88	10116

-----

- model is trained to minimize the loss function
- computing the negative gradient of the loss function with respect to the predicted output
- accuracy: 0.89

## d. Support Vector Machines (SVM)

-----  
Confusion Matrix:  
-----

```
[[6071    0  281]
 [ 219   20  118]
 [ 874    1 2532]]
```

-----  
Classification Report:  
-----

	precision	recall	f1-score	support
ADOPT	0.85	0.96	0.90	6352
DIE	0.95	0.06	0.11	357
RTO	0.86	0.74	0.80	3407
accuracy			0.85	10116
macro avg	0.89	0.58	0.60	10116
weighted avg	0.86	0.85	0.84	10116

- Effective in high dimensional spaces
- Does not perform well when
  - Data set is large
  - Target classes are overlapping
- required training time is higher
- accuracy: 85%

# e. Ensemble Model

-----  
Confusion Matrix:

-----  
[[6209 0 143]  
 [ 236 20 101]  
 [ 647 1 2759]]

-----  
Classification Report:




-----  
precision recall f1-score support

ADOPT	0.88	0.98	0.92	6352
DIE	0.95	0.06	0.11	357
RTO	0.92	0.81	0.86	3407
accuracy			0.89	10116
macro avg	0.92	0.61	0.63	10116
weighted avg	0.89	0.89	0.87	10116

- ensemble method combines all three and the final performance
- precision had improved especially for DIED outcome
- accuracy: 89%



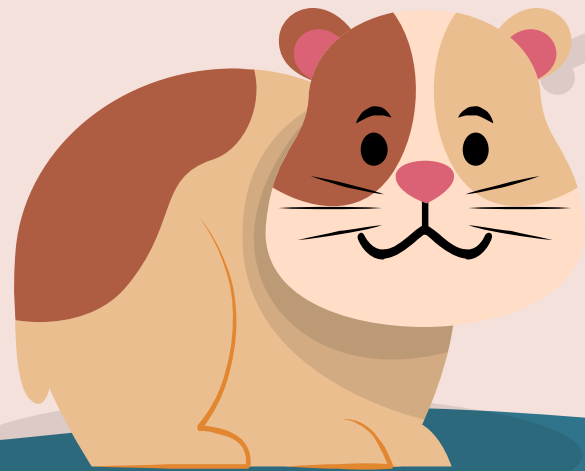
# Model Comparison

- Gradient Boosting Algorithm is the most efficient model
  - Accuracy of 89% and consistent in performance
  - SVM is the worst model
  - Ensemble Model Precision improved dramatically
- 
- 
- 

04

# conclusion

outcomes and insights



# outcomes



## MOST IMPORTANT FEATURES:



1

Age at  
outcome



2

Previous  
history



3

Days stayed at shelter

```
#####  
FEATURE RANKING USING IMPORTANCE SCORE  
#####  
RANK      IMPORTANCE      FEATURE  
-----  
1.         (0.54)         AGEOUTCOME  
2.         (0.15)         PREVHIST  
3.         (0.14)         DAYS_STAY  
4.         (0.08)         IS_PUREBREED  
5.         (0.06)         INTAKETYPE_ABANDONED  
6.         (0.02)         INTAKETYPE_EUTHANASIAREQUEST  
7.         (0.01)         INTAKETYPE_OWNERSURRENDER  
8.         (0.00)         INTAKETYPE_PUBLICASSIST  
9.         (0.00)         INTAKETYPE_STRAY  
#####
```





# insights + recommendations

## insights

- Shelters can note the most important features for better prediction of outcomes
- Outcomes of animals also depend on extrinsic factors such as the shelter's resources

## recommendations

- not rely purely on prediction models to predict deaths of animals
- Push for adoption of animals at a younger age to maximise adoption

# CONCLUSION

## DATA COLLECTION + EDA

- consolidate data set into a clearer and visible format
- Restricting analysis of animals to dogs
- Feature engineering

## MACHINE LEARNING

- Random forest
- Gradient boosting
- Gradient Boosting algorithm was the best ML → highest accuracy of 89%
- SVM model
- Ensemble model

## PROJECT OUTCOMES

- Improve wellbeing of dogs
- Efficient resource allocation
- showing off the dogs that are likely to be adopted
- reduce euthanization rates
- increase rates of adoption

# THANK YOU!

CREDITS: This presentation template was created by [Slidesgo](#), including icons by [Flaticon](#) and infographics & images by [Freepik](#)

