Dataset Size: 10290 rows & 24 columns

LKH'S DATASET PROPOSAL FOR CAPSTONE

https://www.kaggle.com/jinbonnie/animal-data

Predict how likely it is for an animal to be adopted

LABEL?

Multi-class classification:

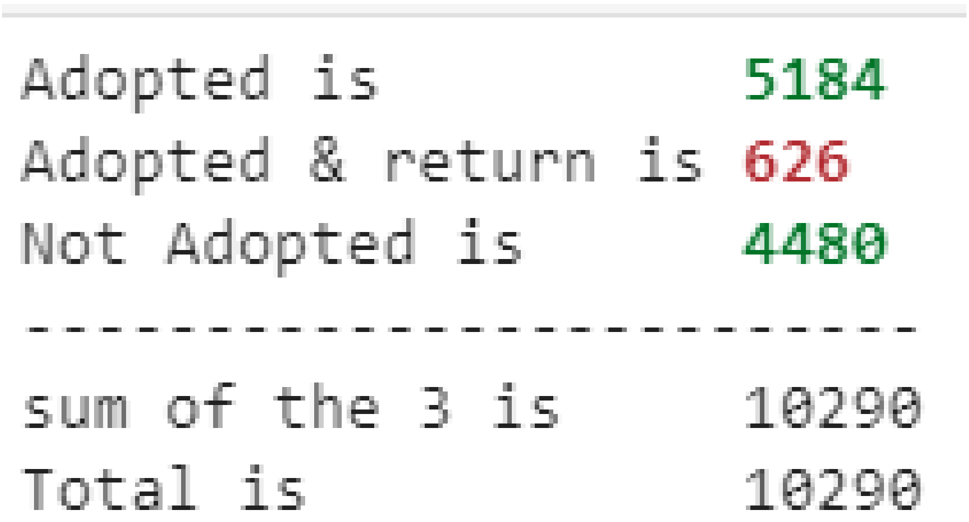
- Adopted,
- adopted & returned
- not adopted

Or... just Binary class:

- Adopted vs not adopted

Will likely do multiclass as I've not explored it before!

LABEL



```
Adopted is            5184
Adopted & return is 626
Not Adopted is        4480
-------------------------
sum of the 3 is       10290
Total is              10290
```

- Total is   10290

- Smote / use class_weight = balanced

COLUMNS

- - id:
- - ID of animal in animal shelter system (Use for initial filtering b4 dropping)
- - intakedate:
- - the date he/she has been taken by the shelter (Easily filter by year etc.) and create new columns
- - intakereason
- - reason for adopting EG. 'Moving' 'Abandoned' 'Incompatible with owner lifestyle' -> One hot for 25 values
- - istransfer
- - has animal been transferred EG. [0 1] one hot it
- - sheltercode
- - the identify code of the shelter ->one hot encode?
- - identichipnumber
- - the micro-chip id of the pet -> Binary, one hot (Chipped / not chipped)
- - animalname
- - animal's name -> considering dropping (no real value) since theyre too unique with 4336 values
- - breedname
- - the breed of the animal -> I can try one hot encoding 799 breeds... we will see
- - basecolour
- - the color of the animal -> can try one hot encoding for 78 colours
- - speciesname
- - Animal Species name -> choosing only cat & dog due to extremely high imbalance
- - animalage
- - Age -> One hot encode 273 values
- - sexname
- - Binary M/F -> One hot encode

COLUMNS

- location
  - section of the shelter -> consider dropping or one hot for 39 values
- movementdate
  - the date they have been moved -> in take to
- movementtype
  - ['Adoption' 'Foster' 'Transfer' 'Reclaimed' 'Released To Wild' 'Stolen' 'Escaped'] -> Where I got my label
- istrial
  - is that trial or confirm change -> Dropping since all are 0s
- returndate
  - Binary 0/1 if animal has been returned -> Where I got my label
- returnedreason
  - why they were been returned -> Drop too hard to make use of the data
- deceaseddate
  - date of passing -> Label encode making not deceased 1 and deceased 0
- deceasedreason
  - reason for passing -> Drop too hard to make use of the data
- diedoffshelter
  - Binary 0/1 death in shelter -> one hot
- puttosleep
  - Binary 0/1 whether put to sleep -> label 0 = put to sleep, 1= sleep
- isdoa
  - Binary 0/1 dead on arrival -> Label 0 = DOA, 1 = alive

COLUMNS

In summary:

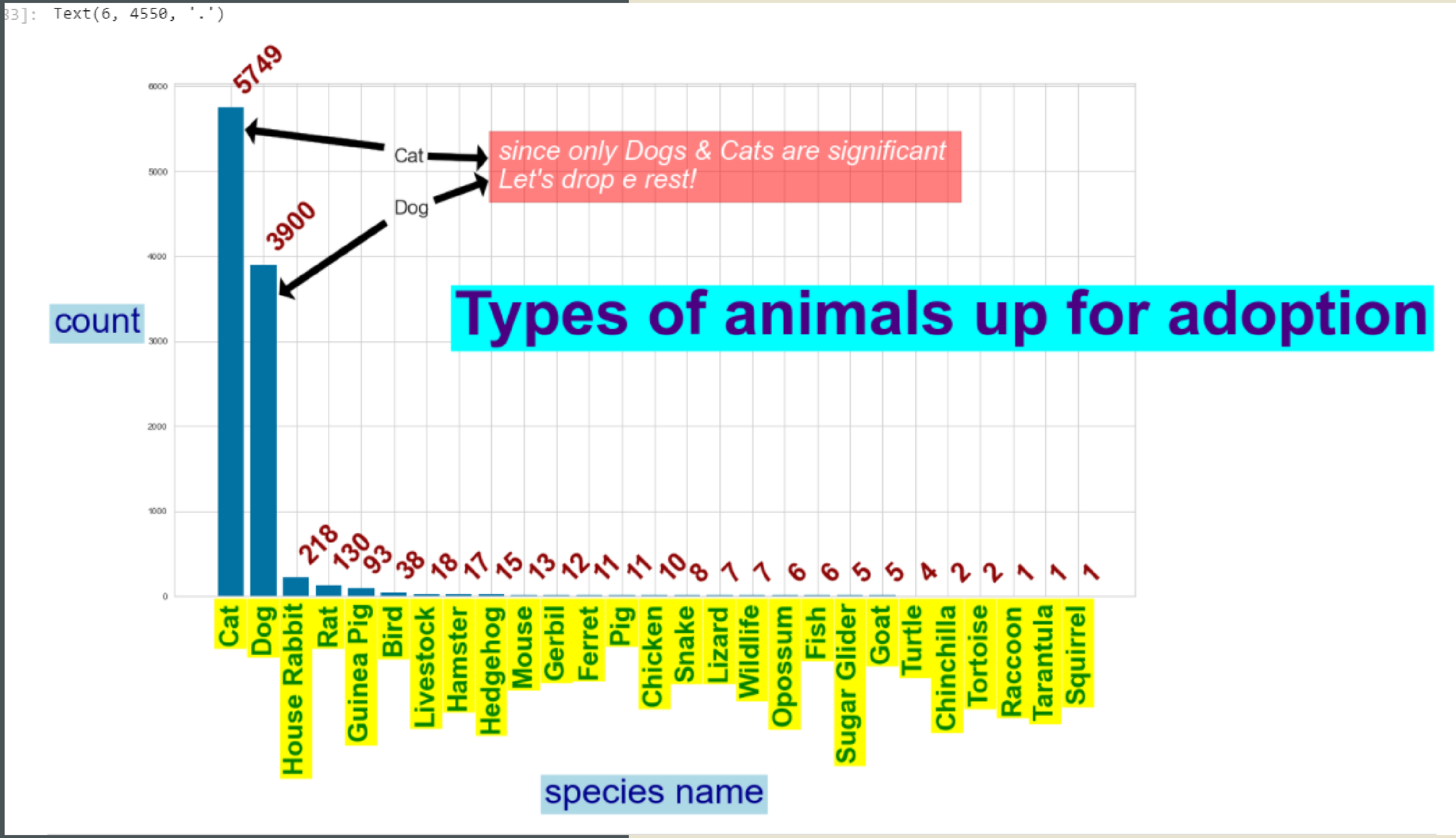
Highly Cardinal/unique features = drop

Categorical data = one hot

Even binary data = one hot

Ordinal data = label encode

SOME EXAMPLES OF AN IMBALANCED FEATURE

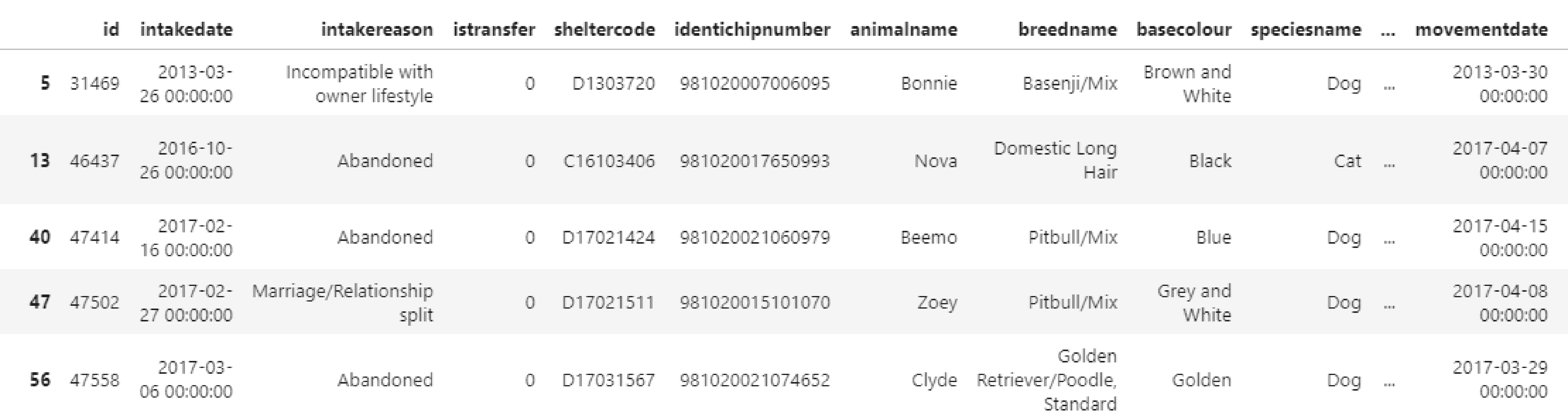


```
33]:  Text(6, 4550, '.')
```

Types of animals up for adoption

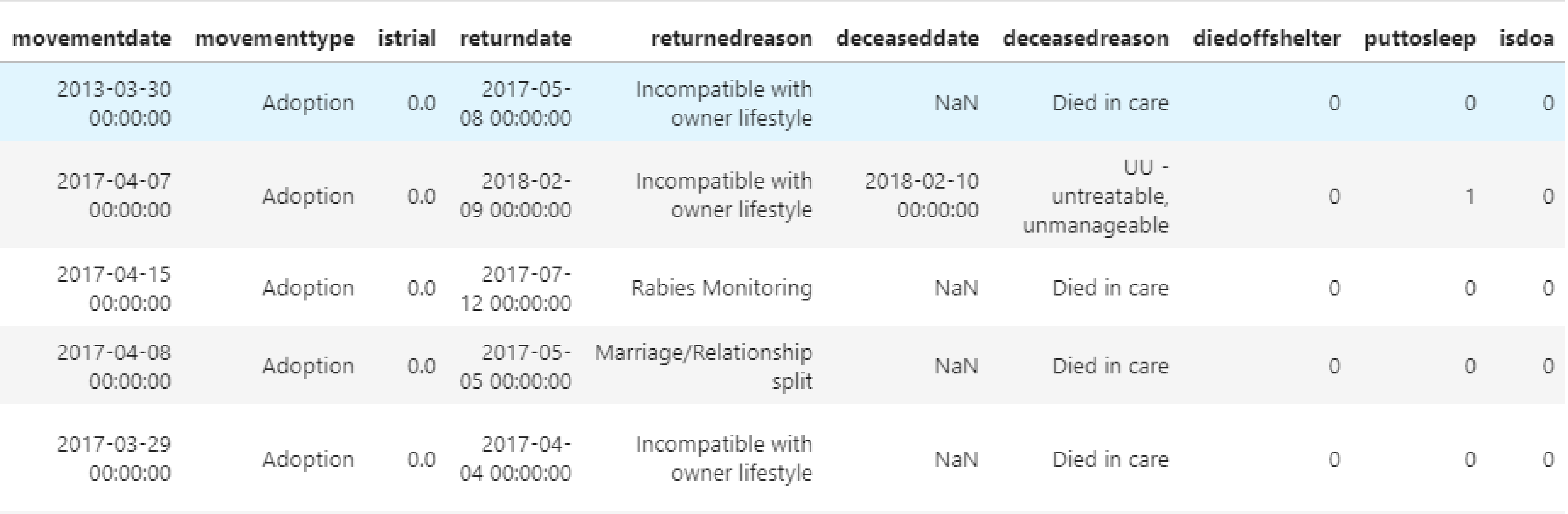
since only Dogs & Cats are significant
Let's drop e rest!

SMALL SAMPLE (HEAD 5)



	id	intakedate	intakereason	istransfer	sheltercode	identichipnumber	animalname	breedname	basecolour	speciesname	...	movementdate
5	31469	2013-03-26 00:00:00	Incompatible with owner lifestyle	0	D1303720	981020007006095	Bonnie	Basenji/Mix	Brown and White	Dog	...	2013-03-30 00:00:00
13	46437	2016-10-26 00:00:00	Abandoned	0	C16103406	981020017650993	Nova	Domestic Long Hair	Black	Cat	...	2017-04-07 00:00:00
40	47414	2017-02-16 00:00:00	Abandoned	0	D17021424	981020021060979	Beemo	Pitbull/Mix	Blue	Dog	...	2017-04-15 00:00:00
47	47502	2017-02-27 00:00:00	Marriage/Relationship split	0	D17021511	981020015101070	Zoey	Pitbull/Mix	Grey and White	Dog	...	2017-04-08 00:00:00
56	47558	2017-03-06 00:00:00	Abandoned	0	D17031567	981020021074652	Clyde	Golden Retriever/Poodle, Standard	Golden	Dog	...	2017-03-29 00:00:00

SMALL SAMPLE (HEAD 5)

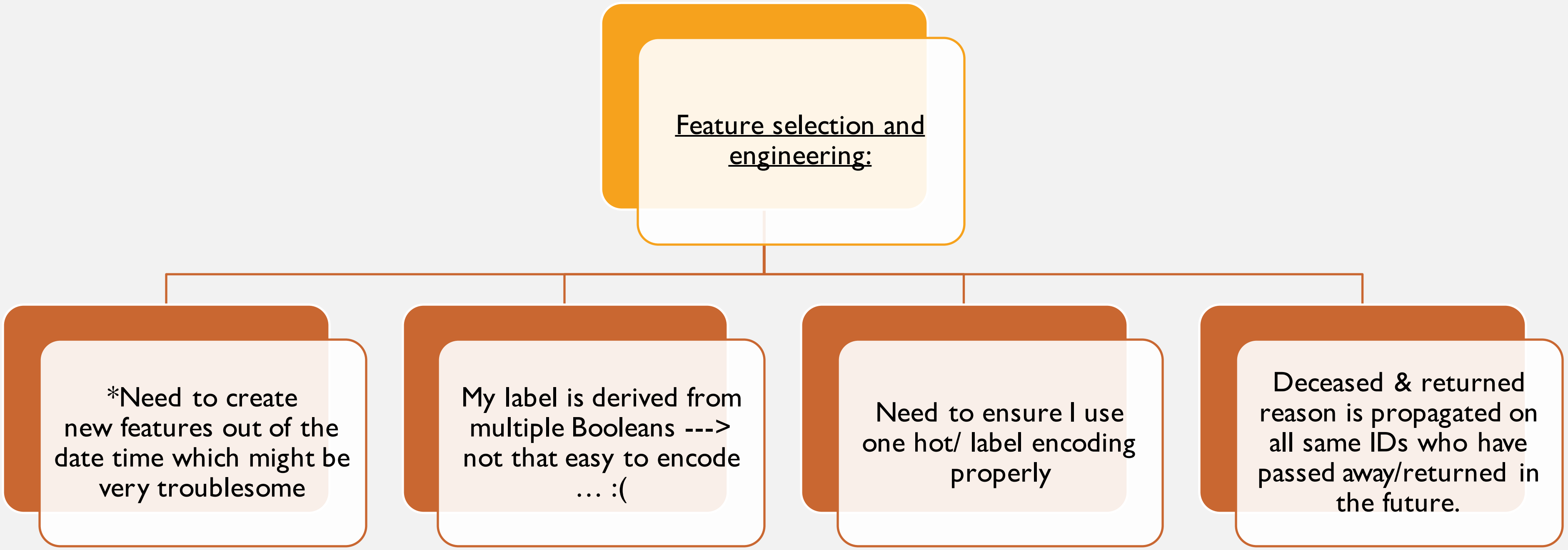


movementdate	movementtype	istrial	returndate	returnedreason	deceaseddate	deceasedreason	diedoffshelter	puttosleep	isdoa
2013-03-30 00:00:00	Adoption	0.0	2017-05-08 00:00:00	Incompatible with owner lifestyle	NaN	Died in care	0	0	0
2017-04-07 00:00:00	Adoption	0.0	2018-02-09 00:00:00	Incompatible with owner lifestyle	2018-02-10 00:00:00	UU - untreatable, unmanageable	0	1	0
2017-04-15 00:00:00	Adoption	0.0	2017-07-12 00:00:00	Rabies Monitoring	NaN	Died in care	0	0	0
2017-04-08 00:00:00	Adoption	0.0	2017-05-05 00:00:00	Marriage/Relationship split	NaN	Died in care	0	0	0
2017-03-29 00:00:00	Adoption	0.0	2017-04-04 00:00:00	Incompatible with owner lifestyle	NaN	Died in care	0	0	0

CHALLENGES

- Many many categorical features -> Need one hot creating even more columns
- -> Might have too many columns -> curse of dimensionality
- HOW TO SOLVE:
- **1) Regularization and Sparsity**
- If supported by the model, I would recommend L1 or ElasticNet regularization to zero-out some features.
- **2) Feature Selection**
- We could try various different feature selection algorithms (e.g., selecting by variance or by greedy search: sequential backward/forward selection, genetic algorithms, etc.)
- 3) Adding dropout layers

CHALLENGES



Feature selection and engineering:

*Need to create new features out of the date time which might be very troublesome

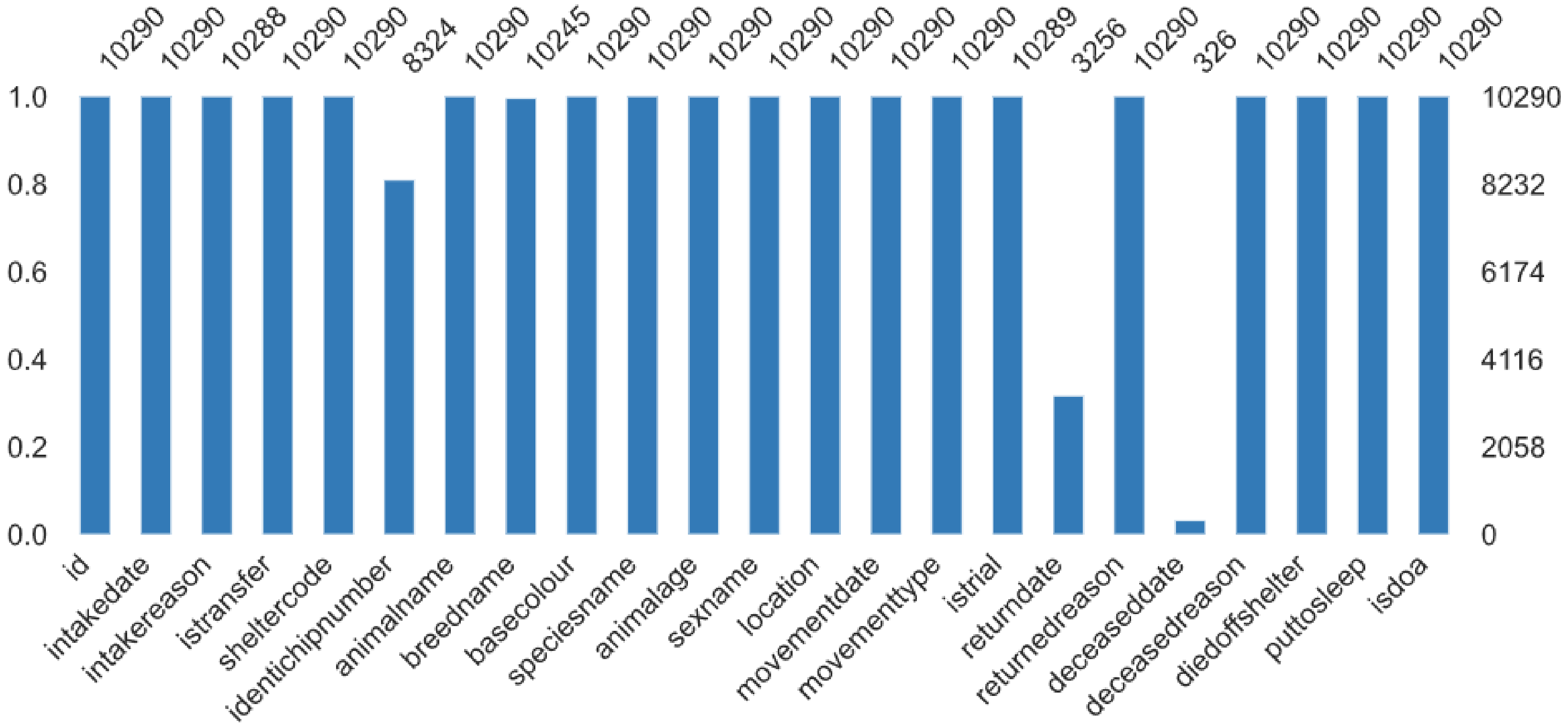
My label is derived from multiple Booleans ---> not that easy to encode … :(

Need to ensure I use one hot/ label encoding properly

Deceased & returned reason is propagated on all same IDs who have passed away/returned in the future.

CHALLENGES

Missing data plot



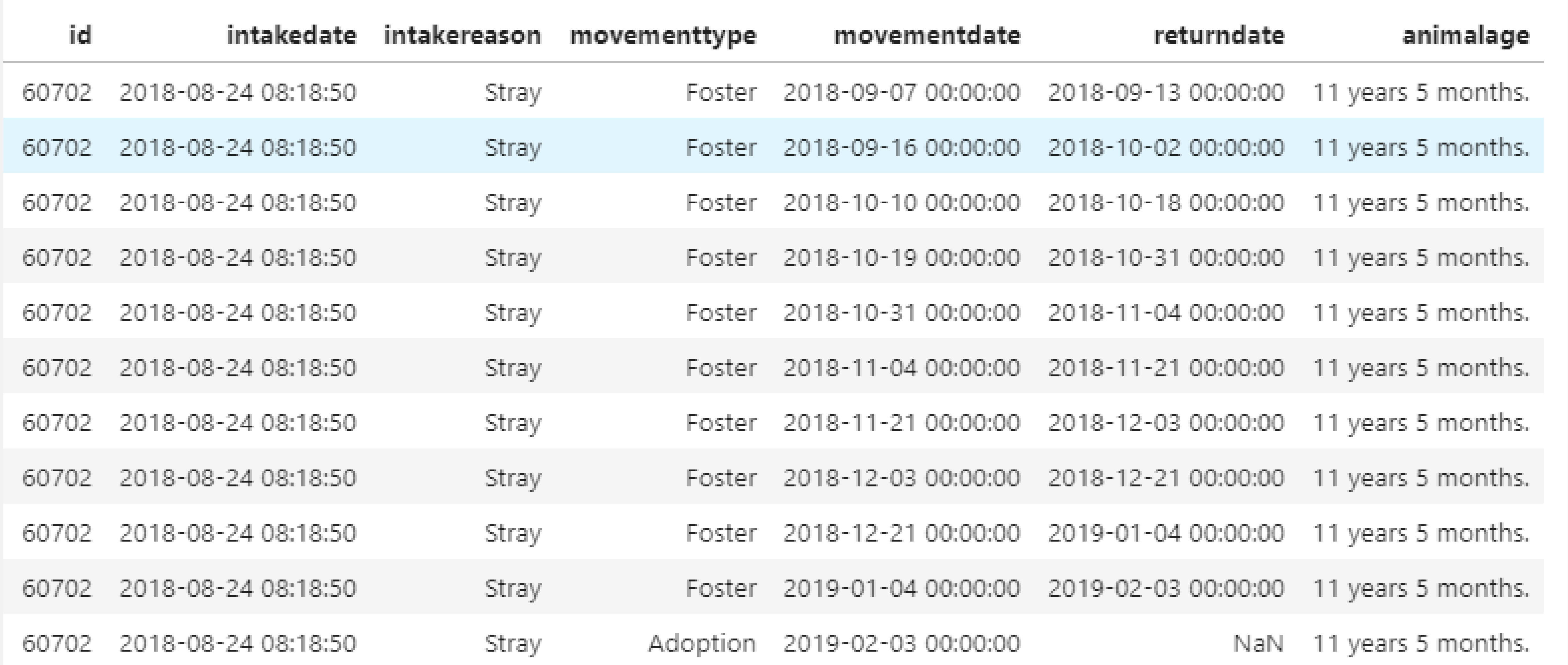
CHALLENGES

Feature selection and engineering:

- *Need to create new features out of the date time which might be very troublesome
- My label is derived from multiple Booleans ---> not that easy to encode … :(
- Need to ensure I use one hot/ label encoding properly

CHALLENGES

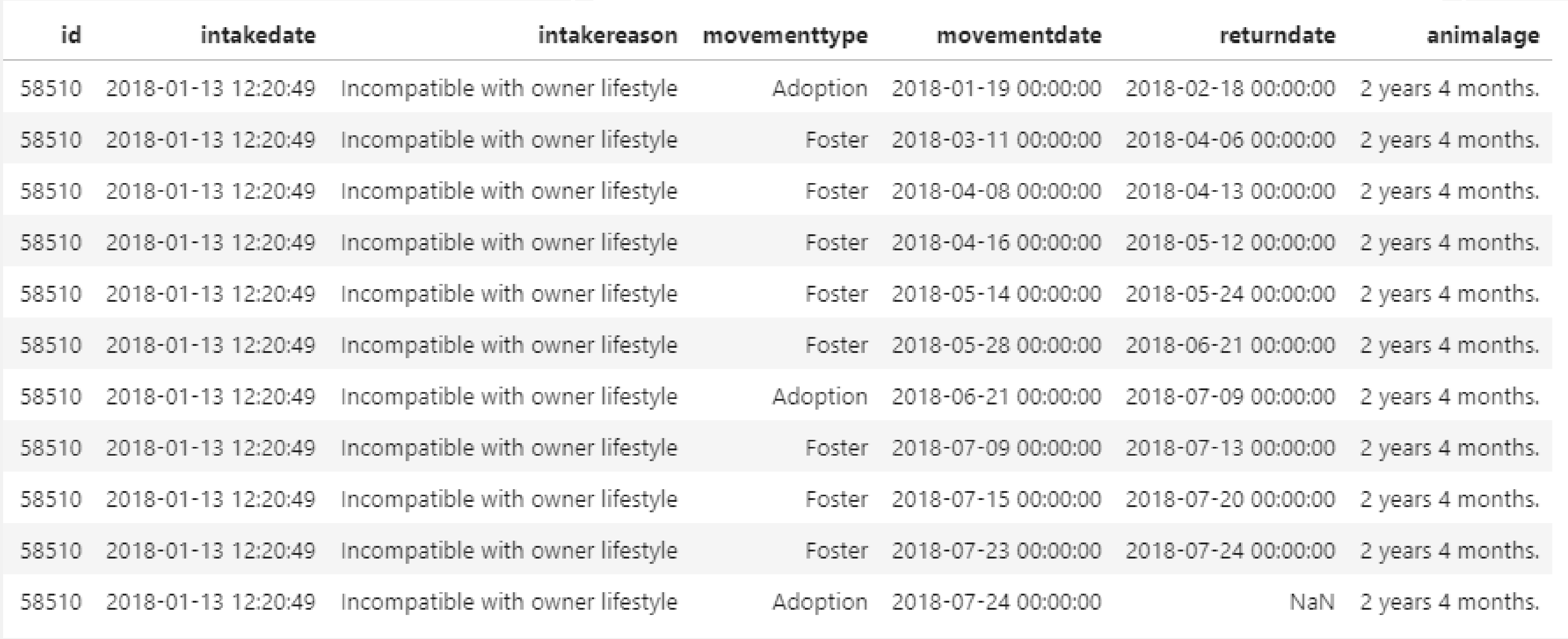
- Animal age doesn’t change and is only on intake date.
- Intake date also doesn’t update to the return date
- It probably is not even worth it to re-categorise the data
- Cannot consider duplicates since they're all different



id	intakedate	intakereason	movementtype	movementdate	returndate	animalage
60702	2018-08-24 08:18:50	Stray	Foster	2018-09-07 00:00:00	2018-09-13 00:00:00	11 years 5 months.
60702	2018-08-24 08:18:50	Stray	Foster	2018-09-16 00:00:00	2018-10-02 00:00:00	11 years 5 months.
60702	2018-08-24 08:18:50	Stray	Foster	2018-10-10 00:00:00	2018-10-18 00:00:00	11 years 5 months.
60702	2018-08-24 08:18:50	Stray	Foster	2018-10-19 00:00:00	2018-10-31 00:00:00	11 years 5 months.
60702	2018-08-24 08:18:50	Stray	Foster	2018-10-31 00:00:00	2018-11-04 00:00:00	11 years 5 months.
60702	2018-08-24 08:18:50	Stray	Foster	2018-11-04 00:00:00	2018-11-21 00:00:00	11 years 5 months.
60702	2018-08-24 08:18:50	Stray	Foster	2018-11-21 00:00:00	2018-12-03 00:00:00	11 years 5 months.
60702	2018-08-24 08:18:50	Stray	Foster	2018-12-03 00:00:00	2018-12-21 00:00:00	11 years 5 months.
60702	2018-08-24 08:18:50	Stray	Foster	2018-12-21 00:00:00	2019-01-04 00:00:00	11 years 5 months.
60702	2018-08-24 08:18:50	Stray	Foster	2019-01-04 00:00:00	2019-02-03 00:00:00	11 years 5 months.
60702	2018-08-24 08:18:50	Stray	Adoption	2019-02-03 00:00:00	NaN	11 years 5 months.

CHALLENGES

- Same same but with different data



id	intakedate	intakereason	movementtype	movementdate	returndate	animalage
58510	2018-01-13 12:20:49	Incompatible with owner lifestyle	Adoption	2018-01-19 00:00:00	2018-02-18 00:00:00	2 years 4 months.
58510	2018-01-13 12:20:49	Incompatible with owner lifestyle	Foster	2018-03-11 00:00:00	2018-04-06 00:00:00	2 years 4 months.
58510	2018-01-13 12:20:49	Incompatible with owner lifestyle	Foster	2018-04-08 00:00:00	2018-04-13 00:00:00	2 years 4 months.
58510	2018-01-13 12:20:49	Incompatible with owner lifestyle	Foster	2018-04-16 00:00:00	2018-05-12 00:00:00	2 years 4 months.
58510	2018-01-13 12:20:49	Incompatible with owner lifestyle	Foster	2018-05-14 00:00:00	2018-05-24 00:00:00	2 years 4 months.
58510	2018-01-13 12:20:49	Incompatible with owner lifestyle	Foster	2018-05-28 00:00:00	2018-06-21 00:00:00	2 years 4 months.
58510	2018-01-13 12:20:49	Incompatible with owner lifestyle	Adoption	2018-06-21 00:00:00	2018-07-09 00:00:00	2 years 4 months.
58510	2018-01-13 12:20:49	Incompatible with owner lifestyle	Foster	2018-07-09 00:00:00	2018-07-13 00:00:00	2 years 4 months.
58510	2018-01-13 12:20:49	Incompatible with owner lifestyle	Foster	2018-07-15 00:00:00	2018-07-20 00:00:00	2 years 4 months.
58510	2018-01-13 12:20:49	Incompatible with owner lifestyle	Foster	2018-07-23 00:00:00	2018-07-24 00:00:00	2 years 4 months.
58510	2018-01-13 12:20:49	Incompatible with owner lifestyle	Adoption	2018-07-24 00:00:00	NaN	2 years 4 months.

HOW

- Attempt AUTO ML with TPOT -> save me all the trouble of doing it manually -> Risk of highly overfitted data
- Also manually do it
- Mutli-class label -> AUC OVO &/or AUC OVR (Might need to manually call it for use in LAzyPredict)
- Hypertune the parameters and test on another dataset.
- Mitigate any under/overfitting.

Sub-goals:

- Do more Fostering sessions improve adoption rates and reduce adoption returns.

- Does animal age affect adoption rates &/or return rates

- is there a common reason for returns

- is there a common reason for abandonment

- Does the shelter itself actually affect the rate of adoption?

- will a binary label be easier for the machine to learn? Will consider it

- explain the bias variance trade off for models used.