THEY ARE EVERYWHERE

Pet Saver 1.0

Springboard Data Science Career Track
February 2019 Cohort Capstone Project

by TUNCOGLU

Contents

- Why?
- How?
  - Data cleaning
  - EDA
  - Statistical Inferences
  - Machine Learning
- What?
- What else?

Why?

No one can pinpoint exactly when humans first started keeping dogs as pets, but estimates range from roughly 13,000 to 30,000 years ago.

Sheep and goats were first domesticated roughly 11,000 years ago, while cats became pets around 7000 B.C. with the advent of agriculture.

Several thousand years later, around 4000 B.C., as trade routes developed, humans began using oxen, donkeys, and camels to transport goods.

What is the very first pet?

Since when human kind have pets?

Spoiler: It is not dinasours

How many pets are there in the world?

- It is estimated that there are 500 million pets and 600 million stray pets living on the planet earth.
- Sixty-eight percent of U.S. households, or about 85 million families, own a pet,
- 2017-2018 National Pet Owners Survey conducted by the American Pet Products Association (APPA).

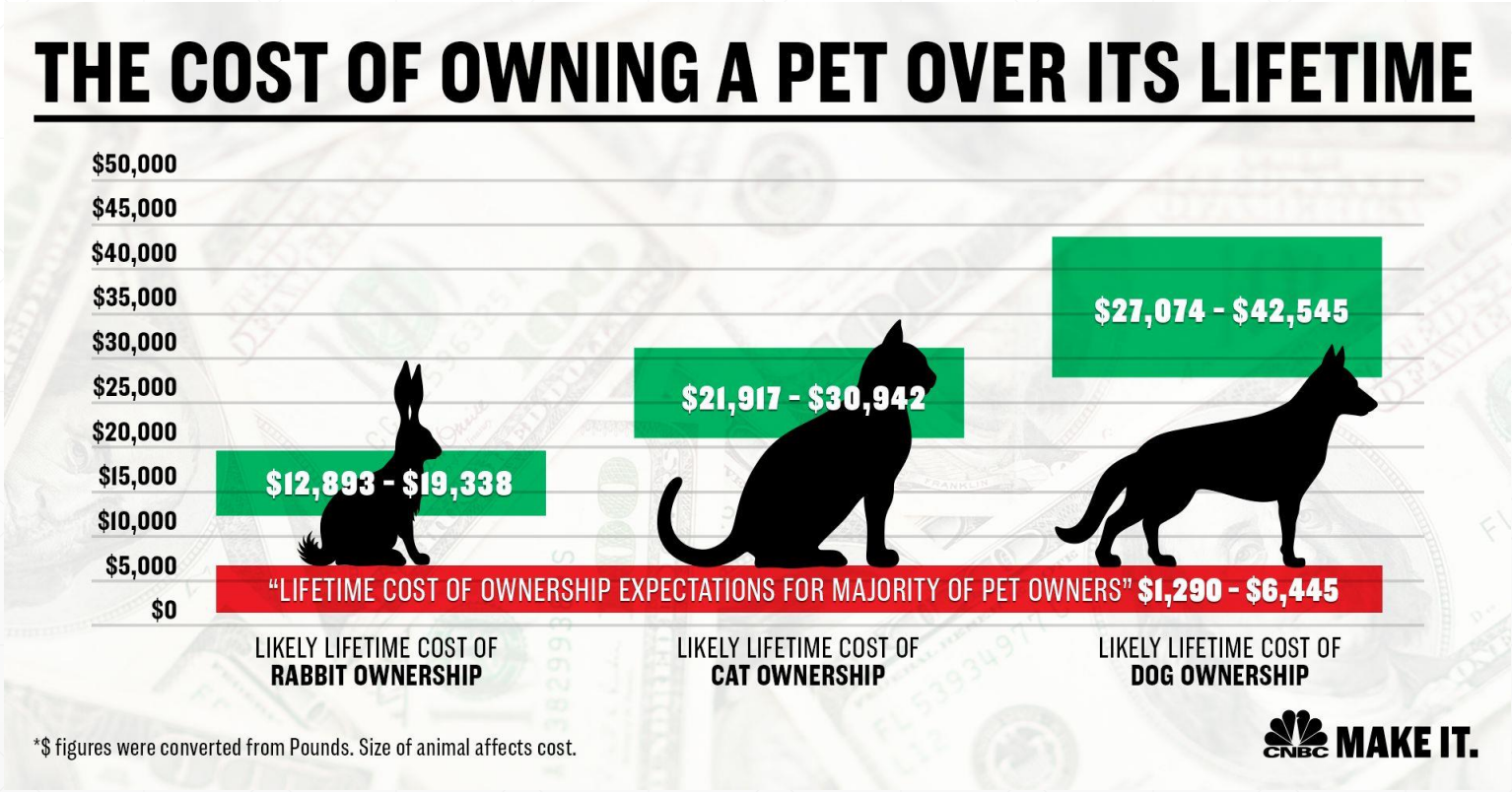
Pet	Number (millions)
Dog	60.2
Cat	47.1
Fish	12.5
Bird	7.9
Reptile	4.7

Why?

- Approximately 6.5 million companion animals enter U.S. animal shelters nationwide every year. Of those, approximately 3.3 million are dogs and 3.2 million are cats.
- Each year, approximately 1.5 million shelter animals are euthanized (670,000 dogs and 860,000 cats).

- The United States represents about **4.4 percent** of the world's population, we can proportionate accordingly and imagine the numbers.

Why?


THE COST OF OWNING A PET OVER ITS LIFETIME
$50,000
$45,000
$40,000
$35,000
$30,000
$25,000
$20,000
$15,000
$10,000
$5,000
$0
$12,893 - $19,338
$21,917 - $30,942
$27,074 - $42,545
"LIFETIME COST OF OWNERSHIP EXPECTATIONS FOR MAJORITY OF PET OWNERS" $1,290 - $6,445
LIKELY LIFETIME COST OF RABBIT OWNERSHIP
LIKELY LIFETIME COST OF CAT OWNERSHIP
LIKELY LIFETIME COST OF DOG OWNERSHIP
*$ figures were converted from Pounds. Size of animal affects cost.
CNBC MAKE IT.

How?

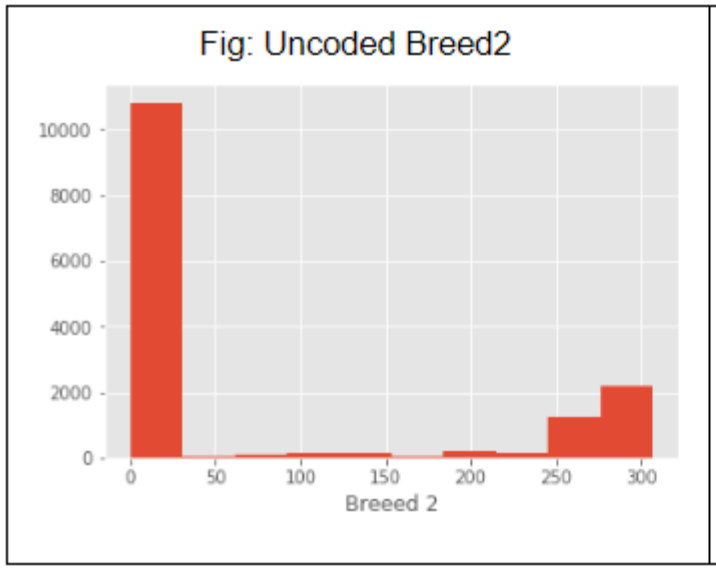
Thank you petfinder.my

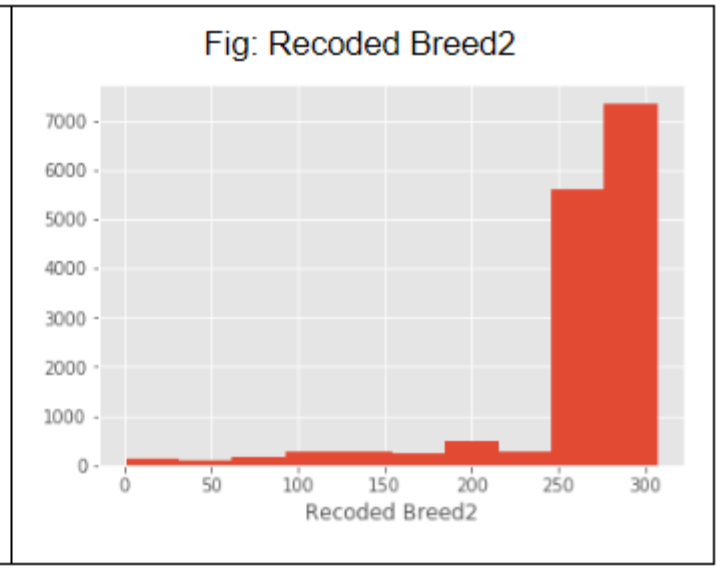
- petfinder.my shared its data set at kaggle.com.
- Data contained 14993 cats and dogs, have profile on the website. Some with pictures, some with descriptions.
- Only, numerical data is used for the analysis.
- 16 categorical and 3 continuous features used for analysis.

```
Data columns (total 24 columns):
Type              14993 non-null int64
Name              13736 non-null object
Age               14993 non-null int64
Breed1            14993 non-null int64
Breed2            14993 non-null int64
Gender            14993 non-null int64
Color1            14993 non-null int64
Color2            14993 non-null int64
Color3            14993 non-null int64
MaturitySize      14993 non-null int64
FurLength         14993 non-null int64
Vaccinated        14993 non-null int64
Dewormed          14993 non-null int64
Sterilized        14993 non-null int64
Health            14993 non-null int64
Quantity          14993 non-null int64
Fee               14993 non-null int64
State             14993 non-null int64
RescuerID         14993 non-null object
VideoAmt          14993 non-null int64
Description       14981 non-null object
PetID             14993 non-null object
PhotoAmt          14993 non-null float64
AdoptionSpeed     14993 non-null int64
dtypes: float64(1), int64(19), object(4)
```

Data Cleaning

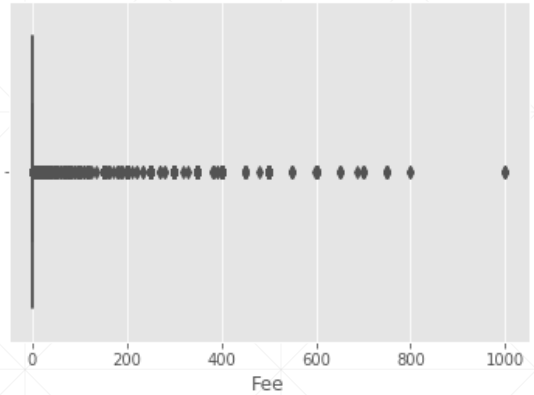
- Website is based on a relational database, numerical data was clean.
- To prevent missing values, recoding in done for some features.

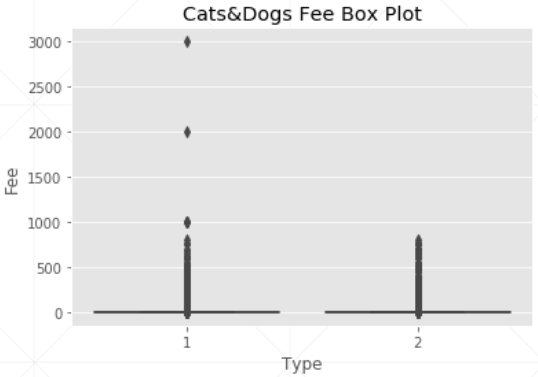


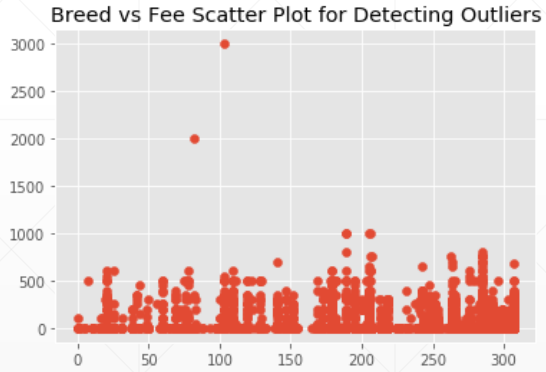


Data Cleaning

- Outliers are detected and removed.







Statistical Analysis

- Two types of pets available in the data set. There is a difference in median. Is it statistically significant?

Type	Count	Mean	Std	Min	25%	50%	75%	Max
1	8119	2.65	1.14	0	2	3	4	4
2	6858	2.40	1.20	0	1	2	4	4

Are you a cat person or dog person?

- The difference between medians can be analyzed with Kruskal Wallis test, aka non-parametric ANOVA. This test can be applied to ordinal categorical variable and calculates via rankings.
  - H0 : Medians of the both groups are equal.
  - H1 : Medians of the both groups are not equal.
- Our test shows p = 3.8183961327530204e-28 < alpha = 0.05 and H0 hypothesis is rejected with 95% confidence. Samples have different medians and they are coming from different populations.

Crosstabs and Feature Independences

Adoption Speed and Color1 Crosstab for DOGS

Color1	1	2	3	4	5	6	7
AdoptionSpeed							
0	66	61	14	0	19	2	7
1	695	420	59	23	130	21	85
2	1101	673	94	26	149	28	90
3	964	640	74	38	132	15	84
4	1113	862	100	62	150	27	95

Adoption Speed and Color1 Crosstab for CATS

Color1	1	2	3	4	5	6	7
AdoptionSpeed							
0	106	58	15	13	15	27	6
1	802	268	152	117	90	135	91
2	933	283	191	119	89	166	92
3	665	214	109	107	53	118	41
4	974	268	138	129	57	143	74

Contingency tables are one of the most important summarizing and preparation tool for categorical variables.

Chi-squared analysis is one of the fundamental test of statistics and tests the independence of the two variable at the contingency table.

Among 400 variable combinations 354 variable couple are independent from each other and 46 variable couple are not independent from each other.

Machine Learning

ML Algorithm	Trimmed Data (Outlier Free)	Whole Data	Change in Accuracy Rate
KNN	35.5%	34.5%	+1%
SVC	36.7%	35.3%	+1.4%
Logistic Regression	34.5%	32.9%	+1.6%
Naive Bayes	32.8%	29.4%	+3.4%
Random Forest	34.5%	35.4%	-0.9%
XGBoost	37.9%	40.5%	-2.6%

*All comparisons are between default parameter models.

Machine Learning

ML Algorithm	Cat Accuracy	Dog Accuracy	Whole Data
KNN	33.0%	38.0%	34.5%
SVC	35.9%	39.2%	35.3%
Logistic Regression	33.8%	34.1%	32.9%
Naive Bayes	21.3%	29.9%	29.4%
Random Forest	32.3%	37.8%	35.4%
XGBoost	37.5%	42.1%	41.1%

*All comparisons are between default parameter models, except XGBoost. XGBoost comparisons are between tuned models.

What?

Results

Tuning hyperparameters is essential for any ML application, with randomized grid search our models tuned with different hyperparameter options.

All models applied with 5-fold cross validation.

ata also splitted into 80% training and 20% test batches.

Results

Initial Random Forest model with default parameters with 5 fold cross validation performed 35.4% accuracy. With our parameter tuning our last model performed 40.82% accuracy with group accuracies of

Group Code	0	1	2	3	4
F1 Score	0.0465	0.3113	0.3917	0.2410	0.5642

Initial XGBoost model with default parameters showed it efficiency before any hyper parameter tuning. Base model accuracy is 40.6% with group scores of:

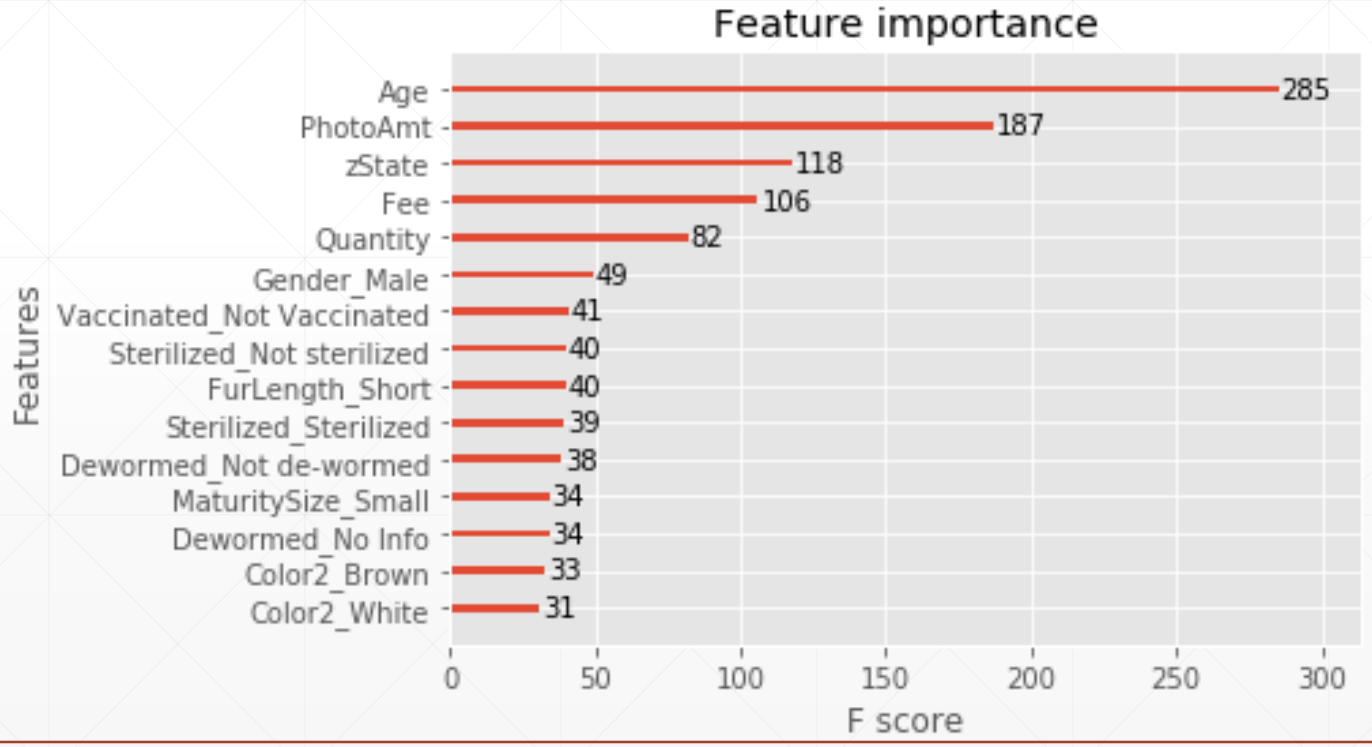
Group Code	0	1	2	3	4
F1 Score	0.0706	0.3339	0.3818	0.2113	0.5648

To increase the model accuracy, we tuned the hyperparameters and be able to optimize model accuracy, our tuned model accuracy is 41.1% and group scores are;

Group Code	0	1	2	3	4
F1 Score	0.0476	0.3436	0.3826	0.2437	0.5688

Lightside of the medallion

With the 41.1% accuracy, we can improve the listings to increase adoption speed and adoptability.



Darkside of the medallion

Each year 1.5 million pets are euthanized only in the USA, worldwide it is possible that it is tens of millions.

If we can predict the animals which will be euthanized eventually, we can guide rescue centers.

If we can reduce the cost of keeping animals, centers can help more animals.

We have over 50% accuracy on not adopted pets, so another program can be started for them to save them from euthanizing.

What else?

More data, cleaner data means better models and better predictions. In our data set data coding needs to be improved and more importantly our target variable has non-equal time blocks. This can be changed and if changed it will improve accuracy.

Image and descriptions can be added, since our models only included numerical data available.

But we still can’t predict