
Looking for Love
Mining Dog Adoption Vocabulary

Mine descriptive texts from a canine adoption website to find words that extend time on site.

- Source: www.puppyfinder.com
- Scrape website
- Available features: name, gender, description, age category (Baby, Young, Adult, Senior), date posted, breed, locale

Sampling

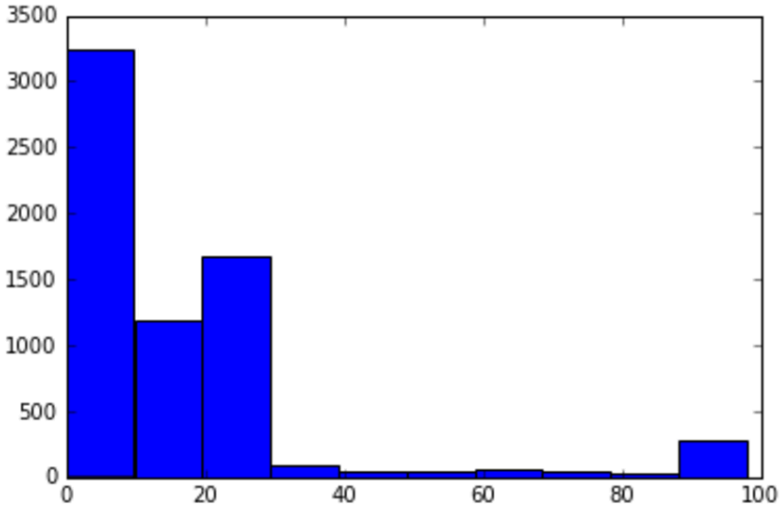
Only dogs for adoption.

Only dogs from California (6500+).

Plan: Sample three times one week apart and look at changes in inventory.

Data Problems

'2016-04-02', '2016-03-29', '2016-03-19', '2016-03-18',
'2016-03-08', '2016-03-07', '2016-03-06', '2016-03-05',
'2016-03-03', '2016-03-02', '2016-03-01', '2016-02-29',
'2016-02-28', '2016-02-27', '2016-02-25', '2016-02-24',
'2016-02-23', '2016-02-22', '2016-02-21', '2016-02-20',
'2016-02-19', '2016-02-18', '2016-02-17', '2016-02-16',...

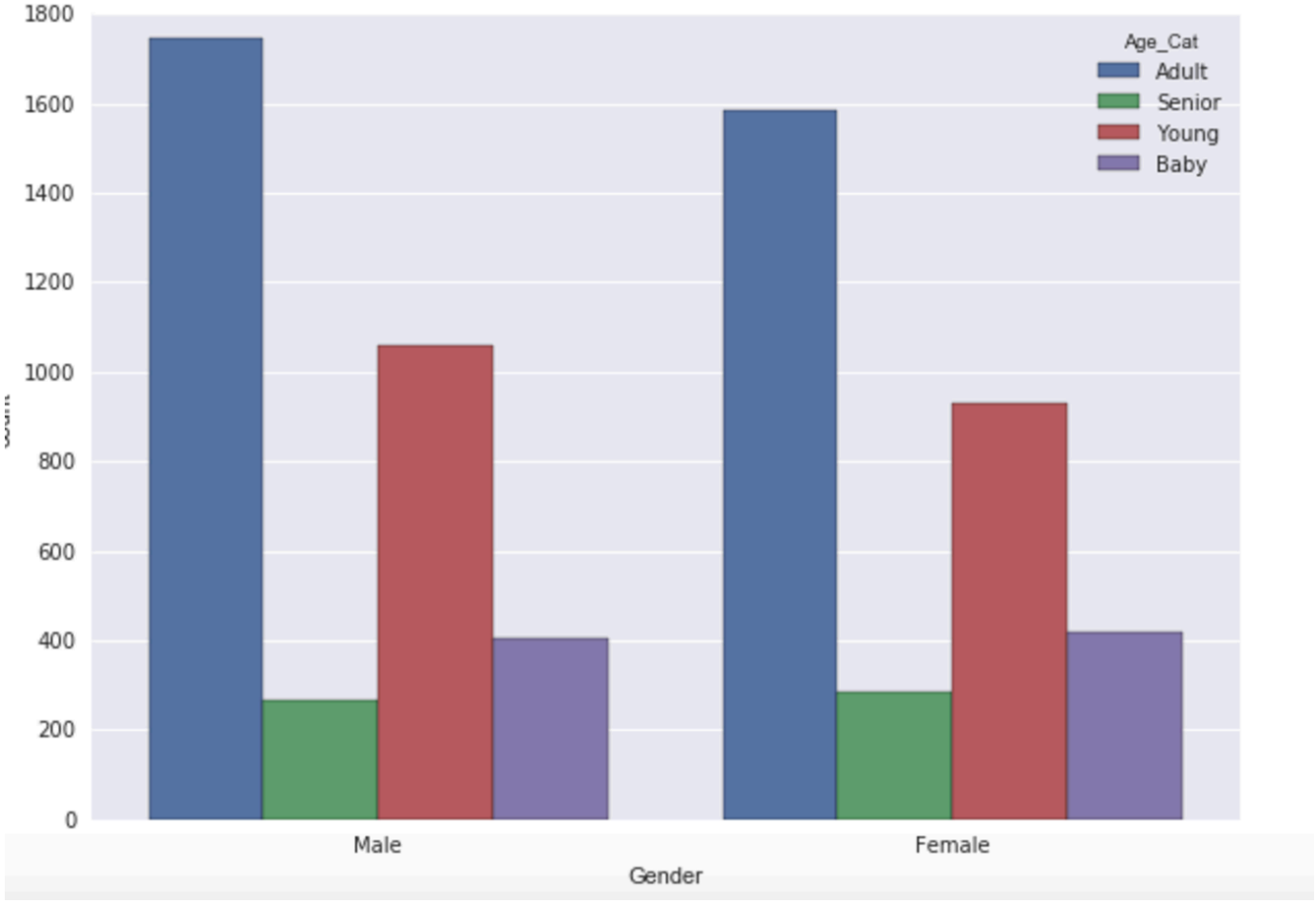


2016-03-29	2035
2016-04-02	1199
2016-03-07	848
2016-03-18	682
2016-03-19	505
2016-03-06	460
2015-12-26	254
2016-03-05	200
2016-03-08	172
2016-02-29	25
2016-02-03	15
2016-02-24	15

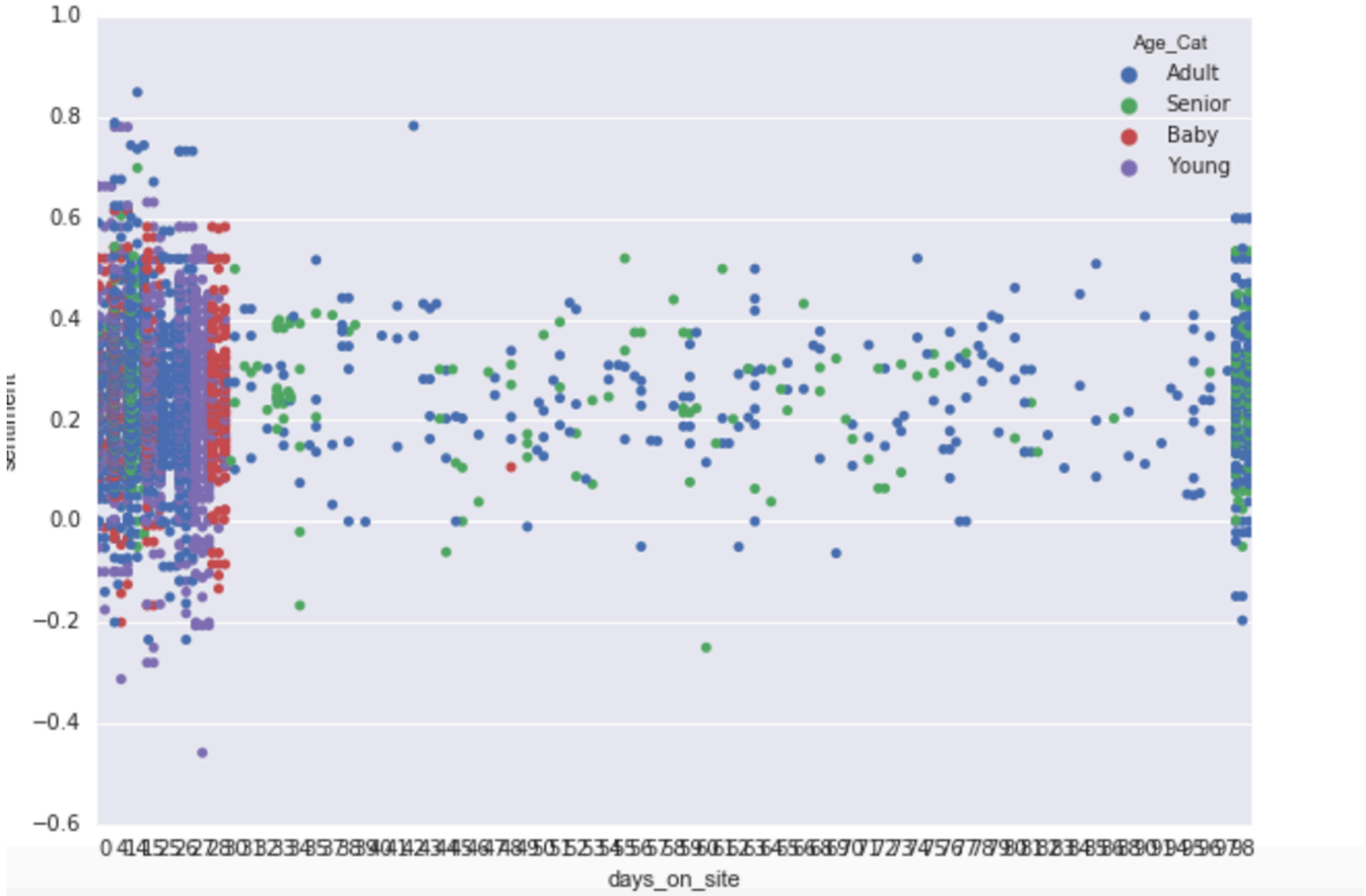
Created Features

- Sentiment (using TextBlob package)
- Subjectivity (using TextBlob package)
- Days on Site = max date – posted date
- Age in years = extracted from description
- Is_male (1 if yes, 0 if no)
- Age categorical : Young, Adult, Senior
- Time on site: 1 if less than 40 days, 0 if longer

Gender by Age



Sentiment by days on site



A different, simple vocabulary

Dog breed.

This is a very special vocabulary but it has meaning and emotional power.

Listing the dogs by breed gave me the idea to mine the breeds for value. Instead of using the breeds as class variables (there are a lot of breeds) I would employ vocabulary tools.

American Pit Bull #1

Chihuahua: The # 2 breed

Top 15 Breeds

Breed	Count
Mutt	913
American Pit Bull Terrier	555
Chihuahua-Unknown Mix	452
Chihuahua	405
American Pit Bull Terrier-Unknown Mix	404
Labrador Retriever-Unknown Mix	300
German Shepherd Dog	200
German Shepherd Dog-Unknown Mix	100
American Bulldog	100
American Staffordshire Terrier-Unknown Mix	96
Dachshund-Unknown Mix	91
American Staffordshire Terrier	89
Boxer	80
Labrador Retriever	80
Chiweenie	76

Age Characteristics

	Sentiment (mean/sd)	Days on Site (mean/sd)
Baby	.25/.12	11/10
Young	.24/.13	15/11
Adult	.24/.12	21./26
Senior	.24/.12	36/37

Lasso to Explore Data

X=['is_male' 'is_rescue' 'sentiment' 'subjectivity' 'agecat_Adult' 'agecat_Senior' 'agecat_Young' 'american' 'australian' 'boxer' 'bull' 'bulldog' 'chihuahua' 'dachshund' 'dog' 'german' 'labrador' 'miniature' 'mix' 'mutt' 'pit' 'poodle' 'retriever' 'shepherd' 'staffordshire' 'terrier' 'unknown']

Y=days_on_site

Dog breed words were extracted using CountVectorizer over the Breed feature with max_features set to 20.

A WYSIWYG Result

agecat_Young = -0.15283189
agecat_Adult = 0.52657824
agecat_Senior = 4.70832842

All other features converged to zero. The intercept is 18.52 .

From this we see that a Senior dog is predicted to be on the site 4.7 days longer than a Baby, but a Young dog will be on the site less time than a Baby.

Estimated Ages for Age Categories

- Baby: median age of 2, max=10
- Young: median age of 2, max=13
- Adult: median age of 4, max=15
- Senior: median age of 8, max=16

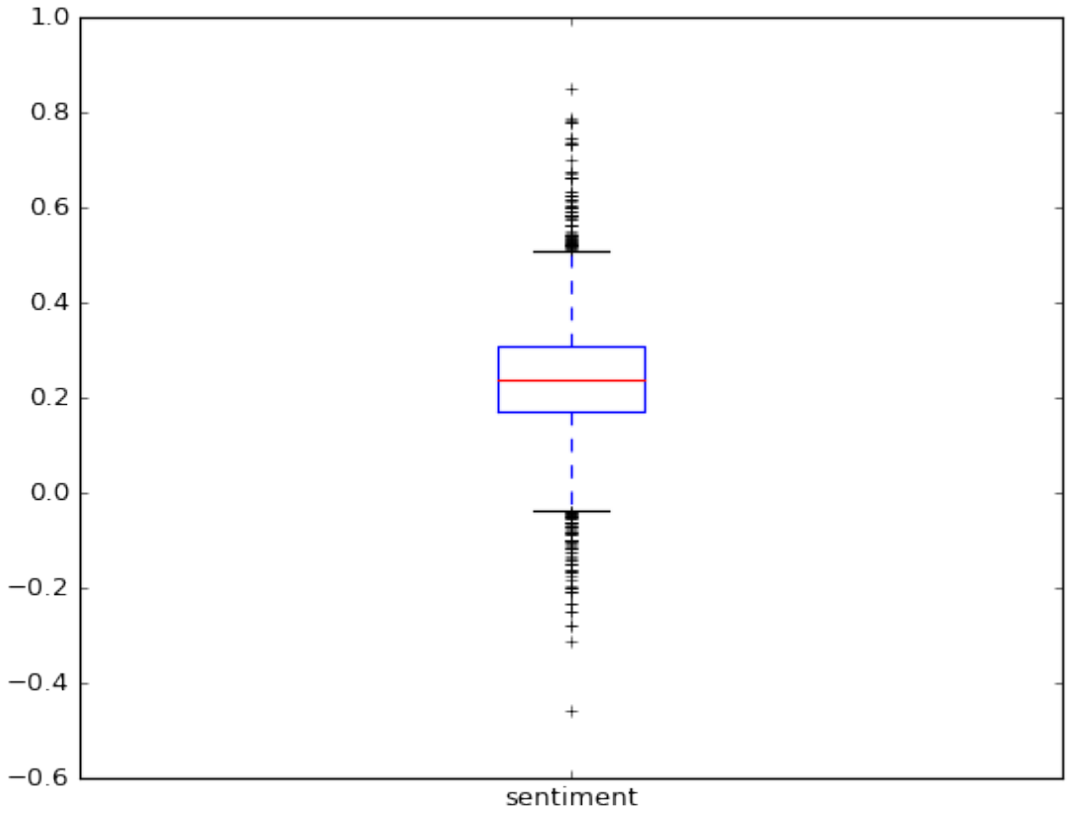
Age as extracted from the Description field.

Logistic Analysis with Sense

- Created a variable called sense which is 1 for a dog with sentiment greater than the mean of all sentiments (.241) , 0 otherwise. A good dog has a sense of 1.

- The mean of sense is .48 so 48% of dogs are good dogs.

Sentiment with Outliers



Simple modeling where y=sense and the features are our 20 breed vocabulary words

Logistic regression:

score=0. 54324724806

Random Forest Classifier 3000 trees: score=0.560718887197

Bernoulli Naive Bayes:

score=0. 546330810922

These results are better than the mean = .48

Logistic Regression Coefficients

word	coef	effect	word	coef	effect
american	0.151	16%	miniature	0.040	4%
australian	0.210	23%	mix	0.001	0%
boxer	0.254	29%	mutt	0.322	38%
bull	-0.275	-24%	pit	0.467	60%
bulldog	0.135	14%	poodle	0.074	8%
chihuahua	-0.405	-33%	retriever	-0.004	0%
dachshund	0.324	38%	shepherd	-0.061	-6%
dog	-0.627	-47%	staffordshire	0.467	60%
german	0.499	65%	terrier	0.124	13%
labrador	0.315	37%	unknown	0.122	13%

The Confusion Matrix

	Predicted “Bad”	Predicted “Good”
Observed “Bad”	1634	1091
Observed “Good”	1153	1335

Predictions

“american” “pit” “bull” “terrier”

Predicted Result = 1 (good) with the probability of a “bad”= 46% and a “good”= 54%

“chihuahua”

Predicted Result = 0 (bad) with the probability of a “bad”= 66% and a “good”= 33%

Thank you and have a nice day

Karla Leibowitz

Sf-DAT 20, April 6, 2016

This is a Chiweenie.