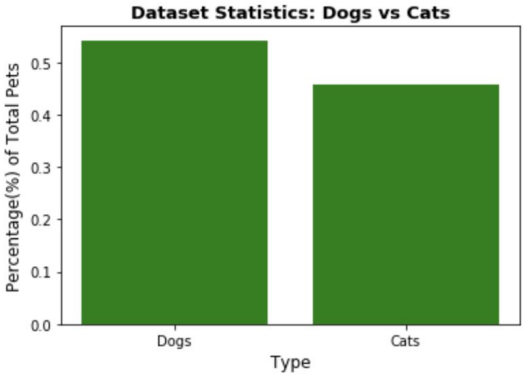
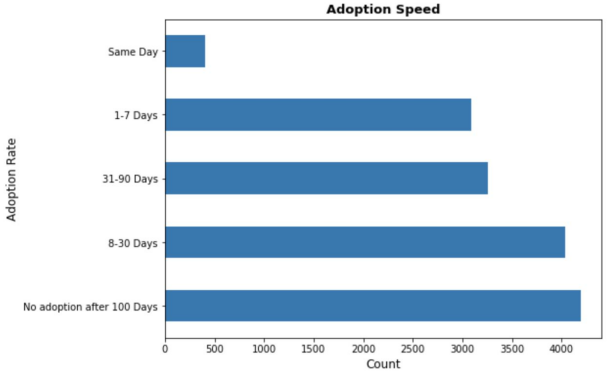
Final Project - Pet Adoption

By Shimeng Cao, Jiarong Li, and Evan Okin

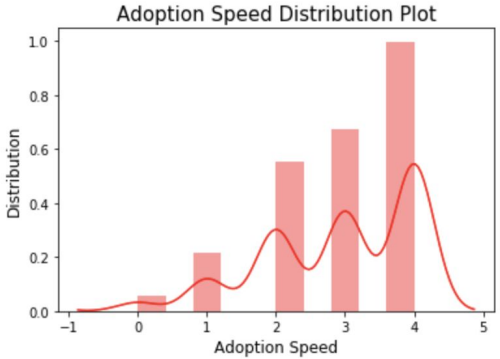
1. Data Exploration and Cleaning



The sad truth is that many pets don't get adopted quickly (or, they don't get adopted at all.



Nearly 28% of all pets in our dataset (which contains about 15,000 rows of data) were not adopted after 100 days.



We filtered on pets who possess are vaccinated, are dewormed, are sterilized, are either healthy or have only a minor injury, have no adoption fee, and age of less than 1 year. Surprisingly, 40% of our sliced dataset were still not adopted.

2. Regression Analysis

OLS Regression Results

Dep. Variable:	adoptionspeed	R-squared:	0.053
Model:	OLS	Adj. R-squared:	0.053
Method:	Least Squares	F-statistic:	76.63
Date:	Fri, 05 Jul 2019	Prob (F-statistic):	8.32e-169
Time:	14:10:59	Log-Likelihood:	-23287.
No. Observations:	14977	AIC:	4.660e+04
Df Residuals:	14965	BIC:	4.669e+04
Df Model:	11		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	1.7848	0.093	19.293	0.000	1.603	1.966
type	-0.1801	0.019	-9.278	0.000	-0.218	-0.142
age	0.0080	0.001	14.026	0.000	0.007	0.009
breed1	0.0030	0.000	17.938	0.000	0.003	0.003
gender	0.0840	0.016	5.270	0.000	0.053	0.115
quantity	0.0444	0.007	5.973	0.000	0.030	0.059
maturitysize	0.0531	0.018	3.036	0.002	0.019	0.087
vaccinated	-0.0965	0.021	-4.534	0.000	-0.138	-0.055
dewormed	0.0803	0.020	4.033	0.000	0.041	0.119
sterilized	-0.1511	0.019	-7.839	0.000	-0.189	-0.113
health	0.1737	0.048	3.654	0.000	0.081	0.267
photoamt	-0.0102	0.003	-3.709	0.000	-0.016	-0.005

Omnibus:	3545.756	Durbin-Watson:	2.006
Prob(Omnibus):	0.000	Jarque-Bera (JB):	688.542
Skew:	-0.156	Prob(JB):	3.05e-150
Kurtosis:	1.997	Cond. No.	2.82e+03

This reg_total model takes multiple independent variables which were tested to be statistically significant in impacting the dependent variable 'adoptionspeed'.

This model explains 5.3% of the total variable in the independent variables, which is still a small amount.

Based on the coefficient of each independent variable, dogs tend to be more likely to be adopted than cats, younger pets tend to be more likely to be adopted than older pets, and higher code primary breeds tend to be more likely to be adopted than lower code primary breeds.

3. ML for the Regression Problem (pet adoption speed)

5.36%

R-Squared of the regression model using 5-fold cross validation

- Y = adoptionspeed
- X = type + age + breed1 + gender + quantity + maturitysize + vaccinated + dewormed + sterilized + health + photoamt

12.51%

R-Squared of the KNN model using 5-fold cross validation

- K = 68

15.37%

R-Squared of the Random Forest model using 5-fold cross validation

- n_estimators=200
- max_depth = 9
- max_features = 5

4. ML for the Classification Problem (binary:adopted vs. not adopted)

72%

Accuracy of the regression model using 5-fold cross validation

- Y = adoption_indicator
- X = type + age + breed1 + gender + quantity + maturitysize + vaccinated + dewormed + sterilized + health + photoamt

74.33%

Accuracy of the KNN model using 5-fold cross validation

- K = 65

75.65%

Accuracy of the Random Forest model using 5-fold cross validation

- n_estimators=200
- max_depth = 11
- max_features = 4