Data Torturers Midterm Project

Petfinder Adoption Speed

Alexis Kaldany, Sahara Ensley, Yixi Liang, Kaiyuan Liang

About the Dataset

1. This dataset is intended to predict factors influences adoption speed of pets .

   （source：kaggle）

2. 14933 observations and 24 variables

```
 [1] "Type"         "Name"         "Age"          "Breed1"
 [5] "Breed2"       "Gender"       "Color1"       "Color2"
 [9] "Color3"       "MaturitySize" "FurLength"    "Vaccinated"
[13] "Dewormed"     "Sterilized"   "Health"       "Quantity"
[17] "Fee"          "State"        "RescuerID"    "VideoAmt"
[21] "Description"  "PetID"        "PhotoAmt"     "AdoptionSpeed"
```

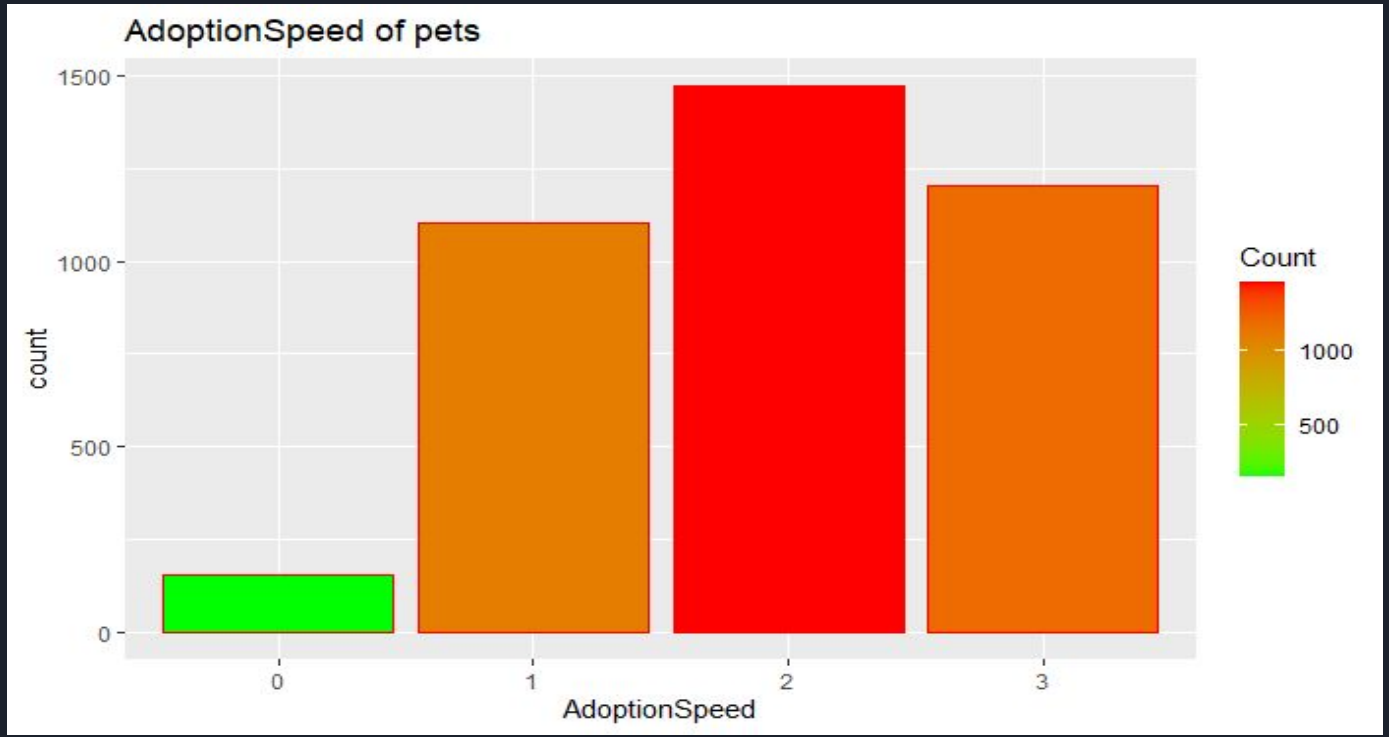
EDA Preparation

1. Convert “Type”, “MaturitySize”, “FurLength”, “Vaccinated”, “Gender” into categorical type
2. Subset the data with only one animal per profile(quantity=1)
3. Subset the necessary columns for EDA: 'Type', 'Age', 'Gender', 'MaturitySize', 'FurLength', 'Vaccinated', 'PhotoAmt', 'AdoptionSpeed'.
4. Graph used:
    a. Histogram
    b. QQ plot
    c. Scatter plot
    d. Box-plot
    e. Pie-Plot

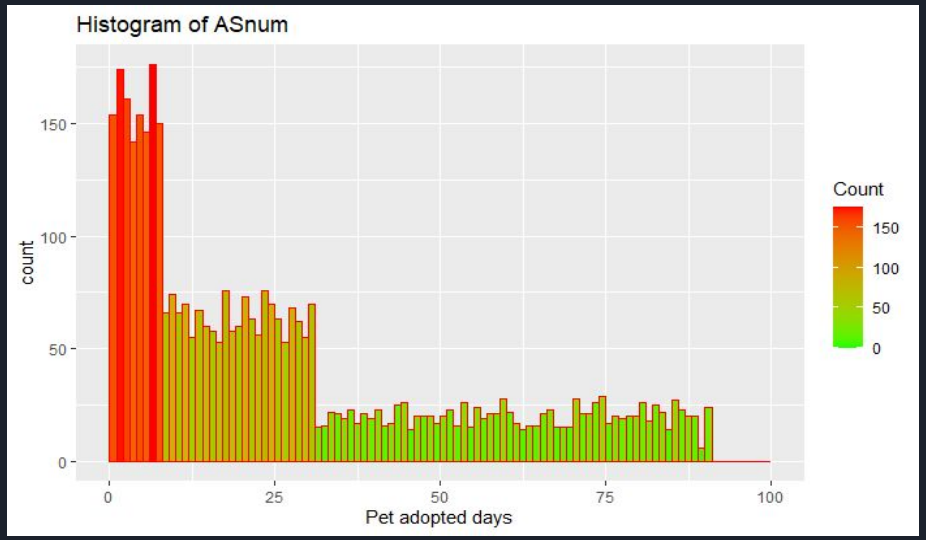
About the variables

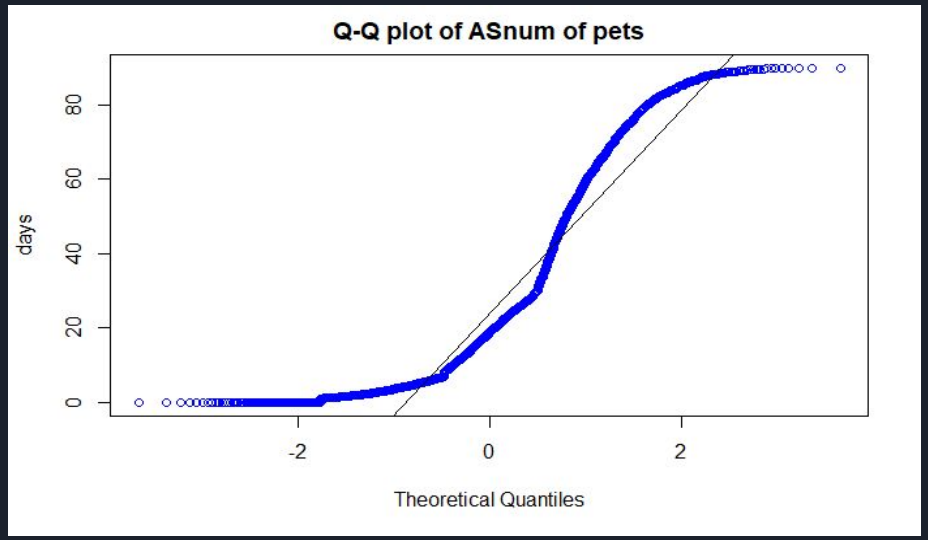
1. 'Type' : the type of animals
2. 'Age' : age of animals in month
3. 'Gender' : gender of animals
4. 'MaturitySize': the size of animals
5. 'FurLength' : the fur length of animals
6. 'Vaccinated' : the status of vaccination
7. 'PhotoAmt' : the total number of photos uploaded for animals
8. 'AdoptionSpeed' : a categorical value to predict the adoption speed of animals. The lower of values, the faster speed of animal adoption speed.

AdoptionSpeed

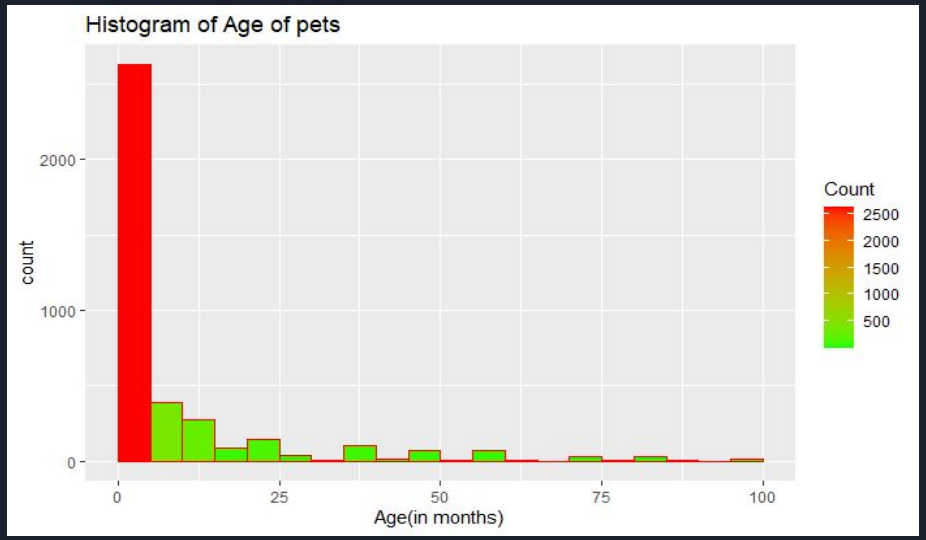


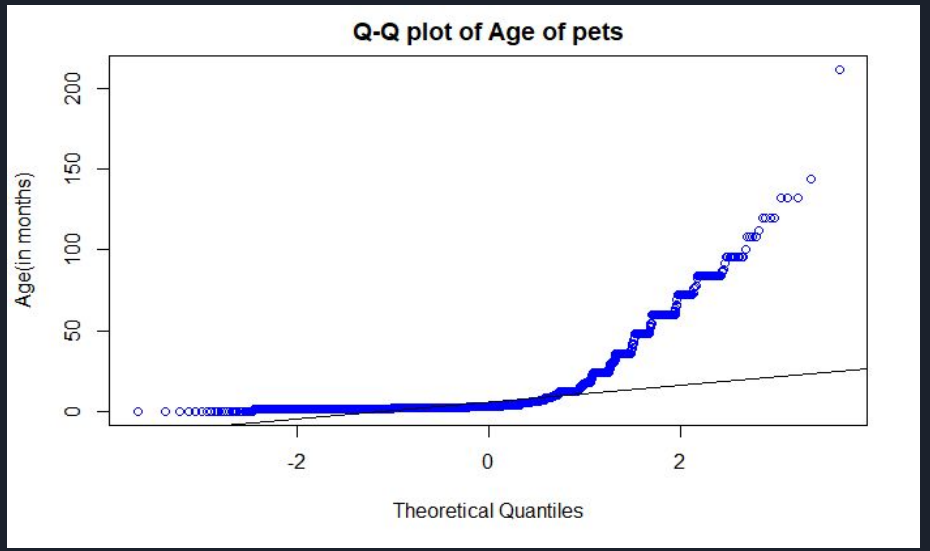
ASnum:
We transformed AdoptionSpeed to numerical variable



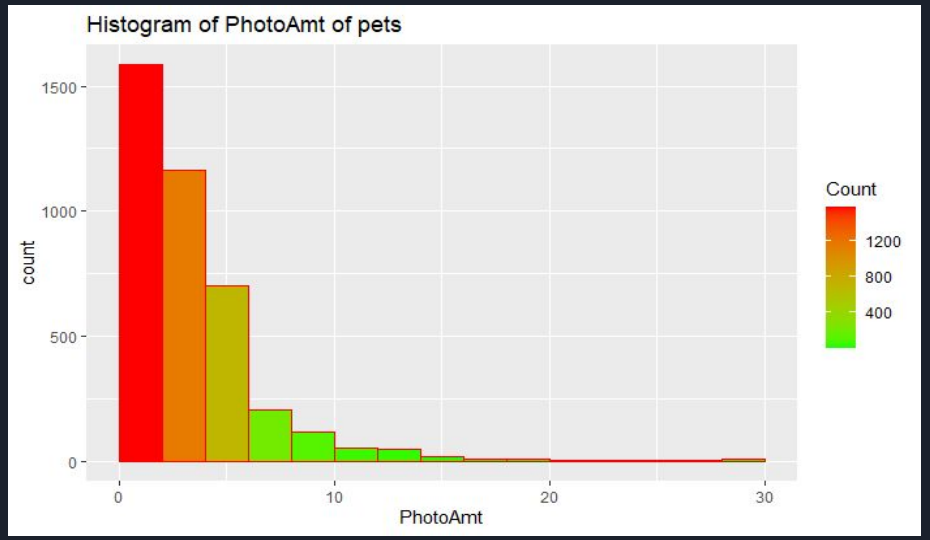


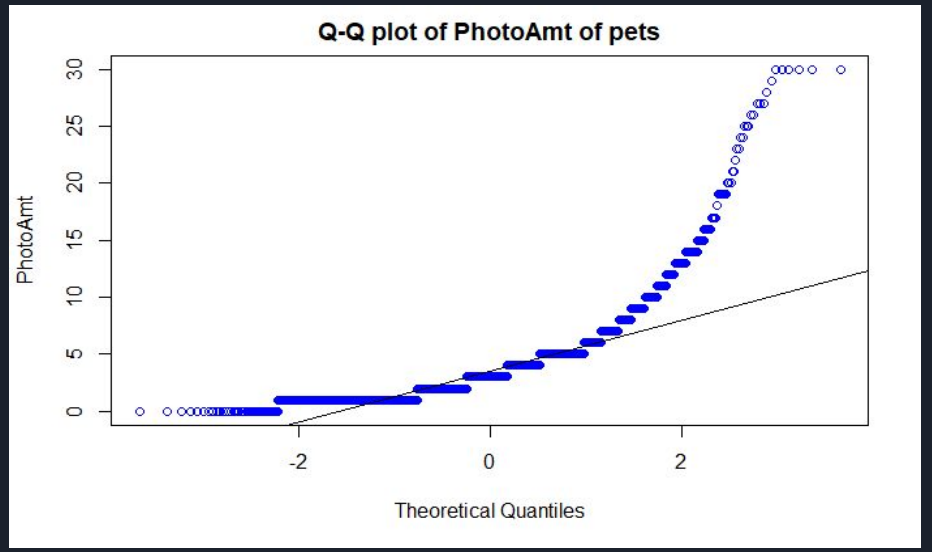
Age (in months)





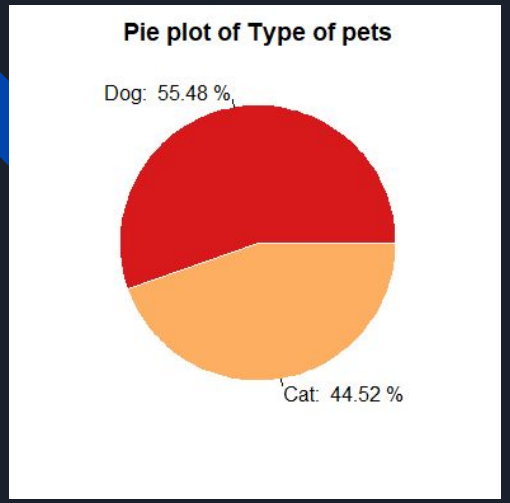
PhotoAmt

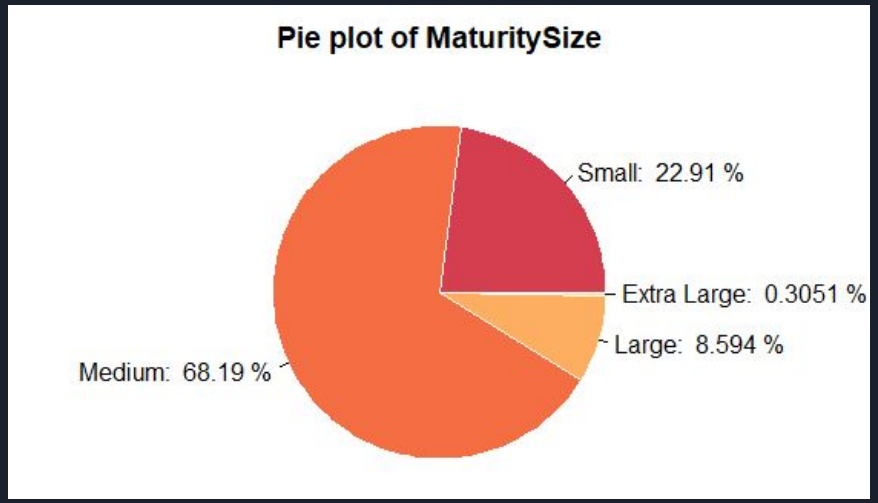

Histogram of PhotoAmt of pets
count
1500
1000
500
0
0
10
20
30
PhotoAmt
Count
1200
800
400

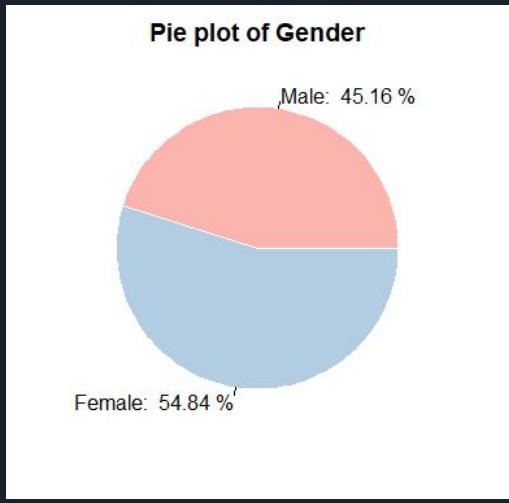

Q-Q plot of PhotoAmt of pets
PhotoAmt
0
5
10
15
20
25
30
-2
0
2
Theoretical Quantiles

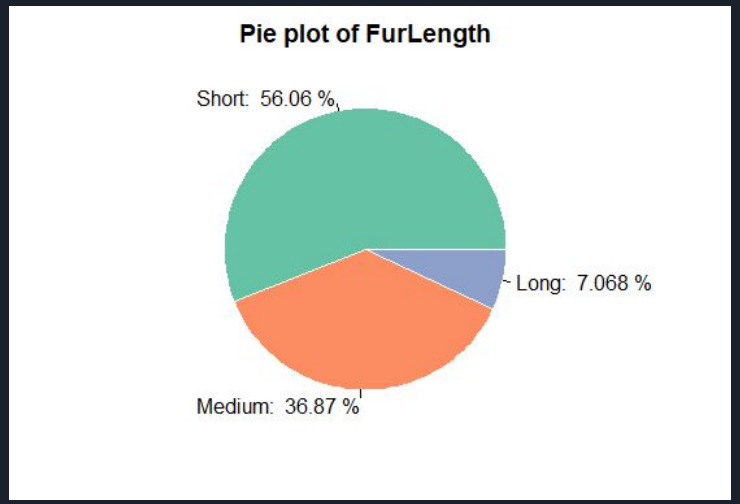
There are five categorical variables:

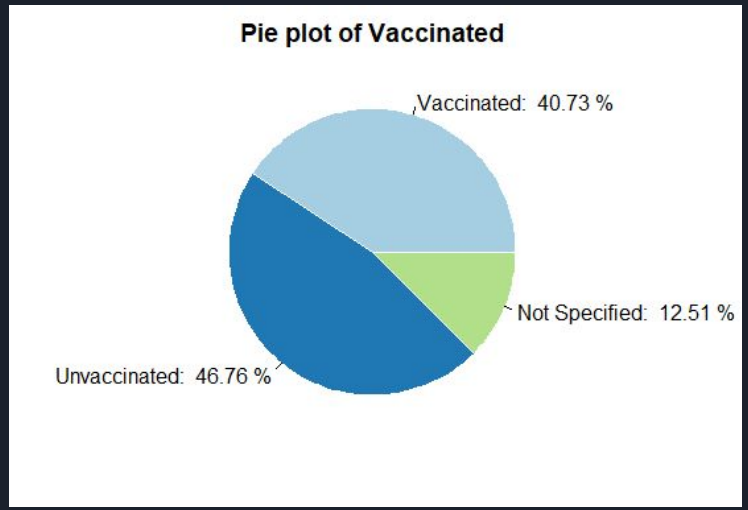
1. Type
2. Gender
3. MaturitySize
4. FurLength
5. Vaccinated


Pie plot of Type of pets
Dog: 55.48 %
Cat: 44.52 %


Pie plot of MaturitySize
Small: 22.91 %
Extra Large: 0.3051 %
Large: 8.594 %
Medium: 68.19 %


Pie plot of Gender
Male: 45.16 %
Female: 54.84 %


Pie plot of FurLength
Short: 56.06 %
Long: 7.068 %
Medium: 36.87 %


Pie plot of Vaccinated
Vaccinated: 40.73 %
Not Specified: 12.51 %
Unvaccinated: 46.76 %

(Categorical) Independent Variables EDA

1. Animal Type
2. Gender
3. Size
4. Fur Length
5. Vaccination Status

SMART: What categorical physical characteristics impact adoption speed?

Assumption Test: Homogeneity of Variance

H0: The variance is the same between the groups
H1: The variance is not the same between the groups

Type

```
data:  ASnum ~ Type
BP = 33, df = 1, p-value = 9e-09
```

Gender

```
data:  ASnum ~ Gender
BP = 8, df = 1, p-value = 0.004
```

Fur Length

```
data:  ASnum ~ FurLength
BP = 7, df = 2, p-value = 0.02
```

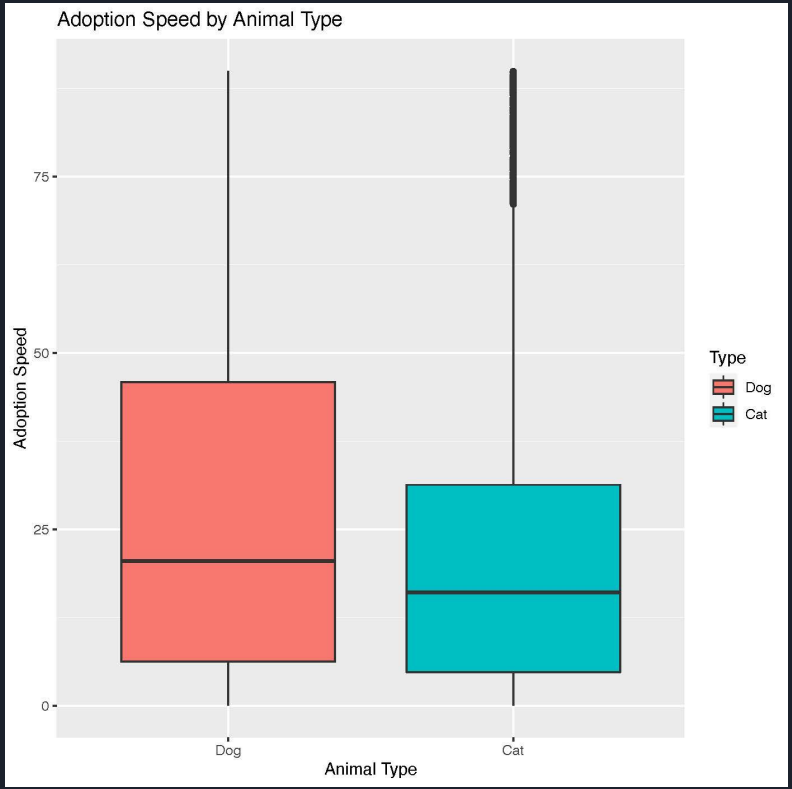
Size

```
data:  ASnum ~ MaturitySize
BP = 21, df = 3, p-value = 9e-05
```

Vaccination

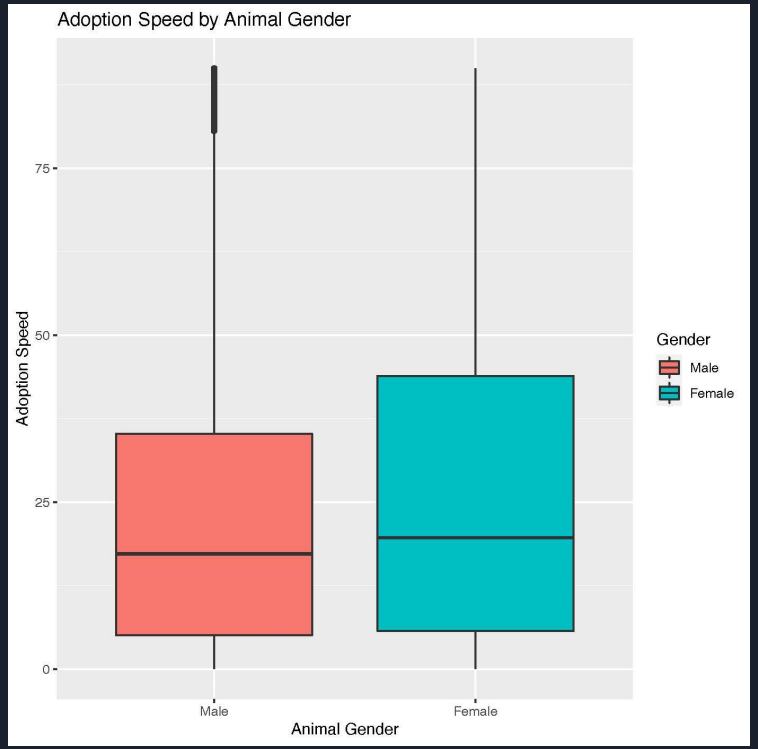
```
data:  ASnum ~ Vaccinated
BP = 51, df = 2, p-value = 9e-12
```

Do dogs get adopted quicker than cats?



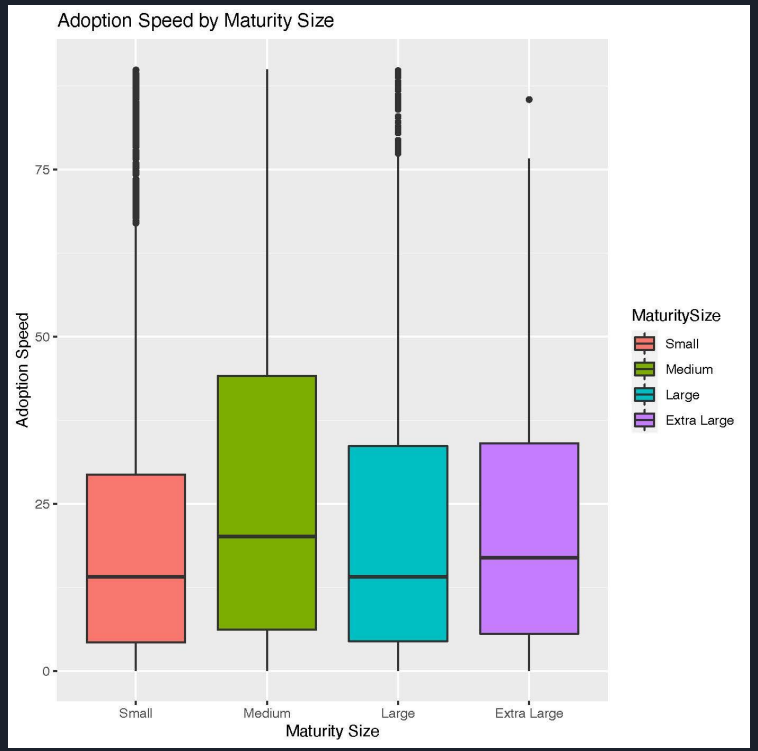

```
        Welch Two Sample t-test

data:  dogs$ASnum and cats$ASnum
t = 10, df = 8280, p-value <2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 4.16 6.30
sample estimates:
mean of x mean of y
     28.7      23.5
```

Do male animals get adopted quicker than female animals?




```
        Welch Two Sample t-test

data:  male$ASnum and female$ASnum
t = -5, df = 8297, p-value = 4e-07
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -3.88 -1.72
sample estimates:
mean of x mean of y
     24.9      27.7
```

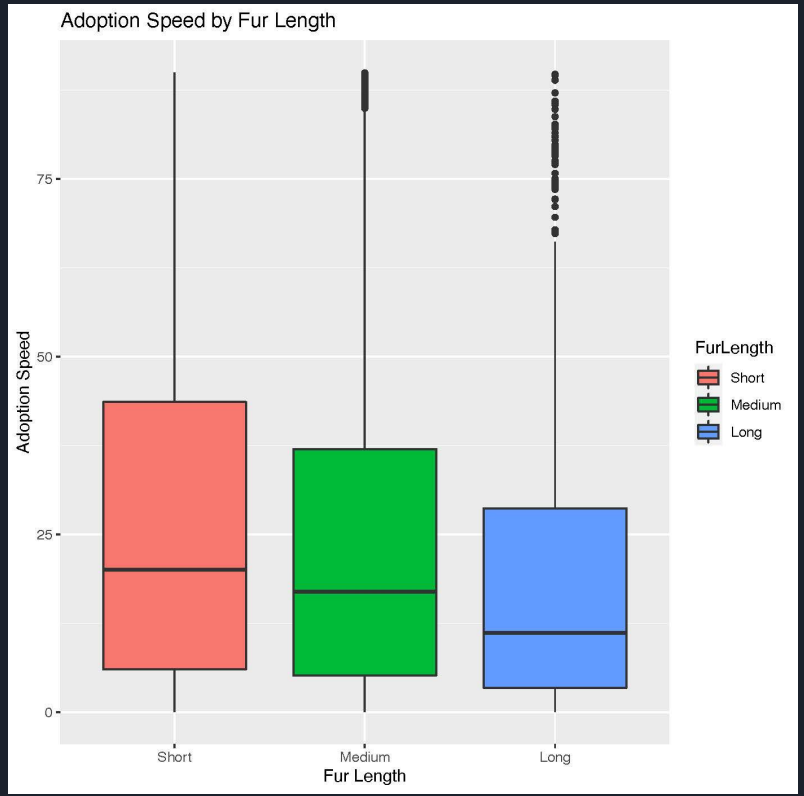
Does size impact adoption speed?




```
Analysis of Variance Table

Response: ASnum
               Df  Sum Sq Mean Sq F value Pr(>F)
MaturitySize    3   53971   17990    28.1 <2e-16 ***
Residuals    8481 5419810     639
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
$MaturitySize
      diff    lwr   upr p adj
2-1  5.631   3.91  7.35 0.000 *
3-1  0.683  -2.12  3.49 0.923
4-1  2.702  -9.66 15.07 0.943
3-2 -4.947  -7.47 -2.43 0.000 *
4-2 -2.929 -15.23  9.38 0.928
4-3  2.019 -10.48 14.52 0.976
```

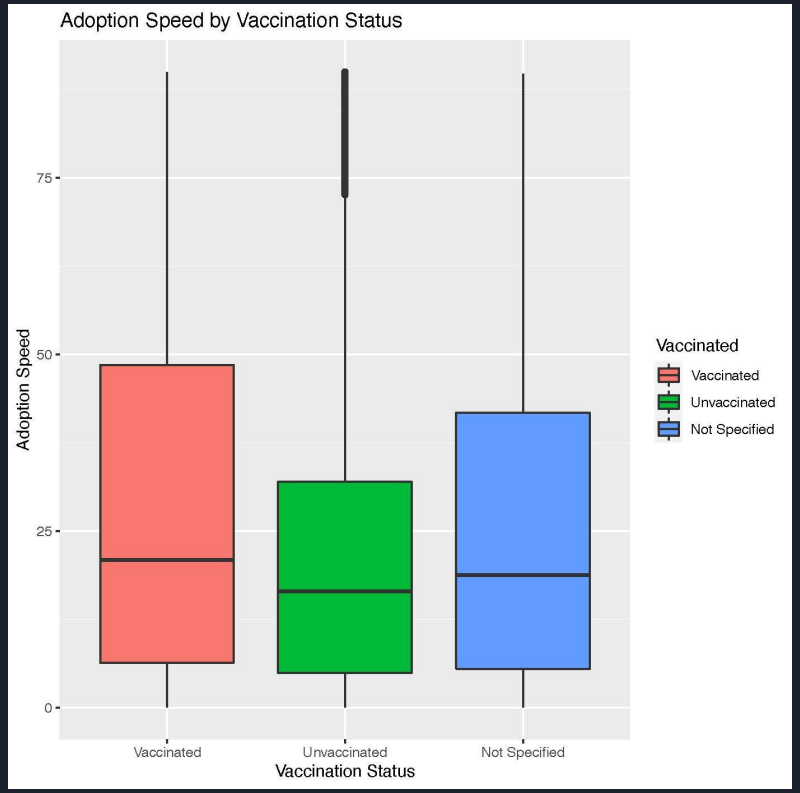
Does Fur Length impact adoption speed?




```
Analysis of Variance Table

Response: ASnum
             Df  Sum Sq Mean Sq F value  Pr(>F)
FurLength     2   31520   15760    24.6 2.3e-11 ***
Residuals  8482 5442260     642
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
$FurLength
      diff   lwr   upr p adj
2-1 -2.71 -4.08 -1.34 0.000*
3-1 -6.70 -9.27 -4.13 0.000*
3-2 -3.99 -6.64 -1.34 0.001*
```

Does Vaccination Status impact adoption speed?



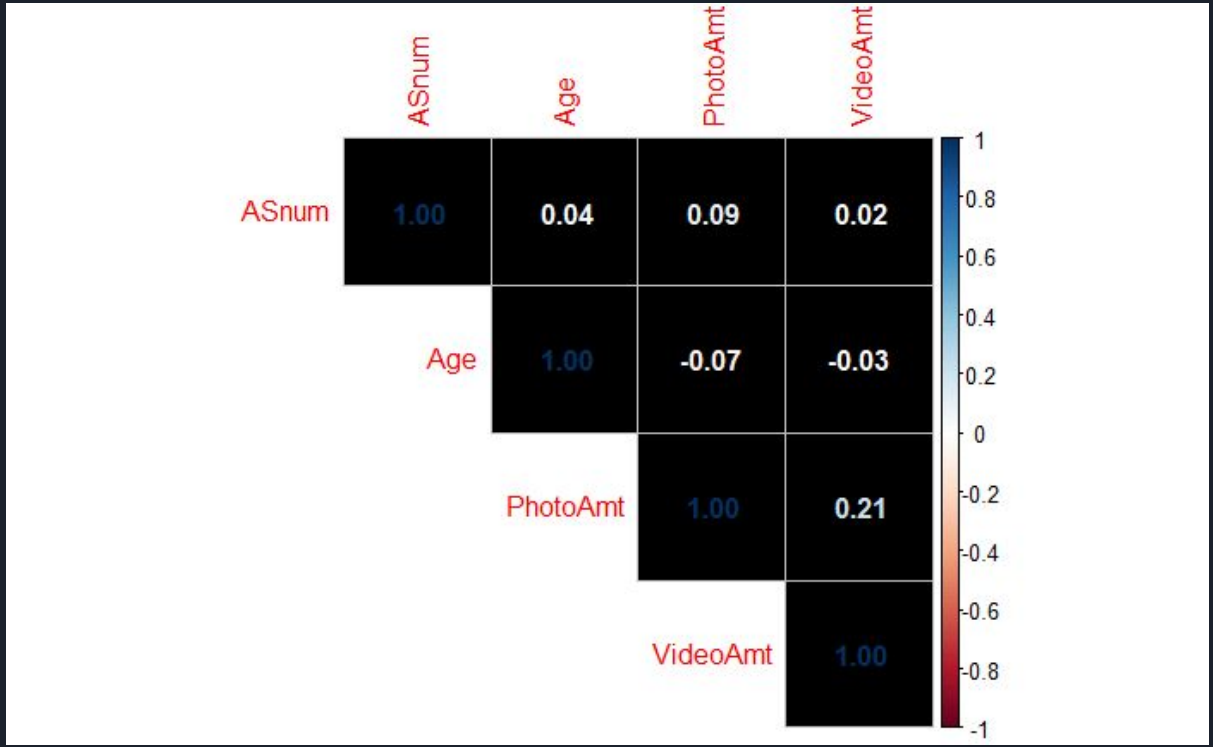

```
Analysis of Variance Table

Response: ASnum
             Df  Sum Sq Mean Sq F value Pr(>F)
Vaccinated    2   60442   30221    47.4 <2e-16 ***
Residuals  8482 5413338     638
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
$Vaccinated
      diff   lwr    upr p adj
2-1 -4.92 -6.31 -3.537 0.000*
3-1 -1.75 -3.85  0.352 0.125
3-2  3.17  1.10  5.249 0.001*
```

1. Age (in months)
2. PhotoAmt (number of photos)
3. VideoAmt (number of videos)

SMART: What numerical variables influence adoption speed?

Correlation Plot



Single Variable Numerical Models

ASnum ~ Age

Characteristic	Beta	95% CI[1]	p-value
(Intercept)	26	25, 27	<0.001
Age	0.05	0.02, 0.08	<0.001

[1] CI = Confidence Interval

ASnum ~ PhotoAmt

Characteristic	Beta	95% CI[1]	p-value
(Intercept)	24	23, 25	<0.001
PhotoAmt	0.73	0.56, 0.89	<0.001

[1] CI = Confidence Interval

ASnum ~ VideoAmt

Characteristic	Beta	95% CI[1]	p-value
(Intercept)	26	26, 27	<0.001
VideoAmt	1.7	0.05, 3.4	0.044

[1] CI = Confidence Interval

Conclusion: Age and number of photos both impact adoption speed, number of videos does not.

SMART: What combination of categorical and numerical variables result in the best predictive model?

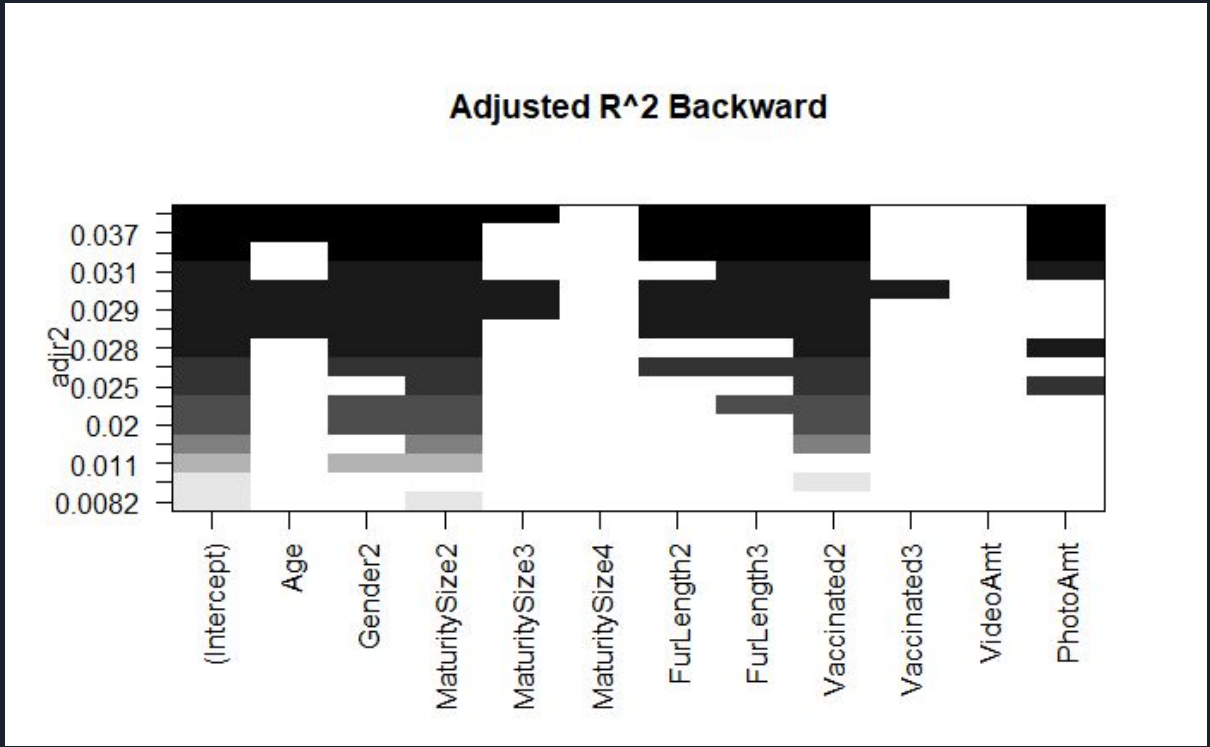
To the right: A full regression of all variables, categorical and numerical.

Characteristic	Beta	95% CI[1]	p-value
(Intercept)	22	20, 24	<0.001
Age	0.06	0.03, 0.10	<0.001
Gender			
1	—	—	
2	2.9	1.8, 4.0	<0.001
MaturitySize			
1	—	—	
2	5.0	3.7, 6.3	<0.001
3	-0.02	-2.2, 2.1	>0.9
4	5.6	-3.7, 15	0.2
FurLength			
1	—	—	
2	-3.0	-4.1, -1.8	<0.001
3	-7.8	-10.0, -5.6	<0.001
Vaccinated			
1	—	—	
2	-5.0	-6.2, -3.8	<0.001
3	-0.45	-2.2, 1.3	0.6
VideoAmt	0.08	-1.6, 1.7	>0.9
PhotoAmt	0.72	0.55, 0.89	<0.001

[1] CI = Confidence Interval

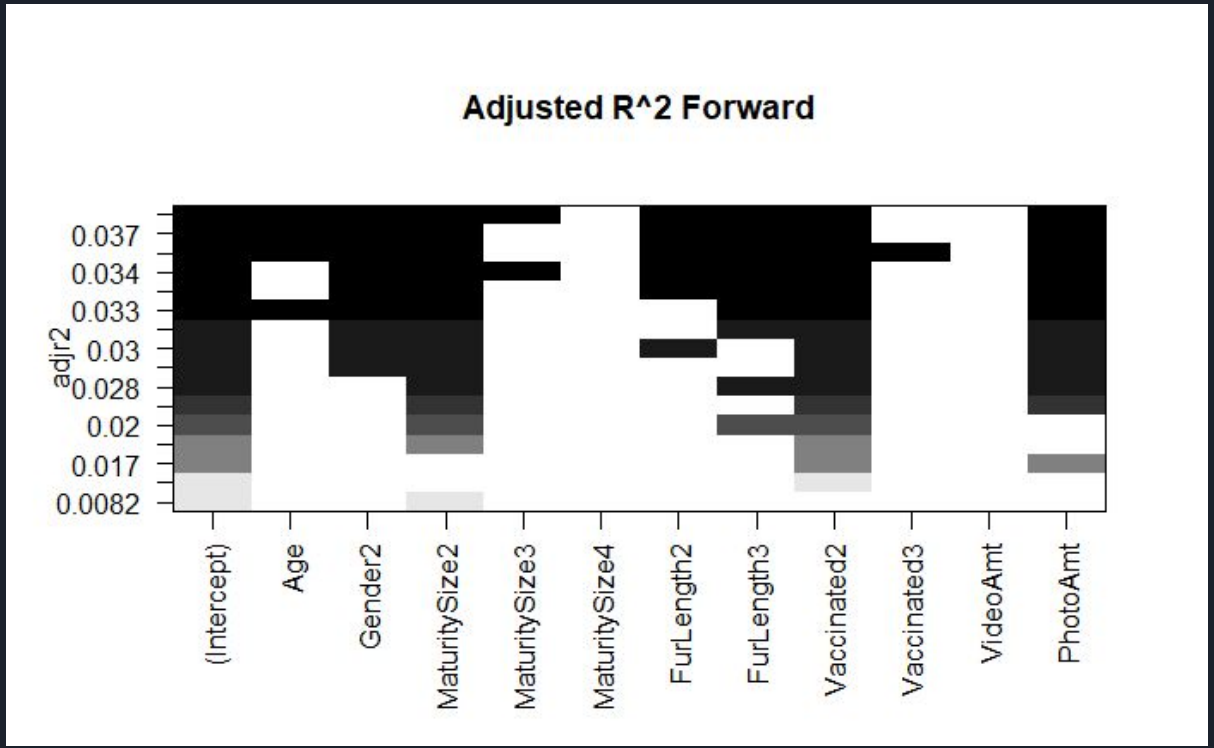
Using Feature Selection to find Best Model

Age, Gender 2, MaturitySize2, MaturitySize3, FurLength2, Furlength3, Vaccinated2, and PhotoAmt



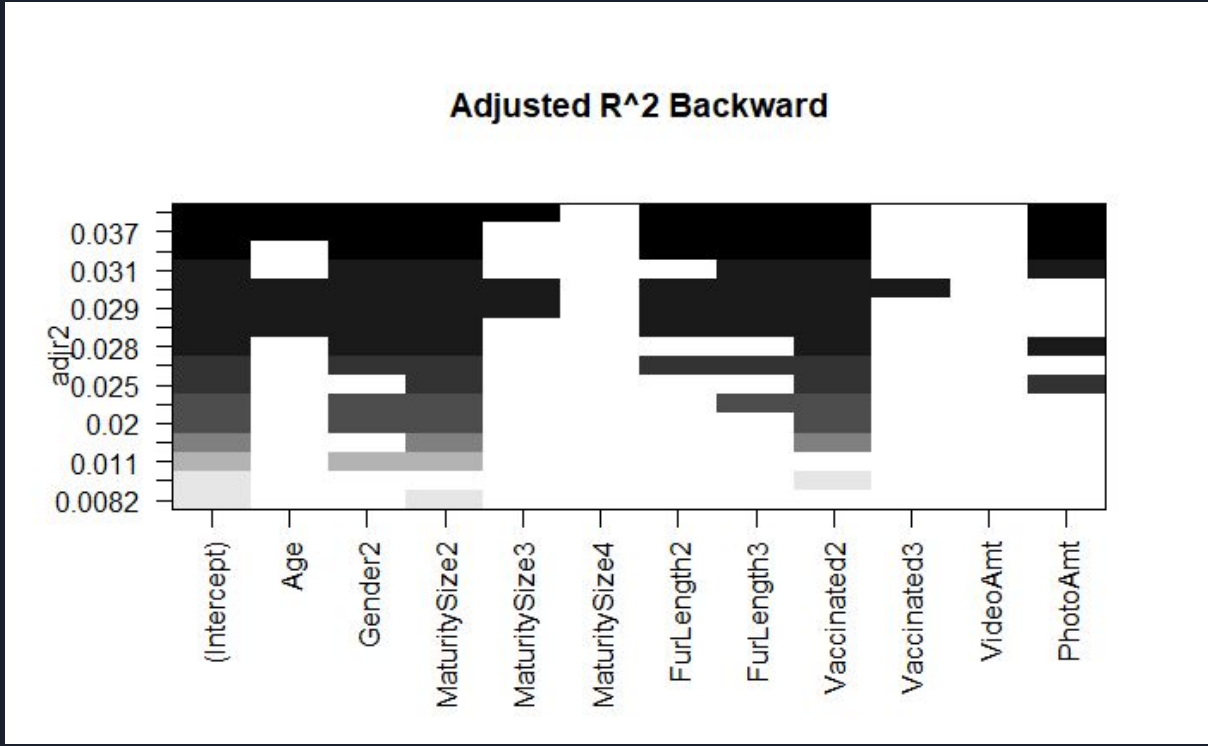
Feature Selection Forward

Age, Gender 2, MaturitySize2, MaturitySize3, FurLength2, Furlength3, Vaccinated2, and PhotoAmt



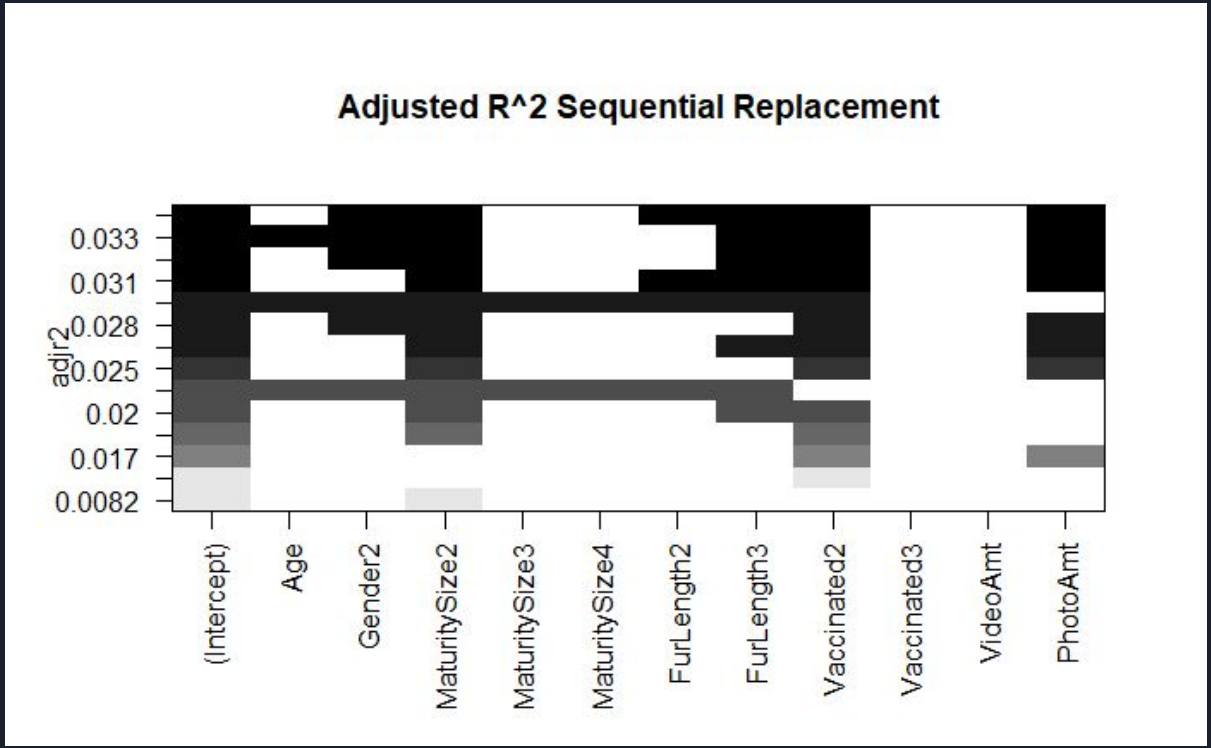
Feature Selection Backwards

Age, Gender 2, MaturitySize2, MaturitySize3, FurLength2, Furlength3, Vaccinated2, and PhotoAmt



Feature Selection Sequential

Gender2, MaturitySize2, FurLength2, FurLength3, Vaccinated2, PhotoAmt



Comparing Two Recommended Models

Characteristic	Beta	95% CI[1]	p-value
(Intercept)	21	19, 23	<0.001
Age	0.08	0.04, 0.12	<0.001
Gender			
1	—	—	
2	2.9	1.8, 4.1	<0.001
MaturitySize			
1	—	—	
2	5.2	3.8, 6.7	<0.001
3	1.1	-1.2, 3.5	0.3
FurLength			
1	—	—	
2	-2.8	-4.0, -1.6	<0.001
3	-6.1	-8.5, -3.6	<0.001
Vaccinated			
1	—	—	
2	-4.6	-5.8, -3.4	<0.001
PhotoAmt	0.78	0.61, 1.0	<0.001

[1] CI = Confidence Interval

Characteristic	Beta	95% CI[1]	p-value
(Intercept)	23	21, 25	<0.001
Gender			
1	—	—	
2	3.0	1.7, 4.2	<0.001
MaturitySize			
1	—	—	
2	4.9	3.5, 6.4	<0.001
FurLength			
1	—	—	
2	-2.8	-4.1, -1.5	<0.001
3	-5.6	-8.4, -2.9	<0.001
Vaccinated			
1	—	—	
2	-5.7	-6.9, -4.5	<0.001
PhotoAmt	0.72	0.54, 0.89	<0.001

[1] CI = Confidence Interval

Left Model: Recommended model from exhaustive, forward, and backwards methods.
Adjusted R-squared: 0.037

Right Model: Recommended by sequential method.

Adjusted R-squared: 0.0343

Conclusion to SMART Question 4:

Best model: ASnum ~ Age + Gender 2 + MaturitySize2 + MaturitySize3 + FurLength2 + Furlength3 + Vaccinated2 + PhotoAmt

Conclusion

1.Do dogs get adopted faster than cats?

- dogs get adopted slower than cats

2. Do physical attributes affect adoption speed?

- Gender,Size, Furlength, Vaccination does impact the adoption speed

3. What numerical variables influence adoption speed?

- Age and PhotoAmt are the only two numerical variables which alone can be considered statistically significant

4. What combination of categorical and numerical variables result in the best predictive model?

- ASnum ~Age + Gender 2 + MaturitySize2 + MaturitySize3 + FurLength2 + Furlength3 + Vaccinated2 + PhotoAmt

Conclusion

Strength

- Large sample size(n=14933 observations), more accurate finding
- Our best model include Age, Gender 2 , MaturitySize2 , MaturitySize3 + FurLength2 ,Furlength3, Vaccinated2, PhotoAmt allow us to accurately predict the adoption speed and improve internal validity

limitation

- The difference in variance does not allow us to draw causal relationship, so we can claim that there is association between predictors and outcomes.
- There might be unidentified confounders exists in causal pathway that interacts with outcome measure. Future study should explore other factors that may impact adoption speed.