Pet.finder 플랫폼 유기동물 입양속도 예측분석

INDEX

프로젝트 설명

개요

‘PetFinder.my’ Adoption Prediction

https://www.kaggle.com/c/petfinder-adoption-prediction

“ 유기동물 입양 플랫폼에 등록된 프로필을 바탕으로 입양 속도 예측 ”

평가 방법: Quadratic Weighted Kappa

* PetFinder.my는 말레이시아의 선도적인 동물 복지 플랫폼으로 15만 마리 이상의 동물 데이터베이스를 보유하고 있다.

데이터 설명

1. Trainset

X 변수

●PetID – 개별 동물 식별코드
●Type – 동물 종 *(1 = Dog, 2 = Cat)*
●Name – 이름
●Age – 나이(개월 수)
●Breed1 – 세부 품종 1
●Breed2 – 세부 품종 2
●Gender – 성별 *(1=Male, 2=Female, 3=group of pets)*
●Color1 – 색깔 1
●Color2 – 색깔 2
●Color3 – 색깔 3
●MaturitySize – 크기 *(1=Small, 2=Medium, 3=Large, 4=Extra Large)*
●FurLength – 털 길이 *(1=Short, 2=Medium, 3=Long)*
●Vaccinated – 백신 접종 여부 *(1=Yes, 2=No, 3=Not Sure)*
●Dewormed – 기생충 미감염 여부 *(1=Yes, 2=No, 3=Not Sure)*
●Sterilized – 중성화 수술 여부 *(1=Yes, 2=No, 3=Not Sure)*
●Health – 건강상태 *(1=Healthy, 2=Minor Injury, 3=Serious Injury)*
●Quantity – 마리 수
●Fee – 입양료 *(0 = Free)*
●State – 말레이시아 시
●RescuerID – 유기동물 구출자의 식별코드
●VideoAmt – 업로드된 동물 비디오 개수
●PhotoAmt – 업로드된 동물 사진 개수
●Description – 플랫폼에 작성된 동물 프로필

Y 변수

AdoptionSpeed

0 – 당일 입양됨

1 – 1 ~ 7일 사이에 입양됨

2 – 8 ~ 30일 사이에 입양됨

3 – 31 ~ 90 일 사이에 입양 됨

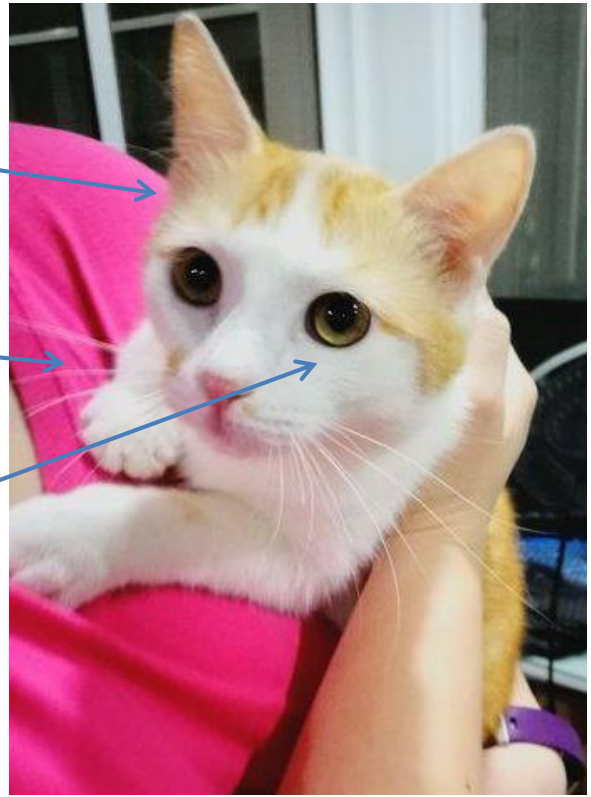
4 – 100일 이상 입양되지 않음

Metadata – Photo

2. Photo(.jpg) + Metadata (.json)

사진에 나타난 사물, 생물 등을 인식


```
{
    "mid": "/m/07k6w8",
    "description": "small to medium sized cats",
    "score": 0.9213904,
    "topicality": 0.9213904
},
{
    "mid": "/m/01l7qd",
    "description": "whiskers",
    "score": 0.91749674,
    "topicality": 0.91749674
},
{
    "mid": "/m/03071",
    "description": "cat like mammal",
    "score": 0.89707345,
    "topicality": 0.89707345
},
{
    "mid": "/m/014sv8",
    "description": "eye",
    "score": 0.80012083,
    "topicality": 0.80012083
},
```

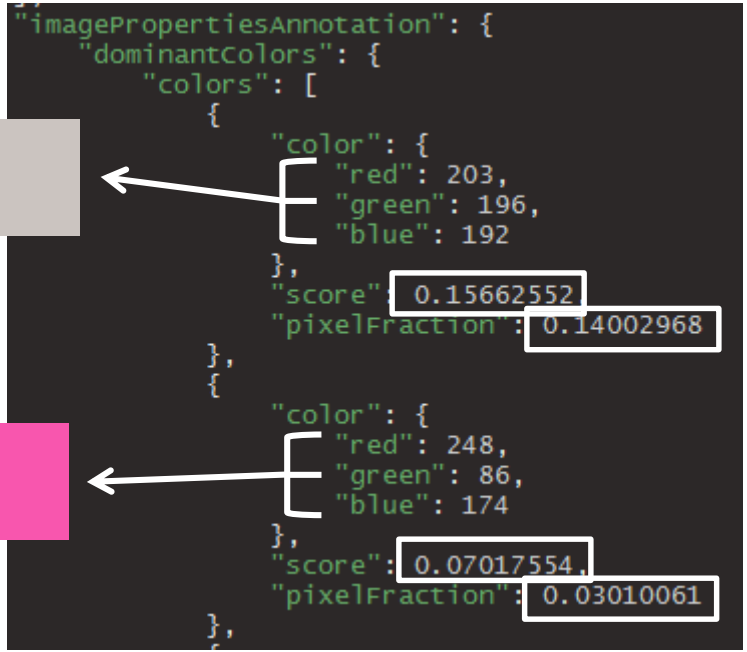



Score 값의 의미 ex 1) 사진 속에 '작은~중간 크기의 개'가 있다고 확신하는 정도: 92%

Metadata – Photo

2. Photo(.jpg) + Metadata (.json)

사진에 주요하게 나타나는 색을 인식

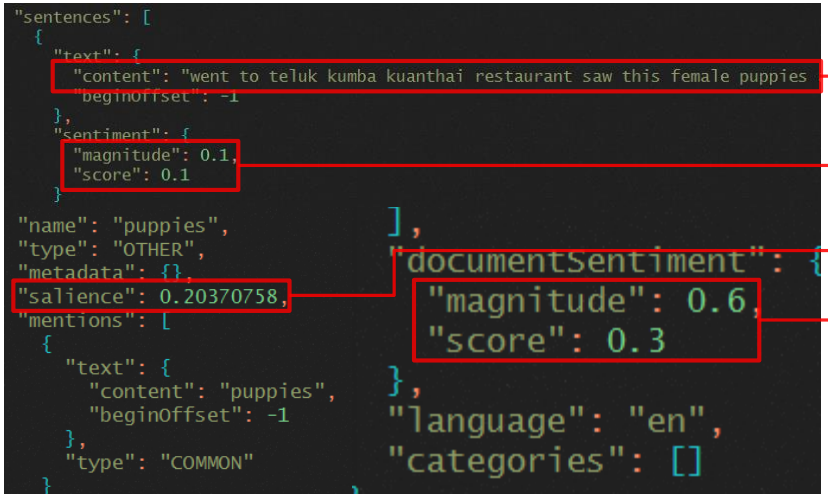


Score 값의 의미 해당 색이라고 확신하는 정도

pixelFraction 값의 의미 전체 픽셀에서 해당 색이 차지하는 비율

Metadata – Sentiment

3. Description Metadata about sentimental analysis using Google API (.json)



감성 분석 문장

위 문장 긍/부정 판단

단어 정확도

전체 문장 긍/부정 판단

모델 평가 방법

Quadratic Weighted Kappa

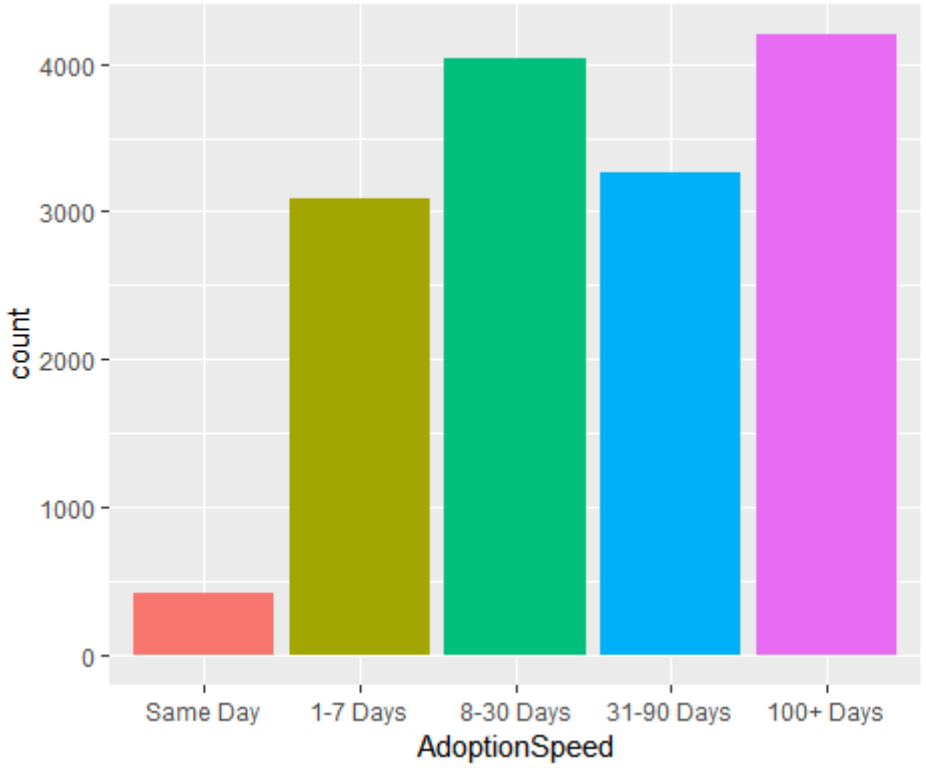
The definition of κ is:

$$\kappa \equiv \frac{p_o - p_e}{1 - p_e} = 1 - \frac{1 - p_o}{1 - p_e},$$

카파값은 정확한 예측과 더불어
일부 신뢰할 수 있는 레이블 간의 일치하는 양을 계량하는 데 사용 함

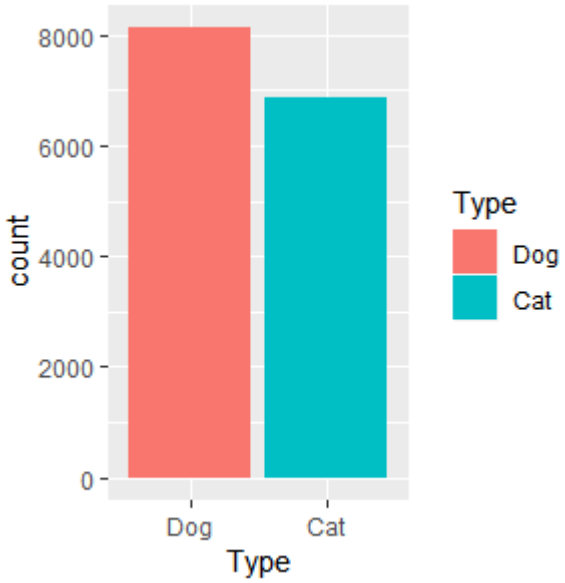
탐색적 데이터 분석

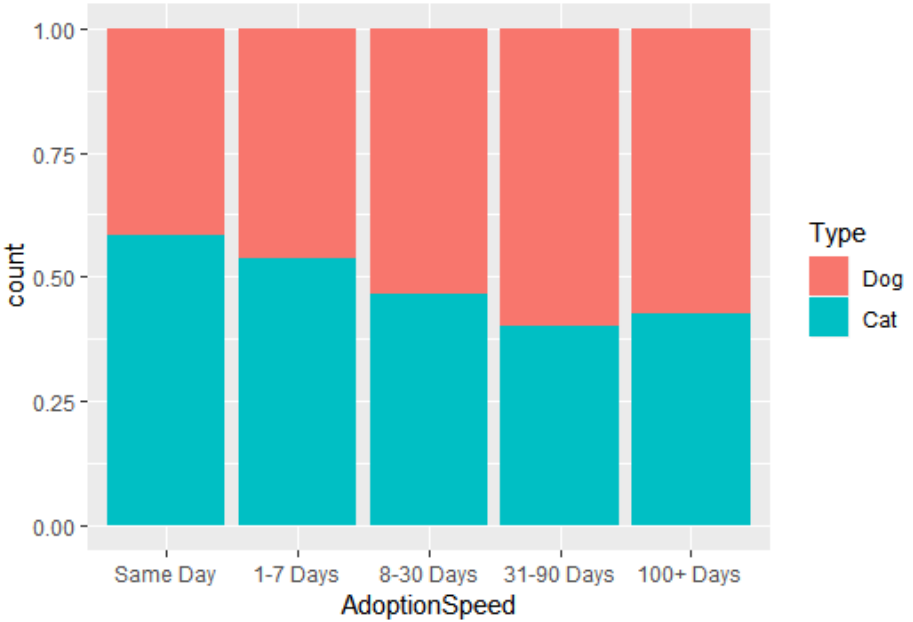
Adoption Speed



Adoption Speed	Pet Amount
Same Day	410
1-7 Days	3090
8-30 Days	4037
31-90 Days	3259
100+ Days	4197

Type : Dog / Cat

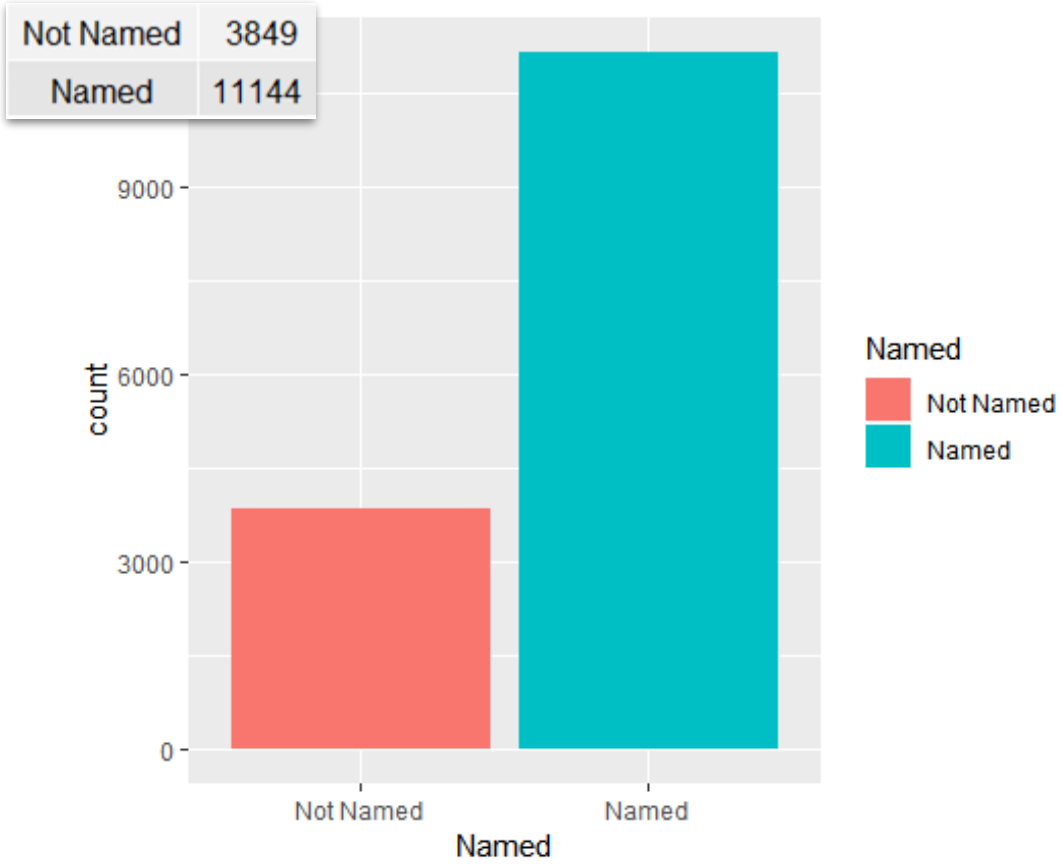




Dog 8132 > Cat 6861

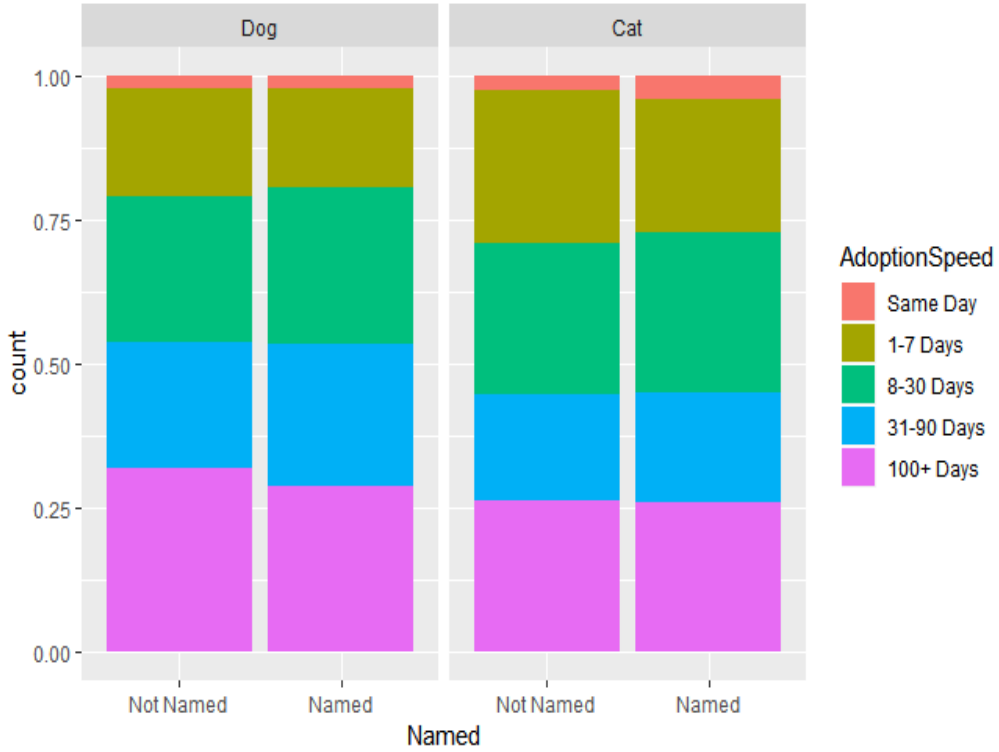
고양이의 입양 속도가 빠른 경향

Named / Not Named



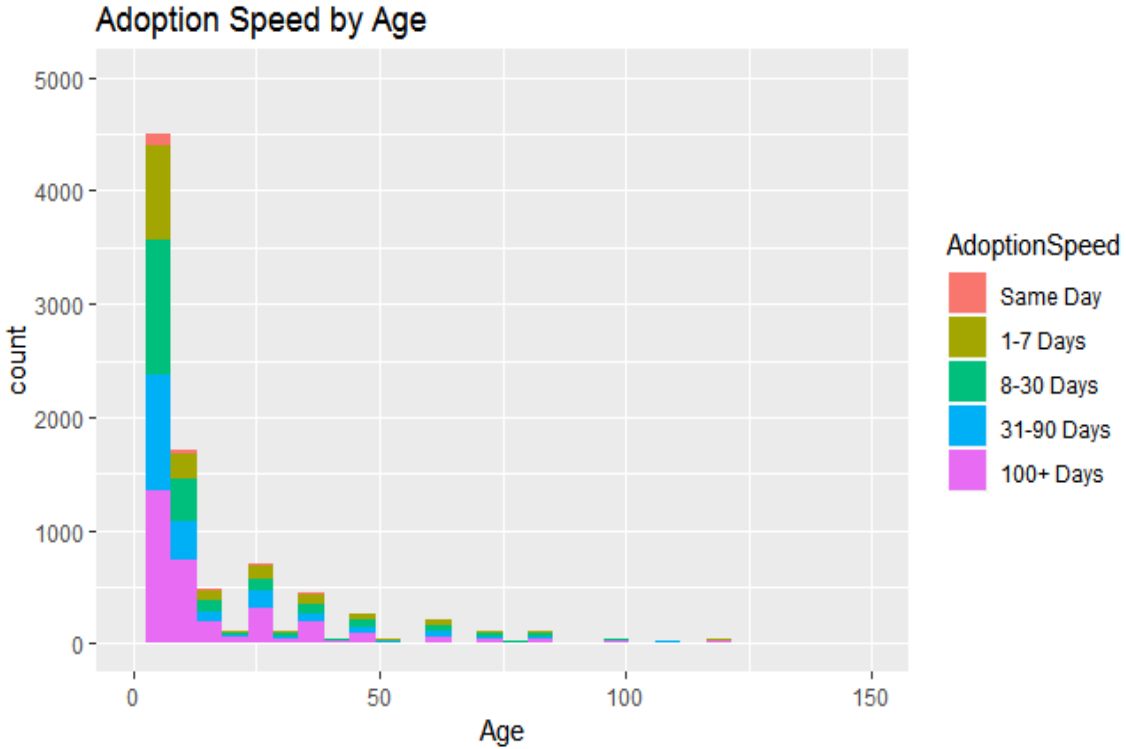
- 단순 종 표기, 나이 등 이름이 아닌 데이터 정제 (ex. 2 Mths old Cute Kitties)
- 이름이 있는지 여부가 감정적으로 어필해 입양 속도에 영향을 미치는가?

Named / Not Named



개/고양이 모두 이름이 있는지 여부가 입양 속도에 거의 영향을 미치지 않는다.

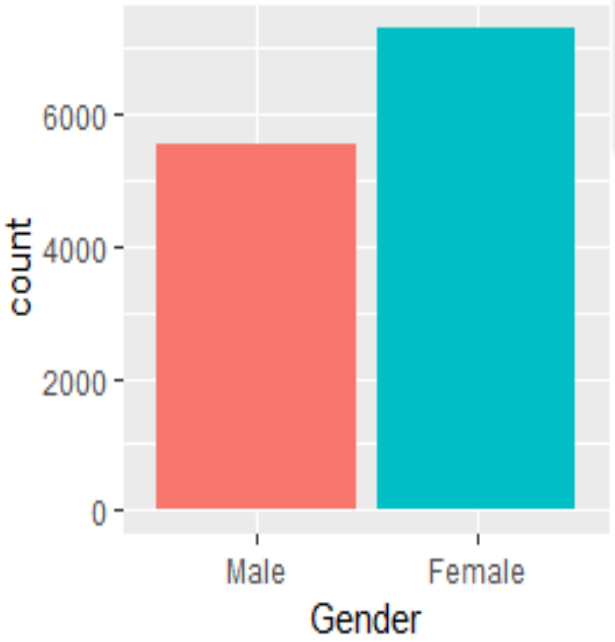
Age



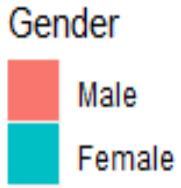
생후 5개월 전후의 어린 개체 수가 많음

어릴수록 입양 속도 빠름

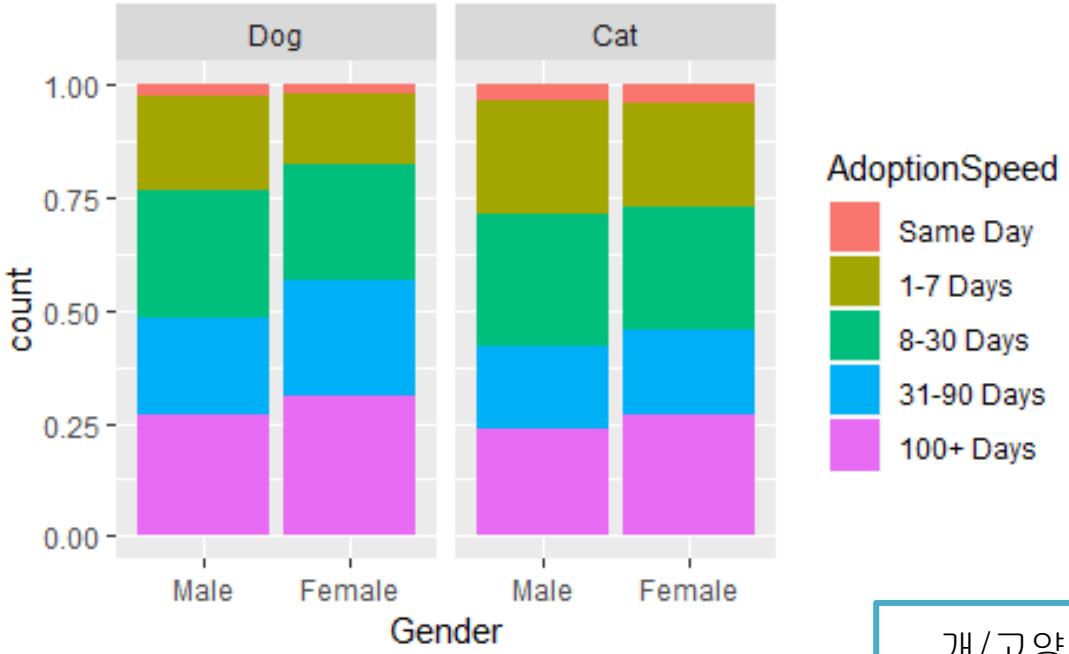
Gender



Gender	Pet Amount
Male	5536
Female	7277
Mixed	2180

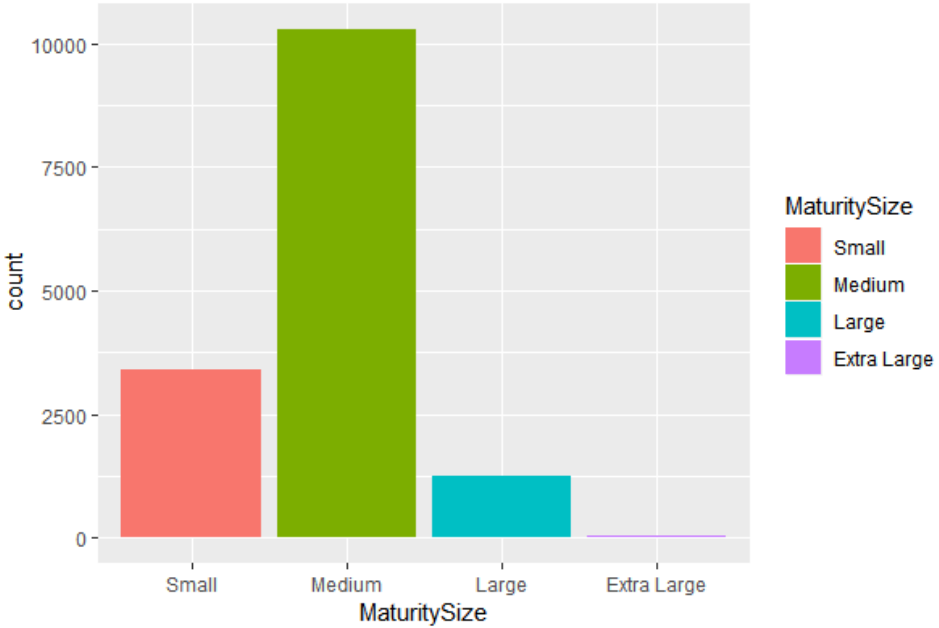


Gender

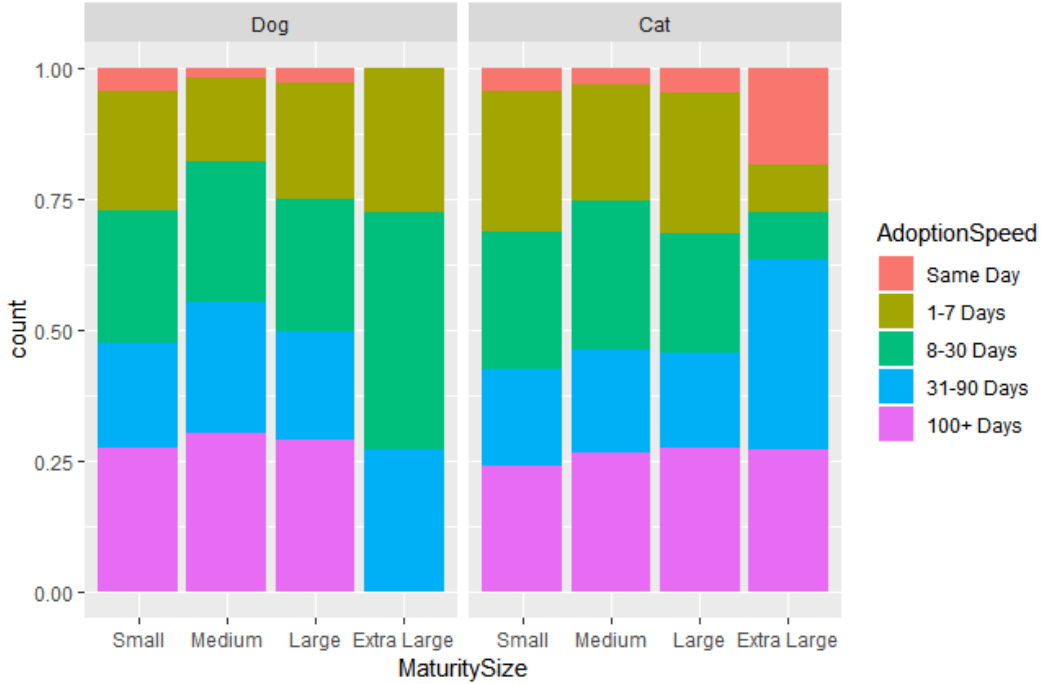


개/고양이 모두
수컷의 입양 속도가
빠름

Maturity Size



Maturity Size

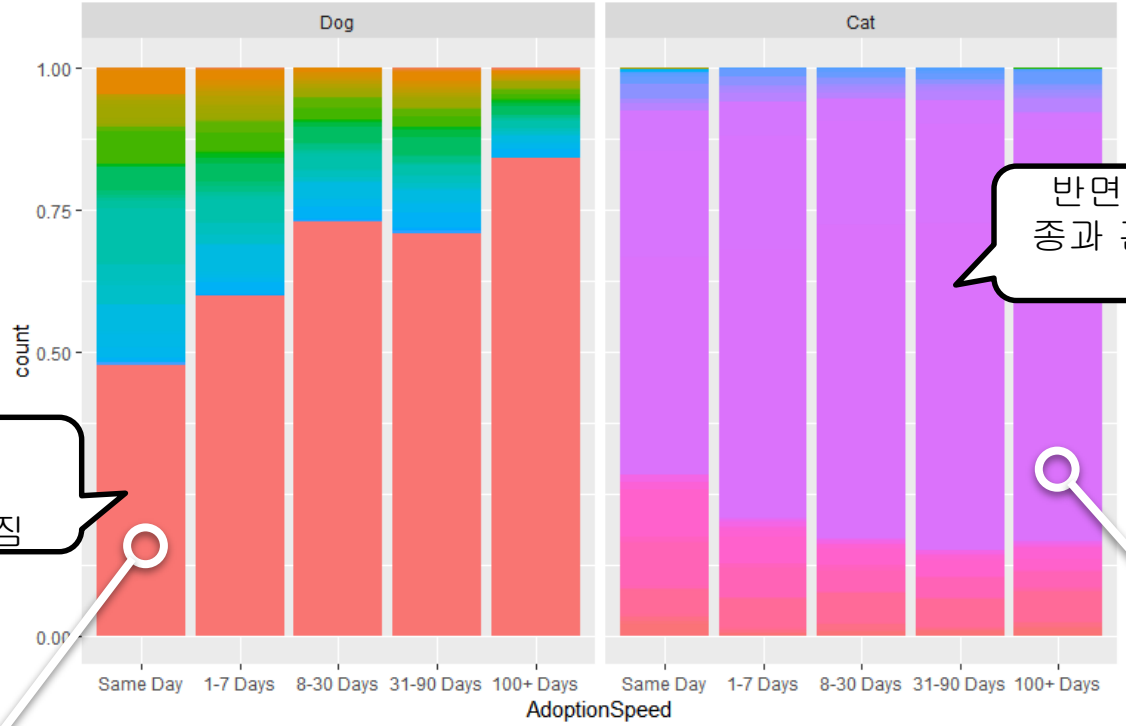


Extra Large Dog
100+ Days 0

Extra Large Cat
Same Day,
100+ Days 비중 큼

뚜렷한 선호/비선호 작용

Breed



입양속도가
늦어질수록
잡종견의 비중 커짐

반면 고양이의 경우
종과 관련없이 일정한
패턴

Mixed Breed
(Dog)

Domestic
Short Hair
(Cat)

* 모든 그래프에서 잡종의 비중이 높은 이유는 선호도가 아니라 잡종의 절대수가 많기 때문

Breed

TOP 10 BREEDS

	Breed Name	Pet Amount	Type
1	Mixed Breed	5927	Dog
2	Domestic Short Hair	3634	Cat
3	Domestic Medium Hair	1258	Cat
4	Tabby	342	Cat
5	Domestic Long Hair	296	Cat
6	Siamese	264	Cat
7	Persian	221	Cat
8	Labrador Retriever	205	Dog
9	Shih Tzu	190	Dog
10	Poodle	167	Dog

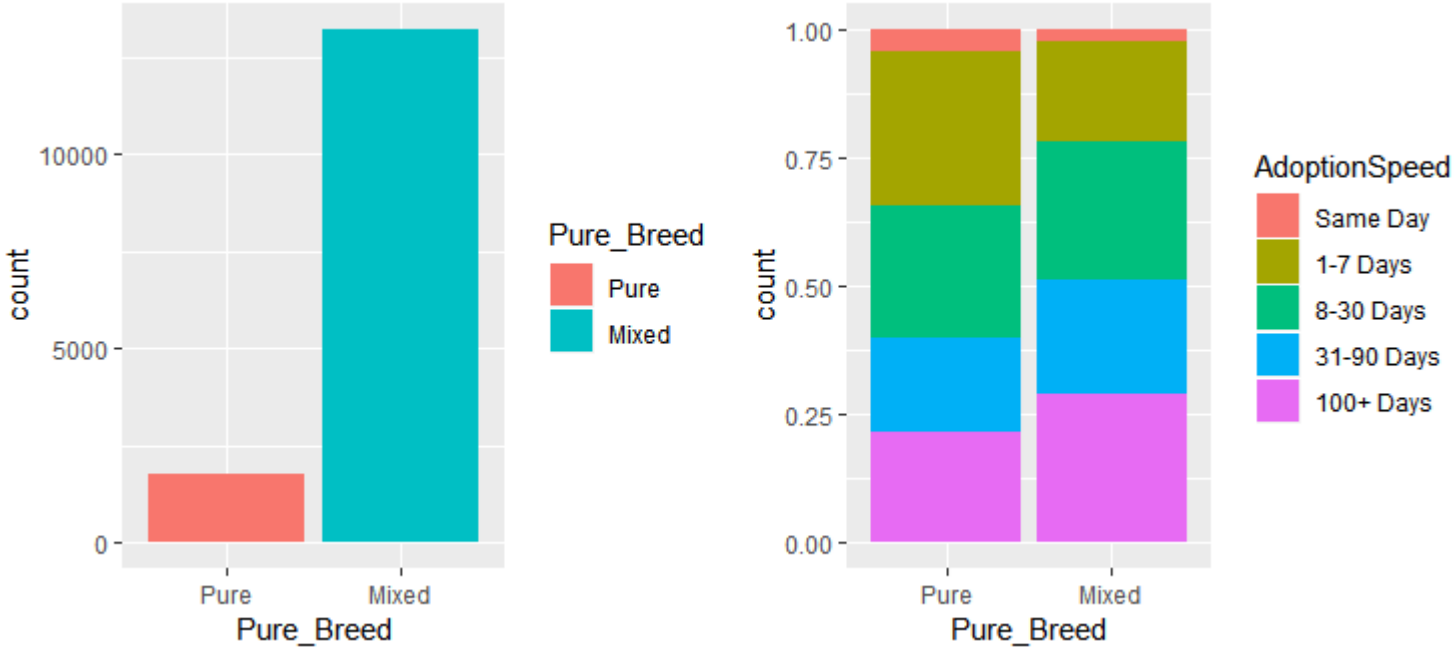
상위 5개 종 모두 잡종

* 선호도보다 절대 개체 수의 영향이 높음

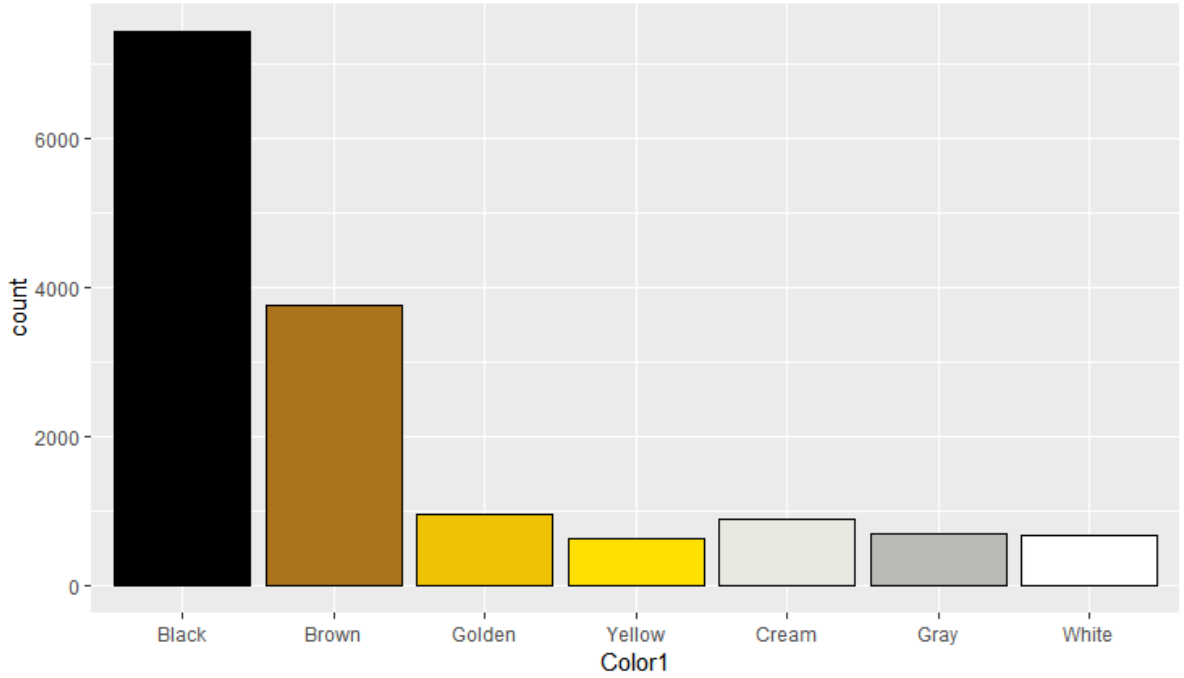
Pure Breed vs Mixed Breed

- 순수종

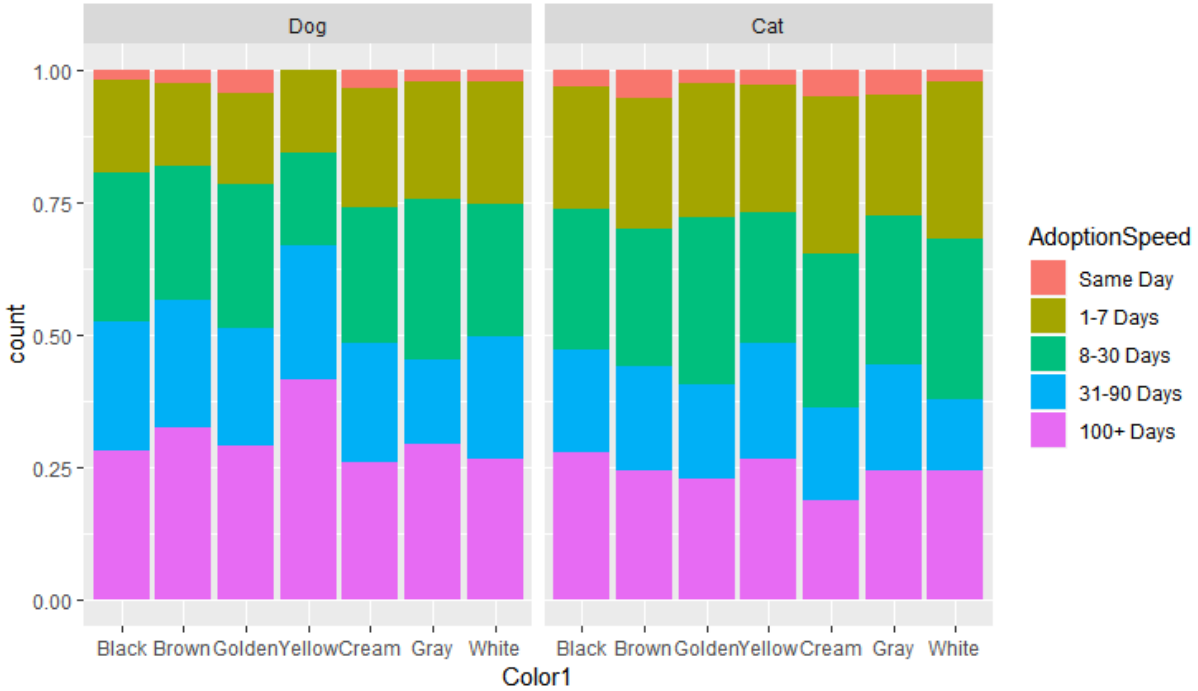
- 비순수종



Color

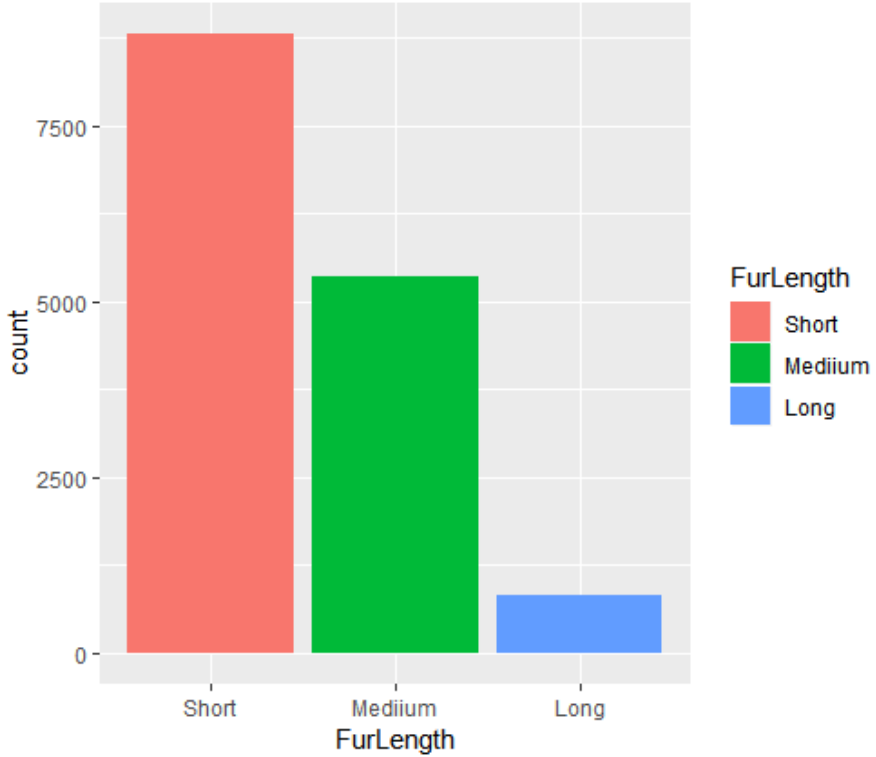

6000
4000
2000
0
count
Black
Brown
Golden
Yellow
Cream
Gray
White
Color1

Color



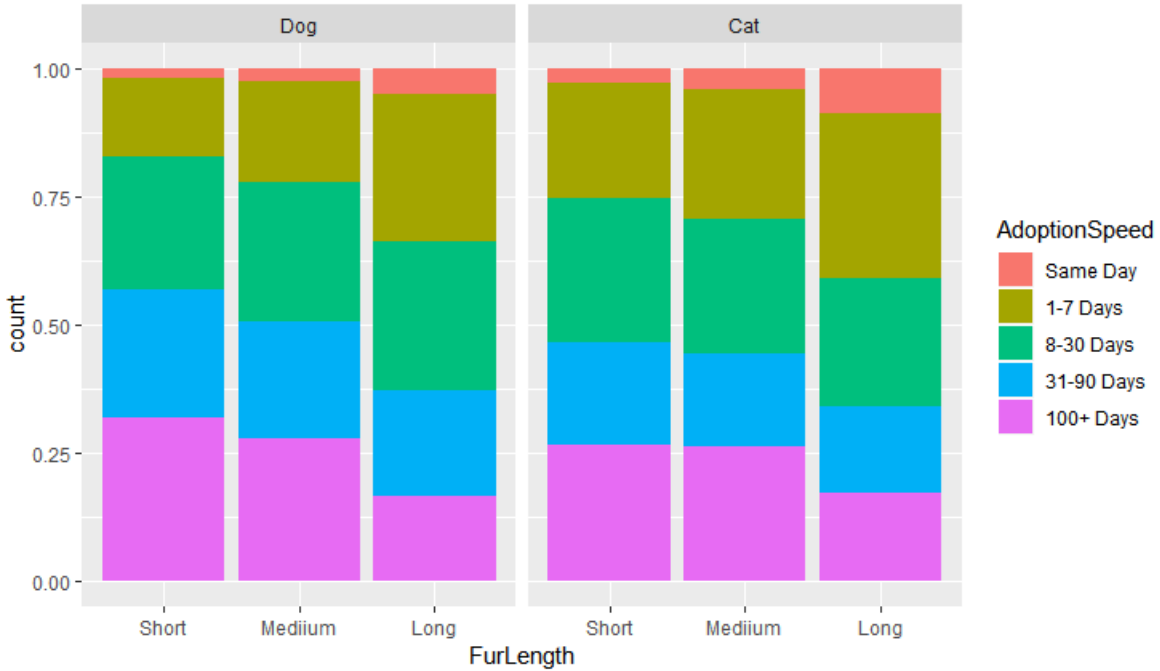
개/고양이 모두 Cream색인 경우 입양 속도가 빠름

Fur Length



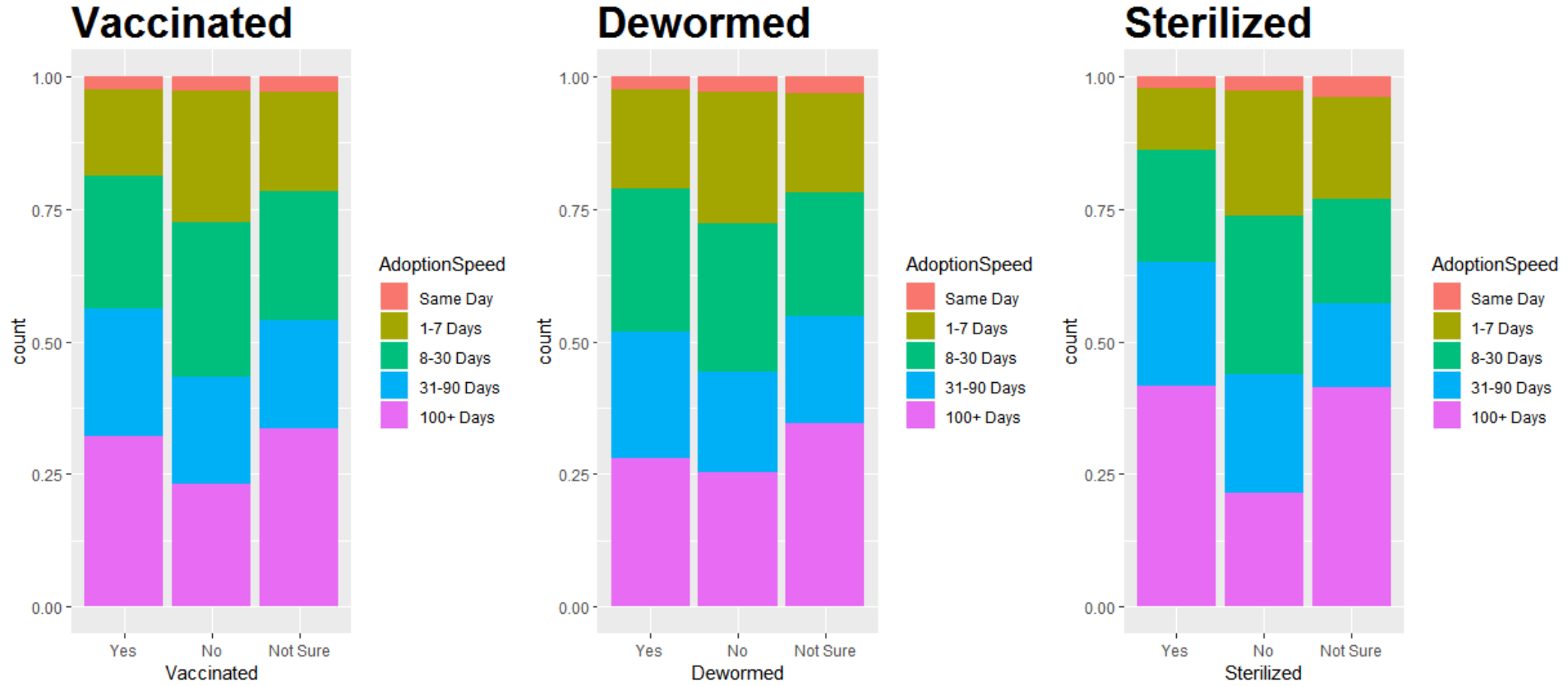
Short > Medium > Long

Fur Length



개/고양이 모두 장모종의
입양속도가 가장 빠름

Health Condition



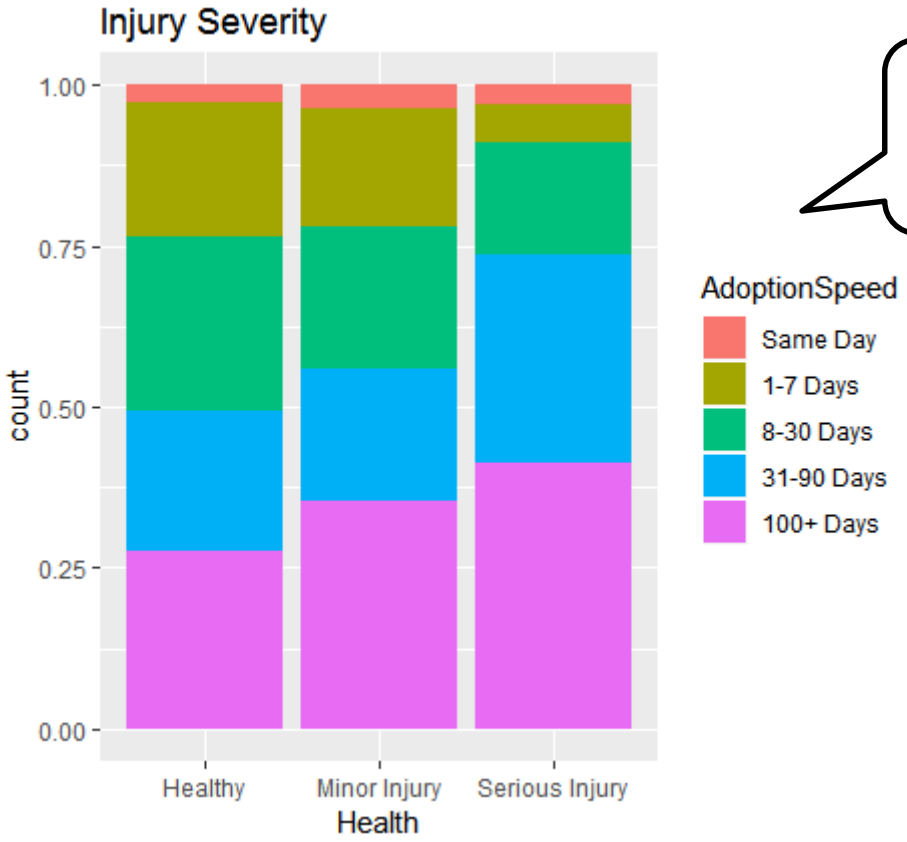
Health Condition

예상
당연히 YES의 선호도가 높을 것

실제
NO의 선호도가 더 높음

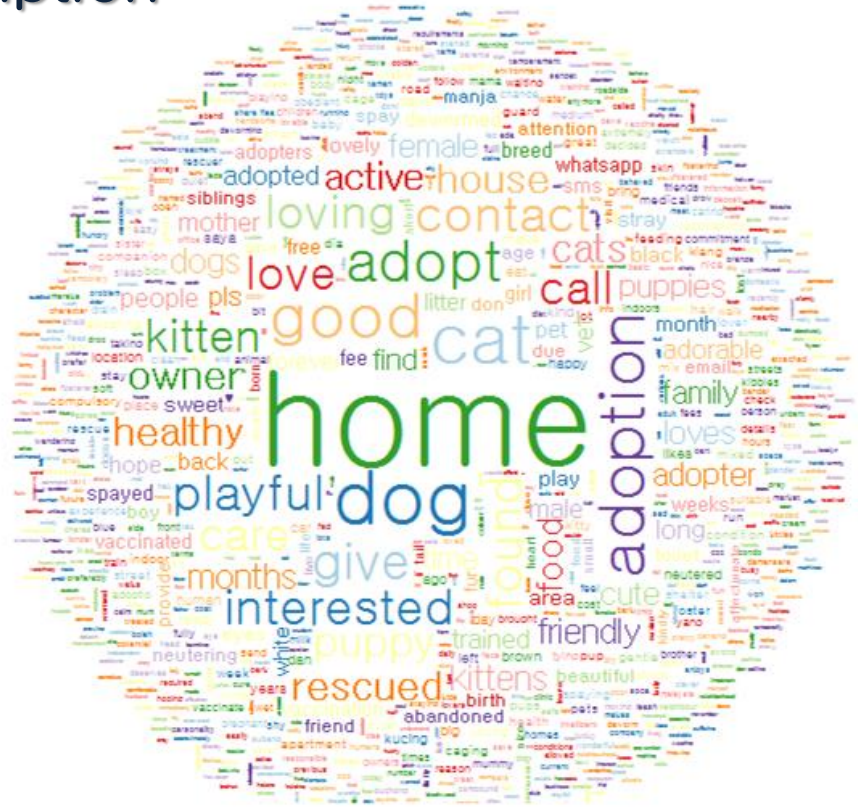
왜?
연령이 Y변수에 영향이 큰 데서 미루어 볼 때
백신 접종 등을 맞기에 아직 어린 경우일 것으로 추측

Health Condition



부상이 심할 수록
입양 속도가
늦어짐

Description



Profile – Description
토픽분석

Adoption Speed 별
상위 단어 차이 적음

Top 5 words

Home	7732
Dog	4517
Cat	4074
Adoption	3647
good	3562

가장 입양이 빠르게 될 확률이 높은 종

분류	결과
Age	young
Type	cat
Gender	male
Name	Named
Maturity Size	Extra Large
Breed	dog는 순종 cat은 품종에 큰 편차 없음
Color	cream color
Fur Length	장모종

-동물 선호도: 이름이 있는 male cat
나이가 어리고 Extra Large
cream color의 장모종

-품종: 스코티쉬폴드

데이터 전처리

파생변수

1. 사진 메타데이터에서 77개 파생변수 생성

1) 동물 신체 관련
whiskers, snout, paw, tail, black mouth cur,
fur, eye, ear, leg, claw, dog clothes

2) 장소 관련
flooring, floor, grass, lawn, animal shelter,
cage, window, furniture

3) 클로즈업 여부
close up

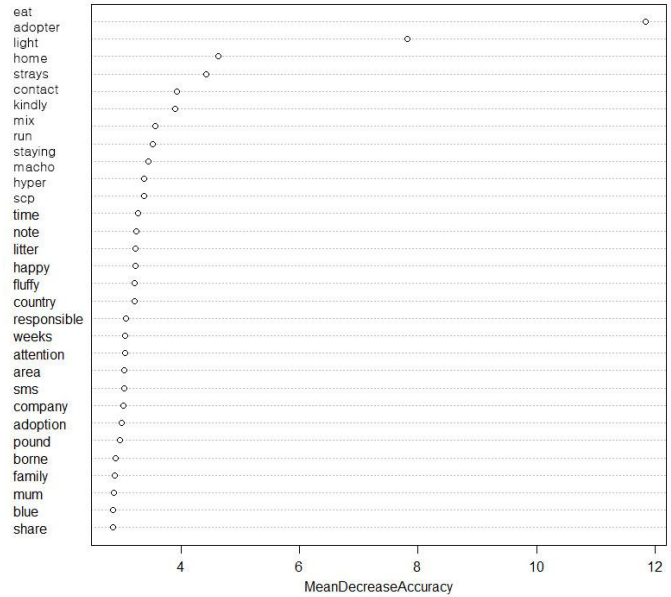
4) 사진 편집 여부
photo caption

5) 사진 색깔 관련
fawn, green, pink, red, black, purple,
blue, white, yellow, black and white

6) 동물 종 및 동물 그룹 (출현 빈도 50회 이상)
sporting group, rare breed dog, companion dog,
puppy, kitten, black cat, dog crate, fauna,
european shorthair, aegean cat, dragon li,
retriever, canaan dog, korat, american shorthair,
huntaway, pariah dog, borador, miniature fox terrier,
jack russell terrier, khao manee, russell terrier,
turkish van, domestic long haired cat,
carolina dog, pinscher, bombay, kennel,
siamese, balinese, burmese, terrier, american wirehair,
chihuahua, feist, norwegian forest cat, ragdoll,
kunming wolfdog, german shepherd dog,
california spangled, potcake dog, turkish angora,
hunting dog, golden retriever, australian mist, schnoodle

파생변수

2. Description 및 감성분석 메타데이터에서 16개 파생변수 생성



1) Description

eat, adopter, light, home, contact, kindly,
mix, run, staying, macho, hyper

(trainset의 description만으로 AdoptionSpeed(y변수)를 randomforest 분류분석한 뒤, 분류 정확도에 기여하는 주요 단어 몇개를 추출하여 파생변수로 사용)

2) Description 감성분석 메타데이터

num_of_sentences (문장 개수)
1st_pos_sentiment (긍정적 문장 중 최고점)
1st_neg_sentiment (부정적 문장 중 최고점)
tot_sentiment (Description의 총 감성점수)

3. Trainset의 name 변수로 이름 유무에 관한 1개 파생변수 생성

NA 처리

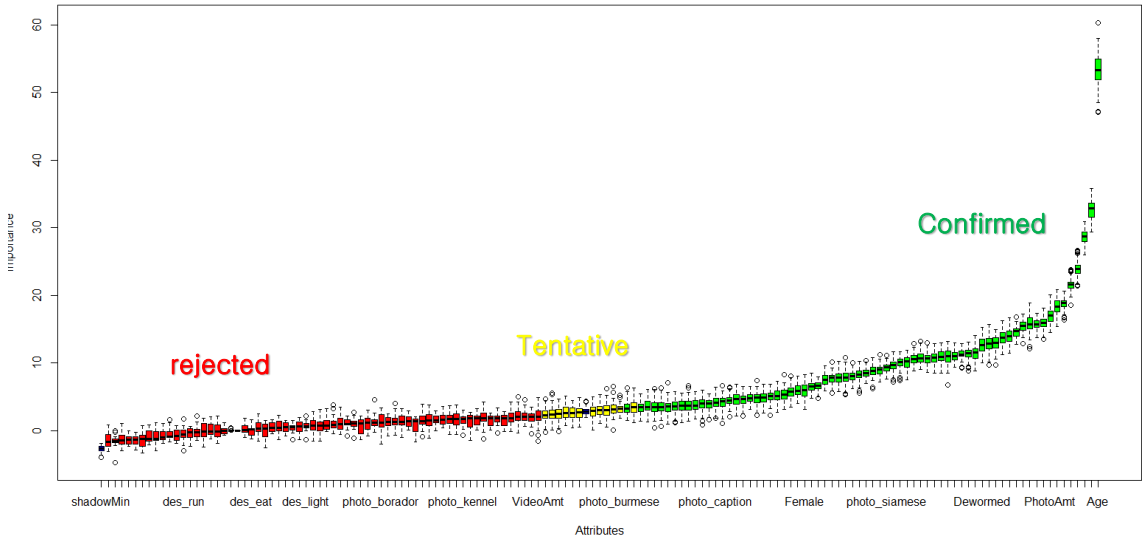
des_num_of_sentences	des_1st_pos_sentiment	des_1st_neg_sentiment	des_tot_sentiment
Min. : 1.000	Min. :0.0000	Min. :-0.8100	Min. :-1.8900
1st Qu.: 2.000	1st Qu.:0.2500	1st Qu.:-0.1600	1st Qu.: 0.0400
Median : 4.000	Median :0.6400	Median : 0.0000	Median : 0.4200
Mean : 5.116	Mean :0.5427	Mean :-0.1171	Mean : 0.6147
3rd Qu.: 7.000	3rd Qu.:0.8100	3rd Qu.: 0.0000	3rd Qu.: 0.8500
Max. :84.000	Max. :0.8100	Max. : 0.0000	Max. : 9.1000
NA's :551	NA's :551	NA's :551	NA's :551

Description 감성 분석 메타데이터가 없는 레코드에 결측치 발생
=> 중앙값으로 대체

모 델 링

변수 선택

1. Boruta 패키지를 이용하여 변수 선택



모델링에 사용할 변수 목록 (Confirmed/Tentative)

Type, RescuerID, Named, Age, Male, Female
Pure_Breed, Single_Color, Breed1 Breed2, Color1_Black
Color1_Brown, Color1_Cream, Color1_Gray Color1_White,
Color2_Brown, Color2_Yellow, Color2_Gray, Color3_White,
MaturitySize_small, MaturitySize_medium,
MaturitySize_large, FurLength_short,
FurLength_medium, FurLength_long, Vaccinated,
Dewormed, Sterilized, Health_healthy,
Health_minor_injury, State_Johor, State_Kelantan,
State_Kuala_Lumpur, State_Melaka,
State_Negeri_Sembilan, State_Perak, State_Pulau_Pinang,
State_Sabah, State_Selangor,
photo_animal_has_whiskers, photo_animal_has_snout,
photo_animal_has_paw, photo_animal_has_tail,
photo_animal_has_fur, Quantity, Fee, VideoAmt,
PhotoAmt, des_num_of_sentences,
des_1st_pos_sentiment, des_1st_neg_sentiment
…. 등 79개

Stacking Ensemble

2) h2o.gbm

trainset을 이용하여 my_gbm 모델 생성

```
my_gbm <- h2o.gbm(x = x,
                  y = y,
                  training_frame = train,
                  distribution = "multinomial",
                  ntrees = 150,
                  max_depth = 10,
                  min_rows = 2,
                  learn_rate=0.01,
                  nfolds = nfolds,
                  fold_assignment = "Modulo",
                  keep_cross_validation_predictions = TRUE,
                  seed = 1)
```

testset 예측
=> 혼동행렬 생성

	real_value				
predicted	0	1	2	3	4
0	4	15	6	2	0
1	48	296	250	165	117
2	28	372	516	331	230
3	15	81	159	205	95
4	38	184	288	257	796

wkappa(t_gbm) = 0.3430419

Stacking Ensemble

2) h2o.randomForest

trainset을 이용하여 my_rf 모델 생성

```
my_rf <- h2o.randomForest(x = x,
                          y = y,
                          training_frame = train,
                          ntrees = 200,
                          nfolds = nfolds,
                          fold_assignment = "Modulo",
                          score_each_iteration = T,
                          keep_cross_validation_predictions = TRUE,
                          seed = 1)
```

testset 예측
=> 혼동행렬 생성

	real_value				
predicted	0	1	2	3	4
0	3	1	0	0	0
1	41	275	231	130	98
2	32	377	515	321	205
3	13	70	140	216	68
4	44	225	333	293	867

wkappa(t_rf) = 0.3391329

Stacking Ensemble

3) h2o.deeplearning

trainset을 이용하여 my_deep 모델 생성

```
my_deep<- h2o.deeplearning(x = x,
                           y = y,
                           training_frame = train,
                           nfolds=nfolds,
                           distribution = "multinomial",
                           activation = 'RectifierWithDropout',
                           hidden=c(30),
                           epochs = 1000,
                           score_validation_samples=10000,
                           stopping_rounds=2,
                           stopping_metric="misclassification",
                           stopping_tolerance=0.01,
                           fold_assignment = "Modulo",
                           keep_cross_validation_predictions = T)
```

testset 예측
=> 혼동행렬 생성

	real_value				
predicted	0	1	2	3	4
0	0	0	1	0	1
1	47	252	193	113	88
2	42	446	541	396	292
3	8	58	130	150	95
4	36	192	354	301	762

wkappa(t_deep) = 0.3205916

Stacking Ensemble

4) gbm, randomforest, deeplearning => ensemble

trainset을 이용하여 my_deep 모델 생성

```
ensemble <- h2o.stackedEnsemble(x = x,
                                y = y,
                                training_frame = train,
                                model_id = "my_ensemble",
                                base_models = list(my_rf, my_gbm, my_deep))
```

testset 예측
=> 혼동행렬 생성

```
          real_value
predicted   0   1   2   3   4
        0   2   1   0   0   0
        1  53 305 256 148 112
        2  30 390 522 332 204
        3  10  62 135 221  84
        4  38 190 306 259 838
>
```

wkappa(t_en) = 0.363616

결 론

진행 과정의 어려움

- 주어진 데이터셋으로 만든 모델의 정확도가 낮았다.
- 파생변수를 포함해 적절한 변수를 선택하는 것이 어려웠다.
- 하드웨어 사양 문제로 코드 실행이 어려웠다.
- 시간 분배에 미숙했다.

개선이 필요한 사항

- 데이터 셋 특성에 맞는 모델을 선택
- 적합한 파라미터 값 조정
- 실행 시간을 고려한 프로젝트 계획 설계

THANK YOU