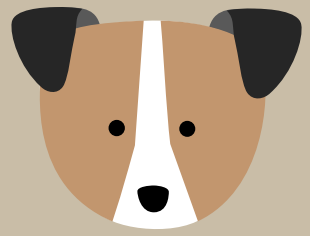



유기동물 입양 예측

3팀 선형대수학

황정현 고경현 김지민 반경림 전효림

목차

데이터셋 완성

추가 전처리 | 변수 검정 | 최종 데이터셋

품종 관련 파생 변수 추가

kind_spec	size_google	size_akc	group_akc	activity_level	group_FCI
믹스견	모름	Unknown	mixed	Unknown	mixed
치와와	소	XS	Toy	Regular exercise	Companions
보더콜리	중	M	Herding	Needs lots of activity	Sheepdogs
킹찰스스파니엘	소	S	Toy	Calm	Companions
요크셔테리어	소	XS	Toy	Regular exercise	Terriers
닥스훈트	중	S	Hound	Regular exercise	Dachshunds
도사	대	L	Others	calm	Pinscher
...	...	...	...	...	...

추후 공개됩니다아아~~

비슷한 특성으로 분류된 변수들은 **검정**으로 적절한 변수 선택 예정

데이터셋 완성

텍스트 데이터 다시

텍스트 데이터(특징) 처리 플로우

질병사전과 긍부정사전은
서로 겹치는게 없도록 처리해줬답니다아

new! 특성 변수

new! 긍정 사전

new! 부정 사전

new! 질병 사전

데이터 정리
숫자 / 영어 / 특수문자 / 공백 제거

맞춤법 정리
띄어쓰기
맞춤법 검사

형태소 분석
(Khaiii 패키지)
명사, 형용사, 동사 선택

필터링
한 글자 단어 제거

데이터셋 완성

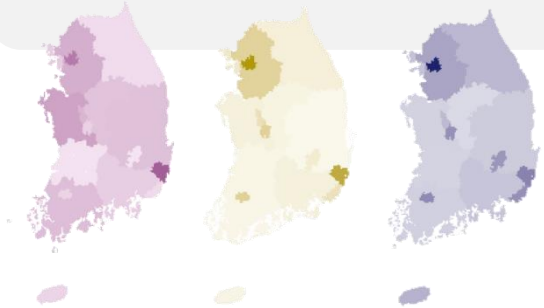
추가 전처리 | 변수 검정 | 최종 데이터셋

지역적 특성 데이터 정제

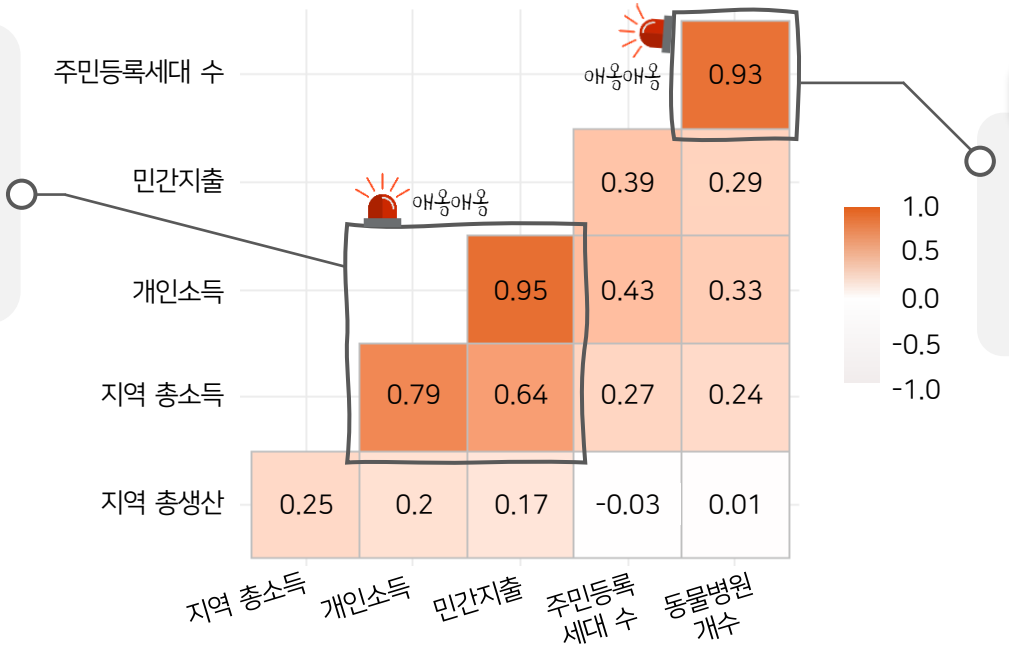
시군구	1인당 지역 내 총생산 (단위: 백만)	1인당 지역 총소득 (단위: 천)	1인당 개인소득 (단위: 천)	1인당 민간지출 (단위: 천)	주민등록세대 수	동물병원 개수
부산광역시 중구	70.16536	29388.16	19680.32	18029.58	23847	3
부산광역시 서구	28.67351	29388.16	19680.32	18029.58	53853	6

경 제

1인당 지역 총소득
1인당 개인소득
1인당 민간지출

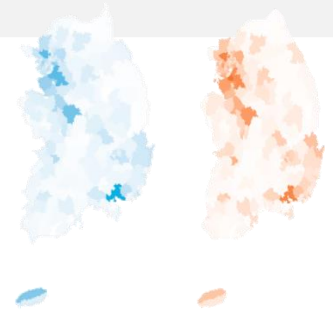


경제 지표 | 수요 지표



수 요

주민등록세대 수
동물병원 개수



데이터셋 완성

범주형 변수 선택

size_google group_FCI

지금까지 강아지 품종 관련 파생 변수 5가지 중 2가지를 제거

같은 기관에서 얻어낸 나머지 품종 관련 변수 세 가지에 대해서도 서로 독립성 검정 진행

헉시...?

	P - value	Cramer's V
품종_크기(AKC) | 품종_분류(AKC)	< 0.001	0.716
품종_크기(AKC) | 품종_활동성	< 0.001	0.702
품종_분류(AKC) | 품종_활동성	< 0.001	0.687

서로 독립도 아니고 심지어 상관도 크다

[크기] [분류] [활동성]

변수 중 하나를 선택해야 하는 상황

	입양 변수와의 Cramer's V 값
품종_크기 (AKC)	0.283
품종_분류 (AKC)	0.305
품종_활동성	0.259

분류 모델링

전체적인 흐름

데이터 처리	연속형 변수	스케일링
	범주형 변수	인코딩
모델링	5 fold CV	
	파라미터 튜닝	최적의 파라미터 조합
최종 모델 선정	F1, acc	

모델링을 할 때, 숫자가 아닌 독립변수는 사용할 수 없어서지이~~

현재 데이터에 불균형이 존재해서 F1을 우선순위로 했어에!!!

결과 해석

로지스틱 회귀 | 랜덤포레스트

로지스틱 회귀 모델 해석

$$\ln\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon$$

$$OR = e^{\beta_p}$$

		오즈비	0.025	0.975	p-value
양적	1인당 경제 지표	1.1138306	1.0792872	1.1494406	1.93e-11
질적	성별_암컷	0.9911635	0.9231118	1.0642177	0.806763
	성별_미상	0.6967774	0.6392530	0.7593137	< 2e-16

Numerical 경제 지표 값이 1단위 증가할 때 입양될 오즈가 1.1138배 증가

Categorical 성별을 알 수 없다면 성별이 수컷일 때보다 입양될 오즈가 0.6968배 낮음

결과 해석

로지스틱 회귀 모델 해석

수치형 변수

변수	오즈비	신뢰구간 하한	신뢰구간 상한	p-value
무게	0.795406	0.778225	0.812858	0
긍정	1.196178	1.174491	1.218280	0
부정	0.837873	0.821044	0.854945	0
1인당 지역 내 총생산	0.955197	0.936549	0.974148	0
1인당 경제 지표	0.991043	0.971736	1.010699	0.36962
동물병원 개수	1.091625	1.071598	1.111996	0

범주형 변수 – 성별

(기준: 암컷)

수컷	0.967[illegible]8	[illegible]93110[illegible]	1.004279	0.08214
미상	0.94[illegible]	[illegible]	1.134230	0.55283

암컷 ≒ 수컷 ≒ 미상

범주형 변수 – 중성화 여부

(기준: X)

미상	0.79[illegible]39	0.[illegible]8289	0.826582	0
O	[illegible]	[illegible]	[illegible].368374	0.00225

중성화 O > 중성화 X > 미상

범주형 변수 – 품종

(기준: herding)

	변수	오즈비	신뢰구간 하한	신뢰구간 상한	p-value
5	hound	0.446776	0.325338	0.611568	0
9	mixed	0.077797	0.061241	0.097818	0
3	non-sporting	0.49374	0.382351	0.632012	0
8	others	0.18633	0.145035	0.237172	0
2	sporting	0.752723	0.565827	0.995455	0.04853
7	terrier	0.339818	0.228447	0.507237	0
6	toy	0.405191	0.315548	0.515472	0
4	working	0.447759	0.33069	0.603261	0

범주형 변수 – 털 색

(기준: 검)

7	검/갈	0.978126	0.887554	1.077868	0.6554
3	검/갈/흰	1.241482	1.100151	1.400418	0.00044
1	검/흰	1.344278	1.219453	1.481931	0
8	갈	0.903008	0.835616	0.976077	0.01005
2	갈/흰	1.268927	1.166621	1.380495	0
4	기타	1.182859	1.063774	1.315179	0.00191
5	흰	1.074246	0.997884	1.156839	0.0575

결과 해석

랜덤포레스트 해석

랜덤포레스트 해석

GLOBAL

전체 모형에 대해
어떤 특징이 중요하고, 어떻게 결정했을까?

- Feature Importance
- Partial Dependence Plot

LOCAL

특정 경우(new observation, test)에 대해
왜 그런 결정을 했을까?

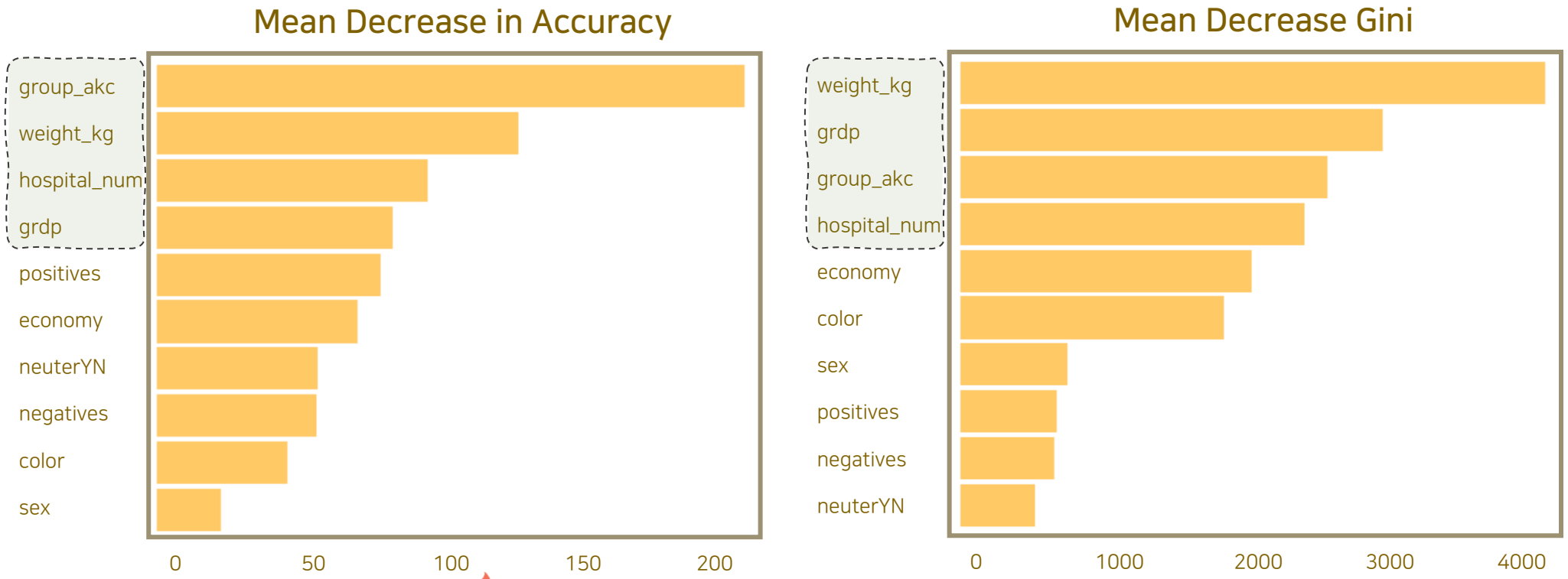
- LIME

결과 해석

로지스틱 회귀

랜덤포레스트

랜덤포레스트 변수 중요도



품종, 체중, 병원개수, 1인당 지역내 총생산의 중요도가 높구나!

- 1주차 복습
- 데이터셋 완성
- 분류 모델링
- 결과 해석

결과 해석

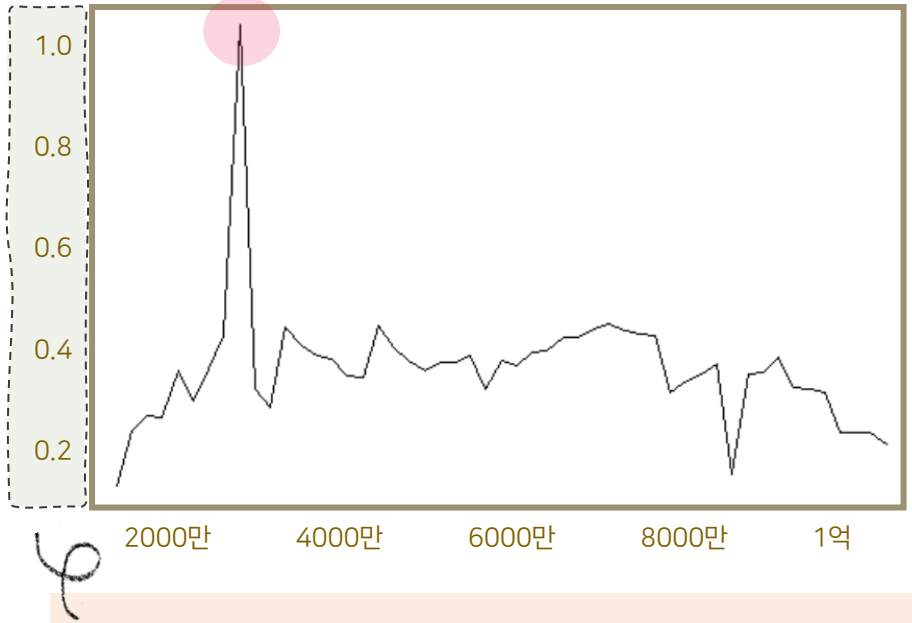
로지스틱 회귀

랜덤포레스트

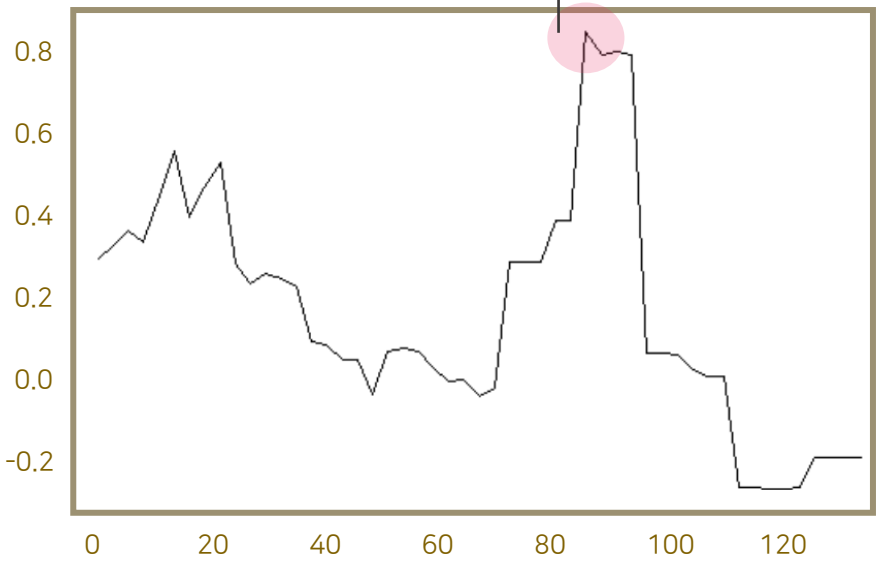
랜덤포레스트 PDP

병원이 90개일 때
예측에 가장 큰 영향을 끼침!

Partial Dependence on “grdp”



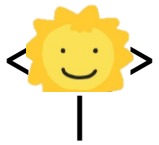
Partial Dependence on “hospital_num”



y축 수치는 ‘상대적인’ 값

즉, 수가 클수록 특정 변수가 종속변수 예측에 크게 영향을 끼침

trend에 집중해야 해~



- 1주차 복습
- 데이터셋 완성
- 분류 모델링
- 결과 해석

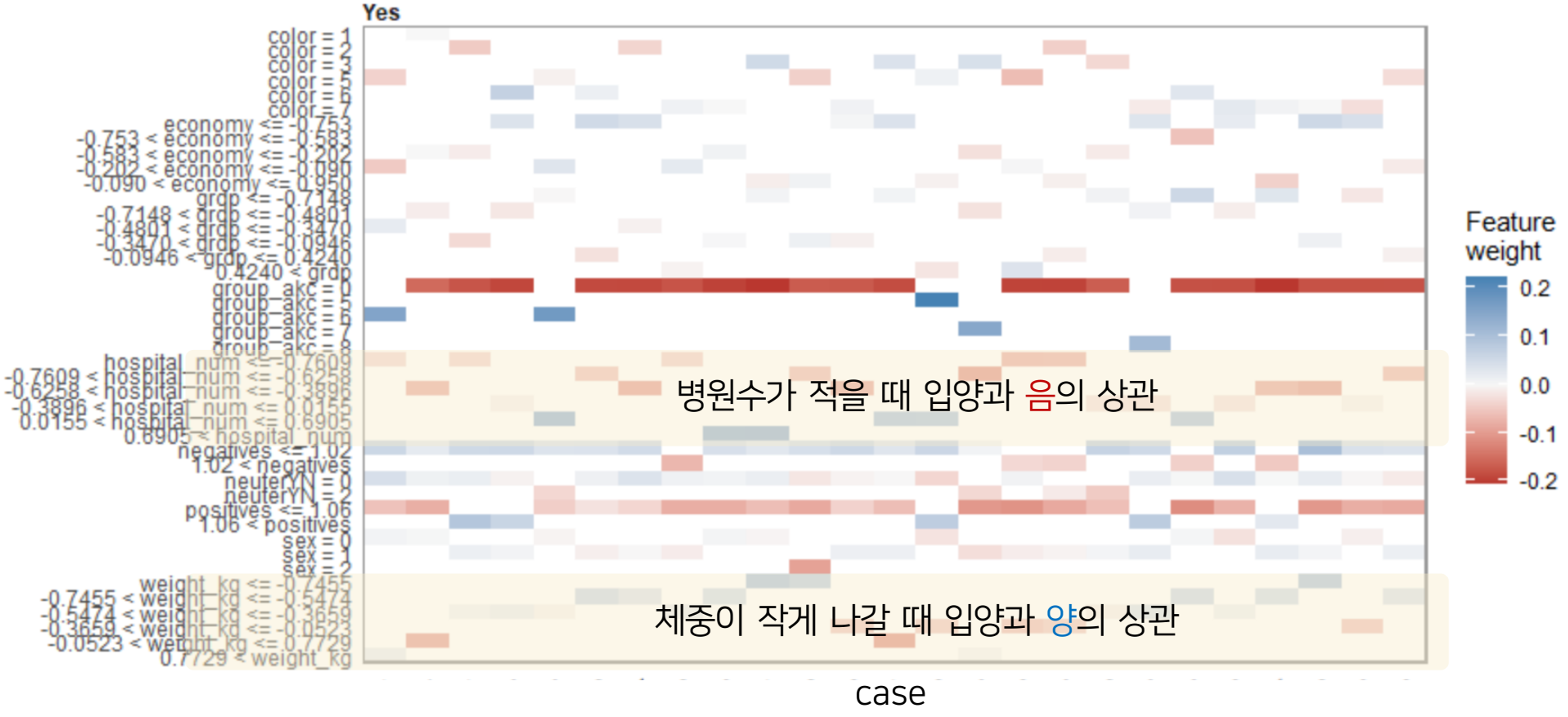
결과 해석

로지스틱 회귀

랜덤포레스트

랜덤포레스트 라임

전체 train셋에 대해 학습시킨 뒤 test셋으로 해석



입양유무와 양의 상관 관계가 있는 변수는 파란색, 음의 상관 관계는 빨간색

마무리

활용 방안 및 제안

분	석	을		진	행	하	며	…	.	

동물보호관리 시스템 공고 기재 가이드라인 마련

기재 형식 통일

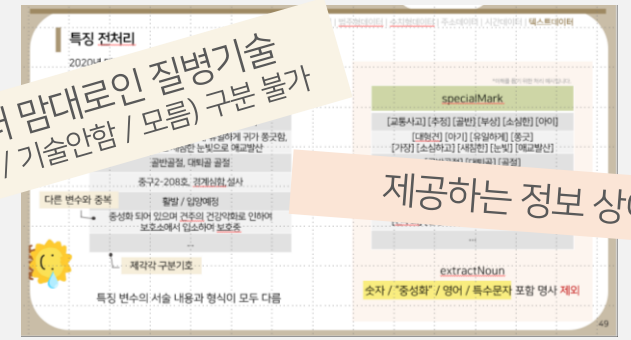
강아지만 털색 Unique 값이 3000개

믹스
혼합
mixed
믹스견

강아지만 품종 Unique 값이 200개

체크박스를 활용해 통일성 있게 기술!!!

특징 기술 가이드라인 제공



센터 맘대로인 질병기술
(없음 / 기술안함 / 모름) 구분 불가

제공하는 정보 상이

칩이 있는아이에요 전화기가 꺼져있네요 남양주에서 여기까지 어쩐일로 온거니?
순둥순둥한 머털도사님...내일 빡빡 밀고 증명사진 바꾸시죠.
겁이 조금 있고 눈이 사시에요 좀더 넓은 세상을 보고 싶었나봐요.

보호소마다 달라달라달라~

성격, 질병, 교육유무는 따로 칸을 마련해도 좋겠다