

유기동물 입양 예측

3팀 선형대수학

황정현 고경현 김지민 반경림 전효림

목차



주제 소개



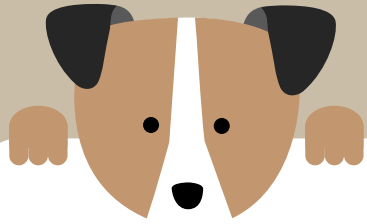
유기 동물 데이터



지역 특성 데이터



다음주 예고



주제 소개



현 반려동물 입양 및 유기 현황 파악



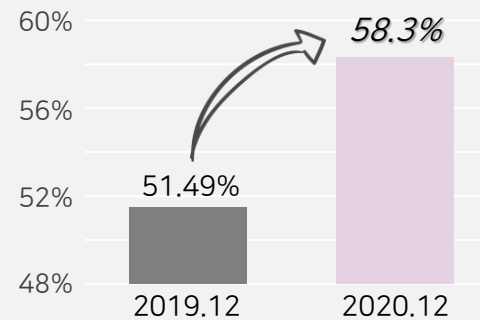
Pet + Family
반려동물을 또 하나의 가족처럼 여긴다는 뜻의 '펫팸',

그리고 신종 코로나 바이러스 감염증의 확산과 함께 늘어난 반려동물 입양 트렌드로
새롭게 등장한 단어 '팬데믹 퍼피'.
Pandemic Puppy

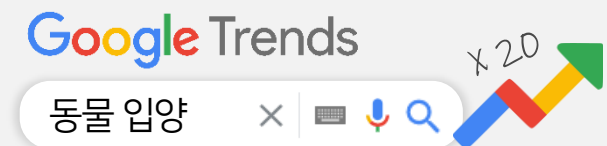
 global trend



유기동물
입양률



반려동물 입양 및 수탁



현 반려동물 입양 및 유기 현황 파악



#펫팸시대

Pet + Family
반려동물을 또 하나의 가족처럼 여긴다는 뜻의 '펫팸',

그리고 신종 코로나 바이러스 감염증의 확산과 함께 늘어난 반려동물 입양 트렌드로
새롭게 등장한 단어 '팬데믹 퍼피'.
Pandemic Puppy



global trend

그러나 그와 동시에 '펫팸시대'의 그늘이 드리워지고 있으니,



농림축산식품부

2020년 8월 말 기준
전국의 보호소에 머무는
유기동물 수

전년 대비 약 6배 증가



지정된 동물 보호센터에서
유기동물을 입양한 사람에 대해 입양비 지원

2020. 09. 16 발표

- 유기동물 보호센터 부족
- 안락사 절차 준수·유기동물 관리 등 미흡

주제 정의



유기동물 입양에
영향을 미치는
주요 변수 파악



??
개와 고양이를 위한
입양 예측 모델 ?
X



입양 예측 O

입양이 더욱 빠른 시일 내에 이루어지도록 함

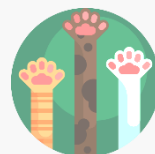


입양 예측 X

보호 기간을 늘려 안락사를 늦추거나 더욱 집중도 있는 관리 진행



200여 개의 유기동물 보호소가 모두 안락사 없는 보호소



유기동물의
행동 교정



예비 반려인을
대상으로 교육

우리도 할 수 있다! 사지 말고 입양하자!

열정 열정 열정!





유기동물데이터



유기동물 데이터 수집

동물보호관리시스템 홈페이지 공고 크롤링

동물보호관리시스템

농림축산식품부가 유기동물관리에서 동물등록에 이르기까지 동물보호 업무 전반을 통합적으로 관리하기 위해 각 시도의 동물보호업무 담당부서와 연계하여 운영하는 시스템



자세히보기

공고번호	경북-경산-2021-00208
접수일자	2021-04-22
품종	한국 고양이
성별	미상
발견장소	자인면
특징	생후30일령 추정...
상태	공고중
공고기간	2021-04-22 ~ 2021-05-03

크롤링



- 공고번호
- 축종
- 품종
- 털색
- 성별
- 중성화
- 특징
- 관할보호센터명
- 보호장소
- 상태

유기동물 데이터 수집

동물보호관리시스템 홈페이지 공고 크롤링

유기동물 입양할 때

체중이랑 나이도 중요할 것 같은데 안 볼고양?



그러개

- 공고번호
- 축종
- 품종
- 털색
- 성별
- 중성화
- 특징
- 관할보호센터명
- 보호장소
- 상태

유기동물데이터

데이터 수집

전처리 및 EDA

최종데이터셋

유기동물 데이터 수집

동물보호관리시스템

10시간이 아니라 99

250시간 99?

내가 지금 계산을 잘못하는 건가 99?

저녁 99 오전 3:26

지금 16쪽, 돌린지 1시간

파이썬에 흔적은 max 14378

그러면 10시간 너무 불가능인데???

오전 3:27



1년치 크롤링 하는데 250시간이면
주분기간 내에 데이터 구할 수 없다아거

동물보호관리시스템

리에서 동물등록에

도의 동물보호업무

나의 부기는 쉬지 않아.

오전 3:46

쉬지 않아서 지금 103쪽이얌 9999

오전 9:24

미쳤어

api인증키가 갑자기 된대^^

```
<?xml version="1.0" encoding="UTF-8" ?>
<response>
  <header>
    <resultCode>00</resultCode>
    <resultMsg>NORMAL SERVICE.</resultMsg>
  </header>
  <body>
    <item>
      <apiKey>2015/04/28</apiKey>
    </item>
  </body>
</response>
```

API하러 당장 가시개

```
<?xml version="1.0" encoding="UTF-8" ?>
<response>
  <header>
    <resultCode>00</resultCode>
    <resultMsg>NORMAL SERVICE.</resultMsg>
  </header>
  <body>
    <item>
      <apiKey>2015/04/28</apiKey>
    </item>
  </body>
</response>
```

오전 9:28

21년도 있네?^^ 범위 조정해서
당장 코딩 다시 해볼게 9999

오전 9:29

변수 나누기 | 범주형데이터 | 수치형데이터 | 주소데이터 | 시간데이터 | 텍스트데이터

유기동물 데이터 수집

동물보호관리시스템 유기동물 조회 서비스 API (2018 ~ 2020)

유기동물 데이터.csv

	보호장소	보호소이름		유기번호		중성화유무	공고종료일	공고번호	공고시작일	관할기관		특징		
Age	careAddr	careNm	colorCd	desertionNo	kindCd	neuterYn	noticeEdt	noticeNo	noticeSdt	orgNm	processState	sexCd	specialMark	weight
2020 (년생)	강원도 원주시 호저면 칠봉로 110-6 (호저면)	횡성유기동물 보호센터	연갈색	442426202 000237	[개] 라브라도 리트리버	U	20210118	강원-횡성- 2021- 00004	20210107	강원도 횡성군	종료(반환)	F	.	18(Kg)
2020 (년생)	강원도 원주시 호저면 칠봉로 110-6 (호저면)	횡성유기동물 보호센터	연갈색	442426202 000236	[개] 라브라도 리트리버	U	20210118	강원-횡성- 2021- 00003	20210107	강원도 횡성군	종료(반환)	F	.	18(Kg)
2021 (년생)	강원도 원주시 호저면 칠봉로 110-6 (호저면)	횡성유기동물 보호센터	흰색/검정	442426202 0002335	[개] 믹스견	U	20210118	강원-횡성- 2021- 00002	20210107	강원도 횡성군	종료(입양)	F	.	2.9(Kg)
2021 (년생)	부산광역시 강 서구 맥도강변 길 752-15 (대 저2동)	부산동물보 호센터	삼색	441405202 003491	[고양이] 한국 고양이	N	20210114	부산-중구- 2021- 00002	20210106	부산광역시 중 구	종료(자연사)	F	중구2-208호, 경계심함, 설사	0.4(Kg)

주제 소개

유기동물데이터

지역특성데이터

다음주 예고

변수 나누기 | 범주형데이터 | 수치형데이터 | 주소데이터 | 시간데이터 | 텍스트데이터

품종 전처리

품종 기준, 개 / 고양이로 분석 대상 한정

kindCd
[개] 라브라도 리트리버
[기타축종] 염소
[개] 믹스견
[고양이] 한국 고양이



품종	세부품종
개	리브라도 리트리버
기타축종	염소
개	믹스견
고양이	한국고양이




품종	세부품종
개	리브라도 리트리버
개	진도
개	믹스견



품종	세부품종
고양이	한국
고양이	쇼컷
고양이	터키시



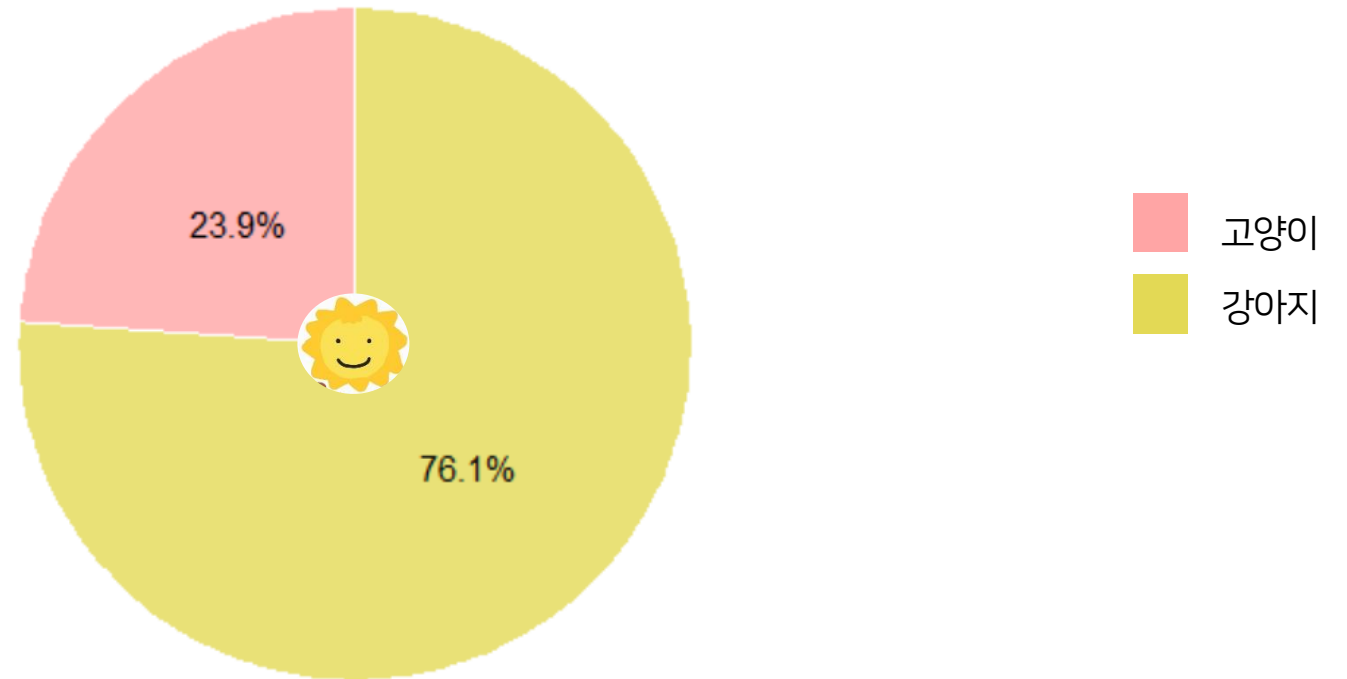
개와 고양이를 제외한
기타축종은 분석에서 제외한다아거

변수 나누기 | 범주형데이터 | 수치형데이터 | 주소데이터 | 시간데이터 | 텍스트데이터

유기동물 시각화

2018.01 ~ 2020. 12

<전체 유기동물 비율>

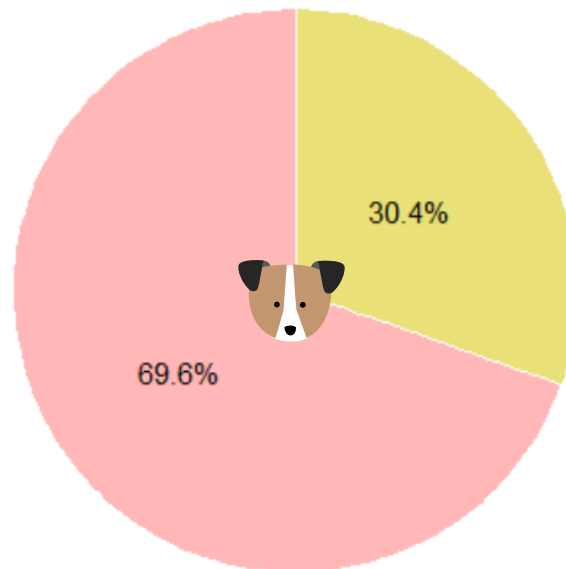


- 주제 소개
- 유기동물데이터
- 지역특성데이터
- 다음주 예고

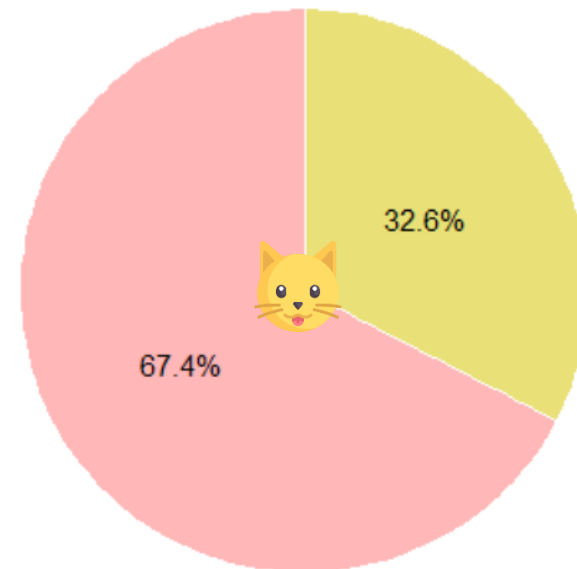
유기동물 시각화

2018.01 ~ 2020. 12

<강아지 유기동물 비율>



<고양이 유기동물 비율>



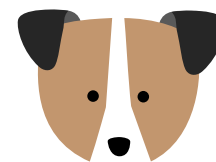
입양 안 됨

입양됨

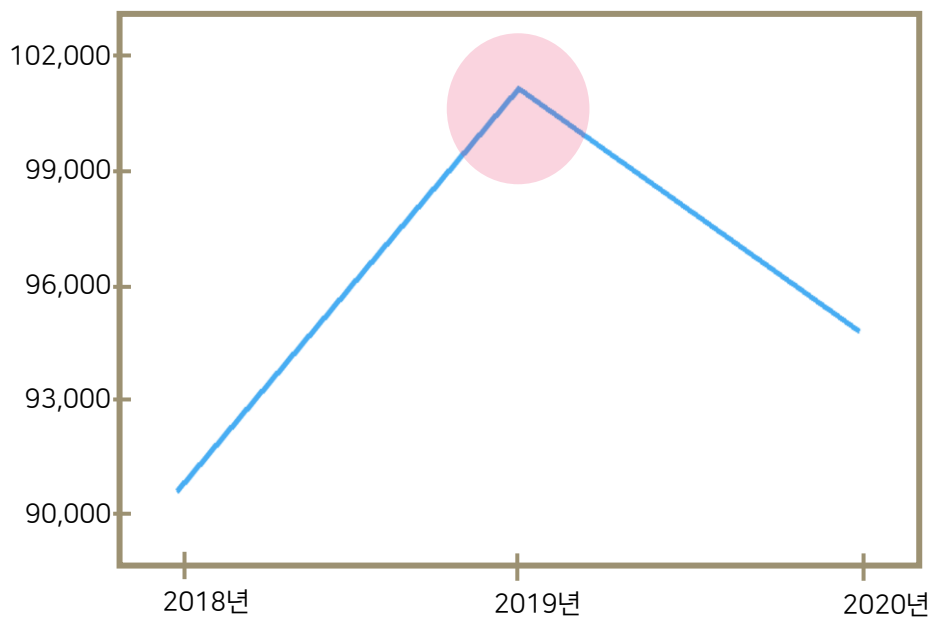
변수 나누기 | 범주형데이터 | 수치형데이터 | 주소데이터 | 시간데이터 | 텍스트데이터

유기동물 시각화 (개)

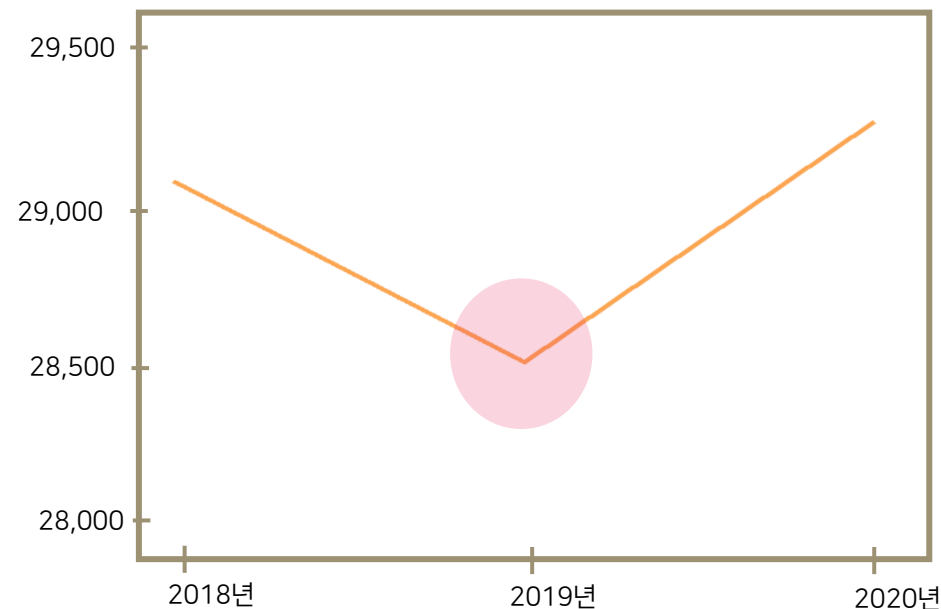
2018.01 ~ 2020. 12



<연도별 유기동물 수>



<연도별 유기동물 입양 수>



유기 수는 2019년이 가장 높은 반면, 입양 수는 2019년이 가장 낮음

주제 소개

유기동물데이터

지역특성데이터

다음주 예고

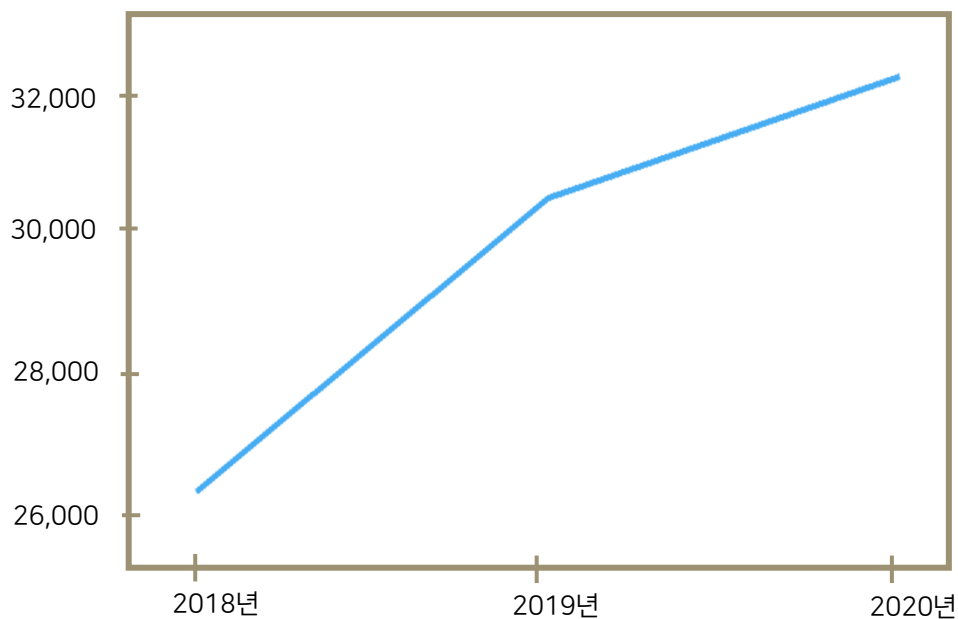
변수 나누기 | 범주형데이터 | 수치형데이터 | 주소데이터 | 시간데이터 | 텍스트데이터

유기동물 시각화 (고양이)

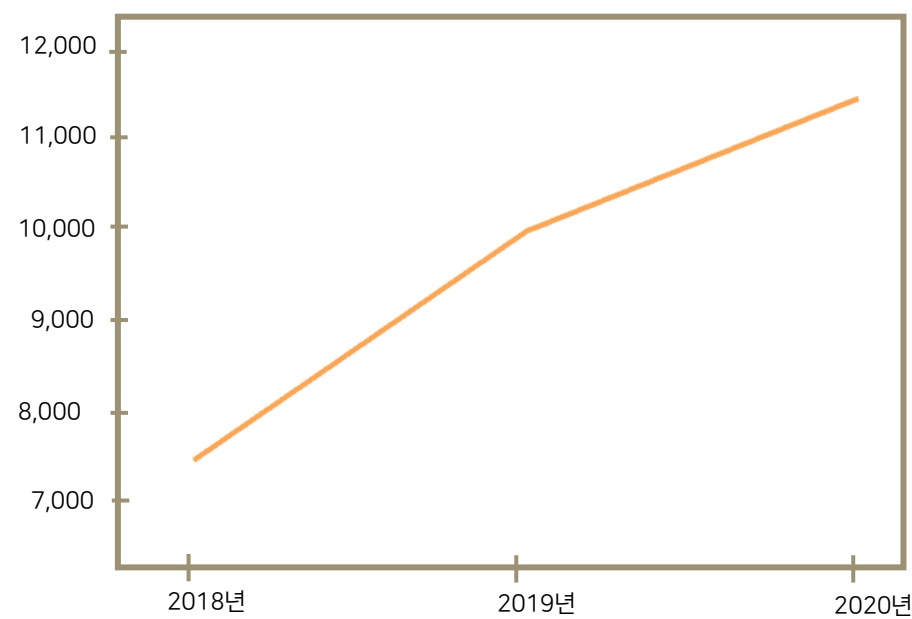
2018.01 ~ 2020. 12



<연도별 유기동물 수>



<연도별 유기동물 입양 수>



유기 수와 입양 수 모두 2019년 이후 기울기가 감소하지만
3년 연속 증가하는 추세

주제 소개

유기동물데이터

지역특성데이터

다음주 예고

변수 나누기 | 범주형데이터 | 수치형데이터 | 주소데이터 | 시간데이터 | 텍스트데이터

상태 전처리

입양여부 파생변수 생성

processState
종료(자연사)
종료(반환)
종료(입양)
종료(안락사)
보호중



상태	입양여부(Y)
자연사	0
반환	0
입양	1
안락사	0
보호중	0

총 7개의 상태 범주
(보호중, 입양, 안락사, 자연사, 반환, 기증, 방사)

- 주제 소개
- 유기동물데이터
- 지역특성데이터
- 다음주 예고

변수 나누기 | 범주형데이터 | 수치형데이터 | 주소데이터 | 시간데이터 | 텍스트데이터

상태 전처리

입양여부 파생변수 생성

processState
종료(자연사)
종료(반환)
종료(입양)
종료(안락사)
보호중



상태	입양여부(Y)
자연사	0
반환	0
입양	1
안락사	0
보호중	0

공고 종료 여부가 아닌 **입양 여부**가 관심 대상

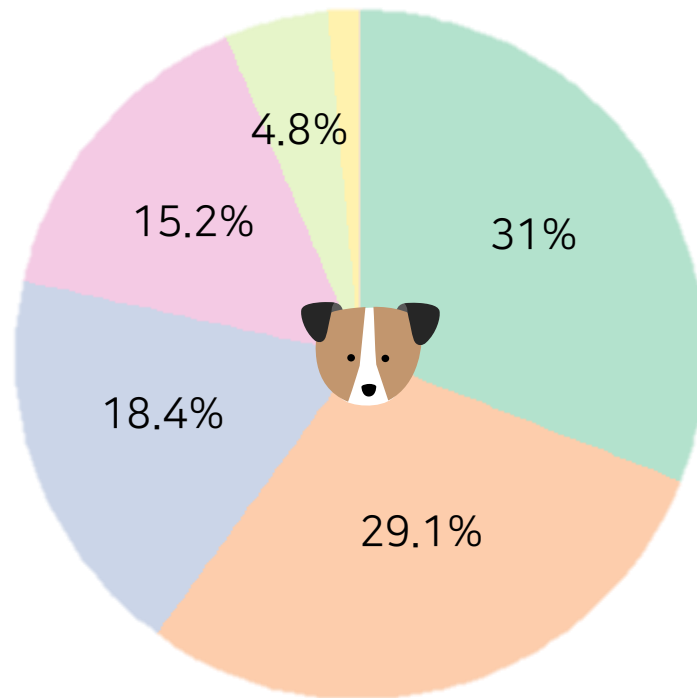


- 주제 소개
- 유기동물데이터
- 지역특성데이터
- 다음주 예고

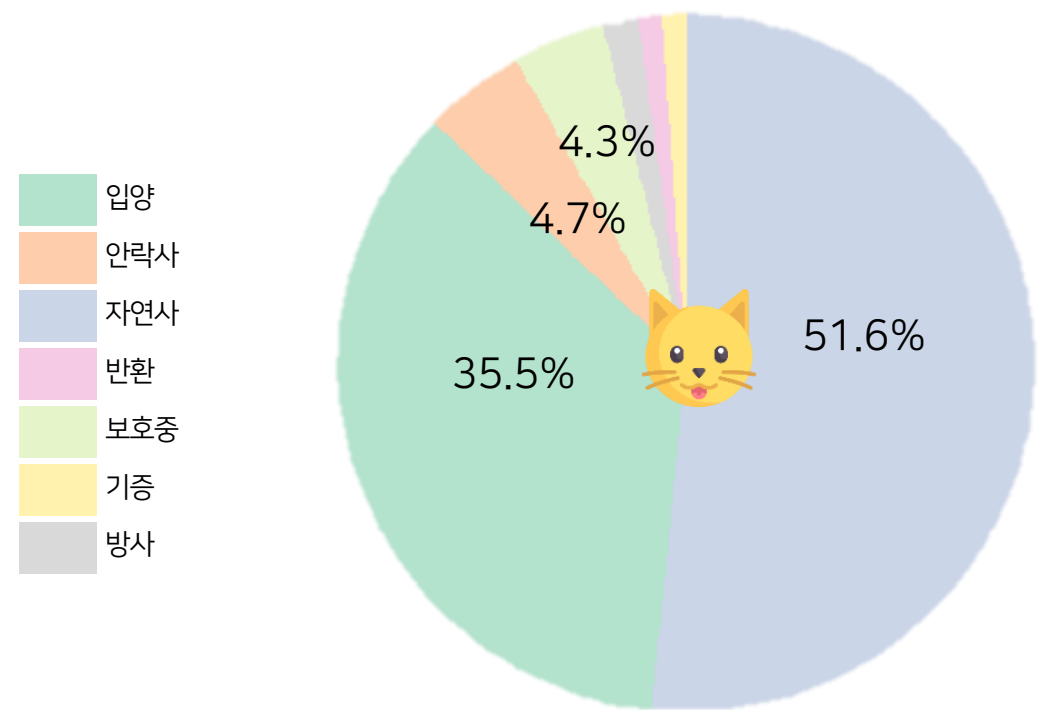
변수 나누기 | 범주형데이터 | 수치형데이터 | 주소데이터 | 시간데이터 | 텍스트데이터

상태 시각화

<강아지 상태 분포 비율>



<고양이 상태 분포 비율>

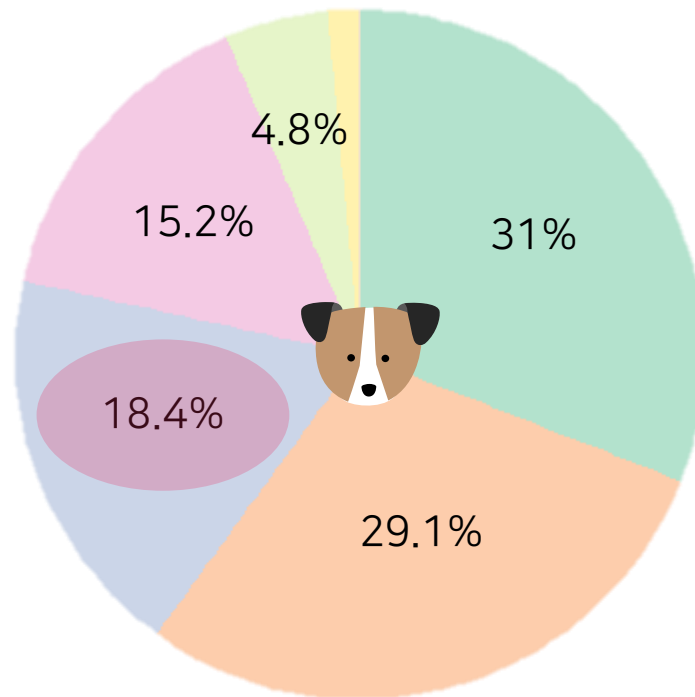


입양
안락사
자연사
반환
보호중
기증
방사

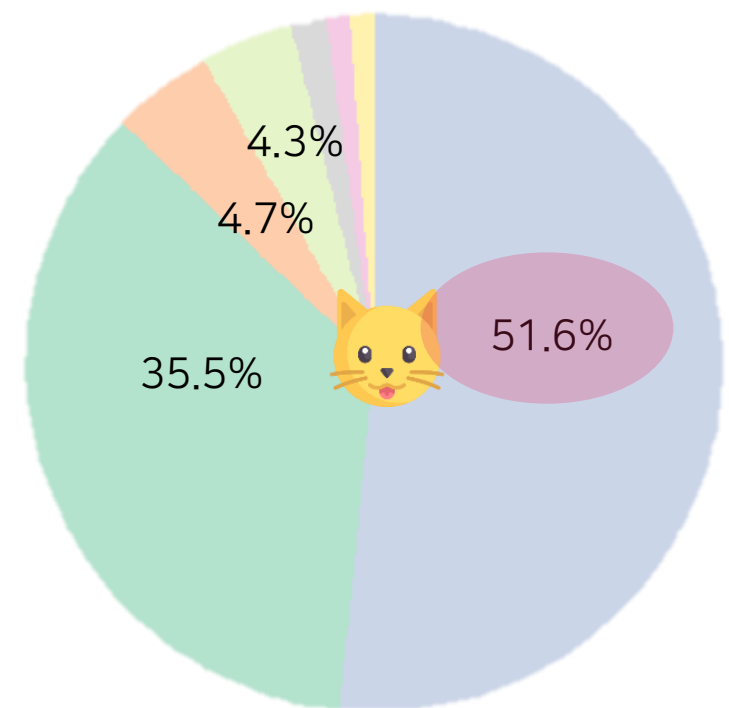
변수 나누기 | 범주형데이터 | 수치형데이터 | 주소데이터 | 시간데이터 | 텍스트데이터

상태 시각화

<강아지 상태 분포 비율>



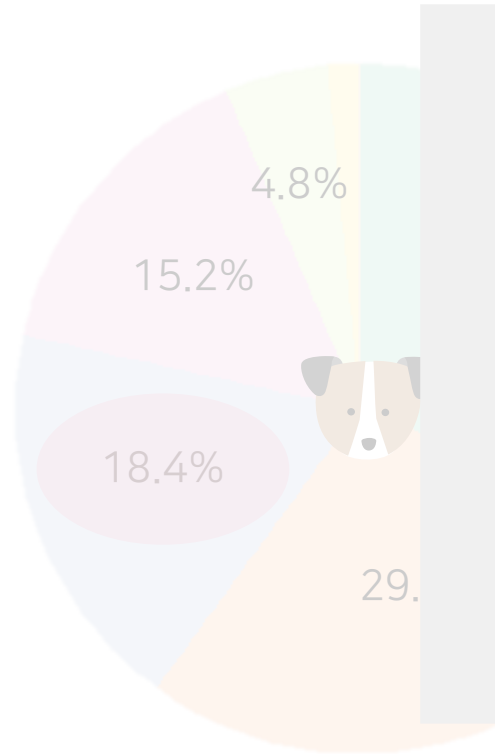
<고양이 상태 분포 비율>



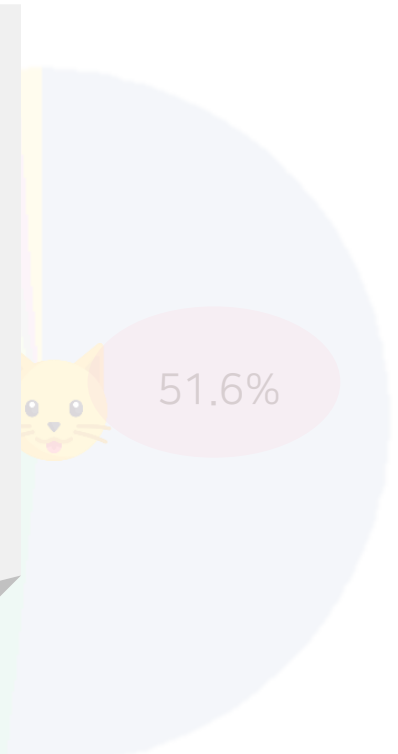
입양
안락사
자연사
반환
보호중
기증
방사

상태 시각화

<강아지 상태 분포 비율>



<고양이 상태 분포 비율>



- 입양되지 않더라도 기증, 방사, 주인 반환 등의 이유로 약 50% 동물 생존
- 강아지와 고양이의 자연사 비율 차이가 큼

중성화여부 전처리

neuterYn
Y
N
U
가마치통닭
4일치료후폐사



중성화여부
Y
N
U
U
U

"Y" : 중성화 O , "N" : 중성화 X , "U" : 미상



- 데이터 기입 상의 오류로
중성화여부에 해당되지 않는 내용 존재
- 전체 128885개의 관측치 중
23개 중성화여부를 "U"로 변환
- 인코딩은 모델링 과정 중에서 고려 예정

주제 소개

유기동물데이터

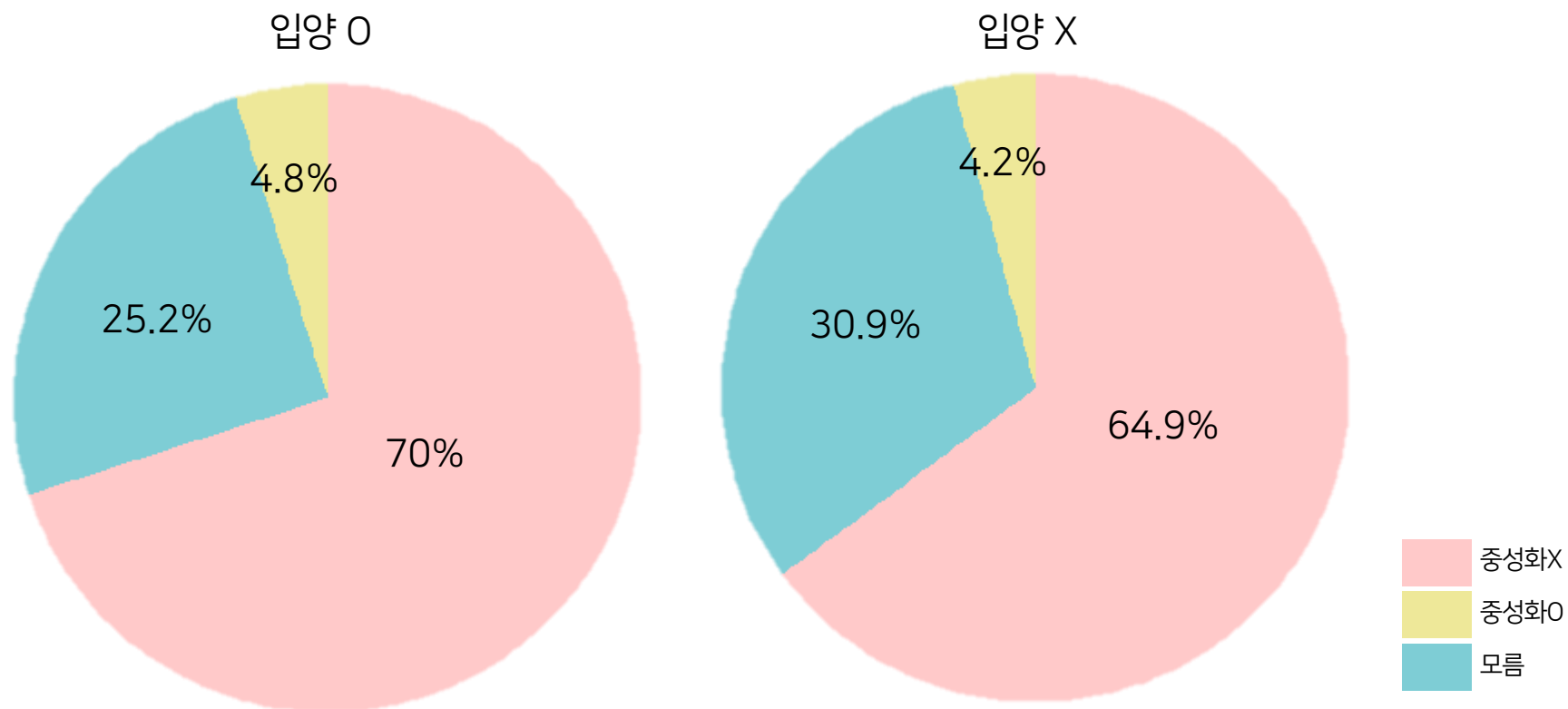
지역특성데이터

다음주 예고

변수 나누기 | 범주형데이터 | 수치형데이터 | 주소데이터 | 시간데이터 | 텍스트데이터

중성화에 따른 입양 여부 🐕

Q. 중성화 여부를 확실히 아는 것이 강아지 입양 시에 영향을 미칠까?



주제 소개

유기동물데이터

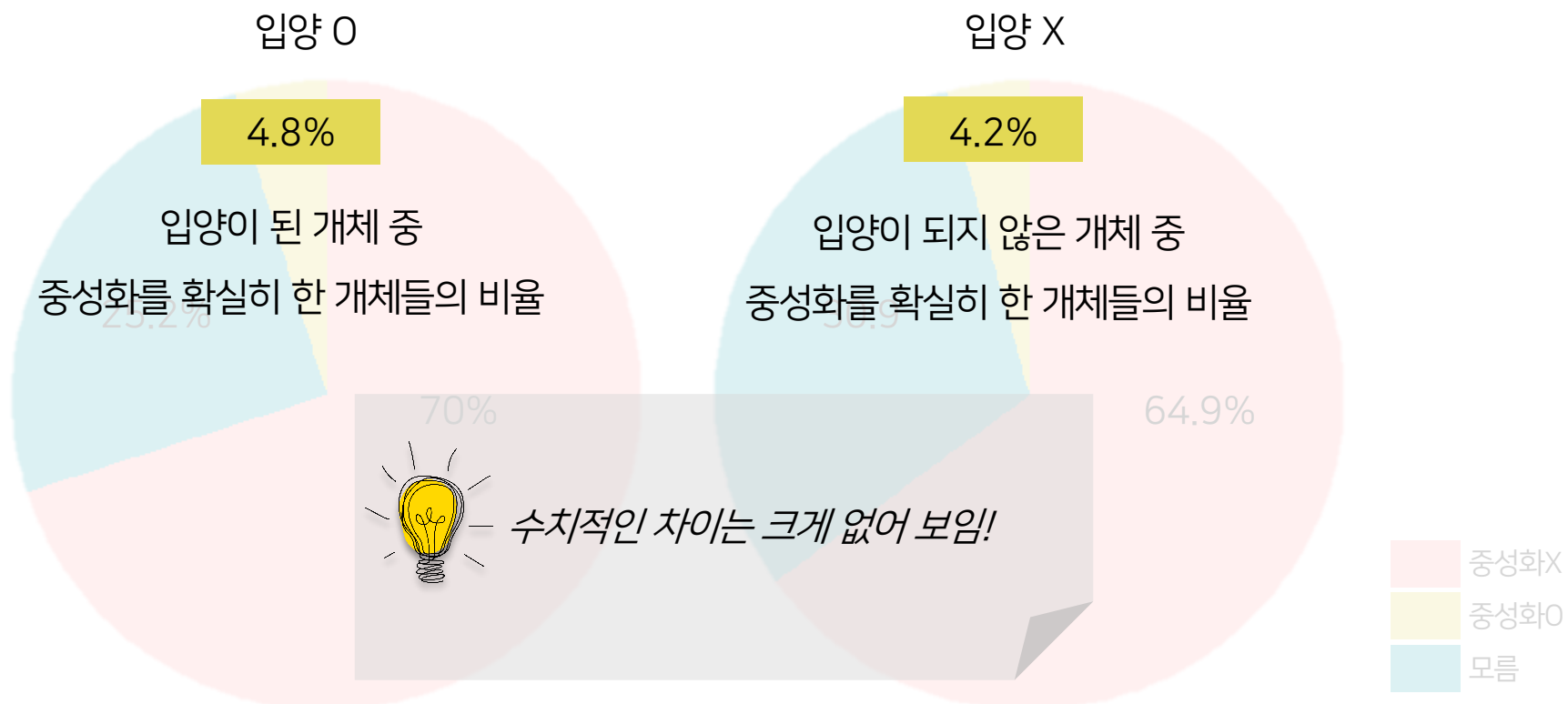
지역특성데이터

다음주 예고

변수 나누기 | 범주형데이터 | 수치형데이터 | 주소데이터 | 시간데이터 | 텍스트데이터

중성화에 따른 입양 여부 🐕

Q. 중성화 여부를 확실히 아는 것이 강아지 입양 시에 영향을 미칠까?



주제 소개

유기동물데이터

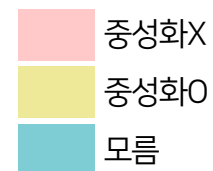
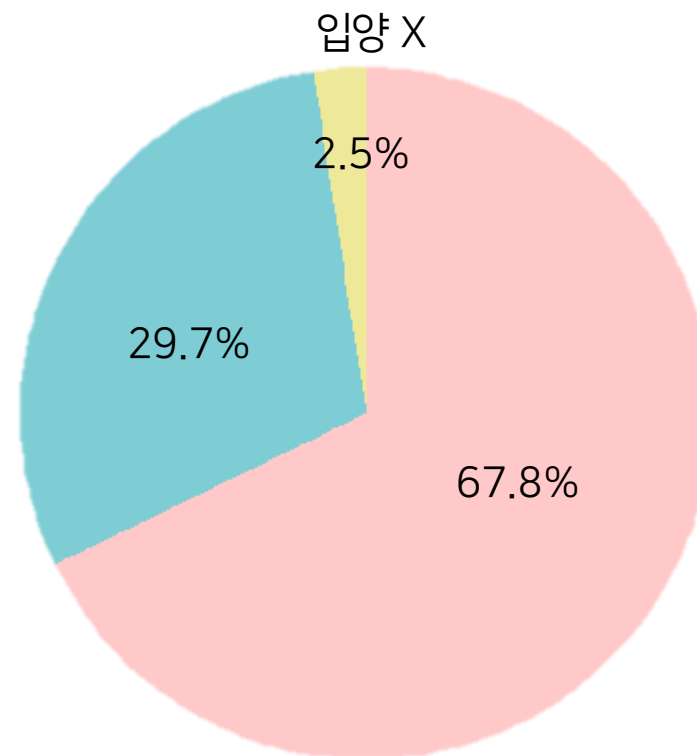
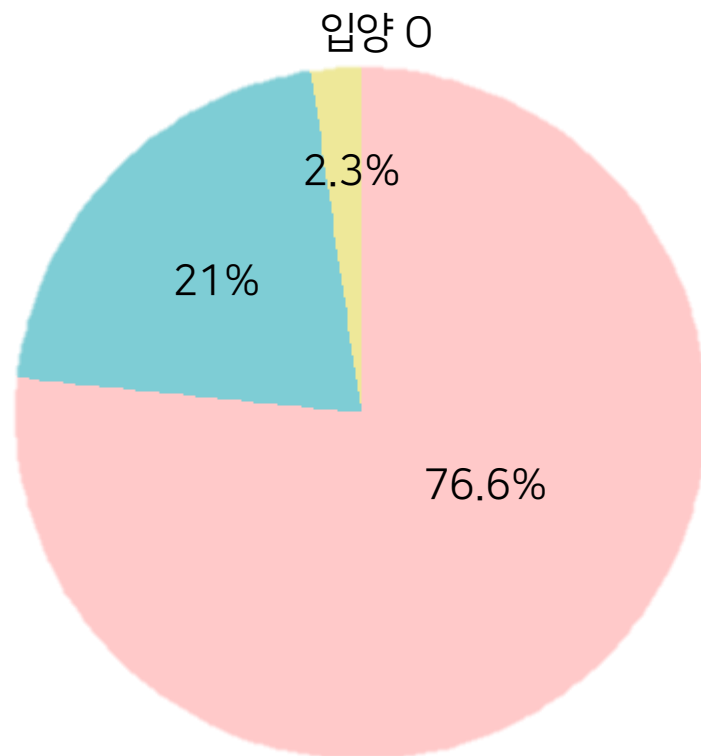
지역특성데이터

다음주 예고

변수 나누기 | 범주형데이터 | 수치형데이터 | 주소데이터 | 시간데이터 | 텍스트데이터

중성화에 따른 입양 여부 🐱

Q. 중성화 여부를 확실히 아는 것이 고양이 입양 시에 영향을 미칠까?



주제 소개

유기동물데이터

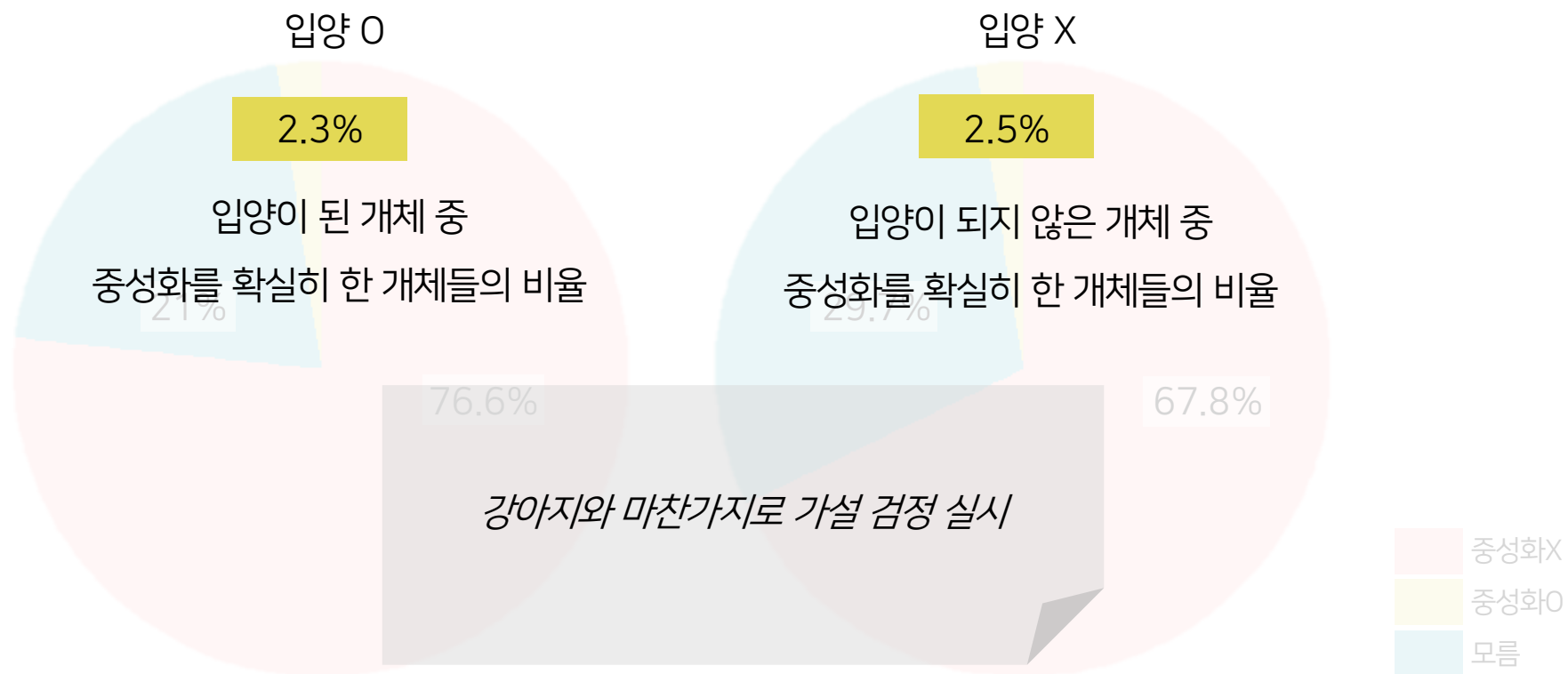
지역특성데이터

다음주 예고

변수 나누기 | 범주형데이터 | 수치형데이터 | 주소데이터 | 시간데이터 | 텍스트데이터

중성화에 따른 입양 여부 🐱

Q. 중성화 여부를 확실히 아는 것이 고양이 입양 시에 영향을 미칠까?



주제 소개

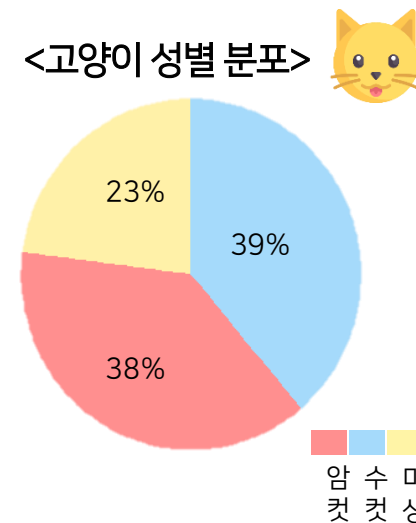
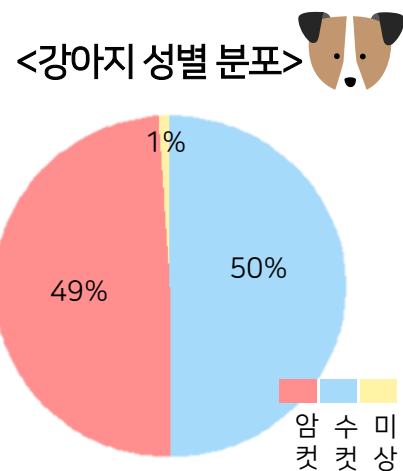
유기동물데이터

지역특성데이터

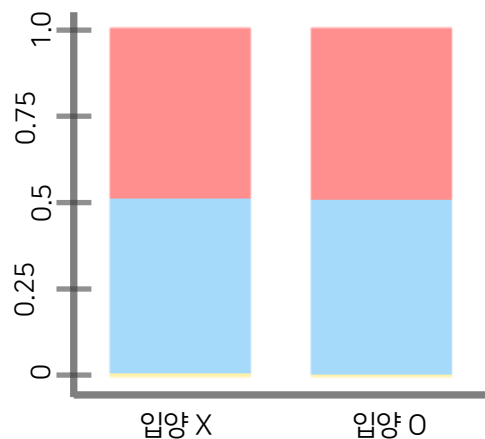
다음주 예고

변수 나누기 | 범주형데이터 | 수치형데이터 | 주소데이터 | 시간데이터 | 텍스트데이터

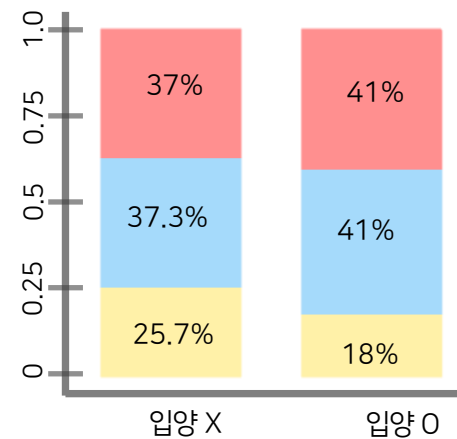
성별 분포 비율



<입양 여부에 따른 강아지 성별 분포>



<입양 여부에 따른 고양이 성별 분포>



주제 소개

유기동물데이터

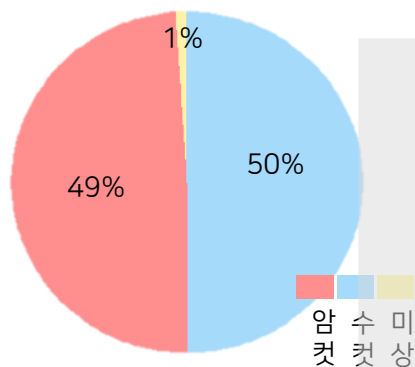
지역특성데이터

다음주 예고

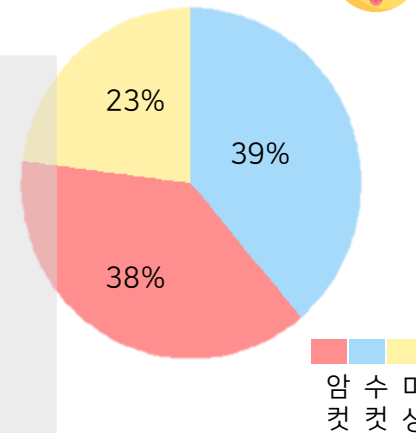
변수 나누기 | 범주형데이터 | 수치형데이터 | 주소데이터 | 시간데이터 | 텍스트데이터

성별 분포 비율

<강아지 성별 분포> 🐕

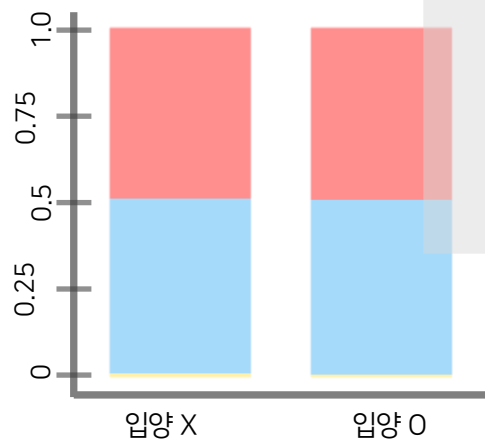


<고양이 성별 분포> 🐈

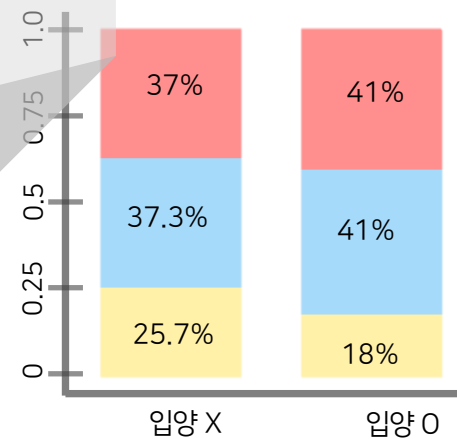


강아지, 고양이 모두
입양 여부에 따른 성별 분포는
전체 성별 분포와 유사

<입양 여부에 따른 강아지 성별 분포>



<입양 여부에 따른 고양이 성별 분포>



변수 나누기 | 범주형데이터 | 수치형데이터 | 주소데이터 | 시간데이터 | 텍스트데이터

체중 & 나이 전처리

parse_number함수

weight_kg
4.3(kg)
5.6(kg)
0.9(kg)
900(kg)
870(kg)



체중
4.3
5.6
0.9
900
870



체중
4.3
5.6
0.9
0.9
0.87

age
2019(년생)
2020(년생)
2018(년생)
2015(년생)
2016(년생)



나이
2019
2020
2018
2015
2016



나이
2
1
3
6
5

2020년 기준

주제 소개

유기동물데이터

지역특성데이터

다음주 예고

변수 나누기 | 범주형데이터 | 수치형데이터 | 주소데이터 | 시간데이터 | 텍스트데이터

체중 & 나이 전처리

parse_number함수

weight_kg
4.3(kg)
5.6(kg)

체중
4.3
5.6

체중
4.3
5.6

특징	체중
(개체관리번호5270)생후1주일	5270(Kg)
(개체관리번호5220)경계심	5220(Kg)
(개체관리번호4098)생후1주	4098(Kg)
(개체관리번호3134)심장사상충	3134(Kg)

0.9
900
870
나이
2019
2020
2018
2015
2016

0.9
0.9
0.87
나이
2
1
3
6
5

- 특징에 담긴 개체관리번호와 같은 숫자가 체중 변수에 존재
- 특징 변수의 다른 설명과 보면 단위의 오류는 아님
- 이 같은 개체는 총 7개로 영향을 미치기에 적다고 판단



기준

- 주제 소개
- 유기동물데이터
- 지역특성데이터
- 다음주 예고

변수 나누기 | 범주형데이터 | 수치형데이터 | 주소데이터 | 시간데이터 | 텍스트데이터

체중 & 나이 전처리

parse_number함수

weight_kg

체중

체중

데이터 7개 삭제!!

특징

체중

(개체관리번호5270)생후1주일

5270(Kg)

(개체관리번호5220)경계심

5220(Kg)

(개체관리번호4098)생후1주일

4098(Kg)

(개체관리번호34)생후1주일

34(Kg)



특징에 담긴 개체관리번호와 같은 숫자가 체중 변수에 존재

특징 변수의 다른 설명과 보던 것의 오류는 아님

이 같은 개체는 총 7개로 영향을 미치기에 적다고 판단

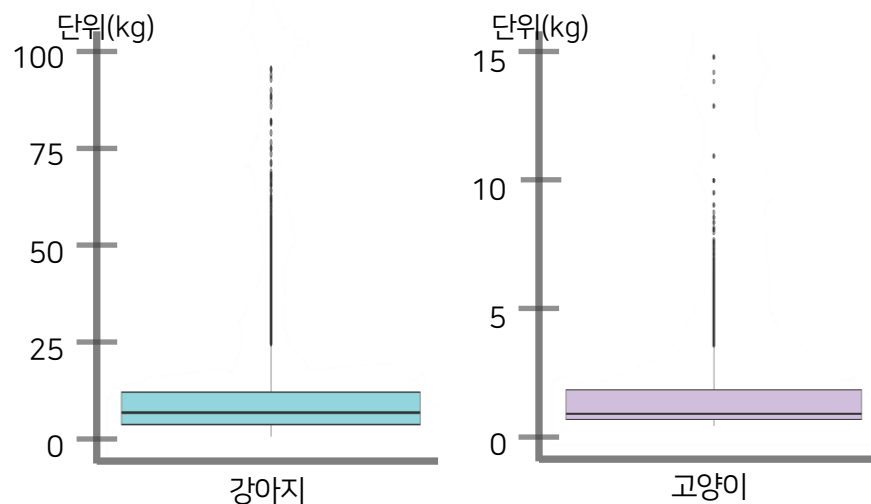


5

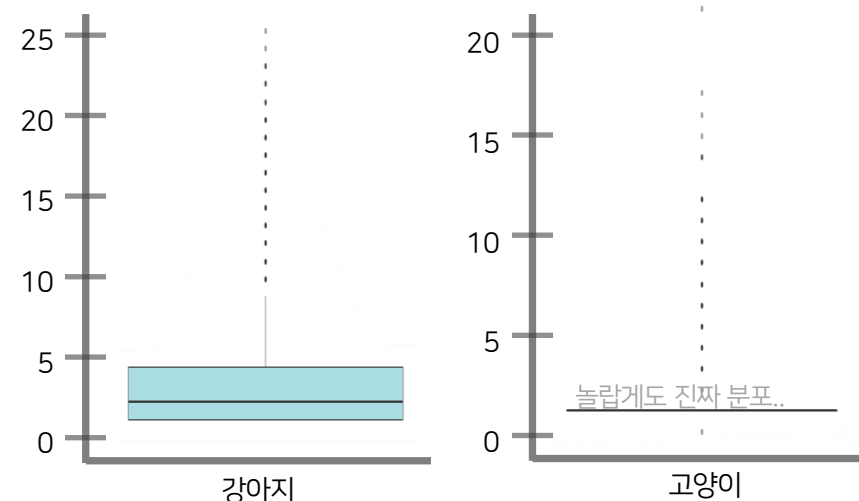
변수 나누기 | 범주형데이터 | 수치형데이터 | 주소데이터 | 시간데이터 | 텍스트데이터

체중 & 나이 시각화

<체중>



<나이>



이상치가 많이 존재 + 한쪽으로 치우친 형태
분석 시 의사결정에 큰 영향을 미칠 수 있는 요인

주제 소개

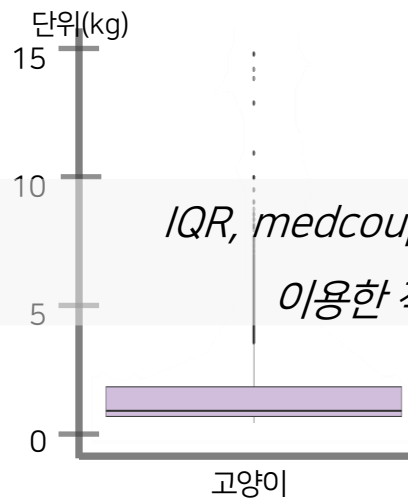
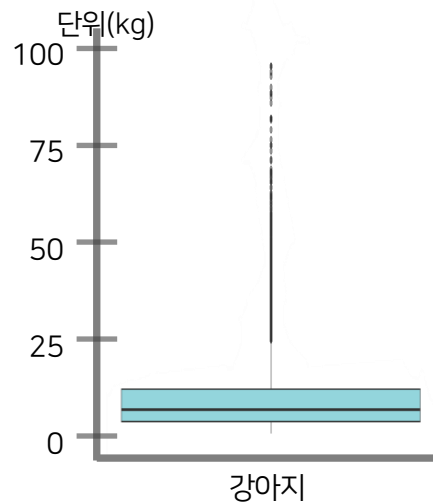
유기동물데이터

지역특성데이터

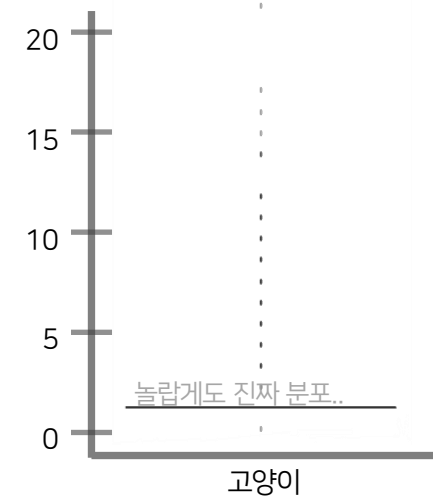
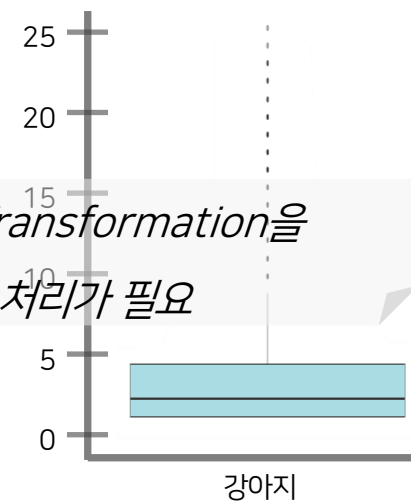
다음주 예고

체중 & 나이 시각화

<체중>



<나이>



*IQR, medcouple, Transformation을
이용한 적절한 처리가 필요*



*이상치가 많이 존재 + 한쪽으로 치우친 형태
분석 시 의사결정에 큰 영향을 미칠 수 있는 요인*

주제 소개

유기동물데이터

지역특성데이터

다음주 예고

이상치 탐색과 이상치 처리

이상치 (Outlier)

관측치들이 주로 모여 있는 곳에서 멀리 떨어져 있는 관측치

이상치를 포함한다면 분석 시에 결과가 왜곡될 수 있기 때문에 데이터의 특성에 적합한 이상치 처리가 필요

이상치 탐색 (Outlier Detection)

단변량, 다변량, 시계열 자료 등 자료의 특성에 따라 이상치를 탐색하는 방법이 달라짐

단변량 자료의 경우 표준화 점수, 가설검정, 사분위수 범위를 주로 활용하여 탐색

 주제 소개

 유기동물데이터

 지역특성데이터

 다음주 예고

이상치 탐색과 이상치 처리

이상치 (Outlier)

이상치 처리

관측치들이 주로 모여있는 곳에서 멀리 떨어진

이상치를 포함한다면 이상치들을 탐색하여 분석가의 판단에 근거하여 대체, 제거, 변환하는 것이 이상치 처리가 필요

자료에 대한 중요한 정보를 제공할 수도 있기 때문에 신중하게 다뤄야 함

단변량, 다변량, 시계열 자료 등 자료의 특성에 따라 이상치를 탐색하는 방법이 달라짐

단변량 자료의 경우 표준화 점수, 가설검정, 사분위수 범위를 주로 활용하여 탐색

주제 소개

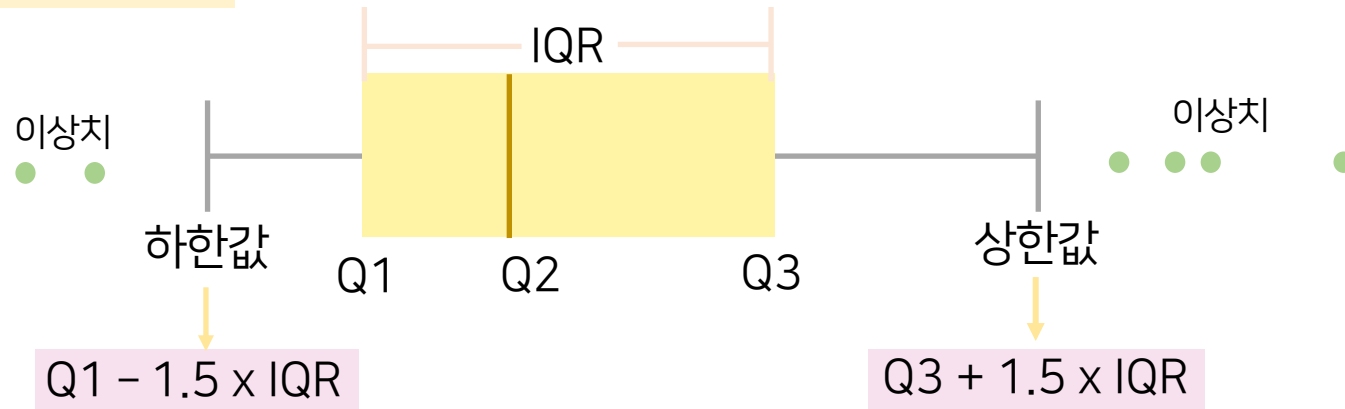
유기동물데이터

지역특성데이터

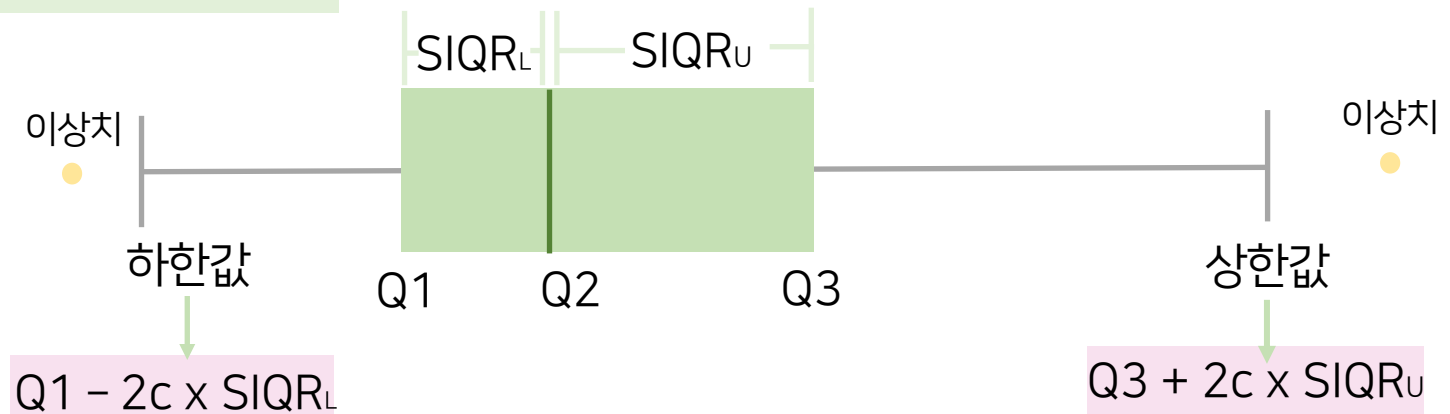
다음주 예고

이상치 처리 - IQR과 SIQR

사분위수범위 (IQR)

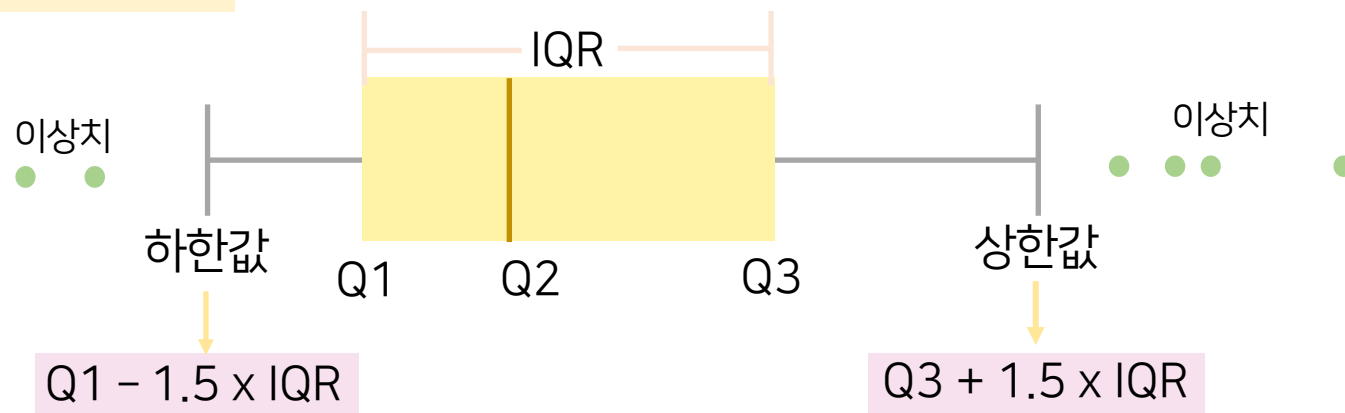


준사분위수범위 (SIQR)



이상치 처리 - IQR과 SIQR

사분위수범위 (IQR)



준사분위수범위 (SIQR)

- 정규분포와 같은 대칭적인 자료의 이상치 탐색에 효과적
- 박스플롯으로 쉽게 구할 수 있음
- 비대칭적인 자료에 대해 많은 이상치를 감지해 데이터 손실을 야기할 수 있음

$Q1 - 2c \times SIQR_L$

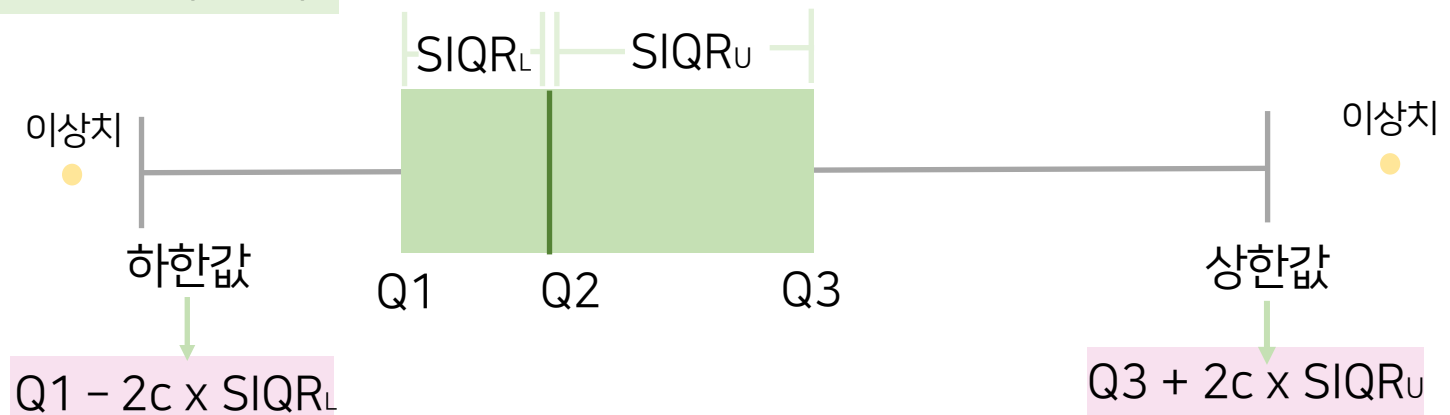
$Q3 + 2c \times SIQR_U$

이상치 처리 - IQR과 SIQR

사분위수범위 (IQR)

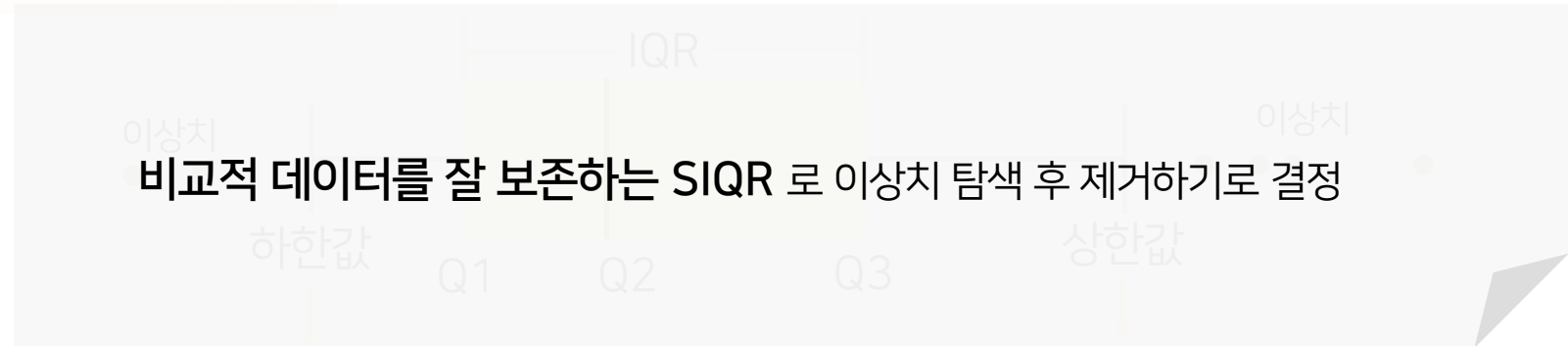
- ✦ IQR로 이상치를 탐색하는 것보다 Robust
- ✦ 비대칭적인 자료에 더 넓은 범위를 정상적인 범주로 허용해 데이터 손실을 방지

준사분위수범위 (SIQR)



이상치 처리 - IQR과 SIQR

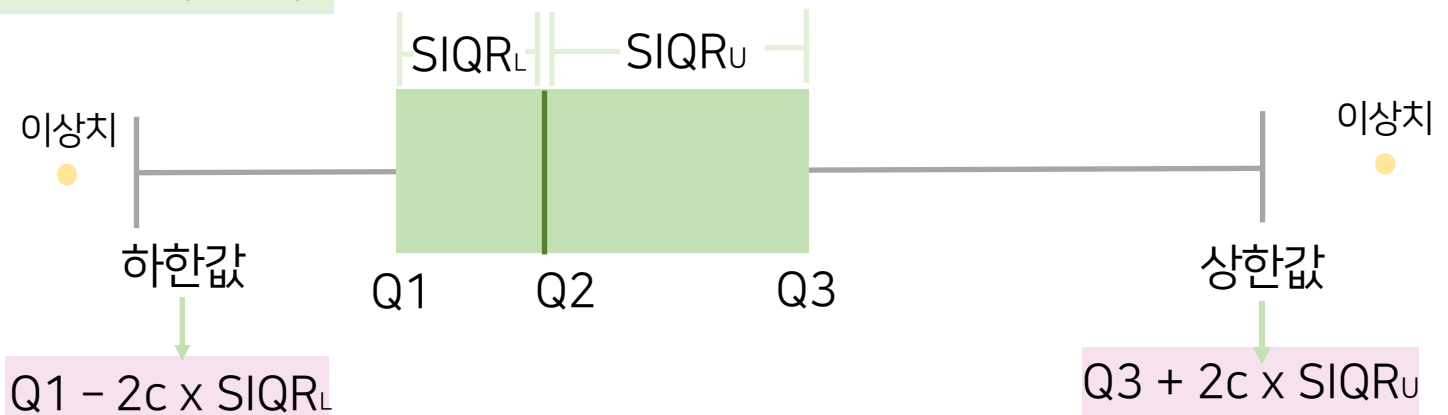
사분위수범위 (IQR)



$Q1 - 1.5 \times IQR$

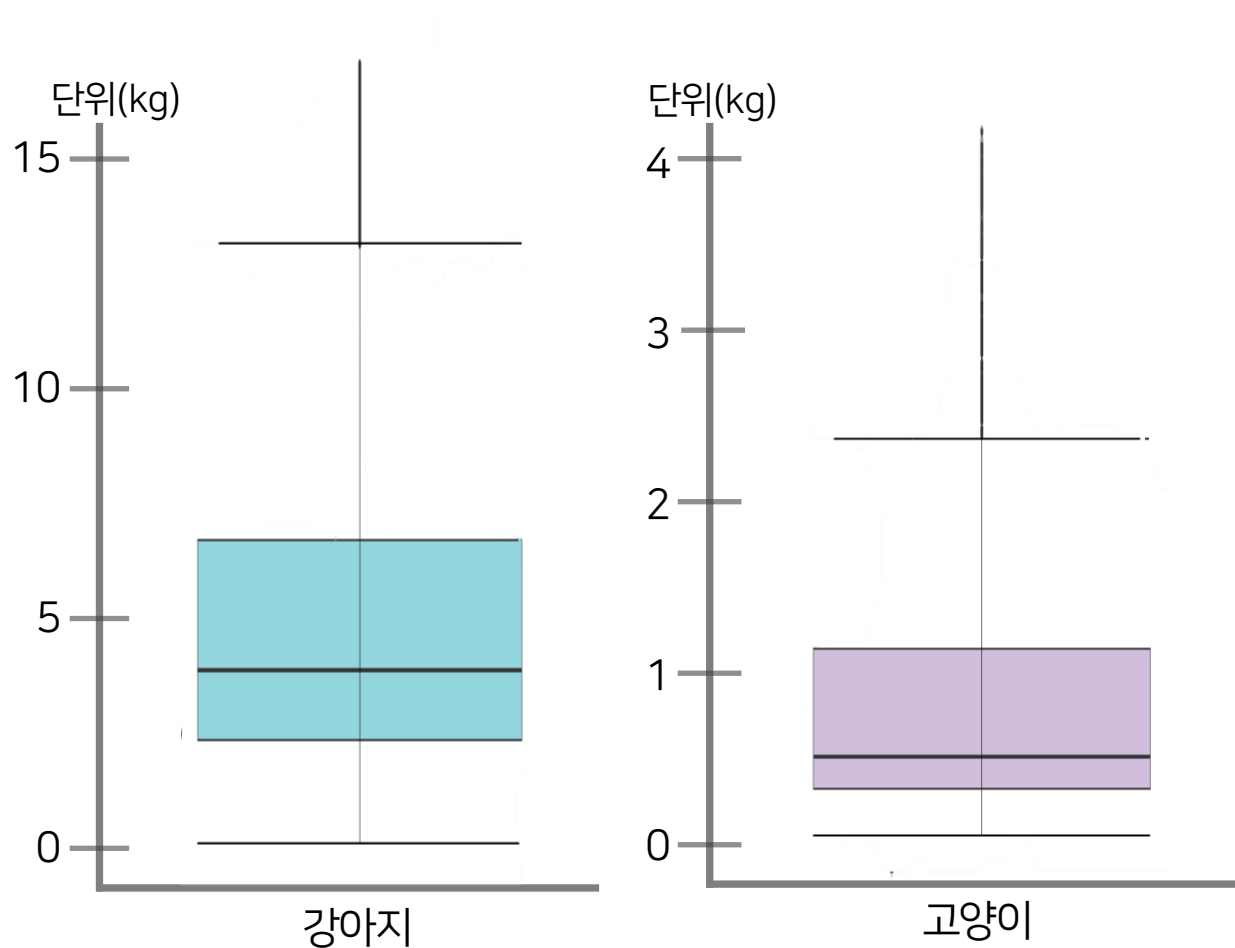
$Q3 + 1.5 \times IQR$

준사분위수범위 (SIQR)



변수 나누기 | 범주형데이터 | 수치형데이터 | 주소데이터 | 시간데이터 | 텍스트데이터

체중 시각화



- 이상치 제거 이전보다 데이터 분포 파악이 쉬워짐
- 하지만 동물 데이터의 특성상 체중이 많이 나가는 개체 존재 가능
- 이상치를 제거하면 이런 특성을 반영 못하기에 데이터 보존 결정



주소데이터 전처리

careAddr
전라북도 정읍시 칠보면 칠보중앙로 77-4 (칠보면, 삼화식당) 정주동물병원
전라북도 군산시 대야면 보덕안정길 108-20 (대야면) 군산도그랜드
경기도 여주시 능서면 능서공원길 34 (능서면)



보호센터
전라북도 정읍시
전라북도 군산시
경기도 여주시

orgNm
경상남도 창원시 의창성산구
경상남도 창원시 진해구
서울특별시 서초구



관할기관
경상남도 창원시
경상남도 창원시
서울특별시 서초구

주제 소개

유기동물데이터

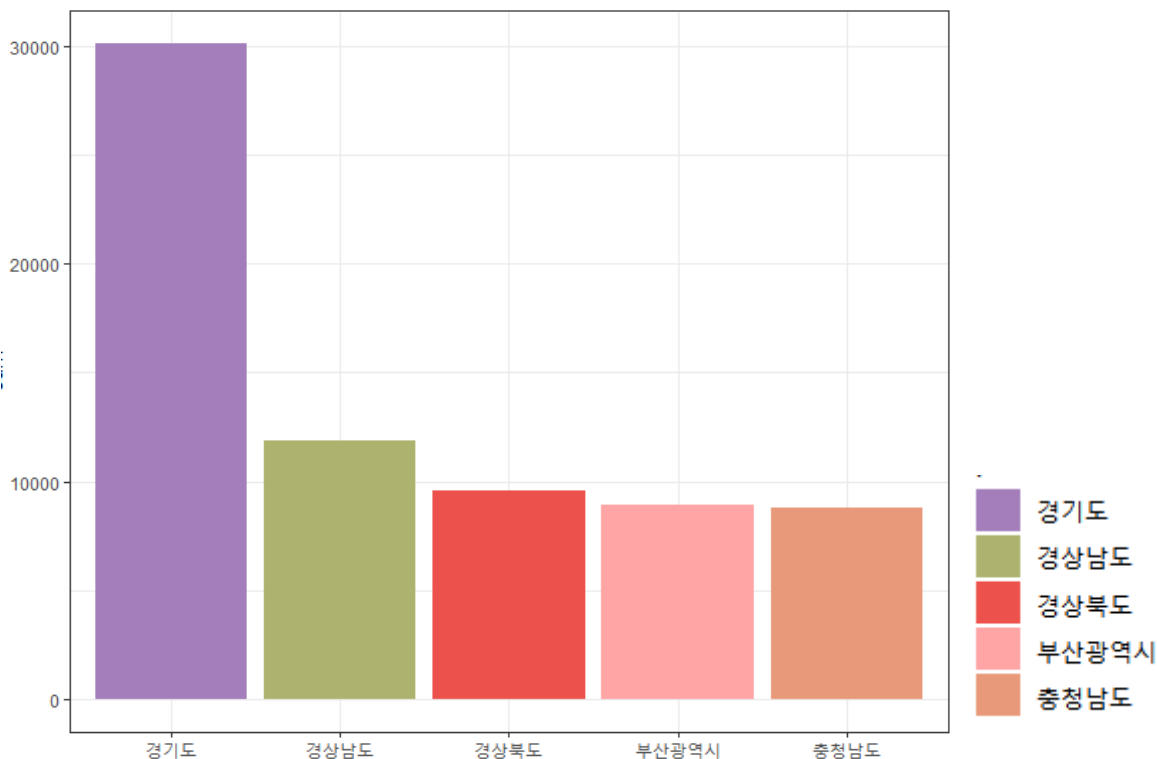
지역특성데이터

다음주 예고

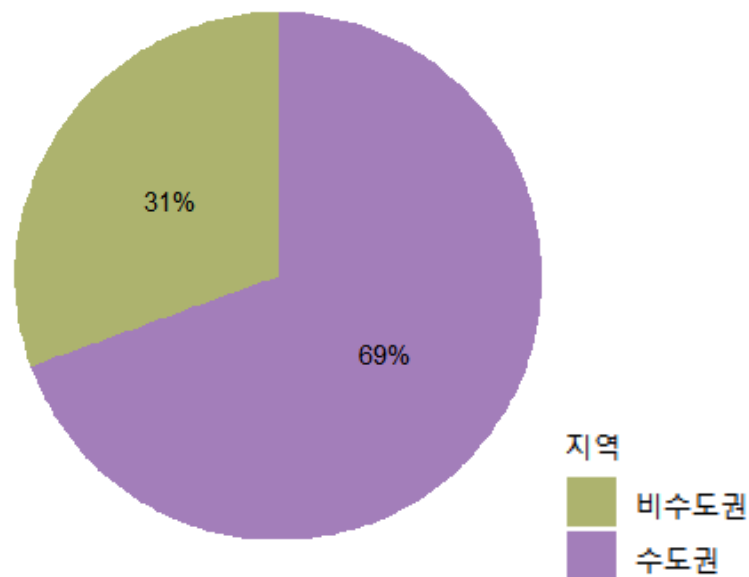
변수 나누기 | 범주형데이터 | 수치형데이터 | 주소데이터 | 시간데이터 | 텍스트데이터

주소데이터 시각화

<상위 다섯 시도별 수 비교>



<수도권/ 비수도권 수 비교>



주제 소개

유기동물데이터

지역특성데이터

다음주 예고

공고일자 전처리

기존 데이터

공고 종료일	공고 시작일
2020-09-12	2020-09-01
2020-03-09	2020-03-03
2020-12-30	2020-12-04
2020-08-20	2020-08-12
2020-01-29	2020-01-12

공고 종료일 - 공고 시작일

= 공고기간(duration) 변수 생성

unique로 살펴본 결과, 음수 존재

홈페이지 대조 결과,

시간의 역행 존재

음수인 행 제거!



주제 소개

유기동물데이터

지역특성데이터

다음주 예고

변수 나누기 | 범주형데이터 | 수치형데이터 | 주소데이터 | 시간데이터 | 텍스트데이터

공고일자 전처리

기존 데이터

공고 종료일	공고 시작일
2020-09-12	2020-09-01
2020-03-09	2020-03-03
2020-12-30	2020-12-04
2020-08-20	2020-08-12
2020-01-29	2020-01-12



공고기간
11 days
6 days
26 days
8 days
17 days

공고기간(duration) 변수 생성

- 주제 소개
- 유기동물데이터
- 지역특성데이터
- 다음주 예고

변수 나누기 | 범주형데이터 | 수치형데이터 | 주소데이터 | 시간데이터 | 텍스트데이터

공고일자 전처리

기존 데이터

공고 종료일
2020-05-01
2020-03-09
2020-09-01
2020-10-13
2020-12-30

계절 변수 생성

계절
Spring
Winter
Summer
Fall
Winter



3개월 단위!

유기동물이 어느 계절에 입양되는지
경향성을 파악하고자 만든 변수
공고 종료일 기준으로 계절 변수 생성!

주제 소개

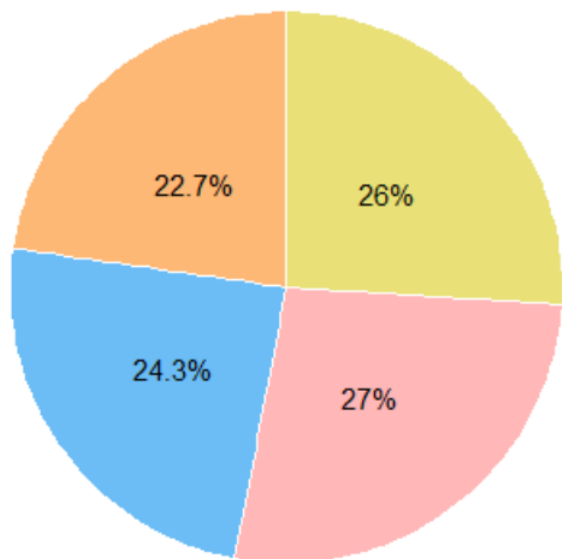
유기동물데이터

지역특성데이터

다음주 예고

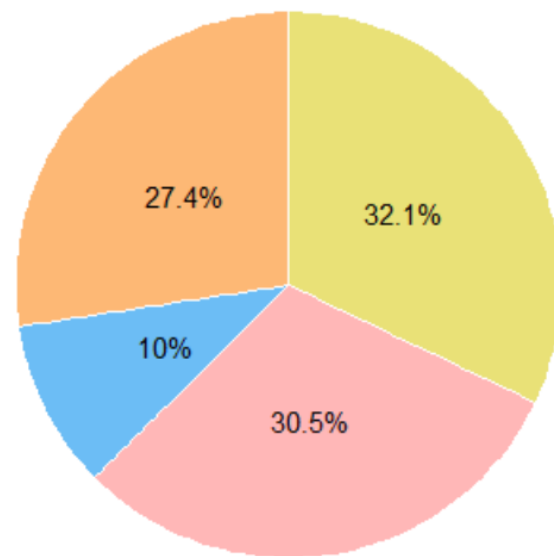
입양 계절

<강아지 입양 계절> 🐶



비율 차가 뚜렷하지 않음

<고양이 입양 계절> 🐱



여름이 32% 가장 높고,
겨울이 10% 가장 낮음

주제 소개


유기동물데이터

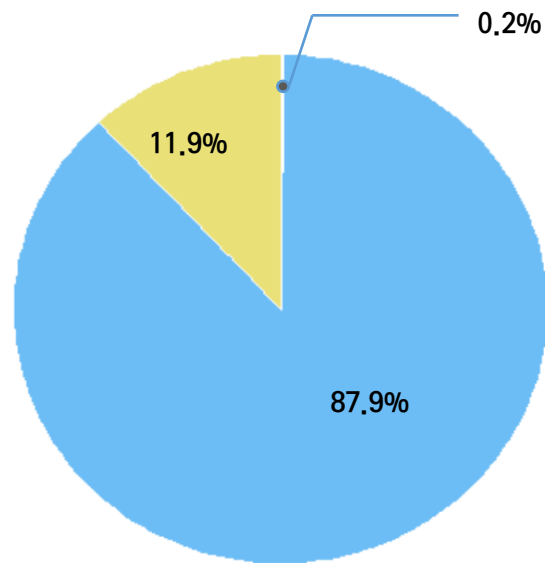
지역특성데이터

다음주 예고


변수 나누기 | 범주형데이터 | 수치형데이터 | 주소데이터 | 시간데이터 | 텍스트데이터

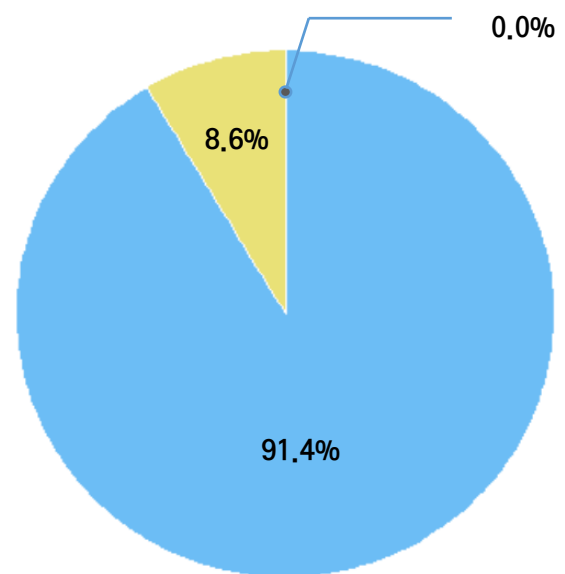
입양 기간

<강아지 입양 기간> 

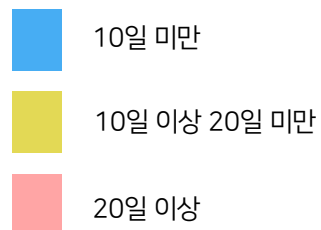


약 90% 10일 이내 입양
나머지 10% 20일 이내 입양

<고양이 입양 기간> 



약 92% 10일 이내 입양
나머지 8% 20일 이내 입양



입양된 강아지/고양이 기준 공고기간!

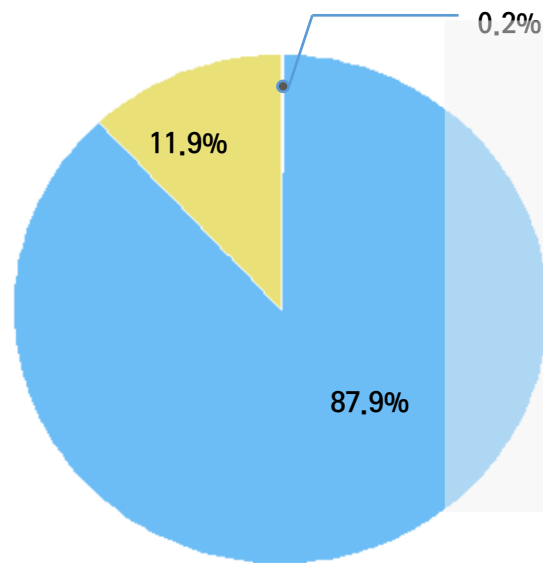
변수 나누기 | 범주형데이터 | 수치형데이터 | 주소데이터 | 시간데이터 | 텍스트데이터

입양 기간



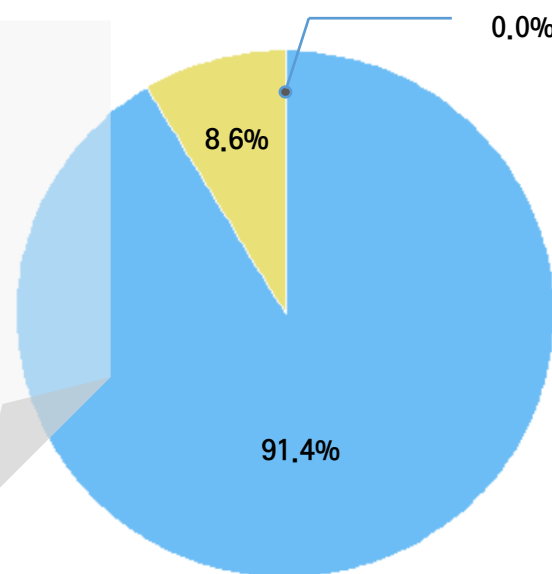
입양된 강아지/고양이 기준 공고기간!

<강아지 입양 기간>



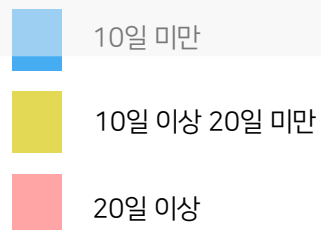
약 90% 10일 이내 입양
나머지 10% 20일 이내 입양

<고양이 입양 기간>



약 92% 10일 이내 입양
나머지 8% 20일 이내 입양

모두 입양기간은
20일 넘기지 않음



주제 소개

유기동물데이터

지역특성데이터

다음주 예고

특징 전처리

2020년 데이터의 특징 변수를 명사 리스트로 전처리

specialMark
교통사고추정, 골반 부상, 소심한 아이
3개월, 6마리입소, 대형견 아기, 유일하게 귀가 쫄긋함, 가장 소심하고 새침한 눈빛으로 애교발산
골반골절, 대퇴골 골절
중구2-208호, 경계심함, 설사
다른 변수와 중복 활발 / 입양예정
중성화 되어 있으며 견주의 건강악화로 인하여 보호소에서 입소하여 보호중
...

다른 변수와 중복

제각각 구분기호

특징 변수의 서술 내용과 형식이 모두 다름

*이해를 돕기 위한 처리 예시입니다.

specialMark
[교통사고] [추정] [골반] [부상] [소심한] [아이]
[대형견] [아기] [유일하게] [쫄긋]
[가장] [소심하고] [새침한] [눈빛] [애교발산]
[골반골절] [대퇴골] [골절]
[경계심함] [설사]
[활발] [입양예정]
[견주의] [건강악화] [보호소] [입소] [보호중]
...

extractNoun

숫자 / "중성화" / 영어 / 특수문자 포함 명사 제외

50

질병기술유무 변수 만들기

한국과학기술정보연구원_동물질병데이터_20201203.csv

질병명	축종	정의	...
갑상선기능저하증(Hypothyroidism)	개, 고양이	갑상선에서 T3, T4의 부적절한 생산, 분비로 세포대사활성 및 행동기능의 저하가 초래되는 질환	...
갑상선기능항진증 (Hyperthyroidism)	개, 고양이	T3와 T4의 분비와 과도한 생산물에 결과로 나타난 질병	...
코로나 바이러스 감염증, 개(Canine Coronavirus Infection)	개	개 코로나 바이러스(CCV)에 의해 발생하며 구토와 설사를 수반하는 장염을 유발하는 질병	...
파보바이러스 감염증, 개(Canine Parvoviral Infection)	개	주로 3주미만의 어린 개에서 출혈성 장염을 일으키는 급성 위장관 질환	...

1529 x 16

질병명 기술에 사용된 **명사 추출**
영어나 숫자가 들어가지 않은 것만 필터링

unique

질병 단어 사전.txt



질병기술유무 변수 만들기

한국과학기술정보연구원 동물질병데이터 20201203.csv

단어 사전 만들 때 **축종도** 고려한**고양**?

동물의 병명을 기술하는데 자주 쓰이는 어휘를
파악하려는 목적으로 사전을 만든거라서!!
어떤 동물이 특정 병에 걸리는지 같은 정보는 신경쓰지 않았어!



질병명 기술에 사용된 **명사** 추출
영어나 숫자가 들어가지 않은 것만 필터링



unique

질병 단어 사전.txt

질병기술유무 변수 만들기

질병 단어 사전.txt

간뇌
간대성근경련
간병
간선충피부
간섬유화
간세포
간질
간혈적
간혈성
간흡충
...
힘줄윤활막



질병기술유무

1

0

질병 단어 사전과 겹치는 명사가 한 개 이상 있으면 **1**
사전에 있는 명사가 없으면 **0**

주제 소개

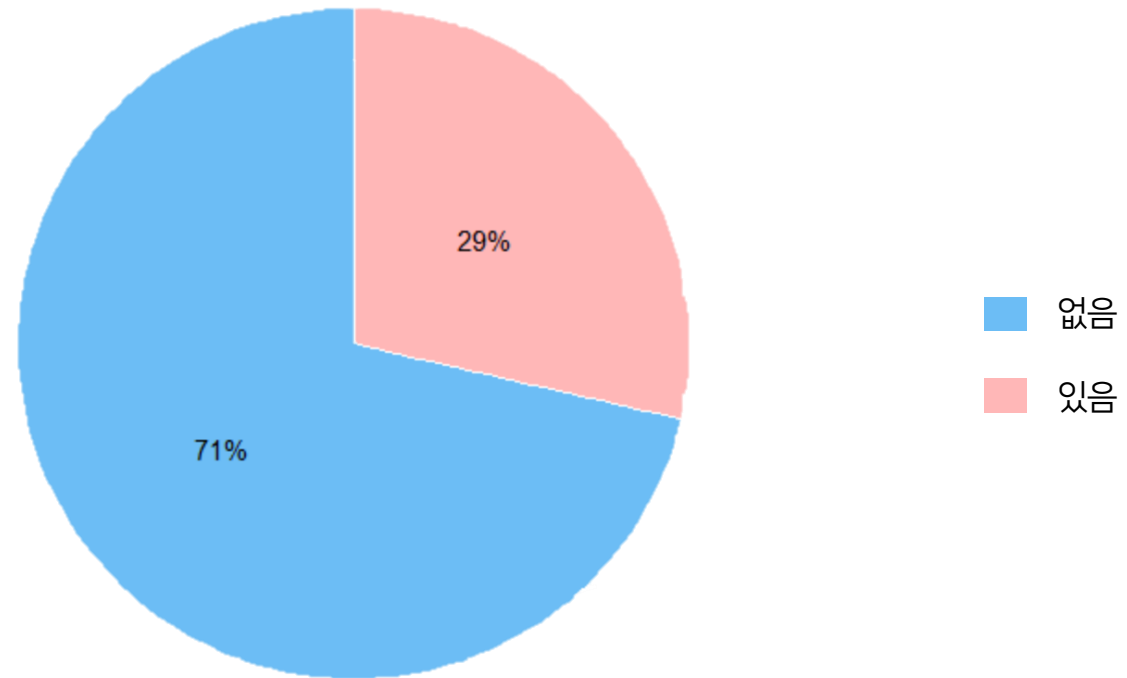
유기동물데이터

지역특성데이터

다음주 예고

질병기술유무 변수 시각화

<2020 유기동물 전체 질병변수 비율>



주제 소개

유기동물데이터

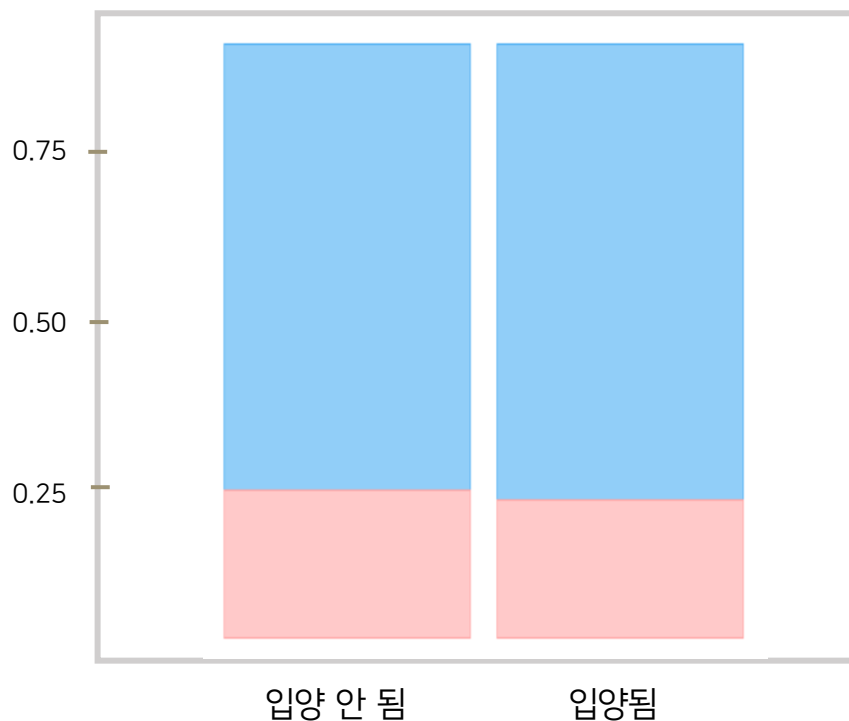
지역특성데이터

다음주 예고

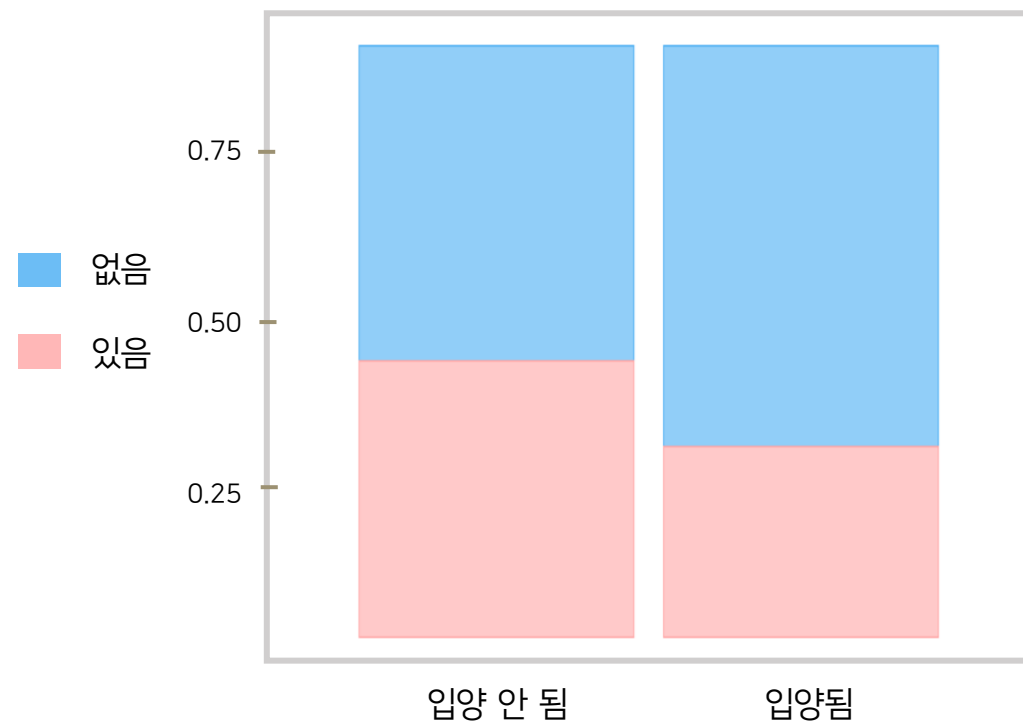
질병기술유무 변수 시각화



<강아지 질병 변수 비율>



<고양이 질병 변수 비율>



주제 소개

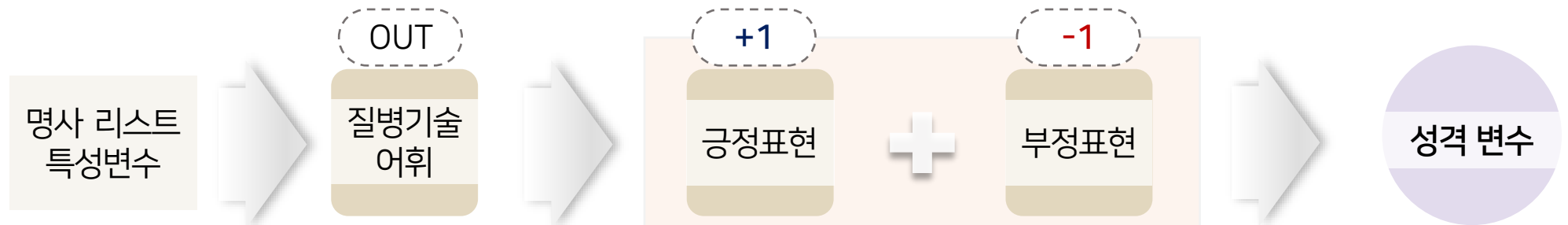
유기동물데이터

지역특성데이터

다음주 예고

변수 나누기 | 범주형데이터 | 수치형데이터 | 주소데이터 | 시간데이터 | 텍스트데이터

성격 변수 만들기



감성분석

텍스트에 나타난 사람들의 태도, 의견, 성향과 같은 주관적인 데이터를 분석하는 자연어 처리 기술
각 단어가 가지는 감정을 부정(-1) 혹은 긍정(+1)의 점수로 판단
댓글의 긍/부정 판단, 리뷰 분석, 콜센터 메시지분석 등에 사용

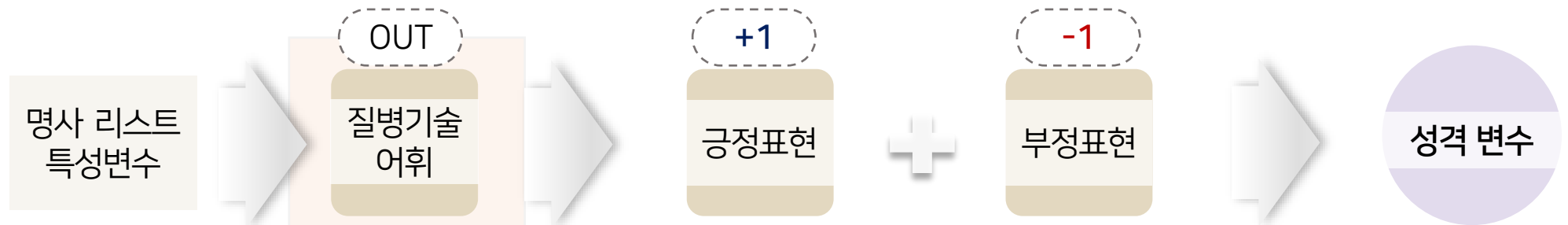
데이터 수집

주관성 탐지

극성탐지

변수 나누기 | 범주형데이터 | 수치형데이터 | 주소데이터 | 시간데이터 | 텍스트데이터

성격 변수 만들기



감성분석

질병기술에 사용되는 어휘는 부정으로 탐지될 가능성이 큼
질병기술유무에 이어 중복되어 사용되는 것을 방지

질병 단어 사전의 병명

피부염

부정어사전의 부정표현

성격 변수 만들기

군산대학교 한국어 감성사전.txt

긍정 단어 사전.txt

가격이 싸다
가까이 사귀어
가까이하다
...
적극적이다
승리
승리하다
유명하다

부정 단어 사전.txt

가난
가난뱅이
가난살이
...
의혹
내팽개치다
황령
불안증



이모티콘, 특수문자가 들어간 단어는 뺐어!
내가가내가가



명사만 나오게 처리도 해줬어!
내가가내가가

성격 변수 만들기

질병 단어 사전.txt와
겹치는 어휘 정리

● all_positives.txt

● all_negatives.txt

데이터별 극성탐지

[경계함] [귀여움] [순함]

[귀여움] [순함]

+2

[경계함]

-1

성격

1

0

-1

1 긍정

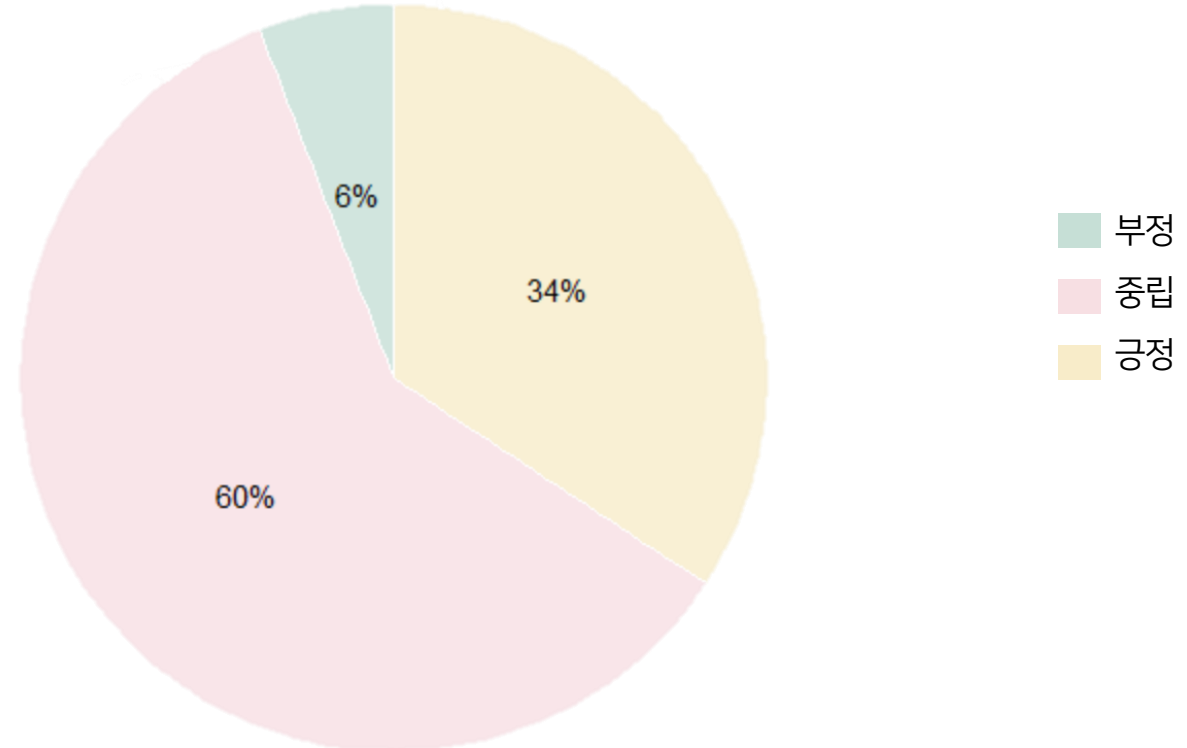
0 중립

-1 부정

긍정

성격 변수 만들기

<2020 유기동물 전체 성격변수 비율>



주제 소개

유기동물데이터

지역특성데이터

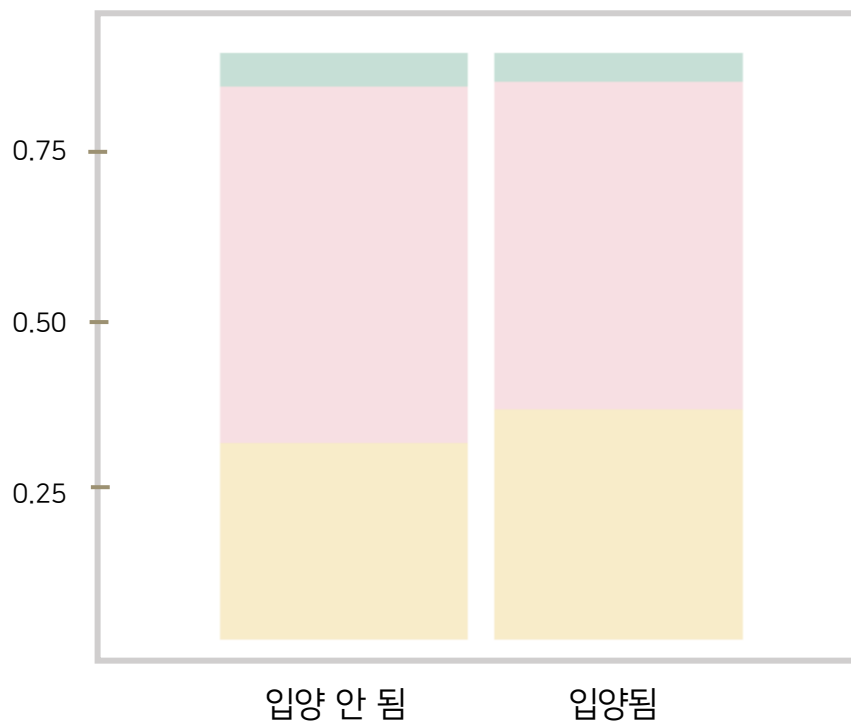
다음주 예고

변수 나누기 | 범주형데이터 | 수치형데이터 | 주소데이터 | 시간데이터 | 텍스트데이터

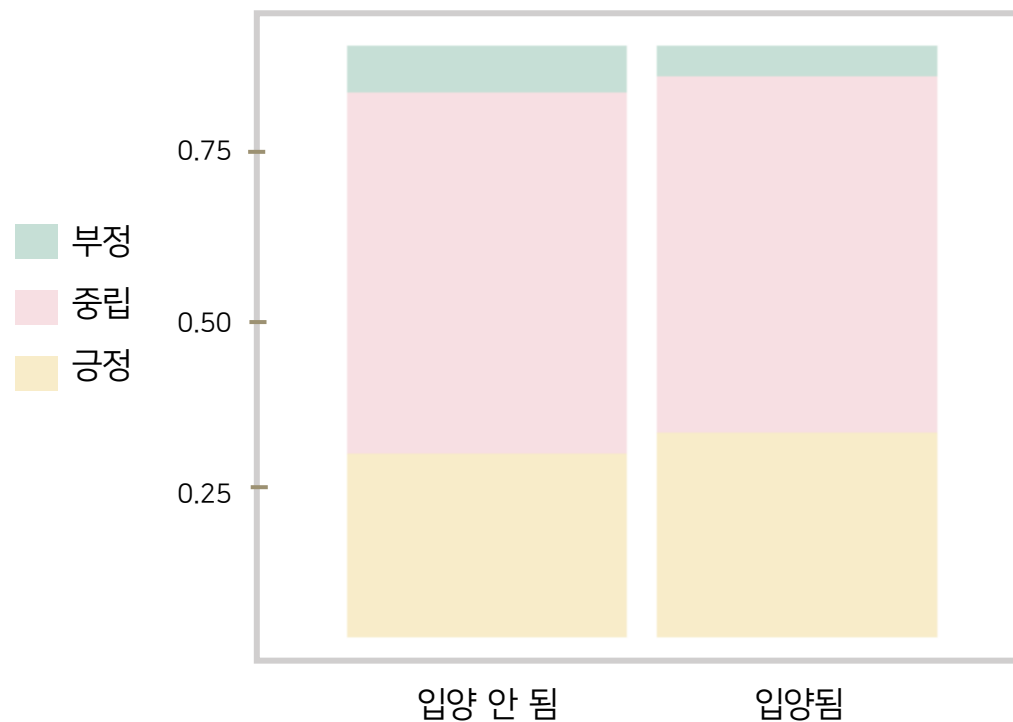
성격 변수 만들기



<강아지 성격 변수 비율>



<고양이 성격 변수 비율>



주제 소개

유기동물데이터

지역특성데이터

다음주 예고

최종 데이터셋 소개

2020.01 ~ 2020.12

Age	careAddr	careNm		Kind	Kind_spec	...	Weight	...	State	adoptYN	Duration	Season	Disease	mood
7	강원도 원주시	횡성유기동물보호센터	...	개	말티즈	...	5.5		입양	1	11	Winter	0	0
7	강원도 원주시	횡성유기동물보호센터	...	개	웰시	...	10.0		입양	1	11	Winter	0	0
6	강원도 원주시	횡성유기동물보호센터	...	개	믹스견	...	5.0		입양	1	7	Winter	1	0
6	부산광역시 해운대구	부산동물보호센터	...	개	믹스견	...	3.0		입양	1	11	Winter	1	0

 주제 소개

 유기동물데이터

 지역특성데이터

 다음주 예고



지역특성데이터



지역특성데이터

데이터 수집

전처리 및 EDA

최종데이터셋

데이터 소개



경제



1인당 지역내총생산(GRDP)

KOSIS 국가통계포털

지역별_경제활동별_지역내총생산 데이터 (2017)

시군구



1인당 지역총소득

행정구역_시군구별_주민등록인구 데이터 (2017)



1인당 개인소득



1인당 민간소비

시도별_1인당_지역총소득_개인소득_민간소비 데이터 (2019)

시도



인구



주민등록세대 수

행정구역_시군구별_주민등록세대수 데이터 (2020.12)

시군구

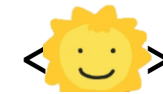


서비스



동물병원 개수

동물보호관리시스템 내 동물병원 정보 크롤링

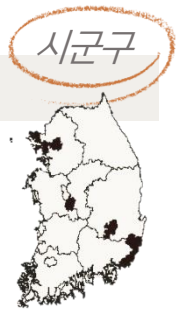


시군구



시군구 정보를 '대전광역시 중구'와 같은 형식으로 통일

'중구'가 무려
6개의 시도 안에..!



경제 지표 - (1) 1인당 지역 내 총생산(GRDP)

1인당
지역 내
총생산

=

2017

지역 내 총생산

÷

2017

주민등록인구



통계청

- 산업 구조 및 규모
- 경제적 성장률



지역 내 총생산(GRDP)이란?

- ① 해당 지역에 사업장을 둔 기업에서 발생한 소득

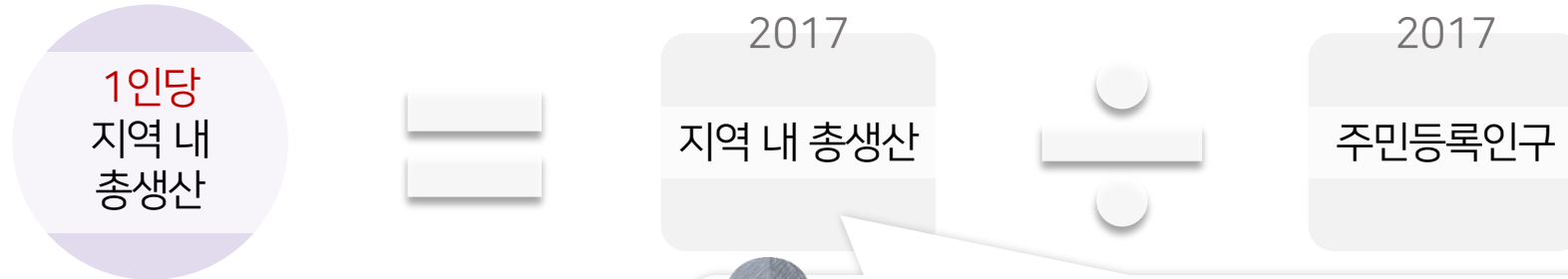
주제 소개

유기동물데이터

지역특성데이터

다음주 예고

경제 지표 - (1) 1인당 지역 내 총생산(GRDP)



통계청

- 산업 구조 및 규모
- 경제적 성장률



지역 내 총생산(GRDP)이란?

- ⊙ 해당 지역에 사업장을 둔 기업에서 발생한 소득

경제 지표 - (2) 1인당 지역 총소득 · 개인소득 · 민간소비



시군구 단위의 소득 데이터를 얻고 싶었지만 ...

통계청 「가계조사」

근로자 가구와 사업자 가구의 소득을 종합한 통계

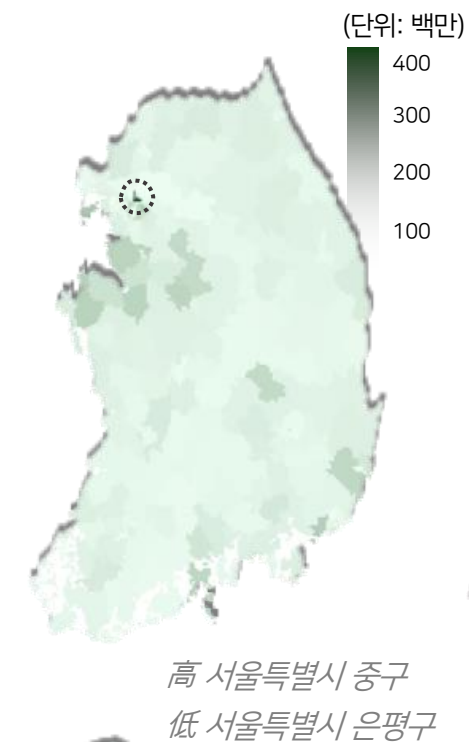
2000년 이후 폐지

생산, 소비, 물가 등의 기초통계를 바탕으로 추계한 해당 지역의 소득

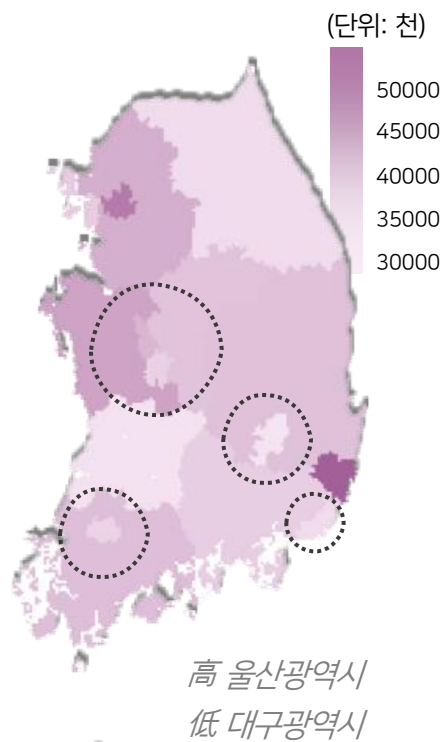
시군구	지역 내 총생산	지역 총소득	개인소득	민간소비
전라남도 목포시	17.49479	35532.06	18710.62	16103.89
전라남도 순천시	21.34757	35532.06	18710.62	16103.89
전라남도 나주시	38.92824	35532.06	18710.62	16103.89

같은 시도에 속해 있다면 같은 값 부여

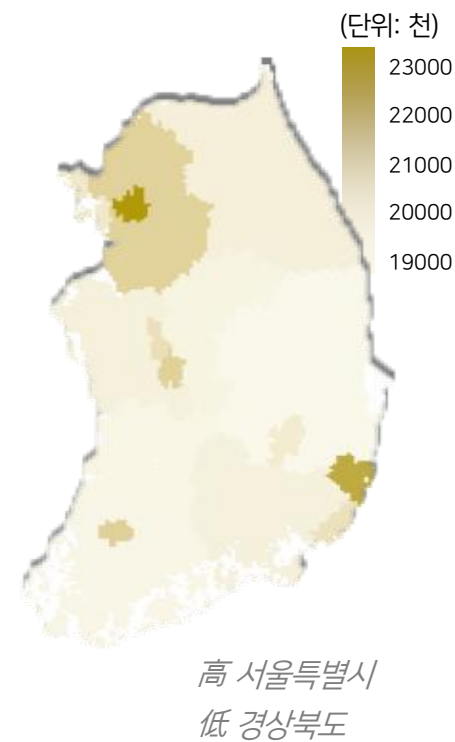
경제 지표 관련 변수 시각화



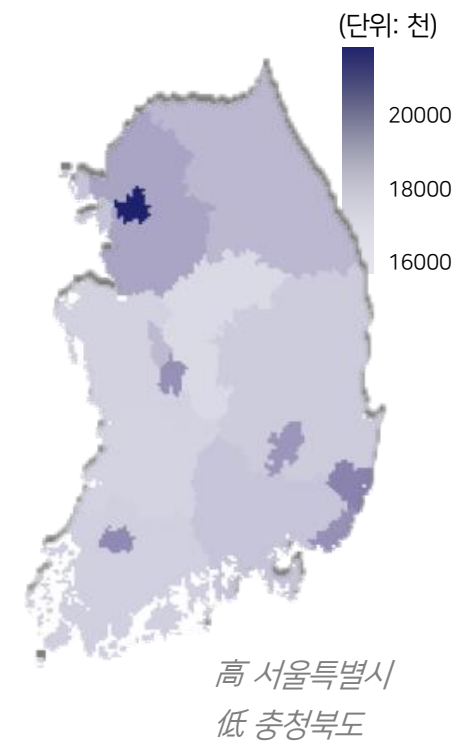
1인당 지역 내 총생산



1인당 지역 총소득



1인당 개인소득



1인당 민간소비

주제 소개

유기동물데이터

지역특성데이터

다음주 예고

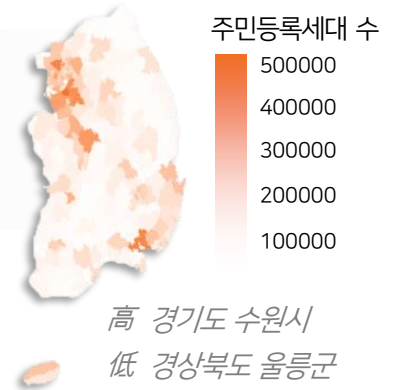
인구 지표 - 주민등록세대 수



가족의 구성원으로 들이는 것이기에
인구 수가 아닌 세대 수로 고려

익히 알고 있는 도시들에
상대적으로 많은 가구가
거주 중인 것으로 보임

- 유기동물 데이터에 맞춰 2020년 12월 데이터로 수집



주제 소개

유기동물데이터

지역특성데이터

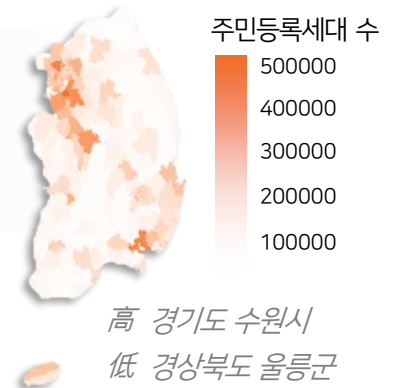
다음주 예고

인구 지표 - 주민등록세대 수



가족의 구성원으로 들이는 것이기에
인구 수가 아닌 세대 수로 고려

익히 알고 있는 도시들에
상대적으로 많은 가구가
거주 중인 것으로 보임

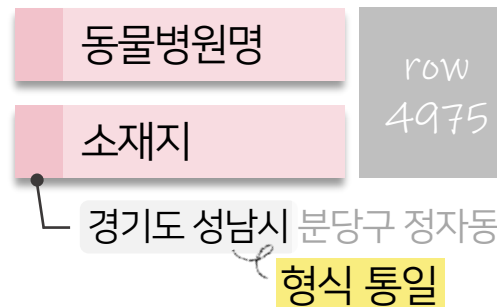


- 유기동물 데이터에 맞춰 2020년 12월 데이터로 수집

서비스 지표 - 동물병원 개수



- 유기된 동물에 대한 꾸준한 관리 용이



NA 104개

직접 검색
후 기재

총 4952개

주민등록세대 수의 분포와 유사

高 경상남도 창원시 (129개)
低 충청북도 단양군 (1개)
경상북도 울릉군
인천광역시 옹진군
전라남도 신안군



최종 데이터셋 소개

시군구	(단위: 백만) 1인당 지역 내 총생산	(단위: 천) 1인당 지역 총소득	(단위: 천) 1인당 개인소득	(단위: 천) 1인당 민간소비	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
부산광역시 동구	53.24438	29388.16	19680.32	18029.58	46003	5
부산광역시 영도구	20.56649	29388.16	19680.32	18029.58	54903	5

2017 시군구 2019 시도 2020.12 시군구



데이터상의 한계

- 최근 데이터가 아니며 시점이 일치하지 않음
- 보다 세밀한 단위의 데이터를 수집하지 못함



추가적 전처리 방안

- '개인소득'과 '민간소비'의 분포가 유사함
- 상관관계를 고려해 새로운 경제 지표 개발



다음주 예고



다음주 예고

일단 똑딱똑딱 만들 수 있는
모든 변수를 만든 선대팀!!
와다다다 코딩 공장을 돌리다가 정신을 차리니
그들 앞에는 변수 태풍이 다가오고 있었어요...
높은 파도에 정신을 잃기 일보 직전이라네요?



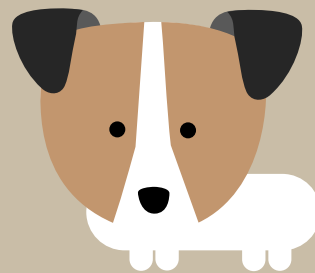
- 주제 소개
- 유기동물데이터
- 지역특성데이터
- 다음주 예고

다음주 예고

하지만 선대는 아직 희망을 잃지 않았어요!
미래의 본인들이 어떻게든 해줄거라 믿고 있거든요^^
다음주엔 어떤 변수가 의미있는지 **검정**하고
클러스터링도 좀 해보고 **예측 모델**을 개발하면서
입양 선호 유기동물의 특성을 파악해볼게요!
이를 바탕으로 유기동물 관리에 대해
멋진 의견까지 낸다면 정말정말 좋겠어요

- 주제 소개
- 유기동물데이터
- 지역특성데이터
- 다음주 예고





감사합니다