

# 서울시 공공자전거 수요 예측



Sponge

# LIFE IS LIKE RIDING A BICYCLE.

To keep your balance  
you must keep moving.

Albert Einstein

SYMPHONY OF LOVE  
Photo by Unsplash

# CONTENTS

- 01 분석개요
- 02 데이터 정의서
- 03 데이터 검증 결과서
- 04 데이터 탐색
- 05 데이터 분석 보고서
- 06 결론

VIVALDI

비전을 발견하고 디자인하라.

## 자전거

두 발로 가는 자전거는  
동력이 없으면 쓰러집니다.

두 발로 걷는 사람은  
희망이 없으면 쓰러집니다.



# 01

## 분석개요

- 분석 목적 및 필요성

## 분석 목적 및 필요성

분석 목적	서울시 공공자전거 수요 예측
분석 필요성	<p>공공자전거 시스템은 Rental 및 Return Back의 전체 프로세스가 자동으로 이루어지는 전통적인 자전거 렌탈의 차세대 제품이다. 이 시스템은 버스나 지하철과 같은 다른 운송 서비스와는 달리 여행, 출발 및 도착 위치의 시간이 시스템에 명시적으로 기록되는 원리이다.</p> <p>이러한 시스템은 교통, 환경 및 건강 문제에서 중요한 역할을 하기 때문에 현재 전 세계적으로 500개 이상의 자전거 공유 프로그램이 구성되어 있다.</p> <p>따라서 이러한 시스템을 좀 더 효과적으로 체계화시키기 위해 공유자전거 대여 수요를 예측해 볼 필요가 있다.</p>
보유데이터 현황	<p>서울시 공공자전거(따릉이) 대여 현황 (2016.01.01 ~ 2017.12.31)</p> <p>서울시 구별 대기환경 데이터 (2016.01.01 ~ 2017.12.31)</p> <p>서울시 읍면동별 주민등록 인구 데이터 (10세~69세) (2016.01 ~ 2017.12)</p> <p>서울시 구별 날씨 데이터 (2016.01.01 ~ 2017.12.31)</p> <p>서울시 아파트 매매가격지수 데이터 (종합주택유형) (2016.01 ~ 2017.12)</p>

## 분석 목적 및 필요성

### ◆ 국내 공공자전거 수요의 감소와 증가 사례

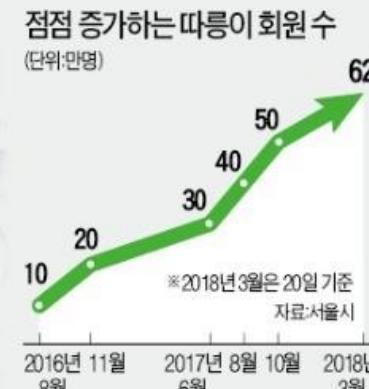
#### 창원 누비자 이용 갈수록 준다

개통 첫해 2008년 회원 1598명  
2013년 7만명 넘긴 후 감소세  
인구감소·미세먼지 등 영향탓

기사입력 : 2017-04-19 22:00:00

출처 : 경남신문

처음 개통 때 1598명이 누비자 회원으로 가입했다. 2009년 3만7759명, 2010년 4만8419명, 2011년 6만9284명, 2012년 6만8823명, 2013년 7만1304명으로 대체로 해마다 늘었다. 그러나 2014년 6만1676명, 2015년 6만1097명, 2016년 5만3940명까지 떨어졌다. 비회원인 1일 이용자 역시 2013년 23만5801명에서 2016년 19만5169명으로 감소했다.



창원시 생태교통과 관계자는 “이용률이 떨어진 가장 큰 원인으로 날씨나 미세먼지 등 환경적인 영향으로 보고 있다. 그 외 인구 감소나 지역 재개발 등 복합적인 영향을 받았다”면서 “신규 사업 등 회원을 확보하기 위해 노력하고 있다. 올 하반기 이후로 이용률이 조금씩 회복할 것으로 기대하고 있다”고 말했다.

## 분석 목적 및 필요성

### ◆ 국내 시행중인 공공자전거 대여 시스템

**고양시  
"피프틴"**

프랑스 파리의 공공자전거  
임대서비스  
'벨리브(Velib)'를  
벤치마킹한 것으로 유명한  
피프틴. 고양시 전역에  
설치된 140개의 피프틴  
파크(대여소)에서 총  
3,000대의 자전거를  
제공하고 있다.

지자체브리핑 KOREA KR

### 서울특별시 "따릉이"

따릉이는 '자전거와  
함께하는 건강한  
도시, 세계적인  
자전거 도시'라는  
슬로건으로  
2015년부터  
서울시에서 운영하는  
공용자전거 서비스다.  
서울시는 올해  
7월까지 자전거를 총  
2만대까지  
확충하기로 계획했다.

### 대전광역시 "타슈"

타라는 뜻의 충청도 사투리에서 착안한 타슈는  
대전광역시 전역의 144개 대여소에서 이용이  
가능하다.

### 순천시 "온누리"

모든 세상을 뜻하는 옛 우리말에서 따온  
온누리는 순천시 총 22곳에 설치되어 운영 중  
이다.

### 창원시 "누비자"

누비다와 자전거의 합성어인 누비자는 전국에서  
최초로 운영된 공용자전거 서비스다. 창원시  
전역 260여개의 무인터미널에서 총 3,000개의  
자전거를 이용 가능하다.

- ✓ 수원시, 제주시, 양산시, 공주시, 세종시, 안산시, 시흥시, 이천시, 과천시, 부천시 등 많은 지역들이  
공공자전거 시스템을 운영 중에 있음

## 분석 목적 및 필요성

- ◆ 신문/SNS를 활용한 웹 스크래핑
- ◆ Keyword : '공공자전거'
- ◆ 분석대상 : 2018년 최근 게시물 총 1161개

언론사/SNS	기사/글 갯수						
연합뉴스	220	동아일보	45	한국일보	54	조선일보	24
뉴시스	181	중앙일보	43	세계일보	48	내일신문	24
아시아투데이	113	한겨례	34	동아일보	48	JTBC	19
매일일보	83	국민일보	32	트위터	42	한국경제TV	9
경향신문	62	문화일보	31	서울신문	46	시사IN	3

## 분석 목적 및 필요성

◆ Keyword : '공공자전거'를 활용한 Wordcloud



따릉이 안전하게 즐기기



## 02

### 데이터 정의서

- 데이터 목록
- 데이터 정의 및 설명

## 데이터 목록

번호	데이터 보유기관	데이터 종류 및 내용	비고
1	정보공개포털	서울시 공공자전거(따릉이) 이용 현황 (2016-01-01 ~ 2017-12-31)	<a href="https://www.open.go.kr">https://www.open.go.kr</a>
2	서울시 대기환경정보	서울시 구 별 대기환경 데이터 (2016-01-01 ~ 2017-12-31)	<a href="http://cleanair.seoul.go.kr">http://cleanair.seoul.go.kr</a>
3	서울시 자동관측기상시스템	서울시 구 별 날씨 데이터 (2016-01-01 ~ 2017-12-31)	<a href="http://aws.seoul.go.kr">http://aws.seoul.go.kr</a>
4	한국 감정원 (Korea Appraisal Board)	서울시 아파트 매매가격지수 (종합주택유형) (2016-01-01 ~ 2017-12-31)	<a href="http://www.kab.co.kr">http://www.kab.co.kr</a>
5	국가통계포털	서울시 읍, 면, 동 별 주민등록 인구데이터 (10세 ~ 69세) (2016-01-01 ~ 2017-12-31)	<a href="http://kosis.kr">http://kosis.kr</a>

- ✓ 서울시 아파트 매매가격지수 : 서울시 공공자전거 이용특성에 관한 연구  
(서울도시연구 제17권 제4호 2016. 12, 논문 pp. 77 ~ 91 77)

- ✓ 서울시 공공자전거 수요를 예측하기 위해 서울시 공공자전거 이용 현황 데이터와 기상 및 대기환경 데이터, 서울시 아파트 매매가격지수 데이터를 융합하여 분석하기로 함

## 데이터 정의 및 설명

기본정보		법적사항	
개요	서울시 공공자전거(따릉이) 이용 현황	개인정보 항목	개인정보 없음
데이터 보유기관	정보공개포털( <a href="https://www.open.go.kr">https://www.open.go.kr</a> )	비식별화 여부	해당사항 없음
파일명	따릉이 대여소별 대여내역(2016년).xlsx, 따릉이 대여소별 대여내역(2017년).xlsx		
데이터 상세항목			
번호	열 이름	유형	설명
1	대여소 그룹	범주형	서울시 25개 구 ex) 광진구, 마포구, ...
2	대여소 명	범주형	각 구에 대한 대여소 명 ex) 500. 어린이대공원역 3번 출구 앞, ...
3	대여 일자/월	범주형	대여 년, 월, 일 ex) 2016-01-01, ..., 2017-12-31
4	대여 건수	이산형	공공자전거 대여 횟수 ex) 0, 7, 0, ...

- ✓ 각 구 별 데이터 갯수에 차이가 존재 : 용산구, 양천구 (2016-07-18 ~ 2017-12-31)  
 동대문구 (2016-07-13 ~ 2017-12-31)  
 은평구 (2016-09-06 ~ 2017-12-31)

## 데이터 정의 및 설명

기본정보		법적사항	
개요	서울시 구 별 대기환경 데이터	개인정보 항목	개인정보 없음
데이터 보유기관	서울시 대기환경정보( <a href="http://cleanair.seoul.go.kr">http://cleanair.seoul.go.kr</a> )	비식별화 여부	해당사항 없음
파일명	2016_eco.slx, 2017_eco.slx		
데이터 상세항목			
번호	열 이름	유형	설명
1	날짜	범주형	측정 년, 월, 일 ex) 2016-01-01, ..., 2017-12-31
2	측정소명	범주형	서울시 25개 구 ex) 광진구, 마포구, ...
3	미세먼지(PM-10(µg/m³))	연속형	미세먼지 수치 ex) 48, 58, 61, ...
4	초미세먼지(PM-25(µg/m³))	연속형	초미세먼지 수치 ex) 26, 40, 48, ...
5	오존(O3(ppm))	연속형	오존 수치 ex) 0.024, 0.003, 0.002, ...
6	이산화질소(NO2(ppm))	연속형	이산화질소 수치 ex) 0.031, 0.051, 0.061, ...
7	일산화탄소(CO(ppm))	연속형	일산화탄소 수치 ex) 0.5, 1, 1.1, ...
8	아황산가스(SO2(ppm))	연속형	아황산가스 수치 ex) 0.005, 0.006, 0.007, ...

## 데이터 정의 및 설명

기본정보		법적사항	
개요	서울시 구 별 날씨 데이터	개인정보 항목	개인정보 없음
데이터 보유기관	서울시 자동관측기상시스템( <a href="http://aws.seoul.go.kr">http://aws.seoul.go.kr</a> )	비식별화 여부	해당사항 없음
파일명	광진(2016-01).xls, ..., 중구(2017-12).xls (서울시 11개의 구 별 데이터 2016-01-01 ~ 2017-12-31)		
데이터 상세항목			
번호	열 이름	유형	설명
1	날짜	범주형	측정 년, 월, 일 ex) 2016-01-01, ..., 2017-12-31
2	평균(풍속(m/s), 풍향(DEG))	연속 / 이산형	일 별 평균 풍속 및 풍향 ex) 0.8, 1.1, 1.3, ... /233, 226, 255, ...
3	최대(풍속(m/s), 풍향(DEG))	연속 / 이산형	일 별 최대 풍속 및 풍향 ex) 2.7, 2.9, 5, ... /254, 251, 290, ...
4	순간최고 (풍속(m/s), 풍향(DEG))	연속 / 이산형	일 별 순간최고 풍속 및 풍향 ex) 3, 3.3, 5.7, ... /258, 235, 306, ...
5	기온(°C)(평균, 최저, 최고)	연속형	일 별 평균, 최저, 최고 기온 ex) 1.5, 5.6, ... /-2.6, 2.4, ... /4, 9.2, ...
6	일강수(mm)	연속형	일 별 강수량 ex) 0, 0.5, ...
7	습도(%)(평균, 최저, 최고)	연속형	일 별 습도 ex) 대부분 결측값(자동관측시스템으로 측정불가 판단)
8	일사(MJ/m <sup>2</sup> )	연속형	일 별 일사량 ex) 0 (자동관측시스템으로 측정불가 판단)
9	일조(hour)	이산형	일 별 일조시간 ex) 0 (자동관측시스템으로 측정불가 판단)

## 데이터 정의 및 설명

기본정보		법적사항	
개요	서울시 아파트 매매가격지수 (종합주택유형)	개인정보 항목	개인정보 없음
데이터 보유기관	한국 감정원( <a href="http://www.kab.co.kr">http://www.kab.co.kr</a> )	비식별화 여부	해당사항 없음
파일명	월간_매매가격지수_아파트.xlsx		
데이터 상세항목			
번호	열 이름	유형	설명
1	지역	범주형	서울시 25개 구 ex) 광진구, 마포구, ...
2	날짜	범주형	측정 년, 월 ex) 2016-01, ..., 2017-12
3	매매가격지수	연속형	아파트 매매가격지수 ex) 94.5, 95.2, 93.8, ...

## 데이터 정의 및 설명

기본정보		법적사항	
개요	서울시 읍, 면, 동 별 주민등록 인구 데이터 (10세 ~ 69세)	개인정보 항목	개인정보 없음
데이터 보유기관	국가통계포털( <a href="http://kosis.kr">http://kosis.kr</a> )	비식별화 여부	해당사항 없음
파일명	서울특별시_읍면동별_5세별_주민등록인구_20180404155345.csv		
데이터 상세항목			
번호	열 이름	유형	설명
1	행정구역(동읍면)별	범주형	서울시 25개 구 ex) 광진구, 마포구, ...
2	날짜	범주형	측정 년, 월 ex) 2016-01, ..., 2017-12
3	10 - 14세	연속형	주민등록인구(10세 ~ 14세) ex) 5874, 5833, 5779, ...
4	15 - 19세	연속형	주민등록인구(15세 ~ 19세) ex) 8504, 8433, 8724, ...
5	20 – 24세	연속형	주민등록인구(20세 ~ 24세) ex) 11428, 11462, 11666, ...
6	25 – 29세	연속형	주민등록인구(25세 ~ 29세) ex) 11877, 11894, 11932, ...
:	:	:	:
14	65 – 69세	연속형	주민등록인구(65세 ~ 69세) ex) 7695, 7720, 7680, ...
15	100+	연속형	주민등록인구(100세 이상) ex) 214, 215, 220, ...

# 따릉이는 어떻게 생겼나요?



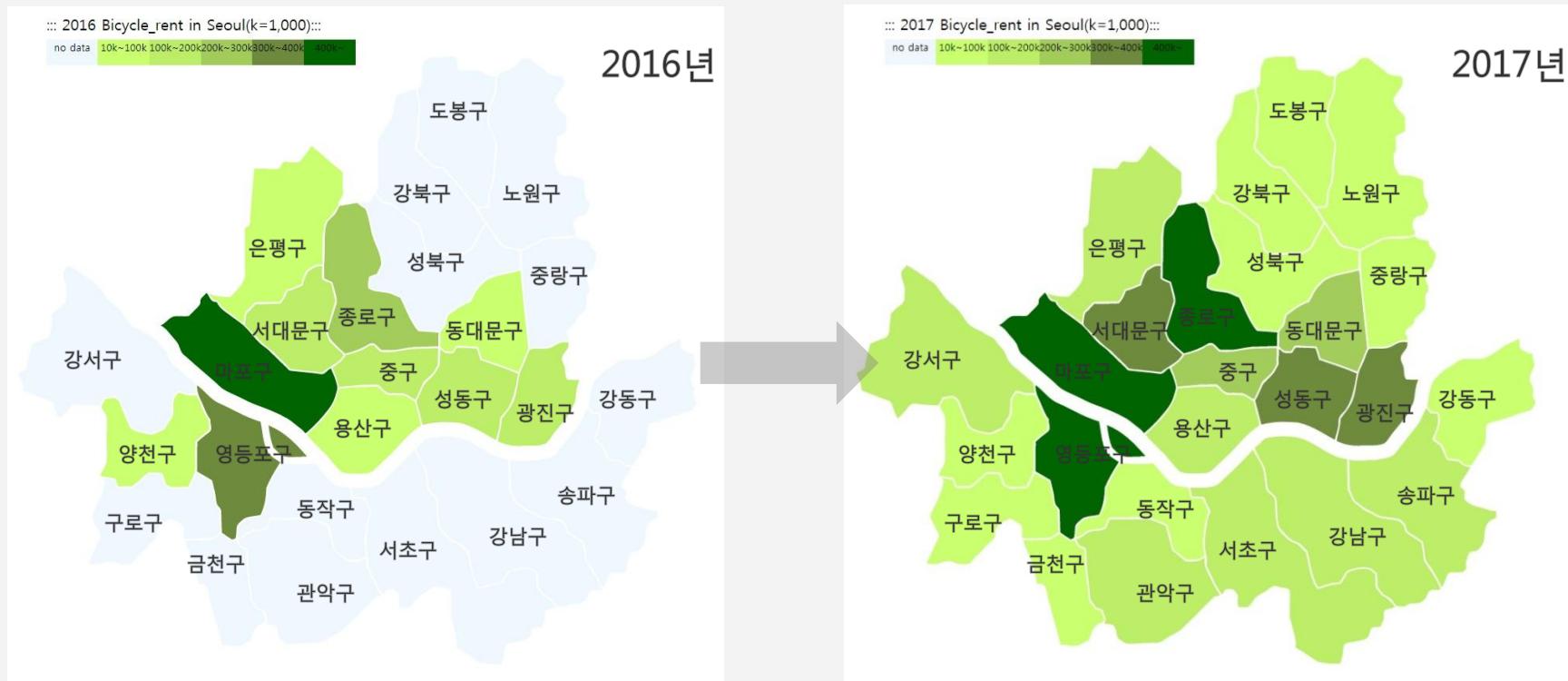
## 03

### 데이터 검증 결과서

- 데이터 연계안
- 최종 데이터

## 데이터 연계안

### ◆ 서울시 공공자전거 이용 현황



- ✓ 서울시 공공자전거는 450개소, 5,600대(2016년)에서 1,500개소, 20,000대(2017년)로 확대
- ✓ 본 연구는 서울시 11개 구 기준으로 2016년, 2017년 '서울시 공공자전거 이용 현황' 데이터 병합

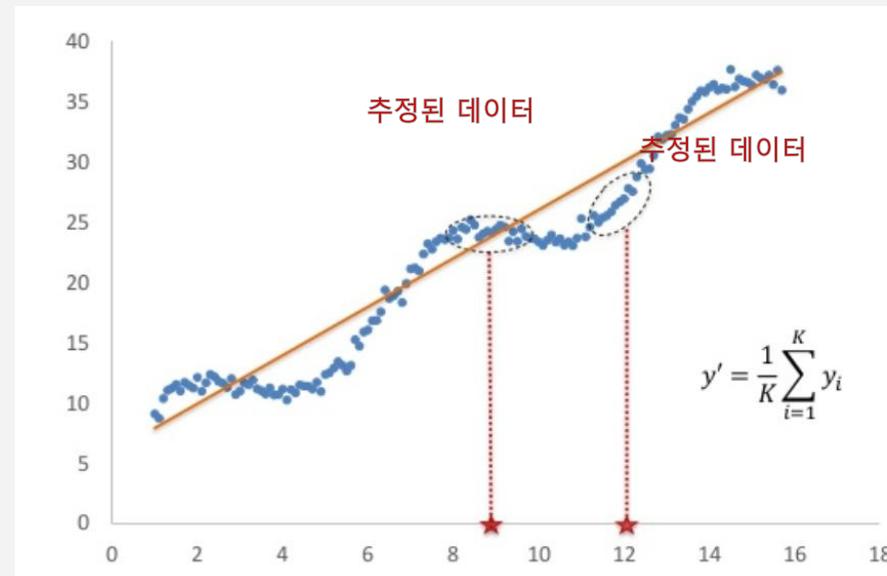
## 데이터 연계안

수집 데이터	변수 선택/파생변수 생성	결측값 유무 파악	병합/결측값 보정 방법
<p>1. 서울시 공공자전거 이용 현황  </p> <p>2. 서울시 구 별 대기환경 데이터  </p> <p>3. 서울시 구 별 날씨 데이터  </p> <p>4. 서울시 아파트 매매가격지수  </p> <p>5. 서울시 읍, 면, 동 별 주민등록 인구 데이터  </p>	<p>1. 서울시 공공자전거 이용 현황          대여소 그룹 -&gt; <b>area</b>,          대여 일자/월 -&gt; <b>날짜</b>,          대여건수 -&gt; <b>rentcnt</b>          날짜 -&gt; <b>month, day, weekday, dayofweek</b> 파생변수 생성</p> <p>2. 서울시 구 별 대기환경 데이터          날짜, 미세먼지, 초미세먼지, 오존, 이산화질소, 일산화탄소, 아황산가스</p> <p>3. 서울시 구 별 날씨 데이터          날짜, 평균풍속, 최대풍속, 순간최고풍속, 기온평균, 강수일강수, 습도평균</p> <p>4. 서울시 아파트 매매가격지수          날짜,          매매가격지수 -&gt; <b>아파트가격</b></p> <p>5. 서울시 읍, 면, 동 별 주민등록 인구 데이터          날짜,          10 - 14세 ~ 65 - 69세 -&gt; <b>인구</b>          파생변수 생성</p>	<p>1. 서울시 공공자전거 이용 현황          서울시 전체 25개 구 중에 2016년도 14개 구 결측값 존재</p> <p>2. 서울시 구 별 대기환경 데이터          미세먼지, 초미세먼지, 오존, 이산화질소, 일산화탄소, 아황산가스 결측값 존재</p> <p>3. 서울시 구 별 날씨 데이터          평균풍속, 최대풍속, 순간최고 풍속, 기온평균, 강수일강수, 습도평균 결측값 존재</p> <p>4. 서울시 아파트 매매가격지수          아파트가격 : 월 별 데이터          월에 따른 일 별 결측값 존재</p> <p>5. 서울시 읍, 면, 동 별 주민등록 인구 데이터          이구 : 월 별 데이터          월에 따른 일 별 결측값 존재</p>	<p>1. 서울시 공공자전거 이용 현황          Inner join (조인키 : area) : 서울시 11개 구 병합 -&gt;          일 별로 rentcnt 합 -&gt;          서울시 11개 구 별로 나눔</p> <p>2. 서울시 구 별 대기환경 데이터          Inner join (조인키 : 날짜) -&gt;          k-nearest neighbors Regression 방법으로 결측값 보정</p> <p>3. 서울시 구 별 날씨 데이터          Inner join (조인키 : 날짜) -&gt;          k-nearest neighbors Regression 방법으로 결측값 보정</p> <p>4. 서울시 아파트 매매가격지수          Full outer join (조인키 : 날짜) -&gt;          일 별 결측값을 월 별 값으로 대체</p> <p>5. 서울시 읍, 면, 동 별 주민등록 인구 데이터          Full outer join (조인키 : 날짜) -&gt;          일 별 결측값을 월 별 값으로 대체</p>

## K-NN Regression Imputation

- ✓ 데이터 결측값 존재 시, 예측하고자 하는 데이터로부터 가장 가까운 K개의 이웃을 찾은 뒤, 이들 이웃으로부터 예측하고자 하는 데이터의 분류를 정하여 값을 대체하는 방법
- ✓ K-NN Regression에서 K-NN 알고리즘은 연속 변수를 추정하는데 쓰임
- ✓  $n_{neighbors} = K$ , 교차검증(k - fold cross validation)을 하기 위한 k 설정 필요

K-NN Regression



### ◆ K-NN Regression 작동 원리

1. 추정하고자 하는 데이터로부터 분류된 데이터까지의 유clidean 거리(default)를 계산
2. 거리가 계산된 데이터를 오름차순으로 정렬
3. RMSE에 기반한 최근접 이웃들을 발견하는데 최적의 수를 찾음 (이 과정은 교차검증을 통해 이루어짐)
4. K-최근접 다변량 이웃들에서 거리의 역수를 가중치로 한 가중 평균을 계산

- ✓ 본 연구는  $n_{neighbors} = 3$ ,  $k = 10$  (10 – fold cross validation)으로 결측값 보정

(2016년 기상청 외 빅데이터 분석 서비스)

## 최종 데이터

일별 데이터 = 해당 월별 데이터

K-NN Regression imputation

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1	날짜	area	rentcnt	인구	아파트가구	평균풍속	최대풍속	순간최고풍속	기온평균	강수일강수습도평균	미세먼지	조미세먼지오존	이산화질소	일산화탄소	아황산가스	month	day	weekday	dayofweek		
2	2016-01-01	종로구	93	128540	94.5	0.3	3	4.5	2.8	0	66.9	68	51	0.005	0.047	0.7	0.006	1	1	1	
3	2016-01-01	광진구	33	303607	94	0.8	2.7	3.9	2.1	0	74.6	85	64	0.004	0.046	0.9	0.006	1	1	1	
4	2016-01-01	마포구	197	323075	93.4	1.5	3.6	4.8	2.6	0	74	59	49	0.008	0.042	1	0.006	1	1	1	
5	2016-01-01	성동구	40	248415	94.6	0.8	4	5.5	2.9	0	75.9	77	39	0.004	0.047	1	0.005	1	1	1	
6	2016-01-01	중구	39	103841	95.2	0.8	2.7	3	1.5	0	67	61	35	0.005	0.065	0.8	0.007	1	1	1	
7	2016-01-01	서대문구	41	239015	94.2	1.2	3.3	4.4	2.1	0	68.3	69	39	0.007	0.06	0.9	0.005	1	1	1	
8	2016-01-01	영등포구	80	316451	92.7	1	3.6	4	3.3	0	61.98889	80	40	0.006	0.035	1.1	0.006	1	1	1	
7153	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	
7154	2017-12-30	마포구	527	311598	100.3	1.3	2.7	3.6	1.9	0.5	78.8	130	99	0.004	0.055	1.2	0.005	12	30	0	
7155	2017-12-30	양천구	55	399424	102	1.3	2.7	3.6	1.9	0.5	78.8	115	102	0.004	0.057	1.2	0.005	12	30	0	
7156	2017-12-30	은평구	159	401107	100.2	1.3	2.7	3.6	1.9	0.5	78.8	121	89	0.006	0.029	1.1	0.004	12	30	0	
7157	2017-12-30	종로구	373	127305	100.2	1.3	2.7	3.6	1.9	0.5	78.8	112	88	0.003	0.059	1.3	0.006	12	30	0	
7158	2017-12-30	성동구	245	253544	101.6	1.3	2.7	3.6	1.9	0.5	78.8	130	84	0.003	0.046	0.7	0.006	12	30	0	
7159	2017-12-30	서대문구	260	256678	100.3	1.3	2.7	3.6	1.9	0.5	78.8	127	95.66667	0.005667	0.048333	0.833333	0.006333	12	30	0	
7160	2017-12-30	광진구	260	304786	101.6	1.3	2.7	3.6	1.9	0.5	78.8	136	96	0.005	0.06	1	0.005	12	30	0	
7161	2017-12-31	마포구	450	311598	100.3	2.4	5.9	8.2	1.1	0.5	56.5	63	40	0.018	0.023	0.7	0.005	12	31	0	
7162	2017-12-31	서대문구	202	256678	100.3	2.4	5.9	8.2	1.1	0.5	56.5	61.66667	38	0.019667	0.020333	0.466667	0.004667	12	31	0	
7163	2017-12-31	중구	113	103194	100.6	2.4	5.9	8.2	1.1	0.5	56.5	53	34	0.022	0.022	0.6	0.004	12	31	0	
7164	2017-12-31	영등포구	297	305918	101	2.4	5.9	8.2	1.1	0.5	56.5	70	42	0.018	0.021	0.7	0.007	12	31	0	
7165	2017-12-31	성동구	228	253544	101.6	2.4	5.9	8.2	1.1	0.5	56.5	62	37	0.019	0.016	0.4	0.005	12	31	0	
7166	2017-12-31	은평구	113	401107	100.2	2.4	5.9	8.2	1.1	0.5	56.5	67	33	0.024	0.018	0.8	0.003	12	31	0	
7167	2017-12-31	양천구	46	399424	102	2.4	5.9	8.2	1.1	0.5	56.5	58	41	0.014	0.028	0.7	0.005	12	31	0	
7168	2017-12-31	동대문구	154	289307	100.2	2.4	5.9	8.2	1.1	0.5	56.5	69	34	0.02	0.018	0.6	0.005	12	31	0	
7169	2017-12-31	광진구	230	304786	101.6	2.4	5.9	8.2	1.1	0.5	56.5	63	37	0.023	0.023	0.6	0.004	12	31	0	
7170	2017-12-31	용산구	81	188138	100.2	2.4	5.9	8.2	1.1	0.5	56.5	60	38	0.017	0.022	0.4	0.005	12	31	0	

✓ n = 7168, 변수의 총 개수(p) = 21 - 2(날짜, area) = 19 (구 별로 데이터를 나누어 분석 진행)

✓ Target Variable = rentcnt

## 최종 데이터

상세항목			
NO.	열이름	유형	설명
1.	날짜	범주형	년, 월 ex) 2016-01-01, ..., 2017-12-31
2.	area	범주형	서울시 11개 구 ex) 광진구, 마포구, ...
3.	rentcnt	이산형	공공자전거 대여 횟수 ex) 93, 33, 197, ...
4.	인구	범주형	주민등록인구(10세 ~ 69세) ex) 128540, 308607, 323075, ...
5.	아파트가격	연속형	아파트 매매가격지수 ex) 94.5, 94, 93.4, ...
6.	평균풍속	연속형	일 별 평균 풍속 ex) 0.3, 0.8, 1.5, ...
7.	최대풍속	연속형	일 별 최대 풍속 ex) 3, 2.7, 3.6, ...
8.	순간최고풍속	연속형	일 별 순간최고 풍속 ex) 4.5, 3.9, 4.8, ...
9.	기온평균	연속형	일 별 평균 기온 ex) 2.8, 2.1, 2.6, ...
10.	강수일강수	연속형	일 별 강수량 ex) 0, 0, 0, ...
11.	습도평균	연속형	일 별 평균 습도 ex) 66.9, 74.6, 74, ...

## 최종 데이터

상세항목			
NO.	열이름	유형	설명
12.	미세먼지	연속형	미세먼지 수치 ex) 48, 58, 61, ...
13.	초미세먼지	연속형	초미세먼지 수치 ex) 26, 40, 48, ...
14.	오존	연속형	오존 수치 ex) 0.024, 0.003, 0.002, ...
15.	이산화질소	연속형	이산화질소 수치 ex) 0.031, 0.051, 0.061, ...
16.	일산화탄소	연속형	일산화탄소 수치 ex) 0.5, 1, 1.1, ...
17.	아황산가스	연속형	아황산가스 수치 ex) 0.005, 0.006, 0.007, ...
18.	month	범주형	월 ex) 1, 2, ..., 12
19.	day	범주형	일 ex) 1, 2, .., 31
20.	weekday	범주형	평일/주말 여부 ex) 평일 : 1, 주말 : 0
21.	dayofweek	범주형	요일 ex) 월요일 : 0 ~ 일요일 : 6

- ✓ 11개의 구별로 데이터를 나누어 분석 하기로 함 (분석 목적 : 구별 rentcnt를 예측하는 알고리즘 개발)

# 2살 따릉이 파헤치기

여가용은 OK,  
대중교통은 아자…

서울시의 따릉이는  
지난 2015년 9월에 첫 선을  
보였습니다. 이제 두 돌을 앞두고  
있는 따릉이는 서울시민의  
일상에서 어떤 의미일까요?

따릉이의 실태, 현황, 그리고  
파급효과를 인포그래픽을  
통해 살펴봤습니다.

전승엽.이상서 기자 kirin@yna.co.kr

김유정.신아현 인턴기자

YONHAP NEWS AGENCY

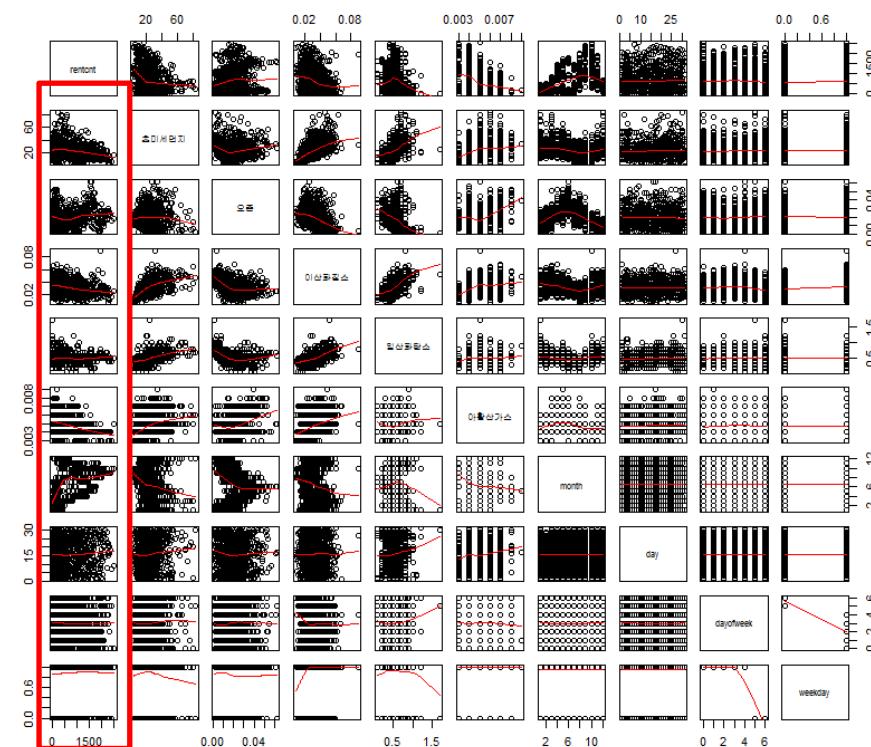
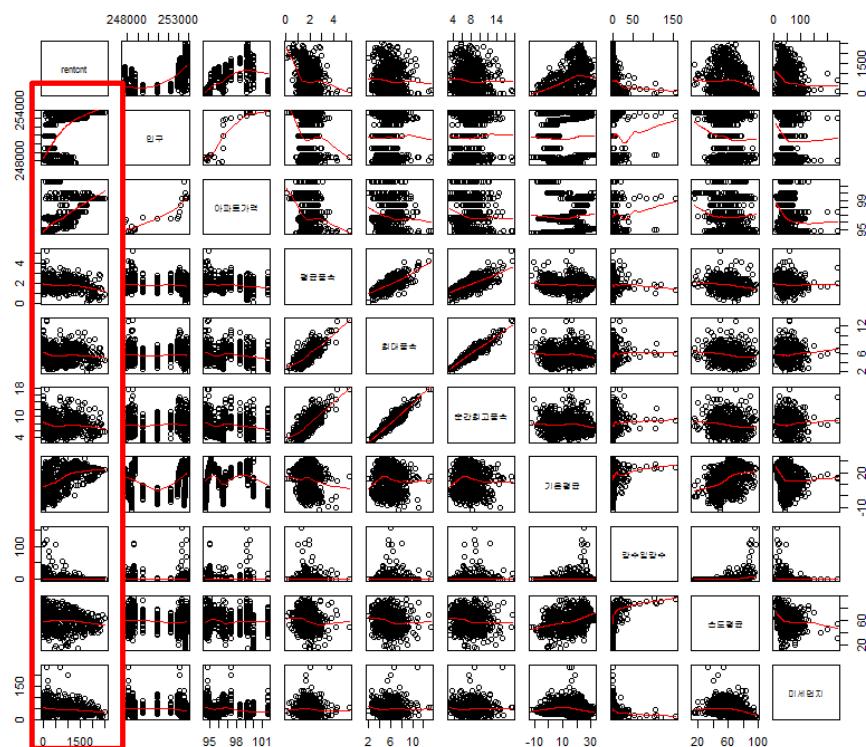


## 04

### 데이터 탐색

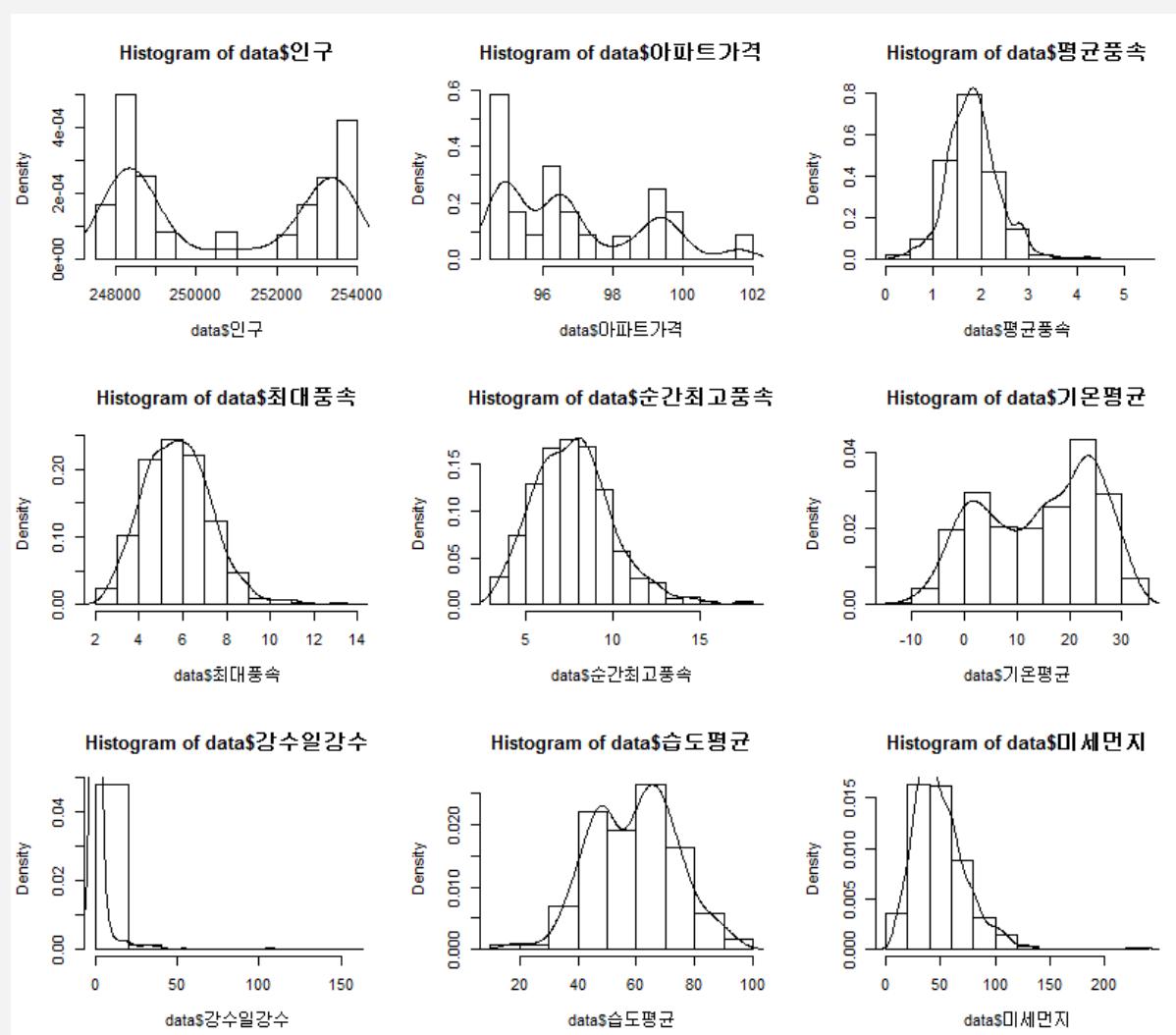
- 산점도
- 히스토그램
- 상관 분석

## 산점도

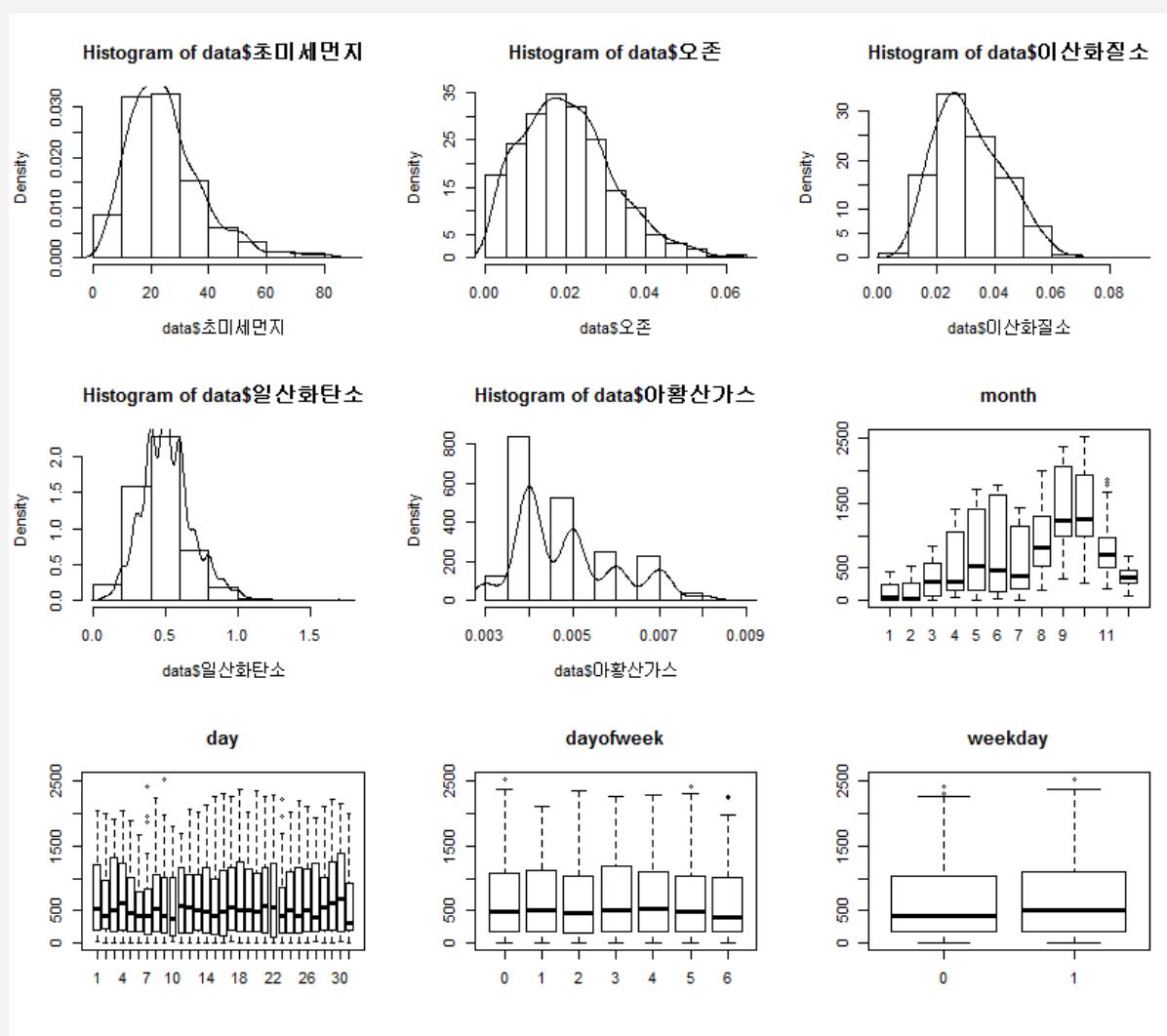


✓ 반응 변수와 설명 변수의 관계가 대부분 비선형 관계라는 것을 알 수 있음

## 히스토그램



## 히스토그램



## 상관관계 행렬

coefficient of correlation : pearson

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
1		rentcnt	인구	아파트가구	평균풍속	최대풍속	순간최고총기온평균	강수일강수습도평균	미세먼지	초미세먼지	오존	이산화질소	일산화탄소	아황산가스	month	day		dayofweek	weekday	
2	rentcnt	1	0.864849	0.867003	-0.30292	-0.25962	-0.24403	0.496241	-0.05	-0.01757	-0.60704	-0.53298	-0.02006	-0.4791	-0.18961	-0.75593	0.696611	-0.05954	-0.04712	0.027667
3	인구	1.77E-06	1	0.944129	-0.23139	-0.1537	-0.13625	0.142274	0.026048	-0.15293	-0.50916	-0.43157	-0.12829	-0.42719	-0.11351	-0.83242	0.527317	-0.07966	-0.0225	-0.00146
4	아파트가구	1.56E-06	1.28E-09	1	-0.39641	-0.3565	-0.33563	0.15483	0.063019	-0.01709	-0.57193	-0.42168	-0.32782	-0.35671	0.006145	-0.81653	0.76105	-0.05898	-0.01359	-0.0036
5	평균풍속	0.207447	0.340505	0.092905	1	0.963752	0.961947	-0.1036	0.050116	-0.36168	-0.06241	-0.3778	0.687624	-0.55122	-0.73683	0.075358	-0.5428	-0.08256	-0.01882	-0.00988
6	최대풍속	0.283095	0.529848	0.134081	3.47E-11	1	0.998625	-0.11517	0.073855	-0.38006	-0.03393	-0.36182	0.691743	-0.53038	-0.70644	0.036303	-0.59302	-0.11862	-0.03458	-0.00057
7	순간최고총기온	0.314015	0.578097	0.160083	5.21E-11	0	1	-0.10644	0.090941	-0.36965	-0.05251	-0.37867	0.685925	-0.54734	-0.71914	0.015702	-0.57109	-0.1343	-0.03894	0.0029
8	기온평균	0.030693	0.561223	0.526795	0.672971	0.638709	0.66452	1	0.262394	0.498672	-0.49971	-0.46824	0.442728	-0.48191	-0.43701	-0.09261	0.368845	-0.19645	0.025553	-0.06616
9	강수일강수	0.83892	0.915701	0.79772	0.838552	0.763809	0.711187	0.277806	1	0.598655	-0.49764	-0.42894	-0.03051	-0.36916	-0.26116	-0.2164	0.109185	-0.23177	0.039667	-0.08156
10	습도평균	0.943087	0.531928	0.944631	0.128123	0.108472	0.119317	0.029759	0.006767	1	-0.32767	-0.07957	-0.20455	-0.06416	0.091557	-0.01578	0.288395	-0.21362	0.133083	-0.15593
11	미세먼지	0.005848	0.025979	0.010511	0.799649	0.890308	0.830951	0.029368	0.030153	0.170854	1	0.904541	-0.03238	0.750047	0.518132	0.664114	-0.58084	0.124212	0.071045	-0.08054
12	초미세먼지	0.01879	0.065033	0.072143	0.110771	0.127964	0.109881	0.043186	0.066874	0.746086	1.06E-07	1	-0.30266	0.879888	0.746523	0.589999	-0.36505	0.141405	0.061967	-0.06499
13	오존	0.935035	0.600694	0.170636	0.00114	0.001034	0.001186	0.057663	0.901332	0.400899	0.895309	0.207857	1	-0.54631	-0.72602	0.201469	-0.43331	-0.14444	0.041568	-0.08783
14	이산화질소	0.037946	0.068118	0.133836	0.014434	0.019488	0.015287	0.036677	0.119844	0.794133	0.000217	6.86E-07	0.015519	1	0.853016	0.54537	-0.22356	0.146329	-0.06858	0.105262
15	일산화탄소	0.436883	0.643593	0.980082	0.00032	0.000722	0.000521	0.06136	0.280146	0.709313	0.023055	0.000241	0.000433	3.47E-06	1	0.204009	0.070345	0.180103	0.048473	-0.02869
16	아황산가스	0.000181	9.82E-06	2.00E-05	0.759137	0.882701	0.94913	0.706096	0.373543	0.948877	0.001929	0.007836	0.408183	0.015733	0.402172	1	-0.51096	0.086738	0.008354	-0.01256
17	month	0.000921	0.020336	0.000154	0.016331	0.007449	0.010652	0.120186	0.656357	0.231153	0.009112	0.124342	0.063843	0.357556	0.774752	0.025372	1	-0.03757	-0.00327	-0.00395
18	day	0.808674	0.745801	0.810449	0.736852	0.628627	0.583598	0.420191	0.339705	0.37988	0.612406	0.563642	0.555217	0.549998	0.460625	0.724027	0.878641	1	-0.026	0.006045
19	dayofweek	0.848106	0.927166	0.955977	0.939037	0.888235	0.874237	0.917297	0.871912	0.587033	0.772568	0.801031	0.865826	0.780257	0.84378	0.972923	0.989393	0.915843	1	-0.97523
20	weekday	0.910483	0.995254	0.988337	0.967963	0.998165	0.990598	0.787856	0.73995	0.52384	0.743082	0.791513	0.72068	0.668017	0.907186	0.959302	0.98721	0.980406	1.42E-12	1

P-value

- ✓ 반응변수(rentcnt)와 나머지 설명변수와의 상관관계를 알아보기 위해 상관관계 행렬을 작성
- ✓ 대각성분을 기준으로 위쪽은 상관계수, 아래쪽은 p-value를 표시
- ✓ 추가적으로 변수 선택을 할 필요성이 보임

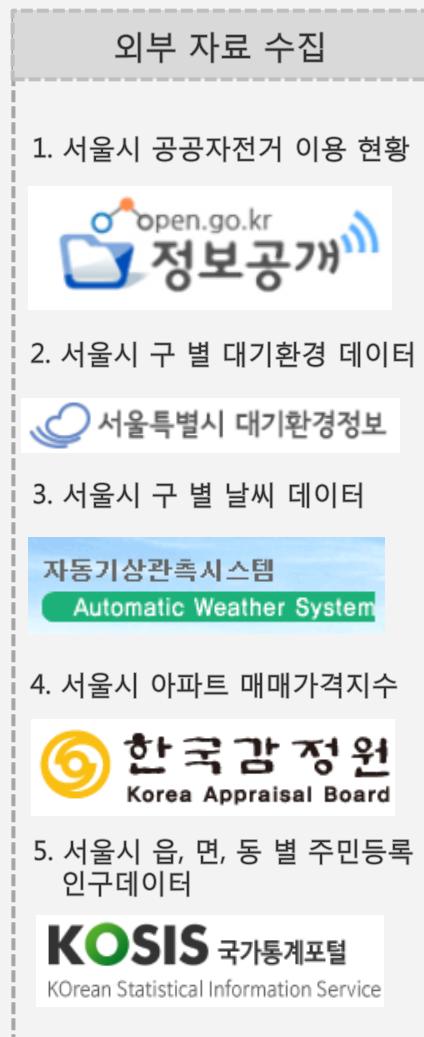


## 05

### 데이터 분석 보고서

- 분석 시나리오
- 변수 선택
- 모델링

## 분석 시나리오



### step 1



- 데이터 병합  
(서울시 11개 구에 대한  
공공자전거 대여 횟수)
- 데이터 전처리  
(결측값 보정)

### step 2

- EDA  
산점도, 히스토그램, 상관분석

- 변수 선택  
stepwise selection

### step 3

- 최종 예측 모델 선정

- 모델링
1. LASSO
  2. Ridge
  3. Elastic net
  4. Decision Tree
  5. SVR(Support Vector Regression)
  6. Random Forest
  7. Gradient Boosting



python



python

## 변수 선택

- ✓ 전체 설명변수(18개)로 Stepwise selection 방법으로 변수 선택 과정 진행 (AIC값이 가장 낮은 기준)
- ✓ 각 설명변수마다 단위가 다르기 때문에 척도화(scaling) 및 중심화(centering) 작업 수행

이름	n	최종 선택된 변수	AIC
광진구	730	인구, 아파트가격, 평균풍속, 최대풍속, 기온평균, 강수일강수, 습도평균, 미세먼지, 초미세먼지, 일산화탄소, 아황산가스, month, day <b>(13개 선택)</b>	-1015.81
동대문구	534	인구, 아파트가격, 평균풍속, 최대풍속, 기온평균, 강수일강수, 습도평균, 초미세먼지, 오존, 이산화질소, 아황산가스, weekday <b>(12개 선택)</b>	-645.45
마포구	730	인구, 아파트가격, 평균풍속, 최대풍속, 기온평균, 강수일강수, 습도평균, 미세먼지, 오존, weekday <b>(10개 선택)</b>	-853.47
서대문구	731	인구, 아파트가격, 평균풍속, 최대풍속, 기온평균, 강수일강수, 습도평균, 초미세먼지, 오존, 아황산가스, day <b>(11개 선택)</b>	-920.28
성동구	730	인구, 아파트가격, 평균풍속, 최대풍속, 순간최고풍속, 기온평균, 강수일강수, 습도평균, 초미세먼지, 아황산가스, month, day, weekday <b>(13개 선택)</b>	-923.06
양천구	531	인구, 아파트가격, 평균풍속, 최대풍속, 기온평균, 강수일강수, 습도평균, 미세먼지, 오존, 아황산가스, month, day, weekday <b>(13개 선택)</b>	-535.6
영등포구	729	인구, 아파트가격, 평균풍속, 순간최고풍속, 기온평균, 강수일강수, 습도평균, 초미세먼지, 오존, 이산화질소, 아황산가스, month, day, weekday <b>(14개 선택)</b>	-1080.28
용산구	531	인구, 아파트가격, 평균풍속, 최대풍속, 기온평균, 강수일강수, 습도평균, 초미세먼지, 오존, 일산화탄소, 아황산가스, month, weekday <b>(13개 선택)</b>	-638.53
은평구	461	아파트가격, 평균풍속, 최대풍속, 기온평균, 강수일강수, 습도평균, 초미세먼지, 오존, 이산화질소, 일산화탄소, 아황산가스, day <b>(12개 선택)</b>	-733.49
종로구	731	인구, 아파트가격, 평균풍속, 최대풍속, 기온평균, 강수일강수, 습도평균, 초미세먼지, 오존, 일산화탄소, day, weekday <b>(12개 선택)</b>	-905.63
중구	730	인구, 아파트가격, 평균풍속, 최대풍속, 기온평균, 강수일강수, 습도평균, 초미세먼지, 오존, 이산화질소, 일산화탄소, 아황산가스, day, weekday <b>(14개 선택)</b>	-1101.2



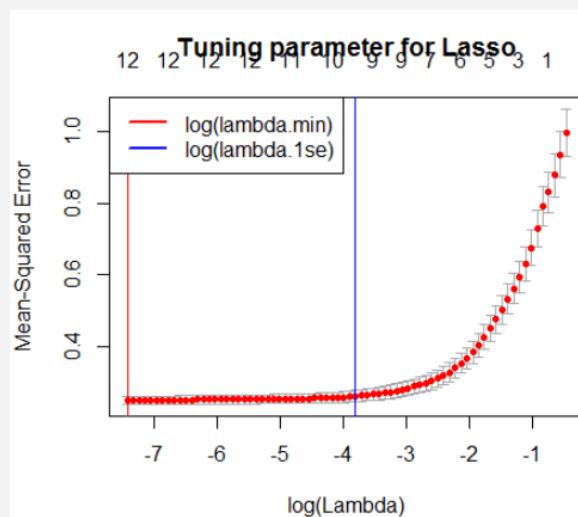
## 모델링

- ✓ 분석 Tool : R studio
- ✓ 날짜, area 변수 제외하고 나머지 변수 모두 사용(19개)
- ✓ 교차검증 (cross validation) : K = 10 (단, Random Forest 제외, why? Out-of-Bag 과정 교차검증과 유사)

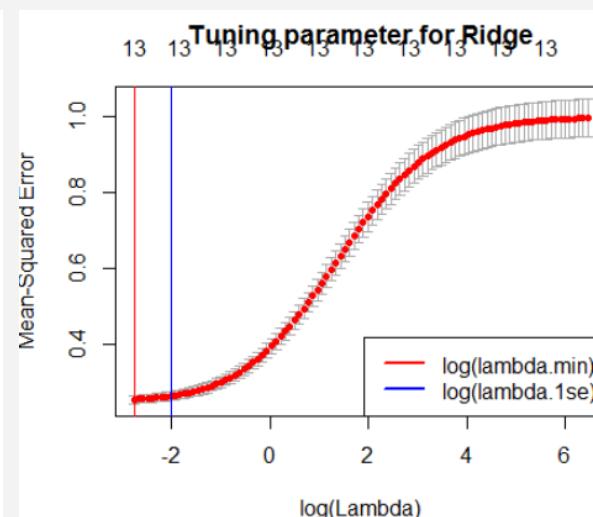
Model	package	Tuning parameter
Lasso, Ridge, Elastic Net	glmnet	alpha = 1 (Lasso) 0 (Ridge) 0.5 (Elastic Net) lambda = lambda.min
Random Forest	randomForest	ntree = 5000 sample size = default maxnodes = default
Support Vector Regressor	e1071	kernel = radial
Decision Tree	tree	prune tree (best = ?)
Gradient Boosting	caret, gbm	distribution = gaussian n.trees = ? shrinkage = ? interaction.depth = ? n.minobsinnode = ?

## Lasso, Ridge, Elastic Net

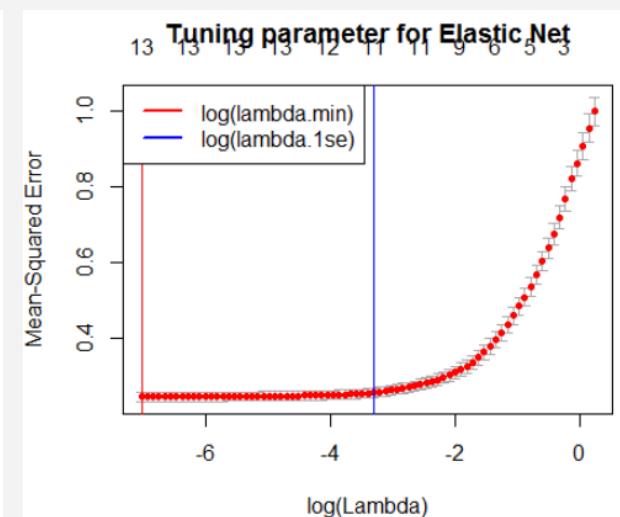
1. Lasso : 통계 모델의 예측 정확도와 해석 가능성을 높이기 위해 변수 선택, 정규화를 수행하는 회귀 모델
  2. Ridge : 기존 선형 회귀의 과대 적합을 일부 해소시킬 수 있는 모델, 다중공선성 제거하기 위하여 쓰임
  3. Elastic Net : Lasso, Ridge의 선형 결합을 통해 패널티를 해소하는 정규화 된 회귀 모델
- ✓ 최적화 된 lambda(min lambda) 값을 찾기 위해 다음과 같은 작업 수행



Lasso



Ridge



Elastic Net

## Lasso, Ridge, Elastic Net

- ✓ tuning parameter : `lambda = lambda.min`으로 추정된 beta 값은 아래와 같음
- ✓ Lasso, Elastic Net은 변수 선택 기능 존재

```
14 x 1 sparse Matrix of class "dgCMatrix"
  s0
(Intercept) -2.238223e-15
인구          5.416909e-01
아파트가격   -1.511092e-01
평균풀속    -1.303030e-01
최대풀속     1.156435e-01
기온평균     4.979808e-01
강수일간수  -1.868920e-01
습도평균     -1.648870e-01
미세먼지      4.399542e-02
초미세먼지   -1.211187e-01
일산화탄소   1.248519e-01
아황산가스   -2.691563e-01
month        2.796963e-01
day          3.505089e-02
```

Lasso - `lambda.min`

```
14 x 1 sparse Matrix of class "dgCMatrix"
  s0
(Intercept) -2.044764e-15
인구          3.678315e-01
아파트가격   5.566745e-02
평균풀속    -1.188943e-01
최대풀속     9.495931e-02
기온평균     4.799886e-01
강수일간수  -1.769029e-01
습도평균     -1.537867e-01
미세먼지      3.240507e-02
초미세먼지   -9.959793e-02
일산화탄소   9.003295e-02
아황산가스   -2.505128e-01
month        1.855527e-01
day          3.544621e-02
```

Ridge - `lambda.min`

```
14 x 1 sparse Matrix of class "dgCMatrix"
  s0
(Intercept) -2.237791e-15
인구          5.418893e-01
아파트가격   -1.515657e-01
평균풀속    -1.307102e-01
최대풀속     1.160210e-01
기온평균     4.978200e-01
강수일간수  -1.869759e-01
습도평균     -1.648154e-01
미세먼지      4.438439e-02
초미세먼지   -1.215943e-01
일산화탄소   1.249718e-01
아황산가스   -2.692331e-01
month        2.799989e-01
day          3.518751e-02
```

Elastic Net - `lambda.min`

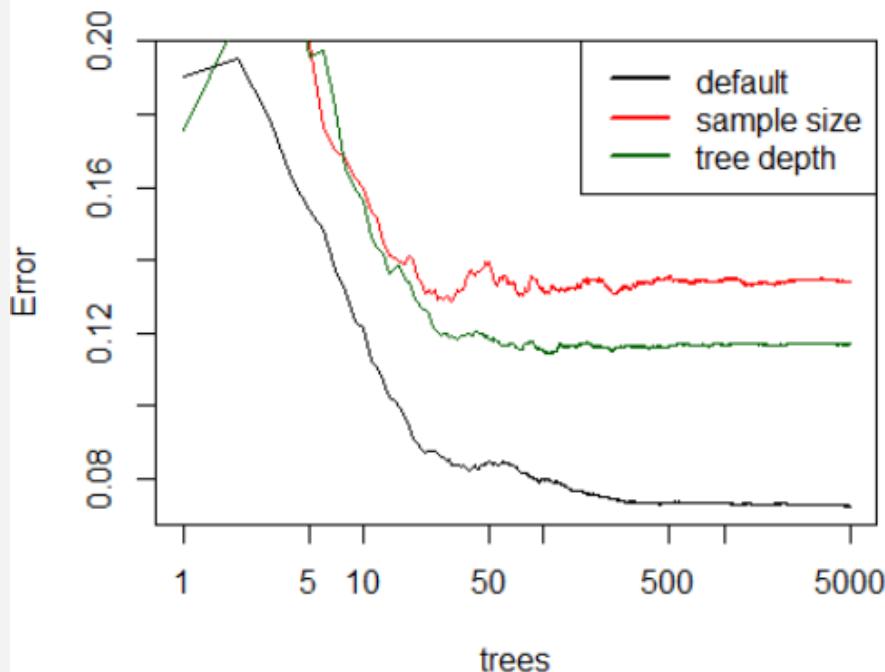
- ✓ . : 추정된 beta 값이 0이라는 의미 (변수의 유의성 판단은 신뢰구간을 통해 가설검정 과정을 거쳐야 함)

## Random Forest

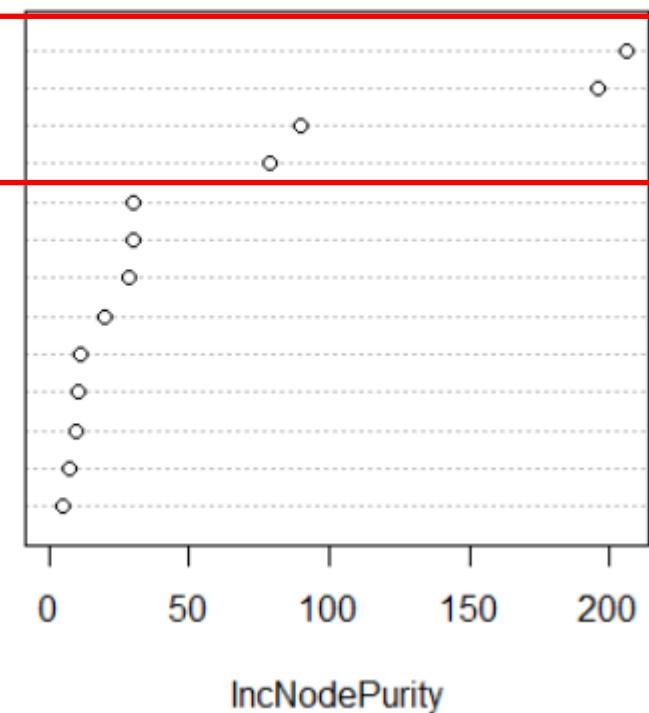
- ✓ 조금씩 다른 여러 결정 트리의 묶음을 통해 평균적 예측을 수행하는 기법
- ✓ 트리 모델의 예측 성능이 유지되면서 과대적합을 줄일 수 있는 모델
- ✓ Tuning parameter : ntree = 5000, sample size = default, maxnodes = default

**fit.rf**

Tuning parameter for Random Forest



아파트가격  
인구  
기온평균  
month  
아황산가스  
습도평균  
강수일강수  
평균풍속  
미세먼지  
초미세먼지  
day  
최대풍속  
일산화탄소



## SVR, Decision Tree

## 1. SVR (Support Vector Regression)

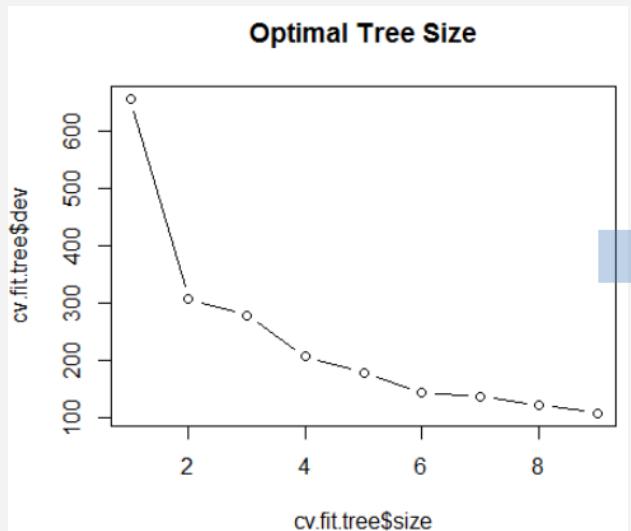
- ✓ kNN과 선형회귀 모델링 기법이 적용된 모델
- ✓ 비선형 예측을 위해서 Kernel 함수 사용
- ✓ 저차원, 고차원 데이터 모두 잘 동작
- ✓ 다음과 같이 Tuning Parameter 과정을 시행



Parameters	
SVM-Type	eps-regression
SVM-Kernel	Radial
cost	1
gamma	0.07692308
epsilon	0.1
Number of Support Vectors	475

## 2. Decision Tree

- ✓ 어떤 항목에 대한 관측값과 목표값을 연결하는 예측 모델
- ✓ 모델의 시각화가 쉽고, 가독성이 높음
- ✓ 최적의 트리 개수 = 9개
- ✓ Prune Tree 통해 결과 도출



```

node), split, n, deviance, yval
* denotes terminal node

1) root 657 654.900 -0.007402
  2) 아파트가격 < -0.0764003 421 111.300 -0.554400
    4) month < 0.283948 279 23.340 -0.806200
      8) 인구 < 0.741473 250 9.433 -0.878000 *
      9) 인구 > 0.741473 29 1.513 -0.187400 *
    5) month > 0.283948 142 35.530 -0.059500
      10) 기온평균 < -0.547173 46 1.708 -0.540900 *
      11) 기온평균 > -0.547173 96 18.060 0.171200 *
  3) 아파트가격 > -0.0764003 236 192.900 0.968300
    6) 기온평균 < -0.6078 38 6.883 -0.200200 *
    7) 기온평균 > -0.6078 198 124.200 1.193000
    14) 강수일강수 < -0.171313 150 58.940 1.452000
      28) month < 0.573448 93 19.950 1.133000 *
      29) month > 0.573448 57 14.150 1.972000 *
    15) 강수일강수 > -0.171313 48 23.600 0.381900
      30) 강수일강수 < 1.79489 39 12.920 0.588700 *
      31) 강수일강수 > 1.79489 9 1.784 -0.514100 *

```

## Gradient Boosting

- ✓ 여러 개의 결정 트리를 묶어 강력한 모델을 만드는 앙상블 기법
- ✓ 이전 트리의 오차를 보완하는 방식을 통해 순차적으로 트리 생성
- ✓ distribution = gaussian, n.trees = 150, interaction.depth = 3, shrinkage = 0.1, n.minobsinnode = 10

### Stochastic Gradient Boosting

730 samples  
13 predictor

No pre-processing  
Resampling: Cross-Validated (25 fold)  
Summary of sample sizes: 699, 699, 702, 701, 700, ...  
Resampling results across tuning parameters:

	interaction.depth	n.trees	RMSE	R squared	MAE
1	50	0.4364688	0.8375803	0.3294610	
1	100	0.3655155	0.8750853	0.2765095	
1	150	0.3452649	0.8843039	0.2575986	
2	50	0.3232031	0.9006129	0.2381190	
2	100	0.2884139	0.9167536	0.2079879	
2	150	0.2779844	0.9216539	0.2001505	
3	50	0.2921554	0.9151216	0.2119544	
3	100	0.2735771	0.9238716	0.1938180	
3	150	0.2699855	0.9258034	0.1906782	

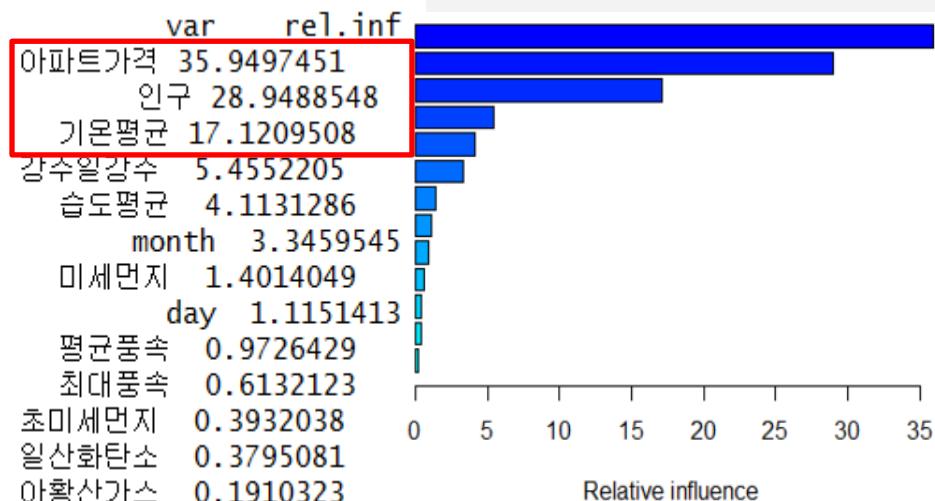
Tuning parameter 'shrinkage' was held constant at a value of 0.1

Tuning parameter 'n.minobsinnode' was held constant at a value of 10

RMSE was used to select the optimal model using the smallest value.

The final values used for the model were n.trees = 150, interaction.depth = 3, shrinkage = 0.1 and n.minobsinnode = 10.

### Tuning parameter





06

결론

## 서울시 공공자전거 따릉이

최종 예측 모형 선정

1일권 : 1000원

기본 대여시간 : 60분

\* 결제 후 60분 내로 반납하고

다시 빌리는 방식으로

추가요금 없이 이용 가능

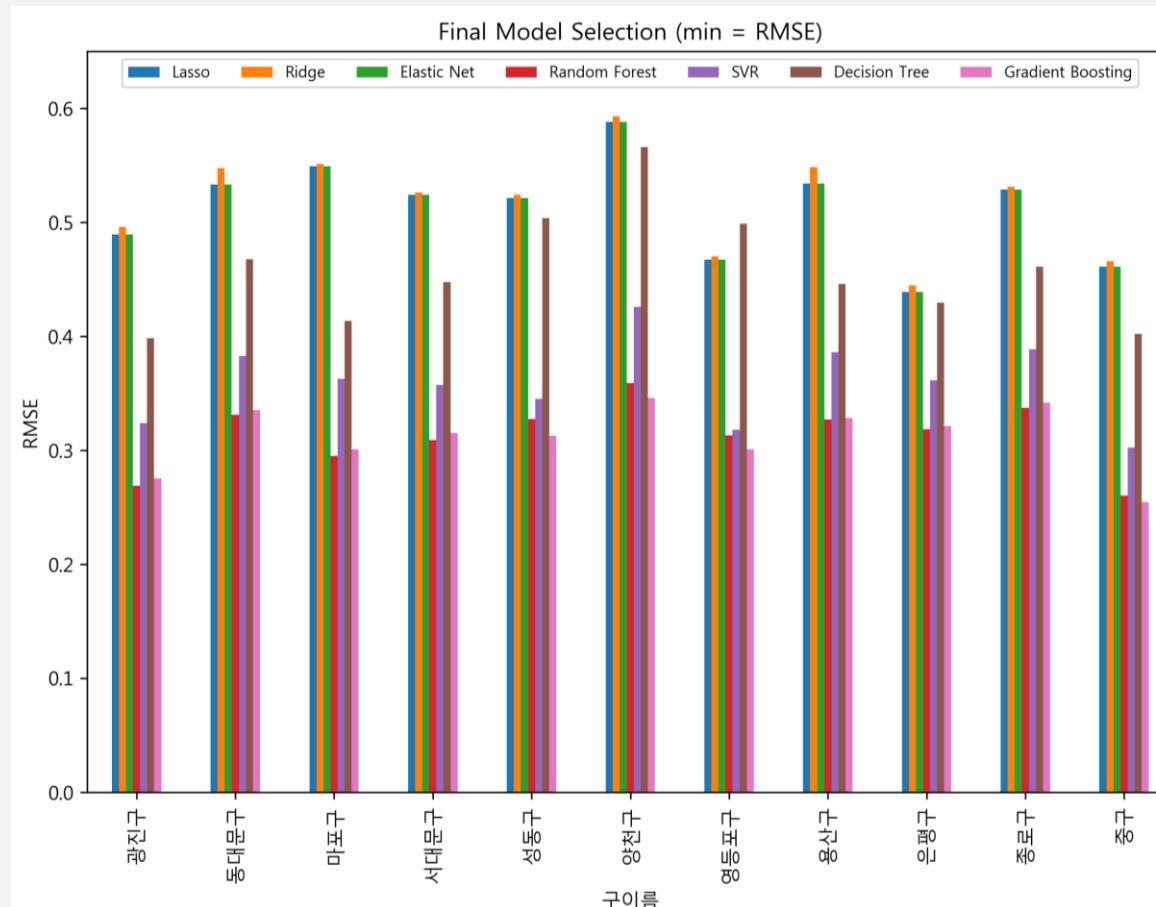
• 서비스 활용방안

• 서비스 기대효과

• 한계점 및 제언

## 최종 예측 모형 선정

- ✓ 최종 예측 모형 선정하기 위해 7개 모형을 비교
- ✓ 모형을 train set을 토대로 예측하여 test set과 비교하여 구한 RMSE가 낮은 값을 예측 모형으로 선정



## 최종 예측 모형 선정

	Lasso	Ridge	Elastic Net	Random Forest	SVR	Decision Tree	Gradient Boosting
광진구	0.4892641	0.4958865	0.4892607	0.268801	0.3236706	0.3983318	0.2751513
동대문구	0.5333251	0.5474841	0.5333262	0.3310775	0.3825591	0.4673729	0.3352258
마포구	0.5490359	0.5510542	0.5490353	0.294888	0.3626913	0.4135126	0.3006528
서대문구	0.5242572	0.5262021	0.5242857	0.308823	0.3571232	0.4474356	0.3149175
성동구	0.5213496	0.5239779	0.5213511	0.3271988	0.3451267	0.5038401	0.3124147
양천구	0.5882213	0.5928757	0.5882219	0.3589086	0.4259162	0.5658102	0.3458456
영등포구	0.4669988	0.4700245	0.4669941	0.3129096	0.3180928	0.4986033	0.3008707
용산구	0.5338995	0.5484463	0.5339435	0.3269265	0.385854	0.4458528	0.3283593
은평구	0.438846	0.4447816	0.4389124	0.3185307	0.361573	0.4295204	0.3213934
종로구	0.5288036	0.531215	0.5288347	0.3373635	0.388332	0.4608266	0.3416726
중구	0.4608508	0.4658427	0.4608403	0.2602673	0.3024133	0.4020753	0.2542885

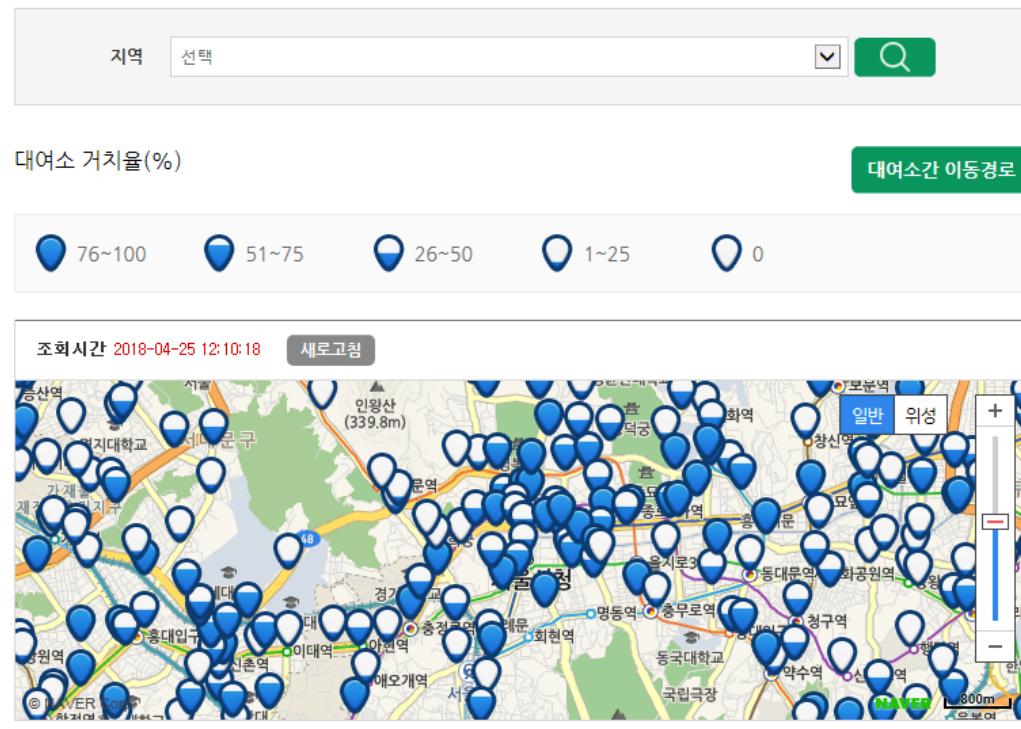
- ✓ 대체적으로 Random Forest가 성능이 가장 우수한 것으로 보임
- ✓ 경우에 따라 Gradient Boosting로 대체해도 무방할 것으로 예상

## 서비스 활용방안

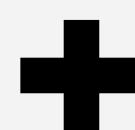
### 대여소 실시간 현황

대여소별 현황을 실시간으로 확인합니다.

[▶ 대여소 조회 > 대여소 실시간 현황](#)



### 따릉이 수요 예측 시스템



## 서비스 기대효과

종 류	대 상	서비스 기대효과
공공적 기대효과	시민	<ul style="list-style-type: none"> <li>‘공공자전거 수요 예측 시스템’으로 효율적으로 공공자전거 이용</li> </ul>
정책적 기대효과	서울시	<ul style="list-style-type: none"> <li>‘서울시 공공자전거 수요 예측 시스템’을 따릉이 홈페이지에 서비스하여 시민들의 원활한 공공자전거 이용에 유용한 정보를 제공</li> <li>사전에 공공자전거 수요가 많은 구와 적은 구를 파악하여 공공자전거를 적재적소에 배치함으로써 효율적인 예산 활용</li> </ul>
	정부	<ul style="list-style-type: none"> <li>교통 체증, 대기 오염 문제에 대해 해결하기 위한 인사이트 제공</li> <li>공공자전거 수요가 적은 지역을 파악하여 문제점 진단에 도움</li> </ul>

## 한계점 및 제언

유형	한계점
외부요인	<ul style="list-style-type: none"><li>서울시 자전거 사고 통계 데이터를 확인하지 못함 → 자전거 사고 발생 건수는 공공자전거 수요 예측에 있어서 중요한 변수로 사료되나 데이터 확인 불가 해당 데이터 확보 시 수요 예측 알고리즘 성능 강화에 도움이 될 것으로 예상</li><li>변수를 확보하는 데 있어서 지나치게 거시적인 요인(날씨, 대기환경)에만 치중 → 설문지를 통하여 '시민들의 공공자전거에 대한 인식' 조사를 통하여 보다 많은 파생변수 확보 즉, 미시적 요인(시민들의 인식)에 대한 접근성도 필요</li></ul>
데이터 확보	<ul style="list-style-type: none"><li>서울시 공공자전거 수요 데이터 양이 충분하지 않음 → 2015년 9월부터 시행 되었기 때문에 충분한 데이터 확보 무리(시간이 지남에 따라 데이터 축적이 해결)</li><li>서울시 각 구별 정확한 날씨, 대기환경 등 기상에 관한 데이터 확인 못함 → 더 정확한 기상정보를 이용하기 위해서는 각 구에서의 실측조사가 필요함</li><li>knn imputation을 이용한 추정치이기 때문에 오차가능성이 존재 → kriging imputation, IDW imputation 등 공간 보간법을 통해 추정이 가능하지만 이 역시 오차가능성이 존재하므로 근본적인 해결책은 아님</li></ul>



Thank you  
Q & A