감기 진료 건수 분석 및 예측

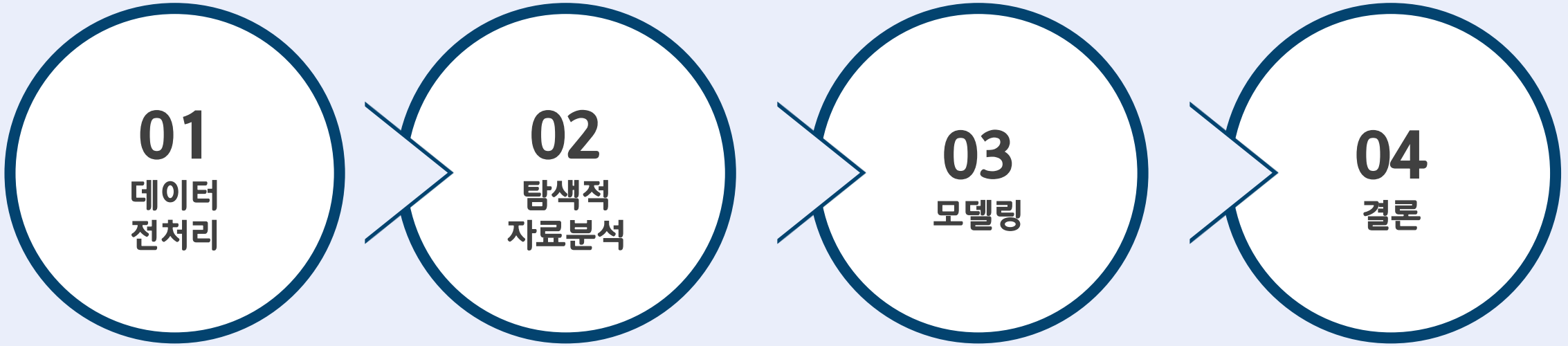
빅데이터 활용을 위한 DB관리 전문가 과정 (C반)

김형순

정승원

황유림

목차



- 데이터 설명
- 데이터 병합
- 결측값 처리
- 불필요한 정보 제거

- 데이터 시각화
- 상관도 분석

- Regression (회귀)
- Classification (분류)
- 예측

- 결론
- 참고 문헌

데이터 전처리

데이터 설명

target!

감기 진료 건수

출처 : 공공 데이터 포털

생활 인구

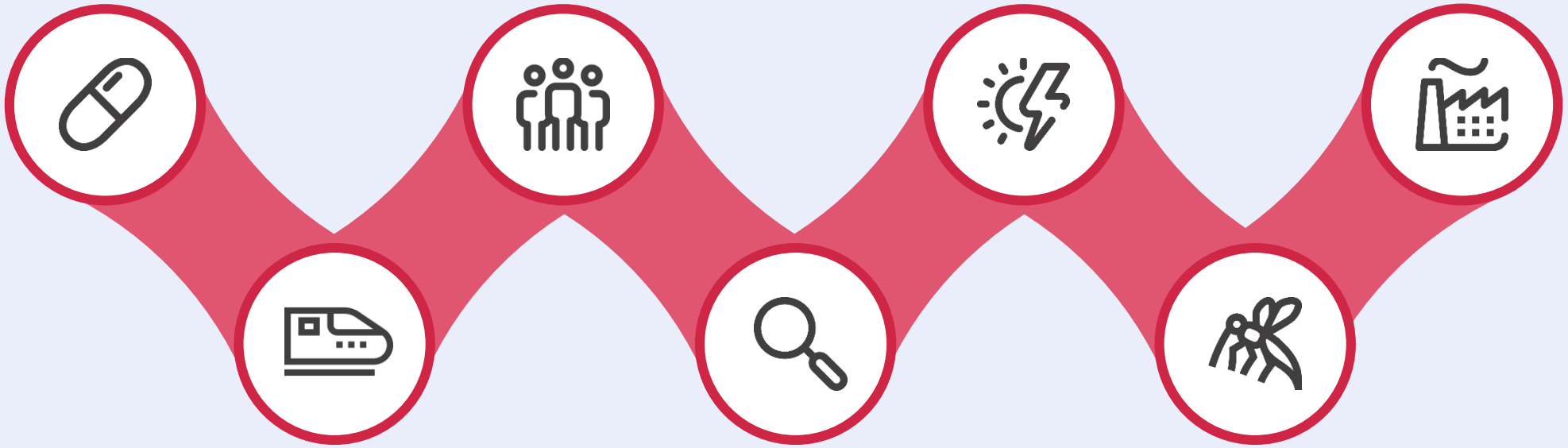
출처 : 서울 열린 데이터 광장

기상 관측

출처 : 기상 자료 개방 포털

대기 오염

출처 : 서울 열린 데이터 광장



지하철 승하자 인구

출처 : 서울 열린 데이터 광장

네이버 '감기' 검색량

출처 : 네이버 데이터랩

모기 지수

출처 : 서울 열린 데이터 광장

target!

감기 진료 건수

(날짜, 구 코드)

기상 관측
- 평균 기온 (°C)
- 강수량 (mm)
- 평균 풍속 (m/s)
- 일교차
- 체감온도

생활 인구
- 내국인 총 생활 인구 수

지하철 승하차 인구
- 총 승객 수

네이버 '감기' 검색량
- 검색량

모기 지수
- 모기 지수

대기 오염
- 미세먼지
- 초미세먼지

일교차 = 최고 기온 (°C) - 최고 기온 (°C)

체감온도 = 13.12 + 0.6215 * 평균기온(°C)" - 11.37 * 평균 풍속(m/s) + 0.3965 * 평균 풍속(m/s) * 평균기온(°C)

결측값 처리

결측값 많음

- 평균 기온 (°C)
- 강수량 (mm)
- 평균 풍속 (m/s)

같은 날, 서울 전체의
평균 값으로 채움

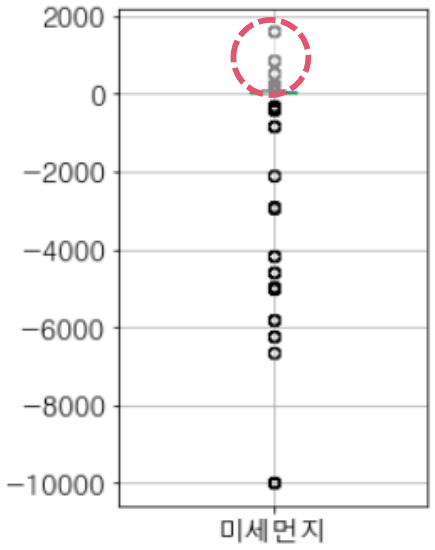
결측값 적음

- 총 생활 인구 수
- 모기 지수

전날, 해당 구의
값으로 채움

이상치 처리

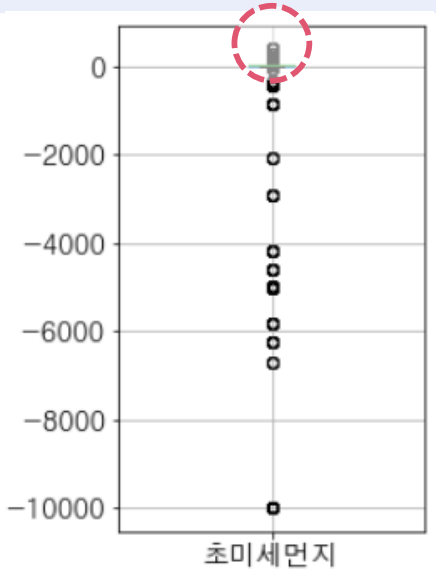
미세먼지



측정 가능한 범위인

0 이상 1000 이하의 값 만 추출

초미세먼지



측정 가능한 범위인

0 이상 1000 이하의 값 만 추출

불필요한 정보 제거

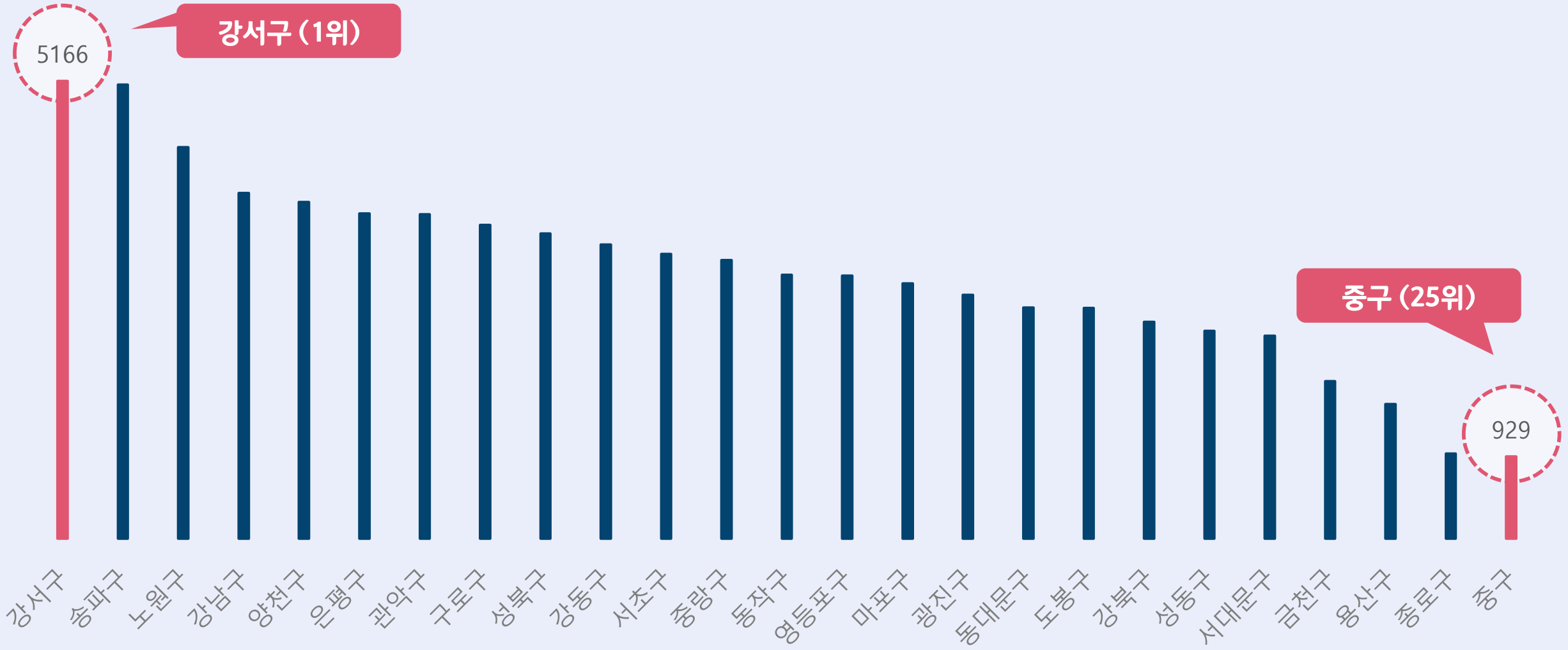
일요일

SUN	MON	TUE	WED	THU	FRI	SAT
1 신정	2	3	4	5	6	7
8	9	10	11	12	13	14
15	16	17	18	19	20	21
22	23	24	25	26	27	28 설날
29	30 대체 휴일	31				

공휴일

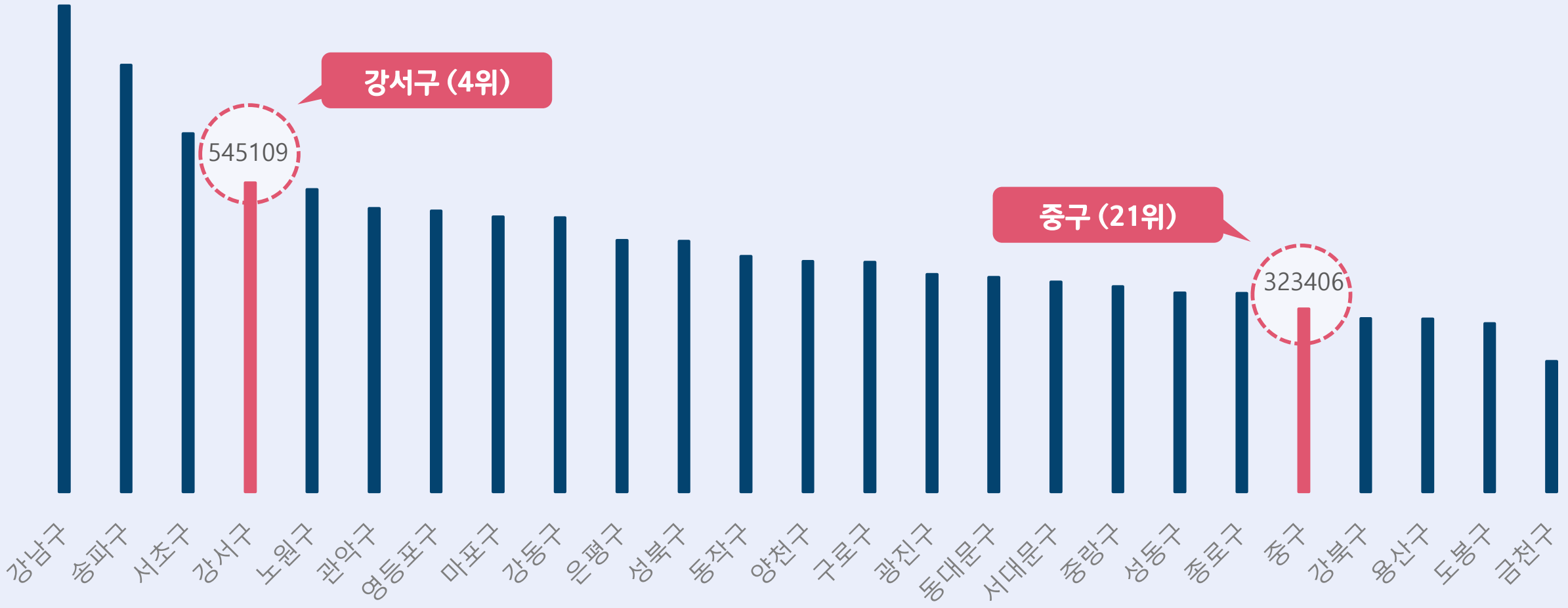
탐색적 자료 분석

시각화



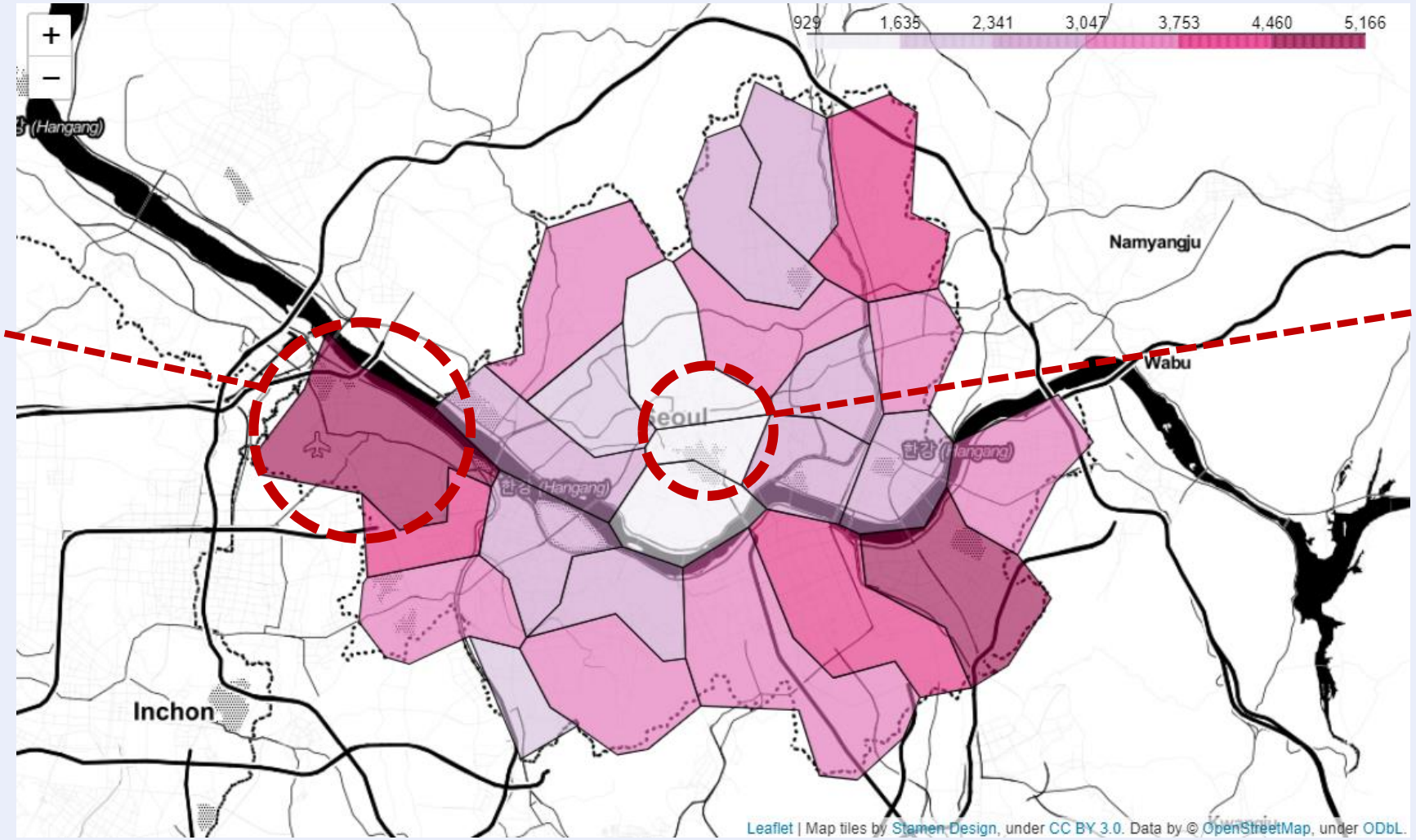
구 별 하루 평균 감기 진료 건수

시각화



구 별 하루 평균 생활 인구

시각화

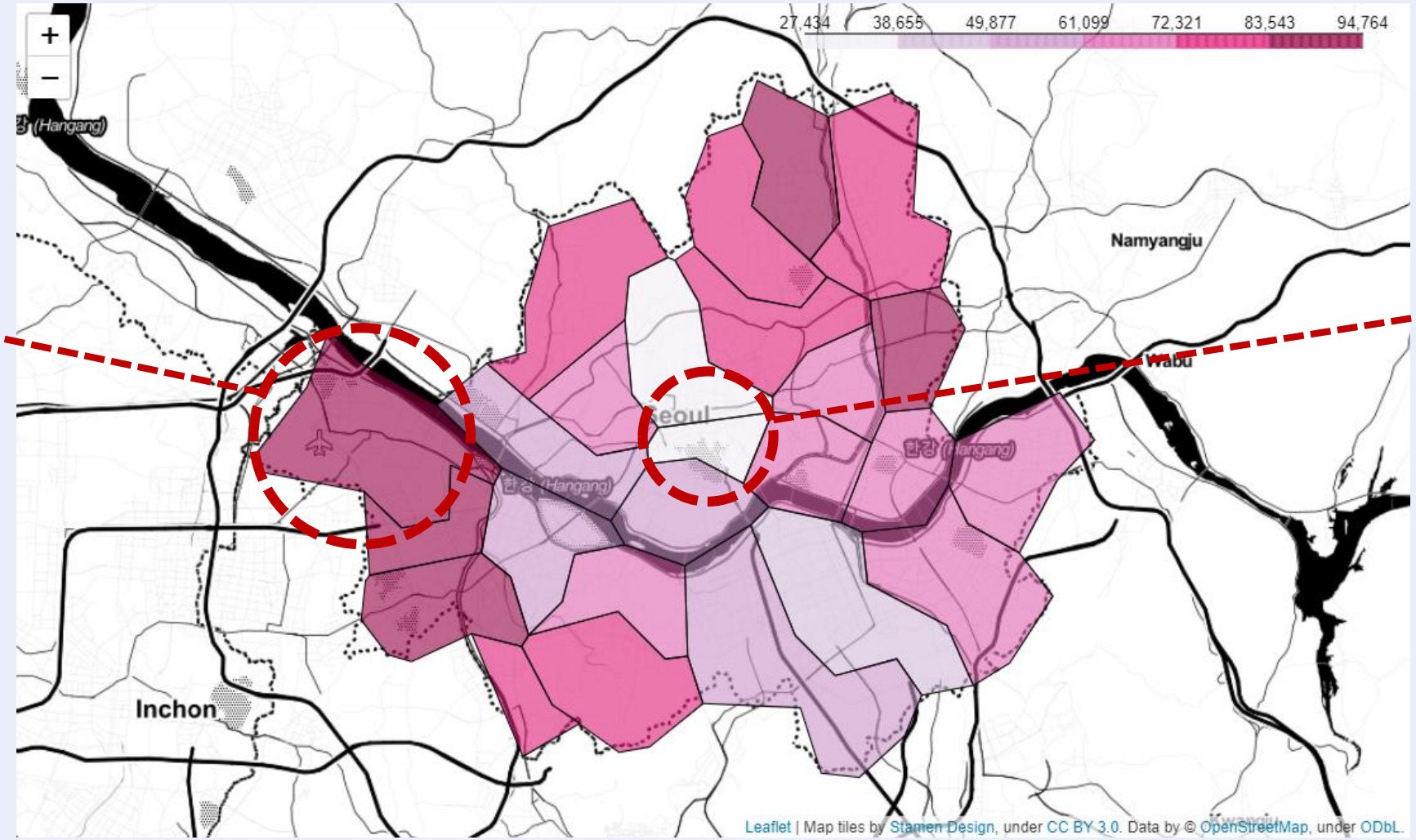


강서구

중구

구 별 하루 평균 감기 진료 건수

시각화

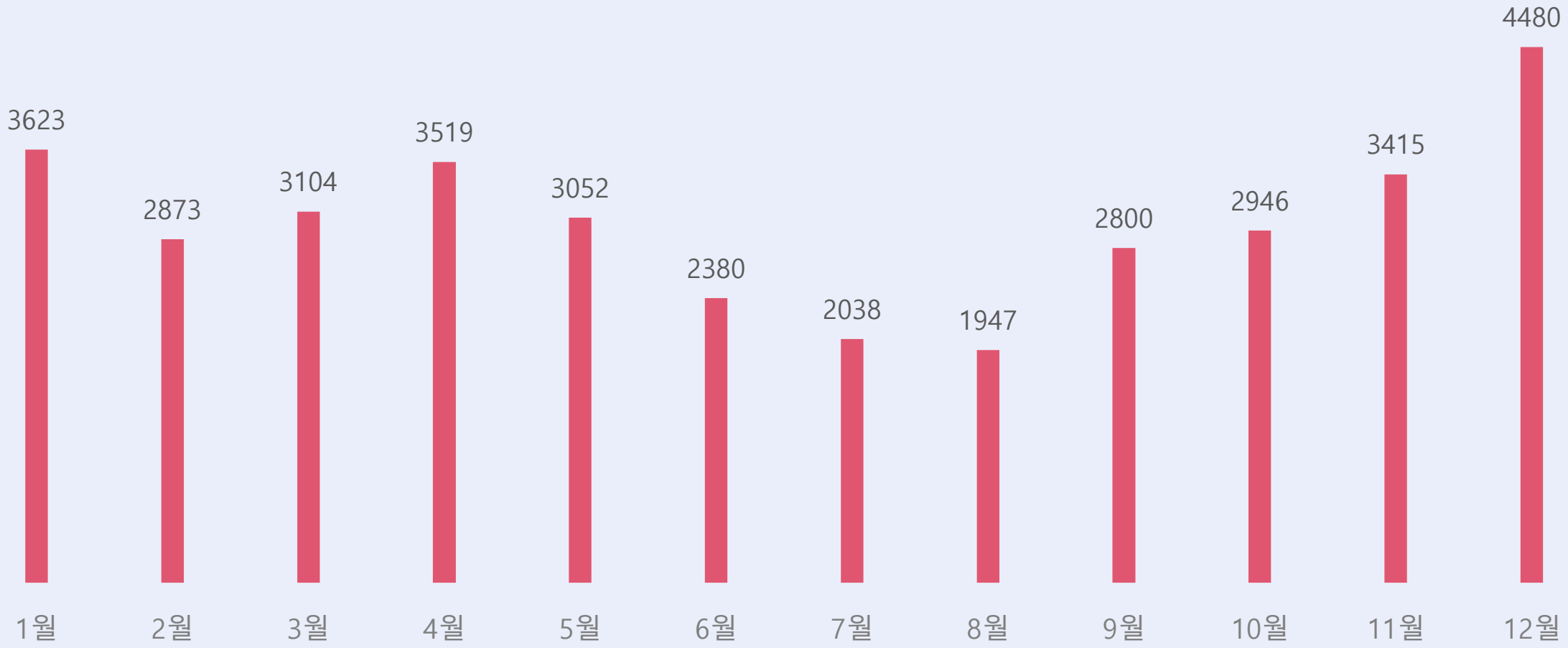


강서구

중구

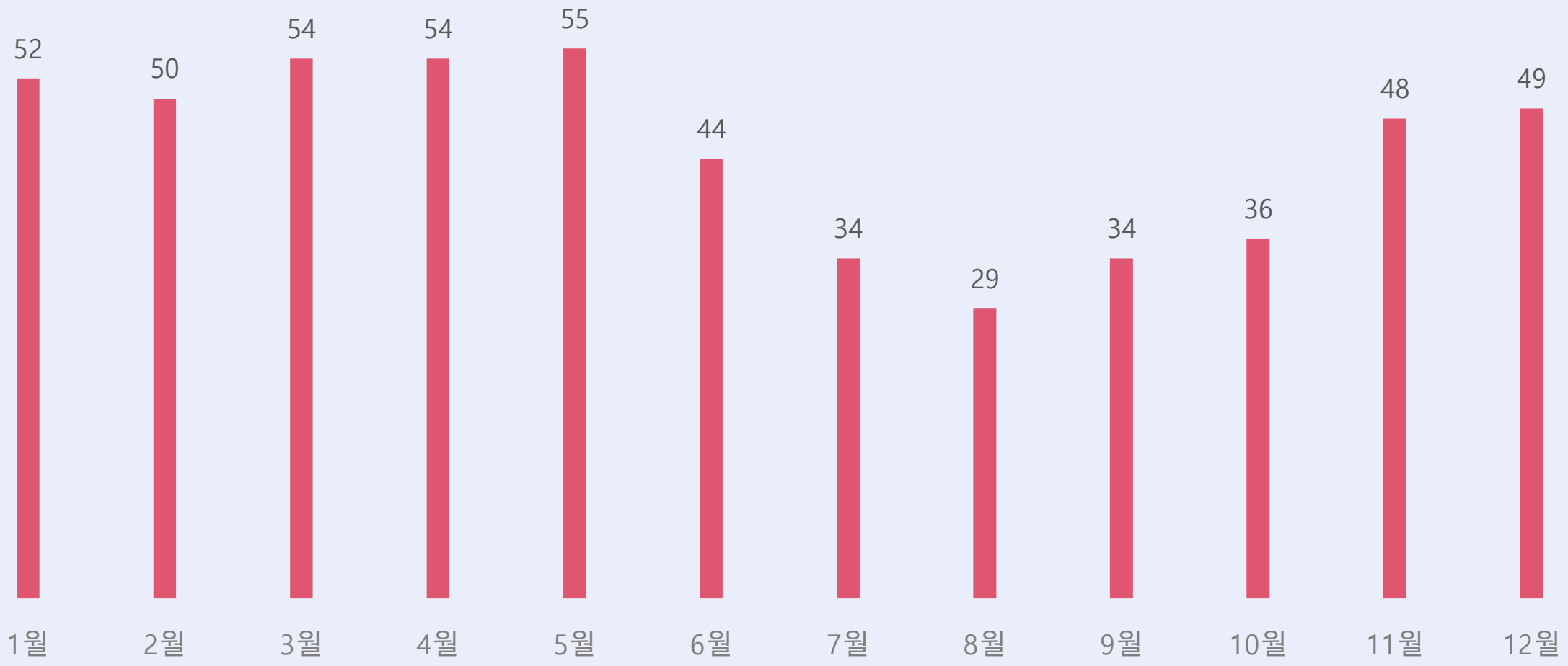
구 별 생활 인구 대비 하루 평균 감기 진료 건수

시각화



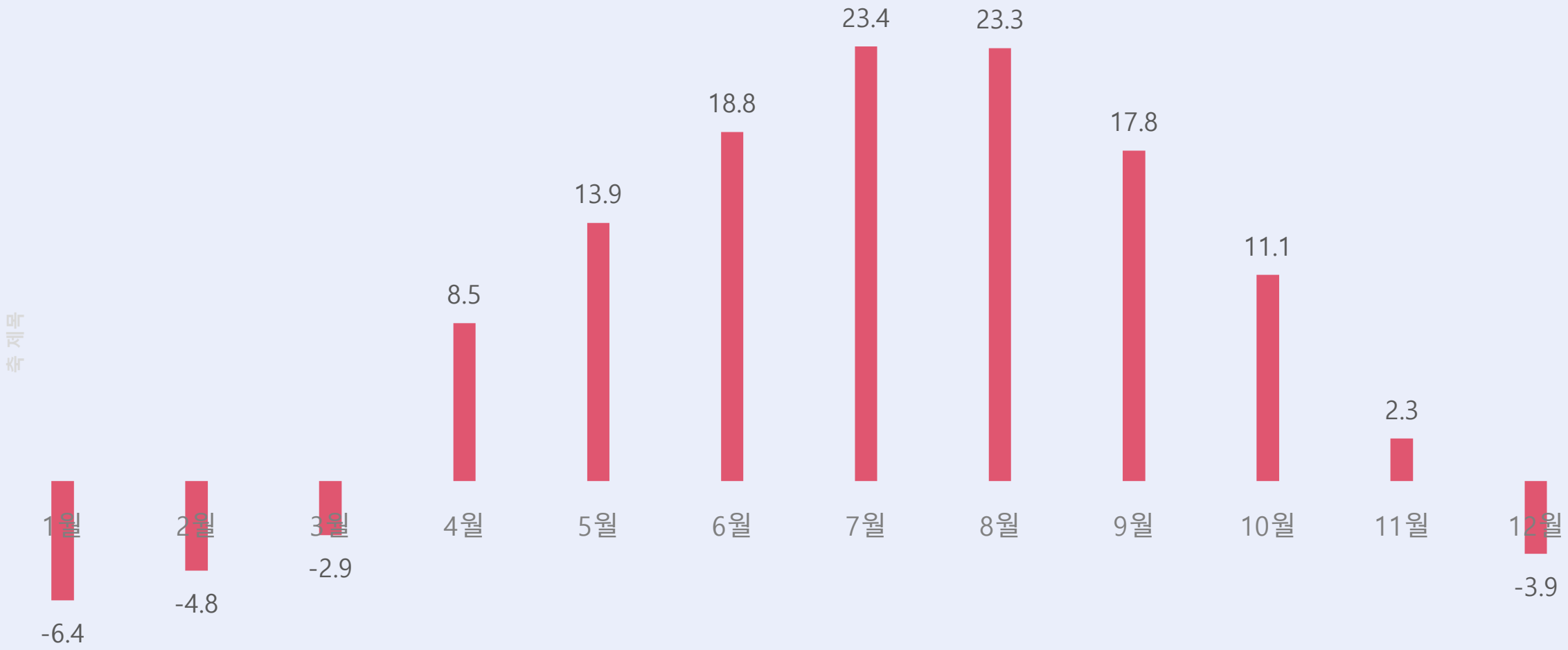
월 별 하루 평균 감기 진료 건수

시각화



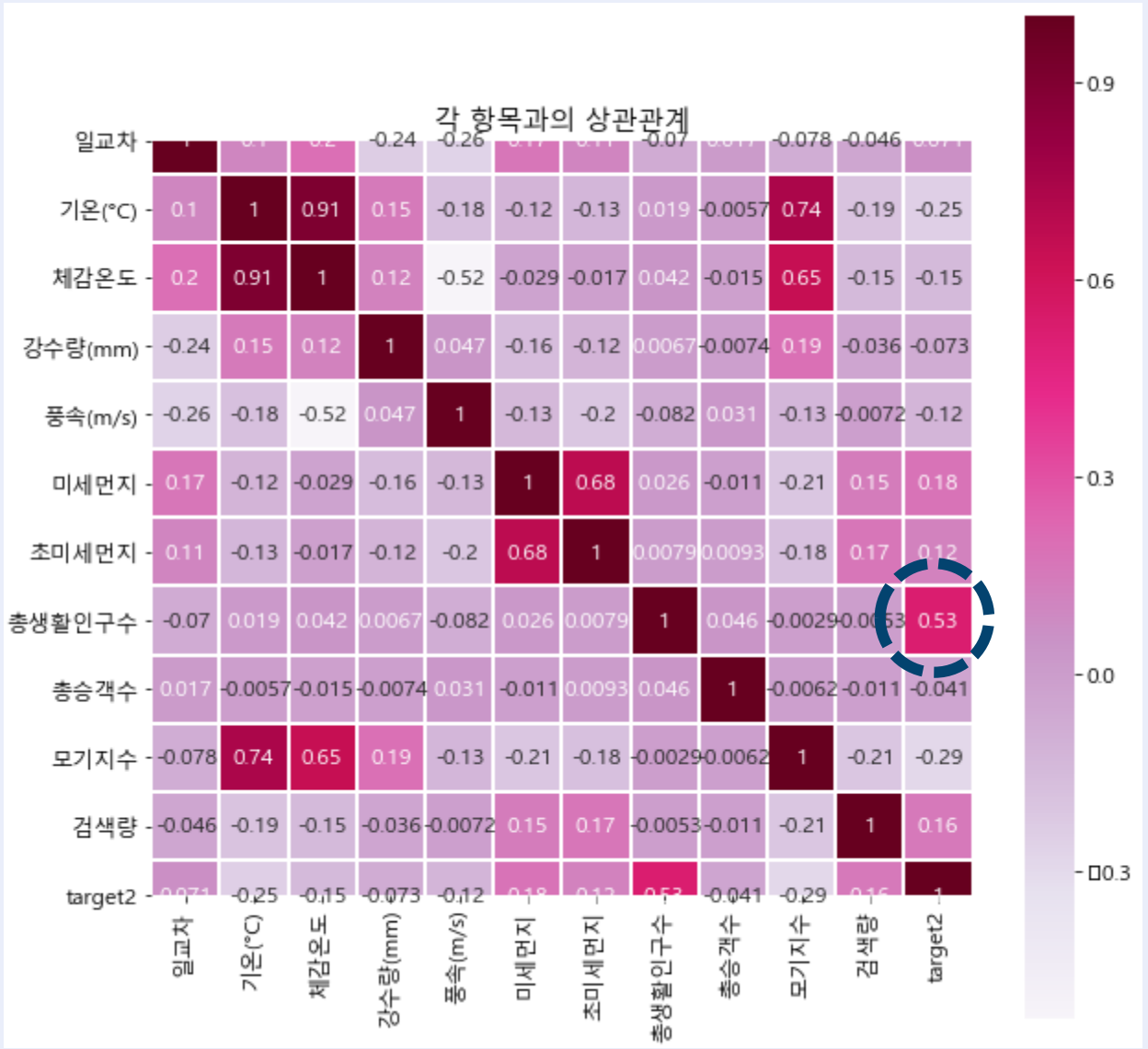
월 별 하루 평균 미세먼지 농도

시각화



월 별 하루 평균 최저 기온 (°C)

상관관계 시각화


각 항목과의 상관관계
일교차
기온(°C)
체감온도
강수량(mm)
풍속(m/s)
미세먼지
초미세먼지
총생활인구수
총승객수
모기지수
검색량
target2
0.9
0.6
0.3
0.0
0.3

모델링

Regression Model

일교차	기온(°C)	체감온도	강수량(mm)	풍속(m/s)	미세먼지	초미세먼지	총생활인구수	총승객수	모기지수	검색량	target
7.7	-0.3	2.593495	0.0	0.9	56.000000	38.000000	384303.8087	23633	5.5	7.90886	3848
19.4	8.7	12.190690	0.0	0.8	82.000000	51.000000	439648.8531	9753	12.1	6.65465	6435
7.2	-3.7	-45.662570	0.0	4.4	37.000000	21.000000	253202.8034	3853	5.5	4.63657	825
8.5	26.3	28.240785	0.0	1.3	51.000000	29.000000	501520.6836	53265	1000.0	3.07548	1798
12.6	12.6	8.202700	0.0	2.0	63.833333	44.833333	872269.2161	33213	5.5	3.97611	3032

학습 데이터: 101481건

일교차	기온(°C)	체감온도	강수량(mm)	풍속(m/s)	미세먼지	초미세먼지	총생활인구수	총승객수	모기지수	검색량	target
13.6	18.9	19.827355	0.0	1.3	110.000000	47.000000	386955.7438	8937	116.8	4.52650	4839
5.8	-9.3	-34.820810	0.0	2.8	32.583333	20.041667	645367.4155	105757	5.5	5.37376	2676
3.7	26.6	27.758770	15.5	2.3	19.000000	11.000000	348832.9839	13336	877.7	2.58180	1575
9.9	21.2	24.517280	16.5	0.6	31.000000	18.000000	414608.2604	24579	285.8	4.35638	3133
10.6	13.7	13.915215	0.0	1.3	42.000000	20.000000	630393.9546	14402	96.7	4.80669	2740

테스트 데이터: 25371건

총 126852건 데이터 중 101481건의 학습 데이터를 이용하여 훈련시켰으며
테스트 데이터 25371건으로 모델의 정확성을 평가하였습니다.

정규화

0	1	2	3	4	5	6	7	8	9	10
0.287081	0.373796	0.735853	0.0	0.121212	0.065392	0.093837	0.257393	0.086423	0.005202	0.057039
0.846890	0.547206	0.812548	0.0	0.106061	0.095752	0.125939	0.329562	0.035663	0.011804	0.044197
0.263158	0.308285	0.350220	0.0	0.651515	0.043205	0.051857	0.086440	0.014087	0.005202	0.023533
0.325359	0.886320	0.940811	0.0	0.181818	0.059553	0.071612	0.410241	0.194788	1.000000	0.007548
0.521531	0.622351	0.780678	0.0	0.287879	0.074539	0.110711	0.893688	0.121457	0.005202	0.016770

학습 데이터

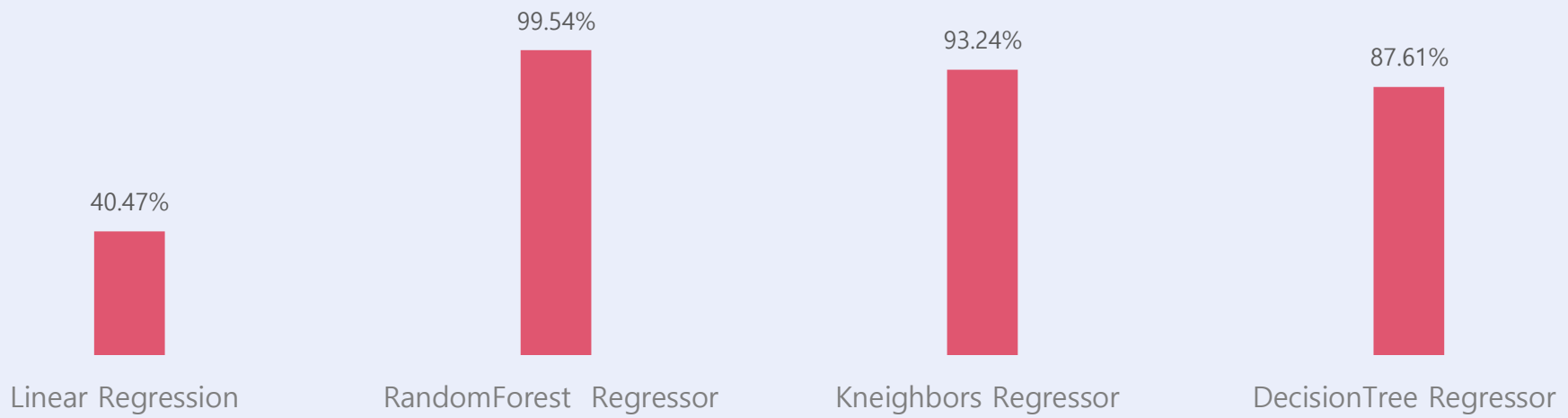
0	1	2	3	4	5	6	7	8	9	10
0.569378	0.743738	0.873575	0.000000	0.181818	0.128448	0.116061	0.260851	0.032679	0.116535	0.022406
0.196172	0.200385	0.436860	0.000000	0.409091	0.038048	0.049491	0.597814	0.386753	0.005202	0.031081
0.095694	0.892100	0.936959	0.095092	0.333333	0.022187	0.027163	0.211140	0.048766	0.877663	0.002493
0.392344	0.788054	0.911055	0.101227	0.075758	0.036199	0.044449	0.296909	0.089882	0.285586	0.020664
0.425837	0.643545	0.826329	0.000000	0.181818	0.049044	0.049388	0.578289	0.052665	0.096429	0.025275

test 데이터

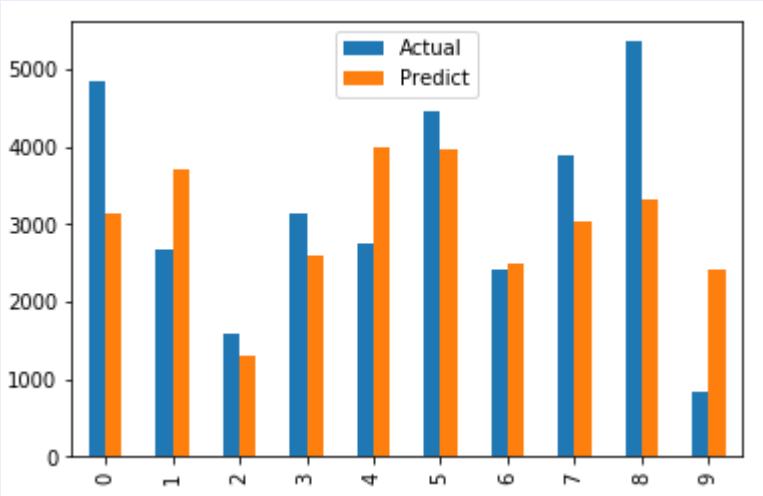
회귀 모델 알고리즘 비교

	Linear Regression	RandomForest Regressor	Kneighbors Regressor	DecisionTree Regressor
MSE	1462940.48	11363.59	166100.80	304554.08
RMSE	1209.52	106.60	407.55	551.86
MAE	909.22	25.53	190.196	367.04
R^2	0.4047	0.9954	0.9324	0.8761

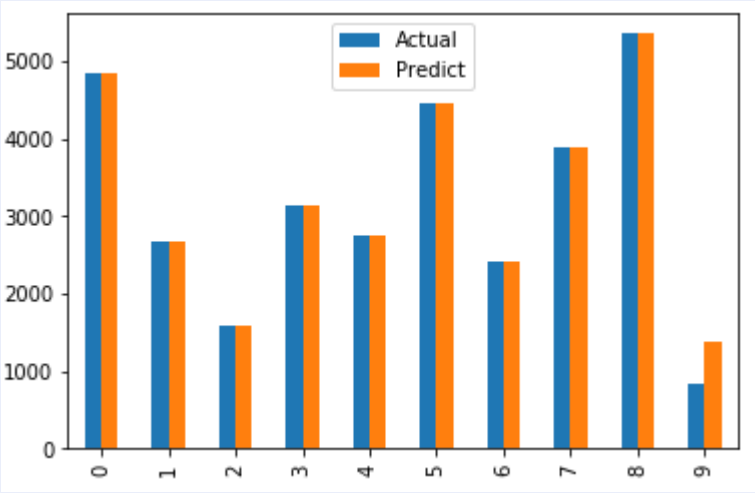
R^2 비교



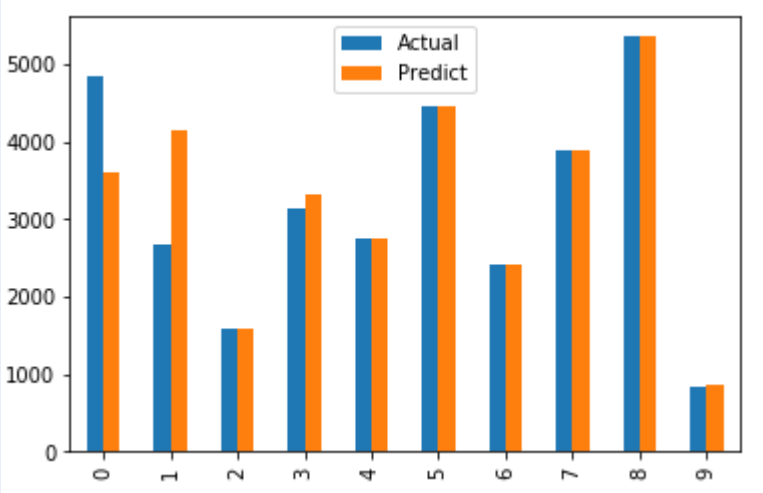
실제값과 예측값의 차이 시각화



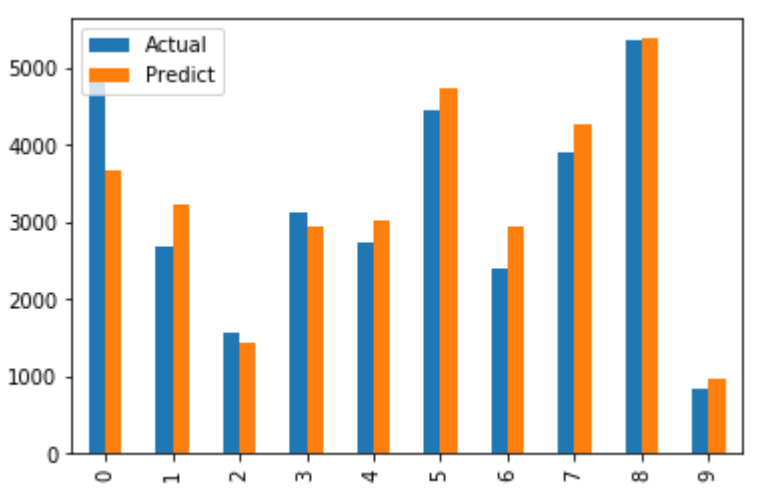
Linear Regression



RandomForest Regressor



Kneighbors Regressor

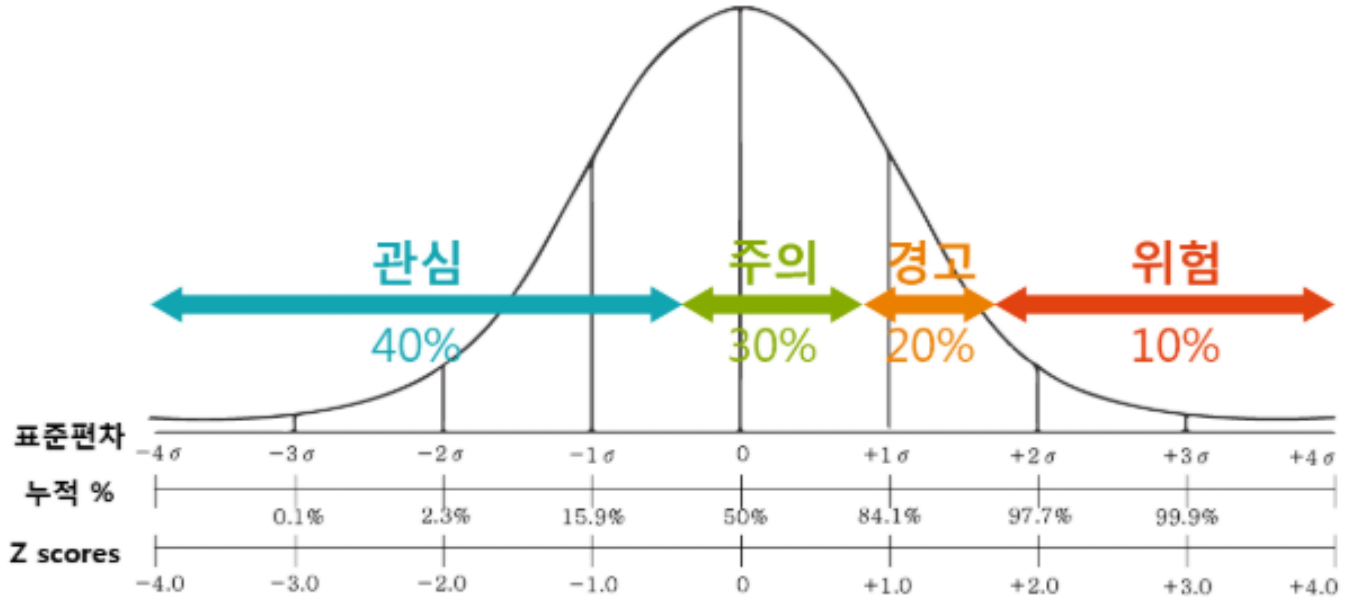


DecisionTree Regressor

Classification Model

단계	구간	구간 확률	누적 확률
● 관심	$-\infty \leq Z \leq -0.11$	40.00%	40.00%
● 주의	$-0.11 < Z \leq 0.58$	30.00%	70.00%
● 경고	$0.58 < Z \leq 1.14$	20.00%	90.00%
● 위험	$1.14 < Z \leq \infty$	10.00%	100%

* $Z=(X-\mu)/\sigma$ where μ=평균, σ=표준편차



출처 : 국민 건강 보험 공단

정규화

0	1	2	3	4	5	6	7	8	9	10
0.287081	0.373796	0.735853	0.0	0.121212	0.065392	0.093837	0.257393	0.086423	0.005202	0.057039
0.846890	0.547206	0.812548	0.0	0.106061	0.095752	0.125939	0.329562	0.035663	0.011804	0.044197
0.263158	0.308285	0.350220	0.0	0.651515	0.043205	0.051857	0.086440	0.014087	0.005202	0.023533
0.325359	0.886320	0.940811	0.0	0.181818	0.059553	0.071612	0.410241	0.194788	1.000000	0.007548
0.521531	0.622351	0.780678	0.0	0.287879	0.074539	0.110711	0.893688	0.121457	0.005202	0.016770

학습 데이터

0	1	2	3	4	5	6	7	8	9	10
0.569378	0.743738	0.873575	0.000000	0.181818	0.128448	0.116061	0.260851	0.032679	0.116535	0.022406
0.196172	0.200385	0.436860	0.000000	0.409091	0.038048	0.049491	0.597814	0.386753	0.005202	0.031081
0.095694	0.892100	0.936959	0.095092	0.333333	0.022187	0.027163	0.211140	0.048766	0.877663	0.002493
0.392344	0.788054	0.911055	0.101227	0.075758	0.036199	0.044449	0.296909	0.089882	0.285586	0.020664
0.425837	0.643545	0.826329	0.000000	0.181818	0.049044	0.049388	0.578289	0.052665	0.096429	0.025275

test 데이터

Classification Model

일교차	기온(°C)	체감온도	강수량(mm)	풍속(m/s)	미세먼지	초미세먼지	총생활인구수	총승객수	모기지수	검색량	class
7.7	-0.3	2.593495	0.0	0.9	56.000000	38.000000	384303.8087	23633	5.5	7.90886	2
19.4	8.7	12.190690	0.0	0.8	82.000000	51.000000	439648.8531	9753	12.1	6.65465	3
7.2	-3.7	-45.662570	0.0	4.4	37.000000	21.000000	253202.8034	3853	5.5	4.63657	0
8.5	26.3	28.240785	0.0	1.3	51.000000	29.000000	501520.6836	53265	1000.0	3.07548	0
12.6	12.6	8.202700	0.0	2.0	63.833333	44.833333	872269.2161	33213	5.5	3.97611	1

학습 데이터: 101481건

일교차	기온(°C)	체감온도	강수량(mm)	풍속(m/s)	미세먼지	초미세먼지	총생활인구수	총승객수	모기지수	검색량	class
13.6	18.9	19.827355	0.0	1.3	110.000000	47.000000	386955.7438	8937	116.8	4.52650	2
5.8	-9.3	-34.820810	0.0	2.8	32.583333	20.041667	645367.4155	105757	5.5	5.37376	1
3.7	26.6	27.758770	15.5	2.3	19.000000	11.000000	348832.9839	13336	877.7	2.58180	0
9.9	21.2	24.517280	16.5	0.6	31.000000	18.000000	414608.2604	24579	285.8	4.35638	1
10.6	13.7	13.915215	0.0	1.3	42.000000	20.000000	630393.9546	14402	96.7	4.80669	1

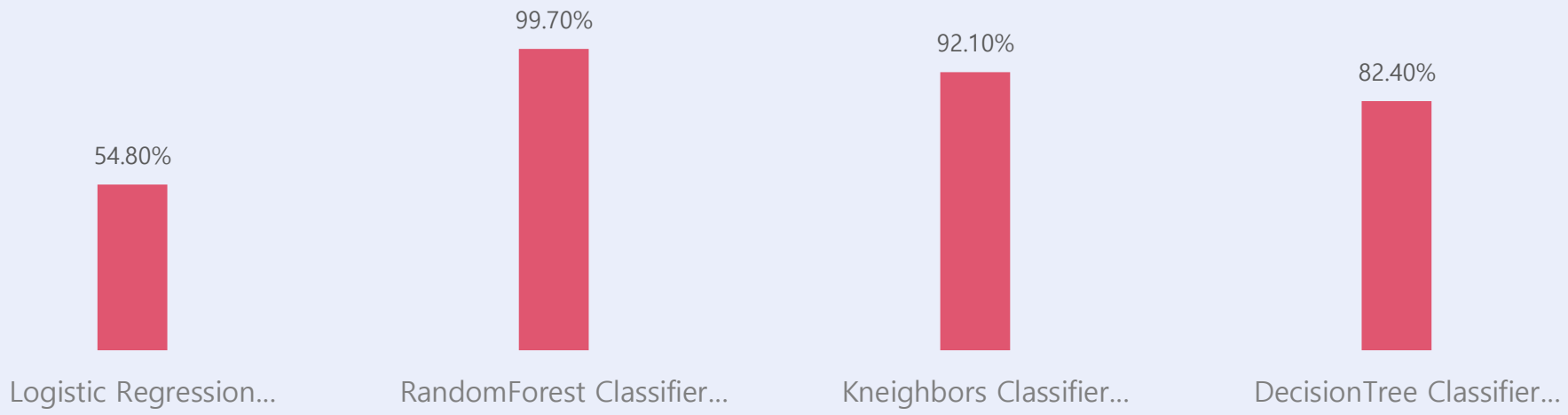
테스트 데이터: 25371건

총 126852건 데이터 중 101481건의 학습 데이터를 이용하여 훈련시켰으며
테스트 데이터 25371건으로 모델의 정확성을 평가하였습니다.

분류 모델 알고리즘 비교

	Logistic Regression	RandomForest Classifier	Kneighbors Classifier	DecisionTree Classifier
accuaracy	0.548	0.997	0.921	0.824
weighted avg precision	0.67	1.00	0.92	0.83
weighted avg recall	0.55	1.00	0.92	0.82

정확도 비교



결론

예측

2월 13일 (회귀)

일교차	기온(°C)	체감온도	강수량(mm)	풍속(m/s)	미세먼지	초미세먼지	총생활인구수	총승객수	모기지수	검색량
6.2	7.4	6.75243	0	1.3	66	48	697349.3157	124346	5.5	24.49889

→ 3122

2월 13일 (분류)

일교차	기온(°C)	체감온도	강수량(mm)	풍속(m/s)	미세먼지	초미세먼지	총생활인구수	총승객수	모기지수	검색량
6.2	7.4	6.75243	0	1.3	66	48	697349.3157	124346	5.5	24.49889

→ 주의

예방 Tip

관심

환기를 자주 시켜 깨끗한 환경을 유지하고 외출 후에는 반드시 손을 씻는 등 평소 손 씻기를 생활화합니다.

주의

기침과 재채기를 할 때에는 반드시 휴지나 손수건으로 가리는 등 기침 에티켓을 지켜주시고 충분한 휴식 및 수분을 섭취합니다.

경고

발열이나 호흡기 증상이 있다면 외출을 삼가되 외출 시에는 마스크를 착용하고 가까운 의료기관에 방문하여 전문의의 진료를 받습니다.

위험

고위험 집단(만성심장폐질환, 천식, 당뇨병 환자, 임산부, 65세 이상 어르신 등)은 중증으로 진행될 수 있으므로 발열, 호흡기 증상이 있으면 인근 의료기관에서 바로 진료를 받으시기 바랍니다.

참고문헌

- 국민건강보험공단
- 데이터 사이언스 스쿨
- 김현준, 정진선, 이관호, 이원찬, 이효철, 테레시아, 이석원. (2017). 기상에 따른 감기위험도 예측 분석. 한국정보과학회 학술발표논문집, (), 1947-1949.

감사합니다