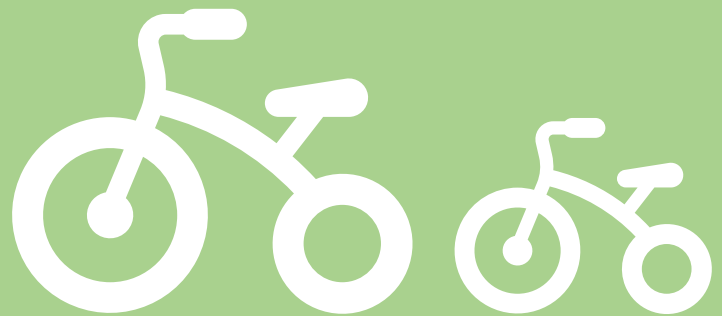


서울시 공공자전거 신규 대여소 이용량 예측

통계자료분석캡스톤디자인 발표 자료

통계학과 2018580028 정세린

분석 주제 및 목적

- 공공자전거 대여소 주변의 환경변수를 이용하여,
  공공자전거 대여소 이용량 예측 모형을 수립한다.

- 모형을 통해 얻은 예측 이용량을 바탕으로,
  공공자전거 신규 대여소 설치 후보지를 선정한다

데이터 구성

- 기존 대여소 데이터
  : 1549개의 거치대에 대한 데이터로,
  39개의 주변 환경 변수와, 일별 평균 대여량, 반납량을 포함한다.

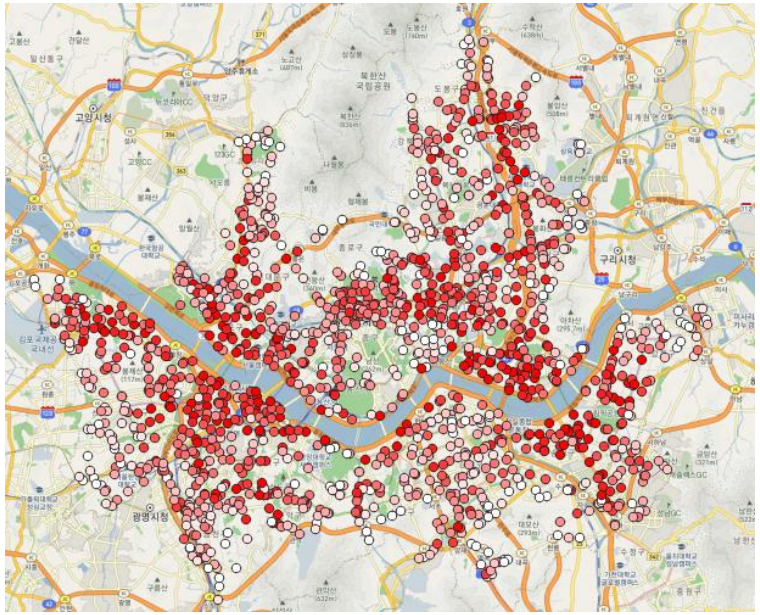
- 신규 거치대 후보지 데이터
  : 18256 곳의 후보지에 대한 데이터로, 일정 반경 이내의
  환경 변수에 대한 정보를 포함한다.

- 대여소의 이용량은 대여량과 반납량의 합으로 정의하였다.

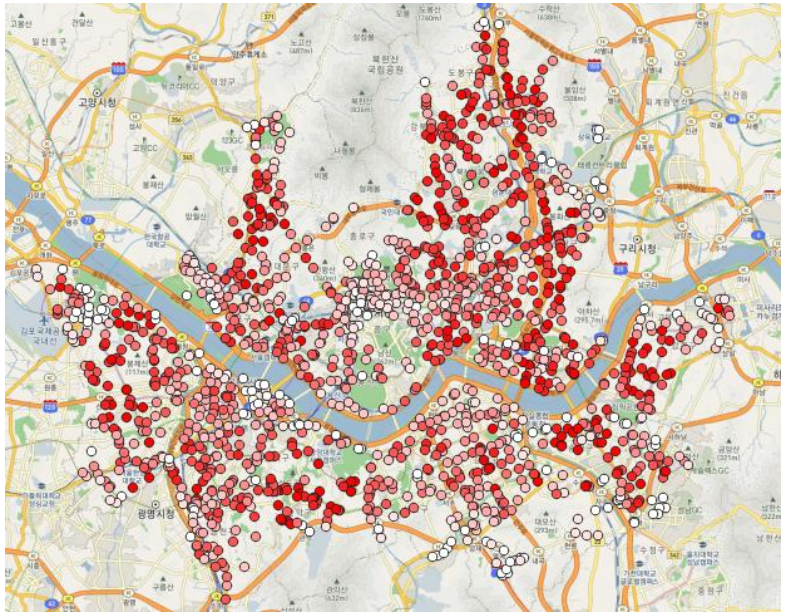
탐색적 자료분석

- 주거인구보다는 유동인구와 이용량의 연관성이 더 높게 나타났다.

총 유동인구
이용량과의 상관계수 : 0.31



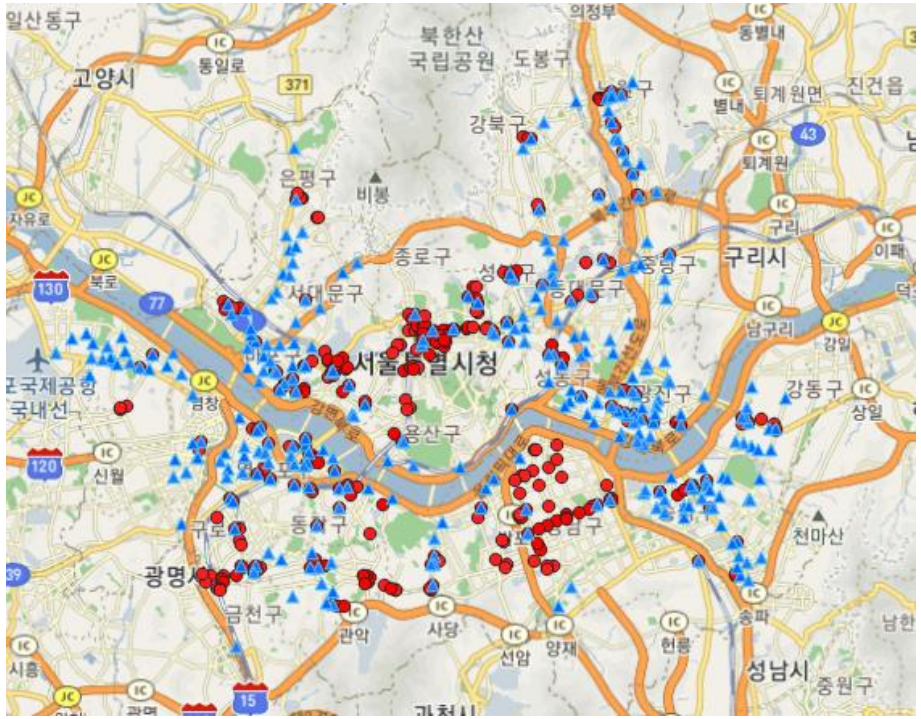
거치대의 이용량



총 주거인구
이용량과의 상관계수 : 0.07

탐색적 자료분석

특히 20대 유동인구와 공공자전거 이용량의 연관성이 높게 나타났다.



▲ 파랑 : 수요량 상위 20% / 빨강 : 20대 유동인구 상위 20%

연령대	상관계수	연령대	상관계수
10대 미만	0.24	40대	0.26
10대	0.29	50대	0.29
20대	0.31	60대	0.28
30대	0.26	70대 이상	0.28

▲ 연령대별 유동인구와 공공자전거 수요량 사이의 상관계수

모형 적합 및 결과

1. $Y = \beta_0 + \sum_{j=1}^{39} \beta_j X_j + \varepsilon$
2. $Y = \beta_0 + \sum_{i=1}^{39} \beta_i X_i + \sum_{i=1}^{39} \beta_{i+39} X_i^2 + \varepsilon$
3. $Y = \beta_0 + \sum_{i=1}^{39} \beta_i X_i + \sum_{i=1}^{39} \sum_{j=1}^{i} \beta_{i(i-1)/2+j+39} X_i X_j + \varepsilon$
4. Random Forest model

- 선형회귀모형은 설명변수를 동일하게 구성하더라도 세가지로 구분하였다.

  a. 변수 선택을 하지 않은 모형

  b. AIC를 기준으로 Stepwise selection을 진행한 모형

  c. BIC를 기준으로 Stepwise selection을 진행한 모형

모형 적합 및 결과

- 모형 평가는 10-fold CV를 이용하여 진행하였다.
  이때 이용량을 예측하는 문제이므로, $\hat{y} = \max(\hat{f(x)}, 0)$을 이용하였다.

- 선형회귀모형 중에서는 각 설명변수의 일차항, 이차항을 모두 포함한
  2 – a 모형이 가장 좋은 성능을 보였다. (교호작용은 포함되지 않는다.)

	1 – a	1 - b	1 – c	2 - a	2 - b	2 - c	3 - a	3- b	3 - c
Train MSE	1798.84	1829.97	1888.08	1592.72	1657.95	1794.06	555.96	1224.14	1708.29
Test MSE	1899.23	1897.73	1941.99	1799.62	1828.35	1934.45	4117.96	2071.35	1906.56

▲ 10-fold CV를 이용하여 구한 선형회귀모형 들의 Train MSE, Test MSE

모형 적합 및 결과

- 최종적으로 선형회귀모형과 Random Forest 모형의 성능을 비교한 결과는 다음과 같다.

- MSE(좌), 이용량 상위 20% 대여소에 대한 오분류율(우)

	선형회귀모형 (2-a)	Random Forest
Train MSE	1592.72	324.66
Test MSE	1799.62	1797.95

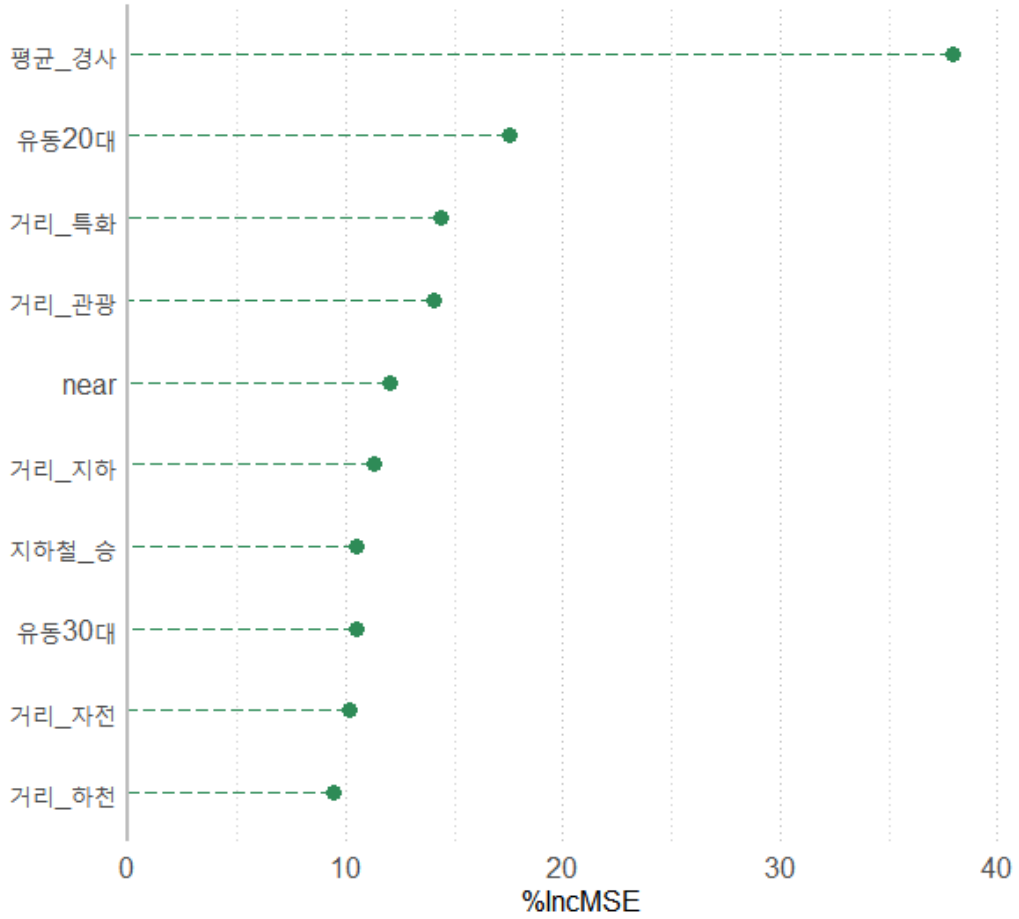
▲ 10-fold CV를 이용한, 각 모형의 Train MSE, Test MSE

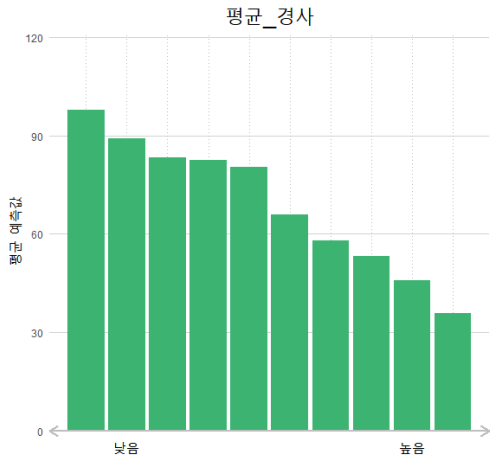
	선형회귀모형 (2-a)	Random Forest
Train Error rate	17.23%	4.32%
Test Error rate	18.70%	18.38%

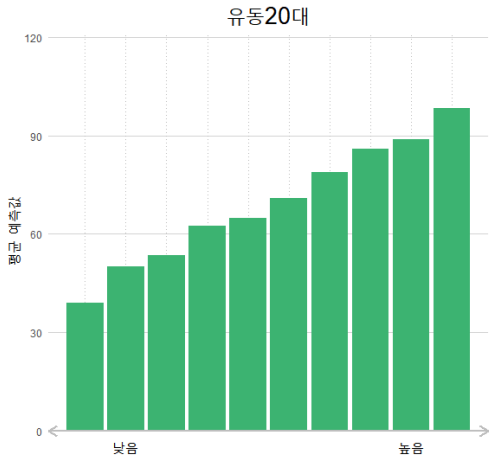
▲ 10-fold CV를 이용한, 상위 20% 관측값에 대한 각 모형의 오분류율

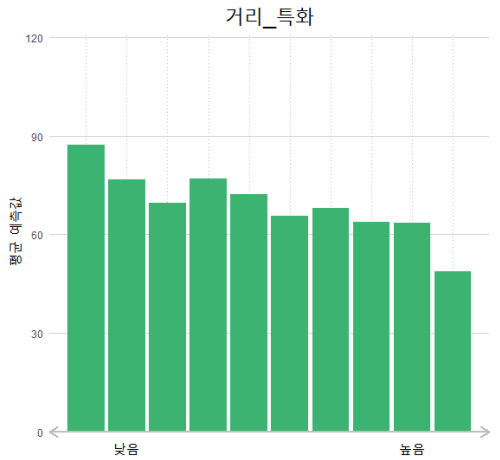
모형 적합 및 결과

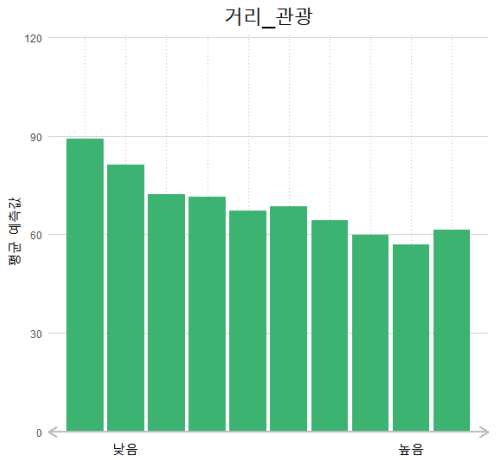
- 설명변수 중요도와 예측값과 설명변수의 관계



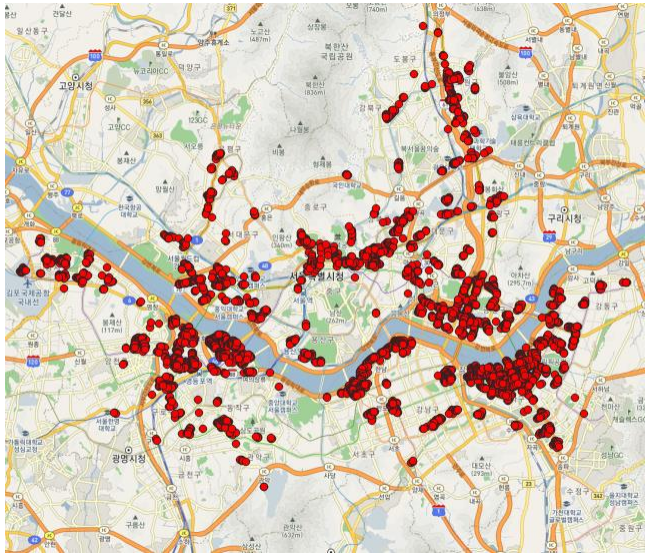




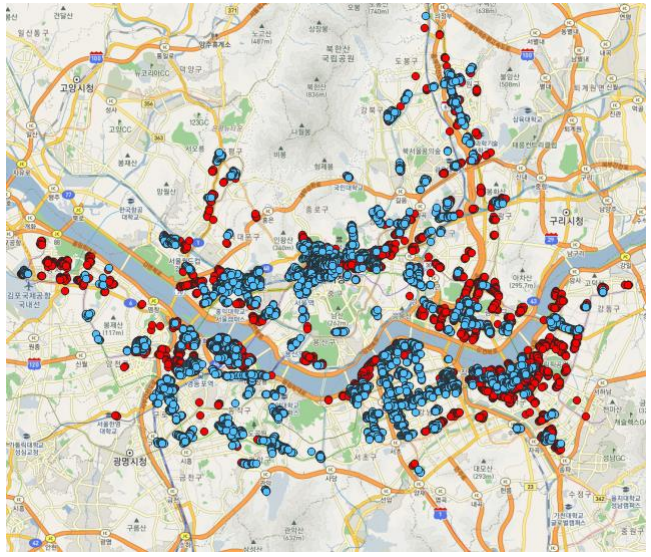




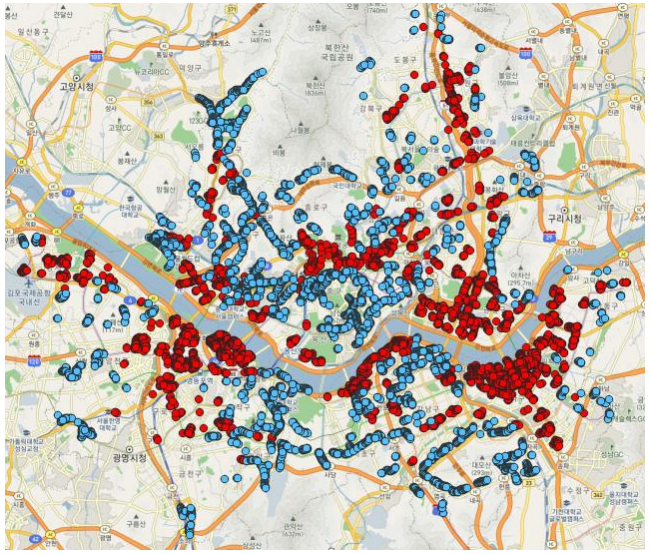
예측 이용량 분석



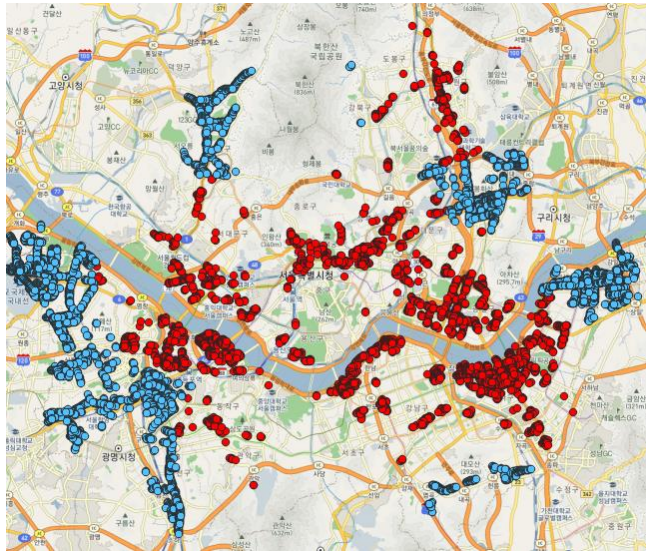
◀ 예측 이용량
상위 20%



◀ 20대 유동인구
상위 20%
(상관계수 : 0.53)



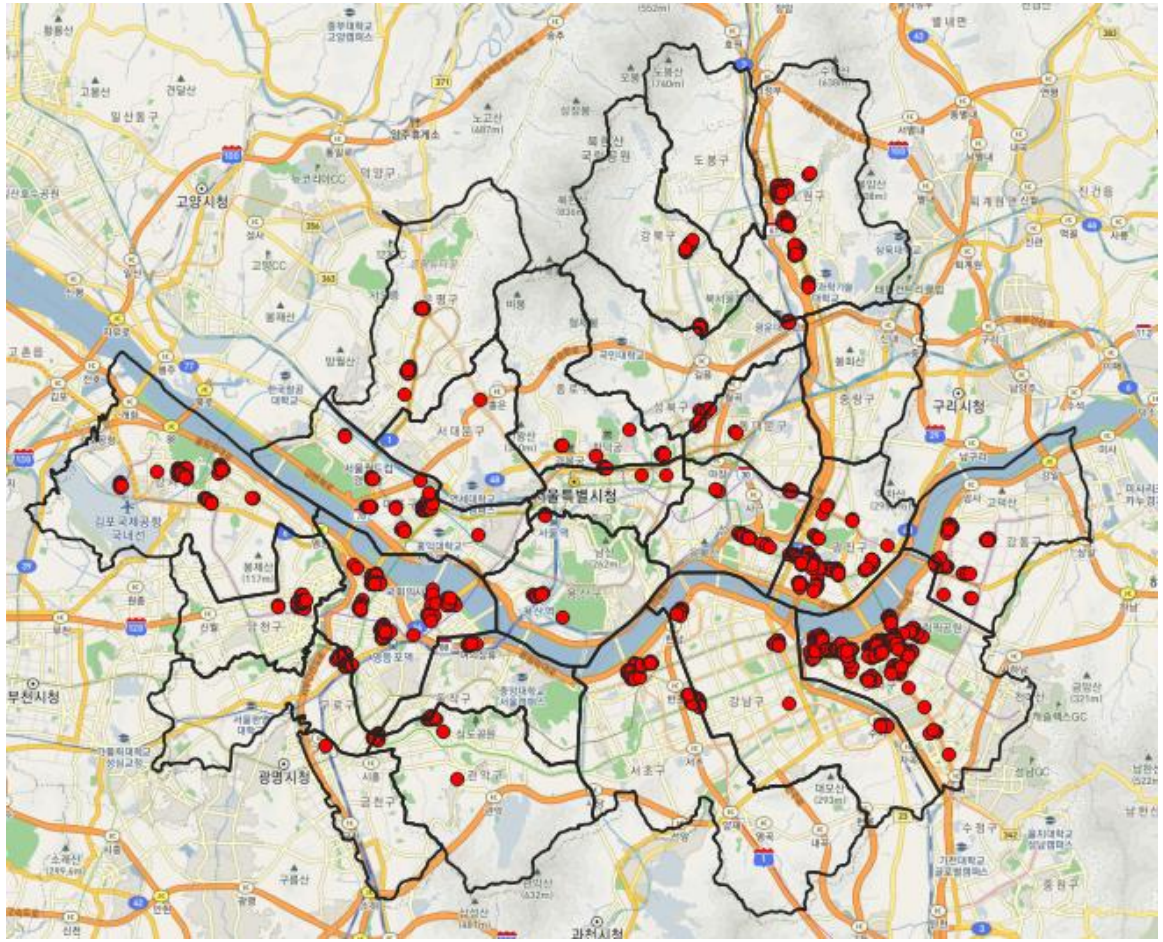
◀ 평균 경사도
상위 20%
(상관계수 : -0.57)



◀ 특화지역과의 거리
상위 20%
(상관계수 : -0.30)

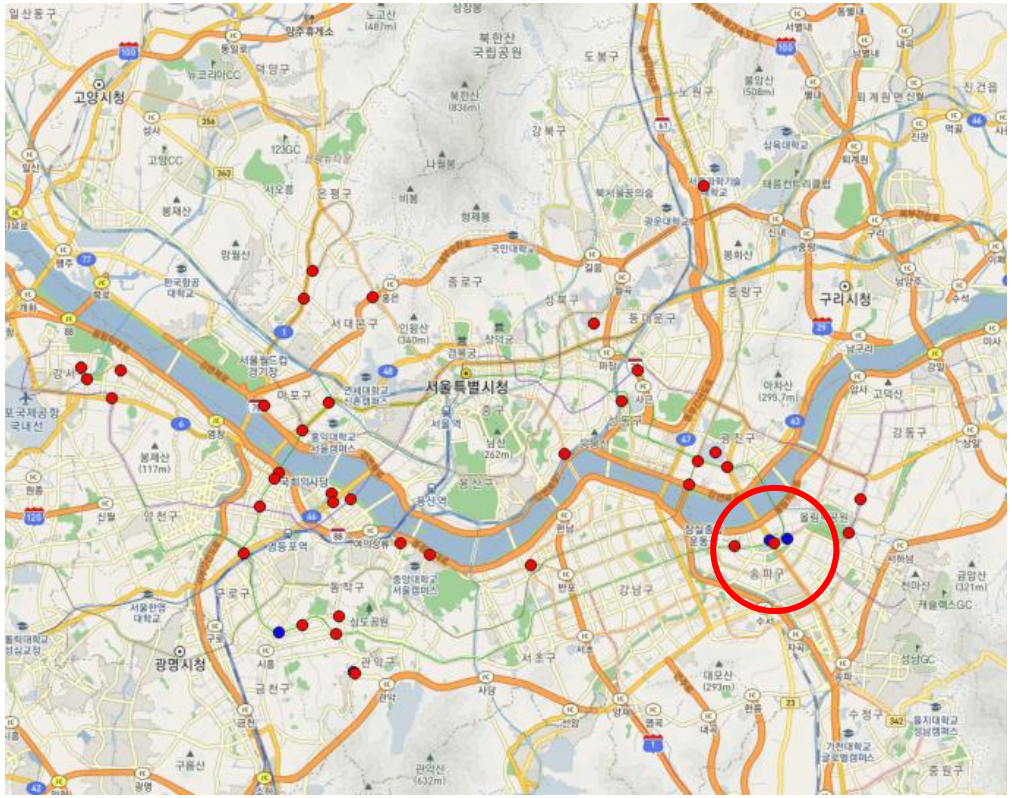
예측 이용량 분석

- 신규 대여소 후보지 중, 예측 이용량 상위 5%



개선할 점

- 잔차의 절대값이 100 이상인 대여소들을 표시한 결과,
  주로 한강 부근의 대여소들이 과소 추정된 것으로 나타났다.



▲ 파랑 : 과대 추정된 거치대 / 빨강 : 과소 추정된 거치대

개선할 점

대여소 번호	평균 경사도	20대 유동인구	예측 이용량	실제 이용량
1231번	0.44	10516.88	260.87	138.85
2608번	0.53	19310.13	214.00	98.82
1210번	0.83	4623.41	189.11	401.78