서울 공공자전거 따릉이 신규대여소 이용량 예측

2017580033 통계학과 길경주

목차

1. EDA
2. 분석한 모델 설명
3. 변수 간의 관계를 바탕으로 한 예측량 설명
4. 서울시 공공자전거 신규대여소 예측

EDA

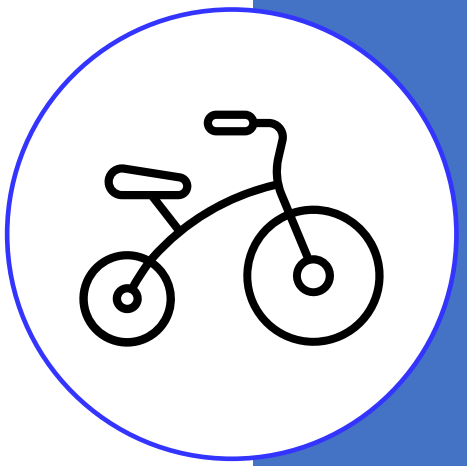
- 결측치 존재
- 변수간 correleation 탐색

  1) 종속변수들 간 상관관계 약 0.99

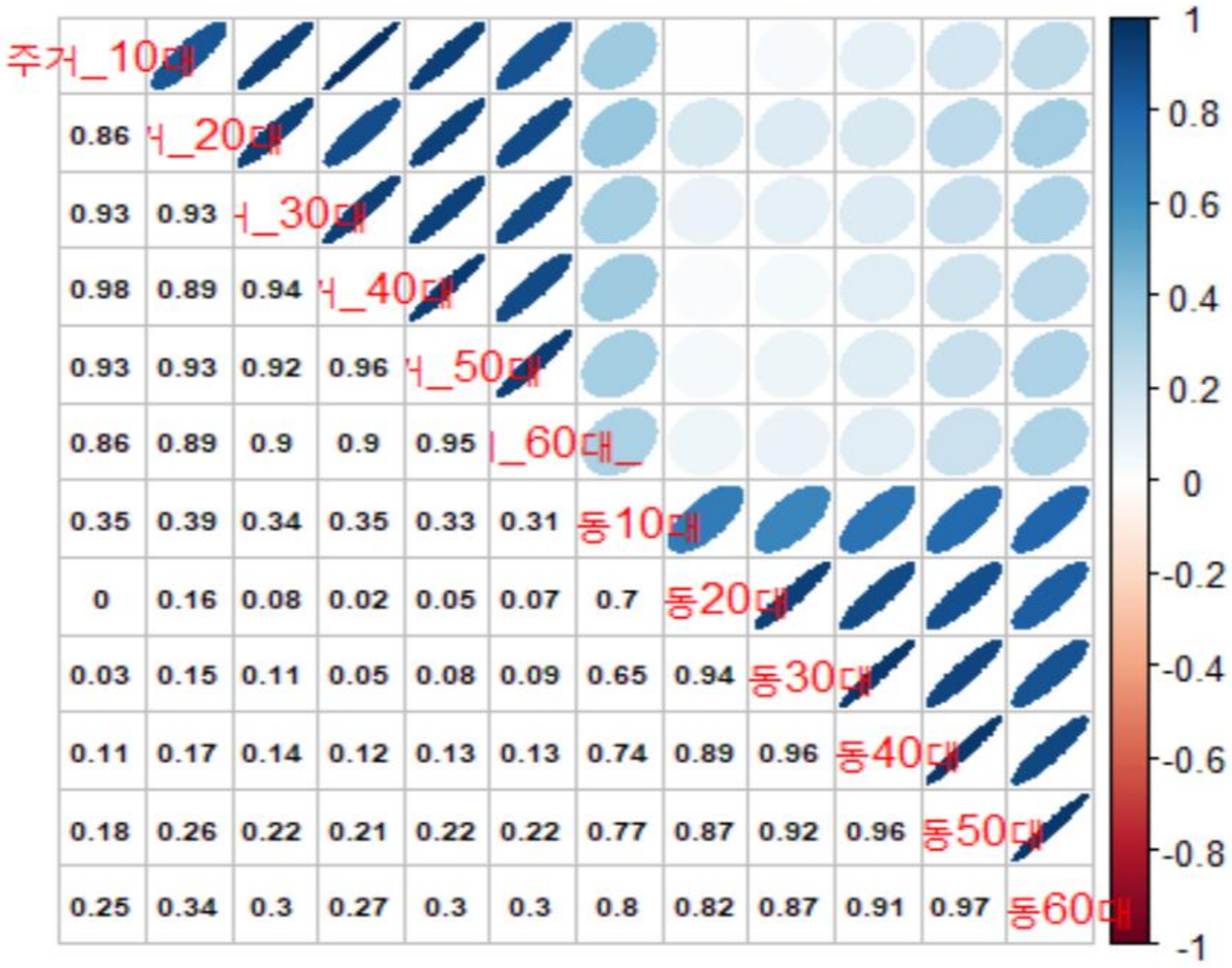
  -> 대여+반납 = 이용량 변수 추가

  2) 독립변수간 상관관계

  -> 주거인구와 유동인구에서의 상관관계를 제외하고 높은 correlation을 보이는 변수들은 없다.



주거인구 & 유동인구 correlation plot



분석한 모델

- 다중회귀모델
- 다중회귀모델&변수선택법
- 교호작용을 포함한 다항회귀모델
- 랜덤 포레스트

1. 다중회귀모델

```
lm(formula = amount ~ 버스_승객 + 버스_경유 + 지하철_승 + 면적_아파 +
    면적_주거_ + 거리_지하 + 거리_자전 + 거리_공공 + 거리_문화 +
    거리_영화 + 거리_관광 + 거리_대학 + 거리_초중 + 거리_상업 +
    거리_의료 + 거리_주차 + 거리_체육 + 거리_공원 + 거리_특화 +
    거리_교통 + 거리_하천 + 평균_경사 + 주거_10대 + 주거_20대 +
    주거_30대 + 주거_40대 + 주거_50대 + 주거_60대_ + near + medium +
    far + 유동10대미 + 유동10대 + 유동20대 + 유동30대 + 유동40대 +
    유동50대 + 유동60대 + 유동70대이, data = join_2)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	5.257e+01	4.324e+00	12.157	< 2e-16	***
버스_승객	6.890e-04	1.428e-04	4.825	1.54e-06	***
버스_경유	-1.793e-01	6.874e-02	-2.609	0.009180	**
지하철_승	1.109e-04	2.522e-05	4.397	1.17e-05	***
면적_아파	2.263e-04	1.311e-04	1.725	0.084686	.
면적_주거_	1.154e-04	7.219e-05	1.599	0.109975	
거리_지하	-4.659e-03	1.729e-03	-2.695	0.007119	**
거리_자전	-6.234e-03	1.603e-03	-3.890	0.000105	***
거리_공공	1.978e-03	2.278e-03	0.868	0.385529	
거리_문화	5.244e-03	1.333e-03	3.934	8.72e-05	***
거리_영화	-3.339e-03	9.911e-04	-3.369	0.000772	***
거리_관광	-2.172e-03	3.375e-04	-6.436	1.64e-10	***
거리_대학	-1.018e-03	5.980e-04	-1.703	0.088800	.
거리_초중	4.479e-03	3.203e-03	1.399	0.162152	
거리_상업	-4.296e-04	9.554e-04	-0.450	0.652991	
거리_의료	1.013e-03	7.452e-04	1.359	0.174275	
거리_주차	4.382e-04	1.723e-03	0.254	0.799334	
거리_체육	4.675e-04	5.106e-04	0.916	0.360064	
거리_공원	-6.011e-03	1.235e-03	-4.866	1.26e-06	***

```
Residual standard error: 21.62 on 1511 degrees of freedom
Multiple R-squared:  0.3835,    Adjusted R-squared:  0.3676
F-statistic: 24.11 on 39 and 1511 DF,  p-value: < 2.2e-16
```

RMSE	Rsquared	MAE
21.84666	0.3583578	14.6641

√ 전체 모델 유의수준 0.05하에서 유의
√ 유의하지 않은 변수들이 존재
√ Adj R-squared : 0.37
√ 다중공선성 존재
√ RMSE : 21.85

2. 다중회귀모델&변수선택법

```
lm(formula = amount ~ 버스_승객 + 버스_경유 + 지하철_승 + 면적_아파 +
    면적_주거_ + 거리_지하 + 거리_자전 + 거리_문화 + 거리_영화 +
    거리_관광 + 거리_대학 + 거리_초중 + 거리_의료 + 거리_공원 +
    거리_특화 + 거리_교통 + 거리_하천 + 평균_경사 + 주거_10대 +
    주거_20대 + 주거_40대 + 주거_50대 + near + medium + 유동10대 +
    유동20대 + 유동30대 + 유동40대 + 유동50대 + 유동60대 + 유동70대이,
    data = join_2)
```

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	5.516e+01	3.400e+00	16.225	< 2e-16	***
버스_승객	6.762e-04	1.416e-04	4.776	1.96e-06	***
버스_경유	-1.747e-01	6.783e-02	-2.576	0.010082	*
지하철_승	1.086e-04	2.496e-05	4.348	1.46e-05	***
면적_아파	2.551e-04	1.289e-04	1.979	0.048002	*
면적_주거_	1.093e-04	6.924e-05	1.578	0.114683	
거리_지하	-4.631e-03	1.678e-03	-2.759	0.005860	**
거리_자전	-6.194e-03	1.576e-03	-3.930	8.87e-05	***
거리_문화	5.619e-03	1.278e-03	4.395	1.18e-05	***
거리_영화	-3.528e-03	9.562e-04	-3.689	0.000233	***
거리_관광	-2.202e-03	3.285e-04	-6.704	2.86e-11	***
거리_대학	-1.098e-03	5.808e-04	-1.890	0.058939	.
거리_초중	4.508e-03	3.117e-03	1.446	0.148364	
거리_의료	1.029e-03	7.185e-04	1.432	0.152249	
거리_공원	-5.893e-03	1.211e-03	-4.866	1.26e-06	***
거리_특화	-8.476e-04	3.570e-04	-2.374	0.017703	*
거리_교통	1.015e-03	1.749e-04	5.803	7.90e-09	***
거리_하천	-4.212e-03	8.347e-04	-5.046	5.07e-07	***
평균_경사	-4.087e+00	3.397e-01	-12.032	< 2e-16	***

```
Residual standard error: 21.59 on 1519 degrees of freedom
Multiple R-squared:  0.3821,    Adjusted R-squared:  0.3695
F-statistic:  30.3 on 31 and 1519 DF,  p-value: < 2.2e-16
```

RMSE	Rsquared	MAE
21.80257	0.3544772	14.66901

√ 전체 모델 유의수준 0.05하에서 유의
√ 유의하지 않은 변수들이 존재
√ Adj R-squared : 0.37
√ 다중공선성 존재
√ RMSE : 21.80

3. 교호작용을 포함한 다항회귀모델

```
lm(formula = amount ~ .^2, data = join_2)

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)             5.833e+01  3.914e+01   1.490  0.13658
버스_승객               5.706e-03  2.441e-03   2.338  0.01964 *
버스_경유              -1.981e+00  9.244e-01  -2.143  0.03240 *
지하철_승              -1.563e-04  4.399e-04  -0.355  0.72240
면적_아파               1.828e-03  1.468e-03   1.246  0.21322
면적_주거_              1.405e-03  8.927e-04   1.574  0.11597
거리_지하              -4.446e-02  2.267e-02  -1.961  0.05028 .
거리_자전              -2.665e-02  2.224e-02  -1.198  0.23123
거리_공공               2.672e-02  2.935e-02   0.910  0.36289
거리_문화               4.705e-03  1.657e-02   0.284  0.77653
거리_영화              -2.140e-02  1.297e-02  -1.650  0.09939 .
거리_관광              -6.748e-03  5.289e-03  -1.276  0.20245
거리_대학               2.606e-04  8.513e-03   0.031  0.97559
거리_초중              -3.696e-02  3.530e-02  -1.047  0.29544
거리_상업               2.266e-02  1.357e-02   1.670  0.09540 .
버스_승객:버스_경유    -1.769e-05  1.064e-05  -1.662  0.09682 .
버스_승객:지하철_승    -1.737e-09  7.004e-09  -0.248  0.80424
버스_승객:면적_아파     3.504e-08  7.100e-08   0.493  0.62181
버스_승객:면적_주거_    5.485e-08  3.810e-08   1.440  0.15040
버스_승객:거리_지하     2.978e-06  1.220e-06   2.440  0.01492 *
버스_승객:거리_자전    -1.716e-06  8.208e-07  -2.091  0.03689 *
버스_승객:거리_공공     1.863e-06  1.702e-06   1.094  0.27423
버스_승객:거리_문화    -1.022e-06  8.755e-07  -1.167  0.24358
버스_승객:거리_영화    -6.067e-07  6.273e-07  -0.967  0.33378

Residual standard error: 18.03 on 770 degrees of freedom
Multiple R-squared:  0.7816,    Adjusted R-squared:  0.5604
F-statistic: 3.534 on 780 and 770 DF,  p-value: < 2.2e-16
```

```
RMSE
37.92061
```

√ 전체 모델 유의수준 0.05하에서 유의
√ 유의하지 않은 변수들이 존재
√ Adj R-squared : 0.56
√ 다중공선성 존재
√ RMSE : 37.92

4. 랜덤포레스트

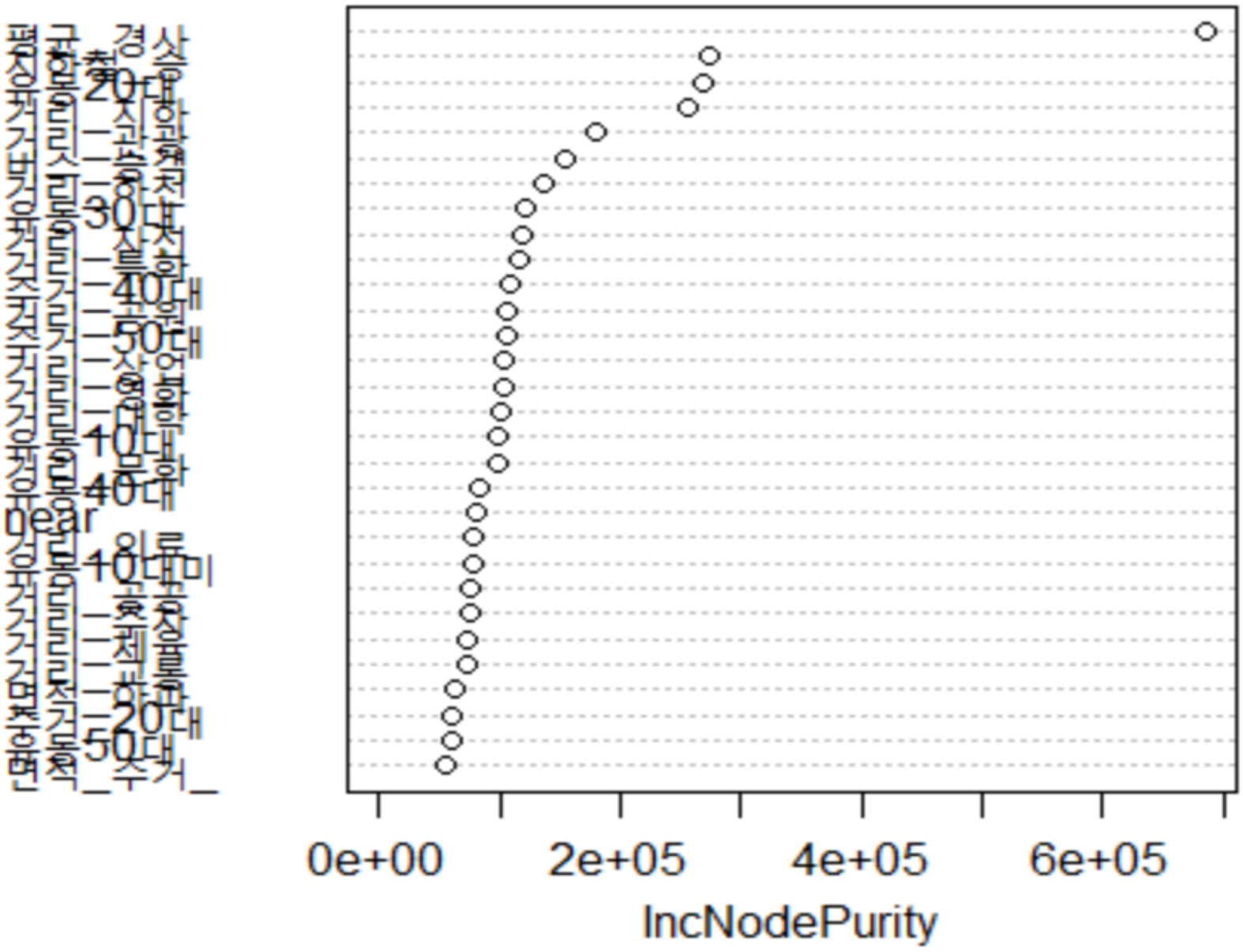
```
> rf1=randomForest(join_2$amount~.,data=join_2,
+                  mtry=20,importance=TRUE)
> print(rf1)

Call:
 randomForest(formula = join_2$amount ~ ., data = join_2,
 mtry = 20,      importance = TRUE)
               Type of random forest: regression
                     Number of trees: 500
No. of variables tried at each split: 20

          Mean of squared residuals: 447.2381
                    % Var explained: 39.46
```

mtry	RMSE	Rsquared	MAE
2	21.80691	0.3802058	14.63908
20	20.96892	0.4065443	13.97045
39	21.31432	0.3901033	14.12012

변수의 상대적 중요도



√ 평균_경사
√ 지하철_승객
√ 유동_20대
√ 거리_지하
√ 거리_관광
√ 버스_승객
√ 거리_하천

모델 비교 및 선택

10 fold cv 이용

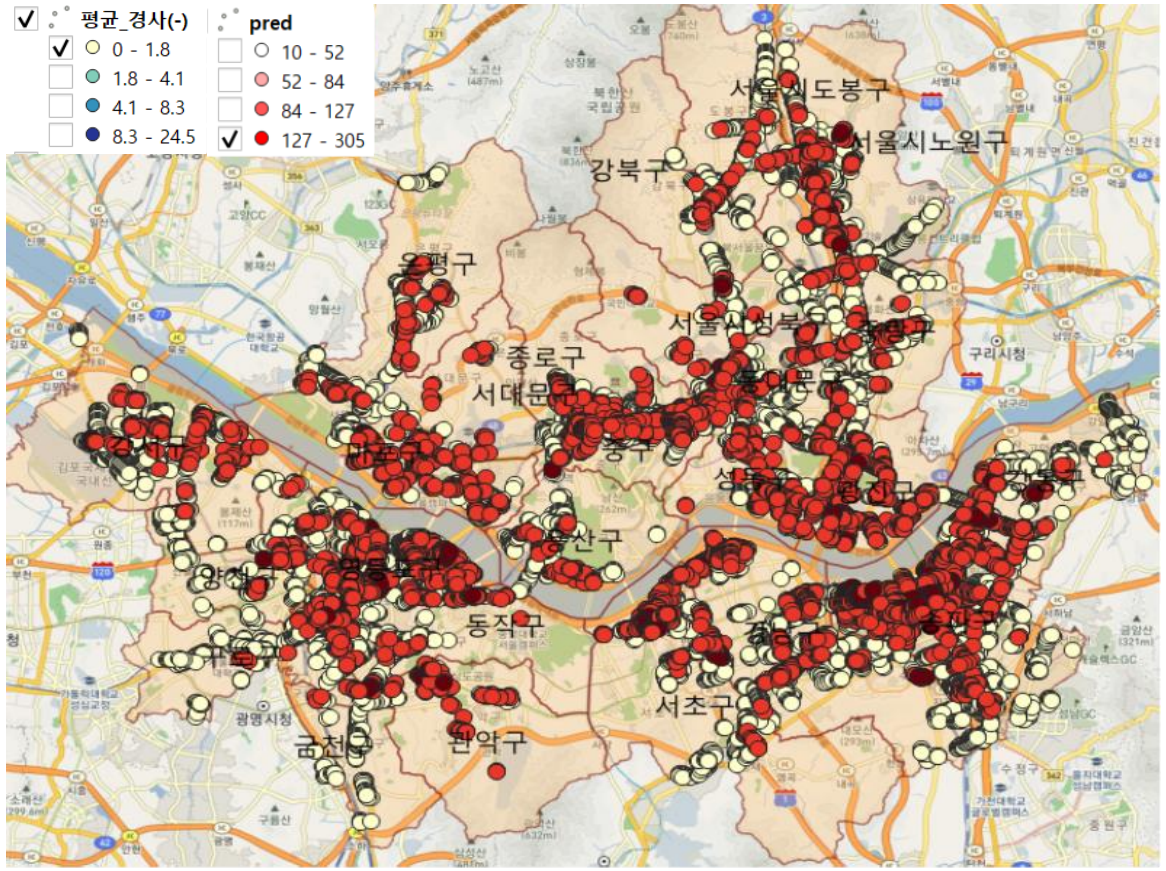
- 다중회귀모델 RMSE : 21.85
- 다중회귀모델&변수선택법 RMSE : 21.80
- 교호작용을 포함한 다항회귀모델 RMSE : 37.92
- 랜덤 포레스트 RMSE : 20.97

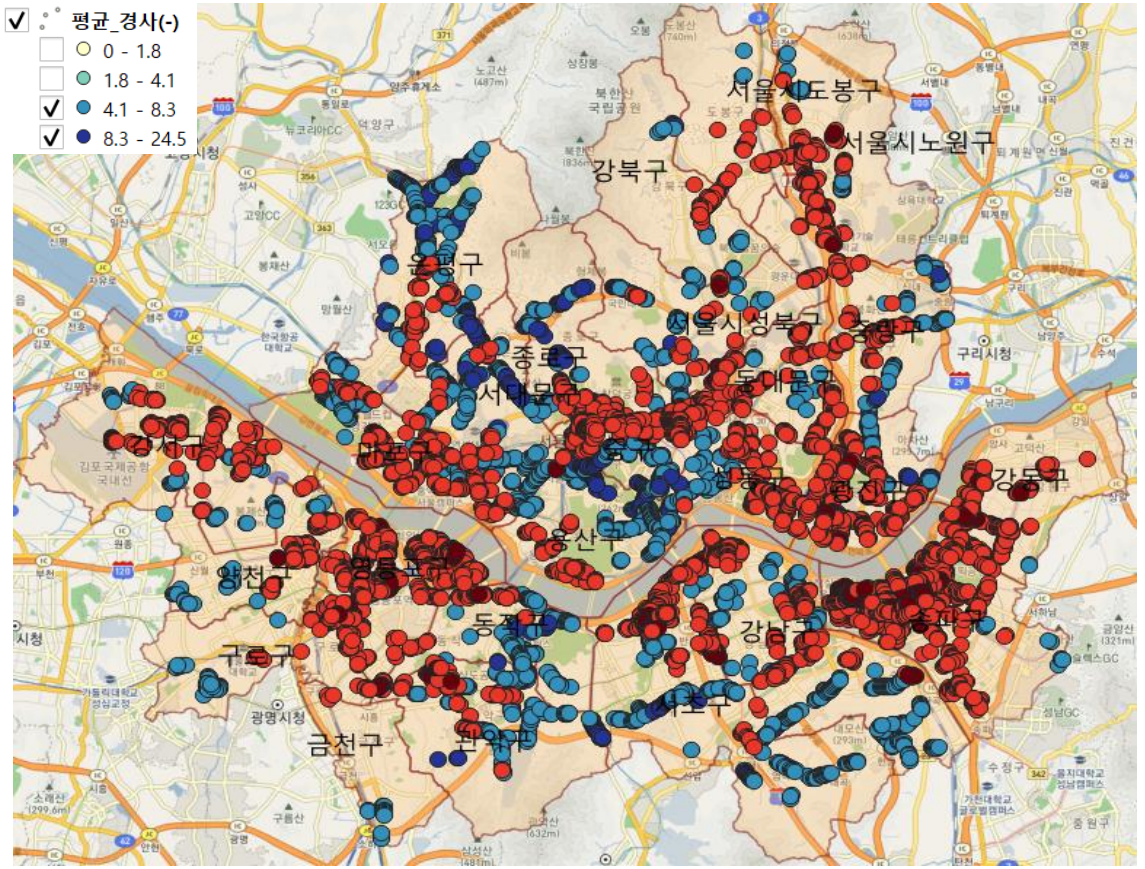
-> 최종적으로 랜덤 포레스트 모델 선택

따릉이 이용량과 높은 상관계수를 가지는 변수들

amount	유동20대	유동30대	유동50대
1.00000000	0.41708721	0.38812334	0.37308166
유동40대	유동60대	유동70대이	유동10대
0.36877411	0.34501763	0.34366766	0.32403085
far	지하철_승	유동10대미	버스_승객
0.29507692	0.28672577	0.28639016	0.28396811
medium	버스_경유	주거_20대	거리_초중
0.18496781	0.13300510	0.10079769	0.09917987
주거_30대	주거_40대	주거_10대	주거_50대
0.09211705	0.06324440	0.05990703	0.05418305
주거_60대_	near	면적_아파	면적_주거_
0.04421036	0.02382641	0.01561772	0.01346000
거리_문화	거리_체육	거리_교통	거리_대학
-0.01380638	-0.03032593	-0.07284303	-0.07382987
거리_공원	거리_주차	거리_의료	거리_하천
-0.07729609	-0.09315773	-0.12596766	-0.14036322
거리_공공	거리_상업	거리_자전	거리_특화
-0.14185619	-0.17442092	-0.23611132	-0.23888320
거리_관광	거리_지하	거리_영화	평균_경사
-0.24874536	-0.32520428	-0.33268498	-0.48534279

예측 총 이용량 vs 평균경사도



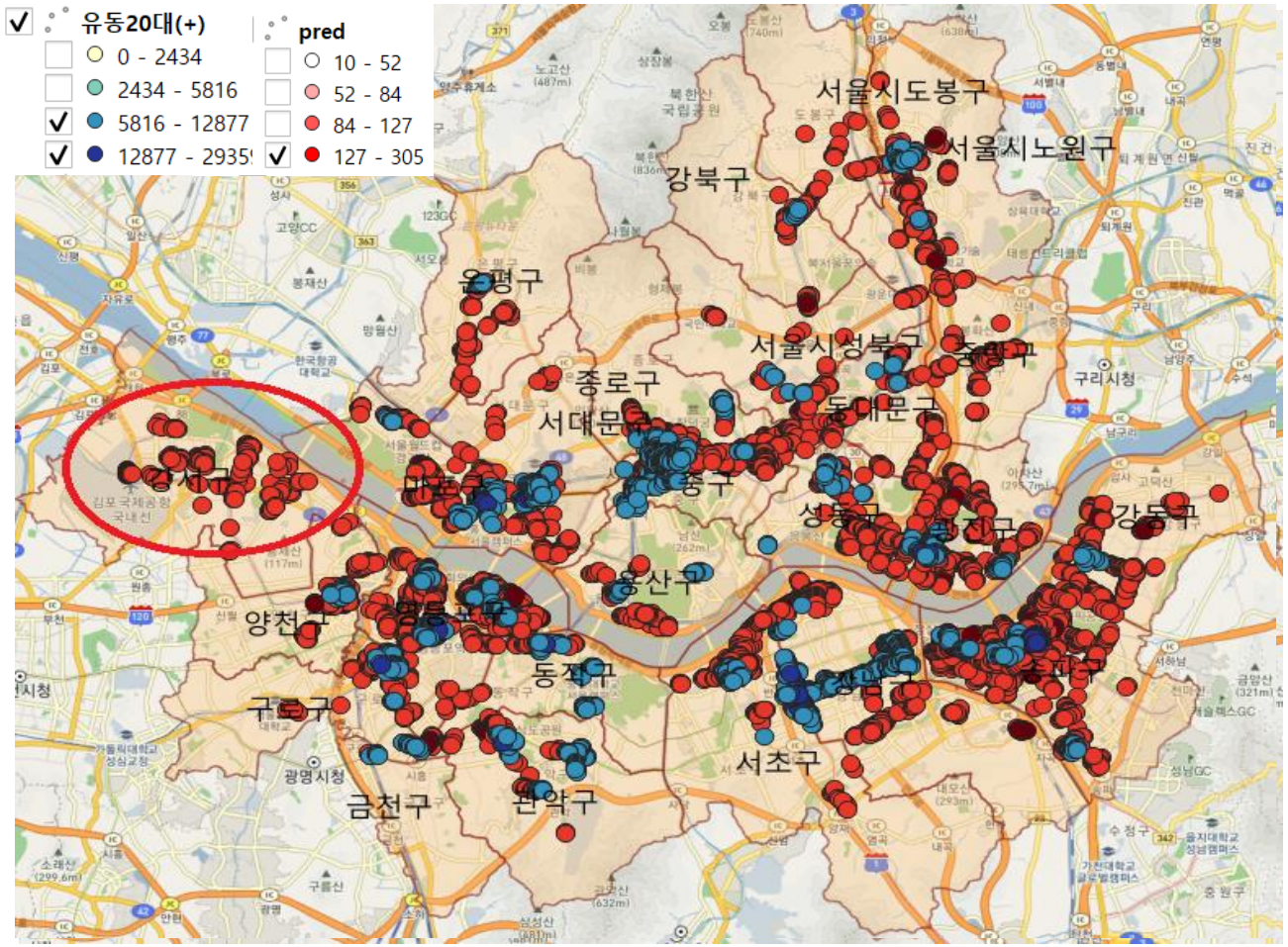


평균 경사도(각도) : 200m 반경 내 경사도의 평균

따릉이 이용량과 평균 경사도의 correlation : - 0.485

-> 200m 반경 내 경사도의 평균이 낮을 수록 따릉이 이용량이 높은 것을 알 수 있다.

예측 총 이용량 vs 유동 20대



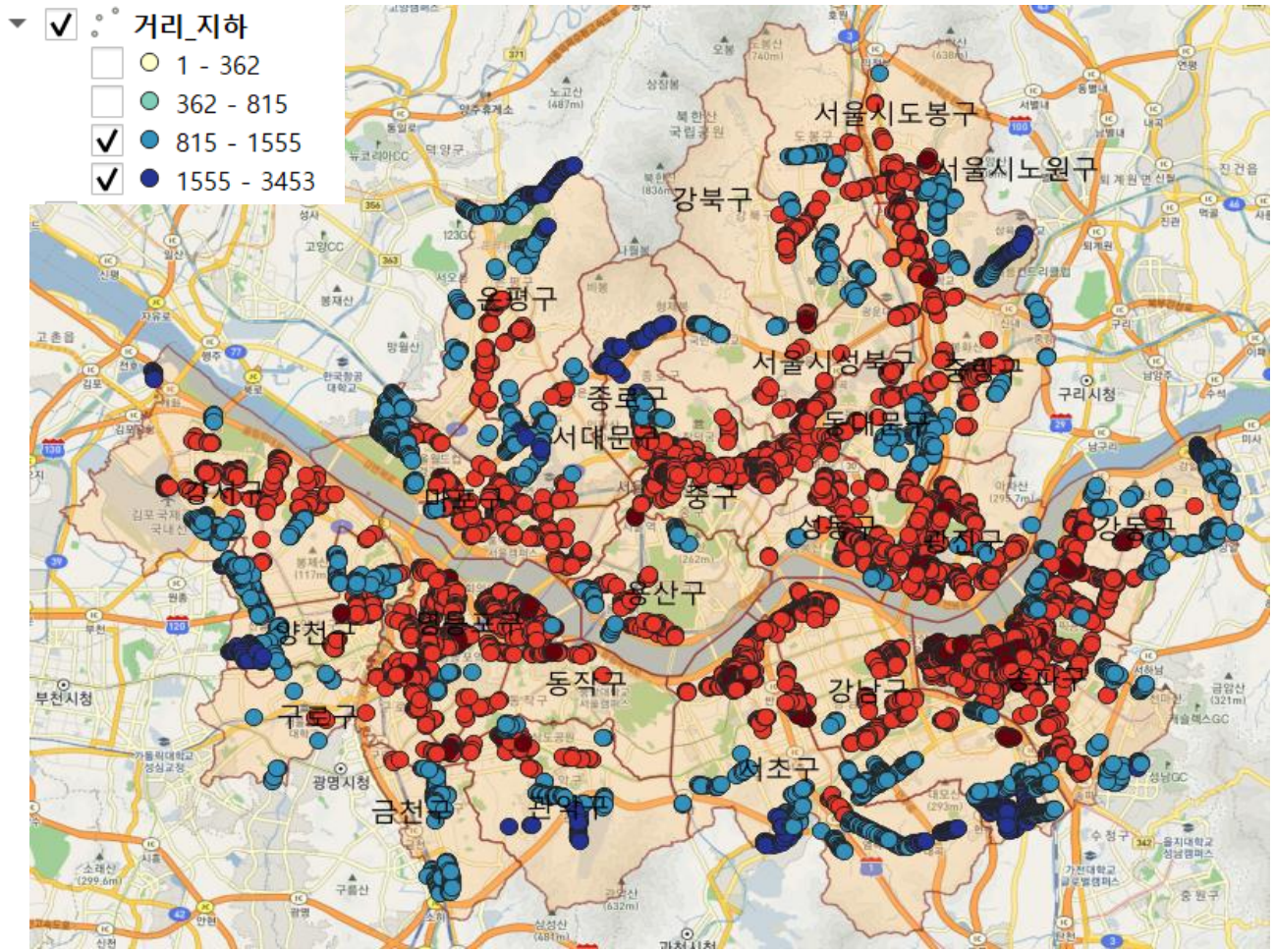
따릉이 이용량과 높은 상관관계를 가지는 변수들

amount	유동20대	유동30대	유동50대	유동40대
1.00000000	0.41708721	0.38812334	0.37308166	0.36877411
유동60대	유동70대이	유동10대	far	지하철_승
0.34501763	0.34366766	0.32403085	0.29507692	0.28672577
유동10대미	버스_승객	medium	버스_경유	주거_20대
0.28639016	0.28396811	0.18496781	0.13300510	0.10079769
거리_초중	주거_30대	주거_40대	주거_10대	주거_50대
0.09917987	0.09211705	0.06324440	0.05990703	0.05418305
주거_60대_	near	면적_아파	면적_주거_	거리_문화
0.04421036	0.02382641	0.01561772	0.01346000	-0.01380638
거리_체육	거리_교통	거리_대학	거리_공원	거리_주차
-0.03032593	-0.07284303	-0.07382987	-0.07729609	-0.09315773
거리_의료	거리_하천	거리_공공	거리_상업	거리_자전
-0.12596766	-0.14036322	-0.14185619	-0.17442092	-0.23611132
거리_특화	거리_관광	거리_지하	거리_영화	평균_경사
-0.23888320	-0.24874536	-0.32520428	-0.33268498	-0.48534279

따릉이 이용량과 유동 20대의 correlation : 0.417

-> 유동 20대의 인구수가 많을수록 따릉이 이용량이 높은 것을 알 수 있다.

예측 총 이용량 vs 거리_지하



따릉이 이용량과 높은 상관관계를 가지는 변수들

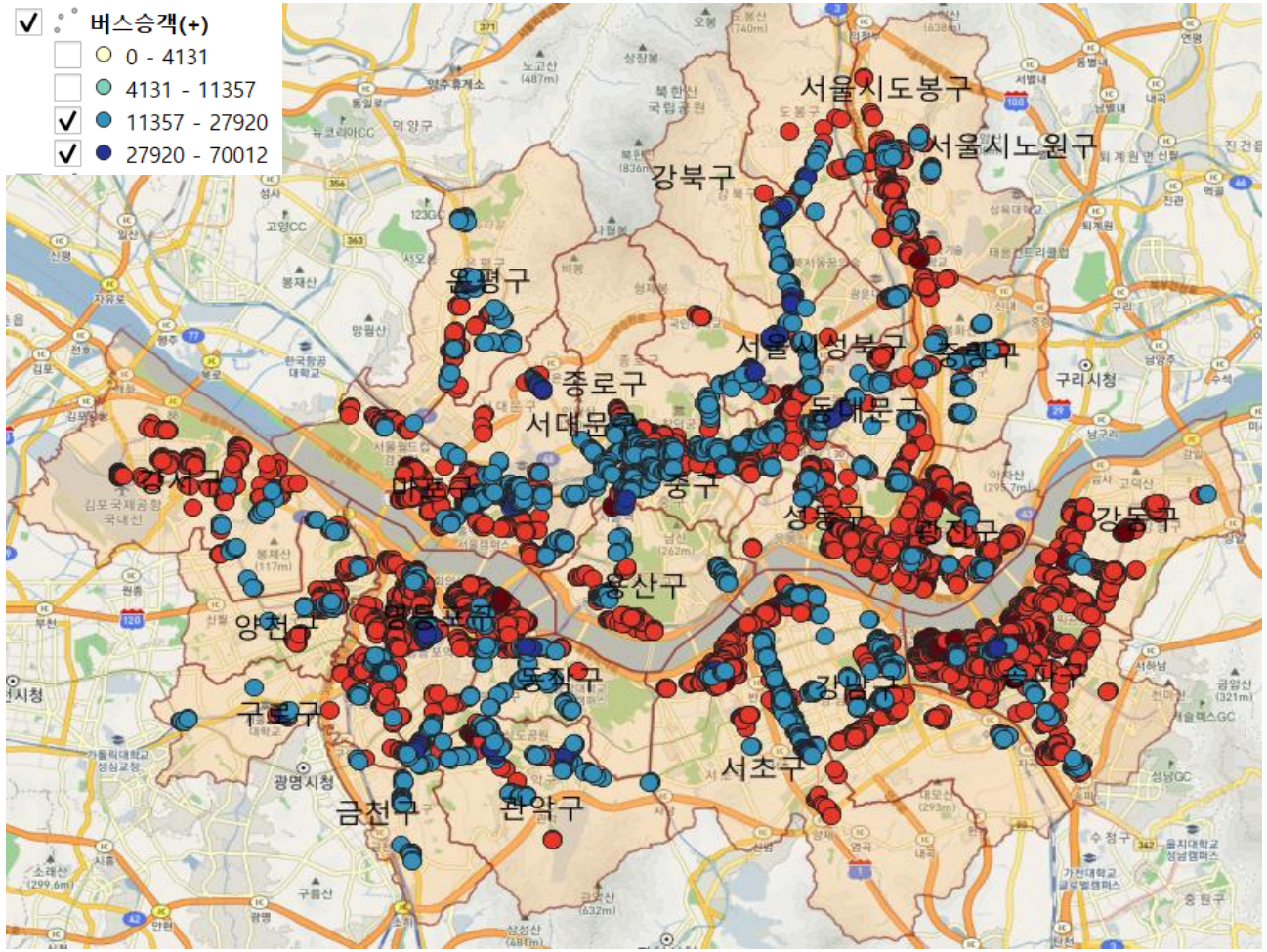
amount	유동20대	유동30대	유동50대	유동40대
1.00000000	0.41708721	0.38812334	0.37308166	0.36877411
유동60대	유동70대이	유동10대	far	지하철_승
0.34501763	0.34366766	0.32403085	0.29507692	0.28672577
유동10대미	버스_승객	medium	버스_경유	주거_20대
0.28639016	0.28396811	0.18496781	0.13300510	0.10079769
거리_초중	주거_30대	주거_40대	주거_10대	주거_50대
0.09917987	0.09211705	0.06324440	0.05990703	0.05418305
주거_60대_	near	면적_아파	면적_주거_	거리_문화
0.04421036	0.02382641	0.01561772	0.01346000	-0.01380638
거리_체육	거리_교통	거리_대학	거리_공원	거리_주차
-0.03032593	-0.07284303	-0.07382987	-0.07729609	-0.09315773
거리_의료	거리_하천	거리_공공	거리_상업	거리_자전
-0.12596766	-0.14036322	-0.14185619	-0.17442092	-0.23611132
거리_특화	거리_관광	거리_지하	거리_영화	평균_경사
-0.23888320	-0.24874536	-0.32520428	-0.33268498	-0.48534279

거리_지하(m) : 대여소로부터 가장 가까운 지하철 역과의 거리

따릉이 이용량과 거리_지하의 correlation : - 0.325

-> 대여소로부터 가장 가까운 지하철 역과의 거리가 가까울수록 따릉이 이용량이 높은 것을 알 수 있다.

예측 총 이용량 vs 버스 승객



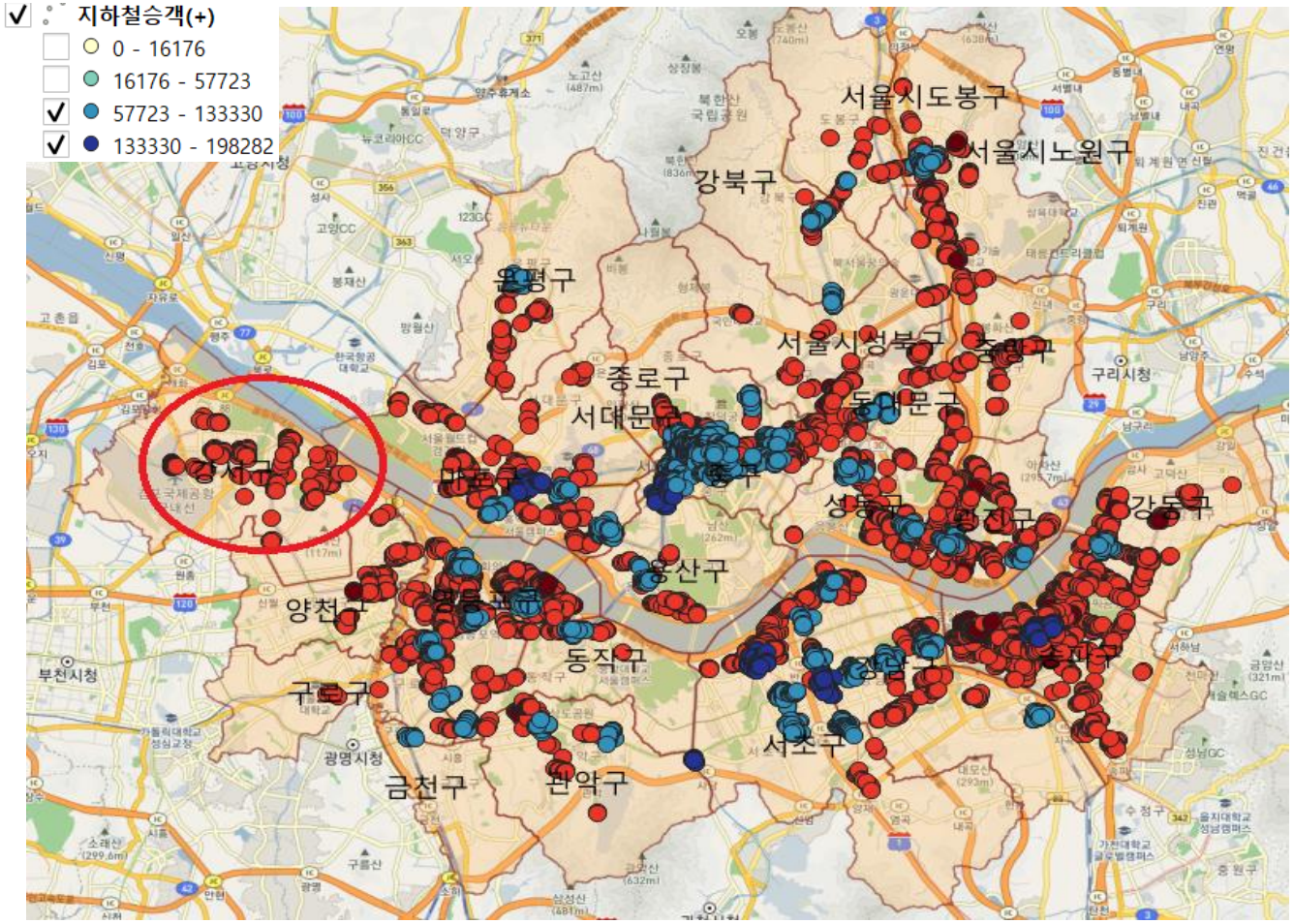
따릉이 이용량과 높은 상관관계를 가지는 변수들

amount	유동20대	유동30대	유동50대	유동40대
1.00000000	0.41708721	0.38812334	0.37308166	0.36877411
유동60대	유동70대이	유동10대	far	지하철_승
0.34501763	0.34366766	0.32403085	0.29507692	0.28672577
유동10대미	버스_승객	medium	버스_경유	주거_20대
0.28639016	0.28396811	0.18496781	0.13300510	0.10079769
거리_초중	주거_30대	주거_40대	주거_10대	주거_50대
0.09917987	0.09211705	0.06324440	0.05990703	0.05418305
주거_60대_	near	면적_아파	면적_주거_	거리_문화
0.04421036	0.02382641	0.01561772	0.01346000	-0.01380638
거리_체육	거리_교통	거리_대학	거리_공원	거리_주차
-0.03032593	-0.07284303	-0.07382987	-0.07729609	-0.09315773
거리_의료	거리_하천	거리_공공	거리_상업	거리_자전
-0.12596766	-0.14036322	-0.14185619	-0.17442092	-0.23611132
거리_특화	거리_관광	거리_지하	거리_영화	평균_경사
-0.23888320	-0.24874536	-0.32520428	-0.33268498	-0.48534279

따릉이 이용량과 버스 승객 수의 correlation : 0.284

-> 버스 승객 수가 많을수록 따릉이 이용량이 높은 것을 알 수 있다.

예측 총 이용량 vs 지하철 승객



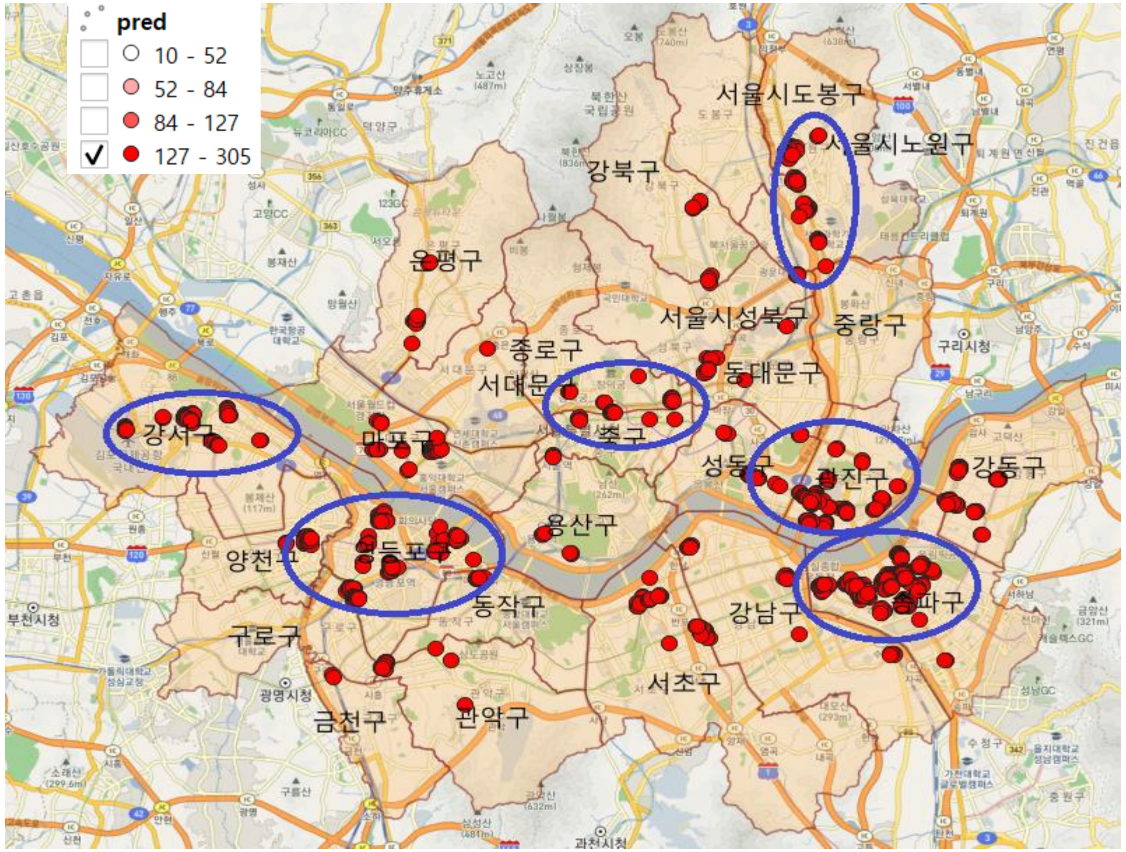
따릉이 이용량과 높은 상관관계를 가지는 변수들

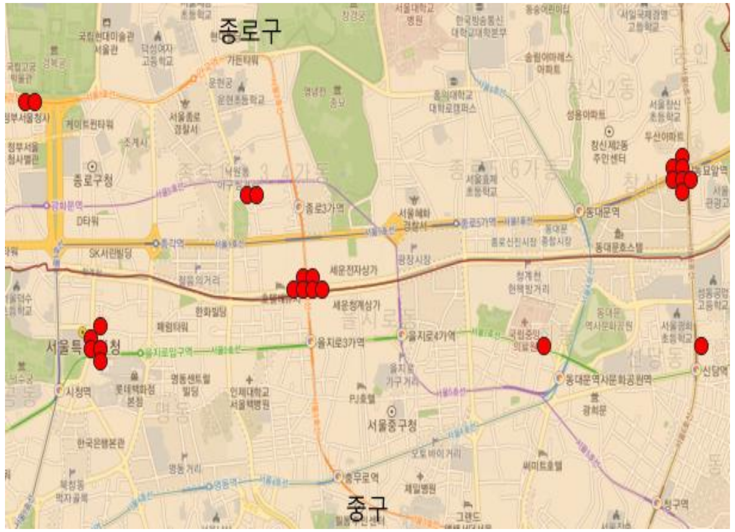
amount	유동20대	유동30대	유동50대	유동40대
1.00000000	0.41708721	0.38812334	0.37308166	0.36877411
유동60대	유동70대이	유동10대	far	지하철_승
0.34501763	0.34366766	0.32403085	0.29507692	0.28672577
유동10대미	버스_승객	medium	버스_경유	주거_20대
0.28639016	0.28396811	0.18496781	0.13300510	0.10079769
거리_초중	주거_30대	주거_40대	주거_10대	주거_50대
0.09917987	0.09211705	0.06324440	0.05990703	0.05418305
주거_60대_	near	면적_아파	면적_주거_	거리_문화
0.04421036	0.02382641	0.01561772	0.01346000	-0.01380638
거리_체육	거리_교통	거리_대학	거리_공원	거리_주차
-0.03032593	-0.07284303	-0.07382987	-0.07729609	-0.09315773
거리_의료	거리_하천	거리_공공	거리_상업	거리_자전
-0.12596766	-0.14036322	-0.14185619	-0.17442092	-0.23611132
거리_특화	거리_관광	거리_지하	거리_영화	평균_경사
-0.23888320	-0.24874536	-0.32520428	-0.33268498	-0.48534279

따릉이 이용량과 지하철 승객 수의 correlation : 0.287

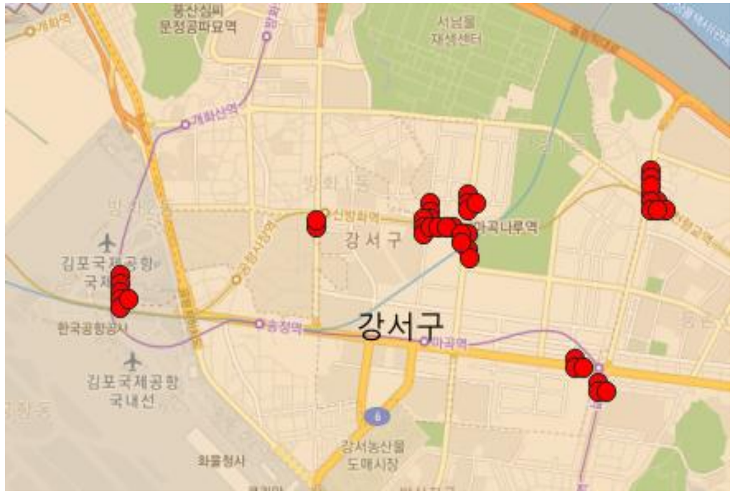
-> 지하철 승객 수가 많을수록 따릉이 이용량이 높은 것을 알 수 있다.

신규대여소 제안

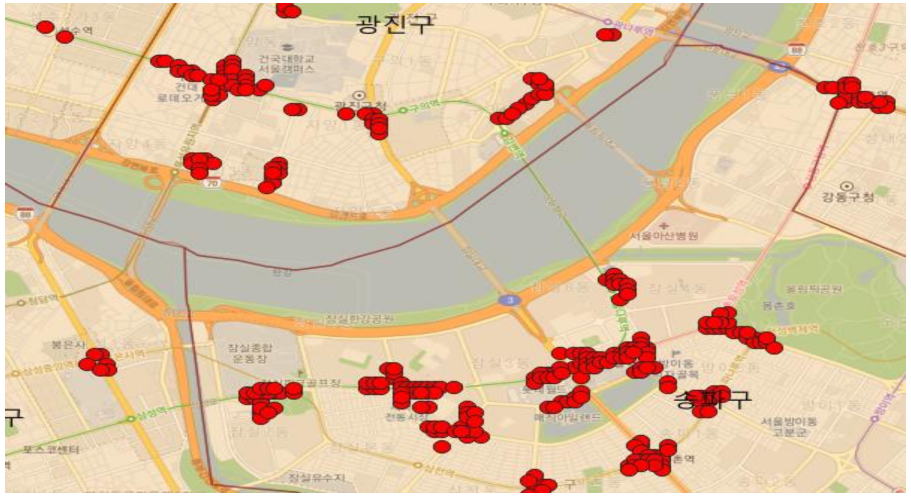

pred
10 - 52
52 - 84
84 - 127
127 - 305
서울시도봉구
서울시노원구
강북구
은평구
서울시성북구
중랑구
종로구
서대문구
동대문구
중구
강서구
마포구
성동구
광진구
강동구
양천구
영등포구
용산구
동작구
송파구
강남구
구로구
서초구
금천구
관악구
부천시청
구리시청
광명시청


종로구
중구


상계6동7동 주민센터
신상중학교
미도아파트
중계무지개 아파트
건영백화점


강서구
김포국제공항 국내선


양천구
영등포구
국회의사당
여의도 한강공원
동작구청


광진구
송파구
강동구청

감사합니다.