서울특별시 공공 자전거 따릉이 데이터 분석

김부겸

목 차

따릉이 데이터 분석

1. 개요

가. 분석배경

1) 따릉이 이용현황 증가

- 2010년 도입 이후 2024년 5월까지 약 1억 9000만건의 이용 횟수 기록
- 2019년과 2023년 이용현황을 비교해보면 주중 이용 건수는 2.5배(1300만건→3300만건), 주말은 2.1배(500만건→1100만건)으로 늘음
- 즉, 따릉이가 단순히 취미수단이 아닌 일상 속의 교통수단으로 정착

2) 전국 자전거 공급 부족

- 아직 전국적으로 공공자전거가 활성화 되어있지 않음
- 수요 대비 공급이 부족한 지역이 있음

<전국 공공자전거 지도>

*2023년 12월 기준

나. 목적

1) 가장 많이 이용되고있는 서울 공공자전거를 기반으로 다른 지역에 공공자전거를 설치하기에 좋은지 알아보려고함

2. 분석 과정

가. 데이터 수집

1) 데이터: 서울 열린데이터 광장
2) 기간: 2023년 12월
3) 구성

- 서울시 공공자전거 대여소 정보
- 서울시 공공자전거 이용 정보_시간대별
- 서울시 공공자전거 이용 정보_월별
- 서울교통공사 역주소 및 전화번호

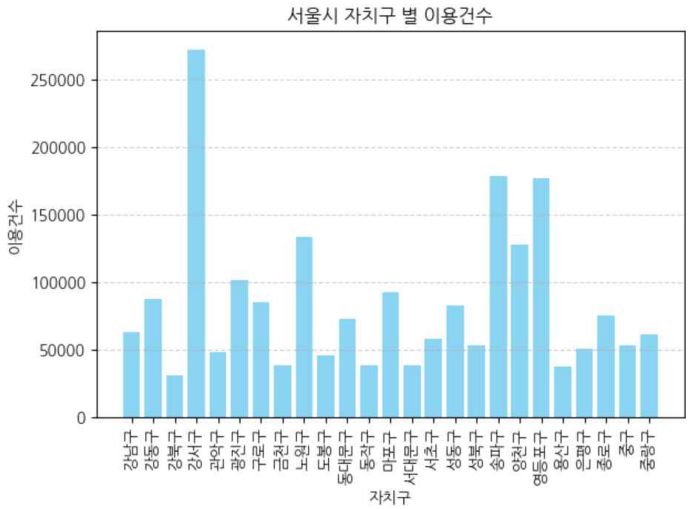
- 서울시 대학 및 전문대학 DB 정보
- 서울시 고등학교 기본정보
- 서울시 인구밀도 (구별) 통계

나. 데이터 전처리

1) 2023년도 12월 데이터만 사용
2) 결측값 제거
3) 필요한 컬럼 추출
4) 데이터 병합

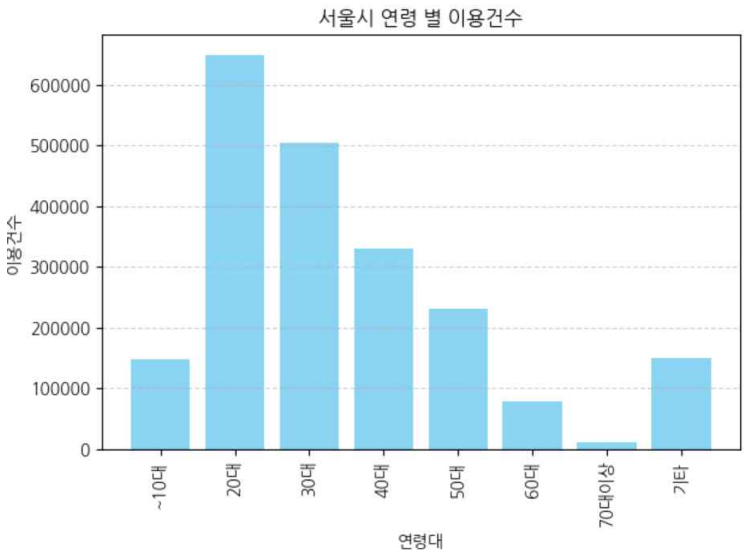
다. 데이터 탐색

1) 자치구별 이용건수



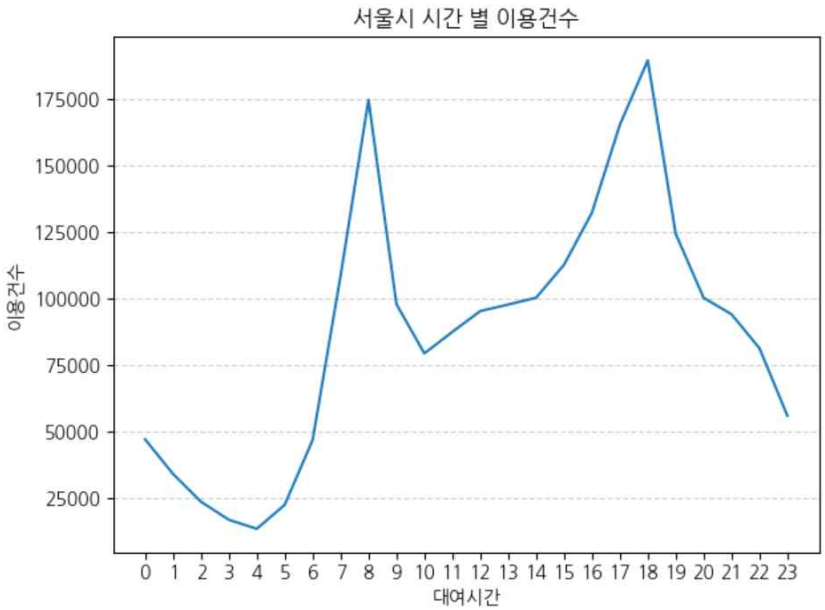
- 강서구의 이용량이 가장 많은 것을 확인할 수 있음

2) 연령 별 이용건수



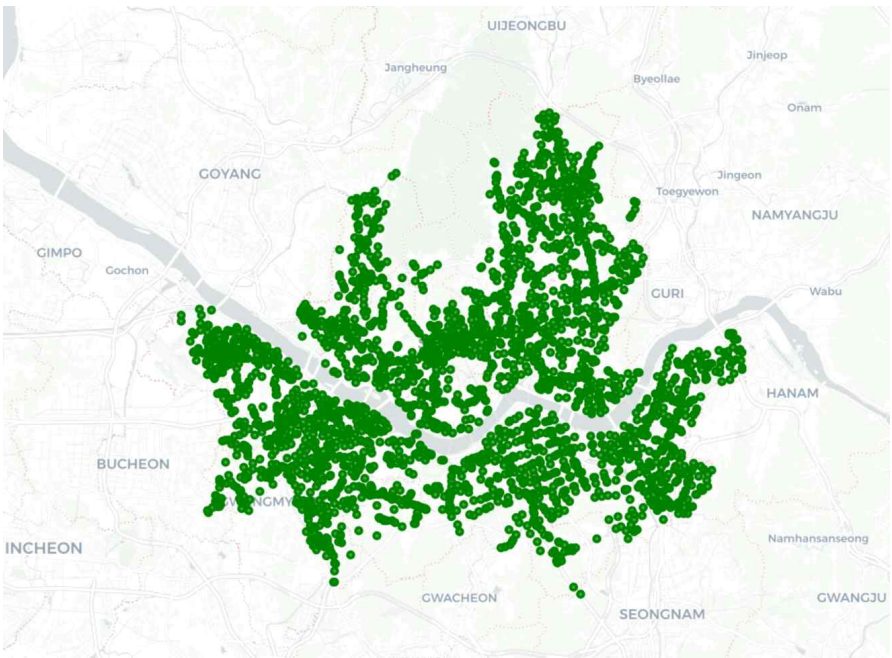
- 20~30대의 이용량이 많은 것을 확인할 수 있음

3) 시간 별 이용 건수

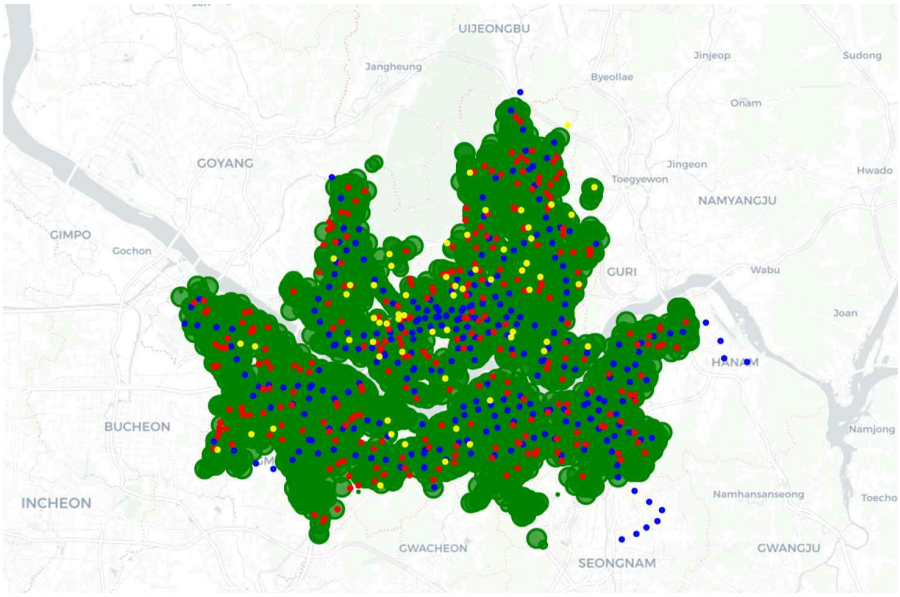


- 8시와 18시의 값이 가장 큰 것을 보아 출퇴근 시간대에 많이 이용하는 것을 알 수 있음

4) 따릉이 대여소 지도 시각화



- 앞선 3가지 시각화를 통해 학교, 지하철역 주변에 대여소가 있을거라 판단함



- 파란색 점(지하철역), 빨간색 점(고등학교), 노란색 점(대학교) 표시
- 지하철 역, 대학교와 고등학교 주변에 대여소가 있는 것으로 보임

라. 데이터 분석

1) 랜덤포레스트 분류

- 랜덤포레스트가 조정을 안해도 성능이 좋게 나오기에 랜덤 포레스트 분류 사용
- 위도, 경도를 이용해 지하철역 인근에 대여소가 있는지 분류

	precision	recall	f1-score
No Rental Shop	0.50	0.20	0.29
Rental Shop	0.93	0.98	0.95
accuracy			0.91
macro avg	0.71	0.59	0.62
weighted avg	0.89	0.91	0.90

- Rental Shop의 정밀도가 0.93으로 잘 예측됨
- 하지만, No Rental Shop의 경우 데이터의 수가 적어 잘 예측하지 못함
- 데이터의 수가 적다는 것은 대부분의 지하철 역 주변에 대여소가 있기 때문
- 이를 통해 지하철역 근처에는 대여소가 있다는 것을 알 수 있음

2) 다중 선형 회귀분석

- 종속변수를 대여소수로 하고, 각 독립변수(이용건수,지하철역 수, 대학교 수, 고등학교 수, 인구밀도)가 영향을 미치는지 확인해보기
- 단계적 회귀법을 이용함

OLS Regression Results

Dep. Variable:	대여소수	R-squared (uncentered):	0.863
Model:	OLS	Adj. R-squared (uncentered):	0.858
Method:	Least Squares	F-statistic:	151.7
Date:	Sun, 01 Sep 2024	Prob (F-statistic):	7.29e-12
Time:	09:50:21	Log-Likelihood:	-129.72
No. Observations:	25	AIC:	261.4
Df Residuals:	24	BIC:	262.7
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
이용건수	0.0011	8.79e-05	12.316	0.000	0.001	0.001

Omnibus:	9.548	Durbin-Watson:	1.450
Prob(Omnibus):	0.008	Jarque-Bera (JB):	9.811
Skew:	-0.798	Prob(JB):	0.00741
Kurtosis:	5.621	Cond. No.	1.00

- 종속변수: 대여소수, 독립변수: 이용건수
- 변수가 유의하고 R-squared 값도 0.863으로 잘 나옴

OLS Regression Results

Dep. Variable:	대여소수	R-squared (uncentered):	0.948
Model:	OLS	Adj. R-squared (uncentered):	0.944
Method:	Least Squares	F-statistic:	209.9
Date:	Sun, 01 Sep 2024	Prob (F-statistic):	1.69e-15
Time:	09:48:55	Log-Likelihood:	-117.63
No. Observations:	25	AIC:	239.3
Df Residuals:	23	BIC:	241.7
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
이용건수	0.0006	9.12e-05	7.003	0.000	0.000	0.001
지하철역수	4.4532	0.727	6.124	0.000	2.949	5.958

Omnibus:	0.282	Durbin-Watson:	2.060
Prob(Omnibus):	0.868	Jarque-Bera (JB):	0.464
Skew:	-0.120	Prob(JB):	0.793
Kurtosis:	2.377	Cond. No.	1.31e+04

- 종속변수: 대여소수, 독립변수: 이용건수, 지하철역수
- 모든 변수 유의하며, R-squared 값이 0.9444 나옴

OLS Regression Results

Dep. Variable:	대여소수	R-squared:	0.766
Model:	OLS	Adj. R-squared:	0.732
Method:	Least Squares	F-statistic:	22.90
Date:	Sun, 01 Sep 2024	Prob (F-statistic):	8.02e-07
Time:	09:34:56	Log-Likelihood:	-109.30
No. Observations:	25	AIC:	226.6
Df Residuals:	21	BIC:	231.5
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	45.7476	10.980	4.166	0.000	22.914	68.582
이용건수	0.0005	7.9e-05	5.860	0.000	0.000	0.001
지하철역수	2.6343	0.682	3.861	0.001	1.215	4.053
대학수	-1.1794	1.694	-0.696	0.494	-4.701	2.343

Omnibus:	4.148	Durbin-Watson:	1.854
Prob(Omnibus):	0.126	Jarque-Bera (JB):	2.407
Skew:	0.504	Prob(JB):	0.300
Kurtosis:	4.139	Cond. No.	2.65e+05

- 종속변수: 대여소수, 독립변수: 이용건수, 지하철역수, 대학수

- 대학수 변수가 유의하지 않아 제거하기로함

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                 대여소수   R-squared:                       0.798
Model:                            OLS   Adj. R-squared:                  0.769
Method:                 Least Squares   F-statistic:                     27.63
Date:                Sun, 01 Sep 2024   Prob (F-statistic):           1.75e-07
Time:                        09:34:56   Log-Likelihood:                -107.47
No. Observations:                  25   AIC:                             222.9
Df Residuals:                      21   BIC:                             227.8
Df Model:                           3
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         30.2063     10.661      2.833      0.010       8.036      52.377
이용건수          0.0004   8.33e-05      4.690      0.000       0.000       0.001
지하철역수         2.2598      0.664      3.404      0.003       0.879       3.640
고등학교수         1.7851      0.906      1.970      0.062      -0.099       3.669
==============================================================================
Omnibus:                        3.977   Durbin-Watson:                   1.779
Prob(Omnibus):                  0.137   Jarque-Bera (JB):                2.315
Skew:                           0.679   Prob(JB):                        0.314
Kurtosis:                       3.613   Cond. No.                     2.77e+05
==============================================================================
```

- 고등학교수의 변수도 유의하지 않아 제거하기로함

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                 대여소수   R-squared:                       0.834
Model:                            OLS   Adj. R-squared:                  0.811
Method:                 Least Squares   F-statistic:                     35.23
Date:                Sun, 01 Sep 2024   Prob (F-statistic):           2.22e-08
Time:                        09:35:14   Log-Likelihood:                -104.99
No. Observations:                  25   AIC:                             218.0
Df Residuals:                      21   BIC:                             222.9
Df Model:                           3
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         13.8095     12.199      1.132      0.270     -11.559      39.178
이용건수          0.0004   7.65e-05      4.606      0.000       0.000       0.001
지하철역수         2.4661      0.577      4.274      0.000       1.266       3.666
인구수           0.0001    3.4e-05      3.057      0.006    3.32e-05       0.000
==============================================================================
Omnibus:                        2.322   Durbin-Watson:                   1.881
Prob(Omnibus):                  0.313   Jarque-Bera (JB):                1.350
Skew:                           0.564   Prob(JB):                        0.509
Kurtosis:                       3.158   Cond. No.                     1.44e+06
==============================================================================
```

- 모든 변수가 유의하고, R-squared값도 0.834로 모형을 잘 설명함

3. 결론

가. 분석 결과

- 다중 선형 회귀 분석을 통해 대여소 수에 영향을 미치는 변수는 이용건수, 지하철역 수와 인구 수 라는 것을 확인할 수 있음
- 학교 주변과 대여소가 연관이 있을 것이라 판단했지만 유의하지 않을 것을 보아 영향을 미치지 않음
- 결론적으로 이용건수, 지하철역수, 인구수가 많은 지역에 대여소의 수가 많다.

나. 활용 방안

- 공공 자전거가 없는 지역에서 이를 지표로 사용하여 공공 자전거 대여소를 설치할 수 있음
- 이용건수의 경우, 20-30대와 출퇴근 시간대에 이용을 많이하는 것을 확인하여 이용건수가 많을 위치를 판단할 수 있음
- 추가적인 데이터를 수집하여 더 많은 인사이트를 도출해 활용할 수 있을 것이라 예상함