

DATASET

BIKE SHARING

This dataset contains the hourly and daily count of rental bikes between years 2011 and 2012 in Capital bikeshare system with the corresponding weather and seasonal information.

[HTTPS://ARCHIVE.ICS.UCI.EDU/ML/DATASETS/B
IKE+SHARING+DATASET](https://archive.ics.uci.edu/ml/datasets/Bike+Sharing+Dataset)



PROJECT

프로젝트 배경

1. 자전거 공유 프로그램으로 친환경 도시 조성

Bike Share Programs in the U.S. Foster Greener Cities

MAY 19, 2023 11:21 AM

As biking continues to gain popularity in the United States, several programs across the country make it easier for people to hop on and head to their destination.

BY KRISTI EATON

GOOD RESOURCES

ENVIRONMENT

2. 통근자들의 발빠른 대중교통으로

CityLab | Transportation

In Times of Crisis, Bikeshare Rolls On

When transit systems stumble, municipal bikesharing services can keep commuters moving — if cities know how to plan properly.

"일상의 문화가 된 공유 자전거 시스템"



ANALYTICS AND VISUALIZATION PROJECTS..

PROJECT

프로젝트 정의 및 목적



"공유 자전거 대여 횟수에 영향을 주는 요인분석"



자전거 대여 횟수에 영향을 주는 요인을 식별하고, 이를 통해 인사이트를 도출



PROJECT

수집 과정과 결과





Machine Learning Repository
머신러닝 및 지능시스템 연구단

현재 테스트 중인 새로운 UCI Machine Learning Repository의 베타 버전을 확인하십시오

자전거 공유 데이터 세트 데이터 세트
다운로드: [데이터 폴더](#), [데이터 세트 설명](#)

<https://archive.ics.uci.edu/ml/datasets/bike+sharing+dataset>



PROJECT

데이터들의 구조, 관계 및 특징

< 범주형 데이터 >

#	Column	Non-Null Count	Dtype
---	-----	-----	-----
0	instant	17379 non-null	int64
1	dteday	17379 non-null	object
2	season	17379 non-null	int64
3	yr	17379 non-null	int64
4	mnth	17379 non-null	int64
5	hr	17379 non-null	int64
6	holiday	17379 non-null	int64
7	weekday	17379 non-null	int64
8	workingday	17379 non-null	int64
9	weathersit	17379 non-null	int64

< 연속형 데이터 >

10	temp	17379 non-null	float64
11	atemp	17379 non-null	float64
12	hum	17379 non-null	float64
13	windspeed	17379 non-null	float64
14	casual	17379 non-null	int64
15	registered	17379 non-null	int64
16	cnt	17379 non-null	int64

종속변수 : cnt (자전거 대여횟수)



PROJECT

데이터들의 구조, 관계 및 특징

< 종속변수와 관련있는 것 >

```
1 hour_df[['casual', 'registered', 'cnt']]
✓ 0.0s
```

	casual	registered	cnt
0	3	13	16
1	8	32	40

<0과 1의 값>

```
1 hour_df['yr'].unique()
✓ 0.0s
array([0, 1], dtype=int64)
```

```
1 hour_df['workingday'].unique()
✓ 0.0s
array([0, 1], dtype=int64)
```

```
1 hour_df['holiday'].unique()
✓ 0.0s
array([0, 1], dtype=int64)
```

<1과 4의 값>

```
1 hour_df['season'].unique()
✓ 0.0s
array([1, 2, 3, 4], dtype=int64)
```

```
1 hour_df['weathersit'].unique()
✓ 0.0s
array([1, 2, 3, 4], dtype=int64)
```



PROJECT

질문 정의

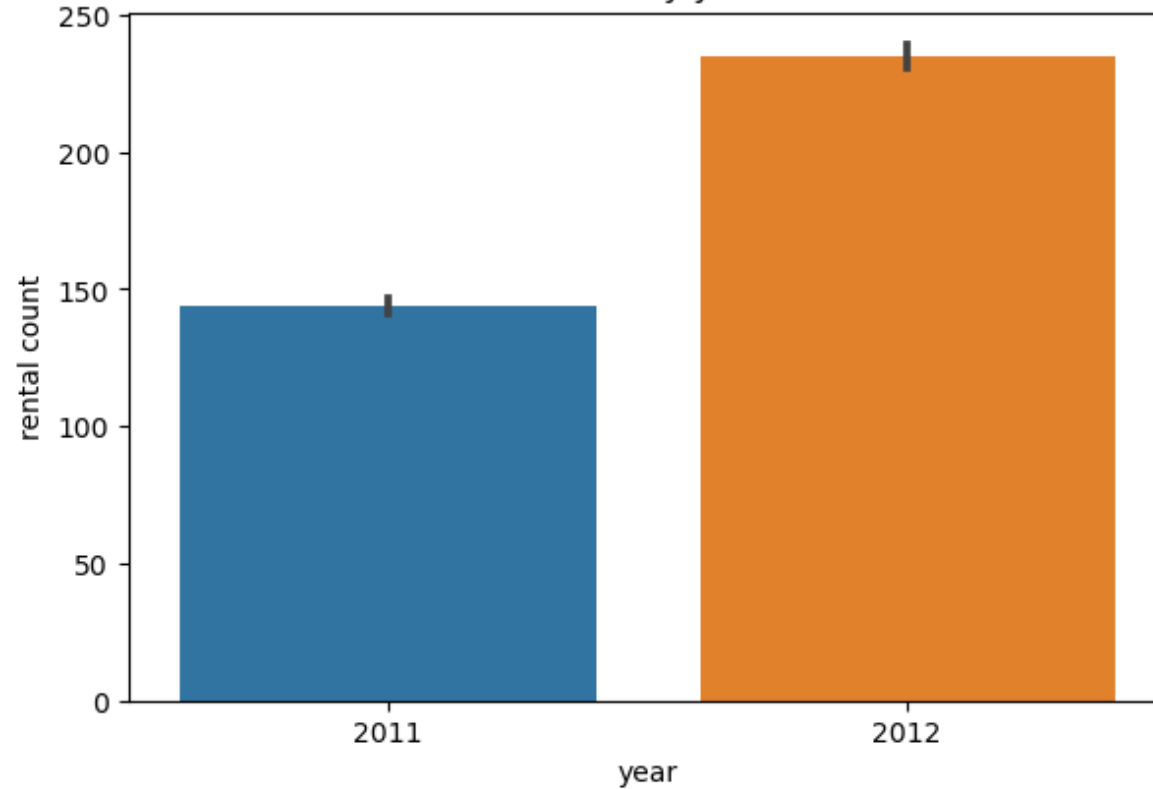
- 자전거 대여횟수가 가장 많은 년, 월, 요일, 시는?
- 휴일과 근무일의 총 자전거 대여 차이는?
- 총 대여 횟수와 시즌 사이에 상관관계가 있을까?
- 자전거 대여 횟수가 기상 조건에 따른 차이가 있을까?
- 온도, 습도, 풍속은 대여횟수에 유의미한 영향을 줄까?



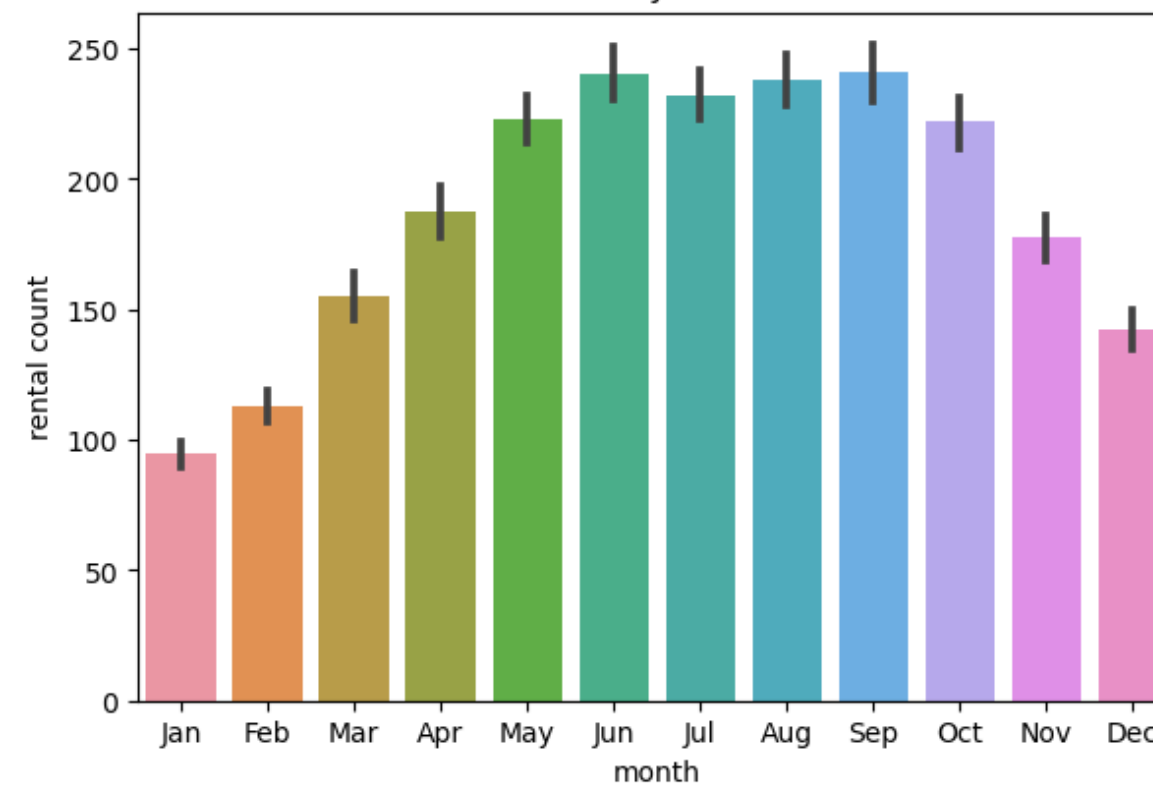
데이터 시각화



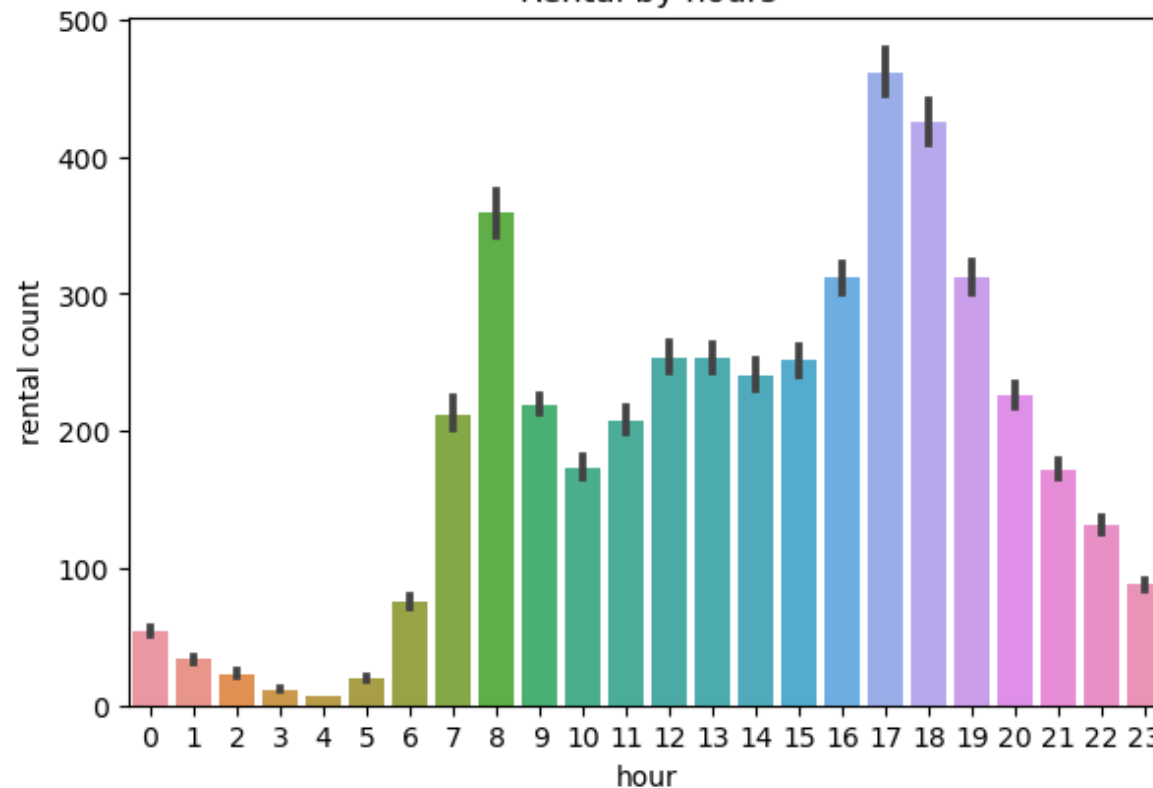
Rental by year



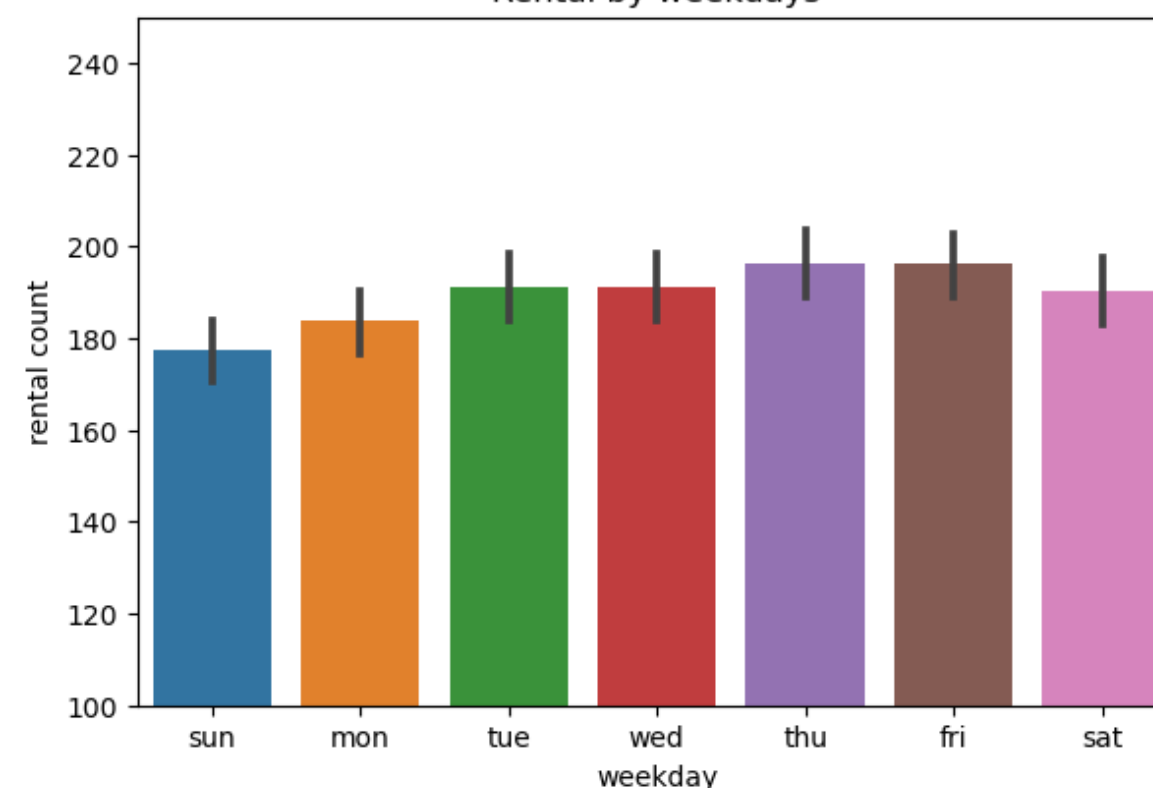
Rental by months



Rental by hours



Rental by weekdays



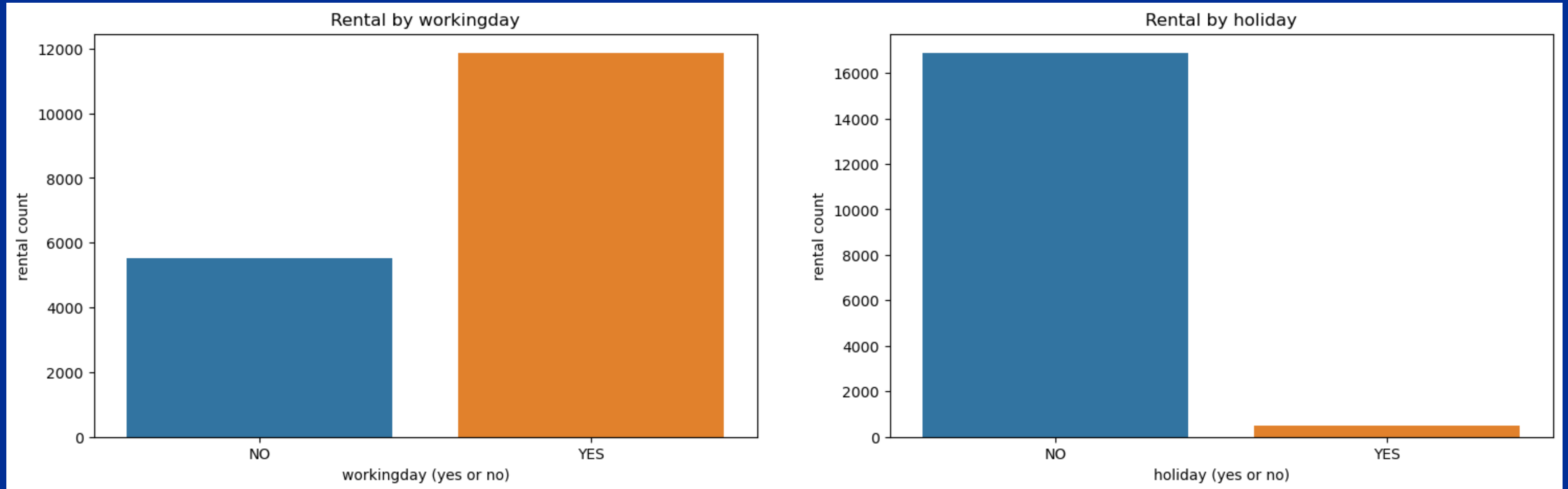
<자전거 대여량이 많은 시기>

6~9월

8시 출근시간

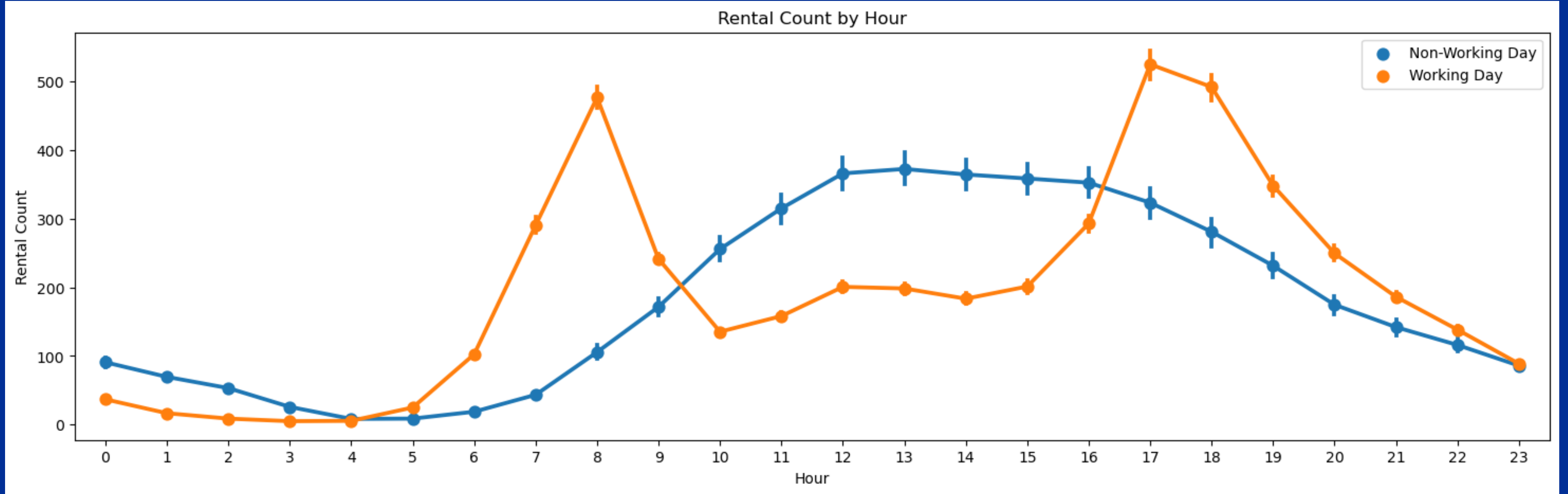
17-18시 퇴근시간

데이터 시각화



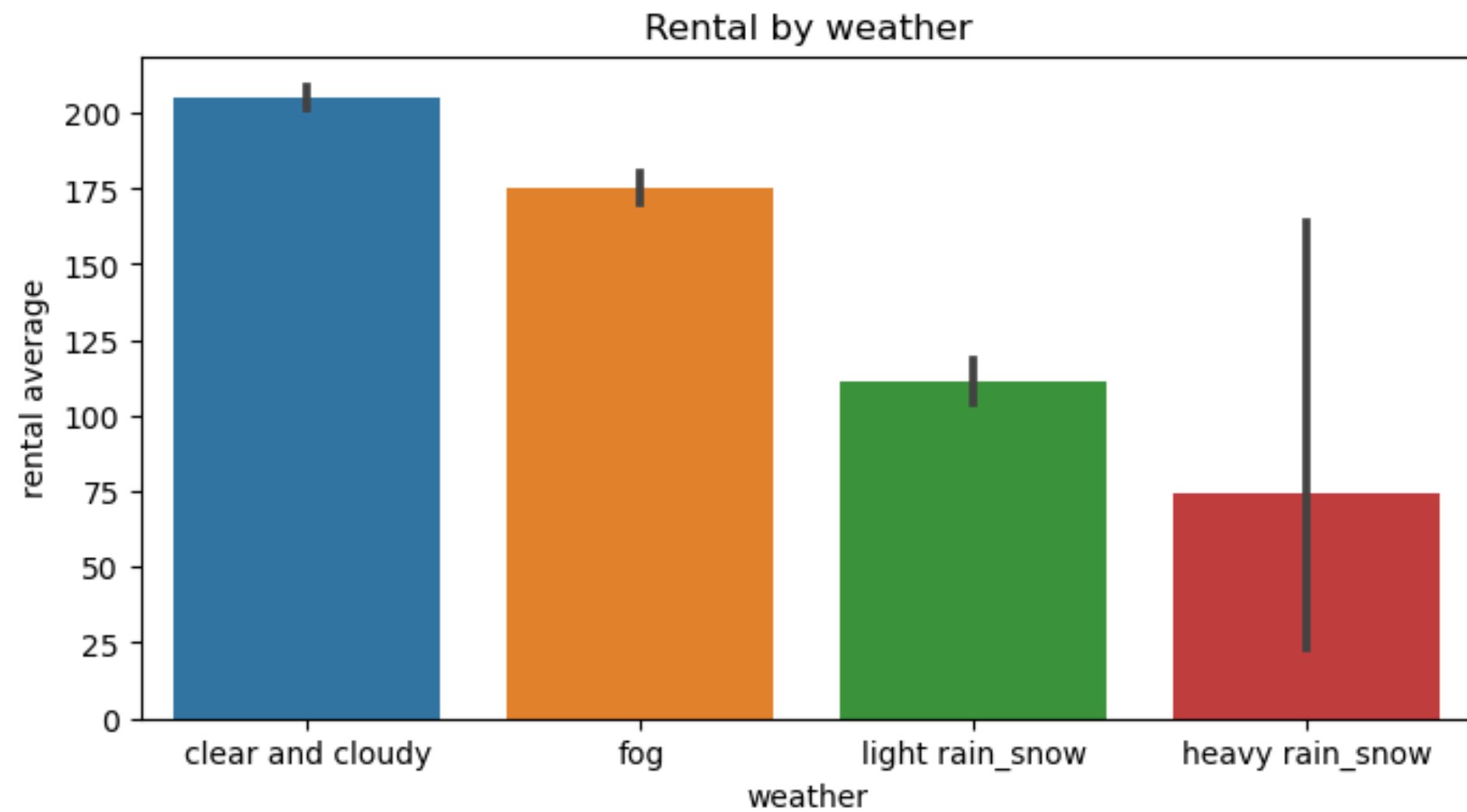
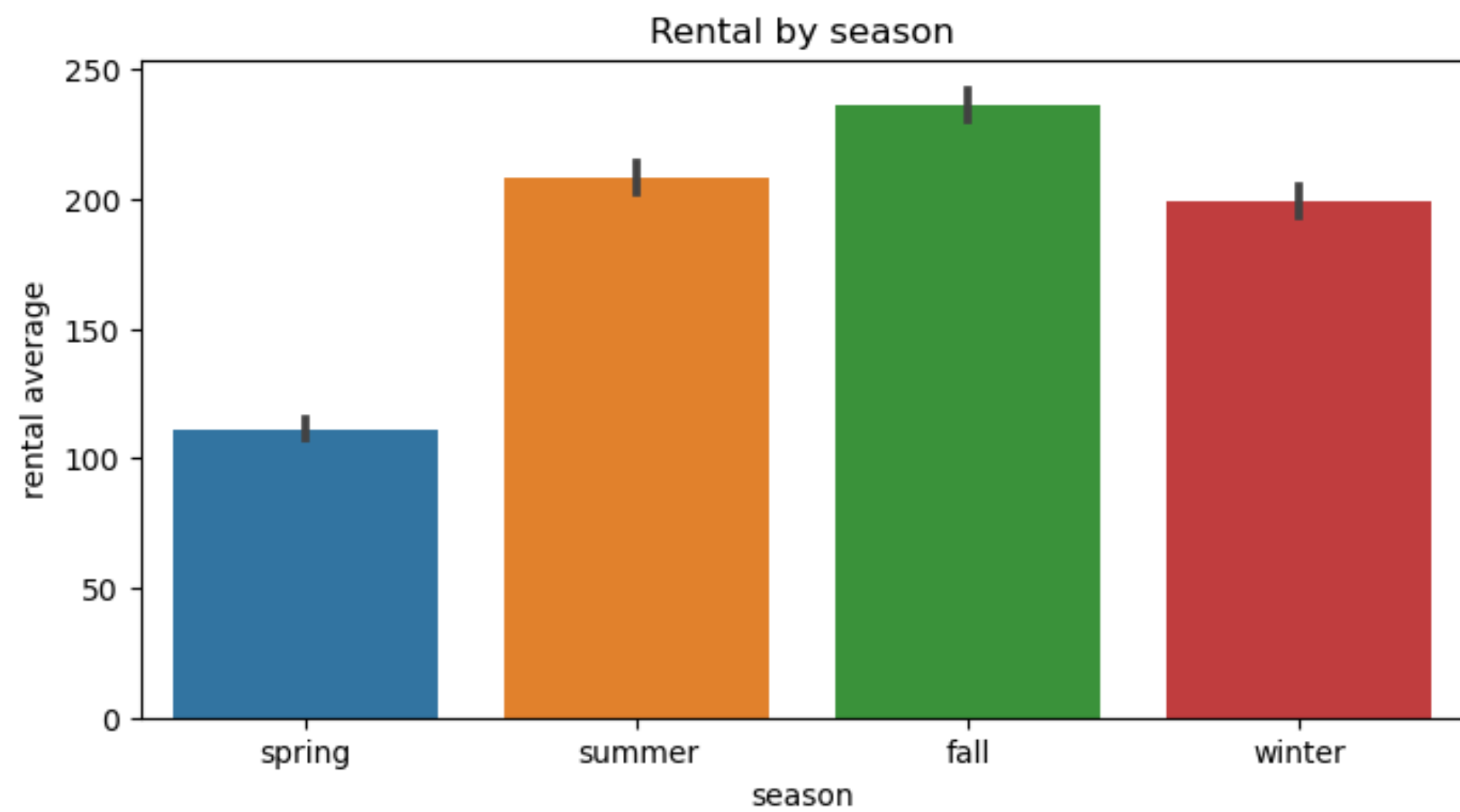
출근일 > 대여건수,
휴일이 아닌 경우 > 대여건수가 더 많은 것을 확인

데이터 시각화



출근여부에 따른 시간대별 대여횟수

데이터 시각화



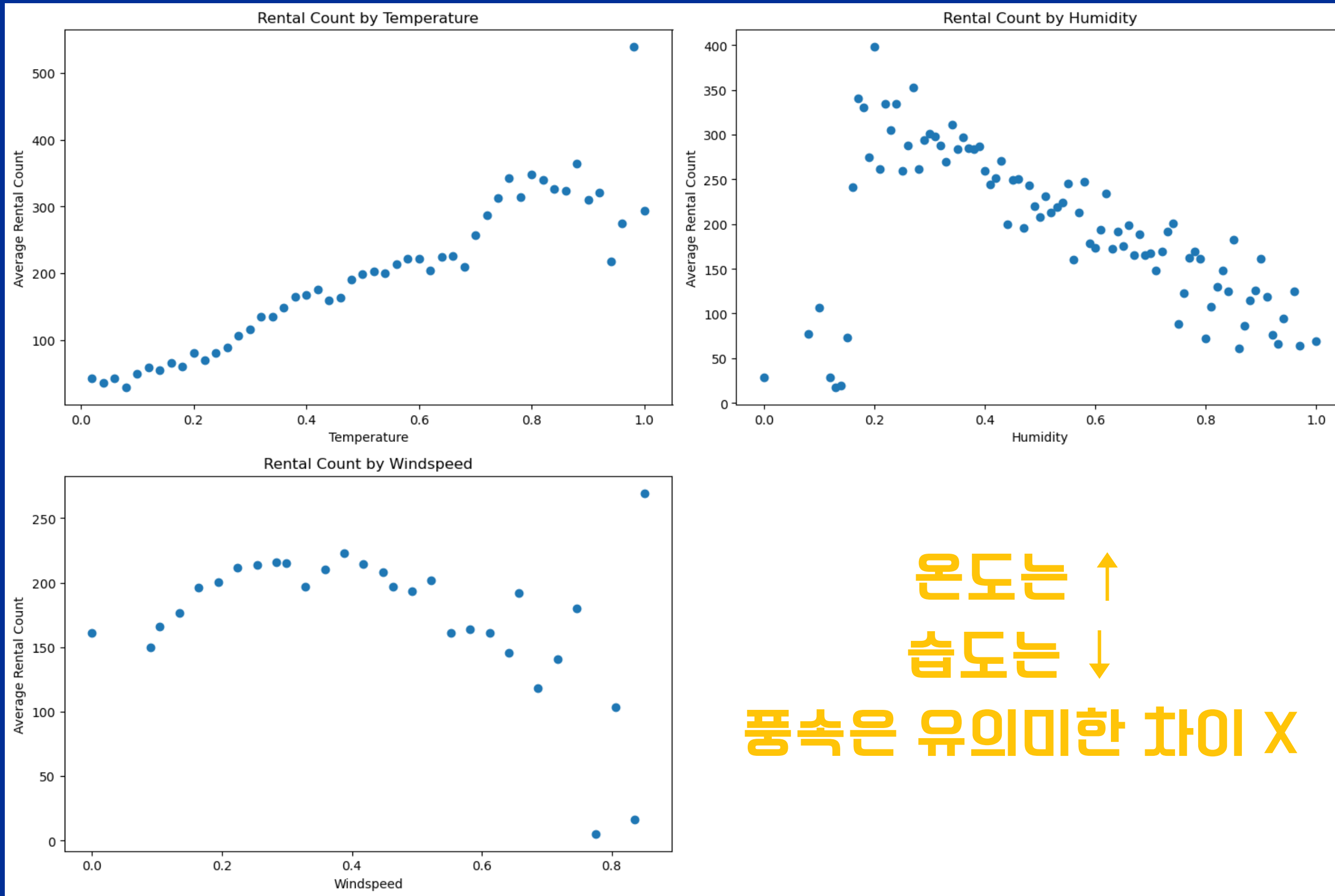
날씨가 좋을 수록 대여건수 많았으며,

시즌의 경우

봄 < 겨울 < 여름 < 가을 순으로

대여건수가 더 많은 것을 확인

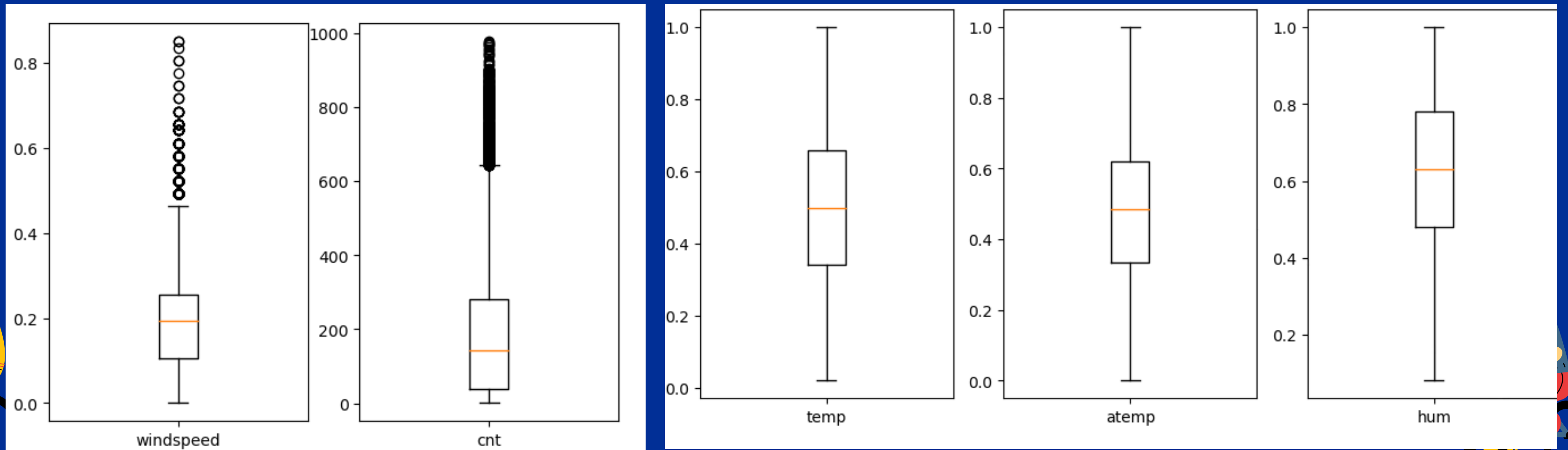
데이터 시각화



PROJECT

검정의 과정과 결과

연속형 변수들의 분포 확인



PROJECT

검정의 과정과 결과

<범주형 변수>

- t-test

독립표본 t-Test : Ttest_indResult(statistic=4.07884548432246, pvalue=4.5461689487652855e-

- ANOVA(일원 분산분석 및 사후검정)

일원 분산분석 : F_onewayResult(statistic=409.1810372630525, pvalue=7.40107139971279e-257)

분산 분석의 사후분석 : Tukey's HSD Pairwise Group Comparisons (95.0% Confidence Interval)

Comparison	Statistic	p-value	Lower CI	Upper CI
(0 - 1)	-18.440	0.189	-40.125	3.245
(0 - 2)	-60.986	0.000	-82.164	-39.808
(0 - 3)	-92.836	0.000	-114.144	-71.528
(0 - 4)	-128.482	0.000	-149.608	-107.357
(0 - 5)	-146.091	0.000	-167.387	-124.794
(0 - 6)	-137.395	0.000	-158.520	-116.270
(0 - 7)	-112.672	0.000	-134.012	-91.332
(1 - 2)	-42.546	0.000	-63.892	-21.200
(1 - 3)	-74.396	0.000	-95.742	-53.050
(1 - 4)	-110.042	0.000	-131.388	-88.696
(1 - 5)	-127.651	0.000	-148.997	-106.305
(1 - 6)	-118.955	0.000	-140.301	-97.609
(1 - 7)	-94.232	0.000	-115.578	-72.886
(2 - 3)	-31.850	0.000	-53.196	-10.504
(2 - 4)	-67.496	0.000	-88.842	-46.150
(2 - 5)	-85.105	0.000	-106.451	-63.759
(2 - 6)	-76.409	0.000	-97.755	-55.063
(2 - 7)	-51.686	0.000	-73.032	-30.340
(3 - 4)	-36.660	0.000	-58.006	-15.314
(3 - 5)	-54.269	0.000	-75.615	-32.923
(3 - 6)	-45.573	0.000	-66.919	-24.227
(3 - 7)	-20.850	0.000	-42.196	-0.504
(4 - 5)	-17.605	0.000	-38.951	3.741
(4 - 6)	-9.909	0.000	-31.255	11.437
(4 - 7)	15.814	0.000	-6.532	38.160
(5 - 6)	24.510	0.000	3.164	45.856
(5 - 7)	49.233	0.000	27.887	70.579
(6 - 7)	24.510	0.000	3.164	45.856



분석 과정과 결과



	Variable	VIF
0	const	58.120092
1	temp	2.811897
2	hr	1.143030
3	hum	1.392318
4	yr	1.013348
5	season_1	2.371754
6	season_3	1.897591
7	weathersit_3	1.158823
8	windspeed	1.147116
9	holiday	1.013304
10	season_4	1.696261
11	weekday	1.012754

OLS Regression Results							
=====							
Dep. Variable:	cnt		R-squared:	0.398			
Model:	OLS		Adj. R-squared:	0.398			
Method:	Least Squares		F-statistic:	1043.			
Date:	Thu, 01 Jun 2023		Prob (F-statistic):	0.00			
Time:	17:17:33		Log-Likelihood:	-1.1063e+05			
No. Observations:	17379		AIC:	2.213e+05			
Df Residuals:	17367		BIC:	2.214e+05			
Df Model:	11						
Covariance Type:	nonrobust						
=====							
		coef	std err	t	P> t	[0.025	0.975]

Intercept		-6.5238	8.142	-0.801	0.423	-22.483	9.435
temp		357.2731	9.301	38.413	0.000	339.043	375.504
hr		7.5446	0.165	45.686	0.000	7.221	7.868
hum		-182.6709	6.532	-27.966	0.000	-195.474	-169.868
yr		79.8142	2.150	37.120	0.000	75.600	84.029
season_1		-19.9251	3.829	-5.204	0.000	-27.430	-12.420
season_3		-29.2757	3.359	-8.714	0.000	-35.861	-22.691
weathersit_3		-32.0547	4.198	-7.635	0.000	-40.284	-23.825
windspeed		28.9810	9.350	3.100	0.002	10.654	47.308
holiday		-27.5346	6.431	-4.281	0.000	-40.141	-14.929
season_4		42.7496	3.241	13.191	0.000	36.397	49.102
weekday		1.8990	0.536	3.544	0.000	0.849	2.949
=====							
Omnibus:		3528.488	Durbin-Watson:	0.561			
Prob(Omnibus):		0.000	Jarque-Bera (JB):	7044.890			
Skew:		1.222	Prob(JB):	0.00			
Kurtosis:		4.939	Cond. No.	158.			
=====							
Notes:							
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.							

1에 가까우므로,
다중공선성의 가능성은 낮다

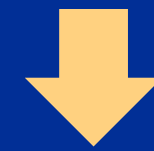
분석 과정과 결과



OLS Regression Results									
=====									
Dep. Variable:	cnt		R-squared:	0.398					
Model:	OLS		Adj. R-squared:	0.398					
Method:	Least Squares		F-statistic:	1043.					
Date:	Thu, 01 Jun 2023		Prob (F-statistic):	0.00					
Time:	17:17:33		Log-Likelihood:	-1.1063e+05					
No. Observations:	17379		AIC:	2.213e+05					
Df Residuals:	17367		BIC:	2.214e+05					
Df Model:	11								
Covariance Type:	nonrobust								
=====									
				coef	std err	t	P> t	[0.025	0.975]

Intercept				-6.5238	8.142	-0.801	0.423	-22.483	9.435
temp				357.2731	9.301	38.413	0.000	339.043	375.504
hr				7.5446	0.165	45.686	0.000	7.221	7.868
hum				-182.6709	6.532	-27.966	0.000	-195.474	-169.868
yr				79.8142	2.150	37.120	0.000	75.600	84.029
season_1				-19.9251	3.829	-5.204	0.000	-27.430	-12.420
season_3				-29.2757	3.359	-8.714	0.000	-35.861	-22.691
weathersit_3				-32.0547	4.198	-7.635	0.000	-40.284	-23.825
windspeed				28.9810	9.350	3.100	0.002	10.654	47.308
holiday				-27.5346	6.431	-4.281	0.000	-40.141	-14.929
season_4				42.7496	3.241	13.191	0.000	36.397	49.102
weekday				1.8990	0.536	3.544	0.000	0.849	2.949
=====									
Omnibus:	3528.488		Durbin-Watson:	0.561					
Prob(Omnibus):	0.000		Jarque-Bera (JB):	7044.890					
Skew:	1.222		Prob(JB):	0.00					
Kurtosis:	4.939		Cond. No.	158.					
=====									
Notes:									
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.									

그러나, ols표
R-squared 0.398
낮은 설명력



기존 변수를 조합,
새로운 변수의 발굴 등

설명력을 높일 수 있는 변수가 필요.

PROJECT

필요성

이투데이 > 사회 > 일반

‘인기 만점’ 따릉이, 누적 이동 거리 3.9억km...“단거리 교통수단 정착”

입력 2023-05-21 11:15

김채빈 기자 chaebi@etoday.co.kr



따릉이 누적 대여 건수 1.4억 건 돌파...지난해 4059만 건
마곡지구·여의도·잠실 등 ‘출퇴근 시간대’ 이용량 높아져

탄소발자국 줄이기
ATFT 2050

SAMSUNG

"꾸준한 수요에 따른 효율적인 운영 및 관리 필요"

[HTTPS://WWW.ETODAY.CO.KR/NEWS/VIEW/2250875.](https://www.etoday.co.kr/news/view/2250875)



PROJECT

향후 계획 및 맺음말



미국의 CAPITAL BIKESHARE에 대한 데이터
한국 따릉이의 표본이 필요



지리적 요소

서로 다른 운영

사용자 특성



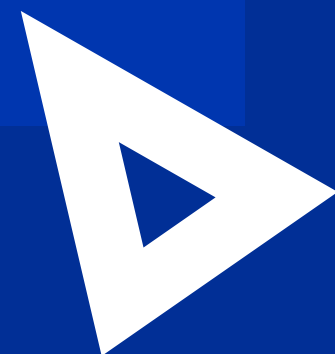
ANALYTICS AND VISUALIZATION PROJECTS..



Q & A



질문 시간입니다.



THANK YOU



감사합니다.

