Practice-HW
DM Process / Variable Selection

논문참조

강세정, 김규현, 신재욱, 장현우

Contents

DM Process /
Variable Selection

01

Research Questions

Introduction to RQ

Data

02

Research Question 1

EDA & VIF

Regression (OLS)

Variable Selection

Shrinkage Comparison & Result

03

Research Question 2

Chi-square test

Result

Research Questions

Contents

DM Process /
Variable Selection

01. Introduction to RQ

(1) RQ1 : 시민들의 여가생활과 문화생활에 미치는 날씨 요인은 무엇일까? → 따릉이, 경복궁 이용객 수의 수요 예측을 시도해보고자 함

(2) RQ2 : 강수 여부가 대기 환경에 어떤 영향을 미칠까? → 강수 여부와 미세먼지, 초미세먼지의 독립 여부를 확인해보고자 함

02. Data

(1) RQ1

X(Numerical Variable)

평균기온 (℃)	최고기온 (℃)	일강수량 (mm)	최대풍속 (m/s)	평균풍속 (m/s)	최소상대 습도(%)	평균상대 습도(%)	가조시간 (hr)
합계일조 시간(hr)	1시간 최다 일사량 (MJ/m2)	합계 일사량 (MJ/m2)	일 최심 적설(cm)	평균전운량 (1/10)	9-9강수 (mm)	안개 계속 시간(hr)	

X(Categorical Variable)

강수여부 (0, 1)	적설여부 (0, 1)	안개여부 (0, 1)	휴일여부 (0, 1)	미세먼지 기준 (1, 2, 3, 4)	초미세먼지 기준 (1, 2, 3, 4)

Y1	Y2
일일 서울시 공공자전거 (따릉이) 대여 건수	일일 경북궁 관람객 수

「대기환경보전법」에 따른 미세먼지, 초미세먼지 기준

물질	단위	산정기준	등급			
			좋음(1)	보통(2)	나쁨(3)	매우나쁨(4)
미세먼지	㎍/㎥	24시간	0~30	31~80	81~150	151~
초미세먼지	㎍/㎥	24시간	0~15	16~35	36~75	76~

(2) RQ2

X(Categorical Variable)		Y
미세먼지 기준 (1, 2, 3, 4)	초미세먼지 기준 (1, 2, 3, 4)	강수여부 (0, 1)

(3) 추가 설명

데이터는 서울 열린 데이터 광장, 공공데이터포털, 기상청 등에서 직접 수집함

데이터 기간 : 2019.01.01 - 2019.12.31

처음에는 2016-2019 데이터를 사용해 프로젝트를 진행했으나, 성능이 좋지 않았음

→ 시계열 데이터의 특성을 고려해 2019년 만의 데이터를 이용하여 성능 개선

Research Question 1

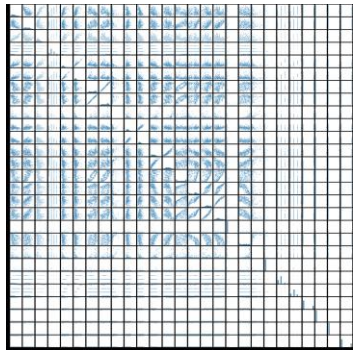
Contents

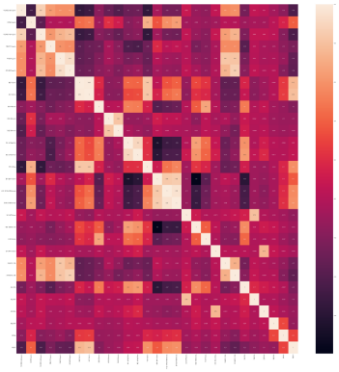
DM Process /
Variable Selection

01. EDA & VIF

(1) EDA

- pairplot과 heatmap을 통해 분포와 상관계수를 확인





(2) VIF

- for loop을 돌며 VIF가 가장 높은 변수를 순차적으로 제거
- VIF 최대값이 10 미만이 될 때까지 반복

[Numerical Variable 중 최종 사용 변수]

이산화질소농도(ppm) / 오존농도(ppm) /
초미세먼지(㎍/㎥) / 평균기온(℃) /
일강수량(mm) / 평균 풍속(m/s) /
일 최심적설(cm) / 평균 전운량(1/10) /
9-9강수(mm) / 안개 계속시간(hr)

```
1번째 VIF 측정
Max VIF feature & value : 최고기온(°C), 703.5177958270519
2번째 VIF 측정
Max VIF feature & value : 평균 상대습도(%), 182.94044233646602
3번째 VIF 측정
Max VIF feature & value : 합계 일조시간(hr), 129.80984464748482
4번째 VIF 측정
Max VIF feature & value : 최소 상대습도(%), 79.98766905033453
5번째 VIF 측정
Max VIF feature & value : 이산화탄소농도(ppm), 76.44009682570287
6번째 VIF 측정
Max VIF feature & value : 최고기온(°C), 57.22876462362821
7번째 VIF 측정
Max VIF feature & value : 평균 풍속(m/s), 52.1022770715405
8번째 VIF 측정
Max VIF feature & value : 이산화탄소농도(ppm), 37.00084280966748
9번째 VIF 측정
Max VIF feature & value : 이산화탄소농도(ppm), 34.35145103240715
10번째 VIF 측정
Max VIF feature & value : 최고기온(°C), 15.72778257651756
11번째 VIF 측정
Max VIF feature & value : 평균기온(°C), 15.347991592950716
12번째 VIF 측정
Max VIF feature & value : 오존농도(ppm), 7.82004429338727
```

02. Regression (OLS)

(1) Y1 (따릉이)

- **F-statistics** = 56.45, **prob(F-statistics)** = 2.01e-84로 도출된 회귀식이 회귀분석 모델 전체에 대해 통계적으로 유의
- **Adj. R-squared** = 0.792으로, 전체의 79% 정도를 설명하고 있음
- **p-value** 0.05 이상인 변수들을 제외하면 통계적으로 유의미하다고 판단

OLS Regression Results

Dep. Variable:	따릉이	R-squared:	0.806
Model:	OLS	Adj. R-squared:	0.792
Method:	Least Squares	F-statistic:	56.45
Date:	Sat, 18 Mar 2023	Prob (F-statistic):	2.01e-84
Time:	04:59:28	Log-Likelihood:	-3154.1
No. Observations:	292	AIC:	6350.
Df Residuals:	271	BIC:	6427.
Df Model:	20		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	6.029e+04	2929.557	20.580	0.000	5.45e+04	6.61e+04
C(기준_미세먼지)[T.2]	4055.7988	2482.598	1.634	0.103	-831.832	8943.430
C(기준_미세먼지)[T.3]	3349.8352	4833.005	0.693	0.489	-6165.173	1.29e+04
C(기준_미세먼지)[T.4]	5440.5260	1.34e+04	0.405	0.686	-2.1e+04	3.19e+04
C(초미세먼지_기준)[T.2]	-4794.0098	2626.989	-1.825	0.069	-9965.910	377.890
C(초미세먼지_기준)[T.3]	-8998.0513	5106.687	-1.762	0.079	-1.91e+04	1055.771
C(초미세먼지_기준)[T.4]	-8319.2676	1.1e+04	-0.755	0.451	-3e+04	1.34e+04
C(강수여부)[T.1]	-1.034e+04	2359.033	-4.385	0.000	-1.5e+04	-5699.421
C(적설여부)[T.1]	-1544.192	1.34e+04	-0.012	0.991	-2.65e+04	2.61e+04
C(안개여부)[T.1]	-4692.5212	1.15e+04	-0.406	0.685	-2.74e+04	1.8e+04
이산화질소농도	2407.2904	1491.774	1.614	0.108	-529.649	5344.230
오존농도	-828.0053	1023.150	-0.809	0.419	-2842.338	1186.327
초미세먼지	-4422.1139	2620.245	-1.688	0.093	-9580.738	736.510
평균기온	2.375e+04	1152.459	20.604	0.000	2.15e+04	2.6e+04
일강수량	-5011.4942	1063.446	-4.713	0.000	-7105.161	-2917.828
평균풍속	1616.0930	965.160	1.674	0.095	-284.071	3516.257
일_최심적설	101.7007	1097.700	0.093	0.926	-2059.403	2262.804
평균_전운량	-2945.7367	999.821	-2.946	0.003	-4914.141	-977.333
강수	-4606.7595	1040.819	-4.426	0.000	-6655.878	-2557.641
안개_계속시간	566.8696	1152.036	0.492	0.623	-1701.209	2834.948
휴일여부	-7338.1835	1619.061	-4.532	0.000	-1.05e+04	-4150.646

Omnibus:	4.201	Durbin-Watson:	2.118
Prob(Omnibus):	0.122	Jarque-Bera (JB):	3.880
Skew:	-0.251	Prob(JB):	0.144
Kurtosis:	3.259	Cond. No.	34.6

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

(2) Y2 (경복궁)

- **F-statistics** = 4.477, **prob(F-statistics)** = 4.15e-09로 도출된 회귀식이 회귀분석 모델 전체에 대해 통계적으로 유의
- **Adj. R-squared** = 0.193으로, 전체의 19% 정도를 설명하고 있음
- **p-value** 0.05 이상인 변수들을 제외하면 통계적으로 유의미하다고 판단

OLS Regression Results

Dep. Variable:	경복궁	R-squared:	0.248
Model:	OLS	Adj. R-squared:	0.193
Method:	Least Squares	F-statistic:	4.477
Date:	Sat, 18 Mar 2023	Prob (F-statistic):	4.15e-09
Time:	04:59:28	Log-Likelihood:	-3082.5
No. Observations:	292	AIC:	6207.
Df Residuals:	271	BIC:	6284.
Df Model:	20		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	1.396e+04	2291.974	6.089	0.000	9444.451	1.85e+04
C(기준_미세먼지)[T.2]	1980.2234	1942.291	1.020	0.309	-1843.674	5804.121
C(기준_미세먼지)[T.3]	1.118e+04	3781.160	2.958	0.003	3739.833	1.86e+04
C(기준_미세먼지)[T.4]	1.827e+04	1.05e+04	1.737	0.084	-2437.305	3.9e+04
C(초미세먼지_기준)[T.2]	-2908.6204	2055.256	-1.415	0.158	-6954.919	1137.679
C(초미세먼지_기준)[T.3]	-1133.9152	3995.279	-0.284	0.777	-8999.646	6731.815
C(초미세먼지_기준)[T.4]	-2190.4181	8620.984	-0.254	0.800	-1.92e+04	1.48e+04
C(강수여부)[T.1]	2741.2888	1845.618	1.485	0.139	-892.284	6374.861
C(적설여부)[T.1]	-1.197e+04	1.04e+04	-1.145	0.253	-3.25e+04	8603.233
C(안개여부)[T.1]	1177.3576	9033.245	0.130	0.896	-1.66e+04	1.9e+04
이산화질소농도	541.0978	1167.108	0.464	0.643	-1756.653	2838.848
오존농도	555.5848	800.474	0.694	0.488	-1020.353	2131.523
초미세먼지	-4455.8635	2049.981	-2.174	0.031	-8491.776	-419.951
평균기온	2040.7448	901.640	2.263	0.024	265.635	3815.855
일강수량	-1364.0188	832.000	-1.639	0.102	-3002.024	273.987
평균풍속	-1499.5313	755.105	-1.986	0.048	-2986.148	-12.914
일_최심적설	519.0404	858.799	0.604	0.546	-1171.725	2209.806
평균_전운량	-1269.6881	782.222	-1.623	0.106	-2809.693	270.317
강수	-830.1623	814.297	-1.019	0.309	-2433.315	772.991
안개_계속시간	17.6893	901.310	0.020	0.984	-1756.770	1792.148
휴일여부	6519.0198	1266.692	5.146	0.000	4025.211	9012.828

Omnibus:	387.505	Durbin-Watson:	2.116
Prob(Omnibus):	0.000	Jarque-Bera (JB):	44007.869
Skew:	6.050	Prob(JB):	0.00
Kurtosis:	61.912	Cond. No.	34.6

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

Research Question 1

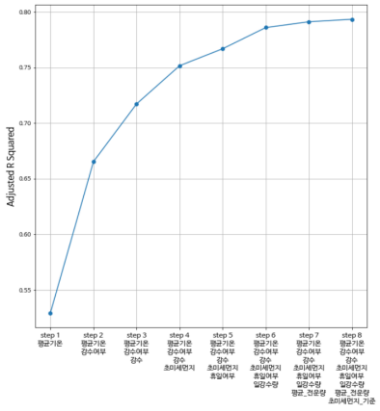
Contents

DM Process /
Variable Selection

03. Variable Selection

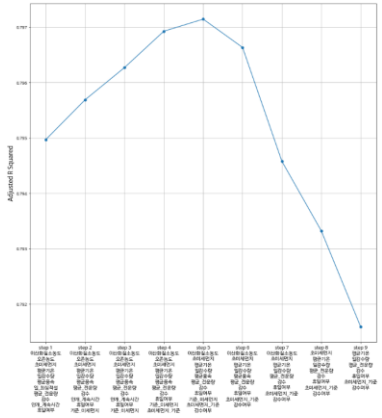
(1) Y1 : 일일 서울시 공공자전거(따릉이) 대여 건수

① 전진 선택법(Forward Selection)



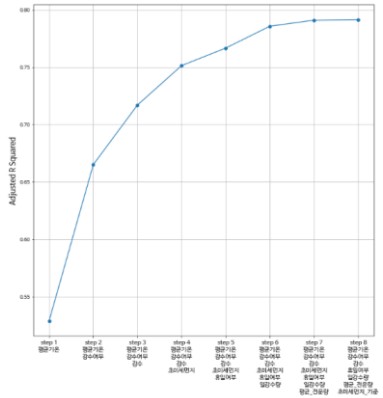
- 총 8개의 변수가 선택됨
- F-statistics = 140.6, prob(F-statistics) = 6.35e-94로 도출된 회귀식이 회귀분석 모델 전체에 대해 통계적으로 유의
- Adj. R-squared = 0.799으로, 전체의 80% 정도를 설명하고 있음
- p-value 0.05 이상인 변수들을 제외하면 통계적으로 유의미하다고 판단

② 후진 선택법(Backward Elimination)



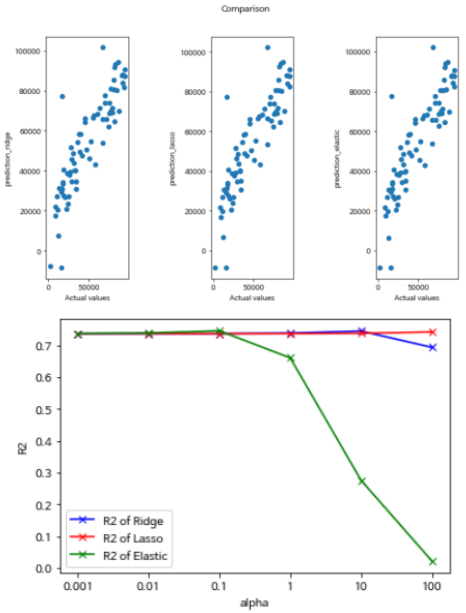
- 총 7개의 변수가 선택됨
- F-statistics = 158.9, prob(F-statistics) = 2.58e-94로 도출된 회귀식이 회귀분석 모델 전체에 대해 통계적으로 유의
- Adj. R-squared = 0.797으로, 전체의 80% 정도를 설명하고 있음
- p-value 0.05 이상인 변수들을 제외하면 통계적으로 유의미하다고 판단

③ 단계별 선택법(Stepwise Selection)



- 총 7개의 변수가 선택됨
- F-statistics = 158.9, prob(F-statistics) = 2.58e-94로 도출된 회귀식이 회귀분석 모델 전체에 대해 통계적으로 유의
- Adj. R-squared = 0.797으로, 전체의 80% 정도를 설명하고 있음
- p-value 0.05 이상인 변수들을 제외하면 통계적으로 유의미하다고 판단

④ Ridge & Lasso & Elastic Net



- 큰 alpha 값에 대해서는 Lasso가 가장 우수
- 모델별 최적 파라미터 상에서의 성능 비교

R-squared : Linear = 0.74, Ridge = 0.74, Lasso = 0.74, Elastic Net = 0.74
MSE : Linear = 202736114.02, Ridge = 198467211.98, Lasso = 198764427.26, Elastic Net = 199819289.10
MAE : Linear = 11019.17, Ridge = 10962.76, Lasso = 10940.26, Elastic Net = 10978.17

Research Question 1

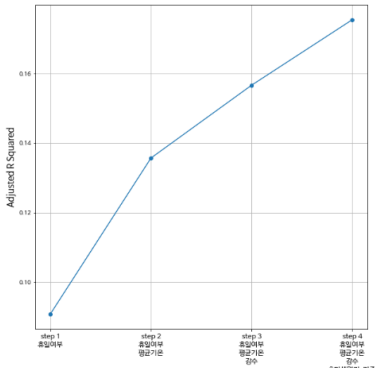
Contents

DM Process /
Variable Selection

03. Variable Selection

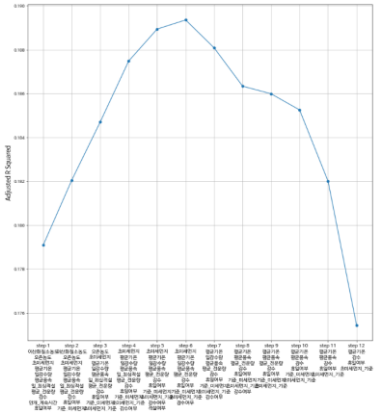
(2) Y2 : 일일 경북궁 관람객 수

① 전진 선택법(Forward Selection)



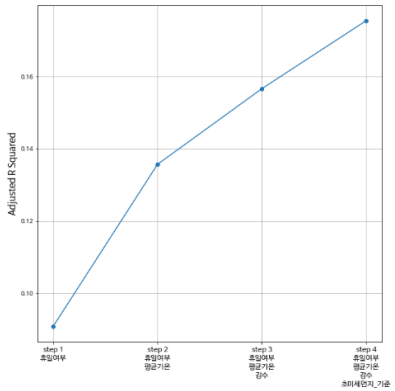
- 총 4개의 변수가 선택됨
- F-statistics = 16.48, prob(F-statistics) = 3.63e-12로 도출된 회귀식이 회귀분석 모델 전체에 대해 통계적으로 유의
- Adj. R-squared = 0.187으로, 전체의 19% 정도를 설명하고 있음
- p-value 0.05 이상인 변수들을 제외하면 통계적으로 유의미하다고 판단

② 후진 선택법(Backward Elimination)



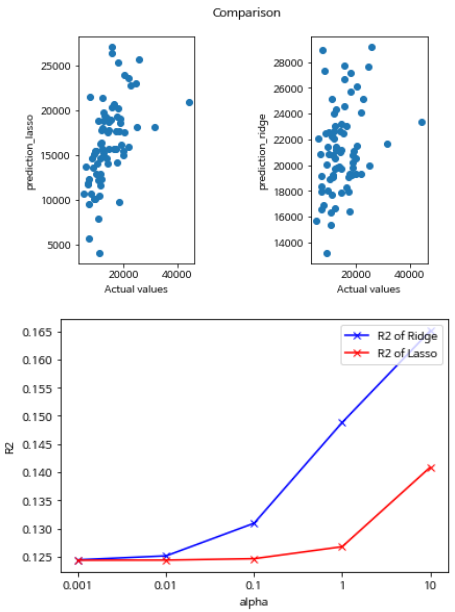
- 총 4개의 변수가 선택됨
- F-statistics = 16.48, prob(F-statistics) = 3.63e-12로 도출된 회귀식이 회귀분석 모델 전체에 대해 통계적으로 유의
- Adj. R-squared = 0.187으로, 전체의 19% 정도를 설명하고 있음
- p-value 0.05 이상인 변수들을 제외하면 통계적으로 유의미하다고 판단

③ 단계별 선택법(Stepwise Selection)



- 총 4개의 변수가 선택됨
- F-statistics = 16.48, prob(F-statistics) = 3.63e-12로 도출된 회귀식이 회귀분석 모델 전체에 대해 통계적으로 유의
- Adj. R-squared = 0.187으로, 전체의 19% 정도를 설명하고 있음
- p-value 0.05 이상인 변수들을 제외하면 통계적으로 유의미하다고 판단

④ Ridge & Lasso Regression



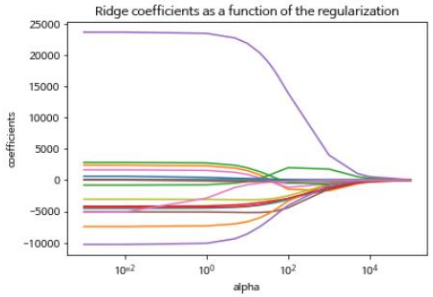
- Ridge의 성능이 Lasso보다 우수
- train accuracy : 0.22 / test accuracy : 0.12
- alpha 값의 증가에 따라 성능의 차이가 커짐

Research Question 1 / 2

Contents

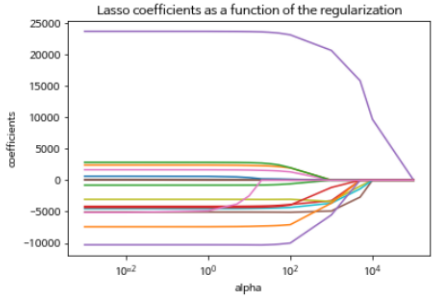
DM Process /
Variable Selection

04. Shrinkage Comparison & Result



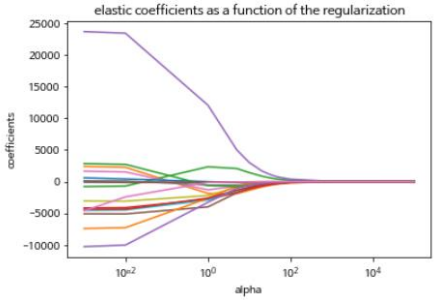
- Y1(따릉이)에 대한 회귀 모델에서 R-squared가 약 0.8 정도로 유의미하게 나와 예측 성능에 있어 우수하다고 판단함

- 다양한 변수 선택법과 더불어 Ridge, Lasso, Elastic Net에 대해 성능을 확인하였고, Shrinkage 효과를 비교 확인함



- Y1(따릉이) 예측에 있어 초기에 우리가 직접 수집한 변수들 중 기온, 강수, 초미세먼지 등의 변수들이 유의미함을 확인함

- Y2(경복궁) 예측 성능은 Y1(따릉이)에 비해 낮지만, 선택된 변수들은 동일함을 확인할 수 있었음



- 분석 이전에는 문화 활동과 여가 활동에 유사한 날씨 변수가 영향을 미칠 것으로 생각하였으나, Y2(경복궁)의 예측력은 Y1(따릉이)에 비해 매우 낮아 여가 활동에 대해서는 날씨 변수 이외에 추가적인 다른 요인을 고려해야 할 필요가 있음

Research Question 2

01. Chi-square Test

(1) 강수여부 - 미세먼지 독립 여부 판단

- chi-square statistic = 26.906, p-value = 0.000006 < 0.05

(2) 강수여부 - 초미세먼지 독립 여부 판단

- chi-square statistic = 14.599, p-value = 0.002194 < 0.05

02. Result

H_0(귀무 가설) = 강우 와 (초)미세먼지는 서로 독립이다

H_A(대립 가설) = 강우 와 (초)미세먼지는 서로 독립이 아니다

→ p-value가 기준치 (0.05)보다 작아 귀무 가설을 기각하고 대립 가설 채택

Chi-square 검정 중 독립성 검정 결과, 상식과 동일한 결론을 얻을 수 있었으며, 자료가 다항분포나 이항분포를 따른다고 가정하며, 기대도수가 5보다 커야 한다는 Chi-square 검정의 가정 만족

Data Source

Contents

DM Process /
Variable Selection

[데이터 출처]

(1) 서울 열린 데이터 광장
- 서울시 일별 평균 대기오염도 정보 http://data.seoul.go.kr/dataList/OA-2218/S/1/datasetView.do#
- 서울시 공공자전거 이용현황 https://data.seoul.go.kr/dataList/OA-14994/F/1/datasetView.do

(2) 공공데이터포털
- 문화재청 궁능유적본부_4대궁 관람객 수 현황https://www.data.go.kr/data/15040885/fileData.do?recommendDataYn=Y

(3) 기상청
- 일별 종관기상관측(ASOS) 자료 https://data.kma.go.kr/data/grnd/selectAsosRltmList.do?pgmNo=36

(4) 2016 ~ 19년 공휴일 데이터
- https://superkts.com/day/holiday/2019

[코드 출처]

(1) VIF 계산 코드
- https://signature95.tistory.com/m/18