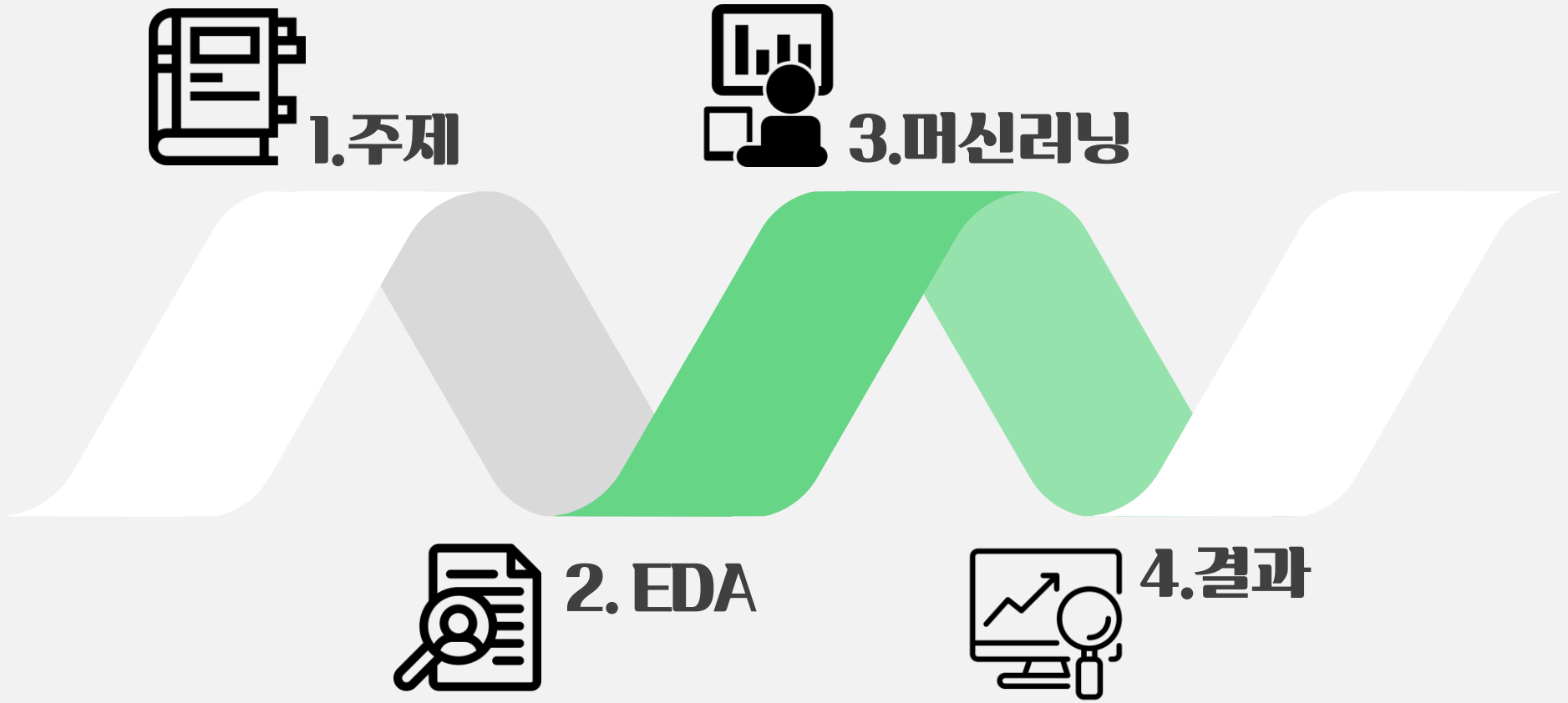
2조

PRESENTATION

퇴근시간 버스 승차인원 예측

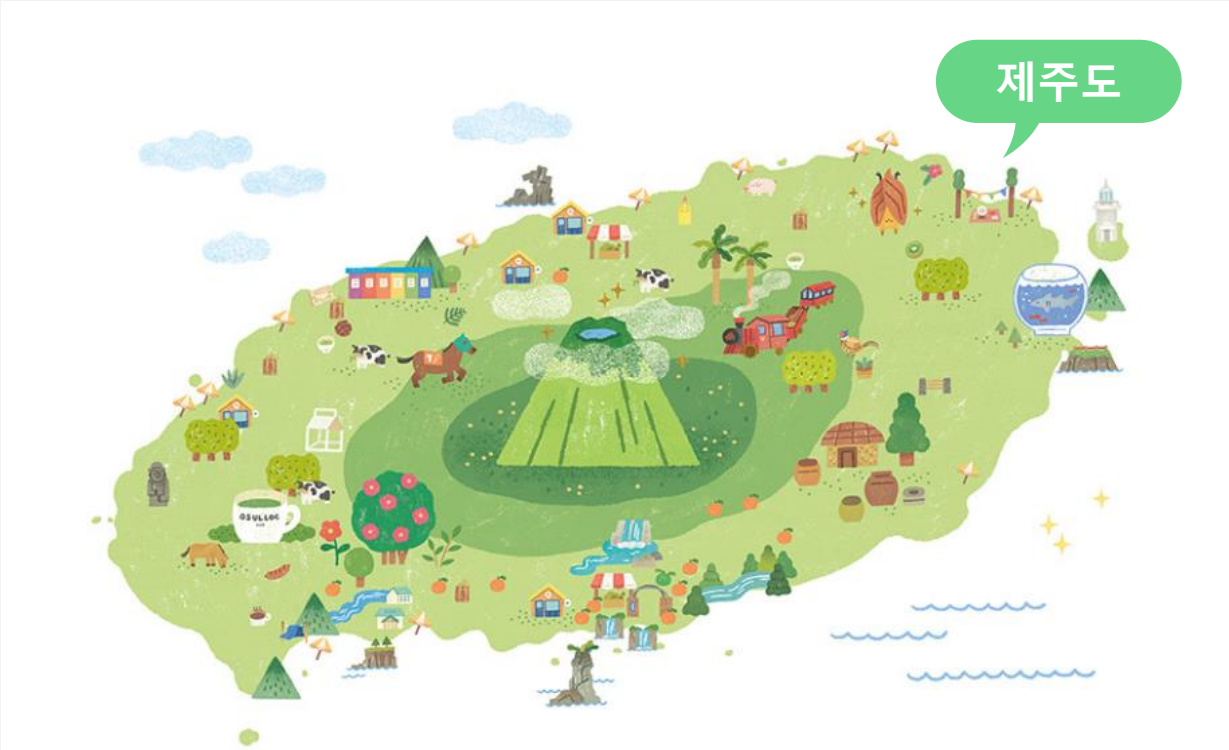
김연지
정유경
강창균

목차



1.주제 배경 및 목적

코로나 시국에 여행을 어디로 가려고 생각하시나요?



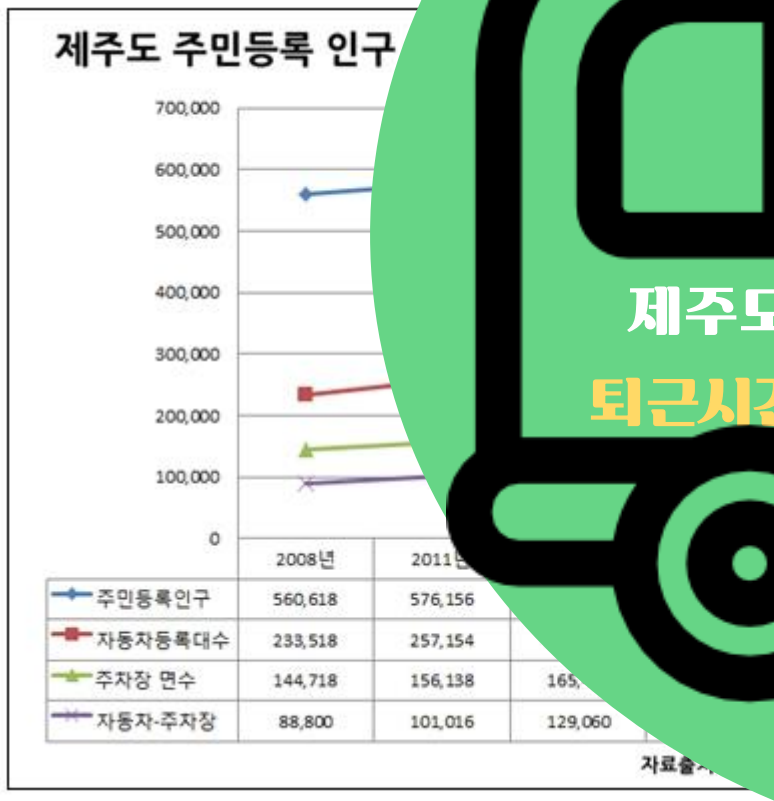
렌트카?

택시?

버스 투어?

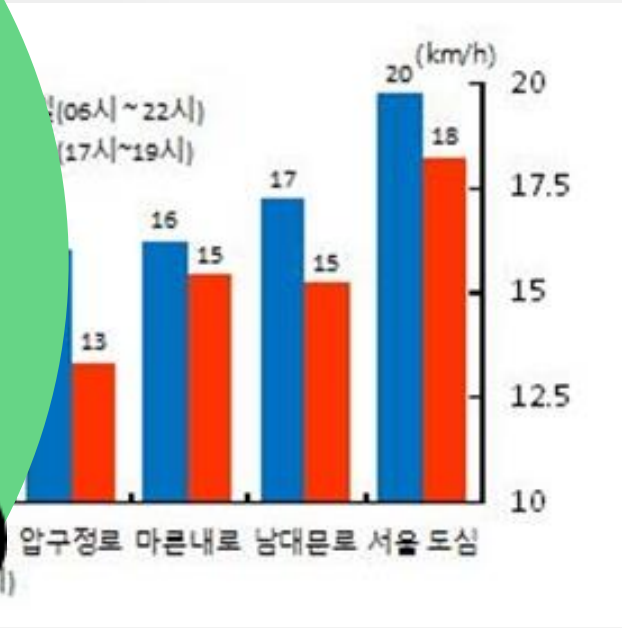
1.주제 배경 및 목적

제주도의 인구 수와 ... 통량?



제주도 버스의 효율적인 운행을 위해

퇴근시간 버스 승차 인원 예측을 해보자!



2.EDA 데이터 컬럼들

TARGET

18-20 ride

퇴근시간 6~8시 버스 승차 인원

In_out

시내버스, 시외버스 구분

X-Y ride

오전 6~8시 / 8~ 10 시 / 10~12 시

승차인원

Longitude / Latitude

버스정류장의 경도와 위도

X-Y takeoff

오전 6~8시 / 8~ 10 시 / 10~12 시

하차인원

Date / Weekday

날짜 : 년도 – 월 – 일

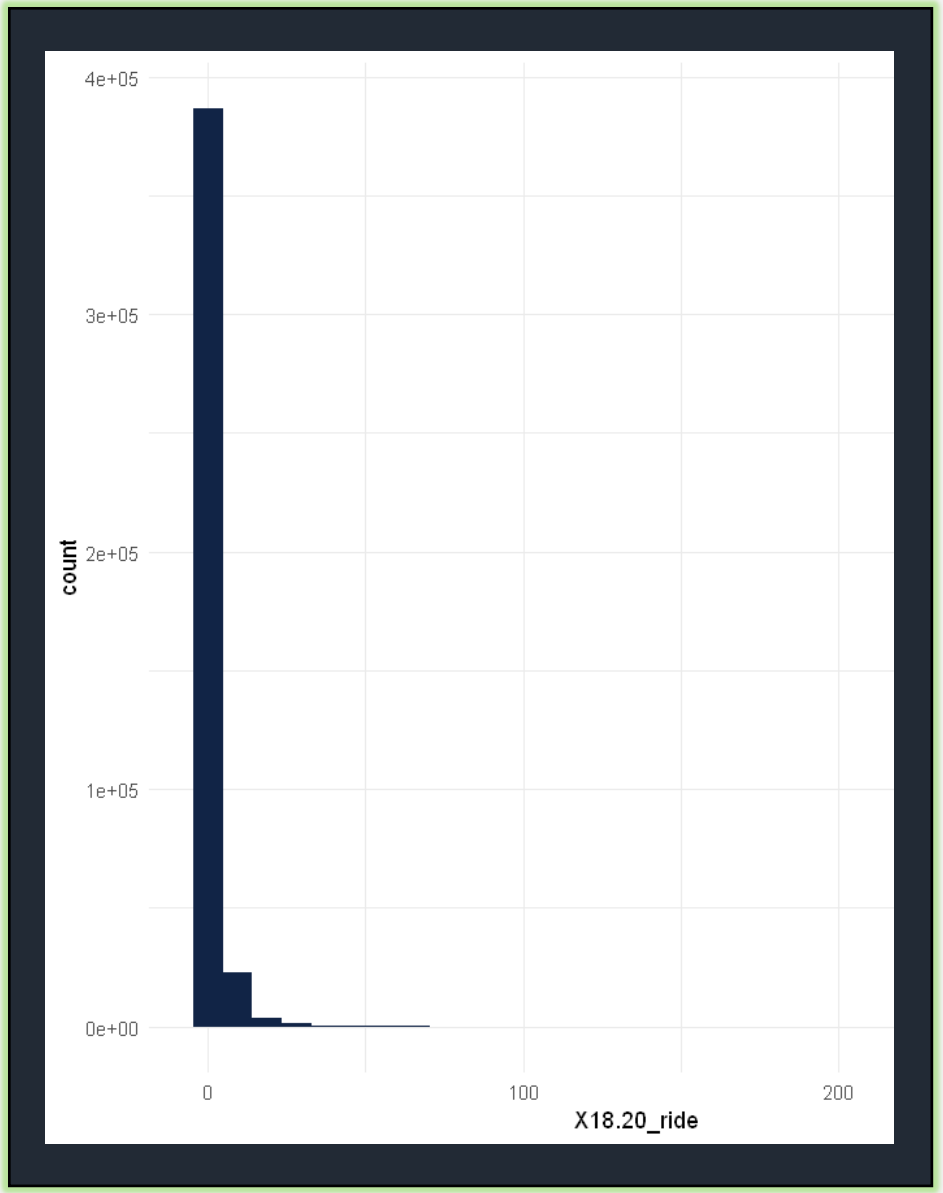
요일 : 월-화-수-목-금-토-일

제주도 4곳과 정류장 거리

제주시/ 고산/ 서귀포/성산과 정류

장과의 거리 차이

2.EDA 데이터 이슈



하차인원

하차 태그를 하지 않은 사람들도 있어서 승차인원과 하자인원이 다를 수 있다.

불균형 데이터

Target 컬럼인 오후 6-8시 승차 인원이 0에 몰려있다.

2.EDA Train & Test 살펴보기

Train과 Test의 null값 확인

colSums(is.na(train))

id	0
date	0
bus_route_id	0
in_out	0
station_code	0
station_name	0
latitude	0
longitude	0
X6.7_ride	0
X7.8_ride	0
X8.9_ride	0
X9.10_ride	0
X10.11_ride	0
X11.12_ride	0
X6.7_takeoff	0
X7.8_takeoff	0
X8.9_takeoff	0
X9.10_takeoff	0
X10.11_takeoff	0
X11.12_takeoff	0
X18.20_ride	0

colSums(is.na(train))

id	0
date	0
bus_route_id	0
in_out	0
station_code	0
station_name	0
latitude	0
longitude	0
X6.7_ride	0
X7.8_ride	0
X8.9_ride	0
X9.10_ride	0
X10.11_ride	0
X11.12_ride	0
X6.7_takeoff	0
X7.8_takeoff	0
X8.9_takeoff	0
X9.10_takeoff	0
X10.11_takeoff	0
X11.12_takeoff	0
X18.20_ride	0

날짜 범위 비교

```
max(as.Date(train$date))
min(as.Date(train$date))
cat('-------------------------------------', fill=T)
max(as.Date(test$date))
min(as.Date(test$date))
```

```
2019-09-30
2019-09-01
-------------------------------------
2019-10-16
2019-10-01
```

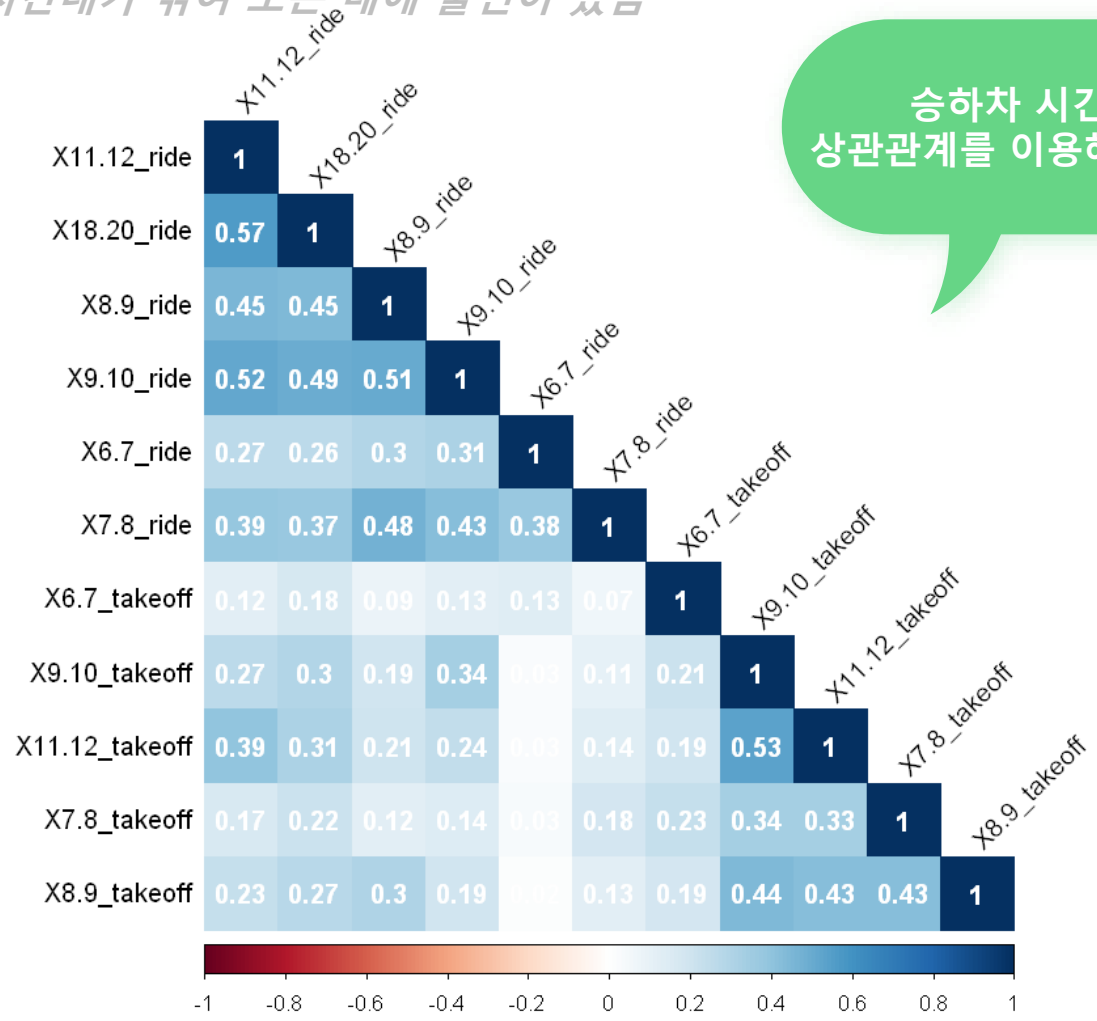
Train date

Test date

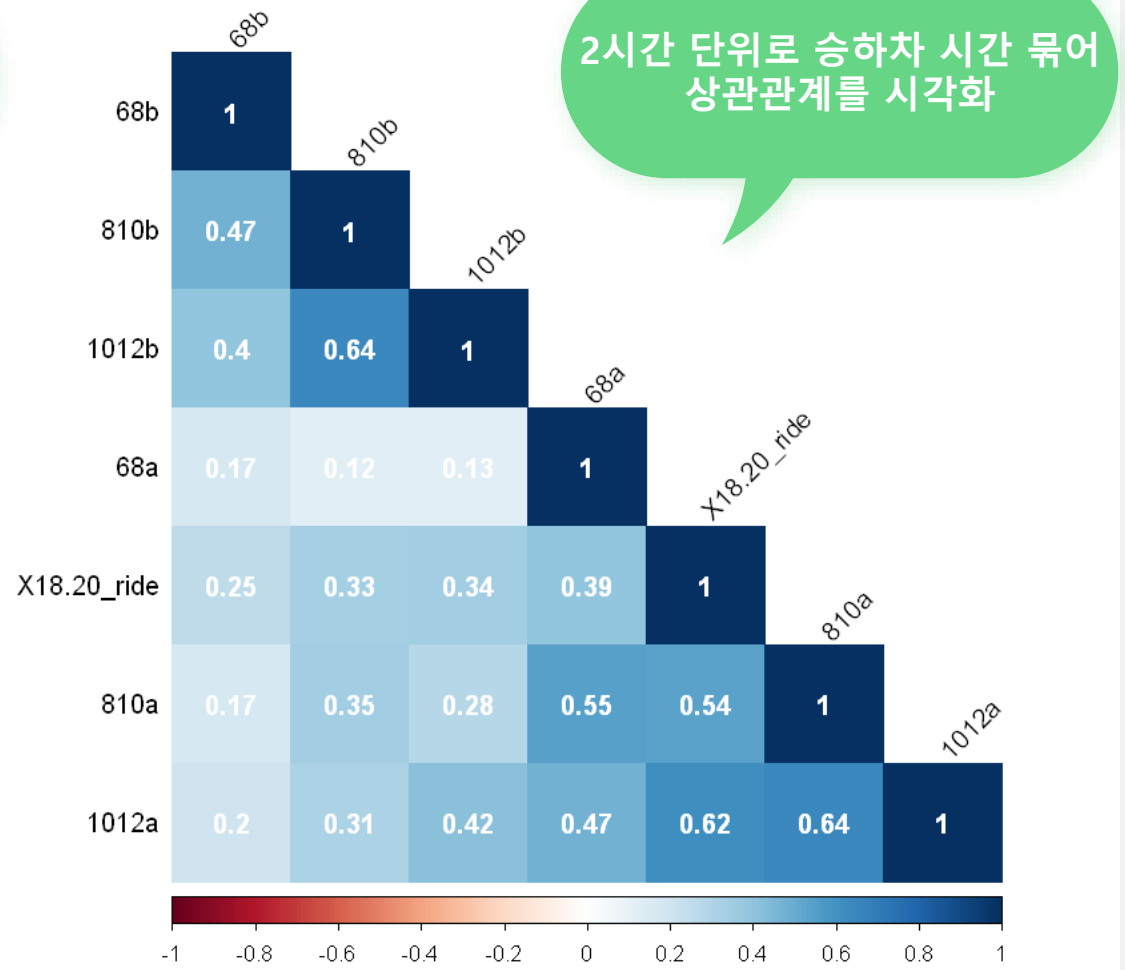
2.EDA 변수들 Heatmap

※ 시간대가 섞여 보는 데에 불편이 있음

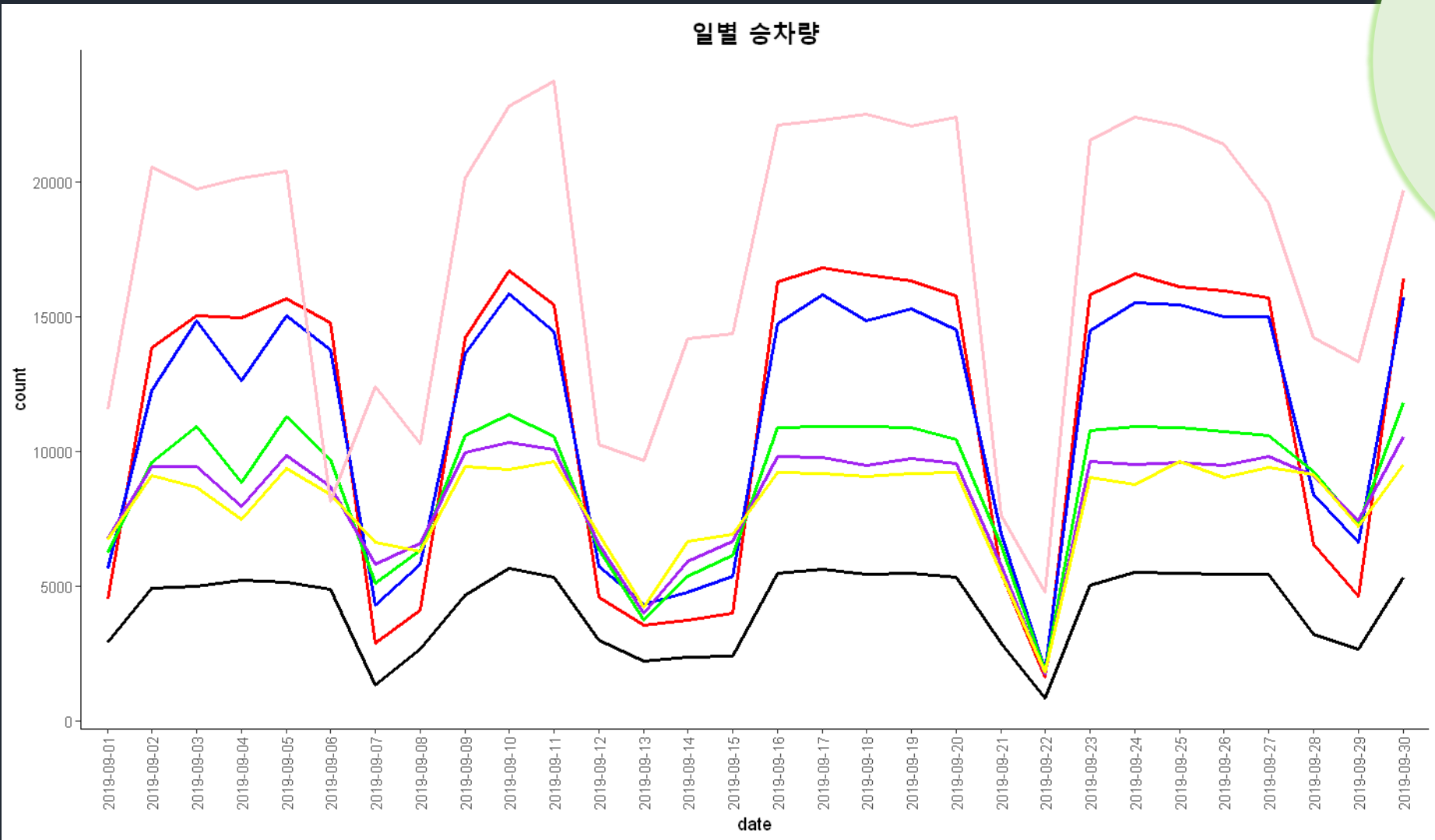
승하차 시간을
상관관계를 이용해 시각화



2시간 단위로 승하차 시간 묶어
상관관계를 시각화



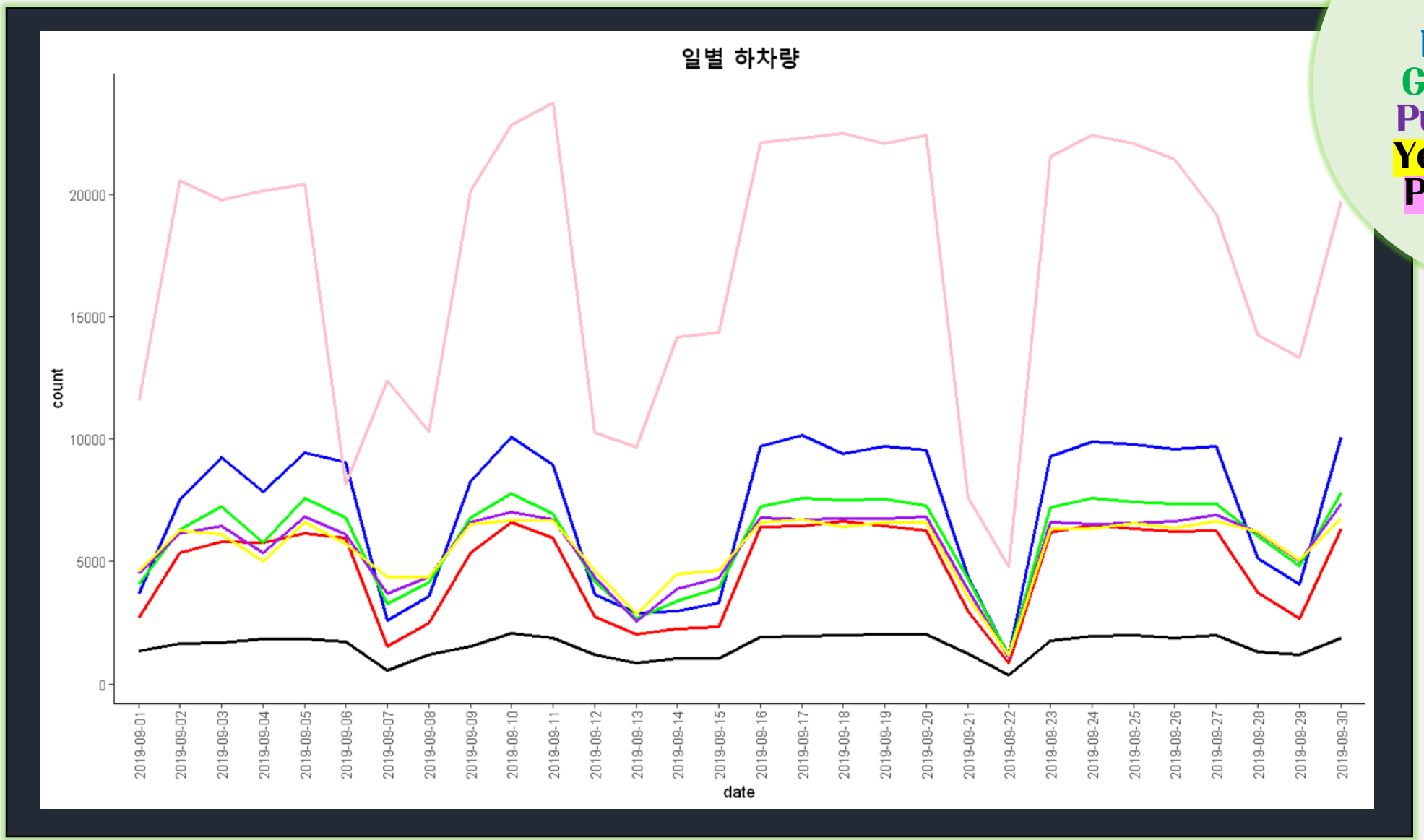
2.EDA 요일별 승차량



Black = 6~7시
Red = 7~8시
Blue = 8~9시
Green = 9~10시
Purple = 10~11시
Yellow = 11~12시
Pink = 18~20시

5일과 2일 주기로 승차량이 급증하고 급감하는 것을 확인

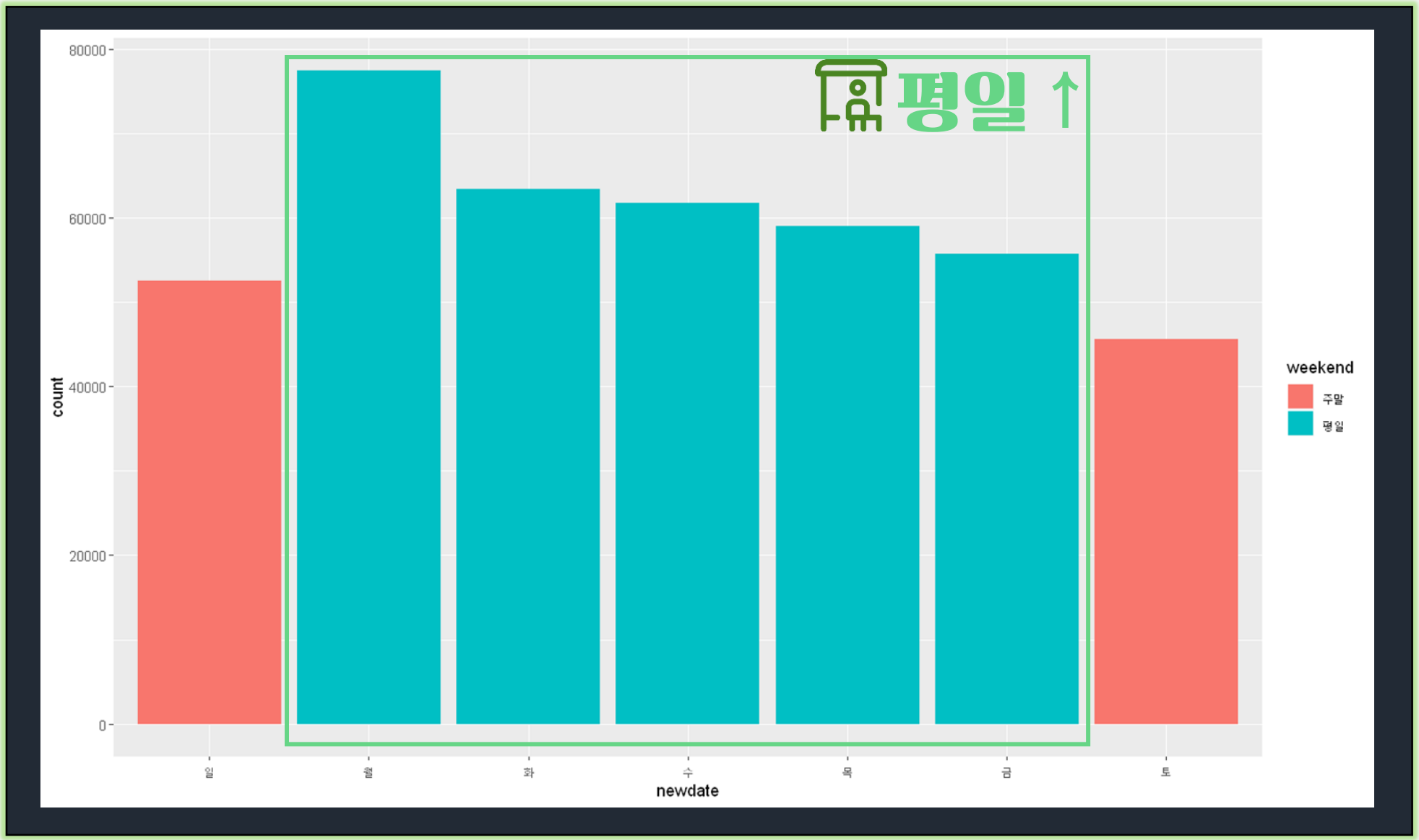
2.EDA 요일별 하차량



Black = 6~7시
Red = 7~8시
Blue = 8~9시
Green = 9~10시
Purple = 10~11시
Yellow = 11~12시
Pink = 18~20시

제주도의 급행 버스를 제외하고 하차 태그를 하지 않아도 추가요금 없다!

2.EDA 평일·주말별 승하차 인원



2.EDA 제주도 관측소 & 버스정류장

관측도

정류장

2.EDA 정류장 확대

추자도

제주 본섬

3.평가 독립변수 유의미성 판단

```
Call:
lm(formula = X18.20_ride ~ ., data = train)

Residuals:
    Min      1Q  Median      3Q     Max
-69.450  -0.641  -0.110   0.219 244.052

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.744e+01  5.051e+00   5.432 5.57e-08 ***
X.1          2.986e-06  3.332e-07   8.962  < 2e-16 ***
weekdays    -1.878e-01  2.077e-02  -9.041  < 2e-16 ***
X           -6.898e-08  4.670e-08  -1.477    0.140
bus_route_id -2.406e-08  1.569e-09 -15.340  < 2e-16 ***
weekday     -3.196e-01  3.179e-02 -10.055  < 2e-16 ***
out         -1.128e-01  4.321e-02  -2.612    0.009 **
ride6_8      1.144e-01  2.391e-03  47.846  < 2e-16 ***
ride8_10     2.358e-01  2.137e-03 110.309  < 2e-16 ***
ride10_12    5.744e-01  2.303e-03 249.362  < 2e-16 ***
off6_8       2.358e-01  4.146e-03  56.866  < 2e-16 ***
off8_10      1.297e-01  2.914e-03  44.494  < 2e-16 ***
off10_12     5.835e-02  3.037e-03  19.212  < 2e-16 ***
latitude     7.501e-01  5.247e-02  14.295  < 2e-16 ***
longitude   -3.999e-01  3.940e-02 -10.149  < 2e-16 ***
on_people    3.034e-05  3.053e-06   9.937  < 2e-16 ***
off_people  -5.328e-05  5.453e-06  -9.770  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.535 on 415406 degrees of freedom
Multiple R-squared:  0.4396,    Adjusted R-squared:  0.4396
F-statistic: 2.036e+04 on 16 and 415406 DF,  p-value: < 2.2e-16
```

P-VALUE

각 독립변수가 얼마나 종속변수에 영향을 미치는지 보여주는 자료

P값이 작을수록 독립변수가 모델에서의 유의미하다.

2.2e-16

p-value<0.05이므로, 해당 독립변수들은

통계적으로 유의미하다고 할 수 있다.

3.평가 회귀 모델 지표

R squared	RMSE
높을수록 좋은 지표	낮을수록 좋은 지표
높을수록 모형의 종속변수와 독립변수 사이의 상관관계가 높아 해당 모델이 유용하다는 뜻	낮을수록 예측값과 실제값의 차이가 없다는 뜻
$R^2 = 1 - \frac{\Sigma(\text{오차}^2)}{\Sigma(\text{편차}^2)}$ 오차 $= (t_i - y_i)$, t_i:실제값 , y_i:예측값 편차 $= (t_i - \overline{t_i})$, t_i:실제값 , $\overline{t_i}$:평균값	$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y})^2}$
예측값 Variance / 실제값 Variance 오차^2은 내가 만든 모델의 에러율 편차^2은 평균으로 예측하는 Zero-R 모델의 에러율	MSE에 비해 이상치에 대한 민감도가 낮음

3.평가 사용 머신러닝

01 LightGBM Regressor — sklearn

02 Linear Regressor — sklearn

03 Logistic Regressor — sklearn

3.평가 머신러닝 결과

결정계수 R^2	RMSE
0.7270988	2.466917

Light GBM

R^2이 0.7정도면 쓸만한 머신러닝이라고 본다. 해당 모델은 0.73정도이므로 사용가능한 모델이다.

결정계수 R^2	RMSE
0.4395837	3.535145

Linear Regression

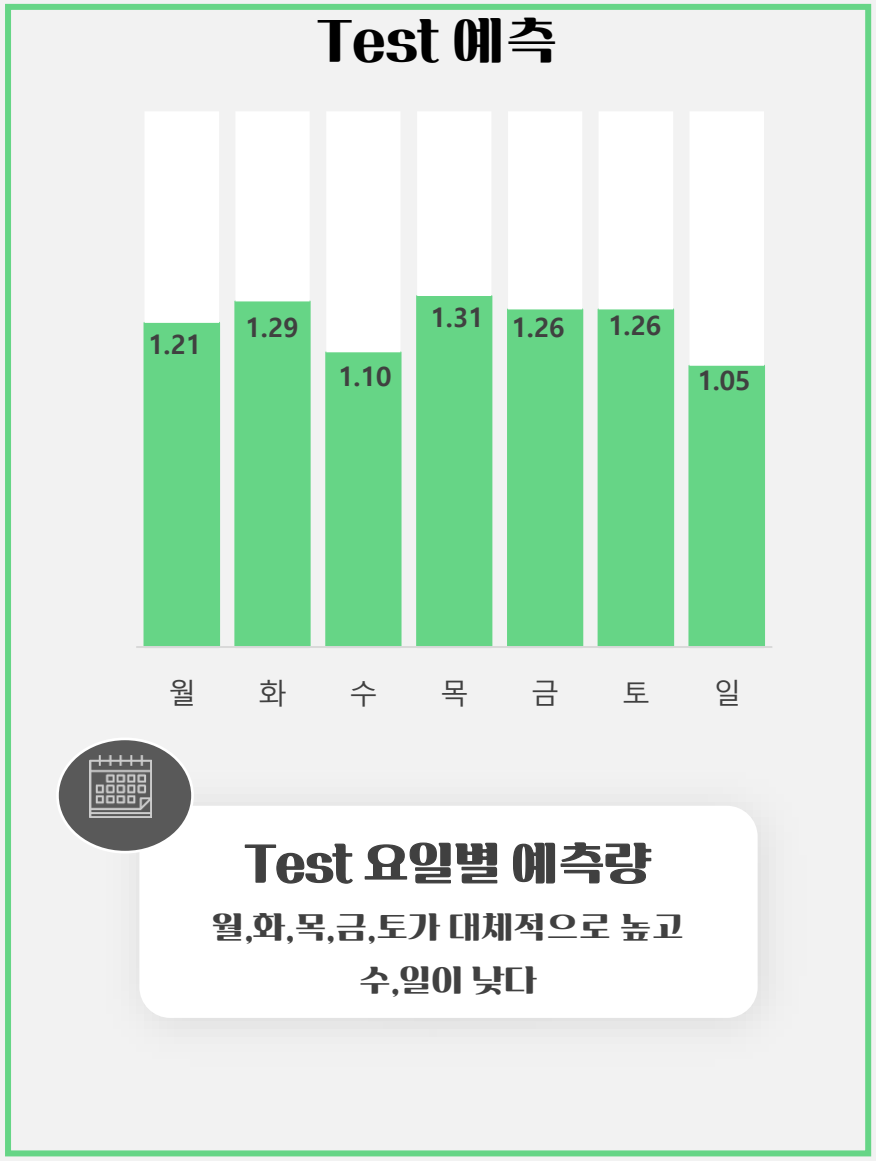
해당 모형의 예측도가 0.44정도 나왔다. Logistic모형보단 낫지만 쓸 수 있다고 볼 수 없다.

결정계수 R^2	RMSE
-0.06918409	4.882903

Logistic Regression

R^2은 0과 1사의 수를 갖는데, 해당 모형의 결정계수는 마이너스 값이 나왔다. 오차가 편차보다 크다는 의미로 굉장히 좋지 못한 모델임을 의미한다.

4.결과 예측 결과



Test 요일별 예측량

월,화,목,금,토가 대체적으로 높고
수,일이 낮다

Train 요일별 예측량

월,화,수,목이 대체적으로 높고
금,토,일이 낮다

4.결과 예측 결과

Test 예측

제주시	고산	서귀포	성산
1.27	0.108	1.23	0.928

Test 지역별 예측량

순위는 Train이랑 똑같지만 가장 많은
제주시는 Train이랑 0.36정도
차이난다.

제주시	고산	서귀포	성산
1.63	0.579	0.77	0.434

Train 지역별 예측량

1등 제주시 2등 서귀포
3등 고산4등 성산

THANK YOU