

서울특별시 공공자전거, 따릉이 대여 수 예측을 위한 탐색적 데이터 분석



서울특별시 마포구 따릉이 데이터 EDA



데이터 전처리 및 탐색 시 유의사항

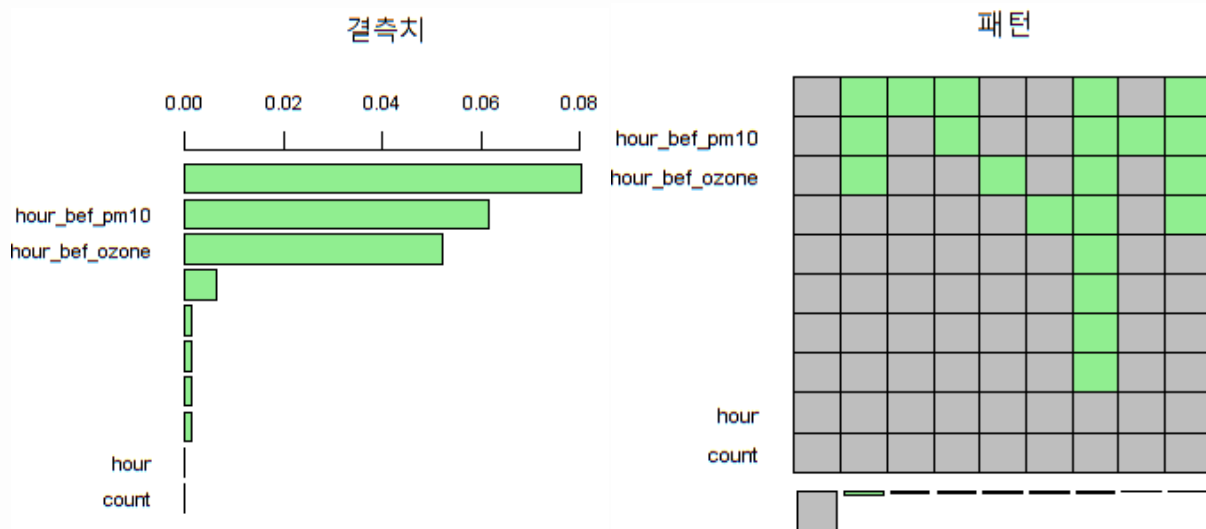
- 해당 도메인 및 프로젝트에 대한 이해 기반 문제 정의
 - 분석 목적을 고려하여 머신러닝 모델을 왜곡할 우려가 있는 결측값 대체 및 이상치 처리
 - 변수 간 관계 파악 및 분석에 용이한 새로운 변수 파생
 - 숫자 또는 문자로 발견하기 어려운 패턴을 직관적으로 전달하기 위한 데이터 시각화
- 시간대 범주형 파생변수
 - 극단적 대기오염물질 농도 논리형 파생변수
timeZone, atmosphere



따릉이 데이터 결측값 처리

데이터 분석 목적

- 서울특별시 마포구의 시간별 기상상황 및 따릉이 대여수 데이터
- 1시간 전 기상상황 데이터로 1시간 후 시간대의 따릉이 대여수 예측 목적의 데이터 분석



결측값 시각화 및 대체

- 결측값을 예측값으로 대체할 변수
 - pm10(117개) pm2.5(90개)**
결측값을 회귀식 추정하여 예측값으로 대체
 - 1시간 전 오존(76개)**
결측값을 다중대체법(mice function)*에 따라 여러 모델 동시에 사용하는 앙상블 기법으로 예측한 값으로 대체
- 결측값이 발생하는 경우의 유형에 따라 처리 기준이 달라진다 (데이터 누락 | 0 | 무작위 발생)

*) Multiple Imputation by Chained Equations

따릉이 데이터 이상치 처리

이상치 시각화 및 처리

- 이상치 검출

- 박스플롯

$Q1 - 1.5 * IQR$ 이하 $Q3 + 1.5 * IQR$ 이상

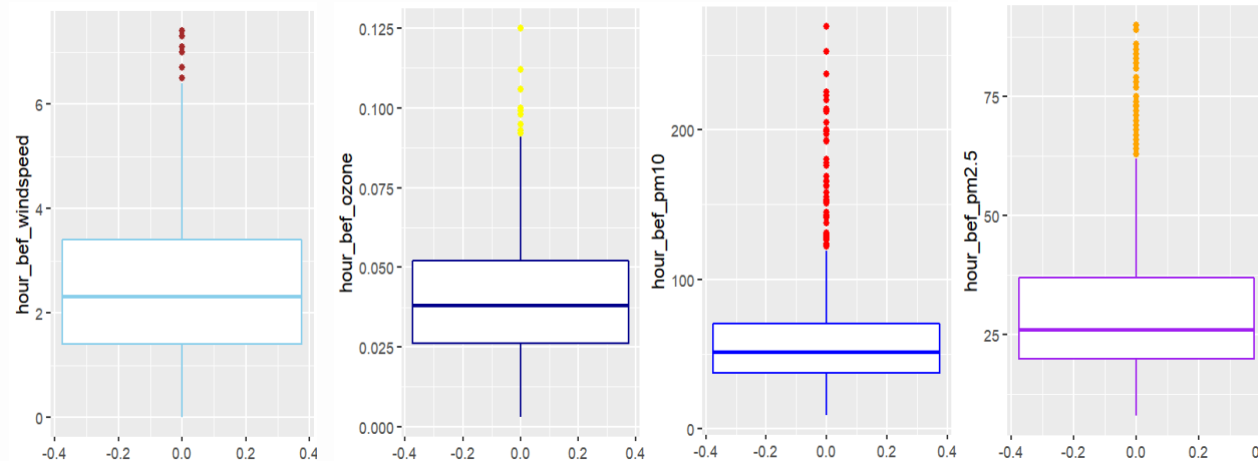
- 표준정규분포를 이용한 두 값 사이의 확률에서 벗어난 확률변수

$\mu - 2.58 * \sigma$ 이하 $\mu + 2.58 * \sigma$ 이상

- 쿡의 거리로 다중회귀식의 그래프 개형에 영향을 미치는 값

다중회귀모형에서 Cook's Distance 값(표준화된 잔차의 합) 확인

$$D_i = \frac{\sum_{j=1}^n (\hat{y}_j - \hat{y}_{j(i)})^2}{pMSE}$$



자연 발생적인 이상치로 판단되어 값에 대한 파악

- 풍속 변수의 이상치 존치 (예시)

오존 이상치, 풍속 이상치, 고온, 낮은 습도, 햇빛 : 저녁 시간대 자전거 280대 대여

미세먼지 이상치, 풍속 이상치, 서늘한 기온, 낮은 습도, 햇빛 : 저녁 시간대 자전거 106대 대여

- 대기오염물질 농도(오존, 미세먼지 입자) 관련 변수의 이상치 여부 관련 파생변수 생성

이상치와 분리하여 시각화 및 모델링 가능



따릉이 데이터 이상치 처리

이상치 시각화 및 처리

- 이상치 검출 $D_i = \frac{\sum_{j=1}^n (\hat{y}_j - \hat{y}_{j(i)})^2}{pMSE}$

쿡의 거리로 다중회귀식의 그래프
개형에 영향을 미치는 값의 행 제거

```
Call:
lm(formula = count ~ hour_bef_temperature + hour_bef_windspeed +
    hour_bef_humidity + hour_bef_ozone + hour_bef_pm10, data = train)
```

Residuals:

	Min	1Q	Median	3Q	Max
	-174.354	-41.626	-8.728	35.187	226.245

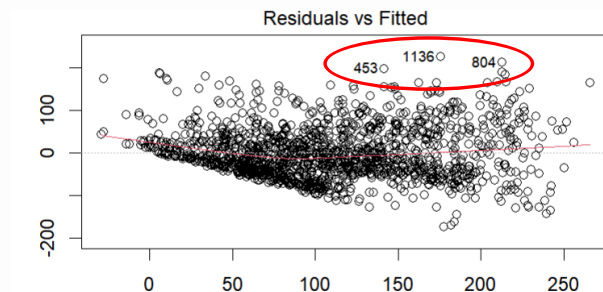
Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	8.25296	10.50492	0.786	0.432
hour_bef_temperature	6.56891	0.38108	17.238	< 2e-16 ***
hour_bef_windspeed	11.00362	1.38899	7.922	4.63e-15 ***
hour_bef_humidity	-0.62616	0.09385	-6.672	3.59e-11 ***
hour_bef_ozone	443.72757	106.01214	4.186	3.02e-05 ***
hour_bef_pm10	-0.37047	0.05105	-7.256	6.47e-13 ***

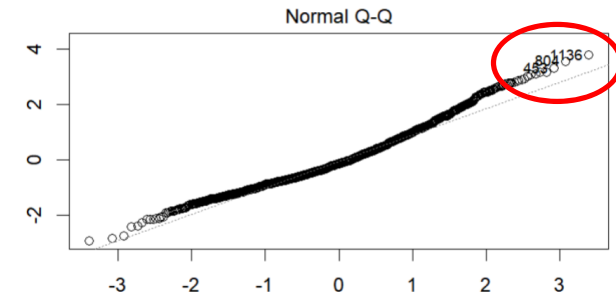
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 59.68 on 1444 degrees of freedom
(결측으로 인하여 9개의 관측치가 삭제되었습니다.)

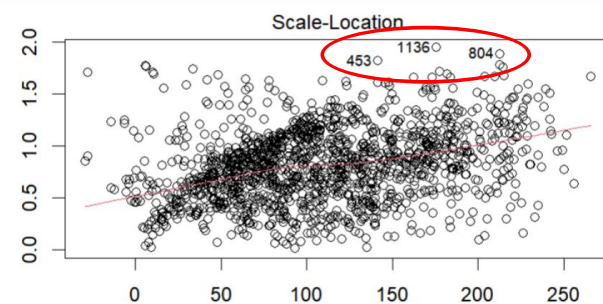
Multiple R-squared: 0.4804, Adjusted R-squared: 0.4786
F-statistic: 267 on 5 and 1444 DF, p-value: < 2.2e-16



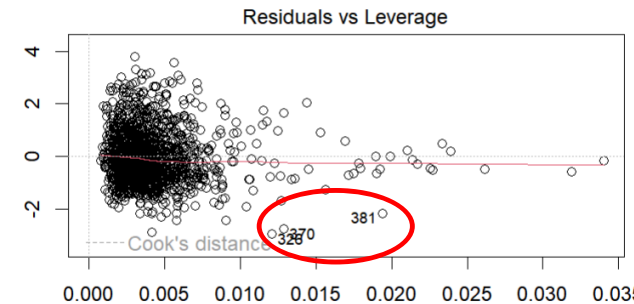
Fitted values
count ~ hour_bef_temperature + hour_bef_windspeed + hour_bef_humidity +



Theoretical Quantiles
count ~ hour_bef_temperature + hour_bef_windspeed + hour_bef_humidity +



Fitted values
count ~ hour_bef_temperature + hour_bef_windspeed + hour_bef_humidity +



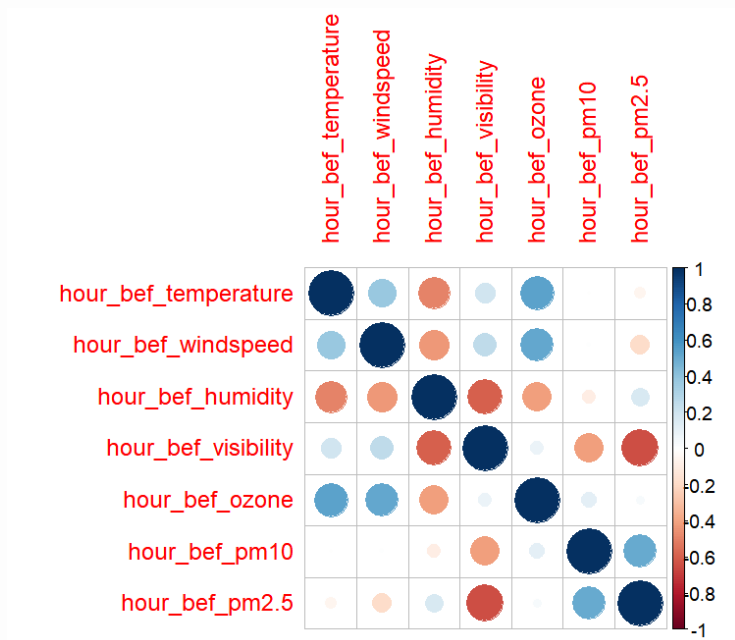
Leverage
count ~ hour_bef_temperature + hour_bef_windspeed + hour_bef_humidity +

pm2.5는 2p-value > 0.05로
상관계수 추정치가 통계적으로 유의하지 않은 바
다중회귀식 추정 시 pm2.5 제거



따릉이 데이터 상관분석

변수 간 상관관계 분석



- temperature와 ozone 간 상관계수 $r = 0.54$
- windspeed와 temperature 간 상관계수 $r = 0.38$
- windspeed와 ozone 간 상관계수 $r = 0.52$
- pm10와 pm2.5 간 상관계수 $r = 0.51$
- humidity와 windspeed 상관계수 $r = -0.43$

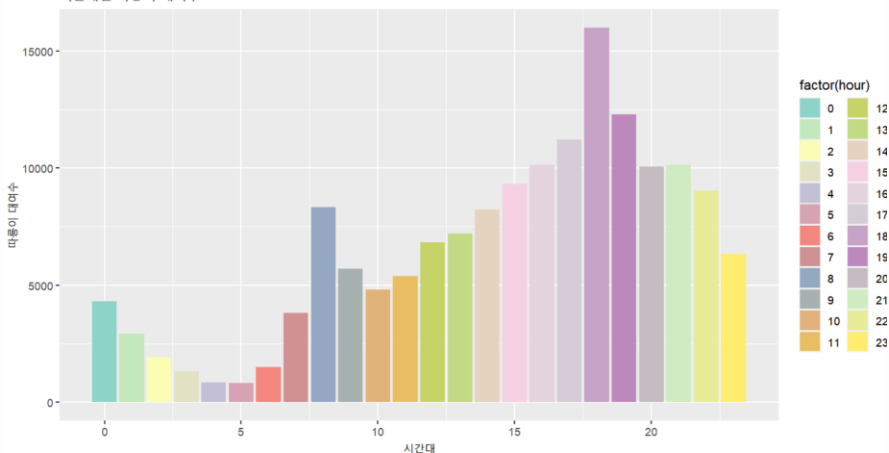
**이상치 검출을 위해 추정할 다중회귀식의 독립변수 선택
(feature selection)**



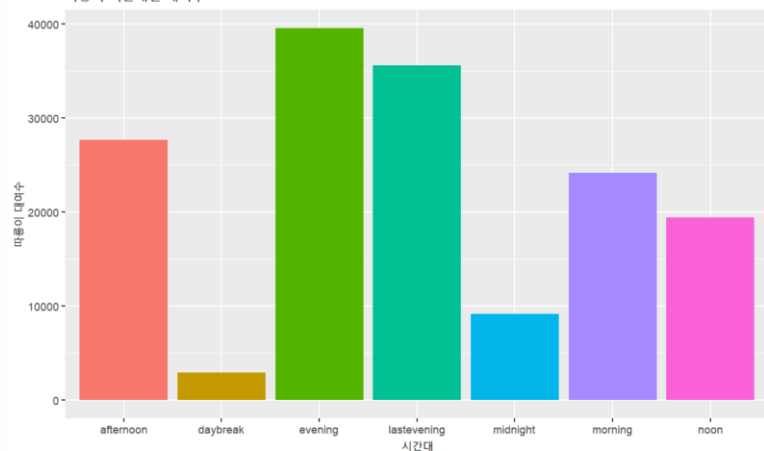
따릉이 데이터 파생변수 생성

시간(hour)에서 시간대(timeZone) 범주형 파생변수 생성

서울특별시 마포구
시간대별 따릉이 대여수



서울특별시 마포구
따릉이 시간대별 대여수



	timeZone	mean.count	count	sum.count
	<fct>	<dbl>	<int>	<dbl>
1	evening	217.	182	39518
2	lastevening	146.	243	35564
3	afternoon	152.	182	27687
4	morning	79.2	305	24153
5	noon	107.	182	19417
6	midnight	50.1	182	9126
7	daybreak	16.0	183	2929

극단적 대기오염물질 농도 여부(atmosphere) 논리형 파생변수 생성

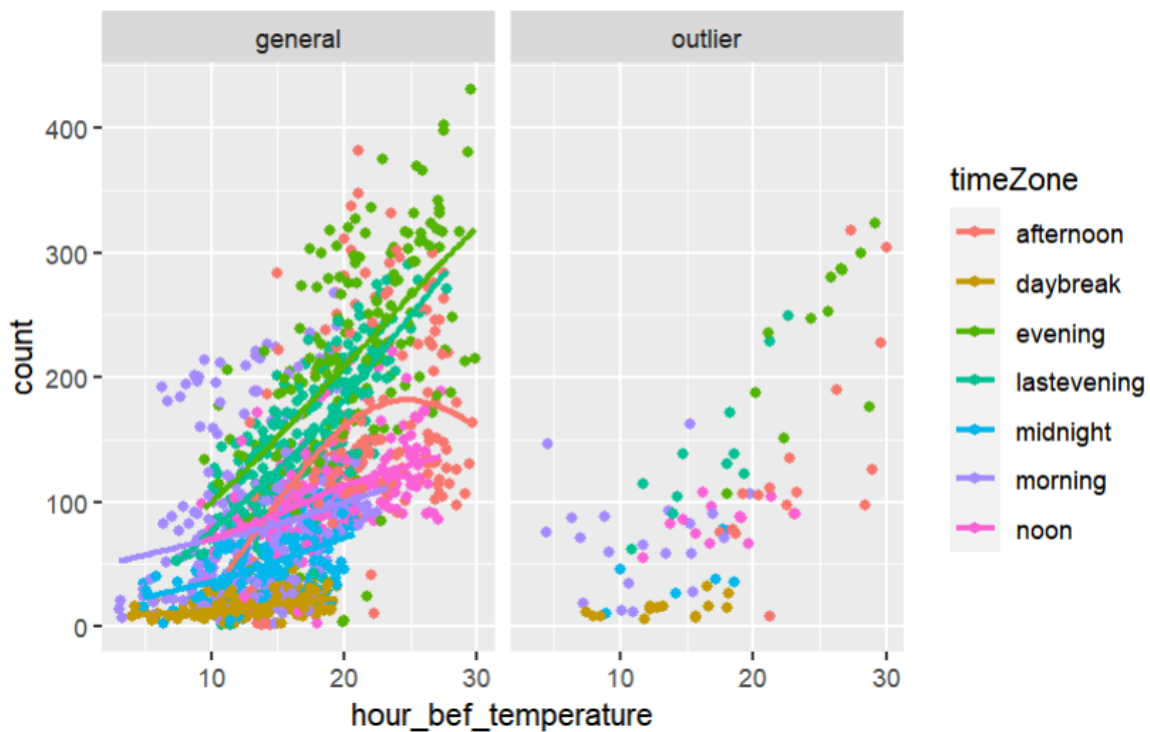
이상치가 자전거 대여수에 영향을 미치는지 예측 목적

```
> table(train$atmosphere)
```

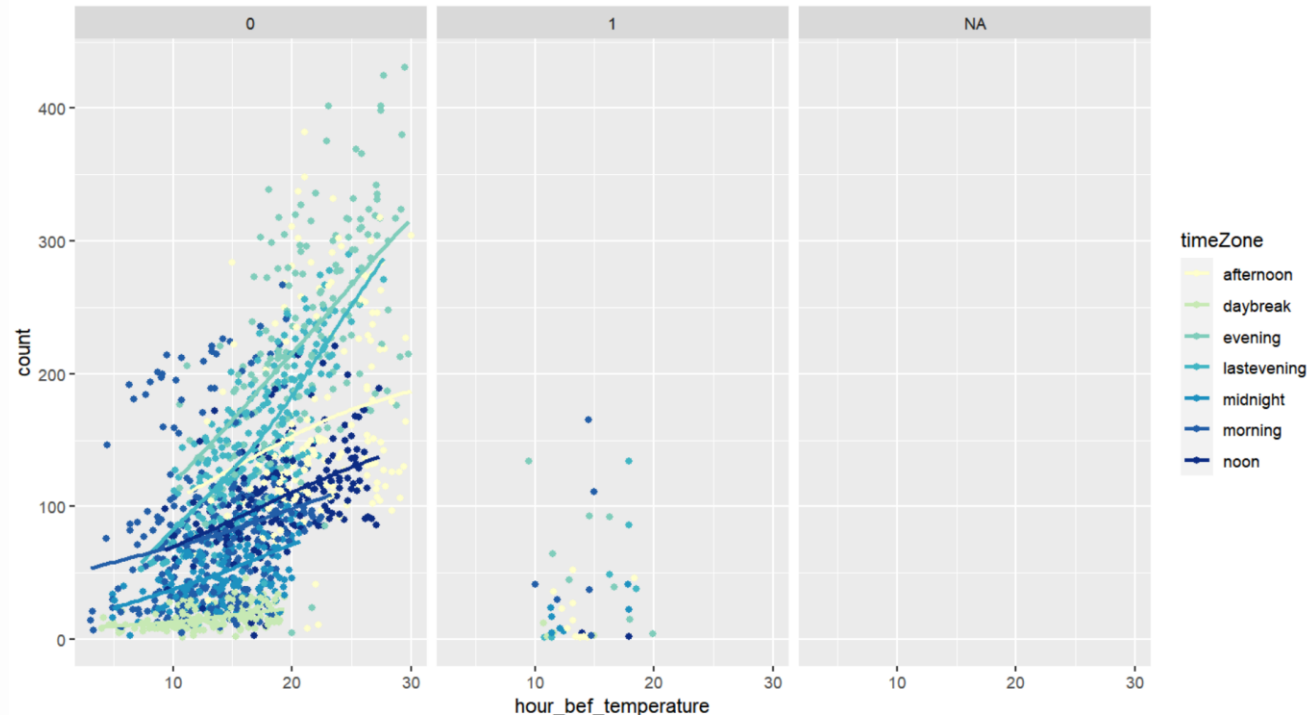
```
general outlier
1365      94
```




따릉이 데이터 시각화 EDA



기온(가로축), 대여수(세로축),
시간대(범주)로 구성된 산점도를
대기오염지수 이상치 여부로 면 분할



기온(가로축), 대여수(세로축),
시간대(범주)로 구성된 산점도를
비가 오는지 여부(0, 1)로 면 분할

서울특별시 마포구 따릉이 데이터 분석 목적



만약 따릉이의 시공간 데이터 수집 기능을 데이터 비즈니스에 활용한다면

- 탄소발자국 감소를 측정하는 데이터 웨어러블 기기, 서울특별시 공공자전거 따릉이
- 신용카드회사, 결제대행기업 등 여신전문금융업계 중 마이데이터 인가 社의 고객에 대한 시공간, 소비, 거래, 생체 데이터 분석 및 활용
- 금융지주 그룹사의 경우 생명보험사에 대하여 데이터 거래하여 생태계 구축