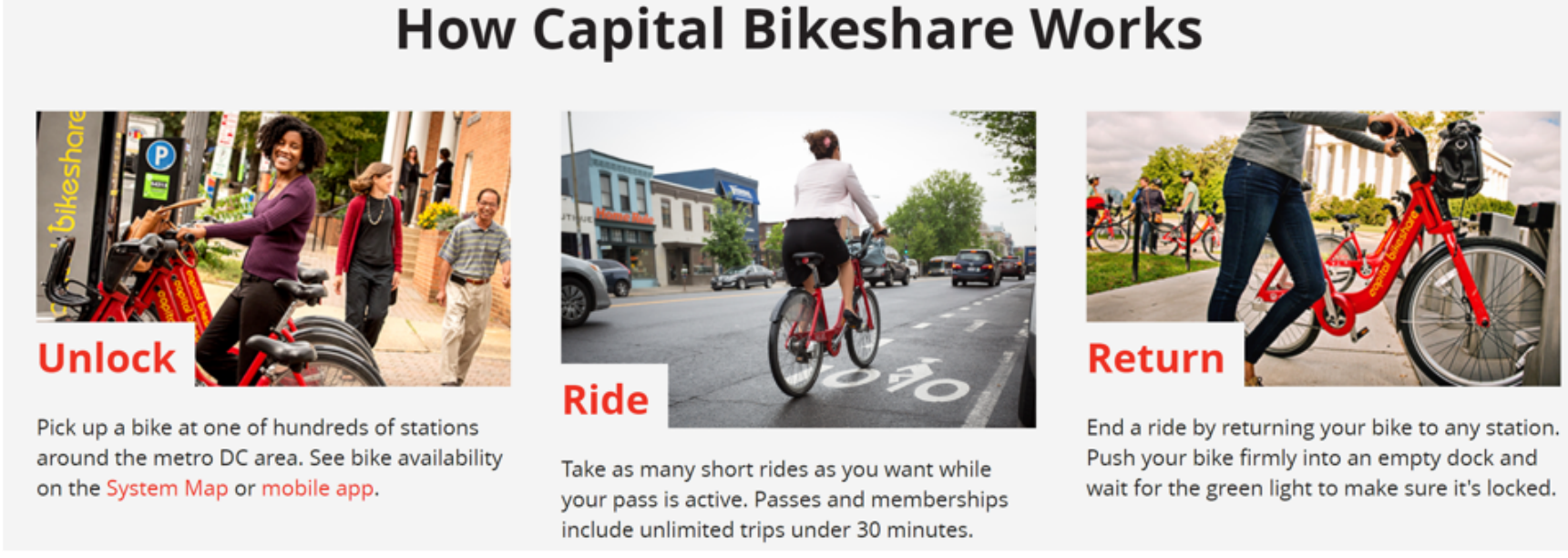
Bike Sharing Demand

현승우

목차

데이터 소개

- 미국 자전거 공유 업체인 Capital Bikeshare의 렌트 정보와
- 워싱턴 DC의 날씨를 결합해서 만든 데이터



데이터 소개

Duration	Start date	End date	Start station number	Start station	End station number	End station	Bike number	Member type
3548	2011-01-01 0:01	2011-01-01 1:00	31620	5th & F St NW	31620	5th & F St NW	W00247	Member
346	2011-01-01 0:02	2011-01-01 0:08	31105	14th & Harvard St NW	31101	14th & V St NW	W00675	Casual
562	2011-01-01 0:06	2011-01-01 0:15	31400	Georgia & New Hampshire Ave NW	31104	Adams Mill & Columbia Rd NW	W00357	Member
434	2011-01-01 0:09	2011-01-01 0:16	31111	10th & U St NW	31503	Florida Ave & R St NW	W00970	Member
233	2011-01-01 0:28	2011-01-01 0:32	31104	Adams Mill & Columbia Rd NW	31106	Calvert & Biltmore St NW	W00346	Casual
158	2011-01-01 0:32	2011-01-01 0:35	31605	3rd & D St SE	31618	4th & East Capitol St NE	W01033	Member
560	2011-01-01 0:35	2011-01-01 0:45	31203	14th & Rhode Island Ave NW	31201	15th & P St NW	W00766	Member
503	2011-01-01 0:36	2011-01-01 0:45	31203	14th & Rhode Island Ave NW	31201	15th & P St NW	W00506	Member
449	2011-01-01 0:45	2011-01-01 0:53	31201	15th & P St NW	31202	14th & R St NW	W00506	Member

데이터 소개

	datetime	season	holiday	workingday	weather	temp	atemp	humidity	windspeed	casual	registered	count
10881	2012-12-19 19:00:00	4	0	1	1	15.58	19.695	50	26.0027	7	329	336
10882	2012-12-19 20:00:00	4	0	1	1	14.76	17.425	57	15.0013	10	231	241
10883	2012-12-19 21:00:00	4	0	1	1	13.94	15.910	61	15.0013	4	164	168
10884	2012-12-19 22:00:00	4	0	1	1	13.94	17.425	61	6.0032	12	117	129
10885	2012-12-19 23:00:00	4	0	1	1	13.12	16.665	66	8.9981	4	84	88

train shape: 10886, 12

test shape: 6493, 9

- test에서는 렌탈 수와 관련된 Casual, Registered, Count가 빠짐

Feature

Feature name	categorical	Description
Datetime	-	날짜 + 시간
Season	1	봄
	2	여름
	3	가을
	4	겨울
Holiday	0, 1	휴일 여부
Workingday	0, 1	일하는 날인지
Weather	1	맑거나 구름이 조금 있음
	2	안개 (구름)
	3	가벼운 눈과 비, 천둥
	4	심한 눈, 안개, 천둥

Feature

Feature name	categorical	Description
Temp	-	기온
Atemp	-	체감기온
Humidity	-	습도
Windspeed	-	바람의 세기
Casual	-	등록되지 않은 회원 수
Registered	-	등록된 회원 수
Count	-	자전거 렌탈 수

초기 설정

```
import pandas as pd
import seaborn as sns
import datetime
import statsmodels.api as sm
from statsmodels.graphics import utils
from sklearn.model_selection import KFold
%matplotlib inline
mpl.rc('figure', figsize=(8, 5))
mpl.rc('figure', dpi=100)
```

EDA

```
train.datetime = pd.to_datetime(train.datetime)
test.datetime = pd.to_datetime(test.datetime)
train['year'] = train.datetime.apply(lambda x: x.year)
test['year'] = test.datetime.apply(lambda x: x.year)
train['month'] = train.datetime.apply(lambda x: x.month)
test['month'] = test.datetime.apply(lambda x: x.month)
train['day'] = train.datetime.apply(lambda x: x.day)
test['day'] = test.datetime.apply(lambda x: x.day)
train['hour'] = train.datetime.apply(lambda x: x.hour)
test['hour'] = test.datetime.apply(lambda x: x.hour)
train['weekday'] = train.datetime.apply(lambda x: x.weekda
test['weekday'] = test.datetime.apply(lambda x: x.weekday
```

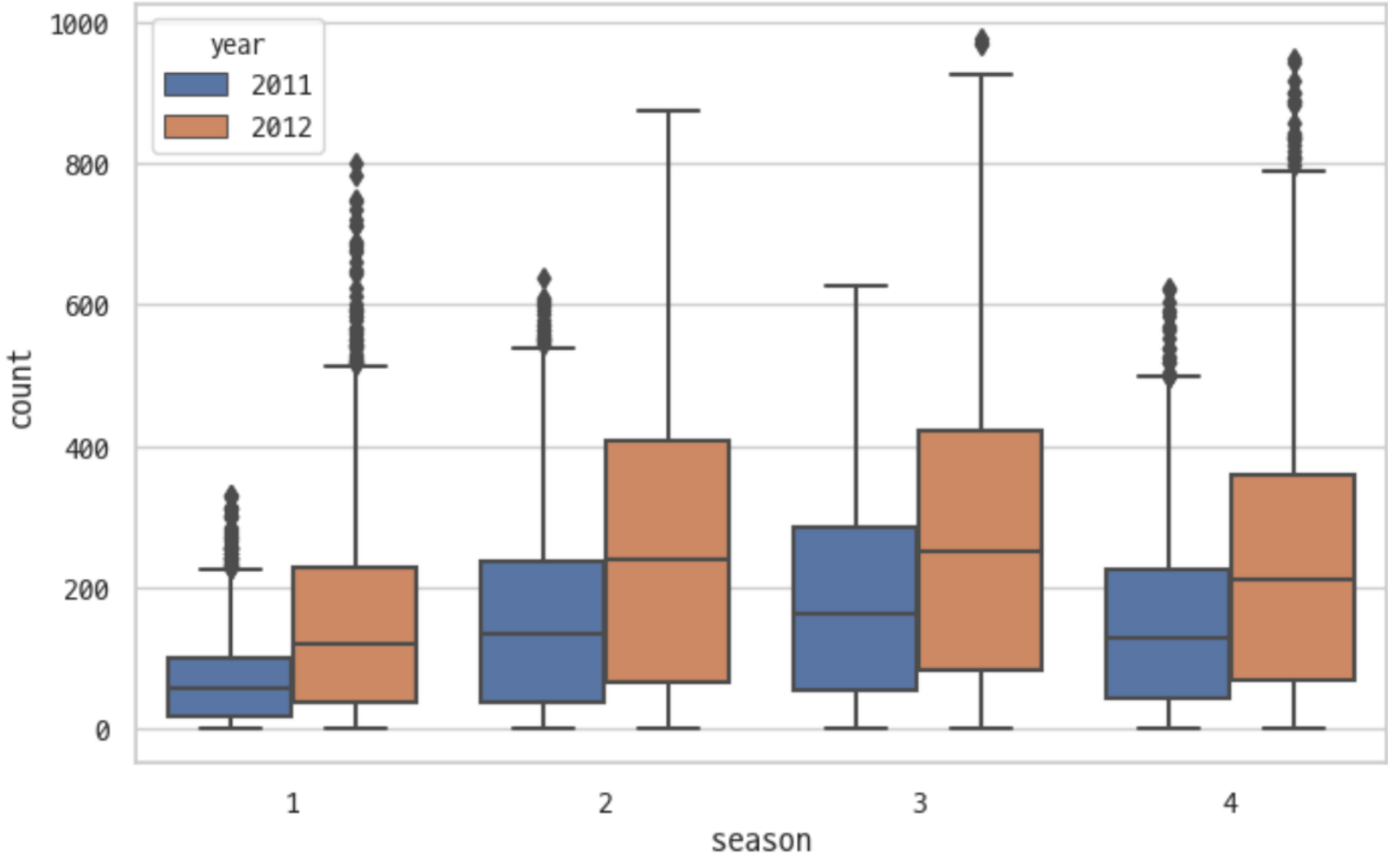
- 문자열 형식인 datetime을 카테고리칼 변수로 만들어주는 코드

```
train.isnull().sum()
```

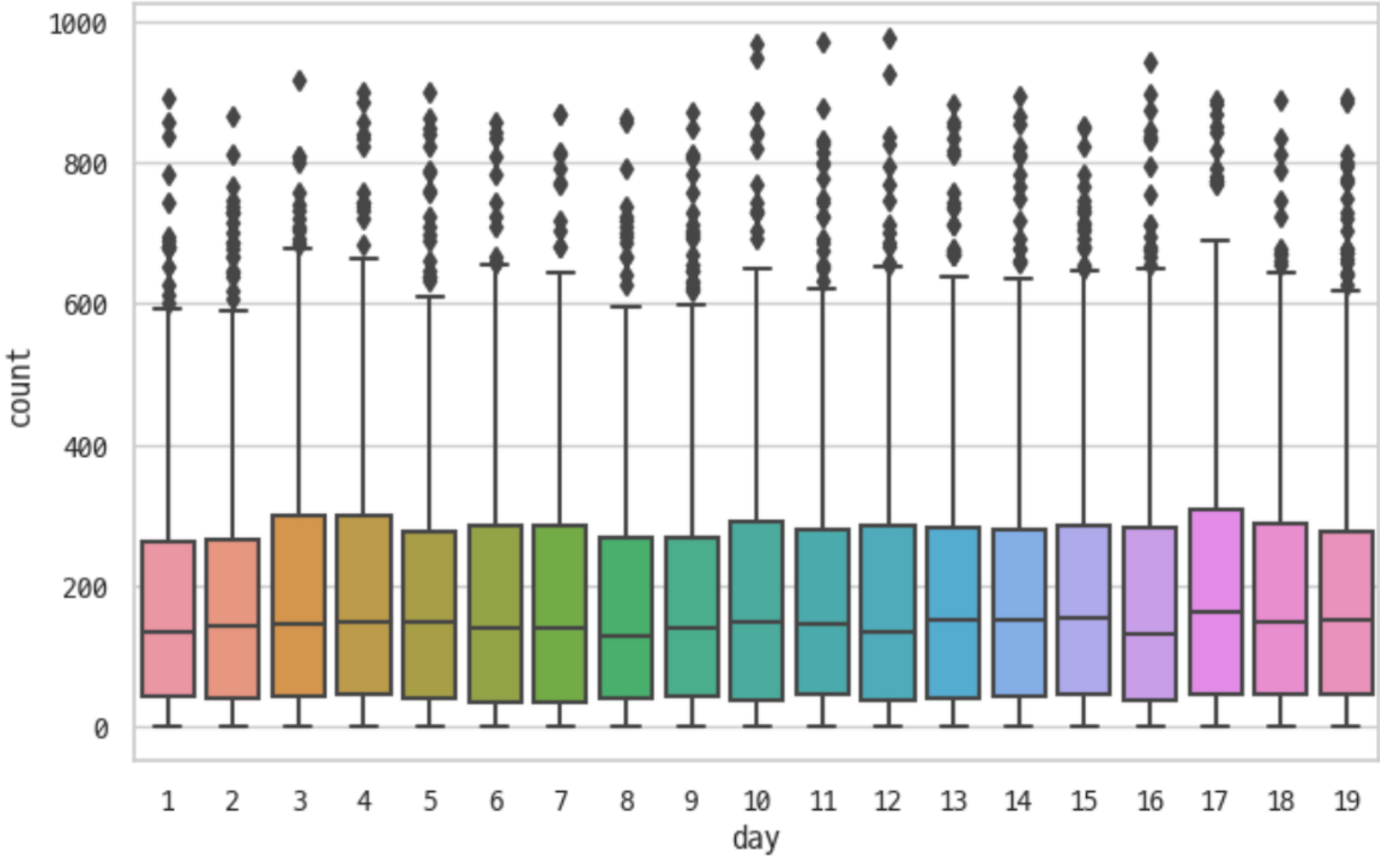
- null 변수는 없음

```
datetime      0
season        0
holiday       0
workingday    0
weather       0
temp          0
atemp         0
humidity      0
windspeed     0
casual        0
registered    0
count         0
day           0
weekday       0
hour          0
year          0
month         0
dtype: int64
```

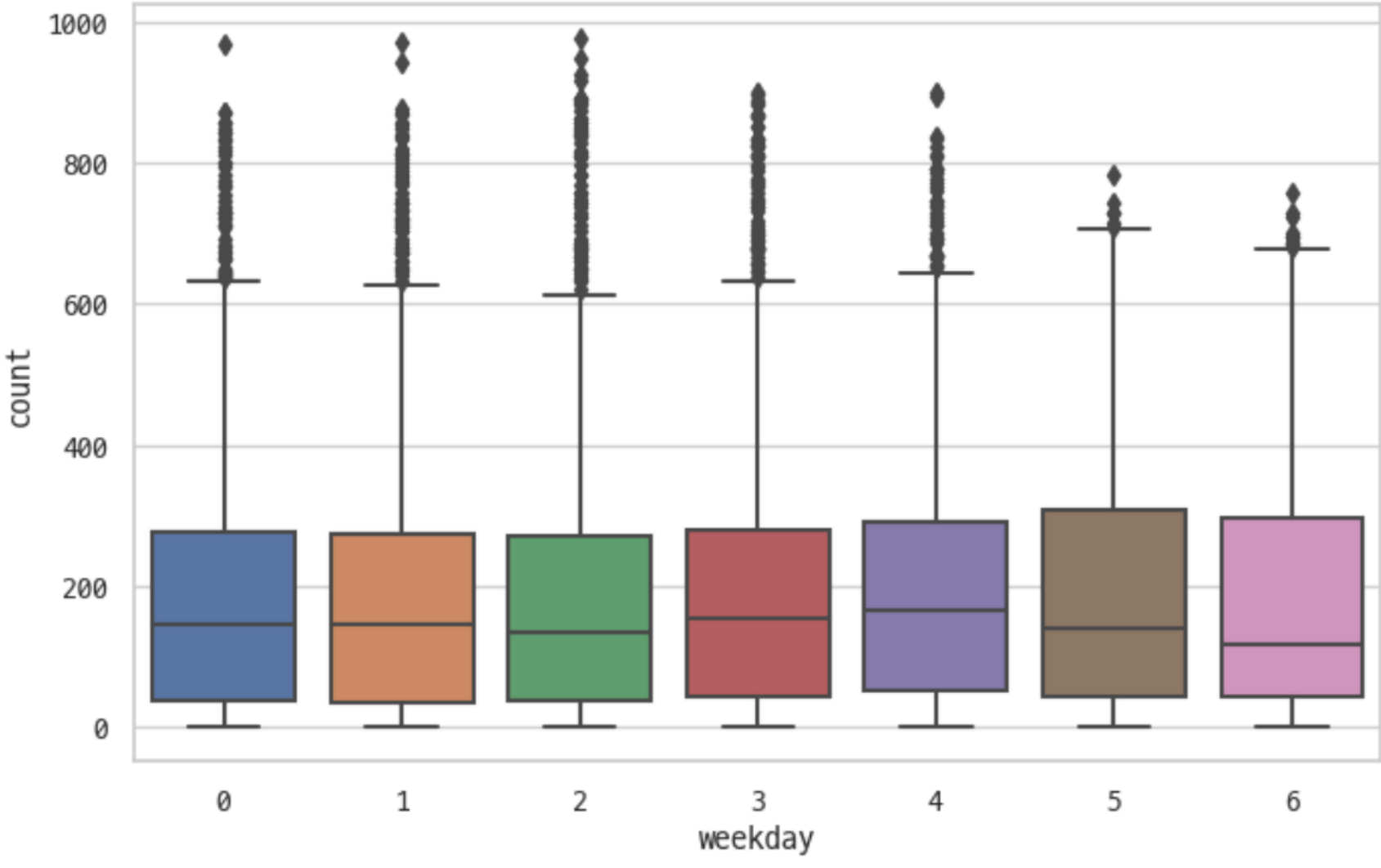
season과 year에 따른 count의 (boxplot)



day별 count수 (boxplot)

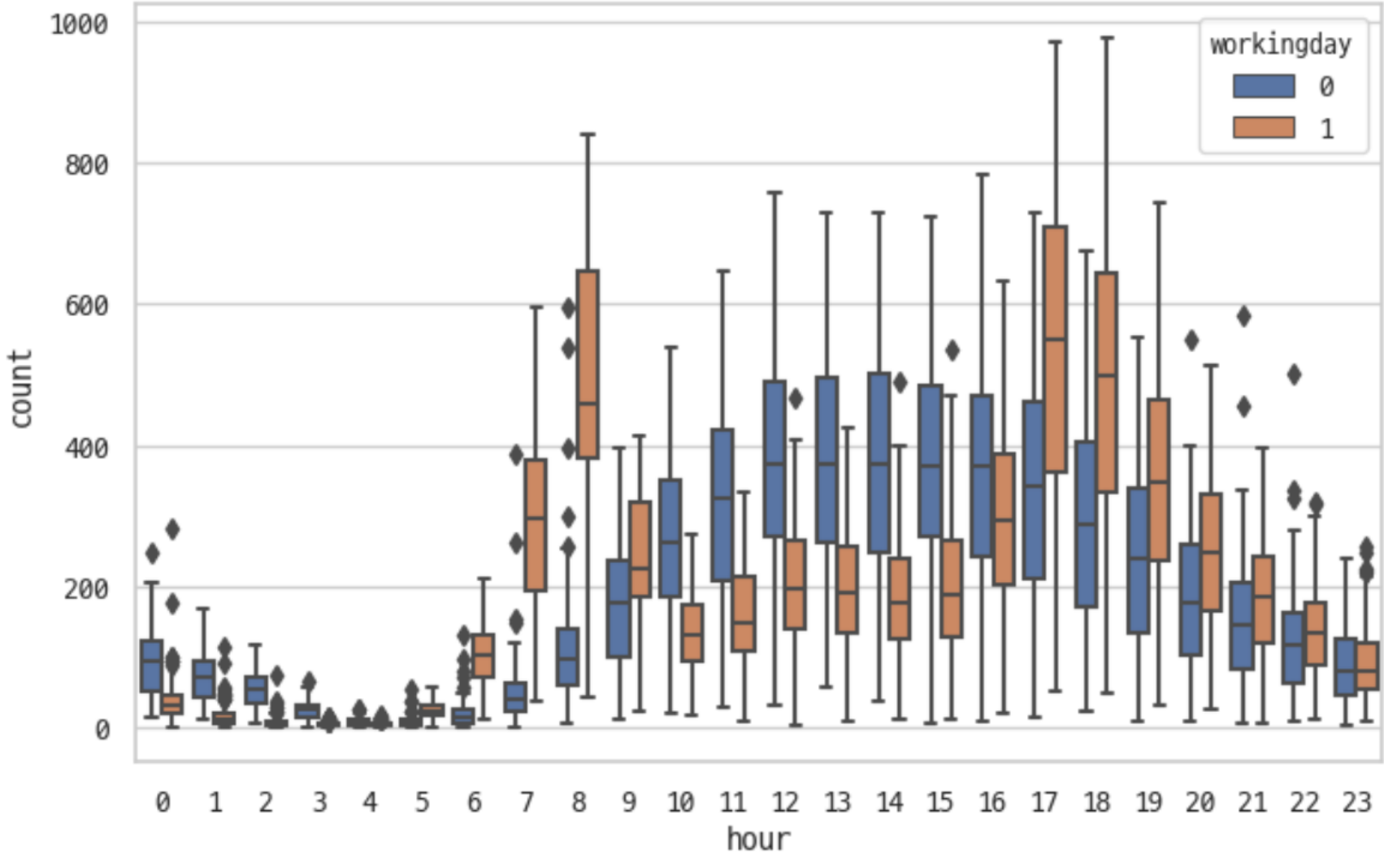


weekday별 count수 (boxplot)



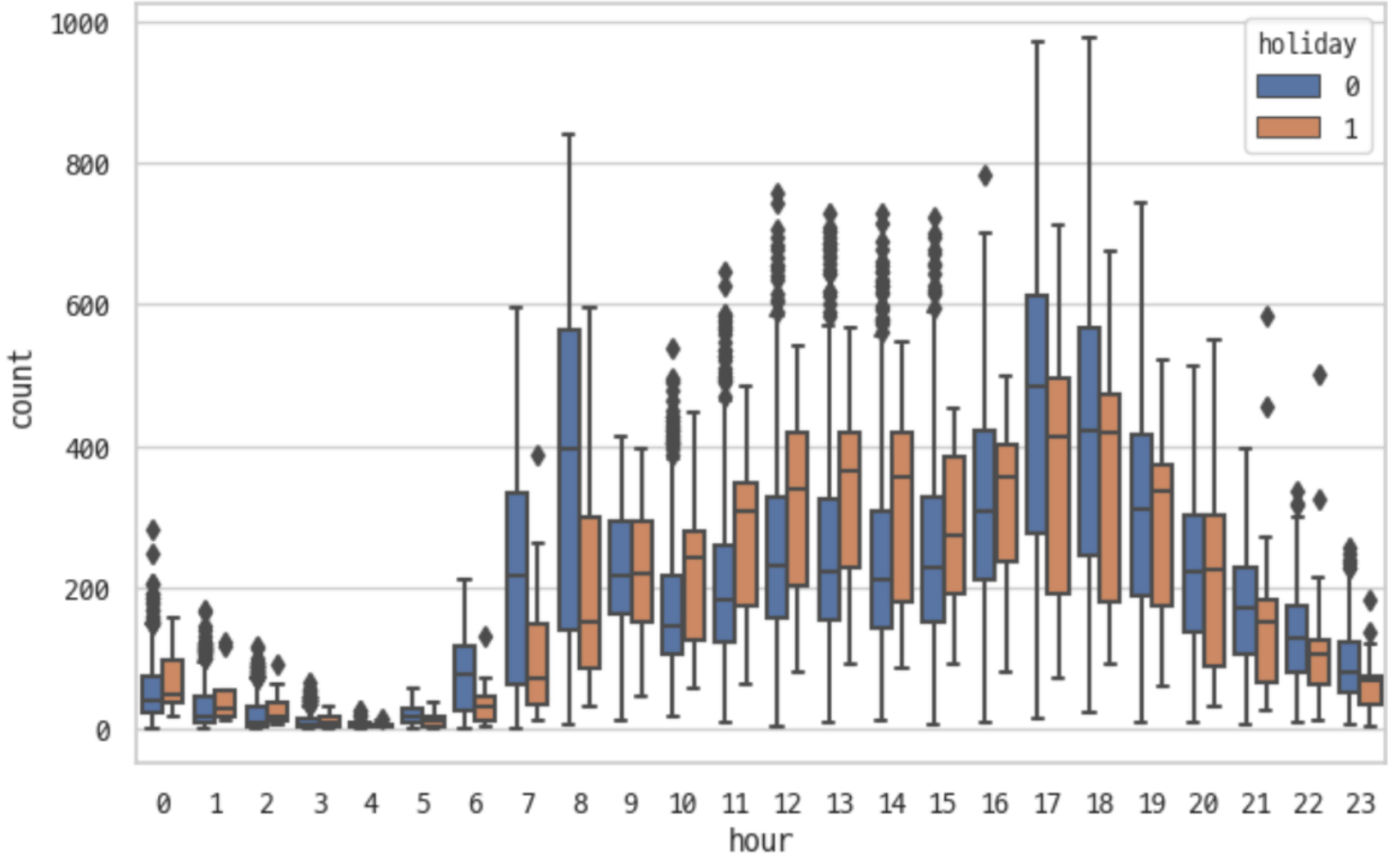
hour별 count수 (boxplot)

- workingday 기준



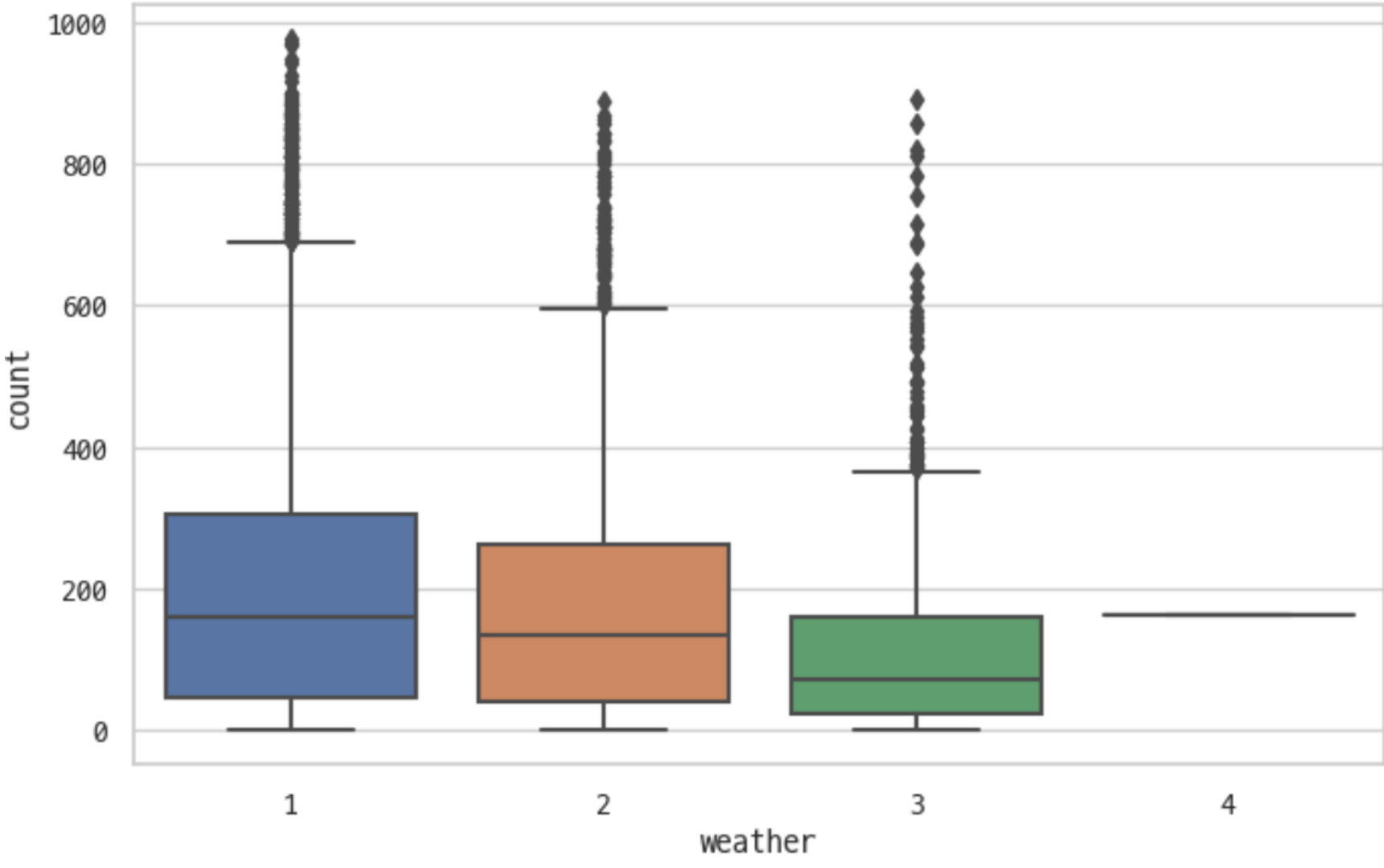
hour별 count수 (boxplot)

- holiday 기준



- 위의 두 그래프로 working day일 때는 출퇴근 시간에 사람들의 자전거 이용이 많았고
- 일하지 않는 날에는 오후 시간대에 자전거 이용량이 많았다
- 또한 holiday와 그래프의 분포가 거의 비슷하고 holiday의 정보가 workingday와 weekday에 포함되므로
- holiday 변수를 OLS 모델에 포함시키지 않는다.

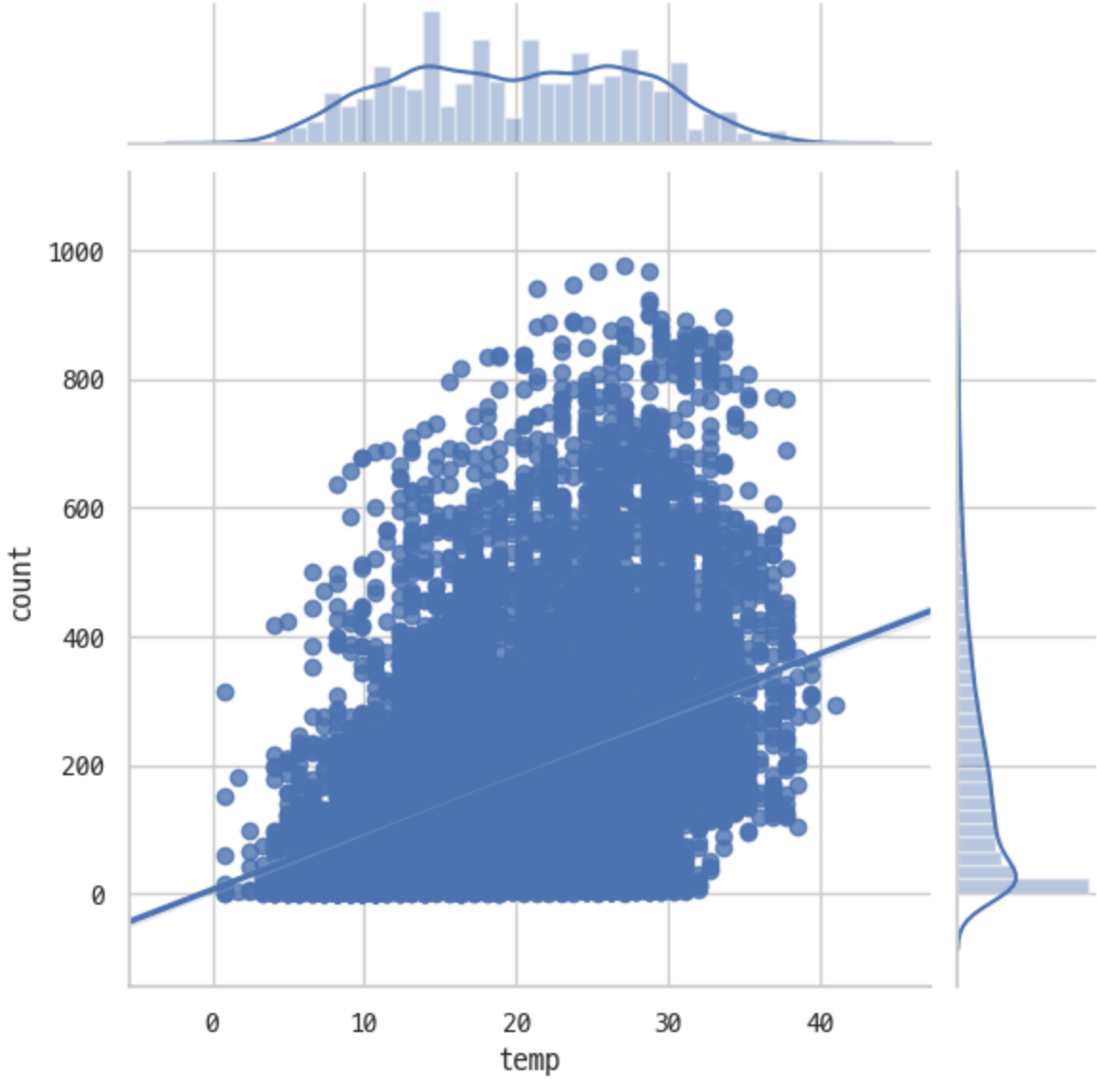
weather에 따른 count수 (countplot)



- 날씨가 제일 안좋은 날인 4번은 train 데이터에 1개 들어있음 (나중에 교차 검증 때 문제가 되었음)

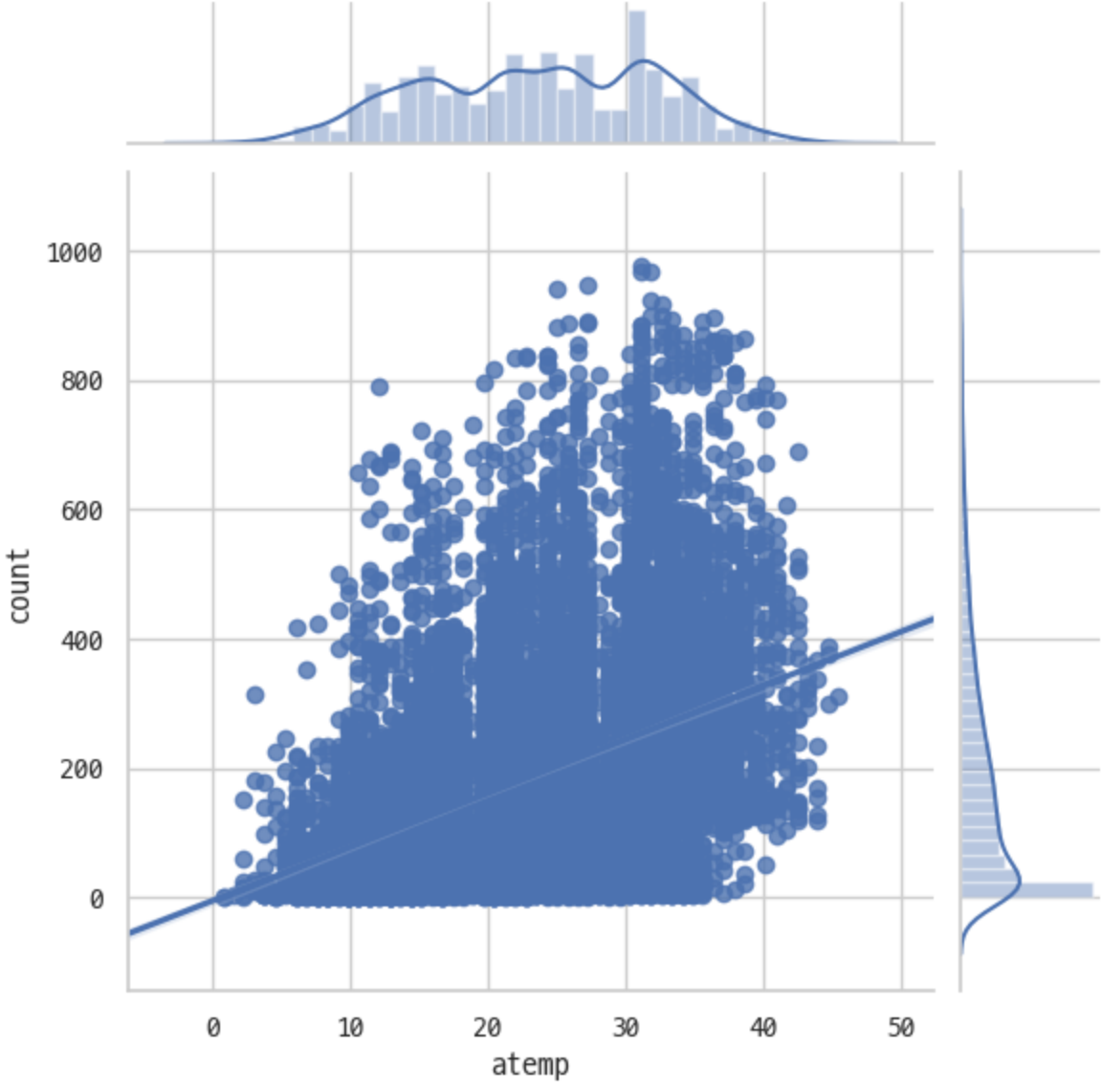
temp와 count의 jointplot

- temp와 count는 0.4의 상관관계를 가짐

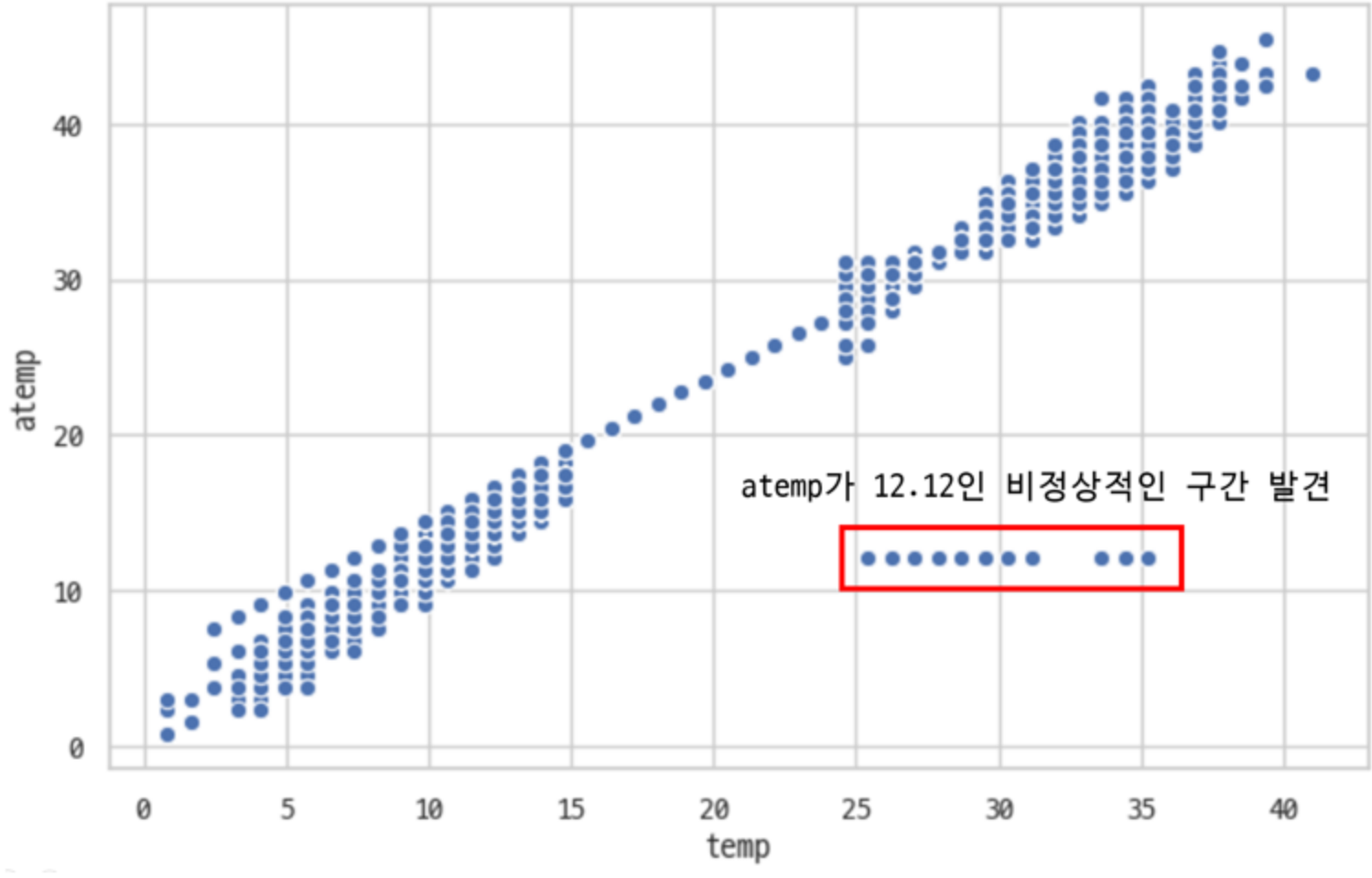


atemp와 count의 jointplot

- atemp와 count는 0.39의 상관관계를 가짐



atemp와 temp의 scatter plot



- atemp가 12.12인 195개의 데이터의 체감온도를 계산하여 넣음

체감온도 계산

- 현재 사용하고 있는 체감온도 산출식은 2001년 8월 캐나다 토론토에서 열린 **Joint Action Group for Temperature Indices(JAG/TI)회의**에서 발표된 것으로 미국과 캐나다 등 북아메리카 국가들을 중심으로 **최근에 가장 널리 사용되고 있음**

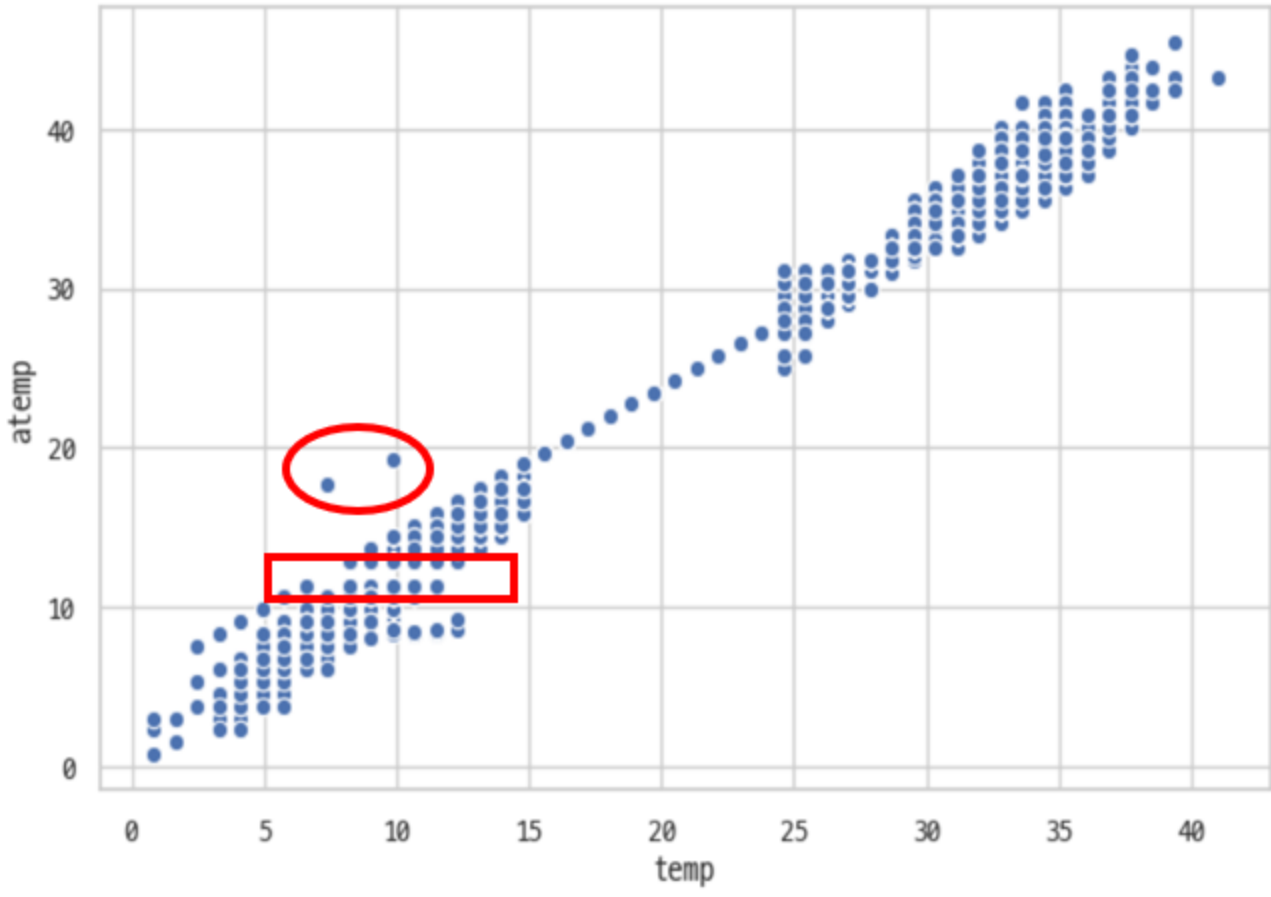
$$\text{atemp} = 13.12 + 0.625T - 11.37V^{0.16} + 0.3965V^{0.16}T$$

$$T : \text{temp (celcius)}$$

$$V : \text{windspeed (km/h)}$$

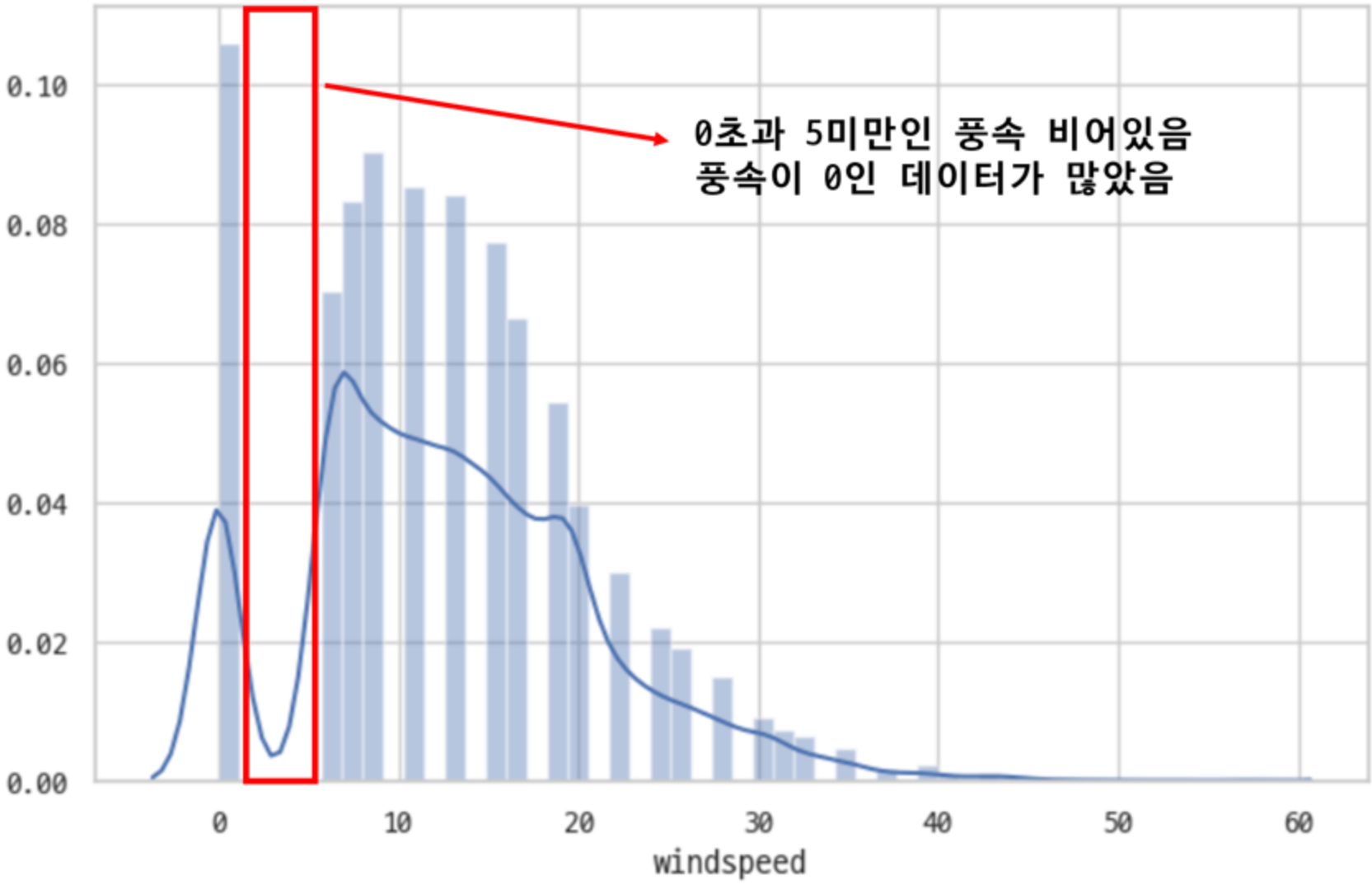
- 체감온도 산출법

atemp와 temp의 scatter plot



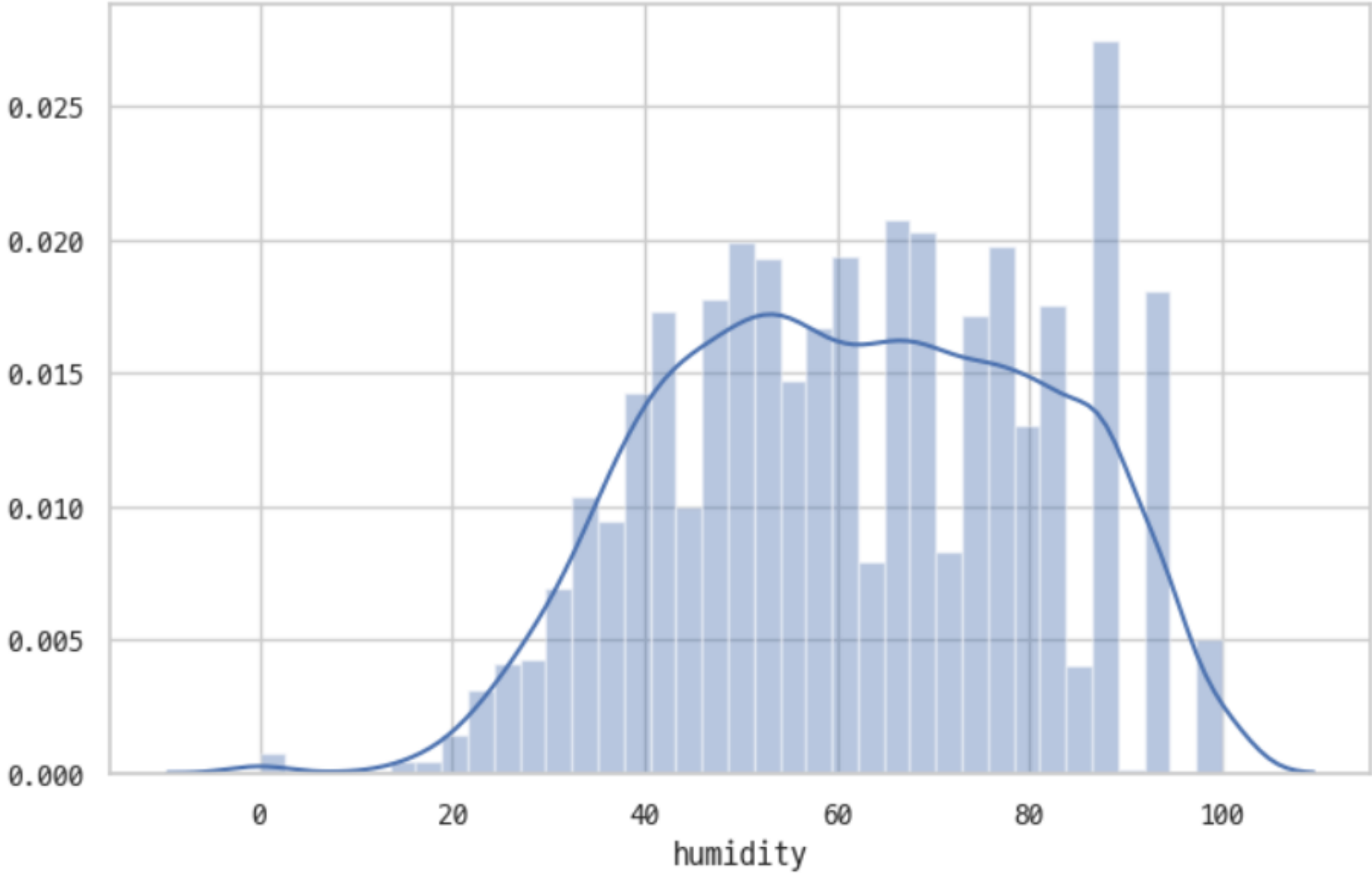
- atemp가 12.12인 데이터가 모두 바뀜
- 체감 온도와 기온의 상관관계가 0.99므로 atemp만 OLS 모델에 포함시킨다.

바람세기



- 체감온도 계산법에 바람세기가 들어가므로 바람세기를 OLS 모델에 추가하지 않는다.

습도의 distplot



- 0인 습도는 존재할 수 없다.

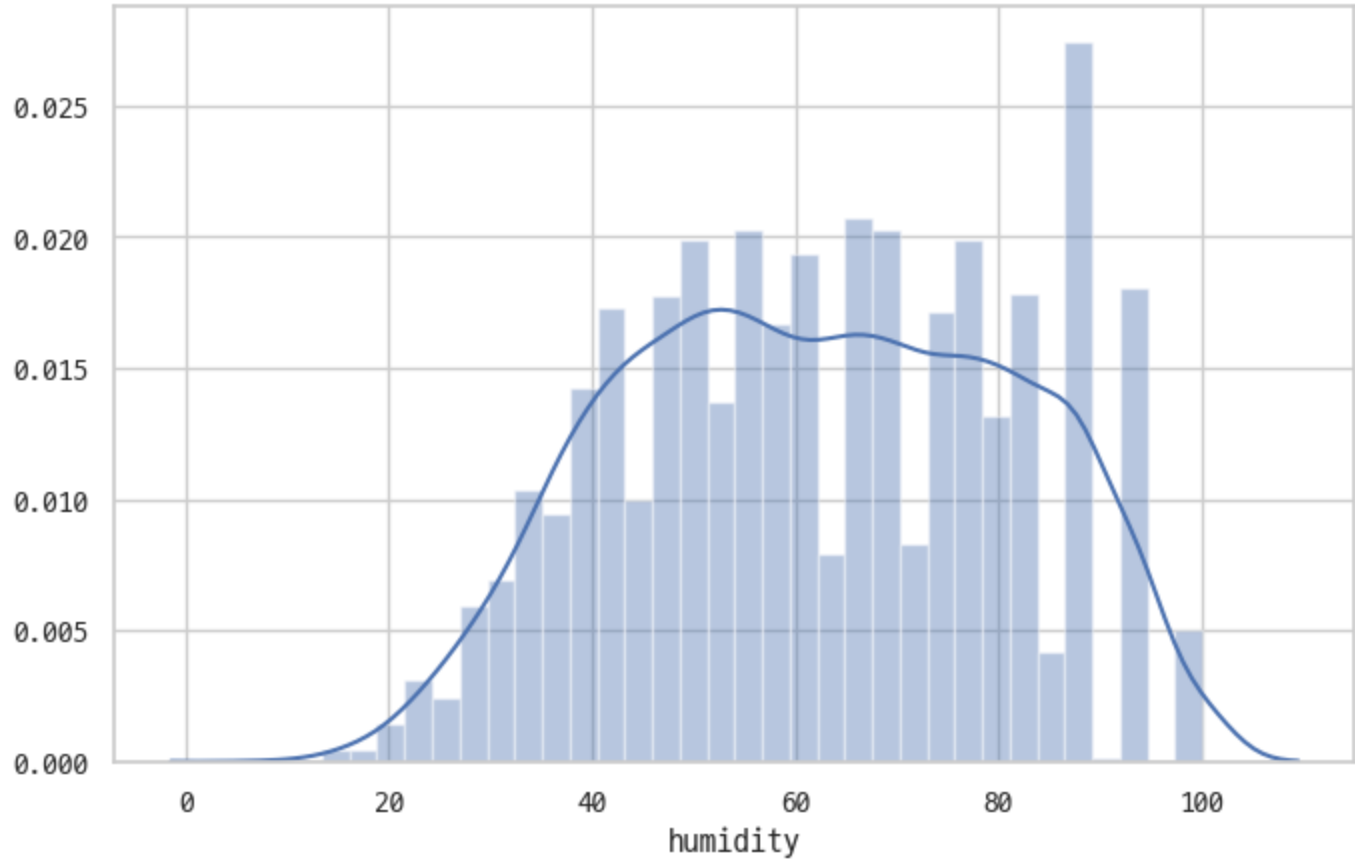
습도가 0인 데이터 처리

- 2011년 3월 10일 22개의 데이터가 0
- 3월 10일과 같은 날씨 3의 시간별 습도 평균을 각각 구하여 데이터에 넣어 주었다.

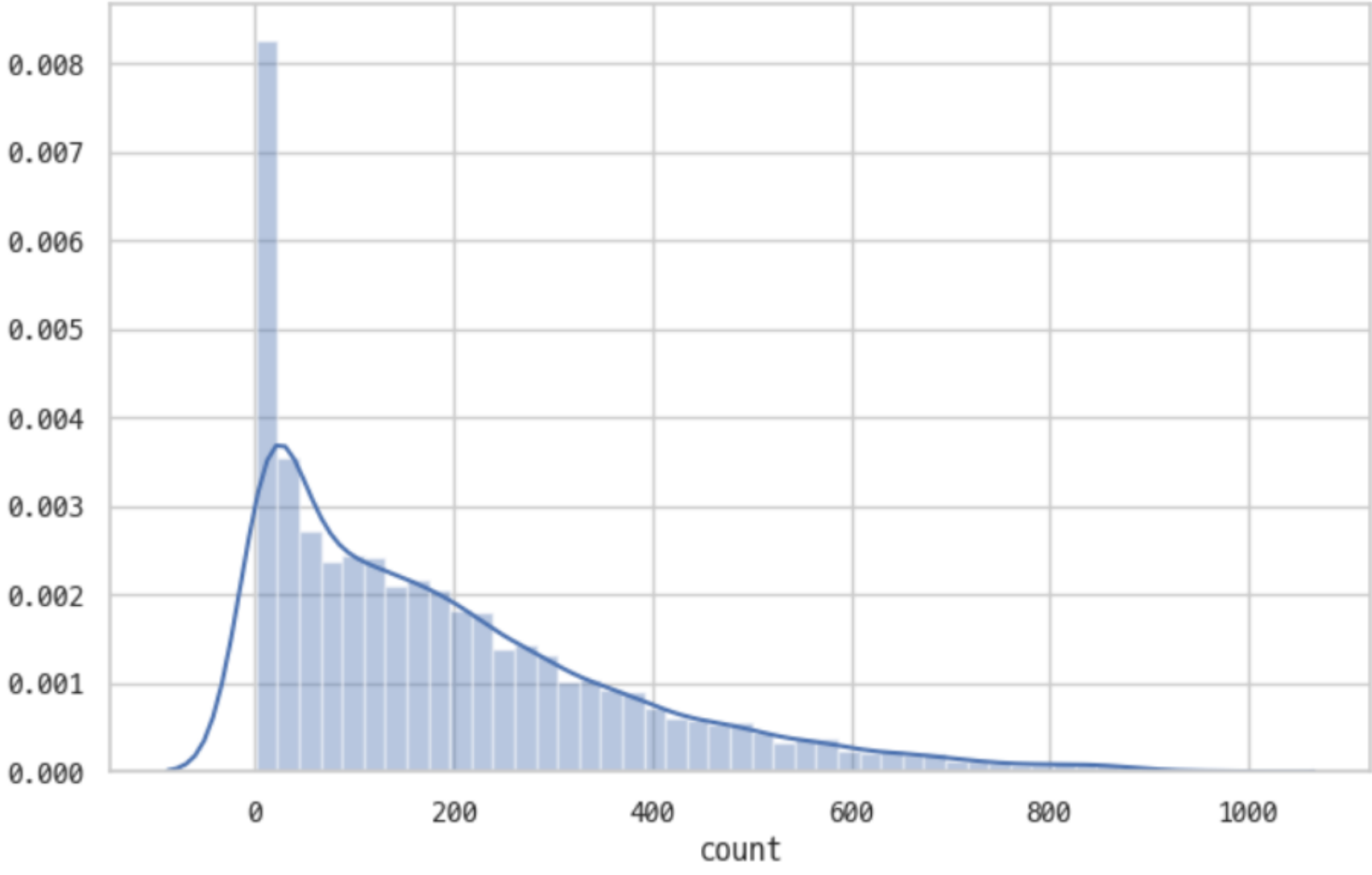
습도가 0인 데이터 처리

시간	원래 3월 10일의 습도 데이터	날씨가 3인 데이터의 시간별 평균 습도
0	0	82.821429
1	0	82.473684
2	0	84.156250
3	존재하지 않음	86.733333
4	존재하지 않음	86.083333
5	0	84.121212
6	0	83.642857
7	0	84.697674
8	0	83.378378
9	0	81.875000
10	0	80.718750
11	0	80.538462
12	0	78.448276
13	0	76.296296
14	0	76.529412
15	0	77.095238
16	0	79.125000
17	0	77.183673
18	0	75.723404
19	0	82.416667
20	0	82.447368
21	0	80.562500
22	0	85.181818
23	0	82.372093

처리 후 습도의 distplot

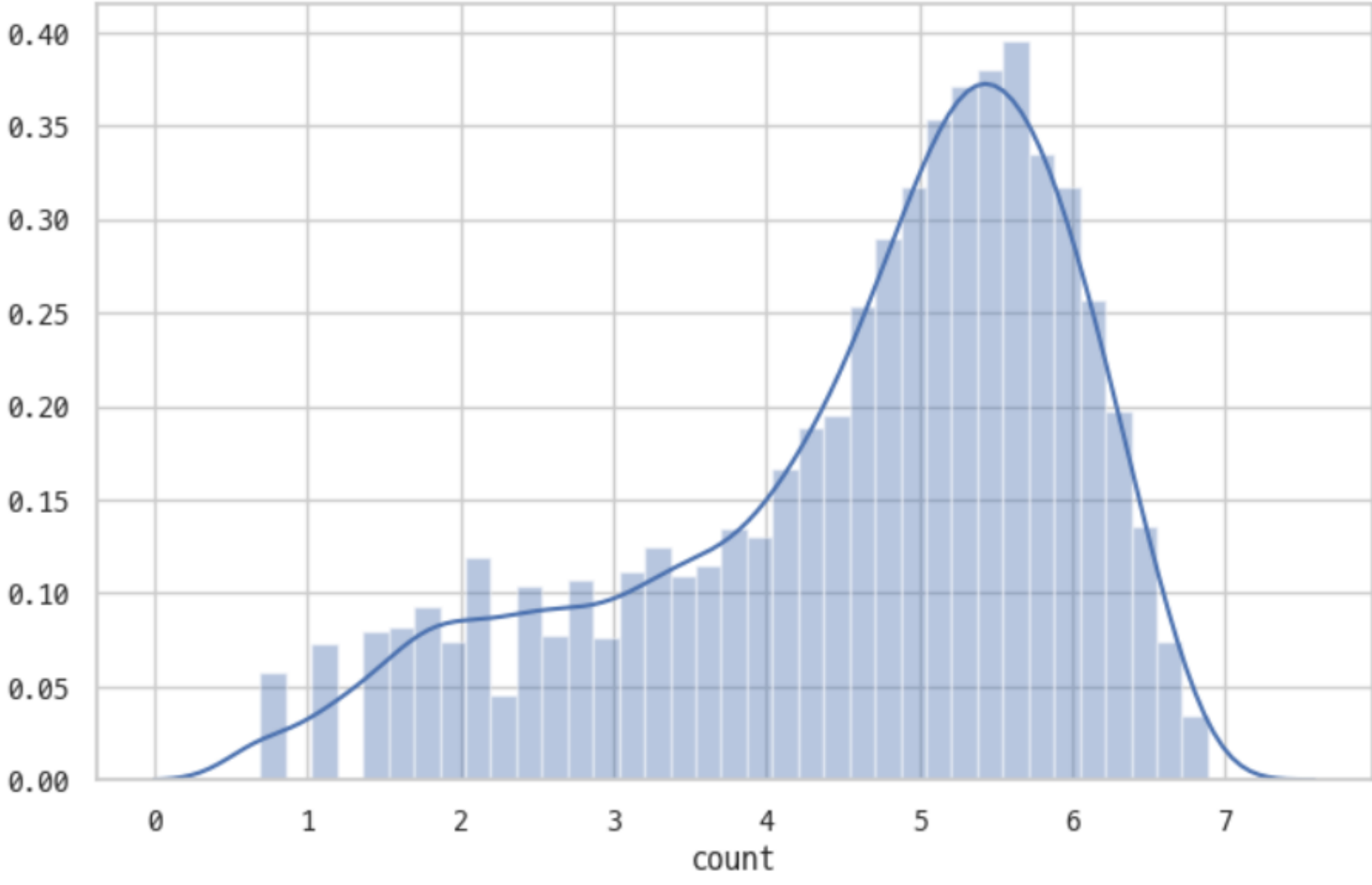


count의 distplot



- 분포가 왼쪽으로 치우쳐있어 log를 줘서 정규분포 형태에 가깝게 만들어 준다.

np.log1p(count)의 distplot



OLS.from_formula

변수 설명	변수
종속변수 Y	np.log1p(count)
독립변수 X1	C(season)
독립변수 X2	C(workingday)
독립변수 X3	scale(atemp)
독립변수 X4	scale(humidity)
독립변수 X5	C(weekday)
독립변수 X6	C(weather)
독립변수 X7	C(hour)
독립변수 X8	C(month)
독립변수 X9	C(year)
상수항 제거	+ 0

```
* model = sm.OLS.from_formula('np.log1p(count) ~ C(season) + C(workingday) +
scale(atemp) + scale(humidity) + C(weekday) + C(weather) + C(hour) + C(month) +
C(year) + 0', data=train)
```

OLS Summary

OLS Regression Results

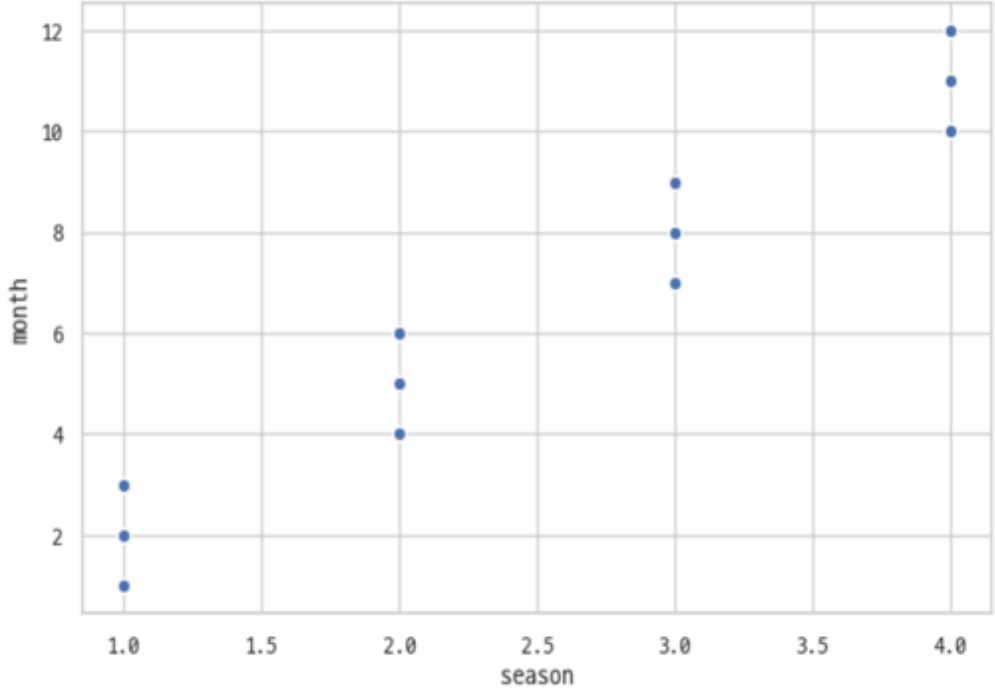
Dep. Variable:	np.log1p(count)	R-squared:	0.835
Model:	OLS	Adj. R-squared:	0.834
Method:	Least Squares	F-statistic:	1163.
Date:	Wed, 12 Dec 2018	Prob (F-statistic):	0.00
Time:	19:26:57	Log-Likelihood:	-9467.9
No. Observations:	10886	AIC:	1.903e+04
Df Residuals:	10838	BIC:	1.938e+04
Df Model:	47		
Covariance Type:	nonrobust		

Omnibus:	769.045	Durbin-Watson:	0.547
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1531.095
Skew:	-0.490	Prob(JB):	0.00
Kurtosis:	4.554	Cond. No.	7.62e+15

- 조건수가 7.62e + 15로 굉장히 높다.
- 수정.

season과 month의 상관관계

- 당연히 month가 있으면 season은 넣어줄 필요가 없다는 것을 생각하지 못했다.



	season	month
season	1.000000	0.971524
month	0.971524	1.000000

수정한 OLS

OLS Regression Results

Dep. Variable:	np.log1p(count)	R-squared:	0.834
Model:	OLS	Adj. R-squared:	0.833
Method:	Least Squares	F-statistic:	1159.
Date:	Sat, 15 Dec 2018	Prob (F-statistic):	0.00
Time:	17:17:06	Log-Likelihood:	-9482.1
No. Observations:	10886	AIC:	1.906e+04
Df Residuals:	10838	BIC:	1.941e+04
Df Model:	47		
Covariance Type:	nonrobust		
Omnibus:	777.940	Durbin-Watson:	0.547
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1547.143
Skew:	-0.495	Prob(JB):	0.00
Kurtosis:	4.559	Cond. No.	117.

- 조건수가 117로 크게 줄었다.

계수 부분

	coef	std err	t	P>|t|	[0.025	0.975]
C(workingday)[0]	3.0579	0.050	61.396	0.000	2.960	3.155
C(workingday)[1]	3.0196	0.042	72.462	0.000	2.938	3.101
C(month)[T.2]	0.1648	0.028	5.919	0.000	0.110	0.219
C(month)[T.3]	0.2445	0.029	8.302	0.000	0.187	0.302
C(month)[T.4]	0.4458	0.031	14.248	0.000	0.384	0.507
C(month)[T.5]	0.6740	0.035	19.182	0.000	0.605	0.743
C(month)[T.6]	0.6272	0.039	16.018	0.000	0.550	0.704
C(month)[T.7]	0.5360	0.043	12.398	0.000	0.451	0.621
C(month)[T.8]	0.5534	0.042	13.163	0.000	0.471	0.636
C(month)[T.9]	0.6444	0.038	16.962	0.000	0.570	0.719
C(month)[T.10]	0.7568	0.033	22.597	0.000	0.691	0.822
C(month)[T.11]	0.7156	0.029	24.676	0.000	0.659	0.772
C(month)[T.12]	0.6741	0.029	23.340	0.000	0.618	0.731
C(weather)[T.2]	-0.0549	0.014	-3.990	0.000	-0.082	-0.028
C(weather)[T.3]	-0.5416	0.023	-23.474	0.000	-0.587	-0.496
C(weather)[T.4]	-0.0816	0.581	-0.140	0.888	-1.220	1.057
C(weekday)[T.1]	-0.0066	0.022	-0.305	0.760	-0.049	0.036
C(weekday)[T.2]	0.0062	0.021	0.288	0.773	-0.036	0.048
C(weekday)[T.3]	0.0685	0.022	3.173	0.002	0.026	0.111
C(weekday)[T.4]	0.1585	0.021	7.409	0.000	0.117	0.200
C(weekday)[T.5]	0.1312	0.037	3.571	0.000	0.059	0.203
C(weekday)[T.6]	0.0240	0.037	0.654	0.513	-0.048	0.096
C(hour)[T.1]	-0.6044	0.038	-15.719	0.000	-0.680	-0.529
C(hour)[T.2]	-1.1001	0.039	-28.503	0.000	-1.176	-1.024
C(hour)[T.3]	-1.5941	0.039	-40.909	0.000	-1.671	-1.518
C(hour)[T.4]	-1.8621	0.039	-47.991	0.000	-1.938	-1.786
C(hour)[T.5]	-0.9009	0.039	-23.337	0.000	-0.977	-0.825
C(hour)[T.6]	0.2789	0.039	7.231	0.000	0.203	0.354
C(hour)[T.7]	1.2466	0.038	32.381	0.000	1.171	1.322
C(hour)[T.8]	1.8713	0.038	48.688	0.000	1.796	1.947
C(hour)[T.9]	1.5419	0.038	40.097	0.000	1.467	1.617
C(hour)[T.10]	1.2190	0.039	31.589	0.000	1.143	1.295
C(hour)[T.11]	1.3369	0.039	34.404	0.000	1.261	1.413
C(hour)[T.12]	1.5242	0.039	38.928	0.000	1.447	1.601
C(hour)[T.13]	1.4976	0.039	37.943	0.000	1.420	1.575
C(hour)[T.14]	1.4151	0.040	35.647	0.000	1.337	1.493
C(hour)[T.15]	1.4741	0.040	37.079	0.000	1.396	1.552
C(hour)[T.16]	1.7344	0.040	43.746	0.000	1.657	1.812
C(hour)[T.17]	2.1461	0.039	54.419	0.000	2.069	2.223
C(hour)[T.18]	2.0618	0.039	52.580	0.000	1.985	2.139
C(hour)[T.19]	1.7713	0.039	45.623	0.000	1.695	1.847
C(hour)[T.20]	1.4724	0.039	38.099	0.000	1.397	1.548
C(hour)[T.21]	1.2170	0.039	31.606	0.000	1.142	1.293
C(hour)[T.22]	0.9753	0.038	25.373	0.000	0.900	1.051
C(hour)[T.23]	0.5832	0.038	15.182	0.000	0.508	0.658
C(year)[T.2012]	0.4824	0.011	42.713	0.000	0.460	0.505
scale(atemp)	0.2179	0.012	17.578	0.000	0.194	0.242
scale(humidity)	-0.0501	0.008	-6.626	0.000	-0.065	-0.035

교차 검증

- weather 4인 데이터가 1개 밖에 존재하지 않기 때문에
- 그 데이터가 test데이터에 들어갈 때 문제가 될 수 있다.
- 그렇기 때문에 행 번호 5631의 데이터를 삭제한 후 교차검증을 실시 하였다.

```
학습 R2 = 0.83266068, 검증 R2 = 0.72814599
학습 R2 = 0.83666127, 검증 R2 = 0.70220668
학습 R2 = 0.83532784, 검증 R2 = 0.72562044
학습 R2 = 0.83735682, 검증 R2 = 0.71543570
학습 R2 = 0.82934134, 검증 R2 = 0.74183288
```

결론 및 아쉬운 점

- R-square 결과 0.835가 나왔는데 더이상 올리지 못함
- 변수 간의 상관관계를 찾아내어 모델에 반영하지 못함
- VIF를 실행하지 않아 조건수를 내리지 못함 (season과 month의 상관관계를 간과 하였다)
- 대체 할 데이터의 filtering에 실수가 있던 점 (atemp 12.12)

감사합니다.