스프린트 미션 3

개요

분석 목표

💡 1. 자전거 대여 패턴을 분석하여 자전거 배치 및 운영 전략을 최적화하고, 대여 수요를 정확히 예측하는 것
2. 대여 시스템의 효율성을 높이고 사용자 만족도를 증가시키는 방법을 찾는 것
3. RMSLE (Root Mean Squared Logarithmic Error)를 최대한 낮추는 것

데이터 셋

자전거 대여 시스템

📂 test.csv

train.csv

데이터 정보

컬럼명	데이터 타입	설명
datetime	datetime	자전거 대여 기록의 날짜 및 시간. 예시: 2011-01-01 00:00:00
season	int	계절 (1: 봄, 2: 여름, 3: 가을, 4: 겨울)
holiday	int	공휴일 여부 (0: 평일, 1: 공휴일)
workingday	int	근무일 여부 (0: 주말/공휴일, 1: 근무일)
weather	int	날씨 상황 (1: 맑음, 2: 구름낌/안개, 3: 약간의 비/눈, 4: 폭우/폭설)
temp	float	실측 온도 (섭씨)
atemp	float	체감 온도 (섭씨)
humidity	int	습도 (%)
windspeed	float	풍속 (m/s)
casual	int	등록되지 않은 사용자의 대여 수
registered	int	등록된 사용자의 대여 수
count	int	총 대여 수 (종속 변수)

> **train.csv** 파일에는 `count` 컬럼이 포함되어 있으며, 예측 대상인 종속 변수입니다.

> **test.csv** 파일에는 `casual`, `registered`, `count` 컬럼이 포함되어 있지 않습니다.

> **casual**과 **registered**는 자전거 대여 수요를 예측하는데 참고하실만한 자료이며, `count`는 두 컬럼간의 합입니다.

데이터 분석

데이터 속성 파악

데이터 정보

```
# 훈련 데이터
train_df = pd.read_csv("/content/drive/MyDrive/.../train.csv")

train_df.info()
```

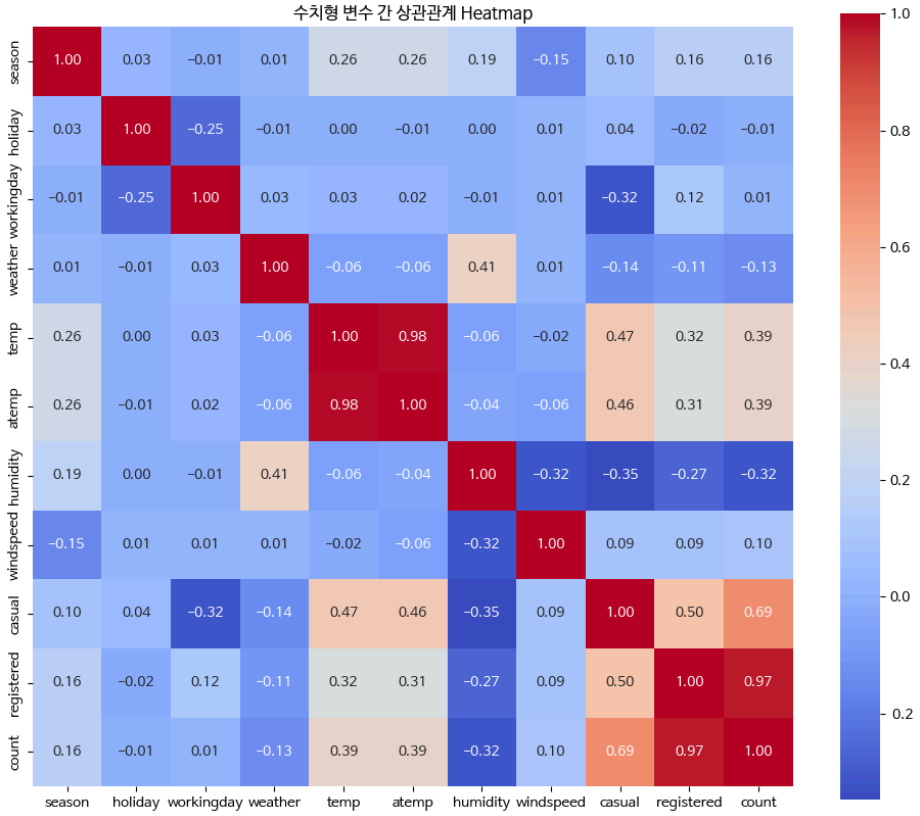
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10886 entries, 0 to 10885
Data columns (total 12 columns):
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   datetime    10886 non-null  object
 1   season      10886 non-null  int64
 2   holiday     10886 non-null  int64
 3   workingday  10886 non-null  int64
 4   weather     10886 non-null  int64
 5   temp        10886 non-null  float64
 6   atemp       10886 non-null  float64
 7   humidity    10886 non-null  int64
 8   windspeed   10886 non-null  float64
 9   casual      10886 non-null  int64
 10  registered  10886 non-null  int64
 11  count       10886 non-null  int64
dtypes: float64(3), int64(8), object(1)
memory usage: 1020.7+ KB
```

데이터 요약 통계량

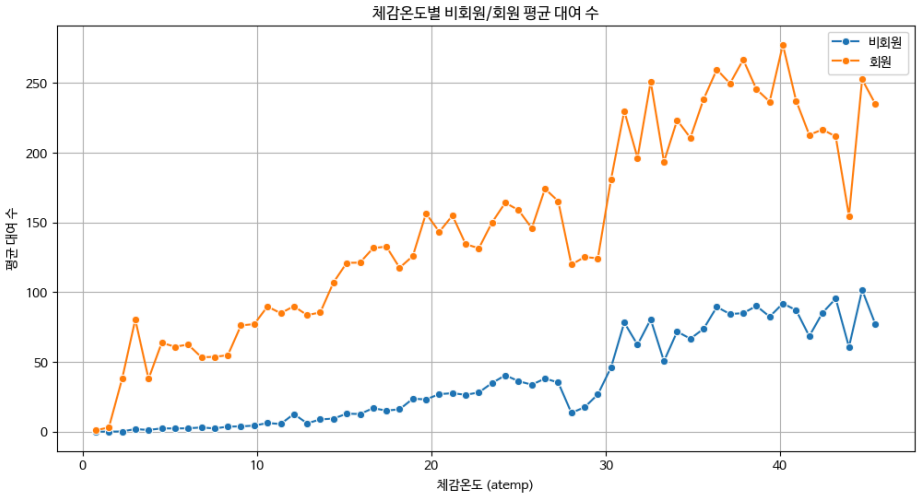
```
train_df.describe()
```

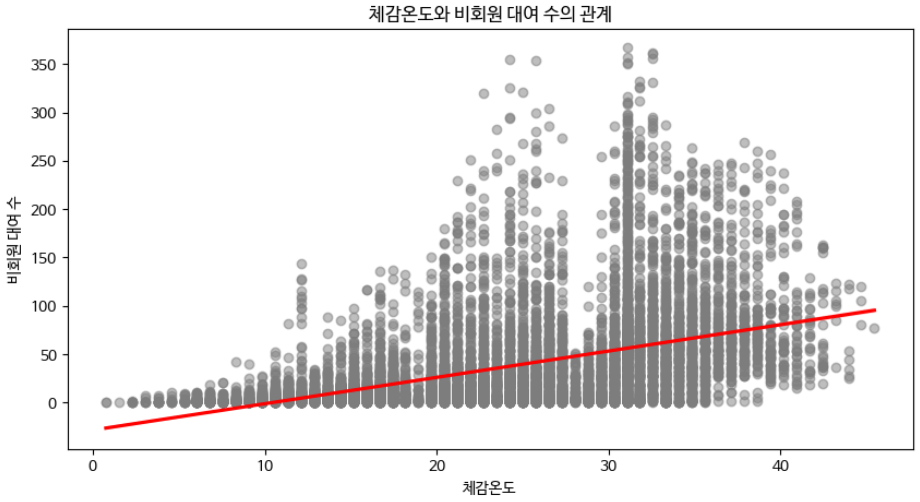
	season	holiday	workingday	weather	temp	atemp	humidity	windspeed	casual	registered	count
count	2665.0	2665.000000	2665.000000	2665.000000	2665.000000	2665.000000	2665.000000	2665.000000	2665.000000	2665.000000	2665.000000
mean	1.0	0.026642	0.678049	1.411632	12.506462	15.200004	56.709193	14.618354	15.594371	101.424765	117.019137
std	0.0	0.161064	0.467312	0.621028	5.196104	6.102418	19.628299	9.165139	31.322157	108.271187	125.494362
min	1.0	0.000000	0.000000	1.000000	0.820000	0.760000	0.000000	0.000000	0.000000	0.000000	1.000000
25%	1.0	0.000000	0.000000	1.000000	9.020000	10.605000	41.000000	7.001500	1.000000	22.000000	25.000000
50%	1.0	0.000000	1.000000	1.000000	12.300000	14.395000	54.000000	12.998000	5.000000	70.000000	79.000000
75%	1.0	0.000000	1.000000	2.000000	15.580000	19.695000	70.000000	19.999500	15.000000	142.000000	165.000000
max	1.0	1.000000	1.000000	3.000000	29.520000	32.575000	100.000000	51.998700	367.000000	681.000000	801.000000

수치형 데이터 연관성 파악 - heatmap()



체감온도(atemp)와 회원(registered) / 비회원(casual) 데이터 연관 파악

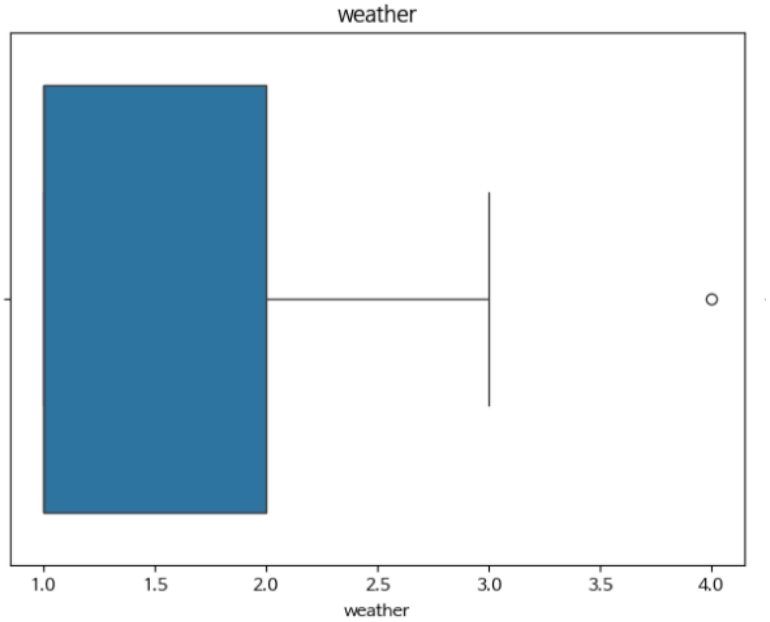




💡 **체감 온도가 높아질수록 회원 및 비회원들의 자전거 대여율이 높아지는 경향이 있음**

데이터 이상치 분석

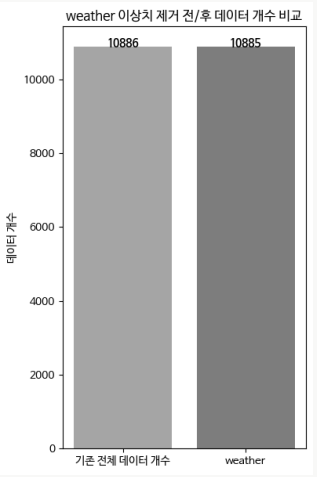
1. weather




```
train_df[train_df['weather'] == 4]
```

	datetime	season	holiday	workingday	weather	temp	atemp	humidity	windspeed	casual	registered	count
5631	2012-01-09 18:00:00	1	0	1	4	8.2	11.365	86	6.0032	6	158	164

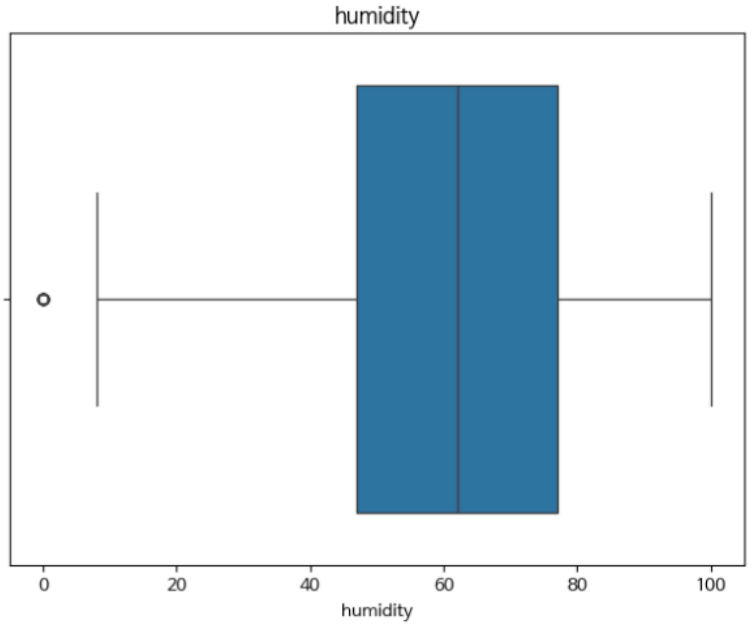
- 계절: 봄
- 날씨: 폭우/폭설
- 기온: 8 ~ 11도
- 습도: 86% (엄청 습함)
- 전체 자전거 대여 대수: 164



→ 봄에 폭우: 이상기후 현상으로 가정

→ 폭우가 내렸는데 164대의 자전거 대여 발생?: 전체 데이터 대비 비중이 적으니 삭제

2. humidity

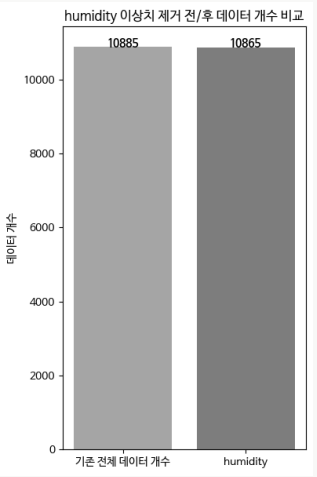



```
train_df[train_df['humidity'] == 0]
```

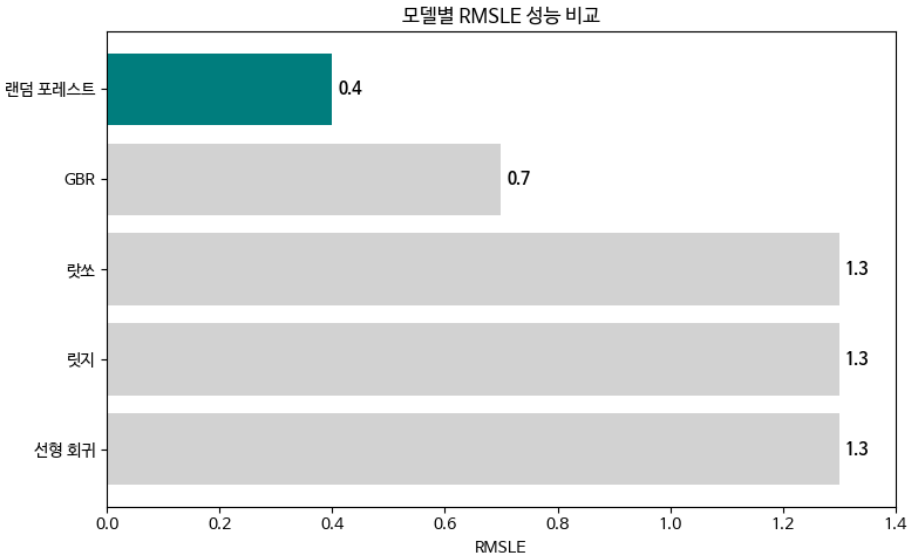
	datetime	season	holiday	workingday	weather	temp	atemp	humidity	windspeed	casual	registered	count
1091	2011-03-10 00:00:00	1	0	1	3	13.94	15.910	0	16.9979	3	0	3
1092	2011-03-10 01:00:00	1	0	1	3	13.94	15.910	0	16.9979	0	2	2
1093	2011-03-10 02:00:00	1	0	1	3	13.94	15.910	0	16.9979	0	1	1
1094	2011-03-10 05:00:00	1	0	1	3	14.76	17.425	0	12.9980	1	2	3
1095	2011-03-10 06:00:00	1	0	1	3	14.76	16.665	0	22.0028	0	12	12
1096	2011-03-10 07:00:00	1	0	1	3	15.58	19.695	0	15.0013	1	36	37
1097	2011-03-10 08:00:00	1	0	1	3	15.58	19.695	0	19.0012	1	43	44
1098	2011-03-10 09:00:00	1	0	1	3	16.40	20.455	0	15.0013	1	23	24
1099	2011-03-10 10:00:00	1	0	1	3	16.40	20.455	0	11.0014	0	17	17
1100	2011-03-10 11:00:00	1	0	1	3	16.40	20.455	0	16.9979	6	5	11
1101	2011-03-10 12:00:00	1	0	1	3	17.22	21.210	0	15.0013	4	30	34
1102	2011-03-10 13:00:00	1	0	1	3	17.22	21.210	0	15.0013	1	11	12
1103	2011-03-10 14:00:00	1	0	1	3	18.04	21.970	0	19.9995	0	12	12
1104	2011-03-10 15:00:00	1	0	1	3	18.04	21.970	0	15.0013	3	11	14
1105	2011-03-10 16:00:00	1	0	1	3	17.22	21.210	0	16.9979	1	20	21
1106	2011-03-10 17:00:00	1	0	1	2	18.04	21.970	0	26.0027	2	109	111
1107	2011-03-10 18:00:00	1	0	1	3	18.04	21.970	0	23.9994	2	80	82
1108	2011-03-10 19:00:00	1	0	1	3	18.04	21.970	0	39.0007	5	51	56
1109	2011-03-10 20:00:00	1	0	1	3	14.76	16.665	0	22.0028	9	29	38
1110	2011-03-10 21:00:00	1	0	1	3	14.76	17.425	0	15.0013	1	27	28
1111	2011-03-10 22:00:00	1	0	1	2	13.94	16.665	0	8.9981	4	30	34
1112	2011-03-10 23:00:00	1	0	1	3	13.94	17.425	0	6.0032	1	26	27

습도가 0 이면서, 날씨가 3 or 4인 경우

→ 말이 안되는 데이터라 판단하여 삭제



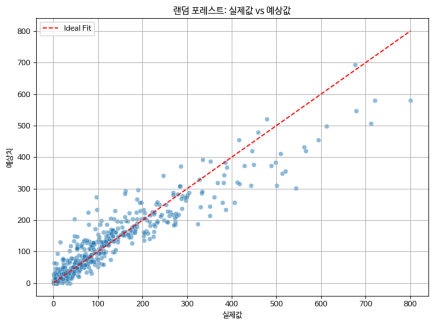
모델별 성능 비교



모델링 결과

긍정적 모델링

- **산점도의 점들이 대체로 대각선 주변에 밀집**되어 오차가 적음.
- 저수요 구간(0~200)에 특히 예측이 잘 되어 있음:



아쉬운점

- 수요가 높은 일부 구간 (400 이상)에서 과소 예측 경향 존재
- 분산이 큼

제언

1. 출퇴근 시간대의 러시 아워 (Rush Hour) 공략

- 출퇴근 패스권
  - 출퇴근 시간(06:00 ~ 12:00 / 17:00 ~ 20:00)에 운행 거리 및 시간에 구애받지 않는 월 단위 구독 시스템 도입
- 1+n 이벤트
  - 계정에 등록된 친구, 지인과 같은 시간에 함께 대여 시 할인권 제공

2. 날씨 및 체감 온도에 따른 공략

- 자전거를 타기 가장 쾌적한 기온을 설정하여 해당 기온에 도달하면 자전거 대여 권유 알림 발송
  - 해당 알림을 통해 대여시 대여료 할인 및 위치기반 자전거 드라이브 코스 추천 시스템 개발