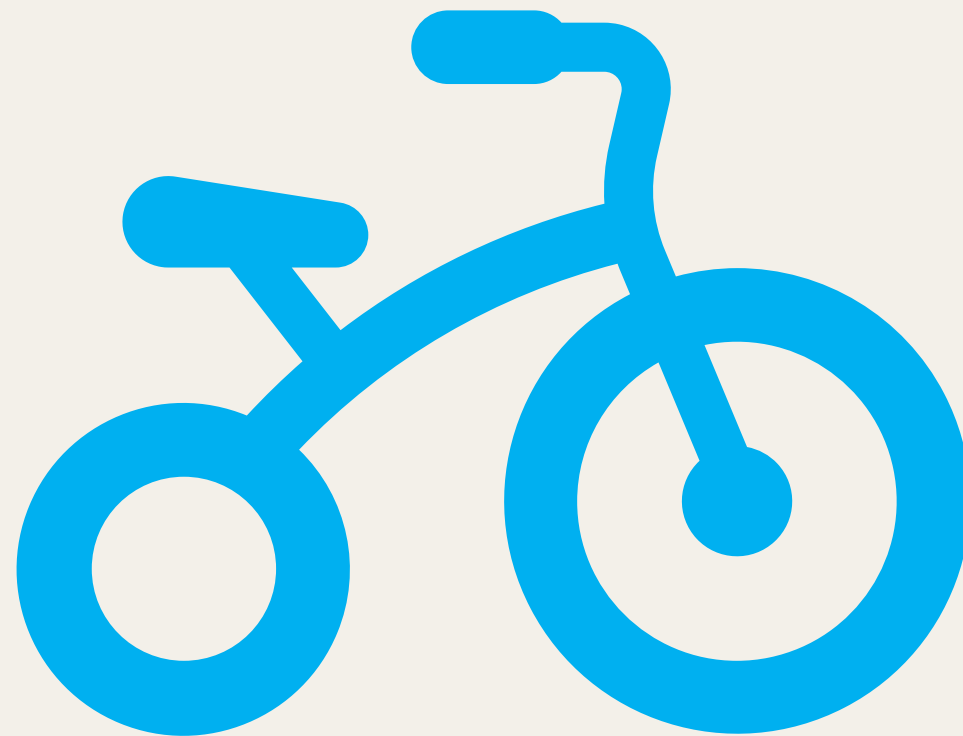


자전거 대여 수요 예측
Kaggle 4조



목차

1. 데이터 설명
2. EDA 및 전처리
3. 모델
4. Kaggle 제출

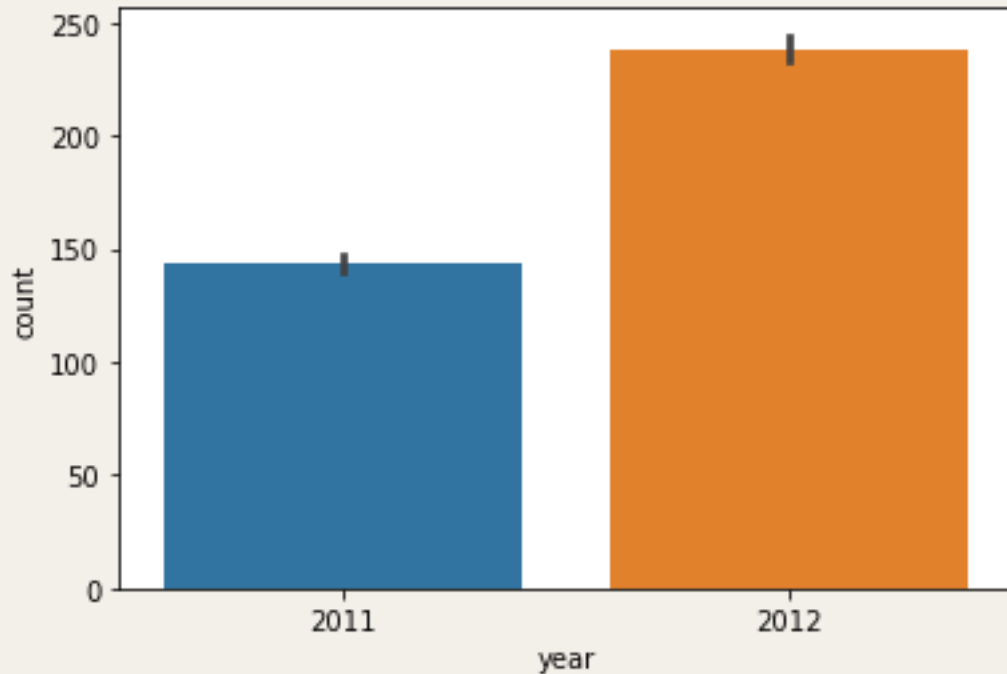
1. 데이터 설명

자전거 수요 예측 대어 <https://www.kaggle.com/c/bike-sharing-demand/>

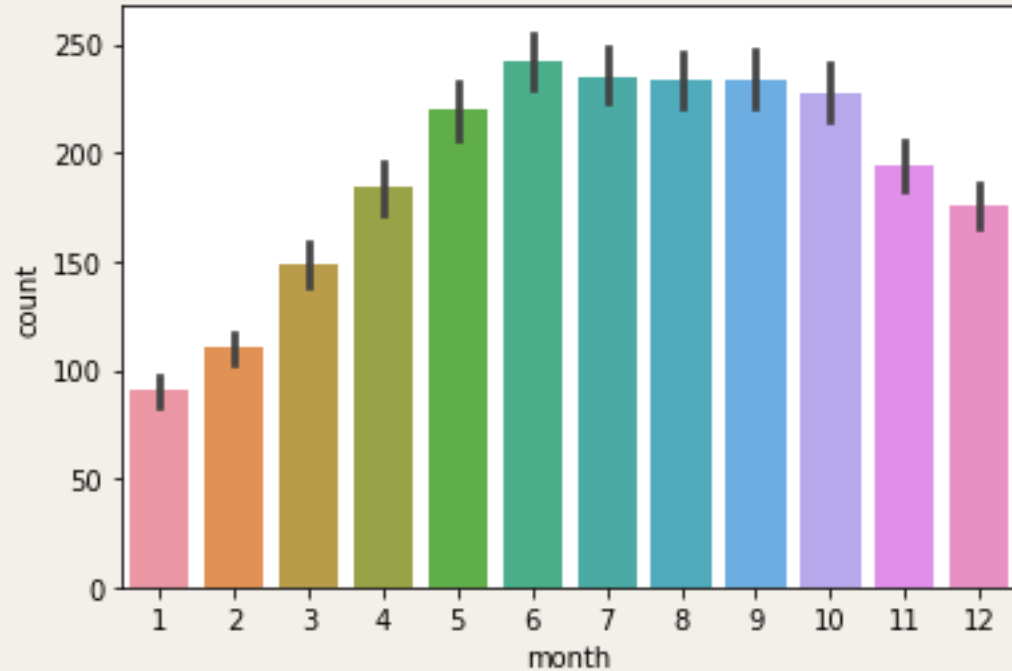
Column 명	데이터 내용
Datetime	시간(YYYY-MM-DD 00:00:00)
Season	봄(1) 여름(2) 가을(3) 겨울(4)
Holiday	공휴일(1) 공휴일 아님(0)
Workingday	근무일(1) 근무일 아님(0)
Weather	맑음(1) 약간 흐림(2) 약간의 눈, 비(3) 비, 우박(4)
Temp	온도(섭씨)
Atemp	체감온도(섭씨)
Humidity	습도
Windspeed	풍속
Casual	비회원의 자전거 대여량
Registered	회원의 자전거 대여량
Count	총 자전거 대여량(회원 + 비회원)

2. EDA 및 전처리

연도별 자전거 대여 수요량



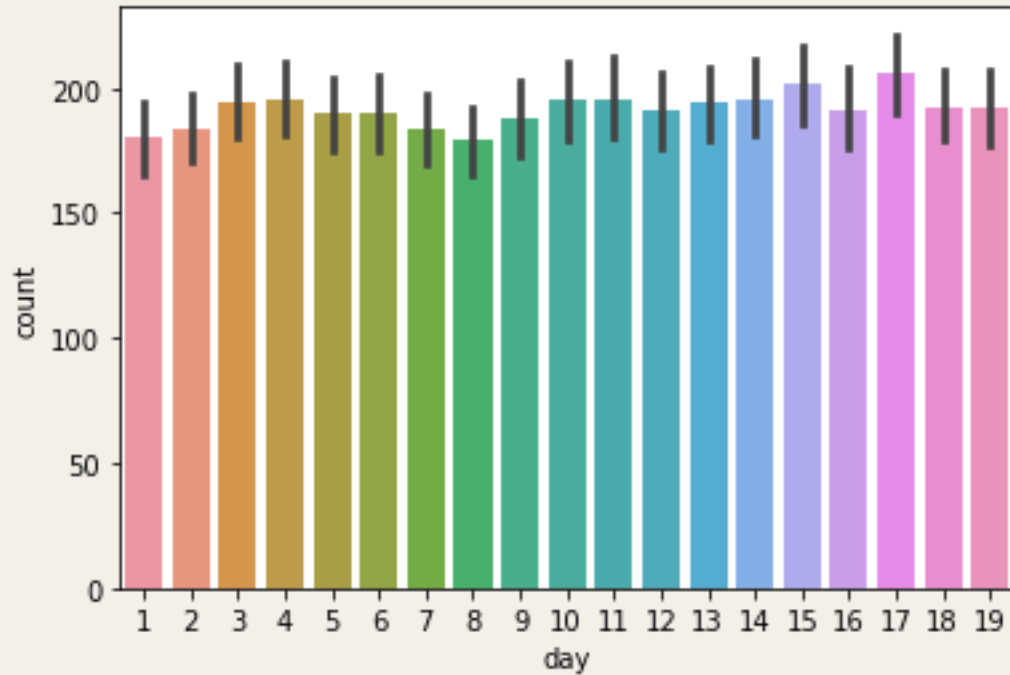
월별 자전거 대여 수요량



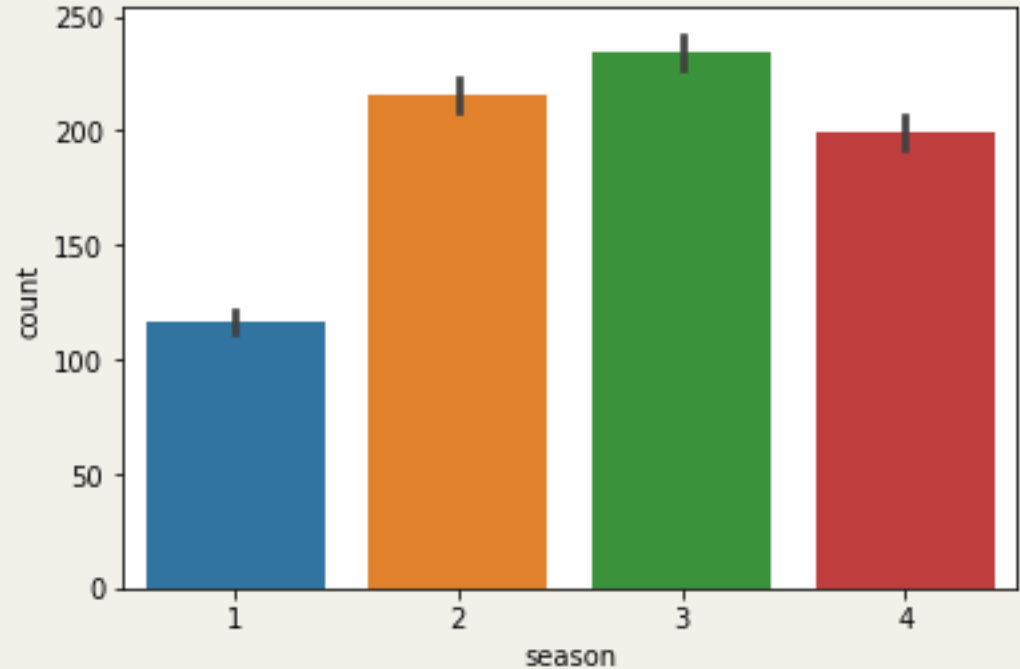
- 2012년의 대여량이 더 많다
- 6~8월에 수요가 많고 12~1월에는 비교적 적은 편이다

2. EDA 및 전처리

일별 자전거 대여 수요량



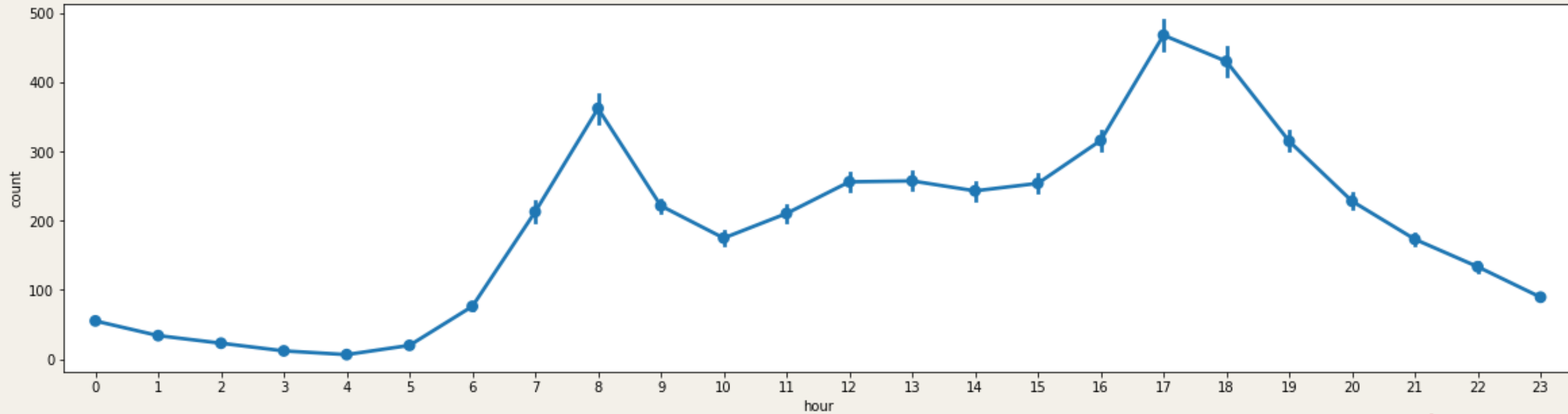
계절별 자전거 대여 수요량



- 일별 수요량에는 경향성이 나타나지 않는 것으로 보인다
- 여름(2), 가을(3)에 비교적 수요가 많은 것으로 보인다

2. EDA 및 전처리

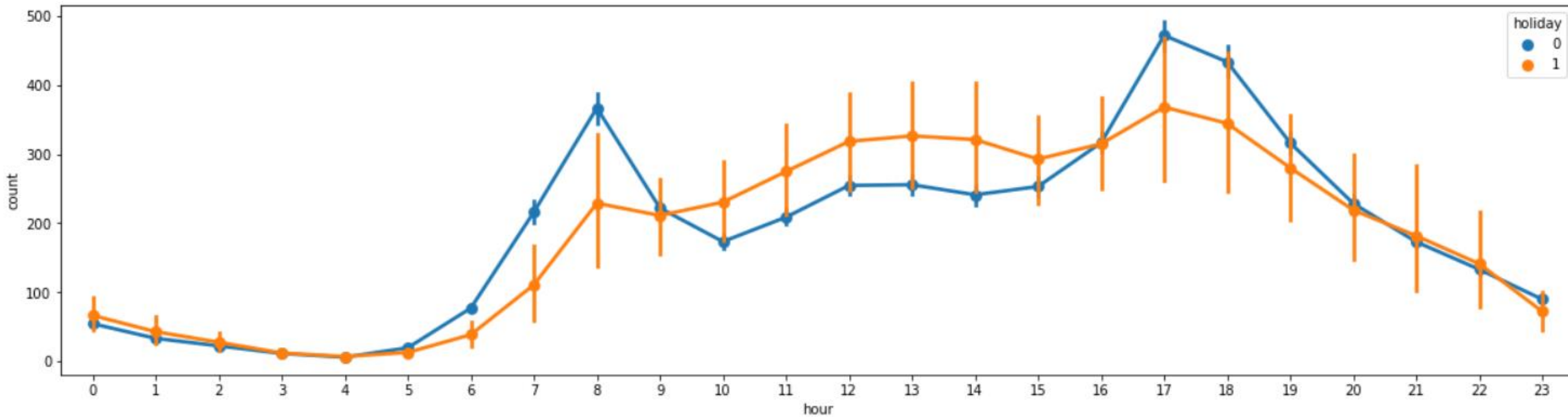
시간대별 자전거 대여 수요량



- 출퇴근 시간대에 수요가 증가함을 알 수 있다

2. EDA 및 전처리

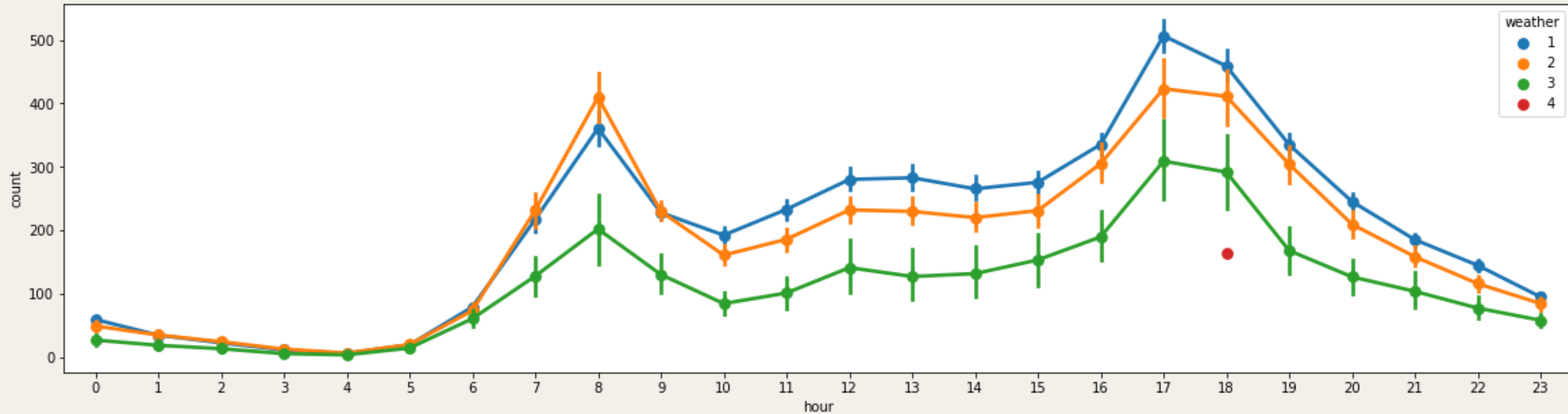
휴일 여부에 따른 시간대별 자전거 대여 수요량



- 휴일일 때(1) 오후 시간에 수요가 증가하는 것으로 보인다
- 휴일이 아닐 때(0): 출퇴근 시간대에 수요가 증가하는 것으로 보인다

2. EDA 및 전처리

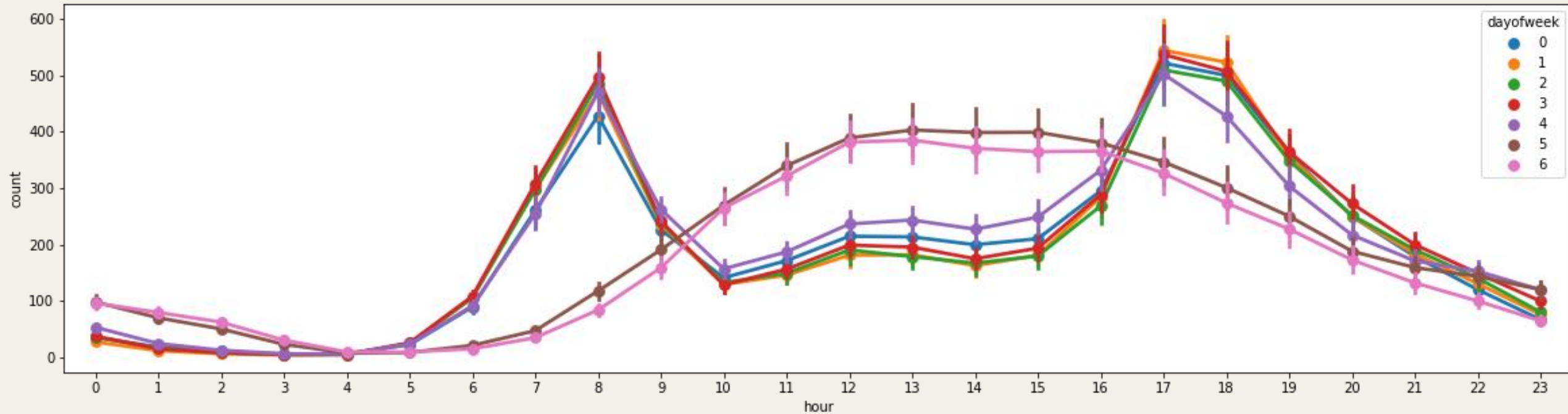
날씨에 따른 시간대별 자전거 대여 수요량



- 비와 우박(4)이 오는 날의 데이터가 거의 없다

2. EDA 및 전처리

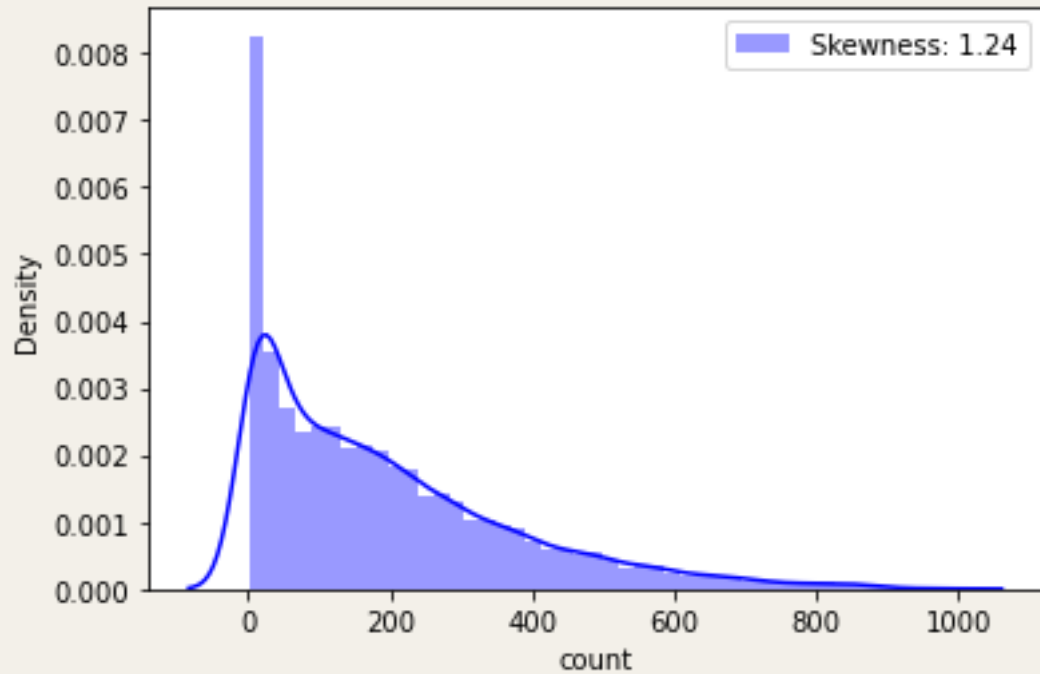
요일에 따른 시간대별 자전거 대여 수요량



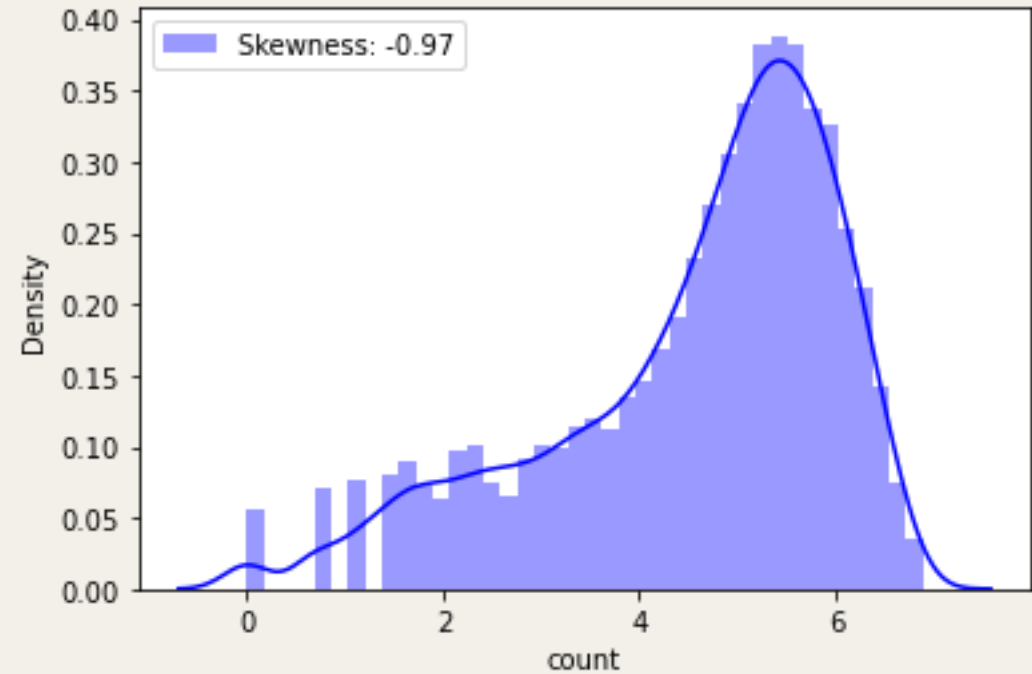
- 요일별 수요에는 평일(0~4)/주말(5, 6)의 차이가 존재하는 것으로 보인다

2. EDA 및 전처리

Count의 왜도



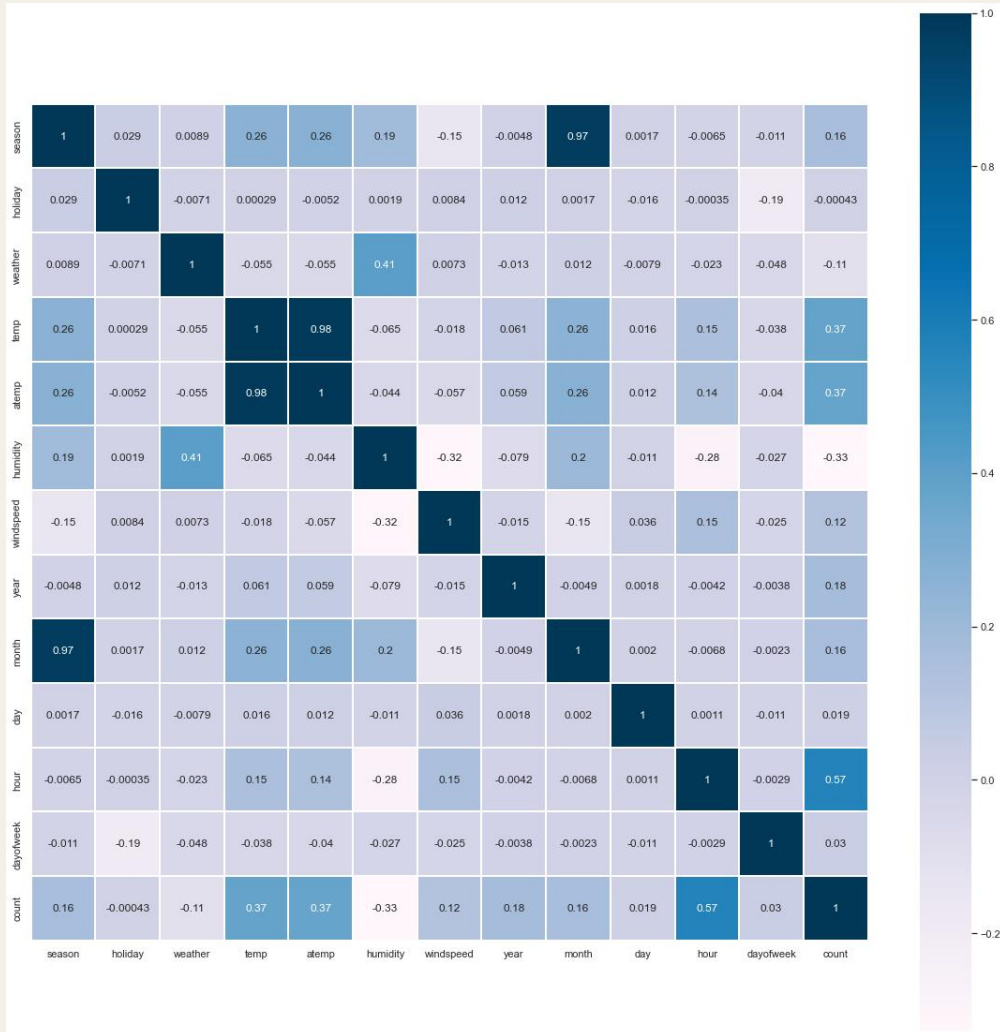
Log count의 왜도



- Count에 0값이 너무 많이 존재하므로 로그를 씌워 스케일링을 하였다
- 로그를 씌운 결과 비교적 정규분포 형태를 갖추는 것을 볼 수 있다

2. EDA 및 전처리

변수별 상관관계 히트맵



- temp와 atemp 변수의 상관관계가 매우 높다
- month와 season의 상관관계가 매우 높다



상관관계가 매우 높은 변수로 인해 다중공산성이 의심되어 atemp, season 변수 제거

2. EDA 및 전처리

그 밖의 전처리 과정

- Datefield, Casual, Registered 열 제거
- MinMaxScaler를 이용하여 모든 열의 값을 0과 1 사이의 값으로 스케일링
- Train 데이터를 8 : 2의 비율로 나눠서 훈련 세트와 검증 세트를 만듦

3. 모델

- Scikit-Learn의 LinearRegression 함수를 사용하여 선형 회귀 모델 사용
- Pytorch로 인공신경망 모델 사용

평가지표	선형 회귀 모델	인공 신경망 모델
RMSE	Train: 1.074, Valid: 1.068	Train: 0.423, Valid: 0.442
R2 Score	Train: 0.480, Valid: 0.483	Train: 0.919, Test: 0.912

- RMSE: 평균 제곱 오차의 제곱근을 씌운 값. 범위는 0 이상, 작을수록 좋음
- R2 Score: 실제 값의 분산 대비 예측 값의 분산 비율. 범위는 0~1, 클수록 좋음



인공 신경망 모델의 성능 지표가 훨씬 좋게 나타난 것을 확인할 수 있다

3. 모델

모델 성능을 더 높이기 위해 개선할 점

- 연속형 변수(습도, 온도 등)에 대한 EDA
- 이상치 제거
- Windspeed(풍속) = 0 결측치 처리

4. Kaggle 제출

제출 파일 만들기

```
submission = pd.DataFrame()  
submission['datetime'] = sample['datetime']  
submission['count'] = pred  
submission.to_csv("Bike.csv", index=False)
```

Test 데이터 예측 값

Bike Sharing Demand

Forecast use of a city bikeshare system



Kaggle · 3,242 teams · 8 years ago

[Overview](#) [Data](#) [Code](#) [Discussion](#) [Leaderboard](#) [Rules](#) [Team](#)

[Submissions](#)


[Late Submission](#)

...

Late Submission 클릭

4. Kaggle 제출

YOUR RECENT SUBMISSION

 **Bike.csv** Score: 0.52734

Submitted by Hyeokju Park · Submitted a minute ago

↓ [Jump to your leaderboard position](#)

만들어진 csv파일을 올리면 채점이 되고 Leaderboard에서 순위를 확인할 수 있다.