선형회귀 모델 설명, 기온 데이터 EDA

김은교

선형회귀 모델

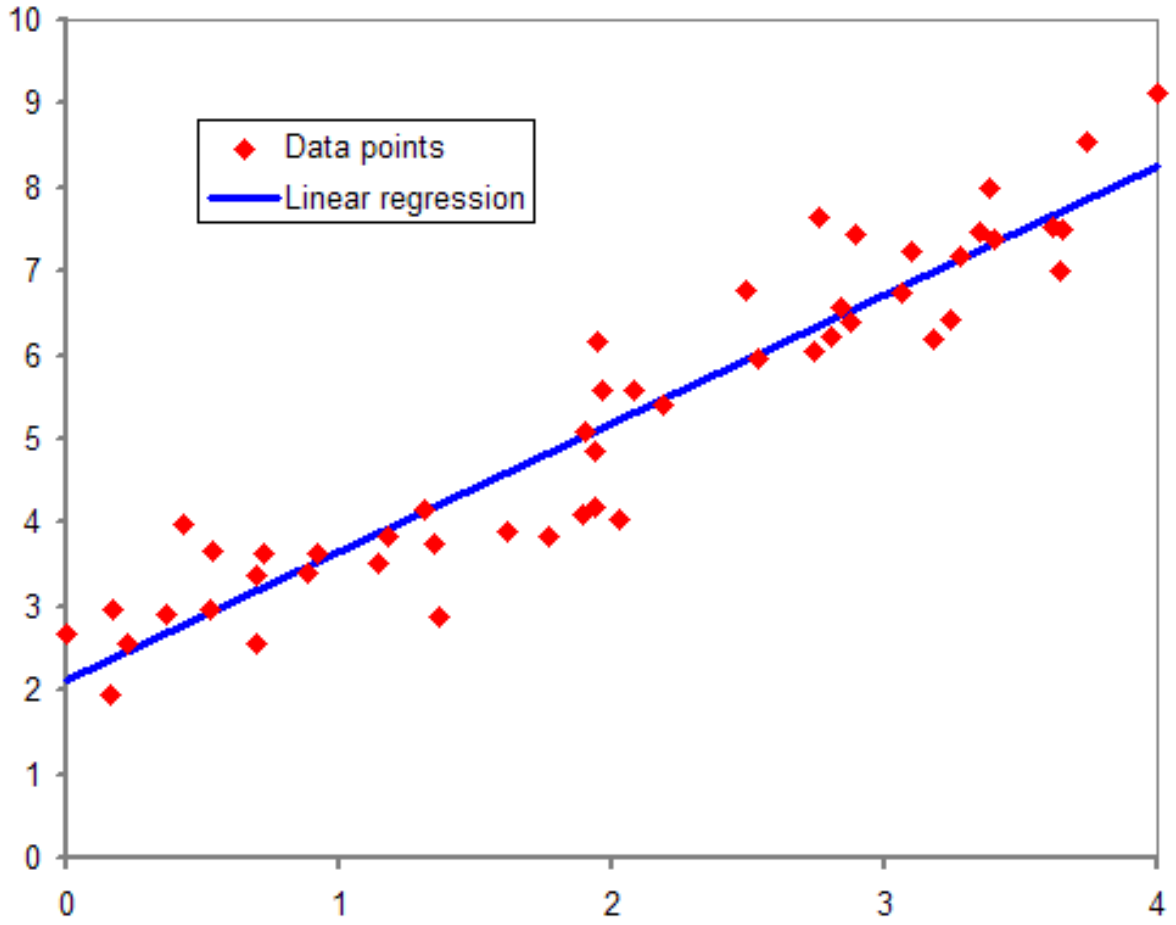


Image 출처: Wikipedia '독립변수 1개와 종속변수 1개를 가진 선형 회귀의 예', https://ko.wikipedia.org/wiki/%EC%84%A0%ED%98%95_%ED%9A%8C%EA%B7%80

우리가 예측하려는 값은 '따릉이 대여 수' 이므로,

수치형 데이터에 해당함.

→ baseline 코드에서는 주어진 feature(x)과 target(y)인 대여 수 간의 관계를 학습하여 예측값과 실제 값의 차이가 최소가 되는 line을 학습하는 선형 회귀 모델을 선택하였음.

Model 형태:

$$y = w_1 x_1 + w_2 x_2 + \cdots + w_n x_n + b$$

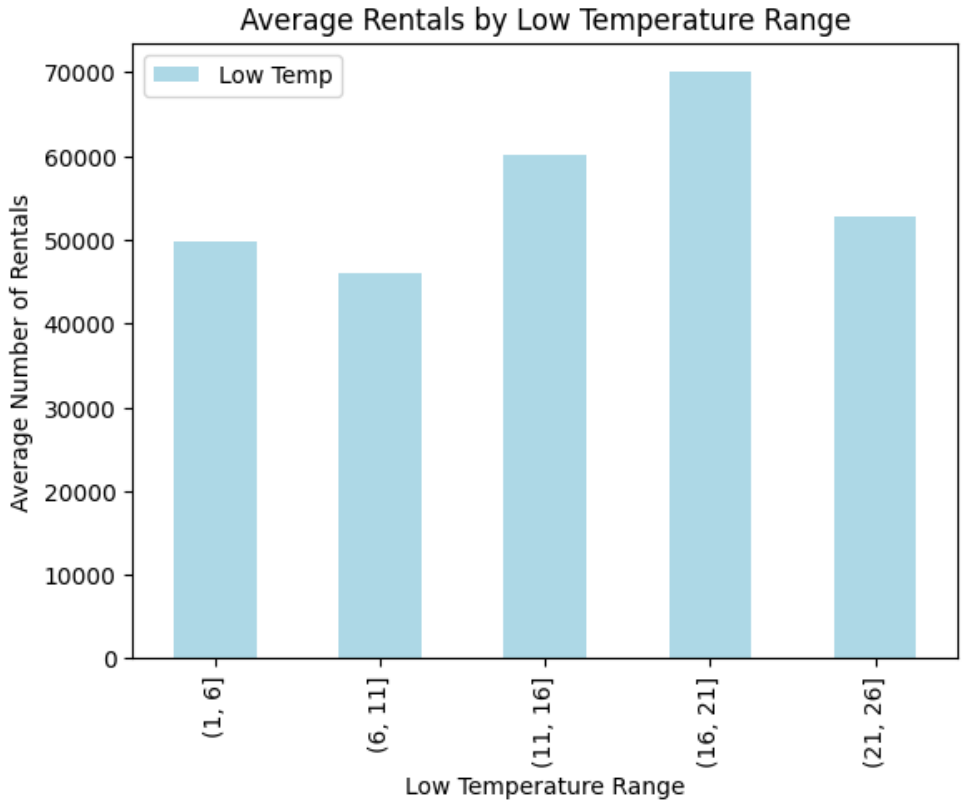
특성 x에 대한 각각의 **가중치 w**와 **절편 b** 학습

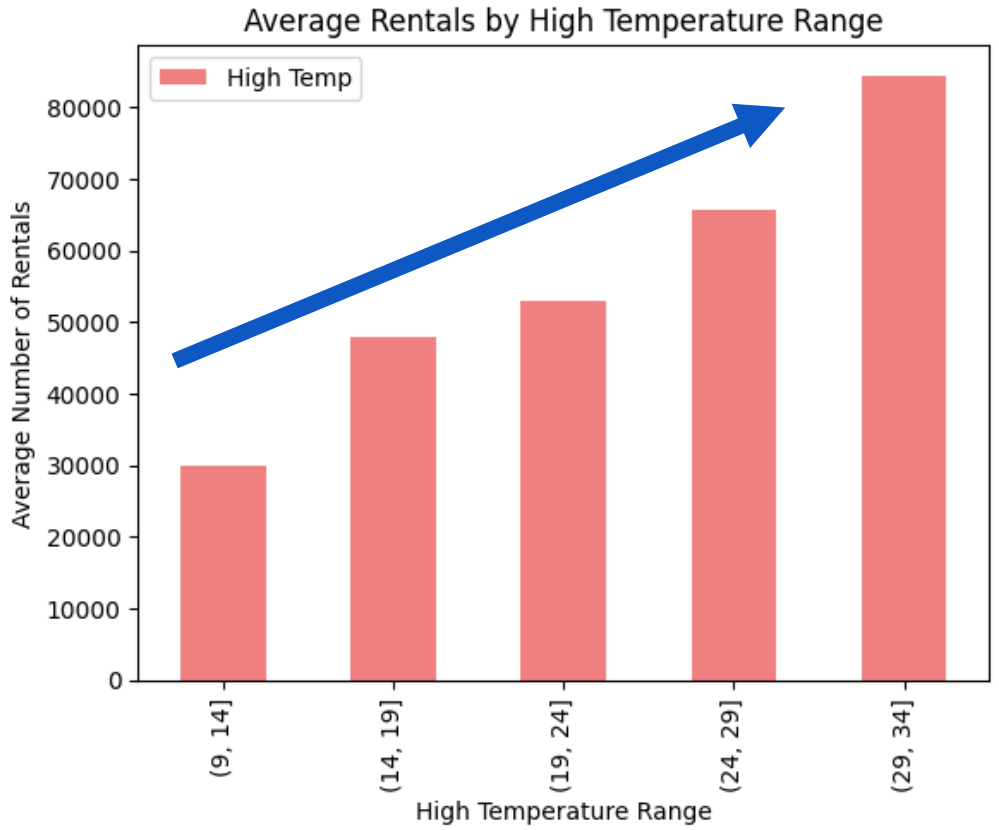
선형회귀 모델 주요 장단점

- 장점
  - 데이터가 선형성을 가질 때 좋은 성능을 가짐
  - 연속적인 수치를 예측할 때 주로 사용
  - 직관적으로 해석 가능
  - 빠르고 효율적
  - 가중치 값을 통해 특성의 중요도 확인 가능
- 단점
  - 비선형적 관계에서 성능이 낮음
  - 데이터의 스케일, 이상치에 민감
  - 차원이 너무 많을 때 overfitting(과적합) 문제 발생할 수 있음
  - 다중공선성 문제 (변수들 간 높은 상관관계를 가질 때 모델의 안정성과 성능 저하)

Low temp & High temp 조사

방법: low temp, high temp 각각 최솟값부터 최댓값까지 구간을 5도씩 나눠서 시각화
목표: 온도가 가장 높은 구간에서 대여 수가 가장 많은지, 증가하는 추세를 띄는지 확인하기 위함



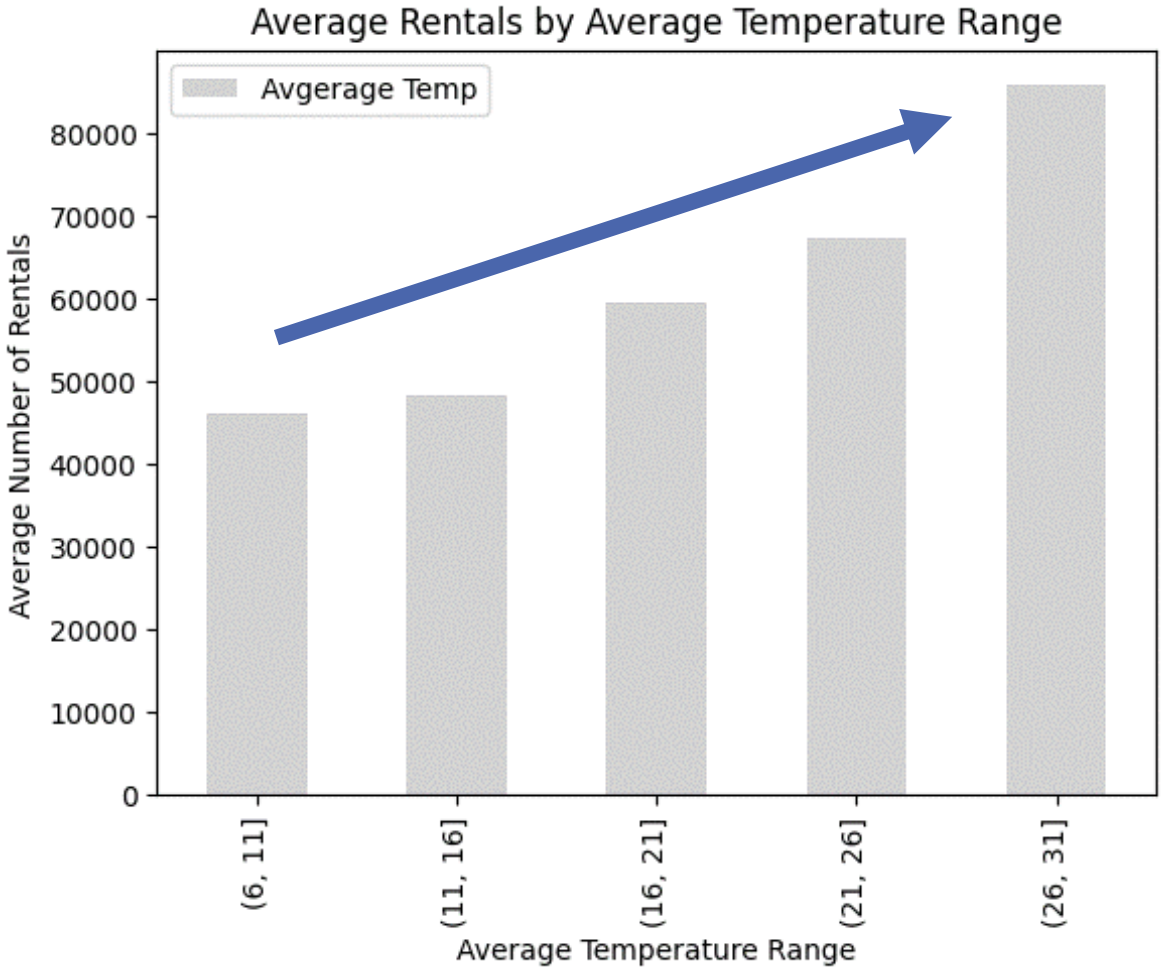


결과: low temp에서는 유의미한 상관관계를 찾지 못하였지만,
High temp에서는 온도가 가장 높은 구간에서 대여 수가 가장 많음을 확인할 수 있었음.
또한 최고기온이 증가할수록 대여 수도 많은 경향이 있는 것을 확인함.

평균 기온 (avg temp) 조사

방법: (low temp + high temp)/2로 새로운 칼럼 avg temp를 생성 후 동일한 방식으로 시각화

목표: 평균 기온이 상승할수록 대여 수가 증가하는 추세가 있는지 확인하기 위함



✓결론

시각화 그래프에서 증가하는 추세를 확인 할 수 있으므로,

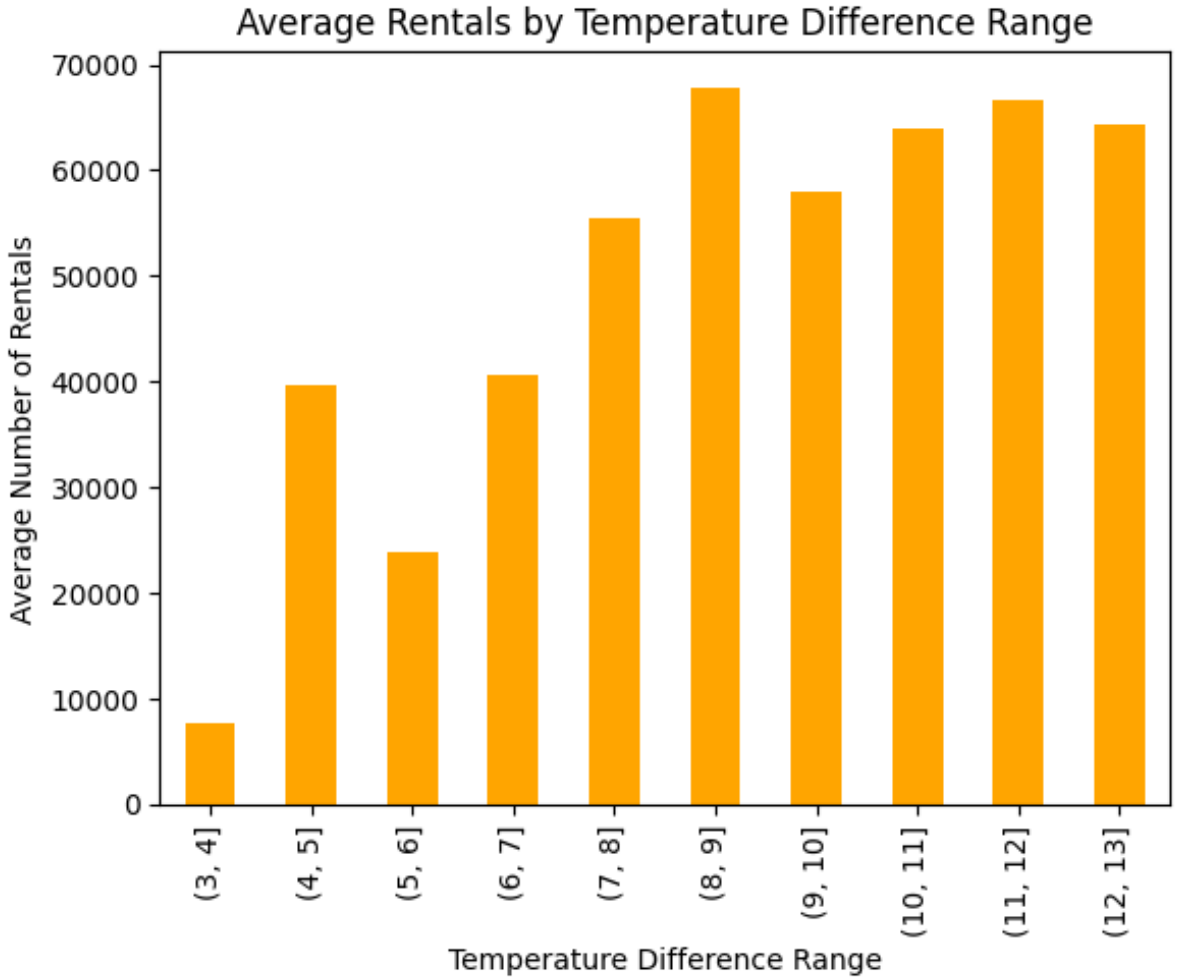
주어진 데이터 내에서는 기온이 상승함에 따라 대여 수도 증가함을 확인

일교차 조사

방법: high temp - low temp로 새로운 칼럼을 생성하고,

해당 칼럼의 최솟값부터 최댓값까지 1도 간격으로 구간을 나눠 시각화

목표: 일교차와 대여 수의 관계를 알아보기 위함



✓결론

특별히 유의미한 상관관계는 찾지 못하였으나, 일교차가 적을 때 대여 수가 확연히 적은 것을 확인

→ 일교차와 다른 변수들 간의 추가 조사 진행

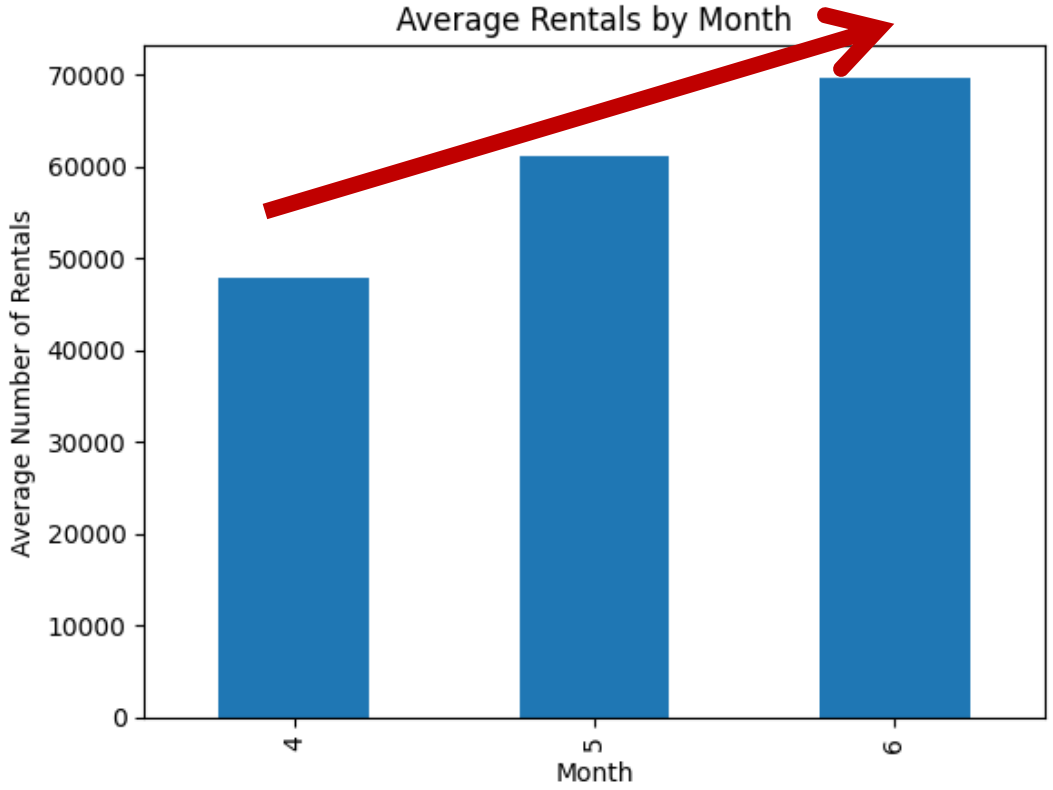
: 일교차와 강수 확률이 뚜렷한 음의 상관관계를 나타냄 (-0.59)

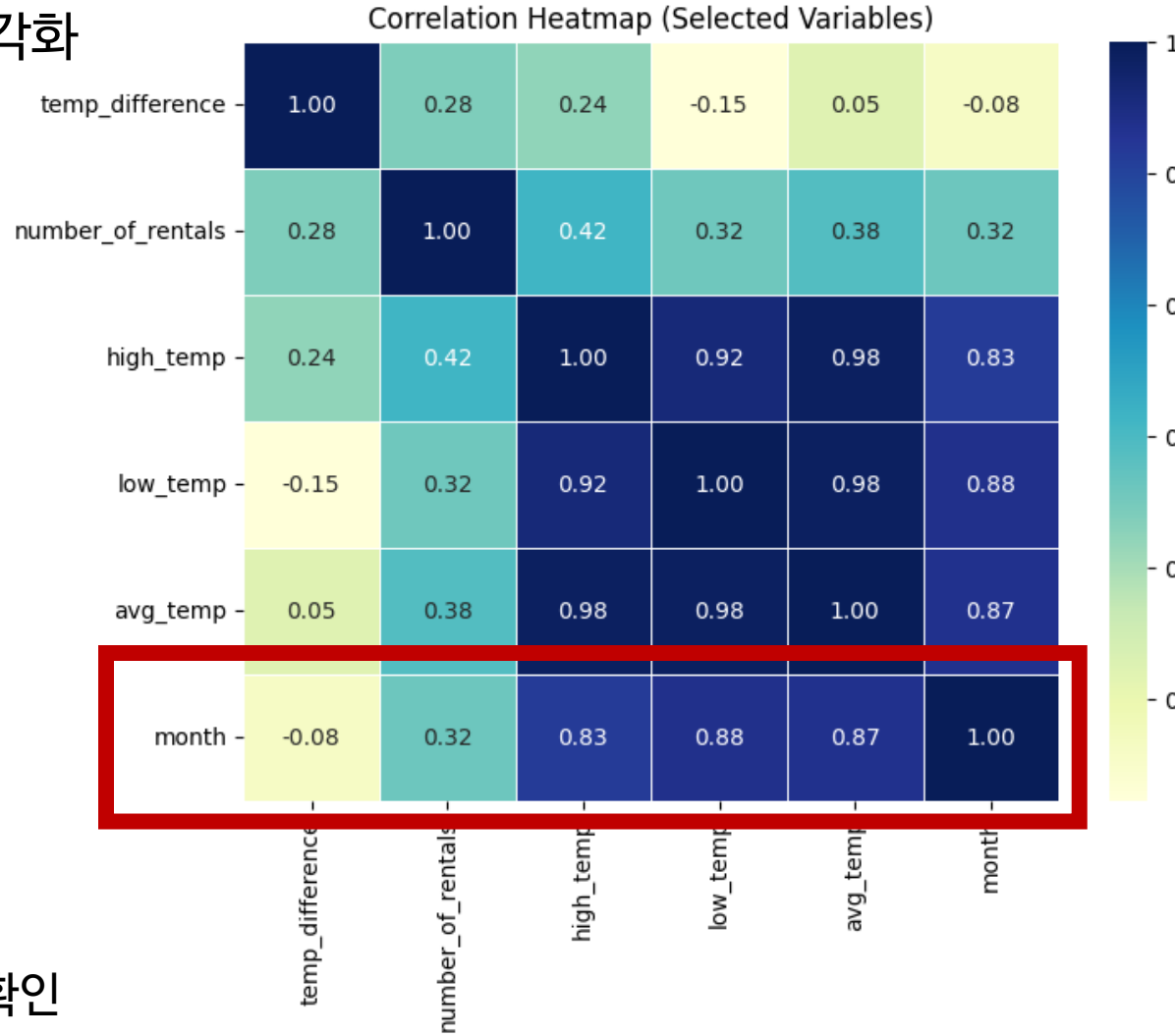
∴ 일교차가 작을 때는 강수 확률이 높은 날이었을 것으로 예상

→ 강수확률과 대여 수에 유의미한 상관관계가 있음을 조사할 필요성이 있음.

월별 평균 대여량 조사

방법: 주어진 데이터를 월별로 그룹화하여 평균 대여 수를 시각화





Heatmap에서 기온 관련 데이터와 강한 양적 상관관계를 나타냄을 확인

→ 6월로 갈수록 대여 수가 증가하는 이유는 기온이 상승하기 때문일 것으로 생각할 수 있음

메모

- 내용 요약용으로 만든 ppt이니 그대로 쓰셔도 되고 적당히 추려서 쓰셔도 돼요
- 일교차와 강수확률 관련 인사이트는 갑자기 생각나서 해 본거라 강수확률과 관련 지어서 하셔도 되고 그냥 빼셔도 될 것 같습니다
- 다음에 해야 할 것: 그래서 전처리를 어떻게 하겠다~ 어떤 모델을 선택해 보겠다~
- 베이스라인에서는 선형회귀를 썼는데 우리는 다른 모델을 써서 성능을 다르게 해 보겠다~로 진행해도 좋겠네요