서울시 공유자전거 '따릉이'의 수요 예측

김영우

대주제

어떠한 날씨요소가 따릉이 수요와 관련 있는지 분석하여 인사이트를 도출한다.

가설 수립

- 가설 1 : 달(month)이 자전거 대여량(count)와 연관이 있는가?
- 가설 2 : 요일(day)이 자전거 대여량(count)와 연관이 있는가?
- 가설 3 : 시간대(hour)가 자전거 대여량(count)와 연관이 있는가?
- 가설 4 : 미세먼지(PM10)이 자전거 대여량(count)와 연관이 있는가?
- 가설 4 : 온도(temperature)가 자전거 대여량(count)와 연관이 있는가?
- 가설 5 : 강우 여부(precipitation)가 자전거 대여량(count)와 연관이 있는가?
- 가설 6 : 풍속(windspeed)가 자전거 대여량(count)와 연관이 있는가?
- 가설 7 : 습도(humidity)가 자전거 대여량(count)와 연관이 있는가?

단변량 분석 – 달 (month)

- 4월, 5월, 6월 …. 11월 데이터 건수가 (29개 ~ 30개) 같기 때문에 단변량 분석은 하지 않았다.

단변량 분석 – 요일(day)

- 일요일, 월요일 … 토요일 데이터 분포가 똑같기 때문에 단변량 분석은 하지 않았다.

단변량 분석 – 시간대 (hour)

- 0 ~ 23의 데이터 분포가 똑같기 때문에 단변량 분석은 하지 않았다.

단변량 분석 – 미세먼지 (PM10)

- 결측치 NaN 처리

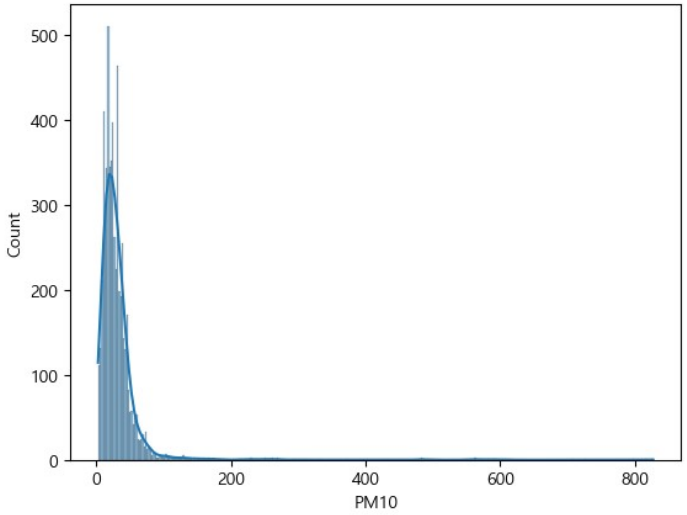
  결측치 NaN 102개는, 미세먼지(PM10)의 평균값으로 대치했다.

```
# 결측치는 미세먼지(PM10)의 평균으로 대치한다.
PM10_not_null_datas = df[df['PM10'].notna()]
PM10_average = PM10_not_null_datas["PM10"].sum() / PM10_not_null_datas.shape[0]
df["PM10"] = df["PM10"].fillna(PM10_average)
df["PM10"].isna().sum()
```

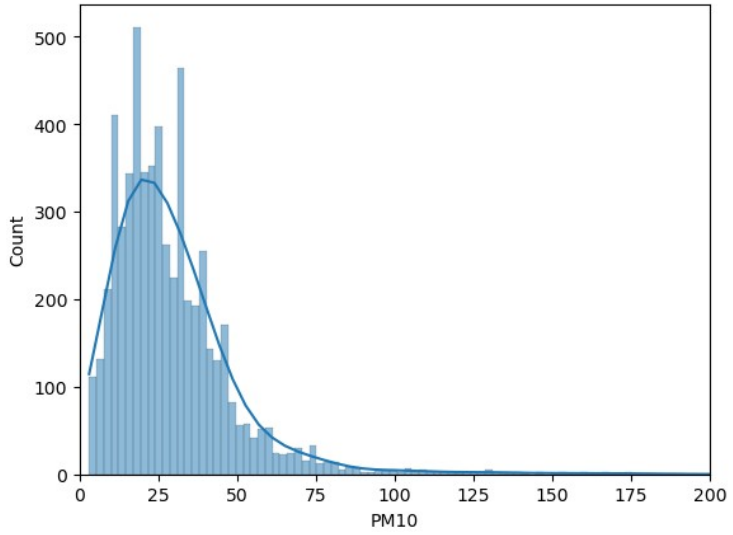
```
0
```

  따라서 PM10의 결측치 NaN은 0이 된 것을 알 수 있다.

단변량 분석 – 미세먼지 (PM10)



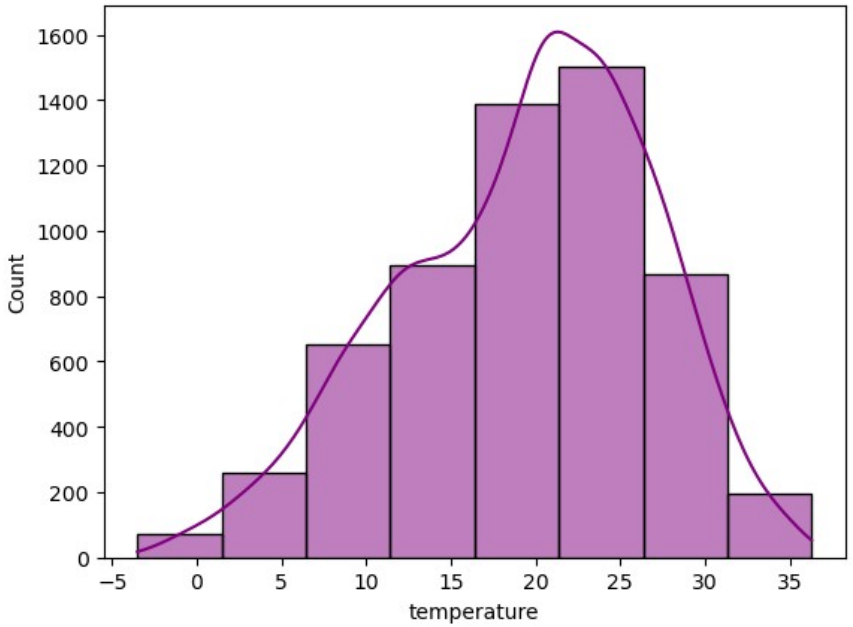
구간을 좁힐 필요성이 생겼다.



Xlim(0, 200)으로 설정
미세먼지 좋음 (0 <= < 31),
보통 (31 <= < 80)이 대부분 분포

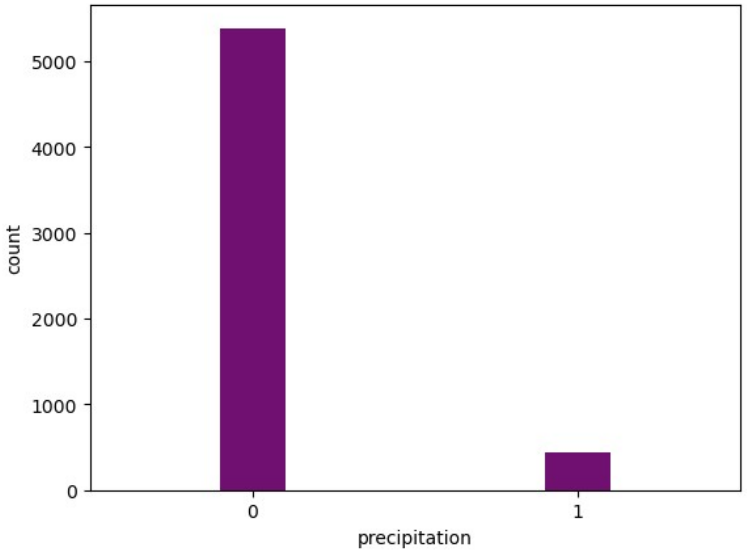
미세먼지 나쁜 데이터가 거의 없는 것을 알 수 있다.
-> 미세먼지가 한창 창궐하는 12월 ~ 3월달
데이터가 없기 때문

단변량 분석 – 온도 (temperature)



16도 ~ 26도 정도 데이터가 많다.
영하의 데이터는 거의 없다시피한데
-> 12월, 1월, 2월의 겨울 시간 데이터가
없어서 라고 판단한다.

단변량 분석 – 강우 여부 (preciptation)

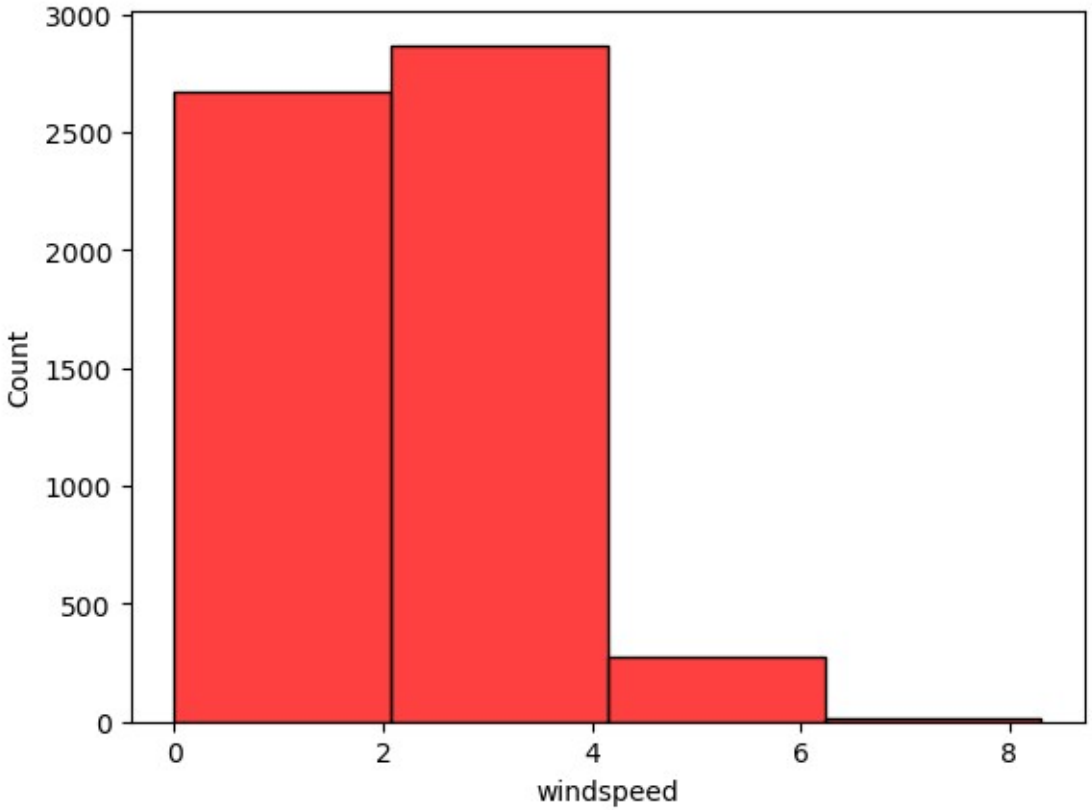



```python
# 전체 데이터
row = df.shape[0]

# 비가 오지 않는 데이터 건수(확률)
precipitaion_0 = df.loc[df["precipitation"] == 0].shape[0]
display(precipitaion_0 / row * 100)

# 비가 오는 데이터 건수(확률)
precipitaion_1 = df.loc[df["precipitation"] == 1].shape[0]
display(precipitaion_1 / row * 100)
```

```
92.44894456838854
7.551055431611465
```

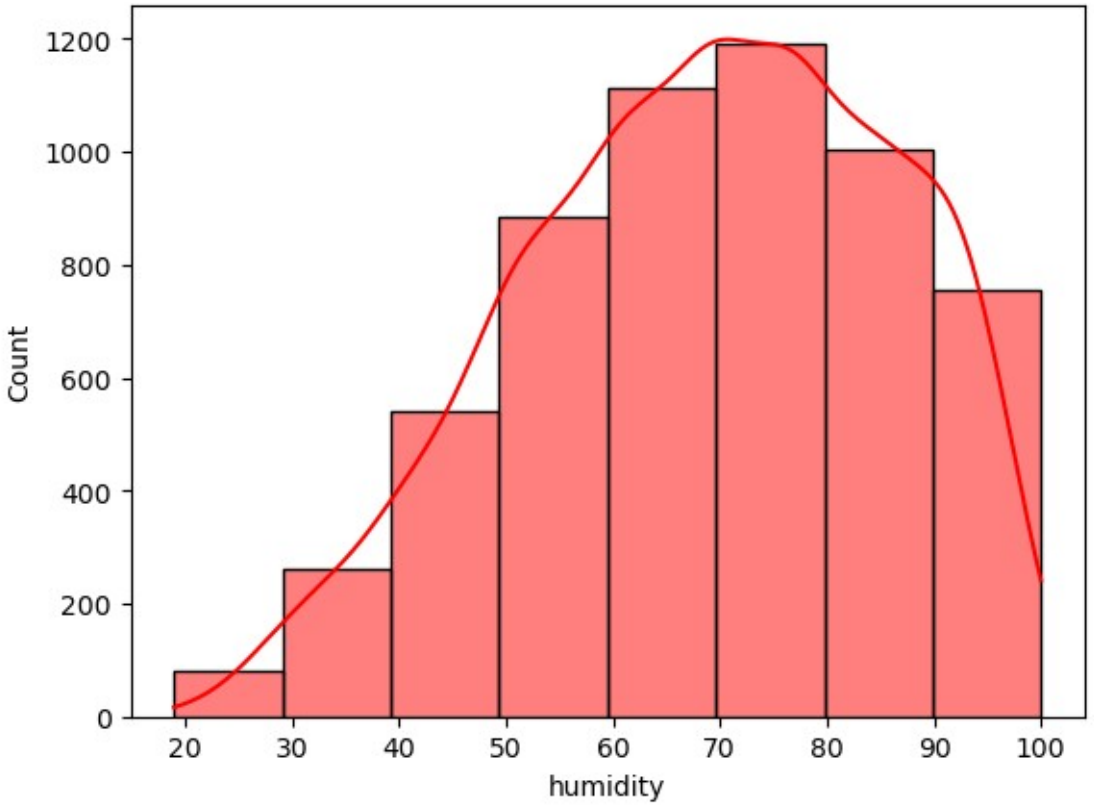
전체 데이터 중 약 92퍼는 비가 오지 않은 데이터, 약 8퍼는 비가 온 데이터로 구성되어 있다.

단변량 분석 – 풍속 (windspeed)



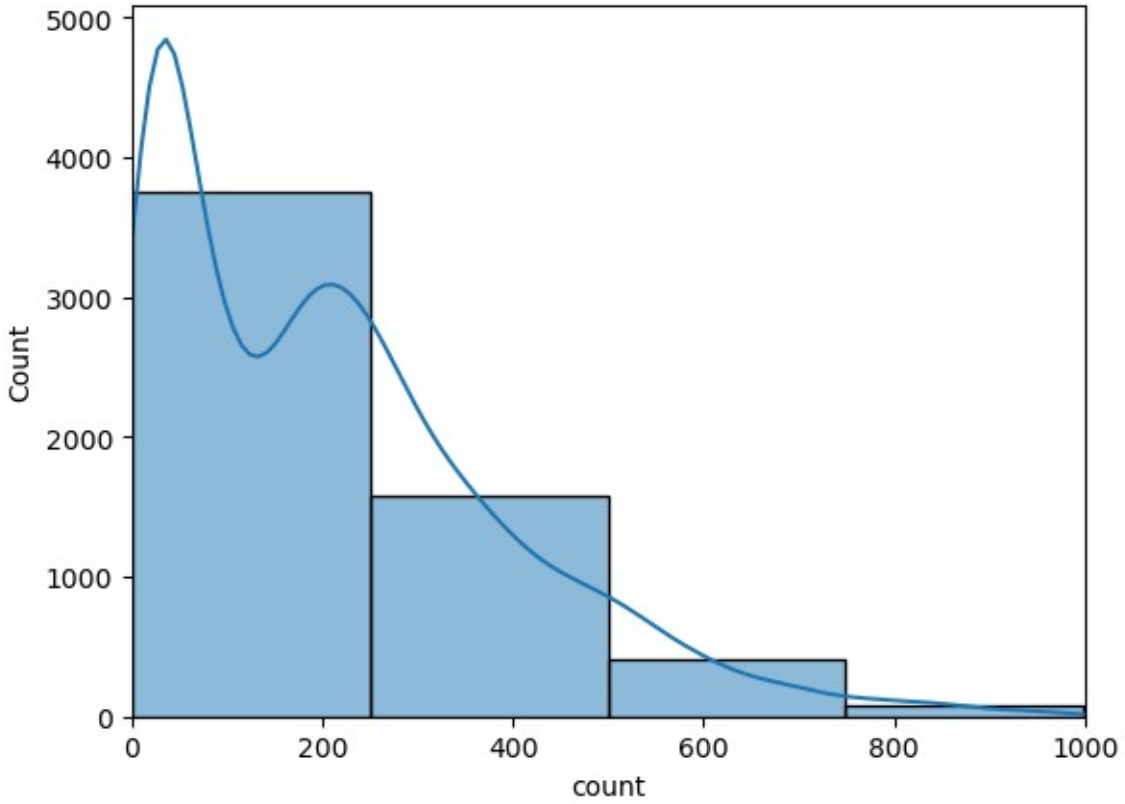
바람이 없거나 약한 정도(0 ~ 4)의 데이터 건수가 많다.

단변량 분석 – 습도 (humidity)



습도는 60 ~ 90인
즉 습합, 매우 습합 데이터가 가장 많은 것을 알 수 있다.

단변량 분석 – 자전거 대여량 (count)

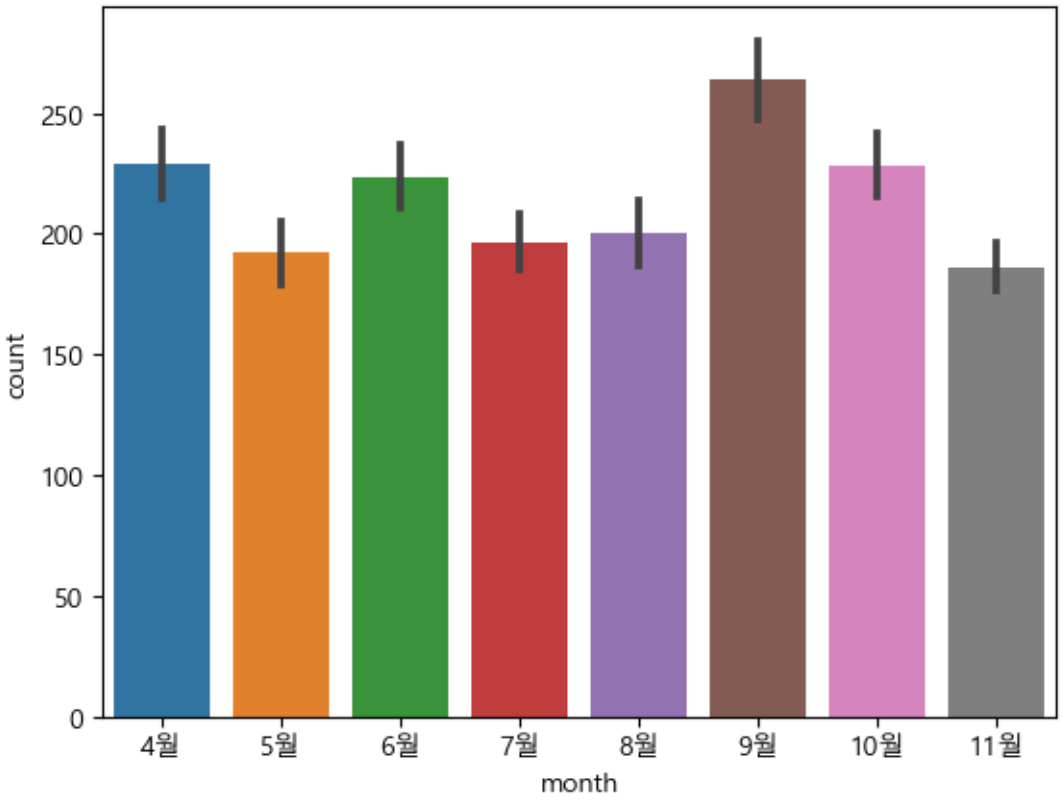


자전거 대여량 0 ~ 200인 구간이 데이터 건수가 많다.

가설 수립 (재확인)

- 가설 1 : 달(month)이 자전거 대여량(count)와 연관이 있는가?
- 가설 2 : 요일(day)이 자전거 대여량(count)와 연관이 있는가?
- 가설 3 : 시간대(hour)가 자전거 대여량(count)와 연관이 있는가?
- 가설 4 : 미세먼지(PM10)이 자전거 대여량(count)와 연관이 있는가?
- 가설 4 : 온도(temperature)가 자전거 대여량(count)와 연관이 있는가?
- 가설 5 : 강우 여부(precipitation)가 자전거 대여량(count)와 연관이 있는가?
- 가설 6 : 풍속(windspeed)가 자전거 대여량(count)와 연관이 있는가?
- 가설 7 : 습도(humidity)가 자전거 대여량(count)와 연관이 있는가?

이변량 분석(달(month) ->자전거 대여량(count)

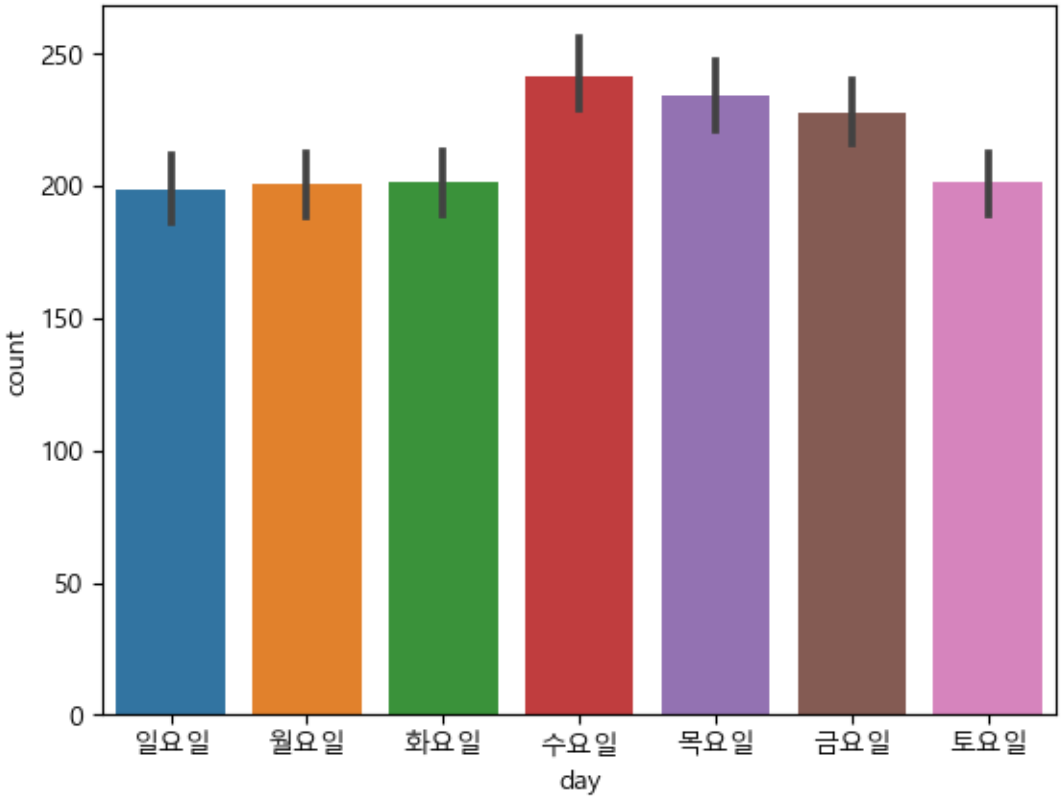



```
F_onewayResult(statistic=14.418555266868342, pvalue=1.0250409939149756e-18)
```

date 열에 있는 값을 이용해
4월, 5월 .. 11월 이루는 새로운 열(month) 추가 후
이변량 분석

F_onewayResult를 통해 달(month)와 자전거 대여량은
연관이 있다고 판단
4, 6, 9, 10월이 자전거 대여량이 많은 것으로 추론
8월에서 9월로 넘어가는 사이에 자전거 점검을 하는 것이
좋아보인다.

이변량 분석(요일(day) ->자전거 대여량(count)

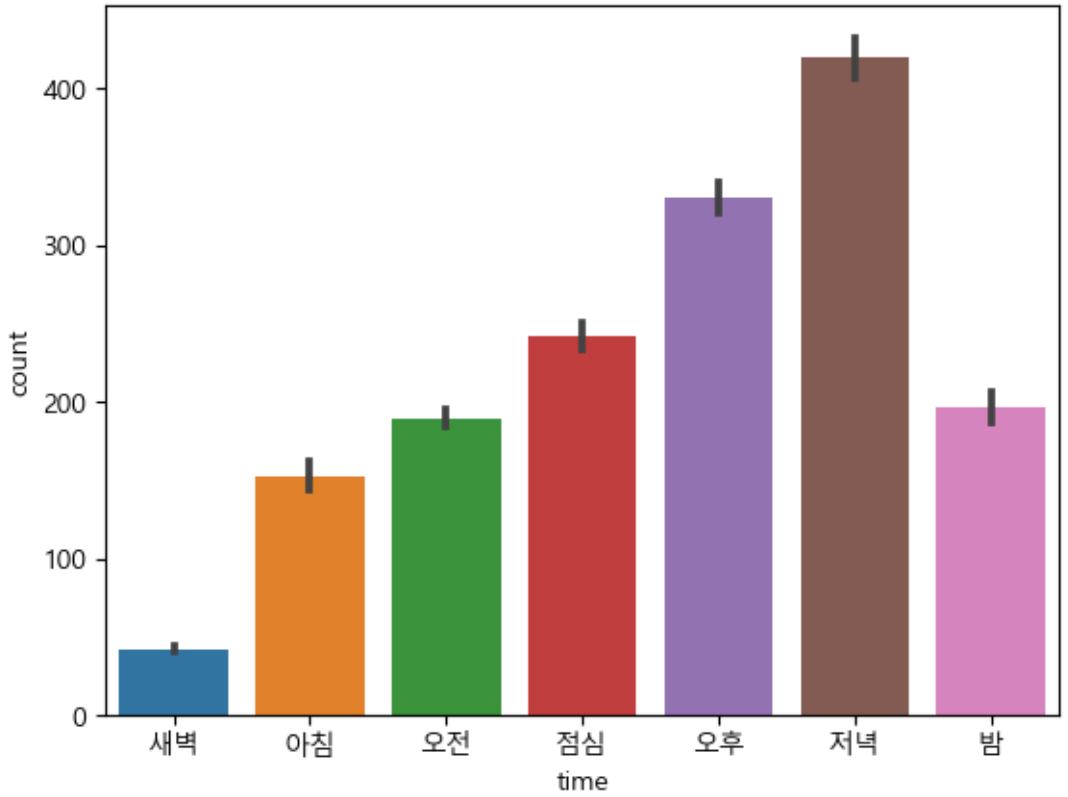



```
F_onewayResult(statistic=8.317297883731852, pvalue=5.378942897976812e-09)
```

date 열에 있는 값에서 dt.weekday를 통해
0, 1, 2, 3, 4, 5, 6을 각각 월요일, 화요일, … 일요일로 변환
새로운 열(day) 추가 후 이변량 분석

F_onewayResult를 통해 요일(day)와 자전거 대여량은
연관이 있다고 판단
수요일, 목요일, 금요일이 자전거 대여량이 많은 것으로 추론
하지만 비교적 약한 연관이 있다고 판단

이변량 분석(시간(time) ->자전거 대여량(count)




```
F_onewayResult(statistic=971.6523606007573, pvalue=0.0)
```

시간대(hour)를 새벽(0 <= < 6), 아침(6 <= < 9), 오전 (9 <= < 12), 점심 (ex. 12 <= < 14), 오후(ex. 14 <= < 18), 저녁 (ex. 18 <= < 22), 밤 (ex. 22 <= < 24) 으로 범주화 후 이변량 분석

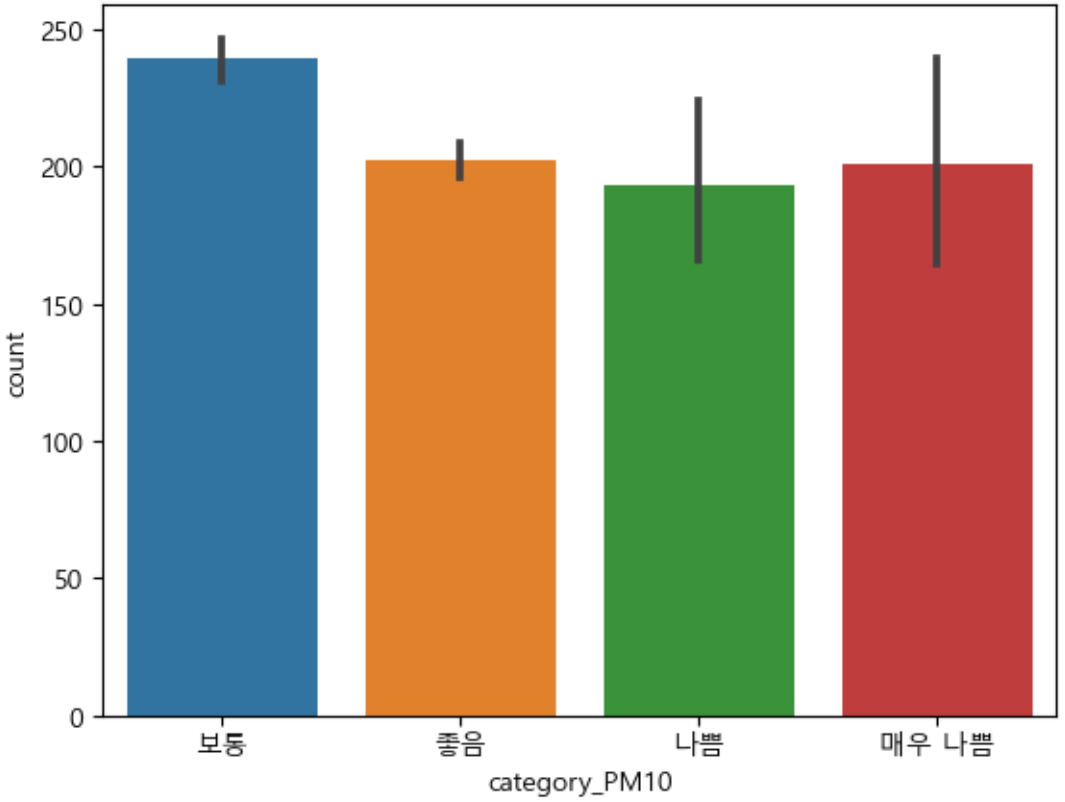
F_onewayResult를 통해 시간(time)과 자전거 대여량은 매우 연관이 있다고 판단

오후, 저녁이 자전거 대여량이 많은 것으로 추론

새벽 -> 아침으로 넘어갈 때 자전거 대여량 급상승 하는데 아마도 출근할 때 사람들이 자전거를 많이 사용하지 않을까 추론

점심 -> 오후
오후 -> 저녁으로 넘어갈 때도 자전거 대여량 급상승하는데 아마도 퇴근할 때도 사람들이 자전거를 많이 사용하지 않을까 추론

이변량 분석(미세먼지 정도(category_PM10) ->자전거 대여량(count)



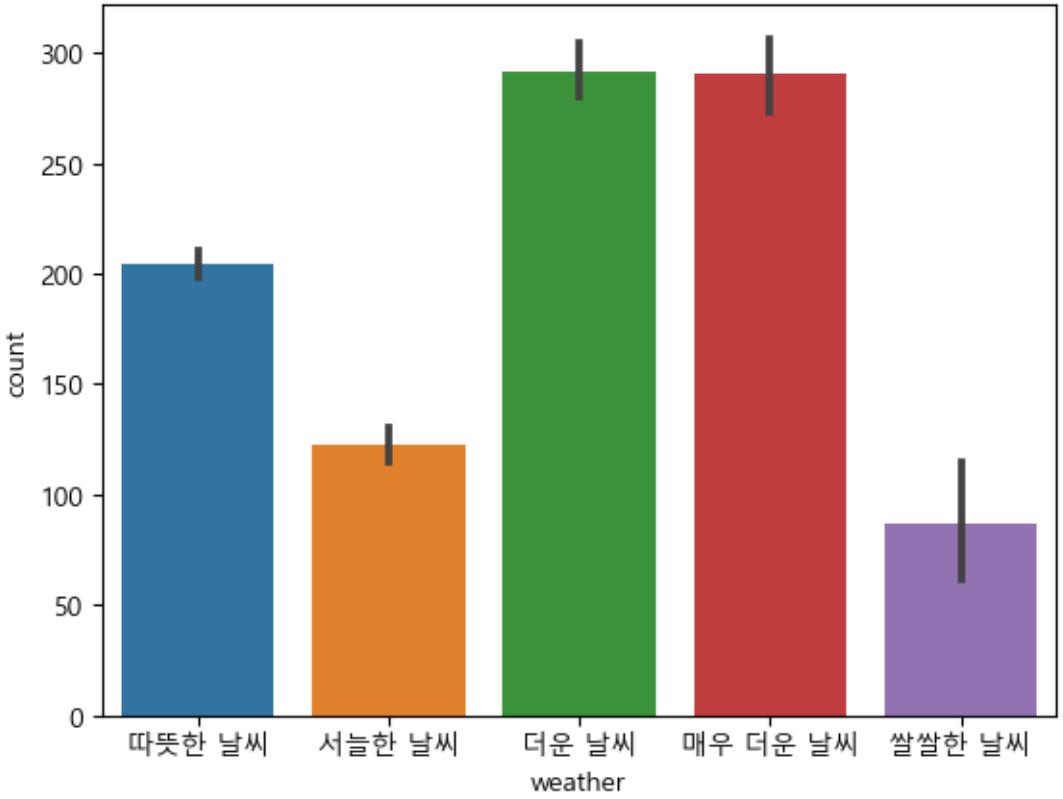

```
F_onewayResult(statistic=18.396308492448572, pvalue=7.0933081064829544e-12)
```

미세먼지(PM10)을 좋음 (0 <= < 31), 보통 (31 <= < 81),
나쁨 (81 <= < 151), 매우 나쁨 (151 <= < 250),
위험 (250 <= < 828)으로 범주화 한 후 이변량 분석

F_onewayResult를 통해 미세먼지(PM10)와
자전거 대여량은 연관이 있다고 판단

미세먼지 농도가 보통일 때 자전거 대여량이 많은 것으로 추론

이변량 분석(온도 정도(weather) ->자전거 대여량(count)




```
F_onewayResult(statistic=121.55416173385122, pvalue=8.911711252667439e-100)
```

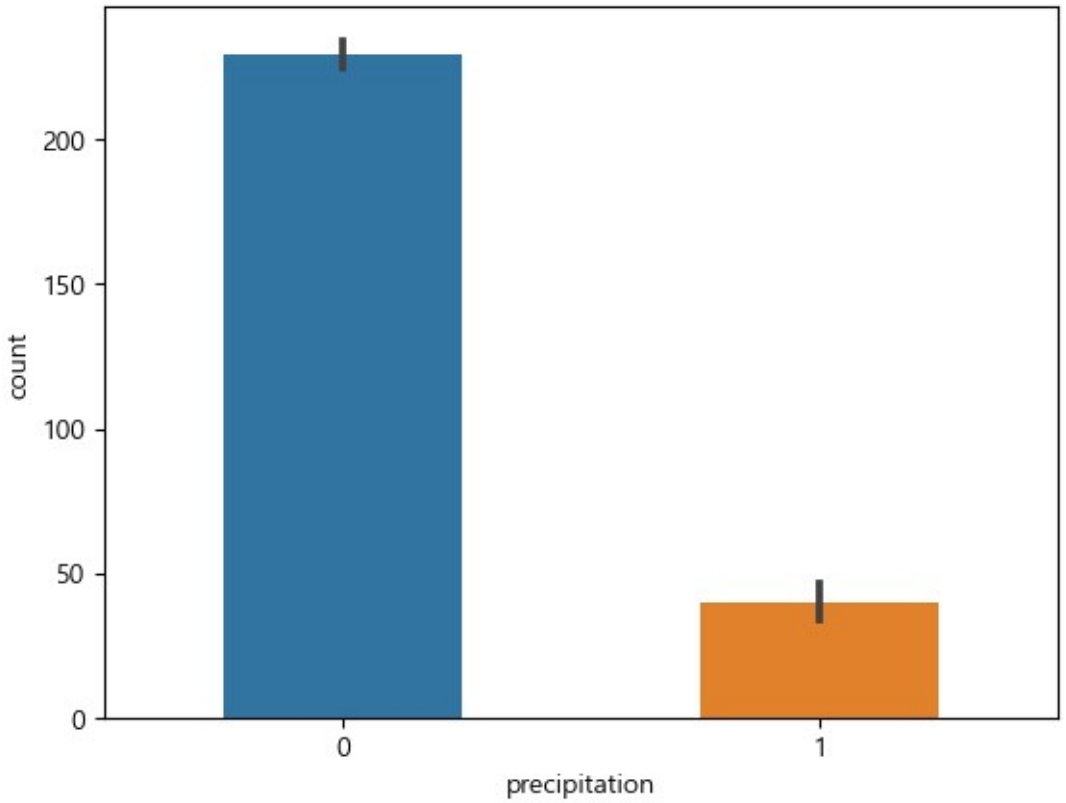
온도(temperature)을 쌀쌀한 날씨: (-3.5 <= < 0),
서늘한 날씨(0 <= < 10), 따뜻한 날씨: (10 <= < 25),
더운 날씨: (25 <= < 30), 매우 더운 날씨: (30 <= < 36.4)
범주화 한 후 이변량 분석

F_onewayResult를 통해 온도 정도(weather)와 자전거 대여량은 강한 연관이 있다고 판단

더운 날씨, 매우 더운 날씨일 때 자전거 대여량이 많은 것으로 추론

서늘한 날씨일 때 자전거를 전체적으로 수리하는 것이 좋다고 판단
(서늘한 날씨 -> 따뜻한 날씨 -> 더운 날씨로 갈수록 기하급수적으로 자전거 대여량이 높아지기 때문)

이변량 분석(강우 여부(precipitation) ->자전거 대여량(count)

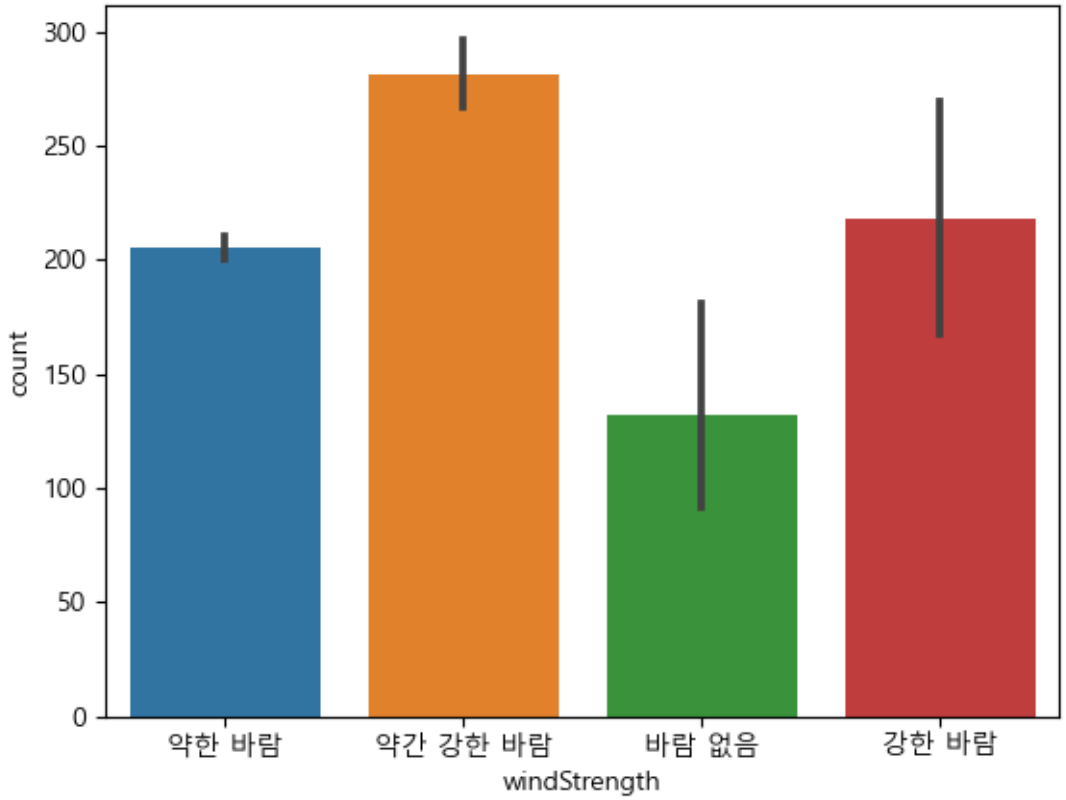



```
Ttest_indResult(statistic=21.389614151911022, pvalue=8.86239184041254e-98)
```

Ttest_indResult를 통해 강우 여부(precipitation)와 자전거 대여량은 연관이 있다고 판단

기본적으로 해가 있을 때 자전거를 많이 탄다고 판단

이변량 분석(풍속 정도(windStrength) ->자전거 대여량(count)



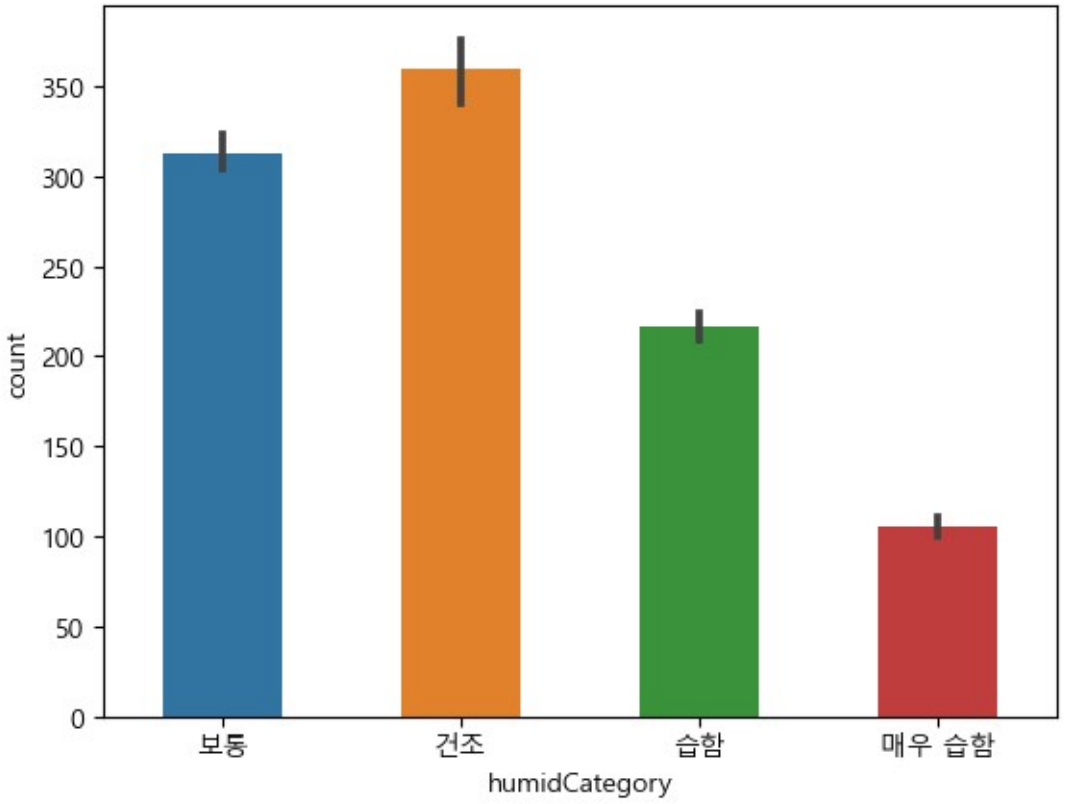

```
F_onewayResult(statistic=41.54021634414594, pvalue=1.4897443950234392e-26)
```

풍속을 바람 없음(0 <= < 0.3), 약한 바람: (0.3 <= < 3.4),
약간 강한 바람: (3.4 <= < 8.0), 강한 바람: (8.0 <= < 10.0)으로
분류하여 범주화 한 후 이변량 분석

F_onewayResult를 통해 풍속 정도(windStrength)와
자전거 대여량은 연관이 있다고 판단

약간 강한 바람이 불 때 자전거 대여량이 많은 것으로 추론

이변량 분석(습도 정도(humidityCategory) ->자전거 대여량(count)




```
F_onewayResult(statistic=521.498087710674, pvalue=3.492954240514575e-300)
```

습도를 건조: (19 <= < 41), 보통: (41 <= < 60),
습함: (60 <= < 80), 매우 습함: (80 <= < 101)으로
범주화 한 후 이변량 분석

F_onewayResult를 통해 습도 정도(humidityCategory)와 자전거 대여량은 강한 연관이 있다고 판단

습도가 건조 ~ 보통일 때 자전거 대여량이 높으며 이는 날씨가 비교적 따뜻할 때와 일맥상통하다.

Insight 도출 (우선 고려사항)

- 출근, 퇴근 시 사람들이 자전거를 많이 이용한다.
  따라서 아침에는 거주지역 주변에 자전거를 많이 배치하고
  오후에서 저녁 넘어가는 시점에 회사 주변에 자전거를 많이 배치한다.

- 25도가 넘어가는 더운 날씨일 때 사람들은 자전거를 많이 이용한다.
  따라서 서늘한 날씨 (0 ~ 10)일 때 자전거를 전체적으로 수리하고,
  따뜻한 날씨, 더운 날씨일 때 사람들이 자전거를 문제 없이 사용하도록 해야 한다.

- 온도와 연관되는 습도
  습도가 보통, 건조일 때 즉 기온이 어느정도 있을 때 자전거를 많이 이용하는 것을 알 수 있다.

Insight 도출 (차기 고려사항)

- 4월, 6월 9월, 10월에 자전거를 많이 이용한다.
- 수요일, 목요일, 금요일에 자전거를 많이 이용한다.
- 기본적으로 비가 오지 않아야 한다.
- 미세먼지가 이슈가 되지 않는 농도가 자전거 대여가 높음을 알 수 있다.
- 사람들이 바람이 어느정도 불 때 라고 느끼는 풍속 정도가 자전거 대여가 높음을 알 수 있다.