인공지능

서울지역의 모기 활동량 분석 및 예측

2020105045_정초의

발표일 : 2023.11.22.

>> 목차

1. 개요 및 필요성

- 지구온난화, 이상기후가 나날이 심화됨에 따라 환경이 크게 바뀌고 있다.
- 올해 11월 초까지 한낮 기온이 20도를 오르내리는 초여름 날씨였고, 이상 고온 현상과 맞물려 모기 등 해충의 번식량이 급증해 큰 문제가 되고 있다.

이렇게 추운데 모기가?..."12월초까지 계속 물릴 것" 왜?

머니투데이 | 김지성 기자, 정진솔 기자

입동 코앞인데 모기 기승…"기후변화 탓에 활동 기간↑"

등록 2023.11.03 12:56:21 | 수정 2023.11.03 13:57:15

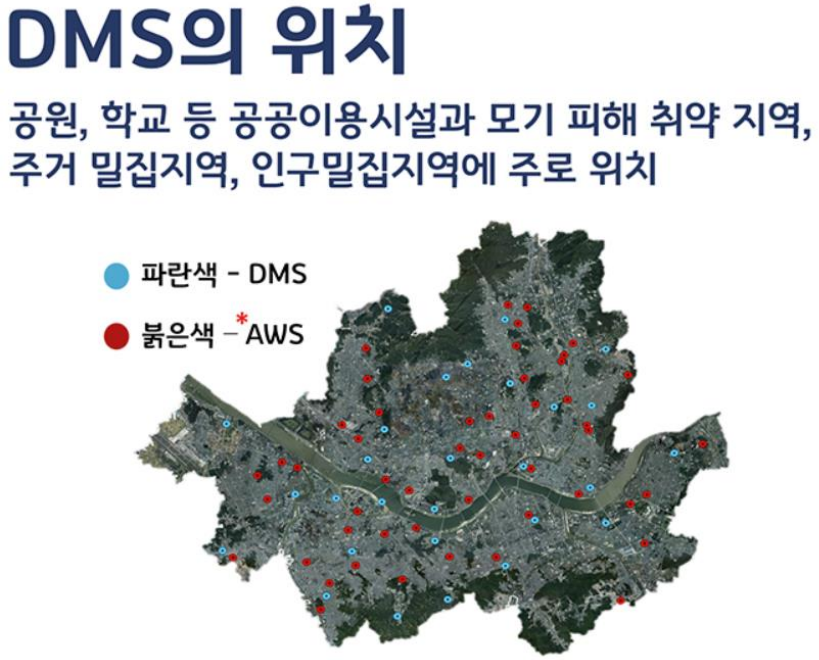
사회

한강공원에 송충이 닮은 해충 '습격'…기후변화에 피해 지속

송고시간 2023-11-05 09:55:12

2. 관련 내용

- 7일 서울시에 따르면 관내 디지털 모기 측정기(DMS) 51개를 통해 채집한 모기 수는 지난달 둘째 주 기준 총 933마리다.
- 9월 마지막 주 607마리보다 오히려 1.5배가량 늘었다. 지난해 같은 기간(357마리)에 비해서도 약 2.6배 증가했다.



출처 < 이렇게 추운데 모기가?…"12월초까지 계속 물릴 것" 왜? >, 김지성, 정진솔 기자, 머니투데이, 2023.11.08. , https://news.mt.co.kr/mtview.php?no=2023110722275399994

2. 관련 내용

- 질병관리청이 지난달 발표한 권역별 기후 변화 매개체 감시 현황에 따르면 지난달 1일부터 7일까지 전국 도심·철새도래지의 모기 트랩지수*는 47.1개체로 지난해(28.8개체)보다 63.6% 증가했다. 5년 평균치(41.8)와 비교해도 12.7% 늘었다.
- 도심은 같은 기간 트랩지수가 72.5개체로 작년에 비해 두 배 늘었다.

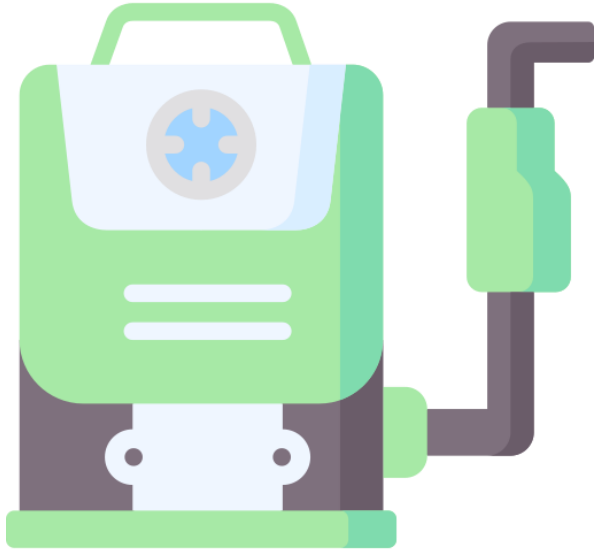
*트랩지수 : 하룻밤 모기 유인 포집기(트랩) 한 대에서 잡힌 모기 개체수

출처 < 입동 코앞인데, 초여름 같은 11월…모기까지 극성 >, 김우섭 , 한국경제, 2023.11.02. ,
https://www.hankyung.com/article/2023110288871

3. 기대 효과

- 목표 : 온도, 강수량 등 환경요인에 따른 모기 활동량 예측 모델

방제, 방역 활동을 촉진하여
피해를 줄일 수 있다.



기관, 사회 차원에서 시설을
효율적으로 관리할 수 있다.

4. 데이터셋

출처
https://www.kaggle.com/datasets/kukuroo3/mosquito-indicator-in-seoul-korea/data

KUKUROO3 · UPDATED 2 YEARS AGO

73 | New Notebook | Download (14 kB)

Mosquito Indicator in Seoul, Korea

2016~ 2019, daily data with Temperature, precipitation

About Dataset

- Context
  - This dataset deals with Mosquito Indicator and Temperature, precipitation in Seoul, South Korea.
  - Mosquito Indicator is a number of Mosquito per Specific area.

- Content
  - This data provides six cols (date, mosquito_Indica, rain(mm), mean_T(°C), min_T(°C), max_T(°C))
  - Data were measured every day between 2016-05 and 2019-12.
  - Data were measured for Specific area in Seoul.

4. 데이터셋

출처
https://www.kaggle.com/datasets/kukuroo3/mosquito-indicator-in-seoul-korea/data

KUKUROO3 · UPDATED 2 YEARS AGO

73 | New Notebook | Download (14 kB)

Mosquito Indicator in Seoul, Korea

2016~ 2019, daily data with Temperature, precipitation

- **date :** YYYY-MM-DD - **날짜**
- **mosquito_Indicator :** number of mosquito per area – **모기 활동지수**
- **rain(mm) :** daily precipitation - **일강수량**
- **mean_T(°C) :** mean temperature of day - **일평균기온**
- **min_T(°C) :** min temperature of day - **일최저기온**
- **max_T(°C) :** max temperature of day - **일최고기온**

5. 데이터 분석

1. 라이브러리 불러오기

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

2. csv 파일 로드

```
df = pd.read_csv('/kaggle/input/mosquito-indicator-in-seoul-korea/mosquito_Indicator.csv')
```

5. 데이터 분석

df.head()

	date	mosquito_Indicator	rain(mm)	mean_T(℃)	min_T(℃)	max_T(℃)
0	2016-05-01	254.4	0.0	18.8	12.2	26.0
1	2016-05-02	273.5	16.5	21.1	16.5	28.4
2	2016-05-03	304.0	27.0	12.9	8.9	17.6
3	2016-05-04	256.2	0.0	15.7	10.2	20.6
4	2016-05-05	243.8	7.5	18.9	10.2	26.9

df.describe().T

	count	mean	std	min	25%	50%	75%	max
mosquito_Indicator	1342.0	251.991803	295.871336	0.0	5.5	91.9	480.400	1000.0
rain(mm)	1342.0	3.539866	13.868106	0.0	0.0	0.0	0.400	144.5
mean_T(℃)	1342.0	14.166021	10.943990	-14.8	4.5	16.5	23.300	33.7
min_T(℃)	1342.0	10.005663	11.109489	-17.8	0.3	11.5	19.500	30.3
max_T(℃)	1342.0	19.096870	11.063394	-10.7	9.3	21.9	28.175	39.6

df.info()

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1342 entries, 0 to 1341
Data columns (total 6 columns):
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
 0   date                1342 non-null   object
 1   mosquito_Indicator  1342 non-null   float64
 2   rain(mm)            1342 non-null   float64
 3   mean_T(℃)           1342 non-null   float64
 4   min_T(℃)            1342 non-null   float64
 5   max_T(℃)            1342 non-null   float64
dtypes: float64(5), object(1)
memory usage: 63.0+ KB
```

6. 데이터 전처리

1. date를 string에서 date 타입으로 변환

```
df['date'] = pd.to_datetime(df['date'])
```

2. 결측치 확인

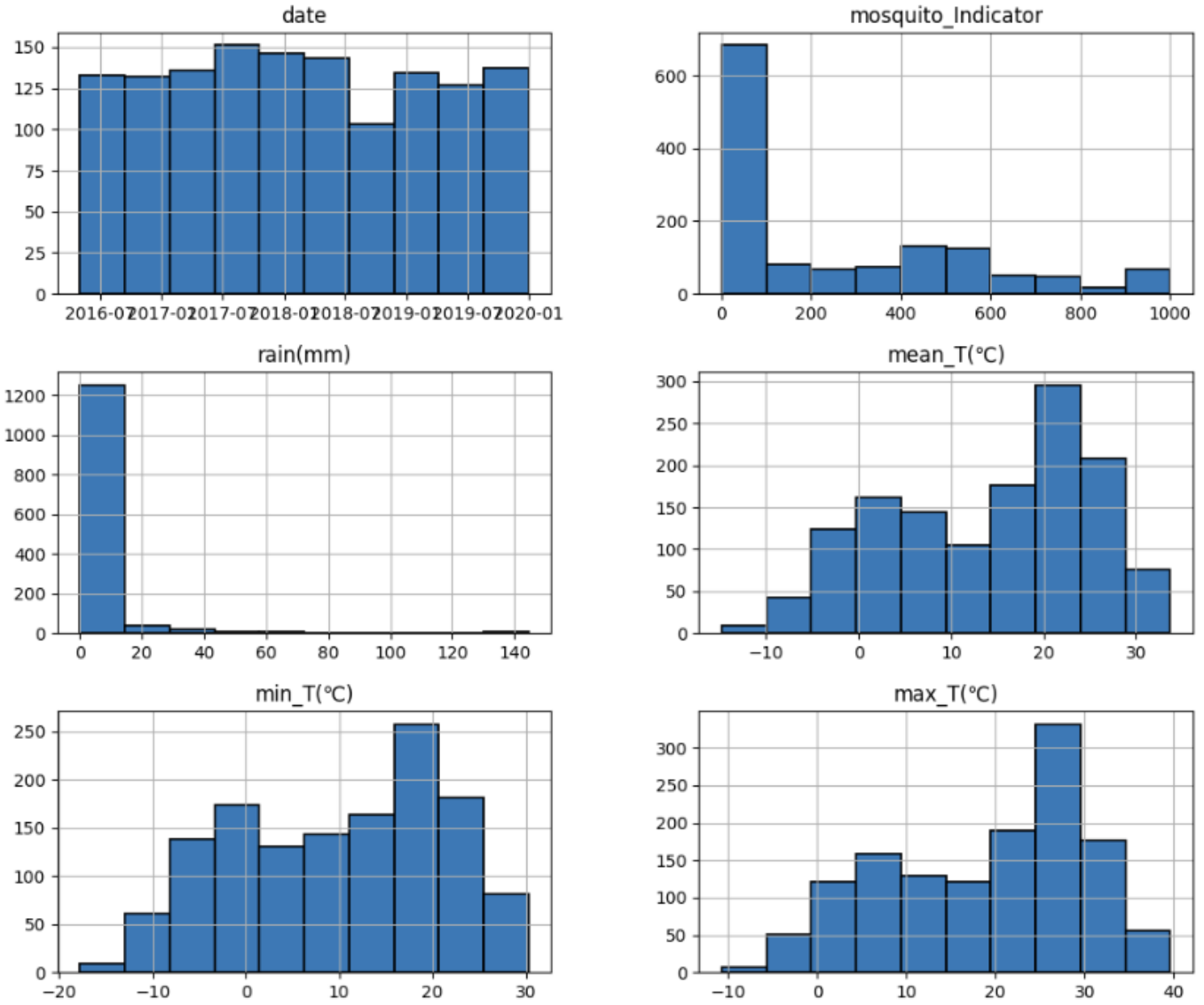
```
df.isnull().sum()
```

```
date                0
mosquito_Indicator  0
rain(mm)            0
mean_T(℃)           0
min_T(℃)            0
max_T(℃)            0
dtype: int64
```

7. 상관관계 분석

추가한 부분

```
df.hist(edgecolor='black', linewidth=1.2)
fig = plt.gcf()
fig.set_size_inches(12,10)
plt.show()
```

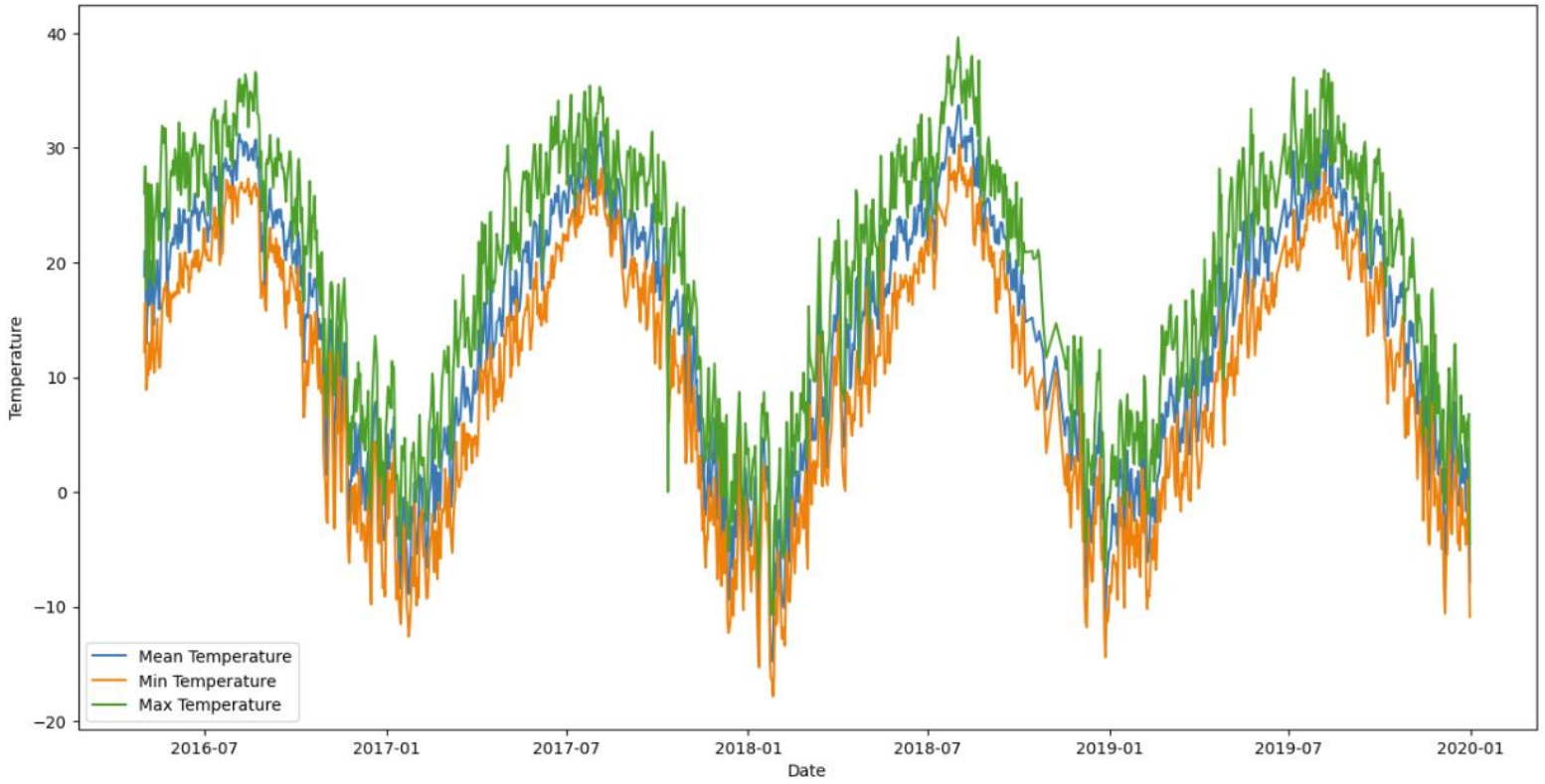


7. 상관관계 분석

추가한 부분

```
plt.figure(figsize=(16,8))
sns.lineplot(data=df, x='date', y='max_T(℃)', label='Max Temperature')
sns.lineplot(data=df, x='date', y='mean_T(℃)', label='Mean Temperature')
sns.lineplot(data=df, x='date', y='min_T(℃)', label='Min Temperature')

plt.xlabel('Date')
plt.ylabel('Temperature')
plt.legend()
plt.show()
```

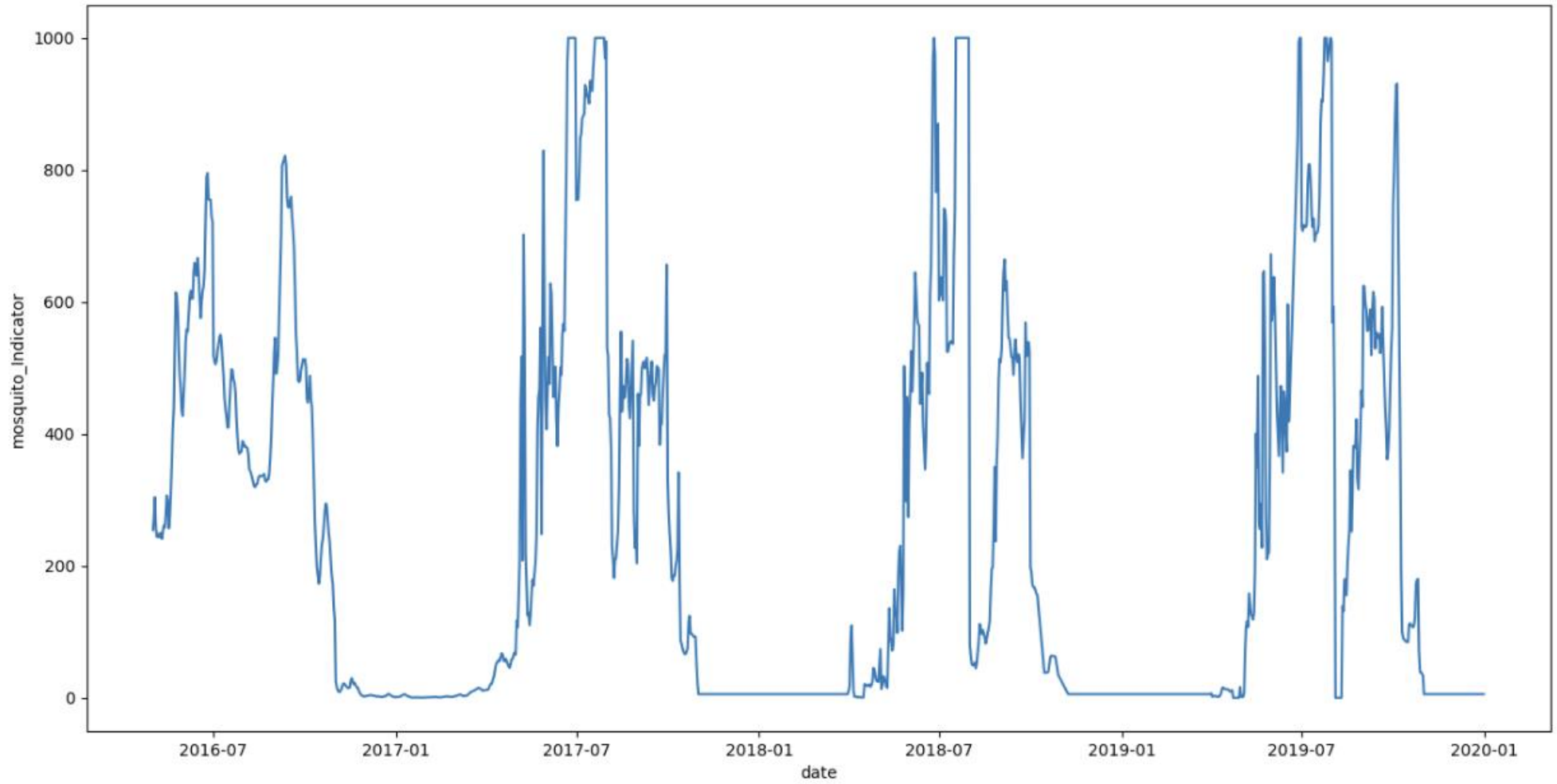


- 연도별 온도변화 추이는 비슷한 양상을 띄고 있다.

7. 상관관계 분석

```
plt.figure(figsize=(16,8))
sns.lineplot(data=df, x="date", y="mosquito_Indicator")
plt.show()
```

- 날짜에 따른 모기 활동지수
- 겨울에는 모기 활동지수가 급감한다.

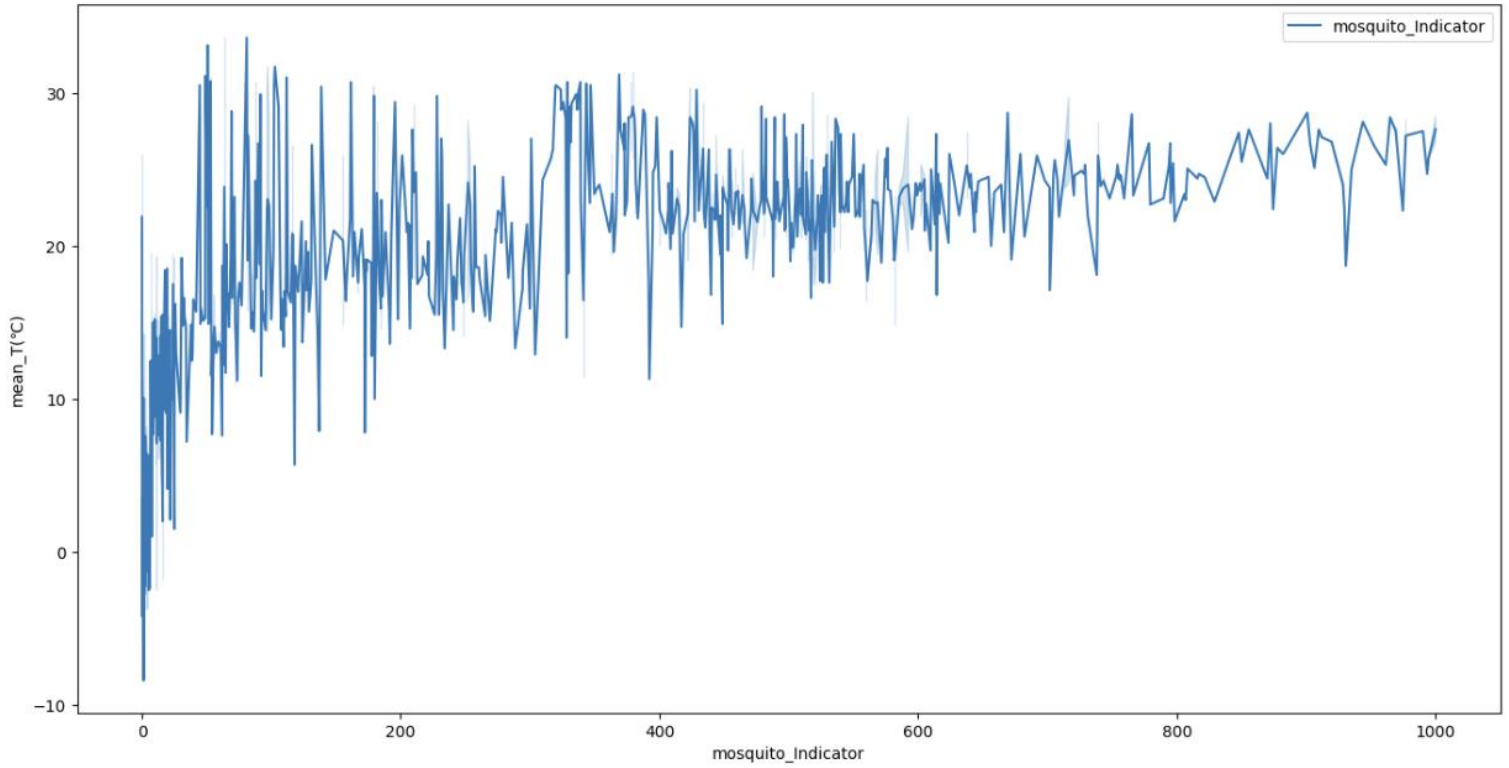


7. 상관관계 분석

추가한 부분

```
plt.figure(figsize=(16,8))
sns.lineplot(data=df, x='mosquito_Indicator', y='mean_T(℃)', label='mosquito_Indicator')

plt.xlabel('mosquito_Indicator')
plt.ylabel('mean_T(℃)')
plt.legend()
plt.show()
```



- 평균 온도에 대한 모기 활동지수
- 대략 20~30℃일 때 모기 활동지수가 높다.
- 온도가 너무 높거나 낮아도 모기 활동 지수는 감소한다.

7. 상관관계 분석

```
df_max_dates = df.loc[df['mosquito_Indicator'] == df['mosquito_Indicator'].max()]
df_max_dates.head()
```

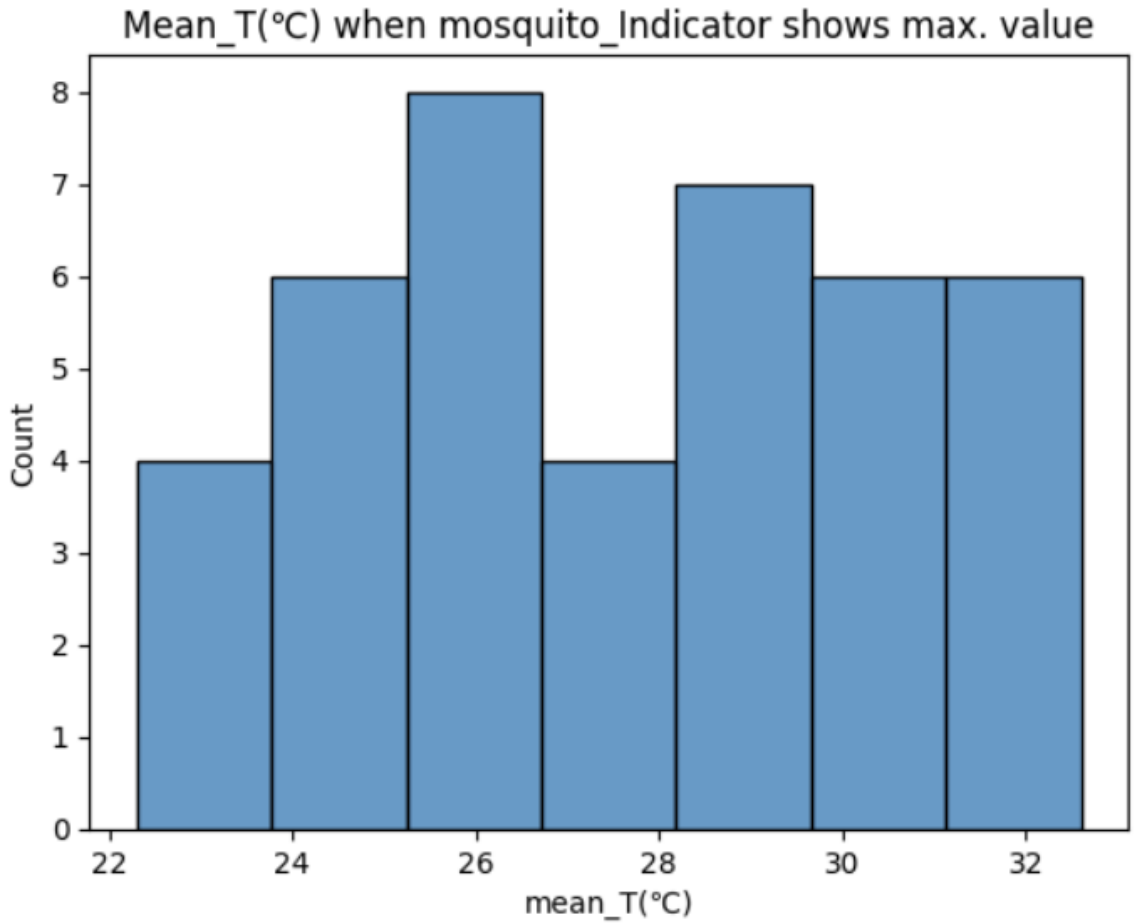
	date	mosquito_Indicator	rain(mm)	mean_T(℃)	min_T(℃)	max_T(℃)	year
418	2017-06-23	1000.0	0.0	26.7	21.3	34.1	2017
419	2017-06-24	1000.0	3.5	24.5	21.5	27.8	2017
420	2017-06-24	1000.0	3.5	24.5	21.5	27.8	2017
421	2017-06-25	1000.0	1.5	23.8	21.0	28.4	2017
422	2017-06-26	1000.0	29.0	23.1	20.1	29.9	2017
423	2017-06-26	1000.0	29.0	23.1	20.1	29.9	2017
424	2017-06-27	1000.0	0.0	25.2	20.8	30.5	2017
425	2017-06-28	1000.0	0.0	26.0	22.5	30.5	2017
426	2017-06-29	1000.0	0.0	26.2	21.9	31.5	2017
427	2017-06-30	1000.0	0.0	25.9	22.2	30.9	2017

- 모기 활동수치가 최대인 데이터를 가져와 새로운 데이터 프레임을 만든다.

7. 상관관계 분석

추가한 부분

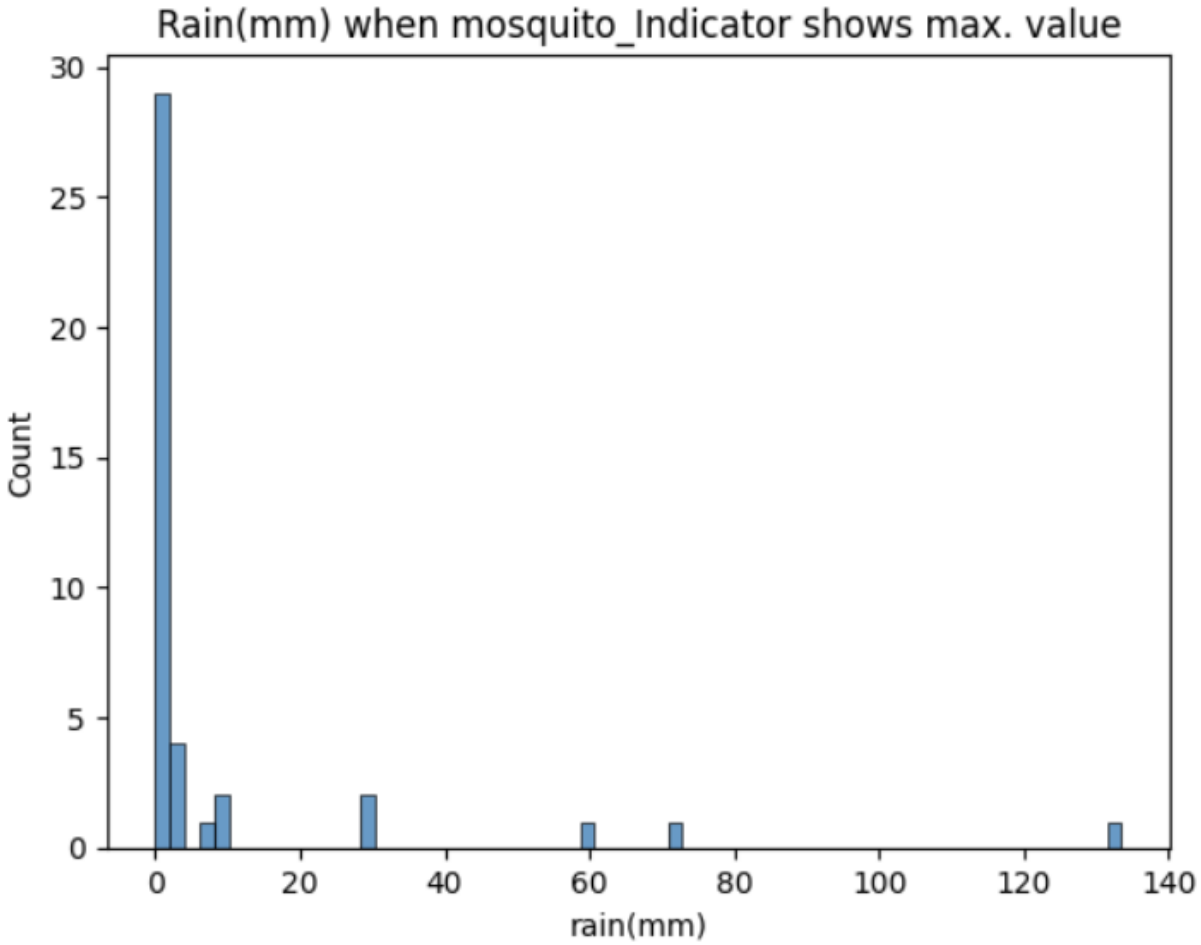
```
g2 = sns.histplot(df_max_dates['mean_T(℃)'])
g2.set(title = 'Mean_T(℃) when mosquito_Indicator shows max. value')
plt.show()
```



- 모기 활동지수가 최대인 데이터에 대해 평균 온도 분포
- 온도가 약 22~32 ℃ 일 때 모기 활동지수가 가장 높다.

7. 상관관계 분석

```
g1 = sns.histplot(df_max_dates['rain(mm)'])
g1.set(title = 'Rain(mm) when mosquito_Indicator shows max. value')
plt.show()
```



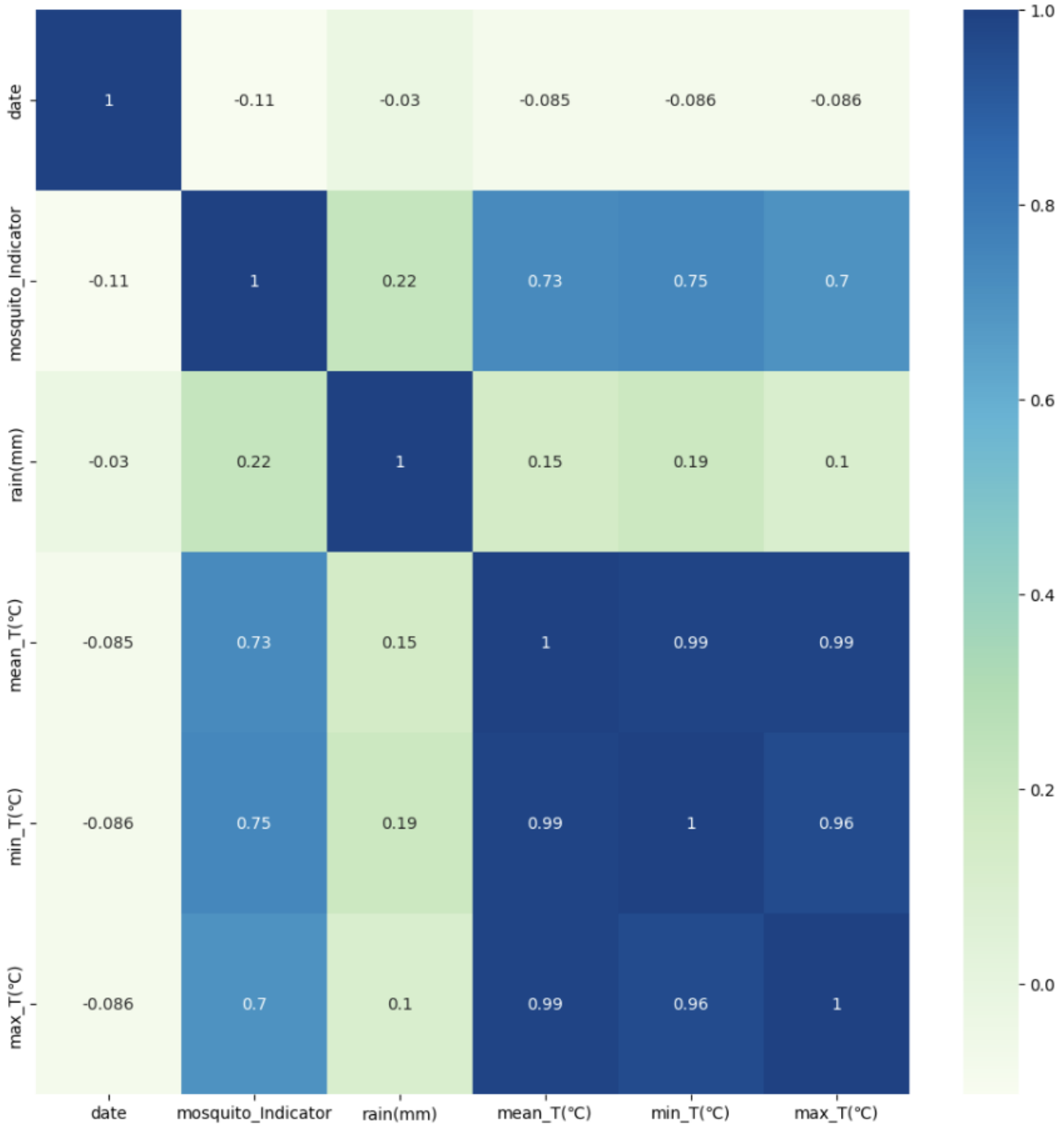
- 모기 활동지수가 최대인 데이터에 대해 강수량 분포
- 강수량이 낮을 수록 모기가 활동하기 쉬울 것이다.

7. 상관관계 분석

```
#Calculating Correlation
corr=df.corr()

#Plotting Correlation
plt.figure(figsize=(12,12))
sns.heatmap(corr,annot=True,cmap="GnBu")
```

- 모기 활동지수와 관련성이 높은 속성은 일최저기온, 일평균기온, 일최고기온이다.



8. 데이터 분할

```
from sklearn.model_selection import train_test_split    # 데이터 분할 모듈

train, test = train_test_split(df, train_size = 0.8)

# 학습용 문제, 정답
train_X = train[['min_T(℃)']]
train_y = train.mosquito_Indicator

# 테스트용 문제, 정답
test_X = test[['min_T(℃)']]
test_y = test.mosquito_Indicator
```

- 학습용 데이터와 테스트용 데이터를 8 : 2 의 비율로 분리
- 문제는 온도, 정답은 모기 활동지수로 설정

9. 모델 학습

추가한 부분

```
from sklearn.tree import DecisionTreeRegressor
from sklearn.neighbors import KNeighborsRegressor
from sklearn.linear_model import LinearRegression
from sklearn.ensemble import RandomForestRegressor
from sklearn.ensemble import GradientBoostingRegressor
```

- 예측 알고리즘 모델 학습
- 모델 정확도 비교 : GradientBoostingRegressor > RandomForestRegressor > KNeighborsRegressor > DecisionTreeRegressor > LinearRegression

```
DT = DecisionTreeRegressor()
DT.fit(train_X, train_y)
score = DT.score(test_X, test_y)
print('Score:', format(score,'.3f'))
```

Score : 0.739

```
KN = KNeighborsRegressor()
KN.fit(train_X, train_y)
score = KN.score(test_X, test_y)
print('Score:', format(score,'.3f'))
```

Score : 0.750

```
RF = RandomForestRegressor()
RF.fit(train_X, train_y)
score = RF.score(test_X, test_y)
print('Score:', format(score,'.3f'))
```

Score : 0.774

```
LR = LinearRegression()
LR.fit(train_X, train_y)
score = LR.score(test_X, test_y)
print('Score:', format(score,'.3f'))
```

Score : 0.606

```
GB = GradientBoostingRegressor()
GB.fit(train_X, train_y)
score = GB.score(test_X, test_y)
print('Score:', format(score,'.3f'))
```

Score : 0.807

10. 소감

- 관심 가는 주제를 선택해 데이터 분석, 머신러닝을 학습할 수 있는 좋은 기회였다.
- 관심 있는 소재라 선정했지만, 컬럼이 적은 데이터셋을 이용하여 모기 활동량에 영향을 끼치는 요인을 다방면으로 분석하지 못한 것 같아 아쉬웠다.
- 관측지의 습도, 인구밀집도, 지리 형태, 방역 상태 등 더 다양한 항목의 데이터가 수집된 데이터셋이 있다면 모기 활동량에 대한 상관관계를 다시금 분석하고 싶다.

감사합니다