



# 기후 데이터를 활용한 모기 발생 예보

10조 Synergy

2021-07-20

김형림, 고아름, 남예은



# 1. 프로젝트 개요

# 구성원 및 역할

## 1. 프로젝트 개요



**Synergy**

깃허브 마스코트 '옥토캣',  
고양이인 줄 알았는데 문어였다?!

다양한 사람들이 모여 시너지를 내는 팀!

김형림



프로젝트 총괄

데이터 수집 및  
전처리

시각화

예측 모델링

고아름



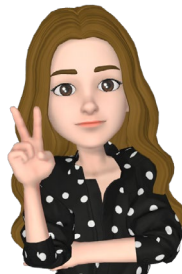
데이터 수집

데이터 분석

시각화

모델 검증

남예은



데이터 수집

데이터 분석

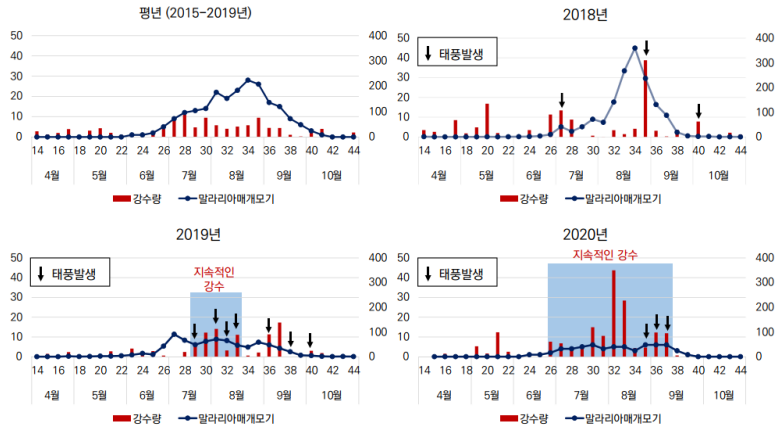
시각화

가설 검증

# 기후 변화와 모기매체 질병

## 1. 프로젝트 개요

### 기후 변화와 모기



- 일평균기온 혹은 최고기온이 1℃ 상승함에 따라 일주일 후 모기 성체 개체수가 27% 증가할 가능성이 있으며 상대습도와 강수량은 양의 상관성을 보임
- 기후 변화로 기온 상승 시 매개모기의 밀도 증가가 예상되며 이로 인해서 환자수가 다시 증가할 가능성

출처=  
2017, 「주요환경변화에 따른 미래 감염병의 발생 양상」, 질병관리본부 미래감염병대비과, p.1024~1028  
2020, 「한국 기후 변화 평가보고서」, 환경부  
채수미, 2014, 「기온과 지역 특성이 말라리아 발생에 미치는 영향」

### 모기매개 질병

최근 5년간 말라리아·일본뇌염 발생 현황 (단위: 명)



#### 모기 매개 감염병

국내 말라리아, 일본뇌염

해외 지카바이러스감염증, 황열, 뎡기열, 웨스트나일열, 치쿤구니아열

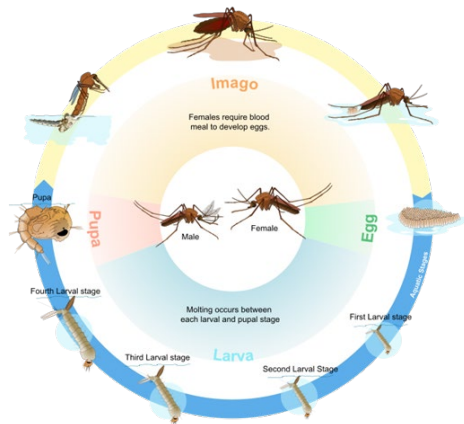


- 기후 변화에 의해 우리나라에서 발생될 모기매개 감염병 : 삼일열 말라리아, 일본뇌염
- 해외 유입 예상 감염병 : 지카, 뎡기열, 황열, 웨스트나일열, 치쿤구니아열, 열대열 말라리아 등
- 말라리아 환자가 밀집되어 있는 서울, 경기, 인천, 강원에서 기온 1℃ 상승 시 발생 위험 10.8%, 12.7%, 14.2%, 20.8% 증가

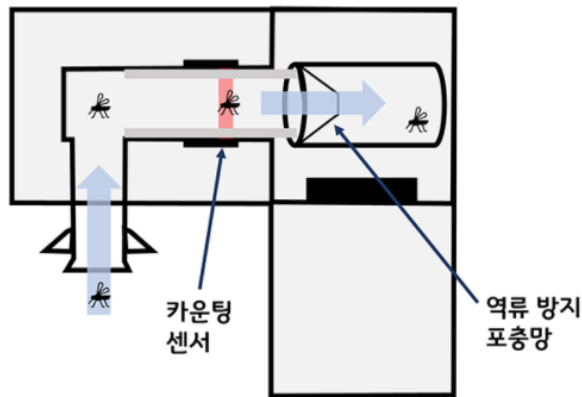
# 모기 발생 요인, 모기채집 방법, 서울시 모기예보

## 1. 프로젝트 개요

### 모기의 생활사



### 모기채집 방법



### 서울시 모기예보제

## 서울시 모기예보제

환경요인을 고려해 모기활동지수를 산정하고,  
4단계 모기 발생 예보를 제공하는 서비스



- 1단계** 쾌적, 일 평균 모기 개체수 0~30
- 2단계** 관심, 일 평균 모기 개체수 30~60
- 3단계** 주의, 일 평균 모기 개체수 60~90
- 4단계** 불쾌, 일 평균 모기 개체수 90~

- 유충 기간 : 1 ~ 2주
- 평균 산란수 : 30 ~ 100개/1회
- 종류 : 빨간집모기, 작은 빨간집모기, 동양집모기, 반점날개모기, 금빛숲모기, 한국숲모기, 흰줄숲모기, 등줄숲모기, 토고숲모기 등

- DMS(디지털 모기 측정기)를 활용하여 모기 개체수를 계측
- 유인제 방출구에서 CO<sub>2</sub>를 방출시켜 유인된 모기가 흡입구 근처에 오면 흡입 / 진공방식으로 흡입된 모기가 카운팅 센서를 지나 자동계수

- 서울지역 모기 발생 상황을 수치화하여 모기 발생 단계별 시민행동요령을 알려주는 일일 모기 발생 예보서비스

### 기후변화는 모기의 성장과 발달에 영향을 주어 모기 감염병 위험을 증가시킨다.



- > 기후변화로 인한 기온상승으로 인해 모기가 성충이 되는 시간이 단축되지 않을까?
- > 그럼 이에 따라 모기의 개체수도 증가하지 않을까?

#### <가설수립>

**H<sub>0</sub>** : 모기 성장 기간을 반영한 기후 데이터와  
모기 개체수 증가량과 **상관관계가 없다.**

**H<sub>1</sub>** : 모기 성장 기간을 반영한 기후 데이터와  
모기 개체수 증가량과 **상관관계가 있다.**

#### <가설검정>

입력 변수	P-value
temp(평균온도(°C))	≈0(1.02e-266)
rain_per_day(일강수량(mm))	≈0(2.99e-268)
wind(평균 풍속(m/s))	≈0(4.27e-268)
humidity(평균 상대습도(%))	≈0(1.16e-268)
sunshine(합계 일사량(MJ/m2))	≈0(1.43e-267)

종속변수 : mosquito (모기 개체수)

모든 유의 확률이 0.05 이하 → 귀무가설 기각

「모기 성장 기간을 반영한 기후데이터와  
모기 개체수 증가량과 **상관관계가 있다.**」



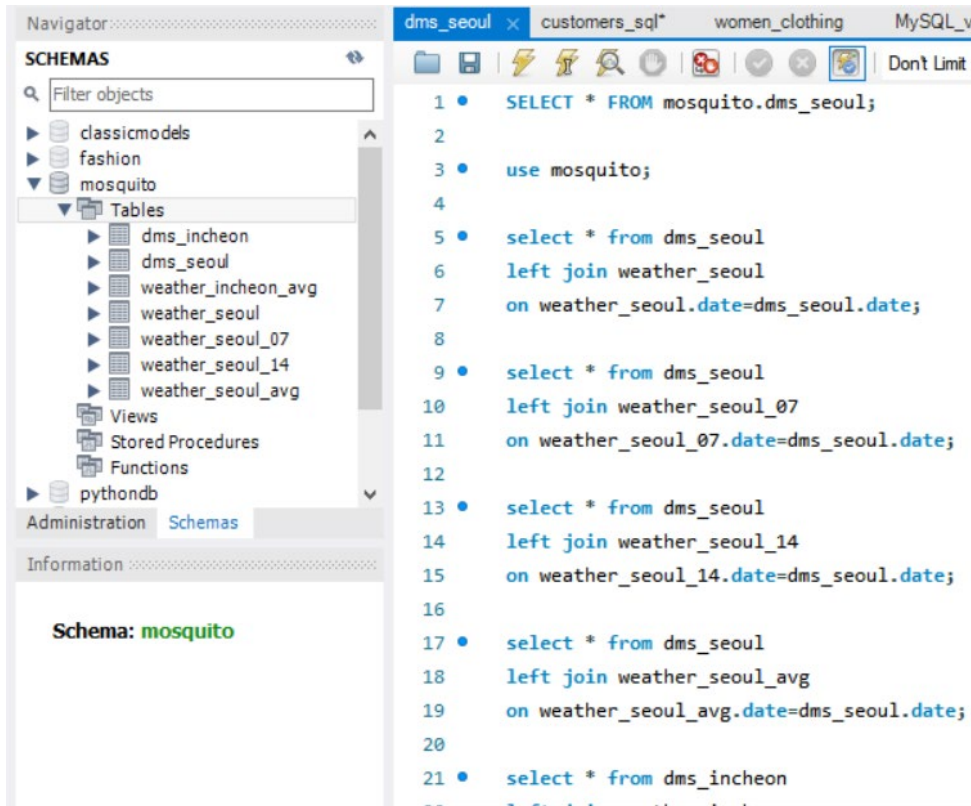
## 2. 프로세싱

이름	내용	기간	출처	비고
서울특별시 DMS 설치 현황	서울특별시의 DMS 설치 위치	2021년	서울특별시 보건환경연구원	구별 2개씩 총 50개
서울특별시 DMS 포집내역	모기를 포함한 일자별 DMS 벌레 포집량	2015년 01월 ~ 2020년 12월	서울특별시 보건환경연구원	매년 04월 ~ 11월
서울특별시 기후	일자별 종상기상관측(ASOS) 기후 데이터	2015년 01월 ~ 2020년 12월	기상청	



loc_num	지점	rain_per_day	일강수량(mm)	steam_pressure	평균 증기압(hPa)
loc_name	지점명	accum_rain	누적 강수량(mm)	sunshine_time	합계 일조시간(hr)
temp	평균기온(°C)	wind	평균 풍속(m/s)	sunshine	합계 일사량(MJ/m2)
l_temp	최저기온(°C)	dew	평균 이슬점온도(°C)	ground_surface_temp	평균 지면온도(°C)
h_temp	최고기온(°C)	humidity	평균 상대습도(%)	※ accum_rain은 rain_per_day의 누적 합계	





The screenshot shows a MySQL database interface. On the left, the 'Navigator' pane displays the 'mosquito' schema with tables: dms\_incheon, dms\_seoul, weather\_incheon\_avg, weather\_seoul, weather\_seoul\_07, weather\_seoul\_14, and weather\_seoul\_avg. The main pane shows SQL queries for the 'mosquito' database:

```
1 • SELECT * FROM mosquito.dms_seoul;
2
3 • use mosquito;
4
5 • select * from dms_seoul
6   left join weather_seoul
7   on weather_seoul.date=dms_seoul.date;
8
9 • select * from dms_seoul
10  left join weather_seoul_07
11  on weather_seoul_07.date=dms_seoul.date;
12
13 • select * from dms_seoul
14  left join weather_seoul_14
15  on weather_seoul_14.date=dms_seoul.date;
16
17 • select * from dms_seoul
18  left join weather_seoul_avg
19  on weather_seoul_avg.date=dms_seoul.date;
20
21 • select * from dms_incheon
```

서울특별시  
DMS  
포집내역



일별 기후 데이터

7일전 일별 기후  
데이터

14일전 일별  
기후 데이터

7일 평균 기후  
데이터

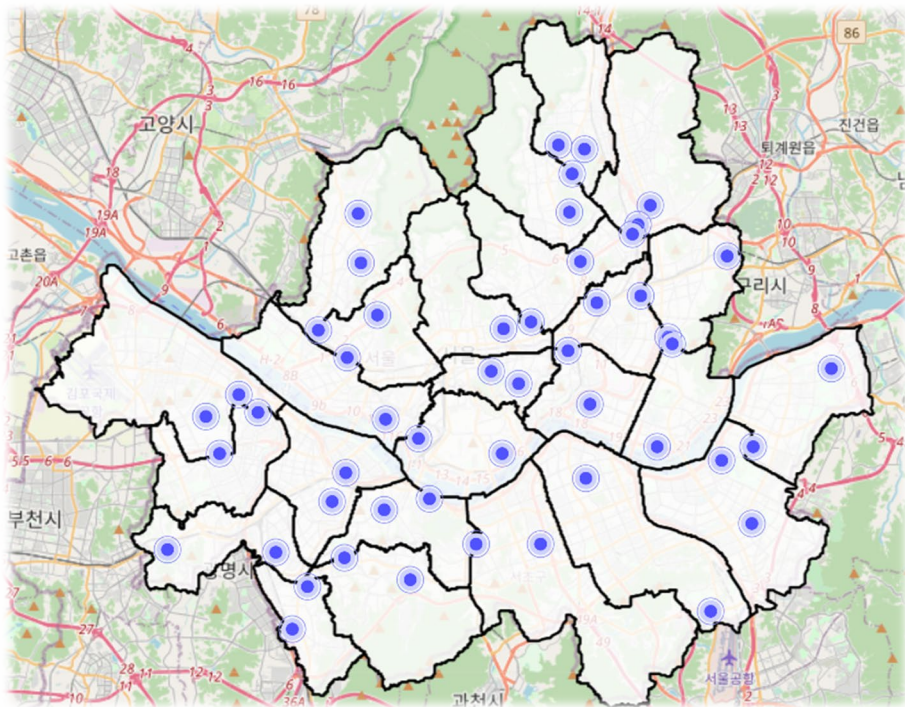
인천시  
DMS  
포집내역



14일전 일별  
기후 데이터

# 서울시 DMS(디지털 모기 측정기) 위치

## 2. 프로세싱



• 서울은 25개의 자치구가 있으며 구별로 DMS가 2개씩 설치되어 있음(총 50개)

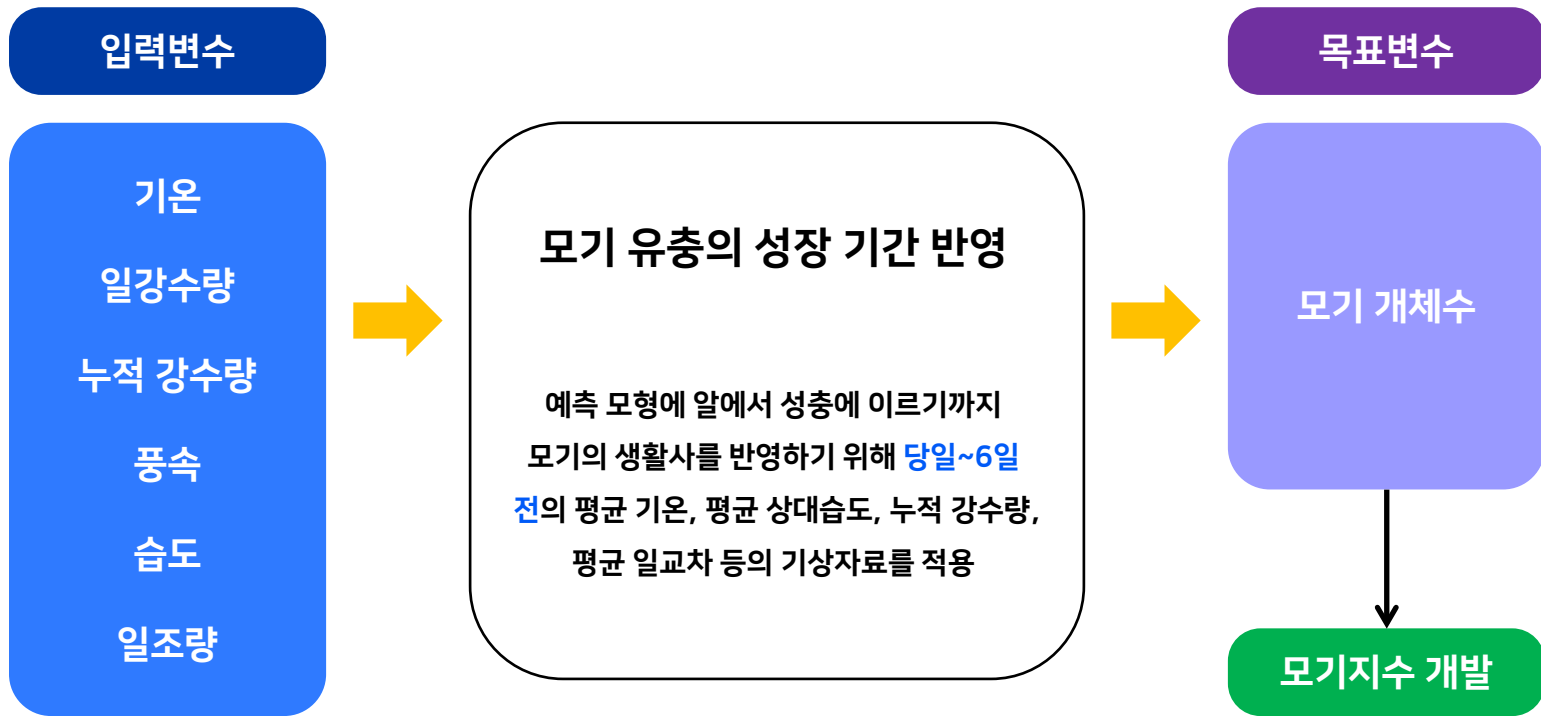
• 지리 타입은 3개(공원, 수변부, 주거지)로 분류

• DMS 설치 현황 시각화 자료를 통해 구별, 지리 타입별로 골고루 설치되어 있어 서울 전역의 모기 발생량을 분석하기에 적합하다고 판단

# 데이터 전처리

## 연구 모형

## 2. 프로세싱



# 서울특별시 기후 데이터

## 2. 프로세싱

### 분석 데이터 구축

	date	temp	rain_per_day	wind	humidity	sunshine		date	temp	rain_per_day	wind	humidity	sunshine
0	2015-01-01	-7.7	NaN	4.6	41.4	9.79	0	2015-01-07	-2.700000	0.485714	2.985714	53.157143	8.558571
1	2015-01-02	-6.0	NaN	3.2	45.9	9.07	1	2015-01-08	-2.328571	0.485714	2.642857	53.785714	8.601429
2	2015-01-03	-2.7	NaN	1.9	56.1	8.66	2	2015-01-09	-1.771429				
3	2015-01-04	2.5	NaN	2.0	70.1	5.32	3	2015-01-10	-1.428571				
4	2015-01-05	3.7	0.4	2.4	73.1	6.48	4	2015-01-11	-1.842857				
5	2015-01-06	-3.2	3.0	4.3	52.4	10.47	5	2015-01-12	-2.757143				
6	2015-01-07	-5.5	NaN	2.5	33.1	10.12	6	2015-01-13	-2.271429				
7	2015-01-08	-5.1	NaN	2.2	45.8	10.09	7	2015-01-14	-1.128571	0.000000	2.171429	48.128571	8.484286
8	2015-01-09	-2.1	NaN	2.0	58.3	8.74	8	2015-01-15	0.000000	0.000000	2.157143	49.242857	8.252857
9	2015-01-10	-0.3	NaN	2.3	56.5	9.41	9	2015-01-16	0.414286	0.042857	2.442857	51.342857	7.402857
10	2015-01-11	-0.4	NaN	3.6	57.8	9.60	10	2015-01-17	-0.128571	0.042857	2.657143	49.842857	7.614286
11	2015-01-12	-2.7	NaN	1.8	37.8	10.05	11	2015-01-18	-0.257143	0.614286	2.471429	51.357143	6.965714
12	2015-01-13	0.2	NaN	1.8	39.6	8.14	12	2015-01-19	0.114286	0.628571	2.671429	55.685714	6.604286
13	2015-01-14	2.5	NaN	1.5	41.1	3.36	13	2015-01-20	-0.042857	0.628571	2.685714	58.000000	6.840000
14	2015-01-15	2.8	NaN	2.1	53.6	8.47	14	2015-01-21	-0.028571	0.628571	2.757143	60.457143	7.170000

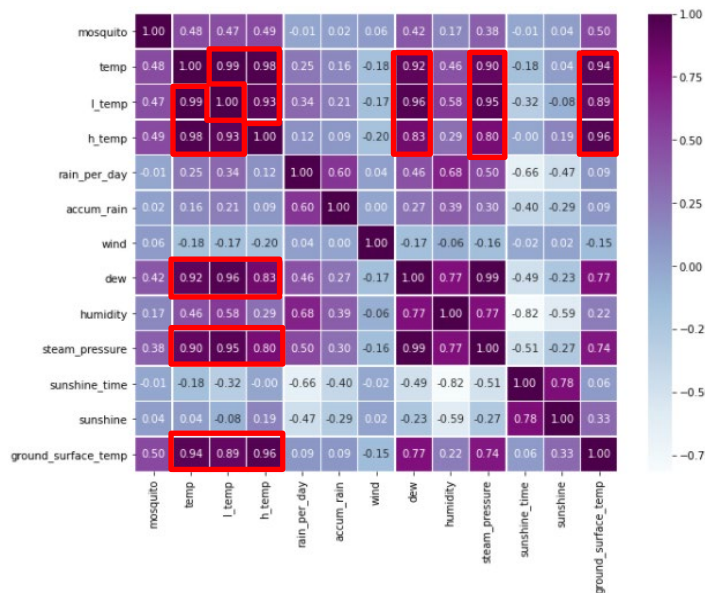
각 변수는 모기 유충에서  
성충으로의 성장 기간을 고려하여  
당일을 포함한 일주일 데이터의  
평균값 도출

# 서울특별시 기후 데이터

## 2. 프로세싱

### 다중공선성

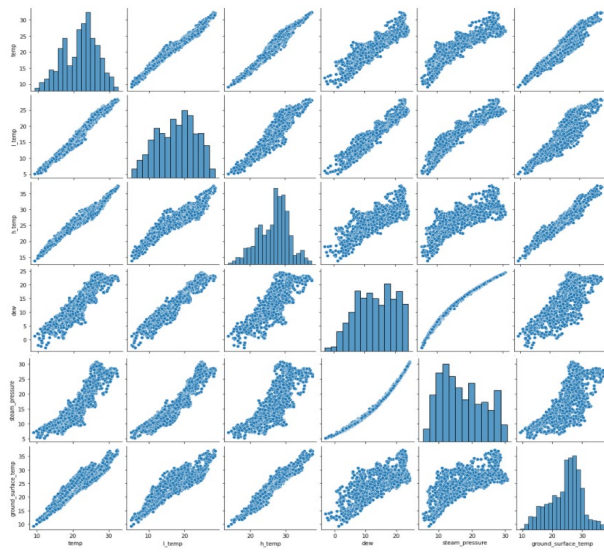
Heatmap - Pearson



Heatmap으로 각 변수 사이의 상관관계를 확인

온도 관련 변수와 일조량 관련 변수들의 다중공선성이 우려

Pair Plot



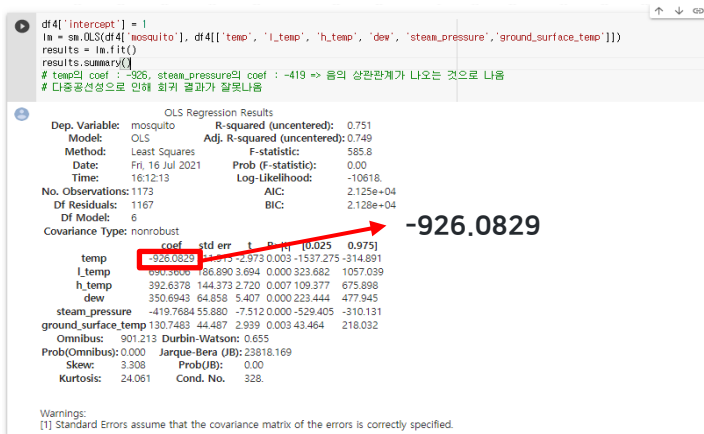
Pair Plot으로 높은 양의 상관관계, 다중공선성 확인

# 서울특별시 기후 데이터

## 2. 프로세싱

### 다중공선성

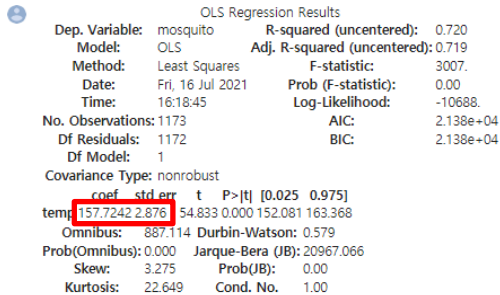
#### VIF



#### drop을 통한 다중공선성 해결

```
# VIF 10 이상 변수 drop을 통한 다중공선성 해결
df4['intercept'] = 1
lm = sm.OLS(df4[['mosquito']], df4[['temp']])
results = lm.fit()
results.summary()

# 다중공선성 문제를 해결하니 올바른 결과가 도출
```



```
# VIF도 재계산
y, X = dmatrices('mosquito ~ temp', df4, return_type='dataframe')
vif = pd.DataFrame()
vif['VIF Factor'] = [variance_inflation_factor(X.values, i) for i in range(X.shape[1])]
vif['features'] = X.columns
vif
```

VIF Factor	features
0	Intercept
1	temp

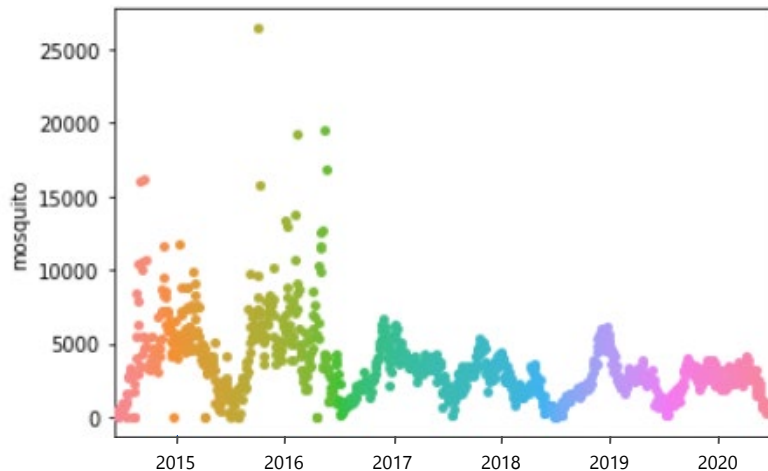
temp(평균기온(°C))만 분석에 사용하기로 결정

# 서울특별시 DMS 포집내역

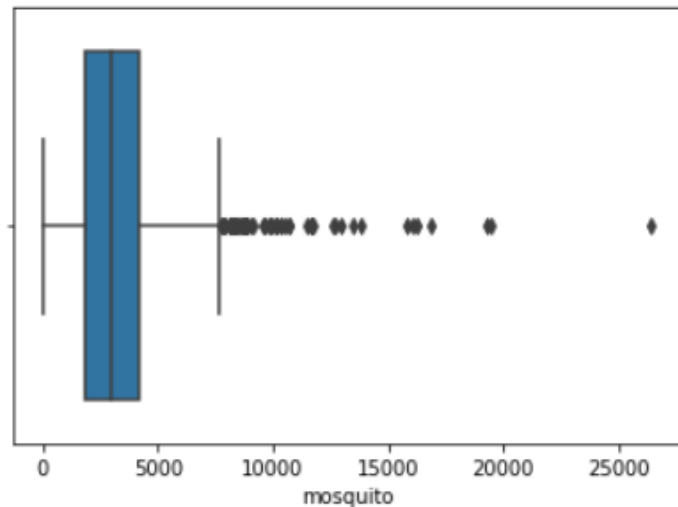
## 2. 프로세싱

### 이상치 확인

Strip Plot



Box Plot



Strip Plot을 통해 2015년 ~ 2016년에 몇몇 이상치가 발생하는 것을 확인

2017년 이후에는 데이터를 공개하기 전에 이상치를 제거했을 것으로 추정

정확한 분석을 위해서 이상치를 제거하지 않기로 결정

### 이상치 & 결측치 처리와 데이터 결합

데이터	변수	이상치 & 결측치 처리
서울특별시 DMS 포집내역	mosquito (모기 포집량)	기타 벌레 포집량은 분석에서 제외
		2017년 이후로는 보건환경연구원에서 이상치를 제거한 것으로 추정 → 이상치를 제거하지 않고 분석 진행
서울특별시 기후	rain_per_day (일강수량(mm))	0.1 이하 → 0으로 변환
		결측치는 비가 전혀 오지 않았음을 의미 → 0으로 변환
	나머지 기후 변수 (rain_per_day 포함)	모기 유충의 성장 기간 고려 → 당일을 포함한 이전 7일 평균값 계산 결측치는 앞뒤 3일의 값으로 변환

<mosquito>			<rain_per_day>		<나머지 기후 변수>			
	date	mosquito		rain_per_day		temp	l_temp	h_temp
0	2015-04-06	199		5.285714		12.214286	8.757143	16.500000
1	2015-04-07	146		4.928571		11.571429	8.114286	16.114286
2	2015-04-08	90	+	4.571429	+	10.914286	7.185714	15.557143
3	2015-04-09	172		0.571429		10.171429	6.514286	14.771429
4	2015-04-10	249		0.071429		10.314286	6.185714	15.357143

분석 데이터 구축  
&  
Standardization

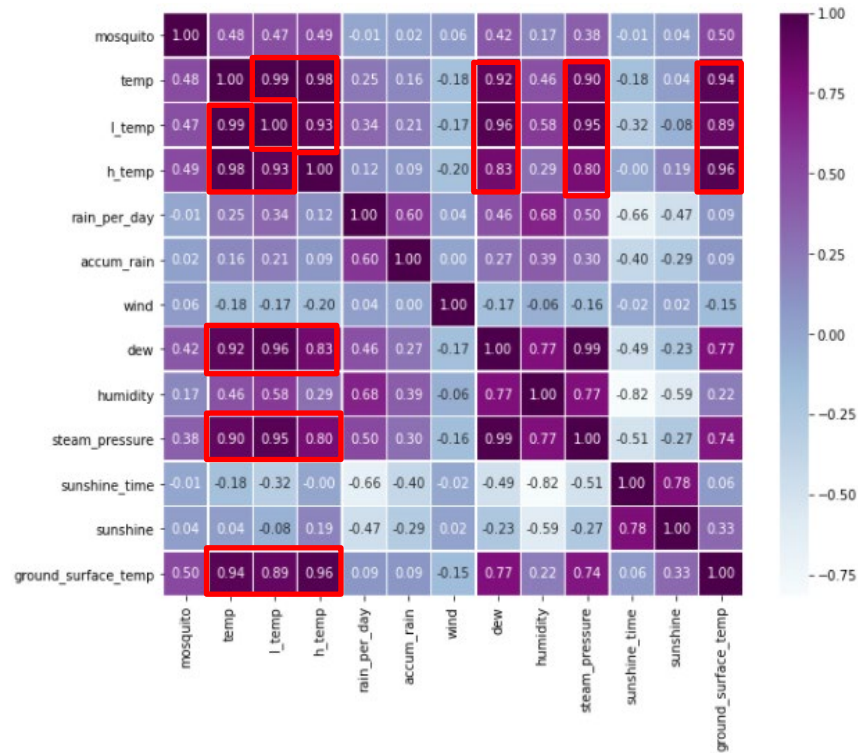


# 데이터 프로세싱

## 2. 프로세싱

### 전처리 전후 비교

전처리 전 Heatmap



전처리 후 Heatmap



### 모델 선정

데이터 형태  
파악

Grid Search  
CV

RMSE 비교

모든 데이터가 연속적인 숫자로 구성

∴ 트리 기반 Regressor 모델 사용

Grid Search CV로  
LGBM, GBR, RF 모델의  
최적의 하이퍼 파라미터 조합 확인



모델	RMSE
LGBM(Light Gradient Boosting Machine)	0.634
GBR(Gradient Boosting Regressor)	0.692
RF(Random Forest)	0.690



LGBM 채택!

※ 결정트리(Decision Tree)는 분류와 회귀 작업이 모두 가능한 머신러닝 알고리즘

### 모델 선정

<원본 데이터를 사용한 경우> <데이터를 7일 앞당긴 경우>

모델	RMSE
LGBM	0.688
GBR	0.688
RF	0.687

모델	RMSE
LGBM	0.717
GBR	0.709
RF	0.696

모기 유충의 성장 기간을 반영할 때  
단순히 기후 데이터를 7일씩 앞당긴 경우,

RMSE 값이 오히려 증가  
→ 데이터 구축 방법 변경 필요

<데이터의 7일 평균값을 적용한 경우>

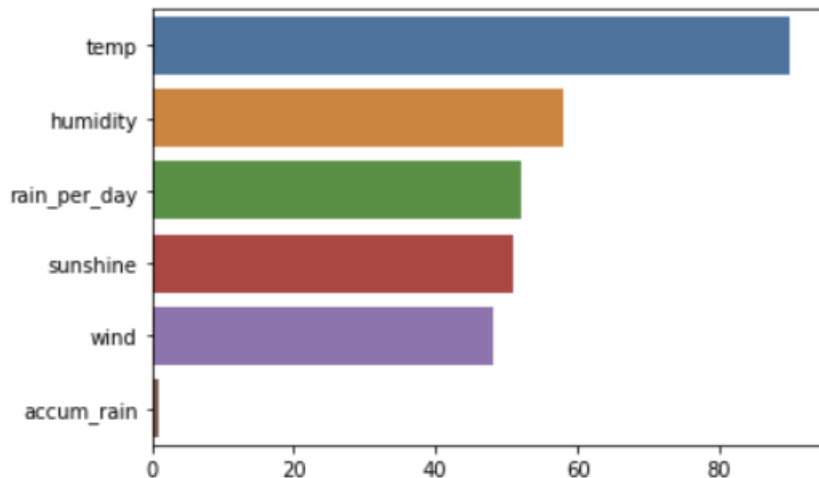
모델	RMSE
LGBM	0.634
GBR	0.692
RF	0.690

누적된 기후 데이터의 영향을 적용하기 위해  
기후 데이터의 7일 평균값을 적용한 경우,

RMSE 값이 감소  
→ 보다 정확한 예측력

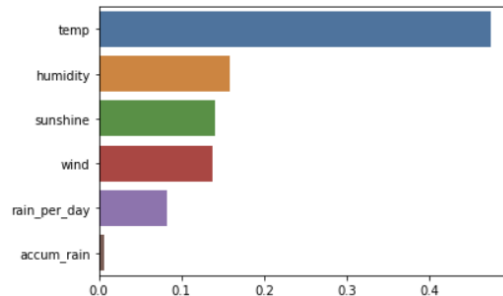
### 분석 결과

Feature Importance(LGBM)

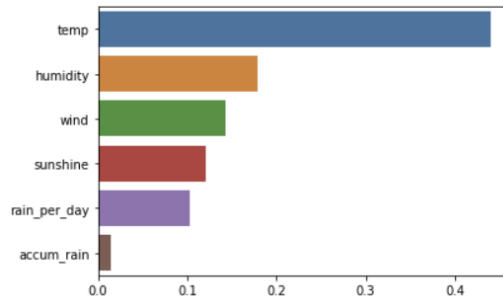


- 다른 변수들에 비해 temp의 중요성이 크게 나타남
- 새롭게 도출한 accum\_rain의 영향력은 미미한 수준

Feature Importance(GBR)



Feature Importance(RFR)



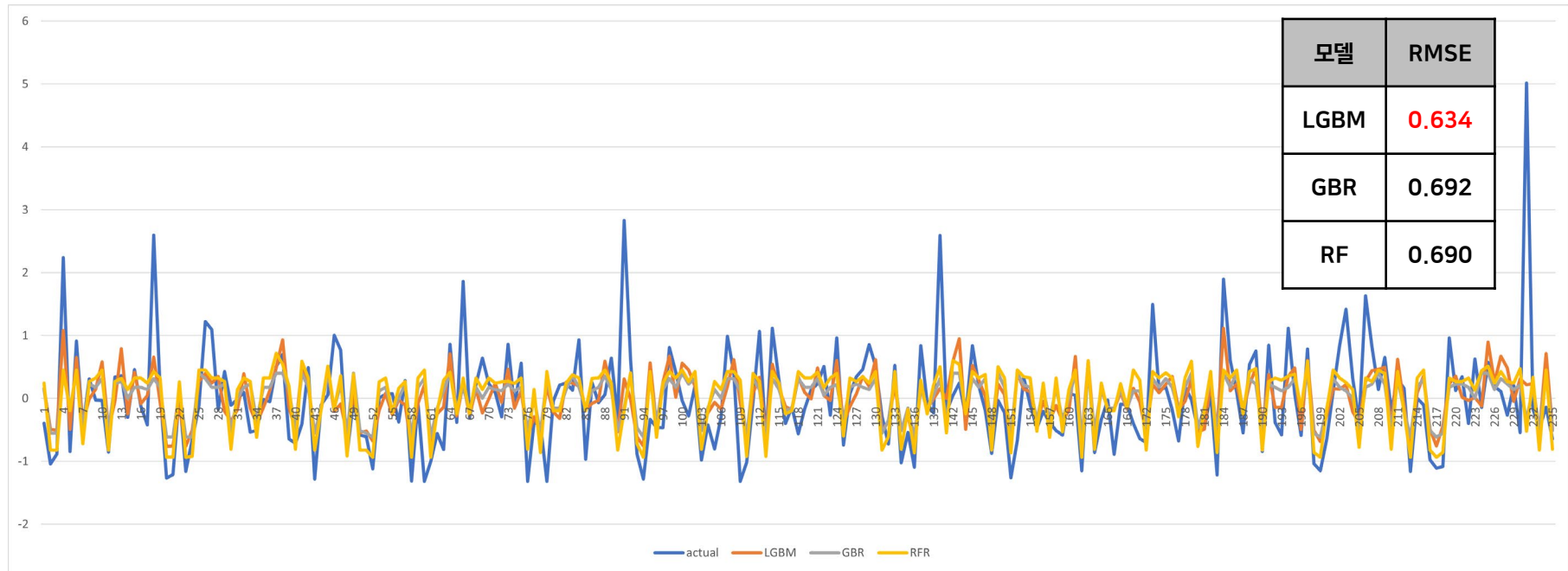
다른 모델에서도 결과가 동일함을 확인

# 데이터 분석

## 2. 프로세싱

### 분석 결과

#### <Test Set 예측 결과>



### 분석 결과

<Test Set 예측 결과> 실제값과 LGBM 결과만 표시, 시간순 정렬



실제값과 비슷한 추이를 보이며, 계절(평균 기온)에 따라 일정한 주기성을 가짐

모델	RMSE
LGBM	0.634

### 지수 개발

#### <선행연구의 구별 모기 활동지수 적용 기준>

표 22. 분류 기준에 따른 단계별 모기 활동지수

단계	기준	1일 채집개체 수	단계	기준	1일 채집개체 수
1	미출현	0	6	15분당 1마리	60~89
2	900분당 1마리	1~9	7	10분당 1마리	90~149
3	90분당 1마리	10~19	8	6분당 1마리	150~224
4	45분당 1마리	20~29	9	4분당 1마리	225~449
5	30분당 1마리	30~59	10	2분당 1마리	450 이상

출처=정해관 외, 2019, 「기상자료와 GIS 활용 수도권 모기 활동지수 개발」, 수도권기상청 기후서비스과

#### <서울 전체의 모기 활동지수 적용 기준>

단계	1일 채집개체 수
쾌적	0 ~ 250
관심	251 ~ 750
주의	751 ~ 2250
불쾌	2251 ~ 5625
심각	5626 ~

기존 10단계를 두 단계씩 묶어 5단계로 만들고,

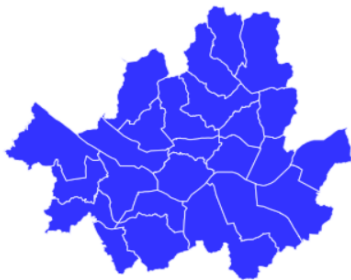
서울 전체 25개 구의 수치를 계산하기 위해 단계별로 ×25

# 모기지수 시각화

실제 적용 예시

Folium

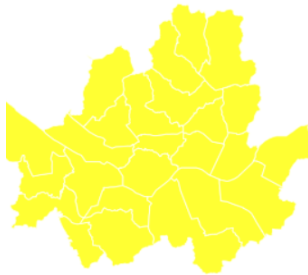
쾌적



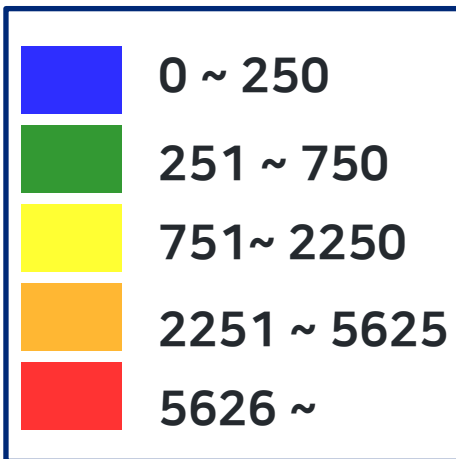
관심



주의



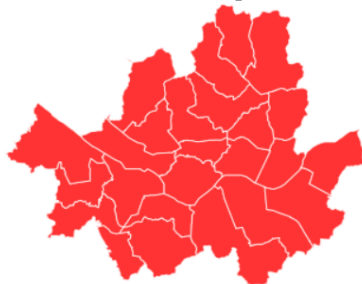
모기 개체수에 따른 모기지수



불쾌



심각







### **3. 기대효과 및 한계점**

질병 및  
전염병 관리

기온 상승으로 예측되는 모기 증가와 관련  
매개질환 감염병 위험에 선제적 대응

모기예보제의 정확성 제고와  
전국적으로 기술 확대 및 적용

기술 확대  
적용 및 개발

방역 사업의  
효율화 도모

모기 활동성이 높은 지역에 방역을  
집중함으로써 환경오염 위험 감소

방역 약품 감소에 따른 예산 절감 효과

예산 절감  
효과

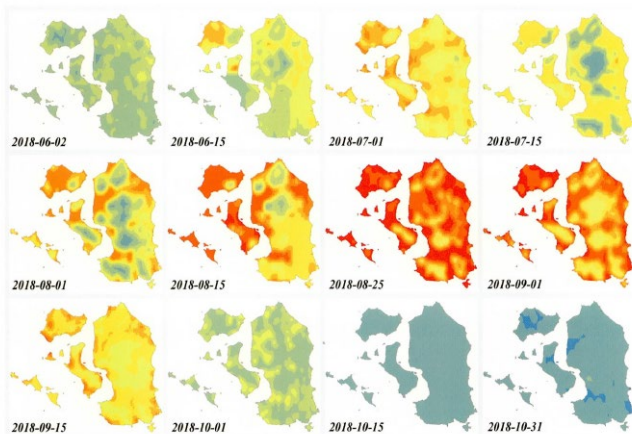
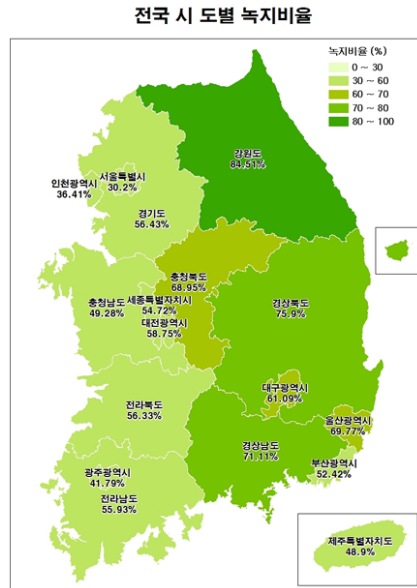


그림 68. 2018년 인천광역시 강화군 모기 활동지수 예측 결과

인천광역시 비도심 지역인 강화군은 도심보다 **해안**에 가까운 교외지역에서 모기 활동지수가 높았다. 그러나 모기 발생이 가장 많은 8월에는 교외 지역뿐만 아니라 도심지역에서도 모기 활동지수가 매우 높은 양상을 보였고, 기온이 감소하는 10월부터 전 지역의 모기 활동지수가 감소하는 경향을 보였다.

출처=정해관 외, 2019, 「기상자료와 GIS 활용 수도권 모기 활동지수 개발」, 수도권기상청 기후서비스과, p,85






출처=환경부

- 서울특별시는 서울 전체 합산 모기 데이터만 공개하기 때문에 구별 지리적인 특성을 반영하지 못함
- 인천광역시의 경우 **해안가에 모기가 집중적으로 발생** → 본 분석에서 **지리 데이터를 추가해야 정확한 분석 가능**
- 구별 DMS 데이터가 공개된다면 지리 데이터도 적용할 수 있어 좀 더 정확한 예측이 가능할 것으로 기대



## 4. 느낀점

김형민		<p>욕심 많고 사서 고생하는 조장을 아무런 불만 없이 열심히 잘 따라와준 팀원분들께 정말 감사드립니다. 수업 시간에 배운 내용을 실제로 활용해볼 수 있어서 뿌듯했고, 서로 서로 부족한 부분을 메워주면서 정말정말정말 많은 것을 배웠습니다. 무엇보다도 협업의 중요성에 대해서 깨달을 수 있었던 경험이었고, 그저 팀원들과 웃고 즐기면서 진행한 약 2주 간의 프로젝트 기간이 행복했습니다.</p>
고아름		<p>데이터 시각화 프로젝트를 통해 데이터 수집부터 분석까지 전반적인 과정을 직접 수행해볼 수 있었습니다. 팀원들과 다양한 의견을 나누면서 새로운 변수와 관련하여 심도 깊게 논의할 수 있었고 미처 생각하지 못한 여러 부분을 고려할 수 있었습니다. 데이터 전처리부터 모델 선정, 분석까지 너무나 막연했지만 팀원들 덕분에 많은 것을 배우고 해낼 수 있었습니다.</p>
김예이		<p>이번 프로젝트를 통해서 통계, 분석 등 어려웠던 부분에 대한 이해도 증가와 다양한 툴을 익히고 사용할 수 있어서 좋았습니다. 또한 즉각적인 피드백과 함께 문제를 해결했을 때마다 서로 격려하며 결과물을 만들어 가는 과정에서 팀원들에게서도 많은 것을 배울 수 있었습니다. 부족했던 저를 많이 도와준 형림님과 아름님과 함께여서 더 특별하고 값진 경험이었던 것 같습니다. 너무나 뜻 깊은 시간이었습니다.</p>

# 참고문헌

- 김태규 외, 2021, 「주간 건강과 질병 제14권 제14호 - 2020년 국내 일본뇌염 매개모기 발생 감시 현황」, 질병관리청, p.802, p.1016~1017
- 이동우, 2019, 「모기 활동성 예측을 위한 기상 데이터에서 인접성 향상을 위한 연구」, 국민대학교, p.1~8
- 임영미 외, 2019, 「동네예보자료를 활용한 수도권 모기예측 지수 개발」, 한국기상학회, p.1
- 황세영 외, 2020, 「머신러닝 기반 기후 데이터를 활용한 모기 개체 수 예측」, 순천대학교, p.1~3
- 정해관 외, 2019, 「기상자료와 GIS 활용 수도권 모기 활동지수 개발」, 수도권기상청 기후서비스과
- 2017, 「주요환경변화에 따른 미래 감염병의 발생 양상」, 질병관리본부 미래감염병대비과, p.1024~1028
- 2020, 「한국 기후 변화 평가보고서」, 환경부
- 채수미, 2014, 「기온과 지역 특성이 말라리아 발생에 미치는 영향」

# Thanks

Do you have any question?



**STOP**  
**Mosquito!!**