<독립변수 수집 및 정제 내용 정리>

1. **기후(온도, 습도, 강수량)**

**기후 데이터 수집: 기상청(지우: 데이터 수집했던 주소 첨부하기)**

관련 논문: 기후 변화 데이터를 활용한 조류인플루엔자 발생 예측 방법

* 1. **독립변수 채택 과정 (**파일 출처: 깃-jiwoo/Data\_preprocess(weather).ipynb)
     1. 온도 가설 검정 과정
* 가설 설정

###### *(평균 온도 검증)*

1. 가정: 계절별 HPAI가 발생하지 않은 날짜의 평균 온도는 HPAI 발생한 날짜의 평균 온도와 차이가 있을 것이다.

2. 귀무가설: 두 변수의 차이는 0의 근접할 것이다.

3. 대립가설: 두 변수의 차이가 0에 근접하지 않을 것이다.

* 가설 검정

-정규성 검정->독립성 검정->등분산성 검정->가설검정(Mann-Whitney U Test)을 통해 유의 수준 5% 내 경우, 귀무가설 기각 ·대립가설 채택

텍스트이(가) 표시된 사진

자동 생성된 설명

<파일 출처: 깃-jiwoo/Data\_preprocess(weather).ipynb>

가설 설정 후 Mann-Whitney U test 이용하여 유의수준 5% 이내 확인. 온도를 독립변수로 채택.

* + 1. 습도 가설 검정 과정
* 가설 검정

###### *(평균 상대습도 검증)*

1. 가정: 겨울을 기준으로 HPAI가 발생하지 않은 날짜의 평균 상대습도은 HPAI 발생한 날짜의 평균 상대습도보다 낮을 것이다.

2. 귀무가설: 두 변수의 차이는 0의 근접할 것이다.

3. 대립가설: 두 변수의 차이가 0에 근접하지 않을 것이다.

* 가설 검정 과정

-정규성 검정->독립성 검정->등분산성 검정->가설검정(Mann-Whitney U Test)을 통해 유의 수준 5% 내 경우, 귀무가설 기각 ·대립가설 채택

텍스트이(가) 표시된 사진

자동 생성된 설명

<파일 출처: 깃-jiwoo/Data\_preprocess(weather).ipynb>

가설 설정 후 Mann-Whitney U test 이용하여 유의수준 5% 이내 확인. 습도를 독립변수로 채택.

* + 1. 강수량 가설 검정 과정
* 가설 설정

###### *(일강수량 검증)*

1. 가정: 여름을 기준으로 HPAI가 발생하지 않은 날짜의 일강수량은 HPAI 발생한 날짜의 일강수량과 차이가 있다.

2. 귀무가설: 두 변수의 차이는 0의 근접할 것이다.

3. 대립가설: 두 변수의 차이가 0에 근접하지 않을 것이다.

* 가설 검정 과정

-정규성 검정->독립성 검정->등분산성 검정->가설검정(Mann-Whitney U Test)을 통해 유의 수준 5% 내 경우, 귀무가설 기각 ·대립가설 채택

텍스트이(가) 표시된 사진

자동 생성된 설명<파일 출처: 깃-jiwoo/Data\_preprocess(weather).ipynb>

가설 설정 후 Mann-Whitney U test 이용하여 유의수준 5% 이내 확인. 강수량을 독립변수로 채택.

*++추가) 평균풍속 유의 수준 5% 내였으나, 독립변수로 채택하지 않음->이유:* ***(지우가 작성****)*

###### *(평균 풍속 검증)*

1. 가정: 겨울을 기준으로 HPAI가 발생하지 않은 날짜의 평균 풍속은 HPAI 발생한 날짜의 평균 풍속보다 높을 것이다.

2. 귀무가설: 두 변수의 차이는 0의 근접할 것이다.

3. 대립가설: 두 변수의 차이가 0에 근접하지 않을 것이다.

텍스트이(가) 표시된 사진

자동 생성된 설명

* 1. **기후 RISK TABLE SET 제작 (지우오빠가 기재 요망)/추후 파일 추가**

사분위 수 범위(IQR)

위험 점수-범위에 포함되면

-

1. **거리**

**데이터 수집: 우리나라 시군구(공공데이터포털), 시군구 좌표(카카오 api)**

**2-1 독립변수 채택 과정**

-거리는 사전 자료검토 과정에서 중요한 요소임이 확인되어, 가설검정 진행하지 않고 독립변수 채택.

**2-2 거리 RISK TABLE SET 제작 (지우오빠가 기재 요망)/추후 파일 추가**

-각 월 별로 가장 많은 건수가 발생한 시군구 기준으로 10km 당 발생건수 확인

-

1. **철새수**

**데이터 수집: 환경부 국립생물자원관(철새수), 발생건수(**고병원성조류인플루엔자 발생목록(2000~2021) 출처 기재 요망)

**3-1. 독립변수 채택 과정**

3-1-1) 철새와 HPAI 상관관계 확인

텍스트이(가) 표시된 사진

자동 생성된 설명

->철새가 나타나는 시기인 1~3월 AND 10-12월(철새 유) 발생건수와 철새가 나타나지 않는 4월~9월(철새무) 발생건수 비교.

-> 철새가 나타나는 시기의 발생건수가 철새 나타나지 않는 시기의 발생건수 보다 2배 이상 더 많음을 확인.

3-1-2) 철새 수 가설검정 과정

* 가설 설정

가정: 철새가 나타나는 월과 철새가 나타나지 않는 월보다 발생건수 차이가 있을 것이다.

귀무가설: 철새가 나타나는 월의 발생건수와 철새가 나타나지 않는 월의 발생건수 차이는 0이다.

대립가설: 철새가 나타나는 월의 발생건수와 철새가 나타나지 않는 월의 발생건수 차이는 0이 아니다.

* 가설 검정 과정

- 정규성 검정->독립성 검정->등분산성 검정->가설검정(Mann-Whitney U Test)을 통해 유의 수준 5% 내 경우, 귀무가설 기각 ·대립가설 채택

텍스트이(가) 표시된 사진

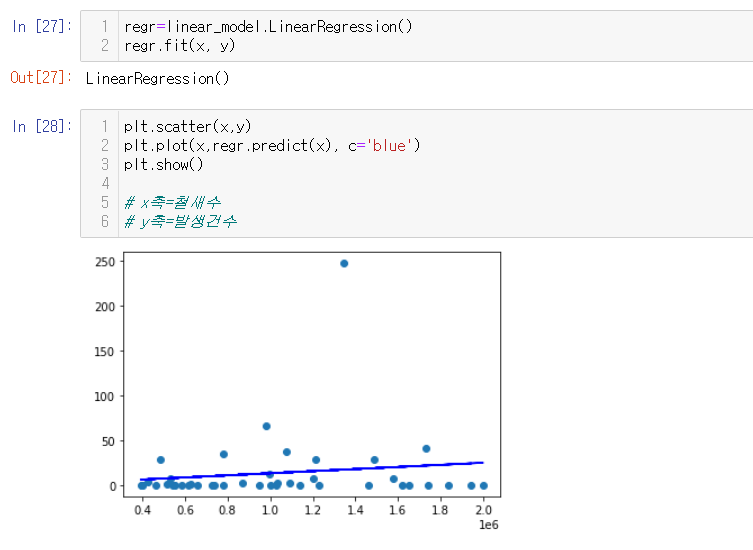
자동 생성된 설명

<출처-추후 깃에 업로드한 후 기재 하겠음(현지)>

가설 설정 후 Mann-Whitney U test 이용하여 유의수준 5% 이내 확인. 철새 수 독립변수로 채택.

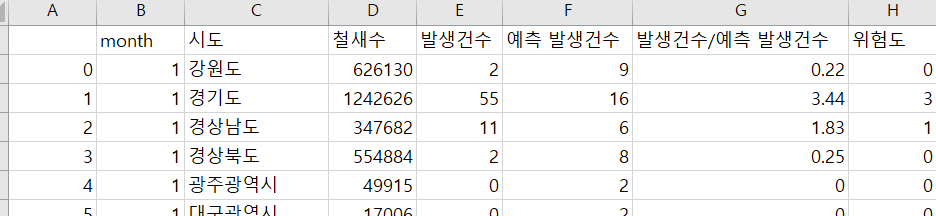
**3-2 철새 수 RISK TABLE SET 제작**

3-2-2 선형관계 확인. (X축: 철새수,Y축: 발생건수)



3-2-3 위험등급 기준 정하기

1. 선형회귀식의 예측발생건수에서 실제 발생건수 비율 확인(발생건수/예측 발생건수)



1. 발생건수/예측 발생건수 비율로 등급 나누기

테이블이(가) 표시된 사진

자동 생성된 설명

1. 위험도 등급 부여하기

테이블이(가) 표시된 사진

자동 생성된 설명

1. **데이터 정제**

**관련 내용 파일(깃-지우- Train\_data\_set.ipynb)**

1. 다중공선성[[1]](#footnote-1) 확인: 시도, 시군구, 상대습도가 다른 변수 대비 높아 데이터 분석 시 신뢰도가 떨어질 것이라 판단.

텍스트이(가) 표시된 사진

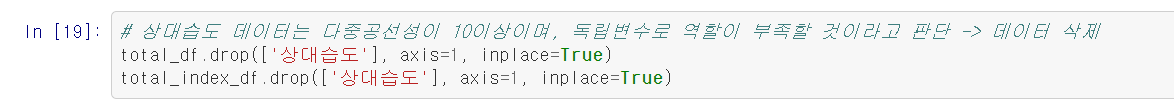
자동 생성된 설명.

1. 다중공선성 제거
2. 시도, 시군구 번호 합치기

텍스트이(가) 표시된 사진

자동 생성된 설명

1. 상대습도 제거



1. 다중공선성 재확인 (상대습도 삭제/시군구 index 다중공선성 지수 하락)

테이블이(가) 표시된 사진

자동 생성된 설명

1. 회귀 분석에서 사용된 모형의 일부 예측 변수가 다른 예측 변수와 상관 정도가 높아, 데이터 분석 시 부정적인 영향을 미치는 현상 [↑](#footnote-ref-1)