소방청 National Fire Agency 기상청

과제 3 최종 공모안

기상현상과 화재 발생에 대한 상관분석

접수 번호	240205	팀 명 / 팀 원	나이승~ 염승원, 이인서, 장한나

1. 분석 배경 및 주제

화재는 여러 요인이 복합적으로 작용하여 발생하는 재해 중 하나이다. 그 중에서도 기상 현상은 중요한 요소 중 하나로 여겨진다. 기상청과 소방청에서도 이를 고려하여 함께 협업해 화재 발생 확률이 높은 기상특보 시에 화재 위험 경보 시스템을 운영하는 시도가 있었다. 하지만 발령 기준이 명확하지 않아 실효성 있는 시행이 어려웠다. 따라서 본 공모안에서는 기상 관측 자료만으로 화재 발생을 예측하고 이를 통해 보다 효과적인 예방 시스템을 구축할 수 있는 방안을 모색하고자 한다.

이를 위해 특보와 화재 발생의 관계성을 파악하는 것을 넘어, 화재 발생 시의 기상 상태를 분석하여 보다 일반화된 기상 조건과 화재 발생 간의 관계를 연구한다. 또한 기상 데이터와 화재 데이터를 결합하여 다양한 통계 분석 및 기계 학습 기법을 활용함으로써, 화재 발생의 패턴과 주요 기상 요인을 식별하고자 한다.

2. 데이터 정의

본 공모안에서는 날씨마루에서 제공해준 소방 데이터, 기상 데이터, 기상특보 데이터를 활용하였고, 2018년부터 2023년까지 6년간 데이터를 다루고 있다. 소방 데이터는 연월일, 시간(시,분) 단위의 화재발생 정보를 담고 있으며, 기상 데이터는 날짜, 지역, 기상 정보(기온, 강수량, 풍속, 상대습도)를 담고 있고, 기상특보 데이터는 날짜, 지역, 기상 특보 정보 (특보 종류, 특보 단계, 특보 명령)을 담고 있다.

기상 데이터의 기상 관련 변수들은 모두 결측치를 가지고 있고, 이는 -99로 표시되어있다. 또한 기상 데이터와 기상 특보 데이터 모두 중복되는 데이터(행들의 변수의 값들은 서로 달라도 결국 같은 의미를 가지는 데이터)들을 가지고 있다. 이러한 값들에 대한 자세한 처리는 3절인 데이터 전처리에서 더욱 자세히 다룰 예정이고, 제출한 코드파일에서 확인할 수 있다.

3. 데이터 전처리

소방 데이터, 기상 데이터, 기상특보 데이터를 각각 df1, df2, df3로 정의하여 데이터 전처리를 시행하였다.

3.1 접두사 제거

코드 작성의 용이성을 위해 df1의 fire_firefighting, df2의 fire_weather, df3의 fire_weather_special라는 접두사가 반복되므로 gsub() 함수로 제거하였다.

3.2 불필요 변수 제거

3.2.1 fire_type_2

df1의 fire_type_2의 빈 문자열은 fire_type_1의 임야와 기타에 해당했다. 또한 fire_type_1에서 임야와 기타를 제외한 요소는 fire_type_2와 일대일 대응됨을 확인하였다. 따라서 실질적으로 fire_type_2가 갖는 의미는 없다고 판단하여 제거하였다.

3.2.2 stn

행정 구역(district_1, district_2) 기준으로 상관관계를 판단하고자 했기에, df2에서 AWS 지점 코드인 stn 열을 제거하였다.

3.3 data type 변경

3.3.1 tm 열을 POSIXct type으로 변경

각 데이터프레임에는 시간 정보를 담고있는 character type 열이 존재하기에, 시간 정보의 특성을 활용할 수 있도록 POSIXct type으로 변경하였다. 따라서 df1, df2의 tm 열과, df3의 tm_fc, tm_ef열을 'YYYY-MM-DD HH:MM:SS' 형태로 저장하였다.

3.3.2 그 외 character 변수를 factor type으로 변경

특정 열을 기준으로 다른 요소를 고려할 때 character 변수보다는 factor 변수가 적합하기에, factor 변수로 변경하였다.

	변경한 열
df1	year, district_1, district_2, fire_type_1, ignition_factor_category_1, ignition_factor_category_2, location_category_1, location_category_2, location_category_3
df2	district_1, district_2
df3	district_1, district_2, stn, reg_id, wrn, lvl, cmd

3.4 중복 데이터 제거

3.4.1 df2

df2에서 같은 날짜(tm)와 같은 지역(district_1, district_2)에 여러 개의 기상 데이터가 존재하는 것을 확인하였다. 이러한 경우 중복되는 행이 해당 행의 결측치(-99)를 일부 보완해준다는 것을 알 수 있었다.

따라서 중복되는 한 쌍의 행에서 변수별로 두 값을 비교해 하나의 행으로 병합하는 과정을 거쳤다. 결측치가 없는 변수일 경우 평균값으로, 결측치가 하나인 경우 존재하는 관측값으로 처리해주었고, 결측치가 둘 다 존재하는 경우 그대로 결측치로 남겨두었다.

3.4.2 df3

발표된 특보가 중요하다고 생각하여 cmd가 1인 데이터를 제외하고 모두 삭제하였다. 다만 발표된 특보의 해제 발표가 해당 특보의 발표시간과 발효시간 사이에 일어난다면, 그 특보는 발효되지 못한 것이므로, 그러한 경우도 모두 제거해주었다.

이후 tm_fc, tm_ef, district_1, wrn 변수만 남긴 후 중복 데이터를 모두 제거하여 districts_1별로 발표된 특보를 중복없이 파악할 수 있도록 하였다.

3.5 결측치 처리

df2에 존재하는 결측치를 용이하게 제거하기 위해 우선 -99인 값을 모두 NaN으로 변경해주었다. 시계열 특성을 가진 기상데이터이므로 지역별 선형보간법[1]으로 결측치들을 처리해주었다. 하지만 전북특별자치도, 세종특별자치시, 부산광역시 지역은 다른 지역에 비해 많은 결측치를 포함하고 있어 선형보간으로 해결할 수 없었고, 따라서 이 지역들은 기상이 가장 유사한 지역을 이용해 결측치 처리를 해주었다.

해당 지역들에 대해 기상이 유사한 지역은 유클리드 거리, 맨하탄 거리를 기준으로 유사도를 계산하였을 때 두 거리 측도에 의해 모두 가까운 지역으로 선정하였다. 그 결과 전북은 경상북도, 충청남도, 대전광역시와 기상 환경이 유사하였고, 세종은 충청북도, 충청남도와 유사하며, 부산은 울산광역시와 가장 유사하였다.

$$\text{유클리드 거리 : } d(x,y) = \sqrt{(x_1 - y_1)^2 + \ \dots \ + (x_p - y_p)^2} = \sqrt{(x-y)^T (x-y)}$$

$$\text{맨하탄 거리 : } d(x,y) = \sum_{i=1}^{p} |x_i - y_i|$$

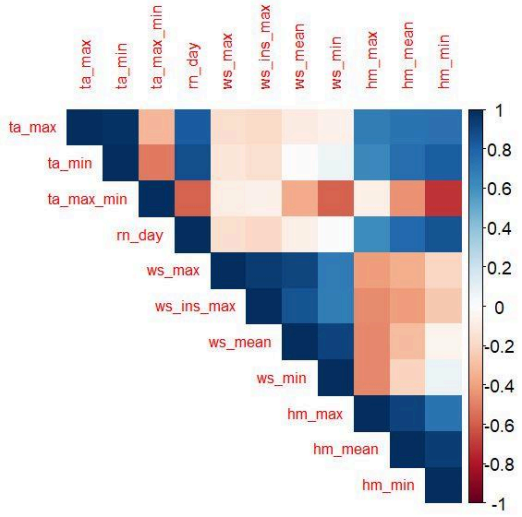
유사 지역의 기상 데이터를 이용해 스플라인 보간법[2]을 진행하였고, 세 개의 지역을 제외한 나머지 지역들은 각 지역별로 선형보간을 통해 결측치를 대체해주었다. 마지막으로 전북, 세종, 부산, 그리고 나머지 지역의 결측치 처리한 데이터프레임을 모두 병합하여 df2에 대입해주었다.

[1] 선형 보간법 (Linear Interpolation) : 끝점의 값이 주어졌을 때 그 사이에 위치한 값을 추정하기 위하여 직선 거리에 따라 선형적으로 계산하는 방법

[2] 스플라인 보간법 (Spline Interpolation) : 전체 구간을 소구간별로 나누어 저차수의 다항식으로 매끄러운 함수를 구하는 방법

4. 상관 분석

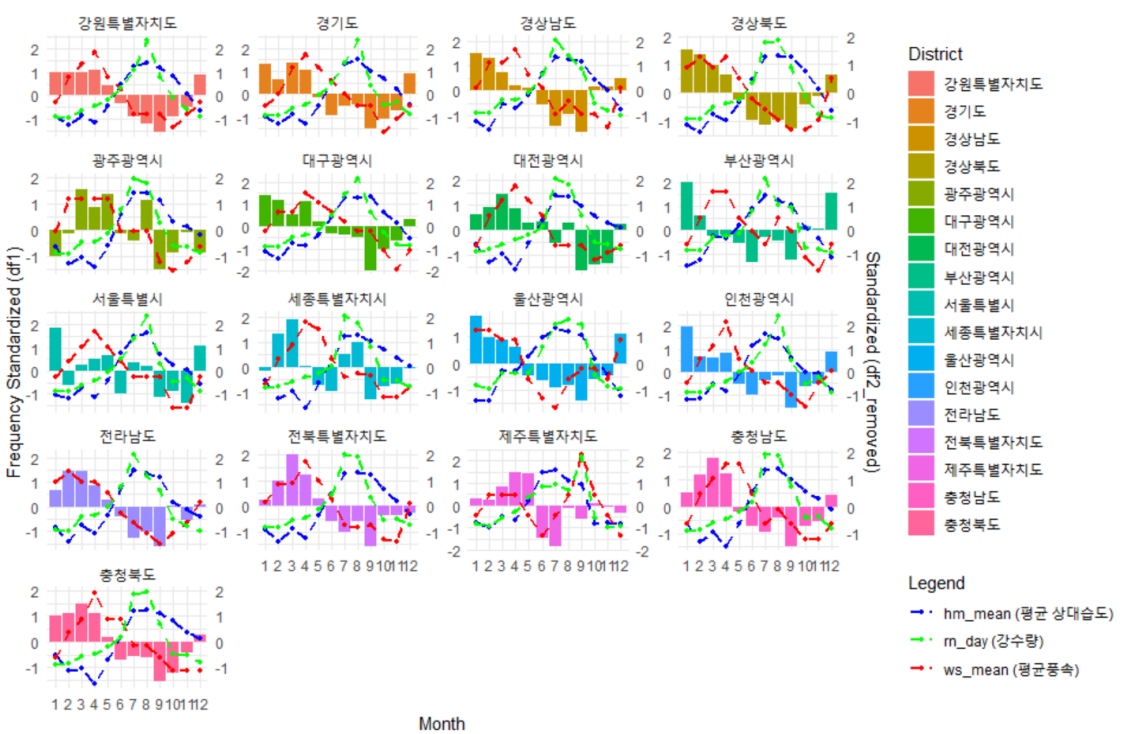
4.1 데이터 축소



기상 현상과 화재 발생과의 관계성을 찾기 위해 상관분석을 진행하였다. 이 때 기상 변수는 rn_day, ws_mean, hm_mean만 사용하여 살펴보았다. 변수 간의 상관관계가 높아 발생하는 다중공선성 문제를 해결하기 위해 VIF 값이 가장 높은 기온 관련 변수는 제거하였다. 나머지 강수량, 풍속, 상대습도 변수들 또한 서로 상관계수가 높게 확인되어, 그 중 일 강수량, 평균 풍속, 평균 상대습도만 사용하였다. 특보 데이터와 관련이 있을 최소, 최댓값보다는 평균값을 선택하였고, 순간 최대 풍속은 전체적인 추이를 보는 데에 유용하지 않을 것이라 판단해 제거하였다.

4.2 기상변수와 화재 빈도수의 상관성

지역별로 세 가지 기상 변수들과 화재 빈도수 간의 상관관계를 확인했을 때 결과는 (표1)과 같았다. 대부분의 지역에서 강수량과 상대습도는 화재빈도수와 음의 상관관계를, 평균풍속은 화재빈도수와 양의 상관관계를 가짐을 확인하였다. 또한 화재 빈도수와 기상 변수들간의 회귀분석을 통해 이러한 상관성이 유의미함을 한 번 더 입증하였다.



(그래프1) 월별, 지역별 빈도(막대그래프)와 기상 변수 간 그래프

(그래프1)은 정규화한 월별 화재빈도수와 평균 상대습도, 강수량, 평균풍속을 지역별로 나타낸 것이다. 대부분의 지역에서 빈도수와 평균풍속의 그래프 추이는 비슷하고, 빈도수와 평균 상대습도, 강수량은 반대로 움직인다는 것을 확인할 수 있다.

4.3 기상변수와 화재 피해 정도의 상관성

지역별로 세 가지 기상 변수들과 화재 피해 정도(property_damage) 간의 상관관계를 확인했을 때 결과는 (표2)와 같았다. 기상 변수와 재산 피해 정도의 직접적인 상관관계는 크지 않았다.

다만 월별 재산 피해정도의 추이를 보았을 때, 여름과 겨울에 특히 피해 정도가 크다는 것을 확인할 수 있었다. 우리는 이를 화재가 가장 많이 일어나는 장소가 건축이고, 미상을 제외하고 가장 많은 발화요인이 부주의 및 전기적 요인이라는 것을 고려하여, 여름과 겨울철 냉난방과 관련된 사고로 인함이라고 판단하였다.

여름과 겨울철 냉난방 사고는 기온과 밀접한 관련이 있으므로, 기상변수 중 기온은 화재 피해 정도와 간접적인 상관성이 있다.

	rn_day	ws_mean	hm_mean
강원특별자치도	-0.7898	0.8664	-0.9542
경기도	-0.5267	0.5997	-0.8302
경상남도	-0.8830	0.6615	-0.9511
경상북도	-0.8205	0.8640	-0.9482
광주광역시	0.1549	0.7120	-0.2723
대구광역시	-0.5042	0.5906	-0.8557
대전광역시	-0.2320	0.7169	-0.6705
부산광역시	-0.6376	-0.2283	-0.8098
서울특별시	-0.1248	0.3670	-0.3337
세종특별자치시	0.0648	0.3904	-0.3226
울산광역시	-0.7406	0.8366	-0.8940
인천광역시	-0.4998	0.6805	-0.7042
전라남도	-0.7190	0.9139	-0.9162
전북특별자치도	-0.5411	0.8571	-0.8295
제주특별자치도	-0.4531	0.2546	-0.6716
충청남도	-0.6131	0.6156	-0.8680
충청북도	-0.5684	0.7509	-0.9125

	rn_day	ws_mean	hm_mean
강원특별자치도	0.1110	-0.6977	0.3617
경기도	-0.5207	-0.5681	-0.2311
경상남도	-0.1805	-0.0904	-0.2255
경상북도	-0.4882	0.3528	-0.4364
광주광역시	-0.3657	-0.2899	-0.1276
대구광역시	0.2795	-0.7429	0.5814
대전광역시	0.0024	-0.5534	0.3466
부산광역시	0.3149	-0.1528	0.2074
서울특별시	-0.1881	0.0080	-0.3591
세종특별자치시	-0.1146	-0.2293	0.0089
울산광역시	0.1303	-0.3031	0.0269
인천광역시	-0.5493	0.2340	-0.5729
전라남도	0.0706	-0.3700	0.3499
전북특별자치도	-0.3777	-0.0704	-0.2475
제주특별자치도	-0.2572	0.5042	-0.1332
충청남도	-0.3839	-0.7099	0.0475
충청북도	0.2554	-0.8098	0.6743

(표1 (왼)) 월별, 지역별 화재 발생 빈도수(fire_count)와 기상 변수 간의 상관계수

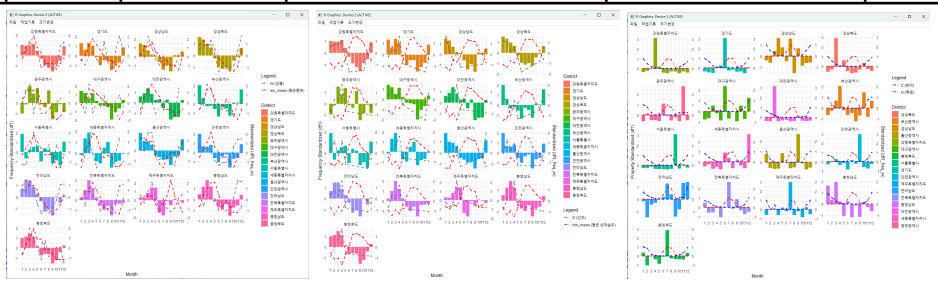
(표2 (오)) 월별, 지역별 화재 재산 피해 정도(median_property_damage)와 기상 변수 간의 상관계수

4.4 기상 특보와 화재발생의 상관성

4.2와 4.3에서의 상관분석을 통해 화재 빈도수는 강수량, 평균 풍속, 평균 상대습도와 관련이 있고, 화재 피해 정도는 기온과 관련이 있음을 파악할 수 있었다.

여러 시각화를 통해 평균 풍속은 강풍특보(W)와, 평균 상대습도는 건조 특보(D)와, 기온은 폭염(H), 한파(C) 특보와 밀접한 관련이 있다는 것을 알 수 있었으며, 이를 통해 기상 특보와 화재 발생의 상관성도 따져볼 수 있었다.

	특보	기상 요소	대체로 많이 발생하는 월	특히 유효한 지역	대처
화재 빈도	강풍	ws_mean	3~5월	강원특별자치도	화재 발생 조심, 인력 배치 증가
	건조	hm_mean	4월 (2~4월)	강원, 대구, 부산, 제주, 충북	
재산 피해	폭염	(온도)	6-8월	전남	냉난방기 사용 조심, 인력 배치 증가
	한파		12-1월	전남	



(그래프2 (왼)) 월별, 지역별 화재 발생 빈도수와 강풍, ws_mean 정규화 그래프

(그래프3 (중)) 월별, 지역별 화재 발생 빈도수와 건조, hm_mean 정규화 그래프

(그래프3 (오)) 월별, 지역별 화재 재산 피해 정도와 한파, 폭염 정규화 그래프

5. 활용 방안 및 기대효과

기상과 화재 데이터의 상관 분석에서 화재 발생 빈도와 화재 발생 피해 재산 규모를 기준으로 보았을 때 상대 습도, 풍속, 강수가 월별 분포에서 유의한 상관을 가진다는 것을 볼 수 있었다. 이에 따라 각 지역 자치에서는 해당 지역에 화재가 발생할 확률이 높거나, 재산 피해 규모가 클 것이라 예상되는 월에 화재에 대한 경계를 강화할 필요가 있다. 따라서 정부는 이를 고려해 소방 인력을 효율적으로 활용하여 화재 피해를 줄이는 효과를 가질 것이다. 또한 캠페인, 광고 등을 통해 강풍, 건조, 폭염, 한파 등의 특보가 자주 내려지는 시기에는 국민들이 화재의 위험성을 알 수 있도록 홍보해야 한다.

이 분석을 통해 기상 요인과 화재 발생의 상관관계를 명확히 규명하고, 보다 신뢰성 있는 화재 예방 경보 시스템을 설계하는 데 기여하고자 합니다. 궁극적으로, 이러한 연구는 화재 예방 및 대응 시스템의 효율성을 향상시키고, 재산 피해와 인명 피해를 줄이는 데 중요한 기초 자료를 제공할 것입니다.