

날씨와 유형 별 산악사고의 연관 관계

팀명 : 라차타

팀원 : 이연주, 임정주, 최성필, 홍기대

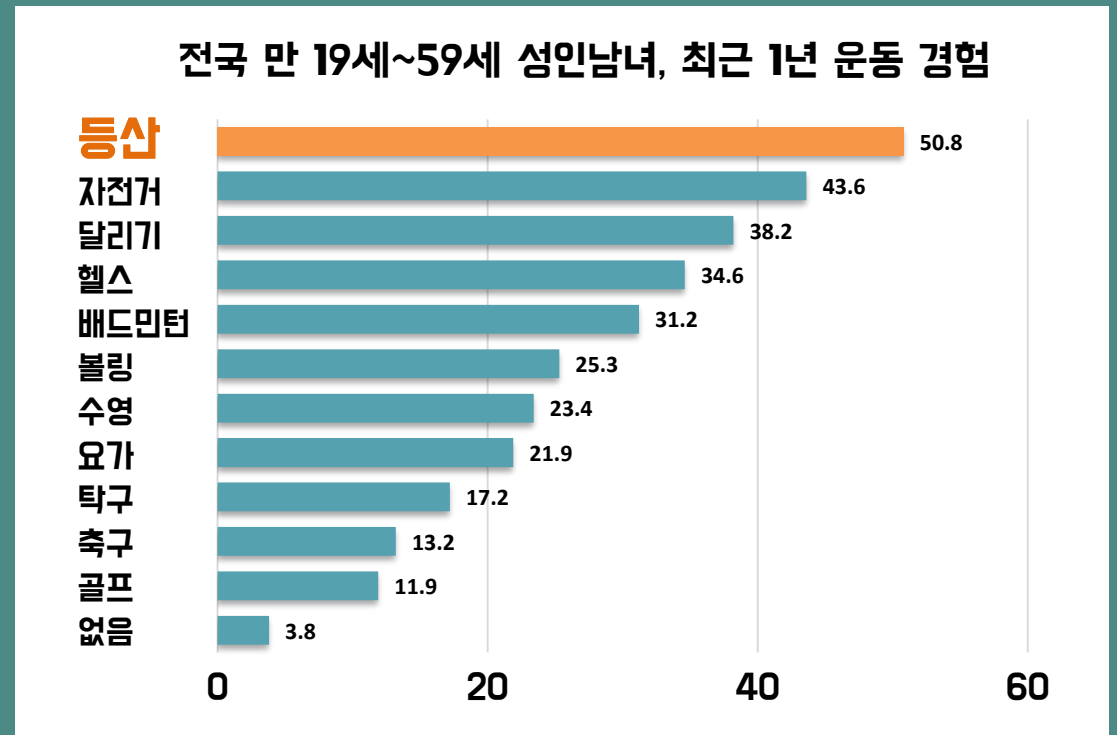
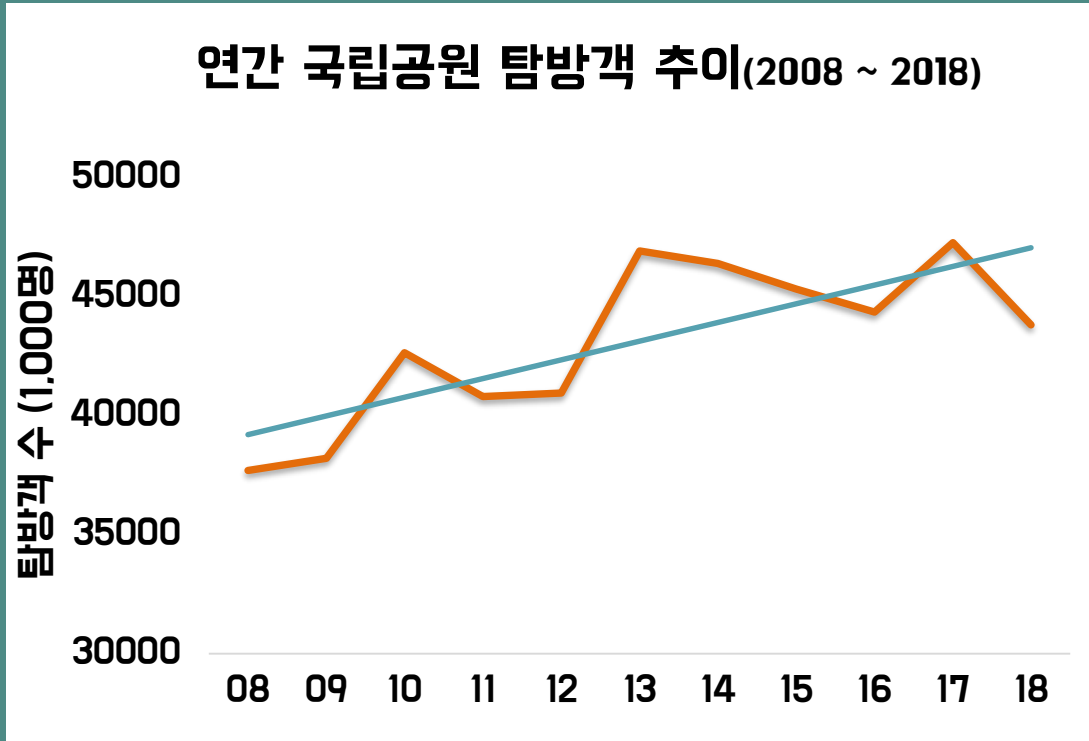


목 차

- 1. 공모 배경**
 - | 주제 선정 배경 및 이유
- 2. 데이터 정의**
 - | 데이터 수집 | 데이터 탐색
- 3. 데이터 전처리 및 분석 계획**
 - | 데이터 전처리 | 데이터 전처리 결과 | 분석 계획
- 4. 분석**
 - | 로지스틱 분석 및 결과
- 5. 최종 결과**
 - | 서비스 활용 방안 | 기대 효과

1 공모배경

1. 주제 선정 배경 및 이유



대한민국 취미 1위 등산

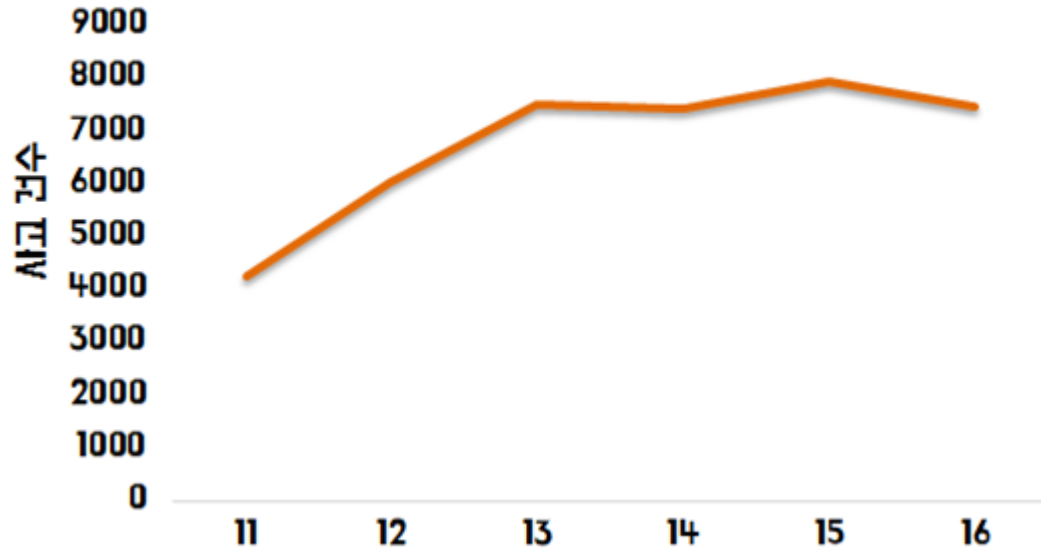
자료 출처 :

1) 국립공원 통계 연보

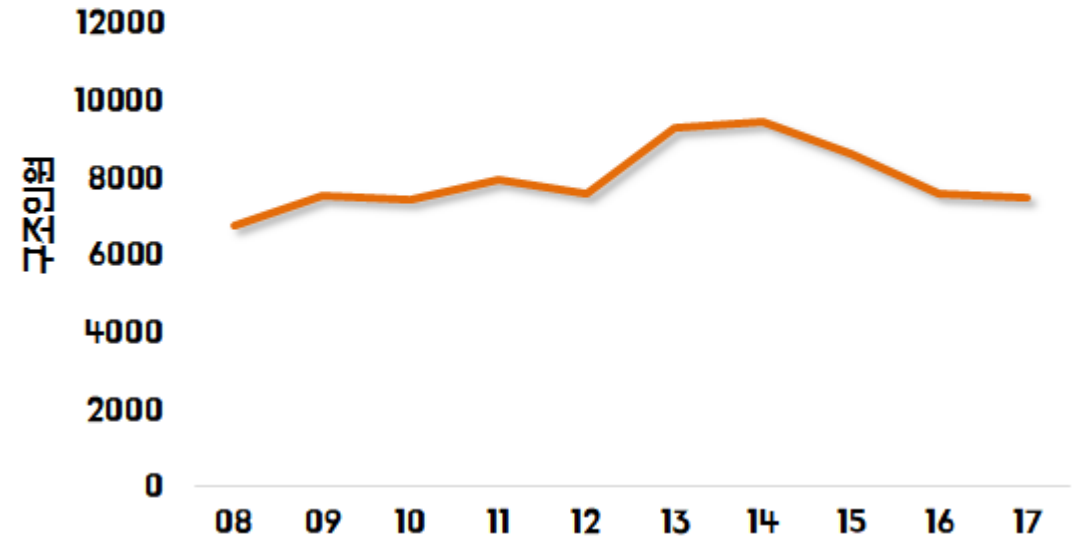
2) 김세진, 데이터숨, "최근 1년 기준 많이 해 본 운동은 '등산'", 2017.03.07

1. 주제 선정 배경 및 이유

연간 등산사고 발생 건수(2011 ~ 2016)



연간 산악 사고 구조 인원 현황(2008 ~ 2017)



증가하는 등산 사고

자료 출처 :

3) 이영민, 열린뉴스, 행안부, "춡고 미끄러운 겨울산행, 안전사고 조심하세요!", 2018.01.14

4) 전병권, 남해시대, 5월, 산행 사고가 많은 달", 2017.05.16

5) 소방청

2 데이터 정의

2.1 데이터 수집



기상 데이터

- 출처 : 기상청 (기상자료 개방 포털 종관 기상 관측)
- 지역 : 서울, 경기, 경북, 경남
- 기간 : 2013~2018
- 요소: 기온, 강수, 바람, 기압, 습도, 일사, 일조, 눈, 구름, 시정, 지면상태, 증발량 등



산악 사고 데이터

- 출처 : 소방청
- 지역 : 서울, 경기, 경북, 경남
- 기간 : 2013~2018
- 요소: 구조 보고서 번호, 소방서명, 센터명, 등록일시, 신고일시, 사고원인, 사고장소, 활동개요

2.1 데이터 수집 - 데이터 타입

기상 데이터

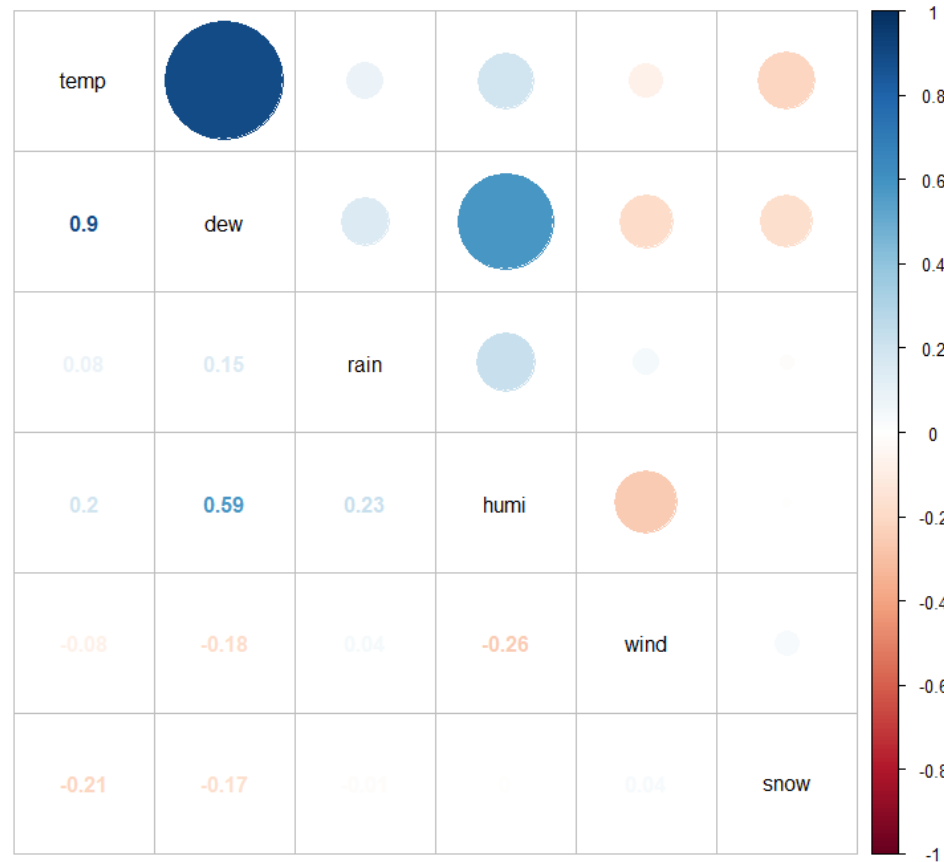
변수	변수 타입
날짜	Date
기온	Float
습도	Float
온도	Float
풍속	Float
강수량	Float
이슬점	Float
적설량	Float

산악 사고 데이터

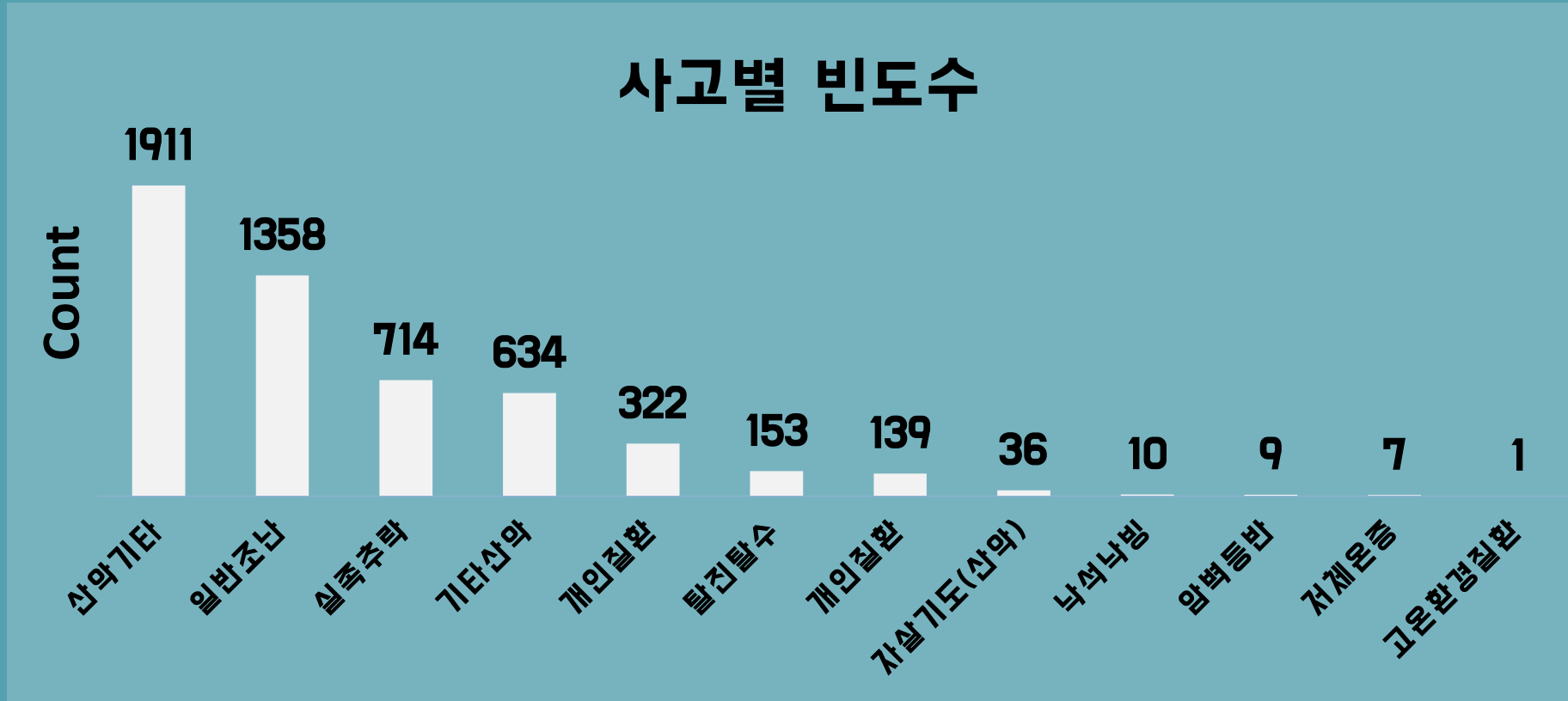
변수	변수 타입
사고일시	Date
신고일시	Date
사고원인	String

날씨와 산악 사고 데이터의 자료형

2.2 데이터 탐색 - 날씨 변수간 상관 관계



2.2 데이터 탐색



산악 사고 중 가장 빈도가 높은 상위 6가지의 원인 선택

3 데이터 전처리

3.1 데이터 전처리 - 분석 데이터 선정

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
지점	일시	기온(°C)	강수량(mn	풍속(m/s)	풍향(16방	습도(%)	증기압(hP	이슬점온도	현기습기압(h	해면기압(h	일조(hr)	일사(MJ/n	적설(cm)	3시간신적	전운량(10	중하층운량	운형(운형	최저운고(1
108	#####	-3.2		0.5	110	40	1.9	-14.9	1015.4	1026.4					0	0		
108	#####	-3.3		0.7	360	41	2	-14.7	1015.1	1026.1						0		
108	#####	-3.7		0.9	270	42	2	-14.7	1015.2	1026.3						0		

	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1	소방서명	센터명	등록일시	신고일시	사고원인	사고장소	활동개요											
2	강원소방본1	항공대	201301191203	20130119121400	개인질환	등산	○ 구조출동/n태백산 유일사 호흡곤란 요구조자 1명 발생에 따라 강원002이륙하여 도착하여 구조대원 문도권 하강하여 선착한 태백											
3	강원소방본1	항공대	201301191234	20130119123800	산악기타	정상	○ 구조출동/n태백산 장군봉 하지 근육경련 요구조자 1명 발생에 따라 강원002이륙하여 도착하여 구조대원 문도권 하강하여 호이스											
4	강원소방본1	항공대	201301191514	20130119151800	산악기타	정상	○ 구조출동/n만산동 계곡 만산정상 조측무류 통증 요구조자 1명 발생에 따라 강원002이륙하여 도착하여 구조대원 문도권 하강하여											
5	강원소방본1	항공대	201301191528	20130119163300	산악기타	정상	○ 구조출동/n태백산 천제단 인근 양측하지 통증 요구조자 1명 발생에 따라 강원002이륙하여 도착하여 구조대원 문도권 하강하여											
6	강원소방본1	항공대	201301201433	20130120144400	개인질환	정상	○ 구조출동/n계방산 정상 05-09지점 양측하지 통증 요구조자 1명 발생에 따라 강원002이륙하여 도착하여 패드장 랜딩구조 완료/n											

1. 불필요한 컬럼 제거

- 사고 데이터 : 소방서명, 센터명, 등록일시, 사고장소, 활동개요
- 날씨 데이터 : 지점, 풍향, 증기압, 일사, 일조 등등

3.2 데이터 전처리 - 컬럼 명 변경 및 병합

날씨 데이터

변경 전	변경 후
날짜	DATE
기온	TEMP
강수량	RAIN
적설	SNOW
습도	HUMI
이슬점 온도	DEW
풍속	WIND

지역별 사고 데이터

변경 전	변경 후
날짜	DATE
사고 원인	CAUSE

1. 지역별 사고 데이터를 컬럼 명 기준으로 INNER JOIN
2. 날씨와 사고 두 데이터의 날짜 표기 형식을 동일하게 맞추는 후,
날씨데이터의 날짜 컬럼 기준으로 OUTER JOIN

3.3 데이터 전처리 - 누락 값 처리

A	B	C	D	E	F	G
Cause	humi	temp	wind	rain	dew	snow
산악기타	32			0	-9.25	0
산악기타				0		4.5
산악기타				0		4.5
실족추락	40.5			0	-5.2	0
실족추락	40.5			0	-5.2	0
일반조난	25			0	-11.6	0
일반조난	27			0	-11.7	0
일반조난	34			0	-8.8	0

Cause	humi	temp	wind	rain	dew	snow
실족추락	51.5	30.525	2.175	0	21.575	0
개인질환	51.25	30.575	1.225	0	21.775	0
산악기타	51.5	30.525	2.35	0	21.825	0
산악기타	58	30.6	2.075	0	22.75	0
산악기타	51.25	30.5	1.975	0	21.675	0
산악기타	51.25	30.5	1.975	0	21.675	0
산악기타	51.25	30.5	1.975	0	21.675	0
실족추락	51.25	30.5	1.975	0	21.675	0

1. Snow, Rain 의 누락 값은 기상청 데이터와 비교 결과 0인 경우 누락인 것을 확인.
=> 0으로 대체
2. 나머지 컬럼은 평균 대체
: 이전과 이후 값의 평균 값으로 대체하여 누락 값 제거

3.4 데이터 전처리 - 데이터 범주화

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
Cause	humi	temp	wind	rain	dew	snow	dist	other	fall	ill	exhausting	climb	rockslide	hypothermia
일반조난	34.5	-12.725	3	0	-23	11.25	1	0	0	0	0	0	0	0
일반조난	46.25	-5.675	1.775	0	-14.65	8.925	1	0	0	0	0	0	0	0
산악기타	62.25	-2.9	2.475	0	-7.35	3.7	0	1	0	0	0	0	0	0
일반조난	59.5	-0.3	1.8	0	-6	3.525	1	0	0	0	0	0	0	0
일반조난	55	0.7	1.4	0	-5	3.45	1	0	0	0	0	0	0	0

	A	B	C	D	E	F	G
1	temp	wind	rain	humi	dew	snow	accident
2	5.25	1.9	0	58	-6.6	0	0
3	5.15	2.4	0	50.5	-6.6	0	1
4	4.5	1.9	0	48.5	-6.1	0	0
5	4.15	2.95	0	55.5	-4.4	0	0

1. 산악사고 발생 여부에 따라 0(발생하지 않음)과 1(발생)로 범주화 하여 accident 컬럼 생성
2. 각 산악사고 원인별로 발생 여부에 따라 0과 1로 범주화 하여 사고 원인 명으로 새로운 컬럼(dist, other, fall, ill, exhaustion, climb, rockslide, hypothermia) 생성

3.5 최종 데이터

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
Cause	humi	temp	wind	rain	dew	snow	dist	other	fall	ill	exhausting	climb	rockslide	ypothermia
일반조난	34.5	-12.725	3	0	-23	11.25	1	0	0	0	0	0	0	0
일반조난	46.25	-5.675	1.775	0	-14.65	8.925	1	0	0	0	0	0	0	0
산악기타	62.25	-2.9	2.475	0	-7.35	3.7	0	1	0	0	0	0	0	0
일반조난	59.5	-0.3	1.8	0	-6	3.525	1	0	0	0	0	0	0	0
일반조난	55	0.7	1.4	0	-5	3.45	1	0	0	0	0	0	0	0
일반조난	46.5	1.875	1.9	0	-6.1	3.075	1	0	0	0	0	0	0	0
일반조난	44	2.425	1.725	0	-6.55	2.425	1	0	0	0	0	0	0	0
일반조난	44	2.425	1.725	0	-6.55	2.425	1	0	0	0	0	0	0	0
일반조난	43.75	2.55	1.8	0	-6.475	2.3	1	0	0	0	0	0	0	0

	A	B	C	D	E	F	G
1	temp	wind	rain	humi	dew	snow	accident
2	5.25	1.9	0	58	-6.6	0	0
3	5.15	2.4	0	50.5	-6.6	0	1
4	4.5	1.9	0	48.5	-6.1	0	0
5	4.15	2.95	0	55.5	-4.4	0	0
6	4	2.45	0	55	-4.4	0	1
7	3.1	1.7	0	58	-4.05	0	1
8	2.75	2.05	0	63.5	-3.35	0	0
9	2.7	2.2	0	68.5	-2.75	0	1

3.6 분석 계획

날씨와 사고 발생 사이에 연관성이 있을까?

연관성이 있다면,

각각의 사고 원인에 어떤 날씨 요소가 영향을 끼칠까?

4 분석 결과

4.1.1 날씨 vs 사고 발생 연관 분석 - 가설

귀무 가설

: 날씨가 사고 발생 여부에 영향을 끼치지 않는다.

대립 가설

: 날씨가 사고 발생 여부에 영향을 끼친다.

4.1.2 날씨 vs 사고 발생 연관 분석 - 분석

1. 변수

X : temp, wind, rain, humi, dew, snow

Y : accident (0, 1)

2. 전체 데이터를 3:7비율로 train, test set으로 나눔

3. 로지스틱 회귀분석

4.1.3 날씨 vs 사고 발생 연관 분석 - 로지스틱 분석 결과

Logit Regression Results						
=====						
Dep. Variable:	accident	No. Observations:	147100			
Model:	Logit	Df Residuals:	147094			
Method:	MLE	Df Model:	5			
Date:	Wed, 10 Jul 2019	Pseudo R-squ.:	0.08361			
Time:	15:45:25	Log-Likelihood:	-40419.			
converged:	True	LL-Null:	-44106.			
		LLR p-value:	0.000			
=====						
	coef	std err	z	P> z	[0.025	0.975]

temp	0.0030	0.002	1.892	0.058	-0.000	0.006
wind	0.0638	0.009	6.932	0.000	0.044	0.082
rain	0.0248	0.012	2.002	0.045	0.001	0.049
humi	-0.0495	0.000	-139.164	0.000	-0.050	-0.049
dew	0.0586	0.002	35.033	0.000	0.055	0.062
snow	0.1139	0.011	10.394	0.000	0.092	0.135
=====						

오즈 비

```
* Odds Ratio :  
temp      1.002997  
wind       1.065924  
rain       1.025160  
humi       0.951664  
dew        1.060337  
snow       1.120664  
dtype: float64
```

Temp를 제외한 나머지 날씨 요인이 **유의**하다
각 날씨 요인의 Odds Ratio로 사고 발생 확률에 영향력을 파악하였다

4.1.4 날씨 vs 사고 발생 연관 분석 - 검증1. K-fold Cross Validation

Logistic Regression

Accuracy : 0.9091096553146265

Score of K-fold Cross Validation :

[0.91080897 0.90815772 0.91305235 0.91169273 0.91210061]

Mean: 0.91

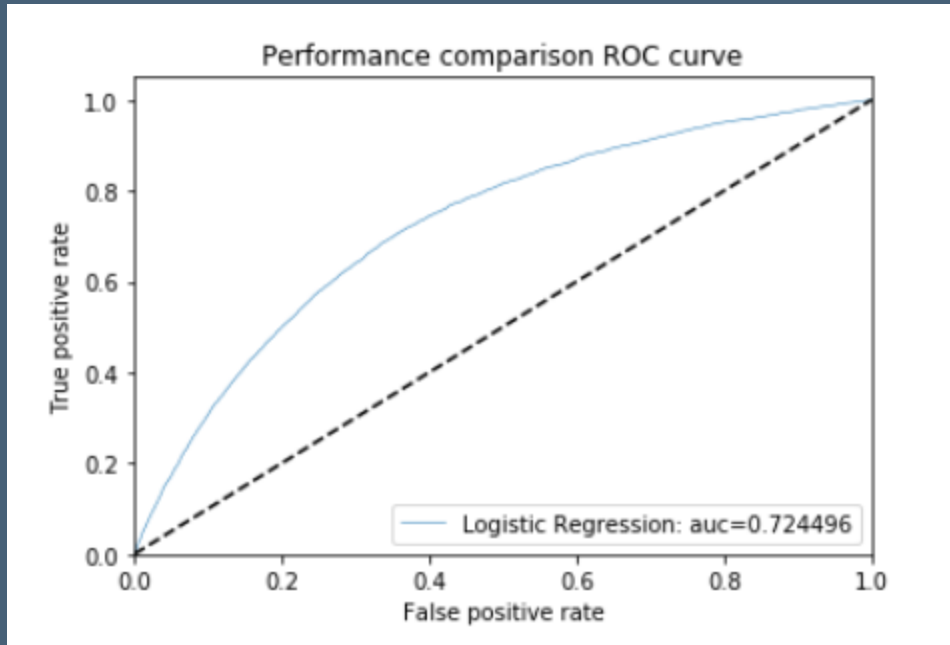
K-fold Cross Validation (K겹 교차 검증)

: 데이터를 k개의 조각으로 나눠

1개 부분을 testing으로, 나머지 k-1개 부분을 training으로 사용해 검증.
testing을 돌려가면서 k번 반복해 검증하는 교차 검증 방법.

K-fold Cross Validation 으로 **5번**의 검증을 반복하여 모델을 검증하였다.

4.1.4 날씨 vs 사고 발생 연관 분석 - 검증2. ROC - Curve



1) ROC - Curve :

FPR과 TPR을 각각 x, y축으로 놓은 그래프

* FPR : 1-특이도 (= 0을 1로 잘못 예측한 비율)

* TPR : 민감도 (= 1을 1로 예측한 비율)

2) AUC(the Area Under a ROC Curve) :

ROC 커브의 밑면적을 구한 값.

1에 가까울수록 성능이 좋다.

ROC - CURVE를 통해 **모델의 정확성**을 검증하였다.
AUC = 0.72

4.2.1 날씨 vs 사고 발생 유형별 연관 분석 - 가설

귀무 가설

: 각 사고 별로 영향을 끼치는 날씨요인이 같다.

대립 가설

: 각 사고 별로 영향을 끼치는 날씨요인이 같지 않다.

4.2.2 날씨 vs 사고 발생 유형별 연관 분석 - 분석

1. 변수

X : temp, wind, rain, humi, dew, snow

Y : dist, other, fall, ill, exhausting, climb, rockslide, hypothermia,
exhaustion 발생 유무 (0 / 1)

2. 로지스틱 회귀분석

4.2.3 날씨 vs 사고 발생 유형별 연관 분석

- stepwise 로지스틱 회귀분석 with AIC

```
Start:  AIC=2124  
climb ~ temp + wind + rain + dew + snow + humi
```

		Df	Deviance	AIC
-	wind	1	2111.3	2123.3
-	rain	1	2111.4	2123.4
<none>			2110.0	2124.0
-	snow	1	2113.8	2125.8
-	temp	1	2121.4	2133.4
-	dew	1	2128.5	2140.5
-	humi	1	2130.9	2142.9

```
Step:  AIC=2122.54  
climb ~ temp + dew + snow + humi
```

		Df	Deviance	AIC
<none>			2112.5	2122.5
-	snow	1	2116.3	2124.3
-	temp	1	2124.7	2132.7
-	dew	1	2132.0	2140.0
-	humi	1	2136.8	2144.8

AIC 를 통해 각 사고 유형에 영향을 끼치는 날씨 요인 파악,
회귀 모형 **최적화**를 실시하였다

4.2.3 날씨 vs 사고 발생 유형별 연관 분석 - 로지스틱 분석 결과

```
Call:
glm(formula = climb ~ temp + dew + snow + humi, family = "binomial",
    data = df)
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-0.2921	-0.1195	-0.1026	-0.0843	3.5800

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-0.62220	1.10202	-0.565	0.572344
temp	-0.14335	0.04211	-3.404	0.000664 ***
dew	0.19290	0.04518	4.270	1.96e-05 ***
snow	-92.77519	3130.93307	-0.030	0.976361
humi	-0.07143	0.01511	-4.721	2.27e-06 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

< 0.05

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 2157.8 on 31647 degrees of freedom
Residual deviance: 2112.5 on 31643 degrees of freedom
AIC: 2122.5

Number of Fisher Scoring iterations: 23

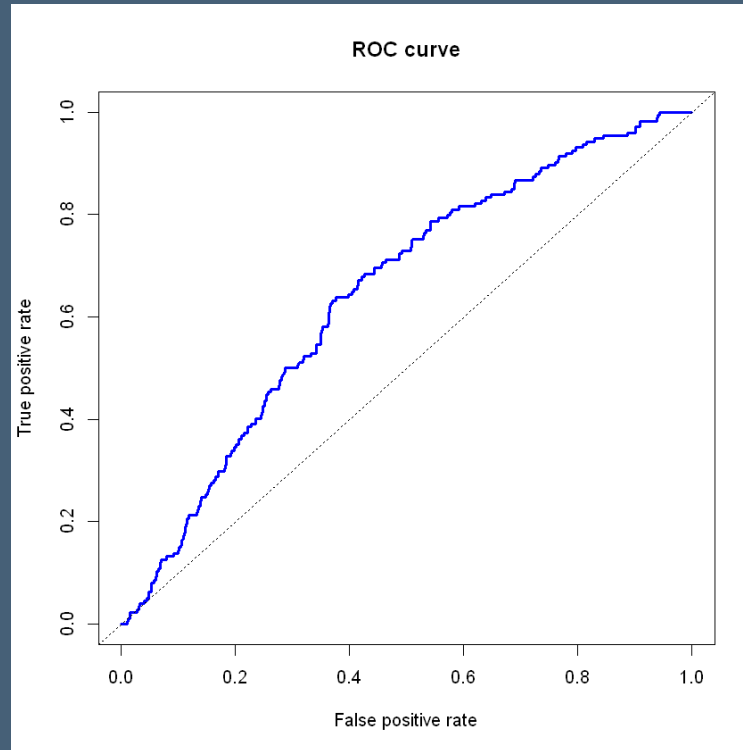
오즈 비

* Odds Ratio :

(Intercept)	0.53676089226258
temp	0.866453565274097
dew	1.21275930989304
snow	5.10796096218099e-41
humi	0.931056991496511

Climb 사고 요인에는 temp, dew, humi 날씨 요소가 **유의**하다
각 날씨 요인의 Odds Ratio로 사고 발생 확률에 영향력을 파악하였다

4.2.4 날씨 vs 사고 발생 유형별 연관 분석 - 검증. ROC Curve



ROC - CURVE를 통해 **모델의 정확성**을 검증하였다.
AUC = 0.60

4.2.4 날씨 vs 사고 발생 유형별 연관 분석 - 로지스틱 분석 결과

1. 로지스틱 회귀분석 결과 - 각 사고 유형에 영향을 끼치는 날씨요소

암벽 등반 - 기온, 이슬점 온도, 습도

탈진 탈수 - 강수량, 이슬점 온도, 습도

저 체온증 - 이슬점 온도, 습도

일반 조난 - 적설량, 풍속, 기온, 습도

실족 추락 - 습도, 기온, 풍속

개인 질환 - 기온, 강수량, 이슬점 온도, 습도

산악 기타 - 기온, 이슬점 온도, 적설량, 습도

낙석 낙빙 - 강수량

2. ROC Curve 결과

암벽 등반, 탈진 탈수, 저 체온증은 **높은 모델 정확성**을 보였지만,
나머지 사고 원인에 대한 ROC Curve 결과는 비교적 낮은 정확성을 보였다.

5 최종 결과

5.1 활용 방안

1. 다양한 등산 정보와 산악 별 기상 정보 **알림 앱 서비스 개발**

- 오늘의 등산 위험도, 안전 등산로, 고도에 따른 날씨 제공, 등산 장비 추천, 사고 현황 제공

2. 산악 구조대와의 협력을 통해, 등산객에 대한 **안전망 구축 및 사고 예방**

3. 날씨에 따른 지역별 응급 처치 도구의 **효율적 배치**

5.2 기대 효과

1. 오늘의 산악 날씨와 위험도 제공으로 사고율 하락
2. 산악 구조 장비의 부족함 해결과 효율적 배치 및 자체 구급함 활용
3. 유동적인 소방 인력 충원 및 산악 구조대 배치
4. 등산 위험도 및 사고 현황을 통해 산악 사고에 대한 경각심 부여

THANKS

참고 문헌

- 1) 2019 국립공원 기본 통계, 연간 탐방객 추이, 5쪽
- 2) 김세진, 데이터솜, “최근 1년 기준 많이 해 본 운동은 ‘등산’”, 2017.03.07,
<http://www.datasom.co.kr/news/articleView.html?idxno=1902>
- 3) 이영민, 열린뉴스, “행안부, “춡고 미끄러운 겨울산행, 안전사고 조심하세요!””, 2018.01.14, <http://www.opennews.kr/1759>
- 4) 전병권, 남해시대, “5월, 산행 사고가 많은 달”, 2017.05.16,
<http://m.nhtimes.co.kr/news/articleView.html?idxno=38754>
- 5) 2018 소방청 통계연보, 연도별(사고종류) 구조인원 현황(2008~2017), 135쪽