【붙임 2】

데이터 분석 최종결과보고서

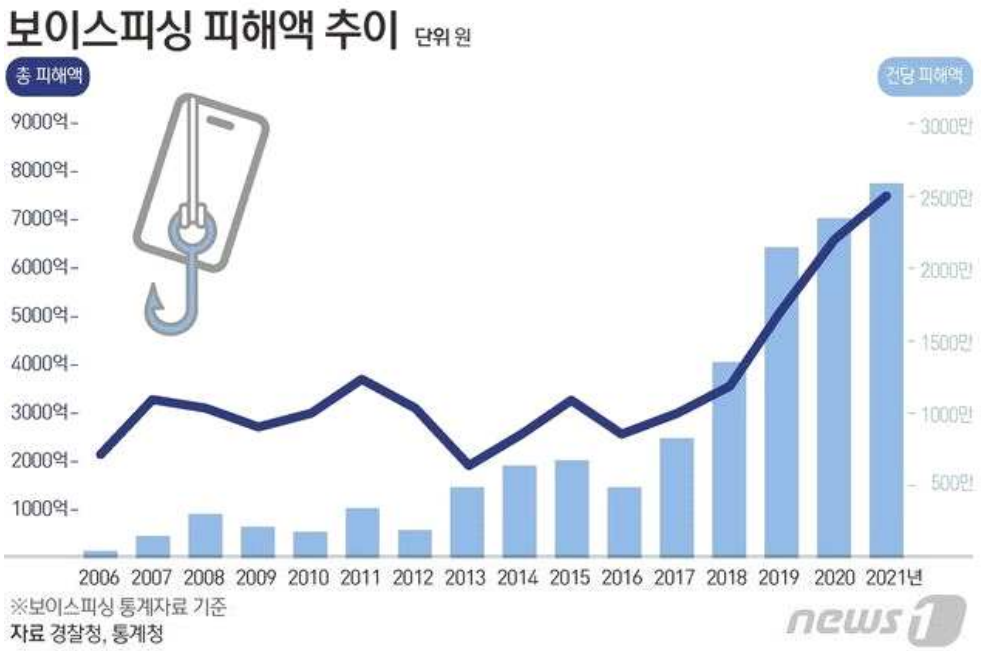
Ⅰ. 참가자 정보

제 목	보이스피싱 데이터 분석을 통해 피해 축소 및 예방법 구축	
팀 명	Meta - Cognition	
성 명	국성우	
연락처	휴대폰	010 - ■■■■ - ■■■■
	E-mail	■■■■@■■■■.com

Ⅱ. 개요

1. 분석 배경

시대가 발전할수록 개인의 금융사기 피해는 점점 심각해지고 있음. 특히 전화 금융사기는 최근 심각한 문제로 떠오르고 있으며, 피해자의 금융안전까지도 위협하고 있음. 보이스피싱은 범행 대상자에게 전화를 걸어 금융감독원이나 수사기관, 지인 위장, 대출 사기등을 허위로 말하면서 협박하여, 불안감을 먼저 조성하고, 송금을 요구하거나 특정 개인정보를 수집하는 사기 수법을 말하는데, 최근 계속되는 예방 교육 및 정부 차원에서 오랜 기간 예방책을 홍보, 관련 규정을 강화하는 등 관련조치가 향상되고 있으나, 날이 갈수록 범죄 수법이 교묘하게 진화하고 발전하기 때문에, 완전히 근절하기 어렵고 꾸준히 발생하고 있는 현실임.



(보이스피싱 통계자료 출처: 통계청)

최근까지 코로나19 정부 지원금 지원을 빙자하여 피해자를 현혹하거나, 스마트폰에 원격 조정 애플리케이션 설치를 유도하는 등 새로운 수법들이 발생. 뿐만 아니라 그에 따른 피해액 추이도 증가하고 있는 추세임. 연구분야도 발맞춰 제도적인 측면에서 보이스피싱에 대한 대응책을 꾸준히 제시해 왔지만, 국내에서는 금융 관련, 통신 서비스 이용률이 높은 만큼 누구나 쉽게 피해자가 될 수 있음. 개인적인 차원에서 보이스피싱은 피해자의 재산 손실은 물론이고, 발생 이후 우울증, 심리적 무기력함 및 불안감 등 다양한 부작용을 발생시키며, 가정 파괴나 국부 유출 나아가, 사회의 경제적 문제까지 초래할 수 있음. 따라서 소비자의 특징과 사건 발생 연관관계를 파악하고 효과적인 예방책을 마련해, 보이스피싱 피해를 낮추기 위함에 목적을 둠.

2. 분석 · 시각화 목적

2.1 분석 · 시각화 목적

- 경찰청제공 보이스피싱 피해 데이터들의 지표(접수완료 일시, 성별, 발생지점, 동일사건여부), 발생지점의 이미지 공간좌표를 활용하여, 보이스피싱 발생 지점의 패턴과 연계분석함.

- 정확한 분석을 위해 새로운 데이터 컬럼들을 생성 후 서로 융합하고, 사건발생 상관관계 예측.

- 분석한 데이터셋과 금융사기 탐지에 활용되고 있는 이상거래탐지시스템(FDS)을 융합하여, 발생방지 및 보이스피싱의 대응방안 솔루션 제시

2.2 활용방안

● 본 데이터 분석의 결과물은 보이스피싱과 관련된 연구 전반에 활용이 가능하며, 궁극적으로는 보이스피싱 단절 분야에서 사용될 것으로 예상

- 또한 데이터 부족으로 어려움을 겪고 있는 경찰청, 연구자에게 데이터를 제공함으로써 관련 AI 서비스, 솔루션 활성화에 기여할 것으로 기대함.

● 학습용 데이터 구축과 이를 통한 예측 및 분석은 맞춤형 보이스피싱 단절에 다가갈 수 있고, 이는 보이스피싱 검거 성적 향상과 사회경제학적인 부담을 덜 수 있음.

● 실제 경찰청 데이터 外 공공데이터 분석을 통해 기존 보이스피싱 예방 지침을 뒷받침하거나 새로운 가이드라인 제정에 도움을 줄 수 있음.

- 또한 일시 패턴과 위치기반 분석을 추가하여 새로운 보이스피싱 관련 규정에 도움이 되는 기초 자료로 활용 가능

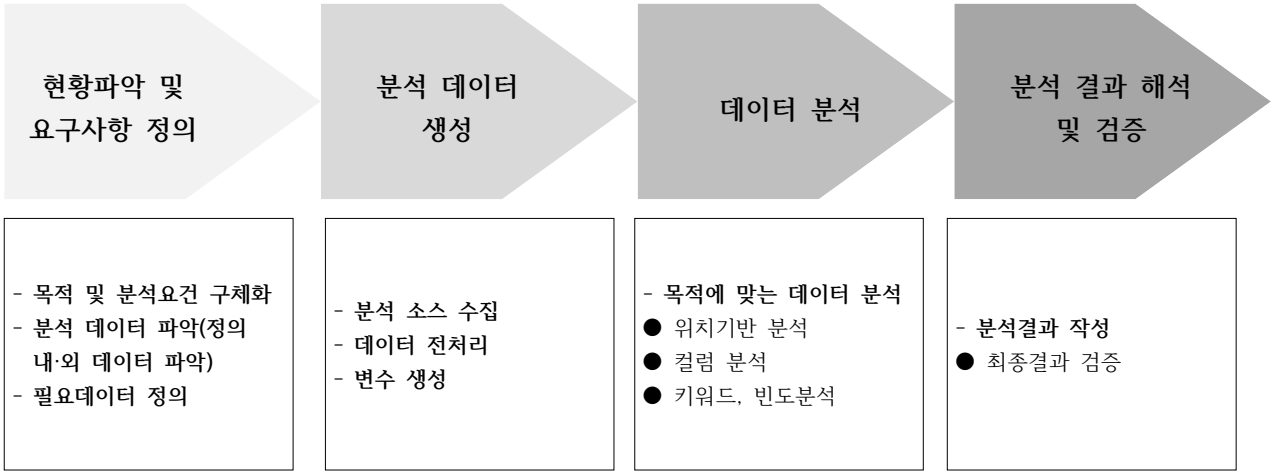
● 기존의 보이스피싱 관련 해석의 어느 정도 차이를 극복하고, 불필요한 분석을 낮추어 사건 검거를 하는 데에 도움이 될 것임.

- 구축된 양질의 데이터 셋을 활용하여 경찰청을 위한 사건 해석에 대한 교육 및 수련에도 활용될 수 있겠음.

● 구축된 데이터는 보이스피싱을 비롯한 타 사건에도 접목하여, 여러 융합범죄 연구의 활성화에 기여할 수 있음.

Ⅲ. 분석/시각화 결과 상세내용

1. 분석 프로세스



2. 분석 데이터 생성

2.1 분석데이터

<table>
<tr><th>데이터명</th><th>데이터설명</th></tr>
<tr><td>경찰청
신고데이터</td><td>- KP 2020, 2021 데이터 각 7만여건, 100만여건
<table>
<tr><th>컬럼ID</th><th>컬럼설명</th></tr>
<tr><td>RECV_DEPT_NM</td><td>접수부서 코드</td></tr>
<tr><td>RECV_CPLT_DM</td><td>접수완료일시</td></tr>
<tr><td>NPA_CL</td><td>경찰청구분 [그룹코드:32]</td></tr>
<tr><td>EVT_STAT_CD</td><td>사건상태코드 [그룹코드: 01: 사건상태]</td></tr>
<tr><td>EVT_CL_CD</td><td>사건종별코드 [그룹코드: 02]</td></tr>
<tr><td>RPTER_SEX</td><td>신고 성별 - 1: 남자 2: 여자 3: 불상</td></tr>
<tr><td>HPPN_PNU_ADDR</td><td>발생지점(PNU)</td></tr>
<tr><td>HPPN_X</td><td>발생좌표X</td></tr>
<tr><td>HPPN_Y</td><td>발생좌표Y</td></tr>
<tr><td>SME_EVT_YN</td><td>동일사건여부</td></tr>
</table>
- 보이스 피싱 사건종별코드(215)만 활용. 관련 없는 데이터 drop
- 활용데이터 NEW_KP2020 1,660건, NEW_KP2021 33,829건
<table>
<tr><th>컬럼ID</th><th>컬럼설명</th></tr>
<tr><td>RECV_DEPT_NM</td><td>접수부서 코드</td></tr>
<tr><td>RECV_CPLT_DM</td><td>접수완료일시</td></tr>
<tr><td>NPA_CL</td><td>경찰청구분 [그룹코드:32]</td></tr>
<tr><td>EVT_STAT_CD</td><td>사건상태코드 [그룹코드: 01: 사건상태]</td></tr>
<tr><td>EVT_CL_CD</td><td>사건종별코드 [그룹코드: 02]</td></tr>
<tr><td>RPTER_SEX</td><td>신고 성별 - 1: 남자 2: 여자 3: 불상</td></tr>
<tr><td>HPPN_PNU_ADDR</td><td>발생지점(PNU)</td></tr>
</table></td></tr>
</table>

	HPPN_X: 발생좌표X HPPN_Y: 발생좌표Y SME_EVT_YN: 동일사건여부 year: 사건발생년도 month: 사건발생월 day: 사건발생일 hour: 사건발생시간 minute: 사건발생분 second: 사건발생초 dayofweek: 사건발생년도(0–월요일 ~ 6-일요일) - 10개 컬럼의 원본 데이터 중 분석에 필요한 7개 컬럼 추가 생성 - year, month, day, hour, minute, second, dayofweek 변수 추가 생성
경찰청 전화금융사기 피해자 연령별 현황 데이터	- 경찰청 전화금융사기 피해자 연령별 현황 데이터(공공데이터) 매핑 **컬럼ID** / **컬럼설명** 구분 / 사건 발생 년도 구분 합계 / 발생 건수 총 합계 20대 이하 / 20대 이하 보이스피싱 발생건수 30대 / 30대 보이스피싱 발생건수 40대 / 40대 보이스피싱 발생건수 50대 / 50대 보이스피싱 발생건수 60대 / 60대 보이스피싱 발생건수 70대이상 / 70대 이상 보이스피싱 발생건수
경찰청 보이스피싱 현황 데이터	- 경찰청 보이스 피싱 현황 데이터(공공데이터) 매핑 **컬럼ID** / **컬럼설명** 구분 / 사건 발생 년도 구분 기관사칭형_발생건수 / 기관사칭형 발생건수 현황 기관사칭형_피해액_억원 / 기관사칭형 피해액 현황 기관사칭형_검거건수 / 기관사칭형 검거건 수 현황 기관사칭형_검거인원 / 기관사칭형 검거인원 현황 대출사기형_발생건수 / 대출사기형 발생건 수 현황 대출사기형_피해액_억원 / 대출사기형 피해액 현황 대출사기형_검거건수 / 대출사기형 검거건 수 현황 대출사기형_검거인원 / 대출사기형 검거인원 현황

2.2 데이터 전처리

구분	처리내용
경찰청 신고데이터	가시화 위해 [‘HPPN_PNU_ADDR‘] 컬럼 발생지점 세부주소 제외 ex) 충청남도 보령시 청라면 나원리(행정:청라면) 910-7 → 충청남도 보령시 청라면 (python 자연어 처리 사용)
경찰청 신고데이터	HPPN_X, HPPN_Y 결측 값으로 위치 파악이 불가능한 데이터들 중 결측값 예측 (구글 위치기반 api 활용 x좌표 결측 → x값 주변 가장 많이 발생한 밀집 지역 y좌표 예측 ➔ train 데이터 수가 많아짐.(예측에 용이))

2.3 변수 생성

변수	생성로직
보이스 피싱 발생 년, 월, 일, 시간, 분, 초	[‘year’, ‘month’, ‘day’, ‘hour’, ‘minute’, ‘second’] 변수 생성 보이스 피싱 발생 년, 월, 일, 시간, 분, 초([‘RECV_CPLT_DM’] → python : parsedate 사용)
보이스 피싱 발생 요일	[‘dayofweek’] 변수 생성 → [‘RECV_CPLT_DM’].dt.dayofweek 사건 발생 요일 변수 생성(0 - 월요일 ~ 6 - 일요일)
보이스 피싱 발생 지역	[‘location’] 변수 생성 python 자연어 처리 [‘HPPN_PNU_ADDR‘] → split(), join() 사용 ex) 충청남도 보령시 청라면 나원리(행정:청라면) 910-7 → 충청남도 보령시 청라면

3. 데이터 분석

3.1 일반 데이터 현황

```
import pandas as pd
import matplotlib as mpl
import matplotlib.pyplot as plt
import seaborn as sns
import os
from scipy import stats

import missingno as msno
plt.style.use('seaborn')

import warnings
warnings.filterwarnings("ignore")

df_train_20 = pd.read_csv('new_sample_kp2020.csv', parse_dates=['RECV_CPLT_DM'], encoding='utf-8')
df_train_21 = pd.read_csv('new_sample_kp2021.csv', parse_dates=['RECV_CPLT_DM'], encoding='utf-8')
df_train = pd.concat([df_train_21, df_train_20])

# parse_dates는 날짜, 시간 변수를 datetime 변수로 변환하기 위함
# KP2020 + KP2021 데이터프레임 concat -> df_train

# ['RECV_DEPT_NM', 'RECV_CPLT_DM', 'NPA_CL', 'EVT_STAT_CD', 'EVT_CL_CD', 'RPTER_SEX', 'HPPN_PNU_ADDR', 'HPPN_X', 'HPPN_Y', 'SME_EVT_YN']
# ['접수부서 코드', '접수완료일시', '경찰청구분', '사건상태코드', '사건종별코드', '신고성별', '발생지점(PNU)', '발생좌표X', '발생좌표Y', '동일사건여부']
```

기본적인 모듈과 필요한 모듈들을 import. 보이스 피싱과 관련된 데이터를 추출 후, 학습데이터를 더 많고 자세히 보기위해 KP2021, KP2020을 concat. 총 35,489건의 데이터 생성 (df_train = kp2021 + kp2020)

```
df_train.count()
✓ 0.0s
RECV_DEPT_NM     35489
RECV_CPLT_DM     35489
NPA_CL           35489
EVT_STAT_CD      35489
EVT_CL_CD        35489
RPTER_SEX        35489
HPPN_PNU_ADDR    27250
HPPN_X           29794
HPPN_Y           29794
SME_EVT_YN        5360
dtype: int64
```

각 컬럼별 로우 수를 살펴보면 사건 발생지점, 발생좌표x, y, 동일사건여부 컬럼에 결측값이 있다는 사실을 알 수 있음.

3.2 EDA를 통한 데이터 분석

발생지점의 위치를 정확히 시각화하기 위해 train 데이터에서 x, y좌표를 별도로 coords 데이터를 생성. x, y 데이터의 결측값은 python google 지도 api 사용해 결측치 예측 후 대체.

```
from urllib.parse import quote
from urllib.request import Request, urlopen
import ssl
import json

location_url = quote('충청남도 천안시 동남구 청당동(청룡동)718')
api_key = "API_key"

url = 'https://maps.googleapis.com/maps/api/geocode/json?address='+ location_url +'&key=' + API_key + '&language=ko&regio
req = Request(url, headers={ 'X-Mashape-Key': api_key})
```

(1) 발생위치 분석

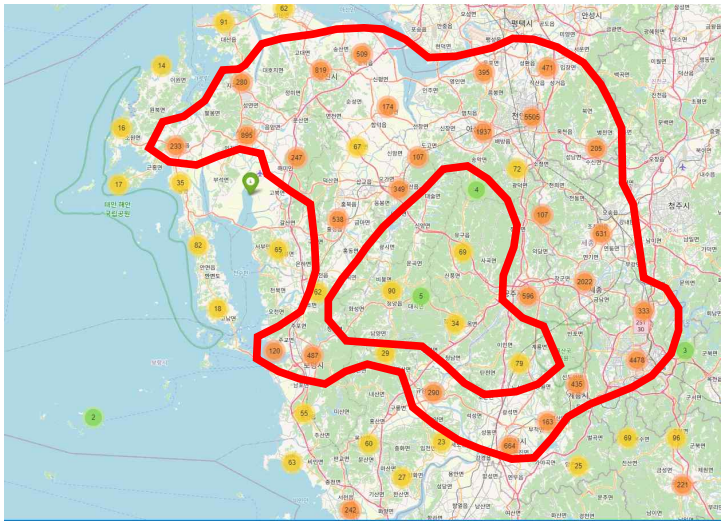
발생 위치를 직관적으로 보기 위해 파이썬에서 제공하는 folium 라이브러리를 사용하여 EDA를 진행.

```
from folium.plugins import MarkerCluster



m = folium.Map(
    location=[latitude, longitude],
    zoom_start=15
)

coords = coords.dropna()

marker_cluster = MarkerCluster().add_to(m)

for lat, long in zip(coords['HPPN_Y'], coords['HPPN_X']):
    folium.Marker([lat, long], icon = folium.Icon(color="green")).add_to(marker_cluster)
m
```

세부주소까지 위치를 보기엔 불필요한 소요가 예상되어 직관적인 시각화를 위해 세부주소 제외, 읍면동 기준으로 나눈 [‘location’] 변수 생성(python: 문자열 가공) (ex. 충청남도 보령시 청라면 나원리(행정:청라면) 910-7 → 충청남도 보령시 청라면)

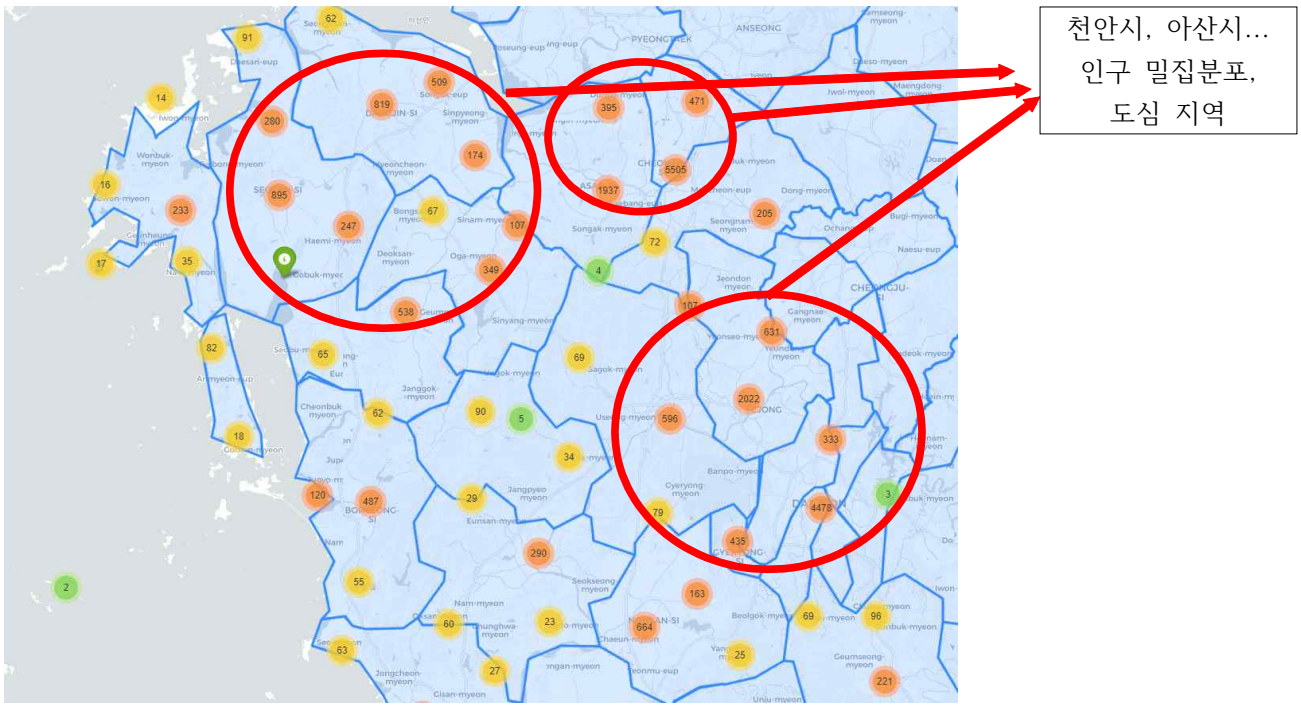
```
test_list = list(df_train['HPPN_PNU_ADDR'])
result_list = []

for test in test_list:
    temp = str(test).split()
    temp.pop()
    last = ' '.join(temp)
    result_list.append(last)

df_train['location'] = result_list
df_train.head(5)
# 상세주소 제거
```

필자는 “비교적 고령화의 인구가 많은 산간 지역 주변, 외부와 교류가 없는 지역이 보이스 피싱이 많이 발생할 것이다. ”라는 가설을 세우고 분석.



시각화를 해보니, 유동인구가 많은 지역에서 많이 발생.

좀 더 직관적으로 발생지역 확인을 위해, 시군구 읍면동을 기준으로 위치 좌표를 구분해놓은 korea_geo.json 파일과 연동시켜 행정구역 divide.



(2) 결론

이로서, "특정 지역(산간지역, 외부와 별로 교류가 없는 지역, 고령화 인구가 밀집한 지역)이 보이스피싱 발생확률이 더 높다"라는 가설은 설득력이 없고, 지역이 어디든 유동인구가 많고, 도심 지역이 많이 발생한다는 insight를 도출.

3.3 각 컬럼별 데이터 분석

(1) 동일사건 재발생률 분석

```
df_train['SME_EVT_YN'].notnull().sum()
# 동일사건 재발생 사건의 수 5,360건
# 동일사건 발생 확률 15.1%
✓ 0.0s
5360
```

다시 train 데이터로 돌아와, 동일사건이 다시 발생한다면 ['SME_EVT_YN']컬럼에 'y'가 채워지고 발생하지 않는다면, Nan 값(결측값)이 채워짐. 동일 사건이 발생한 건 수를 확인해보니, 35,489건 중 5,360건이 확인. 따라서 동일사건이 발생할 확률은 15.1%로서 10명 중 1.5명꼴로 다시 보이스 피싱에 당한다고 볼 수 있음.

(2) 사건 발생 datetime 분석

```
df_train["year"] = df_train["RECV_CPLT_DM"].dt.year
df_train["month"] = df_train["RECV_CPLT_DM"].dt.month
df_train["day"] = df_train["RECV_CPLT_DM"].dt.day
df_train["hour"] = df_train["RECV_CPLT_DM"].dt.hour
df_train["minute"] = df_train["RECV_CPLT_DM"].dt.minute
df_train["second"] = df_train["RECV_CPLT_DM"].dt.second

df_train.shape

#RECV_CPLT_DM 변수를 편하게 알아보기 위해 년, 월, 일, 시, 분, 초 단위로 나눠주는 과정
```

['RECV_CPLT'] 접수완료일시 컬럼을 발생한 년, 월, 일, 시, 분, 초와 상관관계를 알아보기 위해 나눠주고, 각 새로운 컬럼들을 생성.

먼저 ['year'] 변수를 살펴보면, 년도와 상관없이 보이스피싱은 꾸준히 일어나고 있는 것으로 확인.

```
print('2018년도 보이스피싱 발생 건 수 : ' + str(sum(df_train['year'] == 2018)))
print('2019년도 보이스피싱 발생 건 수 : ' + str(sum(df_train['year'] == 2019)))
print('2020년도 보이스피싱 발생 건 수 : ' + str(sum(df_train['year'] == 2020)))
print('2021년도 보이스피싱 발생 건 수 : ' + str(sum(df_train['year'] == 2021)))
print('2022년도 보이스피싱 발생 건 수 : ' + str(sum(df_train['year'] == 2022)))
```

✓ 0.0s

```
2018년도 보이스피싱 발생 건 수 : 1205
2019년도 보이스피싱 발생 건 수 : 1118
2020년도 보이스피싱 발생 건 수 : 1059
2021년도 보이스피싱 발생 건 수 : 1279
2022년도 보이스피싱 발생 건 수 : 1234
```

['month'] 변수. 대체로 보이스피싱은 연말(12월) 연초(1월) 순으로 가장 많이 발생하고, 상반기 3월, 4월 순으로 많이 발생.

```
df_train['month'].value_counts()
# 12월(연말), 1월(연초), 3월, 4월 순으로 피해량이 많았다.
```

✓ 0.0s

```
12    4340
1     3530
3     3299
4     3104
5     3095
11    2892
8     2874
7     2663
6     2576
10    2542
2     2335
9     2239
Name: month, dtype: int64
```

['hour'] 변수([minute], [second] 지엽적 변수 배제). 주로 보이스피싱이 일어나는 시간은, 오후시간 11 ~ 17시. 즉 바쁘고, 활동량이 많은 시간에 많이 일어난다는 insigh 도출.

```
df_train['hour'].value_counts()

#점심시간부터 오후시간 발생률이 높음
```

✓ 0.1s

```
11    3816
12    3784
13    3752
14    3639
10    3435
15    3373
16    2711
17    2347
9     2115
18    1881
19    1268
20     900
21     596
8      495
22     452
23     287
7      147
0      145
1       93
6       70
2       66
5       47
3       41
4       29
Name: hour, dtype: int64
```

(3) 발생 요일 분석

```
df_train['dayofweek'].value_counts()

# 0 = 월요일 ~ 6 = 일요일
# 큰 차이는 안보이지만 4, 3, 6(금, 목, 일)의 발생량이 상대적으로 높음.
```

✓ 0.0s

```
4    5292
3    5132
6    5125
5    5096
0    5019
2    5018
1    4807
Name: dayofweek, dtype: int64
```

0 ~ 6까지의 범주의 개수, 0 = 월요일 ~ 6 = 일요일. 큰 차이는 없지만, 금요일, 목요일, 일요일 순으로 발생량이 좀 더 많은 것으로 보아 일주일 후반부에, 비교적 경계가 느슨해질 때 많이 발생할 것이라 예측.

(4) 발생 성별 분석

```
df_train['RPTER_SEX'].value_counts()

#남자가 더 많이 보이스피싱 더 발생
```

✓ 0.0s

```
1.0    16839
2.0    14024
3.0     4626
Name: RPTER_SEX, dtype: int64
```

신원 미상 4,626건을 제외하면, 남자(55%)가 여자(45%)보다 더 많이 보이스피싱을 당함.

3.4 데이터 피해 현황 분석

(1) 보이스피싱 발생 종류 분석

구분	기관사칭형(A)		대출사기형(B)		A+B	
	발생건수	피해금액 (억)	발생건수	피해금액 (억)	발생건수	피해금액 (억)
2016	3,384	541	13,656	927	17,040	1,468
2017	5,685	267	18,574	1,503	24,259	2,470
2018	6,221	1,430	27,911	2,610	34,132	4,040
2019	7,219	2,506	30,448	3,892	37,667	6,398
2020	5,006	1,492	16,008	3,036	21,014	4,528
소계	27,515	6,936	106,597	11,968	134,112	18,904

(자료: 공공데이터 포털)

국내에서는 보이스피싱 관련 제도적 보안과 예방 교육, 노력에도 불구, 보이스피싱은 감소하지 않는 추세. 이에따라 대출 사기를 사칭하는 보이스피싱의 발생률이 더 높으며 이에 대한 대응 방안이 필요.

(2) 보이스피싱 발생 연령 분석

```
df_train
```

✓ 0.0s

	구분	합계	20대이하	30대	40대	50대	60대	70대이상
0	2016년	17040	3209	3735	4542	3834	1261	459
1	2017년	24259	5273	4887	6473	5412	1807	407
2	2018년	34132	4480	6483	9842	9313	3389	625
3	2019년	37667	3855	6041	10264	11825	4617	1065
4	2020년	31681	5323	4406	7704	9217	4188	843

(자료: 공공데이터 포털)

필자는 “나이가 많을수록 더욱 보이스피싱에 취약하다”는 가설을 세우고 공공데이터에서 데이터를 수집, 연관관계 분석. 오히려 4, 50대의 발생률이 현저히 높은 것을 파악할 수 있었음.

4. 분석결과 요약

● 위치기반 분석은 발생 지역 外(folium 라이브러리, 구글 api, request, pandas... 활용) 사각지대 추출, 주변 환경 분석 등을 수행하였으며, 분석결과 도심과 고립되어 있어 인구가 적고, 고령화의 인구가 많은, 산간지역은 오히려 낮은 보이스피싱 발생량을 보였고, 인구 밀집, 도심지역이 높은 사건 발생량을 보임.

● 컬럼기반 분석은 사건 발생 접수 일자를 기반으로, 년/월/일/시간대/요일과 사건발생과의 상관관계를 분석. 연도에따른 사건 발생량은 관계없이 일어나고 있지만, 월/시간대/요일에서 유의미한 insight 도출. 대체로 보이스피싱은 연말(12월), 연초(1월) 즉, 경제활동과 은행거래량이 많은 시기에 가장 많이 발생이 되었고, 다음으로 3, 4, 5월...순으로 상반기가 하반기보다 높은 발생량을 보임. 시간과의 상관관계는 바쁘고, 비교적 활동량이 많은 오후 시간대(11, 12, 13, 14시...)에 집중적으로 많이 일어난다는 결과 확인. 또한, 큰 차이는 안 보이지만 일주일의 전반부보다는, 경계가 느슨해진 금요일, 목요일, 일요일 순으로 후반부에 많이 일어남.

● 보이스피싱 피해자의 나이 / 사건의 유형 분석을 통해 “나이가 많을수록 피해 확률이 높다”는 가설은 무의미하고, 주로 4, 50대의 발생률이 높았으며 기관의 사칭보다는 대출을 빙자하는 사기가 사건 발생 확률이 더 높다고 확인됨.

5. 결과 해석 및 시사점

본 공모전에서는 경찰청에서 제공한 보이스피싱 발생 현황 데이터와 보이스피싱 관련 외부 데이터들을 가공, 분석하여 유의미한 데이터를 도출함. 이하 내용에서는 도출한 데이터를 바탕으로 결과 해석과 보이스피싱 피해 감소를 위한 방안을 제시.

첫째, 보이스피싱 피해 발생 지역을 읍면동 기준으로 나눔. 발생 지역 현황을 지도에 표현하여 가시성은 좋아졌지만, 주로 전화 또는 메세지, SNS를 통해 범죄가 일어나는 보이스피싱의 특성상 발생 지역 현황을 통해 피해 지역을 예측하기에는 어려움이 있음. 피해 발생 지점을 예측하기 보다 전자금융거래 시 단말기 정보와 접속 정보, 거래 정보 등을 수집 및 분석하여 이상 금융거래를 차단하는 기술인 FDS(이상거래탐지시스템)의 알고리즘 모델 업데이트를 통해 금융기관과의 긴밀한 협력으로 보이스피싱을 예방하는 것이 필요.

둘째, 동일사건 코드를 이용해 동일인의 보이스피싱 재발생율이 15.2%인 것을 확인.

이는 10명중 1.5명꼴로 같은 범죄에 당한다는 의미를 갖고 있음. 진화하는 보이스피싱의 방식으로 인해 피해자가 보이스피싱을 인지하고 식별하는데 어려움을 갖는 것으로 보임.
진화하는 방식을 모두 데이터베이스에 적재, 관리를 통해 기존 데이터를 분석하고 인공지능으로 새로운 유형을 예측해 그에 따른 예방책 전파 요구.

셋째, 월별 피해 발생량 분석 결과 연말(12월) 연초(1월)에 발생량이 높은 사실을 확인.
보이스피싱의 유형별 비율을 살펴보면 대출사기, 경찰, 검찰 등 국가기관 사칭, 가족납치유형, 기타 순으로 나타남. 12월은 대부분 연말정산 등 공적으로 처리해야 할 일이 많고, 1월에는 설날이 껴있어 금융거래가 활발해짐과 동시에 보이스 피싱도 연말연초에 성행한다고 생각됨.
이에 따라 연말연초에 보이스피싱 피해예방을 위해 예방법 적극적 홍보와 금융기관의 이상거래 집중 감시 등 고도의 관리 요구.

넷째, 보이스피싱 발생 시각에 따른 분석, 발생 요일별 분석, 성별 분석 진행.
발생 시각 분석을 보면 보이스피싱은 대부분 10~17시 오후시간의 피해가 많은 것으로 확인. 이는 대부분의 사람들이 활발한 경제활동을 하는 시간대에 피해가 주로 일어남. 발생 요일별 분석으로는 큰 차이는 없지만, 경계가 느슨해진 일주일 후반부에 주로 일어난다는 것을 확인. 성별 분석으로 신원미상 4,626건을 제외하고 백분율로 계산하였을 때, 남자(55%)가 여자(45%)보다 10%정도 더 피해가 많은 사실이 확인됨.

다섯째, 연령대별 보이스피싱 피해 발생량을 분석. 나이가 많을 수록 보이스피싱에 취약할 것이라고 가설을 설정하여 분석하였지만, 오히려 40,50대의 피해 발생량이 눈에 띄게 많았음. 연령대 별로 주 피해 유형을 살펴보면, 20대는 범죄 연루를 빙자한 국가기관 사칭사기, 30, 40대는 저금리 대출을 빙자한 금융기관 사칭사기, 50대 이상은 가족, 지인 사칭 사기에 주로 피해를 입는 것으로 확인. 더 효율적인 예방책 홍보 효과를 위해 연령대별 주요 사용 매체를 통해 홍보가 요구. 20, 30대에겐 유튜브 광고와 예방 컨텐츠 제작을 통해 경각심을 전파하고, 40, 50대 이상에겐 TV광고와 보이스피싱 주의 SNS 게시물 제작 등 보다 적극적인 예방책 홍보를 통해 보이스피싱 피해 감소를 기대. 보이스피싱은 경찰청 뿐만이 아닌, 금융기관, 통신기관, 정부 등 여러 기관들의 긴밀한 협력을 통해야만 근절할 수 있을것이라 생각됨. 본 공모전에서 분석하여 도출한 시사점을 바탕으로 적극적 예방을 통해 보이스피싱 감소의 기여했으면 함.

Ⅳ. 기타

○ 건의 사항

- 보이스피싱에는 범행 수법에 따라 공무원 사칭, 대출사기, 지인사칭 등 여러가지 종류가 있음. 하지만 이번 경진대회에서 경찰청에서 제공한 데이터에는 보이스피싱의 종류가 세분화 되어있지 않았음. 세분화 하여 제공하게 된다면 연령대, 성별에 따른 취약 유형을 파악하기 더 용이할 것이라 기대됨.

- 예측하는 데에 있어서 타겟변수가 명확했으면 함. [2024년의 충남, 대전, 세종 지역의 발생량 회귀 예측] 이런식으로 예측 ai 모델을 만들 수 있도록 데이터셋을 제공할 때 train용 데이터(발생량이 있는 데이터)와 test용 데이터(예측할 발생량을 비워둔 데이터)를 제공해주었으면 함.

○ 활용 데이터 및 참고 문헌 출처

신상철 2018. 대만 보이스피싱 범죄조직 국내 활동 분석, 아시아연구, 21(3), 151-191.
이승용 2020. 빅데이터와 FDS를 활용한 보이스피싱 피해 예측 방법 연구, 시큐리티연구 = Korean security journal no.62, pp.185 - 203
김민정, 김은미 보이스피싱 피해 경험 및 영향요인 분석. 소비자문제연구, 52(1), 53-72.
한겨레 보이스피싱 '20대 이하는 검찰, 30·40대는 금융사, 50대 이상은 가족 사칭에 취약'
동아일보(2021.07.01). 보이스피싱, 연령대별 '약한 고리' 파고든다
행정안전부 [빅데이터 분석으로 보이스피싱 막고, 불합리한 법령 정비한다]
공공데이터 www.data.go.kr