



## 6조 (final) - 데이터로 막는 다음 테러

👤멋쟁이사자처럼 노션 : <https://likelion.notion.site/Mid-Final-PJT-1b544860a4f480d9b760c2b0f755e33a>



### 목차

주제

데이터

변수

135개 모든 변수

주요 변수 18

**attacktype1\_txt**

**weaptype1\_txt**

region

파생 변수

테러의 정의

**success** 의 정의

가설 설정

선정 이유 (문제 정의)

데이터 전처리

1. 연도 필터링 (1990~)

2. 주요 변수 추출

3. 결측치 처리

4. 이상치 파악

수행 작업

피드백

분석(예측) 방향 아이디어

우리의 식스센스로 해당 나라의 테러성공을 예측해서 테러를 미리 대비한다!!

**3/31 아이디어**

**4/1 아이디어**

4/3 회의록

4/4 회의록 - 변수 선택 & 지역 선택 및 모델링 방안 논의

비교 분석 & 프로젝트 방향

머신러닝 파이프라인

1. 인코딩 방식

2. 불균형 처리 기법

3. 학습 모델 종류

4. 성능 평가 지표

참고 자료

프로젝트 타임라인 (전체 노션과 동일)



### 주제

데이터	설명
글로벌 테러리즘 데이터베이스 (Global Terrorism Database, GTD)	1970년부터 2020년까지 전 세계에서 발생한 <b>테러 사건</b> 정보를 망라한 대규모 공개 데이터베이스입니다 ( <a href="#">Global Terrorism Database - Web Resource - Catalog</a> ). 테러 발생 일자, 국가/지역, 공격 수단, 사상자 수, 배후 단체 등 상세한 필드가 포함된 거대한 CSV 데이터셋으로, 18만 건이 넘는 사건 기록을 담고 있습니다. <b>분석 아이디어: 연도별 테러 발생 추이와 지역별 분포</b> 를 분석하여 시대에 따른 테러 양상의 변화를 살펴볼 수 있습니다. 예를 들어 특정 기간에 테러 사건이 급증하거나 감소한 패턴을 찾고 그 사회적 배경을 해석하거

나,  
공격 유형이나 표적 유형에 따라 군집화를 시도해 보는 등의  
심층 분석 프로젝트를 진행할 수 있습니다.



## 데이터

### ▼ 데이터

6조(final) 공유폴더

### ▼ DATA URL

<https://www.start.umd.edu/download-global-terrorism-database>

### ▼ 사용한 데이터

- 사이트에서 제공하는 1970~2020년 + 2021년 데이터를 세로 병합한 후, 1990년부터 시작하는 필터를 적용한 **gtd\_1990-2021 (1).xlsx**를 사용



## 변수

### ▼ 135개 모든 변수

#### ① 사건 식별 및 날짜 정보

- **eventid**: 사건 고유 식별자 (연도와 일련번호 포함)
- **iyear, imonth, iday**: 사건이 발생한 연도, 월, 일
- **approxdate**: 정확한 날짜가 아닌 경우의 근사 날짜
- **extended**: 사건 기간이 확장되었는지 여부 (예: 단일 일 vs. 다일 사건)
- **resolution**: 사건 해결 상태나 결과에 대한 정보

#### ② 지역 및 위치 정보

- **country, country\_txt**: 국가 코드와 국가명
- **region, region\_txt**: 지역 코드와 지역명 (예: 중동, 아시아 등)
- **provstate**: 주 또는 도 단위 지역
- **city**: 도시명
- **latitude, longitude**: 사건 발생 위치의 위도와 경도
- **specificity, vicinity, location**: 사건 위치의 상세도, 인접지역 정보 등

#### ③ 사건 특성 및 분류 기준

- **summary**: 사건에 대한 간단한 요약
- **crit1, crit2, crit3**: 테러 사건으로 분류하기 위한 기준 값들
- **doubtterr**: 사건이 테러인지에 대한 의심 여부
- **alternative, alternative\_txt**: 대체 분류나 추가 설명
- **multiple**: 단일 사건인지, 복합 사건인지를 나타냄

#### ④ 공격 형태 관련 정보

- **success**: 공격 성공 여부 (예: 목표 달성 여부)
- **suicide**: 자살 공격 여부
- **attacktype1, attacktype1\_txt**: 주요 공격 방식 (예: 폭발, 총격 등) 및 그 설명

- **attacktype2, attacktype2\_txt / attacktype3, attacktype3\_txt:** 추가 공격 방식이 있을 경우의 정보

## ⑤ 목표 대상 정보

- **targetype1, targetype1\_txt:** 주요 표적 유형 (예: 정부기관, 민간인 등)과 설명
- **targsubtype1, targsubtype1\_txt:** 표적의 세부 유형
- **corp1, target1:** 표적이 속한 조직 및 구체적 표적 정보
- **natlty1, natlty1\_txt:** 표적의 국적
- (targetype2, targsubtype2, corp2, target2, natlty2 등 유사한 변수들이 2차, 3차 표적에 대해 존재)

## ⑥ 가해자(테러 집단) 정보

- **gname, gsubname:** 주 가해자(테러리스트 집단)와 그 하위 집단 정보
- **gname2, gsubname2 / gname3, gsubname3:** 추가 가해자 정보 (여러 집단 관련 시)
- **motive:** 공격의 동기나 의도
- **guncertain1, guncertain2, guncertain3:** 가해자 정보의 불확실성 여부
- **individual:** 개인 단독 공격 여부
- **nperps, nperpcap:** 가해자 수와 체포된 인원 수
- **claimed, claimmode, claimmode\_txt:** 공격에 대한 책임 주장 여부 및 주장 방식
- (claim2, claimmode2, claimmode2\_txt, claim3, claimmode3, claimmode3\_txt, compclaim 등 추가 주장 관련 변수)

## ⑦ 사용 무기 정보

- **weaptype1, weaptype1\_txt:** 사용된 주요 무기 유형과 그 설명 (예: 화약물, 총기 등)
- **weapsubtype1, weapsubtype1\_txt:** 무기 세부 유형
- **weaptype2, weaptype4, weapsubtype2, weapsubtype4:** 추가 무기 정보(최대 4종까지)

## ⑧ 인명 피해 정보

- **nkill, nkillus, nkillter:** 사망자 수 (피해자, 미국 관련, 가해자 구분)
- **nwound, nwoundus, nwoundte:** 부상자 수 (피해자, 미국 관련, 가해자 구분)

## ⑨ 재산 피해 정보

- **property:** 재산 피해 발생 여부
- **propextent, propextent\_txt:** 재산 피해 정도 및 서술형 설명
- **propvalue:** 재산 피해 금액
- **propcomment:** 재산 피해에 대한 추가 설명

## ⑩ 인질 및 납치 관련 정보

- **ishostkid:** 인질 또는 납치 여부
- **nhostkid, nhostkidus:** 인질 또는 납치된 인원 수 (전체, 미국 관련 등)
- **nhours, ndays:** 인질 상황 지속 시간 (시간, 일)
- **divert:** 사건 후 상황 전환 여부
- **kidhijcountry:** 인질 또는 납치 사건이 발생한 국가
- **ransom, ransomamt, ransomamtus, ransompaid, ransompaidus, ransomnote:** 몸값 요구 및 지급 관련 정보

## ⑪ 인질 사건 결과 및 기타 정보

- **hostkidoutcome, hostkidoutcome\_txt:** 인질 상황의 결과와 그에 따른 설명
- **nreleased:** 석방된 인질 수
- **addnotes:** 추가적인 설명이나 특이사항

- **scite1, scite2, scite3**: 사건 관련 참고 기사 출처
- **dbsource**: 데이터 출처
- **INT\_LOG, INT\_IDEO, INT\_MISC, INT\_ANY**: 국제 테러와 관련된 추가 지표
- **related**: 연관된 사건들의 eventid (콤마로 구분)

#### ▼ 주요 변수 18

변수	data	설명 (codebook)
사건 번호	<b>eventid</b>	- 사건 식별 번호
연도, 월, 일	<b>year, imonth, iday</b>	<b>[iday]</b> - 사건이 발생한 월(month) 중 정확한 일(day)을 숫자로 표현 사건이 여러 날에 걸쳐 발생한 경우, iday에는 사건이 시작된 날짜의 일(day)이 기록 - 1970년부터 2011년 사이에 발생한 공격 중 정확한 날짜를 알 수 없는 경우, <b>0</b> 으로 기록 - 2011년 이후 발생한 공격의 경우 정확한 날짜를 알 수 없을 때, 이 필드에는 출처 자료에서 보고된 가능한 날짜 범위의 중간값이 기록
성공 여부	<b>success</b>	- 0(실패), 1(성공)으로 테러의 성공 여부를 판단 - 구체적인 정의는 <테러 성공(success)변수의 정의> 참고
사건 발생 국가명	<b>country_txt</b>	- 테러가 발생한 국가
사건 발생 지역	<b>region_txt</b>	- 테러가 발생한 지역을 식별 - 테러에 대해 코딩된 국가에 따라, 지역은 12개 범주로 나누어진다.
국가 내 행정 구역	<b>provstate</b>	- Province / Administrative Region / State - 사건이 발생한 시점 기준으로, 해당 사건이 일어난 <b>1차 하위 행정 구역(광역 단위: 주, 도, 특별시 등)</b> 의 이름을 기록한 것
사건이 발생한 구체적인 도시	<b>city</b>	- 테러가 발생한 도시, 마을, 또는 촌락의 이름을 포함 - 만약 해당 정보가 알려지지 않은 경우에는, 사건과 관련하여 찾을 수 있는 provstate(1차 행정구역) 아래의 가장 작은 행정 구역(예: 구역)의 이름이 기록
위도, 경도	<b>latitude, longitude</b>	테러가 발생한 도시의 위도와 경도를 기록
공격 유형	<b>attacktype1_txt</b>	- 공격의 일반적인 방법을 나타내며, 사용된 전술의 대분류를 반영 - 예를 들어, <b>폭발물을 이용한 암살</b> 이 발생한 경우에는, 폭파/폭발(Bombing/Explosion)이 아닌 <b>암살(Assassination)</b> 으로 분류
주요 공격 대상 유형	<b>targettype1_txt</b>	- 사건의 공격 대상 또는 피해자의 일반적인 유형을 기록 - 피해자가 특정 인물(예: 고위 인사)과의 관계 때문에 공격당한 경우, 해당 관계에 기반한 동기가 반영되어 타겟 유형이 결정 - 예를 들어, <b>어떤 정부 고위 관료의 가족</b> 이 그 인물과의 관계 때문에 공격당한 경우, 타겟 유형은 <b>정부(Government)</b> 로 분류
피해자의 국적	<b>natlity1_txt</b>	- <b>공격 대상이 된 사람 또는 대상의 국적</b> - 이는 사건이 발생한 국가와 반드시 일치하지는 않으며, 대부분의 경우에는 같지만 예외도 존재한다. - 예를 들어, 어떤 국가에서 발생한 사건이라 하더라도, 그 <b>공격 대상이 외국 국적을 가진 개인 또는 단체</b> 일 수 있다. - 비행기 납치 사건(hijacking)의 경우, 탑승객의 국적이 아니라 비행기의 등록 국적이 기록
무기 유형	<b>weaptype1_txt</b>	- 사건에 사용된 무기의 일반적인 유형을 뜻함
사망자 수	<b>nkill</b>	- 총 사망자 수 (Total Number of Fatalities) - 해당 사건에서 직접적으로 사망한 피해자와 공격자(가해자)를 모두 포함한 총 확정 사망자 수를 기록

부상자 수	nwound	- 총 부상자 수 - 피해자와 가해자 모두를 포함한 사건으로 인한 확정된 비치명적 부상자 수를 기록합니다.
테러 단체 이름	gname	- Perpetrator Group Name (가해자 그룹 이름) - 공격을 수행한 그룹의 이름 - 출처 자료에 정식 가해자 그룹 또는 조직의 이름이 보고되지 않은 경우, 이 필드에는 가해자의 일반적인 정체성을 나타내는 정보(예: "개신교 극단주의자")가 포함될 수 있다. - 이 범주들은 개별적인 독립체를 나타내지 않으며, 서로 배타적이지도 않다(예: "학생 급진주의자"와 "좌파 무장단체"가 동일한 사람들을 설명할 수 있음).

#### ▼ attacktype1\_txt

data	변수	설명
Bombing/Explosion	폭탄/폭발 공격	폭발물을 설치하거나 투척하여 발생시키는 공격
Armed Assault	무장 습격	총기나 무기 등을 사용한 직접 공격
Assassination	암살	특정 인물을 표적으로 한 계획적 살해
Facility/Infrastructure Attack	시설/인프라 공격	전력 설비, 통신기지, 교통 인프라 등 시설이나 시스템을 목표로 한 파괴 공격
Hostage Taking (Kidnapping)	인질 납치	특정 인물을 몰래 납치하여 협상 수단으로 활용
Unknown	미상	공격 방식이 기록되지 않았거나 명확하지 않은 경우
Unarmed Assault	비무장 공격	무기 없이 신체적 폭력(구타, 질식 등)을 동반한 공격
Hijacking	탈취	차량, 선박, 항공기 등 이동 수단을 강제로 점거하거나 이동 경로를 바꾸는 행위
Hostage Taking (Barricade Incident)	인질극 (바리케이드 사건)	범인이 건물 등에 진입해 인질을 잡고 농성하며 경찰과 대치하는 상황




#### ▼ weaptype1\_txt

data	변수	설명
Explosives/Bombs/Dynamite	폭발물 / 폭탄 / 다이내마이트	폭발물, 수류탄, 다이내마이트 등을 사용한 공격. 차량 폭탄, 자살폭탄 등도 이 범주에 포함
Firearms	총기류	총기류(권총, 소총 등)를 사용한 테러 습격, 암살, 총격전 등이 여기에 해당
Unknown	미상	사용된 무기가 기록되지 않았거나 분류 불가능한 경우
Incendiary	방화 무기	방화 공격을 의미 화염병, 가연성 액체, 불붙은 장치로 건물·차량을 불태우는 공격
Melee	근접 무기	근접 무기(칼, 도끼 등)를 사용한 공격
Chemical	화학 무기	유독 화학 물질 사용 예: 독가스, 염산, 기타 화학 약품을 통한 공격
Vehicle (not to include vehicle-borne explosives)	차량 (폭탄 제외)	차량 자체를 무기로 사용한 돌진 공격 예: 군중 속 차량 돌진. 차량에 폭탄이 설치된 경우는 여기 포함되지 않음
Sabotage Equipment	시설 파괴 장비	시설 파괴용 도구 또는 기계 장비 예: 철도 레일 파괴, 발전기 손상 등 직접적 공격보다는 시스템 교란 목적
Other	기타	위에 분류되지 않는 기타 무기 사용
Fake Weapons	모조 무기	예: 장난감 총, 폭탄처럼 보이게 만든 가짜 장치 등 위협 효과를 의도한 경우
Biological	생물학 무기	바이러스, 세균, 병원균 등을 이용한 생물학적 테러 예: 탄저균, 바이러스 살포 등
Radiological	방사능 무기	방사능 물질을 사용한 공격






#### ▼ region

#### GTD 기준 12개 지역 설명





## 1. North America

-  미국 (USA),  캐나다,  멕시코
- **특징:** 고소득국 중심, 총기 테러와 차량 테러 등 비교적 선진국형 테러 양상
- **비고:** 멕시코는 지리상 중남미지만, 문화·경제적으로 북미로 분류됨






## 2. Central America & Caribbean

-  과테말라,  온두라스,  엘살바도르,  쿠바,  자메이카 등
- **특징:** 마약 카르텔, 조직범죄 기반 테러 많음
- **비고:** 중남미 일부와 해양국가 포함됨





## 3. South America

-  브라질,  콜롬비아,  아르헨티나,  페루 등
- **특징:** 게릴라/반정부 단체 활동, FARC 등
- **비고:** 중남미의 남부 국가들






## 4. Western Europe

-  프랑스,  독일,  영국,  이탈리아,  스페인 등
- **특징:** 이슬람 극단주의 테러 + 정치적 시위형 테러 공존
- **비고:** 유럽연합 중심의 선진국 집단





## 5. Eastern Europe

-  폴란드,  러시아,  우크라이나,  루마니아 등
- **특징:** 민족 갈등/분리주의, 체첸 문제 등
- **비고:** 과거 공산권 영향 받은 국가들




## 6. Middle East & North Africa (MENA)

-  이집트,  사우디,  시리아,  이라크,  리비아 등
- **특징:** 종교/정치 이슈 중심, 극단주의, IS, 알카에다 등 활발
- **비고:** 아랍어권 + 이슬람국가가 주류






## 7. Sub-Saharan Africa

-  나이지리아,  소말리아,  말리,  케냐 등
- **특징:** 부족 갈등, IS/알샤바브 활동, 교육·종교 문제
- **비고:** 사하라 이남 아프리카 국가 포함

## 8. Central Asia

-  카자흐스탄,  우즈베키스탄,  타지키스탄 등
- **특징:** 구소련권, 이슬람권 기반, 분리주의 테러 가능성
- **비고:** 중앙 아시아 내륙국들로 구성

## 9. South Asia

-  인도,  파키스탄,  방글라데시,  네팔,  스리랑카 등
- **특징:** 종교 분쟁 (힌두교-이슬람), 분리주의 테러 활발
- **비고:** 인도 중심의 테러 핫스팟

## 10. Southeast Asia

- 🇵🇭 필리핀, 🇮🇩 인도네시아, 🇹🇷 태국, 🇲🇾 말레이시아 등
- 특징: 이슬람 극단주의, 지역 분리주의, 공산반군 등 혼재
- 비교: 섬나라 포함, 마르크스주의 성향도 일부 존재

## 11. East Asia

- 🇨🇳 중국, 🇰🇷 한국, 🇯🇵 일본, 🇲🇻 몽골 등
- 특징: 전반적으로 테러 발생 낮음, 하지만 소수민족 문제(위구르 등)는 존재
- 비교: 경제 선진국 다수 포함, 북한은 종종 예외 처리됨

## 12. Australasia & Oceania

- 🇺🇸 호주, 🇳🇿 뉴질랜드, 태평양 도서국 (피지, 사모아 등)
- 특징: 테러 발생 적음, 서구권과의 문화적 유사성
- 비교: 지리적으로는 멀리 떨어져지만, 서구 선진국 분류에 포함됨

### ▼ 파생 변수

1. total\_victim (피해자 수)

```
df['total_victim'] = df['nkill'] + df['nwound']
```

2. weekday (요일)

```
# 요일 파생변수 생성
iraq_df["date"] = pd.to_datetime(dict(year=iraq_df.iyear, month=iraq_df.imonth, day=iraq_df.iday), errors='coerce')
iraq_df["weekday"] = iraq_df["date"].dt.day_name()
iraq_df['weekday']

# 요일 변수 결측치 존재! (정확한 요일을 모르는 경우 0으로 표시되었기 때문에 결측치로 나타남)
print(iraq_df[iraq_df['weekday'].isna()])

# → weekday를 표시할 수 없으므로 unknown으로 대체
iraq_df['weekday'] = iraq_df['weekday'].fillna('Unknown')
```



## 테러의 정의

### 국가법령정보센터(2023.05.20. 검색) 국민 보호와 공공안전을 위한 테러방지법

“테러란 국가·지방자치단체 또는 외국 정부(외국 지방자치단체와 조약 또는 그 밖의 국제적인 협약에 따라 설립된 국제기구를 포함한다)의 권한 행사를 방해하거나 의무 없는 일을 하게 할 목적 또는 공중을 협박할 목적으로 하는 다양한 행위” (뉴노멀 시대의 테러 및 대체 논문)

### 테러의 사전적 정의

: 테러의 사전적 정의 : 폭력을 사용하여 상대를 위협하거나 공포에 빠뜨리게 하는 행위. 순화어는 ‘폭력’, ‘폭행’

### 테러 관련 논문

KBB SCHOLAR 국제+테러+추세에+따른+테러+예방방안+분석+다중이용시설을+중심으로.pdf

뉴노멀 시대의 테러 및 대테.pdf

[https://drive.google.com/file/d/1FV0X2MBhrzNkVWg94to2LsgXYLcojJmd/view?usp=drive\\_link](https://drive.google.com/file/d/1FV0X2MBhrzNkVWg94to2LsgXYLcojJmd/view?usp=drive_link)



## success 의 정의

GTD(Global Terrorism Database)에서 말하는 success 란?

success는 테러 공격이 의도된 계획대로 실행되었는가를 나타냄.

즉, 물리적으로 공격이 실행되었느냐, 사람을 죽이거나 다치게 했느냐와는 별개

예: 성공 (success=1)인데 사상자(nkill, nwound)=0인 경우:

- 폭발물이 터졌지만 아무도 근처에 없어서 피해가 없었음.
- 정부기관 건물 앞에 터뜨려서 위협 메시지를 전달하는 게 목적이었고, 실제로 메시지는 전달됨.
- 차량 폭탄이 터졌지만 시간 설정이 잘못되어 밤에 터져 아무도 다치지 않음. 하지만 계획된 장소에서 성공적으로 터짐.

→ 이런 경우는 사상자는 없지만, 폭발 테러가 '의도대로 실행'되었기 때문에 success=1이 될 수 있다.



## 가설 설정

### ▼ 문규빈

- 1990년대 이후 테러의 수는 점점 증가하고 있다.
- 테러는 지역별로 차이가 있고, 특히 중동/북아프리카, 남아시아와 같은 정치적,경제적 불안이 큰 지역에서 많이 발생한다.
- 과거 군사시설, 정부기관 등을 목표로 삼았으나 최근엔 민간인을 대상으로한 상업시설, 대중교통 시스템등을 목표로 한다.

### ▼ 이동주

- 시간의 흐름에 따라 테러의 성공률이 떨어질 것이다.
- 특정 조직의 테러의 사상자가 많을 것이다.

### ▼ 장선희

- 시간의 흐름에 따라 테러 무기와 유형, 장소가 변화할 것이다.

### ▼ 최혜은

- 중동 지역(Iraq, Pakistan 등)에서 테러 사건이 가장 많이 발생했을 것이다.
- 테러 피해(사망자 수)는 사용된 무기 유형과 밀접한 관련이 있을 것이다.





## 선정 이유 (문제 정의)

- 전 세계에서 발생한 테러 사건의 **패턴, 영향도, 지역적 분포, 공격 유형** 등을 분석하여 다양한 인사이트를 분석하고자 한다.
- 도출한 인사이트를 바탕으로 **테러 대응 능력을 향상시키고 미리 예측**하는 것을 목표로 한다.



## 데이터 전처리

### ▼ 1. 연도 필터링 (1990~)

- gtd\_1990-2021 로 저장 후 작업 시작!

```
# 'iyear'가 1990 이상인 데이터만 필터링
df_1990 = df[df['iyear'] >= 1990]
```

### ▼ 2. 주요 변수 추출

- 주요 변수 18개를 중심으로 분석!

```
# 추출할 변수 목록
columns_to_extract = [
    'eventid',
    'iyear', 'imonth', 'iday',
    'success',
    'country_txt', 'region_txt', 'provstate',
    'latitude', 'longitude',
    'attacktype1_txt', 'targettype1_txt', 'natlty1_txt', 'weaptype1_txt',
    'nkill', 'nwound',
    'gname'
]

# 해당 변수들만 추출
df[columns_to_extract]
```

**nkillter** , **nwoundte** 변수 자체를 주요 변수에 포함시키지 않음

- 0의 비율이 너무 많아 유의미한 결과를 도출하지 못할 것 같다고 판단하여 주요 변수에 포함 X

```
# 구간화
bins = [0, 1, 5, 10, 1000]
labels = ['0', '1-4', '5-9', '10+']
nkillter_bin = pd.cut(df_1990['nkillter'], bins=[-1]+bins[1:], labels=labels)
nwoundte_bin = pd.cut(df_1990['nwoundte'], bins=[-1]+bins[1:], labels=labels)

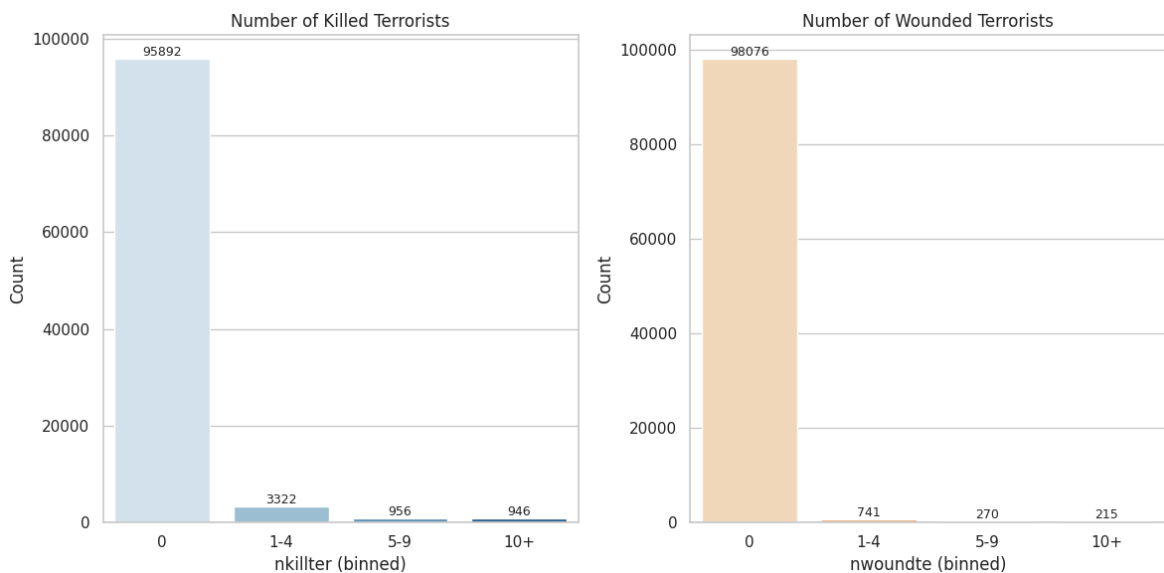
# 수치 표시 함수
def annotate_barplot(ax):
    for bar in ax.patches:
        height = bar.get_height()
        if height > 0:
            ax.annotate(f'{int(height)}', xy=(bar.get_x() + bar.get_width() / 2, height),
                        ha='center', va='bottom', fontsize=9)

# 시각화
```

```
plt.figure(figsize=(12, 6))

# nkillter 그래프
ax1 = plt.subplot(1, 2, 1)
sns.countplot(x=nkillter_bin, palette='Blues', ax=ax1)
ax1.set_title("Number of Killed Terrorists")
ax1.set_xlabel("nkillter (binned)")
ax1.set_ylabel("Count")
annotate_barplot(ax1)

# nwoundte 그래프
ax2 = plt.subplot(1, 2, 2)
sns.countplot(x=nwoundte_bin, palette='Oranges', ax=ax2)
ax2.set_title("Number of Wounded Terrorists")
ax2.set_xlabel("nwoundte (binned)")
ax2.set_ylabel("Count")
annotate_barplot(ax2)
plt.show()
```



### ▼ 3. 결측치 처리

#### ▼ 1. 주요 변수의 결측치 개수 & 비율 확인

```
# 결측치 건수 계산
null_counts = df[columns_to_extract].isnull().sum()

# 결측치 비율 계산 (백분율로)
null_ratios = (df[columns_to_extract].isnull().mean() * 100).round(2)

# 결과 데이터프레임 생성
null_summary = pd.DataFrame({
    'Null Count': null_counts,
    'Null Ratio (%)': null_ratios
})
```

	0
Unnamed: 0	0
eventid	0
year	0
imonth	0
iday	0
country_txt	0
region_txt	0
provstate	0
latitude	2531
longitude	2532
city	427
success	0
summary	26227
attacktype1_txt	0
targettype1_txt	0
weaptype1_txt	0
nkill	7270
nkillter	30061
nwound	13210
nwoundte	32645
natlty1_txt	1869
gname	0

1990~2021 데이터셋 활용시 결측치

- 결측치 처리가 필요한 변수
  - city, provstate, latitude, longitude (위치적 정보)
    - city, provstate에는 unknown이라는 정보가 존재
  - natlty1\_txt (가해자의 국적)
  - nkill, nwound (사망자 수, 부상자 수)

## ▼ 2. 위치정보 관련 변수 결측치 처리

(내부 데이터 참조 보간 → 외부 API보간 → 남아있는 결측값들은 모두 Unknown으로 채워서 시각화에서 활용)

### (1) city처리

#### ▼ 왜 city를 먼저 처리?

1. 상대적으로 적은 결측치
2. 상세한 주소이기 때문에 이 결측치를 채우면 provstate를 더 용이하게 채울 수 있음

#### • 내부 참조기반 보간

# 1. city가 결측 또는 'Unknown'이면서 위경도, provstate, country는 있는 행 필터링

```
mask_city_unknown = (
    ((df_filled['city'].isnull()) | (df_filled['city'] == 'Unknown')) &
    df_filled['provstate'].notnull() &
    df_filled['country_txt'].notnull() &
    df_filled['latitude'].notnull() &
    df_filled['longitude'].notnull()
)
df_city_target = df_filled[mask_city_unknown].copy()
print(f"📍 내부 참조 기반 city 보간 대상 행 수: {len(df_city_target)}")
```

# 2. city가 제대로 있는 참조 테이블 구성

```
df_city_reference = df_filled[
    df_filled['city'].notnull() &
    (df_filled['city'] != 'Unknown') &
    df_filled['provstate'].notnull() &
    df_filled['country_txt'].notnull() &
    df_filled['latitude'].notnull() &
    df_filled['longitude'].notnull()
][['country_txt', 'provstate', 'latitude', 'longitude', 'city']].drop_duplicates()
```

```

# 3. 좌표 + 지역 기준으로 병합 (city 보간용)
merged = pd.merge(
    df_city_target,
    df_city_reference,
    on=['country_txt', 'provstate', 'latitude', 'longitude'],
    how='left',
    suffixes=('_', '_imputed')
)

# 4. 보간 반영
to_update = merged['city_imputed'].notnull()
print(f"✅ 내부 참조 기반 city 보간 성공 건수: {to_update.sum()}")
df_filled.loc[merged[to_update].index, 'city'] = merged.loc[to_update, 'city_imputed'].values

# 5. 보간 후 Unknown/결측 상태 확인
city_unknown_after = ((df_filled['city'].isnull()) | (df_filled['city'] == 'Unknown')).sum()
print(f"📍 city 보간 후 Unknown/결측 수: {city_unknown_after}")

```

#### • 외부 API 활용 보간(역지오코딩)

```

import requests
import time

# 1. 역지오코딩 함수 정의 (영어 반환, city 우선 → town → village → municipality)
def reverse_geocode(lat, lon):
    try:
        url = f"https://nominatim.openstreetmap.org/reverse?lat={lat}&lon={lon}&format=json"
        headers = {
            "User-Agent": "geo-reverse-city-bot",
            "Accept-Language": "en" # 영어로 반환되게 설정
        }
        response = requests.get(url, headers=headers, timeout=10)
        if response.ok and response.json():
            address = response.json().get("address", {})
            city = address.get("city") or address.get("town") or address.get("village") or address.get("municipality")
            return city
        except Exception as e:
            print(f"❌ Error at ({lat}, {lon}) → {e}")
        return None

# 2. city가 Unknown/결측이고 위경도, provstate, country가 있는 행 필터링
mask_geo_city = (
    ((df_filled['city'].isnull()) | (df_filled['city'] == 'Unknown')) &
    df_filled['provstate'].notnull() &
    df_filled['country_txt'].notnull() &
    df_filled['latitude'].notnull() &
    df_filled['longitude'].notnull()
)

df_city_geo_target = df_filled[mask_geo_city].copy()
print(f"📍 역지오코딩 city 보간 대상 행 수: {len(df_city_geo_target)}")

# 3. 역지오코딩 실행
city_results = []
for idx, row in df_city_geo_target.iterrows():

```

```
lat, lon = row['latitude'], row['longitude']
city = reverse_geocode(lat, lon)
city_results.append(city)
print(f"[{idx}] ({lat}, {lon}) → {city}")
time.sleep(1) # 요청 제한 피하기 위해 슬립
```

#### # 4. 결과 반영

```
df_filled.loc[mask_geo_city, 'city'] = city_results
print("✅ 역지오코딩 기반 city 보간 완료!")
```

#### • 보간되지 않은 3424개의 행에 대해서 재 역지오코딩

```
# 경위도와 provstate가 모두 없는 행
group1 = df_city_missing[
    df_city_missing['provstate'].isnull() |
    df_city_missing['latitude'].isnull() |
    df_city_missing['longitude'].isnull()
]

# 경위도와 provstate가 모두 있는 행 (보간 가능한데 실패한 케이스)
group2 = df_city_missing[
    df_city_missing['provstate'].notnull() &
    df_city_missing['latitude'].notnull() &
    df_city_missing['longitude'].notnull()
]

print(f"🌿 그룹1 (불가능한 보간): {len(group1)}건" ) #RESULTS: 🌿 그룹1 (불가능한 보간): 1010건
print(f"🔍 그룹2 (보간 가능했는데 실패): {len(group2)}건" ) #results: 🔍 그룹2 (보간 가능했는데 실패): 2414건
```

#### • 확장 보간

```
import requests
import time

# 1. 역지오코딩 함수 (확장 필드까지 고려)
def reverse_geocode_fallback(lat, lon):
    try:
        url = f"https://nominatim.openstreetmap.org/reverse?lat={lat}&lon={lon}&format=json"
        headers = {
            "User-Agent": "geo-city-fallback-bot",
            "Accept-Language": "en"
        }
        response = requests.get(url, headers=headers, timeout=10)
        if response.ok and response.json():
            address = response.json().get("address", {})
            # city 우선, 없으면 town → village → municipality → county → region → state_district
            city = (
                address.get("city") or
                address.get("town") or
                address.get("village") or
                address.get("municipality") or
                address.get("county") or
                address.get("region") or
                address.get("state_district")
            )
    except:
```

```

        return city
    except Exception as e:
        print(f"❌ Error at ({lat}, {lon}) → {e}")
    return None

# 2. group2 추출: 보간 가능한데 실패했던 city들만
city_missing_mask = (df_filled['city'].isnull()) | (df_filled['city'] == 'Unknown')
group2_mask = (
    city_missing_mask &
    df_filled['provstate'].notnull() &
    df_filled['latitude'].notnull() &
    df_filled['longitude'].notnull()
)
df_group2 = df_filled[group2_mask].copy()
print(f"🔄 확장 역지오코딩 대상 행 수 (group2): {len(df_group2)}")

# 3. 보간 수행
city_results_fallback = []
for idx, row in df_group2.iterrows():
    lat, lon = row['latitude'], row['longitude']
    city = reverse_geocode_fallback(lat, lon)
    city_results_fallback.append(city)
    print(f"[{idx}] ({lat}, {lon}) → {city}")
    time.sleep(1) # 속도 제한 방지용

# 4. 결과 반영
df_filled.loc[group2_mask, 'city'] = city_results_fallback
print(f"✅ 확장형 역지오코딩 기반 city 보간 완료!")

```

📁 확장 보간 후 남은 city 결측/Unknown 수: 1921

## • provstate 처리

```

# 2. 보간 대상: provstate가 결측 또는 'Unknown'이면서 city/lat/lon 존재하는 행
mask_provstate_unknown = (
    ((df_filled['provstate'].isnull()) | (df_filled['provstate'] == 'Unknown')) &
    df_filled['city'].notnull() &
    df_filled['latitude'].notnull() &
    df_filled['longitude'].notnull()
)
df_prov_target = df_filled[mask_provstate_unknown].copy()
before_count = len(df_prov_target)

# 3. 참조 테이블: 신뢰 가능한 city + 위경도 조합에서 provstate가 존재하는 행만
df_prov_reference = df_filled[
    df_filled['provstate'].notnull() &
    (df_filled['provstate'] != 'Unknown') &
    df_filled['city'].notnull() &
    (df_filled['city'] != 'Unknown') &
    df_filled['latitude'].notnull() &
    df_filled['longitude'].notnull()
][['city', 'latitude', 'longitude', 'provstate']].drop_duplicates()

# 4. 병합 및 보간
merged = pd.merge(
    df_prov_target,
    df_prov_reference,

```

```

on=['city', 'latitude', 'longitude'],
how='left',
suffixes=('_', '_imputed')
)

# 5. 보간 성공 여부
to_update = merged['provstate_imputed'].notnull()
filled_count = to_update.sum()

# 6. 원본에 보간 결과 반영
df_filled.loc[merged[to_update].index, 'provstate'] = merged.loc[to_update, 'provstate_imputed'].values

# 7. 보간 후 결측/Unknown 상태 확인
after_count = ((df_filled['provstate'].isnull()) | (df_filled['provstate'] == 'Unknown')).sum()

# import ace_tools as tools; tools.display_dataframe_to_user(name="provstate 보간 결과", dataframe=merged[to
before_count, filled_count, after_count

```

보간 전 : 2497개 → 보간 후:2434

## (2) 위/경도값 결측치 보간

summary함수의 위치 전치사들을 정규표현식을 통해 감지해 그 뒤의 장소들을 뽑아내려고 함.

ex: at the ~church, in the xxx center

- 정규표현식을 활용한 장소 추출

```

import re

def extract_place(summary):
    if pd.isnull(summary):
        return None

    # ① 괄호 안 장소 추출: e.g. (Baghdad)
    match = re.search(r"((.*?)\)", summary)
    if match:
        return match.group(1).strip()

    # ② 'in XXXX' 패턴 추출
    match = re.search(r"\bin\s+([A-Z][a-zA-Z\s-]+)", summary)
    if match:
        return match.group(1).strip()

    # ③ 'at XXXX' 패턴 추출
    match = re.search(r"\bat\s+([A-Z][a-zA-Z\s-]+)", summary)
    if match:
        return match.group(1).strip()

    return None

```

- 보간 수행

```
import requests
```

```
def geocode_place(place_name):
    try:
        url = f"https://nominatim.openstreetmap.org/search?q={place_name}&format=json&limit=1"
        headers = {"User-Agent": "geo-latlon-bot"}
        response = requests.get(url, headers=headers, timeout=10)
        if response.ok and response.json():
            data = response.json()[0]
            return float(data['lat']), float(data['lon'])
    except Exception as e:
        print(f"❌ Geocode error for '{place_name}': {e}")
    return None, None
```

- 하지만 보간 실패
- 정규표현식에서 추출된 단어들을 살펴보니 테러조직 이름, 나라이름, 단체이름임  
즉, 위경도를 추측할 수 있도록 하는 단서가 되지 못함  
➡ 결측값들을 모두 unknown으로 변환

- 결측값 → unknown변환(위치변수 결측지 제거 완료 ✅)

```
# 오직 결측치(NaN)만 'Unknown'으로 대체
df_filled['provstate'] = df_filled['provstate'].fillna('Unknown')
df_filled['city'] = df_filled['city'].fillna('Unknown')
df_filled['latitude'] = df_filled['latitude'].fillna('Unknown').astype(str)
df_filled['longitude'] = df_filled['longitude'].fillna('Unknown').astype(str)
```

### cf) 위치정보 변수 전처리 코드

Google Colab

 <https://colab.research.google.com/drive/1XL8QCgFF7rtZiZAcgGJz6Usnr0cfdls3?usp=sharing>



### ▼ 3. 사상자 수 변수 결측치 처리

**nkil**, **nwound** 결측치 대체

- 0과 0이 아닌 값을 나눠서 봄
- success (0과 1) 성공 여부를 나눠서 봄
- 공격 유형을 나눠서 봄
- 3가지 기준 적용해 중앙값으로 대체한다.

```
# 사망자 수 대체
df['nkil'] = df['nkil'].fillna(
    df.groupby(['success', 'attacktype1_txt'])['nkil']
    .transform(lambda x: x[x > 0].median() if (x > 0).any() else 0)
)

# 부상자 수 대체
df['nwound'] = df['nwound'].fillna(
    df.groupby(['success', 'attacktype1_txt'])['nwound']
    .transform(lambda x: x[x > 0].median() if (x > 0).any() else 0)
)
```



▼ 4. 피해자 국적 변수(natlty1\_txt) 결측치 처리

**natlty1\_txt 결측치 대체**

- natlty1\_txt는 타겟 대상의 국적을 뜻하므로 타겟 대상 유형을 기준으로 파악
- targtype1\_txt이 unknown인 경우 → natlty1\_txt도 unknown으로 대체
- country(테러 발생 국가)와 natlty(타겟 국적)이 일치하는 경우를 타겟 대상 유형을 기준으로 파악

		proportion
targtype1_txt	nation_match	
Abortion Related	True	99.39
	False	0.61
Airports & Aircraft	True	82.97
	False	17.03
Business	True	87.47
	False	12.53
Educational Institution	True	96.92
	False	3.08
Food or Water Supply	True	95.85
	False	4.15
Government (Diplomatic)	False	90.39
	True	9.61
Government (General)	True	97.59
	False	2.41
Journalists & Media	True	89.58
	False	10.42
Maritime	True	64.41
	False	35.59
Military	True	87.12
	False	12.88
NGO	False	55.56
	True	44.44
Other	True	75.52
	False	24.48
Police	True	97.00
	False	3.00
Private Citizens & Property	True	93.65
	False	6.35

Religious Figures/Institutions	True	90.31
	False	9.69
Telecommunication	True	98.29
	False	1.71
Terrorists/Non-State Militia	True	96.34
	False	3.66
Tourists	False	66.21
	True	33.79
Transportation	True	95.47
	False	4.53
Unknown	False	100.00
	True	98.00
Utilities	False	2.00
	True	97.50
Violent Political Party	True	97.50
	False	2.50

- 테러 발생 국가와 타겟 국적이 얼마나 일치하는지 타겟 유형별로 비율을 파악
- 이 비율이 97% 이상인 타겟 유형만 natlty를 country로 대체
- True 비율이 97% 이상인 범주 → ['Abortion Related', 'Government (General)', 'Police', 'Telecommunication', 'Utilities', 'Violent Political Party']

```
# 사건 발생 국가와 공격 대상 국적이 같은지 여부를 나타내는 새로운 컬럼 생성
df['nation_match'] = df['country_txt'] == df['natlty1_txt']

# 타겟 대상이 unknown인 경우 natlty도 unknown으로 대체
df.loc[df['targettype1_txt'] == 'Unknown', 'natlty1_txt'] = 'Unknown'

# country = natlty 경우가 97% 이상인 타겟 유형만 natlty의 결측치를 country로 대체
# 1. 그룹별로 nation_match True 비율 계산 (퍼센트 단위)
result_pct = df.groupby('targettype1_txt')['nation_match'].value_counts(normalize=True).mul(100).round(2)

# 2. True 비율만 추출 (MultiIndex의 nation_match 레벨에서 True 값 선택)
true_pct = result_pct.xs(True, level='nation_match')

# 3. True 비율이 97% 이상인 범주의 리스트 생성
categories_high = true_pct[true_pct >= 97].index.tolist()
print("True 비율이 97% 이상인 범주:", categories_high)

# 4. 결측치(또는 NaN)인 'country' 컬럼만 업데이트:
mask_update = df['targettype1_txt'].isin(categories_high) & df['country_txt'].isna()
df.loc[mask_update, 'country_txt'] = 'natlty1_txt'

# 'natlty1_txt' 컬럼의 남은 결측치를 "unknown"으로 채움
df['natlty1_txt'] = df['natlty1_txt'].fillna('Unknown')

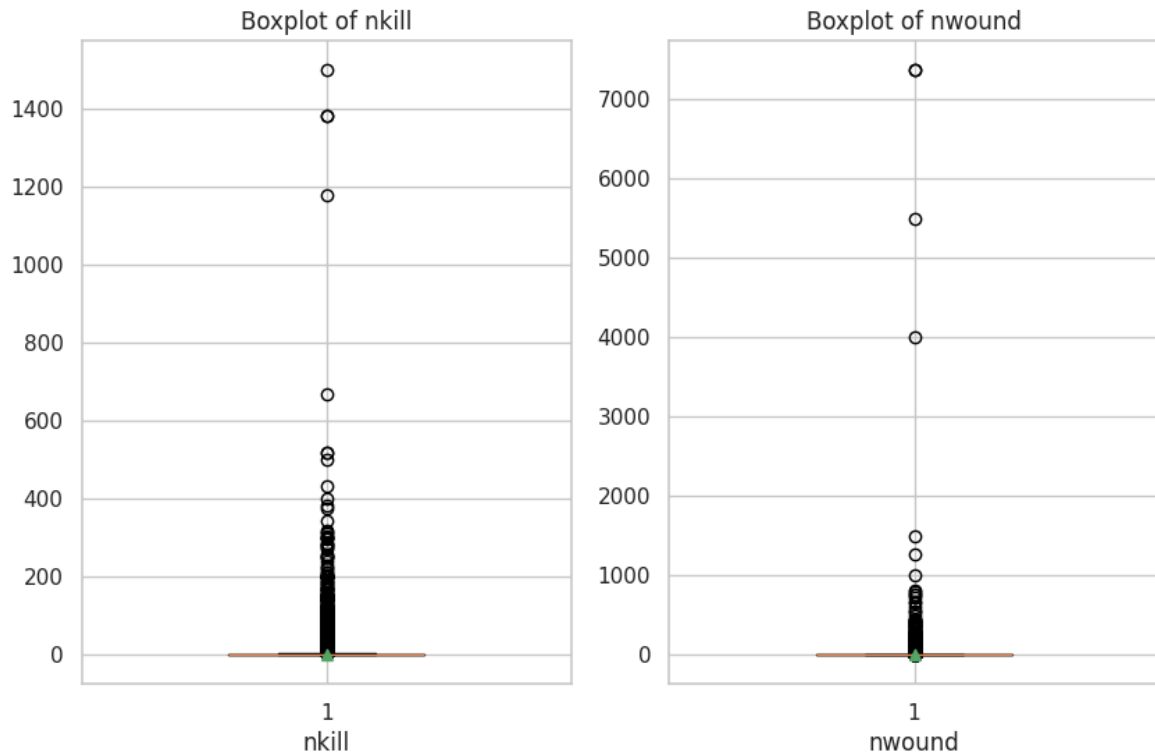
# 'nation_match' 컬럼 삭제
df.drop('nation_match', axis=1, inplace=True)
```

#### ▼ 4. 이상치 파악

- boxplot을 통해 이상치 판단

```
# 사망한 / 부상당한 사람
group1 = ['nkill', 'nwound']

plt.figure(figsize=(10, 6))
for i, var in enumerate(group1, 1):
    plt.subplot(1, 2, i)
    plt.boxplot(df_1990[var].dropna(), vert=True, showmeans=True, widths=0.4)
    plt.title(f"Boxplot of {var}")
    plt.xlabel(var)
plt.show()
```



- 구체적으로 IQR을 수치로 확인

```
# IQR 기반 이상치 통계 요약
import pandas as pd

# 사망주 및 관련 변수 추출
plot_vars = ['nkill', 'nwound', 'nkillter', 'nwoundte']

# 이상치 요약
outlier_summary = {}

for col in plot_vars:
    q1 = df_1990[col].quantile(0.25)
    q3 = df_1990[col].quantile(0.75)
    iqr = q3 - q1
    lower_bound = q1 - 1.5 * iqr
    upper_bound = q3 + 1.5 * iqr
    outliers = df_1990[(df_1990[col] < lower_bound) | (df_1990[col] > upper_bound)]

    outlier_summary[col] = {
```

```

'Q1': q1,
'Q3': q3,
'IQR': iqr,
'Lower Bound': lower_bound,
'Upper Bound': upper_bound,
'Outlier Count': outliers.shape[0],
'Outlier %': round(100 * outliers.shape[0] / df_1990.shape[0], 2)
}

```

```

# 결과 데이터프레임 생성
outlier_df = pd.DataFrame(outlier_summary)
outlier_df

```

	nkill	nwound	nkillter	nwoundte
<b>Q1</b>	0.00	0.00	0.00	0.00
<b>Q3</b>	2.00	3.00	0.00	0.00
<b>IQR</b>	2.00	3.00	0.00	0.00
<b>Lower Bound</b>	-3.00	-4.50	0.00	0.00
<b>Upper Bound</b>	5.00	7.50	0.00	0.00
<b>Outlier Count</b>	12281.00	12465.00	10972.00	1932.00
<b>Outlier %</b>	9.49	9.64	8.48	1.49



## 수행 작업

### ▼ 최혜은

#### ▼ 1주차

##### ▼ 3/24

##### 6조 전처리 - 변수처리

- 데이터가 너무 많아 2013년부터 시작하는 데이터 필터!
- 6조 전처리 - 연도 필터링
- 분석에 유의미한 컬럼만 추출!

```

columns_to_extract = [
    'eventid',
    'iyear', 'imonth', 'iday',
    'country_txt', 'region_txt', 'provstate', 'city',
    'latitude', 'longitude',
    'attacktype1_txt', 'targettype1_txt', 'weaptype1_txt',
    'nkill', 'nwound',
    'gname'
]

```

##### ▼ 3/25

##### 6조(final) - 결측치, 이상치, 파생변수

- 결측치 처리 판단 : nkill, nwound는 결측치를 어떻게 처리해야할지 왜도와, 개수를 보고 판단 후 제거하기로 결정

- 이상치 파악 : 두 변수 이상치가 많았지만, 나중에 처리가 필요하다면 로그 변환 혹은 정규화 할 예정
- 파생 변수 : nkill+nwound를 하나의 컬럼으로 묶는 변수 생성

#### ▼ 3/26

##### 6조(final) - 가설

- 결측치 처리를 결정!
- nkill, nwound, 위도, 경도 결측치 제거
- 가설 검증! (시각화 위주)
- 가설(1) : 중동 지역(Iraq, Pakistan 등)에서 테러 사건이 가장 많이 발생했을 것이다.
- 지역별
  - 가설에 맞게 중동 & 북아프리카 지역의 테러 발생이 제일 많았다.
- 나라별
  - 이라크, 파키스탄, 아프가니스탄의 테러 발생이 많았다.
- 연도별 테러 추이 (중동지역, 나라별)
- 중동&북아프리카 지역 중 이라크가 차지하는 비율이 54.4% → 중동지역, 이라크 변화 추이가 비슷함

#### ▼ 3/27

##### 6조(final) - 가설

- 가설(2) : 테러 피해(사망자 수)는 사용된 무기 유형과 밀접한 관련이 있을 것이다.
- 가설 2에 대해서 시각화 진행
- 무기 유형 및 무기 유형별 사상자 수 파악
- 9.11테러로 인한 Vehicle 이상치 파악
- 공격 유형 = Unknown에 대한 분포 파악
- 피드백을 통해 수정 및 추가
- 추가적인 EDA : 범주형 변수 상관관계 파악

#### ▼ 3/28

##### 6조(최혜은) - 추론통계

- 피드백을 중심으로 추가 그래프 생성
- 중동&북아프리카 지역 내 국가 막대그래프 생성
- 추론 통계 가설1 - 카이제곱 검정 적합도 검정

#### ▼ 2주차

#### ▼ 3/31

##### 6조(최혜은) - nkill, nwound 처리

- nkill, nwound의 결측치 : 성공 여부, 0과0이 아닌 값, 공격 유형
- 3가지 기준 중앙값으로 결측치 대체
- natlty1\_txt 결측치 고민

#### ▼ 4/1

##### 6조(최혜은) - 결측치 처리

- natlty 결측치 대체 완료
- nkill, nwound 결측치 대체 완료
- provstate 결측치 대체

##### 6조(최혜은) - 추론 통계

- 가설1 추론통계검정 추가

#### ▼ 4/2

#### 6조(최혜은) - 추론 통계

- 가설1 연계 : 지역 ↔ 공격 성공 카이제곱 독립성 검정
- 가설1 연계 : 잔차분석
- 가설 2 정규성 검정
- 가설 2 일원분석
- 가설 2 H-test

#### ▼ 4/3

#### 6조(최혜은) - 이라크 머신러닝 (정리X)

- 머신러닝 예측을 위한 범위를 지정 : 이라크 담당!
- 이라크와 테러 성공률에 대해서 EDA 및 모델링
- 정리 필요!

#### ▼ 4/4

- 이라크에 대한 변수별 EDA 추가 진행
- provstate, city에 대해서 범주형 인코딩 다양하게 적용
- OneHot 인코딩된 개별 피쳐 vs 기존 피쳐 중요도 파악

#### ▼ 3주차

#### ▼ 4/7

#### 6조(최혜은) - 이라크 머신러닝

- decision tree, SVM, MLP, KNN, Voting, Steaking 머신러닝 진행
- 모델 성능 비교
- lightGBM이 제일 높게 나왔으나 0(실패)f1과 1(성공)f1 스코어가 차이가 크다.
- 이를 줄이기 위한 가중치 조정, threshold 조정 했지만 결과 유의미하지 않음

#### ▼ 4/8

- 각 모델 성능 비교 마무리
- Shap 도출
- feature importance를 바탕으로 제일 많이 일어난 조합 파악
- PPT 작업 시작!
- 이라크 관련 정책 고민

#### ▼ 4/9

- PPT 자료 준비 (이라크 파트!)
- 필요한 그래프 한글화

#### ▼ 4/10

#### 6조(final)-데이터로 막는 다음 테러

<https://github.com/DAB-4th/Final-Project/tree/team6/6팀/최혜은>

- 깃 코드 제출
- 노션 정리
- PPT 정리

#### ▼ 문규빈

#### ▼ 1주차

#### ▼ 3/24

[https://colab.research.google.com/drive/1Afp1LDRoSJgQfVlxxCZr2G8Qw9\\_2PNAS#scrollTo=Aj5uQ36HICju](https://colab.research.google.com/drive/1Afp1LDRoSJgQfVlxxCZr2G8Qw9_2PNAS#scrollTo=Aj5uQ36HICju)

#### ▼ 3/25

[https://colab.research.google.com/drive/1Afp1LDRoSJgQfVlxxCZr2G8Qw9\\_2PNAS#scrollTo=1vVQY\\_ehm5\\_V](https://colab.research.google.com/drive/1Afp1LDRoSJgQfVlxxCZr2G8Qw9_2PNAS#scrollTo=1vVQY_ehm5_V)

▼ 3/26

[https://colab.research.google.com/drive/1Afp1LDRoSJgQfVlxxCZr2G8Qw9\\_2PNAS#scrollTo=JSSbIxNazNFj](https://colab.research.google.com/drive/1Afp1LDRoSJgQfVlxxCZr2G8Qw9_2PNAS#scrollTo=JSSbIxNazNFj)

▼ 3/27

[https://colab.research.google.com/drive/1Afp1LDRoSJgQfVlxxCZr2G8Qw9\\_2PNAS#scrollTo=MqiVq6eEugyL](https://colab.research.google.com/drive/1Afp1LDRoSJgQfVlxxCZr2G8Qw9_2PNAS#scrollTo=MqiVq6eEugyL)

▼ 이동주

▼ 1주차

▼ 3/24

[https://colab.research.google.com/drive/12vsMldc2u\\_bqe4m6UKp9pGlxldZZpa0?usp=sharing](https://colab.research.google.com/drive/12vsMldc2u_bqe4m6UKp9pGlxldZZpa0?usp=sharing)

▼ 3/26

<https://colab.research.google.com/drive/1QURvIN4Zfjl-3kBvyzEkCQ7jzE1glFWm?usp=sharing>

▼ 3/28

<https://colab.research.google.com/drive/1QURvIN4Zfjl-3kBvyzEkCQ7jzE1glFWm?usp=sharing>

▼ 2주차

▼ 4/2

<https://colab.research.google.com/drive/1QURvIN4Zfjl-3kBvyzEkCQ7jzE1glFWm?usp=sharing>


▼ 장선희


▼ 1주차

▼ 3/24

- 2013년부터 데이터 필터링
- 


Google Colab


 [https://colab.research.google.com/drive/1vKflVUtSSzPQkqBP\\_efoKqu7VzclFfgZ](https://colab.research.google.com/drive/1vKflVUtSSzPQkqBP_efoKqu7VzclFfgZ)



▼ 3/26

Google Colab

 <https://colab.research.google.com/drive/1Hnazk2-znQv7OxodeV4K0CVbxFIDnmMs?usp=sharing>



▼ 3/28

▼ 3주차

▼ 4/8

<https://colab.research.google.com/drive/1A9iQwoXrCvmQJGVvoht6A54W2SLH3DII?usp=sharing>

- SHAP결과 해석 및 인사이트 도출
- @동주 이 @문규빈 @혜은



**피드백**

▼ 규빈님께 남기는 피드백

▼ 3/27 (이동주)

- 중동/북아프리카의 테러 수가 많은 이유가 무엇일까
- 지역별 사건당 평균 피해량이 적거나 많은 곳이 있을까 같은데 특별한 이유가 있을까
- 공격 대상 별 주요 공격 유형은 무엇인지 알면 효율적인 예방에 도움이 될까

▼ 3/28 (최혜은)

[가설1]

1. 라인그래프 2개 (빨강,파랑)

- 2001년은 9.11테러로 인한 급등, 2004년은 이라크 전쟁으로 인한 급등이라고 보는데 그 외에 1998년, 2006년에는 어떤 사건으로 인해 급등했는지 궁금하네요

2. 공격유형별 분석 막대그래프 (빨강,파랑)

- 폭발물 무장공격과 달리 암살은 사건수에 비해 피해가 크지 않은데, 이 원인이 성공률이 낮아서 인지? 궁금하네요. 성공률과 비교해서 보면 좋을 것 같아요.

3. 지역별 최빈 테러 공격 (인터랙티브)

- 어떤 테러가 제일 많이 발생했는지는 보기 쉬운데, 각 지역 전체 테러 중 어느정도의 비율을 차지하는지 보여준다면 더 시각적으로 보기 좋을 것 같아요.
- 사하라 이남, 북아메리카, 오세아니아는 폭발 사고보다 인프라 공격이 많은 이유가 궁금한데 이걸 좀 자료 찾는게 어려울 수 있겠네요.  
혹은 무기 유형과 관련지어 보는 것도 좋을 것 같아요.

4. 테러 사건 규모 top4

- 도시를 접목해서 굵직한 사건을 한번에 볼 수 있는 그래프 아이디어가 너무 좋아요!  
관련 설명도 깔끔하고 이 그래프는 발표자료에 넣고 싶어요!

[가설2]

1. 연도별 테러 발생 애니메이션

- 이 애니메이션도 신기하고 너무 깔끔해서 발표자료에 넣고 싶어요!  
그래프가 다양해서 정말 재밌어요.

[가설3]

1. 정부 vs 민간 라인그래프

- 정부 vs 민간으로 그룹화 아이디어가 너무 좋아요. 분류 예측의 타겟값으로 사용해도 좋을 것 같네요!!

2. 공격 대상별 공격 유형 막대그래프

- 위에 그룹핑한 정부 vs 민간을 가지고 와서  
정부-공격유형 순위 & 민간 - 공격유형 순위 -> 이것도 다중막대그래프 좋을 것 같아요.
- **전체적으로**, 깔끔하게 정리 시각화 도출까지 잘 되어있다는 느낌을 받았습니다.
- ~여기까지입니다~

▼ 동주님께 남기는 피드백

▼ 3/27 (문규빈)

- 성공률에 대한 정의가 무엇인가.
- 왜 국가를 미국,영국,터키,프랑스만 선정했는가. 실제로 가장 많은 난민을 수용한 국가는 튀르키예, 요르단, 콜롬비아 순이다.
- 성공률과 공격유형 그래프를 봤을때, 그럼 폭탄은 발생은 많아지지만 성공률은 낮아진다는 결론인가.왜?
- 터키에서 25년에 테러가 급증하고 또 바로 급락하는데 이유?
- 1990년대 터키에는 난민이 적는데도 테러가 급상하고 급락한다?
- 난민이 증가할수록 테러수가 감소하긴 하지만, 유럽보단 테러가 많이 발생하는곳도 확인해보면 좋을듯.

▼ 3/28 (장선희)



1. GTI스코어가 높은 국가와 실제 테러공격수가 많은 국가들을 봤을 때, GTI스코어에서는 Muslim > Marxist > Hindu > Other Christian인데, 실제 테러 공격수가 많은 국가 상위 20개국을 봤을 땐, 무슬림과 크리스천이 대부분인 것 같아서, 이를 통해서 "Marxist, Hindu, Other Christian의 종교를 가진 사람들이 테러를 일으키면 위험한 테러를 일으키는 사람들이다." 라는 것을 추측해볼 수 있을 것 같아요
  2. 특정 종교별 선호하는 무기유형이 있는지 궁금해졌어요
  3. 지난번에 영상을 보니까 테러단체들이 그들의 종교와 신념에서 비롯되던데, 종교별로 테러 단체가 뚜렷하게 나뉘어질지도 궁금하네요
  4. 아까 규빈오빠가 말했던 것처럼 난민데이터에서 4개국을 선정한 이유에 대한 근거가 확실히 보이면 좋겠습니다~! 제가 알고 있는 지식에서는 유럽권이 난민 수용이 매우 활발했던 지역이라 프랑스 뿐 아닌 다른 유럽권 지역에서의 영향도 궁금합니다
- 끝입니두 ~

#### ▼ 4/2 (최혜은)

전체적으로, 기울기, pvalue, 결정계수에 대한 설명을 덧붙이면 좋을 것 같아요!

- 기울기 -0.0025:  
→ 연도(ityear)가 1년 증가할수록 테러 성공률은 약 0.25%p 감소하는 경향을 보인다.  
→ "시간이 흐를수록 테러 성공률이 낮아진다" 가설1과 방향이 일치
- p-value = 0.00078 < 0.05:  
→ 이 관계는 통계적으로 유의미  
→ 즉, 테러 성공률 감소가 우연에 의한 것이 아닐 가능성이 높다.
- $R^2 = 0.327$ :  
→ 전체 성공률 변화 중 약 32.7%는 연도에 의해 설명  
→ 나머지 67.3%는 다른 요인들(예: 국가, 무기 유형, 공격 방식 등)에 의해 좌우됨
- 위에 내용은 가설1번의 결과 바탕으로 해석을 써봤습니다! 이런식으로 쓰면 더 좋은 자료가 될 것 같아요~!

#### [가설1]

- 테러 성공률 변화가 선형이 아닐 수 있으므로 다항 회귀로 비선형 추세를 한 번 해보는 건 어떨까요?

#### [가설2]

- 제가 ANOVA를 진행하기 전에 정규성 검정을 진행했는데, 이 가설도 정규성 검증을 진행해보시는건 어떨까요?  
저는 from scipy.stats import shapiro를 이용했습니다!

#### [가설3]

- 히트맵을 통해서 종교와 타겟유형을 한눈에 볼 수 있어서 좋아요!

#### [가설4]

- 난민수와 테러발생수를 비교한 그래프가 흥미로워요!
- 통계적으로는 유의하지만, 상관계수와 결정계수가 낮아서 설명력이 약하다는 판단이 들어요.  
추가적인 잔차분석을 진행하면 좋을 것 같습니다.
- ~여기까지입니다~

#### ▼ 선희님께 남기는 피드백

##### ▼ 3/27 (최혜은)

- **가설1 무기유형별 연도 비율 변화 라인 그래프**에서 marker='o' 옵션, grid 옵션을 부여하면 더 시각적으로 보기 좋은 그래프가 될 것 같아요.
- **가설1-2 무기 유형과 공격 대상**

제가 조사했을 때,

**Radiological** 사건이 통틀어서 10건 밖에 없었는데  
크게 다

**정부 대상**이라는게 신기하네요.

근

**사건들에 대해서 알아보면 좋을 것 같아요**

→ 놀랍게도 데이터가 전부 일본 정부 데이터였습니다. 같은날에 발생한 테러도 있어서 세부사항을 확인하기 위해, 원본 페이지를 파일을 확인해봤는데, 동일한 날짜에 다른 장소에서 동일한 무기를 이용한 테러였습니다. 뉴스를 찾아봤는데 기사를 찾진 못했습니다.

- 빨간 막대 차트 (1) - 총기 관련 필터링

무기 유형을 나라별 기준으로 본 것이 좋은 관점아네요!



- 빨간 막대 차트 (2) - 생물학 무기

제가 조사했을 때,

생물학 무기 사건이 25건이었는데 70%가 미국이라는게 흥미롭고,

미국이 주로 쓰는 무기가 무엇인지도 궁금하네요!

이런 테러 관련에서는 동아시아가 많이 언급되지 않는데, 일본이 있다는게 흥미롭습니다.

일본에서 있었던 생물학 무기 이용 테러가 궁금하네요!

- 테러 유형별 연도 비율 변화

1995년 이후로 암살이 줄어들고 폭발과 무장습격이 증가했는데 그 원인이 궁금하네요.

점점 줄어드는 추세가 아니고

급격하게 주황색 그래프가 뚝 떨어진 원인이 궁금합니다.

- 폭발물 테러만 필터링 파이차트

아라크와 others만 구분되는 색상을 쓴다면 더 좋을 것 같아요! ✓

- 전체적으로 시각화가 다양해서 흥미로웠고

저랑 겹치는 부분이 많았는데 저와 관점이 달라서 보는 재미가 있었습니다! 💖

- 추가적으로, 지역별로 보실 때 테블로 이용해서 지도 그래프도 넣으면 좋을 것 같아요! (저도 적용해보려고 합니다) ✓ (만들어 봤답니당)

[final 년도별 무기유형 변화.twb](#)

▼ 3/28 (이동주)

- 사람이 많이 모이는 요일처럼 세계적 행사인 올림픽이나 월드컵이 있는 년도의 테러 수도 연관이 있는지 궁금합니다.
- 중동지역의 차량 테러가 많은 이유가 궁금합니다. 무기의 접근성 때문일까요
- Fake weapons 테러에 비행선이 주요 표적인데 그런 경우에 어택타입이 하이재킹인걸까요? 하이재킹에 사용되는 무기중 fake weapon인 경우가 많다고 볼 수 있을까요?

▼ 4/2 (이동주)

- 요일별 테러 분포에 대해서 교회, 성당에 모이는 주말에 밀집해 있지 않을까 하셨는데 그렇다면 종교 데이터를 이용해 천주교나 기독교 국가 등 종교별 요일 테러 분포도 확인해보면 영향이 있는지 파악할 수 있을꺼 같아요

▼ 해은님께 남기는 피드백

▼ 3/27 (장선희)

- ✓<가설1>

상위 n개의 나라에서 어떤 유형의 공격이 많이 일어나는지도 알아보면 좋을 것 같아요

➡ 누적그래프로 표현

- ✓<무기유형별 피해량>

부상자와 사망자를 합친 피해량으로 그래프를 그리셨는데, 사망자가 더 많은 유형과, 부상자가 더 많은 유형으로 나누면 어떤 인사이트가 나올 수 있을지 고민하게 됐어요

➡ 사망자별 유형, 부상자별 유형을 막대그래프로 표현

- ✓<무기유형별 평균 피해량>

폭발물이 광범위하게 대량살상할 수 있는 무기로 1위를 차지할 것이라고 생각했는데, vehicle이 1위인게 의외네요

➡ 9.11 테러로 인해서 비이상적인 사상자가 나왔고 이 이상치로 인해 평균피해량이 올라갔습니다!

- ✓<추가적인 EDA>

total victim이 nwound + nkill로 이루어진 것이기 때문에 이 변수들간의 상관계수들이 높게 나오는 것은 당연하다는 생각이 들어요

혹시 이 상관계수를 확인해봐야겠다고 생각하신 이유가 있을까요??

→ 크게 다른 사유가 있는 것은 아니고 수치형 변수를 묶어서 상관관계를 볼까? 하는 마음으로 만든건데 total\_victim이라는 파생변수에 대해서는 생각하지 않았어요 😞 피처가 더 생기면 추가로 넣으면 좋을 것 같다는 생각뿐....!

#### ▼ 3/28 (문규빈)

- ✅ 위 도출에서 연노운 삭제한다고 했는데 뭔가 너무 데이터가 많은데...? 가설과 벗어난 그냥 해본걸까요?  
→ 연노운 데이터를 삭제한 이유 혹은 연노운에 대한 질문이 들어올까 싶어서 다양한 결과를 만들어놓고 나중에 필요할 때 사용하려고 합니다! (가설과는 무관)
- ✅ 남아시아, 동남아시아 구분한거 생각도 못했는데 좋네요.  
👍
- ✅ 2번가설에 테러피해(사망자 수) 인데 피해자수 가 좋지않을까 하는 생각임다.  
→ 바꿨습니다!
- ✅ ### 중동&북아프리카 지역 내에서 테러 건수가 많은 국가 이거 파이로 비율도 좋은데 막대그래프도 있는거 어떨까요  
  
→ 추가했습니다!
- ✅ 전체적으로 한글화하면 더 좋을거 같아요.  
→ 제가 한글 폰트 충돌이 일어났는데 해결을 못해서 영어로 진행했습니다 ππ 발표자료에 쓰일 그래프가 있으면 간단 영어로 바꿔놓겠습니다!

#### ▼ 4/2 (장선희)

- ✅ 가설1 지역에 대한 가설에서의 잔차분석
  - 기댓값보다 크고 작음의 기준을 2로 잡으셨는데, 원래 기준이 그런건가요?  
제가 chat선생님과 함께 공부했을 때는 아래처럼 알려주셨는데, 이거랑은 다른건가요??

### 📌 사용하는 잔차의 종류

보통은 표준화 잔차 (Standardized Residuals) 또는 정규화된 잔차 (Adjusted Residuals) 를 써.

- 값이  $\pm 1.96$  이상이면 → 95% 유의 수준에서 의미 있는 차이
- 값이  $\pm 2.58$  이상이면 → 99% 유의 수준에서 더 강한 차이

→ 95% 유의수준으로 판단한 것 맞습니다! 대신 폭 넓게 1.96대신 근사치인  $\pm 2$ 로 설정했습니다! 1.96으로도 코드를 돌려보겠습니다! 👍👍

- 성공률에 관한 가설

✅ 1-1. 갑자기 헛갈리는건데 지역별 성공률이란 해당 지역에서 테러가 일어났을 때, 그것이 성공할 확률을 말하는건가요??

→ 네! 성공한 확률을 말합니다! → 성공한 테러 사건 수 / 전체 테러 사건 수

성공률에 대한 구체적인 설명을 추가하겠습니다!

✅ 1-2. 만약 그런거라면 가장 낮은 서유럽에서는 어떻게 성공률이 낮을 수 있는 건지도 궁금하다! (데이터로 예측이 가능할랑가..~)

→ 서유럽에 대해서 성공 사건 수, 전체 사건 수, 연도별 성공 추이에 대해서 추가적인 시각화 도출을 진행해보겠습니다!

2. 제가 무기유형 컬럼을 골라서 그런지 각 지역에서 어떤 무기유형이 성공률이 높고 실패율이 높은지도 궁금하네용 ㅎㅎ 이게 각각 다르면 각 지역의 문화나 특성별로 성공률이 달라지는것을 확인할 수 있을 것 같기도 하고, 그러면 방지 방법도 생각날 것 같기도 하고...!?



## 분석(예측) 방향 아이디어

**우리의 식스센스로 해당 나라의 테러성공을 예측해서 테러를 미리 대비한다!!**

### ▼ 3/31 아이디어

- 범주형이 많다.
- 국가를 추려내서 특정 국가에서 발생할 예측
- 테러 유형 분류 예측
- 규모 예측 (사상자)
- 특정 월에 발생할 예측
- 위도 경도가 있으므로 발생 지역을 예측
- 월 일 도시 석세스 지역 예측 모델 생성 -> 댓글 가져온 다음에 NLP (LEG) 요소 추출  
ex) 댓글에서 파키스탄에서 테러가 발생할 것 같다고 예측 -> 우리 데이터의 '파키스탄'을 찾고 그를 분석

### ▼ 4/1 아이디어

- 저도 1번으로 할 경우에는 중동지역으로 보면 근거가 있어서 좋고,  
한국으로 보면 우리나라니까 궁금해서 했다 할 수 있어서 이거 2개 생각하고 있었어요
- 특정 국가에서 사건 발생 예측하면서 디테일하게 장소나 시간분석까지 해보는 예측모델을 만들어보고 싶습니다
- 규모 예측할때 공휴일이나 여름휴가 시즌 이런것도 고려를 해봤고
- 제가 세운 가설 중에서 월별로 무기유형별 차이가 있다 라는 가설이 있어서 그걸 다시 보고 왔는데
- 이걸 바탕으로 월별 예측을 해보려고한다. 저런 상관분석을 나라별, 공격유형별 이런식으로 확장하면
- 되지않을까요?? 무기유형도 충분히 고려할만한 사항이니까
- 공격유형도 해보고  
차이가 얼마나 있는지 확인 한번 해보는것도  
괜찮을까요  
지역별 무기유형도 보면 좋을꺼 같아요
- 어떤 요소들이 테러의 성공을 결정짓는가? 뭐 이런거도 될까요
- 저는 분석 결과를 바탕으로 어느 지역에 더 많은 예방 및 대응 자원을 투입해야 하는지?

### ▼ 4/3 회의록

1. 각자가 선택한 나라의 전체 테러 성공을 예측한다. 성공률이 높으면 문제!!

⇒ 성공률이 높게 나오는 요인들을 탐색 (기존 EDA + 추가적인 EDA)

- 무기 유형
- 공격 유형
- 타겟 대상 유형
- 월(month)
- 요일(day)
- 종교

이를 통해 나온 결과들을 종합해서 각 나라 내에서

장소 / 사상자 수 / 시기 / 무기 유형 / 공격 유형 / 공격 대상 등을 예측해서 방안을 세우는 것을 목표로한다.

## 2. 전체적인 목표

- 테러 피해를 줄인다.
- 예측 모델 : 이전 ↔ 이후 얼마까지 피해를 줄일 수 있는지
- 발생율이 높은곳에 대비하기
- 사상자 중의 사망자가 차지하는 비율

피해 최소화 :  $y = nkill, nwound$

성공을 예측해서 예방 :  $y = success$

## 3. 일정 조율

- ~4/8(화) : 머신러닝 마무리
- 4/9(수)~10(목) : 발표자료 준비 + 대본
- 4/11(금) 오전 : 발표 준비 마무리!!

## ▼ 4/4 회의록 - 변수 선택 & 지역 선택 및 모델링 방안 논의

### 1. 예측 목표 설정

- **분류 모델**: 테러의 성공 여부(success) 예측 (0/1)
- **회귀 모델**: 테러 발생 시의 사상자 수(nkill, nwound 또는 이 둘의 합인 casualties)를 예측  
→ 이를 통해 "현재 사상자 수 대비, 개입 후 얼마나 줄일 수 있을지" 시나리오 분석이 가능해짐

### 2. Feature Engineering 및 인코딩

- **범주형 변수**
  - **타깃 인코딩 또는 임베딩**: 만약 범주 수가 너무 많아 차원이 급격히 늘어난다면 고려할 수 있지만, 현재는 Label Encoding으로 충분할 것 같습니다.
- **수치형 변수**: **nkill**, **nwound**, 그리고 파생 변수 **casualties** (예:  $nkill + nwound$ )는 그대로 사용하거나, 필요 시 스케일링(StandardScaler, MinMaxScaler) 적용

### 3. 모델링 및 시나리오 분석

- **분류 모델**로 테러 성공 여부를 예측하고,
- **회귀 모델**을 통해 사상자 수를 예측

이를 통해 "현재의 테러 상황(예: 특정 장소, 시기, 무기/공격/타깃 유형)에 대해 모델이 예측하는 사상자 수"와

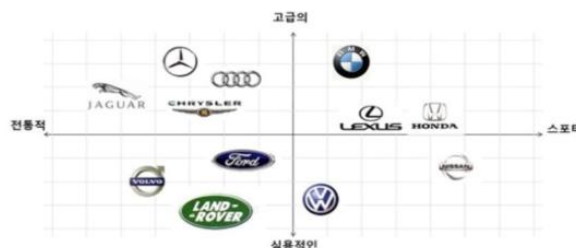
"개입(정책, 방어 대책 등)을 가정한 후 변화를 어떻게 줄일 수 있는지"를 시뮬레이션 해보는 것이 목적.

예를 들어,

- **현재 시나리오**: 모델이 예측한 사상자 수가 100명
- **개입 후 시나리오**: 입력 변수(예: 보안 강화, 특정 공격 유형 감소 등)의 변화를 가정하면 예측이 70명으로 나온다면,  
→ 약 30%의 사상자 감소 효과를 기대할 수 있다고 해석할 수 있겠다!!

++동주오빠 의견

여러 국가를 할 거라면 아래 포지셔닝처럼 특징을 뽑아서 배치하는 형식으로 한 뒤, 각각에 맞는 대응전략 도출하는것도 좋을 듯~!





## 비교 분석 & 프로젝트 방향

### ▼ 공통 비교 기준

	비교 기준	요약 설명
①	테러 성공률	성공한 공격 비율
②	공격 유형	자살폭탄, 총격, 차량 등
③	국가 대응 역량	보안 인프라, 예방 능력
④	지역·문화적 배경	중동/서구권, 정치체제 등
⑤	테러 조직 특성	중앙 집중 vs 분산형 조직
⑥	테러 목적	정치, 종교, 사회 불만 등

### ▼ 두 국가를 중심으로 비교분석

#### ▼ 1. 이라크 vs 터키

항목	이라크	터키	비교 요점
성공률	높음	상대적으로 낮음	이라크가 내전으로 인한 성공률 우세
공격 유형	자살폭탄, 폭탄테러	폭탄, 총격	자살폭탄 비중에서 차이
대응 역량	매우 낮음 (전쟁·혼란)	중간	국가 통제력 차이
문화/지역	중동	중동/유럽 중간	지역 문화 유사성 존재
조직	ISIS, Al-Qaeda	PKK	조직 집중도 vs 분산성 차이
목적	정치/종교 극단주의	분리주의 중심	동기는 명확히 다름

→ 실질적인 지역 기반 비교가 가능, 단 유사성이 높아 명확한 인사이트 도출은 다소 약할 수 있음.

#### ▼ 2. 이라크 vs 영국

항목	이라크	영국	비교 요점
성공률	높음	낮음	대응 시스템 차이
공격 유형	자살폭탄, 폭탄	차량 돌진, 총기, 자폭	공격 방식의 현대화 차이
대응 역량	낮음	매우 강함	보안, 감시 체계 차이 큼
문화/지역	중동	서구권	명확한 대비 가능
조직	ISIS, 알카에다	IRA(과거), 소규모	대형 vs 국지적 조직
목적	종교/정치	정치/사회적 메시지	다름

→ 명확한 대비 구도가 뚜렷, 테러 환경·정치 구조·대응이 극단적으로 다름, 인사이트 도출에 적합할 수 있음

#### ▼ 3. 이라크 vs 미국

항목	이라크	미국	비교 요점
성공률	높음	낮거나 중간	혼란 상황 차이
공격 유형	자살폭탄, 폭탄	총기, 차량, 방화	공격 스타일 대조적
대응 역량	낮음	매우 강함	대응/예방 시스템 강도 차이
문화/지역	중동	서구권	대비 뚜렷
조직	ISIS 중심	단독범/무소속	분산형 vs 조직형
목적	종교·정치	개인 불만·사회문제	동기 구조 다름

→ 가장 선명한 대비 가능, 전통적 테러국가 vs 내적 불만 기반 테러국

#### ▼ 4. 터키 vs 영국

항목	터키	영국	비교 요점
성공률	낮거나 중간	낮음	큰 차이는 없음
공격 유형	폭탄, 총격	차량, 자폭, 총기	약간의 차이
대응 역량	중간	강함	차이 존재

문화/지역	중동/유럽 경계	서구권	문화적 대비 있음
조직	PKK 등	IRA(과거) 등	유사한 분리주의 성격
목적	분리 독립	정치/이념	동기 비교 가능

→ 대비 가능한 항목 존재하지만 차이가 다소 약함, 유럽권 내부 비교 느낌

#### ▼ 5. 터키 vs 미국

항목	터키	미국	비교 요점
성공률	중간	낮거나 중간	큰 차이 없음
공격 유형	폭탄, 총기	총기, 방화, 차량	일부 차이
대응 역량	중간	강함	보안 시스템 차이
문화/지역	중동/유럽 경계	서구권	대비 가능
조직	PKK	무소속, 내부 테러	구조적 차이 있음
목적	분리주의	사회적 분노	비교 가능

→ 차이보다는 비슷한 점이 많아, 극단적 비교보다 '내부 테러 변화' 같은 주제에 적합

#### ▼ 6. 영국 vs 미국

항목	영국	미국	비교 요점
성공률	낮음	낮거나 중간	유사
공격 유형	차량, 자폭	총기, 차량	스타일은 유사하나 총기 중심 차이
대응 역량	강함	강함	유사
문화/지역	서구권	서구권	거의 동일
조직	IRA/무소속	무소속	유사
목적	정치/종교	사회적 불만	차이 존재

→ 대비보다는 유사점이 많음, '서구권 내 테러 변화'라는 주제에 적합

#### ▼ 네 국가 전체 비교분석

##### 1. [비교 테이블]

기준	이라크(최혜은)	터키(이동주)	영국(장선희)	미국(문규빈)
① 성공률	매우 높음 (내전 상황)	중간	낮음	낮거나 중간
② 공격 유형	자살폭탄, 무차별 폭탄	폭탄, 총기	차량 돌진, 자폭	총기, 방화, 차량
③ 대응 역량	매우 낮음 (혼란)	중간 (내부 단속 존재)	높음	매우 높음
④ 지역/문화	중동 (정치·종교 불안)	중동-유럽 접경	서구권	서구권
⑤ 조직	ISIS, 알카에다 (강한 중앙 조직)	PKK 등 (내부 반군)	IRA/소규모 (과거 중심)	무소속 또는 개인 단독범
⑥ 목적	종교·정치	민족주의·분리주의	정치·사회 메시지	사회 불만, 이념적 분노

##### 2. [각 나라 특징]

##### ● 이라크

- 전형적 분쟁지역 테러 국가
- 조직 집중도 높고 자살폭탄 비율 높음
- 정치적·종교적 이유 강함
- 대응력 낮아 성공률 매우 높음

##### ● 터키

- 중간지대 특성: 유럽/중동 경계에 위치
- 민족·분리주의적 테러 중심
- 보안은 있으나 이라크보다는 양호

##### ● 영국

- 서구권 전통 테러 경험국
- 과거 IRA 중심 → 최근엔 차량·자폭 중심으로 변화
- 사회적 메시지 중심, 성공률 낮음

##### ● 미국

- 개인형·내부형 테러 중심
- 대형조직보다는 "개인 불만 기반" 테러
- 총기류 이용 테러 많음, 대응 시스템 매우 강함

##### 3. [비교 시사점]

비교 포인트	시사점
--------	-----

성공률	혼란지역일수록 성공률 ↑ (이라크), 보안체계 견고할수록 ↓ (미국, 영국)
공격 수단	문화·법적 환경에 따라 달라짐 (총기 보유 가능 여부 등)
조직 구조	중동은 중앙 집중 조직, 서구는 분산형 또는 개인 단독 범죄
대응 역량	대응력의 차이가 피해 규모와 성공률에 직접 영향
목적 차이	중동은 정치·종교적, 서구는 개인적·사회적 동기 많음

#### ▼ 전체적인 방향

#### ✓ 목표 요약

각 국가(이라크, 터키, 영국, 미국)의 테러 데이터를 바탕으로

- ✓ 테러 성공 여부를 예측하고
- ✓ 주요 영향 요인 도출 및
- ✓ 국가 맞춤형 정책/대응 전략을 제시하는 것

#### ✓ 최종 보고서/발표에서의 결론 및 인사이트 도출 방식

##### ▼ 1. 국가별 예측 성능 요약

- Accuracy / F1-score / ROC-AUC 등 지표를 정리하고
- 모델이 신뢰할 만한지 평가

국가	Accuracy	F1-score	ROC-AUC
이라크	0.92	0.95	0.83
터키	0.89	0.91	0.81
...	...	...	...

##### ▼ 2. Feature Importance 비교

- 국가별로 어떤 피처가 성공 여부에 중요한 영향을 주는지 비교
- 이를 통해 "국가별 취약한 테러 유형" 파악 가능

[예시]

국가	주요 요인
이라크	target type , city , weapon type
터키	attack type , region , imonth
영국	weekday , provstate , targtype1_txt
미국	city , attack type , weaptype1_txt

##### ▼ 3. 공통 패턴 & 특이점 도출

- 모든 국가에서 공통적으로 중요한 피처는 무엇인가?
- 반대로, 특정 국가에서만 중요한 피처는?

[예시]

- "attacktype1\_txt"는 전 국가 공통 중요 변수지만, "weekday"는 영국에서만 의미 있는 변수로 나타남.

##### ▼ 4. 국가별 정책 인사이트 제안

###### 🇮🇷 이라크

- 성공률 높은 유형: Assassination + Firearms + Baghdad
- 대응 제안: 특정 지역·무기 조합 감시 강화

###### 🇹🇷 터키

- 특정 월( 6~8월 )에 테러 성공률 급증
- 대응 제안: 여름철 경계 태세 강화



## 🇬🇧 영국

- **요일별 성공률** 차이가 큼 (예: 월·화 ↑)
- 대응 제안: 평일 보안 인력 재배치

## 🇺🇸 미국

- **target = Government** 의 성공률이 높음
- 대응 제안: 공공기관 보안 강화 필요

### ▼ 5. 결론 정리 예시

본 연구는 GTD 데이터를 기반으로 4개국의 테러 성공 여부를 예측하고 주요 영향을 미치는 요인을 분석함으로써, **국가 맞춤형 테러 대응 전략 수립의 데이터 기반 인사이트를 제공하였다.**

특히, **city**, **target type**, **attack type** 이 전반적으로 중요한 요인이며, 국가별로 **계절성/지역성/무기 유형별 특이한 패턴**도 존재하였다.

이러한 결과는 **\*\*정책적 의사결정(예산 배분, 감시 강화 구역 설정 등)\*\***에 활용될 수 있다.

### ? 추가 가능하면 좋은 내용

- SHAP 기반 인사이트 (각 feature가 성공 확률에 어떤 방향으로 작용하는지)
- 위험도 지도 (예: 성공률 높은 지역 시각화)
- Threshold 조정에 따른 경계 수준 시뮬레이션

### ▼ 비교 분석 구조(5단계 비교 분석)

#### ▼ 1. 모델 성능 비교

→ 각국 예측 정확도/정밀도 비교

국가	Accuracy	F1-score	ROC-AUC	PR AUC
🇮🇷 이라크	0.92	0.95	0.83	0.96
🇹🇷 터키	0.89	0.91	0.81	0.93
🇬🇧 영국	0.86	0.89	0.79	0.91
🇺🇸 미국	0.88	0.90	0.80	0.92

📝 해석 예시:

이라크는 클래스 불균형에도 불구하고 높은 성능을 보이며, 모델이 복잡한 성공 패턴을 잘 포착했음을 의미.

#### ▼ 2. 중요 피처 비교

→ 어떤 피처가 국가별로 성공에 영향을 주는가?

피처명	이라크	터키	영국	미국
city	✓	✓	✓	✓
targettype1_txt	✓	✓	✓	✓
weekday	✗	✗	✓	✗
imonth	✓	✓	✗	✗
attacktype1	✓	✓	✓	✓
weaptype1	✓	✗	✗	✓

📝 해석 예시:

공통적으로 중요 변수는 **"target type"**, **"attack type"**이며, 영국은 **요일**, 이라크·터키는 **월(month)**에 영향이 있음.

#### ▼ 3. 고위험 조합 비교

→ 국가별로 성공률이 유독 높은 유형은?

[예시]





국가	조합 사례 (공격 + 타겟 + 지역)	성공률
 이라크	Assassination + Private Citizen + Baghdad	96%
 터키	Bombing + Government + Istanbul	94%
 영국	Armed Assault + Police + London	89%
 미국	Unarmed Assault + Government + Washington	87%

 **인사이트:**

이런 고위험 조합을 중심으로 **예방/감시 자원 우선 배분 가능**

#### ▼ 4. 정책 제안 비교

→ 국가별로 구체적인 대응 전략 도출

국가	제안 전략
 이라크	특정 도시와 공격 조합에 대한 감시 체계 강화
 터키	특정 계절(6~8월)에 자원 집중
 영국	평일 출근 시간대 보안 인력 재배치
 미국	정부 기관 대상 접근 통제 강화

#### ▼ 5. 결론

→ 테러 대응의 데이터 기반화

이 분석은 GTD 데이터를 활용해

각 국가의 **테러 성공 패턴을 예측하고 비교함**으로써,

**국가별 상황에 맞는 예방 전략과 자원 분배 방안**을 제시한다.

#### ▼ 한 줄 요약:

☒ 4개 국가를 비교하는 방식은 "예측 성능, 중요 변수, 고위험 조합, 정책 전략" 4가지 기준으로 분석하면 체계적으로 인사이트 도출이 가능



## 머신러닝 파이프라인

### 1. 인코딩 방식

- 범주형 데이터를 수치형 데이터로 변환하기 위해 사용하는 인코딩

인코딩 방식	설명
Label Encoding	<ul style="list-style-type: none"> <li>- 설명 : 범주형 값을 정수(0, 1, 2, ...)로 매핑</li> <li>- 예시: {'Red': 0, 'Blue': 1, 'Green': 2}</li> <li>- 장점 : 간단하고 메모리 효율적</li> <li>- 단점 : 숫자 간의 순서가 의미 있는 것으로 오해될 수 있음 → 트리 기반 모델에서는 괜찮지만, 선형 모델에서는 주의 필요</li> </ul>
One-Hot Encoding	<ul style="list-style-type: none"> <li>- 설명 : 각 범주를 이진 벡터로 표현. (각 범주를 새로운 열로 만들고, 해당 열에만 1을 부여)</li> <li>- 예시 : Color: Red → [1, 0, 0]</li> <li>- 장점 : 범주 간의 순서 정보가 없음을 명확히 표현.</li> <li>- 단점 : 차원이 급격히 늘어남 → 컬럼수 증가</li> </ul>
Ordinal Encoding (순위 인코딩)	<ul style="list-style-type: none"> <li>- 설명 : 범주의 순서가 명확할 때, 그 순서에 따라 숫자를 부여.</li> <li>- 예시: {'Low': 0, 'Medium': 1, 'High': 2}</li> <li>- 사용 조건: 순위가 의미 있는 경우만 사용 (예: 품질 등급, 학력 등)</li> </ul>
Frequency Encoding (빈도 인코딩)	<ul style="list-style-type: none"> <li>- 설명 : 각 범주를 해당 값이 나타나는 빈도로 대체.</li> <li>- 예시: {'Red': 0.4, 'Blue': 0.3, 'Green': 0.3}</li> </ul>

	<ul style="list-style-type: none"> <li>- 장점: 차원이 늘어나지 않음.</li> <li>- 단점: 빈도에 따라 모델이 잘못된 중요도를 학습할 수 있음.</li> </ul>
Target Encoding	<ul style="list-style-type: none"> <li>- 설명: 각 범주에 대해 타겟 값의 평균값으로 대체.</li> <li>- 예시: {'Red': 0.8, 'Blue': 0.5, 'Green': 0.3} (타겟이 이진일 때)</li> <li>- 장점: 정보량 많고 차원 축소됨.</li> <li>- 단점: 데이터 누수(leakage) 가능성 존재 → 훈련/검증 분리 후 적용 필요</li> </ul>

## 2. 불균형 처리 기법

구분	기법	설명
리샘플링(Resampling) 계열 (리샘플링은 사전 데이터 조정 단계에서 적용)	Undersampling	<ul style="list-style-type: none"> <li>- 다수 클래스의 샘플을 줄임. 정보 손실 가능성 있음.</li> <li>- 장점: 학습 시간 단축</li> <li>- 단점: 정보 손실 위</li> </ul>
	Oversampling	<ul style="list-style-type: none"> <li>- 소수 클래스의 샘플을 복제. 과적합 위험 존재.</li> <li>- 장점: 간단하고 효과적</li> <li>- 단점: 과적합 위험</li> </ul>
	SMOTE	<ul style="list-style-type: none"> <li>- 소수 클래스 샘플 사이에 새로운 가상 샘플 생성. 일반적으로 성능 향상에 유리.</li> <li>- 장점: 과적합 완화</li> <li>- 단점: 노이즈 증폭 가능성</li> </ul>
	ADASYN	<ul style="list-style-type: none"> <li>- SMOTE + 어려운 샘플에 더 많은 가상 샘플 생성</li> <li>- 장점: 유연성이 상승함</li> <li>- 단점: 불안정할 수 있음</li> </ul>
	Tomek Links	<ul style="list-style-type: none"> <li>- 클래스 경계에 있는 샘플 제거 (정제 효과)</li> <li>- 장점: Noise 제거</li> <li>- 단점: 과도한 정보 손실 가능</li> </ul>
Class Weight 조정		<ul style="list-style-type: none"> <li>- 모델 학습 시 손실 함수에 클래스 가중치를 조정.</li> <li>- Scikit-learn 계열 모델에서 class_weight='balanced' 옵션 사용 가능</li> <li>- 장점: 리샘플링 없이도 불균형 완화</li> <li>- 단점: 데이터 불균형이 심할 경우 한계 존재</li> </ul>

## 3. 학습 모델 종류

- 지도학습

구분	모델	설명
선형 계열	Linear/Logistic Regression	선형 분류 모델로 해석력이 좋음. 단순하고 빠름.
	SVM (Support Vector Machin)	클래스 간 최대 margin을 찾는 결정 경계를 생성. 고차원 분류에 강하지만, 커널 선택과 계산 비용이 변수.
트리 계열	Decision Tree	조건 기반 분류로 직관적이지만 과적합 위험 있음.
	Random Forest	다수의 결정 트리를 앙상블하여 과적합 감소 및 성능 향상.
	XGBoost	Gradient Boosting 방식. 빠르고 정교한 예측 가능.
	LightGBM	XGBoost보다 더 빠르고 메모리 효율적인 Boosting 방식.
거리 기반 비모수	KNN (K-Nearest Neighbors)	주변 이웃의 다수 클래스를 기준으로 분류. 학습은 빠르나 예측 시 계산 비용이 높고, 스케일에 민감함.

### ▼ 비지도학습

구분	모델	설명
군집화 (Clustering)	KMeans	중심점 기준 거리 최소화
	DBSCAN	밀도 기반, 군집 개수 자동 결정
	계층적 군집(Hierarchical)	덴드로그램, 병합/분할 방식
차원 축소 (Dimensionality Reduction)	PCA	분산 최대 보존, 선형
	t-SNE	비선형, 시각화용 (거리/유사도 보존)
	UMAP	t-SNE보다 빠르고 구조 보존 더 잘함
이상 탐지 (Anomaly Detection)	LOF (Local Outlier Factor)	밀도 기반 이상점 탐지
	Isolation Forest	트리 기반, 빠름

	One-Class SVM	경계 학습 기반 이상 탐지
잠재 변수 학습 (Representation Learning)	AutoEncoder	입력 → 압축 → 재구성
	GAN (Generative Adversarial Network)	생성자 vs 판별자 경쟁
	VAE (Variational AutoEncoder)	확률 기반 AutoEncoder

#### 4. 성능 평가 지표

지표	설명
Accuracy	전체 중 맞춘 비율. 불균형 데이터에서는 신뢰 어려움.
Precision	양성 예측 중 실제 양성 비율. 정밀도.
Recall	실제 양성 중 모델이 맞춘 비율. 재현율.
F1 Score	정밀도와 재현율의 조화 평균. 불균형 데이터에 유리.
ROC-AUC	양성/음성 분류 성능을 종합적으로 평가하는 지표.



#### 참고 자료

1. 미국과 영국의 테러조직 지정 및 관리에 대한 고찰

[미국과 영국의 테러조직 지정.pdf](#)

2. 요인테러범죄의 실태분석과 그 대책방안에 관한 연구 - 독일 요인경호 업무와 비교하여 -

[요인테러범죄의 실태분석과 그.pdf](#)

3. 테러 대응 로컬 거버넌스 활성화 방안에 관한 연구: 미국의 9·11테러와 영국의 7·7테러 사례를 중심으로

[KBB\\_SCHOLAR 테러 대응 로컬 거버넌스 활성화 방안에 관한 연구.pdf](#)

4. 세계 종교 데이터

<https://www.kaggle.com/datasets/edoardoba/world-flags>



#### 프로젝트 타임라인 (전체 노선과 동일)

##### 파이널 타임라인

Aa 이름	날짜	태그
▶ <b>파이널(1주차) - 전처리, 시각화, 가설검증</b>	@2025년 3월 24일 → 2025년 3월 28일	
📄 <b>1일차</b>	@2025년 3월 24일	
🎯 <b>2일차</b>	@2025년 3월 25일	
📊 <b>3일차</b>	@2025년 3월 26일	
📝 <b>4일차</b>	@2025년 3월 27일	
👤 <b>5일차</b>	@2025년 3월 28일	

Aa 이름	📅 날짜	☰ 태그
▶ <u>파이널 (2주차) - 결측치 마무리, 가설 추론 통계 검증, 머신러닝 시작</u>	@2025년 3월 31일 → 2025년 4월 4일	
🧠 6일차	@2025년 3월 31일	
🧠 7일차	@2025년 4월 1일	
▶ <u>파이널 (3주차) - 머신 러닝, 발표 준비</u>	@2025년 4월 7일 → 2025년 4월 10일	
🧠 9일차	@2025년 4월 3일	
🧠 8일차	@2025년 4월 2일	
🧠 14일차	@2025년 4월 10일	
🧠 10일차	@2025년 4월 4일	
제목 없음		
🧠 11일차	@2025년 4월 7일	
🔥 12일차	@2025년 4월 8일	
🖨️ 13일차	@2025년 4월 9일	

## 6팀 업무 보드

Aa 이름	👤 담당자	🔄 상태
✅ <u>주제 선정</u>		완료
✅ <u>주요 변수 추출</u>		완료
✅ <u>가설 검증</u>		완료
✅ <u>시각화 도출</u>		완료
✅ <u>결측치 처리</u>		완료
✅ <u>추론 통계</u>		완료
✅ <u>타겟 변수 설정</u>		완료
✅ <u>머신 러닝</u>		완료
✅ <u>예측 방향 정하기</u>		완료
✅ <u>PPT 준비</u>		완료
✅ <u>대본 준비</u>		완료