STREET Data 분석 및 연구 진행 과정

발표자 : 202000919 손건희

목차

table of contents

1

논문 Review

Part 1

논문 리뷰

1. 무슨 데이터인가?

2. 데이터셋의 목표

3. 데이터 수집 과정

논문리뷰

무슨 데이터인가?

약 400만장 이상

Part 1

논문 리뷰

1. 무슨 데이터인가?

2. 데이터셋의 목표

3. 데이터 수집 과정

2. 데이터셋의 목표

기준 연구의 한계점

기존의 교통 데이터셋들은 대부분 **도심 지역**이나 **고속도로** 교통흐름에 초점을 맞추고 있어, **교외 지역의** 교통 패턴을 잘 반영하지못하는 **한계가 존재**

주요 목표

지능형 교통 시스템 및 **스마트 시티 구축**에 필요한 교통데이터를 제공하고 연구자들에게 **벤치마크 데이터셋으로 활용**될 수 있도록 하는 것

논문 리뷰

1. 무슨 데이터인가?

2. 데이터셋의 목표

3. 데이터 수집 과정

3. 데이터 수집 과정

전체 기간	2.5개월
사진 간격	10분
사진 방향	최소 2방향 ~ 최대 4 방향

- [ ] graphs.zip
- [ ] incidents.zip
- [ ] roadmasks.zip
- [ ] trafficcounts.zip
- [ ] trafficstate.zip
- [ ] vehicleannotations.zip
- [ ] viewclassifiers.zip

추가적으로 데이터 정보 넣을지 고민 중

데이터 정보는 바로 사고 데이터 파라미터 등 추가적인 정보

2

데이터 EDA 및 전처리

요약 | Overview

과제 1

과제 2

과제 3

과제 1 데이터 소개

	eventid	eventtype	road	roaddirection	MaxOftrafficimpact	startcrossstreet	endcrossstreet	responsestart	createtime
0	929916	Stall	I94	3	1	IL-60 (Town Line Rd)	IL-60 (Town Line Rd)	NaT	2018-08-20 00:03:30
1	929917	Accident	I94	2	1	Old US Hwy 41	Old US Hwy 41	NaT	2018-08-20 00:03:39
2	929918	Stall	IL Rte 21	2	1	Casey Rd	Casey Rd	NaT	2018-08-20 00:07:39
3	929919	Other	I94	4	1	Everett Rd	Everett Rd	NaT	2018-08-20 00:08:12
4	929920	Weather	NaN	1	0	NaN	NaN	NaT	2018-08-20 00:29:31

	eventid	eventtype	road	roaddirection	MaxOftrafficimpact	startcrossstreet	endcrossstreet	responsestart	createtime
12538	945924	Roadwork	Kenosha Rd	1	4	IL Rte 173	21st St	2018-09-10 09:01:41	2018-08-24 09:01:26
12539	945924	Roadwork	Kenosha Rd	1	4	IL Rte 173	IL Rte 173	2018-09-10 09:01:41	2018-08-24 09:01:26
12540	945924	Roadwork	Kilbourne Rd	1	4	IL Rte 173	21st St	2018-09-10 09:01:41	2018-08-24 09:01:26
12541	945924	Roadwork	Kilbourne Rd	1	4	IL Rte 173	IL Rte 173	2018-09-10 09:01:41	2018-08-24 09:01:26

과제 1 데이터 소개

```
df.info()
```

✓ 0.0s

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 12543 entries, 0 to 12542
Data columns (total 9 columns):
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
 0   eventid             12543 non-null  int64
 1   eventtype           12543 non-null  object
 2   road                11680 non-null  object
 3   roaddirection       12543 non-null  int64
 4   MaxOftrafficimpact  12543 non-null  int64
 5   startcrossstreet    7584 non-null   object
 6   endcrossstreet      7584 non-null   object
 7   responsestart       1574 non-null   datetime64[ns]
 8   createtime          12543 non-null  datetime64[ns]
```

1. 9개의 컬럼
2. 정수, 문자열, 시간 타임 존재
3. Null 값이 존재하는 것 확인
4. 총 12542개
5. 사고 데이터 같은 경우는 약 한달 치 (8월20일부터 9월20일)

과제 1 데이터 소개

```
df.describe()
```

✓ 0.0s

	eventid	roaddirection	MaxOftrafficimpact
count	12543.000000	12543.000000	12543.000000
mean	935674.119031	1.869409	1.928167
std	3347.696768	1.460207	0.941702
min	929916.000000	0.000000	0.000000
25%	932779.500000	1.000000	1.000000
50%	935683.000000	2.000000	2.000000
75%	938564.500000	3.000000	3.000000
max	958076.000000	4.000000	4.000000

1. Eventid : 그냥 93000가 많이 존재, 사고가 90000 발생이라고 보기는 어려움 확인

과제 1 데이터 소개

각 열별 널 값의 수 계산

```
# 각 열별 널 값의 수 계산
null_counts = df.isnull().sum()

print(null_counts)
```

✓ 0.0s

```
eventid                    0
eventtype                  0
road                     863
roaddirection              0
MaxOftrafficimpact         0
startcrossstreet        4959
endcrossstreet          4959
responsestart          10969
createtime                 0
dtype: int64
```

```
drop_df = df.drop(columns= ["startcrossstreet","endcrossstreet","responsestart" ])
```

✓ 0.0s

1. Null값 확인 : 거의 50%에 가까운 것과 80% 넘는 경우는 데이터 분석 시 필수적인 열이 아니여서 삭제 후 진행

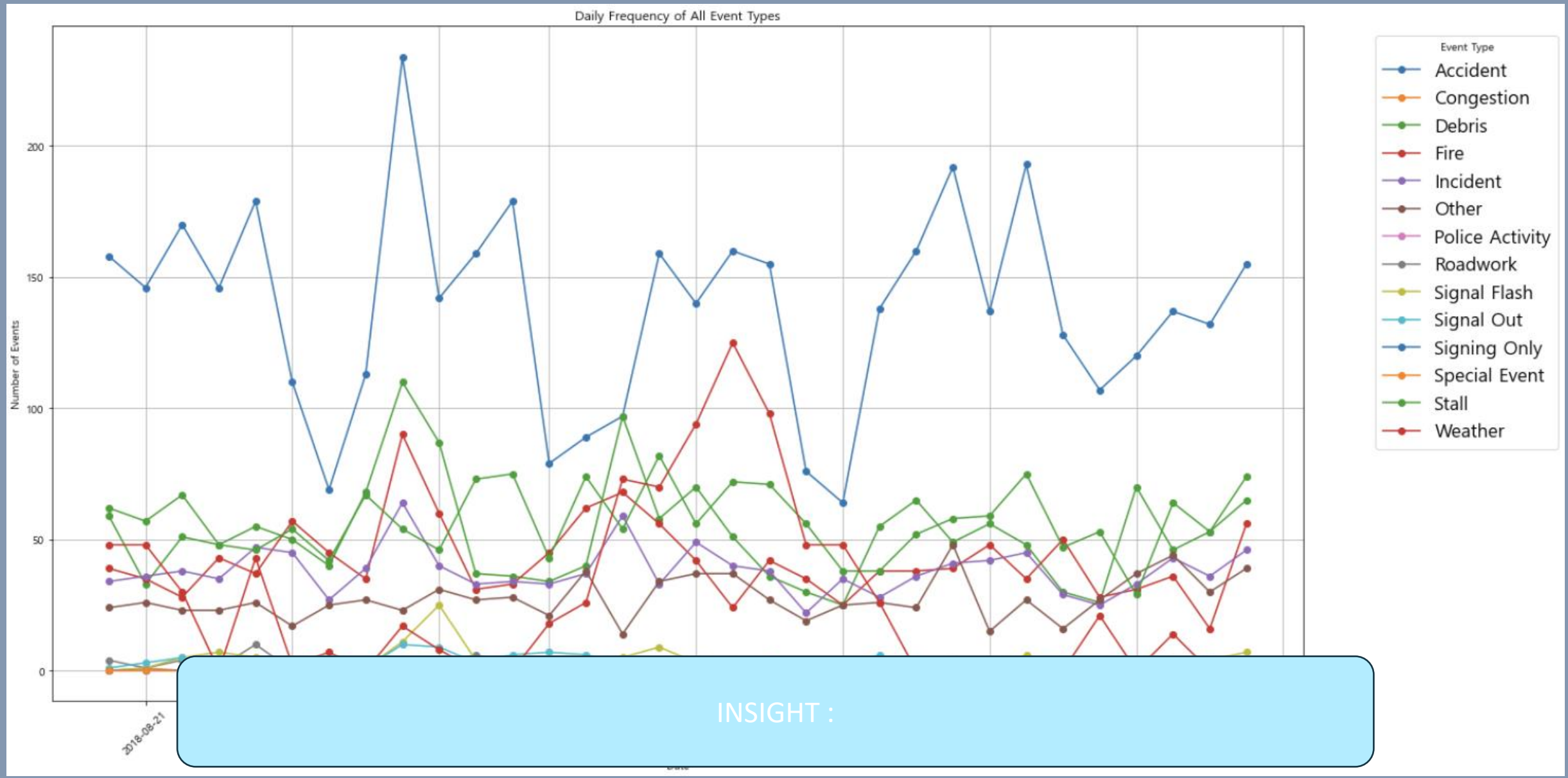
과제 1 데이터 소개

```
eventtype 열의 고유 값과 빈도수:
Accident            4423
Stall               1830
Debris              1687
Fire                1347
Incident            1222
Other                885
Weather              864
Signal Flash         109
Signal Out            85
Roadwork              52
Congestion            17
Police Activity       11
Signing Only           8
Special Event          3
Name: eventtype, dtype: int64
```

1. Eventtype : 어떤 이벤트가 존재했는지 확인하고 횟수 확인
   -> 사고가 가장 많은 것을 확인

과제 1

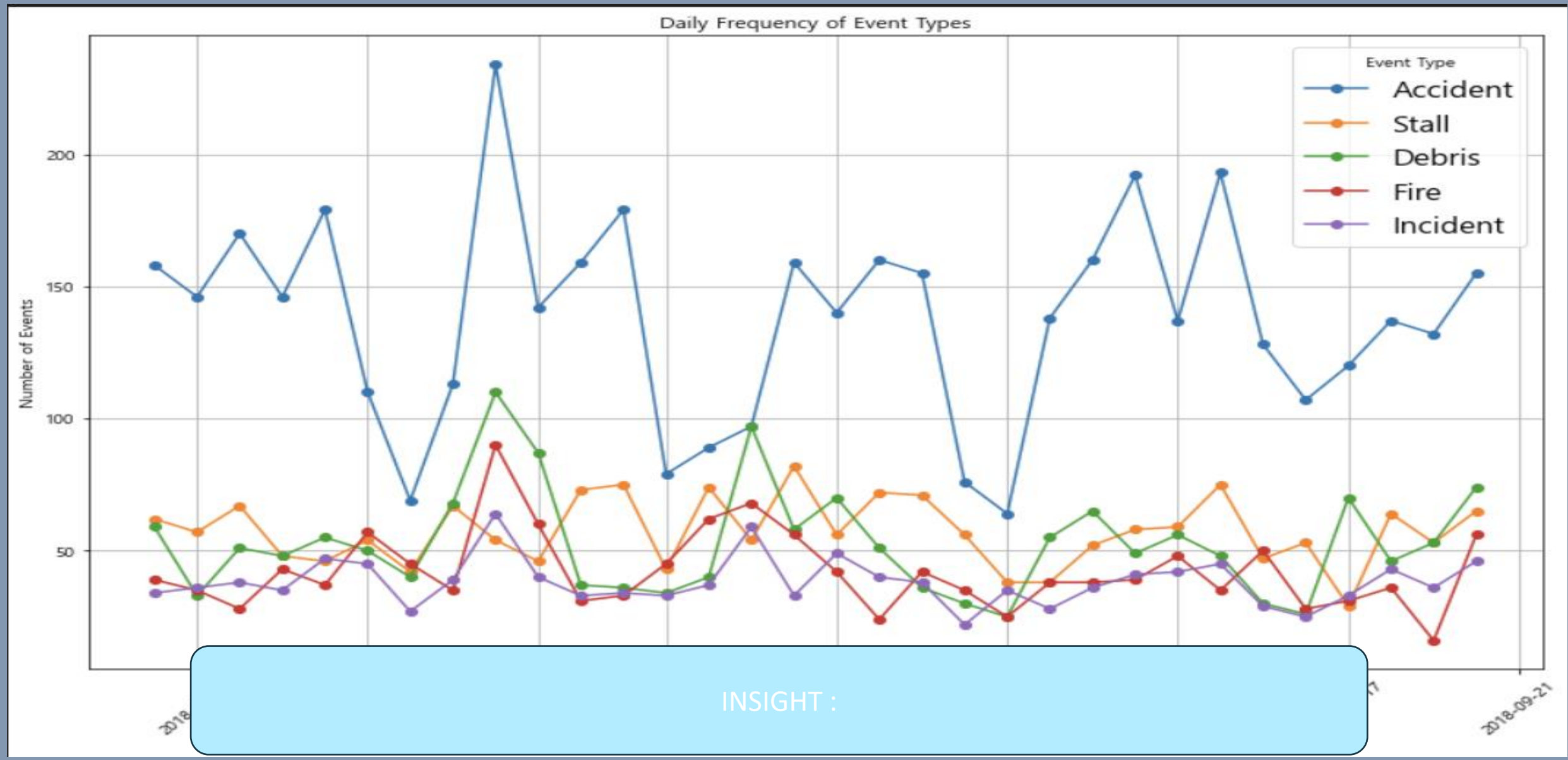
시간에 따른 이벤트 타입 별 수치화 x 축 시간 y축 사건 타입 4가지(꺽은선 그래프)



INSIGHT :

과제 1

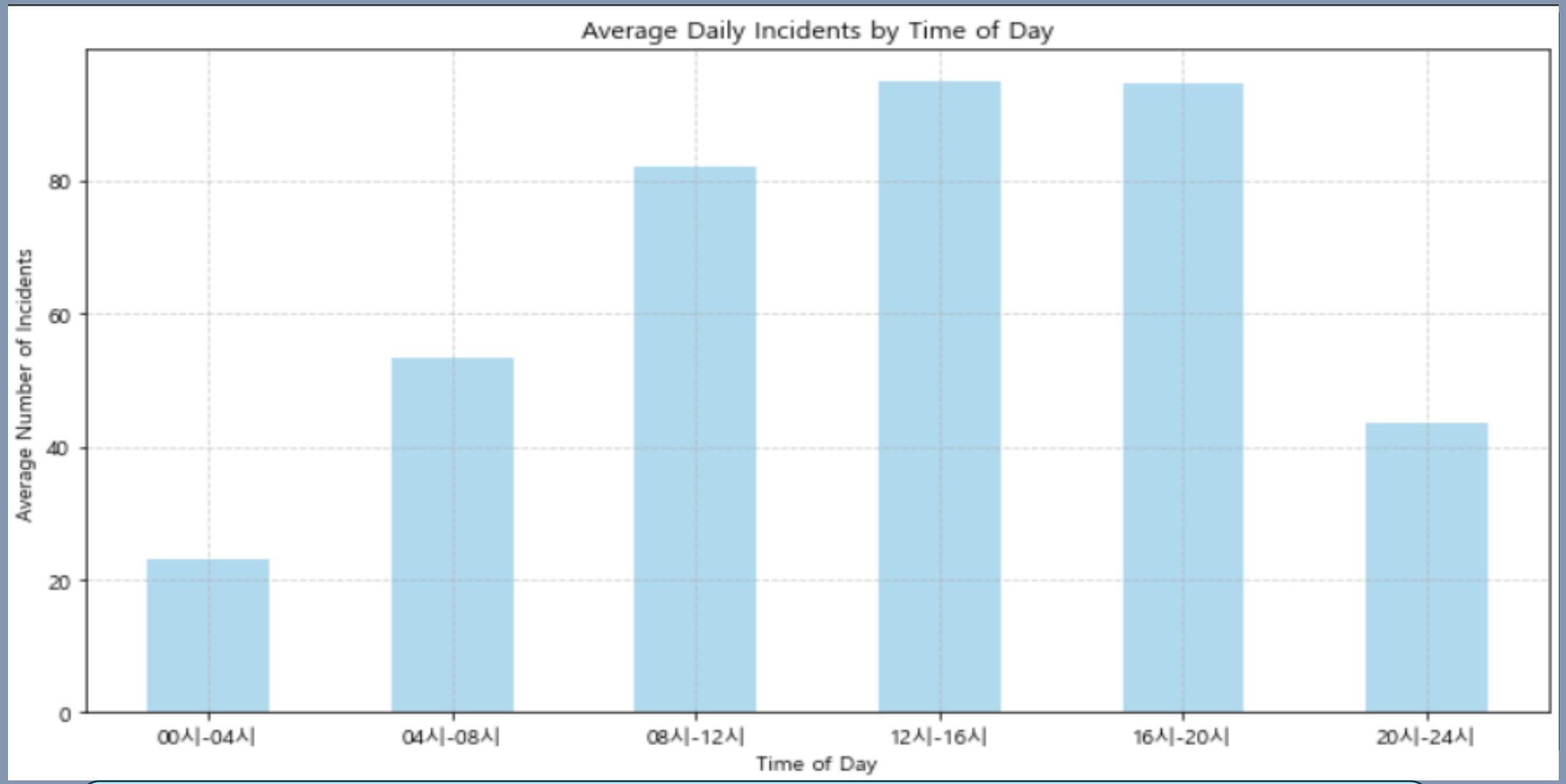
시간에 따른 이벤트 타입 별 수치화 x 축 시간 y축 사건 타입 4가지(꺽은선 그래프)



INSIGHT :

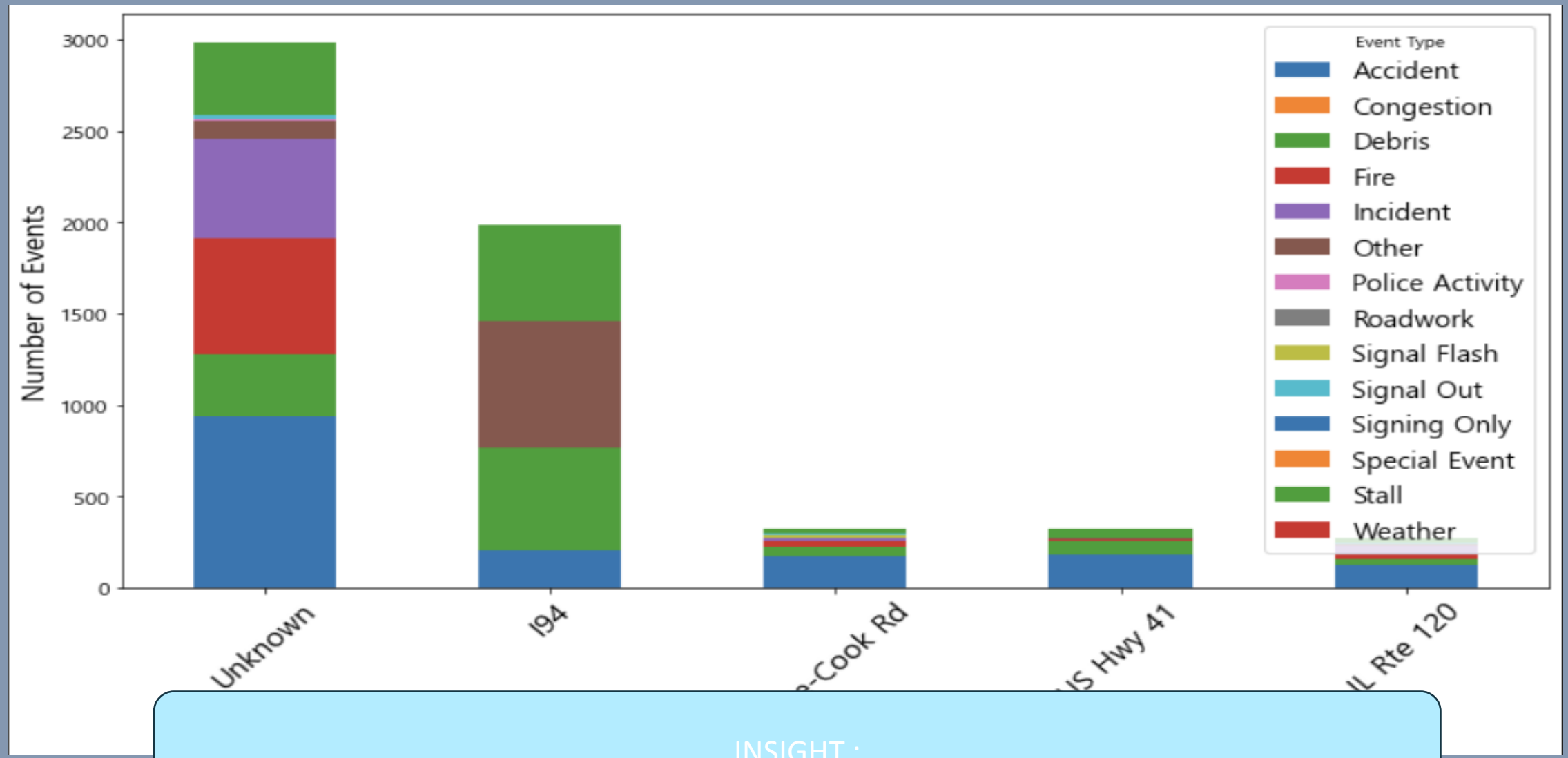
과제 2

시간에 따른 교통 사고 수치화 x축 시간(00시 04시 08시 12시 16시 20시 00시)(일 평균내기), y축 사건 발생 수 (bar 차트)



INSIGHT :

과제 3 장소에 따른 교통 사건 타입 수치화 x축 지역 이름 y축 사건 타입(bar 차트)



INSIGHT :

Part 1

데이터 전처리

3

문제점

문제점

1. 사고 데이터셋은 약 30일정도의 기록만 존재 -> 현재의 존재하는 데이터와 일치하지않는 문제

2. 사건 데이터 엑셀과 row data 사진 이름과 다르다. 즉 이름 형식을 같게 전처리 필요

3. GPU 용량 부족

4.사고 데이터 엑셀에서 지역 부분 컬럼에서 unknown data 정말 많다는 점

4

향후 연구 계획

향후 연구 지향점

4번 5번 성공 후 ->

5번에 대한 이해 필요해 질의 ? ?

교통 상황에 대한 정보를

얻고 사고 상황에 대한 인지

과제 4

과제 5

과제 6

Q & A