Multidimensional analytical framework for risky driving behaviors of commercial vehicles using data cube and interpretable machine learning techniques

데이터 큐브와 해석 가능한 머신 러닝 기술을 사용하여 상용차의 위험 운전 행동에 대한 다차원 분석 프레임워크

Joobin Jin

Dept. of AI and Bigdata, SCH Univ.

jjb0821@sch.ac.kr

Contents

1. OLAP technology
2. Data
3. Star schema and Concept hierarchies
4. Experiments and Results

OLAP (Online Analytical Processing: 온라인 분석 처리)

데이터 웨어하우스와 OLAP도구는 다차원적 데이터 모델을 기반으로 함

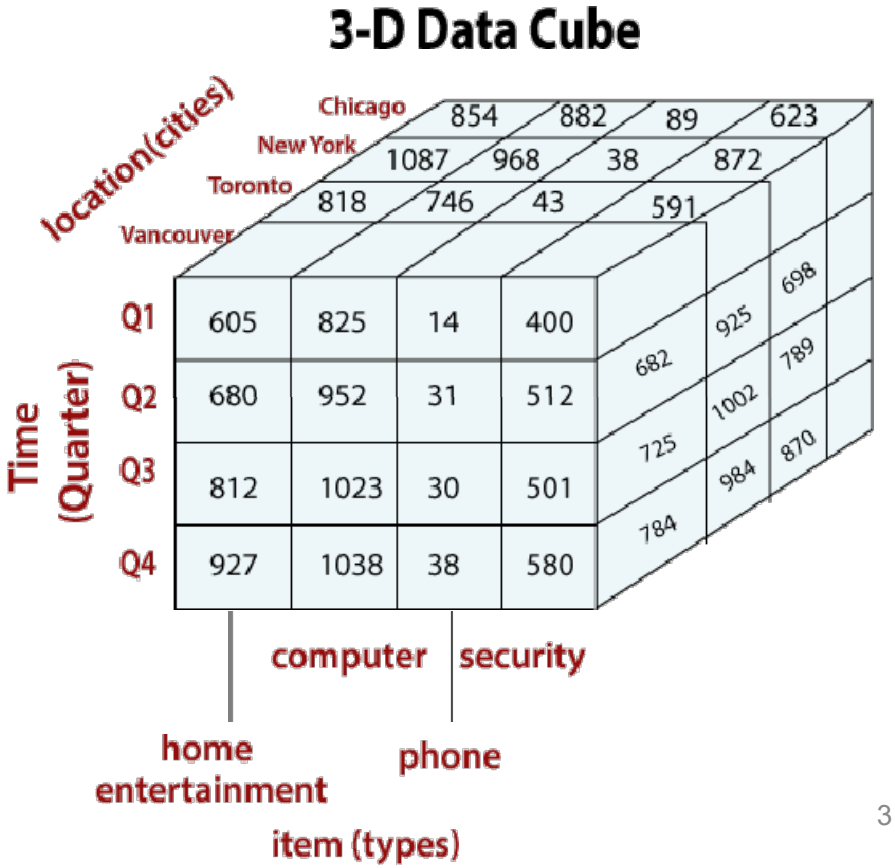
이 모델은 데이터 큐브의 형태로 데이터를 바라봄

Data cube

- 일반적으로 n차원 데이터를 모델링함
- 차원과 사실에 의해 정의됨
- 데이터를 여러 차원으로 모델링하고 볼 수 있음

3-D view of Sales Data

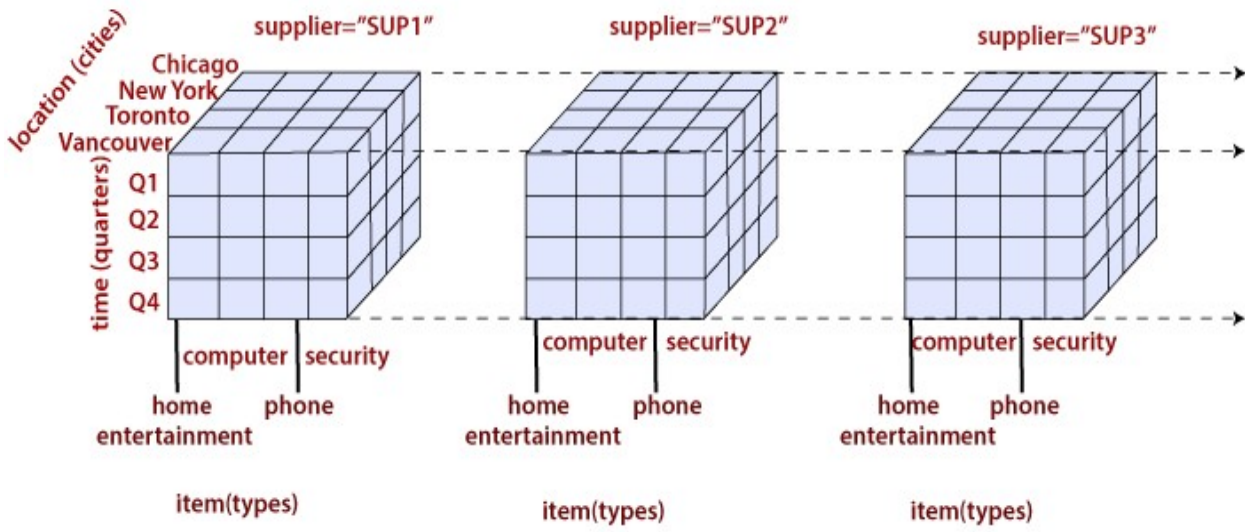
location ="Chicago"					**location** ="New York"				**location** ="Toronto"			
item					item				item			
time	home ent.	comp.	phone	sec.	time	home comp.	phone	sec.	home ent.	comp.	phone	sec.
Q1	854	882	89	623	1087	968	38	872	818	746	43	591
Q2	943	890	64	698	1130	1024	41	925	894	769	52	682
Q3	1032	924	59	789	1034	1048	45	1002	940	795	58	728
Q4	1129	992	63	870	1142	1091	54	984	978	864	59	784

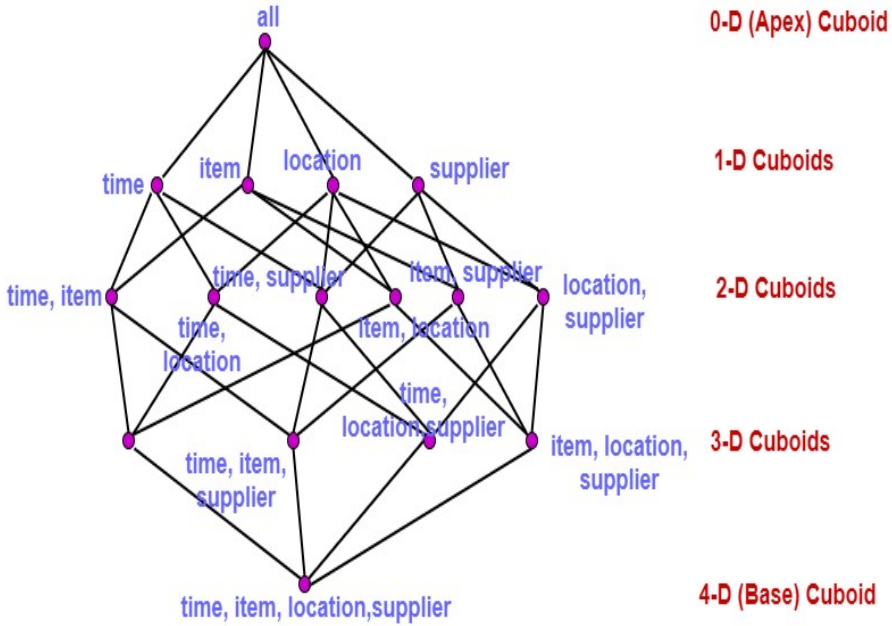


OLAP (Online Analytical Processing: 온라인 분석 처리)

Data cube

- 일반적으로 n차원으로 모델링 -> 3차원으로 제한하지 않음





OLAP (Online Analytical Processing: 온라인 분석 처리)

Star schema

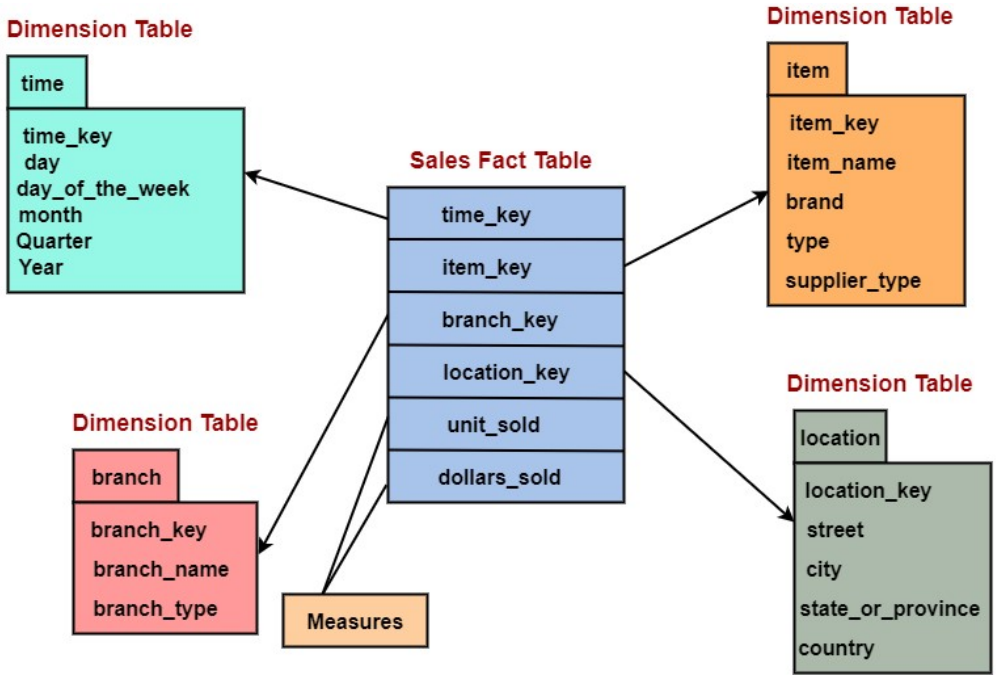
- 데이터가 **사실(Fact)**과 **차원(Dimension)**으로 구성되는 차원 모델의 기본 형태

사실테이블(Fact Table)

- 사실을 포함하고 차원에 연결된 테이블
- 사실을 포함하는 열과 차원에 대한 외래키를 포함
- 기본키는 모든 외래키로 구성된 복합키

차원테이블(Dimension Table)

- 데이터를 분류, 하나 이상의 계층 구조로 구성
- 속성은 일반적으로 설명이 포함된 텍스트 값
- 기본키는 사실테이블의 복합 기본키의 일부



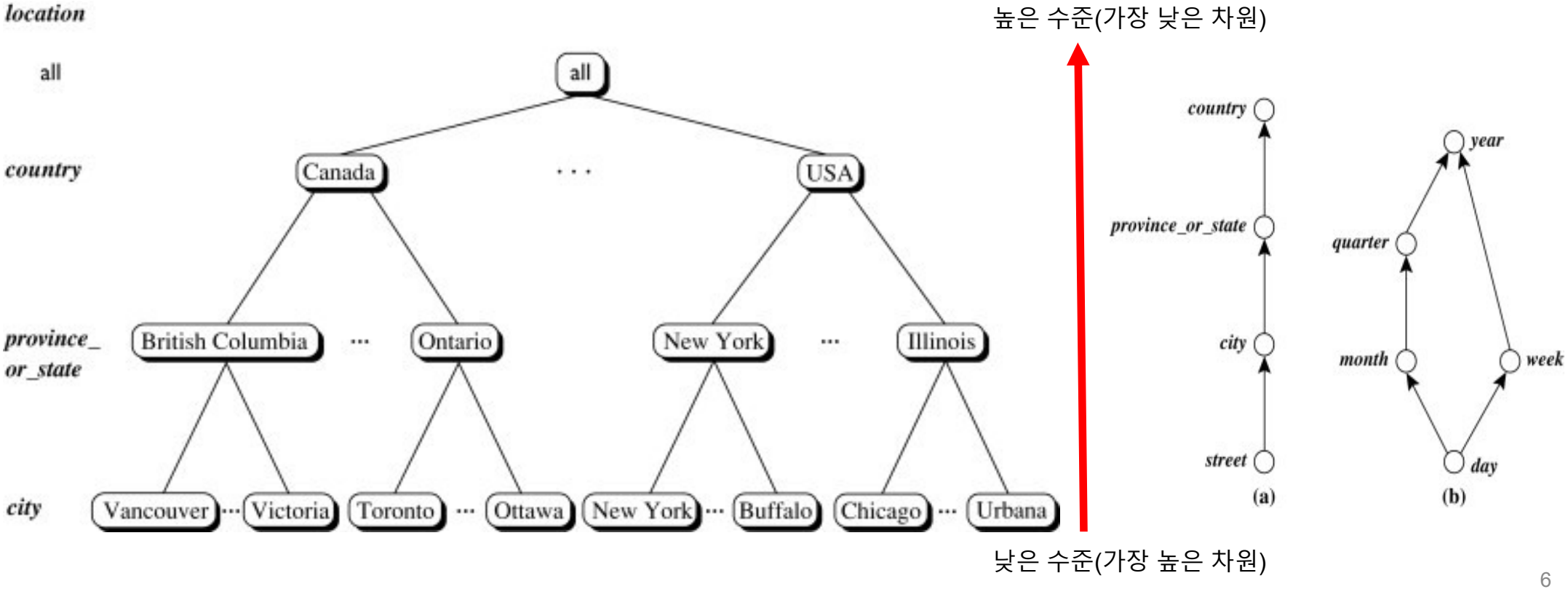
오른쪽 그림은 4개의 차원테이블과 1개의 사실테이블로 정의

이외에도 Snowflake schema, Fact constellation이 있음

OLAP (Online Analytical Processing: 온라인 분석 처리)

Concept hierarchies

- 낮은 수준의 개념 집합에서 더 높은 수준의 보다 일반적인 개념으로의 매핑 시퀀스를 정의



OLAP (Online Analytical Processing: 온라인 분석 처리)

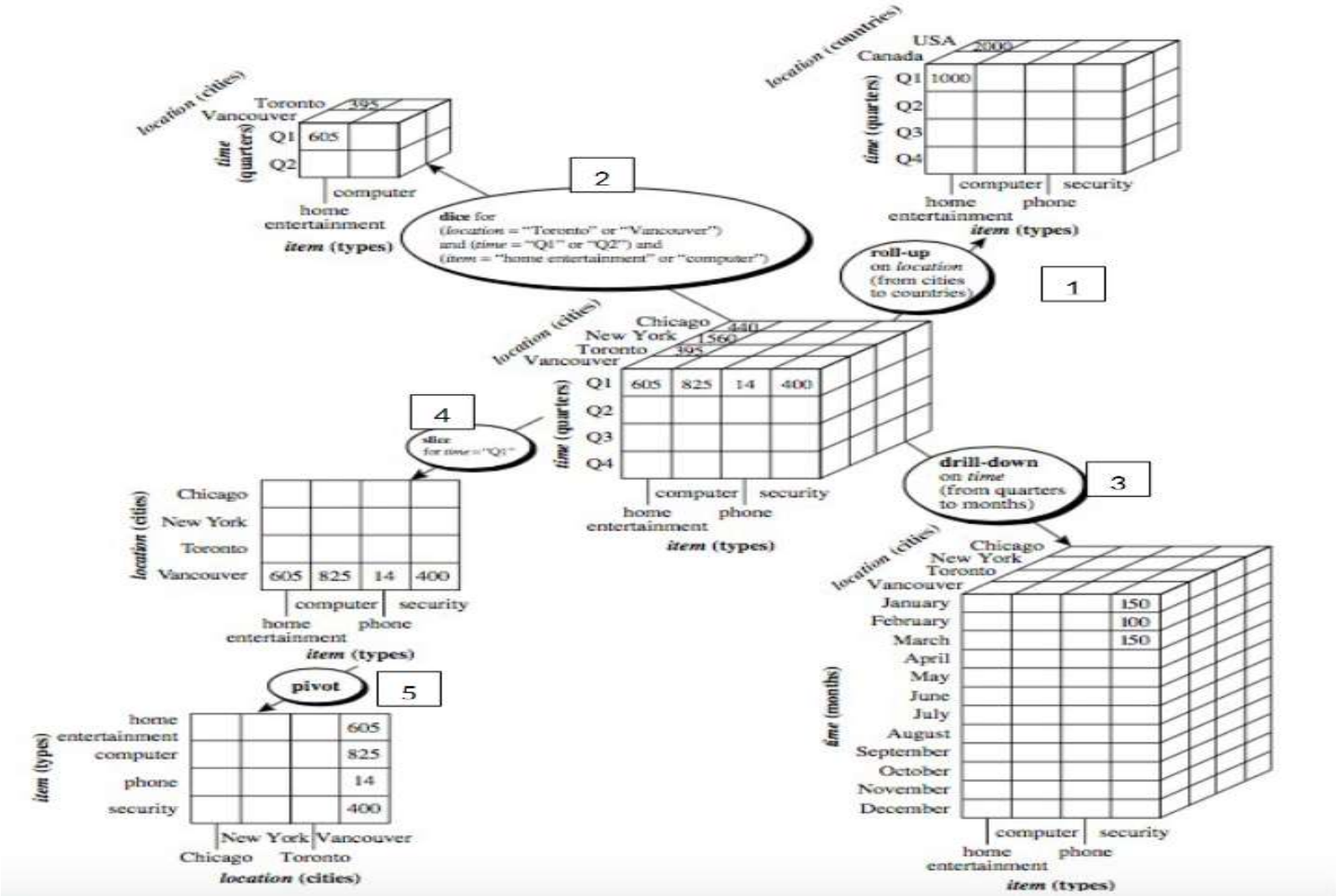
OLAP에서 개념 계층이 어떻게 적용되는지

- 다차원 모델에서 데이터는 여러 차원으로 구성됨
- 각 차원에서는 개념 계층에 의해 정의된 여러 수준의 추상화가 포함 #추상화 = 세부사항 수준
- 다양한 관점에서 데이터를 볼 수 있는 유연성을 제공
- 이러한 다양한 보기를 구체화하기 위해 많은 OLAP Data cube 연산이 존재

OLAP (Online Analytical Processing: 온라인 분석 처리)

OLAP Operations

1. Roll-up
2. Dice
3. Drill-down
4. Slice
5. Pivot



Data

본 연구에서 사용한 데이터

- 2019년 대전광역시의 상업용차량 운행 데이터(DTG)
  (날짜, 시간, 운수회사코드, 차량번호, 위도, 경도, 11대위험운전유무) 약 1초단위
- 2019년 전국 교통사고 데이터
  (92개의 columns 존재)
- 대전광역시의 법정동 경계면 Polygon데이터
  (법정동 코드, 법정동명, Ploygon)

Data Preprocessing

DTG 데이터 전처리

- 1개월씩 12개의 2019년 데이터 존재

  -> 1월 데이터에 다른 월도 껴있고 섞여있어서 Datetime변환 후 각 월별로 재분할

  -> 20km과속여부 + 40km과속여부 + 60km과속여부 = 과속으로 정의

  -> 시간 변수가 HHMISSSS형식의 8자리로 나타나서 HH만으로 시간 변수 재추출

  -> 같은 운송회사 코드이지만 다르게 분류돼 있어서 {01374, 1374}처럼 통일화

  -> 위도, 경도 변수를 이용해 공간 데이터 포인트 생성

DTG데이터 용량이 매우 크므로 전처리 후 필요하지 않은 변수들 제거

DTG Data

	날짜	시간	운수회사	과속	급가속	급출발	급감속	급정거	급좌회전	급우회전	급유턴	급앞지르기	급차선변경
0	2019-01-01	0	25959	0	0	0	0	0	0	0	0	0	0
1	2019-01-01	1	17797	0	0	0	0	0	0	0	0	0	0
2	2019-01-01	1	17797	1	0	0	0	0	0	0	0	0	0
3	2019-01-01	1	17797	0	0	0	0	0	0	0	0	0	0
4	2019-01-01	1	17797	0	0	0	0	0	0	0	0	0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...
16754543	2019-01-31	17	01374	0	0	0	0	0	0	0	0	0	0
16754544	2019-01-31	17	01374	0	0	0	0	0	0	0	0	0	0
16754545	2019-01-31	17	01374	0	0	0	0	0	0	0	0	0	0
16754546	2019-01-31	15	01374	0	0	0	0	0	0	0	0	0	0
16754547	2019-01-31	19	01374	0	0	0	0	0	0	0	0	0	0

Data Preprocessing

2019년 전국 교통사고 데이터 전처리

- 92개의 columns 존재

  ->날짜, 시간, 요일, 법정동 코드, 사고장소, 사고심각도, 사망자수, 중상자수, 경상자수, 부상자수,

  위도, 경도 만 추출

  -> 대전광역시의 법정동 코드는 30XX..로 시작하므로 30으로 시작하는 법정동 코드 필터링 (8337개)

  -> 위도, 경도 변수를 이용한 공간 데이터 포인트 생성

Accident Data

	날짜	시간	요일	법정동코드	사고심각도	사망자수	중상자수	경상자수	부상자수
0	2019-01-01	18	3	3.020014e+09	3	0	0	1	0
1	2019-01-01	18	3	3.017011e+09	3	0	0	2	0
2	2019-01-01	13	3	3.014010e+09	3	0	0	1	0
3	2019-01-01	12	3	3.011010e+09	2	0	1	0	0
4	2019-01-01	20	3	3.017011e+09	3	0	0	1	0
...	...	...	...	...	...	...	...	...	...
8332	2019-12-31	20	3	3.014011e+09	3	0	0	3	2
8333	2019-12-31	16	3	3.017011e+09	2	0	1	0	0
8334	2019-12-31	15	3	3.011011e+09	2	0	1	7	0
8335	2019-12-31	22	3	3.020014e+09	3	0	0	1	0
8336	2019-12-31	13	3	3.020012e+09	3	0	0	3	

Polygon Data

	법정동 코드	시군 필드	법정동	geometry
0	30110301	30110	원동	POLYGON ((127.43429 36.32904, 127.43431 36.329...
1	30110302	30110	인동	POLYGON ((127.43909 36.32427, 127.43898 36.324...
2	30110303	30110	효동	POLYGON ((127.44155 36.31926, 127.44162 36.319...
3	30110304	30110	천동	POLYGON ((127.44675 36.31952, 127.44680 36.319...
4	30110305	30110	가오동	POLYGON ((127.45415 36.31216, 127.45425 36.312...
...	...	...	...	...
172	30230122	30230	부수동	POLYGON ((127.47616 36.44571, 127.47677 36.445...
173	30230123	30230	황호동	POLYGON ((127.50279 36.45550, 127.50280 36.455...
174	30230124	30230	삼정동	POLYGON ((127.46258 36.45616, 127.46329 36.455...
175	30230125	30230	미호동	POLYGON ((127.48423 36.47580, 127.48428 36.475...
176	30230126	30230	신탄진동	POLYGON ((127.43213 36.45662, 127.43379 36.456...

Total Data

DTG데이터와 2019년 전국 교통사고 데이터 merge

- DTG데이터의 공간포인터와 법정동 경계면 데이터를 공간조인
- 교통사고 데이터의 공간포인터와 법정동 경계면 데이터를 공간 조인
- 날짜, 시간, 법정동코드를 기준으로 공간조인된 2개의 데이터를 merge

Total Data

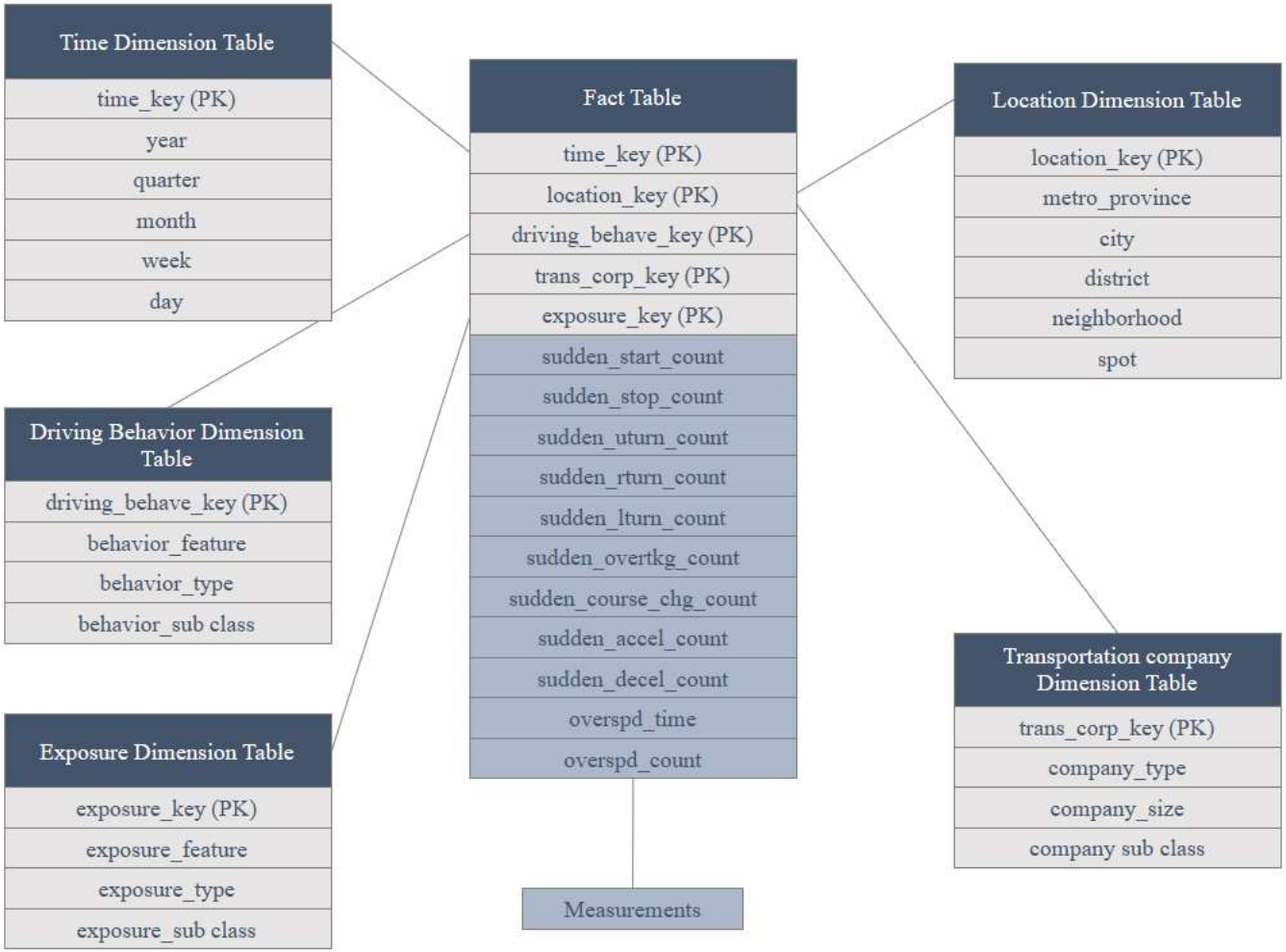
날짜	시간	법정동	과속	장기과속	급가속	급출발	급감속	급정거	급좌회전	급우회전	급유턴	급앞지르기	급차선변경
2019-01-01	3	방현동	1	0	1	0	3	0	0	1	0	0	0
	12	가오동	7	0	62	0	13	3	6	3	2	5	7
	13	구암동	71	0	121	0	24	4	6	3	0	1	8
		목동	33	0	39	0	5	1	0	0	0	0	1
	18	둔산동	167	0	863	2	97	37	23	9	2	8	39
		반석동	0	0	29	0	2	1	6	2	0	0	8
		탄방동	46	0	356	2	33	6	7	7	1	1	12
	19	둔산동	230	0	863	3	103	29	24	10	3	11	43
	20	탄반동	94	0	390	1	47	14	5	1	0	2	9
2019-01-02	0	둔산동	121	0	192	1	27	5	1	1	0	2	0
		장대동	6	0	27	0	2	0	1	3	0	0	0
	6	봉명동	26	0	144	0	37	6	15	5	0	4	25
		송촌동	30	0	79	0	26	4	5	8	0	8	17

Star schema

본 연구에서 제안하는 스타 스키마

- 5개의 차원 테이블 -> Time, Location, Driving Behavior, Transportation company, Exposure
- 1개의 사실 테이블 -> 급출발, 급정거, 급유턴, 급우회전, 급좌회전, 급앞지르기, 급차선변경, 급가속, 급감속, 과속, 장기과속(11대위험운전)의 측정값(Distributive)

Star schema


Time Dimension Table
time_key (PK)
year
quarter
month
week
day
Fact Table
time_key (PK)
location_key (PK)
driving_behave_key (PK)
trans_corp_key (PK)
exposure_key (PK)
sudden_start_count
sudden_stop_count
sudden_uturn_count
sudden_rturn_count
sudden_lturn_count
sudden_overtkg_count
sudden_course_chg_count
sudden_accel_count
sudden_decel_count
overspd_time
overspd_count
Location Dimension Table
location_key (PK)
metro_province
city
district
neighborhood
spot
Driving Behavior Dimension Table
driving_behave_key (PK)
behavior_feature
behavior_type
behavior_sub class
Exposure Dimension Table
exposure_key (PK)
exposure_feature
exposure_type
exposure_sub class
Transportation company Dimension Table
trans_corp_key (PK)
company_type
company_size
company sub class
Measurements

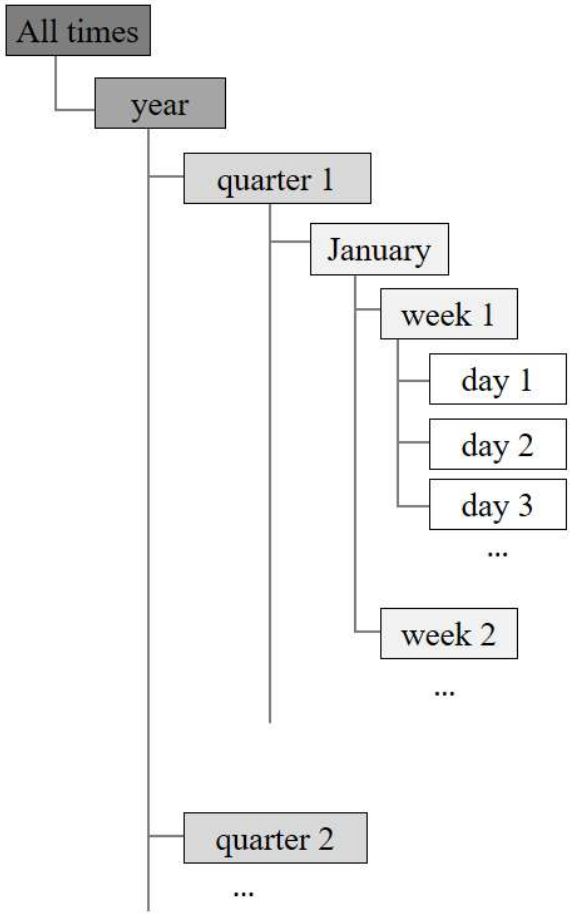
Concept hierarchies

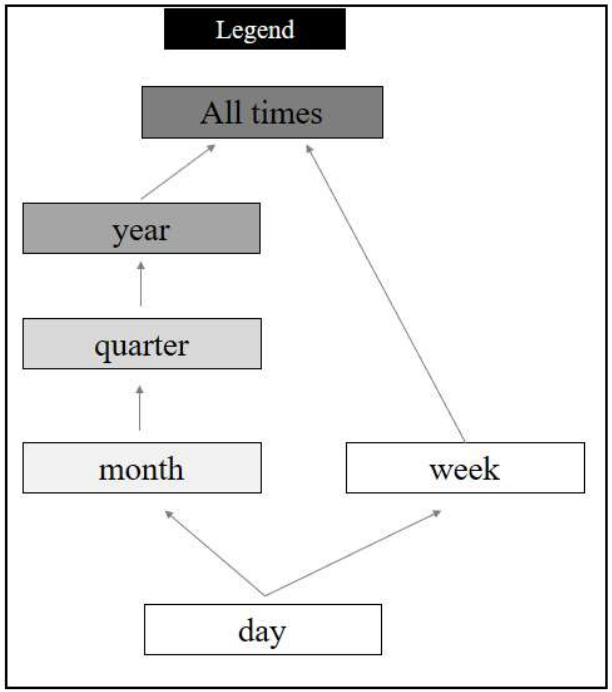
본 연구에서 제안하는 개념계층도

- 5개의 차원에 대한 개념계층도와 Legend를 표시

Concept hierarchies

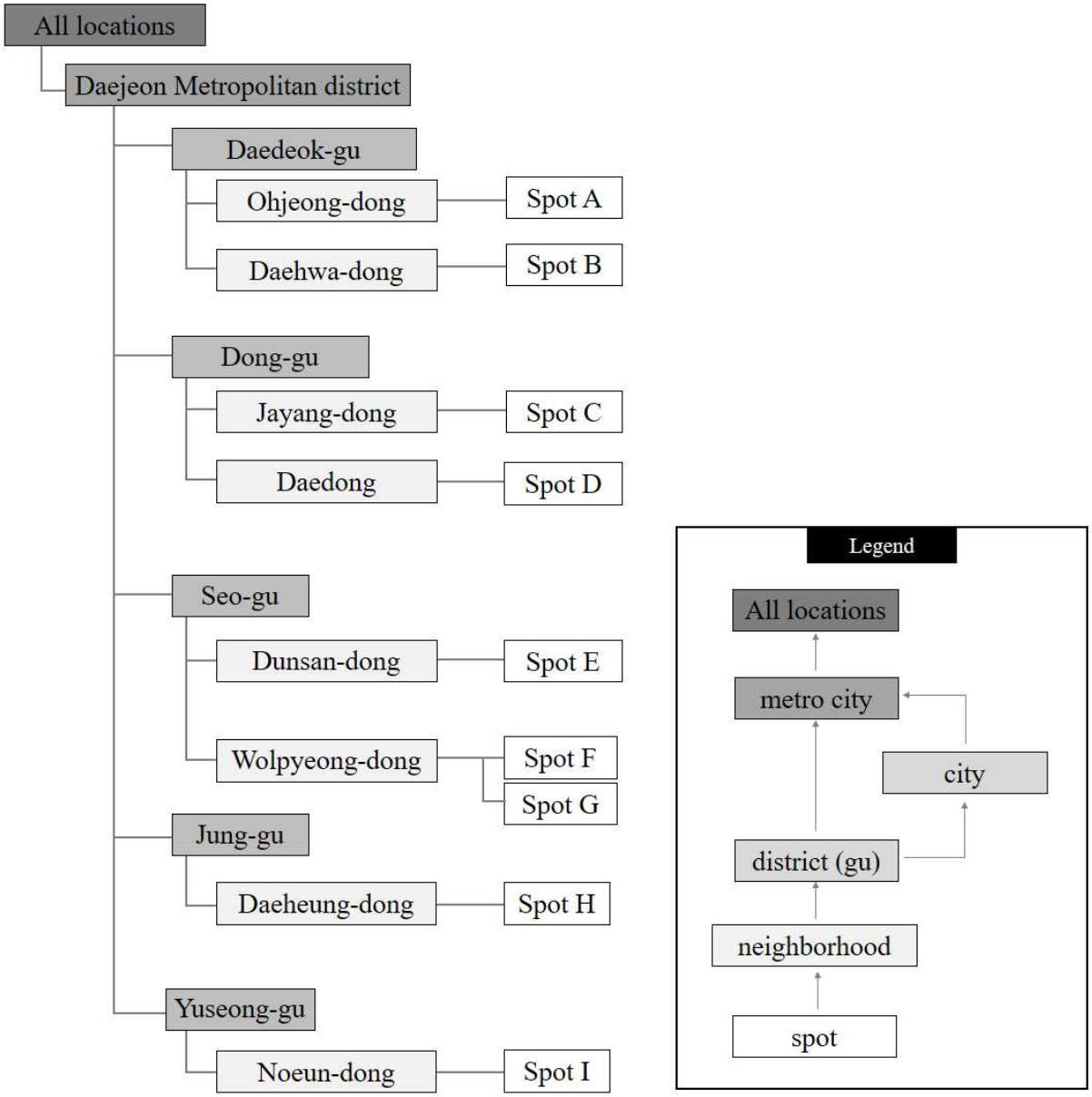
Time





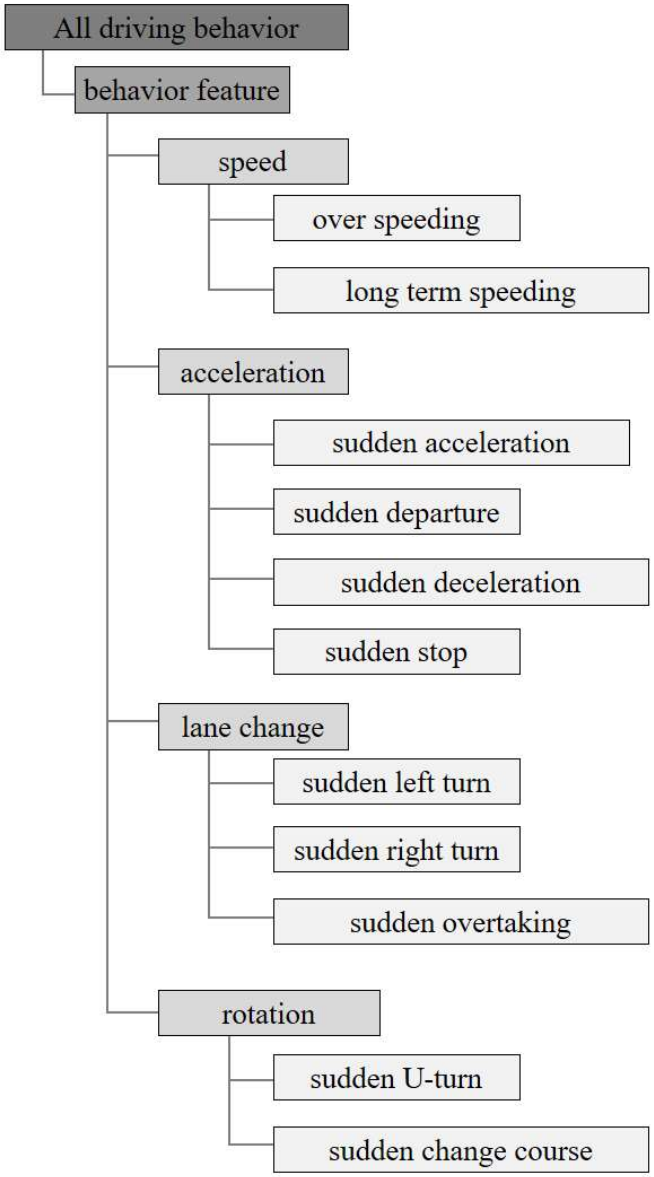
Concept hierarchies

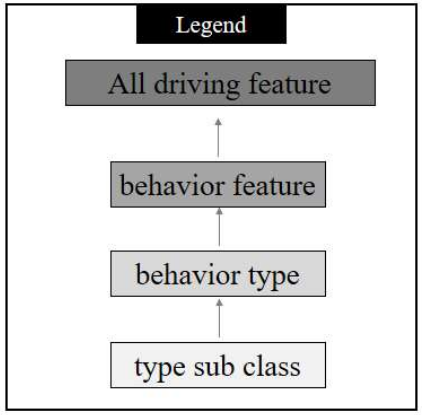
Location



Concept hierarchies

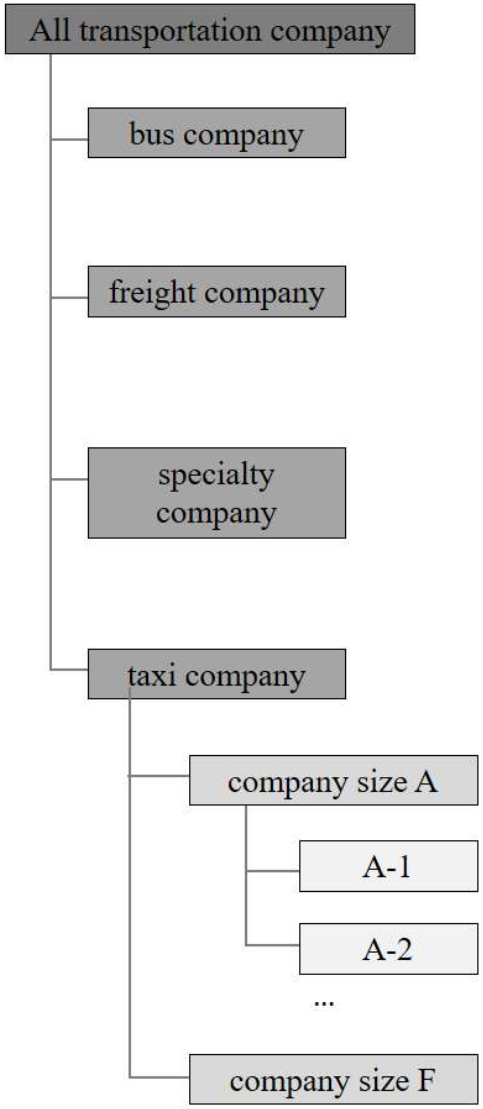
Driving Behavior

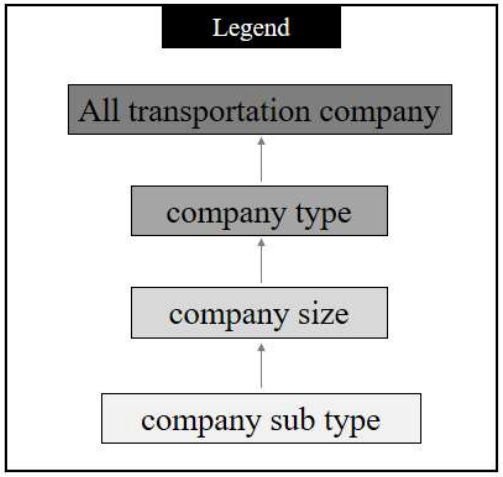




Concept hierarchies

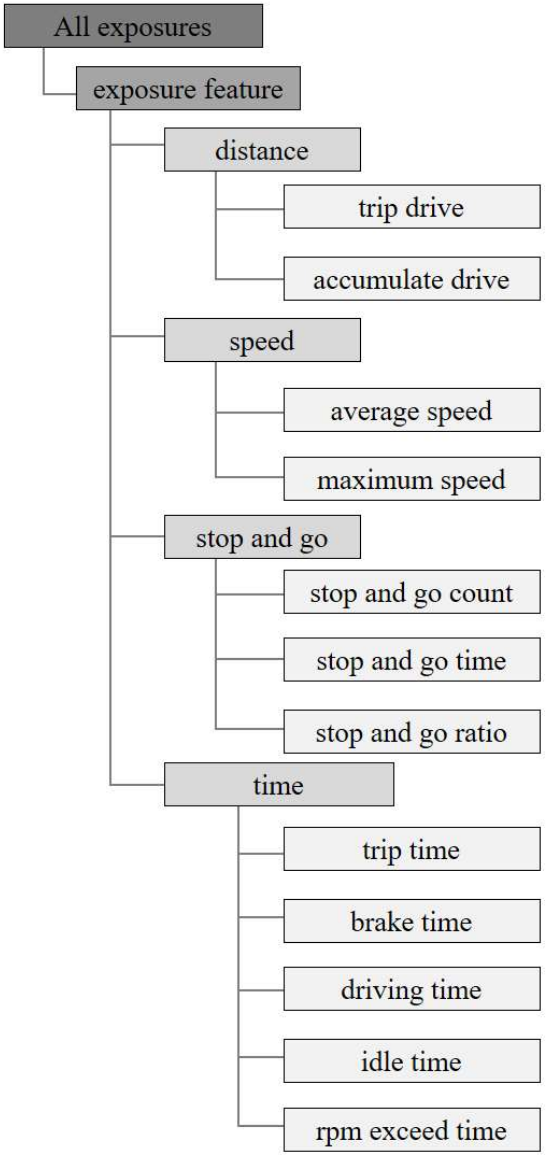
Transportation company

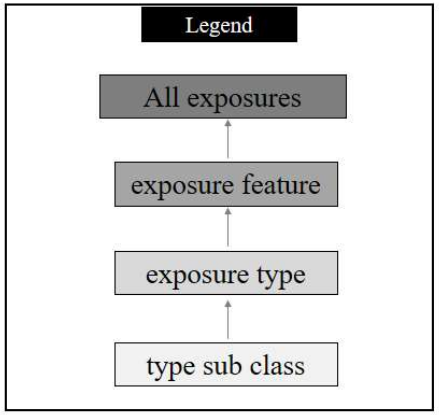




Concept hierarchies

Exposures





Scenario

시간별 분석

- 년 -> (분기) -> 월 -> (주) -> 일 -> 시간 의 형태로 추상화 수준을 달리하여 세부사항을 볼 수 있게 drill-down작업 진행 후 분석

공간별 분석

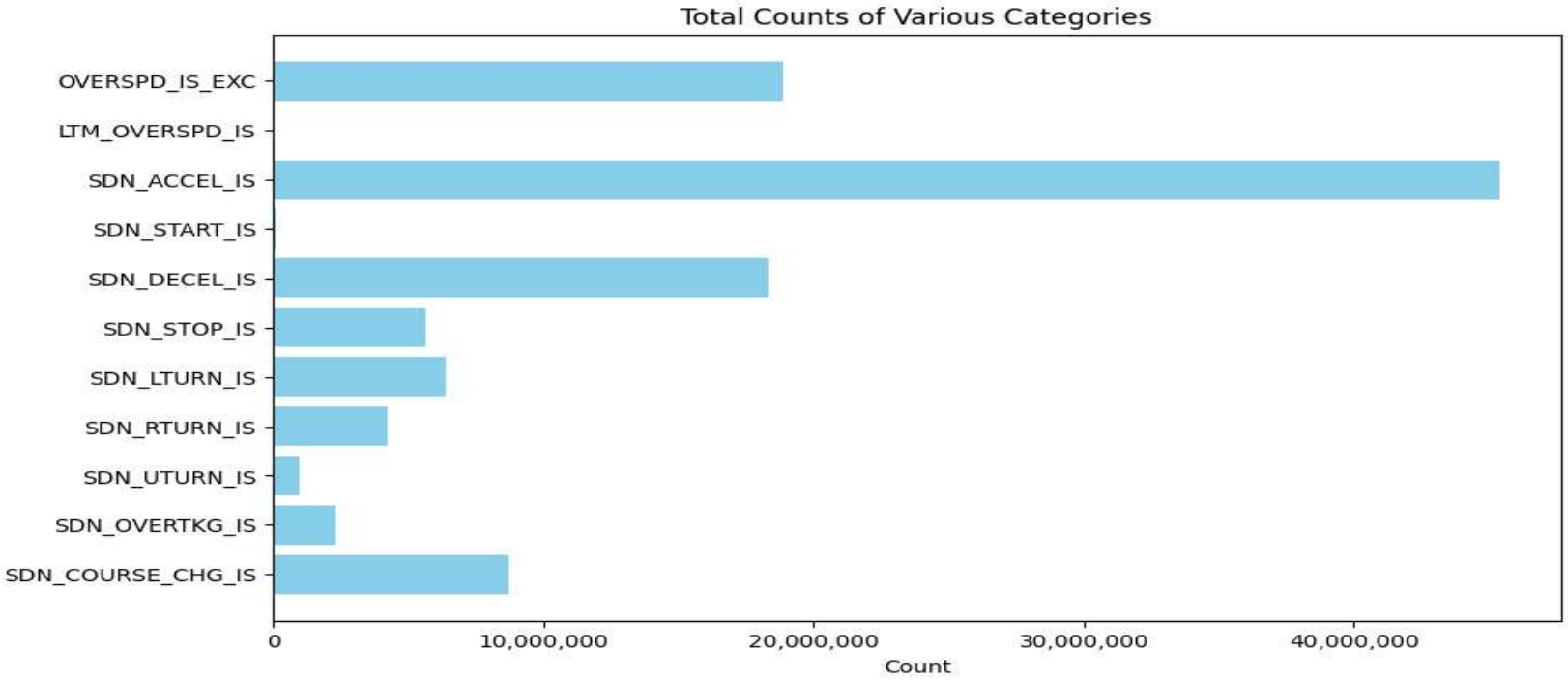
- 대전 -> 구 -> 동 의 형태로 추상화 수준을 달리하여 세부사항을 볼 수 있게 drill-down작업 진행 후 분석

시간별 + 공간별 분석

- 어떤 시간대에 어느 장소에서 교통사고가 자주 발생하는지 결과 도출 예상

Experiments and result

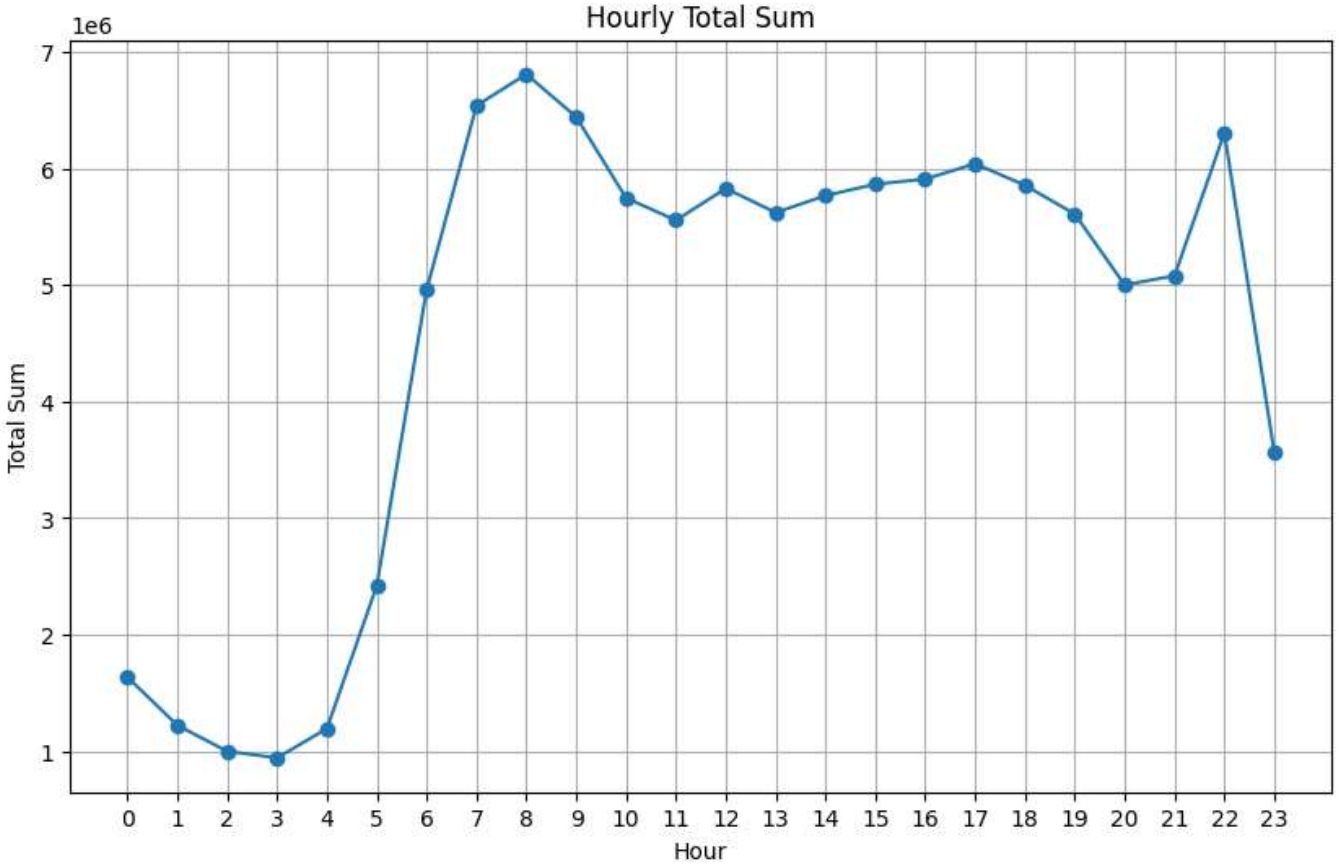
11대 위험운전 집계



- 급가속이 45,393,506회로 가장 빈도수가 높고 다음으로 과속 18,852,883회 급감속 18,296,282회

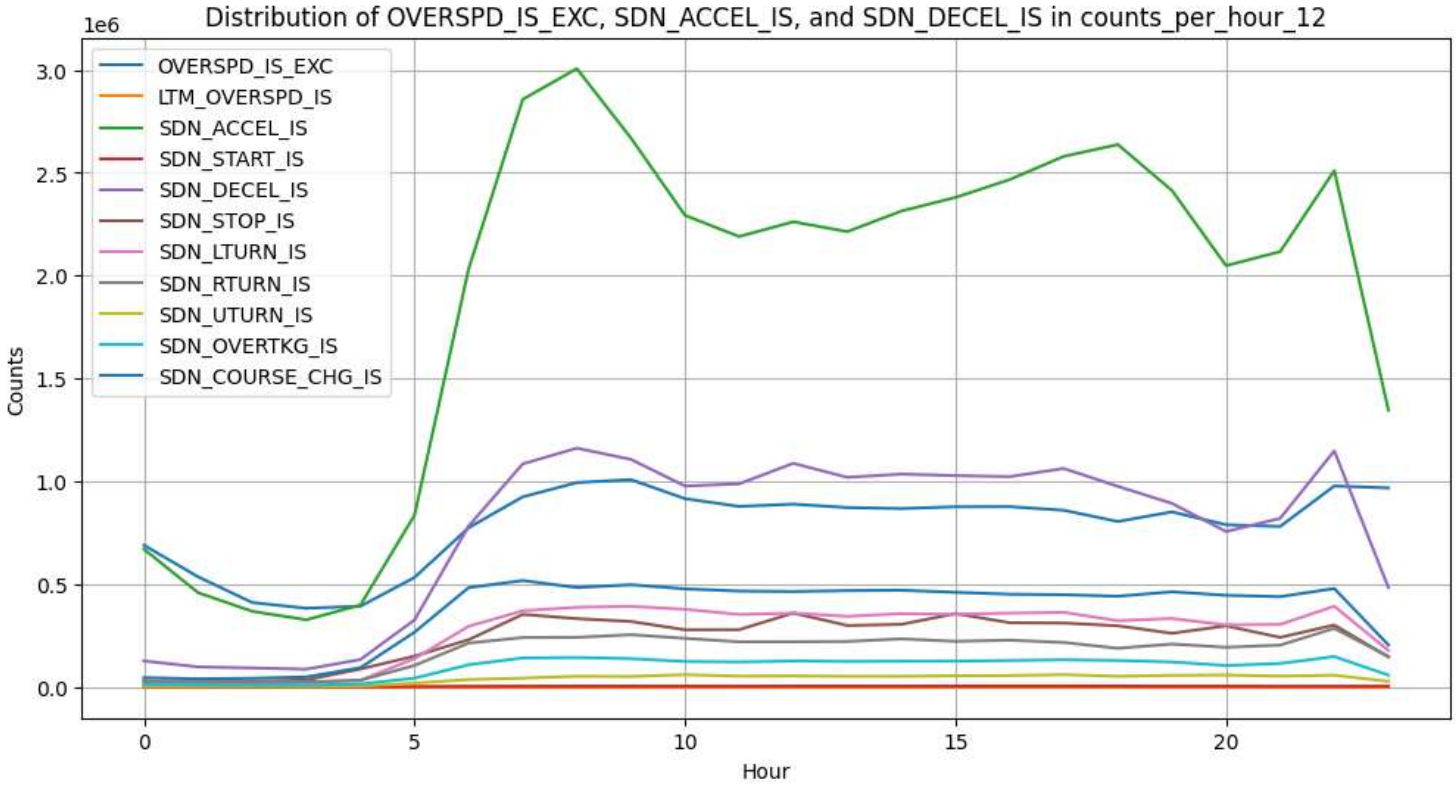
Experiments and result

시간별 11대 위험운전 총집계



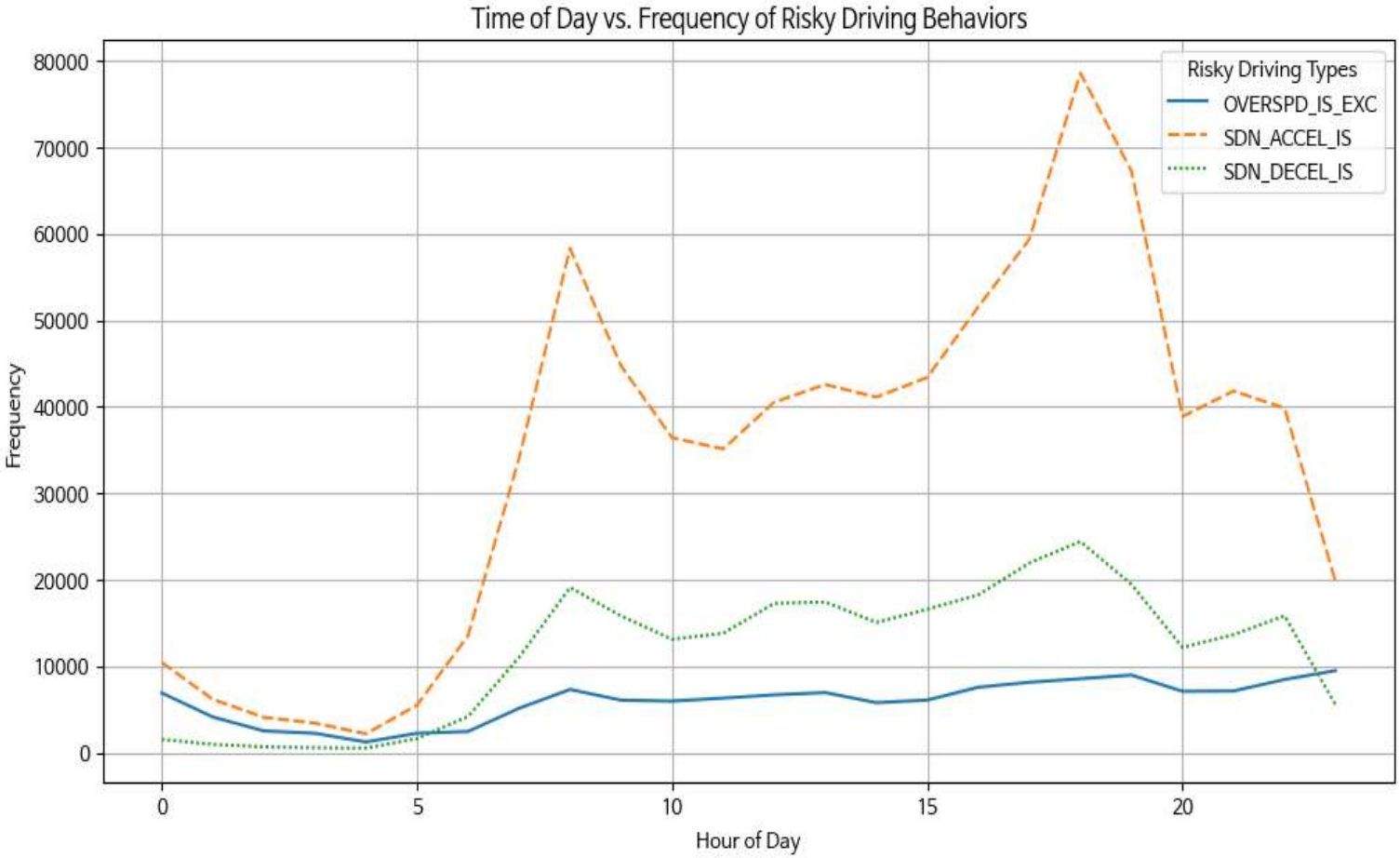
Experiments and result

시간별 11대 위험운전 총집계



Experiments and result

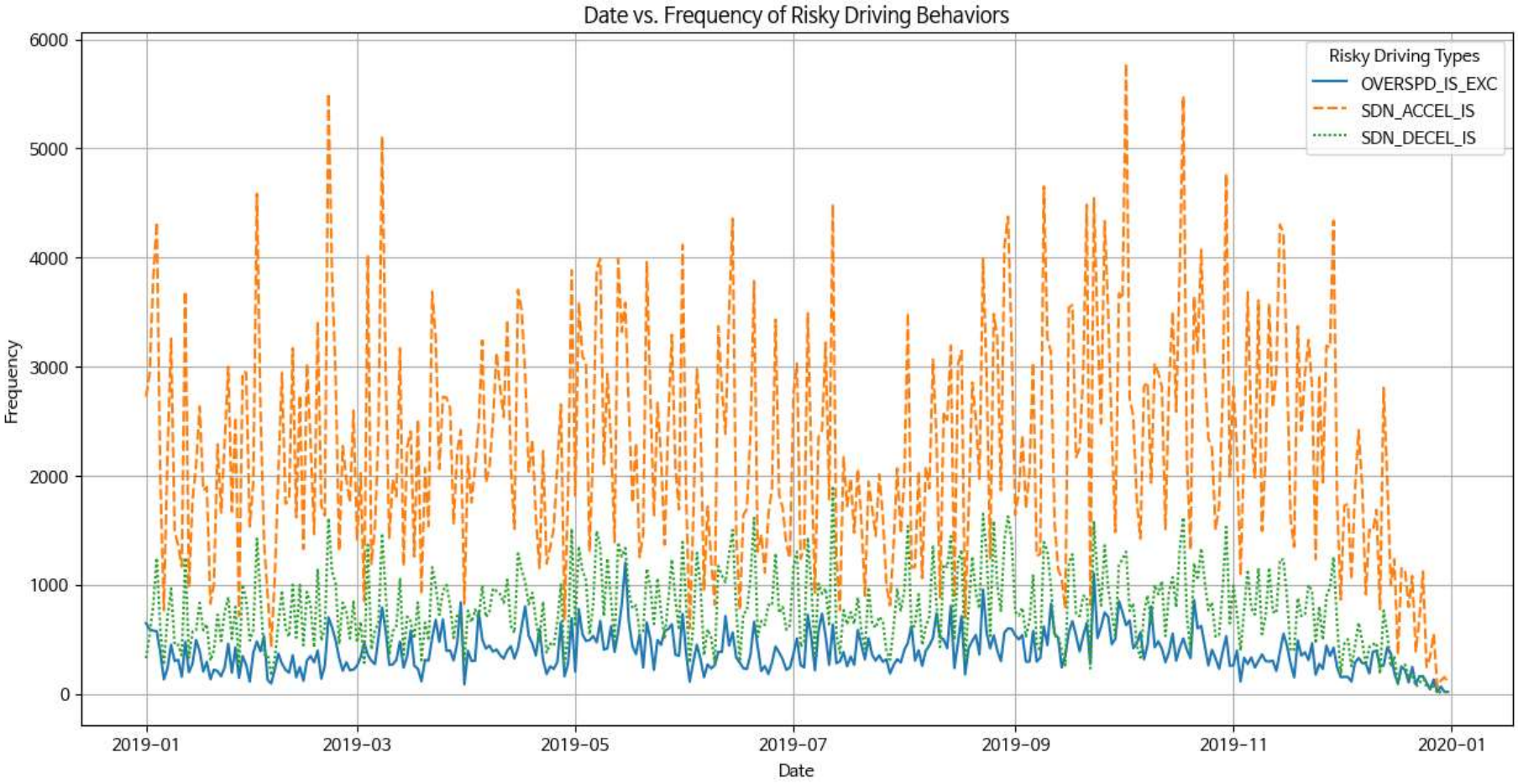
교통사고 발생 시간대의 주요 11대 위험운전 집계



- 18시에 급가속이 78,603회로 가장 높고, 역시 18시에 급감속이 24,428회로 가장 높게 나타남
- 하지만 과속은 비교적 차량의 통행이 없는 23시에 9,494회로 가장 높게 나타남.

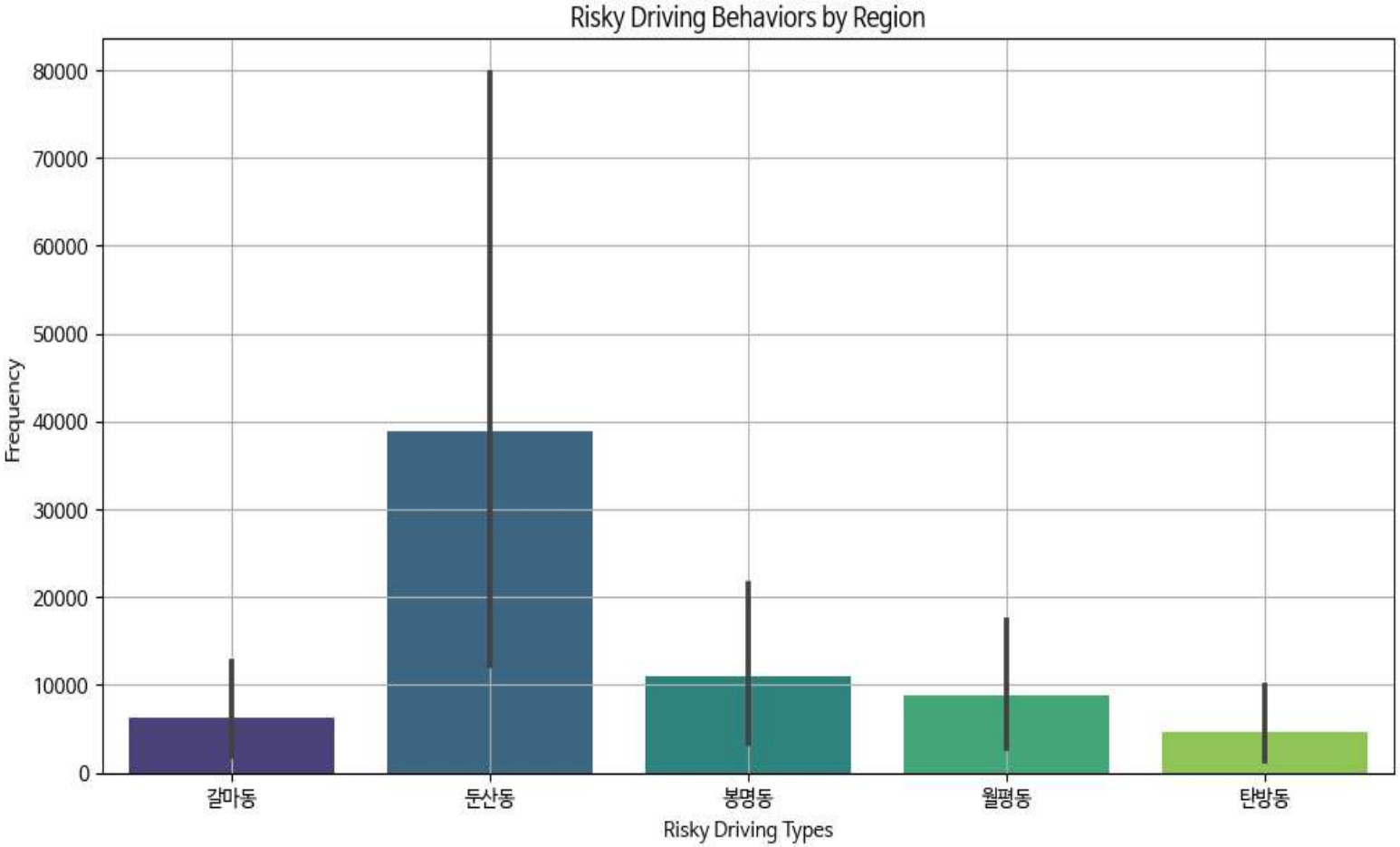
Experiments and result

11대 위험운전유형들의 날짜별 교통사고 발생 추이



Experiments and result

교통사고 빈도수 Top5 법정동 집계



- 교통사고 발생이 가장 많이 일어나는 법정동 상위5개의 지역을 추출해보니 둔산동의 11대위험운전 합계가 426,839회로 가장 높음
- 봉명동 120,002회
- 월평동 97,829회
- 갈마동 68,340회
- 탄방동 50,645회

End of slide