Exploratory Data Analysis

- StudentID: 21700411

- Name: ahndongsu

- 1st Major: 경영학부

- 2nd Major: 데이터사이언스

댈러스 도시의 시간, 공간, 요일 데이터를 활용하여 범죄 많이 일어날 시간, 장소 예상-> 범죄예방대시보드 구축 및 활용, 인력 적재적소 배치

1. Data overview

Descriptives statistics on overall data (sample size, number of variables, data type, data range, distribution, etc.)

- data shape : 663249행 ,107열
- 데이터 타입: float64(1), int16(1), int8(1), object(104)
- 추후에 필요하다면, 데이터 타입을 변경하면서 사용할 것임.
- 데이터 ''(na값)는 총 14228539로 많았음. ''(na) 으로 빈값들이 많았음. ''은 추후, 사용할 컬럼들을 이용하면서 제거할 것임.
- duplicated로 확인한 결과 중복값은 없었음.
- 'city'열을 살펴보니 댈러스가 아닌 도시들이 섞여있어서 제거했음. [''(na)도 많지 않아 제거, 'dallas'->'DALLAS'로 바꾸고 제거.]
- 제거한 후 , 658043 가 남음.
- 밑에 테이블은 head()한 모습임.

index	incidentnum	UCR_ctype	cnt	ctype	year	servicenumberid	watch
0	000001-	NaN	1		2019	000001-2019-01	3

index	incidentnum	UCR_ctype	cnt	ctype	year	servicenumberid	watch
	2019						
1	000001-2020	NaN	1		2020	000001-2020-01	1
2	000002-2015	1.0	1	MURDER	2015	000002-2015-01	3
3	000002-2018	NaN	1	FOUND	2018	000002-2018-01	1
4	000002-2019	NaN	1		2019	000002-2019-01	1

2. Univariate analysis

Presentation of key variables from various aspects

2.1 day1oftheweek

- day1oftheweek: 범죄발생일. 즉,처음 사건이 발생한 요일을 의미.
- dtype : object
- 총 데이터 갯수 : 658043, na값 존재하지 않음.
- da1y1oftheweek는 Mon~Sun 데이터 총 7개의 unique의 값이 존재.
- 무슨 요일날 범죄사건이 가장 많이 일어났는지 보기 원해, 요일별 범죄사건 Count.

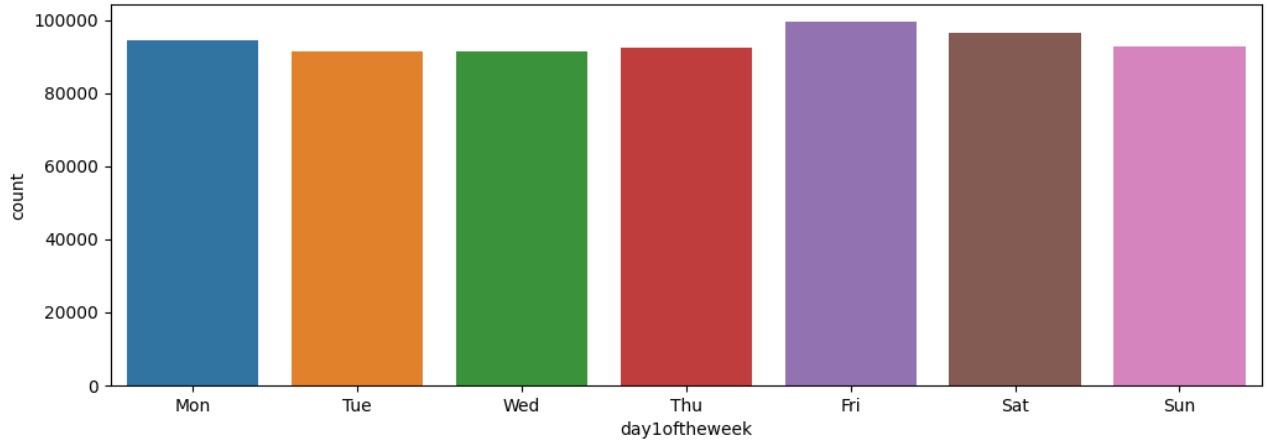

Figure 1. 요일별 범죄 사건 수

- 금요일날 범죄수가 가장 많았고, 화요일날 범죄가 가장 적었음. 요일마다 범죄수가 다름을 catch.

2.2 'typeofincident'

- 'typeofincident': 발생사건유형을 기록.
- dtype : object
- 총 데이터 갯수 658043, na값 존재하지 않음
- 발생사건 유형은 1,055개가 존재한다.
- 대표적으로 보면, BMV(자동차 절도) , UNAUTHORIZED USE OF MOTOR VEH - AUTOMOBILE(자동차 불법사용) , BURGLARY OF HABITATION - FORCED ENTRY(주택 침입 절도)가 있다.
- value_counts()로 상위 범죄들을 간단히 살펴보면, BMV 가 다른 것에 비해 훨씬 많았음.
  *BMV : 75126
  *UNAUTHORIZED USE OF MOTOR VEH - *AUTOMOBILE : 34101
  *FOUND PROPERTY (OFFENSE) :24166
  *BURGLARY OF HABITATION - FORCED ENTRY
  : 23431
  *PUBLIC INTOXICATION: 20736

2.3 'division'

- 'division': 경찰 관할구역을 의미
- dtype : object
- 관할 구역 : 14개 존재
- na값 236개 존재. 극소수이므로 제거함.
- division을 value_counst()해보니, Northeast, central이 범죄가 많았음. 지역마다 범죄수가 다름을 catch.
  *NORTHEAST :109710
  *CENTRAL :100276
  *SOUTHEAST : 97765
  *NORTHWEST : 97755

*SOUTHWEST :96847
*SOUTH CENTRAL : 85511
*NORTH CENTRAL : 69985

2.3 'time1ofoccurence_h'

- 'time1ofoccurence'의 파생변수데이터
- 원래 데이터 'time1ofoccurence' : 사건 발생한 시간을 의미(ex: '22:00' 처럼 hour,minute존재)
- dtype변경 : object -> datetime
- na값 존재하지 않음.
- time1ofoccurence_h로 이름을 두어, 파생변수로 사용. 원래 데이터 타입이 object type 이어서 datetime으로 변경 후, mintue은 사용하지 않고, hour 만 추출하여 사용
- 원데이터 형태 : '23:00' -> 23(hour)
- 0-23 시 까지 존재. 총 24개 unique값 존재.

2.4 'xcoordinate'

- 'xcoordinate': (위치정보) 지리적 x 좌표를 나타냄.
- dtype 변경: object -> float (지리적 정보 사용을 위해)
- 원데이터에서 663,249개 중 2614 na값(0.003) 존재. 극소수이므로 na값 제거

2.5 'ycordinate'

- 'ycordinate': (위치정보) 지리적 y 좌표를 나타냄.
- dtype 변경: object -> float (지리적 정보 사용을 위해)
- 원데이터에서 663,249개 중 2614 na값 (0.003)존재. 극소수이므로 na값 제거

3. Multivariate analysis

Presenation of hidden patterns between variables (correlation, clustering, etc.)

3.1 'typeofincident'(사건유형)과 day1oftheweek(요일)

- 'typeofincident'(사건유형)과 day1oftheweek(요일)변수 이용
- 대표적으로 댈러스 도시에서 가장 많이 발생하는 상위 5개 사건을 필터링하고, 요일별로 발생빈도를 확인해보니, 범죄 종류별로 범죄가 발생하는 요일이 다른 패턴을 볼 수 있음.
- 상위5개 사건 : BMV(자동차절도), PUBLIC INTOXICATION(공공장소 음주), UNAUTHORIZED USE OF MOTOR(자동차 무단 사용),found proprety(주인없이 발견된 재산)

- BMV, UNAUTORIZED - AUTOMOBILE, FOUND PROPERTY 범죄는 평일, 주말 관계없이 자주 발생한다.
- PUBLIC INTOXICATTION은 주로 토,일 (주말)에 자주 발생한다.
- BURGLAY OF HABITATION은 주로 평일(월-금)에 자주 발생하다.
- Point : 즉, 범죄유형에 따라 자주 발생하는 요일(pattern)이 다름을 시사한다.

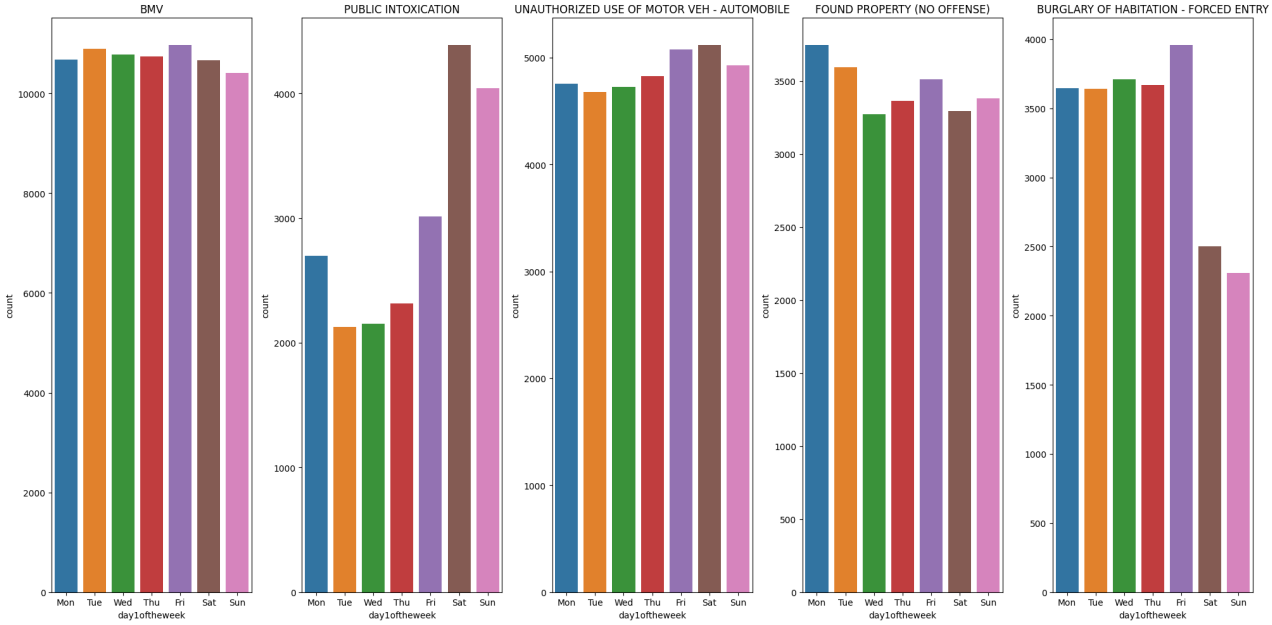

Figure 3.1. 범죄유형,요일별 범죄 사건 수

3.2 typeofincident(사건유형)과 time1ofoccurence_h

- typeofincident(사건유형)과 time1ofoccurence_h(발생시간) 변수 이용
- 댈러스 도시에서 가장 많이 발생하는 상위 5개 사건을 가져와 시간별로 발생빈도를 확인해보니, 범죄 종류별로 범죄가 발생하는 시간대가 다른 패턴을 볼 수 있음.
- 상위5개 사건 : BMV(자동차절도), PUBLIC INTOXICATION(공공장소 음주), UNAUTHORIZED USE OF MOTOR(자동차 무단 사용),found proprety(주인없이 발견된 재산),
- (왼쪽에서부터) 1번째 그래프는 BMV(자동차 절도)는, 주로 저녁 17시이후부터, 범죄가 많이 발생하고, 22시에 가장많은 자동차 절도죄가 가장 많이 일어난다.
- 2번째 그래프는 'PUBLIC INTOXICATION'(공공장소에서 음주)은 저녁시간과 특히 새벽(0-3시)에 주로 발생.
- 3번째 그래프는 'UNAUTHORIZED USE OF MOTOR VEH - AUTOMOBILE'(무단 자동차 사용) 저녁(17시이후)부터 많이 발생하고 특히 0시에 가장 많이 발생.
- 4번째 그래프는 'found property(주인 없이 발견된 재산)는 낮에도 많이 발생하고, 오후4-6시, 0시 많이 발생함.
- 5번째 그래프는 'BURGLARY OF HABITATION - FORCED ENTRY'(주거 강제 침입은) 오전시간과 낮시간에 발생. 특히 오전 7-8시에 많음.
- Point : 즉, 범죄 종류에 따라 범죄가 자주 발생하는 시간대가 있다.

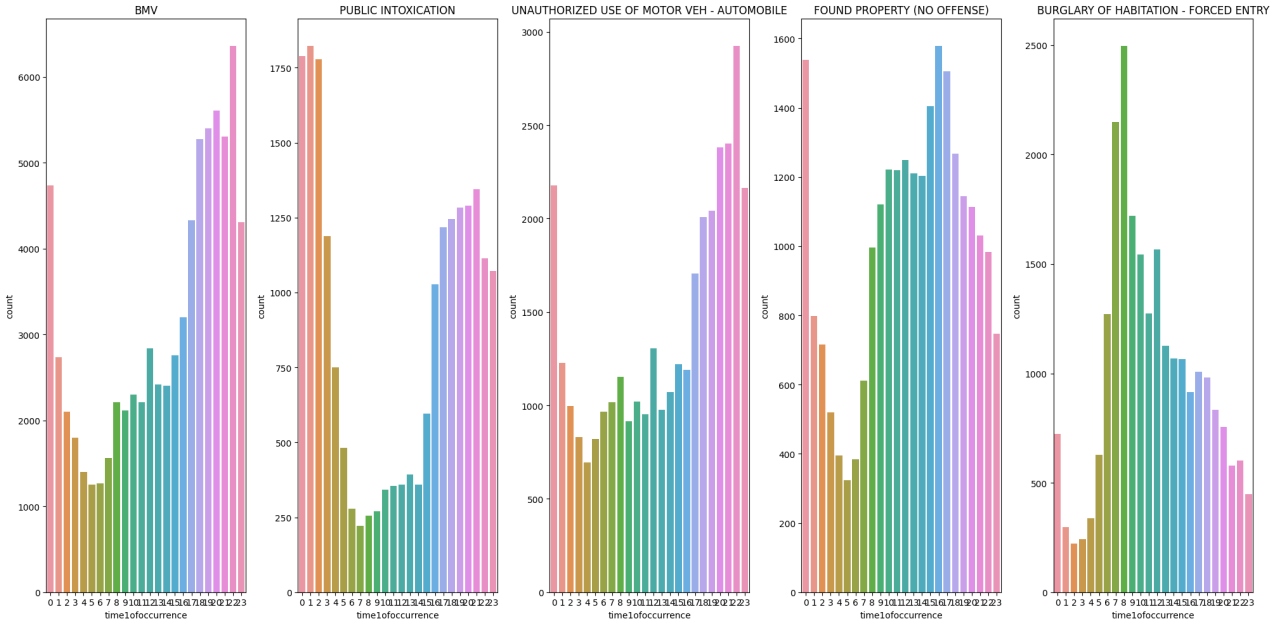

Figure 3.2.범죄유형, 시간대별 범죄 사건 수

3.3 'typeofincident'(사건유형)과 'xcoordinate'(x좌표),'ycoridnate'(y좌표), 'division'(관할구역)

- typeofincident와 x,y,division변수 이용
- figuer3.3.1~ 3.3.3 까지는 댈러스에서 자주 발생하는 상위 3대 범죄들이 발생하는 위치(typeofincident와 x,y,division)를 보여주는 산점도이다.(댈러스 상위 3대범죄 : BMV,UNAUTORIZE USE - Automobile, Burglary of habitaion)
- figure3.3.1을 보면, bmv가 자주 발생하는 위치를 살펴볼 수 있다. central과 northeast에 많이 몰려있고,자주 발생하고, southeast는 상대적으로 덜 몰려있고 적게 발생한다.
- figure3.3.2을 보면,UNAUTORIZE USE - Automobile을 발생위치를 살펴볼 수 있다. 상대적으로 UNAUTORIZE USE가 Southwest,SouthCentral이 더 몰려있고, 더 자주발생하고, north central에서 덜 몰려있고, 적게 UNAUTORIZE USE가 발생한다.
- figure3.3.3을 보면, 상대적으로 southewest,south cemtral이 Burglary of habitaion 더 발생하고, north west가 Burglary of habitaion 범죄 발생이 덜 몰려있고 덜 발생한다.
- Point : 즉, 범죄별로 자주 발생하는 위치,관할구역이 있다.

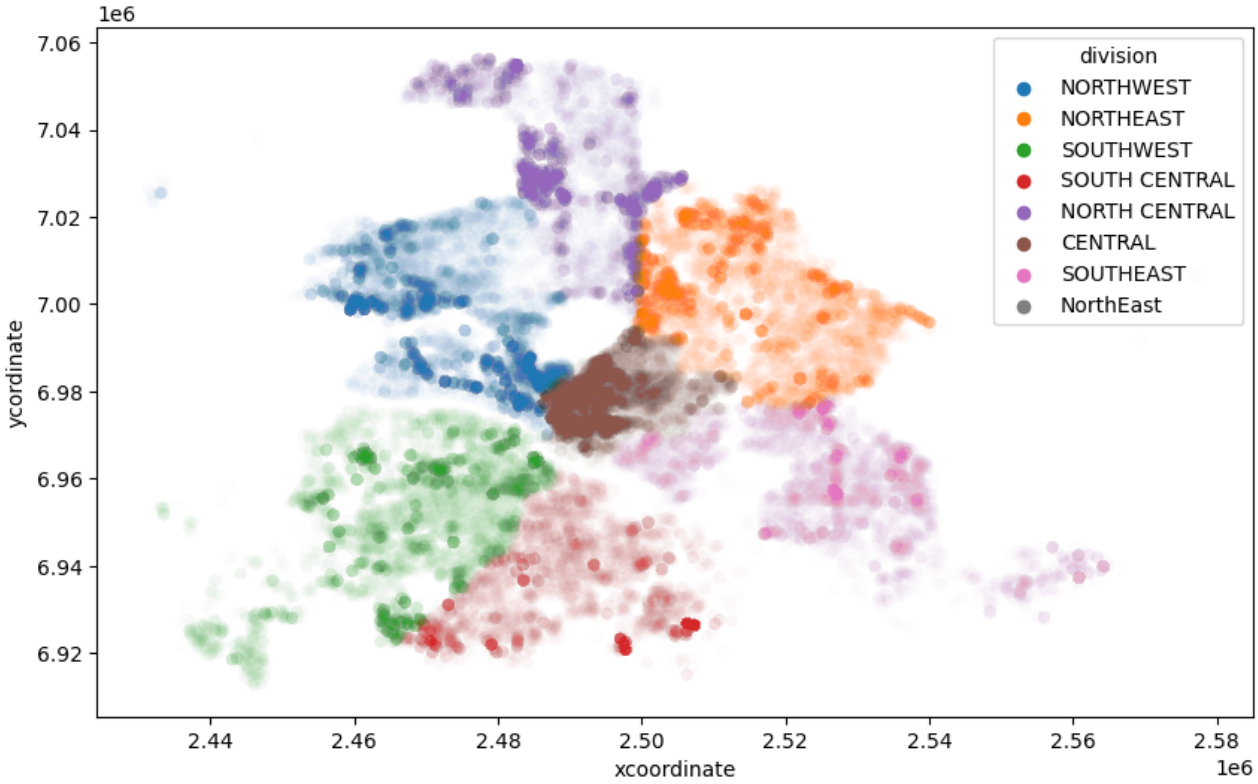

Figure 3.3.1 'BMV'의 위치별 사건 수

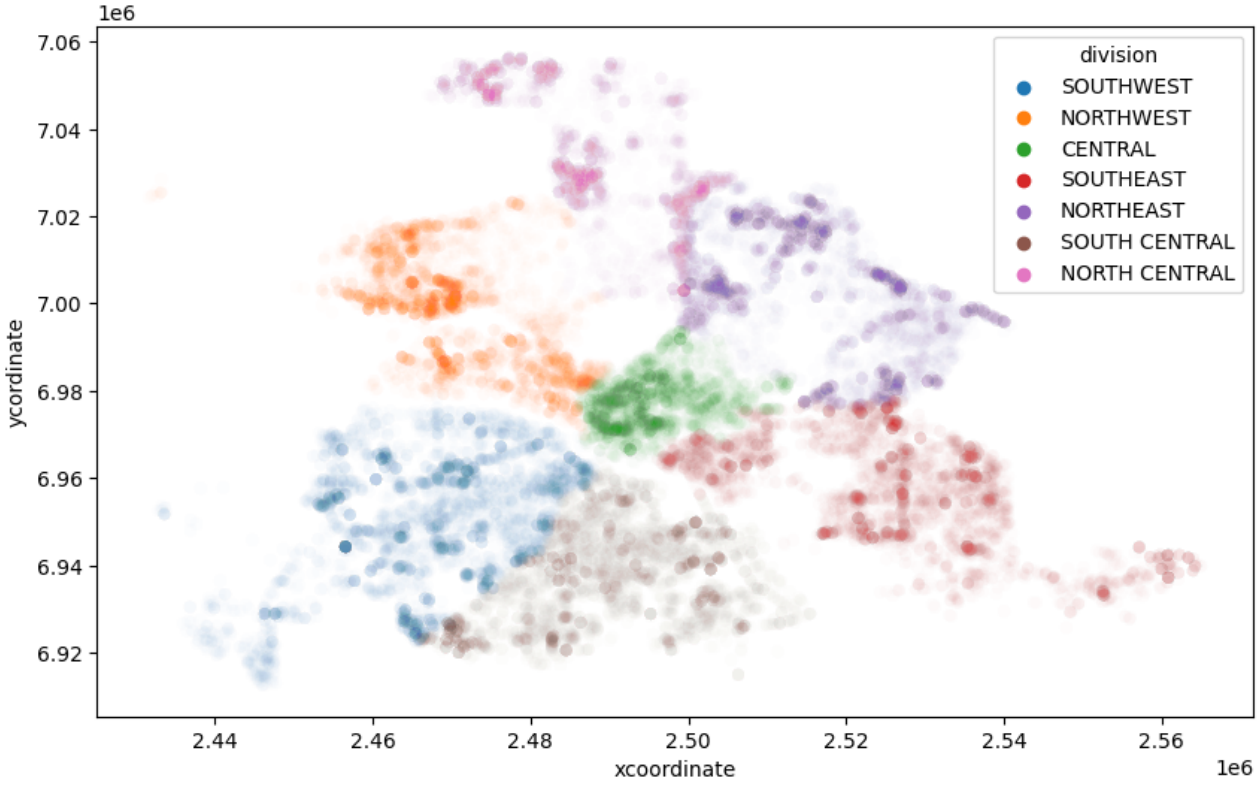

Figure 3.3.2 'UNAUTORIZE USE - Automobile'의 위치별 범죄 사건 수

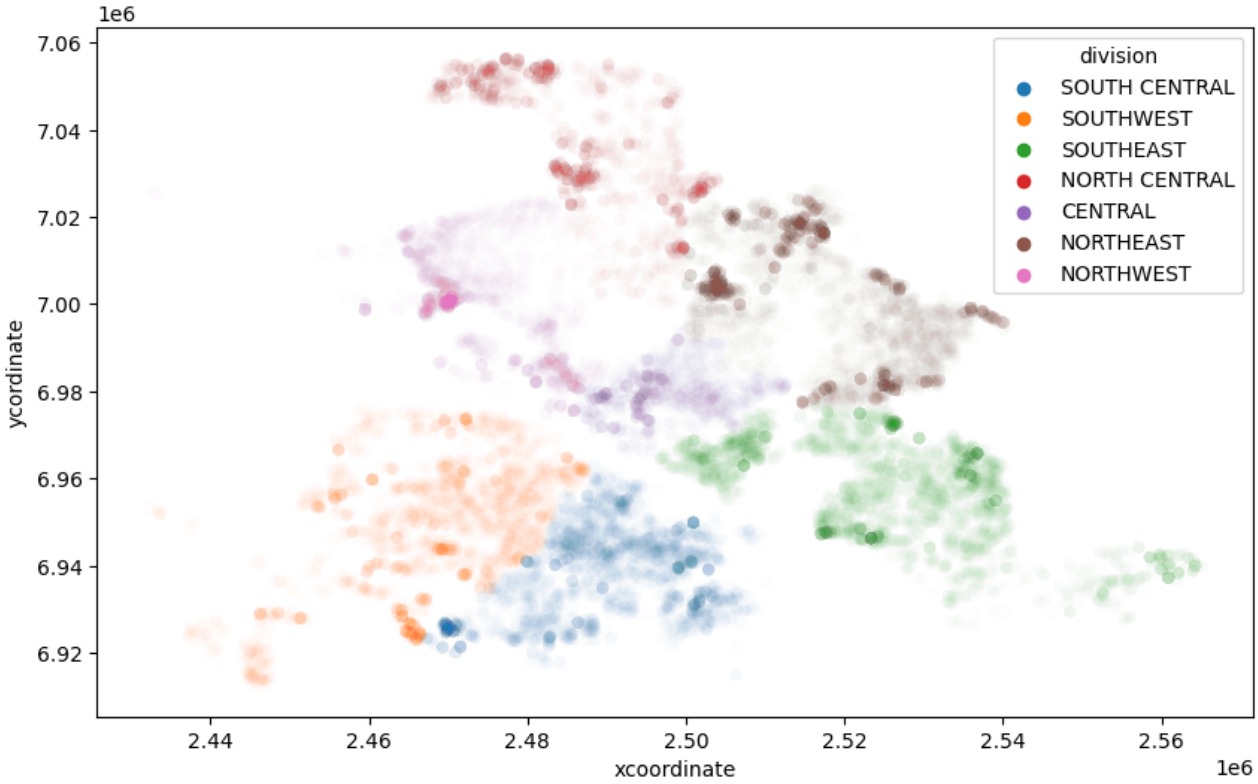

Figure 3.3.3 'Burglary of habitaion' 위치 별 범죄 사건 수

4. Suggestion

- 위 데이터 탐색을 통해, (전부는 아니지만) 범죄유형별로 자주 범죄가 발생하는 시간이 다르고, 자주 범죄가 발생하는 요일이 다르고, 자주 범죄가 발생하는 위치(구역)이 다르다. 즉, 특정시간, 특정공간에 유독 자주 발생하는 범죄들이 존재한다는 것이다. 이 말은 시간과 공간 데이터를 잘 이용한다면, 한 발짝 더 빠르게 인력배치, 모니터링을 할 수 있어 보다 효과적으로 범죄를 예방할 수 있다는 것을 의미한다.

- 위 데이터 기반으로 한기지 예시로 든다면, Burglary of habitaion 범죄는 SOUTH CENTRAL에서 주로 발생. 요일, 시간대는 평일(월-금),오전에 주로 발생. 이러한 예시의 시공간 데이터를 활용한다면, 범죄유형별로 범죄가 자주 발생하는 지역데이터 분석기반 순찰로 지정, 공간데이터 분석기반 cctv 설치, 시간공간 데이터를 활용하여 특정시간 cctv 집중 모니터링, 범죄예방지도(대시보드) 구축 등 해결책을 제시할 수 있다. 실제로 우리나라 울산에서 시공간 빅데이터를 활용하여 범죄예측 시스템 모니터링을 실시하고 있다.

기사링크 첨부 : https://m.kmib.co.kr/view.asp?arcid=0016958731