교통사고 데이터 분석/ 미래 교통사고 예측을 위한 EDA

Introduction

전국의 교통 사고 데이터를 포함한 데이터셋을 이용하여 주어진 지역의 교통사고에 대한 정보를 예측하는 모델을 만드려고 한다. “지역, 사고유형, 도로 형태, 차량 종류 등의 과거 교통사고 데이터를 분석하여, 미래 교통사고에 대한 정보를 예측하는 문제"를 풀고자 한다. 모델링에 앞서, EDA를 통해 교통사고의 원인 및 특성을 파악하는 것을 목표로 하였다.

Data

분석에는 국토교통부에서 공개하는 교통사망사고 및 사고다발지 데이터를 이용하였다

- 데이터 출처 : https://www.data.go.kr/dataset/15003493/fileData.do
- 데이터 설명 : 교통사망사고 데이터의 기간은 12년 1월부터 17년 6월까지이며, 최소 사망자가 1건 이상인 데이터가 기준이 된다. 보조데이터로 활용한 것은 사고다발지 데이터로 각각 무단횡단/보행노인/보행어린이/스쿨존내/자전거 사고다발지 이다. 서울시 도로 링크별 교통 사고발생 수 자료도 참고하였다.

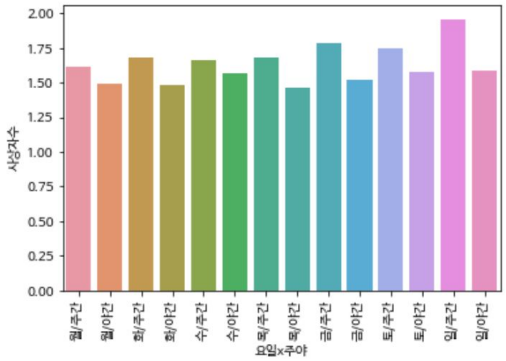
변수 이름 및 설명

- 수치형 : 사상자수, 사망자수, 중상자, 경상자수, 부상신고자수
- 범주형 : 발생시간(주야,요일), 발생지(발생지시도, 발생지시군구), 사고유형(대/중분류), 법규위반, 도로형태, 당사자종별_대분류(1당, 2당)

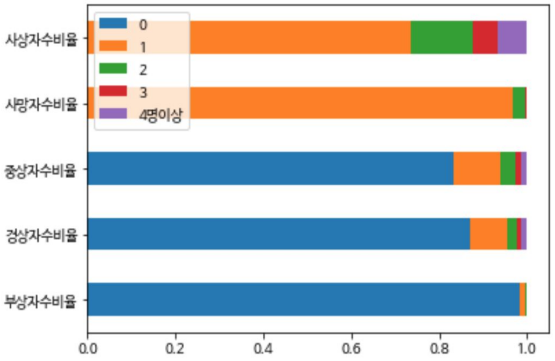
주야	요일	사망자수	사상자수	중상자수	경상자수	부상신고자수	발생지시도	발생지시군구	사고유형_대분류	사고유형_중분류	법규위반	도로형태_대분류	도로형태	당사자종별_1당_대분류	당사자종별_2당_대분류
야간	화	2	3	1	0	0	경기	화성시	차대차	측면충돌	중앙선 침범	단일로	기타단일로	승용차	승합차
야간	화	1	1	0	0	0	전남	영암군	차대사람	차도통행중	과속	단일로	기타단일로	승용차	보행자
야간	화	1	1	0	0	0	전남	곡성군	차량단독	전도전복	안전운전 의무 불이행	단일로	기타단일로	자전거	없음
주간	화	1	5	1	3	0	대구	달성군	차대차	측면충돌	중앙선 침범	단일로	기타단일로	승용차	승합차

Exploratory Data Analysis

요일/주야별 비율



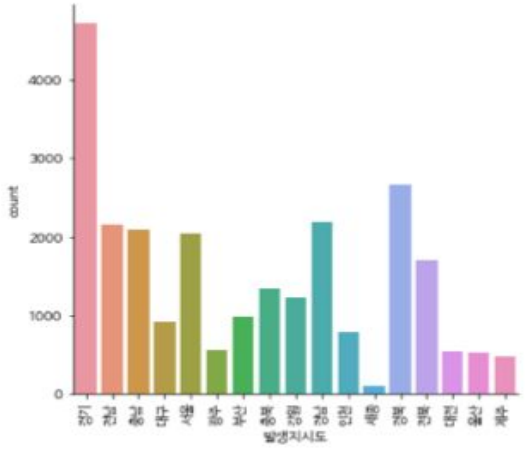
사상자별 비율



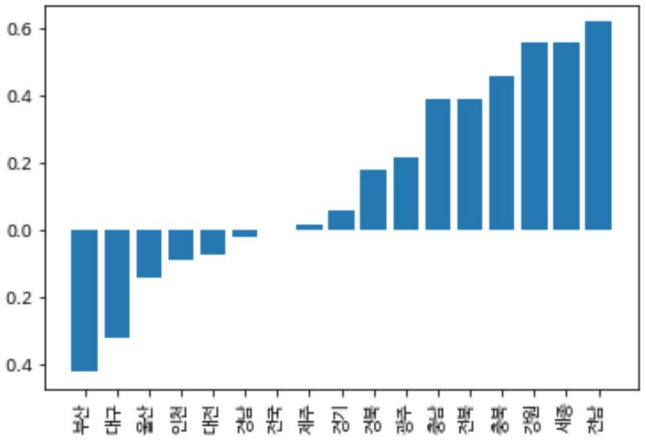
왼쪽 그래프는 전체 데이터의 요일/주야별 비율을 분석한 내용이다. 발생한 사고의 전체비율을 보면 야간(51.5%)이 주간(48.4%)보다 조금더 높다. 그러나 요일/주야별로 좀 더 세분화해서 살펴보면 평일보다는 금,토,일요일 주간이 다소 높게 나온다. 주중의 생활반경이 집과 회사 또는 학교와 같이 익숙한 환경이라면 주말은 나들이나 먼 곳으로도 이동이 가능해 이동반경이 넓어진 것도 한 가지 원인이 될 수 있을 것이다.

오른쪽 그래프는 사상자수 별로 비율을 분석한 내용이다. 우선 사망자수가 0인 데이터는 없다. 모든 교통사고가 아닌 사망자수가 1명 이상인 교통사망사고 데이터이다. 사망자수는 대체적으로 1~2명 이고 중상자수, 경상자수는 0~3명, 부상신고자수는 0~1명 값을 가지기에 모델로 예측시 오차범위가 크지 않을 것이다. 우리가 주의해야할 부분은 특수한 경우, 사고의 피해규모가 범위를 넘어서는 사고들의 특징을 파악하여 이를 변수로 추가하는 것이다.

발생지시도별 사고발생수



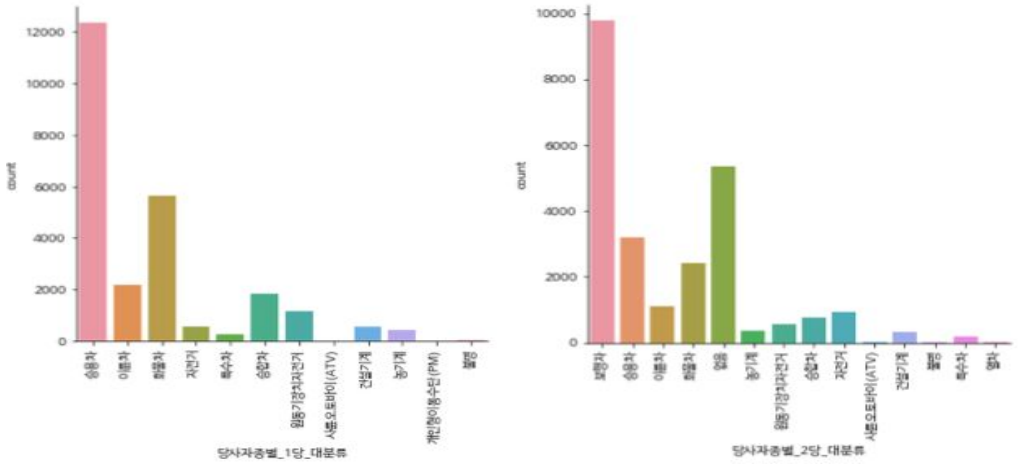
시도별 전국 평균과의 위험도 차이



왼쪽 그래프는 발생지시도별로 전체 사고 건수를 나타낸 것이다. 발생지도별로 총 사고발생건수만 단순 비교하기에는 시도별 인구수 및 도로환경이 동일하지않아 적당하지 않다고 판단했다. 시군구 위험도 데이터를 활용해서 시도별 위험도를 합산 후에 전국 평균 위험도와의 차이를 나타낸 것이 오른쪽 그래프이다. 왼쪽 그래프에서 사고발생수가 가장

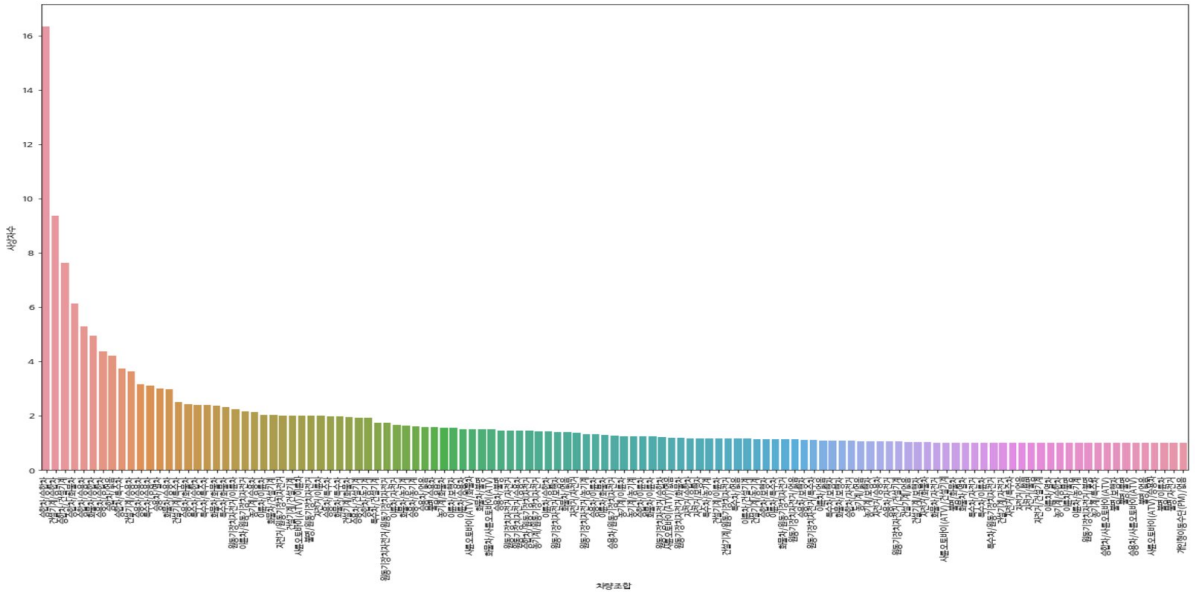
많았던 곳은 경기, 가장 적었던 곳은 세종시였다. 그러나 오른쪽 그래프에서는 전남, 세종, 강원 순서로 위험도가 높고 경기는 전국 평균을 약간 상회하는 모습이다.

당사자종별_1당_대분류 **당사자종별_2당_대분류**



당사자종별 1당은 가해자를 2당은 피해자를 나타낸다. 위 그래프는 당사자종별 1당, 2당별로 사고발생수를 분석한 그래프이다. 1당 대분류에는 승용차, 화물차, 승합차의 순으로 사고발생건수가 많았으며, 2당 대분류에는 보행자, 없음, 승용차 순이였다. 일반적으로 발생하는 다수의 사고가 아닌 사고피해규모가 큰 사고를 찾기 위해선 사고가 발생한 차량간의 조합이 중요하다고 판단했다. 1당과 2당은 가해자와 피해자의 순서를 나타내기때문에 순서를 유지한 차량조합별 사상자수를 분석해보고자 했다.

당사자종별_1당_대분류/ 당사자종별_2당_대분류 조합별 사상자수



위 그래프는 차종 조합별로 평균 사상자수를 나타내보았다. 평균 사상자수가 가장 높은 조합은 탑승인원이 많은 승합차와 승합차간의 사고이며, 그 다음 조합들은 승합차가 포함되고 건설기계, 화물차, 승용차, 없음의 순서이다. 평균 사상자수가 낮은 조합은 탑승인원이 상대적으로 적은 자전거, 원동기장치자전거, 이륜차를 포함하고 있다. 사상자수가 3이하로 낮아지기 시작하는 건설기계/화물차 조합 이후의 조합들을 별도로 차량조합_저위험 으로

분류하고 반대로 3이상인 조합은 차량조합_고위험으로 분류해서 분석을 해보는 것도 의미가 있을것이라 판단된다.

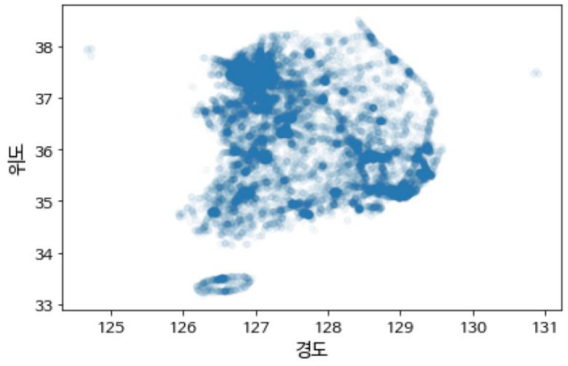
Exploring Traffic Accident

몇 가지 가설을 통해 교통사망사고 데이터에 대해 자세히 알아보자.

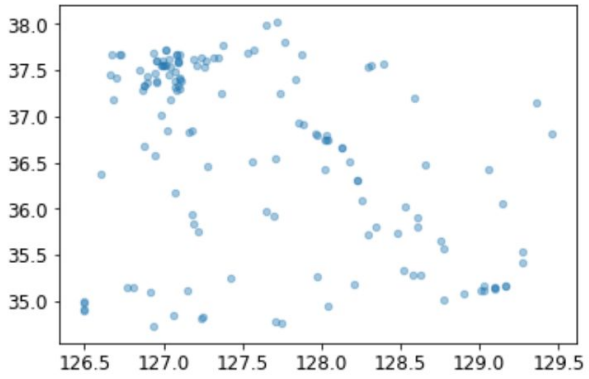
-도로의 구조, 특성이 사고에 영향을 줄 수 있을 것이다

사고가 가장 많이 발생하는 도로형태는 단일로-기타단일로, 교차로-교차로내 이다. 기타/불명도 상대적으로 높은 수치를 나타내고, 특수한 경우인 단일로-터널안, 단일로-교량위 케이스도 분석해 보려고한다. 단순합인 사상자수만으로는 사망자 5명이 사고와 부상자 4명인 사고가 동일하다. 따라서 사고의 피해정도를 수치로 나타내기 위해 서로다른 가중치를 넣어서 사고피해정도 변수를 새롭게 생성, 분석을 진행하였다. (사망자수에는 4, 중상자수 3, 경상자수 2, 부상신고자수 1을 곱한후 합산) 주로 발생하는 사고유형타입은 승용차-보행자가 많았다.

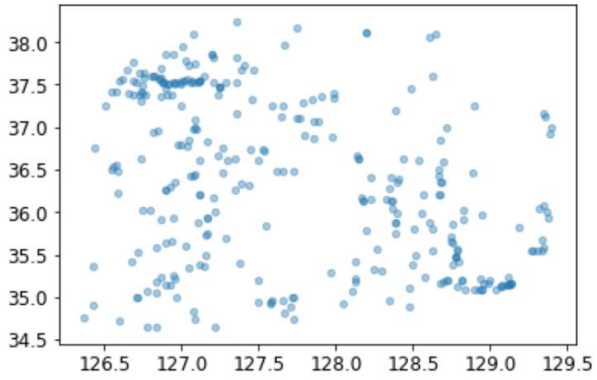
교통사망사고 전체의 위경도 plot



터널안 사고 plot



교량위 사고 plot



위 plot 은 특수한 도로형태에서 사고의 분포가 어떻게 발생되는지에 대한 내용이다.
왼쪽 터널안 사고는 주로 서울,경기를 중심으로 차대차, 추돌사고 주로 발생한다. 특이한 점은 요일기준으로 목요일과 수요일은 하루 차이지만 사망사고발생에선 차이를 보인다. 사고가 가장 많이 발생하는 요일은 목 > 금 > 토 순으로 나타난다. 오른쪽 교량위 사고는

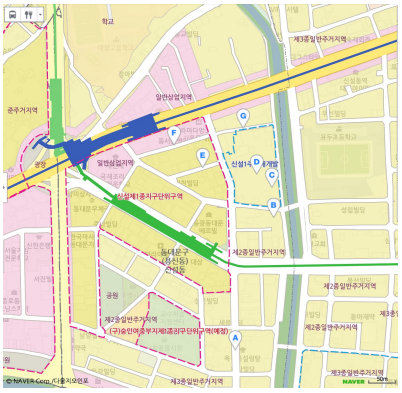
경북지역에서 가장 많이 발생하고 차대차가 비율이 가장 높지만 차량단독 사고도 터널보다는 비율이 높다. 터널과는 다르게 야간, 토요일에 관련 사고가 많이 발생한다.

서울시 도로링크별 교통사고 데이터 > 위험도가 높은 도로

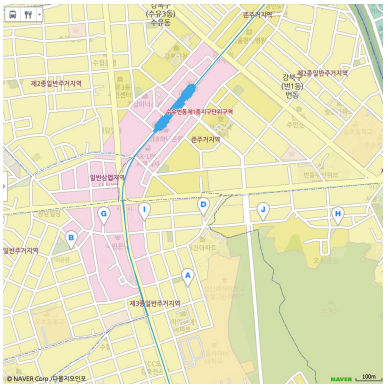
	링크ID	위험도	위험등급	시군구	읍면동	도로명	도로길이_m	차로수_편도	사고건수	사망자수	중상자수	경상자수	부상신고자수
13698	1030049800	502.81	4	성동구	상왕십리동	하정로	54	1	40	0	14	33	7
2323	1080007301	182.27	4	강북구	미아동	덕릉로	71	4	20	0	10	14	1
19438	2180000801	180.86	4	은평구	수색동	승전로	25	3	5	2	4	4	0
1463	1240000302	173.99	4	강동구	성내동	천호대로	197	5	28	1	18	55	6
5106	1200004701	171.38	4	관악구	봉천동	남부순환로	572	4	90	3	21	89	5
4027	1200004701	171.38	4	관악구	봉천동	남부순환로	572	4	90	3	21	89	5
19478	1000002100	158.55	4	종로구	창신동	종로	611	3	73	0	27	57	18
14803	1230008303	156.43	4	송파구	가락동	중대로	200	3	27	0	17	20	2
17899	1180040301	146.69	4	영등포구	당산동6가	당산로	50	2	8	0	11	65	8
17198	1180040801	145.05	4	영등포구	양평동4가	선유로	69	5	20	0	10	20	4
2292	1080005203	141.87	4	강북구	미아동	도봉로	226	3	28	1	17	17	2

남부순환로와 종로를 제외하고 도로길이가 200m 이하이다. (전체 도로길이 평균은 273.98m)
지도상으로는 중심지인 것만 알수 있고 특별한 특징은 알 수 없었다.

지적편집도_하정로



지적편집도_덕릉로



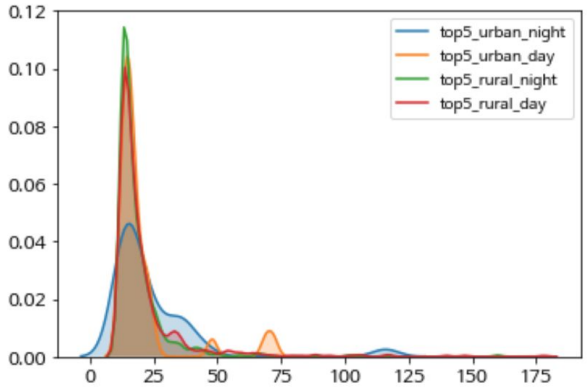
지적편집도_천호대로

네이버 지도의 지적편집도 기능을 활용해서 위험도가 높은 도로를 다시 살펴보았다.
위험도가 502.81로 가장 높은 하정로는 제2종일반주거지역(지도상 녹색 영역)에
위치(A)해있고 12시방향 일반상업지역
(지도상 분홍색 영역)과 만나는 모습이다. 위험도 182.27로 2번째로 높았던 덕릉로는
제2종일반주거지역 가운데 부분을 일반상업지역이 지나가는 모습이다. 천호대로 역시
제2종일반주거지역과 일반상업지역의 경계에 위치해있다. 종합해보면 위험도가 높은 도로는
일반주거지역과 상업지역이 만나거나 경계를 이루는 특징을 보인다. 원인을 분석해보면
주거지와 상업지역 사이의 인구 왕래가 잦고, 상업지역은 대부분 중심지에 형성되어서 다른
곳에 비해 인구밀도도 높을 것이다.

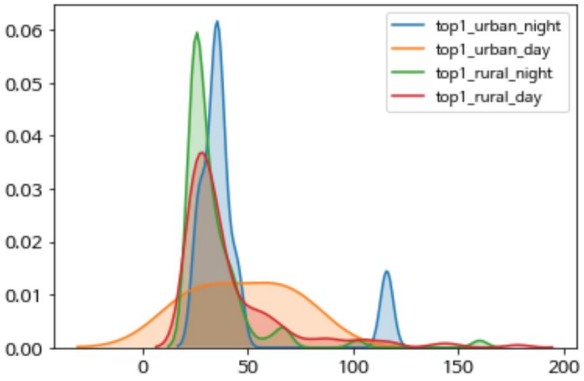
- 도시/지방/주간/야간 이라면 도시주간보다 지방야간이 규모가 큰 사고가 자주 발생할까?

도시와 지방을 어떤 기준으로 나눌것인지, 규모가 큰 사고를 어떻게 규정할 것인지에 차이를 두어 실험을 해보았다.

실험1.도시=서울 / 사고피해정도 상위5%



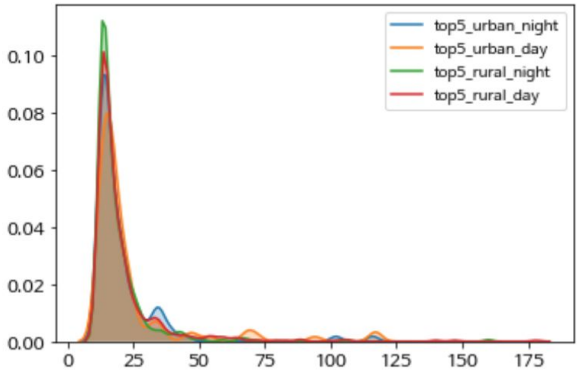
실험2.도시=서울 / 사고피해정도 상위1%



먼저 실험1을 살펴보면 도시여부에 상관없이 규모가 큰 사고(사고피해정도 상위5프로)는 주간에 더 많이 발생한다. 처음에 가정했던 규모가 큰 사고가 도시 주간(34건) 보단 지방 야간(497건)에서 더 자주 발생하는 건 맞다. 그러나 전체사고 대비 상위5프로 사고가 차지하는 비중으로 보면, 지방 주간사고가 6.10%, 도시 야간사고가 2.59%로 지방 주간에서 사고가 발생했을 때 큰 사고일 가능성이 더 높다

규모가 큰 사고의 기준을 사고피해정도 상위1프로로 변경한 실험2를 살펴보면, 도시에선 사고피해정도 평균값이 주야간 모두 지방보다 높으나 피해정도가 상대적으로 큰 사고들은 지방에서 더 많이 발생한다.
전체사고 대비 상위1프로 사고가 차지하는 비중으로 보면, 지방 주간사고가 1.34%, 도시 주간사고가 0.58%로 역시 지방 주간에서 사고가 발생했을 때 큰 사고일 가능성이 더 높다

실험3.도시=5대도시 / 사고피해정도 상위5%



실험3은 도시의 기준을 5대도시(서울,부산,인천,대전,대구)로 변경해서 사고피해정도 상위 5프로 사고를 살펴보았다. 발생한 사건수(625건) 나 전체사고 대비 비중(6.31%)로 지방

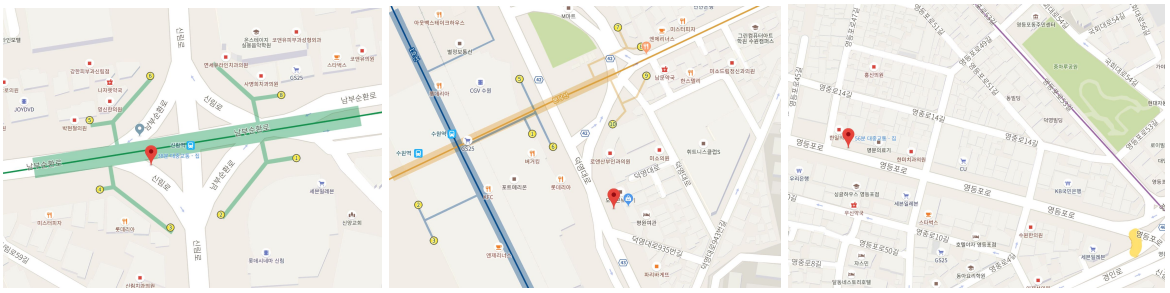
주간이 높고, 도시 야간은 전체사고 대비 비중이(3.44%) 낮게 나왔다. 사고피해정도가 큰 규모의 교통사망사고들은 지방 주간에서 뚜렷하게 높은 발생빈도를 보인다. 이러한 원인은 무엇일까? 실험1 데이터에서 지방*주간*사고피해정도가 높은 사고들을 살펴보면 단일로-기타단일로에서 주로 사고가 발생하였다. 위경도 좌표로 해당위치를 살펴보면 고속도로에 위치해 있고 승합차가 포함된 추돌사고이다보니 피해정도가 큰 것을 알 수 있었다.

-전라남도 곡성군(호남고속도로) -대전광역시 대덕구(경부고속도로) -인천광역시 부평구(서울외곽순환)

-보행자편, 보행자 입장에서 언제 사고 위험이 높을까?

무단횡단사고다발지에서 발생건수, 사고피해정도가 높은 지역부터 살펴보았다.

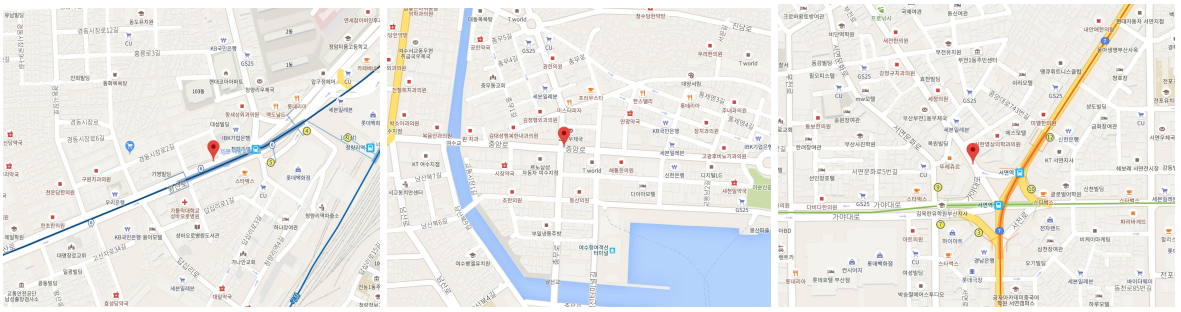
- 서울특별시 관악구 신림동 -경기도 수원시 팔달구 매산로1가 -서울특별시 영등포구 영등포동3가



무단횡단사고가 자주 발생하는 곳의 주요 특징은 횡단보도와 횡단보도 사이의 거리가 멀거나 육교/지하도상가는 있으나 횡단보도까지의 거리가 먼 곳이 많았다. 다리나 순환로가 끝나는 곳에서 나온 차량과 무단횡단하는 사람이 만나는 교차점으로 추정되는 곳도 있었다.
특이사항은 서울특별시 관악구 신림동(신림역부근)은 발생건수 및 사고피해정도가 가장 높은 곳으로 횡단보도가 많은데도 무단횡단이 많이 발생하였다. 검색결과에 따르면 무단횡단은 횡단보도 위에서 신호를 위반하고 건너는 경우도 포함한다고 한다. 횡단보도 설치간격도 무단횡단 사고에 직간접적으로 영향을 미친다고 생각된다.

보행노인사고다발지에서 발생건수, 사고피해정도가 높은 지역부터 살펴보았다.

-서울특별시 동대문구 청량리동 -전라남도 여수시 교동 -부산광역시 부산진구 부전동

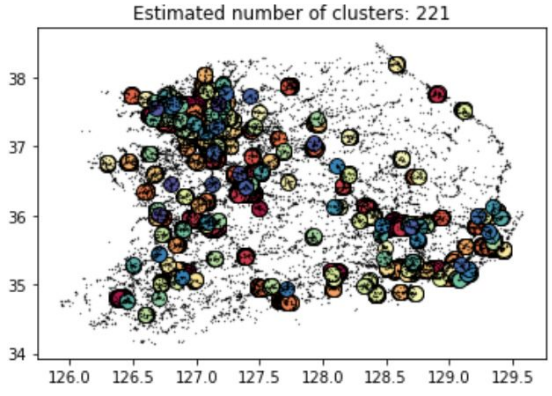


보행노인사고가 가장 많이 발생하는 서울특별시 동대문구 청량리동을 먼저 분석해보았다. 사고발생지 주변에 위치한 지역적 특징들은 청량리역, 재래시장, 농수산물시장, 성바오로병원이 위치해있었다.
전라남도 여수시 교동에는 교동시장과 여객터미널이, 부산광역시 부산진구 부전동에는 서면역, 백화점, 종합시장이 있었다. 종합해보면 노인사고다발지에는 공통적으로 전통시장이 있는 부근이었다. 노인들이 많이 통행하는 곳이 시장이다보니 보행시 사고도 많이 발생하는 것으로 추정된다.

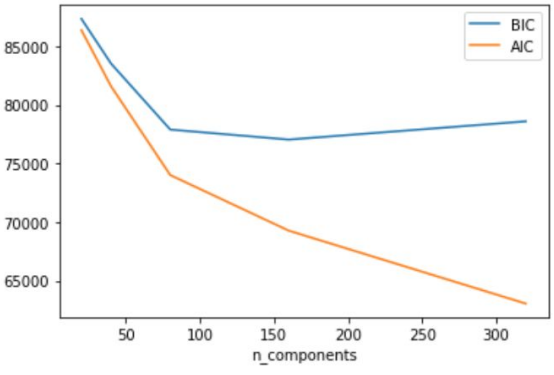
-위도 경도를 이용해서 사고다발지별로 군집화/ 새로운 변수 생성해보기

train 데이터셋에 있는 위도경도 정보를 이용해서 test 데이터셋에서도 활용할 수 있는 새로운 변수를 생성해보고자 한다. 데이터셋에 있는 발생지시도와 발생지시군구를 합쳐서 발생지 변수를 생성하면 총 231개의 발생지가 나온다.

- DBSCAN을 이용한 군집화



- AIC, BIC 함수

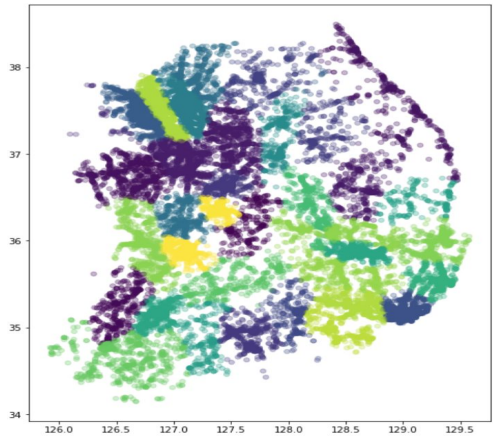


왼쪽에 있는 그림은 DBSCAN을 이용해서 사고데이터를 군집화 한것으로 색이 있는 원은 각 군집의 중심을, 상대적으로 작고 검은색 점은 군집에 포함되지 않는 노이즈를 나타낸다. 각 군집의 중심을 기준으로 발생지 231개가 어떤 군집에 포함되는지, 포함된다면 어느정도 확률로 포함되는지를 새로운 변수로 생성하고자 한다. 여기서 잠시 고민해보아야할 부분은 발생지 변수는 총 231개의 구역이 나왔다.

행정구역상 기준으로 한다면 위 기준으로 하는 것이 맞으나 사고다발지를 계산해서 데이터 자체에 의미있는 변수로 활용하기엔 적합하지 않을 수 있다. GMM과 같은 generative 모델의 경우 데이터에 주어지는 고유의 확률분포에 따라 정보량을 측정할 수 있는 측정기준이 제공된다. AIC와 BIC는 각각 Akaike information criterion, Bayesian information criterion 을 나타낸다. 오른쪽에 있는 그래프에서
교통사망사고 데이터에 AIC, BIC 함수를 이용해 적절한 components의 수를 찾아보면 BIC함수는 우리가 160개 정도를 선택하는게 적절한 것을 알려준다.

앞서 찾은 적정 components 의 수(160) 를 Gaussian Mixture 모델에 적용해보면 아래와 같은 모습이다.
가우시안 모델이다보니 DBSCAN과 달리 따로 노이즈로 표시되지않고 동일한 클러스터라면 동일한 색으로 표시가 된다.

-Gaussian Mixture Model



- 발생지별 클러스터번호 및 클러스터속할확률

	발생지	클러스터번호	클러스터속할확률
0	강원강릉시	10	0.990362
1	강원고성군	10	0.912299
2	강원동해시	10	0.998834
3	강원삼척시	10	0.992975
4	강원속초시	10	0.939397
5	강원양구군	33	0.883091
6	강원양양군	10	0.948229
8	강원영월군	31	0.868520
11	강원원주시	78	0.860871
12	강원인제군	10	0.641060

Conclusion

본 분석은 주어진 지역의 교통사고에 대한 정보를 예측하기 위한 EDA이며 위 결과들을 바탕으로 하여 과거 교통사고들의 특성과 원인을 유추해 볼 수 있었다. 전반적으로 교통사망사고의 경우 요일과 사고유형이 다른 변수들에 관련이 높게 나왔다. 차량에 따라 탑승인원수의 범위가 정해지다보니 특정 차량조합의 경우 다른 조합에 비해 평균사상자수가 높게 나타남을 확인 할 수 있었다. 위 분석을 바탕으로 어떤 특성들에 집중하여야 하며 어떤 교차/파생 변수를 추가하여야 할지를 설정할 수 있을 것이다.
수치형 변수를 예측하는데 Random Forest 모델을, 범주형 변수를 예측하는데 Bayesian Net 모델을 사용하고자 한다.