목차

I. 서론

1. 프로젝트 개발 동기

옛 말 중에 먼저 가는데 순서 없다는 말이 있습니다. 우리는 언제 어디서 어떤 사건 사고가 발생할지 알 수 없습니다. 어제까지만 해도 건강하기만 하던 사람들이 갑작스런 사고를 당해 다치기도, 불행하게는 생을 마감하기도 합니다. 이러한 불확실한 미래는 우리들을 불안에 떨게 하며, 각종 소문이나 미신들을 믿게 만듭니다. 이렇듯 미래의 일은 단 0.1초 이후도 먼저 알 수 없습니다. 하지만 역사를 통해 우리가 미래를 대비하듯이, 모든 일은 과거의 일들을 통해서 어느 정도 예측이 가능합니다.

그래서 생각하였습니다. 각종 지역들에서 일어난 과거 사건 사고들에 대한 공공 데이터들과, 자연재해, 날씨 등 사건 사고에 관련될 만한 모든 데이터들을 수집해 통계를 낸다면, 당장 오늘이나 내일에 조금이라도 안전하게 다닐수 있지 않을까? 이렇게 시작한 아이디어는, 나의 알 수 없는 미래를 조금이나마 확률을 통해 알려준다는 재미를 선사함과 동시에 전체 지역적인 통계를 다시 한 번 가공하여서 떠돌고 있는 지역적 루머들이 (ex. '마계인천', '대구 = 고담시티' 등) 실제로 있음직한 말이라는 것을 검증해내는 아이디어로 나아가게 되었습니다.

2. 프로젝트의 목표

어플리케이션의 푸시알람을 통해 오늘의 운세처럼 매일 오늘은 내가 얼마나 안전한가에 대한 정보를 알림 받습니다. 어플리케이션을 통해 오늘 내가 사고가 날 확률, 좋지 않은 일을 당할 확률 등 여러가지 확률을 확인 할 수 있을 뿐만 아니라, 지역별 위험도 더 보기 항목을 통하여 지역별 위험도가 어느 정도인지 시각적으로 바로 확인 할 수 있습니다. 각종 사건 사고들이 발생한 데이터들을 종합하여 통계를 내 신뢰도 있는 지역별 위험도를 제공하고, 그 통계를 바탕으로 낸 각종사건들의 확률을 통하여 사용자들에게 재미를 선사하는 것이 최종 목표입니다.

II. 본론

1. 프로젝트의 진행 방법

1-1. 시도한 다양한 방법

프로젝트의 목표였던 각종 사건 사고들의 확률들을 구하기 위하여 구체적으로 어떤 확률을 구할 것인지에 대해 먼저 토의하였습니다. 사망확률, 다칠 확률, 병에 들확률 등의 확률들에 대해 나왔습니다. 다음은, 그 확률들을 구하기 위해 관련된 데이터종류를 찾기 시작했습니다. 공공 데이터 포탈을 이용하여 여러 가지 사건 사고에 관련된 데이터들을 수집하였습니다. 그 중 데이터가 가장 많았던 교통사고에 관련된 데이터 셋, 각종 범죄 통계 데이터 셋을 가지고 구하려고 했던 여러 가지 확률 중 생존할 확률에 대해서 구하려 시도하였습니다.

1-2. 봉착한 문제

공공 데이터 포털에서 제공하는 풍부한 데이터 셋을 가지고 충분히 데이터의 연관성을 찾을 수 있을 것이라 생각했지만, 그렇지 않았습니다. 본래 이 프로젝트에서의 중심축이 되는 생존확률을 구하기 위해 SPSS를 이용하여 데이터의 연관성을 찾아보려 했지만, 연관성을 위해 필요한 양식의 데이터를 찾을 수 없었습니다. 특히, 범죄데이터의 날짜 항목을 보았을 때, 어떤 데이터 셋은 요일만 표시되어있는 반면에 다른 데이터 셋은 월 통합으로 몇 건이 발생하였는지 표시되어있어 데이터 가공이 불가능한 상황이었습니다. 분산되어있는 데이터를 수작업으로 어떻게든 원하는 양식대로 가공하여 연관성을 찾으려 시도했지만, 위와 같은 문제와 시간상의 문제로 생존확률에 대한 데이터의 연관성을 찾는 것은 불가능하였습니다.

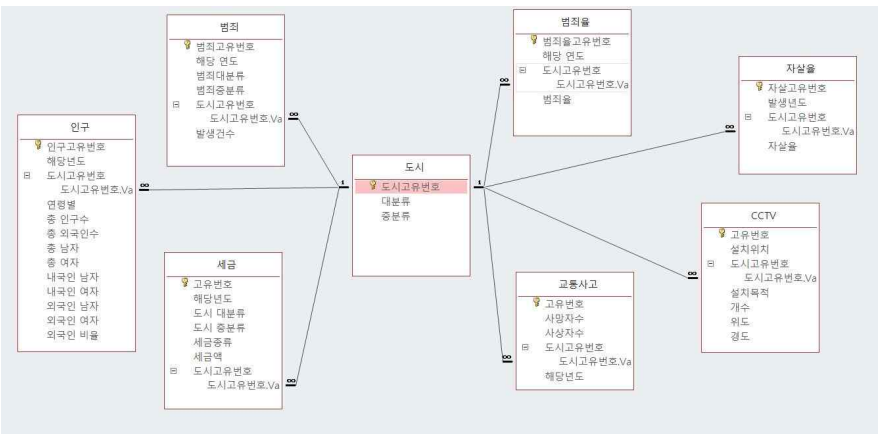
1-3. 해결 방안

위와 같은 문제점을 해결하기 위해 프로젝트의 방향을 다르게 설정하였습니다. 연관성을 찾기 힘든 생존확률보다, 연관성을 찾아보기 쉬운 범죄율을 기반으로 하여 다른 데이터와의 연관성을 알아내는 데이터의 분석으로 프로젝트의 방향을 잡았습니다.

1-4. 데이터 분석

1-4-1 사전준비

본격적인 데이터 분석에 앞서, SPSS (Statistical Package for Social Science)툴을 사용하기로 하였습니다. SPSS를 통한 빅 데이터 분석을 위해서, 지금껏 모으고 가공하였던 데이터 셋들을 시군구를 기준으로 합쳐 통합 데이터베이스를 만들었습니다. 통합한 데이터베이스의 ER다이어그램은 아래 <ER 다이어그램>그림과 같습니다. 이를 조인한 SQL문은 뒤에 부록에 수록하였습니다.



<ER 다이어그램>

이 데이터베이스로 SPSS에서 제공하는 상관분석과 회귀분석방법 두 가지를 이용하였습니다. 상관분석은 데이터의 크기를 가지고 서로 어떤 상관관계에 있는지 분석하는 방법입니다. 회귀분석은 데이터끼리 서로 인과관계에 있는지 알아보는 방법입니다. 즉, 데이터끼리 서로 영향을 주는 관계에 있는지에 대한 것을 알 수 있습니다.

어떤 데이터끼리 서로 연관성이 있는지와 어떤 요소가 범죄율에 가장 영향을 미쳤는지에 대해 알기 위해서 분석할 데이터의 목록을 정하였습니다. 엮어본 데이터의 목록은 아래와 같습니다. 아래 목록을 기준으로 SPSS에서 제공하는 두 분석방법을 적용시켜 보았습니다.

범죄율 데이터	외국인 인구수 비율, 자살율, 교통사고사망 및 부상, cctv 수, 세금데이터
전체 인구수 데이터	범죄율, 외국인 인구수 비율, 자살율, 교통사고 사망, cctv 수, 세금데이터
세금 데이터	범죄율, 전체 인구수, 자살율, 교통사고 사망 및 부상, cctv 수

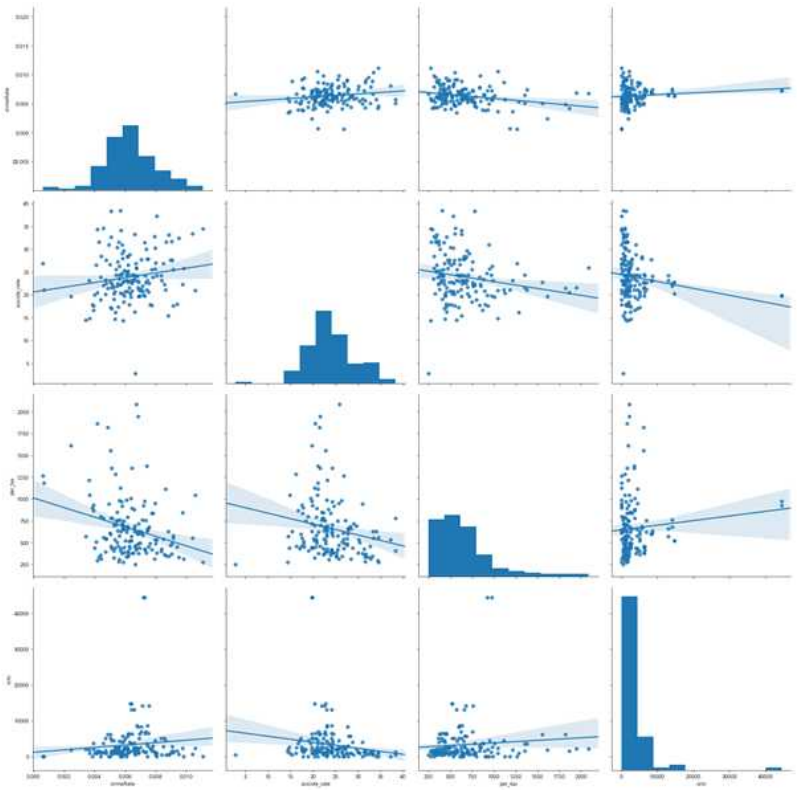
1-4-2 상관분석 결과

상관분석은 독립변수와 종속변수 간의 값들이 어떠한 선형적 관계를 갖고 있는지 분석하는 방법입니다. 자료 분석을 통하여 피어슨 상관계수를 구해 어떠한 상관관계를 가지고 있는지 알아보았습니다. 상관계수는 -1~1 사이의 값을 지니며,
절대 값이 0.1 이하면 상관관계가 존재하지 않고, 0.3 이하면 약한 상관관계, 그 이상은 더 강한 상관관계를 보여준다고 합니다. 또한 부호에 따라 음인지 양인지에대한 관계를 나타냅니다.

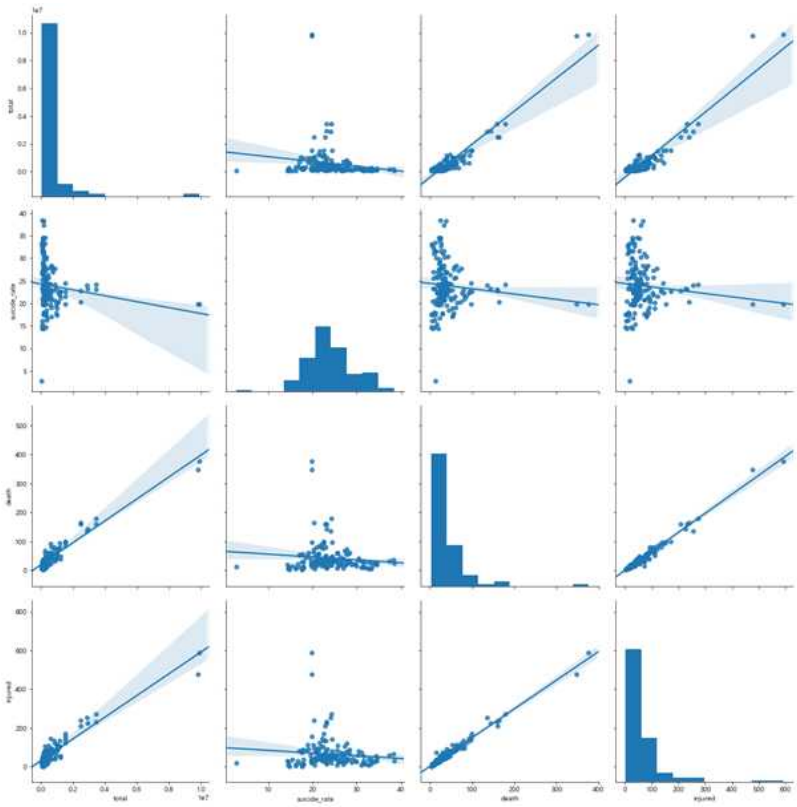
탐색적 자료 분석방법을 통하여 얻어낸 데이터 간의 상관관계는 아래 표와 같습니다.

데이터 연관		상관관계	강한 정도
범죄율	자살율	+	약함
	소득수준	-	약함
	CCTV	+	약함
인구수	자살율	-	약함
	교통사고 사망, 부상	+	매우강함
외국인	자살율	+	약함
	소득수준	+	약함
자살율	소득수준	-	약함
	교통사고 사망, 부상	-	약함
	CCTV	-	약함
교통사고 사망, 부상	CCTV	+	매우강함

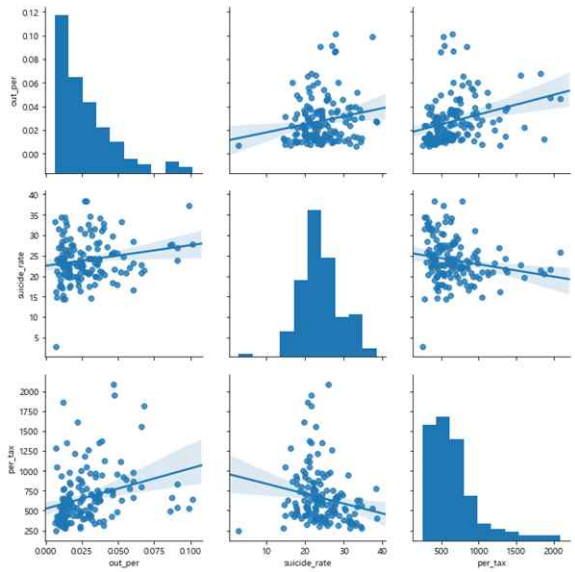
아래 그림들은 Seaborn 패키지의 Pairplot을 이용하여 각 연관 지은 데이터 중 상관관계를 지닌 데이터마다 시각화하여 나타낸 그래프들입니다.



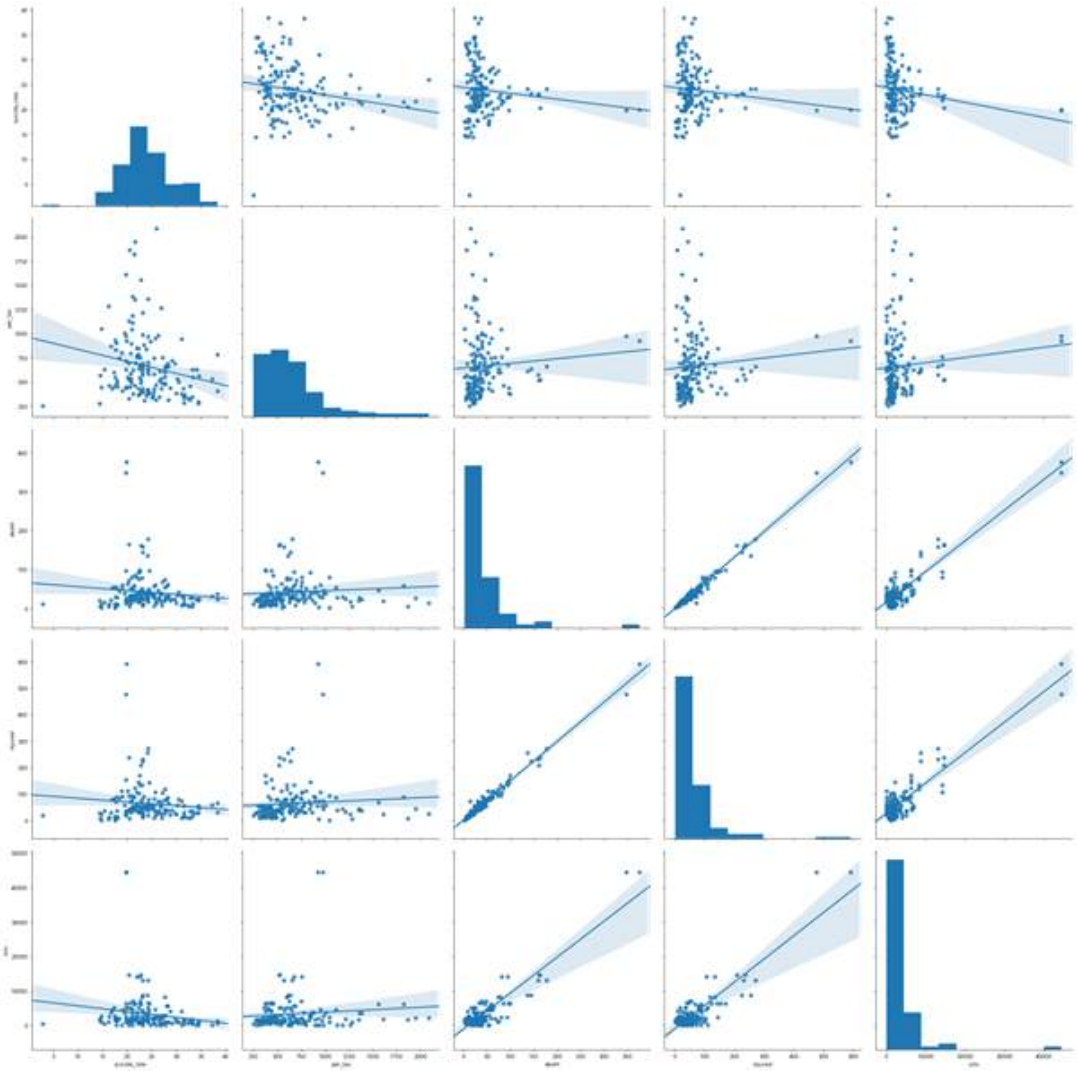
<범죄율과 자살율, 소득수준, CCTV개수>



<인구수와 자살율, 교통사고 사망자 및 부상자>



<외국인 비율과 자살율, 소득수준>



<자살율과 소득수준, 사망자, 부상자, CCTV개수>

1-4-3 회귀분석 결과

분석결과 이전에, 통계수치에 대한 이해가 필요하였습니다. 회귀분석모델에서 사용하는 통계 수치 중, durbin-watson 값은 0~4의 수치를 지니며, 2에 가까울수록 적합한 회귀분석 모델임을 알 수 있습니다. 또, 유의확률이 0.05 이하이면 서로 인과관계가 존재하는 데이터라고 볼 수 있습니다.

첫 번째 분석결과는 아래 <분석결과 1>과 같습니다. 먼저 R 값이 0.339로 독립변수와 종속변수가 약한 상관관계를 가짐을 알 수 있습니다. 그리고 durbin-watson 값이 1.423으로 약하긴 하지만 적합한 회귀모형임을 알 수 있습니다. 독립변수와 종속변수의 유의 관계에서, 유의확률 0.001로 범죄율이 1인당 내는 세금에 영향을 주는 것으로 나타났습니다. 베타값을 보면 -0.276으로 약하기는 하지만 범죄율이 올라감에 따라 1인당 내는 세금이 낮아진다는 결론을 얻을 수 있습니다.

모형 요약[b]

모형	R	R 제곱	수정된 R 제곱	추정값의 표준 오차	통계량 변화량					Durbin-Watson
					R 제곱 변화량	F 변화량	자유도1	자유도2	유의확률 F 변화량	
1	.339[a]	.115	.073	.001647106	.115	2.762	7	149	.010	1.423

a. 예측자: (상수), 인당 세금, death, suicide_rate, out_per, cctv, total, injured

b. 종속변수: crimeRate

ANOVA[a]

모형		제곱합	자유도	평균제곱	F	유의확률
1	회귀	.000	7	.000	2.762	.010[b]
	잔차	.000	149	.000		
	전체	.000	156			

a. 종속변수: crimeRate

b. 예측자: (상수), 인당 세금, death, suicide_rate, out_per, cctv, total, injured

계수[a]

모형		비표준화 계수		표준화 계수	t	유의확률
		B	표준오차	베타		
1	(상수)	.006	.001		8.044	.000
	total	5.315E-11	.000	.038	.114	.909
	out_per	.013	.007	.146	1.752	.082
	suicide_rate	3.385E-5	.000	.104	1.260	.210
	death	7.344E-6	.000	.208	.373	.709
	cctv	5.956E-8	.000	.194	.777	.438
	injured	-7.158E-6	.000	-.305	-.611	.542
	인당 세금	-1.391E-6	.000	-.276	-3.290	.001

a. 종속변수: crimeRate

<분석결과 1>

두 번째 분석결과는 아래 <분석결과 2>와 같습니다. R 값과 R의 제곱 값을 보면 두 값 모두 1에 근접해 독립변수와 종속변수간의 강한 상관관계가 있음을 알 수 있습니다 또한 durbin-watson 값도 2에 근접해 좋은 회귀모형입니다. 먼저 cctv와 범죄율, 자살율은 유의확률이 0.05를 넘어 인구수에 영향을 받지 않음을 알 수 있고, 사망자 수와 1인당 내는 세금의 양은 유의확률 0으로 서로 연관이 있으나, 이는 매우 상식적인 내용입니다. 유의확률이 0.04로 인구수에 의해 외국인의 비율이 영향을 받으나, 베타 값을 볼 때 음의 상관관계를 가집니다. 외국인의 비율은 전체 인구 수가 늘어날수록 줄어드는 다소 특이한 결과가 나왔습니다.

모형 요약[b]

모형	R	R 제곱	수정된 R 제곱	추정값의 표준 오차	통계량 변화량					Durbin-Watson
					R 제곱 변화량	F 변화량	자유도1	자유도2	유의확률 F 변화량	
1	.991[a]	.982	.981	165989.3759	.982	1369.291	6	150	.000	1.993

a. 예측자: (상수), out_per, cctv, crimeRate, suicide_rate, death, tax

b. 종속변수: total

ANOVA[a]

모형		제곱합	자유도	평균제곱	F	유의확률
1	회귀	2.264E+14	6	3.773E+13	1369.291	.000[b]
	잔차	4.133E+12	150	2.755E+10		
	전체	2.305E+14	156			

a. 종속변수: total

b. 예측자: (상수), out_per, cctv, crimeRate, suicide_rate, death, tax

계수[a]

모형		비표준화 계수		표준화 계수	t	유의확률
		B	표준오차	베타		
1	(상수)	-71259.105	75504.481		-.944	.347
	death	8313.024	697.460	.332	11.919	.000
	tax	.001	.000	.587	17.784	.000
	cctv	20.886	7.414	.096	2.817	.006
	crimeRate	14081818.68	7983155.454	.020	1.764	.080
	suicide_rate	-2738.255	2649.769	-.012	-1.033	.303
	out_per	-2055527.507	702791.077	-.033	-2.925	.004

a. 종속변수: total

<분석결과 2>

세 번째 분석결과는 아래 <분석결과 3>과 같습니다. R 값이 0.339로 독립변수와 종속변수 간에 약한 상관관계를 가짐을 알 수 있습니다. 그리고 durbin-watson 값이 1.880 으로 적합한 회귀모형임을 알 수 있습니다. 유의확률을 비교해 보았을 때, 0.05수치를 가진 범죄율이 1인당 내는 세금의 양에 의해 영향을 받고 있음을 볼 수 있습니다. 나머지 데이터들은 유의확률이 대부분 지나치게 높아 의미를 발견하지 못하였습니다.

첫 번째 분석결과에서 범죄율이 1인당 내는 세금의 양에 영향을 준다는 결론을 보았을 때, 세금과 범죄율은 서로 확실한 유의 관계가 있음을 알 수 있습니다.

모형 요약[b]

모형	R	R 제곱	수정된 R 제곱	추정값의 표준 오차	통계량 변화량					Durbin-Watson
					R 제곱 변화량	F 변화량	자유도1	자유도2	유의확률 F 변화량	
1	.311[a]	.097	.061	329.45504	.097	2.675	6	150	.017	1.894

a. 예측자: (상수), injured, crimeRate, suicide_rate, cctv, total, death

b. 종속변수: 인당세금

ANOVA[a]

모형		제곱합	자유도	평균제곱	F	유의확률
1	회귀	1742329.241	6	290388.207	2.675	.017[b]
	잔차	16281093.67	150	108540.624		
	전체	18023422.91	156			

a. 종속변수: 인당세금

b. 예측자: (상수), injured, crimeRate, suicide_rate, cctv, total, death

계수[a]

모형		비표준화 계수		표준화 계수	t	유의확률
		B	표준오차	베타		
1	(상수)	1139.487	150.157		7.589	.000
	total	-5.403E-5	.000	-.193	-.581	.562
	suicide_rate	-9.519	5.195	-.148	-1.833	.069
	crimeRate	-44942.331	15799.629	-.226	-2.845	.005
	cctv	.014	.015	.235	.936	.351
	death	-2.154	3.932	-.307	-.548	.585
	injured	1.681	2.342	.360	.718	.474

a. 종속변수: 인당세금

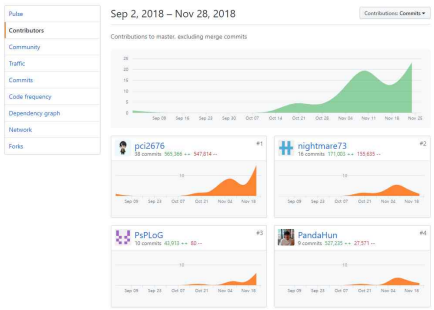
<분석결과 3>

2. 프로젝트의 협업 방법

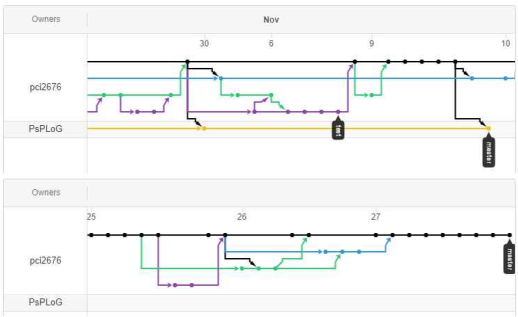
2-1. 기본적인 Git 기능 사용

많은 양의 데이터들을 수집해야 하고, 가공해야 하며 여러 가지 문서들을 수월하게 관리하기 위하여 Git을 많이 활용하였습니다. 개인 각자가 수집하고 가공한 뒤 따로 저장하면 어떤 것이 가장 최근의 데이터였는지 구분할 수 없게 되며, 불필요한 일을 하게 됩니다. 하지만 Git을통해 하나의 원격 저장소를 두고 작업 내역 들을 볼 수 있다면, 그러한 일들을 발생하지 않습니다. 그래서 이 팀은, 여러 가지 데이터들을 수집하여 데이터를 공유하였으며, 수작업으로 가공하여야 할 데이터들을 작업하고 올려놓아 쉽게 공유하였고, 체계적으로 관리하였습니다.

아래 <그림 1>은 각 팀원들이 이 원격 저장소에 commit을 한 횟수를 그래프로 나타낸 그림입니다. <그림 2>는 master와 다른 branch들이 어떤 식으로 갈려 나오고, 합병되었는지 시각적으로 나타내주며, 얼마나 commit을 진행해왔는지도 시각적으로 보여주는 그림입니다.



<그림 1>

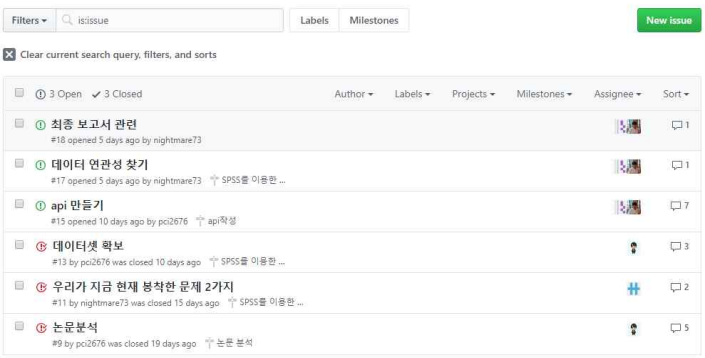


<그림 2>

2-2. 추가적인 Github 기능 사용

대부분의 회의나 의견제시 등은 대면 대화 또는 채팅으로 이루어졌습니다. 이러한 방식으로 소통할 경우 쉽게 잊혀 질 가능성이 매우 높아 문서화시키려 노력하였습니다. Github에서 이를 도와주는 기능을 몇 가지 추가적으로 이용하여 매우 효율적으로 협업을 진행하였습니다.

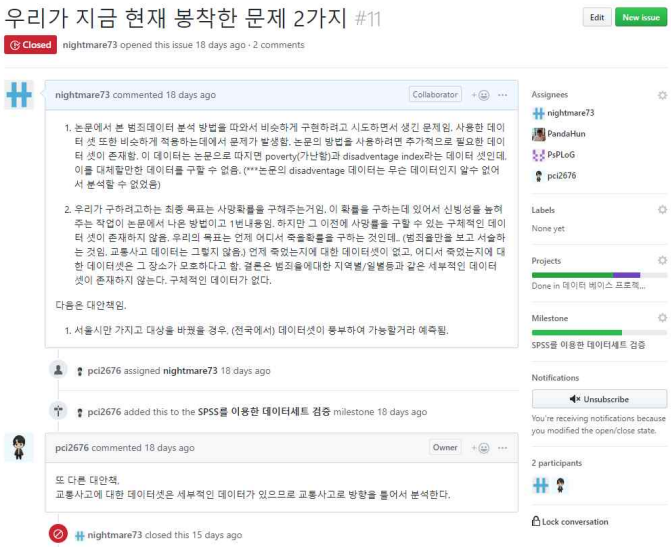
사용한 기능 중 하나는 Issue등록 기능입니다. 아래 <그림 3>과 같이 논의하고 싶은 주제를 제목으로 하여 글을 작성하여 팀원들과 github상에서 토의할 수 있는 기능입니다. 반



<그림 3>

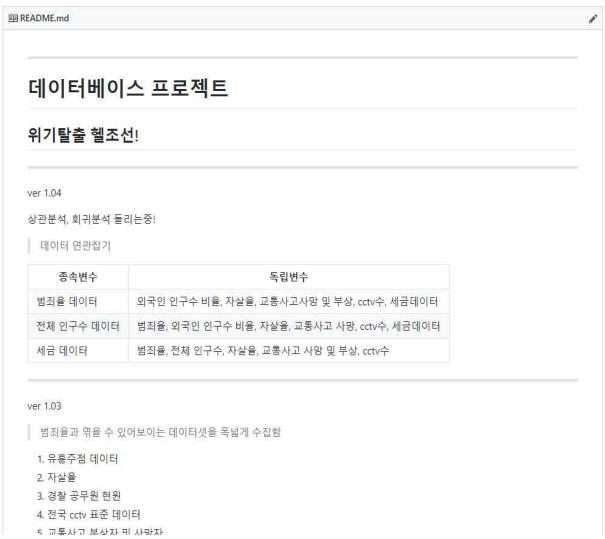
영하고 싶은 내용에 대해 자유롭게 작성하여 각 팀원이 무엇을 원하는지 쉽게 알 수 있었습니다. 또, 인터넷을 통해 언제든 확인 할 수 있고, 저장되어있으므로 내용이 유실될 걱정이 없었습니다.

등록된 Issue를 클릭하게 되면, <그림 4>와 같이 issue에 대한 내용이 나오게 되며, 아래에 팀원이 하고싶은말을 댓글처럼 달 수 있습니다. 또, <그림 4>의 우측 중앙에 있는 project, milestone의 주제를 달아서 구분하는 것 또한 가능합니다.



<그림 4>

마지막으로 Github에 들어가면 바로 볼 수 있는 README 파일을 활용하였습니다. 이 README 파일은 마크다운 형식으로 이루어져 있습니다. 문서의 양식을 편집하는 문법을 이용하여, 프로젝트의 큰 줄기가 업데이트 될 때마다 이 파일에 어떤 변화점이 생겼는지 적어놓았습니다. 아래 <그림 5>처럼 지속적으로 업데이트 하였으며, 모두가 프로젝트에 대해 동일한 생각을 할 수 있게 도와주었습니다.



<그림 5>

III. 결론

1. 종합 결론

본론의 상관분석에서 각 데이터들은 약하지만 대부분 상관관계를 보여주었습니다. 이는 탐색적 자료 분석방법에서 우리의 프로젝트에 그리 큰 기여를 하지 못했음을 의미합니다. 회귀분석에서는 각 데이터들이 어떤 상호 연관성을 보이는지 알아보았습니다. 이 자료들을 꽤 구체적인 수치들을 가지고 있었으며, 우리의 프로젝트인 데이터 연관성을 찾기에 아주 적합하였습니다.

범죄율과 관련이 있는 요인을 찾고 분석하여 구체적으로 어떤 관련성이 있는지 찾는 것이 이 프로젝트의 목표였습니다. CCTV의 개수, 교통사고 부상자 및 사망자, 인구수, 소득수준, 자살율을 범죄율과 관련이 있는 요인이라고 추정하였습니다. 위에 서술한 데이터 분석 방법 두 가지를 통하여 데이터를 분석해 보았을 때, 1인당 내는 세금의 양 이외의 데이터는 범죄율과 연관성을 보이지 못하였습니다.

본론의 회귀분석에서 세금 데이터가 범죄율에 영향을 주며, 범죄율 또한 1인당 내는 세금에 영향을 미치는 것을 확인하였습니다. 세금을 많이 내는 지역은 곧 소득수준이 높다고 생각하였습니다. 이는 소득수준이 높은 지역일수록 범죄율이 낮게 나옴을 증명하는 분석이었습니다.

범죄율과 연관을 지을 데이터 셋을 찾는 것이 가장 어려운 과제였습니다. 관련이 있을 것으로 추정되는 데이터를 찾는 것도 쉽지 않지만, 분석을 위해 알맞은 형식의 데이터 셋과 양이 존재하는지가 가장 문제였습니다. 우리는 이 프로젝트를 수행함으로써 적은 종류의 데이터 셋을 범죄율에 적용시켰지만, 이는 제한된 시간과 데이터를 수집했기 때문이라고 생각됩니다. 그러므로 충분한 시간을 가지고 더 많은 데이터를 수집해 나간다면, 범죄율과 연관되어있는 데이터 셋을 발견하여 충분히 극복할 수 있을 것이라 생각됩니다.

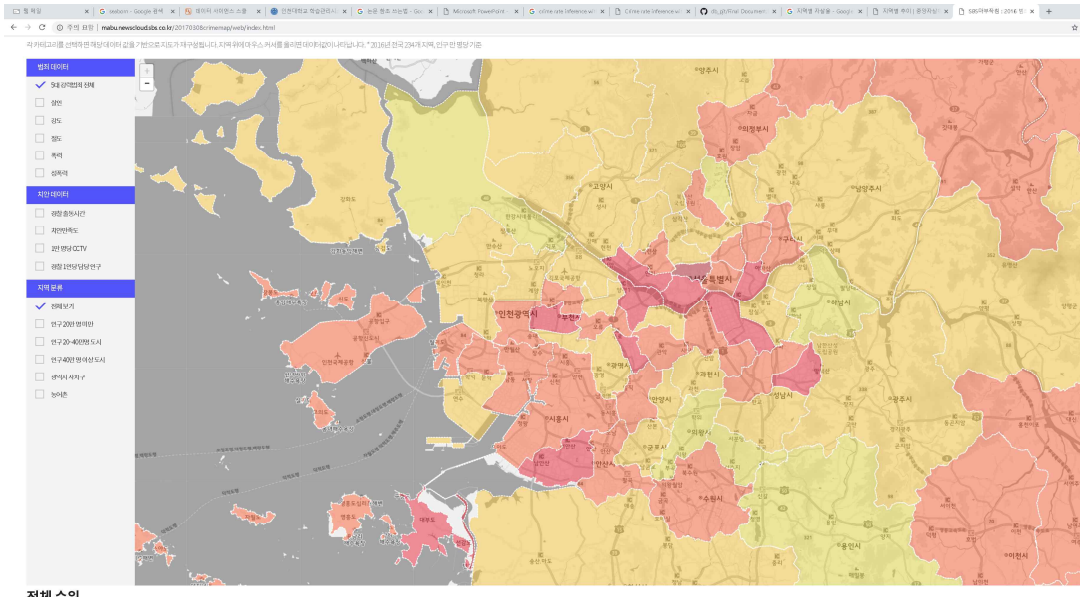
2. 참고 문헌 및 부록

2-1. 참고 문헌

-공공 데이터 포털 : https://www.data.go.kr

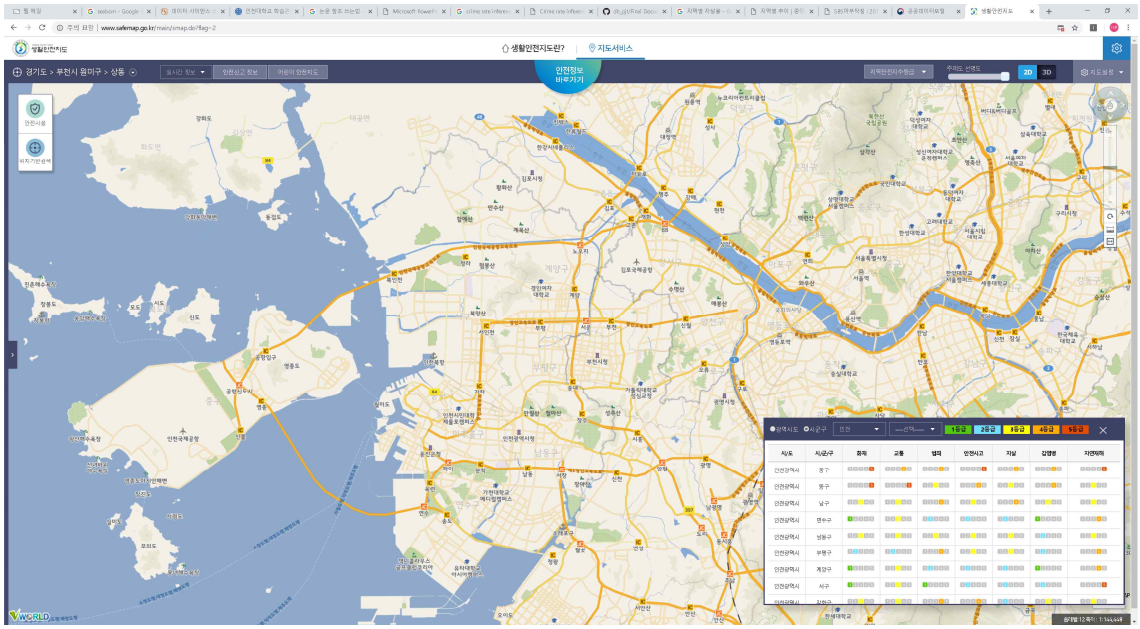
사용 데이터셋	세부 데이터셋	링크	기타
사고발생총괄현황	사고발생 현황(연간)	https://www.data.go.kr/dataset/15014225/fileData.do	매년 발간되는 재난연감 통계자료를 기초로 한 사고발생현황자료
교통사고정보	무단횡단사고다발지, 자전거사고다발지, 교통사망사고정보	https://www.data.go.kr/dataset/15003493/fileData.do	교통사고정보를 위치데이터 기반 제공 교통사고 항목별 정보제공
[범죄통계] 발생시간·요일	범죄 발생시간 및 요일	https://www.data.go.kr/dataset/3074459/fileData.do	
[범죄통계] 발생 및 검거 현황 (지방경찰청별)	[범죄통계] 발생 및 검거 현황 (지방경찰청별),[범죄통계] 2015 년 발생시간·요일	https://www.data.go.kr/dataset/3074451/fileData.do	전국 경찰관서에 고소, 고발, 인지 등으로 형사입건된 사건의 발생, 검거, 피의자에 대한 죄종별 분석 현황
[범죄통계] 범죄발생장소 (장소별)		https://www.data.go.kr/dataset/3074463/fileData.do	
[범죄통계] 범죄발생지 (지역별)		https://www.data.go.kr/dataset/3074462/fileData.do	
세금			
인구			
자살율	중앙 자살예방센터-자살통계-지역별추이	http://www.spckorea.or.kr/new/sub03/sub02.php	전국의 자살율을 지역별 연도별로 10 만명당 자살율

- SBS데이터저널리즘팀 <마부작침>의 2016 범죄여지도 :
https://mabu.newscloud.sbs.co.kr/20170308crimemap/web/index.html



-국립 재난 연구원<생활안전지도>

http://www.safemap.go.kr/main/smap.do?flag=2



- Hongjian Wang, Daniel Kifer, Corina Graif, Zhenhui Li.(2016) Crime Rate Inference with Big Data

2-2. 부록

-1)통합 데이터베이스 쿼리문

```
SELECT *
FROM (SELECT main.*,sub.cctv
        FROM (SELECT main.*,sub.death,sub.injured
                FROM (SELECT main.*,sub.tax
                        FROM (SELECT main.*,sub.suicide_rate
                                FROM (SELECT
main.crimeRate,sub.city_idx,sub.total,sub.out_per,sub.total_m,sub.total_w,sub.out_m,sub
.out_w
                                        FROM (SELECT *
                                                FROM crimeRate
                                                WHERE year=2016) as main
                                    INNER JOIN (SELECT *
                                                FROM population
                                                WHERE year=2016 AND age = '합계') as
sub
                                            ON main.city_idx=sub.city_idx) as main
                            INNER JOIN (SELECT city_idx,suicide_rate
                                        FROM suicide
                                        WHERE year = 2016) as sub
                                    ON main.city_idx=sub.city_idx) as main
                    INNER JOIN (SELECT sum(tax) as tax, city_idx
                                FROM tax
                                WHERE year=2016
                                GROUP BY city_idx) as sub
                    ON main.city_idx=sub.city_idx) as main
            INNER JOIN (SELECT city_idx,SUM(death) as 'death',SUM(injured) as 'injured'
                        FROM trafficAccident
                        WHERE year =2016
                        GROUP BY city_idx) as sub
                    ON main.city_idx=sub.city_idx) as main
    INNER JOIN (SELECT city_idx, sum(count) as cctv
                FROM cctv
                GROUP BY city_idx) as sub
            ON main.city_idx=sub.city_idx) as main
INNER JOIN (city)
ON main.city_idx=city_id;
```