

Big_Data_Project

201844101 추인혁

github 주소

https://github.com/inhyukChoo/201844101_BigData_Project

지역별 범죄 현황 분석

1. Language & Libraries

1. Python
2. pandas : 데이터 분석
3. matplotlib & seaborn : 시각화

2. 데이터 출처

공공 데이터 포털

<https://www.data.go.kr/data/15010720/fileData.do>

지역별 범죄 통계 데이터

<https://www.data.go.kr/data/3074461/fileData.do>

범죄 검거 기간 데이터

해당 분석에서 사용 된 데이터는 공공 데이터 포털 경찰청에서 제공하는 경찰청_범죄 발생 지역별 통계 데이터와 마찬가지로 경찰청에서 제공하는 경찰청_범죄발생부터 검거까지의 기간 데이터입니다.

데이터 수집 기간

해당 데이터들은 2022년 12월 31일을 시작으로 2023년 9월 13일까지의 데이터를 담고 있습니다. 데이터의 최신 갱신은 매년 9월 13일날 이루어지며, 데이터 업데이트를 통해 최신 데이터를 확인할 수 있습니다.

3. 배경

지역별 범죄 현황을 분석하면 다양한 정보를 파악할 수 있습니다.

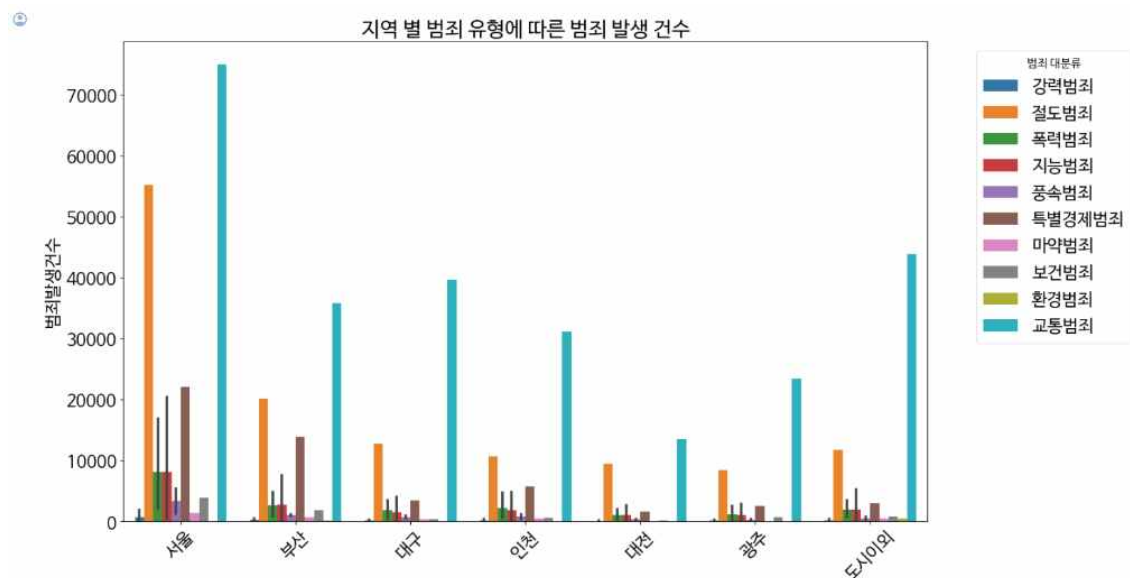
일반적으로 이런 분석은 지역 내에서 발생한 다양한 범죄의 종류와 빈도, 특정 범죄의 경향 등을 이해하는 데 도움이 됩니다.

예로 어떤 지역에서 어떤 종류의 범죄가 얼마나 자주 발생하는지를 파악할 수 있습니다. 다양한 범죄 유형 간의 상관관계를 조사하여 특정 범죄가 다른 범죄에 영향을 미치는지, 특정 지역에서 특정 범죄 유형이 더 많이 발생하는지를 파악할 수 있습니다.

범죄에 대한 검거율과 수사 속도를 지역별로 비교하여 효과적인 수사 및 예방 활동이 이루어지는 곳과 그렇지 않은 곳을 확인할 수 있습니다.

4. 데이터 분석

우선 대한민국 지역별 범죄 유형에 따른 범죄 발생 건수가 가장 높은 곳은 어느 지역이며 어느 범죄가 가장 큰 비중을 차지하고 있을까요?



우선 각 지역별로 가장 범죄율이 높은 범죄 항목은 교통범죄이며, 서울 지역이 가장 범죄 발생 건수가 높은 것을 확인할 수 있었습니다.

그렇다면 두 번째로 높은 비율의 범죄는 무엇일까요? 두 번째로는 절도 범죄가 높은 비율을 차지하고 있었습니다. 그 뒤로는 특별경제범죄, 폭력범죄, 지능범죄, 풍속범죄, 보건범죄, 마약범죄, 강력범죄, 환경범죄의 순으로 비율이 나타났고 가장 범죄율이 높은 지역은 서울인 것을 확인할 수 있었습니다.

해당 차트는 지역별 범죄 발생률 비교를 위해 Seaborn 라이브러리를 사용하여 바 차트(bar plot)를 사용하여 데이터 시각화를 진행하였습니다.

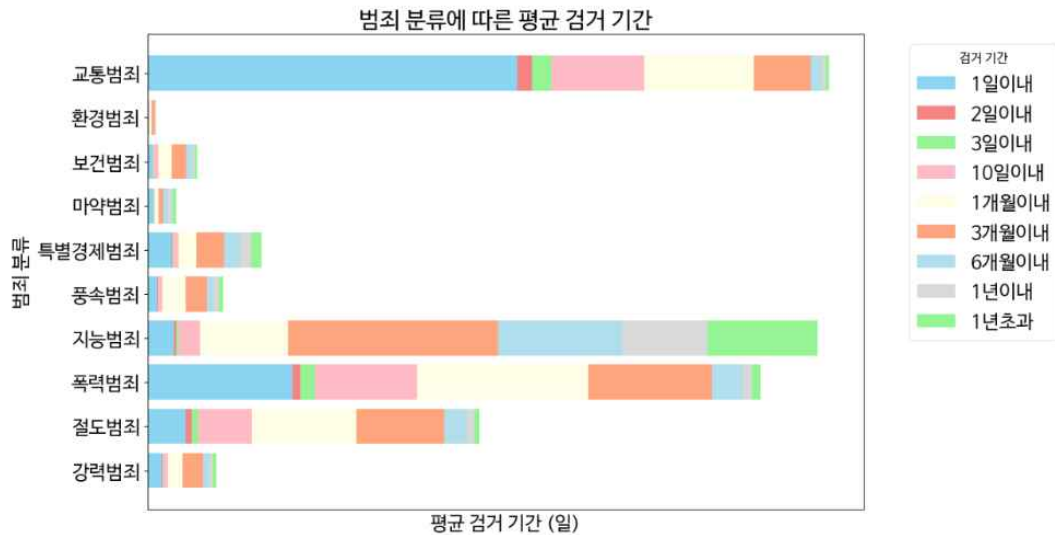
그렇다면 이번엔 모든 지역 통합 상위 10 범죄 유형을 토대로 해당 데이터를 비율로 나타내 확인해볼까요?



모든 지역을 통합한 상위 10 범죄 유형을 전체 비율로 나타냈을 때 교통 범죄가 34.9%로 가장 많았습니다. 다음으로 지능범죄와 절도범죄가 17.1%로 동일한 비율을 보여주고 있고, 폭력범죄-폭행 10.3% 특별경제범죄 6.9% 폭력범죄-손괴 3.8% 지능범죄 3.4% 폭력범죄-상해 3.1% 폭력범죄-폭력행위등 2.3% 풍속범죄 1.3%로 지역 별 상위 10의 범죄 비율을 확인할 수 있었습니다.

해당 차트는 범죄 유형 별로 발생 건수를 합산하고, 상위 N개의 범죄 유형에 대한 비율을 파이 차트(pie chart)로 데이터 시각화를 진행했습니다.

다음으로는 범죄 분류에 따른 평균 검거 기간을 범죄 검거 기간 데이터를 통해 알아보겠습니다.



범죄 분류에 따른 평균 검거 기간은 교통 범죄는 1일 이내에 검거가 많이 진행되었고, 풍속범죄와 폭력범죄 절도범죄는 1개월 이내 환경범죄와 보건범죄 마약범죄 특별경제범죄 지능범죄 강력범죄는 3개월 이내로 검거가 많이 진행되었음을 확인할 수 있었습니다.

해당 차트는 범죄 분류에 따른 평균 검거 기간을 시각화하기 위한 가로 막대 그래프(bar-plot)을 사용하여 데이터 시각화를 진행했습니다.

5. 데이터 수집, 데이터 전처리, 데이터 변환 과정

● 데이터 수집

`crime_data = pd.read_csv('범죄_지역별_통계.csv', encoding='euc-kr')` 해당 코드는 `pd.read_csv` 함수를 사용하여 '범죄_지역별_통계.csv' 파일을 불러와 `crime_data`라는 데이터프레임에 저장합니다. 그리고 파일 경로와 인코딩 방식을 지정하여 데이터를 읽어옵니다.

● 데이터 전처리

`crime_data_melted = crime_data.melt(id_vars=['범죄대분류', '범죄중분류'], var_name='도시', value_name='범죄발생건수')` `melt` 함수를 사용하여 데이터프레임을 재구조화합니다. '범죄대분류'와 '범죄중분류' 열을 기준으로 '도시'와 '범죄발생건수' 열로 변환합니다. 이는 데이터를 시각화하기 용이한 형태로 변환하는 과정으로, 데이터프레임의 구조를 변경하고 있습니다.

● 데이터 필터링

`violent_crimes = crime_data[crime_data['범죄대분류'] == '강력범죄']`
crime_data 데이터프레임에서 '범죄대분류'가 '강력범죄'인 행들만을 선택하여 violent_crimes에 저장합니다. 이는 주어진 데이터에서 '강력범죄'에 해당하는 부분만을 추출하는 과정으로, 특정 조건을 만족하는 데이터를 필터링하는 역할을 합니다.

● 데이터 변환 과정

해당 과정에서 데이터를 재구조화하여 시각화에 적합한 형태로 변환하고 있습니다. 데이터 수집, 전처리, 가공 모두를 아우르는 하나의 과정으로 볼 수 있습니다. 또한 데이터가 과도하게 많은 점 너무 많이 세분화 되어있는 데이터들은 직접 csv파일에서 데이터를 조금 제거하여 더욱 시각적으로 보기 쉽게 변환했습니다.

6. 통계 및 분석 결과

6-1. 지역별 범죄율 현황 분석

통계 분석 결과

서울, 경기, 부산 등 대도시에서 범죄 발생 건수가 상대적으로 높게 나타남.
산간 지역이나 소규모 도시에서는 상대적으로 범죄 발생 건수가 낮음.

6-2. 지역별 범죄 평균 검거 기간 분석

통계 분석 결과

일부 지역에서는 특정 범죄에 대한 검거 기간이 다른 지역에 비해 상대적으로 높게 나타남. 대체로 검거 기간이 짧은 지역이 많으나, 일부 범죄에 대해서는 검거가 어려운 지역이 확인됨.

7. 프로젝트 결론

1. 지역별 범죄 현황 분석

1-1 범죄 발생 건수

프로젝트를 통해 수집한 지역별 범죄 발생 건수 데이터를 분석한 결과, 대도시인 서울, 경기, 부산 등에서 범죄 발생 건수가 상대적으로 높게 나타났습니다. 특히 교통 범죄가 가장 많은 비중을 차지하고 있으며, 두 번째로는 절도 범죄가 높은 비율을 차지하고 있습니다. 다양한 범죄 유형 간의 상관관계를 통해 어떤 범죄가 다른 범죄에 미치는 영향을 파악할 수 있었습니다.

1.2 범죄 발생률 시각화

데이터 시각화를 통해 지역별 범죄 발생률을 확인한 결과, 서울 지역이 가장 범죄 발생 건수가 높은 것을 확인할 수 있었습니다. 교통 범죄와 절도 범죄가 두드러지게 나타나며, 이를 통해 각 지역의 범죄 특성을 시각적으로 확인할 수 있었습니다.

2. 범죄 평균 검거 기간 분석

2-1 검거 기간

범죄 분류에 따른 평균 검거 기간을 분석한 결과, 교통 범죄는 대체로 1일 이내에 검거되는 경향이 있었습니다. 그 외에도 특별경제범죄, 폭력범죄, 절도범죄 등은 1개월 이내에 검거되는 비율이 높았습니다. 이를 통해 특정 범죄에 대한 검거 기간이 다른 지역에 비해 상대적으로 높게 나타나는 경향을 확인할 수 있었습니다.

2-2 검거 기간 시각화

범죄 분류에 따른 평균 검거 기간을 가로 막대 그래프로 시각화하여 확인한 결과, 각 범죄에 대한 검거 기간의 차이가 나타났습니다. 교통 범죄의 빠른 검거 속도와 일부 범죄에 대한 상대적으로 느린 검거 속도가 시각적으로 확인 가능했습니다.

3. 결론

프로젝트를 통해 수집한 데이터를 기반으로 한 지역별 범죄 현황 분석 결과, 대도시에서의 범죄 발생이 상대적으로 높으며, 특정 범죄에 대한 검거 기간도 지역에 따라 다르게 나타났습니다. 이를 통해 지역별 범죄 예방 및 대응에 대한 인사이트를 얻을 수 있었습니다. 향후 더 많은 데이터와 깊은 분석을 통해 보다 정확하고 실질적인 정책 제안이 가능할 것으로 기대됩니다.