데이터_마이닝_미니과제_공공데이터분석(경찰청_범죄자 생활정도, 혼인관계...)

본 실험에서 사용된 전체 코드는 GitHub 저장소에서 확인할 수 있습니다:
[GitHub 링크](https://github.com/Foordle/PSU_DataMining_202055593)

Date: 2024. 12. 28
Student ID: 202055593
Name: 전승윤

1. 서론

본 보고서는 공공 데이터를 활용하여 데이터 마이닝 기술을 적용한 분석 과정을 설명하고, 그 결과를 도출하는 것을 목표로 한다. 특히 클러스터링 기법과 시각화를 통해 '경찰청_범죄자 생활정도, 혼인관계 및 부모관계'데이터를 이해하고, 각 클러스터의 특징을 분석한다.

2. 본론

2-1. 이론적 배경

데이터 마이닝은 대규모 데이터에서 유용한 패턴과 정보를 추출하는 과정으로, 클러스터링은 대표적인 비지도 학습 방법 중 하나이다. 클러스터링은 데이터 간의 유사성을 기반으로 그룹화하여 숨겨진 구조를 파악할 수 있도록 한다. 본 과제에서는 K-Means 클러스터링을 사용하여 데이터를 분석하며, 이를 통해 데이터 분포와 특성을 이해하고자 한다.

2-2 본론 2: 주요 코드 리뷰 및 실험 과정

코드는 아래와 같이 8단계로 구성된다.

1) 데이터 로드

CSV 파일을 데이터프레임 형태로 불러오는 역할을 한다. 한글 데이터를 처리하기 위해 encoding='cp949'을 사용하였다.

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.cluster import KMeans
from sklearn.preprocessing import StandardScaler

# 1. 데이터 불러오기
file_path = '경찰청_범죄자 생활정도, 혼인관계 및 부모관계_12_31_2020.csv'
data = pd.read_csv(file_path, encoding='cp949')
```

코드1

2) 클러스트링을 위한 숫자형 데이터 선택

클러스터링에 적합한 데이터만 선택하기 위해 숫자형 컬럼을 필터링하였다.

```
# 2. 클러스터링을 위한 숫자형 데이터 선택
numeric_columns = data.select_dtypes(include='number').columns
clustering_data = data[numeric_columns]
print(numeric_columns)
```

코드2

3) 데이터 정규화

변수 간의 스케일 차이를 줄이기 위해 데이터를 정규화, StandardScaler를 사용하여 평균이 0, 표준편차가 1인 데이터로 변환하였다.

```
# 3. 데이터 정규화
scaler = StandardScaler()
scaled_data = scaler.fit_transform(clustering_data)
```

코드3

4) K-Means 클러스트링 수행(K = 4)

K-Means 알고리즘을 사용하여 클러스터 레이블을 각 데이터에 할당

```
# 4. KMeans 클러스터링 수행
kmeans = KMeans(n_clusters=4, random_state=42)  # 클러스터 개수는 4로 설정
data['Cluster'] = kmeans.fit_predict(scaled_data)
```

코드4

5) 범죄대분류와의 클러스터 관의 관계분석 시각화

클러스터링 결과와 범죄대분류 간의 관계를 분석, pd.crosstab을 사용하여 범죄대분류와 클러스터의 빈도를 교차표로 계산하였다.

```
# 5. 범죄대분류와 클러스터 간 관계 분석
cross_tab = pd.crosstab(data['범죄대분류'], data['Cluster'])
```

코드5

6) 클러스터 분포 시각화

박스플롯을 통해 각 특성별 클러스터의 분포를 비교

```
import seaborn as sns
import matplotlib.pyplot as plt
```

```
import matplotlib.font_manager as fm

# 한글 폰트 설정
font_path = 'H2GPRM.TTF'
font_prop = fm.FontProperties(fname=font_path)
plt.rc('font', family=font_prop.get_name())

# 특정 변수의 분포를 클러스터별로 시각화하는 루프
plt.figure(figsize=(12, 6))  # 초기 그래프 설정

# 데이터의 숫자형 컬럼 확인
numeric_columns = data.select_dtypes(include='number').columns

# 루프를 통해 모든 y 값에 대해 클러스터별 시각화
for col in numeric_columns:
    if col != 'Cluster':  # Cluster는 x축으로 사용되므로 제외
        plt.figure(figsize=(12, 6))
        sns.boxplot(x='Cluster', y=col, data=data)
        plt.title(f'클러스터별 {col} 분포')
        plt.xlabel('Cluster')
        plt.ylabel(col)
        plt.show()
```

코드6

7) 범죄대분류와 클러스터 간 분포 시각화

범죄대분류와 클러스터 간의 분포를 히트맵으로 시각화

```
# 범죄대분류와 클러스터 간 분포
crime_distribution = pd.crosstab(data['Cluster'], data['범죄대분류'], normalize='index')

# 히트맵 시각화
plt.figure(figsize=(12, 8))
sns.heatmap(crime_distribution, annot=True, cmap='Blues', fmt='.2f')
plt.title('클러스터별 범죄대분류 분포')
plt.xlabel('범죄대분류')
plt.ylabel('Cluster')
plt.show()
```

코드7

8) 클러스터별 평균값 계산

클러스터별 주요 특징을 확인하기 위해 평균 값을 계산

```
import pandas as pd

# 숫자형 데이터만 선택하고 Cluster 열 추가
numeric_columns = data.select_dtypes(include=['number']).columns.tolist()  # 숫자형 열 목록 가져오기
if 'Cluster' not in numeric_columns:  # Cluster 열이 숫자형 데이터에 없을 경우 추가
    numeric_columns.append('Cluster')

# 숫자형 데이터 선택
numeric_data = data[numeric_columns]

# 클러스터별 평균 값 계산
cluster_characteristics = numeric_data.groupby('Cluster').mean()

# 결과 출력
print(cluster_characteristics)

# 결과를 엑셀 파일로 저장
output_file_path = 'cluster_characteristics.xlsx'
cluster_characteristics.to_excel(output_file_path, sheet_name='Cluster_Averages')
print(f"클러스터별 평균 값이 '{output_file_path}' 파일로 저장되었습니다.")
```

코드8

3. 실험 결과

1) 각 그래프를 통해 알 수 있는 사실

- 박스플롯 분석 + 클러스터별 평균 값 계산

Cluster 0:
생활정도: 하류와 중류 계층이 주요 구성원이며, 상류는 거의 포함되지 않는다.
혼인관계: 유배우자가 다수를 차지하며, 이혼율은 다른 클러스터에 비해 낮은 편이다.
미혼자 부모 관계: 실부모가 주를 이루며, 계부모 관계는 미미하다.
전반적으로 경제적으로 하위 계층에 속하며, 안정적인 가족 관계를 유지하는 구성원이 많다.

Cluster 1:
생활정도: 중류와 하류 계층이 혼합되어 있으며, 상류 비율이 약간 더 높다.
혼인관계: 유배우자와 이혼율이 모두 높은 편이다.
미혼자 부모 관계: 실부모와 계부모 모두 균형 있게 포함되어 있다.
경제적 다양성이 존재하며, 가족 형태가 다양하게 나타난다.

Cluster 2:
생활정도: 대부분 하류 계층으로 구성되며, 중류와 상류 계층은 거의 없다.
혼인관계: 유배우자가 적고, 이혼율이 매우 낮다.
미혼자 부모 관계: 실부모 관계가 거의 대부분을 차지한다.
전반적으로 가장 낮은 경제 수준이며, 단일한 가족 구조가 주를 이루는 특징이 있다.

Cluster 3:
생활정도: 상류 계층의 비중이 가장 높으며, 중류와 하류 계층도 혼합되어 있다.
혼인관계: 이혼율이 가장 높으며, 동거와 유배우자 관계도 상당히 높다.
미혼자 부모 관계: 다양한 부모 관계가 포함되어 있으며, 실모무부 관계가 두드러진다.
경제적으로 상위 계층이 주를 이루며, 가족 관계가 복잡한 양상을 보인다.

- 히트맵 분석
Cluster 0
폭력범죄가 가장 높은 비중을 차지하며 또한 대다수의 폭력범죄, 보건범죄, 경제범죄, 지능범죄가 이 cluster에 속함

Cluster 1
대다수의 교통범죄, 기타범죄, 절도범죄와 일부의 폭력범죄가 이 cluster에 속함

Cluster 2
대다수의 강력, 노동, 마약, 병력, 선거, 안보, 풍속, 환경범죄와 일부의 폭력, 지능범죄가 이 cluster에 속함

Cluster 3
지능범죄가 가장 높은 비중을 차지하며 또한 대다수의 지능범죄가 cluster에 속함

2) 최종 결론
- Cluster 0
특징:
중하위 생활 수준(하류 및 중류 계층)에 속하며, 상류 계층은 거의 포함되지 않음.
혼인 상태에서 유배우자의 비율이 상대적으로 높고, 이혼율이 낮음.
미혼자 부모 관계(무부모) 및 관련 변수의 값이 낮은 편.
주요 범죄대분류: 폭력범죄, 보건범죄, 경제범죄, 지능범죄.

결론:
중하위 계층에서 안정적인 혼인 상태를 가진 개인들이 폭력범죄와 지능범죄와 연관이 있음. 이는 특정 생활 수준과 안정성 부족이 이러한 범죄를 유발할 가능성이 있음을 시사.

- Cluster 1
특징:
생활정도(중류 및 상류) 계층에 속하며, 상류 계층 특성이 두드러짐.
혼인 상태가 다양하며, 유배우자, 이혼, 동거 등 폭넓게 분포함.
미혼자 부모 관계(실부모, 계부모 등)에서 높은 평균 값을 보임.
주요 범죄대분류: 교통범죄, 기타범죄, 절도범죄.

결론:
경제적 활동이 활발한 중류 및 상류 계층에서 교통범죄와 기타범죄가 두드러짐. 다양한 가족 형태와 혼인 상태가 범죄와 연관성을 가질 가능성이 있음.

- Cluster 2
특징:
전반적으로 낮은 생활 수준(하류 계층)에 속하며, 중류와 상류 계층은 거의 포함되지 않음.
유배우자 비율이 적고, 이혼율이 매우 낮음.
미혼자 부모 관계에서는 실부모가 주를 이룸.
주요 범죄대분류: 강력범죄, 마약범죄, 노동범죄, 환경범죄.

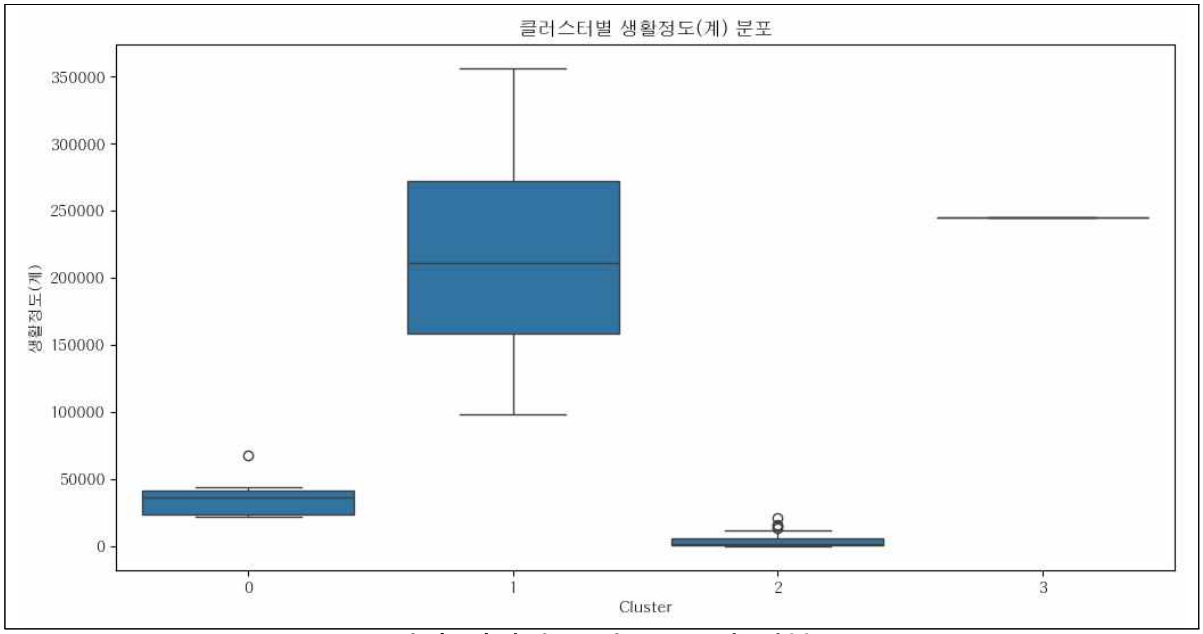
결론:
낮은 경제 수준과 단일한 가족 구조를 가진 개인들이 강력범죄와 환경범죄와 연관이 있음. 이는 경제적 불안정성과 관련된 특수한 요인이 영향을 미쳤을 가능성이 높음.

- Cluster 3
특징:
극단적으로 높은 생활 수준(상류 및 미상)과 관련이 깊음.
미혼자 부모 관계(무부모) 변수가 높게 나타나며, 가족 구조가 복잡한 경향을 보임.
혼인 상태에서는 동거와 이혼 비율이 상대적으로 높음.
주요 범죄대분류: 지능범죄.

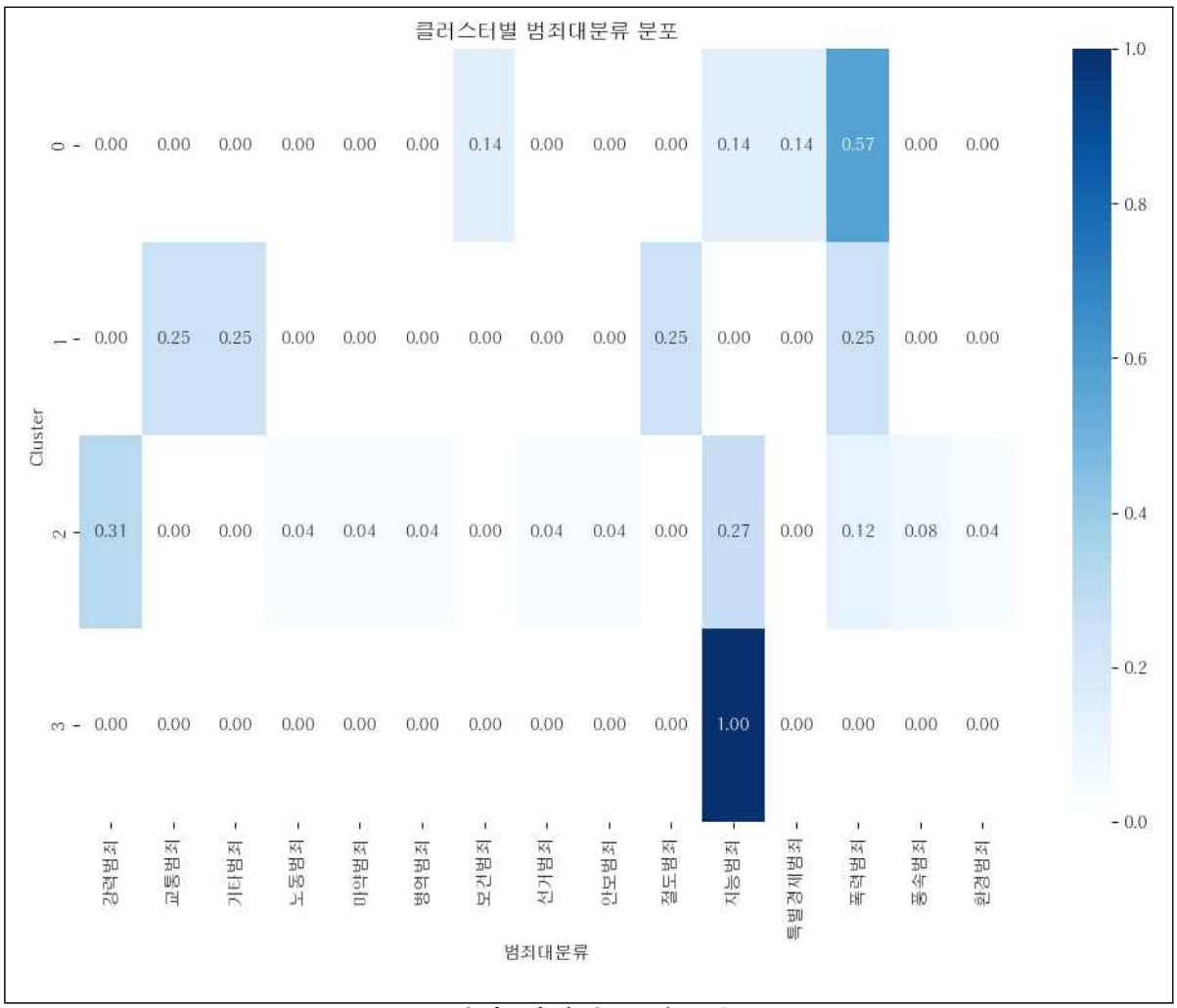
결론:
상류 계층의 극단적인 생활 수준에서 지능범죄가 두드러짐. 이는 경제적 불균형, 가족 관계 문제, 교육 수준과 같은 요인이 주요 원인일 가능성이 높음.

3) 한계
박스플롯을 살펴보면 각 클러스터 간의 구분이 생각만큼 뚜렷하지 않아 아쉬움이 있음.
k값을 증가시키는 방향으로 시도했으나, 메모리 부족 문제로 인해 VScode가 종료되어

추가적인 결과를 확인하지 못함.



결과 이미지1. 박스플롯의 일부



결과 이미지2. 히트맵