공공데이터를 이용한 사회문제 발견과 해결책 모색

최수연 교수

mibm400@hanmail.net

목차

학습 목표

공공 데이터를 읽어와 데이터의 특징을 파악할 수 있다.

- 서울시 자치구별 CCTV 현황 데이터
- 서울시 범죄현황 데이터

읽어온 데이터로부터 분석에 필요한 부분을 정제할 수 있다.

정제된 데이터를 분석하고 시각화 할 수 있다.

분석한 데이터로부터 새로운 인사이트를 발견할 수 있다.

서울시 자치구별 CCTV 현황 파악하기

1. 데이터 읽어오기

import pandas as pd # 데이터 관리와 정제 기능을 가진 라이브러리

- 데이터 읽어오기

  # 인코딩방식 :'cp949'(MS office에서 저장한 파일 형식) /'utf-8'(그 외 일반적인 경우)

  # .csv 파일을 불러올 때

  - 변수명 = pd.read_csv('파일경로명', sep= ',', encoding='인코딩방식')
  - sep 옵션은 생략하면 ',' 로 인식
  - header = 숫자 옵션은 위에서 몇 째줄 부터 읽어올지 지정(줄 수는 0부터 시작)

1. 데이터 읽어오기

import pandas as pd # 데이터 관리와 정제 기능을 가진 라이브러리

- 데이터 읽어오기

  # 인코딩방식 :'cp949'(MS office에서 저장한 파일 형식) /'utf-8'(그 외 일반적인 경우)

  # .txt 파일을 불러올 때

  - 변수명 = pd.read_csv('파일경로명', sep= '₩t', encoding='인코딩방식')

  # .xlsx 파일을 불러올 때

  - 변수명 = pd.read_excel('파일경로명')

  - header = 숫자 옵션은 위에서 몇 째줄 부터 읽어올지 지정(줄 수는 0부터 시작)

2. 데이터 살펴보기

- 데이터에서 일부 내용 보기
  - 변수명 : 전체 데이터 보기
  - 변수명.**head()** : 위에서 5행 보기 / 변수명.head(3) : 위에서 3행 보기
  - 변수명.**tail()** : 아래서 5행 보기 / 변수명.tail(3) : 아래서에서 3행 보기

2. 데이터 살펴보기

- 데이터 형식 보기
  - 변수명.**shape** : 행,열 수 보기
  - 변수명.**index** : index 범위 보기
  - 변수명.**columns** : 열이름 보기
- 데이터 정보 보기
  - 변수명.describe() : 숫자형 데이터의 통계치 계산
  - 변수명.info() : 데이터 타입, 각 아이템 개수, 누락데이터 수 등 확인

3. 데이터 정리하기

- 데이터에서 열이름 변경하기
  - 변수명.rename(columns = {'열이름': '새로운 열이름'}, inplace= True)

    # 데이터가 저장된 변수명의 열이름을 새로운 열이름으로 변경

    # inplace = True 옵션은 원본데이터를 변경함
- 행 데이터 삭제
  - 변수명.drop(index='행번호', axis=0) : index가 0인 행 삭제

    # 여러행 삭제 : 변수명.drop(index=[0,1,2], axis=0) : index가 0,1,2인 행(3줄) 삭제

    # inplace= True 옵션을 추가하면 원본을 변경함

3. 데이터 정리하기

- 열 데이터 삭제
  - 변수명.drop(colums='열이름', axis=1) : '열이름'열 삭제

    # 여러열 삭제 :

    변수명.drop(colums=['열이름1','열이름2'], axis=1) : '열이름1','열이름2'열 삭제

    # inplace= True 옵션을 추가하면 원본을 변경함
- 인덱스 리셋
  - 변수명.reset_index(drop=True, inplace=True)

    #drop=True 옵션은 기존 인덱스는 버리고 새로 인덱스 설정

4. 데이터 자세히 보기(CCTV 현황)

CCTV의 전체 개수가 가장 적은/가장 많은 상위 5개 구는 어디일까?

- 변수명.sort_values(by='정렬기준 열이름', ascending=True)

  # ascending = True : 오름차순, False : 내림차순

최근 3년간 CCTV 증가율을 계산하여 '최근증가율'을 알아보자.

- 변수명['열이름'] + 변수명['열이름'] : 사칙연산 가능
- 변수명['최근증가율'] = 열단위 연산식 : 해당 열이름이 없으면 새로운 열 생성

최근증가율'이 가장 높은 상위 5개 구는 어디일까?

- 변수명.sort_values(by='최근증가율', ascending=False).head()

서울시 인구현황 분석

1. 서울시 인구현황 데이터 읽어오기
2. 인구 현황 데이터 정리하기
3. 인구현황 데이터 자세히 보기
    1) 인구수가 가장 높은 상위 5개 구는 어디일까?
    2) 여성인구와 남성인구 비율을 알아보자
    3) 여성비율이 가장 높은 상위 5개 구는 어디일까?
    4) 남성비율이 가장 높은 상위 5개 구는 어디일까?

5,6. 인구현황 데이터 읽어와서 정리하기

5. 인구현황 데이터 읽어오기

- 1. CCTV현황 읽어오기와 동일

행정구역_시군구_별__성별_인구수_서울.xlsx

pd.read_excel(경로명)

- Colab에서는 파일의 경로명을 해당 파일 옆에 점 세개 버튼의 경로 복사

6. 인구현황 정리하기

- 3. CCTV현황 데이터 정리하기 과정과 동일

7. 데이터 자세히 보기(인구 현황)

- 인구수가 가장 높은/낮은 상위 5개 구는 어디일까?
  - 변수명.sort_values(by='정렬기준 열이름', ascending=True)

    # ascending = True : 오름차순, False : 내림차순
- 여성인구와 남성인구 비율을 알아보자.
  - 변수명['새로운열이름'] = 변수명['열이름'] / 변수명['열이름'] * 100

- 여성인구 비율이 가장 높은 상위 5개 구는 어디일까?

- 남성인구 비율이 가장 높은 상위 5개 구는 어디일까?

데이터 합치기

1. 인구 대비 CCTV 현황을 확인해 보자

2. 인덱스 이름을 구이름으로 바꿔보자

3. 서울시 자치구별 CCTV 현황 분석하기

    1) CCTV 수에 가장 영향을 미치는 항목은 무엇일까?

    2) CCTV 수에 가장 영향을 미치는 항목의 수가 가장 큰 구는 CCTV 수도 가장 큰가?

8. 데이터 합치기

인구 대비 CCTV 현황을 확인해 보자.

- 변수명= pd.merge(변수명, 변수명, on='구별')

  # on = '병합기준 열이름'

- 서울시 CCTV 설치운영 현황(자치구)-연도별.csv
- 행정구역_시군구_별__성별_인구수_서울.xlsx

8. 데이터 합치기

- 인덱스 이름을 구이름으로 바꿔보자
  - 변수명.set_index('인덱스로 사용할 열이름', inplace= True)

    # inplace = True 옵션은 원본데이터를 변경함
  - 변수명.reset_index(drop=True, inplace= True)

    #drop=True 기존에 사용하던 인덱스를 버리고 새로 0부터 새로 리셋

8. 데이터 합치기

- 합친 데이터 엑셀파일로 저장하기
  - 변수명 = pd.ExcelWriter('./CCTV_pop.xlsx', engine='xlsxwriter')
  - data_result.to_excel(excel_writer, index=True)
  - excel_writer.save()
  - # inplace = True 옵션은 원본데이터를 변경함
  - 변수명.reset_index(drop=True, inplace= True)

    #drop=True 기존에 사용하던 인덱스를 버리고 새로 0부터 새로 리셋

9. 자치구별 CCTV현황 분석하기

CCTV 수에 가장 영향을 미치는 항목은 무엇일까?

- import numpy as np #복잡한 수학연산을 지원하는 라이브러리
  - np.corrcoef(변수명['열이름'], 변수명['열이름'])

    # correlation-coefficient 계산 함수

    # 대각선을 중심으로 좌우 대칭

    # 상관계수의 절대값이 0.3이하이면 약한 상관관계, 0.5에 가까우면 보통, 1에 가까울 수록 높은 상관관계

    # 음수이면 양의 상관관계, 양수이면 양의 상관관계

CCTV 수에 가장 영향을 미치는 항목의 수가 가장 큰 구는 CCTV 수도 가장 큰가?

서울시 자치구별 CCTV 현황 시각화

1. CCTV수가 가장 많은 구는 어디인가?
2. CCTV의 비율(인구수에 대비 CCTV 수)이 가장 높은 구는 어디인가?
3. 인구 수에 대한 CCTV수의 일반적인 분포를 살펴보자
4. 일반적인 기준에서 인구 수에 비해 CCTV 수가 많은 구와 적은 구는 어디인가

10. 시각화 하기

CCTV수가 가장 많은 구는 어디인가?

CCTV 비율 (인구수에 대비 CCTV수)이 가장 높은 구는 어디인가?

인구 수에 대한 CCTV수의 일반적인 분포를 살펴보자

10. 시각화 하기

일반적인 기준에서 인구 수에 비해 CCTV 수가 많은 구와 적은 구는 어디인가?

- 데이터들을 대표하는 대표선 긋기
  - fp1 = np.ployfit(변수명['열이름'] , 변수명['열이름'], 1) # 대표선(fp1)의 기울기, 절편을 저장
  - f1 = np.ploy1d(fp1) # 직선의 y값 생성
  - fx = np.linspace(100000, 700000, 100) # 지정한 범위 내에서 데이터의 개수만큼 같은 간격의 숫자 생성, x축 숫자 생성
  - plt.plot(fx, f1(fx), '--', lw=2, color=g) # 선 그래프 그리기

    # 순서대로, x축 데이터, y축 데이터, ls는 선 스타일, lw는 선 굵기, color는 선색(그린)