인천광역시 소득분포 조사

데이터 전처리 및 분석_김민수

목차

01

개요

개요

- **프로젝트명 _ 인천광역시 소득분포 조사**
- **수행기간 _ 2023. 05. 16 ~ 2023. 05. 26**
- **기여도 _ 100%**
- **시스템환경**
  - *Windows 10 Pro*
  - *파이썬3.8.10*
  - *아나콘다23.3.1*
  - *Jupyter notebook*

02

분석배경 및 필요성

분석배경 및 필요성

소비 트렌드의 변화에 의문

호캉스, 고급외제차, 명품 등 고소득자들의 전유물로 여겨졌던 사치품이 일상이 되고
소비 트렌드가 고급화되는 상황에서 가계소득이 뒷받침이 되는지에 대한 의문

가계 경제 상황의 이해

특정 지자체의 객관적인 소득데이터를 통해 소득분포를 이해하고 가계의 체감 경기를
간접적으로 바라보며 가계 경제상황과 소비 트렌드를 이해

균형발전에 활용

소득 격차에 따라 체감하는 경기가 다르고 소비 행태가 다름을 이해함.
해당 데이터를 재정/금융정책 제안 등에 활용하여 경제의 균형 발전에 활용

03

데이터 전처리

데이터 전처리

1) 컬럼명 변경

```
## 컬럼명 변경 : 분위구분-> 소득분위, 평균처분가능소득->평균가처분소득
income.rename(columns={'분위구분': '소득분위', '평균처분가능소득': '평균가처분소득'},
inplace=True)
income[:2]
```

- 일부 컬럼을 이해하기 편한 컬럼으로 이름 변경
- income변수에 저장한 소득정보 dataframe의 칼럼들을 rename
- ‘분위구분’은 ‘소득분위’로, ‘평균처분가능소득’은 ‘평균가처분소득’으로 변경
- 원본에 반영하기 위해 inplace=True코드 추가

데이터 전처리

2) 열 삭제

```
## 칼럼 삭제 : 기준년월, 행정동코드, 시도명, 중위소득, 중위처분가능소득, 소득분위배율, 소득점유율, 소득경계값, 소득미산출인구수,
소득미산출인구포함평균소득
income.drop(columns=['기준년월', '행정동코드', '시도명', '중위소득', '중위처분가능소득', '소득분위배율',
                     '소득점유율', '소득경계값', '소득미산출인구수', '소득미산출인구포함평균소득'],
            inplace=True)
income[:2]
```

- 데이터 분석에 사용하지 않는 컬럼 삭제
- 소득분석에 영향을 미치지 않는 칼럼을 drop해서 삭제
- 원본에 반영하기 위해 inplace=True코드 추가

데이터 전처리

3) 행 삭제

```
## 행 삭제 : 성별, 연령별, 직업구분, 분위구분이 '0'(소계)인 행
zero_gender = income[income['성별'] == 0 ].index
income.drop(zero_gender, inplace=True)
```

```
## 시군구명이 '전체'인 행
all_district = income[income['시군구명'] == '전체' ].index
income.drop(all_district, inplace=True)
```

- 중복산입을 방지하기 위해 '소계'값을 마스킹으로 추출 후 삭제(drop)함
- 원본에 반영하기 위해 inplace=True코드 추가

데이터 전처리

4) 결측치 삭제

```
## 결측치 여부 확인 : 평균소득, 평균가처분소득
income.isna().sum()
```

```
## 결측치 삭제 : 평균소득, 평균가처분소득
income.dropna(how='any', inplace=True)
income[:7]
```

- 결측치 확인 후 데이터 분석에 불필요하다고 판단되는 행 삭제
- 결측치(True : 1, False : 0)의 합계(sum)를 통해 결측치 확인
- 결측치가 하나라도 존재하는 경우 평균값이 변하므로 해당 행 전체를 삭제하기 위해 how='any'코드를 추가
- 원본에 반영하기 위해 inplace=True코드 추가

데이터 전처리

5) 파일 저장

```
income.to_csv('income.csv', index=False, encoding='utf-8')
print('저장완료')
```

- 전처리한 파일은 csv파일로 저장
- 인덱스 삭제(index=False), 저장 시 인코딩은 utf-8

04

데이터 분석

데이터 분석-가설

[가설1] 평균소득은 연령이 높아짐에 따라 지속적으로 상승하다가 50대에 정점을 찍고 하락 반전할 것이다.

```
#연령별 평균소득의 평균
g_age_income = income.groupby('연령별').평균소득.mean()
g_age_income.round(2)
```

```
#연령별 평균소득 시각화
plt.figure(figsize=(10,5), facecolor='#f5f5f5')
plt.rc('font', family = 'D2coding')

g_age_income.plot(kind='line'
                  , rot=0,
                  color='Gold',
                 )
plt.xlabel('연령', labelpad=15, size=12, loc='right')
plt.ylabel('평균급여(만원)', labelpad=10, size=12, loc='top')
plt.title('[연령별 평균소득]', size=15)
plt.grid(axis='y', alpha=0.5)

y_ticks = plt.gca().get_yticks()
plt.gca().set_yticks(y_ticks)
plt.gca().set_yticklabels([f'{x:,.0f}' for x in y_ticks])

plt.savefig('연령별평균소득.png', dpi=120)
```

- '연령별'로 그룹짓고 '평균소득'의 평균(mean)을 산출
- matplotlib를 사용한 시각화①
- figure 크기&색상 지정, 폰트설정
- x축, y축 라벨, 타이틀, 격자 지정
- y축 틱 설정
  - 리스트내포 - f문자열 천자리 표시
- 이미지파일 저장

데이터 분석-가설

[가설1] 평균소득은 연령이 높아짐에 따라 지속적으로 상승하다가 50대에 정점을 찍고 하락 반전할 것이다.

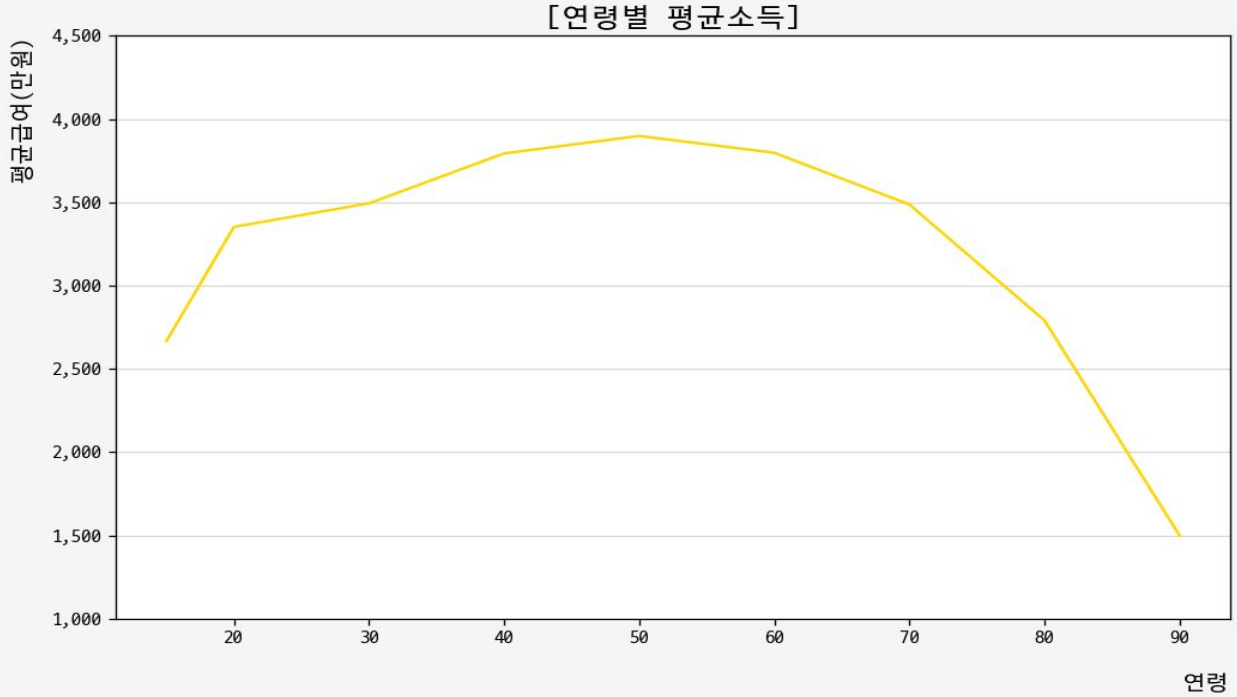
```
#성별을 구분한 연령별 평균소득
plt.figure(figsize=(10,5), facecolor='#f5f5f5')
plt.rc('font', family = 'D2coding')

sns.lineplot(data=income,
             x='연령별',
             y='평균소득',
             hue='성별',
             palette='pastel',
             errorbar=None
            )

plt.title('[연령별 및 성별 평균소득]', size=15)
plt.xlabel('연령', labelpad=15, size=12, loc='right')
plt.ylabel('평균급여(만원)', labelpad=10, size=12, loc='top')
plt.grid(axis='y', alpha=0.5)

y_ticks = plt.gca().get_yticks()
plt.gca().set_yticks(y_ticks)
plt.gca().set_yticklabels([f'{x:,.0f}' for x in y_ticks])

plt.legend(['남성','여성'], fontsize=12)

plt.savefig('연령별 및 성별 평균소득.png', dpi=120)
```

- 성별에 따른 연령별 평균소득
- seaborn을 사용한 시각화②
- 범례 재설정(1→남성, 2→여성)

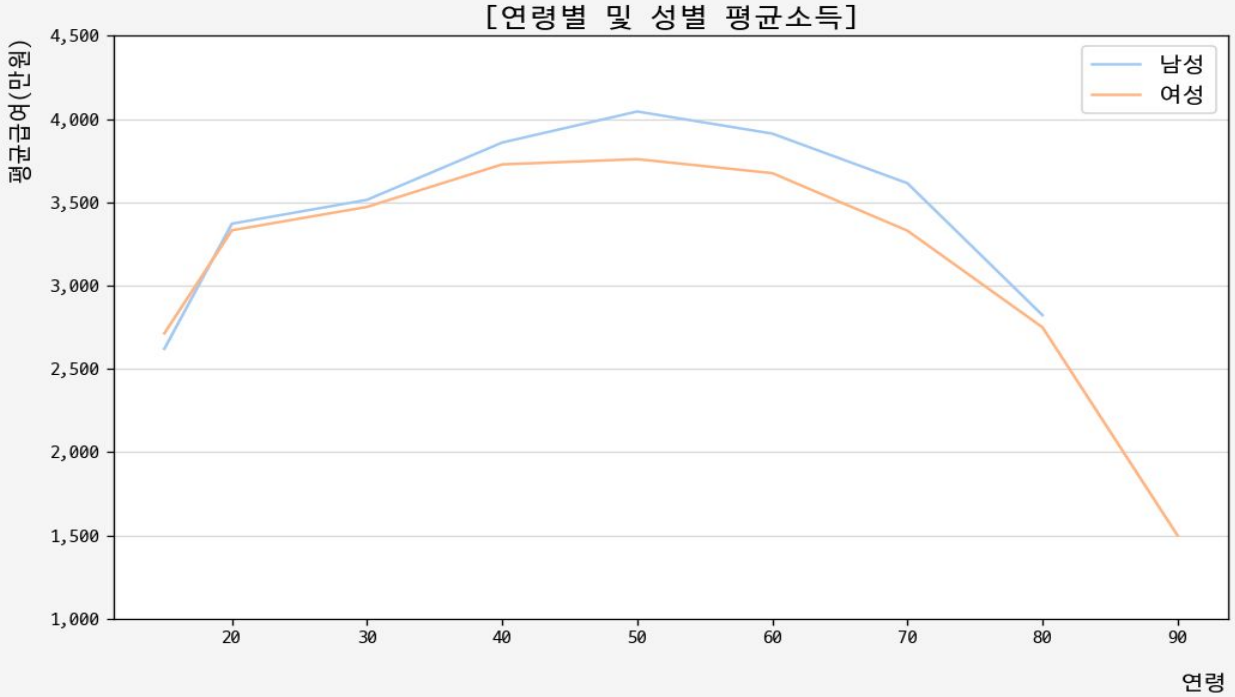
데이터 분석-결과

- 10대에서 20대로 넘어가는 구간의 평균소득은 **약 25%p 증가**하여 모든 구간 중에서 가장 크게 증가한다.
- 20대 이후 평균소득은 서서히 증가하여 **50대에 약3,899만원**으로 정점을 찍는다.
- 50대 이후부터 상승이 하락으로 반전되고 **급격한 하락**을 보이며 90세 이후 약 1,500만원으로 최저점을 찍는다.
- 성별이 연령대별 평균소득의 추세에 미치는 영향은 **미미**하다.

데이터 분석-결과


[연령별 평균소득]
평균금액(만원)
4,500
4,000
3,500
3,000
2,500
2,000
1,500
1,000
20
30
40
50
60
70
80
90
연령

데이터 분석-결과


[연령별 및 성별 평균소득]
평균소득(만원)
4,500
4,000
3,500
3,000
2,500
2,000
1,500
1,000
20
30
40
50
60
70
80
90
연령
남성
여성

데이터 분석-가설

[가설2] 직업 별 평균급여는 상장사 근로소득자가 가장 높을 것이다.

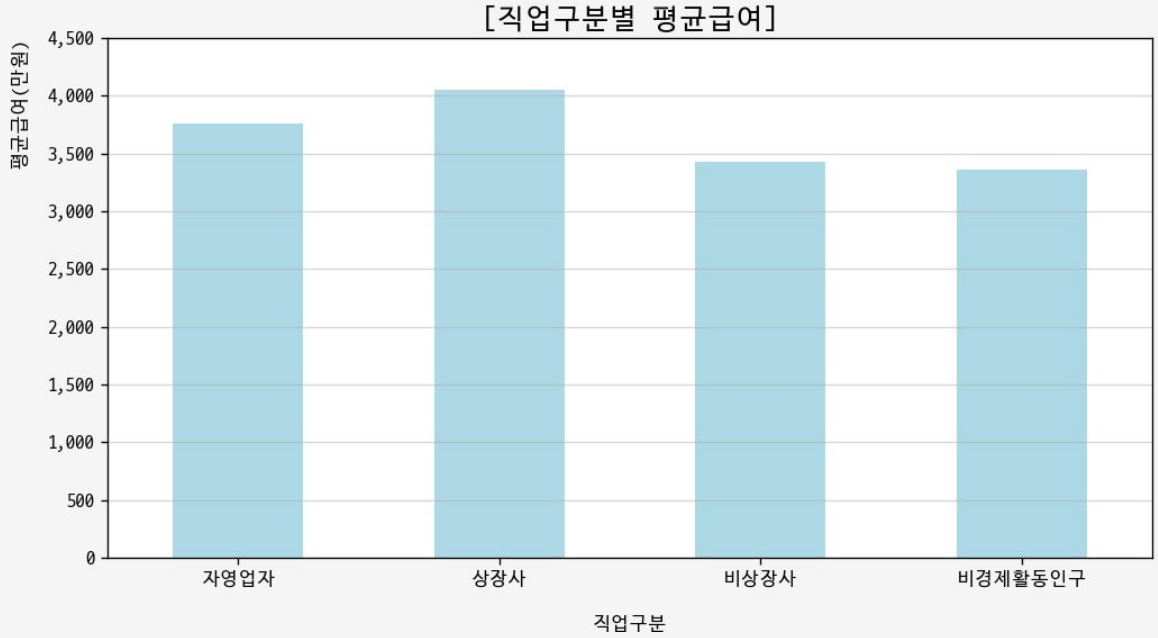
```
## 직업별 평균소득
for idx in range(1,5):
    g_job_income = income.groupby(['직업구분']).get_group(idx).평균소득.mean().round(2)
    print(g_job_income)


plt.figure(figsize=(12,8), facecolor='#F5F5F5')
income.groupby('직업구분').평균소득.mean().plot(kind='bar',
                                        rot=0,
                                        color='LightBlue',
                                        figsize=(10,5))

plt.xlabel('직업구분', labelpad=15)
plt.ylabel('평균급여(만원)', labelpad=10, loc='top')
plt.title('[직업구분별 평균급여]', size=15)

plt.xticks(np.arange(0,4,1), ('자영업자', '상장사', '비상장사', '비경제활동인구'))
plt.grid(axis='y', alpha=0.5)

y_ticks = plt.gca().get_yticks()
plt.gca().set_yticks(y_ticks)
plt.gca().set_yticklabels([f'{x:,.0f}' for x in y_ticks])

plt.savefig('직업구분별 평균급여.png', dpi=120)
```

- 직업구분별로 그룹화 후 직업별(1~4) 평균소득의 평균 계산
- x축 틱 재설정(직업별)

데이터 분석-결과

- 직업별 평균소득은 자영업자 **약 3,764만원**, 상장사 근로소득자 **약 4,051만원**, 비상장사 근로소득자 **약 3,429만원,** 비경제활동인구 **약 3,364만원**으로 나타났다.
- 평균소득의 순위는 상장사 근로소득자, 자영업자, 비상장사 근로소득자, 비경제활동인구 순이다.
- 상장사 근로소득자의 평균급여와 비경제활동인구의 평균급여 차이금액은 약 687만원으로 상장사 근로소득자가 약 20%p정도 더 높다.
- 근로소득자의 경우 **상장사 근로소득자**의 평균소득이 비상장사 근로소득자의 평균소득보다 약 622만원(약18%p) **더 높다.**

데이터 분석-결과


[직업구분별 평균급여]
평균급여(만원)
4,500
4,000
3,500
3,000
2,500
2,000
1,500
1,000
500
0
자영업자
상장사
비상장사
비경제활동인구
직업구분

데이터 분석-가설

[가설3] 성별 평균소득은 여성이 남성보다 낮을 것이다.

```
#성별 평균소득
income[['성별', '평균소득']].groupby('성별').평균소득.mean()
```

```
plt.figure(figsize=(12,8), facecolor='#F5F5F5')

fig = sns.barplot(data=income,
            x='성별',
            y='평균소득',
            palette='pastel',
            errorbar=None
           ).set_title("[성별 평균소득]")
fig.figure.set_size_inches(6, 4)
plt.ylabel('평균급여(만원)', labelpad=10)
plt.xticks(np.arange(0,2,1), ('남성', '여성'))
plt.grid(axis='y', alpha=0.5)

y_ticks = plt.gca().get_yticks()
plt.gca().set_yticks(y_ticks)
plt.gca().set_yticklabels([f'{x:,.0f}' for x in y_ticks])

plt.savefig('성별 평균소득.png', dpi=120)
```

- 성별과 평균소득을 추출 후 성별로 그룹화하여 평균소득의 평균 산출
- seaborn을 이용해 시각화
- bar 그래프, 색상(palette)지정

데이터 분석-결과

- **여성**의 평균소득은 **약 3,500만원**이고, **남성**의 평균소득은 **약 3,637만원**이다.
- 남여간의 평균소득 격차는 약 137만원으로 **남성**이 평균소득이 여성의 평균소득보다 근소하게 **높다.**

데이터 분석-결과


[성별 평균소득]
4,000
3,500
3,000
2,500
2,000
1,500
1,000
500
0
평균금액(만원)
남성
여성
성별

데이터 분석-가설

[가설4] 소득분위 4~7분위에 가장 많은 소득산출인구가 분포할 것이다.

```
## 소득분위별 소득산출인구수의 합
g_rank = income[['소득분위','소득산출인구수']].groupby('소득분
위').sum().sort_values(by='소득산출인구수', ascending=False)
g_rank
```

```
g_rank.plot(kind='pie',
            subplots=True,
            colormap='rainbow',
            autopct='%.f%%',
            title='[소득분위별 소득산출인구 비율]',
            explode=[0, 0, 0, 0, 0, 0, 0, 0, 0, 0.1],
            shadow=True
            )
plt.legend().remove()
plt.gca().axes.yaxis.set_visible(False)
plt.savefig('소득분위별 소득산출인구수 비율.png', dpi=120)
```

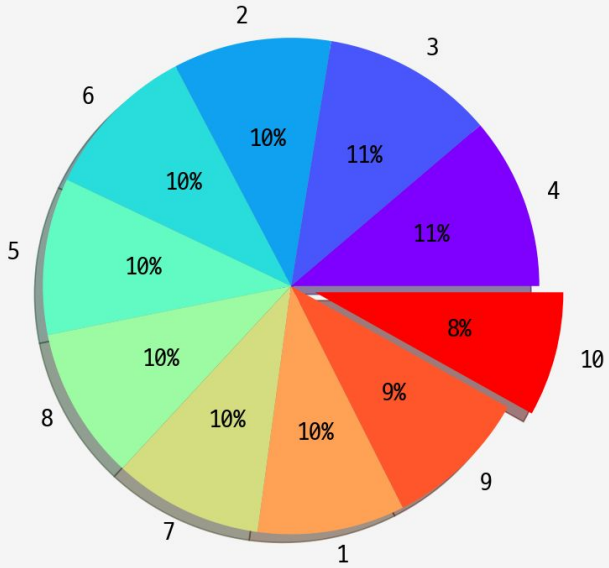
- 소득분위와 소득산출인구수를 추출한 후 소득분위별로 그룹화하고 소득산출인구수를 합(sum)한 후 소득산출인구수 기준 내림차순 정렬 (sort_values(ascending=False))
- pie그래프, 색 지정(colormap)
- 퍼센트 표시(autopct)
- 조각 분리하여 강조(explode)
- 레이블 제거, 축 표시 제거

데이터 분석-결과

- 소득산출인구는 각 소득분위마다 약 10% 내외의 분포율을 나타냈다.
- 소득분위 **4분위**의 인구가 **248,314명(11%)**으로 가장 많았고, 소득분위 **10분위**의 인구가 **179,797명(8%)**으로 가장 적었다.
- 특정 분위에 소득산출인구가 밀집돼있지 않고 **고르게** 분포돼있다.

데이터 분석-결과

[소득분위별 소득산출인구 비율]



데이터 분석-가설

[가설5] 연령이 30세 미만인 비경제활동인구 중에서 소득분위가 10분위인 사람들은 연수구에 가장 많을 것이다.

```
## 30대 미만 비경제활동인구수 : 331,076
u30_inactive = income[(income.연령별 < 30) & (income.직업구분 == 4)]
u30_inactive['소득산출인구수'].sum()
```

```
331076
```

```
## 30대 미만 비경제활동인구 중 소득분위가 10위인 사람들 : 3,620
g_u30_inactive_top10 = u30_inactive.groupby('소득분위').get_group(10).sort_values(by='시군구명')
g_u30_inactive_top10['소득산출인구수'].sum()
```

```
3620
```

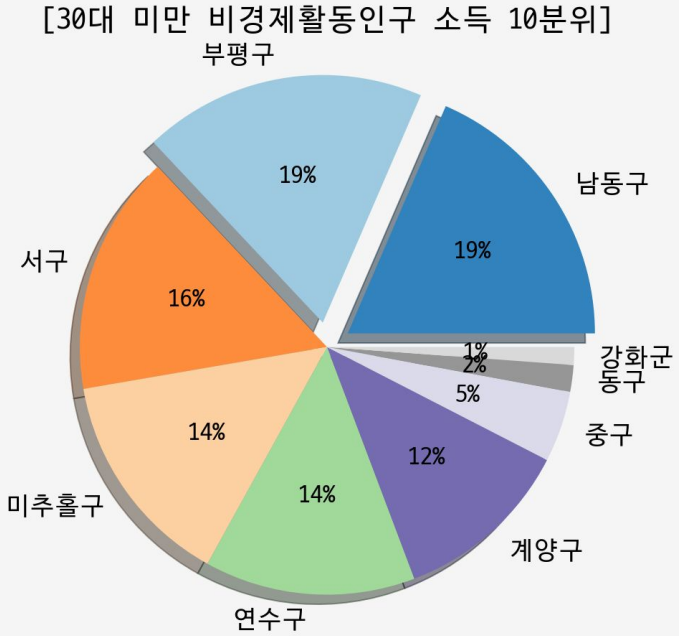
```
plt.rc('font', size = 18)

young_and_rich = income[(income.연령별 < 30) & (income.직업구분 == 4) & (income.소득분위 == 10)]

gold_spoon = young_and_rich[['시군구명', '소득산출인구수']].groupby('시군구명').소득산출인구
수.sum().sort_values(ascending=False)

gold_spoon.plot(kind='pie',
                y='소득산출인구수',
                x='시군구명',
                colormap='tab20c',
                autopct='%.f%%',
                title='[30대 미만 비경제활동인구 소득 10분위]',
                explode=[0.1, 0.1, 0, 0, 0, 0, 0, 0, 0, 0],
                shadow=True
                )
plt.gca().axes.yaxis.set_visible(False)

plt.savefig('30대 미만 비경제활동인구 소득 10분위.png', dpi=140)
```

- 연령이 30미만이고 직업구분이 4인 값을 추출한 후 소득분위별 그룹화
- 소득분위 10분위 그룹을 가져와서 시군구명 순으로 오름차순 정렬 후 소득산출인구수의 합 산출

데이터 분석-결과

- **30세 미만**의 **비경제활동인구** 331,076명 중에서 소득분위가 **10분위**인 사람은 **3,620명**이 존재한다.
- **남동구(19%)**와 **부평구(19%)**가 공동 1위를 기록했다.
- 서구(16%)는 3위, 미추홀구(14%)와 **연수구(14%)**가 공동 5위, 계양구(12%)가 6위를 차지했다.
- 중구(5%), 동구(2%), 강화군(1%), 옹진군(0%)은 모두 10%미만으로 그 뒤를 이었다.

데이터 분석-결과


[30대 미만 비경제활동인구 소득 10분위]
부평구
19%
남동구
19%
서구
16%
강화군
1%
동구
2%
중구
5%
계양구
12%
연수구
14%
미추홀구
14%

05

한계점

한계점

- 직업구분의 직업군이 자영업자와 근로소득자로만 분류돼있음. 소득세법에 따른 소득분류나 기타 기준으로 세분화한다면 더 구체적인 자료가 될 것으로 기대
- 특정 지자체만을 대상으로 분석한 자료이므로 해당 결과를 다른 지자체로 확대하여 적용하기 어려울 수 있음. 가령 대기업이나 산업단지가 집중된 지역과 그렇지 않은 지역은 소득분포가 다를 수 있음
- 소득은 추청된 자료이므로 불명확할 수 있음. 가령 탈세나 명의대여 등으로 소득을 빼돌리거나 제3자의 소득이 본인에게 잡히는 경우 실질적인 소득이 다를 수 있음

감사합니다!