

데이터 수집 계획 및 보고서

[데이터 수집] 프로젝트 계획

프로젝트명	인천광역시 소득 분포 조사
능력단위분류	데이터 전처리 및 분석
수행 기간	23.05.16~23.05.26

■ 상세 정보

구분	설명 및 의견
데이터 수집 목적	인천광역시 경제활동인구의 소득데이터 수집을 통해 인천시 가계경제상황을 객관적으로 이해하고자 함. 향후 해당 데이터를 인천광역시의 재정/금융정책 제언 등에 활용
데이터 수집 계획	- url : https://www.data.go.kr/data/15076577/fileData.do - 인천광역시_소득 데이터 - 인천데이터포털 사이트의 공공데이터 검색 후 인천시민 소득 분포 데이터 수집. 인천광역시와 NICE신용평가에서 만든 소득 데이터(CSV파일) 활용
데이터 수집 시스템 환경	- Windows 10 Pro - 파이썬3.8.10 - 아나콘다23.3.1
데이터 수집 후 저장 방법	- Pandas - DataFrame - Seaborn, Matplotlib - CSV파일 및 JPG파일로 산출물 저장
데이터 수집 후 정제 계획	소득 데이터를 확인하여 전처리 진행. 성별, 연령, 직업구분 등의 칼럼에서 중복되는 데이터 삭제. 일부 컬럼(분위구분, 평균처분가능소득) 명칭 변경. 소득관련 결측치 열 삭제. 가설에 대한 분석 결과는 그래프 등으로 시각화

[데이터 수집] 프로젝트 보고서

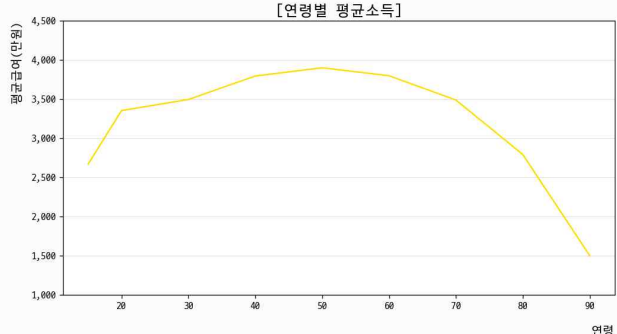
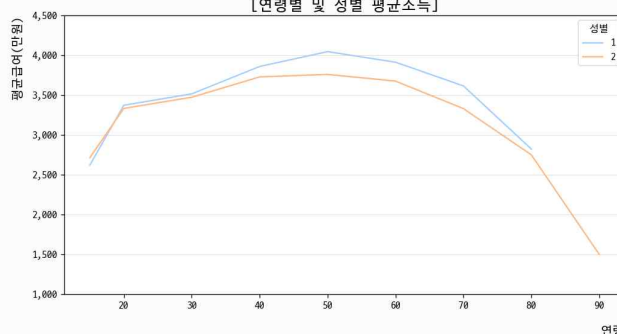
■ 데이터 전처리 작업

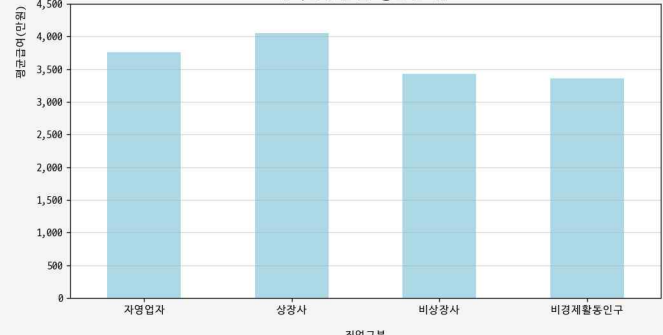
구분	설명
전처리	<p>‘인천광역시_소득 데이터’에 대한 전처리 작업 수행</p> <p>1) 컬럼명 변경 : 일부 컬럼을 이해하기 편한 컬럼으로 이름 변경</p> <ul style="list-style-type: none"> - income변수에 저장한 소득정보 dataframe의 칼럼들을 rename - ‘분위구분’은 ‘소득분위’로, ‘평균처분가능소득’은 ‘평균가처분소득’으로 변경 - 원본에 반영하기 위해 inplace=True코드 추가 <pre>## 컬럼명 변경 : 분위구분-> 소득분위, 평균처분가능소득->평균가처분소득 income.rename(columns={'분위구분': '소득분위', '평균처분가능소득': '평균가처분소득'}, inplace=True) income[:2]</pre> <p>2) 컬럼 삭제 : 데이터 분석에 사용하지 않는 컬럼 삭제</p> <ul style="list-style-type: none"> - 소득분석에 영향을 미치지 않는 칼럼을 drop해서 삭제 - 원본에 반영하기 위해 inplace=True코드 추가 <pre>## 컬럼 삭제 : 기존년월, 행정동코드, 시도명, 중위소득, 중위처분가능소득, 소득분위배율, 소득경계값, 소득미산출인구수, 소득미산출인구포함 평균소득 income.drop(columns=['기존년월', '행정동코드', '시도명', '중위소득', '중위처분가능소득', '소득분위배율', '소득경계값', '소득미산출인구수', '소득미산출인구포함평균소득'], inplace=True)</pre> <p>3) 행 삭제 : 중복 산입 방지</p> <ul style="list-style-type: none"> - 행의 중복산출을 방지하기 위해 ‘소계’값을 마스킹을 통해 추출 후 삭제(drop)함 - 원본에 반영하기 위해 inplace=True코드 추가 <pre>## 행 삭제 : 성별, 연령별, 직업구분, 분위구분이 '0'(소계)인 행, 시군구명이 '전체'인 행 zero_gender = income[income['성별'] == 0].index income.drop(zero_gender, inplace=True) zero_age = income[income['연령별'] == 0].index income.drop(zero_age, inplace=True) zero_job = income[income['직업구분'] == 0].index income.drop(zero_job, inplace=True) zero_rank = income[income['소득분위'] == 0].index income.drop(zero_rank, inplace=True) all_district = income[income['시군구명'] == '전체'].index income.drop(all_district, inplace=True)</pre>


구분	설명
	<p>4) 결측치 삭제 : 결측치 확인 후 데이터 분석에 불필요하다고 판단되는 행 삭제</p> <ul style="list-style-type: none"> - 결측치 확인. True : 1, False : 0이므로 합계를 통해 결측치 확인 <pre>## 결측치 여부 확인 : 평균소득, 평균가처분소득 income.isna().sum()</pre> <ul style="list-style-type: none"> - 결측치가 하나라도 존재하는 경우 소득의 평균 등이 변하기 때문에 해당 행 전체를 삭제하기 위해 how='any'코드를 추가 - 원본에 반영하기 위해 inplace=True코드 추가 <pre>## 결측치 삭제 : 평균소득, 평균가처분소득 income.dropna(how='any', inplace=True) income[:7]</pre> <p>5) 저장 : 전처리한 파일은 csv파일로 저장</p> <ul style="list-style-type: none"> - 인덱스 삭제, 저장 시 인코딩은 utf-8 <pre>income.to_csv('income.csv', index=False, encoding='utf-8') print('저장완료')</pre>
기타 설명 및 의견	<ul style="list-style-type: none"> - 자료설명 <p>: 해당 데이터는 인천시와 NICE 신용평가사에서 기업매칭 사업을 통해 제작한 자료. 2020년 6월 말 기준 직전 1개년도의 인천시 경제활동인구에 대한 신용카드 및 체크카드 소비를 분석하여 소득데이터를 추정한 자료(행정구별 소득데이터로 성별/10세단위 연령별/직업별/소득분위별)</p> <ul style="list-style-type: none"> - 느낀점 <p>1) 데이터를 파악하는 과정에서 구체적이지 못한 데이터들이 있음. 가령 직업구분의 직업군은 자영업자와 근로소득자만으로 분류돼있어서 자세한 분류를 알 수 없음</p> <p>2) 분석에 필요한 데이터를 찾는 과정 자체가 쉽지 않음. 가공이 많이 된 자료는 데이터 분석에 필요한 자료를 제공해주지 못하고, 가공이 안 된 자료는 불필요한 부분을 삭제하면 데이터 수집이 어려운 상태가 됨</p>

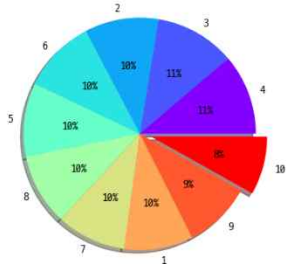
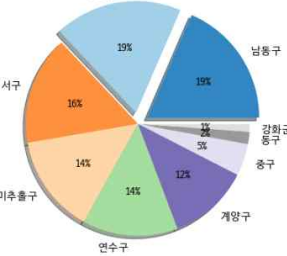
■ 데이터 분석 결과

구분	설명																																																						
	<p>5.2 [가설1] 평균소득은 연령이 높아짐에 따라 지속적으로 상승하다가 50대에 정점을 찍고 하락 반전할 것이다.</p> <div><pre>#연령별 평균소득의 평균 g_age_income = income.groupby('연령별').평균소득.mean() g_age_income.round(2)</pre><pre>#연령별 성별 평균소득의 평균 g_age_income = income.groupby(['연령별', '성별']).평균소득.mean() g_age_income.round(2)</pre></div> <div><table><tr><th>연령별</th><th>성별</th><th>평균소득</th></tr><tr><td>15</td><td>1</td><td>2621.01</td></tr><tr><td>20</td><td>2</td><td>2714.16</td></tr><tr><td>20</td><td>1</td><td>3372.11</td></tr><tr><td>20</td><td>2</td><td>3332.60</td></tr><tr><td>30</td><td>1</td><td>3515.31</td></tr><tr><td>30</td><td>2</td><td>3472.80</td></tr><tr><td>40</td><td>1</td><td>3959.20</td></tr><tr><td>40</td><td>2</td><td>3728.10</td></tr><tr><td>50</td><td>1</td><td>4045.73</td></tr><tr><td>50</td><td>2</td><td>3759.45</td></tr><tr><td>60</td><td>1</td><td>3912.68</td></tr><tr><td>60</td><td>2</td><td>3675.58</td></tr><tr><td>70</td><td>1</td><td>3615.21</td></tr><tr><td>70</td><td>2</td><td>3530.22</td></tr><tr><td>80</td><td>1</td><td>2822.84</td></tr><tr><td>80</td><td>2</td><td>2749.92</td></tr><tr><td>90</td><td>2</td><td>1500.00</td></tr></table></div> <div><p>Name: 평균소득, dtype: float64</p><p>Name: 평균소득, dtype: float64</p></div>	연령별	성별	평균소득	15	1	2621.01	20	2	2714.16	20	1	3372.11	20	2	3332.60	30	1	3515.31	30	2	3472.80	40	1	3959.20	40	2	3728.10	50	1	4045.73	50	2	3759.45	60	1	3912.68	60	2	3675.58	70	1	3615.21	70	2	3530.22	80	1	2822.84	80	2	2749.92	90	2	1500.00
연령별	성별	평균소득																																																					
15	1	2621.01																																																					
20	2	2714.16																																																					
20	1	3372.11																																																					
20	2	3332.60																																																					
30	1	3515.31																																																					
30	2	3472.80																																																					
40	1	3959.20																																																					
40	2	3728.10																																																					
50	1	4045.73																																																					
50	2	3759.45																																																					
60	1	3912.68																																																					
60	2	3675.58																																																					
70	1	3615.21																																																					
70	2	3530.22																																																					
80	1	2822.84																																																					
80	2	2749.92																																																					
90	2	1500.00																																																					
분석 결과																																																							

구분	설명
	<p>시각화 1</p> <pre>#연령별 평균소득 시각화 plt.figure(figsize=(10,5), facecolor='ivory') 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)</pre>  <p>상세한 구분화 연령별 평균소득</p> <pre>#상세한 구분화 연령별 평균소득 plt.figure(figsize=(10,5), facecolor='ivory') 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.savefig('연령별 및 성별 평균소득.png', dpi=120)</pre> 

구분	설명
	<p>[가설2] 직업별 평균급여는 상장사 근로소득자가 가장 높을 것이다.</p> <p>가설 분석</p> <pre>## 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)</pre> <pre>3764.48 4051.04 3428.88 3364.16</pre> <p>시각화</p> <pre>plt.figure(figsize=(12,8), facecolor='ivory') income.groupby('직업구분').평균소득.mean().plot(kind='bar', rot=0, color='Gold', 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)</pre>  <p>[가설3] 성별 평균소득은 여성이 남성보다 낮을 것이다.</p> <p>가설 분석</p> <pre>#성별 평균소득 income[['성별', '평균소득']].groupby('성별').평균소득.mean()</pre> <pre>성별 1 3637.771037 2 3500.065878 Name: 평균소득, dtype: float64</pre>

구분	설명																						
시각화	<pre> plt.figure(figsize=(12,8), facecolor='ivory') 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(['x:{}'.format(x) for x in y_ticks]) plt.savefig('성별 평균소득.png', dpi=120) </pre>  <p>[성별 평균소득]</p>																						
[가설4] 소득분위 4~7분위에 가장 많은 소득산출인구가 분포할 것이다.																							
가설 분석	<pre> # 소득산출인구수 : 2,214,047명 income.소득산출인구수.sum() 2214047 ## 소득분위별 소득산출인구수의 합 g_rank = income[['소득분위', '소득산출인구수']].groupby('소득분위').sum().sort_values(by='소득산출인구수', ascending=False) g_rank </pre> <table border="1"> <thead> <tr> <th>소득분위</th> <th>소득산출인구수</th> </tr> </thead> <tbody> <tr><td>4</td><td>248314</td></tr> <tr><td>3</td><td>248109</td></tr> <tr><td>2</td><td>228064</td></tr> <tr><td>6</td><td>226826</td></tr> <tr><td>5</td><td>226242</td></tr> <tr><td>8</td><td>220400</td></tr> <tr><td>7</td><td>215330</td></tr> <tr><td>1</td><td>212547</td></tr> <tr><td>9</td><td>209198</td></tr> <tr><td>10</td><td>179797</td></tr> </tbody> </table>	소득분위	소득산출인구수	4	248314	3	248109	2	228064	6	226826	5	226242	8	220400	7	215330	1	212547	9	209198	10	179797
소득분위	소득산출인구수																						
4	248314																						
3	248109																						
2	228064																						
6	226826																						
5	226242																						
8	220400																						
7	215330																						
1	212547																						
9	209198																						
10	179797																						

구분	설명
시각화	<pre> g_rank.plot(kind='pie', subplots=True, y='소득산출인구수', colormap='rainbow', autopct='%1.1%', 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) </pre> <p>[소득분위별 소득산출인구 비율]</p> 
[가설5] 연령이 30세 미만인 비경제활동인구 중에서 소득분위가 10분위인 사람들은 연수구에 가장 많을 것이다.	
가설 분석	<pre> ## 30대 미만 비경제활동인구수 : 331,076 u30_inactive = income[(income.연령별 < 30) & (income.직업구분 == 4)] u30_inactive['소득산출인구수'].sum() 331076 ## 30대 미만 비경제활동인구 중 소득분위가 10위의 사람들 : 3,820 g_u30_inactive_top10 = u30_inactive.groupby('소득분위').get_group(10).sort_values(by='시군구명') g_u30_inactive_top10['소득산출인구수'].sum() 3620 </pre>
시각화	<pre> g_u30_inactive_top10_su = g_u30_inactive_top10.groupby(['시군구명']).sum().sort_values(by='소득산출인구수', ascending=False) g_u30_inactive_top10_su.plot(kind='pie', y='소득산출인구수', colormap='tab20', autopct='%1.1%', shadow=True, explode=[0.1, 0.1, 0, 0, 0, 0, 0, 0, 0, 0], title='[30대 미만 비경제활동인구 소득분위 10분위]',) plt.legend().remove() plt.gca().axes.yaxis.set_visible(False) plt.savefig('30대 미만 비경제활동인구 소득분위 10분위.png', dpi=120) </pre> <p>[30대 미만 비경제활동인구 소득분위 10분위]</p> 

구분	설명
분석 결과 (의견)	<p>[가설1] 평균소득은 연령이 높아짐에 따라 지속적으로 상승하다가 50대에 정점을 찍고 하락 반전할 것이다.</p> <ul style="list-style-type: none"> - 10대에서 20대로 넘어가는 구간의 평균소득은 약 25%p 증가하여 모든 구간 중에서 가장 크게 증가한다. - 20대 이후 평균소득은 서서히 증가하여 50대에 약3,899만원으로 정점을 찍는다. - 50대 이후부터 상승이 하락으로 반전되고 급격한 하락을 보이며 90세 이후 약 1,500만원으로 최저점을 찍는다. - 성별이 연령대별 평균소득의 추세에 미치는 영향은 미미하다.
	<p>[가설2] 직업 별 평균급여는 상장사 근로소득자가 가장 높을 것이다.</p> <ul style="list-style-type: none"> - 직업별 평균소득은 자영업자 약 3,764만원, 상장사 근로소득자 약 4,051만원, 비상장사 근로소득자 약 3,429만원, 비경제활동인구 약 3,364만원으로 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 근로소득자의 평균소득이 비상장사 근로소득자의 평균소득보다 약 622만원(약18%p) 더 높다.
	<p>[가설3] 성별 평균소득은 여성이 남성보다 낮을 것이다.</p> <ul style="list-style-type: none"> - 여성의 평균소득은 약 3,500만원이고, 남성의 평균소득은 약 3,637만원이다. - 남녀간의 평균소득 격차는 약 137만원으로 남성이 평균소득이 여성의 평균소득보다 근소하게 높다.
	<p>[가설4] 소득분위 4~7분위에 가장 많은 소득산출인구가 분포할 것이다.</p> <ul style="list-style-type: none"> - 소득산출인구는 각 소득분위마다 약 10% 내외의 분포율을 나타냈다. - 소득분위 4분위의 인구가 248,314명(11%)으로 가장 많았고, 소득분위 10분위의 인구가 179,797명(8%)으로 가장 적었다. - 특정 분위에 소득산출인구가 밀집돼있지 않고 고르게 분포돼있다.
	<p>[가설5] 연령이 30세 미만인 비경제활동인구 중에서 소득분위가 10분위인 사람들은 연수구에 가장 많을 것이다.</p> <ul style="list-style-type: none"> - 30세 미만의 비경제활동인구 331,076명 중에서 소득분위가 10분위인 사람은 3,620명이 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.