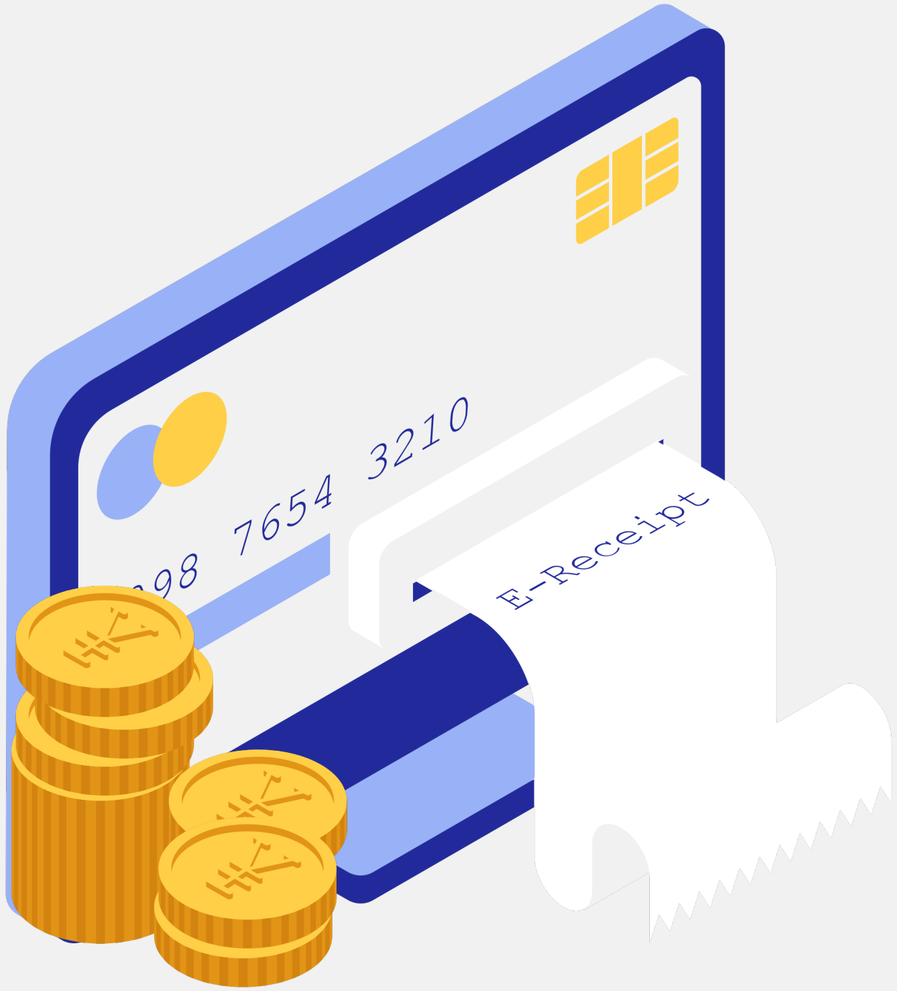


다변량 통계분석

가계금융복지조사 자료를 활용한 경상소득 설명 및 회귀모형 구축

15조
20182838 김상준
20180814 김현석
20190472 진준용

목차

프로젝트 배경과 주제

· 주제 선정 이유: 가계금융복지조사 데이터를 통해 경상 소득과 여러 설명변수 간의 관계를 분석하여 가계금융복지에 대한 이해를 높이고, 최근 일어나는 다양한 경제 현상 등을 쉽게 이해하기 위함

· 연구 목표: 경상 소득과 여러 설명변수(예: 성별, 교육수준, 직업 등) 간의 상관관계를 분석하여 어떤 변수가 경상소득과 관련이 있는지 확인한다. 회귀분석을 통해 가계 소득에 영향을 주는 주요 요인을 식별한다.

· 위의 연구 목표를 토대로 가계금융복지조사 데이터를 활용한 회귀분석 모델을 구축하여 가계금융복지에 대한 실질적인 인사이트를 도출하고자 한다.

데이터 소개와 변수 정의

- 본 데이터는 통계청 MDIS 서비스에 등록된 가계금융복지조사 데이터

변수명	형태	내용	변수값 설명

반응변수:

경상소득(보완)	숫자	소득	단위: 만원

설명변수:

자산_금융자산_저축금액	숫자	저축	단위: 만원
자산_실물자산_부동산금액	숫자	부동산	단위: 만원
부채_금융부채_신용대출금액	숫자	신용대출	단위: 만원
부채_금융부채_담보대출금액	숫자	담보대출	단위: 만원
지출_소비지출_식료품(외식비포함)	숫자	식료지출	단위: 만원
지출_소비지출_주거비	숫자	주거지출	단위: 만원
지출_소비지출_교육비(보육료포함)	숫자	교육지출	단위: 만원
지출_비소비지출_세금(보완)	숫자	세금	소득세, 재산세, 자동차세, 기타세금 등
만연령	숫자	나이	
가구원수	숫자	가구원수	가구주와 주거 또는 소득과 지출 등 생계를 같이 하는 사람
가구주_성별코드	문자	성별	1: 남자; 2: 여자
가구주_교육정도_학력코드	문자	학력	1. 안 받음(미취학 포함) 2. 초등학교 3. 중학교 4. 고등학교 5. 대학(3년제 이하) 6. 대학교(4년제 이상) 7. 대학원 석사 8. 대학원박사 이상
가구주_직업대분류코드	문자	직업대분류	1. 관리자 2. 전문가 및 관련 종사자 3. 사무 종사자 4. 서비스 종사자 5. 판매 종사자 6. 농림어업 숙련종사자 7.기능원 및 관련 기능 종사자 8. 장치,기계조작 및 조립 종사자 9. 단순노무 종사자 A. 군인
수도권여부	문자	수도권여부	G1: 수도권; G2: 비수도권

데이터 탐색

```
data = pd.read_csv('2022.csv', encoding='CP949')
data.rename(columns={ '가구주_성별코드':'성별', '가구주_교육정도_학력코드':'학력', '가구주_직업대분류코드':'직업',
                      '가구주_만연령':'연령', '자산_금융자산_저축금액':'저축', '자산_실물자산_부동산금액':'부동산',
                      '부채_금융부채_담보대출금액':'담보대출', '부채_금융부채_신용대출금액':'신용대출', '경상소득(보완)':'소득',
                      '지출_소비지출_식료품(외식비포함)':'식료지출', '지출_소비지출_주거비':'주거지출',
                      '지출_소비지출_교육비(보육료포함)':'교육지출', '지출_비소비지출_세금(보완)':'세금'},inplace=True)
data['학력'] = data['학력'].astype('category')
data['직업'].fillna('기타',inplace=True)
```

```
int_columns = ['연령', '가구원수', '저축', '부동산', '담보대출', '신용대출', '소득', '식료지출', '주거지출', '교육지출', '세금']
scaler = StandardScaler()
data[int_columns] = scaler.fit_transform(data[int_columns])
```

```
data.head()
```

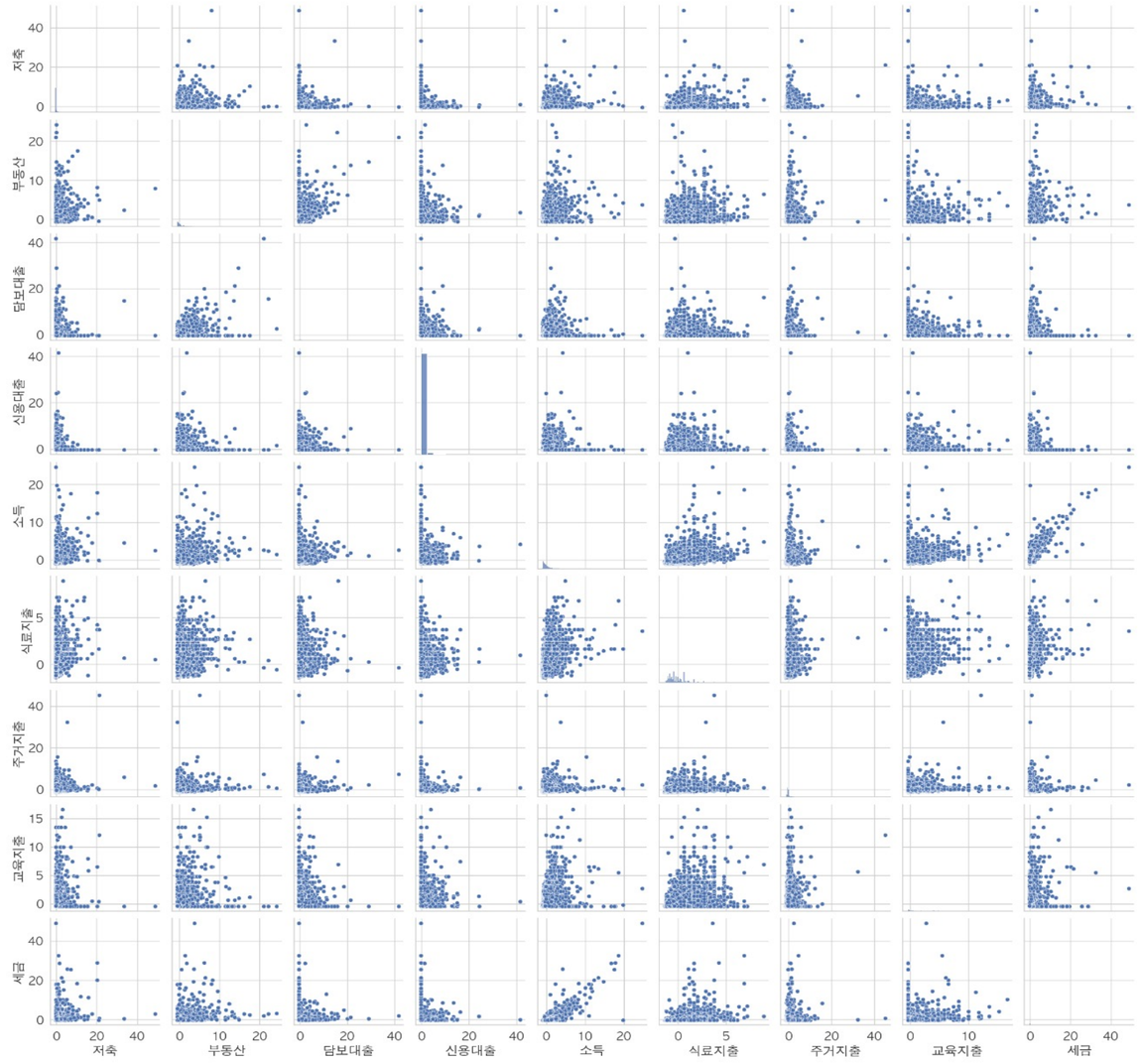
	수도권여부	성별	가구원수	학력	직업	연령	저축	부동산	담보대출	신용대출	소득	식료지출	주거지출	교육지출	세금
0	G1	1	-1.120841	1	기타	0.364231	-0.428647	-0.538143	-0.325902	-0.252578	-0.749248	-0.293596	-0.053373	-0.407146	-0.256492
1	G1	1	-1.120841	1	기타	1.021987	-0.302715	-0.261454	-0.325902	-0.252578	-0.670925	-0.655862	-0.481841	-0.407146	-0.239660
2	G1	1	-1.120841	1	기타	1.482415	-0.418201	-0.538143	-0.325902	-0.252578	-0.821298	-0.569608	-0.095517	-0.407146	-0.256492
3	G1	1	-1.120841	1	기타	1.548191	-0.384080	-0.538143	-0.325902	-0.252578	-0.825366	-1.225136	-0.587202	-0.407146	-0.255760
4	G1	1	-1.120841	1	기타	1.613966	0.122483	-0.022039	-0.325902	-0.252578	-0.786883	-0.811118	-0.622322	-0.407146	-0.225755

- 변수명 단순화
- 학력 변수 데이터 타입 변경
- 직업 변수 결측치 '기타'로 대체
- 연속형 변수의 표준화 진행

데이터 탐색

```
1 sns.pairplot(data_int)
2 plt.show
```

```
<function matplotlib.pyplot.show(close=None, block=None)>
```

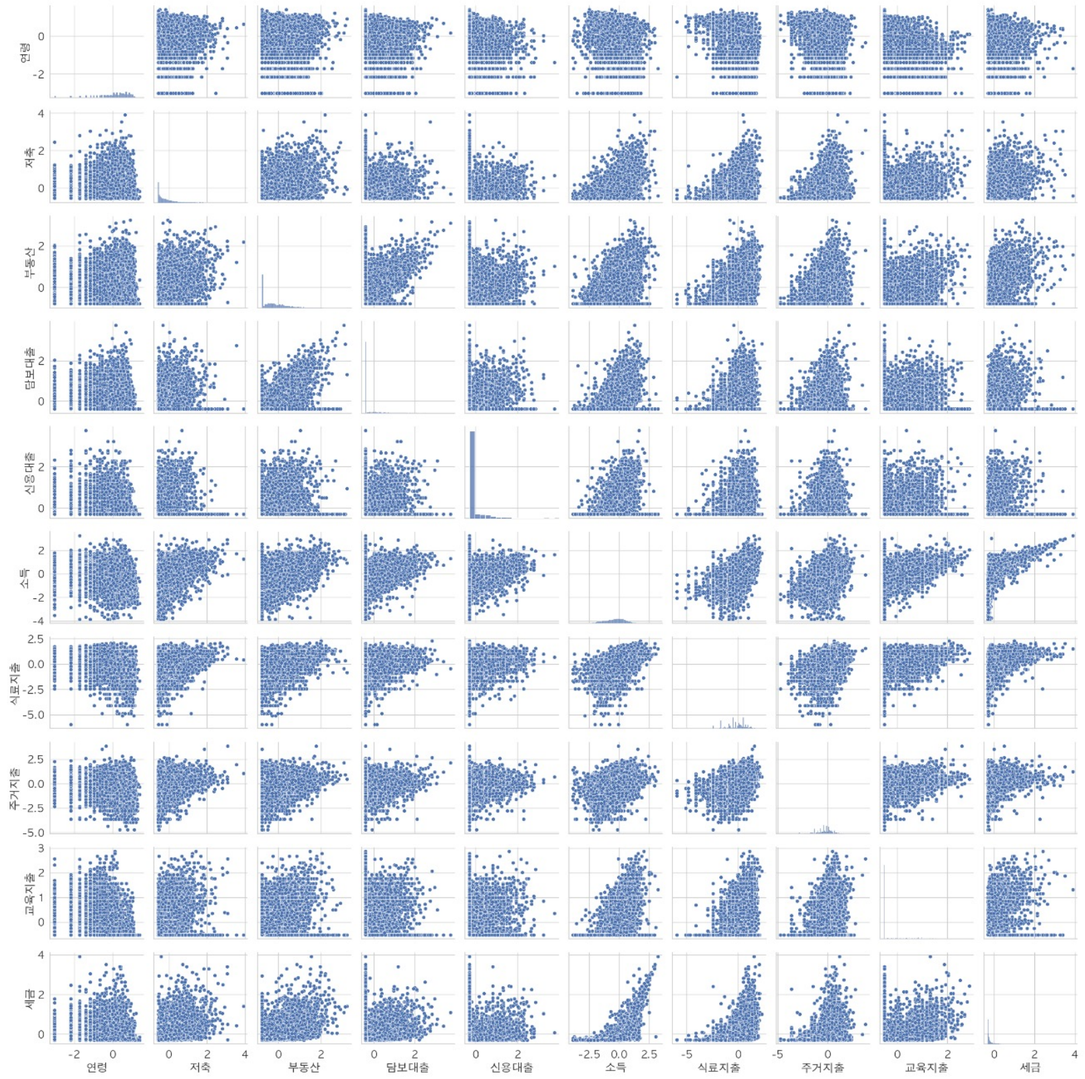


· 산점도를 살펴보면 대부분의 데이터들이 앞쪽에 몰려 있는 것을 확인할 수 있음

데이터 탐색

```
: 1 sns.pairplot(data_log[int_columns])
  2 plt.show
```

```
: <function matplotlib.pyplot.show(close=None, block=None)>
```

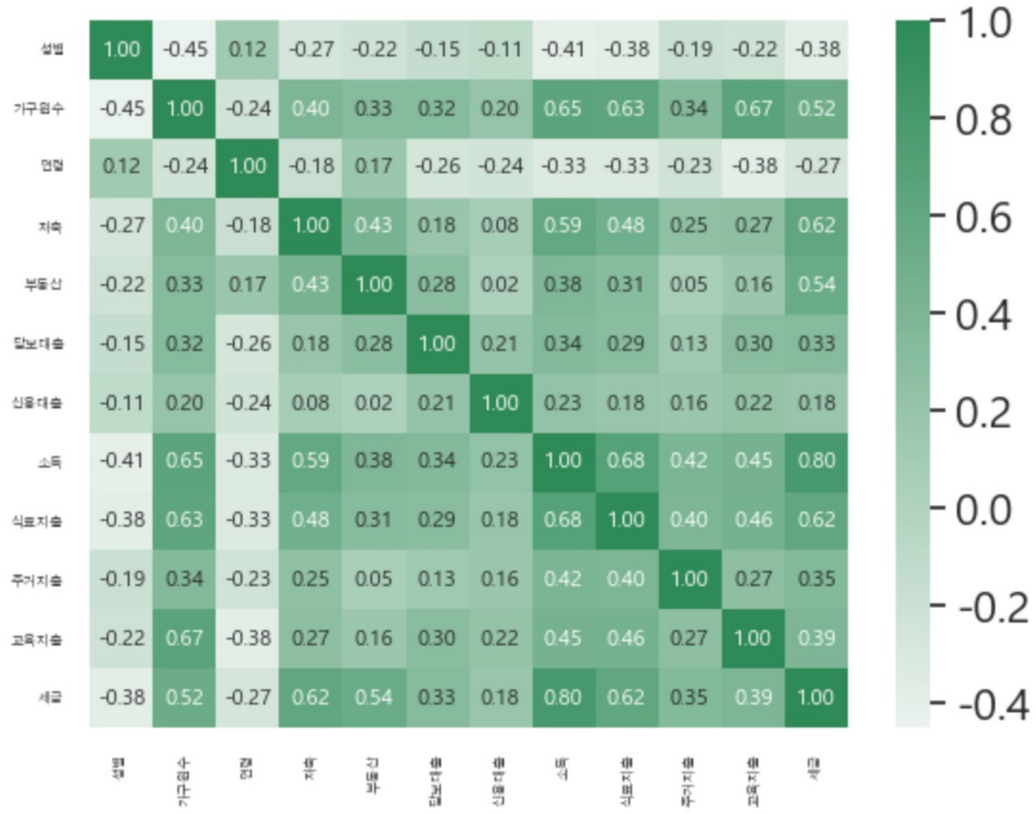


· 로그 변환 이후에 데이터가 앞쪽에 몰려 있는 현상이 줄어든 것을 확인할 수 있음

데이터 탐색

```
cmap = sns.light_palette("seagreen", as_cmap = True)
heatmap = sns.heatmap(data_log.corr(), annot=True,annot_kws={'size': 8},fmt='.2f', cmap=cmap)
heatmap.set_xticklabels(heatmap.get_xticklabels(), fontsize=5)
heatmap.set_yticklabels(heatmap.get_yticklabels(), fontsize=5)
plt.savefig("data_log_linear")

#모든 변수에 대한 상관관계 확인
```



- 반응변수 소득과 설명변수들 간의 상관관계를 보면 세금이 제일 소득과 상관관계가 높고, 신용대출이 소득과 상관관계가 가장 낮음

데이터 분석을 통한 문제해결

Dep. Variable:	소득	**R-squared:**	0.762
Model:	OLS	**Adj. R-squared:**	0.762
Method:	Least Squares	**F-statistic:**	1984.
Date:	Fri, 15 Dec 2023	**Prob (F-statistic):**	0.00
Time:	14:45:53	**Log-Likelihood:**	-10866.
No. Observations:	17954	**AIC:**	2.179e+04
Df Residuals:	17924	**BIC:**	2.203e+04
Df Model:	29		
Covariance Type:	nonrobust		

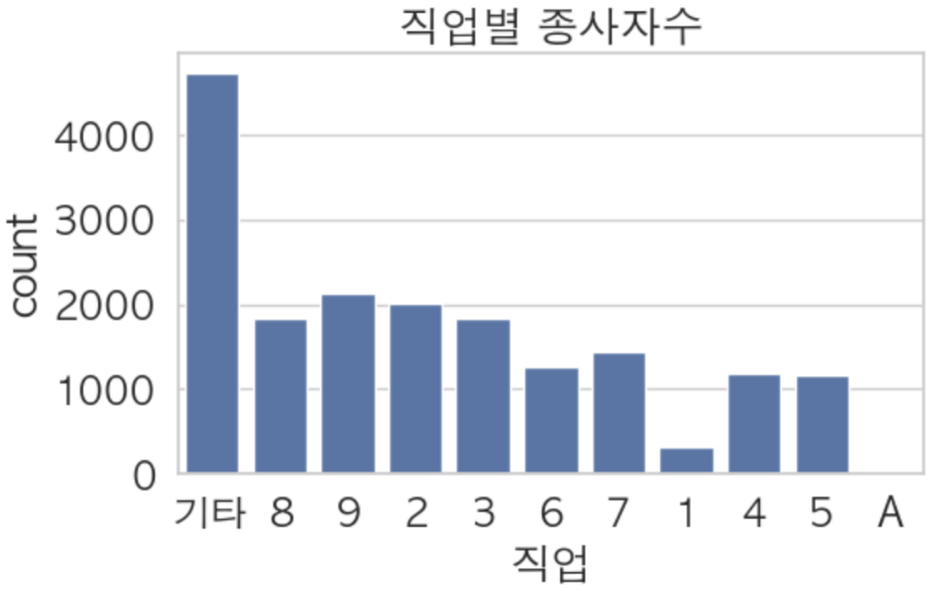
	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	4.8165	0.092	52.175	0.000	4.636	4.997
직업[T.2]	-0.0253	0.027	-0.935	0.350	-0.078	0.028
직업[T.3]	-0.0322	0.027	-1.191	0.234	-0.085	0.021
직업[T.4]	-0.0742	0.029	-2.591	0.010	-0.130	-0.018
직업[T.5]	-0.0704	0.028	-2.478	0.013	-0.126	-0.015
직업[T.6]	-0.0368	0.029	-1.260	0.208	-0.094	0.020
직업[T.7]	-0.0879	0.028	-3.147	0.002	-0.143	-0.033
직업[T.8]	-0.0367	0.027	-1.336	0.182	-0.090	0.017
직업[T.9]	-0.1033	0.028	-3.722	0.000	-0.158	-0.049
직업[T.A]	-0.0158	0.096	-0.165	0.869	-0.204	0.172
직업[T.기타]	-0.3344	0.027	-12.341	0.000	-0.388	-0.281
수도권여부[T.G2]	0.0141	0.007	1.911	0.056	-0.000	0.029
C(학력, Treatment(reference=6))[T.1]	-0.0793	0.020	-3.929	0.000	-0.119	-0.040
C(학력, Treatment(reference=6))[T.2]	-0.0586	0.015	-3.972	0.000	-0.088	-0.030
C(학력, Treatment(reference=6))[T.3]	-0.0151	0.014	-1.066	0.287	-0.043	0.013
C(학력, Treatment(reference=6))[T.4]	-0.0154	0.010	-1.481	0.139	-0.036	0.005
C(학력, Treatment(reference=6))[T.5]	-0.0192	0.012	-1.565	0.118	-0.043	0.005
C(학력, Treatment(reference=6))[T.7]	0.0680	0.018	3.758	0.000	0.033	0.103
C(학력, Treatment(reference=6))[T.8]	0.1765	0.030	5.982	0.000	0.119	0.234
성별	-0.0186	0.022	-0.838	0.402	-0.062	0.025
가구원수	0.6306	0.016	40.548	0.000	0.600	0.661
연령	0.0653	0.018	3.651	0.000	0.030	0.100
저축	0.0419	0.002	18.760	0.000	0.038	0.046
부동산	-0.0157	0.001	-16.472	0.000	-0.018	-0.014
담보대출	0.0075	0.001	8.394	0.000	0.006	0.009
신용대출	0.0083	0.001	7.335	0.000	0.006	0.011
식료지출	0.1669	0.007	23.493	0.000	0.153	0.181
주거지출	0.0722	0.006	13.021	0.000	0.061	0.083
교육지출	-0.0193	0.002	-11.464	0.000	-0.023	-0.016
세금	0.2271	0.003	79.166	0.000	0.221	0.233

Omnibus:	6231.028	**Durbin-Watson:**	1.946
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	63732.670
Skew:	-1.372	**Prob(JB):**	0.00
Kurtosis:	11.813	**Cond. No.**	508.

· R-squared: 0.762
· 유의성을 어긋나는 변수: 직업[T.2], 직업[T.3], 직업[T.6], 직업[T.8], 직업[T.A], 수도권여부, 학력[T.3], 학력[T.4], 학력[T.5], 성별

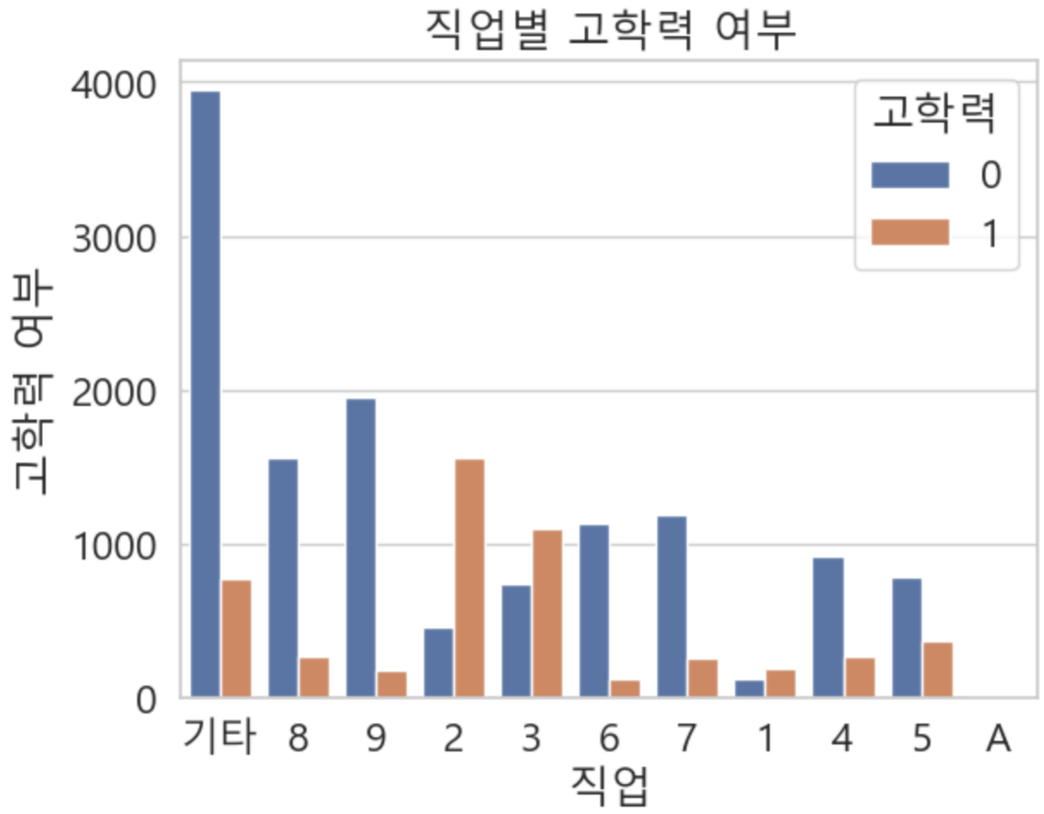
데이터 분석을 통한 문제해결

```
VIF of Intercept : 106.3837097189233
VIF of 직업[T.2] : 6.708083839942073
VIF of 직업[T.3] : 6.201924108007746
VIF of 직업[T.4] : 4.683874495612615
VIF of 직업[T.5] : 4.4820229663325035
VIF of 직업[T.6] : 5.092073469328004
VIF of 직업[T.7] : 5.344201670612241
VIF of 직업[T.8] : 6.399369250618629
VIF of 직업[T.9] : 7.37586727748963
VIF of 직업[T.A] : 1.0750654729051399
VIF of 직업[T.기타] : 12.988670848172294
VIF of 수도권여부[T.G2] : 1.0915565167414603
VIF of 학력[T.2] : 3.3234404980005183
VIF of 학력[T.3] : 3.2250868703709705
VIF of 학력[T.4] : 6.403677786459711
VIF of 학력[T.5] : 4.043006267934295
VIF of 학력[T.6] : 6.487458550945975
VIF of 학력[T.7] : 2.311125655897616
VIF of 학력[T.8] : 1.5393276397240663
VIF of 성별 : 1.4226738111024482
VIF of 가구원수 : 2.007487169389681
VIF of 연령 : 2.285361422755931
VIF of 저축 : 1.2863636333241804
VIF of 부동산 : 1.8127390548689721
VIF of 담보대출 : 1.3688080220473429
VIF of 신용대출 : 1.0646442444379207
VIF of 식료지출 : 2.0104282178943285
VIF of 주거지출 : 1.150456886677737
VIF of 교육지출 : 1.5662135195840556
VIF of 세금 : 1.3188326575310647
```



- 기타 직업에 대한 VIF 값이 12.988로 높은 편이지만 기타에 해당되는 관측치가 4000개가 넘기 때문에 제거하는 것이 좋다고 보기는 힘듦
- 직업과 학력의 VIF 값이 4~6정도를 나타냄

데이터 분석을 통한 문제해결



· 상대적으로 전문직으로 볼 수 있는 1, 2, 3번 직업의 고학력자 비율이 더 높은 것을 확인 가능

분석결과 도출 및 시사점

소득과 세금의 상관관계

회귀분석 모델을 통해 소득과 세금이 상대적으로 꽤 강한 양의 상관관계를 보인다는 것을 확인할 수 있음 소득세로 인해 나타나는 현상으로 추측 가능

소득과 학력의 상관관계

학력의 reference level을 대학교(4년제 이상)으로 두고 회귀분석 모델을 적합한 결과 사람들이 일반적으로 생각하는 것과 마찬가지로 학력이 낮을수록 소득이 더 적고, 학력이 높을수록 소득이 더 높은 것을 확인 가능

소득과 수도권 여부의 상관관계

수도권 여부의 reference level을 G1(수도권)에 두고 회귀분석 모델을 적합한 결과 사람들이 일반적으로 생각하는 것과 달리 오히려 비수도권의 소득이 더 높게 나타나는 것을 확인 가능