1인당 GDP와 간접 지표들의 연관관계 분석 및 예측

201300938 최영조

1. 분석 배경과 목적
    (1) 분석 배경
    (2) 분석 목적
    (3) 분석 방법

2. 자료 수집

3. 시각화 및 연관 관계 분석
    (1) 범주형 자료의 시각화
    (2) 연속형 자료의 시각화
    (3) 특이사항 및 개선시 수치 예측

4. 결론

□ 자료 출처

1. 분석 배경과 목적

(1) 분석 배경

○ 1인당 GDP는 국가의 경제력을 나타내는 중요 지표로서의 의미를 갖는다.

○ GDP는 국가 총 생산량을 고려하는 지표로서 해당 국가 내에서 생산된 재화와 용역의 시장가치 의 합이다. 이를 측정하기 위해서는 당연히 경제지표가 고려되고 직접적인 연관관계가 있을 것이다. 그러나 생산과 거래는 안정된 사회제반 위에서 이루어지므로 간접적인 지리, 사회 등 지표들과도 깊 은 연관관계가 있으리라 추측할 수 있다.

○ 정책을 수립할 때 정권의 성격에 따라, 국민의 요구에 따라 그 방향성이 갈릴 수 있다. 올바른 방 향성이라도 다른 가치를 중요시하는 여러 의견들 간에 충돌이 발생할 수도 있다. 그러나 서로 다른 지표들이 깊은 연관관계가 있고 어느 하나를 개선했을 때 다른 무언가도 개선될 수 있음을 보이면 사 회적 합의에 이르기 쉬워질 것이다.

(2) 분석 목적

○ 여러 국가들의 지표를 기반으로 각 항목에 따른 1인당 GDP와의 상관계수, 그래프의 모양 등 관계 를 살펴봄으로써 그 상호관계를 명확하게 나타낸다.

○ 대한민국과 다른 국가들의 수치를 비교해봄으로써 개선방향을 모색한다.

○ 1인당 GDP 상승을 위해 간접지표 중 가장 우선적으로 고려해야할 것은 무엇인지 파악하고 명시 적인 결론을 얻는다.

(3) 분석 방법

○ 신뢰할 수 있는 조사기관의 자료를 웹상에서 스크래핑한다. ('자료수집'에서 어느정도 상술한다.)

○ 스크래핑한 데이터는 하나의 DataFrame으로 합쳐 시각화한다.

○ 뚜렷한 값이 나타나는 지표가 있다면 예측 모델을 이용해 개선 후의 1인당 GDP 상승폭을 참고 수준에서만 예측해본다.

○ 수집하는 자료들은 다음과 같다.

- 면적, 인구, 평균수명, 출산율, 자살률, 실업률, 에이즈 감염률, 위치지역, 언론자유, 민주주의지수, 청렴도, 정치체 제, 징병제 여부, 석유생산량, 내륙국 여부, 인지도, 최상위 교육기관 보유

○ 수집하는데 시간이 오래 걸릴 수 있으니 엑셀 파일로 저장을 병행하며 진행한다.

2. 자료 수집

사용한 라이브러리들은 아래와 같다.

```
In [2]: import pandas as pd
        import matplotlib.pyplot as plt
        import numpy as np
        import seaborn as sns
        from selenium import webdriver
        from bs4 import BeautifulSoup
        from tqdm import tqdm_notebook
        %matplotlib inline
```

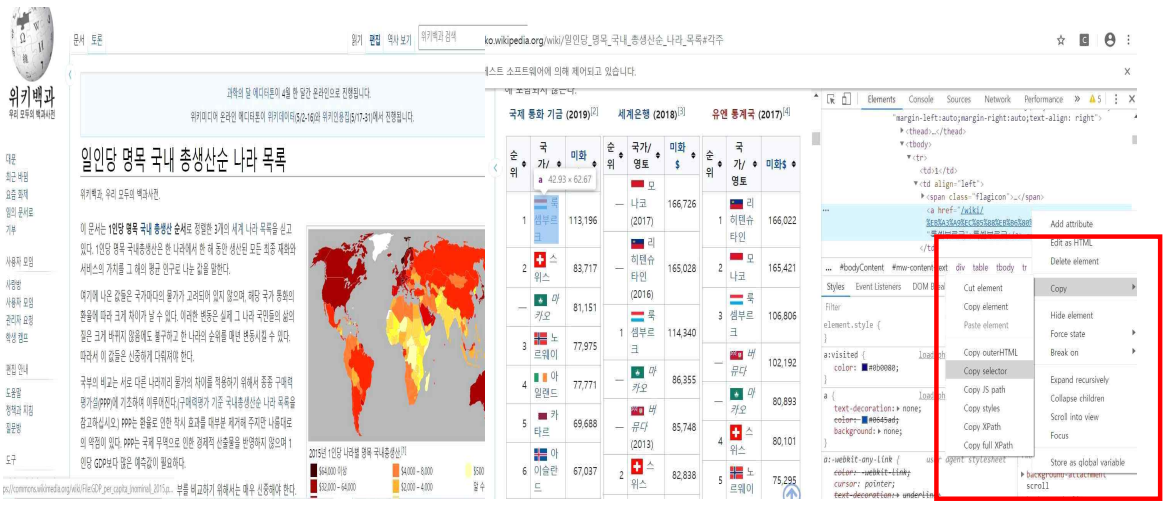
request를 이용하여 스크랩할 수도 있지만 페이지를 바꿔가며 수집해야할 자료들도 있기 때문에 크롤링 페이지 컨트롤이 용이한 selenium의 webdriver를 사용하기로 한다.
tqdm 모듈의 경우 장시간 for문을 이용해 스크래핑할 경우를 위해 임포트하였다.

(지표마다의 수집은 각 페이지마다 약간씩 차이가 있는 코드를 통해 얻어온다. 그러나 전체적인 방법은 비슷하므로 자료 수집에 있어서 초반 과정만 설명하고 나머지는 세세하게 설명하지는 않도록 한다.)

우선 1인당 GDP 정보를 얻어온다.

```
In [5]: url = 'https://ko.wikipedia.org/wiki/일인당_명목_국내_총생산순_나라_목록'
        driver.get(url)
```

이와 같은 명령을 내려 크롤링 창을 아래와 같이 해당 페이지에 접속시킨다.



원하는 정보에 해당하는 HTML을 이용하거나 더 단순하게는 select 정보를 복사해와 웹에서 글자나 숫 자 등을 얻어올 수 있다.

```
In [4]: # 위키백과 확인 결과 마지막 순위는 남수단이므로 남수단이라는 이름이 나올 때까지 실행하자.
        i=1
        nation = ""
        nation_list=[]
        while nation != "남수단":
            info = driver.find_element_by_css_selectorW
            ('#mw-content-text > div > table > tbody > tr:nth-child(2) > td:nth-child(1) > table > tbody > tr:nth-child({})'.format(i))
            nation_info = info.text.split(" ")
            nation = nation_info[1]
            nation_list.append(nation_info)
            i += 1
        nation_list[181:187]

Out[4]: [['178', '시에라리온', '547'],
         ['179', '아프가니스탄', '513'],
         ['180', '콩고', '민주', '공화국', '501'],
         ['181', '모잠비크', '484'],
         ['182', '마다가스카르', '464'],
         ['183', '중앙아프리카', '공화국', '448']]
```

반복문을 사용하여 정보를 얻어오는데 전체 개수가 몇 개인지 정확히 파악이 어려우므로 마지막 항목이 나타날 때까지 반복을 수행하게 한다.

```
In [5]: # ['180', '콩고', '민주', '공화국', '501'] 처럼 나라 이름이 잘 못 띄어쓰기 된 경우가 있으므로 이를 수정해주자.

        for i in range(len(nation_list)):
            nation = nation_list[i]
            if len(nation) != 3:
                name = " ".join(nation[1:-1])
                rank = nation[0]
                GDP_per_capita = nation[-1]
                nation_list[i] = [rank,name,GDP_per_capita]

In [6]: nation_list[181:187]

Out[6]: [['178', '시에라리온', '547'],
         ['179', '아프가니스탄', '513'],
         ['180', '콩고 민주 공화국', '501'],
         ['181', '모잠비크', '484'],
         ['182', '마다가스카르', '464'],
         ['183', '중앙아프리카 공화국', '448']]
```

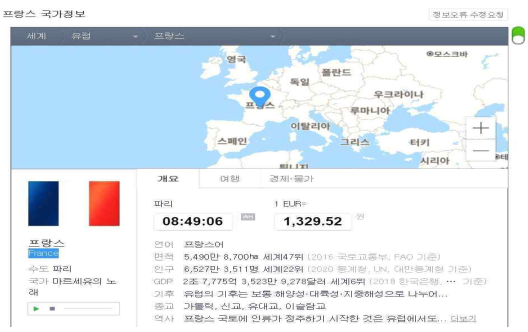
정보가 수집될 때는 표기가 흐트러지지 않도록 수정을 병행해가면서 진행한다.

얻어진 정보는 마카오, 홍콩 등이 섞여있기도한데 UN 가입국을 기준으로만 수집을 하도록한다. UN 가 입국 역시 동일한 방법으로 얻어온다. UN가입국 중에서 1인당 GDP까지 잘 집계되어있는 국가들만을 기 준으로 차후 자료들을 수집하기로 한다.

```
In [27]: first_result = pd.merge(GDP_df,UN_df,on='국가',how='inner')
         first_result
```

Out[27]:

	순위	국가	1인당_GDP	영문
0	1	룩셈부르크	113,196	Luxembourg
1	2	스위스	83,717	Switzerland
2	3	노르웨이	77,975	Norway
3	4	아일랜드	77,771	Ireland
4	5	카타르	69,688	Qatar
...	...	...	...	...
180	184	니제르	405	Niger
181	185	말라위	371	Malawi
182	186	에리트레아	343	Eritrea



영문명의 경우 네이버에 국가 검색시 나타나는 정보를 활용했다.(오른쪽)

병합을 진행해 감에 있어서 각 항목 페이지마다 국가의 표기가 약간씩 차이가 있기 때문에 신경을 써주 면서 진행했다.

```
In [25]: # 가끔 국가명이 똑같이 써있지 않아 문제가 생기는 경우가 있다. GDP가 집계된 나라 중 UN회원국에 들어있지 않은 나라들을 출력한다.

         un_list = list(UN_df.국가)
         GDP_df[GDP_df['국가'].isin(un_list)==False]
```

Out[25]:

	순위	국가	1인당_GDP
2	—	마카오	81,151
14	—	홍콩	49,334
28	—	푸에르토리코	31,538
37	35	타이완	24,878
68	66	중국	10,099
109	107	코소보	4,442

다른 국가들은 UN 회원국이 아니므로 포함되지 않아도 상관 없지만 중국의 경우 어떤 항목에는 '중국', 어떤 항목에는 '중화인민공화국'이라 표기되어있어 일치하지 않는 일이 종종 발생한다.

```
In [40]: first_result = first_result.append({'순위':85,'국가':'중화인민공화국','1인당_GDP':'10,099','영문':'China'},ignore_index=True)
         first_result
```

```
In [76]: first_result.sort_values(by='순위',inplace=True)
         # 중국에 대한 index가 이상한 상황이므로 인덱스를 재구성
         first_result.reset_index(drop=True,inplace=True)
```

```
In [83]: first_result
```

Out[83]:

	순위	국가	1인당_GDP	영문
0	1	룩셈부르크	113,196	Luxembourg
1	2	스위스	83,717	Switzerland
2	3	노르웨이	77,975	Norway
3	4	아일랜드	77,771	Ireland
4	5	카타르	69,688	Qatar
...	...	...	...	...
181	184	니제르	405	Niger
182	185	말라위	371	Malawi
183	186	에리트레아	343	Eritrea
184	187	부룬디	310	Burundi
185	188	남수단	275	Republic of South Sudan

186 rows × 4 columns

```
In [84]: first_result.to_csv('중간과정/first_result.csv',index=False)
         first_result.to_excel('중간과정/first_result2.xlsx',index=False)
```

이와 같이 일치하지 않는 표기에 대해서는 따로 처리해주면서 진행했다. 또 합쳐진 데이터는 엑셀 파일 로 저장해가면서 자료를 수집하였다.

다른 자료들의 경우 대부분 위키백과 페이지를 활용하였다. 위키 페이지 중에서도 신뢰할 수 있는 1차 조사기관이 있는 경우만 사용하였다. 다만 '인지도'의 경우 구글에 영문명으로 검색 시 얻어지는 결과 수 를 기준으로 하였다. '위치지역'의 경우 네이버 지식 백과 '유엔의 지역그룹' 항목을 참조하였다. 미국, 캐나다, 이스라엘 등은 서유럽권으로 분류되어있으며 지역 간 격차를 확실히 하기 위해 한중일은 아시아 대신 동아시아로 따로 분류하였다.
'석유생산량'은 석유매장량이 아닌 가공 혹은 생산량에 관한 것으로서 대한민국 역시 어느정도 수치를 갖 는다. '내륙국 여부'는 국가면적 대비 해안선 비율로써 0인 국가는 내륙국, 100이 넘어가는 국가는 바다 가 충분히 있다는 뜻에서 100으로 통일하였다. '최상위 교육기관 보유'는 '글로벌대학리더포럼'이라는 커 뮤니티에서 선정된 최상위 26개 대학을 얼마나 갖고 있는지에 관한 지표다. '정치체제'의 경우 민주주의 와 공산주의, 전제군주제를 분류했다. 왕에게 실권이 없다는 점에서 입헌군주제는 민주주의와 같이 분류 했다.

모든 지표들의 스크래핑은 이미 상술된 부분과 아주 비슷한 방식으로 진행하였기에 세부적인 과정설명은 생략한다.
모든 자료를 수집한 데이터는 다음과 같다.

	순위	국가	1인당_GDP	영문	인지도	지역	정치체제	최상위_교육기관_수	복무기간	인구	...	기대수명	출산율	10만명당_자살률	언론자유지수	민주주의지수	청렴도	실업률	면적 대비 해안선 비율	에이즈감염률(%)	석유생산량
0	1	룩셈부르크	113196.0	Luxembourg	515000000	서유럽권&기타	민주주의	0	0	613894.0	...	81.34	1.5	8.5	15.66	8.81	81.0	6.5	0.000	0.0	0.0
1	2	스위스	83717.0	Switzerland	712000000	서유럽권&기타	민주주의	2	0	8542323.0	...	82.66	1.5	10.7	10.52	9.03	85.0	5.2	0.000	0.0	0.0
2	3	노르웨이	77975.0	Norway	572000000	서유럽권&기타	민주주의	0	0	5334762.0	...	81.32	1.7	9.3	7.82	9.87	84.0	4.0	100.000	0.0	1647975.0
3	4	아일랜드	77771.0	Ireland	1090000000	서유럽권&기타	민주주의	0	0	4857000.0	...	80.57	1.9	11.1	15.00	9.15	73.0	6.0	21.019	0.2	0.0
4	5	카타르	69688.0	Qatar	717000000	아시아권	전제군주제	0	1년 이하	2772294.0	...	78.88	1.9	5.7	42.51	3.19	62.0	0.4	49.226	0.0	1522902.0

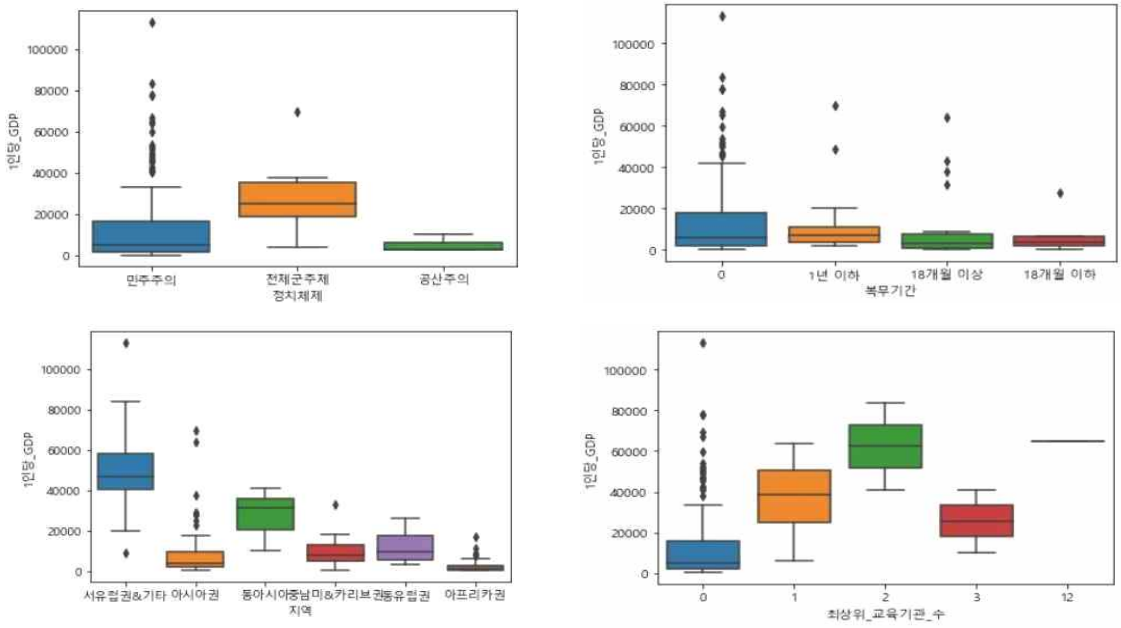
5 rows × 21 columns

DataFrame은 1인당 GDP순으로 정렬되어있으며 크기가 큰 관계로 head(5)까지 출력한 것을 가져왔다. (해당 DataFrame은 all.xlsx이라는 엑셀파일로 저장되어있으니 참고해주시기 바랍니다.)

3. 시각화 및 연관 관계 분석

- 해당 DataFrame에서 '1인당_GDP' 컬럼을 기준으로 각 항목들과의 연관성을 살펴본다.

(1) 범주형 자료의 시각화



- **위치지역** : 지역에 따른 편차는 매우 크다고 할 수 있다. 서유럽이 제일 앞서나가는데 서유럽권의 중 위수만 보더라도 다른 지역들보다 훨씬 높은 GDP를 갖는다. 아시아의 경우에는 한국과 일본이 제외되었 음에도 국가마다 편차가 상당히 크다는 것을 확인할 수 있다. 아프리카의 경우 대부분의 1인당 GDP가 아주 낮고 지역 내에서 상위 25%를 넘어가는 국가들마저도 수치가 매우 좋지 않다.

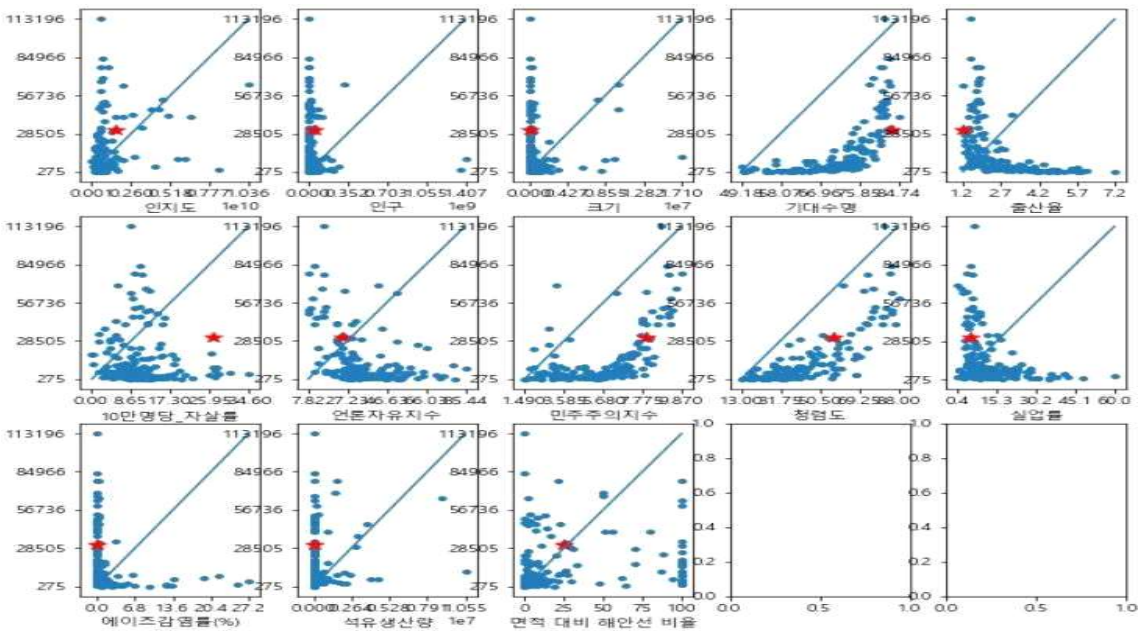
- **정치체제** : 정치체제는 민주주의 국가보다 전제군주제가 높다는 약간 이상한 결론이 나온다. 다만 이것은 절대왕정을 유지하는 산유국들로 인한 수치이다. 또 민주주의 국가의 스펙트럼이 너무 넓기 때문에 평균만 따졌을 때는 전제군주제가 앞서 나가는 것으로 보이는 것이다. 실제로 boxplot을 보면 gdp 수치가 아주 높은 국가들은 거의 민주주의 국가에 있다. 공산국가가 비교적 못사는 것은 뚜렷해 보인다.

- **복무기간** : 징병제가 아닌 국가들 중 1인당 GDP가 높은 국가가 많지만 대부분의 국가가 징병제를 택 하지 않기 때문에 그렇게 보이는 것이다. 중위수를 본다면 복무기간은 1인당 GDP와 큰 관련이 없어보인 다.

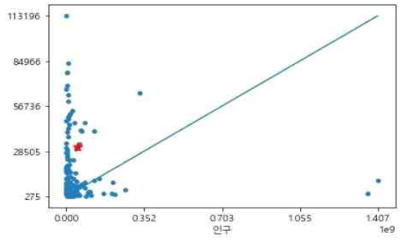
- **최상위 교육기관** : 최고의 교육기관 모임에는 가입하지 않아도 1인당 GDP가 높은 나라들은 많지만 일 단 최상위 교육기관이 있으면 대체로 (남아공 제외) 1인당 GDP가 높은 것을 확인할 수 있다.

(2) 연속형 자료의 시각화

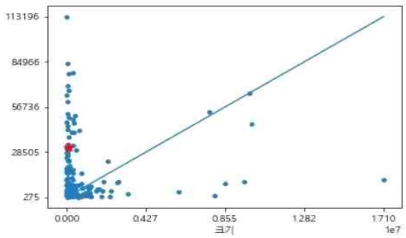
모든 연속형 자료들과 1인당 GDP의 관계를 시각화하면 다음과 같다. 그래프에 나타나는 직선은 지표간 상관계수가 1이라 가정했을 때의 모양이다. 즉 직선을 중심으로 점들이 뭉쳐있다면 상관계수가 높다는 의미이며 직선과 대칭 모양으로 뭉쳐있으면 음의 상관계수가 높다고 생각해볼 수 있다.
빨간 색 별은 대한민국이다.



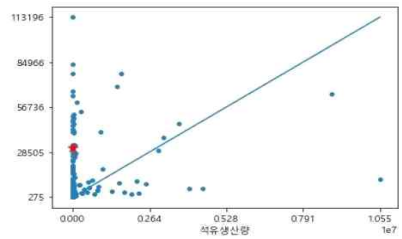
위 그래프를 하나하나씩 살펴보도록 한다.



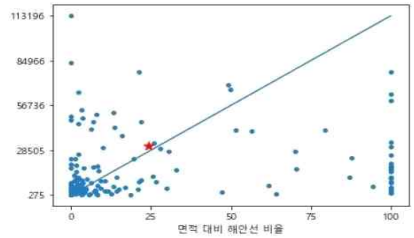
1인당 GDP와 인구의 상관계수는 -0.03638911553048926 입니다.



1인당 GDP와 크기의 상관계수는 0.056890324120920974 입니다.



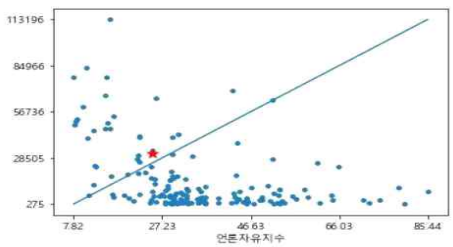
1인당 GDP와 석유생산량의 상관계수는 0.15277831340796355 입니다.



1인당 GDP와 면적 대비 해안선 비율의 상관계수는 0.1555671406342624 입니다.

-인구, 면적 크기, 석유생산량, 자살률은 1인당 GDP와 별다른 연관성이 발견되지 않는다.

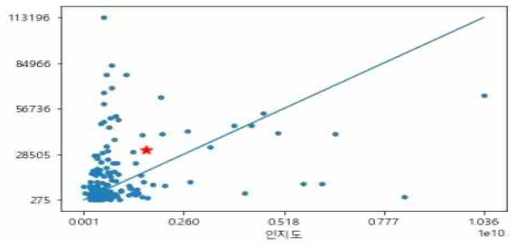
이 중 인구와 면적의 경우 아주 수치가 높은 국가(러시아, 캐나다, 중국, 인도 등)와 그렇지 않은 국가들 사이의 격차가 매우 크기 때문에 유의미한 비교가 되기 어려워 보인다.



1인당 GDP와 언론자유지수의 상관계수는 -0.421931719174487 입니다.

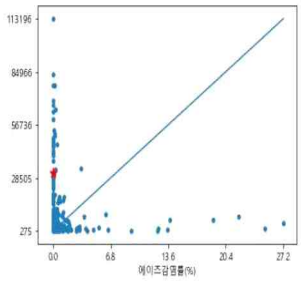
95	자메이카	11.13	5461.0
17	벨기에	12.07	45176.0
57	코스타리카	12.24	12015.0
34	에스토니아	12.27	23524.0
36	포르투갈	12.63	23031.0
15	독일	14.60	46564.0
5	아이슬란드	14.71	67037.0
3	아일랜드	15.00	77771.0
12	오스트리아	15.33	50023.0
0	룩셈부르크	15.66	113196.0
16	캐나다	15.69	46213.0
48	우루과이	16.06	17029.0
86	수리남	16.38	6311.0
104	사모아	16.55	4501.0
9	오스트레일리아	16.55	53825.0
90	나미비아	18.95	5842.0
42	라트비아	19.53	18172.0
117	카보베르데	19.81	3599.0
137	가나	20.61	2223.0
31	키프로스	21.74	27720.0
28	스페인	21.99	29961.0
40	리투아니아	22.05	19267.0
88	남아프리카 공화국	22.19	6100.0

- **언론자유지수** : 언론자유지수는 값이 작을수록 좋은 지표이다. 언론자유가 보장된 국가들 중에 GDP도 높은 국가가 많다. 단 언론자유지수 상위 30에 들어가는 국가들(오른쪽) 중에는 1인당 GDP가 낮은 나라 들도 꽤 포함되어있다. 상관계수의 크기는 아주 높다고 하기는 어렵다.

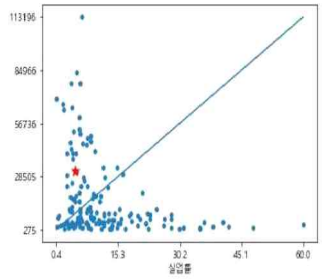


1인당 GDP와 인지도의 상관계수는 0.29673430003550416 입니다.

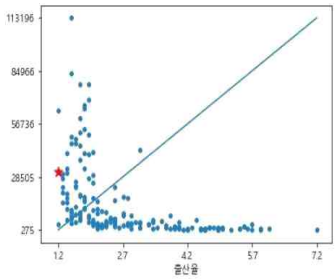
- **인지도** : 인지도는 국가 브랜드적인 측면에서 연관이 있을 것으로 추측했으나 큰 연관성은 없어보인다. 안 좋은 쪽으로 많이 검색되는 나라들도 있고, 룩셈부르크 같은 유럽 소국들처럼 높은 1인당 GDP임에도 사람들에게 익숙하지 않은 국가들도 있다.



1인당 GDP와 에이즈감염률(%)의 상관계수는 -0.17887414053177159 입니다.

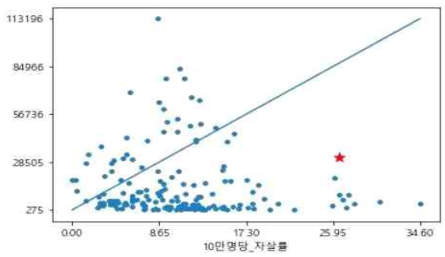


1인당 GDP와 실업률의 상관계수는 -0.3186854036643336 입니다.



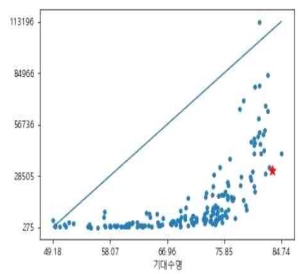
1인당 GDP와 출산율의 상관계수는 -0.48606288670288755 입니다.

- **에이즈 감염률, 실업률, 출산율** : 상관계수는 별로 크지 않으나 어느 정도의 연관성은 보인다. 해당 지 표가 0에 가까운 나라들은 1인당 GDP가 높은 나라도 있고 낮은 나라도 있으나, 해당 지표가 높은 나라 들은 모두 1인당 GDP가 낮은 나라들이다. 낮은 에이즈 감염률, 실업률, 출산율은 높은 1인당 GDP에 대 해 충분조건이 되지는 못하지만 필요조건은 된다는 것을 확인할 수 있다. 해당 지표가 높은 국가들은 제 기능을 발휘하지 못할 가능성이 높다.

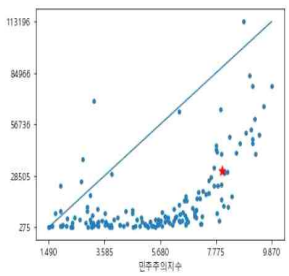


1인당 GDP와 10만명당_자살률의 상관계수는 -0.06777523258018017 입니다.

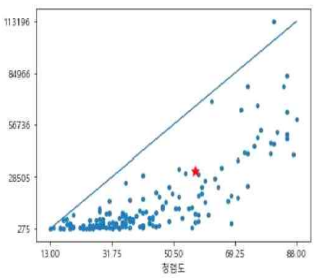
- **자살률** : 자살률은 대한민국이 심각한 수준이나 1인당 GDP와의 연관성 측면에서는 별다른 관계는 없 어보인다.



1인당 GDP와 기대수명의 상관계수는 0.6401493033129791 입니다.



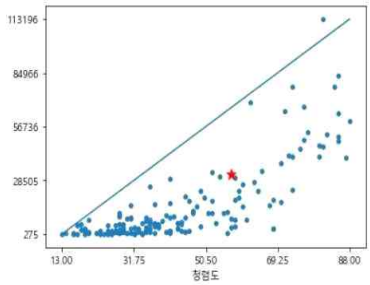
1인당 GDP와 민주주의지수의 상관계수는 0.5868675157418239 입니다.



1인당 GDP와 청렴도의 상관계수는 0.7915323370933969 입니다.

- **기대수명, 민주주의지수, 청렴도** : 기대수명, 민주주의 지수, 청렴도는 1인당 GDP와 큰 상관계수를 갖는다. 특히 청렴도의 경우 0.79가 넘어가는 큰 상관계수를 보여준다. 그래프의 모양을 보면 일차함수보다는 지수함수의 모양으로 분포가 되어있는데 어느 정도 선까지는 GDP와 연관이 비교적 적다가도 연관이 매우 커지는 지점이 있을 것이라고 추측이 가능하다. 발전 중인 개발도상국과 선진국의 차이라고도 볼 수 있을 듯하다.

(3) 특이사항 및 개선시 수치 예측

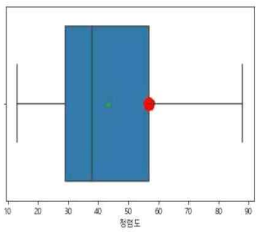


1인당 GDP와 청렴도의 상관계수는 0.7915323370933969 입니다.

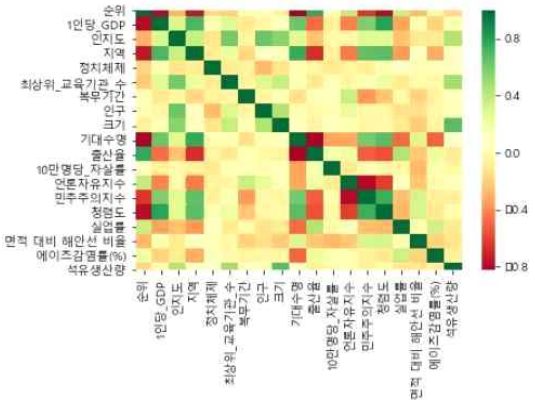
살펴본 그래프 중 가장 눈여겨볼 지표는 '청렴도'이다. 1인당 GDP와의 상관계수가 0.5 이상인 지표 중 기대수명과 민주주의지수는 대한민국 역시 상당히 상위권이다. 그러나 청렴도의 경우 가장 상관관계가 0.79로써 제일 높음에도 불구하고 대한민국의 수치가 그다지 높지 못하다.

다음은 청렴도의 수치의 first quartile, median, mean, third quartile과 함께 나타낸 boxplot이다.

```
In [56]: sns.boxplot(df['청렴도'],showmeans=True)
         plt.scatter(df['청렴도'].iloc[26],y=0,c='red',marker='*',linewidth=10)
Out[56]: <matplotlib.collections.PathCollection at 0x1f5ba95f4c8>
```



대한민국의 수치는 빨간색으로 표시되었는데 1/4 가량의 국가들이 대한민국보다 높은 청렴도를 갖고 있는 것을 확인할 수 있다. 청렴도의 수치가 지수함수 모양으로 퍼져있었으니 청렴도 개선 시 1인당 GDP 상승률이 매우 클 것으로 기대할 수 있다.



추가로 청렴도와 상관계수의 절댓값이 큰 다른 지표로는 지역, 기대수명, 언론자유지수, 민주주의지수, 기대수명, 출산율 등이 있으므로 해당 지표들을 개선해 청렴도를 같이 개선하는 방향을 생각해볼 수도 있다.

그렇다면 청렴도를 OECD 평균수준으로 끌어올렸을 때 1인당 GDP는 어떨지 예측해본다.

(단, 해당 데이터셋이 완벽하게 들어맞는 모델이 아닐 수 있으므로 다음 예측은 그저 참고수준이다.)

다음은 대한민국과 OECD 평균 수치를 비교해놓은 표이다.

국가	대한민국	NaN
1인당_GDP	31431	41112.7
인지도	1.61e+09	1.91739e+09
지역	4	4.3
정치체제	1	1
최상위_교육기관_수	1	0.6
복무기간	3	0.3
인구	5.17806e+07	3.62629e+07
크기	100378	1.00879e+06
기대수명	83.31	79.9
출산율	1.2	1.7
10만명당_자살률	26.5	11.7
언론자유지수	24.94	20.5
민주주의지수	8	8.1
청렴도	57	68.4
실업률	4.9	6.8
면적 대비 해안선 비율	24.242	23.4
에이즈감염률(%)	0	0.1
석유생산량	1000	501384

해당 수치들은 높아야 좋은 것도 있고 낮아야 좋은 수치들도 있다.
지리적인 요건을 제외하고 대한민국은 '기대수명, 실업률, 최상위 교육기관' 수치에서 앞서고 있고 '자살율, 민주주의 지수, 언론자유지수, 복무기간, 석유생산량, 인지도, 출산율, 청렴도' 측면에서는 뒤처지고 있다.
해당 수치들은 측정 기준이 다르므로 무엇이 더 심각하게 차이가 있다는 식으로의 해석은 하기 힘들다.

이 중 청렴도를 개선했을 때의 대한민국의 GDP 순위를 예측해보겠다.

빈 값의 경우 위의 히트맵 이미지에서처럼 지표간 상관계수가 높은 다른 지표들을 파악해 그것을 기준으 로 그룹을 만들어 그 평균값을 채워넣었다.

해당 데이터셋을 랜덤포레스트에 학습 시키고 성과를 측정했을 때 0.89가 나왔다.

```
In [20]: print(tree.score(X_test,y_rank_test))
         0.8936215137498238
```

```
In [23]: # 대한민국에 대한 예측값이다. 5순위 정도 차이가 난다.
         dd = X.iloc[26]
         dd2=np.array(dd)
         tree.predict(dd2.reshape(-1,17))
Out[23]: array([32.73333333])
```

대한민국의 실제 순위는 27위이며 해당 모델이 예측한 순위는 32.73위 정도이다.

다음은 랜덤포레스트 모델이 예측한 OECD 평균 수준의 청렴도를 가진 대한민국의 1인당 GDP 순위이 다.

```
In [40]: # 청렴도의 개선

         korea_updated = np.array([1280000000, 4, 1, 1, 3, 51780579, 100378, 83.31,
                1.2, 26.5, 24.94, 8.0, 57.0+11.3509, 4.9, 24.242, 0.0, 1000], dtype=object)
         tree.predict(korea_updated.reshape(-1,17))
Out[40]: array([25.6])
```

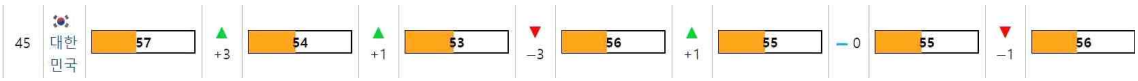
25.6위로 예측되었다.

원래 대한민국의 수준을 5위 정도 낮게 예측한 것을 감안한다면 청렴도가 OECD 평균만 되더라도 20위 수준으로 올라갈 수 있을 것이라고 생각해 볼 수 있다.

4. 결론(Discussion & Conclusion)

가장 눈에 띄는 결론은 1인당 GDP를 높이기 위해서는 청렴도를 개선할 필요가 있다는 것이다.

정경유착, 뇌물수수, 자녀 부정청탁, 부동산 투기 등에 대해 연일 뉴스가 끊이지 않고 고위 공직자의 윤 리적 해이에 관하여서 항상 도마에 올라있다.



이것은 위키피디아에서 가져온 대한민국의 년도별 부패인식지수(CPI, 수집한 자료의 청렴도 컬럼이 이 부패인식지수임) 이다. 제일 오른쪽이 2012년 수치이고 가장 왼쪽이 2018년 수치인데 크게 개선되는 것이 없는 것을 알 수 있다.

사람에 따라서 공직자들이 능력만 있으면 개인의 부정에 대해서 너무 민감해서는 안된다 얘기하는 경우도 있는데 데이터의 분석을 통해서 이런 것들이 실질적으로 대한민국의 발전을 가로막는 장애물이 맞고 그 영향력 역시 아주 크다는 사실을 확인할 수 있었다. 위에 상술했듯 청렴도의 분포는 지수함수 모양으 로써 개선됐을 때 예상되는 1인당 GDP 상승폭이 매우 클 것으로 생각해 볼 수 있다.

그 외 위의 분석들에 바탕으로 각 지표들 내에서 대한민국의 수치를 고찰해본다면 다음과 같다.

- **위치지역** : 주변국가들이 누구인가 자체만으로 1인당 GDP에 상당한 영향을 미칠 수 있다. 아시아 국 가들에 비해서는 크게 웃돌고 있으나 서유럽 국가들에게는 살짝 미치지 못한다. 국가 감정을 배제하고 주변국들과의 동반 성장을 추구하면 시너지가 있을 듯 하다.

- **정치체제** : 민주주의 국가들은 1인당 GDP의 스펙트럼이 아주 넓고 대한민국은 평균치와 관계없이 민주주의 국가 그룹 중 상위권에 있으니 다른 정치체제와 비교하는 것은 의미가 없어보인다.

- **복무기간** : 대한민국이 징병제 국가라 예민한 부분이지만 1인당 GDP와의 연관 자체는 별로 없다.

- **최상위 교육기관** : 대다수 국가의 수치가 0이기는 하지만 일단 가지고 있는 국가는 1인당 GDP가 대체 로 높다. 대한민국의 카이스트가 최상위 26개 대학 중 하나에 속하고 있다. 미국이 압도적으로 많고 나머지는 0~3개 수준이므로 대한민국의 경제적 위치에서 대학의 역할을 긍정적으로 바라볼 수 있다.

- **인지도, 인구, 면적, 석유생산, 해안선 비율**은 대한민국이 1인당 GDP를 위해 고려할 사항들은 아닌 듯 하다.

- **자살률**의 경우 대한민국이 상당히 높지만 1인당 GDP의 관점에서만 볼 때는 연관성이 거의 없다.

- 낮은 **에이즈 감염률, 실업률**은 높은 1인당 GDP의 필요조건이지만 대한민국은 나쁘지 않은 수치를 갖 고 있기 때문에 크게 고려할 필요는 없어보인다.

- **출산율**은 심각하다는 논의가 있지만 1인당 GDP가 아닌 다른 관점에서 논의되어야할 문제인 듯하다.

- **언론자유지수**는 1인당 GDP와의 직접적인 연관관계는 아주 크지 않지만 1인당 GDP와의 상관계수가 높은 청렴도와 민주주의지수와의 상관계수가 높으므로 개선할 필요가 있어보인다.

- **민주주의지수, 기대수명**은 1인당 GDP와의 연관관계도 높고 청렴도처럼 지수함수 모양을 갖지만 대한 민국의 현재 수치가 괜찮은 편이다.

이상으로 대한민국의 개선 방향을 모색한다는 원래 목적에 부합하게, 데이터를 통해 얻은 결론을 정리할 수 있다.

다만 지표들의 GDP와의 연관성을 충분히 담아내기에는 질적인 측면에서 충분하지 못했던 것들이 있다.

- 존재하는 국가 이상의 데이터를 모집할 수 없다보니 데이터의 개수가 충분하지 않았다.(186개)
- 각 국가의 상황을 충분히 반영하기에는 지표수가 부족했다.
- 영토 면적은 양적인 부분만이 아니라 평지와 산지 비율, 천연자원 분포 정도 등 질적인 측면에서 측 정할 다른 것이 있었으면 관련된 세부적 관찰이 가능했을 듯 싶다.

□ 자료 출처

- 1인당 GDP : 국제통화기금(2019년)
기준 국가들은 1인당 GDP 자료를 구할 수 있는 국가 중 UN 가입국으로 한다.
- 인지도 : 구글에 국가명을 영문으로 검색했을 때 나오는 결과 수를 기준으로 한다.
- 면적 : CIA 월드팩트북(2005년),'면적' 항목에는 출처가 없으나 '해안선' 항목에 표시돼 있다.
- 인구 : CIA(2010년)
- 평균수명 : UN(2018년)
- 출산율 : CIA 월드팩트북(2018년)
- 자살률 : WHO(2017년)
- 실업률 : 국가별로 취합한 기관과 년도가 약간씩 다르지만 위키피디아 '실업률에 따른 나라 목록' 항목을 보면 출처를 일일이 확인할 수 있다.
- 에이즈감염자수 : CIA 월드팩트북(2016~2018년)
- 위치지역 : 네이버백과 '유엔의 지역그룹' 항목에는 외교부 제공 UN국들의 지역별 그룹이 나와있다. 단 미국, 캐나다, 이스라엘 등은 서유럽권과 같이 묶여있다.
- 언론자유 : RSF(2019년)
- 민주주의지수 : The Economist(2017년)
- 부패지수 : 국제투명성기구(TI)(2018년)
- 정치체제 : 위키백과 '공산국가','전제군주제' 항목
- 징병제 여부 : 위키백과 '징병제' 항목
- 석유매장 : U.S. Energy Information Administration(2016년)
- 내륙국 여부 : 해안선 길이를 전체 면적으로 나눈다. CIA 월드팩트북(2005년)
- 최상위 교육기관 보유 : '글로벌대학리더포럼' 참가 26개 대학의 소속 국가들

(징병제와 정치체제는 신뢰있는 조사기관이 굳이 필요하지 않아 위키피디아의 항목만 적어놓았다.)

cf) proposal에서는 예측모델을 주로 데이터를 다루겠다 하였으나 교수님의 지적사항을 있어 예측모델은 참고 수준에서만 약간 사용하고 EDA 위주의 방향으로 보고서를 작성했습니다.