데이터분석 및 시각화 과제2

과목	**데이터분석 및 시각화**
지도교수	**이현봉 교수님**
학과	**산업시스템공학과**
이름	**박영웅 오광혁 이윤주 정지훈**
학번	**2017112502 2017112570** **2017112520 2017112548**
제출일	**2021. 11. 29.**

프로젝트명: GDP 는 국가의 행복을 반영하는가?

1. 서론

1.1 GDP란?

GDP(Gross Domestic Product: 국내총생산)는 일정 기간에 한 국가 안에서 생산된 최종 생산물에 대한 시장 가치의 합을 말한다. 즉, 한 국가의 영역 내에서 가계, 기업, 정부 등 모든 경제주체가 일정기간 동안 생산한 재화 및 서비스의 부가가치를 시장가격으로 평가하여 합산한 것이다. GDP가 높아진다는 것은 그 국가가 더 많은 가치를 생산했다는 것을 의미하고, 이는 그 나라의 경제가 성장했다는 것을 말한다. GDP의 구성 항목은 교육 서비스, 운송 배달, 농장 수확, 공장 생산 등으로 다양하며 다음의 기준을 만족하면 포함된다. 기준은 '국경 안에서 생산된 것', '일정 기간내에 생산된 것', '판매용으로 최종 생산된 것', '새롭게 생산된 것'이다. 이렇듯 GDP는 각 국가의 총 생산량을 뜻하는 지표이기 때문에 그 나라의 경제 규모와 성장 속도를 보여주는 데 사용되고 있다.

1.2 GDP가 높은 국가일수록 행복할까?

GDP가 높아진다는 것은 그 국가가 더 많은 가치를 생산했다는 것을 의미하고, 통상적으로 이는 그 나라의 경제가 성장했음을 의미한다. 많은 경제학자와 정치인은 GDP를 국가의 경제 성장의 척도로 삼고, 이를 기준으로 경제성장을 이야기한다. 이에 따라 본 팀은 "GDP가 높은 국가일수록 행복할까?" 라는 질문에 대해 "그렇다"고 생각했다. GDP의 증가로 경제가 성장한다는 것은 노동, 자본과 같은 생산요소의 질적 증대와 기술의 발전을 의미하기 때문이다. 국가 경제가 풍요로움에 따라 국민들의 행복 또한 높아질 것이라고 예상한다. 따라서 GDP가 높을수록 행복 지수가 높을 것이라는 귀무가설을 세우고, 이에 대한 EDA및 시각화를 진행하여 상관관계를 분석해보고자 한다. 행복 지수 외에 상관관계를 가지는 변수가 있는지 또한 확인해보았다.

2. 본론

2.1 사용 데이터

사용 데이터	데이터 설명	데이터 출처
GDP (Current US$)	단일 연도 공식 환율을 사용하여 변환된 국가별 GDP	The World Bank
GDP growth (annual %)	국가별 연간 GDP 성장률	
CO2 emissions (kt)	국가별 이산화탄소 배출량	
Life expectancy at birth, total (years)	국가별 출생 시 기대수명	

World Happiness Report	-세계 행복 상태에 대한 조사 - ladder score, economic production, social support, freedom, absence of corruption, and generosity - 각 요인에 대한 세계 최저의 국가 평균과 동등한 가치를 가진 가상의 국가 보다 삶의 질을 높이는데 기여하는 정도를 추정하여 수치로 나타냄	World Happiness Report
How's Life Well-Being (Current Well-being average and deprivation)	OECD 국가의 Well-being과 관련된 다양한 지표들을 나타낸 데이터	OECD.stat

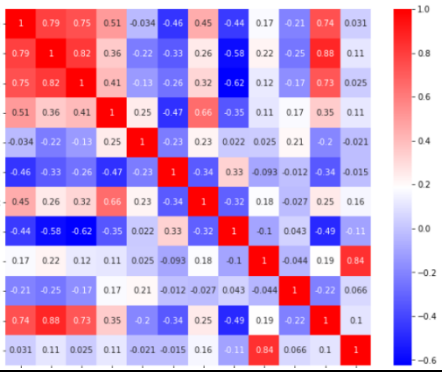
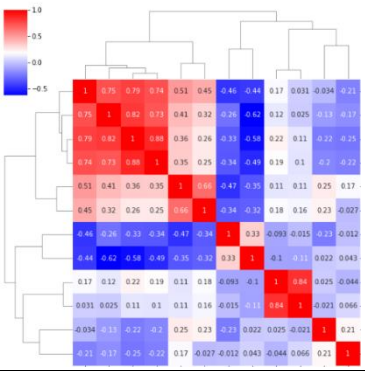
2.2 데이터 전처리

데이터 전처리에 앞서 사용할 데이터셋(csv파일)을 모두 불러와 데이터 프레임으로 저장하였다. 데이터프레임을 살펴보았을 때 The World Bank에서 가져온 데이터의 경우 최근에 가까운 데이터일 수록 결측치가 많다는 것을 확인할 수 있었다. 그래서 본 팀은 비교적 최근이면서 결측치가 최소로 있는 2018년 기준 데이터를 사용하기로 하였다. The World Bank 데이터에 따라서 World Happiness Report의 데이터 프레임 또한 2018년 기준 데이터만을 추출하여 사용하였다. 이렇게 2018년 기준으로 전처리된 데이터 프레임들을 Pandas의 merge를 이용하여 "total"이라는 통합 데이터 프레임으로 통합하였다. 마지막 과정으로 보다 다양한 시각화와 기존의 데이터 프레임에 없던 명목변수인 "대륙" 칼럼을 추가하여 시각화에 사용할 최종 데이터 프레임을 완성하였다.

2.3 데이터 시각화

2.3.1 GDP는 어떤 지표와 관련이 있을까?

아래의 시각화는 데이터 프레임의 모든 변수 간의 상관관계를 heatmap을 통해 나타내었다.

그림1. 변수간 상관관계를 나타낸 heatmap	그림2. 상관관계 별로 묶은 heatmap
	

위의 heatmap을 참고하여 각 변수들의 상관관계를 살펴보면 GDP는 CO2 배출량과 가장 큰 양의 상관관계를 가지고 있다. 하지만 CO2 배출량을 제외한 변수들과의 상관계수는 -0.093~0.22로 GDP는 나머지 변수들과 유의미한 상관관계를 갖지 못하는 것을 발견하였다. 이것을 통해 GDP는 단순히 생산성을 나타내는 지표일 뿐 국가를 나타내는 다양한 지표들을 대표할 수 없다고 할 수 있다.

2.3.2 GDP가 국가의 행복을 반영할 수 있을까?

서론에서 본 팀은 GDP와 행복은 어느정도 상관관계가 있을 것으로 예상하였다. 하지만 아래의 시각화를 참고하였을 때 본 팀의 귀무가설이 옳지 않음을 알 수 있었다. GDP Top10과 Bottom10을 비교하였을 때 Top10의 경우 GDP가 Top5인 영국만이 행복 점수가 10위로 간신히 Top10에 포함되었고, 나머지 국가는 전부 다르다는 것을 확인할 수 있다. Bottom10도 마찬가지의 결과로 나와 GDP가 국가의 행복을 반영할 수 없다는 것이 확인되었다.

그림3. GDP Top10, Bottom10 그래프 시각화

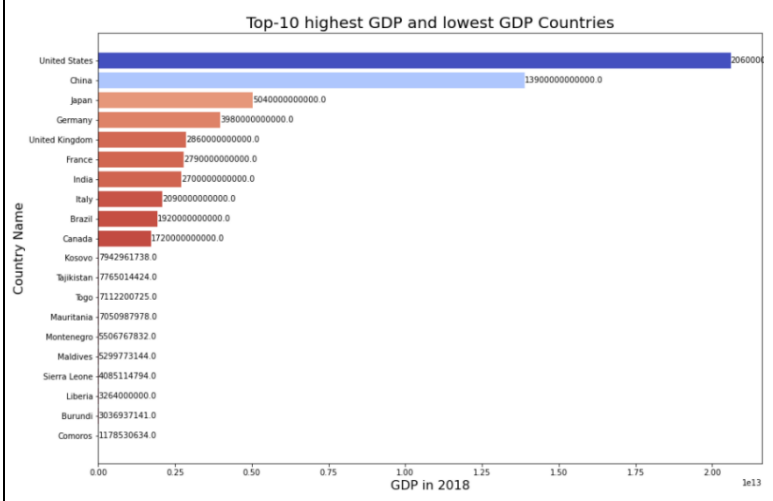


그림4. Ladder Score(행복 점수) Top10, Bottom10 그래프 시각화

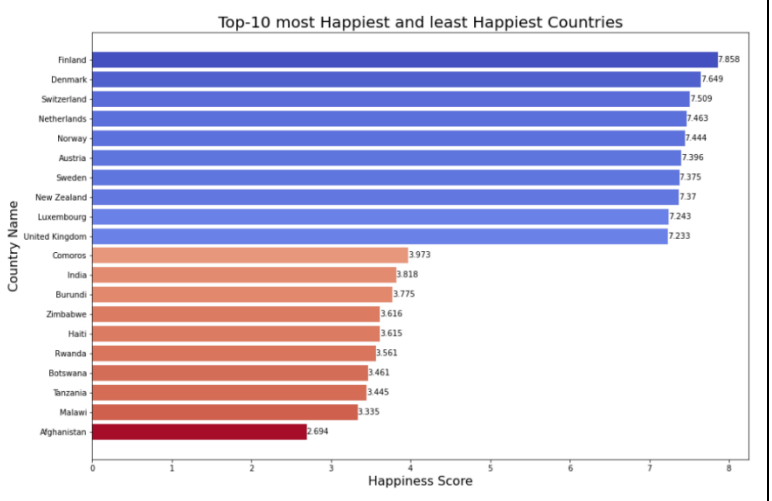


그림5. GDP와 행복 점수의 Scatter Plot

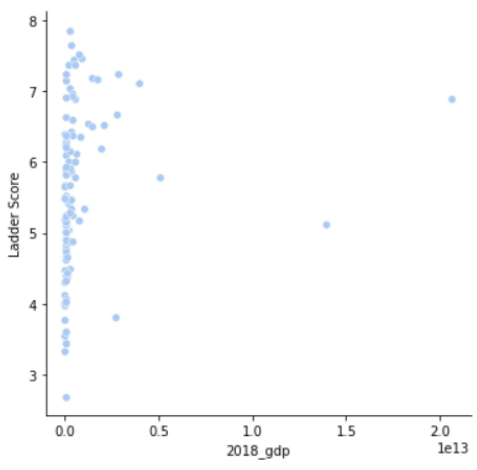
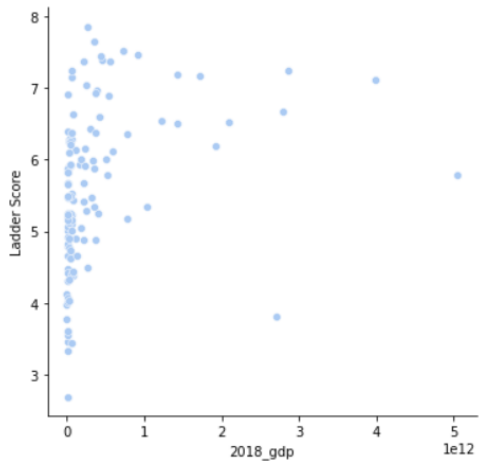


그림6. GDP가 비교적 매우 높은 미국, 중국 제거



위의 그림5는 GDP를 x축에, 행복점수를 y축에 두고 시각화 한 Scatter plot이다. Scatter plot으로 확인해 보아도 해당 두 변수의 상관관계는 확인하기 힘들다. 그림6은 다른 국가들에 비해 GDP가 이상치라 생각될 만큼 높은 미국과 중국을 제거한 후 시각화 한 Scatter plot이다. 마찬가지로 두 변수간 상관관계가 높다고 볼 수 없는 시각화 결과이다.

3. 결론

3.1 결과 분석

본론에서 분석한 결과와 아래 plotly를 이용하여 세계지도를 GDP와 행복점수 지도를 시각화한 아래 그림에 의해, "GDP가 높은 국가일수록 행복할 것"이라는 본 팀의 귀무가설은 틀렸다는 것을 알 수 있었다. GDP는 국가의 경제 척도이지만 빈부격차, 기타 외부 조건, 여가 및 가사노동 등을 모두 고려하지 못한다는 한계점이 있다는 결론을 내렸다. 이에 따라 본 팀은 우리의 데이터를 활용하여 새로운 대안을 제시하고자 한다.

그림7. Plotly를 활용한 GDP 지도 시각화

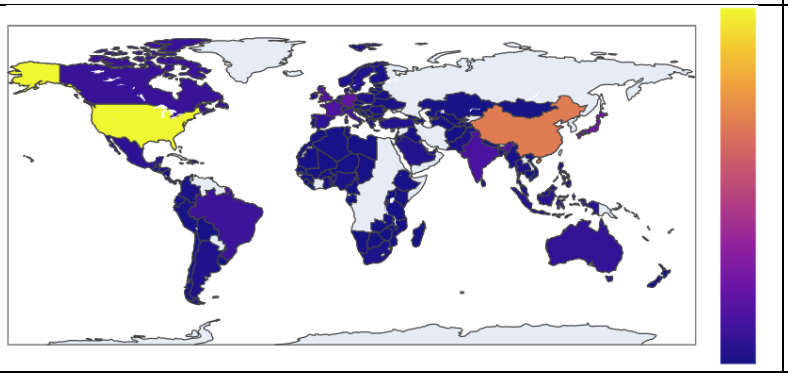
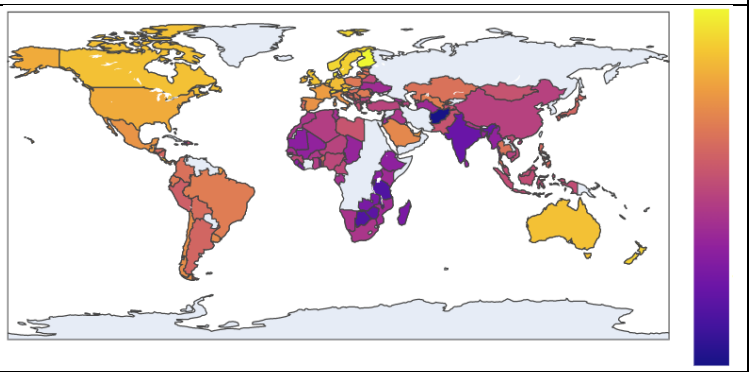


그림8. Plotly를 활용한 Ladder Score(행복점수)지도 시각화



3.2 대안 제시

3.2.1 대안 모색

행복지수(Happy Planet Index)란 영국의 신경재단이 도입한 국내 총생산과 국민의 삶의 만족도 등 여러 무형의 기준을 합하여 인간의 행복과 삶의 질을 포괄적으로 고려하여 수치로 표현한 지표이다. 행복지수를 구한 공식은 다음과 같다. "행복지수 = 웰빙 * 기대수명 * 부의평등/생태발자국" 여기서 생태 발자국이란 인간이 자연에 살면서 남긴 영향을 나타낸다. 생태 발자국이 클수록 지구에 악영향을 남긴 것으로 계산된다. 생태발자국이 중요한 이유는 인간의 지속가능성을 대변하기 때문이다. 환경을 고려하지 않는 무분별한 개발은 인류의 미래를 위협할 수 있어, 생태 발자국이 크면 그만큼 경제의 지속가능성이 떨어진다. 본 팀은 위 행복지수를 이용하여 우리의 데이터 프레임에서 사용할 수 있는 데이터로 치환하여 국가 행복을 나타내는 대안을 제시하고자 한다.

3.2.2 대안 제시 과정

대안 제시를 위해 'OECD.stat'의 OECD 국가의 Well-being과 관련된 다양한 지표들을 나타낸 데이터를 사용하였다. 기존 데이터 프레임의 경우 null값이 다수 존재하여 2010년 이후 데이터들의 평균 값으로 해당 국가의 지표 별 value값을 대체하였다. 그 후 국가를 x축, 웰빙 지표들을 y축에 오도록 데이터프레임을 수정하였다. 각 지표별로 단위가 다르므로, 모든 데이터를 0과 1사이로 정규화를 진행하여 각 지표가 동일한 기준에서 고려될 수 있도록 하였다. 또한 행복과는 크게 상관없는 지표들을 제거하고, 긍정적인 지표는 더하고 부정적인 지표는 빼서, 국가별 최종 well-being 총점을 구하였다.

행복지수(Happy Planet Index) 를 구하는 공식인 "행복지수 = 웰빙 * 기대수명 * 부의평등/생태발자국"에 웰빙은 우리가 구한 well-being 총점을, 기대수명엔 The World Bank의 데이터를, 부의평등엔 GDP, 생태발자국엔 CO2 배출량을 대입하여 본 팀에서 제시하는 데이터 분석을 통한 새로운 행복지수를 도출해냈다.

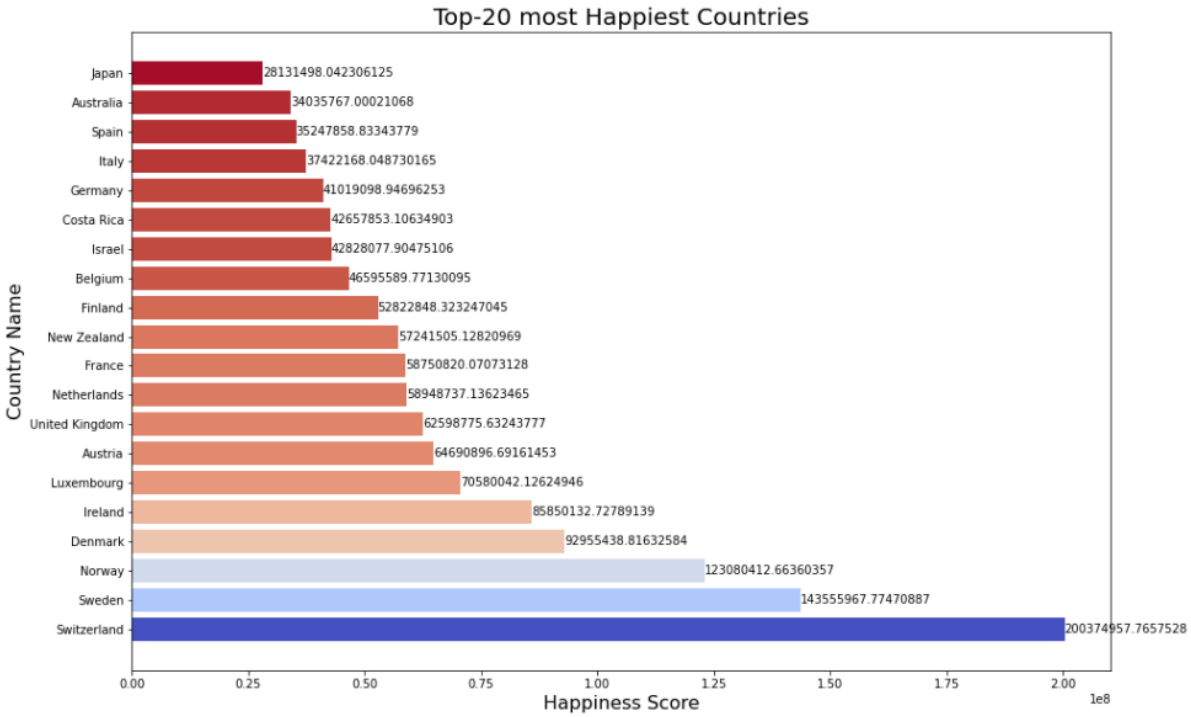


그림9. 새롭게 제시한 행복지수를 기준으로 구한 행복한 나라 Top20

위 그림9는 본 팀이 새롭게 제시한 행복지수를 기준으로 구한 행복한 나라 Top20를 시각화한 그래프이다. 스위스-스웨덴-노르웨이-덴마크-아일랜드-룩셈부르크 순으로 행복지수가 높은 것으로 확인되었다.

3.3 최종 결과

본 팀은 GDP가 어느정도 행복과 상관관계가 있을 것으로 가정 하고 데이터 분석 및 시각화를 진행하였다. 하지만 시각화를 통해 GDP는 국가의 경제를 나타내는 일종의 지표일 뿐 행복도와는 매우 상관관계가 낮고, 그 밖에 다양한 지표들과도 CO2 배출량을 제외하고는 상관관계가 없다는 결론을 얻어 귀무가설을 기각하였다. 이에 따라 본 팀은 GDP를 이용하여 새로운 행복 지수를 제안하고자 하였고, 직접 수집하고 전처리한 데이터를 활용해 유의미한 결과를 도출할 수 있었다. 아래의 그림10은 기존 Total 데이터프레임에서 Ladder Score를 기준으로 Top20 국가를 시각화한 그래프이다. 새롭게 제시한 행복지수Top20과 비교하였을 때 20개 중 16개 국가가 서로 겹쳐 80%가 일치한다는 결론을 내었다. 지금까지의 데이터 분석 및 시각화를 통해 한 국가의 행복을 반영할 수 있는 지표인 행복지수를 GDP의 대안으로 제시하고자 한다.

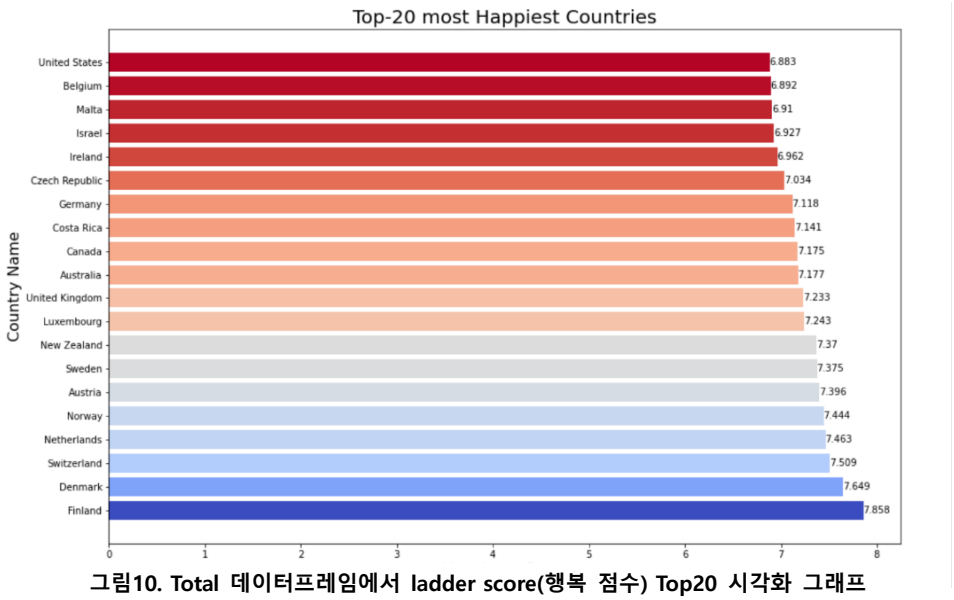


그림10. Total 데이터프레임에서 ladder score(행복 점수) Top20 시각화 그래프

3.4 기대효과 및 한계점

3.4.1 의의 및 기대효과

GDP는 국가 경제를 나타내지만, 국가 성장 및 국민의 행복도에는 크게 반영되지 않는다는 것을 데이터 시각화를 통해 확인하였다. 이에 따라 GDP와 Ladder Score를 비롯한 다양한 변수들을 고려한 새로운 대안을 제시함으로써 GDP로는 고려하지 못한 국민의 삶까지 나타내었다.

3.4.2 한계점 및 추후 계획

The World Bank 데이터에 모든 국가가 포함되어 있지 않았고, 최근 연도의 데이터는 결측치가 매우 많았다. 또한 행복지수 계산시 사용한 well-being 데이터의 경우 OECD국가만 포함된 데이터 프레임이었으며, 지표 설정에 주관성이 들어갔다는 한계점이 존재했다. 추후 결측치가 아닌 완벽한 데이터와 OECD국가 외에 다른 국가의 데이터를 이용하고, 더욱 면밀한 지표 설정을 하게 된다면 더욱 발전된 데이터 분석 및 시각화 프로젝트로 확장 할 수 있을 것이다.