다양한 시각으로 바라본 전세계 여성의 삶

-폭력, 교육, 자살, 출산, 사회 진출, 경제 지표를 중심으로-

과목	데이터분석및시각화
지도교수	이현봉 교수님
팀원 정보	정보통신공학과 2018112168 권나영 통계학과 2019113421 최은진
제출일	2021.11.29

주제 선정 이유

올해 중순 아프가니스탄 내전 발생로 인해 특별 기여자 체류를 허용한다는 뉴스를 보았습니다. 그 때 아프가니스탄이란 나라에 대해 찾아봤을 때 그 나라의 여성들은 남성 동반자 없이 외출이 불가능하고 교육과 경제 활동도 금지된 상태라는 것을 알게 되었습니다. 여성이 최초로 선거 참가권을 얻은 지 약 120년인데, 아직도 주체적으로 자신의 삶을 살 수 없는 여성들이 많다는 것에 안타까움을 느끼며 다양한 지표들을 이용해 전세계 여성의 삶을 분석했습니다.

Issue1. 여성이 남편의 폭력이 정당하다고 생각하는 것은 교육 수준과 관계가 있을까?

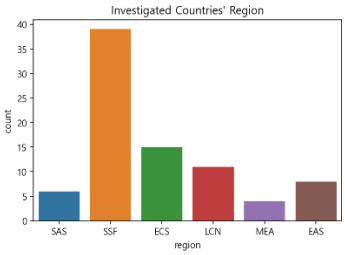
- 사용한 지표

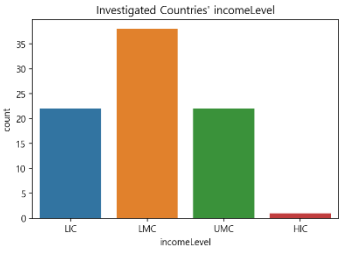
지표 코드	설명
SG.VAW.ARGU.ZS	'남편과 논쟁 시 맞아도 된다'고 믿는 여자의 비율
SG.VAW.BURN.ZS	'요리를 태웠을 때 남편에게 맞아도 된다'고 믿는 여자의 비율
SG.VAW.GOES.ZS	'말 없이 외출 시 남편에게 맞아도 된다.' 고 믿는 여자의 비율
SG.VAW.NEGL.ZS	'아이들을 무시하면 남편에게 맞아도 된다'고 믿는 여자의 비율
SG.VAW.REFU.ZS	'성관계를 거부하면 남편에게 맞아도 된다'고 믿는 여자의 비율
SE.TER.CUAT.BA.MA.ZS	25 세 이상 남성 중 '학사' 학위 이상을 가진 남자의 비율
SE.TER.CUAT.BA.FE.ZS	25 세 이상 여성 중 '학사' 학위 이상을 가진 여자의 비율

- 전처리 과정

각 지표에 관한 데이터를 wbgapi 를 이용해 DataFrame 형식으로 받아왔습니다. 이 때, 년도 열이 'YR2020'처럼 string 형으로 들어오는데, 그냥 두면 시각화 시 불편한 부분이 있었습니다. 이에 gedaTools.py 안에 timeToInt 라는 lambda 함수 기반 메서드를 만들어 년도 열을 int64 형으로 변환해주는 전처리를 해주었습니다. 폭력의 정당화에 관한 데이터는 결측치가 예상보다 많아서, 나라와 년도를 인덱스로 설정하였을 때, 겨우 196 행에 그쳤습니다. 원래는 결측치가 하나라도 있는 행은 모두 제거할 계획이었으나, 그렇게 되면 데이터의 수가 너무 줄고, 대체하는 방법도 오히려 오차를 증가시킬 것 같아서, 존재하는 데이터에 대해서만 처리하였습니다. 이렇게 예상과의 차이점이 있는 것처럼, 진행 과정에서 초반의 계획보다도 더 다양한 데이터들과 결합해보며, 해당 데이터가 보다 재미있는 상관관계를 갖는 데이터가 무엇인지 찾아보기 위해 기존에 기획된 교육 수준 데이터뿐만 아니라 대륙이나 소득 수준 데이터도 함께 살펴보았습니다. 이러한 대륙이나 소득수준을 알아보기 위해서는 편리하게 라이브러리를 활용했는데, wbjapi 의 economy 데이터 베이스에서 원하는 나라들의 대륙이나 소득수준을 가져와서 사용했습니다. 다양한 관계성을 살펴보기 위해 데이터 프레임을 자체적으로 변형함에 멈추지 않고 피봇 테이블 함수를 이용해 그룹 값을 구하기도 하고, 특정 값을 기준으로 두 데이터 프레임을 합치고 싶을 때, 데이터 프레임들끼리의 Join 을 구현하기 위해 merge 를 사용하는 등 특정 시각화 상황에 따라 필요한 전처리 과정을 거쳤습니다.

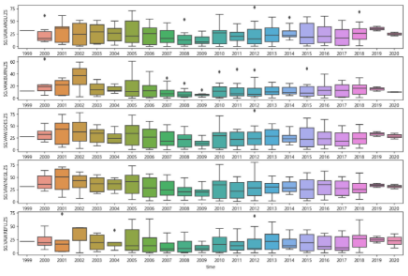
- EDA

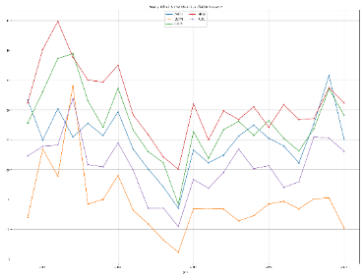




1 차 : 데이터의 양이 국가별로 표현하기에는 적다는 생각이 들어, 대륙이나 소득 수준으로 묶어 상관관계를 나타내면 더 유의미해질 것이라는 생각에 count plot 를 이용해 대륙별, 소득 수준별로 막대 그래프로 나타냈습니다.

결과 : 사하라 사남 이남의 아프리카 대륙과 중하위소득의 데이터가 가장 많았습니다. 조사 대상들을 한눈에 살펴보자, 대륙별 추이나 상관관계가 더 궁금해졌고, 이를 중점적으로 시각화하기로 하였습니다.





2 차 : 남편의 폭력을 정당하다고 생각하는 여성 비율의 연도별 추이를 보기 위해 전체 데이터에 대한 boxplot 으로 확인한 결과 이상치가 많았습니다. 그래서 각 폭력 이유별로 평균을 내어 년도에 따라 꺾은선그래프로 나타냈습니다.

결과 : 00 년대에 가장 높은 비율을 보였지만, 현대로 오면서 크게 줄어들지 않았습니다

3 차 - 1 : 대륙별로 5 가지 폭력 정당화 데이터의 평균을 구한 후, 수치가 다르기에 평균을 각 데이터의 최대값으로 나누어 상관계수를 구해서 heat map 을 나타냈습니다.

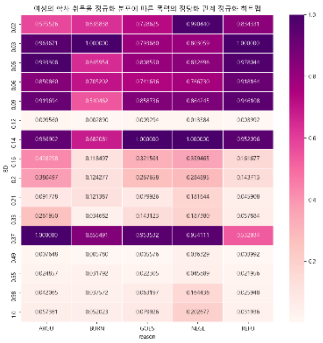
결과 : 남편의 폭력 이유에 대해 어떤 이유에서든 중남미 대륙의 폭력 정당화 수치가 가장 낮았고, 중동과 북아프리카에서 성관계 거부 시 폭력에 대해서만 수긍하는 편이었습니다.

3 차 - 2 : 남편의 폭력 이유별로 정당화(수긍)의 수치를 정렬한 후, 상위 10 개만 뽑아낸 데이터프레임을 가지고 국가, 나라별 마크를 다르게 하여 scatter plot 로 나타냈습니다.

결과 : 최대값은 모두 사하라 사막 서남 아프리카 대륙에서 나왔고, 특히 기니 국가가 최대값을 가지고 있었습니다. 전체 기간의 상위값 10 개가 현대에도 계속 나오는 것을 보아 폭력을 아직도 수용하는 여성들이 많은 것으로 보입니다.

4 차, 5 차 : 폭력에 대해 조사된 나라에 대해서만 대륙별 남녀 학사 비율을 피봇 테이블을 함수를 이용해 대륙별 평균값을 구하고, 대륙별로 비교하기 위해서 절대적인 남녀 학사 비율보다 상대적인 학사 취득율이 더 차이를 극대화 시킬 수 있을 것이라고 생각하여 평균값의 최대값을 1 로 스케일링 후 히트맵으로 나타냈습니다. 그리고 전세계 인구에 대해 학사 이상 취득율이 시간의 흐름에 따라 어떻게 되는지 궁금하여 년도별로 평균을 내어 막대그래프로 나타냈습니다.

결과 : 북미 지역의 남녀 학사 취득률이 가장 높으며, 사하라 서남 아프리카 대륙의 학사 취득률이 남녀 모두 가장 낮았습니다. 그리고 남녀 모두 최근에 올수록 학사 취득율이 높아졌고 여성의 학사 취득률도 매우 증가하고 있습니다. 북미 지역의 학사 이상 취득률을 남녀 각각 1 로 두었음에도 사하라 서남 아프리카의 학사 취득률이 남자는 북미의 12%, 여자는 북미의 5% 수준으로 매우 큰 차이를 볼 수 있었습니다.



6 차 : 폭력에 관한 데이터가 존재하는 국가, 연도에 대한 여성의 학사 취득율 데이터를 매칭하여 가져온 후 폭력의 정당화에 대한 관계와 상관관계를 보기 위해 히트맵으로 나타냈습니다

결과 : 예상처럼 여성의 학사 이상 학위 취득률과 폭력의 정당화는 반대 관계였습니다. 남성의 학사 취득율이 오히려 예상이 안 되는 부분이었는데, 여성의 결과와 비슷하였으나, 오히려 학사 취득율이 가장 높을 때 예외 현상도 있었습니다. 결론적으로, 한 나라의 교육 수준 자체가 여성 자체의 인권 의식이나 자존감을 결정 짓는데 어느 정도의 영향을 미칠 수도 있겠다 하는 생각을 갖게 되었습니다.

Issue2. 폭력의 정당화와 여성의 자살 수가 관계가 있을까?

• 사용한 지표 : Issue1 폭력의 정당화 데이터 + 아래의 추가 데이터

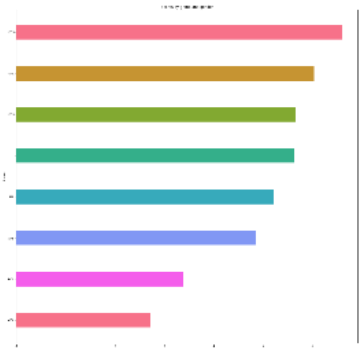
지표 코드	설명
SH.STA.SUIC.FE.P5	여성 100,000 명당 자살수

• 전처리 과정

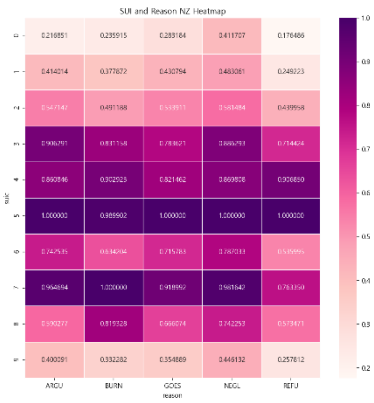
Issue1 과 동일한 방법으로 5 가지 이유에 대한 남편의 폭력 정당화 데이터를 가져왔습니다. 여성의 자살 수 데이터는 폭력 데이터가 존재하는 국가에 대해서만 가져온 후 대륙별 자살 수 평균 값 계산하여

사용하기도 하고, 폭력 데이터와 일치하는 년도에 대해서만 남겨 매칭 후 나머지는 제거한 후 사용하는 두가지 방법으로 활용했습니다.

• EDA



1 차 : 대륙별 10 만명당 여성 자살 수를 subplot 함수를 이용해 여러 개의 막대그래프를 이용해 나타냈습니다.
결과 : 자살 수 평균이 유럽과 중앙 아시아에서 가장 높고 중동과 북아프리카가 가장 낮았고 그 수는 2 배나 차이납니다.



2 차 : 5 가지 이유에 대한 남편의 폭력을 정당하다고 생각하는 여성의 비율과 자살 수의 상관관계를 보기 위해 자살수를 10 단계로 나누어 히트맵으로 나타냈습니다.
결과 : 남편의 폭력에 대한 정당화가 높을수록, 여성의 자존감 부재와 문화·도덕적 지식 습득의 기회 결여를 동시에 겪을 것이라 예상하며 이러한 환경이 여성의 자살 수를 높게 할 것이라 생각했지만 오히려 유럽, 중앙 아시아의 자살 수가 높은 결과를 보며 유의미한 상관관계가 없다는 결론을 도출할 수 있었습니다. 여성의 자살 수가 높은 대륙에 대하여 자살 수와 비례하는 지표가 무엇이 있는지, 인과 관계가 어디에 있는지 기회가 된다면 더 분석하고 싶습니다.

Issue3. 청소년 출산 수와 전체 여성의 출산 수는 상관 관계가 있을까?

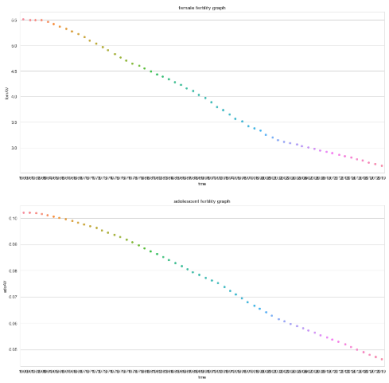
• 사용한 지표

지표 코드	설명
SP.ADO.TFRT	15-19 세 여성의 1000 명이 출산하는 아이의 수
SP.DYN.TFRT.IN	전체 나이의 여성 1 명이 출산하는 아이의 수

• 전처리 과정

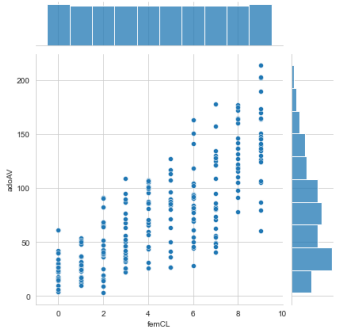
두개의 지표에 대한 데이터를 전체 값이 없는 행은 제외하고 데이터프레임 형식으로 각각 저장했습니다. 두 데이터프레임 결합 후 컬럼명을 지정하고 인덱스 초기화 후 전세계 두가지 출산 수에 대해 년도별로 평균을 계산했습니다. 청소년의 출산 수는 전체 나이 여성 1 명의 출산 아이 수와 비교할 수 있도록 1000 으로 나누었습니다. 또한 국가별 청소년과 전체 나이의 여성 출산 수의 모든 기간에 대하 평균을 구하고 그래프의 해석 편리를 위해 총 여성 출산 수를 10 단계로 범주화 하여 청소년 출산 수와 비교했습니다.

• EDA



1 차 : 모든 연령에서 여성 1 명이 출산하는 아이의 수와 청소년 1 명이 출산하는 아이의 수의 시계열 분석을 위해 모든 나라의 평균을 구하고 연도별로 strip plot 을 나타냈습니다.
결과 : 여성 1 명당, 청소년 1 명당 모두 연도에 반비례하여 출산 수가 떨어지고 있습니다. 데이터가 수집된 초기(60 년대)와 비교하여 현대 여성의 출산 수는 절반입니다. 전세계 인구는 아직도 폭발적으로 증가 중인데 출산 수는 확연히 감소하고 있는 것을 보아 전세계 인구의 증가는 의학의 발전으로 수명이 길어졌기 때문이라는 결론을 낼 수 있었습니다.

2 차 : 전체 연령의 여성 출산 수를 10 단계로 나타내어 청소년 출산 수를 비교하기 위해 joint plot 과 heat map 으로 비교해보았습니다.
결과 : 여성 1 명의 출산 수가 높은 나라는 청소년 출산 수도 높은 편인 것을 보아 양의 상관 관계를 가질 것이라는 예상과 일치했습니다.



Issue4. 청소년의 출산율과 여아의 취학률은 관계가 있을까?

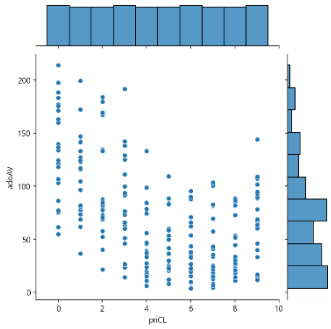
• 사용한 지표

지표 코드	설명
SP.ADO.TFRT	15-19 세 여성의 1000 명이 출산하는 아이의 수
SE.ENR.PRSC.FM.ZS	초등 및 중등 교육에서 남아에 대한 여아 비율 (%)

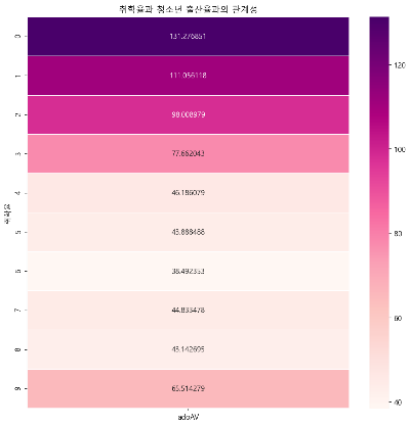
• 전처리 과정

전체 값이 없는 행은 제외하고 DataFrame 형식으로 저장. 국가별 취학률과 청소년의 출산 수 평균을 계산하고 그래프 해석의 편리성을 위해 여아의 취학률은 10 단계로 범주화 한 후, 청소년 출산 수와 비교해보았습니다.

• EDA



1 차 : 여아의 취학률과 청소년의 출산수를 joint plot 으로 상관관계 시각화하였습니다.
결과 : 대체적으로 넓게 분포되어 있지만 청소년 출산 수가 높으면 취학률이 낮은 편이고, 취학률이 높으면 청소년 출산 수는 낮은 편입니다.



2 차: 청소년 출산 수와 여아의 취학률을 heat map 으로 비교
결과 : 취학률이 낮으면 출산 수가 높은 것은 확실합니다. 하지만 취학률이 가장 높았을 때 청소년 출산 수가 낮지 많은 않은 의외의 경우도 있었습니다. 출산 때문에 학업을 연장하지 못하는 여아들이 많을 것이라 생각했는데 의외의 결과였습니다. 취학률은 초등학교, 중학교 기준이고 청소년 출산 수는 15-19 세 기준이라 우리나라 기준으로 생각하면 고등학교 진학을 포함하지 못한 것으로 보입니다. 고등학교 취학률을 포함한다면 다른 결과가 나왔을 수도 있고, 교육을 많이 받을 수 있는 국가는 복지 제도가 잘 마련되어 있어 출산 여부에 관계없이 여아가 배움을 지속할 수 있도록 지원받을 수 있겠다는 예상도 해보았습니다.

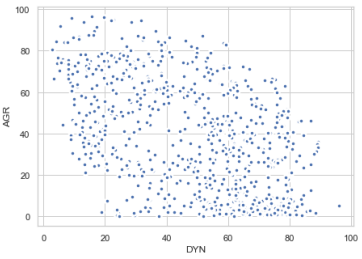
Issue5. 여아의 피임보급률과 여성의 사회 진출이 관계가 있을까?

지표 코드	설명
SP.DYN.CONU.ZS	15 세 – 49 세 사이 모든 방법을 사용한 여성의 피임 보급율
SG.GEN.PARL.ZS	전체 국회의원 중 여성 의석 비율(%)
SL.AGR.EMPL.FE.ZS	모든 여성 노동자 중 농업계 종사율(%)
SL.IND.EMPL.FE.ZS	모든 여성 노동자 중 산업계 종사율(%)
SL.SRV.EMPL.FE.ZS	모든 여성 노동자 중 서비스업계 종사율(%)

• 전처리 과정

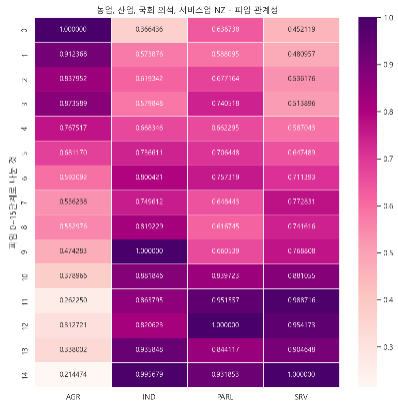
피임 보급율 및 4가지의 여성 고용률 지표들을 데이터 프레임에 담아, 1차 시각화에서는 필요한 열만 잘라 사용하였고, 2 차 시각화에서는 피임 보급율을 기준으로 전체 데이터를 동일한 수량을 갖을 수 있도록 15 개로 쪼개어 분위수를 구한 후, 해당 분위수를 인덱스로 하는 피봇 테이블을 만들었습니다.

• EDA



1 차 : 각 분야의 15 세 ~ 49 세 사이 피임 보급율을 x 축, 각 분야별 여성 고용률을 y 축에 놓고 총 4 개의 scatter plot 으로 표현했습니다.
결과 : 각 분야별로 그 분포의 모양이 모두 다르지만, 농업만 눈에 띄게 피임 보급률이 낮을수록 종사율이 높습니다. 나머지 분야는 피임 보급률이 높아짐에 따라 점점 해당 분야의 비율이 높아집니다. 다시 말해, 피임 보급율이 증가함에 따라 여성의 농업 종사 비율은 줄고, 국회나 산업계, 서비스계의 종사 비율이 높아짐을 알 수 있었습니다. 농업에 종사하면 육체 노동 강도가 높아서 임신할 경우 일을 하기가 쉽지 않을텐데 피임률이 낮은 이유가 궁금했습니다.

2 차 : 피임 보급률 15 단계를 기준으로 각 고용률 데이터들의 평균값을 피봇 테이블로 만든 후, 이를 히트맵으로 시각화 하였습니다.



결과 : 피임률이 증가할수록 농업 종사 비율은 줄고, 국회·산업·서비스직 종사율은 증가하는 것을 확인할 수 있었습니다. 여성이 국회의원을 차지하는 비율은 전체 의원 중 여자가 차지하는 비율인 반면 나머지 고용률에 대한 데이터는 전체 여성 노동자 수 중 해당 분야 종사율을 모델링한 추정치이기 때문에, 해당 분야에서 일하고 있는 전체 노동자 수 중 여성이 차지하는 비율이었으면 더 정확하게 비교할 수 있었을 것이라는 아쉬움이 남습니다.

Issue6. 여성의 사회 진출과 국가의 경제 발전은 상관관계가 있을까?

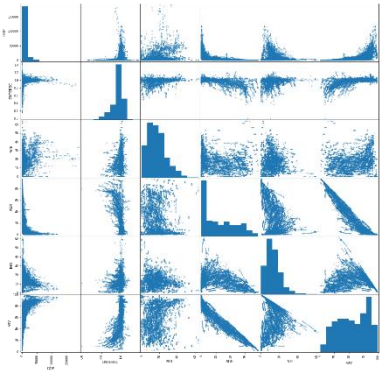
지표 코드	설명
SE.ENR.PRSC.FM.ZS	초등 및 중등 교육에서 남아에 대한 여아의 비율(%)
SP.DYN.CONU.ZS	15-49 세 여성의 모든 방법의 피임 보급률(%)
SG.GEN.PARL.ZS	전체 국회의원 중 여성 의석 비율(%)
SL.AGR.EMPL.FE.ZS	모든 여성 노동자 중 농업계 종사율(%)
SL.IND.EMPL.FE.ZS	모든 여성 노동자 중 산업계 종사율(%)
SL.SRV.EMPL.FE.ZS	모든 여성 노동자 중 서비스업계 종사율(%)
NY.GDP.PCAP.CD	1 인당 GDP(US$)

• 전처리 과정

처음에는 NY.GDP.MKTP.CD GDP 지표의 데이터를 이용하려고 했는데 이 지표는 1 년동안 한 국가에서 생산된 재화와 용역의 시장 가치를 합한 것으로 인구수가 많은 나라는 아무래도 GDP 가 클 것이라는 생각이 들었습니다. 그래서 NY.GDP.PCAP.CD (한 나라에서 한 해 동안 생산된 모든 최종 재화와 서비스의 가치를 그 해의 평균 인구로 나눈 값)를 사용했습니다.

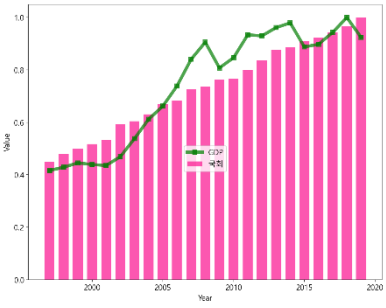
1인당 GDP와 여성 피임 보급률 및 고용률 지표를 한 데이터 프레임에 담아, 결측값이 많기 때문에 7가지 지표에 대한 값이 모두 채워진 국가, 연도 행만 남기고 삭제 후 이 상태 그대로 사용하거나, 년도별 각 지표에 대해 평균을 계산하고 최대값을 1 로 스케일링 하고, 1 인당 GDP 를 15 단계로 범주화 하여 사용하기도 했습니다.

• EDA

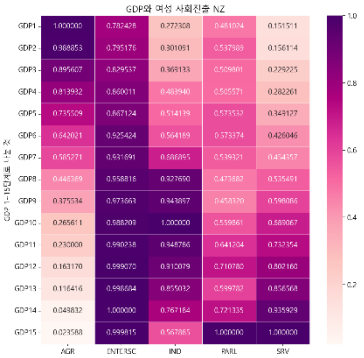


1 차 : 각 지표 간 관계를 보기 위해 scatter matrix 함수를 이용해 산점도 행렬로 시각화 했습니다.
결과 : 1 인당 GDP 를 y 축으로, 다른 요소들을 x 축으로 나타낸 그래프들을 보았을 때, 농업, 산업 종사율이 낮은 곳에 GDP 가 큰 값들로 분포하는 것으로 보아, 대체적으로 두 값이 반비례할 것으로 예상할 수 있었고, 여성 의석 비율은 중간 쯤에서 높은 값이 등장하고, 오히려 의석율이 큰 경우가 없는 것으로 보아, 어느정도 비례하다가 반대 양상을 갖는 느낌이 들었습니다. 마지막으로, 여아 취학율이나 서비스 종사율과는 비례하는 양상으로 판단하였습니다. 이 외의 지표들 간 관계들도 흥미로웠습니다.

2 차 : 연도를 기준으로 평균을 계산한 피봇 테이블을 만든 후 모든 값을 최대값이 1 이 되도록 스케일링 한 후, 증가율을 뚜렷하게 보기 위해 1 인당 GDP 에만 2 를 곱한 후 다른 지표는 bar plot 으로, 1 인당 GDP 는 꺾은선으로 그려 연도별 추이를 시각화하였습니다.



결과 : 시간의 흐름에 따라 각 지표 값의 변화를 분석한 결과 1 인당 GDP 의 평균은 거의 매년 증가하는 양상을 보이며, 2019 년에 최대값을 갖는 것을 알 수 있었습니다. 농업과 산업을 제외하고는 모두 증가하고 있으며, 여성의 취학률은 미세하지만 꾸준히 높아졌으며 국회 여성 의석 비율이 가파르게 상승하고 있는 모습을 볼 수 있어습니다. 20 년간 2 배나 상승했습니다.



3 차 : 1 인당 GDP 를 1~15 단계로 나누고 여러 지표들을 최대값을 1 로 두고 스케일링 한 후, 히트 맵을 그려 시각화 해보았습니다.
결과 : 농업과 산업을 제외하고는, 1 인당 GDP 와 양의 상관관계를 가지고 있었으며 여성의 서비스업 종사율, 국회 의석 비율, 취학률 순서로 증가하고 있었습니다. 산업 고용률은 어느정도 1 인당 GDP 와 비례하지만, 1 인당 GDP 의 10 단계에서 최대값을 찍자 다시 감소하는 것을 알 수 있었습니다. 또한 농업 종사율은 확실하게 1 인당 GDP 가 증가함에 따라 감소하는 부분임을 알 수 있었습니다.

Conclusion. 다양한 시각으로 전세계 여성의 삶을 분석한 EDA 를 마치며

• 느낀 점

편견이나 막연한 생각들을 객관적인 데이터로 직접 시각화 한 후 검증할 수 있는 시간이었습니다.
결측치가 많아 아쉬운 점이 많았지만 존재하는 데이터로만 분석하는 것도 충분히 의미 있었습니다.
남편의 폭력을 정당화하는 데이터의 업데이트가 다음에 있을 때는, 남편의 폭력을 수용하는 여성의 수가 낮아질 수 있기를 간절히 바랍니다.

• 다음에 분석해 보고 싶은 주제

1) GDP 의 증가와 인플레이션의 관계
2) 자살수가 높은 국가를 선정하여 자살률과 양의 상관 관계를 가지는 지표는 무엇인지, 그리고 인과관계가 있는지 궁금합니다.
3) 전세계 여성의 삶에 대한 지표들을 바라보았으니 다음에는 한 국가, 특히 한국에 대해서 깊이 분석해보고 싶습니다. 우리나라의 경우 남아의 수와 비교하여 여아의 진학률이 높은 편인데 국회의원 수는 전세계 평균보다 적어서 시간이 지나면 높아질지 궁금합니다.