1. 정의

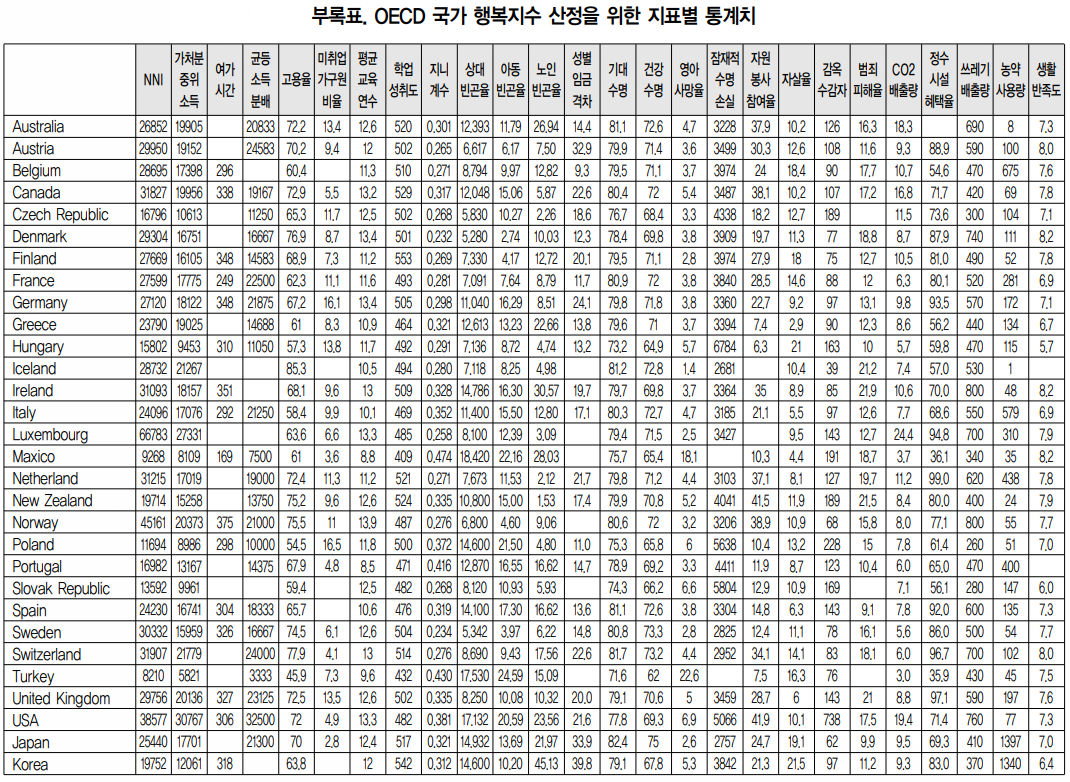
이번 최종과제를 통해 수업시간에 배운 내용들을 적용해보고, 의미 결과가 도출될 겨우, 그 결과에 대한 해설을 추가하는 것이다. 이번 과제를 통해 ‘국가별 행복지수’에 대한 통계자료를 통해 다양한 의미관계를 도출해 내보는 것을 목표로 하고 더 나아가 의미 관계를 통한 새로운 정보를 획득하는 것을 목표로 한다.

1. 데이터 수집

Kaggle 사이트를 통해 데이터를 수집할 계획이다.

1. 데이터 관리 방법 내용 설명

행복지수는 다음과 같은 자료들을 통해 종합적으로 지수를 산출해 낸다.



이번 과제를 통해 5년간 행복지수가 증가한 나라는 어떤 변화가 있었는지, 감소한 나라들은 어떤 변화가 있었는지 분석해보는 시간을 가질 계획이다.

3. 데이터 탐색

<https://www.kaggle.com/roshansharma/world-happiness-report?select=2019.csv>

연도별 국가 행복지수 통계자료 5건을 다운받아 2015-2019 진행   
4. 데이터 전처리

연도별 행복지수 데이터에는 다음과 같은 값들이 표기되어 있었다.

* Country (나라)
* Region (지역)
* Happiness Rank (행복 순위)
* Happiness Score (행복지수)
* Standard Error (표준 오류)
* Economy (GDP per Capita) (경제성 - 인구당 GDP)
* Family Health (Life Expectancy) (가족 건강)
* Freedom (자유도)
* Trust (Government Corruption) (신뢰도 – 정부 붕괴)
* Generosity (아량)
* Dystopia Residual (디스토피아 잔차)

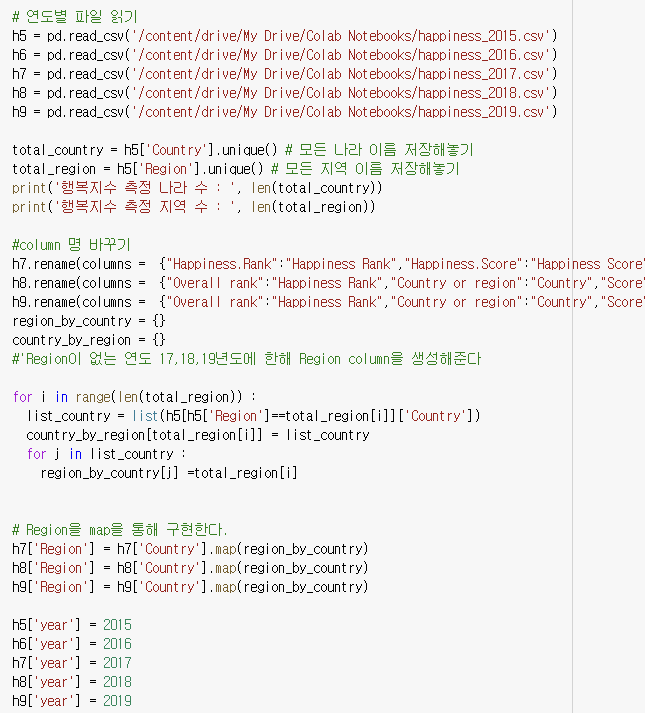
하지만, 연도별로 해당 column들이 생략되기도 하고 다른 명칭으로 불리기도 한다. 따라서 이를 2015년 column 기준으로 19년까지 공통된 값만 이용을 할 계획이며 이외의 column들은 분석에서 제외하였다.

공통된 컬럼 값

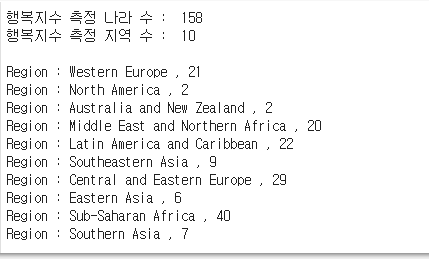
* Country
* Region(없어서 추가할 계획)
* Happiness Rank
* Happiness Score
* Economy (GDP per Capita)
* Family Health (Life Expectancy)
* Freedom
* Trust (Government Corruption)
* Generosity

5. 데이터 분석 방법 결정 및 결정 사유

데이터는 전부 수치화 되어있기 때문에, 컬럼 간의 관계와 지역별 행복지수에 대한 분석을 진행하고, 행복지수에 대한 다양한 분석을 진행할 계획이다. 연도별 수치를 나눠 보여줘야 하기 때문에 대부분 barplot을 이용하여 시각화 할 예정이다.

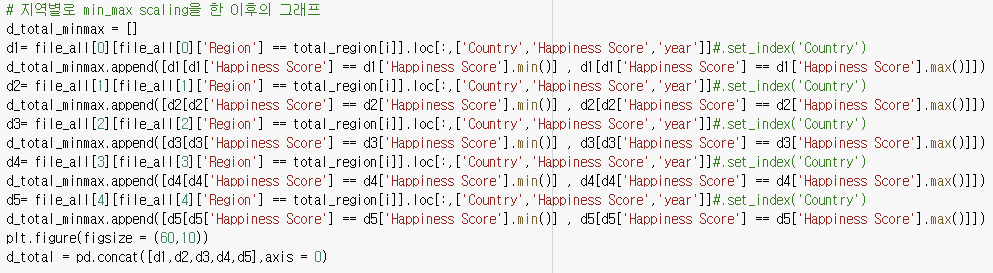
6. 데이터 분석 시행 (결정된 분석 방법에 따라서 진행) – 7. 데이터 시각화  


2015-2019년도 5개의 행복지수 파일들을 읽어 다양한 컬럼명을 일치시키며, year와 region 컬럼을 새로 정의해주고 시작한다.

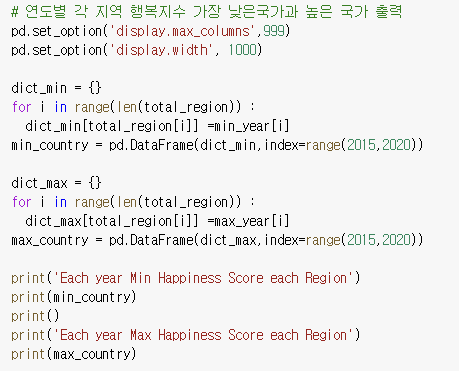


행복지수를 측정한 나라는 총 158 국가이며 이는 연도에 따라 증감한다.

측정지역은 10개 지역이며, 다음과 같은 지역들이 참여하였다.



각 지역별로 데이터를 그룹화하여 시각화하고, 각 지역별 최대, 최소 국가들을 연도별로 출력한다.



각 연도별로 지역별 최대, 최소 행복지수 나라를 한 번에 보여준다.

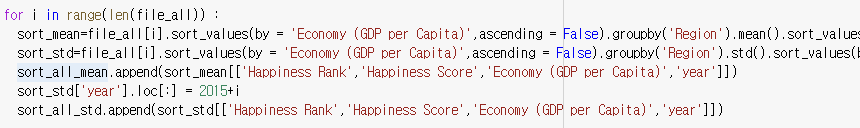


대개 지역별 최대 최소 행복지수 나라는 변하지 않고 일관적으로 나오는 것을 볼 수 있다. 하지만,

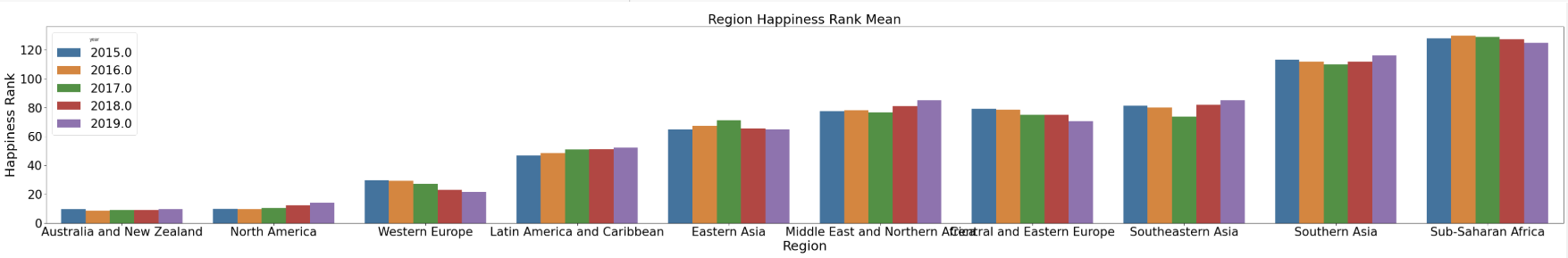
최소 행복지수 국가가 크게 바뀐 지역은 Western Europe과 Sub-Saharan Africa가 있고

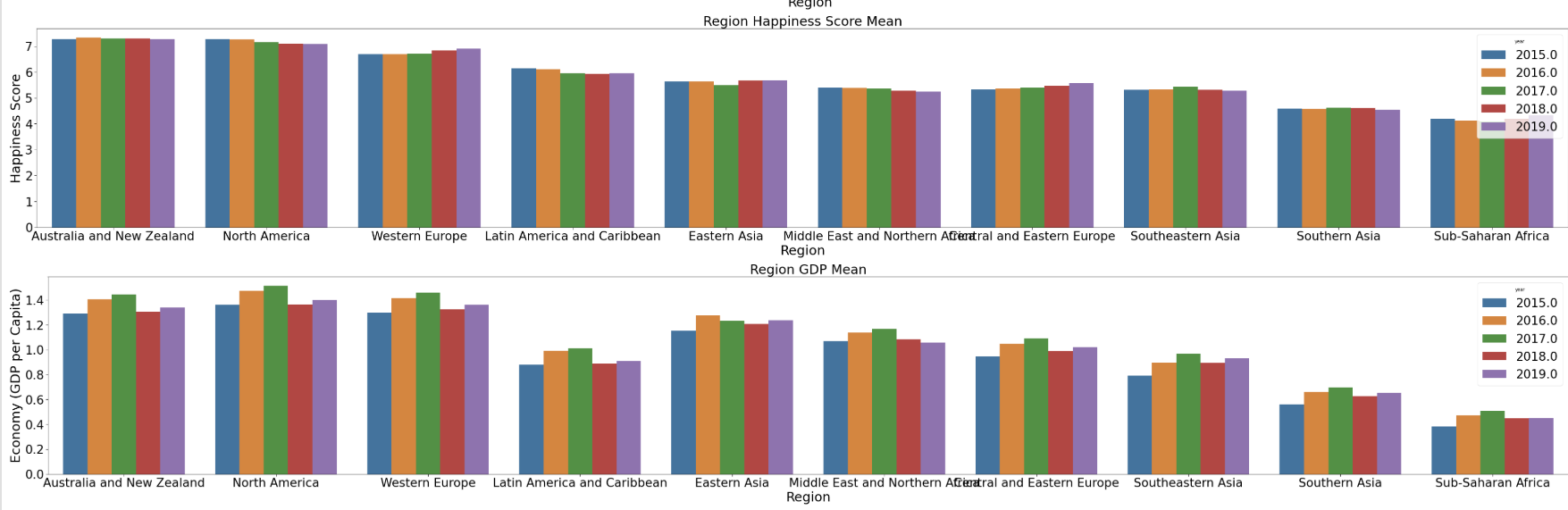
최대 행복지수 국가가 크게 바뀐 지역은 Western Europe이다.

이런 결과를 보아, Western Europe지역의 국가들의 순위변동이 연도별로 자주 일어난 것을 알 수 있다.



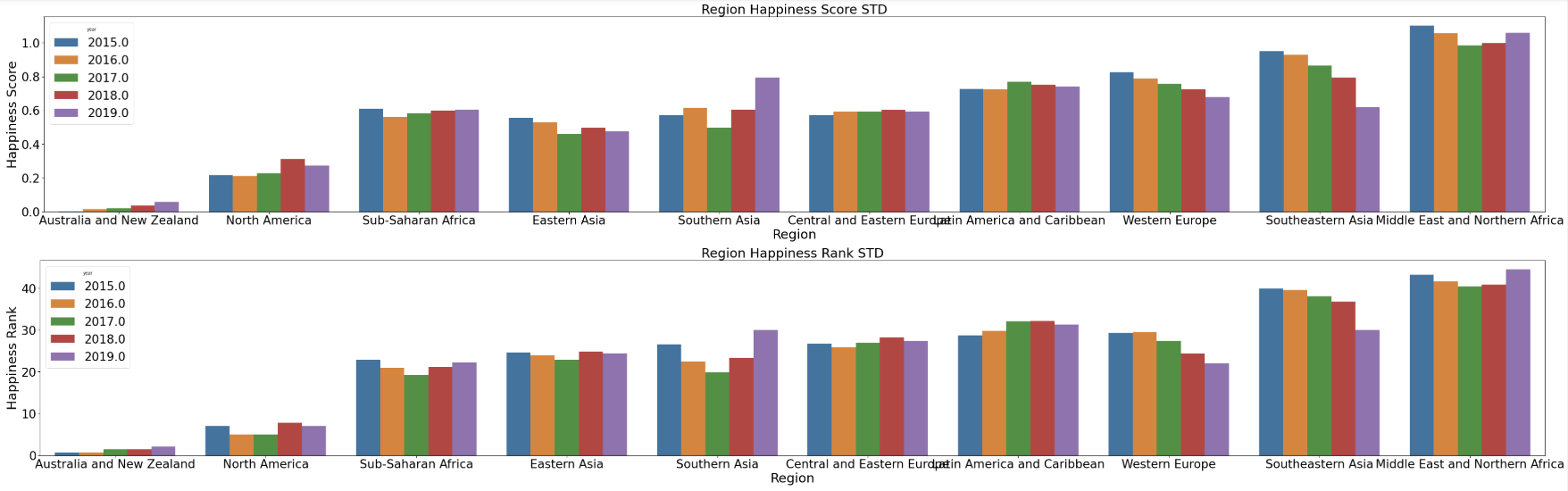
이후, 지역별로 그룹화하여 GDP 기준으로 내림차순 정렬을 진행하여 각 지역의 평균 행복 등수, 행복 지수, GDP 표준편차 행복 등수, 행복 지수, GDP에 대해 알아보았다.

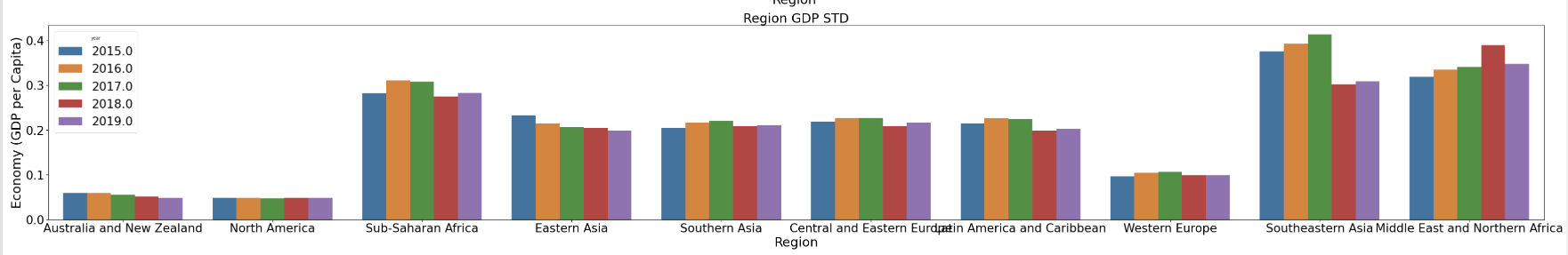




연도별로 행복 지수 및 순위와 GDP 평균을 내어 통계를 내 본 결과, 소득이 높을수록 행복지수가 보편적으로 높아졌고, Western Europe과 Central and Eastern Europe의 행복 지수는 계속 높아졌다.

하지만 특이 케이스로 Latin America and Caribbean의 GDP가 상대적으로 낮음에도 불구하고 행복 순위가 높았다



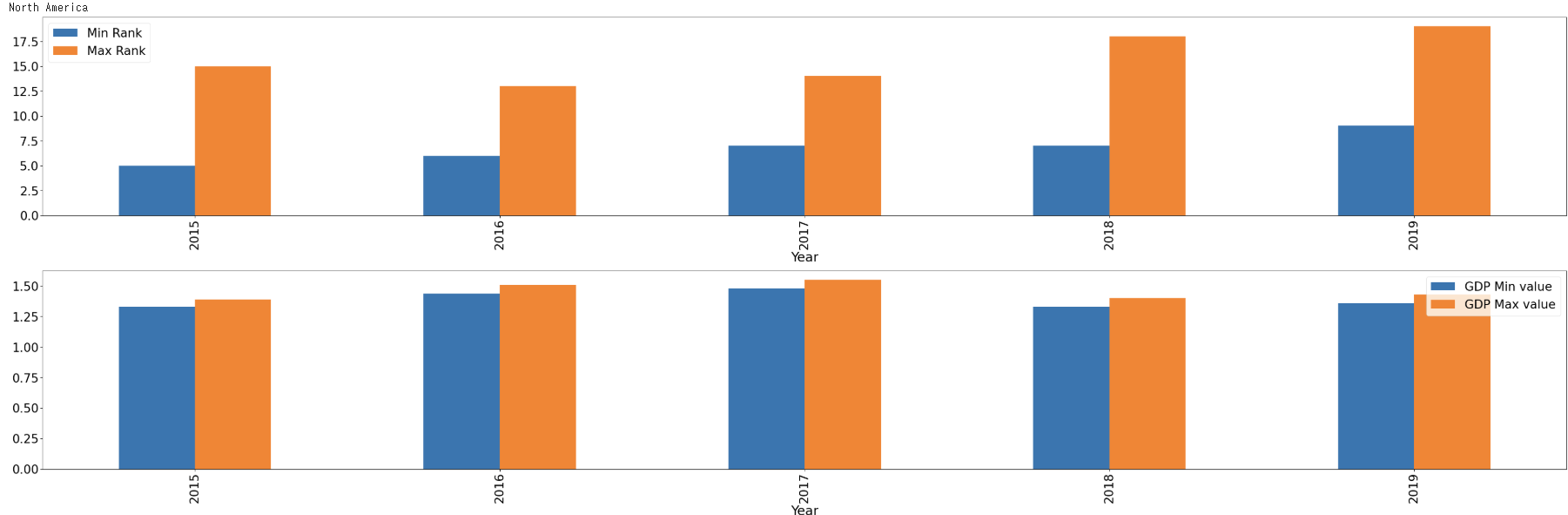


Western Europe의 GDP는 표준편차가 작지만 지역별 평균 GDP에서 상위권인 것으로 보아 전반적으로 GDP가 높은 지역이다

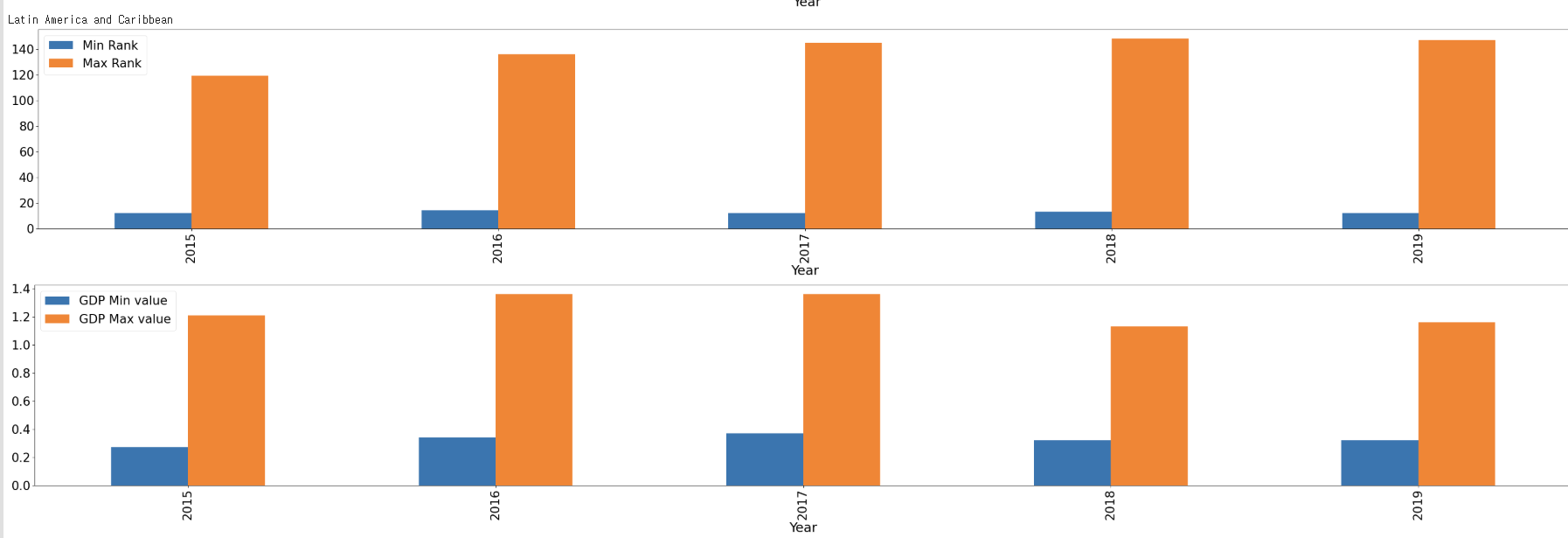
또한, Middle East and Northern Africa를 보았을 때, 중간 순위이지만 표준편차가 심하였다.

* 소득이 높을수록 행복지수가 높음
* Latin America and Caribbean지역은 평균 소득이 낮으나, 행복 순위가 높음
* North America의 GDP 표준편차가 작지만 순위가 높은 것으로 보아 대체적으로 다 잘사는 국가이다.
* Middle East and Northern Africa는 중간 순위이지만 표준편차가 심하다.
* 편차가 줄어드는 것은 지역간 격차가 줄어듦을 의미한다.

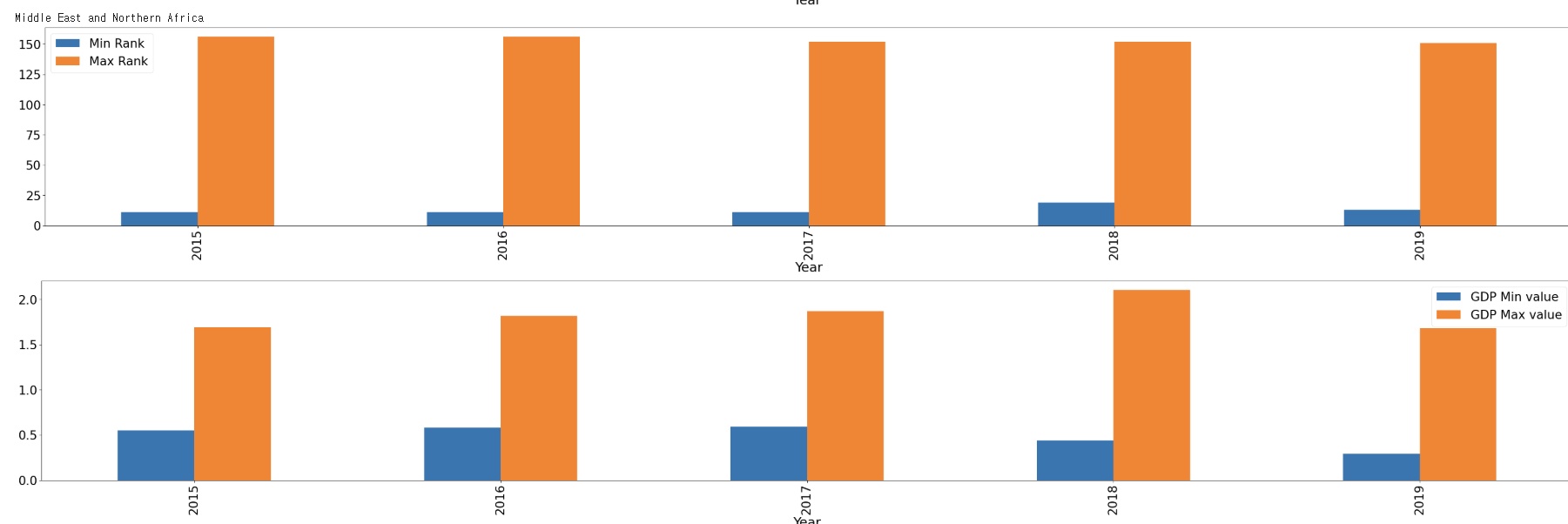
해당 정보들에 대해 의문을 가져 심층적으로 분석하였다.



North America 지역을 보면 생각보다 GDP지수가 낮은 국가와 높은 국가의 차이가 생각보다 크게 나지 않는 것을 알 수 있으며, 행복 순위 또한 2015년에는 격차가 조금 있었으나 시간이 지날수록 차이는 줄어들고 있음을 알 수 있다. 하지만 대체적으로 순위가 전체적으로 오르며 GDP가 17년을 기점으로 낮아진 것으로 보아, 지역 자체가 경제적으로 경쟁력이 약화된 것을 알 수 있다.



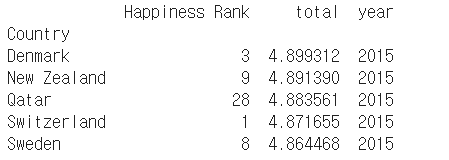
North America와 달리 Latin America and Caribbean 지역은 평균점이 낮으나 소득이 높은 이유는 행복 순위와 GDP가 낮은 국가와 높은 국가와의 차이가 극심하다는 것을 알 수 있다.



Middle East and Northern Africa 지역은 표준편차가 가장 큰 지역으로 나타났다. 분류하여 본 결과, Latin America and Caribbean 지역보다 더 극심한 차이가 보이는 것으로 나타났다. 또한, 이 지역은 지속적으로 차이가 극심하게 유지되는 특징이 보이는 것으로 보인다.

이러한 결과를 보았을 때, 지역마다 나라별 빈익빈 부익부가 크다는 것을 알 수 있었다.

이제 행복 지수에 대해 분석해 보았다. 아무리 생각해도 지수로 나타낸 통계이다 보니, 이를 전부 합하면 행복 등수가 되지 않을까? 하는 마음에 minmax 스케일링을 진행한 후, 이를 합산한 total로 top 5국가를 선정하였다.

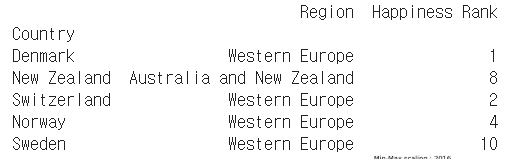




다음과 같이 대개, 전체를 합산하면 등수가 높은 것으로 나타난다. 하지만 완벽한 수치로 계산되지 못한 것은 기존 행복지수는 7가지를 고려해 산정하지만, 현재 공통된 값들만 이용하여 Dystopia Residual 값을 추가하지 못했다. 이를 이용한다면 더 가까운 수치를 산정할 수 있지 않을까 추론된다.

수업에서 배운 z-score로 다음 내용을 동일하게 진행하니 좀 더 순위에 가까운 결과가 나타났다.

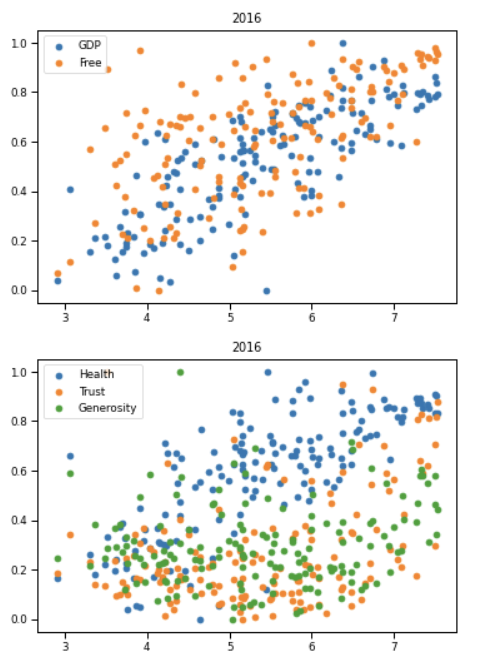




그냥 min-max scaling 보다는 z-score를 통해 진행하는 것이 더 행복지수에 가까운 결과를 도출해내는 것 같다.

그럼 어떤 데이터가 가장 행복 지수와 연관이 있을까?

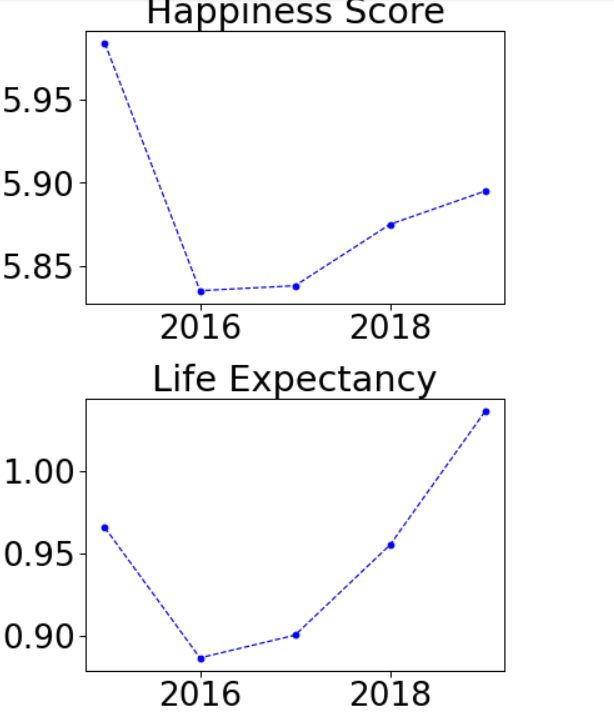
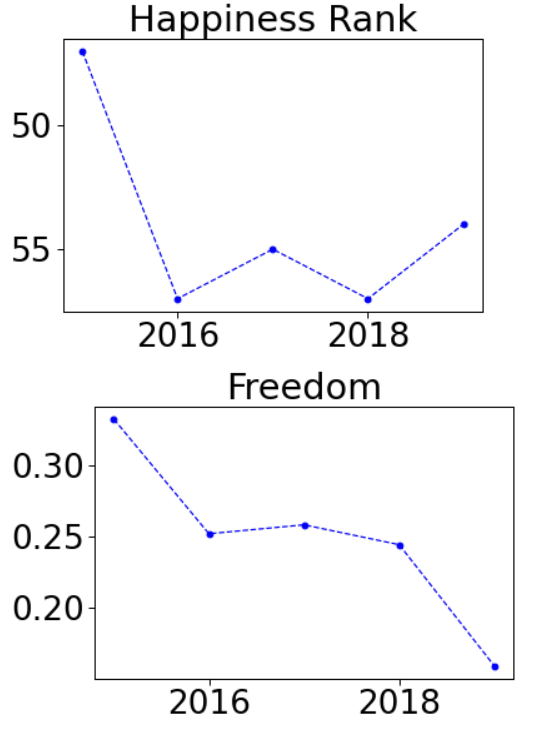
Scatter를 통해 산포도를 내어 행복지수와 각 컬럼간 상관관계를 나타내 보았다.

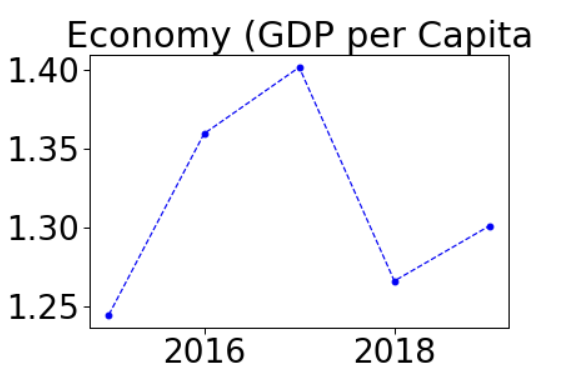




연도별 자료 분포도는 연도별로 비슷한 추세를 보였다. 기본적으로 GDP와 Freedom 지수가 행복 지수와 대부분 정비례하는 관계를 가졌으며, 생각 외로 건강지수가 행복 지수와 높은 상관관계를 가졌다. 생각보다 Trust와 Generosity 컬럼은 분포도가 밀집되어 있다는 것을 알 수 있었다. 행복 지수를 나타냄에 있어 이 두 컬럼은 큰 상관관계가 없어 보이는데 행복 지수를 계산하는데 한 요소로 사용되고 있다는 것의 의아함을 느꼈다.

추가적으로, 갑자기 한국의 데이터가 궁금해 한국은 5년동안 어떤 지표를 가졌는지 그래프를 이용하여 시각화 하였다.





해당 데이터를 통해 우리나라는 자유도는 시간이 지날수록 지표가 떨어지지만, 행복지수가 2015년에서 2016년으로 급격히 떨어진 이후, 다시 조금씩 증가하고 있음을 알 수 있다.

8. 분석 결과에 대한 해설

5년치 데이터를 통해 서로 간의 상관관계를 조사하면서 다음과 같은 정보를 알 수 있었다.

1. North America 지역은 대체적으로 GDP 지수가 높고 행복 지수가 높은 국가들로 이루어져 있다.
2. Latin America and Caribbean 지역은 평균 GDP에 비해 평균 행복 순위가 높은 지역이다.
3. 각 지역별로 빈익빈 부익부가 확실히 나타났고, Middle East and Northern Africa가 가장 큰 편차가 있던 것으로 나타났다.
4. 행복 지수와 크게 연관된 요소는 GDP, Freedom, 건강 세 개 요소가 가장 큰 연관을 가졌다.
5. 연도별로 Southeastern Asia 과 Western Europe 지역은 시간이 지날수록 지역 국가간 격차가 줄어들고 있다.
6. Z-score를 통해 각 지표를 합산한 것이 행복 지수와 가까운 결과물이 나타났다.
7. 한국은 건강 지표가 지속적으로 증가하지만, 자유도가 지속적으로 하락하고 있다. 전반적으로 하락한 이후 다시 증가하고 있는 추세이다.

행복지수를 산출함에 있어 특정 요소들이 큰 역할을 담당하지 못하는 것 같아 해당 부분에 있어 지표를 조금 수정하거나 행복지수 산출 기준을 일부 수정하는 것이 각 나라들의 지표를 더 잘 나타내지 않을까 생각되며, 우리나라가 행복지수가 높아지기 위해서는 자유도 지수를 높일 방안을 모색하여야 되지 않을까 생각한다.

9. 느낀점

되며 해당 과제를 하며 groupby 함수와 plot, map 등 다양한 pandas 문법을 사용하면서 데이터에 대한 다양한 시도를 진행하였지만 한 것에 비해 분석 결과물이 많이 나오지 않을 것 같아 아쉬웠다. 이번 데이터 분석을 통해 다른 데이터 분석에 도전해 봐야겠다는 생각이 들었고 더 나아가 이용할 수 있는 데이터로 정제해보도록 노력해야겠다.