Report

김준호

1. 1번 시각화의 목적과 효과를 평가하고 개선점을 제안해주세요. 목적과 효과 두 가지 이상을 설명해주세요. 개선점 두 가지 이상을 설명해주시고 개선점을 반영한 코드를 작성해주세요.

(1) 목적과 효과 설명

목적:

-상관관계 분석: 이 시각화는 다양한 변수들 사이의 상관관계를 시각적으로 분석하여, 어떤 변수들이 서로 강하게 연관되어 있는지를 파악하기 위해 사용된다.

-데이터 이해: 변숫값들 간의 관계를 이해함으로써 데이터셋의 구조적 특성을 파악하고, 이후의 데이터 분석이나 모델링에 중요한 인사이트를 얻기 위해 사용된다.

효과:

-직관적 이해: 열 지도(heatmap)를 통해 각 변수들 간의 상관계수를 색상의 강약으로 직관적으로 이해할 수 있다.

-변수 선택: 상관계수가 높은 변수들 사이의 관계를 파악하여, 중요 변수를 선택하거나 다중 공선성을 피하기 위한 전략을 세울 수 있다.

(2) 개선점 및 개선 코드

개선점:

-가독성 향상: 현재 색상 팔레트는 고대비이지만, 색상 범위가 충분히 다양하지 않아 일부 상관관계가 명확하게 드러나지 않을 수 있다. 더 명확한 색상 팔레트를 사용하는 것이 좋다.

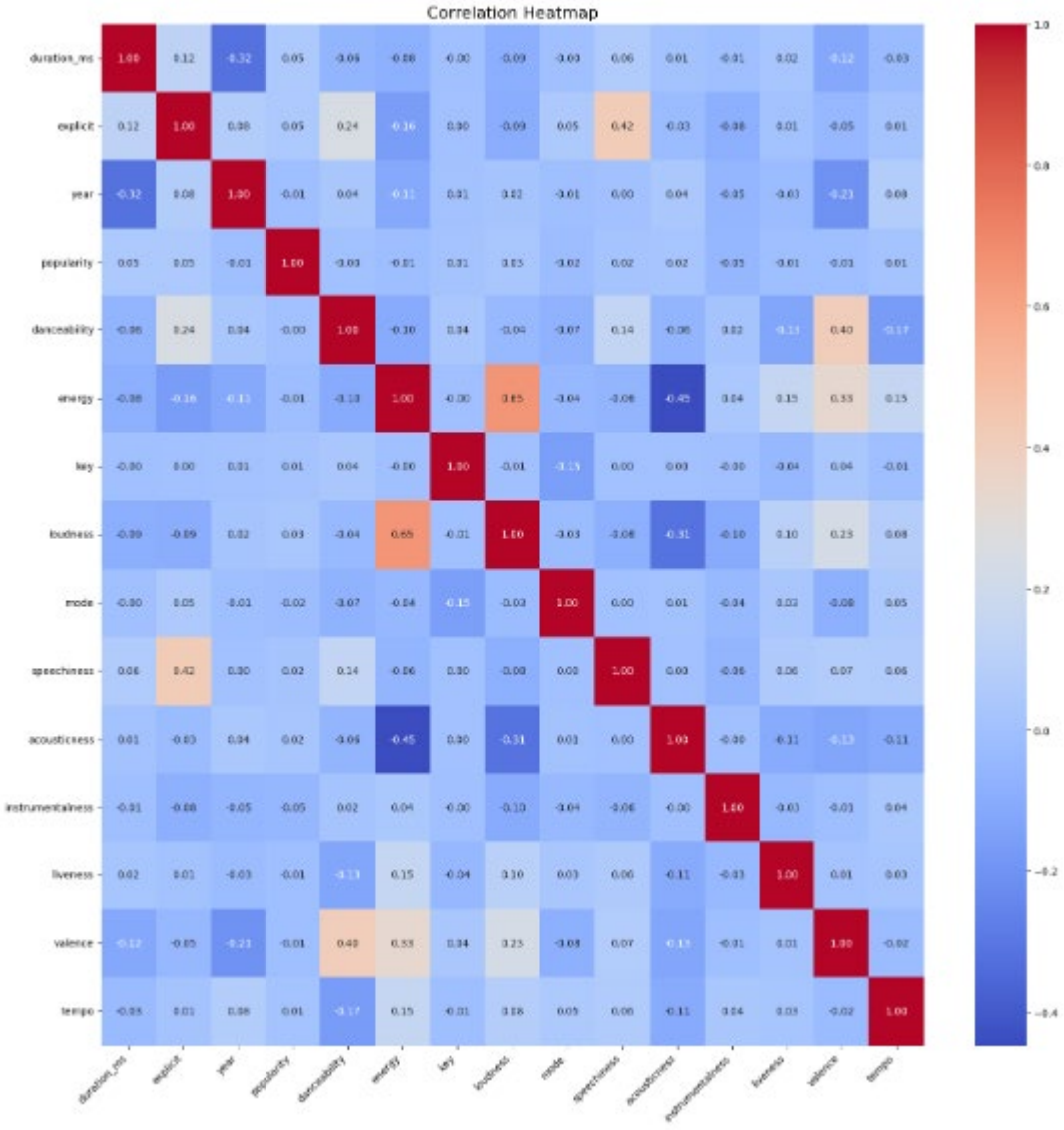
-크기 조절: 히트맵의 크기를 조금 더 키우거나 각 셀의 텍스트 크기를 조절하여 가독성을 향상시킬 수 있다.

-숫자 형식 조절: 소수점 이하 자릿수를 조절하여 숫자를 더 깔끔하게 표시할 수 있다.

개선 코드:

코드는 ipynb 파일에 있다. 이 코드는 히트맵의 색상 팔레트를 'coolwarm'으로 변경하여 상관관계를 더 명확하게 표현하고, 숫자 형식을 소수점 이하 두 자리로 조정하여 가독성을 향상시켰다. 또

한, 히트맵의 크기를 조정하고 축 레이블의 크기를 조절하여 전체적인 가독성을 높였다.



2. 2번 시각화의 목적과 효과를 평가하고 개선점을 제안해주세요. 목적과 효과 한 가지 이상을 설명해주세요. 개선점을 두 가지 이상을 설명해주시고 개선점을 반영한 코드를 작성해주세요.

(1) 목적과 효과 설명

목적:

-변수 간 관계 분석: 이 시각화는 instrumentalness와 popularity 간의 관계를 분석하고, 이 관계가 duration_ms와 explicit 변수에 따라 어떻게 변하는지를 시각적으로 나타내기 위해 사용된다.

효과:

-다차원 데이터 표현: 다양한 색상과 크기를 사용하여 다차원 데이터를 한눈에 이해할 수 있도록 돕는다.

-패턴 및 이상치 발견: instrumentalness와 popularity 간의 패턴과 특정 구간의 이상치를 발견하는 데 효과적이다.

(2) 개선점 및 개선 코드

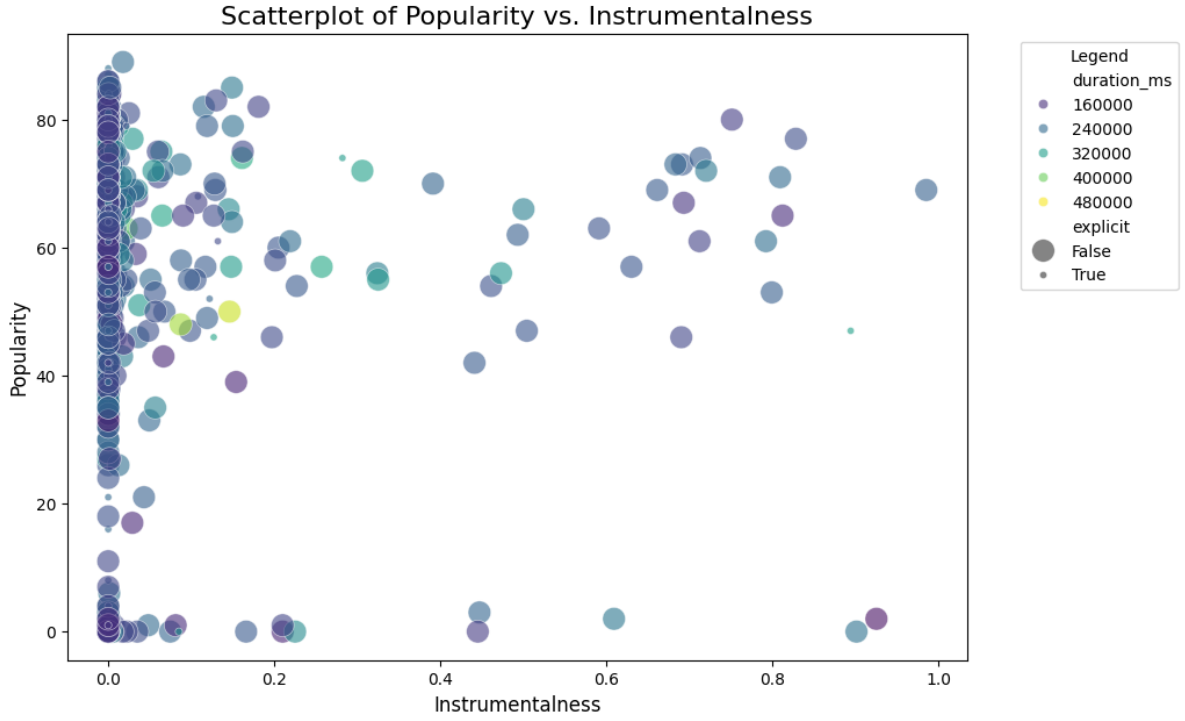
개선점:

-가독성 향상: 현재 색상 팔레트는 시각적으로 잘 구별되지 않으므로, 더 명확한 색상 팔레트를 사용하여 데이터 포인트를 구별할 필요가 있다.

-범례 및 제목: 범례와 제목을 더 명확하게 하고, 축 라벨을 추가하여 그래프의 의미를 더 잘 전달할 필요가 있다.

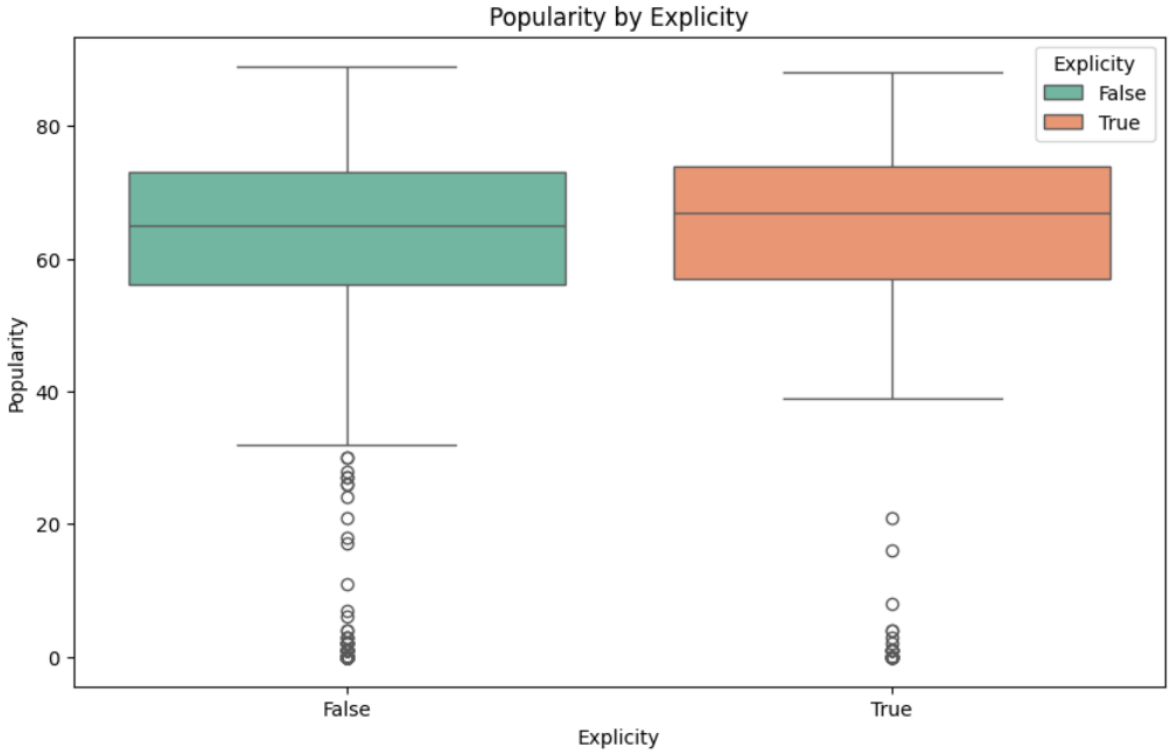
개선 코드:

코드는 ipynb 파일에 있다. 이 코드는 산점도의 크기와 투명도를 조정하고, 색상 팔레트를 'viridis'로 변경하여 데이터 포인트 간의 차이를 더 명확하게 구분할 수 있게 한다. 또한, 제목과 축 라벨을 추가하여 그래프의 의미를 더 잘 전달할 수 있도록 하였다.



3. explicity가 popularity에 영향을 주는지 주지 않는지 판단하고 시각화를 통해 이를 정당화하세요. 판단과 정당화에 대한 설명을 작성해주세요.

(1) 시각화 결과



(2) 판단과 정당화

explicit 변수가 popularity에 영향을 주는지 판단하기 위해, explicit 변수에 따른 popularity의 분포를 boxplot을 통해 시각화하였다.

boxplot에서 explicity가 True인 경우와 False인 경우의 popularity 분포를 비교해보자. 만약 두 그룹 간의 중앙값, 사분위 범위, 이상치 등이 유사하다면 explicity가 popularity에 큰 영향을 주지 않는다고 판단할 수 있다. 반면, 명확한 차이가 존재하면 영향을 준다고 볼 수 있다.

위의 시각화 결과를 보면, 두 그룹 간 분포의 명확한 차이가 존재하지 않는 것을 확인할 수 있다. 따라서, explicity가 popularity에 미치는 영향은 매우 미미하다는 결론을 낼 수 있다.

4. 0725_visualization.ipynb에서 spotify 데이터를 시각화하여 내릴 수 있는 결론을 설명해주세요.

Spotify 데이터를 다양하게 시각화한 결과 다음과 같은 결론을 내릴 수 있다.

(1) 상관관계 분석 결과:

Energy와 Loudness는 높은 양의 상관관계를 가지며, 이는 에너지가 높은 곡일수록 소리의 강도도 높다는 것을 의미한다.

Danceability와 Valence도 양의 상관관계를 가지며, 이는 춤추기 좋은 곡일수록 긍정적인 분위기를 가진다는 것을 나타낸다.

Instrumentalness와 Popularity는 음의 상관관계를 가지며, 기악곡일수록 인기가 낮다는 경향이 있다.

(2) 산점도 분석 결과:

Instrumentalness와 Popularity 간의 산점도에서 대부분의 데이터가 Instrumentalness가 낮은 구간에 몰려 있으며, Popularity가 다양한 범위에 분포되어 있다. 이는 기악곡의 경우 인기가 높은 곡이 적다는 것을 보여준다.

Duration_ms와 Explicit 변수를 색상과 크기로 나타낸 결과, 곡의 길이나 명시적인 콘텐츠가 인기에 큰 영향을 미치지 않는다는 것을 시사한다.

(3) Explicit 변수의 영향:

Explicit 변수에 따른 Popularity의 분포를 box plot, violin plot, strip plot으로 비교한 결과, explicit한 콘텐츠 여부가 인기의 중앙값이나 분포에 큰 영향을 주지 않는 것으로 나타났다. 이는 explicit 여부가 곡의 인기에 결정적인 요소가 아니라는 것을 의미한다.

(4) 다른 변수들 간의 관계:

Danceability와 Energy는 양의 상관관계를 가지며, 춤추기 좋은 곡일수록 에너지가 높은 경향이 있다.

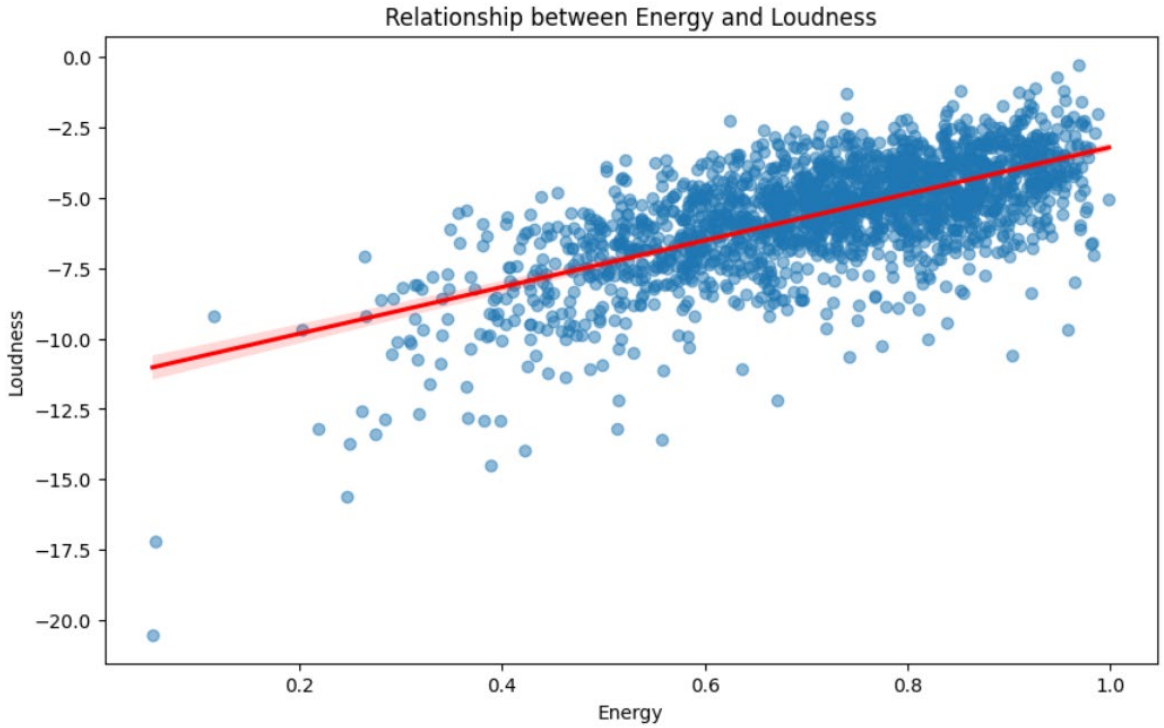
Acousticness와 Energy는 음의 상관관계를 가지며, 어쿠스틱한 곡일수록 에너지가 낮다는 것을 나타낸다.

결론적으로, Spotify 데이터의 시각화 분석을 통해 인기 있는 곡의 특성을 이해하고, 여러 변수들 간의 관계를 파악할 수 있었다. 기악곡의 인기가 상대적으로 낮고, explicit 여부는 인기에 큰 영향

을 미치지 않으며, 곡의 에너지와 춤추기 좋은 특성이 강하게 연관되어 있다는 것을 알 수 있다.

5. 물음4에서 내린 결론을 정당화하기 위해 적절한 시각화를 한 개 이상 추가해주세요. 정당화에 대한 설명을 작성해주세요.

(1) Energy와 Loudness 간의 높은 상관관계를 정당화하기 위한 **산점도와 회귀선을 포함한 시각화**를 구현했다.

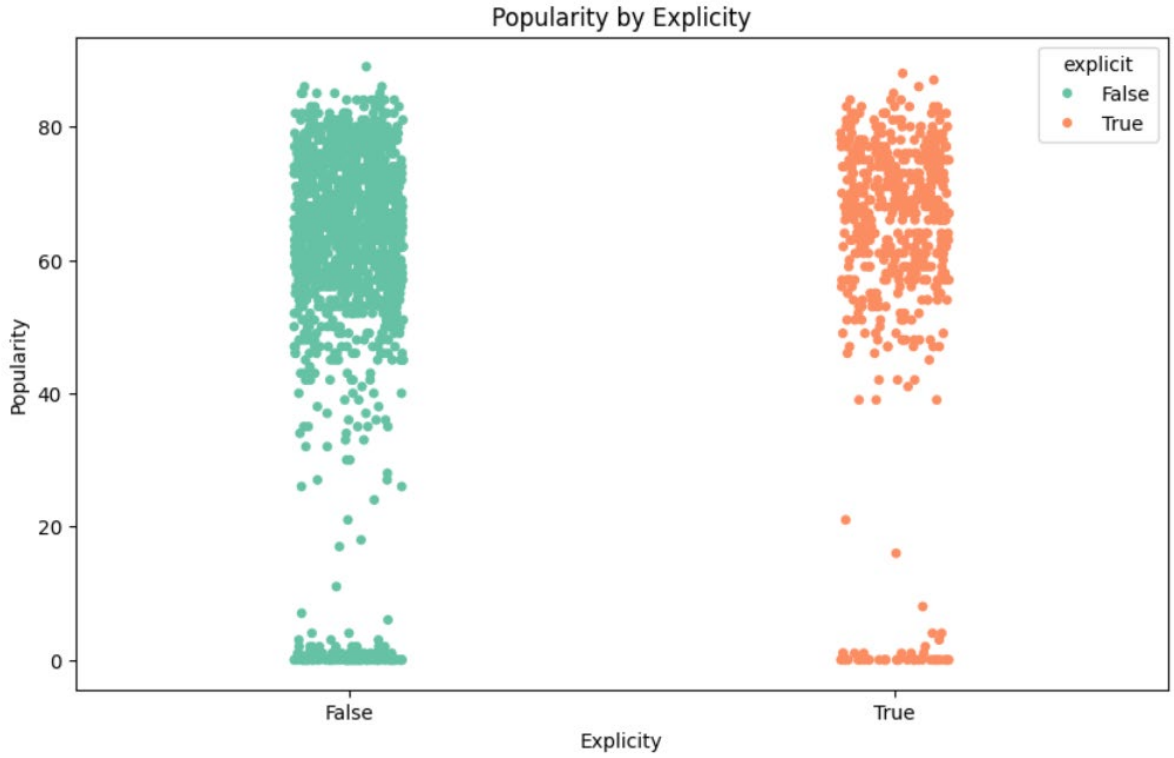


위 산점도는 두 변수 간의 데이터 포인트를 시각적으로 나타내며, 회귀선은 그들의 관계를 더욱 명확히 보여준다.

정당화: 시각화 결과, Energy와 Loudness 사이에 명확한 양의 상관관계가 존재하는 것을 확인할 수 있다. 이는 에너지가 높은 곡일수록 소리의 강도도 높다는 것을 의미한다.

회귀선은 두 변수 간의 관계가 선형적임을 보여주며, 이는 두 변수 사이의 강한 연관성을 시사한다.

(2) explicit와 popularity 간의 관계를 시각화하기 위해 box plot이 아닌 **strip plot**을 추가적으로 구현해보았다. strip plot은 각 데이터 포인트를 개별적으로 보여주어 데이터의 분산과 개별 포인트를 확인하는 데 유용하다.



위 시각화를 통해 explicit가 True인 경우와 False인 경우의 popularity 분포를 비교하고, 명확한 차이가 존재하는지 확인할 수 있다.

정당화: 두 그룹 간의 분포와 데이터 포인트의 분포가 유사하여 명확한 차이가 존재하지 않으므로 explicit가 popularity에 큰 영향을 주지 않는다고 판단할 수 있다.

life_expectancy_visualization.ipynb에 대한 report

(1) 검증/답하고자 하는 가설 혹은 질문

질문: "경제적 지표와 보건 지표 등이 기대 수명(Life Expectancy)에 미치는 영향은 무엇인가?"

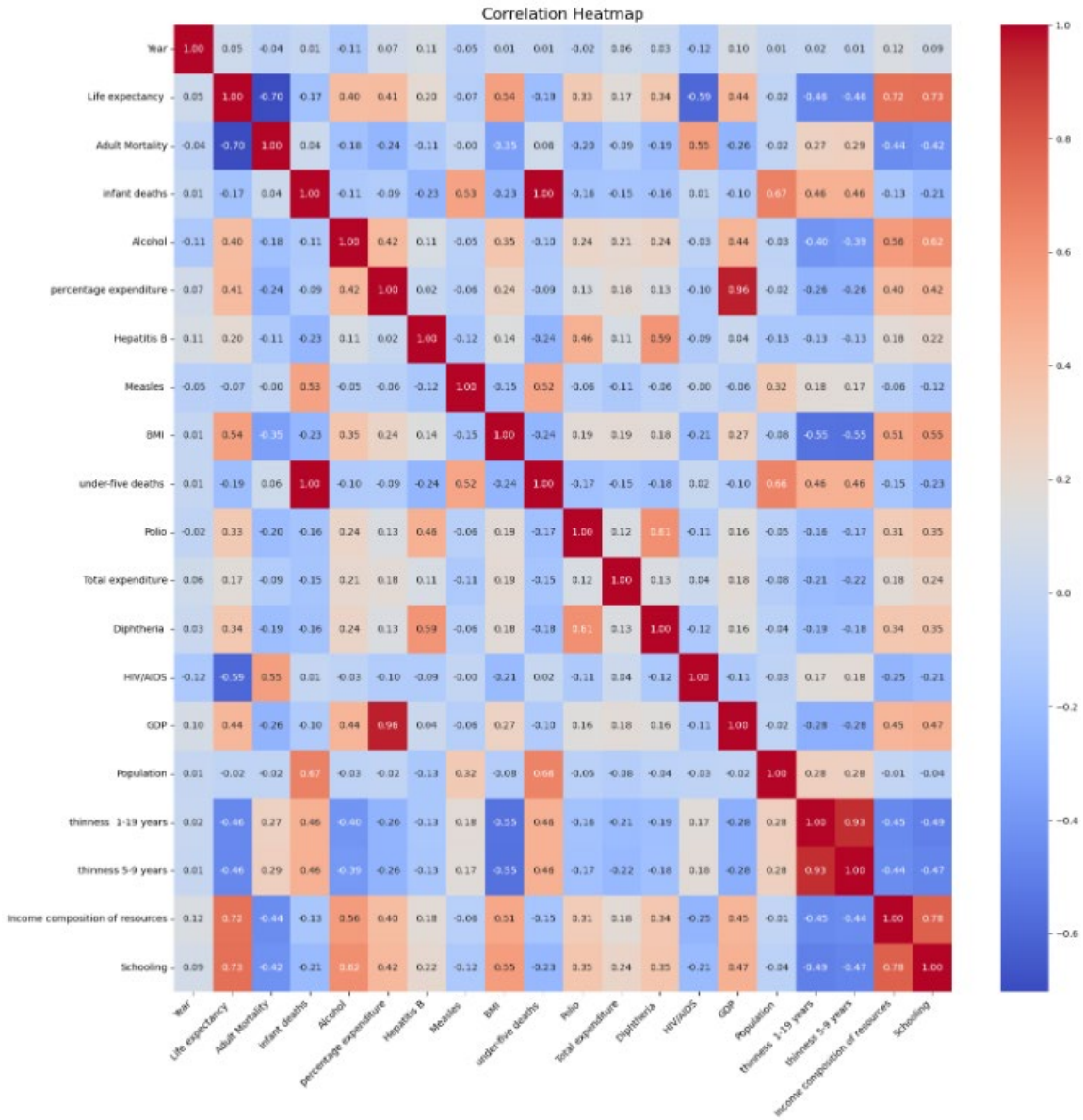
가설: GDP가 높을수록, 소득구성 지표가 높을수록 기대수명이 높을 것이다, 교육(Schooling) 지표가 높을수록 기대수명이 높을 것이다. 성인 사망률이 높을수록 기대수명이 낮을 것이다.

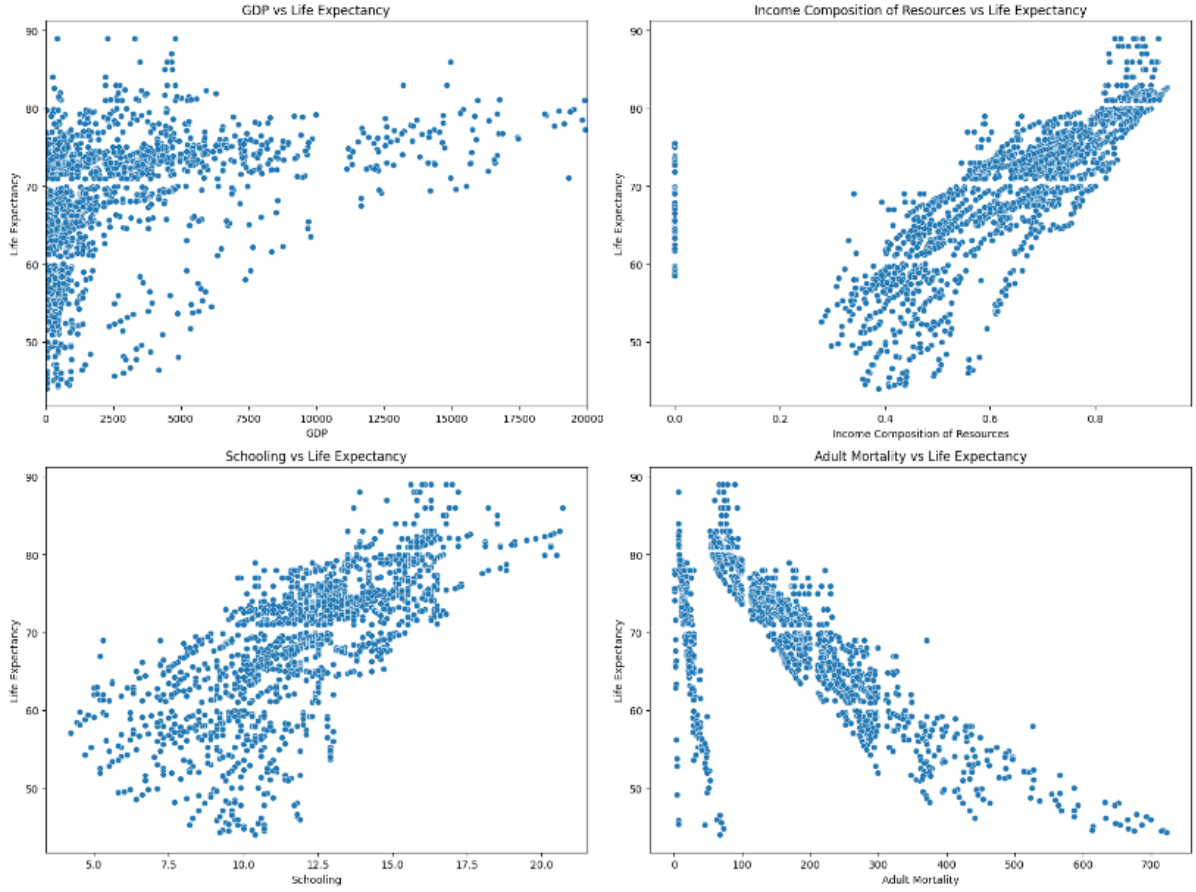
(2) 살펴보거나 고려해야 하는 독립변수, 종속변수, 데이터의 특성

종속변수: Life expectancy (수치형 자료)

독립변수: GDP, Income composition of resources, Schooling, Adult Mortality (모두 수치형 자료)

(3) 완료한 시각화와 가설/질문에 대한 결론





가설을 모두 채택할 수 있다. 경제적 지표와 보건적 지표가 우수할수록 기대수명이 증가하는 경향을 보인다. 자세한 설명은 (4)번에서 한다.

(4) 시각화에서 얻을 수 있는 인사이트

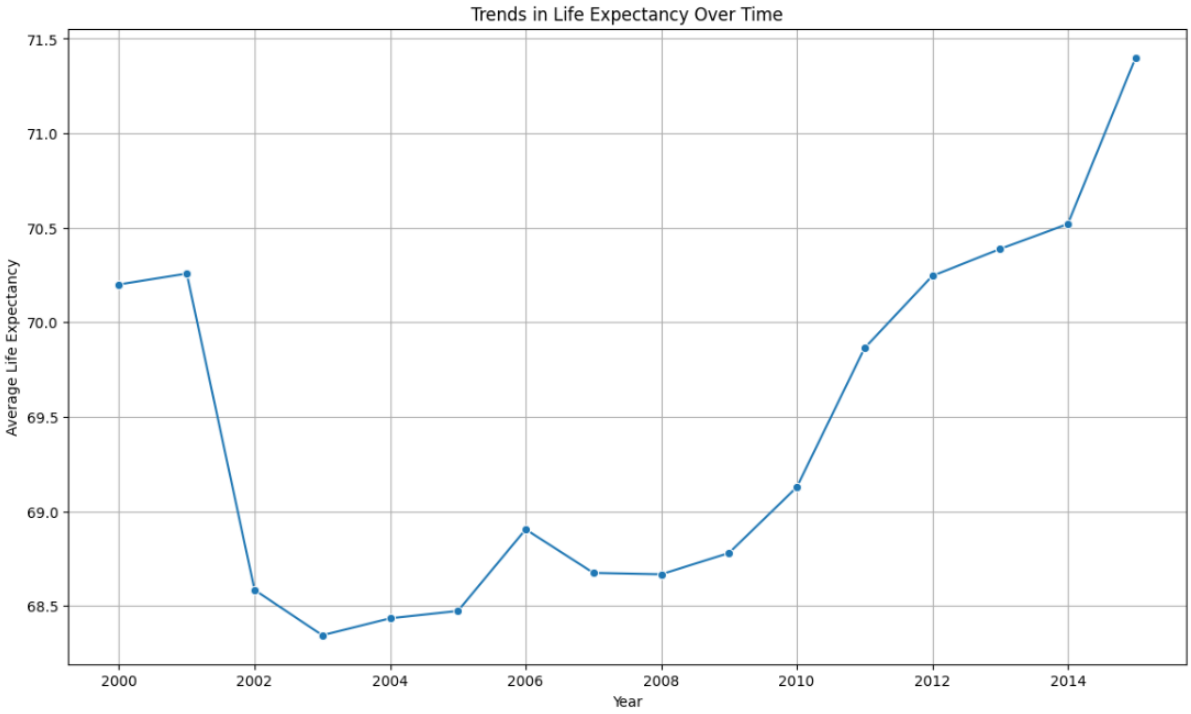
GDP와 기대 수명: GDP와 기대 수명 사이에는 양의 상관관계가 나타난다. 경제적으로 더 부유한 국가일수록 기대 수명이 더 높게 나타나는 경향이 있으나, 상관관계가 0.44로 그 정도가 아주 뚜렷하진 않다. Scatter plot 또한 어느 정도의 양의 선형관계를 보여주나 역시 아주 또렷하진 않다. 실제 GDP 수치는 12만 정도까지 있지만 시각화 결과 대부분의 표본이 20000 아래이므로 그 이상은 이상치로 판단하여 GDP 범위는 0~20000으로 제한하여 나타냈다.

소득 구성 지표와 기대 수명: 소득 구성 지표가 높은 국가일수록 기대 수명이 더 높다. 상관관계도 0.72로 높고, Scatter plot 또한 뚜렷한 양의 선형관계를 보인다. 이는 자원의 소득 분배가 건강에 중요한 역할을 한다는 것을 시사한다.

교육(Schooling) 지표와 기대 수명: 두 지표 또한 양의 상관관계가 나타난다. 소득 구성 지표와 비슷한 0.73 정도의 높은 양의 상관관계가 나타나고, Scatter plot 또한 뚜렷한 양의 선형관계를 보인다. 이는 높은 교육 수준 또한 기대 수명에 중요한 역할을 한다는 결론으로 이어진다.

성인 사망률과 기대 수명: 성인 사망률이 낮을수록 기대 수명이 높다. 이는 성인의 건강 관리가 중요하다는 것을 나타낸다. 이는 직관적으로 당연한 결과이다. -0.70의 강한 음의 상관관계를 보이고, Scatter plot에서도 뚜렷한 음의 선형관계를 보여준다.

결론적으로, 경제적 지표와 보건 지표는 기대 수명에 큰 영향을 미치며, 부유한 국가일수록, 그리고 높은 교육 수준을 갖추고 건강 관리가 잘 이루어지는 국가일수록 기대 수명이 높다는 것을 알 수 있다. 이를 통해 정책 입안자들은 국민의 건강 증진을 위해 경제적 안정과 보건 자원의 효율적 분배가 중요함을 이해할 수 있다.



추가적으로, 시간이 지날수록 전체적으로 기대수명이 증가하는 추세를 보아 전세계적으로 선진화에 따른 기대수명 증가가 이루어지고 있음을 확인할 수 있다.