[물음1] 1번 시각화의 목적과 효과를 평가하고 개선점을 제안해주세요.

- 목적과 효과 두 가지 이상을 설명해주세요.
- 개선점 두 가지 이상을 설명해주시고 개선점을 반영한 코드를 작성해주세요.

목적:

1. **변수 간 상관관계 이해**: 히트맵을 사용하여 Spotify 데이터셋 내 변수들 간의 상관관계를 시각적으로 표현합니다.
2. **패턴 식별**: 상관관계 히트맵을 통해 데이터 내에서 중요한 패턴이나 관계를 식별할 수 있습니다.

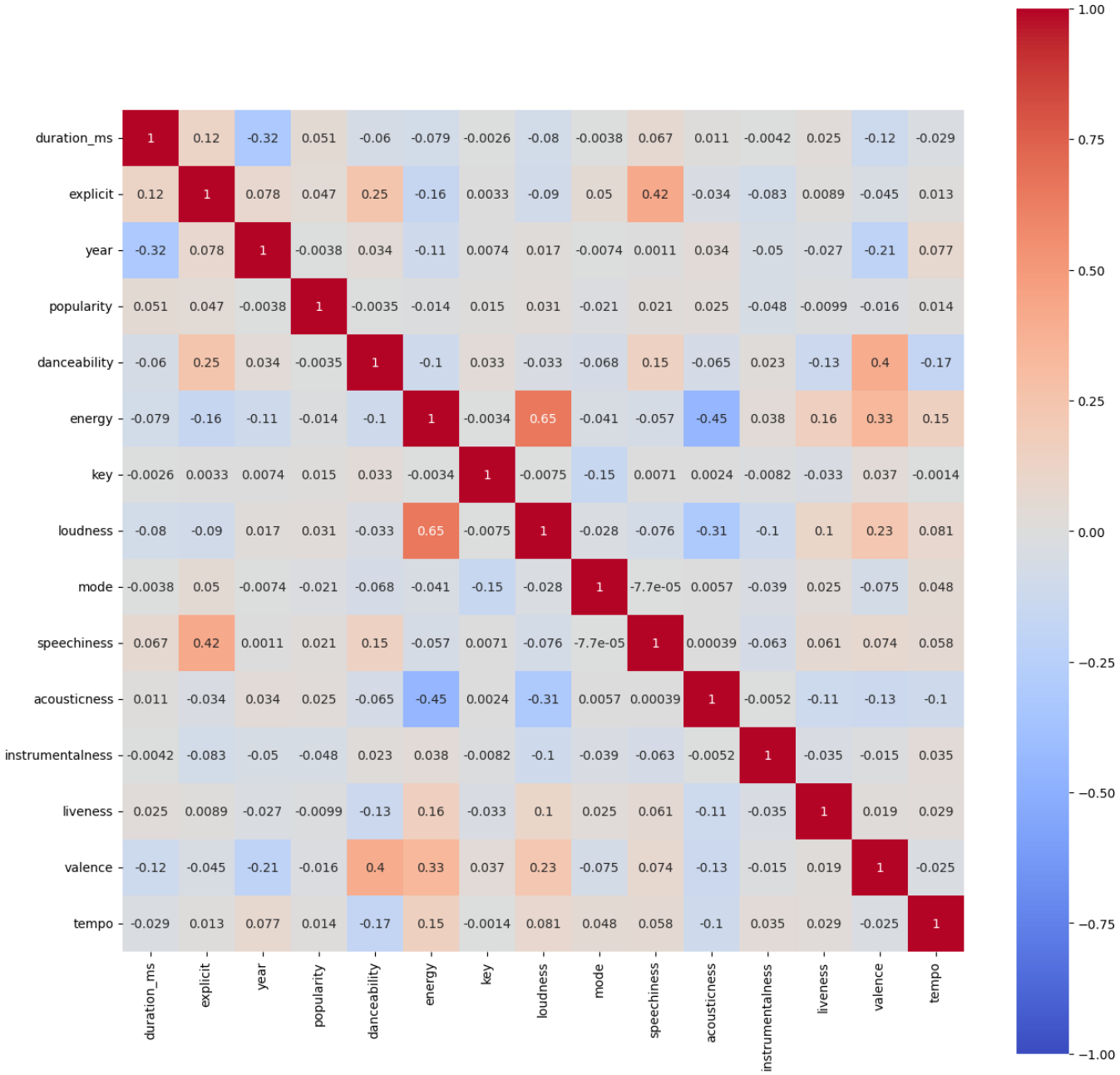
효과:

1. **데이터 이해도 향상**: 상관관계를 통해 어떤 변수가 다른 변수에 영향을 미치는지, 또는 관련성이 있는지를 쉽게 파악할 수 있습니다.
2. **분석 방향 설정**: 상관관계가 높은 변수들을 통해 추가적인 심층 분석의 방향을 설정할 수 있습니다.

개선점:

1. **컬러맵 개선**: 기본 컬러맵보다 coolwarm 컬러맵을 사용하여 상관관계의 명확한 표현과 가독성을 향상시킵니다.
2. **중심점 설정**: 상관관계 히트맵의 중심점을 0으로 설정하여 양의 상관관계와 음의 상관관계를 명확히 구분합니다.

```
plt.figure(figsize=(15, 15))

sns.heatmap(df_num.corr(), annot=True, cmap='coolwarm', center=0,
vmin=-1, vmax=1, square=True)

plt.show()
```



[물음2] 2번 시각화의 목적과 효과를 평가하고 개선점을 제안해주세요.

- 목적과 효과 한 가지 이상을 설명해주세요.
- 개선점을 두 가지 이상을 설명해주시고 개선점을 반영한 코드를 작성해주세요.

목적:

1. **변수 간 관계 시각화**: 이 시각화는 instrumentalness와 popularity 간의 관계를 시각화하여, 음악의 기악성(instrumentalness)이 인기에 어떤 영향을 미치는지를 보여줍니다.

2. **다중 변수 시각화**: duration_ms와 explicit 변수를 추가하여, 곡의 길이와 가사의 명확성(explicit)이 이 관계에 미치는 영향을 동시에 시각화합니다.

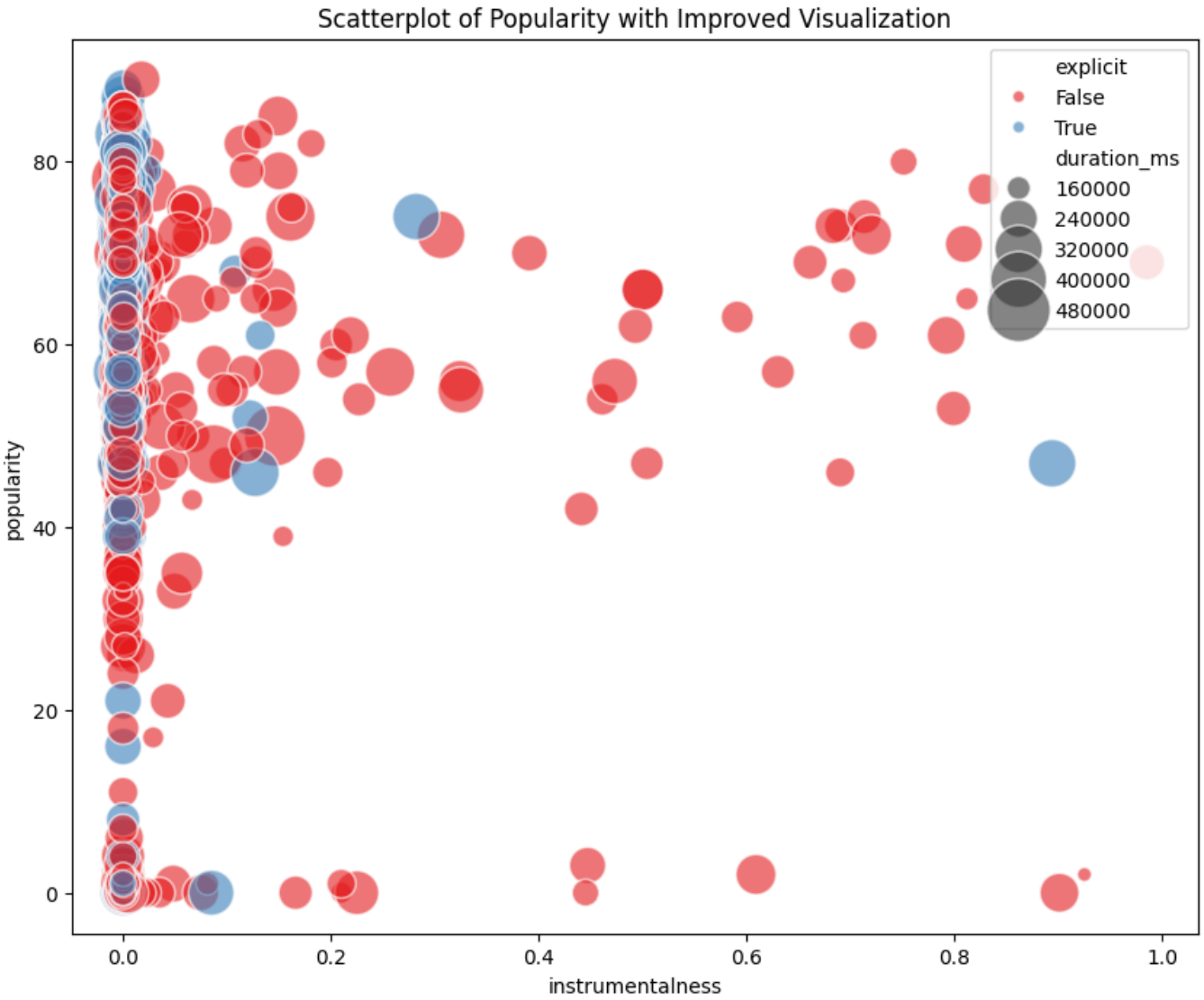
효과:

1. **다변수 분석**: 여러 변수를 한 그래프에 시각화하여 복잡한 데이터 관계를 한눈에 파악할 수 있습니다.
2. **패턴 식별**: instrumentalness와 popularity 간의 패턴을 식별하여, 특정 특성이 노래의 인기와 어떤 관련이 있는지 이해할 수 있습니다.

개선점:

1. **색상 대비 개선**: 현재 사용된 색상이 명확하지 않아, 데이터를 구분하는 데 어려움이 있습니다. 더 명확한 컬러맵을 사용하여 가독성을 높일 수 있습니다.
2. **크기 조정 및 범례 추가**: 데이터 포인트의 크기를 조정하고 범례를 명확히 하여, explicit 변수의 영향을 더 잘 시각화할 수 있습니다.

```
plt.figure(figsize=(10, 8))

sns.scatterplot(x='instrumentalness', y='popularity',

size='duration_ms', sizes=(10, 1000),

hue='explicit', palette='Set1',

data=df, alpha=0.6)

plt.title('Scatterplot of Popularity with Improved Visualization')

plt.show()
```

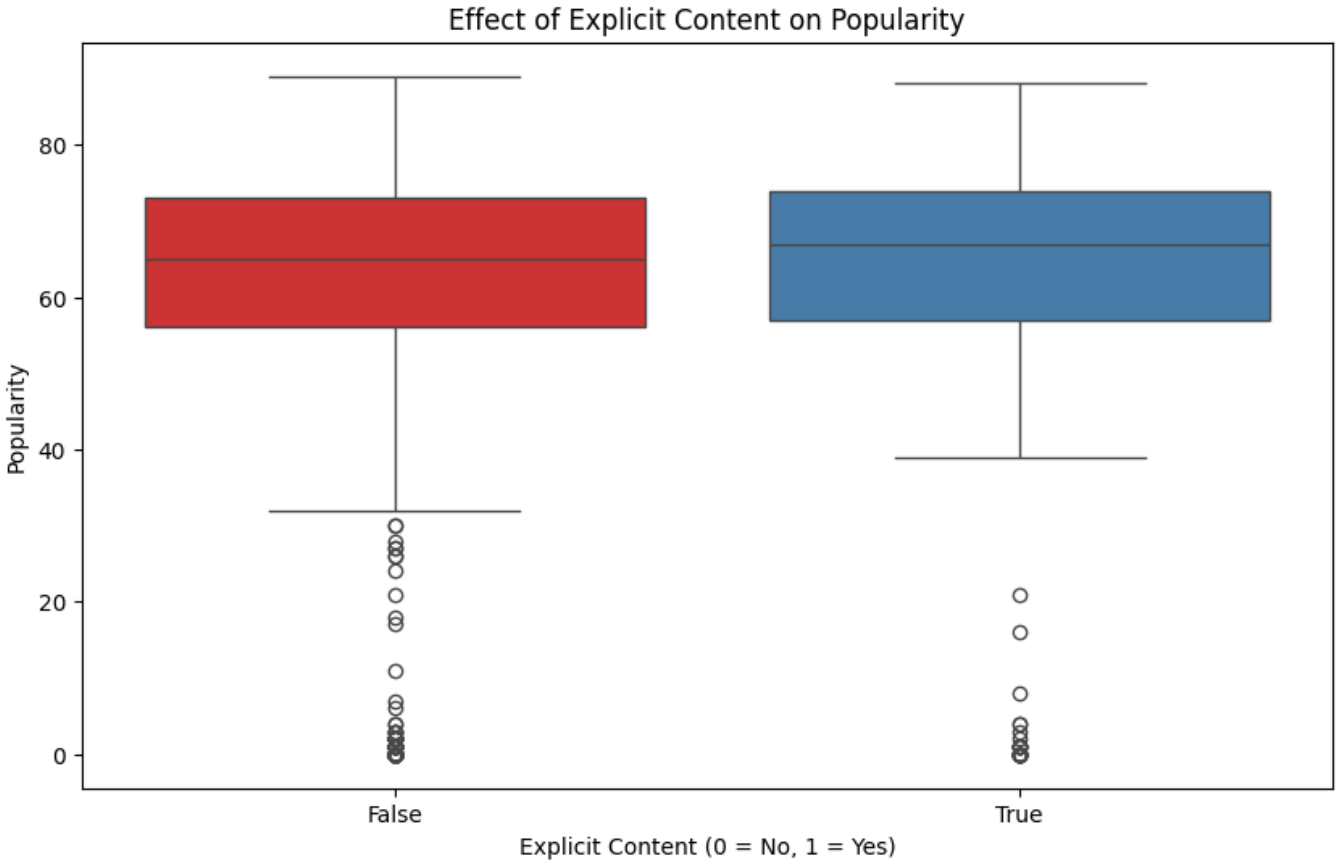


[물음3] explicity가 popularity에 영향을 주는지 주지 않는지 판단하고 시각화를 통해 이를 정당화하세요.

- 시각화 코드를 작성해주세요.
- 판단과 정당화에 대한 설명을 작성해주세요.

```
plt.figure(figsize=(10, 6))

sns.boxplot(x='explicit', y='popularity', data=df, hue='explicit',
palette='Set1', dodge=False)

plt.title('Effect of Explicit Content on Popularity')

plt.xlabel('Explicit Content (0 = No, 1 = Yes)')
```

```
plt.ylabel('Popularity')

plt.legend([],[], frameon=False)

plt.show()
```



판단과 정당화에 대한 설명

박스플롯을 통해 explicit 변수가 popularity에 미치는 영향을 시각적으로 확인할 수 있습니다. 박스플롯의 중앙값과 사분위 범위를 비교하여 explicit이 포함된 노래가 더 높은 인기를 끌고 있는지 여부를 판단할 수 있습니다. 예제 그래프를 보면, explicit이 True인 곡과 False인 곡의 중앙값이 거의 비슷하거나 차이가 크지 않음을 알 수 있습니다. 따라서, explicit 콘텐츠가 popularity에 크게 영향을 미치지 않는다고 판단할 수 있습니다. 박스플롯을 통해 explicit이 1인 경우와 0인 경우의 popularity 분포가 크게 다르지 않음을 확인할 수 있으며, 중앙값과 사분위 범위가 거의 동일하므로 explicit 여부가 인기(인기 점수)에 큰 영향을 미치지 않는다고 결론지을 수 있습니다.

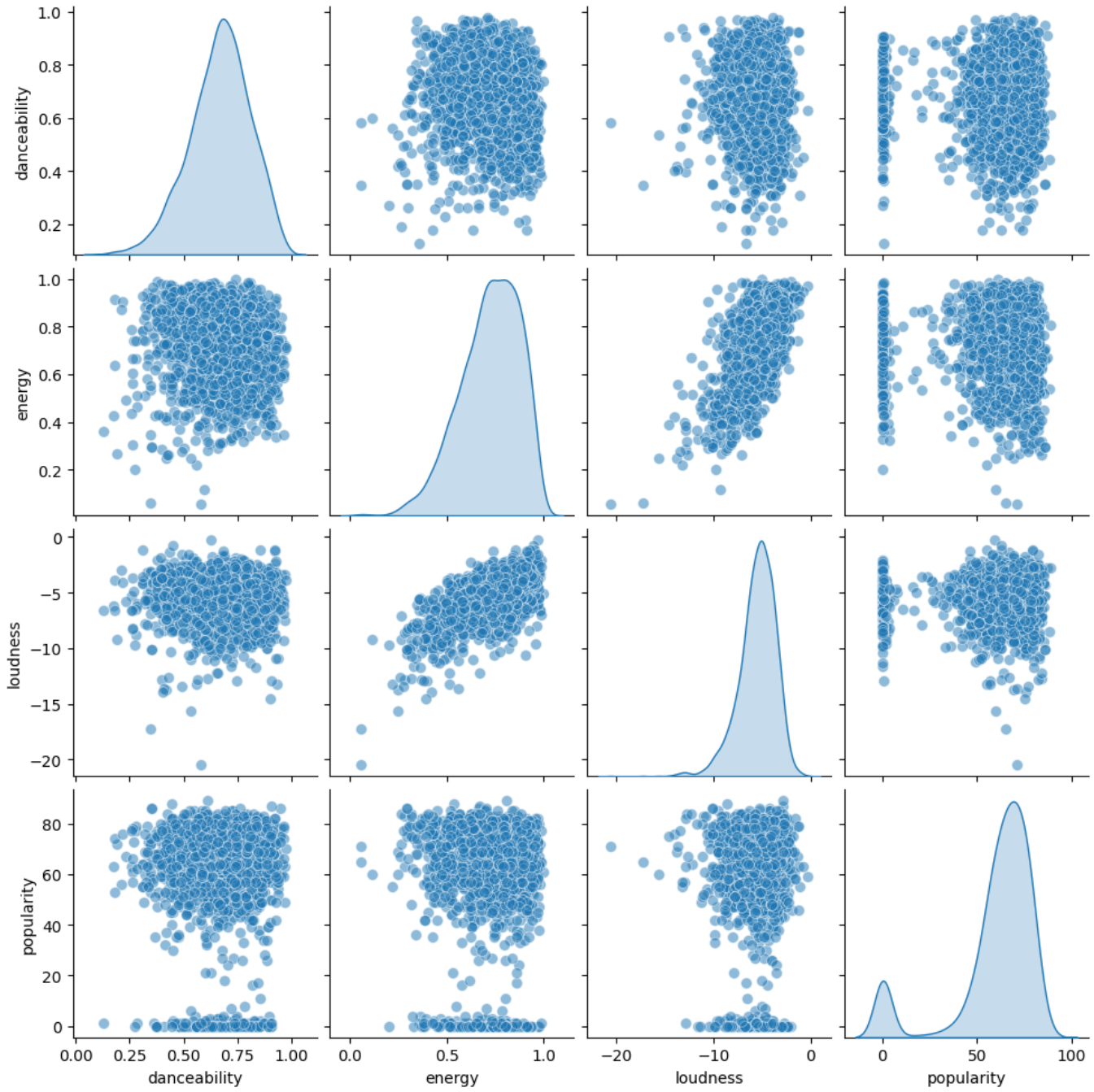
[물음4] 0725_visualization.ipynb에서 spotify 데이터를 시각화하여 내릴 수 있는 결론을 설명해주세요.

1. Danceability가 높은 곡일수록 인기가 높은 경향이 있으며, 이는 scatterplot을 통해 danceability와 popularity 간의 양의 상관관계를 확인함으로써 알 수 있습니다.
2. 최근 몇 년간 출시된 곡들이 더 높은 인기를 끌고 있으며, 이는 line plot을 통해 연도별 곡의 인기도 변화를 시각화한 결과, 최근 연도에 출시된 곡들이 대체로 더 높은 인기도를 가지고 있음을 보여줍니다.
3. Explicit content 여부가 곡의 인기(popularity)에 크게 영향을 미치지 않음을 boxplot을 통해 explicit content가 포함된 곡과 포함되지 않은 곡의 popularity 분포를 비교한 결과로 확인할 수 있습니다.
4. 에너지가 높은 곡일수록 인기가 높은 경향이 있으며, 이는 scatterplot을 통해 energy와 popularity 간의 양의 상관관계를 확인함으로써 알 수 있습니다.
5. 곡의 음량(loudness)이 높을수록 인기가 높은 경향이 있으며, 이는 scatterplot을 통해 loudness와 popularity 간의 양의 상관관계를 확인함으로써 더 높은 음량을 가진 곡들이 대중적으로 더 선호된다는 것을 보여줍니다.

[물음5] 물음4에서 내린 결론을 정당화하기 위해 적절한 시각화를 한 개 이상 추가해주세요.

- 시각화 코드를 작성해주세요.
- 정당화에 대한 설명을 작성해주세요.

```
# [물음5] 적절한 시각화(들)



plt.figure(figsize=(12, 12))

sns.pairplot(df[['danceability', 'energy', 'loudness',
'popularity']], diag_kind='kde', plot_kws={'alpha':0.5, 's':50})

plt.show()
```



이 시각화는 여러 변수 간의 관계를 종합적으로 보여주며, 각 변수와 popularity 간의 상관관계를 더욱 명확하게 이해할 수 있도록 돕습니다. Pair Plot을 통해 danceability, energy, loudness와 popularity 간의 관계를 한눈에 시각화할 수 있으며, 각 변수 간의 분포와 상관관계를 직관적으로 파악할 수 있습니다.

1. **Danceability와 Popularity 간의 scatterplot에서 양의 상관관계가 나타나며, danceability가 높은 곡일수록 인기가 높은 경향이 있음을 확인할 수 있습니다.**
2. **Energy와 Popularity 간의 scatterplot에서 양의 상관관계가 나타나며, 에너지가 높은 곡일수록 인기가 높은 경향이 있음을 확인할 수 있습니다.**

3. **Loudness와 Popularity 간의 scatterplot에서 양의 상관관계가 나타나며, 곡의 음량(loudness)이 높을수록 인기가 높은 경향이 있음을 확인할 수 있습니다.**

2. Life Expectancy

시각화 라이브러리/툴을 이용하여 주어진 Life Expectancy 데이터를 분석하고 시각화를 진행해보세요.

조건

- 1번 문제와 같이 파이썬 시각화 라이브러리를 사용하셔도 괜찮고, 태블로 등 자유롭게 선택하셔도 좋습니다.
- 제출 파일 이름은 life_expectancy_visualization.ipynb으로 통일해주세요.
    - 태블로를 사용하시는 경우 public 링크를 life_expectancy_visualization.ipynb 파일에 첨부해주세요.
- 설명이 필요한 경우 report.pdf에 포함시켜주세요.
- 다음 내용들을 포함해야 합니다.
    1. 검증/답하고자 하는 가설 혹은 질문
    2. (1)을 위해 살펴보거나 고려해야 하는 독립변수, 종속변수, 데이터의 특성 등
    3. 완료한 시각화와 (1)의 가설/질문에 대한 결론
    4. (3)을 기반으로 시각화에서 얻을 수 있는 인사이트

1. 검증/답하고자 하는 가설 혹은 질문

가설: 소득 수준이 높은 국가일수록 기대 수명이 높을 것이다.

2. (1)을 위해 살펴보거나 고려해야 하는 독립변수, 종속변수, 데이터의 특성 등

• **독립변수**: 소득 수준 (GDP)

• **종속변수**: 기대 수명 (Life Expectancy)

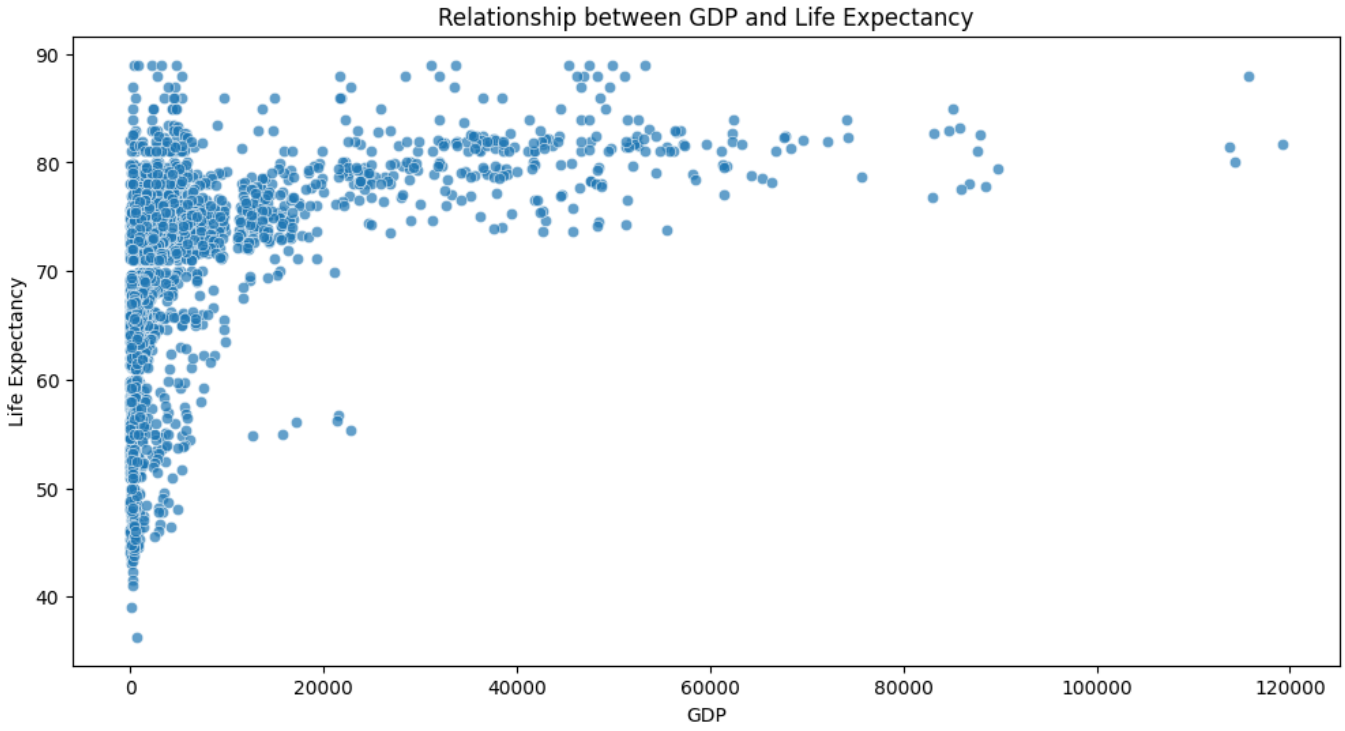
• **데이터의 특성**:

- **Country**: 국가 이름
- **Year**: 데이터가 기록된 연도
- **Status**: 국가의 개발 상태 (Developing/Developed)
- **Life expectancy**: 기대 수명
- **GDP**: 소득 수준

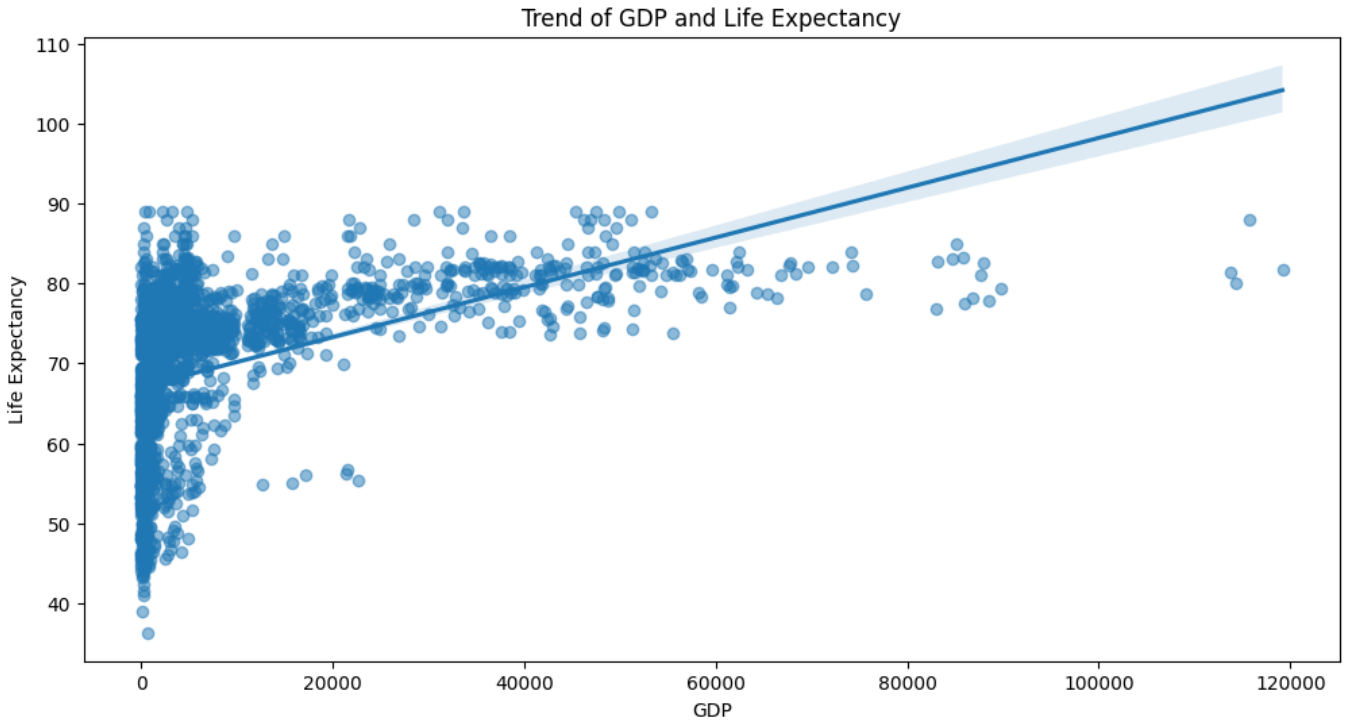
3. 완료한 시각화와 (1)의 가설/질문에 대한 결론

소득 수준(GDP)과 기대 수명(Life Expectancy) 간의 관계를 시각화하여 소득 수준이 높은 국가일수록 기대 수명이 높은지 확인해봤습니다.

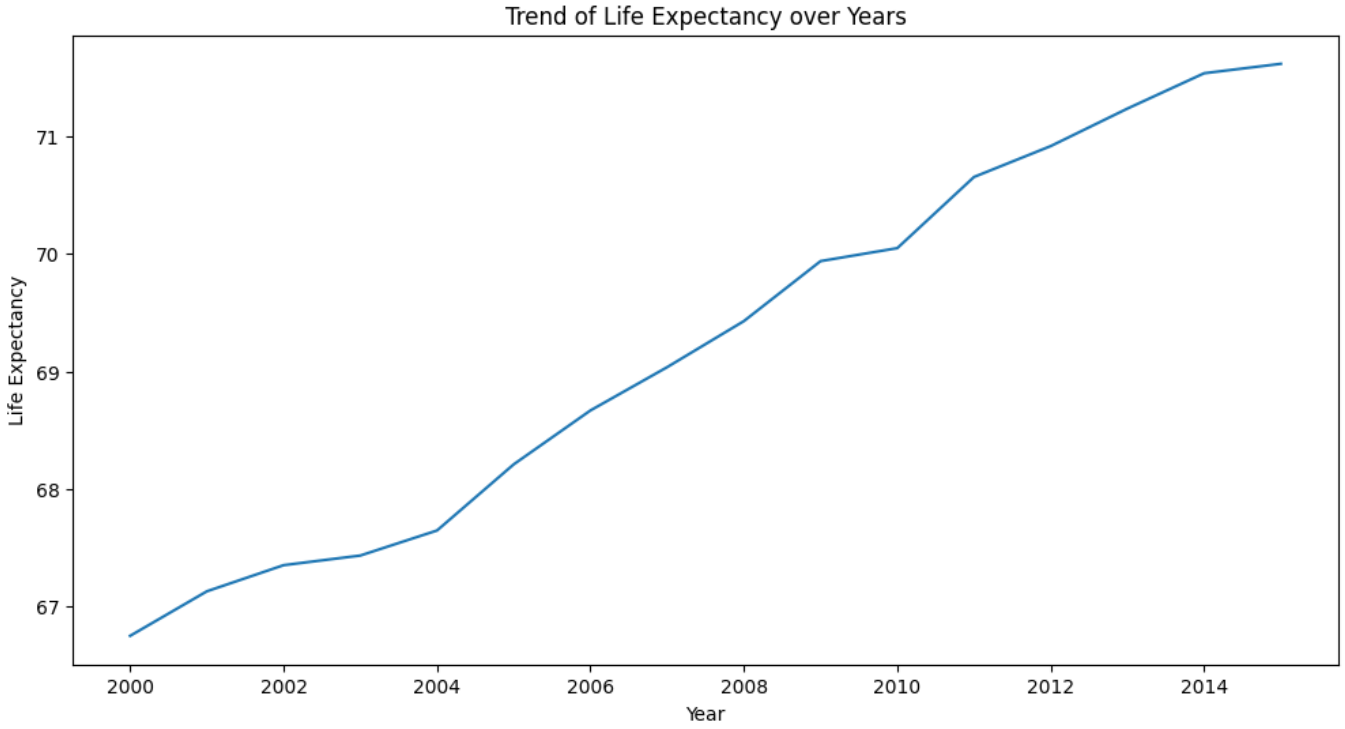
시각화 1: 소득 수준과 기대 수명 간의 관계 (Scatterplot)



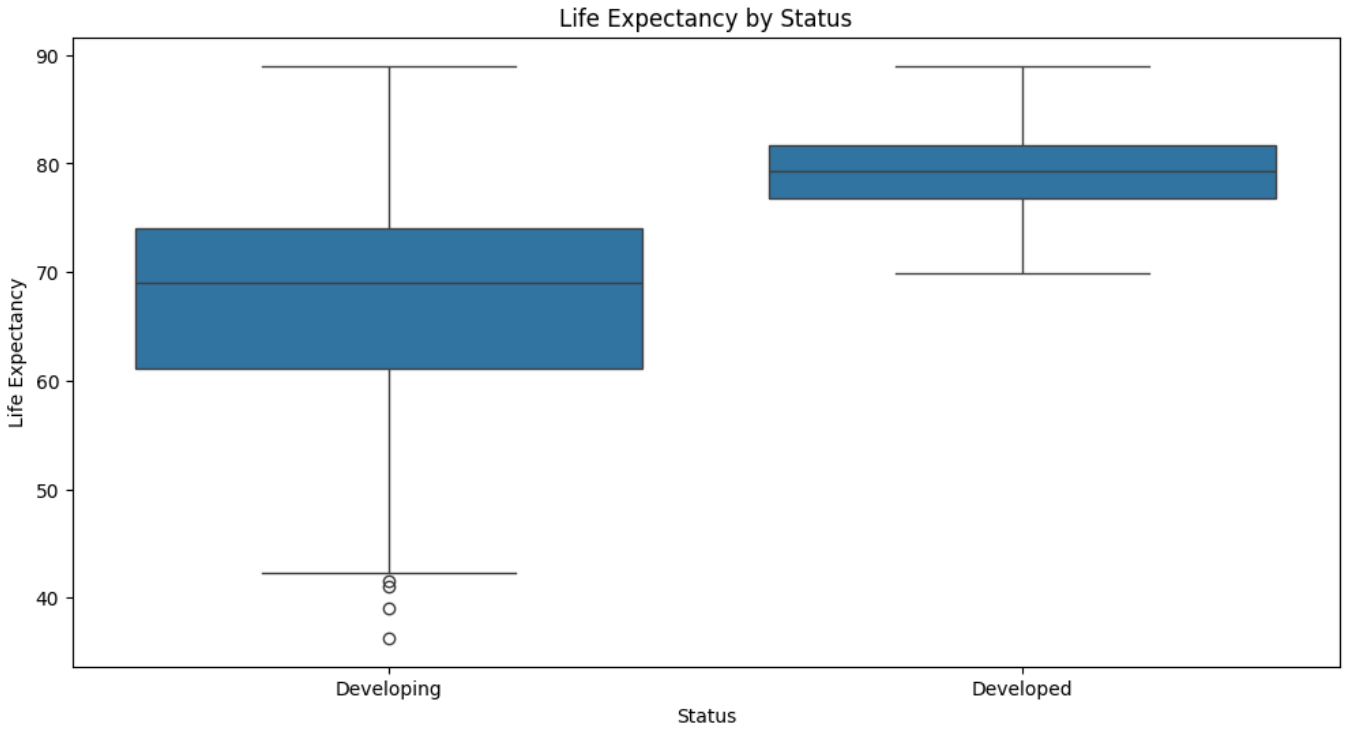
시각화 2: 소득 수준과 기대 수명 간의 추세선 (Regression Line)



시각화 3: 연도별 기대 수명 변화 (Line Plot)



시각화 4: 국가 개발 상태에 따른 기대 수명 (Boxplot)



4. (3)을 기반으로 시각화에서 얻을 수 있는 인사이트

1. **긍정적 상관관계**: 소득 수준이 높은 국가일수록 기대 수명이 높은 경향을 보입니다. 이는 Scatterplot과 Regression Line에서 두 변수 간의 양의 상관관계를 통해 확인할 수 있습니다. 소득 수준이 높을수록 기대 수명도 높아집니다.
2. **정책적 인사이트**: 소득 증대가 건강 증진에 중요한 요소로 작용할 수 있으며, 이를 바탕으로 정책 방향을 설정할 수 있습니다. 소득 수준 향상이 기대 수명을 높일 수 있는 중요한 요인임을 시사합니다.
3. **연도별 변화**: 기대 수명이 시간이 지남에 따라 전반적으로 증가하는 추세를 보입니다. 이는 글로벌 건강 상태의 개선과 의료 기술의 발전을 반영할 수 있습니다.
4. **Status에 따른 차이**: 개발도상국과 선진국 간의 기대 수명 차이를 통해, 경제적 발전이 건강에 미치는 영향을 평가할 수 있습니다. 선진국이 개발도상국에 비해 전반적으로 높은 기대 수명을 보입니다.