



report

1. Top Hits Spotify

[물음1] 1번 시각화의 목적과 효과 평가 및 개선점 제안

시각화의 목적 평가

1. **데이터 이해 증진:** Spotify 음악 요소 간의 상관관계를 시각화하여 복잡한 관계를 더 쉽게 이해할 수 있도록 하는 데 목적이 있습니다. 이를 통해 특정 음악 요소가 다른 요소와 어떻게 관련되어 있는지, 예를 들어 템포와 에너지가 높은 상관관계를 가지고 있는지를 직관적으로 파악할 수 있습니다.
2. **의사 결정 지원:** 상관관계 시각화는 음악 제작자나 마케팅 담당자가 특정 요소가 청취자 반응에 미치는 영향을 이해하는 데 도움을 줄 수 있습니다. 예를 들어, 감정과 에너지가 높은 상관관계를 가진 곡이 더 인기가 있다면, 이를 바탕으로 음악 제작 전략을 세울 수 있습니다.

시각화의 효과 평가

1. **통찰력 제공:** 시각화는 숫자 데이터의 관계를 시각적으로 표현하여 데이터에서 중요한 통찰력을 빠르게 얻을 수 있게 해줍니다. 히트맵을 통해 상관관계가 강한 요소들을 쉽게 식별할 수 있습니다.

2. **커뮤니케이션 개선:** 다양한 이해관계자에게 데이터의 관계를 명확하게 전달할 수 있는 도구로 작용합니다. 특히, 비전문가도 쉽게 이해할 수 있도록 도와줍니다.
3. **Spotify 음악 요소 간의 관계 파악:** 히트맵은 각 요소들 간의 상관관계를 색상과 숫자로 직관적으로 표현하여, 청취자의 선호도와 관련된 요소들을 한눈에 파악할 수 있도록 돕습니다.

개선점 제안

1. **색상 팔레트 최적화:** sns.diverging_palette를 사용하여 양의 상관 관계를 파란색, 음의 상관 관계를 빨간색으로 시각화했습니다. 색상의 채도는 상관 계수의 절대값에 따라 조절되어 상관 관계의 강도를 명확하게 표현합니다.
2. **마스킹과 대각선 처리:** 대각선 요소는 항상 자기 자신과의 상관 관계(상관 계수 1)를 나타내므로 중요도가 낮습니다. 이를 시각적으로 덜 강조하기 위해, 대각선 요소에 마스킹을 적용하여 하얀색으로 처리했습니다.
3. **데이터 값 단순화:** annot=True와 fmt=".2f"를 설정하여 소수점 둘째 자리까지 상관 계수 값을 표시하는 것으로 단순화하여 복잡한 히트맵에서의 가독성을 개선하였습니다.
4. **제목 추가:** 그래프에 적절한 제목을 추가하여 시각화의 목적과 내용을 명확히 전달하였습니다. 이는 시각화의 맥락을 제공하고, 사용자에게 현재 보고 있는 데이터가 무엇인지를 쉽게 이해할 수 있게 합니다.

[물음2] 2번 시각화의 목적과 효과 평가 및 개선점 제안

시각화의 목적과 효과

목적

1. **데이터 이해 증진:** instrumentalness와 popularity 사이의 관계를 시각적으로 표현하여 데이터의 분포와 상관관계를 이해하고자 합니다.
2. **특성 비교:** duration_ms와 explicit 등의 추가적인 변수와 popularity 간의 상관관계를 탐색하기 위한 목적으로, 이러한 변수가 다른 변수들에 미치는 영향을 시각적으로 확인할 수 있습니다.

효과

1. **다차원 데이터 표현:** 색상(hue)과 크기(size)를 사용하여 다차원 데이터를 하나의 시각화에 통합함으로써, 복잡한 데이터 세트를 한눈에 파악할 수 있습니다.

2. **상관관계 및 트렌드 파악:** instrumentality와 popularity 사이의 트렌드를 파악하고, 특정 특성(예: duration_ms나 explicit 여부)에 따라 분포가 어떻게 달라지는지 분석할 수 있습니다.

개선점 제안

1. **데이터 포인트 겹침 방지- 반투명 적용:** 크기가 작은 데이터 포인트가 겹쳐서 보이지 않는 경우가 발생할 수 있습니다. 이를 해결하기 위해 데이터 포인트에 반투명도를 적용하면 포인트가 겹치더라도 모두 보이게 할 수 있습니다.
2. **컬러 팔레트 최적화:** 이전 시각화는 모호한 색깔 구성을 하였고, 정도를 범주화 시켜서 나타내었기 때문에 차이 비교가 힘듭니다. 따라서 색깔 구성을 바꾸고 정도에 따라 표시되게 형태를 바꾸었습니다.
3. **라벨링 추가:** 색상 바에 라벨을 추가하여 새로 바뀐 색상 바가 나타내는 의미를 명확히 하였습니다.
4. **범주형 변수:** explicit 변수는 바이너리한 범주형이므로 모양을 달리하여 명확하게 비교될 수 있게 하였습니다.

물음 3: Explicit 여부가 Popularity에 미치는 영향 평가

- Explicit 여부가 음악의 인기에 미치는 영향을 판단하기 위해, explicit 변수를 기준으로 나누어 각 그룹의 popularity 평균과 분포를 박스플롯을 사용하여 비교하였습니다.
- 박스플롯은 explicit 여부에 따른 popularity의 분포를 시각화한 것으로 각 박스는 데이터의 중간값, 사분위수, 그리고 이상치를 보여줍니다.

시각화 결과, 두 그룹의 중간값과 분포가 유사하여 큰 영향을 미치지 않는다고 판단하였습니다.

물음 4: spotify 데이터를 시각화하여 내릴 수 있는 결론

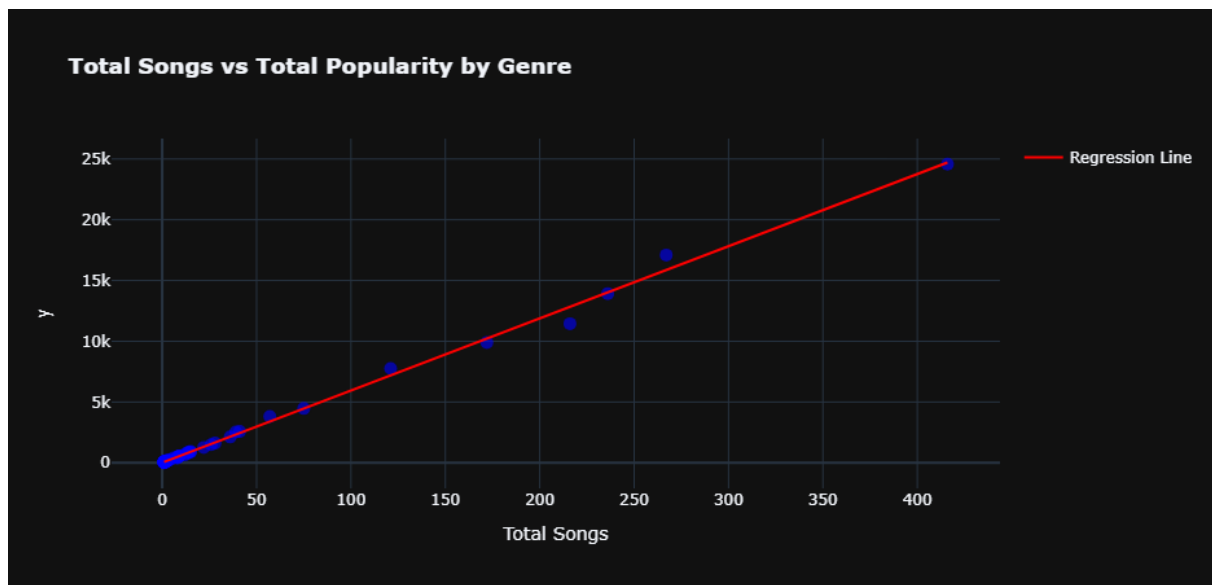
- **음악 요소 간의 상관관계:** 1번 시각화에서 상관관계 분석을 통해 특정 음악적 요소들이 서로 어떻게 연관되어 있는지 파악할 수 있습니다. 예를 들어, loudness는 energy와 큰 양의 상관관계를 보이는 반면, acousticness와는 큰 음의 상관관계를 보입니다. 이는 더 큰 소리를 내는 음악일수록 에너지가 높고, 어쿠스틱한 성질은 낮다는 것을 의미합니다. popularity에 직접적 영향이 있는 요소는 없었습니다.
- **장르 분포:** 장르 시각화를 통해 Spotify 데이터에서 총 음악의 장르별 갯수와 인기 있는 장르를 파악할 수 있었습니다.

- **Popularity에 영향을 미치는 요소:** 2번 시각화를 통해 popularity에 영향을 미치는 다양한 요소들을 확인할 수 있었습니다. popularity에 영향을 끼치는 요소는 없었습니다.

물음5: 물음 4의 결론과 정당화

- total songs based on genres, popular genres based on popularity 두가지의 시각화가 있었는데,
특정 장르의 곡 수가 많을수록 그 장르의 총 popularity도 높아지는지 궁금하였습니다.
- 따라서 Total Songs vs Total Popularity by Genre 산점도를 그려, 각 장르의 곡 수와 총 popularity를 함께 나타내고 두 변수 간의 관계를 시각화하였습니다.

시각화: 장르별 곡 수와 총 인기도 비교



- Total Songs vs Total Popularity by Genre 산점도는 각 장르의 곡 수와 총 popularity를 한 화면에 나타내어, 두 변수 간의 관계를 명확히 보여줍니다.
- 이 그래프에는 회귀선이 추가하여 두 변수 간의 경향성을 시각적으로 확인하였습니다. 회귀선의 기울기가 양수라면, 곡 수가 많을수록 인기도도 높아지는 경향을 시사합니다.

결론의 정당화:

- 시각화 결과와 회귀선 분석을 통해, 장르별 곡 수와 인기도 간의 상관관계가 확인되었습니다. 이는 인기 있는 장르일수록 해당 장르의 곡이 많이 생산되고, 반대로 곡 수가 많은 장르일수록 인기도가 높아질 가능성이 있다는 결론을 내릴 수 있습니다.

2. Life Expectancy

가설

독립변수: 종속변수를 제외한 변수들

종속변수: Life expectancy

다양한 요인들은 Life expectancy에 긍정, 부정적인 영향이 있거나 아무 영향도 끼치지 못하거나 혼합된 영향을 끼칠 것이다.

긍정적일 것 같은 요소

1. **GDP:** 경제적으로 발전한 국가들은 더 나은 보건 및 의료 서비스를 제공할 수 있기에 기대 수명이 높을 것이다.
2. **Health Expenditure:** 보건 의료에 대한 지출이 많을수록, 의료 서비스에 대한 접근성이 향상되기 때문에 기대 수명이 높을 것이다.
3. **Income Composition of Resources:** 자원의 소득 구성은 일반적으로 경제적 안정성을 나타내며, 이는 기대 수명에 긍정적인 영향을 미칠 것이다.
4. **Schooling:** 교육 수준이 높을수록 건강에 대한 인식과 의료 접근성이 향상될 수 있어 기대 수명에 긍정적인 영향을 미칠 수 있다.

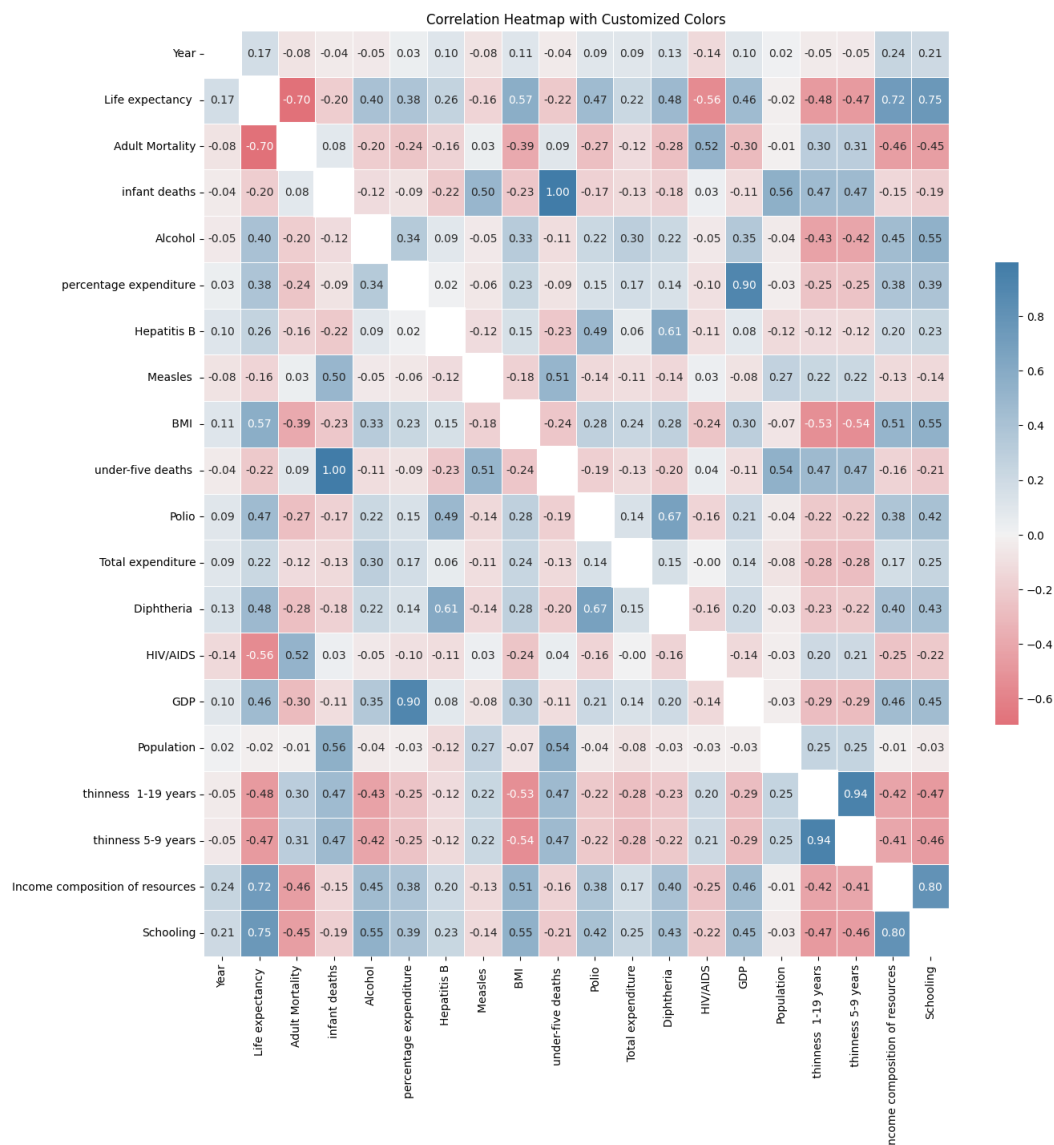
부정적일 것 같은 요소

1. **Adult Mortality:** 성인 사망률이 높을수록 기대 수명이 짧을 것이다.
2. **Infant Deaths:** 유아 사망률이 높을수록 기대 수명이 짧을 것이다.
3. **HIV/AIDS:** HIV/AIDS의 유병률이 높을수록 기대 수명이 짧을 것이다.
4. **Alcohol:** 높은 알코올 소비는 건강 문제를 일으킬 수 있으며, 이는 기대 수명에 부정적인 영향을 미칠 것이다.
5. **Diphtheria:** 디프테리아와 같은 전염병의 발생은 기대 수명에 부정적인 영향을 미칠 것이다.
6. **Measles:** 홍역과 같은 질병의 발생률이 높으면 기대 수명이 짧을 것이다.
7. **Polio:** 소아마비와 같은 질병의 발생은 기대 수명에 부정적인 영향을 미칠 것이다.
8. **Under-Five Deaths:** 5세 이하 사망률이 높을수록 기대 수명이 짧을 것이다.
9. **Thinness 1-19 years:** 1-19세의 저체중 비율이 높을수록 기대 수명이 짧을 것이다.
10. **Thinness 5-9 years:** 5-9세의 저체중 비율이 높을수록 기대 수명이 짧을 것이다.

상관없음 또는 혼합된 영향일 것 같은 요소

1. **Percentage Expenditure:** 전체 지출의 비율이 기대 수명에 미치는 영향은 관련없을 것이다.
2. **BMI:** 평균 체질량지수(BMI)와 기대 수명 간의 관계는 국가의 식습관, 문화에 따라 다를 것이다.
3. **Total Expenditure:** 총 지출은 경제적 발전을 나타낼 수 있으나, 지출이 보건 분야에 얼마나 사용되는지에 따라 다를 수 있다.
4. **Population:** 인구 규모가 기대 수명에 미치는 영향은 상대적인 인구 밀도와 보건 인프라에 따라 달라질 수 있다.

시각화 결과



시각화에서 얻을 수 있는 인사이트

강한 양의 상관관계(corr값 0.5 이상)

1. Income composition of resources
2. Schooling
3. BMI

약한 양의 상관관계(corr값 0.2 이상 0.5미만)

1. Alcohol percentage
2. expenditure
3. Hepatitis B
4. Polio
5. Total expenditure
6. Diphtheria
7. GDP
8. Year

강한 음의 상관관계(corr값 -0.4 이하)

1. Adult Mortality
2. HIV/AIDS
3. thinness 1-19 years
4. thinness 5-9 years

약한 음의 상관관계(corr값 -0.2 이하 -0.4초과)

1. infant deaths
2. Measles
3. under-five deaths

상관없음 또는 혼합된 영향

1. Population

가설에서의 예상과는 다르게 영향을 끼친 요소들이 있었음을 알 수 있습니다.