1. Spotify

[물음1]

<시각화 목적>

변수들간 상관관계가 어느 정도 되는지 파악하기 위해서 시각화를 진행함

<시각화 효과>

상관계수를 한 눈에 알아볼 수 있게 모든 변수들간 상관관계 정도를 그라데이션으로 제시하여 직관적으로 알아볼 수 있음

<적용해볼 수 있는 개선점>

1. 상관관계 분석에 적합하지 않은 확률변수 제거가 필요해보임. Song, artist, genre는 텍스트 데이터로 상관관계 분석에 직접적으로 활용하기 어렵고, explicit와 mode는 각각 텍스트, 숫자형 데이터이나 모두 범주형 데이터이므로 이 역시 상관관계 분석에 직접적으로 활용하기에 어렵기에 제거하는 것이 적절해보임.

2. 전체 데이터 중 이상치를 제거하는 작업이 필요해보임. 상관관계 분석에서 이상치가 큰 영향을 미칠 수 있으므로 이상치를 제거하는 작업을 선행한 후 분석에 들어감이 적절함. 이번 과제에서 적용한 이상치 탐지 기준은 z-score을 활용함.

[물음 2]

<산점도 목적>

변수 간 상관관계를 직관적으로 볼 수 있게끔 시각화하는 것이 목적

<산점도 효과>

데이터가 어떻게 분포되어있는지 데이터 형태를 쉽게 파악해볼 수 있음

<적용해볼 수 있는 개선점 >

1. 변수 4개를 한꺼번에 상관관계 분석하기 어려우며, explicit는 범주형 데이터이기 때문에 True/False인 데이터를 선제적으로 나눈 뒤 분석한다.

2. 추세선을 삽입하여 상관관계를 보다 명확하게 파악할 수 있도록 한다.

[물음 3]

Explicit이 True, False의 범주형 데이터이므로 True와 False에 해당하는 데이터로 분리한 뒤, 각 그룹의 Popularity의 평균간 유의미한 차이가 있는지 t-test를 통한 가설 검정으로 접근.

H0: Explicit True, False 데이터 간 Popularity avg에 유의미한 차이가 없다.

H1: Explicit True, False 데이터 간 Popularity avg에 유의미한 차이가 있다.

이에 대한 t-통계량, p-value는 각각 2.085, 0.037로 계산됨.

유의 수준 0.05보다 p-value가 작기 때문에 H0을 기각, 다시 말해 유의미한 차이가 있음을 알 수 있음.

다시 말해서, popularity에 대해 explicit가 영향을 준다라고 판단할 수 있다.

[물음4]

Popularity는 다른 변수들간 상관관계가 거의 없다시피하는 변수이지만, 범주형 데이터인 explicit를 기준으로 True, False로 전체 데이터셋을 분류한뒤 확인해보면, loudness와 duration_ms가 각 노래 성향별로 인기도와 의미있는 관계가 있음을 짐작할 수 있다.

이는 산점도를 통해서도 확인해볼 수 있다.

2. Life Expectancy

1. 검증/답하고자 하는 가설 혹은 질문

인적 자원 개발이 기대 수명 증진에 영향을 미칠 수 있는가?

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

독립변수 – Schooling

종속변수 – Life Expectancy

데이터 특성 – Schooling, Life Expectancy 모두 float 타입의 데이터

3. 완료한 시각화와 (1)의 가설/질문에 대한 결론

인적 자원 개발이 기대 수명 증진에 영향을 미친다.

4. (3)을 기반으로 시각화에서 얻을 수 있는 인사이트

주어진 데이터에는 결측치, 오기입 등을 비롯한 그다지 좋지 못한 품질을 보이는 데이터이지만, 상관관계 분석을 기반으로 데이터를 보면 기대 수명 연장에는 꽤나 많은 요소들이 복합적으로 작용하는 결과물이라는 것을 알 수 있다. 정부 차원에서 사회 전체적 교육, 보건 등 다방면적 개선을 위한다면 결국에는 세수를 충분히 확보하는 것이 선제적으로 해결해야하고, 이는 곧 국가의 경제 성장이 수반되어야함을 의미한다. 즉, 국민 기대 수명 연장을 목표로 한다면 경제 발전을 우선 순위로 둬야하는 것이 선제적이어야한다고 생각한다.