다변량데이터분석

과제 1

2021150456 이예지

목차

[Q1] 본인이 스스로 Multiple Linear Regression 모델을 적용하는데 적합한 데이터셋을 선정하고 선정 이유를 설명하시오. 데이터셋 탐색은 아래에서 제시된 Data Repository 를 포함하여 여러 Repository 를 검색해서 결정하시오. 보고서에는 데이터를 다운로드할 수 있는 링크를 반드시 제공하시오.

Multiple Linear Regression 모델 적용을 위해 선정한 데이터셋은 Kaggle 의 Song Popularity Dataset 이다. 해당 데이터셋은 총 15 개의 변수와 18835 개의 관측치를 가지고 있다.
몇 년 전, 앨범을 발매한 적 있는 가수들만 참가할 수 있는 오디션 프로그램인 '싱어게인'을 종종 보곤 했다. 대부분의 참가자들이 뛰어난 실력을 보여주었지만 '싱어게인'에 나오기 전까지는 그들을 아는 사람이 많지 않았다. 왜 이런 일이 일어나는 것일까? 대부분의 경우, 그들의 노래가 소위 말해 히트를 치지 못했던 것이다. 그렇다면 '어떤 노래가 대중적으로 인기가 있을까?'라는 의문에 도달하게 된다. 이러한 의문을 해결하기 위해 Kaggle 에서 제공하는 Song Popularity Dataset 을 분석해보고자 한다.
이에 더해, Multiple Linear Regresison 모델을 적용하기 위해서는 (1) 종속변수가 연속형이어야 하고, (2) 종속변수와 설명변수의 집합 사이의 관계가 선형 결합으로 표현될 수 있음을 가정해야 한다.
선정한 데이터셋의 종속변수는 song_popularity 로, 연속형 변수임을 확인할 수 있다. 종속변수와 설명변수의 집합 사이의 관계가 선형 결합으로 표현될 수 있는지는 아직 확인할 수 없지만, 분석을 위해 선형 결합으로 표현될 수 있다고 가정하기로 한다. 따라서 다중선형회귀분석을 위한 가정들을 만족하므로 다중선형회귀모델을 적용하는데 적합하다고 판단했다.

데이터셋 다운로드 링크: https://www.kaggle.com/datasets/yasserh/song-popularity-dataset

[Q2] 해당 데이터셋의 종속변수와 설명변수는 어떤 것들이 있는가? 분석 전에 아래 세 가지 질문에 대해서 스스로 생각해보고 답변을 하시오.

데이터셋의 종속변수는 song_popularity, 설명변수는 그 외 나머지 변수들이다. 변수들에 대한 설명은 아래와 같다.

변수명	변수 설명
song_name **(설명변수)**	음원 제목
song_popularity **(종속변수)**	음원 흥행 정도. 노래의 재생 횟수, 스트림 수를 기반으로 계산된 백분율 점수이다.
song_duration_ms **(설명변수)**	음원의 길이를 milliseconds 단위로 나타낸 것
acousticness **(설명변수)**	음원이 얼마나 음향적으로 가공되지 않았는지(어쿠스틱인지)를 나타내는 지표. [0, 1]의 값을 가지며, 1 에 가까울수록 어쿠스틱임을 나타낸다.

danceability **(설명변수)**	템포, 리듬 안정성, 비트 강도 및 전반적인 규칙성을 포함한 음악적 요소의 조합에 기반하여 음원이 얼마나 춤에 적합한지를 나타내는 지표. [0, 1]의 값을 가지며, 1 에 가까울수록 춤을 추기에 적합하다는 것을 나타낸다.
energy **(설명변수)**	강렬함과 활동성의 척도. [0, 1]의 값을 가지며, 1 에 가까울수록 더욱 강렬하고 높은 활동성을 지니고 있다는 것을 나타낸다. 에너지는 활발함, 소리의 세기, 음색 등을 포함한다.
instrumentalness **(설명변수)**	음원이 사람의 목소리를 포함하고 있는지 예측하는 지표. "오", "아" 같은 소리는 목소리가 아닌 악기로 처리된다. 값이 1 에 가까울수록 음원에 목소리가 나오지 않을 가능성이 높다. 값이 0.5 를 넘으면 악기의 음원을 나타내도록 의도되었다.
key **(설명변수)**	트랙이 속한 음조. 표준 음계 표기법을 사용하여 정수를 음계와 매칭하였다. 예를 들어 0 은 도, 1 은 도#(레♭), 2 는 레 등이다. 음조가 감지되지 않은 경우 -1 이 입력되어 있다.
liveness **(설명변수)**	녹음 과정에서의 관객 존재 여부. 값이 높을수록 생방송의 음원일 가능성이 높아진다. liveness>0.8 이면 생방송 음원일 가능성이 매우 높다.
loudness **(설명변수)**	데시벨로 표시된 음원의 전반적인 음량. 음량은 전체 음원을 기준으로 평균화되며, 음원 간의 상대적인 음량을 비교하는데 유용하다. 일반적으로 [-60, 0] 사이의 값을 갖는다.
audio_mode **(설명변수)**	선율이 유래된 음계의 조성(장조 또는 단조). 장조는 1 로, 단조는 0 으로 표시된다.
speechiness **(설명변수)**	음원 내 말하는 소리의 존재 여부. 녹음물이 말하는 소리에 가까울수록(토크쇼, 오디오북, 시 등), 속성값은 1.0 에 가까워진다. speechiness>0.66 이면, 아마도 완전히 말로 이루어진 음원을 나타낸다. [0.33, 0.66]의 값은 음악과 말 둘 다 포함되어 있을 수 있는 음원을 나타낸다. 이 경우 랩 음악과 같은 경우가 포함된다. speechiness<0.33 이면, 음악 및 기타 비언어적 트랙을 나타낼 가능성이 높다.
tempo **(설명변수)**	분당 비트수(BPM)로 표현된 음원의 전반적인 추정 박자. 음악 용어에서, 박자는 주어진 부분의 속도이며, 평균 비트 지속 시간에서 직접 파생된다.
time_signature **(설명변수)**	추정된 박자표. 박자표는 매 마디에 비트가 얼마나 있는지를 나타내는 악보 표기 규칙이다. time_signature 는 "3/4"부터 "7/4"까지의 범위로 표시되며, 각각 3 부터 8 까지의 숫자로 표현된다.
audio_valence **(설명변수)**	음원이 전달하는 음악적 긍정성을 나타내는 지표. [0.0, 1.0]의 값을 가진다. 값이 높은 음원일수록 더 긍정적이고(행복, 즐거움, 환희 등), 값이 낮을수록 더 부정적이다(슬픔, 우울, 화남 등).

1) 이 데이터는 종속변수와 설명변수들 사이에 실제로 “선형 관계”가 있다고 가정할 수 있겠는가? 가정할 수 있음/없음 판단에 대한 본인의 생각을 서술하시오.

종속변수와 설명변수들 사이에 선형관계가 있을 것이라고 가정하기 어렵다. 빌보드 차트나 멜론 차트를 보면, 상위권에는 특정 장르만 존재하는 것이 아니라, 다양한 장르가 공존하고 있다. 장르가 다르면 ‘acousticness’, ‘danceability’, ‘energy’, ‘instrumentalness’, ‘speechiness’, ‘tempo’와 같은, 멜로디와 관련이 높은 변수들의 값에 차이가 있을 것이다. 하지만 종속변수의 값은 비슷한 정도로 측정될 것이다. 또, 장르가 비슷하면 ‘acousticness’, ‘danceability’, ‘energy’, ‘instrumentalness’, ‘speechiness’, ‘tempo’와 같은, 멜로디와 관련이 높은 변수들의 값의 차이가 크지 않을 것이다. 그러나 비슷한 장르의 노래이더라도 각각의 흥행 성적은 천차만별이다. 게다가 ‘song_name’, ‘song_duration_ms’을 생각해보면, 음원의 제목으로 음원의 흥행 성적이 차이가 나지는 않을 것이며, 음원의 길이 역시 음원의 흥행 성적과 큰 관련이 없어 보인다. 따라서 설명변수들과 종속변수 사이에 선형관계가 없을 것이라고 판단했다.

2) 이 데이터에서 제공된 설명변수들 중에서 높은 상관관계가 있을 것으로 예상되는 변수들을 어떤 것들이 있는가? 왜 그렇게 생각하는가?

종속변수와 높은 상관관계가 있을 것으로 예상되는 설명변수는 없다. 설명변수들은 음원의 흥행 성적을 예측하는 것보다 음원의 장르가 무엇인지 예측하는 것에 더 적합하다고 생각한다. 인기 있는 음원을 보면, 하나의 장르만 존재하는 것이 아니다. 아이돌 음악, 발라드, 랩 등 다양한 장르가 존재하고, 사람들의 관심을 끌지 못한 음악에도 아이돌 음악, 발라드, 랩 등의 장르가 모두 존재한다. 따라서 각 장르 내에서는 설명변수들이 비슷한 값을 가질 수 있지만, 그것이 음원의 흥행 성적을 설명한다고 보기는 어려울 것이라 판단했다.

추가적으로 설명변수 간 상관관계를 생각해보면, {acousticness, energy}, {danceability, energy}, {danceability, tempo}, {energy, tempo}, {energy, loudness}, {liveness, energy}, {audio_mode(혹은 key), audio_valence} 변수들 사이에 높은 상관관계가 있을 것으로 예상된다.

- {acousticness, energy}: 일반적으로 어쿠스틱 곡으로 분류되는 노래들은 전자음이 거의 들어가지 않는다. 또한, 노래를 파워풀하게 부르는 대신 부드럽고 잔잔하게 부르고, 미성을 지닌 경우가 많다. 따라서 ‘acousticness’가 높으면, 즉 어쿠스틱한 곡일수록 ‘energy’가 낮을 것이라 판단했다.
- {danceability, energy}: 대개 춤은 높은 활동성을 요하며, ‘danceability’는 춤을 추기에 적합할수록 높은 값을 가져 ‘energy’와 높은 양의 상관관계를 지녔을 것으로 생각된다.
- {danceability, tempo}: ‘danceability’가 고려하는 다양한 음악적 요소들 중 ‘tempo’가 포함되어 있어, 둘 사이에 높은 상관관계가 있을 것이라 판단했다.

- {energy, tempo}: 템포가 빠를수록 음악이 활발해진다. 따라서 'tempo'와 'energy' 사이에 높은 상관관계가 존재할 것이라 생각된다.
- {energy, loudness}: 'energy'의 값을 측정하는데 있어 활발함, 소리의 세기 등이 고려되는데, 'loudness'는 음악의 전반적인 음량을 고려한다. 음악이 활발하면 전반적으로 큰 소리들이 지속되고, 이로 인해 'energy'의 값이 증가할수록 'loudness'의 값도 증가해 두 변수 사이에 높은 상관관계가 있을 것이라 생각했다.
- {liveness, energy}: 'liveness'는 녹음 과정에서의 관객의 존재 여부를 판단하는 지표이며, 값이 높을수록 생방송 음원일 가능성이 높다고 했는데, 이는 관객들의 함성 소리 등이 녹음 과정에서 포함되었기 때문으로 보인다. 이러한 소리들은 활발한 소리로 분류되어 'energy'의 값을 증가시키는 방향으로 영향을 미칠 수 있기 때문이다.
- {audio_mode(혹은 key), audio_valence}: 일반적으로 장조는 밝고 즐거운 분위기를, 단조는 어두운 분위기를 나타낸다고 알려져있다. 따라서 'audio_mode'와 음악을 긍정성을 나타내는 'audio_valence' 사이에 상관관계가 있을 것으로 판단했다.

3) 제공된 설명변수들 중에서 종속변수를 예측하는데 필요하지 않을 것으로 예상되는 변수들은 어떤 것들이 있는가? 왜 그렇게 생각하는가?

'song_name', 'song_duration_ms', 'audio_mode', 'tempo' 변수가 필요하지 않을 것으로 예상된다.

- 'song_name': 음원의 제목 그 자체로는 음원의 흥행에 있어 큰 영향을 미치지 못할 것이라고 생각하기 때문이다.
- 'song_duration_ms': 재생할 음원을 고르는데 있어 음원의 길이를 고려하는 경우는 거의 없어, 음원의 흥행 정도를 판단하는데 있어 유의미한 지표가 되지 못할 것으로 판단했다.
- 'audio_mode': 'key' 변수는 음조를 나타내는데, 이는 이미 장조와 단조의 구분을 포함한 개념이다. 따라서 'audio_mode' 변수는 새로운 정보를 제공하는 것이 아닌, 중복된 정보를 제공해 종속변수를 예측하는데 있어 그 효과가 크지 않을 것으로 보인다.
- 'tempo': 'danceability' 변수에서 다양한 음악적 요소의 조합을 고려하는데, 그 고려 대상 중 하나로 템포가 있다. 또, 템포는 곡의 빠르기를 결정하는데, 곡의 빠르기는 음원의 활발함과 관련이 있고, 이는 'energy'에서 고려되는 요소 중 하나이다. 따라서 'tempo'변수가 이미 다른 변수에서 고려되고 있으므로 종속변수를 예측하는데 있어 추가적으로 사용할 필요가 없다고 판단했다.

[Q3] 개별 입력 변수들에 대하여 각각 다음과 같은 단변량 통계량을 계산하고 Box plot 을 도시하시오: Mean, Standard deviation, Skewness, Kurtosis. 전체 변수

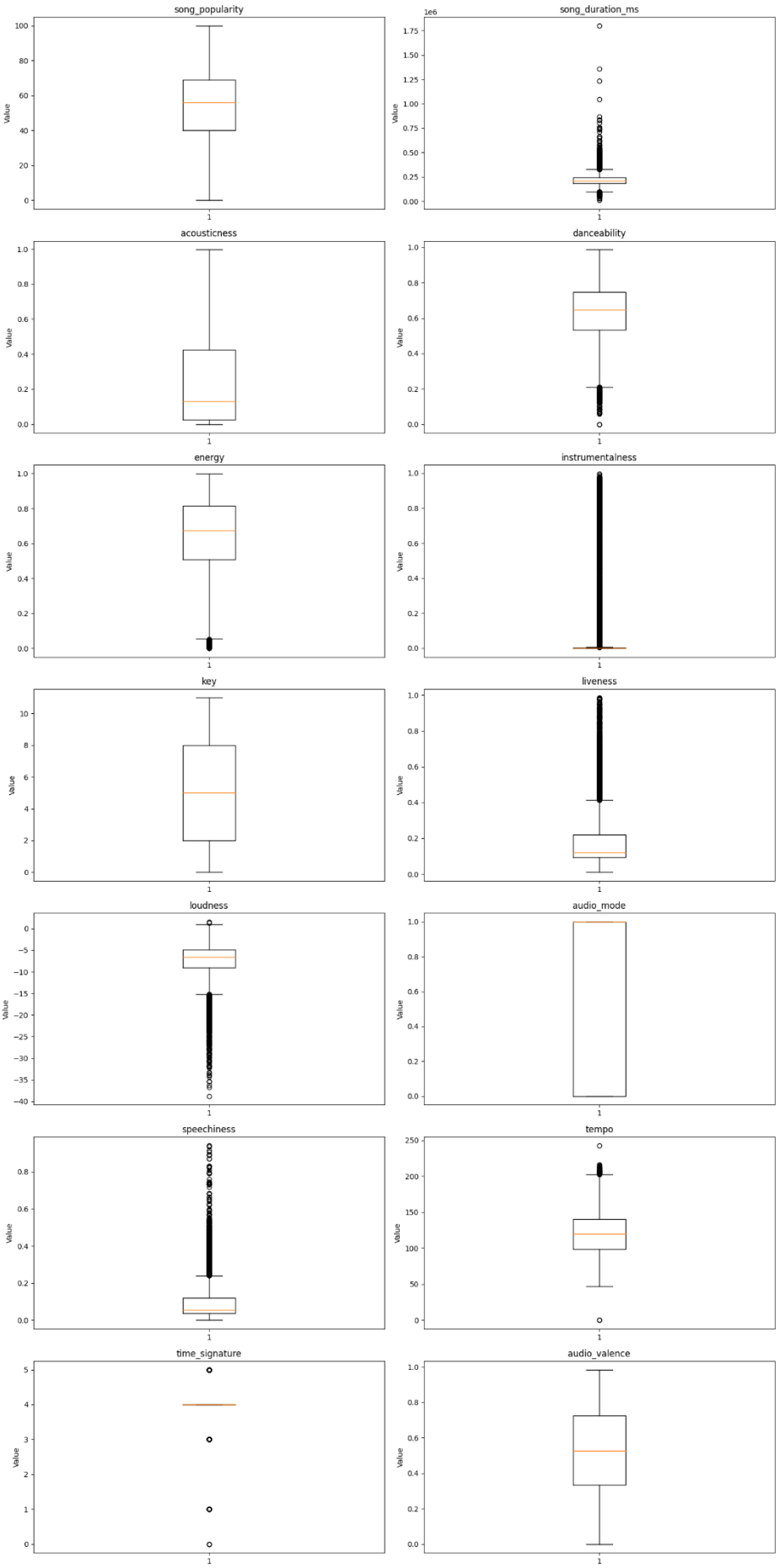
중에서 정규분포를 따른다고 할 수 있는 변수들은 몇 개인가? 정규분포를 따른다고 가정한 근거는 무엇인가?

종속변수를 예측하는 것보다 각 관측치를 구분하기 위한 id 개념에 더 가까운 'song_name'은 제거한 후 진행하도록 한다. 나머지 설명변수들 중 범주형 변수는 존재하지 않으므로 1-of-C coding 은 적용하지 않는다.

개별 입력 변수들에 대한 단변량 통계량을 계산하면 다음과 같다.

	Mean	Standard deviation	Skewness	Kurtosis
song_popularity	52.991877	21.905073	-0.501448	-0.169377
song_duration_ms	218211.587576	59885.950751	3.257218	46.696232
acousticness	0.258539	0.288711	1.071079	-0.096569
danceability	0.633348	0.156719	-0.391688	-0.075095
energy	0.644995	0.214095	-0.620688	-0.138156
instrumentalness	0.078008	0.221585	2.984939	7.561338
key	5.289196	3.614499	-0.002520	-1.311436
liveness	0.179650	0.143980	2.215246	5.788064
loudness	-7.447435	3.827730	-1.929357	6.520430
audio_mode	0.628139	0.483302	-0.530266	-1.718818
speechiness	0.102099	0.104376	2.270837	6.502932
tempo	121.073154	28.713693	0.442819	-0.217777
time_signature	3.959119	0.298525	-4.978946	45.541917
audio_valence	0.527967	0.244625	-0.016422	-0.977730

개별 입력 변수들의 box plot 시각화 결과는 다음과 같다.


song_popularity
song_duration_ms
1e6
acousticness
danceability
energy
instrumentalness
key
liveness
loudness
audio_mode
speechiness
tempo
time_signature
audio_valence
Value
1

정규분포를 가정하기 위해서 왜도와 첨도(본 분석에서는 첨도를 구하기 위해 Fisher 의 정의를 사용한다.)가 모두 0 이어야 하지만, 보편적으로는 왜도의 절댓값이 3 보다 작고, 첨도의 절댓값이 8 보다 작으면 정규분포를 따른다고 해석할 수 있다. 따라서 정규분포를 따른다고 할 수 있는 변수들은 'song_popularity', 'acousticness', 'danceability', 'energy', 'instrumentalness', 'key', 'liveness', 'loudness', 'audio_mode', 'speechiness', 'tempo', 'audio_valence'로, 총 12 개이다.

Box plot 을 확인해보았을 때, 데이터들이 정규분포를 따르면 이상치가 없을 것이라 예상할 수 있다. 따라서 이상치가 없는 Box plot 을 가졌을 경우 정규분포를 따른다고 가정하면, 정규분포를 따른다고 할 수 있는 변수는 'song_popularity', 'acousticness', 'key', 'audio_mode', 'audio_valence', 총 5 개이다. 이 변수들은 모두 왜도와 첨도를 기준으로 각 변수가 정규분포를 따르는지 확인해보았을 때 포함되었던 변수들이다. 이는 Box plot 과 왜도, 첨도 모두 데이터의 분포를 통해 나오는 결과이며, Box plot 에 조금 더 엄격한 기준을 적용했기 때문으로 보인다.

추가적으로 정규성 검정을 위해 Anderson-Darling Test 를 진행하였다. Shapiro-Wilk Test 는 표본 크기가 5000 을 넘어가면 정확하지 않을 수 있으므로, 표본 크기에 제약이 없고 Shapiro-Wilk Test 다음으로 검정력이 높은 Anderson-Darling Test 를 선정하였다.
Anderson-Darling Test 의 가설은 다음과 같다.

H_0: 데이터가 정규분포를 따른다. vs. H_a: 데이터가 정규분포를 따르지 않는다.

	Column	Test Statistic	Critical Value(alpha = 0.05)	Test Result
0	song_popularity	135.004781	0.787	귀무가설을 기각한다. (데이터가 정규분포를 따르지 않는다.)
1	song_duration_ms	344.806377	0.787	귀무가설을 기각한다. (데이터가 정규분포를 따르지 않는다.)
2	acousticness	1218.562425	0.787	귀무가설을 기각한다. (데이터가 정규분포를 따르지 않는다.)
3	danceability	39.858309	0.787	귀무가설을 기각한다. (데이터가 정규분포를 따르지 않는다.)
4	energy	156.701859	0.787	귀무가설을 기각한다. (데이터가 정규분포를 따르지 않는다.)
5	instrumentalness	5214.375908	0.787	귀무가설을 기각한다. (데이터가 정규분포를 따르지 않는다.)
6	key	406.970166	0.787	귀무가설을 기각한다. (데이터가 정규분포를 따르지 않는다.)
7	liveness	1424.139689	0.787	귀무가설을 기각한다. (데이터가 정규분포를 따르지 않는다.)
8	loudness	521.969280	0.787	귀무가설을 기각한다. (데이터가 정규분포를 따르지 않는다.)
9	audio_mode	3681.182391	0.787	귀무가설을 기각한다. (데이터가 정규분포를 따르지 않는다.)
10	speechiness	1978.577383	0.787	귀무가설을 기각한다. (데이터가 정규분포를 따르지 않는다.)
11	tempo	90.804324	0.787	귀무가설을 기각한다. (데이터가 정규분포를 따르지 않는다.)
12	time_signature	6181.027531	0.787	귀무가설을 기각한다. (데이터가 정규분포를 따르지 않는다.)
13	audio_valence	98.373212	0.787	귀무가설을 기각한다. (데이터가 정규분포를 따르지 않는다.)

입력 변수들에 대하여 Anderson-Darling Test 를 진행한 결과, 모든 변수가 정규분포를 따르지 않는다.

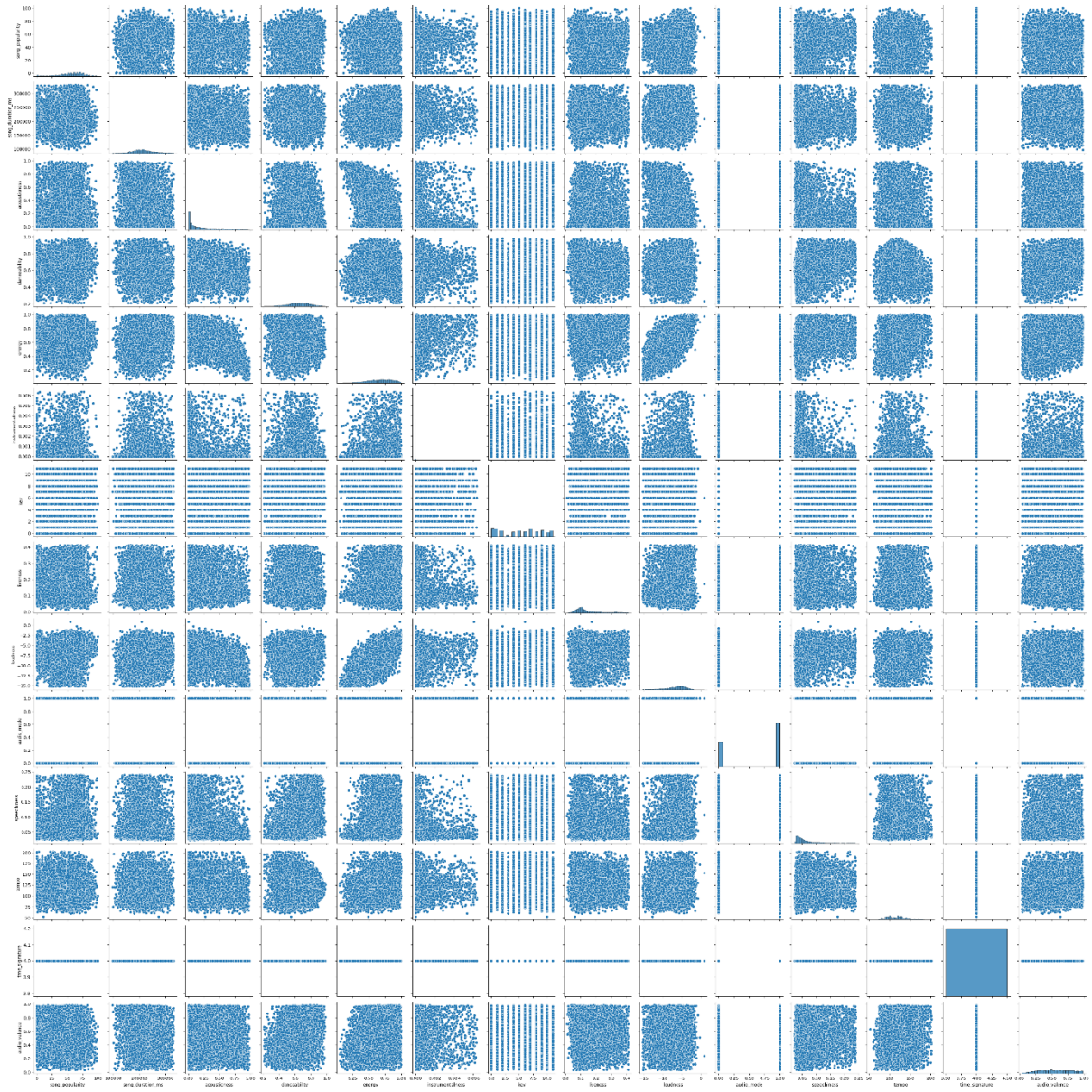
[Q4] [Q3]의 Box plot 을 근거로 각 변수들에 대한 이상치(너무 값이 크거나 작은 객체) 조건을 정의하고, 해당하는 객체들을 데이터셋에서 제거해 보시오.

[Q3]의 Box plot 에서 나타난 것처럼, $Q3 + 1.5 * IQR$보다 위쪽에 있거나, $Q1 - 1.5 * IQR$보다 아래쪽에 있는 값을 이상치로 정의하였다. 여기서 $Q3$는 제 3 사분위수, $Q1$은 제 1 사분위수, $IQR = Q3 - Q1$이다. 정의한 이상치 범위에 해당하는 객체들을 제거한 결과, 데이터셋의 크기가 18835 에서 11004 로 감소하였다. 이는 다중회귀분석을 진행하기에 충분한 크기이므로 추가적인 이상치 범위 조절 없이 분석을 진행한다.

다음 각 물음에 대해서는 [Q4]에서 제거된 객체들을 제외시킨 데이터프레임을 사용하여 답하시오.

[Q5] 가능한 모든 두 쌍의 입력 변수 조합에 대한 산점도(scatter plot) 등을 도시하여 입력변수간 상관성에 대한 분석을 수행해 보시오. 어떤 두 조합의 변수들이 서로 강한 상관관계가 있다고 할 수 있는가? 이렇게 강한 상관관계가 발생한 변수들은 상식적으로도 상관관계가 높은 변수들이라고 할 수 있는가?

가능한 모든 두 쌍의 입력 변수 조합에 대한 산점도를 그리고자 seaborn 라이브러리의 pairplot 함수를 이용하였다. 그 결과는 다음과 같다.

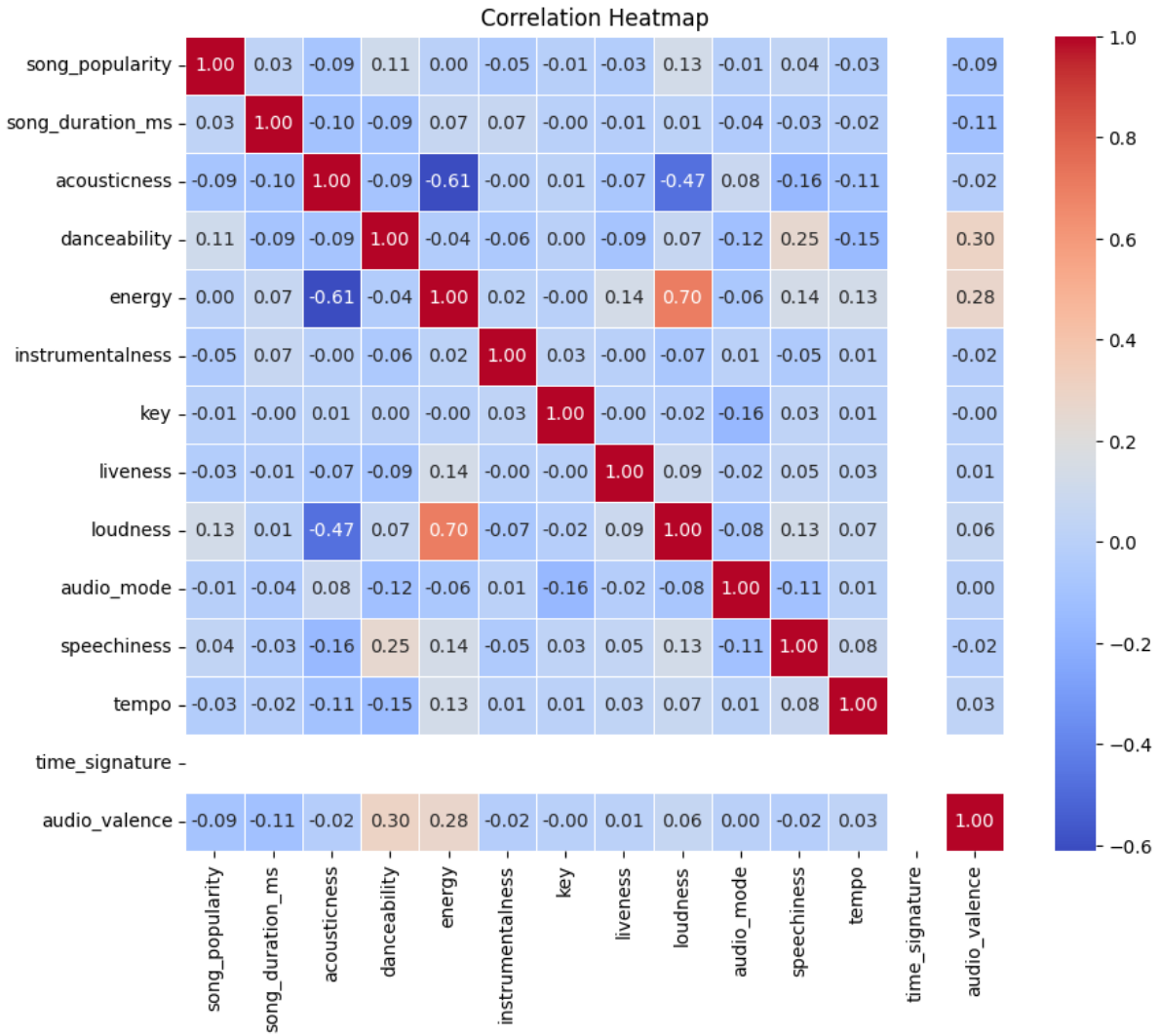


산점도를 통해 육안으로 확인해본 결과, {energy, loudness}는 양의 상관관계를, {energy, acousticness}는 음의 상관관계를 보이고 있다.

- {energy, loudness}: 'energy'는 강렬함과 활동성의 척도로, 1 에 가까울수록 더욱 강렬하고 높은 활동성을 나타낸다. 일반적으로 음악을 들을 때, 강렬하고 활발한 음악은 조용하고 잔잔한 음악보다 다양한 소리들이 밀집되어 있고, 소리가 역시 더 세다. 예를 들어 밴드 음악은 전자 기타, 전자 피아노 등의 다양한 악기들이 노래 내내 소리를 낸다. 그러나 발라드 음악을 생각해보면 대개 적은 수의 악기를 사용하며, 동시에 악기를 연주하기보다는 피아노와 같은 악기로 멜로디를 연주하고, 간간히 필요한 부분에 새로운 악기를 등장시키는 식이다. 따라서 'energy'와 음원의 전반적인 음량을 나타내는 'loudness'는 상식적으로도 높은 양의 상관관계를 가져야 한다.
- {energy, acousticness}: 'acousticness'는 음원이 얼마나 음향적으로 가공되지 않았는지를 나타낸다. 간단히 생각해보면, 어쿠스틱 음악이라는 것은 대개 사람의 목소리와 잔잔한 기타

혹은 피아노 연주로 이루어져 있다. 이는 어쿠스틱 음악이 당연히 높은 'acousticness' 값을 가지나, 높은 'energy' 값을 갖기 어려울 것이라고 예측할 수 있다. 밴드 음악의 경우는 다양한 전자 악기들을 사용한다. 이는 음향적으로 가공된 소리를 가지고 있기 때문에, 'acousticness' 값이 높을 수 없다. 그러나 'energy'의 값은 높을 것이다. 따라서 'energy'와 'acousticness'는 당연히 음의 상관관계를 가져야 한다는 것을 알 수 있다.

추가적으로, 변수들 간의 상관계수를 시각화한 히트맵은 다음과 같다.



상관계수는 변수 간의 선형관계를 측정하는 단위이며, [-1, 1]의 값을 가질 수 있다. 값이 1 에 가까울수록 양의 상관관계, -1 에 가까울수록 음의 상관관계를 가진다. 값의 절댓값이 클수록 선형성은 강하며, 값이 0 에 가까우면 선형성이 존재하지 않는다고 판단할 수 있다.

상관계수의 절댓값	**의미**	**해당하는 변수쌍**
[0.0, 0.1)	상관관계가 거의 없다.	그 외 모든 변수쌍.

[0.1, 0.5)	약한 음/양의 상관관계가 있다.	- 음: {acousticness, song_duration_ms}, audio_valence_song_duration_ms}, {loudness, acousticness}, {speechiness, acousticness}, {tempo, acousticness}, {audio_mode, danceability}, {tempo, danceability}, {audio_mode, key}, {speechiness, audio_mode} - 양: {danceability, song_popularity}, {loudness, song_popularity}, {speechiness, danceability}, {audio_valence, danceability}, {liveness, energy}, {speechiness, energy}, {tempo, energy}, {audio_valence, energy}, {speechiness, loudness}
[0.5, 0.7)	음/양의 상관관계가 있다.	- 음: {energy, acousticness}
[0.7, 0.9)	높은 음/양의 상관관계가 있다.	- 양: {energy, loudness}
[0.9, 1.0]	매우 높은 음/양의 상관관계가 있다.	없음.

대부분의 변수들 사이에 상관관계가 없거나 약한 것으로 나타났다.

[Q6] 전체 데이터셋을 70%의 학습 데이터와 30%의 테스트 데이터로 분할한 후 모든 변수를 사용하여 MLR 모델을 학습한 뒤, $Adjusted\ R^2$ 값을 통해 데이터의 선형성(linearity)을 판단해 보시오. Residual plot 과 Q-Q Plot 을 도시하고 Ordinary Least Square 방식의 Solution 이 만족해야 하는 가정들이 만족될만한 수준인지 정성적으로 판단해 보시오.

OLS Regression Results

Dep. Variable:	song_popularity	R-squared:	0.054
Model:	OLS	Adj. R-squared:	0.052
Method:	Least Squares	F-statistic:	36.26
Date:	Sat, 30 Mar 2024	Prob (F-statistic):	3.90e-83
Time:	04:45:33	Log-Likelihood:	-34532.
No. Observations:	7702	AIC:	6.909e+04
Df Residuals:	7689	BIC:	6.918e+04
Df Model:	12		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	4.0637	0.201	20.174	0.000	3.669	4.459
song_duration_ms	1.247e-05	6.23e-06	2.000	0.046	2.49e-07	2.47e-05
acousticness	-6.0438	1.275	-4.741	0.000	-8.543	-3.545
danceability	18.4916	2.037	9.077	0.000	14.498	22.485
energy	-16.3452	2.367	-6.907	0.000	-20.984	-11.706
instrumentalness	-811.1961	254.531	-3.187	0.001	-1310.146	-312.247
key	-0.0013	0.069	-0.018	0.985	-0.137	0.134
liveness	-2.8560	2.743	-1.041	0.298	-8.233	2.521
loudness	1.5214	0.134	11.362	0.000	1.259	1.784
audio_mode	0.6742	0.525	1.285	0.199	-0.355	1.703
speechiness	-1.7127	5.100	-0.336	0.737	-11.711	8.285
tempo	-0.0044	0.009	-0.473	0.636	-0.023	0.014
time_signature	16.2549	0.806	20.174	0.000	14.675	17.834
audio_valence	-8.9099	1.230	-7.246	0.000	-11.320	-6.499

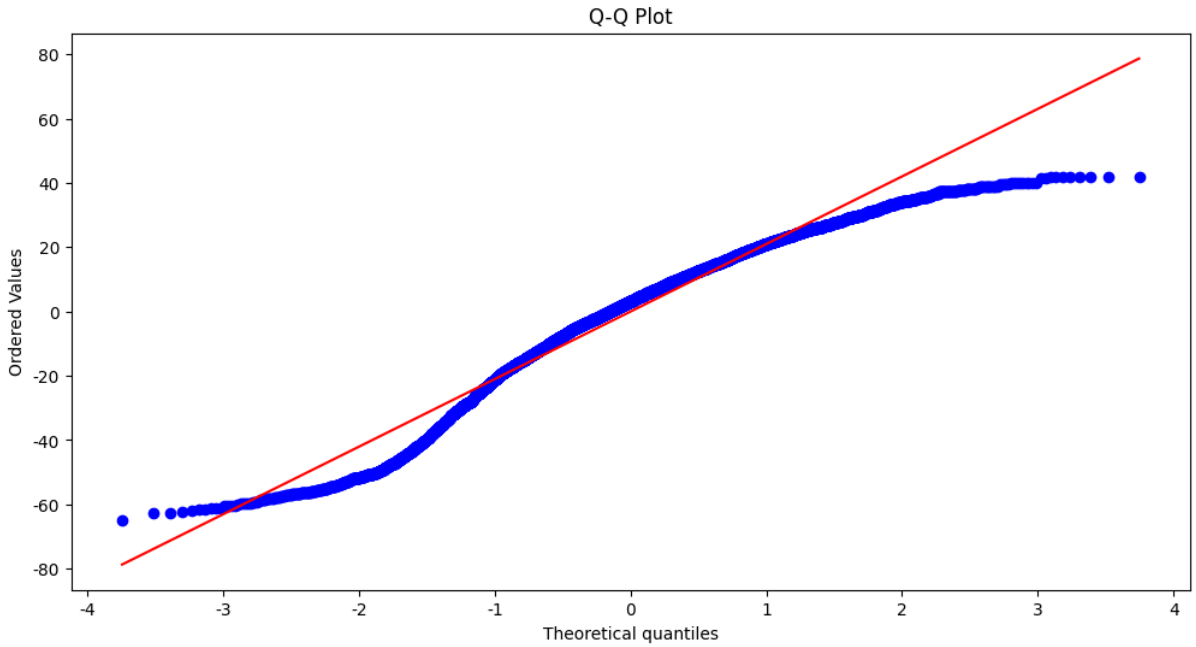
Omnibus:	499.821	Durbin-Watson:	1.974
Prob(Omnibus):	0.000	Jarque-Bera (JB):	603.012
Skew:	-0.685	Prob(JB):	1.14e-131
Kurtosis:	3.041	Cond. No.	1.86e+19

$Adjusted\ R^2$ 값은 0.052 로, 데이터에 선형성이 존재하지 않는다. 따라서 선형성을 만족해야 사용할 수 있는 선형회귀가 아닌, 다른 모델을 통해 종속변수를 예측해야 할 것이다.
추가적으로, R^2는 0.054 로, 단순히 종속변수의 평균값을 사용했을 때 대비 독립변수들을 사용함으로써 5.4% 정도의 성능 향상이 이루어졌다. 분야의 특수성을 고려해보아도, 현재의 설명변수만으로는 종속변수를 예측하기 어렵다는 결론을 내릴 수 있다.

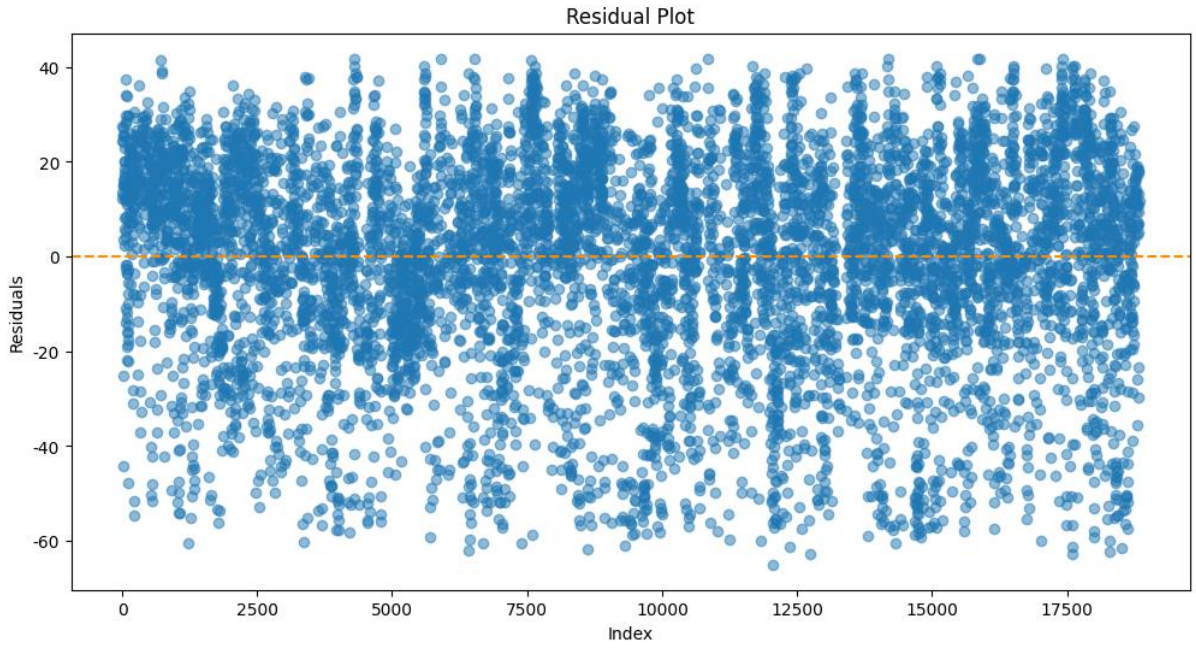
Ordinary Least Square 방식의 Solution 이 만족해야 하는 가정은 다음과 같다.

1. 설명변수와 종속변수 사이에 선형관계가 성립한다.
2. 오차항 ε이 정규분포를 따른다.
3. 각 관측치들은 서로 독립이어야 한다.
4. 종속변수 Y에 대한 오차항(residual)은 설명변수 값의 범위에 관계없이 일정해야 한다.

Q-Q Plot 을 통해 2 번 가정을, Residual Plot 을 통해 4 번 가정을 만족하는지 살펴보고자 한다.
다음의 Q-Q Plot 을 보면, 점들이 직선 위에 분포해 있는 것이 아니라, 직선을 벗어나 S 자 형태의 곡선을 그리고 있는 것을 볼 수 있다. 따라서 Q-Q Plot 을 통해서는 설명변수와 종속변수 사이에 선형관계가 있다고 이야기하기 어렵다.



4 번 가정은 잔차의 등분산성에 대한 것으로, 잔차들이 무작위로 분포되어 있어야 한다는 것을 의미한다. 다음의 Residual Plot 을 보면, 잔차들이 특정한 분포 형태를 띠지 않고, 무작위로 분포되어 있으므로 4 번 가정을 만족한다고 볼 수 있다.



따라서 Ordinary Least Square 방식의 Solution 의 4 번 가정은 만족하나 2 번 가정은 만족하지 못한다고 결론지을 수 있다.

[Q7] 유의수준 0.01 에서 모형 구축에 통계적으로 유의미한 변수들은 어떤 것들이 있는가? 해당 변수들은 종속변수와 양/음 중에서 어떤 상관관계를 갖고 있는가?

앞서 제시했던 [Q6]의 모델 학습 결과를 이용해서 통계적으로 유의미한 변수들을 알아볼 것이다.

먼저, 변수가 통계적으로 유의미한지를 검정하기 위한 가설은 다음과 같다.

$$H_0: \beta_i = 0 \quad \text{vs.} \quad H_a: \beta_i \neq 0$$

$p - value < \alpha = 0.01$이면 귀무가설을 기각할 수 있다.

	coef	std err	t	P>|t|	[0.025	0.975]
const	4.0637	0.201	20.174	0.000	3.669	4.459
song_duration_ms	1.247e-05	6.23e-06	2.000	0.046	2.49e-07	2.47e-05
acousticness	-6.0438	1.275	-4.741	0.000	-8.543	-3.545
danceability	18.4916	2.037	9.077	0.000	14.498	22.485
energy	-16.3452	2.367	-6.907	0.000	-20.984	-11.706
instrumentalness	-811.1961	254.531	-3.187	0.001	-1310.146	-312.247
key	-0.0013	0.069	-0.018	0.985	-0.137	0.134
liveness	-2.8560	2.743	-1.041	0.298	-8.233	2.521
loudness	1.5214	0.134	11.362	0.000	1.259	1.784
audio_mode	0.6742	0.525	1.285	0.199	-0.355	1.703
speechiness	-1.7127	5.100	-0.336	0.737	-11.711	8.285
tempo	-0.0044	0.009	-0.473	0.636	-0.023	0.014
time_signature	16.2549	0.806	20.174	0.000	14.675	17.834
audio_valence	-8.9099	1.230	-7.246	0.000	-11.320	-6.499

위 내용을 참조하면 다음과 같은 결과를 얻을 수 있다.

변수명	귀무가설 기각 여부	해석
song_duration_ms	귀무가설 기각	회귀계수가 유의미하다. song_duration_ms 가 1 단위 증가하면 song_population 이 1.247e-05 만큼 증가하는 양의 상관관계를 갖는다.
acousticness	귀무가설 기각	회귀계수가 유의미하다. acousticness 가 1 단위 증가하면 song_population 이 6.0438 만큼 감소하는 음의 상관관계를 갖는다.
danceability	귀무가설 기각	회귀계수가 유의미하다. danceability 가 1 단위 증가하면 song_population 이 18.4916 만큼 증가하는 양의 상관관계를 갖는다.
energy	귀무가설 기각	회귀계수가 유의미하다. energy 가 1 단위 증가하면 song_population 이 16.3452 만큼 감소하는 음의 상관관계를 갖는다.
instrumentalness	귀무가설 기각	회귀계수가 유의미하다. instrumentalness 가 1 단위 증가하면 song_population 이 811.1961 만큼 감소하는 음의 상관관계를 갖는다.
key	귀무가설 기각 불가	회귀계수가 유의미하지 않다.

liveness	귀무가설 기각 불가	회귀계수가 유의미하지 않다.
loudness	귀무가설 기각	회귀계수가 유의미하다. loudness 가 1 단위 증가하면 song_population 이 1.5214 만큼 증가하는 양의 상관관계를 갖는다.
audio_mode	귀무가설 기각 불가	회귀계수가 유의미하지 않다.
speechiness	귀무가설 기각 불가	회귀계수가 유의미하지 않다.
tempo	귀무가설 기각 불가	회귀계수가 유의미하지 않다.
time_signature	귀무가설 기각	회귀계수가 유의미하다. time_signature 가 1 단위 증가하면 song_population 이 16.2549 만큼 증가하는 양의 상관관계를 갖는다.
audio_valence	귀무가설 기각	회귀계수가 유의미하다. audio_valence 가 1 단위 증가하면 song_population 이 8.9099 만큼 감소하는 음의 상관관계를 갖는다.

[Q8] Test 데이터셋에 대하여 MAE, MAPE, RMSE 를 계산하고 그에 대한 해석을 해 보시오.

다음은 seed=123 으로 고정한 뒤 분할한 test 데이터셋에 대해 MAE, MAPE, RMSE 를 계산한 결과이다.

	MAE	MAPE	RMSE
song_population	17.129326	2.828409e+15	21.66449

seed 를 고정하지 않고 10 회 반복 실험한 결과는 다음과 같다.

	MAE	MAPE	RMSE
song_population	17.0964 (0.1257)	2747718340001399.0000 (368550287957285.0625)	21.5344 (0.1745)

각 값의 형태는 10 회 반복 실험 결과의 평균(표준편차)이다.

- MAE 값을 통해 실제 음원의 흥행 성적과 예측 흥행 성적 사이의 절대적인 오차의 평균이 약 17 정도 된다는 것을 알 수 있다. 음원의 흥행 성적이 백분율 점수임을 고려하면, 회귀모델이 음원의 흥행 성적을 잘 예측하지 못한다고 판단할 수 있다.
- MAPE 는 실제 값 대비 얼마나 예측 차이가 있는지를 비율로 측정한 결과이다. 도출된 MAPE 값을 통해 약 2747718340001399% 정도의 차이가 있다는 것을 알 수 있다. 흥행 성적이

모두 양수임을 고려하면, 실제로는 흥행 성적이 낮은데 흥행 성적이 높을 것이라고 예측한 경우가 많아서 매우 큰 값을 가지게 되었다고 해석할 수 있다.

- RMSE 는 실제 값과 예측 값 사이의 오차에 제곱합의 제곱근을 취한 값이다. RMSE 값을 통해 약 22 정도의 오차가 발생한다고 생각할 수 있다. MAE 와 마찬가지로 회귀모델이 음원의 흥행 성적을 잘 예측하지 못한다고 판단할 수 있다.

음원의 흥행 성적을 생각해보면, 절대적인 차이보다는 상대적 차이가 중요하다. 따라서 가장 중점적으로 고려해야 할 지표는 MAPE 다. 상대 오차는 약 2747718340001399%로 매우 높은 값을 가져 회귀모델의 성능이 좋지 못함을 알 수 있다.

[Q9] 만약 원래 변수 수의 절반 이하로 입력 변수를 사용하여 모델을 구축해야 할 경우 어떤 변수들을 선택하겠는가? [Q5]와 [Q7]의 답변을 바탕으로 본인이 선택한 변수들에 대한 근거를 제시하시오.

원래 변수 수의 절반 이하로 입력 변수를 사용하기 위해 최소 7 개의 설명변수를 제거할 것이다.
먼저, [Q7]에서 유의미하지 않다고 판단했던 변수인 'key', 'liveness', 'audio_mode', 'speechiness', 'tempo'를 제거한다. 해당 변수들은 p-value 값을 통해 가설검정을 진행한 결과, 회귀계수가 0 이라는 귀무가설을 기각하지 못해, 종속변수인 'song_popularity'를 설명하는데 아무런 영향을 미치지 못할 것이라 판단했기 때문이다.
다음으로는 [Q5]에서 종속변수인 'song_popularity'와의 상관성이 낮았던 변수를 제거할 것이다. 앞서 제거했던 변수를 제외하고, 두개의 변수를 추가적으로 제거하면, 종속변수와의 상관계수가 0.00 인 'energy', 종속변수와의 상관계수가 0.03 인 'song_duration_ms'를 제거한다. 종속변수와의 상관계수가 낮다는 것은, 결국 해당 변수를 통해 종속변수를 설명하지 못한다는 의미이기 때문이다.
추가적으로, [Q5]에서 비교적 높은, |0.5| 이상의 상관계수를 가졌던 설명변수들의 쌍을 살펴보고, 두 변수 중 하나를 제거할 것이다. 설명변수 간의 상관관계가 높다는 것은, 설명변수들이 서로를 잘 설명한다는 이야기이고, 이는 둘 중 하나의 변수를 제거해도 종속변수를 예측하는데 있어 성능의 차이가 그리 크지 않다는 것을 의미한다. 먼저, 0.7 의 상관계수를 가졌던 {loudness, energy}를 보면, 'loudness'와 'song_popularity'의 상관관계가 'energy'와 'song_popularity'의 상관관계보다 높으므로, 'energy' 변수를 제거한다. 'energy'보다는 'loudness'가 'song_popularity'를 예측하는데 있어 더 높은 설명력을 가지고 있기 때문이다. 마찬가지로, {energy, acousticness} 역시 'acousticness'와 'song_popularity'의 상관관계가 'energy'와 'song_popularity'의 상관관계보다 높으므로, 'energy' 변수를 제거한다.
마지막으로, 다른 변수와의 상관계수가 계산되지 않았던 'time_signature'의 값을 확인해본 결과, 모든 관측치가 동일한 값을 가진 것으로 나타나 설명력이 없다고 판단해 입력변수에서 제외하기로 하였다.
따라서, 위 결과들을 종합해보면, 선택된 변수는 'acousticness', 'danceability', 'instrumentalness', 'loudness', 'audio_valence'이다.

[Q10] [Q9]에서 선택한 변수들만을 사용하여 MLR 모형을 다시 학습하고 $Adjusted\ R^2$, Test 데이터셋에 대한 MAE, MAPE, RMSE 를 산출한 뒤, 두 모형(모든 변수 사용 vs. 선택된 변수만 사용)을 비교해 보시오.

[Q9]에서 선택한 변수들만을 사용하여 학습한 모형의 결과는 다음과 같다.

OLS Regression Results

Dep. Variable:	song_popularity	R-squared:	0.047
Model:	OLS	Adj. R-squared:	0.046
Method:	Least Squares	F-statistic:	75.08
Date:	Mon, 01 Apr 2024	Prob (F-statistic):	4.64e-77
Time:	09:45:57	Log-Likelihood:	-34560.
No. Observations:	7702	AIC:	6.913e+04
Df Residuals:	7696	BIC:	6.917e+04
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	54.6658	1.347	40.584	0.000	52.025	57.306
acousticness	-1.6966	1.104	-1.537	0.124	-3.860	0.467
danceability	22.4052	1.799	12.456	0.000	18.879	25.931
instrumentalness	-908.7508	253.848	-3.580	0.000	-1406.362	-411.140
loudness	0.9019	0.103	8.762	0.000	0.700	1.104
audio_valence	-12.7543	1.084	-11.767	0.000	-14.879	-10.630

Omnibus:	494.254	Durbin-Watson:	1.972
Prob(Omnibus):	0.000	Jarque-Bera (JB):	595.170
Skew:	-0.681	Prob(JB):	5.76e-130
Kurtosis:	3.030	Cond. No.	7.53e+03

기존 모델의 $Adjusted\ R^2$ 값은 0.052 로, 새롭게 학습한 모델의 $Adjusted\ R^2$ 값보다 0.06 높은 것을 볼 수 있다. 이는 유의한 변수들 역시 함께 제거되었기 때문으로 보인다. 새롭게 학습한 모델의 R^2 값은 0.047 로, 단순히 종속변수의 평균값을 사용했을 때 대비 독립변수들을 사용함으로써 4.7% 정도의 성능 향상이 이루어졌다고 해석할 수 있다. R^2 값 역시 기존 모델에 비해 낮아졌는데, 이는 변수가 줄어들었기 때문이다. 또, 기존 모델에서는 유의한 변수였던 'acousticness'가 새로운 모델에서는 유의수준 0.01 에서 유의하지 않은 변수로 바뀐 것을 확인할 수 있다. 다른 네 변수보다 종속변수에 대한 설명력이 낮았던 것으로 보인다.

다음은 seed=123 으로 고정한 뒤 분할한 test 데이터셋에 대해 MAE, MAPE, RMSE 를 계산한 결과이다. 기존 모델과의 비교를 위해, 동일한 데이터셋에서 실험을 진행하였다. 기존 모델 역시 새롭게 실험을 진행해보았다.

	MAE	MAPE	RMSE
song_population_original	17.129326	2.828409e+15	21.66449

	MAE	MAPE	RMSE
song_population_new	17.251332	2.832460e+15	21.759797

새로운 모델이 기존 모델에 비해 세가지 지표의 값이 모두 증가하였다. 그러나 모두 증가한 비율이 매우 작아 데이터셋의 영향이 있을 수 있다고 판단하여 seed 를 고정하지 않고 10 회 반복 실험을 진행하였다.

	MAE	MAPE	RMSE
song_population_original	17.1269 (0.1665)	2844474914593865.5000 (341836455458609.4375)	21.6298 (0.2105)

	MAE	MAPE	RMSE
song_population_new	17.2278 (0.1618)	2859893660990447.5000 (337512654780189.3125)	21.7126 (0.2123)

각 값의 형태는 10 회 반복 실험 결과의 평균(표준편차)이다.
MAE, MAPE, RMSE 는 모두 새로운 모델이 기존 모델에 비해 높아졌고, 표준편차는 RMSE 를 제외하고는 낮아진 것을 확인할 수 있다. 성능 지표만을 고려해서 모델을 선정하는 경우에는 기존 모델을 선택하는 것이 좋겠지만, 백분율 점수를 예측하는데 변수를 줄임으로써 늘어난 오차가 소수점 단위라는 것은 두 모델 사이에 설명력 차이가 거의 없다는 것으로 보아도 무방할 것이다. 이는 애초에 기존 모델과 새로운 모델 모두 종속변수를 예측하는데 있어 매우 낮은 성능을 보였기 때문으로 추측된다. 따라서 만약 두 모델 중 하나를 선택한다면 새로운 모델을 선택할 것이다. 기존 모델과의 설명력 차이가 별로 없으나, 표준편차가 작아 더 안정적이고, 변수의 개수가 적어 해석을 하는데 있어서도 새로운 모델이 더 용이하기 때문이다.

[Extra Question] 이 외 해당 데이터셋을 통해 MLR 관점에서 가능한 추가적인 분석을 웹에서 검색해서 수행하고 그 결과를 해석해 보시오.

[Q9]에서 다중공선성을 확인하지 않았기에, 먼저 추가적으로 다중공선성을 확인하려 한다.

	Feature	VIF_Factor
0	song_duration_ms	1.048195
1	acousticness	1.752620
2	danceability	1.437796
3	energy	3.267782
4	instrumentalness	1.023853
5	key	1.030429
6	liveness	1.032174
7	loudness	2.226669
8	audio_mode	1.062166
9	speechiness	1.155384
10	tempo	1.058182
11	time_signature	196.413006
12	audio_valence	1.429269

VIF 값을 확인해본 결과, [Q9]에서 이미 제거하였던 'time_signature'만 다중공선성이 존재하는 것으로 나타났다. 이는 이상치 제거를 통해 모든 관측치의 'time_signature' 값이 동일해졌기 때문으로 예상된다. 다중공선성으로는 [Q9], [Q10]과 다른 모델을 만들기 어려우므로 Forward Selection, Backward Elimination 기법을 통해 새롭게 변수를 선택해보려 한다.

먼저 Forward Selection 결과, 'danceability', 'instrumentalness', 'key', 'audio_mode', 'time_signature', 'audio_valence', 총 6 개의 변수가 선택되었다.

OLS Regression Results

Dep. Variable:	song_popularity	R-squared:	0.032
Model:	OLS	Adj. R-squared:	0.031
Method:	Least Squares	F-statistic:	41.78
Date:	Mon, 01 Apr 2024	Prob (F-statistic):	2.03e-50
Time:	12:37:56	Log-Likelihood:	-34620.
No. Observations:	7702	AIC:	6.925e+04
Df Residuals:	7695	BIC:	6.930e+04
Df Model:	6		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-3.013e+13	2.05e+13	-1.468	0.142	-7.04e+13	1.01e+13
danceability	23.6460	1.820	12.990	0.000	20.078	27.214
instrumentalness	-1079.4473	254.956	-4.234	0.000	-1579.230	-579.664
key	-0.0295	0.070	-0.421	0.673	-0.167	0.108
audio_mode	0.2451	0.526	0.466	0.641	-0.787	1.277
time_signature	7.532e+12	5.13e+12	1.468	0.142	-2.53e+12	1.76e+13
audio_valence	-12.3351	1.092	-11.292	0.000	-14.476	-10.194

Omnibus:	438.442	Durbin-Watson:	1.975
Prob(Omnibus):	0.000	Jarque-Bera (JB):	516.889
Skew:	-0.634	Prob(JB):	5.74e-113
Kurtosis:	2.946	Cond. No.	6.35e+14

기존 모델과 [Q10]의 모델보다 R^2 값이 낮으며, 5 개의 변수만을 사용했던 [Q10]의 모델보다 $Adjusted\ R^2$ 값이 낮은 것을 확인할 수 있다. 이는 애초에 모델의 설명력이 낮아, 새로운 변수를 추가하더라도 설명력이 높아지기 어려운 상황이었기 때문으로 보인다.

	MAE	MAPE	RMSE
song_population_original	17.1188 (0.1589)	2806039425713910.5000 (291199210455070.7500)	21.5853 (0.1864)

	MAE	MAPE	RMSE
song_population_forward	17.4007 (0.1749)	2757342386097753.0000 (221013564327930.6562)	21.7742 (0.1571)

MAE 와 RMSE 는 증가, MAPE 는 감소한 것을 볼 수 있다. 기존 모델에 비해 Forward Selection 을 적용한 모델의 MAE 가 높고 MAPE 가 낮은 것을 고려해보면, Forward Selection 을 적용한 모델이 실제 종속변수의 값이 작을 때, 기존 모델보다 종속변수의 값을 더 작게 추정해서 그런 것으로 추측된다.

Backward Elimination 결과, 'energy', 'instrumentalness', 'key', 'loudness', 'audio_mode', 'tempo', 총 6 개의 변수가 선택되었다.

OLS Regression Results

Dep. Variable:	song_popularity	R-squared:	0.032
Model:	OLS	Adj. R-squared:	0.031
Method:	Least Squares	F-statistic:	42.24
Date:	Mon, 01 Apr 2024	Prob (F-statistic):	5.48e-51
Time:	12:40:22	Log-Likelihood:	-34619.
No. Observations:	7702	AIC:	6.925e+04
Df Residuals:	7695	BIC:	6.930e+04
Df Model:	6		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	83.6661	2.246	37.255	0.000	79.264	88.068
energy	-19.4154	1.891	-10.265	0.000	-23.123	-15.708
instrumentalness	-742.0995	256.185	-2.897	0.004	-1244.292	-239.907
key	-0.0185	0.070	-0.264	0.791	-0.156	0.119
loudness	1.9444	0.130	15.008	0.000	1.690	2.198
audio_mode	-0.2277	0.524	-0.435	0.664	-1.254	0.799
tempo	-0.0158	0.009	-1.701	0.089	-0.034	0.002

Omnibus:	447.294	Durbin-Watson:	1.979
Prob(Omnibus):	0.000	Jarque-Bera (JB):	529.390
Skew:	-0.642	Prob(JB):	1.11e-115
Kurtosis:	2.981	Cond. No.	1.28e+05

Backward Elimination 역시 기존 모델과 [Q10]의 모델보다 R^2 값이 낮으며, 5 개의 변수만을 사용했던 [Q10]의 모델보다 $Adjusted\ R^2$ 값이 낮은 것을 확인할 수 있다. Forward Selection 과 마찬가지로 설명변수들이 종속변수를 잘 설명하지 못해 발생한 일인 것으로 추정된다.

	MAE	MAPE	RMSE
song_population_original	17.0971 (0.1690)	2805717178656000.5000 (335525701833399.8125)	21.5496 (0.1990)

	MAE	MAPE	RMSE
song_population_backward	17.2455 (0.1807)	2787086237413284.5000 (400585544478677.9375)	21.6966 (0.2089)

MAE 와 RMSE 는 증가, MAPE 는 감소한 것을 볼 수 있다. Forward Selection 과 마찬가지로, 기존 모델에 비해 Backward Elimination 을 적용한 모델의 MAE 가 높고 MAPE 가 낮은 것을 고려해보면, Backward Elimination 을 적용한 모델이 실제 종속변수의 값이 작을 때, 기존 모델보다 종속변수의 값을 더 작게 추정해서 그런 것으로 추측된다.