Spotify top 200 dataset

Introduction

Motivation: We are trying to find out if there are some interesting trending underlying the features of the top 200 songs in the dataset.

Research Questions:

1. What is the relationship between energy, streams?

2. Is there a difference for songs’ tempo between summer and fall?

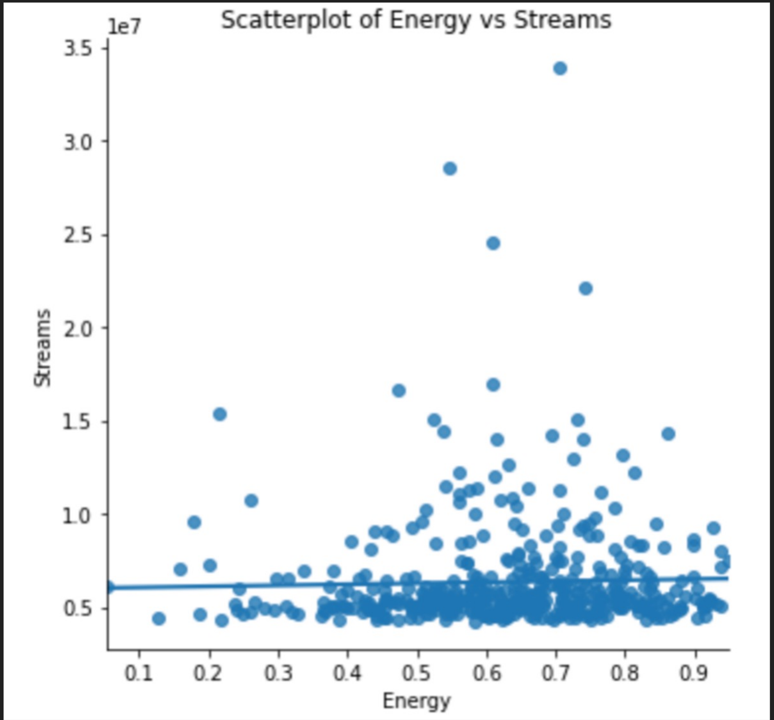
3. Is there a linear relationship between a song’s highest charted position and the song’s artist features?

4. Is there a linear relationship between a song’s highest charted position and the song’s artist features?

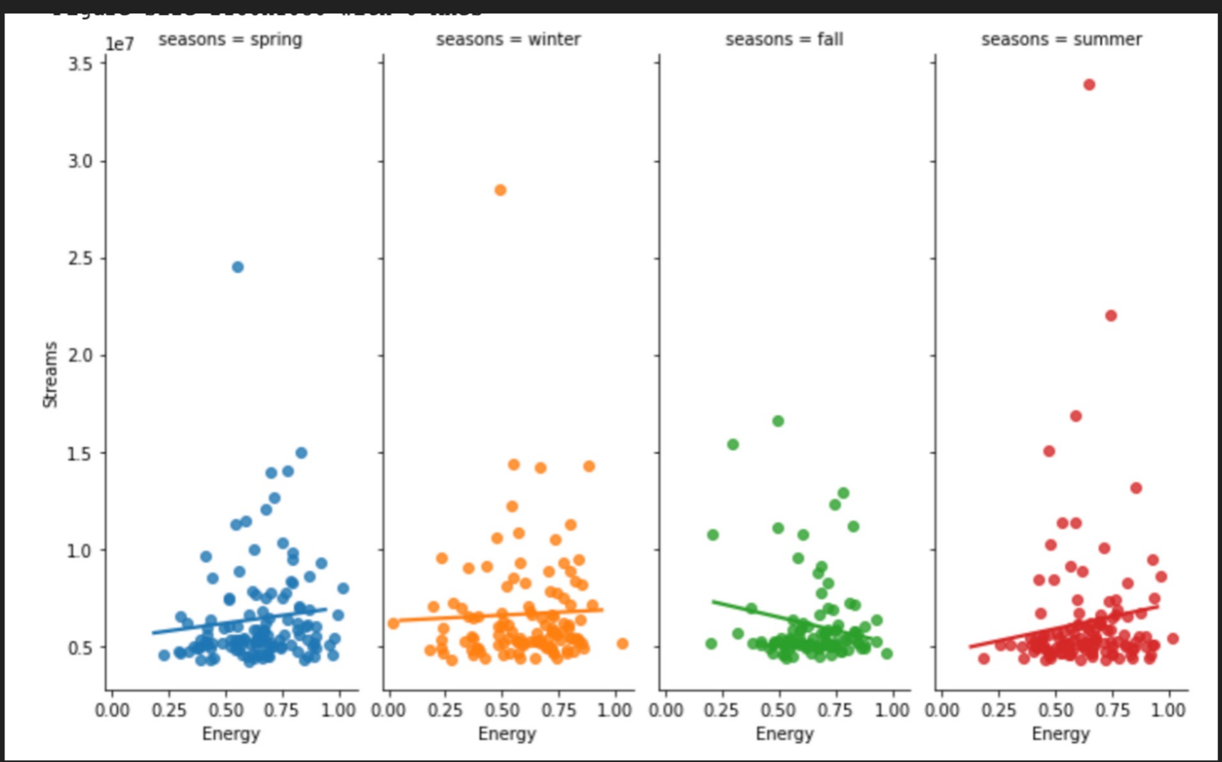
5. Can we detect any underlying association between “hip hop” or “rap” song and music component

Descriptive Analytics

What is the relationship between energy, streams?



How does this relationship change among seasons?



Summary Stats for different seasons

Energy

	Energy							
	count	mean	std	min	25%	50%	75%	max
seasons								
fall	101.0	0.645149	0.148752	0.214	0.54200	0.640	0.758	0.904
spring	125.0	0.637504	0.157339	0.186	0.54500	0.632	0.748	0.939
summer	129.0	0.640287	0.149718	0.128	0.53100	0.648	0.740	0.948
winter	112.0	0.597196	0.187088	0.054	0.48975	0.633	0.731	0.938

Streams

	Streams							
	count	mean	std	min	25%	50%	75%	max
seasons								
fall	101.0	6.090321e+06	2.223230e+06	4425610.0	5005018.00	5242347.0	5914661.00	16613649.0
spring	125.0	6.394110e+06	2.702761e+06	4252898.0	4966878.00	5383125.0	6809438.00	24551591.0
summer	129.0	6.245144e+06	3.473929e+06	4314080.0	4900492.00	5228433.0	6144736.00	33948454.0
winter	112.0	6.647179e+06	2.976003e+06	4293009.0	5114516.25	5701746.5	7077461.75	28509534.0

Inference Test

Is there a difference for songs' tempo between summer and fall?

Hypothesis

- $H_0 : \mu_{Summer} = \mu_{Fall}$
- $H_A : \mu_{Summer} \neq \mu_{Fall}$

Conditions Check

1. Random Sample
2. n1 < 10% of the population and n2 < 10% of the population
3. n1 > 30 and n2 > 30 or ~~population 1 and population 2 are normally distributed~~

Test Result

P-value = 0.2

With the alpha of 0.05, we do not reject the null hypothesis and there is no difference between the tempo of summer and fall.

Linear Regression

Is there a linear relationship between a song's highest charted position and the song's artist features?

Highest_Charting_Position
8
92
181
12
32
...

Response variable

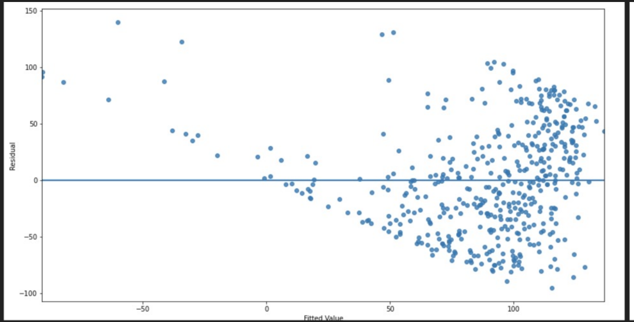
Number_of_Times_Charted	Artist_Followers	Streams	Popularity
24	1398563.0	8832945	87.0
2	5436999.0	5018592	70.0
1	42227614.0	6657404	52.0
29	36142273.0	5242347	82.0
10	11821805.0	5386512	67.0

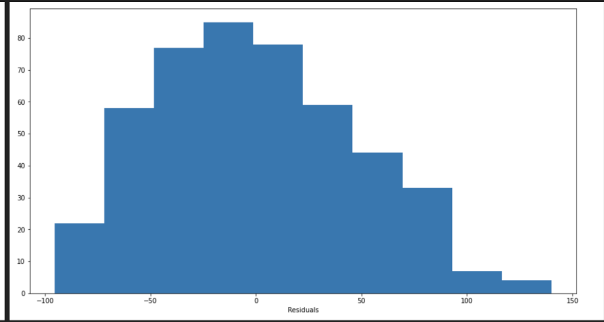
Explanatory variables

Checking linear regression conditions

1. Linearity
2. Constant variance
3. Normality
4. Residual independence
5. Multicollinearity

Try log transformation?



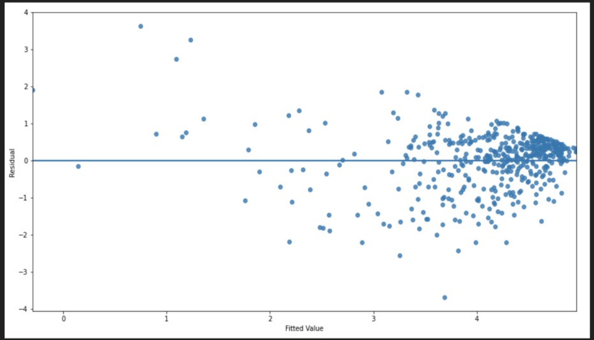


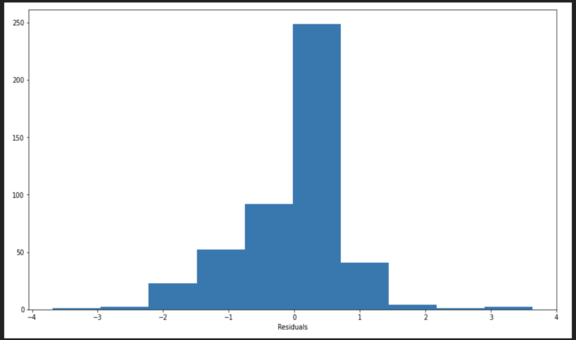
	Number_of_Times_Charted	Streams	Artist_Followers	Popularity
Number_of_Times_Charted	1.000000	-0.036494	0.080957	0.474509
Streams	-0.036494	1.000000	0.072609	0.119730
Artist_Followers	0.080957	0.072609	1.000000	0.037811
Popularity	0.474509	0.119730	0.037811	1.000000

Checking linear regression conditions after transformation

1. Linearity
2. Constant variance
3. Normality
4. Residual independence
5. Multicollinearity

Conditions still not met





	Number_of_Times_Charted	Streams	Artist_Followers	Popularity
Number_of_Times_Charted	1.000000	-0.036494	0.080957	0.474509
Streams	-0.036494	1.000000	0.072609	0.119730
Artist_Followers	0.080957	0.072609	1.000000	0.037811
Popularity	0.474509	0.119730	0.037811	1.000000

Variability of response variable explained by the model

```
In [36]: print("R-squared Value of the Original Model: %s"%(mlr.rsquared))
         print("R-squared Value of the Transformed Model: %s"%(mlr2.rsquared))

         R-squared Value of the Original Model: 0.35823288513223694
         R-squared Value of the Transformed Model: 0.47679605972515504
```

- R-squared value for the **original model** was 0.358
- R-squared value for the **transformed model** was 0.476

- We can examine that the transformation was effective of explaining more variability of the response variable compared to the original model
- However, it is difficult to say that both models have good performance

Which slopes in our model do we have sufficient evidence to suggest are non-zero in the population model?

Hypothesis Test

- $H_0 : \beta_{Number\ of\ Times\ Charted} = \beta_{Streams} = \beta_{Artist\ Followers} = \beta_{Popularity} = 0$
- H_A: At least one of the slopes is not zero.

OLS Regression Results

Dep. Variable:	log_Highest_Charting_Position	**R-squared:**	0.477
Model:	OLS	**Adj. R-squared:**	0.472
Method:	Least Squares	**F-statistic:**	105.3
Date:	Sun, 05 Dec 2021	**Prob (F-statistic):**	1.15e-63
Time:	17:17:58	**Log-Likelihood:**	-557.20
No. Observations:	467	**AIC:**	1124.
Df Residuals:	462	**BIC:**	1145.
Df Model:	4		
Covariance Type:	nonrobust		

- Since the p-value is smaller than the significance level (0.05), we **reject** our null hypothesis
- We can conclude that at least one of the slopes in the corresponding population is not zero.

Answering our research question

Is there a linear relationship between a song’s highest charted position and the song’s artist features?

- Yes, there is at least some linear relation
- However, the relation is not strong
- Should be careful with the results since linear regression conditions are not met

Logistic Regression

Too many types of genres…

```
In [40]: df['Genre_1'].unique()

Out[40]: array(['pop', 'boy band', 'latin', 'melodic rap', 'k-pop',
                'colombian pop', 'hip hop', 'trap chileno', 'deep german hip hop',
                'comic', 'conscious hip hop', 'indie rock italiano', 'atl hip hop',
                'canadian pop', 'trap latino', 'sertanejo', 'italian hip hop',
                'modern alternative rock', 'german hip hop', 'edm', 'dance pop',
                'francoton', 'acoustic pop', 'brooklyn drill', 'alternative r&b',
                'puerto rican pop', 'latin pop', 'east coast hip hop',
                'houston rap', 'uk hip hop', 'trap queen', 'australian pop',
                'sad rap', 'trap argentino', 'chicago rap', 'adult standards',
                'australian rock', 'london rap', 'canadian contemporary r&b',
                'lgbtq+ hip hop', 'german cloud rap', 'dfw rap',
                'north carolina hip hop', 'alt z', 'nyc rap', 'brostep',
                'dominican pop', 'french hip hop', 'classic uk pop',
                'memphis hip hop', 'rap', 'neo mellow', 'australian hip hop',
                'canadian hip hop', 'modern rock', 'cali rap', 'pop soul',
                'detroit hip hop', 'forro', 'r&b brasileiro', 'australian dance',
                'melodic metalcore', 'mariachi', 'electropop', 'dance rock',
                'albanian hip hop', 'eurovision', 'florida rap', 'meme rap',
                'art pop', 'chicago soul', 'pop rap', 'contemporary country',
                'belgian hip hop', 'argentine hip hop', 'british soul',
                'sertanejo pop', 'emo rap', 'viral rap', 'funk carioca',
                'gauze pop', 'reggaeton', 'a cappella', 'celtic', 'electro house',
                'album rock', 'ohio hip hop', 'italian adult pop', 'bedroom pop',
                'garage rock', 'musical advocacy', 'brega funk', 'afroswing',
                'afrofuturism', 'german drill', 'k-pop girl group', 'new wave pop',
                'big room', 'icelandic pop', 'australian psych'], dtype=object)
```

Equation of the Final Model

$$\log\left(\frac{\hat{p}_{Hiphop}}{1-\hat{p}_{Hiphop}}\right) = -4.3158 + 4.0778x_{Danceability} - 0.1279x_{Loudness} + 7.9076x_{Speechiness} - 1.0820x_{Acousticness} + 2.1592x_{Liveness} - 2.8531x_{Valence}$$

Test Statistic & p-value

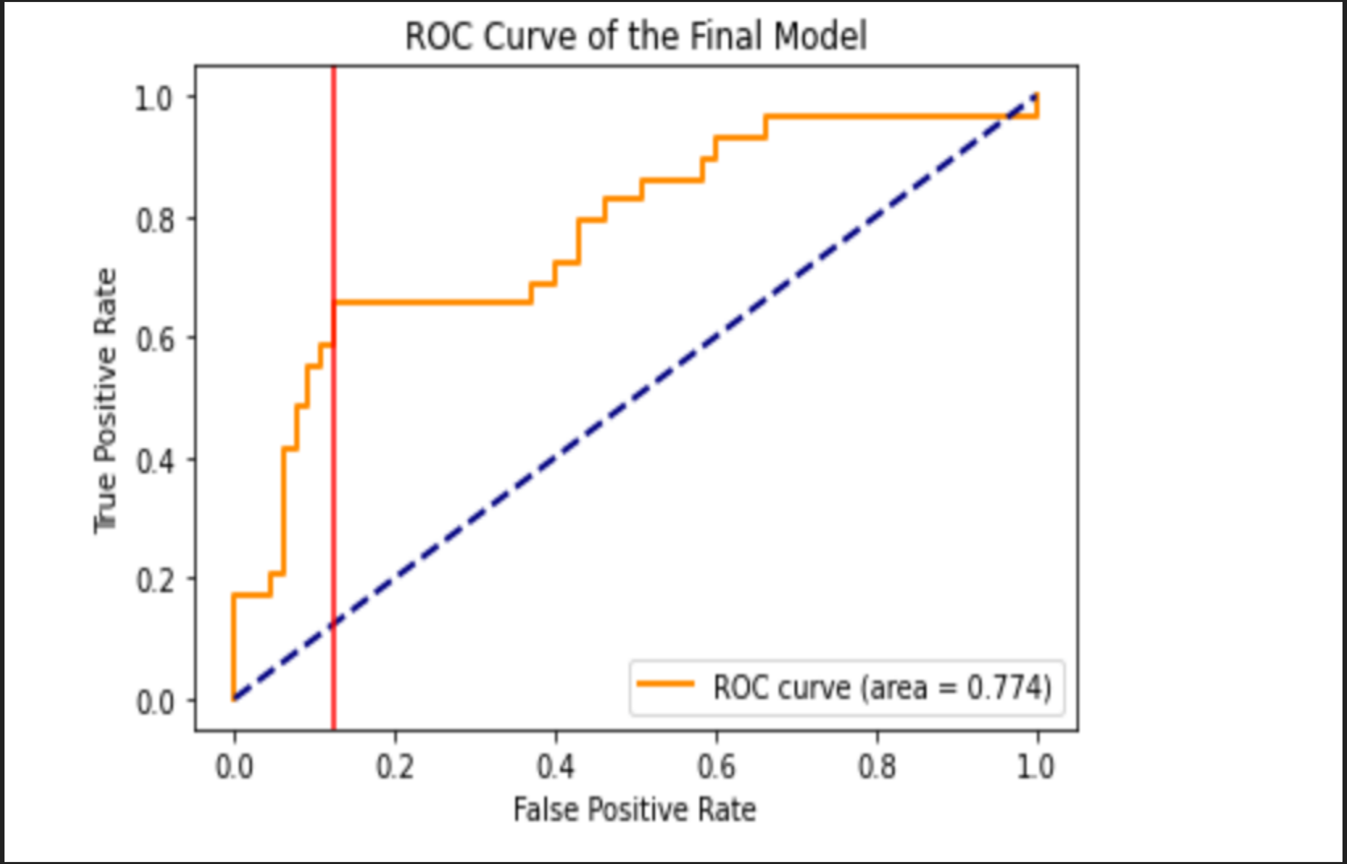
In [55]:
```
test_stat = -2*(final_mod.llf - full_mod.llf)
test_stat
```

Out[55]: 1.6627923135308151

In [56]:
```
from scipy.stats import chi2

p_val = 1 - chi2.cdf(test_stat, df = 3)
p_val
```

Out[56]: 0.6452373459230079


ROC Curve of the Final Model
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.2
0.4
0.6
0.8
1.0
ROC curve (area = 0.774)

Conclusion