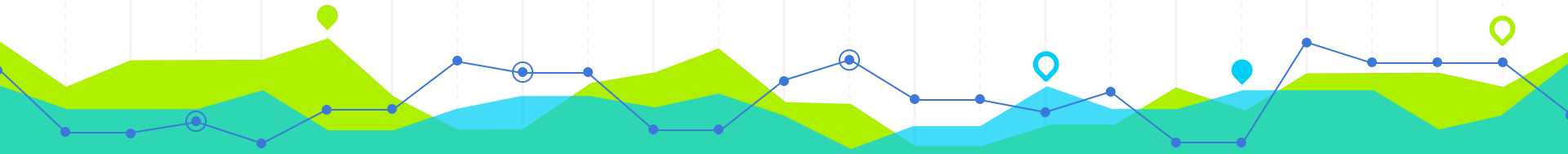
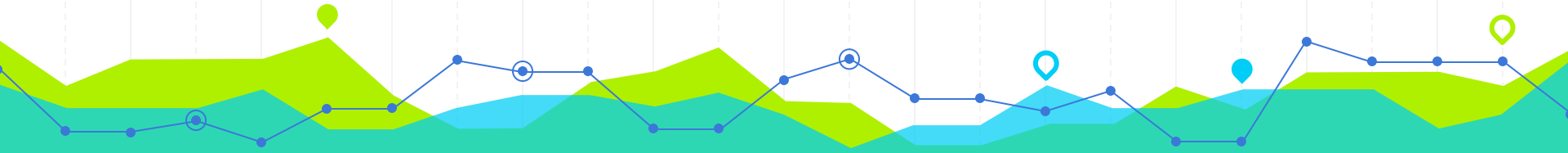


Song Popularity Analysis

By Cecilia Iacometta, Nicholas Tiveron, Ginevra Cepparulo

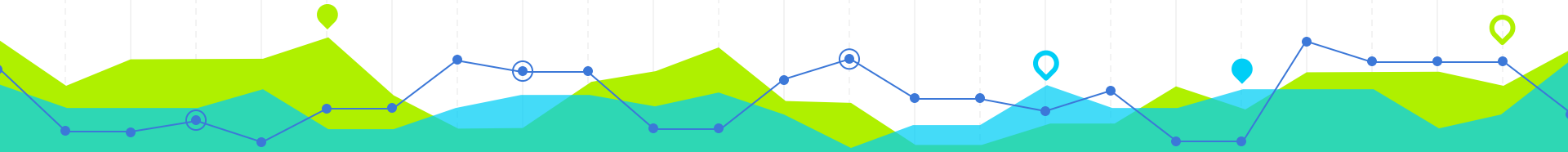
ROADMAP


Introduction and RQ
1
Variables
2
Methodology
3
Analysis
4
Conclusions
5
References
6



What is our project about?

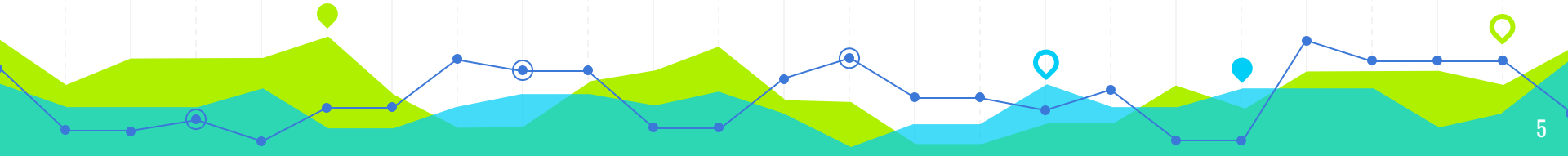
1



Relationship between

Popularity of songs and song metrics.

“The underlying assumption behind HSS is that popular songs are similar with respect to a set of features that make them appealing to a majority of people. These features could then be exploited by learning machines in order to predict whether a song will rise to a high position in the chart” - someone



Why we are interested and why is it important?

- Relevance: for musicians, music listeners as well as streaming platforms and industry
  - Did you ever wonder what characterizes top songs in the billboards?
  - Spotify pays record companies loyalties based on how many streams they produce.
  - Insights about musical tastes of the population from which the data is drawn.

Why Spotify?

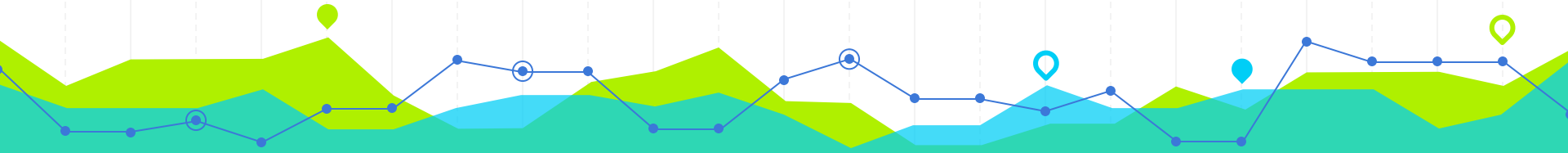
The platform totally dominates the streaming market.

200 mil users

That’s a lot of users following artists, browsing top charts and hunting for music.

Research Question:

To what extent can a song's popularity be explained by its audio features?



Variables 2

Independent Variables

Attribute	Scale	Explanation
Acousticness	0 - 1	Likelihood of a track being acoustic
Danceability	0 - 1	Suitability for dancing
Duration	ms	Length of the track
Energy	0 - 1	Measure of intensity and activity
Instrumentalness	0 - 1	Likelihood of having no vocals
Liveness	0 - 1	Likelihood of detecting an audience in the recording
Loudness	-60 - 0 (dB)	Loudness of a track
Mode	0 or 1	Indicates whether a song is Major (1) or minor (0)
Speechiness	0 - 1	Presence of spoken words in the track
Tempo	BPM	Estimated tempo of a track
Key	0 - 11	Key of the track, in pitch class notation
Valence	0 - 1	The musical positivity conveyed

Table 1: Features tracked

Dependent variable: Popularity

- Spotify metric for song popularity 0-100 based
  - total number of plays the track has
  - how recent those plays are
- Songs that are being played a lot now have a higher popularity than songs that were played a lot in the past.

LET'S HYPOTHESIZE

H1: Acousticness is negatively correlated to song popularity.

Given the high current trend in electronic music in Italy in recent years it is believed that acousticness will show negative relationship with popularity.

H2: Danceability is positively correlated to song popularity

Since electronic music is generally also danceable, it is believed that danceability will show a positive relationship with popularity.

H3: Duration is negatively correlated to song popularity

Most songs today are within a restricted range; those that last beyond generally have lower popularity.

H4: Energy is positively correlated to song popularity

Since today music is in many times an accompaniment to physical activities it makes sense that many popular songs are energetic.

rumentalness is ely correlated to song ity.

Given that lyrics allow the person to sing along to the song and hence to remember it more easily it is expected to be inversely related to song popularity.

C and E are positively ted to song popularity.

Experts say that the keys G,C and E are most commonly used and will therefore be seen in popular songs.

...some more HYPOTHESES

H7: Liveness is negatively correlated to song popularity.

Since the presence of an audience and hence background noise is thought to lower the quality of the audio recording, it is believed that liveness will be negatively correlated with song popularity.

H8: Loudness is positively correlated to song popularity.

It is thought that loudness is the most common way to show expressiveness so we could then assume that the louder the song the more popular.

H9: The Major mode is positively correlated to song popularity

Songs written in major scales are cheerful whereas ones written in minor scales tend to be sad. It is believed that most people prefer to listen to happy music.

eechiness is negatively ted to song popularity

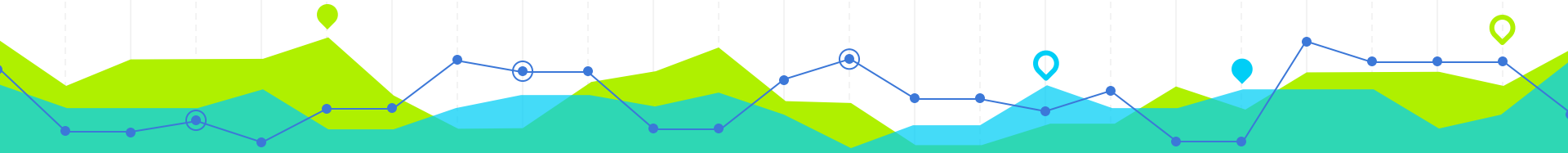
It is believed that a medium level of speechiness is most popular nowadays and hence that speechiness is negatively correlated to song popularity.

H11: Tempo is not correlated to song popularity.

Given the assumption that there are both hip-hop and ballad tracks in the top charts, we believe that this variable will not show a correlation to song popularity.

H12: Valence is positively correlated to song popularity.

Since it is believed that cheerful songs are more popular, it is hypothesized that popularity correlates positively with respect to song popularity.



Methodology 3

Sample and Data collection

50 top songs for each of the **10** most popular genres in the Italian market for the year **2021**.

A single song could be labelled with more than one genre, leading to **468** songs total.

Ordinary Least Squares

Minimizes SSE

$$RSS = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

$$= \sum_{i=1}^{n} (Y_i - \hat{b}_0 - \hat{b}_1 X_i)^2$$

Assumptions of the linear model

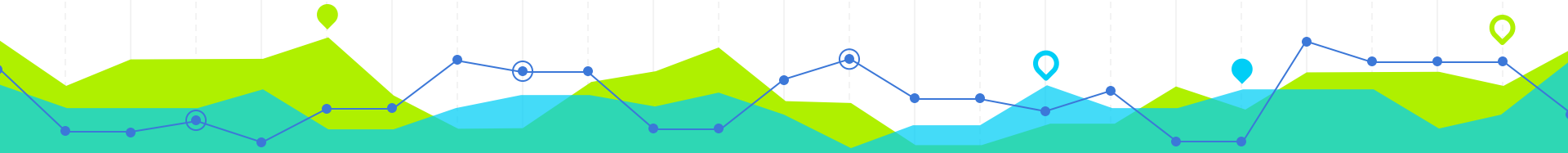
Linear Relationships
Checked by the entire analysis process which concludes that it is respected by conducting several tests.

Multivariate normality
The distribution of popularity was plotted and it showed to closely resemble a normal distribution.

No multicollinearity
Checked that each of the explanatory variables we chose did not have high correlation coefficients with any other variable with some exceptions.

No auto-correlation
Auto-correlation is only an issue in the case of time series.

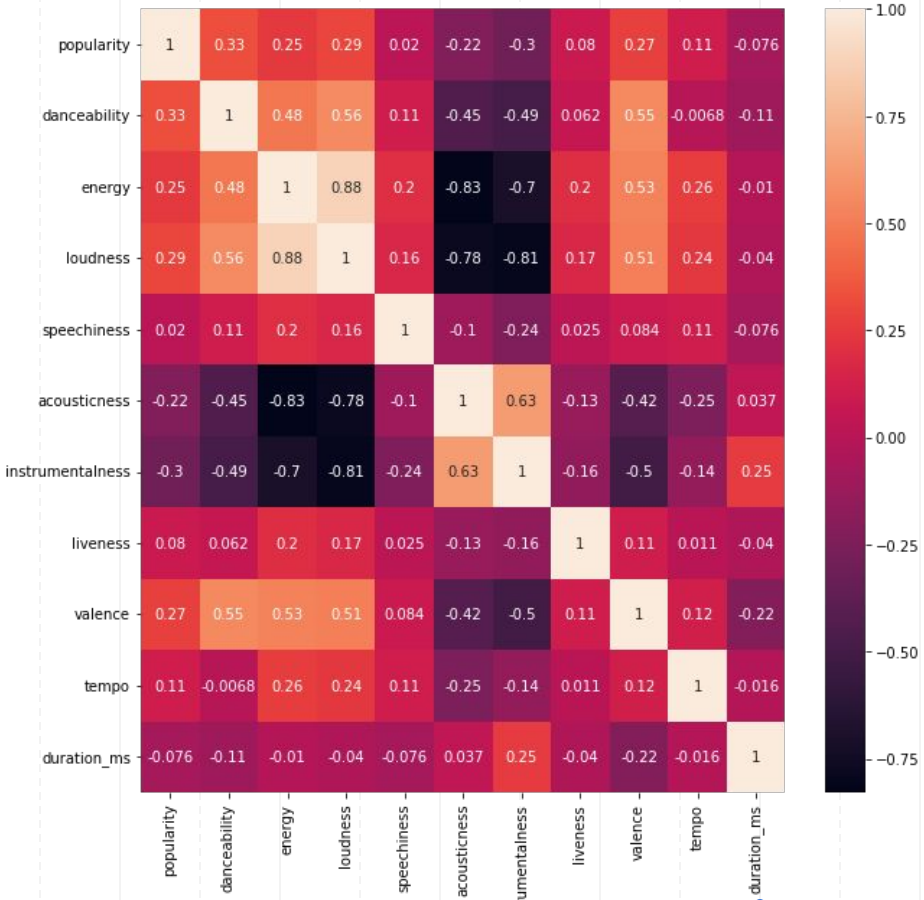
Homoscedasticity
Observed vs residuals plots all showed the final linear model, which violates it.



Analysis 4


Correlation analysis
Exploratory data analysis
Regression analysis

Correlation Matrix

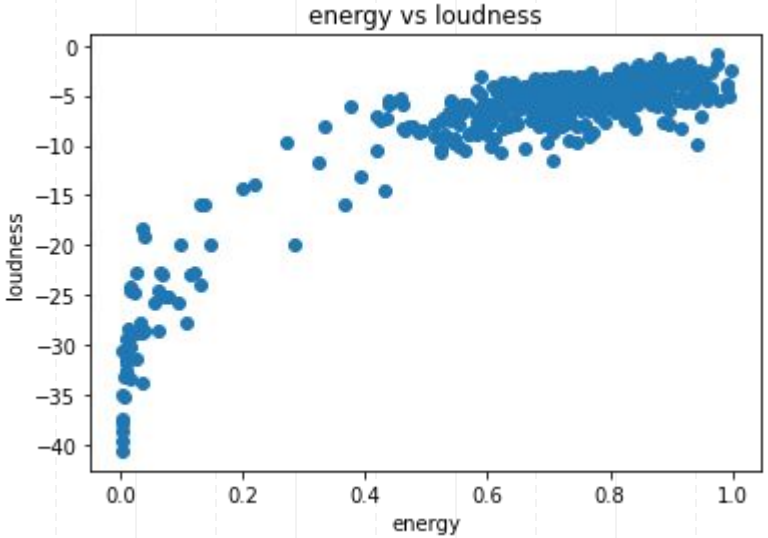


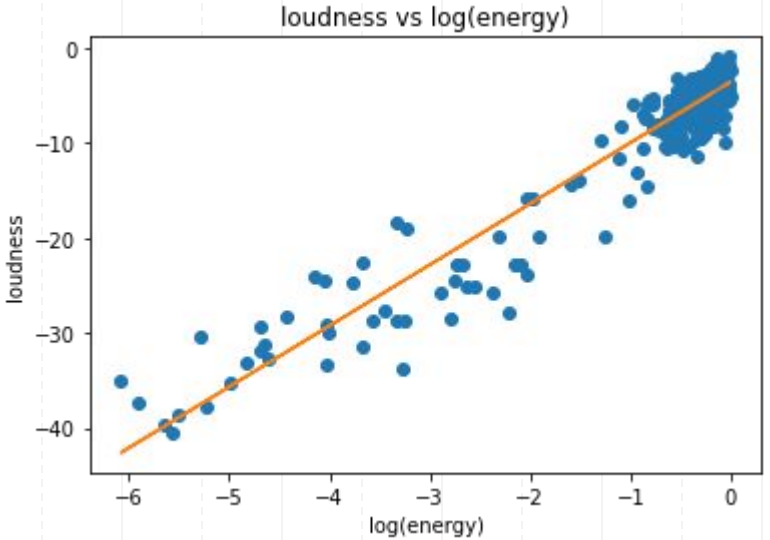
CHOSEN VARIABLES

Correlation coefficients for the variables we chose:

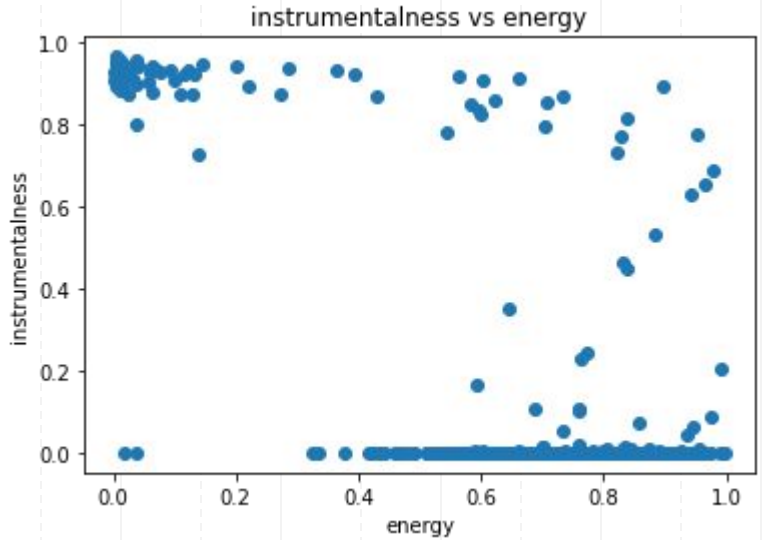
- **Danceability: 0.32**
- **Energy: 0.24**
- **Loudness: 0.29**
- **Valence: 0.27**
- **Instrumentalness: -0.3**

Loudness vs Energy

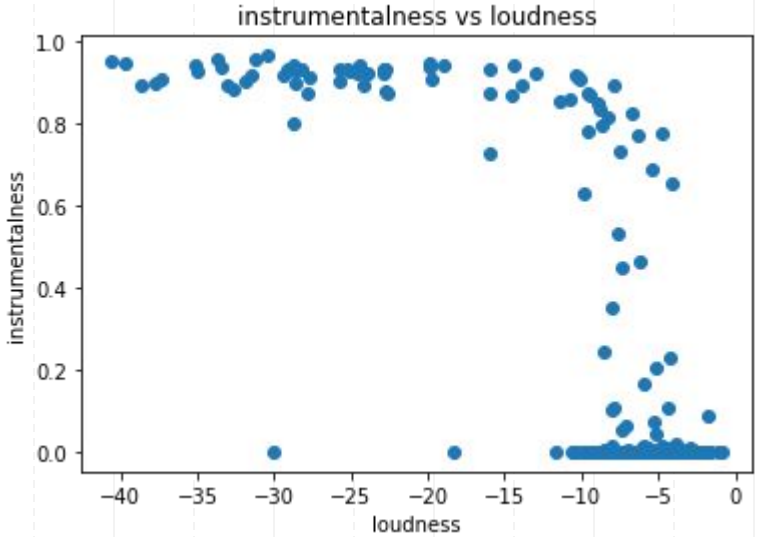




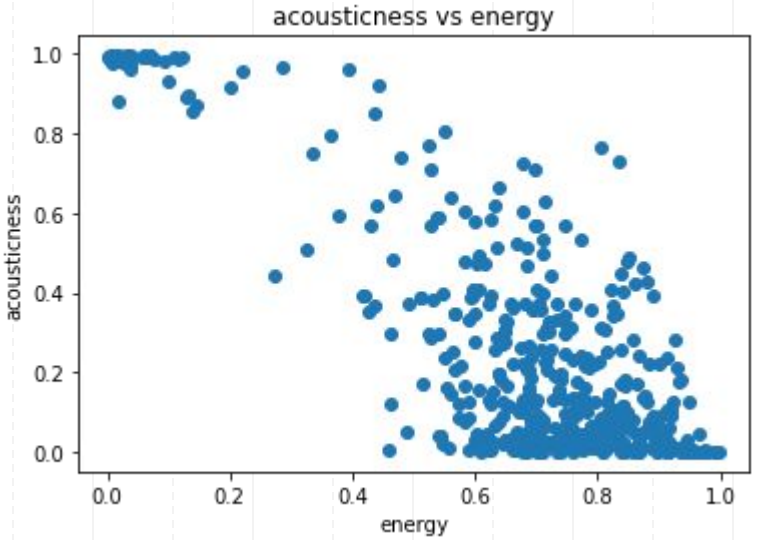
Instrumentalness vs Energy

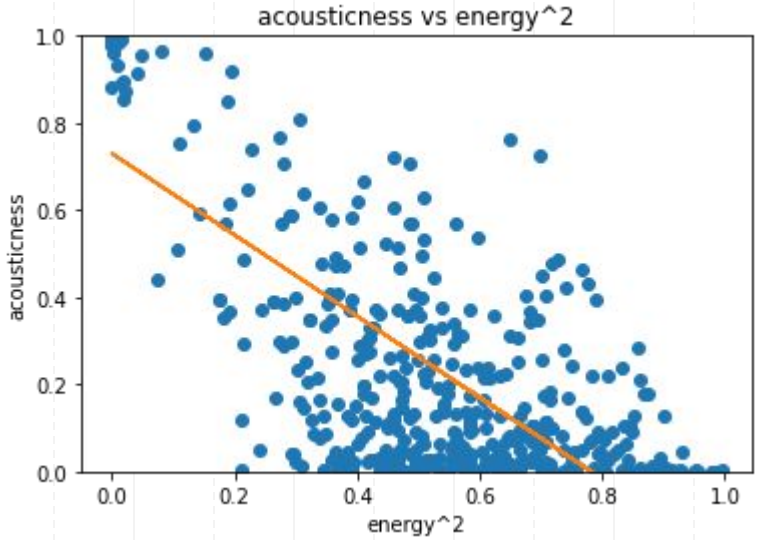


Instrumentalness vs Loudness



Acousticness vs Energy





First model's results

	coef	std err	t	P>|t|
Intercept	54.9891	5.197	10.580	0.000
danceability	14.5856	4.309	3.385	0.001
loudness	0.1462	0.196	0.744	0.457
instrumentalness	-5.3294	2.917	-1.827	0.068
valence	4.7311	3.031	1.561	0.119
energy	-3.0865	4.777	-0.646	0.519

R-squared:	0.137
Adj. R-squared:	0.128
F-statistic:	14.65
Prob (F-statistic):	2.39e-13

First model's results

$$\widehat{\mu} = 54.4721 + 15.5545 * x_1 + 0.1270 * x_2 - 5.5977 * x_3 + 4.7753 * x_4 - 3.2302 * x_5$$

where:

$x_1 = danceability$
$x_2 = loudness$
$x_3 = instrumentalness$
$x_4 = valence$
$x_5 = energy.$

Second model's results

	coef	std err	t	P>|t|
Intercept	71.7610	6.752	10.628	0.000
danceability	12.5250	4.362	2.871	0.004
loudness	1.1581	0.436	2.654	0.008
instrumentalness	-43.3837	11.750	-3.692	0.000
valence	3.4702	3.031	1.145	0.253
energy	-11.9551	6.548	-1.826	0.069
energy:loudness	0.7432	0.630	1.180	0.239
instrumentalness:loudness	-1.8632	0.586	-3.181	0.002
instrumentalness:energy	37.3804	12.616	2.963	0.003

R-squared:	0.179
Adj. R-squared:	0.165
F-statistic:	12.53
Prob (F-statistic):	2.58e-16

Mallows $c_p = 9$

Second model's results

$$\widehat{\mu} = 71.5089 + 13.3781 * x_1 + 1.1888 * x_2 - 44.5080 * x_3 + 3.5454 * x_4 - 13.2385 * x_5 + 0.5031 * x_6 - 1.9270 * x_7 + \\ +37.1912 * x_8$$

where:

$x_1 = danceability$
$x_2 = loudness$
$x_3 = instrumentalness$
$x_4 = valence$
$x_5 = energy$
$x_6 = energy * loudness$
$x_7 = instrumentalness * loudness$
$x_8 = instrumentalness * energy.$

Third model's results

	coef	std err	t	P>|t|
Intercept	72.9609	6.607	11.042	0.000
danceability	14.4018	4.012	3.589	0.000
loudness	1.4836	0.344	4.313	0.000
instrumentalness	-46.9167	11.349	-4.134	0.000
energy	-14.0559	5.638	-2.493	0.013
instrumentalness:loudness	-2.2678	0.486	-4.663	0.000
instrumentalness:energy	34.9281	12.542	2.785	0.006

R-squared:	0.174
Adj. R-squared:	0.163
F-statistic:	16.19
Prob (F-statistic):	6.12e-17

Mallows $c_p = 7$

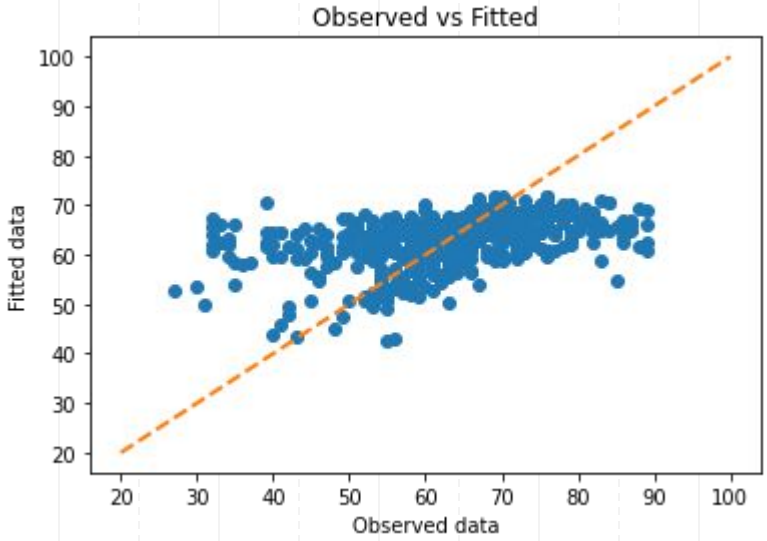
Third model's results

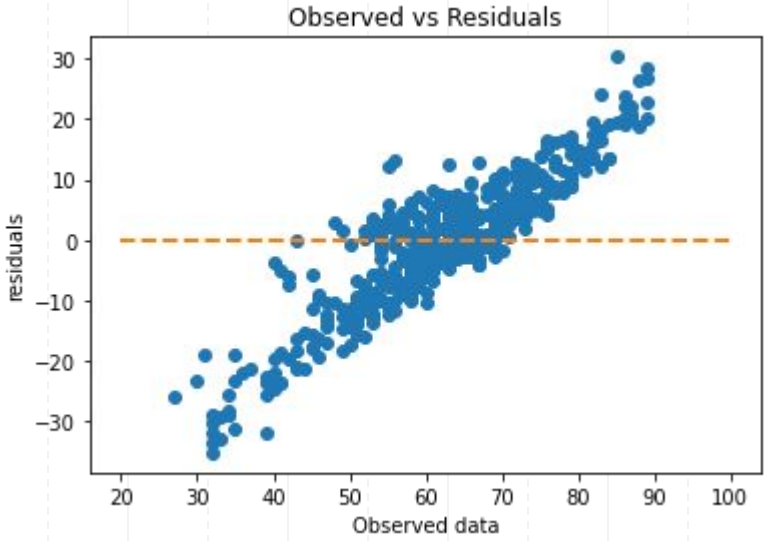
$$\widehat{\mu} = 72.9609 + 14.4018 * x_1 + 1.4836 * x_2 - 46.9157 * x_3 - 14.0559 * x_4 - 2.2678 * x_5 + 34.9281 * x_6$$

where:

$x_1 = danceability$
$x_2 = loudness$
$x_3 = instrumentalness$
$x_4 = energy$
$x_5 = instrumentalness * loudness$
$x_6 = instrumentalness * energy.$

Second model's evaluation





The final model (standardized coefficients)

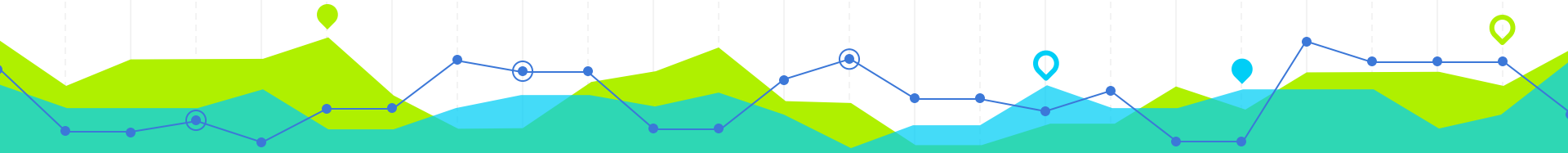
$$\widehat{\mu} = -0.2045 + 0.1646 * x_1 + 0.8080 * x_2 - 0.1042 * x_3 + 0.0644 * x_4 - 0.2460 * x_5 + 0.1043 * x_6 - 0.3406 * x_7 + \\ +0.2311 * x_8$$

where:

$x_1 = danceability$
$x_2 = loudness$
$x_3 = instrumentalness$
$x_4 = valence$
$x_5 = energy$
$x_6 = energy * loudness$
$x_7 = instrumentalness * loudness$
$x_8 = instrumentalness * energy$

CI's for the final model

- 95% confidence interval of (61.50, 63.56) for the mean of popularity
- 95% prediction interval of (40.22, 84.85) for a new observation Y_0



Conclusion 5

To sum up...

Given our model, it is concluded that it was not possible to explain a song's popularity by the analysis of its audio features.

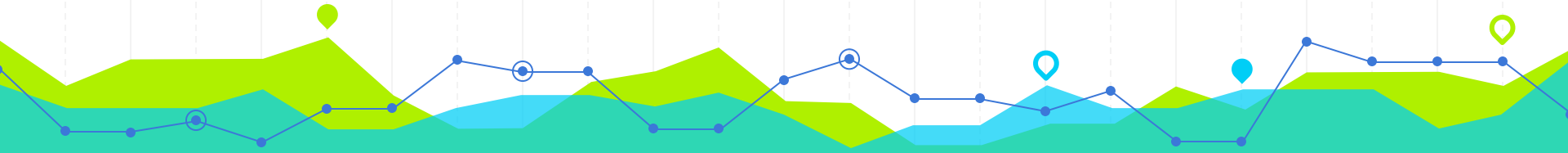
Reasons:

- the metric used to describe a song's Popularity
- transformations for some independent variables
- genre focused analysis would have been better

Implications

Further research:

- other models other than OLS
- prediction models such as Random Forests or SVM, song popularity for out-of-sample data.
- very significant for music streaming platforms such as Spotify as well as the music Industry in general



References

6

- Nijkamp, Rutger.Prediction of Product Success: Explaining Song Popularity by Audio Features from Spotify Data, University of Twente, 10 July 2018, https://essay.utwente.nl/75422/1/NIJKAMP_BA_IBA.pdf.Accessed 6 Sept. 2022.
- Çimen, Ahmet, and Enis Kayış. A Longitudinal Model for Song Popularity Prediction, Ozyegin University, 2021, https://www.scitepress.org/Papers/2021/106077/106077.pdf. Accessed 6 Sept. 2022.
- Suh, Brendan Joseph. International Music Preferences: An Analysis of the Determinants of Song Popularity on Spotify for the U.S., Norway, Taiwan, Ecuador, and Costa Rica, Claremont Colleges, 29 Apr. 2019, https://core.ac.uk/download/pdf/216833222.pdf. Accessed 6 Sept. 2022.
- "F Statistic / F Value: Simple Definition and Interpretation." Statistics How To, 20 Nov. 2021, https://www.statisticshowto.com/probability-and-statistics/f-statistic-value-test/. Accessed 6 Sept. 2022.
- Fogarty, Taylor, et al. "Predicting the Future of Music." Towards Data Science, 24 May 2019, https://towardsdatascience.com/predicting-the-future-of-music-c2ca274aea9f. Accessed 6 Sept. 2022.
- Frost, Jim. "How to Interpret p-Values and Coefficients in Regression Analysis." Statistics By Jim, Jim Frost, 22 July 2022, https://statisticsbyjim.com/regression/interpret-coefficients-p-values-regression/ Accessed 6 Sept. 2022.
- Pham, James, et al. Predicting Song Popularity, http://cs229.stanford.edu/proj2015/140_report.pdf . Accessed 6 Sept. 2022.
- Watts, Cameron. "Extracting Song Data from the Spotify API Using Python." Extracting Song Data from the Spotify API Using Python, Towards Data Science, 10 Feb. 2022, https://towardsdatascience.com/extracting-song-data-from-the-spotify-api-using-python-b1e79388d50. Accessed 6 Sept. 2022.
- Ni, Y., Santos-Rodriguez, R., Mcvicar, M., & De Bie, T. (2015). Hit song science once again a science?, 1-2.

THANKS!

Any questions?