DECEMBER 1ST, 2022

Regression

NAVIGATING AND MODELING SPOTIFY DATA

Trends

MGSC 310 FINAL PROJECT

Music

Sarah Fieck & Sreya Vadlamudi

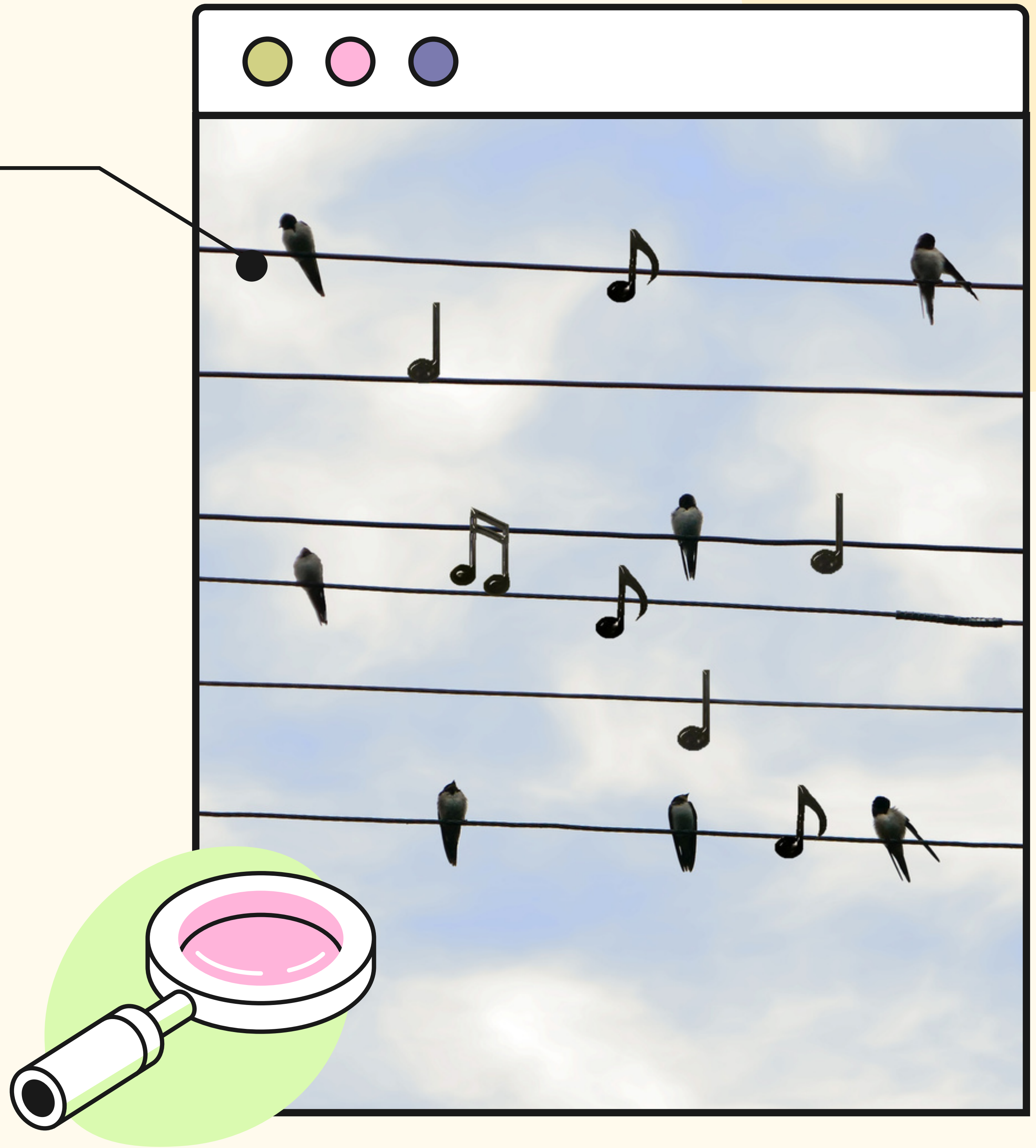
Decision Tree

MOTIVATION

1. MUTUAL INTEREST IN MUSIC ANALYTICS
2. BUSINESS STANDPOINT

DESCRIPTION OF SPOTIFY DATA

- Dataframe showing songs in Spotify Weekly Top Chart
- Variables
  - **Identification Data** - URI, track_name, artist_name, etc.
  - **Scores** - danceability, energy, key, loudness, etc.



SPOTIFY DATA

uri	artist_names	track_name	peak_ranl	weeks_on	danceabil	energy	key	loudness	mode	speechine	acousticn	instrumer	li
spotify:tra	Glass Animals	Heat Waves	1	65	0.761	0.525	11	-6.9	1	0.0944	0.44	6.70E-06	
spotify:tra	The Kid LAROI, Justin Bie	STAY (with Justin Bieber)	1	37	0.591	0.764	1	-5.484	1	0.0483	0.0383	0	
spotify:tra	Anitta	Envolver	3	3	0.812	0.736	4	-5.421	0	0.0833	0.152	0.00254	
spotify:tra	Imagine Dragons, JID, Ar	Enemy (with JID) - from the s	3	21	0.728	0.783	11	-4.424	0	0.266	0.237	0	
spotify:tra	GAYLE	abcdefu	1	19	0.695	0.54	4	-5.692	1	0.0493	0.299	0	
spotify:tra	Elton John, Dua Lipa, PN	Cold Heart - PNAU Remix	4	32	0.795	0.8	1	-6.32	1	0.0309	0.0354	7.25E-05	
spotify:tra	Becky G, KAROL G	MAMIII	5	6	0.843	0.7	4	-3.563	0	0.0803	0.0934	0	
spotify:tra	Lil Nas X, Jack Harlow	INDUSTRY BABY (feat. Jack Ha	2	35	0.741	0.691	10	-7.395	0	0.0672	0.0221	0	
spotify:tra	Rauw Alejandro, Chench	Desesperados	8	15	0.869	0.694	1	-3.35	0	0.0783	0.356	0.00125	
spotify:tra	Ed Sheeran	Shivers	4	28	0.788	0.859	2	-2.724	1	0.0856	0.281	0	
spotify:tra	Doja Cat	Woman	6	37	0.824	0.764	5	-4.175	0	0.0854	0.0888	0.00294	
spotify:tra	Adele	Easy On Me	1	23	0.604	0.366	5	-7.519	1	0.0282	0.578	0	

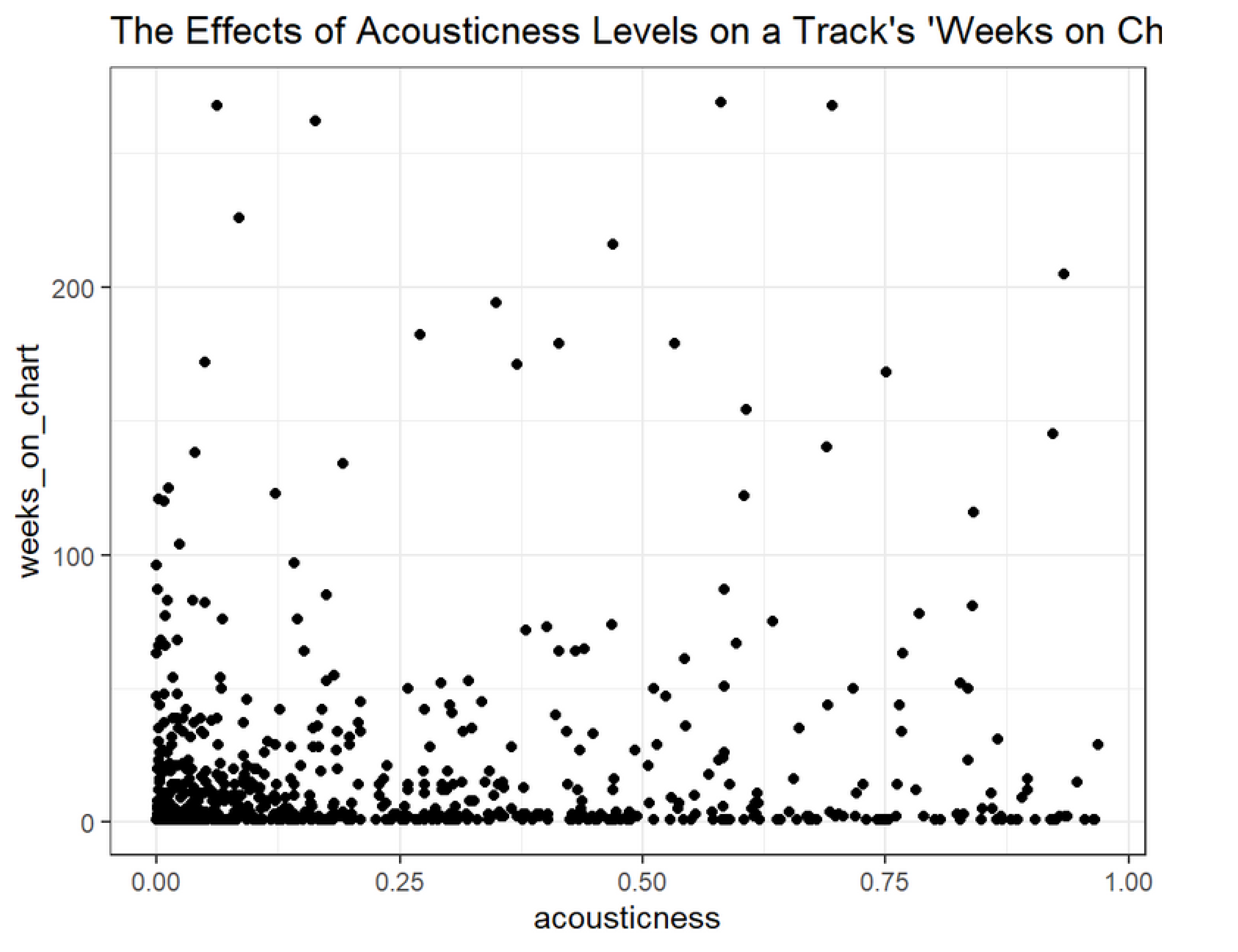
SPOTIFY DATA

Summary Statistics

Variable	N	Mean	Std. Dev.	Min	Pctl. 25	Pctl. 75	Max
peak_rank	646	65.923	57.005	1	15	108.75	200
weeks_on_chart	646	19.498	37.814	1	1	19.75	269
danceability	646	0.674	0.152	0.193	0.569	0.791	0.985
energy	646	0.641	0.165	0.022	0.532	0.769	0.972
key	646	5.087	3.622	0	1	8	11
loudness	646	-6.356	2.627	-31.16	-7.716	-4.596	-0.514
mode	646	0.577	0.494	0	0	1	1
speechiness	646	0.11	0.102	0.023	0.041	0.136	0.611
acousticness	646	0.256	0.263	0	0.04	0.424	0.969
instrumentalness	646	0.014	0.078	0	0	0	0.908
liveness	646	0.176	0.134	0.026	0.093	0.232	0.968
tempo	646	121.092	28.268	66.165	98.428	139.968	205.863
time_signature	646	3.927	0.351	1	4	4	5
duration_ms	646	203629.859	54966.666	36935	169901.5	229213.5	613027

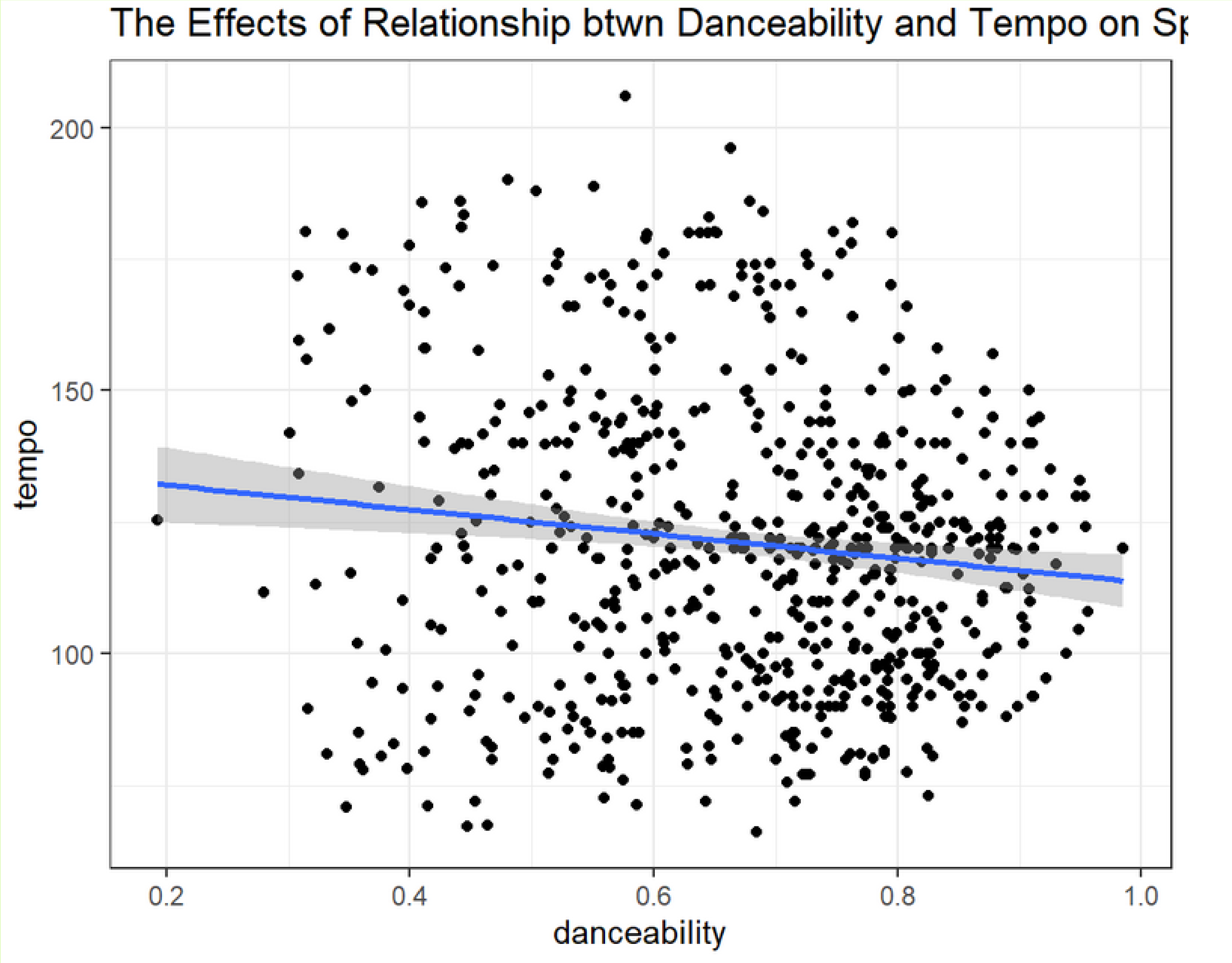
SPOTIFY DATA SUMMARIES

ACOUSTICNESS VS. TRACKS WEEKS ON CHARTS



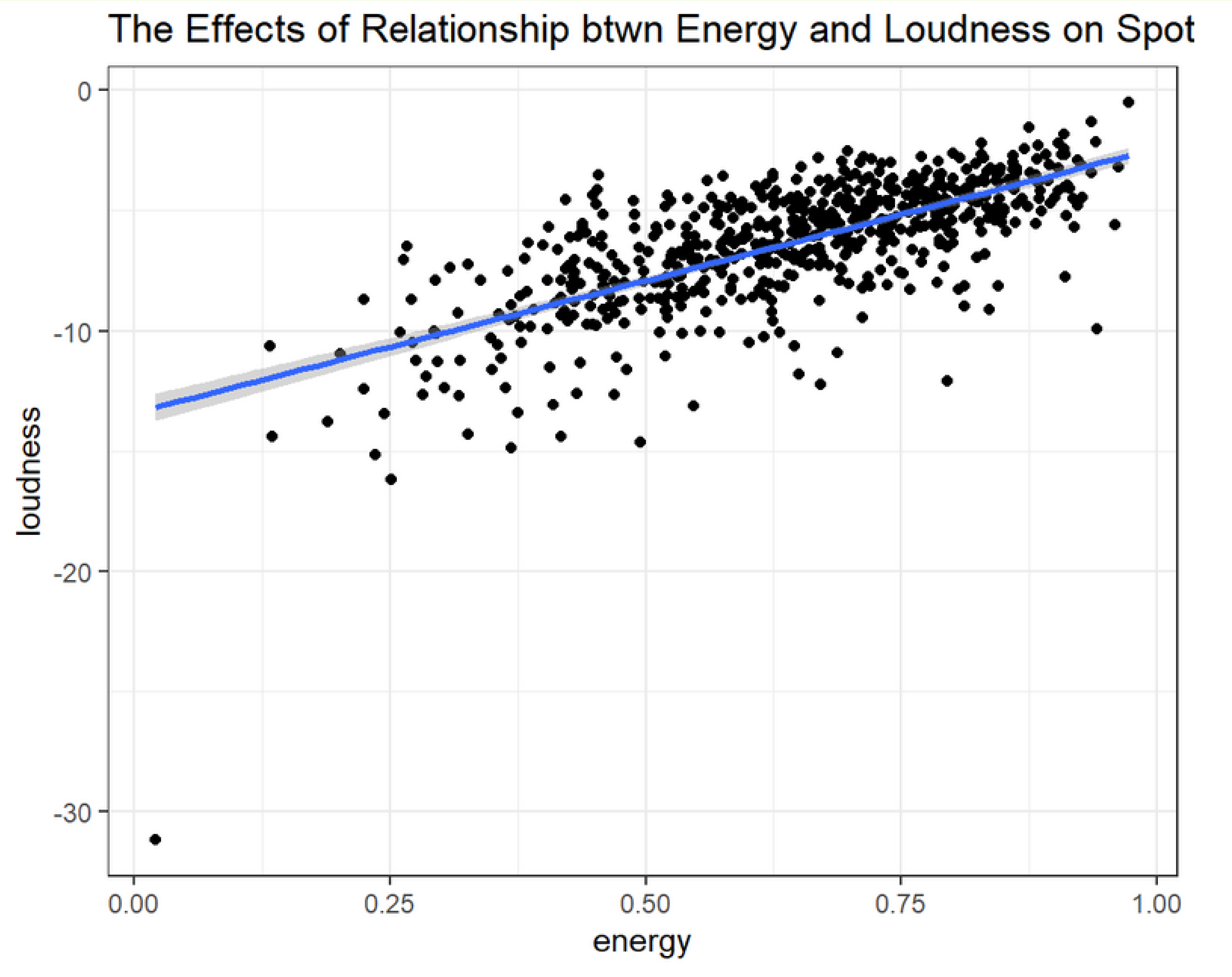
SPOTIFY DATA SUMMARIES

DANCEABILITY VS. TEMPO



SPOTIFY DATA SUMMARIES

ENERGY VS. LOUDNESS



LINEAR REGRESSION

- Chose LR due to its ability to predict continuous variables
  - Chose predictors to help predict a song's weeks on the Spotify charts

- Log transformed dependent variable because weeks on chart can be incredibly varied
  - Improves accuracy, risks interpretation

```r
# Linear Regression Model for Weeks on Char

spot_mod <-lm(formula = log(weeks_on_chart)
                   danceability +
                   energy +
                   loudness +
                   speechiness +
                   acousticness +
                   instrumentalness +
                   liveness +
                   tempo,
                   data = sd)

summary(spot_mod)
```

LINEAR REGRESSION

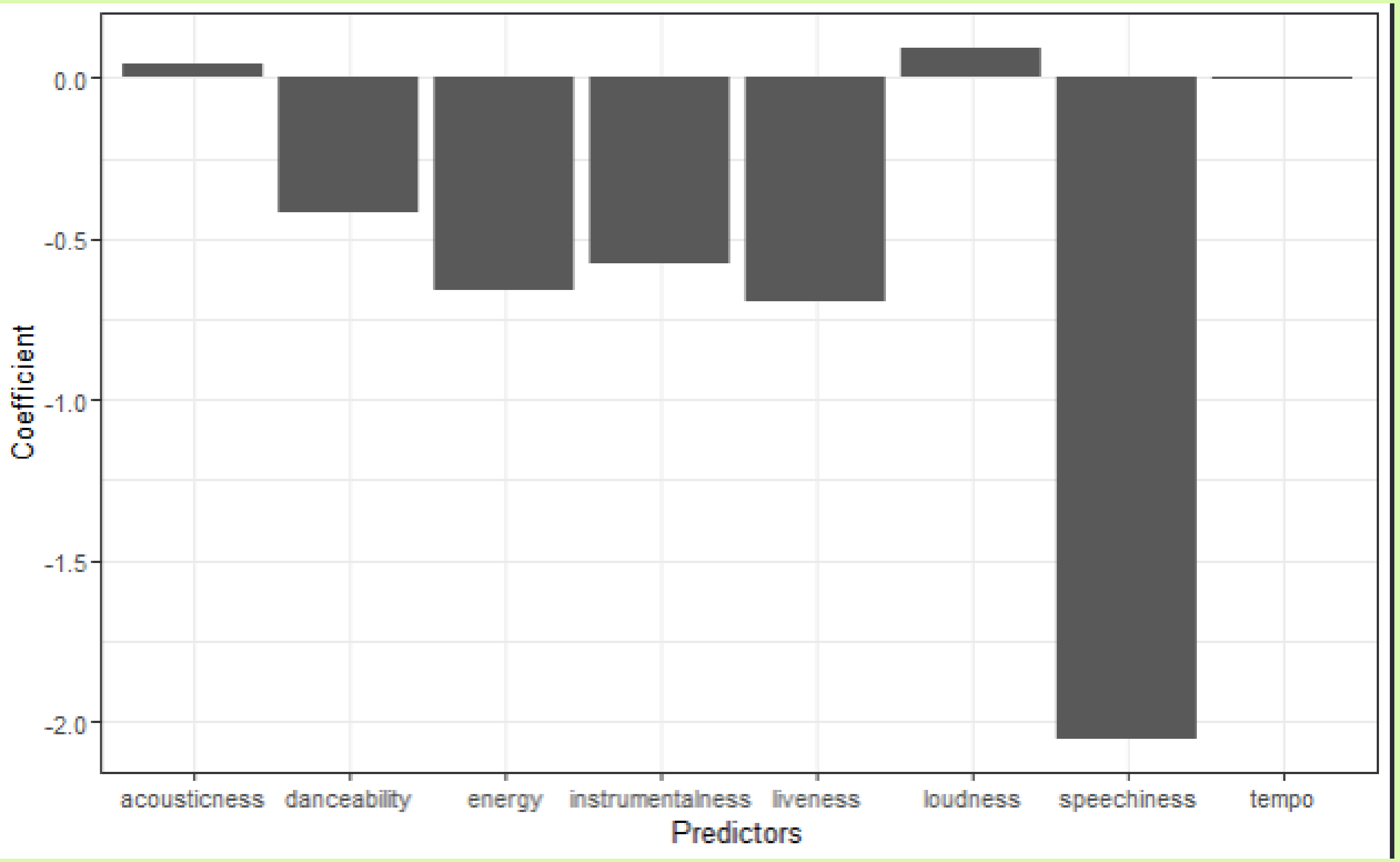
```
Call:
lm(formula = log(weeks_on_chart) ~ danceability + energy + loudness +
    speechiness + acousticness + instrumentalness + liveness +
    tempo, data = sd)

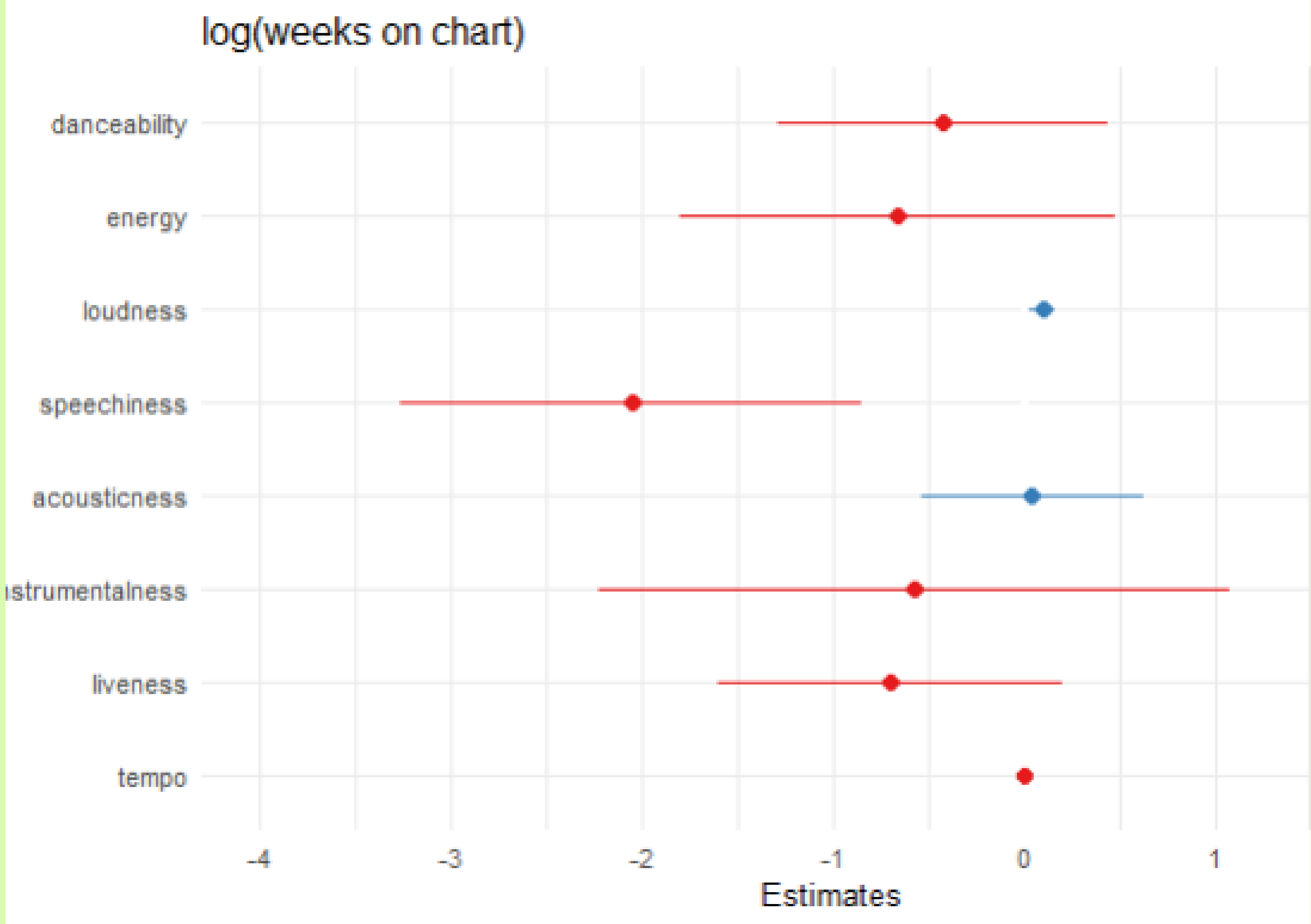
Residuals:
    Min      1Q  Median      3Q     Max
-2.3186 -1.4253 -0.1357  1.1632  4.1687

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       3.5078948  0.7231648   4.851 1.55e-06 ***
danceability     -0.4202763  0.4401329  -0.955 0.339998
energy           -0.6596616  0.5791169  -1.139 0.255097
loudness          0.0956661  0.0342240   2.795 0.005342 **
speechiness      -2.0533180  0.6143221  -3.342 0.000879 ***
acousticness      0.0437226  0.2941453   0.149 0.881882
instrumentalness -0.5769564  0.8394172  -0.687 0.492125
liveness         -0.6962217  0.4588896  -1.517 0.129715
tempo            -0.0008873  0.0021986  -0.404 0.686678
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.543 on 637 degrees of freedom
Multiple R-squared:  0.04184,	Adjusted R-squared:  0.0298
F-statistic: 3.477 on 8 and 637 DF,  p-value: 0.0006148
```

LINEAR REGRESSION COEFFICIENTS





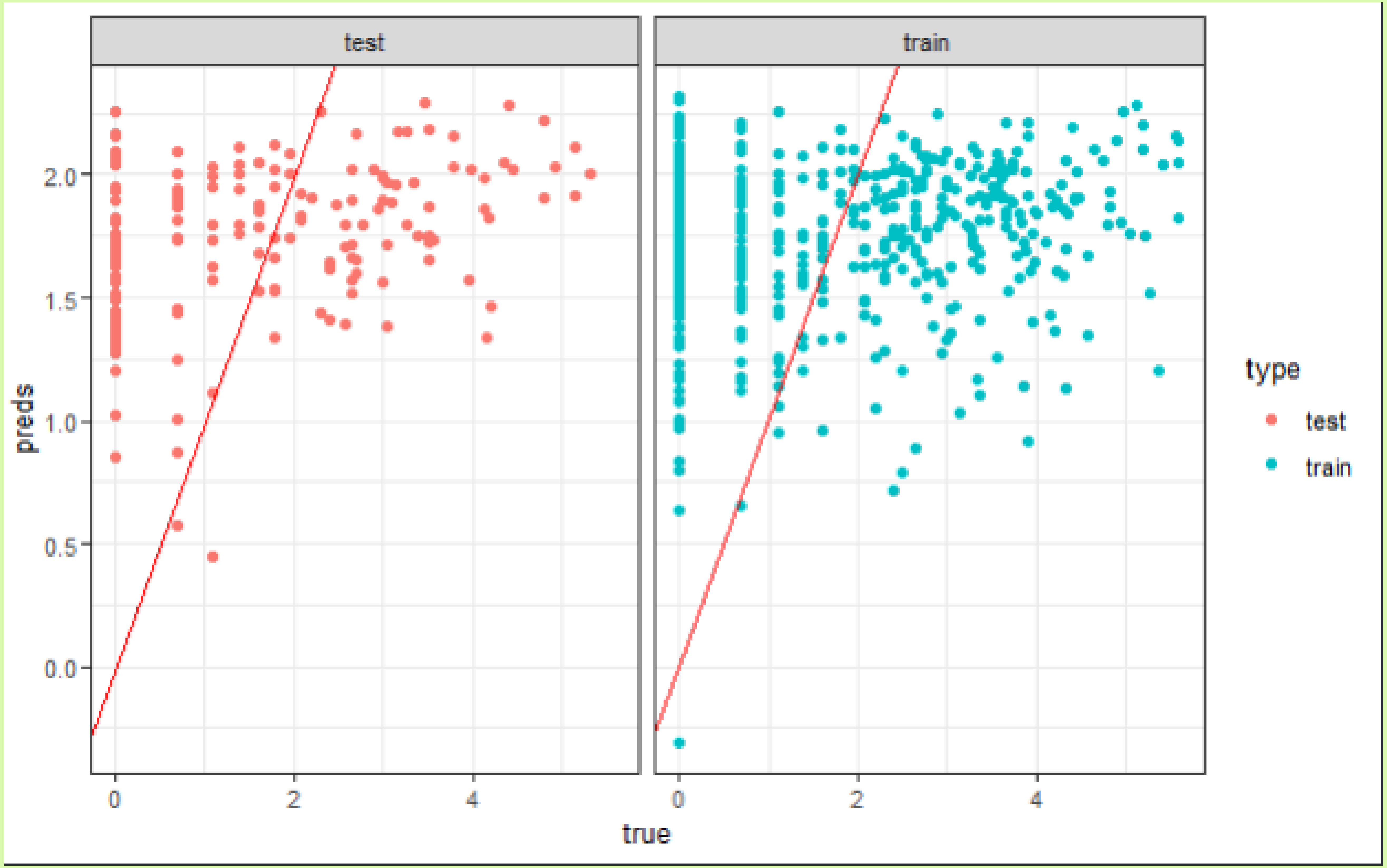
MODEL VALIDATION

- Used Train - Test Split of 75% to see how our model fits

preds <dbl>	true <dbl>	type <chr>
1.8713167	2.3978953	train
1.7603743	1.3862944	train
2.0055879	0.6931472	train
1.9540165	3.4965076	train
1.4782390	1.6094379	train
2.0224466	1.9459101	train
1.6862184	1.0986123	train
1.8119638	0.6931472	train
1.4380306	0.6931472	train
1.8329252	0.0000000	train

preds <dbl>	true <dbl>	type <chr>
2.0939041	2.9957323	test
1.6738290	0.6931472	test
2.1570175	2.4849066	test
2.1855508	0.6931472	test
1.5202594	2.6390573	test
1.8947454	2.9957323	test
2.0872961	3.5553481	test
1.9934101	2.0794415	test
1.5654566	2.9444390	test
2.2564866	4.9767337	test

MODEL VALIDATION



MODEL VALIDATION METRICS

Median Average Error

- **Test:** 1.210288
- **Train:** 1.375389

Mean Average Error

- **Test:** 1.258735
- **Train:** 1.3464

RMSE

- **Test:** 1.456105
- **Train:** 1.557226

DECISION TREE

````
#Decision Tree
```{r}
spotify_split <- initial_split(sd, prop = 0.8)
spotify_train <- training(spotify_split)
spotify_test <- testing(spotify_split)

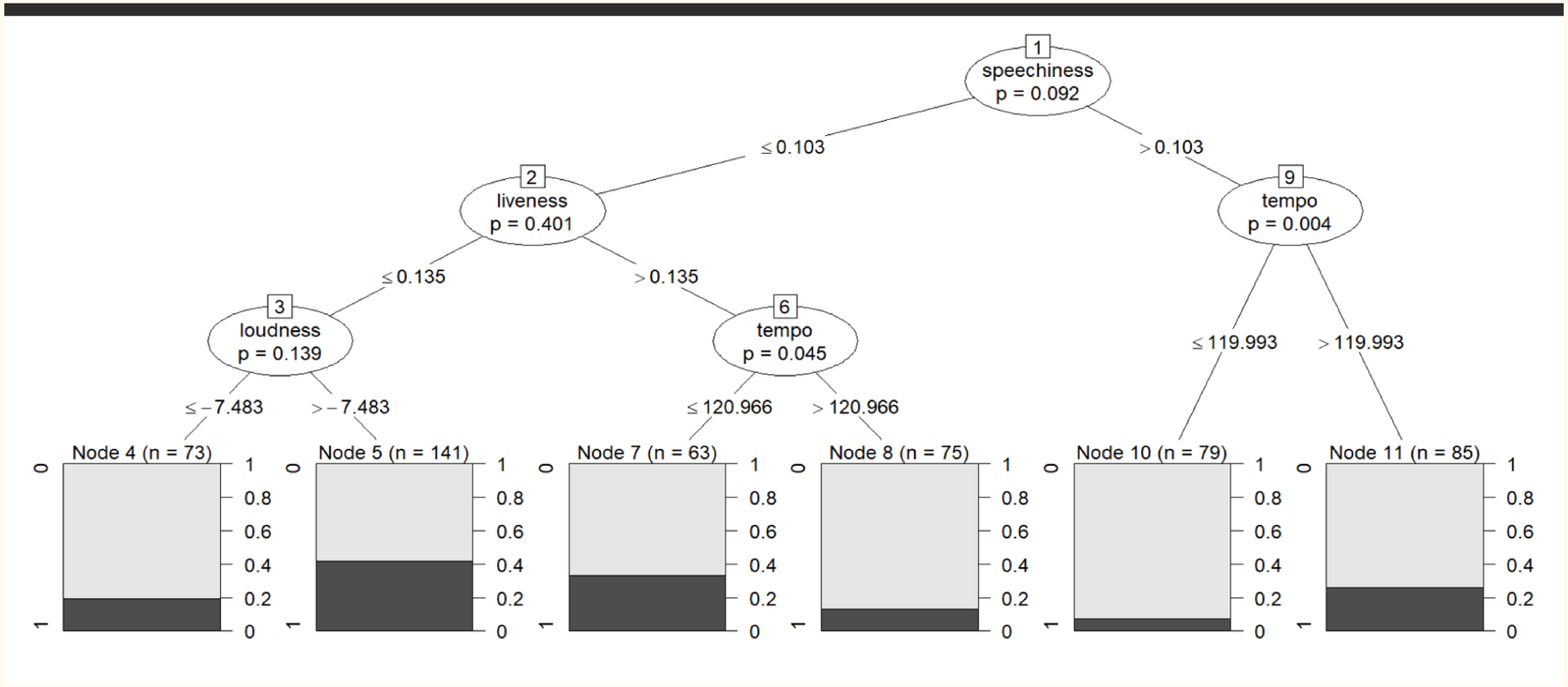
spotify_df <- spotify_train %>% as_tibble() %>%
  mutate(weeks_on_chart = if_else(weeks_on_chart>=20, "1", "0"), weeks_on_chart = as.factor(weeks_on_chart))

spot_mod2 <- ctree(weeks_on_chart ~ speechiness + acousticness +
                      instrumentalness +energy + loudness + danceability + tempo + liveness, data = spotify_df,  control =
partykit::ctree_control(alpha=0.5, minbucket = 60))

plot(spot_mod2)
````

- Chose Decision Tree in order to predict what variables most affect whether or not a song will be on the charts for more than or equal to 20 weeks.
- Cleaned the data by creating a binary variable for weeks on chart with 1 and 0
- Split into training and testing sets
````

- Most significant value was speechiness
- Speechiness was correlated most with tempo, liveliness, and loudness
- Danceability, energy, acousticness, and instrumentalness were not as important



COMPARISON & CONCLUSION

LINEAR

Based on the metrics on the training and testing sets for this model, we can conclude that it is pretty accurate.

It should be noted that we did log the dependent variable. This increases accuracy but risks interpretation.

DECISION TREE

This model helped predict which variables were most significant in predicting whether or not a song would last more than 20 weeks on the charts, but the percentages were low so not the best indicator.

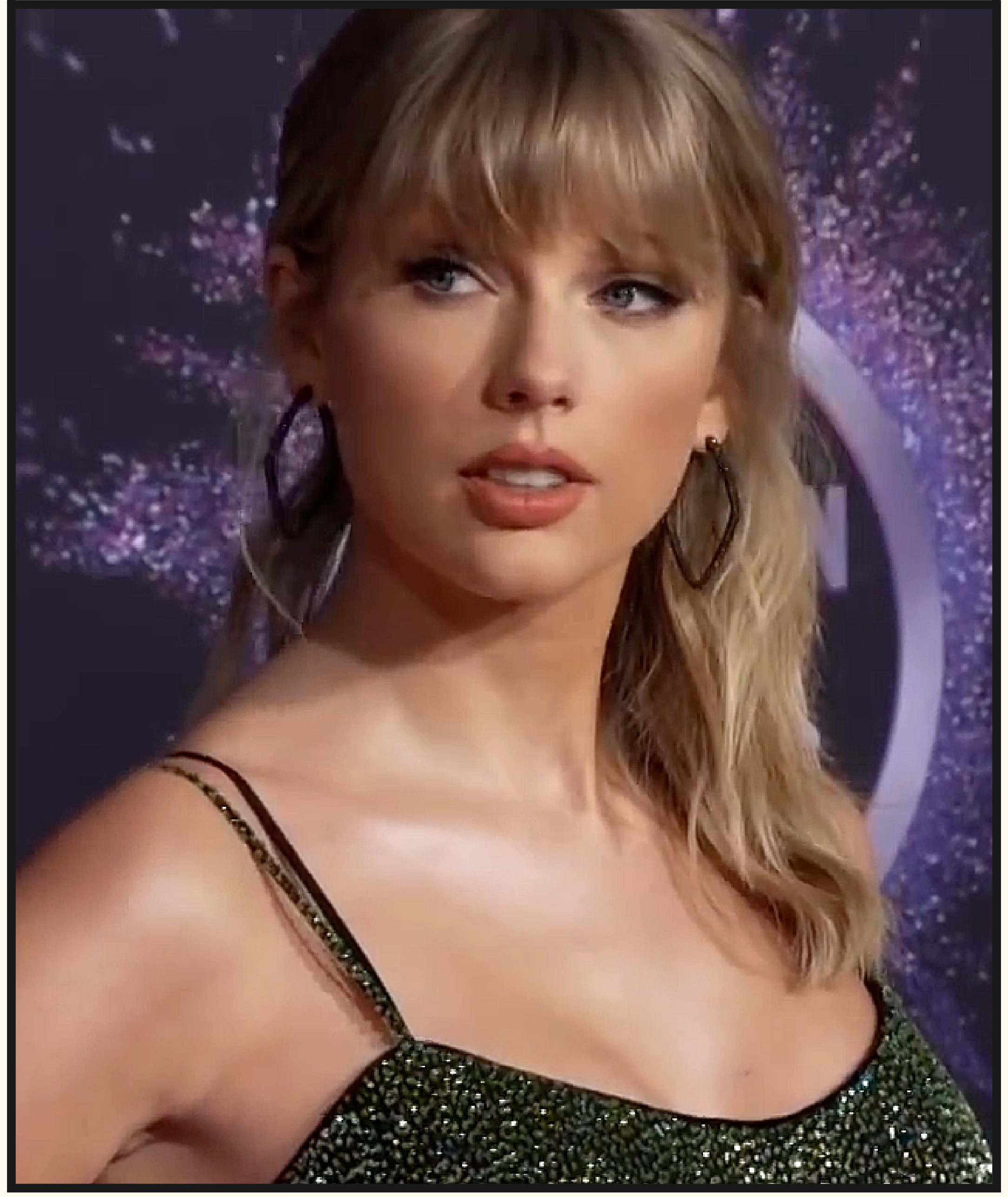
CONCLUSION

Overall, the linear model was better and would be the better model to use by the music industry to predict the weeks on chart for a song, which is why we used it to test Taylor Swift songs data using it as a test set.

LINEAR

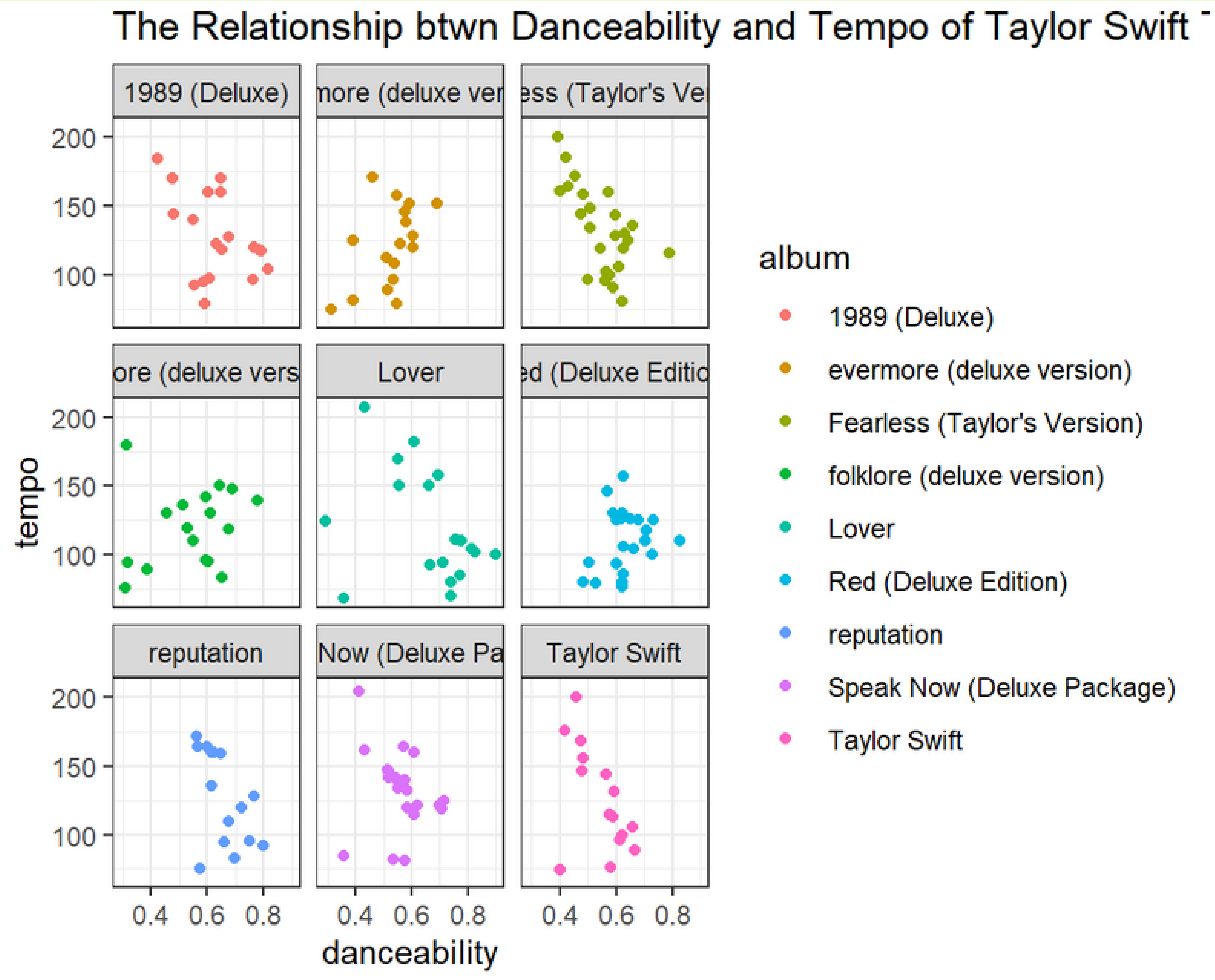
DATA (TAYLOR'S VERSION)

- Taylor Swift music data
- Similar to Spotify data, contains identification & similar scores
- Debut album - Fearless (Taylor's Version)



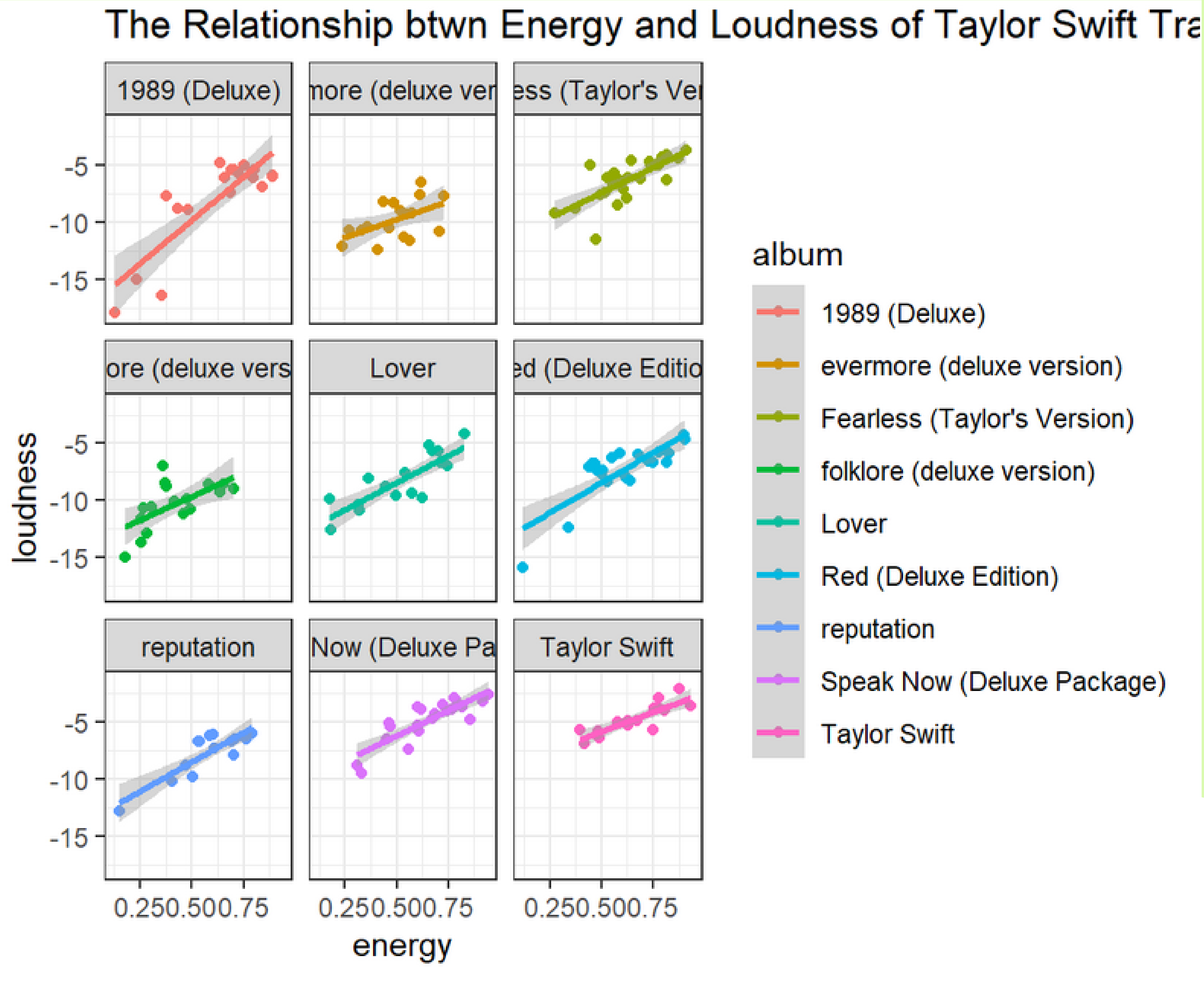
TS DATA SUMMARIES

DANCEABILITY OF TS TRACKS VS TEMPO PER EACH ALBUM



TS DATA SUMMARIES

ENERGY OF TS TRACKS VS LOUDNESS PER EACH ALBUM



RUNNING TS DATA THROUGH LINEAR REGRESSION

name <chr>	artist <chr>	preds <dbl>	type <chr>
Tim McGraw	Taylor Swift	2.1439655	test
Picture To Burn	Taylor Swift	2.2327049	test
Teardrops On My Guitar - Radio Single Remix	Taylor Swift	2.1014311	test
A Place in this World	Taylor Swift	2.0884957	test
Cold As You	Taylor Swift	2.1758325	test
The Outside	Taylor Swift	2.0141127	test
Tied Together with a Smile	Taylor Swift	2.2248547	test
Stay Beautiful	Taylor Swift	2.1138817	test
Should've Said No	Taylor Swift	2.0901508	test
Mary's Song (Oh My My My)	Taylor Swift	2.1674470	test

TS SONGS ON THE SPOTIFY CHARTS

track_name <chr>	weeks_on_chart <dbl>
Blank Space	20
All Too Well (10 Minute Version) (Taylor's Version) (From The Vault)	19
august	7
Don’t Blame Me	13
Enchanted	10
This Love (Taylor’s Version)	2
Wildest Dreams (Taylor's Version)	11
Carolina - From The Motion Picture “Where The Crawdads Sing”	1

track_name <chr>	weeks_on_chart <dbl>
Blank Space	2.9957323
All Too Well (10 Minute Version) (Taylor's Version) (From The Vault)	2.9444390
august	1.9459101
Don’t Blame Me	2.5649494
Enchanted	2.3025851
This Love (Taylor’s Version)	0.6931472
Wildest Dreams (Taylor's Version)	2.3978953
Carolina - From The Motion Picture “Where The Crawdads Sing”	0.0000000

COMPARED TO TESTING WITH TS DATA

track_name <chr>	weeks_on_chart <dbl>
Blank Space	2.9957323
All Too Well (10 Minute Version) (Taylor's Version) (From The Vault)	2.9444390
august	1.9459101
Don't Blame Me	2.5649494
Enchanted	2.3025851
This Love (Taylor's Version)	0.6931472
Wildest Dreams (Taylor's Version)	2.3978953
Carolina - From The Motion Picture "Where The Crawdads Sing"	0.0000000

name <chr>	preds <dbl>	type <chr>
Blank Space	1.951882	test
august	1.850105	test
Enchanted	2.257491	test

NOTE:

- It is difficult to truly predict what songs will become popular
- Many aspects of music cannot be measured as easily with a statistical model
    - Personal connection to the song, cultural significance, lyrical differentiation, subject matter, etc.
- TikTok has also impacted popular music
    - Amplifying smaller artists and revitalizing older tracks

LINKS & REFERENCES

- **Dataframe:** https://www.kaggle.com/datasets/sveta151/spotify-top-chart-songs-2022
- **Taylor Dataframe:** https://www.kaggle.com/datasets/thespacefreak/taylor-swift-spotify-data?select=spotify_taylorswift.csv
- **Our Repository:** https://github.com/sreyavadlamudi/MGSC310_Project

Emamzadeh, Arash. “Why Do Some Songs Become Popular? | Psychology Today.” Psycology Today, 7 June 2018, https://www.psychologytoday.com/us/blog/finding-new-home/201806/why-do-some-songs-become-popular.

Venkat, Mia. “TikTok Has Changed Music — and the Industry Is Hustling to Catch Up.” NPR, 22 May 2022. NPR, https://www.npr.org/2022/05/22/1080632810/tiktok-music-industry-gayle-abcdefu-sia-tai-verdes-celine-dion.