Bella Almeter, Rani Misra, Marco Teniente
STA 4713: Project

Introduction

Spotify is one of the leading digital music streaming platforms in today's entertainment industry. It enables over 574 million users access to hundreds of millions of songs and artists worldwide ("About Spotify."). One of the most critically acclaimed artists on the platform, Taylor Swift, has consistently been successful since her debut in 2007. Of the aforementioned 574 million users, over 108 million are Swift's monthly listeners, encompassing approximately 20% of users on the platform ("Taylor Swift."). With a career that has spanned three different decades and music that has transitioned through multiple different genres such as country, pop, and alternative, Swift's popularity and likeness seem to be exponentially increasing. With each reinvention of her musical persona, Swift has proven herself to be a pioneer in the music industry.

When it comes to Spotify alone, Swift continues to show that her only competition is herself. In October 2023, Swift became the most-streamed artist in a single day in Spotify history, and *1989 (Taylor's Version)* became Spotify's most-streamed album in a single day in 2023 thus far (Willman). While this feat is impressive enough independently, to even further prove her gravitas, the last artist to set the record for being the most-streamed in a single day was Swift herself with the 2022 release of her album *Midnights* (Willman).

As if her musical achievements were not enough, the soundwaves of her impact can be seen across many other industries such as travel, tourism, and just general consumer spending. Following the inception of her highly anticipated *The Eras Tour*, Swift's events have propelled the economies, or 'Swiftonomies' as some might say, to the level of those prior to the pandemic. For instance, The Federal Reserve Bank of Philadelphia stated that Swift's tour helped stimulate travel and tourism in the area, making the month of her visit the largest for hotel revenue in the city since the onset of the pandemic (Kuzub). Similarly, Swift's arrival to Las Vegas generated the highest consumer and tourism spending post-pandemic (Kuzub). It was even estimated that her two shows in Colorado in July 2023 could have boosted the state's gross domestic product with $140 million in consumer spending (Kuzub). Regardless of whether or not the average American might describe themselves as a Swiftie (the term which colloquially describes a fan of Swift), it is pretty certain that her presence has swiftly ingrained into American culture.

One might wonder what encourages such mass appeal. While there are a multitude of social and behavioral factors that could be analyzed to determine the cause, perhaps an inquiry into her artistry is a good start. Thus, to determine what exactly is the key to her popularity, a linear regression analysis will be conducted on Spotify's musical algorithmic factors to establish which predictors are the perfect mix towards success.

Discussion of the Data Set

This data has been collected from Spotify's application programming interface on Kaggle. It provides the name of Swift's songs, albums, their release dates, track number, Spotify ID and URL. Additionally, it measures the acousticness, danceability, energy, instrumentalness, liveness, loudness, speechiness, tempo, and valence. The popularity and duration of the tracks (in milliseconds) is also calculated in the data.

The original data set contained 530 observations. During cleaning, we retained any albums that were owned by Swift herself or were recorded as "Taylor's Version" to gain ownership. If a "Taylor's Version" of the album did not exist at the time of analysis, the original version was kept instead. After cleaning, 205 observations were kept and the 10 continuous variables to be analyzed are described below ("Get Track's Audio Features."):

- Acousticness is a confidence measure from 0.0 to 1.0 of whether the track is acoustic. 1.0 represents high confidence that the track is acoustic.
- Danceability describes how suitable a track is for dancing based on a combination of musical elements including tempo, rhythm stability, beat strength, and overall regularity. A value of 0.0 is least danceable and 1.0 is most danceable.
- Energy is a measure from 0.0 to 1.0 and represents a perceptual measure of intensity and activity. Perceptual features contributing to this attribute include dynamic range, perceived loudness, timbre, onset rate, and general entropy.
- Instrumentalness predicts whether a track contains no vocals. "Ooh" and "aah" sounds are treated as instrumental in this context. Rap or spoken word tracks are clearly "vocal". The closer the instrumentalness value is to 1.0, the greater likelihood the track contains no vocal content. Values above 0.5 are intended to represent instrumental tracks, but confidence is higher as the value approaches 1.0.
- Liveness detects the presence of an audience in the recording. Higher liveness values represent an increased probability that the track was performed live. A value above 0.8 provides strong likelihood that the track is live.

- Loudness describes the overall loudness of a track in decibels (dB). Loudness values are averaged across the entire track and are useful for comparing relative loudness of tracks. Loudness is the quality of a sound that is the primary psychological correlate of physical strength (amplitude). Values typically range between -60 and 0 db. Based on Spotify's API, the closer a value is to 0, the louder the track is ("Loudness Normalization").
- Speechiness detects the presence of spoken words in a track. The more exclusively speech-like the recording (e.g. talk show, audio book, poetry), the closer to 1.0 the attribute value. Values above 0.66 describe tracks that are probably made entirely of spoken words. Values between 0.33 and 0.66 describe tracks that may contain both music and speech, either in sections or layered, including such cases as rap music. Values below 0.33 most likely represent music and other non-speech-like tracks.
- Tempo describes the overall estimated tempo of a track in beats per minute (BPM). In musical terminology, tempo is the speed or pace of a given piece and derives directly from the average beat duration.
- Valence is a measure from 0.0 to 1.0 describing the musical positiveness conveyed by a track. Tracks with high valence sound more positive (e.g. happy, cheerful, euphoric), while tracks with low valence sound more negative (e.g. sad, depressed, angry).
- Popularity measures the popularity of the song from 0 to 100.

Modeling

In terms of the model selection process, a backwards selection method was employed to evaluate which predictors were most influential in model significance. The respective method eliminated liveness, speechiness, instrumentalness, and acousticness in favor of factors like danceability, loudness, tempo, valence, and energy. This provided an R^2 value of 0.9857 and an adjusted R^2 value of 0.9854.

Summary of Backward Elimination

Step	Variable Removed	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	liveness	8	0.0000	0.9861	7.0044	0.00	0.9471
2	speechiness	7	0.0000	0.9861	5.1708	0.17	0.6830
3	instrumentalness	6	0.0002	0.9859	5.3625	2.21	0.1385
4	acousticness	5	0.0002	0.9857	5.8962	2.54	0.1125

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	1143114	228623	2766.92	<.0001
Error	200	16525	82.62725		
Uncorrected Total	205	1159639			

Root MSE	9.08995	R-Square	0.9857
Dependent Mean	74.87317	Adj R-Sq	0.9854
Coeff Var	12.14047		

Results

The prediction equation is popularity = 40.98652*(danceability) - 2.28497*(loudness) + 0.07642*(tempo) - 7.62950*(valence) + 46.46287*(energy).

This indicates that danceability, tempo, and energy have a positive effect on popularity whereas loudness and valence have a negative effect on popularity.

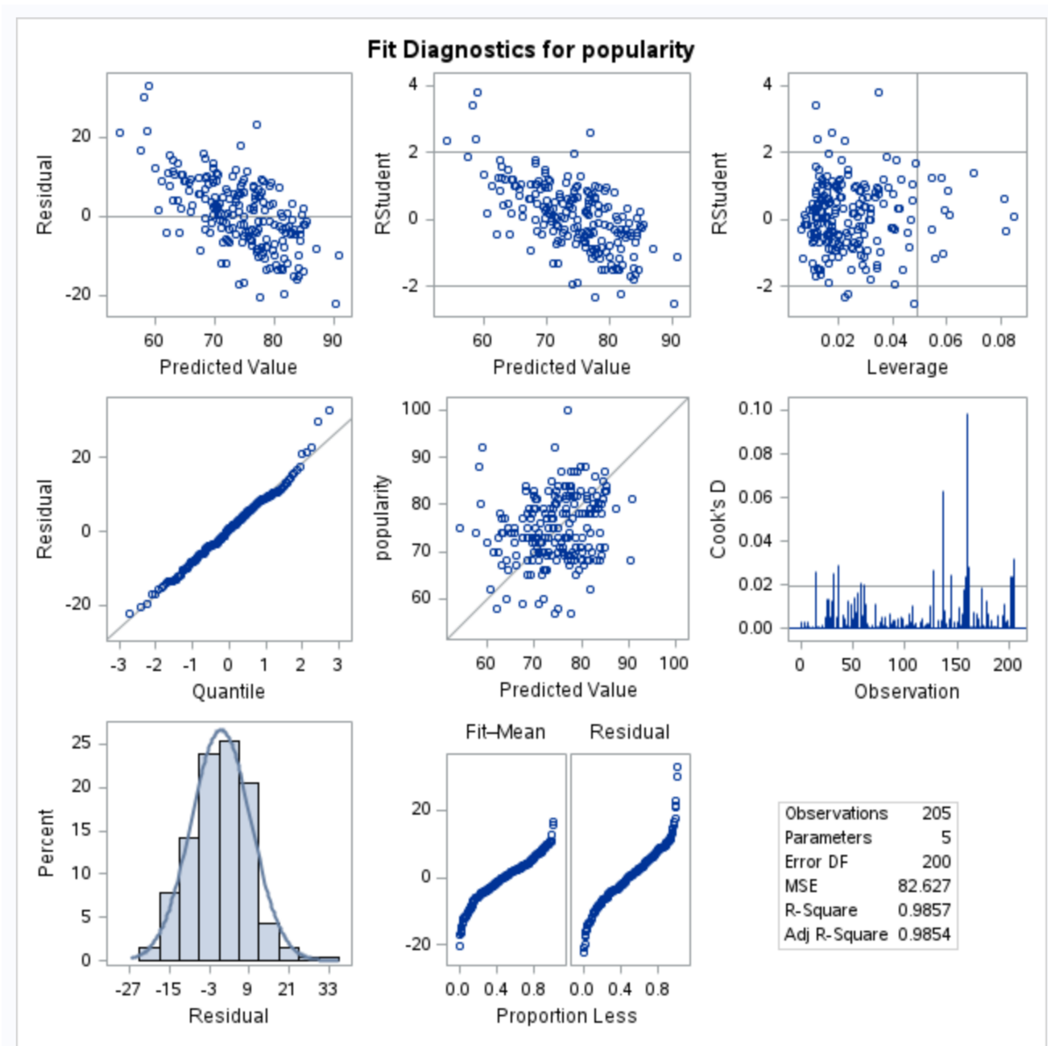
Parameter Estimates								
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation	95% Confidence Limits	
danceability	1	40.98652	5.18618	7.90	<.0001	23.64845	30.75991	51.21312
loudness	1	-2.28497	0.26180	-8.73	<.0001	10.87989	-2.80120	-1.76873
tempo	1	0.07642	0.01916	3.99	<.0001	15.04549	0.03863	0.11420
valence	1	-7.62950	3.86058	-1.98	0.0495	7.13811	-15.24215	-0.01684
energy	1	46.46287	4.50150	10.32	<.0001	18.27194	37.58637	55.33937

After conducting a diagnostic analysis of the utilized model, we conclude observations 31 (*Enchanted (Taylor's Version)*) and 160 (*Lover*) to have the highest impacts in reducing model precision, given the highest R-studentized residual values. However, given neither of these observations have simultaneously a high leverage and R-student values from the provided plots, we choose to keep both points in the model as valid observations. Despite this conclusion, further inspection should be conducted on these two points, given their abnormal deviations and influence from the corresponding observations.

The REG Procedure
Model: MODEL1
Dependent Variable: popularity

Output Statistics

Obs	Dependent Variable	Predicted Value	Std Error Mean Predict	95% CL Mean		95% CL Predict		Residual	Std Error Residual	Student Residual	Cook's D	RStudent	Hat Diag H	Cov Ratio	DFFITS	DFBETAS danceability	DFBETAS loudness	DFBETAS tempo	DFBETAS valence	DFBETAS energy
31	88	58.0844	0.9768	56.1582	60.0106	40.0568	76.1121	29.9156	9.037	3.310	0.026	3.3963	0.0115	0.7827	0.3671	0.2065	0.1569	-0.0426	-0.2165	0.0877
160	92	59.0492	1.6983	55.7004	62.3980	40.8146	77.2838	32.9508	8.930	3.690	0.098	3.8128	0.0349	0.7468	0.7251	-0.5122	-0.6001	-0.4608	0.2536	0.4728



The threshold for leverage is 2(5+1)/205 = 0.0585. The threshold for Cook's D is 4/205 = 0.0195. The threshold for DFBETA is $2/\sqrt{205}$ = 0.1397. The threshold for DFFIT is $2\sqrt{5/205}$ = 0.3123. The threshold for COVRATIO is 3(5)/205 = 0.0732.

Upon looking at the Hat Diag H, Cook's D, DFBETA, DFFIT, and COVRATIO values, while some observations appear to exceed their respective thresholds, none hold enough high leverage over the obtained model; all observed outliers may be treated as valid observations with minimal issue.

Discussion and Concluding Remarks

A limitation on our dataset is that the songs analyzed in our model contain albums that are 'Taylor's Version', as opposed to the original recordings. We chose these because the re-recorded albums contain more songs, thus giving more observations to work with. This may make the prediction of popularity inaccurate, as the older version would have more streams as they've been out longer and the newer versions are much less successful in comparison despite them being commercially successful since the original release. If streaming popularity for both the older version of the songs and Taylor's Version were to be combined, a more accurate depiction of the predictors could be gathered. For interpretations' sake, we chose not to include an intercept in our model. It would not make logical sense to include it, as there is no way to interpret a lack of all variables. Including an intercept would indicate that a song has inherent popularity due to musical factors. This case might be true if one were to also include Swift's overall likeness as a factor but since we are only analyzing auditory variables, the intercept is better off being excluded.

Based on our model, it is found that the success of Taylor Swift's discography is best influenced by the energetic type factors like (valence, danceability, etc.) and less influenced by more slow-paced factors (acousticness, speechiness, etc.). In other words, for Swift to obtain even greater mainstream acclaim, concentrating on more upbeat themes should be a priority with lively beats, big energy, and high tempo. Thus far, Swift has been able to create hits that exhibit these characteristics while also incorporating exceptional songwriting. Her ability to accompany complex lyricism with classic pop tactics enable her to always be on top of the charts. This model could serve as a learning tool for other mainstream artists on Spotify who hope to match Swift's success. It is to be stated that Swift has mastered the game of adoration all too well.

References

"About Spotify." Spotify, Spotify, 24 Oct. 2023, newsroom.spotify.com/company-info/#:~:text=We%20are%20the%20world's%20most,in%20more%20than%20180%20markets. Accessed 27 Nov. 2023.

"Get Track's Audio Features." Web API Reference | Spotify for Developers, Spotify, developer.spotify.com/documentation/web-api/reference/get-audio-features. Accessed 27 Nov. 2023.

Kuzub, Alena. "'Swiftonomics,' or the Smart Business Choices Taylor Swift Makes That Affect the U.S. Economy." *Northeastern Global News*, Northeastern University, 17 Nov. 2023, news.northeastern.edu/2023/08/11/taylor-swift-economy-impact/. Accessed 27 Nov. 2023.

"Loudness Normalization." Spotify, Spotify, support.spotify.com/us/artists/article/loudness-normalization/. Accessed 16 Nov. 2023.

Priester, J. (2023, November 7). Taylor Swift Spotify Dataset (Version 12). Kaggle www.kaggle.com/datasets/jarredpriester/taylor-swift-spotify-dataset

"Taylor Swift." *Spotify*, open.spotify.com/artist/06HL4z0CvFAxyc27GXpf02. Accessed 27 Nov. 2023.

Willman, Chris. "Taylor Swift Beats Her Own Spotify Record for Most Single-Day Streams for an Artist with '1989 (Taylor's Version)' Release." Variety, Variety, 28 Oct. 2023, variety.com/2023/music/news/taylor-swift-spotify-record-most-streams-single-artist-1989-taylors-version-1235771881/. Accessed 27 Nov. 2023.

Bella Almeter, Rani Misra, Marco Teniente
STA 4713: Project

Appendix

```
PROC IMPORT OUT=tss
   DATAFILE="/home/u63069202/Multivariate Analysis/Project/taylor_swift_spotify.csv"
   DBMS=CSV
   REPLACE;
   GETNAMES=YES;
RUN;

data tss;
set tss;
drop var1 release_date track_number id uri duration_ms;
if album in ('Live From Clear Channel Stripped', 'Fearless', 'Fearless (International Vers',
'Fearless Platinum Edition', 'Speak Now (Deluxe Edition)', 'Speak Now', 'Speak Now World Tour
Live', 'Red', 'Red (Deluxe Edition)',  '1989', '1989 (Deluxe Edition)',  'reputation Stadium Tour
Surprise',  'folklore', 'folklore: the long pond studio s',  'evermore', 'Midnights', 'Midnights (3am
Edition)', "1989 (Taylor's Version)") then delete; run;

proc print data=tss;
proc reg data=tss;
        MODEL popularity = acousticness danceability energy instrumentalness liveness
  loudness speechiness tempo valence /
  selection = backward cli clm clb vif influence P R noint;
OUTPUT OUT=NEW P=PRED R=RES;
RUN;
```