Kyle Conrad, Satoshi Ido
STAT 52500 - Fall 2022
12/01/2022

Final Project Report

Introduction:

We will be analyzing the “Spotify Top 200 Dataset”, found on Kaggle at https://www.kaggle.com/datasets/younver/spotify-top-200-dataset?select=spotify-top-200-dataset.csv. This dataset consists of Spotify’s global weekly top 200 songs between 2017 and 2021 compiled into a single dataset. It consists of 74,661 observations of 40 variables extracted from Spotify API Reference. These variables are:

- **Track id**: spotify id for the track
- **Track name**: name of the track
- **Track popularity** (Quantitative): popularity of the track calculated by spotify
- **Track number**: track’s index relative to its album
- **Album id**: spotify id for the album that the track is from
- **Album name**: name of the album that the track is from
- **Album img**: link to the cover image of album that the track is from
- **Album type**: type of the album (eg. single, album)
- **Album label**: track’s record label
- **Album track number**: number of the tracks in the album that the track is from
- **Album popularity** (Quantitative): popularity of the album calculated by spotify
- **Artist num**: number of artists in the track
- **Artist names**: names of all artists who participated in the track (separated by comma)
- **Artist id**: spotify artist id for the artist individual
- **Artist name**: one of the artists who participated in the track (tracks with multiple artists are split into separate rows for each artist)
- **Artist img**: link to the artist individual's image
- **Artist followers**: follower amount of artist
- **Artist popularity** (Quantitative): popularity of the artist calculated by spotify
- **Artist genres**: artist’s genres
- **Rank** (Quantitative): ranking of the track on the chart
- **Week**: end of week the track was in charts as date format
- **Streams** (Quantitative): number of streams in that week
- **Collab** (Quantitative): if the participation of the track is multiple or not (False if there is only one artist, else True)
- **Explicit** (Quantitative): explicit situation of the track (True if explicit, False otherwise)
- **Release date**: release date of the album (thus track)
- **Danceability** (Quantitative):
- **Energy** (Quantitative): Energy is a measure from 0.0 to 1.0 and represents a perceptual measure of intensity and activity.

- **Key** (Qualitative): The key the track is in. Integers map to pitches using standard Pitch Class notation. E.g. 0 = C, 1 = C♯/D♭, 2 = D, and so on. If no key was detected, the value is -1.
- **Mode** (Qualitative): Mode indicates the modality (major or minor) of a track. Major is represented by 1 and minor is 0.
- **time signature** (Qualitative): An estimated time signature. The time signature ranges from 3 to 7 indicating time signatures of "3/4", to "7/4".
- **Loudness** (*Quantitative)*: The overall loudness of a track in decibels (dB). Loudness values are averaged across the entire track.
- **Speechiness** (*Quantitative)*: Speechiness detects the presence of spoken words in a track. The more exclusively speech-like the recording, 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.
- **Acousticness** *(Quantitative)*: A confidence measure from 0.0 to 1.0 of whether the track is acoustic. 1.0 represents high confidence the track is acoustic.
- **Instrumentalness** (*Quantitative)*: 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** (*Quantitative)*: 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.
- **Valence** (*Quantitative)*: A measure from 0.0 to 1.0 describing the musical positiveness conveyed by a track. Tracks with high valence sound more positive, while tracks with low valence sound more negative.
- **Tempo** (*Quantitative)*: The overall estimated tempo of a track in beats per minute (BPM).
- **Duration** (*Quantitative)*: The duration of the track in milliseconds.
- **Pivot** (Qualitative): When multiple artists are split into separate rows, this value takes 0 for the first artist and 1 for the rest.
- **Track_index**(Quantitative): track index for create_dataset script

Our analysis is to be split into two, largely independent, sections: Model Selection and Factor Analysis. Within the Model Selection section we will be determining the best possible model, out of a subset of the available variables, at explaining the variation within the amount of time between the release date of a song and the week it appears on the weekly Spotify Global Top 200 List. We will then be interpreting and conducting a final analysis on the chosen model. After we do model selection, we then incorporate some categorical variables: collab, explicit, mode, key, into the given model. We check their difference in means, if they meet the normality and assumptions, and then, check the interaction to reach some conclusion.

Model Selection:

We are interested in building the best possible model containing a subset of the following variables at explaining the variability of the age of a song on the weekly global top 200 list (days_since_release). The variables of interest are: rank of song during a given week (*srank*), *danceability*, *energy*, *loudness*, *speechiness*, *acousticness*, *instrumentalness*, *liveness*, *valence*, *tempo*, and *duration*. Additionally, we will need to ensure that it is not necessary to incorporate info on the higher order terms of these variables.

We start by checking the Normality Assumption of our dependent variable, *days _since_release*, by generating both a Histogram and a QQ-Plot of the variable.

Figure 1: Histogram of days_since_release

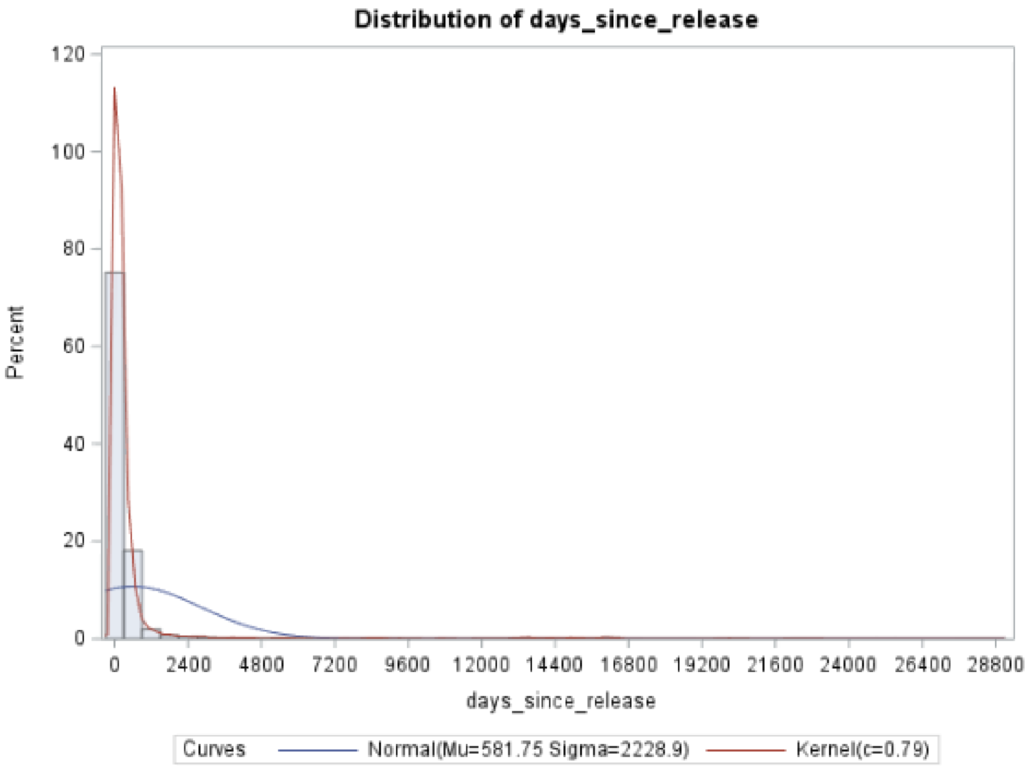
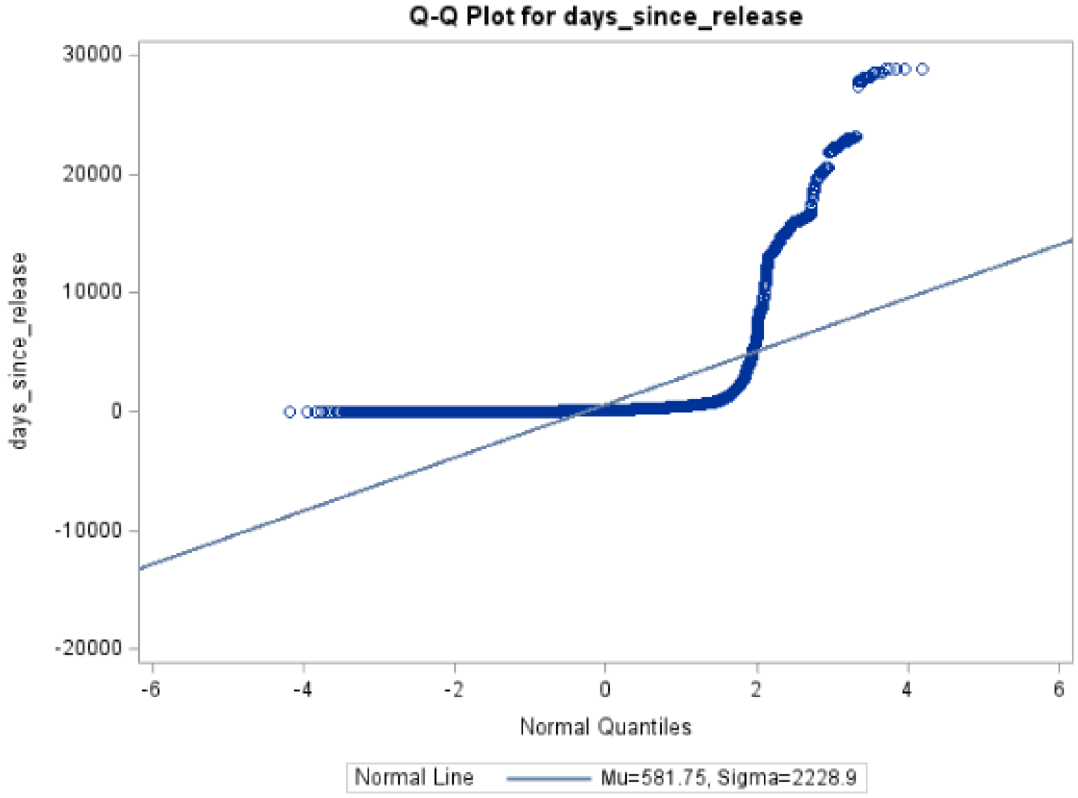
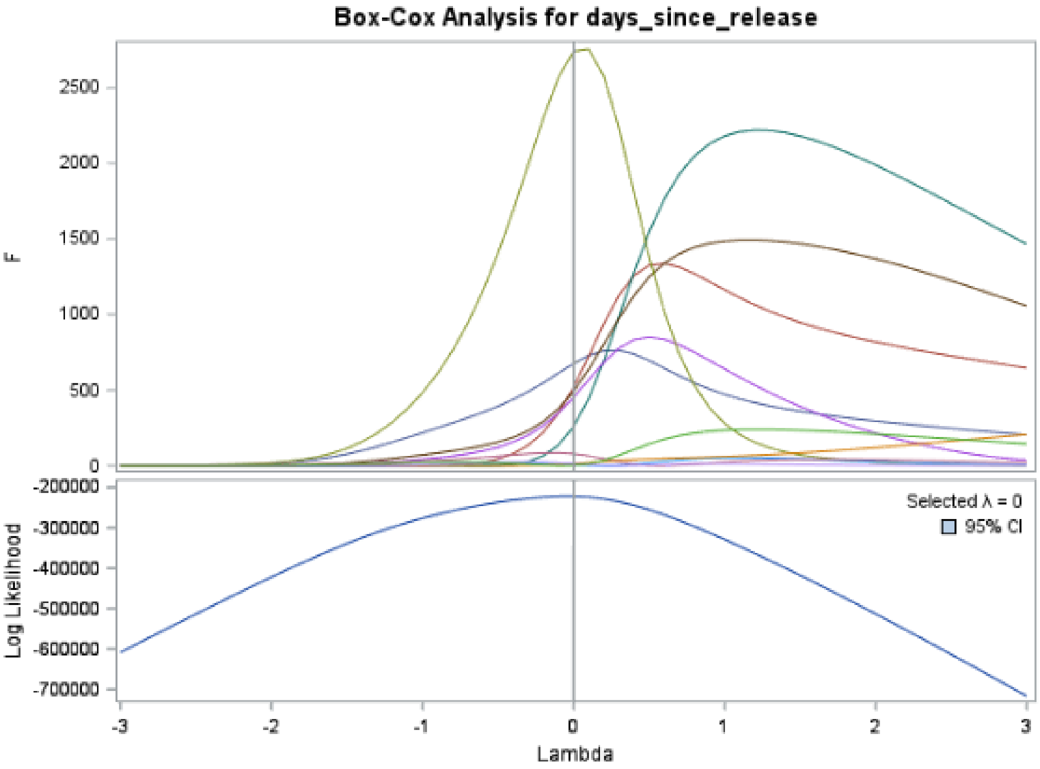


Figure 2: QQ-Plot of days_since_release



It is obvious from the above plots that there are major deviations from normality within our dependent variable. The histogram shows that the data has a strong right-skewness, while the QQ-Plot shows clearly the data does not follow an approximately normal distribution. Thus, we will use the Box-Cox Transformation to normalize our dependent variable before we continue our analysis.

Figure 3: Box-Cox Analysis of days_since_release



As seen in the above Box-Cox Analysis Plot the optimal value of λ for our transformation was found to be $\lambda = 0$. We will thus apply the Box-Cox Transformation using the transformation, $Y = Log(Y)$, to *days_since_release*. We will generate a histogram and QQ-plot of the transformed variable to confirm that the violation of normality was resolved.

Figure 4: Histogram of Box-Cox Transformed (λ = 0) days_since_release

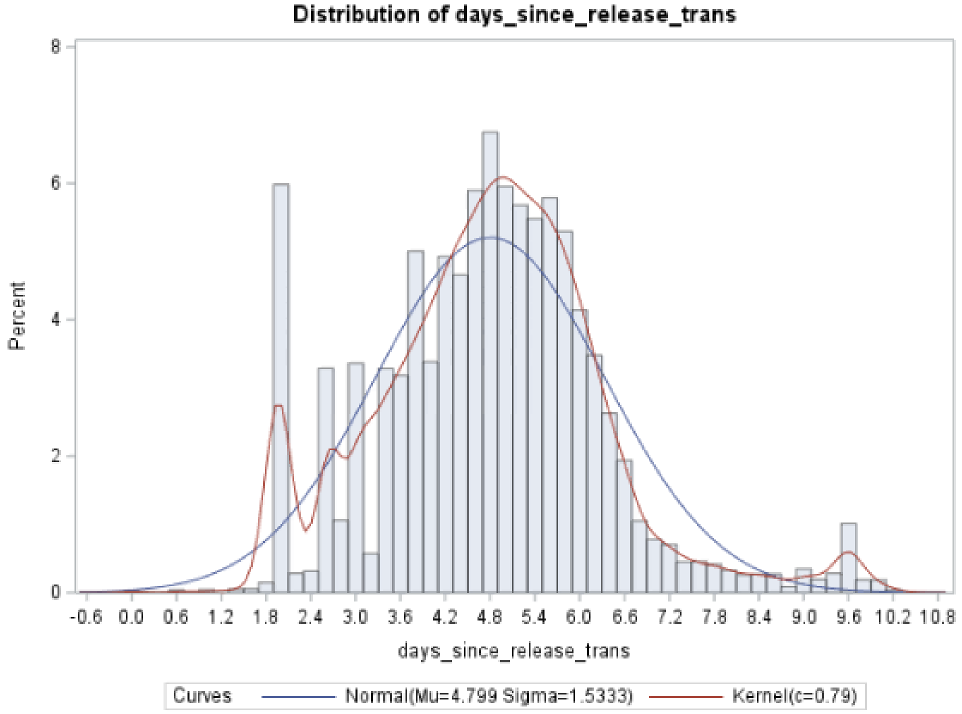
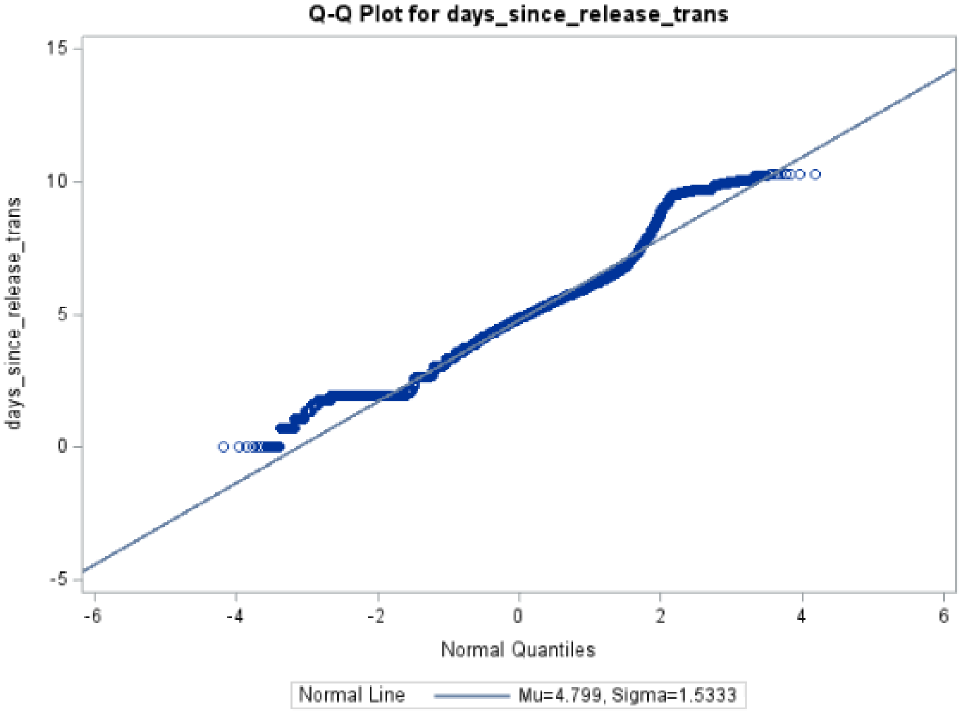


Figure 5: QQ-Plot of Box-Cox Transformed (λ = 0) days_since_release



We can see that there are still some slight deviations from normality near the tails of the data, however these are negligible. Thus, it appears that the Box-Cox Transformation was sufficient for resolving the violation of the assumption of normality.

We now generate the Scatterplot Matrix and table of Pairwise Pearson Correlations for the variables of interest to check for possible multicollinearity and evidence for including higher-order terms of the data.

Figure 6: Scatterplot Matrix of Box-Cox Transformed (λ = 0) Data

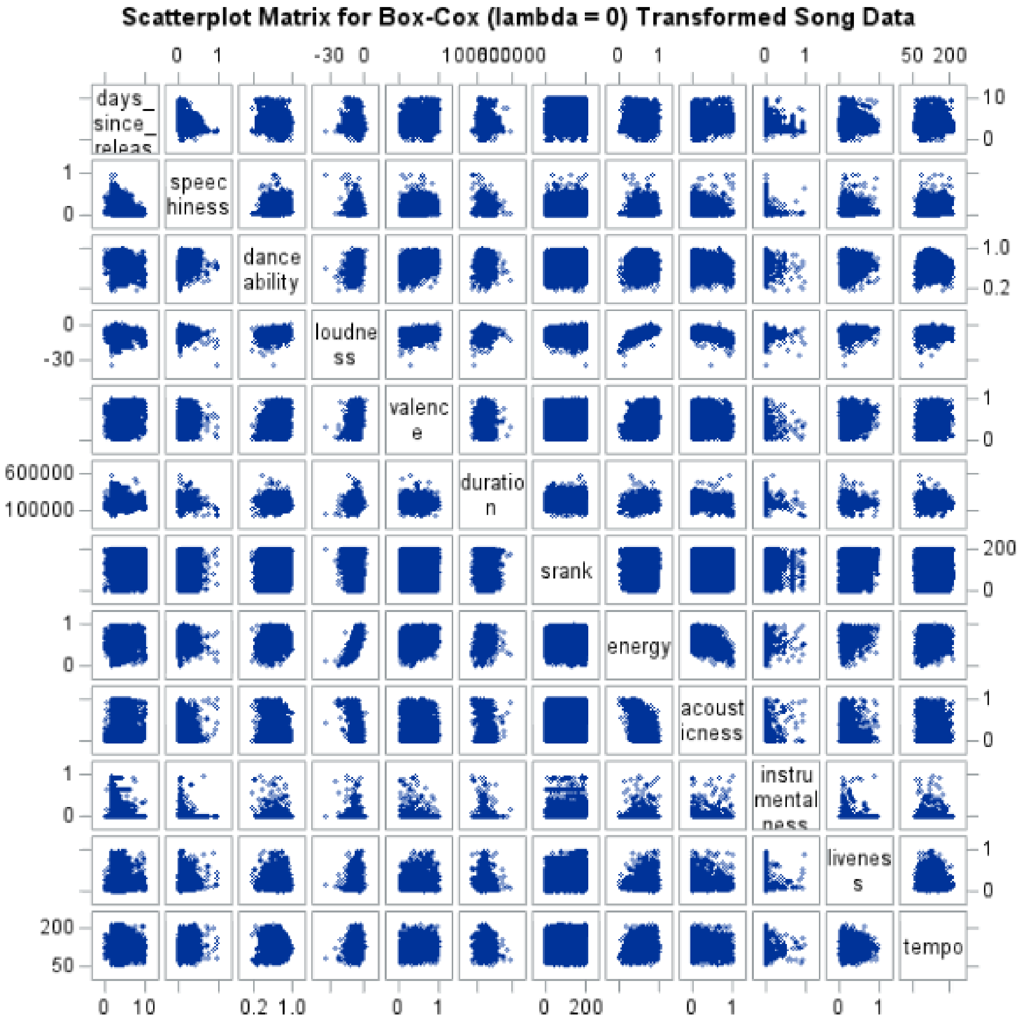
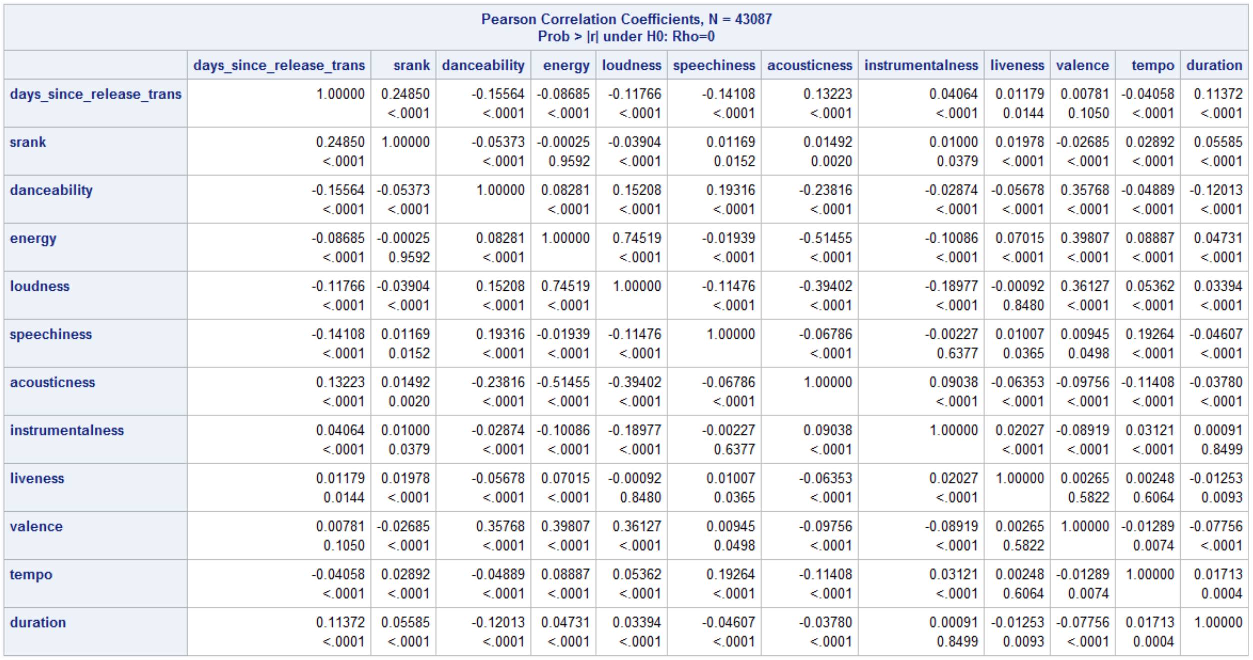


Figure 7: Pairwise Pearson Correlation Table of Data



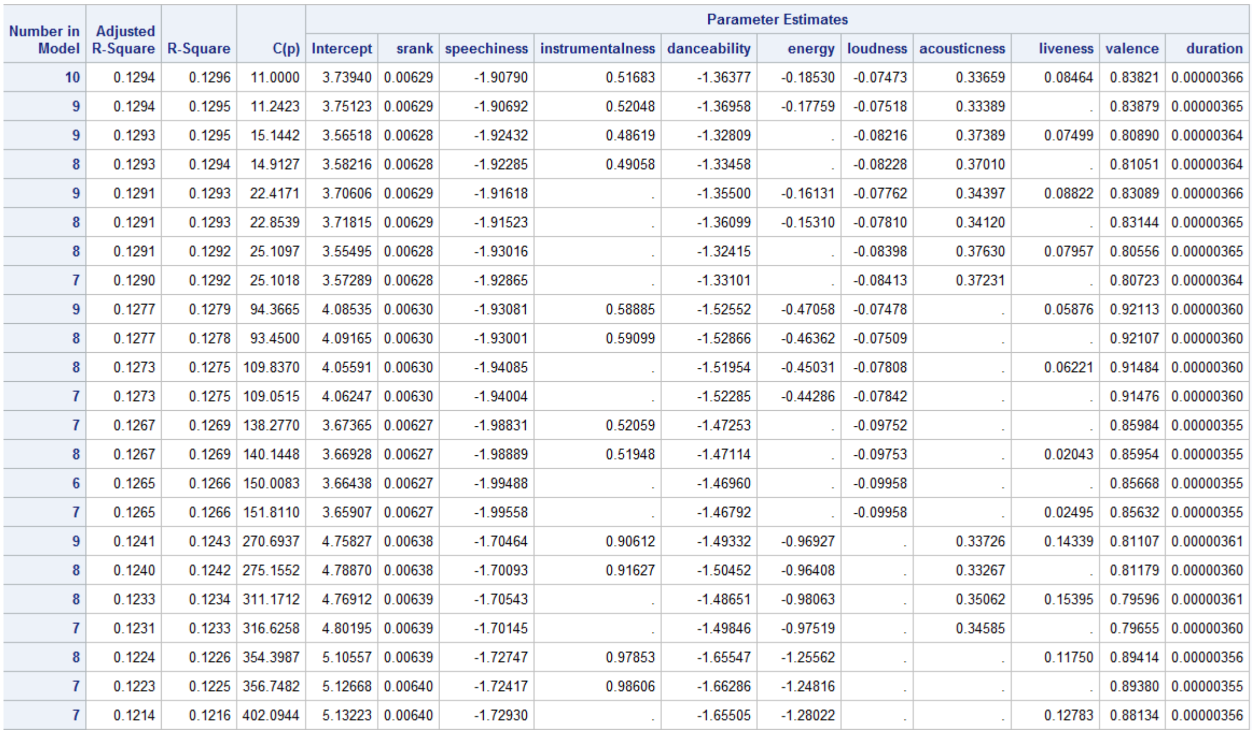
Pearson Correlation Coefficients, N = 43087
Prob > |r| under H0: Rho=0

	days_since_release_trans	srank	danceability	energy	loudness	speechiness	acousticness	instrumentalness	liveness	valence	tempo	duration
days_since_release_trans	1.00000	0.24850 <.0001	-0.15564 <.0001	-0.08685 <.0001	-0.11766 <.0001	-0.14108 <.0001	0.13223 <.0001	0.04064 <.0001	0.01179 0.0144	0.00781 0.1050	-0.04058 <.0001	0.11372 <.0001
srank	0.24850 <.0001	1.00000	-0.05373 <.0001	-0.00025 0.9592	-0.03904 <.0001	0.01169 0.0152	0.01492 0.0020	0.01000 0.0379	0.01978 <.0001	-0.02685 <.0001	0.02892 <.0001	0.05585 <.0001
danceability	-0.15564 <.0001	-0.05373 <.0001	1.00000	0.08281 <.0001	0.15208 <.0001	0.19316 <.0001	-0.23816 <.0001	-0.02874 <.0001	-0.05678 <.0001	0.35768 <.0001	-0.04889 <.0001	-0.12013 <.0001
energy	-0.08685 <.0001	-0.00025 0.9592	0.08281 <.0001	1.00000	0.74519 <.0001	-0.01939 <.0001	-0.51455 <.0001	-0.10086 <.0001	0.07015 <.0001	0.39807 <.0001	0.08887 <.0001	0.04731 <.0001
loudness	-0.11766 <.0001	-0.03904 <.0001	0.15208 <.0001	0.74519 <.0001	1.00000	-0.11476 <.0001	-0.39402 <.0001	-0.18977 <.0001	-0.00092 0.8480	0.36127 <.0001	0.05362 <.0001	0.03394 <.0001
speechiness	-0.14108 <.0001	0.01169 0.0152	0.19316 <.0001	-0.01939 <.0001	-0.11476 <.0001	1.00000	-0.06786 <.0001	-0.00227 0.6377	0.01007 0.0365	0.00945 0.0498	0.19264 <.0001	-0.04607 <.0001
acousticness	0.13223 <.0001	0.01492 0.0020	-0.23816 <.0001	-0.51455 <.0001	-0.39402 <.0001	-0.06786 <.0001	1.00000	0.09038 <.0001	-0.06353 <.0001	-0.09756 <.0001	-0.11408 <.0001	-0.03780 <.0001
instrumentalness	0.04064 <.0001	0.01000 0.0379	-0.02874 <.0001	-0.10086 <.0001	-0.18977 <.0001	-0.00227 0.6377	0.09038 <.0001	1.00000	0.02027 <.0001	-0.08919 <.0001	0.03121 <.0001	0.00091 0.8499
liveness	0.01179 0.0144	0.01978 <.0001	-0.05678 <.0001	0.07015 <.0001	-0.00092 0.8480	0.01007 0.0365	-0.06353 <.0001	0.02027 <.0001	1.00000	0.00265 0.5822	0.00248 0.6064	-0.01253 0.0093
valence	0.00781 0.1050	-0.02685 <.0001	0.35768 <.0001	0.39807 <.0001	0.36127 <.0001	0.00945 0.0498	-0.09756 <.0001	-0.08919 <.0001	0.00265 0.5822	1.00000	-0.01289 0.0074	-0.07756 <.0001
tempo	-0.04058 <.0001	0.02892 <.0001	-0.04889 <.0001	0.08887 <.0001	0.05362 <.0001	0.19264 <.0001	-0.11408 <.0001	0.03121 <.0001	0.00248 0.6064	-0.01289 0.0074	1.00000	0.01713 0.0004
duration	0.11372 <.0001	0.05585 <.0001	-0.12013 <.0001	0.04731 <.0001	0.03394 <.0001	-0.04607 <.0001	-0.03780 <.0001	0.00091 0.8499	-0.01253 0.0093	-0.07756 <.0001	0.01713 0.0004	1.00000

From the Scatterplot Matrix we can note that there does not appear to be evidence for including higher-order terms of the predictor variables. Additionally, from both the Pearson Correlation table and matrix there appears to be a significant amount of between the predictor variables. It will thus be necessary to ensure that multicollinearity is not a serious concern later in our model. Finally, we can see from the Pearson Correlation table that all the variables appear to be correlated with *days_since_release* at a 15% significance level.

We will use All Subset Selection with Mallow's Cp Criterion and Adjusted R-Squared to determine a subset of the models which are “best” according to these criteria. We’ll justify our final choice of model by calculating the PRESS criterion for each of the “best” models.

Figure 8: All Subset Selection with R^2_{adj} and Mallow's Cp Criteria



Number in Model	Adjusted R-Square	R-Square	C(p)	Parameter Estimates										
				Intercept	srank	speechiness	instrumentalness	danceability	energy	loudness	acousticness	liveness	valence	duration
10	0.1294	0.1296	11.0000	3.73940	0.00629	-1.90790	0.51683	-1.36377	-0.18530	-0.07473	0.33659	0.08464	0.83821	0.00000366
9	0.1294	0.1295	11.2423	3.75123	0.00629	-1.90692	0.52048	-1.36958	-0.17759	-0.07518	0.33389	.	0.83879	0.00000365
9	0.1293	0.1295	15.1442	3.56518	0.00628	-1.92432	0.48619	-1.32809	.	-0.08216	0.37389	0.07499	0.80890	0.00000364
8	0.1293	0.1294	14.9127	3.58216	0.00628	-1.92285	0.49058	-1.33458	.	-0.08228	0.37010	.	0.81051	0.00000364
9	0.1291	0.1293	22.4171	3.70606	0.00629	-1.91618	.	-1.35500	-0.16131	-0.07762	0.34397	0.08822	0.83089	0.00000366
8	0.1291	0.1293	22.8539	3.71815	0.00629	-1.91523	.	-1.36099	-0.15310	-0.07810	0.34120	.	0.83144	0.00000365
8	0.1291	0.1292	25.1097	3.55495	0.00628	-1.93016	.	-1.32415	.	-0.08398	0.37630	0.07957	0.80556	0.00000365
7	0.1290	0.1292	25.1018	3.57289	0.00628	-1.92865	.	-1.33101	.	-0.08413	0.37231	.	0.80723	0.00000364
9	0.1277	0.1279	94.3665	4.08535	0.00630	-1.93081	0.58885	-1.52552	-0.47058	-0.07478	.	0.05876	0.92113	0.00000360
8	0.1277	0.1278	93.4500	4.09165	0.00630	-1.93001	0.59099	-1.52866	-0.46362	-0.07509	.	.	0.92107	0.00000360
8	0.1273	0.1275	109.8370	4.05591	0.00630	-1.94085	.	-1.51954	-0.45031	-0.07808	.	0.06221	0.91484	0.00000360
7	0.1273	0.1275	109.0515	4.06247	0.00630	-1.94004	.	-1.52285	-0.44286	-0.07842	.	.	0.91476	0.00000360
7	0.1267	0.1269	138.2770	3.67365	0.00627	-1.98831	0.52059	-1.47253	.	-0.09752	.	.	0.85984	0.00000355
8	0.1267	0.1269	140.1448	3.66928	0.00627	-1.98889	0.51948	-1.47114	.	-0.09753	.	0.02043	0.85954	0.00000355
6	0.1265	0.1266	150.0083	3.66438	0.00627	-1.99488	.	-1.46960	.	-0.09958	.	.	0.85668	0.00000355
7	0.1265	0.1266	151.8110	3.65907	0.00627	-1.99558	.	-1.46792	.	-0.09958	.	0.02495	0.85632	0.00000355
9	0.1241	0.1243	270.6937	4.75827	0.00638	-1.70464	0.90612	-1.49332	-0.96927	.	0.33726	0.14339	0.81107	0.00000361
8	0.1240	0.1242	275.1552	4.78870	0.00638	-1.70093	0.91627	-1.50452	-0.96408	.	0.33267	.	0.81179	0.00000360
8	0.1233	0.1234	311.1712	4.76912	0.00639	-1.70543	.	-1.48651	-0.98063	.	0.35062	0.15395	0.79596	0.00000361
7	0.1231	0.1233	316.6258	4.80195	0.00639	-1.70145	.	-1.49846	-0.97519	.	0.34585	.	0.79655	0.00000360
8	0.1224	0.1226	354.3987	5.10557	0.00639	-1.72747	0.97853	-1.65547	-1.25562	.	.	0.11750	0.89414	0.00000356
7	0.1223	0.1225	356.7482	5.12668	0.00640	-1.72417	0.98606	-1.66286	-1.24816	.	.	.	0.89380	0.00000355
7	0.1214	0.1216	402.0944	5.13223	0.00640	-1.72930	.	-1.65505	-1.28022	.	.	0.12783	0.88134	0.00000356

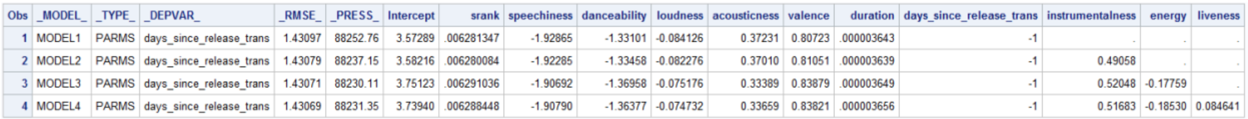
NOTE: For conciseness, only showing the top 23 models generated via All Subset Selection.

We can see that there appear to be a few models which provide a good fit for our data. The chosen "best" models are as follows:

Model 1: All predictors except for instrumentalness, energy, and liveness
Model 2: All predictors except for energy and liveness
Model 3: All predictors except for liveness
Model 4: All predictors

The above models were chosen because they were all similar in R^2_{Adj} value and were individually the "best" model among all models of their respective sizes according to their R^2_{Adj} and Mallow's Cp criterion. It appears from both Mallow's Cp and R^2_{Adj} that the best model out of those identified as possible "best" models will be Model 3. However, to be certain of our model choice we will calculate the PRESS criterion of each model and use that information to help guide our choice of best model. Note that since Mallow's Cp is guaranteed to be equal to the number of parameters when all predictors are included in the model, we disregard it in the case of Model 4.

Figure 9: PRESS Criterion for Selected Models



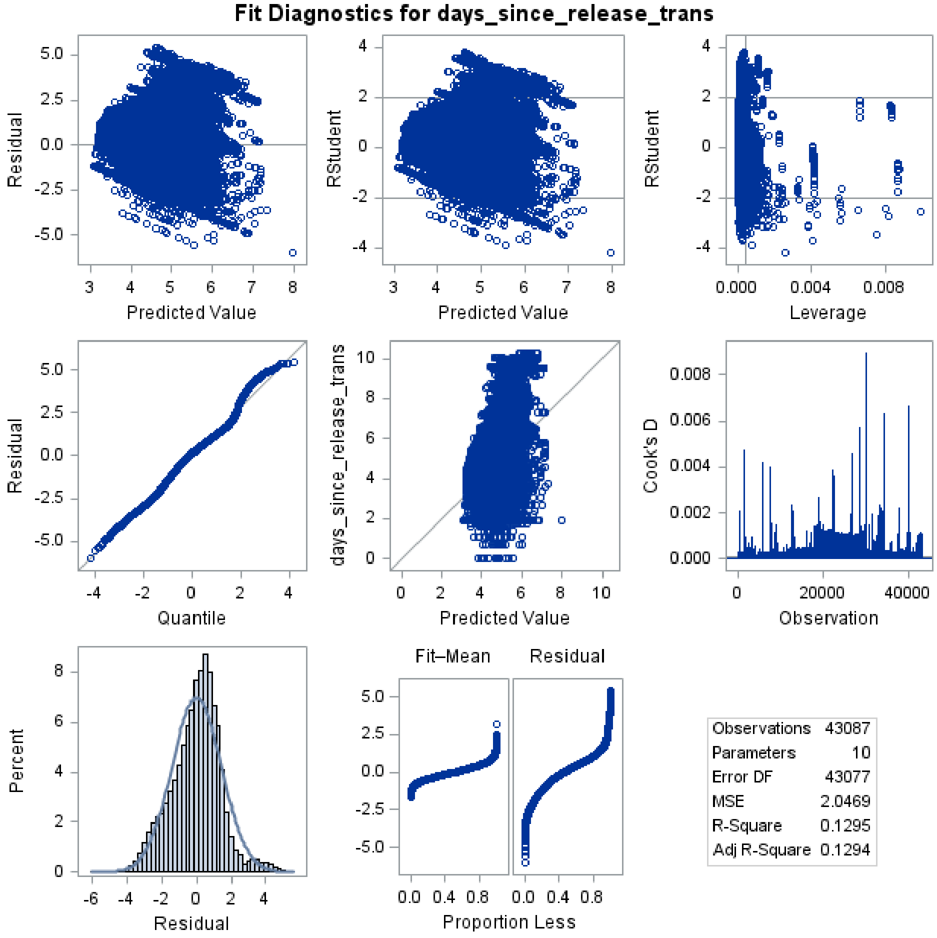
Obs	_MODEL_	_TYPE_	_DEPVAR_	_RMSE_	_PRESS_	Intercept	srank	speechiness	danceability	loudness	acousticness	valence	duration	days_since_release_trans	instrumentalness	energy	liveness
1	MODEL1	PARMS	days_since_release_trans	1.43097	88252.76	3.57289	.006281347	-1.92865	-1.33101	-0.084126	0.37231	0.80723	.000003643	-1	.	.	.
2	MODEL2	PARMS	days_since_release_trans	1.43079	88237.15	3.58216	.006280084	-1.92285	-1.33458	-0.082276	0.37010	0.81051	.000003639	-1	0.49058	.	.
3	MODEL3	PARMS	days_since_release_trans	1.43071	88230.11	3.75123	.006291036	-1.90692	-1.36958	-0.075176	0.33389	0.83879	.000003649	-1	0.52048	-0.17759	.
4	MODEL4	PARMS	days_since_release_trans	1.43069	88231.35	3.73940	.006288448	-1.90790	-1.36377	-0.074732	0.33659	0.83821	.000003656	-1	0.51683	-0.18530	0.084641

It was found that the PRESS criterion (see above) of the model containing all predictors except for liveness (Model 3) was minimal, thus Model 3 is our best model according to the PRESS criterion. This conclusion is in agreement with our conclusion based purely off of Mallow's Cp and $R^2{}_{Adj}$, thus we will be moving forward using Model 3 as our chosen "best" model.

The following can now be noted about Model 3 in comparison with the other tested models: Its PRESS criterion was minimal, so it was the model whose fitted values best predicted the observed responses. Its Mallow's Cp criterion was minimal and close to the number of parameters in the full model, so it is an approximately unbiased model with minimal bias of all possible models. Additionally, its $R^2{}_{Adj} = 0.1294$ so approximately 12.94 % (adjusted for number of parameters) of the variation in the number of days between a song's release and its appearance on the Spotify weekly top 200 list is explained by the variables in our model.

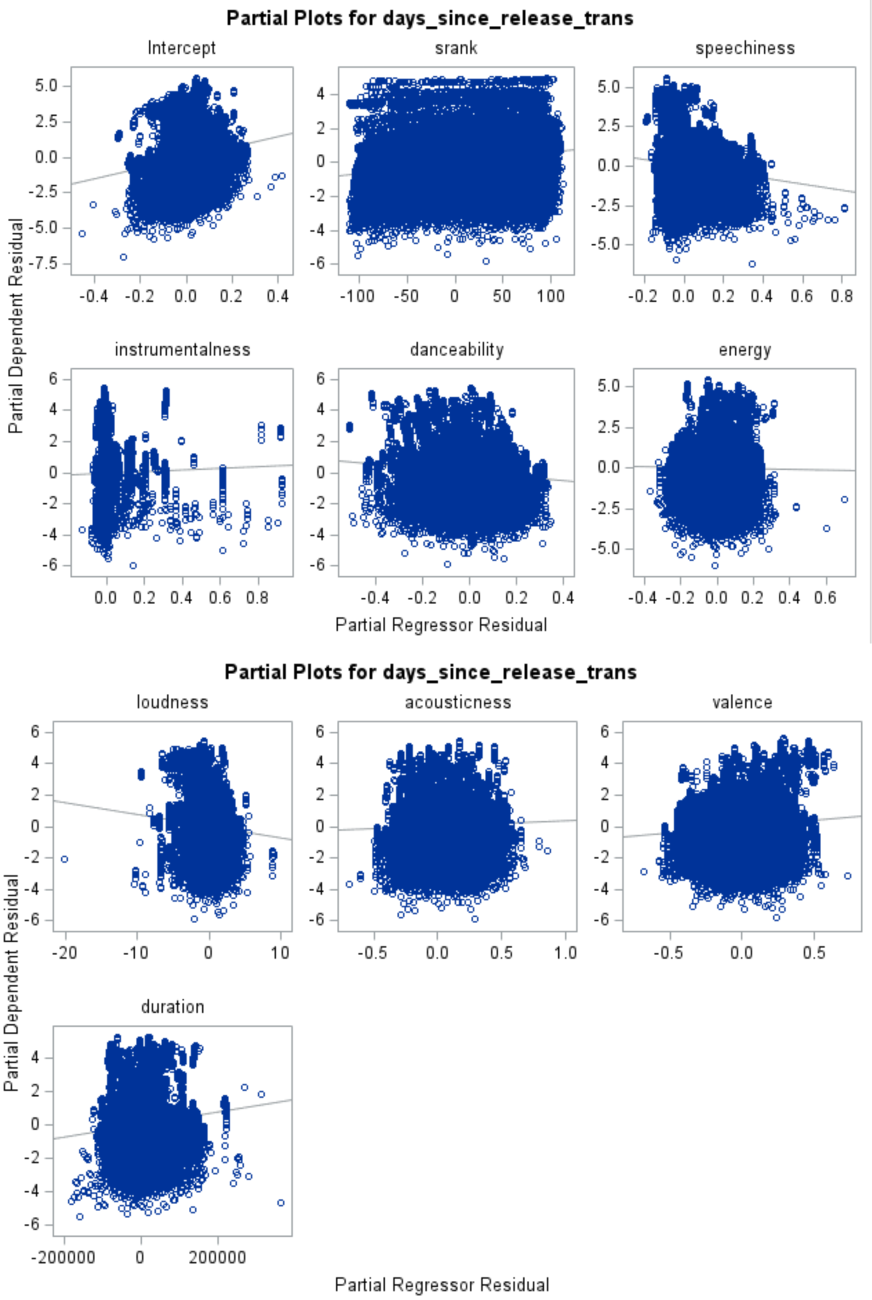
We now check the assumptions of Independence, Normality, and Constant Variance of the error terms for our chosen "best" model. We will be using Partial Regression Plots for each variable and both a Histogram and QQ-plot of the residuals to perform these diagnostics. Finally we will be looking at the plot of the residuals vs the predicted value to check for any clear patterns in our residuals.

Figure 10: Fit Diagnostics for Selected Model of Best Fit



Observations	43087
Parameters	10
Error DF	43077
MSE	2.0469
R-Square	0.1295
Adj R-Square	0.1294

Figure 11: Partial Residual Plots for Predictors



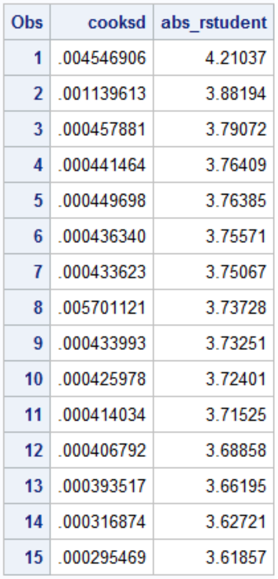
Looking at the Histogram and QQ-plot of the residuals of our model it appears that the data is approximately normally distributed. However, we can see from the plot of the residuals vs the predicted values that there is a clear linear trend in our residuals. This is due to our response variable taking only integer values (before transformation), thus this is a violation of normality. We do not have the tools necessary to overcome this violation in our current repertoire so we will simply note that the assumption the error terms are Normally distributed is not met. For the sake of analysis / practice we will, however, continue on under the assumption that it is met.

We can see from the Partial Residual plots that there is no clear pattern in the residuals in relation to any of the predictor variables. Therefore we can say that the assumption of Independence of the Error Terms is met for our chosen model. Finally, the plot of the residuals vs the fitted values shows that the variance of the residuals appears to be constant. Thus the assumption of Constant Variance is met for our model.

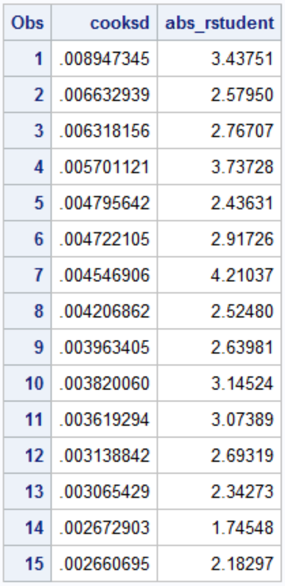
We can see that all assumptions of the error terms (Normality, Independence, and Constant Variance) are met for our model (Note: normality is assumed). Now the final diagnostics we must perform are checks for outliers, influential points, and multicollinearity between the predictors.

To check for potentially influential and possibly outlying cases, we calculated the Studentized Deleted Residuals and Cook's Distance for each case. We then sorted the cases in two ways: descending order of the absolute value of the Studentized Deleted Residual and descending order of Cook's Distance.

Figures 12 (left) and 13 (right): Top 15 cases with largest magnitude Studentized Deleted Residual (left), and with largest Cook's Distance (right).



Obs	cooksd	abs_rstudent
1	.004546906	4.21037
2	.001139613	3.88194
3	.000457881	3.79072
4	.000441464	3.76409
5	.000449698	3.76385
6	.000436340	3.75571
7	.000433623	3.75067
8	.005701121	3.73728
9	.000433993	3.73251
10	.000425978	3.72401
11	.000414034	3.71525
12	.000406792	3.68858
13	.000393517	3.66195
14	.000316874	3.62721
15	.000295469	3.61857

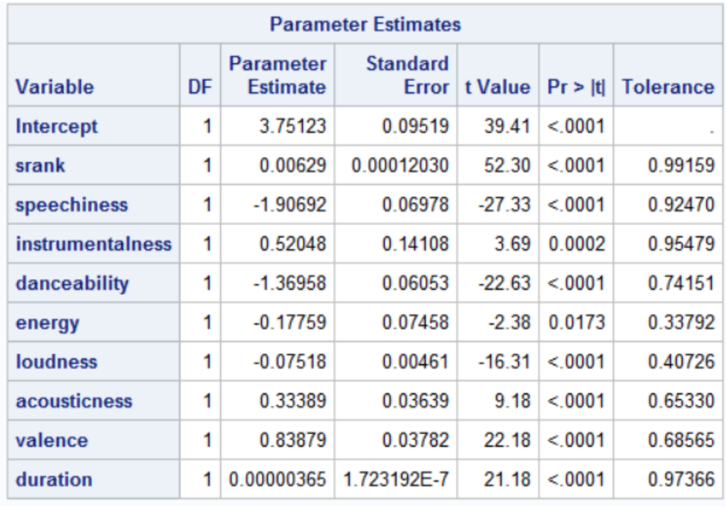


Obs	cooksd	abs_rstudent
1	.008947345	3.43751
2	.006632939	2.57950
3	.006318156	2.76707
4	.005701121	3.73728
5	.004795642	2.43631
6	.004722105	2.91726
7	.004546906	4.21037
8	.004206862	2.52480
9	.003963405	2.63981
10	.003820060	3.14524
11	.003619294	3.07389
12	.003138842	2.69319
13	.003065429	2.34273
14	.002672903	1.74548
15	.002660695	2.18297

It can be seen from the above sorted tables that there are no cases with Cook's Distance > .9341963 $(= F(.5, 10, 43077))$ and there are no cases with |Studentized Deleted Residual| > 4.86297 $(= t(1 - \frac{.05}{2*43087}, 43076))$. Thus, we conclude that there are no potential outlying points nor potentially influential cases.

We will now check for multicollinearity between our predictor variables via calculating the Variance Inflation Factor values for our model.

Figure 14: Table of Calculated VIF Values for Selected Model



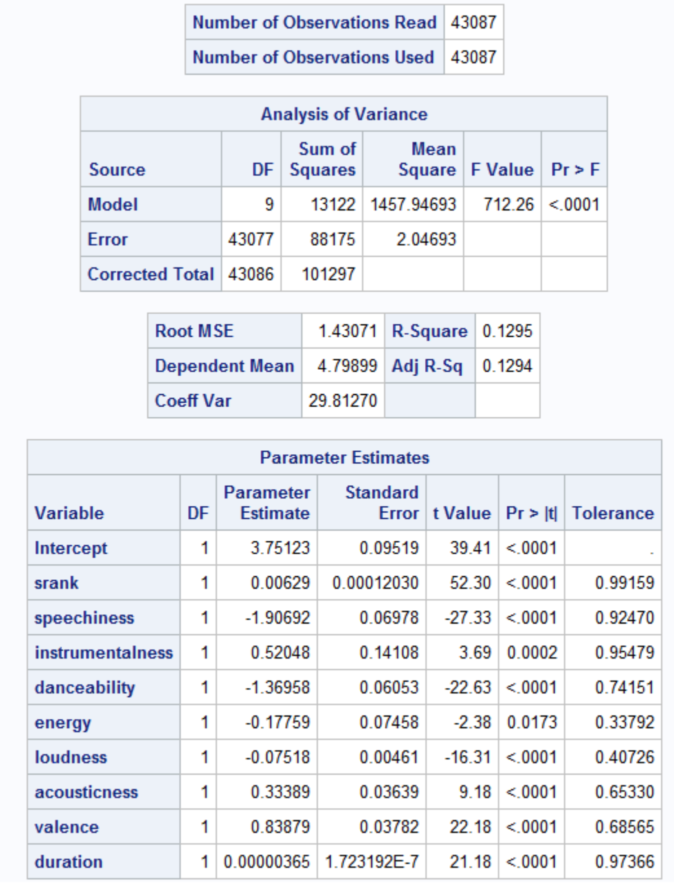
Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Tolerance
Intercept	1	3.75123	0.09519	39.41	<.0001	.
srank	1	0.00629	0.00012030	52.30	<.0001	0.99159
speechiness	1	-1.90692	0.06978	-27.33	<.0001	0.92470
instrumentalness	1	0.52048	0.14108	3.69	0.0002	0.95479
danceability	1	-1.36958	0.06053	-22.63	<.0001	0.74151
energy	1	-0.17759	0.07458	-2.38	0.0173	0.33792
loudness	1	-0.07518	0.00461	-16.31	<.0001	0.40726
acousticness	1	0.33389	0.03639	9.18	<.0001	0.65330
valence	1	0.83879	0.03782	22.18	<.0001	0.68565
duration	1	0.00000365	1.723192E-7	21.18	<.0001	0.97366

From the table above we are able to calculate the VIF trivially by noting that $Tolerance = \frac{1}{VIF}$. Using this conversion we see that the VIF for each predictor is less than 10 and the mean VIF, $\underline{VIF}$, for our predictors is 1.54631. Since no single predictor has a VIF greater than 10 and $\underline{VIF}$ is relatively close to 1, we can state that there is no indication of serious multicollinearity issues within our dataset.

Since there are no serious multicollinearity issues between our predictors and there are no influential cases, no additional remedial procedures need to be performed. Thus, we have come to the final model which is the best model according to Mallow’s Cp, PRESS criterion, and Adjusted R-Squared. Finally we will run the linear regression procedure in SAS to retrieve the final model’s parameter estimates and statistics.

Figure 15: Least Squares Regression output for Model 3



Number of Observations Read	43087
Number of Observations Used	43087

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	9	13122	1457.94693	712.26	<.0001
Error	43077	88175	2.04693		
Corrected Total	43086	101297			

Root MSE	1.43071	R-Square	0.1295
Dependent Mean	4.79899	Adj R-Sq	0.1294
Coeff Var	29.81270		

Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Tolerance
Intercept	1	3.75123	0.09519	39.41	<.0001	.
srank	1	0.00629	0.00012030	52.30	<.0001	0.99159
speechiness	1	-1.90692	0.06978	-27.33	<.0001	0.92470
instrumentalness	1	0.52048	0.14108	3.69	0.0002	0.95479
danceability	1	-1.36958	0.06053	-22.63	<.0001	0.74151
energy	1	-0.17759	0.07458	-2.38	0.0173	0.33792
loudness	1	-0.07518	0.00461	-16.31	<.0001	0.40726
acousticness	1	0.33389	0.03639	9.18	<.0001	0.65330
valence	1	0.83879	0.03782	22.18	<.0001	0.68565
duration	1	0.00000365	1.723192E-7	21.18	<.0001	0.97366

We can see from the above regression output, that the estimated least squares regression line for our chosen model is:

$$Log(\hat{y}) = 3.75123 + .00629X_1 - 1.90692X_2 + .52048X_3 - 1.36958X_4 - .17759X_5 - .07518X_6 + \ldots$$
$$\ldots + .33389X_7 + .83879X_8 + .00000365X_9,$$

Where:

$y = days\ since\ release,\ X_1 = srank,\ X_2 = speechiness,\ X_3 = instrumentality, X_4 = danceability,\ X_5 = energy,\ X_6 = loudness,\ X_7 = acousticness,\ X_8 = valence,\ X_9 = duration$

Due to our response variable being logarithmically transformed, we can interpret the meaning of each predictor's regression coefficient as follows: a one unit increase in the predictor variable, when all other predictors are held constant, will result in a multiplicative increase in the response variable of e^{β} units. Thus, for example, an increase in the *rank* of a song by 1, when all other predictor variables are held constant, will result in a multiplicative increase in the predicted *days since song release* by $e^{.00629}$ days. Additionally, note that the predicted intercept of $e^{3.75123}$

has no meaningful interpretation since it would imply that the rank of the song on the Spotify weekly top 200 list would be zero, which is an impossibility. Lasty, we can note that the factor with the largest effect on the days between a song's release and its appearance on the weekly top 200 list is speechiness with a multiplicative decrease of $e^{1.90692}$ days per unit increase in speechiness.

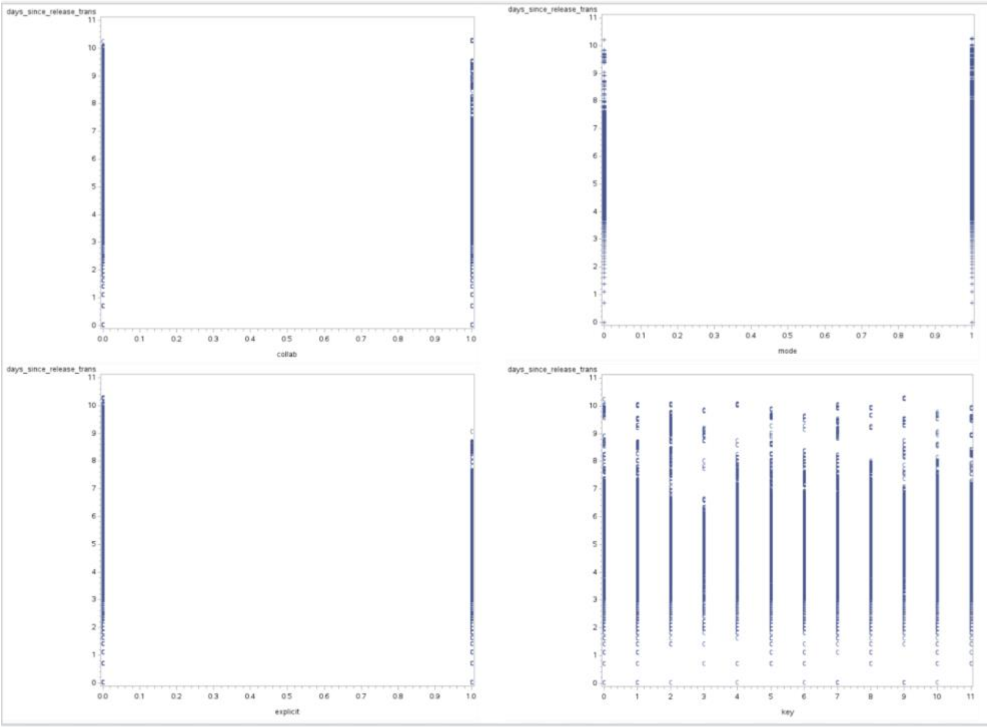
Factor Analysis:

Next, we are interested in seeing if any categorical variables contain information useful in explaining the variability of "days_since_release" in addition to the model we created above. These variables are: artists' collaboration (Collab), music explicit (explicit), the key the track is in (key), the modality (major or minor) of a track (Mode).

We first work on the factor analysis with these four categorical variables. Then, we focus on some of the interaction between categorical variables and numeric variables by showing the interaction plot and one-way or two-way ANOVA. By doing so, we can find the useful categorical variables and their interactions.

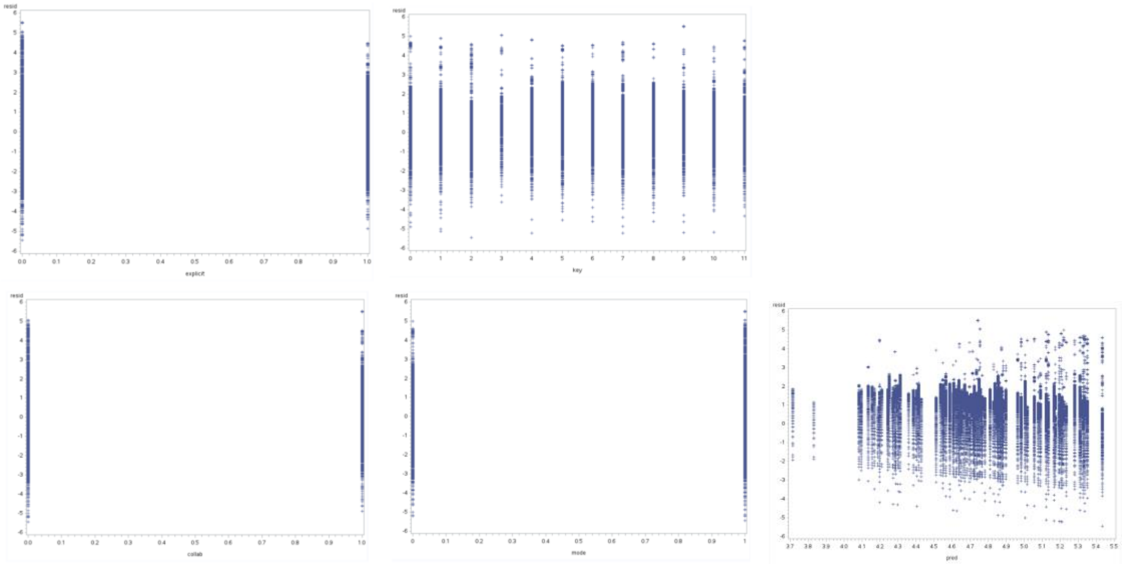
We start by checking the Normality Assumption of factor variables.

Figure 16: Scatter plots of factor variables



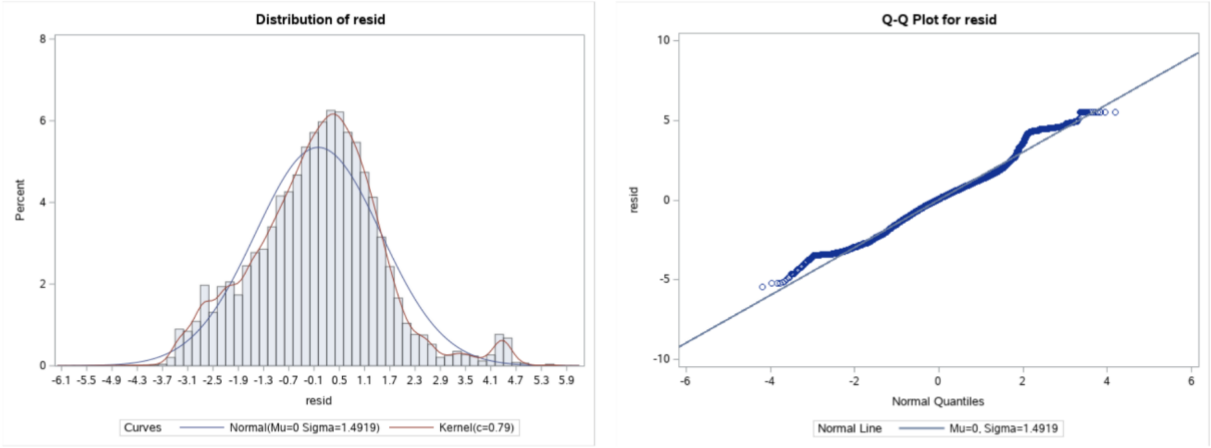
There are few deviations among scatter plots of all factor variables, meaning there are few outliers and independence is fulfilled.

Figure 17: Residual plots of factor variables



Also, from the plots above, there is no odd variance either. Thus, we can assume that among categorical variables, the assumption that the underlying residuals are normally distributed, or approximately so.

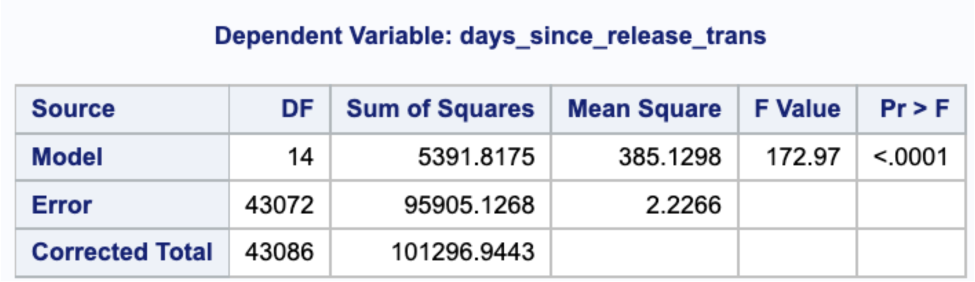
Figure 18: Histogram and QQ plot for residuals



In both plots above, while there are still some slight deviations from normality near the tails of the data, they are negligible. Thus, it appears that the Normality Assumption of the Independent variable is now met. Overall, factor variables meet the normality and aptness for the model.

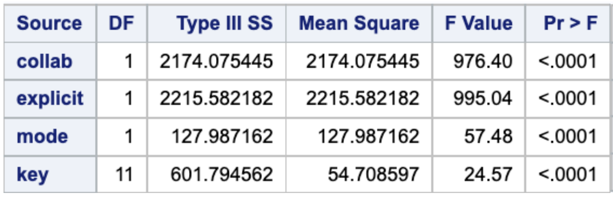
Now, we move on the ANOVA with factor variables.

Figure 19: ANOVA table of categorical variables

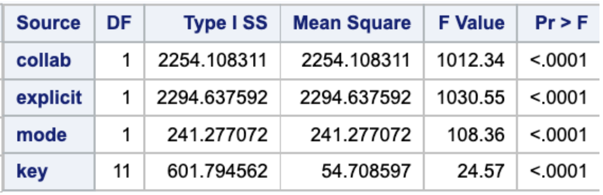


Dependent Variable: days_since_release_trans

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	14	5391.8175	385.1298	172.97	<.0001
Error	43072	95905.1268	2.2266		
Corrected Total	43086	101296.9443			



Source	DF	Type III SS	Mean Square	F Value	Pr > F
collab	1	2174.075445	2174.075445	976.40	<.0001
explicit	1	2215.582182	2215.582182	995.04	<.0001
mode	1	127.987162	127.987162	57.48	<.0001
key	11	601.794562	54.708597	24.57	<.0001

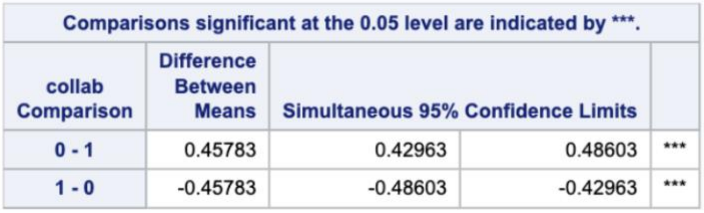


Source	DF	Type I SS	Mean Square	F Value	Pr > F
collab	1	2254.108311	2254.108311	1012.34	<.0001
explicit	1	2294.637592	2294.637592	1030.55	<.0001
mode	1	241.277072	241.277072	108.36	<.0001
key	11	601.794562	54.708597	24.57	<.0001

As we see the ANOVA table and TYPE {I | III} table above, we can see that overall, the model with categorical variables is significant at F-test; F(14, 43072, 0.95). The significant value is < .0001. Type {I | III} are all significant as well, meaning they have some sort of information useful in explaining the variability of "days_since_release." Furthermore, Type I and type III are very similar with values, meaning the variation of the model is well balanced.

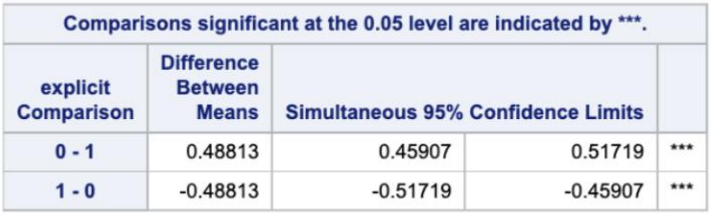
Now, we will consider every factor level means. We conduct a simultaneous hypothesis test with the Tukey method.

Figure 20: Simultaneous comparisons of every levels by Tukey method with α=0.05



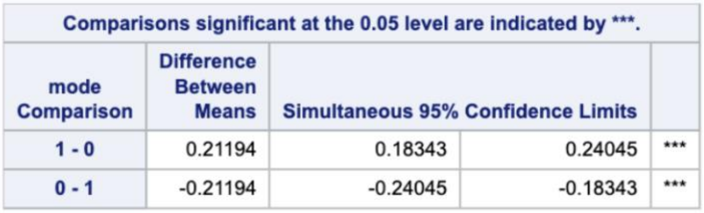
Comparisons significant at the 0.05 level are indicated by ***.

collab Comparison	Difference Between Means	Simultaneous 95% Confidence Limits		
0 - 1	0.45783	0.42963	0.48603	***
1 - 0	-0.45783	-0.48603	-0.42963	***



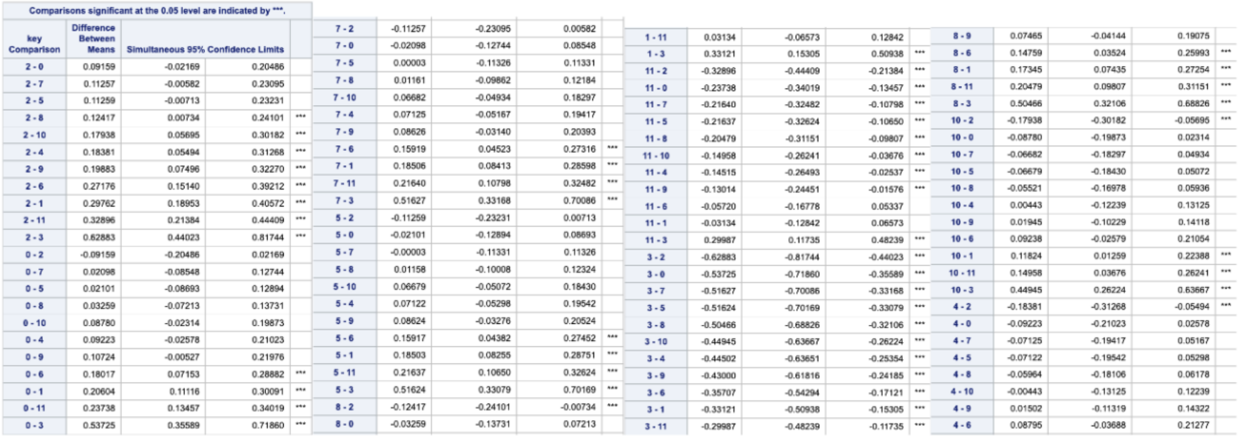
Comparisons significant at the 0.05 level are indicated by ***.

explicit Comparison	Difference Between Means	Simultaneous 95% Confidence Limits		
0 - 1	0.48813	0.45907	0.51719	***
1 - 0	-0.48813	-0.51719	-0.45907	***



Comparisons significant at the 0.05 level are indicated by ***.

mode Comparison	Difference Between Means	Simultaneous 95% Confidence Limits		
1 - 0	0.21194	0.18343	0.24045	***
0 - 1	-0.21194	-0.24045	-0.18343	***



Comparisons significant at the 0.05 level are indicated by ***.

key Comparison	Difference Between Means	Simultaneous 95% Confidence Limits		
2 - 0	0.09159	-0.02169	0.20486	
2 - 7	0.11257	-0.00582	0.23095	
2 - 5	0.11259	-0.00713	0.23231	
2 - 8	0.12417	0.00734	0.24101	***
2 - 10	0.17938	0.05695	0.30182	***
2 - 4	0.18381	0.05494	0.31268	***
2 - 9	0.19883	0.07496	0.32270	***
2 - 6	0.27176	0.15140	0.39212	***
2 - 1	0.29762	0.18953	0.40572	***
2 - 11	0.32896	0.21384	0.44409	***
2 - 3	0.62883	0.44023	0.81744	***
0 - 2	-0.09159	-0.20486	0.02169	
0 - 7	0.02098	-0.08548	0.12744	
0 - 5	0.02101	-0.08693	0.12894	
0 - 8	0.03259	-0.07213	0.13731	
0 - 10	0.08780	-0.02314	0.19873	
0 - 4	0.09223	-0.02578	0.21023	
0 - 9	0.10724	-0.00527	0.21976	
0 - 6	0.18017	0.07153	0.28882	***
0 - 1	0.20604	0.11116	0.30091	***
0 - 11	0.23738	0.13457	0.34019	***
0 - 3	0.53725	0.35589	0.71860	***
7 - 2	-0.11257	-0.23095	0.00582	
7 - 0	-0.02098	-0.12744	0.08548	
7 - 5	0.00003	-0.11326	0.11331	
7 - 8	0.01161	-0.09862	0.12184	
7 - 10	0.06682	-0.04934	0.18297	
7 - 4	0.07125	-0.05167	0.19417	
7 - 9	0.08626	-0.03140	0.20393	
7 - 6	0.15919	0.04523	0.27316	***
7 - 1	0.18506	0.08413	0.28598	***
7 - 11	0.21640	0.10798	0.32482	***
7 - 3	0.51627	0.33168	0.70086	***
5 - 2	-0.11259	-0.23231	0.00713	
5 - 0	-0.02101	-0.12894	0.08693	
5 - 7	-0.00003	-0.11331	0.11326	
5 - 8	0.01158	-0.10008	0.12324	
5 - 10	0.06679	-0.05072	0.18430	
5 - 4	0.07122	-0.05298	0.19542	
5 - 9	0.08624	-0.03276	0.20524	
5 - 6	0.15917	0.04382	0.27452	***
5 - 1	0.18503	0.08255	0.28751	***
5 - 11	0.21637	0.10650	0.32624	***
5 - 3	0.51624	0.33079	0.70169	***
8 - 2	-0.12417	-0.24101	-0.00734	***
8 - 0	-0.03259	-0.13731	0.07213	
1 - 11	0.03134	-0.06573	0.12842	
1 - 3	0.33121	0.15305	0.50938	***
11 - 2	-0.32896	-0.44409	-0.21384	***
11 - 0	-0.23738	-0.34019	-0.13457	***
11 - 7	-0.21640	-0.32482	-0.10798	***
11 - 5	-0.21637	-0.32624	-0.10650	***
11 - 8	-0.20479	-0.31151	-0.09807	***
11 - 10	-0.14958	-0.26241	-0.03676	***
11 - 4	-0.14515	-0.26493	-0.02537	***
11 - 9	-0.13014	-0.24451	-0.01576	***
11 - 6	-0.05720	-0.16778	0.05337	
11 - 1	-0.03134	-0.12842	0.06573	
11 - 3	0.29987	0.11735	0.48239	***
3 - 2	-0.62883	-0.81744	-0.44023	***
3 - 0	-0.53725	-0.71860	-0.35589	***
3 - 7	-0.51627	-0.70086	-0.33168	***
3 - 5	-0.51624	-0.70169	-0.33079	***
3 - 8	-0.50466	-0.68826	-0.32106	***
3 - 10	-0.44945	-0.63667	-0.26224	***
3 - 4	-0.44502	-0.63651	-0.25354	***
3 - 9	-0.43000	-0.61816	-0.24185	***
3 - 6	-0.35707	-0.54294	-0.17121	***
3 - 1	-0.33121	-0.50938	-0.15305	***
3 - 11	-0.29987	-0.48239	-0.11735	***
8 - 9	0.07465	-0.04144	0.19075	
8 - 6	0.14759	0.03524	0.25993	***
8 - 1	0.17345	0.07435	0.27254	***
8 - 11	0.20479	0.09807	0.31151	***
8 - 3	0.50466	0.32106	0.68826	***
10 - 2	-0.17938	-0.30182	-0.05695	***
10 - 0	-0.08780	-0.19873	0.02314	
10 - 7	-0.06682	-0.18297	0.04934	
10 - 5	-0.06679	-0.18430	0.05072	
10 - 8	-0.05521	-0.16978	0.05936	
10 - 4	0.00443	-0.12239	0.13125	
10 - 9	0.01945	-0.10229	0.14118	
10 - 6	0.09238	-0.02579	0.21054	
10 - 1	0.11824	0.01259	0.22388	***
10 - 11	0.14958	0.03676	0.26241	***
10 - 3	0.44945	0.26224	0.63667	***
4 - 2	-0.18381	-0.31268	-0.05494	***
4 - 0	-0.09223	-0.21023	0.02578	
4 - 7	-0.07125	-0.19417	0.05167	
4 - 5	-0.07122	-0.19542	0.05298	
4 - 8	-0.05964	-0.18106	0.06178	
4 - 10	-0.00443	-0.13125	0.12239	
4 - 9	0.01502	-0.11319	0.14322	
4 - 6	0.08795	-0.03688	0.21277	

We can see that “collab,” “explicit, ” “mode” variables comparison are significant. Half of “key” variables comparison are significant as well. Even though the other half are insignificant, we can assume that the “key” variable is an effective categorical data.

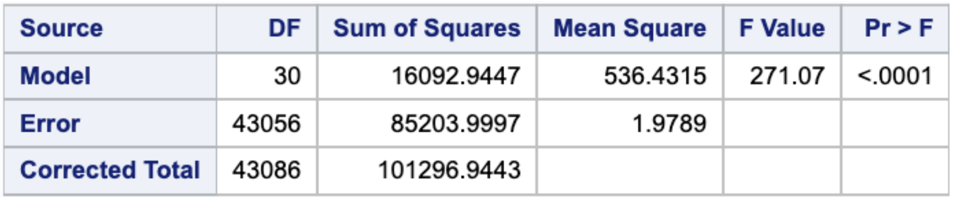
Interaction model:

Next, we will check some of the interaction between categorical variables and numeric variables by showing the interaction plot and one-way or two-way ANOVA.

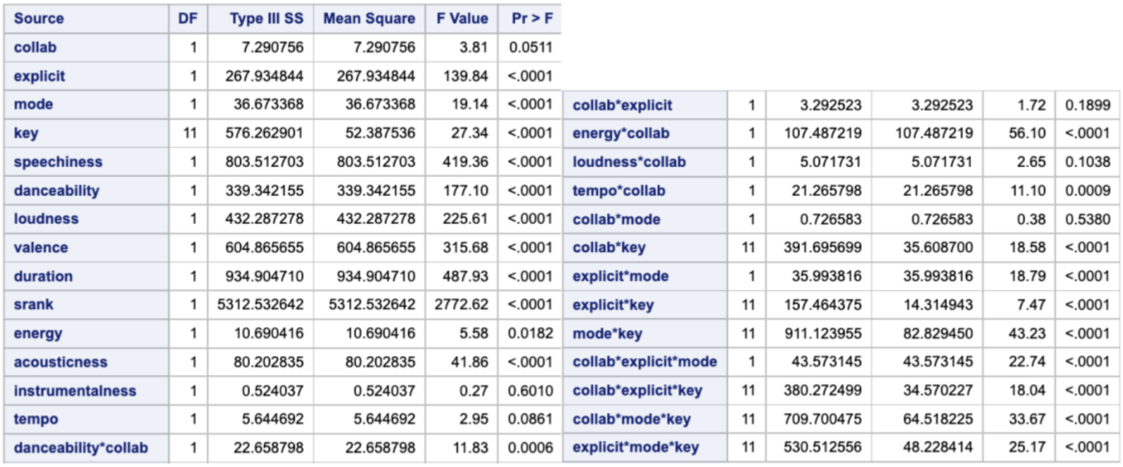
We cannot conduct an analysis of all interaction within 15 variables. Therefore, we only consider some variables of interest.

We first build the interaction model centering on the “collab” variable.We assume the “collab” variable is effective in terms of a model prediction since these days, many popular artists collaborate with other artists and generate hit songs. Considering we handle the data from Spotify which is especially favored among young generations more than older generations, those collaborated music can stay at Spotify top 200 longer than non-collab music.

Figure 21: ANOVA table for an interaction



Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	30	16092.9447	536.4315	271.07	<.0001
Error	43056	85203.9997	1.9789		
Corrected Total	43086	101296.9443			

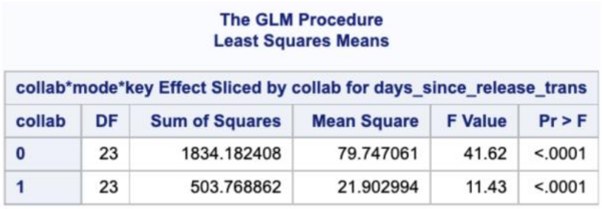


Source	DF	Type III SS	Mean Square	F Value	Pr > F
collab	1	7.290756	7.290756	3.81	0.0511
explicit	1	267.934844	267.934844	139.84	<.0001
mode	1	36.673368	36.673368	19.14	<.0001
key	11	576.262901	52.387536	27.34	<.0001
speechiness	1	803.512703	803.512703	419.36	<.0001
danceability	1	339.342155	339.342155	177.10	<.0001
loudness	1	432.287278	432.287278	225.61	<.0001
valence	1	604.865655	604.865655	315.68	<.0001
duration	1	934.904710	934.904710	487.93	<.0001
srank	1	5312.532642	5312.532642	2772.62	<.0001
energy	1	10.690416	10.690416	5.58	0.0182
acousticness	1	80.202835	80.202835	41.86	<.0001
instrumentalness	1	0.524037	0.524037	0.27	0.6010
tempo	1	5.644692	5.644692	2.95	0.0861
danceability*collab	1	22.658798	22.658798	11.83	0.0006
collab*explicit	1	3.292523	3.292523	1.72	0.1899
energy*collab	1	107.487219	107.487219	56.10	<.0001
loudness*collab	1	5.071731	5.071731	2.65	0.1038
tempo*collab	1	21.265798	21.265798	11.10	0.0009
collab*mode	1	0.726583	0.726583	0.38	0.5380
collab*key	11	391.695699	35.608700	18.58	<.0001
explicit*mode	1	35.993816	35.993816	18.79	<.0001
explicit*key	11	157.464375	14.314943	7.47	<.0001
mode*key	11	911.123955	82.829450	43.23	<.0001
collab*explicit*mode	1	43.573145	43.573145	22.74	<.0001
collab*explicit*key	11	380.272499	34.570227	18.04	<.0001
collab*mode*key	11	709.700475	64.518225	33.67	<.0001
explicit*mode*key	11	530.512556	48.228414	25.17	<.0001

We can see the overall F-test statistics (271.07) is large enough to be significant. Yet, as we check the simultaneous F-test table, there are some insignificant variables. Some of the quantitative variables such as “instrumentalness,” “tempo,” were significant when the model only contained quantitative variables, but now they are insignificant.

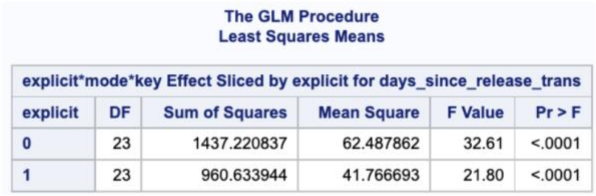
Before we define the model, we can analyze some of those significant interactions.

Figure 22: Expected Mean comparison



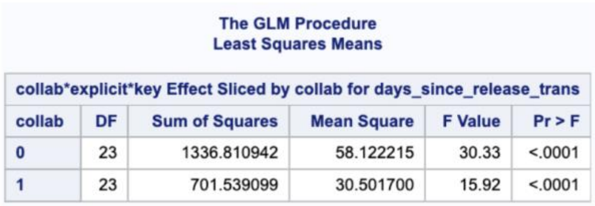
The GLM Procedure
Least Squares Means

collab*mode*key Effect Sliced by collab for days_since_release_trans					
collab	DF	Sum of Squares	Mean Square	F Value	Pr > F
0	23	1834.182408	79.747061	41.62	<.0001
1	23	503.768862	21.902994	11.43	<.0001



The GLM Procedure
Least Squares Means

explicit*mode*key Effect Sliced by explicit for days_since_release_trans					
explicit	DF	Sum of Squares	Mean Square	F Value	Pr > F
0	23	1437.220837	62.487862	32.61	<.0001
1	23	960.633944	41.766693	21.80	<.0001



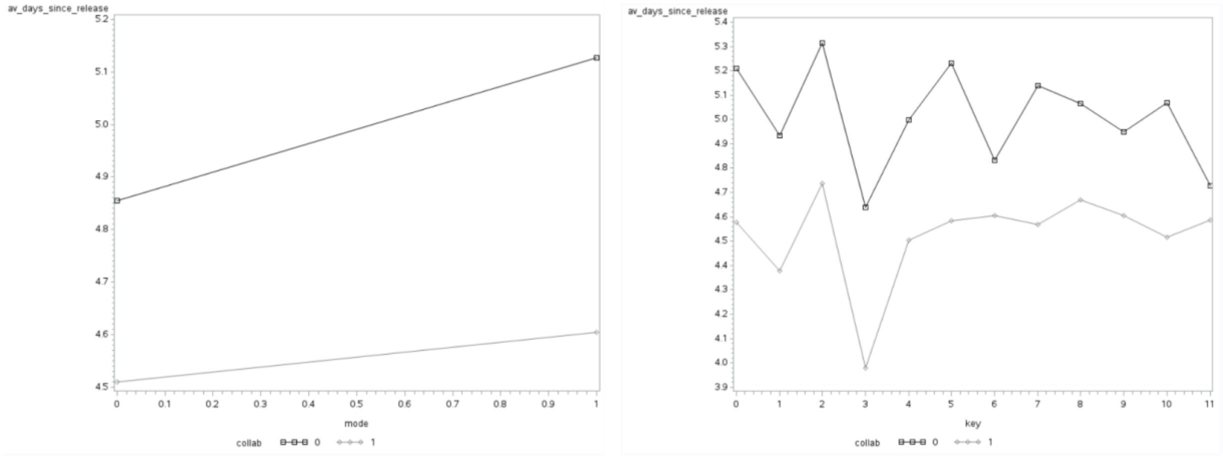
The GLM Procedure
Least Squares Means

collab*explicit*key Effect Sliced by collab for days_since_release_trans					
collab	DF	Sum of Squares	Mean Square	F Value	Pr > F
0	23	1336.810942	58.122215	30.33	<.0001
1	23	701.539099	30.501700	15.92	<.0001

Even though we do not show all of the analyses for each level of one factor by using the "*slice"* option, most of the conditional interactions are significant. "*Collab*explicit*," "*Collab*Loudness*," "*Collab*Mode*" are only two insignificant interactions in this model.

We can also check some of the categorical interactions in interaction plots. Left plot is "*collab*mode,*" which is insignificant and right plot is "*Collab*Key*" which is significant. Left plot seems to have some interaction but as test statistics says, it is still insignificant. Right plot obviously has some interaction between different levels.

Figure 23: Interaction plot of average_days_since_release (y_value) and mode while controlling collab; same plot of average_days_since_release (y_value) and key while controlling collab;



Based on the factor analysis, the beneficial variables that we can present here is below:

- **Quantitative Variables**: *single, speechness, danceability, loudness, valence, duration, srank, energy, acousticness, tempo,*
- **Categorical Variables**: *collab, explicit, mode, key*
- **Interaction:** *danceability*collab, energy*collab, tempo*collab, collab*key, explicit*mode, explicit*key, mode*key, collab*explicit*mode, collab*explicit*key, collab*mode*key, explicit*mode*key*

It is interesting to see “collab*explicit” are not significant here. Many artists these days collaborate with other major artists and make some explicit music (one that has curse words or language or art that is generally deemed sexual, violent, or offensive in nature, which gains popular especially among young generations). We can assume that even though those songs can be great hits for a certain period of time, those songs less likely remain on top 200 for long. Overall, interaction between “collab” and other variables are significant, showing they have some information.

Conclusion

We found that the best model (out of all models containing the predictors specified at the beginning of the Model Selection section) for predicting the number of days between a songs release and its appearance on the Spotify top 200 list was given by the model containing the predictors: rank of song during a given week (*srank*), *danceability, energy, loudness, speechiness, acousticness, instrumentalness, liveness, valence, tempo,* and *duration.* This was chosen as the best model according to Mallow’s Cp, Adjusted R-Squared, and PRESS criterion.Additionally, we found that the chosen model is approximately unbiased and accounts for about 12.79% of the variability in our dependent variable. Finally, we can note that the predictor which is most influential in predicting the number of days between a song’s release and its appearance on the Spotify top 200 list was *speechiness*.

In the factor analysis, we found there is little violation of normal assumptions among all categorical variables. We analyzed them and checked the interaction mainly with “collab” variables since collaborated songs are likely to be hits. After the analysis, we got the idea that “collab” in general has a significant effect on building a model while interacting variables between “collab” and “explicit” is not significant, which is surprising for us. As well as the “collab” variable, many categorical variables themselves have additional information to our model. We can use those significant categorical variables and their interactions to explain further and give some hints to the artists to stay on top 200 longer.

Appendix:

Model Selection:

```
* Change work directory;
DATA _NULL_;
        rc=dlgcdir("\\nas01.itap.purdue.edu\puhome\my documents\My SAS Files\9.4\Final_Project");
        PUT rc=;
RUN;

* Import dataset to be analyzed;
OPTIONS OBS=max;
PROC IMPORT DATAFILE="spotify-top-200-dataset.csv"
        OUT=songs_in
        DBMS=csv
        REPLACE;
        DELIMITER=";";

        GETNAMES=yes;
RUN;

/* Create new dataset by adding new variable to track the time since release
   of a song and its appearance on the week's top 200 list. Additionally, drop irrelevant
   variables to simplify analysis and change name of rank variable to avoid syntax conflicts.
   Lastly, removed all observations with non-positive values of days_since_release
*/
DATA songs_dupes; SET songs_in;
        days_since_release = intck('day', release_date, week);
        srank = rank;
        IF days_since_release < 1 THEN DELETE;
        DROP rank track_id track_number album_id album_img artist_id artist_img track_index rank;

* Create new dataset by removing all duplicate entries of each song from song_dupes;
DATA songs_nodupes; SET songs_dupes;
        IF pivot=1 THEN DELETE;
        DROP pivot;

* Check normality assumption of days_since_release (dependent) variable;
PROC UNIVARIATE;
        VAR days_since_release;
        HISTOGRAM days_since_release / NORMAL KERNEL (L=2);
        QQPLOT days_since_release / NORMAL(L=1 mu=est sigma=est);
RUN; QUIT;

* Performing Box-Cox Transformation with the Grand Model to ensure that the normality
```

```
  assumption of the residuals is met;
PROC TRANSREG;
          MODEL BOXCOX(days_since_release/lambda=-3 to 3 by .1) = IDENTITY(speechiness)
                    IDENTITY(danceability) IDENTITY(loudness) IDENTITY(valence)
IDENTITY(duration) IDENTITY(srank)
                    IDENTITY(energy) IDENTITY(acousticness) IDENTITY(instrumentalness)
IDENTITY(liveness) IDENTITY(tempo);
RUN; QUIT;

DATA nodupes_trans; SET songs_nodupes;
          days_since_release_trans = LOG(days_since_release);
          DROP days_since_release;

* Check Normality Assumption of transformed dependent variable;
PROC UNIVARIATE;
          VAR days_since_release_trans;
          HISTOGRAM days_since_release_trans / NORMAL KERNEL (L=2);
          QQPLOT days_since_release_trans / NORMAL (L=1 mu=est sigma=est);
RUN; QUIT;


* Check scatterplot matrix for if any assumptions are violated.;
PROC SGSCATTER;
          TITLE "Scatterplot Matrix for Box-Cox (lambda = 0) Transformed Song Data";
          MATRIX days_since_release_trans speechiness danceability loudness valence duration srank
energy acousticness instrumentalness liveness tempo;
RUN; TITLE;

* Justify modeling choices via pairwise correlation matrix;
PROC CORR;
          VAR days_since_release_trans srank danceability energy loudness speechiness acousticness
instrumentalness liveness valence tempo duration;
RUN; QUIT;


* Perform All Subset Selection with Adjusted R-Squared, Mallow’s CP;
PROC REG;
          MODEL days_since_release_trans = srank speechiness instrumentalness danceability energy loudness
acousticness liveness valence duration / selection = adjrsq cp PRESS b;
RUN; QUIT;

* Check the PRESS criterion of the best models with 7, 8, 9, and 10 variables;
PROC REG OUTEST=sumstats1 PRESS;
          MODEL days_since_release_trans = srank speechiness danceability loudness acousticness valence
duration;
```

```
    MODEL days_since_release_trans = srank speechiness instrumentalness danceability loudness
acousticness valence duration;
    MODEL days_since_release_trans = srank speechiness instrumentalness danceability energy loudness
acousticness valence duration;
    MODEL days_since_release_trans = srank speechiness instrumentalness danceability energy loudness
acousticness liveness valence duration;

PROC PRINT DATA=sumstats1; RUN; QUIT;

* Check Assumptions of choosen model using diagnostic plots;
PROC REG DATA = nodupes_trans PLOTS(MAXPOINTS=NONE);
    MODEL days_since_release_trans = srank speechiness instrumentalness danceability energy loudness
acousticness valence duration / PARTIAL;
RUN; QUIT;

* Calculate analysis for residuals and influence statistics, saving to new dataset;
PROC REG DATA=nodupes_trans NOPRINT;
    MODEL days_since_release_trans = srank speechiness instrumentalness danceability energy loudness
acousticness valence duration / R INFLUENCE;
    OUTPUT OUT=reg_stats1 RSTUDENT=rstudent COOKD=cooksd;
RUN; QUIT;

DATA reg_stats1; SET reg_stats1;
    abs_rstudent = ABS(rstudent);
    DROP rstudent;

* Using previously calculated statistics, check for potential outliers and influential points;
PROC SORT; BY DECENDING cooksd;
PROC PRINT DATA=reg_stats1 (OBS=15); VAR cooksd abs_rstudent; RUN; QUIT;

PROC SORT; BY DESCENDING abs_rstudent;
PROC PRINT DATA=reg_stats1 (OBS=15); VAR cooksd abs_rstudent; RUN; QUIT;

* Calculate the VIF for the parameters to check for multicollinearity issues;
PROC REG DATA=nodupes_trans;
    MODEL days_since_release_trans = srank speechiness instrumentalness danceability energy loudness
acousticness valence duration / TOL;
RUN; QUIT;
```

Factor Analysis:

```
%let path=/home/u62387331/STAT525;
libname stat525 base "/home/u62387331/STAT525";

data stat525.nodupes; set stat525.nodupes (rename=(collab=oldcollab explicit=oldexplicit));
    if oldcollab = 'True' then collab = 1; else collab = 0;
    if oldexplicit = 'True' then explicit = 1; else explicit = 0;
    drop oldcollab oldexplicit;
```

```
run;

/* transformation */
* Box-Cox transformation;
PROC TRANSREG data=stat525.nodupes;
        MODEL BOXCOX(days_since_release/lambda=-3 to 3 by .1) = IDENTITY(speechiness)
        IDENTITY(danceability) IDENTITY(loudness) IDENTITY(valence) IDENTITY(duration)
        IDENTITY(srank) IDENTITY(energy) IDENTITY(acousticness) IDENTITY(instrumentalness)
        IDENTITY(liveness) IDENTITY(tempo)
        identity(collab) identity(explicit) identity(Mode) identity(key);
run;

* log transformation ;
data stat525.nodupes_trans; set stat525.nodupes;
        days_since_release_trans = log(days_since_release);
        drop days_since_release;
run; quit;


/* count the number of value by each category */
* count the key;
proc sql;
   select key, count(*) as total_count
   from stat525.nodupes_trans
   group by key;
run;

* count the time_signature;
proc sql;
   select time_signature, count(*) as total_count
   from stat525.nodupes_trans
   group by time_signature;
run; quit;


/* factor analysis assumption check */
* Scatterplot;
symbol1 v=cirle i=none; symbol2 v=diamond i=join c=blue;
proc gplot data=stat525.nodupes_trans;
        plot days_since_release_trans*collab/frame;

proc gplot data=stat525.nodupes_trans;
        plot days_since_release_trans*explicit/frame;
```

```
proc gplot data=stat525.nodupes_trans;
	plot days_since_release_trans*key/frame;

proc gplot data=stat525.nodupes_trans;
	plot days_since_release_trans*mode/frame;
run; quit;

* ANOVA;
proc glm data=stat525.nodupes_trans;
	class collab explicit Mode key;
	model days_since_release_trans=collab explicit Mode key;
	means collab explicit Mode key/ tukey;
	lsmeans collab explicit Mode key/ stderr;
	output out=stat525.nodupes_trans_out r=resid p=pred;
run;

* residual plot;
proc gplot data=stat525.nodupes_trans_out;
	plot resid*(collab explicit mode key pred);
run; quit;

* Histogram & QQplot;
proc univariate noprint data=stat525.nodupes_trans_out;
	histogram resid / normal kernel(L=2);
	qqplot resid / normal (L=1 mu=est sigma=est);
run; quit;


* Pairwise comparisons with Tukey procedure;
proc glm data=stat525.nodupes_trans;
	class collab explicit Mode key;
	model days_since_release_trans=collab explicit Mode key;
	means collab explicit Mode key/ tukey CLDIFF;
run; quit;


/* interaction model */
proc glm data=stat525.nodupes_trans;
	class collab explicit Mode key;
	model days_since_release_trans = collab | explicit | Mode | key speechiness danceability loudness
valence duration srank energy acousticness instrumentalness tempo;
run; quit;

proc glm data=stat525.nodupes_trans;
```

```
        class collab explicit Mode key;
        model days_since_release_trans
        = collab explicit Mode key speechiness danceability loudness valence duration srank energy acousticness instrumentalness tempo
        collab*danceability collab*energy collab*loudness collab*tempo
        collab*explicit collab*Mode collab*key explicit*Mode explicit*key Mode*key
        collab*explicit*Mode collab*explicit*key collab*Mode*key explicit*Mode*key
        ;
run; quit;

* Analysis of interactions;
proc glm data=stat525.nodupes_trans;
        class collab explicit Mode key;
        model days_since_release_trans
        = collab explicit Mode key speechiness danceability loudness valence duration srank energy acousticness instrumentalness tempo
        collab*danceability collab*energy collab*loudness collab*tempo
        collab*explicit collab*Mode collab*key explicit*Mode explicit*key Mode*key
        collab*explicit*Mode collab*explicit*key collab*Mode*key explicit*Mode*key
        ;
        lsmeans explicit*key*mode / slice=explicit;
run; quit;

* interaction plot;
proc sort data=stat525.nodupes_trans; by collab key;
proc means data=stat525.nodupes_trans;
        var days_since_release_trans; by collab key;
        output out=stat525.nodupes_trans_out2 mean=av_days_since_release;

symbol1 v=square i=join c=black; symbol2 v=diamond i=join c=bgr;
proc gplot data=stat525.nodupes_trans_out2;
        plot av_days_since_release*key=collab/frame;
run; quit;

proc sort data=stat525.nodupes_trans; by collab mode;
proc means data=stat525.nodupes_trans;
        var days_since_release_trans; by collab mode;
        output out=stat525.nodupes_trans_out2 mean=av_days_since_release;

symbol1 v=square i=join c=black; symbol2 v=diamond i=join c=bgr;
proc gplot data=stat525.nodupes_trans_out2;
        plot av_days_since_release*mode=collab/frame;
run; quit;
```