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](https://developer.spotify.com/documentation/web-api/reference/#/operations/get-audio-features). 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*

Distribution of 
2 
60 
2400 
4800 
7200 
9600 
12000 
14400 16800 19200 
days_since release 
.75 Sigma-2228.9) 
21600 24000 26400 28800 
Curves 

*Figure 2: QQ-Plot of days\_since\_release*

Q-Q Plot for 
30000 
20000 
10000 
-1 0000 
.20000 
Normal Line 
Normal Ouantiles 
Mu=581.75, Sigma-2228.9 

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*

Box-Cox Analysis for 
-2000tn 
-c, .3000m 
_4000cn 
sooom 
-6001m 
Lambda 
Selected A = O 
a 95%Cl 

As seen in the above Box-Cox Analysis Plot the optimal value of for our transformation was found to be . We will thus apply the Box-Cox Transformation using the transformation, , 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 () days\_since\_release*

Distribution of 
-0.6 0.0 
06 1.2 
Curves 
1.8 
2.4 
3.0 3.6 4.2 
4.8 5.4 
6.0 6.6 
days_since_release trans 
Sigma-I .5333) 
7.2 
7.8 8.4 9.0 9.6 10.2 10.8 

*Figure 5: QQ-Plot of Box-Cox Transformed () days\_since\_release*

Q-Q Plot for days _ 
Normal Ouantiles 
Normal Line 
Mu=4.799, Sigma-I .5333 

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 () Data*

Od 山 
00 : OS 
00 ~ 0 
00000900 
00000 
000009 
0u8 
L 
SS8ul 
388dS 
300 6uos p ~ 0 u ( 0 = epqwel) x00 • x09 」 0 乂 ! | 山 ~ 

*Figure 7: Pairwise Pearson Correlation Table of Data*

Pearson Correlation Coefficients, N = 43087 
Prob > Irl under HO: Rho=0 
days_since_release_trans 
1 _ooooo 
0_24850 
-0_15564 
-0_08685 
-0_11766 
_0_14108 
0_13223 
0_04064 
0_01179 
0_0144 
0_00781 
0_1050 
-0_04058 
0_11372 
srank 
0_24850 
1 _ooooo 
-0_05373 
-0_00025 
0_9592 
-0_03904 
0_01169 
0_0152 
0_01492 
0_0020 
0_01000 
0_0379 
0_01978 
-0_02685 
0_02892 
0_05585 
danceability 
-0_15564 
-0_05373 
1 _ooooo 
0_08281 
0_15208 
0_19316 
-0_23816 
-0_02874 
-0_05678 
0_35768 
-0_04889 
-0_12013 
energy 
-0_08685 
-0_00025 
0_9592 
0_08281 
1 _ooooo 
0_74519 
-0_01939 
-0_51455 
-0_10086 
0_07015 
0_39807 
0_08887 
0_04731 
loudness speechiness acousticness 
days_since_release_trans 
srank 
danceability 
energy 
loudness 
speech i ness 
acousticness 
instrumentalness 
liveness 
valence 
tempo 
duration 
-0_11766 
-0_03904 
0_15208 
0_74519 
1 _ooooo 
-0_11476 
-0_39402 
-0_18977 
-0_00092 
0_8480 
0_36127 
0_05362 
0_03394 
_0_14108 
0_01169 
0_0152 
0_19316 
-0_01939 
-0_11476 
1 _ooooo 
-0_06786 
-0_00227 
0_6377 
0_01007 
0_0365 
0_00945 
0_0498 
0_19264 
-0_04607 
0_13223 
0_01492 
0_0020 
-0_23816 
-0_51455 
-0_39402 
-0_06786 
1 _ooooo 
0_09038 
-0_06353 
-0_09756 
-0_11408 
-0_03780 
instrumentalness 
0_04064 
0_01000 
0_0379 
-0_02874 
-0_10086 
-0_18977 
-0_00227 
0_6377 
0_09038 
1 _ooooo 
0_02027 
-0_08919 
0_03121 
0_00091 
o_84gg 
liveness 
0_01179 
0_0144 
0_01978 
-0_05678 
0_07015 
-0_00092 
0_8480 
0_01007 
0_0365 
-0_06353 
0_02027 
1 _ooooo 
0_00265 
0_5822 
0_00248 
0_6064 
-0_01253 
o_ooga 
valence 
0_00781 
0_1050 
-0_02685 
0_35768 
0_39807 
0_36127 
0_00945 
0_0498 
-0_09756 
-0_08919 
0_00265 
0_5822 
1 _ooooo 
-0_01289 
0_0074 
-0_07756 
tempo 
-0_04058 
0_02892 
-0_04889 
0_08887 
0_05362 
0_19264 
-0_11408 
0_03121 
0_00248 
0_6064 
-0_01289 
0_0074 
1 _ooooo 
0_01713 
0_0004 
duration 
0_11372 
0_05585 
-0_12013 
0_04731 
0_03394 
-0_04607 
-0_03780 
0_00091 
0_8499 
-0_01253 
0_0093 
-0_07756 
0_01713 
0_0004 
1 _ooooo 

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 and Mallow’s Cp Criteria*

Parameter Estimates 
Number in 
Model 
10 
9 
9 
8 
9 
8 
8 
7 
9 
8 
8 
7 
7 
8 
6 
7 
9 
8 
8 
7 
8 
7 
7 
Adj usted 
R-Square 
0_1294 
0_1294 
0_1293 
0_1293 
0_ 1291 
0_1291 
0_1291 
0_ 1290 
0_1277 
0_1277 
0_1273 
0_1273 
0_1267 
0_1267 
0_1265 
0_1265 
0_1241 
0_1240 
0_1233 
0_1231 
0_1224 
0_1223 
0_1214 
R-Square 
0_1296 
0_1295 
0_1295 
0_1294 
0_1293 
0_1293 
0_1292 
0_1292 
0_1279 
0_1278 
0_1275 
0_1275 
0_1269 
0_1269 
0_1266 
0_1266 
0_1243 
0_1242 
0_1234 
0_1233 
0_1226 
0_1225 
0_1216 
ll_oooo 
11_2423 
1442 
14_9127 
22_4171 
22_8539 
25_1097 
25_1018 
94_3665 
93_4500 
109_8370 
109_0515 
138_2770 
140_1448 
150_0083 
151_8110 
270_6937 
275_1552 
311_1712 
316_6258 
356_7482 
402_0944 
Intercept 
3_73940 
3_75123 
3_56518 
3_58216 
3_70606 
3_71815 
3_55495 
3_57289 
4_08535 
4_09165 
4_05591 
4_06247 
3_67365 
3_66928 
3_66438 
3_65907 
4_75827 
4_78870 
4_76912 
4_80195 
5_10557 
5_12668 
5_13223 
srank 
0_00629 
0_00629 
0_00628 
0_00628 
0_00629 
0_00629 
0_00628 
0_00628 
0_00630 
0_00630 
0_00630 
0_00630 
0_00627 
0_00627 
0_00627 
0_00627 
0_00638 
0_00638 
0_00639 
0_00639 
0_00639 
0_00640 
0_00640 
speechiness 
-1_90790 
-1_90692 
1_92432 
1 _ 92285 
-1_91618 
-1_91523 
-1_93016 
1 _ 92865 
-1_93081 
-1_93001 
1 94085 
1 94004 
-1_98831 
-1 _ 98889 
-1_99488 
-1_99558 
-1_70464 
-1_70093 
-1_70543 
-1_70145 
-1_72747 
-1_72417 
-1_72930 
instrumentalness 
0_51683 
0_52048 
0_48619 
0_49058 
0_58885 
o_sgogg 
0_52059 
0_51948 
0_90612 
0_91627 
0_97853 
0_98606 
danceability 
1_36377 
-1_36958 
-1_32809 
-1_36099 
-1_32415 
-1_33101 
_ I _ 52552 
-1_52866 
-1_51954 
-1_52285 
-1 47253 
-1_47114 
1 _46960 
1_46792 
-1 49332 
-1_50452 
-1 48651 
-1 49846 
1 65547 
1 _66286 
1 _6ssos 
energy 
-0_18530 
-0_17759 
-0_16131 
-0_15310 
_O_470S8 
-0_46362 
-0_45031 
_O _ 44286 
-0_96927 
-0_96408 
-0_98063 
-0_97519 
1_25562 
-1_24816 
-1_28022 
loudness 
-0_07473 
-0_07518 
-0_08216 
-0_08228 
-0_07762 
-0_07810 
-0_08398 
-0_08413 
-0_07478 
-0_07509 
-0_07808 
-0_07842 
-0_09752 
-0_09753 
-0_09958 
-0_09958 
acousticness 
0_33659 
0_33389 
0_37389 
0_37010 
0_34397 
0_34120 
0_37630 
0_37231 
0_33726 
0_33267 
0_35062 
0_34585 
liveness 
0_08464 
0_07499 
0_08822 
0_07957 
0_05876 
0_06221 
0_02043 
0_02495 
0_14339 
0_15395 
0_11750 
0_12783 
valence 
0_83821 
0_83879 
0_80890 
0_81051 
0_83089 
0_83144 
0_80556 
0_80723 
0_92113 
0_92107 
0_91484 
0_91476 
0_85984 
0_85954 
0_85668 
0_85632 
0_81107 
0_81179 
0_79596 
0_79655 
0_89414 
0_89380 
0_88134 
duration 
0_00000366 
0_00000365 
0_00000364 
0_00000364 
0.00000366 
0_00000365 
0_00000365 
0.00000364 
0_00000360 
0_00000360 
0_00000360 
0_00000360 
0_00000355 
0_00000355 
0_00000355 
o_oooooass 
0_00000361 
o. 00000360 
0_00000361 
0_00000360 
0_00000356 
0_00000355 
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 value and were individually the “best” model among all models of their respective sizes according to their and Mallow’s Cp criterion. It appears from both Mallow’s Cp and 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 
2 
3 
4 
MODEL 
MODELI 
MODEL2 
MODEL3 
MODEL4 
TYPE 
PAR MS 
PAR MS 
PAR MS 
PAR MS 
DEPVAR 
release 
release 
release 
release 
trans 
trans 
trans 
trans 
RMSE 
1 _43097 
1 _43071 
1 _43069 
PRESS 
88262.76 
88237.15 
88230.11 
88231.35 
Intercept 
3.57289 
3.68216 
3.76123 
3.73940 
srank 
_006281347 
_006280084 
_ 006291036 
_006288448 
speech iness 
-1_92865 
1 _ 92285 
-1_90692 
-1_90790 
danceability 
-1.33101 
1_36377 
loudness 
-0.084126 
-0.082276 
-0.076176 
-0.074732 
acousticness 
0.37231 
0.37010 
0.33389 
0.33669 
valence 
0.80723 
0.81051 
0.83879 
0.83821 
duration 
_ 000003643 
_000003639 
_ 000003649 
_000003666 
days_ 
since 
release 
trans 
days_ 
days 
days_ 
days_ 
since 
since 
since 
since 
instrumentalness 
0.49068 
0.62048 
0.51683 
energy 
-0.17769 
-m 18630 
liveness 
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 , 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 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*

Fit Diagnostics for days _ 
since 
release 
00 
Predicted Value 
Quantile 
Residual 
trans 
0 000 
0 008 
0 006 
0 004 
0 002 
0 000 
10 
00 
Predicted Value 
8 
Predicted Value 
0 004 
0 008 
Leverage 
20000 
40000 
Observation 
Flt—Mean 
Residual 
Observations 
Parameters 
Error DF 
R-Square 
Adj R-Square 
43087 
43077 
2 0469 
on 295 
on 294 
OA 08 00 OA 08 
Proportion Less 

*Figure 11: Partial Residual Plots for Predictors*

Partial Plots for days_since_release_trans 
Intercept 
-02 00 
instrumentalness 
srank 
danceability 
-02 00 
Partial Regressor Residual 
speechiness 
-02 
00 
02 04 
energy 
06 
08 
00 
02 
04 
06 08 
-04 -02 00 02 
04 06 

Partial Plots for days_since_release_trans 
valence 
-200000 
loudness 
duration 
200000 
acousticness 
Partial Regressor Residual 

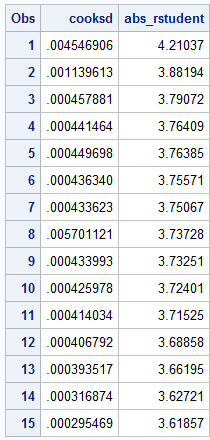
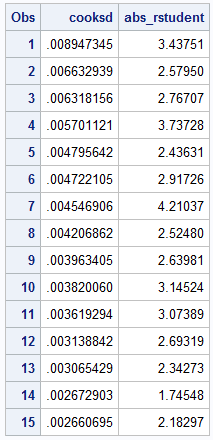
Looking at the Histogram and QQ-plot of the residuals of our model it appears that the data is approximately normally distributed. However, we can see from the plot of the residuals vs the predicted values that there is a clear linear trend in our residuals. This is due to our response variable taking 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).*

It can be seen from the above sorted tables that there are no cases with Cook's Distance > .9341963 and there are no cases with |Studentized Deleted Residual| > 4.86297 . 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.

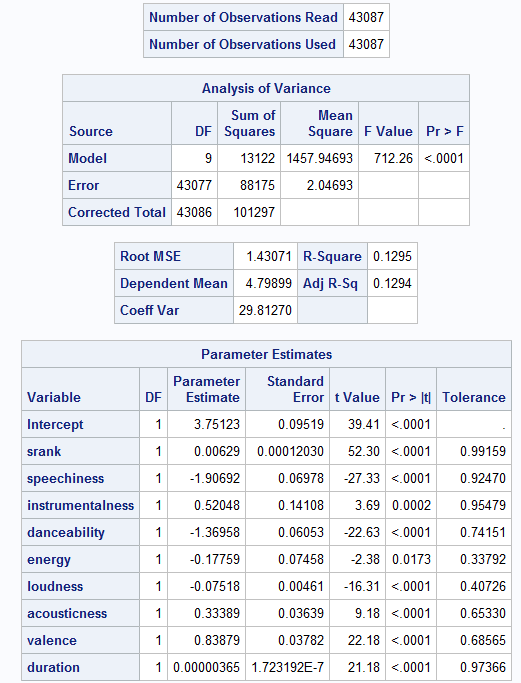
*Figure 14: Table of Calculated VIF Values for Selected Model*

Parameter Estimates 
Variable 
Intercept 
srank 
speechiness 
i nstrumentalness 
danceability 
energy 
loudness 
acousticness 
valence 
duration 
Parameter 
DF 
Estimate 
3.75123 
OT0629 
-1.90692 
0_52048 
-1.36958 
-0.17759 
-om518 
0,33389 
0_83879 
1 0.00000365 
Standard 
Error t Value 
0.09519 
0.00012030 
0.06978 
o. 14108 
0_06053 
0.07458 
0_03639 
0_03782 
1.723192E-7 
39.41 
52.30 
-27.33 
3.69 
-22_63 
-2.38 
-16.31 
9.18 
22_18 
21.18 
<.0001 
0001 
<.0001 
0.0002 
0.0173 
<.0001 
<.0001 
<.0001 
Tolerance 
0.99159 
0.92470 
0.95479 
0_74151 
0.33792 
0.40726 
0.65330 
0_68565 
0.97366 

From the table above we are able to calculate the VIF trivially by noting that . Using this conversion we see that the VIF for each predictor is less than 10 and the mean VIF, , for our predictors is 1.54631. Since no single predictor has a VIF greater than 10 and 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*



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

,

Where:

, , , ,, , , , ,

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 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 days. Additionally, note that the predicted intercept of 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 days per unit increase in speechiness.

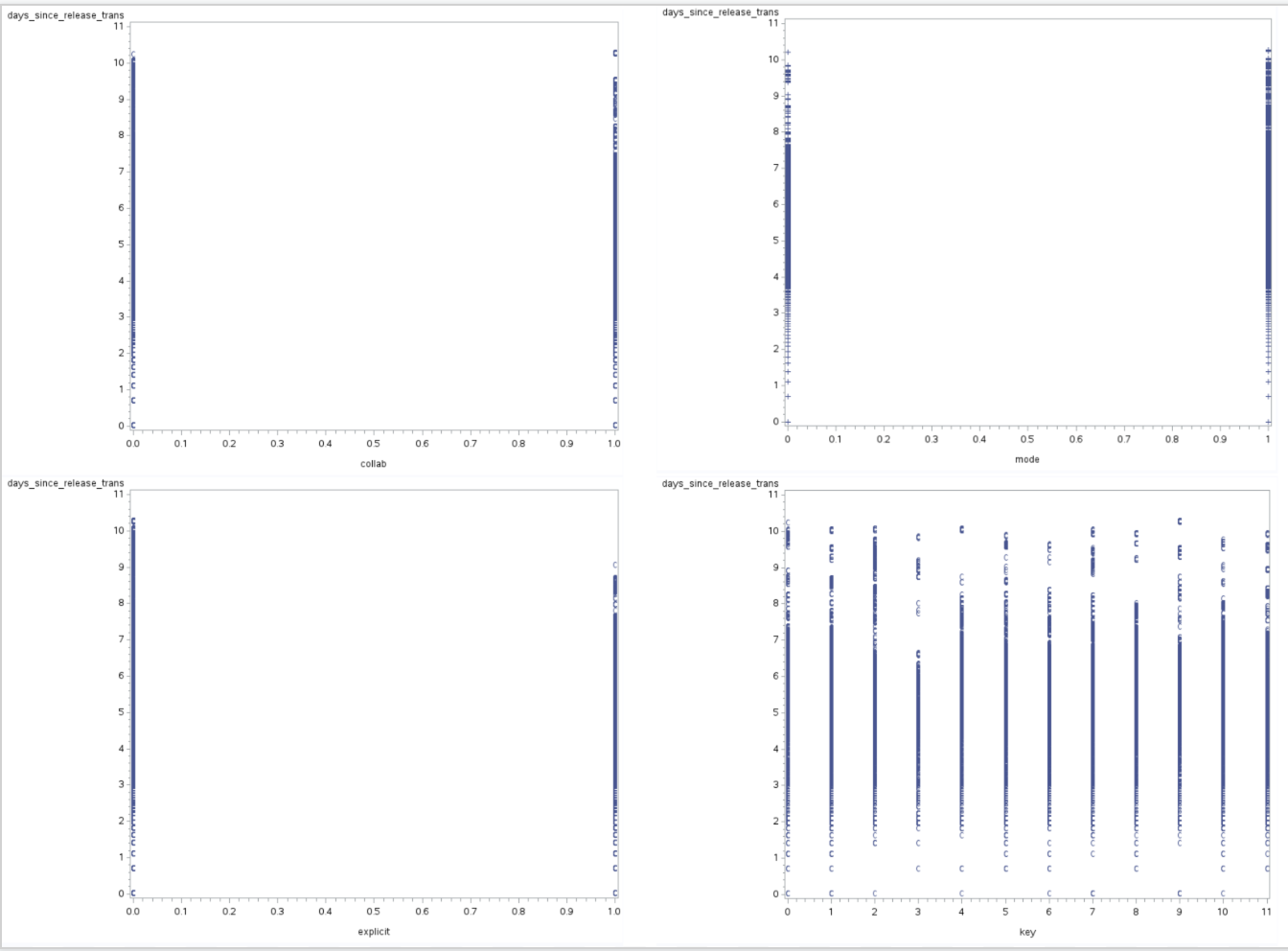
**Factor Analysis:**

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.

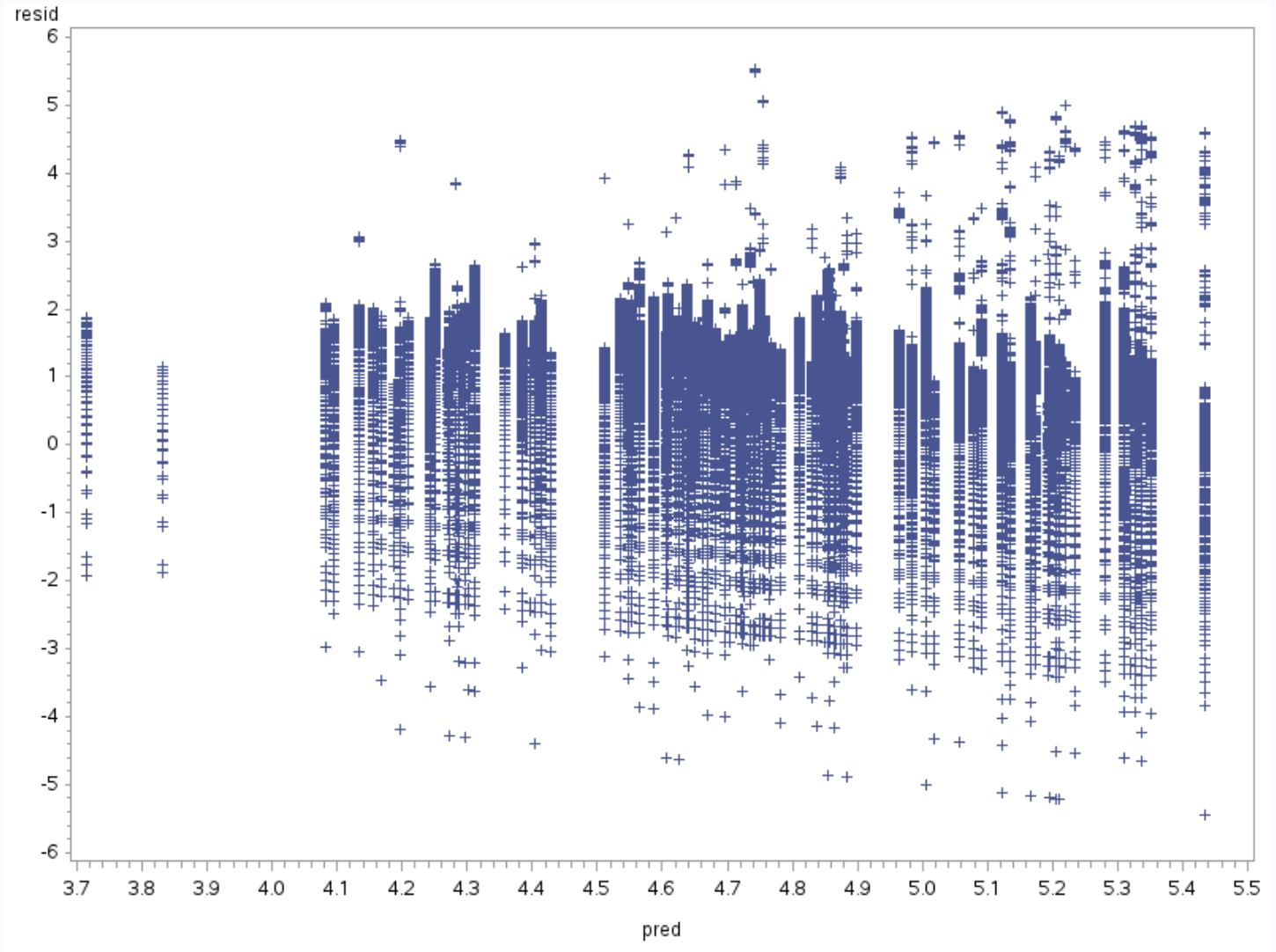
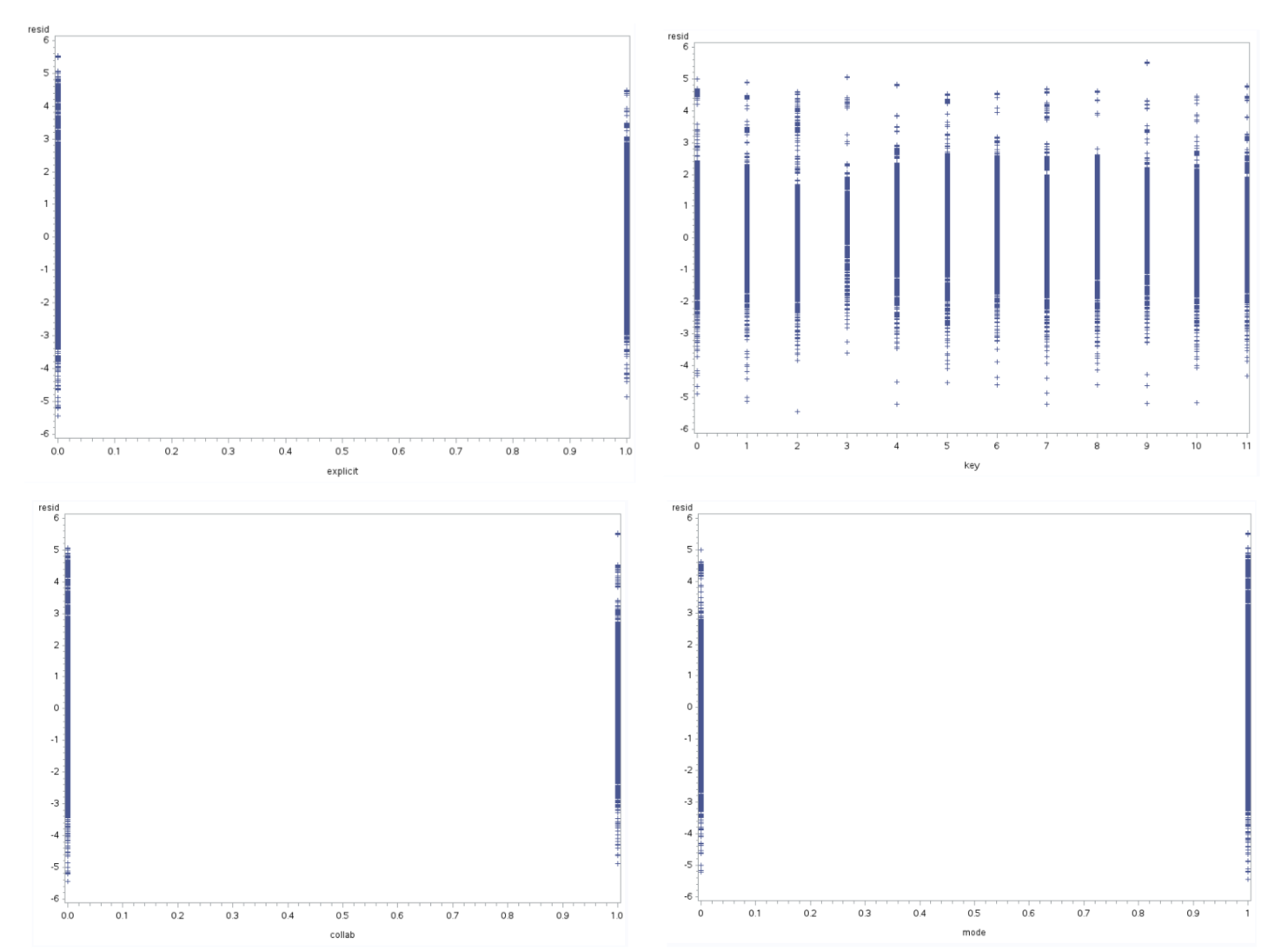
We start by checking the Normality Assumption of factor variables.

*Figure 16: Scatter plots of factor variables*

**

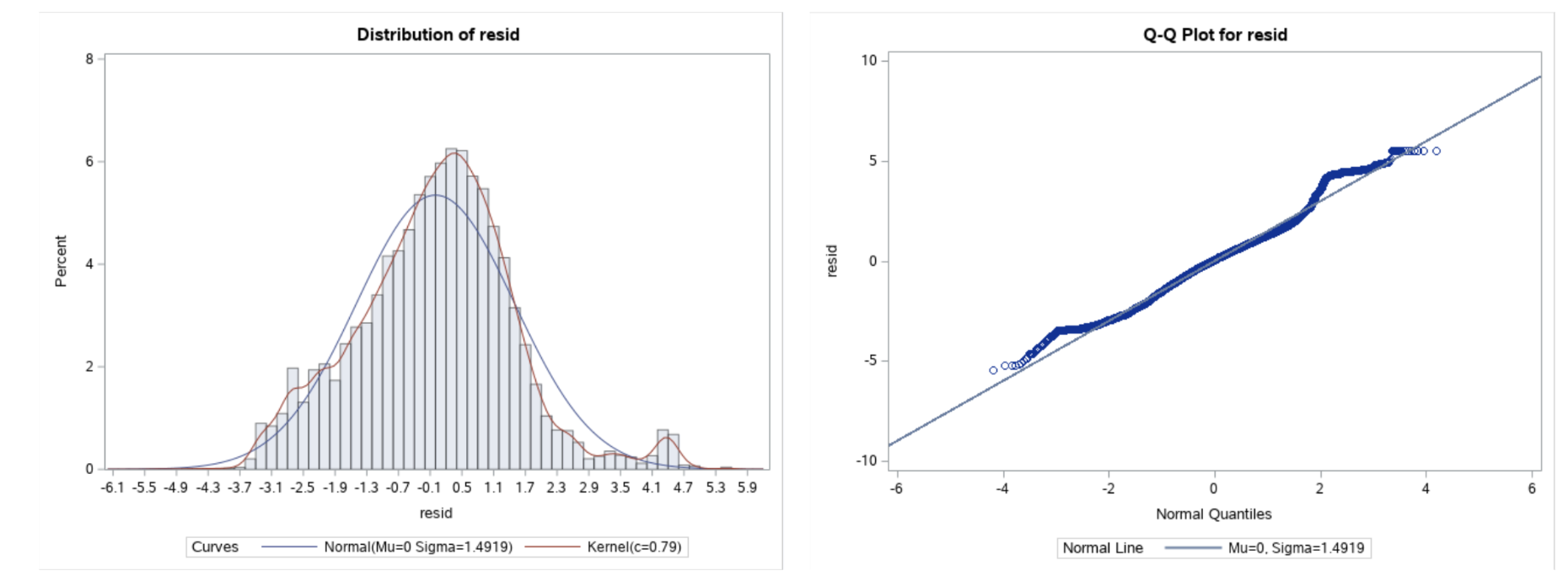
There are few deviations among scatter plots of all factor variables, meaning there are few outliers and independence is fulfilled.

*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.

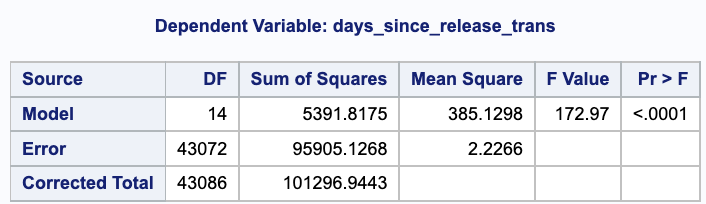
*Figure 18: Histogram and QQ plot for residuals*

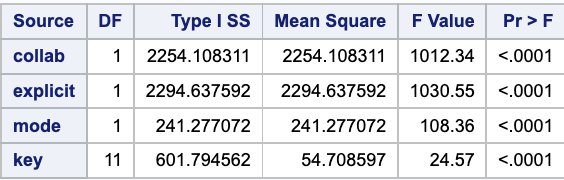
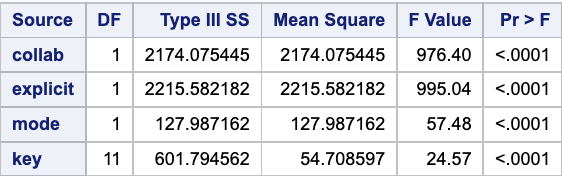


In both plots above, while there are still some slight deviations from normality near the tails of the data, they are negligible. Thus, it appears that the Normality Assumption of the Independent variable is now met. Overall, factor variables meet the normality and aptness for the model.

Now, we move on the ANOVA with factor variables.

*Figure 19: ANOVA table of categorical variables*

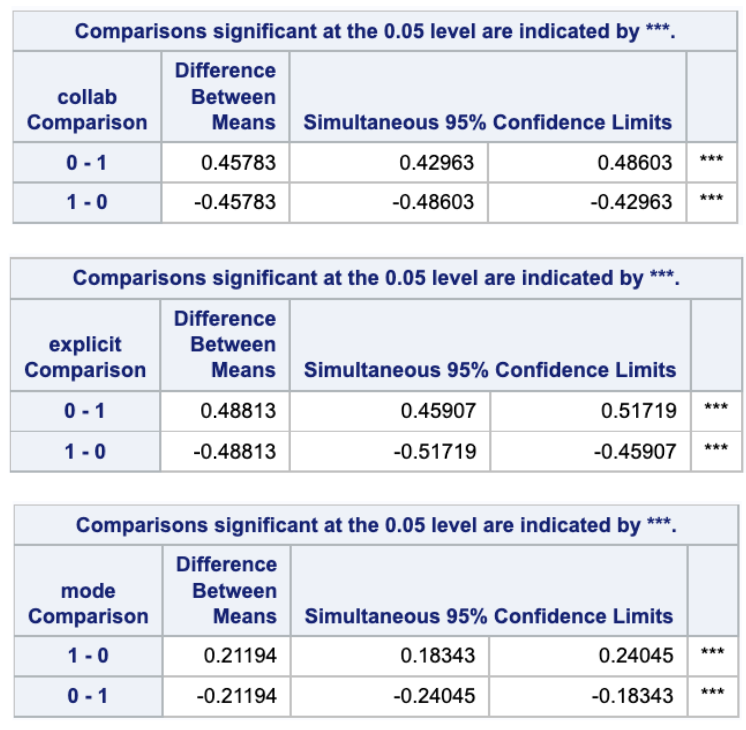


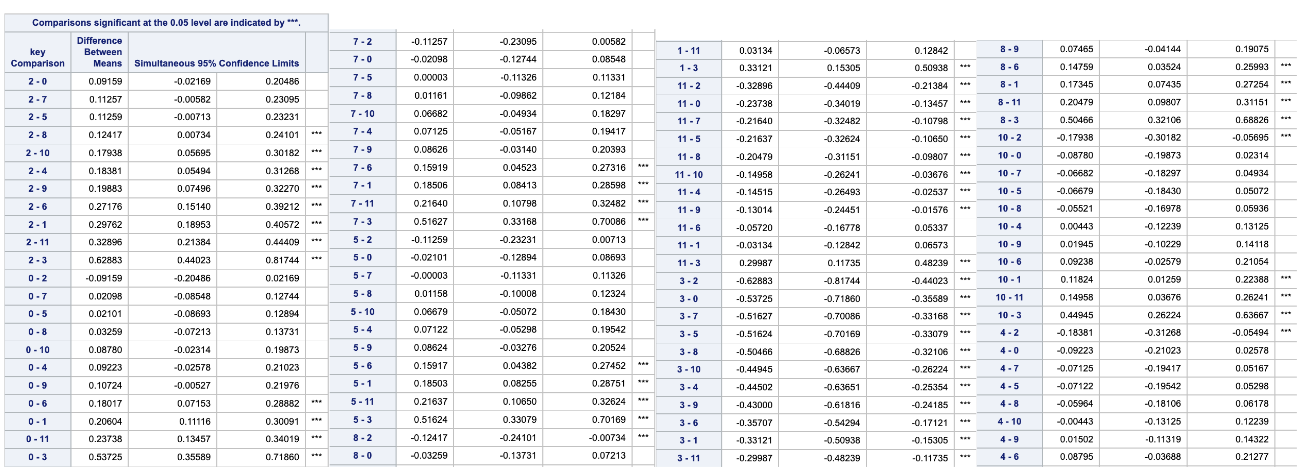


As we see the ANOVA table and TYPE {I | III} table above, we can see that overall, the model with categorical variables is significant at F-test; F(14, 43072, 0.95). The significant value is < .0001. Type {I | III} are all significant as well, meaning they have some sort of information useful in explaining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





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.

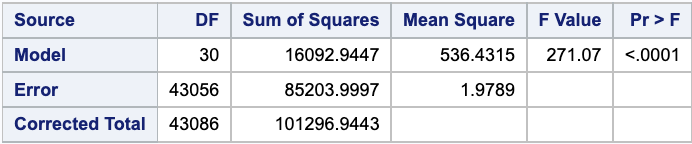
**Interaction model:**

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*

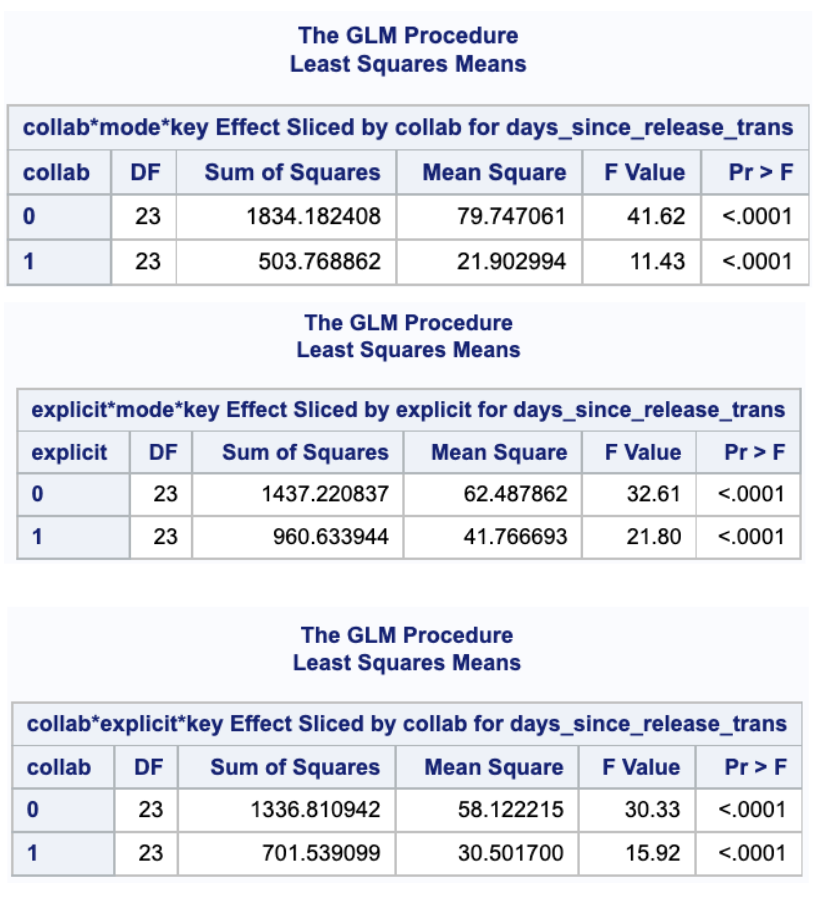




We can see the overall F-test statistics (271.07) is large enough to be significant. Yet, as we check the simultaneous F-test table, there are some insignificant variables. Some of the quantitative variables such as “instrumentalness,” “tempo,” were significant when the model only contained quantitative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

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

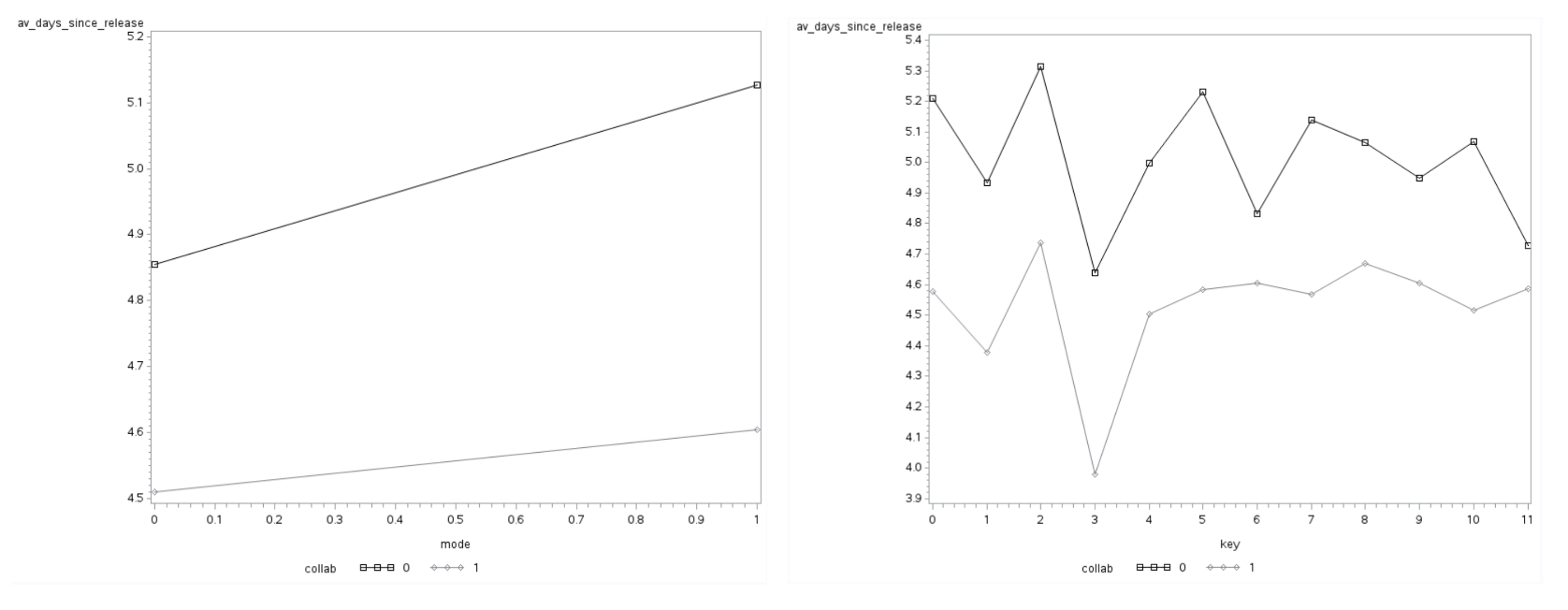
*Figure 22: Expected Mean comparison*



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;