UNIVERSIDAD
DE GRANADA

Analysis of Song Popularity Based on Spotify and YouTube Data

Natalia De Vega Suárez Bárcena
Alba Diez Santos
Lidia Nievas Dueñas
Gonzalo Romero Gallego

Granada, Thursday 5th June, 2025

Contents

1 Introduction

Our first task was to find a dataset to analyse, in our case we have chosen one based on songs of various artist in the world, since all the members of the group like different type of music and love discovering new songs. Our objective is to establish a model that let us define which variables could explain the popularity of the songs. Thanks to the contents of the course we are able to make this study.

We are going to start by describing the dataset in the following section, so we can understand the results we will obtain. In this part we will also study if there is possible linear relationship between the response (popularity) and the covariates and if there is possible co-linearity among covariates.

Then we will provide the mathematical formulation of the chosen model, explaining everything of it and including the assumptions to be made.

The next step will be to make a statistical analysis, that is describing the goodness of the fit, making inference related to the model and possible simplification of it.

To finalize, we will make conclusions of the results we have obtained in the previous section. The idea is to summarize the analysis performed and highlight the most relevant results. We will also write about the limitations of our analysis and suggest possible improvements or extensions.

To do all of this, we rely on information found in some books about R [1],specifically about about regression [2] and other statistical methods [3].

2 The dataset

As we have mentioned, our dataset is about music. We got it from Kaggle [4] (one of the link provided for the coursework to download datasets). It contains songs of different artists in the world, and for each one is present:
- Several Spotify statistic, including the amount of streams.
- Number of views and likes of the official music video on YouTube.

Firstly we are going to visualize the variables:

```
spotyout<- read.csv("Spotify_Youtube.csv", as.is=TRUE)
```

The dataset contains 26 variables for each song collected from Spotify and YouTube. These include:

Track, Artist, Album, Album type: Basic song and album info.

Url spotify, Uri: Links to the song/artist on Spotify.

Danceability, Energy, Valence, Tempo: Musical features (e.g., mood, pace, rhythm).

Key, Loudness, Speechiness, Acousticness, Instrumentalness, Liveness: Audio attributes describing sound quality, vocals, live performance, etc.

Duration ms: Track length in milliseconds.

Stream: Number of Spotify streams.

Url youtube, Title, Channel: YouTube video info.

Views, Likes, Comments: Video engagement metrics.

Description: Video description on YouTube.

Licensed: Whether the video is officially licensed.

Official video: Indicates if it's the official video of the song.

As we don't have a parameter explaining the popularity, for us it makes sense to establish that the popularity of a song is based on the likes it has, so we are going to reformulate the question we want to answer: how do the variables involved in the dataset contribute on the likes of a song?.

To answer this question, we begin by working only with singles. We made this decision because it makes more sense to compare the number of views or likes for singles, as in the case of albums, one track may become significantly more popular than the rest. This can lead listeners to focus on just that one song and overlook the others. In contrast, when an artist releases a single, all the attention is typically concentrated on that individual track.

We are also going to remove the variable X, which is simply a song identifier. Since it does not contain any meaningful information related to the number of views, it is not useful for explaining or predicting the response variable.

```
spotyout<- subset(spotyout, Album_type=="single")
spotyout<- subset(spotyout, select = -c(Album_type,X,Key) )
```

After visualizing the plot that displays the relationships between all variables (we do not include it here because it is too large due to the presence of 27 variables), we can make the following observations:
- There may be a potential linear relationship between the response variable (Likes) and some of the covariates (which we will later select).
- There appears to be some multicollinearity among the covariates, meaning that certain variables are highly correlated and could be redundant. In such cases, we may consider removing them.

From the plot, we observe that indeed Likes might show a linear dependence on some of the other variables. This motivates our decision to use a multiple linear regression model. As a first step, we restrict the analysis to numeric variables only.

```
numerics <- spotyout[sapply(spotyout, is.numeric)]
```

Then, we can remove the variables which are highly correlated between them. Let us study this by calculating the pairwise Pearson coefficients, and focus on couples with absolute value higher than 0.5:

```
#Pearson coefficient between each variable
cor_mat <- cor(numerics, use = "complete.obs")
# We search the ones near to 1
indices <- which(abs(cor_mat) > 0.5 & abs(cor_mat) < 1 &
                   upper.tri(cor_mat), arr.ind = TRUE)
for (i in 1:nrow(indices)) {
 var1 <- rownames(cor_mat)[indices[i, 1]]
 var2 <- colnames(cor_mat)[indices[i, 2]]
 r <- cor_mat[indices[i, 1], indices[i, 2]]
  cat(sprintf("%s ~ %s = %.2f\n", var1, var2, r))
 }
```

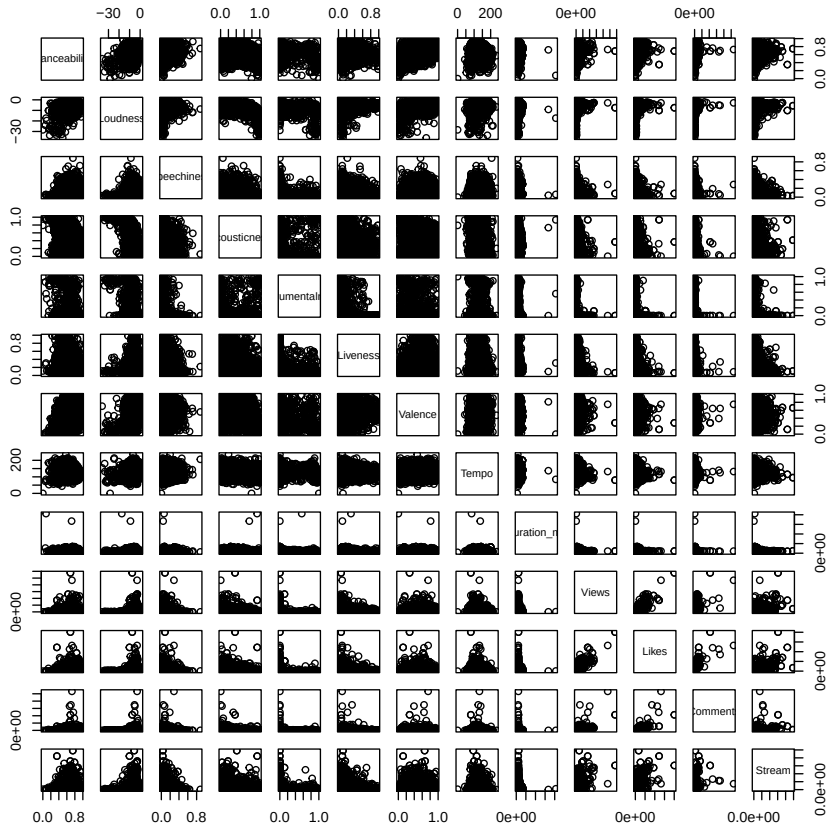
```
## Energy ~ Loudness = 0.73
## Energy ~ Acousticness = -0.60
## Views ~ Likes = 0.89
## Views ~ Comments = 0.64
## Likes ~ Comments = 0.75
## Views ~ Stream = 0.53
## Likes ~ Stream = 0.57
```

Energy is highly correlated with Loudness, and a little less with Acousticness, too. This lead us to decide on removing it. Thus, the dataset would be :

```
dataset<- subset(numerics, select = -Energy)
```

Now we can visualize the scatter plot :

```
plot(dataset)
```



3 The model

We proceed to provide the mathematical formulation of the choosen model. As we announced, it will be a multilinear regression model of the type :

$$Y_i = \beta_0 + x_{i1}\beta_1 + x_{i2}\beta_2 + \cdots + x_{ik}\beta_k + \varepsilon_i, \quad i = 1, \ldots, n \tag{1}$$

where $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ are independent random errors.
In our case, the response variable is **Likes** , and the $k = 12$ covariates are : *Danceability*, *Loudness*, *Speechiness*, *Acousticness*, *Instrumentalness*, *Liveness*, *Valence*, *Tempo*, *Duration_ms*, *Views*, *Comments*, *Stream*.

```
lmlikes <- lm(Likes~Danceability+Loudness+ Speechiness+Acousticness+
                Instrumentalness+Liveness+Valence+Tempo+Duration_ms+
                Views+Comments+Stream ,dataset)
```

Likes $= 8.889 \times 10^4 + 1.599 \times 10^5 x_1 + 5.999 \times 10^3 x_2 + 1.913 \times 10^5 x_3 + 1.423 \times 10^5 x_4 - 7.322 \times 10^4 x_5 - 4.374 \times 10^4 x_6 - 2.133 \times 10^5 x_7 - 1.880 \times 10^2 x_8 + 1.499 \times 10^{-2} x_9 + 4.494 \times 10^3 x_{10} + 4.306 \times x_{11} + 1.479 \times 10^3 x_{12}$

with $x_1, \ldots, x_{12}$ the covariates we have previously mentioned respectively .

4 Statistical analysis

Above we have fitted the model to the data. However we need to measure how well it actually describes the information, this is, the goodness of fit.

```
summary(lmlikes)
```

```
##
## Call:
## lm(formula = Likes ~ Danceability + Loudness + Speechiness +
##     Acousticness + Instrumentalness + Liveness + Valence + Tempo +
##     Duration_ms + Views + Comments + Stream, data = dataset)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
```

```
## -18195199   -138802    -53586     32258  10973812
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       8.889e+04  8.347e+04   1.065  0.28699
## Danceability      1.599e+05  7.965e+04   2.008  0.04474 *
## Loudness          5.999e+03  3.760e+03   1.596  0.11066
## Speechiness       1.913e+05  1.057e+05   1.809  0.07050 .
## Acousticness      1.423e+05  4.425e+04   3.216  0.00131 **
## Instrumentalness -7.322e+04  6.729e+04  -1.088  0.27661
## Liveness         -4.374e+04  6.399e+04  -0.684  0.49429
## Valence          -2.133e+05  4.920e+04  -4.337 1.48e-05 ***
## Tempo            -1.880e+02  3.649e+02  -0.515  0.60642
## Duration_ms       1.499e-02  1.009e-01   0.149  0.88183
## Views             4.494e-03  5.885e-05  76.363  < 2e-16 ***
## Comments          4.306e+00  9.366e-02  45.975  < 2e-16 ***
## Stream            1.479e-03  5.838e-05  25.337  < 2e-16 ***
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 673500 on 4676 degrees of freedom
##   (315 observations deleted due to missingness)
## Multiple R-squared:  0.8702, Adjusted R-squared:  0.8699
## F-statistic:  2613 on 12 and 4676 DF,  p-value: < 2.2e-16
```

The adjusted percentage (*Adjusted R-squared*) of variance of the response explained by the covariates through the model is 86.99%. As we now, the actual value (*Multiple R-squared*) may be quite inflated, but in our case they are practically the same. This all means it is really a good model. Apart from the goodness of fit, we are showing some results related to the inference techniques performed on the model (after its fitting).

4.1 Inference on the model

4.1.1 Regression test: overall significance of the model

As a first step, we check the answers given on the reasonableness of the model itself (even though we already have some ideas). The regression test formulates this null hypothesis: $H_0 : \beta_1 = \beta_2 = .. = \beta_{12} = 0$ that is, the 12 covariates do not have any effect. We look at the value of the test statistic F = 2613, and the corresponding p-value, that is approximately 0. The conclusion is to reject the null hypothesis, therefore the set of covariates we are considering can be used to explain the response (they are not all null).

4.1.2 Testing the effect of just one covariate: individual significance

Now we want to know if we can drop just one covariate, this means to get rid of some of the variables which have less influence on the response. This will be important if we want to reduce the dimensions of the model. To deduce it, we set the following hypothesis test: $H_0 : \beta_j = 0 \quad vs \quad H_1 : \beta_j \neq 0$. Considering a significance level of 0.05 we would have to reject the null hypothesis for all the covariates except for Loudness, Speecheness, Instrumentalness, Liveness, Tempo and Duration. We are not taking any final decission, but this suggests us that these variables may not be needed to explain the response.

4.1.3 Testing the significance of the intercept β_0

When defining the model we have included an intercept β_0 and we want to check if it is correct to include this parameter. We contrast now: $H_0 : \beta_0 = 0 \quad vs \quad H_1 : \beta_0 \neq 0$. In this case, considering again a significance level of 0.05, we can not reject the null hypothesis, since the p-value (0.28699) is not lower. This information leads us to think of removing it.

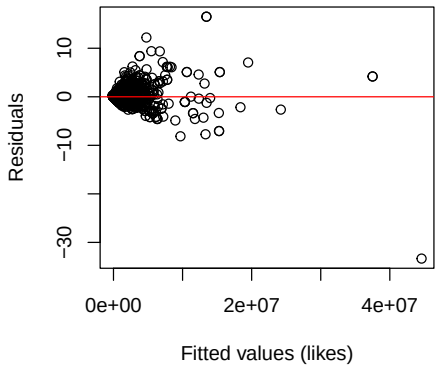
4.2 Regression diagnostics: residual analysis

We have concluded that the model could be simplified because some covariates seem to have very little effect on the popularity (Likes). However, the estimation and inference results supporting those conclusions rely on several assumptions that we have not evaluated yet.

4.2.1 Constant variance assumption

We have assumed that errors are normally distributed, are independent and have equal (constant) variance.

```
residualslikes <- rstandard(lmlikes)
fittedvalueslikes <- fitted.values(lmlikes)
plot(fittedvalueslikes, residualslikes, xlab="Fitted values (likes)",
     ylab="Residuals")
abline(h=0, col = 'red')
```



The residuals appear to be randomly scattered around zero, suggesting that the assumption of constant variance is reasonably satisfied. Residuals that lie far from the zero line or stand out from the overall pattern may indicate the presence of outliers.

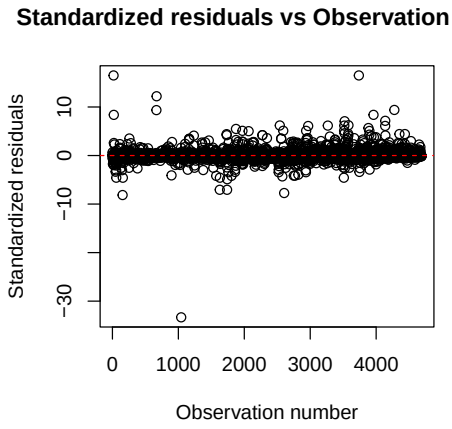
4.2.2 Independence assumption

We have also assumed that the conditional mean of the response is a linear function of the covariates.

```
library(lmtest)
plot(1:length(residualslikes), residualslikes,
     xlab = "Observation number",
     ylab = "Standardized residuals",
     main = "Standardized residuals vs Observation index")
abline(h = 0, col = "red", lty = 2)
```

```
dwtest(lmlikes)

##
##  Durbin-Watson test
##
## data:  lmlikes
## DW = 1.7038, p-value < 2.2e-16
## alternative hypothesis: true autocorrelation is greater than 0
```



In the plot of standardized residuals versus the observation index, most residuals appear randomly scattered around zero, forming a horizontal band centered at zero. This supports the assumption of constant variance. However, a few points lie far from this band, indicating the presence of potential outliers. The Durbin-Watson statistic is 1.70 with a very small p-value, indicating significant positive autocorrelation in the residuals.

4.2.3 Normality assumption

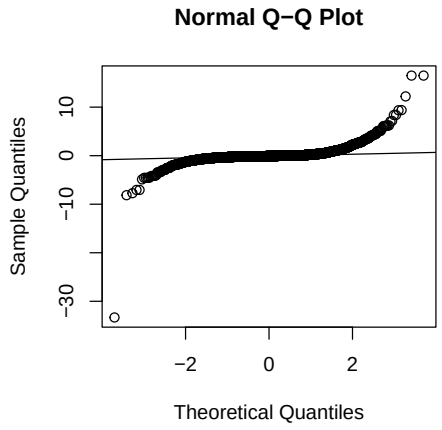
We want to verify the normality of the residuals.

```
ks.test(residualslikes ,"pnorm")

##
##  Asymptotic one-sample Kolmogorov-Smirnov test
##
## data:  residualslikes
```

```
## D = 0.26305, p-value < 2.2e-16
## alternative hypothesis: two-sided

qqnorm(residualslikes)
qqline(residualslikes)
```



The Kolmogorov-Smirnov test suggests that the residuals are not normally distributed (p-value less than 2.2e-16), and the Q-Q plot shows departures from the reference line, especially in the tails. This indicates a violation of the normality assumption.

4.2.4 Unusual observation

As we have already said, the model is not bad but it does not work for the outliers. So our suggestion is to repeat the model without the outliers :

```
outliers<-which(abs(residualslikes) > 3);
dataset2 <- dataset[-c(outliers), ]
lmlikes2 <- lm(Likes~Danceability+Loudness+ Speechiness
               +Acousticness+Instrumentalness+Liveness+Valence+Tempo+
                 Duration_ms + Views+Comments+Stream, dataset2)
summary(lmlikes2)

##
## Call:
## lm(formula = Likes ~ Danceability + Loudness + Speechiness +
```

```
##     Acousticness + Instrumentalness + Liveness + Valence + Tempo +
##     Duration_ms + Views + Comments + Stream, data = dataset2)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -18085631   -136043    -56756     29763  11082927
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       7.565e+04  8.061e+04   0.938  0.34805
## Danceability      1.901e+05  7.696e+04   2.469  0.01357 *
## Loudness          6.129e+03  3.632e+03   1.687  0.09158 .
## Speechiness       2.025e+05  1.018e+05   1.989  0.04671 *
## Acousticness      1.206e+05  4.272e+04   2.823  0.00477 **
## Instrumentalness -6.180e+04  6.499e+04  -0.951  0.34172
## Liveness         -3.126e+04  6.163e+04  -0.507  0.61199
## Valence          -2.131e+05  4.750e+04  -4.487  7.4e-06 ***
## Tempo            -2.246e+02  3.518e+02  -0.638  0.52327
## Duration_ms       3.386e-02  9.682e-02   0.350  0.72652
## Views             4.551e-03  5.798e-05  78.487  < 2e-16 ***
## Comments          4.238e+00  9.693e-02  43.721  < 2e-16 ***
## Stream            1.410e-03  5.746e-05  24.540  < 2e-16 ***
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 644700 on 4594 degrees of freedom
##   (314 observations deleted due to missingness)
## Multiple R-squared:  0.8732, Adjusted R-squared:  0.8729
## F-statistic:  2637 on 12 and 4594 DF,  p-value: < 2.2e-16
```

We get a higher adjusted and multiple R-squared, which aligns with our thinking.

5 Conclusions

Nowadays, many years after the invention and globalization of the Internet, music is more accessible than ever for everyone. Very often, as with any other type of art, people identify and gets attached with some particular band, singer or gender. Nevertheless, there will always be certain songs that happen to seduce almost every kind of ear, and possibly many persons around you know them.

But how do they do it? Where does this success come from? Is it measurable? If so, on what does it depend? Are these factors measurable too? Some of these were the questions we made to ourselves at the beginning of this coursework, when we decided to focus it on music. In our try to solve these problems, we could not do other thing than using the statistics tools we have learned, and make conclusions under some confidence level (as always in inference).

At first, we looked for a good dataset, that adjusted to our needs and approach. The one we selected, presented on the first part, satisfied our requirements. It includes mainly numeric variables, some of them seem to be pretty related with global popularity of songs (such as likes, views or streams), and others are closer to purely music properties of the songs (which could be also relevant). Then we also preformed some modifications on the data, since we were going to work only with numeric values and singles (in order to balance the scenario).

After that, we continued evaluating our problem during section 2, and the path we chose was to fit a model which could explain the popularity, represented by the likes of the song, with the information on the rest of the variables (we could have also chosen views or streams). The next step was to decide on which kind of model we were going to use, and the observation of the scatter plots led us to try a multilinear one.

We fitted and showed the model (on part 3), but we needed to evaluate the accuracy of it. We performed this task by testing the data on many hypothesis about the model (throughout the last section). The results were successful but not perfect, in the way they showed that the choice was finely adjusted but some assumptions were not satisfied by the data. Finally, as a consequence of this information, we proposed how the model (by removing variables) and the data (by removing outliers) could be optimized to find even a better fit that fulfills the rest of the requirements for a multilinear model.

As we have shown, popularity of music can actually be measured and predicted. There may be other better models to study it, but a simple linear one is enough to get conclusions. Since music is everywhere nowadays, the industry that manages its distribution is growing day by day, resulting in songs actually getting closer to marketing than art. If streams or views can be bought (and they can) we have a formula for making a song popular (that can depend on other musical features such

as danceability or acousticness), and then there is actually a formula for making money (in form of songs).

The big record labels know this and are taking advantage of it (they surely update this research constantly). This is the most important reason why people today fight or argue about big pop stars not deserving their commercial success (that may not be so licit), while little artist can not get any recognition or fame, unless they follow that magic formula.

6 References

References

[1] Crawley. *The R Book, 2nd ed.* Wiley-Blackwell, London, 2012.

[2] Crawley. *Statistics: An Introduction Using R, 2nd ed.* Wiley, London, 2015.

[3] J.E. Gentle. *Computational Statistics. Statistics and Computing.* Springer, London, 2009.

[4] Marco Guarisco, Salvatore Rastelli and Marco Sallustio. Kaggle dataset: Spotify and Youtube. `https://www.kaggle.com/datasets/salvatorerastelli/spotify-and-youtube/data`, 2023.