THE MUSIC INDUSTRY: past and future directions

A case study of music formats over the last five decades and of streaming services in recent years

Alvise Dei Rossi, Elisa Tremolada

Context

The music industry, like many others, has changed completely in the last four decades due to **technological advancements**.

Record labels need to keep up with **ever-changing music formats** and corresponding **shifts in consumer preferences**.

That's where we come in!

Using **three different datasets** and **four different classes of statistical models**, we will give our clients and overview **of future industry directions**, **market leaders** and even **consumers' song choices**.

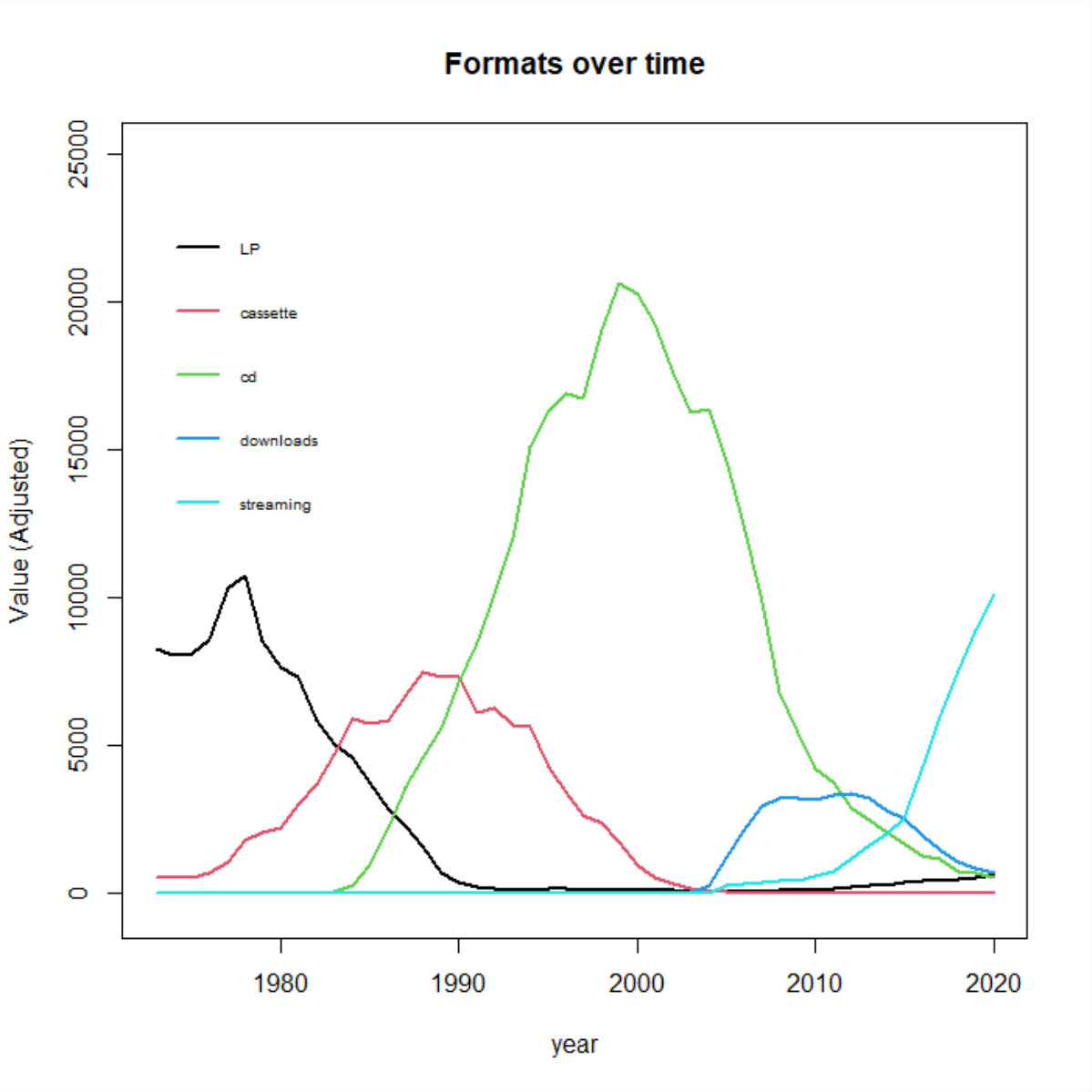
Research Questions

1. Which music format will dominate the market in the next 2 or 3 years?
   - Diffusion models, UCRCD models
2. Who is a potential market leader among the producers of this format? Will they keep growing in the coming years?
   - Simple linear regression, ARIMA
3. What makes a song popular among consumers who buy that particular format?
   - Multiple linear regression, GAM, GBM

Question 1:

Which music format will dominate the market in the next 2 or 3 years?

❖**EDA**: For this analysis, we worked on a dataset provided by the Recording Industry Association of America (RIAA) . This dataset contains **U.S. sales of recorded music from 1973 to 2020**, measured both in units sold/year and in yearly revenue. In the dataset, music formats are divided into 23 categories. We merged similar categories for simplicity*, obtaining **5 categories: LPs, cassettes, CDs, downloads and streaming services**. In order to avoid difficulties due to the different meaning of "unit" for the various formats, we use as a metric the **revenue (adjusted for inflation), measured in millions of dollars**.



Overview: diffusion of music formats up until 2020

We developed a diffusion model for each single format – both as a simple BASS model and with a SARIMAX refinement.

Approximate parameter values for the BASS models (m, p, q) can be observed in the table below.

Note that all values shall be multiplied by 1 million in order to get the real value in USD.

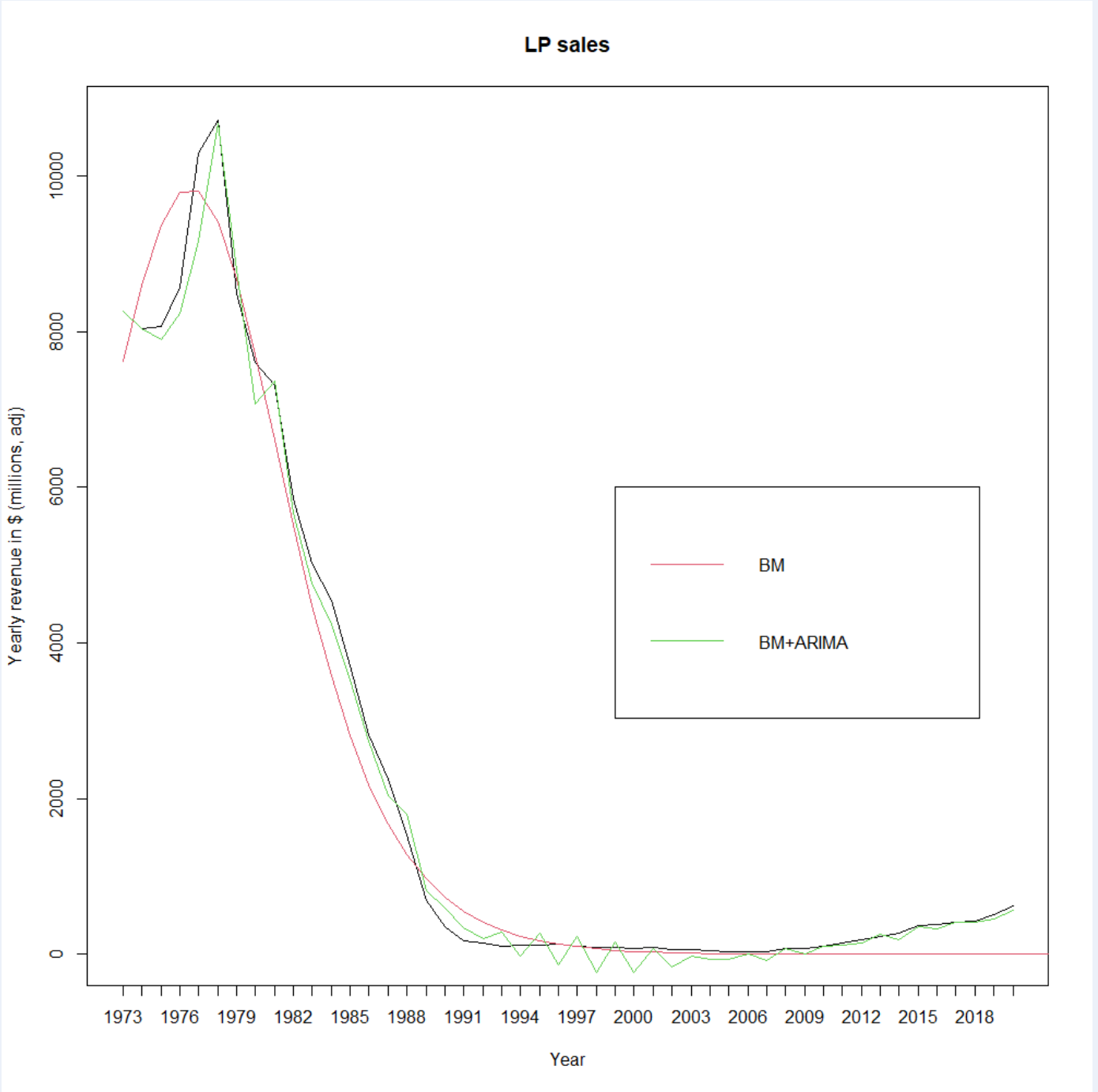
Category / Model parameters	m	P	q
LP	106362	0.0613	0.2318
Cassette	108629	0.0032	0.2704
CD	**320583**	0.0034	0.2514
Downloadable formats	**38452**	0.0233	**0.3246***
Streaming services	163694	0.0004	**0.3452***

* The **increase in the q parameter** with the appearance of downloadable formats might be due to the increased communication speed due to the beginning of the Internet era.

Diffusion of the LP

All parameters resulted significant for a simple BASS model.

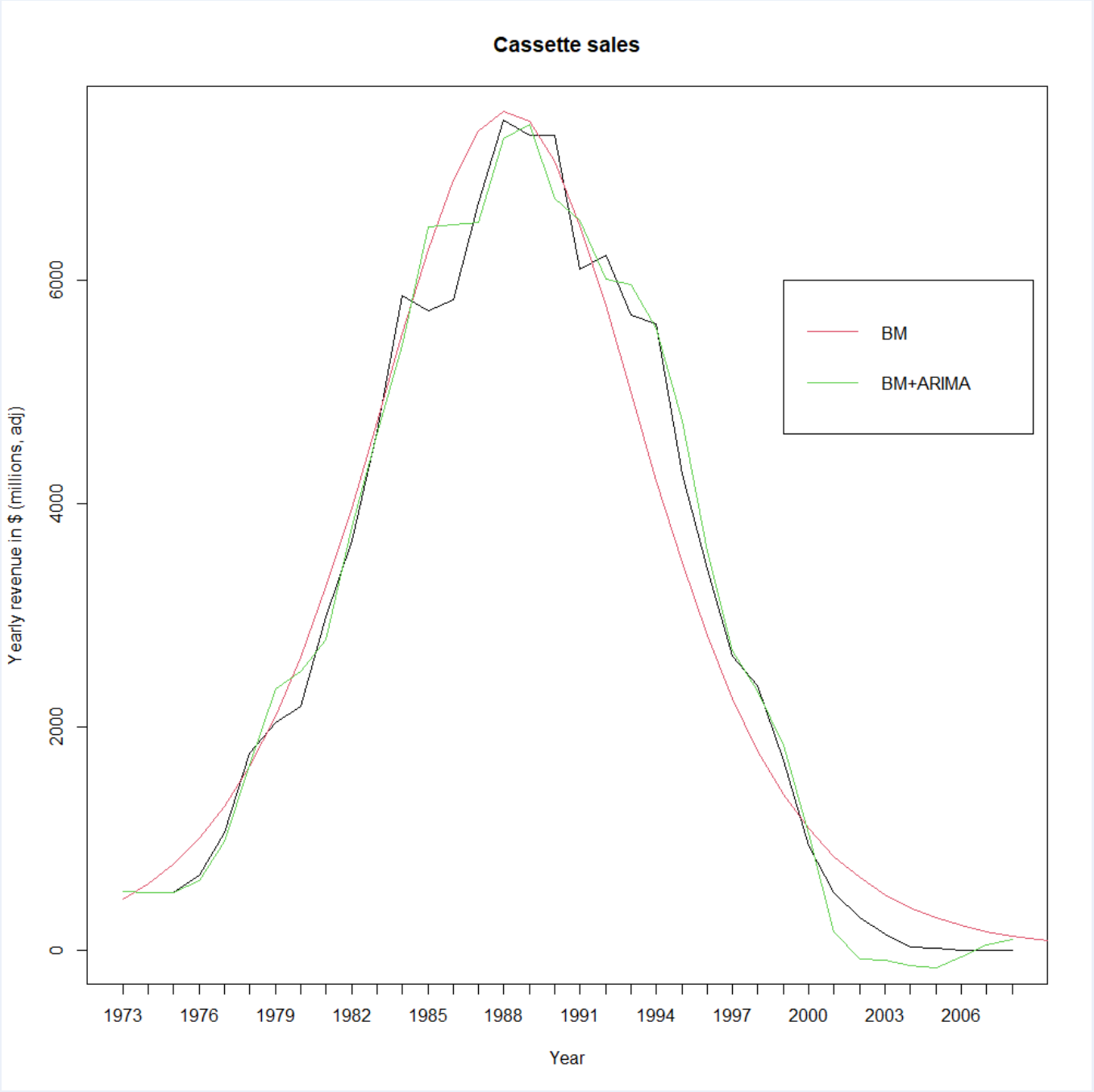
Category / Model parameters	m	P	q
LP	106362	0.0613	0.2318



Diffusion of the cassette

All parameters resulted significant for a simple BASS model.

Category / Model parameters	m	P	q
Cassette	108629	0.0032	0.2704

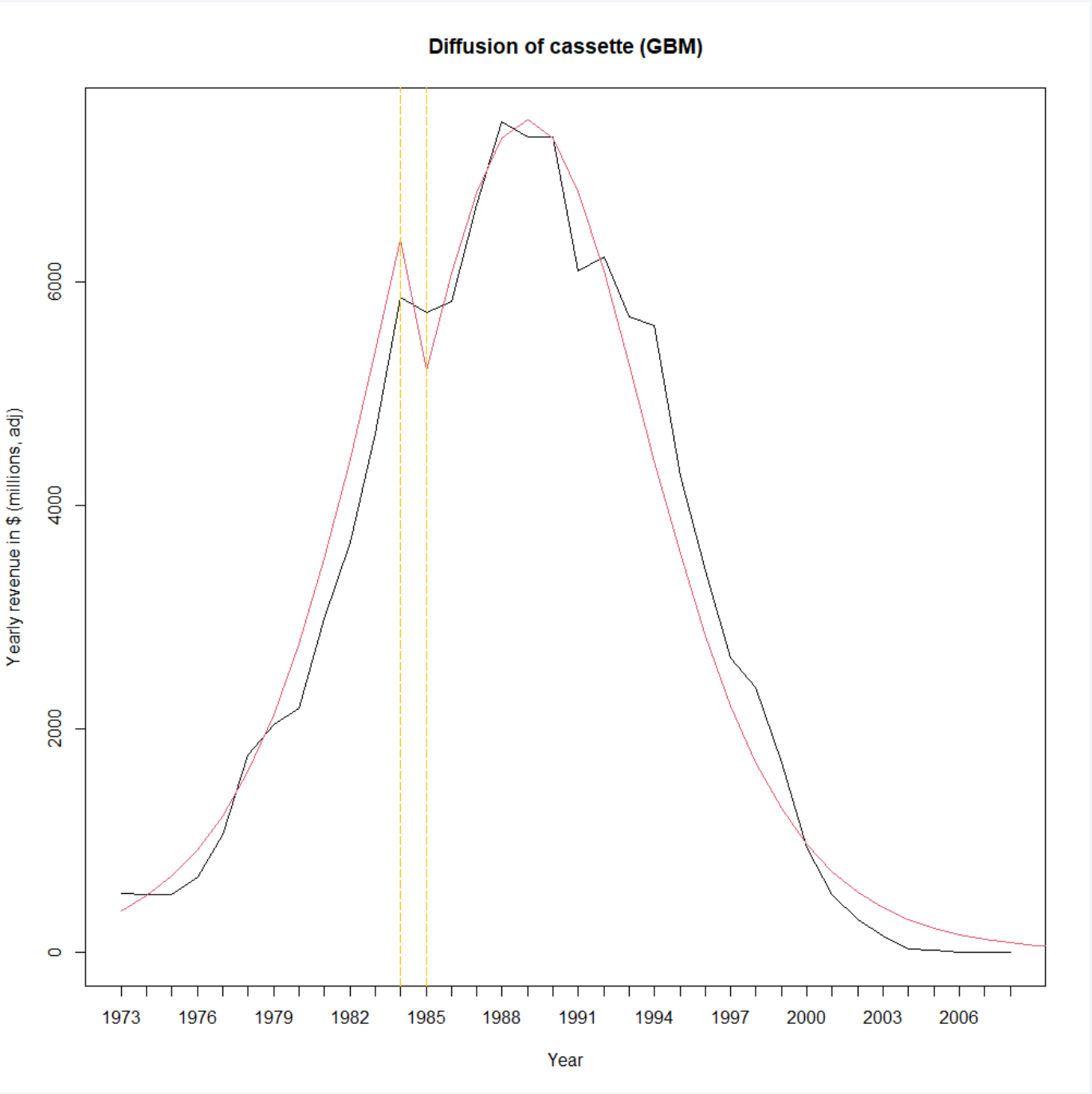


Cassette diffusion - GBM

Type of shock: Exponential

Note that all values shall be multiplied by 1 million in order to get the real value in USD.

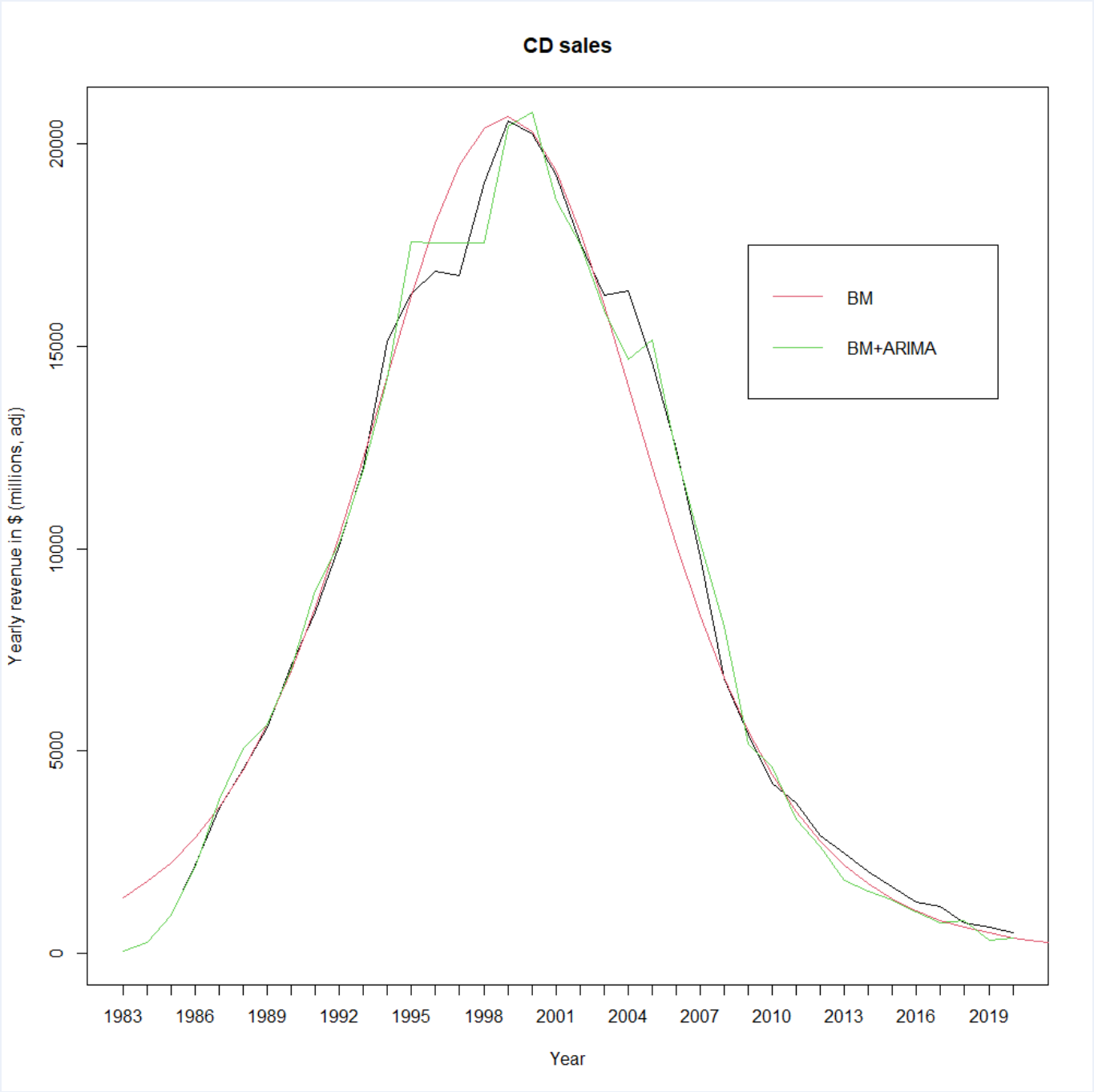
Parameters	Values
Adjusted R-squared	0.48
m	107990
p	0.0026
q	0.3066
a1	12
b1	- 0.2381
c1	- 0.3107



Diffusion of the CD

All parameters resulted significant for a simple BASS model.

Category / Model parameters	m	P	q
CD	320583	0.0034	0.2514

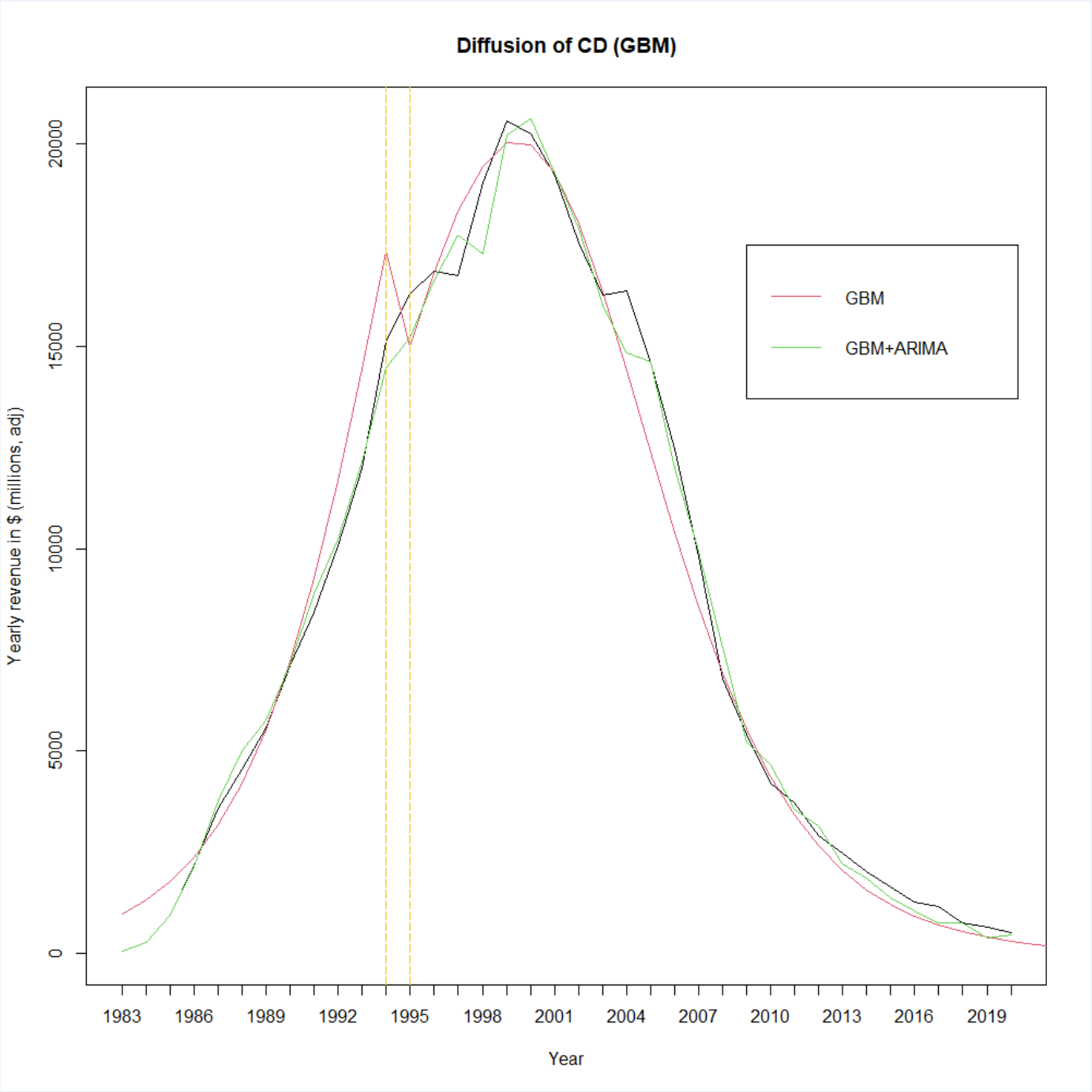


CD diffusion - GBM

Type of shock: Exponential

Note that all values shall be multiplied by 1 million in order to get the real value in USD.

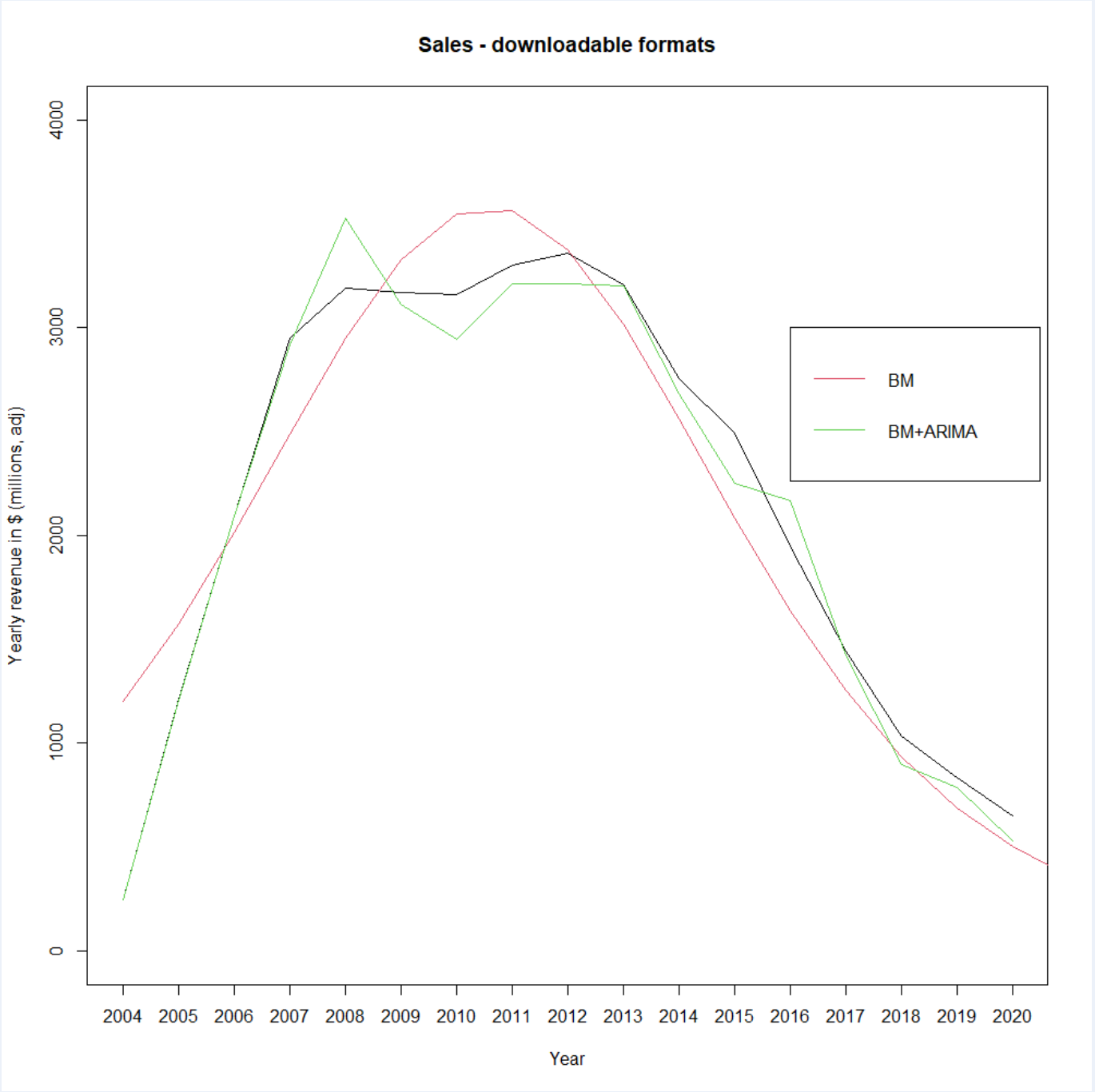
Parameters	Values
Adjusted R-squared	0.69
m	319541
p	0.0022
q	0.3048
a1	12
b1	- 0.0594
c1	- 0.2486



Diffusion of downloadable formats (legal)

All parameters resulted significant for a simple BASS model.

Category / Model parameters	m	P	q
Downloadable formats	38452	0.0233	0.3246

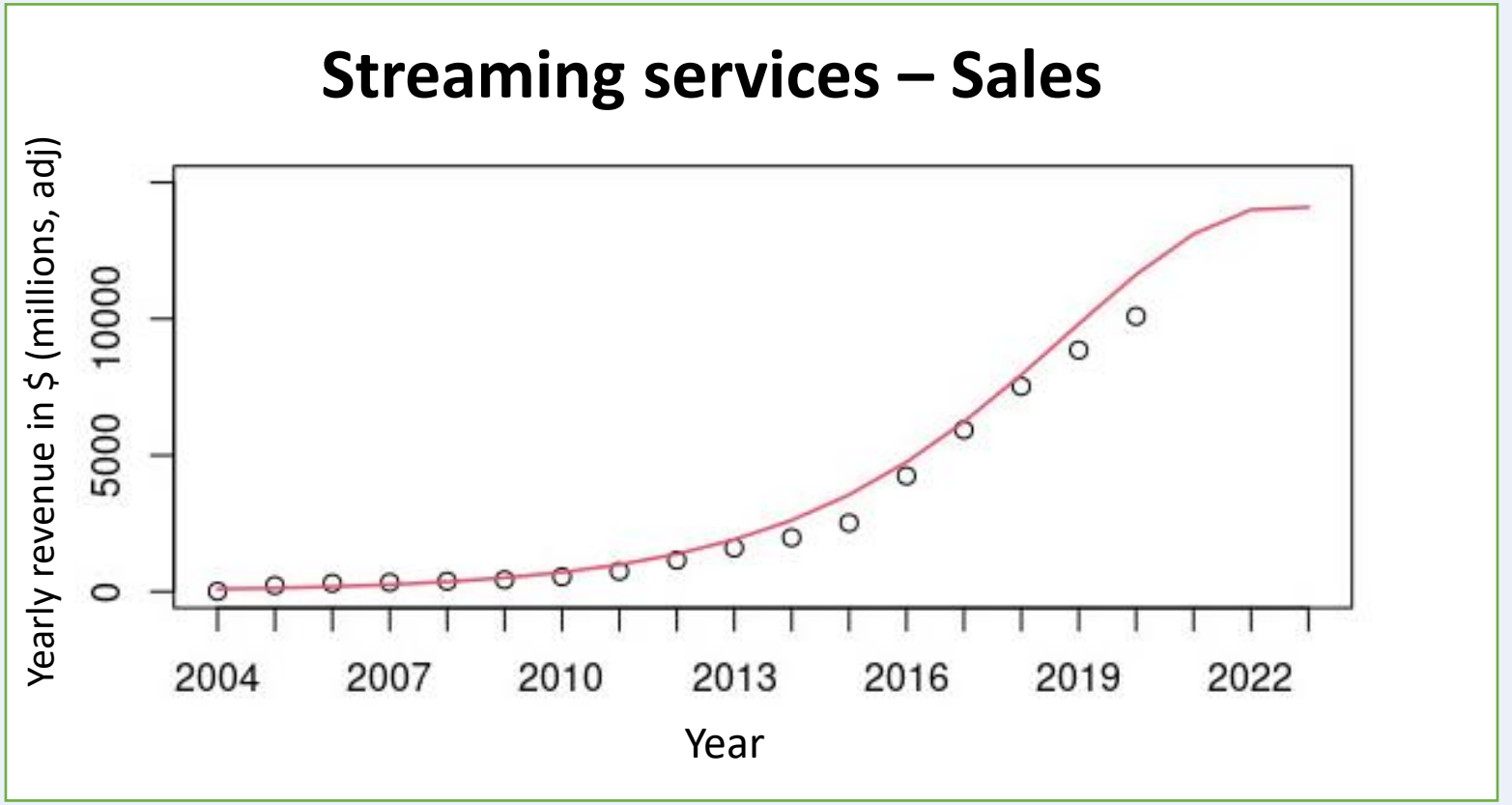


Diffusion of streaming services

All parameters resulted significant for a simple BASS model.

The graph also shows the predicted values for the next 3 years.

Category / Model parameters	m	P	q
Streaming services	163694	0.0004	0.3452



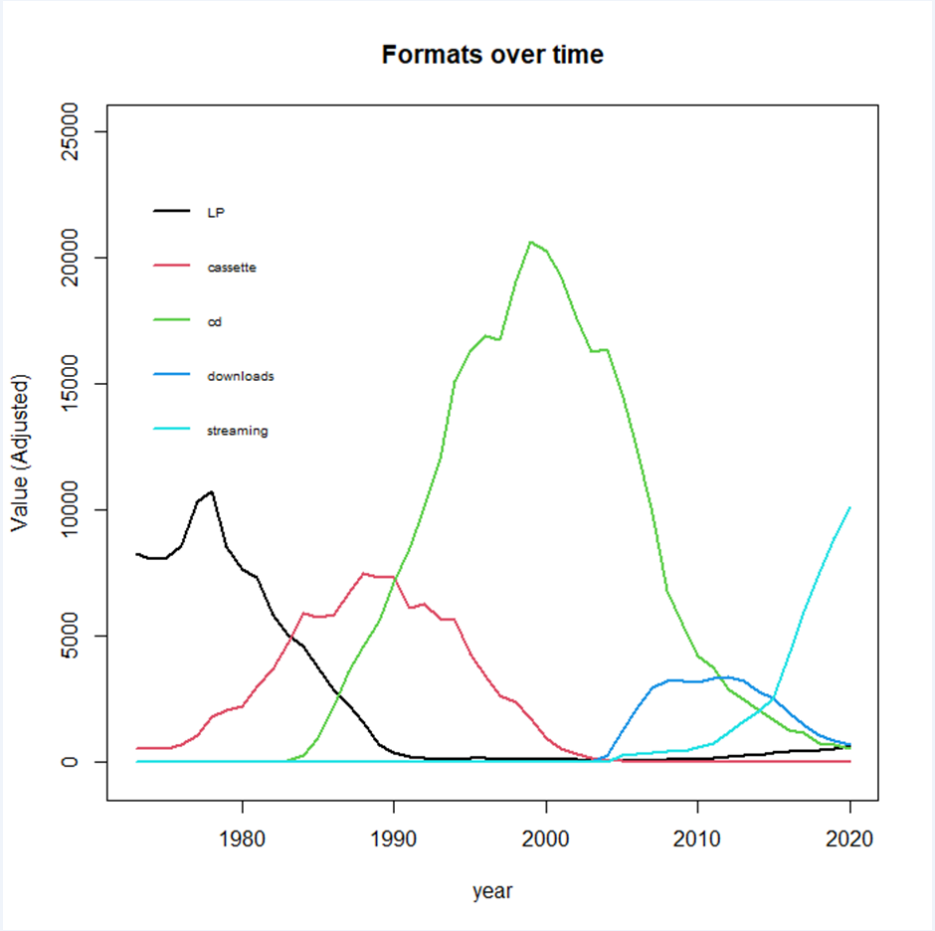
Overview: competition between music formats

After exploring the nature of product diffusion for the 5 different music formats considered, we move on to analysing their **competition dynamics**.

Competitions between formats was modeled, by pairs, using the **UCRCD (Unbalanced Competition Regime Change Diachronic) model**.

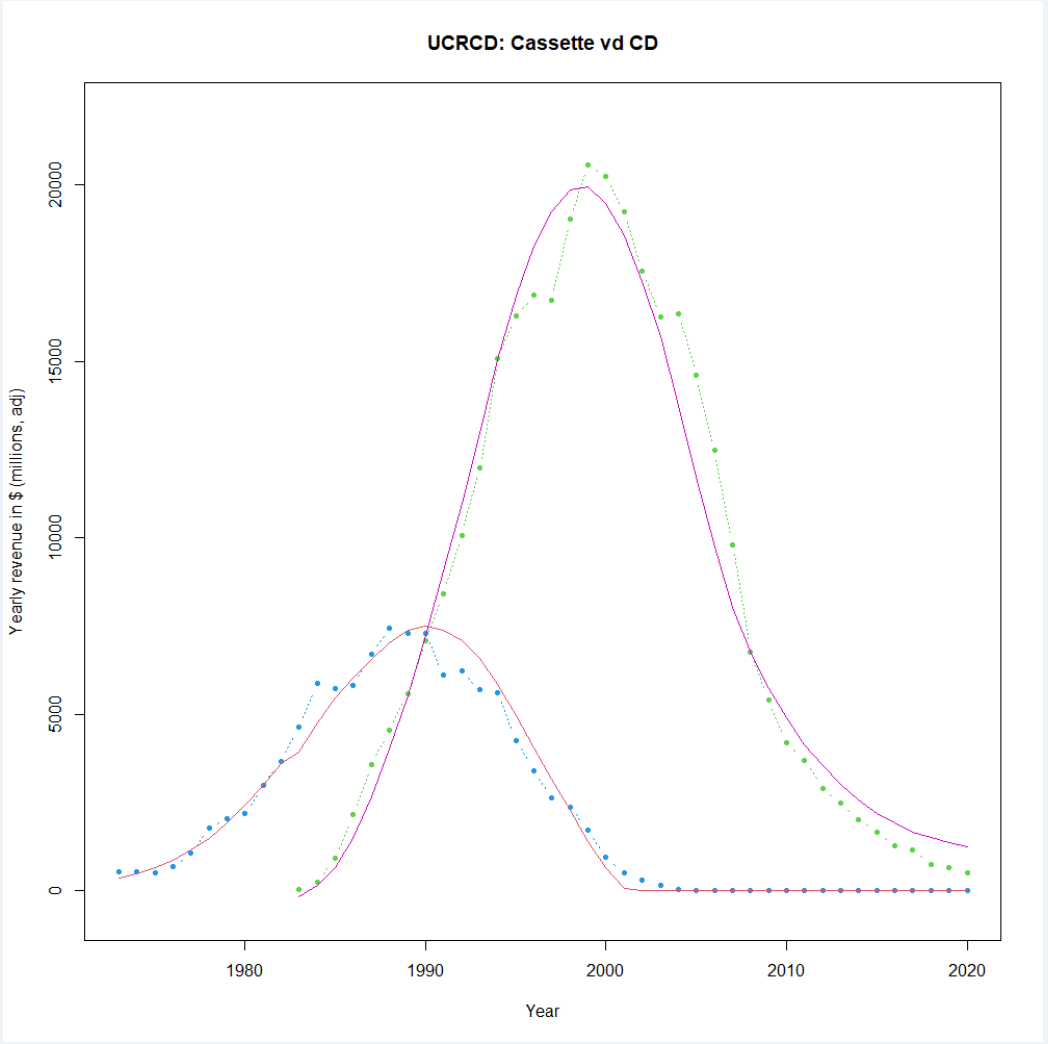
In this context, we noticed **the incumbent format gradually gets replaced by the entrant format**.

As evident from the graph on the right, the duopolistic assumption is only partially respected. Yet, we can still gain some insights from the study of the parameters.



Competition example: Cassette vs CD (UCRCD)

Parameter	Interpretation	Values
m1	Market potential product 1	**67775**
p1a	Innovation coefficient product 1 (before competition)	**0.0046**
q1a	Imitation coefficient product 1 (before competition)	**0.3010**
mc	Combined market potential (after competition)	**415414**
p1c	Innovation coefficient product 1 (after competition)	**0.0078**
q1c + δ	Within – imitation (product 1)	**0.1581**
q1c	Cross – imitation (product 1)	**-0.0868**
p2c	Innovation coefficient product 2	**-0.0009**
q2c	Within – imitation (product 2)	**0.2888**
q2c -ɣ	Cross – imitation (product 2)	**0.0439**



Competition example: Cassette vs CD (UCRCD)

The competition model between CD and cassettes constitutes an example of an **almost-complete duopolistic cycle**.

Observations:

- Parameters δ and ɣ are assumed equal in this case
- Parameter **p2c non-significant**, may indicate **initial resistance to adoption**
- Parameters p1a, q1a and q2c coherent to those observed for standard BM
- **Cross-imitation parameters indicate two different contributions between formats**: competition for CDs w.r.t. cassettes, while a positive effect is suggested for cassettes w.r.t. CDs.

Parameter	Interpretation	Values
m1	Market potential product 1	**67775**
p1a	Innovation coefficient product 1 (before competition)	**0.0046**
q1a	Imitation coefficient product 1 (before competition)	**0.3010**
mc	Combined market potential (after competition)	**415414**
p1c	Innovation coefficient product 1 (after competition)	**0.0078**
q1c + δ	Within – imitation (product 1)	**0.1581**
q1c	Cross – imitation (product 1)	**-0.0868**
p2c	Innovation coefficient product 2	**-0.0009**
q2c	Within – imitation (product 2)	**0.2888**
q2c -ɣ	Cross – imitation (product 2)	**0.0439**

Competition between CDs and downloads: the issue of piracy

- Officially the **first legal revenue for downloads is reported by RIAA for 2004** by ITunes.
- **The MP3 format was released to the public in the early ‘90s** and gained traction in the late ‘90s, initially through small forums/websites.
- **In the second half of 1999 the first large peer-to-peer filesharing network, Napster, is launched**. Widespread adoption rapidly takes off.
- Numerous clones and imitators of Napster quickly appeared online and had a huge impact throughout the first part of 00’s (e.g. LimeWire, Torrent).
- Reaction to the phenomenon by major record companies was slow and inadequate, resulting in huge and prolonged losses.
- **Appropriate laws and sanctions were put into effect around 2005**.

Competition between CDs and downloads: **the issue of piracy**

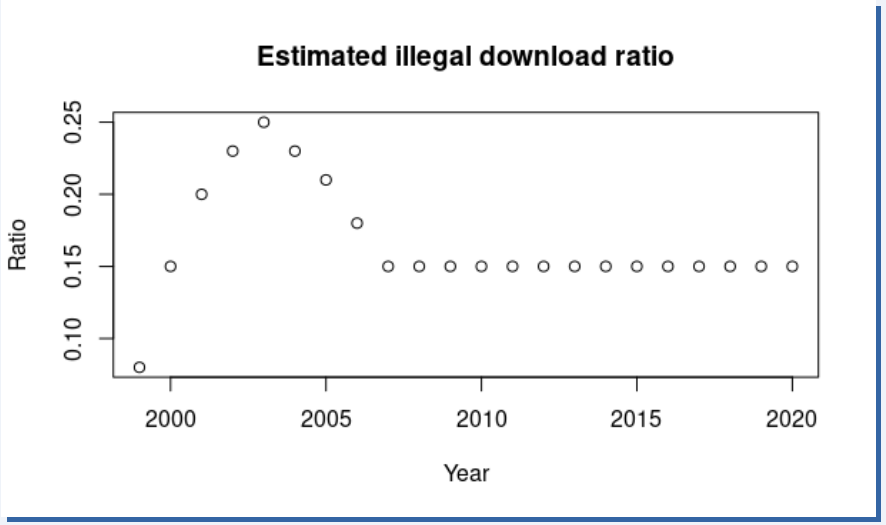
- **Models trained on official data fail to capture the real implications and behaviour of the market in the decade between between 1997-2007**.
- Studies conducted on the true effect of piracy and copyright policy suggest that:

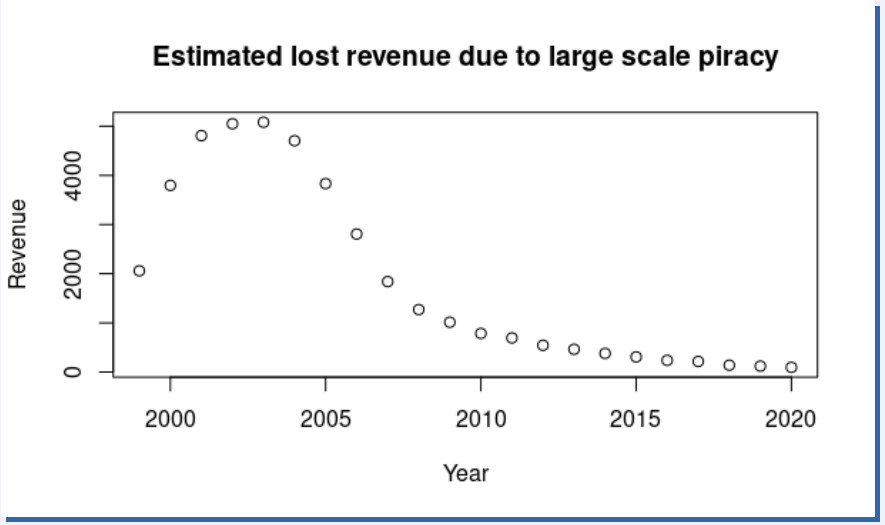
- ***" The estimated displacement rates range up to 30 percent in some studies**, but an average number of these estimated rates can be set to 20 percent." (N. Muller, 2014)*
- *"The results suggest that p2p usage reduces the probability of buying music by 30 percent." (A. Zentner., 2006)*
- *"...the average number of files available from each search (each album) decreased by almost 30 percent following recent legal action against individual file sharers..." (S. Bhattacharjee et Al. 2006)*
- *"Following the announcement of the RIAA to sue P2P users during the summer of 2003, most file-sharing networks have seen their number of users drop by 10 to 30 percent. " (M. Peitz, P. Waelbroeck, 2005)*

- Piracy still remains, although in smaller proportion, an open issue.

Competition between CDs and downloads: an estimation of the scale of piracy

- We employed an **estimate of illegal downloads ratio**, starting from the introduction of large peer-to-peer networks. This was derived from research papers (see previous slide).
- Thus, we were also able to create **an estimate of the CD revenue which was lost to piracy** over the period considered.
- With this integrated data it was then possible to appropriately study the relationship between downloadable format and other recent formats.

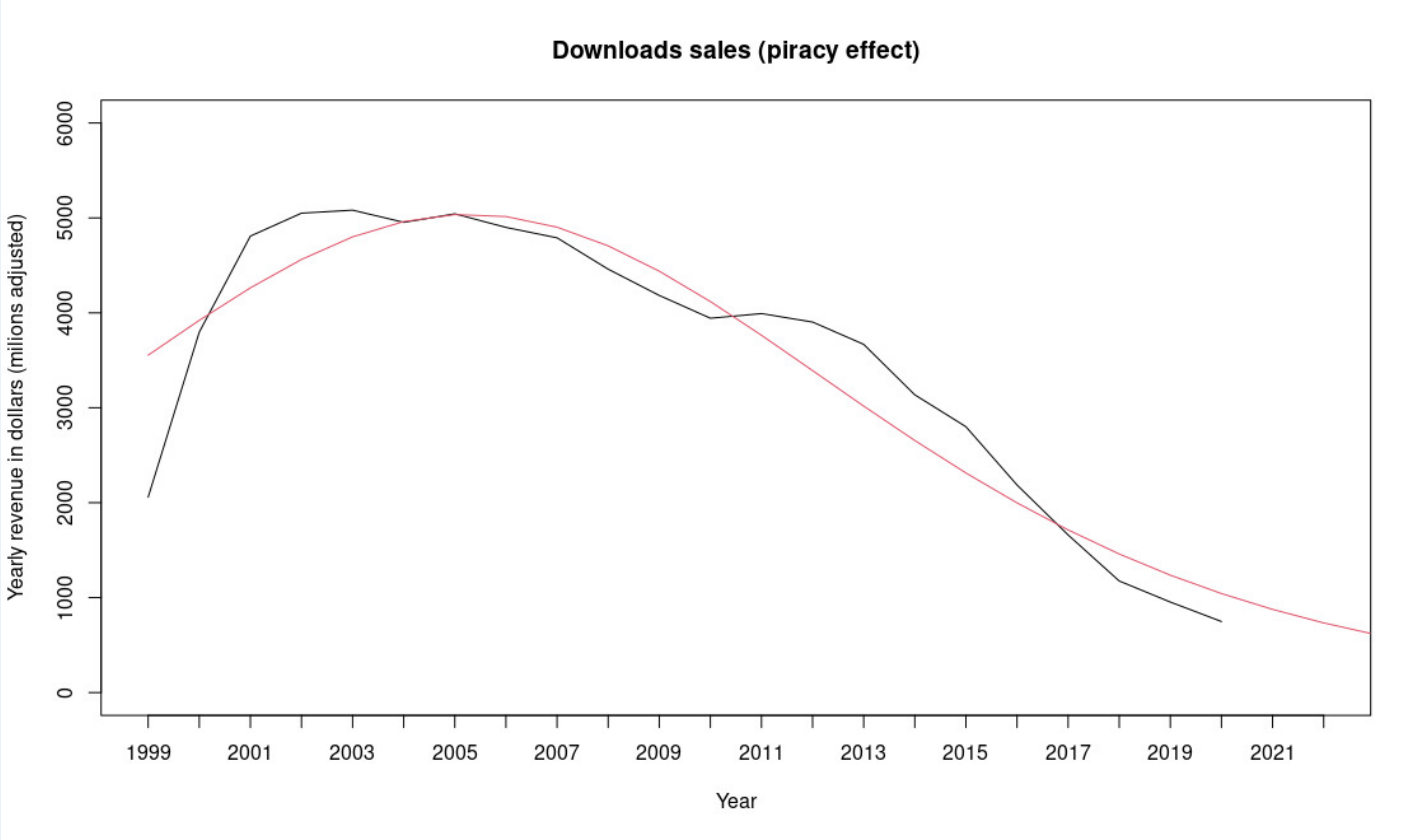




Diffusion of downloadable formats (legal + piracy effect)

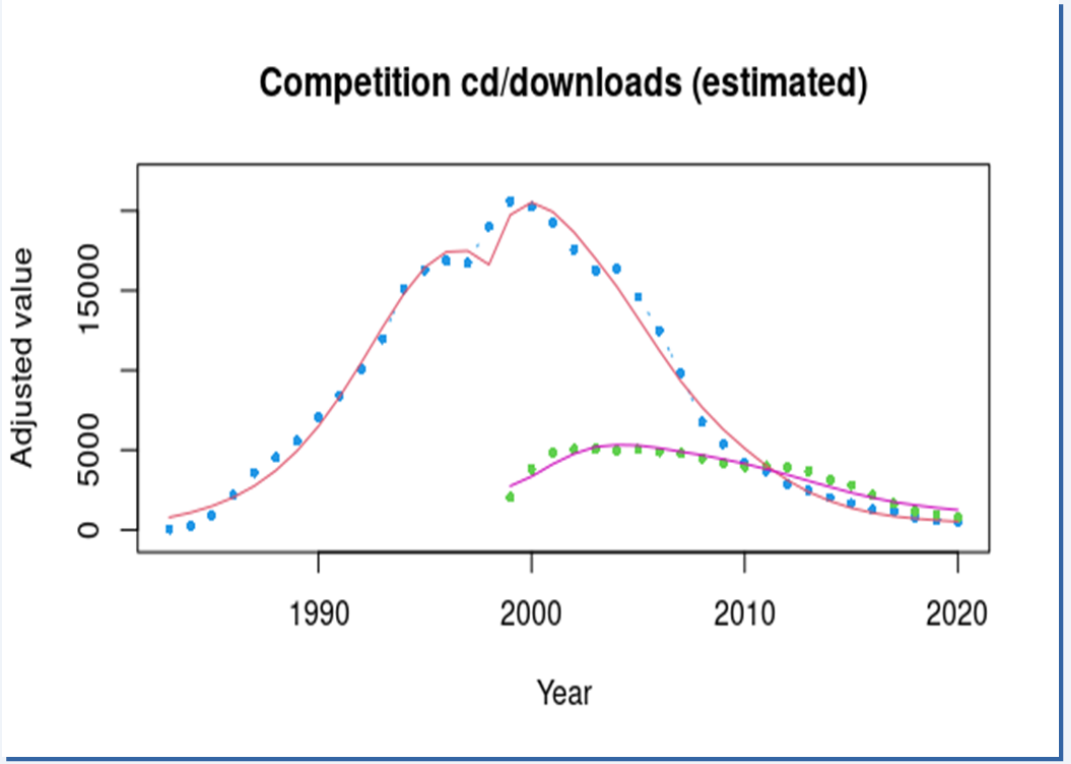
All parameters resulted significant for a simple BASS model.

Category / Model parameters	m	p	q
Downloadable formats	83683	0.0380	0.1556



CD vs Downloadable formats (UCRCD)

Parameter	Interpretation	Values
m1	Market potential product 1	211183
p1a	Innovation coefficient product 1 (before competition)	0.0031
q1a	Imitation coefficient product 1 (before competition)	0.3274
mc	Combined market potential (after competition)	268292
p1c	Innovation coefficient product 1 (after competition)	0.0628
q1c + δ	Within – imitation (product 1)	0.3045
q1c	Cross – imitation (product 1)	-0.7571
p2c	Innovation coefficient product 2	0.0103
q2c	Within – imitation (product 2)	0.4583
q2c -ɣ	Cross – imitation (product 2)	-0.0341



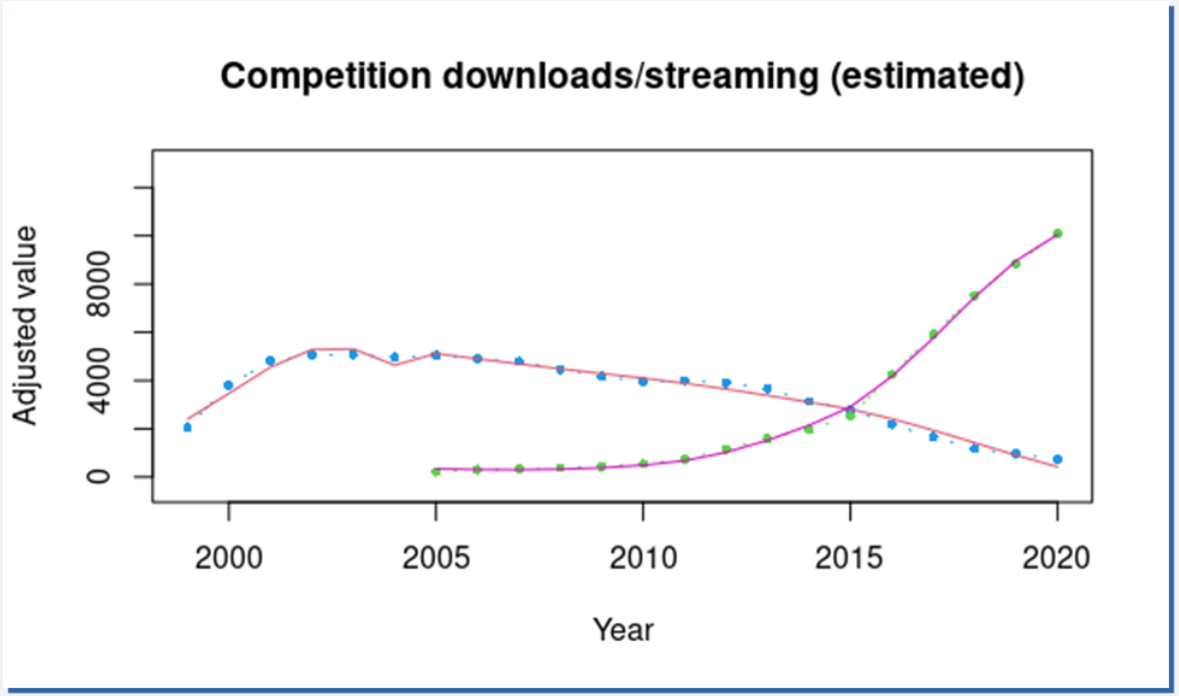
CD vs Downloadable formats (UCRCD)

- All parameters are significant
- **p2c significant and larger than usual**
- **Cross-imitation coefficients both negative**, competition between formats
- Combined m coefficients in line with BMs and added data
- The model appropriately captures the beginning of the shock to the CD market before the appearance of large-scale download platforms on the market (aka before 1999)

Parameter	Interpretation	Values
m1	Market potential product 1	211183
p1a	Innovation coefficient product 1 (before competition)	0.0031
q1a	Imitation coefficient product 1 (before competition)	0.3274
mc	Combined market potential (after competition)	268292
p1c	Innovation coefficient product 1 (after competition)	0.0628
q1c + δ	Within – imitation (product 1)	0.3045
q1c	Cross – imitation (product 1)	-0.7571
p2c	Innovation coefficient product 2	0.0103
q2c	Within – imitation (product 2)	0.4583
q2c -ɣ	Cross – imitation (product 2)	-0.0341

Downloadable formats vs streaming services (UCRCD)

Parameter	Interpretation	Values
m1	Market potential product 1	35794
p1a	Innovation coefficient product 1 (before competition)	0.0540
q1a	Imitation coefficient product 1 (before competition)	0.4927
mc	Combined market potential (after competition)	168121
p1c	Innovation coefficient product 1 (after competition)	0.0317
q1c + δ	Within – imitation (product 1)	-0.0064
q1c	Cross – imitation (product 1)	-0.0850
p2c	Innovation coefficient product 2	0.0025
q2c	Within – imitation (product 2)	0.5505
q2c -ɣ	Cross – imitation (product 2)	-0.0382



Downloadable formats vs streaming services (UCRCD)

- All parameters are significant.
- **Cross-imitation coefficients both negative.**
- p2c significant but lower than previous model for downloads.
- Parameters essentially in line with BMs.
- **Streaming format doesn't appear to have reached peak yet.**

Parameter	Interpretation	Values
m1	Market potential product 1	35794
p1a	Innovation coefficient product 1 (before competition)	0.0540
q1a	Imitation coefficient product 1 (before competition)	0.4927
mc	Combined market potential (after competition)	168121
p1c	Innovation coefficient product 1 (after competition)	0.0317
q1c + δ	Within – imitation (product 1)	-0.0064
q1c	Cross – imitation (product 1)	-0.0850
p2c	Innovation coefficient product 2	0.0025
q2c	Within – imitation (product 2)	0.5505
q2c -ɣ	Cross – imitation (product 2)	-0.0382

What does the history of music industry tell us?

- **The growth of the revenue from a product is usually stopped by the entrance in the market of a competitor.**
- **Piracy had a huge effect in the past on total sales in the music market.** Its effect is reduced nowadays but it's still a factor to consider.
- Music became much cheaper in terms of money per song/album in recent years.
- Adjusted revenue has not caught up to the 90's levels, possibly suggesting there's margin for future growth in the market.
- **Today the dominant products in the market are streaming services.**
- Streaming doesn't appear to have peaked yet, and no clear competitor is present in the market either.

Question 2:
Who is a potential market leader, among producers of this format?

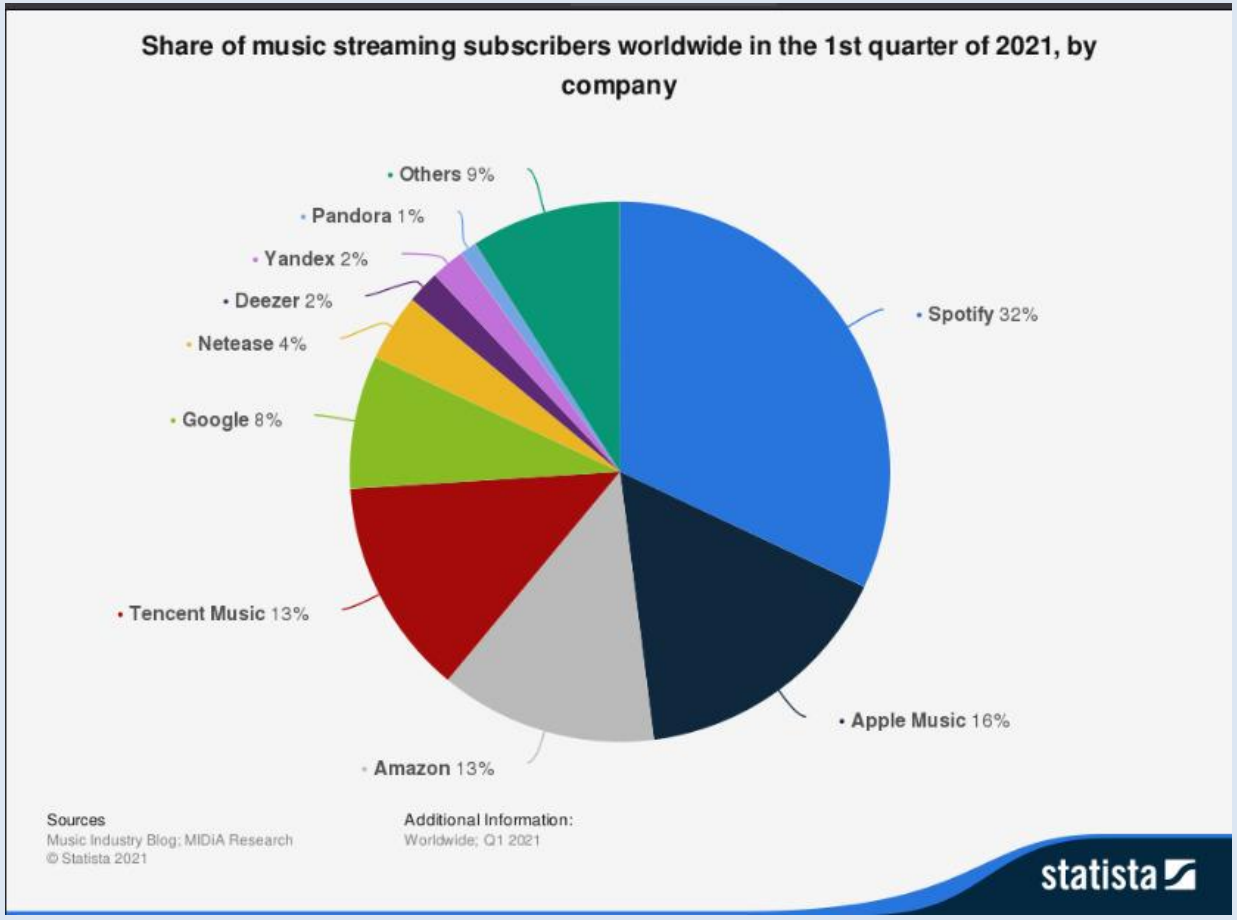
Once streaming services have been identified as the current dominant music format, we were interested in understanding who are the lead players on the streaming market – and identifying a potential business partner for our clients.

As shown in the pie chart to the right, Spotify is currently the market leader for paid streaming services. Moreover, they release their data to the public quarterly.

For the analysis of Spotify's future growth prospects, we worked on 2 datasets:

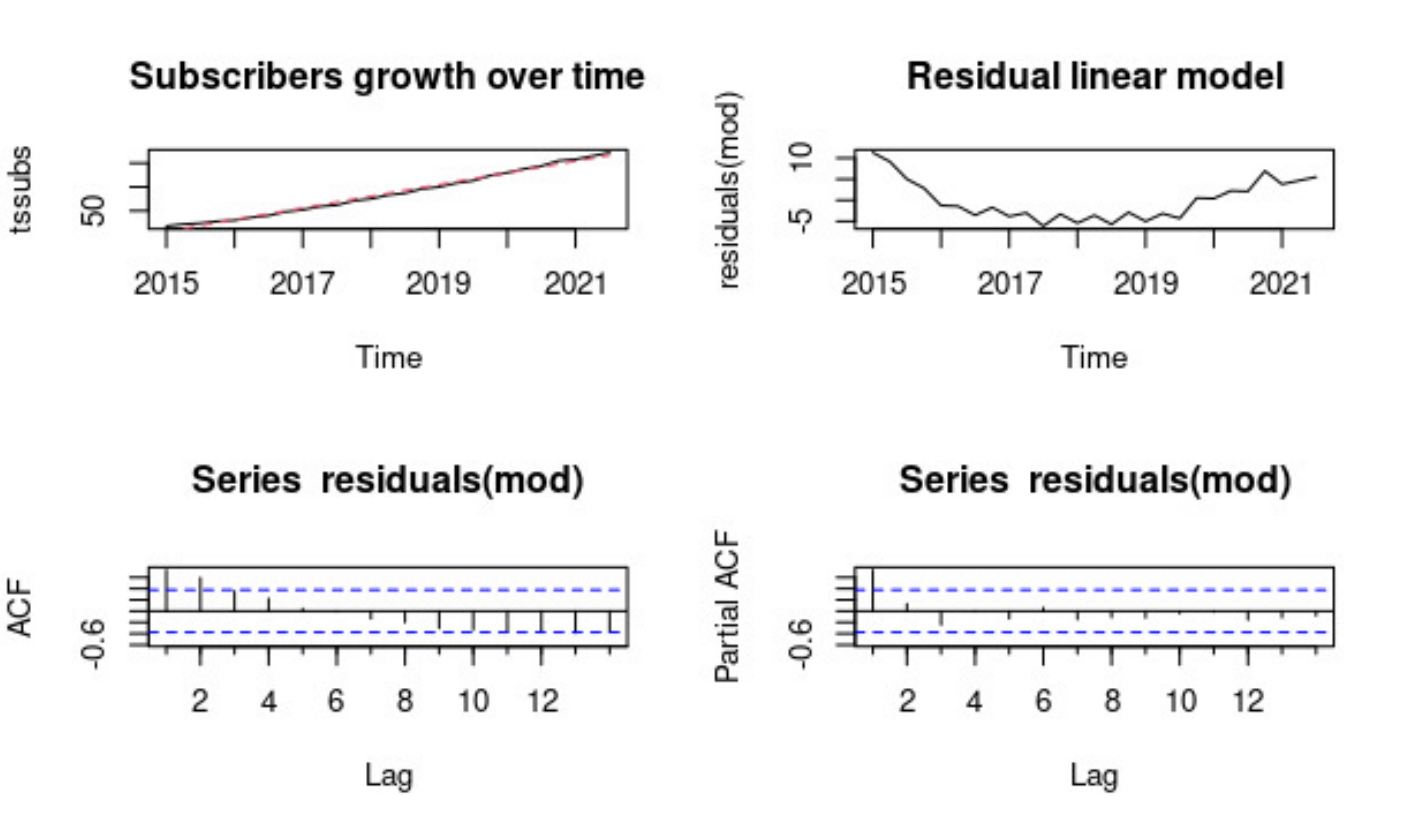
- **Dataset 1**: Quarterly revenue (measured in million of €)
- **Dataset 2**: Number of subscribers (with paid subscription) for each quarter

Both datasets regard the period 2015-2021



Spotify overview: Number of subscribers

Below you can observe a **linear regression model for the growth of Spotify subscribers** and the relative residual plot. Parameters regarding seasonality were added, but they resulted non-significant.
Although the R-squared shows a good fit, **the presence of structure in the residuals suggested to use an ARIMA model in order to improve performance.**

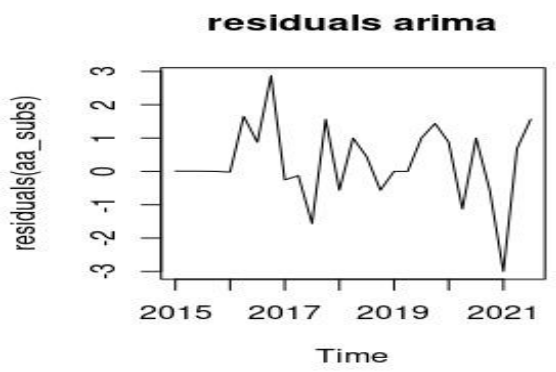


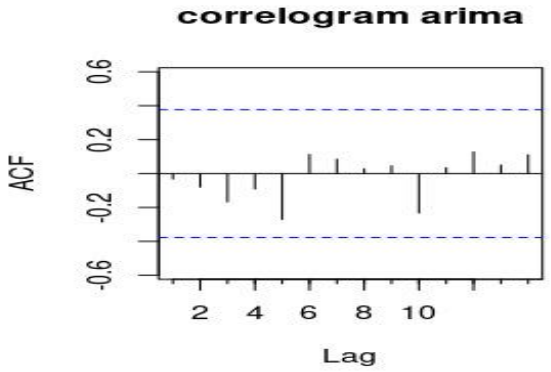
Parameter	Value
Intercept	0.5499
Trend	6.1459
R2	0.9904

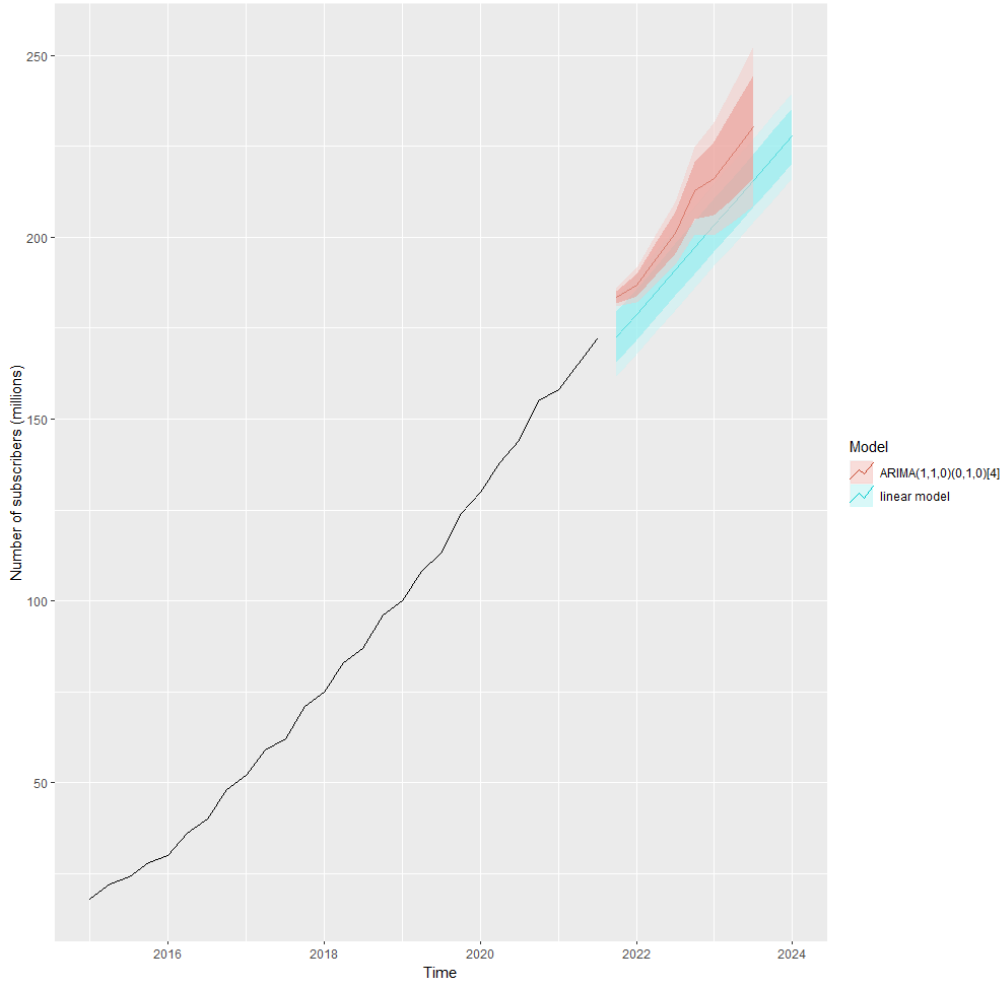
…will Spotify's user base keep growing in the coming years?

In order to obtain reliable predictions for Spotify's future growth, we applied an **ARIMA(1,1,0)(0,1,0)$_4$ model**, selected via AIC and RMSE comparison. Both show significant improvement wrt the linear model. This model suggests even faster growth compared to the linear one, implying **Spotify will keep increasing its subscribers in coming years**.
*RMSE was calculated over the last 7 observations (test set).

Metric/Model	Linear	ARIMA
AIC	166.33	78.18
RMSE*	86.9	13.80



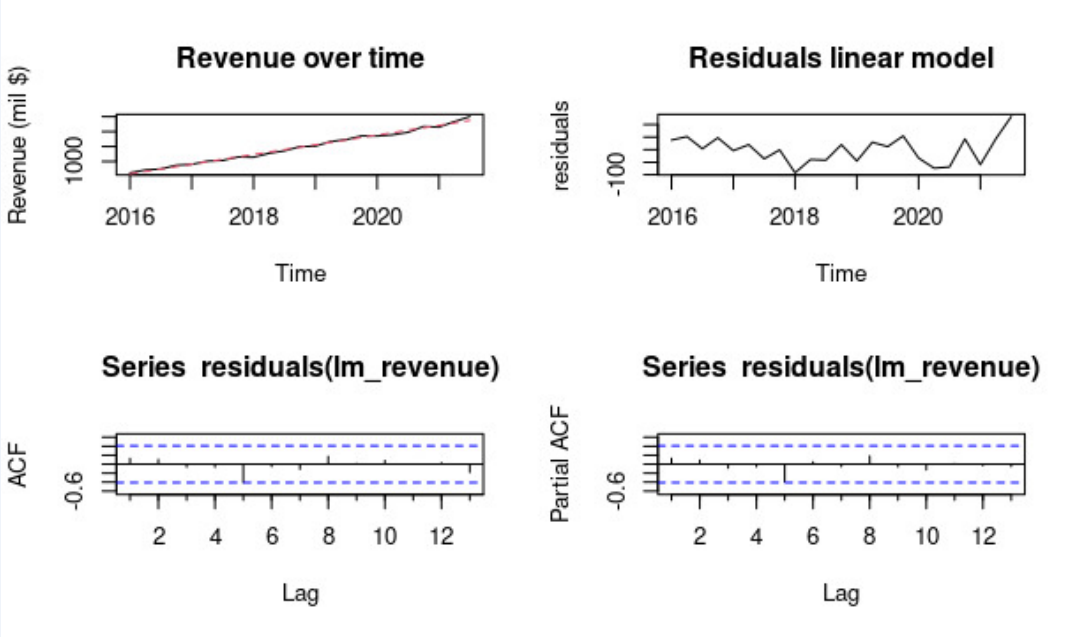


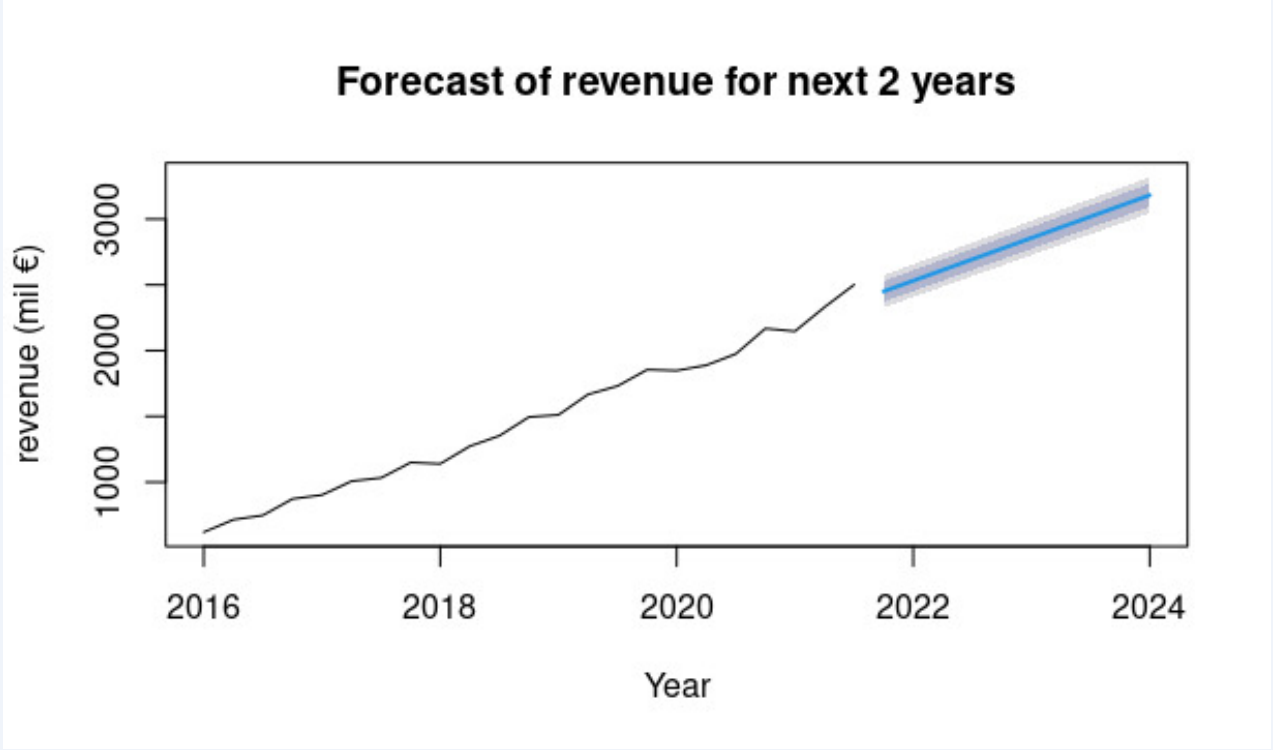


Spotify overview: Revenue

Below you can observe a linear **model for the growth of Spotify revenue** and the relative residual plot. Parameters regarding seasonality were added, but they resulted non-significant. The residuals show no structure, hence we conclude **a linear regression model is sufficient for prediction**, shown at the bottom-right of the slide.

Parameter	Value
Intercept	499.42
Trend	81.29
R2	0.9908





Conclusions: who is a potential market leader in the streaming services area?

Spotify is the dominant market player for streaming services nowadays. Data shows we can expect it to **continue its fast-paced growth for the next 2 or 3 years**.

Hence, we decide to suggest Spotify as a business partner for our clients.

Question 3:

What makes a song popular among consumers who use Spotify?

Following the identification of streaming services as the leading format on the market, and of Spotify as a potential business partner, we move on to **investigate the popularity trends of songs released on the Spotify platform.**

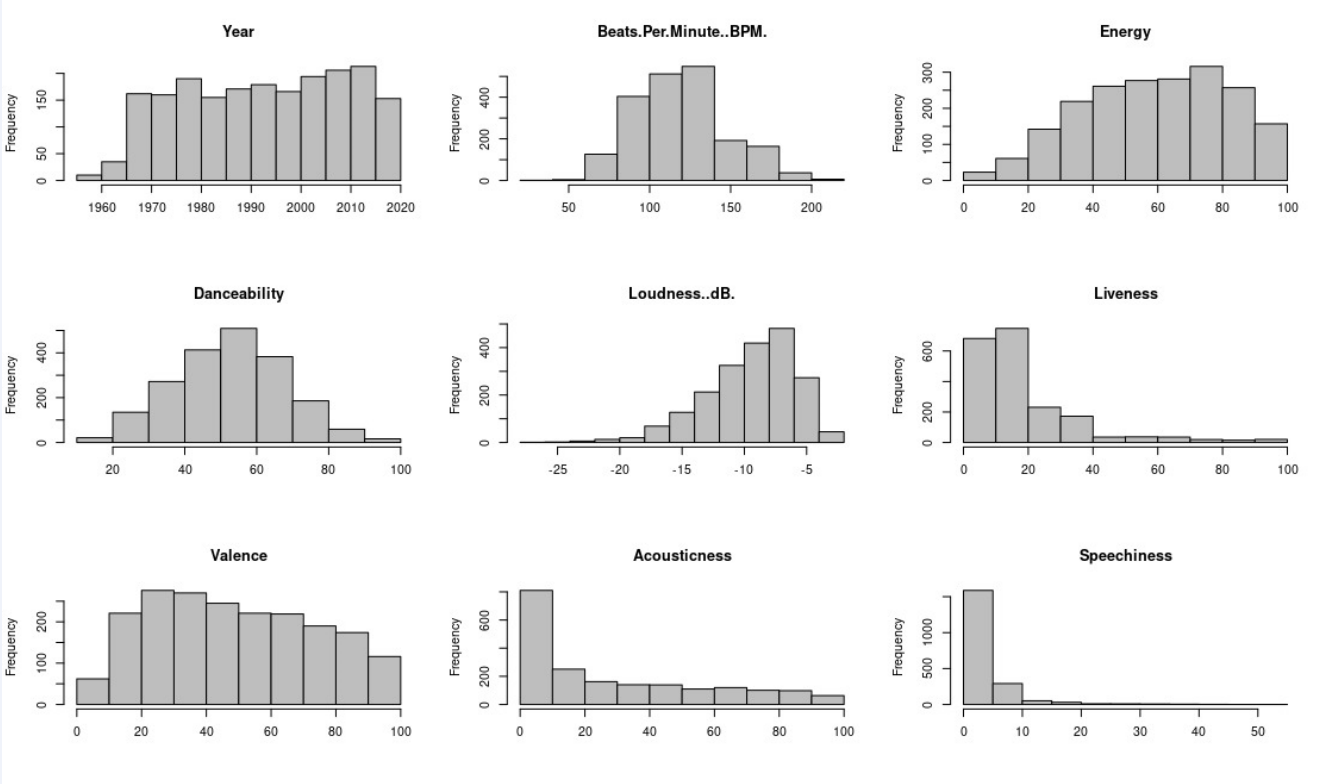
To this aim, we employed a **dataset of 2000 songs provided by Spotify API**, which reports all variables relative to each song, used by Spotify to construct playlists. Many of these variables had no relevance for our purposes, so they were removed in the preliminary phase. Below, you can find **a list of the variables used and their interpretation**.

- **Genre**: Genre of the track*
- **Year**: Release Year of the track
- **Beats per Minute(BPM):** The tempo of the song
- **Energy**: The energy of a song - the higher the value, the more energetic the song
- **Danceability**: The higher the value, the easier it is to dance to this song.
- **Loudness**: The higher the value, the louder the song.
- **Valence**: The higher the value, the more positive mood for the song.
- **Length**: The duration of the song.
- **Acoustic**: The higher the value the more acoustic the song is.
- **Speechiness**: The higher the value the more spoken words the song contains
- **Liveness**: probability that the song was recorded with a live audience
- **Popularity**: The higher the value the more popular the song is.

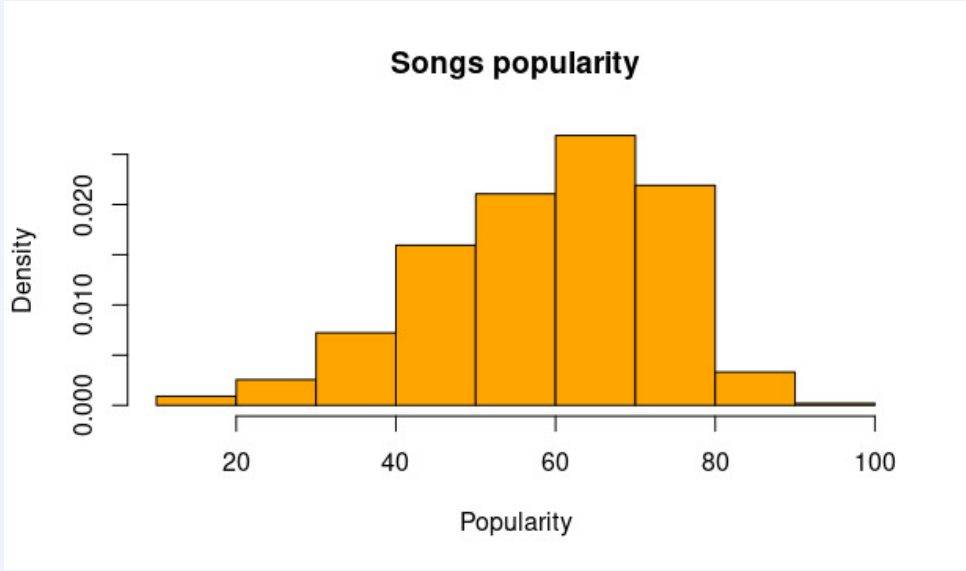
* The Genre variable contained approximately 150 possible genres for each song. For simplicity, we converted these into 12 genres: pop, indie, rock, country, dance, hip-hop, metal, electronica, folk, soul, blues, other

Spotify songs dataset - EDA

Explanatory variables:



Target variable:



- Speechiness: the distribution is clearly skewed to the right, hence we perfomed a logarithmic transformation on this variable

Generalized Additive Model (GAM)

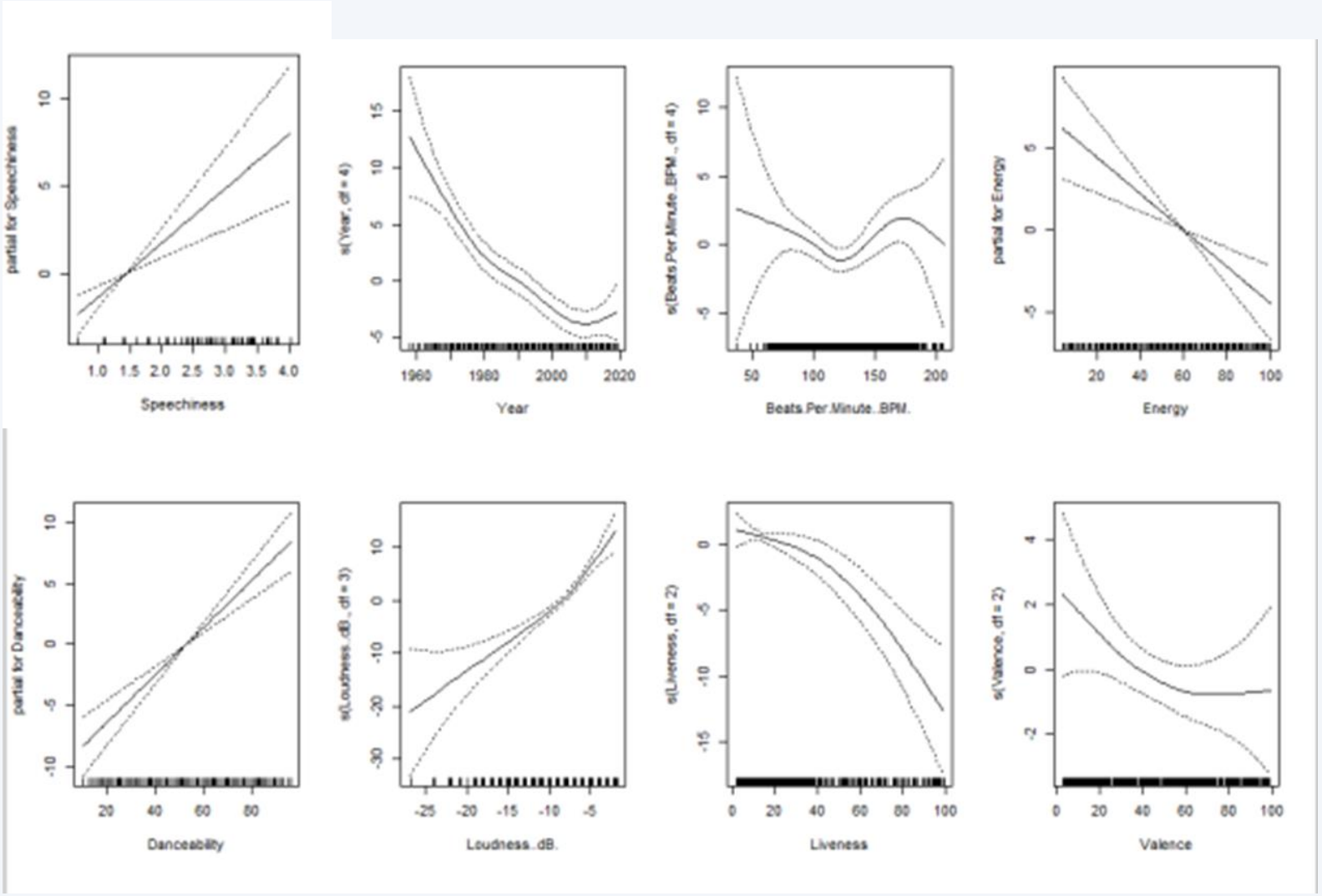
We divided the dataset with a **80%-20% split between training and test set.**

After fitting a **multiple regression model** on the data, which will be used as a **benchmark**, we fitted a GAM on the dataset.

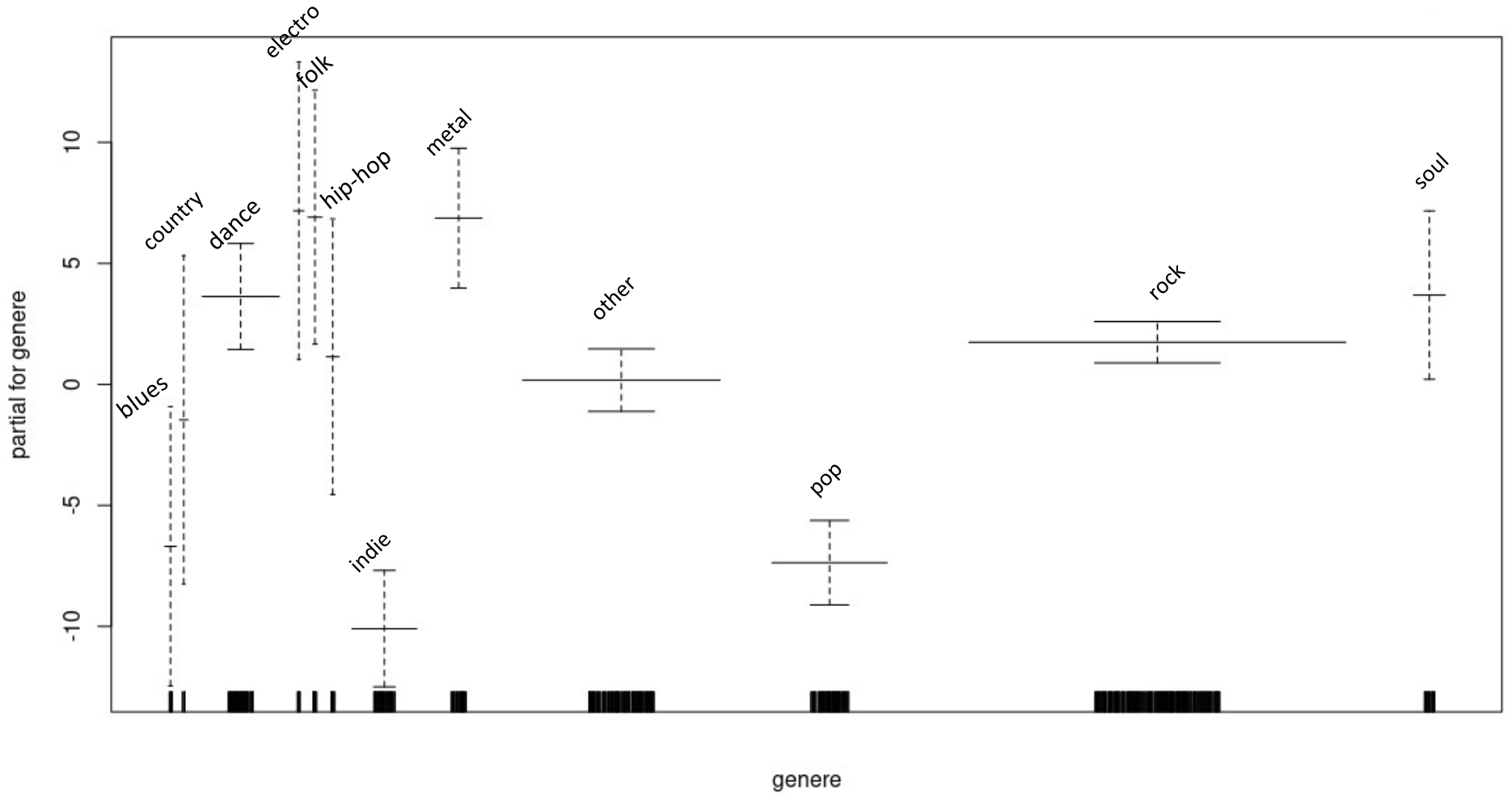
We used an automated step procedure in order to select the best model.

Model / Metric	AIC	Deviance
MLR	12705	60329
GAM	12679	59876

* Variables “length” and “BPM” resulted non-significant both in MLR and GAM

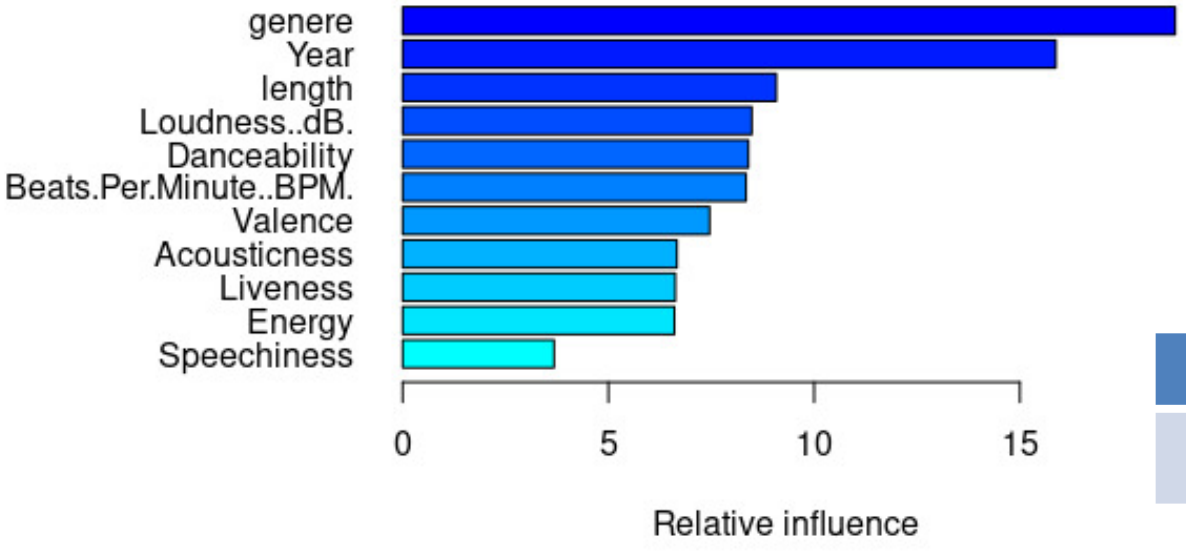


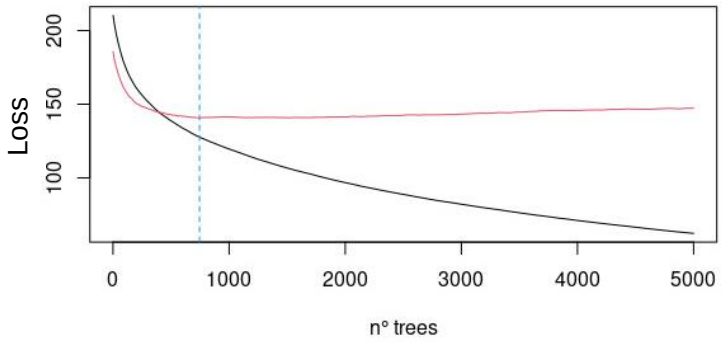
Marginal effect plot – "genre" variable: we were particularly interested in the interpretation of this plot, since it could be useful in order to suggest artist choices to our clients at the hypothetical record label.



Gradient Boosting Machine

Since GAM didn't much improve performance from our benchmark model, we decided to **fit a GBM model to our dataset.**
Below you can observe the **relative influence plot** and **training error plot** for our model, alongside a table which indicates that **GBM has the best performance** wrt both the benchmark and the GAM model.
We tried GBM with 2 different loss functions: absolute (1) and quadratic (2). Best performance was obtained with quadratic loss function (2) and this is the one for which we show graphs.



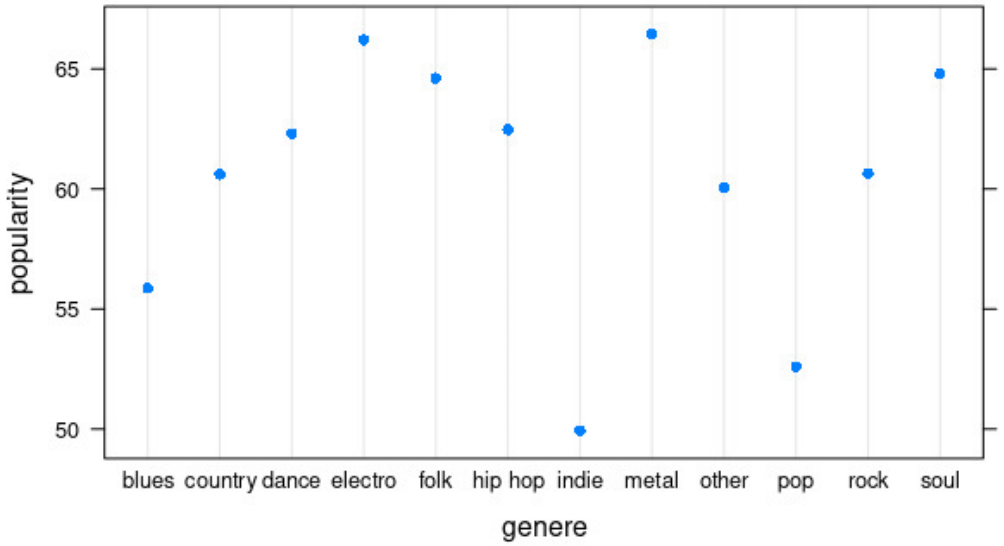


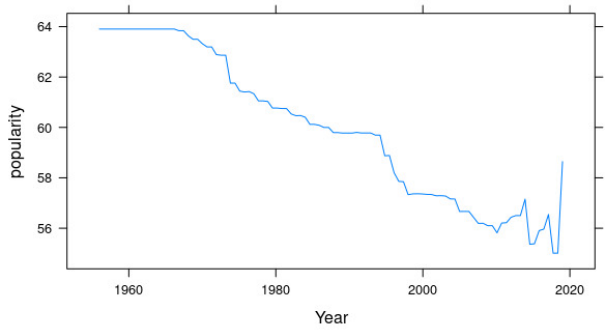
Par	Value
Max depth	4
shrinkage	0.01
Loss function	Quadratic
N of trees	789

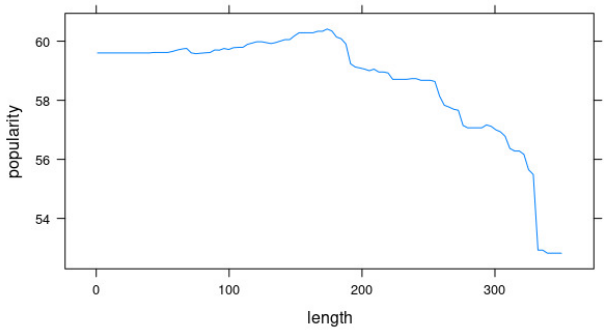
Metric/Model	MLR	GAM	GBM_1	GBM_2
Deviance	60329	59876	58617	55862

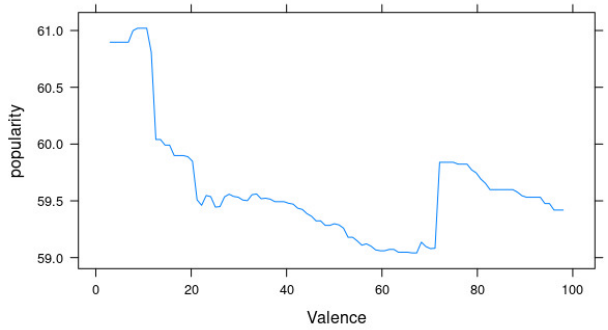
Partial dependence plots (GBM)

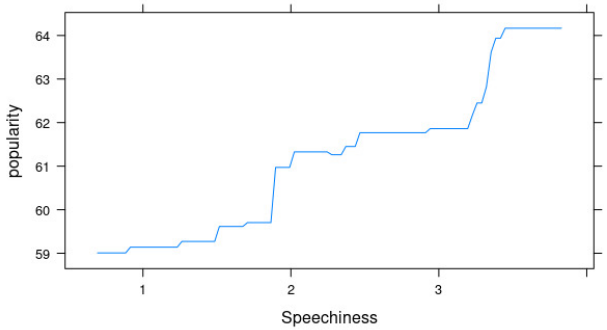
- **Genre**: in line with GAM results, we see the most popular genres are electro, metal and soul.
- **Year**: it negatively influences popularity, with the exception of very recent releases
- **Length**: over a certain length, a drop in popularity is observed
- **Valence**: songs with a negative mood are more popular
- **Speechiness**: as the number of words goes up, popularity increases
- **Liveness**: live recording decreases popularity
- **(Bivariate) Danceability + Loudness**: the louder and more danceable a song is, the more popular

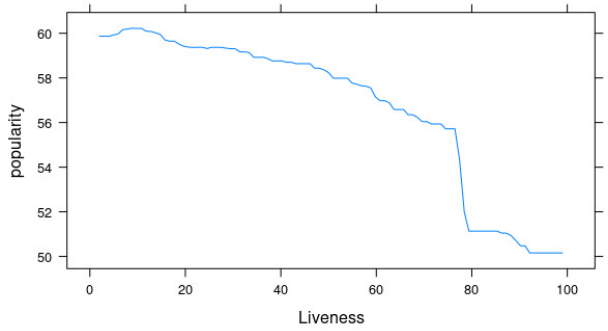


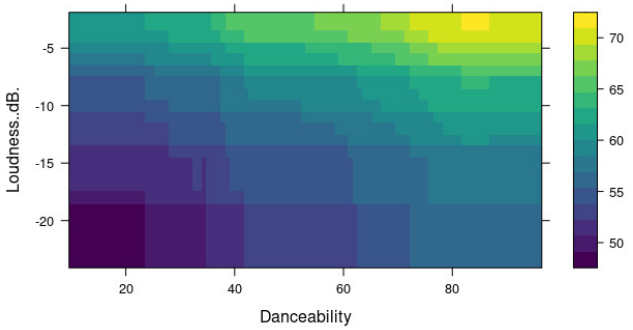












Conclusions: What makes a song popular among consumers who use Spotify?

Our analysis indicated that the factors which relate the most to the popularity of a song on Spotify are its genre, its length (when >3mins), its valence (i.e. mood) and its loudness/danceability.

Hence, we would suggest our clients to focus on buying rights to songs with those qualities, or even signing artists who share these traits.

Here, we assume popularity is a proxy for the number of times a song is listened to on the platform (an easily satisfied assumption, since Spotify API documentation affirms that the value of the variable is calculated on the basis of the number of times the track is listened to).

Conclusions

1. Which music format will dominate the market in the next 2 or 3 years?
   - All diffusion models applied suggest streaming is set out to be the dominant format in the coming years, even considering the problem of piracy.
2. Who is a potential market leader among the producers of this format? Will they keep growing in the coming years?
   - Spotify is already the market leader for streaming services and analysis via ARIMA and linear model suggests it will continue to be a relevant player in the market.
3. What makes a song popular among consumers who buy that particular format?
   - Both GAM and GBM models suggest that the factors which relate the most to the popularity of a song on Spotify are its genre, its length (when >3mins), its valence (i.e. mood) and its loudness/danceability.