Discogs Dataset trend of vinyl and CD with prediction. Decision Trees.

COL Data Science Bootcamp Course

by: Samer Ahmad Diaz

INDEX

1. **Data Set Selection and Understanding**
2. **Problem Definition**
3. **Exploratory Data Analysis (EDA)**
4. **Model Selection and Evaluation**
5. **Feedback and Summarizing**

1. Data Set Selection and Understanding

- The Discogs database (discogs.com) is the biggest and most comprehensive music database online. It hosts a catalogue of over 13 million recordings through all genres and formats, as well as an immense database where everyone can buy and sell music.

1. Data Set Selection and Understanding

- The Discogs database (discogs.com) is the biggest and most comprehensive music database online. It hosts a catalogue of over 13 million recordings through all genres and formats, as well as an immense database where everyone can buy and sell music.

- I have been buying vinyl records for the last years, I have used the website as the main place to discover and buy music which is my passion, this is one of the main reason to select and understanding this data set.

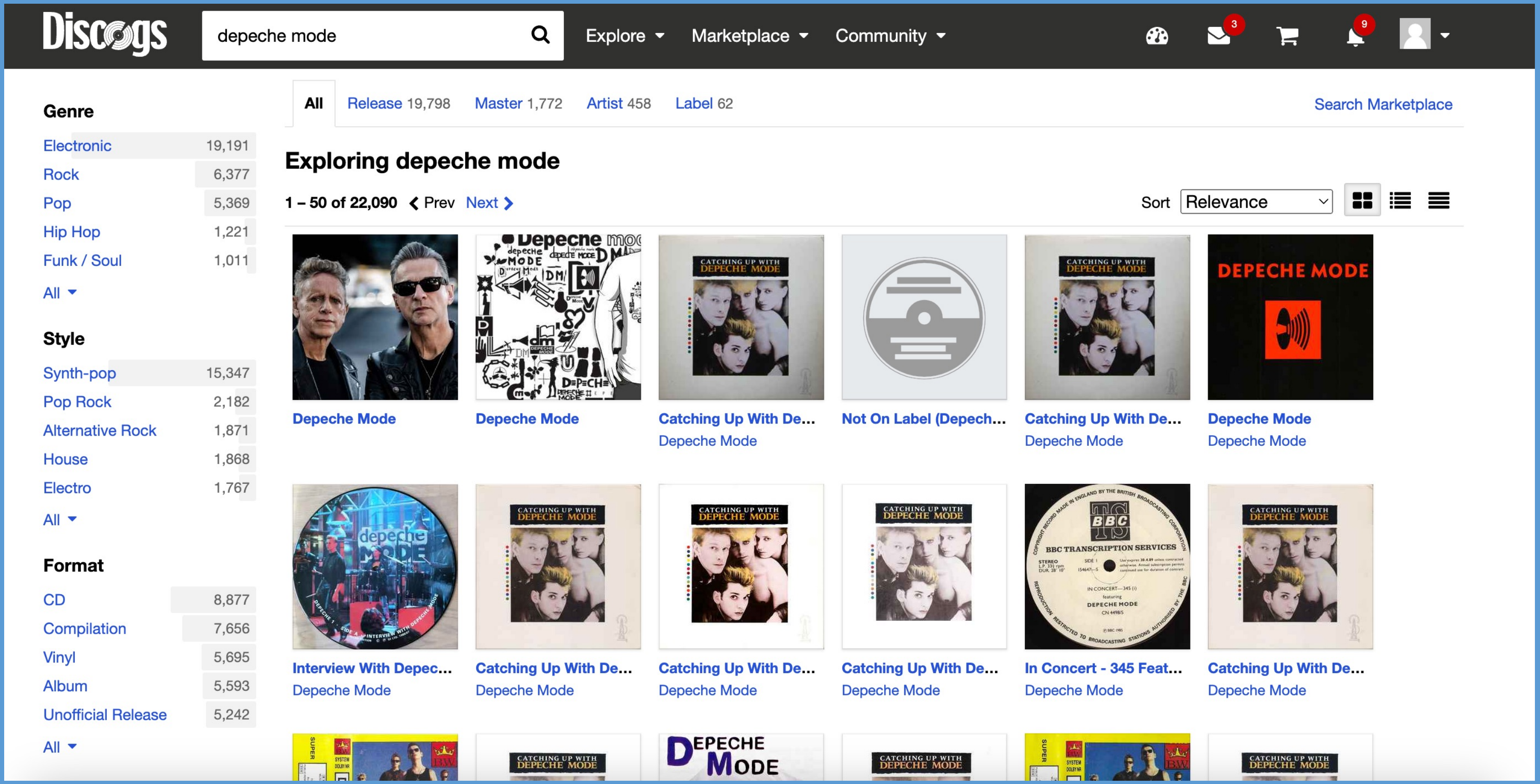
1. Data Set Selection and Understanding

- The Discogs database (discogs.com) is the biggest and most comprehensive music database online. It hosts a catalogue of over 13 million recordings through all genres and formats, as well as an immense database where everyone can buy and sell music.

- I have been buying vinyl records for the last years, I have used the website as the main place to discover and buy music which is my passion, this is one of the main reason to select and understanding this data set.

- The dataset has the following attributes: artist, title, label, country, format, release date , genre, styles, have, want, num_ratings, average_rating, lowest_price, median_price, highest_price.

1. Data Set Selection and Understanding



All | Release 19,798 | Master 1,772 | Artist 458 | Label 62 | Search Marketplace

Genre

Electronic	19,191
Rock	6,377
Pop	5,369
Hip Hop	1,221
Funk / Soul	1,011
All	

Style

Synth-pop	15,347
Pop Rock	2,182
Alternative Rock	1,871
House	1,868
Electro	1,767
All	

Format

CD	8,877
Compilation	7,656
Vinyl	5,695
Album	5,593
Unofficial Release	5,242
All	

Exploring depeche mode

1 – 50 of 22,090 ‹ Prev Next ›

Sort Relevance

Depeche Mode

Depeche Mode

Catching Up With De...
Depeche Mode

Not On Label (Depech...

Catching Up With De...
Depeche Mode

Depeche Mode
Depeche Mode

Interview With Depec...
Depeche Mode

Catching Up With De...
Depeche Mode

Catching Up With De...
Depeche Mode

Catching Up With De...
Depeche Mode

In Concert - 345 Feat...
Depeche Mode

Catching Up With De...
Depeche Mode

2. Problem Definition

Objectives

In order to figure out the market of turntables and CD, looking into the trend of vinyl and CD sales is an idea.

1. How many vinyl and CD were released per year from 2000 to 2019?

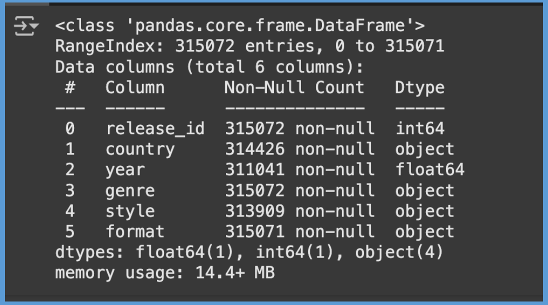
2. Is there a correlation between vinyl records and CD?

3. What is the forecast for the vinyl and CD released through 2025?

4. Decision Tree

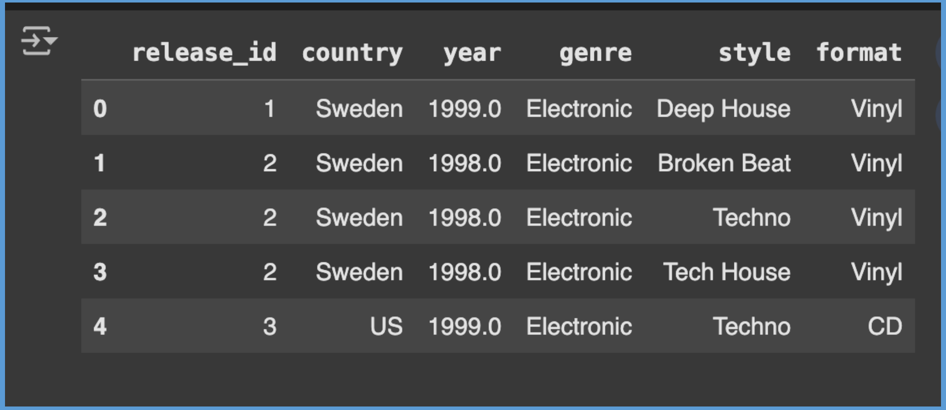
3. Exploratory Data Analysis (EDA)

df.info()

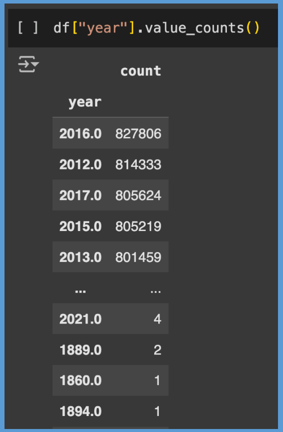


```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 315072 entries, 0 to 315071
Data columns (total 6 columns):
 #   Column      Non-Null Count   Dtype
---  ------      --------------   -----
 0   release_id  315072 non-null  int64
 1   country     314426 non-null  object
 2   year        311041 non-null  float64
 3   genre       315072 non-null  object
 4   style       313909 non-null  object
 5   format      315071 non-null  object
dtypes: float64(1), int64(1), object(4)
memory usage: 14.4+ MB
```

df.head()



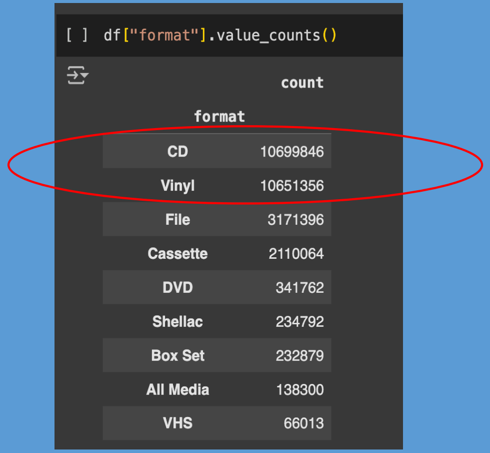
	release_id	country	year	genre	style	format
0	1	Sweden	1999.0	Electronic	Deep House	Vinyl
1	2	Sweden	1998.0	Electronic	Broken Beat	Vinyl
2	2	Sweden	1998.0	Electronic	Techno	Vinyl
3	2	Sweden	1998.0	Electronic	Tech House	Vinyl
4	3	US	1999.0	Electronic	Techno	CD



```
[ ] df["year"].value_counts()
```

	count
year	
2016.0	827806
2012.0	814333
2017.0	805624
2015.0	805219
2013.0	801459
...	...
2021.0	4
1889.0	2
1860.0	1
1894.0	1

1. Exploratory
2. Clean data
3. Filter CD &Vinyl
4. Combining per year

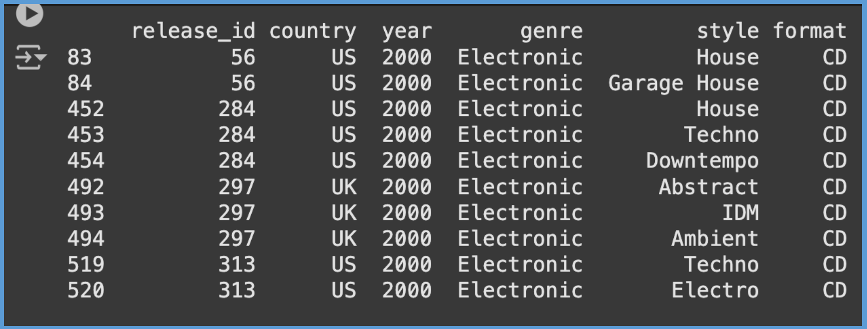


```
[ ] df["format"].value_counts()
```

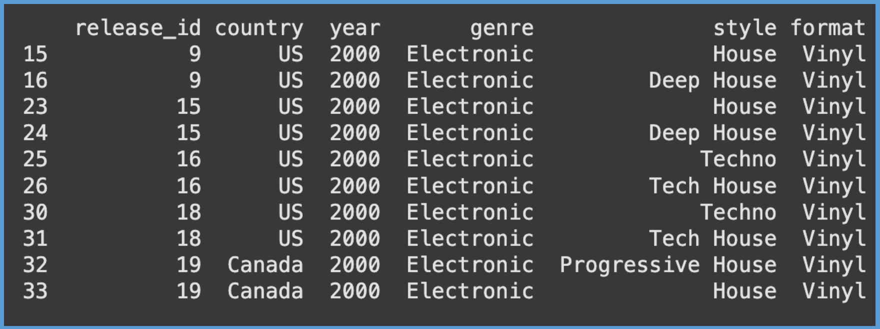
	count
format	
CD	10699846
Vinyl	10651356
File	3171396
Cassette	2110064
DVD	341762
Shellac	234792
Box Set	232879
All Media	138300
VHS	66013

3. Exploratory Data Analysis (EDA)

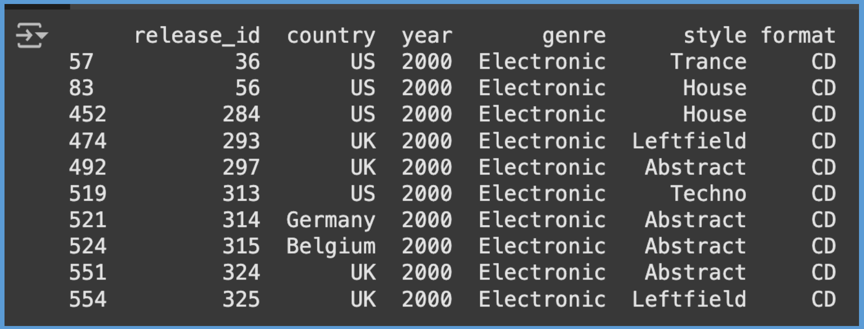
Remove duplicates values and combining num realese_id / year



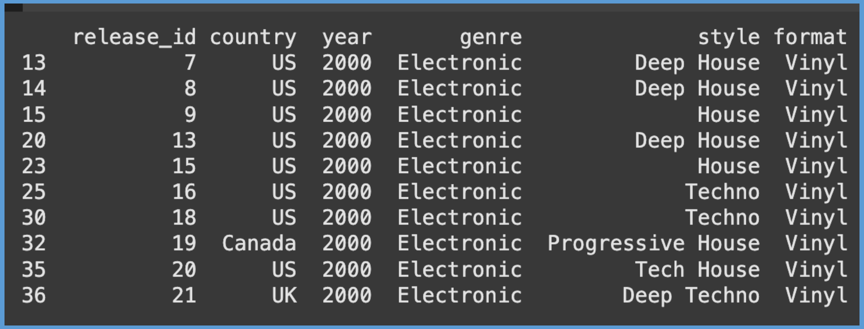
```
     release_id country  year       genre         style format
83           56      US  2000  Electronic         House     CD
84           56      US  2000  Electronic  Garage House     CD
452         284      US  2000  Electronic         House     CD
453         284      US  2000  Electronic        Techno     CD
454         284      US  2000  Electronic     Downtempo     CD
492         297      UK  2000  Electronic      Abstract     CD
493         297      UK  2000  Electronic           IDM     CD
494         297      UK  2000  Electronic       Ambient     CD
519         313      US  2000  Electronic        Techno     CD
520         313      US  2000  Electronic       Electro     CD
```



```
    release_id country  year       genre              style format
15           9      US  2000  Electronic              House  Vinyl
16           9      US  2000  Electronic         Deep House  Vinyl
23          15      US  2000  Electronic              House  Vinyl
24          15      US  2000  Electronic         Deep House  Vinyl
25          16      US  2000  Electronic             Techno  Vinyl
26          16      US  2000  Electronic         Tech House  Vinyl
30          18      US  2000  Electronic             Techno  Vinyl
31          18      US  2000  Electronic         Tech House  Vinyl
32          19  Canada  2000  Electronic  Progressive House  Vinyl
33          19  Canada  2000  Electronic              House  Vinyl
```



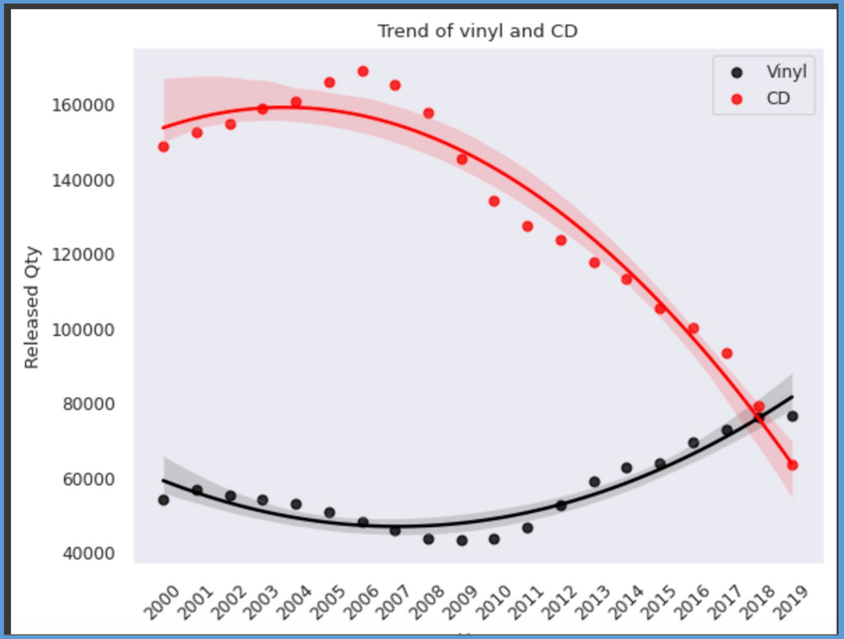
```
     release_id  country  year       genre      style format
57           36       US  2000  Electronic     Trance     CD
83           56       US  2000  Electronic      House     CD
452         284       US  2000  Electronic      House     CD
474         293       UK  2000  Electronic  Leftfield     CD
492         297       UK  2000  Electronic   Abstract     CD
519         313       US  2000  Electronic     Techno     CD
521         314  Germany  2000  Electronic   Abstract     CD
524         315  Belgium  2000  Electronic   Abstract     CD
551         324       UK  2000  Electronic   Abstract     CD
554         325       UK  2000  Electronic  Leftfield     CD
```

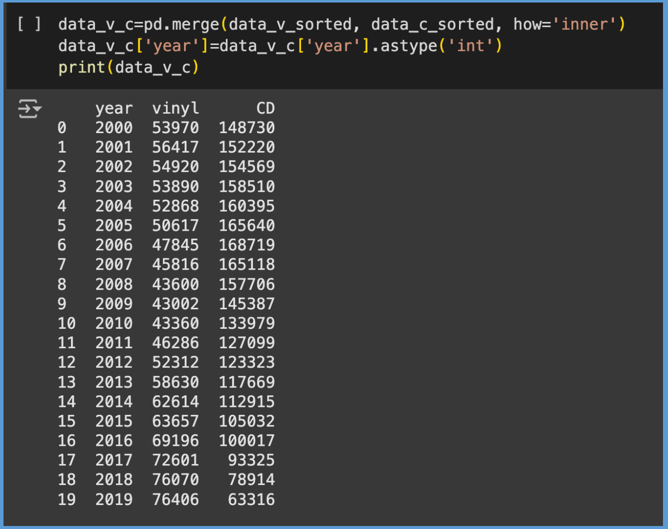


```
    release_id country  year       genre              style format
13           7      US  2000  Electronic         Deep House  Vinyl
14           8      US  2000  Electronic         Deep House  Vinyl
15           9      US  2000  Electronic              House  Vinyl
20          13      US  2000  Electronic         Deep House  Vinyl
23          15      US  2000  Electronic              House  Vinyl
25          16      US  2000  Electronic             Techno  Vinyl
30          18      US  2000  Electronic             Techno  Vinyl
32          19  Canada  2000  Electronic  Progressive House  Vinyl
35          20      US  2000  Electronic         Tech House  Vinyl
36          21      UK  2000  Electronic        Deep Techno  Vinyl
```

4. Model Selection and Evaluation

Plot the data to see the released vinyl and CD were changing



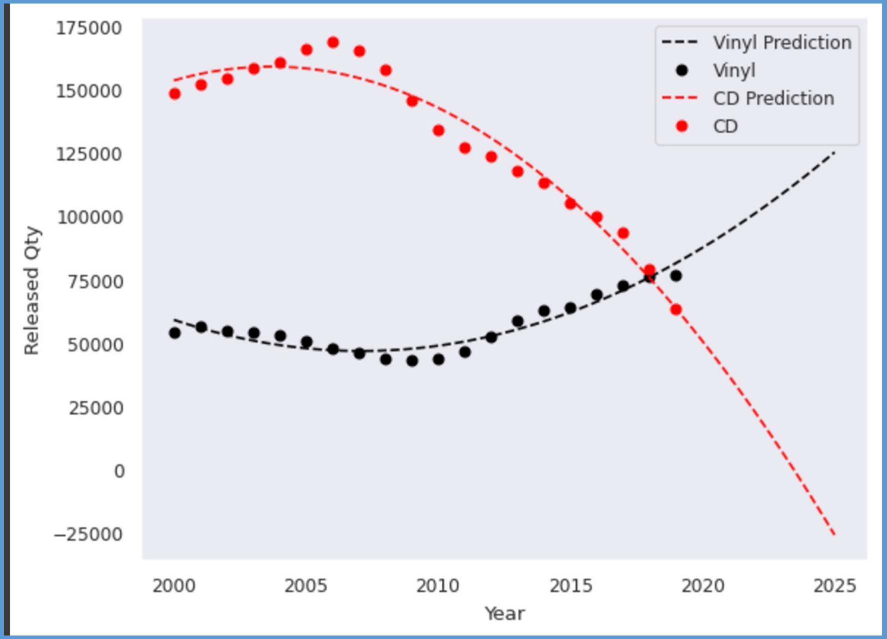


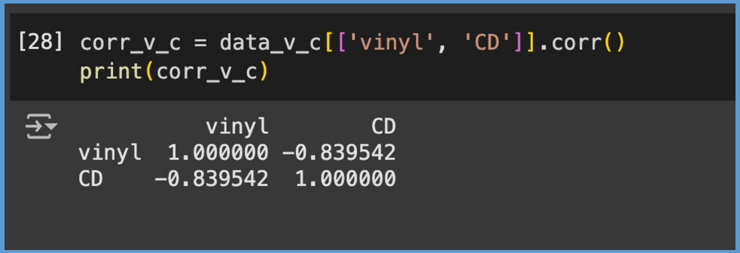
```
[ ] data_v_c=pd.merge(data_v_sorted, data_c_sorted, how='inner')
    data_v_c['year']=data_v_c['year'].astype('int')
    print(data_v_c)
```

```
    year  vinyl      CD
0   2000  53970  148730
1   2001  56417  152220
2   2002  54920  154569
3   2003  53890  158510
4   2004  52868  160395
5   2005  50617  165640
6   2006  47845  168719
7   2007  45816  165118
8   2008  43600  157706
9   2009  43002  145387
10  2010  43360  133979
11  2011  46286  127099
12  2012  52312  123323
13  2013  58630  117669
14  2014  62614  112915
15  2015  63657  105032
16  2016  69196  100017
17  2017  72601   93325
18  2018  76070   78914
19  2019  76406   63316
```

4. Model Selection and Evaluation

Regression model polynomial In order to forecast how many vinyl and CD may be released in the next 5 years.





```
[28] corr_v_c = data_v_c[['vinyl', 'CD']].corr()
     print(corr_v_c)

               vinyl        CD
     vinyl  1.000000 -0.839542
     CD    -0.839542  1.000000
```

The number of records for vinyls are increasing, although the number of CDs are decreasing.

There's an inverse relationship between the variables vinyls and CDs.

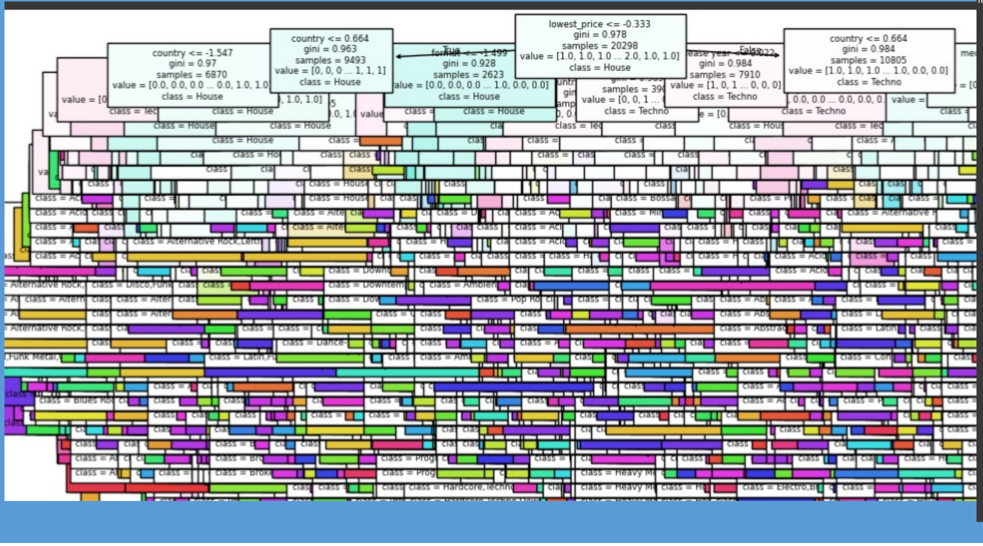
4. Model Selection and Evaluation

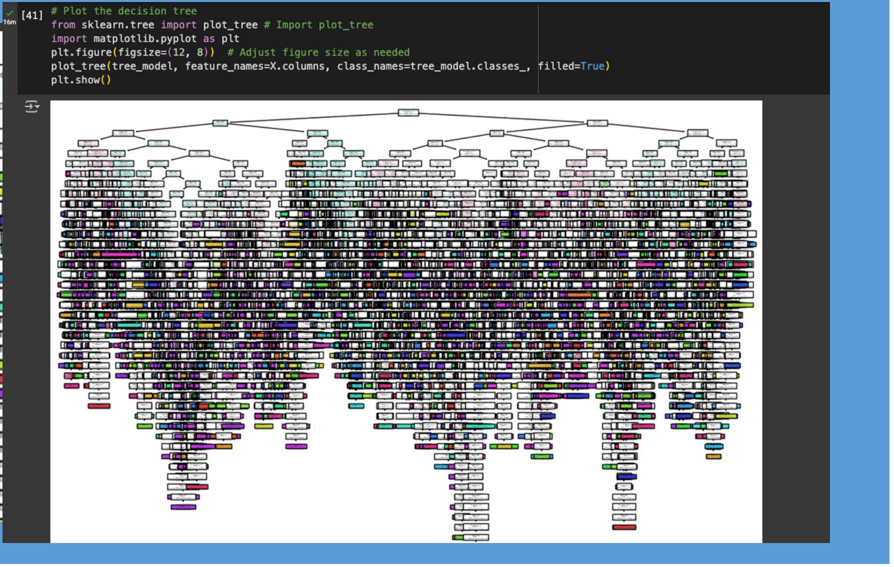
Decision Tree

The 'styles' feature in the dataset likely represents a list of music styles associated with each record. This is a categorical feature that could be valuable for predicting other aspects of the record, such as genre, popularity, or price. Possible decision tree tasks with 'styles'.

Conditions: we only take top 5 countries.

Due to the quantities of matrix the decision tree is not a model selection to be considered.





```
[41] # Plot the decision tree
     from sklearn.tree import plot_tree # Import plot_tree
     import matplotlib.pyplot as plt
     plt.figure(figsize=(12, 8))  # Adjust figure size as needed
     plot_tree(tree_model, feature_names=X.columns, class_names=tree_model.classes_, filled=True)
     plt.show()
```

5. Feedback and Summarizing

The number of records for vinyls are increasing, although the number of CDs are decreasing.

There's an inverse relationship between the variables vinyls and CDs.

Vinyls is increasing the quantities released exponential.

CDs are not expecting to be selling anymore.

People are preferring vinyls collectors or media than CDs.