Atividade 05 - Relatório Final

Equipe

Nome: Donda
Repositório GitLab da equipe: https://gitlab.com/jr_waine/csb-53
Membros:

Waine Barbosa de Oliveira Junior, 1905120,
jr_waine, waine@alunos.utfpr.edu.br, Eng. Comp., UTFPR
Eduardo Yoshio da Rocha, 1508733,
eduardo2798, eduardo.yoshio@outlook.com, Eng. Comp., UTFPR

Tema

Analisar os fatores determinantes para indicação de melhor álbum no Grammy Awards.

Motivação

É muito comum ouvir críticas acerca do Grammy e suas indicações para melhor música e álbum do ano, alegando que os álbuns são indicados não pela qualidade, mas sim porque foram populares ou sucesso de vendas.

Uma análise criteriosa dos dados de álbuns, como nota da crítica, número de vendas/ouvintes do álbum, deve ajudar a checar os argumentos dessa discussão e seu respaldo na realidade.

Perguntas de pesquisa

- Quais são as características em comum dos álbuns indicados e vencedores do Grammy? Popularidade, relevância/influência, crítica positiva, gênero musical?
- Quais são as semelhanças e diferenças entre álbuns indicados ao Grammy e aqueles aclamados pela crítica ou populares?

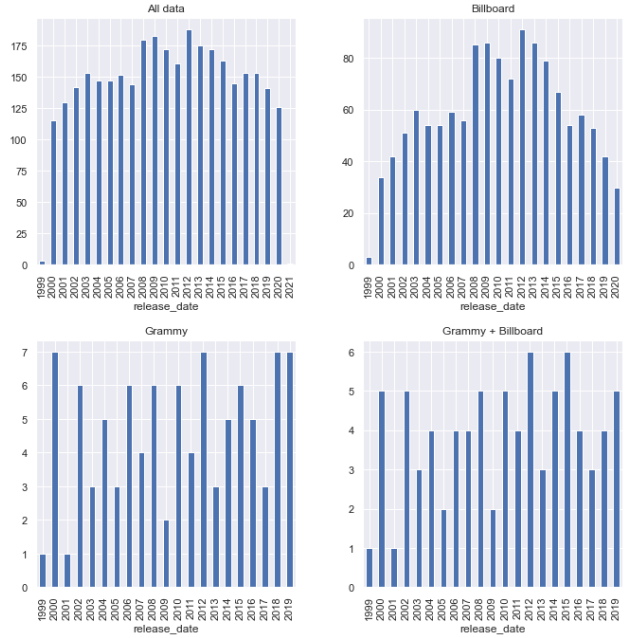
Hipóteses

- Os álbuns indicados e vencedores do Grammy tem muito mais relação com sua popularidade que sua qualidade técnica (crítica positiva)
- Os gêneros dos álbuns indicados ao Grammy têm pouca diversidade com relação aos álbuns aclamados pela crítica

Dados

Para as métricas necessárias, coletamos dados de 3 fontes diferentes, Billboard, Metacritic e dos álbuns indicados ao Grammy. Para todas as fontes foram utilizados os anos de 2000 até 2020. Na Billboard buscamos os top 200 álbuns mais vendidos dos EUA no ano. No Metacritic buscamos os top 100 álbuns melhores avaliados do ano e também os álbuns indicados ao Grammy. Para o Grammy, os dados foram obtidos manualmente, sendo obtido o ano, artista, álbum e se o álbum venceu ou não.

A partir do nome dos álbuns e dos artistas, fizemos um parser para a URL do Metacritic. Com isso baixamos as páginas de todos álbuns encontrados e, a partir de web scrapping, retiramos as seguintes informações: nome do artista; nome do álbum; link para imagem do álbum; gêneros; data de lançamento; metascore (nota da crítica); userscore (nota do usuário).
Com todos esses dados, agrupamos os resultados em um JSON para processamento. Os álbuns que não foram encontrados no Metacritic não foram utilizados.

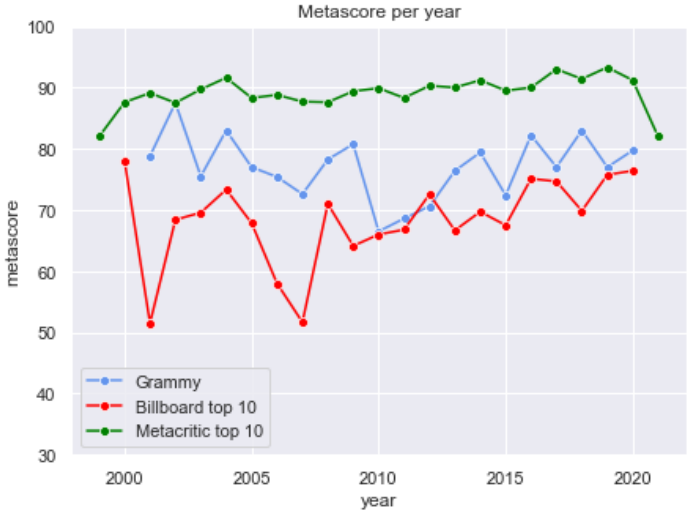


Número de álbuns por ano para cada base de dado

O ideal para Billboard seria 200 álbuns por ano, enquanto que para Grammy aproximadamente 5. Consideramos que o número de dados obtidos (cerca de 50% do total para Billboard e 90% do total para o Grammy) são o suficiente para nossas análises.

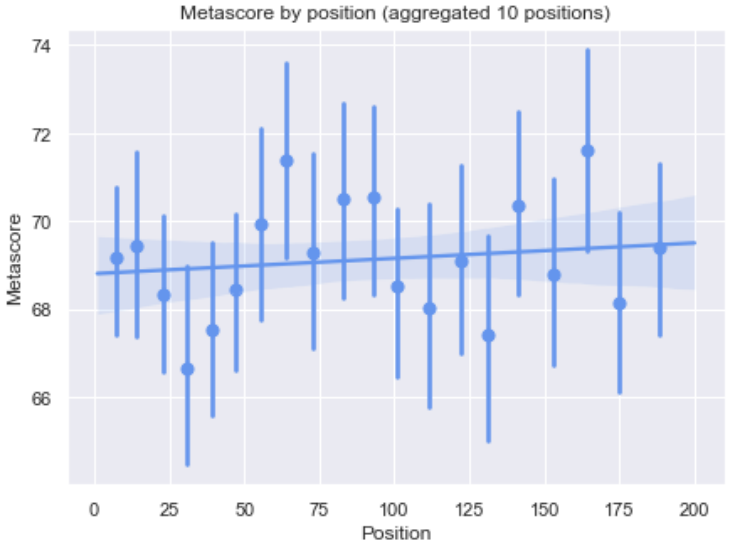
Análise exploratória

A partir dos álbuns, algumas análises básicas foram feitas. A média do metascore em função do ano para o top 10 Billboard, top 10 Metacritic e os indicados ao Grammy.

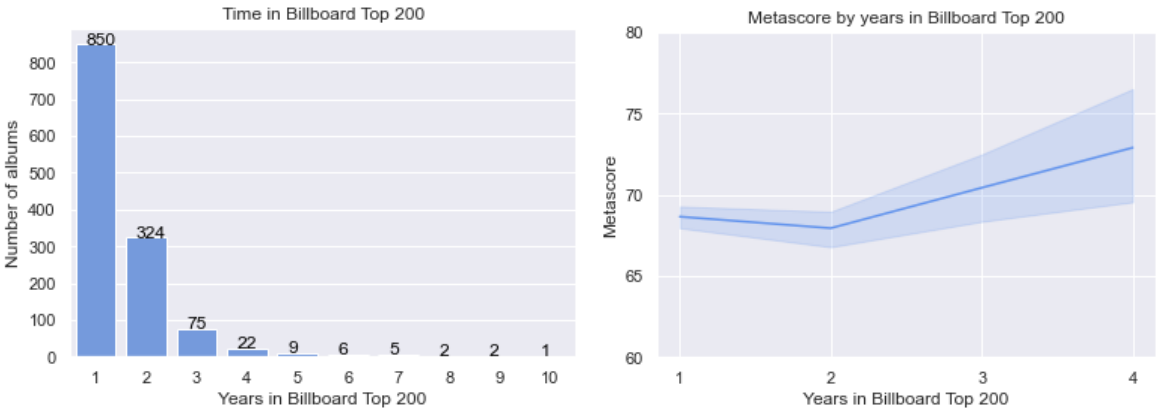


Média do metascore em função do ano

É possível observar que a Billboard tem uma média abaixo do Grammy na grande maioria dos anos, porém de 2010 até 2020 essas médias se aproximam e apresentam uma leve tendência de aumento.



Metascore em função da posição na Billboard



Metascore e número de anos na Billboard

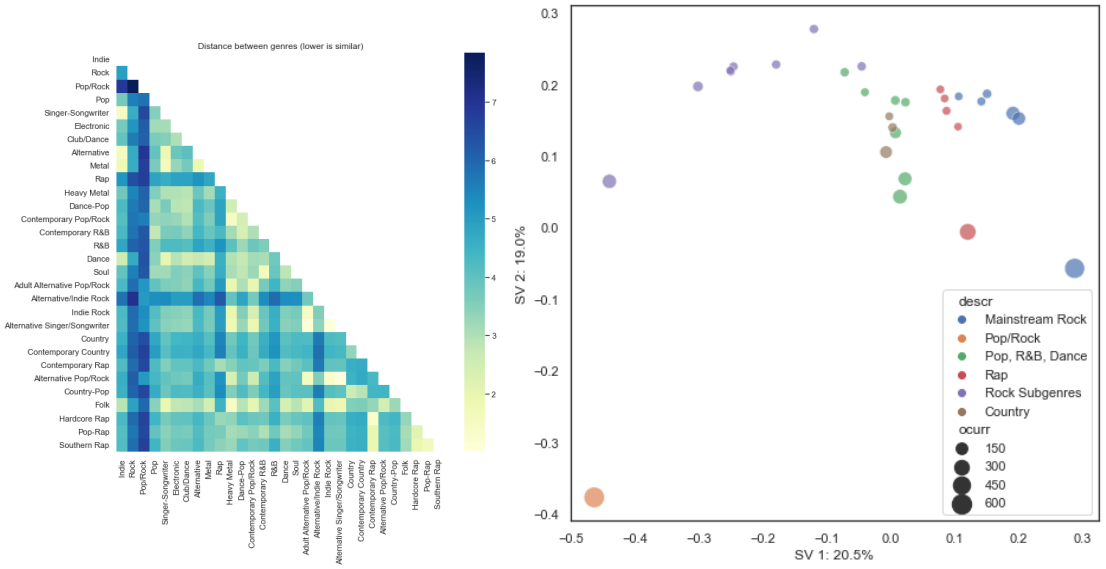
Posição na Billboard por ano dos álbuns indicados ao Grammy e top 10 Metacritic

Com relação ao metascore e a posição na Billboard, não é possível observar nenhuma relação entre os dois, com a regressão linear apresentando um R de 0.019 e um P de 0.40. Já ao se fazer uma regressão relacionando o metascore e o número de anos na Billboard, é possível ver uma relação entre ambos, com a nota aumentando conforme o maior número de anos.
Uma última análise comparando a posição na Billboard dos álbuns top 10 Metacritic e dos indicados ao Grammy mostra que há uma presença muito maior dos álbuns indicados. Isso leva a crer que há uma relação muito forte entre popularidade e indicação ao Grammy, ou vice-versa. Mais de 90% dos álbuns do Grammy estão na Billboard, enquanto que para o top 10 Metacritic (com o dobro de álbuns), apenas cerca de 30% estão no top 200.

Modelos e Resultados

Gêneros em cada fonte de dados

Para análise dos gêneros, primeiramente foi feita uma filtragem com os top 30 gêneros mais comuns. Após isso, uma matriz de coocorrência dos gêneros em relação aos álbuns que estão presentes. A partir disso, foram calculados os vetores (ou posição) de cada gênero e então normalizados a partir do desvio padrão e da média.

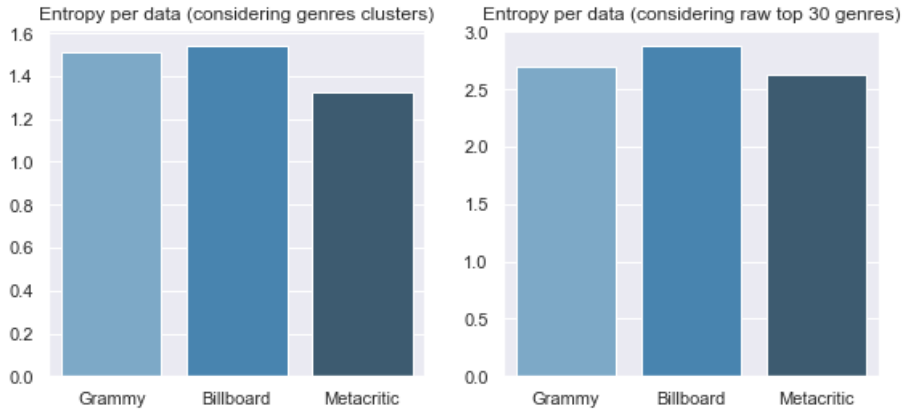


Matriz de distância entre gêneros e representação dos clusters a partir dos dois autovetores de maior autovalor utilizando SDV

Com o valor do vetor de cada gênero, foi utilizado o algoritmo KMeans para clusterização deles, definindo o número de clusters como 6 (definido após análises quantitativas da inércia e qualitativas).
Os clusters encontrados podem ser caracterizados como:

- **Subgêneros do Rock**: Heavy Metal, Contemporary Pop/Rock, Adult Alternative Pop/Rock, Alternative/Indie Rock, Indie Rock, Alternative Singer/Songwriter, Alternative Pop/Rock, Folk;
- **Pop/Rock**: Pop/Rock;
- **Rap:** Rap, Contemporary Rap, Hardcore Rap, Pop-Rap, Southern Rap;
- **Country:** Country, Contemporary Country, Country-Pop;
- **Pop, R&B, Dance:** Pop, Electronic, Club/Dance, Dance-Pop, Contemporary R&B, R&B, Soul;
- **Rock Mainstream:** Indie, Rock, Singer-Songwriter, Alternative, Metal, Dance.

O outlier observado foi o Pop/Rock, pois a maior parte dos álbuns que esse é presente é o único, fazendo com que o gênero esteja distante de todos os outros.
A partir desses clusters e dos gêneros, foi calculada a entropia de cada base de dados, utilizando o top 10 pela Billboard e pelo Metacritic. Quanto menor a entropia, menor a diversidade.



Entropia dos dados (esquerda com clusters, direita com gêneros)

É possível perceber que o Metacritic tem a menor diversidade em ambos os casos, enquanto a Billboard e o Grammy tem entropias semelhantes.

Fatores para indicação ao Grammy

Foi feita uma regressão logística com o intuito de descobrir quais fatores influenciam mais na indicação de um álbum ao grammy. Foram considerados em princípio 3 fatores para a indicação dos álbuns ao Grammy: Nota do Metacritic (Metascore), posição na Billboard e mês de lançamento. A partir disso, foram feitas primeiramente 3 regressões utilizando cada fator individualmente, a fim de fazer uma análise sem a interferência dos demais fatores.

Com a regressão a partir do Metascore, observou-se que este fator não é significativo, tendo um *P-value* de 0.845, acima do valor máximo de 0.05 que indica significância. Em seguida, foi feita a regressão a partir da posição na Billboard, observando que este fator é significativo, tendo um *P-value* de 0.000 (arredondado) e acima de 0.05. Por fim, foi feita a regressão a partir do fator de mês de lançamento, verificando que este não é significativo, tendo um *P-value* de 0.314.

Após as regressões utilizando os fatores individualmente, foi feita nova uma regressão logística, desta vez com todos os fatores (figura abaixo). Analisando os resultados, foi visto que nessa regressão o Metascore torna-se significativo, indicando uma possível interação entre a nota do Metascore e a posição na Billboard. Foi visto também que o mês de lançamento continua sem significância. Por fim, observou-se que o R é baixo na regressão logística realizada.

Com isso, foi possível calcular o *Odds Ratio* de cada fator. Nele, viu-se que o Metascore tem o maior valor, seguido pela posição na Billboard e por fim o mês de lançamento. Para o Metascore, o valor obtido de 1.088116 significa basicamente que, para cada unidade que aumenta da nota do Metascore, a chance do respectivo álbum ser indicado aumenta em 8.81%. Para os outros fatores isto ocorre de forma análoga, levando em consideração que valores abaixo de 1 diminuem as chances do álbum ser indicado ao invés de aumentar.

Logit Regression Results

Dep. Variable:	grammy_indication	No. Observations:	3246
Model:	Logit	Df Residuals:	3242
Method:	MLE	Df Model:	3
Date:	Sat, 11 Dec 2021	Pseudo R-squ.:	0.2369
Time:	19:08:32	Log-Likelihood:	-332.73
converged:	True	LL-Null:	-436.05
Covariance Type:	nonrobust	LLR p-value:	1.554e-44

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-7.4348	0.986	-7.538	0.000	-9.368	-5.502
metascore	0.0844	0.013	6.622	0.000	0.059	0.109
best_rank_billboard	-0.0223	0.002	-13.118	0.000	-0.026	-0.019
release_month	-0.0049	0.033	-0.148	0.882	-0.070	0.060

	5%	95%	Odds Ratio
Intercept	0.000085	0.004080	0.000590
metascore	1.061257	1.115655	1.088116
best_rank_billboard	0.974676	0.981197	0.977931
release_month	0.932664	1.061739	0.995111

Também foram feitos dois gráficos *boxplot* com os quais foi possível concluir que em álbuns indicados ao grammy, as posições na Billboard tendem a ser melhores do que em álbuns não indicados. Já para a nota no Metascore, não foi possível observar tal tendência, estando na mesma faixa de valores tanto os álbuns indicados ao grammy quanto os não indicados ao grammy.

Respostas às hipóteses

- **Os álbuns indicados e vencedores do Grammy tem muito mais relação com sua popularidade que sua qualidade técnica (crítica positiva)?**

Sim, a popularidade tem uma grande influência na indicação ao Grammy (ou vice-versa), maior que a avaliação crítica

- **Os gêneros dos álbuns indicados ao Grammy têm pouca diversidade com relação aos álbuns aclamados pela crítica?**

Não, na verdade a diversidade entre álbuns aclamados é menor que álbuns indicados ao Grammy ou populares (na Billboard)

Limitações

- Uma maior base de dados, com mais informações e maior intervalo de tempo, poderia permitir uma análise mais precisa
- A análise e clusterização de gênero foi feita apenas utilizando as ocorrências nos álbuns. Adicionar outras informações para essa clusterização permitiria uma melhor caracterização dos clusters.
- A modelagem de popularidade e de avaliação crítica dos álbuns é relativamente simples. Considerar outros fatores além da posição da Billboard e a nota no Metacritic poderia refletir melhor esses fatores.
- Há vários outros caminhos de análises a partir dos dados obtidos que podem ser explorados ainda (e.g. periodicidade de lançamento de álbuns por artista e influência disso nos fatores apresentados).