به نام خدا

حسین رضائی (۹۸۲۲۲۰۴۴) و آرمان کاظمی (۹۸۲۲۲۰۷۸)

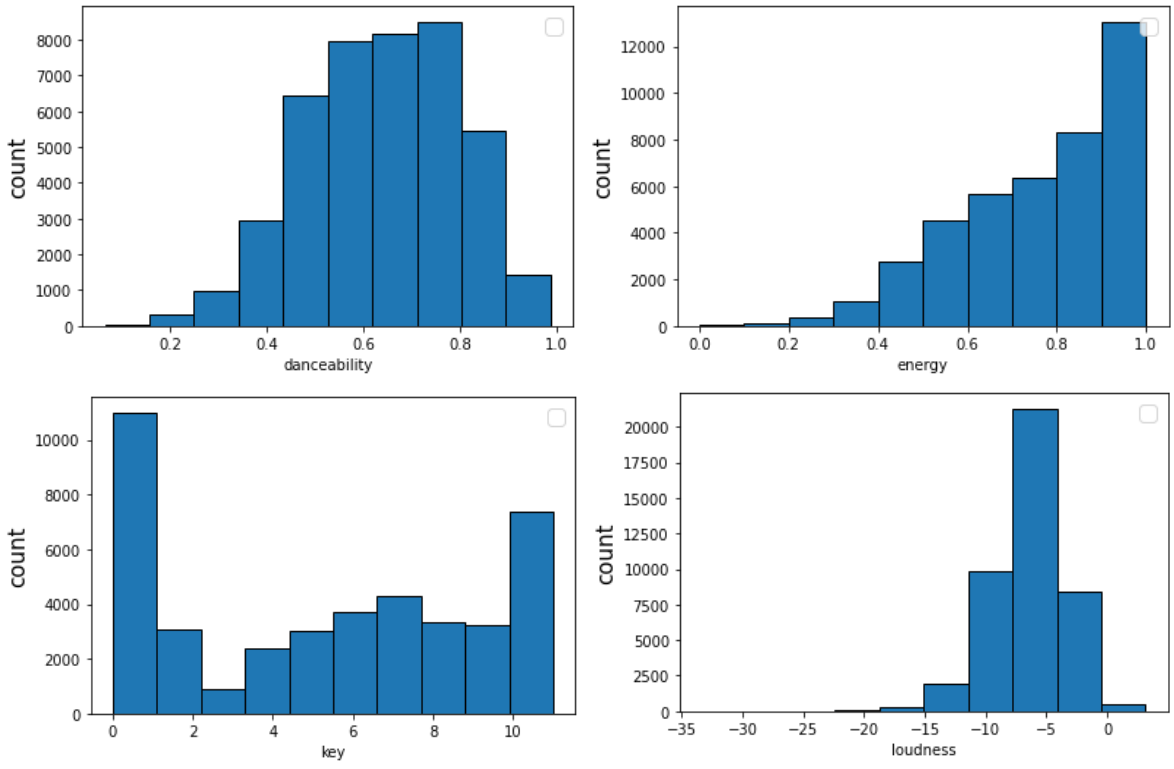
تاریخ: ۱۹/ ۰۹/ ۱۴۰۰

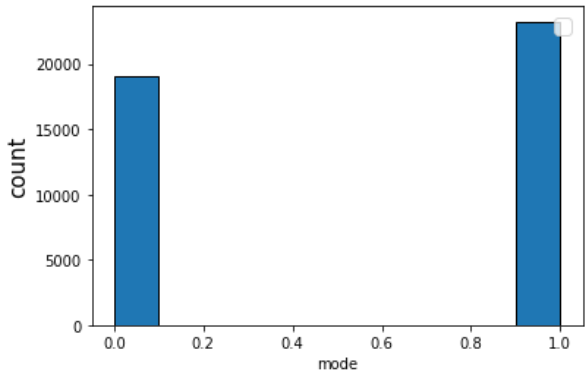
گزارش تمرین سری سوم

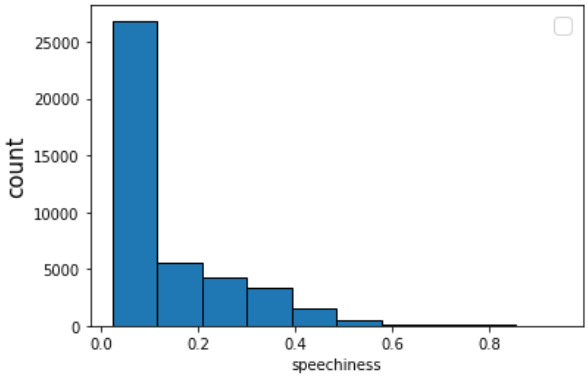
درس: مبانی علوم داده

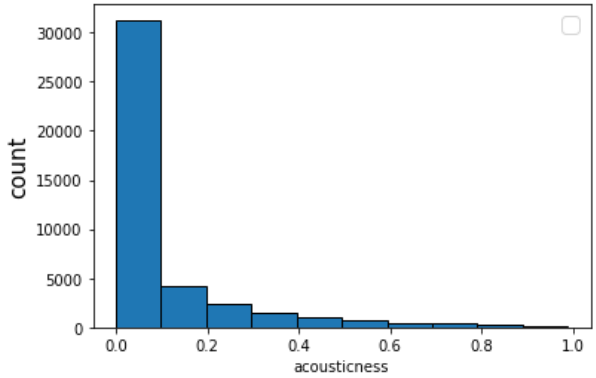
استاد درس: دکتر خردپیشه

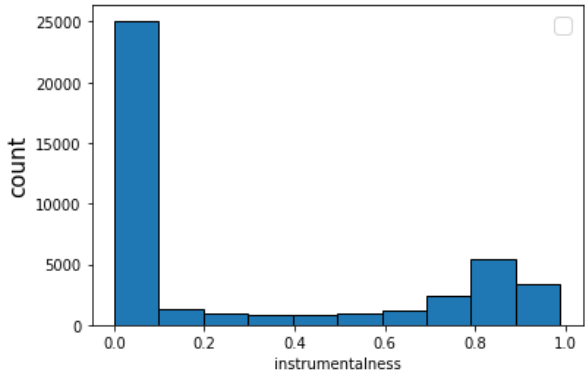
- بررسی دقیق داده‌ها در تمیز کردن و استفاده از آن‌ها اهمیت زیادی دارد. داده‌های ما دارای ۲۲ ستون و ۴۲۳۰۵ سطر است.
- سه ستون song_name, Unnamed: 0, title دارای مقادیر null نسبتا زیادی هستند، همچنین با توجه به ماهیتشان در خوشه بندی اهمیتی ندارند پس آن‌ها را حذف می‌کنیم.
- حال در بررسی سایر ستون‌ها ابتدا نمودار آن‌ها را رسم می‌کنیم:

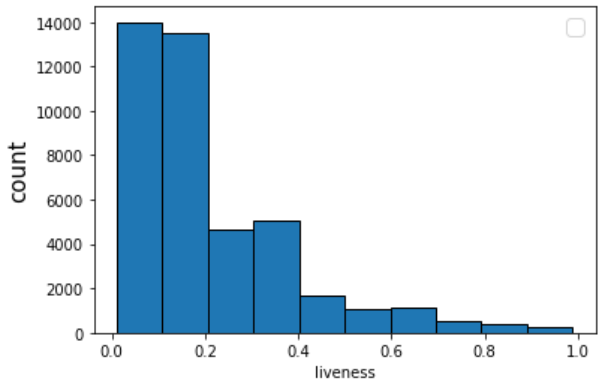


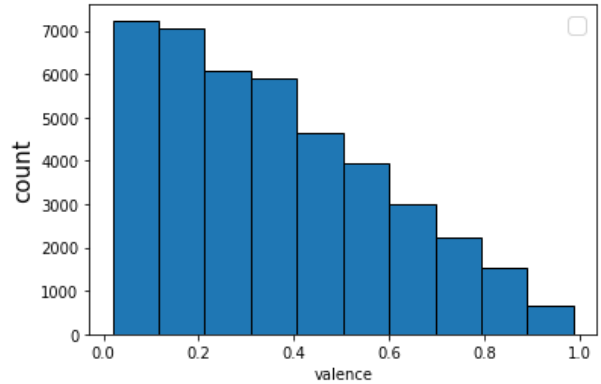

count
mode

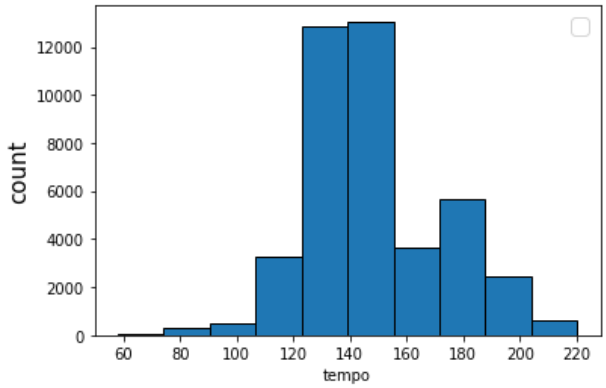

count
speechiness

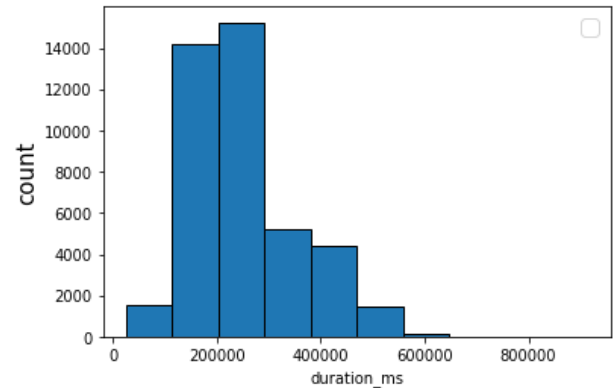

count
acousticness

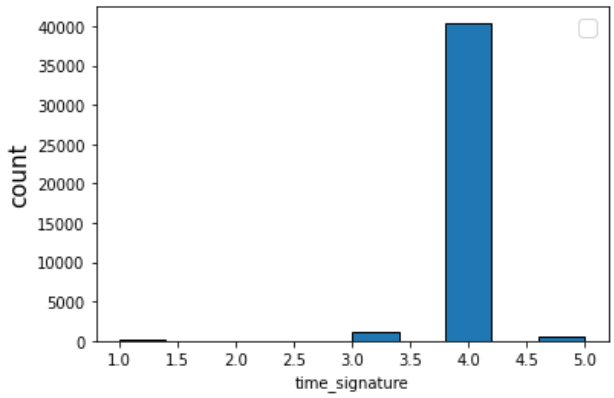

count
instrumentalness


count
liveness

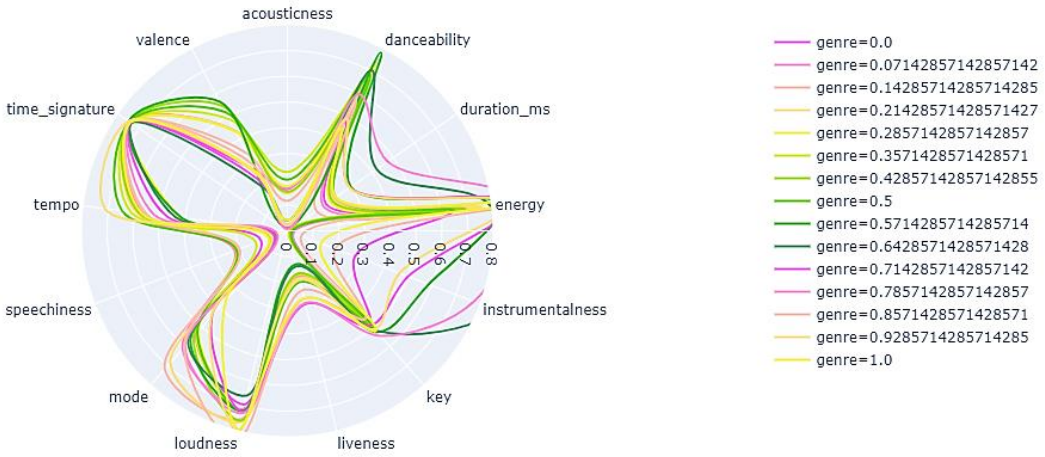

count
valence


count
tempo

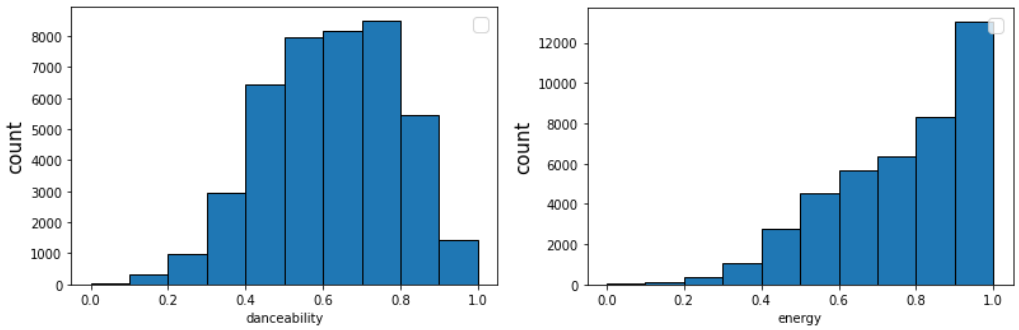

count
duration_ms

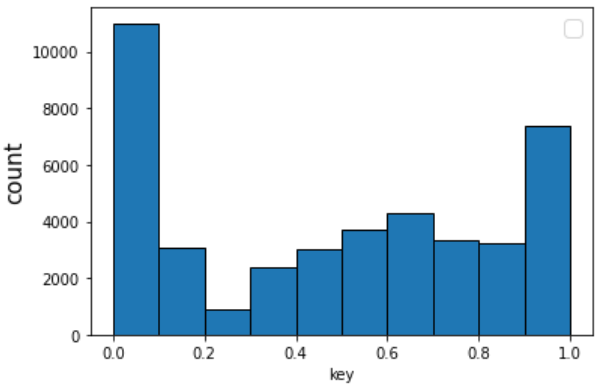


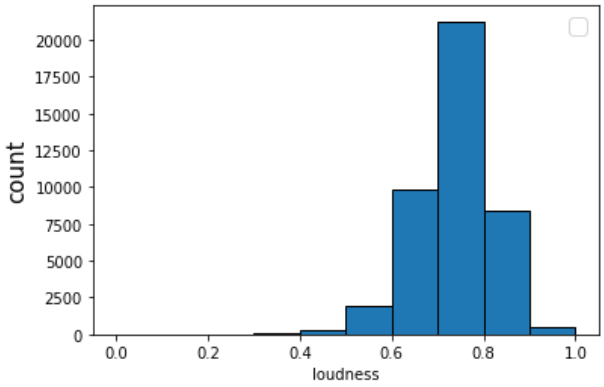
- همانطور که مشخص است داده‌ها نرمال و استاندارد شده نیستند و باید آن‌ها را نرمال کنیم سپس نمودارها را رسم کنیم. منتهی ابتدا نگاهی به چارت میانگین موسیقی‌ها به تفکیک ژانرها توجه می‌کنیم.

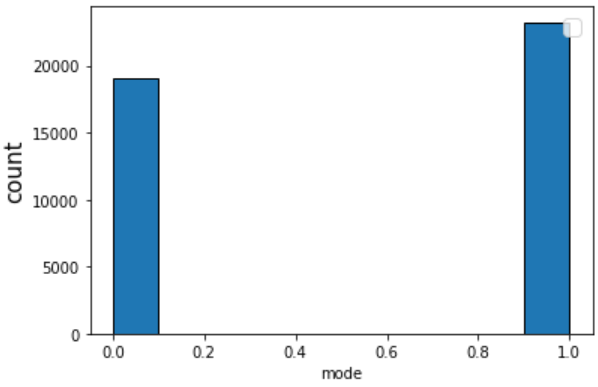


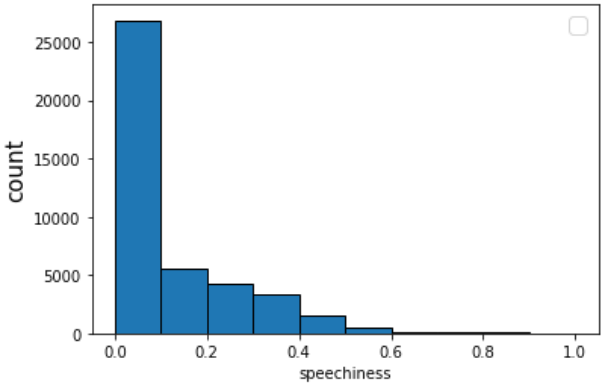
- همانطور که مشخص است time_signature ژانرهای مختلف تقریبا برابر است پس ستون مناسبی برای خوشه بندی نیست. هچنین مشخصه duration_ms با توجه به ماهیتش نباید اهمیت بخش باشد و می‌توانیم آن را در نظر نگیریم. حال به نمودارهای نرمال شده می‌پردازیم.

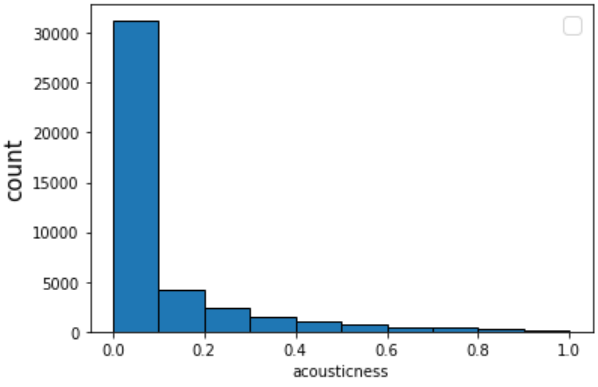


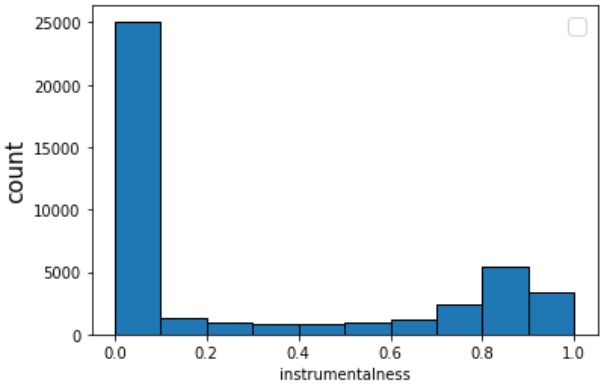

count
key

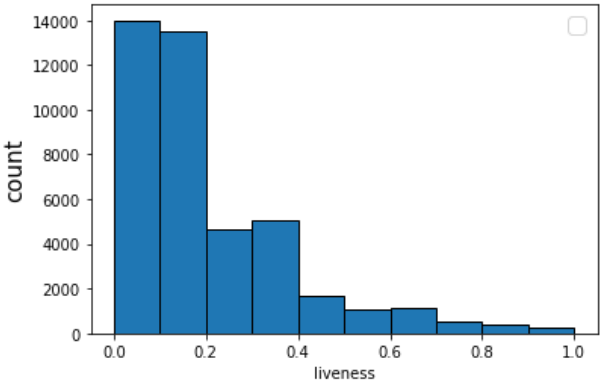

count
loudness

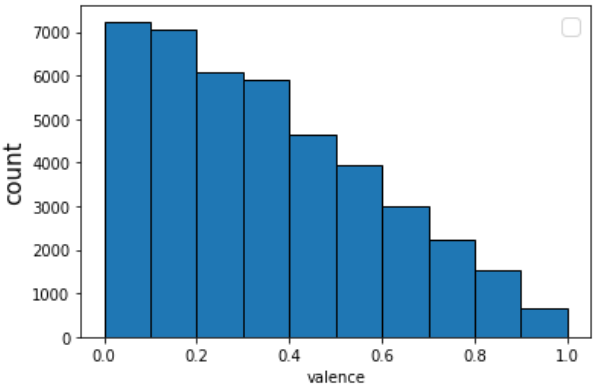

count
mode

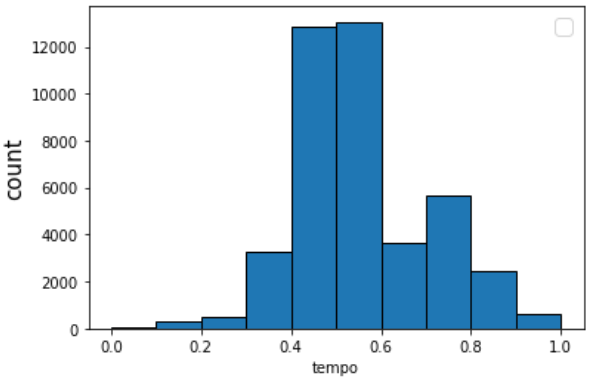

count
speechiness


count
acousticness


count
instrumentalness


count
liveness


count
valence

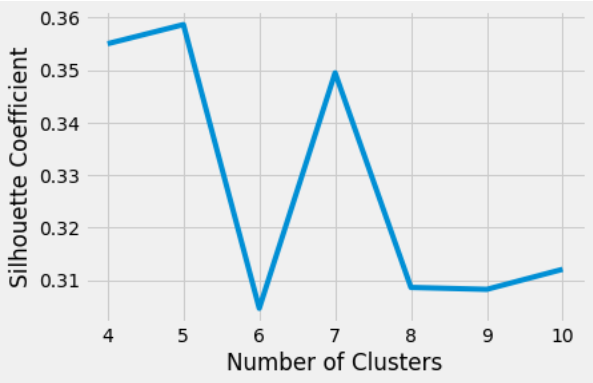


- حال بین ستون‌های باقی مانده خوشه بندی می‌کنیم.

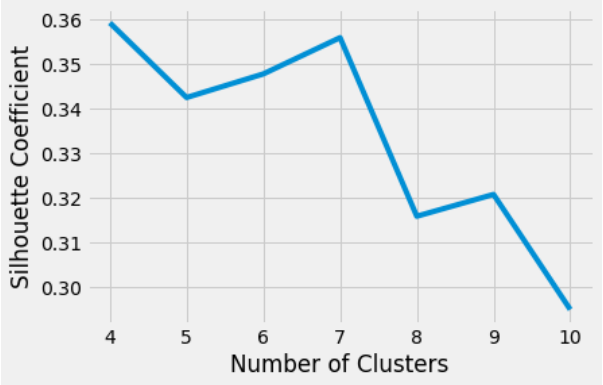
همانطور که مشخص است خوشه بندی با این تعداد ستون مناسب نیست و مقداری که به هر خوشه بندی نسبت داده شده است مقدار بسیار پایینی‌است که نشان می‌دهد خوشه بندی ما کیفیت مناسبی ندارد، پس به حذف چند ستون می‌پردازیم.

- حال بین ستون‌های زیر انتخاب‌های سه تایی و چهارتایی را بررسی می‌کنیم و در هر مورد مقدار k-means را با تعداد خوشه‌های ۴ تا ۱۱ تست می‌کنیم تا به انتخاب‌های مناسب برسیم.

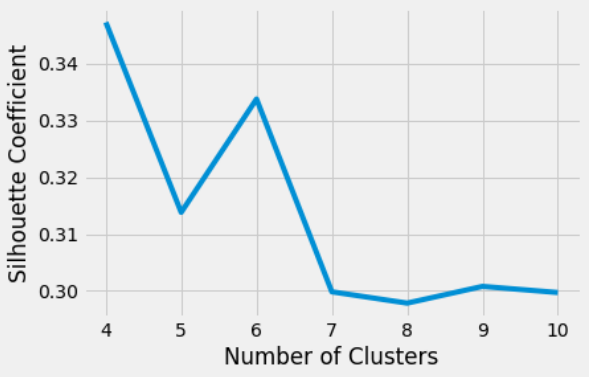
'energy', 'loudness', 'speechiness', 'valence', 'tempo', 'instrumentalness', 'acousticness'



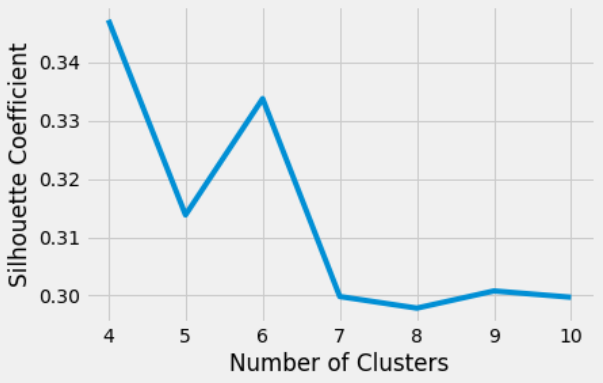
('energy', 'speechiness', 'valence')



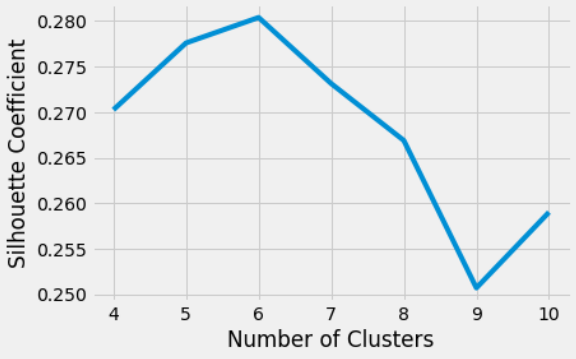
('energy', 'speechiness', 'tempo')



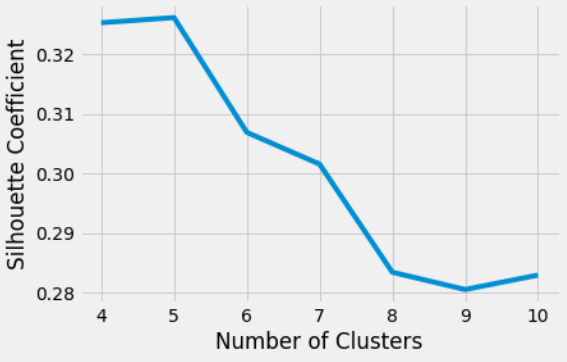
('danceability', 'speechiness', 'tempo')



('energy', 'loudness', 'valence')

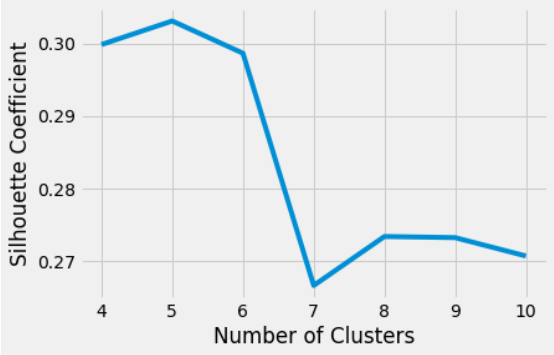


('danceability', 'energy', 'valence')



('energy', 'tempo', 'valence')

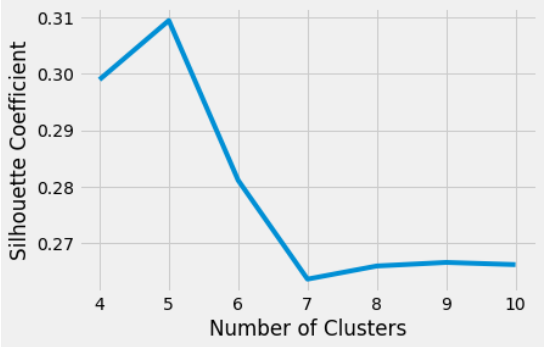
('danceability', 'energy', 'loloudness')



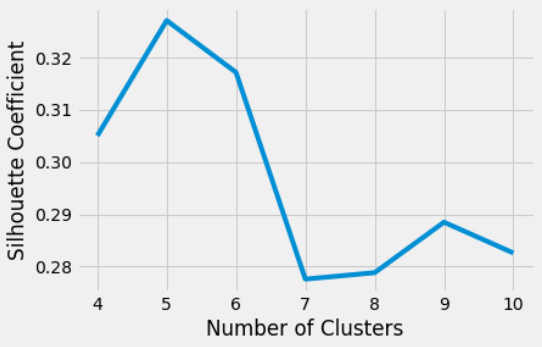
('danceability', 'energy', 'tempo')

('energy', 'loudness', 'tempo')

('loudness', 'tempo', 'valence')



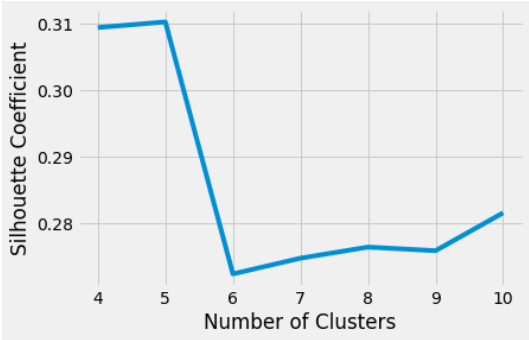
('danceability', 'loudness', 'speechiness')



('danceability', 'energy', 'speechiness')

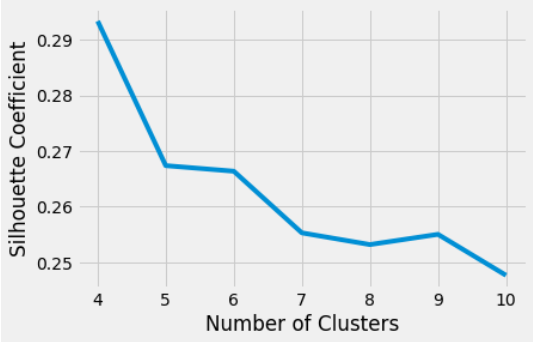
('energy', 'loudness', 'speechiness')

('danceability', 'tempo', 'valence')



('loudness', 'speechiness', 'valence')

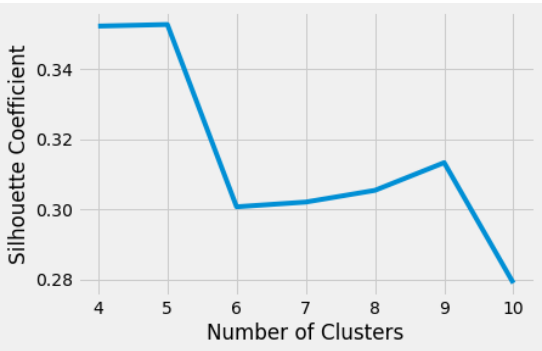
('danceability', 'loudness', 'valence')



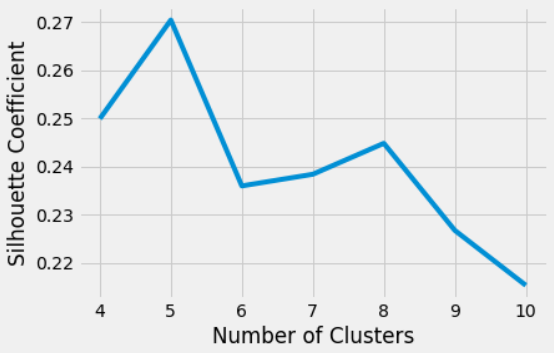
('danceability', 'loudness', 'tempo')

('loudness', 'speechiness', 'tempo')

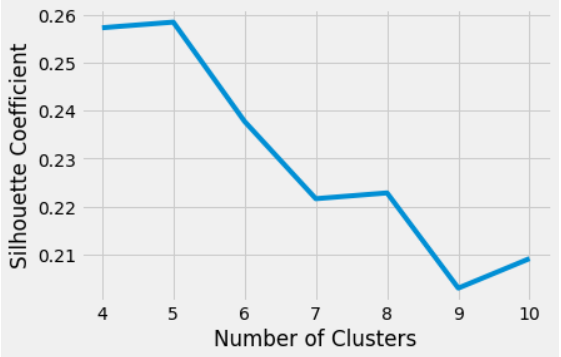
('danceability', 'speechiness', 'valence')



('speechiness', 'tempo', 'valence')

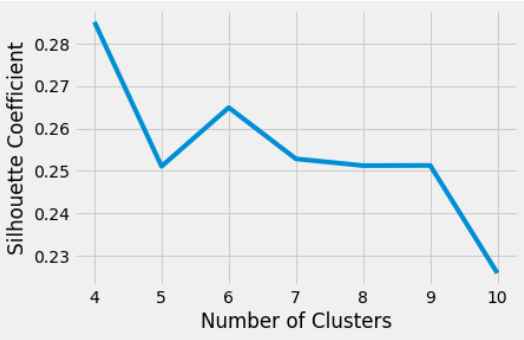


('danceability', 'energy', 'loudness', 'speechiness')



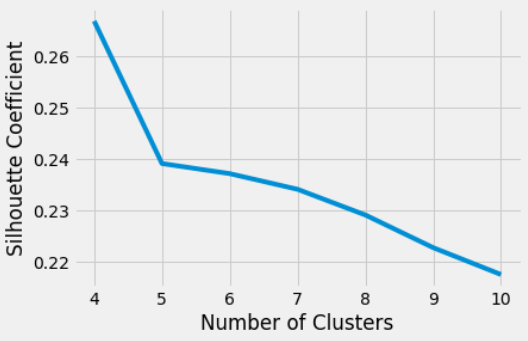
('danceability', 'loudness', 'speechiness', 'valence')

('energy', 'speechiness', 'tempo', 'valence')

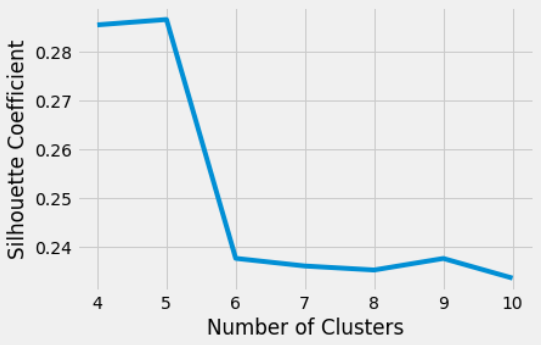


('danceability', 'energy', 'speechiness', 'tempo')

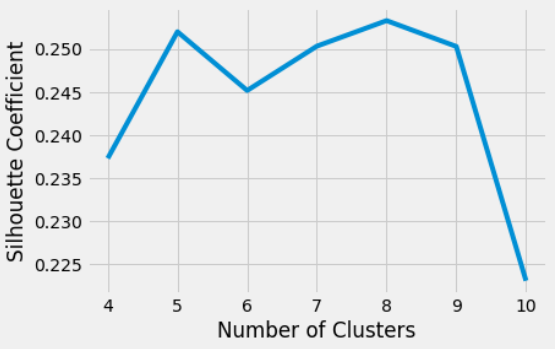
('energy', 'loudness', 'speechiness', 'tempo')



('energy', 'loudness', 'tempo', 'valence')

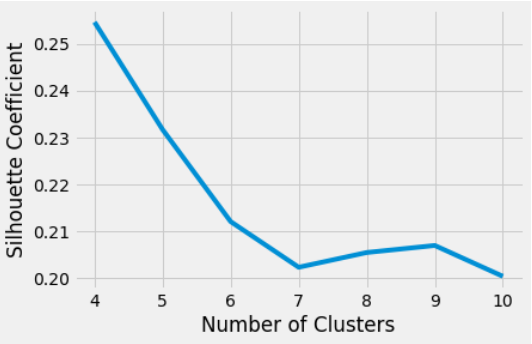


('danceability', 'energy', 'speechiness', 'valence')



('danceability', 'speechiness', 'tempo', 'valence')

('energy', 'loudness', 'speechiness', 'valence')



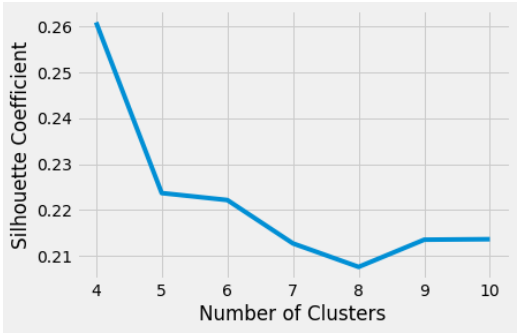
('danceability', 'loudness', 'tempo', 'valence')

('danceability', 'energy', 'loudness', 'valence')

('danceability', 'loudness', 'speechiness', 'tempo')

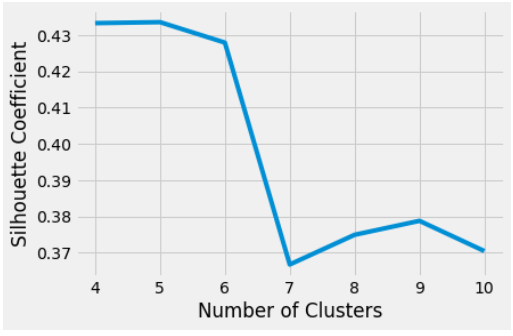
('danceability', 'energy', 'tempo', 'valence')

('loudness', 'speechiness', 'tempo', 'valence')



('danceability', 'energy', 'loudness', 'tempo')

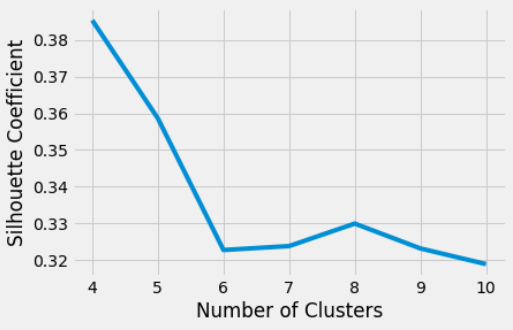
('energy', 'tempo', 'acousticness')



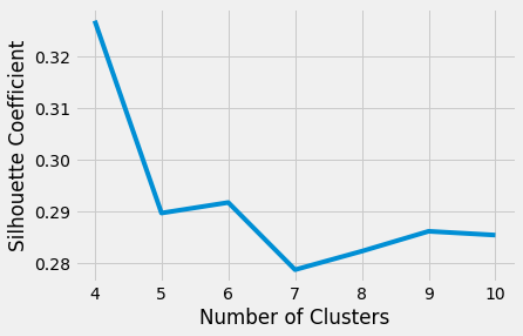
('energy', 'speechiness', 'acousticness')

('tempo', 'valence', 'acousticness')

('loudness', 'tempo', 'acousticness')



('energy', 'valence', 'acousticness')

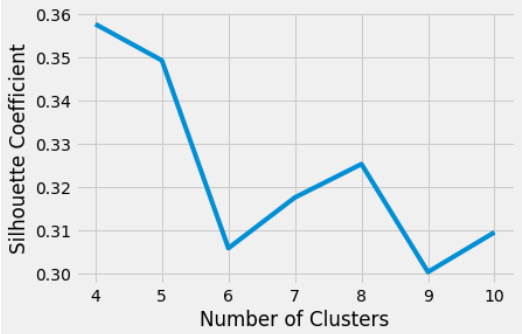


('loudness', 'valence', 'acousticness')

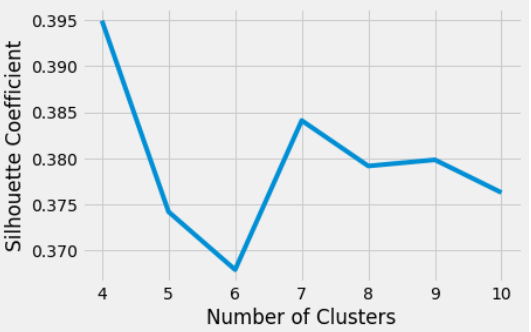
('speechiness', 'valence', 'acousticness')

('speechiness', 'tempo', 'acousticness')

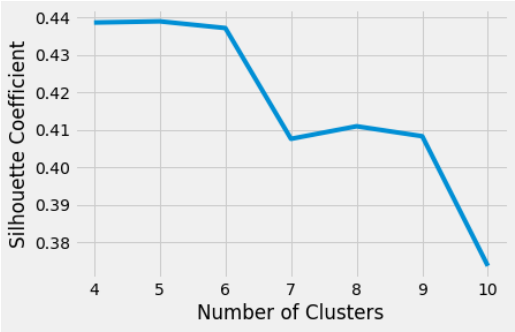
('energy', 'loudness', 'acousticness')



('loudness', 'speechiness', 'acousticness')



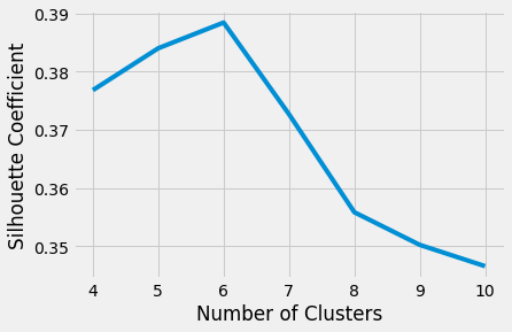
('energy', 'tempo', 'instrumentalness')



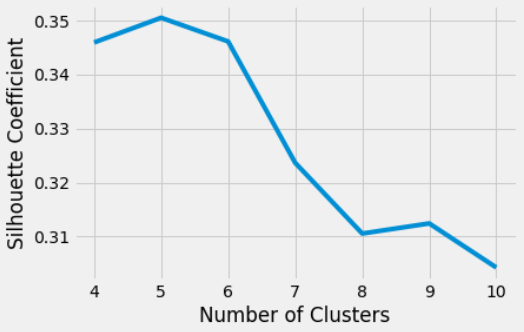
('energy', 'speechiness', 'instrumentalness')

('tempo', 'valence', 'instrumentalness')

('loudness', 'tempo', 'instrumentalness')

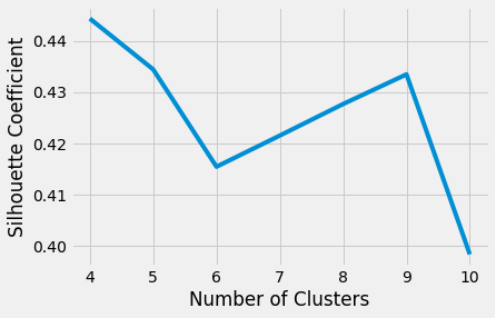


('energy', 'valence', 'instrumentalness')

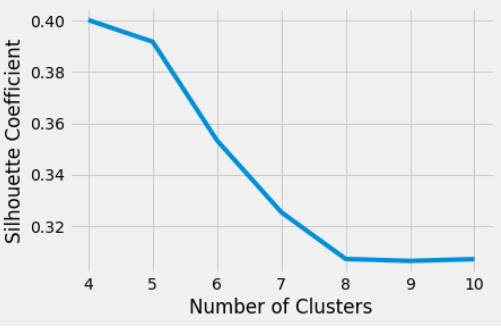


('loudness', 'valence', 'instrumentalness')

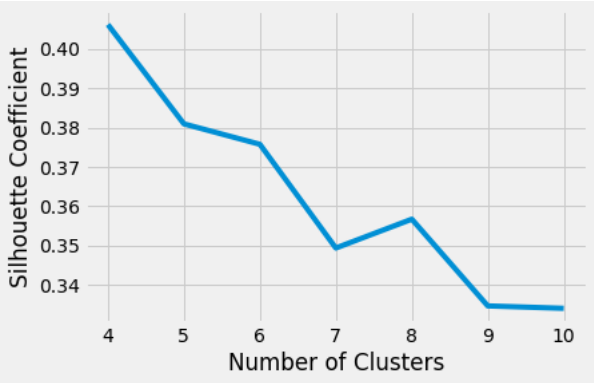
('speechiness', 'valence', 'instrumentalness')



('speechiness', 'tempo', 'instrumentalness')

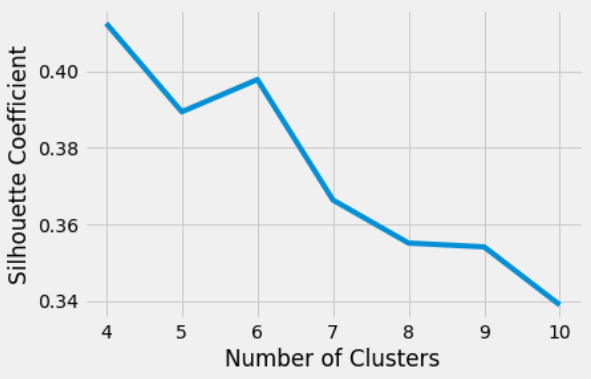


('energy', 'loudness', 'instrumentalness')

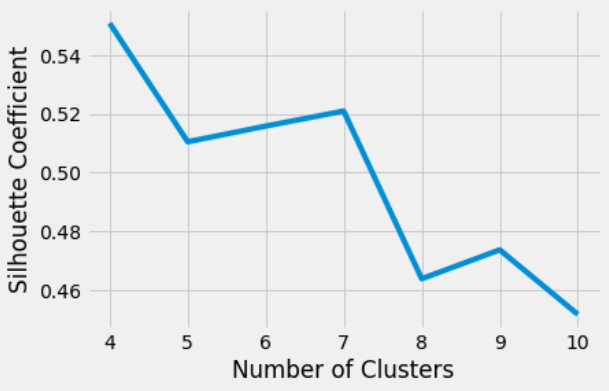


('loudness', 'speechiness', 'instrumentalness')

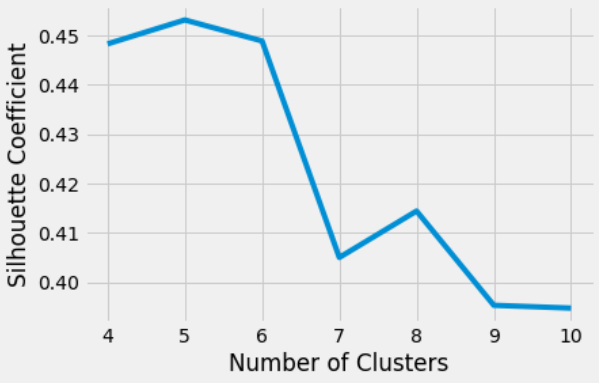
('energy', 'instrumentalness', 'acousticness')



('loudness', 'instrumentalness', 'acousticness')



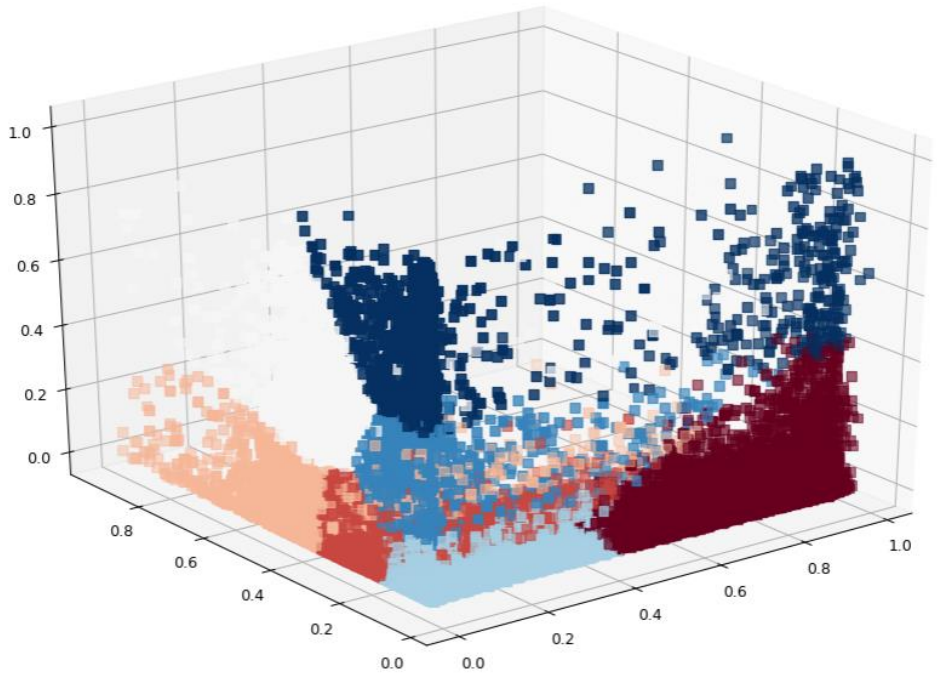
('speechiness', 'instrumentalness', 'acousticness')



('valence', 'instrumentalness', 'acousticness')

('tempo', 'instrumentalness', 'acousticness')

- همانطور که مشخص است بهترین انتخاب مربوط به خوشه بندی بر اساس سه ستون ' speechiness', 'instrumentalness', 'acousticness'' است. پس k-means را با ۷ خوشه بر اساس این سه ستون حساب کرده و مقادیر kmeans.pkl و scaler.pkl را ذخیره می‌کنیم. خوشه بندی هم اکنون داده‌ها به شکل زیر است.



- در نهایت ۲۵ موسیقی برتر هر خوشه را انتخاب می‌کنیم و تا از آن‌ها برای recommendation استفاده کنیم.
- در ادامه پلی لیست ورودی را گرفته و موسیقی‌های آن را با kmeans.pkl و scaler.pkl خوشه بندی می‌کنیم تا مشخص شود مربوط به کدام خوشه هستند.
- سپس تعداد موسیقی‌های هر خوشه را بررسی می‌کنیم. چون ۷ خوشه داریم اگر نسبت کمتر از یک هفتم باشد آن خوشه در نظر گرفته نمی‌شود و باقی خوشه‌ها از بیشترین تعداد موسیقی به کمترین مرتب می‌شوند و به همان نسبت پلی لیست از آن خوشه ارائه می‌دهیم. بدین شکل مشخص می‌شود که از هر خوشه چند پلی لیست ارائه می‌شود.
- موسیقی‌های هر خوشه با توجه به تعداد پلی لیست‌هایی که قرار است از آن خوشه خارج شود انتخاب می‌شود به این شکل که به طور مثال اگر ۳ پلی لیست از یک خوشه داشته باشیم، موسیقی‌های ضریب ۳ در پلی لیست اول، 3k+1 در دومی و 3k+2 در سومی قرار می‌گیرند.