Clustering - What is the difference in hierarchical or partitional in clustering?

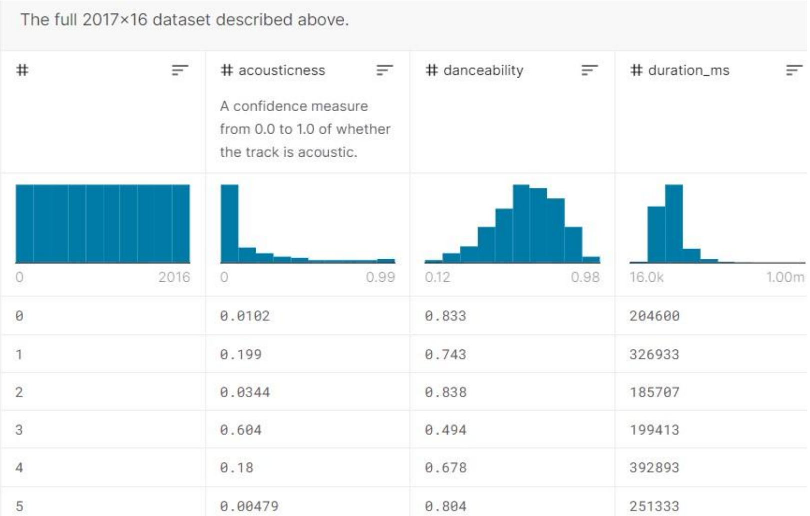
By: Edric He, Peggy Lin

Introduction

- Investigate the usability between hierarchical and partitional clustering.

- Evaluate the data in different method: Normalization, delete outlier, finding accuracy.

Dataset



The full 2017×16 dataset described above.

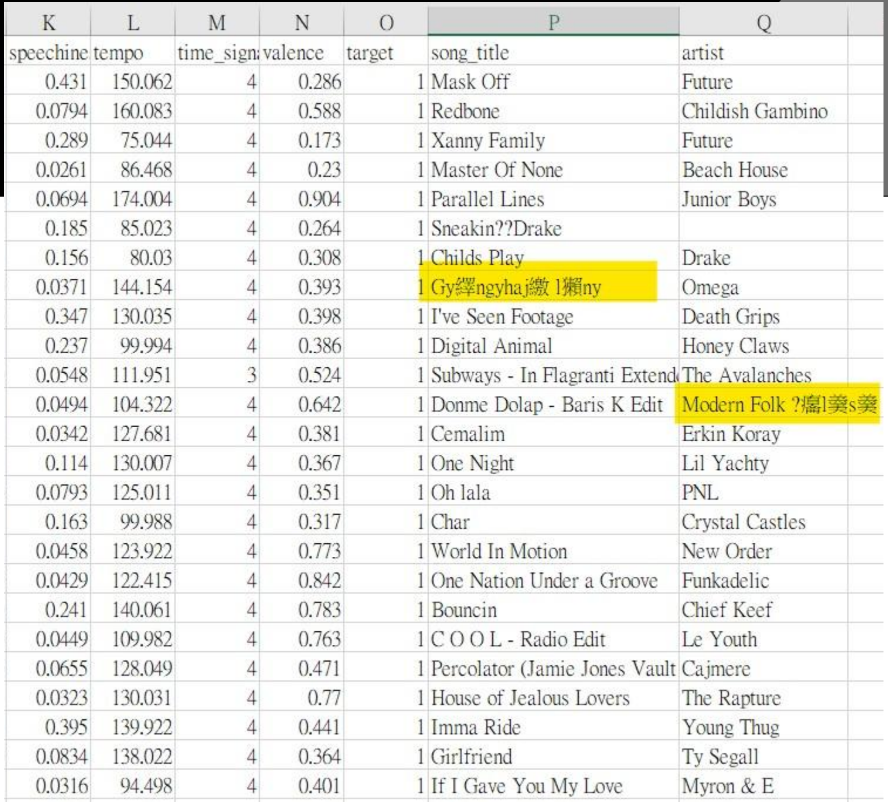
#	# acousticness	# danceability	# duration_ms
	A confidence measure from 0.0 to 1.0 of whether the track is acoustic.		
0	0.0102	0.833	204600
1	0.199	0.743	326933
2	0.0344	0.838	185707
3	0.604	0.494	199413
4	0.18	0.678	392893
5	0.00479	0.804	251333

Figure. 1

- (Figure 1) **16** columns. **13** of which are song attributes
- one column for song name, one for artist
- and a column called "target" which is the label for the song.(will be deleted later)
- Here are the 13 track attributes: acousticness, danceability, duration(*ms*), *energy, instrumentalness, key, liveness, loudness, mode, speechiness, tempo, time* signature, valence

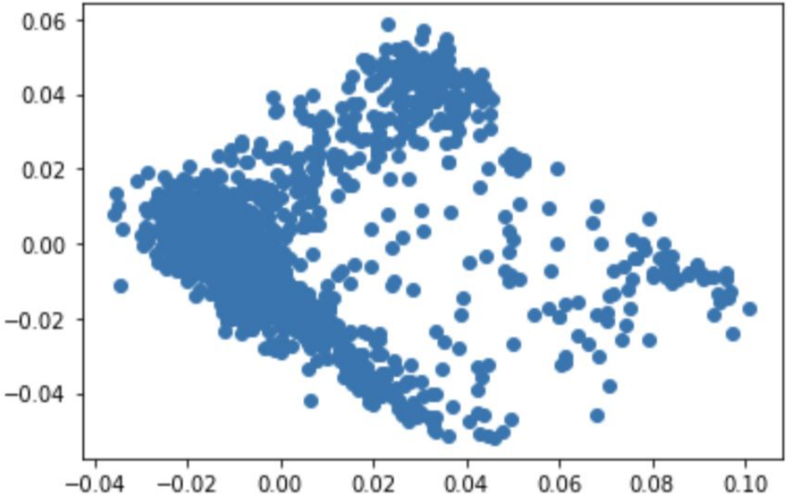
Preprocessing

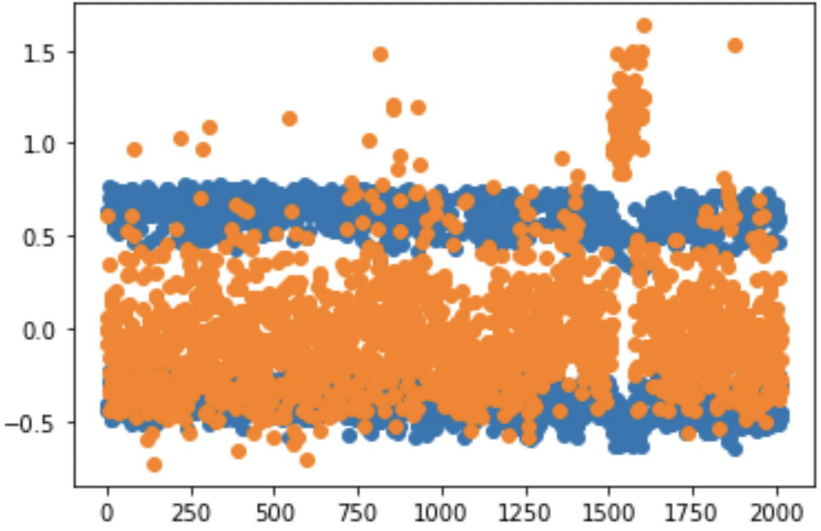
- Potential issues with data
  - Missing data, errors, inconsistency,availability



K	L	M	N	O	P	Q
speechine	tempo	time_sign	valence	target	song_title	artist
0.431	150.062	4	0.286	1	Mask Off	Future
0.0794	160.083	4	0.588	1	Redbone	Childish Gambino
0.289	75.044	4	0.173	1	Xanny Family	Future
0.0261	86.468	4	0.23	1	Master Of None	Beach House
0.0694	174.004	4	0.904	1	Parallel Lines	Junior Boys
0.185	85.023	4	0.264	1	Sneakin??Drake	
0.156	80.03	4	0.308	1	Childs Play	Drake
0.0371	144.154	4	0.393	1	Gy繹ngyhaj繳 l獺ny	Omega
0.347	130.035	4	0.398	1	I've Seen Footage	Death Grips
0.237	99.994	4	0.386	1	Digital Animal	Honey Claws
0.0548	111.951	3	0.524	1	Subways - In Flagranti Extend	The Avalanches
0.0494	104.322	4	0.642	1	Donme Dolap - Baris K Edit	Modern Folk ?癟l羹s羹
0.0342	127.681	4	0.381	1	Cemalim	Erkin Koray
0.114	130.007	4	0.367	1	One Night	Lil Yachty
0.0793	125.011	4	0.351	1	Oh lala	PNL
0.163	99.988	4	0.317	1	Char	Crystal Castles
0.0458	123.922	4	0.773	1	World In Motion	New Order
0.0429	122.415	4	0.842	1	One Nation Under a Groove	Funkadelic
0.241	140.061	4	0.783	1	Bouncin	Chief Keef
0.0449	109.982	4	0.763	1	C O O L - Radio Edit	Le Youth
0.0655	128.049	4	0.471	1	Percolator (Jamie Jones Vault	Cajmere
0.0323	130.031	4	0.77	1	House of Jealous Lovers	The Rapture
0.395	139.922	4	0.441	1	Imma Ride	Young Thug
0.0834	138.022	4	0.364	1	Girlfriend	Ty Segall
0.0316	94.498	4	0.401	1	If I Gave You My Love	Myron & E

Visualize - PCA





Normalize

- Sklearn
  - MinMaxscaler

$$x_{scaled} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

  - StandardScaler

Standardization:

$$z = \frac{x-\mu}{\sigma}$$

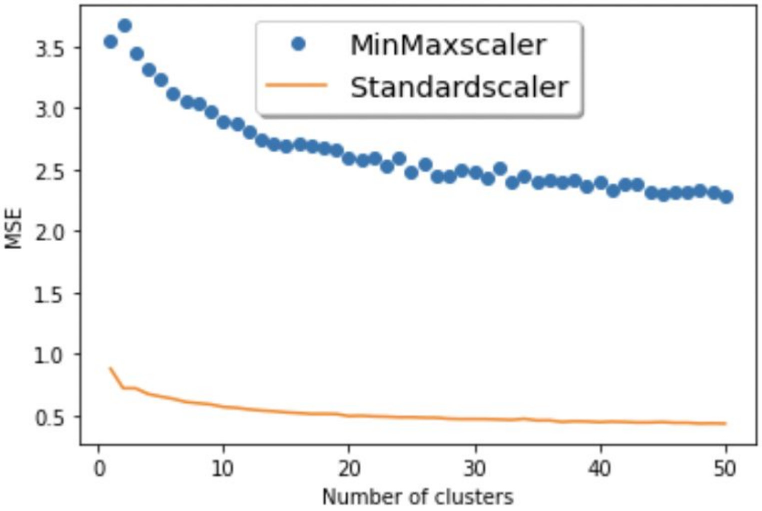
with mean:

$$\mu = \frac{1}{N}\sum_{i=1}^{N}(x_i)$$

and standard deviation:

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}$$

Normalize - in K-means

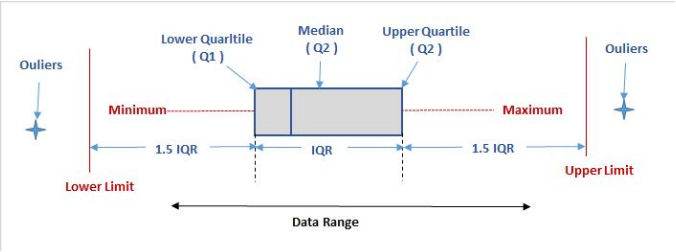


- Have different scaler so the square error will be different.

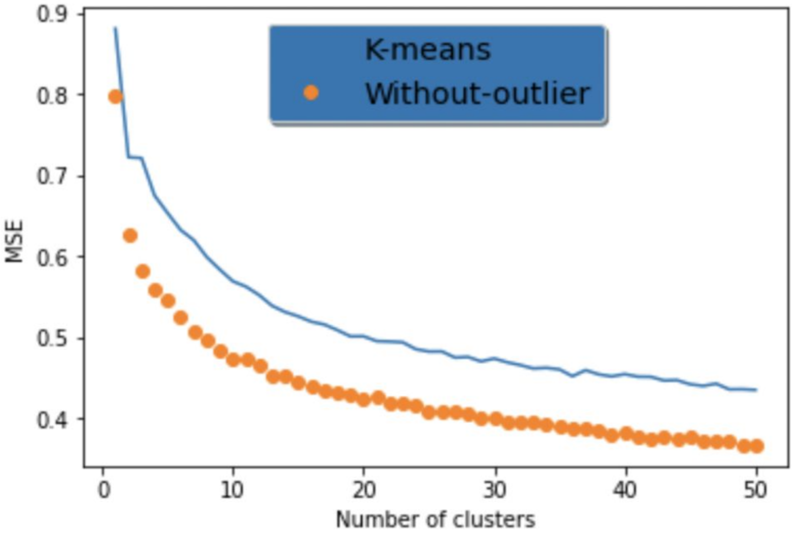
Outlier: IQR - distance based

- Interquartile Range Rule
  - Normal outlier:
    - Number bigger than upper quartile + 1.5 IQR or smaller than lower quartile + 1.5 IQR
  - Strong outlier:
    - Number bigger than upper quartile + 3 IQR or smaller than lower quartile + 3 IQR

- standard deviation graph



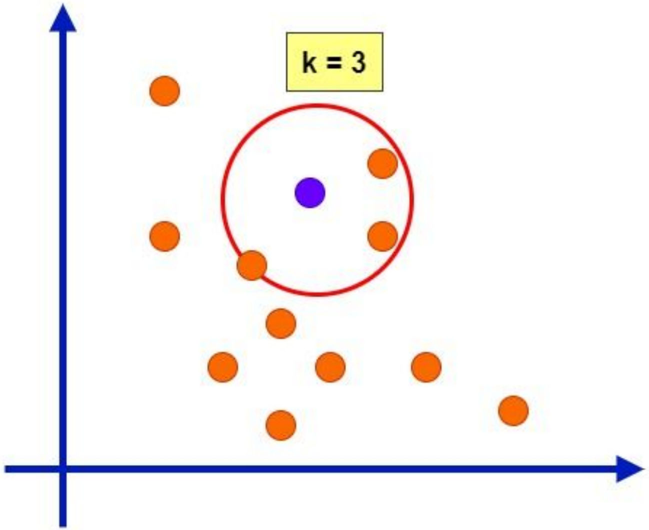
Without outlier - IQR



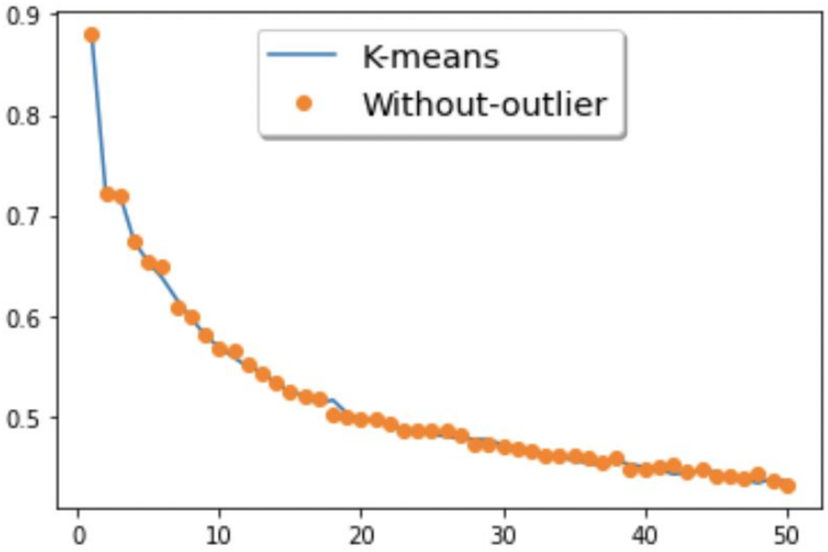
- Blue curve - with
- Orange curve - unnormalized
- Error decrease after remove the outlier.

Outlier: Local outlier factor - density based

- Local outlier factor using similar way like KNN, it find the neighbor around the data point which is also the density of it.
- It can be clustering-based and density-based.



Without outlier - LOF



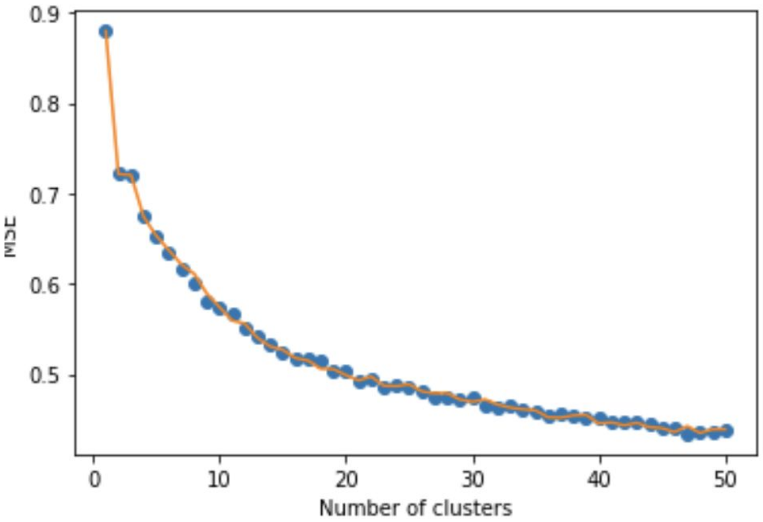
- Basically the same
- Means: The density and distance of each point are close.

Partitional clustering

K-means is the commonly used partitional clustering.

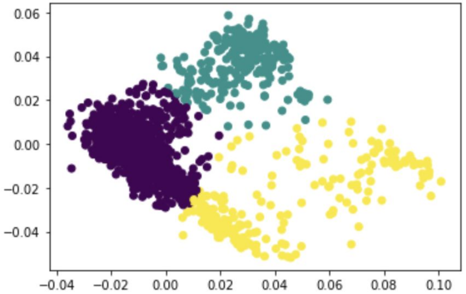
- Each partition have a centroid, data point assigned to each partition measure by different criteria

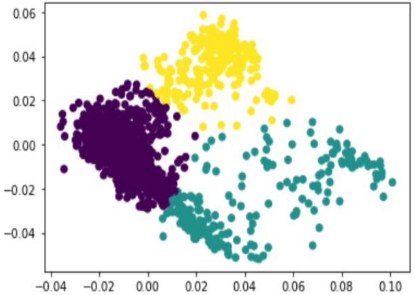
Comparison:K-means & K-means++



Curve = K-means

Point = K-means++



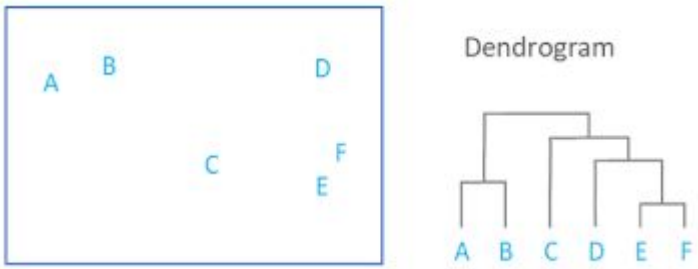


Hierarchical clustering

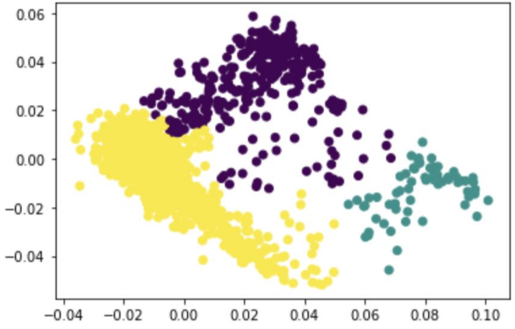
Continuing merge the smaller clusters or separate bigger clusters.

The separation or merging in this algorithm is meaningful.

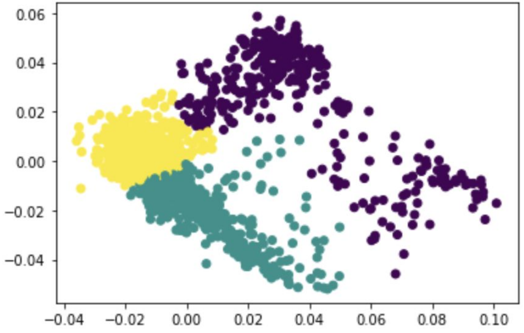
Usually visualized as dendrogram which shows the hierarchy of data.



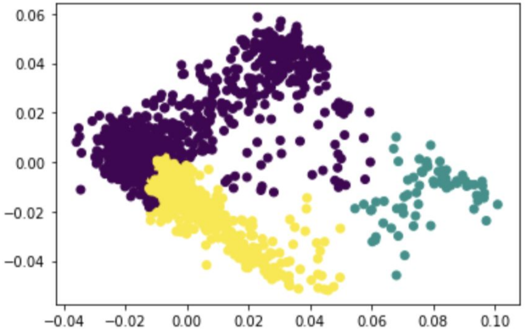
Comparison: Different link(PCA)



- Group link

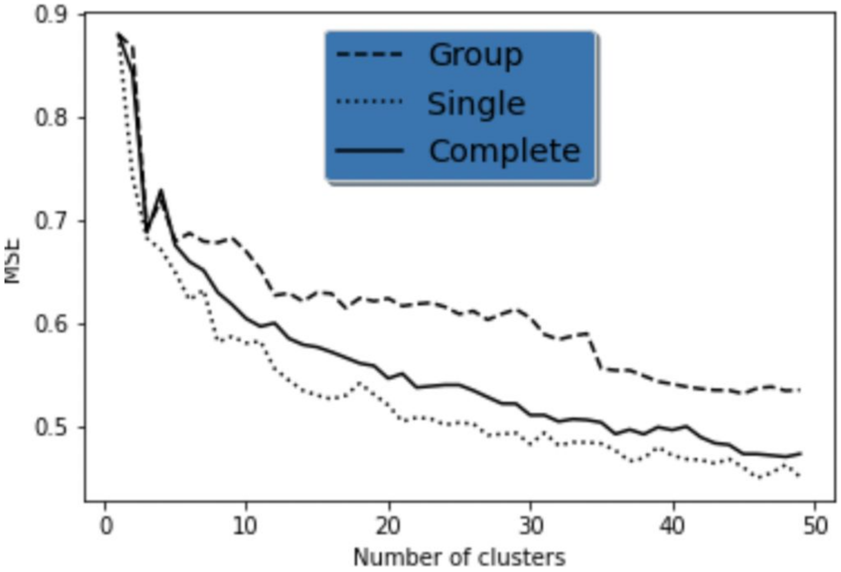


- Single link



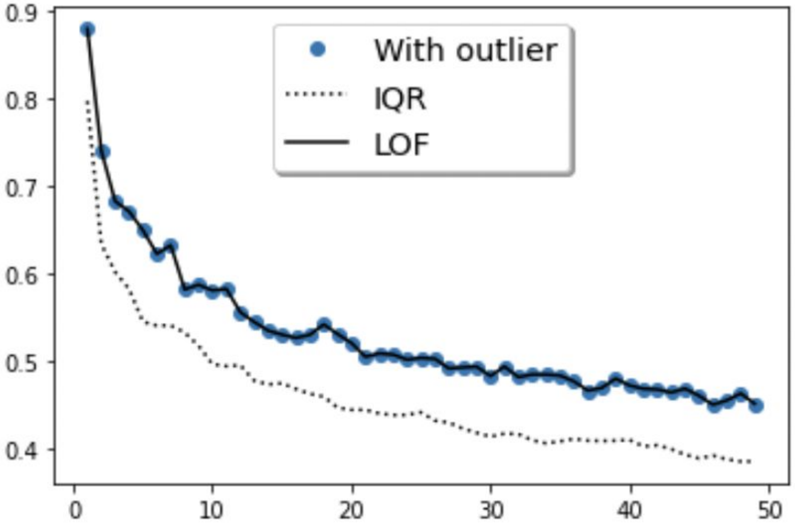
- Complete link

Compare the error



Single link in hierarchical clustering in this dataset have less error than other.

Hierarchical clustering



- IQR can minimize the error of hierarchical clustering in single link.

Method

K-means

- The data that have process the outlier with IQR.

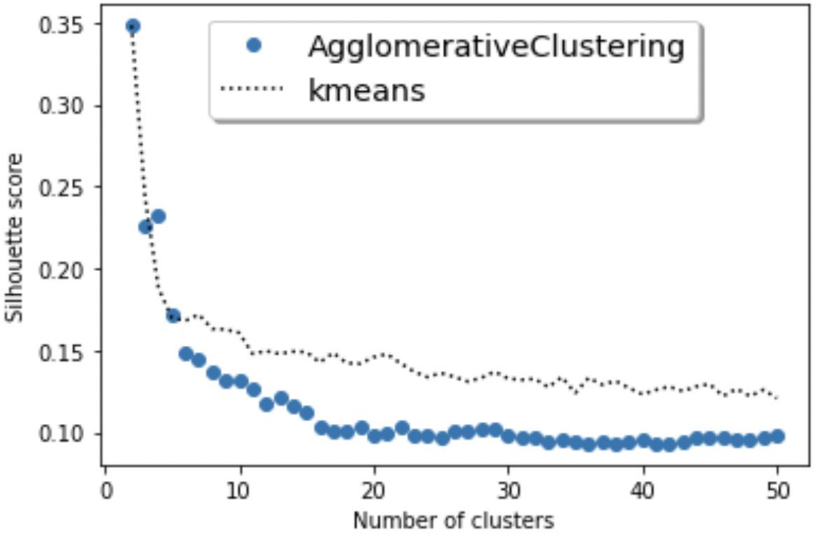
Agglomerative Hierarchical Clustering

- Single link
- Process the outlier with IQR

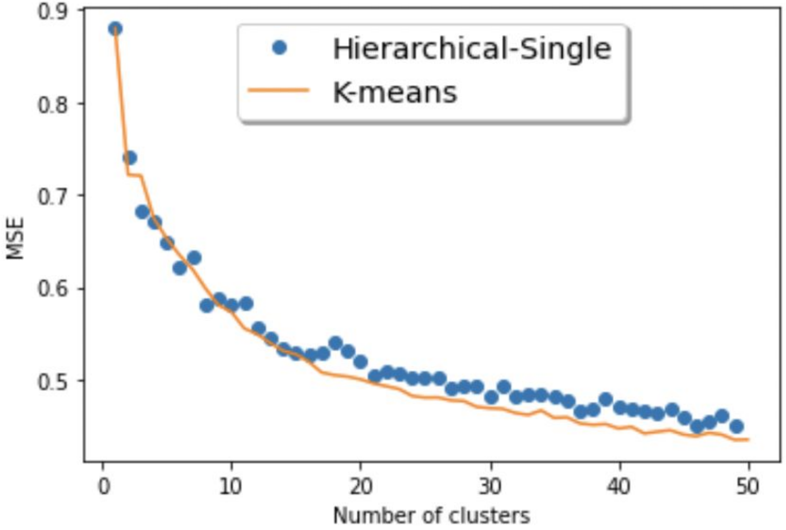
Method:Average error & Silhouette score

- After processing two algorithm compare it with the error each algorithm make in grouping.
- Average Squared error
  - In all the clusters calculate the euclidean distance to its centroid.
- Dunn index
  - Finding the difference between biggest and smallest cluster.

Silhouette score - different clustering



- hierarchical clustering have


0.9
0.8
0.7
0.6
0.5
MSE
0
10
20
30
40
50
Number of clusters
● Hierarchical-Single
K-means

Future investigation: recommendation system

- Clustering is good at separating, it can help is identify user group and music types. Each of this will able to help in assigned customers interest and find future recommendation types.
- Also in some recommendation system they will compare the similarity between customer interest and product.

Citation

Indika. (2011, May 29). *Difference between hierarchical and partitional clustering*. Compare the Difference Between Similar Terms. Retrieved September 10, 2021, from https://www.differencebetween.com/difference-between-hierarchical-and-vs-partitional-clustering/.

begin4learn. (n.d.). *Boxplot*. Boxplot · Begin to Learn R. Retrieved September 10, 2021, from https://begin4learn.gitbooks.io/begin-to-learn-r/content/R/Graphics_BasicGraphics_boxplot.html.

Ansari, Z., Azeem, M. F., Ahmed, W., & Babu, A. V. (2011, June). Quantitative Evaluation of Performance and Validity Indices for Clustering the Web Navigational Sessions. Retrieved September 10, 2021, from https://arxiv.org/pdf/1507.03340.pdf.

Kalafinaian. (2019, July 13). *L2范数归一化概念和优势*. L2范数归一化概念和优势- Kalafinaian - 博客园. Retrieved from https://www.cnblogs.com/Kalafinaian/p/11180519.html.

seraloukseralouk 24.2k55 gold badges8585 silver badges107107 bronze badges. (2020, June 3). *Can someone explain to me how MINMAXSCALER() WORKS?* Stack Overflow. Retrieved September 13, 2021, from https://stackoverflow.com/questions/62178888/can-someone-explain-to-me-how-minmaxscaler-works.

seraloukseralouk 24.2k55 gold badges8585 silver badges107107 bronze badges. (2017, June 18). *Can anyone explain me STANDARDSCALER?* Stack Overflow. Retrieved September 13, 2021, from https://stackoverflow.com/questions/40758562/can-anyone-explain-me-standardscaler.

Metin, F. (n.d.). *LOCAL outlier FACTOR (SOLUTION by hand AND Implementation): Data science and machine learning.* Kaggle. Retrieved September 13, 2021, from https://www.kaggle.com/general/183478.