In [1]:

```
from sklearn.cluster import KMeans
from sklearn.preprocessing import MinMaxScaler
from tqdm import tqdm
import pandas as pd
import numpy as np
import os
tqdm.pandas()
```

```
/Users/jonas/opt/anaconda3/envs/dl/lib/python3.8/site-packages/tqdm/
std.py:668: FutureWarning: The Panel class is removed from pandas. A
ccessing it from the top-level namespace will also be removed in the
next version
  from pandas import Panel
```

Load Dataset

In [2]:

```
file = 'billboard_with_attributes_67741.csv'
df = pd.read_csv(file)
print(df.shape)
df.sample(3)
```

```
(67741, 26)
```

Out[2]:

	Artists	Name	Weekly.rank	Peak.position	Weeks.on.chart	Week	Date	G
14326	Britney Spears, Iggy Azalea	Pretty Girls	97	29.0	8.0	2015-07-10	May 4, 2015	Producer,Ele Pop,Rap
42392	Hurricane Chris	A Bay Bay	85	85.0	2.0	2007-06-27	May 28, 2007	
58580	Korn	Here To Stay	87	72.0	7.0	2002-07-02	June 11, 2002	Metal, Rock,Altern Meta Metal,I

3 rows × 26 columns

In [3]:

```
# 이름, 가수가 같은 곡 제거
dropped = df.drop_duplicates(subset=['Artists', 'Name'], keep='first', inplace=F
alse)
dropped.reset_index(inplace=True)
print('length of dropped: {:,}'.format(len(dropped)))
dropped.sample(3)
```

```
length of dropped: 4,314
```

Out[3]:

	index	Artists	Name	Weekly.rank	Peak.position	Weeks.on.chart	Week	Da
3739	56412	Foo Fighters	All My Life	82	43.0	20.0	2003-03-11	May 1 20
2228	31162	Paramore	The Only Exception	53	24.0	20.0	2010-10-21	Septemb 29, 20
430	4121	Post Malone	Blame It On Me	93	47.0	2.0	2018-05-19	April 2 20

3 rows × 27 columns

Explore attributes & Select to use

In [4]:

```
attrs = dropped.columns[13:]
print('# of attributes: {}'.format(len(attrs)))
print(attrs)
```

```
# of attributes: 14
Index(['acousticness', 'danceability', 'duration_ms', 'energy', 'exp
licit',
       'instrumentalness', 'key', 'liveness', 'loudness', 'mode', 'p
opularity',
       'speechiness', 'tempo', 'valence'],
      dtype='object')
```

In [5]:

```
dropped[attrs].describe()
```

Out[5]:

	acousticness	danceability	duration_ms	energy	explicit	instrumentalness
count	4314.000000	4314.000000	4314.000000	4314.000000	4314.000000	4314.000000
mean	0.163621	0.623372	229969.027585	0.694390	0.265415	0.008349
std	0.204126	0.142267	44468.120384	0.172276	0.441605	0.062264
min	0.000002	0.113000	73813.000000	0.056100	0.000000	0.000000
25%	0.016725	0.528000	202046.750000	0.580000	0.000000	0.000000
50%	0.076700	0.624000	225858.500000	0.720000	0.000000	0.000000
75%	0.236000	0.721000	252270.000000	0.830000	1.000000	0.000021
max	0.978000	0.986000	688453.000000	0.996000	1.000000	0.989000

In [6]:

```
dropped[attrs].info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 4314 entries, 0 to 4313
Data columns (total 14 columns):
 #   Column            Non-Null Count  Dtype  
---  ------            --------------  -----  
 0   acousticness      4314 non-null   float64
 1   danceability      4314 non-null   float64
 2   duration_ms       4314 non-null   float64
 3   energy            4314 non-null   float64
 4   explicit          4314 non-null   float64
 5   instrumentalness  4314 non-null   float64
 6   key               4314 non-null   float64
 7   liveness          4314 non-null   float64
 8   loudness          4314 non-null   float64
 9   mode              4314 non-null   float64
 10  popularity        4314 non-null   float64
 11  speechiness       4314 non-null   float64
 12  tempo             4314 non-null   float64
 13  valence           4314 non-null   float64
dtypes: float64(14)
memory usage: 472.0 KB
```

In [7]:

```
print(dropped['explicit'].value_counts())
print(dropped['mode'].value_counts())
```

```
0.0    3169
1.0    1145
Name: explicit, dtype: int64
1.0    2942
0.0    1372
Name: mode, dtype: int64
```

NOTE

- duration_ms, popularity는 곡을 clustering할 만한 attributes로 볼 수 없다 판단하여 attrs에서 제거
- explicit과 mode는 0과 1로만 이뤄진 카테고리형 변수? -> 제거

In [8]:

```
attrs = ['acousticness', 'danceability', 'energy', 'instrumentalness', 'key', 'l
iveness', 'loudness',
       'speechiness', 'tempo', 'valence']
print('# of attributes: {}'.format(len(attrs)))
```

```
# of attributes: 10
```

Preprocessing

- Normalizing attrs columns

In [9]:

```
# before normalizing
dropped[attrs].head(3)
```

Out[9]:

	acousticness	danceability	energy	instrumentalness	key	liveness	loudness	speechiness
0	0.3280	0.701	0.425	0.13	7.0	0.1000	-10.965	0.3750
1	0.0501	0.900	0.400	0.00	0.0	0.0876	-8.443	0.1240
2	0.0427	0.842	0.734	0.00	1.0	0.1060	-5.065	0.0588

In [10]:

```
# after normalizing
normed = dropped.copy()
mms = MinMaxScaler()
mms.fit(normed[attrs])
normed[attrs] = mms.transform(normed[attrs])
normed[attrs].head(3)
```

Out[10]:

	acousticness	danceability	energy	instrumentalness	key	liveness	loudness	spee
0	0.335377	0.673540	0.392489	0.131446	0.636364	0.085605	0.534486	0
1	0.051225	0.901489	0.365890	0.000000	0.000000	0.072451	0.646277	0
2	0.043658	0.835052	0.721247	0.000000	0.090909	0.091970	0.796011	0

In [11]:

```
normed[attrs].describe()
```

Out[11]:

	acousticness	danceability	energy	instrumentalness	key	liveness	
count	4314.000000	4314.000000	4314.000000	4314.000000	4314.000000	4314.000000	4
mean	0.167299	0.584619	0.679105	0.008442	0.475071	0.169985	
std	0.208719	0.162963	0.183292	0.062957	0.326122	0.141926	
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	
25%	0.017099	0.475372	0.557400	0.000000	0.181818	0.080195	
50%	0.078423	0.585338	0.706352	0.000000	0.454545	0.113186	
75%	0.241307	0.696449	0.823385	0.000022	0.727273	0.225629	
max	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	

In [12]:

```
normed[attrs].describe().loc['std'].sort_values()
```

Out[12]:

```
instrumentalness    0.062957
loudness            0.096704
liveness            0.141926
speechiness         0.143317
danceability        0.162963
energy              0.183292
tempo               0.183697
acousticness        0.208719
valence             0.237016
key                 0.326122
Name: std, dtype: float64
```

NOTE

- instrumentalness: all low
- loudness: all high
- liveness: too low
- speechiness: too low
- danceability: central
- energy: central
- tempo: central
- acousticness: little low
- valence: central
- key: suitable

In [13]:

```
attrs1 = ['acousticness', 'danceability', 'energy', 'instrumentalness', 'key',
'liveness', 'loudness',
       'speechiness', 'tempo', 'valence'] # all
attrs2 = ['acousticness', 'danceability', 'energy', 'key', 'liveness',
       'speechiness', 'tempo', 'valence'] # inst. loud.
attrs3 = ['acousticness', 'danceability', 'energy', 'key', 'tempo', 'valence'] #
live. speech.
```

Clustering

TEST LIST

1. attrs1, k=7
2. attrs1, k=5
3. attrs2, k=7
4. attrs2, k=5
5. attrs3, k=5: attr3은 feature 수가 6개이므로 k=7 작업은 하지 않음

In [14]:

```
import matplotlib.pyplot as plt
from math import pi
from matplotlib.path import Path
from matplotlib.spines import Spine
from matplotlib.transforms import Affine2D
```

In [15]:

```
def get_clusters(df, attrs, k=5):
    'append the cluster column'
    song_vectors = df[attrs].to_numpy()
    kmeans = KMeans(n_clusters=k, max_iter=10, n_init=1, verbose=False)
    clusters = kmeans.fit_predict(song_vectors)
    df_ = df.copy()
    df_['cluster'] = clusters
    return df_

def radar(df, attrs):
    num_attrs = len(attrs)
    num_clusters = len(df)

    angles = [x/float(num_attrs)*(2*pi) for x in range(num_attrs)]
    angles += angles[:1]

    my_palette = plt.cm.get_cmap('Set2', len(df))
    if num_clusters == 5:
        fig = plt.figure(figsize=(20,15))
    else:
        fig = plt.figure(figsize=(20,20))
    fig.set_facecolor('white')

    for i, row in df.iterrows():
        color = my_palette(i)
        data = df.loc[i].tolist()
        data += data[:1]

        if num_clusters == 5:
            ax = plt.subplot(2, 3, i+1, polar=True)
        else:
            ax = plt.subplot(3, 3, i+1, polar=True)
        ax.set_theta_offset(pi/2)
        ax.set_theta_direction(-1)

        plt.xticks(angles[:-1], attrs, fontsize=13)
        ax.tick_params(axis='x', which='major', pad=15)

        ax.set_rlabel_position(0)
        plt.ylim(0,1)

        ax.plot(angles, data, color=color, linewidth=2, linestyle='solid')
        ax.fill(angles, data, color=color, alpha=0.3)

        plt.title("Cluster "+str(i+1), size=20, color=color, x=-0.1, y=1.2, ha=
'left')

    plt.tight_layout(pad=3)
    plt.show()
```

In [16]:

```
df1 = get_clusters(normed, attrs1, k=7)
df1_radar = pd.pivot_table(df1, index='cluster', values=attrs1, aggfunc='mean')
radar(df1_radar, attrs1)
```

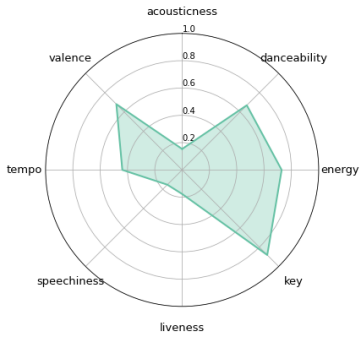
In [17]:

```
df2 = get_clusters(normed, attrs1, k=5)
df2_radar = pd.pivot_table(df2, index='cluster', values=attrs1, aggfunc='mean')
radar(df2_radar, attrs1)
```

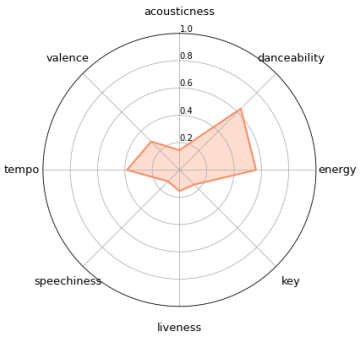
In [18]:

```
df3 = get_clusters(normed, attrs2, k=7)
df3_radar = pd.pivot_table(df3, index='cluster', values=attrs2, aggfunc='mean')
radar(df3_radar, attrs2)
```

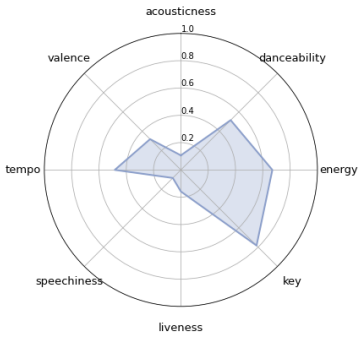
Cluster 1



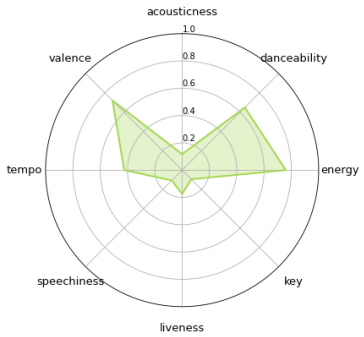
Cluster 2



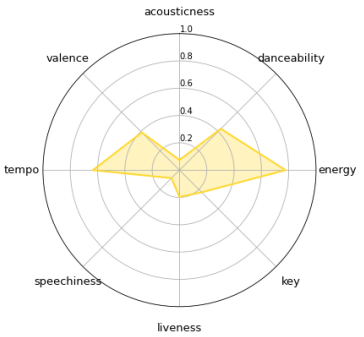
Cluster 3



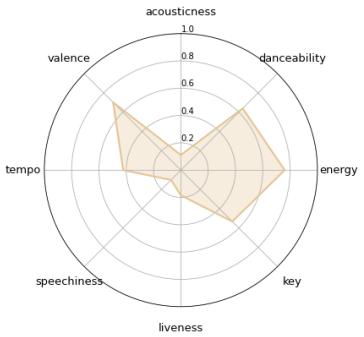
Cluster 4



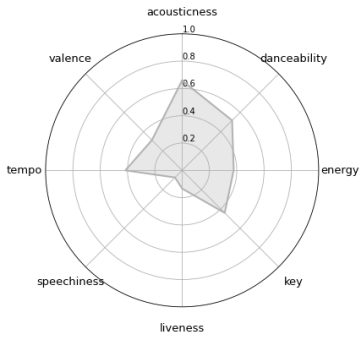
Cluster 5



Cluster 6

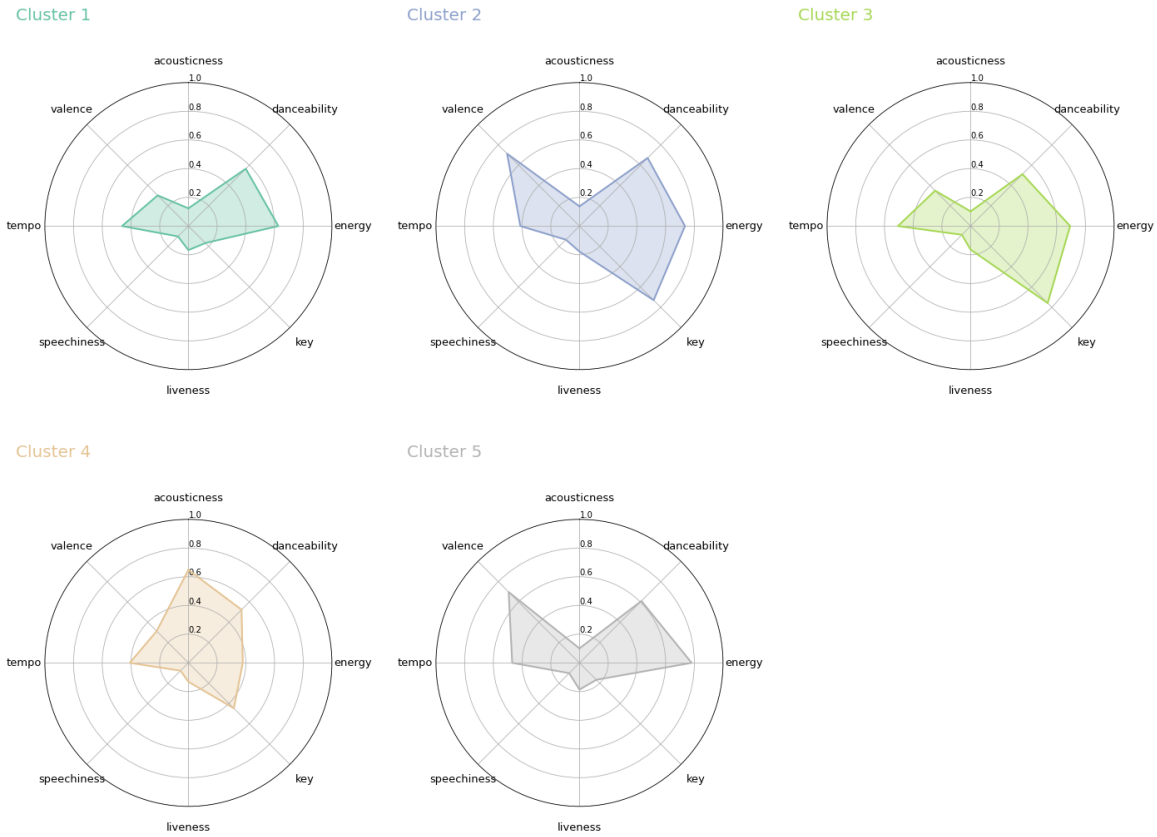


Cluster 7



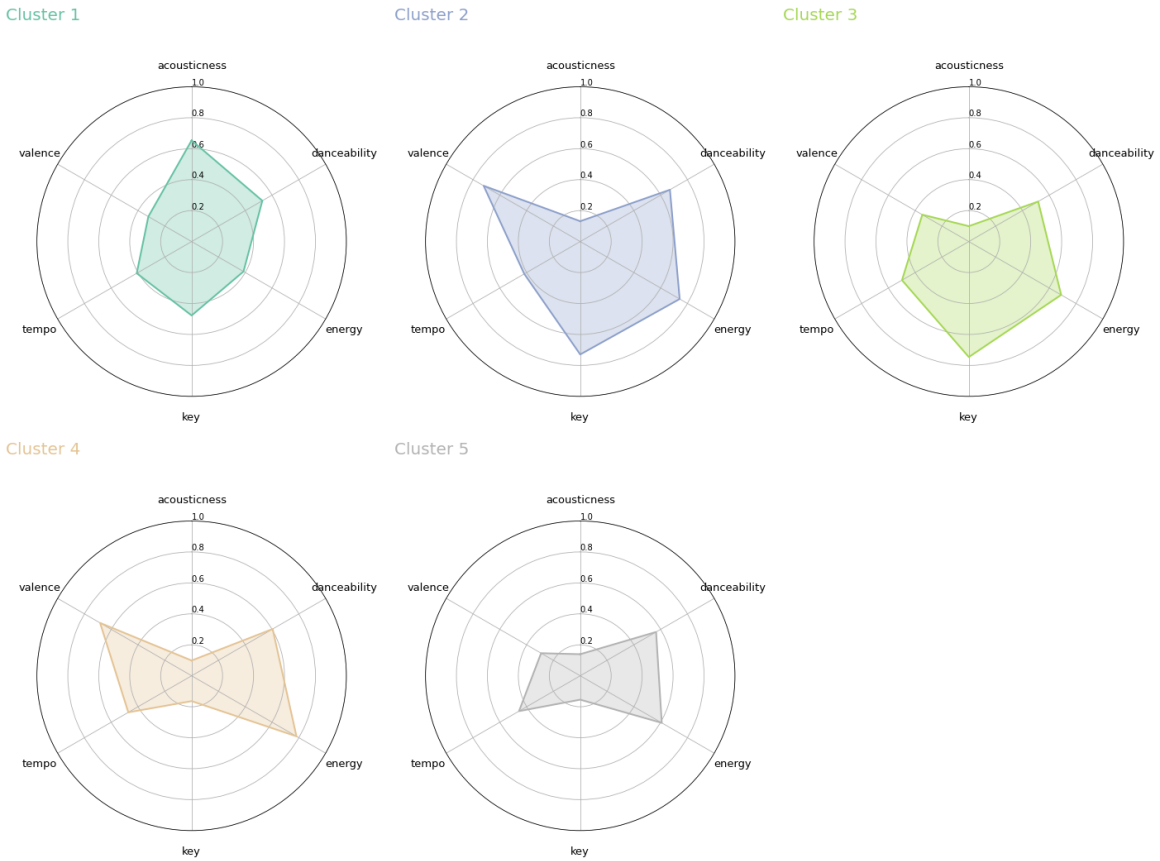
In [19]:

```
df4 = get_clusters(normed, attrs2, k=5)
df4_radar = pd.pivot_table(df4, index='cluster', values=attrs2, aggfunc='mean')
radar(df4_radar, attrs2)
```



In [20]:

```
df5 = get_clusters(normed, attrs3, k=5)
df5_radar = pd.pivot_table(df5, index='cluster', values=attrs3, aggfunc='mean')
radar(df5_radar, attrs3)
```



Explore Each Cluster

1. count
2. most frequent artists
3. top ranked songs

LIST

1. test4: attrs2, k=5
2. test5: attrs3, k=5

In [135]:

```
from collections import Counter

class cluster_inform:

    def __init__(self, df):
        self.df = df

    def get_counting(self):
        self.total = len(self.df)
        df_ = pd.pivot_table(self.df, index='cluster', values='Name', aggfunc='c
ount')
        df_.columns = ['count']
        df_['rate(%)'] = round(100*df_['count'] / self.total)
        print('total count: {:,}'.format(self.total))
        return df_

    def get_artists(self, k=5):
        self.artists = list()
        self.num_clusters = len(self.df['cluster'].unique())

        for i in range(self.num_clusters):
            df_ = self.df.loc[self.df['cluster']==i, ['Artists', 'Name']]
            counter = Counter()
            counter.update(df_.Artists)
            self.artists.append(counter.most_common()[:k])

        for i in range(len(self.artists)):
            print('*** Frequent Artists in Cluster {} ***'.format(str(i+1)))
            print(self.artists[i])
            print(' ')

        #return self.artists

    def get_songs(self, k=10):
        self.songs = list()

        for i in range(self.num_clusters):
            df_ = self.df.loc[self.df['cluster']==i, ['Artists', 'Name', 'Weekl
y.rank']]
            sorted_table = df_.sort_values(by='Weekly.rank')[:k].reset_index(dro
p=True)
#             sorted_table = df_.sort_values(by='Weekly.rank')[:k]  # change ind
ex to rank (1,k+1)
#             sorted_table['rank'] = range(1,k+1)
#             sorted_table.set_index('rank', inplace=True)
            self.songs.append(sorted_table)

        for i in range(len(self.songs)):
            print('*** Top Ranked Songs in Cluster {} ***'.format(str(i+1)))
            print(self.songs[i])
            print(' ')

        # return self.songs
```

test4: attrs2 & k=5

In [136]:

```
t4 = cluster_inform(df4)
t4.get_counting()
```

```
total count: 4,314
```

Out[136]:

cluster	count	rate(%)
0	877	20.0
1	1078	25.0
2	1022	24.0
3	412	10.0
4	925	21.0

In [137]:

```
t4.get_artists()
```

```
*** Frequent Artists in Cluster 1 ***
[('Drake', 40), ('Taylor Swift', 21), ('Kanye West', 13), ('Eminem',
11), ('Rihanna', 11)]

*** Frequent Artists in Cluster 2 ***
[('Eminem', 16), ('Nicki Minaj', 11), ('Taylor Swift', 11), ('Britne
y Spears', 11), ('Drake', 10)]

*** Frequent Artists in Cluster 3 ***
[('Drake', 27), ('Taylor Swift', 20), ('Lil Wayne', 15), ('Jason Ald
ean', 13), ('Kanye West', 12)]

*** Frequent Artists in Cluster 4 ***
[('Billie Eilish', 12), ('XXXTENTACION', 10), ('Ed Sheeran', 10),
('Drake', 8), ('Ariana Grande', 7)]

*** Frequent Artists in Cluster 5 ***
[('Keith Urban', 15), ('Justin Bieber', 12), ('Blake Shelton', 11),
('Jonas Brothers', 10), ('Taylor Swift', 10)]
```

In [138]:

```
t4.get_songs()
```

```
*** Top Ranked Songs in Cluster 1 ***
              Artists                 Name  Weekly.rank
0              Khalid                 Talk            4
1               Lizzo          Truth Hurts           11
2       Jonas Brothers  A Little Bit Longer           11
3         Taylor Swift       State Of Grace           13
4            Lil Nas X               Panini           16
5           Kanye West                 Gone           18
6         Taylor Swift            Superstar           26
7         Taylor Swift              Innocent           27
8  5 Seconds Of Summer      Kiss Me Kiss Me           28
9                Migos      Walk It Talk It           29

*** Top Ranked Songs in Cluster 2 ***
               Artists                                  Name  Weekly.r
ank
0         Billie Eilish                              Bad Guy
3
1                Prince                       When Doves Cry
8
2            Brenda Lee  Rockin' Around The Christmas Tree
9
3         Justin Bieber                        Heartbreaker
13
4       Michael Jackson                         Billie Jean
14
5                Halsey                          Without Me
17
6    Lil Nas X, Cardi B                               Rodeo
22
7  Marshmello, Bastille                              Happier
25
8                Prince                                1999
27
9         Morgan Wallen                   Whiskey Glasses
28

*** Top Ranked Songs in Cluster 3 ***
               Artists                                 Name  Weekly.rank
0         Mariah Carey  All I Want For Christmas Is You            3
1            Lady Gaga                                Hair           12
2         Rich The Kid                          Plug Walk           15
3        Blake Shelton                     God's Country           18
4           Ed Sheeran                             Perfect           21
5  5 Seconds Of Summer          Everything I Didn't Say           24
6         Taylor Swift                           Superman           26
7           Katy Perry                 Never Really Over           27
8           Aloe Blacc                            The Man           33
9           City Girls                             Act Up           34

*** Top Ranked Songs in Cluster 4 ***
           Artists                                      Name  Weekl
y.rank
0      Bobby Helms                          Jingle Bell Rock
8
1    Nat King Cole                       The Christmas Song
11
2    Andy Williams  It's The Most Wonderful Time Of The Year
13
3    Ariana Grande                                 7 Rings
23
```

```
4         Lee Brice                                      Rumor
26
5        Gene Autry                   Here Comes Santa Claus
28
6        Kanye West               I Thought About Killing You
28
7          Coldplay                                   Midnight
29
8       Alicia Keys                          A Woman's Worth
30
9   Whitney Houston                   I Will Always Love You
30

*** Top Ranked Songs in Cluster 5 ***
               Artists                      Name  Weekly.rank
0       Jonas Brothers                    Sucker            6
1                DaBaby                      Suge            7
2         One Direction                     Diana           11
3          Taylor Swift   You Need To Calm Down           13
4          Shawn Mendes     If I Can't Have You           14
5                Prince     Little Red Corvette           20
6           Nicki Minaj                  MEGATRON           20
7               Ava Max         Sweet But Psycho           21
8   Panic! At The Disco  Hey Look Ma, I Made It           24
9             Ed Sheeran             Shape Of You           24
```

test5: attrs3 & k=5

In [139]:

```
t5 = cluster_inform(df5)
t5.get_counting()
```

```
total count: 4,314
```

Out[139]:

cluster	count	rate(%)
0	409	9.0
1	1070	25.0
2	1060	25.0
3	960	22.0
4	815	19.0

In [140]:

```
t5.get_artists()
```

```
*** Frequent Artists in Cluster 1 ***
[('Billie Eilish', 13), ('XXXTENTACION', 11), ('Drake', 9), ('Ed She
eran', 9), ('Ariana Grande', 7)]

*** Frequent Artists in Cluster 2 ***
[('Eminem', 16), ('Britney Spears', 13), ('Nicki Minaj', 11), ('Tayl
or Swift', 11), ('Drake', 9)]

*** Frequent Artists in Cluster 3 ***
[('Drake', 26), ('Taylor Swift', 22), ('Lil Wayne', 15), ('Jason Ald
ean', 13), ('Kanye West', 13)]

*** Frequent Artists in Cluster 4 ***
[('Keith Urban', 15), ('Taylor Swift', 12), ('Justin Bieber', 12),
('Kenny Chesney', 11), ('Jonas Brothers', 10)]

*** Frequent Artists in Cluster 5 ***
[('Drake', 41), ('Taylor Swift', 17), ('Kanye West', 13), ('Rihann
a', 11), ('Eminem', 9)]
```

In [141]:

```
t5.get_songs()
```

```
*** Top Ranked Songs in Cluster 1 ***
          Artists                                      Name  Weekl
y.rank
0    Billie Eilish                                 Bad Guy
3
1      Bobby Helms                          Jingle Bell Rock
8
2    Nat King Cole                       The Christmas Song
11
3    Andy Williams  It's The Most Wonderful Time Of The Year
13
4    Ariana Grande                                 7 Rings
23
5        Lee Brice                                   Rumor
26
6       Kanye West              I Thought About Killing You
28
7       Gene Autry                   Here Comes Santa Claus
28
8         Coldplay                                Midnight
29
9  Whitney Houston                   I Will Always Love You
30

*** Top Ranked Songs in Cluster 2 ***
               Artists                                Name  Weekly.r
ank
0                Prince                    When Doves Cry
8
1            Brenda Lee  Rockin' Around The Christmas Tree
9
2       Michael Jackson                       Billie Jean
14
3                Halsey                       Without Me
17
4    Lil Nas X, Cardi B                             Rodeo
22
5  Marshmello, Bastille                          Happier
25
6                Prince                            1999
27
7         Morgan Wallen                  Whiskey Glasses
28
8               J. Cole                  4 Your Eyez Only
29
9       Imagine Dragons                         Believer
29

*** Top Ranked Songs in Cluster 3 ***
               Artists                               Name  Weekly.rank
0         Mariah Carey  All I Want For Christmas Is You             3
1            Lady Gaga                              Hair            12
2        Justin Bieber                     Heartbreaker            13
3         Rich The Kid                        Plug Walk            15
4        Blake Shelton                   God's Country            18
5           Ed Sheeran                          Perfect            21
6  5 Seconds Of Summer        Everything I Didn't Say             24
7         Taylor Swift                          Superman            26
8           Katy Perry               Never Really Over            27
9        Justin Bieber                      Hold Tight            29
```

```
*** Top Ranked Songs in Cluster 4 ***
               Artists                     Name  Weekly.rank
0       Jonas Brothers                   Sucker            6
1               DaBaby                     Suge            7
2        One Direction                    Diana           11
3         Taylor Swift   You Need To Calm Down           13
4         Shawn Mendes     If I Can't Have You           14
5          Nicki Minaj                 MEGATRON           20
6               Prince     Little Red Corvette           20
7             Ava Max        Sweet But Psycho           21
8  Panic! At The Disco  Hey Look Ma, I Made It           24
9           Ed Sheeran            Shape Of You           24

*** Top Ranked Songs in Cluster 5 ***
          Artists                     Name  Weekly.rank
0          Khalid                     Talk            4
1           Lizzo             Truth Hurts           11
2  Jonas Brothers    A Little Bit Longer           11
3    Taylor Swift         State Of Grace           13
4       Lil Nas X                 Panini           16
5      Kanye West                   Gone           18
6    Taylor Swift              Superstar           26
7    Taylor Swift               Innocent           27
8           Migos        Walk It Talk It           29
9    Taylor Swift  Come In With The Rain           30
```