Overview:

This analysis aims to understand the trend of public tastes in music in 3 main East Asia countries: China, Korea, and Japan. A comparison among artists from these countries was done in this notebook.

- **Finding I**: Japanese and Korean artists tends to have higher popularity and influence regionally and world-wide
- **Finding II**: Follower base for Chinese artists are significantly less than that of Japanese and Korean artists
- **Finding III**: Rap is trending in Japan and Korea in the past few years whereas in China it receives less attention
- **Finding IV**: J-pop has been one of the most popular genres in East Asia in the past 20+ years.
- **Finding V**: There is a rising trend of 'Chinese viral music' (songs that gone viral on the Internet, specifically TikTok and will quickly vanish). However, artists who benefit from this types of music tend to be 'short-lived' in terms of their long-term influence and popularity. Thus, this kind of artists need to find ways to retain public attention and actually accumulate follower bases (e.g. increase music quality, be innovative) to avoid being 'nobody' after the trend is gone.

What I hope I can do:

- I hope to collect data on rough demographic of followers for each artists, and analyzed their influence globally. Although I have said Japanese and Korean artists are popular world-wide, that is half data-supported and half observation. Additional data is needed to solidify this.

Note:

1. This analysis assumes that popularity of artists indicates public music tastes
2. It further assumes that top track is representative to the genres artists are involved in
3. Potential deviation in analysis given the fact that there are many alternative music apps that are popular in mainland China i.e., not a lot of people use Spotify there, which may affect calculation of popularity metrics.

In [1]:

```
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
```

In [2]:
```
dfc = pd.read_csv('chinese_top100_artist.csv', index_col = 0).reset_ir
dfe = pd.read_csv('east_asia_top_artists.csv', index_col = 0).reset_ir
dfj = pd.read_csv('japanese_top100_artist.csv', index_col = 0).reset_i
dfk = pd.read_csv('korean_top100_artist.csv', index_col = 0).reset_inc

df_all = {'China': dfc, 'Japan': dfj, 'Korea': dfk, 'East Asia': dfe}
```

In [3]:
```
# east_asia_artist has an extra column 'query genre', not very informa
for df in df_all.values():
    print(df.shape)
```

```
(100, 13)
(100, 13)
(100, 13)
(700, 14)
```

In [4]:
```
# example head
dfj.head(2)
```

Out[4]:

	artist_name	popularity	followers	artist_link	genres
0	YOASOBI	76	5945744	https://open.spotify.com/artist/64tJ2EAv1R6UaZ...	['j-pop', 'japanese teen pop']
1	Vaundy	70	1731531	https://open.spotify.com/artist/2IUl3m1H1EQ7Qf...	['j-pop', 'japanese soul']

In [5]:
```
dfe.head(2)
```

Out[5]:

	artist_name	popularity	followers	artist_link	genres	t
0	BTS	88	67507448	https://open.spotify.com/artist/3Nrfpe0tUJi4K4...	['k-pop', 'k-pop boy group', 'pop']	
1	BLACKPINK	82	43385244	https://open.spotify.com/artist/41MozSoPIsD1dJ...	['k-pop', 'k-pop girl group', 'pop']	

Features (dfc, dfj, dfk, dfe)

- artist_name: unique values
- popularity: int, calculated by Spotify, [0, 100]

- followers: int, # of followers
- artist_link: **Drop**
- genres: used list to contain different genres
- top_track: name of the top track for given artist **Drop?**
- top_track_album: name of top track's album **Drop?**
- top_track_popularity: int, [0, 100]
- top_track_release_date: obj, include different format (e.g. idx 100), should be datetime, extract year only
- top_track_duration_ms: int, duration in milisecond consider convert into second/min
- top_track_explicit: bool, 80%+ 'False' in all four datasets **Drop**
- top_track_album_link: **Drop**
- top_track_link: **Drop**

Features (dfe)

- 'query_genre': Not very informative **Drop**

In [6]:

```
for df in df_all.values():
    # extract year
    df['top_track_release_year'] = df['top_track_release_date'].str.sl

    # add duration minute
    df['top_track_duration_min'] = (df['top_track_duration_ms']/60000)

    # drop columns
    df.drop(['artist_link', 'top_track_album_link', 'top_track_link',
         'top_track', 'top_track_album','top_track_release_date'], axi

dfe.drop('query_genre', axis = 1, inplace = True)
```

PART II. Analysis

I want to explore the following questions based on what are given in datasets:

1. How are popularity of artists in China, Japan, and Korea compared to each other? How are they compared to the East Asia Standard?
2. Assuming popularity of artists indicates public music tastes, are there any similarity in terms of genres?
3. How has the trend changed over year in these countries?
4. Are there any outlier-artists in dataset? (i.e. in any way different than others)
5. Can we segement these artists? How is each group different?

Note: There are some artists in east asia dataset that are also from China, Japan, or Korea. Hence, we shouldn't overinterpret the result but focus on the general trend, similarity, and differences.

Note2: There might be deviations for Chinese artists as access limitation exists in mainland China.

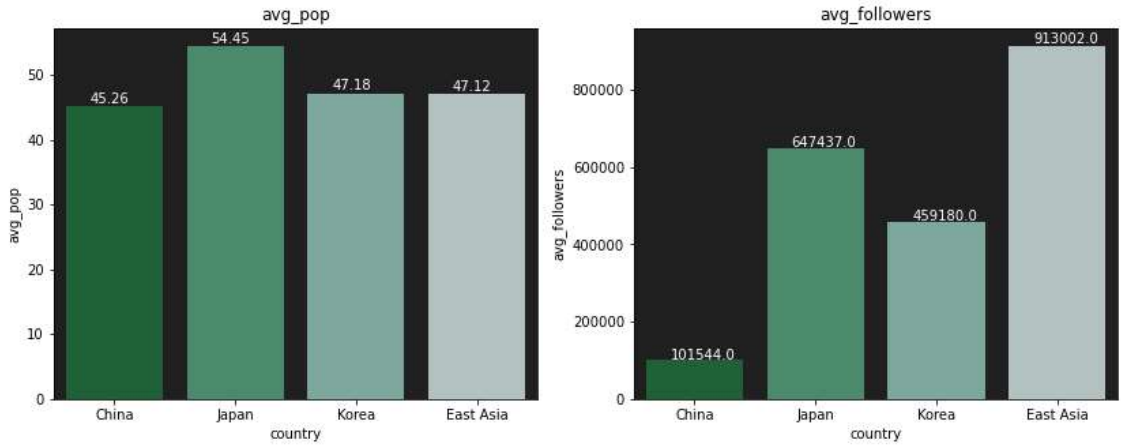
In [7]:
```
1 cty_avg = pd.DataFrame({'country': df_all.keys()})
2 cty_avg['avg_pop'] = [df_all[name]['popularity'].mean().round(2) for r
3 cty_avg['avg_followers'] = [df_all[name]['followers'].mean().round() f
4 cty_avg
```

Out[7]:

	country	avg_pop	avg_followers
0	China	45.26	101544.0
1	Japan	54.45	647437.0
2	Korea	47.18	459180.0
3	East Asia	47.12	913002.0

In [49]:

```
fig, axes = plt.subplots(1, 2, figsize = (14, 5))

fig.set_facecolor('white')

ls = ['avg_pop', 'avg_followers']

for j in range(len(ls)):
    p = sns.barplot(data = cty_avg, x = 'country', y = ls[j], ax = axe
                    palette = 'BuGn_r', alpha = 0.8)
    p.set(title = ls[j])
    p.set_facecolor('#202020')

    y = cty_avg[ls[j]].tolist()
    for i in range(len(y)):
        p.text(i - 0.2, y[i] + y[i]/100, y[i], color = 'white')

# sns.barplot(data = cty_avg, x = 'country', y = 'avg_followers', ax =
```



1. How are popularity of artists in China, Japan, and Korea compared to each other? How are they compared to the East Asia Standard?

Japanese artists have the highest avaerage popularity and # of followers among the three. Korean and Chinese artists have fairly close popularity (to the east asia standard as well) whereas Japan stands out by 15% higher than the East Asia average.

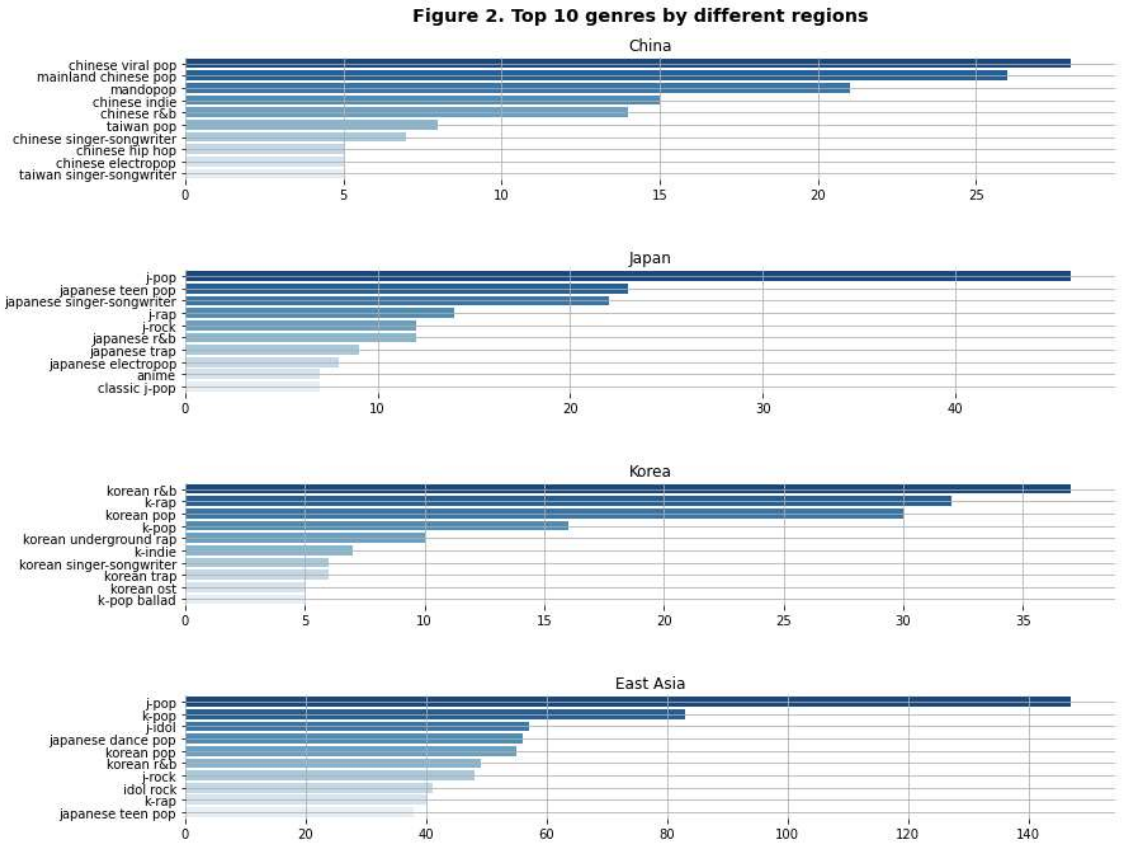
The difference is more significant in average # followers. **Chinese artists has the lowest avg # followers (77% and 84% lower than Korea and Japan repsectively)**

In [9]:

```
# extract and count genres and total popularity from each artists, ret
def get_genre_counts(df):
    df = df.reset_index(drop=True)   #make sure filtered df also work
    genres = dict()
    popularity = dict()

    for i in range(len(df)):
        genre_list = df.genres[i].strip("['']").split("', '")
        for g in genre_list:
            if not g in genres.keys():
                genres[g] = 1
                popularity[g] = df.top_track_popularity[i]
            else:
                genres[g] += 1
                popularity[g] += df.top_track_popularity[i]
    genres = pd.DataFrame({'genres': genres.keys(),
                           'total_popularity': popularity.values(),
                           'count': genres.values()}).sort_values(by =

    return genres
```

In [42]:

```
fig, axes = plt.subplots(4, 1, figsize = (12, 10))

fig.tight_layout(pad = 5)
fig.suptitle('Figure 2. Top 10 genres by different regions', weight =
fig.set_facecolor('white')

j = 0

for i in df_all:
    genres = get_genre_counts(df_all[i]).head(10)
    p = sns.barplot(data = genres, x = 'count', y = 'genres', ax = axe
    p.set(title = f'{i}', xlabel = '', ylabel = '')
    p.grid(visible = True)

    axes[j].spines['right'].set_color(None)
    axes[j].spines['top'].set_color(None)
    axes[j].spines['bottom'].set_color(None)
    axes[j].spines['left'].set_color(None)

    j += 1
```



2. Assuming popularity of artists indicates public music tastes, are there any similarity in terms of genres?

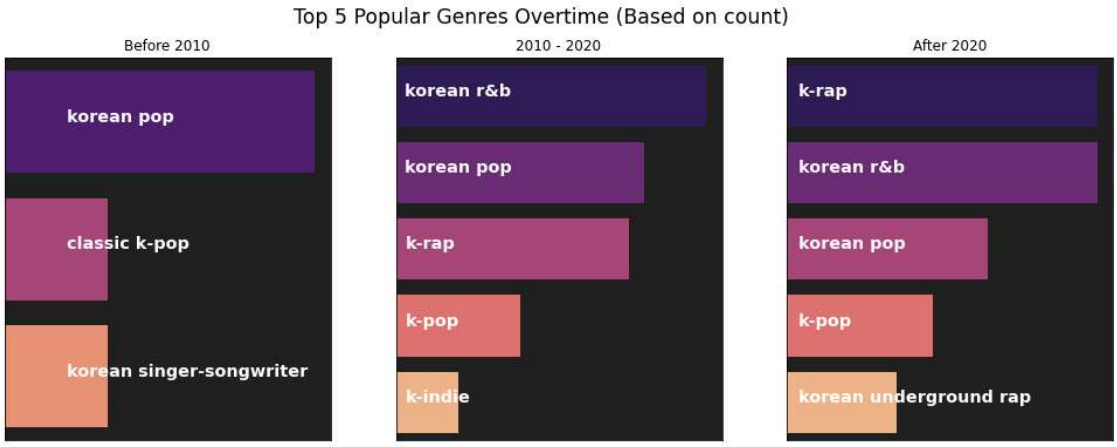
- From East Asia list, it's very impressive to see top 10 genres are all from Japan and Korean, suggesting their high influence in east Asia (and world-wide)
- Rap appears to be more popular in Japan and Korean than in China (or that they are more popular in these 2 countries)

- Common genres in all three genres: pop & r&b
- Special genres in Japan: anime, teen pop
- Special genres in China: viral pop (?), mandopop
- trap which is a type of hip pop also makes its way into top 10 list for Japan and Korea

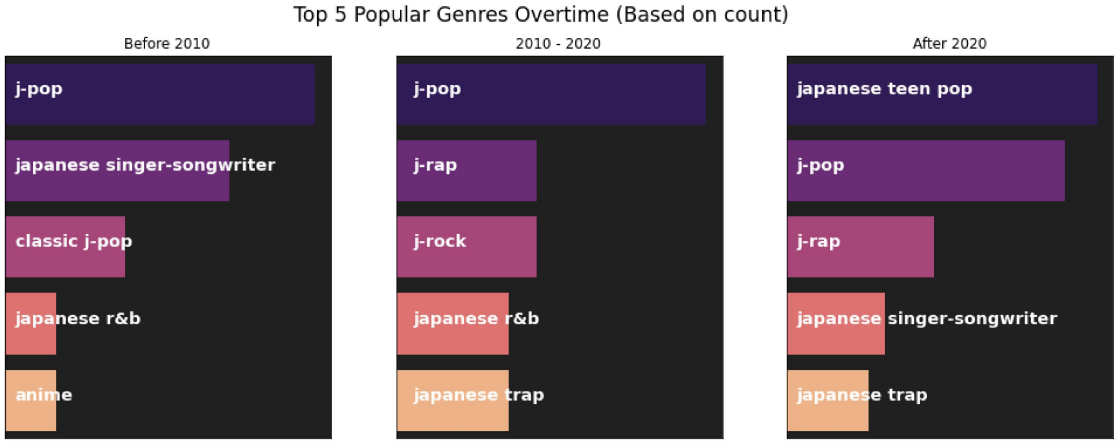
In [45]:

```
def plot_top_genres_overtime(df, by = 'count'): #another option is 'to

    before_2010 = get_genre_counts(df[df['top_track_release_year'] <=
    between_2010_and_2020 = get_genre_counts(df[(df['top_track_release
                                                 (df['top_track_release
    after_2020 = get_genre_counts(df[(df['top_track_release_year'] > '

    period = ['Before 2010', '2010 - 2020', 'After 2020']
    datasets = [before_2010, between_2010_and_2020, after_2020]

    fig, axes = plt.subplots(1, 3, figsize = (17, 6))
    fig.set_facecolor('white')

    fig.suptitle(f'Top 5 Popular Genres Overtime (Based on {by})', for

    for i in range(len(datasets)):
        dataset = datasets[i]
        data = dataset.head(5)
        p = sns.barplot(data = data, x = 'count', y = 'genres', orient

        axes[i].xaxis.set_visible(False)
        axes[i].yaxis.set_visible(False)

        p.set_yticklabels('')
        p.set_xticklabels('')
        p.set(title = period[i], ylabel = '', xlabel = '')
        p.set_facecolor('#202020')

        text = list(data['genres'].values)
        for j in range(len(text)):
            axes[i].text(x = 0.6, y = j, s = text[j], fontsize = 'x-la

    return None
```

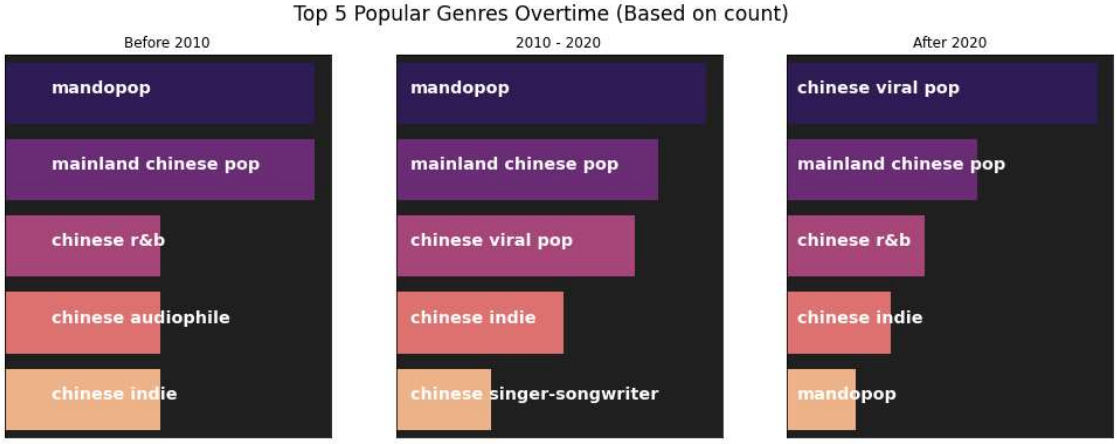
In [46]:
```
1 plot_top_genres_overtime(dfk, by = 'count')
```



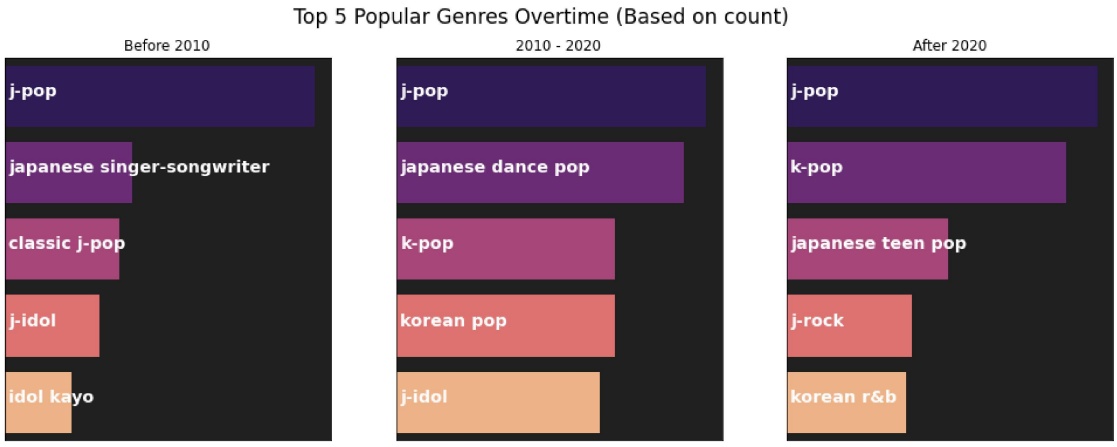
In [13]:
```
1 plot_top_genres_overtime(dfj, by = 'count')
```



In [47]:
```
1 plot_top_genres_overtime(dfc, by = 'count')
```



In [15]:
```
1 plot_top_genres_overtime(dfe, by = 'count')
```



3. How has the trend changed over years in these countries?

Note: we should be aware that there are some overlapping genres such as 'Korean Pop' and 'k-pop'

Korean

- Korean pop is gradually giving its way to k-rap. There is a rising trend in rap as there are two types of them (rap, underground rap) appearing in the top 5 list after 2023

Japan

- While there is also a rising trend of rap in Japan, j-pop has an unassailable leading advantages. Only that there is changes with in j-pop. A rising of japanese teen pop is seen after 2020.

China

- While the rank may have slightly changed over time, the main categories the public enjoyed listening have no big changes (e.g. pop, r&b, indie).
- It's noticeble that mandopop has dropped to the 5th place after 2020 and chinese viral pop has slowly moved its way up the first place. 这与近几年的音乐流行趋势相符，可能由于抖音等短视频的出现，'网红音乐'变得越来越流行，相反流行华语音乐逐渐式微。 However, due to insufficient data, a firm conclusion can not be drawn.

East Asia

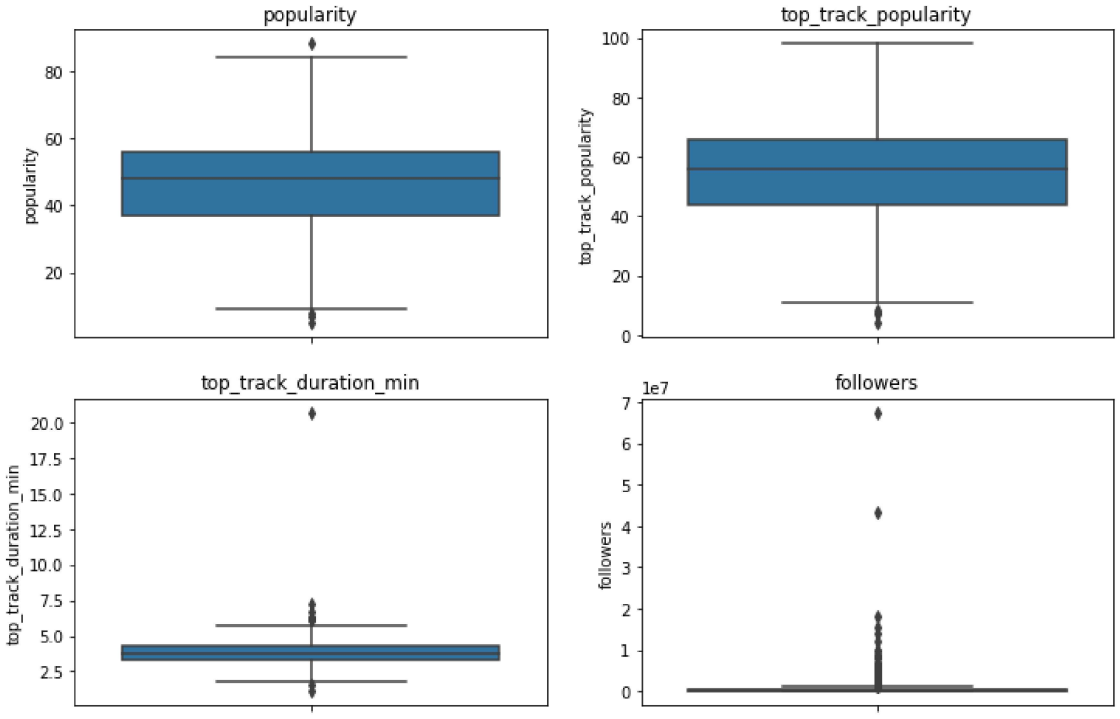
- Leading poisition of j-pop even shows up on the East Asia list, suggesting that it's very well received regionally as well.
- Unfortunately, Chinese aritists didn't make it to the list.

In [16]:

```
# add a column called 'country' before concat
for country in df_all:
    df_all[country]['country'] = country

dfALL = pd.concat(df_all, axis = 0).reset_index(drop = True).drop_dupl
```

In [17]:

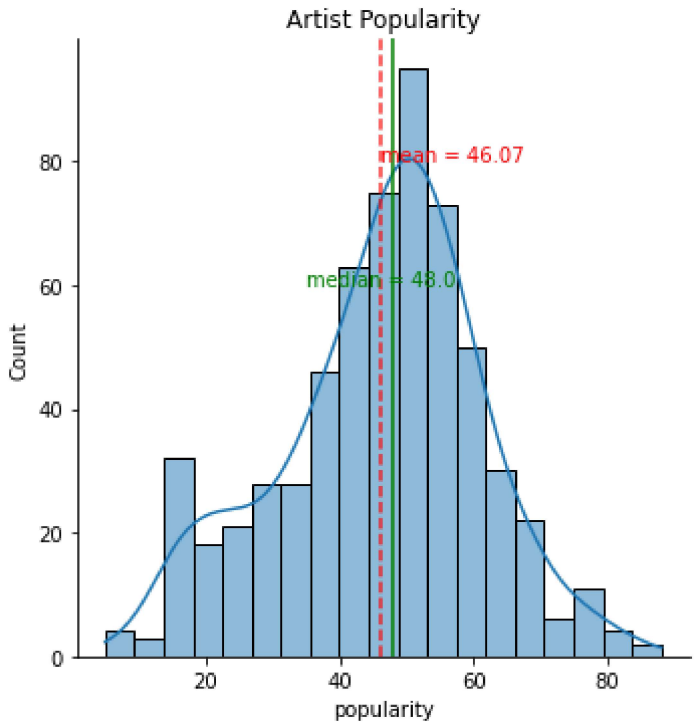
```
# Univariate Analysis on outliers
fig, axes = plt.subplots(2, 2, figsize = (12, 8))
y = ['popularity', 'top_track_popularity', 'top_track_duration_min', '

for i in range(len(y)):
    sns.boxplot(data = dfALL, y = y[i], ax = axes[i//2, i%2])
    axes[i//2, i%2].set_title(y[i])
```



In [18]:

```
p = sns.displot(dfALL['popularity'], kde = True)
p.set(title = 'Artist Popularity')

avg = dfALL['popularity'].mean().round(2)
plt.axvline(avg, color = 'red', linestyle = '--', alpha = 0.8)
plt.text(46, 80, f"mean = {avg}", color = 'red')

med = dfALL['popularity'].median().round(2)
plt.axvline(med , color = 'green')
plt.text(35, 60, f"median = {med}", color = 'green')
```

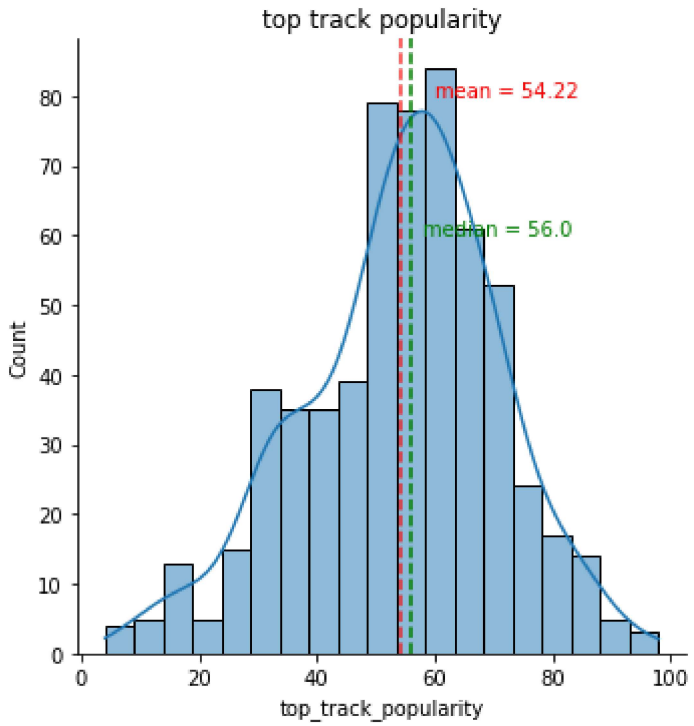
Out[18]: Text(35, 60, 'median = 48.0')



In [19]:

```
p = sns.displot(dfALL['top_track_popularity'], kde = True)
p.set(title = 'top track popularity')

avg = dfALL['top_track_popularity'].mean().round(2)
plt.axvline(avg, color = 'red', linestyle = '--', alpha = 0.8)
plt.text(60, 80, f"mean = {avg}", color = 'red')

plt.axvline(dfALL['top_track_popularity'].median(), color = 'green', l
plt.text(58, 60, f"median = {dfALL['top_track_popularity'].median().ro
```

Out[19]: Text(58, 60, 'median = 56.0')



In [20]:

```
dfALL[dfALL['top_track_duration_min'] > 15]
```

Out[20]:

	artist_name	popularity	followers	genres	top_track_popularity	top_track_duration_ms
234	The Quiett	44	109376	['k-rap', 'korean r&b', 'korean trap']	48.0	1240555.0

In [21]:
```
1 dfALL[dfALL['followers'] > 3e7]
```

Out[21]:

	artist_name	popularity	followers	genres	top_track_popularity	top_track_duration_ms
300	BTS	88	67507448	['k-pop', 'k-pop boy group', 'pop']	88.0	154486.0
301	BLACKPINK	82	43385244	['k-pop', 'k-pop girl group', 'pop']	79.0	175889.0

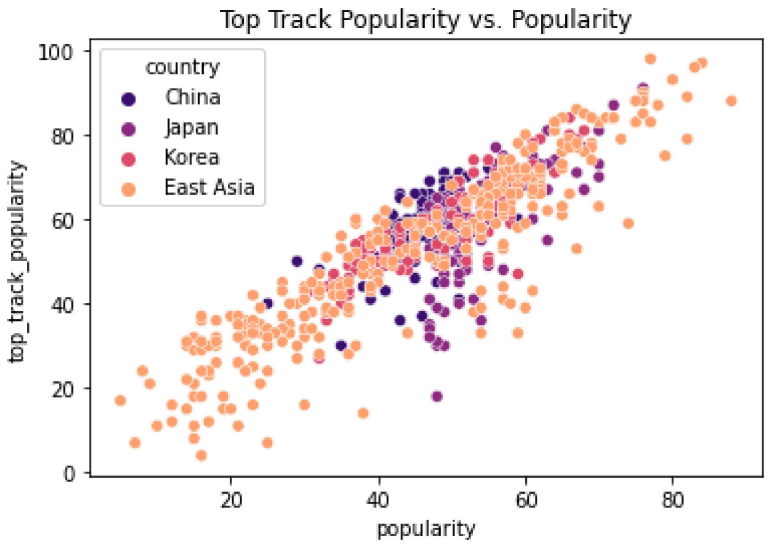
4. Are there any outlier-artists in dataset? (i.e. in any way different than others)

There are 3 noticeable outliers from the plots above.

- Duration of top track by **the Quiett** has more than 20 minutes
- **BTS** and **BLACKPINK** have significantly larger follower bases compared to the rest

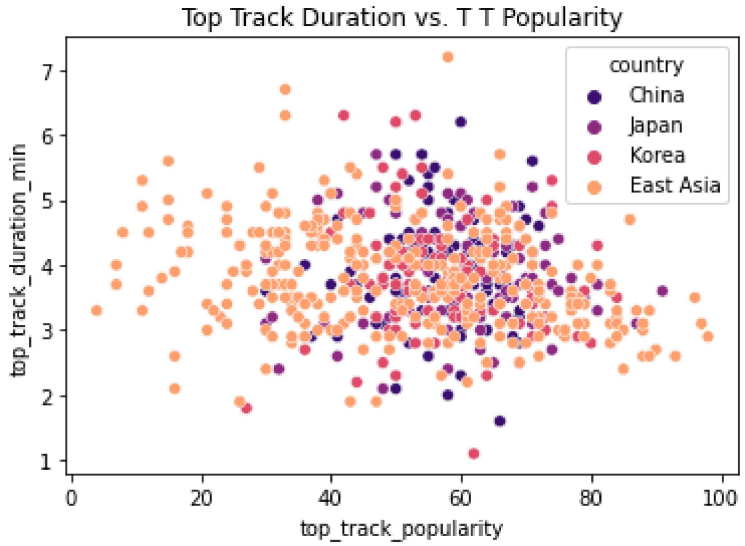
In [22]:
```
1 # Bivariate: Are there any outlier-artists in data set?
2 p = sns.scatterplot(data = dfALL, x = 'popularity', y = 'top_track_pop
3                     hue = 'country', palette = 'magma')
4 p.set(title = 'Top Track Popularity vs. Popularity')
```

Out[22]:
```
[Text(0.5, 1.0, 'Top Track Popularity vs. Popularity')]
```



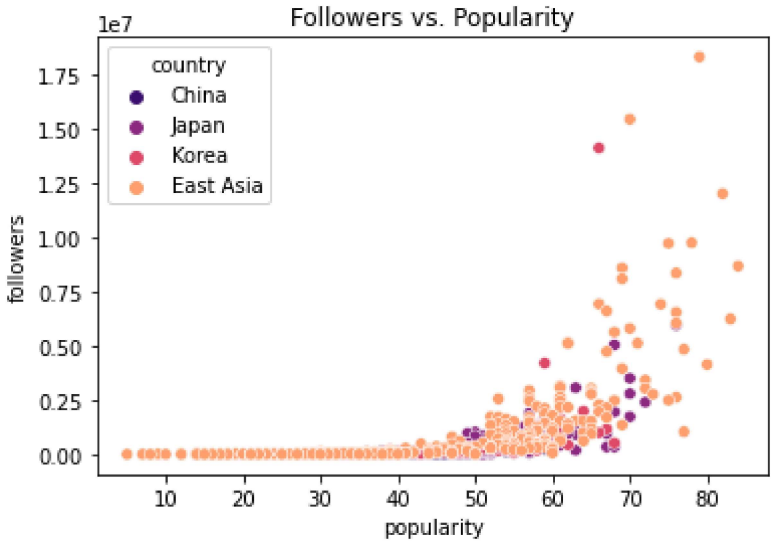
In [23]:
```
# removed outlier
p = sns.scatterplot(data = dfALL[dfALL['top_track_duration_min'] < 10]
                    hue = 'country', palette = 'magma')
p.set(title = 'Top Track Duration vs. T T Popularity')
```

Out[23]: [Text(0.5, 1.0, 'Top Track Duration vs. T T Popularity')]



In [24]:
```
# removed 2 outliers
p = sns.scatterplot(data = dfALL[dfALL['followers'] < 3e7], x = 'popul
                    hue = 'country', palette = 'magma')
p.set(title = 'Followers vs. Popularity')
```

Out[24]: [Text(0.5, 1.0, 'Followers vs. Popularity')]



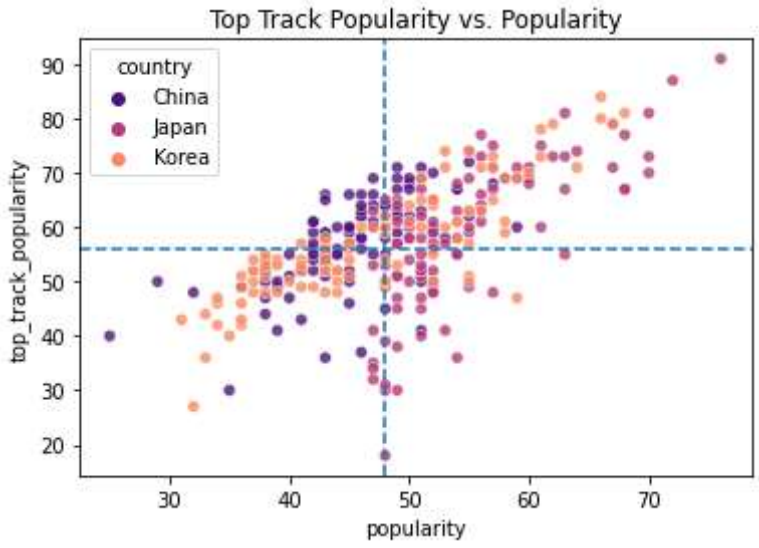
Add-on 4

- By plotting popularity vs. top-track-popularity, we can clearly see the upwards trend. These two variables are definitely positively correlated with each other but causal inference can't be made

- It looks like there is an exponential relationship between followers and popularity, i.e. the higher the popularity, more followers are required to increase popularity score. This could correspond to how Spotify calculate this metrics.
- Another outlier: **V** from Korean has disporportionally high number of followers compared to its popularity score

In [48]:
```
# because there are mix of artists from different countries in East As
fig, ax = plt.subplots()

fig.set_facecolor('white')

p = sns.scatterplot(data = dfALL[dfALL['country'] != 'East Asia'],
                    x = 'popularity', y = 'top_track_popularity',
                    hue = 'country', palette = 'magma', alpha = 0.8)
p.set(title = 'Top Track Popularity vs. Popularity')

plt.axvline(dfALL['popularity'].median(), linestyle = '--')
plt.axhline(dfALL['top_track_popularity'].median(), linestyle = '--')
```

Out[48]: <matplotlib.lines.Line2D at 0x2877e5913c8>



5. Artists Segmentation

For artists in these three countries, I split them into 4 quadrants by median of popularity score and top track popularity score.

- 1st quadrant: 'On Fire'
- 2nd quadrant: 'Flash' artists, their songs or music may become trending at one point but it doesn't last long
- 3rd quadrant: Less popular artists (among all popular ones)
- 4th quadrant: 'High potential'

Something I found interesting:

1. A lot of Chinese artists are on 'Flash' quadrant, which confirm my previous findings. A lot of artists may have one of their songs gone viral on the Internet at a point but soon it will be

replaced by another song. This popularity didn't help them to accumulate their follower/fan bases or click/like (depending on how Spotify calculate popularity score). (People may listen to it because it was trending at that time not because they truely enjoy their music)
2. Almost all top 100 Japanese artists are in the 1st or 4th quadrant, which also align with the fact that Japanese music is well received regionally and even world-wide.