Data Understanding

The data of spotify daily top 200 songs from 2017 - 2021 is offered by spotify. The data has 317,262 rows. It is an unorganized dataframe. Each column includes position, track name, artist, streams, date, and ,genres data. So I have to extract the information to organize my dataset.

Import Packages

In [1]:
```
import pandas as pd
import pickle
```

Load dataset

In [2]:
```
df = pd.read_csv('../data/raw/data.csv', sep='#', parse_dates=['Date'])
```

In [3]:
```
df.head()
```

Out[3]:

	Position	Track Name	Artist	Streams	Date	Genre
0	1	Starboy	The Weeknd	3135625	2017-01-01	['canadian pop', 'canadian contemporary r&b', ...
1	2	Closer	The Chainsmokers	3015525	2017-01-01	['pop', 'pop dance', 'tropical house', 'edm', ...
2	3	Let Me Love You	DJ Snake	2545384	2017-01-01	['pop', 'electronic trap', 'dance pop', 'edm',...
3	4	Rockabye (feat. Sean Paul & Anne-Marie)	Clean Bandit	2356604	2017-01-01	['pop', 'uk dance', 'dance pop', 'uk funky', '...
4	5	One Dance	Drake	2259887	2017-01-01	['toronto rap', 'canadian pop', 'canadian hip ...

Data Information

In [4]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 321200 entries, 0 to 321199
Data columns (total 6 columns):
 #   Column      Non-Null Count   Dtype         
---  ------      --------------   -----         
 0   Position    321200 non-null  int64         
 1   Track Name  321182 non-null  object        
 2   Artist      321182 non-null  object        
 3   Streams     321200 non-null  int64         
 4   Date        321200 non-null  datetime64[ns]
 5   Genre       321182 non-null  object        
dtypes: datetime64[ns](1), int64(2), object(3)
memory usage: 14.7+ MB
```

In [5]:
```
unique_artists = set(df['Artist'].values)
```

In [13]:
```
df.groupby('Artist').sum()['Streams'].sort_values()
```

Out[13]:
```
Artist
M.I.A.                341003
Gente De Zona         377647
Nano                  411014
Bonnie Tyler          450769
Snow Patrol           457635
                      ...
Ariana Grande     8209907401
Billie Eilish     8570405790
Drake             9158708130
Ed Sheeran       11450684279
Post Malone      13513461417
Name: Streams, Length: 1127, dtype: int64
```

In [7]:
```
print('{} artists are in my dataset'.format(len(list(unique_artists))))
```

```
1128 artists are in my dataset
```

In [8]:
```
print('{} streams are average of my dataset'.format(round(df['Streams'].mean(), 0)))
```

```
1188494.0 streams are average of my dataset
```

In [ ]:
```

```

In [9]:
```
with open('../data/final/spotify_df.pickle', 'wb') as spotify:
    pickle.dump(df, spotify)
    spotify.close()
```

In [ ]:
```

```