Data Preparation

For this project, i need time series of streams for each artist. I predict 3 month future growth of streams and create new dataframe with artists rank and 3 month future growth. Finally, cluster the data for analysis to recommendation.

Import Packages

In [1]:
```
import pandas as pd
import matplotlib.pyplot as plt
import pickle
```

Load Data

In [2]:
```
with open('../data/final/spotify_df.pickle', 'rb') as spotify:
    df = pickle.load(spotify)
    spotify.close()
```

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 Cleaning

In [4]:
```
df.isna().sum()
```

Out[4]:
```
Position       0
Track Name    18
Artist        18
Streams        0
Date           0
Genre         18
dtype: int64
```

In [5]:
```
df.dropna(inplace=True)
```

Feature Engineering

In [6]:
```
df['Points'] = (201 - df.Position)/200
```

In [7]:
```
df.head()
```

Out[7]:

	Position	Track Name	Artist	Streams	Date	Genre	Points
0	1	Starboy	The Weeknd	3135625	2017-01-01	['canadian pop', 'canadian contemporary r&b', ...	1.000
1	2	Closer	The Chainsmokers	3015525	2017-01-01	['pop', 'pop dance', 'tropical house', 'edm', ...	0.995
2	3	Let Me Love You	DJ Snake	2545384	2017-01-01	['pop', 'electronic trap', 'dance pop', 'edm',...	0.990
3	4	Rockabye (feat. Sean Paul & Anne-Marie)	Clean Bandit	2356604	2017-01-01	['pop', 'uk dance', 'dance pop', 'uk funky', '...	0.985
4	5	One Dance	Drake	2259887	2017-01-01	['toronto rap', 'canadian pop', 'canadian hip ...	0.980

Create time series for each artists

Create Empty Time Series Frame

In [8]:
```
empty_df = df.groupby('Date').mean()
```

In [9]:
```
empty_df['Streams'] = 0
```

In [10]:
```
empty_df
```

Out[10]:

Date	Position	Streams	Points
2017-01-01	100.5	0	0.5025
2017-01-02	100.5	0	0.5025
2017-01-03	100.5	0	0.5025
2017-01-04	100.5	0	0.5025
2017-01-05	100.5	0	0.5025
...	...	...	...
2021-07-13	100.5	0	0.5025
2021-07-14	100.5	0	0.5025
2021-07-15	100.5	0	0.5025
2021-07-16	100.5	0	0.5025
2021-07-17	100.5	0	0.5025

1606 rows × 3 columns

Rank data frame

In [11]:
```
rank_df = df.groupby(['Artist'])['Points'].sum().sort_values(ascending=False)
rank_df = pd.DataFrame(rank_df, columns=['Rank'], index=rank_df.index)
rank_df['Rank'] = range(1, 1128)
```

In [12]:
```
rank_df.head()
```

Out[12]:

Artist	Rank
Post Malone	1
Ed Sheeran	2
Billie Eilish	3
Drake	4
Ariana Grande	5

Streams data frame

In [15]:
```
artists_dict = {}
for artist in rank_df.index:
    sum_df = df.loc[df.Artist == artist, ['Date', 'Streams']].groupby(['Date'])['Streams'].sum()
    fill_df = empty_df.copy()
    fill_df.loc[sum_df.index, 'Streams'] = sum_df.copy()
    artists_dict[artist] = fill_df['Streams'].copy().cumsum().resample('W').max()
```

Save the datasets

In [16]:
```
with open('../data/final/artists_dict.pickle', 'wb') as artists, open('../data/final/rank_df.pickle', 'wb') as rank:
    pickle.dump(artists_dict, artists)
    pickle.dump(rank_df[:100], rank)
    artists.close()
    rank.close()
```

In [ ]:
```

```