Syracuse University

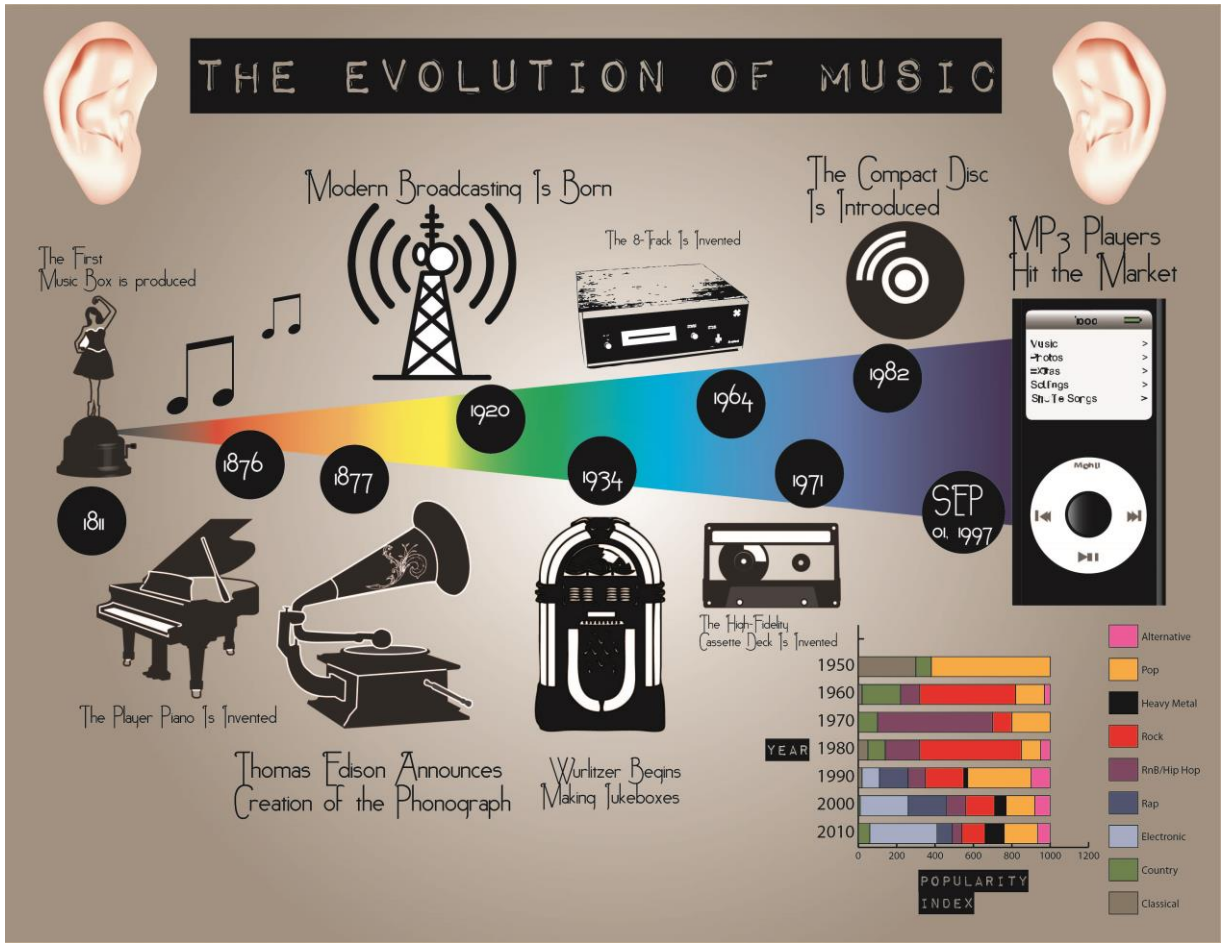
Predicting Song Popularity

Project Introduction

- The evolution of the music industry, driven through technology, has drastically changed the genres and accessibility to music
- More recent developments allow for song attributes to be extracted from an audio file

Our Goal:

- Allow for artists to be able to accurately predict the popularity of a song through an analysis of the song attributes



https://www.google.com/url?sa=i&url=https%3A%2F%2Fsearch.zonealarm.com%2F%3Fq%3Devolution%2520of%2520music&psig=AOvVaw3-3dYiDhBiTaVjAy6FEDVX&ust=1623783591373000&source=images&cd=vfe&ved=0CA0QjhxqFwoTCIjRhrjnl_ECFQAAAAAdAAAAABAd

The Data Sets

Sources

- Kaggle
  - song_data.csv
  - song_info.csv
- Web Scraping Performed in R Studio
  - master_song_T.csv
  - master_artist_T.csv

```
# read in the 4 datasets used for this project
# the first two data sets can be downloaded from kaggle
# https://www.kaggle.com/edalrami/19000-spotify-songs/discussion/73524

song_data=pd.read_csv('song_data.csv')
song_info=pd.read_csv('song_info.csv')

# the second two datasets are provided seperately
# these two datasets come from web scraping wikipedia

master_artist=pd.read_csv('master_artist_T.csv')
master_song=pd.read_csv('master_song_T.csv', encoding = ('ISO-8859-1'))

# used ISO 8859-1 because without it, i received a UTF-8 error.
# The ISO 8859-1 is a single byte encoding that can represent the first 256 Unicode characters
```

Data Definition

Song_data

```
# song_name          - the name of the song
# song_popularity    - the higher the value the more popular the song is
# song_duration_ms   - the length of the song measured in milliseconds
# acousticness       - the higher the value the more acoustic the song is
# danceability       - the higher the value, the easier it is to dance to
# energy             - the higher the value, the more energtic the song is
# instrumentalness   - the higher the value, the more instrumental the song is
# key                - description not provided
# liveness           - the higher the value, more likely a live recording
# loudness           - the higher the value, the louder, measured in dB
# audio_mode         - description not provided
# speechiness        - the higher the value the more spoken word in the song
# tempo              - the tempo of the song measured in beats per minute
# time_signature     - description not provided
# audio_valence      - the higher the value, the more positive mood
```

Song_info

```
# song_name          - the name of the song
# artist_name        - the name of the corresponding artist(s)
# album_names        - the name of the corresponding album(s)
# playlist           - the name of the corresponding playlist(s)
```

Master_song

```
# song_id           - identifier for the songs
# song_name          - the name of the song
# artist_name       - artist of the song
# song_single        - binary whether the song is a single or not
# song_released      - the year that the song was released in
# song_genre         - the corresponding genre(s) for the song
# song_label         - the corresponding label(s) for the song
# song_songwriter    - the corresponding songwriter(s) for the song
# song_producer      - the corresponding producer(s) for the song
```

Master_artist

```
# artist_name        - the name of the artist
# birthday           - the date that the artist was born
# country            - 2 values - either USA or foreign
# startyear          - the year that the artist started making music
```

Data Preparation

- Merge all 4 data frames into one comprehensive data frame labeled ‘song_main’
- Remove all NA and duplicate line items
- Final data frame size: 5,260 rows x 21 columns

```
# combine all the dataframes into one
song_main = master_song.merge(song_info, how = 'outer', on = 'song_name')  # merge master song and song info
song_main = song_main.merge(song_data, how = 'outer', on = 'song_name')  # merge song main with song data

# view the shape of the song main df
song_main.shape
```

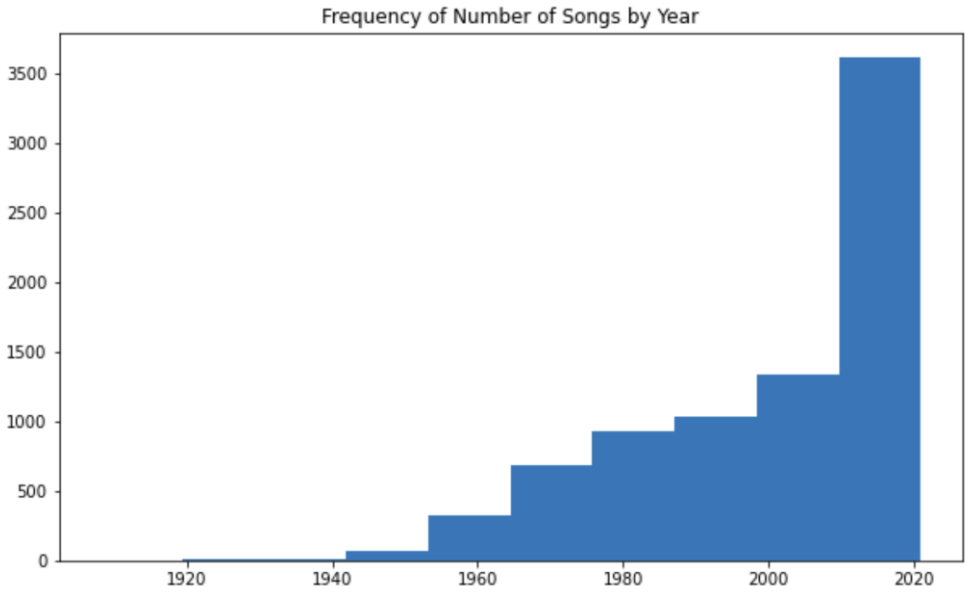
```
(10174, 27)
```

```
# clean up the song_main df

# keep only selected columns
# these are the columns that will be dropped
song_main = song_main.drop(['song_id',
                            'artist_name_x',
                            'song_label',
                            'song_songwriter',
                            'song_producer',
                            '_merge'], axis = 1)

# remove all NAs
# we only want to keep data where the information was available from wikipedia
song_main = song_main.dropna()

# view the cleaned df dimensions
song_main.shape
```

```
(5260, 21)
```

Exploratory Data Analysis & Visualization

Song Distribution

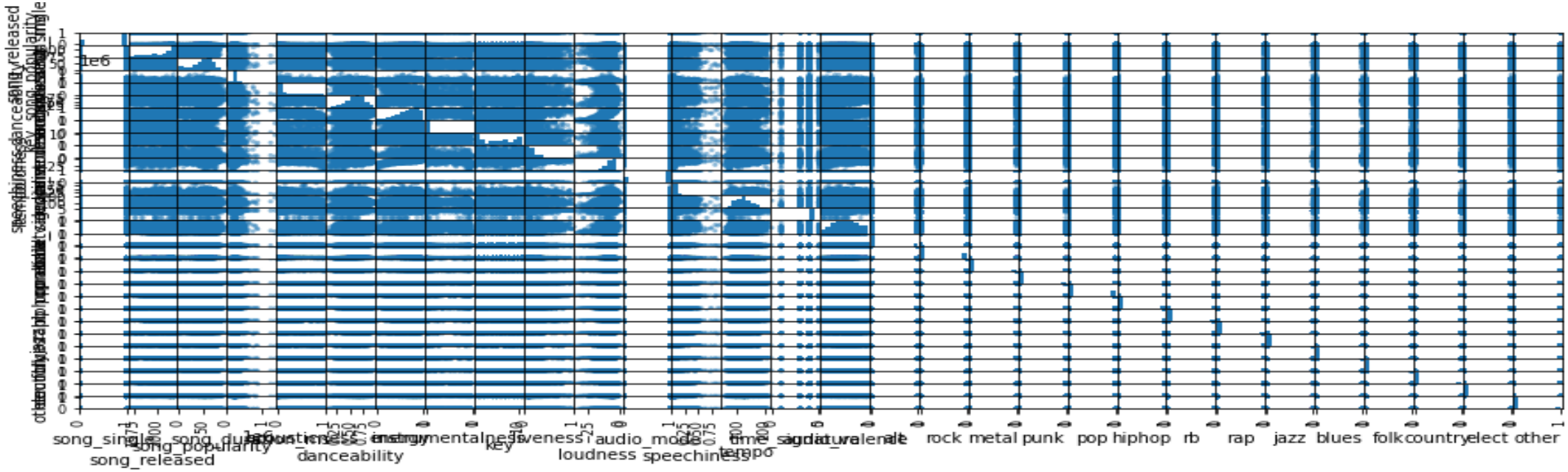


Observations

- There are more newer generation songs than older generation songs
- There is a particularly larger number of songs between 2010 and 2020
- There is a large spike in the number of songs released in 2017-2018
- The bias is probably because the creator of the data prefers newer songs

Exploratory Data Analysis & Visualization

Scatter Plot Matrix



Observations

- There are too many variables compact together to see any relationship

Question 1:
What song genre has the most popular songs?

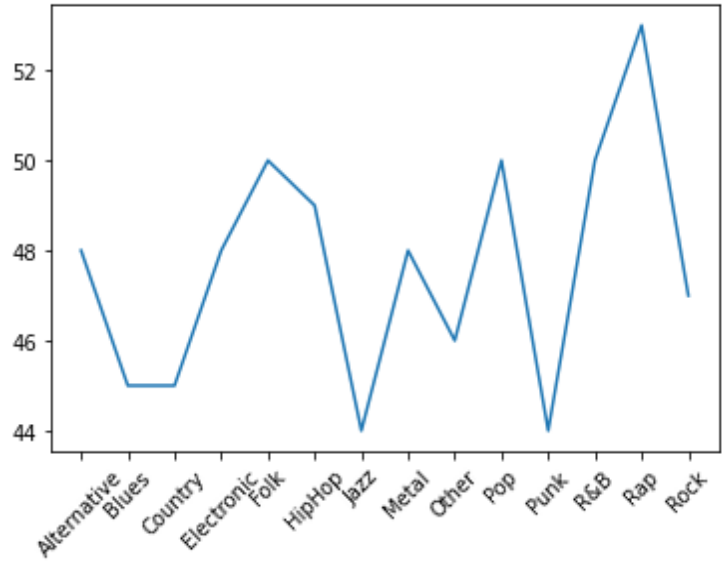
Steps to Answer:

- Unpack the song genre
  - Created a function which searches for a substring

- 14 genres were found for which multiple songs could fit within
  - The function iterated through each of the songs and corresponding genres and if it contained the genre, it will provide a 1 for true or a 0 for false

Results: Rap is the most popular

```
{'Alternative': 48,
 'Rock': 47,
 'Metal': 48,
 'Punk': 44,
 'Pop': 50,
 'HipHop': 49,
 'R&B': 50,
 'Rap': 53,
 'Jazz': 44,
 'Blues': 45,
 'Folk': 50,
 'Country': 45,
 'Electronic': 48,
 'Other': 46}
```



Question 2:
How has song popularity changed over time?

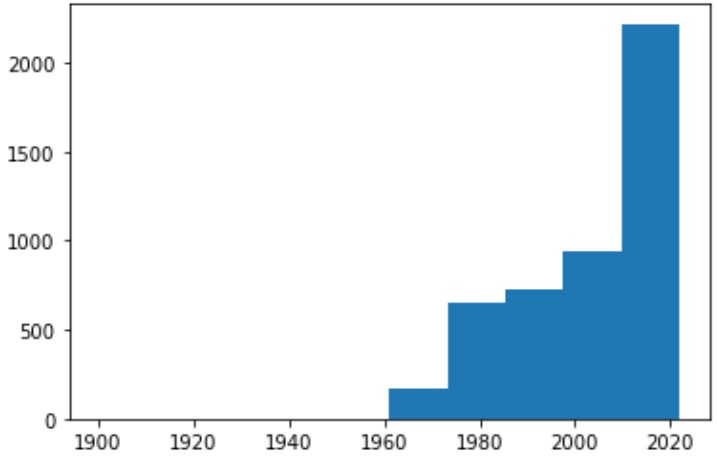
Steps to Answer:

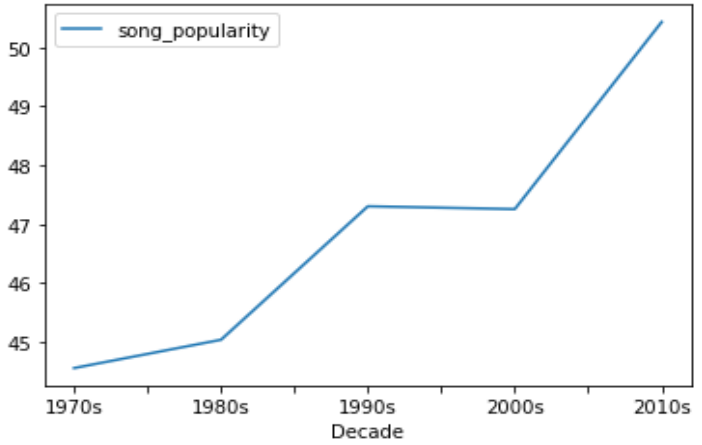
- First, we plotted a histogram to show the popularity of songs each year between 1969 – 2021
- Then, we bin the released year by decades to gain insight on the average popularity by decade

Observations

- The average song popularity was steadily increasing between the 1970s -1990s before plateauing between the 1990 and 2000
- Since then, it has been on the rise, increasing to 50.54% in the 2010s

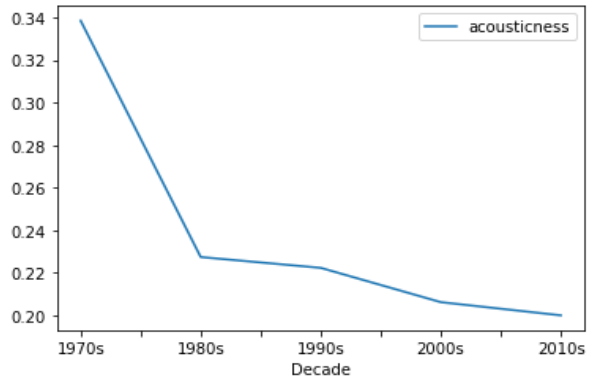
Results

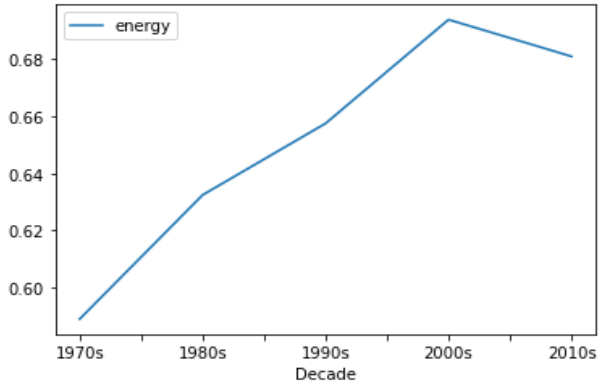


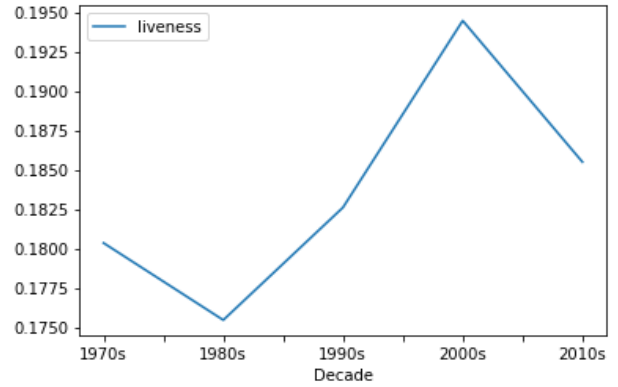


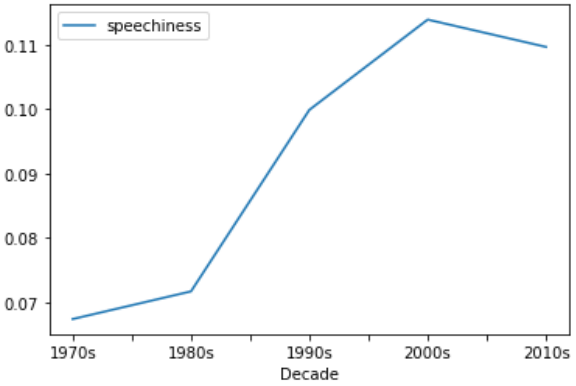
Question 3:
How have song attributes changed over years?

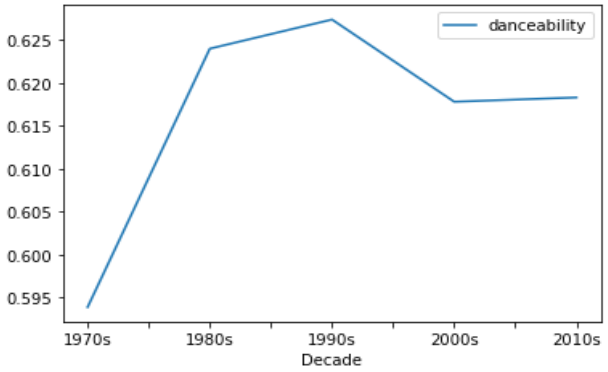
Individual Attribute Visuals

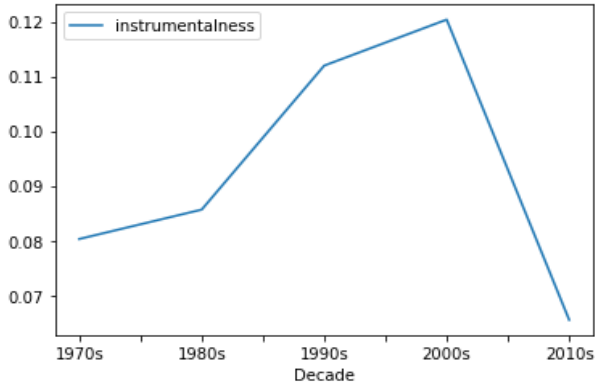


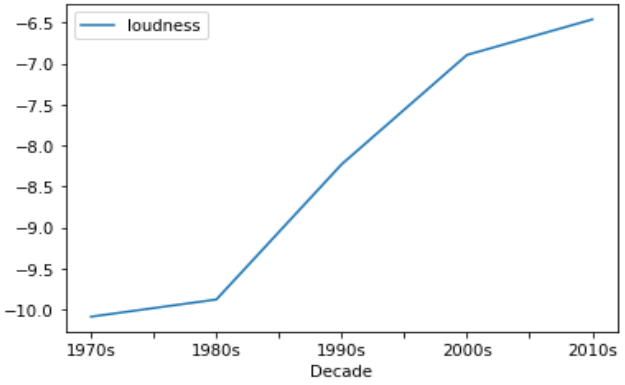






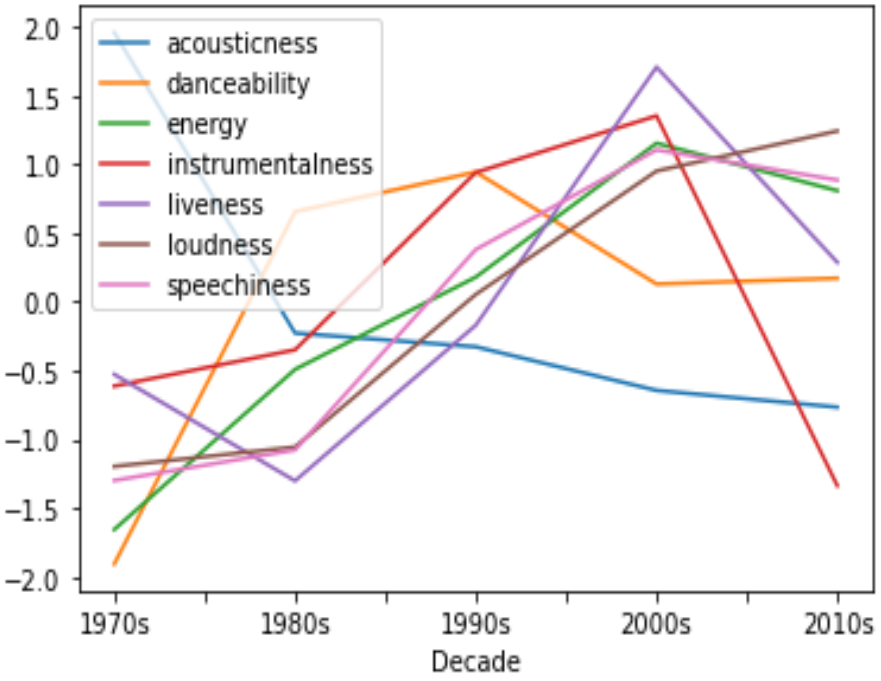






Question 3:
How have song attributes changed over years?

Collective Attribute Visual



Observations

- Acousticness is the one attribute that has been on the decline throughout the decades
  - This is likely the result of electronic instruments becoming more prominent
- Instrumentalness, Liveness, and Speechiness all increased over time before declining in the 2010s
- Energy and Loudness have been heavily increasing over the years.

Question 4:
Are there certain song attributes that correlate with popularity?

Steps to Answer:

- Bin the popular songs from the continuous variable to discrete variables
  - Not Popular, Medium Popularity, High Popularity

- Used the seaborn package based on matplotlib to visually showed the relationship interface between each attribute and song popularity

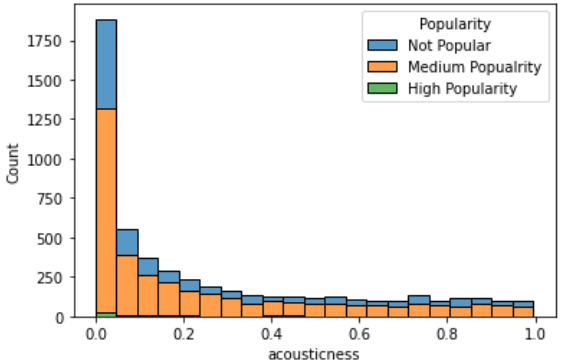
```
#Binning function

#function created to bin the attributes
def binningFunction(col, cut_points, labels=None):
    minval=col.min()
    maxval=col.max()
    break_points= [minval]+cut_points+[maxval]
    print(break_points)
    if not labels:
        labels = range(len(cut_points)+1)
    colBin=pd.cut(col,bins=break_points, labels=labels, include_lowest=True)
    return colBin
```

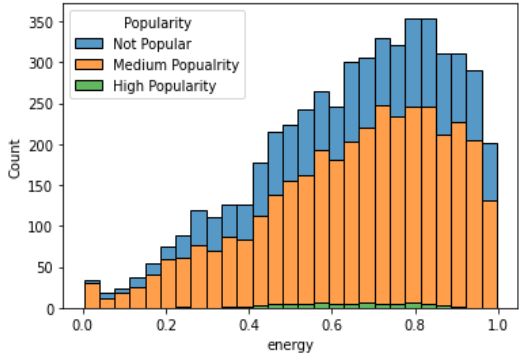
```
# bin song_popularity for later use
cut_points=[40,80];
labels=['Not Popular', 'Medium Popualrity', 'High Popularity']
song_main['Popularity']=binningFunction(song_main['song_popularity'], cut_points, labels)
song_main
```

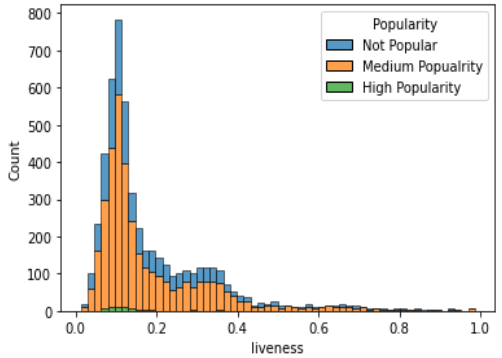
```
[0, 40, 80, 93]
```

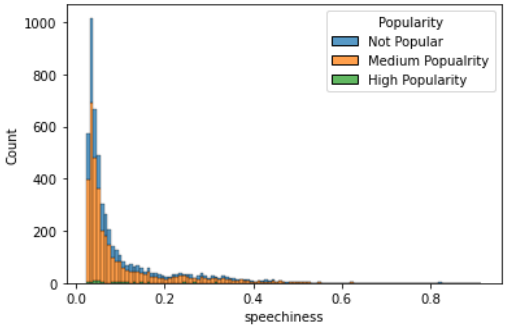
Question 4:
Are there certain song attributes that correlate with popularity?

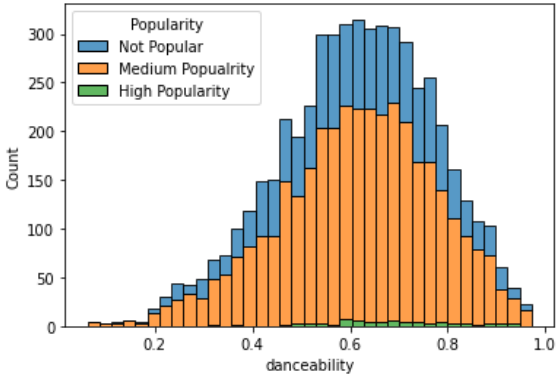
Visualizations

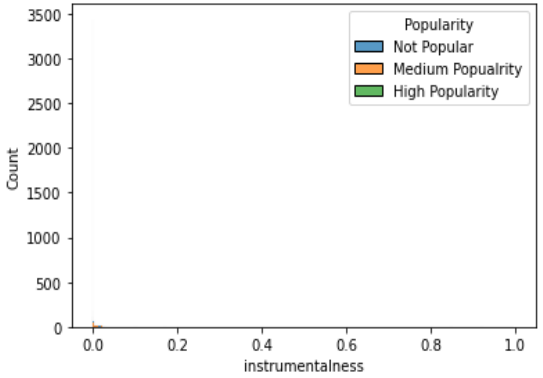


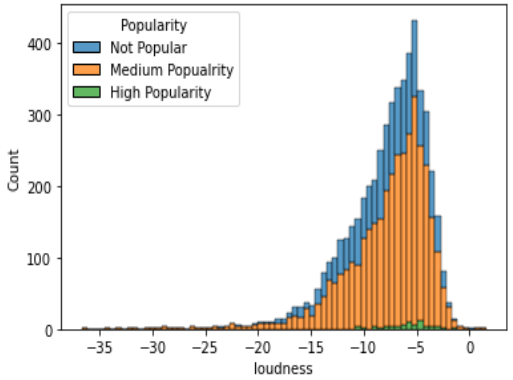


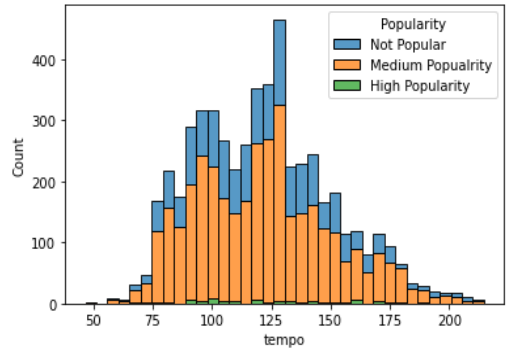


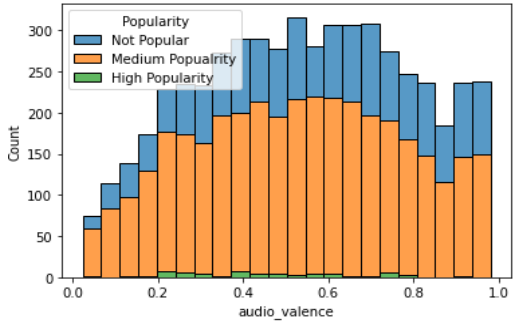












Conclusion

- Popular song do have common attributes such as low acousticness, high danceability, high energy, medium tempo
- Song popularity has changed drastically over the decades, due to culture and generation interest at the time of the song

Next Steps:

- Further collection and cleansing of song data
- Analyze cultural events as they relate to the song release date

https://s.wsj.net/public/resources/images/BN-DF456_STREAM_G_20140612162817.jpg

Syracuse University

Thank you!