CMPT 732 Project Report: Trends in Song Popularity from Global Phenomena

Tianyi Wu, Rebekah Wong, Qingrui (Rachel) Li

I. Problem Definition

Our project is an exploration of recent trends in Spotify song popularity based on global phenomena, which aims to answer the following research questions:

1. **To what extent is the popularity of songs affected by economic factors—namely GDP, unemployment rates, and inflation?** For instance, the COVID-19 lockdowns caused economic fluctuations in 2020, so we want to investigate whether song preferences have changed during this time.
2. **Have songs written in certain languages risen in popularity internationally?** From a North American perspective, English songs are mainstream. However, genres in other languages, such as K-pop, have spiked in popularity in recent years and found success in global music charts, despite Korean being a language native to only Korea.
3. **Is there a correlation between the popularity of song moods—such as happy music or sad music—to the average happiness of people in different countries?**

We chose a total of 18 countries to base our analyses on. To ensure that our samples were more representative, we selected nine developed countries—the United Kingdom, Japan, Canada, Singapore, Australia, Norway, Belgium, Hong Kong and the United States—and nine developing countries—Argentina, Thailand, Brazil, Honduras, Malaysia, Philippines, Mexico, Bulgaria and Turkey. Unfortunately, there is a lack of Spotify data from underdeveloped countries, but we believe that our selected countries were diverse enough, both geographically and linguistically, to extrapolate to the rest of the world. Other challenges imposed by this problem include categorizing songs into genres, identifying the languages of song lyrics, and classifying the moods of different songs.

II. Methodology

a. Top Spotify Songs

First, we used a dataset from Kaggle that contains weekly Spotify track chart data from 2013 to 2023 as our source for extracting the desired Spotify data. The dataset originally included 10 columns such as the date, track ID, artist names, track duration, artist genres, weekly stream counts, and more. The Extract-Transform-Load (ETL) process was performed using Spark.

During the transformation stage, the date column was reformatted into “year” and “week of the year,” and data from irrelevant years and countries was filtered out. By aggregating the number of weekly streams, we obtained the total number of streams for each song. To ensure a fair representation of popularity, we retained only songs that appeared on the chart at least three times within a year. Next, the average weekly streams for each song was calculated by dividing its total annual streams by the number of weeks it appeared on the chart. This metric was used to rank the annual top 200 songs for each of the aforementioned 18 countries and globally.

As there were many unique genres listed in the top songs’ artist genres, we used Spark to group certain genre tags together based on keywords. Then, we printed out leftover genres that had not been included by any of our defined genre categories and repeated this process to manually add more keywords until most genres were accounted for. For example, “Francoton” is a French pop genre that we included in the “pop” category and “zhenskiy rep” is a Russian genre for “women’s rap” that we included in the “rap” category. For each song, we assigned a boolean value for whether this category was included in the list of artist genres, which would later be turned into counts of 1 or 0 to avoid a genre being counted more than once if the artist’s genre list contained duplicate words.

b. Economic Attributes and Happiness Data

Then, we downloaded Excel data for each economic attribute—GDP per capita (USD), unemployment rate (%), and inflation rate (%)—from Statista for every country. Each file included data about the annual economic attribute for that country from around the 1990s until now. We also downloaded happiness data from the World Happiness Report, which included the annual happiness scores out of 10 for every country. During the ETL step, we used Spark to filter out the years to our desired range of 2018 to 2022, add a column for the

country name, and combine all three economic attributes with the happiness score. The result was written into Parquet files hive-partitioned by country to use in the subsequent analysis tasks.

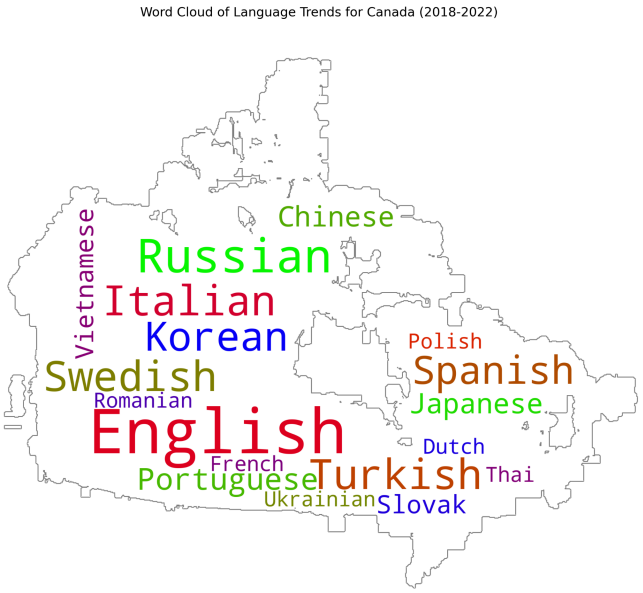
For Task 1, we used this data to calculate the Pearson correlation matrix between genres and economic attributes for the top 150 songs' artist genres, then for the top 10 song genres to compare. To do so, we combined the genre counts and economic attributes for each country and year and turned them into vectors through PySpark ML's VectorAssembler to use in the ML Correlation function, because the correlation function could not be used without vectors. Then, we plotted the results through Seaborn as a heatmap to visualize the correlation, since Seaborn generally produces visually appealing graphs. We also wanted to plot a stacked bar chart of the genre counts mapped with lines of the economic attribute trends, and Matplotlib seemed to be the primary library to do so in Python. However, after some research, it did not seem possible to plot our results without turning our Spark DataFrames into Pandas DataFrames, which we initially wanted to avoid as Pandas would be less efficient to run. To mitigate this, we left it as the last step of our analysis and made sure to close the Matplotlib plot after saving every graph to prevent memory issues and to ensure that it was still scalable even if we added more countries to our list.

c. Song Lyrics and Language Data

In Task 2, our analysis primarily relied on three data sources: the aforementioned Top 200 Songs dataset, which had been pre-cleaned; the Genius Song Lyrics Kaggle dataset, which contained a substantial number of songs with their corresponding lyrics; and the Genius API, which we used to retrieve missing song lyrics not included in the Kaggle dataset. The ETL process started with merging the Top 200 Songs dataset with the Genius Song Lyrics dataset by matching artist names and song titles, then partitioning the data. This process was executed using AWS EMR to handle large-scale data efficiently, with results stored in Amazon S3. Songs were then categorized into two groups: songs with lyrics and songs without lyrics. For the "songs without lyrics" category, missing lyrics were retrieved using the Genius API. Once we obtained the lyrics, we performed language detection using the pycld2 and langdetect libraries. The results from both libraries were cross-validated to ensure reliable and accurate language identification. Finally, the processed data—including metadata, track IDs, lyrics, and language information—was stored in a data cache on Amazon S3. The data was organized into a structured hierarchy, partitioned by language, ensuring efficient retrieval and scalability for future analysis.

```
s3://project-spotify-songs/
    ├── parquet_by_lan/
    │   ├── language="en"/
    │   │   ├── songs.parquet
    │   ├── language="it"/
    │   |   ├── songs.parquet
    .......
```

In Spark, the language data was joined with the top Spotify songs data we obtained previously and grouped by year, country, and language for language distribution analysis. Utilizing Matplotlib and Seaborn, we created heatmaps to demonstrate the global language usage among the annual top 200 songs from 2018 to 2022. These heatmaps effectively highlighted trends and shifts in language preferences over time. Moreover, we created country-shaped word clouds to visually represent the popularity of different languages in music for each country.



d. Sentiment Analysis of Songs

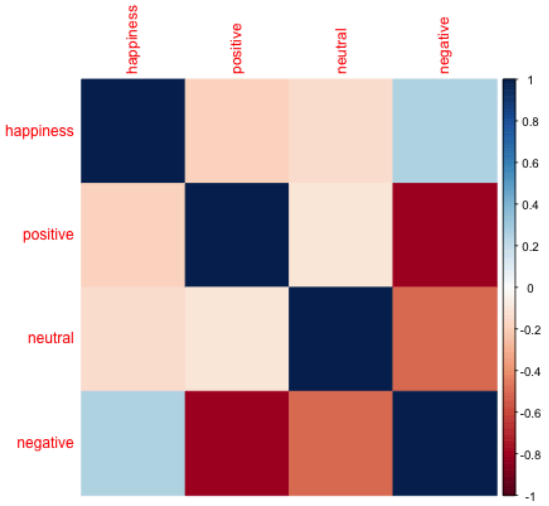
To ensure that our sentiment analysis was meaningful and statistically relevant in Task 3, we established a threshold requiring at least 10 data points with lyrics in any given language for the language to be considered. Categories with less than 10 data points were excluded from the sentiment analysis to prevent biased or unreliable results due to insufficient data. For the sentiment analysis, we analyzed the lyrics and classified them into three categories: “positive”, “neutral”, and “negative”. To do so, we utilized pre-trained sentiment analysis models on Hugging Face. However, because many of the pre-trained models were language-specific, using a single model for multilingual data would be difficult. Thus, we selected and tested multiple language-specific models to ensure accuracy.

To make sure that we comprehensively covered each language, we compared different models for them. For languages with multiple suitable models, we conducted one-on-one tests using example sentences representing a range of emotions, including very positive, generally positive, neutral, generally negative, and very negative. This allowed us to identify the best model for each language. If the models provided outputs in a different format than what we needed, such as providing a 1 to 5 star rating on emotions, we recalibrated the output. For instance, we would test the model with example sentences and map the model’s output to our desired categories of “positive”, “neutral”, and “negative”. This process ensured that the selected models were consistent with our sentiment categorization goals. Overall, we analyzed and used 13 different models to ensure complete language coverage and generate accurate sentiment results.

```
s3://project-spotify-songs/
├── mood_data/
|   ├── language="en"/
|   |   ├── songs.parquet
|   ├── language="it"/
|   |   ├── songs.parquet
.......
```

The pre-processing step for Task 3 was done in Spark. Firstly, we performed an inner join on the sentiment data and the top songs. Due to the limitation of the lyrics data, there were some missing values so we needed to determine a threshold for the number of songs to be used in the analysis later. After proper aggregation, it was found that the lowest track count was 114 for Thailand in 2018, hence we only retain 110 songs per year for each country to maintain consistency. Next, the frequency of each mood category (positive, negative, neutral) was calculated, resulting in a DataFrame with five columns: country, year, number of positive songs, number of negative songs, and number of neutral songs. This DataFrame was then joined with the happiness data, forming the final dataset for analysis.

Given the manageable size of the dataset, a small data tool, R, was chosen for its statistical capabilities to conduct the analysis. A correlation matrix was calculated to investigate relationships between variables. Results revealed a strong correlation between "positive" and "negative" moods, while the happiness score showed little to no correlation with song moods. To address multicollinearity, a new variable, "comb," was defined as a linear combination of "positive" and "negative" moods. Also, we explored the impact of a country's development status (developed vs. developing) on happiness using a boxplot and linear regression model. Since the results indicate that developed countries have a higher happiness score on average, it is logical to examine the relationships between happiness and song sentiment for developed and developing countries separately. A new variable, "in_one," was created as a linear combination of all three moods for modelling purposes.



III. Problems

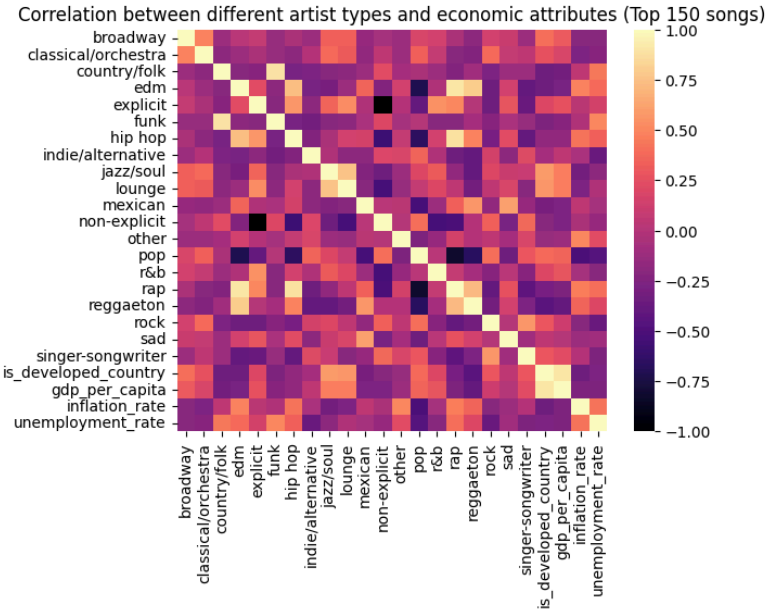
Budget constraints were a major challenge that affected our approach to processing the data. In order to manage costs effectively, most of our data processing was carried out on local computers during the initial stages to minimize our reliance on cloud computing services for tasks that could be performed locally. However, all the provided code and data subsets were tested on Amazon EMR for the lyrics and language detection, as well as SageMaker for the lyrics-based machine learning sentiment analysis, to ensure future compatibility and scalability with larger datasets. This approach balanced cost-effectiveness with readiness for database expansion and scalability.

During the ETL phase, the first problem we encountered was that Spark did not natively support reading Excel files, but the economic and happiness data were not available to download in Spark-supported formats such as CSV or JSON. After some further research, we decided to use the Crealytics Spark Excel library to read Excel files into DataFrames directly. We later realized that the data provided by Spotify only lists the artist genres associated with every song, which could include an upwards of at least 10 different genre tags, but not the genre of the song itself. In an attempt to solve this, we also compiled a smaller dataset containing 900 entries from manually looking up song genres from Rate Your Music for the top 10 songs to have both to analyze. We would have liked to scrape this data from the website directly using the RymScraper API, but unfortunately, Rate Your Music added an extra layer of protection to their website earlier this year to block scraping, so the API is no longer functional.

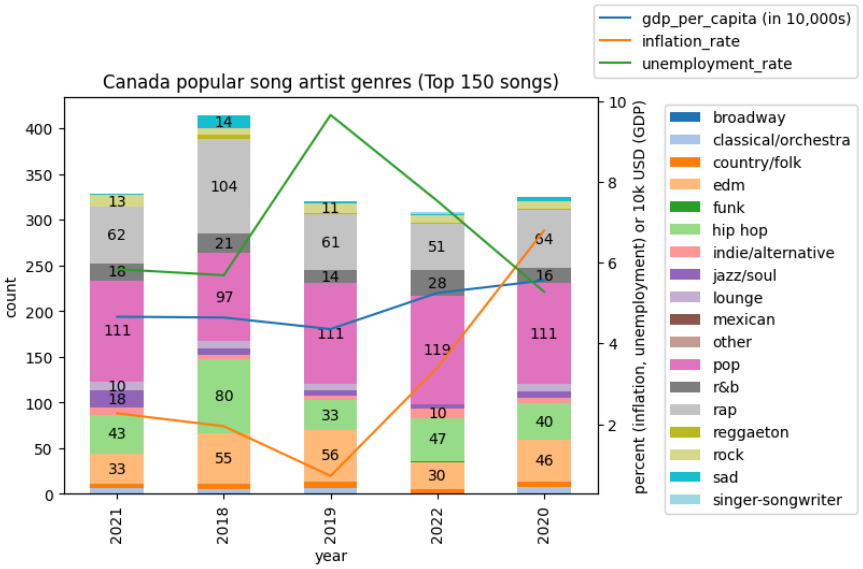
Another problem was caused by the Genius API's rate limitation and budget constraints, as it was very difficult to efficiently fetch lyrics at a larger scale. To mitigate these issues, we first connected the data to an existing database containing lyrics, identified songs without lyrics, and executed API requests locally. This approach allowed us to control the request rate and implement a latency mechanism to manage rate limitations. Processing these API requests directly on AWS EMR would have significantly increased costs due to prolonged waiting times for the API to reset, making it impractical with our student budget.

IV. Results

From working on the implementation, there did not appear to be very strong correlations between the economic status of a country and how popular certain genres were in Task 1; they were relatively consistent from year to year. Interestingly, the resulting heatmap for the top 150 songs' artist genres suggested that Reggaeton, for instance, is quite negatively correlated with GDP per capita, and more highly correlated with inflation and unemployment rates. This is not necessarily true across all countries, because Reggaeton is a genre with close ties to Latin American culture. Reggaeton was the dominant genre in Argentina, Honduras, and Mexico, but nowhere else—which all happen to be developing countries with poorer economic conditions.

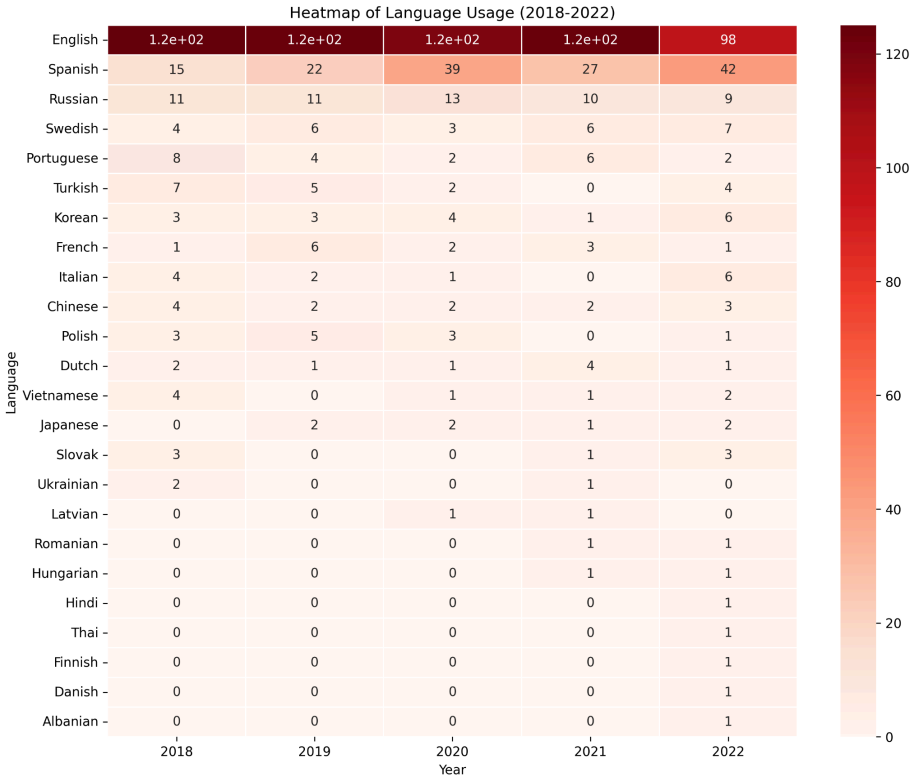


Similarly, pop music was highly prevalent in most countries, but the heatmap showed that it was positively correlated with GDP per capita, and negatively correlated with inflation and unemployment rate. On the other hand, pop music was generally more revered in developed countries, which is correlated with better economic conditions. The musical genres of smaller, developing countries seemed to use more keywords that were unrelated to pop, but pop was still generally loved worldwide. Despite our efforts, the specific top 10 song genre heatmap was less useful because the manual dataset had too little data to find meaningful insights.

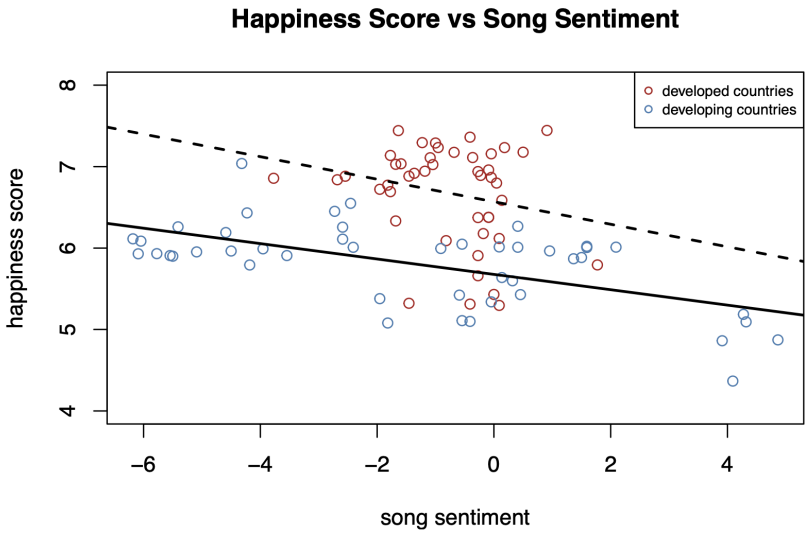


For Task 2, the heatmap revealed that while Spanish songs gained rising popularity, English remained the dominant language. Other languages did not exhibit clear trends over the years. However, there was a noticeable increase in the diversity of languages represented in the top 200 songs. From gathering top 10 song data in Task 1, we noticed that there were cases of music in foreign languages becoming more popular in certain countries, such as Hong Kong—a country with Chinese and English as their official languages—having four

Korean pop songs in their top 10 in 2020 during the peak of COVID-19. We believe that the diversity in music will only improve with time as some of the recent top songs have featured multiple languages through collaborations with artists across continents. Examples of this include “My Universe” by Coldplay and BTS and “Ice Cream” by Blackpink and Selena Gomez, both of which include a blend of English and Korean lyrics.



For Task 3, the happiness scores showed little correlation with song moods overall. However, based on the observation that developed countries tend to have higher happiness scores on average, both the fitted linear regression model and the correlation matrices suggested that a moderate negative correlation exists in developing countries: happier developing countries tended to listen to more negative songs. In summary, there were no significant correlations observed in general or in developed countries.



Future extensions of our work could include expanding our top song data to other music streaming platforms, such as YouTube Music or Apple Music, as Spotify listeners are not necessarily representative of the entire population. In addition, we could add more countries to analyze so that our samples become more accurate as well. For future dataset expansions and larger budgets, we could consider using more data caches created on S3 along with distributed computing frameworks such as AMS EMR to process and integrate large datasets efficiently. This approach would ensure that our solution is robust, cost-effective, and scalable for

future data growth. In future updates, our workflow could consist of checking whether the song and lyrics already exist in the data cache; if not, the missing lyrics would be retrieved through the API to detect the relevant language before updating it in the data cache. Doing so would help us efficiently process new data while minimizing any redundancy. The same applies to our sentiment analysis task, which could also be similarly scaled to handle larger datasets through a data cache architecture on Amazon S3. By integrating with Amazon SageMaker, sentiment analysis could be efficiently applied to new data not already present in our existing lists and stored directly in the data cache, ensuring seamless scalability and efficient processing.

V. Project Summary

Project Area	**Points (Total: 20)**
Getting the data	4
ETL	4
Problem	1.5
Algorithmic work	1.5
Bigness/parallelization	1
UI	2
Visualization	2
Technologies	4