DATA VISUALIZATION

MEMUDU Alimatou Sadia LEE Hyelim

December 12, 2022

Table of Content

1 Presentation

The WASABI Song Corpus is a large corpus of songs enriched with metadata extracted from music databases on the Web, and resulting from the processing of song lyrics and from audio analysis. A summary of the data contained in The WASABI Corpus raw data used for this project is as follow:

- Album Data : (7798, 27)
- Artist Data : (3000, 50)
- Song Data : (76027, 60)

The github repository of this project contains all the files used for this project.

1.1 Users

People who are interested in the distribution of music genres in the world during a certain period of time.

1.2 Goal

The mutual goal of this project is to provide a general repartition of wasabi data in terms of genre of music in the world from 1980 to 2015 to the users.

2 Data Preprocessing

2.1 Feature Selection

We fetch our raw data from three files which regrouped the 3000 data from the albums, song and artist from the wasabi dataset, there are album_all_artists_3000.rds, songs_all_artists_3000.rds, wasabi_all_artists_3000.rds which can be downloaded from this link. From each file, we selected our preferred variable which we find useful for our project.

- Album
  - id_artist
  - id_album
  - genre
  - publicationDate
  - deezerFans
- Artist
  - id
  - gender
  - lifeSpan.ended
  - locationInfo
- Song
  - id,
  - id_album
  - publicationDate

- availableCountries

Here, we perform some preprocessing techniques using R in the file named features_selection.R by replacing the missing and empty values by NA, then proceed by saving the updated files into three different csv files named songs_features_final, artists_features_final and albums_features_final which are going to be used for the data cleaning.

- songs_features_final: (47595 rows, 4 columns)
- artists_features_final : (3000 row, 4 columns)
- albums_features_final : (7798 rows, 5 columns)

2.2 Data Cleaning

- Artist

- Splitting locationInfo column into 3 column – country, state, and city

The locationInfo column represents the artist's location information, and the country-state-city is divided into commas. We extracted only country from the string because we needed only country among them and deleted the used locationInfo column.

- Replacing name of some contries for easy geographical location

In the country data extracted earlier, the names of some countries were changed to more useful names. The name changed to be used as a reference key when combining external data and data is as follows.

Before	After
England	United Kingdom
Scotland	
Wales	
Northern Ireland	
The Netherlands	Netherlands

Table 2.1 : list of replaced name in country column

- Rename several columns with tag from artist data

Column id was renamed to id_artist in order to facilitate identification of the source of the overlapping column names when data is merged.

- Song

- Extract year from publication data and rename column

The publicationDate column indicates song release date information, and the year-month-day is divided into '-'. We extracted only a year from the date because we needed only one year among them.

- Create new column for number of availableCountries

The availableCountries column consists of a list of character strings. To use this for visualization, a count_availablecountry column was created by counting strings based on commas.

- Rename several columns with tag from song data

Some columns were renamed from id to id_song and publicationDate to publicationDate_song in order to facilitate identification of the source of the overlapping column names when data is merged.

- Album
  - Rename several columns with tag from album data

Some columns were renamed from publicationDate to publicationDate_album, genre to genre _album, deezerFans to deezerFans_album in order to facilitate identification of the source of the overlapping column names when data is merged.

2.3 Merging Data

After merging the data, the missing values of some columns(gender) were changed to Unknown. In addition, after merging the data, the data including missing values in the country and publication Date columns were deleted. After that process, 41,129 rows of data remained.

2.4 Transforming Data

- Group and summarize data by year,gender,country,genre_album,lifeSpan.ended

We selected several columns to be used for future visualization from the merged data and summarized data for each category. Since the deezerFans_album column has different attributes used for each visualization, columns using sum and mean were created separately. The methods used for each column in summarizing are as follows.

Method		**Column name(before)**	**Column name(after)**
group_by		year, gender, country, genre, lifeSpan.ended	year, gender, country, genre, lifeSpan.ended
summarize	n_distinct	id_song	count_song
	n_distinct	id_album	count_album
	n_distinct	id_artist	count_artist
	sum	deezerFans_album	deezer_fans
	mean	count_availablecountry	average_availableCountries
	mean	deezerFans_album	average_fans

Table 2.2 : list of grouped and summarised features

- Create New column – lat, lon, continent

The latitude and longitude of each country's capital was required to place the bubbles on the map. This was done using a dataset containing all of the countries in the world, along with their capitals and the coordinates of the capital. The cleaned data was given three new columns, latitude, longitude, and continent.

- Categorize values of genre – 182 to 9

There are many genres of wasabi data sets. There are a total of 182 unique variables in the genre of our dataset, summarized in 3000 rows. This has a problem in that visibility is poor due to too few values for each category during visualization. Therefore, categorizing was conducted based on the fact that some genres originated from main genres. For example, "Rock" has subgenres such as "Soft Rock", "Rap Rock", and "Pop Rock". Therefore, we clustered dependent genres under 9 representative genres for visualization and replaced missing values by unknown.

Main Genre	Sub Genre	Count
Rock	Rock, Soft Rock, Rap Rock, Pop Rock, Piano Rock, Symphonic Rock, Electronic Rock, Progressive Rock, Indie Rock, Industrial Rock, Gothic Rock, Glam Rock, Hard Rock, Deutschrock, Experimental Rock, Folk Rock, Art Rock, J-Rock, Psychedelic Rock, Alternative Rock	636
Hip Hop	Hip Hop, Trip Hop, Southern Hip Hop, Trip Hop, Hardcore Hip Hop, Christian Hip Hop, Australian Hip Hop	87
Pop	Pop, Electropop, Synthpop, Psychedelic Pop, Power Pop, Experimental Pop, Pop Punk, Indie Pop, French Pop	267
Folk	Folk, Filk	45
Metal	Black Metal, Glam Metal, Melodic Death Metal, Metalcore, Progressive Metal, Power Metal, Folk Metal, Death Metal, Heavy Metal, Gothic Metal, Nu Metal	228
Jazz	Jazz	47
Country	Country	91
R&B	R&B	43
Others	Genres not in the above category.	740

Table 2.3 : a list of Category of Genre

- Re-Categorize values of continent – Central America to North America

As a result of exploring the unique value of the continent column, the category of Central America seemed somewhat ambiguous. Therefore, we categorized the continent to North America instead of Central America. We changed Canada, Guatemala, Mexico, and the United States, which are divided into Central America, to North America.

The data cleaning process was performed in the data_cleaning.R file, At the end of the data cleaning phase, we saved the transformation into a csv file named final_data composed of 3048 rows and 14 columns which are year, gender, country, genre, lifeSpan.ended, count_song, count_album , count_artist, deezer_fans, average_availableCountries, average_fans, lat,lon and continent.

2.5 Data Preprocessing Workflow

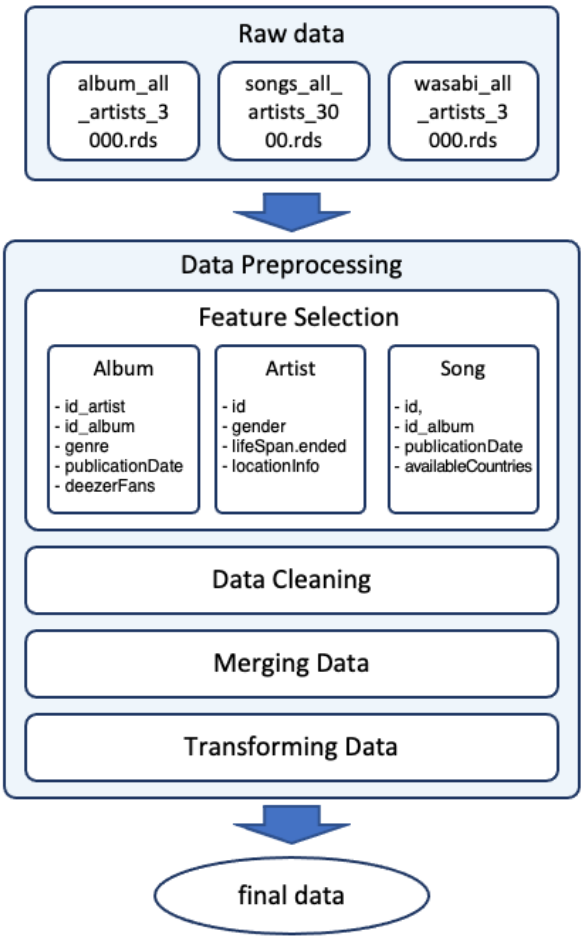

Figure 2.4 : Data preprocessing Workflow.

3 Data visualization By Memudu Alimatou Sadia

3.1 Overview

3.1.1 Users

My users are preferably people who are interested in digging up the distribution of genres of music and the number of Deezer fans during a certain period in the world to acquire knowledge about music in the World. Deezer is a French online music streaming service for music listening.

3.1.2 Goal

My visualization goal is to provide useful information about the repartition of wasabi dataset in terms of genre of music by continent in the world, from 1980 to 2015. This information groups the number of songs, albums , artists and Deezer fans by genre of music albums around the world.
In order to achieve this goal, bubble map and stacked bar graph visualization techniques were chosen for this project. The two visualizations share the same filtering input. This application was built from scratch using R shiny and composed of 2 main sections as displayed in figure 3.1 , the feature setting section by the left which allows the user to select the features depending on their preferences shown in figure 3.2, The bottom part of the data filtering section is the display of the number of artists, song and albums present in the data after the filtering performed and the visualization section where the visualization is displayed. The visualization section is composed of 2 tabs where each tab displays a different visualization technique.

- Visualization Techniques : Bubble Map and Stacked Bar Graph
- Environment : R
- Utilized libraries :
    - library(shiny) : Implementing R-shiny Applications
    - library(highcharter): User for the bubble map implementation.
    - library(dplyr):filtering and selecting data to be used for visualization.
    - library(ggplot2):For producing the stacked bar graph.
    - library(plotly): for displaying the stacked bar graph.

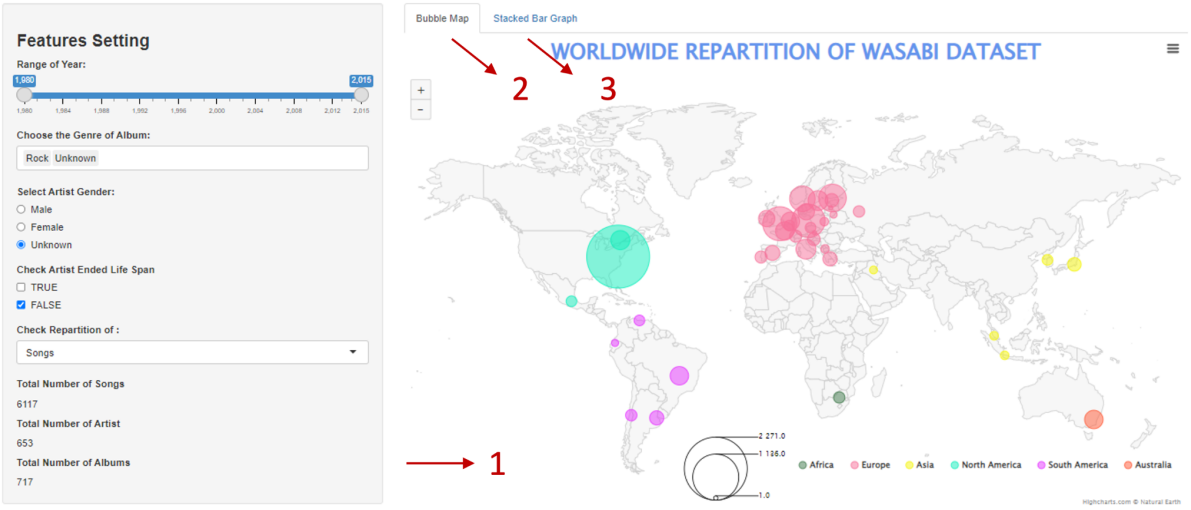

Figure 3.1: Application Interface

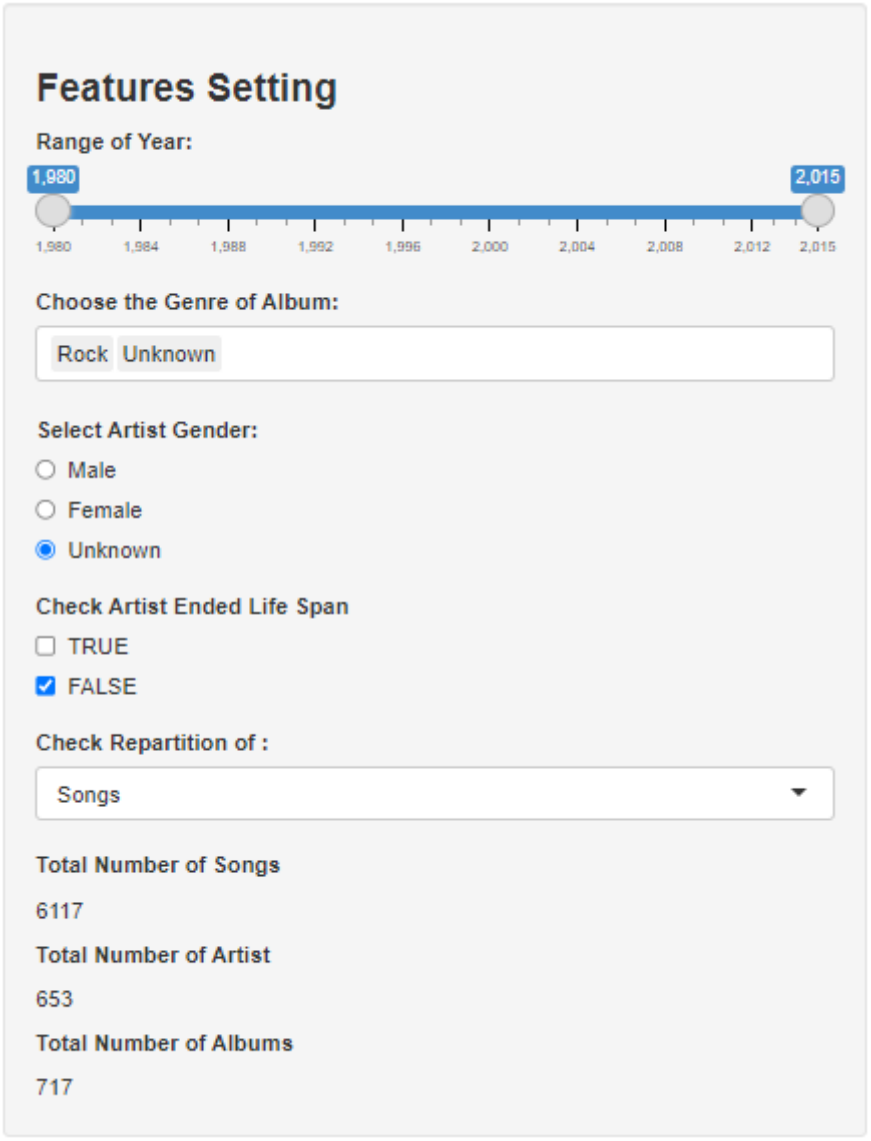

Figure 3.2: Feature Settings

3.2 Visualization Techniques

3.2.1 Bubble Map visualization

The visualization technique displays the number of songs, albums and artists all around the world where each continent is associated with a different color. The number of songs per country is identified by the size of the bubble. where each bubble belongs to a specific country and continent as shown in figure 3.3. Using the mouse you can zoom into any part of the part for easy visualization and by putting the mouse over a bubble you can see the continent, the country, the number of artists, songs of the country that correspond to that bubble as shown in figure 3.4.

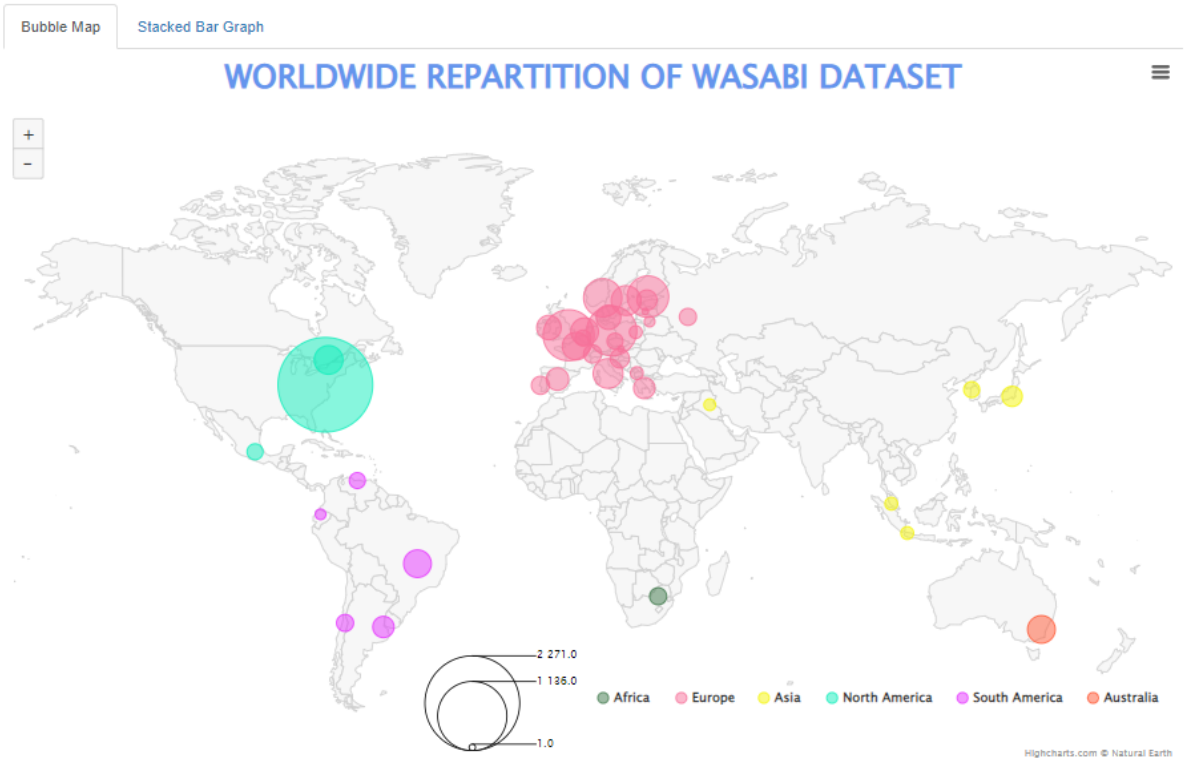

Figure 3.3: Bubble Map

3.2.1 a- User task

Table 1 details the actions available for the user in order to ease his navigation of the bubble map.

User task	Feature	Details
Overview		An overview of the number of songs and albums released with a specific number of artists that produced it.
Filter	Year Range	Display information according to the year by country
	Genre	Select or remove the genre of album from the dropdown
	Gender	Choose the gender of the artist
	Life ended	Show the information about the artists whose career ended or not
	Check repartition	Choose to display information about only songs or albums ot both
Point	Bubble	To see the information about the name od the continent, the number of artists, songs and albums
Click	Bubble Map	To view a Map , the legend to view only the repartition of a specific continent , the + and - to zoom in and out.

Table 3.1: User task of Bubble Map

3.2.1 b- Visualization Mapping

- *size of the Bubble:* represents the number of songs of each country
- *color of the bubble:* represents the continent of the country.

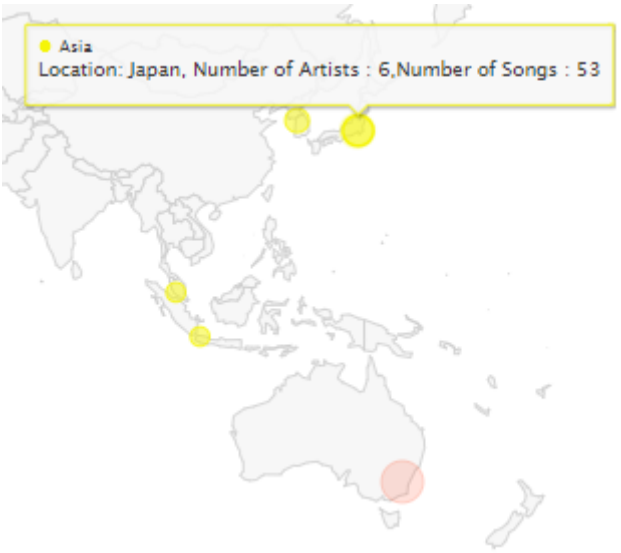

Figure 3.4: Map Bubble Tooltip

3.2.2 Stacked Bar Graph visualization

In the aim of providing necessary information to the users, the stacked bar graph visualization technique is to allow the user to have a glance of the number of Deezer fans each album from the result of his filtering generated. This visualization technique shows the information about the number of Deezer fans each album from the range of year specified in the Feature Setting section of the application results grouped into 10. The minimum range of the year for this is 10 years. By putting the mouse over each bar the number of years and the exact number of Deezer fans per continent in the world is displayed as shown in figure 3.5.

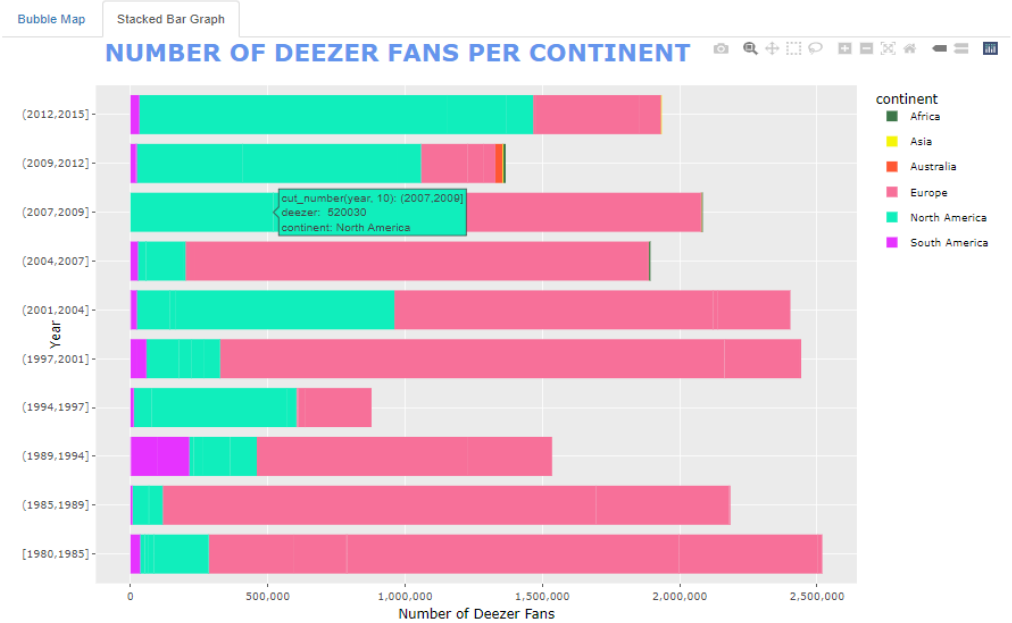

Figure 3.5: Deezer fans stacked bar Plot

3.2.2 a- User Task

Table 2 details the actions available for the user in order to ease his navigation of the stacked bar graph.

User task	Feature	Details
Overview		An overview of the number of Deezer Fans per albums released over the years.
Filter	Year Range	Display information according to the year by country
	Genre	Select the genre of album
	Gender	Choose the gender of the artist
	Life ended	Show the information about the artists whose career ended or not
Point	Bar	To see the information about the number of Deezer fans and the year
Click	Stacked Bar Graph	To view the bar graph, take a picture of the graph for future purposes and zoom in and out of the graph.

Table 3.2: User task of Stacked Bar Graph

3.2.2 b- Visualization Mapping

- *color* : represents the continent the pointed bar corresponds to.
- *size:* represents the number of Deezer fans per continent.

4 Data Visualization by LEE Hyelim

4.1 Overview

4.1.1 Visualization Goal

My visualization goal is to show the preference for music by country according to the year. Through this visualization, the user can get information about preferences of music through the number of music released, the average of available countries, and the average number of fans per country and year. It also aims to visualize each figure by country to make it easier to compare, making information more intuitive to understand.

4.1.2 Visualization Techniques

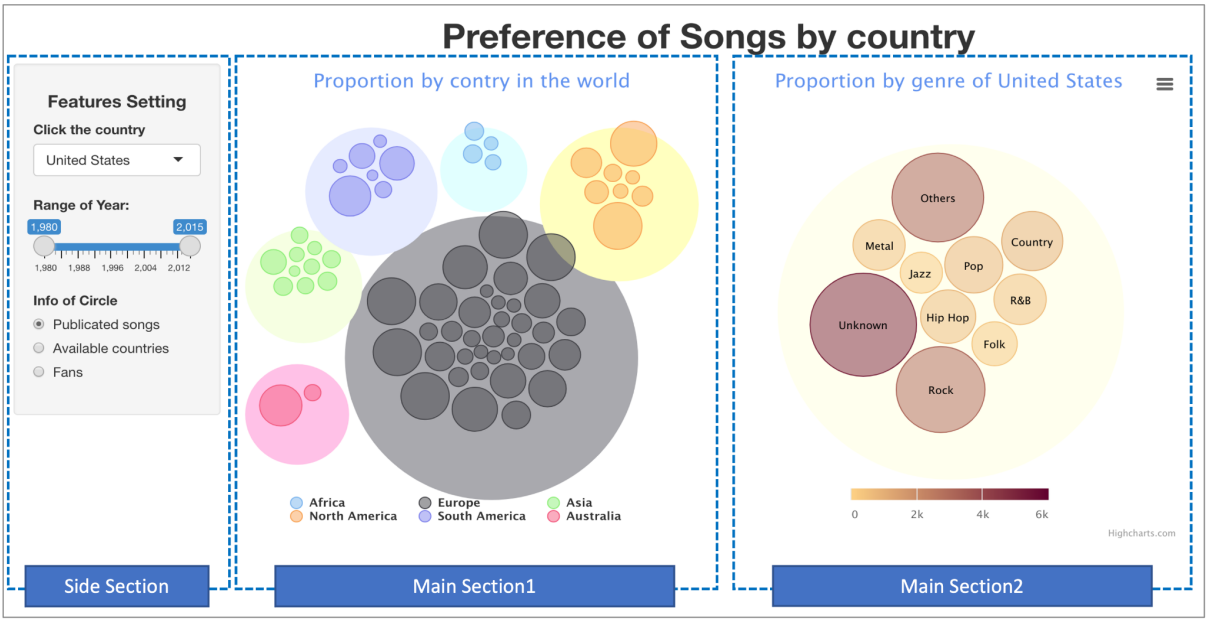

Figure 4.1 : Overview of the Visualization

This project chose Circular Packing visualization techniques to achieve Visualization goals.
Circular packing allows to visualize a hierarchic organization. It is an equivalent of a treemap or a dendrogram, where each node of the tree is represented as a circle and its sub-nodes are represented as circles inside of it.

The application was built using R-shiny and consists of a side section for filtering data and a main section divided into two sections. The two visualizations in main section share same filtering input.

- Visualization Techniques : Circular Packing
- Environment : R
- Utilized libraries :
  - library(shiny) : Implementing R-shiny Applications
  - library(dplyr) : filtering and selecting data to be used for visualization.
  - library(highcharter) : Implementing circle packing in main section 1 (d3 highchart implementation method)
  - library(hpackedbubble) : Implementing circle packing in main section 2
  - library(purrr) : Utilized to convert Hierarchical DataFrame to json form

4.1.2 User tasks

The user can perform the following tasks through the visualization

User Task	Feature	Details
Overview	-	An overview shows the bubbles of number of songs released over all of the years and all continents.
Filter	Country	Show Information by genre in a particular country.(Main2)
	Year	Show only information for a specific period of time.
	Info of Circle	Allow the user to set an attribute represented by the bubble size.
Point	Bubble	Display a tooltip explaining its information.
Click	Legend (Main1)	Information on a particular continent may be excluded or included. (Main1)

Figure 4.2 : A list of User Task

4.1.3 Visualization Mapping

The visual information shown in this visualization is as follows

- Size of circle – number of publicated songs, Average of available countries, Average of Fans
- Color of circle – Classification of Continent(main1), Number of the values by genre(main2)
- Text in tooltip – Information of the selected circle

4.2 Side Section : Features Setting

In the side section, the user may set a feature of visualization. The selection of country is reflected in the visualization of main section 2, helping users to grasp information of the specific country about publicated songs, available countries, and fans by genre.

The selection of range of year supports users who want to grasp information at a specific time by filtering and showing the data at that time (both main section1 and main section2 are reflected).

The selection for “info of circle” reflects the property of determining the size of the circle. Users can select attributes for published songs, available countries, and fans, and the visualizations for main1 and main2 show a circle packing plot representing the attributes.

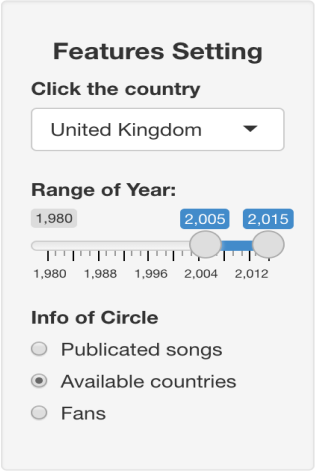

Figure 4.3 : Side panel for feature setting

4.3 Main Section 1 : Circle Packing by Continent-Country

The visualization of Main section 1 shows information on number of songs, average of available countries, and average of fans by continent and country according to feature setting. Through this visualization, the user can easily compare the size of the bubble by attribute. Through this visualization, users can easily grasp the following information.

- Which country's songs were published the most (or least) during a specific period?
- Which continent's songs were published the most (or least) during a specific period?
- Which country's song were the most (or least) available globally during a specific period?
- Which continent's songs were the most (or least) available globally during a specific period?
- Which country's songs have the most (or least) fans worldwide during a specific period?
- Which continent's songs have the most (or least) fans during a specific period?

The user can change the range of year and attribute of circle by adjusting the feature of the side section.(Figure 4.4)

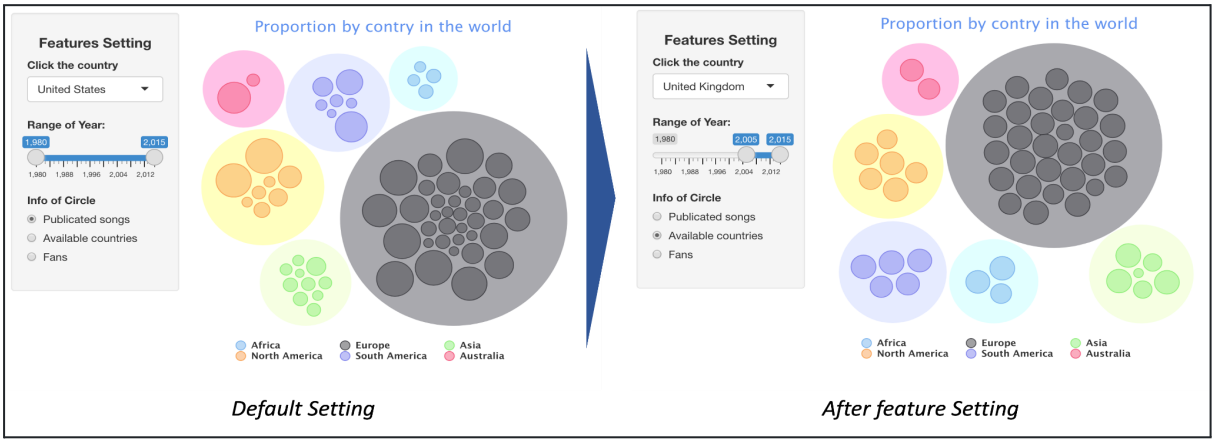

Figure 4.4 : Comparison of plot in main section1 after feature setting

This visualization provides an interaction experience in several ways.(Figure 4.5)

1) If the user puts the cursor on the circle, the information on the circle is displayed as a tooltip box.

2) If the user puts the cursor on a specific continent of the legend, then the circle is emphasized..

3) If the user clicks once on a specific continent of the legend, the circles of the continent are removed in the visualization. And if the users double-click it, circles of the continent are added again.

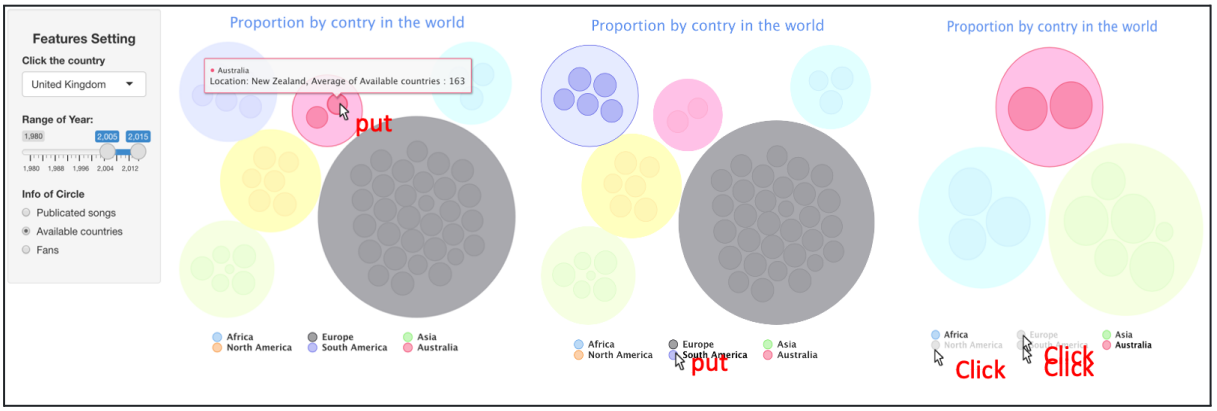

Figure 4.5 : Comparison of plot in main section1 after user interacting

4.4 Main Section 2 : Circle Packing by Country-Genre

The visualization of Main section 2 shows information on the number of songs, average of available countries, and average of fans by genre in a specific country according to feature setting. Through this visualization, users can easily compare the proportion of each country's genre by attribute. The user can easily grasp the following information.

- Which genre of song was the most (or least) released in the United States (example) during a specific period?
- Which genre of song among the songs released in the United States (example) was the most (or least) available worldwide during a specific period?
- Which genre of song among the songs released in the United States (example) had the most (or least) fans worldwide during a specific period?

The user can change the country, the range of year and attribute of the circles by adjusting the feature of the side section.(Figure 4.6)

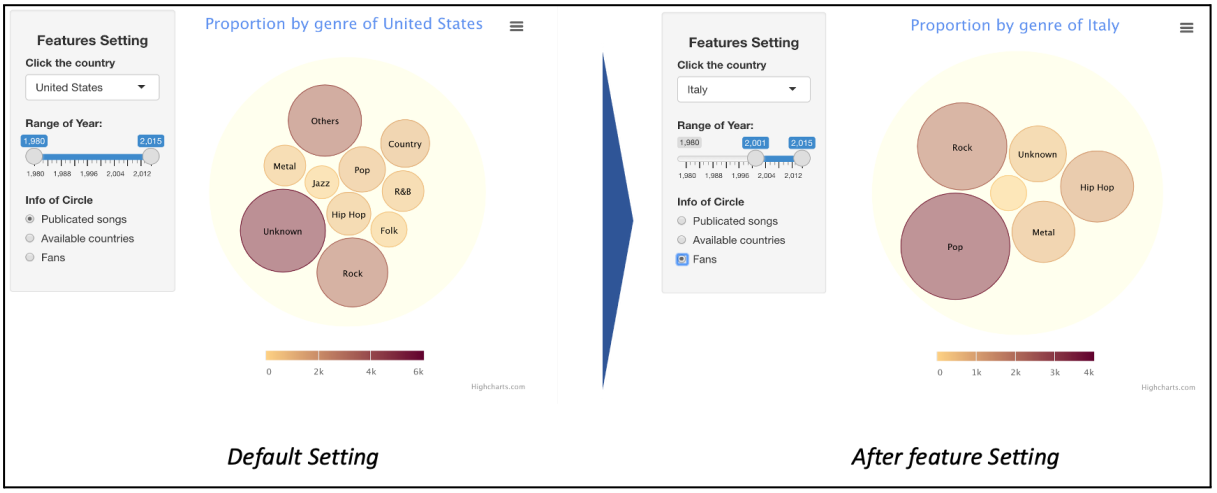

Figure 4.6 : Comparison of plot in main section2 after feature setting

This visualization provides an interaction experience in a way.(Figure 4.7) If the user puts the cursor on the circle, the information on the circle is displayed as a tooltip box.

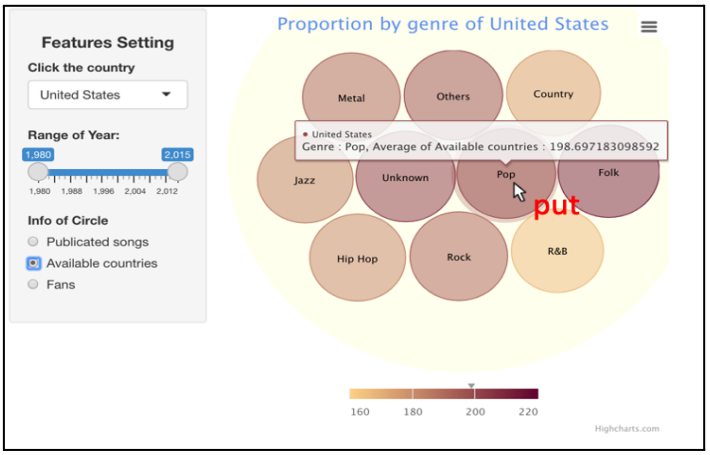

Figure 4.7 : Comparison of plot in main section2 after user interacting