UINVERSITY NICE COTE D'AZUR
MSC DATA SCIENCE AND AI

DATA VISUALIZATION

Final project report

Advisor: Macro Winkler

SOPHIA ANTIPOLIS, DECEMBER 2022

Member list

No.	Name	Surname
1	Anjana	BHAT
2	Huyen Trang	NGUYEN

Contents

1 Introduction

1.1 Idea context

Album is the principal form of musical recording all over the world. Over the time, the music market has been changed significantly, including genres, languages, contents and type of recording. In the previous phases, there were 3 dominant kinds of musical recording, such as long-playing record (LP), the audio-cassette, and the compact disc. Due to the improvement of technology, music nowadays can be released on streaming platform, which creates a strong impact to traditional album format market. In other words, the revolution of album publication has a strong relation to the musical history and also to the technology development.

The main idea of this project is inspired by the album era from the late 20th century to the early 21th century, in order to see the trend of musical consumption in a timeline. The album era could be divided into 6 phases:

- Pre-history (1950s): The beginning of long-playing record (LP), which was known as "record album" in early music industry.

- Beginning of Rock era (1960s): Rock came in to musical industry, and be significantly presented in album records.

- Golden age of LP (1970s): The dramatic rise of musical album all over the world, with various languages and genres.

- Start of CDs and cassettes (1980s-1990s): The reduce of LP albums, in contrast with the strong increase of CDs and cassettes records.

- Pop and urban albums (2000s): The new genres (pop and urban) came to the musical album market and received a lot of positive reception from audiences.

- Streaming era (2010s-present): The technology brings a brand new breath into album industry by appearance of online musical platform.

A intelligent data visualization is performed in this project, to bring an interactive sight-seeing in album era for users.

1.2 Target users

The target users of this project are music-lovers who want to see how albums have been changed over the time. This project also aims to the music producers and artists, who want to analyse the album market, in order to catch up the audience tastes in music, and to improve their future album publication. In order to attract these target users, many detailed information have been showed in this data visualization, in terms of genres, artists, countries and number of deezerFans (a platform where musical fans can follow their favorite artist). A lot of intelligent tools are applied in order to offer an interactive and user-friendly data visualization for the users.

User task	Description
Overview	The album era is presented in one unique data visualization, including various additional information.
Select detailed information	Users can select any detailed information based on their interests, by clicking at any information in the visualization. The detailed information might be countries, artists, languages and number of deezerFans, etc.
Filter information	Filtering information is possible in this data visualization, similarly to the select detailed information task.
Zoom	The world map plot allows users to zoom in and zoom out countries, resulting in filter information about those selected contries also.

Table 1: Intelligent visualization for users

1.3 Data set

WASABI Song Corpus is an enriched data set, presented in terms of CSV files extracted from online music databases. WASABI Song Corpus contains 3 major parts of data about songs, artists and albums. Some useful links for more information about WASABI Song Corpus: WASABI - API, WASABI Song Corpus description, WASABI Interactive Navigator.

For the purpose of this project, the album data set in WASABI is used, containing 208k albums with additional information of publication date, countries, languages, genres, title, length, etc. The most useful data are selected and analysed from this album data set in order to have a clear and informative visualization. The progress of analyzing data will be discussed in the following part of the report.

2 Data Processing

WASABI Albums dataset has been processed to remove missing values. Required variables are extracted and the dataframe is then transformed into the required format using R programming and Power BI.

2.1 Tools used

1. RStudio is used to remove missing values and to extract the necessary variables and to format them, that has been used to visualize the data.
2. Power BI is used to add conditional columns and to visualize the data.

2.2 Workflows

Data processing part is divided into 6 steps. Figure 1 shows the steps involved in data processing.

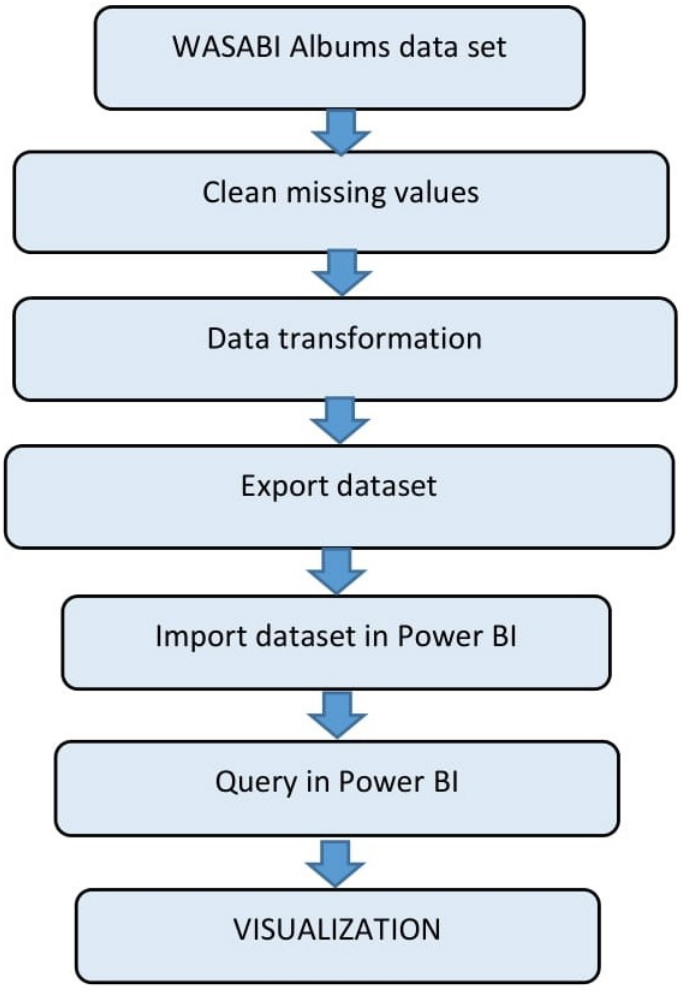

Figure 1: Data Processing Workflow

1. WASABI Albums Dataset:
   WASABI has 3 datasets namely Songs, Artists and Albums. To visualize the evolution of albums' era among different countries, from the year 1950 to 2016, Albums dataset has been choosed. Figure 2 shows the R snippet to read the Albums dataset.

```
library(dplyr)
```

```
df_albums<-readRDS("albums_all_artists_3000.rds")
```

Figure 2: Reading the album dataset

2. Clean missing values:
   First step was to clean the missing values using RStudio. So, first of all, null and empty values were changed in the dataset to NA's. Then the NA's in column publicationDate was removed as it is one of the main feature that has been used in the visualization. Figure 3 shows the R snippet of cleaning missing values.

```
df_albums[df_albums == 'NULL'] <- NA
df_albums[df_albums == ''] <- NA
#removing NA's from column publicationDate
df <- filter(df_albums, !is.na(publicationDate))
```

Figure 3: Cleaning missing values

3. Data transformation:
   Columns are selected from the cleaned dataset and are then transformed. Columns selected from Albums dataset are publicationDate, country, language, genre, name, deezerFans. Then, data type of publicationDate and deezerFans are changed to numeric. In addition, values in column "country" are replaced from country code to country name. Also, values in column "language" are replaced from language code to language name. Figure 4 shows the Rsnippet of data transformation.

```
#selecting the required columns
df <- df %>% select(publicationDate, country, language, genre, name, deezerFans)
#changing datatype of publicationDate and deezerFans to numeric
df$publicationDate<-as.numeric(df$publicationDate)
```

```
df$deezerFans<-as.numeric(df$deezerFans)
#Changing values in column country from country code to country name
df$country[df$country=='DE']<-'Germany'
df$country[df$country=='US']<-'United States'
df$country[df$country=='GB']<-'United Kingdom'
df$country[df$country=='JP']<-'Japan'
df$country[df$country=='CA']<-'Canada'
df$country[df$country=='AU']<-'Australia'
df$country[df$country=='IT']<-'Italy'
df$country[df$country=='FI']<-'Finland'
df$country[df$country=='XW']<-'World Wide'
df$country[df$country=='BR']<-'Brazil'
df$country[df$country=='NO']<-'Norway'
df$country[df$country=='AT']<-'Austria'
df$country[df$country=='ES']<-'Spain'
df$country[df$country=='FR']<-'France'
df$country[df$country=='SE']<-'Sweden'
df$country[df$country=='XE']<-'Europe North'
df$country[df$country=='CL']<-'Chile'
df$country[df$country=='NL']<-'Netherlands'
df$country[df$country=='PT']<-'Portugal'
df$country[df$country=='RU']<-'Russia'
df$country[df$country=='DK']<-'Denmark'
df$country[df$country=='BE']<-'Belgium'
df$country[df$country=='MX']<-'Mexico'
df$country[df$country=='CZ']<-'Czech Republic'
df$country[df$country=='RO']<-'Romania'
df$country[df$country=='PL']<-'Poland'
df$country[df$country=='EE']<-'Estonia'
df$country[df$country=='FO']<-'Faroe Islands'
df$country[df$country=='AR']<-'Argentina'
df$country[df$country=='ZA']<-'South Africa'
df$country[df$country=='IN']<-'India'
df$country[df$country=='VE']<-'Venezuela'
df$country[df$country=='GR']<-'Greece'
df$country[df$country=='UA']<-'Ukraine'
df$country[df$country=='HR']<-'Croatia'
df$country[df$country=='YU']<-'Yugoslavia'
df$country[df$country=='RS']<-'Serbia'
df$country[df$country=='SI']<-'Slovenia'
#Changing values in column language from language code to language name
df$language[df$language=='eng']<-'English'
df$language[df$language=='fra']<-'French'
df$language[df$language=='fin']<-'Finnish'
df$language[df$language=='por']<-'Portuguese'
df$language[df$language=='deu']<-'Deutsh'
df$language[df$language=='mul']<-'Multiple'
df$language[df$language=='jpn']<-'Japanese'
df$language[df$language=='vie']<-'Vietnamese'
df$language[df$language=='nor']<-'Norwegian'
df$language[df$language=='spa']<-'Spanish'
df$language[df$language=='ksh']<-'Colognian'
df$language[df$language=='nld']<-'Nederlands'
df$language[df$language=='swe']<-'Swedish'
df$language[df$language=='rus']<-'Russian'
df$language[df$language=='ita']<-'Italian'
df$language[df$language=='dan']<-'Danish'
df$language[df$language=='ces']<-'Cestina'
df$language[df$language=='pol']<-'Polish'
df$language[df$language=='est']<-'Espanol'
df$language[df$language=='fao']<-'Faroese'
df$language[df$language=='afr']<-'Afrikaans'
df$language[df$language=='mkd']<-'Macedonian'
df$language[df$language=='hin']<-'Hindi'
df$language[df$language=='ron']<-'Romanian'
df$language[df$language=='hrv']<-'Croatian'
df$language[df$language=='srp']<-'Serbian'
df$language[df$language=='slv']<-'Slovenian'
```

Figure 4: Data transformation

4. Export dataset:
   Processed dataset is then exported as Albums_Era.csv. Figure 5 shows the Rsnippet of exporting dataset. Figure 6 shows the top 10 rows of the final dataset after R coding.

```
#Exporting the dataset
write.csv(df,"Albums_era_Fin.csv", row.names = FALSE)
```

Figure 5: Exporting dataset

	publicationDate	country	language	genre	name	deezerFans
1	1995	Germany	English	Trip Hop	Tricky	7706
2	1996	United States	English	Trip Hop	Tricky	646
3	1996	United States	English	Trip Hop	Tricky	2485
4	1996	United States	English	Trip Hop	Tricky	2485
5	1998	United Kingdom	English	Trip Hop	Tricky	1250
6	1999	Japan	English	Trip Hop	Tricky	3261
7	2001	United Kingdom	English	Trip Hop	Tricky	662
8	2001	United States	English	Trip Hop	Tricky	4609
9	2002	United States	English	Trip Hop	Tricky	5019
10	2003	United States	English	Trip Hop	Tricky	2150

Figure 6: Dataset after R coding

5. Import dataset in Power BI:
   For visualizing the data, Power BI is used. So, Albums_Era.csv is imported in Power BI.

6. Query in Power BI:
   Further modifications are made in Power BI to make the visualization more clear and understandable. First of all, columns with NAs' are replaced as null. Then, a conditional column is created in order to split the genre into different Eras. So the final dataset contains 7 columns namely, publicationDate, country, language, genre, name, deezerFans, YearEra. Figure 7 shows the queries applied to the dataset in Power BI.

```
= Csv.Document(File.Contents("C:\Users\NAGESH\Desktop\My Notes M1\Data Visualization\Albums_era_Fin.csv"),[Delimiter=",", Columns=6,
    Encoding=65001, QuoteStyle=QuoteStyle.None])
```

```
= Table.PromoteHeaders(Source, [PromoteAllScalars=true])
```

```
= Table.ReplaceValue(#"Removed Blank Rows","",null,Replacer.ReplaceValue,{"country"})
```

```
= Table.ReplaceValue(#"Replaced Value","",null,Replacer.ReplaceValue,{"language"})
```

```
= Table.ReplaceValue(#"Replaced Value1","",null,Replacer.ReplaceValue,{"genre"})
```

```
= Table.ReplaceValue(#"Replaced Value2","",null,Replacer.ReplaceValue,{"deezerFans"})
```

```
= Table.ReplaceValue(#"Replaced Value3","",null,Replacer.ReplaceValue,{"name"})
```

```
= Table.AddColumn(#"Replaced Value4", "Year Era", each if [publicationDate] >= 2010 then "Streaming Era" else if [publicationDate] >= 2000
    then "Pop and Urban Era" else if [publicationDate] >= 1980 then "Casette and CDs Era" else if [publicationDate] >= 1970 then "Album's
    Golden Era" else if [publicationDate] >= 1960 then "Rock Era" else "Pre-history")
```

Figure 7: Query in Power BI

7. Visualization:
   Final part but not the least is to visualize the data in Power BI. This project uses PowerBI, which is an interactive data visualization software product developed by Microsoft. The figure 8 shows a short-view about how to use PowerBI to plot data. On the right sight, in the "Visualizations" part, all the available diverse choices of plot are offered, such as bar chart, donut chart, line chart, etc. In the bottom of "Visualization" part, there are choices about x-axis, y-axis, Legend and Values that can be chose to plot. This could be more explained in the following part of mapping. Besides, the "Fields" part allows users to insert and choosing the data that they want to plot. The Figure 8 performs the way how the count of deezerFans for albums are created, by choosing a stacked column bar plot, and selecting date,

deezerFans and era data.

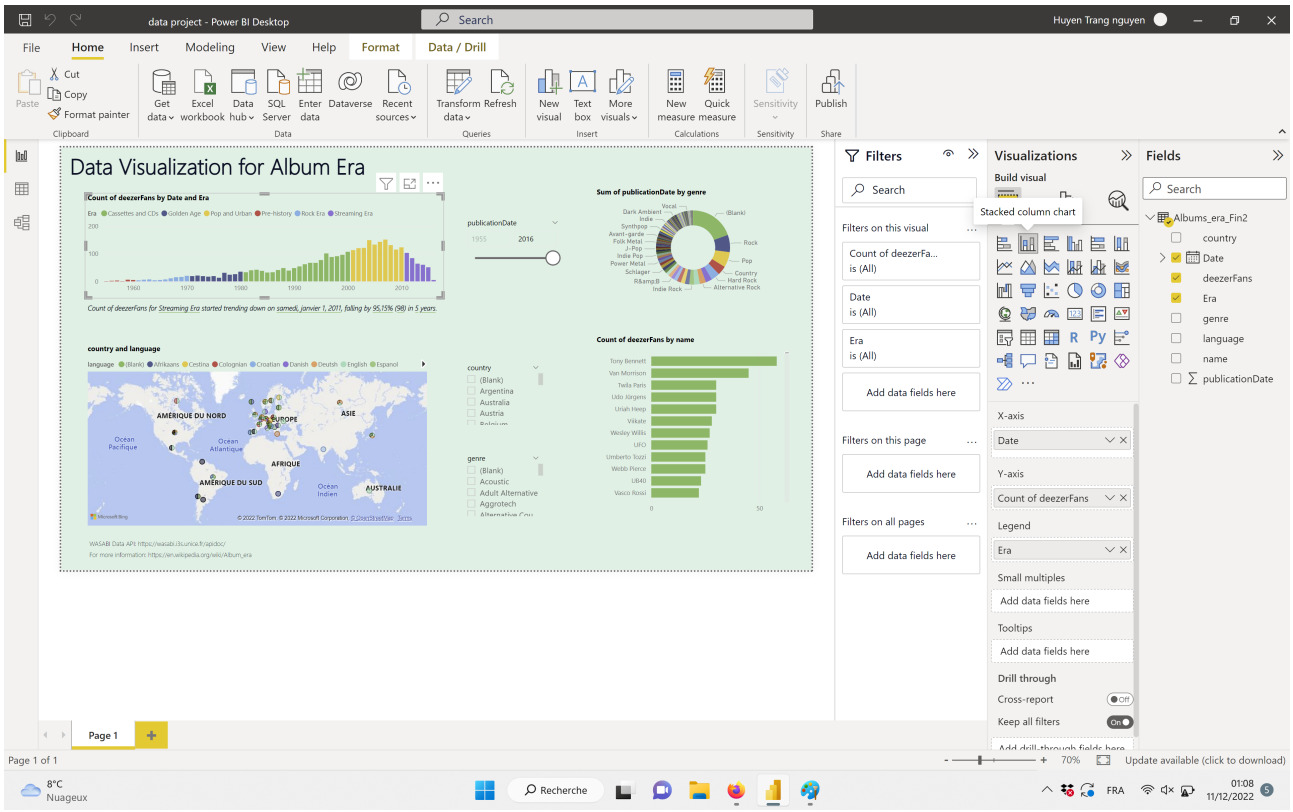

Figure 8: Example for visualization method

Detailed explanation of visualization part is mentioned in section 3 and section 4. Figure 9 shows the top 10 rows of final dataset that has been used in visualizing the data.

	publicationDate	country	language	genre	name	deezerFans	Era
1	1955	null	null	Country	Webb Pierce	3	Pre-history
2	1956	null	null	Pop	Tonina Torrielli	16	Pre-history
3	1956	null	null	Country	Webb Pierce	2	Pre-history
4	1957	null	null	Country	Webb Pierce	16	Pre-history
5	1957	United States	English	Jazz	Tony Bennett	42	Pre-history
6	1957	null	null	Jazz	Tony Bennett	1	Pre-history
7	1958	null	null	Jazz	Tony Bennett	4	Pre-history
8	1959	null	null	Jazz	Tony Bennett	5	Pre-history
9	1959	null	null	Country	Webb Pierce	1	Pre-history
10	1959	United States	English	Jazz	Tony Bennett	107	Pre-history

Figure 9: Final dataset for visualization

3 Data Visual Mapping

Data visual mapping is all about how the dataset is mapped to the plots and various features of the plots such as filters applied, tooltips and plot interactivity.
Moreover, All the choosen charts are interactive. So, it is easy to visualize the data for one particular genre by clicking on any one of the charts and other charts will highlight only the required visualization part.Further, Subsection 3.1 and 3.2 explains about the selected charts, tool-tips and filters applied for the individual group members.

3.1 Data visual mapping (Anjana)

3 plots for visualization namely, clustered bar chart, line chart and donut chart are considered. Brief data visual mapping is shown in Figure 10.

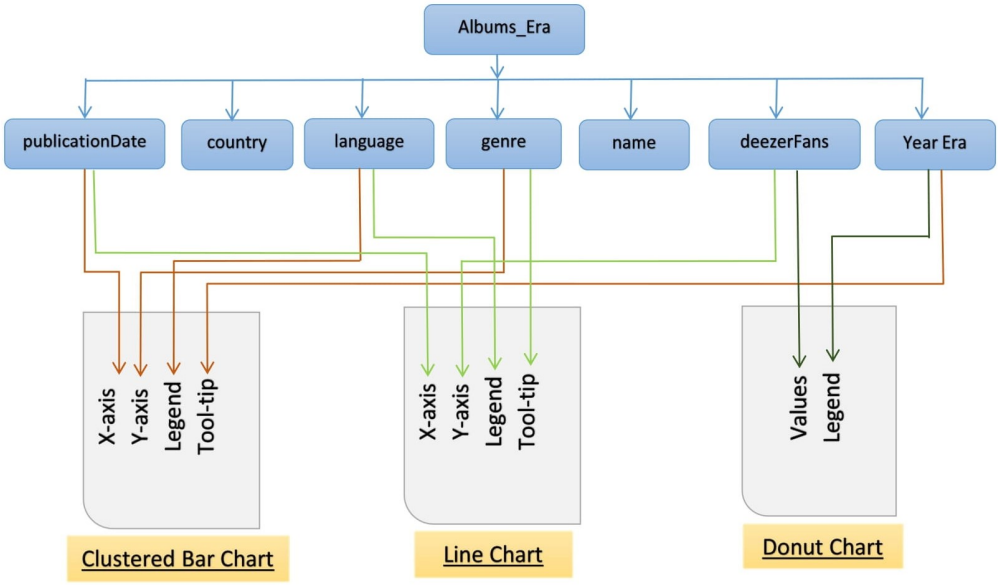

Figure 10: Data Visual Mapping (Anjana)

- Clustered Bar chart shows the evolution of genre based on publicationDate and, language is choosen as legend. Here, tool-tips gives the information about genre, language, start year, end year, and the era that the genre belongs to.
- Line chart shows the growth of deezerFans with respect to publicationDate. Also, here language is choosen as legend. In addition, tool-tips gives the information about number of deezerfans, language and genre.
- Donut chart shows the growth of deezerFans per Era. Also, tool-tips gives the information about number of deezerFans and Era.

Filters are provided to choose country, language and year. Here, preference is given to select only one country at a time as the idea to visualize the evolution of albums era per country. In regards to the choosen country, one can select multiple languages for visualization. Also, Year can be selected with the help of slider bar. Tooltips are provided so that it will be helpful for the user to look into any particular information on the plot.

3.2 Data visual mapping (Huyen Trang)

For the overall data visualization, there are 4 main plots that are performed:

- A stacked column bar plot, where the number of deezerFans are counted by each year and each era. (Count of deezerFans data is y-axis, and x-axis is for year data. The colors of bar column (Legend) depends on era data).
- A world map, where reveals the language used in musical album all over the world. Countries' names are used as Location and languages is used as Legend to plot this world map.
- A donut plot shows all information about music genre in all the released album. The publicationDate is Values and genre is Legend in this donut visualization.
- A clustered bar chart presents the name of artists and the sum of their deezerFans. The y-axis is the name of artist, in ascending ranking or deezerFans. The x-axis performs the total number of deezerFans. Data of artist name and deezerFans are used to create this bar chart plot.

There are 3 filters (slicers) for publication date, country and genre, which allow users to filter all the information as interests of user. Brief data visual mapping is shown in Figure 11.

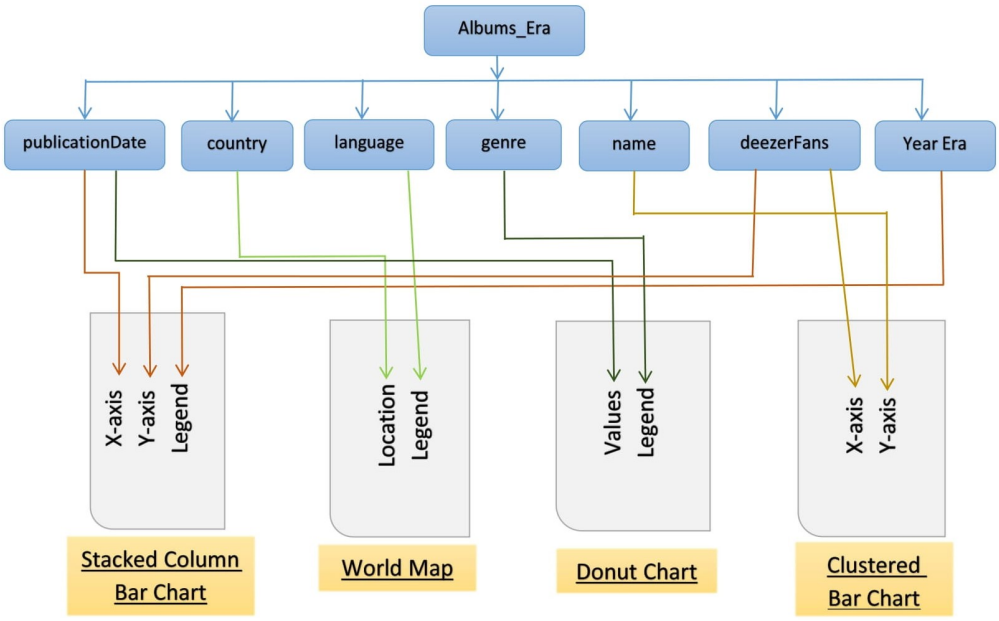

Figure 11: Data Visual Mapping (Huyen Trang)

4 Data Visualization

4.1 Data visualization by NGUYEN Huyen Trang

4.1.1 Overview

The Figure 12 below shows an overview of album era all over the world. The visualization contains 4 major information plots:

- Top-left of the visualization: Count of deezerFans for albums by era using a stacked column bar plot. This plot brings a clear sight-seeing how album has been changing over the 6 different phases: Pre-history, Beginning of Rock era, Golden age of LP, Start of CDs and cassettes, Pop and Urban, Streaming era.

- Bottom-left of the visualization: Languages of albums released in each country using a world map plot. A global look of diverse languages in musical album publish is shown, thanks to this world map.

- Top-right of the visualization: Sum of albums for each musical genre using a donut chart. The relation between album market and genre of music are strongly connected.

- Bottom-right of the visualization: Number of albums released for each artist's name using a clustered bar chart. The name of artists are arranged by the total number of their deezerFans. For overall, Tony Bennet is the most-favorite artist in deezerFans platform.

The other filter fools have been used in this plot, including: 3 slicers for publication date, country and genre, which allow users to filter all the information as their interests.

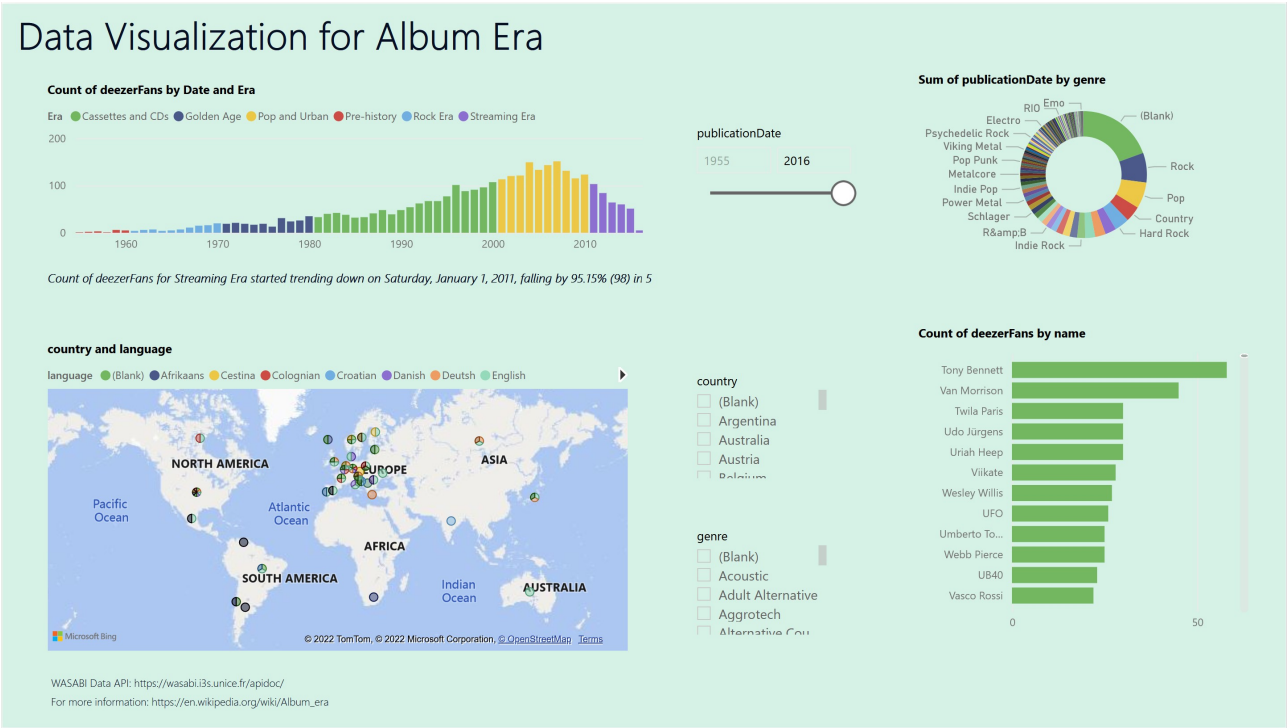

Figure 12: Overview of data visualization

4.1.2 An intelligent visualization

This data visualization adapts to all three requirements of interactive, filtering and toolTips.

Interactive visualization

If the users choose name of a country and a genre of music, all the plots and information will change to specify that country and genre. The interactive factor can be apply for any plot and information in this project. The figure 13 gives an example of visualization for country music in United States.

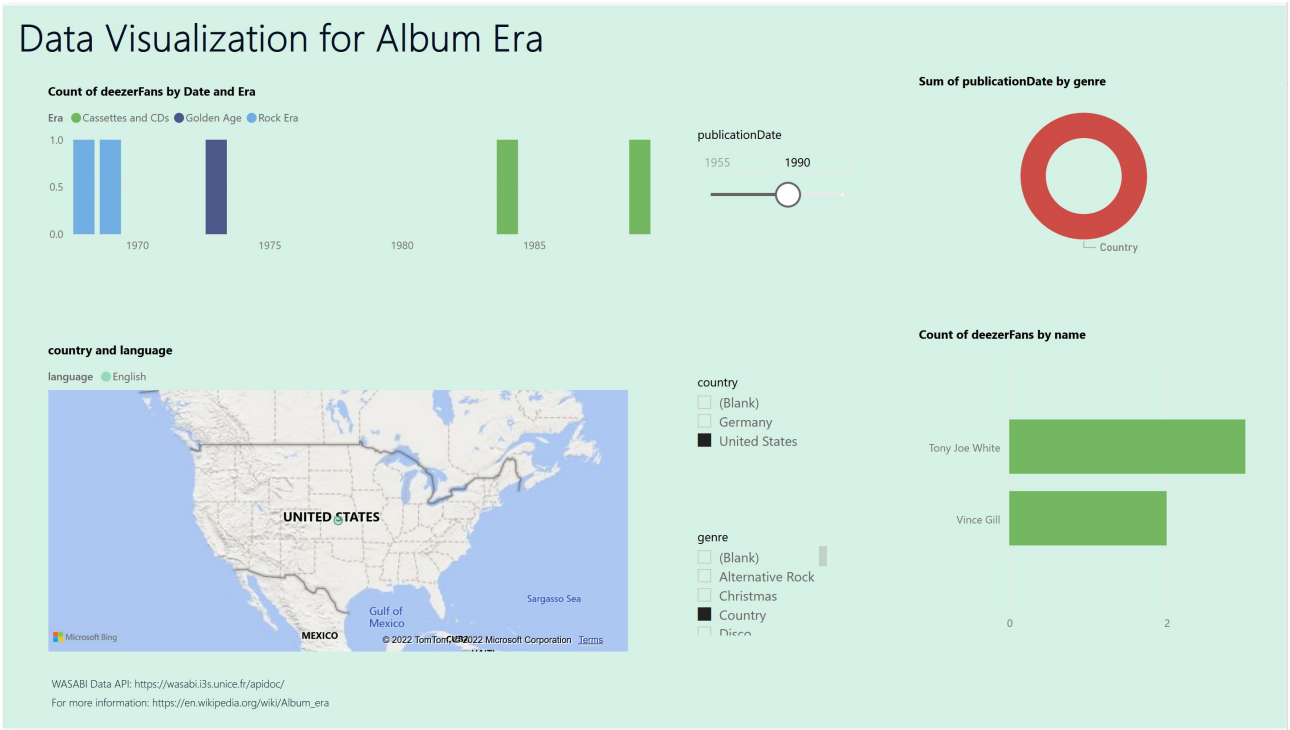

Figure 13: Interactive visualization example for US and country music

Filtering capacity

The data visualization can be filtered as the aim of users. The figure 14 gives an example of filtering capacity in the project. There are 4 example are revealed: filter for Golden Age era, filter for Rock genre, filter for artist Tony Bennett and filter for United States and country music.

ToolTips

The ToolTips capacity can be presented when users select any plot in the data visualization. The figure 15 gives an example for Australia, when the user choose position, the name of country will be appear, with information about language of album published in that country.

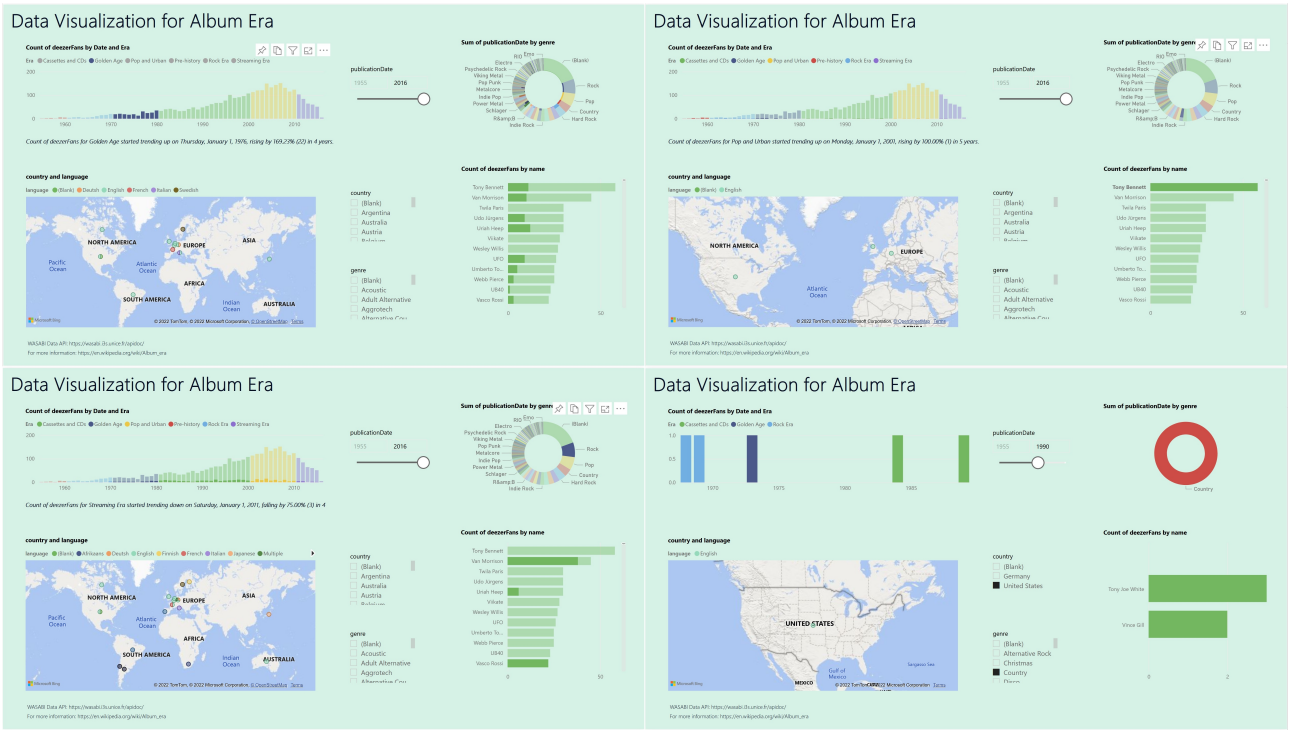

Figure 14: Filtering capacity example

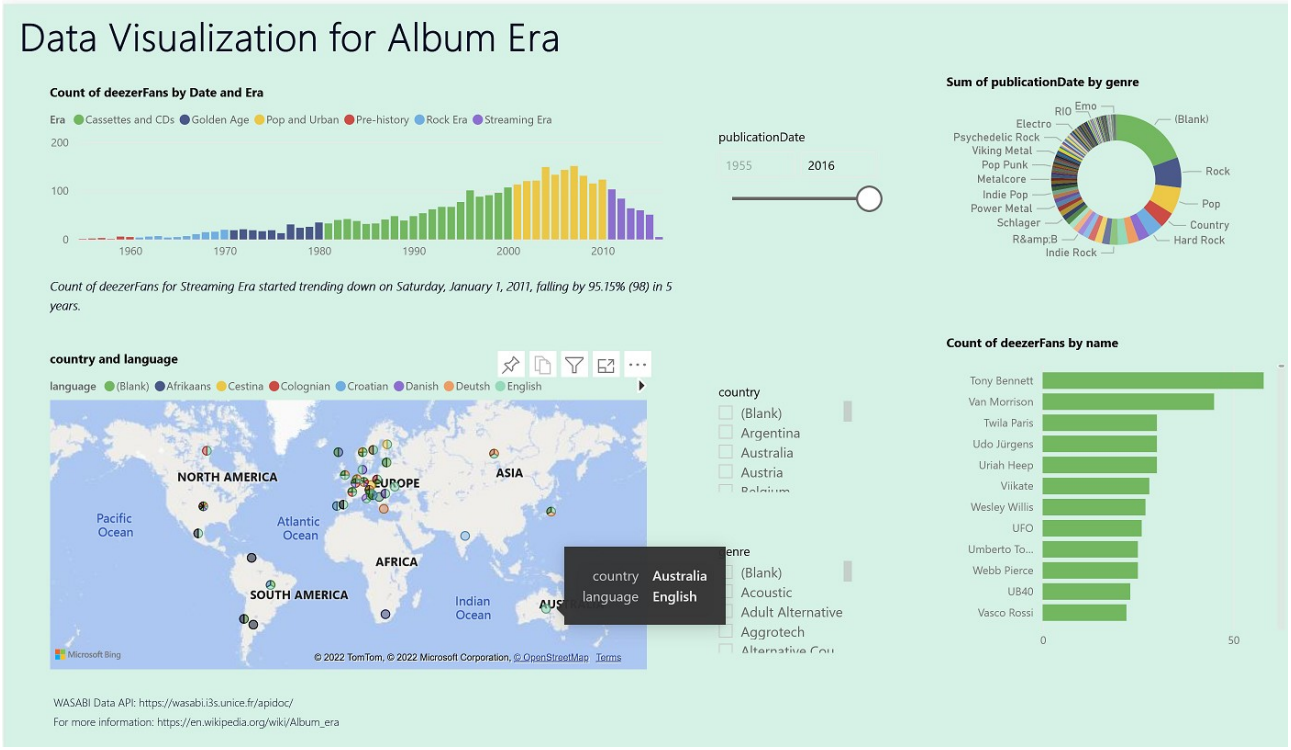

Figure 15: ToolTips example for Australia

4.2 Data visualization by Anjana BHAT

The Figure 16 shows an overview of evolution of album era per selected country by the user. The visualization has 3 information plots:

- Genre vs Year per country: Clustered bar chart shows the evolution of genre for a selected country. Also, language is given as legend so, plot defines the evolution for each language that is selected.
- DeezerFans per year: Line chart shows the growth of deezerFans for a selected country. In addition, a line for growth of deezerfans is presented for each language choosen.
- DeezerFans per Era: Donut chart is choosen to know the percentage of deezerFans per Era. So there are 6 eras as mentioned in Section 1.1

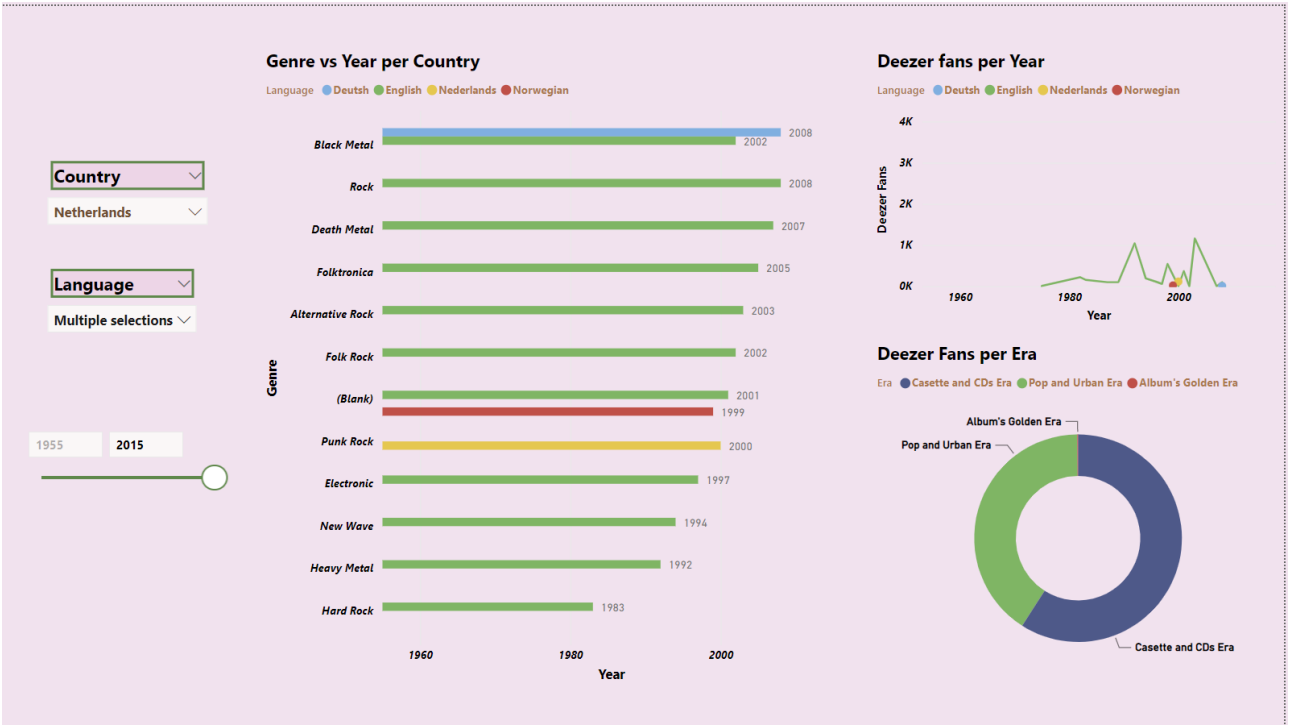

Figure 16: Evolution of Album Era

Figure 17 shows the evolution of album era for country France that has albums in languages french and english and the year is filtered from 1955 to 2016. Figure 18 shows various examples of visualization by filtering country, language and year and also the intereactivity among the plots. An example for tooltip can be seen in figure 19

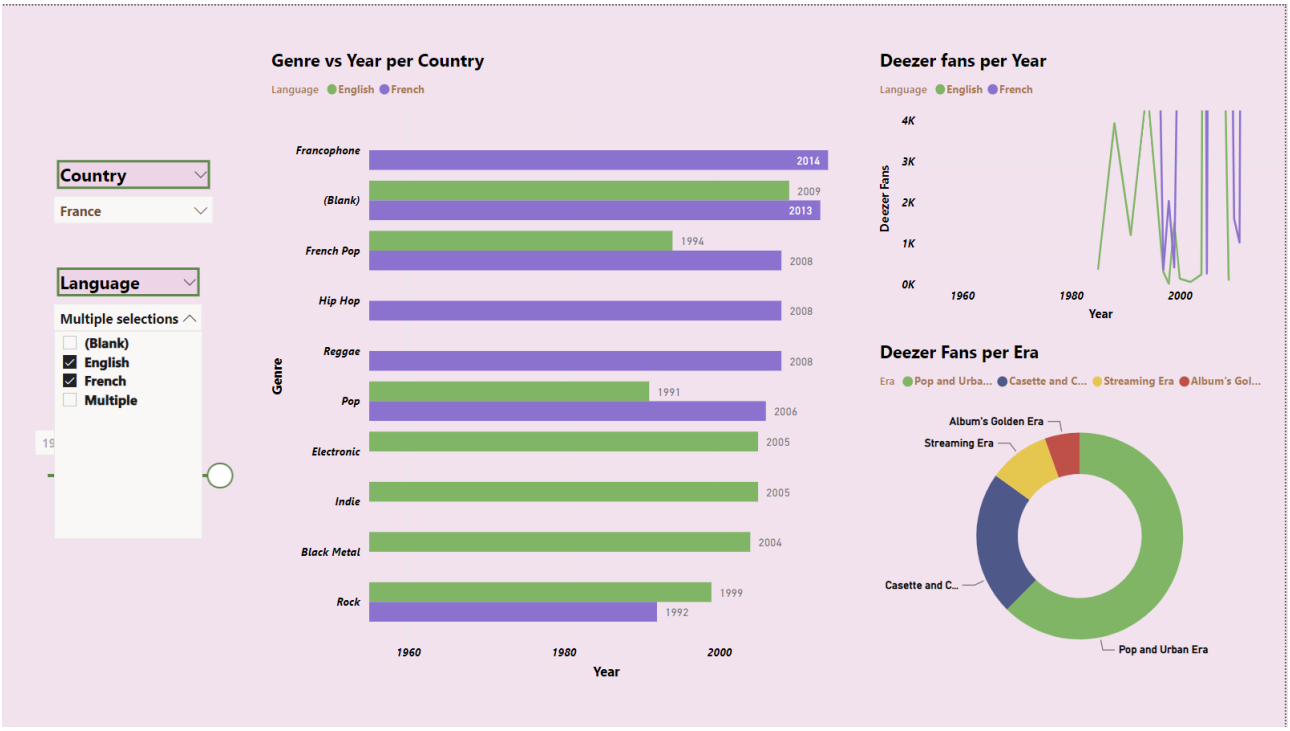

Figure 17: Evolution of Album Era for France

Figure 18: Examples for filtering and interactivity

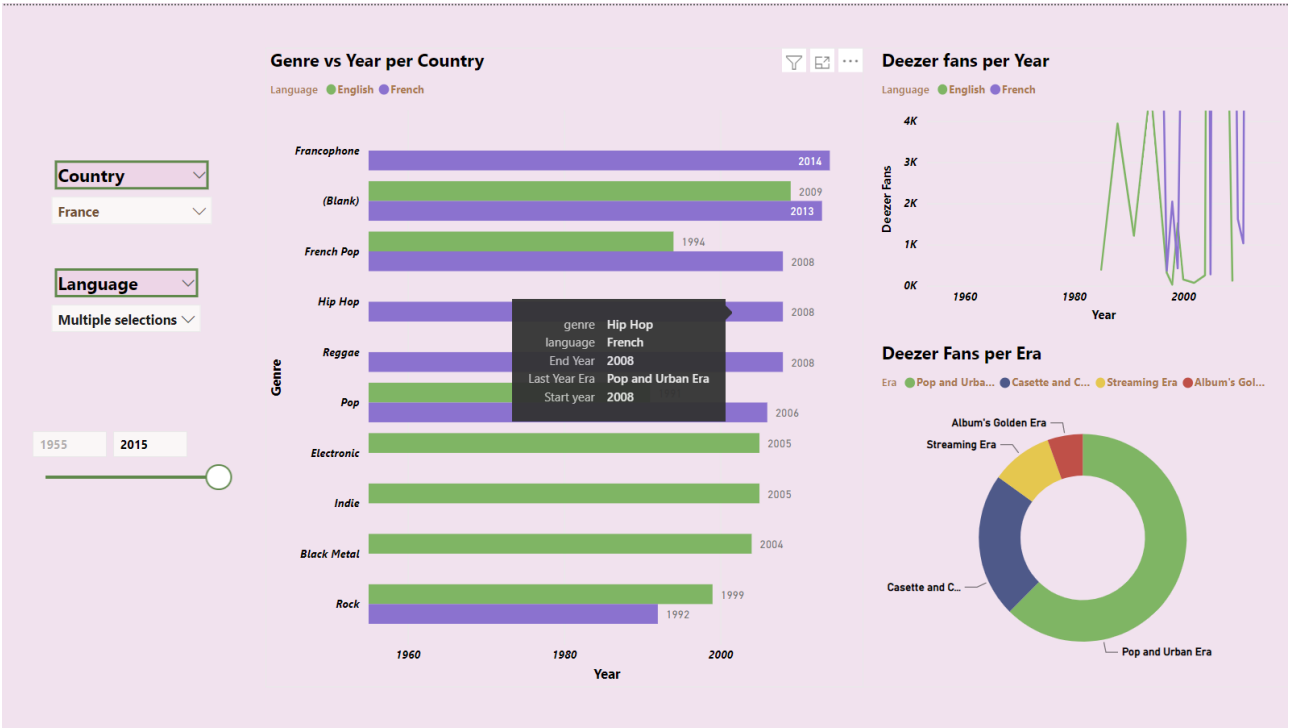

Figure 19: Example for tooltip

5 Conclusion and future propositions

This project brings an intelligent data visualization to users, who can be music-lovers, music producers or artists, to help them have a global detailed sight-seeing about album era from the lately 20th century to the early of 21th century. PowerBI is the data visualization tool that is used in this project, which totally presents capacity of being interactive, filtering and ToolTips. In the future, an animation of data visualization is expected for improving this project.