



UNIVERSITY OF COTE D'AZUR

MSC DATA SCIENCE AND AI

Data Visualization Project

Author:

Sourav Rai

Ayoub Youssofi

Supervisor:

Prof. Marco Winckler

January 20, 2022

Abstract

Research on data visualization has been a major development in a number of different fields. This development includes investigating ways in applying visualization techniques for more efficient use, interpretation, and presentation of the data. A graphical presentation is generated from the data content and view by a user. The primary aim of the work summarized in this paper is to create visualization for WASABI dataset. WASABI data describes more than 2 million commercial songs. It can be exploited by music search engines, music professionals (e.g., journalist, artists, music teachers, streaming music companies...etc.) or scientist willing analyze most popular music. This paper shows that each of the proposed applications significantly contributes on giving insight about dataset by identifying patterns and relationships.

Finally, it should also be considered as a development of low-cost data visualization tool where the license is not required.

Keywords: Data Visualization, information visualization, Data processing, Big data

Chapter 1

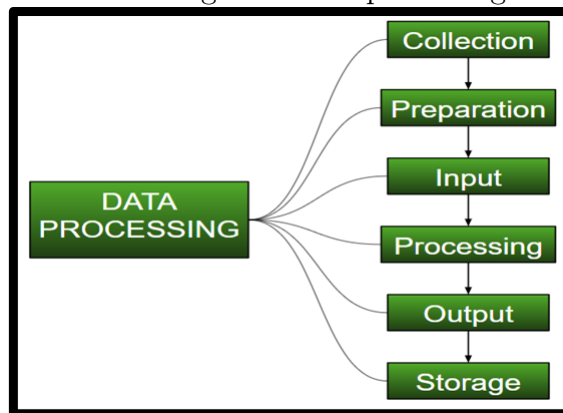
Introduction

Advances in science and technology of computing have engendered unprecedented improvement in science and engineering research. The use of visual representation, such as diagrams and models, has been part of this improvement. Their use makes the possibility to interact with and represent complex observations into visual representations. Data visualization is concerned with different stages from the data processing, the design, to the development of graphical representation. It provides an effective presentation of unorganized dataset. Through the development of our visualizations, we went into different stages that are going to be described in the following sections.

1.1 DATA PREPROCESSING(Ayoub Youssofi)

To implement the visualizations, different stages are required beforehand. The most important step is Data processing. In this stage, the dataset is prepared in order to be used into the visualization.

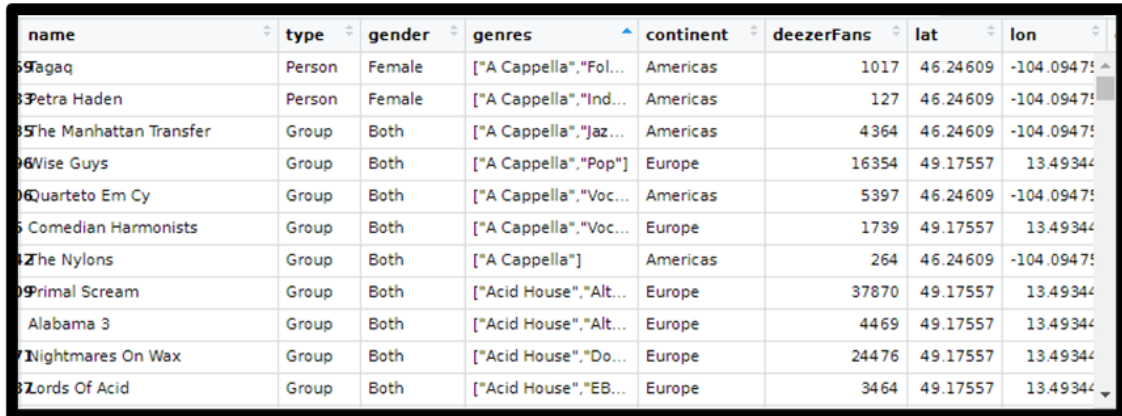
Figure 1.1: Different stages of Data processing of the project



After the collection of the dataset file, the features needed for the visualisation are selected. The new dataframe contains more missing values data. To deal with this problem the missing values has been replaced with the mean value of each column. The final output is a dataframe with no missing values. Besides, I had to add

to columns “Latitude” and “Longitude”. These columns are needed for the sack of implementing the bubble in the right continent.

Figure 1.2: Screenshot of the dataframe deezerFans of the visualization



name	type	gender	genres	continent	deezerFans	lat	lon
59Agagq	Person	Female	["A Cappella","Fol...	Americas	1017	46.24609	-104.09475
83Petra Haden	Person	Female	["A Cappella","Ind...	Americas	127	46.24609	-104.09475
85The Manhattan Transfer	Group	Both	["A Cappella","Jaz...	Americas	4364	46.24609	-104.09475
96Vise Guys	Group	Both	["A Cappella","Pop"]	Europe	16354	49.17557	13.49344
96Quarteto Em Cy	Group	Both	["A Cappella","Voc...	Americas	5397	46.24609	-104.09475
6 Comedian Harmonists	Group	Both	["A Cappella","Voc...	Europe	1739	49.17557	13.49344
92The Nylons	Group	Both	["A Cappella"]	Americas	264	46.24609	-104.09475
99Primal Scream	Group	Both	["Acid House","Alt...	Europe	37870	49.17557	13.49344
Alabama 3	Group	Both	["Acid House","Alt...	Europe	4469	49.17557	13.49344
7Nightmares On Wax	Group	Both	["Acid House","Do...	Europe	24476	49.17557	13.49344
8Lords Of Acid	Group	Both	["Acid House","EB...	Europe	3464	49.17557	13.49344

1.2 DATA PREPROCESSING(Sourav Rai)

This data preprocessing is done on python. After loading the albums dataset, the rows are grouped by genres and country and their frequency is calculated. The frequency of all the genres by country is used to generate a new dataframe. Next, I update the new dataframe with their full names. Then, the longitude and latitude of the countries is added to the dataframe corresponding to the countries(using a seperate csv file). This column is needed to set the locations of the circles corresponding to the countries. The genres are grouped according to their popularity. More famous genres are given a higher letter(A) and less famous genres are given a lower letter(H). This is done to change the colours of the genres in the bubble map.

Figure 1.3: Screenshot of the dataframe used in the visualization

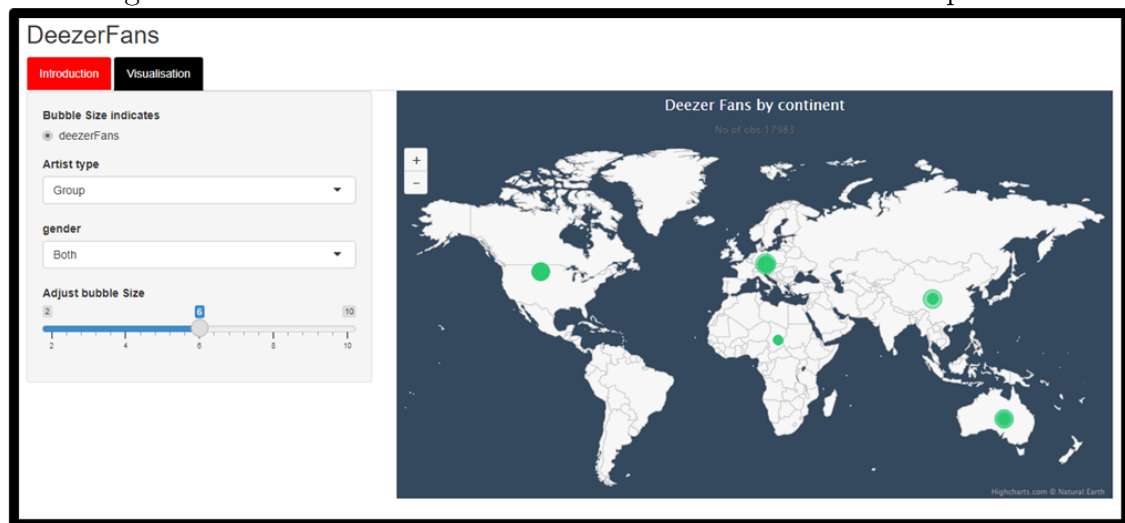
	country	genre	Freq	Full_name	longitude	latitude	group
0	GB	Rock	607	United Kingdom	-3.435973	55.378051	A
1	NL	Pop	54	Netherlands	5.291266	52.132633	B
2	US	Rock	1914	United States	-95.712891	37.090240	A
3	DE	Pop	309	Germany	10.451526	51.165691	B
4	SE	Pop	112	Sweden	18.643501	60.128161	B
5	PT	Fado	12	Portugal	-8.224454	39.399872	H
6	BR	MPB	224	Brazil	-51.925280	-14.235004	H
7	CA	Rock	113	Canada	-106.346771	56.130366	A
8	GR	Alternative Rock	7	Greece	21.824312	39.074208	C
9	DK	Rock	27	Denmark	9.501785	56.263920	A

Chapter 2

Data Visualization(Ayoub Youssofi)

We choose to represent “location” of the distribution of our dataset into a bubble map. This representation helps to represent the numeric values on a territory. It displays one bubble per geographic coordinate. The coordinate could be assigned to a country or a region or a continent. In my case, the visualization displays the Deezer fans by continent.

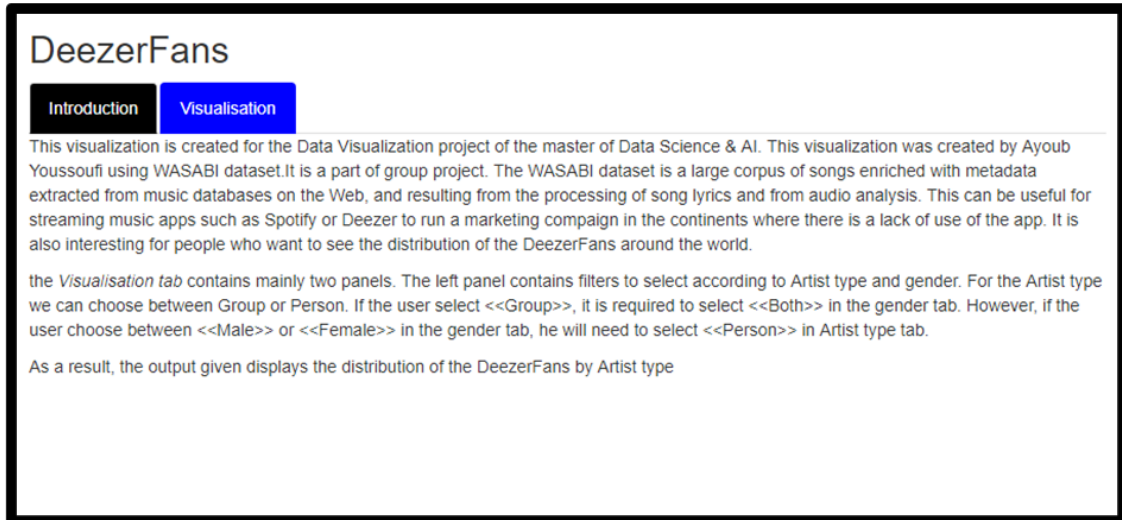
Figure 2.1: The vizualisation of the Deezer Fans in bubble map in R



In the figure above, the user needs to select the filters displayed in the left panel in order to compute the number of deezer fans according to “Artist type” and “Gender”

- If the user wants to display the number of deezer fans for only artists of type “Group”, the user will need to select the input “Group” from the slider. Moreover, it will also be required to select “Both” in gender as the matter of fact in group of artist we might find both female and male.
- If the user wants to display the number of deezerfans only for an individual artist, he will need to select in the second filter the options : male / Female. As a result, the user can display the number of deezer fans for sepecific gender of artists (Male/female).

Figure 2.2: The vizualisation of the tab panel “introduction” made in R



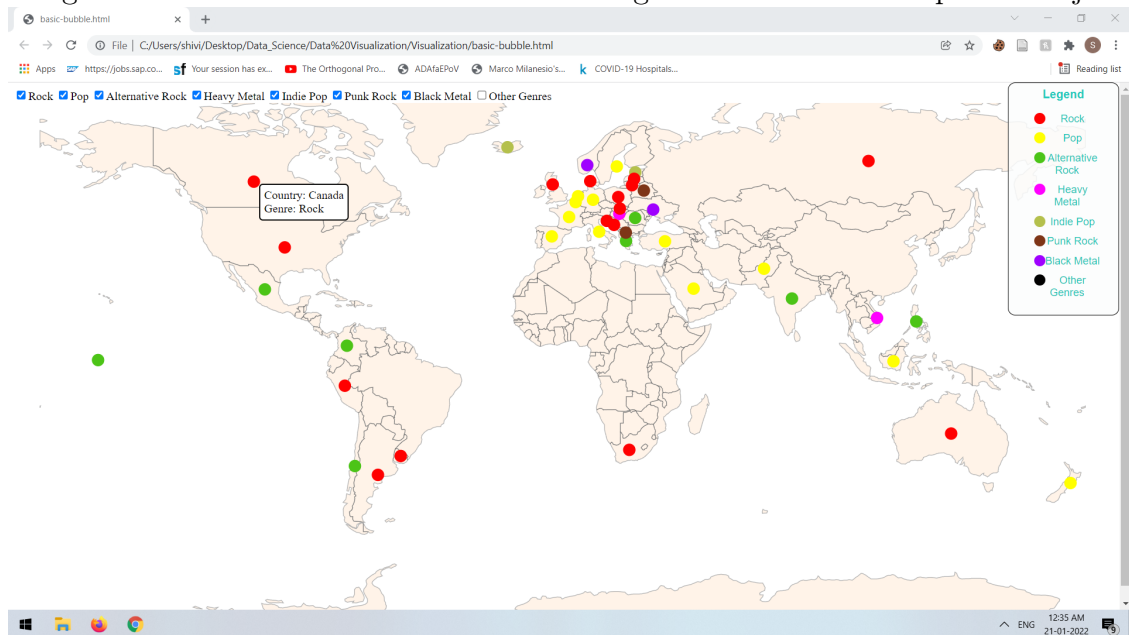
In the visualization, tab panel “Introduction” has been implemented. The purpose of this panel is to explain to the user how to navigate through the dashboard. It helps to explain the required fields to fill in to display the desired output.

Chapter 3

Data Visualization(Sourav Rai)

In this visualization I display the most popular genres in a country. This visualization is made with D3js. The visualization contains all the countries that are represented in the WASABI dataset. The final dataframe from the preprocessing is organized in a certain way and loaded in the html file. Then with the help of javascript, the bubble map is coded and all the details are integrated into the visualization.

Figure 3.1: The visualisation of the famous genres in a bubble map with D3js



The user can experience various aspects in this visualization. Those are described below.

- The user can hover over the circles to know the name of the genre which is famous in the country. Also the user can see the full name of the country.
- The user can select any one genre or any group of genres from the selection panel. The colour coding in the visualization is based on the popularity of the

genres. More famous genres are of on the left hand side of the panel and less famous genres are on the right hand side of the panel.

- Also the user can see from the legends what colour is represented by a genre. This makes it easier to select a genre of choice.
- Also the user can zoom in the visualization to observe more clearly in the places where the circles are not clearly differentiated.

Chapter 4

Conclusion

During the data visualization lectures held at Université Côte d’Azur, the concepts used in the visualization are covered. These concepts are put into practice in project “Data visualization” where each student part of the master’s degree of Data Science & AI has to provide a visual of wasabi dataset. This class provided the student the ability to implement a wide range of visualization techniques. The main focus is an interactive visualization that allows the user to interact with the visualization in order to filter the data being displayed or to change display parameters.