Université Côte d'Azur

Professor: Marco Winckler

Project: Visualizing production of album's

Author's: Mohammed Danish Mustafa
Madzella Yannick
Md Abdul Mazed Siddiki

December 12, 2022

Contents:

1.Introduction

Data visualization techniques are used to represent the large set of data in the simple format for the better understating and summarize the total data into a simple picture or in the simple format.

There are so many data visualization techniques are available to represent different types of data in the different formats and in different representations with so many colors and graphical mode.

In this project I choose World Map, Pie Chart, & Gantt to represent data because World Map, Pie Chart & Gantt have the capability to show all data by categorizing and segregating all data based on the values. So, we can be able to understand the total strength based on the area the category covered.

2.Data Set

The WASABI dataset is from the music data base has all song lyrics and from audio analysis. The dataset contains 1.73M songs with lyrics at different levels and huge number of artists and their bands, origins, type of the band, group information etc.

There are so many features to work and explore in this data set. Wasabi data set we can avail in the three different segmented formats namely as follows.

Songs data set – 2.1 m samples.

Album data set – 208 k samples.

Artist data set. – 77 k samples.

3.Objective

The goal of this project to visualize the:

- Acceptance of Rock Album over a period of time in entire World (“World Map’’).

- Percentage of people liking different Genre of Album (“Pie Chart’’).

- Number of Different Genre of Album Composed in a single year by different countries(‘’Gantt’’).

4.Attribute

To fulfil the above functionality the following attribute are required form WASABI data set:

- Country

-Genre

-Id Artist

-Language

-Name

-Title

-Deezer Fans

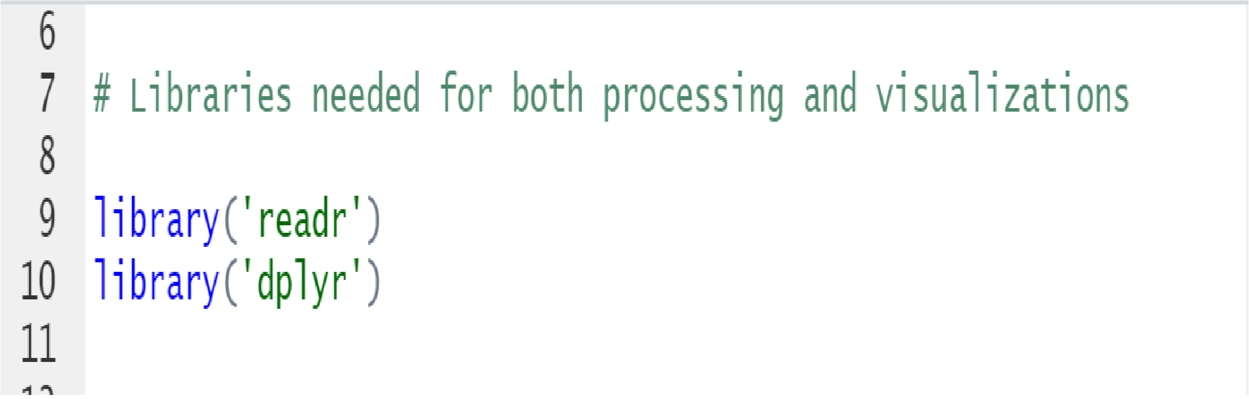
These attributes will be grouped to find the total counts and then will be passed to the visualization to represent the diagram.

5.Data Pre-Processing (worked all together)

By using the R programming language, filtered the required attributes from WASABI data set in the following format. Formatted all values into required and acceptable format. Then converted the data in to file and named as “iadata.csv”.

5.1 Library:

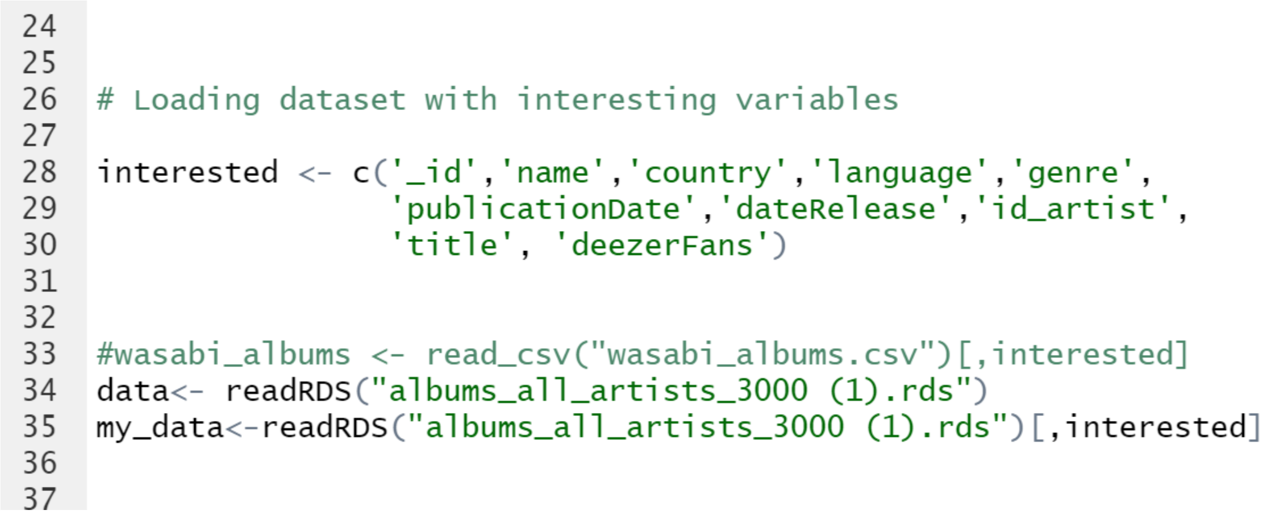
We have imported different libraries for execution of our wasabi data set such as read, dplyr.



```

# Libraries needed for both processing and visualizations

library('readr')
library('dplyr')

```

5.2 Importing Data:

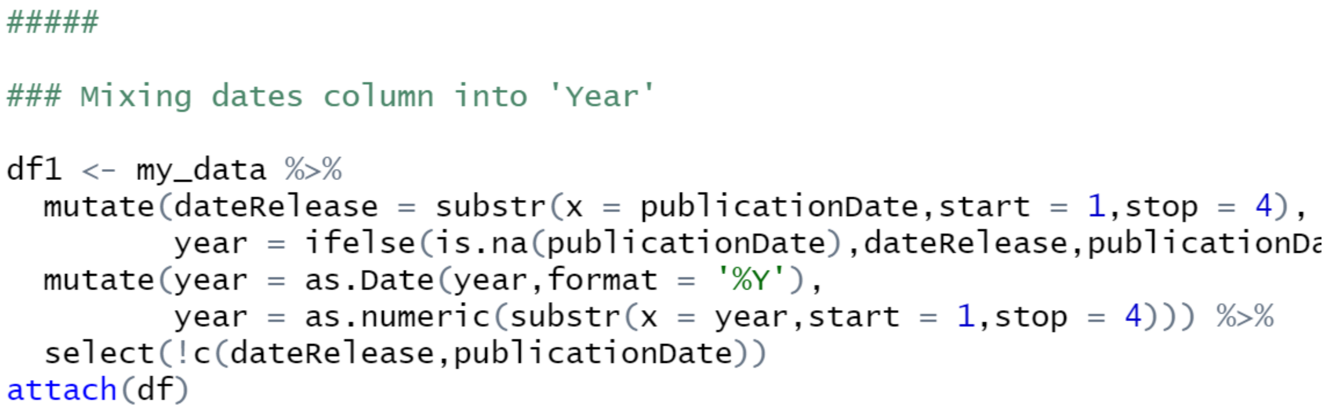
Importing data from the wasabi.rds data set and considering only required variables such as id, name, country, language, genre, publication, daterelease, title etc. for execution of our project.



```


# Loading dataset with interesting variables

interested <- c('_id','name','country','language','genre',
                'publicationDate','dateRelease','id_artist',
                'title', 'deezerFans')


#wasabi_albums <- read_csv("wasabi_albums.csv")[,interested]
data<- readRDS("albums_all_artists_3000 (1).rds")
my_data<-readRDS("albums_all_artists_3000 (1).rds")[,interested]


```

5.3 Missing Date's:

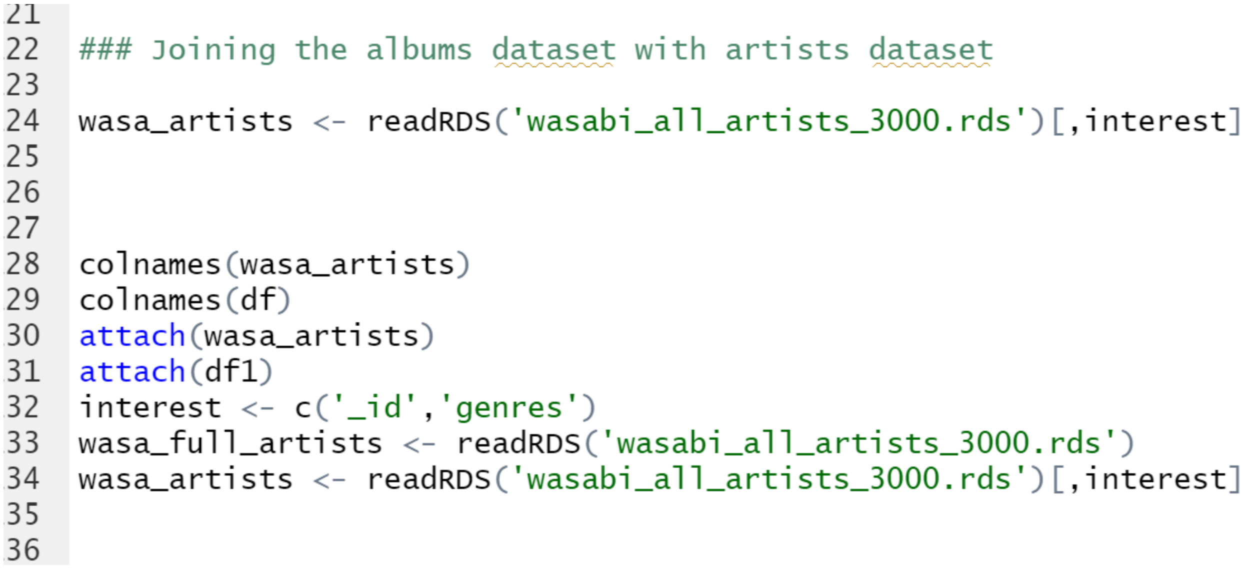
We have some date missing in the data set, so we have used only year (1st character) and replaced missing data("null'') value with the year.



```
#####

### Mixing dates column into 'Year'

df1 <- my_data %>%
  mutate(dateRelease = substr(x = publicationDate,start = 1,stop = 4),
         year = ifelse(is.na(publicationDate),dateRelease,publicationDa
  mutate(year = as.Date(year,format = '%Y'),
         year = as.numeric(substr(x = year,start = 1,stop = 4))) %>%
  select(!c(dateRelease,publicationDate))
attach(df)
```

5.4 Joining two data sets:

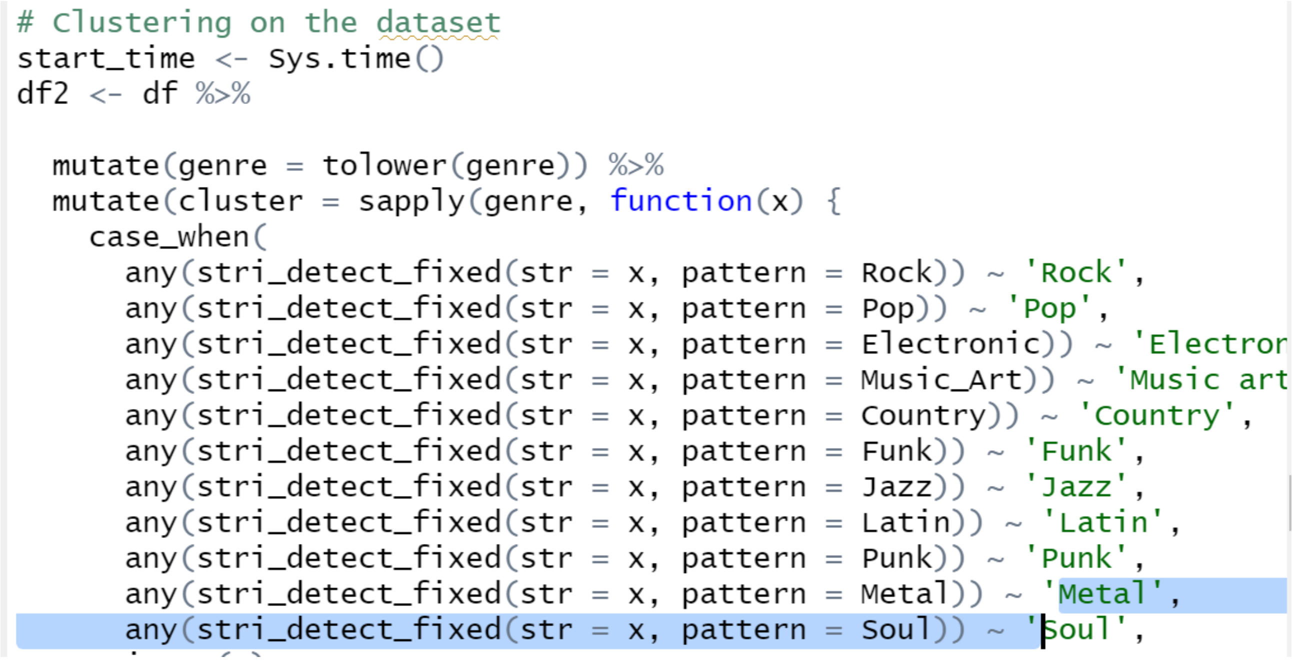
We have joined two data set such as in Artist data set we have taken both Id and genre and joined with Album data set, so as to get more informative data set.



```

### Joining the albums dataset with artists dataset

wasa_artists <- readRDS('wasabi_all_artists_3000.rds')[,interest]



colnames(wasa_artists)
colnames(df)
attach(wasa_artists)
attach(df1)
interest <- c('_id','genres')
wasa_full_artists <- readRDS('wasabi_all_artists_3000.rds')
wasa_artists <- readRDS('wasabi_all_artists_3000.rds')[,interest]


```

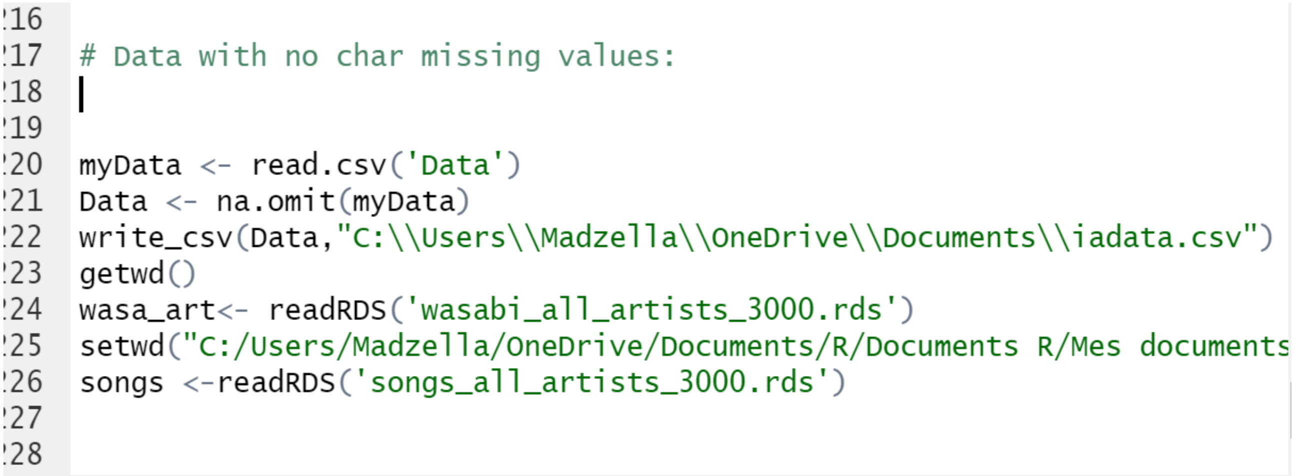
5.5 Clustering:

Clustering is a process where we have grouped all the albums based on different genre such as rock, pop, funk, jazz, Latin, metal, soul etc.

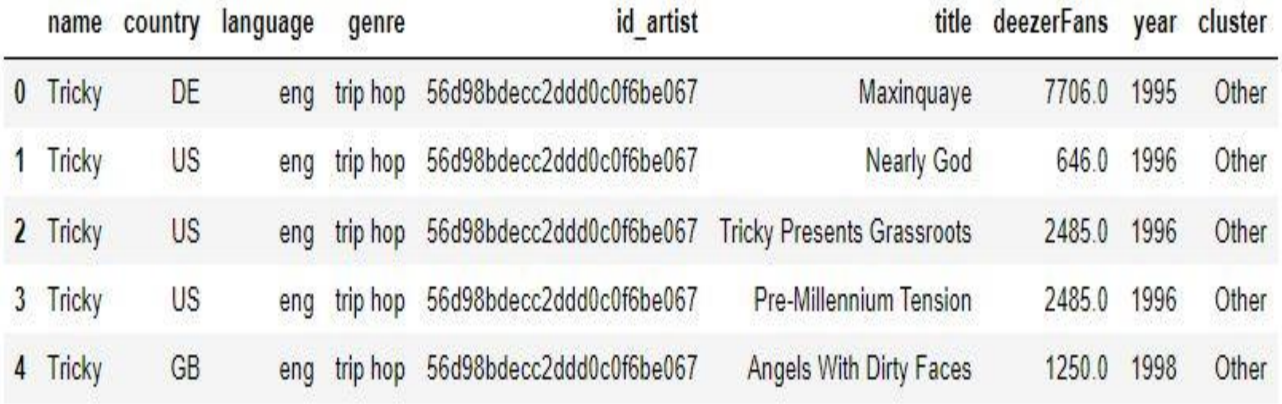


```
# Clustering on the dataset
start_time <- Sys.time()
df2 <- df %>%

  mutate(genre = tolower(genre)) %>%
  mutate(cluster = sapply(genre, function(x) {
    case_when(
      any(stri_detect_fixed(str = x, pattern = Rock)) ~ 'Rock',
      any(stri_detect_fixed(str = x, pattern = Pop)) ~ 'Pop',
      any(stri_detect_fixed(str = x, pattern = Electronic)) ~ 'Electron
      any(stri_detect_fixed(str = x, pattern = Music_Art)) ~ 'Music art
      any(stri_detect_fixed(str = x, pattern = Country)) ~ 'Country',
      any(stri_detect_fixed(str = x, pattern = Funk)) ~ 'Funk',
      any(stri_detect_fixed(str = x, pattern = Jazz)) ~ 'Jazz',
      any(stri_detect_fixed(str = x, pattern = Latin)) ~ 'Latin',
      any(stri_detect_fixed(str = x, pattern = Punk)) ~ 'Punk',
      any(stri_detect_fixed(str = x, pattern = Metal)) ~ 'Metal',
      any(stri_detect_fixed(str = x, pattern = Soul)) ~ 'Soul',
```

5.6 Generation Of Final Data Set:



```

# Data with no char missing values:


myData <- read.csv('Data')
Data <- na.omit(myData)
write_csv(Data,"C:\\Users\\Madzella\\OneDrive\\Documents\\iadata.csv")
getwd()
wasa_art<- readRDS('wasabi_all_artists_3000.rds')
setwd("C:/Users/Madzella/OneDrive/Documents/R/Documents R/Mes documents
songs <-readRDS('songs_all_artists_3000.rds')


```



	name	country	language	genre	id_artist	title	deezerFans	year	cluster
0	Tricky	DE	eng	trip hop	56d98bdecc2ddd0c0f6be067	Maxinquaye	7706.0	1995	Other
1	Tricky	US	eng	trip hop	56d98bdecc2ddd0c0f6be067	Nearly God	646.0	1996	Other
2	Tricky	US	eng	trip hop	56d98bdecc2ddd0c0f6be067	Tricky Presents Grassroots	2485.0	1996	Other
3	Tricky	US	eng	trip hop	56d98bdecc2ddd0c0f6be067	Pre-Millennium Tension	2485.0	1996	Other
4	Tricky	GB	eng	trip hop	56d98bdecc2ddd0c0f6be067	Angels With Dirty Faces	1250.0	1998	Other

6.Data visualization

Considering the above objective and complexity, World Map, Pie Chart & Gantt is the best fit for this requirement with notice of the dataset and attributes.

Tableau is a visual analytics platform transforming the way we use data to solve problems, empowering people and organization to make the most of their data, and we have used Tableau to visualize the wasabi data set.

World Map, Pie Chart & Gantt are an alternative way of visualizing the structure of a Diagram while also displaying quantities for each category.

6.1 World Map:(Mohammed Danish Mustafa)

Here we have visualized Acceptance of Rock Album over a period of time in entire World from 1964 till 2016. And how it attracted people across the continents and when and where it started it’s initial foot prints and spreading of it across the world.

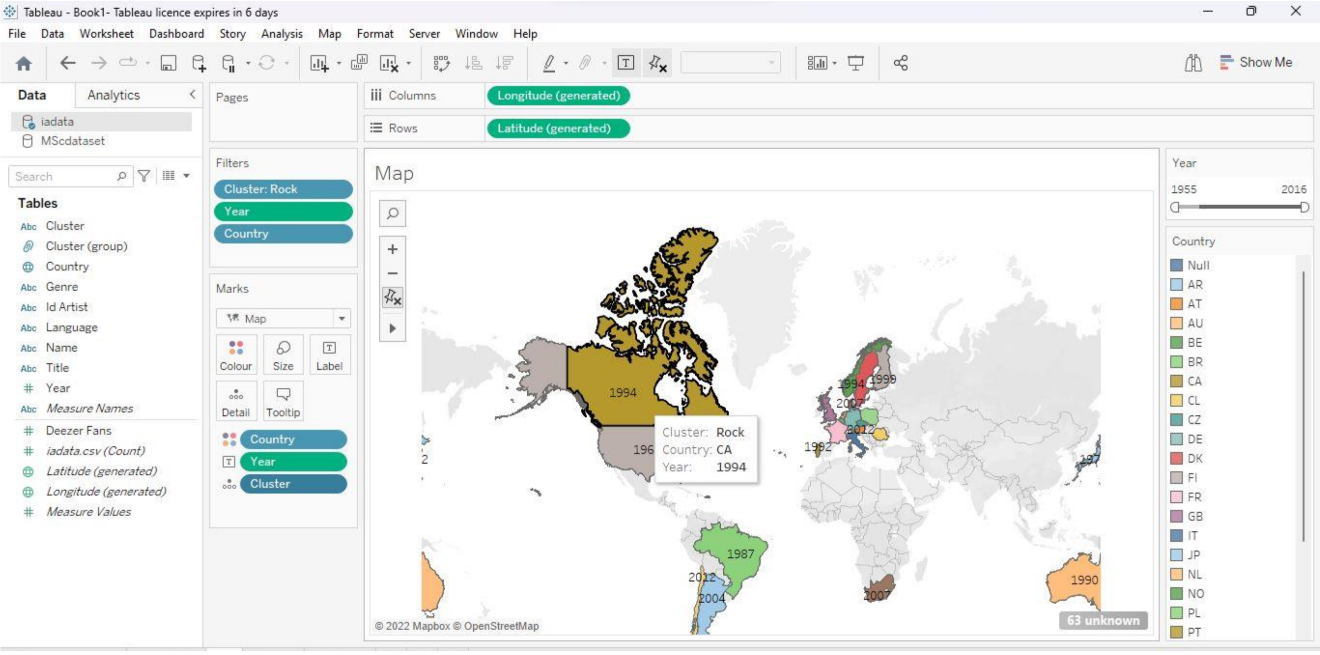


Tableau - Book1- Tableau licence expires in 6 days

Columns: Longitude (generated)

Rows: Latitude (generated)

Filters: Cluster: Rock, Year, Country

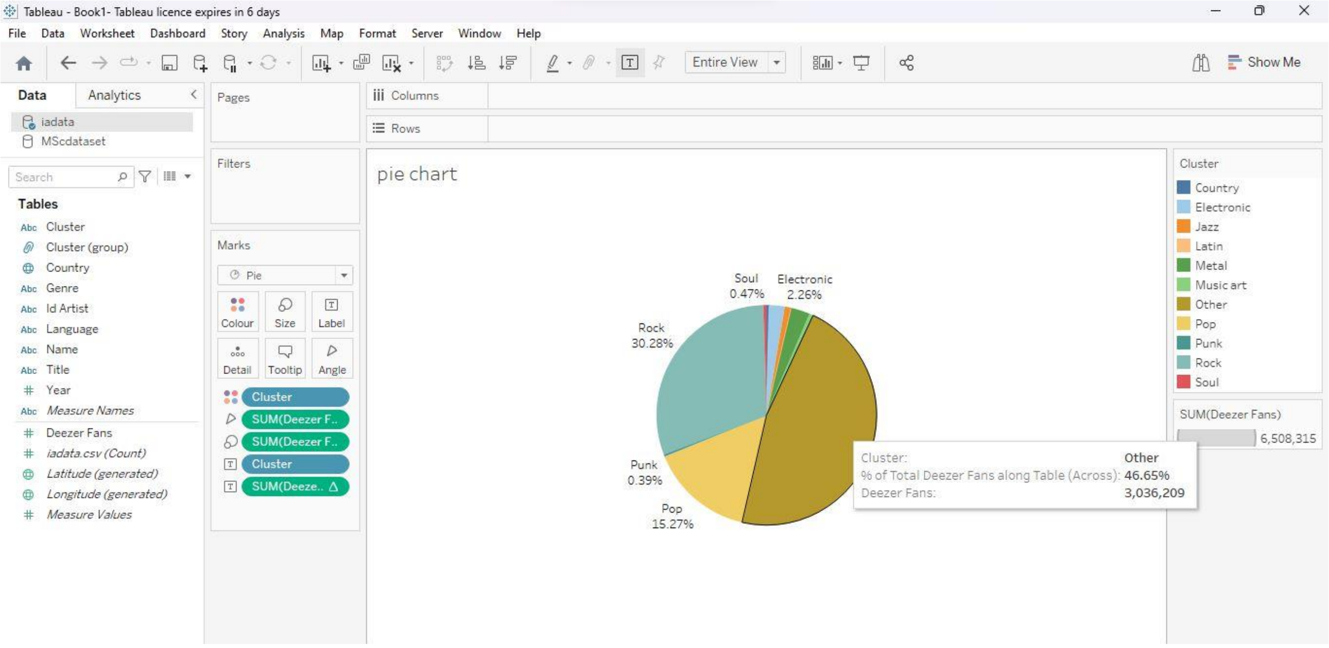
Marks: Map — Colour: Country; Label: Year; Detail: Cluster

Cluster: Rock
Country: CA
Year: 1994

Year: 1955 – 2016

6.2 Pie Chart:(MD Abdul Mazed Siddiki)

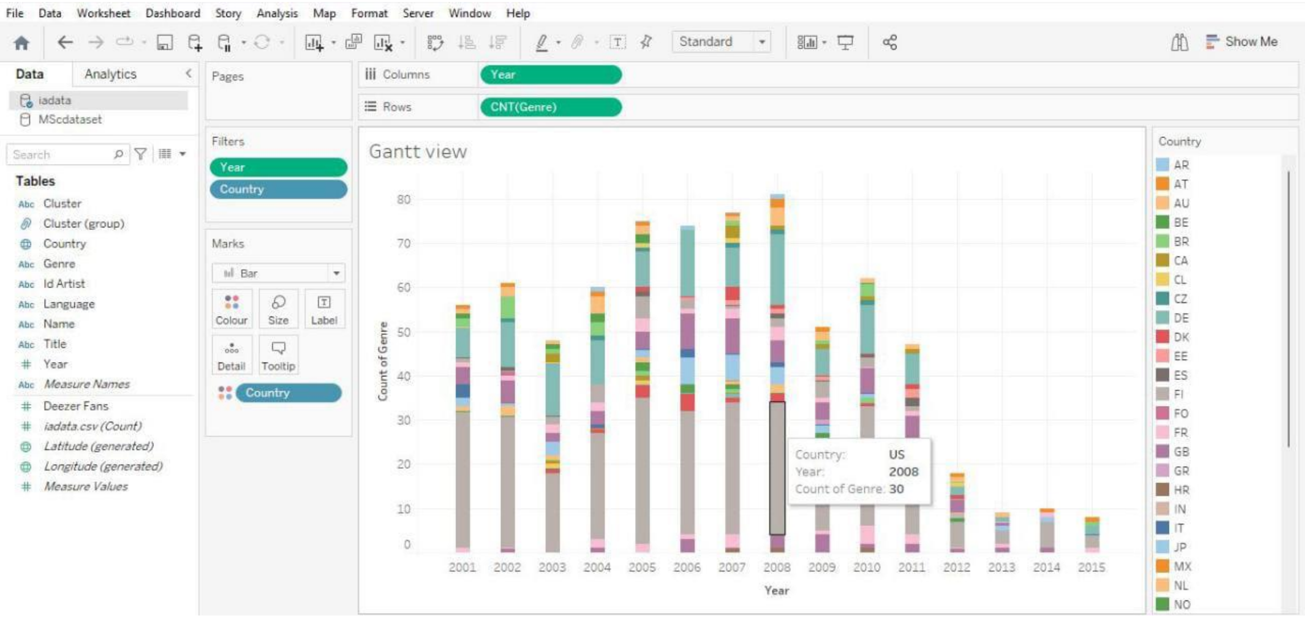
We have visualized the Percentage of people liking different Genre of Album and their ration of distribution based upon the Deezer Fans and cluster and its individual percentage of people liking the genre.



Cluster: Other
% of Total Deezer Fans along Table (Across): 46.65%
Deezer Fans: 3,036,209

6.3 Gantt: (Madzella Yannick)

Gantt is a visualizing technique which we have used to visualize number of different Genre of Album Composed in a single year by different countries.



Columns: Year

Rows: CNT(Genre)

Country: US
Year: 2008
Count of Genre: 30

7.Conclusion

- Considering Imbalance unknown data, visualization showing huge unknown part. On removal of unknown data, the remaining data is very minimal.

- Based on the data and the objective, data visualization should be chosen. Every data visualization technique has its own qualities and capabilities which may not be satisfied by other techniques.

- Data should be processed properly, otherwise while implementation the visualization, manipulating the data will be a complex task.