ntg-Zhuo Leng

Introduction

For this assignment, I analyze the Spotify 2017 playlist data based on most streamed songs throughout the year to understand the music preferences of customers in different regions using different audio features. In this dataset, preference for different music types and elements is reflected by an index and then grouped by countries. The measuring attributes in music include lyrics, genres and audio features etc.. As I was intrigued by how music preference varies by region, I developed several graphics to compare the difference among different countries more intuitively. For comparison purpose, I used data from 2017 and focused on the following measurements: o Danceability - How suitable a track is for dancing based on a combination of musical elements. o Instrumentalness - Predicts whether a track contains no vocals. o Loudness - The overall loudness of a track in decibels (dB). o Speachiness: Detects the presence of spoken words in a track based on values from 0 to 1.The more exclusively speech-like the recording (e.g. talk show, audio book, poetry), the closer to 1.0 the attribute value.

```
library(ggplot2)
library(ggmap)
```

```
## Google Maps API Terms of Service: http://developers.google.com/maps/terms.
```

```
## Please cite ggmap if you use it: see citation("ggmap") for details.
```

```
library(maps)
library(mapdata)
library(dplyr)    # yes, i could have not done this and just used 'subset' instead of 'filter'
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(ggalt)
#library(ggthemes) #


char = read.csv('country_features.csv')
ab = read.csv('ab.csv')
ab$ab=sapply(ab$ab,tolower)
df = merge(char,ab,by.x='region',by.y='ab')
head(df)
```

```
##   region danceability    energy      key  loudness      mode speechiness
## 1     ar    0.7257500 0.7360417 5.125000 -4.623083 0.5833333  0.07887500
## 2     at    0.6912941 0.6830980 5.098039 -5.353980 0.6274510  0.08312353
## 3     au    0.6981930 0.6519298 5.035088 -5.897421 0.6315789  0.11570702
## 4     be    0.6965500 0.6725333 5.116667 -5.350983 0.5833333  0.08354500
## 5     bo    0.7126316 0.6986316 4.947368 -5.114579 0.5526316  0.07873947
## 6     br    0.7057000 0.7089667 5.333333 -5.168367 0.5666667  0.09570667
```

```
##   acousticness instrumentalness  liveness   valence    tempo duration_ms
## 1    0.1794958      0.000025700 0.1446542 0.6524167 114.9790    230313.8
## 2    0.1582490      0.007525614 0.1453961 0.5402980 115.8196    212378.0
## 3    0.1740240      0.008335915 0.1496614 0.5167754 118.1546    220848.4
## 4    0.1676795      0.007813299 0.1485467 0.5499867 116.9160    217056.3
## 5    0.1474534      0.005920552 0.1604289 0.5818211 116.4790    220954.1
## 6    0.1701033      0.007422486 0.1660567 0.5766067 123.9841    212595.5
##   time_signature   country
## 1       3.958333 Argentina
## 2       3.980392   Austria
## 3       3.982456 Australia
## 4       3.983333   Belgium
## 5       3.973684   Bolivia
## 6       4.000000    Brazil
```

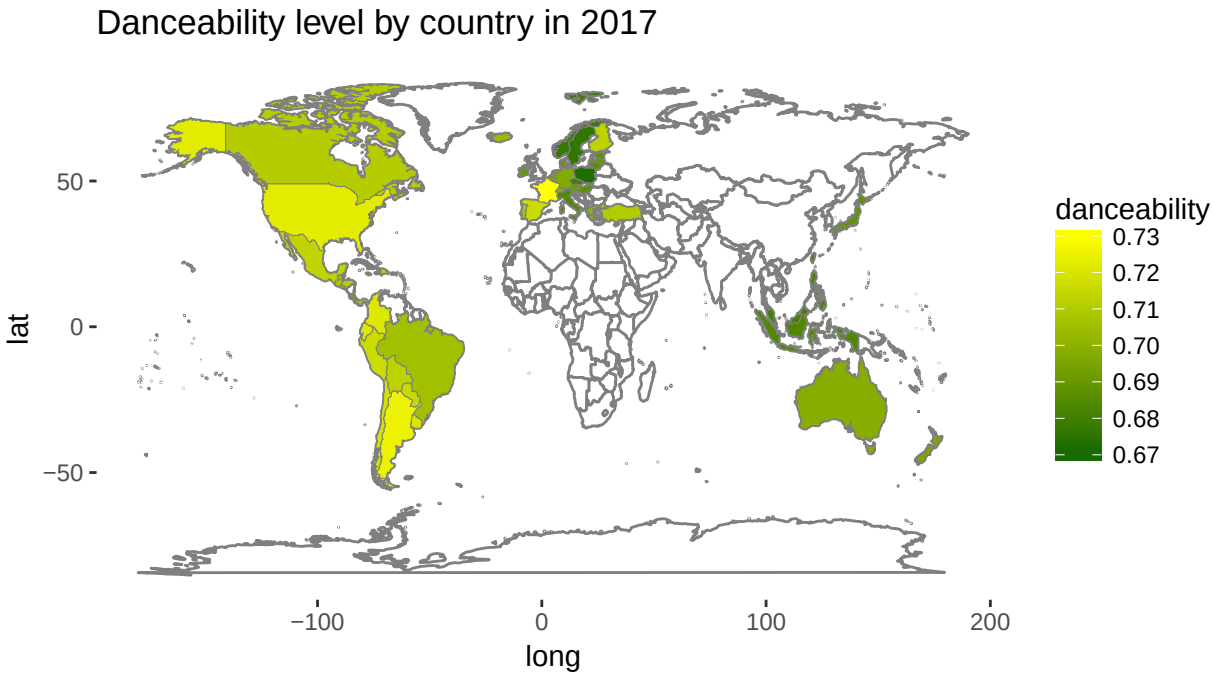
Danceability map

```
world <- map_data("world")

gg1 <- ggplot() +
  geom_polygon(data = world, aes(x = long, y = lat, group = group),
               fill = "white", color = "#7f7f7f") +
  coord_fixed(1.3)

dff <- data.frame(region=df$country,
                  value=df$danceability,
                  stringsAsFactors=FALSE)

p <- gg1 + geom_map(data=dff, map=world,
                   aes(fill=value, map_id=region),
                   colour="#7f7f7f", size=0.15) +scale_fill_gradient(low="darkgreen", high="yellow", name
    plot.background = element_blank(),
    panel.background = element_blank(),
    panel.grid.major = element_blank(),
    panel.grid.minor = element_blank(),
    panel.border = element_blank()) +
  labs(x = "long", y = "lat",
       title = "Danceability level by country in 2017")
p
```

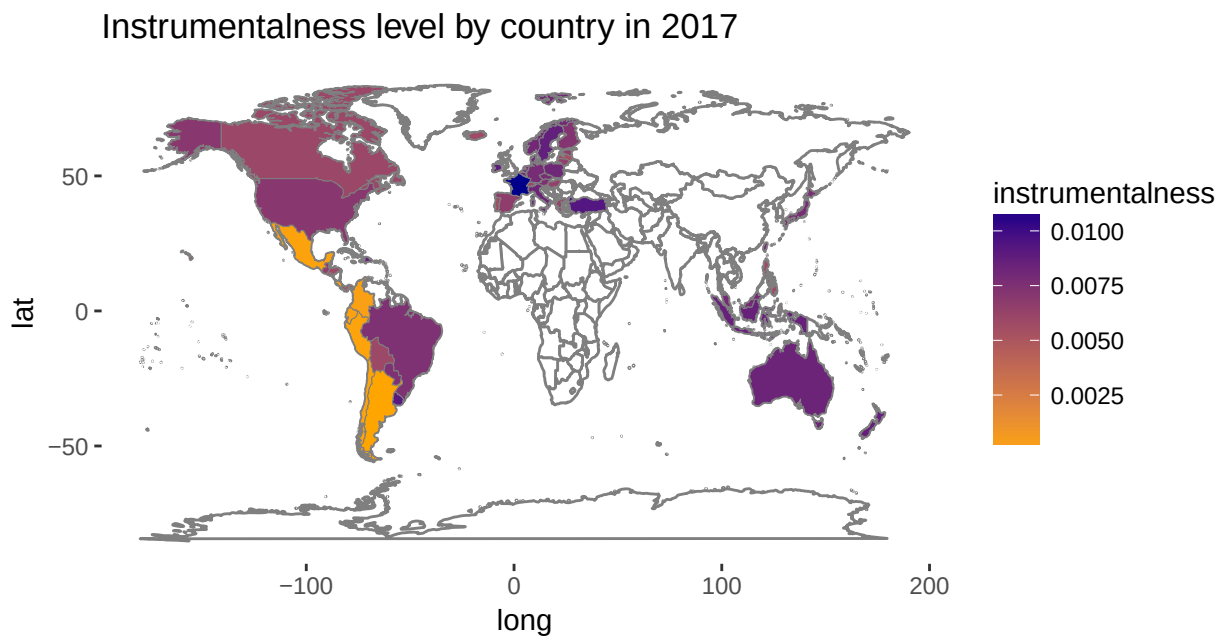


instrumentalness map

```
gg2 <- ggplot() +
  geom_polygon(data = world, aes(x = long, y = lat, group = group),
               fill = "white", color = "#7f7f7f") +
  coord_fixed(1.3)

dff <- data.frame(region=df$country,
                  value=df$instrumentalness,
                  stringsAsFactors=FALSE)

p2 <- gg2 + geom_map(data=dff, map=world,
                   aes(fill=value, map_id=region),
                   colour="#7f7f7f", size=0.15) +scale_fill_gradient(low="orange", high="darkblue", name=
    plot.background = element_blank(),
    panel.background = element_blank(),
    panel.grid.major = element_blank(),
    panel.grid.minor = element_blank(),
    panel.border = element_blank()) +
  labs(x = "long", y = "lat",
       title = "Instrumentalness level by country in 2017")
p2
```

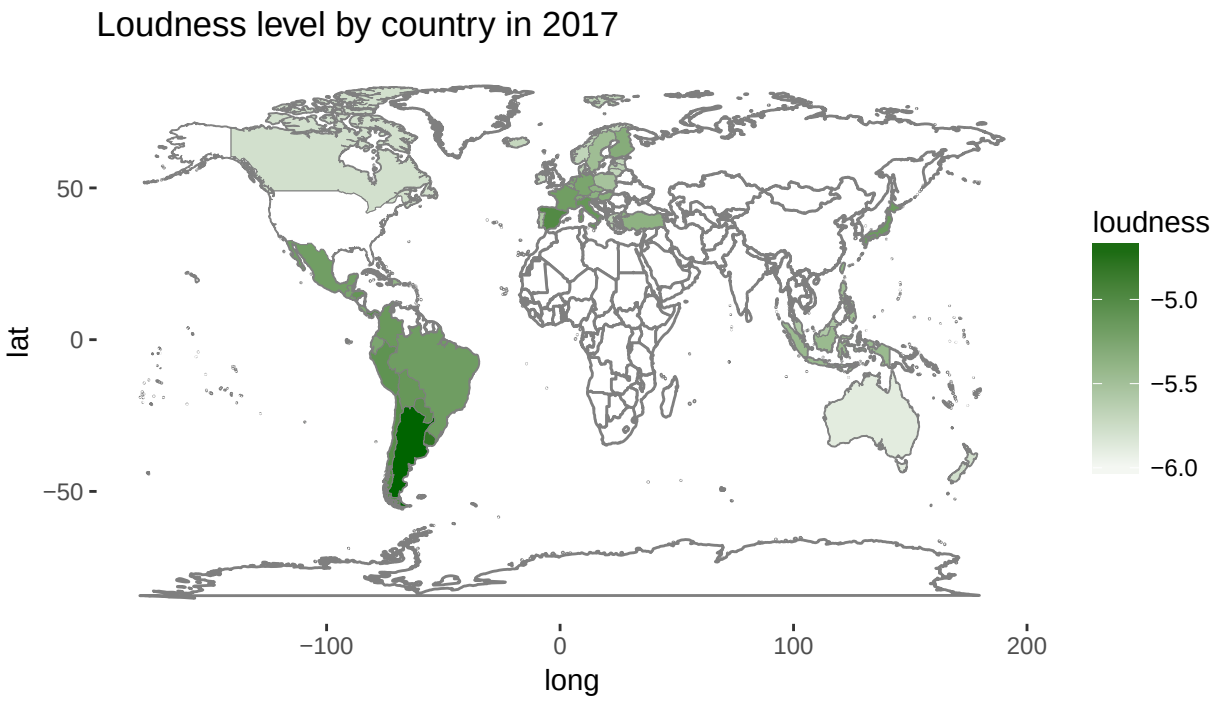


loudness map

```
gg3 <- ggplot() +
  geom_polygon(data = world, aes(x = long, y = lat, group = group),
               fill = "white", color = "#7f7f7f") +
  coord_fixed(1.3)

dff <- data.frame(region=df$country,
                  value=df$loudness,
                  stringsAsFactors=FALSE)

p3 <- gg3 + geom_map(data=dff, map=world,
                   aes(fill=value, map_id=region),
                   colour="#7f7f7f", size=0.15) +scale_fill_gradient(low="white", high="darkgreen", name=
    plot.background = element_blank(),
    panel.background = element_blank(),
    panel.grid.major = element_blank(),
    panel.grid.minor = element_blank(),
    panel.border = element_blank()) +
  labs(x = "long", y = "lat",
       title = "Loudness level by country in 2017")
p3
```

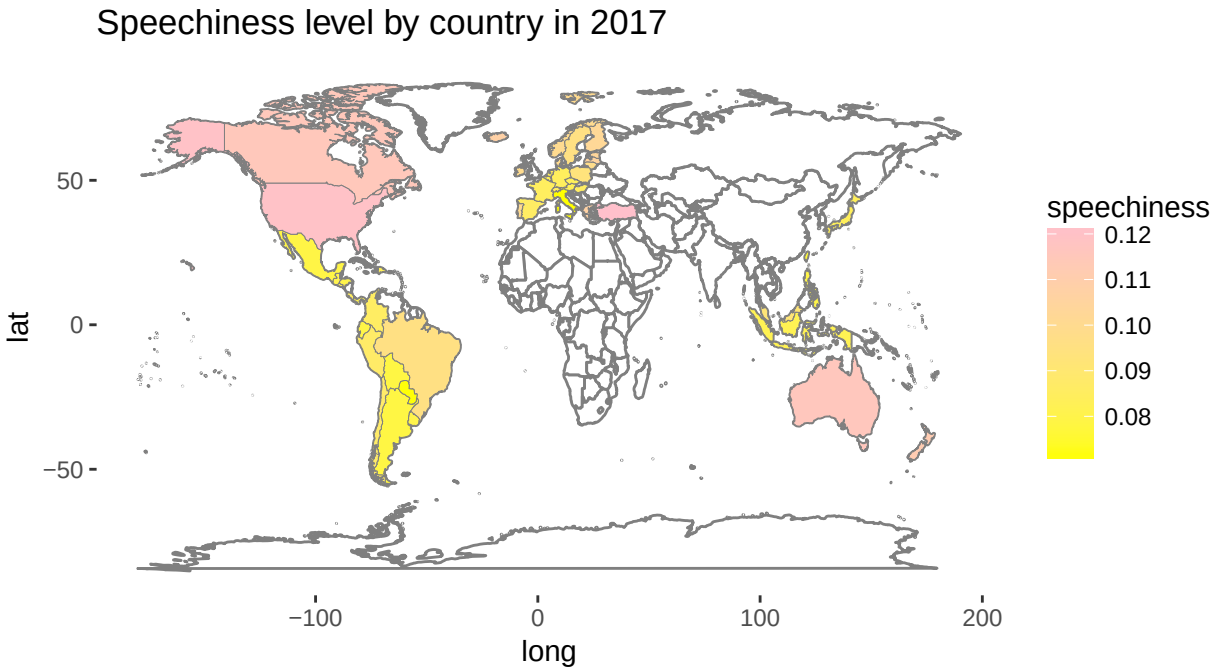


speechiness map

```
gg3 <- ggplot() +
  geom_polygon(data = world, aes(x = long, y = lat, group = group),
               fill = "white", color = "#7f7f7f") +
  coord_fixed(1.3)

dff <- data.frame(region=df$country,
                  value=df$speechiness,
                  stringsAsFactors=FALSE)

p3 <- gg3 + geom_map(data=dff, map=world,
                   aes(fill=value, map_id=region),
                   colour="#7f7f7f", size=0.15) +scale_fill_gradient(low="yellow", high="pink", name="sp
    plot.background = element_blank(),
    panel.background = element_blank(),
    panel.grid.major = element_blank(),
    panel.grid.minor = element_blank(),
    panel.border = element_blank()) +
  labs(x = "long", y = "lat",
       title = "Speechiness level by country in 2017")
p3
```



What is the story?

From the graphic I created, I was able to generate the following insights for each attribute o Danceability ??? Latin America and North America countries tend to score high on dancebility, whereas South Asia and Europe countries are relatively lower, except France. o Instrumentalness ??? most of the countries have similar scores on instrumentalness, but most of Latin American countries are below average. o Loudness ??? people in Latin American and Europe countries prefer louder music, but ingeneral, the difference among all the countries are not that obvious. o Speachiness: the US, Canada and Turkey score significantly higher on speachiness than the rest of the countries.

At same time, after comparing the same country across different measurements, I discovered that Brisiel is very different from the surrounding countries: while it is low on danceability, Brisiel has significantly higher index on speachiness and instrumentalness. Another interesting country is France: it score high on all of the measurement except speachiness, which shows very different patterns from the rest of European countires.

Why did you select this graphical form?

For this analysis, I chose the geo special heat map as the form of data visualization. There are several reasons why I selected this graphical form. First of all, this type of maps can be easily interpreted. Compared to traditional heat map, geo special heat map help the audience make quick comparison among the region by providing visual stimulation, which generate a great high level overview of the data. Secondly, the geo special heat maps provide additional information for readers who are not familiar with the countries in this data set. Instead of spending additional time searching the location and size of the countries, the readers can easily

gather the information from my graphics. Furthermore, this graphical form can help discover time-based trend from future analysis if adding the year attribute to each of the four measurements.

What challenges did you encounter in creating the visualization?

Although I find that the geo special heat map is a great graphic form to communicate the story, there are several challenges and limitations. The first challenge I faced when I was creating the graphics is that some of the country names cannot be detected by R. I need to spend extra time going through the country names in my dataset to make sure they can be matched. Another challenges is that heat map can be hard to interpreted when the difference among the regions are small. Besides, heat map is not able to show sample density from the data set. If it is not used correctly, the graphic can be misleading. To avoid this problem, I have checked outlier and sample size for each region before creating the visual.