1 Data

Spotify is one of the largest music streaming platform with the number of subscribers in the world. In the 21st century, where physical music is gradually declining, the Spotify dominates music consumption and allows the user to access the largest music collections in history, as well as podcasts and other audio content. According to the company's Q1 report published in 2018, there are 207 million active Spotify users per month, 96 of whom are Spotify Premium subscribers.

Our spotify data set includes various information about the most listened 200 songs per day in 53 countries from 1st January 2017 to 9th January 2018. It contains over 2 million rows, consisting of 6629 artists, 18598 songs counting a total of one hundred five billion streams. This dataset allows us to analyze many topics that we can associate with music. We can aslo how the behaviour of listening to music changes according to country and time.

We have received data from kaggle (https://www.kaggle.com/nadintamer/top-tracks-of-2017), which collected it from Spotify's regional chart data (https://spotifycharts.com/regional) by using Spotify Web API.

Column Variables:

1) Date: The date variable consists of 371 days, 53 weeks between 1st January 2017 and 9th January 2018. When we import data into R, dates and times are usually stored as character or factor by default due to symbols such as "-", ":" and "/". Using the str or class functions tell us how this variable stored. If dates stored as character or factor that means we can't calculate or summarize elapsed times. We used R's core function as.Date to convert the Date variable to the date format.

```r
library(dplyr)
library(magrittr)
library(knitr)

data <- read.csv("data/data.csv", encoding="UTF-8")
class(data$Date)
```

```
## [1] "factor"
```

```r
data %<>% mutate(Date=as.Date(Date,format= "%m/%d/%Y" ))

class(data$Date)
```

```
## [1] "Date"
```

2) TrackName: TrackName variable shows the track names in the daily top 200. There are 18598 different track names in total.

3) Artist: Artist vaiable indicates to the artist to which the relevant song belongs. Total number of different artists 6629.

4) TrackId: TrackId variable consists unique values for each track and refers to the track ids at the Spotify Web API. As we would see in the experimental section, we used these ids when extarcting data from Spotify Track API.

5) Region : Region variable contsist ISO 3166-1 two-letter country codes of 53 countries and global

6) Streams : Streams variable shows the number of times the song is played in the relevant date and region. Among the top 200, Luis Fonsi has the most stream number (11381520) with Despacito (Featuring Daddy Yankee) track in global.

Date	TrackName	Artist	TrackId	Region	Streams
2017-01-01	Hasta la luna	#TocoParaVos	1A3yLIqelu1D7MCtw06Cu7	ar	35806
2017-01-01	Una historia	#TocoParaVos	6Ol4gn2vFyn4dckha1fGaN	ar	21108
2017-01-01	Sólo necesito	#TocoParaVos	7kK1cPOZ9RxZcgamiYysxw	ar	14397
2017-01-01	Pájaro Cantor	Abel Pintos	16rQTTMtbn0ODLyflr0O6Q	ar	25525
2017-01-01	Le Hace Falta un Beso	Agapornis	300xUYXWhFA6Q1sbNvVItP	ar	45251
2017-01-01	Baila	Agapornis	3iS4VBJ3Bg1HyL3e6WNpgi	ar	42127
2017-01-01	Si Te Vas	Agapornis	3HoTiqEEY2SmTeq1XGINo3	ar	24819
2017-01-01	Dale Tu Amor	Agapornis	3W9HjL7lmbYXL4kQ1kicH3	ar	24345
2017-01-01	La Llave	Agapornis	1v78qANb9b176fgEzR2eaW	ar	20952
2017-01-01	Me Estoy Enamorando	Agapornis	0mX14Kpe4VTYmkqRgGslT5	ar	20900
2017-01-01	Ella	Agapornis	6LxbbZYulsV3tCZvKpJgSh	ar	20643
2017-01-01	Hoy	Agapornis	11aJTEmPwKKoKRTpsudE1H	ar	19864
2017-01-01	I Will Be There	Agapornis	60LbXuVjG11t0cIendTtHh	ar	16424
2017-01-01	Persiana Americana	Agapornis	4amGeAvkzFunJTHZ0KJZba	ar	15501
2017-01-01	Tus Ojos No Me Ven	Agapornis	5lZzrsIShvSlXlYlJiaiEf	ar	14483
2017-01-01	Te Vas - Version Cumbia	aLee DJ	0j8zA8KrITjLVoAm3H3X6u	ar	23981
2017-01-01	Deja Que Te Bese	Alejandro Sanz	0s6yB0ZFl1LvYHAnGhXFga	ar	15920
2017-01-01	Una Cita - Remix	Alkilados	6v3QCw4ca5cR4VkzpG0PHy	ar	18103
2017-01-01	Yo Tomo Licor	Amar Azul	5EHRc2WJASyZg3alJqsbqU	ar	19889
2017-01-01	Nunca me Faltes	Antonio Rios	3YCcsuRdJArE4rsAG5V3sW	ar	18253

2 Theory

2.1 Correlation

Correlation is a statistical method used to test relationship between two variables or the relationship of a variable with two or more variables, to measure the degree and direction of this relationship. Correlation does not find the best-fit line through the data points, as in the linear regression statistic. It just produces the correlation coefficient that shows how much a one variable tends to change with the change of the other variable. The correlation coefficient is indicated by the τ symbol and takes values between -1 and +1. Positive values indicate that variables go up or go down in paralel, negative values indicate one decreases while the other decreases. If the correlation is positive it is called as positive or direct correlation. If it is negative,called inverse or contrary correlation. A correlation does not mean that there is a causal relationship.

If the τ is close to 1 or -1, it can be infered the relationship between variables is strong. If τ is 0, there is no linear relationship between these variables.

In statistics, usually three main types of correlation is measured: Pearson correlation, Kendall rank correlation and Spearman correlation.

http://www.oicstatcom.org/file/TEXTBOOK-CORRELATION-AND-REGRESSION-ANALYSIS-EGYPT-EN.pdf

2.1.1 Pearson Correlation

Pearson correlation is the correlation statistics used to measure the degree and direction of the relationship between linearly related variables. Before the correlation coefficient is calculated, it is necessary to check whether there is a linear relationship by using the scatter graph. The magnitude of τ indicates how close the data points are on a straight line in the scattering plot.

The fact that τ is close to -1 indicates that there is a very strong negative linear relationship between these variables and a very strong positive linear relationship if it is close to + 1. The fact that τ is greater than

0.7 allows us to interpret that the linear relationship is strong. The number of data is important when evaluating the correlation coefficient. If the number of data increases, the result can be more reliable as the impact of the coincidental causes will decrease.

2.2 Paired t-test

2.3 Markov Chain Model