COMP2501 Proposal

Dataset introduction

This dataset contains 100 most streamed songs on *Spotify* with their features extracted using the Spotify API. There are two given files named `Features.csv` and `Streams.csv` both sorted by their rankings. The first one renders some relevant features about every top song including `name`, `duration`, `key`, etc., and an unique identifier `ID`. The other file contains some basic information of every song, such as its name `Song`, its singer `Artist`, number of song plays `Streams`, and its release time `Release Date`. The data is available on Kaggle.

Data preprocessing

Before I conduct some exploratory data analysis on the data, I need to first preprocess it, which removes missing or inconsistent data values resulting from human or computer error. Thus, it can improve the accuracy and reliability of the data. Firstly, as there are two files given as dataset, we need first to merge them to form a single table for better analysis afterwards. Also, the olumn `Release Date` is represented with both numbers and characters, which makes it difficult to take comparisons and numerical calculations. Thus, I need to convert it to another data type, i.e., `Date`. Furthermore, since there is no missing data, we will not drop any observations.

Interested questions

The profit development trend of the music industry has changed significantly in recent years, mainly due to the influence of digital technology and the internet. From this dataset, I am curious about some basic information about these top songs. Typically, I would like to find:

- How many artists made it to the top 100 most streamed songs of all time on *Spotify*?
- Who are the artists with more songs in the top list?
- Who are the artists with more streams in total in the top list?
- What is the newest song and the oldest song respectively in the list?
- Which year did the most number of songs released on the list?

To go deeper, I am also intereseted in some intrinsic links among different varibles, including but not limited to the following tasks:

- Is there a pattern in terms of `key` and `mode` for these top songs?
- Do these songs gravitate around an `tempo`?
- Do these songs gravitate around an `optimal duration`?
- Is there any relationship between any pair of attributes of top songs? Which pair has the most positive or negative correlation?
- Is there any other interesting findings of the distributions of characteristics of songs in the list?
- Can the songs be clustered as groups based on their features?

Finally, I would like to model the **streams** solely with the features given in the dataset. To be more precise, I want to explore:

- Can I predict the `Streams (Billions)` using different machine learning models? How is the fit for each model?

Methods

- **Platform**: RStudio (4.2.2)
- **Packages**: datasets, tidyr, dplyr, rvest, stringr, lubridate, ggplot2, dslabs, caret, tidyverse, gridExtra, purrr, randomForest (versions are temporarily not decided)
- **Solutions**: data preprocessing, data visualization, clustering, linear regression, random forest

Feature works

In the future, we could collect the top streamed songs in a monthly or quarterly basis instead of all-time. Thus, we can find the potential transitions of pop songs. This can be recognized as the current trend of music. For customers, they could easily know the recent hot songs. Meanwhile, record companies and producers can use some analysis as a guide to better grasp the current trends to innovate and provide better music to everyone.

References

Suh, B.J. (2019). International Music Preferences: An Analysis of the Determinants of Song Popularity on Spotify for the U.S., Norway, Taiwan, Ecuador, and Costa Rica.