Digital Receipt

This receipt acknowledges that Turnitin received your paper. Below you will find the receipt information regarding your submission.

The first page of your submissions is displayed below.

Submission author:	Philip Steinke
Assignment title:	Assignment 3:
Submission title:	Assignment 3 preprocessing music…
File name:	MATH2349_Assignment_3_AshBe…
File size:	0
Page count:	21
Word count:	5,655
Character count:	32,958
Submission date:	03-Jun-2018 11:41PM (UTC+1000)
Submission ID:	971740063

MATH2349 Semester 1, 2018

CODE

Assignment 3

Ben Cole s3412349, Ashleigh Olney s3686808, Phil Steinke s3725547, Elleni Toumpas s3708633

Links: - Github (https://github.com/tgrn/data-preprocessing-assignment3/) - assignment notes (https://drive.google.com/open?id=1dgDg_sO-G4ZX_N3BX03XbCPhGFFAeAZ5ZI5QlNoFND4) - [How do work on R Markdown and Rpubs]

Required packages

Executive Summary

We found a datasource with 9 tables that gave us the potential to explore our abilities to merge datasets. However upon investigating the full scope of all the information and variables we had access to we realised playing and cleaning with ALL of the data would make this project a lot bigger than it needed to be. Therfore before proceeding with the preprocessing a decision had to be made to select a subset of this data to play with.

Once the data was subsetted problems were experienced even from the import, with the multi-index nature of the original data creating issues with column names and data types on import. Actions were required to fix this, correct data types and rename a few factors.

Understanding the data to get preliminarily insight into data that may form impossible records and contain missing values was performed. Moving in to making our data abide by the tidy data concepts, genre_ids column was separated then gathered in to a new column so that each genre id for a song was recorded to a separate row. Latitude and longitude was also gathered to abide by the tidy data rules. Dealing with missing values, and outliers followed by steps to transform.

Data

The dataset was originally found on the UCI Machine Learning Repository and came from a project called FMA: A Dataset For Music Analysis Data Set [LINK: https://archive.ics.uci.edu/ml/datasets/FMA%3A+A+Dataset+For+Music+Analysis (https://archive.ics.uci.edu/ml/datasets/FMA%3A+A+Dataset+For+Music+Analysis)].

The original data is split in to nine csvs: the tracks.csv, genres.csv, features.csv, echonest.csv along with the raw_albums.csv, raw_artists.csv, raw_echonest.csv, raw_genres.csv and raw_tracks.csv.

Out of all nine csvs we had decided to only concentrate on three of them. The echonest.csv, the tracks.csv and the genres.csv. Not all variables were ultimately dealt with, but the data descriptions of the three are below.

TRACK.CSV is a dataset with 53 variables and 106,574 records. The columns we decided to work with were: - track_id [integer] an id used for track - album.type [character] a factor/ categorical data classifying the album type - album.title [character] the title of the album - album.tracks [integer] referencing the track on the album - album.listens [integer] the number of album listens - artist.name [string] the artist releasing the album/ song - artist.latitude [float] the artist's latitude location - artist.longitude [float] the artist's longitude location - track.interest [integer] - track.duration [integer] the length of the song in seconds - track.genre_top [string] the top genre tag describing the song - track.genres [integer] a list of all the other description tags allocated to genre - track.composer [string] the composer working on the track - track.genres_all [string] all the genre descriptors

GENRES.CSV is a dataset with 5 variables and 163 records. - genre_ID [integer] an id used to match genres. - #tracks [integer] number of tracks recorded against this genre - parent [integer] - title [character] the title of the genre - top_level

ECHONEST.CSV has 250 variables and 13,129 records - audiofeatures (acousticness, danceability, energy, instrumentalness, liveness, speechiness, tempo, value columns) [double] all values measured by the echonest project to capture the features of a song - metadata.album_date [date] the album date the album was recorded - metadata.album_name [character] the album title the track belonged to - metadata.artist_latitude [double] the artist's latitudal location - metadata.artist_location [character] a string representating the artist's location (city, state, country) - metadata.artist_longitude [double] the artist's longitidal location - metadata.artist_name [character] - the name of the artist releasing the track - metadata.release [character] - the title of the release - ranks.artist_discovery_rank, artist_hottness_rank, song_currency_rank, artist_familiarity, artist_hottness, song_currency, song_hottness [integer] all values related to the popularity, social feeds, ranks of songs and the artist - temporalfeatures (000 - 250) temporal features that were recorded using the echonest project to capture the features of a song.

```
echonest%>% head(1) %>% kable("html") %>% kable_styling(full_width=F, bootstrap_options="striped")
```

X1	audio_features	audio_features_1	audio_features_2	audio_features_3	audio_features_4	audio_features_5	audio_features_6	audio_features_7	met
NA	acousticness	danceability	energy	instrumentalness	liveness	speechiness	tempo	valence	alb

```
tracks %>% head(1) %>% kable("html") %>% kable_styling(full_width=F, bootstrap_options="striped")
```

X1	album	album_1	album_2	album_3	album_4	album_5	album_6	album_7	album_8	album_9	album_10	album_11	album_
NA	comments	date_created	date_released	engineer	favorites	id	information	listens	producer	tags	title	tracks	type