etl project

**‘top music artists’**

**Alexis Palmer, Brynna Bridges, Jewell Foster**

## Table of Contents

1. Introduction
   1. [Summary](#_introduction)
   2. [Scope](#_introduction)
   3. [Technologies and Resource contributions](#_introduction)
   4. [Definitions, Acronyms and Abbreviations](#_introduction)
2. ETL Details
   1. [Data Import/Extract Sources and Method](#_ETL_DETAILS)
   2. [Data Acquisition](#_ETL_DETAILS)
   3. [Data Transform](#_ETL_DETAILS)
   4. [Data Integrity](#_ETL_DETAILS)
   5. [Data Refresh Frequency](#_ETL_DETAILS)
   6. [Data Security](#_ETL_DETAILS)
   7. [Data Loading and Availability](#_ETL_DETAILS)
3. Data Quality

## introduction

* 1. **Summary**

The goal of this project was to extract data from two or more sources like Kaggle and perform Extraction, Transformation and Load (ETL) on the top music artists of today for a music company. To assist our client our team gathered information from Billboard and Last.FM on the most popular artist (by plays and weeks on top charts) to determine whether its relevant information to promote an artist by genre and number of listeners.

* 1. **Scope**

The Billboard dataset comes from the weekly updated HOT100 section in Billboards charts from 2015 to 2019. Included in this dataset is the top 100 song titles for every week, the artist that performed the songs, how many weeks the song was at number one, how many weeks the song was on the top100 chart at all, and the peak rank the song achieved on the chart.

The music-artists-popularity dataset consists of over 1.4 Million musical artists present in MusicBrainz database -- their names, tags, and popularity (listeners/scrobbles), based on data scraped from Last.fm. This dataset contains the MusicBrainz every artist name from Last.fm, the country the artist is from according artist country, based on Last.fm tags, artist tags on MusicBrainz, artist tags on Last.fm, number of listeners on Last.fm, number of scrobbles on Last.fm, and the last column is marked true if more than one artist shares the same Last.fm page.

From the Billboard dataset we used the artist, weeks on #1 and weeks on chart. And from music-artists-popularity we choose to keep artist last.fm, country last.fm, tags last.fm, listeners last.fm and scrobbles last.fm.

**1.3 Technologies and Resource Contributions**

Due to the short timeline, we gathered datasets and collaborated on which sets we thought would be more useful to create the idea needed for the project. As a team we determined the columns needed, and cleaned the data using Jupyter Lab and which databases we were going to transform the data into. We used Python data to import our data to a csv and structure it into a Pandas DataFrame. We then used Postgres to load our new dataframes and perform queries.

Group Members: Brynna Bridges, Jewell Foster and Alexis Palmer

**1.4 Definitions, Acronyms and Abbreviations**

ETL: Extract, Transform and Load

PD: Pandas

Scrobbles: Number of plays for each song

Tags: Genre

## ETL DETAILS

* 1. **Data Import/Extract Sources and Method**

We used 2 data sets from a free public platform Kaggle which provided the Billboard and Last.fm data.

The data of interest include:

Music Popularity Artist

* Mbid
* Artist\_mb
* Artist\_lastfm
* Country\_mb
* Country\_lastfm
* Tags\_mb
* Tags\_lastfm
* Listeners\_lastfm
* Scrobbles\_lastfm
* Ambigous\_artist

Billboard

* Song
* Artist
* Weeks on #1
* Weeks on Chart
* Peak Rank

The data that we would like is country of origin, genre, number of listeners, weeks on Billboard charts, and other relevant information to decide what type of artists would be most successful to promote at the current time. The music artist popularity dataset was too large to upload to GitHub so you can access the data in the link provided below.

Here are the following sources we used for our datasets:

* <https://www.kaggle.com/saberianz/billboard-charts>
* <https://www.kaggle.com/pieca111/music-artists-popularity>
  1. **Data Acquisition**

Going forward this database should be updated weekly by scraping Billboard and last.fm top 100 charts. The data will be updated with each weekly scraping. We would suggest making a new dataset for the next five years starting in 2020 for the Billboard data.

The music artist popularity dataset requires a lot of cleaning and attention to what is actually is in each column.

* 1. **Data Transform**

To transform the data needed we performed the following:

* Created tables using Pandas functions in Jupyter Lab to upload the Spotify and Billboard csv files we used.
* Created dataframes with the columns needed
* Removed columns we didn’t need
* Dropped the null values
* Removed the special characters and spaces we found as we were cleaning the data and determined the columns needed to be changed to gather the correct data.
* Dropped the duplicate Artist and merged both datasets on Artist.
  + Split the artist column on featuring and keep the first value
  + Keep the first value if there were multiple artist
* Split the Genre column on the first semi-colon and keep the first value
  + If the value was seen live we used the second value
* Grouped by Genre to get the count and added it back to the Artist dataframe
  + Dropped all genres with counts of less than 750
* Split the Country column on the first semi-colon and keep the first value
* Renamed the columns as follows
  + 'artist\_lastfm': 'Artist'
  + 'country\_lastfm': 'Country'
  + 'tags\_lastfm': 'Genre'
  + ‘listeners\_lastfm': 'Listener Count'
  + 'scrobbles\_lastfm': 'Play Count'

Initial Music Artist Popularity DataFrame:

A screenshot of a computer screen

Description automatically generated

Final Music Artist Popularity DataFrame:

A screenshot of a cell phone

Description automatically generated

Initial Billboard DataFrame:

A screenshot of a cell phone

Description automatically generated

Final Billboard Dataframe:

A screenshot of a cell phone

Description automatically generated

* 1. **Data Integrity**

The Last.fm dataset had several duplicate Artist values that needed to be manipulated and cleaned. Some of the Country data was missing. The tags values are from users versus experts so they are not as reliable for comparisons. Some of the Artist names had special characters especially from the MusicBrainz dataset which is why we chose Last.fm.

**2.5 Data Refresh Frequency**

The datasets we have collected are updated weekly. This database should be updated weekly by scraping Billboard and Last.FM top 100 charts.

**2.6 Data Security**

There are no data security issues or privacy related to these datasets.

**2.7 Data Loading and Availability**

1. Number of rows of data needed from the datasets
2. Unique Identifier in the datasets (Artist)
3. Verified formats of alphanumeric data and all the rows
4. Values for Artist and Genre
5. We setup views in PostgreSQL for the client to pull Top Artist by listener count, weeks on #1, weeks on chart, and play count.

A screenshot of a cell phone

Description automatically generated A screenshot of a cell phone

Description automatically generated

A screenshot of a cell phone

Description automatically generated A screenshot of a cell phone

Description automatically generated

## DATA QUALITY

The Client requested that we gather data on the most popular artist. The KPIs are Top Artist based on listener count, weeks on #1, weeks on chart, and play count. We verified that information by looking at the Wikipedia page for the top artist and confirmed that it’s the same. This verification should be performed monthly.

Based on the values that we have in our database they should be able to gather the data needed weekly and make decisions based on this information. They can use the views in Postgres and let us know if they are needing any additional data.