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.

This section summarized the final objective of the project, the business problem definition (problem statement) and the expected outcome of ETL.

**1.2 Scope**

This section explicitly outlines the disparate data sources that are to be integrated, which components of the overall data science project is in the scope for this initiative and also lists out the components of the data science project that are not in scope here.

**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.

Group Members: Brynna Bridges, Jewell Foster and Alexis Palmer

This section lists out the team members and their contributions towards the ETL initiative. Use this section to also outline (or list) the tech stack used to obtain the final outcome.

**1.4 Definitions, Acronyms and Abbreviations**

List acronyms and terms that need to be defined in this section, such as ETL: Extract, Transform and Load

## ETL DETAILS

* 1. **DataImport/ExtractSourcesandMethod**

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

The data of interest include: ( list all the columns of the data)

* Artist
* Country of origin
* Genre
* Number of listeners
* Play Count
* Weeks on Charts
* Weeks on #1
* 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. This database should be updated weekly by scraping Spotify, Billboard, and last.fm top 100 charts.

Here are the following sources we used for our datasets:

* <https://www.kaggle.com/leonardopena/top-spotify-songs-from-20102019-by-year>
* <https://www.kaggle.com/saberianz/billboard-charts>
* <https://www.kaggle.com/pieca111/music-artists-popularity>

This section provides information about the data and its source. For example, API names and URLs, key parameters available and its subset which will be preserved (loaded). Data extraction protocols (API, FTP, Web scraping etc.), any permissions required to access the said extraction dataset and any restriction placed on the usage and distribution of the acquired dataset.

**2.2 DataAcquisition**

This section outlines the data needed, such as range and if the data is static or dynamic and needs continuous update. Outline the process to obtain again or update the dataset. The formatting and any special attributes about the data the one should be mindful of while obtaining and processing the raw dataset. How to decide on the selection of data while re-obtaining or updating. Discuss, here the dimension of the obtained dataset and if updated what is the project growth rate of the data. Lastly, address any issues or pre-requisites that needs to be cleared prior to getting the data?

* 1. **DataTransform**

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
* 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.
* Grouped the duplicate Artist and merged both datasets on Artist.

In this section address any data transformation that needs to be performed to modify, clean, filter or create existing and new parameters. Address any technical analysis performed, include design specification or data models used (example linear interpolation etc.), and any calculations performed for any newly derived fields.

**2.4 DataIntegrity**

In this section discuss the reliability of the extraction source data (e.g., missing data, dates stored as text, invalid code values, text fields with odd characters, etc.). Address the frequency with which the data sources are updated and if it is necessary to update the local data at the same frequency. Lastly, how if any notification can be received when the source data is updated; and what if any notification will be sent to the internal team when the local dataset is updated.

**2.5 DataRefreshFrequency**

The datasets we have collected are updated weekly. This database should be updated weekly by scraping Spotify, Billboard, and last.fm top 100 charts.

This section explicitly lists the frequency with which this ETL process will refresh the local dataset (Daily, Weekly, Monthly, Quarterly, Semi-Annually, etc.).

**2.6 DataSecurity**

This section discusses any data anonymity and security requirements need to be satisfied. Address any federally mandated HIPAA considerations, any need to build in additional privacy, Encryption, Data masking, Auditing, Backups etc.

**2.7 DataLoadingandAvailability**

This section addresses the data schema and during of data retention. Discuss the interface that will allow your Client/Users to access the data.

1. Number of rows of data needed from the datasets
2. Unique Identifier in the datasets
3. Verified formats of alphanumeric data and all the rows
4. Values for Artist and Genre
5. What categories

## DATA QUALITY

Address in this section success criteria for this project. Summarize the parameter KPIs such as Totals and expected counts. What user acceptance testing was performed and what were the outcomes. What is the recommended site acceptance testing that your client can perform to ensure the expected outcomes meets their expectations?