# Data cleaning for Artists datasets

Artists dataset is extracted from MusicBrainz database which is the reference database of Music related data containing artists, events, songs. The database is provided as an opensource database and monthly dumps are available for download. Getting data analysed and observed the missing data, some cleaning were needed. Moreover, in order to improve the ease of use, transformations were needed.

# Dataset Specification

This dataset contains artists’s information and the unique id of the artists which is the main reference of the artist in other applications and studies. This ***gid*** is known as ***mbid*** globally. In addition to ids, name, sort\_name, gender, type, ended, start\_date, end\_date, begin\_area and end\_area are existing the table. The extracted columns is presented below:

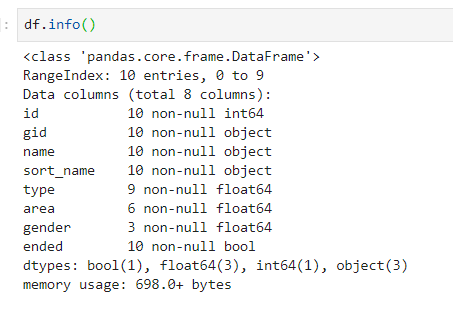


Figure 1 - MusicBrainz Artists dataset info

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| # | Name | Data Type | Column Description | References |
| 1 | Id | Int | Sequential Id in MusicBrainzDB |  |
| 2 | Gid | Guid | Unique Id of the artists |  |
| 3 | Name | String | Full name |  |
| 4 | Sort\_Name | String | Normalized name of t he artists. In some cases the artist name is reversed. (ie. Lewis Capaldi => capaldi\_lewis) |  |
| 5 | Type | Int | Artist\_type foreign key: 1. Person 2.Group | Artists\_Type table |
| 6 | Area | Int | Area foreign key | Area table |
| 7 | Gender | Int | Gender foreign key | Genders table |
| 8 | Ended | Bit | Active artist flag |  |

Figure 2 - Artists columns description - references

This dataset contains 1,610,535 records of artists which 1,514,086 of them are not active and there are 96,449 active artists which can be used for analysis.

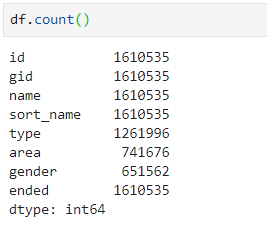


Figure 3 - Artists properties count

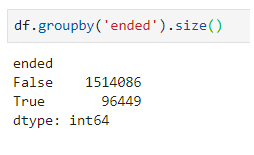


Figure 4 - Active artists statistics

The properties are cleaned and transformed into a new temporary table to then used as artists dimension table in the target data warehouse.

# ETL Process

## Extract

Postgres dump downloaded and loaded into PostgreSQL.

## Transform

Python utilized for transformation process. Data loaded into a Dataframe from PostgreSQL using sqlalchemy library. The datafame is then cleansed and transformed using pandas. The steps are as the following:

1. Searching for Null or Empty Values
   1. Gid (mbid)
   2. Name
   3. Type
   4. Area
   5. Gender
   6. Ended
2. Removing records with Null records in columns
   1. Gid
   2. Name
   3. Ended
3. Replacing Null records with default values
   1. Type
   2. Gender
   3. Area
4. Duplicate artists identified and removed
   1. Exact duplicate record
   2. Duplicate records related to bad data in terms of different collations removed
   3. Duplicate records related to bad data and naming standards removed.

I.e. Lewis Capaldi = > Capaldi Lewis

* 1. Similar records calculated by levenshtein algorithm. Those identified as duplicates removed.

I.e. wizkid => WizKid , Wizkid



## Load

Cleaned data (dataframe) is loaded into a staging database in PostgreSQL for future usage.

# References

<https://musicbrainz.org/doc/MusicBrainz_Database/Download>