# Information Visualization

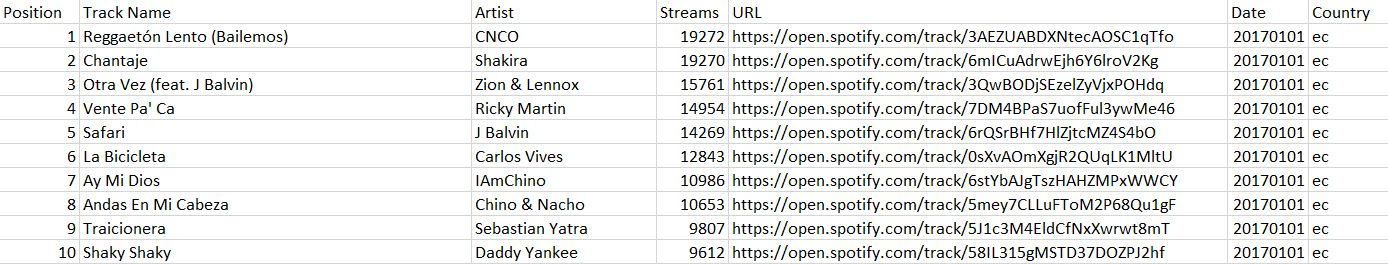
# CHECKPOINT II: Data cleaning and processing

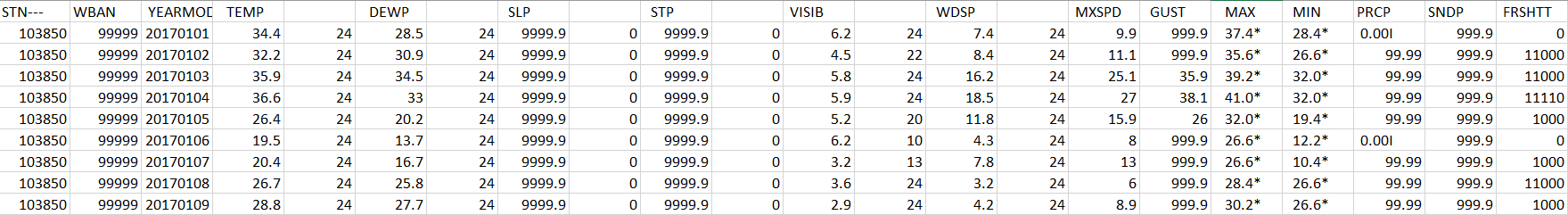
G17 - A

**1. Initial Dataset**

Our initial dataset consisted of two different sets of .csv files:

* **spotify.csv**, which corresponds to the top 200 songs streamed each day on Spotify, for each of the 53 countries it available in, from 2017/01/01 to 2018/01/09, with 369MB and 4.028.400 rows;
* 53 different **xx.csv** files, which corresponds to the weather data in each of the 53 countries Spotify is available in, whereas “xx” matches the ISO 3166-1 country code (*de.csv* corresponds to Germany’s weather data, *pt.csv* corresponds to Portugal’s weather data, etc.), with a total of 2MB and 19.345 rows.

The following images represent examples of each of the sets above described:

*spotify.csvde.csv*

**2. Selected/Derived Data**

We have refined the original dataset into the following list of tables:

* **weather.csv** was generated from all the 53 xx.csv files (xx = country code, like pt — Portugal, es — Spain…), and contains all the processed information about the weather conditions in each country in 2017, with 13.5MB and 19.345 rows;
* **processed\_spotify.csv** was generated from spotify.csv and contains mostly the same information, but processed in order to uniformize data, with 151MB and 933.607 rows;
* **full\_dataset.csv** was generated from weather.csv and spotify.csv and contains all the processed information about the weather conditions and the most streamed songs in each day, with 178MB and 933.607 rows;
* **songs\_temp.csv** was derived from full\_dataset.csv and contains all the information about the songs, and it is sorted by streams and temperature, with 7.57MB and 49.798 rows.
* **sunny.csv** was derived from full\_dataset.csv and contains all the information about the songs when it’s sunny, and it is sorted by number of streams, with 18KB and 114 rows.
* **raining.csv** was derived from full\_dataset.csv and contains all the information about the songs when it’s raining, and it is sorted by number of streams, with 205KB and 7.235 rows.
* **indicators.csv** was derived from full\_dataset.csv and contains all the information about the listening habits according to the weather conditions, and it is sorted by number of streams, with 11MB and 67.438 rows.
* **streams\_conditions.csv** was derived from full\_dataset.csv and contains all the information regarding the number of streams of each country according to its weather conditions, and it is sorted by number of streams, with 10KB and 389 rows.
  + - 1. **3. Data abstraction**

|  |  |  |  |
| --- | --- | --- | --- |
| * + - 1. Attribute | * + - 1. Tables where it appears | * + - 1. Type | * + - 1. Semantics |
| * + - 1. Date | * + - 1. weather.csv, processed\_spotify.csv, full\_dataset.csv, songs\_temp.csv | * + - 1. Hierarchical | * + - 1. The date in the format YYYYMMDD, corresponding to the day measured. |
| * + - 1. Temperature | * + - 1. weather.csv, full\_dataset.csv, songs\_temp.csv | * + - 1. Continuous | * + - 1. The mean temperature (in Celsius) for the day measured. |
| * + - 1. Visibility | * + - 1. weather.csv, full\_dataset.csv | * + - 1. Continuous | The mean visibility (in kilometres) for the day measured. |
| * + - 1. Windspeed | weather.csv, full\_dataset.csv | * + - 1. Ratio | * + - 1. The mean wind speed (in kilometres per hour) for the day measured. |
| * + - 1. Total Precipitation | * + - 1. weather.csv, full\_dataset.csv | * + - 1. Continuous | * + - 1. The total precipitation reported in the day (in centimetres). |
| * + - 1. Fog | * + - 1. weather.csv, full\_dataset.csv | * + - 1. Ordinal | * + - 1. An indicator for fog — if its 1, that day had fog; if its 0, it didn’t. |
| * + - 1. Rain | * + - 1. weather.csv, full\_dataset.csv | * + - 1. Ordinal | * + - 1. An indicator for rain — if its 1, it rained that day; if its 0, it didn’t. |
| * + - 1. Snow | * + - 1. weather.csv, full\_dataset.csv | * + - 1. Ordinal | * + - 1. An indicator for snow — if its 1, it snowed that day; if its 0, it didn’t. |
| * + - 1. Hail | * + - 1. weather.csv, full\_dataset.csv | * + - 1. Ordinal | * + - 1. An indicator for hail — if its 1, it hailed that day; if its 0, it didn’t. |
| * + - 1. Thunder | * + - 1. weather.csv, full\_dataset.csv | * + - 1. Ordinal | * + - 1. An indicator for thunder — if its 1, there were thunders that day; if it’s 0, there weren’t. |
| * + - 1. Tornado | * + - 1. weather.csv, full\_dataset.csv | * + - 1. Ordinal | * + - 1. An indicator for tornado — if its 1, there were tornados that day; if it’s 0, there weren’t. |
| * + - 1. Country | * + - 1. weather.csv, processed\_spotify.csv, full\_dataset.csv, songs\_temp.csv | * + - 1. Nominal | * + - 1. The country associated to the weather and music data. |
| * + - 1. Track name | processed\_spotify.csv, full\_dataset.csv, songs\_temp.csv | * + - 1. Nominal | The name of the song. |
| * + - 1. Artist | * + - 1. processed\_spotify.csv, full\_dataset.csv, songs\_temp.csv | * + - 1. Nominal | * + - 1. The song’s artist. |
| * + - 1. Streams | * + - 1. processed\_spotify.csv, full\_dataset.csv, songs\_temp.csv | * + - 1. Continuous | * + - 1. The total number of streams on Spotify, on a given day, on a given country. |
| * + - 1. URL | * + - 1. processed\_spotify.csv, full\_dataset.csv, songs\_temp.csv | * + - 1. Nominal | * + - 1. The URL link to directly play a given song on Spotify. |

**4. Dataset processing**

Given that we had 53 .csv flies corresponding to the weather conditions of each of the 53 countries where Spotify is available in, the first step to accomplish was to add a column in each file, with its corresponding country code, in order to make work easier when merging with the music dataset.

After that, we merged all the 53 files into 1 single weather.csv file, containing all the weather data in just 1 file. To wrap up the weather dataset, we removed the columns that had extra data we don’t consider relevant to our visualization, such has dew point, the station ID, sea level pressure, among others. We also had to split the indicator column, in order to access each digit directly. The last step was to merge the music dataset with the weather dataset. Firstly, we reduced the Spotify’s dataset size by ¾, by considering only the top 50 songs, instead of the top 200While processing this, we found out some problems regarding with the music dataset. It turns out Spotify only ranks tracks on its daily song ranking if it has at least 1,000 streams in that given day. That being said, in smaller countries where Spotify might not be broadly use, like Luxemburg, where it’s hard for songs to have that number of streams, there might be some unavailable data. We also found out that Spotify’s API was down during 3 days in 2017, resulting in the same consequence.

* + - 1. **5. Mapping (Data sample / Questions)**
* On a sunny day, which song is the most listened worldwide?
  + This can be answered by checking the first line in *sunny.csv* as the most listened songs on a sunny day are on top of this table.
* If it’s raining, what artists do people listen the most in Ecuador?
  + *raining.csv* is grouped by countries and number of streams. As such, the most listened artists of each country on a raining day are on top of each country section of this table.
* In what weather conditions is “*Despacito”* most likely to be heard?
  + *indicators.csv* is grouped by song title and number of streams. As such, the weather conditions of each song can be checked by looking at each one of its weather attributes (fog, rain, snow, hail, thunder, tornado).
* In Finland, do people listen to more music if it’s raining or snowing?
  + *streams\_conditions.csv* is grouped by countries and weather attributes (fog, rain, snow, hail, thunder, tornado) It is ordered by number of streams. As such, the weather condition that has the highest number of streams is top of each country section of this table.