

# 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
- 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.)

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

Position	Track Name	Artist	Streams	URL	Date	Country
1	Reggaetón Lento (Bailemos)	CNCO	19272	<a href="https://open.spotify.com/track/3AEZUABDXNtecAOSC1qTfo">https://open.spotify.com/track/3AEZUABDXNtecAOSC1qTfo</a>	20170101	ec
2	Chantaje	Shakira	19270	<a href="https://open.spotify.com/track/6mICuAdrwEjh6Y6lroV2Kg">https://open.spotify.com/track/6mICuAdrwEjh6Y6lroV2Kg</a>	20170101	ec
3	Otra Vez (feat. J Balvin)	Zion & Lennox	15761	<a href="https://open.spotify.com/track/3QwBODjSEzelZyVjxPOHdq">https://open.spotify.com/track/3QwBODjSEzelZyVjxPOHdq</a>	20170101	ec
4	Vente Pa' Ca	Ricky Martin	14954	<a href="https://open.spotify.com/track/7DM4BPas7uofFul3ywMe46">https://open.spotify.com/track/7DM4BPas7uofFul3ywMe46</a>	20170101	ec
5	Safari	J Balvin	14269	<a href="https://open.spotify.com/track/6rQsrBHf7HIZtcMZ4S4bO">https://open.spotify.com/track/6rQsrBHf7HIZtcMZ4S4bO</a>	20170101	ec
6	La Bicicleta	Carlos Vives	12843	<a href="https://open.spotify.com/track/0sXvAOMXgjR2QUqLK1MltU">https://open.spotify.com/track/0sXvAOMXgjR2QUqLK1MltU</a>	20170101	ec
7	Ay Mi Dios	IAmChino	10986	<a href="https://open.spotify.com/track/6stYbAJgTszHAHZMPxWWCY">https://open.spotify.com/track/6stYbAJgTszHAHZMPxWWCY</a>	20170101	ec
8	Andas En Mi Cabeza	Chino & Nacho	10653	<a href="https://open.spotify.com/track/5mey7CLLuFTom2P68Qu1gF">https://open.spotify.com/track/5mey7CLLuFTom2P68Qu1gF</a>	20170101	ec
9	Traicionera	Sebastian Yatra	9807	<a href="https://open.spotify.com/track/5J1c3M4EldCfNxxWrwrt8mT">https://open.spotify.com/track/5J1c3M4EldCfNxxWrwrt8mT</a>	20170101	ec
10	Shaky Shaky	Daddy Yankee	9612	<a href="https://open.spotify.com/track/58IL315gMSTD37DOZPJ2hf">https://open.spotify.com/track/58IL315gMSTD37DOZPJ2hf</a>	20170101	ec

*spotify.csv*

STN---	WBAN	YEARMO	TEMP		DEWP		SLP		STP		VISIB		WDSP		MXSPD	GUST	MAX	MIN	PRCP	SNDP	FRSHTT
103850	99999	20170101	34.4	24	28.5	24	9999.9	0	9999.9	0	6.2	24	7.4	24	9.9	999.9	37.4*	28.4*	0.001	999.9	0
103850	99999	20170102	32.2	24	30.9	24	9999.9	0	9999.9	0	4.5	22	8.4	24	11.1	999.9	35.6*	26.6*	99.99	999.9	11000
103850	99999	20170103	35.9	24	34.5	24	9999.9	0	9999.9	0	5.8	24	16.2	24	25.1	35.9	39.2*	32.0*	99.99	999.9	11000
103850	99999	20170104	36.6	24	33	24	9999.9	0	9999.9	0	5.9	24	18.5	24	27	38.1	41.0*	32.0*	99.99	999.9	11110
103850	99999	20170105	26.4	24	20.2	24	9999.9	0	9999.9	0	5.2	20	11.8	24	15.9	26	32.0*	19.4*	99.99	999.9	1000
103850	99999	20170106	19.5	24	13.7	24	9999.9	0	9999.9	0	6.2	10	4.3	24	8	999.9	26.6*	12.2*	0.001	999.9	0
103850	99999	20170107	20.4	24	16.7	24	9999.9	0	9999.9	0	3.2	13	7.8	24	13	999.9	26.6*	10.4*	99.99	999.9	1000
103850	99999	20170108	26.7	24	25.8	24	9999.9	0	9999.9	0	3.6	24	3.2	24	6	999.9	28.4*	26.6*	99.99	999.9	11000
103850	99999	20170109	28.8	24	27.7	24	9999.9	0	9999.9	0	2.9	24	4.2	24	8.9	999.9	30.2*	26.6*	99.99	999.9	1000

*de.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.

### 3. Data abstraction

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

### 4. Dataset processing

Given that we had 53 .csv files 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 as 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  $\frac{3}{4}$ , by considering only the top 50 songs, instead of the top 200. While 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.

### 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\_top.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?
  - This can be answered by checking the first lines in raining\_ecuador.csv as the most listened artists on a raining day on Ecuador are on top of this table
- In what weather conditions is "Despacito" most likely to be heard?
  - By checking the *despacito\_indicators.csv*, the first lines correspond to the weather conditions where people listened more to Despacito
- Between Portugal (winter) and Australia (summer), where was "All I Want For Christmas Is You" most streamed during Christmas?
  - This can be checked by analyzing the *chistmas\_eve.csv* file.
- How likely is "Let It Snow" to be streamed during snow days?
  - This can be checked by analyzing the *letitsnow.csv* file.