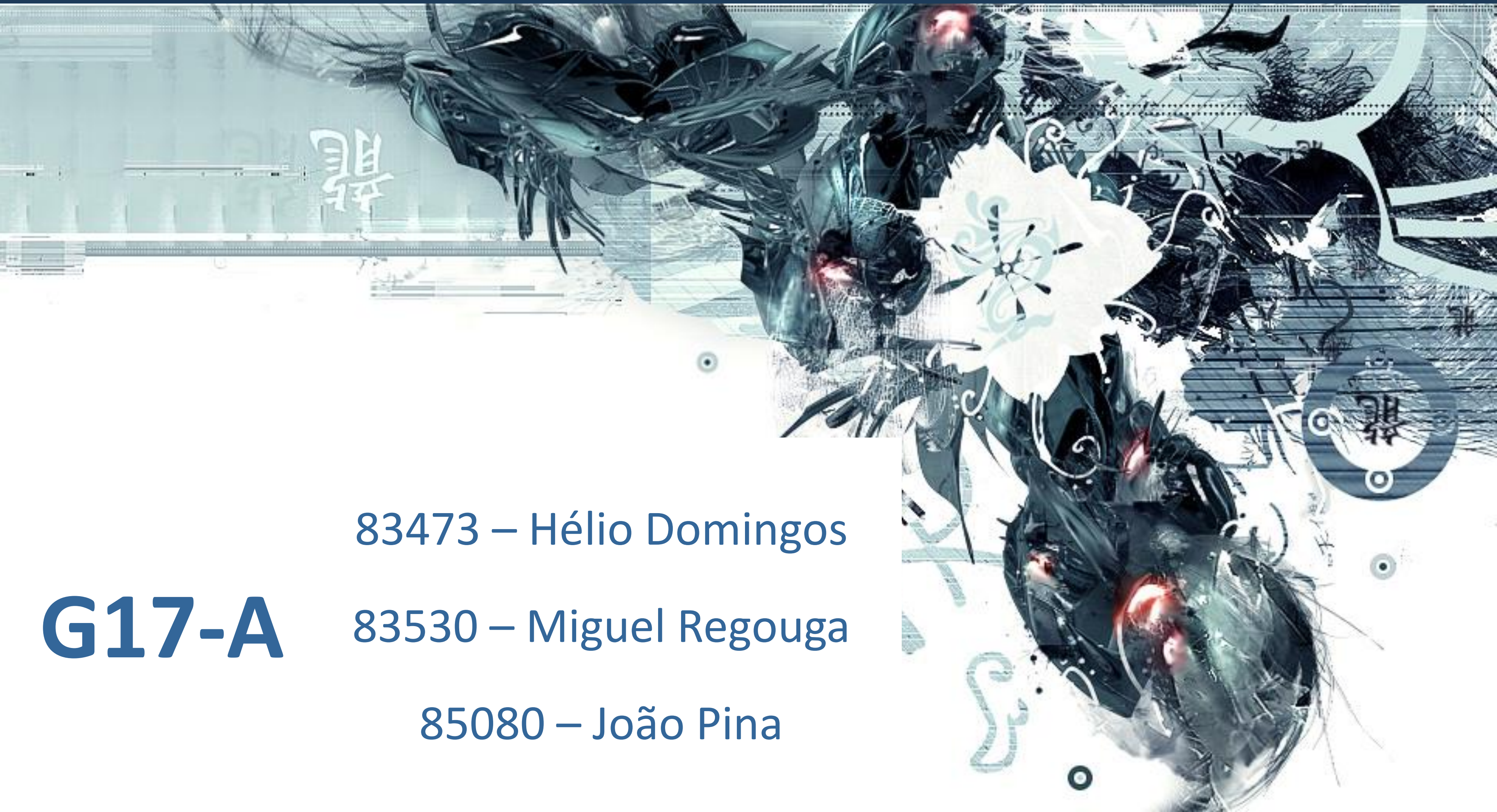
Information Visualization
Project Proposal and Dataset



83473 – Hélio Domingos

83530 – Miguel Regouga

85080 – João Pina

G17-A

01

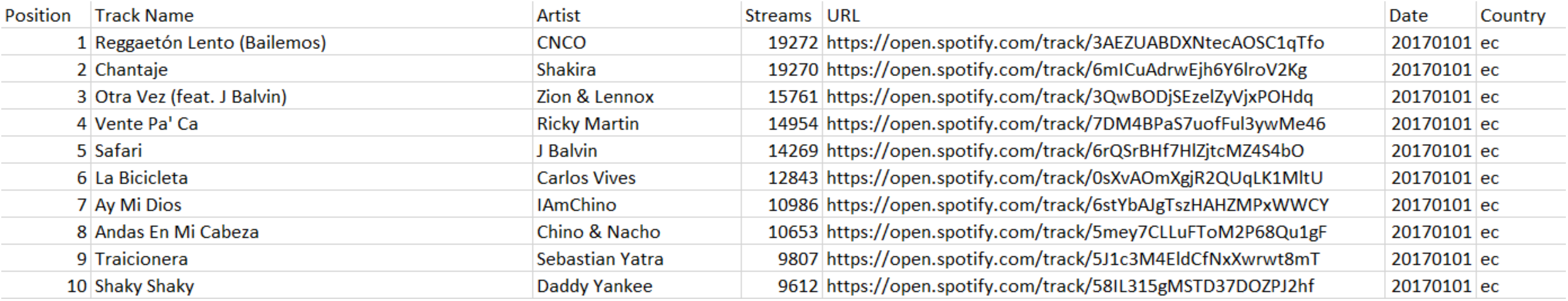
INITIAL DATASET

Initial Dataset

Description — Spotify Dataset

- *spotify.csv*
- Top 200 songs streamed each day, for each of the 53 countries Spotify is available in, from 2017/01/01 to 2018/01/09

Initial Dataset



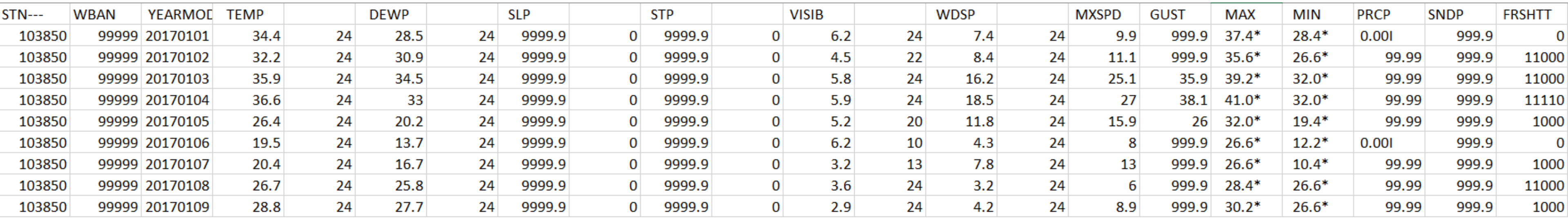
Position	Track Name	Artist	Streams	URL	Date	Country
1	Reggaetón Lento (Bailemos)	CNCO	19272	https://open.spotify.com/track/3AEZUABDXNtecAOSC1qTfo	20170101	ec
2	Chantaje	Shakira	19270	https://open.spotify.com/track/6mICuAdrwEjh6Y6lroV2Kg	20170101	ec
3	Otra Vez (feat. J Balvin)	Zion & Lennox	15761	https://open.spotify.com/track/3QwBODjSEzelZyVjxPOHdq	20170101	ec
4	Vente Pa' Ca	Ricky Martin	14954	https://open.spotify.com/track/7DM4BPaS7uofFul3ywMe46	20170101	ec
5	Safari	J Balvin	14269	https://open.spotify.com/track/6rQSrBHf7HlZjtcMZ4S4bO	20170101	ec
6	La Bicicleta	Carlos Vives	12843	https://open.spotify.com/track/0sXvAOmXgjR2QUqLK1MltU	20170101	ec
7	Ay Mi Dios	IAmChino	10986	https://open.spotify.com/track/6stYbAJgTszHAHZMPxWWCY	20170101	ec
8	Andas En Mi Cabeza	Chino & Nacho	10653	https://open.spotify.com/track/5mey7CLLuFToM2P68Qu1gF	20170101	ec
9	Traicionera	Sebastian Yatra	9807	https://open.spotify.com/track/5J1c3M4EldCfNxXwrwt8mT	20170101	ec
10	Shaky Shaky	Daddy Yankee	9612	https://open.spotify.com/track/58IL315gMSTD37DOZPJ2hf	20170101	ec

Initial Dataset

Description — Weather Dataset

- 53 *xx.csv* files
- xx = country code (pt, es, fr, it, …)
- 2017 daily weather conditions for the most populous city of each of the 53 countries Spotify is available in

Initial Dataset



STN---	WBAN	YEARMOD	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*	0.00I	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*	0.00I	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

02

SELECTED / DERIVED DATA

Derived data

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;

processed_spotify.csv

was generated from spotify.csv and contains mostly the same information, but processed in order to uniformize data;

Derived data

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;

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.

03

DATA ABSTRACTION

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.
Wind Speed	weather.csv, full_dataset.csv	Ordinal	The mean wind speed (in kilometres per hour) for the day measured.
Precipitation	weather.csv, full_dataset.csv	Ordinal	The total precipitation reported in the day (in centimetres).

Data abstraction

Attribute	Tables where it appears	Type	Semantics
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.

Data abstraction

Attribute	Tables where it appears	Type	Semantics
PT 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.

04

DATASET PROCESSING

Dataset processing

Dataset cleaning description

Weather dataset

- Removed excessive data not needed to our visualization (sea level pressure, dew point, station ID…)
- Added country code
- Merged all 53 .csv files into one
- Created a column for each indicator

Dataset processing

Dataset cleaning description

Spotify dataset

- Normalized the date format (to match the weather dataset)
- Reduced to top 50 (instead of top 200)
- Merged with weather dataset

Dataset processing

Problems found:

Spotify only ranks if streams > 1000

- Although we reduced from top 200 to top 50, there are a few smaller countries (such as Luxembourg) that don’t have as many data available as other countries. In the minimum, top 15

Spotify’s API was down during 3 days

- Removed the lines corresponding to those 3 days, since there is no data available

Dataset processing

Problems found:

Songs removed from Spotify

- In a very rare case, we found out that a song which was on the top 50 was removed from the platform - so did we from our dataset

Weather indicators

- Our weather dataset had a number with 6 digits, corresponding to various weather factors (rain, snow, fog…), but we couldn’t access them directly – we created a new column for each digit

05

MAPPING

Mapping

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

Mapping

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

Mapping

How likely is “*Let It Snow*” to be streamed during snow days?

This can be checked by analyzing the *letitsnow.csv* file