ETL (Extract, Transform, and Load) project proposal
Group 9

Main topic: Different recycling rates in NYC in 2017

EXTRACT
Databases:

- DSNY Waste Characterization: Citywide Subsort (2017)
  - Relational database
  - Source: https://data.cityofnewyork.us/Environment/DSNY-Waste-Characterization-Citywide-Subsort/phkb-tkts
  - License: Public domain
  - Size: 1832 rows, 9 columns

- DSNY Waste Characterization: Mainsort (2017)
  - Relational database
  - Source: https://data.cityofnewyork.us/City-Government/DSNY-Waste-Characterization-Mainsort/k3ks-jzek
  - License: Public domain
  - Size: 560 rows, 9 columns

- Recycling Diversion and Capture Rates
  - Relational database
  - Source: https://www.kaggle.com/new-york-city/nyc-recycling-diversion-and-capture-rates
  - License: Public domain
  - Size: 2832 rows, 9 columns

- DSNY Frequencies
  - Relational database
  - Source: https://data.cityofnewyork.us/City-Government/DSNY-Frequencies/rv63-53db
  - License: Public domain
  - Size: 610 rows, 13 columns

All the databases will be extracted using different methods that will be defined in the construction process (APIs with JSON response or reading CSV files).

TRANSFORM
Goal: To combine the databases for getting all the possible information about recycling per material (paper, organic, MGP, C&D, special waste, E-Waste, others) in 2017. This information could include: total aggregate percent and total refuse percent in the Subsort and Mainsort, capture rates, collection, and frequency dates. In the following figures, it is possible to see the expected result and how information is going to be taken from each dataset.

Initial databases columns:

Table: DSNY Waste Characterization: Citywide Subsort (2017)	**Table:** DSNY Waste Characterization: Mainsort (2017)	**Table:** Recycling Diversion and Capture Rates	**Table:** DSNY Frequencies
Material	Material	Zone	DISTRICT
Aggregate Percent	Aggregate Percent	District	FID
Refuse Percent	Refuse Percent	Fiscal Month Number	FREQUENCY
MGP Percent	MGP Percent	Fiscal Year	FREQ_BULK
Paper Percent	Paper Percent	Month Name	FREQ_ORGANICS
Organic Percent	Organic Percent	Diversion Rate-Total (Total Recycling / Total Waste)	FREQ_RECYCLING
Material Group	Material Group	Capture Rate-Paper (Total Paper / Max Paper)	FREQ_REFUSE
DSNY Diversion Summary Category	DSNY Diversion Summary Category	Capture Rate-MGP (Total MGP / Max MGP)	GlobalID
Location	Location	Capture Rate-Total ((Total Recycling - Leaves (Recycling)) / (Max Paper + Max MGP))x100	SCHEDULECODE
			SECTION
			SHAPE_Area
			SHAPE_Length
			multipolygon

Expected result:

Materials database
Total Aggregate Percent Subsort
Total Refuse Percent Subsort
Total Aggregate Percent Mainsort
Total Refuse Percent Mainsort
Capture rate
Collection dates
Frecuency

The transformation will include:

- Filtering: 2 of the databases have information only related to 2017, which means that it will be necessary to filter the other 2 databases for this year. Also, filtering by type of material will be necessary.
- Selection: It will be necessary to select and group all the types of material in new categories because not all the databases have the same categories.
- Joining: creating a final database with tables per material.
- Other processes will be identified with the summarization of the information and some cleaning.

LOAD

According to the data analysis, it will be possible to identify if the final database will be relational or non-relational for each material in 2017. For now, what is expected is to get a table or collection per material, the group of all tables or collections will create a loadable database for future use.