ETL Proposal

Mika Nagamine & Jessica Nugent

Goal :

Use greenhouse gas emissions data and climate data, and find the relationships between the amount of the emission and changes in weather. Specifically, we'll be using the data from New York City and develop a data file that allows us to examine whether there are any relationships between the amount of the gas emitted and the temperatures of the city from 1990 to 2015.

The sources of data that you will extract from :

- Historical average weather data for New York City: https://www.weather.gov/media/okx/Climate/CentralPark/monthlyannualtemp.pdf (found on this site: https://www.weather.gov/okx/CentralParkHistorical)
    - Because this is a PDF, we copied and pasted the data into an Excel file to analyze.
- Greenhouse gas emissions data for New York State: https://data.world/data-ny-gov/djfn-trk4
- (We also found pollution data for the U.S.: https://data.world/garyhoov/declining-pollution.)

The type of transformation needed for this data (cleaning, joining, filtering, aggregating, etc) :

- We will need to clean the data and remove any extra information from our table.
- The historical average weather data covers a larger time frame than the emissions data, so we will filter out the extra years of temperature data.
- We will join the emissions and temperature data on year to see the relationship between the amount of emissions and changes in the weather.

The type of final production database to load the data into (relational or non-relational).

- We will produce a relational database.

The final tables or collections that will be used in the production database :

- We will have one table for the historical average temperature by year.
- We will have at least one other table for the greenhouse gas emissions by year.