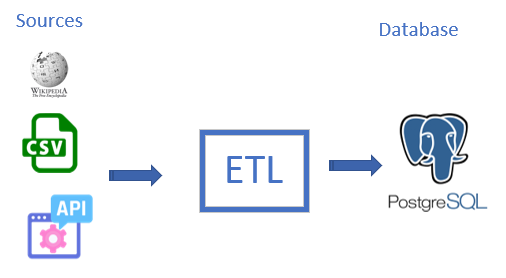
Project name: ETL for different sources and files

Project Proposal

This project will analyze crimes in cities in America that have almost the same Toronto’s population. In this case, we chose some data set from Toronto, Vancouver, and Chicago.

Workflow



**Extract**

The data sets were extracted from different sources, for instance:

Wikipedia: This was a source where we found Toronto census from 2001 to 2016

<https://en.wikipedia.org/wiki/Demographics_of_Toronto#cite_note-2011censuspop-18>

CSV: Two CSV files were downloaded from these links that track crime from Chicago (USA) and Vancouver (Canada) cities:

<https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-present/ijzp-q8t2/data>

https://data.vancouver.ca/datacatalogue/crime-data.htm

API return JSON: This an API created by me (Manoel) which can be viewed with more detail here. Moreover, this API returns some main crimes in Toronto from 2004 to 2018

http://manoelburgos.azurewebsites.net/api/Crimes

**Transform**

HTML -> Database

This image below was a table found in an HTML page (Wiki) which was cleansing to extract data from the correct columns and rows. Furthermore, the values found inside this HTML was cast to float data type, then they could be inserted to database.



JSON -> Database

We iterate with the necessary keys to insert into database. Moreover, the key called Orcurrence\_Date has values in this format “2004-01-01T05:00:00:000Z”. For this reason, all values of this key were split in day, month and year inside the python code before inserted into database.



CSV -> Database

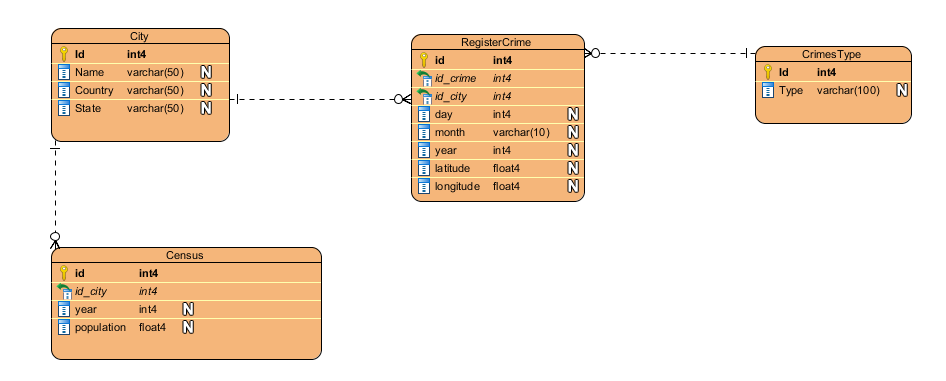
The number of lines of each CSV has more than 200k rows.



**Load**

At the beginning of this project, I designed this ERD below to retrieve our data set.

It was designed to follow the 1st,2nd,3rd normal forms.



**Results**

Based on the results loaded in our database, we can do some analysis, for instance, discover how many crimes happened in a specific year in each city.

Below is a sample, however, it is possible to see more queries here:

select c.name,

c.country,

rg.year,

cr.type,

count(cr.type)

from city c

inner join registercrime rg

on c.id = rg.id\_city

inner join crimestype cr

on rg.id\_crime = cr.id

group by c.name,

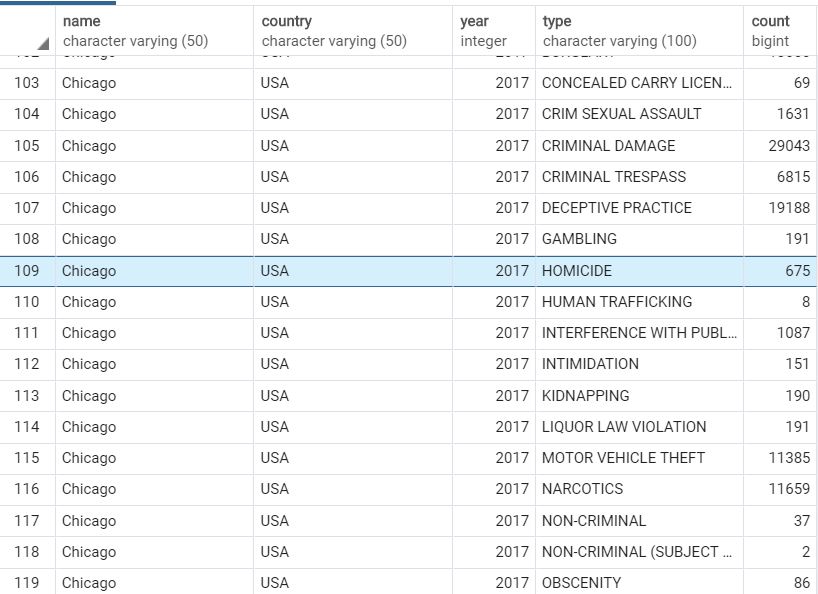
c.country,

rg.year,

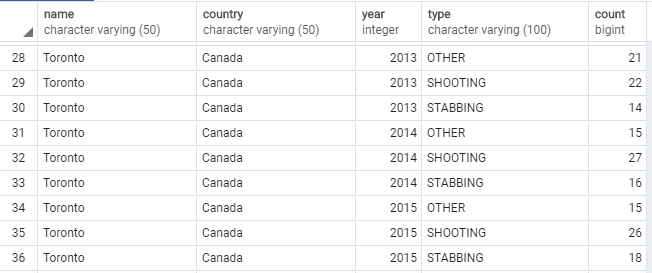
cr.type

order by rg.year

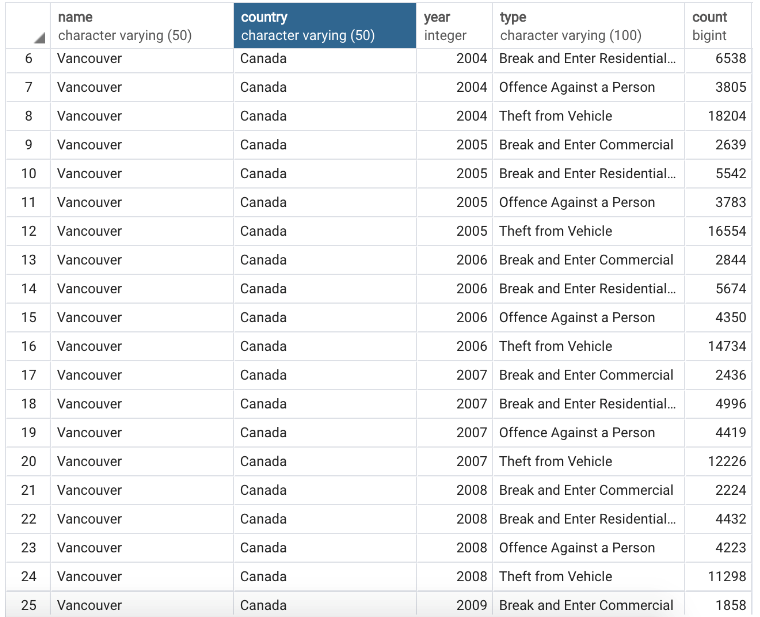
Chicago:



Toronto:



Vancouver:



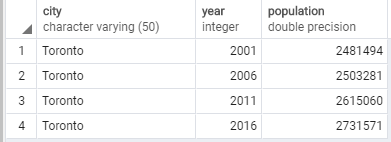
select (select name

from city where id = c.id\_city) as City,

c.year,

c.population

from census c;



Features:

PostgreSQL database

ERD designed in Visual Paradigm

Python with libraries:

Pandas,

Sqlalchemy

Json,

Requests

By:

Manoel,

Mir,

Banafshe