Data Analytics Bootcamp 12 June 12, 2021

Tecnologico de Monterrey

Team Members:

Gustavo Maldonado

Juan Cortez

Marcelo Flores

Carlos Gomez

ETL- Project Report

Housing loans Analysis for Mexican Municipalities.

**Project Proposal**

The objective is to analyze housing credit loans that have general data from the applicants and combine it with housing and spending data from municipalities and states in Mexico to get better insigh

**ETL Process**

To achieve the objective for the project the following steps were taken:

**Extraction**

1. Data was acquired from two different CSV sources: 1) from INEGI (Instituto Nacional de Estadística y Geografía) and 2)SNIIV (Sistema Nacional de Información e Indicadores de Vivienda).

Sources:[[1]](#footnote-1)

https://www.inegi.org.mx/programas/ccpv/2020/default.html#Datos\_abiertos

https://sniiv.conavi.gob.mx/reportes/datos\_abiertos.aspx

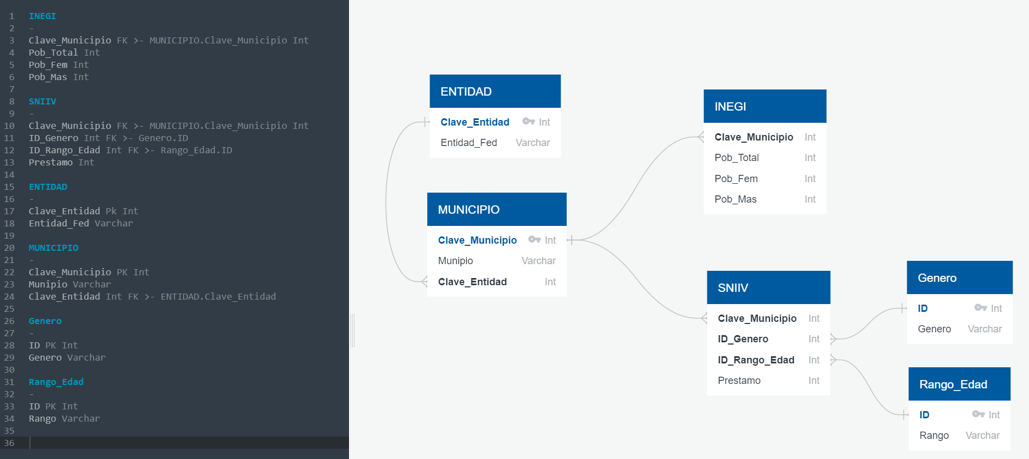
INEGI Data set contains 232 columns x 195,663 rows. This Dataset contains housing and population data from the 2020 census.

SNIIV Data sets contains 16 columns x 355,375 rows. This Dataset contains financing and loans awarded for housing actions (Acquisition and maintenance of homes, payment of mortgages, etc.).

1. Concatenate elements from both data sets were chosen for their similarities
   1. “Clave de Entidad”
   2. “Clave de Municipio”

**Transformation**

1. Creation of a Schema for the relational database utilizing app QuickDBD.



1. Data was transformed for readability using Jupyter Notebook, in the following order:
   * + Decide on the data model.
     + Load Datasets to pandas as Data Frames.
     + Filter columns for analysis.
     + Change and match the names of columns.
     + Normalize data by creating different tables from the original datasets.
     + The following books were created within Jupyter notebooks:
2. Entidad\_table
3. municipality\_table
4. end\_inegi\_df
5. sniiv\_final\_df
6. gender\_df
7. rango\_edad\_df

* Removal of rows with summary data by municipality.

**Load**

The team determined the best direction to move towards was to use a relational database to maintain integrity of the information provided in the initial datasets, avoiding changes that could affect the database, when new information is loaded or changed.

1. Create Database on pgAdmin.
2. Run schema file (queries.sql) to add tables and constraints.
3. Connect Jupyter notebook to SQL databases.
4. Loaded information into database.

1. Link the files can be found in ReadMe.md File. [↑](#footnote-ref-1)