Tito KOH, Kimmeng HONG, Iwan DEROUET, Adel Nève

Dataset for Real Estate Property Value Analysis in France

Group 4

Dataset: https://files.data.gouv.fr/geo-dvf/latest/csv/2023/

Overview: The dataset, sourced from the French government's open data platform, details geolocalized property transaction values across France for the year 2023. This comprehensive dataset includes 3,799,407 observations and 40 variables, capturing a variety of information related to property characteristics, location, and transaction details.

Key Attributes:

- **Transaction Details:** Includes variables such as `id_mutation` (unique transaction ID), `date_mutation` (date of transaction), `nature_mutation` (type of transaction), and `valeur_fonciere` (property value, our target variable).
- **Location Information:** Features detailed geolocation information through `code_postal` (postal code), `code_commune` and `nom_commune` (INSEE commune code and name), as well as `longitude` and `latitude` of the property.
- **Property Specifications:** Variables include built area (`surface_reelle_bati`), number of rooms (`nombre_pieces_principales`), and land area (`surface_terrain`). The dataset also contains details of individual lots (`lot_1_surface_carrez` through `lot_5_surface_carrez`), and property type information (`type_local`).

Data Preprocessing:

Since there are a lot of missing values in the dataset, we decided to drop columns that have more than 50% of missing values and then we started to drop row by row if there's a missing value in one of the columns. After dropping all the missing values and data duplication, we have 893234 rows of observation and 22 columns.
The dataset being too big, we intend on taking 1000 random observations to run our project.

Analysis Objective:

The primary aim of this project is to analyze the factors that influence property values (`valeur_fonciere`) in France using PCA. By reducing dimensionality, we seek to identify the most significant underlying factors among the property's attributes and location data that explain the variability in property values.

The target variable is `valeur_fonciere` which is the price of the land. For the potential covariables are `surface_terrain`, `longitude` and `latitude`.

This dataset provides a rich foundation for dimensionality reduction techniques, offering insights into the main components that impact property values in France, facilitating efficient analysis and potential future modeling.