A study about the French counties and house prices

1. Contents

2. Introduction

Introduction where you discuss the business problem and who would be interested in this project.

Disclaimer:

I made this project to practice my coding skills and the data science tools I have learned from *IBM Data Science Professional Certificate* class and other various tutorials. The results of this study should be taken as mere insight.

Context:

This year, the COVID-19 pandemic hit France severely and induced a national lockdown of more than 2 months in the country. A big proportion of France population live in apartments, and since the lockdown, many are looking to move into actual houses with maybe a garden. And some people even started to work remotely and are willing to relocated in places where house prices are more affordable.

Project goal:

This project is meant to answer this question:

Where is a "good place" to buy a house in France and at what price?

To answer this question, first we need to define "good place". For this study, we will define "good" with those criteria:

- Population density
- Population median age
- Population median income
- Real estate price
- Nightlife

We will play with those criteria to filter places in France and pick a town or a city.

Secondly, we will estimate more precisely how much will a house cost in a specific town/city according to those factors:

- The size of the house
- The amount of rooms
- The land area

The project may interest:

- People who are willing to relocate and buy a house
- Real estate agencies
- Real estate developers

3. About the data

Data where you describe the data that will be used to solve the problem and the source of the data.

The data was collected on the internet, mainly on France's government website.

The French territory is subdivided in many smaller administrative areas. The smallest administrative area is called a *Commune*, it is the English equivalent of a county. A c*ommune* consists on average of fifteen square kilometres.

The data obtained from the French government website's is at the precision of *communes*.

All the data obtained is in French, therefore here is the translation of some the terms encountered:

Commune: French word for county

Code INSEE: County's Id as per the French government's statistics department (INSEE)

Revenue: Household income

Valeur foncière: Real estate value

Vente: Sale

Maison: House

Surface Carrez: Living area

Surface terrain: Land area

Nombre de pièces principales: Number of main rooms

Here are the data used for this project with some details about how they were used in this project.

Commune's geographical data of 2013

- Source : https://www.data.gouv.fr/en/datasets/correspondance-code-insee-code-postal/
- File name: correspondance-code-insee-code-postal.csv
- Provider: Région Île-de-France
- Number of rows: 36 742
- Contains: Geographical information for each Code INSEE (county Id)
- Main feature used:
    - *Communes'* names
    - Latitudes and longitudes
    - Communes' surface
- Used for:
    - Locating the *Communes*
    - Estimating the *Communes* population density
- Pre-processing:
    - A quick data-pre-processing with allowed us to extract the information needed

- Notebook: https://nbviewer.jupyter.org/github/Ashish-3/House-prices-in-France/blob/master/Preprocessing_communes_stats.ipynb

French population age census of 2016

- Source : https://www.data.gouv.fr/fr/datasets/population-municipale-t-popmun-com/
- File name: t_popmun_com.csv
- Provider: Atlasanté
- Number of rows: 7 million (approximately)
- Contains: One row for each combination of age, sex, and *Commune*
- Main feature used:
    - Age count per sex per *Commune*
    - *Code INSEE*
- Used for estimating in each *Commune*
    - The population
    - The density of population
    - Statistics about the age (median age for instance)
- Pre-processing:
    - A quick data-pre-processing with some sums allowed us to extract the information needed
- Notebook: https://nbviewer.jupyter.org/github/Ashish-3/House-prices-in-France/blob/master/Preprocessing_communes_stats.ipynb

2017 French household incomes

- Source : https://www.insee.fr/fr/statistiques/4507225?sommaire=4507229
- File name: cc_filosofi_2017_COM.CSV
- Provider: INSEE –Filosofi
- Number of rows: 34 931
- Contains: One row per *Commune* that describes households' incomes
- Main feature used:
    - *Revenue* median
    - *Revenue* inequalities
    - Number of households per *Commune*
- Notebook: https://nbviewer.jupyter.org/github/Ashish-3/House-prices-in-France/blob/master/Preprocessing_communes_stats.ipynb

Real estate transaction records 2015 to 2019

- Source : https://www.data.gouv.fr/en/datasets/demandes-de-valeurs-foncieres/
- Files names: valeursfoncieres-2019.txt
  valeursfoncieres-2018.txt
  valeursfoncieres-2017.txt
  valeursfoncieres-2016.txt
  valeursfoncieres-2015.txt
- Provider: Ministère de l'économie et des finances

- Number of rows: 14.8 million in total (approximately)
- Contains:
    - One row for each real estate transactions
- Main feature used:
    - *Commune*
    - Real estate value *(Valeur foncière)*
    - Type of estate: house (*maison*), flat, outbuilding (*dépendance*), office…
    - Living area (*Surface Carrez)*
    - Land area (*Surface terrain)*
    - Number of main rooms (*Nombre de pièces principales*)
- Used for:
    - Estimating the price of houses per meter square for each *Commune*
    - Build a price estimator for houses in a specific area (Clermont-Ferrand)
- Pre-processing:
    - We filtered the transaction and kept only transactions:
        - concerning sales
        - concerning houses
            - with no outbuildings
            - and with less or equal to 1000m² of land surface
            - and with less or equal to 200m² of living area
            - and with less or equal to 10 main rooms
    - Many duplicates existed in the data frame, for instance outbuildings of a house sale were listed as a separate transaction but the house price. A gross approach was used to remove most if not all the duplicates.
    - Some empty land surface fields were replaced with the living area
    - Some empty living area fields were replaced with the building's base surface
    - The name of the *Communes* in the dataframe had to be corrected so it matches the other dataframes. This step allows us to merge different dataframes together.
    - The remaining dataframe consisted of only 1.6 million transactions. Those transactions were then used to:
        - generate a 32 567 rows dataframe of statistics of houses per *Commune*
        - build a house price estimator in Clermont-Ferrand
- Notebook: https://nbviewer.jupyter.org/github/Ashish-3/House-prices-in-France/blob/master/Preprocessing_house_transactions.ipynb

Foursquare's database

- Source : Foursquare API
- Provider : Foursquare Labs Inc.
- Description: Foursquare has a database of restaurants, bars, companies, and other venues that can be accessed through API calls. With a free Developer account, a limited amount of calls can be made.

- Used for: Getting the number of bars in *Communes* as a nightlife indicator. To minimize the number of API calls, Foursquare's database was used only on a pre-sorted list of *Communes*.
- API calls used: Search endpoint using the category Id of "bars"
- For more info: *https://en.wikipedia.org/wiki/Foursquare_City_Guide*

France Geojson

- Source: https://github.com/gregoiredavid/france-geojson
- File name: communes-version-simplifiee.geojson
- Provider: Gregoire David
- Contains: Polygon lines for each *Communes*
- Used for: Plotting choropleth maps

4. Methodology

Methodology section which represents the main component of the report where you discuss and describe any exploratory data analysis that you did, any inferential statistical testing that you performed, if any, and what machine learnings were used and why.

Notebooks

The project details can be found in four separate notebooks:

- Pre-processing data about the communes
  https://nbviewer.jupyter.org/github/Ashish-3/House-prices-in-France/blob/master/Preprocessing_communes_stats.ipynb

- Pre-processing the real estate transaction records
  https://nbviewer.jupyter.org/github/Ashish-3/House-prices-in-France/blob/master/Preprocessing_house_transactions.ipynb

- Data analysis
  https://nbviewer.jupyter.org/github/Ashish-3/House-prices-in-France/blob/master/Communes_analysis.ipynb

- Estimating house prices in Clermont-Ferrand
  https://nbviewer.jupyter.org/github/Ashish-3/House-prices-in-France/blob/master/House_prices_modeling.ipynb

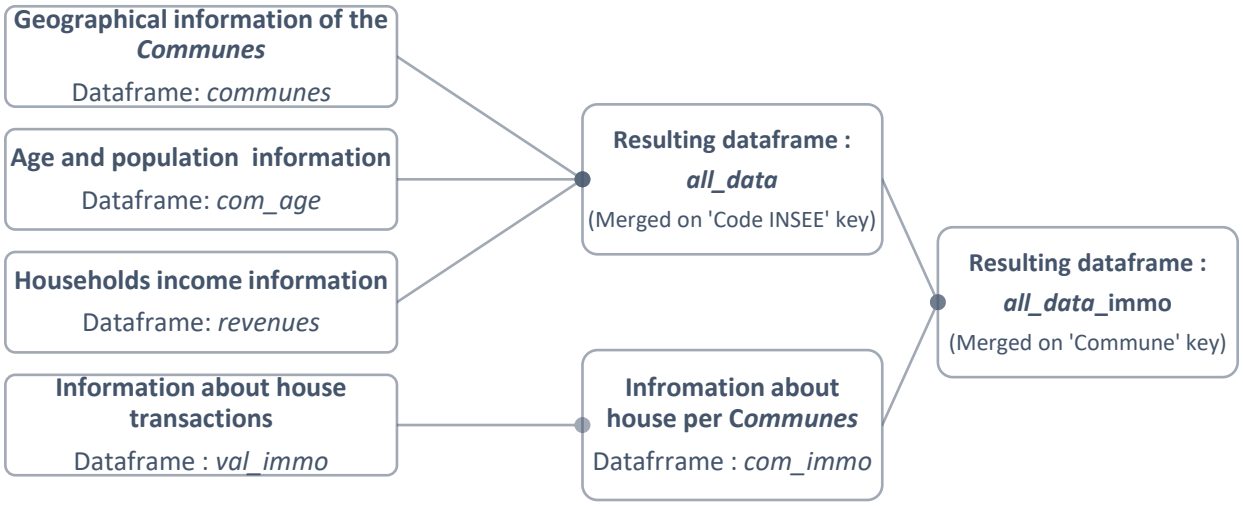
Overview of the data preparation

First, we started pre-processes de data about:

- The geographical information of the *Communes*
- The population and its age statistics
- The income information of households

And then we merged them all in a single dataframe.

Secondly, we studied the real estate transactions. It was pre-processed separately because of its size and the amount of the work needed. This was then used to create a house statistics dataframe. Thirdly, we combined all the data gathered in one a single dataframe.



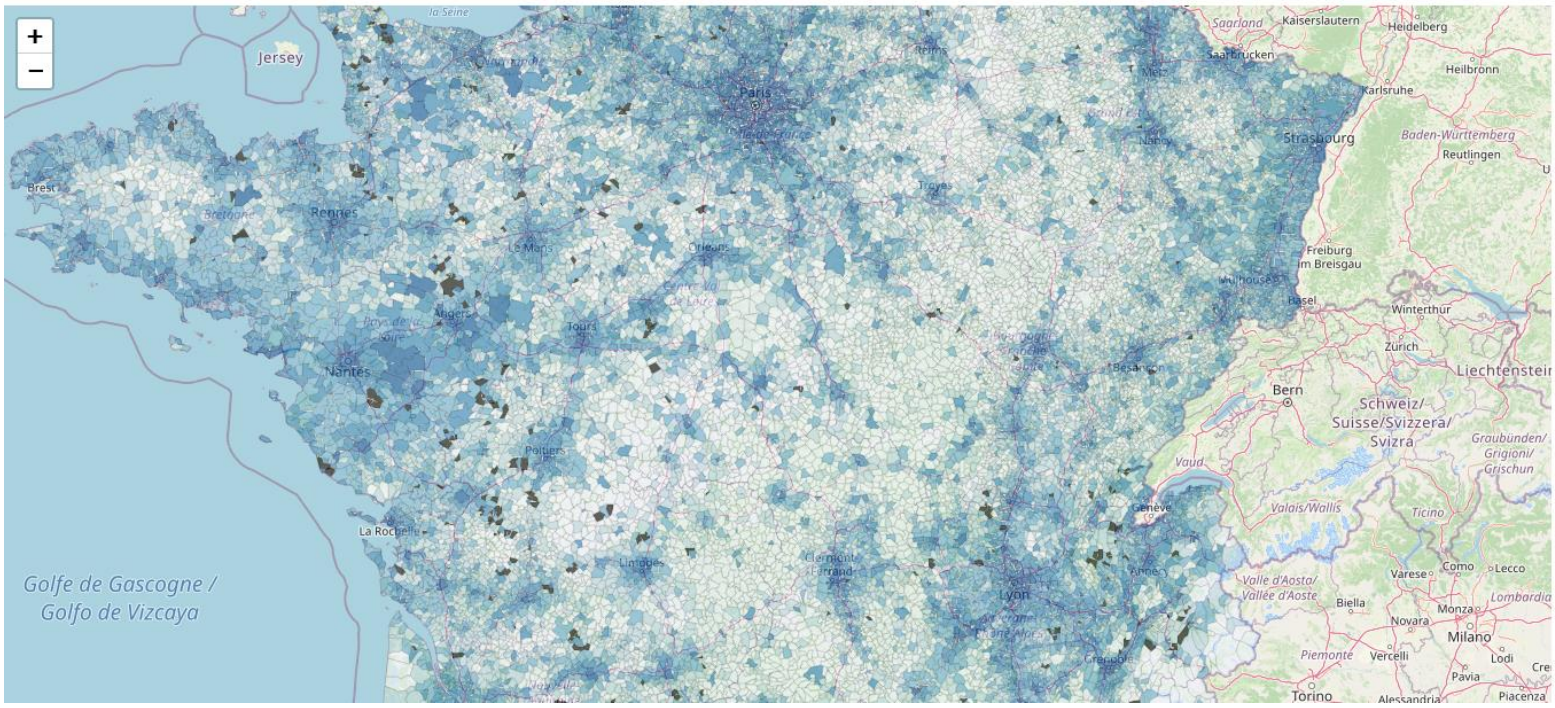
Exploratory data analysis

Once all the preparation done; we can now start to visualise how the different indicators varies across the French territory. With the help of a Geojson file mapping all the Comunnes of France, here is an extraction of the obtained maps:

- Map of the population and / or population density
- Map of revenues median
- Map of revenues inequalities
- Map of house prices per metre square

Representation of the feature : density

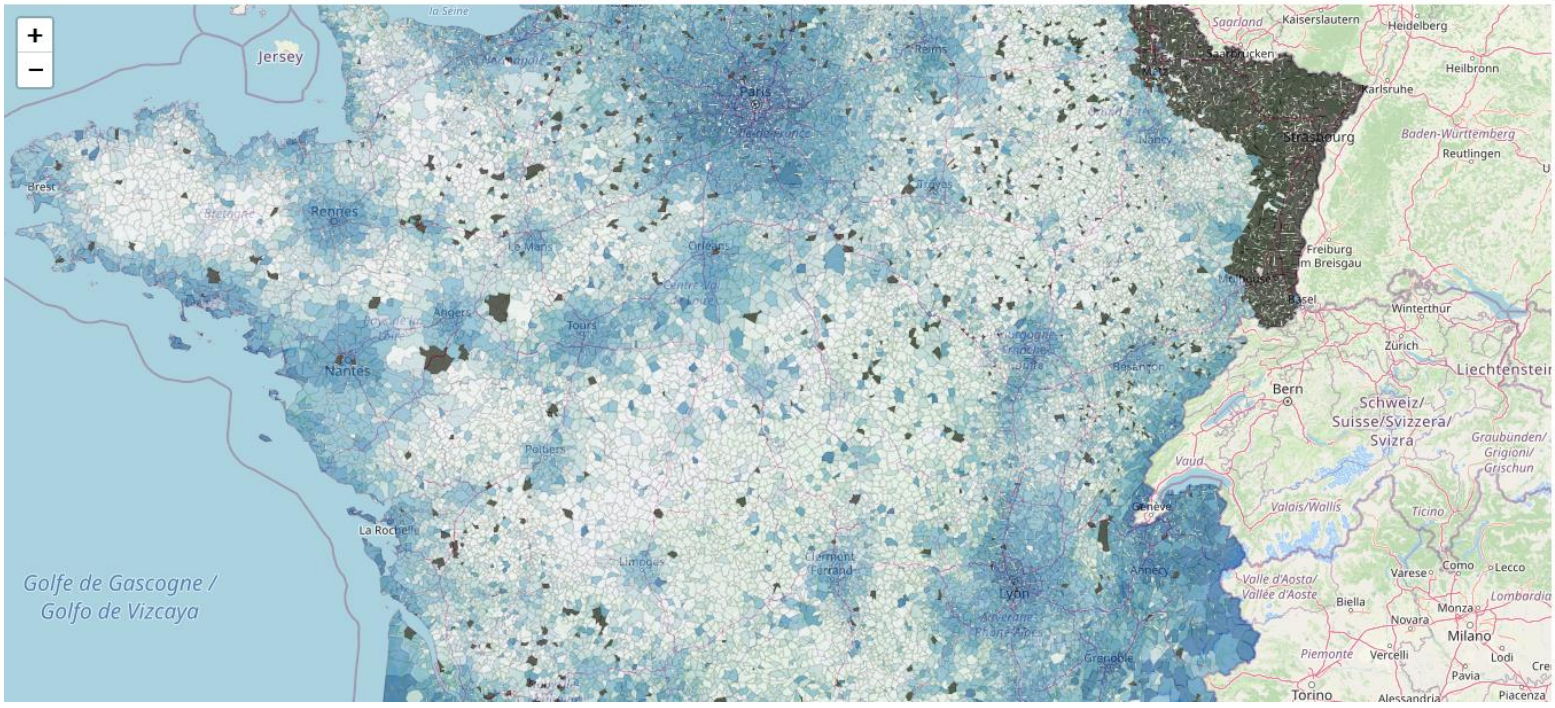
Bins used : [0, 19, 41, 99, 536, 40169] (Warning : the map is heavy and may take time to load)



-

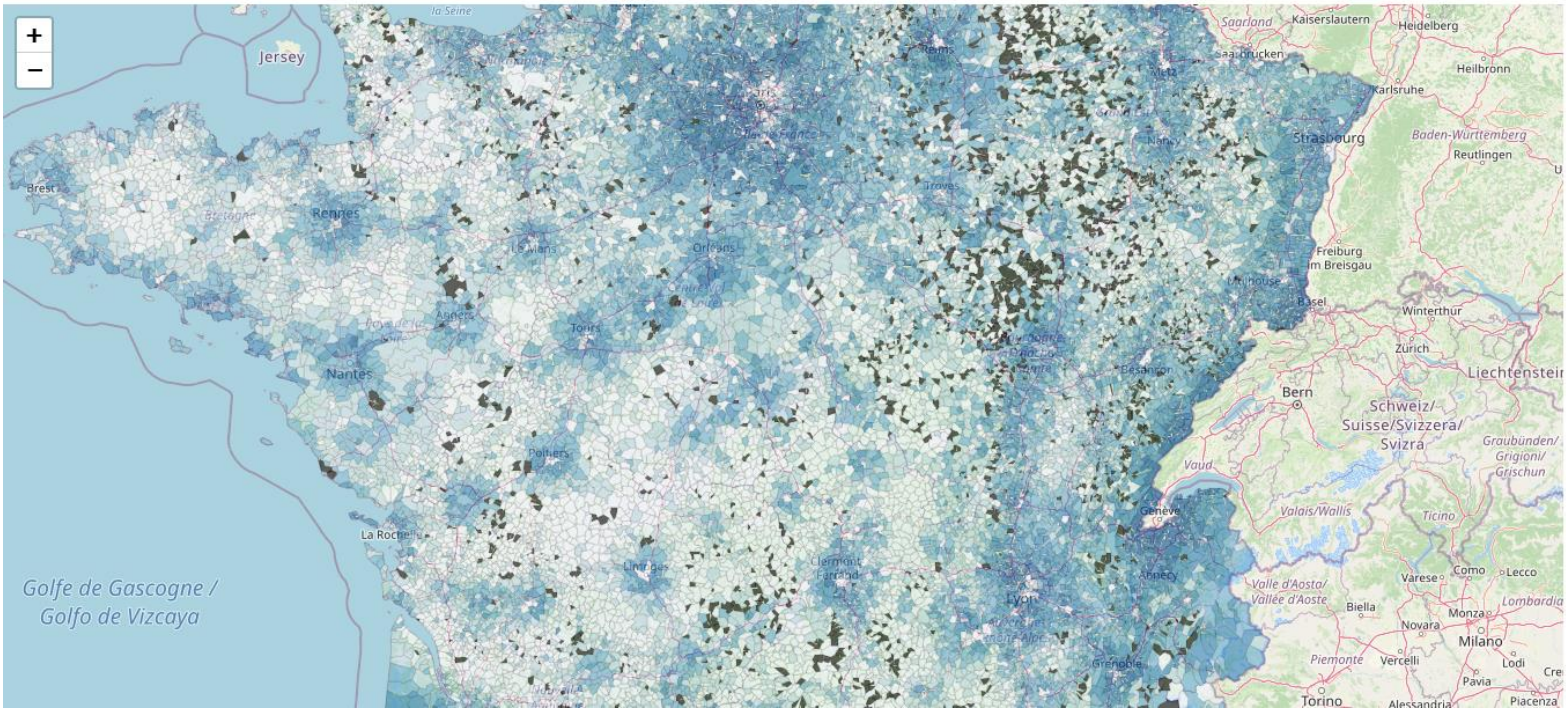
Representation of the feature : Prix_m2

Bins used : [0, 1030, 1358, 1792, 3028, 376914] (Warning : the map is heavy and may take time to load)



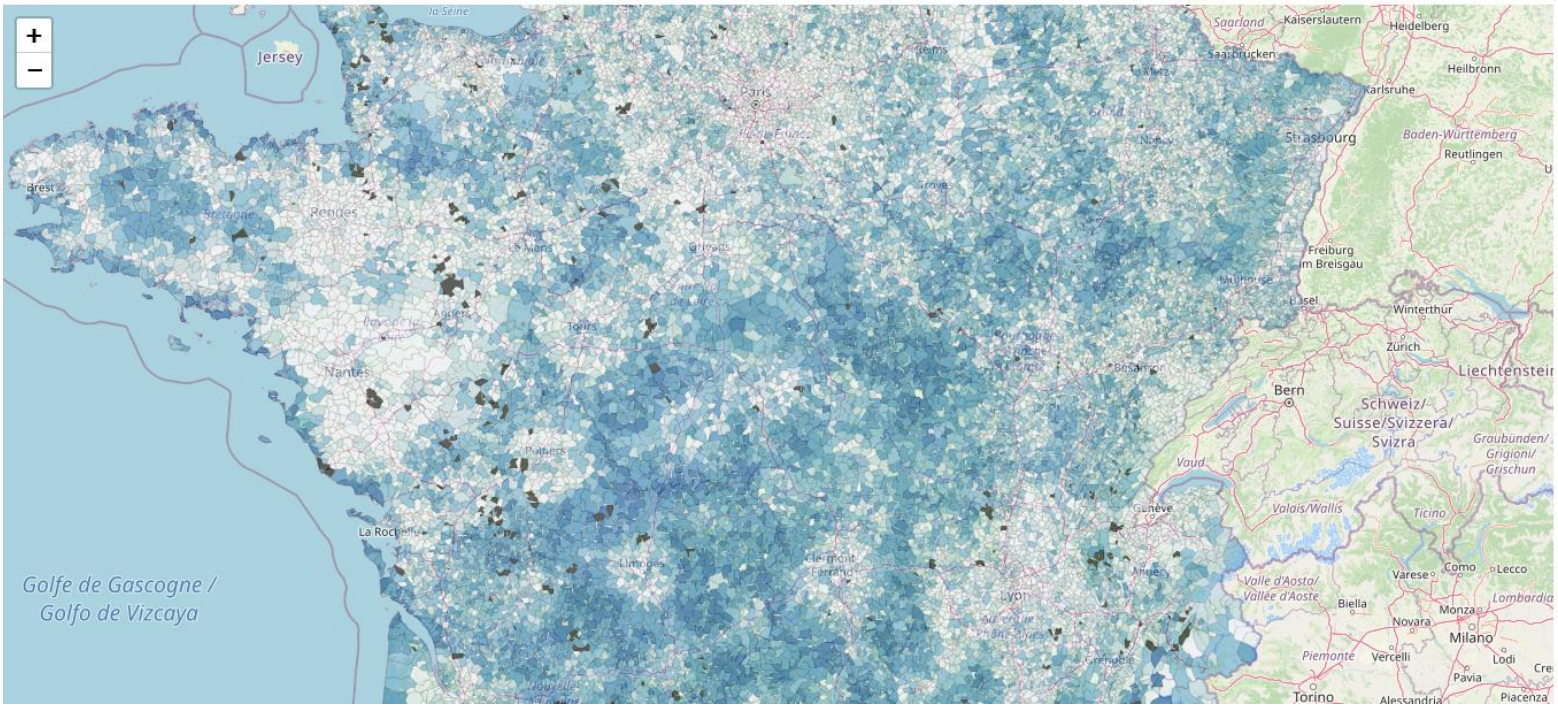
Representation of the feature : revenue_median

Bins used : [11070, 19320, 20760, 22560, 26330, 48310] (Warning : the map is heavy and may take time to load)



Representation of the feature : age_median

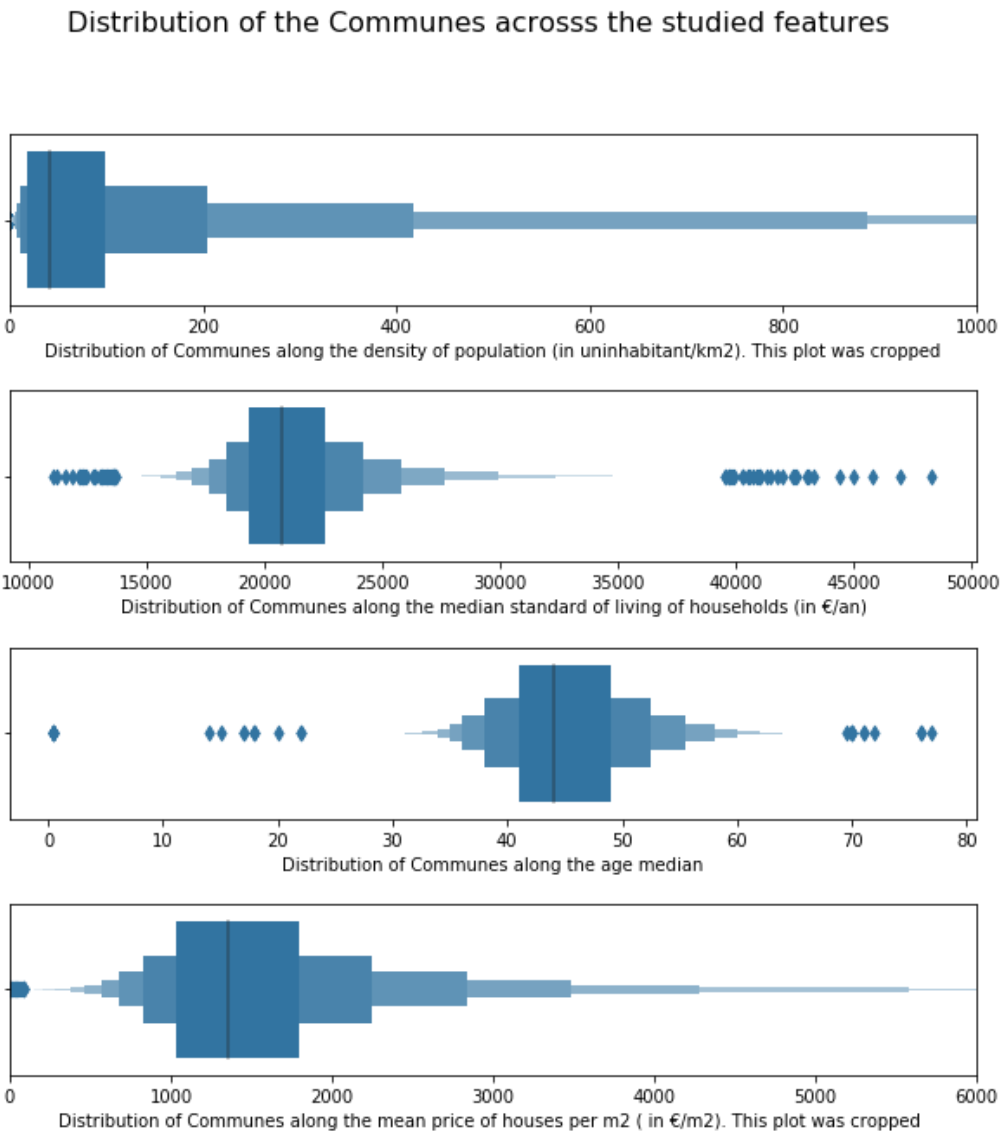
Bins used : [0, 41, 44, 49, 56, 77] (Warning : the map is heavy and may take time to load)



The interactive maps (html versions) can be downloaded from here for better viewing experience:

https://github.com/Ashish-3/House-prices-in-France/tree/master/maps

To better understand how the Communes are distributed along the studied features, we have built a small figure that show box plots of all features. The style of the box plot used here is called "Letter value plots".



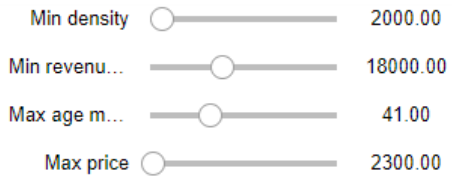
Searching for a decent place

We can now start searching for a place to buy a house with the different indicator we have. For this purpose, we have built a small tool that allows us to filter the *Communes* very easily. Here is a screenshot of the tool.

If we filter the data with those criteria's, we can view places:

- With a dense enough population — minimum of 2000 person per km2
- With a population no to old — maximum age median of 41 years old
- Where people have a decent income — minimum standard of living of 18k€
- With a fair house price — max. average of 2300€/m² for houses

With those parameters we can see that only 35 places satisfy those criteria. Here is a screenshot of the tool:




```
32 result found
Parameters : filter_result(density= 2000.0 , revenue= 18000.0 , age= 41.0 , price= 2300.0 )
```

	Code INSEE	Code Postal	Commune	Code Département	Code Région	Superficie_km2	lat	lng	age_r
30844	62041	62000	ARRAS	62	31.0	11.93	50.289896	2.765873	38.30
23974	29019	29200	BREST	29	53.0	49.12	48.400500	-4.502791	38.45

Using Foursquare's database to get an insight of night life

Once we have a limited number of *Communes*, we can now use the Foursquare API to gather information about those areas. We did a Foursquare search of available bars in the list of *Communes* found previously. The search was conducted with the category Id for bars, in a radius of 3km from the centre of the *Communes*.

If we consider the presence of bar as a good indicator of nightlife, we can see a good contrast between some *Communes*. Clermont-Ferrand stands out with 50 bars which is the maximum amount of venues that the Foursquare API can return in single call. We can say that, though all the filtering, Clermont-Ferrand is the winner!

Commune	population	N	nb_venues
CLERMONT-FERRAND	142686.0	1382.0	50
BREST	139342.0	2938.0	48
LE MANS	142991.0	5398.0	35
CROIX	21271.0	1093.0	32
SAINT-MAURICE	14312.0	55.0	28
LOOS	22076.0	817.0	28
SAINT-ANDRE-LES-VERGERS	12116.0	436.0	26
LE HAVRE	170352.0	4106.0	23
POITIERS	87961.0	2087.0	22
WATTRELOS	41341.0	1960.0	20

Estimation of house prices in a town

We have seen that Clermont-Ferrand stands out considering our filtering criteria. Let us study the actual prices of house in Clermont-Ferrand.

A quick analysis show that the price of the houses is correlated to the number of rooms, the land area and obviously the living area. From those features and using machine learning, we have built a model that predict the house prices. Different model can be used for such prediction, but the most common ones are multiple linear regression, polynomial regression, and ridge regression.

To choose between those models, we have split the 1209 house transactions of Clermont-Ferrand in a training set and a testing set with 70/30 ratio. For this study the training set was used to train the different models, but it was also used to tune the hyper parameters such as : data standardisation, the polynomial order (for polynomial regression) and the parameter alpha (for ridge regression).

We have tested each model with the coefficient of determination R^2 and the MSE score. The study shows that:

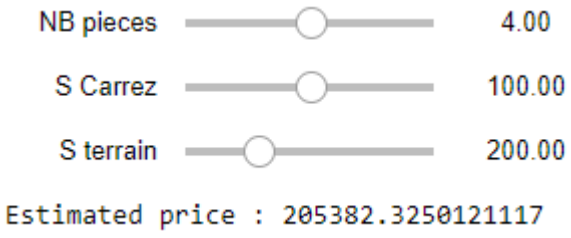
- The polynomial regression model performs better with a degree of 1 (which corresponds to a linear model), and data standardisation have almost no effect on the model.
- The ridge regression performs better with alpha coefficient of 10 000.

Here are the scores obtained bellow:

	R2 score	MSE score
Linear regression	0.404699	6.043176e+09
Polynomial regression	0.404699	6.043176e+09
Ridge regression	0.406139	6.028566e+09

Therefor in this case, the best model is also the simplest model: a linear regression model.

After training this model on all Clermont-Ferrand's house transactions, we have built a small tool that can predict the price of this *Commune* from the living area, the land area, and the number of rooms.



5. Results

Results section where you discuss the results.

Where is a "good place" to buy a house in France and at what cost?

In this study we pointed out around 35 "good places" to buy a house and built a proper model to estimate how much a house can cost. In this case, we have seen that Clermont-Ferrand stands out of the other *Communes*.

The parameters used for choosing a *Commune* can be freely modified to find a place that can better fit a person's preference. And the price estimator model can also estimate the price of different kind of houses depending on a person's preferences.

However, we have seen that the price estimator model has a coefficient of determination R^2 of 0.40. This score remains modest but is enough to give a quick overview of house prices.

6. Discussion

Discussion section where you discuss any observations you noted and any recommendations you can make based on the results.

This approached gave a proper overview of real estate in France, particularly houses. One of the greatest assets of this study is its flexibility. With the exact same tools, we could very easily study apartments or offices for instance.

However, a more in-depth study can be done by optimising many aspects of this work. For instance, many transaction observations were lost during the data cleaning process. A less strict cleaning process could give more accurate figures and better price prediction.

One of the other aspects to be improved is the modestly low R^2 score of the price estimator. This score can be easily explained by the fact that the features used to train the model are not enough to explain the price of a property. For example, the age or state of the property, the safety of the borough, the proximity to transit facilities or city centres, and other factors can strongly influence the price of a property. We can note that the data we have also include the address of the properties. One can easily use a geo-localisation services to analyse distance of properties to the city centres for better prediction.

It is interesting to note that with the data collected, it could also be easy to analyse other topics such the correlation between social and geographical indicators. For example, the correlation of the poorest households with the density of population, or the correlation of the age population and the household incomes…

7. Conclusion

Conclusion section where you conclude the report.

This study shows that with some geographical data and basic data science skills we can extract precious information about real estate choices. In my opinion this project succeeded in two ways. Firstly, the goal of the project was achieved even if there is room for improvement. Secondly, on a more personal level, this project not only allowed me to practice the skills I have learned from my classes, but it and pushed me to learn more. Hoping that someday I will be able to work on even more meaningful projects.