Programming for big data assessment

MASTER 2 EBDS **MASTER 2 APE**

Realized by:

Freedel Zinsou-Ply

Rachad Liassou

Hadare Idrissou

Supervised by:

Jonathan Stern

Laurent Centisoy

Olivier Philip

January 2022

Content

Introduction and problematic

A land value is the value of a piece of land, a property based on its future construction potential. In other words, it is the price at which the building can be sold on a particular market with regard to supply and demand. During a real estate transaction it is therefore imperative to know the property value of a building for both parties, on the one hand for the buyer in order to have an idea of the real value of the asset and not to be scammed, and on the one hand for the seller to set a correct price and make a satisfactory profit from the sale.

In France the land value of a plot of land or real estate can be influenced by several characteristics namely: the region, the type of land (apartment, house, agricultural land, etc ...), the area and many other characteristics. One remarkable thing is that over time, the value of a property increases enormously. Because of this it is therefore important to be able to predict the value of a property. Since the value of a property depends on its land value, it would be more interesting to be able to predict the land value from the various characteristics of the property or land.

The objective of this work is therefore to predict from statistical, econometric and machine learning methods, the land value according to the type of real estate. To do this we will rely on the databases from a French Administrative Direction, namely the Directorate General of Public Finance (DGFiP). The databases date from 2016 to 2021 and lists all real estate sales made over the last six years, in metropolitan France and in the overseas departments and territories except in Mayotte and Alsace-Moselle.

1- Description of the dataset

The data collected on the government website are txt files that contain the different information in relation to the land sale over the yearsfrom 2016 to 2021. Each database contains 43 variables. The information in each database covers all the departments and communes of France, but we have decided to work only on the Toulouse region. What justifies this choice is that we checked the occurrence of cities in the databases. We have noticed that Toulouse oftenshows. A fois our chosen study area we merged the different bases by filtering in relation to Toulouse. After the merger, we cleaned up our database, we noticed a lot of missing values in the database, as well as null values. So we removed these missing values. After that we tackled the data visualization part of our database. The following table and graphs report the different information in relation to our database.

2- Methodology

Multiple linear regression

Linear regression is a linear model that makes it possible to make estimates in the future from information from the past. In this linear regression model, we have several variables, one of which is an explanatory variable and the others which are explained variables. This tool is used for stock market technical analysis but also for budget management. It is often calculated using the least squares method, which reduces errors by adding information.

Linear regression based on CPA

Principal component analysis is an exploratory statistical tool. It allows you to explore multivariate data, that is, data with several variables. It is therefore considered a

multivariate statistical analysis or a multivariate technique, allowing to reduce a set of initial variables into a few factors or main components that are new variables. The initial variables are thus reduced into a small number of new variables or main components, while retaining as much information as possible. There are therefore a small number of factors that explain most of the variance contained in the initial variables. The information contained in the original variables is thus extracted, visualized and synthesized into a few new variables, from a cross between several numerical variables. The dimensions of a multi-variable data are thus reduced to a few main components that can be visualized graphically. The CPA results in a graphical representation of the data (point cloud) according to these factors or main components in the form of axes. These main axes or components summarize as well as possible all the initial variables. They are linear combinations of the initial, hierarchical and independent variables.

Support Vector Regresssion

Support vector regression (SVR) is very different from other regression models. It uses the Support Vector Machine (SVM) algorithm to predict a continuous variable. Since SVR performs linear regression in a higher dimension, this function is crucial. There are many types of nuclei such as polynomial nucleus, Gaussian nucleus, sigmoid nucleus, etc. In Support Vector Machine, a hyperplane is a line used to separate two classes of data in a dimension greater than the actual dimension. In SVR, a hyperplane is a line used to predict a continuous value. While other linear regression models attempt to minimize the error between the predicted value and the actual value, support vector regression attempts to fit the best line into a predefined error value or threshold. What SVR does in this sense, tries to classify all prediction lines into two types, those that go through the error limit (space separated by two parallel lines) and those that do not. Lines that

do not exceed the error limit are not considered to be the difference between the predicted value and the actual value exceeded the error threshold, ε(epsilon). Passing lines are considered a potential support vector to predict the value of an unknown.

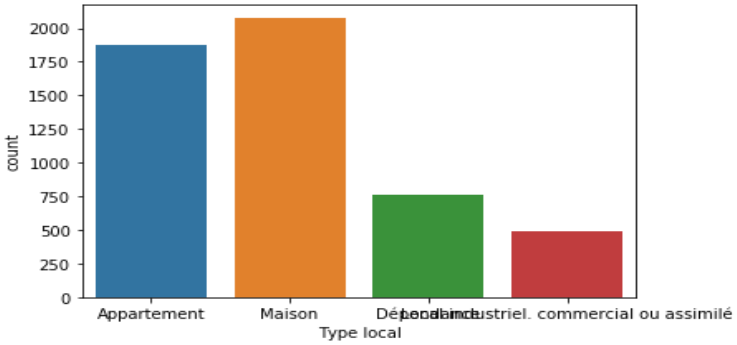
Data analysis and presentation of results

Table 1: Distribution of mutation types by occurrence

	Count	percentage
Vente	5110	98.080614
Echange	77	1.477927
Adjudication	12	0.230326
Vente en l'état futur d'achèvement	10	0.191939
Vente terrain à bâtir	1	0.019194

Table 1 shows us that the most common transactions in terms of real estate are the sale of land which occupies more than 98.08%. All the others, i.e. the Auction, the Exchange, the sale of building land, the Expropriation represents only a small part, barely 2%.

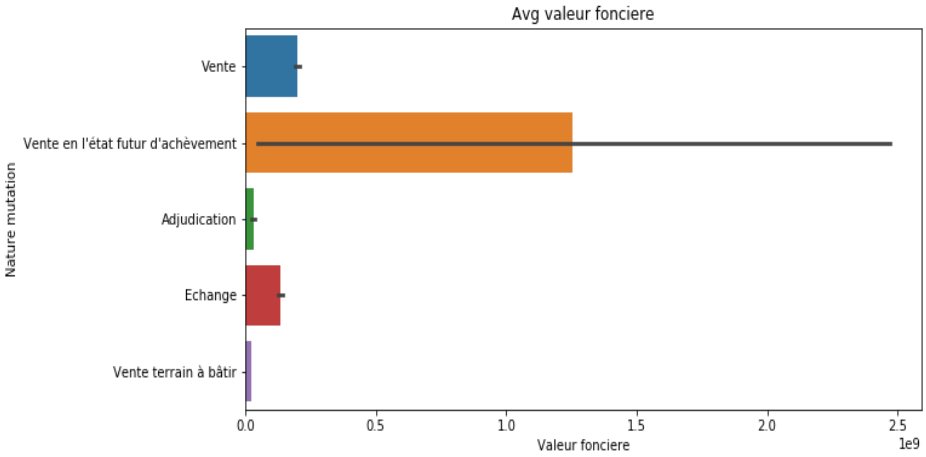
Figure 1: Occurrence of the different local types.



Source: Produced by the authors.

The following Figure 1 shows the types of premises that are the best-selling. From the results of this graph we deduce that the best-selling premises are the apartment and house.

Figure 2: Land value according to the type of transfer.



Source: Produced by the authors.

Figure 2 shows us the distribution of land value according to the nature of the change. This graph clearly shows us that the land value of sales in the future state of completion is higher than all the others.

Figure 3: Land value by type of premises.

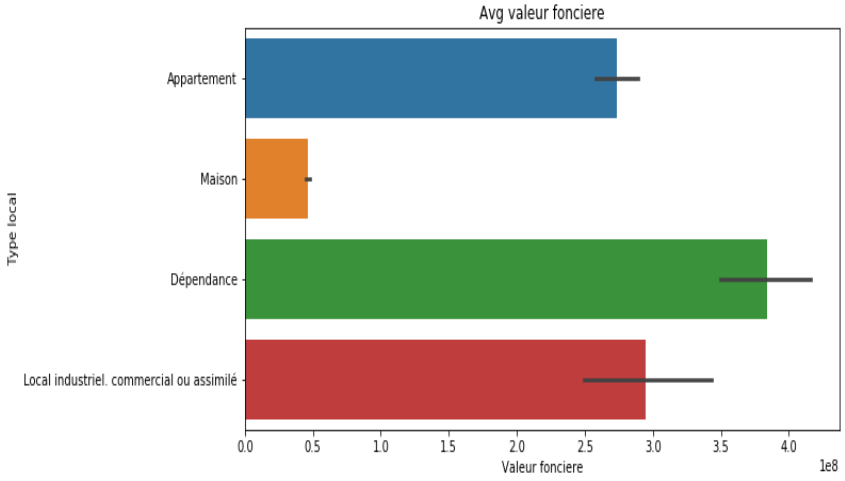
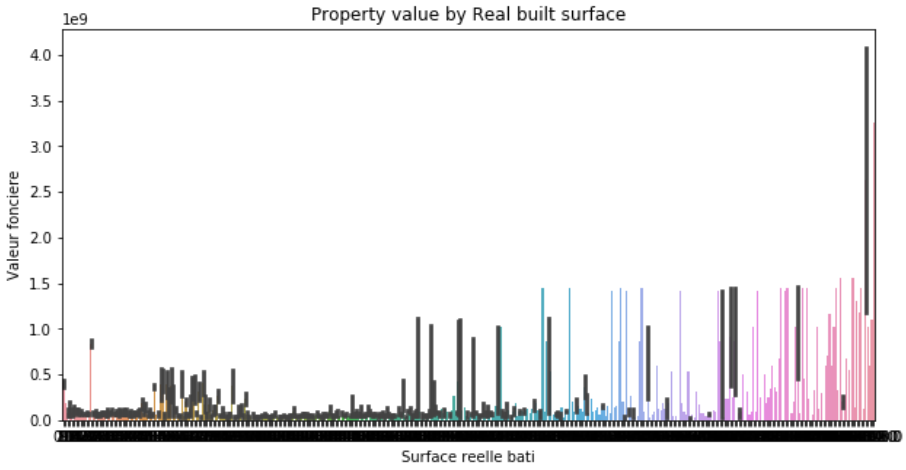


Figure 3 shows the distribution of the land value according to the type of premises. This graph clearly shows us that the sale of industrial, commercial or similar premises and dependency have a higher land value than houses, apartments.

Figure 4: Land value as a function of the actual built area.



The following graph shows the land value as a functionof the actual built area. From this graph we can remember that on average the larger the surface the higher the land value. This confirms the previous results concerning the type of premises, since it is known that the usines generally have more surface area than houses and apartments.

3- Modeling of a relational database

We have separated the database into 4 tables:

- The first table called Localization containing all the variables to locate or situate the land, apartment or house and having for primary key the department code
- The second table called Mutation containing all the necessary information concerning the sale, or any other operation carried out on the land, the house or the apartment, this table has for primary key id of the mutation, this variable is not in the database, but we think that it would be interesting to have a unique identifier associated with each operation, that will allow to find each operation details.
- The third table called Local Characteristic contains all the characteristics of the land, house or apartment, the primary key of this table is local id.

- The last table Administration is the table that contains all the information related to the administrative documents of the houses or lands, the primary key is code service CH.

All these tables are linked together by the id local

Figure 5: Design of the database

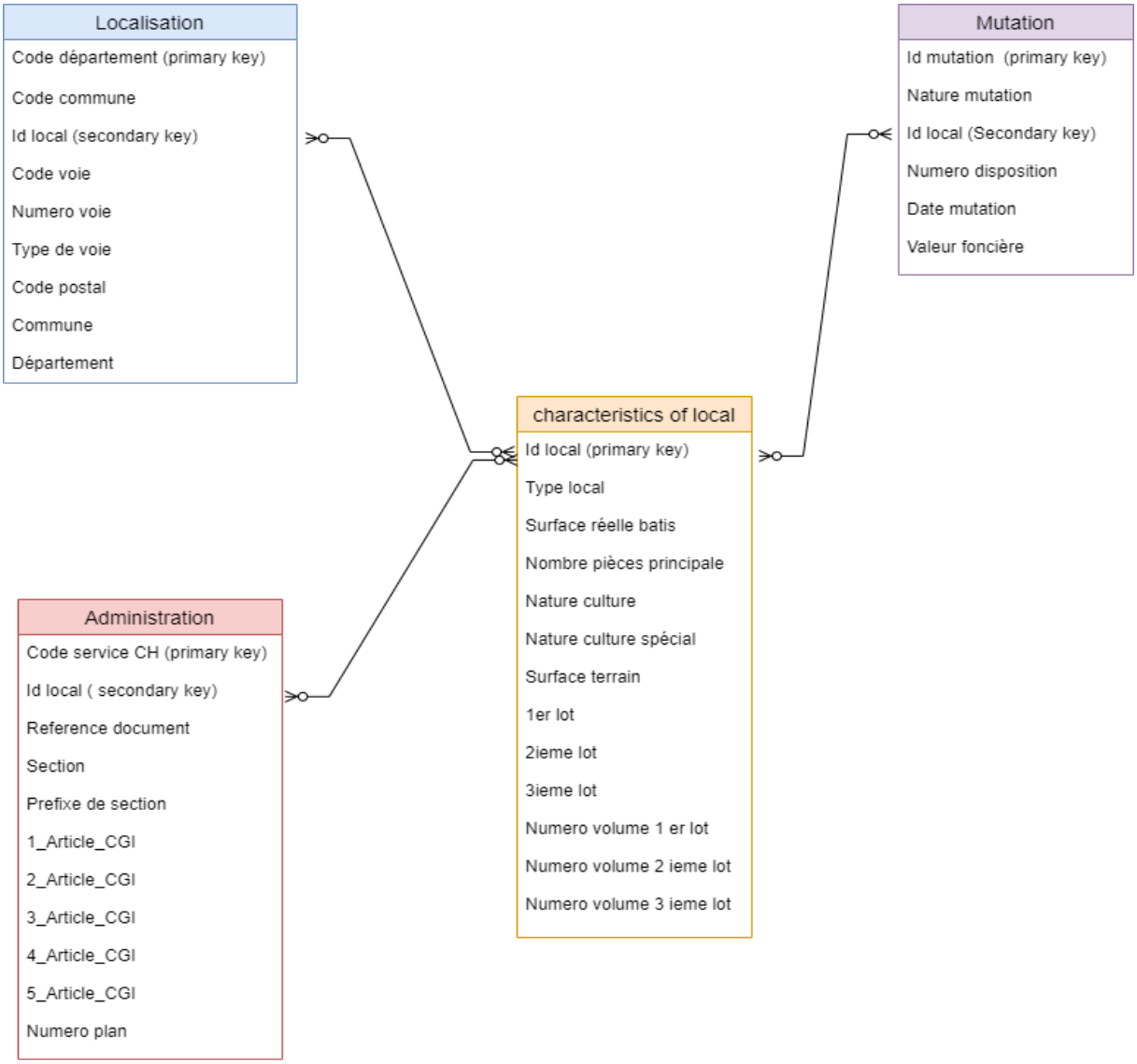


Figure 6: Description of the database

Variables	Types	Description
Localisation		
Code département	numerical	Allows to associate each land with a unique department
Id local	numerical	Allows to associate each land to a unique identifiant
Code commune	numerical	Allows to associate each land with a unique commune
Code voie	numerical	Allows to associate each land with a street
Numero de voie	numerical	Allows to associate each land with a street's number
Type de voie	alphabetic	Allows to know the type of road (street, boulevard)
Voie	alphabetic	Allows to know the name of the street
Code postal	numerical	the postal code of the land
Commune	alphabetic	The name of the commune
Mutation		
Nature mutation	alphabetic	Allows to know the type of mutation(Sale, exchange, expropriation, etc.)
Id local	numerical	Allows to associate each land to a unique identifiant
Date de mutation	date	Date of signature of the act
valeur foncière	numerical	amount or valuation declared in the context of a transfer for valuable consideration.
characteristics of local		
id local	numerical	Allow to associate each land or local to a unique idendifiant
Type of local	aphabetic	Allow to know the type of local(apartment , house)
Surface réeelle batis	numerical	Allow to know the surface in m^`2 of the local
Nombres de pièces	numerical	Allows to see the number of rooms in each apartment or house
Nature culture	alphabetic	Allows you to see the type of cultivation done in the field
Nature culture special	alphabetic	Allows you to see the type of cultivation done in the field
Surface terrain	numerical	Allows to see in m2 the total surface of the land
1 er lot	alphabetic	Allow to identify a part of a building and thus to associate a specific property right to it .
2 ieme lot	alphabetic	Allow to identify a part of a building and thus to associate a specific property right to it .
3 ieme lot	alphabetic	Allow to identify a part of a building and thus to associate a specific property right to it .
Numero volume 1er lot	numerical	Allow to identify the first lot
Numero volume 2 ieme lot	numerical	Allow to identify the second lot
Numero volume 3ieme lot	numerical	Allow to identify the third lot
Surface 1 er lot	numerical	Allow to identify the surface in m^2 of the first lot
Surface 2ieme lot	numerical	Allow to identify the surface in m^2 of the second lot
Administration		
Code service	numerical	Administrative document related to the land
Id local	numerical	Allows to associate each land to a unique identifiant
Reference document	alphanumerical	Administrative document related to the land
Section	alphabetic	Administrative document related to the land
Prefixe de section	numerical	Administrative document related to the land
Numero de plan	numerical	Administrative document related to the land
1 Articles CGI	numerical	Administrative document related to the land
2 Articles CGI	numerical	Administrative document related to the land
3 Artilcles CGI	numerical	Administrative document related to the land
4 Articles CGI	numerical	Administrative document related to the land

4- Manipulation and exploitation of data with the Python programming language

Model 1: Multiple linear regression

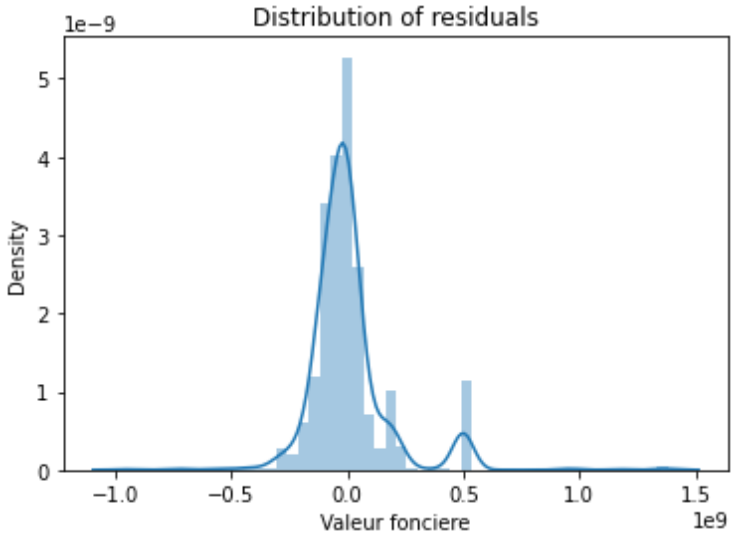
Table 2: Linear Regression Estimation Results

Variables	Coefficient	P value
Const	9.112e+07	0,182
Nature mutation_Echange	-1.315e+08	0.072
Nature mutation Sale	5.631e+07	0.408
Nature mutation Sale in the future state of completion	1.189e+09	0.000***
Nature mutation sale building land	6.658e+06	0.975
Local type Dependency	-6.48e+06	0.595
Local type Industrial local. Commercial or similar	-9.876e+07	0.000***
Local type_Maison	-5.158e+07	0.000***
Actual built area	1.663e+04	0.008***
Land area	8.6e+04	0.000***
Number of main rooms	-2.067e+07	0.000***
Number of lots	-3.307e+07	0.458
R2	0.67	

From the results of the estimation of the model by ordinary least squares, we retain that the R-square of the model is 67%, which reflects a good predictive power. Also, several variables including "Type local_Local industrial", "Type local_Maison", "Real building surface", "Nature mutation_Vente in the future state of completion", "Ground surface", "Number of main pieces"are significant at the threshold of 5%.

The area is positively correlated with the property value while the number of main rooms and the number is negatively correlated.

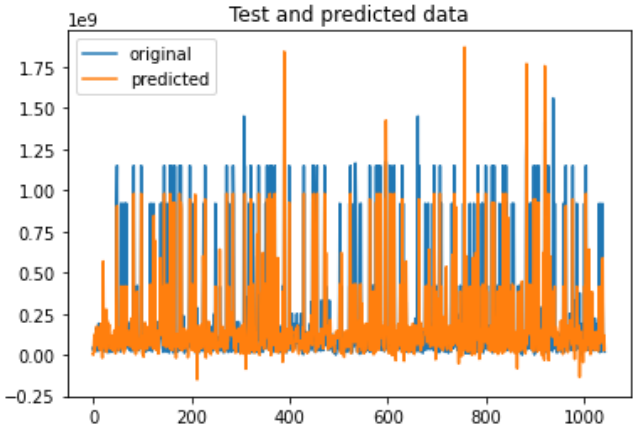
Figure 7: Distribution of linear regression residuals



Source: Produced by the authors.

The curve looks like the normal distribution, the distribution of residues follows approximately a normal distribution.

Figure 8: Comparative analysis of predicted and observed values based on linear regression

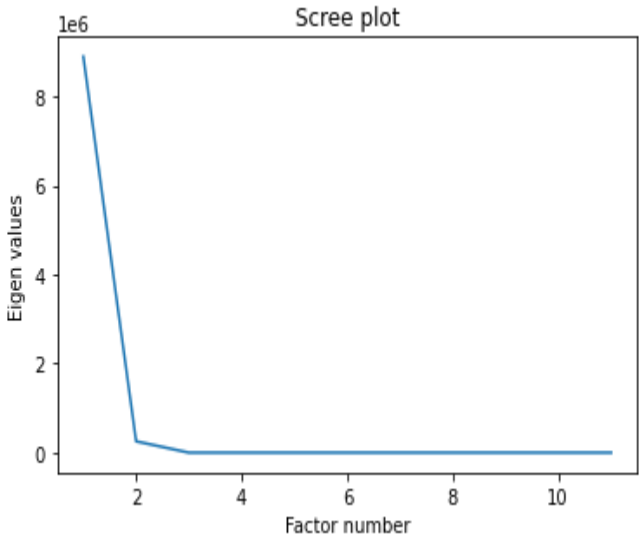


Source: Produced by the authors.

We notice through the graph that there are no large differences between the predicted and observed values. This confirms the good fit quality of our model.

Model 2: Multiple linear regression based on ACP

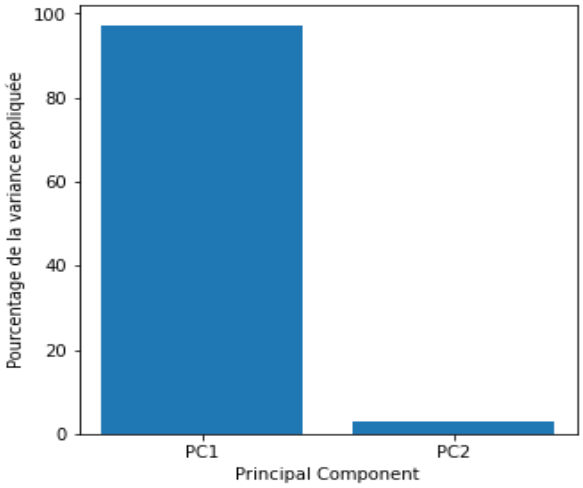
Figure 9: Detection of the number of optimal principal components



Source: Produced by the authors

From the analysis of the graph representative of the eigenvalues as a function of the number of factors, it is observed that the optimal number of principal components is 2.

Figure 10: Amount of information explained by the two factorial axes used



Source: Produced by the authors

The first two factorial axes contain more than 99% of the information extracted from initial data. Also, note that the R-Square associated with linear regression based on ACP

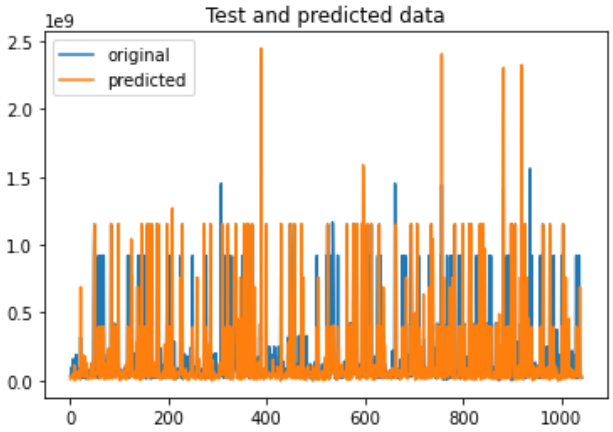
is about 63%. This seems less than that of the first model, but the ACP is still a good recourse when you have large data.

Model 3: Support Vector Regression

We first chose the right kernel for our data. After analysis as shown in the table below, we observed that the linear core was the most suitable for our data

Kernel	Values
Linear	0.56
Poly	-0.17
rbf	-0.17
Sigmoid	-0.17

Figure 11: Comparative analysis of predicted and observed values based on SVR.



Source: Produced by the authors.

We notice through the graph that as in the case of linear regression, there are no large differences between the predicted and observed values. This confirms the good fit quality of our model also for the SVR with an R-square of about 56%.

Conclusion

This assignment allowed us to deepen our knowledge in database management, mainly on the formation of several tables from a database, and processing on the statistical, econometric and machine learning levels. The formation of several tables allowed us to avoid duplication and to better structure our database, the data processing and analysis part, how to better describe a database, how to process it, analyze it, make econometric and machine learning models. All this work was carried out in two languages, sql and Python, two essential languages in the world of programming, analysis and data processing.