Data analysis tools (SAS and R)

Project - 3rd year BFA

Jérôme Lepagnol Pierre Lepagnol

Preliminaries

Expected for : 18th November 2024 (23h59 GMT+1)

In teams of 2 (to maximum 4) students, submit :

- All source code (SAS or R scripts, qmd, etc.)
- Professional report document in HTML or in PDF format (ideally generated/knitted by the qmd/rmd).
- Input files (if modified)
- Output files (if applicable)

Normally, no input files change is needed, but if you imperatively need to modify the input data files (in XLS/CSV file for example) :

- Please provide these modified files with a note explaining your changes to these files.
- All option consisting in scripting the modification of data is preferred

During all the project you should check for outliers, remove row if needed preferably by coding.

Please have a look at meta files, describing in English the meta-data documentation.

You may use this information to interpret variables present in your charts, models and tables. There is no need to load this file.

The data sets files to be used are the result of web-scrapping.

Those data files are delivered only for *this* academic use and shall in no way be used for any other purpose

.

The files (meta-data and data set) can be found under the public git repository here : link to github project 2024

Goals & Assessment

The goal of the project is to make a professional report on the prices of real estate in the Grand Paris based on the data provided. Beyond understanding the data structure, providing key knowledge on the data within the set, you should model (predict) the price of real estate in order to discover/identify the best investment opportunity, for your management.

You will be **evaluated** on the following :

- **Data preparation**: Did the students properly prepare the data, including checking for missing values, dealing with outliers, and transforming variables if necessary?
- **Appropriate statistical method**: Did the students choose the appropriate statistical method for the type of data ?
- **Interpretation of results**: Did the students provide a clear and accurate interpretation of the results, including explanations of statistical measures, hypothesis tests, and confidence intervals?
- **Communication of results**: Did the students effectively communicate their results through visualizations, tables, and written explanations, using appropriate statistical language and conventions?

Document your understanding of the "data model" behind the table

The files provided describe data with an underlying data model. The objective here is to provide a normalized model (Class diagram in UML format)) of the data *behind* the file.

The source of the information is : the contents (headers of) files :

- real_estate_data_2024.csv (main table)
- property_details_table.csv
- table_id_photos.csv
- table_merge_id_transport_stations.csv
- transport_stations_lines.csv

plus the content of file meta data (describing the data of the main table):

- realestate_metadata.csv

(1) Identify the entities, associations and attributes from the data file**s**, and imagine some entities, associations and attribute that are not in the data files, but could have usefull for the understanding of the data. Make a clear difference between what is there and what is not.

(2) Based on above considerations, draw a data model that explains the link between objects. The corresponding data model should at least explain the structure of the current provided files.

Import and qualification of the data

Import and correct the data sets

(3) Import the data sets and check that the characters with accent is readable in the data frame

- real_estate_data_2024.csv (main table)
- property_details_table.csv
- table_id_photos.csv
- table_merge_id_transport_stations.csv
- transport_stations_lines.csv

Be carefull : some files are comma separated, some are semi-colon separated.

(4) The data sets contain some duplicated rows : write the code to remove them. You should end with reduced data sets. In the following we will consider those data sets with no duplicates.

(5) In the mail file, do the same to `reference`, in order to split the reference in new columns

```
- ref_part0 <- substr(real_estate_data_2024$reference, 1,3)
- ref_part1 <- substr(real_estate_data_2024$reference, 5,10)
- ref_part2 <- substr(real_estate_data_2024$reference, 11,15)
```

Discuss the relevancy of the different parts created with regards to our goal. Keep the full `real_estate_data_2024$reference` for joining key and relevant new data made upper.

Similarly, in the main file create a new variable `departement` has been created by extracting the 2 first digits the `code.postal`.

(6) Join the tables coming from :

- real_estate_data_2024.csv (main table)
- property_details_table.csv

in order to have a main table enriched with data coming from the "property details table".

(7) Please add a metric counting the number of pictures in the mail data set. For the following steps, the main table enriched with this data will be considered.

To do this, consider using the data in the table :

- table_id_photos.csv

(8) Please add metrics counting the number of transportation lines in the main data set, split by type of transportation (RER, Train, Tram, Metro, and eventually the total for all types). For the following steps, the main table enriched with this data will be considered.

To do this, consider using the data in the tables :

- table_merge_id_transport_stations.csv
- transport_stations_lines.csv

(9) Enrich the data model with the 2 additionnal peices of information computed above (number of picture, number of transportation lines per type)

(10) Please remove any data from "the main table enriched" outside of the Grand-Paris area (departements 75,77, 78, 91, 92, 93, 94, 95) . For the following steps, the main table with reduced data set will be considered.

Uni-variate Statistical description

Starting a this point, the table to be considered is still the unique table coming from the main table, with enrichment done in steps 4/5/6/7/8 and filtering at step 10.

(11) As a prerequisite, list all variables your are going to describe and determine which are:

- **continuous variable** (numeric) :
- **categorical variable** (can be made out of figure, thus)
- **Free text** (non categorical variable) -> to be ignore in the rest of the exercise.
- **other** if any , please precise if it's relevant or not.

(12) Make a full statistical description of each individual variable by

- a dedicated graph (data viz) and
- a summary of the most important values (average, mode…) and
- an plain English explanation of your vision on the individual variable

Adapt your graph, the geometry of the plot to the variable your are plotting.

Do not build non-sense plots (such as bar-plot for continuous data or scatter plot for categorical variable).

Bi-variate Statistical description

(13) Make a full description of some pairs of data:

- '`price`' versus all other relevant variables (continuous and categorical)
- all other continuous variables, with each other (mays include correlation analysis and scatter plots)
- all categorical variable, with each other (under contingency tables)

(14) Select any 2 categorical variables and any 2 continuous variables (other than `price`):

- make the bi-variate analysis of all the 4 combinations (2*2) with data viz
- for each summarize important values and
- for each provide appropriate comment on the key information to be retained

Statistical Analysis :

Selection of variables

(15) As a prerequisite, remove any variable that is determined (caused) -even partially- by the `price`. Indeed, a model would not be relevant if the explanatory variable is the **result** of the explained variable.

Explain what variable your dropped and why.

(16) Select all **continuous** variables and **check whether the price can be predicted** by them.

- What method to select the best explanatory **continuous** variables ? (explain 'best', explain 'method')

At the end:

- Rank the explanatory **continuous** variables, from most to the least explanatory
- Retain the 3 "**best**" explanatory variables
- **Give an textual explanation and *formula* about your model.**

Any comment based on a figure should be the result of a *computation.*

Any computation, model calculation shall be based on piece of *code* provided along.

(17) Select each of the **categorical** (discrete) variables, tell whether **it explains the `price`.**

Please consider the following questions:

- retain only the meaningful explanatory **categorical** variables (with an explanatory power)
- and eventually group the similar values (modalities) ; explain how and why you grouped the modalities of the **categorical** variables.

At the end:

- Rank the explanatory **categorical** variables from most to the least explanatory
- Retain the 3 "**best**" explanatory **categorical** variables
- Explain what "**best**" explanatory **categorical** variables means here (consider the p-value of the impact of the variable and modality).
- **Give a textual explanation and *formula* about your model.**

Any comment based on a figure should be the result of a *computation*.

Any computation, model calculation shall be based on piece of *code* provided along.

Development of linear models

(18) As a prerequisite , split the data set in 2 part :

The table to be considered is still the unique table coming from the main table, with enrichment done in step 4/5/6/7/8 and filtering at step 10.

- `data.test` with the 1/3 of the data set. (please choose the lines with `ref_part1(modulo 3) = 0` ; eg 5055 is selected, 1023 is select, 1012 is not selected, 5044 is not selected)
- `data.train` with the remaining 2/3 of the data set.

The creation of the various models will use the `data.train`. The measure of the model performance will use the `data.test`.

(19) Based on the results of previous section (selection of continuous & categorical variables):

Make a **combined model** with both selected *continuous* and selected *categorical* variables.

Please consider both metrics (R^2 and RMSE) for the model comparison.

Among the expected task, please consider the following questions:

- Give a textual explanation and formula about your model (model_combined).
- Evaluate and comment the R^2 and RMSE of your model and any other performance measure (R^2 and RMSE on data.train, R^2 and RMSE on data.test).

Compare this **combined model** with:

- a model where only the continuous variables are used (model_cont)
    - give the formula, the R^2 and RMSE-or other- on data.train and the R^2 and RMSE on data.test
- a model where only the categorical variables are used (model_cat)
    - give the formula, the R^2 and RMSE-or other- on data.train and the R^2 and RMSE on data.test

On which parameter shall you decide with model to use ?

At the end, what model do you choose ?

Model assumption verification

Reminder on the linear model assumptions : see Here

(20) Can you tell something about the assumptions of your linear model (your chosen model) in relation to the current data set (validation and invalidation of assumptions)? Among the expected task, please consider the following questions:

- Text explaining the assumption
- Code to evaluate the assumption (statistical test, graphical output)
- Text explaining the result of the code and a conclusion on the respect of the assumption

Check at minimal:

- the normality of residual for the chosen model and
- provide some data visualization on the homostedaticity.

Conclusion : select the best investment opportunity

Simple investment question

(21) Considering that the model you have chosen is a better (more accurate) estimator of the value than the "market" price (i.e. the price offered), we want to use the model to determine which are the most profitable investments in terms of prices haircut.

Example :

If your model indicates that the predicted price for house in line 1 is `500.000 €` and the offered price is `450.000 €`, there is a haircut of `50.000 €`, that is to say `9%` expectancy of gain (`50.000/450.000`). This value can be called the **predicted profitability rate**.

Here is the situation :

Your company has 1M€ to invest, the question is (based on your model and on the data set limite to Grand-Paris prepared upper) :

“Please list the investments (up to our investment capability) you would recommand to make the best bargain (in terms of profitability rate, as defined upper). you should consider the data set for the Grand -Paris for this selection (not only data.train or data.test)”

In terms of methodology if an amount remains available there are 2 options :

1. Pick the next possible investment (price offered below your remaining)
2. If no asset with positive profitability is available, keep the cash.

Expected results:

- List of profitable investments, ranked by predicted profitability rate
- Estimation on the potential gain the company can make when the 1M€ is fully invested (there is no gain on cash).
- Comment the results.
- Display the preferred picture (from the possible picture in reference in file table_id_photos.csv) of the first best investement opportunity you selected.