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 (generated/knitted by the qmd/rmd).
- Input files (if modified)
- Output files (if any)

Normally, no input file 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 your project you should check for outliers, remove row if needed (In case of NULL/UNDETERMINED values, etc) preferably by coding.

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

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

The single data set to be used is the result of web-scrapping.

This data file is fake and 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 based on the data provided. Beyond understanding the data structure, providing key knowledge on the data within the set, you should model 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?

Import and qualification of the data

Import and correct the data set

(1) Import the data set , and check that the character with accent is readable in the data frame

(2) The data set contains multiples duplicated rows : write the code to remove them. You should end with a half reduced data set. In the following we will consider this data set with no duplicates.

(3) For our information : the `code.postal` has being used, by extracting the 2 first digit, to make a new variable `departement`

Do the same to `ref`, in order to split the reference in new columns

```
- ref_part0 <- substr(bien$ref, 1,6)
- ref_part1 <- substr(bien$ref, 7,10)
- ref_part2 <- substr(bien$ref, 12,15)
```

Discuss the relevancy of the different parts created with regards to our goal.

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

Obviously, the file in not in a normalized form : the objective here is provided a normalized (3FN) model of the data *behind* the file

(4) Identify the columns which are/look dependent from other columns

Those can be considered as separated "entities" within a model

(5) Identify the column which are/look composed of multiple individual element (including in other columns)

Those can be split (from a conceptual model position) and each shall be a unique "attribute" or "entity"

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

Uni-variate Statistical description

(7) As a prerequisite, list all data your are going to describe and determine which are

- **continuous data** (from numeric) : are there some numerical data that should be considered as categorical ?
- **categorical data**
- Free text (non categorical data) -> to be ignore in the rest of the exercise.
- any other , please precise if it's relevant or not.

(8) Make a full description of each individual data by

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

Adapt you graph, the geometry of the plot to the data your are plotting. **Do not build** non-sense plots (such as bar-plot for continuous data or scatter plot for categorical data).

Bi-variate Statistical description

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

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

(10) Select any 2 categorical data and any 2 continuous data (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

(11) 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.

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

Please answer the following questions:

- Which **continuous** variable are eligible as explanatory variables ? Why ?
- How to select the best explanatory **continuous** variables ? (explain 'best', explain 'method')

At the end:

- Rank the explanatory **continuous** variables
- 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.

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

Among the expected task, please consider the following questions:

- retain only the meaningful explanatory **categorical** variables
- 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
- Retain the 3 "**best**" explanatory **categorical** variables
- Explain what "**best**" explanatory **categorical** variables means here (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

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

- `data.train` with the first 2/3 of the data set.
- `data.test` with the last 1/3 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`.

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

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

Among the expected task, please consider the following questions:

The metrics

- Select a metric for comparing all models below (R^2 or equivalent)
- Give a textual explanation and formula about your model (model_combined).
- Evaluate and comment the R^2 (or any other appropriate metric) of your model and any other performance measure (R^2-or other- on data.train, R^2-or other- 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-or other- on data.train and the R^2-or other- on data.test

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

(15) Can you tell us 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

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 EUR` and the offered price is `450.000 EUR`, there is a haircut of `50.000 EUR`, 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 1 million euros to invest, the question is (based on your model and on the data set provided) :

“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).”

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.