UNIVERSITÉ PARIS 1
PANTHÉON SORBONNE

MASTER II TIDE

RAPPORT DE PROJET

X

⛩ JAPAN
HOSTEL 🛏
DATASET

Gefflot - Mostamandy - Pénichon
2022 – 2023

1- DATA

We have a dataset "Hostel.csv" with 342 observations and 16 variables, the data being related to hostels located in Japan. This dataset contains about 4% of missing values.

We excluded the index variable *'v1'* from our study and transformed the distance variable by removing the characters to keep only the numeric part of interest. We standardized all the data by putting them in lower case for the sake of convenience in handling the data. We also decided to remove two outliers that were likely to distort our results in terms of visualization but also modeling. Indeed, the prices of these two hostels were extremely high compared to the others. We believe that these are either luxury hotels mixed with the hostels (thus by mistake), or a pure error on the prices of these two hostels (after checking on the net, we think this is an error because their prices are inconsistent relatively to the average prices of hostels in Japan). We therefore worked with a dataset of 340 observations instead of 342.

2- OUR METHODOLOGY

We segmented our work into several parts:

- ✓ a general descriptive analysis to explore the dataset (using the *summary*, *describe*, *df_status*, *glimpse* commands) and thus identify: character/numeric types, number or proportion of zeros, number or proportion of missing values, number or proportion of outliers
- ✓ data standardization (types, upper/lower case)
- ✓ analysis of each variables independently via the creation of a ***quant_analysis*** function taking a variable as input and giving as output a histogram, a boxplot, a summary of descriptive statistics and a density curve
- ✓ processing : treatment of missing values and outliers
- ✓ after processing : analysis of each variable independently of the others
- ✓ global analysis (relevant connections of variables in the form of appropriate graphs taking into account the type of quantitative/categorical variable) via the creation of functions: ***boxplot_groupby*, *best_city*, *choice_city*, *b_h_i_c*, *in_spider_chart* and *all_spider_chart*, *corr_scat_plot* and *corr_plot***
- ✓ to go further : significance and independence tests between variables and supervised learning models to predict hostel's prices
- ✓ layout of the *Shiny* application and integration of functions

➢ Treatment of missing values

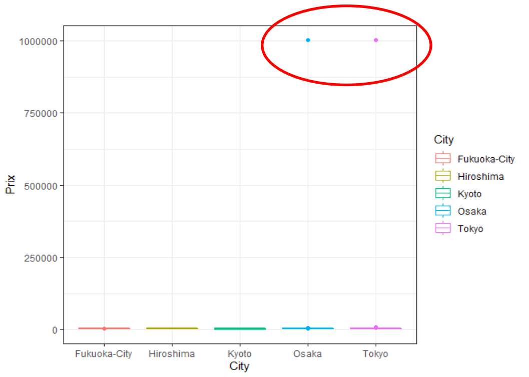
From our first analysis results, we noticed that some values were missing, using the *df_status* command, or that we obtained a partially empty correlation matrix.

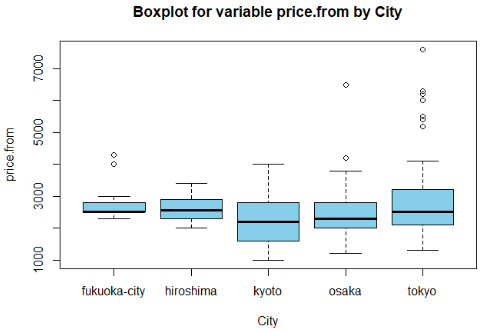
We imputed the missing values for *summary.score* by its median (grouped by *City* and *Distance* for greater precision), and we imputed all the missing values for the different scores by the median of the observations for that score (grouped by *City* and by *Distance* also). Finally, we had missing values for *rating.band* which is a categorical variable. This one being correlated to *summary.score* (a rating corresponds to a slice of quantitative score), we therefore referred for each missing value to the *summary.score* of the concerned observation. We also had missing data on the geographical data (*lon* and *lat*): we imputed by the median of *lon* and *lat* of hostel grouped by cities and by rounded distance (except for two where we grouped by city only because there were missing values for them).

Finally we created a file of the dataset after imputation *hostel_imputation*, then we worked with it for the rest of the analysis.

➢ Treatment of outliers

In the same way, we had noticed the presence of outliers from our first results. Indeed, the graphs were unreadable because they were crushed on values with extremely large scales of abscissa and ordinate (exponential).





Graphique 1: price by city boxplots with extreme outliers *Graphique 2 : price by city boxplots without extreme outliers*

There are two options for dealing with outliers: correcting them or removing them. Deleting outliers can lead to the deletion of relevant values. We decided not to delete or impute our outliers but rather to set a limit on our graphs using the neighborhood of the affected observations to make them readable and because we lack the information to be able to remove or correct the outliers.

In order to detect outliers of the variables *lon* and *lat*, we decided to make a comparison to the minimum latitude (123) and the maximum longitude (154) of Japan and thus to remove the hostels which would not be located in Japan (the hostels having a latitude lower than that of Japan, which is impossible for a hostel in Japan). We have also removed the outliers for the price because we thought it were mistakes (as we previously said).

3- OUR ISSUE

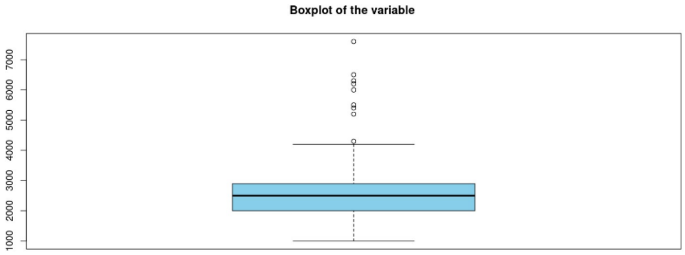
Faced with our dataset, we asked ourselves the question of what the end user of the application would like to achieve. A person looking for a hostel generally wants to maximize its value for money, thus maximizing its score (while having the choice of distance from the city center, which is subjective) and minimizing its price. A user may also pay more attention to some scores than others, but will almost always look for the highest rated hostel possible given his budget and criteria. The user may also either already have a destination city in mind, or choose its city based on amenities or otherwise.

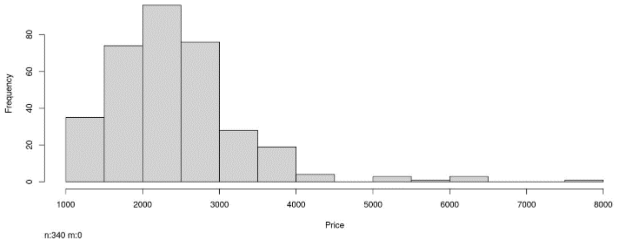
Therefore, we put ourselves in the end user perspective and we concluded the following problematic: How to optimize the user's choice given his budget?

4- OUR ANALYSIS

All the descriptive statistics elements below allow you to obtain the dispersion of values for each variable in different ways.

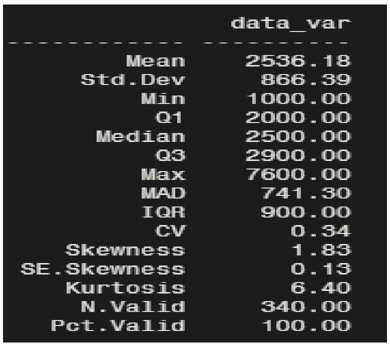
- The boxplot represents one or more boxes (depending on whether you categorize the analysis) that allow you to visualize the extreme values (at the end) as well as the quantiles (first and third one, median)
- The histogram in which each bar corresponds to a number of people for each value
- The Kernel density curve provides a smooth representation of the values distributions
- Summary statistics





Graphique 4 : price boxplot

Graphique 5 : price histogram



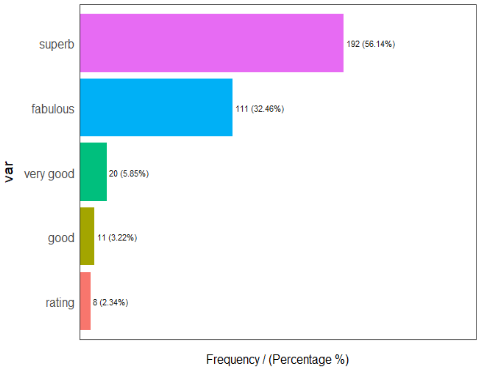
```
                  data_var
-----------  ----------
       Mean     2536.18
    Std.Dev      866.39
        Min     1000.00
         Q1     2000.00
     Median     2500.00
         Q3     2900.00
        Max     7600.00
        MAD      741.30
        IQR      900.00
         CV        0.34
   Skewness        1.83
SE.Skewness        0.13
   Kurtosis        6.40
    N.Valid      340.00
  Pct.Valid      100.00
```

Graphique 6 : statistic summary

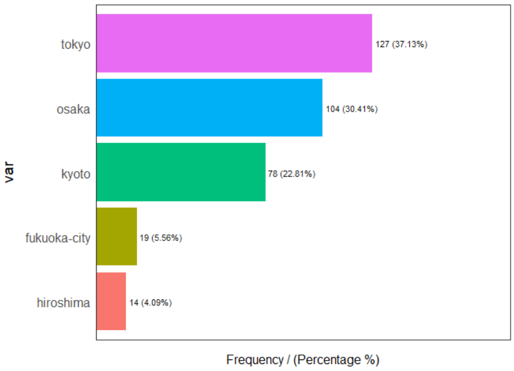
Here we can see how the variable *price.from* is distributed in several ways. We can see in general that it is unevenly distributed - concentrated towards the left (lower values). We can see that the median is at 2500. The skewness coefficient is positive here which tells us that the distribution is spread to the right (where we find the outliers). We can also see this inequality of distribution in the difference between the mean and the median visible thanks to the summary table. The kurtosis coefficient in this case tells us that the distribution is flatter than the normal distribution (due to positive coefficient).

➢ Categorical variables

It seemed relevant to establish graphs representing the frequency by modality.



Graphique 7: ratings frequency

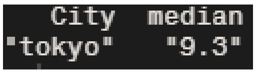


Graphique 8 : cities frequency

We can see that the majority of the hostels are rated as “superb”. Indeed, more than half percent of the hostels have a score above 8 out of 10. We also can see that the majority of the hostels are located in Tokyo, which seems coherent because it is the capital and it receives many tourists, whereas the minority is located in Hiroshima.

➢ Crossed categorical and quantitative variables

- Output for the best city based on a specific score (see our *Shiny* app – recommendation tab)

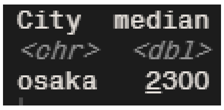


```
   City   median
"tokyo"    "9.3"
```

For example, if somebody wants the city with the best cleanliness, it would be Tokyo with a median score of 9.3.

Graphique 9 : best city output example

- Output for the median budget by city



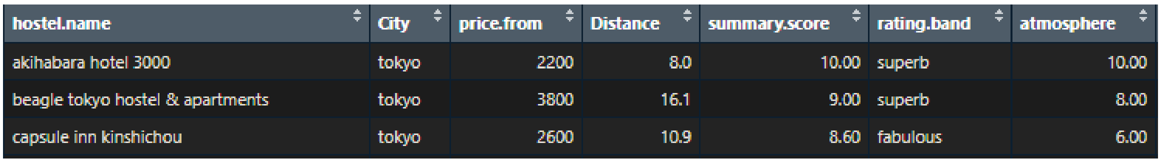
```
City   median
<chr>   <dbl>
osaka    2300
```

For example, the median budget for Osaka is 2300.

Graphique 10 : median budget by city output example

- Output for the function *b_h_i_c* (from the best to the worst hostel for a specific city)

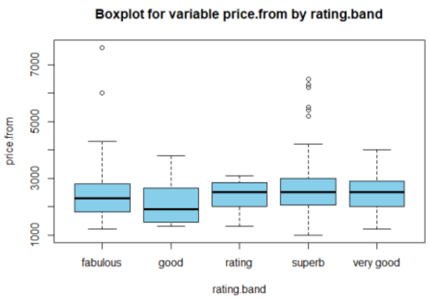
- Output for the function *b_h_i_b_c* (from the best to the worst hostel in the best city according to a specific score)

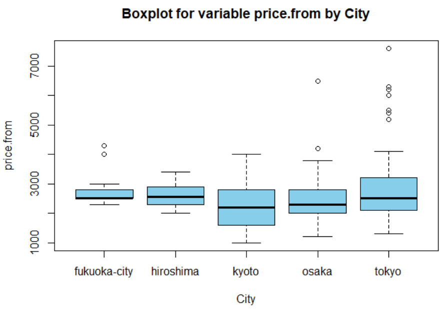


hostel.name	City	price.from	Distance	summary.score	rating.band	atmosphere
akihabara hotel 3000	tokyo	2200	8.0	10.00	superb	10.00
beagle tokyo hostel & apartments	tokyo	3800	16.1	9.00	superb	8.00
capsule inn kinshichou	tokyo	2600	10.9	8.60	fabulous	6.00

Graphique 11 : ouput for the b_h_i_c and b_h_i_b_c functions

- Boxplots





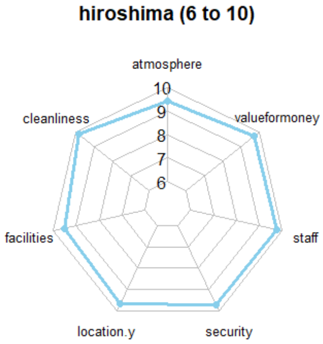
Graphique 12: price by rating boxplots

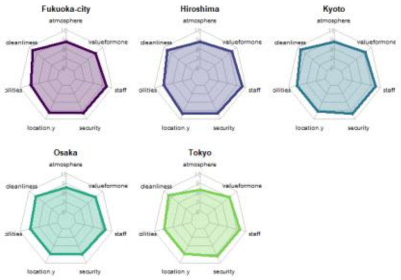
Graphique 13: price by city boxplots

We can see that there is no obvious link between rating and price which is surprising. Indeed, the median price for *rating* rated hostels is higher than the median price for *fabulous* rated hostels for example. We can think that people who pay more have higher expectations and thus rate the hostels more harshly. Nevertheless, we can still see that the distribution of the price for *fabulous* rated hostels is more spread out (toward higher values) than for *rating* rated hostels.

We can see on the second graph that in some cities the hostels are more often expansive like in Tokyo (which is the capital) whereas in some others the hostels are more often cheap.

- *Kiviat* diagram : detailed scores for each city





Graphique 15 : spider charts for all cities

Graphique 14 : spider chart by city

This graph is relevant in order to see how to explain the score for a city. For example, we can see that the score that pulls the average score up for hostels in Hiroshima is cleanliness, whereas what pulls it down is more the atmosphere one. On one hand, we could have change the scale for these graphs in order to make them more precise but on the other hand there are few fluctuations.

➢ **Crossed quantitative variables**

- Plots to see the correlation between two variables only (*summary.score* x *price.from* & *distance* & *price.from*)

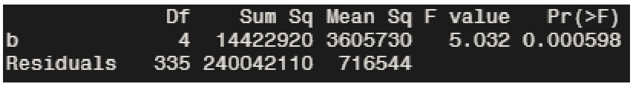
The relation between the distance and the price is not strictly linear. The majority of the hostels are located less than 20km from the city center and the majority have a price below 4000. Nevertheless, we can see that the more expansive hostels are closed to the city center.
The correlation is negative but not significant (p-value=0.48).

- Correlation matrix between all quantitative variables

5- EXPANSION (TESTS AND MACHINE LEARNING MODEL)

- Independence and significance tests

- Significance test

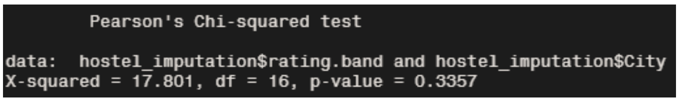


```
             Df    Sum Sq Mean Sq F value   Pr(>F)
b             4  14422920 3605730   5.032 0.000598
Residuals   335 240042110  716544
```

Graphique 16: significance test output example

This test allows us to see if there is a significant difference in price between the cities. We can see from the p-value that the difference is significant (at 1%). We can link this test to the boxplot showing the relationship between price and city (see above). Nevertheless, we cannot bring any clear conclusion with the graph (even if we see that there is a clear difference between Tokyo and Kyoto for example).

- Independence test



```
        Pearson's Chi-squared test

data:  hostel_imputation$rating.band and hostel_imputation$City
X-squared = 17.801, df = 16, p-value = 0.3357
```

Graphique 17:independence test output example

This test allows us to see if there is independence or not between the rating and the city. We can see from the p-value that we do not reject the H0 hypothesis (that there is independence) at 5%. There is therefore no significant link between the rating and the city. The same conclusion can be found in the graph (in boxplot or scatterplot, see above) showing the link between the rating and the price.

- Machine learning models - supervised learning: linear regression, Ridge regression, Lasso regression, Elasticnet regression, SVM regression

It seemed appropriate to use machine learning to predict the prices of hostels in Japan. Indeed, we have the price data as well as the data related to all the explanatory variables of this price. As price is a continuous variable and as we have the price data, we thought of a supervised learning model (the training is done on the price data) of the regressive type.
Before building the model, some preliminary steps are necessary:

- ✓ Imputation of missing values and outliers
- ✓ Encoding (*One Hot Encoding* for *City* which is a categorical variable and manual encoding for *rating.band* to keep the hierarchy of values)
- ✓ Normalization of the data (in our case of regressions) (after a *Kolmogorov-Smirnov* test to know if each variable has a normal distribution or not)
- ✓ Verification of correlations between variables before deleting a variable in case of strong correlation (for example we suspected a strong correlation between *rating.band* and *summary.score*)
- ✓ Variable selection : verification of correlations between the variables and the target *price.from* in order to better know which variables we should keep or not – removal of 8 variables
- ✓ Regularization : linear regression algorithm works by selecting coefficients for each independent variable that minimizes a loss function. However, if the coefficients are large, they can lead to over-fitting on the training dataset, and such a model will not generalize well on the unseen test data. To overcome this shortcoming, we do regularization, which penalizes large coefficients

Conclusion : The results are very bad. The R2 is very low (no more than 10%). This is mainly due to the fact that there are few explanatory variables for the price and that the ratings are too approximate (strong concentration towards the same values). Also, we had a small dataset with few variables and few observations.