INSTITUT NATIONAL DES SCIENCES APPLIQUÉES TOULOUSE

Department of Applied Mathematics

Data analysis: Vélib data

Analysis of the loading profiles of the Vélib stations in Paris

---

Juan Ayala

Jeong Hwan Ko

Alice Laloue

Aldo Mellado Aguilar

4th Year Applied Mathematics - Groups A&B

Date of submission: May 5th 2021

Contents

1 Descriptive Statistics **1**

1.1 Basic database descriptions 1

1.2 A first look at the behaviour of the loading through time 1

1.3 Elementary conclusions after descriptive analysis 3

2 Multidimensional Analysis **4**

2.1 Correlation between station loading and date 4

2.2 Loading profiles based on the altitude 4

2.3 Conclusions from multidimensional analysis 5

3 Principal Component Analysis **6**

3.1 Reducing data dimension: number of principal components 6

3.2 Interpreting the principal components 7

3.3 Conclusion of Principal Component Analysis 8

4 Clustering on the original data **9**

4.1 Ascending Hierarchical Classification 9

4.2 K-Means 10

4.3 Clustering of raw data conclusions 12

5 Clustering on the coefficients of a suitable functional basis **13**

5.1 Ascending Hierarchical Classification 13

5.2 K-Means 13

5.3 Gaussian mixture model 14

5.4 Clustering conclusions 15

6 Conclusion **15**

For this project, we have the data from the loading profiles of the 1189 Paris Vélib (public bicycles) stations, scattered around the city. The loading corresponds to the filling percentage of a given station: a loading of 1 means that the bike station is full, and 0 means that there are no bicycles left. The loading is measured every hour from September 2nd 2014 at 00:00 am to Sunday 9th at 11:59 pm. The goal of this project is to detect common customer behaviours, which we will call clusters. Our analysis will be done using **R** and **Python**.

1 Descriptive Statistics

1.1 Basic database descriptions

We have two data sets. The first one, called `velibLoading`, is the data set containing 168 variables and 1189 individuals. Each variable corresponds to the time of measurement of the station's loading, that is, one measurement per hour for 7 days. Every variable in this data set is quantitative.

The second one, called `velibAdds`, contains the same 1189 individual, in the same order, but this time it consists of the stations' $name, longitude$, and $latitude$, and a variable called *bonus* that indicates if the station is located on a hill (1), or not (0). The variable *bonus* and the coordinates of the stations are quantitative variables, and the name of the station is qualitative. Around 10.68% of the stations are on hills.

1.2 A first look at the behaviour of the loading through time

In order to get a general idea of the loading's behaviour, we start by plotting the loading profile's evolution for the first 16 stations on the list.

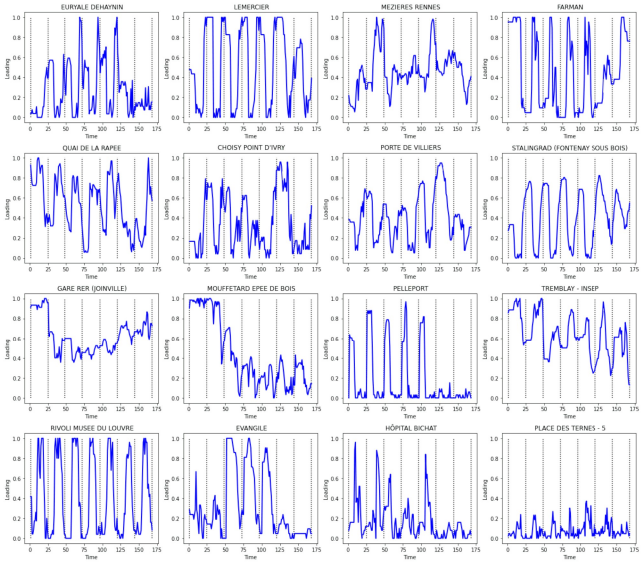

Figure 1: Loading profile evolution through time.

These graphs show that, in general, the loading is periodical, following a similar pattern each day. The seasonality is as following: a high loading at night, but lower during the day. This can be explained by the fact that the bikes are used more during the day to move around in the city, for instance, during rush hour, whereas during the night the activity is generally lower.

However, the popularity of the stations vary. For example, there is a clear difference in the loading profiles of the 12th and the 16th stations compared to all of the other ones that are plotted above, which we will explore further on.

Then we plot 168 box plots side by side corresponding to each variable, i.e. the time of the week. These box plots are also separated by blue vertical lines to represent the seven days of the week.

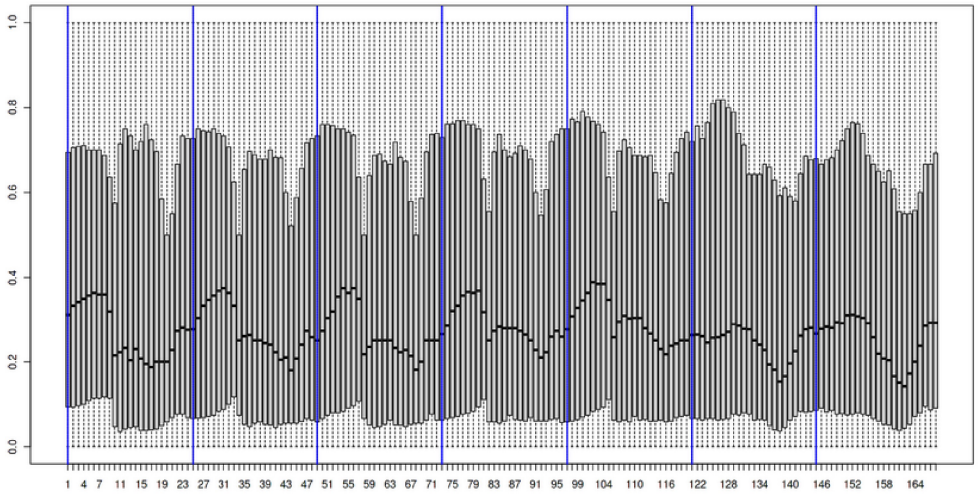

Figure 2: Loading through 7 days.

This graph shows that the median behaves a certain way during the weekdays, and a different way during the weekend. This reflects how bikes are used differently during the week days from the weekends. During the weekdays we see that the median reaches a peak at around 8:00 AM and then it suddenly drops. This might be due to the large amount of people who use the bikes to go to work or to school at that time, i.e. rush hour. During the weekend the median reaches a minimum value at around 5:00 PM, which might occur because of all the people who like to go out on the weekends. Also, in figure 2, we can notice that the loading on Saturday and Sunday looks very much alike.

Furthermore, we try to find any correlation between two chosen variables. Thus, we plot the loading of a given time versus the loading with a slight shift in time to see if there is a linear correlation.

Most of the variables are strongly correlated, except for those comparing the time couples (9, 10) and (10, 11). As explained above using the box plots, this drastic change from one hour to the next

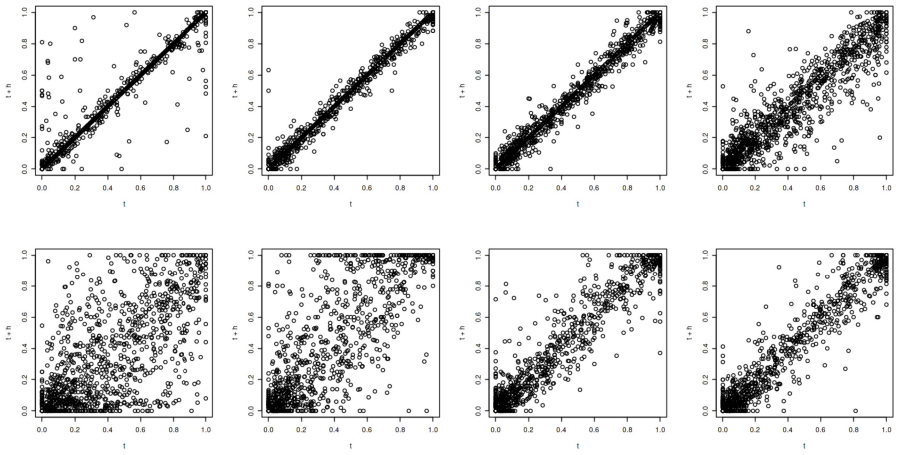

Figure 3: Difference of loading between two consecutive hours.

happens when most people go from sleeping to going to work.

1.3 Elementary conclusions after descriptive analysis

Our first analysis on the variables strongly indicates that, in general, the loading varies periodically through time: it is higher during night-times and lower during the day. Finally, the loading on the weekends has different properties: the behavior seems to be more random and erratic.

2 Multidimensional Analysis

2.1 Correlation between station loading and date

To visualize the linear correlation between all variables in `velibLoading`, we plotted the correlogram. Two variables will be strongly correlated if the absolute value of their correlation is close to 1, and the opposite for a correlation close to 0.

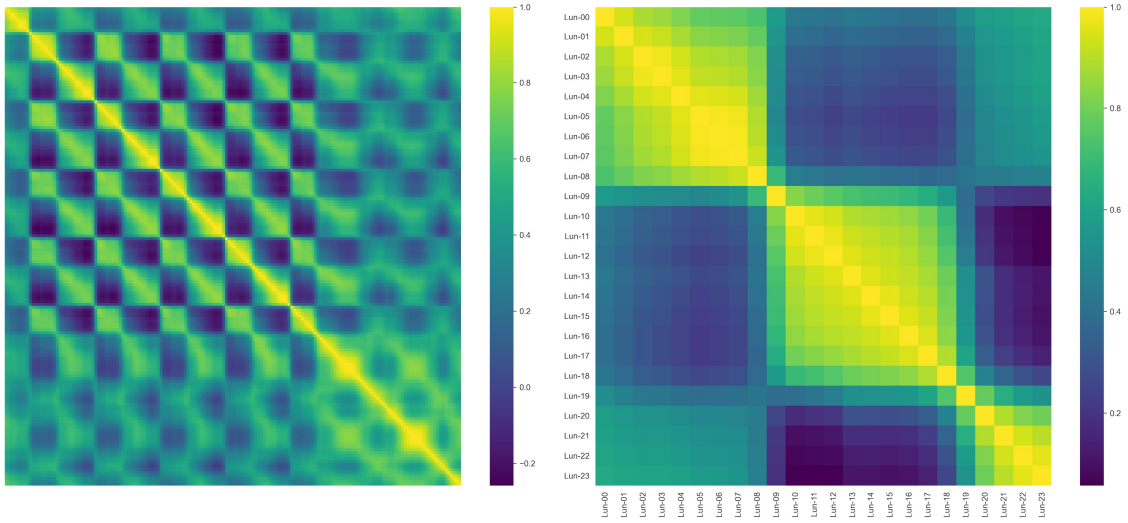

Figure 4: Left: correlation for all 168 variables. Right: correlation between the loading of all hours of the first day (zoom).

To analyse the correlation, we use the `corrplot` library from **R** to have a visual representation of the correlation between variables. The first figure corresponds to the whole week and the second one represents only the first day, Monday.

The correlation plot on the left shows that the weekend loading profiles are less correlated because the bikes are used less and more randomly during the weekend, in contrast to their high activity during the week. This is highlighted by the blurriness in the bottom right-hand corner

Moreover, by focusing on the first day (right plot), we can see that the values near the diagonal have a very strong correlation. We observe that day-time hours, i.e., hours between 8 AM and 6 PM, are strongly linked, which is why the color is very dark; just as nighttime hours, but these two hour groups are not correlated.

2.2 Loading profiles based on the altitude

Using the variable `bonus`, we can determine whether a station is on a hill or not. We first suspect that the usage will be lower than the stations that are not on a hill, assuming that customers do not want to ride a bicycle up a hill.

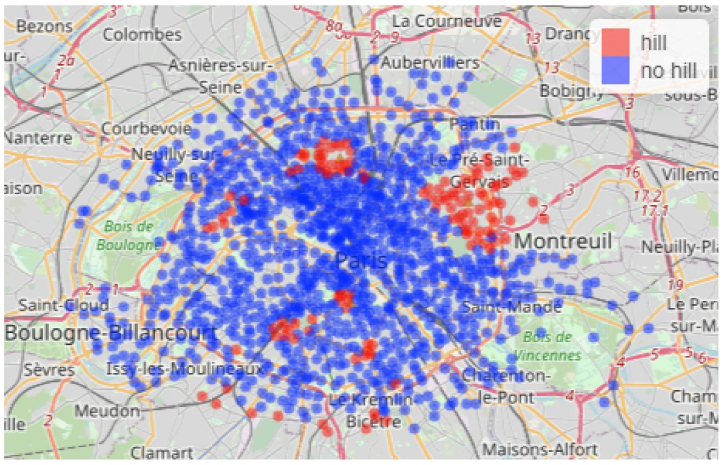

Figure 5: Plot of all stations in Paris.

In order to compare the loading profiles between these groups of stations, we use the box plots of the loading profiles throughout the week.

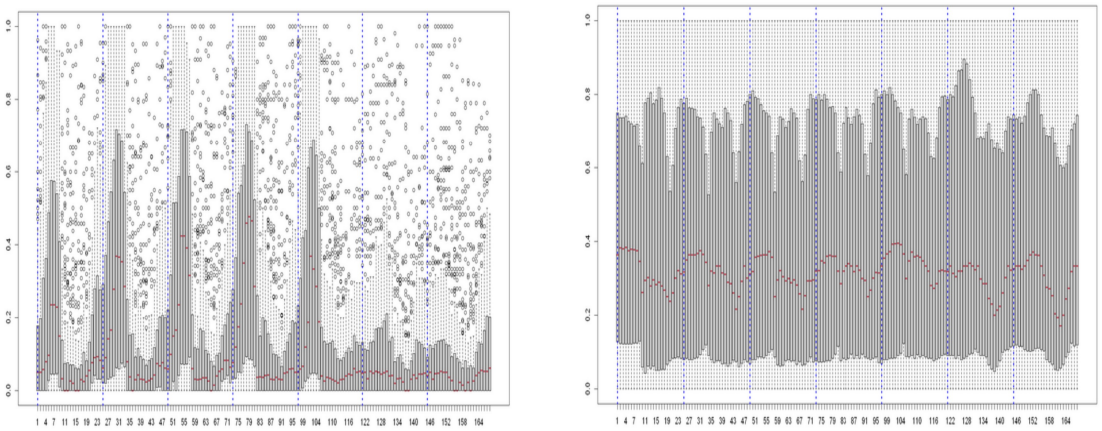

Figure 6: Loading through 7 days of stations on a hill and on plain terrain resp.

The first graph is not as clear as the second one since it has a large amount of outliers and not enough data (around 125 stations). This is due to the fact that the use of bicycles on the weekends is erratic. However, we can see that there is a certain pattern during the weekdays, and another one during the weekend.

2.3 Conclusions from multidimensional analysis

In this section, we studied the correlation between the loading profiles of the Vélib bike stations in Paris. We can see that the customer behaviour follows a certain pattern that corresponds to rush hour. Furthermore, the behaviour is more random on the weekends, but it still obeys a periodicity on both days. The same conclusion applies to the stations located on a hill.

3 Principal Component Analysis

Because we have 168 variables, we perform a Principal Component Analysis (PCA) to better understand their meaning and reduce the data's dimensions. Our data does not need to be scaled before performing PCA as the loading is between 0 and 1 and the dispersion is fairly constant. However, we chose to scale the data before performing PCA just to be sure.

3.1 Reducing data dimension: number of principal components

To determine the number of principal components to keep, we display the explained variance ratios versus the principal components computed.

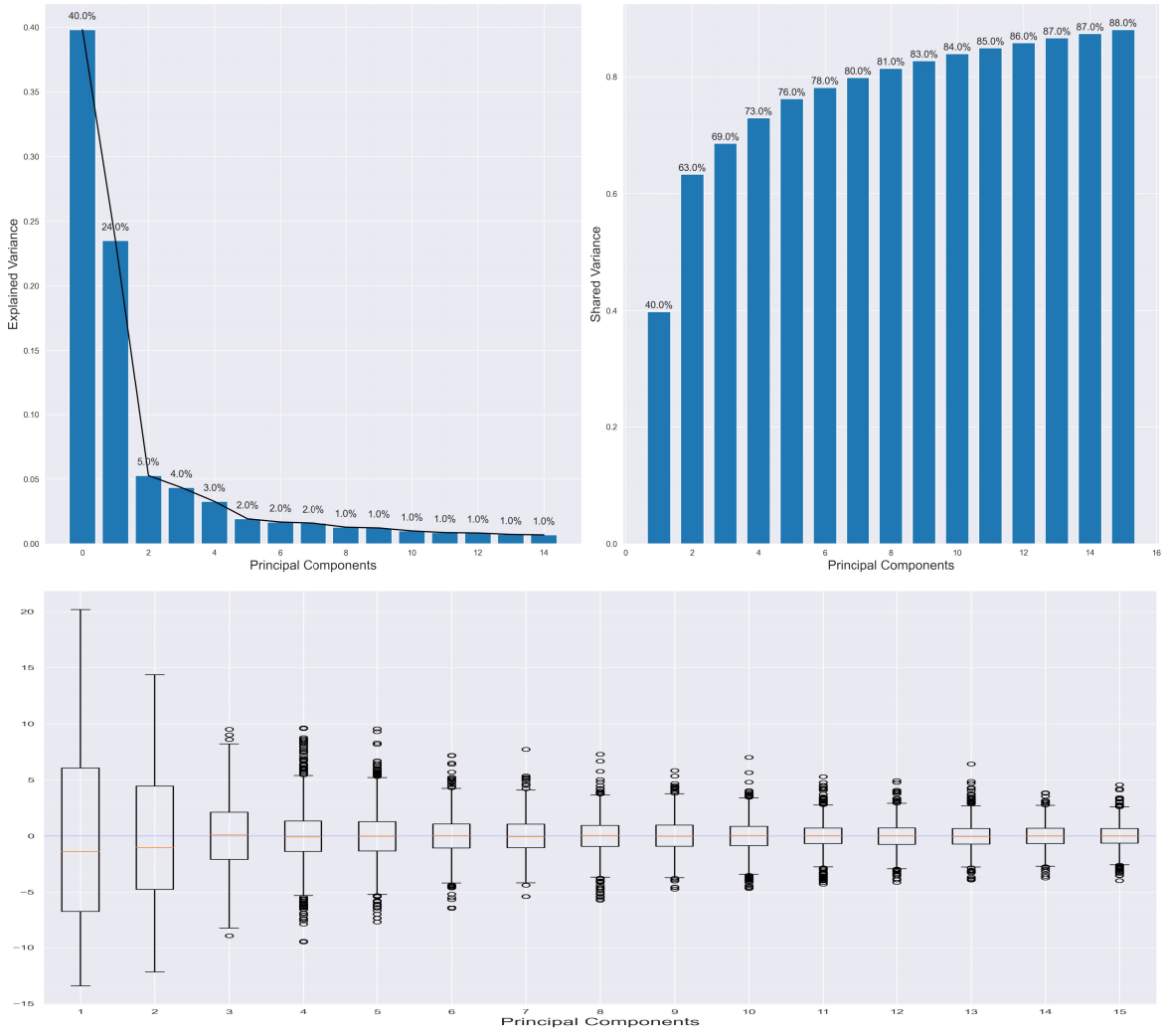

Figure 7: Top: Percentage of Variance explained by the first fifteen principal components of the PCA applied on the loading data. Bottom: Box plots of the first fifteen principal components.

As shown by figure 7, most of the variance is contained in the first two components (63%), but some important information still remains in some of the following components. Indeed, 80% of the variance is explained by the first seven components.

On the graph of explained variance ratios (figure 7, top left), the curve of explained variance ratios is bent over the sixth principal component.

Figure 7, bottom, shows that the first two components are the ones with the largest variance. After the sixth principal component, the explained variance is low and decreases very slowly. The medians are centered around 0 after the third principal component and the extent of the box plots is constant after the fifth principal component: the dispersion stabilizes.

Consequently, we chose to work with the first five principal components, given that after the fifth one, the accumulated sum of explained variance doesn't increase significantly.

3.2 Interpreting the principal components

The circles of correlation give a first interpretation of the first three principal components.

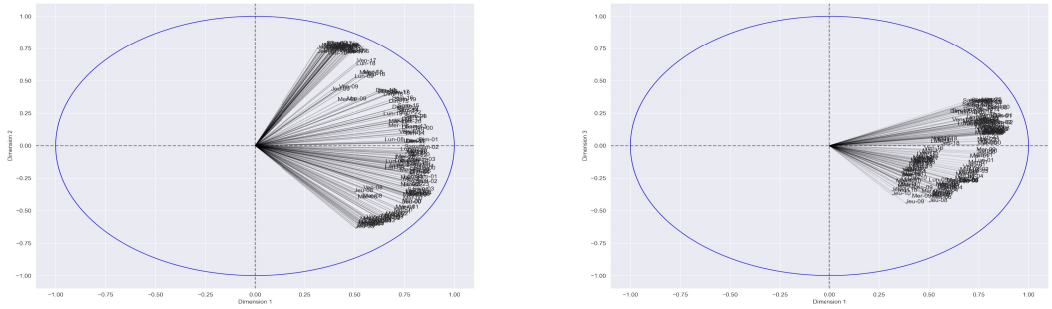

Figure 8: Left: Circle of correlation between variables on the first and second principal components. Right: Correlation on the first and third principal components.

First, on the left panel of figure 8, most of the arrows are close to the circle of correlation of radius 1, which means that most variables are close to their projections on the first two components. The correlation circle between the first and second components can be used safely.

All variables take positive values along the first dimension and the abscissas of most variables are contained between 0.4 and 0.6 which means that the first dimension corresponds to half the sum of loadings of the week. As for the second dimension, positive values tend to correspond to day-time hours (from 2 PM to 6 PM), while negative values correspond to nighttime hours (11 PM to 5 AM). The second dimension contrasts between daytime loading and nighttime loading.

On the circle of correlation between the first and third components, most variables with positive coordinates correspond to weekends, while most variables with negative coordinates correspond to weekdays. However, this analysis is mostly inconclusive because the ordinates of all the variables are low, which means that the correlation with the third dimension is not very strong.

Therefore, we can deduce that :

- The first principal component represents the average loading of the Vélib stations, which is a positive value between 0 and 1,
- The second principal component is the contrast between night and day,
- The third principal component corresponds to the difference between weekdays and weekends.

Additionally, on figure 9, the graph of individuals projected on the first two components depicts a potential cluster in the bottom left-hand corner mainly composed of bike stations located on a hill.

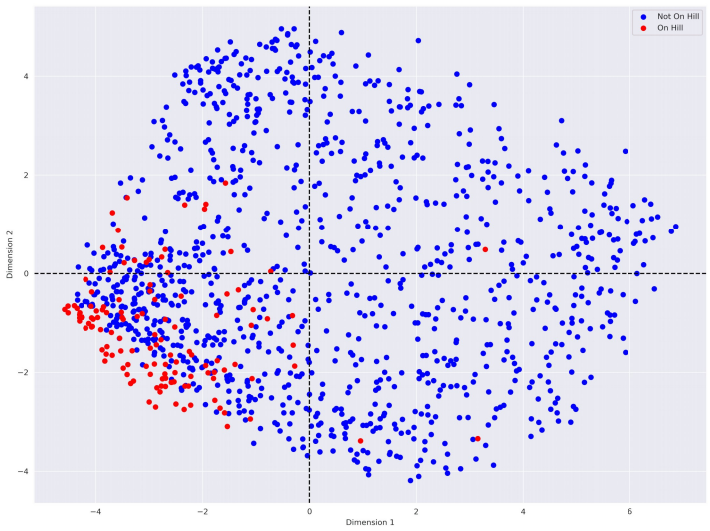

Figure 9: Individuals projected on the first two dimensions of the PCA computed of the loading data. In red (blue), Vélib stations (not) located on a hill (resp.).

3.3 Conclusion of Principal Component Analysis

Thanks to the PCA procedure, we found that the first axis of projection corresponds to the overall loading of the stations, and the second one to the variation in bike rentals between night and day. The third component could represent the distinction between weekdays and weekends, although further analysis is needed to back this claim.

4 Clustering on the original data

This section aims to detect loading clusters in the original data, corresponding to common customer use.

4.1 Ascending Hierarchical Classification

The Ascending Hierarchical Classification (AHC) method allows us to divide our data set (here, the loadings) into similar clusters, based on several metrics and methods. In our case, we use the `euclidean distance` metric and `Ward's` method. We explain why later in this section.

Firstly, we apply the AHC method on the raw data in order to detect clusters and to determine the best number of clusters to retrieve.

We apply the AHC algorithm using Ward's method and the euclidean metric: first, we chose Ward's method for clustering because it consists in evaluating the distance between clusters as how much the sum of squares increases when merging them. We want to keep this cost as low as possible. Then, we chose the euclidean distance because, in this case, it corresponds to the squared distance between two curves (station loading). This means that two similar loadings have a small area between their curves.

According to the center panel in figure 10, five or eight clusters can be retrieved from the raw data. Indeed, a first jump in the height is observed between seven and eight classes. The second jump is between four and five classes. Our results indicate that we should choose five clusters. When cutting the dendrogram at 5 classes (height around 40), their size seems large enough, and groups look fairly homogeneous. Only the third class appears to be much larger than the others.

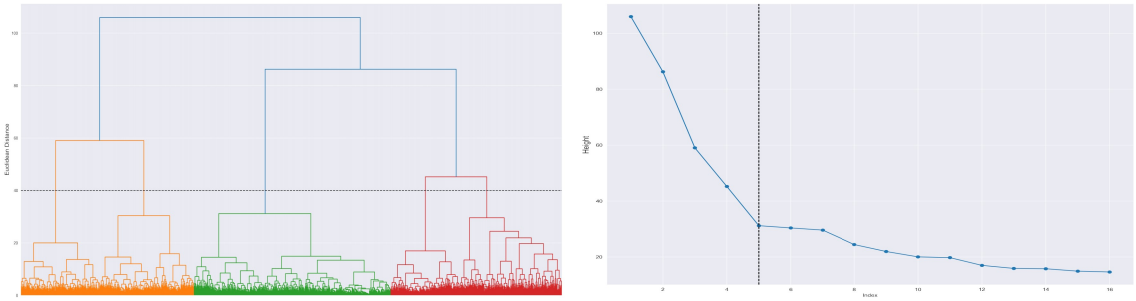

Figure 10: Results of hierarchical clustering (5 classes). Left: Dendrogram. Right: Height versus number of classes.

We describe the behaviour of the 5 clusters of Vélib stations on figure 11:

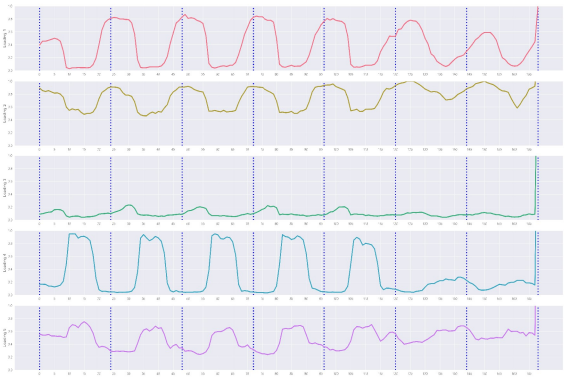

Figure 11: Results of hierarchical clustering: Mean loading of classes. Blue dotted lines are plotted on each day at midnight.

- Cluster 1: Unloading in the morning and loading in the evening. These stations are loaded at night and unloaded during the day.
- Cluster 2: Unloading in the morning and loading in the evening on weekdays (loading $> 50\%$). These stations are mostly completely loaded in the weekend.
- Cluster 3: Stations almost always unloaded (loading $< 20\%$).
- Cluster 4: Loading in the morning and unloading in the evening on weekdays. These stations are unloaded in the weekend.
- Cluster 5: Loaded in the morning and unloaded in the evening on weekdays only (loading $> 30\%$).

Compared to figure 2, the variability around each class center is reduced. Most classes are thus homogeneous, except for the second and fifth classes as shown by the box plots of each class.

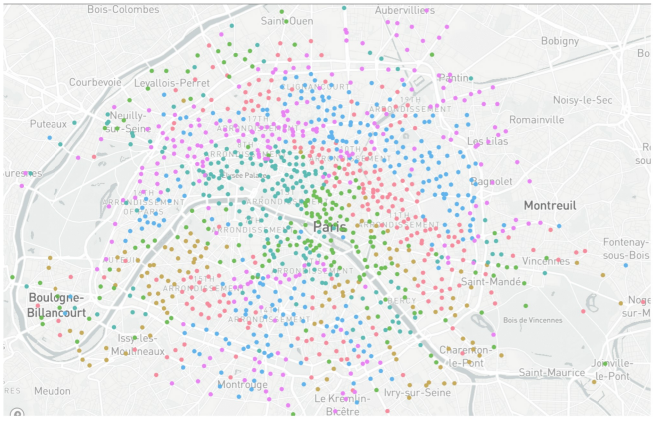

Figure 12: Results of hierarchical clustering (5 classes). Vélib stations coloured according to their class.

Thanks to our analysis, we obtain 5 clusters for the AHC method.

4.2 K-Means

Secondly, we use the K-Means algorithm to the raw data to detect clusters of Vélib stations. This method allows us to separate the data with straight lines.

According to the silhouette coefficients and inertia of classes on figure 13, we could choose either $k = 4$ or $k = 6$ classes. Indeed, almost all classes have a a silhouette coefficient above the mean silhouette coefficient, and they also have a similar width, meaning that they have a similar number of individuals. We chose $k = 6$ in order to have enough clusters to better represent Vélib stations. These figures allow us to have a basic idea of how many clusters to choose, but further research is needed, since this is only a visual method and could be erroneous.

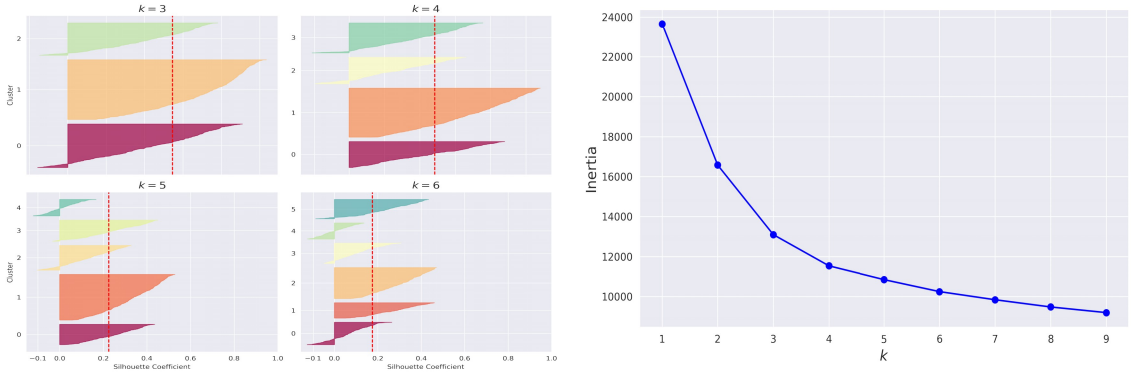

Figure 13: K-Means applied to the raw data. Left: Silhouettes of classes for $k = 3, 4, 5$ and 6. Right: Inertia of classes.

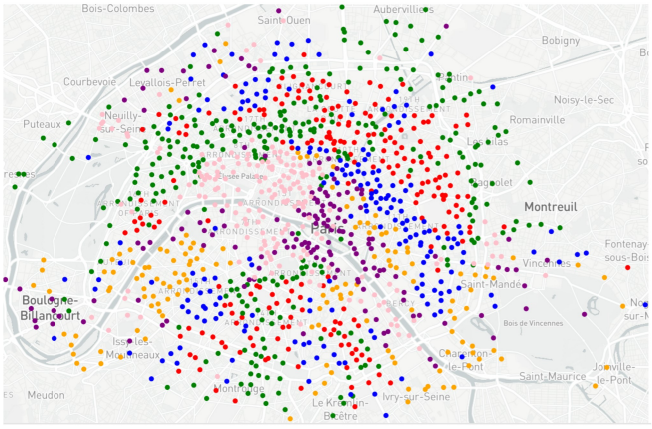

Figure 14: Clusters obtained with K-Means applied to the raw data with $k = 6$. Vélib stations coloured according to their class.

The figure 14 shows the Vélib stations coloured according to their respective cluster (total of 6 clusters). The plots on figure 15 depict the median loading of each cluster. We can observe that these are different from each other, which highlights the fact that the clusters are separated well. We describe more precisely the behaviour of the 6 clusters of Vélib stations:

- Cluster 1 corresponds to the Vélib stations loaded at night and unloaded during the day.
- Cluster 2 corresponds to the stations that are almost always completely loaded.

- Cluster 3 corresponds to the opposite of the previous cluster. These stations are almost always empty at every hour of the week.
- Cluster 4 corresponds to the Vélib stations loaded on weekdays in the morning and the evening but unloaded during the day. These stations are loaded in the weekend.
- Cluster 5 corresponds to the stations loaded during day and unloaded at night. The behaviour is the exact opposite of the one described for the first cluster. These stations are never completely unloaded (loading $> 30\%$). Cluster 5 corresponds to the Vélib stations on the banks of the Seine.
- Cluster 6 corresponds to the stations loaded during the day on weekdays and unloaded in the morning, evening, and night. These stations are almost completely unloaded in the weekend. The center of class 6 corresponds to an evening usage on weekdays and a weekend usage. This behaviour is the opposite of the fourth cluster. Cluster 6 corresponds to the tourist districts of Paris: the 1^{st}, the 7^{th} and the 8^{th} arrondissements.

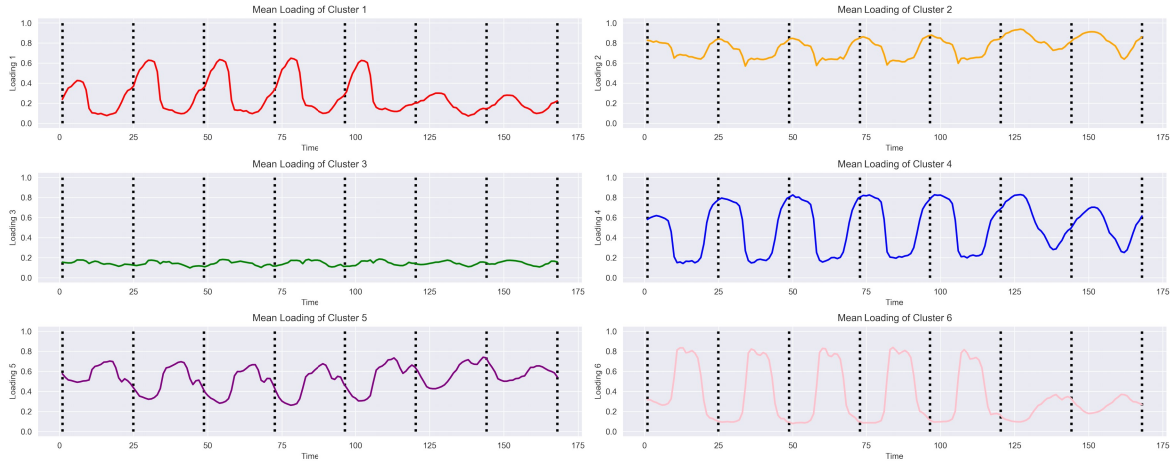

Figure 15: Clusters obtained with K-Means applied to the raw data with $k = 6$. Representation of the center of the class as a function of time: loading of the cluster barycenters. Black dotted lines are plotted each day at midnight.

Compared to figure 2, the variability around each class center is reduced. The K-Means clustering method reduced the within-class variance. Most classes are, as such, homogeneous, except for the second and fifth classes or weekends in the 6th class, where the errors are larger.

4.3 Clustering of raw data conclusions

Thanks to our analysis, we obtained 5 clusters for the AHC method and 6 clusters for the K-Means method. As we said at the beginning of this section, these clusters correspond to similar customer behaviour, loadings and uses of Vélib stations. These clusters are dissimilar, so we can affirm that we found a proper separation of customer habits.

5 Clustering on the coefficients of a suitable functional basis

In this section, we use the same methods as before, but now on the projected data found through the PCA.

5.1 Ascending Hierarchical Classification

We apply the Ascending Hierarchical Classification method on the projected data using the PCA. We only use the first 5 principal components, since they are the ones that contain most of the information. Figure 16 depicts the results: we should select 6 or 4 clusters; however, to be consistent with the previous clustering method, we decided to keep 5 clusters.

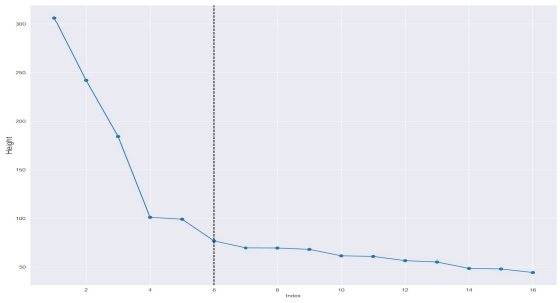

Figure 16: AHC applied to the projected data using Ward's method and the euclidean distance.

To compare these results with the ones corresponding to the raw data, we created a cross-table to see which individuals are in the same cluster for both methods, and which ones are in different clusters, as presented in figure 17. We observe that, overall, the clusters are almost the same, but there are large differences between clusters 3 and 4 between both methods. We can deduce that the AHC algorithm gives different results on the raw data and on the PCA results.

	0	1	2	3	4
classe...					
0	137	13	1	0	0
1	1	151	0	2	2
2	27	0	371	23	23
3	0	0	0	150	150
4	0	19	1	125	125

Figure 17: Cross table of AHC method applied to the raw and the PCA data.

We try applying the AHC to the PCA data and cutting the tree in order to get 6 classes. We observe that the 6 clusters obtained largely correspond to the 5 clusters obtained on the raw data. Only one cluster was divided in two, which has an impact on the variability: it is slightly reduced when applying the AHC to the PCA data.

5.2 K-Means

We continue using the first 5 principal components, and we apply the K-Means method to the data to find 6 clusters. According to figure 19, the best number of clusters to choose could be 5 or 6. We choose $k = 6$ in order to compare both K-Means clustering methods by using a cross table (figure

18), and we conclude that there is no significant difference between the K-Means method applied to the raw or to the PCA data.

	0	1	2	3	4	5
classe...						
0	146	0	0	0	0	0
1	0	292	0	0	0	5
2	5	0	147	0	0	0
3	2	2	0	185	0	1
4	0	0	0	0	186	0
5	1	2	0	0	7	208

Figure 18: Cross table of K-Means applied to raw and PCA data.

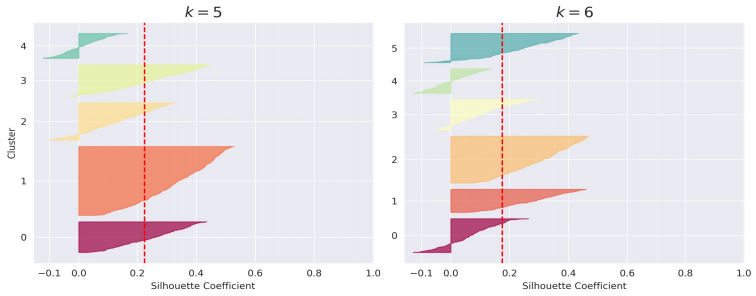

Figure 19: Silhouettes of K-Means applied to the PCA data.

5.3 Gaussian mixture model

Finally, we use the Gaussian mixture model to have a better separation of the data. This method classifies data by its probability of belonging to a certain cluster. It also allows us to separate the data with curves, instead of straight lines, and it can adapt to different shapes of clusters, be it a sphere, an ellipse, or another shape.

As with the other clustering algorithms, we decided to separate the data into 6 clusters. To do so we have to work on two dimensions, so we use the data projected into the plane formed by the first two principal axes. As for the covariance matrices, we use the default method, which does not apply any constraints to the matrices, therefore each cluster can take any particular shape and size. These clusters are shown in figure 20.

Even though we had already established that we would use 6 clusters, we could also find the best number of clusters and the most appropriate type of covariance by minimizing the BIC criterion. By doing so, we found that to minimize the BIC criterion we must use 8 clusters and the diagonal model (parallel ellipsoid clusters, but of different dimension).

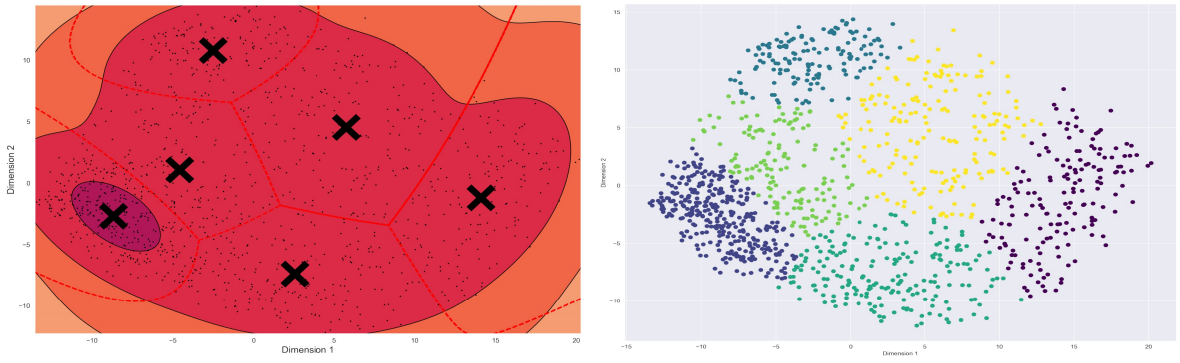

Figure 20: Clustering by Gaussian mixture model.

5.4 Clustering conclusions

After using the previous clustering techniques, we generally find that using 6 clusters is the most efficient way of classifying our data, even though different methods give a different number of clusters. For instance, the Gaussian mixture model provides 8 clusters in order the minimize the BIC criterion, whereas the K-Means method suggests 6. We also find that applying these clustering methods to the raw data and to the PCA data might yield different results. Regarding the K-Means method, we find basically the same result for both sets of data. However, there is a more noticeable difference between the clusters created with the two sets of data when we apply the AHC method.

6 Conclusion

In this project, we studied and used different data analysis methods to extract as much information as possible from the Paris Vélib stations data set.

Our first section gave us a first idea of the behaviour in certain stations, where our study was focused on the correlation between loadings. Then, using Principal Component Analysis, we highlighted the link between some of the variables, which corresponded to the loading during night and day, but also during weekdays and weekends.

The main study of this project was using the clustering methods that we put into practice in our last section. Indeed, we learned that our previous analyses were correct and we found matching behaviours in the loading of the stations.

We would like to acknowledge our gratitude towards Prof. Dr Olivier Roustant for his supervision and advice throughout the project and its write-up. We would also like to thank Mrs. Camille Champion and Mr. Aimen Zerroug for their valuable knowledge on data analysis that we were able to use in this project.

Our code is available on our GitHub repository.