The Seoul bike sharing service
Factor analysis and clustering

Hugo Cornet, Pierre-Emmanuel Diot, Guillaume Le Halper, Djawed Mancer

Contents

Introduction

The first part of our analysis aimed to have a global view on the `SeoulBike` dataset. We used the tools of both descriptive and frequentist statistics in order to find the dataset's main patterns. You will find the main results of the descriptive statistics part in the appendix.

Now, let's dive into data so as to explain the links between the variables more precisely. This part of the analysis aims to explore the results found in the previous part.

The report will be divided into two main axes which are factor analysis and clustering. The former's goal is to reduce a large number of variables into fewer numbers of factors while extracting maximum common variance from these variables. Regarding the latter, it is used to group similar individuals into a limited number of classes which are created by different kind of algorithms.

Code automation

We have automated several functions with the view of simplifying the way our code is built. Among them you can find `mykable` for tables' layout, `ctr_viz` and `point_cloud` for visualizing the principal component analysis' results and `first_look` for the multiple correspondence analysis' visualization.

Importing the dataset : 'SeoulBikeData.csv'

Table 1: Overview of the 'SeoulBike' data frame

Date	Rented.Bike.Count	Hour	Temperature..C.	Humidity...	Wind.speed..m.s.	Visibility..10m.
01/12/2017	254	0	-5.2	37	2.2	2000
01/12/2017	204	1	-5.5	38	0.8	2000
01/12/2017	173	2	-6.0	39	1.0	2000
01/12/2017	107	3	-6.2	40	0.9	2000
01/12/2017	78	4	-6.0	36	2.3	2000
01/12/2017	100	5	-6.4	37	1.5	2000

The dataset's variables

The database is mostly made of quantitative variables which is good news for the principal component analysis. However we need qualitative variables to perform both the correspondence analysis and the multiple correspondence analysis. That's why a few qualitative variables will be created from the quantitative ones using the `class_cut` function, whose role is to cut quantitative variables into classes.

Table 2: Structure of the 'SeoulBike' data frame

col_name	col_index	col_class
Date	1	Date
Rented.Bike.Count	2	integer
Hour	3	integer
Temperature..C.	4	numeric
Humidity...	5	integer
Wind.speed..m.s.	6	numeric
Visibility..10m.	7	integer
Dew.point.temperature..C.	8	numeric
Solar.Radiation..MJ.m2.	9	numeric
Rainfall.mm.	10	numeric
Snowfall..cm.	11	numeric
Seasons	12	factor
Holiday	13	factor
Functioning.Day	14	factor

Principal Component Analysis (PCA)

Principal component analysis is a multivariate statistical technique that uses an orthogonal transformation to convert a set of correlated variables into a set of orthogonal, uncorrelated axes called principal components. The primary motivation behind PCA is to reduce a large number of variables into a smaller number of derived variables while preserving as much of the data's variation as possible.

In this part we will carry out PCA in order to find principal axes which summarize the information contained in the `SeoulBike` dataset. Bear in mind we want to determine the meteorological variables that best explain the number of rented bikes. We expect PCA to make groups out of these variables into few principal axes.

First look at the quantitative variables

Standardized dataset

First of all, PCA cannot be performed if the quantitative variables are not scaled. Scaling the variables means centering each variable then dividing each of them by their respective standard deviation. The standardized matrix is defined as follows:

$$X_{st_j} = \frac{X_j - \mu_j}{\sigma_j}$$

where j is one variable of the dataset (e.g. j can be `Rented.Bike.Count`).

Table 3: Overview of the standardized variables

Rented.Bike.Count	Hour	Temperature..C.	Humidity...	Wind.speed..m.s.
-0.74	-1.66	-1.48	-1.03	0.46
-0.82	-1.52	-1.51	-0.98	-0.90
-0.87	-1.37	-1.55	-0.93	-0.70
-0.97	-1.23	-1.57	-0.89	-0.80
-1.01	-1.08	-1.55	-1.08	0.56
-0.98	-0.94	-1.58	-1.03	-0.22

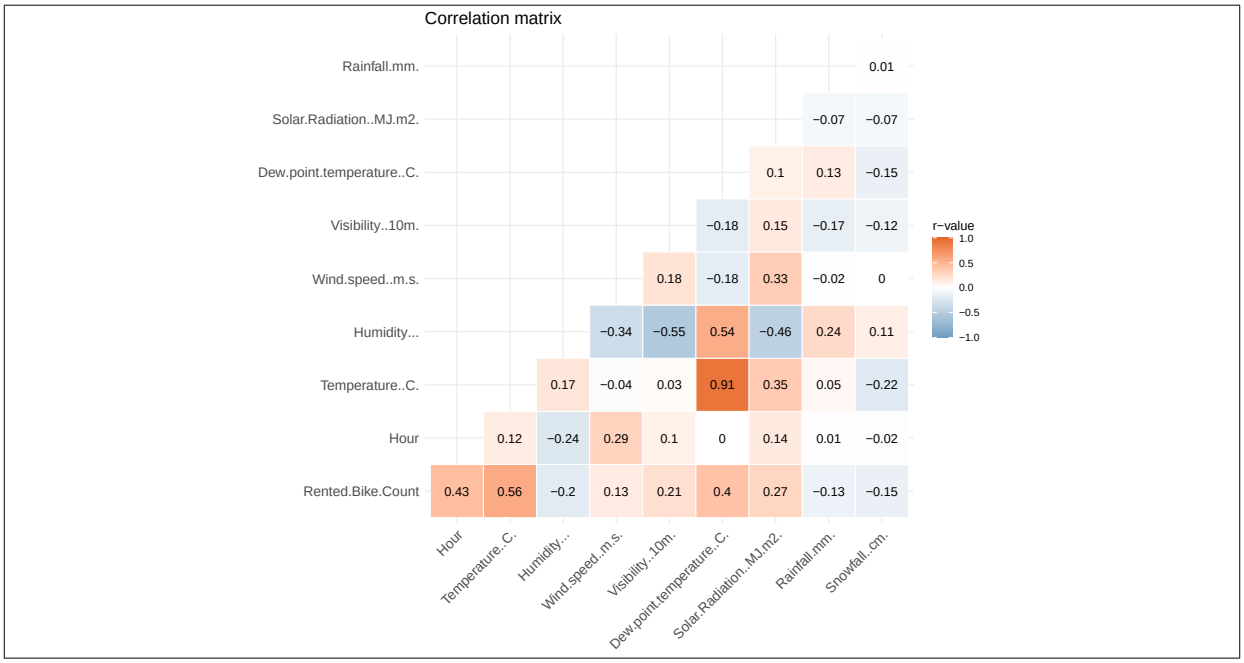
Correlation matrix

The correlation matrix can be obtained by using the dataset of standardized variables. Let's note X_{st} the standardized matrix, ${}^tX_{st}$ its transpose and n the standardized matrix's number of rows which is 8465.

The correlation matrix M_{corr} can be computed as follows:

$$M_{corr} = \frac{1}{n} \, {}^tX_{st}X_{st}$$

The following graph illustrates M_{corr}.



Performing PCA

Once the variables scaled, we can perform PCA. Both the variables `Rented.Bike.Count` and `Hour` are supplementary variables and will help us interpreting the principal components. The only active variables are the meteorological ones.

```
respca <- PCA(pca_data[,1:10], quanti.sup=1:2, graph=F,scale.unit=T)
```

Eigenvalues

Table 4: Eigenvalues and variance explained

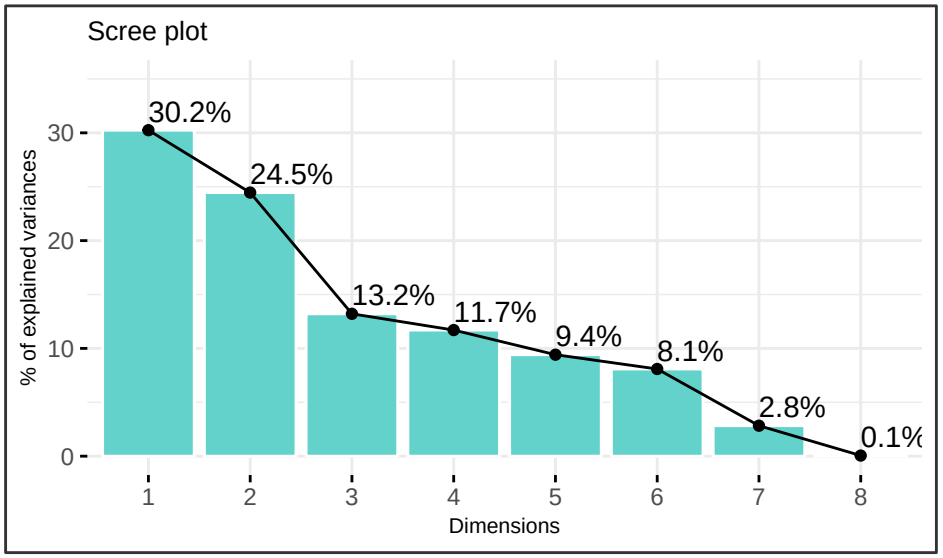
	eigenvalue	variance.percent	cumulative.variance.percent
Dim.1	2.42	30.25	30.25
Dim.2	1.96	24.46	54.70
Dim.3	1.06	13.21	67.91
Dim.4	0.94	11.70	79.61
Dim.5	0.75	9.42	89.03
Dim.6	0.65	8.09	97.12
Dim.7	0.23	2.83	99.94
Dim.8	0.00	0.06	100.00

Each of the ten principal components explain a percentage of the total variation in the dataset. Put together PC_1, PC_2 and PC_3 explain 67.91% of the variance. One could add PC_4 in order to keep almost 80% information. However, the study of the eigenvalues - as stated by the Kaiser-Guttman rule - indicates that only three components can be kept.

$$\forall\ i \in\ [\![1;3]\!]\ ;\ \lambda_i \geq 1$$

where λ_i stands for the eigenvalue associated with the i^{th} principal component.

On the scree plot which represents the eigenvalues found above, the "elbow" is located at the third principal component. This illustrates why only are kept the first three axes.



Variables' analysis

Quality of representation

The quality of representation of a variable j on the s-axis is measured by the percentage of inertia of variable j projected on the s-axis. One knows the closer to one is the value, the better the variable is represented. This numeric indicator can be added up for several axes and is most often calculated for a plan.

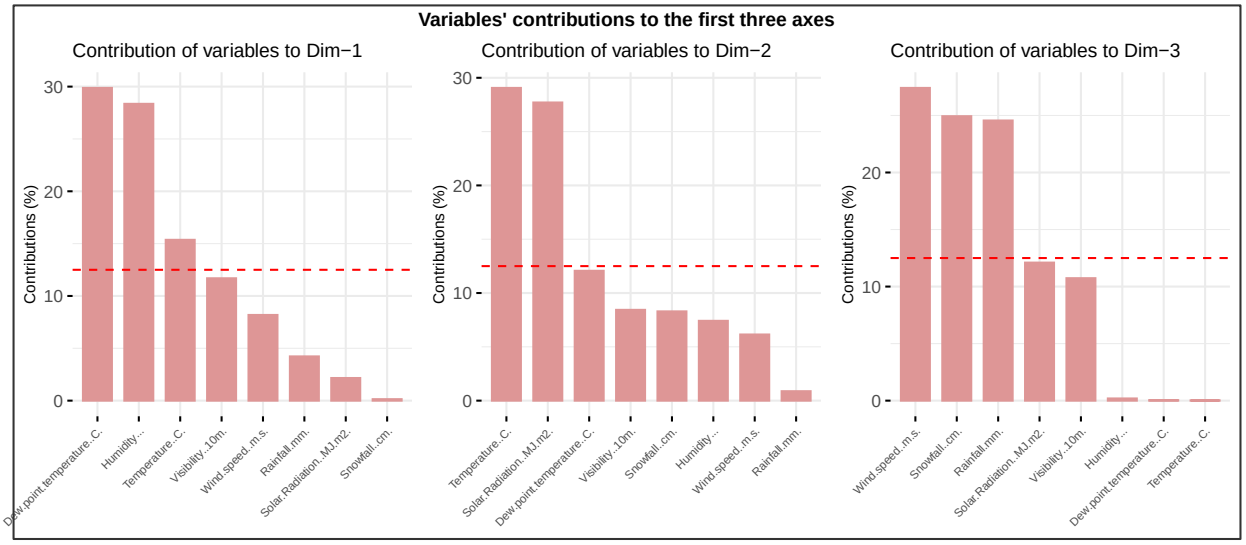
One notes the best represented variables in the first plan are: `Dew.point.temperature..C.`, `Temperature..C.`, and `Humidity...` In the mean time, there are some variables for which the (F_1, F_2) plan is inappropriate: `Wind.speed..m.s.`, `Snowfall..cm.` and `Rainfall.mm.`, and which are better represented in the (F_1, F_3) and (F_2, F_3) plans.

Table 5: Quality of representation of the variables in 3 plans

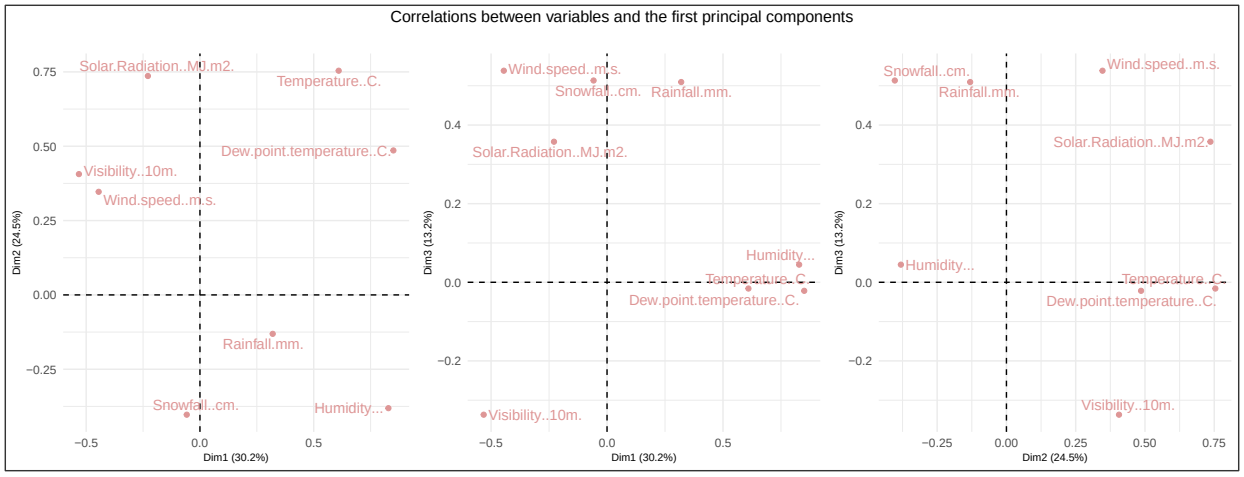
	(F1,F2)	(F1,F3)	(F2,F3)
Temperature..C.	0.94	0.37	0.57
Humidity...	0.83	0.69	0.15
Wind.speed..m.s.	0.32	0.49	0.41
Visibility..10m.	0.45	0.40	0.28
Dew.point.temperature..C.	0.96	0.72	0.24
Solar.Radiation..MJ.m2.	0.59	0.18	0.67
Rainfall.mm.	0.12	0.36	0.28
Snowfall..cm.	0.17	0.27	0.43

Contributions

The dotted lines represent the expected contribution of each variable if contributions were uniform. From the graph, one notes an antagonism between F_2 and F_3. While F_2 is largely described by `Temperature..C.` and `Solar.Radiation..MJ.m2.`, F_3 seems to be associated with bad weather variables with high contributions of `Wind.speed..m.s.`, `Snowfall..cm.` and `Rainfall.mm.`. Graphically, the most contributing variables are close to the edge of the variables' circle plotted below.

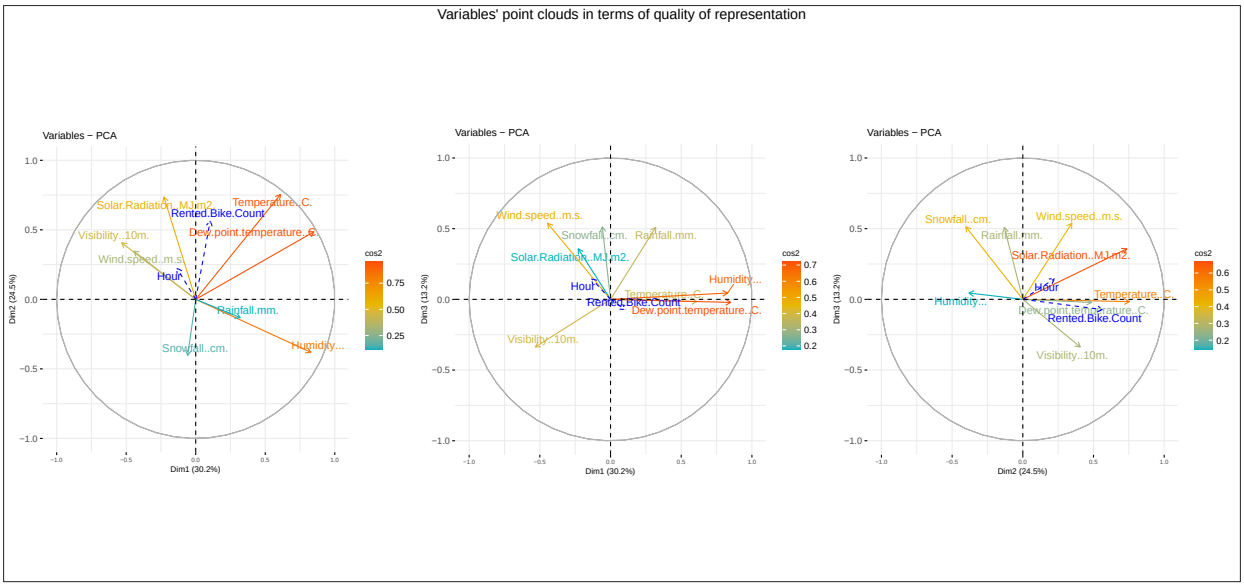


Correlations



Graph of variables

The supplementary variable `Rented.Bike.Count` is projected onto the second component. The latter information allows us pointing out variables pulling up `Rented.Bike.Count` Indeed, `Temperature..C.` and `Solar.Radiation..MJ.m2.` are well-represented and thus have a positive impact on the number of rented bikes. On the other hand regarding the third component, are depicted the negative-impact meteorological variables such as `Rainfall.mm.`, `Snowfall..cm.` and `Wind.speed..m.s.`



Individuals' analysis

Quality of representation

Quality of representation of an individual i on the s axis is measured by the part of inertia of i projected on s.

The following tables highlight the ten individuals who have the **best** quality of representation in each principal component.

Table 6: Dim 1

	cos2
5276	0.98
592	0.96
5310	0.96
808	0.96
5889	0.95
6074	0.95
5291	0.95
5049	0.95
1979	0.95
397	0.95

Table 7: Dim 2

	cos2
5969	0.94
6253	0.94
6018	0.93
4599	0.93
5490	0.93
6041	0.93
6039	0.93
6254	0.93
6113	0.93
5726	0.93

Table 8: Dim 3

	cos2
7672	0.67
3483	0.67
7676	0.67
3697	0.66
3482	0.66
7674	0.66
4127	0.65
7673	0.65
7579	0.64
7677	0.64

The following tables highlight the ten individuals who have the **worst** quality of representation in each principal component.

Table 9: Dim 1

	cos2
1228	0
3663	0
6162	0
1050	0
5465	0
4234	0
2721	0
7669	0
6160	0
2264	0

Table 10: Dim 2

	cos2
974	0
4726	0
7035	0
1835	0
4424	0
3169	0
7701	0
4587	0
3385	0
1410	0

Table 11: Dim 3

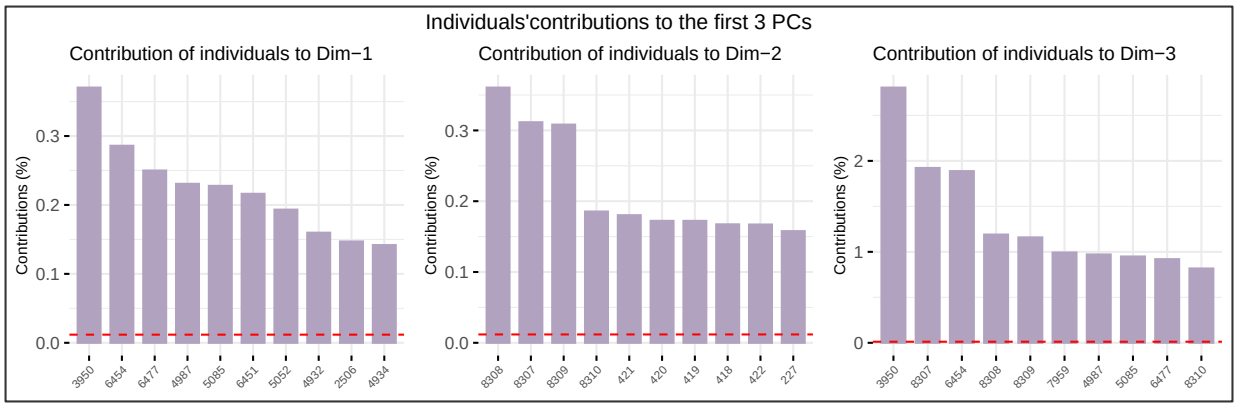
	cos2
5775	0
1004	0
4952	0
7664	0
6945	0
6830	0
2914	0
3862	0
6932	0
1559	0

It is difficult to draw conclusions from the previous tables as each table contain diverse individuals.

Individuals' with the highest contributions

The contribution of an individual i corresponds to the part of inertia he brings to the s-axis.

In these graphs, contributions may seem rather low, but this is due to the large amount of individuals. The red line depicts the uniform situation. We note the contributions to the third component are way higher than the contributions to the first two.



From the table below, we know the first dimension needs less individuals to gather half of the total contribution. Whereas the second, third and fourth dimensions need roughly half of the total amount of individuals to gather half of the total contribution. This makes sens because the first axes better explain the variability between individuals than higher axes.

Table 12: 50 per cent contribution cumulated

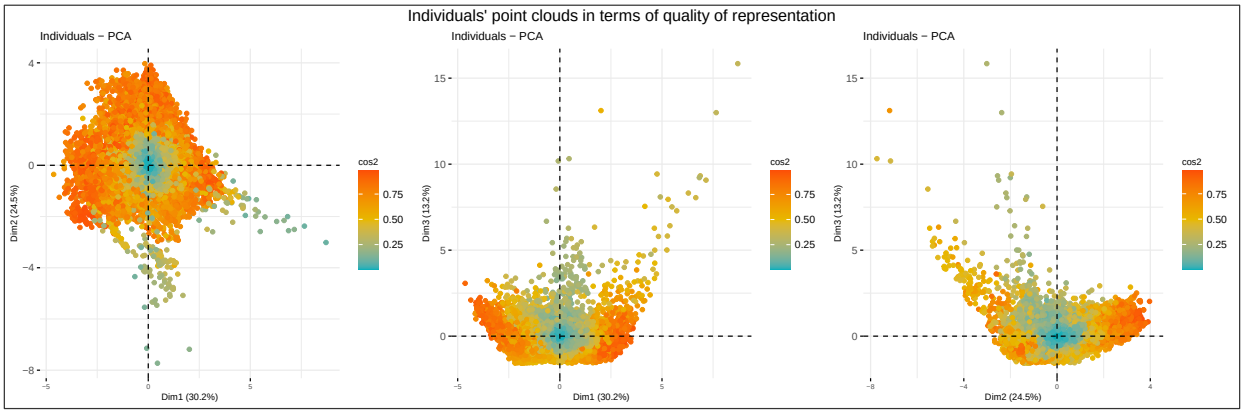
	Dim1	Dim2	Dim3	Dim4	Dim5
Number of individuals	3204	4332	4264	4355	3951

Graph of individuals

These representations not being fine-tuned, we cannot draw very precise conclusions except that well represented individuals are located far from the center of each point cloud. This makes sense because the better the representation of an individual is, the higher its coordinates on the projected axes are (in absolute value). Note that cos^2 stands for the quality of representation as the latter is defined as follows:

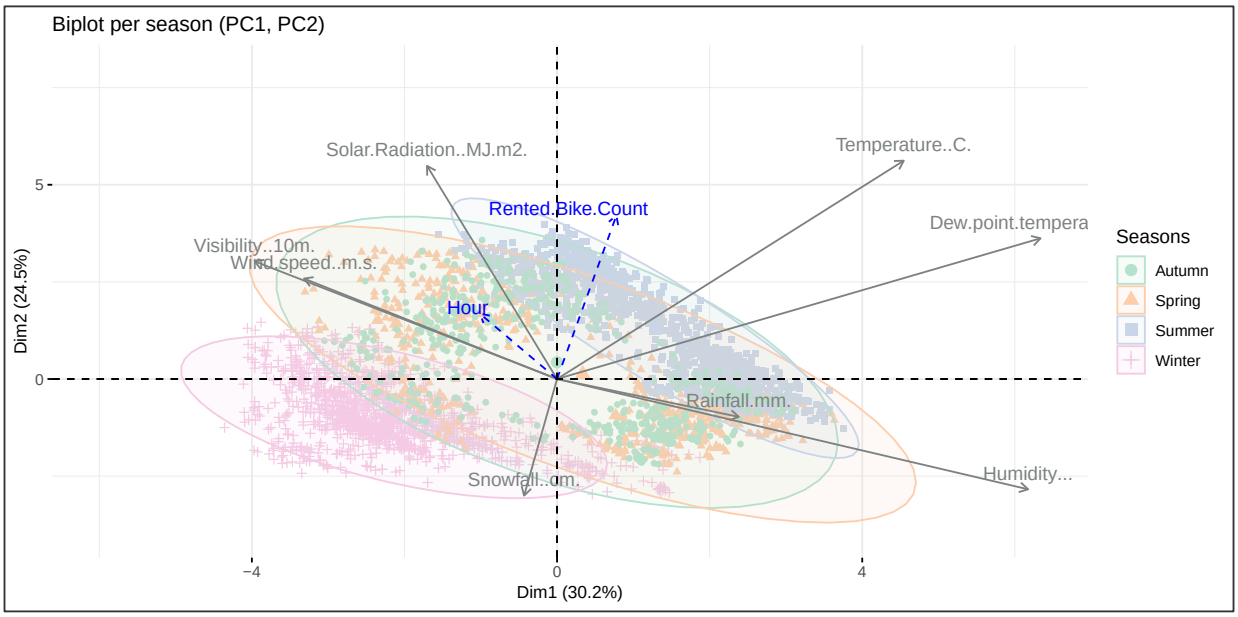
$$qlt_s(i) = \frac{F_s(i)^2}{\sum_i F_s(i)^2} = cos^2(\theta_i^s)$$

This formula is computed by using the Pythagorean theorem in the triangle defined by the vector between the origin of the plan O and the projected point i and the vector defining the s-axis.



Biplots

Last but not least, we decide to represent season groups onto the biplot of variables and individuals so as to bring out how both the variables and the time observations are linked to the seasons. The first striking difference is the dichotomy between Winter and Summer seasons. When looking at the `Dew.point.temperature..C.` and `Temperature..C.` 's projections a discrimination is noticed between these two opposite seasons. In contrast, Spring and Autumn seasons are much more similar according to many projections. We recall considering these two seasons as "average" seasons. In addition, the Winter group does not follow the `Rented.Bike.Count` path as the Summer group does.



Multiple correspondence analysis (MCA)

Multiple correspondence analysis can be seen as a generalization of principal component analysis when the variables to be analyzed are categorical instead of quantitative. In the `SeoulBike` dataset the variables are mostly quantitative except `Seasons`, `Holiday` and `Functionning.Day`. The latter two being useless, we have to create qualitative variables from the quantitative ones in order to perform MCA. We use the `class_cut` function to do so.

Keep in mind that we use this technique in order to summarize the core information that best explains the number of rented bikes in Seoul.

Categorical variables

```
###Function which creates categorical variables from quantitative variables###

break_fun <- function(x){
  if (quantile(x,probs=seq(0,1,0.25))[1] > 0){
    result <- c(0,quantile(x,probs=seq(0,1,0.25))[-1])
    }
```

```
  else{
    result <- quantile(x,probs=seq(0,1,0.25))
    }
  return(result)
}

class_cut <- function(x, breaks=break_fun(x), labels=NULL){
  output <- cut(x,breaks=breaks,labels=labels, include.lowest = TRUE)
  return (output)
}
```

Both the variables `RentedBike.cut` and `Temp.cut` are built on their parent quantitative variables' quartiles. The solar radiation level variable has been cut into three classses in order to create `SolRad.cut`.

Table 13: Random observations of the variables used in the MCA

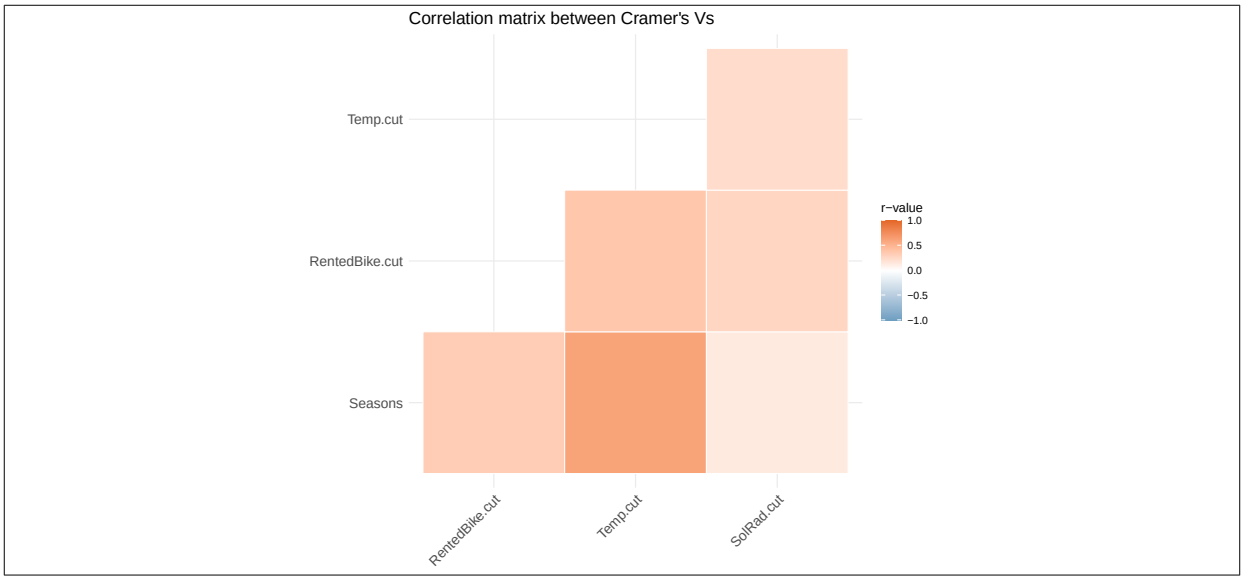
	Seasons	RentedBike.cut	Temp.cut	SolRad.cut
1	Winter	Med Rental	Low Temp.	Low Sol.
1001	Winter	Low Rental	Low Temp.	Low Sol.
2001	Winter	High Rental	Low Temp.	Low Sol.
3001	Spring	Med Rental	Med Temp.	Low Sol.
4001	Spring	Very high Rental	High Temp.	Low Sol.
5001	Summer	Very high Rental	High Temp.	Low Sol.
6001	Summer	High Rental	Very high Temp.	Low Sol.
7001	Autumn	Very high Rental	Very high Temp.	Med Sol.
8001	Autumn	High Rental	Med Temp.	Low Sol.

Links between qualitative variables : Cramér's Vs

Cramér's V is a measure of dependency between two qualitative variables and it is based on Pearson's χ^2 statistic. The higer the χ^2 statistic, the more likely the variables depend on each other, so the higer Cramér's V. It is computed with the following formula:

$$V = \sqrt{\frac{\chi^2/n}{\min(k-1, r-1)}}$$

The correlation matrix based on the Cramér's Vs depicts different positive correlations between the variables. Put another way, each categorical variable has a positive influence on the number of rented bikes.



Performing MCA

The following code line carries out the multiple correspondence on the chosen variables.

```
res.mca <- MCA(mca_data, ncp=10, graph=F, level.ventil = 0.1)
```

We have decide to keep 10 dimensions in the result and we have chosen 1% as frequency below which a rare modality is disaggregated, i.e. its individuals are distributed in the other modalities randomly.

Point clouds of individuals and variables' categories

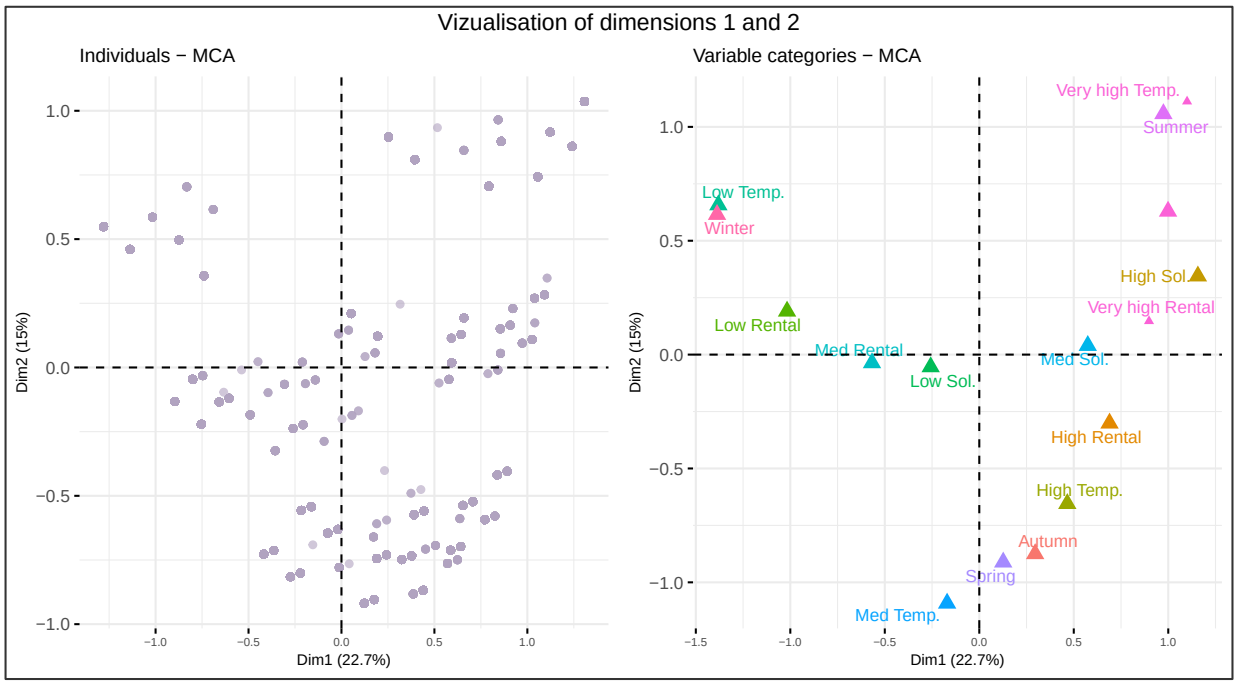
First look at the MCA results

It is important to remember that when we talk about individuals it refers to **time observations** as each row of the dataset is an hourly report from the Seoul bike sharing service.

Dimensions 1 and 2 The individuals point cloud highlights several groups. If each group is considered one big point, we can vizualise a U-shaped curve by linking them. Comparing this plot to the variable categories' one, we see that each group of individuals can be practically associated to one season.

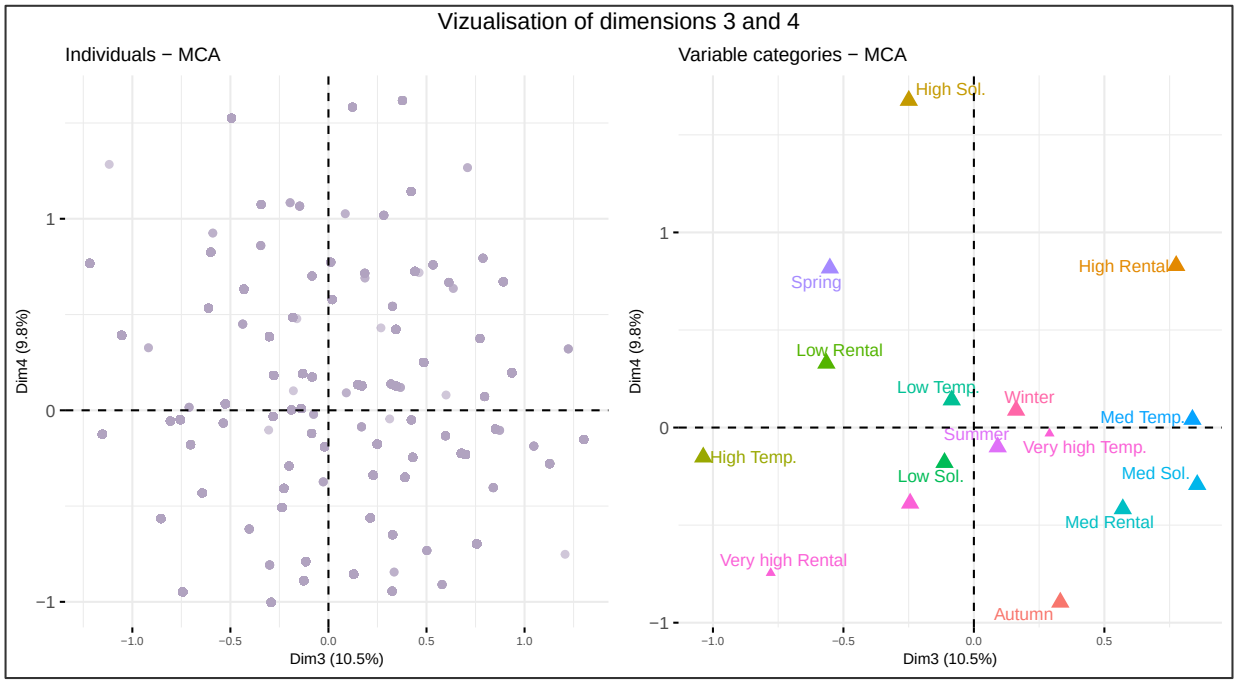
The variable categories' graph is also U-shaped as three out of four variables were formed from quantitative variables. It is called the Guttman effect and it is very obvious for the `Temp.cut` variable. It will be shown more precisely later.

The first axis ranks in ascending order the rental intensity, the temperatures' level and the solar radiation level, it is a **scale factor**. As for the second axis it is an **opposition factor** as it distinguishes moderate values such as `Med Temp` and `High Rental` from extreme values like `Very high Temp` and `Low Temp` (c.f. Guttman effect).

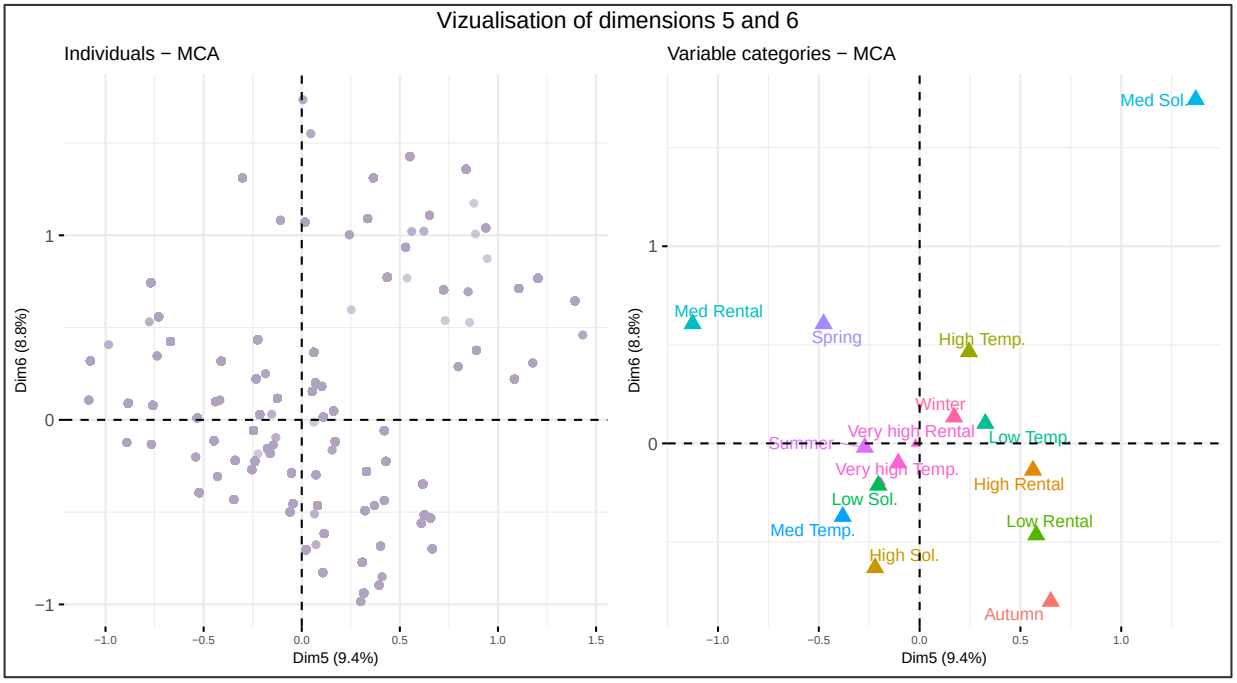


Higher dimensions Just like dimension 2, dimension 3 is an opposition factor for the `RentedBike.cut` variable. One can notice that axis 4 isolates the class `High Sol.` from `Low Sol.` and `Med Sol.`. It may be due to the fact that sunny days are not common place in Seoul.

The individuals' plot is hard to interpret as the point cloud is scattered.



Two groups of individuals can be noticed on the following graph, even though the variable categories' graph being difficult to interpret as it is tightened close to the origin. The `Med Sol.` category is the only one located far from the grouping of categories.



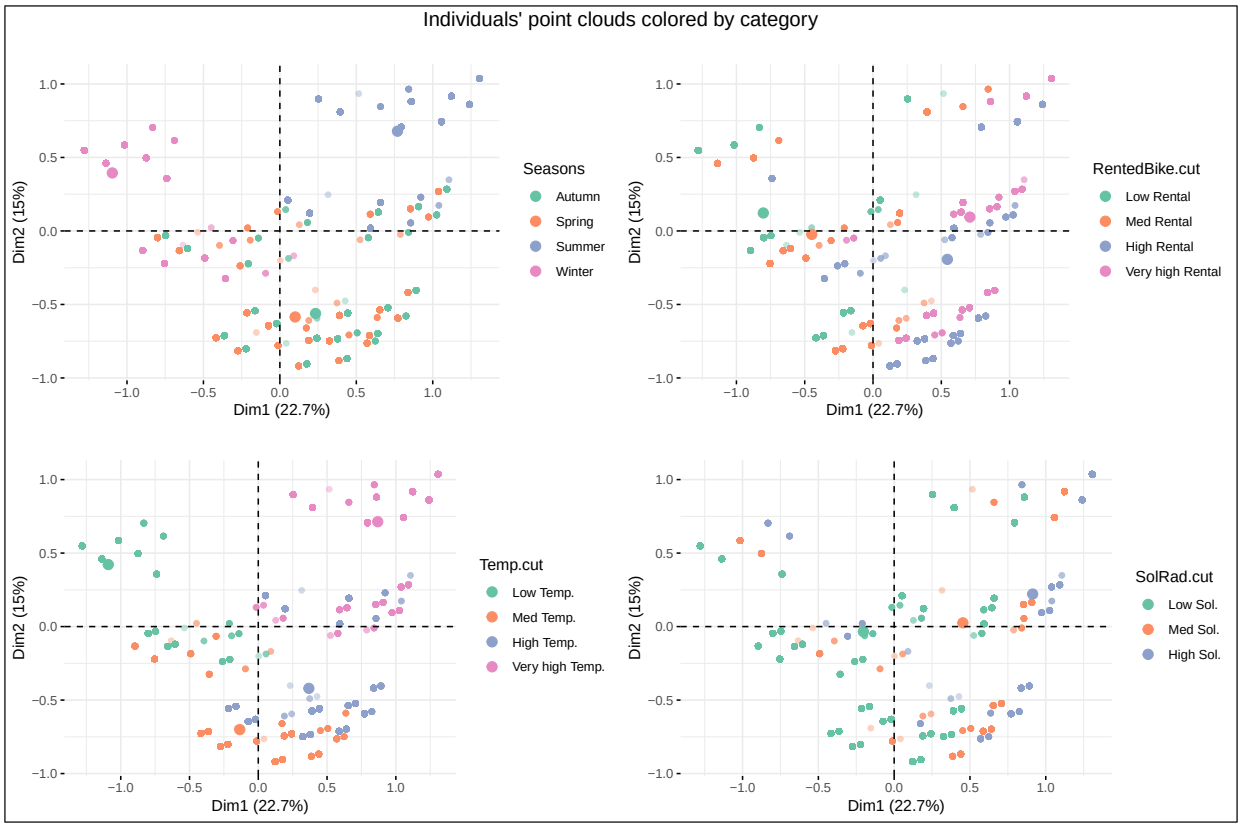
Individuals' point clouds

With the view of having a more fine-tuned representation of the point cloud, we can color each observation by category. It can be useful to identify several groups of time observations.

Top left there are time observations related to low temperatures, Winter season and a low rental level. Whereas both on the right and on the bottom right we can find time observations with high and very high rental levels, pretty high temperatures and mostly related to Spring and Autumn seasons.

The top right points correspond to Summer season with very high temperatures, diverse solar radiation levels and a fluctuating rental intensity.

In that respect, Autumn and Spring appear to be the seasons during which the more bikes are rented. Although the temperatures' level has a positive influence on the number of rented bikes, Summer season stands out as its rental intensity varies.



The MCA's main features

Eigenvalues and variance explained

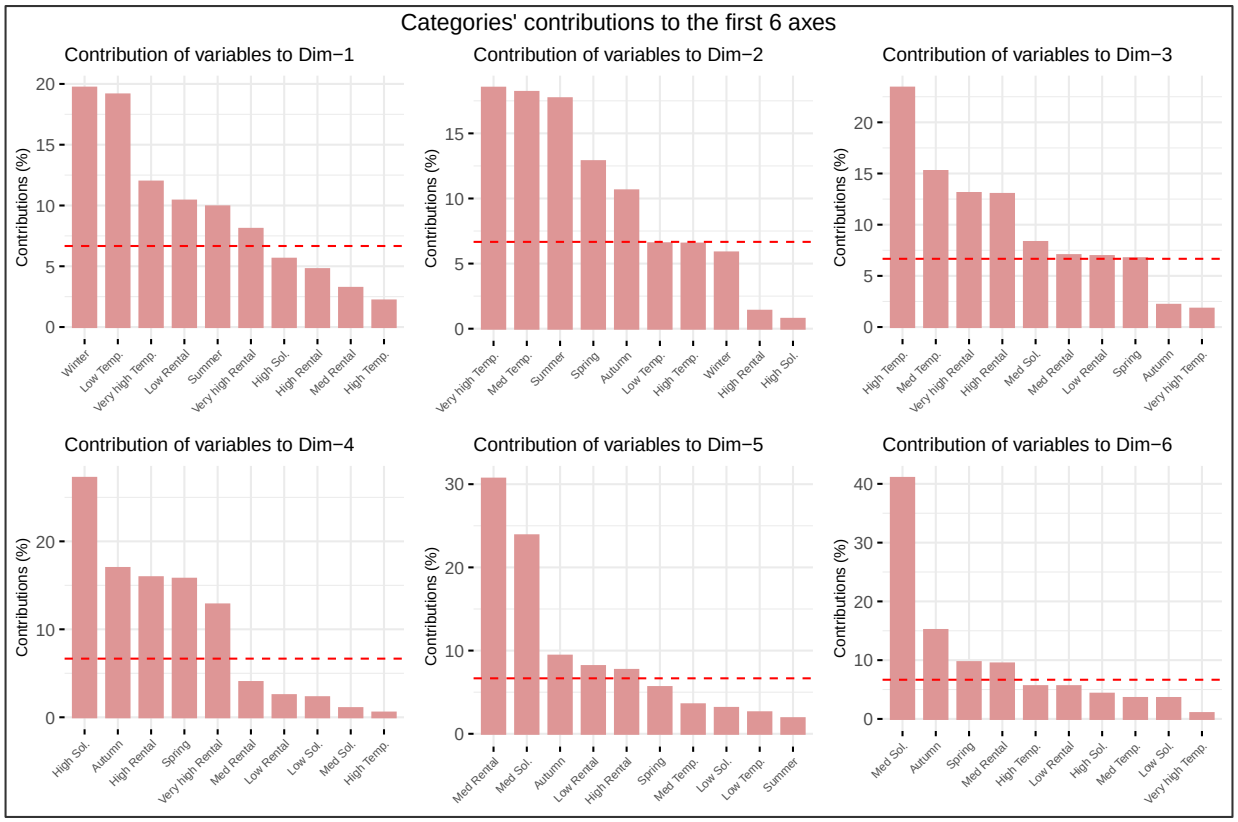
Unlike principal components analysis, MCA usually result in little variability retained by the axis which means that keeping only two or three axis is not enough to summarize most of the information. Here we cumulate almost 50% variance explained just by keeping the three first axes which is quite a good result. It may be due to the fact that we performed MCA on only four variables.

Table 14: Eigenvalues and variance explained

	eigenvalue	variance.percent	cumulative.variance.percent
Dim.1	0.62	22.68	22.68
Dim.2	0.41	14.98	37.66
Dim.3	0.29	10.51	48.17
Dim.4	0.27	9.83	58.00
Dim.5	0.26	9.38	67.38
Dim.6	0.24	8.84	76.22
Dim.7	0.22	7.99	84.21
Dim.8	0.21	7.54	91.74
Dim.9	0.11	3.98	95.72
Dim.10	0.08	3.00	98.72
Dim.11	0.04	1.28	100.00

Contributions

The contribution of one category k (respectively one individual i) to one axis s matches to the proportion of inertia brought by k (respectively i) to s.



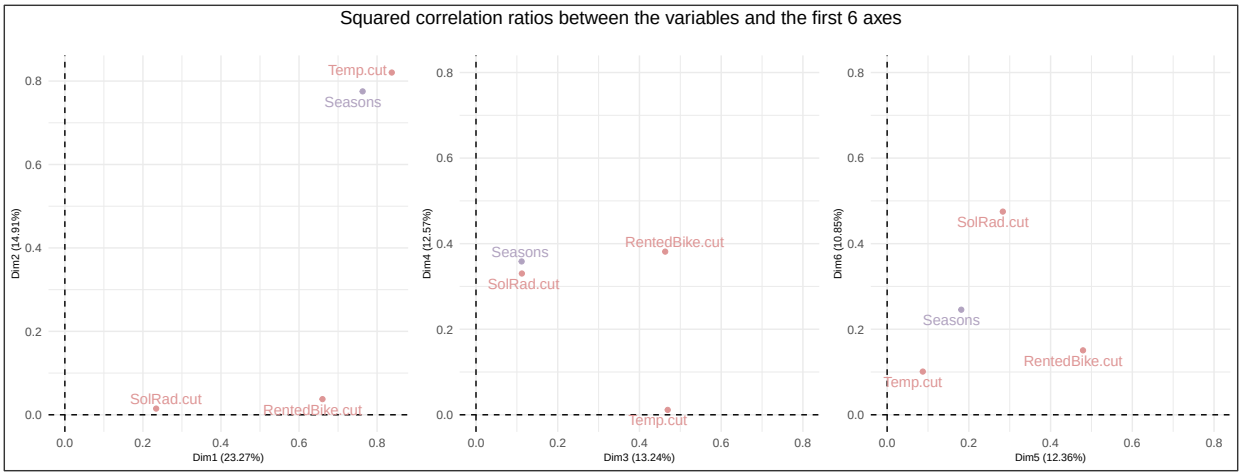
The `list_nb_ind_ctr` function we have created before allows us to determine how many observations we need in order to cumulate at least 50 per cent contribution to one axis s. For instance, about one quarter of the dataset's observations are needed to maintain a proportion of inertia of more than 50 percent in dimension one. This number increases for higher dimensions.

Table 15: 50 per cent contribution cumulated

	Dim1	Dim2	Dim3	Dim4	Dim5	Dim6
Number of individuals	2071	5103	4314	5127	4187	4471

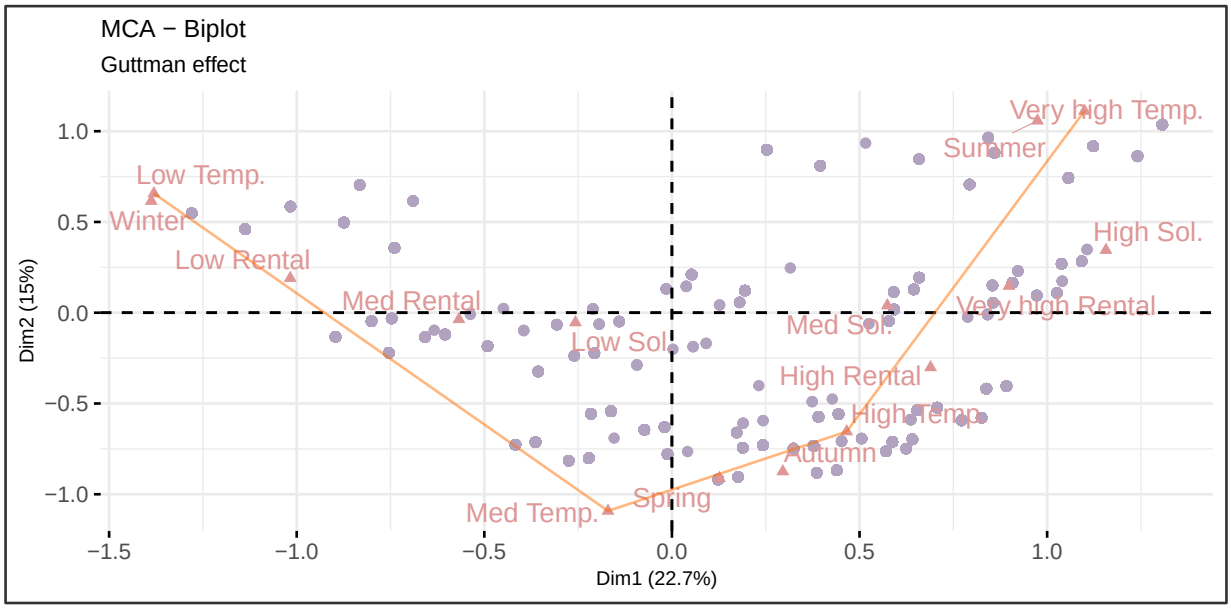
Correlations between the variables and the principal axes

The first axis is strongly correlated to `Temp.cut`, `RentedBike.cut` and `Season` which makes sense because we found a scale factor on both `Temp.cut` and `RentedBike.cut`. Most of the squared correlation ratios are way higher for dimensions one and two than for the other ones, apart from `SolRad.cut` which is 50% correlated to axis 6. This could explain why the `Med Sol` category is located far from the other categories in the $(F_5; F_6)$ plan.



Biplot of individuals and variables

This graph is relevant for two main reasons. On the one hand it links the groups of individuals and the associated categories. Top left we can identify the winter group with low temperatures and a low level of bikes rental, whereas top right we have the summer group. The `Very High Rental` category seems to be at equal distance from both the Summer and the Autumn groups. The Spring group appears to be the middle class. On the other hand the graph perfectly illustrates what Guttman effect is with this well defines U-shaped curve linking the `Temp.cut` variable's categories.



Correspondence analysis (CA)

Correspondence analysis is a multivariate statistical technique similar to PCA. The key difference is that it is not applied on quantitative variables but on categorical ones. CA is used as means of comparing two qualitative variables. It also provides a summary of data. Correspondence analysis uses the χ^2 statistic.

Correspondence analysis is performed on a contingency table C of size $n \times m$ where n is the number of rows and m is the number of columns.

We use this technique in our analysis in order to fine tune the links between the variables taken in pairs and to clarify the results found with multiple correspondence analysis.

Contingency tables

The contingency tables give us a first look at the way the different classes are distributed. By looking at the observed frequencies we see that both the temperatures and the times of day seem to influence the number of rented bikes. However the link between rented bikes and humidity level is less striking as the observed frequencies are more evenly distributed.

Table 16: Rented bikes and temperatures

	Low Temp.	Med Temp.	High Temp.	Very high Temp.
Low Rental	1221	477	334	91
Med Rental	777	713	368	253
High Rental	108	676	615	718
Very high Rental	14	261	811	1028

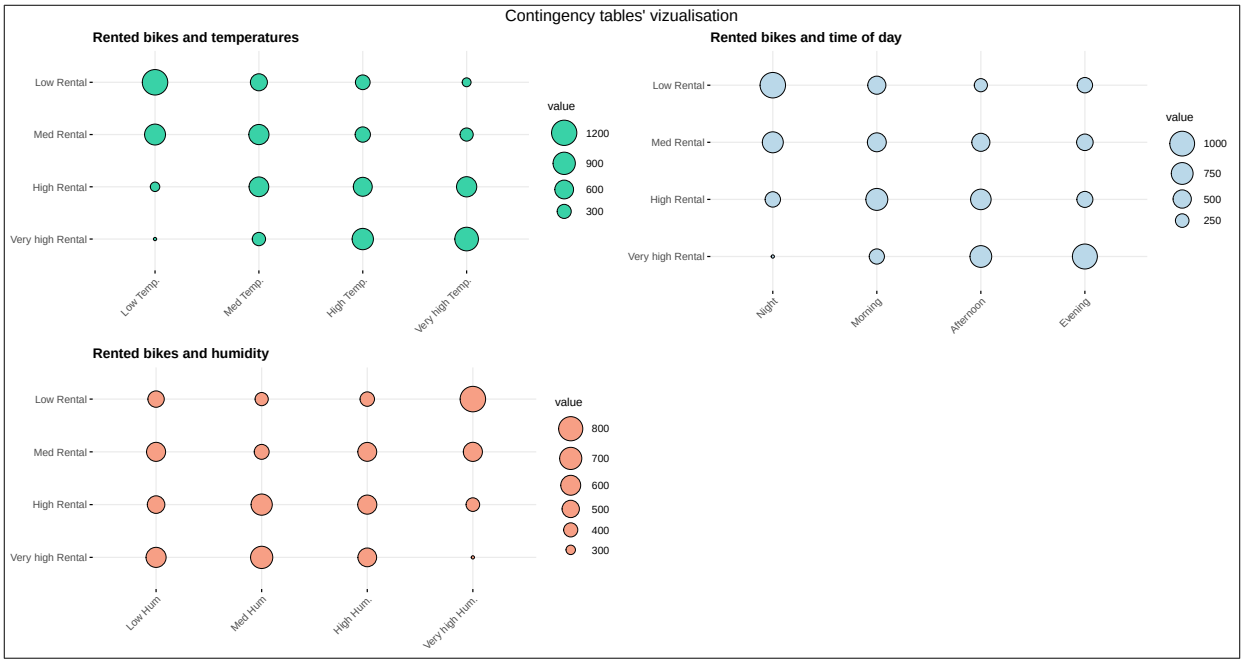
Table 17: Rented bikes and time of day

	Night	Morning	Afternoon	Evening
Low Rental	1065	490	232	336
Med Rental	686	541	488	396
High Rental	335	760	654	368
Very high Rental	26	326	744	1018

Table 18: Rented bikes and humidity

	Low Hum	Med Hum	High Hum.	Very high Hum.
Low Rental	467	376	409	871
Med Rental	564	422	558	567
High Rental	508	656	567	386
Very high Rental	607	703	552	252

Here is another way, may be more telling, to vizualize relationships between the categorical variables. One notices as temperatures increase, there are more and more rented bikes. As regards time of day, the evening clearly stands out with a great amount of rented bikes. Eventually, the less striking effect is the humidity one. It is easy to see people do not rent that many bikes when it is really humid. Aside from that, `Low Hum` and `Med Hum` categories are kind of equivalent in terms of rented bikes, with a slight advantage for the `Med Hum` category.



Independency tests

With the view of testing the likely independency between `RentedBike.cut` and the three other categorical variables, we use the Pearson χ^2 test. First let's check whether the test's validity conditions are met:

- N the total number of frequencies is greater than 50
- Each theoritical frequency is not less than 5

So the χ^2 statistic can be computed for the three tests.

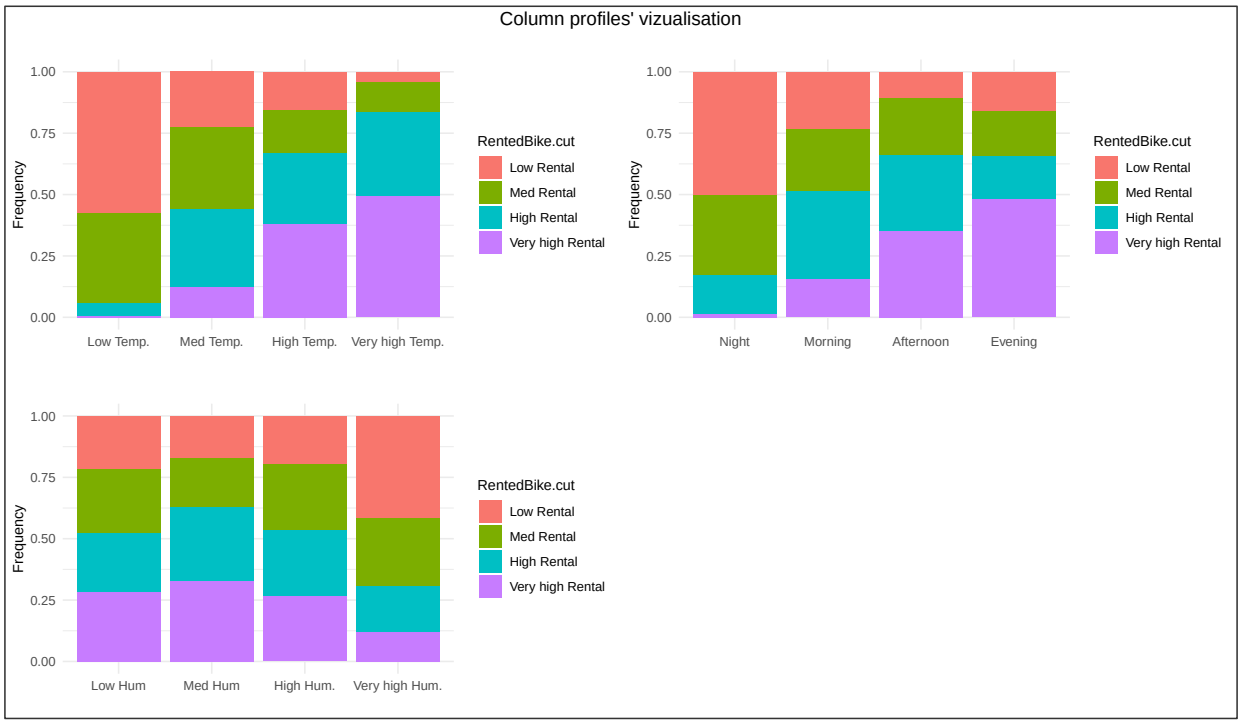
For each test, the null hypothesis that there are no difference between the variables' classes is rejected. The very low p-values and the high values of the χ^2 statistic indicate that the observed frequencies are unlikely to be observed in case of the null hypothesis would be true. So there are significant links between the variables taken in pairs.

Table 19: χ^2 tests on rented bikes

	Statistic	p.value	df
Temp.	3437.69	0	9
Time of day	2212.71	0	9
Humidity	617.51	0	9

The following barplots help us vizualise the links between the variables more easily. The temperatures' barplot displays a staircase shape with the number of rented bikes evolving proportionnally to the temperatures level. In addition, one notices a growing and positively-correlated relationship: the lower the temperatures, the lower the number of rented bikes. Then, as temperature increases, the `Low Rental` proportion decreases, whilst the `High Rental` and `Very high Rental` categories greatly enhance. The barplot on time of day is quite similar to the previous one and also displays a staircase shape, with `Evening` endorsing the number one spot in terms of rented bikes. In the mean time, the `Night` category is not really linked

to high levels of rental. As day goes along, a growth when it comes to rented bikes is noticed. Lastly, the humidity barplot is maybe the most uniform one which mitigates the relationship between this variable and the rented bike count.



Performing CA

Since the χ^2 test indicates no independency between the `RentedBike.cut` variable and the other ones, correspondence analysis can be carried out.

```
resca1 <- CA(CAdf1,graph=F) # RentedBike.cut and Temp.cut
resca2 <- CA(CAdf2,graph=F) # RentedBike.cut and Hour.cut
resca3 <- CA(CAdf3,graph=F) # RentedBike.cut and Hum.cut
```

Eigenvalues

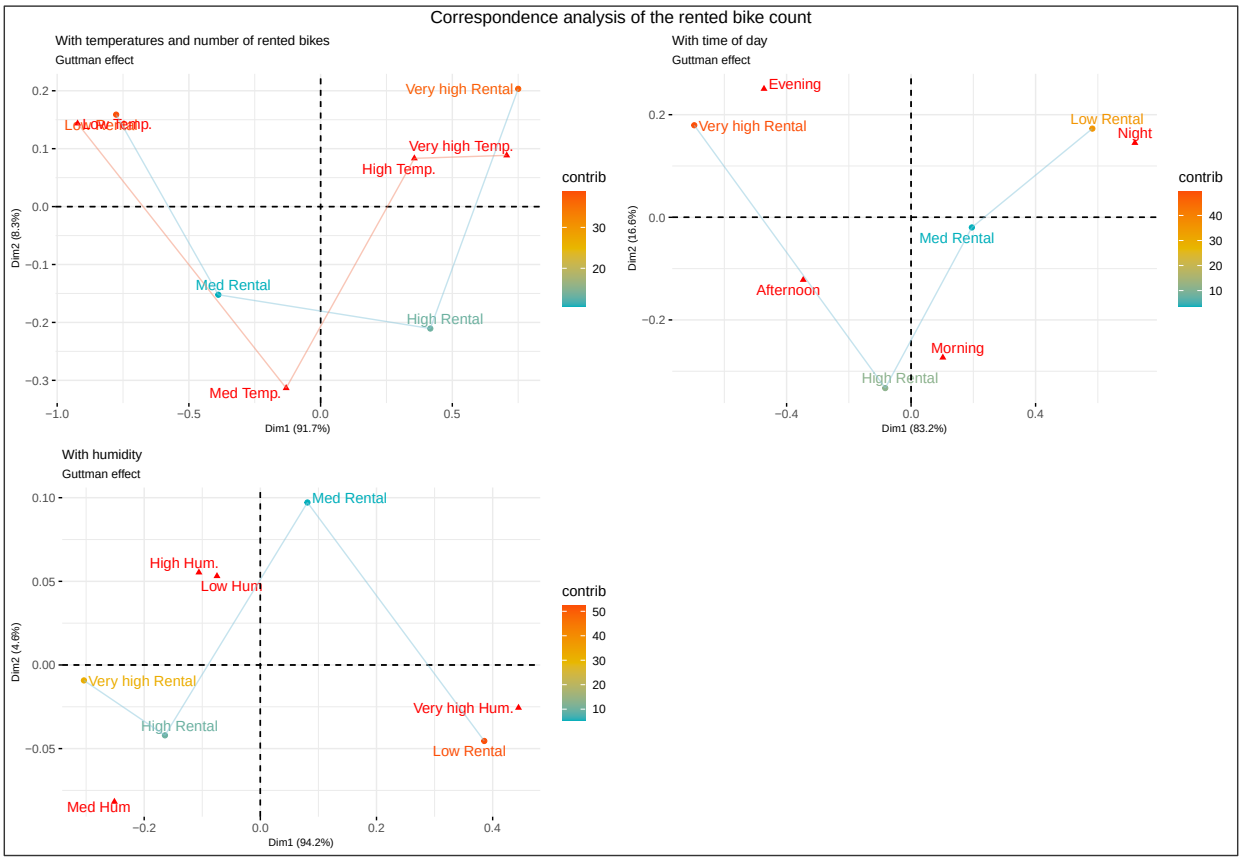
Here are displayed the three eigenvalues, per cent variances and per cent cumulative variances tables. For the `Temp.cut` variable, there is a huge amount - more than 90% - of information saved by the first axis. In other words, the first principal compenent tends to explain almost the entire link between `RentedBike.cut` and `Temp.cut`. The second and third axes gather the residual information. As far as `Hour.cut` is concerned, the first axis explains somewhat less information than for `Temp.cut`. We thus have more residual information even if more than 80% variance is explained on axis one. Regarding `Hum.cut`, the first axis almost explains the entirety of information, with only 5% noise on the upper dimensions.

Table 20: Eigenvalues - Rented bikes and temperatures, time of day and humidity

	Temp.cut			Hour.cut			Hum.cut		
	Dim.1	Dim.2	Dim.3	Dim.1	Dim.2	Dim.3	Dim.1	Dim.2	Dim.3
eigenvalue	0.37	0.03	0.00	0.22	0.04	0.00	0.07	0.00	0.00
variance.percent	91.74	8.25	0.01	83.21	16.58	0.21	94.19	4.57	1.24
cumulative.variance.percent	91.74	99.99	100.00	83.21	99.79	100.00	94.19	98.76	100.00

Biplots

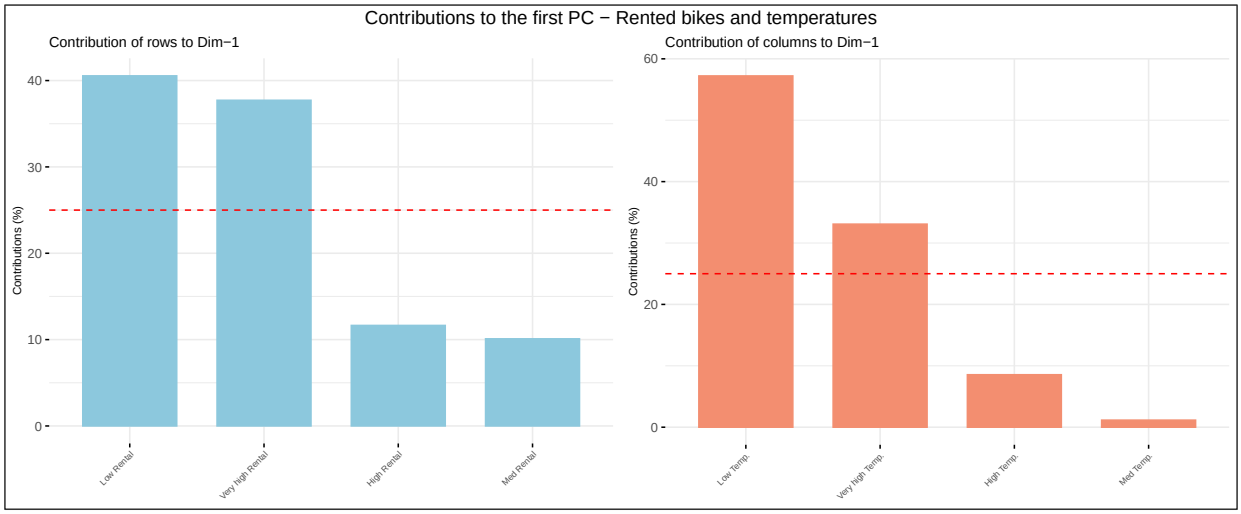
We distinguish both U-shaped and inverted U-shaped curves. The first two graphs are quite similar as the first axis discriminates between temperature levels (respectively time of day) and rental levels whereas the second axis contrasts extreme and average values. We see that each `Rentedbike.cut` category is associated with a level of temperature (graph 1) and a time of day (graph 2): it is the Guttman effet. It is really obvious for the `Low Rental` and `Very high Rental` classes. As for the inverted U-shaped curve graph, it depicts that a very high humidity level causes a low rental level whilst a medium humidity level seems to be the perfect condition for having high and very high levels of rental. The relationship is less strong though.



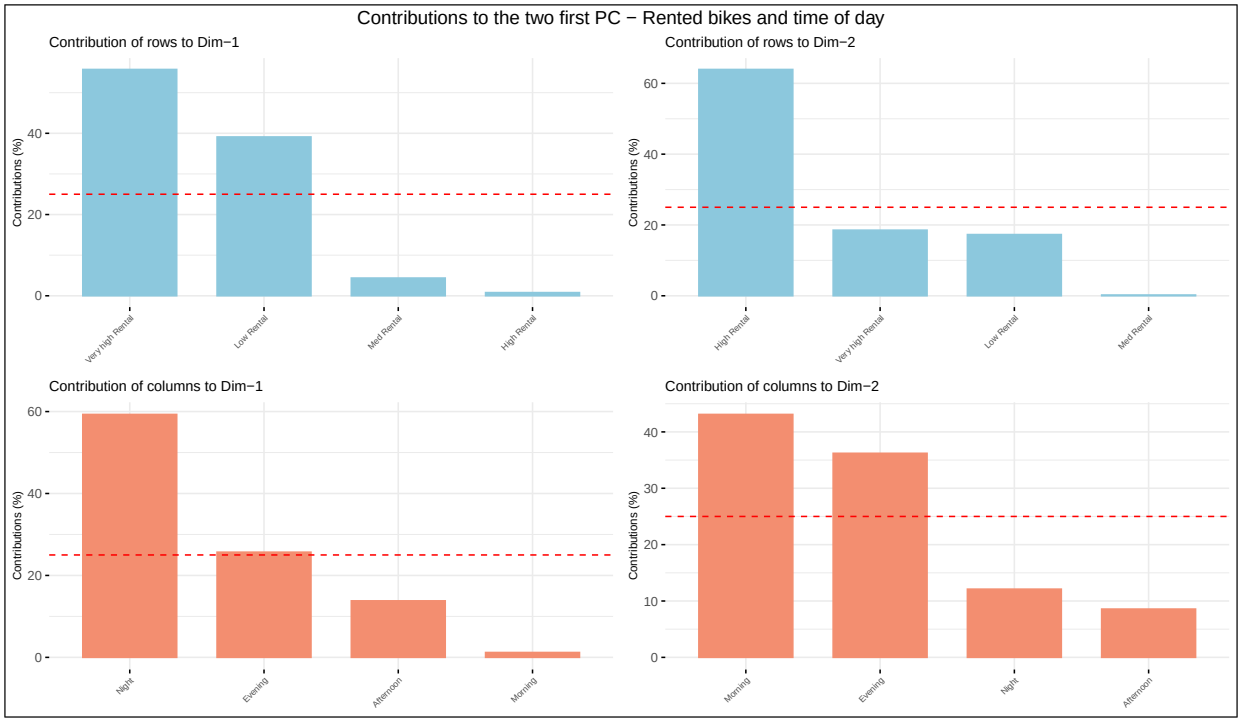
Contributions

The following graphs give sense to the previous biplots as the categories which contribute the most to the first two principal components are plotted for each CA.

For the CA on rented bikes and temperatures, extreme categories are above the red line which depicts the uniform situation. One knows that the rows and columns which contribute more than others are the most important when it comes to variability kept by the principal components.

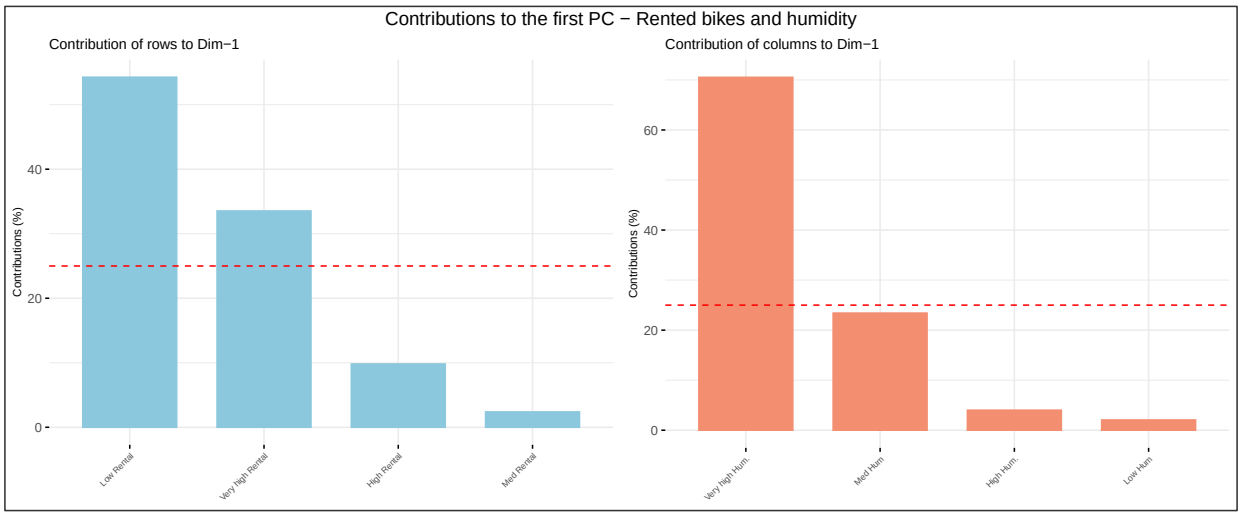


As regards time of day, one notices the huge contribution of the `Night` category to the first dimension. This category attracts towards itself the `Low Rental` category as shown by the second biplot. Here both dimensions one and two are plotted because the second axis still keeps 16.5% variance.



These last graphs of contributions demonstrate our previous assumptions: a very high humidity level has a negative impact on the bike rental level. The `Very high Hum` category attracts towards itself the `Low`

`Rental` one. What's more the first axis sorts the rental intensity in decreasing order which means it is a scale factor.



Clustering using the MCA's results

Starting from the results of the multiple correspondence analysis, we aim to group individuals into several subsets. We have been able to make associations between different time observations thanks to the MCA. That's why we chose to perform clustering on the individuals' coordinates found with MCA rather than those found with PCA. Now we will clearly identify these associations by using hierarchical clustering.

First and foremost the technique used in this part is led by `R` algorithms. Thus the clusters we get will be discovered, not computed by us.

Table 21: Cumulative inertia - MCA

	Dim.1	Dim.2	Dim.3	Dim.4	Dim.5	Dim.6	Dim.7	Dim.8	Dim.9	Dim.10	Dim.11
Inertia (%)	22.68	37.66	48.17	58	67.38	76.22	84.21	91.74	95.72	98.72	100

The previous table indicates that eight principal axes can be kept in order to preserve more than 90 % cumulative inertia. The 10 % left can be considered noise on the higher dimensions. That's why the `clust_data` data frame has been created.

```
clust_data <- res.mca$ind$coord[,1:8] %>% as.data.frame()
```

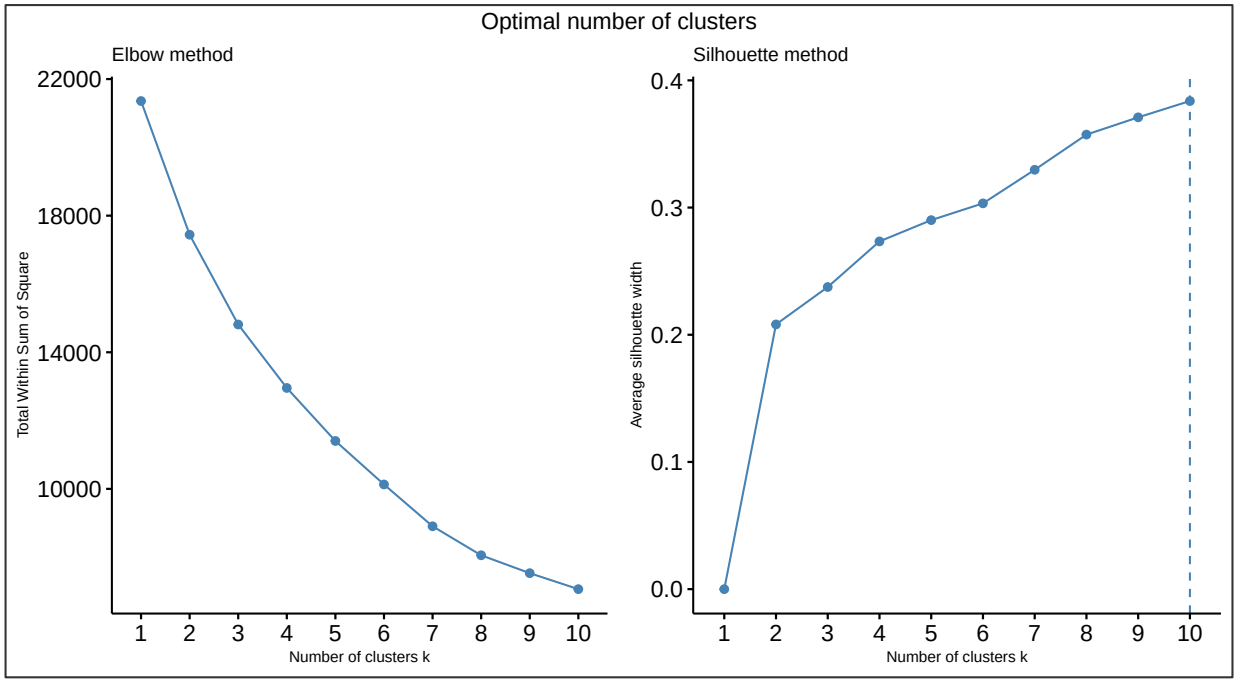
The optimal number of clusters

We use two methods to determine the number of clusters k. The first one amounts to find the optimal k such as the total within sum of square, i.e. the variation into one cluster, be low enough without k being too high. It is called the "elbow" method because the curve is elbow-shaped where the optimal number of clusters is located. The second one named the "silhouette" method is a measure of how similar an object is to its own cluster compared to other clusters. The silhouette ranges from -1 to 1, where a high value

indicates that the object is well matched to its own cluster and poorly matched to neighboring clusters. The average silhouette of one data point i is defined as follows:

$$s_i = \frac{b_i - a_i}{max\{a_i; b_i\}}$$

where a_i stands for the average distance between point i and the other points of the same cluster and b_i being the average distance between point i and the nearest cluster's points.



On the "elbow" method no elbow stands out but there are two points where the curve breaks which are $k = 3$ and $k = 7$. On the other hand the "silhouette" depicts an optimal number of ten clusters.

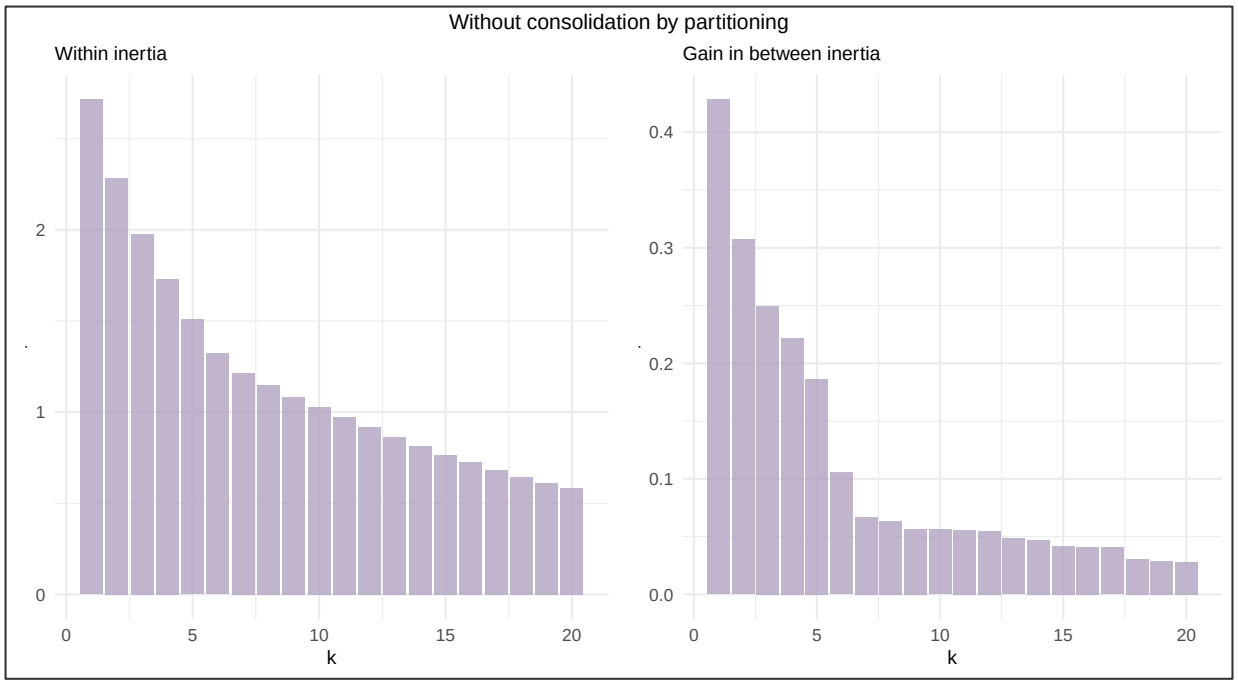
Hierarchical clustering

Without consolidation

We will fondly perform a hierarchical clustering without consolidation indicating to the algorithm that we want ten clusters.

```
hc1 <- HCPC(res.mca, consol = FALSE, nb.clust = 10, graph = FALSE)
```

Let's recall we want to have clusters with similar individuals inside, but an important variability between clusters. It amounts to minimize within inertia while maximizing between inertia.



On the first plot, we see that the decrease in the within inertia is noticeable until $k = 5$. Then, the marginal decrease is almost constant. On the second plot, the gain in between inertia is quite high until $k = 5$, then there is a break.

After having done a clustering with $k = 5$, `R` indicated that the natural and optimal number of clusters was three.

As we have a large number of individuals, we ask `R` to perform a Kmeans preprocessing (with 50 clusters) before the hierarchical clustering in order for the dendrogram to be displayed. This preprocessing by partioning is very useful if the number of individuals is high. Note that consolidation cannot be performed if `kk` is different from Inf and some graphs are not drawn. So, the dendrogram's graph is related to a non-consolidated hierarchical clustering.

```
hc2 <- HCPC(res.mca, kk=50,consol = FALSE, nb.clust = 3, graph = FALSE)
```

With consolidation

Once the dendrogram can be plotted, we decide to carry out a hierarchical agglomerative (HAC) clustering with consolidation by partioning. This time we do not ask for a Kmeans preprocessing because it would prevent us for having an appropriate clusters' vizualisation and the consolidation could not be performed.

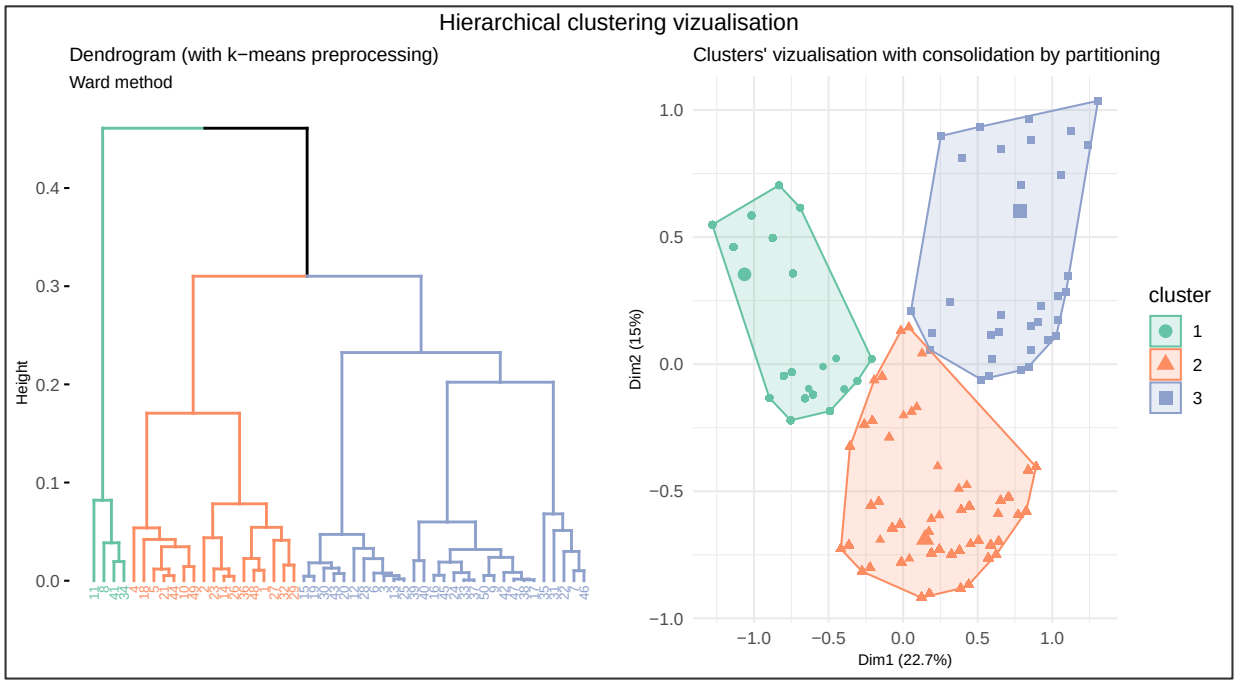
```
hc_conso <- HCPC(res.mca, nb.clust=3, consol = TRUE,  graph = FALSE)
```

The consolidation has been effective as it led to a 15 % increase in between inertia, which means more variability between clusters.

Table 22: Gain in between inertia

	Before consolidation	After consolidation
Between inertia	0.7355	0.8831

The hierarchical clustering has been efficient to the extent that we can vizualise three distinct groups of observations. The larger points of each cluster represent their respective centers, i.e. the individuals who are most representative of each group. The red and the blue groups are both made of more individuals than the green one as shown by the two graphs. The figures written below the dendrogram are related to the 50 groups obtained by the K-means preprocessing. Each of theses groups contain a large number of individuals. As for the clusters' vizualisation, it is not directly linked to the dendrogramn as it has not been obtained after a K-means preprocessing on the HAC. So the dendrogramn gives rather an idea of how the groups of individuals are distributed in each cluster than the exact individuals' distribution.



Clusters' description

The following tables are useful to give sense to the three groups depicted above.

96.93% of time observations which are strongly linked to a low level of temperatures are located in **cluster 1** and they account for 86.09% of this group whereas this category only represents one quarter of all observations. In the meantime there is no observation with a very high rental level and 2.39% with a high rental level in cluster 1, whereas these two categories represent 50% of all observations if put together. So cluster 1 describes observations related to Winter season with a low or medium level of rental and low temperatures.

Looking at the second table, we see that **cluster 2** brings together Spring and Autumn related observations because 85.69% and 82.65% of these categories are located in this cluster. Put together, more than 99% of the cluster 2's individuals come from Spring and Autumns seasons. This group tends to represent large levels of rental as more than 60% of the observations come from `High Rental`and `Very high Rental` categories put together.

As for **cluster 3**, it is the group which brings Summer related observations together as shown by the 100% figure. As a whole, there are more observations with high and very high levels of rental in this cluster than in the others. 54.87% of the `Very high rental` category is located in cluster 1, whereas it only represents one quarter of all observations.

Table 23: Cluster 1

	Cla/Mod	Mod/Cla	Global	p-value	v-test
Temp.cut=Low Temp.	96.93	86.09	25.04	0	Inf
Seasons=Winter	98.47	89.11	25.52	0	Inf
RentedBike.cut=Low Rental	60.76	54.04	25.08	0	37.20
RentedBike.cut=Med Rental	49.27	43.57	24.94	0	24.07
SolRad.cut=Low Sol.	32.79	88.86	76.41	0	17.83
SolRad.cut=Med Sol.	19.01	8.84	13.11	0	-7.54
Temp.cut=Med Temp.	15.61	13.91	25.13	0	-15.53
SolRad.cut=High Sol.	6.20	2.30	10.48	0	-17.36
Seasons=Spring	7.78	7.04	25.52	0	-26.63
Seasons=Autumn	4.75	3.85	22.88	0	-29.38
RentedBike.cut=High Rental	2.69	2.39	25.01	0	-34.61
Temp.cut=Very high Temp.	0.00	0.00	24.69	0	-Inf
Temp.cut=High Temp.	0.00	0.00	25.14	0	-Inf
RentedBike.cut=Very high Rental	0.00	0.00	24.97	0	-Inf
Seasons=Summer	0.00	0.00	26.08	0	-Inf

Table 24: Cluster 2

	Cla/Mod	Mod/Cla	Global	p-value	v-test
Temp.cut=Med Temp.	84.39	51.51	25.13	0.00	Inf
Seasons=Spring	85.69	53.11	25.52	0.00	Inf
Seasons=Autumn	82.65	45.94	22.88	0.00	Inf
Temp.cut=High Temp.	75.80	46.28	25.14	0.00	37.70
RentedBike.cut=High Rental	57.01	34.63	25.01	0.00	17.00
RentedBike.cut=Very high Rental	45.13	27.37	24.97	0.00	4.26
SolRad.cut=Low Sol.	41.76	77.50	76.41	0.05	1.99
SolRad.cut=High Sol.	36.64	9.33	10.48	0.00	-2.91
RentedBike.cut=Med Rental	32.16	19.48	24.94	0.00	-9.80
RentedBike.cut=Low Rental	30.38	18.51	25.08	0.00	-11.81
Temp.cut=Very high Temp.	0.57	0.34	24.69	0.00	-Inf
Temp.cut=Low Temp.	3.07	1.87	25.04	0.00	-Inf
Seasons=Winter	1.53	0.95	25.52	0.00	-Inf
Seasons=Summer	0.00	0.00	26.08	0.00	-Inf

Table 25: Cluster 3

	Cla/Mod	Mod/Cla	Global	p-value	v-test
Temp.cut=Very high Temp.	99.43	80.14	24.69	0	Inf
Seasons=Summer	100.00	85.15	26.08	0	Inf
RentedBike.cut=Very high Rental	54.87	44.74	24.97	0	27.16
SolRad.cut=High Sol.	57.16	19.55	10.48	0	17.35
RentedBike.cut=High Rental	40.29	32.90	25.01	0	10.96
SolRad.cut=Med Sol.	39.64	16.97	13.11	0	6.85
Temp.cut=High Temp.	24.20	19.86	25.14	0	-7.55
RentedBike.cut=Med Rental	18.57	15.12	24.94	0	-14.34
SolRad.cut=Low Sol.	25.45	63.48	76.41	0	-18.17
Seasons=Autumn	12.60	9.41	22.88	0	-20.83
RentedBike.cut=Low Rental	8.86	7.25	25.08	0	-27.23
Seasons=Spring	6.53	5.44	25.52	0	-31.00
Temp.cut=Med Temp.	0.00	0.00	25.13	0	-Inf
Temp.cut=Low Temp.	0.00	0.00	25.04	0	-Inf
Seasons=Winter	0.00	0.00	25.52	0	-Inf

The clusters' parangons

In the three tables below we have the main features of each cluster's parangons which are the closest observations to the center of the cluster. The parangons give sense to the results found in the previous tables.

Table 26: Cluster 1 's parangons

	Seasons	RentedBike.cut	Temp.cut	SolRad.cut
5	Winter	Low Rental	Low Temp.	Low Sol.
2	Winter	Low Rental	Low Temp.	Low Sol.
6	Winter	Low Rental	Low Temp.	Low Sol.
28	Winter	Low Rental	Low Temp.	Low Sol.
29	Winter	Low Rental	Low Temp.	Low Sol.

Table 27: Cluster 2 's parangons

	Seasons	RentedBike.cut	Temp.cut	SolRad.cut
2203	Spring	High Rental	Med Temp.	Low Sol.
2228	Spring	High Rental	Med Temp.	Low Sol.
2229	Spring	High Rental	Med Temp.	Low Sol.
2230	Spring	High Rental	Med Temp.	Low Sol.
2276	Spring	High Rental	Med Temp.	Low Sol.

Table 28: Cluster 3 's parangons

	Seasons	RentedBike.cut	Temp.cut	SolRad.cut
4340	Summer	Very high Rental	Very high Temp.	Low Sol.
4341	Summer	Very high Rental	Very high Temp.	Low Sol.
4342	Summer	Very high Rental	Very high Temp.	Low Sol.
4364	Summer	Very high Rental	Very high Temp.	Low Sol.
4365	Summer	Very high Rental	Very high Temp.	Low Sol.

Dependency between the variables and each cluster

Each cluster depends on the four categorical variables used in the multiple correspondence analysis. It means that these variables are useful to interpret the clusters found by the algorithm. The χ^2 test results depicted below illustrate the interdependence between the cluster variables and the four qualitative variables.

Table 29: Link between the cluster variables and the categorical variables (χ^2 test)

	p.value	df
Seasons	0	6
RentedBike.cut	0	6
Temp.cut	0	6
SolRad.cut	0	4

Before applying hierarchical clustering, we did not know how many clusters exist in the data. The algorithm performed by `R` led us to identify three main groups in the `SeoulBike` dataset. These three groups are quite different one to another. The consolidation by partitioning increased the intercluster distance. Conversely the intracluster distance is quite small meaning the data points inside each cluster are close to each other, i.e. each group of observations is homogeneous. We think the clusters' distribution gives a good insight on the actual Seoul bike sharing service data.

Conclusion

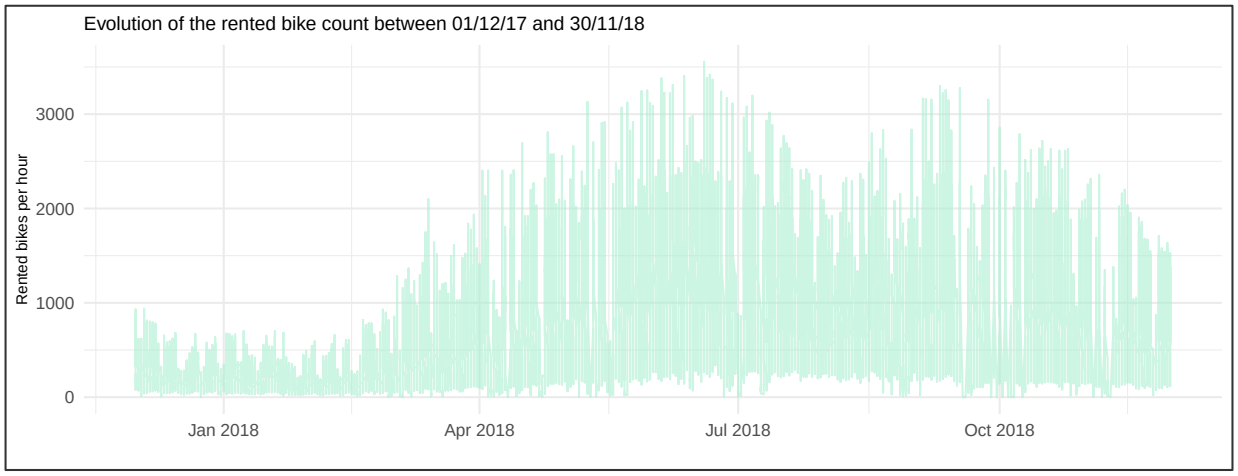
Through the use of factor analysis combined with clustering, we managed to fine-tune the analysis of the `SeoulBike` dataset.

Principal component analysis has allowed us to sort through the meteorological variables which were the most capable of explaining the number of rented bikes. We got out three significant principal components, the second one being strongly related to the bike rental level. Then, thanks to multiple correspondence analysis we have been able to study links between four categorical variables created from the quantitative variables. Although these four variables were all positively correlated, MCA has turned out to be a success with the distinction of several groups of individuals depending on seasons. To end up with factor analysis, has been decided the study of links between the number of rented bikes and temperatures, time of day and level of humidity by the means of correspondence analysis. Were identified the meteorological features which lead to a high rental level. Last but not least, thanks to hierarchical clustering onto MCA's results we have been able to determine three main clusters based on seasons and temperatures. The latter clusters describe and differentiate bike rental levels.

From our analysis we are able to give some advice to the firm running the Seoul bike sharing service. In our opinion, here is how the number of available bikes should be allocated over a year. There should be a decrease in terms of available bikes in Winter as this season is associated with low rental levels. In reverse the number of available bikes should increase between Spring and Autumn seasons with a slight decrease during vacancy periods at mid-summer. The Seoul bike sharing service should step up the number of available bikes in June and September as shown by the third cluster, which is largely representative of both high temperatures and very high rental levels.

Later this year, we may enhance our study with some classification tools, such as random forests and decision-trees which will lead us to develop accurate prediction models.

Appendix



In order to know whether the Seoul bike sharing service has been a success since its start-date, we compute two samples from the `SeoulBike` dataset : the first one dealing with the data related to December 2017 - the bike sharing service's start month - and the second one representing the November 2018 data which is the last month of the database.

Table 30: Overview of the two samples' third rows

Date	Rented.Bike.Count
2017-12-01	254
2017-12-01	204
2017-12-01	173
2018-11-01	584
2018-11-01	524
2018-11-01	362

By using the means of the `St_test` function we automated before, we compare the two samples' means to check whether the number of rented bikes is different between the two periods.

We compute the following t-test with a 5% first species risk.

$$\begin{cases} H_0 : & \mu_1 = \mu_2 \\ H_1 : & \mu_1 > \mu_2 \end{cases}$$

μ_1 stands for the second sample rented bike count's expected value and μ_2 the first one's.

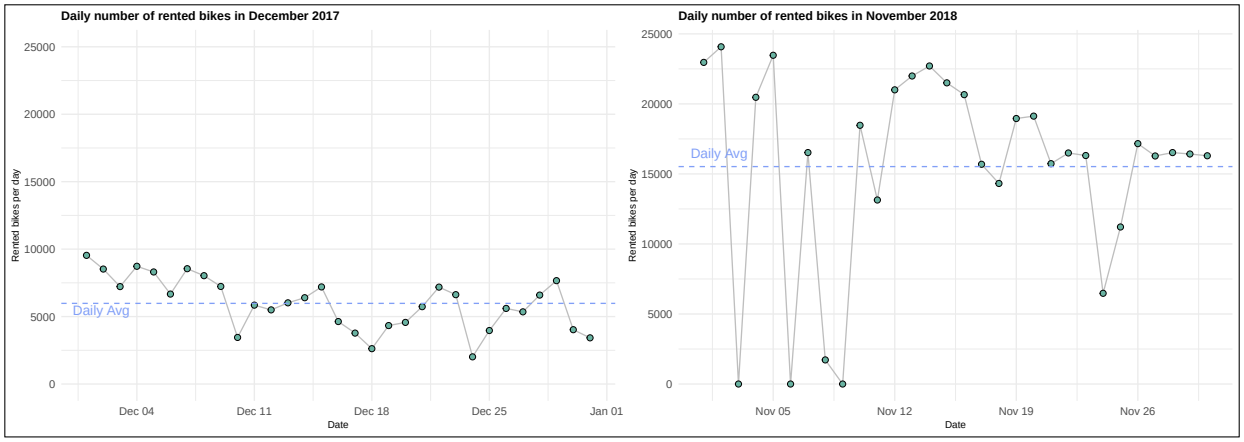
Table 31: Comparison of the 2 samples with a Student test

	mean in Dec 18	mean in Dec 17	p-value
Test results	646.8	249.1	6.779e-80

The t-test's p-value being basically equal to 0, it can be said that μ_1 is significantly higher than μ_2. In other words the number of rented bikes has significantly increased since its start-date.

The following plots depict this positive evolution. The daily count of rented bikes is plotted for both December 2017 and November 2018.

Computing the percent change between the two daily averages we found that the daily average rented bike count has increased by about 189%, that is to say it has almost been tripled over the period.



Resources

Principal component analysis

http://www.sthda.com/english/wiki/factoextra-r-package-easy-multivariate-data-analyses-and-elegant-visualization

https://rpkgs.datanovia.com/factoextra/reference/fviz_pca.html

https://medium.com/@hafezahmad/principal-component-analysis-pca-with-r-6ba954a54b34

https://en.wikipedia.org/wiki/Principal_component_analysis#Compute_the_cumulative_energy_content_for_each_eigenvector

Correspondence analysis

https://towardsdatascience.com/correspondence-analysis-using-r-cd57675ffc3a

Multiple correspondence analysis

http://www.sthda.com/french/articles/38-methodes-des-composantes-principales-dans-r-guide-pratique/75-acm-analyse-des-correspondances-multiples-avec-r-l-essentiel/

https://juliescholler.gitlab.io/publication/m1ade-2021/

Clustering

https://juliescholler.gitlab.io/publication/m1ade-2021/

https://www.statmethods.net/advstats/cluster.html

https://www.datanovia.com/en/courses/hierarchical-clustering-in-r-the-essentials/