Université Paris Cité

L3 MATHÉMATIQUES ET APPLICATIONS
Année 2023 - 2024

ÉTUDE DES VÉLIB À WASHINGTON

Écrit et pensé par :

EBO YEYE William
DERYS Ermilsonn
GALOUL Elias

Dirigé par :

COLLET Jérôme
MOZET Renaud

Table des matières

I. Introduction 3

II. Analyse des jeux de données 4

III. Analyse de la liaison entre la variable cnt et les autres variables du jeu de données 9

IV. Modèles de régressions linéaires simples 12

V. Modèle de régression linéaire multiple 14

VI. Réalisation d'une ACP puis d'une classification ascendante hiérarchique 17

VII. Conclusion 20

I. Introduction

Washington D.C est la capitale des Etats-Unis d'Amérique, 1ère puissance mondiale en termes de PIB. La capitale s'étend sur plus de 177km, soit 70km de plus que la capitale française. Washington tout comme Paris dispose de plusieurs moyens de transports publics notamment le bus ou le métro.

Aujourd'hui, nous allons nous intéresser à l'un de ces nombreux transports publics. En effet Washington dispose de son propre système de partage de vélo du nom de **Capital Bikeshare** crée en 2010 l'équivalant de **Vélib'** (Paris) à Washington.
Ce système permet aux utilisateurs de louer des vélos à partir de stations réparties dans la ville pour des trajets à courtes ou moyennes distances.
Les utilisateurs peuvent déverrouiller un vélo à partir d'une station, faire un trajet, puis le retourner à n'importe quelle autre station du système.
C'est un moyen populaire et pratique de se déplacer dans la région métropolitaine de Washington D.C.
Capital Bikeshare enregistre plus de 3.4 Millions de voyageurs par ans.
Parmi ces voyageurs on en trouve des occasionnels (par exemple des touristes) et des voyageurs récurrents (par exemple des étudiants, travailleurs etc…)

Nous avons 2 jeux de données que l'on traitera à l'aide de SAS afin d'analyser et de tirer certaines conclusions sur leur utilisation en fonctions de plusieurs facteurs que l'on explicitera.

L'objet de ce projet d'étude statistique va donc être d'analyser notre sujet à l'aide des logiciels SAS et R, afin d'en extraire les données et d'en faire ressortir les composantes statistiques.

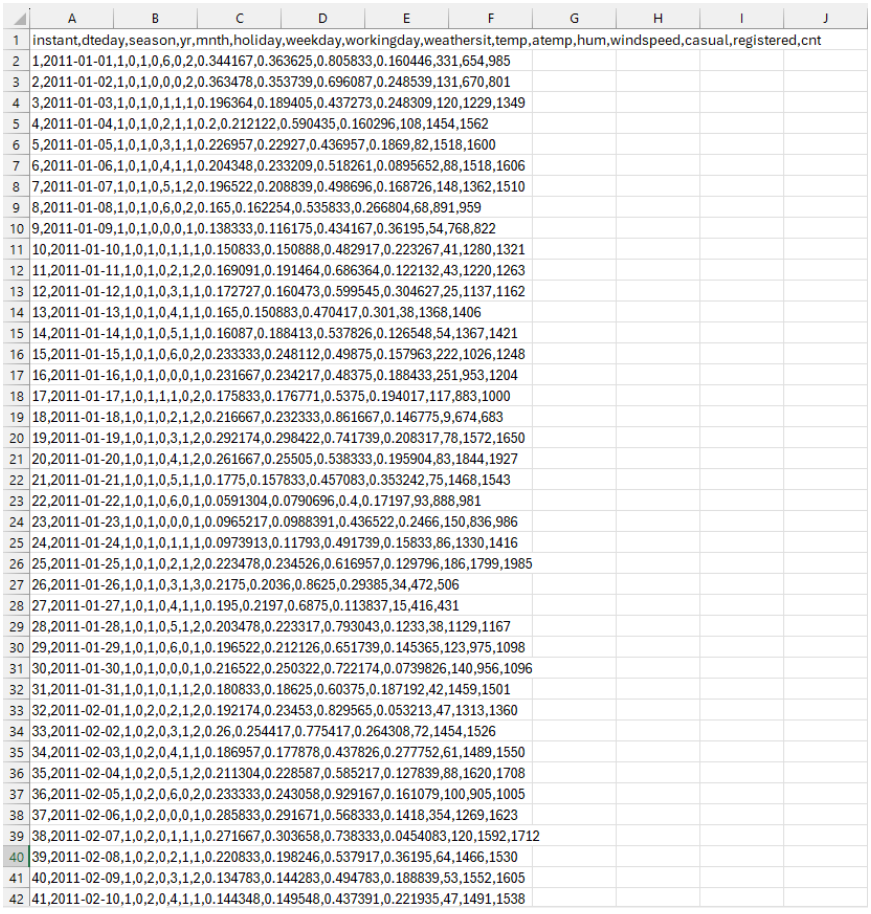
II. Analyse des jeux de données

Dans le cadre de ce projet, nous avons à disposition 3 fichiers :

- 2 jeux de données de type csv : « day.csv » et « hour.csv »
- Un fichier explicatif « readme.txt »

Les données d'utilisations des vélos ont été prises du 1er janvier 2011 au 31 décembre 2012.

Le fichier day.csv contient 731 observations, soit 1 observation par jour. Le fichier hour.csv contient 17379 observations, soit 1 observation par heure sur la période donnée.
Les fichiers se présentent de la sorte lorsqu'on les ouvre sur EXCEL :



	A	B	C	D	E	F	G	H	I	J
1	instant,dteday,season,yr,mnth,holiday,weekday,workingday,weathersit,temp,atemp,hum,windspeed,casual,registered,cnt									
2	1,2011-01-01,1,0,1,0,6,0,2,0.344167,0.363625,0.805833,0.160446,331,654,985									
3	2,2011-01-02,1,0,1,0,0,0,2,0.363478,0.353739,0.696087,0.248539,131,670,801									
4	3,2011-01-03,1,0,1,0,1,1,1,0.196364,0.189405,0.437273,0.248309,120,1229,1349									
5	4,2011-01-04,1,0,1,0,2,1,1,0.2,0.212122,0.590435,0.160296,108,1454,1562									
6	5,2011-01-05,1,0,1,0,3,1,1,0.226957,0.22927,0.436957,0.1869,82,1518,1600									
7	6,2011-01-06,1,0,1,0,4,1,1,0.204348,0.233209,0.518261,0.0895652,88,1518,1606									
8	7,2011-01-07,1,0,1,0,5,1,2,0.196522,0.208839,0.498696,0.168726,148,1362,1510									
9	8,2011-01-08,1,0,1,0,6,0,2,0.165,0.162254,0.535833,0.266804,68,891,959									
10	9,2011-01-09,1,0,1,0,0,0,1,0.138333,0.116175,0.434167,0.36195,54,768,822									
11	10,2011-01-10,1,0,1,0,1,1,1,0.150833,0.150888,0.482917,0.223267,41,1280,1321									
12	11,2011-01-11,1,0,1,0,2,1,2,0.169091,0.191464,0.686364,0.122132,43,1220,1263									
13	12,2011-01-12,1,0,1,0,3,1,1,0.172727,0.160473,0.599545,0.304627,25,1137,1162									
14	13,2011-01-13,1,0,1,0,4,1,1,0.165,0.150883,0.470417,0.301,38,1368,1406									
15	14,2011-01-14,1,0,1,0,5,1,1,0.16087,0.188413,0.537826,0.126548,54,1367,1421									
16	15,2011-01-15,1,0,1,0,6,0,2,0.233333,0.248112,0.49875,0.157963,222,1026,1248									
17	16,2011-01-16,1,0,1,0,0,0,1,0.231667,0.234217,0.48375,0.188433,251,953,1204									
18	17,2011-01-17,1,0,1,1,1,0,2,0.175833,0.176771,0.5375,0.194017,117,883,1000									
19	18,2011-01-18,1,0,1,0,2,1,2,0.216667,0.232333,0.861667,0.146775,9,674,683									
20	19,2011-01-19,1,0,1,0,3,1,2,0.292174,0.298422,0.741739,0.208317,78,1572,1650									
21	20,2011-01-20,1,0,1,0,4,1,2,0.261667,0.25505,0.538333,0.195904,83,1844,1927									
22	21,2011-01-21,1,0,1,0,5,1,1,0.1775,0.157833,0.457083,0.353242,75,1468,1543									
23	22,2011-01-22,1,0,1,0,6,0,1,0.0591304,0.0790696,0.4,0.17197,93,888,981									
24	23,2011-01-23,1,0,1,0,0,0,1,0.0965217,0.0988391,0.436522,0.2466,150,836,986									
25	24,2011-01-24,1,0,1,0,1,1,1,0.0973913,0.11793,0.491739,0.15833,86,1330,1416									
26	25,2011-01-25,1,0,1,0,2,1,2,0.223478,0.234526,0.616957,0.129796,186,1799,1985									
27	26,2011-01-26,1,0,1,0,3,1,3,0.2175,0.2036,0.8625,0.29385,34,472,506									
28	27,2011-01-27,1,0,1,0,4,1,1,0.195,0.2197,0.6875,0.113837,15,416,431									
29	28,2011-01-28,1,0,1,0,5,1,2,0.203478,0.223317,0.793043,0.1233,38,1129,1167									
30	29,2011-01-29,1,0,1,0,6,0,1,0.196522,0.212126,0.651739,0.145365,123,975,1098									
31	30,2011-01-30,1,0,1,0,0,0,1,0.216522,0.250322,0.722174,0.0739826,140,956,1096									
32	31,2011-01-31,1,0,1,0,1,1,2,0.180833,0.18625,0.60375,0.187192,42,1459,1501									
33	32,2011-02-01,1,0,2,0,2,1,2,0.192174,0.23453,0.829565,0.053213,47,1313,1360									
34	33,2011-02-02,1,0,2,0,3,1,2,0.26,0.254417,0.775417,0.264308,72,1454,1526									
35	34,2011-02-03,1,0,2,0,4,1,1,0.186957,0.177878,0.437826,0.277752,61,1489,1550									
36	35,2011-02-04,1,0,2,0,5,1,2,0.211304,0.228587,0.585217,0.127839,88,1620,1708									
37	36,2011-02-05,1,0,2,0,6,0,2,0.233333,0.243058,0.929167,0.161079,100,905,1005									
38	37,2011-02-06,1,0,2,0,0,0,1,0.285833,0.291671,0.568333,0.1418,354,1269,1623									
39	38,2011-02-07,1,0,2,0,1,1,1,0.271667,0.303658,0.738333,0.0454083,120,1592,1712									
40	39,2011-02-08,1,0,2,0,2,1,1,0.220833,0.198246,0.537917,0.36195,64,1466,1530									
41	40,2011-02-09,1,0,2,0,3,1,2,0.134783,0.144283,0.494783,0.188839,53,1552,1605									
42	41,2011-02-10,1,0,2,0,4,1,1,0.144348,0.149548,0.437391,0.221935,47,1491,1538									

Capture d'écran du fichier day.csv ouvert dans Microsoft Excel

Très compliqué à lire, une importation sur un logiciel adapté comme SAS est indispensable

Lignes totales : 731 Colonnes totales : 16 Lignes 1-100

	instant	dteday	season	yr	mnth	holiday
1	1	11-01-01	1	0	1	0
2	2	11-01-02	1	0	1	0
3	3	11-01-03	1	0	1	0
4	4	11-01-04	1	0	1	0
5	5	11-01-05	1	0	1	0
6	6	11-01-06	1	0	1	0
7	7	11-01-07	1	0	1	0
8	8	11-01-08	1	0	1	0
9	9	11-01-09	1	0	1	0
10	10	11-01-10	1	0	1	0
11	11	11-01-11	1	0	1	0
12	12	11-01-12	1	0	1	0
13	13	11-01-13	1	0	1	0
14	14	11-01-14	1	0	1	0
15	15	11-01-15	1	0	1	0
16	16	11-01-16	1	0	1	0
17	17	11-01-17	1	0	1	1
18	18	11-01-18	1	0	1	0
19	19	11-01-19	1	0	1	0
20	20	11-01-20	1	0	1	0
21	21	11-01-21	1	0	1	0
22	22	11-01-22	1	0	1	0

Capture d'écran du fichier day.csv lu sur SAS

Le fichier day sera le fichier que l'on étudiera étant donné que les deux fichiers sont très similaires et on tire les mêmes conclusions.

Nos deux fichiers sont relativement similaires en termes de variables, Le fichier day.csv comporte 16 variables et le fichier hour.csv en comporte 17.

Voici les différentes variables présentes :

- Instant : index des enregistrements
- dteday : La date du jour
- season : Les saisons numérotés de 1 à 4
  1 = Printemps
  2 = Été
  3 = Automne
  4 = Hiver
- yr : Année numérotés de 0 à 1
  0 = 2011
  1 = 2022
- mnth : Mois numérotés de 1 à 12
- hr (présente uniquement dans le fichier hour.csv) : Heures numérotés de 0 à 23
- holiday : Jour férié
- weekday : Jour de la semaine numérotées de 0 à 6
- workingday : Si le jour n'est ni un week-end ni un jour férié, la valeur est 1, sinon la valeur est 0

- weathersit : Le temps
  - 1 : Clair, Peu de nuages
  - 2 : Brume et nuageux
  - 3 : Neige légère, Pluie légère
  - 4 : Forte pluie + neige abondante, orage, brouillard
- temp : la température normalisée par sa valeur maximale (40)
- atemp : Ressenti normalisé par la valeur maximale (50)
- hum : L'humidité normalisée par sa valeur maximale (100)
- windspeed : La vitesse du vent en mph normalisée par sa valeur maximale (67)
- casual : Le nombre d'utilisateurs occasionnels
- registered : Le nombre d'abonné
- cnt : le nombre total de vélo loué

Nous allons nous intéresser sur l'évolution de la variable cnt
La variable **cnt** est la somme des valeurs de registered et de casual. **Elle sera notre variable dépendante.**
Toutes les autres seront traités en tant que variables explicatives.
En termes de **variables quantitatives** nous avons : cnt, registered, casual, temp, atemp, hum, windspeed
En termes de **variables qualitatives (catégorielles)** nous avons : instant, dteday, season, yr, mnth, hr, holiday, weekday, weathersit.

A l'aide de la procédure MEANS, nous avons plusieurs données concernant la variable cnt.

La procédure MEANS

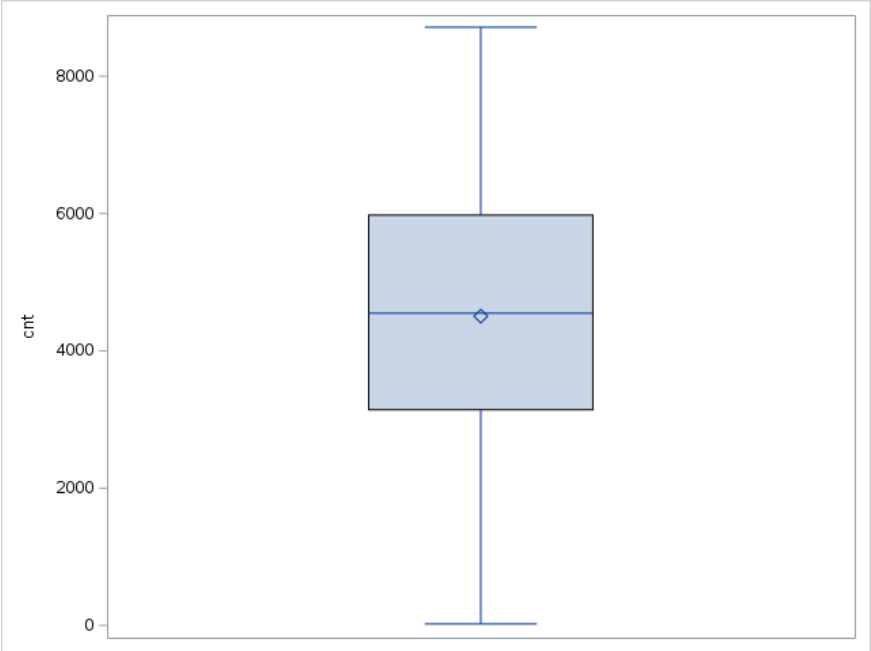
Variable d'analyse : cnt

N	Moyenne	Médiane	Mode	Ec-type	Variance	Erreur type	Minimum	Quartile inférieur	Quartile supérieur	Maximum	Intervalle quartile	Skewness	Kurtosis
731	4504.349	4548.000	1096.000	1937.211	3752788.208	71.650	22.000	3141.000	5976.000	8714.000	2835.000	-0.047	-0.812

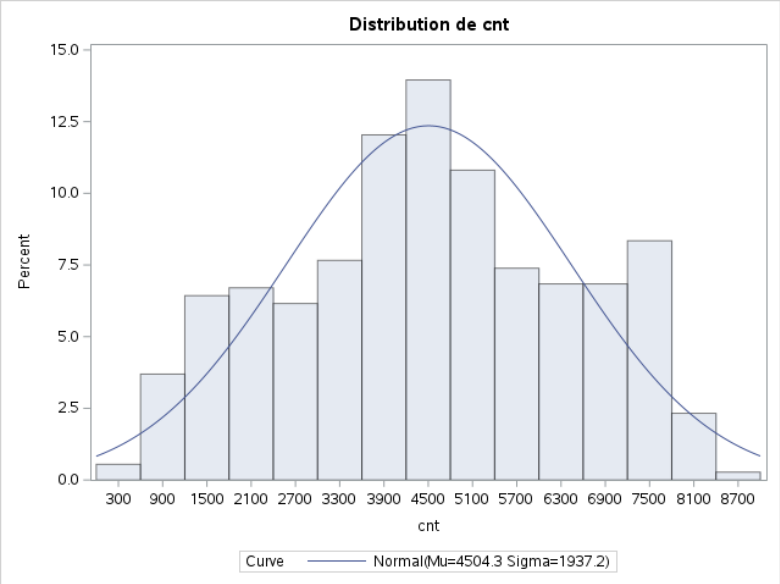
Résultat de la procédures MEANS

En analysant le tableau, on remarque que la moyenne des utilisateurs par jour est de 4504.349, sa médiane de 4548, le minimum de 22 et le pic de cyclistes est de 8714.

La boîte à moustache générée par la procédure SGPLOT nous renseigne sur la répartition globale de la variable cnt, et sur les valeurs aberrantes.



Aucun point se situe en dehors la boite, il y a donc aucune valeur aberrante.



Distribution de la variable cnt

L'histogramme de la distribution de la variable count nous permet de poser **l'hypothèse H0 suivante : « La variable cnt suit une loi normale »** Nous allons la vérifier à l'aide de la procédure *UNIVARIATE*

Parameters for Normal Distribution		
Paramètre	**Symbole**	**Estimation**
Mean	Mu	4504.349
Std Dev	Sigma	1937.211

Goodness-of-Fit Tests for Normal Distribution				
Test	**Statistique**		**p-value**	
Kolmogorov-Smirnov	**D**	0.04705760	**Pr > D**	<0.010
Cramer-von Mises	**W-Sq**	0.30192828	**Pr > W-Sq**	<0.005
Anderson-Darling	**A-Sq**	2.96988261	**Pr > A-Sq**	<0.005

Test de normalité de la variable cnt

Toutes les p-valeurs sont <0.05, donc **l'hypothèse H0 : « La variable cnt suit une loi normale »** est rejetée.

Nous allons donc accepter l'**hypothèse H1 : La variable cnt ne suit pas une loi normale**

III. Analyse de la liaison entre la variable cnt et les autres variables du jeu de données

Analysons maintenant la relation de corrélation entre la variable cnt et les autres variables du jeu de données.

Coefficients de corrélation de Pearson, N = 731 Proba > |r| sous H0: Rho=0	
	cnt
instant	0.62883 <.0001
dteday	0.62883 <.0001
season	0.40610 <.0001
yr	0.56671 <.0001
mnth	0.27998 <.0001
holiday	-0.06835 0.0648
weekday	0.06744 0.0684
workingday	0.06116 0.0985
weathersit	-0.29739 <.0001
temp	0.62749 <.0001
atemp	0.63107 <.0001
hum	-0.10066 0.0065
windspeed	-0.23454 <.0001
casual	0.67280 <.0001
registered	0.94552 <.0001
cnt	1.00000

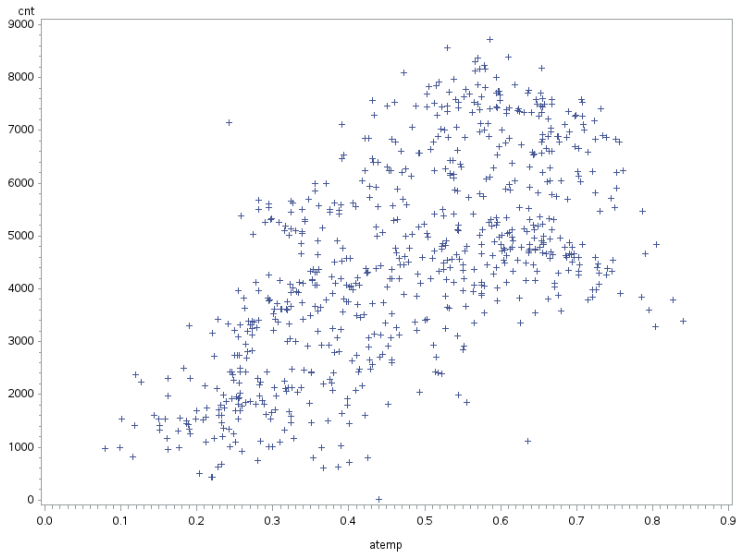
Tableau de corrélation entre les différentes variables

La matrice de corrélation nous affiche pour chaque variable 2 données très importante.

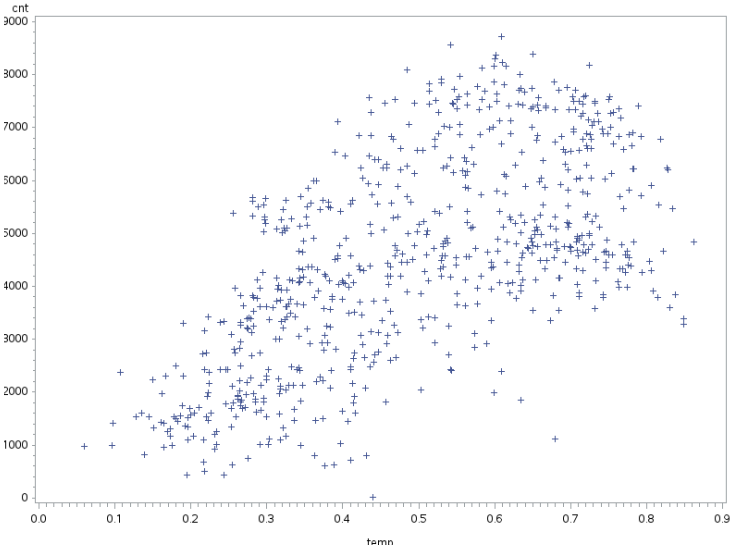
- Le coefficient de corrélation de Pearson : une valeur comprise entre -1 et 1 qui mesure la corrélation entre chaque paire de variable
- La p-value associée au test de Pearson : elle permet d'évaluer une association entre deux variables.
  Si la p-value < 0.05, alors les 2 variables sont corrélées. Elles ne le sont pas dans le cas inverse.

On remarque que la date, les saisons, l'année, le mois, la température et la température ressentie exercent une influence positive sur le nombre d'utilisateurs

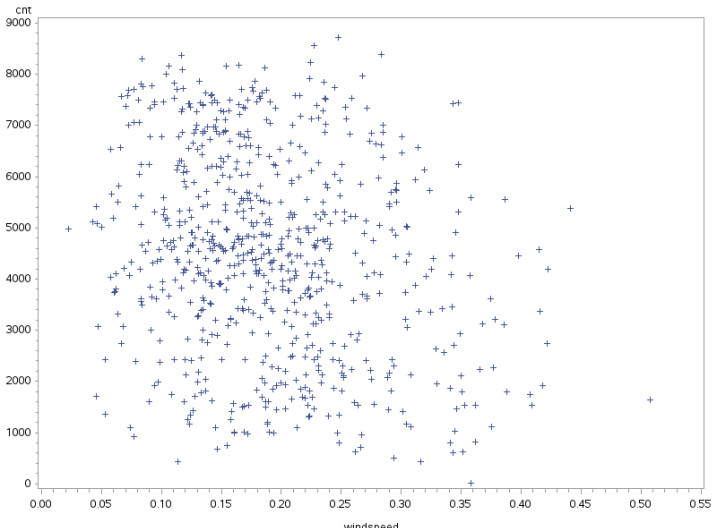
avec une p-value <0.0001.
Le vent, la météo, les vacances et l'humidité, en revanche, toujours avec une p-value <0.0001, exercent une influence négative sur le nombre d'usager.
Au total, ce sont 10 valeurs qui influent plus ou moins notre variable
Les autres ayant une p-value > 0.05, ne sont pas significative.



Nombre d'usager en fonction du ressenti



Nombre d'usager en fonction de la température



Nombre d'usager en fonction de la température

Sur ces 3 graphiques réalisées à l'aide de GPLOT, on voit très bien le lien de corrélation entre les différentes variables sur notre variable dépendante cnt.

Tous nos tests ont été réalisés avec le fichier day.csv, mais il est important de préciser que les résultats sont globalement les mêmes. Le poids de certains coefficients change : les facteurs météorologiques sont plus significatifs lorsqu'on prend une journée entière.
L'heure qui n'est pas présente sur le fichier day.csv exerce une influence significative.

Coefficients de corrélation de Pearson, N = 17379 Proba > |r| sous H0: Rho=0	
	cnt
instant	0.27838 <.0001
dteday	0.27775 <.0001
season	0.17806 <.0001
yr	0.25049 <.0001
mnth	0.12064 <.0001
hr	0.39407 <.0001
holiday	-0.03093 <.0001
weekday	0.02690 0.0004
workingday	0.03028 <.0001
weathersit	-0.14243 <.0001
temp	0.40477 <.0001
atemp	0.40093 <.0001
hum	-0.32291 <.0001
windspeed	0.09323 <.0001
casual	0.69456 <.0001
registered	0.97215 <.0001
cnt	1.00000

Matrice de corrélation de la variable cnt sur le fichier hour.csv

Nous avons mis volontairement de côté les variables casual et registered, la somme de ces 2 variables étant égale à la variable count.

IV. Modèles de régressions linéaires simples

À présent nous allons procéder à plusieurs régressions linéaires sur la variable cnt. Pour rappel, nous avons :
la date, l'heure, les saisons, l'année, le mois, la température, la température ressentie, le vent, la météo, les vacances et l'humidité comme **variable significative**.
Avec 12 variables significatives, nous allons faire le choix d'utiliser 3 variables quantitatives, à savoir : **l'humidité (hum), la vitesse du vent (windspeed) et la température ressenti (atemp).**

Maintenant, posons **H_0 : « La variable ne joue pas un rôle significatif sur le nombre d'usager »**
et **H_1 : « La variable joue un rôle significatif sur le nombre d'usager »**

Nb d'observations lues	731
Nb d'obs. utilisées	731

Analyse de variance

Source	DDL	Somme des carrés	Moyenne quadratique	Valeur F	Pr > F
Modèle	1	1091003307	1091003307	482.45	<.0001
Erreur	729	1648532085	2261361		
Total sommes corrigées	730	2739535392			

Root MSE	1503.78219	R carré	0.3982
Moyenne dépendante	4504.34884	R car. ajust.	0.3974
Coeff Var	33.38512		

Paramètres estimés

Variable	DDL	Valeur estimée des paramètres	Erreur type	Valeur du test t	Pr > |t|
Intercept	1	945.82398	171.29147	5.52	<.0001
atemp	1	7501.83395	341.53825	21.96	<.0001

Régression linéaire avec la variable cnt et atemp

La procédure REG
Modèle : MODEL1
Variable dépendante : cnt

Nb d'observations lues	731
Nb d'obs. utilisées	731

Analyse de variance

Source	DDL	Somme des carrés	Moyenne quadratique	Valeur F	Pr > F
Modèle	1	27757373	27757373	7.46	0.0065
Erreur	729	2711778019	3719860		
Total sommes corrigées	730	2739535392			

Root MSE	1928.69389	R carré	0.0101
Moyenne dépendante	4504.34884	R car. ajust.	0.0088
Coeff Var	42.81848		

Paramètres estimés

Variable	DDL	Valeur estimée des paramètres	Erreur type	Valeur du test t	Pr > |t|
Intercept	1	5363.98645	322.67868	16.62	<.0001
hum	1	-1369.08066	501.19088	-2.73	0.0065

Régression linéaire avec la variable cnt et hum

La procédure REG
Modèle : MODEL1
Variable dépendante : cnt

Nb d'observations lues	731
Nb d'obs. utilisées	731

Analyse de variance

Source	DDL	Somme des carrés	Moyenne quadratique	Valeur F	Pr > F
Modèle	1	150705556	150705556	42.44	<.0001
Erreur	729	2588829836	3551207		
Total sommes corrigées	730	2739535392			

Root MSE	1884.46462	R carré	0.0550
Moyenne dépendante	4504.34884	R car. ajust.	0.0537
Coeff Var	41.83656		

Paramètres estimés

Variable	DDL	Valeur estimée des paramètres	Erreur type	Valeur du test t	Pr > |t|
Intercept	1	5621.15288	185.06238	30.37	<.0001
windspeed	1	-5862.91276	899.98814	-6.51	<.0001

Régression linéaire avec la variable cnt et windspeed

Toutes les p-values sont inférieures à 0.05, donc on rejette l'hypothèse nulle pour toutes les variables et on accepte donc H_1.

Une droite de régression linéaire s'écrit **Y = aX+b** avec

- Y : la variable expliquée, dans notre cas cnt
- a : le coefficient de X qui détermine la pente de la droite.
- X : notre variable explicative
- b : l'intercept

Variable explicative (X)	La pente (a)	Intercept (b)	Équation	R carré	Valeur F
atemp	7501.83	945.82	*Y = 7501.83*X + 945.82*	0.39	482.45
hum	-1369.08	5363.98	*Y = -1369.08*X + 5363.98*	0.01	7.46
windspeed	-5862.91	5621.15	*Y = -5862.91*X + 5621.15*	0.05	30.37

Les valeurs sont arrondies au centième près

Une grande valeur F associé à une P-value inférieur à 0.05 indique généralement que le modèle de régression est statistiquement significatif.
Avec un coefficient de détermination de 0.39 et une valeur F très haute, on peut constater que atemp est la variable qui explique le mieux l'évolution de la location de vélo.
39.82% de la variation de la location est expliquée par ce modèle.
À contrario de l'humidité et la vitesse du vent semble expliquer qu'une toute partie de la location.

Une régression linéaire multiple qui regroupe certaines de nos variables qualitatives et quantitatives semble plus adaptée.

V. Modèle de régression linéaire multiple

Un modèle de régression multiple est une extension du modèle de régression linéaire simple. En effet c'est le même principe ou plusieurs variables indépendantes sont utilisées pour expliquer une variable dépendante. Il s'exprime par une équation qui est la suivante :

$Y = \beta0 + \beta1*X1 + \beta2*X2 + ... + \beta n*Xn + \varepsilon$

Où :

- Y est la variable expliquée, dans notre cas c'est cnt
- X1, X2, ..., Xn sont les variables indépendantes.
- β0, β1, β2, ..., βn sont les coefficients de régression qui représentent l'effet de chaque variable indépendante sur la variable dépendante.
- ε est le terme d'erreur, qui capture les écarts entre les valeurs observées et les valeurs prédites par le modèle.

Avec SAS en plus de la proc reg, on adopte la méthode stepwise afin que le logiciel choisisse automatiquement les variables pertinentes.
Pour commencer, nous incorporons toutes nos variables (qualitatives et quantitatives) et on laisse le logiciel opérer.
On pose :
H_0 : « Le modèle ne convient pas à l'analyse »
et **H_1 : « Le modèle convient à notre analyse »**

Analyse de variance					
Source	DDL	Somme des carrés	Moyenne quadratique	Valeur F	Pr > F
Modèle	9	2157612019	239734669	297.03	<.0001
Erreur	721	581923373	807106		
Total sommes corrigées	730	2739535392			

La valeur de F étant de 297.03 et la P-valeur < 0.001, l'hypothèse H_1 est donc adoptée. On peut donc assurer avec certitude que **le modèle est correct** et que ce dernier **explique donc une part significative de la variance dans la variable dépendante.**

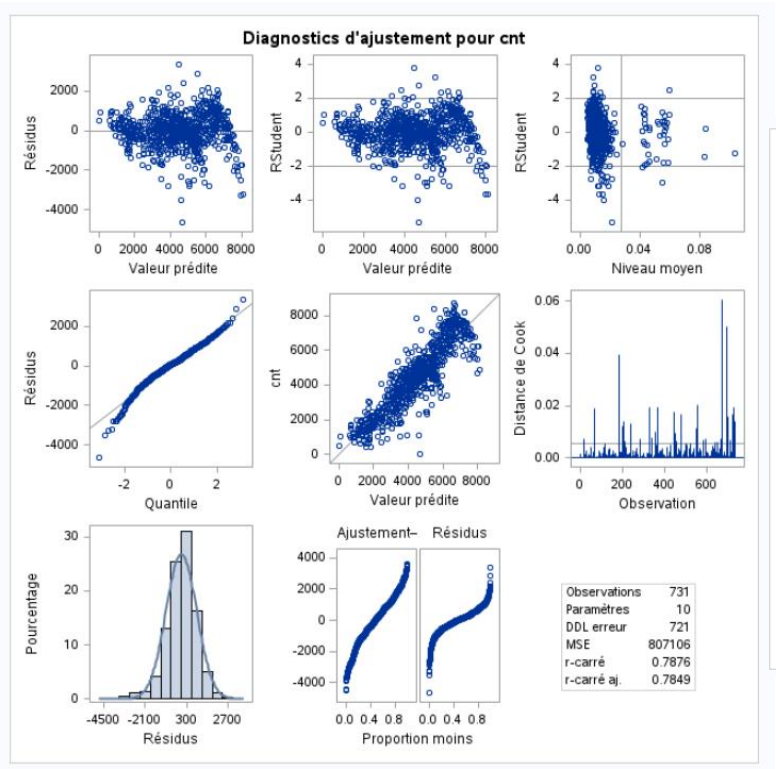
Synthèse de Sélection Stepwise								
Etape	Variable entrée	Variable supprimée	Nombre var. dans	R carré partiel	R carré du modèle	C(p)	Valeur F	Pr > F
1	atemp		1	0.3982	0.3982	1314.81	482.45	<.0001
2	dteday		2	0.2903	0.6886	331.706	678.67	<.0001
3	weathersit		3	0.0487	0.7373	168.418	134.80	<.0001
4	yr		4	0.0151	0.7524	119.223	44.24	<.0001
5	season		5	0.0282	0.7806	25.5155	93.20	<.0001
6	holiday		6	0.0035	0.7841	15.5888	11.79	0.0006
7	hum		7	0.0014	0.7855	12.7166	4.84	0.0281
8	mnth		8	0.0011	0.7866	11.0381	3.67	0.0559
9	workingday		9	0.0010	0.7876	9.7472	3.29	0.0700

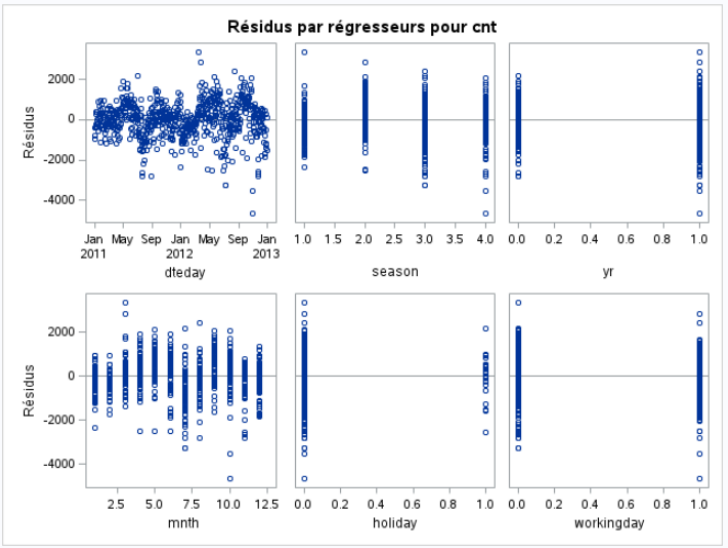
Variable	Valeur estimée des paramètres	Erreur type	SC Type II	Valeur F	Pr > F
Intercept	160345	70440	4182129	5.18	0.0231
dteday	-8.56907	3.78502	4136783	5.13	0.0239
season	522.16831	56.25121	69548498	86.17	<.0001
yr	5188.17167	1387.01160	11292732	13.99	0.0002
mnth	226.81136	116.84594	3041115	3.77	0.0526
holiday	-597.54394	205.91526	6796623	8.42	0.0038
workingday	134.49900	74.12843	2657046	3.29	0.0700
weathersit	-680.15382	78.85707	60043169	74.39	<.0001
atemp	5910.03872	226.54946	549269042	680.54	<.0001
hum	-632.98944	308.96869	3387619	4.20	0.0409

On observe que pour nos 9 variables, seules 7 p-values associées sont < 0.005. Par conséquent, dans notre modèle de régression, nous ne prendrons en compte que les variables : **atemp, dteday, weathersit, yr, season, holiday, hum**. Ici, **environ 78.54% de la variation des valeurs de cnt est expliqué par ces 7 variables**

Ainsi, grâce aux valeurs estimées des paramètres, nous pouvons écrire **l'équation de régression suivante :**

*Cnt = 160345 - 8.56 * (dteday) + 522.16 * (season) + 5188.17 * (yr) + 226.8 * (mnth) – 597.54 * (holiday) + 134.49 * (workingday) – 680.15 * (weatherlist) + 5910.03 * (atemp) – 632.98 * (hum)*

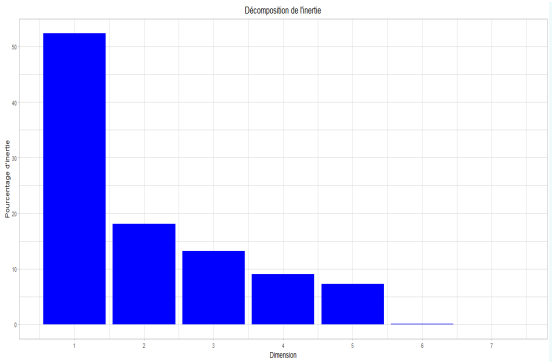




Nous n'observons aucun pattern caractéristique sur les résidus, on peut donc en conclure que notre estimation par régression est correcte.

VI. Réalisation d'une ACP puis d'une classification ascendante hiérarchique

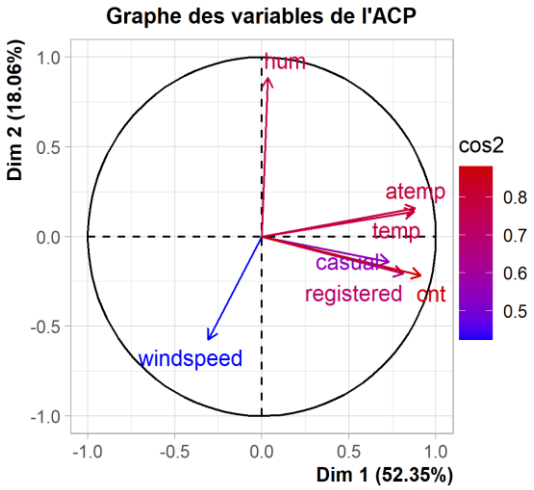
Grâce à la librairie factoextra, nous allons réaliser une ACP normée sur nos 7 variables quantitatives pour que toutes les variables aient le même poids lors de la réalisation de cette ACP. Voici ci-dessous la répartition de la variance expliquée en pourcentage et les valeurs propres qui désignent les dimensions :



eigenvalue	percentage of variance	cumulative percentage of variance
3.66	52.35	52.35
1.26	18.06	70.41
0.92	13.20	83.60
0.63	9.06	92.66
0.51	7.23	99.89
0.01	0.11	100.00
0.00	0.00	100.00

Nous pouvons voir sur ce graphique que les deux premières dimensions expliquent ⅔ de la variance, autrement dit, cela signifie que ⅔ de la variabilité du nuage des individus est représentée dans ce plan. Par conséquent, dans le cadre de notre ACP, nous choisissons arbitrairement de conserver uniquement les deux premières dimensions.

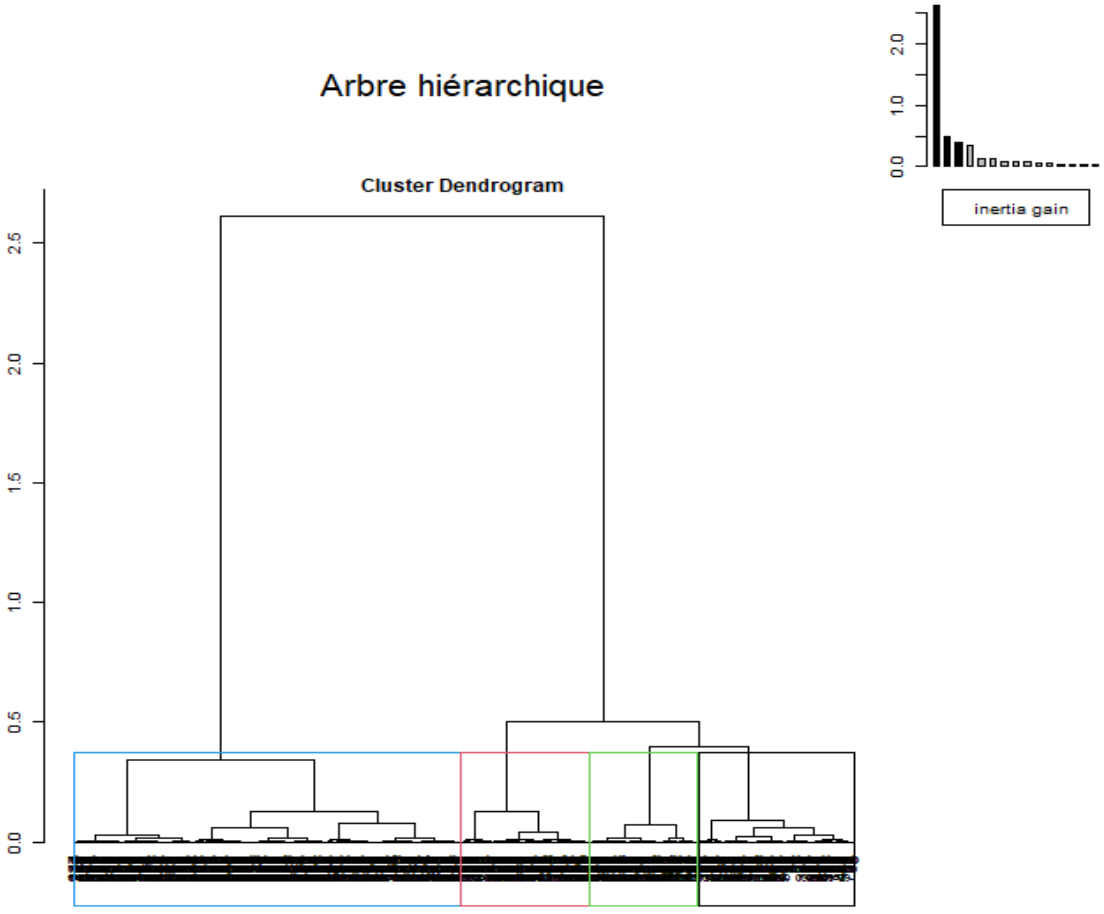
Maintenant, nous allons pouvoir passer à une interprétation graphique de notre analyse en composantes principales :

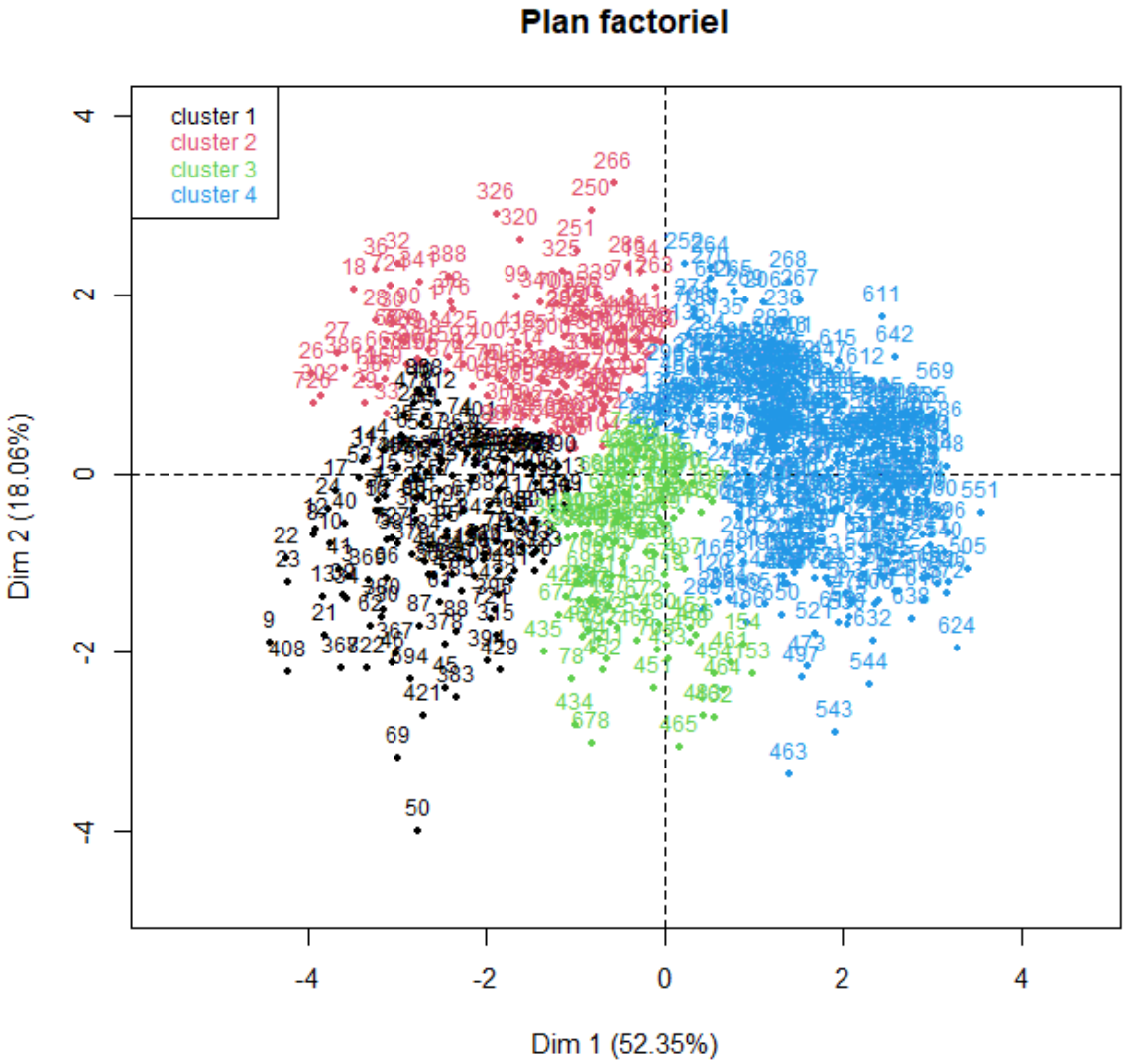


Avec ce graphique, nous pouvons constater une forte corrélation entre la température et la totalité des vélos loués. On peut aussi voir que la vitesse du vent et l'humidité influent beaucoup moins sur la location de vélo. De même, il y a une corrélation entre

les voyageurs occasionnels, récurrents et la somme des deux (pareil pour la température et le ressenti). Ceci et, cependant logique. De plus, on peut voir que les variables *windspeed* et *casual* sont moins importantes (valeur de cos2 basse). On voit aussi que plus il y a de vélos loués plus la température est haute et moins la vitesse de vent est élevée.

Analysons maintenant la classification ascendante hiérarchique de notre jeu de donné.





On remarque une perte d'inertie conséquente jusqu'à la dimension 4. On va donc décider de séparer nos variables en 4 clusters.
Malheureusement, nous avons une trop grande quantité de données pour pouvoir analyser nos données sur le dendrogramme

VII. Conclusion

Tout d'abord nous avons pris l'initiative de traiter spécifiquement le fichier *day.csv* pour plus de clarté.
Ensuite nous avons procédé à une **analyse descriptive des jeux de données** afin de mieux comprendre les données que nous manipulons.
Dans la partie 3 nous avons procédé à une **étude corrélationnelle** entre nos variables et la variable **cnt**.
Avec **la matrice de corrélation** on constate que 10 variables ont un impact significatif sur cnt.

Puis dans la partie 4 et 5 nous avons pu sortir 7 variables significatives grâce aux différents **modèles de régressions linéaires**.
Dans un premier temps nous avons essayé **plusieurs modèles de régressions linéaires simples**, avant de s'adonner à un modèle de régression linéaire multiple.
Ce **modèle de régression multiple** associé à la méthode *STEPWISE* s'avère beaucoup plus efficace.

L'ACP nous donne un résultat qui confirme les résultats trouvés via les régressions linéaires et les corrélations observées.
L'ACP montre cependant que le vent exerce une influence moindre sur la variable cnt et ses composantes casual et registered. On peut malgré tout affirmer via nos autres modèles qu'un temps venteux à une importance significative.

Le vélo étant un moyen de locomotion qui peut associer à un loisir, les facteurs météorologiques et temporels ont forcément un impact sur le nombre de vélo loué.