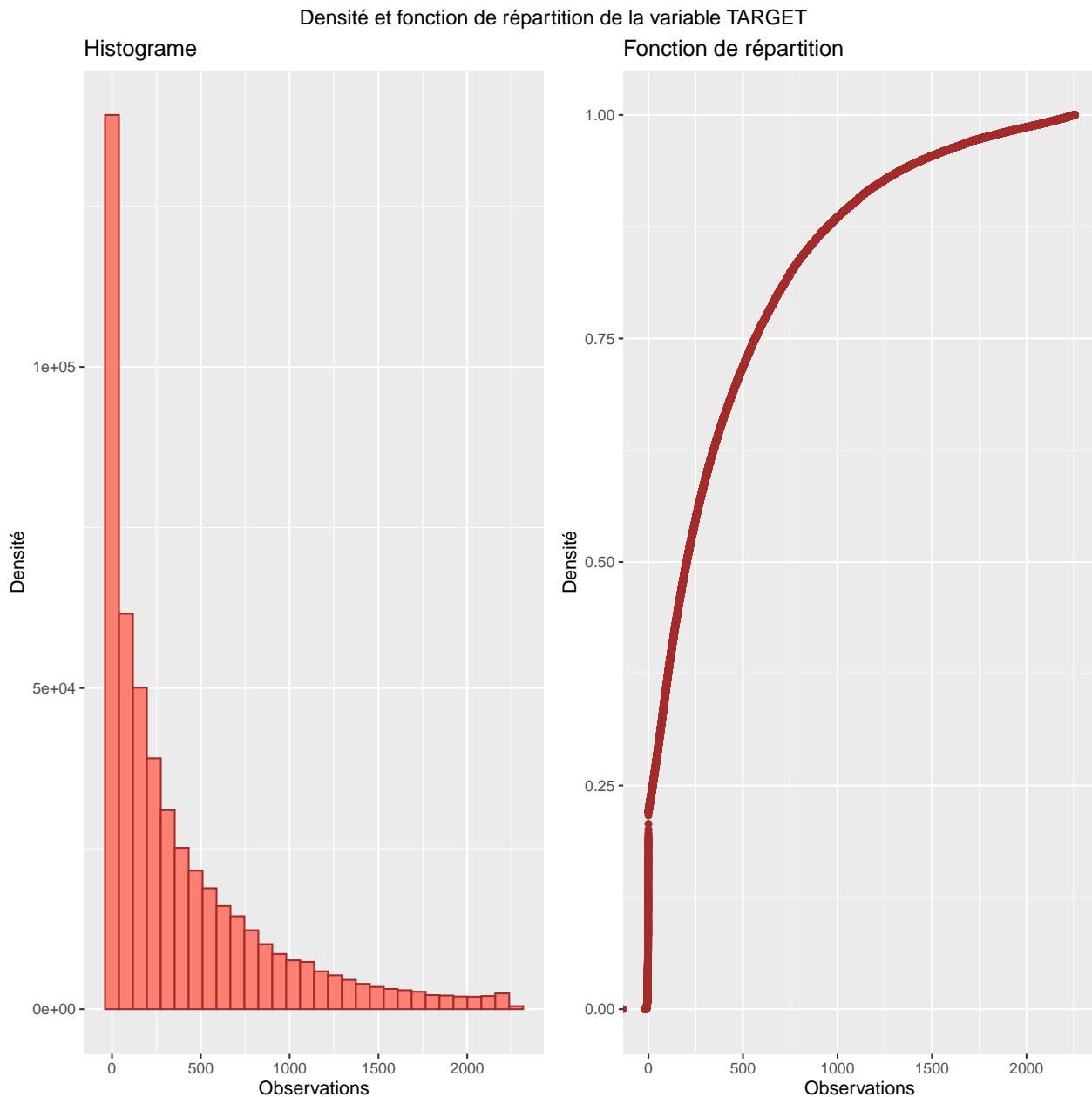
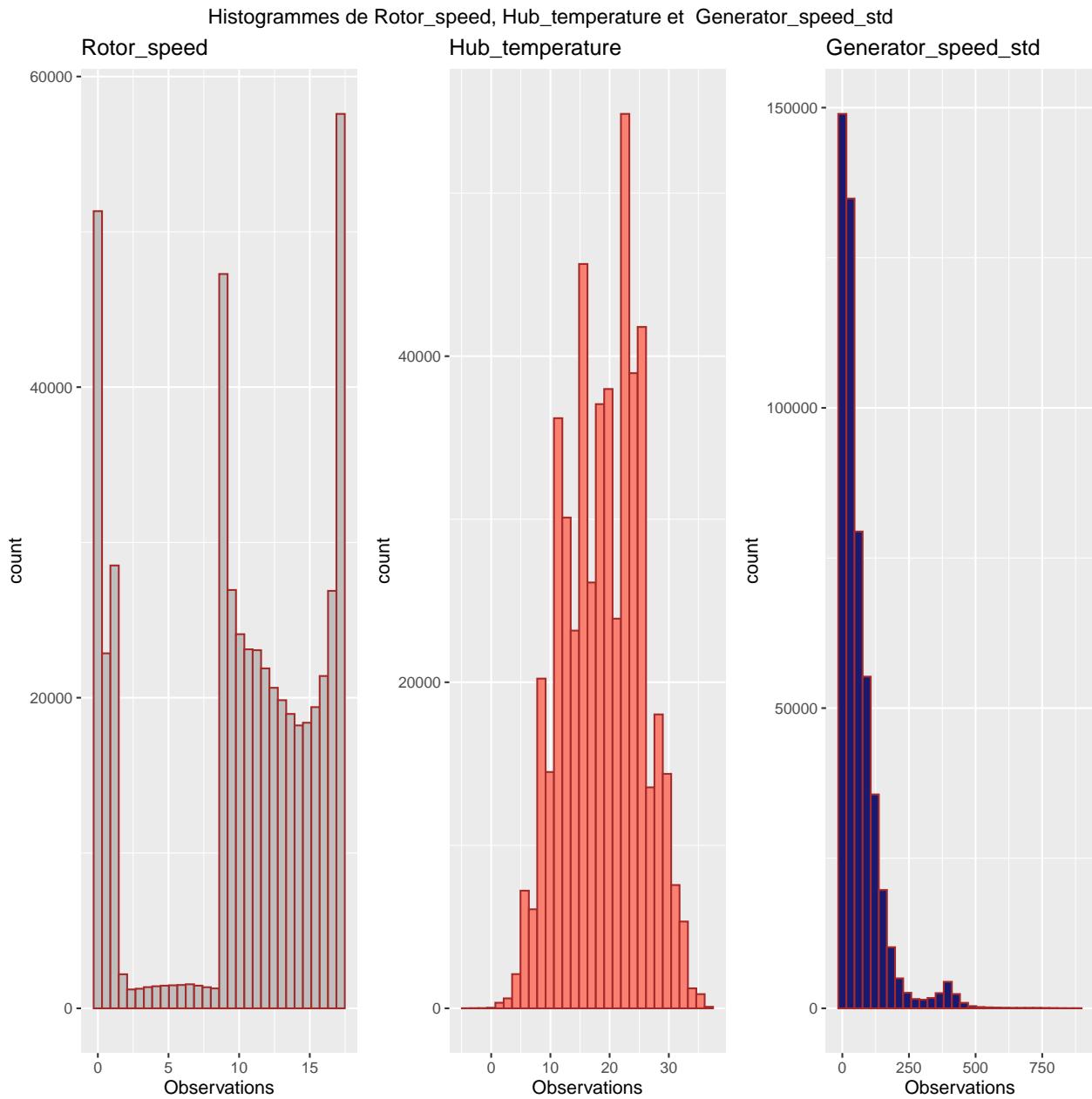


La variable Target



D'après l'histogramme ci-dessus, la variable TARGET prend ses valeurs entre -20 et 2000 , avec une concentration autour de zéro.

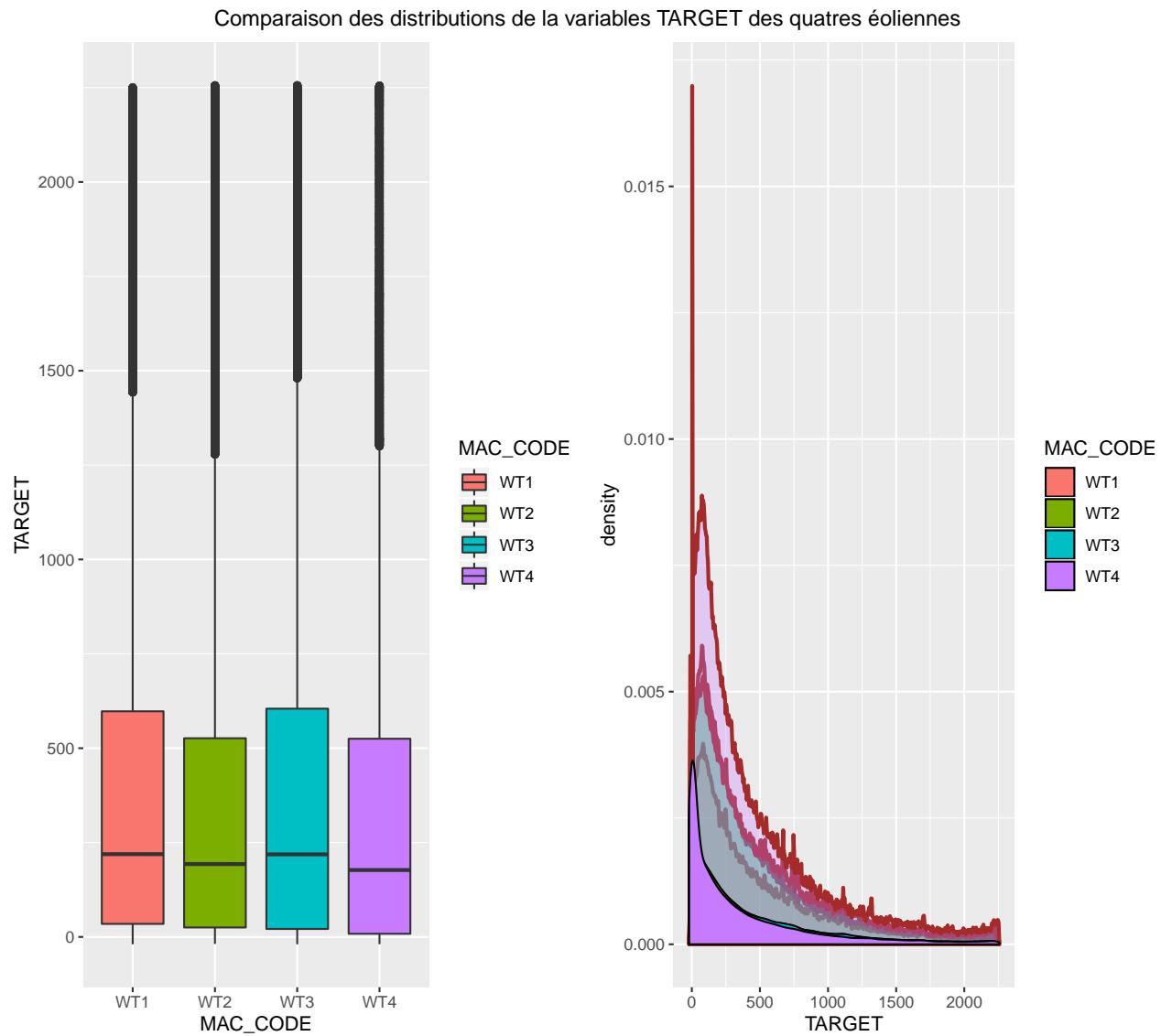


Les graphiques ci-dessus représentent les histogrammes des variables **Rotor_speed**, **Hub_temperature** et **Generator_speed_std**.

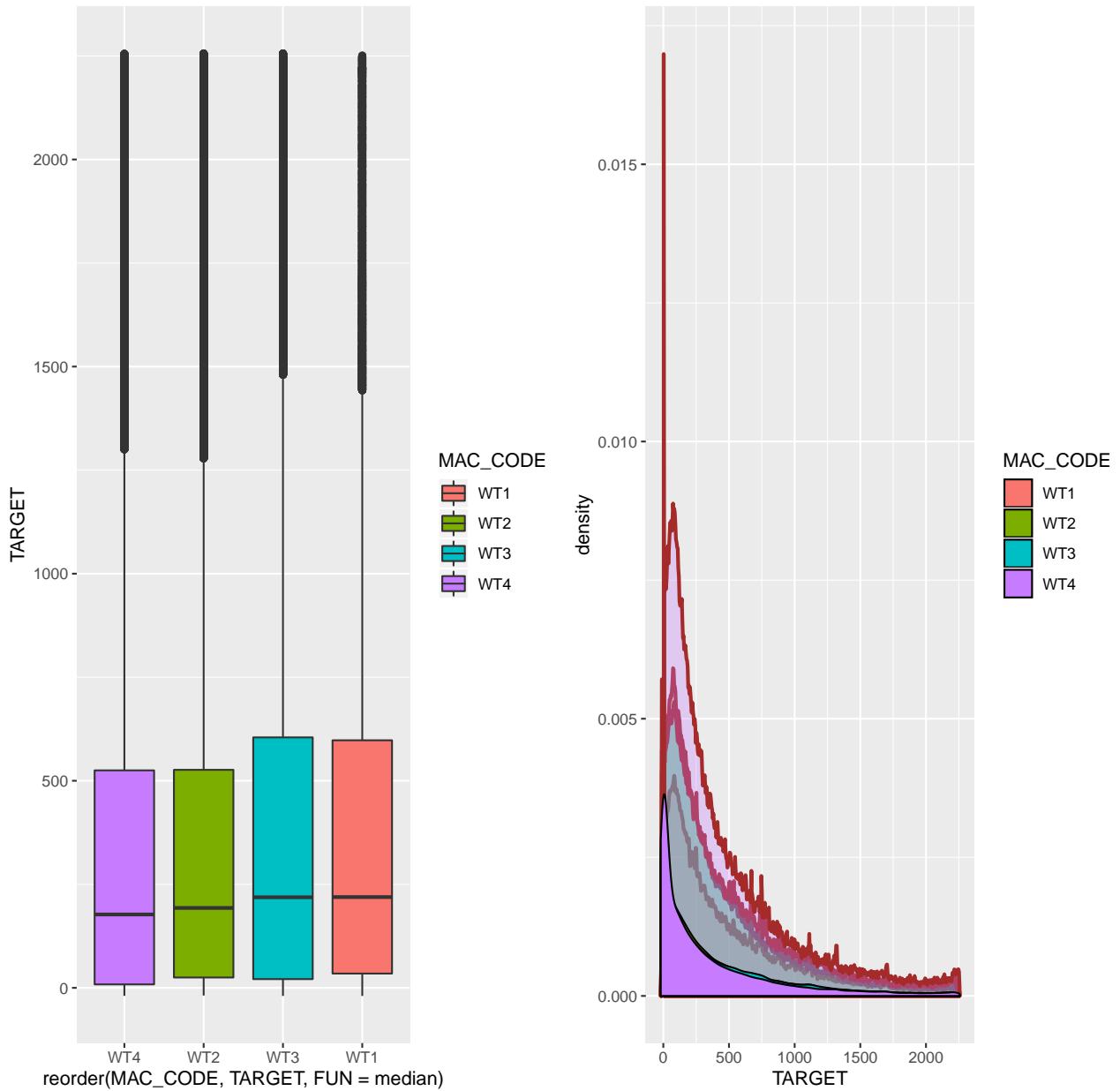
D'abord la variable **Rotor_speed** ne prend que des valeurs positives, dont plus de 80% sont dans $[0, 2.5] \cup [8, 16]$. Ensuite, nous remarquons que **Hub_temperature** est unimodale et asymétrique.

Enfin, la variable **Generator_speed_std** prend que des valeurs positives, de plus elle est unimodale et asymétrique.

Analyse descriptive bivariée



Comparaison des distributions de la variables TARGET des quatres éoliennes



D'après les boxplot et les graphes de densités ci-dessus, nous remarquons que les quantiles de la variable TARGET sont légèrement différents d'une éolienne à l'autre, de même pour les densités. La variable TARGET n'a donc pas la même loi pour les quatres éoliennes.

Analyse descriptive multivariée

ACP normée des données

Variable	Disperssion
Rotor_speed_std	7.550396e-01
Grid_voltage_std	4.583308e+00
Generator_speed_std	7.925428e+01

Variable	Disperssion
Date_time	4.881169e+04

Tableau illustrant la disperssion de quelques variables

Après avoir observé la disperssion des variables, nous avons constaté que les échelles sont très différentes. Nous avons des variables de l'ordre 10^{-1} jusqu'à l'ordre 10^4 , il faut donc normaliser.

Afin d'identifier les variables qui expliquent le mieux notre jeu de données, les variables corrélées et réduire les dimensions, nous allons appliquer une ACP normée. Pour le calcul de l'ACP nous allons utiliser les packages "FactoMiner" et "factoextra".

Les valeurs propre mesurent la quantité de variance expliquée par chaque axe principal, nous allons donc les examiner afin de déterminer le nombre de composantes principales à prendre en compte.

Composante	Valeur propre	Pourcentage de variance	Pourcentage cumulé de variance
comp 1	9.785503e+00	2.878089e+01	28.78089
comp 2	3.958601e+00	1.164294e+01	40.42383
comp 3	3.035529e+00	8.928027e+00	49.35186
comp 4	2.241573e+00	6.592862e+00	55.94472
comp 5	1.518758e+00	4.466934e+00	60.41166
comp 6	1.501787e+00	4.417020e+00	64.82868
comp 7	1.105757e+00	3.252226e+00	68.08090
comp 8	1.048219e+00	3.082998e+00	71.16390
comp 9	1.000882e+00	2.943771e+00	74.10767
comp 10	9.868843e-01	2.902601e+00	77.01027

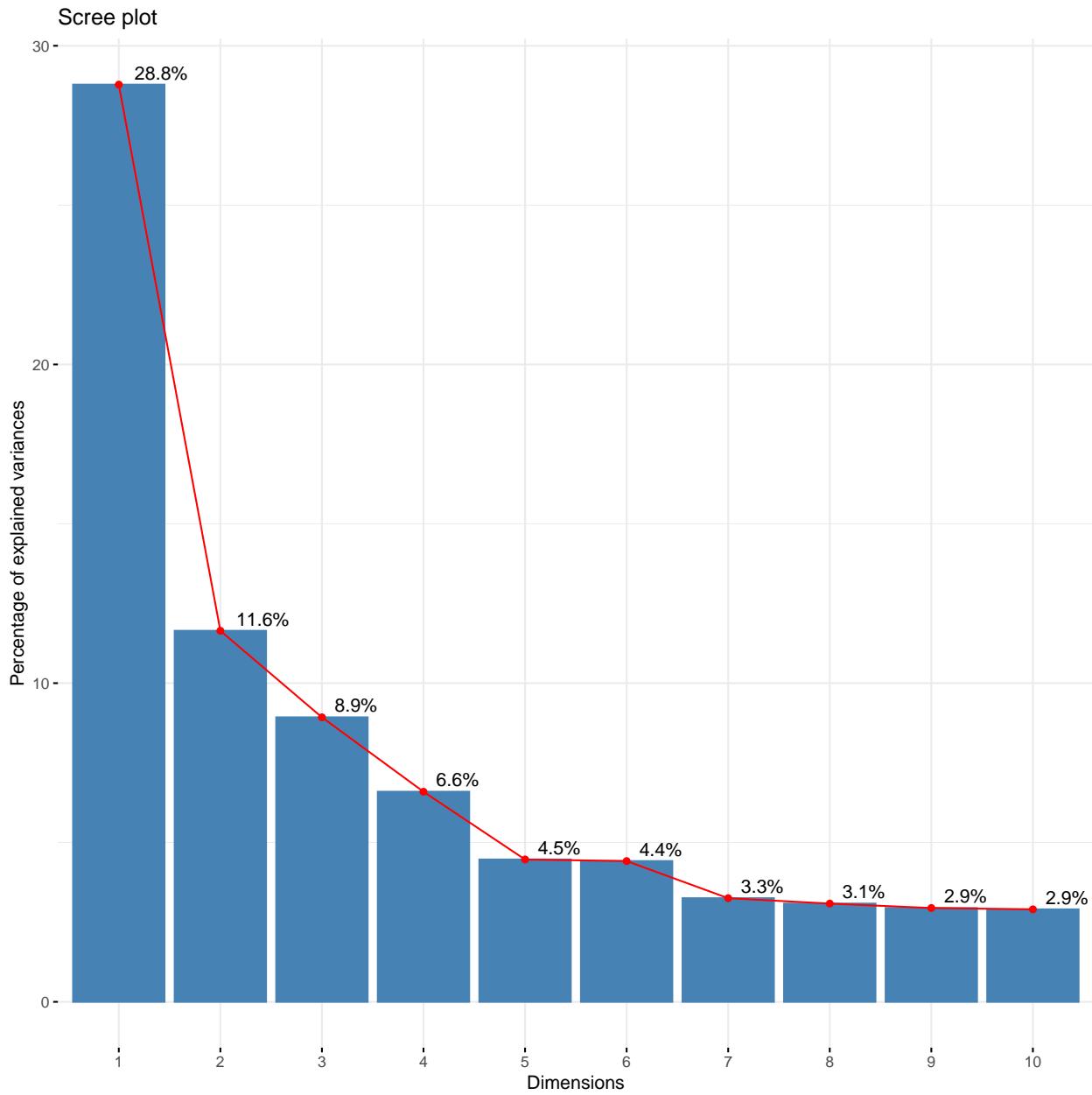
Tableau des valeurs propres de quelques composantes

Dans l'ACP normée, la règle pour choisir la dimension est :

$$q = \max \{l/\lambda_l \geq 1\}$$

D'après le tableau ci-dessus, le nombre d'axes principaux à conserver est : $q = 9$.

Une autre manière de déterminer le nombre de composantes principales est de regarder le graphique des valeurs propres (scree plot).

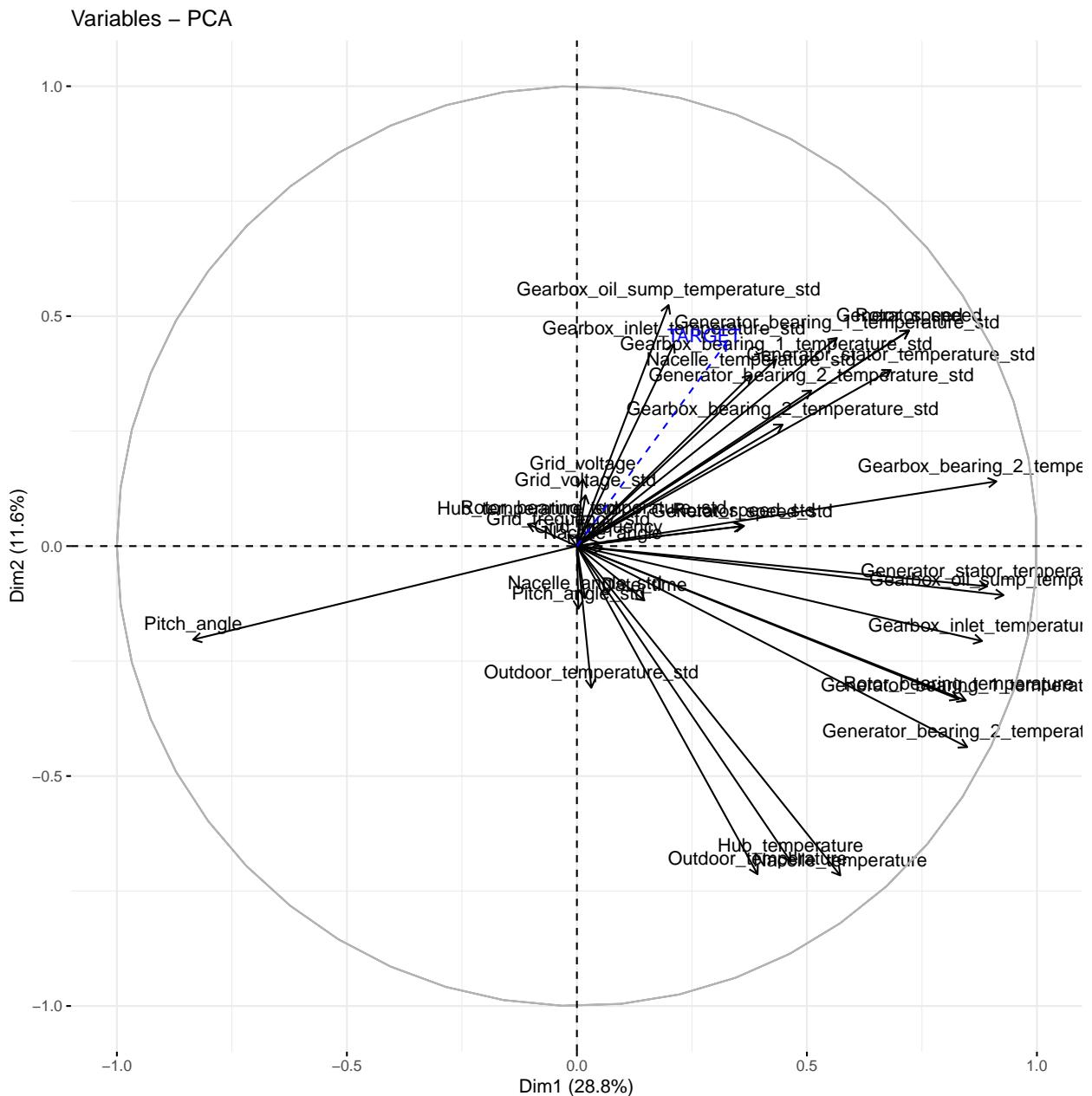


D'après le graphique à partir de $q = 9$ ou $q = 10$ les valeurs propres restantes sont toutes relativement petites, on peut donc choisir de garder les dix premiers axes.

Etude du premier plan factoriel

- Nuage des variables

Nous utiliserons le cercle de corrélation pour visualiser les variables :



Le graphique ci-dessus illustre les relations entre toutes les variables, prenons par exemple :

- Les variables `Gearbox_bearing_2_temperature_std`, `Generator_bearing_2_temperature_std` et `Grid_voltage` sont regroupées, elles sont donc positivement corrélées. En revanche, elles sont négativement corrélées avec la variable `Pitch_angle`.
 - Les variables `Hub_temperature` et `Gearbox_oil_sump_temperature_std` ne sont pas corrélées.
 - Les variables `Hub_temperature` et `Generator_bearing_2_temperature` sont loin de l'origine, elles sont donc bien représentées par l'ACP.
 - Qualité de représentation des variables

La qualité de représentation des variables est mesurée par \cos^2 . Ci-dessus le bar-plot correspondant :