

Analyse des données ENGIE

Alain MANA

23 Octobre 2019

Table des matières

Introduction	2
Traitement des données manquantes	2
Pre-sélection de variables	2
Traitement des valeurs aberrantes	6
Analyse descriptive	7
Analyse descriptive univariée	7
Analyse descriptive bivariée	10
Analyse descriptive multivariée	11
Recodage de la variable MAC_CODE	17
Erreur de prédiction utilisée	17
Sélection exaustive de variables	19
Régression RIDGE	21
Arbre de régression	23
Conclusion	27

Introduction

L'énergie éolienne constitue un moyen propre et renouvelable pour produire de l'électricité, utilisée depuis plusieurs siècles (moulins à vent), cette dernière s'est particulièrement développée depuis l'avènement de l'éolien électrique (seconde moitié du XXe siècle).

Premier acteur éolien en France, ENGIE fait du développement de l'éolien une de ses priorités et vise un objectif de 2 000 MW de puissance installée.

Détecter des écarts anormaux entre la production électrique attendue et la production réalisée aide à mieux optimiser la production de chaque éolienne. Dans l'optique d'améliorer la production d'électricité éolienne nous allons analyser les données fournies par ENGIE sur la plateforme Data challenge dans le cadre du cours "Analyse de données et apprentissage (M2 Ingénierie Statistique et Data Science UPMC&ISUP)".

Dans un premier temps nous étudierons les données, puis nous chercherons à mettre en place un premier modèle, nous comparerons ensuite ce dernier avec d'autres modèles et nous essayerons d'améliorer le modèle sélectionné.

#Descriptif des données

Les données qui seront utilisées pour ce projet sont composées de 617386 observations réparties en deux fichiers. Le premier `input` contient les 78 variables explicatives et le deuxième fichier contient la variable à expliquer `Target`. Chaque ligne de notre jeu de données est définie par un identifiant unique `ID`, lié à une éolienne `MAC_CODE` et un pas de temps `Data_time`.

L'objectif, est de prédire à partir de l'ensemble de données la puissance active des éoliennes.

#Pre-traitement des données

Nous commençons par une brève description des données via les fonctions `summary`, `glimpse`, etc. Nous remarquons que la variable `MAC_CODE` est une variable factorielle, le reste des variables sont quantitatives continues. La variable `Target` étant continue, nous cherchons donc à faire une régression.

Traitemet des données manquantes

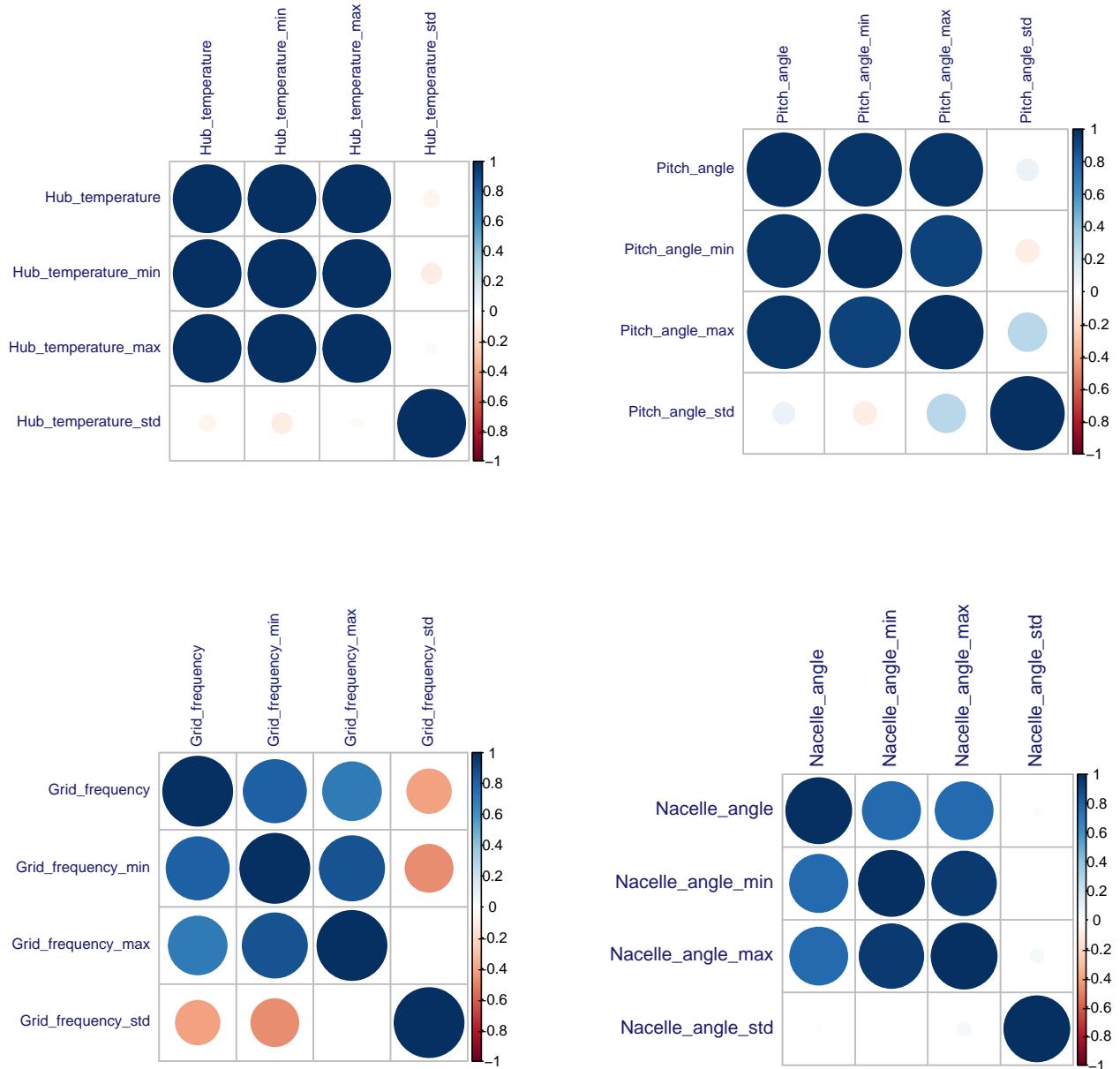
Le jeu de données contient 469944 valeurs manquantes réparties sur 109416 lignes. Il existe plusieurs méthodes pour l'imputation des données manquantes : plus proche voisin médian, Amélia, MissForest ... etc.

Comme le jeu de données est volumineux et la capacité de notre machine est limitée, nous allons tout simplement supprimer toutes les lignes contenant des NA's.

Pre-sélection de variables

Nous avons remarqué que pour toutes les variables nous avons le min (minimum), le max (maximum) et le std (écart-type). D'abord, nous allons prendre quelques exemples de variables : `Pitch_angle`, `Hub_temperature`, `Nacelle_angle` et `Grid_frequency`. Puis, visualiser leurs lien avec leurs `min`, `max` et `std` respectifs.

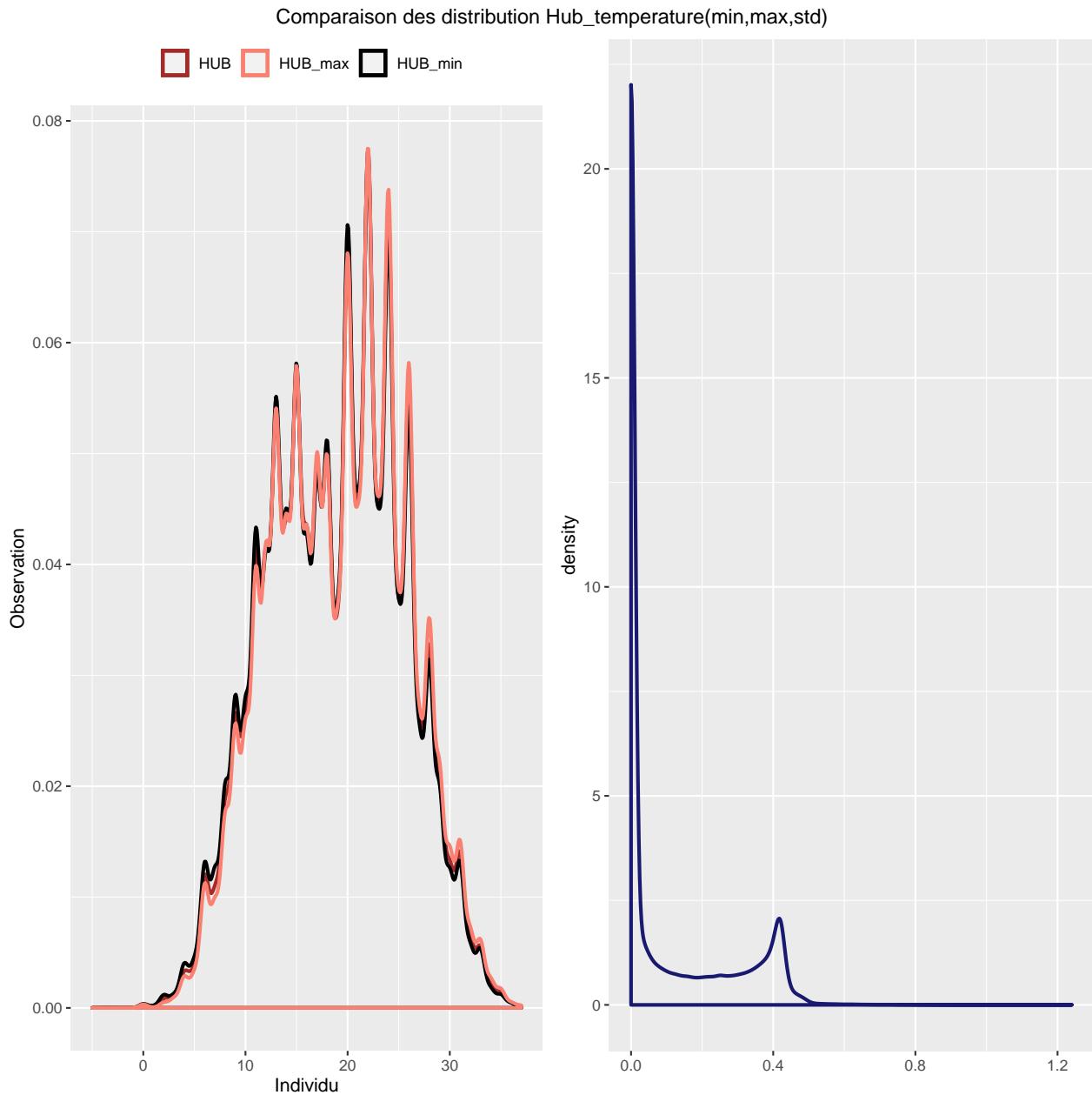
Matrice de corrélation



D'après la matrice de corrélation ci-dessus, les variables `Pitch_angle`, `Hub_temperature`, `Nacelle_angle` et `Grid_frequency` sont très fortement corrélées avec leurs `min` et `max` respectifs. En revanche, ce n'est pas le cas avec leurs `std`.

```
#cor.ci(corr11,method = "kendall")
```

Maintenant, prenons l'exemple de la variable `Hub_temperature` et comparaons sa distribution à celles de `Hub_temperature_min`, `Hub_temperature_max` et `Hub_temperature_std`.

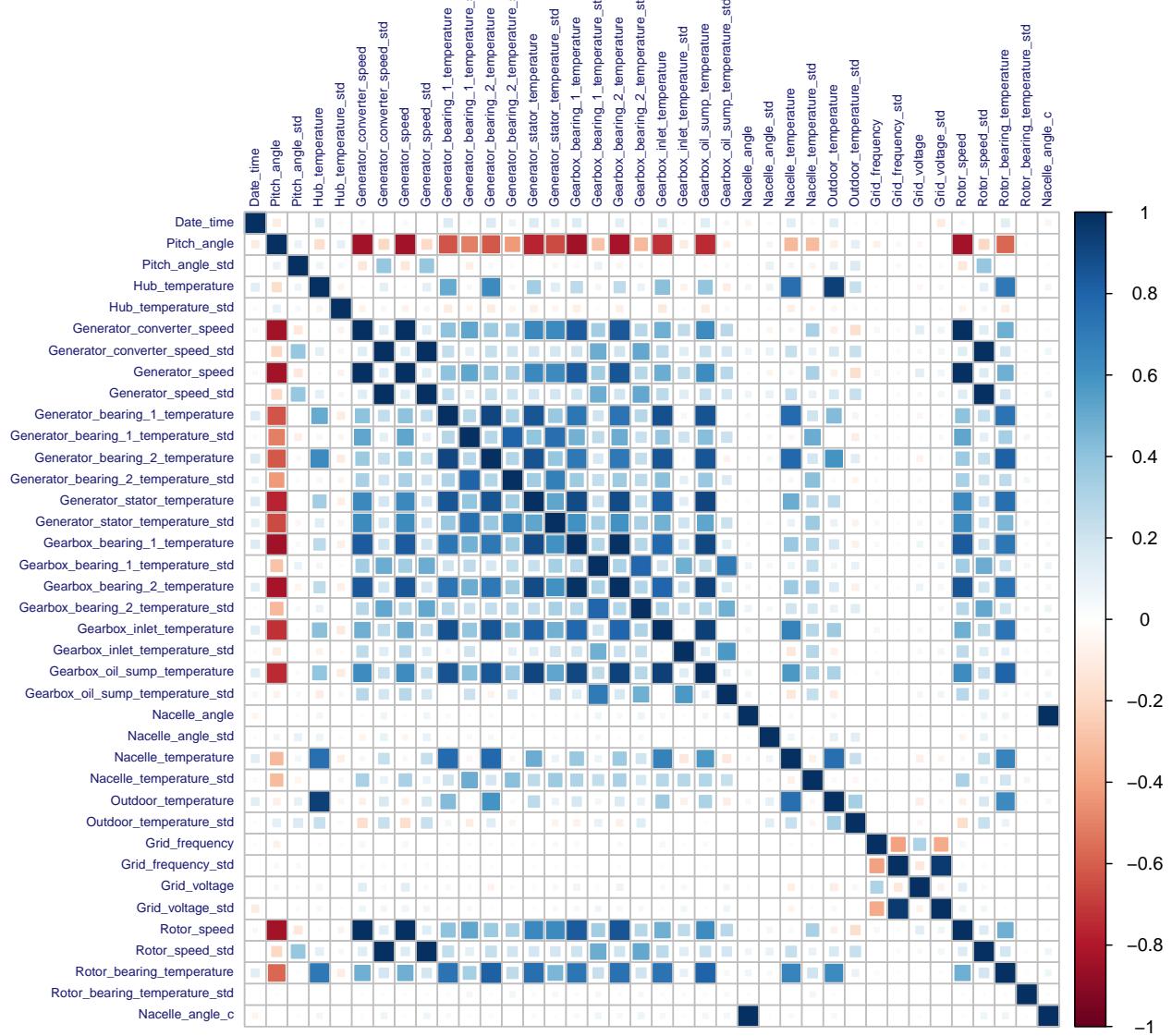


D'après les graphiques ci-dessus, les variables `Hub_temperature`, `Hub_temperature_min` et `Hub_temperature_max` ont une même distribution de densité et prennent leurs valeurs dans l'intervalle [0, 50]. Par contre `Hub_temperature_std` a une distribution différente des précédentes et elle prend ses valeurs dans l'intervalle [0, 0.5]. Afin d'éviter la redondance, il suffit de prendre `Hub_temperature` et `Hub_temperature_std`.

Nous avons eu des résultats sémilaire pour toutes les autres variable, donc pour la suite, nous allons supprimer toutes les variables min et max.

Maintenant, observons la matrice de corrélation des variables restantes après suppression des variables `min` et `max`.

Matrice de corrélation



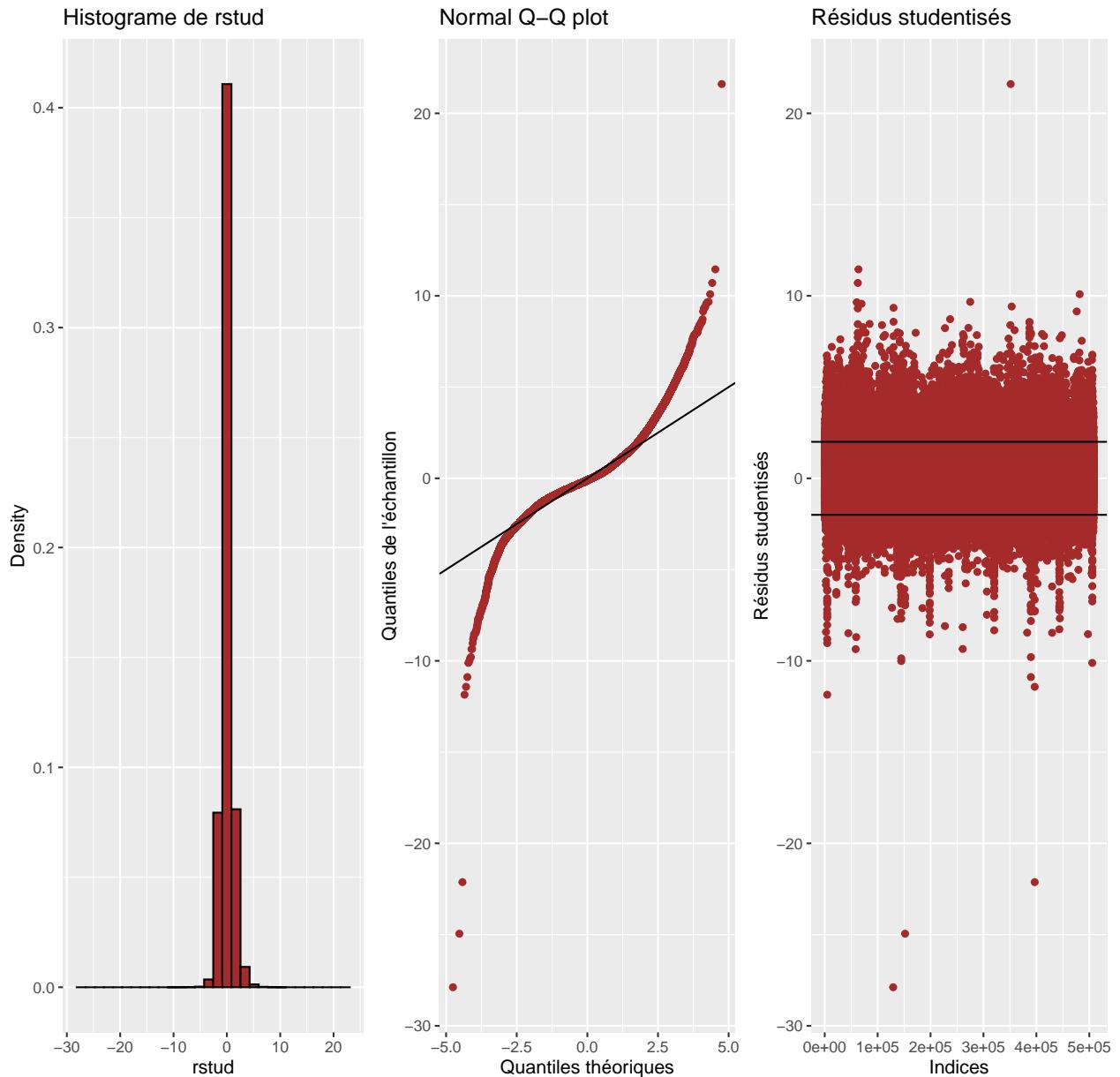
D'après la matrice de corrélation :

- Les variables `Generator_converter_speed` et `Generator_speed` sont très fortement corrélées.
- Les variables `Generator_converter_speed_std` et `Generator_speed_std` sont très fortement corrélées.
- Les variables `Nacelle_angle_c` et `Nacelle_angle` sont très fortement corrélées.

Nous allons donc supprimer les variables : `Generator_converter_speed`, `Generator_converter_speed_std` et `Nacelle_angle_c`.

Traitement des valeurs aberrantes

L'objectif de cette partie est d'étudier tous les points atypiques et éventuellement supprimer des individus.

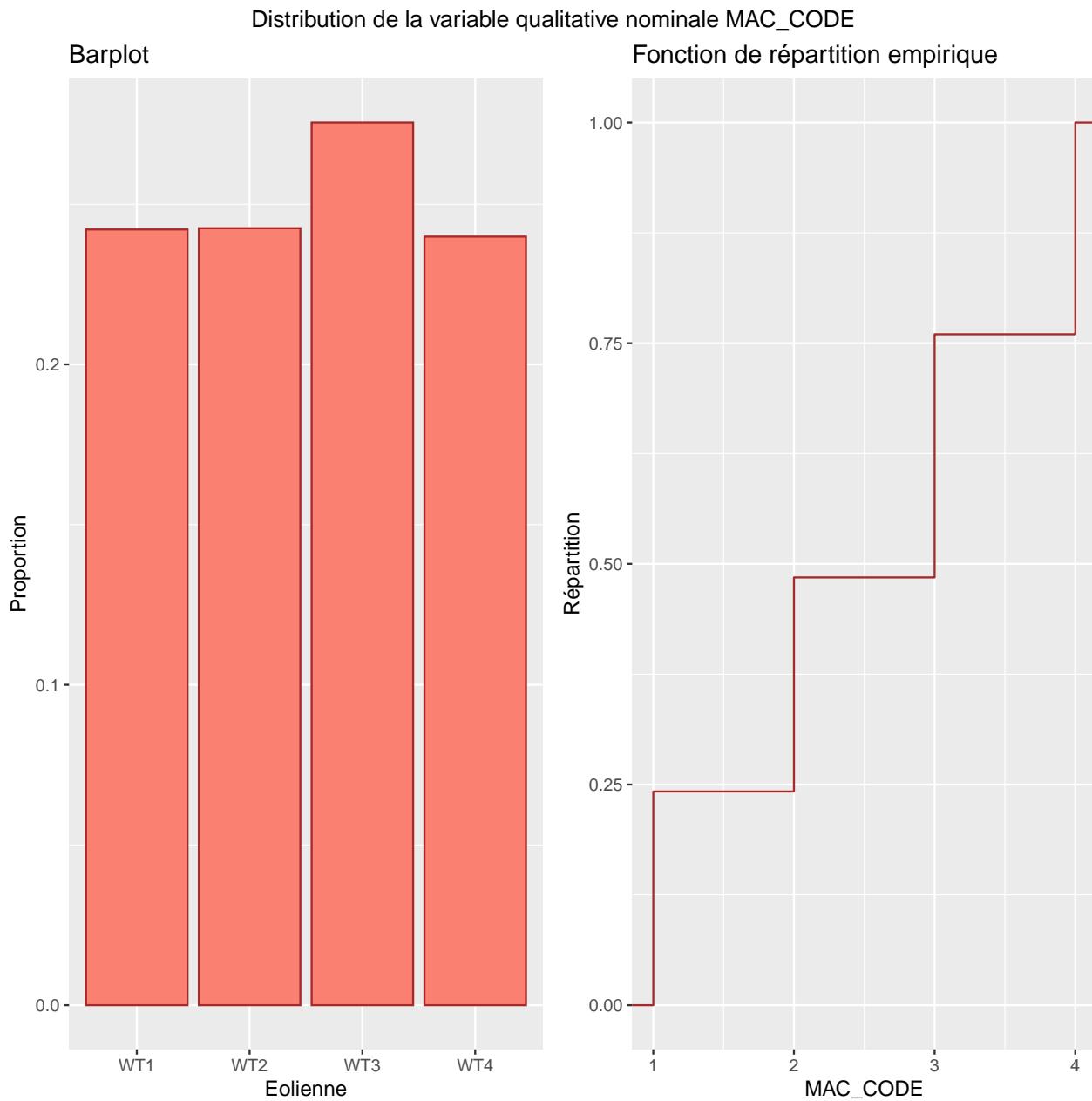


D'après le Normal Q–Q plot, un nombre de points important ne sont pas alignés sur la première bissectrice de plus ils dépassent les bornes $[-2, 2]$. On ne peut pas dire que tous ces points sont aberrants et se permettre de les supprimer.

Analyse descriptive

Analyse descriptive univariée

La variable MAC_CODE



D'après l'histogramme de la variable MAC_CODE qui désigne l'éolienne sur laquelle les données ont été mesurés, on constate que les proportions des quatres éolienne sont presque les mêmes avec une légère hausse pour l'éolienne WT3.