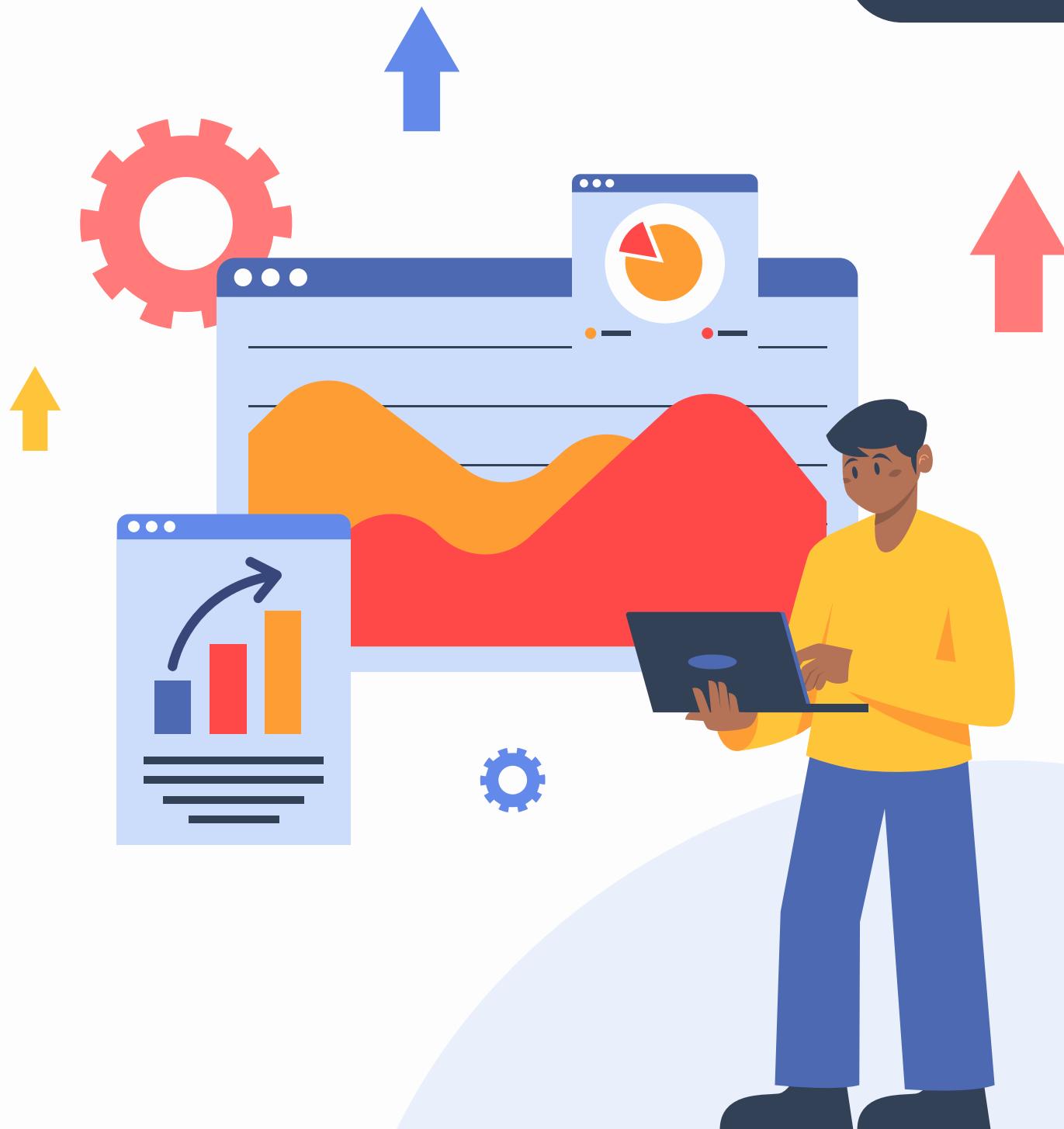


Anova

Regression Correlation analysis

Statistiques



Sommaire

Introduction

Chapitre 1 : Gas Turbine CO and NOx Emission :

Présentation du dataset

Analyse statistique exploratoire et tests

Exploration des données

Regression Linéaire

Préparation des données

Anova

Chapitre 2 : AI4I 2020 Predictive Maintenance :

Présentation du dataset

Analyse statistique exploratoire et tests

Exploration des données

Regression Linéaire

Préparation des données

Anova

Conclusion





Introduction



Introduction

L'analyse statistique constitue un outil indispensable pour comprendre et exploiter des informations complexes.

Ce projet s'inscrit dans cet objectif, en appliquant des techniques statistiques avancées à deux jeux de données distincts :

- 1. Gas Turbine CO and NOx Emission Dataset** : un ensemble de données utilisé pour prédire les émissions de gaz d'une turbine à gaz en fonction de variables environnementales.
- 2. AI4I 2020 Predictive Maintenance Dataset** : qui simule des scénarios industriels de défaillances de machines pour améliorer les stratégies de maintenance prédictive.

Cette présentation détaillera les étapes d'analyse pour chaque jeu de données, discutera des résultats obtenus et proposera des perspectives basées sur les limites rencontrées.



Chapitre 1 :

Gas Turbine CO and
NOx Emission



Présentation du dataset



Analyse de Structure :

Summary :

Le dataset **gt_combined** contient des informations sur diverses variables liées aux turbines à gaz :

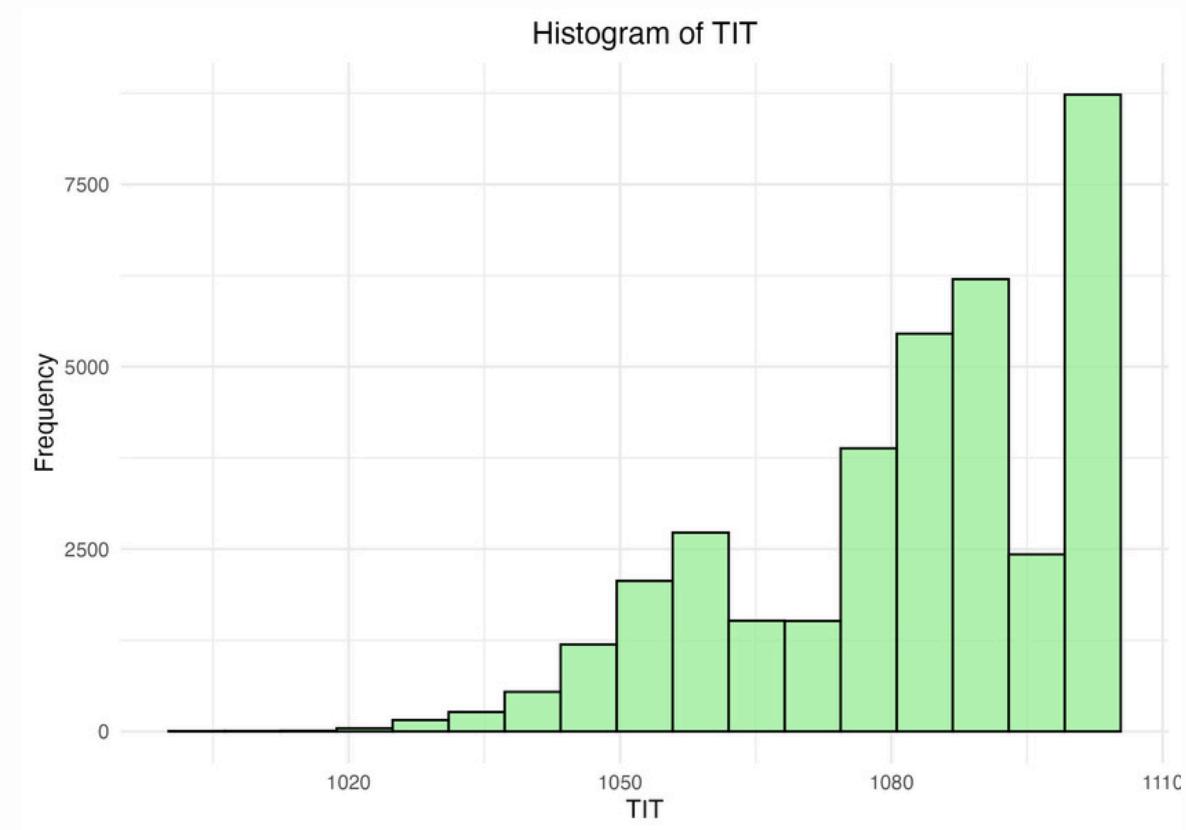
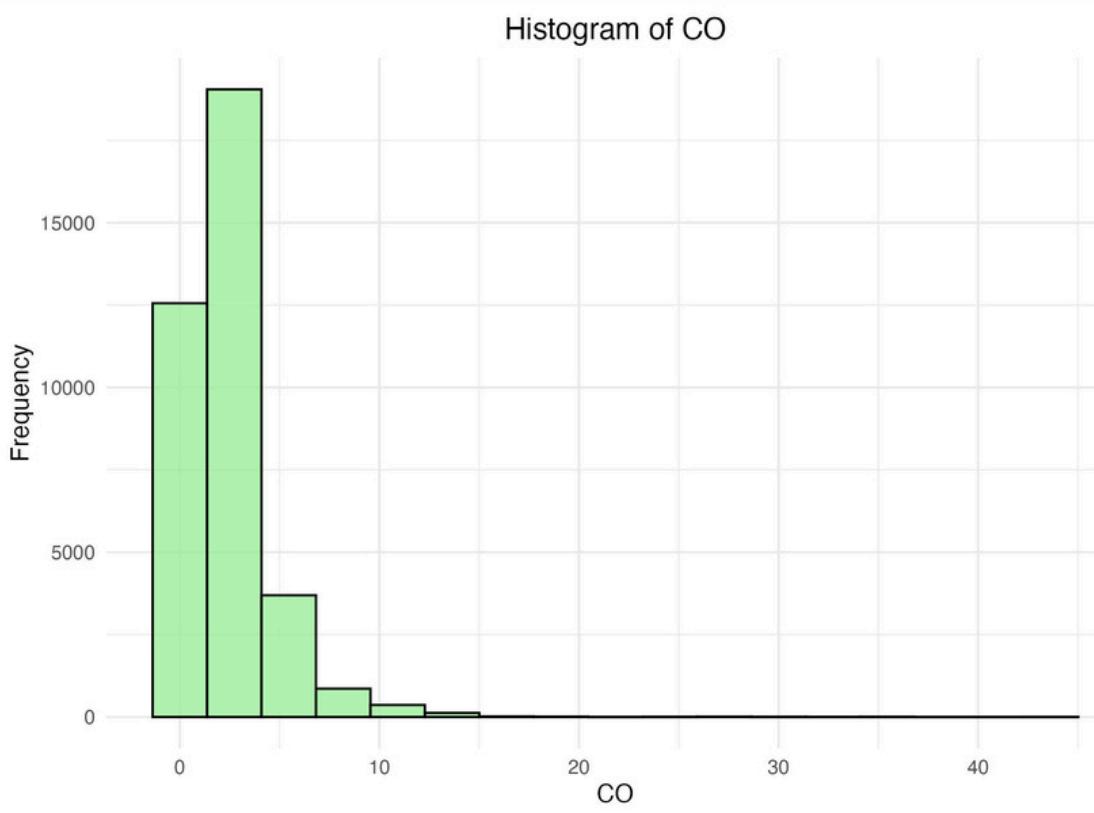
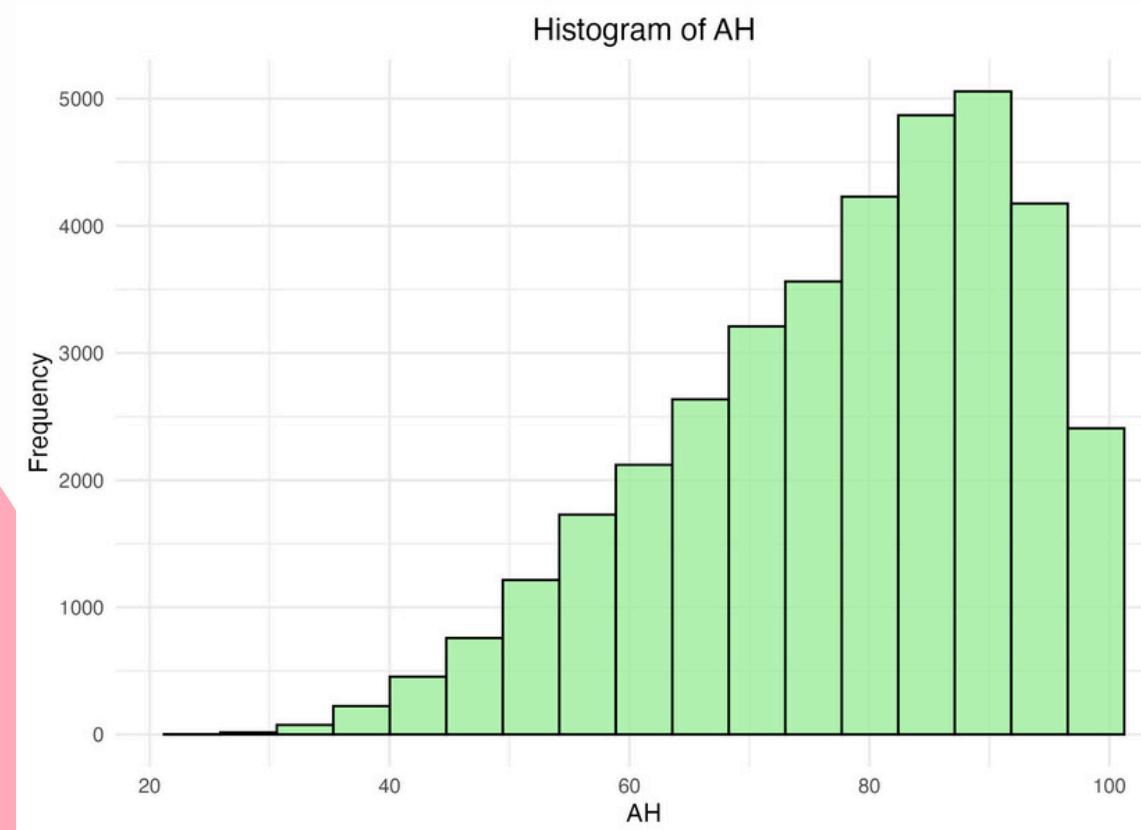
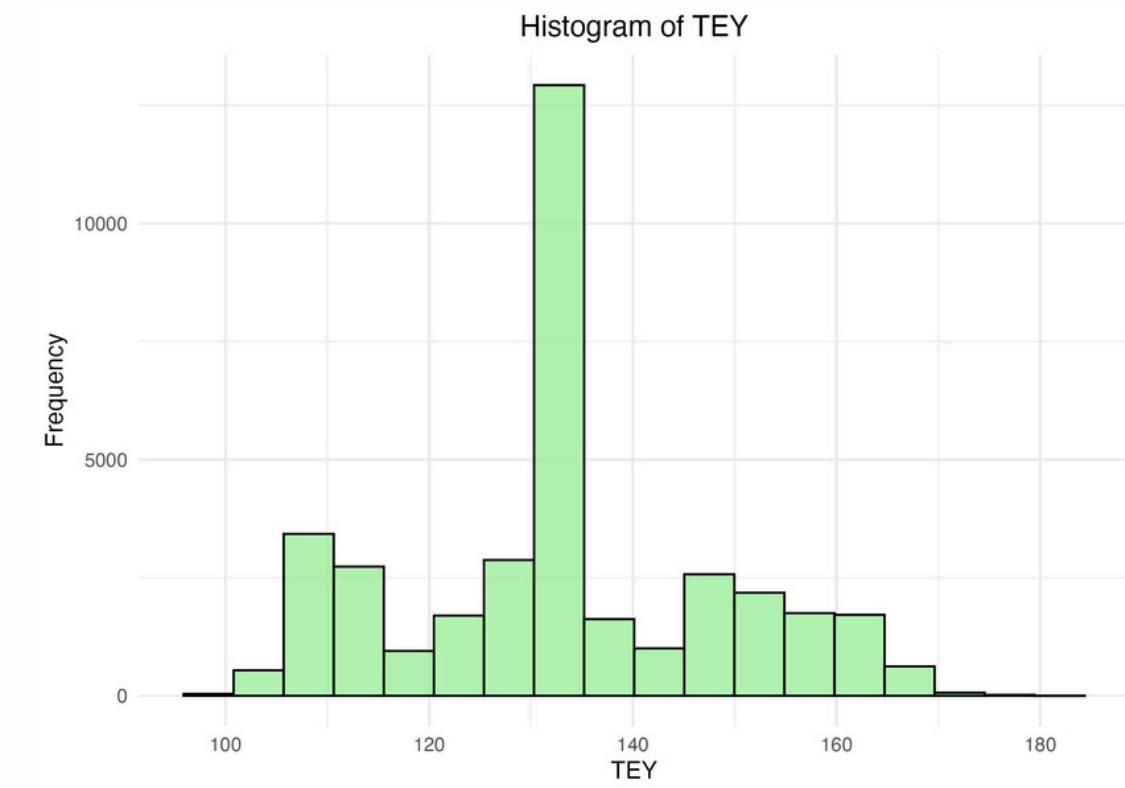
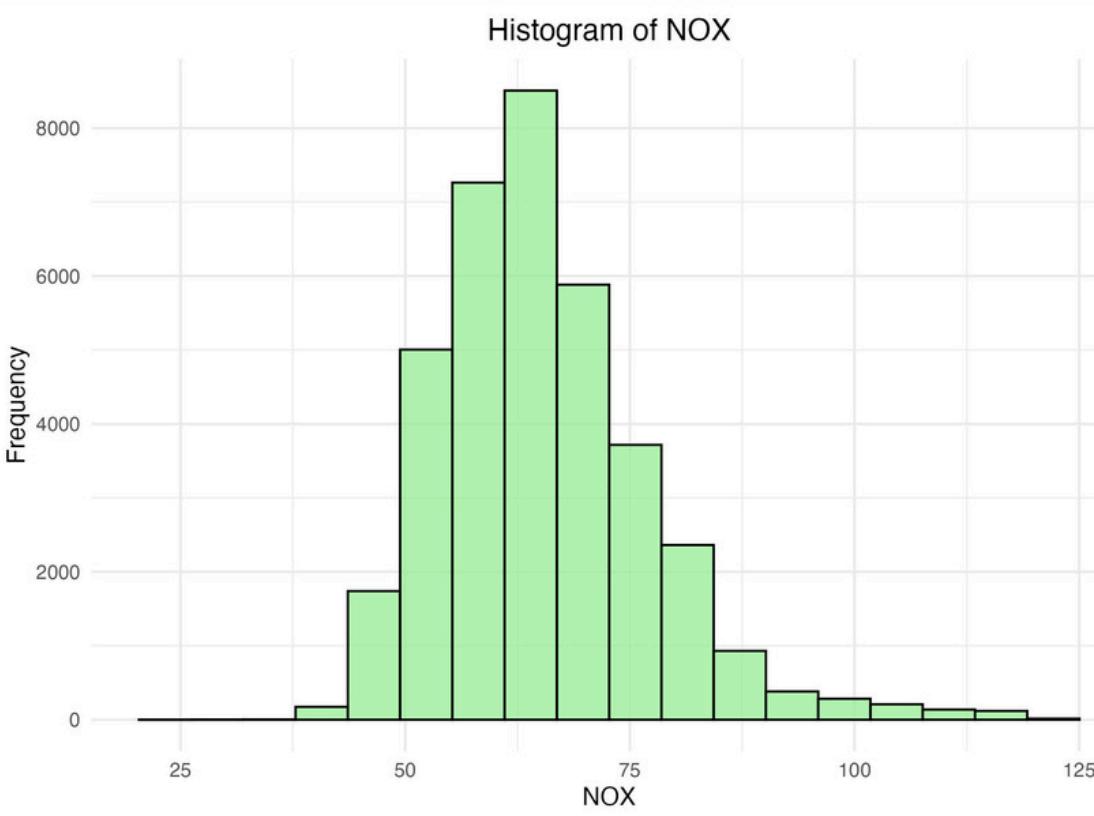
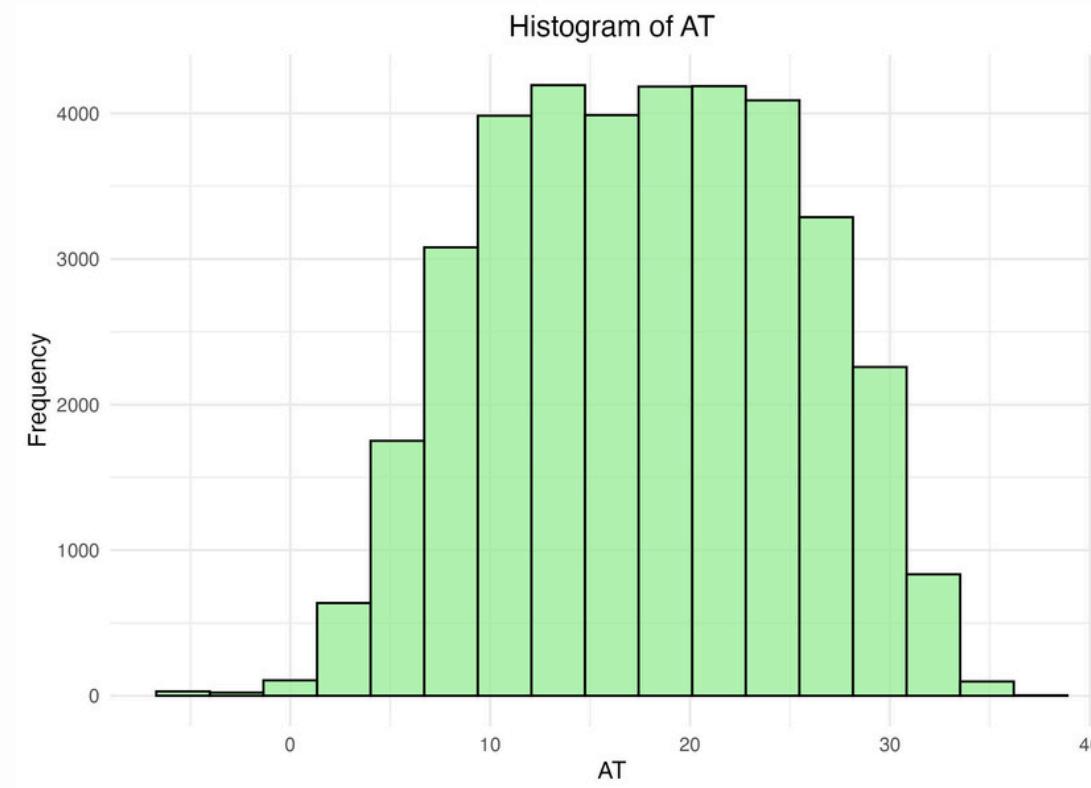
| AT | AP | AH | AFDP | GTEP | TIT | TAT |
|-----------------|-----------------|------------------|----------------|---|-----------------------------------|----------------|
| Min. : -6.235 | Min. : 985.9 | Min. : 24.09 | Min. : 2.087 | Min. : 17.70 | Min. : 1001 | Min. : 511.0 |
| 1st Qu.: 11.781 | 1st Qu.: 1008.8 | 1st Qu.: 68.19 | 1st Qu.: 3.356 | 1st Qu.: 23.13 | 1st Qu.: 1072 | 1st Qu.: 544.7 |
| Median : 17.801 | Median : 1012.6 | Median : 80.47 | Median : 3.938 | Median : 25.10 | Median : 1086 | Median : 549.9 |
| Mean : 17.713 | Mean : 1013.1 | Mean : 77.87 | Mean : 3.926 | Mean : 25.56 | Mean : 1081 | Mean : 546.2 |
| 3rd Qu.: 23.665 | 3rd Qu.: 1017.0 | 3rd Qu.: 89.38 | 3rd Qu.: 4.377 | 3rd Qu.: 29.06 | 3rd Qu.: 1097 | 3rd Qu.: 550.0 |
| Max. : 37.103 | Max. : 1036.6 | Max. : 100.20 | Max. : 7.611 | Max. : 40.72 | Max. : 1101 | Max. : 550.6 |
| TEY | CDP | CO | NOX | 'data.frame': 36733 obs. of 11 variables: | | |
| Min. : 100.0 | Min. : 9.852 | Min. : 0.00039 | Min. : 25.91 | \$ AT : num | 4.59 4.29 3.9 3.74 3.75 ... | |
| 1st Qu.: 124.5 | 1st Qu.: 11.435 | 1st Qu.: 1.18240 | 1st Qu.: 57.16 | \$ AP : num | 1019 1018 1018 1018 1018 ... | |
| Median : 133.7 | Median : 11.965 | Median : 1.71350 | Median : 63.85 | \$ AH : num | 83.7 84.2 84.9 85.4 85.2 ... | |
| Mean : 133.5 | Mean : 12.061 | Mean : 2.37247 | Mean : 65.29 | \$ AFDP: num | 3.58 3.57 3.58 3.58 3.58 ... | |
| 3rd Qu.: 144.1 | 3rd Qu.: 12.855 | 3rd Qu.: 2.84290 | 3rd Qu.: 71.55 | \$ GTEP: num | 24 24 24 23.9 23.9 ... | |
| Max. : 179.5 | Max. : 15.159 | Max. : 44.10300 | Max. : 119.91 | \$ TIT : num | 1086 1086 1086 1086 1086 ... | |
| | | | | \$ TAT : num | 550 550 550 550 550 ... | |
| | | | | \$ TEY : num | 135 135 135 135 135 ... | |
| | | | | \$ CDP : num | 11.9 11.9 12 12 11.9 ... | |
| | | | | \$ CO : num | 0.327 0.448 0.451 0.231 0.267 ... | |
| | | | | \$ NOX : num | 82 82.4 83.8 82.5 82 ... | |

Structure du Dataset :

Le dataset **gt_combined** contient **36,733** observations réparties sur **11 variables numériques**.
Voici les types de données observés :

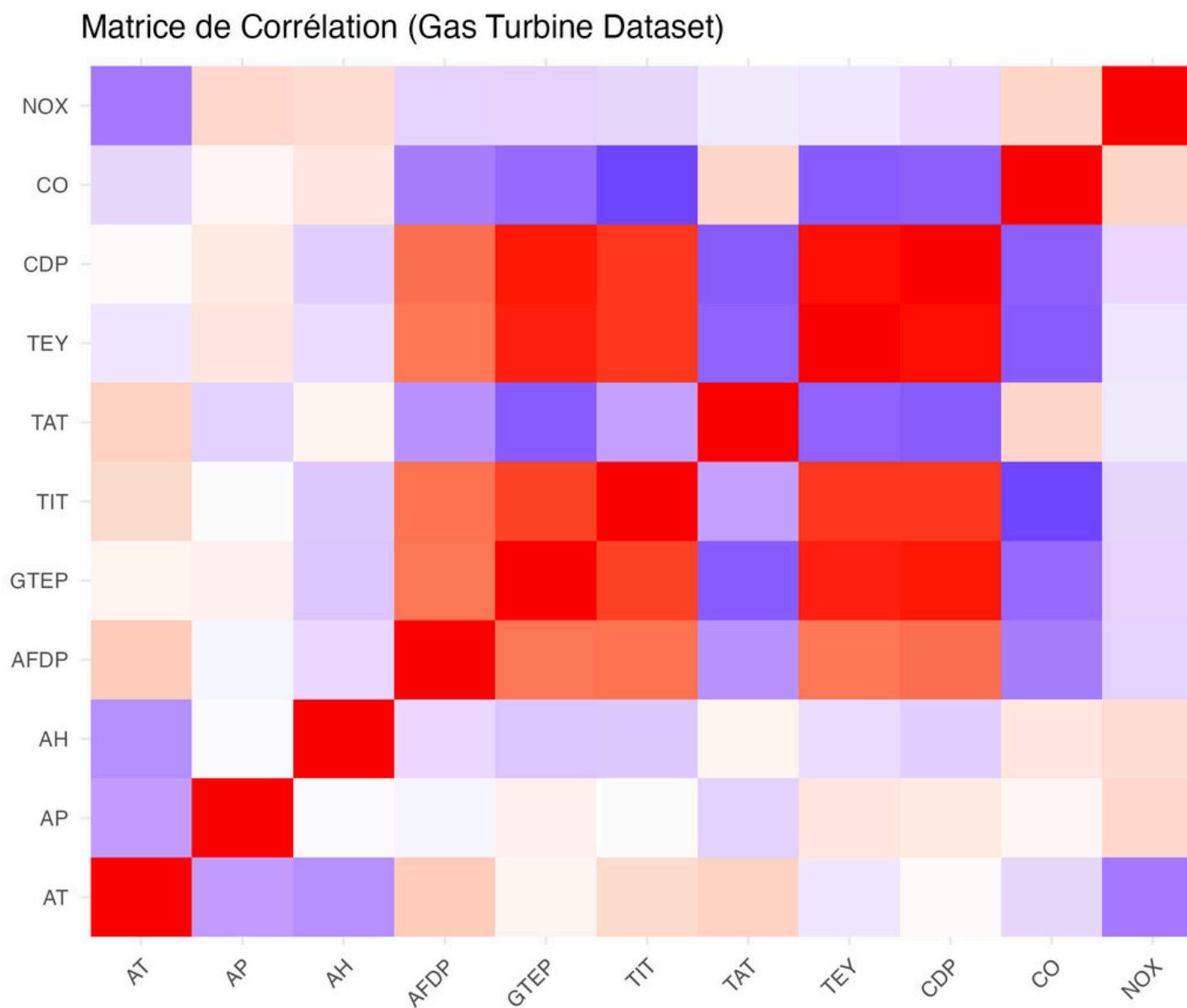
Exploration des données

Histogrames



Matrice de corrélation

La matrice de corrélation montre les relations linéaires entre les différentes variables numériques.
Voici quelques observations clés :



Température ambiante (AT) :

Corrélation positive forte avec TIT (température interne), suggérant que l'augmentation de la température ambiante entraîne une hausse de la température interne de la turbine.

Emissions (CO et NOX) :

CO et NOX montrent une forte corrélation positive (~0.9), ce qui signifie que les niveaux d'émissions de ces gaz sont fortement liés.

Torque (Torque..Nm.) :

Corrélation positive modérée avec Rotational.speed..rpm. (~0.6), indiquant une relation directe entre le couple et la vitesse de rotation.

Préparation des données



Vérification des valeurs manquantes:

| AT | AP | AH | AFDP | GTEP | TIT | TAT | TEY | CDP | CO | NOX |
|----|----|----|------|------|-----|-----|-----|-----|----|-----|
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

Pas de valeurs manquantes pour toutes les variables

Afficher les colonnes numériques :

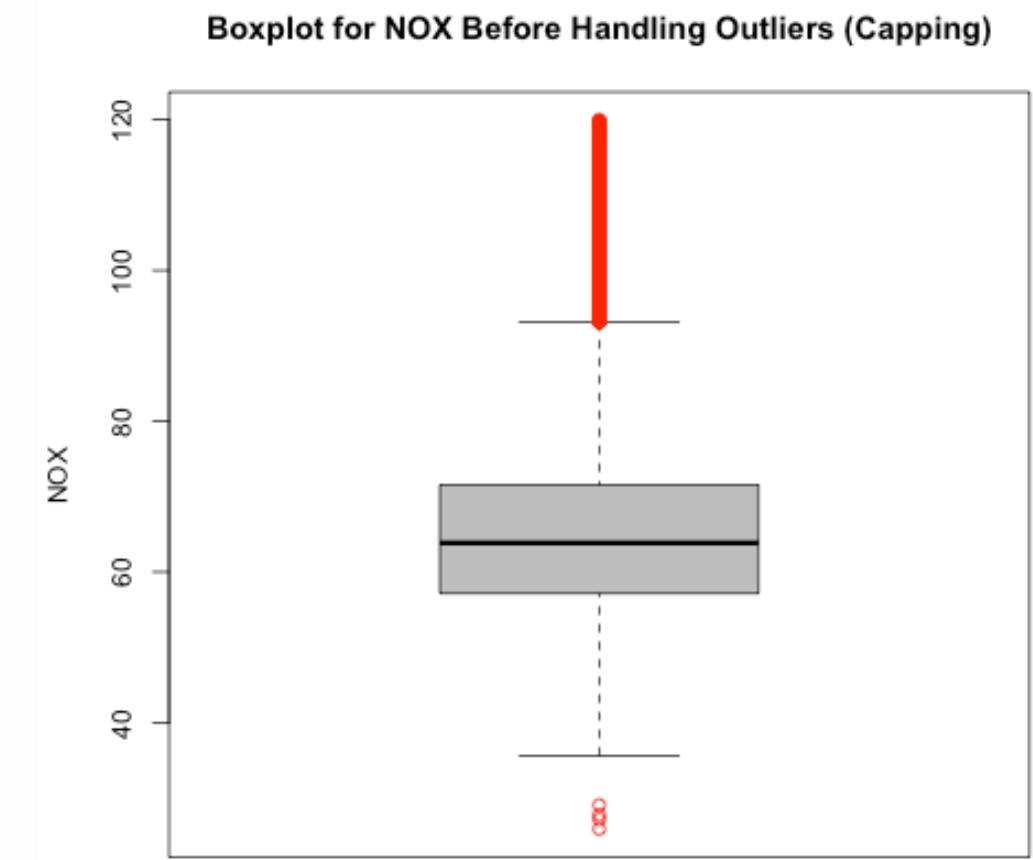
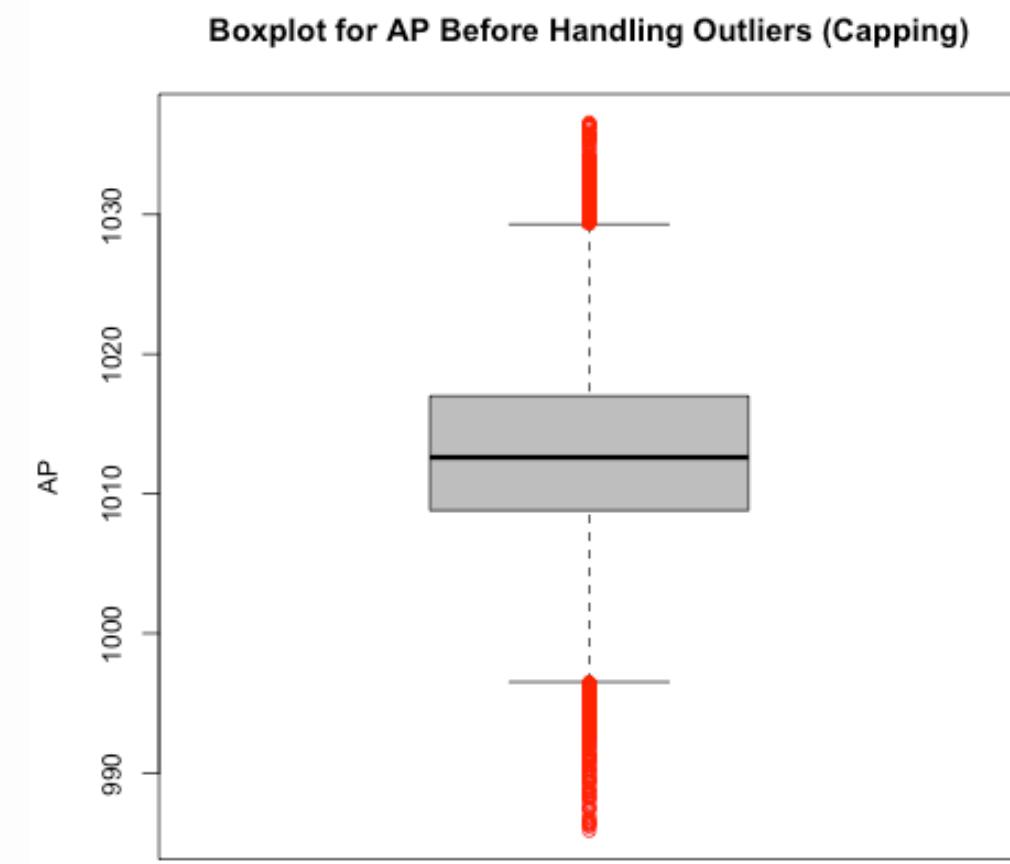
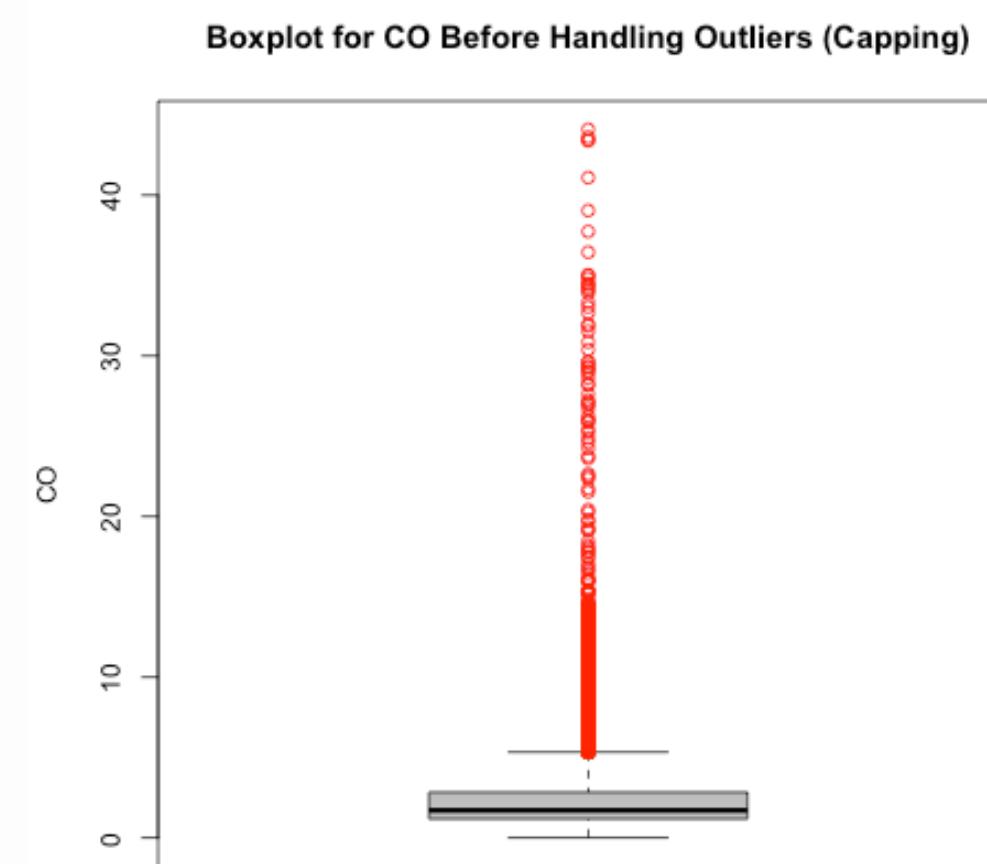
```
[1] "AT"    "AP"    "AH"    "AFDP"  "GTEP"  "TIT"   "TAT"   "TEY"   "CDP"   "CO"    "NOX"
```

Tous les colonnes sont des colonnes numeriques (pas de valeurs categoriques)



Traitement des valeurs aberrantes :

Boxplots Avant le Traitement des Valeurs Aberrantes :



Les boxplots montrent les distributions des différentes variables numériques avec leurs valeurs aberrantes représentées par des points rouges.
Ces points indiquent des observations très éloignées des distributions normales.



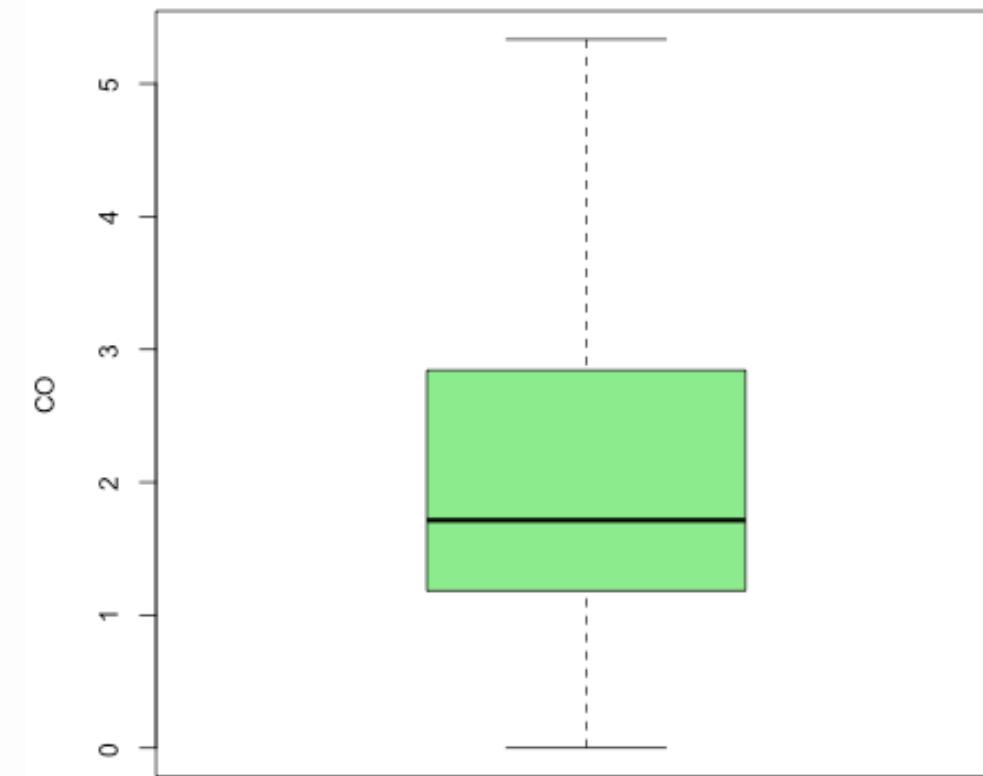
Traitement des valeurs aberrantes :

Boxplots Apres le Traitement des Valeurs Aberrantes :

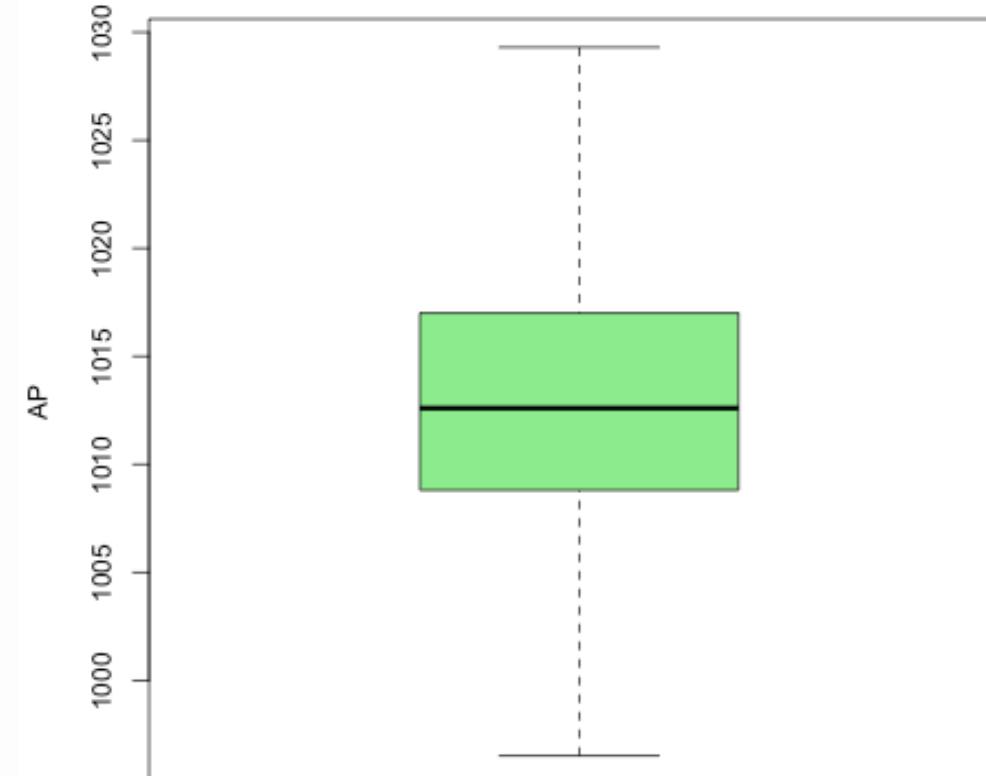
La fonction `cap_outliers(data, column)` a été utilisée pour limiter les valeurs extrêmes en ajustant les valeurs inférieures et supérieures basées sur l'IQR.

Voici comment cela impacte les distributions :

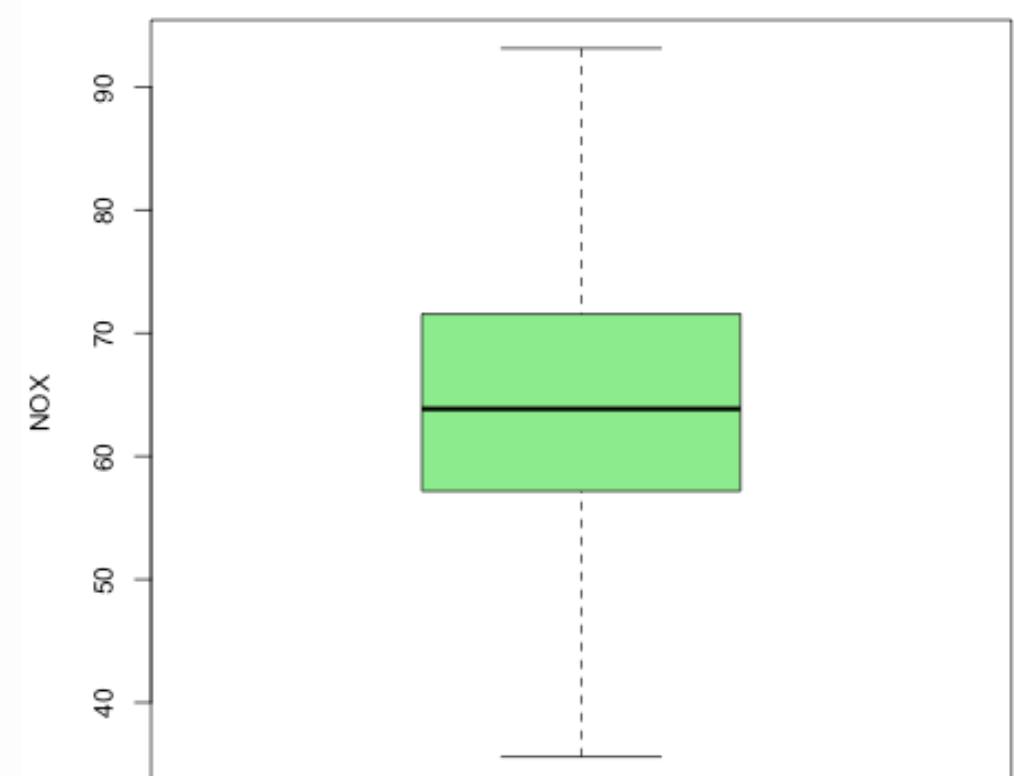
Boxplot for CO (Gas Turbine Dataset) After Handling Outliers (Cappi)



Boxplot for AP (Gas Turbine Dataset) After Handling Outliers (Cappi)



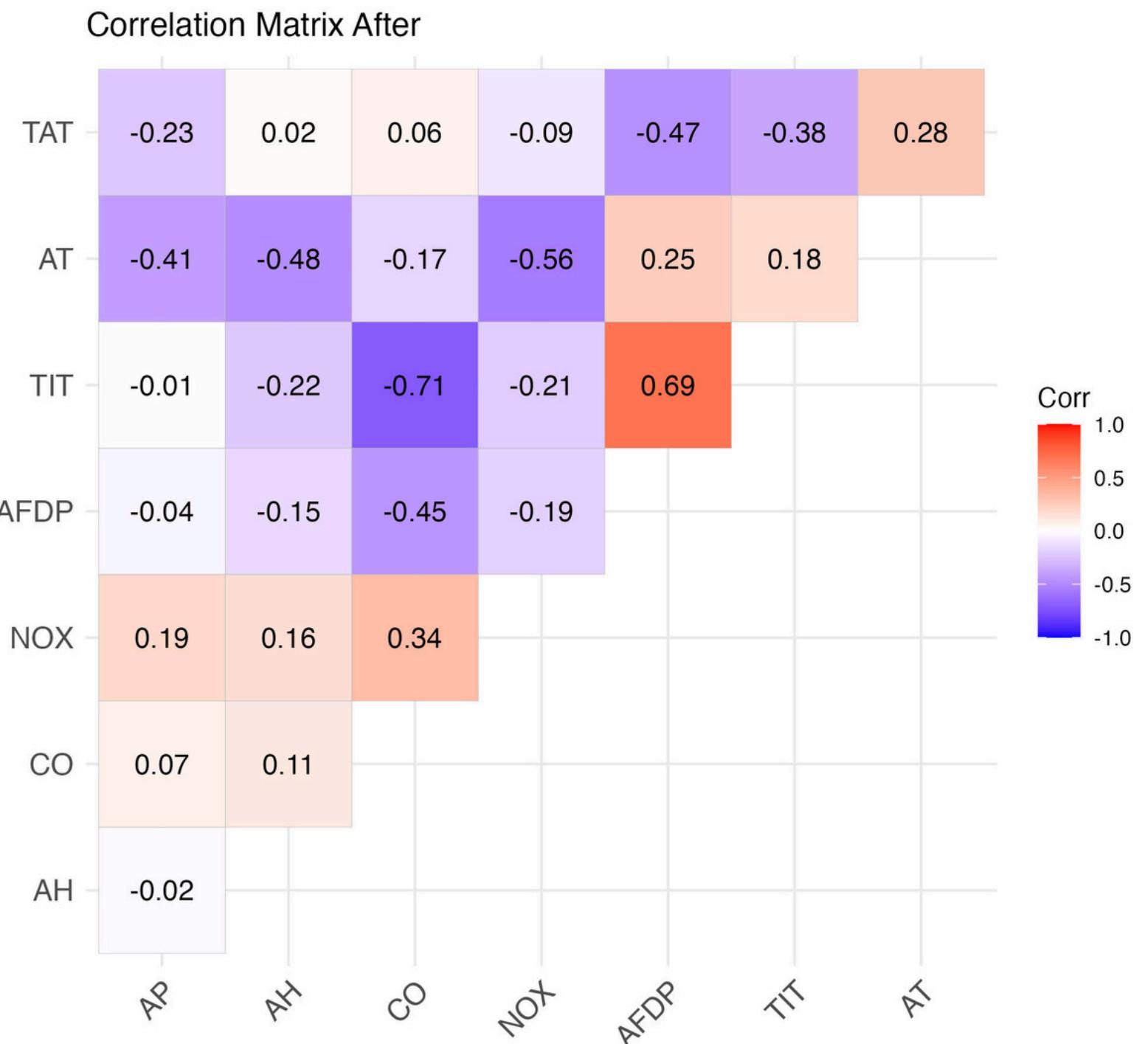
Boxplot for NOX (Gas Turbine Dataset) After Handling Outliers (Cappi)



- Les boxplots montrent des distributions où les valeurs extrêmes ont été ajustées ou capées, éliminant les valeurs qui influençait excessivement les analyses.
- Les plages de valeurs extrêmes sont limitées, rendant les distributions plus stables et plus représentatives.

Supprimer les colonnes fortement corrélées

analyser les corrélations entre les colonnes numériques d'un dataset et à supprimer les colonnes redondantes qui présentent une corrélation élevée avec d'autres (supérieure à 0.8)



Après avoir identifié les colonnes corrélées, elles sont supprimées itérativement jusqu'à ce que la matrice de corrélation réduite ne contienne que des relations faibles ou modérées.

Enfin, une heatmap est générée pour visualiser les corrélations restantes et sauvegarder les résultats pour une utilisation future.

Cette approche assure un dataset compact et optimisé pour des analyses ultérieures.

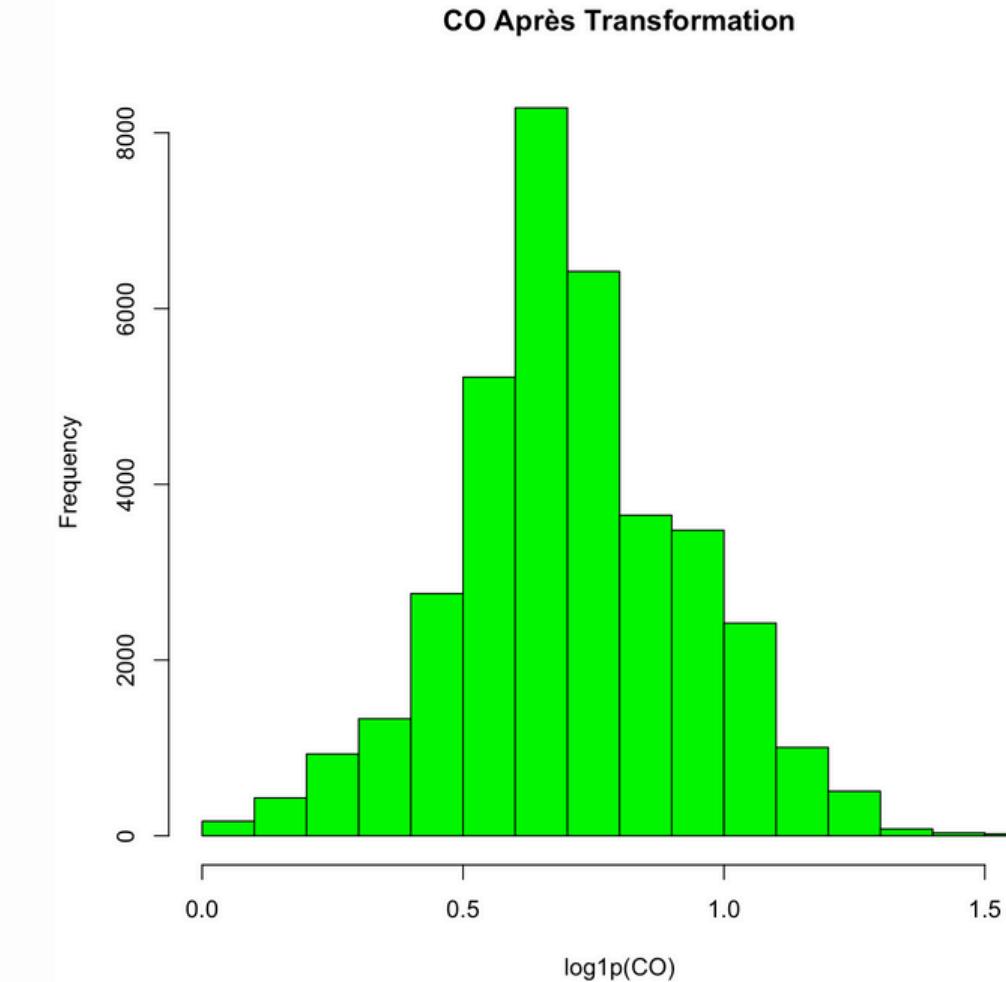
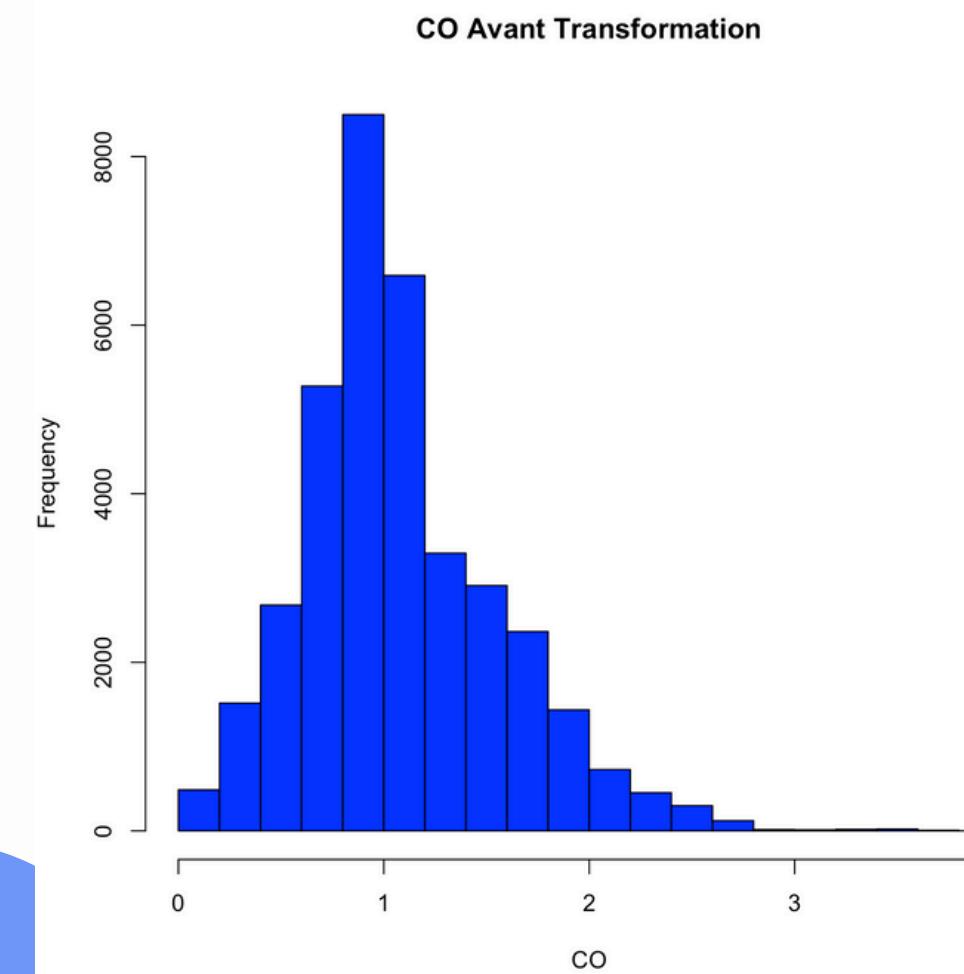
Transformation de la variable cible (CO):

Cette étape applique une transformation logarithmique à la variable cible CO afin de résoudre les problèmes de non-normalité (skewness) ou de réduction de l'impact des valeurs extrêmes dans les analyses futures.

Resultat attendu :

La distribution de CO devient moins asymétrique et plus proche de la normalité.

Les valeurs extrêmes (grandes ou petites) ont un impact réduit sur les analyses ultérieures.



Test de normalité avec Kolmogorov-Smirnov :

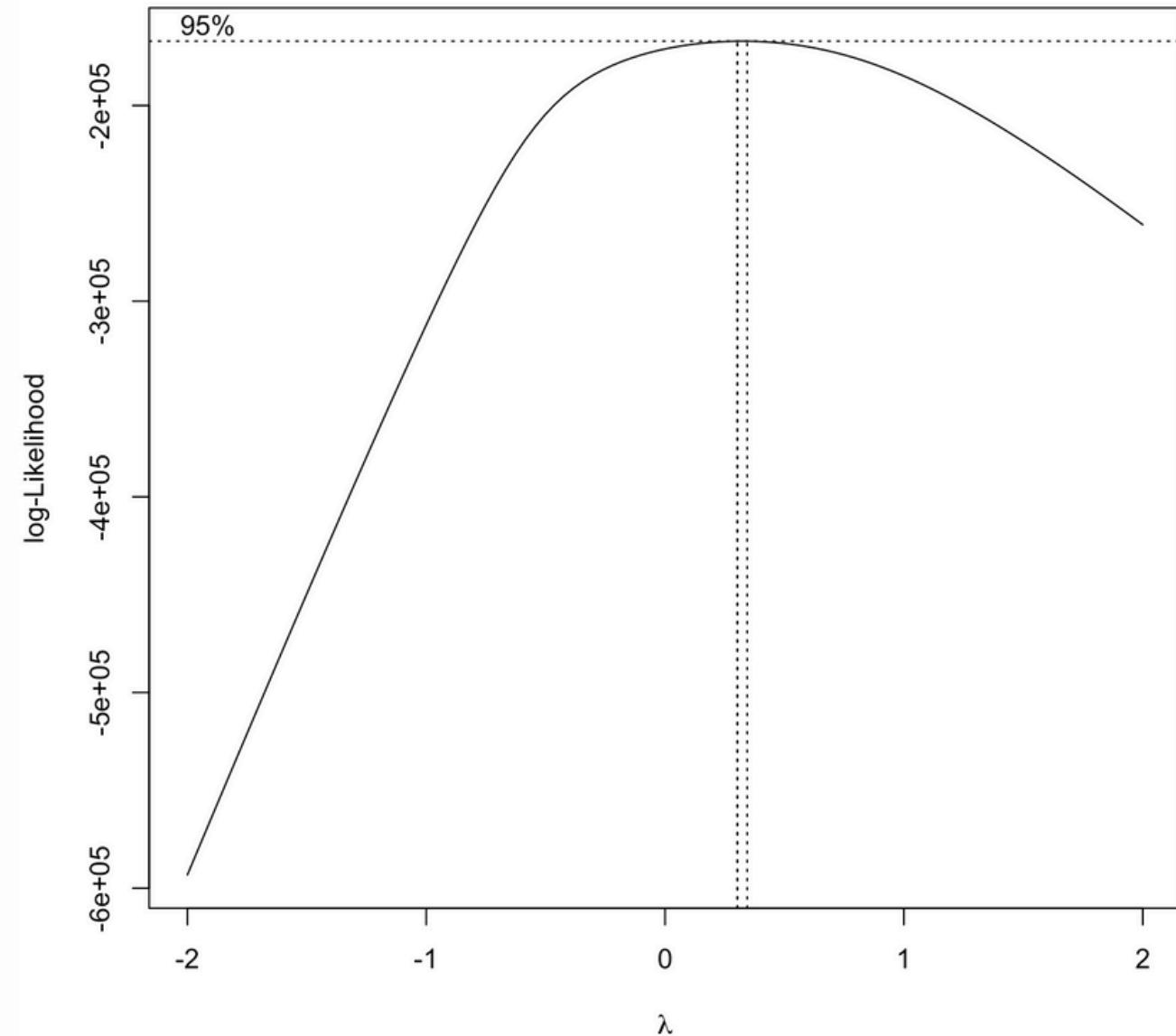
Le Kolmogorov-Smirnov test est un test non-paramétrique largement utilisé pour évaluer la normalité d'une distribution ou pour tester l'ajustement d'une distribution théorique à des données empiriques.

Réultat :

- D= 0.19149 : Bien que ce soit une distance relativement basse, le test montre toujours **une divergence importante par rapport à une distribution normale**.
- **p-value** : Moins de 0.05 signifie que nous **rejetons** fortement l'hypothèse nulle selon laquelle CO suit une distribution normale.
=> La distribution est donc **non normale**.

Asymptotic one-sample Kolmogorov-Smirnov test

```
data: gt_combined$CO  
D = 0.19149, p-value < 2.2e-16  
alternative hypothesis: two-sided
```



Ajout de termes quadratiques :

Permet de capturer des relations plus complexes entre les variables, améliorant ainsi la précision et la robustesse des modèles statistiques.

Division des données :

Permet de diviser le dataset **gt_combined** en deux ensembles : entraînement (80%) et test (20%).

Standardisation :

La standardisation des données consiste à ajuster les valeurs afin qu'elles aient une moyenne de 0 et un écart type 1.

Cette transformation améliore la convergence et la précision des modèles statistiques.

| | AT |
|---------|------------|
| Min. | :-3.208181 |
| 1st Qu. | :-0.795242 |
| Median | : 0.009035 |
| Mean | : 0.000000 |
| 3rd Qu. | : 0.798340 |
| Max. | : 2.598425 |

Analyse statistique exploratoire et tests

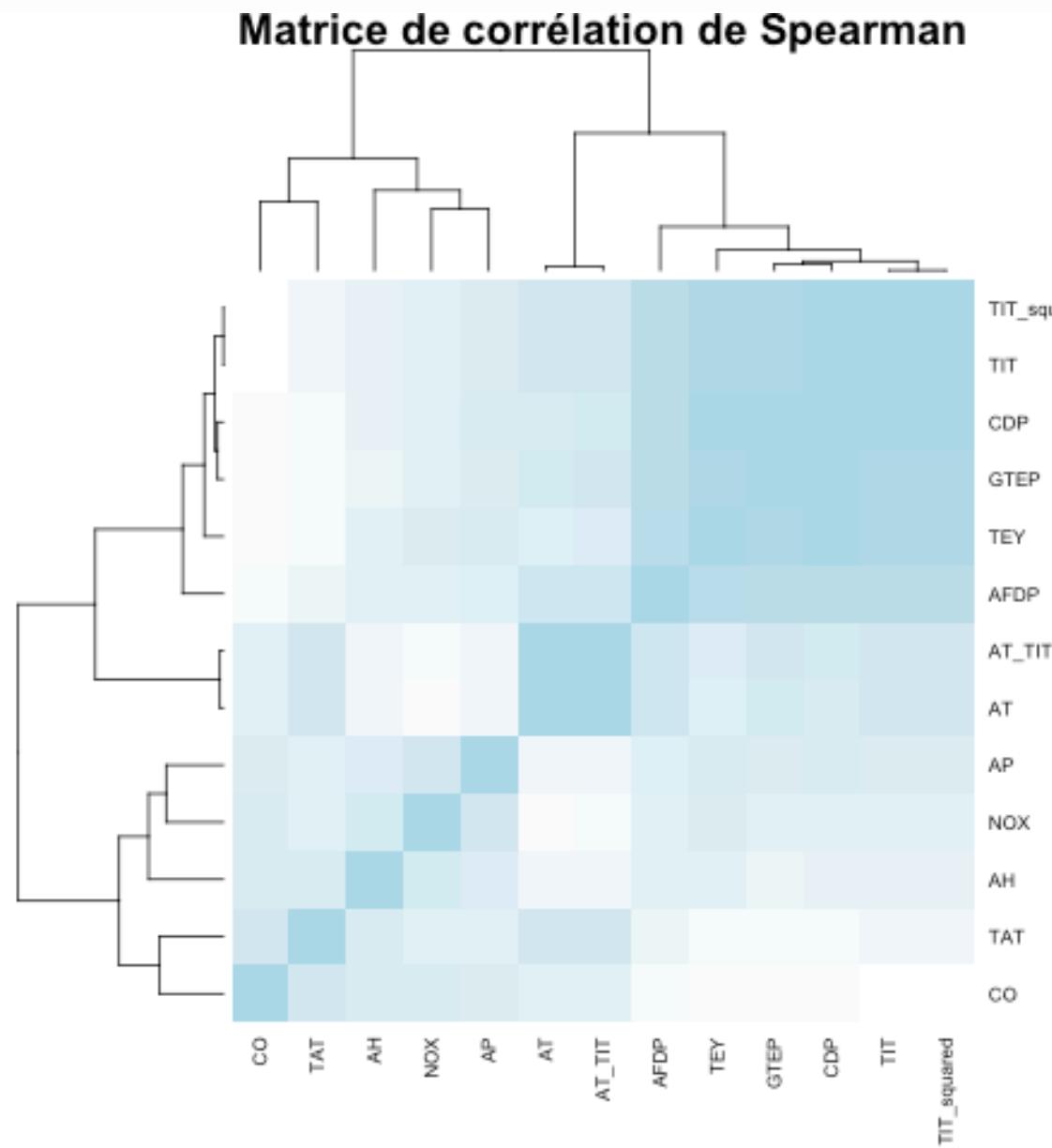


Analyse de corrélation:

Corrélation de Pearson:

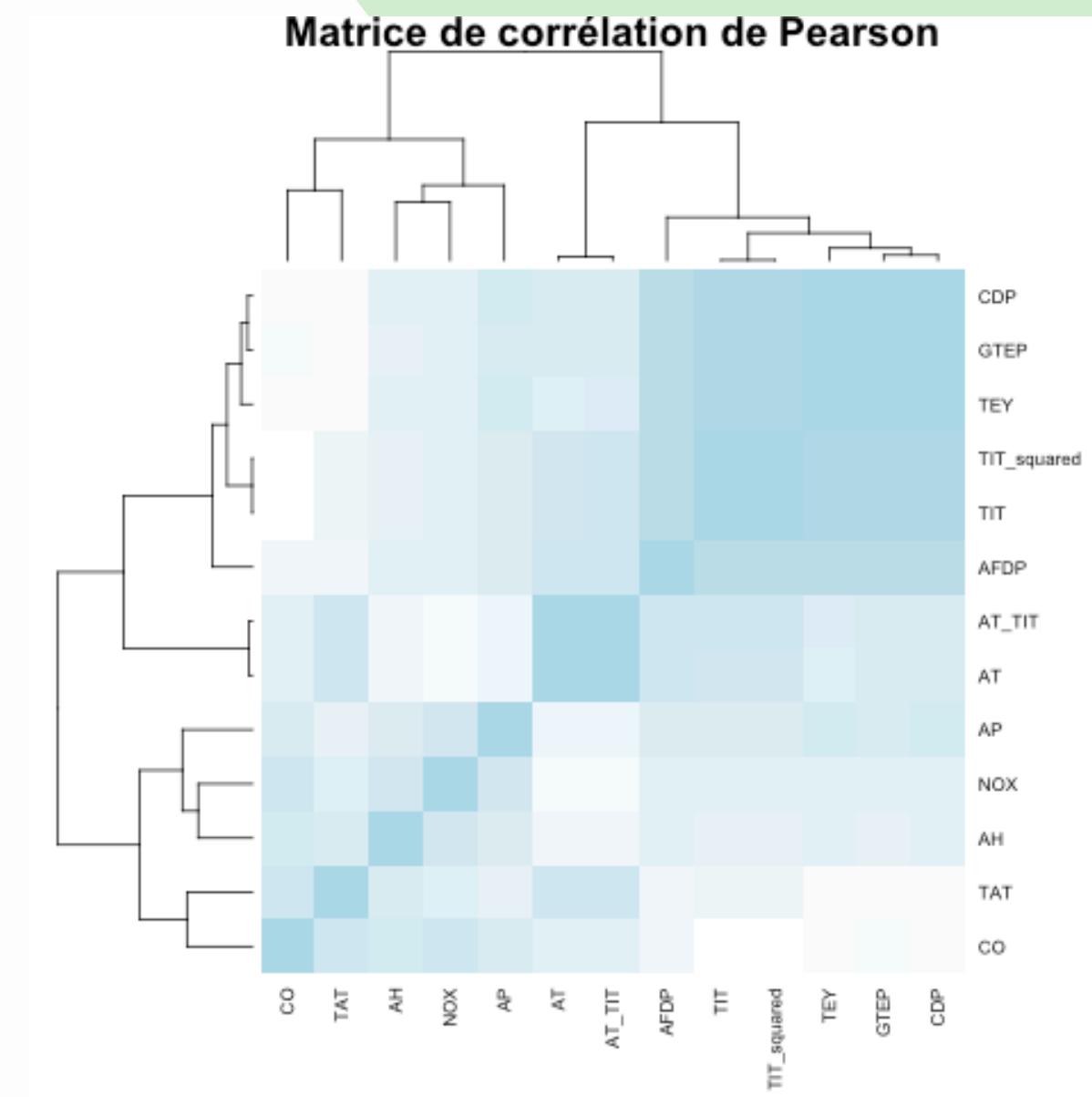
- AT et TIT montrent une corrélation extrêmement forte (0.999), indiquant une relation presque linéaire parfaite.
- CO et NOX ont une corrélation significative de 0.26, ce qui suggère une relation positive modérée.

Corrélation de Spearman :



Contrairement à **Pearson**, qui mesure la corrélation **linéaire**,
Spearman peut être utilisé lorsque les **relations ne sont pas strictement linéaires**, mais plutôt ordinaires ou monotones.

- AT et TIT montrent une forte corrélation positive (0.999), indiquant une relation très proche, même non-linéaire.
- CO et NOX ont une corrélation modérée (0.053), reflétant une relation moins directe ou linéaire.



Nuages de points pour CO par rapport aux autres variables :

Les nuages de points visualisent la relation entre la variable cible CO et chaque autre variable numérique du dataset. Voici les éléments clés à analyser :

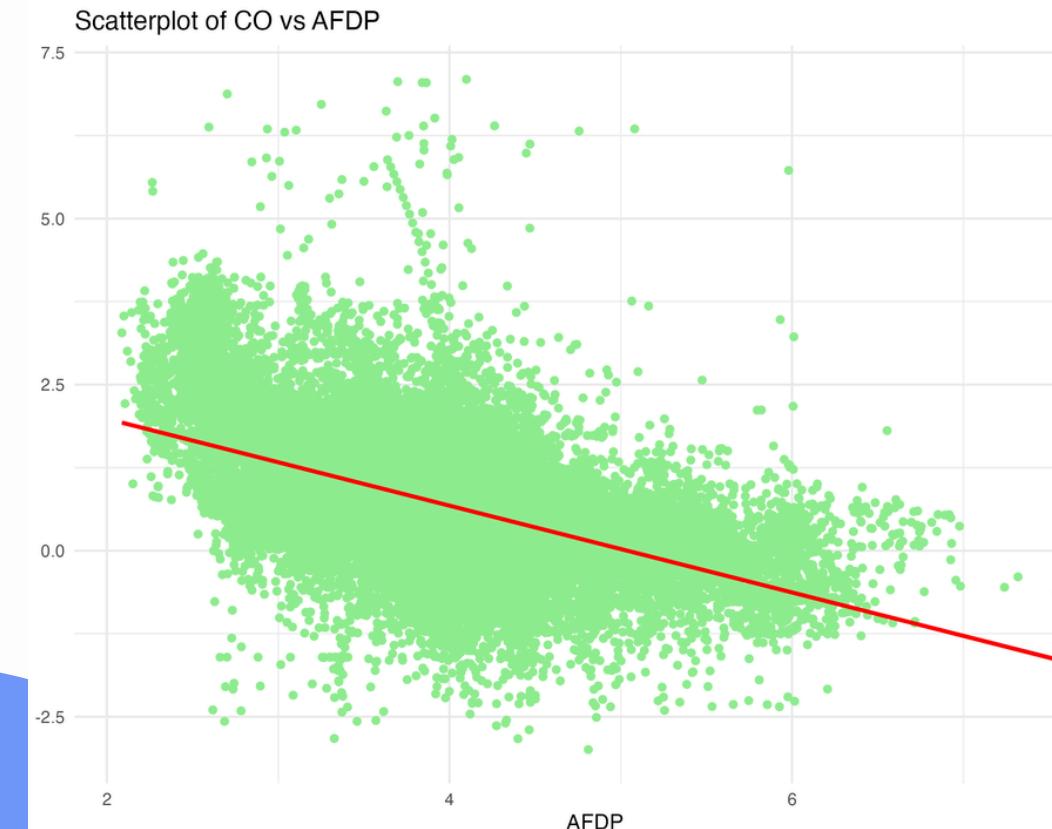
Relation linéaire ou non-linéaire:

- Les graphiques montrent la relation linéaire ou non-linéaire entre CO et une autre variable.
- Le lissage rouge aide à identifier une tendance globale.

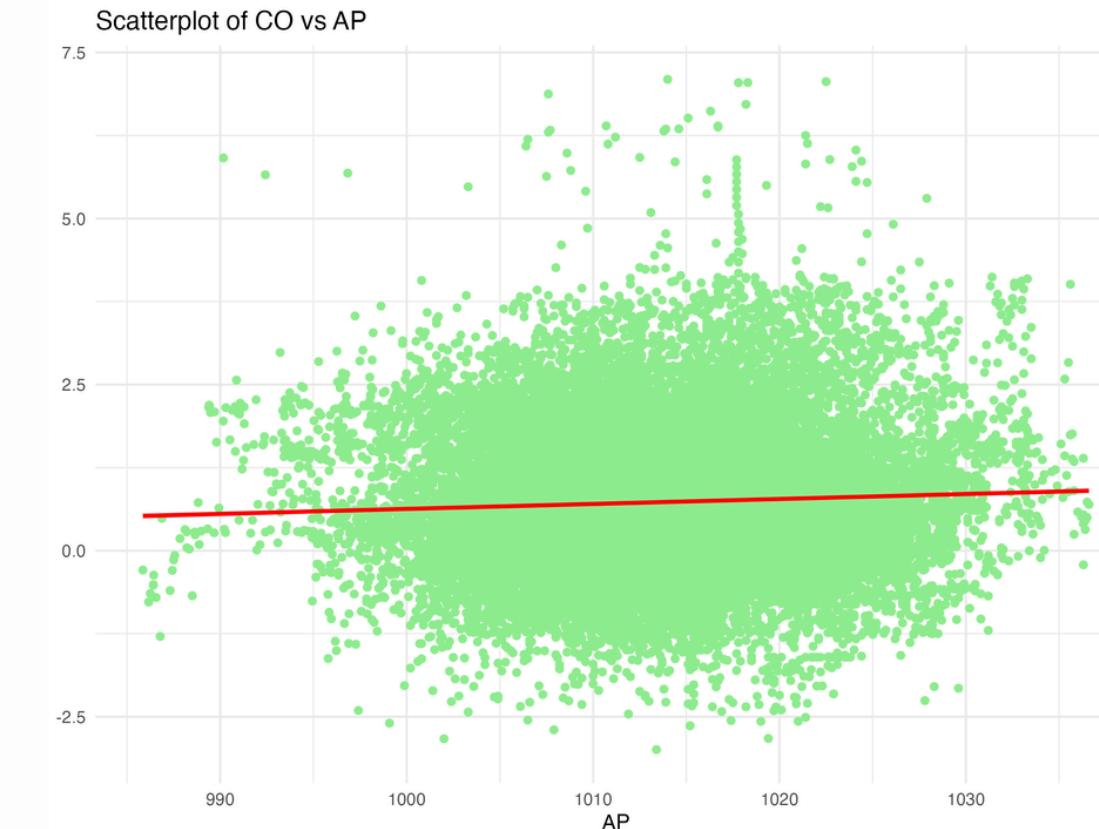
Analyse de la densité des points :

- La densité des points peut indiquer des zones où les relations sont plus fortes.

Variables avec une forte relation :



Variables avec une relation non claire :



Tests Statistiques

T-Tests :

Le test t de Welch est une variante du test t classique, utilisée pour comparer les moyennes de deux groupes indépendants lorsque les variances des groupes sont inégales.

Il teste l'hypothèse nulle selon laquelle les deux groupes ont la même moyenne.

Résultat :

```
Welch Two Sample t-test

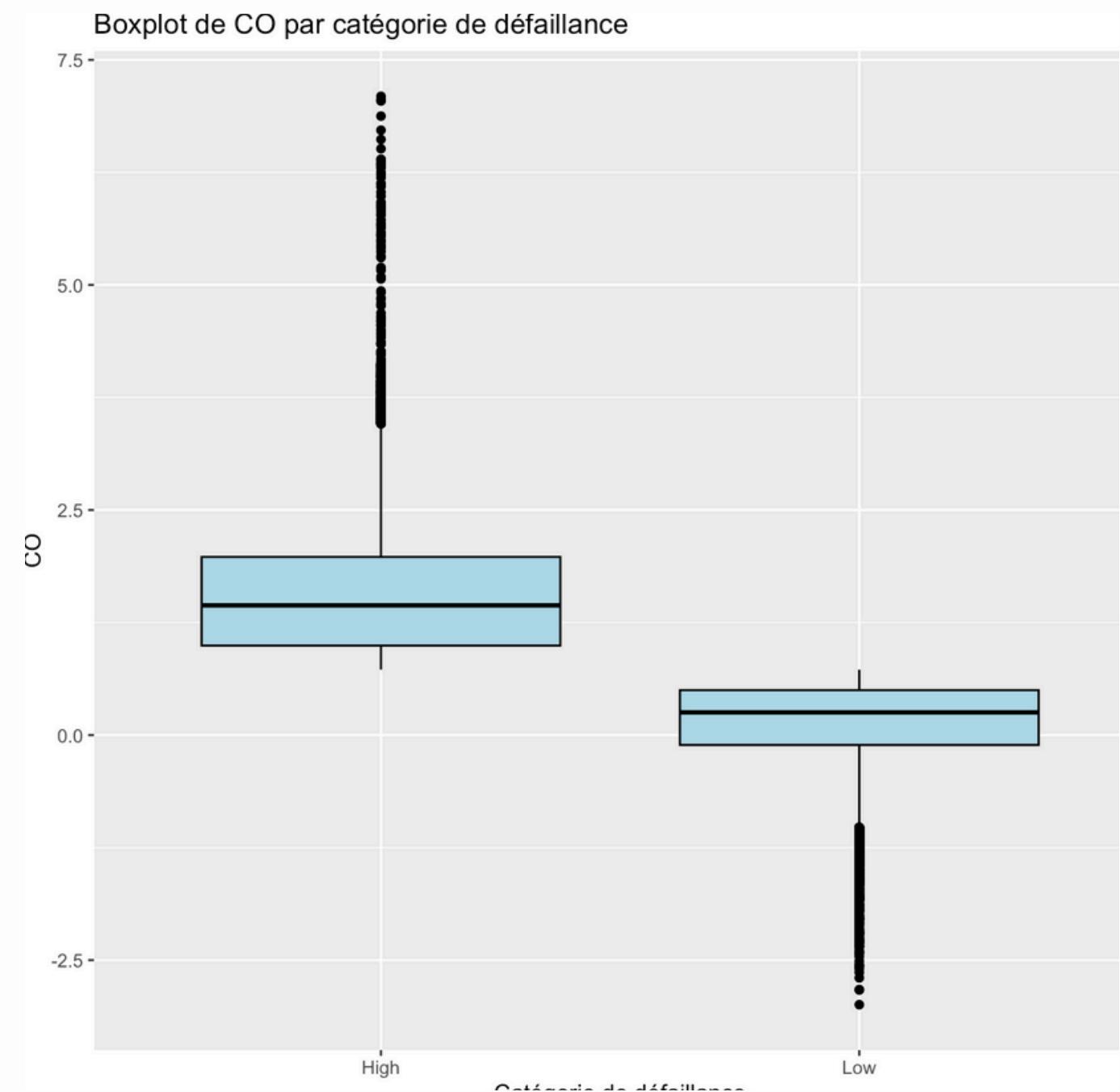
data: CO by FailureCategory
t = 208.41, df = 24236, p-value < 2.2e-16
alternative hypothesis: true difference in means between group High and group Low is not equal to 0
95 percent confidence interval:
1.454959 1.482586
sample estimates:
mean in group High mean in group Low
1.5917060   0.1229336
```

p-value < 2.2e-16 : Une p-value extrêmement faible signifie que **HO est rejetée**.

Cela indique que les **moyennes des deux groupes sont significativement**

Intervalle de confiance (95%)

L'intervalle de confiance montre que la vraie différence entre les moyennes des deux groupes se situe probablement entre ces deux valeurs.



chi Carré test :

Le test du Chi² est utilisé pour examiner si deux variables catégoriques sont indépendantes l'une de l'autre. Il teste l'hypothèse nulle selon laquelle il n'existe pas de relation statistiquement significative entre les deux variables.

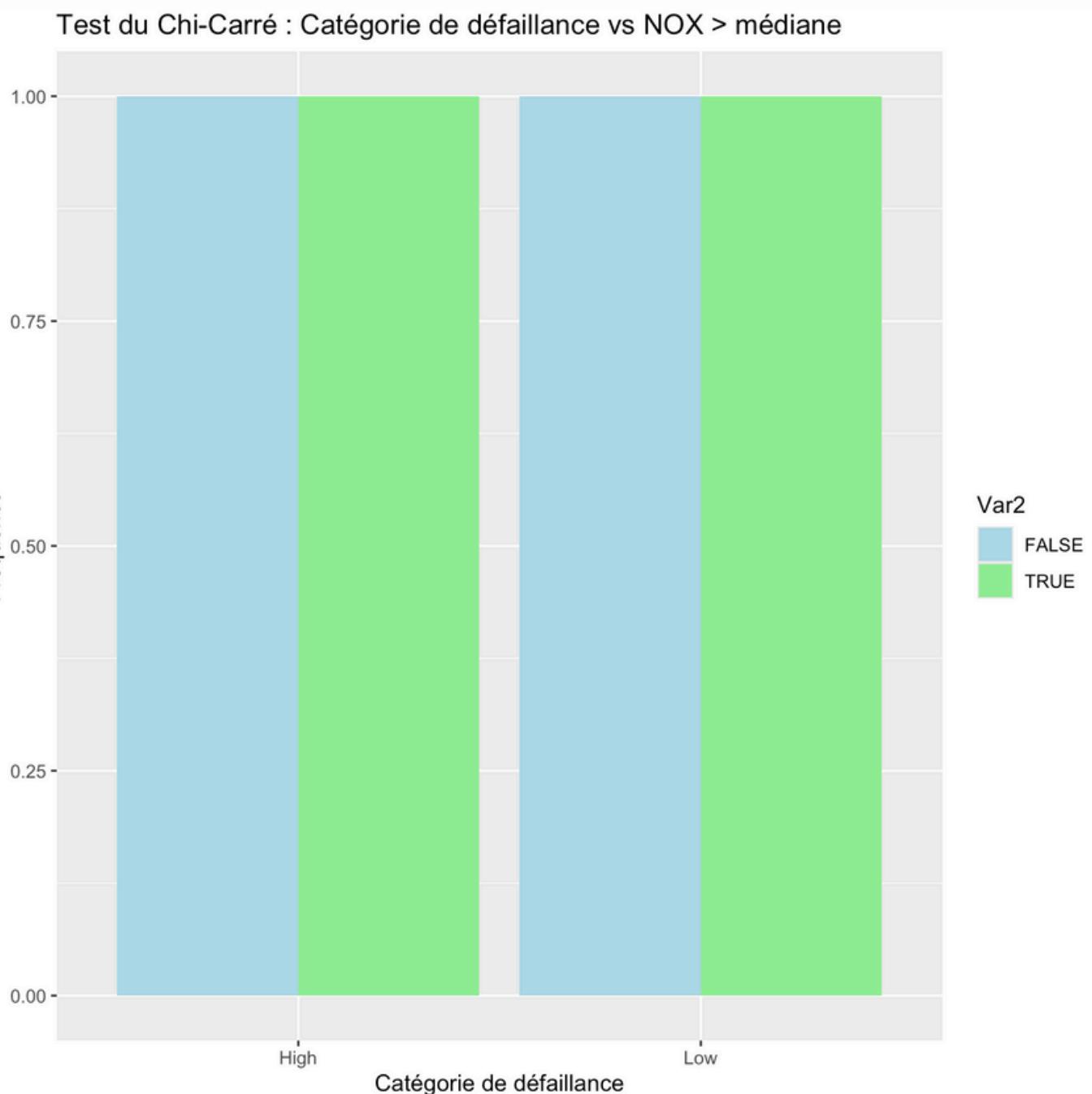
Resultat :

```
Pearson's Chi-squared test with Yates' continuity correction
```

```
data: contingency_table  
X-squared = 0.011369, df = 1, p-value = 0.9151
```

p-value = 0.9151: Une p-value extrêmement élevée signifie que H0 est **retenue**.

il n'y a pas de relation significative entre les deux variables catégoriques étudiées. Ces variables peuvent donc être traitées comme indépendantes dans le cadre de votre analyse, et leur relation n'est probablement pas pertinente pour des modèles prédictifs ou explicatifs.



Test de Kruskal-Wallis :

Le test de Kruskal-Wallis est une méthode non paramétrique utilisée pour comparer les distributions d'une variable continue entre plusieurs groupes.

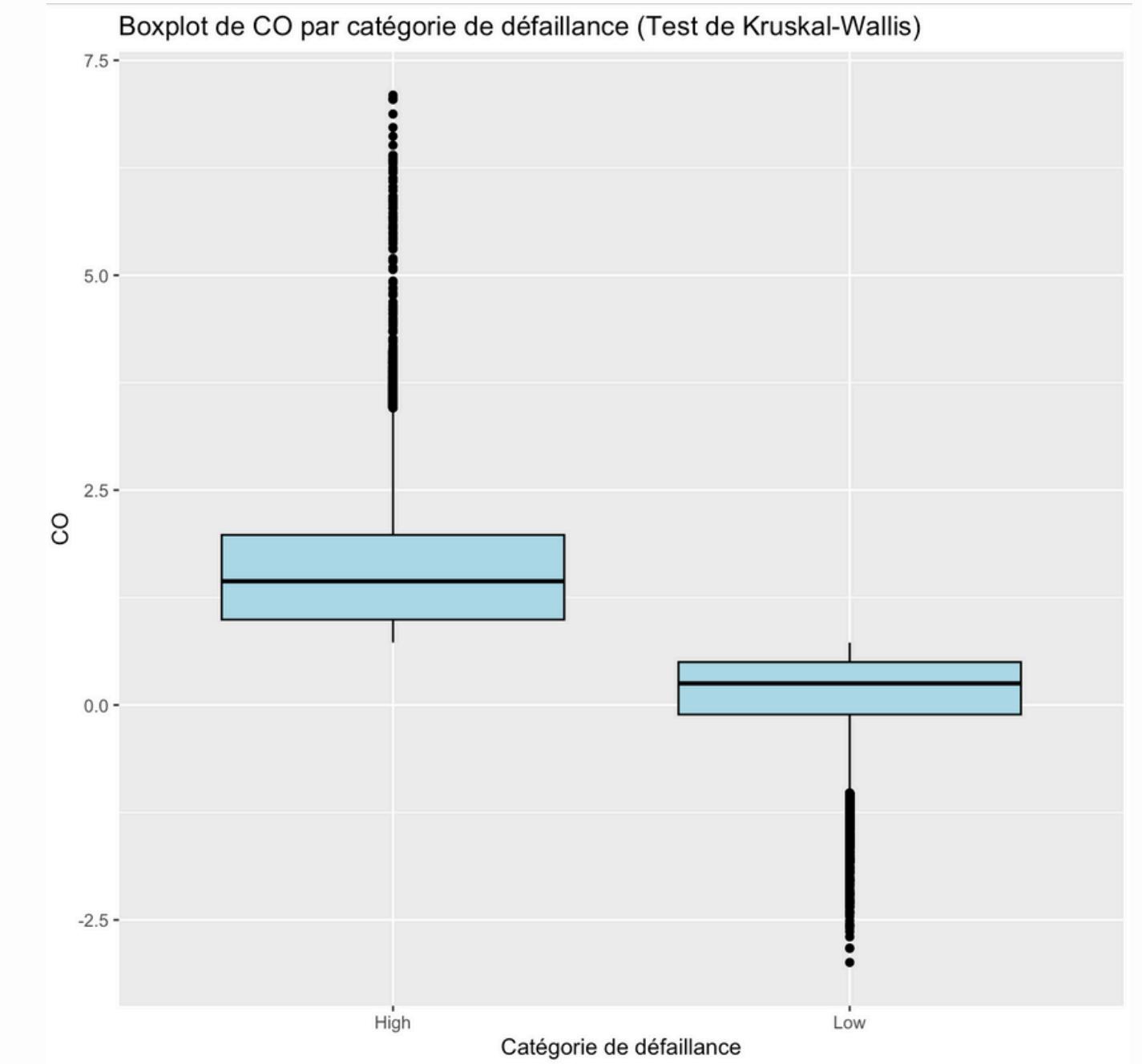
Réultat :

```
Kruskal-Wallis rank sum test  
  
data: CO by FailureCategory  
Kruskal-Wallis chi-squared = 26673, df = 1, p-value < 2.2e-16
```

p-value < 2.2e-16 : Une p-value extrêmement faible signifie que H_0 est **rejetée**. Cela indique que les **moyennes des deux groupes sont significativement**

Intervalle de confiance (95%)

L'intervalle de confiance montre que la vraie différence entre les moyennes des deux groupes se situe probablement entre ces deux valeurs.



Regression Linéaire

stepwise model :

La régression linéaire par étapes est une méthode itérative qui sélectionne les variables explicatives les plus pertinentes pour prédire la variable cible (CO). Elle combine deux approches :

- Ajout (Forward selection) : Ajoute progressivement les variables les plus significatives.
- Suppression (Backward elimination) : Supprime les variables non significatives.

Resultat :

Résidus : La médiane des résidus est proche de zéro, mais des valeurs extrêmes existent.

Coefficients : Tous les coefficients des variables explicatives sont significatifs avec des p-values < 0.05.

R-squared : 58.87% de la variance de CO est expliquée par le modèle.

F-statistic : Le modèle est significatif avec une p-value très faible (< 2.2e-16).

Call:

```
lm(formula = CO ~ GTEP + AT + AP + AH + AFDP + TAT + TEY + NOX,  
   data = train_data)
```

Residuals:

| Min | 1Q | Median | 3Q | Max |
|---------|---------|--------|--------|--------|
| -5.8544 | -0.3310 | 0.0375 | 0.3623 | 6.3192 |

Coefficients:

| | Estimate | Std. Error | t value | Pr(> t) |
|----------------|------------|------------|---------|--------------|
| (Intercept) | 8.245e-17 | 3.742e-03 | 0.000 | 1 |
| GTEP | 6.411e-01 | 1.889e-02 | 33.936 | < 2e-16 *** |
| AT | -1.626e-01 | 7.052e-03 | -23.064 | < 2e-16 *** |
| AP | 2.676e-02 | 4.321e-03 | 6.193 | 5.97e-10 *** |
| AH | -4.193e-02 | 4.806e-03 | -8.725 | < 2e-16 *** |
| AFDP | -9.916e-02 | 5.794e-03 | -17.114 | < 2e-16 *** |
| TAT | -2.998e-01 | 6.212e-03 | -48.267 | < 2e-16 *** |
| TEY | -1.431e+00 | 1.777e-02 | -80.528 | < 2e-16 *** |
| NOX | 8.774e-02 | 4.834e-03 | 18.151 | < 2e-16 *** |
| --- | | | | |
| Signif. codes: | 0 *** | 0.001 ** | 0.01 * | 0.05 . |
| | 1 | | | |

Residual standard error: 0.6414 on 29377 degrees of freedom
Multiple R-squared: 0.5887, Adjusted R-squared: 0.5886
F-statistic: 5256 on 8 and 29377 DF, p-value: < 2.2e-16

stepwise model :

La régression linéaire par étapes est une méthode itérative qui sélectionne les variables explicatives pour prédire la variable cible (CO). Elle combine deux approches :

- Ajout (Forward selection) : Ajoute progressivement les variables les plus significatives.
- Suppression (Backward elimination) : Supprime les variables non significatives.

Résultat :

```
Residual standard error: 0.6414 on 29377 degrees of freedom
Multiple R-squared:  0.5887,   Adjusted R-squared:  0.5886
F-statistic:  5256 on 8 and 29377 DF,  p-value: < 2.2e-16
```

Residuals:

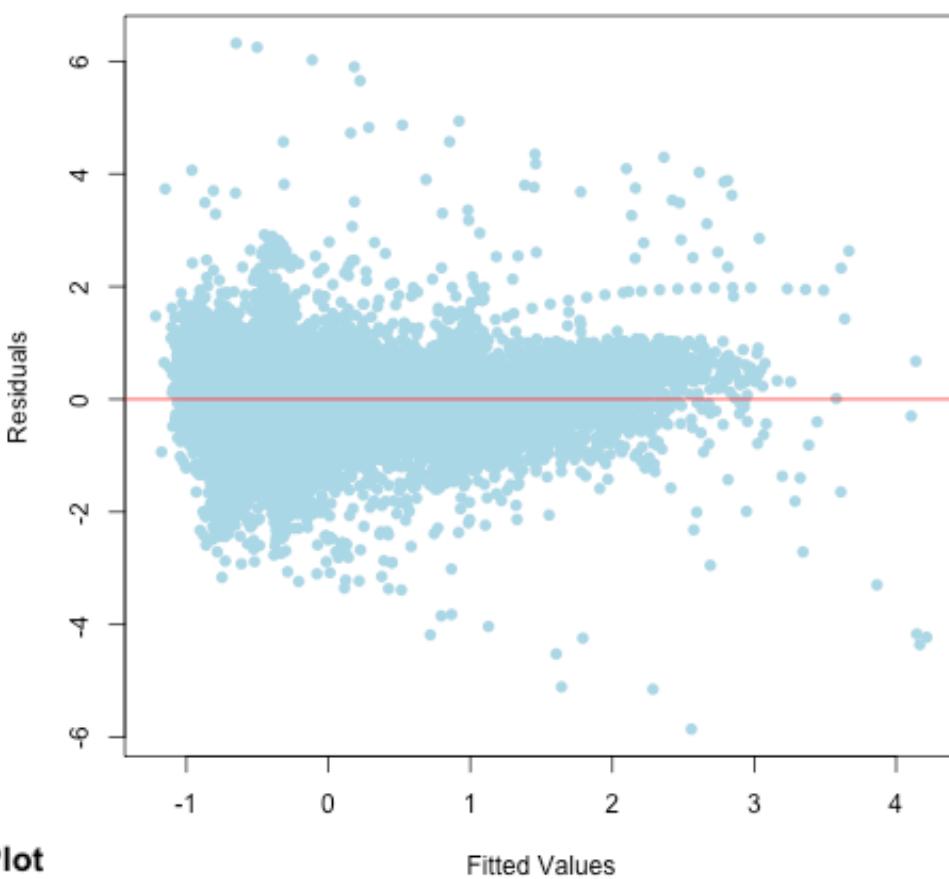
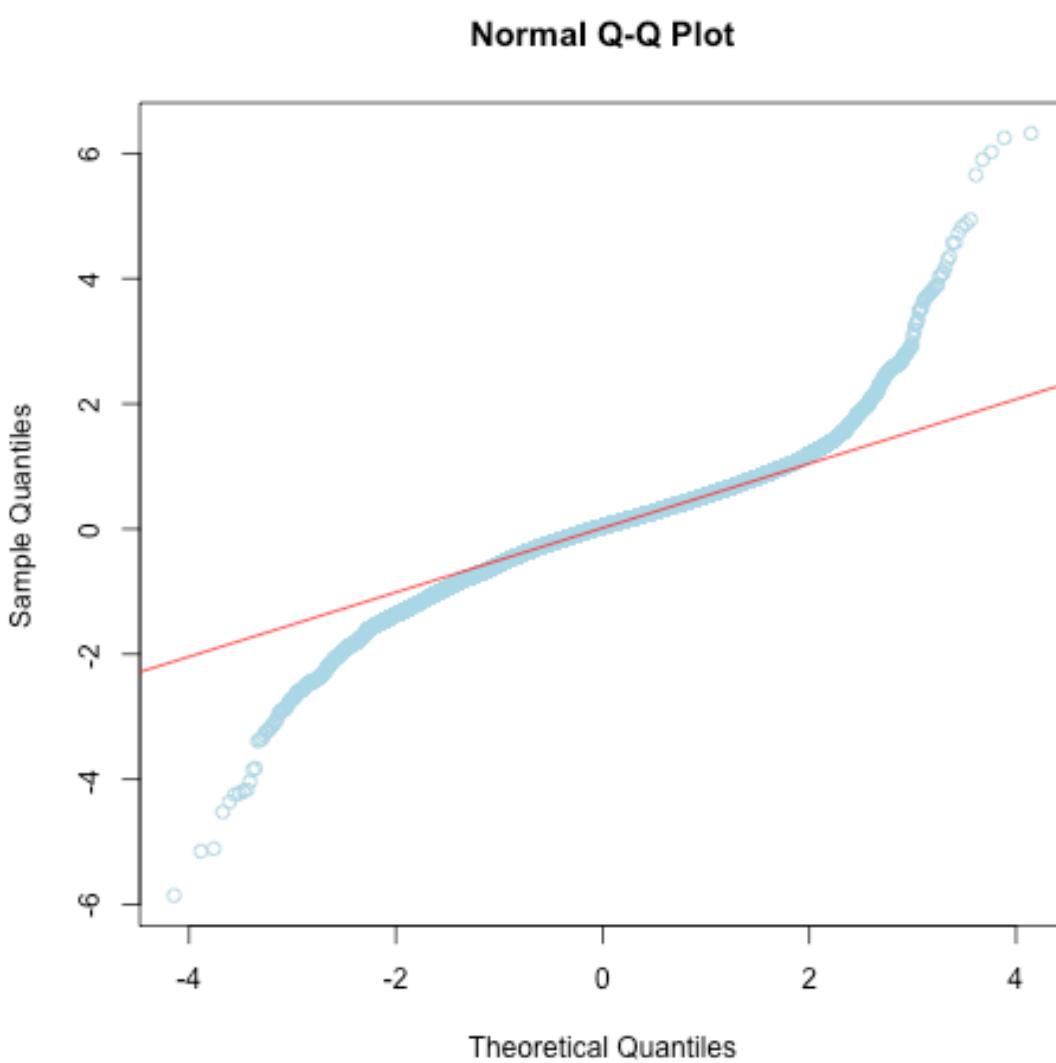
| Min | 1Q | Median | 3Q | Max |
|---------|---------|--------|--------|--------|
| -5.8544 | -0.3310 | 0.0375 | 0.3623 | 6.3192 |

Résidus : La médiane des résidus est proche de zéro, mais des valeurs extrêmes existent.

Coefficients : Tous les coefficients des variables explicatives sont significatifs avec des p-values < 0.05.

R-squared : 58.87% de la variance de CO est expliquée par le modèle.

F-statistic : Le modèle est significatif avec une p-value très faible (< 2.2e-16).



Anova

Anova :

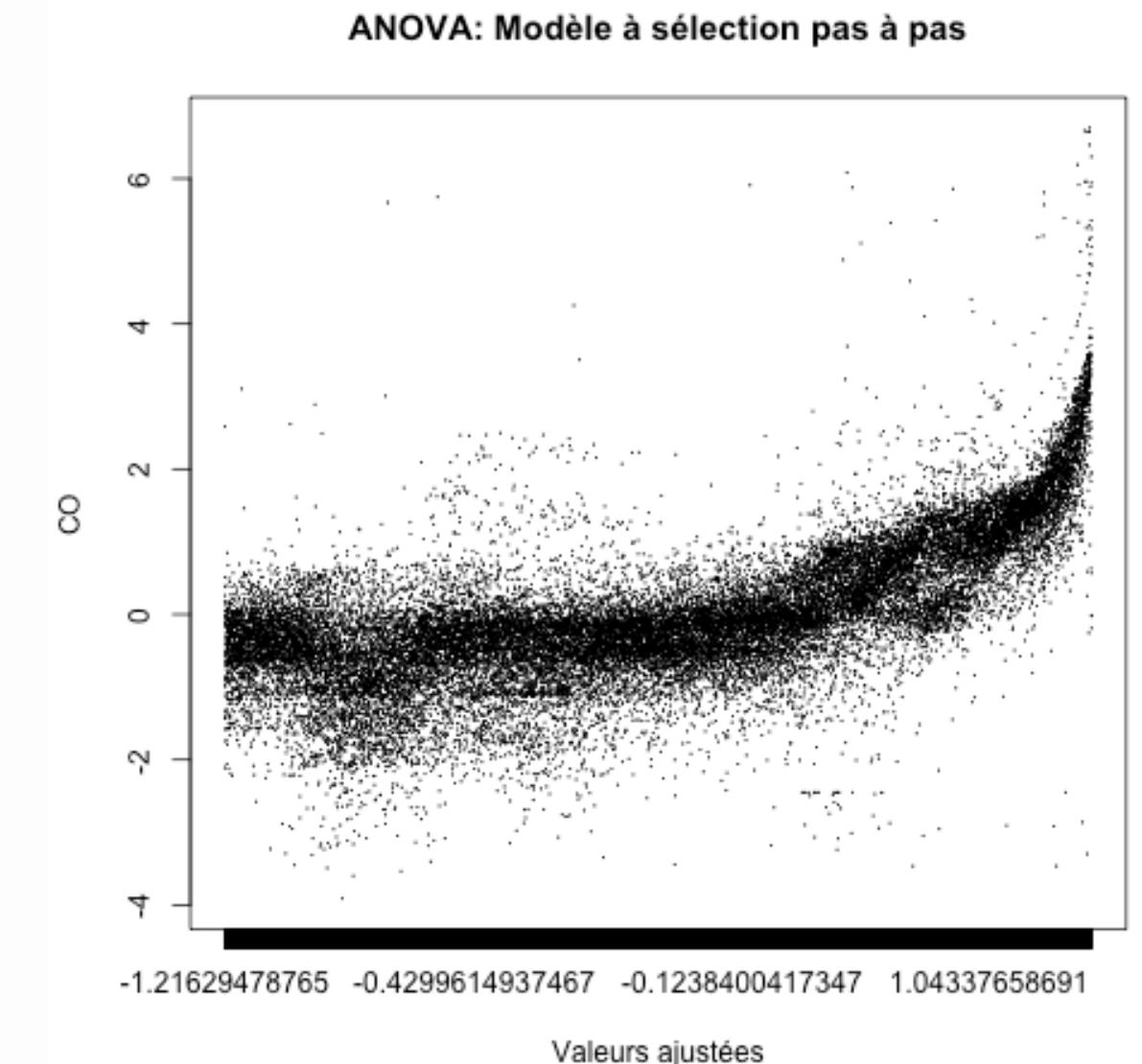
L'analyse de la variance (ANOVA) permet d'évaluer l'effet des différentes variables explicatives sur la variable cible (CO).

Réultat :

| | Df | Sum Sq | Mean Sq | F value | Pr(>F) | |
|-----------|-------|---------|---------|-----------|-----------|-----|
| GTEP | 1 | 10908.6 | 10908.6 | 26516.449 | < 2.2e-16 | *** |
| AT | 1 | 459.0 | 459.0 | 1115.612 | < 2.2e-16 | *** |
| AP | 1 | 39.2 | 39.2 | 95.368 | < 2.2e-16 | *** |
| AH | 1 | 353.8 | 353.8 | 860.020 | < 2.2e-16 | *** |
| AFDP | 1 | 334.6 | 334.6 | 813.264 | < 2.2e-16 | *** |
| TAT | 1 | 2450.0 | 2450.0 | 5955.471 | < 2.2e-16 | *** |
| TEY | 1 | 2618.9 | 2618.9 | 6366.065 | < 2.2e-16 | *** |
| NOX | 1 | 135.5 | 135.5 | 329.473 | < 2.2e-16 | *** |
| Residuals | 29377 | 12085.4 | 0.4 | | | |

Les résultats montrent que toutes les variables explicatives (GTEP, AT, AP, AH, AFDP, TAT, TEY, NOX) ont un effet très significatif sur CO, avec des p-values < 0.001.

Les résidus indiquent que 12085.4 reste non expliqué, mais les prédicteurs expliquent la majorité de la variance de CO.



Comparaison des modèles avec ANOVA :

L'analyse de la variance (ANOVA) permet d'évaluer l'effet des différentes variables explicatives sur la variable cible (CO).

Réultat :

- **Df (Degrees of Freedom)** : Indique le nombre de prédicteurs inclus dans chaque modèle.
- **Sum of Sq** : La somme des carrés représente la variation expliquée par les prédicteurs.
- **F value** : La statistique F mesure l'efficacité du modèle.
- **Pr(>F)** : La p-value montre la significativité des différences entre les modèles.

Model 1: CO ~ GTEP + AT

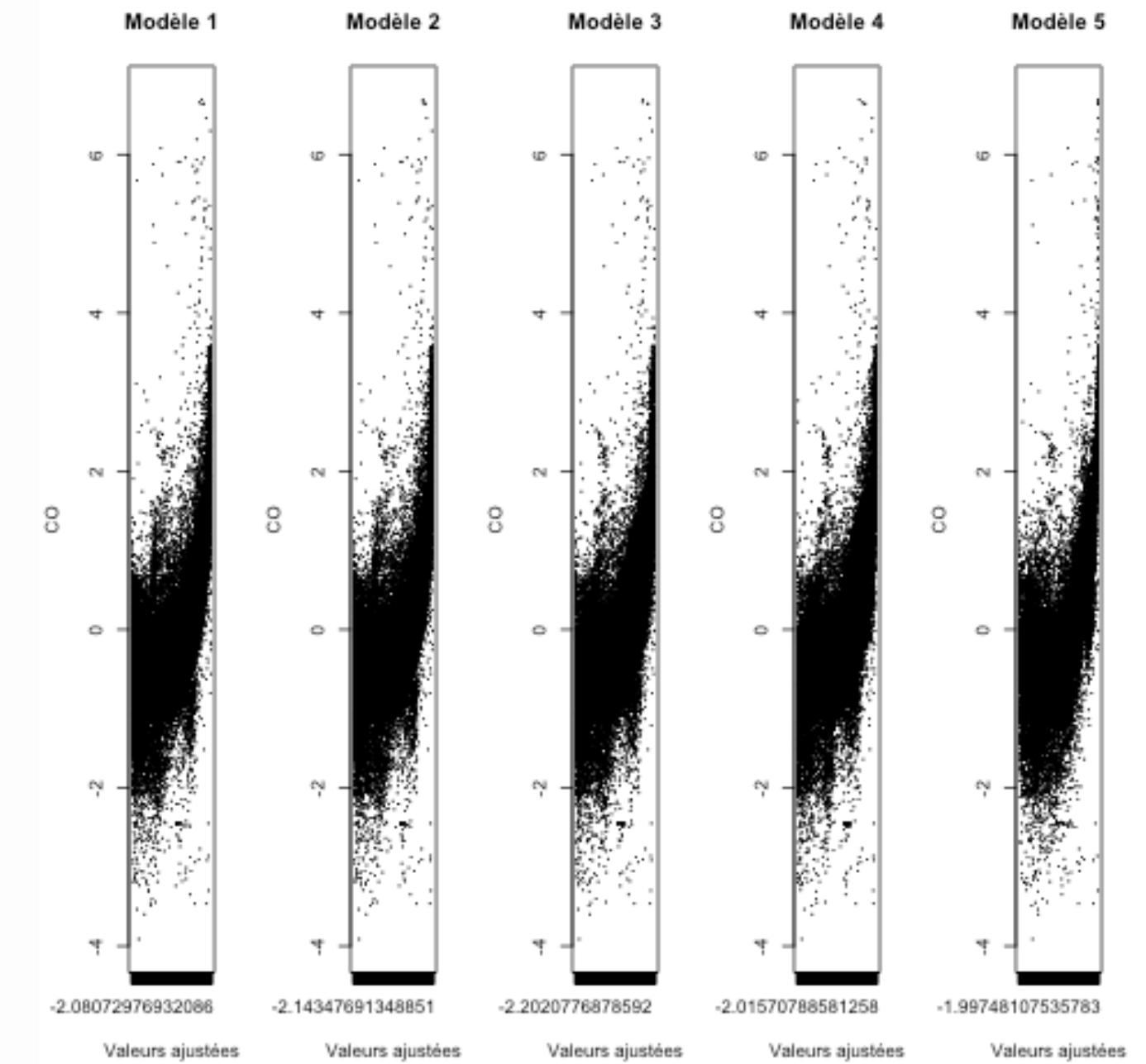
Model 2: CO ~ GTEP + AT + AP

Model 3: CO ~ GTEP + AT + AP + AH

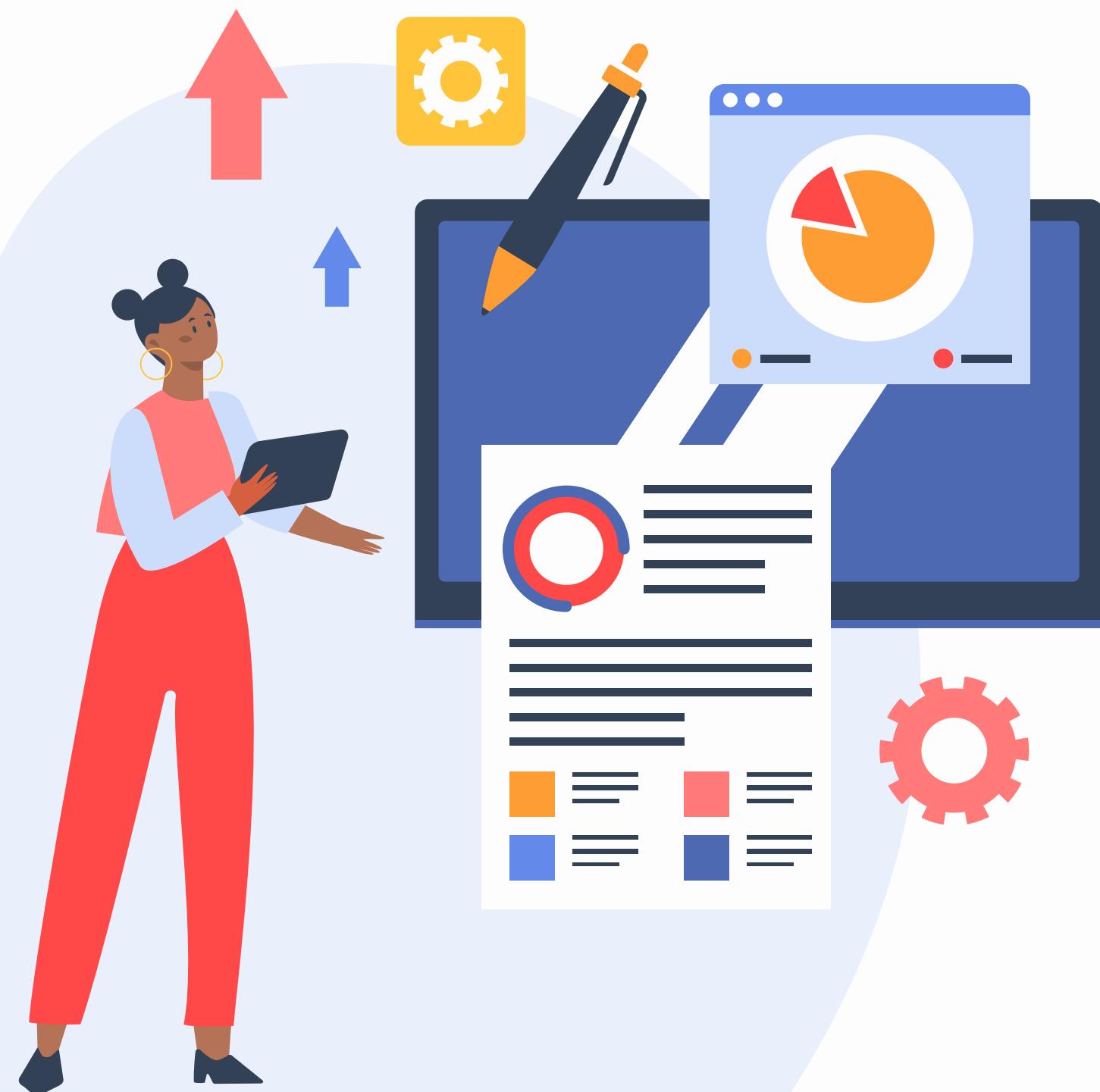
Model 4: CO ~ GTEP + AT + AP + AH + AFDP

Model 5: CO ~ GTEP + AT + AP + AH + AFDP + TAT

| | Res.Df | RSS | Df | Sum of Sq | F | Pr(>F) |
|---|--------|-------|----|-----------|------------------------|--------|
| 1 | 29383 | 18018 | | | | |
| 2 | 29382 | 17978 | 1 | 39.23 | 77.672 < 2.2e-16 *** | |
| 3 | 29381 | 17624 | 1 | 353.80 | 700.436 < 2.2e-16 *** | |
| 4 | 29380 | 17290 | 1 | 334.57 | 662.356 < 2.2e-16 *** | |
| 5 | 29379 | 14840 | 1 | 2450.02 | 4850.387 < 2.2e-16 *** | |



=> l'ajout progressif des prédicteurs améliore les performances du modèle.



Chapitre 2 :

AI4I 2020 Predictive Maintenance

Présentation du dataset



Analyse de Structure du Dataset ai4i2020 :

Summary :

Le dataset **ai4i2020** contient des informations détaillées sur différents aspects de la maintenance prédictive.

| UDI | Product.ID | Type | Air.temperature..K. | Process.temperature..K. | Rotational.speed..rpm. |
|----------------|------------------|------------------|---------------------|-------------------------|------------------------|
| Min. : 1 | Length:10000 | Length:10000 | Min. :295.3 | Min. :305.7 | Min. :1168 |
| 1st Qu.: 2501 | Class :character | Class :character | 1st Qu.:298.3 | 1st Qu.:308.8 | 1st Qu.:1423 |
| Median : 5000 | Mode :character | Mode :character | Median :300.1 | Median :310.1 | Median :1503 |
| Mean : 5000 | | | Mean :300.0 | Mean :310.0 | Mean :1539 |
| 3rd Qu.: 7500 | | | 3rd Qu.:301.5 | 3rd Qu.:311.1 | 3rd Qu.:1612 |
| Max. :10000 | | | Max. :304.5 | Max. :313.8 | Max. :2886 |
| Torque..Nm. | Tool.wear..min. | Machine.failure | TWF | HDF | PWF |
| Min. : 3.80 | Min. : 0 | Min. :0.0000 | Min. :0.0000 | Min. :0.0000 | Min. :0.0000 |
| 1st Qu.:33.20 | 1st Qu.: 53 | 1st Qu.:0.0000 | 1st Qu.:0.0000 | 1st Qu.:0.0000 | 1st Qu.:0.0000 |
| Median :40.10 | Median :108 | Median :0.0000 | Median :0.0000 | Median :0.0000 | Median :0.0000 |
| Mean :39.99 | Mean :108 | Mean :0.0339 | Mean :0.0046 | Mean :0.0115 | Mean :0.0095 |
| 3rd Qu.:46.80 | 3rd Qu.:162 | 3rd Qu.:0.0000 | 3rd Qu.:0.0000 | 3rd Qu.:0.0000 | 3rd Qu.:0.0000 |
| Max. :76.60 | Max. :253 | Max. :1.0000 | Max. :1.0000 | Max. :1.0000 | Max. :1.0000 |
| RNF | | | | | |
| Min. :0.0000 | | | | | |
| 1st Qu.:0.0000 | | | | | |
| Median :0.0000 | | | | | |
| Mean :0.0019 | | | | | |
| 3rd Qu.:0.0000 | | | | | |
| Max. :1.0000 | | | | | |

Structure du Dataset :

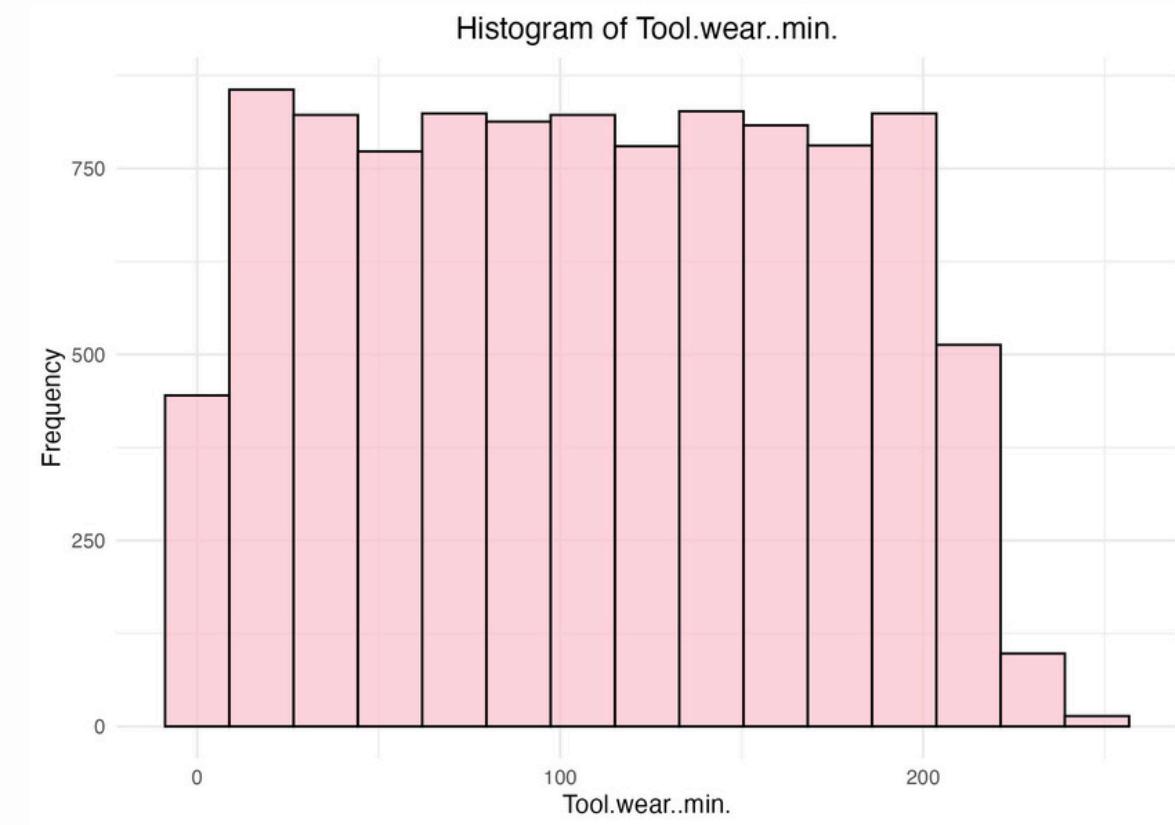
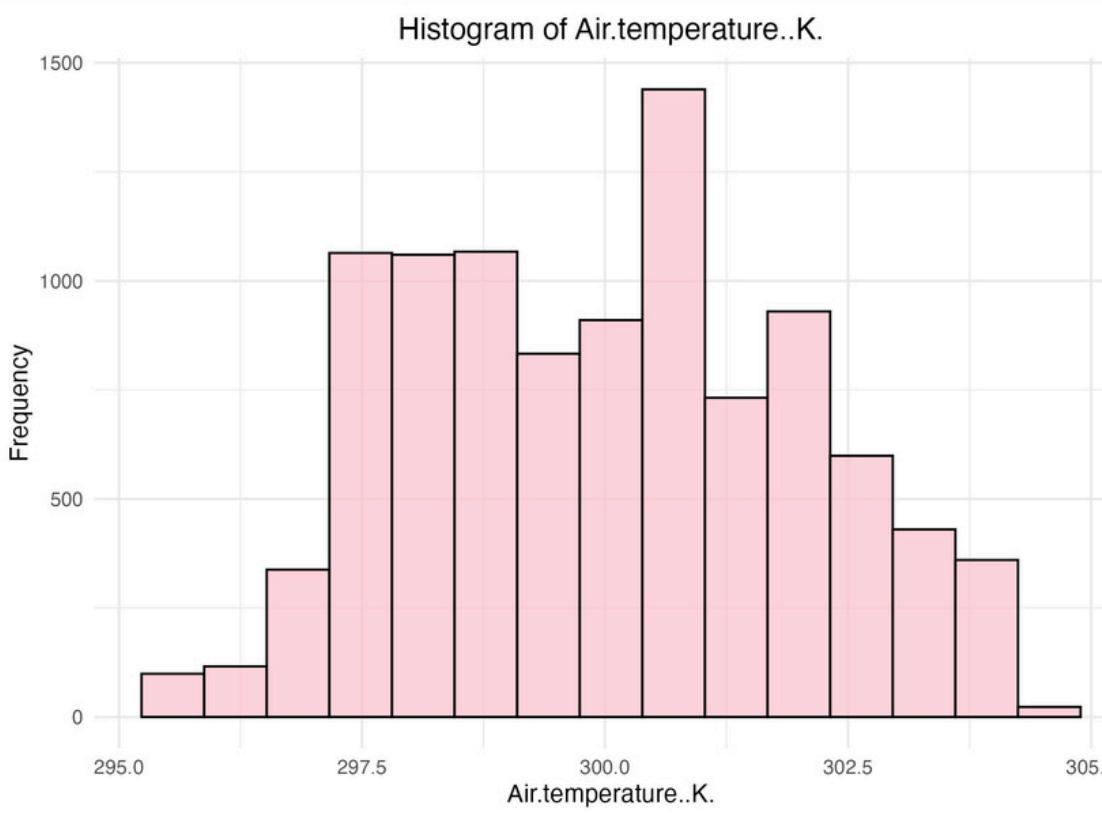
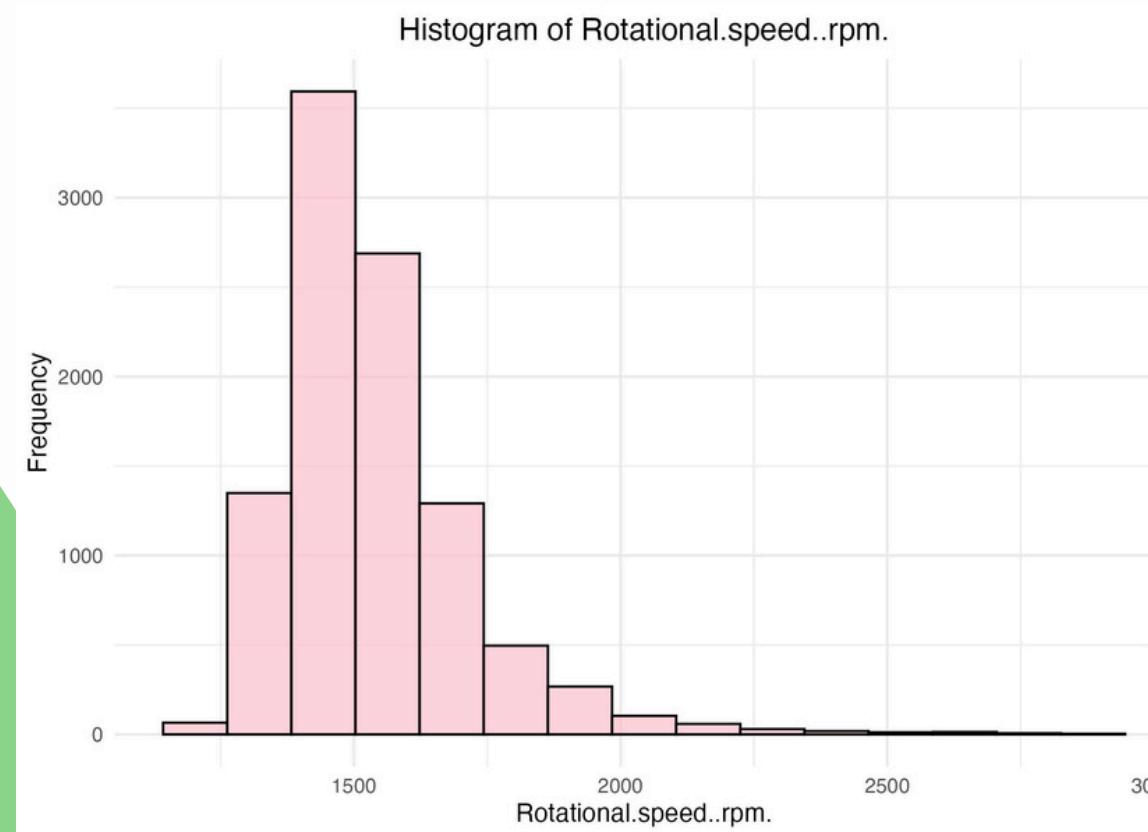
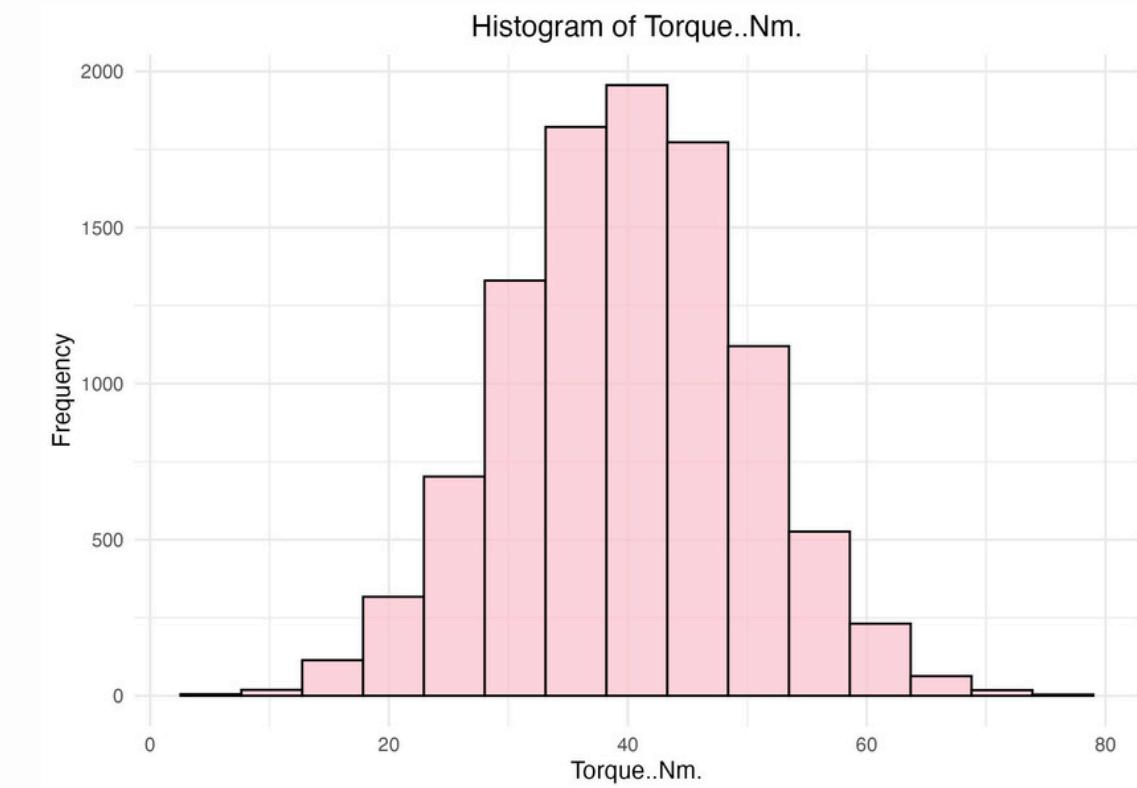
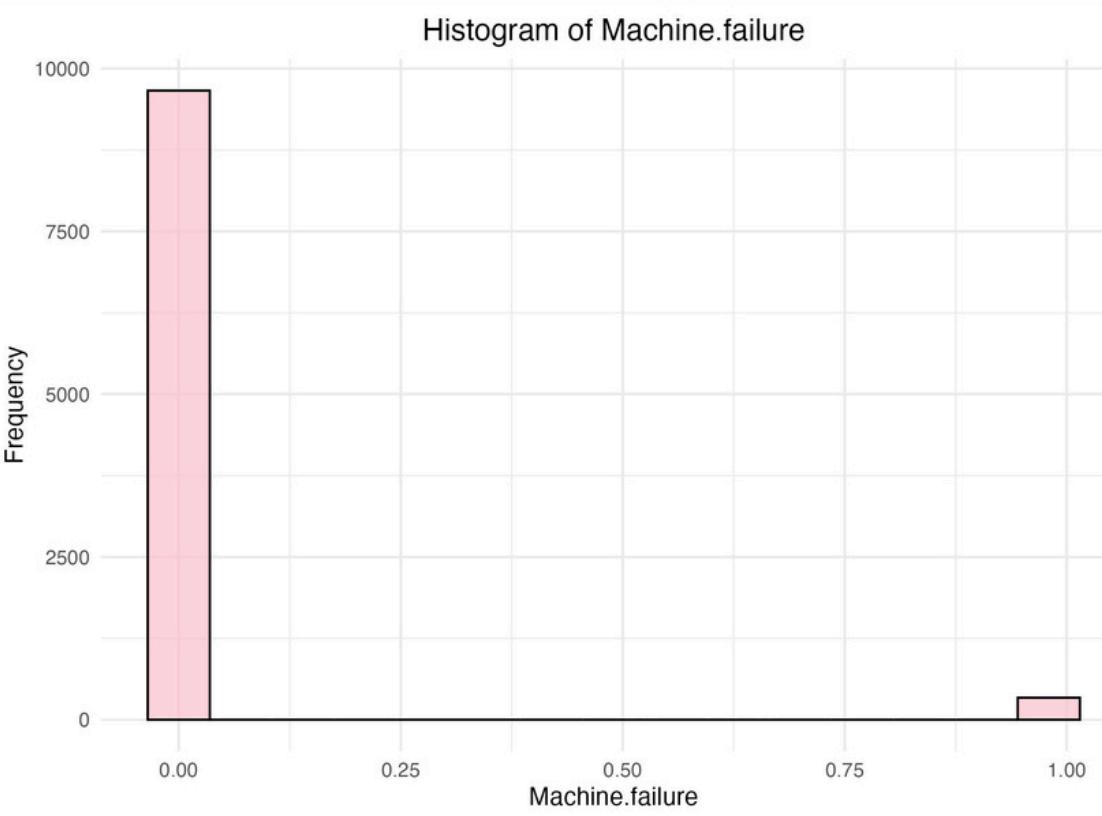
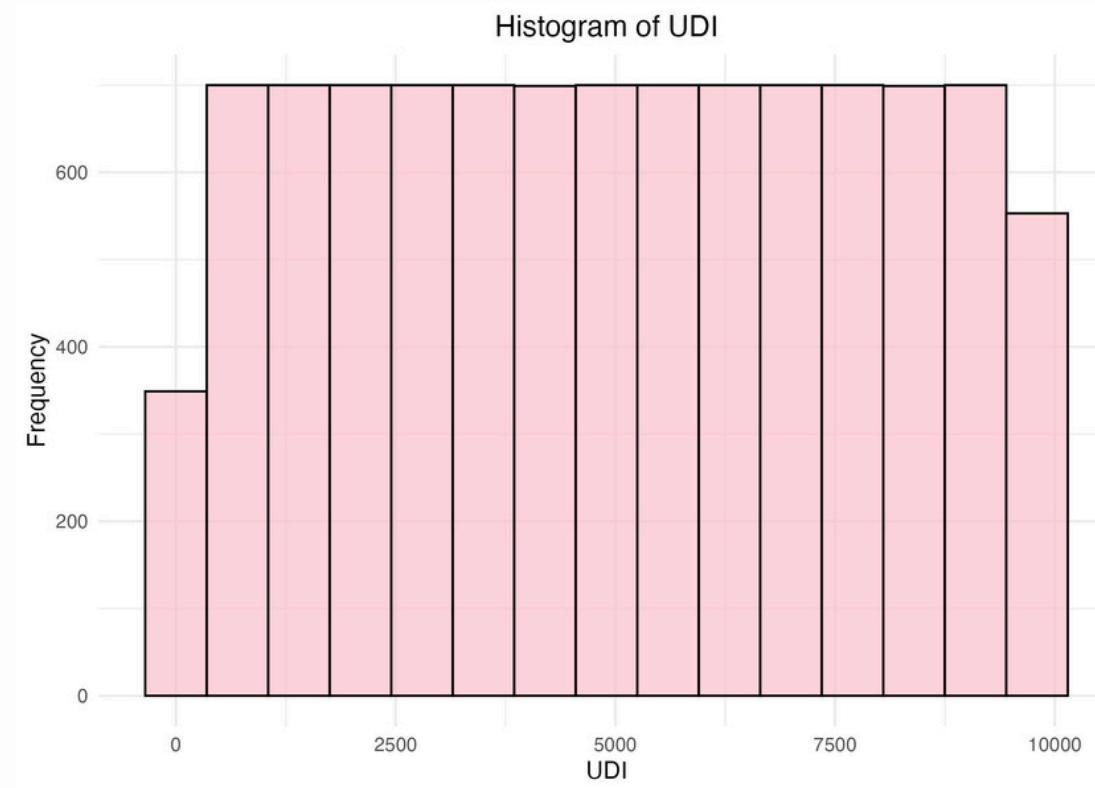
Le dataset **ai4i2020** contient **10,000** observations réparties sur **14 variables numériques et catégoriques**.

Voici les types observés :

| | | |
|----------------------------|-------|---|
| \$ UDI | : int | 1 2 3 4 5 6 7 8 9 10 ... |
| \$ Product.ID | : chr | "M14860" "L47181" "L47182" "L47183" ... |
| \$ Type | : chr | "M" "L" "L" "L" ... |
| \$ Air.temperature..K. | : num | 298 298 298 298 298 ... |
| \$ Process.temperature..K. | : num | 309 309 308 309 309 ... |
| \$ Rotational.speed..rpm. | : int | 1551 1408 1498 1433 1408 1425 1558 1527 1667 ... |
| \$ Torque..Nm. | : num | 42.8 46.3 49.4 39.5 40 41.9 42.4 40.2 28.6 28 ... |
| \$ Tool.wear..min. | : int | 0 3 5 7 9 11 14 16 18 21 ... |
| \$ Machine.failure | : int | 0 0 0 0 0 0 0 0 0 0 ... |
| \$ TWF | : int | 0 0 0 0 0 0 0 0 0 0 ... |
| \$ HDF | : int | 0 0 0 0 0 0 0 0 0 0 ... |
| \$ PWF | : int | 0 0 0 0 0 0 0 0 0 0 ... |
| \$ OSF | : int | 0 0 0 0 0 0 0 0 0 0 ... |
| \$ RNF | : int | 0 0 0 0 0 0 0 0 0 0 ... |

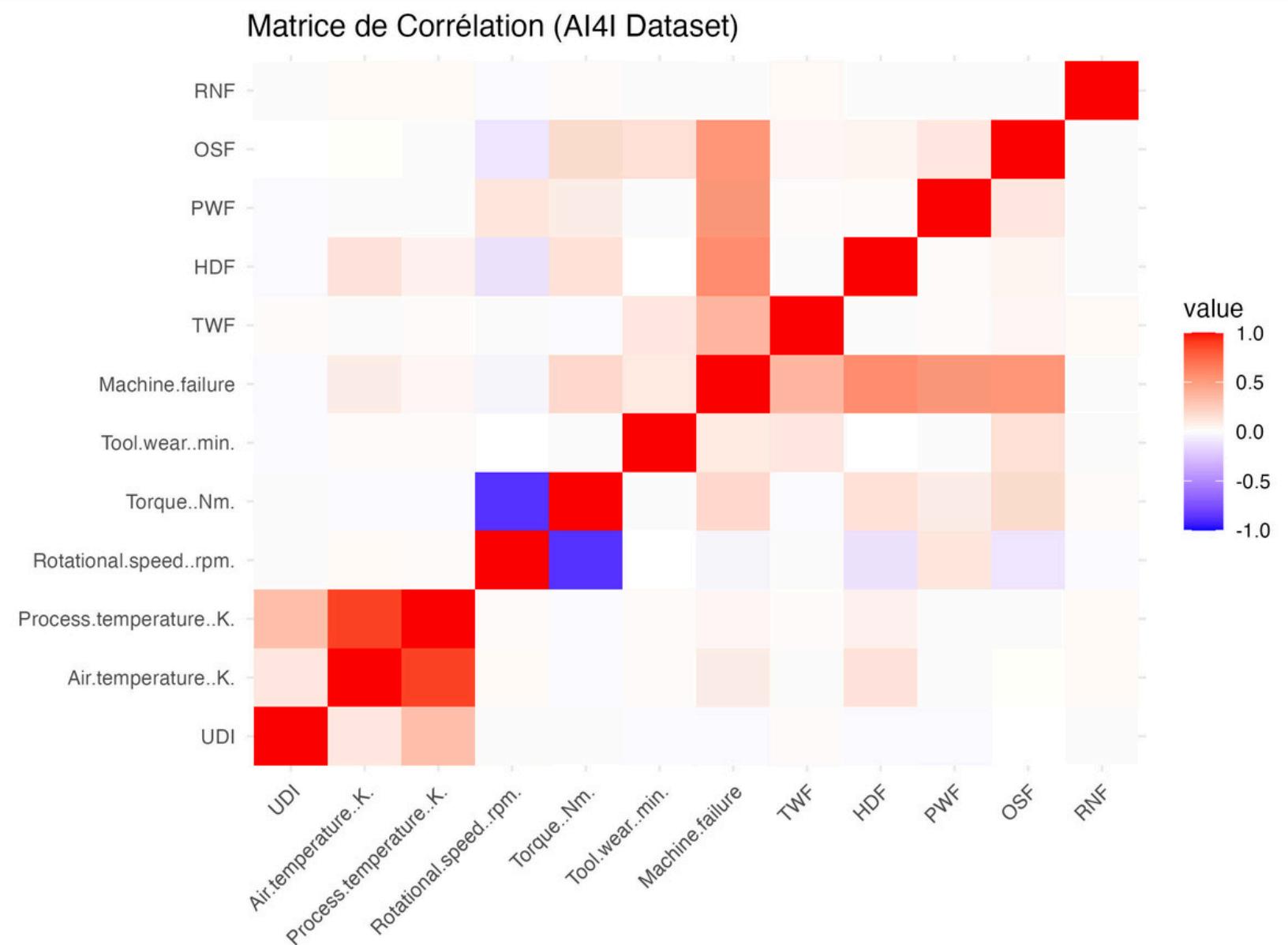
Exploration des données

Histogrames



Matrice de corrélation

La matrice de corrélation montre les relations linéaires entre les différentes variables numériques . Voici quelques observations clés :



Vitesse de rotation (Rotational.speed..rpm.) :
Corrélation positive avec Torque..Nm. (~0.9), indiquant que l'augmentation de la vitesse entraîne généralement une augmentation du couple.

Défaillances (Machine.failure, TWF, HDF, etc.) :
Corrélation élevée entre les différentes types de défaillances, ce qui montre une relation significative entre elles.

Température Processus (Process.temperature..K.) :
Corrélation positive avec Rotational.speed..rpm. (~0.85), indiquant que des températures plus élevées sont associées à des vitesses de rotation plus élevées.

Préparation des données



Vérification des valeurs manquantes:

| | UDI | Product.ID | Type | Air.temperature..K. | Process.temperature..K. |
|------------------------|-----|-------------|-----------------|---------------------|-------------------------|
| | 0 | 0 | 0 | 0 | 0 |
| Rotational.speed..rpm. | | Torque..Nm. | Tool.wear..min. | Machine.failure | TWF |
| | 0 | 0 | 0 | 0 | 0 |
| HDF | | PWF | OSF | RNF | |
| | 0 | 0 | 0 | 0 | |

Pas de valeurs manquantes pour toutes les variables

Afficher les colonnes numériques :

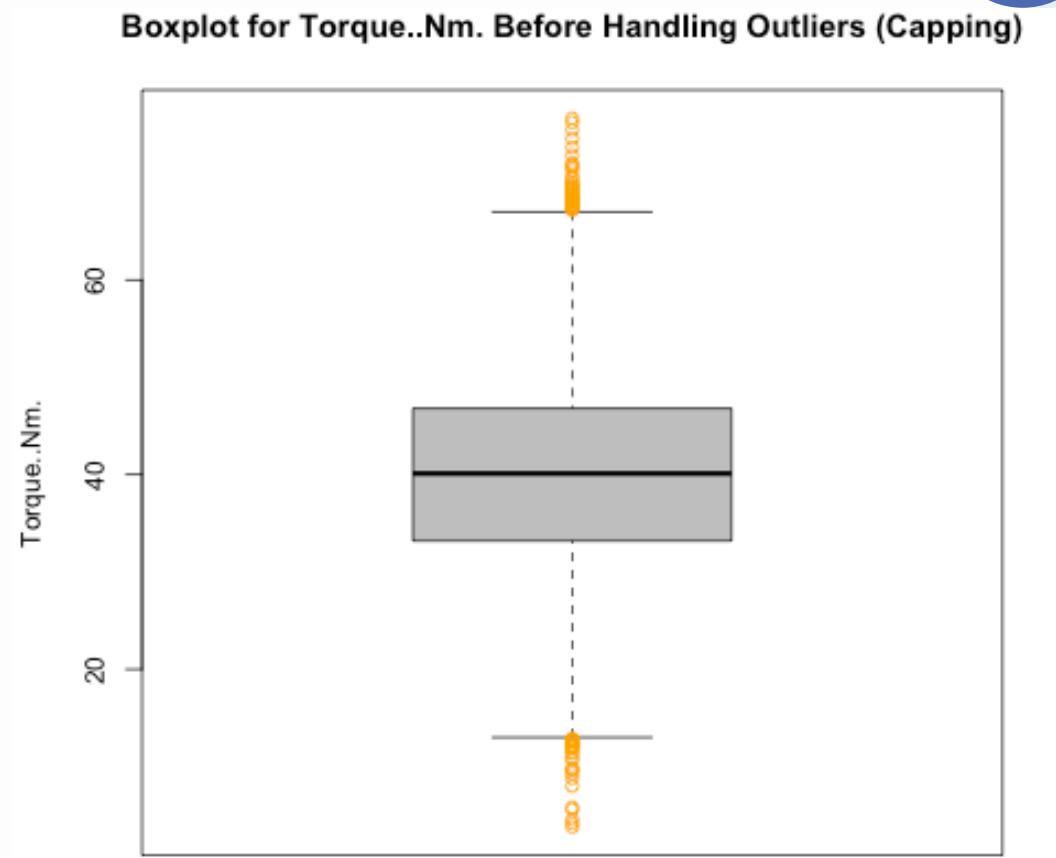
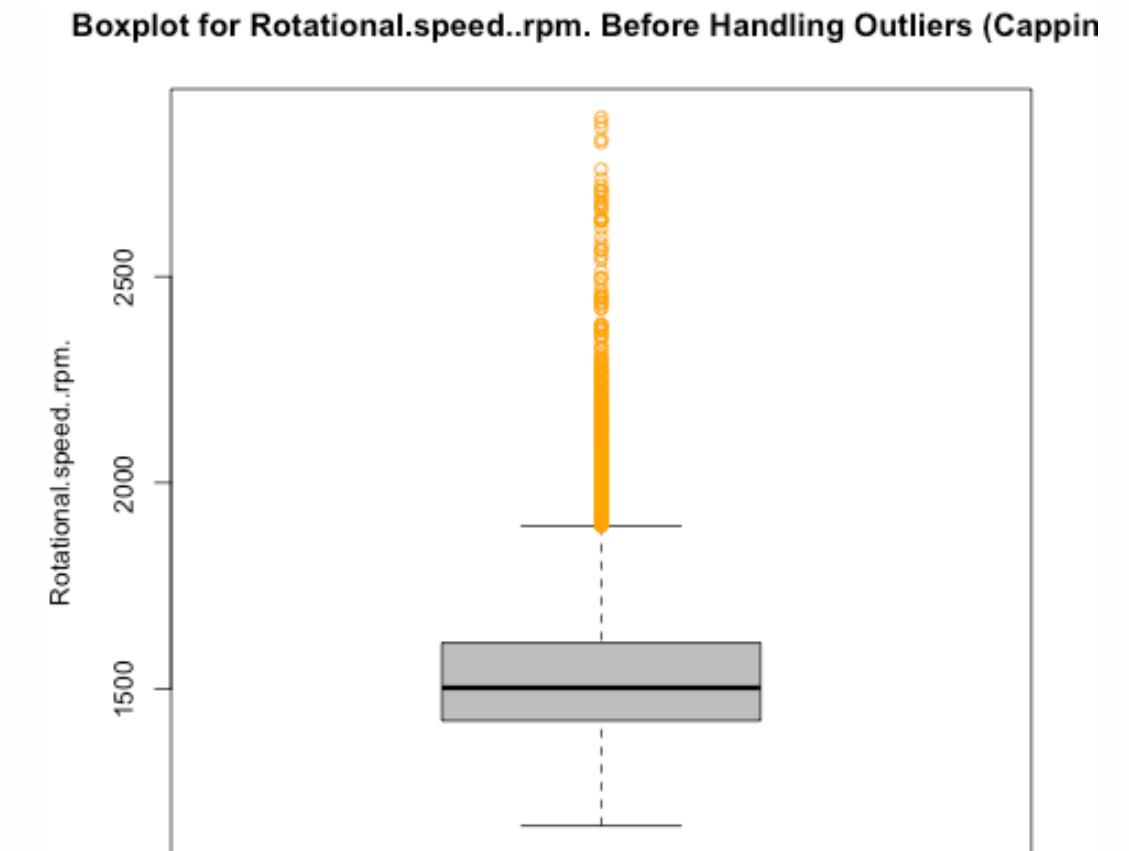
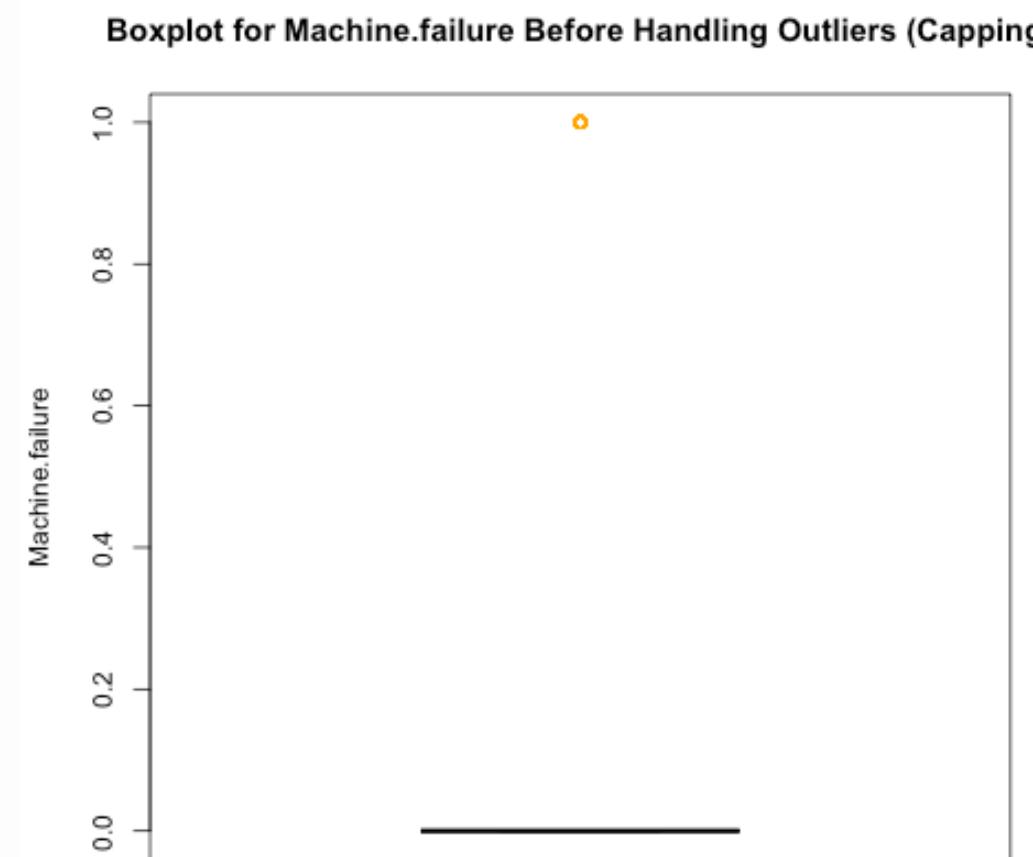
```
[1] "UDI"                      "Air.temperature..K."    "Process.temperature..K." "Rotational.speed..rpm."
[5] "Torque..Nm."                "Tool.wear..min."        "Machine.failure"       "TWF"
[9] "HDF"                       "PWF"                   "OSF"                  "RNF"
```

il ya 12 colonnes numeriques (2 categoriques)



Traitement des valeurs aberrantes :

Boxplots Avant le Traitement des Valeurs Aberrantes :



Les boxplots révèlent des plages de variations de valeurs, où les points orange identifient également les valeurs aberrantes.

La distribution de plusieurs variables est influencée par ces outliers.



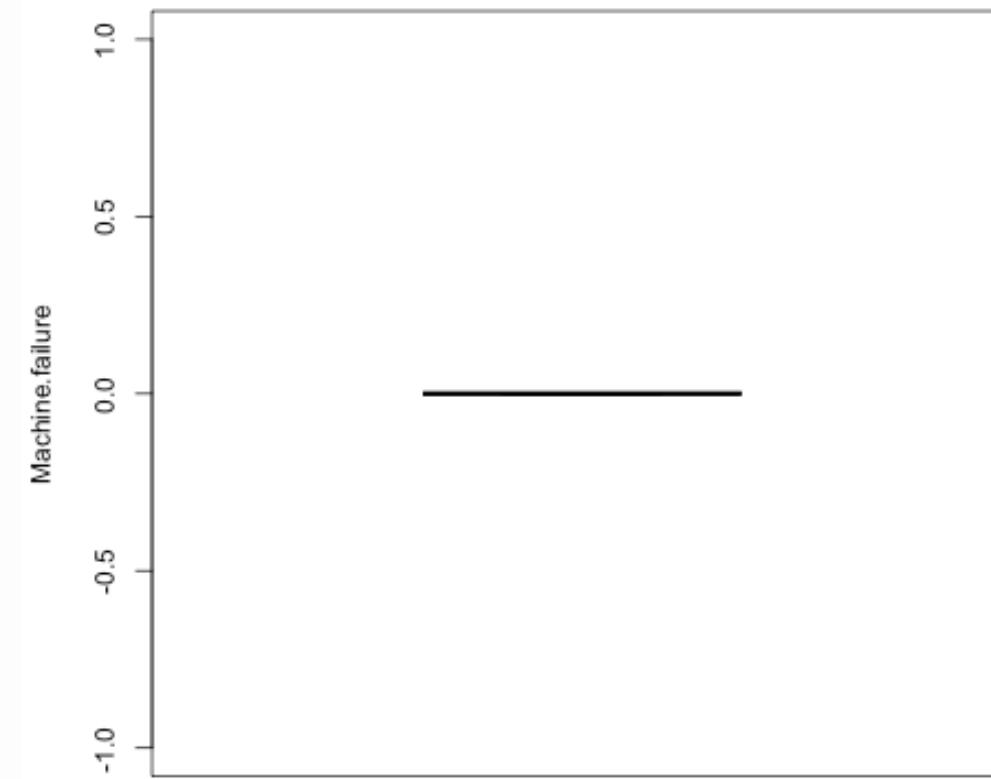
Traitement des valeurs aberrantes :

Boxplots Apres le Traitement des Valeurs Aberrantes :

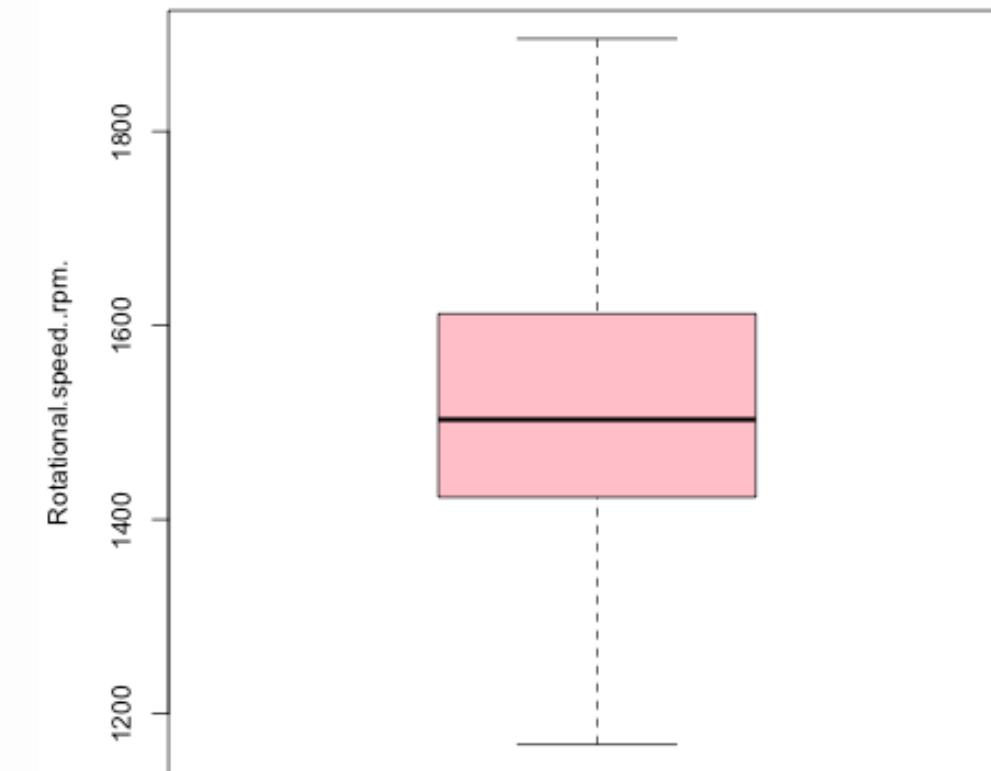
La fonction `cap_outliers(data, column)` a été utilisée pour limiter les valeurs extrêmes en ajustant les valeurs inférieures et supérieures basées sur l'IQR.

Voici comment cela impacte les distributions :

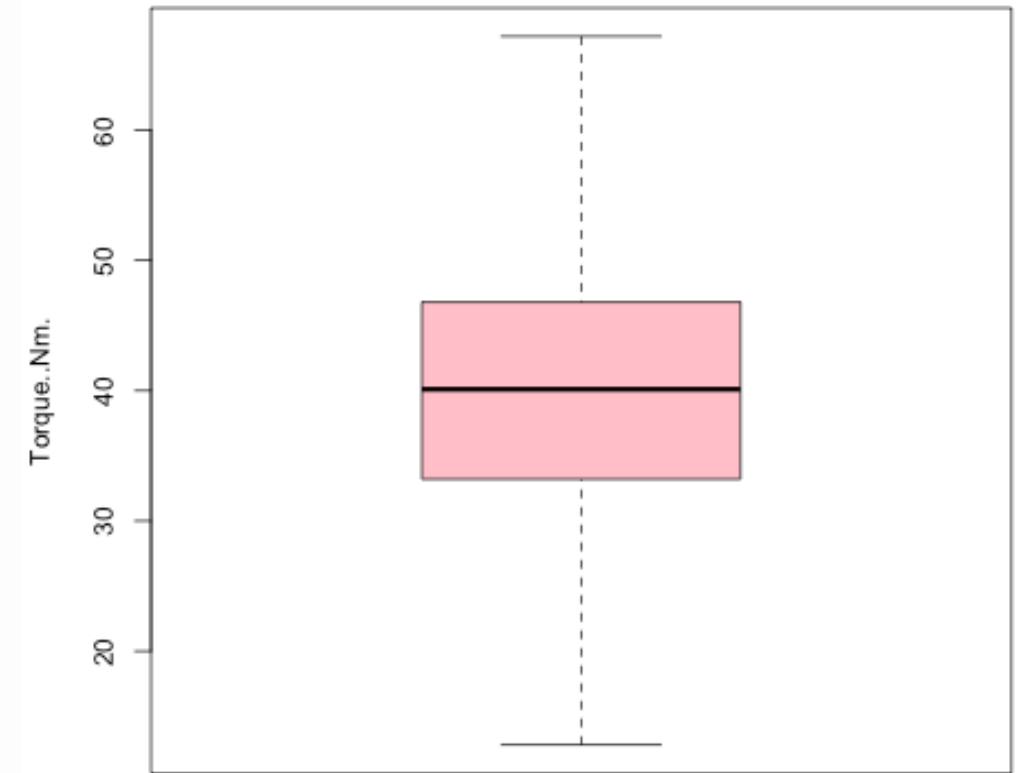
Boxplot for Machine.failure (AI4I Dataset) After Handling Outliers (Cap)



plot for Rotational.speed..rpm. (AI4I Dataset) After Handling Outliers (C)



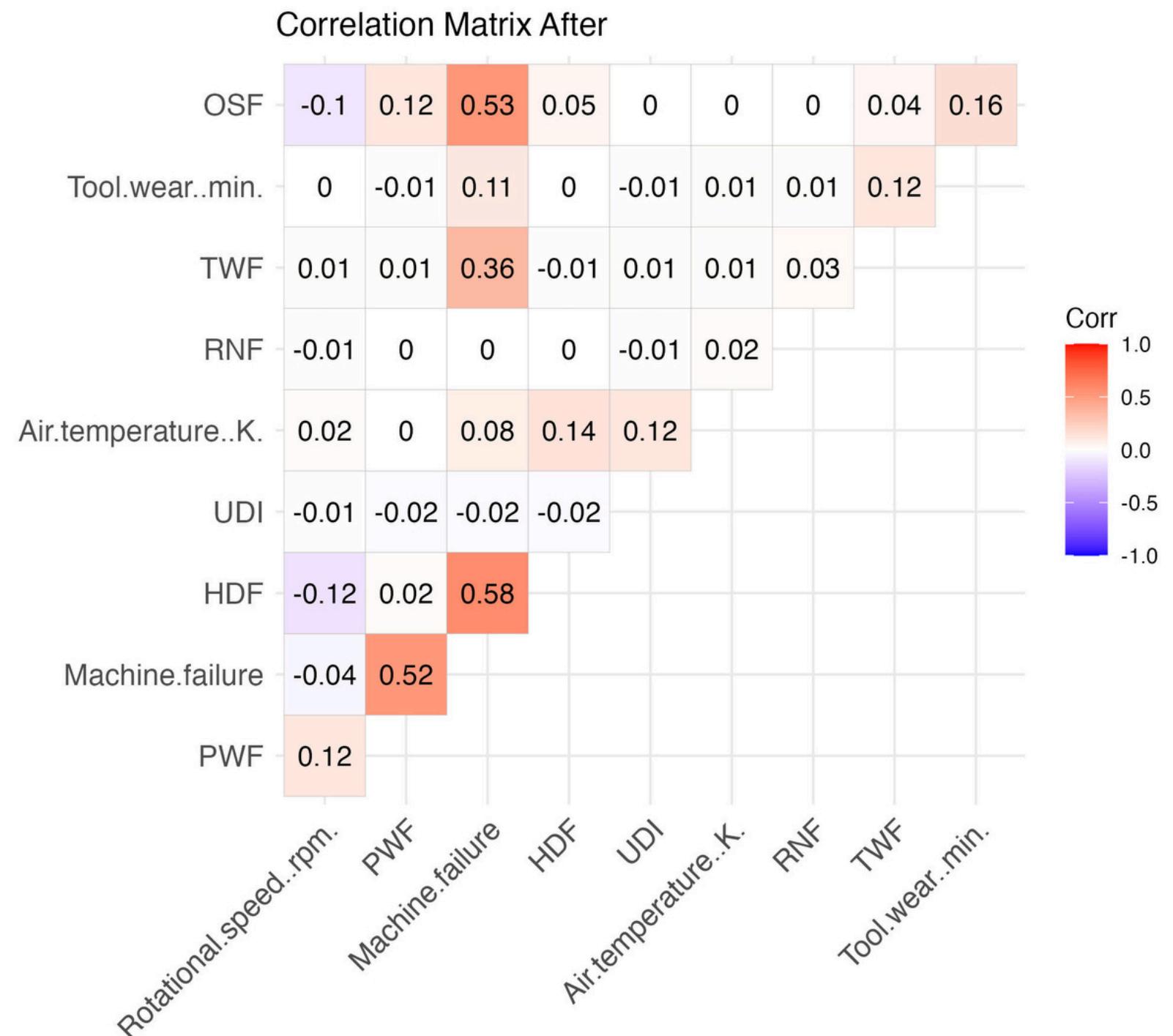
Boxplot for Torque.Nm. (AI4I Dataset) After Handling Outliers (Cap)



Même processus appliqué, les boxplots indiquent une réduction des variations extrêmes, améliorant ainsi la qualité des données analysées.

Supprimer les colonnes fortement corrélées

analyser les corrélations entre les colonnes numériques d'un dataset et à supprimer les colonnes redondantes qui présentent une corrélation élevée avec d'autres (supérieure à 0.8)



Après avoir identifié les colonnes corrélées, elles sont supprimées itérativement jusqu'à ce que la matrice de corrélation réduite ne contienne que des relations faibles ou modérées.

Enfin, une heatmap est générée pour visualiser les corrélations restantes et sauvegarder les résultats pour une utilisation future.

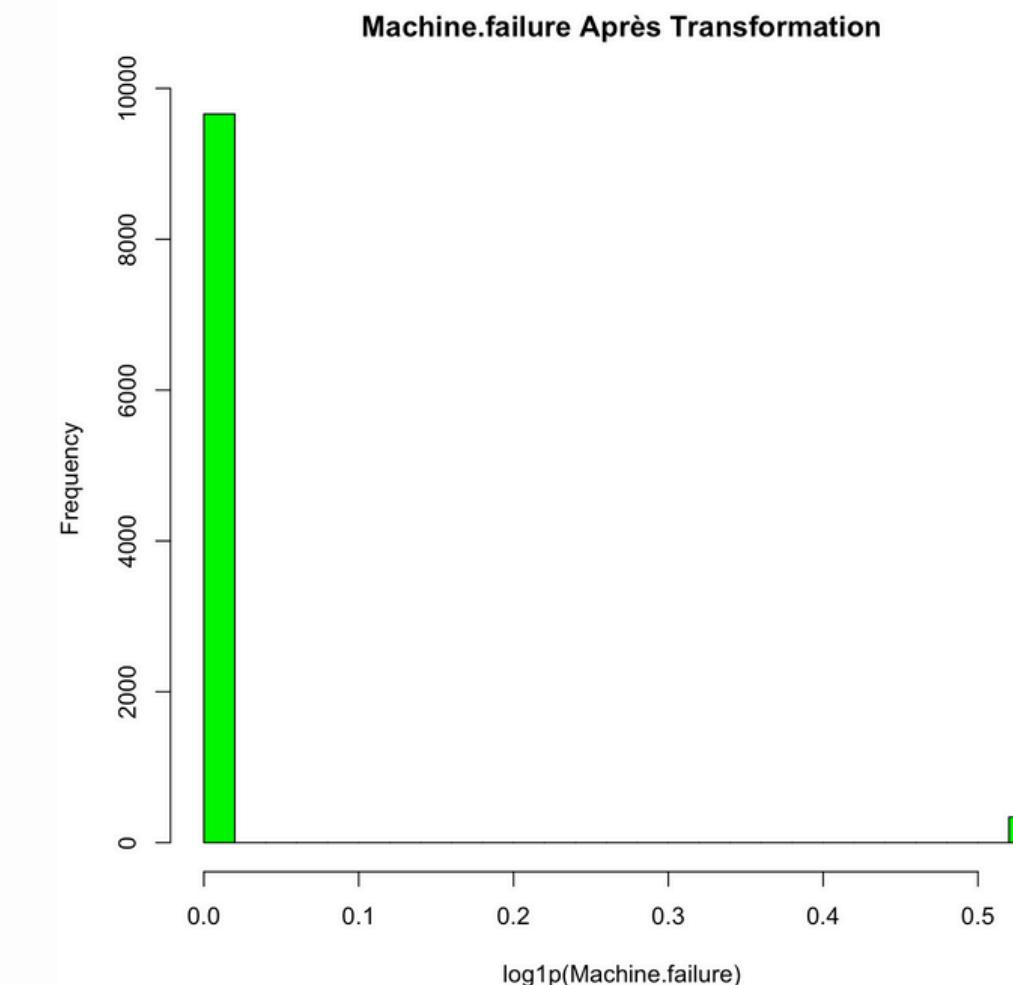
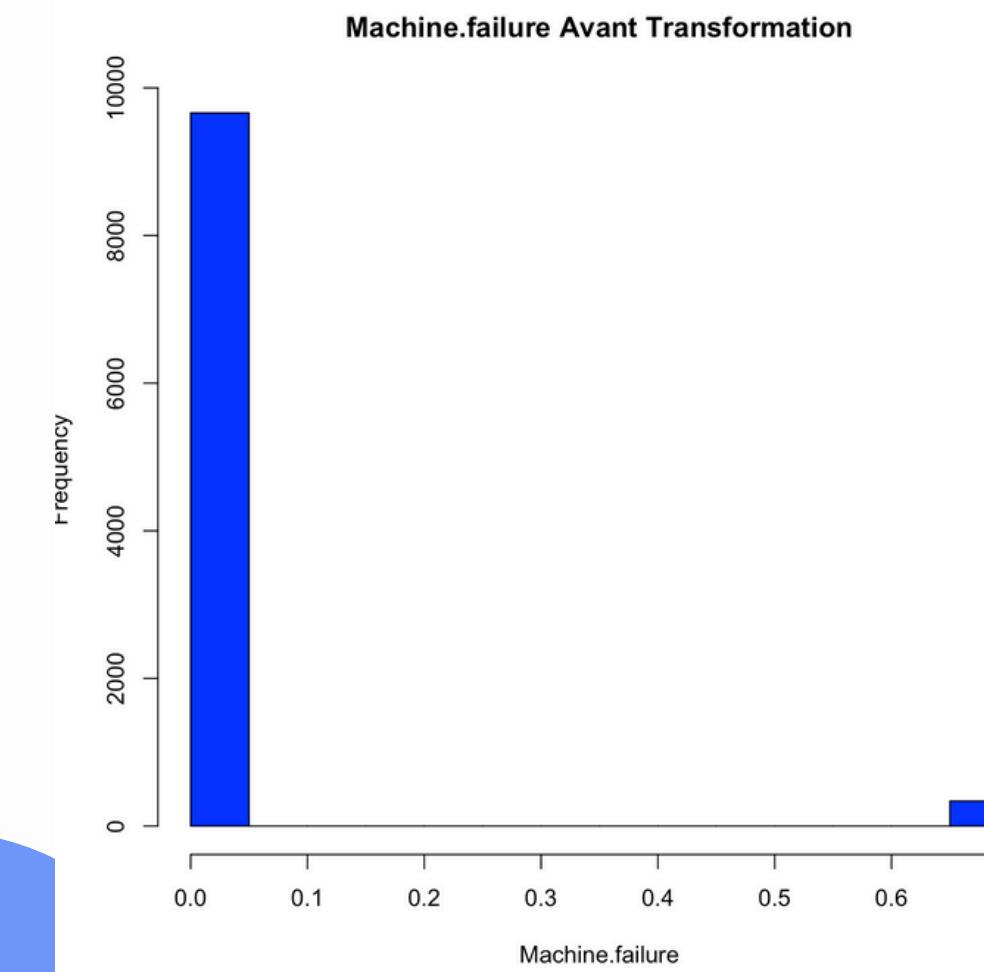
Cette approche assure un dataset compact et optimisé pour des analyses ultérieures.

Transformation de la variable cible (Machine Failure):

La transformation de la variable cible Machine.failure vise à corriger les problèmes de non-normalité qui pourraient compromettre les hypothèses des modèles statistiques comme la régression linéaire.

Resultat attendu :

Améliorer l'adéquation aux hypothèses des modèles statistiques, comme la normalité et l'homogénéité des variances.



Test de normalité avec Kolmogorov-Smirnov :

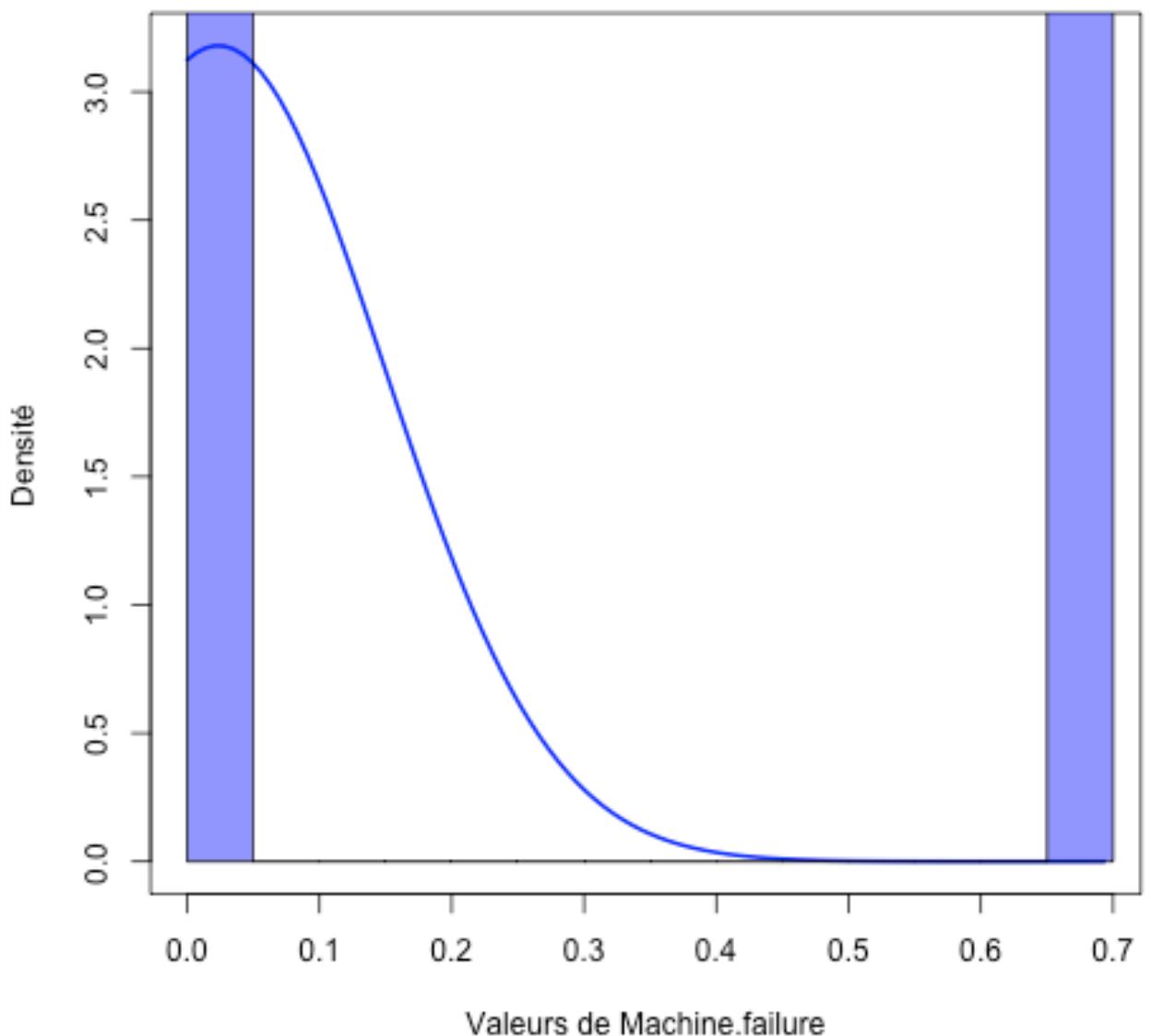
Réultat :

- **D = 0.54039** représente la plus grande différence entre la distribution empirique de Machine.failure et la distribution normale hypothétique.
- **P-value < 2.2e-16 :**
- La p-value **extrêmement faible (< 0.05)** indique que l'hypothèse nulle de normalité est **rejetée**.
- La variable Machine.failure **ne suit pas une distribution normale.**

Asymptotic one-sample Kolmogorov-Smirnov test

```
data: ai4i2020$Machine.failure
D = 0.54039, p-value < 2.2e-16
alternative hypothesis: two-sided
```

Courbe Logarithmique de la Distribution Normale



Ajout de termes quadratiques :

- Ajouter un terme quadratique pour la variable Torque..Nm.. Cela permet d'évaluer si une relation non linéaire entre Torque..Nm. et une variable cible peut améliorer la performance des modèles prédictifs.
- Introduire une interaction entre Air.temperature..K. et Torque..Nm.. Cela permet de capturer l'effet combiné de ces deux variables, qui pourrait être significatif pour expliquer les variations dans la variable cible.

Division des données :

Permet de diviser le dataset [AI2020](#) en deux ensembles : entraînement (80%) et test (20%).

Standardisation :

[La standardisation](#) des données consiste à ajuster les valeurs afin qu'elles aient une moyenne de 0 et un écart type 1.

Cette transformation améliore la convergence et la précision des modèles statistiques.

| AirTorque |
|-------------------|
| Min. :-3.598370 |
| 1st Qu.:-0.682115 |
| Median : 0.007472 |
| Mean : 0.000000 |
| 3rd Qu.: 0.683787 |
| Max. : 3.687074 |

Analyse statistique exploratoire et tests



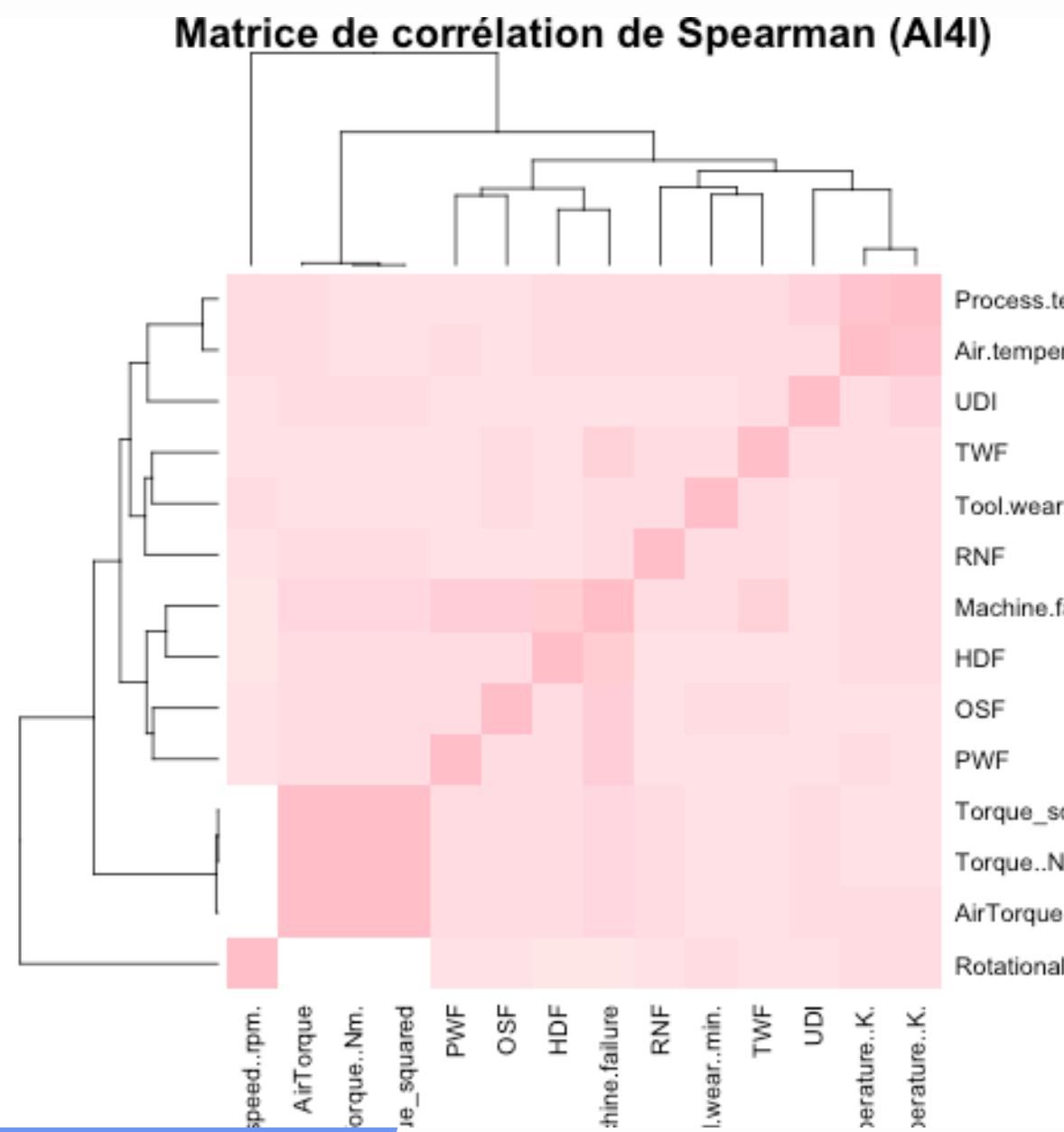
Analyse de corrélation:

Corrélation de Pearson:

Mesure la force et de la direction des relations linéaires entre les variables numériques.

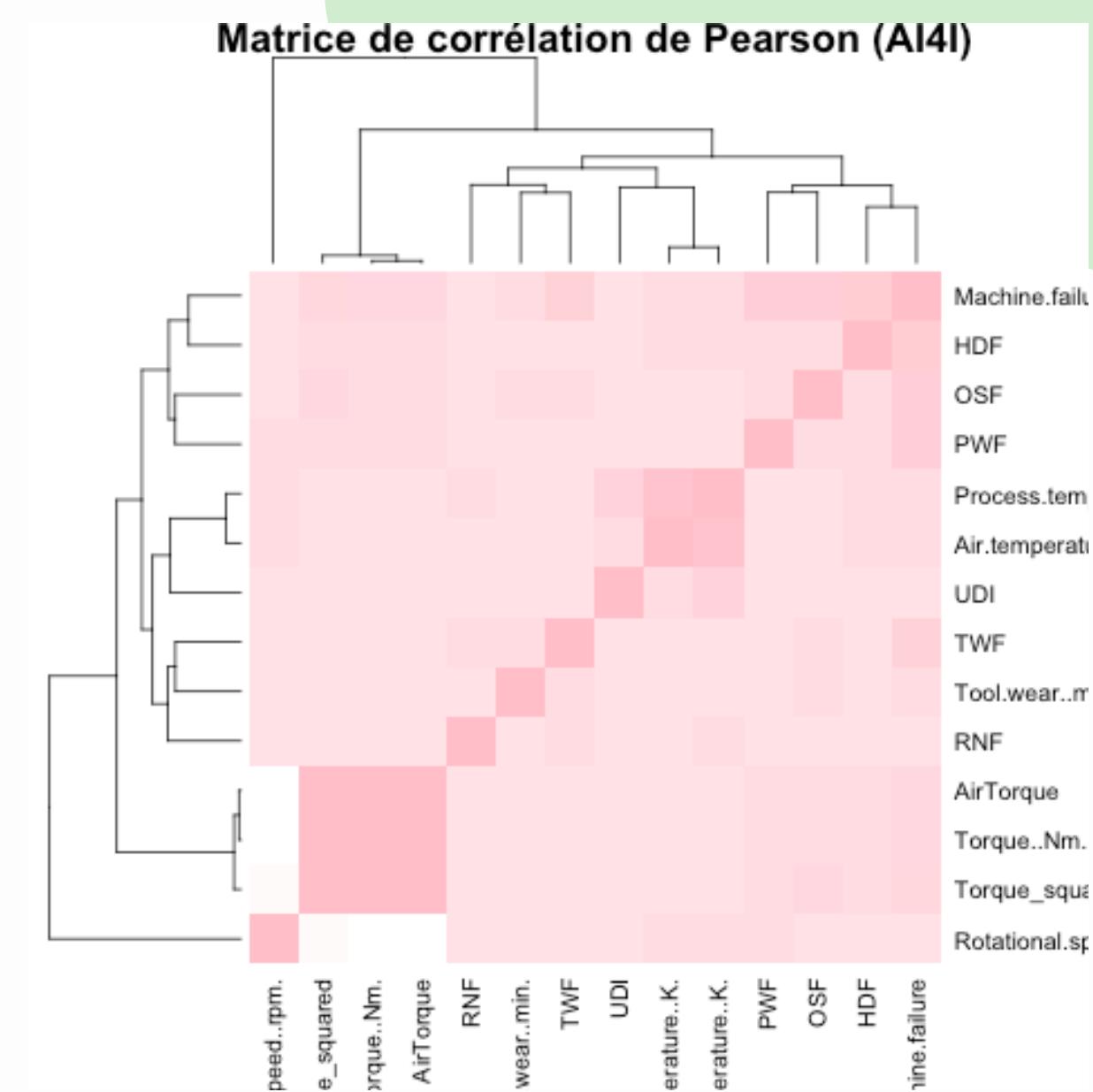
- Rotational Speed et TIT montrent une corrélation extrêmement forte indiquant une relation presque linéaire parfaite.
- AirTorque et Rotational speed ont une corrélation de 0, ce qui suggère une relation inexistante .

Corrélation de Spearman :



Contrairement à **Pearson**, qui mesure la corrélation **linéaire**,
Spearman peut être utilisé lorsque les **relations ne sont pas strictement linéaires**, mais plutôt ordinaires ou monotones.

- AirTemperature et ProcessTemperature montrent une forte corrélation positive (0.999), indiquant une relation très proche, même non-linéaire.
- RNF et Rotational Speed ont une corrélation modérée (0.053), reflétant une relation moins directe ou linéaire.



Nuages de points pour Machine Failure par rapport aux autres variables :

Les nuages de points visualisent la relation entre la variable cible Machine Failure et chaque autre variable numérique du dataset. Voici les éléments clés à analyser :

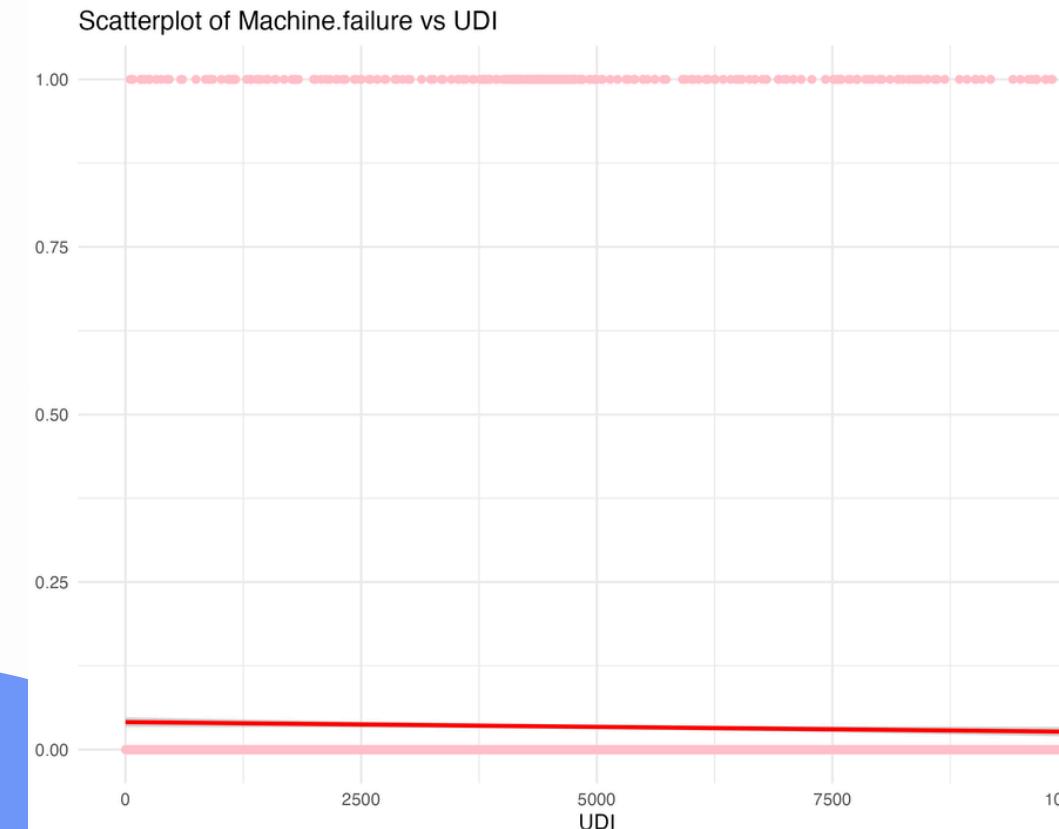
Relation linéaire ou non-linéaire:

- Les graphiques montrent la relation linéaire ou non-linéaire entre Machine Failure et une autre variable.
- Le lissage rouge aide à identifier une tendance globale.

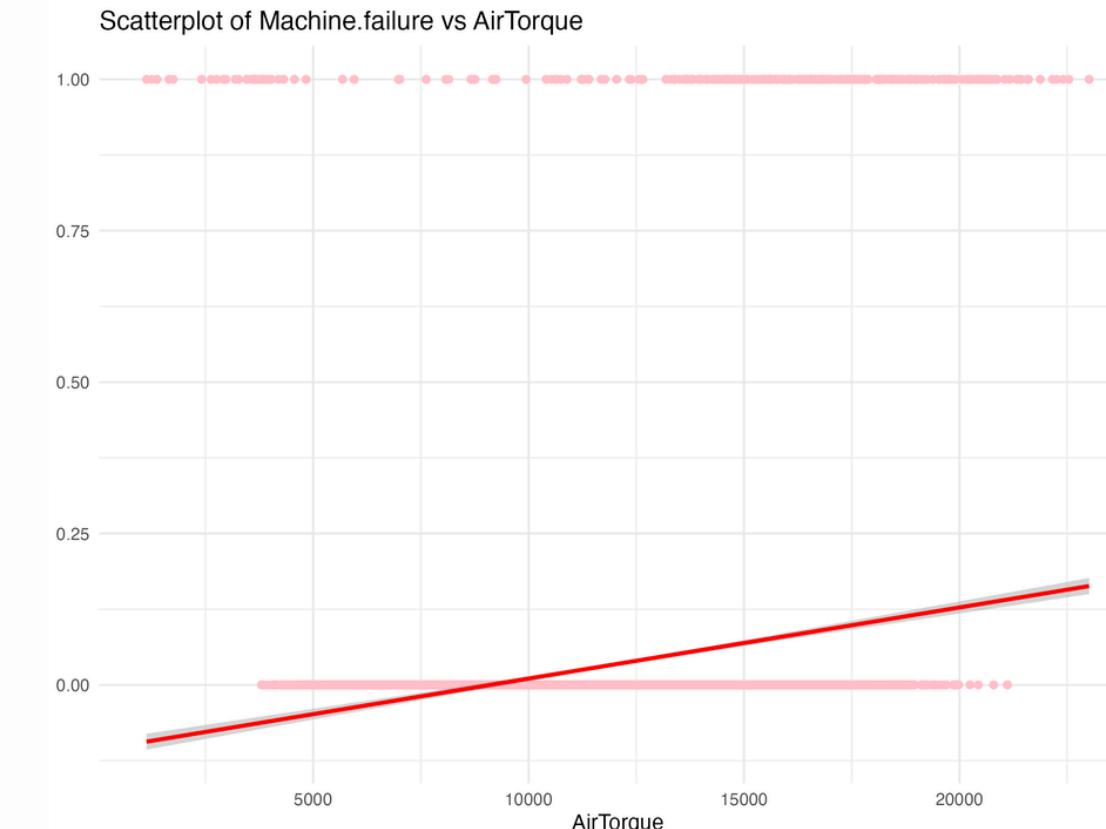
Analyse de la densité des points :

- La densité des points peut indiquer des zones où les relations sont plus fortes.

Variables avec une forte relation :



Variables avec une relation non claire :



Tests Statistiques

wilcox test :

Le test de Wilcoxon est utilisé pour comparer deux échantillons indépendants en utilisant leurs rangs. Contrairement au test t, il ne suppose pas de distribution normale, et il est souvent utilisé pour des données ordinaires ou lorsque les données sont asymétriques ou non paramétriques

Le test montre une différence significative entre les groupes, indiquant que les groupes sont différents en termes de distribution de Machine.failure.

Resultat :

```
Wilcoxon rank sum test with continuity correction

data: Machine.failure by FailureCategory
W = 3275079, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
```

p-value : < 2.2e-16 : La p-value extrêmement faible (< 2.2e-16) indique que les différences entre les deux groupes sont statistiquement significatives. Cela signifie que les groupes "High" et "Low" selon la variable Machine.failure ne sont pas égaux.



chi Carré test :

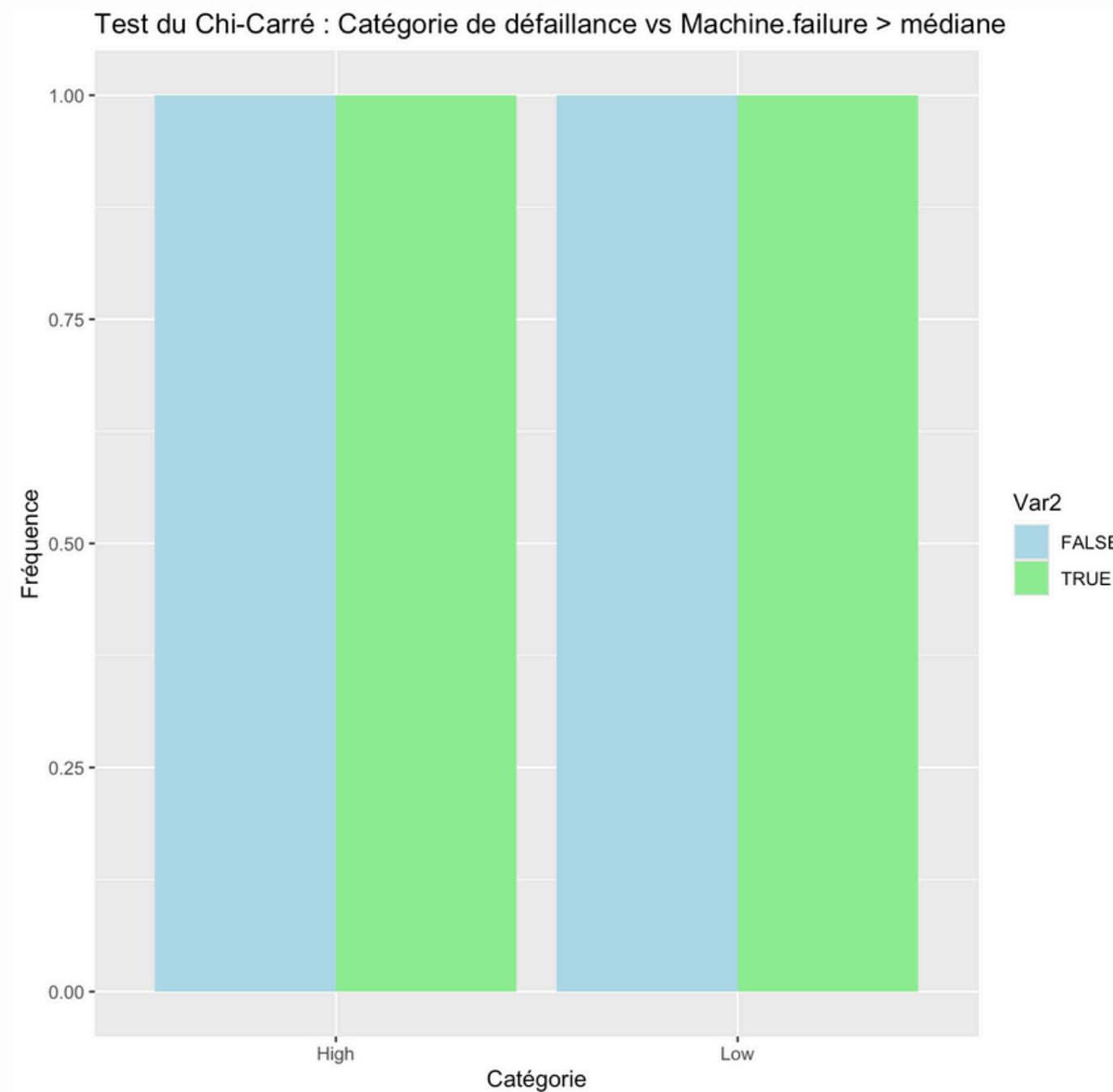
Il est utilisé pour évaluer si une association existe entre deux variables catégoriques. Dans votre cas, vous avez effectué un test entre deux catégories, probablement entre "High" et "Low" pour Machine.failure.

Résultat :

```
Pearson's Chi-squared test with Yates' continuity correction  
data: contingency_table  
X-squared = 9969.5, df = 1, p-value < 2.2e-16
```

p-value < 2.2e-16 : Une p-value extrêmement faible signifie que H₀ est **retenue**.

- montre que la différence entre les deux catégories est significative.
- Cela suggère qu'il existe une association forte entre les niveaux de Machine.failure transformés (logarithmiquement) et les catégories "High" vs "Low".



Test de Kruskal-Wallis :

Ce test non-paramétrique est utilisé pour comparer les différences entre plusieurs groupes indépendants sur une variable continue.

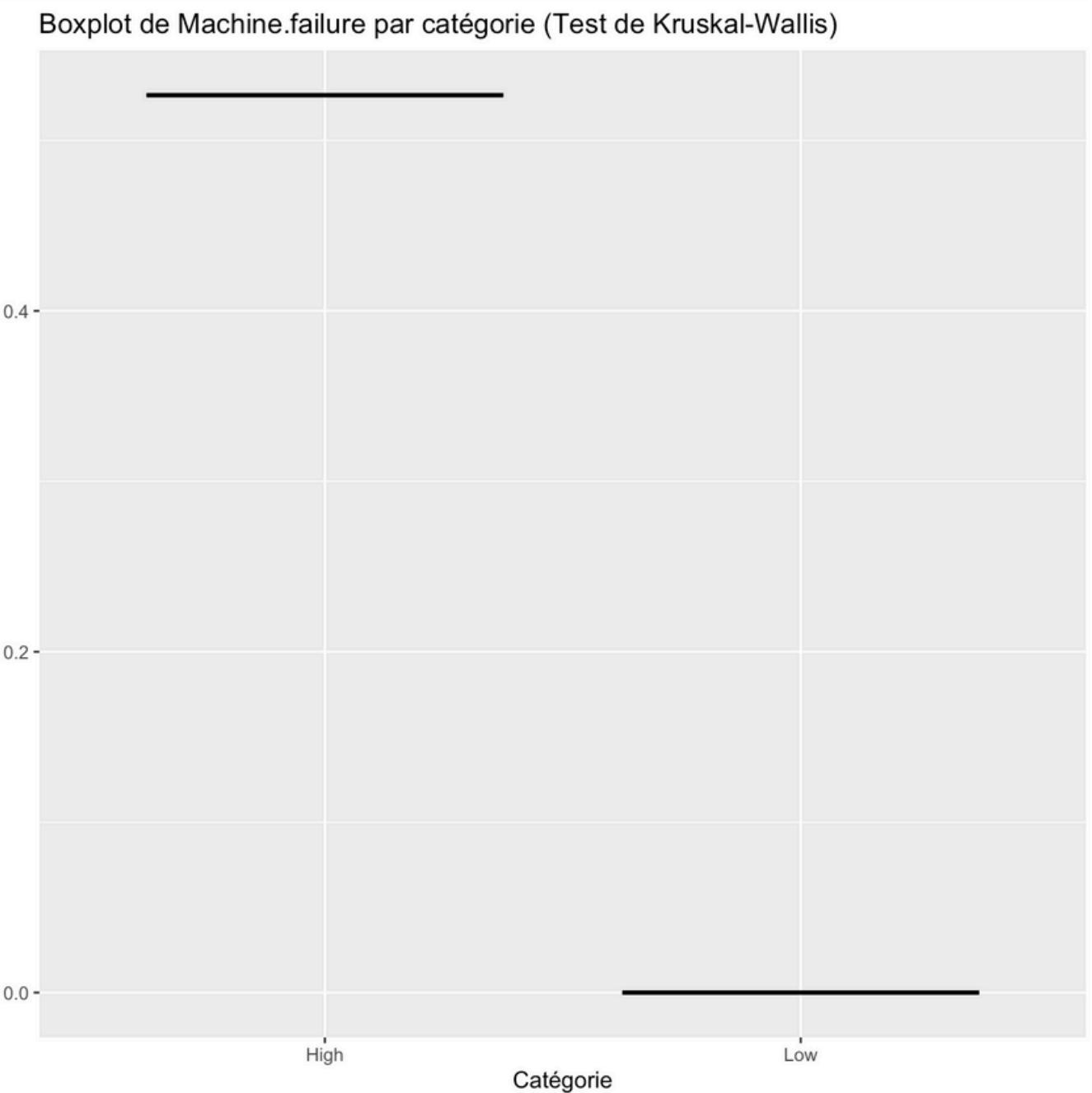
Résultat :

```
Kruskal-Wallis rank sum test

data: Machine.failure by FailureCategory
Kruskal-Wallis chi-squared = 9999, df = 1, p-value < 2.2e-16
```

p-value < 2.2e-16 : Une p-value extrêmement faible signifie que **HO est rejetée**. Cela indique que les **moyennes des deux groupes sont significativement**

- Le p-value extrêmement faible indique que les différences dans Machine.failure entre les catégories "High" et "Low" sont significatives.
- Cela suggère que les deux catégories définies pour Machine.failure (logarithmiquement transformées) présentent des distributions distinctes.



Regression Linéaire

stepwise model :

- L'approche stepwise permet de sélectionner automatiquement les variables en incluant ou excluant des variables à chaque étape, en fonction de leur contribution à l'ajustement du modèle.

Resultat :

- **Intercept** : Le modèle prévoit une valeur initiale de Machine.failure de -1.163 lorsque toutes les variables indépendantes sont nulles.
- **Air.temperature..K.** : Un coefficient positif (0.01872) signifie que lorsque la température de l'air augmente, Machine.failure augmente également.
- **Process.temperature..K.** : Un coefficient négatif (-0.01816) suggère qu'une augmentation de la température de traitement pourrait réduire Machine.failure.

Coefficients:

| | Estimate | Std. Error | t value | Pr(> t) | |
|-------------------------|------------|------------|---------|----------|--------|
| (Intercept) | -1.163e+00 | 4.147e-01 | -2.805 | 0.00504 | ** |
| Air.temperature..K. | 1.872e-02 | 1.961e-03 | 9.548 | < 2e-16 | *** |
| Process.temperature..K. | -1.816e-02 | 2.636e-03 | -6.889 | 6.03e-12 | *** |
| Rotational.speed..rpm. | 4.828e-04 | 2.195e-05 | 21.992 | < 2e-16 | *** |
| Torque..Nm. | 1.088e-02 | 3.919e-04 | 27.749 | < 2e-16 | *** |
| Tool.wear..min. | 3.038e-04 | 2.962e-05 | 10.259 | < 2e-16 | *** |
| --- | | | | | |
| Signif. codes: | 0 | *** | 0.001 | ** | 0.01 * |
| | 0.05 . | 0.1 . | 1 | | |

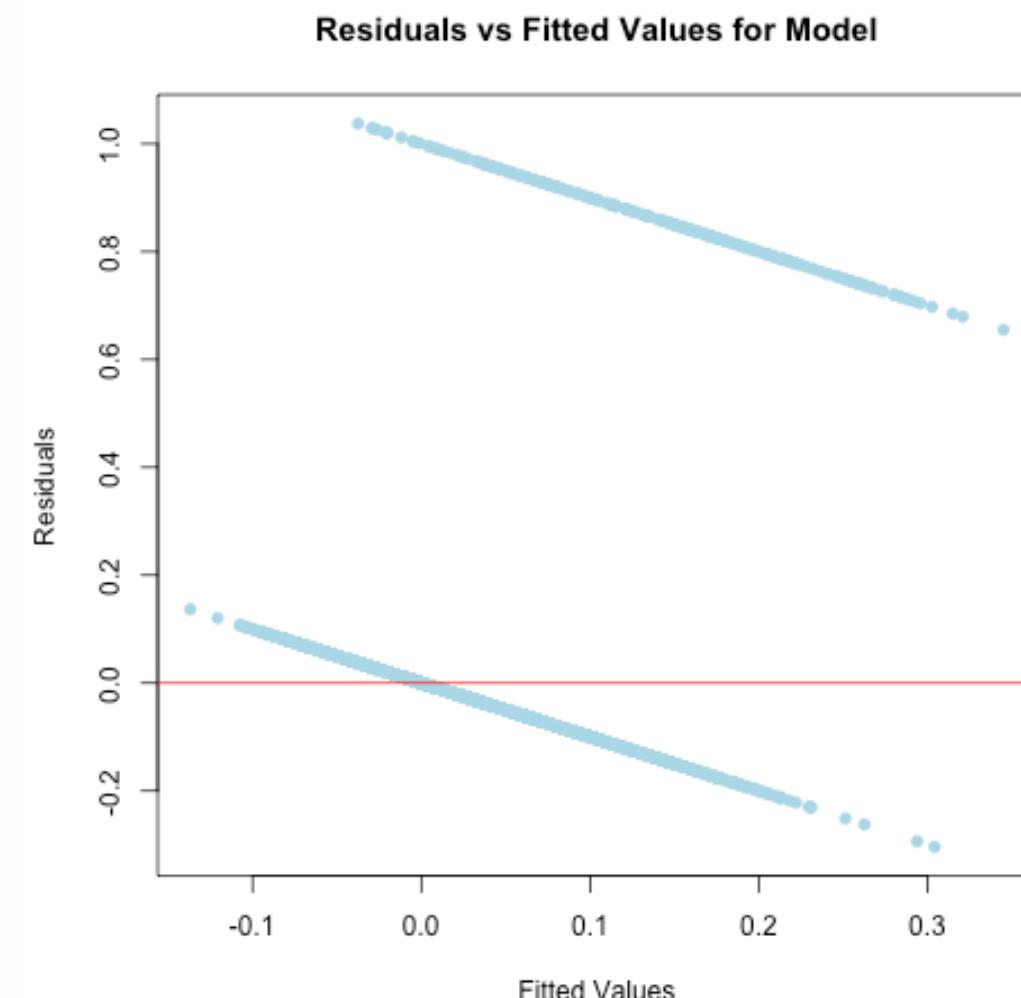
stepwise model :

Resultat :

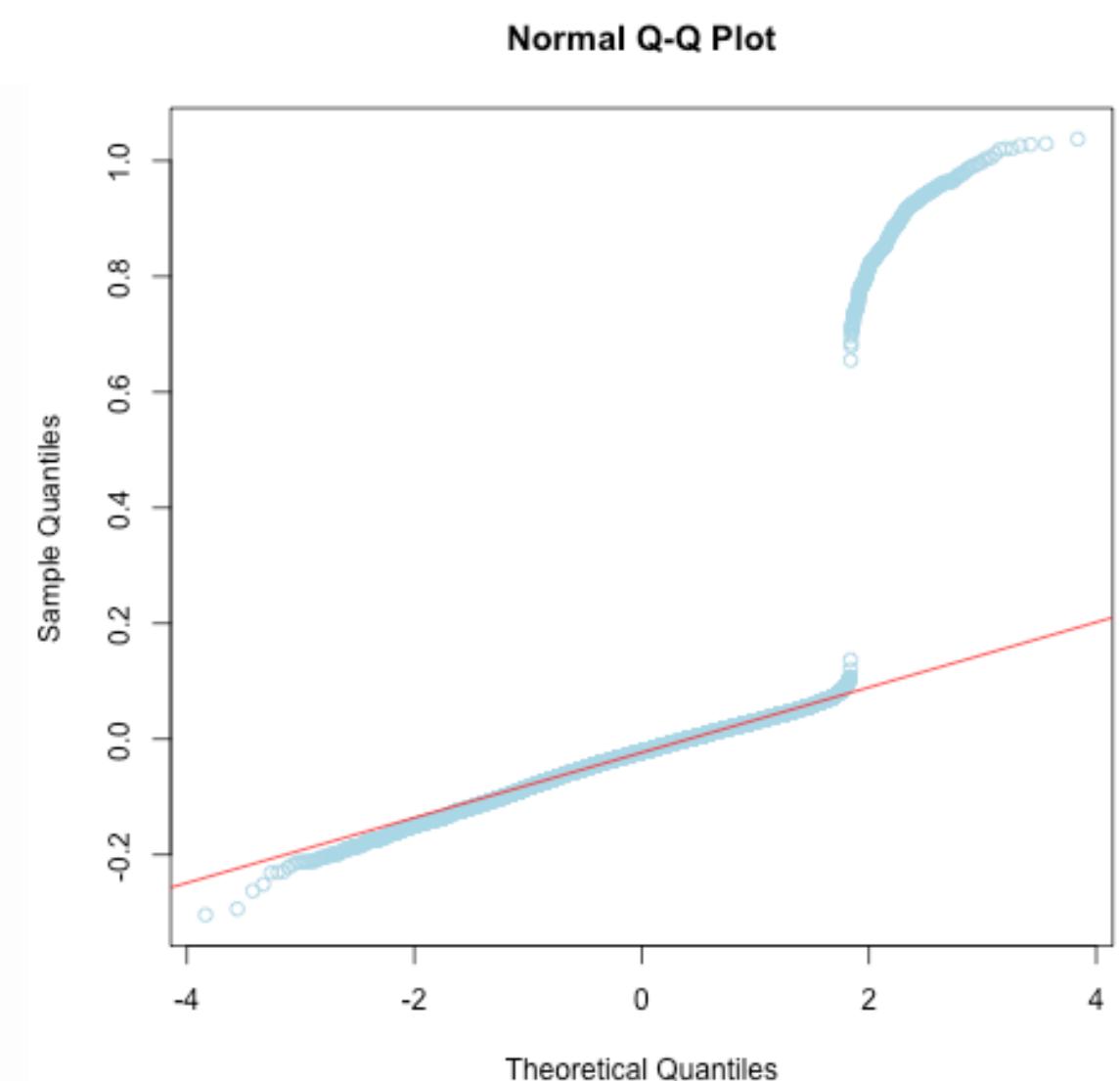
Residuals:

| Min | 1Q | Median | 3Q | Max |
|----------|----------|----------|---------|---------|
| -0.30405 | -0.06130 | -0.02222 | 0.01468 | 1.03747 |

Residual standard error: 0.1684 on 7994 degrees of freedom
Multiple R-squared: 0.1116, Adjusted R-squared: 0.1111



- **Le résidu :** 0.1684, ce qui montre que les résidus sont relativement faibles, indiquant un bon ajustement du modèle.
- **R-squared :** 0.1116 indique que 11.16% de la variation dans Machine.failure peut être expliquée par les variables du modèle, ce qui est modeste mais pertinent.
- **p-value < 0.05**, ce qui signifie qu'ils sont statistiquement significatifs à 95% de confiance.



Anova

Anova :

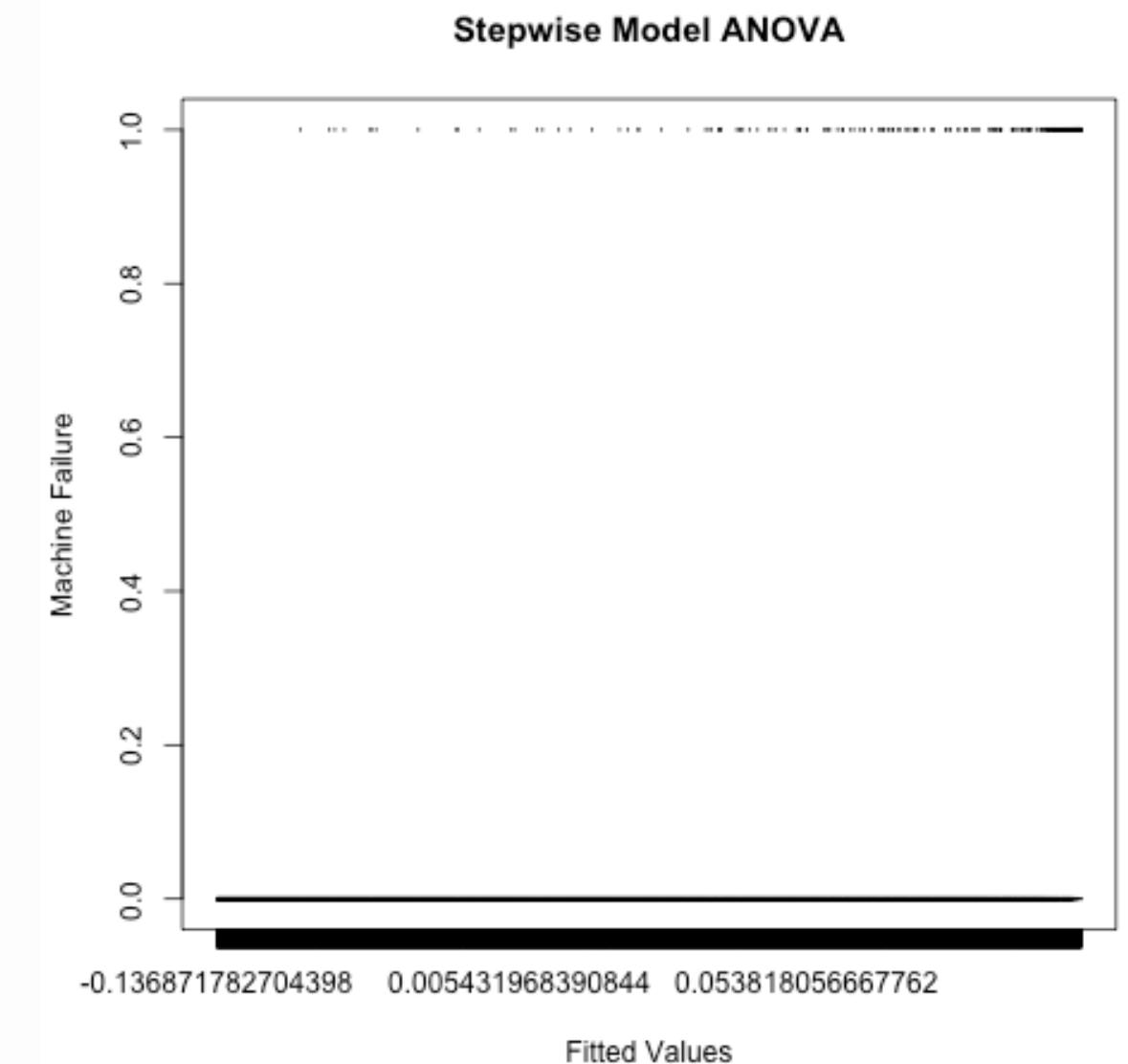
L'analyse de la variance (ANOVA) permet d'évaluer l'effet des différentes variables explicatives sur la variable cible (Machine Failure).

Réultat :

Response: Machine.failure

| | Df | Sum Sq | Mean Sq | F value | Pr(>F) | | | | | | |
|-------------------------|------|---------|---------|---------|---------------|-----|------|-----|-----|---|---|
| Air.temperature..K. | 1 | 1.614 | 1.6136 | 56.876 | 5.152e-14 *** | | | | | | |
| Process.temperature..K. | 1 | 1.422 | 1.4216 | 50.110 | 1.577e-12 *** | | | | | | |
| Rotational.speed..rpm. | 1 | 0.627 | 0.6273 | 22.112 | 2.615e-06 *** | | | | | | |
| Torque..Nm. | 1 | 21.847 | 21.8469 | 770.062 | < 2.2e-16 *** | | | | | | |
| Tool.wear..min. | 1 | 2.986 | 2.9857 | 105.241 | < 2.2e-16 *** | | | | | | |
| Residuals | 7994 | 226.793 | 0.0284 | | | | | | | | |
| --- | | | | | | | | | | | |
| Signif. codes: | 0 | '***' | 0.001 | '**' | 0.01 | '*' | 0.05 | '.' | 0.1 | ' | 1 |

- **Air.temperature..K.** : F-value = 56.88, p-value = 5.15e-14 (significatif)
- **Process.temperature..K.** : F-value = 50.11, p-value = 1.58e-12 (significatif)
- **Rotational.speed..rpm.** : F-value = 22.11, p-value = 2.62e-06 (significatif)
- **Torque..Nm.** : F-value = 770.06, p-value = <2.2e-16 (très significatif)
- **Tool.wear..min.** : F-value = 105.24, p-value = <2.2e-16 (très significatif)
- **Residuals** : indique la variance résiduelle (226.79), très faible comparée aux autres sources de variation.



Comparaison des modèles avec ANOVA :

L'analyse de la variance (ANOVA) permet d'évaluer l'effet des différentes variables explicatives sur la variable cible (Machine Failure).

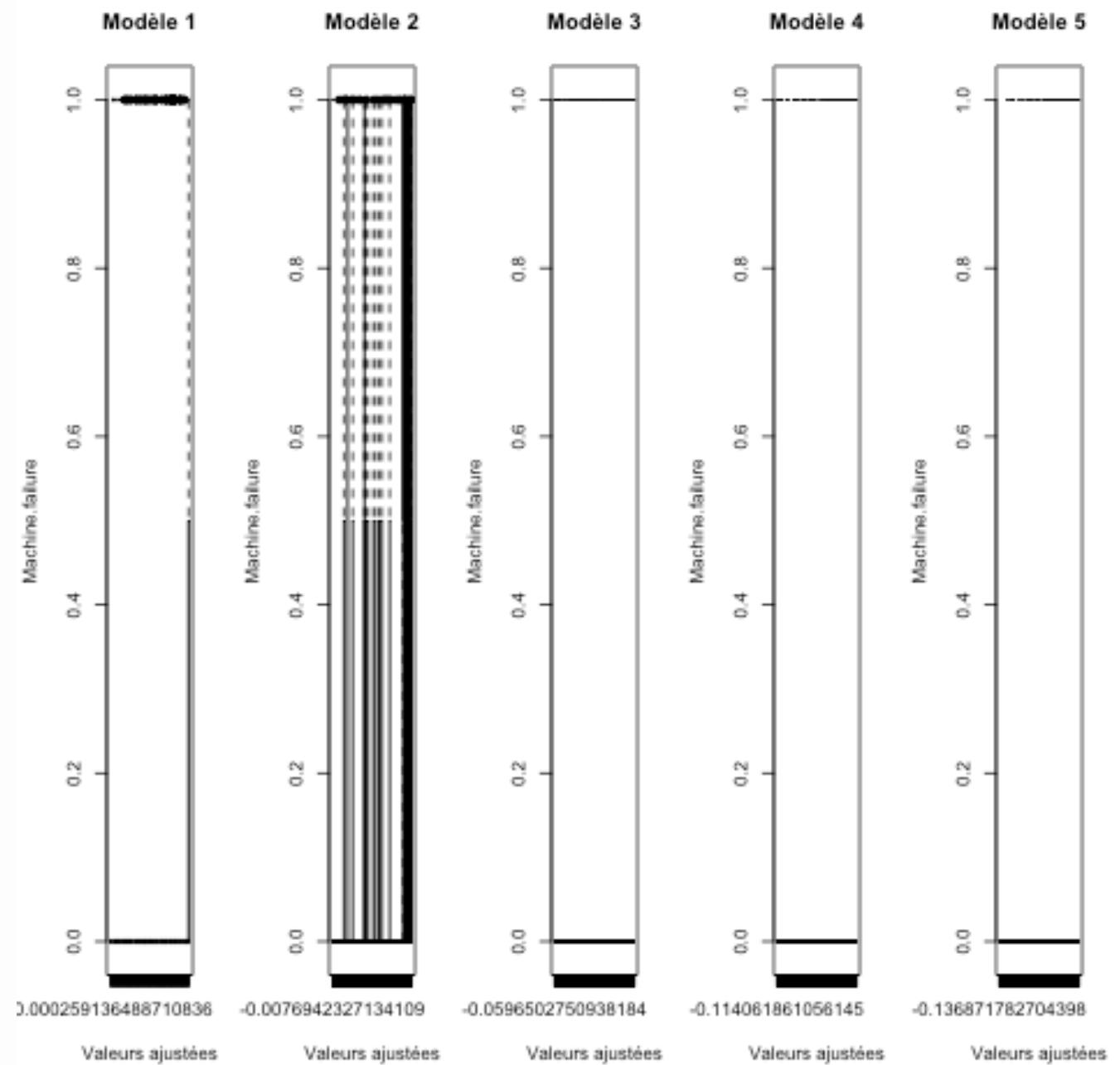
Réultat :

- **Df (Degrees of Freedom)** : Indique le nombre de prédicteurs inclus dans chaque modèle.
- **Sum of Sq** : La somme des carrés représente la variation expliquée par les prédicteurs.
- **F value** : La statistique F mesure l'efficacité du modèle.
- **Pr(>F)** : La p-value montre la significativité des différences entre les modèles.

```
Model 1: Machine.failure ~ Air.temperature..K.  
Model 2: Machine.failure ~ Air.temperature..K. + Process.temperature..K.  
Model 3: Machine.failure ~ Air.temperature..K. + Process.temperature..K. +  
    Rotational.speed..rpm.  
Model 4: Machine.failure ~ Air.temperature..K. + Process.temperature..K. +  
    Rotational.speed..rpm. + Torque..Nm.  
Model 5: Machine.failure ~ Air.temperature..K. + Process.temperature..K. +  
    Rotational.speed..rpm. + Torque..Nm. + Tool.wear..min.
```

| | Res.Df | RSS | Df | Sum of Sq | F | Pr(>F) |
|-----|--------|--------|----|-----------|---------|---------------|
| 1 | 7998 | 253.67 | | | | |
| 2 | 7997 | 252.25 | 1 | 1.4216 | 50.110 | 1.577e-12 *** |
| 3 | 7996 | 251.62 | 1 | 0.6273 | 22.112 | 2.615e-06 *** |
| 4 | 7995 | 229.78 | 1 | 21.8469 | 770.062 | < 2.2e-16 *** |
| 5 | 7994 | 226.79 | 1 | 2.9857 | 105.241 | < 2.2e-16 *** |
| --- | | | | | | |

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1



=> l'ajout progressif des prédicteurs améliore les performances du modèle.



Conclusion

Conclusion



L'analyse statistique appliquée à ces deux jeux de données a permis d'explorer les relations entre les variables et les cibles prédictives respectives.

1. Analyse des émissions de gaz pour les turbines à gaz :

- L'utilisation de modèles ANOVA et de régression a aidé à identifier les contributions significatives des variables environnementales aux émissions de CO et de NOx.
- Ces modèles permettent une meilleure compréhension des facteurs influençant les émissions.

2. Maintenance prédictive pour les défaillances de machines (AI4I 2020) :

- La sélection pas à pas des variables a permis d'améliorer les prédictions de défaillance des machines en intégrant progressivement des variables pertinentes.
- Les résultats montrent une amélioration significative de la précision des modèles .



Merci de votre attention

