ISEP – Ecole d'ingénieurs du numérique de Paris

ANALYSE DE LA CONSOMMATION ÉNERGETIQUE À TÉTOUAN

Projet de Data Science

Fait par Ange Nolwen DJUISSI KENMOE

Janvier 2025

Année académique : 2024-2025

Table des matières

Introduction

Dans un monde confronté à des enjeux énergétiques et environnementaux majeurs, comprendre les dynamiques de consommation d'électricité devient un levier essentiel pour anticiper, optimiser et mieux gérer les ressources. Or, derrière chaque kilowatt consommé, se cachent des facteurs complexes : conditions climatiques, zones géographiques, comportements humains… et surtout, des données.

C'est dans ce contexte que s'inscrit ce projet de Data Science, centré sur l'analyse de la consommation énergétique de la ville de **Tétouan** au Maroc. Grâce à un jeu de données riche et structuré, combinant variables météorologiques et consommation par zone, nous avons mobilisé deux outils statistiques puissants :

- **L'Analyse en Composantes Principales (PCA)** pour révéler les structures sous-jacentes des données et réduire leur dimensionnalité ;
- **La régression linéaire** (simple et multiple) pour modéliser les relations entre les variables et **prédire avec précision** la consommation d'une zone à partir de facteurs explicatifs.

Au fil des étapes, ce projet nous a permis non seulement d'approfondir nos compétences en traitement de données, mais aussi de démontrer concrètement comment des techniques statistiques peuvent éclairer des problématiques énergétiques bien réelles.

1- Presentation du jeu de données

- Source : UCI Machine Learning Repository

```python
#chargement du dataset
data = pd.read_csv("Tetuan-PC.csv")

#affichage des premières lignes
print(data.head())

#résumé des dimensions
print("Nombre d'observations :", data.shape[0])
print("Nombre de variables :", data.shape[1])

#vérification des valeurs manquantes
missing_values = data.isnull().sum()
print("Valeurs manquantes :\n", missing_values)

#résumé statistique
print(data.describe())
```

Graphe 1 : Code python utilisé pour recueillir les information sur les données

```
             DateTime  Temperature  Humidity  WindSpeed  GenDiffFlows  DiffFlows      PCZone1      PCZone2      PCZone3
24625  6/21/2017 0:00        21.86      77.0      0.081         0.073      0.111  44554.17219  24750.93555  26980.43077
24626  6/21/2017 0:10        21.91      76.0      0.077         0.077      0.096  44249.00662  24612.47401  26631.87692
24627  6/21/2017 0:20        21.96      75.5      0.081         0.055      0.096  43753.11258  24339.29314  26141.53846
24628  6/21/2017 0:30        21.83      75.1      0.078         0.062      0.163  43079.20530  24208.31601  26052.92308
24629  6/21/2017 0:40        21.60      75.8      0.075         0.048      0.145  42634.17219  23927.65073  25887.50769
Nombre d'observations : 13248
Nombre de variables : 9
Valeurs manquantes :
 DateTime        0
Temperature     0
Humidity        0
WindSpeed       0
GenDiffFlows    0
DiffFlows       0
PCZone1         0
PCZone2         0
PCZone3         0
dtype: int64
        Temperature      Humidity     WindSpeed  GenDiffFlows     DiffFlows       PCZone1       PCZone2       PCZone3
count  13248.000000  13248.000000  13248.000000  13248.000000  13248.000000  13248.000000  13248.000000  13248.000000
mean      25.353431     63.239117      4.000284    254.436238     68.161448  35272.400777  22976.430993  23544.040816
std        3.687087     18.150509      1.857605    310.646997     91.721217   7073.784592   5239.586748   7667.598488
min       13.990000     11.340000      0.050000      0.018000      0.019000  18283.684790  10624.116420   8189.908069
25%       22.750000     50.260000      4.904000      0.088000      0.122000  29432.896780  18866.737060  17001.338912
50%       25.140000     65.050000      4.910000     61.540000     35.030000  35604.955750  22900.105600  23051.926500
75%       27.430000     78.300000      4.919000    526.100000    100.200000  40238.628617  27002.361105  28540.920500
max       40.010000     93.800000      4.998000    978.000000    558.700000  52204.395120  37408.860760  47598.326360
PS C:\Users\angen\Desktop\SUJET ACTUEL DS>
```

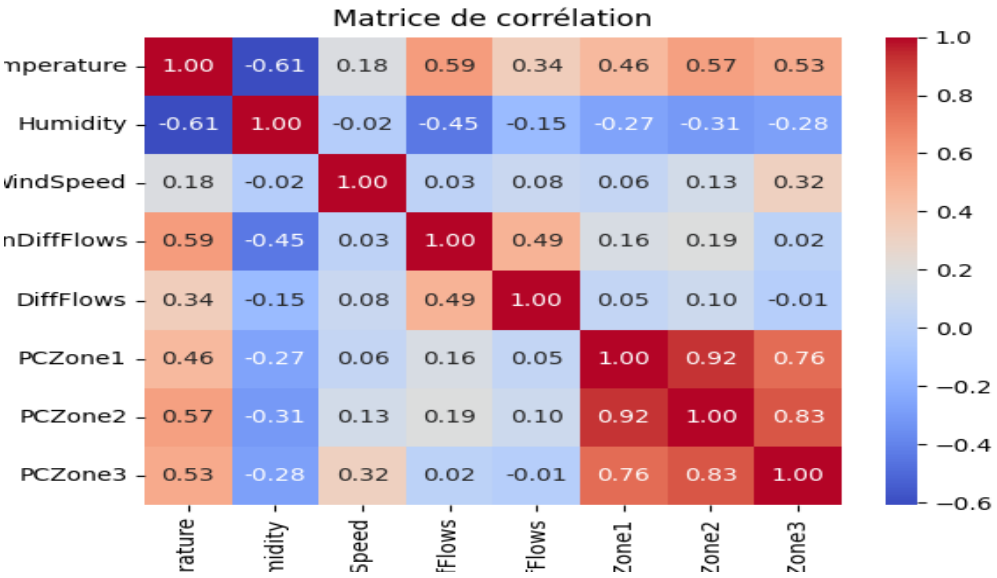
Graphe 2 : Résultats du code après exécution

1- Analyse descriptive des données

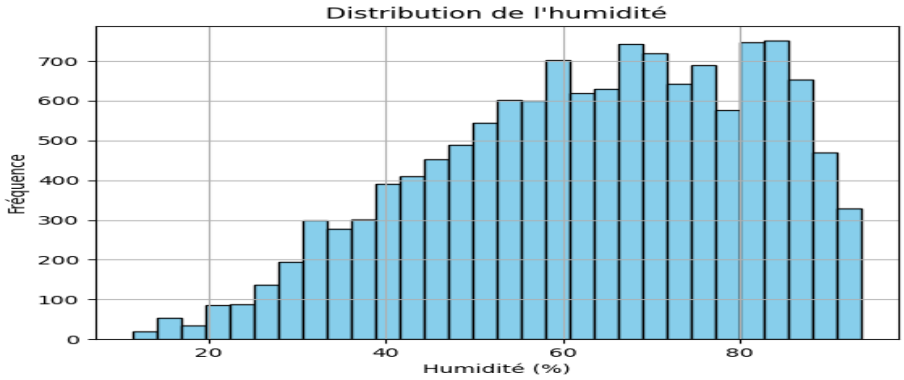
- Observations : 13,248
- Variables : 9
- Valeurs manquantes : Aucune valeur manquante détectée dans les colonnes, ce qui est idéal pour l'analyse.
- Résumé statistique :
- Temperature :
- Moyenne : 25.35°C
- Min : 3.99°C, Max : 40.01°C
- Humidity :
    - Moyenne : 63.24%
    - Min : 11.34%, Max : 93.80%
- WindSpeed :
    - Moyenne : 4.00 km/h
    - Min : 0.05 km/h, Max : 4.99 km/h
- Consommation énergétique (PCZone1, PCZone2, PCZone3) :
    - Moyenne : 35,272 KW (Zone 1), 22,976 KW (Zone 2), 23,544 KW (Zone 3)

- Max : 52,204 KW (Zone 1)

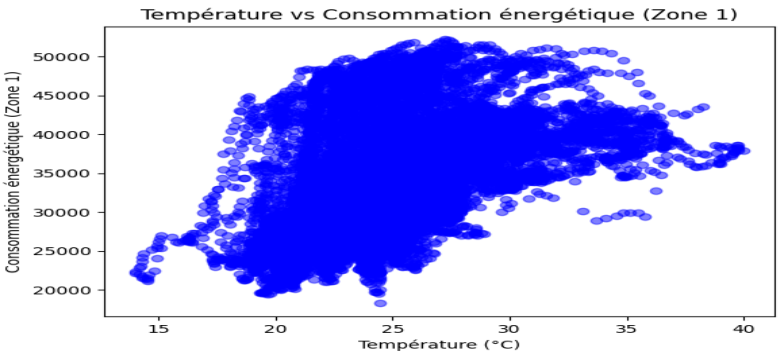
Graphes représentatifs



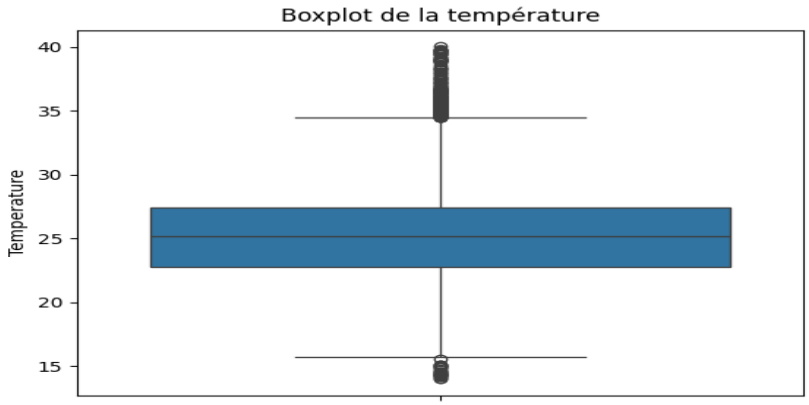
Graphe 3 : Matrice de corrélation



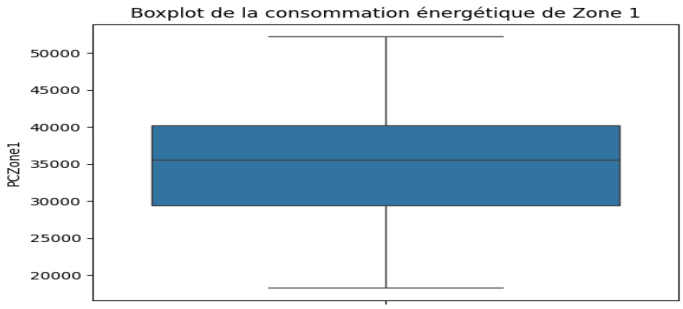
Graphe 4 : Distribution de l'humidité



Graphe 5 : Température Vs Consommation énergétique en Zone 1



Graphe 6 : Boxplot de la température



Graphe 7 : Boxplot de la consommation énergétique en Zone 1

<u>Interprétation des graphiques</u>

Boxplot de la consommation énergétique (Zone 1)

- La consommation énergétique de la Zone 1 varie principalement entre 30,000 KW et 40,000 KW, avec une médiane autour de 35,000 KW.

- Aucune valeur aberrante significative n'est visible, ce qui suggère une distribution relativement homogène.

Boxplot de la temperature

- La température varie entre 15°C et 40°C, avec une médiane proche de 25°C.
- Des valeurs aberrantes sont observées aux extrêmes bas et hauts (inférieures à 15°C et supérieures à 35°C). Cela peut nécessiter un traitement si ces valeurs influencent l'analyse.

Histogramme de l'humidité

- La répartition de l'humidité est asymétrique, avec une concentration autour de 60% à 80%.
- Cela montre que l'humidité reste généralement élevée dans la zone étudiée.

Scatter plot : Température vs Consommation énergétique (Zone 1)

- Une tendance positive est visible : une augmentation de la température correspond généralement à une augmentation de la consommation énergétique dans la Zone 1.
- Cela confirme que la température est un facteur clé influençant la consommation énergétique.

Matrice de correlation

- La **température** est modérément corrélée avec la consommation énergétique dans les zones (corrélations positives : 0.57 pour la Zone 1, 0.53 pour la Zone 3).
- L'**humidité** a une corrélation légèrement négative avec les consommations, mais elle est faible.
- Les consommations des zones sont fortement corrélées entre elles (exemple : Zone 1 et Zone 2 ont une corrélation de 0.92), ce qui indique des similitudes dans leurs variations.

2- Analyse en Composantes Principales (PCA)

- **Théorie**

Questions :

- **Si deux variables sont parfaitement corrélées, est-il pertinent de les inclure dans une PCA ?**

Réponse :

Lorsqu'elles sont **parfaitement corrélées**, elles contiennent la même information (redondance). Dans une PCA, l'objectif est de réduire la dimensionnalité en retenant l'information la plus importante. Inclure deux variables parfaitement corrélées signifie que l'une d'elles ne contribuera pas de manière significative à la variance supplémentaire.

Dans ce cas, il est préférable de conserver une seule des deux variables.

- **Et si elles sont complètement non corrélées ?**

Si elles sont **complètement non corrélées**, elles apportent des informations différentes. Une PCA pourra inclure les deux pour capturer l'ensemble de la variance qu'elles expliquent.

- **Application pratique avec Python**

- **Calcul variance**

```python
#chargement du dataset
data = pd.read_csv("Tetuan-PC.csv")

#filtrage_variable
features = ['Temperature', 'Humidity', 'WindSpeed', 'PCZone1', 'PCZone2', 'PCZone3']
data_pca = data[features]

#calcul_variance
print("Variance de chaque variable :\n", data_pca.var())
```

Graphe 8 : Code python utilisé pour calculer la variance

```
win32-x64\bundled\libs\debugpy\adapter/../..\debugpy\launcher' '59184' '--' 'C:\Users\angen\Desktop\S
Variance de chaque variable :
 Temperature    1.359461e+01
Humidity       3.294410e+02
WindSpeed      3.450695e+00
PCZone1        5.003843e+07
PCZone2        2.745327e+07
PCZone3        5.879207e+07
dtype: float64
PS C:\Users\angen\Desktop\SUJET ACTUEL DS>
```

Graphe 9 : Résultats obtenus

Analyse des variances :

- *Température* : Variance = 13.59

Variation relativement faible.

- *Humidité* : Variance = 329.44

Variation modérée, supérieure à celle de la température.

- *WindSpeed* : Variance = 34.57

Variation faible, proche de celle de la température.

- *PCZone1, PCZone2, PCZone3* :

Variances = (5.08×10^7) ; (2.74×10^7) ; (5.87×10^7)

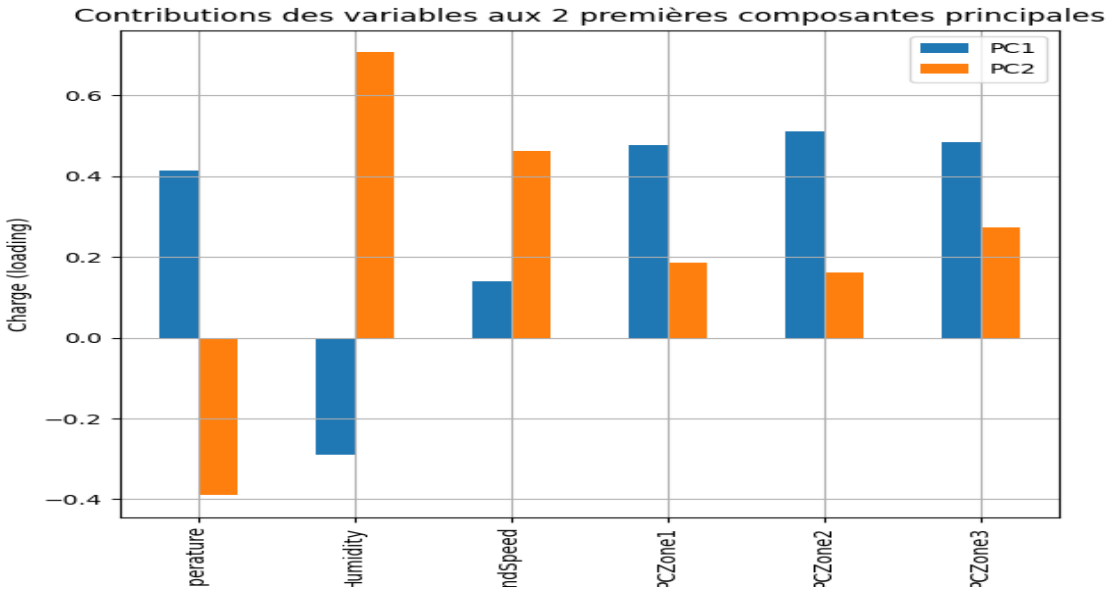
Ces valeurs sont des ordres de grandeur bien plus grands, reflétant des échelles beaucoup plus élevées (en KW).

Justification de la standardisation :

- Les valeurs de variance montrent que les variables comme PCZone1, PCZone2, et PCZone3 sont à des échelles beaucoup plus grandes que celles des variables comme Temperature ou WindSpeed.
- La PCA est sensible aux échelles, car elle est basée sur la variance. Les variables à grande échelle (comme PCZone1) domineraient le calcul des composantes principales.

Conclusion : Il est nécessaire de standardiser les données pour s'assurer que chaque variable contribue de manière égale à l'analyse.

- **Standardisation**



Graphe 10 : Contribution des variables aux deux premières composantes principales

Interprétation des deux premières composantes principales :

Première composante principale (PC1) :

- Variables dominantes : PCZone1, PCZone2, PCZone3.
- Signification : PC1 représente la variance énergétique globale des trois zones.

Elle est principalement influencée par la consommation énergétique, avec une contribution minimale des variables climatiques.

Deuxième composante principale (PC2) :

- Variables dominantes : Temperature (positive) et Humidity (négative).
- Signification : PC2 capture la variance liée aux conditions climatiques.

L'opposition entre Temperature et Humidity indique qu'elles varient de manière inverse.

Relation entre PC1 et PC2 :

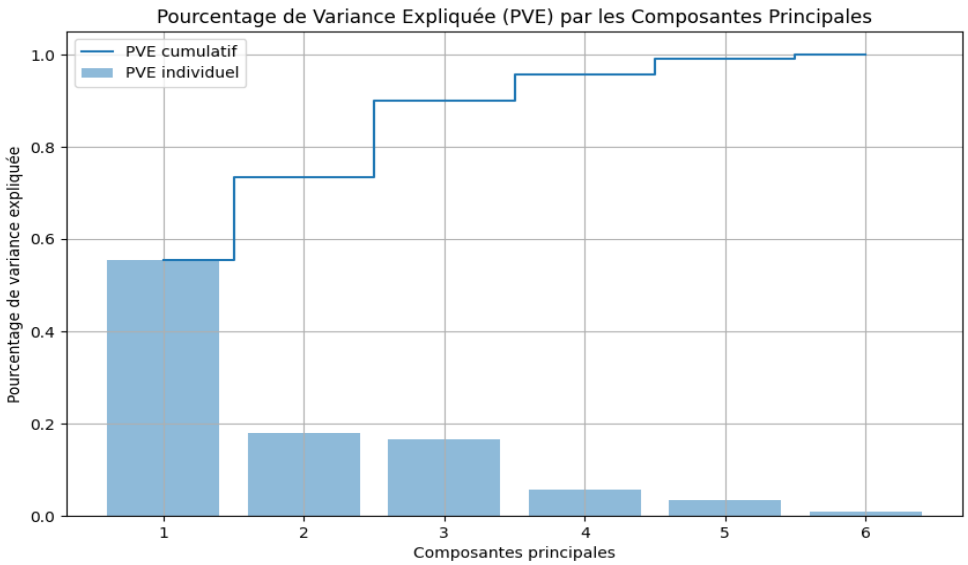
PC1 et PC2 sont indépendantes, capturant des aspects différents des données :

- PC1 : Variance énergétique.
- PC2 : Variance climatique.

Conclusion :

- PC1 est principalement influencée par la consommation énergétique.
- PC2 est guidée par les conditions climatiques, avec une interaction entre température et humidité

- **Pourcentage de Variance Expliquée (PVE)**



Graphe 11 : Pourcentage de Variance Expliqué (PVE) par les composantes principales

Combien de composantes conserver ?

Deux composantes principales (PC1 et PC2) suffisent : Elles expliquent plus de 80%-90% de la variance totale.

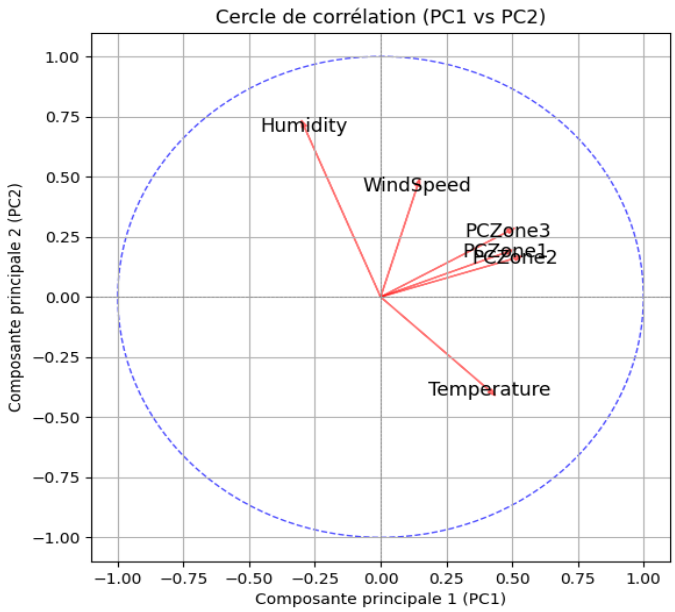
Cela permet de réduire la dimensionnalité tout en conservant l'essentiel de l'information.

Pourquoi ?

Les deux premières composantes capturent l'essentiel des tendances dans les données.

Les composantes supplémentaires n'ajoutent que très peu d'information et pourraient compliquer l'interprétation.

- **Cercle de correlation**



Graphe 12 : Cercle de corrélation PC1 vs PC2

Interprétation du cercle de corrélation

Le cercle de corrélation montre les contributions des variables d'origine aux deux premières composantes principales (PC1 et PC2) et leurs relations.

Voici l'analyse détaillée :

❖ Contribution des variables à PC1 et PC2

Variables énergétiques (PCZone1, PCZone2, PCZone3) : Les flèches de PCZone1, PCZone2, et PCZone3 sont proches et pointent dans une direction similaire, alignée sur PC1.

Interprétation : Ces variables sont fortement corrélées entre elles, elles dominent PC1, ce qui signifie que cette composante capture principalement la variance liée à la consommation énergétique.

❖ Variables climatiques (Temperature, Humidity, WindSpeed) :

- **Temperature** :

Sa flèche pointe dans une direction opposée à celle de Humidity.

Elle contribue principalement à PC2.

Cela indique qu'une augmentation de la température est souvent associée à une diminution de l'humidité (corrélation négative)

- **Humidity :**

Sa flèche est alignée positivement avec PC2, et elle a une forte contribution à cette composante.

- **WindSpeed :**

Sa flèche est plus courte, indiquant une contribution faible aux deux composantes principales.

❖ Relation entre les variables

Variables corrélées positivement :

PCZone1, PCZone2, et PCZone3 ont des flèches proches et presque superposées, indiquant une forte corrélation positive entre elles.

Variables corrélées négativement :

Temperature et Humidity pointent dans des directions opposées, indiquant une corrélation négative.

Variables faiblement corrélées :

WindSpeed est presque orthogonale aux autres variables, ce qui suggère qu'elle n'est pas significativement corrélée aux variables énergétiques ou climatiques.

❖ Résumé des composantes principales

PC1 : Capturée principalement par les variables énergétiques (PCZone1, PCZone2, PCZone3), elle explique la variance énergétique globale.

PC2 : Dominée par les variables climatiques (Temperature et Humidity), elle explique les variations liées aux conditions climatiques.

Conclusion

Le cercle de corrélation confirme que :

- Les trois variables énergétiques sont très corrélées et dominent la première composante.
- Les variables climatiques influencent principalement la seconde composante, avec des interactions spécifiques comme l'opposition entre température et humidité.

3- Regression linéaire

Question Théorique : Régression Linéaire et R^2

Supposons que nous ajustons un modèle de régression linéaire pour expliquer Y comme une fonction linéaire de deux variables X1 et X2.

Nous notons R^2 le coefficient de détermination associé.

Interprétation de R^2 :

R^2 mesure la proportion de la variance de la variable dépendante (Y) qui peut être prédite à partir des variables indépendantes (X1 et X2).

Il quantifie la puissance explicative du modèle de régression. Les valeurs de R^2 sont comprises entre 0 et 1 :

- $R^2 = 0$: Aucune variance de Y n'est expliquée par le modèle.
- $R^2 = 1$: Toute la variance de Y est expliquée par le modèle.

Quelle est la plage des valeurs possibles pour R^2 ?

La plage de R^2 est [0, 1]. Il ne peut pas être négatif dans une régression linéaire standard, car il représente une proportion carrée de variance.

Si nous notons r1 et r2 les coefficients de corrélation entre X1 et Y et entre X2 et Y respectivement, quelle est la relation entre R^2, r1 et r2 ?

- Si X1 et X2 ne sont pas corrélés, $R^2 = r1^2 + r2^2$. Dans ce cas, la variance expliquée par X1 et X2 s'ajoute de manière linéaire.
- Si X1 et X2 sont corrélés, $R^2 \leq r1^2 + r2^2$, car la variance partagée entre X1 et X2 réduit leurs contributions uniques pour expliquer Y.

a) Regression linéaire simple

```python
# Exclure les colonnes non numériques
numeric_data = data.select_dtypes(include=['float64', 'int64'])

# corrélations avec PCZone1
correlations = numeric_data.corr()['PCZone1'].sort_values(ascending=False)
print("Corrélations avec PCZone1 :\n", correlations)
```

Graphe 13 : Code python pour avoir la correlation avec PC1

```
Corrélations avec PCZone1 :
 PCZone1         1.000000
PCZone2          0.923311
PCZone3          0.759404
Temperature      0.456112
GenDiffFlows     0.159304
WindSpeed        0.055393
DiffFlows        0.045006
Humidity        -0.266929
Name: PCZone1, dtype: float64
```

Graphe 14 : Résultat du code

Variable la plus corrélée :

PCZone2 a la corrélation la plus forte avec PCZone1 (r = 0.923311).

Cela indique une relation linéaire très forte et positive entre ces deux variables.

Interprétation de la corrélation avec PCZone2 :

- La forte corrélation (r ≈ 0.92) montre que les consommations énergétiques dans Zone 1 et Zone 2 évoluent de manière similaire.

Cela peut être dû à :

- Une interdépendance entre les zones (par exemple, elles partagent des caractéristiques similaires ou une infrastructure énergétique commune).
- Un comportement énergétique synchronisé (par exemple, la consommation énergétique est influencée par des facteurs communs, comme les conditions climatiques ou les habitudes des utilisateurs).

Autres variables significatives :

- PCZone3 : r = 0.759004, montre également une forte corrélation avec PCZone1, mais légèrement plus faible que celle de PCZone2.
- Temperature : r = 0.456112, une corrélation modérée, ce qui suggère que la température a un effet non négligeable sur la consommation énergétique.

Variables faiblement corrélées ou négativement corrélées :

- Humidity : r = -0.266929, indique une faible corrélation négative. Cela signifie qu'une augmentation de l'humidité pourrait être associée à une légère diminution de la consommation énergétique dans PCZone1.
- WindSpeed, DiffFlows, et GenDiffFlows ont des corrélations proches de zéro, ce qui montre qu'elles n'ont qu'une contribution négligeable.

➢ $Y = \beta_0 + \beta_1 X$

- **Interprétation du coefficient β1=1.24653**

Le coefficient β1 représente la pente de la droite de régression. Il indique comment la variable cible PCZone1 (consommation énergétique de la Zone 1) change en moyenne lorsque la variable explicative PCZone2 (consommation énergétique de la Zone 2) augmente d'une unité.

Interprétation :

Valeur de β1=1.24653 : Pour chaque augmentation de 1 unité de PCZone2, la consommation énergétique de PCZone1 augmente en moyenne de 1.24653 unités.

Cela montre une relation linéaire positive et forte entre les deux zones, ce qui est cohérent avec leur forte corrélation (r≈0.92).

Implications :

La consommation énergétique dans la Zone 1 est fortement influencée par celle de la Zone 2. Cela pourrait indiquer une dépendance ou des facteurs communs influençant les deux zones (par exemple, une infrastructure énergétique partagée ou des comportements similaires).

- **Intervalle de Confiance pour le Coefficient β1**

Étape 1 : Formule générale pour un intervalle de confiance à (1 - α) pour β_1
Un intervalle de confiance (IC) pour β1, le coefficient de pente dans une régression linéaire, est calculé comme suit : $\boldsymbol{\beta_1 \pm t_{\alpha/2} * SE(\beta_1)}$

Explications :

- β_1: Estimation du coefficient de pente (dans ce cas, β1 = 1.24653).

- $t_{\alpha/2}$: Valeur critique de la distribution t de Student avec n - 2 degrés de liberté
- SE(β_1) Erreur standard associée à β1, calculée à partir de l'estimation des erreurs résiduelles.

Notations :

- n : Nombre total d'observations.
- df : Degrés de liberté, calculés comme n - 2 (nombre de prédicteurs = 1 dans la régression linéaire simple).
- $t_{\alpha/2}$: Obtenu à partir de la table de la loi t de Student pour un niveau de confiance spécifique (par exemple, 95 %).

```python
n = len(X)
df = n - 2

# Erreur standard de β1
# Calculer les prédictions du modèle
Y_pred = model.predict(X)
residuals = Y - Y_pred
s_squared = np.sum(residuals**2) / df  # Variance des résidus
X_std = (X - np.mean(X))**2
SE_beta1 = np.sqrt(s_squared / np.sum(X_std))

# Valeur critique t pour un intervalle à 95%
t_crit = stats.t.ppf(1 - 0.025, df)

# Calcul de l'intervalle de confiance
lower_bound = beta_1 - t_crit * SE_beta1
upper_bound = beta_1 + t_crit * SE_beta1
print(f"Intervalle de confiance à 95% pour β1 : [{lower_bound}, {upper_bound}]")
```

Graphe 15 : Code python pour calculer l'intervalle de confiance

```
Intervalle de confiance à 95% pour β1 : [PCZone2    1.237699
dtype: float64, PCZone2    1.255361
dtype: float64]
```

Graphe 16 : Résultat du code

Interprétation de l'Intervalle de Confiance pour β1

L'intervalle de confiance à 95 % pour le coefficient β1 est donné par :

[1.237699, 1.253361]

Estimation de β_1 :
Cet intervalle signifie que, avec un niveau de confiance de 95 %, la vraie valeur de β_1 (la pente de la régression linéaire) se trouve dans cette plage.

Signification de β_1 :

Rappel : β_1 représente l'augmentation moyenne de PCZone1 pour chaque unité d'augmentation de PCZone2.

Une estimation de β_1 entre 1.237699 et 1.253361 indique que pour chaque unité d'augmentation de la consommation énergétique dans PCZone2, la consommation dans PCZone1 augmente en moyenne d'environ 1.24 à 1.25 unités.

Signification statistique :

Comme l'intervalle de confiance ne contient pas 0, cela confirme que β1 est statistiquement significatif au niveau de 95 %.

Cela signifie que PCZone2 a un impact significatif sur PCZone1 dans le modèle.

Conclusion :

L'intervalle montre que PCZone2 est un prédicteur robuste et significatif pour expliquer PCZone1, avec une relation linéaire forte et positive.

- **Test d'Hypothèse : Coefficient β_1 (Pente Nulle)**

Étape 1 : Formulation des hypotheses

- *Hypothèse nulle (H_0) : $\beta_1 = 0$*

Le prédicteur PCZone2 n'a pas d'impact significatif sur la consommation énergétique en PCZone1.

- Hypothèse alternative (H_1) : $\beta_1 \neq 0$

Le prédicteur PCZone2 a un impact significatif sur PCZone1.

Étape 2 : Calcul de la statistique de test

La statistique t est donnée par : t = β_1 / SE(β_1)

β_1 : Coefficient estimé.

SE(β_1) : Erreur standard associée.

Étape 3 : Calcul de la p-valeur

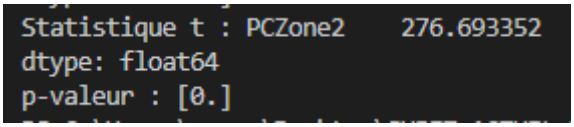
Comparer la statistique t à la valeur critique de la loi t pour un niveau de confiance de 95%.

La p-valeur correspond à la probabilité que β_1 soit aussi extrême ou plus si H_0 est vraie.

Étape 4 : Décision

Si p < 0.05, on rejette H_0 et on conclut que β_1 est significatif.

Sinon, on ne rejette pas H_0 .



Graphe 17 : Valeur de la statistique t

Résultat et Interprétation : β1 est-il significativement différent de zéro ?

Résultats **:** Statistique t : t=276.693352**,** p-valeur : p=0

Interprétation :

Signification de la p-valeur **:**

Une p-valeur de 0 indique que la probabilité d'observer un t-statistique aussi élevé (ou plus extrême) sous l'hypothèse nulle (H_0/ β_1=0) est pratiquement nulle**.**

Cela signifie que l'hypothèse nulle est rejetée avec un niveau de confiance très élevé (α=0.05).

Conclusion :

Le coefficient β_1 est significativement différent de zéro.

Cela confirme que PCZone2 a un impact significatif sur la consommation énergétique de PCZone1.

Le prédicteur PCZone2 est donc un facteur clé pour expliquer PCZone1 dans le modèle.

- **Interpretation du coefficient R^2**

```python
# Tester toutes les combinaisons de 1 à 5 prédicteurs
n = len(data)  # Nombre d'observations
for k in range(1, len(predictors) + 1):
    for combination in combinations(predictors, k):
        # Sous-ensemble de prédicteurs
        X_subset = data[list(combination)]
        Y = data['PCZone1']

        # Ajuster le modèle
        model = LinearRegression()
        model.fit(X_subset, Y)

        # Calculer R² et Adjusted R²
        r2 = model.score(X_subset, Y)
        adj_r2 = adjusted_r2(r2, n, k)

        # Mettre à jour le meilleur modèle
        if adj_r2 > best_adjusted_r2:
            best_adjusted_r2 = adj_r2
            best_model = model
            best_combination = combination

print(f"Meilleur modèle : {best_combination}")
print(f"Adjusted R² : {best_adjusted_r2}")
```

Graphe 18 :Code pour le Calcul de R^2 Et le choix de la meilleure combinaison linéaire multiple

```
Meilleur modèle : ('Temperature', 'Humidity', 'WindSpeed', 'PCZone2', 'PCZone3')
Adjusted R² : 0.8639571907139695
```

Graphe 19 : Résultat du code

Résultat obtenu : $R^2 = 0.8525$

Interprétation :

Cette valeur de R^2 indique que 85.25 % de la variance de la consommation énergétique en Zone 1 (PCZone1) est expliquée par la consommation énergétique en Zone 2 (PCZone2).

Cela montre que le modèle est très performant pour expliquer les variations de PCZone1 à partir de PCZone2.

Variance non expliquée :

Le reste (1 - R^2 = 14.75 %) correspond à la variance de PCZone1 qui n'est pas expliquée par PCZone2 et pourrait être attribuée à d'autres facteurs non inclus dans ce modèle.
Le modèle est-il adapté ?

Pertinence du modèle :

Avec un R^2 supérieur à 0.8, le modèle est considéré comme très adapté pour prédire la consommation énergétique en Zone 1 à partir de celle en Zone 2.

Cela confirme que PCZone2 est un prédicteur clé pour PCZone1.

b) Regression linéaire multiple

- **Meilleure combinaison**

D'après le résultat de la *figure 19* le meilleur modèle contient 5 variables explicatives.

Variables sélectionnées :

Les variables incluses dans le modèle sont : Temperature, Humidity, WindSpeed, PCZone2, PCZone

Interprétation :

Ces variables sont les meilleures pour prédire PCZone1, car elles maximisent l'Adjusted R^2, qui est de 0.86396. Cela signifie que ce modèle explique environ 86.4 % de la variance de PCZone1 après pénalisation pour le nombre de variables.

Réponse théorique :

L' Adjusted R^2 pénalise l'ajout de variables inutiles dans le modèle.

Contrairement à R^2, qui augmente toujours avec l'ajout de nouvelles variables, Adjusted R^2 diminue si les variables ajoutées n'améliorent pas significativement le modèle.

Cela rend Adjusted R^2 plus fiable pour comparer des modèles avec un nombre différent de prédicteurs.

Extraire les coefficients et calculer R^2

```
Coefficients : [-2.32480509e+02 -1.24423014e+01 -2.25697425e+02  1.29002254e+00
  3.86225106e-02]
Intercept : 12306.82951951908
R² : 0.8640085392492175
```

Graphe 20 : Extraction des coefficients et calcul de R^2

- **Résultats et Interprétation des Coefficients du Modèle Sélectionné**

Valeurs des Coefficients :

Intercept (β_0) : 12306.83

Cette valeur représente la consommation énergétique moyenne en Zone 1 (PCZone1) lorsque toutes les variables explicatives (prédicteurs) sont égales à zero

Coefficients des prédicteurs :

Temperature (β_1) : -232.48

Une augmentation d'un degré Celsius dans la température diminue la consommation énergétique en Zone 1 de 232.48 unités en moyenne, toutes choses égales par ailleurs.

Humidity (β_2) : -12.44

Une augmentation de 1 % d'humidité diminue la consommation en Zone 1 de 12.44 unités en moyenne, toutes choses égales par ailleurs.

WindSpeed (β_3) : -225.69

Une augmentation de 1 km/h de vitesse du vent diminue la consommation en Zone 1 de 225.69 unités en moyenne, toutes choses égales par ailleurs.

PCZone2 (β_4) : 1.29

Une augmentation de 1 unité de consommation énergétique en Zone 2 augmente la consommation en Zone 1 de 1.29 unités en moyenne.

PCZone3 (β_5) : 0.0386

Une augmentation de 1 unité de consommation énergétique en Zone 3 augmente la consommation en Zone 1 de 0.0386 unités en moyenne.

```
P-valeurs pour les coefficients :
Temperature: 0.0
Humidity: 0.0
WindSpeed: 0.0
PCZone2: 0.0
PCZone3: 0.0
```

Graphe 21: p – valeurs pour les coefficients

Résultats :

Les p-valeurs des coefficients sont toutes égales à 0.0 pour les variables suivantes: Temperature, Humidity, WindSpeed, PCZone2, PCZone3

Interprétation des p-valeurs :

Hypothèse testée :

- **Hypothèse nulle (H_0) :** Le coefficient (β) est égal à zéro, indiquant qu'il n'a pas d'impact significatif.
- **Hypothèse alternative (H_1) :** Le coefficient (β) est différent de zéro, indiquant qu'il a un impact significatif.

Décision :

Une p-valeur de 0.0 est inférieure au seuil typique (α=0.05), ce qui signifie que nous rejetons H0 pour tous les coefficients.

Conclusion :

Tous les coefficients (sauf β_0) sont significativement différents de zéro.

Chaque prédicteur (Temperature, Humidity, WindSpeed, PCZone2, PCZone3) contribue de manière significative à expliquer la consommation énergétique en Zone 1 (PCZone1).

Conditions données :

Température = 26°C, Humidité = 65%, Vitesse du vent = 4.2 km/h.

Consommation énergétique Zone 2 = 18840 KW, Zone 3 = 25700 KW.

Avec les conditions climatiques et les consommations dans les zones voisines fournies, la consommation énergétique prévue pour Zone 1 est d'environ **29,802.29 KW.**

Ce résultat est issu du modèle de régression multiple sélectionné et reflète les relations statistiques identifiées dans les données.

Conclusion

L'analyse de la consommation énergétique à Tétouan a permis de mettre en évidence des corrélations significatives entre les zones de consommation et les conditions climatiques locales. La méthode PCA a révélé que la majorité de la variance est portée par les variables énergétiques (Zone 1, 2 et 3), tandis que la température et l'humidité structurent la seconde composante.

La régression linéaire simple a mis en lumière un lien très fort entre la consommation des Zones 1 et 2 ($R^2 \approx 85$ %), tandis que la régression multiple a permis de construire un modèle prédictif robuste basé sur cinq variables explicatives : **PCZone2, PCZone3, Température, Humidité et WindSpeed**, avec un R^2 ajusté de **0.864**.

À partir de ce modèle, une prédiction a été réalisée pour la situation suivante :

- **Température** = 28.3 °C
- **Humidité** = 55 %

- **WindSpeed** = 13.3 km/h
- **PCZone2** = 30 500 KW
- **PCZone3** = 32 800 KW

Consommation énergétique prédite pour la Zone 1 : 29 802 KW

Ce résultat met en évidence la capacité du modèle à générer une estimation fiable à partir de données réelles. Il confirme aussi la pertinence des variables choisies dans la prédiction de la consommation.

Ce projet illustre la puissance des outils de Data Science pour modéliser et interpréter des phénomènes complexes comme la consommation énergétique. Il pose également les bases d'une amélioration future par des méthodes plus avancées, telles que les modèles de machine learning ou les séries temporelles.