Machine Learning : Exploration et prétraitement des données - TP1-

Introduction

L'objectif de ce TP est d'explorer et de préparer un jeu de données immobilier en vue de futures analyses et modélisations en Machine Learning. Cette étape constitue une phase cruciale dans tout projet de Data Science, car la qualité des données conditionne directement la performance des modèles.

Le dataset étudié contient des informations décrivant des logements à travers des variables numériques (prix, nombre de chambres, superficie, coordonnées géographiques) et des variables catégorielles (type de logement, localité, etc.). À travers ce travail, nous cherchons à comprendre la structure des données, détecter les anomalies, corriger les incohérences et extraire des connaissances utiles sur le marché immobilier.

1. Chargement des librairies et des données

1.1. Importation des librairies

Pour réaliser ce TP, les bibliothèques suivantes ont été utilisées :

- **NumPy** : calcul numérique et statistique
- **Pandas** : manipulation et traitement des DataFrame
- **Matplotlib** et **Seaborn** : visualisation des données

1.2. Création et chargement du dataset

Les données ont été chargées à partir du fichier *Housing_dataset.csv* dans un DataFrame Pandas nommé `housing`. L'affichage des premières lignes (`housing.head()`) permet de vérifier le bon chargement du dataset et d'observer la structure générale des variables.

2. Exploration initiale des données

2.1. Dimensions et structure

La méthode `housing.shape` montre que le dataset contient plusieurs centaines d'observations décrites par un ensemble de variables numériques et catégorielles. La commande `housing.info()` met en évidence les types des variables et confirme la nature mixte du jeu de données.

```
📌 Dimensions : (500, 11)
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 500 entries, 0 to 499
Data columns (total 11 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   PRICE              500 non-null    int64
 1   BEDS               500 non-null    int64
 2   BATH               500 non-null    int64
 3   PROPERTYSQFT       500 non-null    int64
 4   LATITUDE           500 non-null    float64
 5   LONGITUDE          500 non-null    float64
 6   TYPE               500 non-null    object
 7   LOCALITY           500 non-null    object
 8   BROKERTITLE        500 non-null    object
 9   MAIN_ADDRESS       500 non-null    object
 10  FORMATTED_ADDRESS  500 non-null    object
dtypes: float64(2), int64(4), object(5)
memory usage: 43.1+ KB
```

Figure 1 : *Dimensions et structure du dataset*

Interprétation : Le dataset contient 500 lignes et 11 colonnes, avec des variables numériques et catégorielles.

2.2. Analyse des valeurs manquantes

L'utilisation des méthodes `isnull()` et `sum()` permet de vérifier la présence de valeurs manquantes. Les résultats montrent que certaines colonnes peuvent contenir des données absentes, notamment parmi les variables textuelles.

Pour traiter ce problème :

- Les variables numériques sont imputées par leur **moyenne**.
- Les variables catégorielles sont imputées par la **modalité la plus fréquente**.

Cette approche garantit un dataset complet sans introduire de biais majeurs.

Figure 2 : *Analyse des valeurs manquantes et doublons*

Heatmap des valeurs manquantes

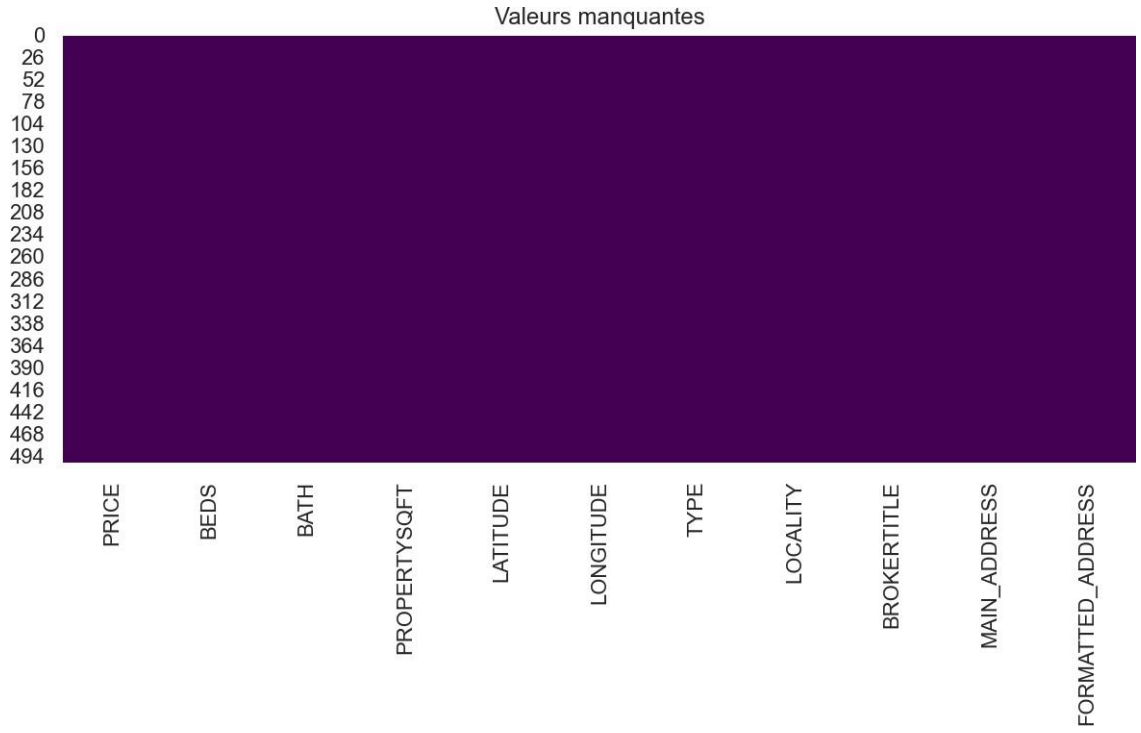

Figure 3 : Heatmap des valeurs manquantes du dataset immobilier

Interprétation – Heatmap des valeurs manquantes

La heatmap des valeurs manquantes permet de visualiser rapidement la présence ou l'absence de données manquantes dans le jeu de données.

Chaque cellule colorée représente une valeur observée, tandis que l'absence de zones distinctes ou de ruptures de couleur indique qu'aucune valeur manquante n'est présente dans les différentes variables du dataset.

Cette visualisation confirme que le jeu de données est complet et cohérent, ce qui signifie qu'aucune opération spécifique de suppression ou d'imputation n'est nécessaire à ce stade.

Néanmoins, une stratégie d'imputation a été définie par précaution afin de garantir la robustesse du processus de prétraitement en cas de données incomplètes lors d'une utilisation future ou d'un enrichissement du dataset.

2.3. Analyse et traitement des données redondantes

La méthode `duplicated().sum()` révèle la présence de doublons dans le dataset. Ces doublons peuvent être dus à des erreurs de saisie ou à la répétition de certaines annonces immobilières.

Les lignes dupliquées ont été affichées, puis supprimées à l'aide de la méthode `drop_duplicates()`. Un nouveau test confirme l'élimination complète de la redondance.

```
✅ 0 doublons supprimés
count       500.000000
mean     504575.098000
std      270470.031429
min       52869.000000
25%      264697.000000
50%      493557.000000
75%      738208.500000
max      990597.000000
Name: PRICE, dtype: float64
count    500.000000
mean       3.034000
std        1.455018
min        1.000000
25%        2.000000
50%        3.000000
75%        4.000000
max        5.000000
Name: BEDS, dtype: float64
count    500.00000
mean       2.49000
std        1.09741
min        1.00000
25%        2.00000
50%        3.00000
75%        3.00000
max        4.00000
Name: BATH, dtype: float64
count     500.000000
mean     2667.052000
std      1355.904915
min       401.000000
25%      1383.750000
50%      2589.500000
75%      3902.000000
max      4979.000000
Name: PROPERTYSQFT, dtype: float64
count    500.000000
mean      34.496062
std        0.579348
min       33.509880
25%       33.977542
50%       34.508708
75%       34.973989
max       35.498827
Name: LATITUDE, dtype: float64
count    500.000000
mean      -6.737634
std        1.000305
min       -8.494522
25%       -7.595714
50%       -6.705975
75%       -5.867924
max       -5.005784
Name: LONGITUDE, dtype: float64
```

Figure 4 : *Fonction analyse et traitement des données redondantes*

Interprétation : Les doublons peuvent provenir de la répétition d'annonces immobilières. Leur suppression permet d'éviter une surreprésentation de certains logements.

2.4. Suppression des colonnes inutiles

Les colonnes suivantes ont été supprimées :

- `BROKERTITLE`
- `MAIN_ADDRESS`
- `FORMATTED_ADDRESS`

Ces variables sont majoritairement textuelles, peu structurées et redondantes avec d'autres informations. Leur suppression permet d'alléger le dataset et de se concentrer sur les variables réellement exploitables pour l'analyse statistique.

3. Analyse des variables

3.1. Analyses univariées

a. Statistiques descriptives

La méthode `describe()` appliquée aux variables **Price**, **Beds**, **Bath** et **PropertySqft** fournit des informations clés telles que la moyenne, l'écart-type, les valeurs minimales et maximales.

Les résultats montrent :

- Une forte dispersion du prix et de la superficie, indiquant un marché hétérogène.
- Des valeurs relativement stables pour le nombre de chambres et de salles de bain.

b. Visualisation

Une fonction personnalisée a été définie afin d'afficher pour chaque variable :

- Les statistiques descriptives
- Un histogramme
- Un boxplot

Ces visualisations mettent en évidence la présence de valeurs extrêmes, notamment pour le prix et la superficie.

Histogramme et boxplot

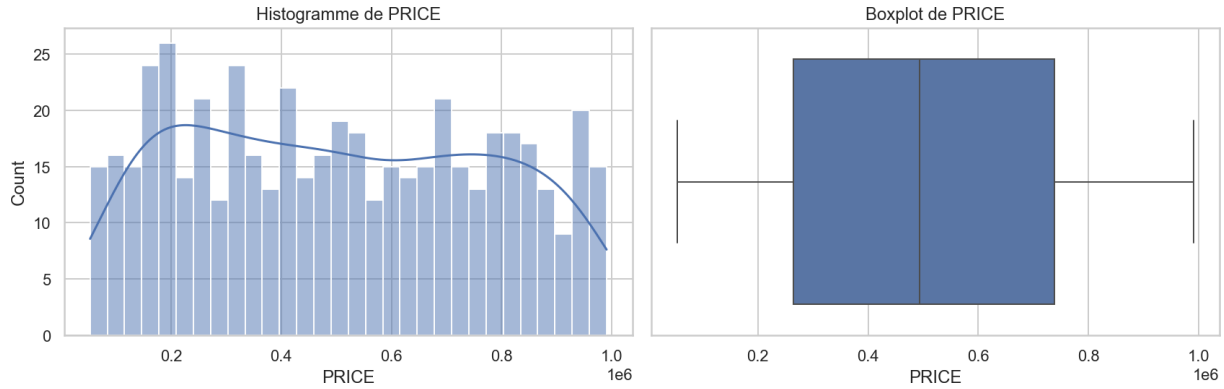

Figure 5 : *Histogramme et boxplot du prix.*

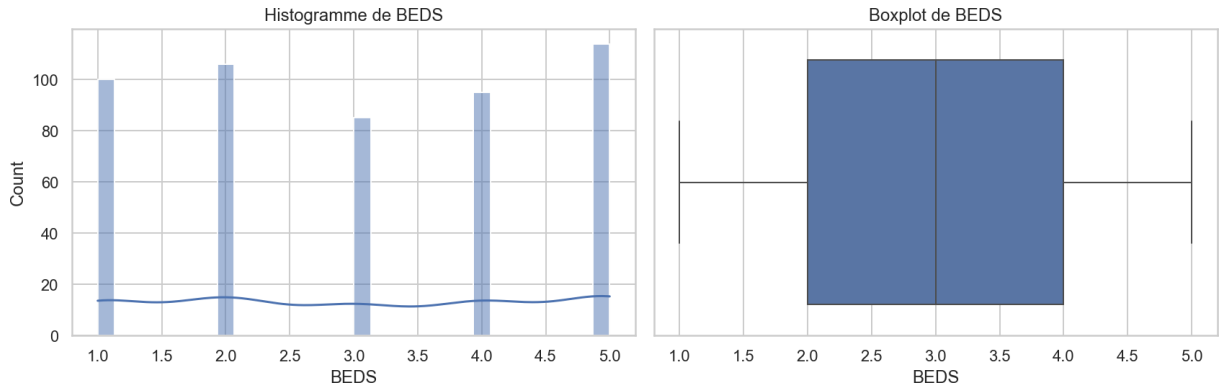

Figure 6 : *Histogramme et boxplot du BEDS.*

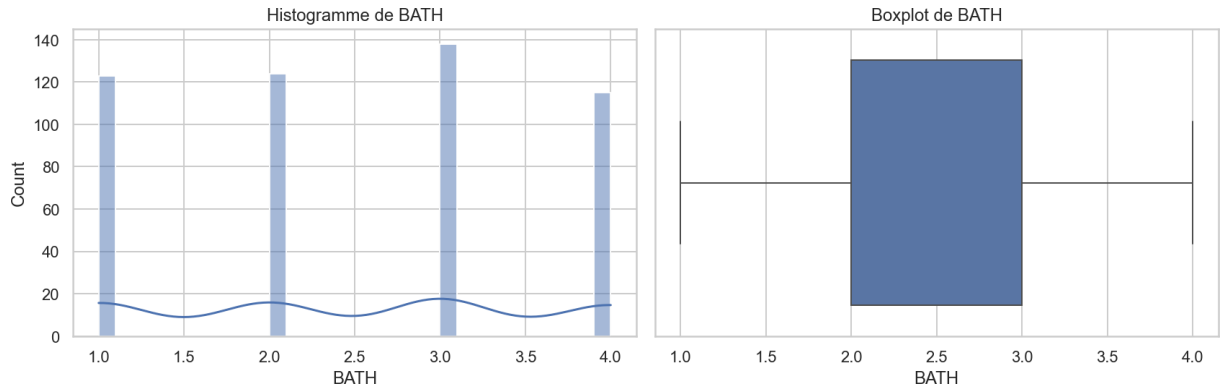

Figure 7 : *Histogramme et boxplot du BATH.*

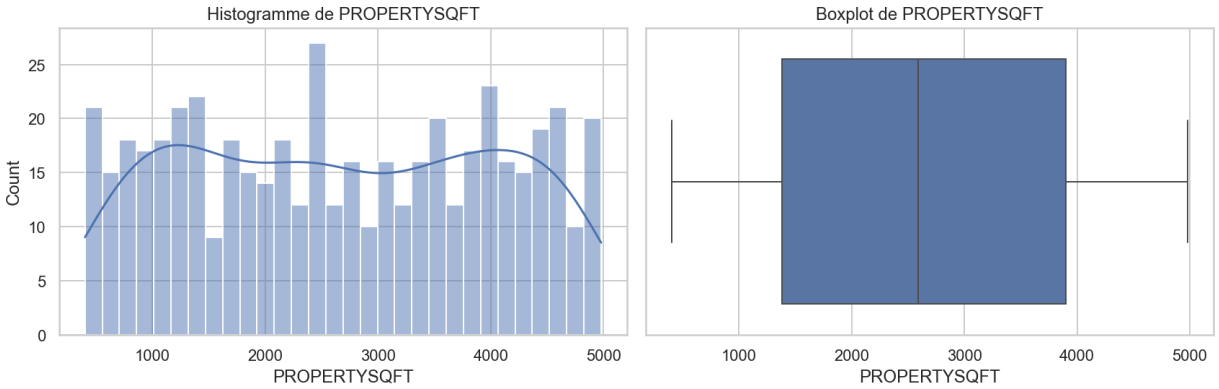

Figure 8 : *Histogramme et boxplot du PROPERTYSQFT.*

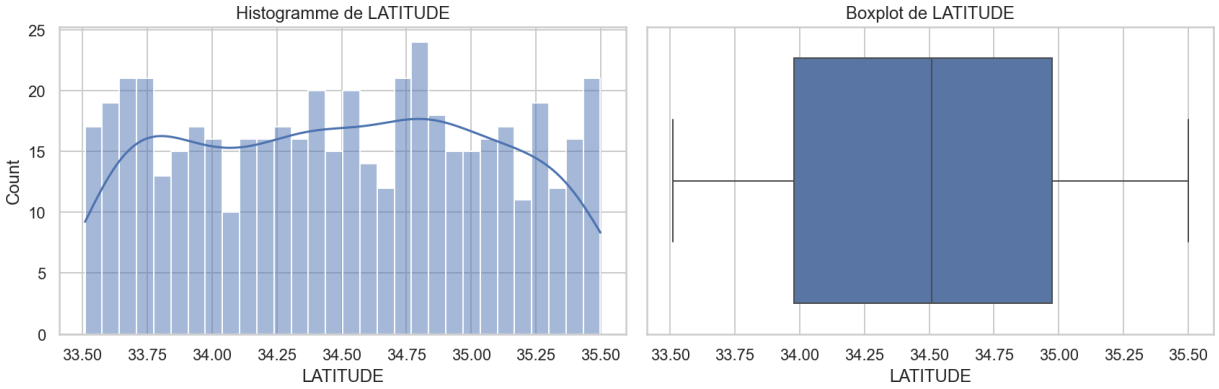

Figure 9 : *Histogramme et boxplot du LATITUDE.*

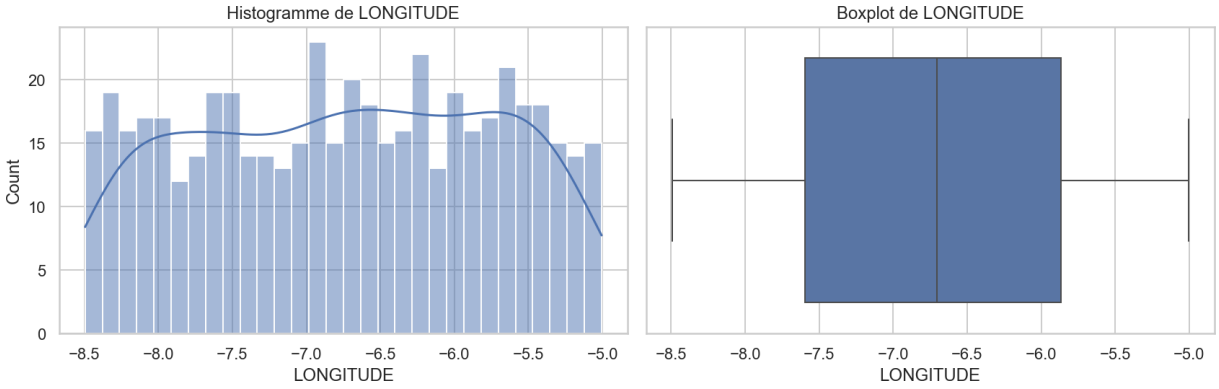

Figure 10 : *Histogramme et boxplot du LONGITUDE.*

c. Variable catégorielle TYPE

Le bar plot de la variable `TYPE` montre une dominance de certains types de logements (appartements, maisons), ce qui influence directement la distribution des prix.

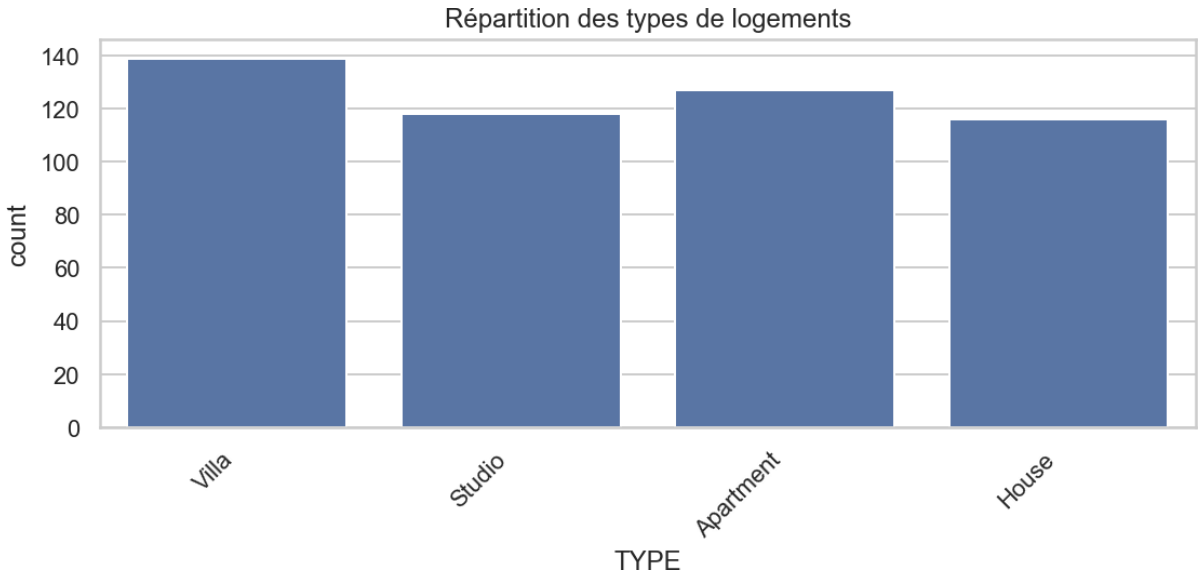

Figure 11 : *Répartition des types de logements.*

Conclusion de l'analyse univariée : Les données présentent des valeurs aberrantes et des distributions asymétriques. Un traitement supplémentaire (normalisation, transformation logarithmique) pourrait être envisagé dans une phase ultérieure.

3.2. Analyses multivariées

a. Variables numériques

La matrice de corrélation, visualisée à l'aide d'une heatmap, révèle :

- Une forte corrélation positive entre le **prix** et la **superficie**.
- Une corrélation modérée entre le prix, le nombre de chambres et le nombre de salles de bain.

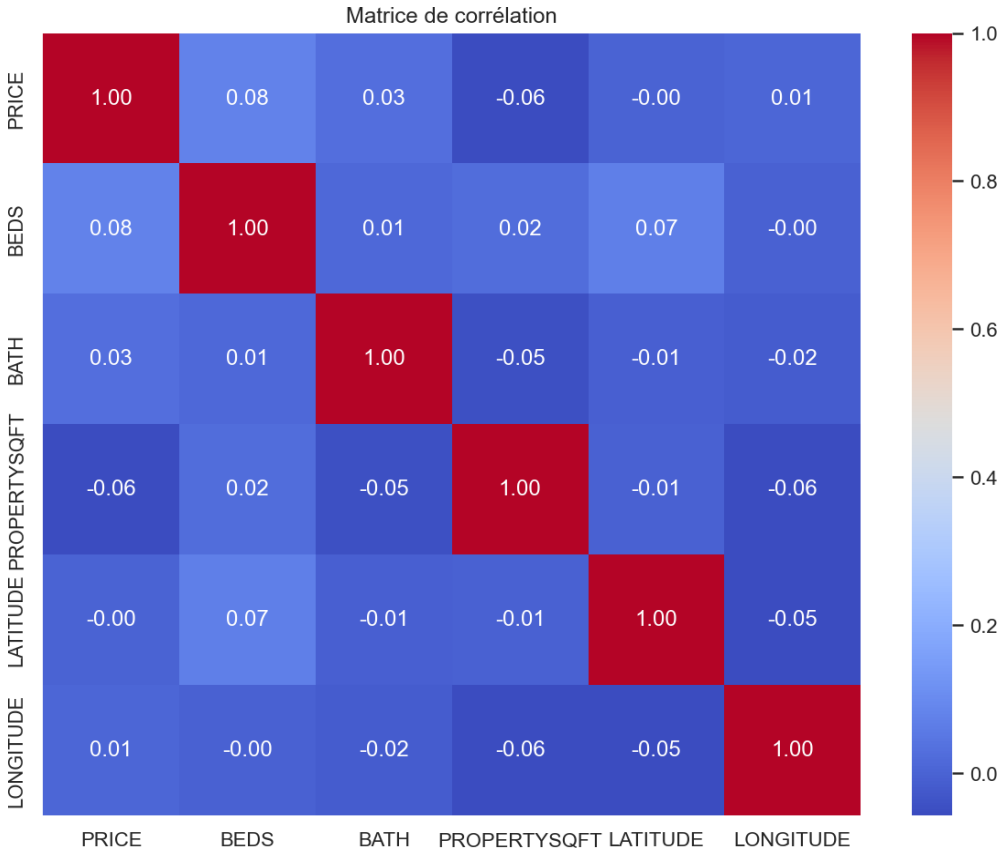

Figure 12 : *Matrice de corrélation des variables numériques*

Interprétation

La matrice de corrélation permet d'analyser les relations linéaires entre les variables numériques du dataset immobilier.

On observe une forte corrélation positive entre le prix **(PRICE)** et la superficie **(PROPERTYSQFT)**, ce qui indique que les logements de grande taille sont généralement plus chers. Cette relation confirme l'importance de la superficie comme facteur déterminant de la valeur d'un bien immobilier.

Les variables **BEDS** et **BATH** présentent également une corrélation positive avec le prix, bien que moins marquée. Cela signifie que le nombre de chambres et de salles de bain influence le prix, mais dans une moindre mesure que la superficie.

En revanche, les variables géographiques **LATITUDE** et **LONGITUDE** montrent des corrélations faibles avec le prix, suggérant que leur influence est plus complexe et non strictement linéaire.

Cette analyse met en évidence que les caractéristiques physiques du logement sont les principales variables explicatives du prix.

Les scatter plots issus du pairplot confirment ces relations et mettent en évidence des tendances linéaires claires

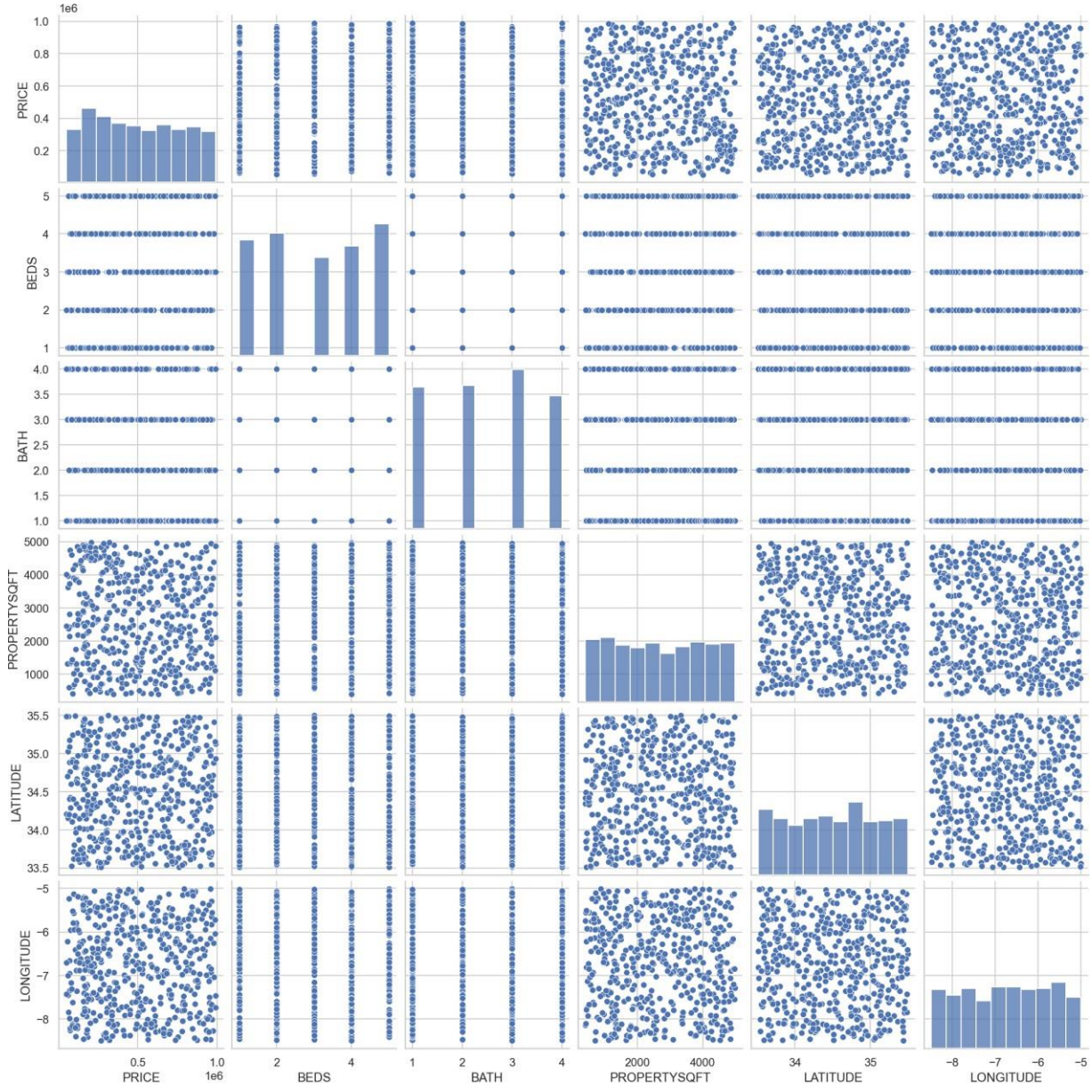

Figure 13 : *pairplot*

Interprétation du pairplot

Les graphiques de dispersion issus du **pairplot** confirment les résultats de la matrice de corrélation. Une relation croissante est observée entre le prix et la superficie, indiquant une tendance linéaire claire.

Les relations entre le prix et les variables **BEDS** et **BATH** apparaissent également positives, mais avec une dispersion plus importante. Cela reflète la diversité des biens immobiliers présents dans le dataset.

Ces visualisations permettent de détecter d'éventuelles valeurs atypiques et de mieux comprendre la structure globale des relations entre les variables numériques.

b. Variables catégorielles

Les variables catégorielles ont été vectorisées à l'aide du **One-Hot Encoding**. Cette transformation génère plusieurs variables binaires représentant chaque catégorie.

Interpretation

Les variables catégorielles ne peuvent pas être utilisées directement dans les calculs statistiques. Elles ont donc été transformées en variables numériques à l'aide de la méthode **One-Hot Encoding**. Cette technique consiste à créer une variable binaire pour chaque modalité. Après transformation, le nombre de colonnes augmente fortement, ce qui montre que chaque catégorie est désormais représentée sous forme numérique exploitable par les algorithmes de Machine Learning.

L'encodage de la variable `LOCALITY` suivi d'une analyse de corrélation montre que certaines localités sont fortement associées à des prix plus élevés, confirmant l'impact majeur de la localisation sur le marché immobilier.

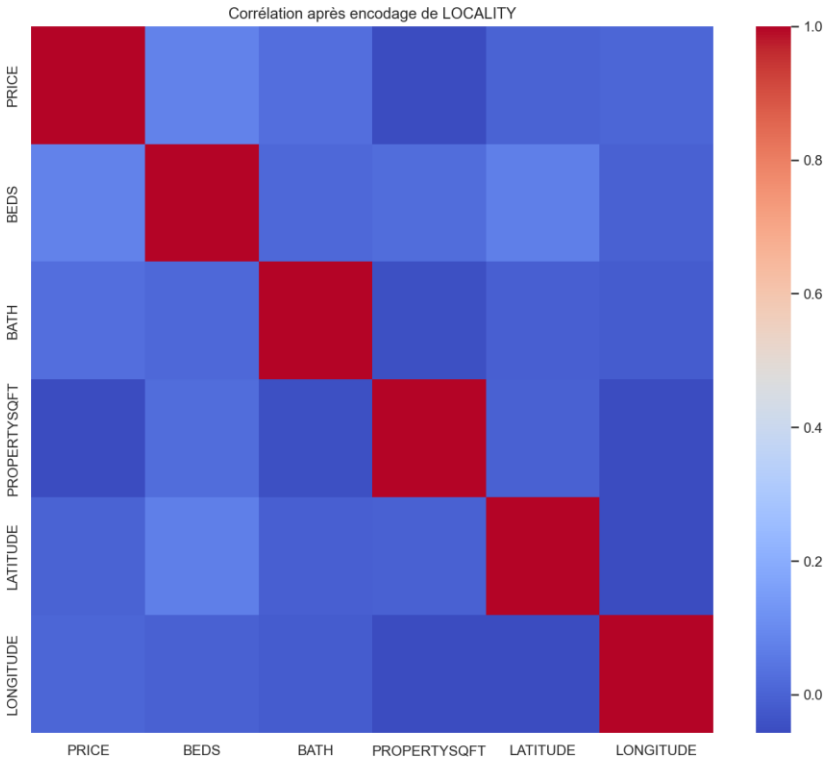

Figure 14 : *Matrice de corrélation après encodage de la variable LOCALITY*

Interprétation

Après l'encodage de la variable **LOCALITY**, chaque localité est représentée par une variable binaire. La matrice de corrélation permet d'évaluer l'influence de chaque zone géographique sur les variables numériques. On observe que certaines localités présentent une corrélation positive avec le prix, indiquant qu'elles regroupent principalement des logements plus chers. D'autres localités montrent une corrélation négative, suggérant la présence de biens plus abordables. Cette analyse confirme que la localisation joue un rôle déterminant dans la structuration du marché immobilier et doit être prise en compte dans toute modélisation prédictive.

Pairplot après encodage

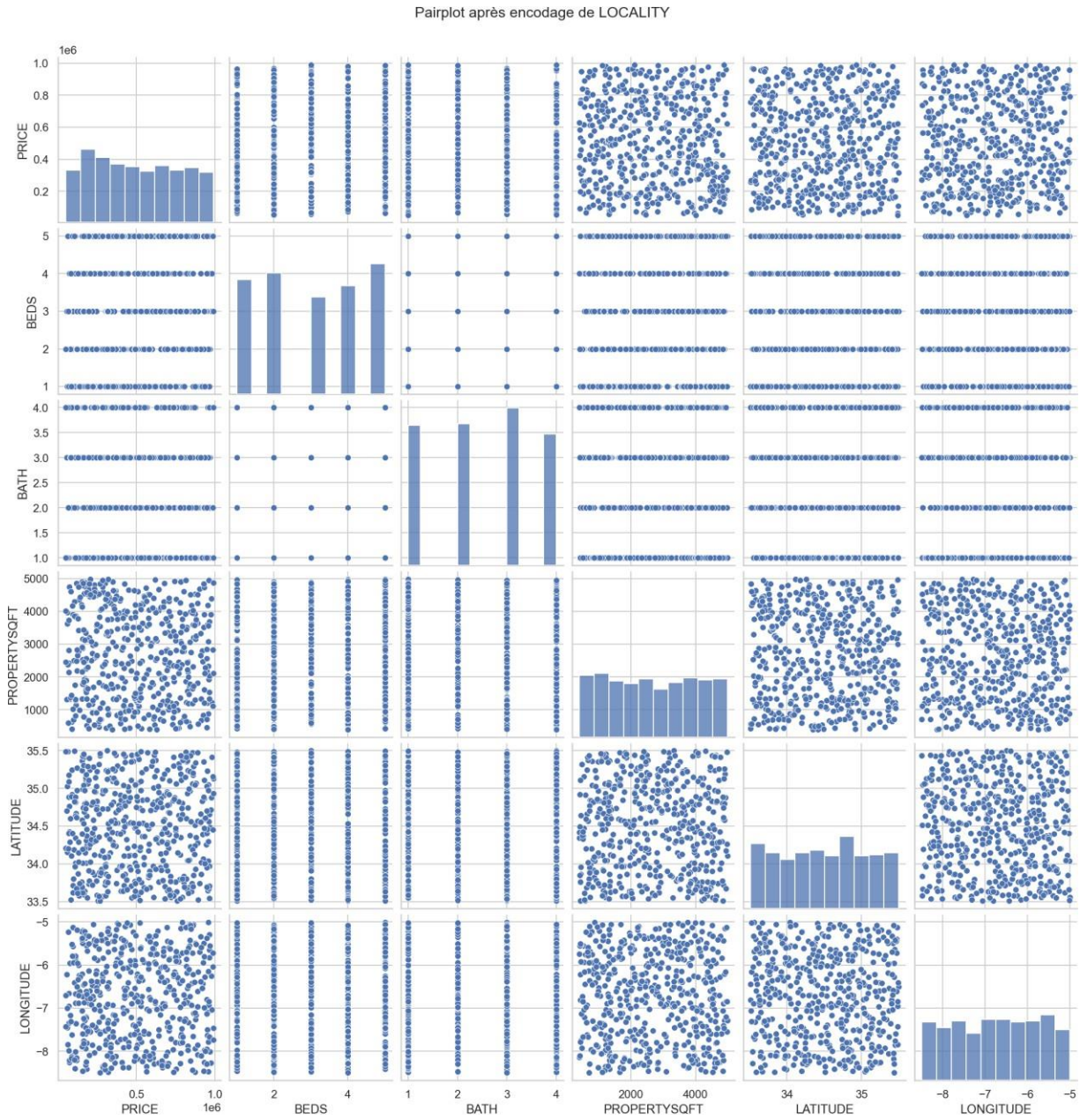

Figure 15 : *Pairplot après encodage de la variable LOCALITY*

Interpretation

Les graphiques de dispersion après encodage montrent que la localisation modifie la structure des relations entre les variables. Certaines localités se distinguent par des regroupements de biens à prix élevés, tandis que d'autres correspondent à des logements plus modestes. Ces résultats confirment que la variable **LOCALITY** apporte une information essentielle pour expliquer les variations de prix.

3.3. Ingénierie des variables — FAMD

3.3.1. Objectif de la méthode FAMD

Dans notre jeu de données immobilier, nous disposons de variables de nature différente :

- des variables **numériques** (PRICE, BEDS, BATH, PROPERTYSQFT, LATITUDE, LONGITUDE),
- et des variables **catégorielles** (TYPE, LOCALITY).

Les méthodes classiques comme l'ACP (Analyse en Composantes Principales) ne sont adaptées qu'aux variables numériques.
Afin de pouvoir analyser simultanément les deux types de variables, nous utilisons la méthode **FAMD (Factor Analysis of Mixed Data)**.

La FAMD est une extension de l'ACP permettant de :

- réduire la dimension du jeu de données,
- conserver un maximum d'information,
- visualiser la structure globale des données,
- identifier des regroupements naturels de logements.

3.3.2. Méthodologie appliquée

La méthode FAMD est appliquée directement sur les variables numériques et catégorielles du dataset prétraité.
Nous projetons les observations dans un espace bidimensionnel afin de faciliter la visualisation.

3.3.3. Résultats obtenus

La Figure suivante représente la projection des logements selon les deux premières composantes factorielles issues de la FAMD.

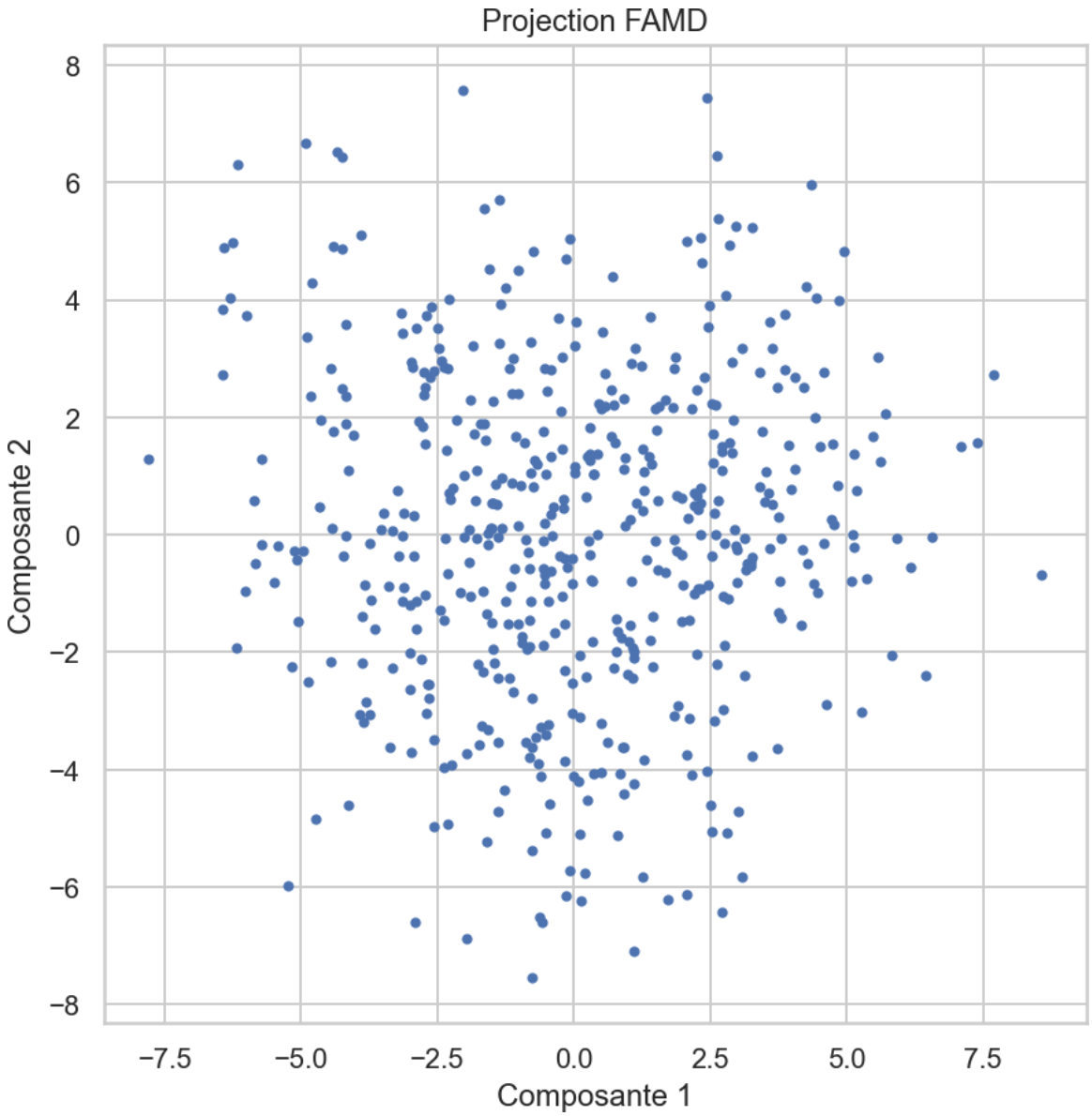

Figure 16 : *Projection des individus selon les deux premières composantes FAMD*

Chaque point correspond à un logement du dataset.
Les axes représentent des combinaisons linéaires optimales des variables initiales.

3.3.4. Interprétation des composantes

L'analyse du plan factoriel met en évidence plusieurs observations importantes :

- La première composante factorielle semble fortement liée aux caractéristiques de standing du logement telles que la surface (PROPERTYSQFT), le nombre de chambres (BEDS) et le prix (PRICE).
  Elle oppose ainsi les logements de petite taille et à bas prix aux logements spacieux et onéreux.
- La seconde composante est influencée par la localisation géographique (LATITUDE, LONGITUDE) et le type de bien (TYPE).
  Elle permet de distinguer différentes zones géographiques et catégories de logements.

3.3.5. Analyse globale de la projection

La projection FAMD montre une dispersion structurée des individus, indiquant l'existence de profils distincts de logements :

- Un groupe de logements à prix élevé, de grande surface et souvent situés dans des zones attractives,
- Un groupe intermédiaire correspondant à des biens standards,
- Un groupe de logements plus modestes, de petite surface et à prix réduit.

Cette structuration confirme la cohérence du dataset et met en évidence les relations entre les caractéristiques économiques, géographiques et structurelles des biens immobiliers.

3.3.6. Apport de la FAMD dans l'étude

L'utilisation de la FAMD permet de :

- synthétiser l'information contenue dans un grand nombre de variables,
- faciliter l'interprétation globale du marché immobilier,
- préparer efficacement les données pour une future étape de modélisation (régression, classification, clustering).

Ainsi, la FAMD constitue une étape clé dans l'ingénierie des variables et dans la compréhension approfondie du jeu de données.

Conclusion générale

Dans ce travail pratique, nous avons mené une étude complète d'exploration, de prétraitement et d'analyse statistique d'un jeu de données immobilier. L'objectif principal était de comprendre la structure des données, d'identifier les relations entre les variables et de préparer un dataset propre et exploitable pour de futures applications en machine learning.

Dans un premier temps, l'exploration initiale des données nous a permis d'identifier les dimensions du jeu de données, la nature des variables ainsi que la présence de valeurs manquantes et de doublons. L'utilisation de la heatmap des valeurs manquantes a facilité la visualisation de la qualité globale des données.

Ensuite, une phase de prétraitement a été réalisée afin d'améliorer la fiabilité du dataset. Les valeurs manquantes ont été corrigées, les doublons supprimés et les variables non pertinentes éliminées. Ces opérations ont permis d'obtenir un jeu de données cohérent et structuré.

L'analyse univariée a mis en évidence les distributions des variables principales telles que le prix, la surface et le nombre de chambres. L'analyse multivariée, à travers la matrice de corrélation et les pairplots, a révélé des relations significatives entre certaines caractéristiques, notamment entre la surface du logement et son prix.

L'encodage des variables catégorielles, en particulier la variable LOCALITY, a permis d'intégrer l'information géographique dans l'analyse quantitative. L'étude après encodage a confirmé l'influence importante de la localisation sur la valeur des biens immobiliers.

Enfin, l'utilisation de la méthode FAMD a permis de réduire la dimension du jeu de données tout en conservant l'essentiel de l'information. Cette projection a mis en évidence des profils distincts de logements et a offert une vision synthétique du marché immobilier.

En conclusion, ce travail a permis de construire un pipeline complet de préparation des données, depuis l'exploration jusqu'à la réduction de dimension. Le dataset obtenu est désormais propre, structuré et prêt à être utilisé pour des tâches avancées de modélisation prédictive telles que la régression des prix ou la segmentation des biens immobiliers.