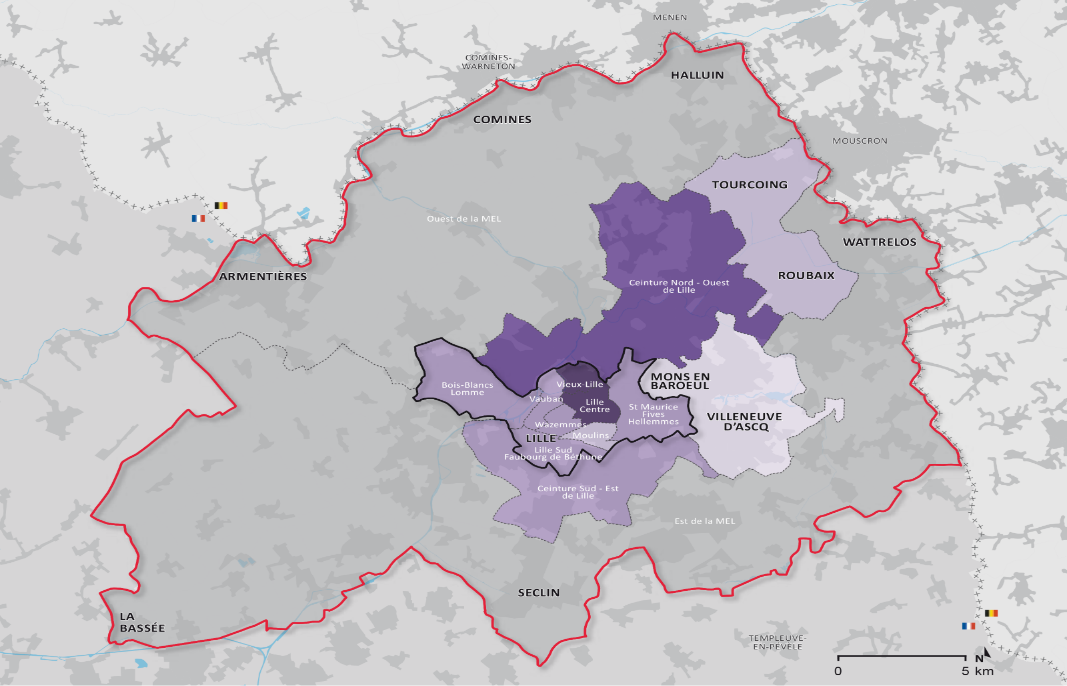




**RAPPORT D’ÉTUDE DE CAS**

*Étude des Loyers dans l'Agglomération Lilloise : Positionnement et Conseil pour IMMO-Bill*



*RÉALISÉ PAR :*

BAQUET Jean

YOMETOWOU Marie-Christelle

**Introduction**

Implantée dans l’agglomération lilloise, Immo Bill est une société spécialisée dans l’intermédiation immobilière, opérant dans l’achat, la vente, ainsi que la location de biens. Afin d’améliorer la qualité des services de cette société, en particulier dans les conseils relatifs à la fixation des loyers, Immo Bill souhaite optimiser ses recommandations aux propriétaires bailleurs.

Cette étude vise à renforcer le rôle de conseil d'IMMO-Bill tout en améliorant sa position stratégique sur le marché lillois

Notre mission consiste à répondre à deux objectifs principaux. Nous devons dans un premier temps mettre en place un système expert capable de calculer le niveau moyen des loyers selon les caractéristiques des logements et les critères pertinents. Dans un second temps nous évaluerons la compétitivité des loyers pratiqués par IMMO Bill par rapport aux loyers du marché à Lille, en identifiant les segments où l’agence se positionne comme concurrentielle ou non.

Ainsi en nous appuyant sur un échantillon de 1 556 logements, nous mettrons en œuvre une approche économétrique pour analyser les données disponibles, détecter les loyers aberrants, et répondre aux attentes de l’agence. Cette étude vise à renforcer le rôle de conseil d'IMMO-Bill tout en améliorant sa position stratégique sur le marché lillois.

La liste et la définition des variables du fichier utilisables dans l’étude et le nombre d’observations total :

Les variables du fichier utilisables dans l’étude :

* Pièces : renseigne sur le nombre de pièces présentes dans les appartements, variant de 1 à 8
* Quartier : Fournit des informations sur les 20 quartiers de la région lilloise inclus dans l’étude.
* Représente le montant du loyer appliqué aux logements, en fonction de leurs caractéristiques et de leur emplacement dans chaque quartier étudié.
* Surface : Indique la superficie des appartements, exprimée en mètres carrés, pour chaque logement inclus dans l’étude.
* Appartenance : Informe sur le statut d’appartenance des appartements à une catégorie ou un type spécifique, tel que défini dans l’étude.
* Prix au mètre carré : Reflète le coût moyen par mètre carré des appartements en fonction de leur quartier et de leurs caractéristiques.

Nous avons rajouté des variables à notre études afin d’obtenir une base de données plus riche nous permettant ainsi d’approfondir notre analyse. Ces variables sont

* Niveau de vie : Indique le revenu médian ou moyen des ménages dans les quartiers étudiés, permettant de caractériser les conditions économiques générales des habitants.
* Part du logement social : Reflète la proportion de logements sociaux dans chaque quartier, exprimée en pourcentage du total des logements disponibles.
* Part des ménages pauvres : Mesure la proportion de ménages vivant sous le seuil de pauvreté, indiquant le niveau de précarité économique dans les différents quartiers étudiés.
* ❖ Nombre d'observation total= 1556

L'analyse des statistiques descriptives dans cette étude se base autour plusieurs parties :

Il peut y avoir dans cette base de données brutes à analyser des valeurs manquantes, aberrantes, de même que des doublons ou des informations mal renseignés dues à des erreurs de saisie ou à des événements imprévus. Ces anomalies doivent être corrigées pour éviter de fausser l’analyse.

1. **Traitement des données**
2. *A)* ***Harmonisation des Modalités Mal Renseignées pour une Analyse Fiable (Viieux- Lille= Vieux Lille)***

La variable *Quartier* présentait des anomalies. En effet des incohérences dans l’orthographe des noms ont été constater . Par exemple, "Vieux Lille" était parfois écrit avec un double "i" ("Vieux Lille"), et "Saint-Maur" remplaçait fréquemment "Saint-Maurice". Ces erreurs de saisie ont été corrigées afin d’harmoniser les données de notre modèle et d’assurer une analyse rigoureuse. Ce nettoyage préliminaire a permis d’éviter toute confusion ou regroupement incorrect lors des étapes suivantes de l’étude.

***B)*** ***Élimination des Doublons afin de faciliter l’optimisation nos données***

Outre les erreurs d’orthographe, des doublons, c'est-à-dire des enregistrements strictement identiques, ont également été identifiés au sein de notre base de données. Un total de 34 doublons a été recensé. Deux outils complémentaires ont été utilisés pour leur détection et leur suppression :

1. **Méthode SAS :** La procédure **PROC SORT** avec l’option **NODUPKEY** a été appliquée. Cette méthode trie les données selon une combinaison de variables clés (*Pièces*, *Loyer*, *Surface*, *Quartier*, *Appartenance*, et *Pm2*) et élimine automatiquement les enregistrements redondants. Cela garantit que seules les observations uniques sont conservées.
2. **Approche Excel :** Une méthode manuelle a également été mise en œuvre à l’aide d’une formule combinée. Une colonne intermédiaire a été créée pour générer un identifiant unique basé sur la concaténation des variables clés, à l’aide de la formule

=A3&"-"&B3&"-"&C3&"-"&D3&"-"&E3&"-"&F3&"-"&TEXTE(B3;"0.00")

Ensuite, une autre colonne a permis d’identifier les doublons avec la formule: =SI(NB.SI(N$3:N3;N3)>1;"Doublon";"Unique")

Cette méthode attribue une étiquette "Doublon" ou "Unique" à chaque ligne, facilitant ainsi leur suppression.

Ces doublons, qui n’ajoutaient aucune information pertinente, ont été retirés pour éviter tout biais dans l’analyse. L’utilisation conjointe de SAS et Excel a permis de croiser les résultats et de confirmer les résultats obtenus, renforçant ainsi la qualité et l’intégrité des données utilisées pour cette étude.

***C) Traitement des données aberrantes et manquantes***

* **Données manquantes**

Aucune donnée manquante n’a été identifiée dans le cadre de cette étude. Pour confirmer cette absence, plusieurs méthodes ont été mises en œuvre, notamment l’utilisation d’Excel. Une formule basée sur la fonction “**SI** “ a été appliquée pour détecter rapidement les cellules vides, garantissant ainsi une vérification efficace des éventuelles valeurs manquantes.

* **Données aberrantes**

Afin de traitées nos variables aberrantes nous avons réalisées des graphiques de boîte à moustaches (*boxplots*) afin de visualiser la distribution des variables clés telles que le loyer (*Loyer*), la surface (*Surface*), et le prix au mètre carré (*Pm2*). Ces graphiques permettent de repérer rapidement les valeurs aberrantes situées en dehors des bornes calculées (souvent définies comme 1,5 fois l’écart interquartile au-dessus du troisième quartile ou en dessous du premier quartile). Une fois identifiées, ces valeurs ont été examinées pour déterminer si elles correspondaient à des erreurs de saisie ou à des cas spécifiques justifiables. Les valeurs non pertinentes ou clairement erronées ont été exclues.

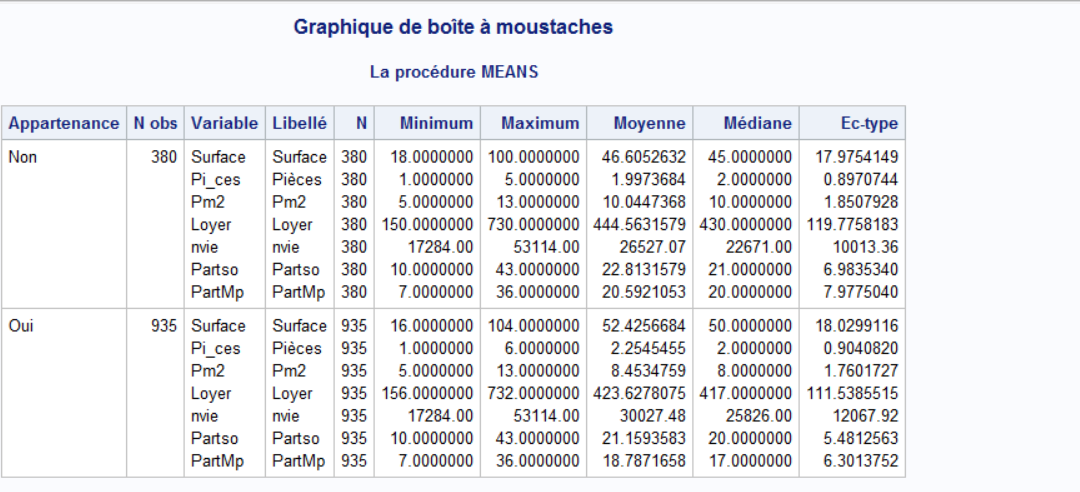
***D) Enrichissement des Données avec des Variables Socio-Démographiques Localisées***

Nous avons enrichi notre analyse en intégrant des données socio-démographiques précises, recueillies à l’échelle de carrés géographiques de 1 km². Au sein de ses données nous avons le **niveau de vie** qui représentent le revenu médian ou moyen des ménages, pour caractériser les conditions économiques générales des habitants), la **part de logements sociaux** (indiquant le pourcentage de logements sociaux par rapport au total des logements disponibles), et la **part des ménages pauvres** (mesurant la proportion de ménages vivant sous le seuil de pauvreté, afin d’évaluer le niveau de précarité économique). Ces données permettent de contextualiser l’analyse en tenant compte des caractéristiques socio-économiques des quartiers étudiés.

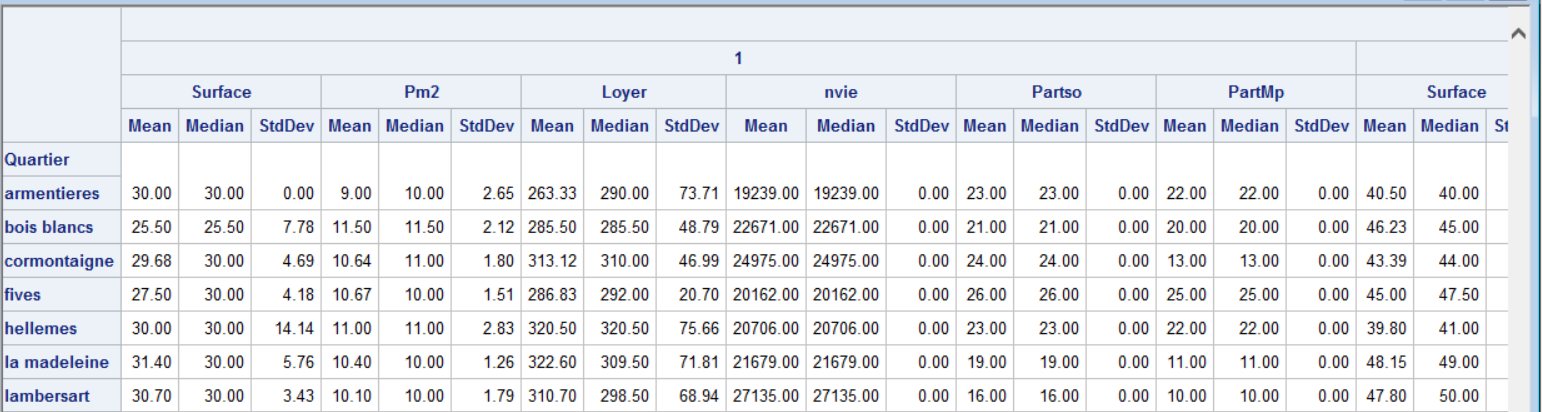
***E) Analyse Statistique Descriptive des Variables***

La procédure **PROC FREQ** a été utilisée pour générer des tables de fréquence, permettant d’analyser la répartition des observations dans différentes catégories, comme les quartiers, et d’obtenir une vue d’ensemble de leur distribution. Ensuite, **PROC MEANS** a fourni des mesures essentielles telles que la moyenne, la médiane, les valeurs minimales et maximales, ainsi que l’écart type pour les variables continues comme le loyer, la surface ou le prix au mètre carré. Cette analyse a été approfondie grâce à l’utilisation de la commande **CLASS**, qui permet d’examiner ces statistiques pour des sous-groupes spécifiques (par exemple, selon le quartier, le nombre de pièces ou l’appartenance à un groupe). En complément, **PROC TABULATE** a produit des tableaux croisés pour combiner des variables continues et catégoriques, offrant une analyse structurée des relations entre les différentes dimensions, comme l’interaction entre les quartiers et les types de logements. Ces statistiques descriptives ont pour objectif de détecter les tendances générales, d'identifier des disparités ou des regroupements particuliers, et de fournir un socle solide pour les analyses ultérieures. Elles permettent également d’évaluer la qualité des données et d’établir des bases comparatives pour explorer des relations plus complexes entre les variables dans le contexte de l’étude.

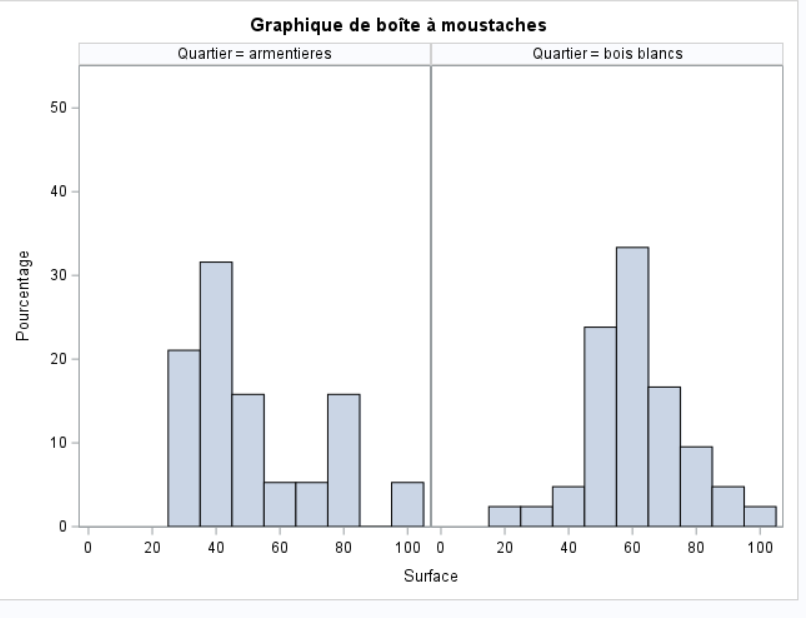
Les statistiques descriptives réalisées à travers les procédures ses nous ont permis de tirer des informations clés sur les variables étudiées. En comparant les observations selon les catégories d’appartenance (*Oui* ou *Non*), des différences notables apparaissent. Par exemple, les logements appartenant au groupe *Oui* affichent une surface moyenne légèrement plus élevée (52,43 m² contre 46,61 m² pour *Non*), ainsi qu’un loyer moyen plus important (423 € contre 444 €). Ces écarts se reflètent également dans l'écart type, soulignant une variabilité plus importante des loyers dans le groupe *Non*.

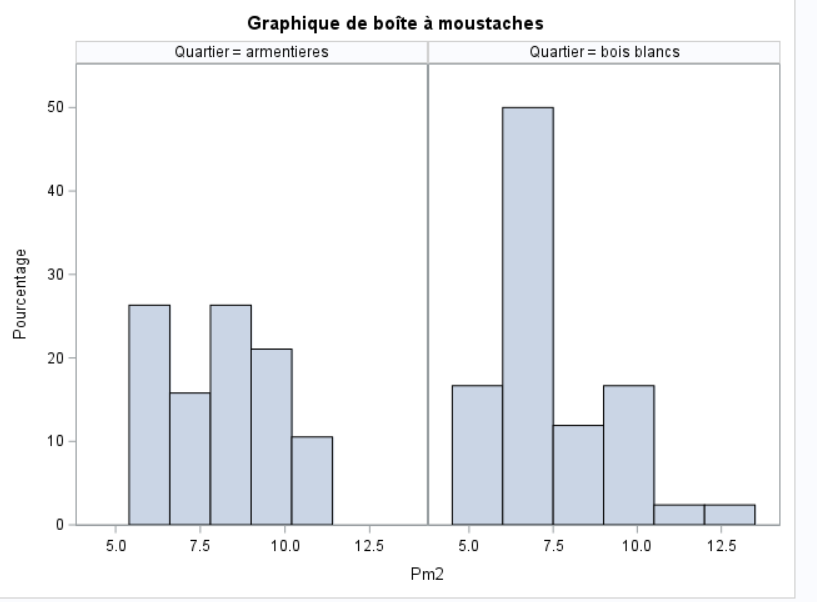


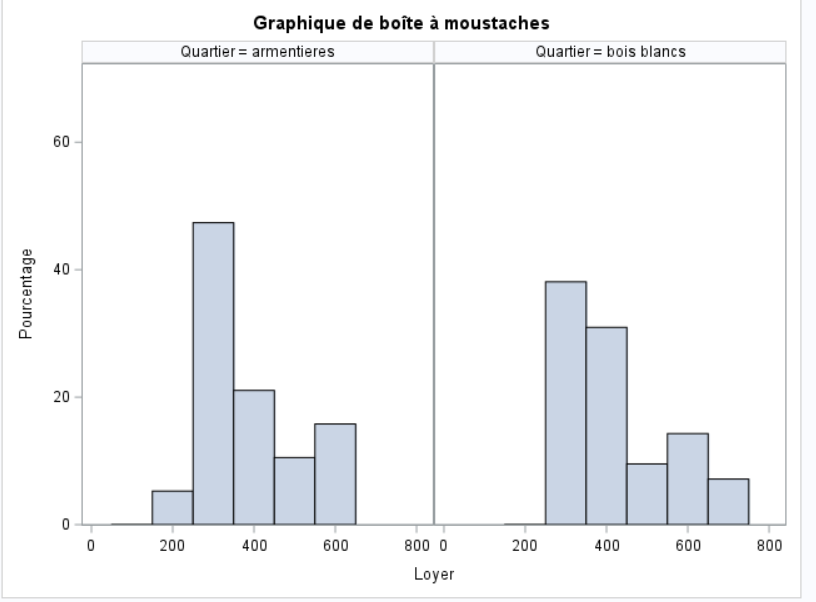
De plus, les statistiques calculées par quartier mettent en lumière des disparités géographiques significatives. Par exemple, le loyer moyen varie entre 263 € à Armentières et 322 € à La Madeleine, tandis que la surface médiane reste relativement stable autour de 30 m² dans la plupart des quartiers. Les variables socio-démographiques, telles que le niveau de vie (*nvie*), montrent également des écarts marqués : le revenu médian est le plus bas à Fives (20 162 €) et le plus élevé à Lambersart (27 135 €). Enfin, les parts des logements sociaux (*Partso*) et des ménages pauvres (*PartMp*) diffèrent sensiblement d’un quartier à l’autre, avec des valeurs maximales à Fives (26 % de logements sociaux) et des valeurs minimales à Lambersart (10 %). Ces observations issues des statistiques descriptives révèlent des tendances et des inégalités marquées dans les caractéristiques immobilières et socio-économiques, offrant ainsi une base solide pour des analyses plus approfondies.



Ajoutées à cela les histogrammes générés à l’aide de la procédure **PROC SGPANEL** permettent d’analyser la distribution des surfaces, du prix au mètre carré (*Pm2*), et des loyers par quartier, en mettant en évidence des disparités significatives. Concernant les surfaces, le quartier d’Armentières présente une répartition concentrée autour de 40 à 50 m², avec peu de grandes surfaces, tandis que Bois Blancs offre une plus grande diversité, avec des surfaces fréquentes entre 50 et 70 m², et quelques très petites ou très grandes surfaces. Pour le prix au mètre carré, Armentières montre une concentration des prix entre 5 et 10 €/m², avec une forte densité autour de 7,5 €/m², alors que Bois Blancs affiche une répartition légèrement plus homogène et des prix globalement plus élevés. Enfin, pour les loyers, les logements d’Armentières se situent majoritairement entre 300 € et 400 €, tandis que ceux de Bois Blancs ont une répartition plus étendue, avec une majorité entre 400 € et 600 €, et quelques logements dépassant 700 €. Ces observations reflètent des différences dans l’offre immobilière et la demande, les quartiers comme Bois Blancs semblant proposer des logements plus diversifiés et globalement plus coûteux que ceux d’Armentières, offrant une base pour des analyses plus approfondies des dynamiques du marché immobilier.







***F) Création de Variables Dummies pour les Catégories de Pièces, Surface, Loyer, Appartenance et Prix au m²***

Afin de garantir une source de données claire, cohérente et exploitable dans SAS, plusieurs variables initiales ont été transformées ou recodées. Ces ajustements étaient essentiels pour permettre des analyses approfondies, notamment la réalisation de régressions statistiques. Voici les transformations et recodages effectués :

#### **1. Catégorisation de la Surface**

La variable *Surface*, initialement exprimée en mètres carrés, a été regroupée en trois catégories principales :

* **Classe 1 (Petite surface)** : Logements de moins de 40 m².
* **Classe 2 (Moyenne surface)** : Logements entre 40 m² et 55 m².
* **Classe 3 (Grande surface)** : Logements de plus de 55 m².

Cette catégorisation permet de simplifier l’interprétation des résultats et de cibler l’analyse en fonction de la taille des logements.

#### **2. Transformation du Prix au mètre carré (Pm2)**

La variable *Pm2* (Prix au mètre carré) a été regroupée en trois classes :

* **Classe 1 (Bas)** : Prix au m² inférieur à 8 €/m².
* **Classe 2 (Moyen)** : Prix au m² compris entre 8 €/m² et 10 €/m².
* **Classe 3 (Élevé)** : Prix au m² supérieur à 10 €/m².

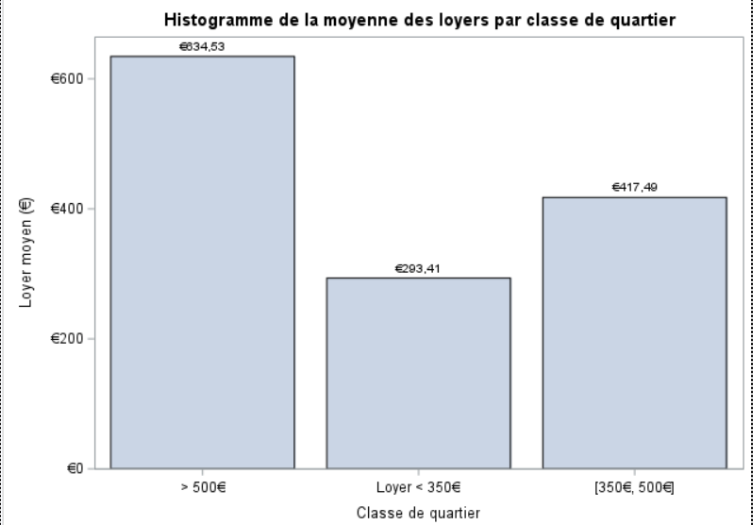
Cette classification facilite l’analyse car elle permet de mettre en avant les différences de prix dans les logements.

#### **3. Catégorisation du Loyer**

La variable *Loyer* a été regroupée en trois classes :

* **Classe 1 (Bas)** : Loyer inférieur à 350 €.
* **Classe 2 (Moyen)** : Loyer compris entre 350 € et 500 €.
* **Classe 3 (Élevé)** : Loyer supérieur à 500 €.

Cette catégorisation est particulièrement utile pour comparer les logements en fonction de leurs niveaux de loyer.



#### **4. Transformation du Nombre de Pièces**

La variable *Pièces* (nombre de pièces, variant de 1 à 8) a été transformée en variables indicatrices (*dummies*) :

* **pieces\_1** : 1 si le logement a 1 pièce, 0 sinon.
* **pieces\_2** : 1 si le logement a 2 pièces, 0 sinon.
* **pieces\_3** : 1 si le logement a 3 pièces, 0 sinon.
* **pieces\_4** : 1 si le logement a 4 pièces, 0 sinon.
* **pieces\_5plus** : 1 si le logement a 5 pièces ou plus, 0 sinon.

Elle facilite l’intégration de cette variable dans les modèles et permet ainsi une analyse spécifique des logements selon leur taille.

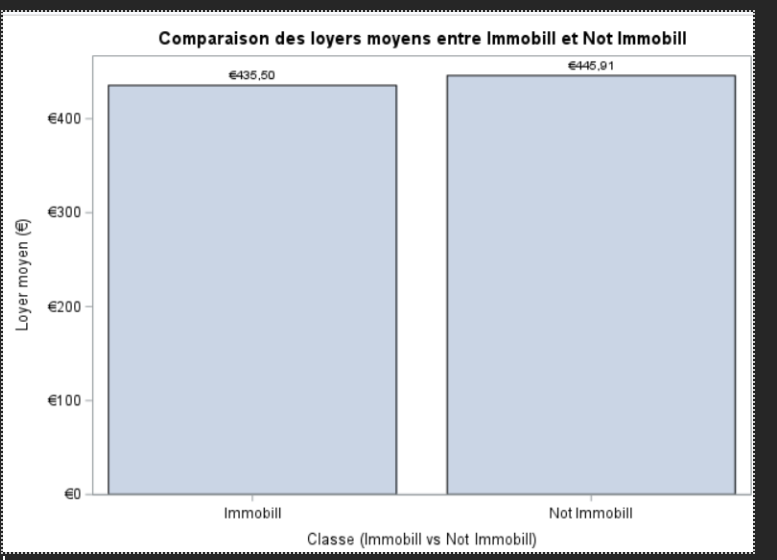
#### **5. Transformation de la Variable Quartier**

La variable *Quartier*, qui renseignait sur les 20 quartiers de l’étude, a été transformée en variables indicatrices (*dummies*), une pour chaque quartier (par exemple, **armentieres**, **bois\_blancs**, **fives**, etc.). Chaque quartier est représenté par une variable prenant la valeur 1 si le logement appartient à ce quartier, et 0 sinon. Cette transformation permet une comparaison simplifiée entre les quartiers.

#### **6. Transformation de la Variable Appartenance**

La variable *Appartenance* (Oui/Non) a été transformée en deux variables distinctes :

* **appartenance\_oui** : 1 si *Appartenance* = Oui, 0 sinon.
* **appartenance\_non** : 1 si *Appartenance* = Non, 0 sinon.



Cela permet de comparer les caractéristiques des logements appartenant ou non à cette catégorie spécifique.

#### **7. Gestion des Valeurs Manquantes**

Pour traiter les valeurs manquantes, les variables *Niveau de Vie* (*nvie*), *Part du Logement Social* (*Partso*), et *Part des Ménages Pauvres* (*PartMp*) ont été complétées avec des valeurs spécifiques :

* **nvie** : Remplacée par la médiane de 24 975.
* **Partso** : Remplacée par 24.
* **PartMp** : Remplacée par 13.

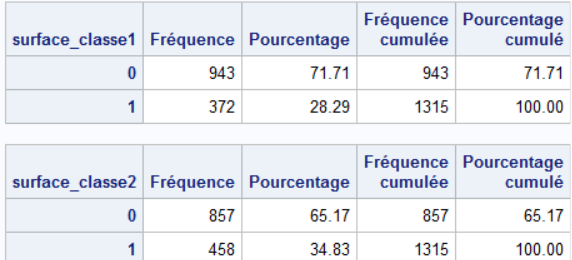
Ce traitement garantit que les données sont complètes et prêtes pour les analyses statistiques.

#### **8. Analyse des Fréquences**

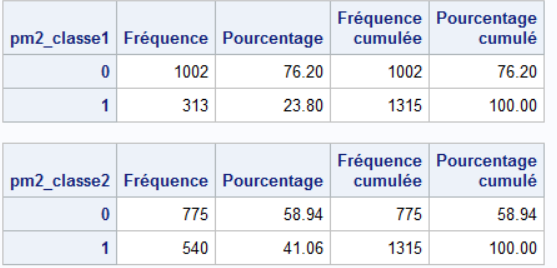
Les fréquences des catégories des variables créées ont été analysées à l’aide de **PROC FREQ**, permettant de valider la répartition des observations dans chaque classe pour les variables comme :

* *surface\_classe1, surface\_classe2, surface\_classe3*.
* *pm2\_classe1, pm2\_classe2, pm2\_classe3*.
* *loyer\_classe1, loyer\_classe2, loyer\_classe3*.
* *pieces\_1, pieces\_2, pieces\_3, pieces\_4, pieces\_5plus*.

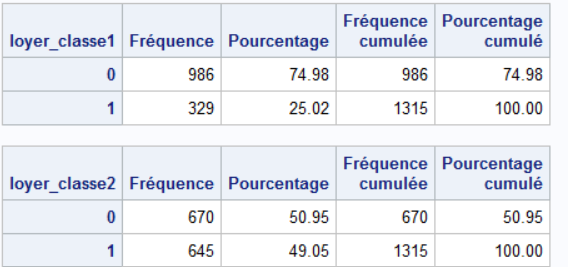
Concernant la variable *Surface*, la majorité des logements (71,71 %) ont une surface supérieure à 40 m², tandis que les petites surfaces (< 40 m²) ne représentent que 28,29 %. Les surfaces intermédiaires (40 m² à 55 m²) constituent un tiers des observations, soulignant une distribution marquée autour des tailles moyennes.



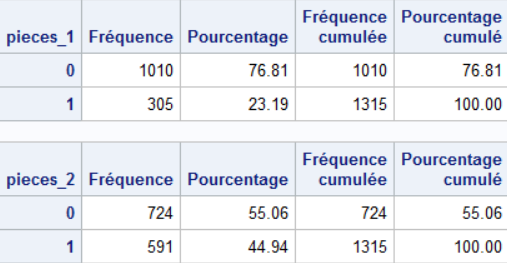
Pour le *Prix au mètre carré (Pm2)*, les logements les moins chers (< 8 €/m²) sont minoritaires avec 23,80 % des observations, tandis que les prix intermédiaires (8 €/m² à 10 €/m²) regroupent 41,06 %, indiquant une concentration importante dans cette catégorie.



En ce qui concerne le *Loyer*, les logements avec des loyers inférieurs à 350 € sont également rares (25,02 %), alors que la classe intermédiaire (350 € à 500 €) est bien représentée, constituant environ 49,05 % des observations, avec une répartition presque égale entre ceux appartenant ou non à cette catégorie.



Pour le *nombre de pièces*, les logements avec une seule pièce représentent une faible proportion (23,19 %), tandis que ceux avec deux pièces sont bien plus fréquents (44,94 %), traduisant une préférence pour des configurations légèrement plus spacieuses.



Ces résultats soulignent des tendances claires dans les données, notamment une préférence pour les logements de taille moyenne, des loyers modérés, et des prix au mètre carré intermédiaires, confirmant la pertinence des catégorisations effectuées.

#### **9. Transformation du Loyer**

La variable *Loyer* a été utilisée telle quelle, sans transformation logarithmique, pour l'intégration dans les modèles statistiques. Cette approche est cohérente avec la nature additive du modèle, où les relations entre les variables sont supposées être directement linéaires, sans nécessiter de transformation pour stabiliser la variance ou atténuer les effets des grandes variations.

### **Statistiques descriptives avant modélisation**

Cette section présente les statistiques descriptives calculées pour les variables indépendantes et la variable d’intérêt. Ces analyses préliminaires fournissent des orientations essentielles pour notre étude, permettant de formuler des hypothèses initiales et de poser les bases nécessaires à la modélisation statistique.

1. **Moyenne** :

La moyenne est obtenue en divisant la somme de toutes les valeurs par le nombre total de données. Elle reflète une tendance centrale des observations.

1. **Médiane** :

La médiane représente la valeur centrale d’un ensemble de données ordonnées.

* 1. Si le nombre de données est impair, elle est la valeur située exactement au milieu.
  2. Si le nombre de données est pair, elle est la moyenne des deux valeurs centrales.

La médiane est particulièrement utile pour analyser des distributions asymétriques ou contenant des valeurs aberrantes.

1. **Variance** :

La variance quantifie la dispersion des données par rapport à leur moyenne. Elle représente la moyenne des carrés des écarts par rapport à la moyenne.

1. **Écart type** :

L’écart type est la racine carrée de la variance. Contrairement à la variance, il est exprimé dans la même unité que les données, ce qui facilite son interprétation. Un écart type élevé indique une dispersion importante des valeurs autour de la moyenne.

1. **Covariance** :

La covariance mesure la relation linéaire entre deux variables.

* 1. Une covariance positive indique que les deux variables évoluent dans le même sens (elles augmentent ou diminuent ensemble).
  2. Une covariance négative indique que lorsque l'une des variables augmente, l'autre tend à diminuer.

La covariance permet ainsi d'identifier des tendances communes entre les variables.

1. **Coefficient de variation (CV)** :

Le coefficient de variation est une mesure relative de la dispersion des données autour de leur moyenne. Calculé en divisant l’écart type par la moyenne, il est généralement exprimé en pourcentage. Le CV permet de comparer la variation relative de variables ayant des unités ou des échelles différentes.

1. **Modélisation**

**A) Présentation des variables et du modèle**

Pour comprendre la nature du marché immobilier nous avons étudié certain facterus influençant le loyer des biens. Ces facteurs caractéristiques propres aux logements, comme leur surface ou leur prix au mètre carré, mais aussi des aspects contextuels tels que la localisation dans un quartier spécifique ou les caractéristiques socio-économiques des habitants. Pour capturer les effets globaux des différentes caractéristiques qui influencent notre variable cible est est important de structurer la base de données en y implémentant de nouvelles variables qu’on appelera variable dummies et de classe.

Il faut distinguer les variables continues, binaires et catégorielles. Elles permettent d'adopter une approche où chaque variable contribue à expliquer une dimension spécifique du loyer. Par exemple, les variables continus mesurent les caractéristiques quantitatives des biens, tandis que les variables binaires pour les quartiers et le nombre de pièces permettent de capturer des différences qualitatives importantes. Enfin, les variables de classes apportent une perspective supplémentaire en regroupant les biens selon des catégories spécifiques, telles que les intervalles de surface ou de prix au mètre carré.

Dans le cadre de cette étude, nous avons construit un modèle de régression linéaire visant à expliquer le loyer en fonction des caractéristiques des biens immobiliers et de leurs quartiers. Ce modèle permet de quantifier les facteurs influençant le loyer et de répondre à la problématique centrale : **"Les biens d'Immobil sont-ils plus chers ou moins chers que ceux de ses concurrents ?"** Pour cela, deux modèles sont estimés séparément : l’un pour les biens appartenant à Immobil et l’autre pour ceux appartenant à ses concurrents. Cette méthode permet de comparer les loyers entre les deux groupes à caractéristiques similaires.

### **Structure et Variables du Modèle**

#### **1. Les caractéristiques continues**

Les variables quantitatives jouent un rôle important dans le modèle, car elles permettent de capturer des effets continus des biens immobiliers. Par exemple, la **surface totale** est une variable de ce type : plus un logement est grand, plus son loyer est élevé. De même, le **prix au mètre carré** est un indicateur de la valeur relative d’un bien, permettant de comparer des logements de tailles différentes.

De plus, des variables socio-économiques comme le **niveau de vie moyen** (nvie) et la **proportion de ménages pauvres** (partMp) apportent une dimension contextuelle. Ces variables permettent d’évaluer comment le loyer est influencé par les caractéristiques des habitants du quartier. Par exemple, un quartier avec un niveau de vie élevé peut justifier des loyers plus importants, tandis qu’une forte proportion de ménages pauvres pourrait limiter les prix pratiqués.

#### **2. Les variables catégorielles des quartiers**

La localisation est un facteur déterminant dans l’évaluation des loyers. Pour capturer cet effet, chaque quartier est représenté par une variable binaire (dummy variable). Par exemple, la variable **armentieres** prend la valeur 1 si le bien est situé dans le quartier d’Armentières, et 0 sinon. Ce codage est répété pour les 22 quartiers étudiés, permettant de quantifier l’impact spécifique de chaque quartier sur le loyer.

Les quartiers présenteent des caractéristiques uniques qui influencent directement le prix des logements. Ces variables permettent également de comparer les quartiers entre eux et d’identifier ceux où Immobil a des loyers significativement différents de ses concurrents.

#### **3. Les variables catégorielles pour le nombre de pièces**

Outre le nombre total de pièces (variable continue), des variables binaires sont utilisées pour indiquer si un logement appartient à une catégorie spécifique (par exemple, 2 pièces, 3 pièces, etc.). Ces variables permettent d’examiner des variations de loyers qui ne seraient pas directement liées à la surface.

#### **4. Les classes pour la surface et le prix au m²**

Pour explorer des relations non linéaires, nous avons regroupé les biens en **classes de surface** et **classes de prix au mètre carré**. Par exemple, les petites surfaces (moins de 12 m²) sont classées différemment des grandes surfaces (plus de 24 m²). De même, les prix au mètre carré faibles, moyens et élevés sont catégorisés.

#### **Présentation des variables utilisées**

Les variables explicatives sont :

1. **Caractéristiques des biens :**
   1. Surface : La superficie totale du bien (en m²), une variable clé dans la détermination du loyer.
   2. Pi\_ces : Le nombre de pièces du logement, qui capture sa fonctionnalité et son attractivité.
   3. Pm2 : Le prix par mètre carré, une mesure standardisée de la valeur du bien.
2. **Aspects socio-économiques :**
   1. NVIE : Niveau de vie moyen des habitants dans la zone, qui reflète le pouvoir d'achat local.
   2. PartMp : La proportion de ménages pauvres, indiquant le profil économique de la population.
   3. Partso : La proportion de logements sociaux dans la zone, souvent corrélée à la demande et à la perception des quartiers.
3. **Localisation géographique :**
   1. Des variables binaires (dummies) ont été créées pour chaque quartier, permettant de capturer les effets spécifiques des 22 zones étudiées. Chaque quartier est représenté par une variable prenant la valeur 1 si le bien est situé dans ce quartier, et 0 sinon (par exemple, armentieres, bois\_blancs, etc.).
4. **Catégorisation des biens :**
   1. Des variables de classe regroupent les biens selon leur surface (surface\_classe1, surface\_classe3) et leur prix au mètre carré (pm2\_classe1, pm2\_classe3). Cela permet d'examiner les effets spécifiques de certaines catégories (par exemple, les très petites surfaces ou les biens avec un faible prix au mètre carré).

#### **Approche méthodologique**

* Les deux modèles utilisent une **sélection pas à pas (stepwise)** pour retenir uniquement les variables les plus significatives.
* Les valeurs ajustées (Loyer\_prédit) seront comparées entre les deux groupes pour analyser les différences de loyers.
* Enfin, les coefficients des variables d'intérêt seront examinés pour déterminer si Immobil applique des prix systématiquement plus élevés ou plus bas que ses concurrents, en tenant compte des caractéristiques des biens et de leur localisation.

**B) Justification du choix du modèle**

* Le modèle de régression linéaire permet de comprendre l'effet spécifique d'Immobil sur les loyers tout en contrôlant les autres facteurs qui influencent les prix, comme la surface, le prix au mètre carré ou la localisation. En construisant deux modèles distincts (un pour Immobil et un pour ses concurrents), il est possible de comparer directement les loyers pratiqués pour des biens présentant des caractéristiques équivalentes. Par exemple, si les coefficients liés à certaines variables, comme la surface ou le quartier, diffèrent significativement entre les deux modèles, cela peut révéler une politique tarifaire particulière d’Immobil.
* Ce modèle offre également des prédictions ajustées (Loyer\_predit) pour chaque bien, permettant de comparer les loyers réels avec ceux attendus. Cette fonctionnalité aide à repérer des écarts importants entre Immobil et ses concurrents, notamment dans certains quartiers ou pour des types spécifiques de logements. En intégrant des variables catégoriques pour les quartiers et des classes pour la surface et le prix au mètre carré, le modèle capture des variations locales et structurelles qui seraient autrement négligées, enrichissant ainsi l’analyse.
* Un autre avantage clé de la régression linéaire est sa simplicité et son interprétation directe. Chaque coefficient du modèle exprime l'effet marginal d'une variable explicative sur le loyer. Par exemple, le coefficient associé à la surface indique combien d'unités monétaires le loyer augmente pour chaque mètre carré supplémentaire, toutes choses égales par ailleurs. Cela facilite l’interprétation et permet d’identifier clairement les facteurs qui influencent les loyers, tout en comparant les pratiques tarifaires d’Immobil à celles de ses concurrents.
* En résumé, ce modèle aide à évaluer si Immobil pratique des loyers plus élevés ou plus bas que ses concurrents pour des biens comparables, tout en identifiant des segments ou des zones géographiques où ces différences sont les plus marquées. Les résidus (Loyer\_reel−Loyer\_predit) permettent en outre de repérer les biens où les loyers diffèrent significativement des attentes, révélant ainsi des anomalies ou des tendances spécifiques.

**C) Modélisation**

**A) Régression pour Immobill**



Comme le montre les résultats ci-dessus, la variable **Surface** a un coefficient de **7.15 unités monétaires**, toutes choses égales par ailleurs. Cet effet est statistiquement significatif, avec une p-valeur <0,0001. Cela reflète l'importance de la superficie dans la détermination du loyer, un résultat intuitif et attendu dans le domaine immobilier.

La variable **Pm2** (prix par mètre carré) est également un déterminant majeur du loyer, avec un coefficient β=36.45. Cela implique qu'une augmentation d'une unité monétaire par mètre carré entraîne une augmentation moyenne de **36.45 unités monétaires** du loyer. La p-valeur associée (0.0001) confirme la pertinence de cette variable. Ces résultats montrent que les caractéristiques physiques et financières des biens immobiliers (surface et prix au mètre carré) sont des facteurs clés pour expliquer les loyers.

**Ci-dessous, les graphiques des résidus sont intéressants à étudier :**

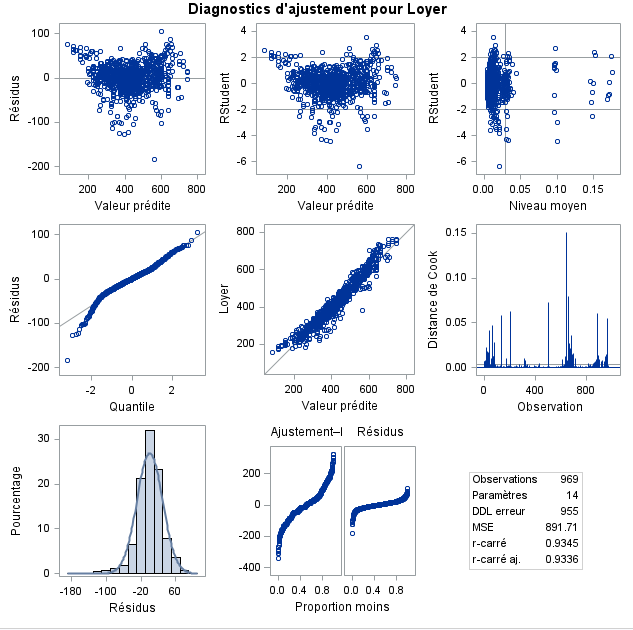
Les résidus, lorsqu'ils sont tracés en fonction des valeurs prédites, montrent un schéma légèrement en "U". En théorie, les résidus devraient être répartis aléatoirement autour de zéro. De plus, les résidus standardisés identifient quelques observations influentes, ce qui est l’effet des valeurs extrêmes. La distribution des résidus semble approximativement normale, comme le montre l’histogramme et le QQ-plot.

Dans un modèle de régression linéaire, l’une des hypothèses fondamentales est que les **résidus** (les différences entre les valeurs observées et les valeurs prédites par le modèle) suivent une distribution normale. Cette hypothèse est cruciale pour plusieurs raisons :

L’hypothèse de normalité des résidus garantit la fiabilité des tests statistiques associés au modèle, comme les tests *t*t pour les coefficients et les tests *F*F pour la qualité globale de l’ajustement. Si les résidus ne suivent pas une distribution normale, les p-valeurs obtenues pourraient être inexactes, ce qui pourrait fausser l’interprétation des résultats.

La normalité des résidus est également essentielle pour calculer des **intervalles de confiance** et des **intervalles de prédiction** fiables. Ces intervalles sont basés sur l’hypothèse que les erreurs suivent une loi normale. Une violation de cette hypothèse pourrait rendre les intervalles inexacts, réduisant ainsi la fiabilité des prédictions issues du modèle.

Globalement, lorsque les résidus suivent une distribution normale, le modèle est considéré comme plus robuste pour faire des prédictions. Cela signifie que les erreurs de prédiction sont symétriques et concentrées autour de zéro, minimisant ainsi le risque de biais systématique dans les estimations.



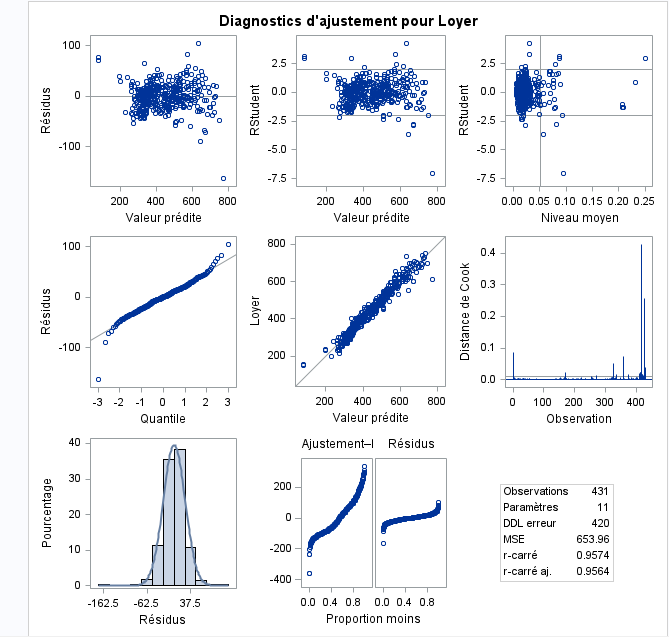
**B) Régression pour les concurrents d’Immobile**



Comme le montrent les résultats ci-dessus, la variable **Surface** a un coefficient estimé de **β=8.06**, ce qui signifie que pour chaque mètre carré supplémentaire, le loyer augmente en moyenne de **8.06 unités monétaires**, toutes choses égales par ailleurs. Cet effet est hautement significatif, avec une p-valeur de **<0.0001**, indiquant l'importance de cette variable dans la détermination du loyer, un résultat cohérent et attendu dans le secteur immobilier.

De même, la variable **Pm2** (prix par mètre carré) s'avère être un déterminant clé du loyer, avec un coefficient estimé de **β=21.52**. Cela suggère qu'une augmentation d'une unité monétaire du prix par mètre carré entraîne une augmentation moyenne de **21.52 unités monétaires** du loyer. La p-valeur associée est également **<0.0001**, confirmant la pertinence de cette variable. Ces résultats montrent clairement que les caractéristiques physiques et financières des biens immobiliers, telles que la surface et le prix par mètre carré, jouent un rôle essentiel dans l'explication des variations de loyer.

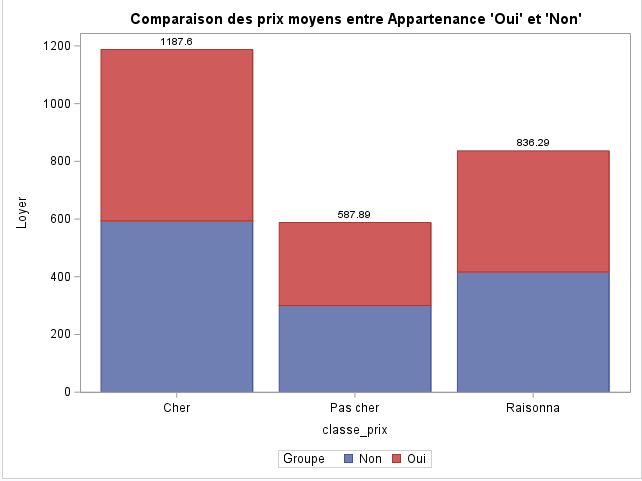
Par ailleurs, certaines variables catégoriques, telles que **pièces\_3** et **pièces\_5plus**, montrent des effets significatifs, avec des coefficients respectivement de **-16.41** et **-143.61**, ce qui peut indiquer une diminution des loyers pour ces catégories spécifiques. La significativité de ces variables est également soutenue par leurs p-valeurs respectives.



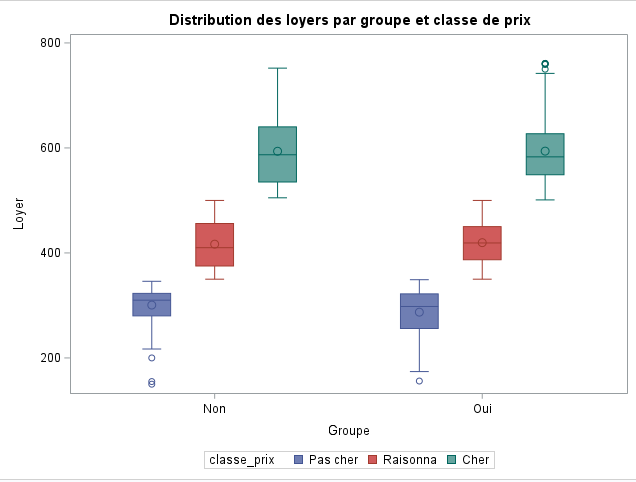
Notre modèle nous présente un excellent ajustement, avec un coefficient de détermination R2=0.9564, indiquant que 95,64% des variations des loyers sont expliquées par les variables du modèle (surface, nombre de pièces, prix au mètre carré, etc.). De plus, les tests statistiques montrent que ces variables sont significatives, avec des p-valeurs très faibles (<0.0001), ce qui confirme leur pertinence dans l'explication des loyers. Par ailleurs, les graphiques des résidus montrent une distribution approximativement normale.

Cependant, nous devons relever certaines limites. Les graphiques des résidus révèlent une légère structure en "U", ce qui suggère que le modèle pourrait ne pas capturer toutes les relations non linéaires ou interactions entre les variables.

En plus de cela , la distance de Cook met en évidence des valeurs aberrantes qui méritent une analyse approfondie. .



Les données montrent une différence entre les loyers moyens de Immobill ("Oui") et non Immobill ("Non"). Dans la catégorie "Cher", les biens Immobill affichent un loyer moyen de **1 187,6 €**, nettement supérieur à celui des biens non Immobill. Pour la catégorie "Raisonnable", les loyers moyens des biens Immobill atteignent **836,29 €**, restant légèrement au-dessus de ceux des biens non Immobill. Même dans la catégorie "Pas cher", Immobill maintient des loyers moyens plus élevés à **587,89 €**, De cette analyse nous pouvons conclure qu'Immobill privilégie des biens de forte valeur.



Les biens appartenant à Immobill ("Oui") propose un large panel en ce qui concerne le loyer, avec une forte présence dans la catégorie "Cher". Cela suggérent une orientation vers des biens haut de gamme ou situés dans des emplacements premium. À l'inverse, les biens opposées à Immobill ("Non") sont majoritairement concentrés dans les catégories "Pas cher" et "Raisonnable", ciblant un marché plus abordable, probablement adapté à des ménages à revenu moyen ou modeste. Cette répartition reflète une stratégie différenciée entre les deux groupes, Immobill propose donc un portefeuille plus varié et potentiellement haut de gamme, tandis que le groupe "Non" privilégie des loyers accessibles.

***ANNEXES***

DATA Immo1.Immob ;

Set Immo1.Immob ;

Pm2 = round(Pm2, 1);

nvie = round(nvie, 1);

Partso = round(Partso, 1);

PartMp = round(PartMp, 1);

Loyer = round(Loyer, 1) ;

Surface = round(Surface,1) ;

run ;

/\*supprimer les doublons \*/

proc sort data=Immo1.Immob nodupkey;

by Pi\_ces Loyer Surface Quartier Appartenance Pm2 ;

run;

data Immo1.Immob ;

set Immo1.Immob;

run;

/\*retirer la premiere ligne\*/

data Immo1.Immob;

set Immo1.Immob(firstobs=2);

run;

/\*traitement des valeurs aberrantes\*/

proc sgplot data=Immo1.Immo;

vbox Loyer / category=Pi\_ces;

title "Graphique de boîte à moustaches";

run;

proc sgplot data=Immo1.Immob;

vbox Surface / category=Pi\_ces;

title "Graphique de boîte à moustaches";

run;

proc sgplot data=Immo1.Immob;

vbox Pm2;

title "Graphique de boîte à moustaches";

run;

/\* Calcul des statistiques descriptives pour Loyer, Surface, et Pm2 \*/

proc univariate data=Immo1.Immob noprint;

var Loyer Surface Pm2; /\* Les variables continues à analyser \*/

output out=stats

p25=Loyer\_Q1 Surface\_Q1 Pm2\_Q1

p75=Loyer\_Q3 Surface\_Q3 Pm2\_Q3;

run;

/\* Ajout des bornes pour chaque variable \*/

data moustache;

set stats;

Loyer\_Lower = Loyer\_Q1 - 1.5 \* (Loyer\_Q3 - Loyer\_Q1);

Loyer\_Upper = Loyer\_Q3 + 1.5 \* (Loyer\_Q3 - Loyer\_Q1);

Surface\_Lower = Surface\_Q1 - 1.5 \* (Surface\_Q3 - Surface\_Q1);

Surface\_Upper = Surface\_Q3 + 1.5 \* (Surface\_Q3 - Surface\_Q1);

Pm2\_Lower = Pm2\_Q1 - 1.5 \* (Pm2\_Q3 - Pm2\_Q1);

Pm2\_Upper = Pm2\_Q3 + 1.5 \* (Pm2\_Q3 - Pm2\_Q1);

run;

data Immo1.Immob;

if \_N\_ = 1 then set moustache;

set Immo1.Immob;

if Loyer >= Loyer\_Lower and Loyer <= Loyer\_Upper and

Surface >= Surface\_Lower and Surface <= Surface\_Upper and

Pm2 >= Pm2\_Lower and Pm2 <= Pm2\_Upper;

run;

/\*Cacher les variables de quartile \*/

Data Immo1.Immob ;

set Immo1.Immob ;

drop Loyer\_Q3 Pm2\_Q3 Surface\_Q3 Surface\_Q3 Loyer\_Q1 Surface\_Q1 Pm2\_Q1 Loyer\_Lower Loyer\_Upper Surface\_Lower Surface\_Upper Pm2\_Lower Pm2\_Upper ;

run ;

/\*statistiqques descriptives\*/

Proc freq Data = Immo1.immob ;

table quartier ;

run ;

PROC MEANS DATA=Immo1.immob N MIN MAX MEAN MEDIAN STDDEV;

VAR surface pi\_ces pm2 loyer nvie Partso partMp;

RUN;

PROC MEANS DATA=Immo1.immob N MIN MAX MEAN MEDIAN STDDEV;

CLASS quartier;

VAR surface pi\_ces pm2 loyer nvie Partso partMp;

RUN;

PROC MEANS DATA=Immo1.immob N MIN MAX MEAN MEDIAN STDDEV;

CLASS pi\_ces;

VAR surface pm2 loyer nvie Partso partMp;

RUN;

PROC MEANS DATA=Immo1.immob N MIN MAX MEAN MEDIAN STDDEV;

CLASS pi\_ces;

VAR surface pm2 loyer nvie Partso partMp;

RUN;

PROC MEANS DATA=Immo1.immob N MIN MAX MEAN MEDIAN STDDEV;

CLASS appartenance;

VAR surface pi\_ces pm2 loyer nvie Partso partMp;

RUN;

PROC TABULATE DATA=Immo1.immob;

CLASS quartier pi\_ces;

VAR surface pm2 loyer nvie Partso partMp;

TABLE quartier,

pi\_ces\*(surface pm2 loyer nvie Partso partMp)\*(MEAN MEDIAN STDDEV);

RUN;

PROC TABULATE DATA=Immo1.immob;

CLASS quartier appartenance;

VAR surface pm2 loyer nvie Partso partMp;

TABLE quartier,

appartenance\*(N MEAN MEDIAN STDDEV)\*(surface pm2 loyer nvie Partso partMp);

RUN;

PROC MEANS DATA=Immo1.immob NOPRINT;

CLASS pi\_ces;

VAR pm2 surface loyer;

OUTPUT OUT=Immo1.min\_max\_pieces

MIN(pm2 surface loyer) = min\_pm2 min\_surface min\_loyer

MAX(pm2 surface loyer) = max\_pm2 max\_surface max\_loyer;

RUN;

PROC SGPANEL DATA=Immo1.immob;

PANELBY quartier;

HISTOGRAM surface / BINWIDTH=10;

RUN;

PROC SGPANEL DATA=Immo1.immob;

PANELBY quartier;

HISTOGRAM pm2 / BINWIDTH=2;

RUN;

PROC SGPANEL DATA=Immo1.immob;

PANELBY quartier;

HISTOGRAM loyer / BINWIDTH=100;

RUN;

/\*test de corrélation et covariance\*/

PROC CORR DATA=Immo1.immob COV;

VAR loyer;

WITH surface pm2 pi\_ces nvie PartMp Partso;

RUN;

PROC CORR DATA=Immo1.immob;

VAR surface loyer pm2 pi\_ces nvie PartMp Partso;

RUN;

/\* Remplacer les valeurs manquantes\*/

data Immo1.Immob;

set Immo1.Immob;

if missing(nvie) then nvie = 24975 ;

if missing(Partso) then Partso = 24 ;

if missing(PartMp) then PartMp = 13 ;

run;

/\*faire les variables dummies et de classes \*/

DATA Immo1.immob;

SET Immo1.immob;

quartier = LOWCASE(quartier);

armentieres = 0;

bois\_blancs = 0;

cormontaigne = 0;

fives = 0;

hellemes = 0;

la\_madeleine = 0;

lambersart = 0;

lille\_centre = 0;

lille\_sud = 0;

vieux\_lille = 0;

Lomme = 0;

loos = 0;

marcq\_en\_baroeul = 0;

mons = 0;

moulins = 0;

ronchin = 0;

saint\_maurice = 0;

tourcoing = 0;

vauban = 0;

villeneuve\_d\_ascq = 0;

wasquehal = 0;

wazemmes = 0;

IF quartier = "armentieres" THEN armentieres = 1;

ELSE IF quartier = "bois blancs" THEN bois\_blancs = 1;

ELSE IF quartier = "cormontaigne" THEN cormontaigne = 1;

ELSE IF quartier = "fives" THEN fives = 1;

ELSE IF quartier = "hellemes" THEN hellemes = 1;

ELSE IF quartier = "la madeleine" THEN la\_madeleine = 1;

ELSE IF quartier = "lambersart" THEN lambersart = 1;

ELSE IF quartier = "lille centre" THEN lille\_centre = 1;

ELSE IF quartier = "lille sud" THEN lille\_sud = 1;

ELSE IF quartier = "vieux lille" THEN vieux\_lille = 1;

ELSE IF quartier = "Lomme" THEN Lomme = 1;

ELSE IF quartier = "loos" THEN loos = 1;

ELSE IF quartier = "marcq en baroeul" THEN marcq\_en\_baroeul = 1;

ELSE IF quartier = "mons" THEN mons = 1;

ELSE IF quartier = "moulins" THEN moulins = 1;

ELSE IF quartier = "ronchin" THEN ronchin = 1;

ELSE IF quartier = "saint maurice" THEN saint\_maurice = 1;

ELSE IF quartier = "tourcoing" THEN tourcoing = 1;

ELSE IF quartier = "vauban" THEN vauban = 1;

ELSE IF quartier = "villeneuve d'ascq" THEN villeneuve\_d\_ascq = 1;

ELSE IF quartier = "wasquehal" THEN wasquehal = 1;

ELSE IF quartier = "wazemmes" THEN wazemmes = 1;

RUN;

DATA Immo1.immob;

SET Immo1.immob;

/\* Création des variables dummies pour l'appartenance \*/

appartenance\_oui = (Appartenance = "Oui");

appartenance\_non = (appartenance = "Non");

RUN;

DATA Immo1.immob;

SET Immo1.immob;

/\* Intervalles pour surface \*/

surface\_classe1 = (surface < 40);

surface\_classe2 = (40 <= surface < 55);

surface\_classe3 = (surface >= 55);

/\* Intervalles pour pm2 \*/

pm2\_classe1 = (pm2 < 8);

pm2\_classe2 = (8 <= pm2 <10);

pm2\_classe3 = (pm2 >= 10);

/\* Intervalles pour loyer \*/

loyer\_classe1 = (loyer < 350);

loyer\_classe2 = (350 <= loyer < 500);

loyer\_classe3 = (loyer >= 500);

Run ;

/\* Classes (1, 2, 3) basées sur les intervalles \*/

DATA Immo1.immob;

SET Immo1.immob;

/\* Création de la variable surface\_classe \*/

IF surface\_classe1 = 1 THEN surface\_classe = 1;

ELSE IF surface\_classe2 = 1 THEN surface\_classe = 2;

ELSE IF surface\_classe3 = 1 THEN surface\_classe = 3;

/\* Création de la variable pm2\_classe \*/

IF pm2\_classe1 = 1 THEN pm2\_classe = 1;

ELSE IF pm2\_classe2 = 1 THEN pm2\_classe = 2;

ELSE IF pm2\_classe3 = 1 THEN pm2\_classe = 3;

/\* Création de la variable loyer\_classe \*/

IF loyer\_classe1 = 1 THEN loyer\_classe = 1;

ELSE IF loyer\_classe2 = 1 THEN loyer\_classe = 2;

ELSE IF loyer\_classe3 = 1 THEN loyer\_classe = 3;

RUN;

DATA Immo1.immob;

SET Immo1.immob;

/\* Création des variables dummies pour le nombre de pièces \*/

pieces\_1 = (pi\_ces = 1);

pieces\_2 = (pi\_ces = 2);

pieces\_3 = (pi\_ces = 3);

pieces\_4 = (pi\_ces = 4);

pieces\_5plus = (pi\_ces IN (5, 6, 7, 8));

RUN;

proc freq data =Immo1.Immob ;

table surface\_classe1 surface\_classe2 surface\_classe3 pm2\_classe1 pm2\_classe2 pm2\_classe3

loyer\_classe1 loyer\_classe2 loyer\_classe3 pieces\_1 pieces\_2 pieces\_3 pieces\_4 pieces\_5plus;

run;

/\*séparation de la série en deux en fonction de l'appartennce \*/

DATA Immo1.immob\_oui Immo1.immob\_non;

SET Immo1.immob;

/\* Séparation en fonction de l'appartenance \*/

IF appartenance\_oui = 1 THEN OUTPUT Immo1.immob\_oui;

ELSE IF appartenance\_non = 1 THEN OUTPUT Immo1.immob\_non;

RUN;

/\* Régression pour les logements appartenant à Immobill \*/

Proc reg data=Immo1.Immob\_oui;

Model Loyer = Surface Pi\_ces Pm2 NVIE PartMp Partso

armentieres bois\_blancs cormontaigne fives hellemes la\_madeleine lambersart lille\_sud vieux\_lille loos

lomme marcq\_en\_baroeul mons moulins ronchin saint\_maurice tourcoing vauban villeneuve\_d\_ascq wasquehal wazemmes

pieces\_2 pieces\_3 pieces\_4 pieces\_5plus surface\_classe1 surface\_classe3 pm2\_classe1 pm2\_classe3

/ selection=stepwise slentry=0.05 slstay=0.05;

OUTPUT OUT=Immo1.Immob1 PREDICTED=Loyer\_predit RESIDUAL=Residu;

run;

/\* Régression pour les logements n'appartenant pas à Immobill \*/

Proc reg data=Immo1.Immob\_non;

Model Loyer = Surface Pi\_ces Pm2 NVIE PartMp Partso

armentieres bois\_blancs cormontaigne fives hellemes la\_madeleine lambersart lille\_sud vieux\_lille loos

lomme marcq\_en\_baroeul moulins ronchin saint\_maurice tourcoing vauban villeneuve\_d\_ascq wasquehal wazemmes

pieces\_2 pieces\_3 pieces\_4 pieces\_5plus surface\_classe1 surface\_classe3 pm2\_classe1 pm2\_classe3

/ selection=stepwise slentry=0.05 slstay=0.05;

OUTPUT OUT=work.Immob\_non\_predit PREDICTED=Loyer\_predit RESIDUAL=Residu;

run;

/\* Fusion des groupes pour analyse comparative \*/

DATA Immo1.Immob\_comparaison;

SET Immo1.Immob\_oui (IN=in\_oui) Immo1.Immob\_non (IN=in\_non);

IF in\_oui THEN Groupe = "Oui";

ELSE IF in\_non THEN Groupe = "Non";

/\* Classification des loyers \*/

IF Loyer < 350 THEN DO;

classe\_prix = "Pas cher";

qualite\_prix = 1;

END;

ELSE IF 350 <= Loyer <= 500 THEN DO;

classe\_prix = "Raisonnable";

qualite\_prix = 2;

END;

ELSE DO;

classe\_prix = "Cher";

qualite\_prix = 3;

END;

RUN;

/\* Visualisations et analyses comparatives \*/

PROC MEANS DATA=Immo1.Immob\_comparaison;

CLASS Groupe classe\_prix;

VAR Loyer;

TITLE "Comparaison des loyers moyens par classe de prix entre les groupes";

RUN;

PROC SGPLOT DATA=Immo1.Immob\_comparaison;

VBAR classe\_prix / GROUP=Groupe RESPONSE=Loyer STAT=MEAN DATALABEL;

TITLE "Comparaison des prix moyens entre Appartenance 'Oui' et 'Non'";

RUN;

PROC SGPLOT DATA=Immo1.Immob\_comparaison;

VBOX Loyer / CATEGORY=Groupe GROUP=classe\_prix;

TITLE "Distribution des loyers par groupe et classe de prix";

RUN;

PROC FREQ DATA=Immo1.Immob\_comparaison;

TABLES Groupe\*classe\_prix / CHISQ;

TITLE "Comparaison des proportions entre les classes de prix pour les groupes";

RUN;

PROC CORR DATA=Immo1.Immob\_comparaison PLOTS(MAXPOINTS=10000);

VAR Loyer qualite\_prix;

TITLE "Corrélations entre loyers et qualité de prix par groupe";

RUN;

proc print; run;