**Rapport de l'analyse statistique réalisée**

**Objectif**

L'objectif est de déterminer les facteurs qui expliquent les prix des logements. Une régression multiple a été effectuée sur les données afin de bien modéliser les prix en fonction des variables associées aux logements et aux propriétaires.

**Description des variables modifiées et créées**

La première étape de mon travail consiste à préparer les données, mais comme certaines variables contiennent beaucoup de virgules et de points, cela a causé de nombreux problèmes pour la lecture des données du programme SAS et du programme R. Je ne peux donc sélectionner manuellement certaines variables importantes via Excel comme variables à traiter. L'échantillon contient 2322 observations.

**Traitement des données manquantes**

Certaines variables ont également été modifiées en conséquence. Par exemple, la variable ***house\_since*** est passée d'une variable d'année à l’âge de propriété. De plus, l'échantillon contient trop de valeurs manquantes. Je voulais à l'origine effectuer une méthode d'imputation pour remplir les blancs, Comme l'une des variables de format avec plus de 600 valeurs manquantes est une combinaison linéaire de plusieurs autres variables, elle ne peut pas être renseignée. Ainsi, afin de garantir que le plus d'observations possible puissent être utilisées, j'ai choisi de supprimer cette variable.

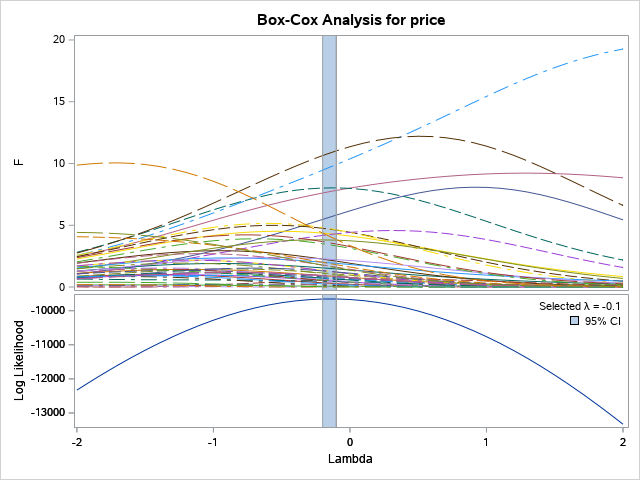
**Choix et description des méthodes statistiques**

Le premier modèle testé pour la régression multiple (voir le livre ***Classical and Modern Regression with Application de RAYMOND H. Myers*** pour plus d'information) est le modèle complet, soit le modèle qui contient les 24 variables présentées précédemment, excluant la variable réponse du ***Prix*** de logement.

Au cours de cette modélisation, la présence de plusieurs observations influentes a été remarquée. Ainsi, afin qu'elles soient moins influentes, chaque observation a été pondérée à l'aide la fonction Huber (voir le livre ***Robust Statistics de Peter J. Huber et Elvezio M. Ronchetti*** pour plus d'information) de la procédure ROBUSTREG de SAS (SAS Studio). Au total, un poids inférieur à 1 a été attribué a 404 observations, soit une proportion de 21.2 %.

En ce qui concerne les postulats du modèle de régression multiple, les résidus de la variable réponse ne respectait pas une distribution normale. En effet, comme Le diagramme boxcox montré dans la figure suivante, Sélection de la valeur lambda La valeur 0 est acceptable, ce qui signifie la transformation de la variable réponse suivante a donc été nécessaire :

**Figure:graphic de box-cox**



Par la suite, le choix du modèle de prédiction du délai judiciaire a été effectué. Voici le modèle général de régression multiple :

|  |
| --- |
| **Y = β0 + β1f1(x1) + β2f2(x2) + · · · + βpfp(xp) + fluctuation,** |

où Y dénote la variable endogène,

x1, . . . , xp représentent les variables exogènes,

f1, . . . , fp sont des transformations connues des variables exogènes et

β0, . . . , βp sont des paramètres de valeur inconnue qu’on estimera à l’aide des données.

La fluctuation indique que le modèle est stochastique, et non déterministe.

Trois méthodes ont été employées pour déterminer les variables qui seraient présentent dans le modèle final. Ces méthodes sont les suivantes : (1) critère BIC en considérant une approche par la méthode algorithmique d'inclusion (forward), (2) méthode de régularisation avec le LASSO à l'aide du critère BIC et (3) méthode basée sur la validation croisée, utilisée avec la méthode algorithmique pas à pas (stepwise).

**Tableau : Tableau des variables explicatives choisies par chacune des méthodes**

|  |  |  |
| --- | --- | --- |
| Méthode 1 | Méthode 2 | Méthode 3 |
| ***neighbourhood\_cleansed*** | ***neighbourhood\_cleansed*** | ***neighbourhood\_cleansed*** |
| ***property\_type*** | ***property\_type*** | room\_type |
| bathrooms\_text | ***minimum\_nights*** | ***longitude*** |
| ***longitude*** | bathrooms\_text | ***accommodates*** |
| ***accommodates*** | ***longitude*** | ***minimum\_nights*** |
| ***minimum\_nights*** | ***accommodates*** | ***availability\_365*** |
| ***availability\_365*** | ***availability\_365*** | ***property\_type*** |

Au total, 7 variables identiques ont été choisies par les trois méthodes. Il s'agit des variables en caractère gras. Donc, les variables du modèle final ont été choisies parmi les variables identiques obtenues avec les trois méthodes. C'est la procédure GLMSELECT de SAS (SAS Studio) qui a été utilisée pour faire le choix du modèle de prédiction. Finalement, le choix des variables à inclure dans le modèle final a été déterminé à l'aide du graphique de l'erreur quadratique moyenne.

**Figure : Graphique de la progression de l'erreur quadratique moyenne de la logarithme du prix de logement**

Chart, line chart

Description automatically generated

La figure montre que les 6 variables choisies précédemment doivent se retrouver dans le modèle final. En effet, le meilleur modèle est celui dont la valeur de l'erreur quadratique moyenne est la plus petite. Cette valeur est atteinte lorsque le modèle contient les 5 variables choisies précédemment.

L'ajout de certaines interactions au modèle de régression a été testé. Bien que certaines d'entre elles étaient significatives, elles complexifiaient le modèle pour seulement une faible augmentation du R². Il a donc été convenu qu'aucune interaction ne serait retenue dans le modèle final.

**Présentation et interprétation des résultats**

Puisque le jeu de données compte un pas grand nombre d'observations, le seuil de signification a été fixé à 5%. La valeur du R² du modèle retenu est de 0.54. Les tableaux à ANNEXE présentent les résultats de la régression multiple obtenus avec la procédure GLM de SAS (SAS STUDIO)

En effet, pour la variable ***Accommodates***, c'est-à dire Le nombre de personnes pouvant être hébergées dans la résidence, est associé à une augmentation moyenne de 0.1357277 dollars du logarithme du prix de logement. Par la formule, il est possible d'interpréter les résultats en terme du prix de logement et non du logarithme du prix de logement. Par exemple, le fait d’ajoute un espace d’une personne pour la résidence est associé à une augmentation moyenne de dollars du prix de logement.

De plus, un test global de régression a été effectué afin de connaître l'influence de l'ensemble des variables contenues dans le modèle final sur le délai judiciaire. Les hypothèses de ce test sont les suivantes :

**Tableau : Tableau du test global de la régression**

| **Source** | **DF** | **Sum of Squares** | **Mean Square** | **F Value** | **Pr > F** |
| --- | --- | --- | --- | --- | --- |
| **Model** | 84 | 584.772086 | 6.961572 | 31.36 | <.0001 |
| **Error** | 2219 | 492.670679 | 0.222024 |  |  |
| **Corrected Total** | 2303 | 1077.442765 |  |  |  |

Le tableau indique la valeur p est inférieure à 0.0001. Au seuil de 1 %, l'hypothèse nulle est rejetée. Cela signifie qu'au moins une des variables influence le prix.

Si nous voulons utiliser la régression linéaire, nous devons d'abord nous assurer que les quatre prémisses sont remplies : ***normalité, homogénéité, linéarité, indépendance***.

Nous pouvons juger que notre modèle final est entièrement conforme à la prémisse ci-dessus en examinant le graphique de QQ-plot et d'autres diagrammes de dispersion affichés à la processus PROC GLM.

**Figure :graphique sorti par PROC GLM**

**Diagram

Description automatically generated with medium confidence**

**Comparaison des variables catégoriques**

**Figure : graphique de comparaison de variable neighbourhood\_cleansed**

**Chart, scatter chart

Description automatically generated**

**Figure : graphique de comparaison de variable property\_type**A picture containing graphical user interface

Description automatically generated

Nous pouvons constater à partir des deux graphiques ci-dessus que différents emplacements de propriété et différentes catégories de maison auront des grands impacts sur le prix de logement.

**Conclusion**

En ce qui concerne notre objectif, les variables qui affectent le prix de location d'une maison sont le type de maison, l'emplacement de la maison, le nombre minimum de nuitées, la longitude, l'hébergement et le nombre de jours pouvant être loués d’un an.

**ANNEXE**

**Tableau : Paramètres estimés du modèle choisis pour la prévision du prix de logement**

| **Parameter** | **Estimate** |  | **Standard Error** | **t Value** | **Pr > |t|** |
| --- | --- | --- | --- | --- | --- |
| **Intercept** | 447.9879471 | B | 99.03267026 | 4.52 | <.0001 |
| **neighbourhood\_cleans Cap-Rouge** | 0.6121956 | B | 0.26034207 | 2.35 | 0.0188 |
| **neighbourhood\_cleans Chutes-Montmorency** | -0.9676409 | B | 0.10388157 | -9.31 | <.0001 |
| **neighbourhood\_cleans Cité Universitaire** | -0.1139592 | B | 0.12512627 | -0.91 | 0.3625 |
| **neighbourhood\_cleans Des Châtels** | 0.6976552 | B | 0.32226013 | 2.16 | 0.0305 |
| **neighbourhood\_cleans Duberger-Les Saules** | -0.1157770 | B | 0.17639366 | -0.66 | 0.5117 |
| **neighbourhood\_cleans Jésuites** | -0.1112356 | B | 0.20073746 | -0.55 | 0.5795 |
| **neighbourhood\_cleans L'Aéroport** | 0.3814537 | B | 0.27248019 | 1.40 | 0.1617 |
| **neighbourhood\_cleans Lac-Saint-Charles** | 0.5818029 | B | 0.26152877 | 2.22 | 0.0262 |
| **neighbourhood\_cleans Lairet** | -0.4093881 | B | 0.08695013 | -4.71 | <.0001 |
| **neighbourhood\_cleans Loretteville** | 0.1567182 | B | 0.24396636 | 0.64 | 0.5207 |
| **neighbourhood\_cleans Maizerets** | -0.3872760 | B | 0.06702927 | -5.78 | <.0001 |
| **neighbourhood\_cleans Montcalm** | -0.2547192 | B | 0.06026367 | -4.23 | <.0001 |
| **neighbourhood\_cleans Neufchâtel-Est/Lebourgneuf** | 0.1920659 | B | 0.16998046 | 1.13 | 0.2586 |
| **neighbourhood\_cleans Notre-Dame-des-Laurentides** | 0.1412979 | B | 0.19720293 | 0.72 | 0.4738 |
| **neighbourhood\_cleans Plateau** | -0.1701624 | B | 0.19089998 | -0.89 | 0.3728 |
| **neighbourhood\_cleans Pointe-de-Sainte-Foy** | 0.0595597 | B | 0.26052429 | 0.23 | 0.8192 |
| **neighbourhood\_cleans Quartier 4-2** | 0.0889785 | B | 0.19043935 | 0.47 | 0.6404 |
| **neighbourhood\_cleans Quartier 4-3** | -0.4563157 | B | 0.16525956 | -2.76 | 0.0058 |
| **neighbourhood\_cleans Quartier 4-5** | -0.2194533 | B | 0.14571063 | -1.51 | 0.1322 |
| **neighbourhood\_cleans Quartier 4-6** | -0.4954187 | B | 0.13585095 | -3.65 | 0.0003 |
| **neighbourhood\_cleans Quartier 5-1** | -0.5028596 | B | 0.15715457 | -3.20 | 0.0014 |
| **neighbourhood\_cleans Quartier 5-2** | -0.5814718 | B | 0.13558124 | -4.29 | <.0001 |
| **neighbourhood\_cleans Quartier 5-4** | -0.6692560 | B | 0.08168379 | -8.19 | <.0001 |
| **neighbourhood\_cleans Saint-Jean-Baptiste** | -0.2888065 | B | 0.04191967 | -6.89 | <.0001 |
| **neighbourhood\_cleans Saint-Louis** | -0.0445592 | B | 0.14710076 | -0.30 | 0.7620 |
| **neighbourhood\_cleans Saint-Roch** | -0.0917584 | B | 0.04122070 | -2.23 | 0.0261 |
| **neighbourhood\_cleans Saint-Sacrement** | -0.2343392 | B | 0.10486882 | -2.23 | 0.0255 |
| **neighbourhood\_cleans Saint-Sauveur** | -0.2764210 | B | 0.05873560 | -4.71 | <.0001 |
| **neighbourhood\_cleans Saint-Émile** | 0.0356589 | B | 0.27797424 | 0.13 | 0.8979 |
| **neighbourhood\_cleans Sillery** | -0.3130118 | B | 0.08778732 | -3.57 | 0.0004 |
| **neighbourhood\_cleans Val-Bélair** | 0.5517965 | B | 0.31999550 | 1.72 | 0.0848 |
| **neighbourhood\_cleans Vanier** | -0.3983639 | B | 0.12311108 | -3.24 | 0.0012 |
| **neighbourhood\_cleans Vieux-Limoilou** | -0.3940957 | B | 0.05590875 | -7.05 | <.0001 |
| **neighbourhood\_cleans Vieux-Moulin** | -0.3722101 | B | 0.08841521 | -4.21 | <.0001 |
| **neighbourhood\_cleans Vieux-Québec/Cap-Blanc/Colline parlementaire** | 0.0000000 | B | . | . | . |
| **property\_type Boat** | 0.9491037 | B | 0.64550443 | 1.47 | 0.1416 |
| **property\_type Camper/RV** | 0.1452289 | B | 0.48716647 | 0.30 | 0.7656 |
| **property\_type Entire apartment** | 0.1078091 | B | 0.34457707 | 0.31 | 0.7544 |
| **property\_type Entire bungalow** | 0.3802747 | B | 0.36389699 | 1.05 | 0.2961 |
| **property\_type Entire cabin** | 0.7322614 | B | 0.58339170 | 1.26 | 0.2095 |
| **property\_type Entire chalet** | 0.6721642 | B | 0.49957238 | 1.35 | 0.1786 |
| **property\_type Entire condominium** | 0.1219770 | B | 0.34630161 | 0.35 | 0.7247 |
| **property\_type Entire cottage** | 0.5664625 | B | 0.36810970 | 1.54 | 0.1240 |
| **property\_type Entire guest suite** | 0.2312521 | B | 0.36577579 | 0.63 | 0.5273 |
| **property\_type Entire guesthouse** | -0.0003554 | B | 0.36859485 | -0.00 | 0.9992 |
| **property\_type Entire home/apt** | 0.3419740 | B | 0.49178419 | 0.70 | 0.4869 |
| **property\_type Entire house** | 0.3375515 | B | 0.34717173 | 0.97 | 0.3310 |
| **property\_type Entire loft** | 0.0404437 | B | 0.34529973 | 0.12 | 0.9068 |
| **property\_type Entire place** | 0.0320468 | B | 0.40868276 | 0.08 | 0.9375 |
| **property\_type Entire serviced apartment** | 0.5588036 | B | 0.35310268 | 1.58 | 0.1137 |
| **property\_type Entire townhouse** | 0.5320417 | B | 0.35563263 | 1.50 | 0.1348 |
| **property\_type Houseboat** | 0.9000879 | B | 0.63455346 | 1.42 | 0.1562 |
| **property\_type Private room** | -0.0343549 | B | 0.58416439 | -0.06 | 0.9531 |
| **property\_type Private room in apartment** | -0.3073686 | B | 0.34630275 | -0.89 | 0.3749 |
| **property\_type Private room in bed and breakfast** | 0.2688943 | B | 0.35677612 | 0.75 | 0.4511 |
| **property\_type Private room in bungalow** | -0.3862043 | B | 0.35924300 | -1.08 | 0.2825 |
| **property\_type Private room in condominium** | -0.2633468 | B | 0.36776884 | -0.72 | 0.4740 |
| **property\_type Private room in cottage** | 0.1114216 | B | 0.44198237 | 0.25 | 0.8010 |
| **property\_type Private room in guest suite** | -0.6527711 | B | 0.63203305 | -1.03 | 0.3018 |
| **property\_type Private room in guesthouse** | 0.0268678 | B | 0.38079453 | 0.07 | 0.9438 |
| **property\_type Private room in hostel** | -0.3537403 | B | 0.37856824 | -0.93 | 0.3502 |
| **property\_type Private room in house** | -0.1898576 | B | 0.34712990 | -0.55 | 0.5845 |
| **property\_type Private room in loft** | 0.3196787 | B | 0.44302043 | 0.72 | 0.4706 |
| **property\_type Private room in minsu** | -0.5865379 | B | 0.58697115 | -1.00 | 0.3178 |
| **property\_type Private room in nature lodge** | 0.1628960 | B | 0.38003106 | 0.43 | 0.6682 |
| **property\_type Private room in serviced apartment** | 0.7722637 | B | 0.37725434 | 2.05 | 0.0408 |
| **property\_type Private room in townhouse** | -0.2055144 | B | 0.35876901 | -0.57 | 0.5668 |
| **property\_type Private room in villa** | -0.3163054 | B | 0.39431438 | -0.80 | 0.4225 |
| **property\_type Room in aparthotel** | 0.3457684 | B | 0.39564800 | 0.87 | 0.3823 |
| **property\_type Room in bed and breakfast** | 0.2811035 | B | 0.40427049 | 0.70 | 0.4869 |
| **property\_type Room in boutique hotel** | 0.4700853 | B | 0.35456170 | 1.33 | 0.1850 |
| **property\_type Room in hostel** | -0.8862849 | B | 0.43945024 | -2.02 | 0.0438 |
| **property\_type Room in hotel** | 0.4021614 | B | 0.35722412 | 1.13 | 0.2604 |
| **property\_type Room in nature lodge** | 0.1917884 | B | 0.58443730 | 0.33 | 0.7428 |
| **property\_type Room in serviced apartment** | 1.0384414 | B | 0.36854002 | 2.82 | 0.0049 |
| **property\_type Shared room** | -0.9127224 | B | 0.58397909 | -1.56 | 0.1182 |
| **property\_type Shared room in apartment** | -0.2852909 | B | 0.51112733 | -0.56 | 0.5768 |
| **property\_type Shared room in condominium** | 0.5996497 | B | 0.49322787 | 1.22 | 0.2242 |
| **property\_type Shared room in dorm** | -0.9220676 | B | 0.58409855 | -1.58 | 0.1146 |
| **property\_type Shared room in hostel** | -1.9632055 | B | 0.39608505 | -4.96 | <.0001 |
| **property\_type Shared room in house** | -0.1436035 | B | 0.58433083 | -0.25 | 0.8059 |
| **property\_type Tiny house** | 0.0000000 | B | . | . | . |
| **longitude** | 6.2301056 |  | 1.39020841 | 4.48 | <.0001 |
| **accommodates** | 0.1357277 |  | 0.00620699 | 21.87 | <.0001 |
| **minimum\_nights** | -0.0019830 |  | 0.00021294 | -9.31 | <.0001 |
| **availability\_365** | 0.0006995 |  | 0.00007934 | 8.82 | <.0001 |