



AH

Ames Housing

Prédiction du prix de vente des maisons – Ames Housing

*Modélisation par régression
linéaire*

Amee Hashley JEUDY
ameehashleyjeudy@gmail.com

Contexte & Objectif

- **Contexte :**
 - Marché immobilier de la ville d'Ames (Iowa, USA).
 - Base de données très utilisée pour les modèles de régression.
- **Objectif :**
 - Prédire le **prix de vente (SalePrice)** à partir de variables structurelles.
 - Comparer un **modèle simple** et un **modèle multiple** de régression linéaire.

Données utilisées

- **Variable cible :**
 - SalePrice (prix de vente en dollars)
- **Variables explicatives :**
 - GrLivArea (surface habitable au sol)
 - GarageArea (surface du garage)
 - LotArea (surface du terrain)
 - LotFrontage (largeur de façade)
- **Préparation :**
 - Nettoyage des valeurs manquantes
 - Vérification des types et des distributions.

Modèle simple

Formule :

$$\text{SalePrice} = b + m * \text{GrLivArea}$$

Résultats :

$R^2 \approx 0.50 \rightarrow$ environ 50 % de la variance expliquée.

Coefficient : +1 sq ft \approx **+118 \$** sur le prix.

Modèle statistiquement **significatif** ($p < 0.05$).

Modèle multiple

Formule :

$\text{SalePrice} \sim \text{GrLivArea} + \text{GarageArea} + \text{LotArea} + \text{LotFrontage}$

Résultats :

$R^2 \approx 0.62 \rightarrow$ amélioration nette du pouvoir explicatif.

LotFrontage non significatif ($p > 0.05$).

Autres variables pertinentes : GrLivArea reste dominante.

Interprétation

- La surface habitable est le facteur **le plus influent**.
- La surface du garage contribue positivement mais moins fortement.
- La surface totale du lot n'explique pas toujours mieux le prix.
- L'ajout de variables améliore la précision, **sans surapprentissage**.

Limites & Recommandations

- **Limites :**

- Données centrées sur une seule ville.
- Variables qualitatives (qualité, quartier, état) non intégrées.
- Risque de corrélation entre surfaces.

- **Pistes d'amélioration :**

- Ajouter des variables catégorielles.
- Tester des modèles non linéaires (Random Forest, XGBoost).
- Mettre en place une validation croisée et un tableau de bord Power BI/Tableau.

Conclusion

- Le modèle multiple offre une **meilleure explication** du prix de vente.
- GrLivArea reste la variable la plus déterminante.
- Les prochaines étapes concernent l'enrichissement et la généralisation du modèle.



AH

Ames Housing

Merci.



Amee Hashley JEUDY



+509 4846-4262



ameehashleyjeudy@gmail.com



<https://www.linkedin.com/in/amee-hashley-jeudy-460449325>