Examen final : analyse exploratoire des données sur Ames Housing

Objectif

Mener une analyse exploratoire, descriptive et inférentielle complète sur le dataset Ames Housing (Kaggle: House Prices – Advanced Regression Techniques). L'étudiant(e) a le libre choix de la méthode d'encodage des variables catégorielles (One-Hot, Label, Target, ou Ordinal avec justification).

Données et préparation

- Télécharger le fichier train.csv du projet House Prices – Advanced Regression Techniques (Ames Housing).
- Lien : https://www.kaggle.com/c/house-prices-advanced-regression-techniques/data
- Décrire brièvement le contexte et les variables principales (SalePrice, GrLivArea, OverallQual, Neighborhood, MSZoning, HouseStyle, etc.).

1. Exploration des données

- Charger le dataset et afficher 10 lignes aléatoires pour un aperçu.
- Identifier et traiter les doublons (compter, justifier la suppression éventuelle).
- Détecter des incohérences logiques (exemples: SalePrice = 0, LotArea = 0, années incohérentes) et expliquer comment vous les avez trouvées par programmation.
- Utiliser info(), describe() (inclure include='all' si pertinent) pour analyser types, cardinalités des catégorielles et statistiques descriptives des numériques.

2. Prétraitement des données

- Valeurs manquantes: lister les colonnes concernées, choisir une stratégie (suppression, imputation par moyenne/médiane/mode ou catégorie "None"), et justifier.

- Valeurs aberrantes: identifier sur les principales numériques (SalePrice, GrLivArea, LotArea, etc.) via boxplots/quantiles; traiter (capping par IQR/quantiles, transformation, ou exclusion) en justifiant l'impact attendu.
- Encodage des catégorielles:
    - Choisir librement la méthode (One-Hot, Label, Target, Ordinal).
    - Justifier le choix (lisibilité, compatibilité avec les tests/statistiques ou modèles).

3. Visualisations

Créer et interpreter chacune:

- Histogramme : distribution de SalePrice.
- Diagramme à barres : fréquence des quartiers (Neighborhood).
- Diagramme circulaire : répartition des styles de maison (HouseStyle).
- Boxplot : SalePrice par niveau de qualité (OverallQual).
- ScatterPlot : relation entre GrLivArea et SalePrice.

4. Loi normale

- Utiliser le test de Shapiro-Wilk pour vérifier si SalePrice suit une loi normale.
- Comparer la p-value au seuil de 0,05 et conclure.

5. Test d'hypothèse (sans utiliser GarageType)

Tester une différence significative de moyennes entre deux groupes pertinents de Neighborhood :

- Formuler H0 et H1 pour tester si les prix moyens des maisons diffèrent significativement entre deux quartiers : **CollgCr vs OldTown**.
- Réaliser un test de Student si les conditions sont remplies, sinon utiliser un test non paramétrique (Mann-Whitney).
- Interpréter les résultats en fonction de la p-value.

6. Statistiques inférentielles

- **Test du Chi-deux** : Construire un tableau de contingence entre les quartiers (Neighborhood) et les types de garage (GarageType). Appliquer le test du Chi-deux pour vérifier s'il existe une relation significative entre ces deux variables catégorielles. Interpréter la p-value pour conclure sur l'existence ou non d'une dépendance statistique entre quartier et type de garage.
- **ANOVA** : Comparer les prix moyens des maisons (SalePrice) selon les différents quartiers (Neighborhood). Utiliser une ANOVA à un facteur pour tester si les différences de prix entre quartiers sont statistiquement significatives. Interpréter la p-value pour conclure sur l'influence du quartier sur le prix des maisons.

7. Algèbre linéaire et probabilités

- Représenter une observation comme vecteur (LotArea, GrLivArea, OverallQual, SalePrice).
- Calculer les distances entre deux observations (euclidienne, manhattan, cosinus).
- Calculer la probabilité conditionnelle : P(SalePrice > seuil | OverallQual ≥ 7).

8. Conclusion

- Résumer les tendances : influence du quartier, de la surface et de la qualité sur le prix.
- Identifier les variables les plus discriminantes pour prédire SalePrice.