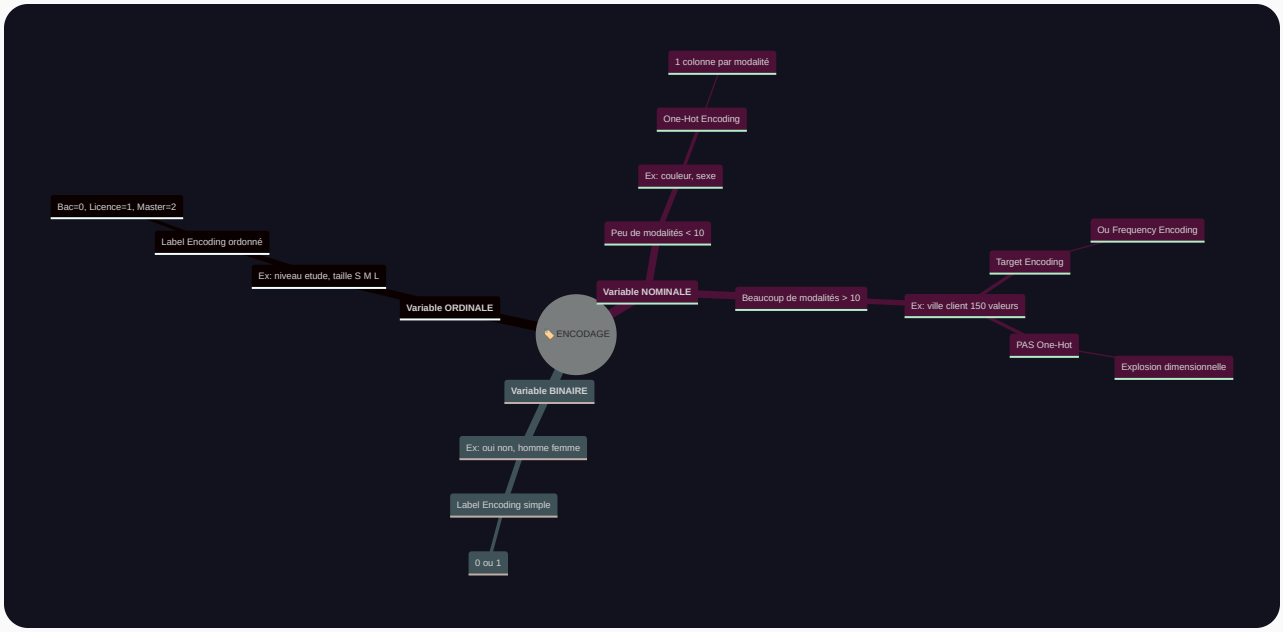
🧠 Mind Maps — "Si tu vois X → Pense Y"

Pré-traitement des données — Rattrapage — Pr. YOUNES DAOUI

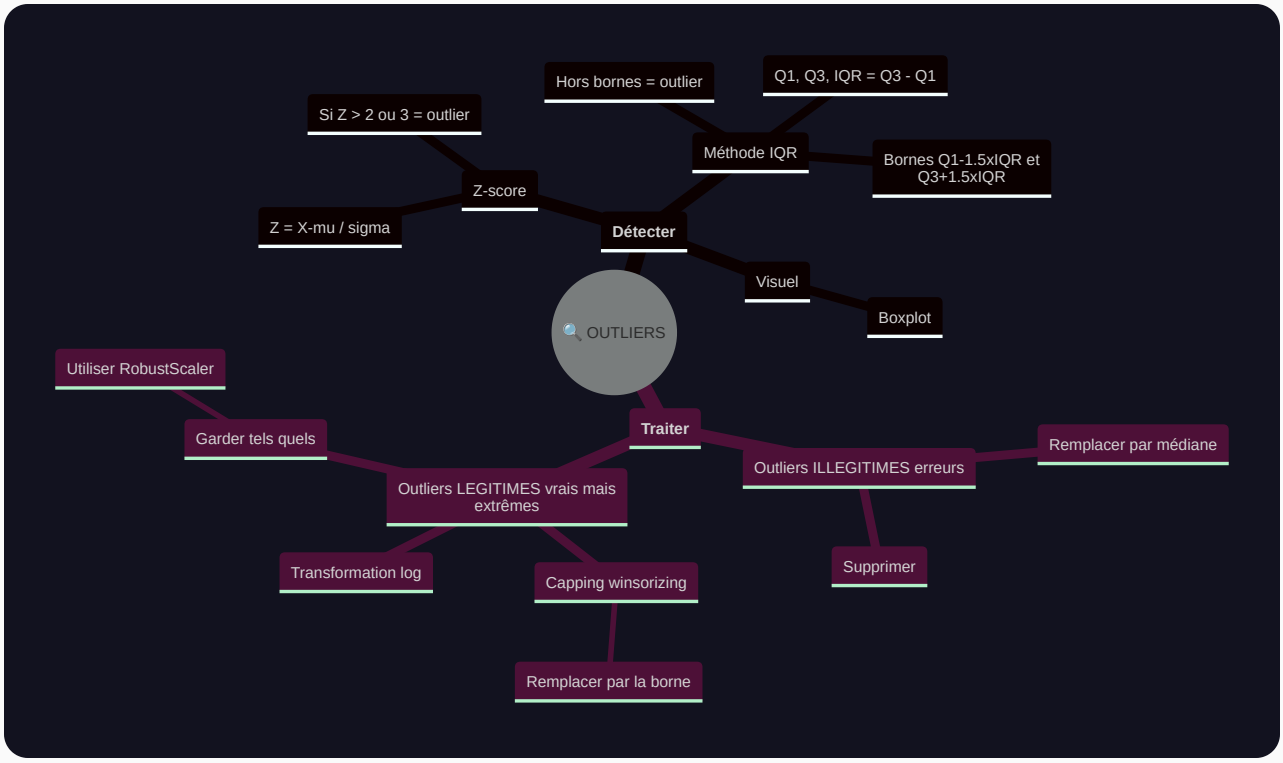
🏷️ 1. Encodage des variables



⚡ Réflexes rapides

Si tu vois en exam...	Pense immédiatement à...
"150 modalités" / "beaucoup de catégories"	❌ One-Hot → ✅ Target Encoding / Frequency Encoding
"ordre naturel" / "hiérarchie" / "Bac < Master"	✅ Label Encoding **ordonné**
"pas d'ordre" / "nominale" / "couleurs"	✅ One-Hot Encoding
"fausse relation d'ordre"	= piège du Label Encoding sur du nominal
"curse of dimensionality"	= trop de colonnes → One-Hot sur trop de modalités

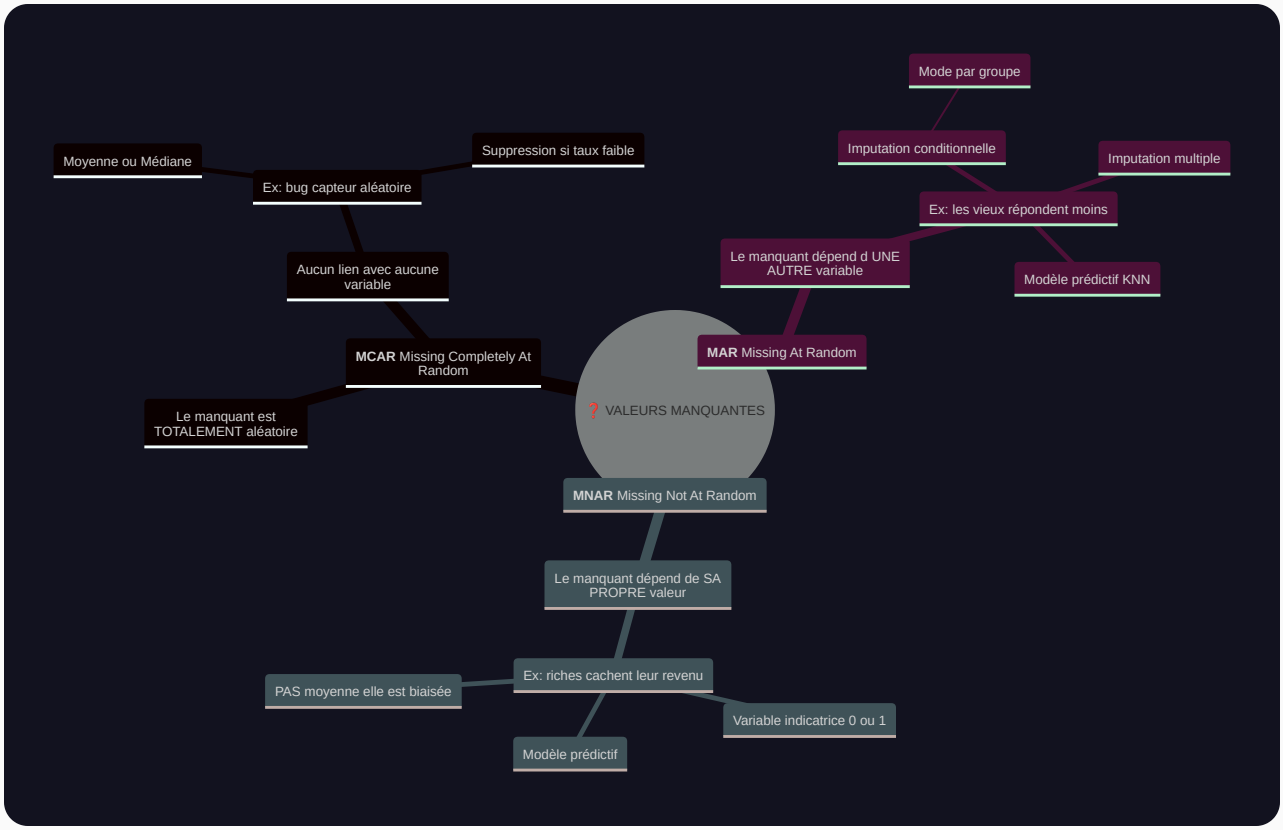
🔍 2. Détection & Traitement des Outliers



⚡ Réflexes rapides

Si tu vois en exam...	Pense immédiatement à...
"calculez les outliers"	→ Méthode IQR : trier, Q1, Q3, bornes
"outliers légitimes" / "valeurs extrêmes mais correctes"	→ **RobustScaler** (médiane + IQR)
"deux techniques de détection"	→ IQR + Z-score (les 2 classiques)
"proposez des stratégies de traitement"	→ Suppression, Capping, Remplacement médiane, Log
Boxplot dans l'énoncé	→ Lire les moustaches = bornes, points isolés = outliers

❓ 3. Valeurs Manquantes (MCAR / MAR / MNAR)

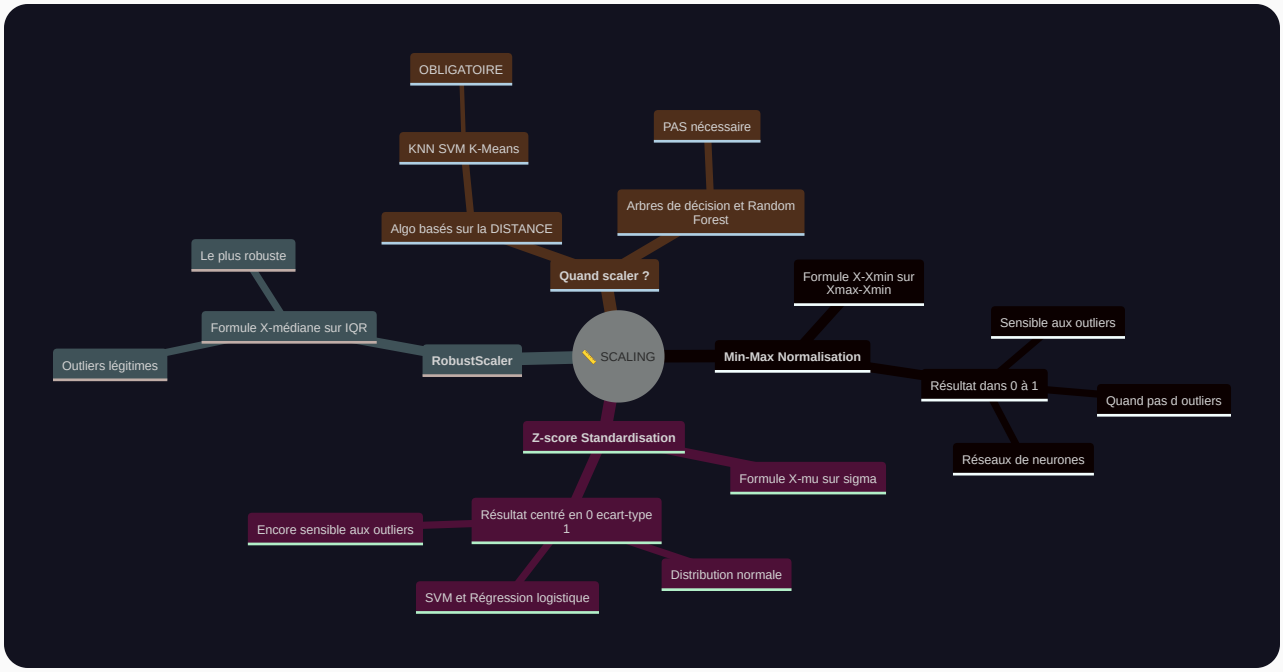


⚡ Réflexes rapides

Si tu vois en exam...	Pense immédiatement à...
"refusent de déclarer **leur revenu**"	→ **MNAR** (dépend de la valeur elle-même)
"dépend de **l'âge** / d'une autre variable"	→ **MAR** (dépend d'une variable observée)
"totalement aléatoire" / "bug" / "hasard"	→ **MCAR**
"92% manquant" / taux très élevé	→ ❌ Pas supprimer → Recoder en binaire ou indicatrice
"distribution asymétrique"	→ Utiliser **médiane** pas moyenne
"variable catégorielle manquante"	→ **Mode** (pas moyenne !)
"n'ont jamais utilisé de coupon"	→ Le manquant = "pas de coupon" → Recoder en 0/1

"créer une variable indicatrice"	→ Quand le fait d'être manquant est **informatif**

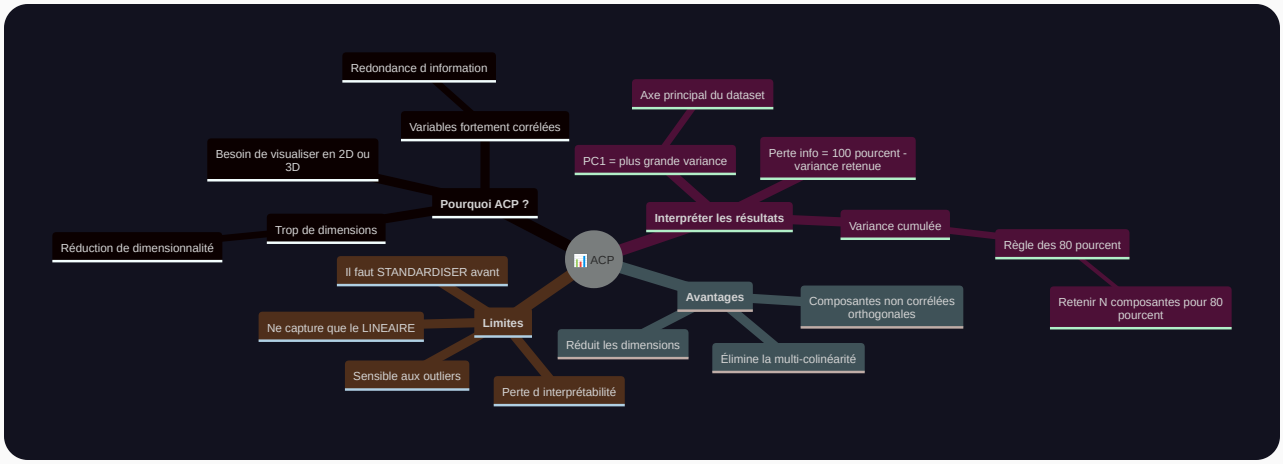
📏 4. Mise à l'Échelle (Scaling)



⚡ Réflexes rapides

Si tu vois en exam...	Pense immédiatement à...
"KNN" / "distance" / "SVM" / "K-Means"	→ Scaling **obligatoire**
"outliers légitimes"	→ **RobustScaler**
"une variable domine les distances"	→ Il faut scaler ! (ex: revenu vs âge)
"Expliquez la formule"	→ Écrire Min-Max ET Z-score
"Arbre de décision" / "Random Forest"	→ Scaling **pas nécessaire**
"appliquez le scaling sur X = ..."	→ Calculer avec la formule (brancher les valeurs)

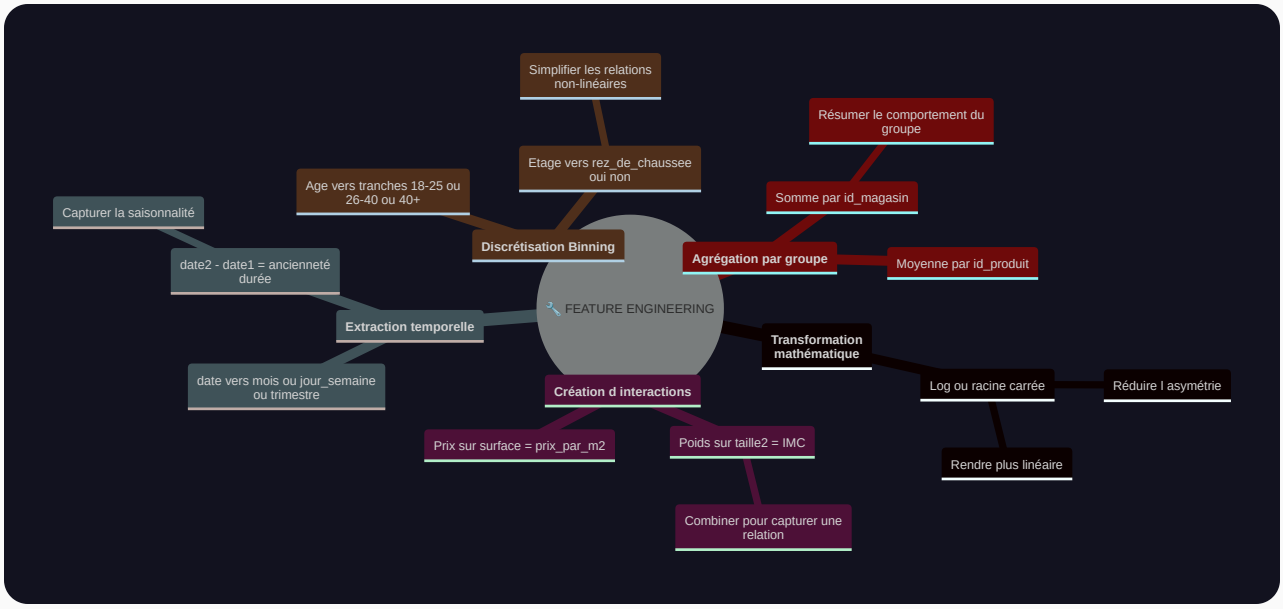
📊 5. ACP (Analyse en Composantes Principales)



⚡ Réflexes rapides

Si tu vois en exam...	Pense immédiatement à...
"hautement corrélés" / "redondance"	→ ACP pour éliminer la redondance
"espace réduit" / "visualiser en 2D"	→ Retenir PC1 + PC2
"PC1 = 60%, PC2 = 22%, PC3 = 9%"	→ Cumuler ! 60→82→91. Dire combien pour ≥80%
"interprétez les résultats"	→ 4 choses : variance cumulée, nb composantes, perte info, interprétation métier
"perte d'information"	→ = 100% - variance cumulée des PCs retenues
"anomalies" + ACP	→ Points éloignés dans l'espace réduit = anomalies
"limites de l'ACP"	→ Linéaire, perte interprétabilité, sensible outliers

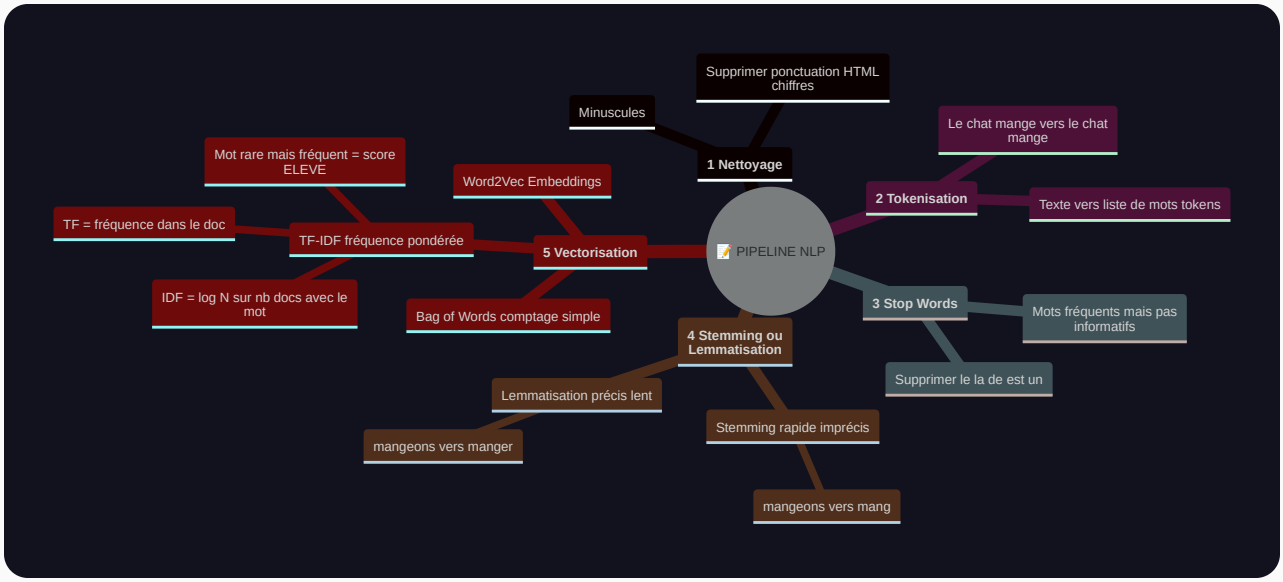
🔧 6. Feature Engineering



⚡ Réflexes rapides

Si tu vois en exam...	Pense immédiatement à...
Une variable `date`	→ Extraire : mois, jour_semaine, année, trimestre, est_weekend
Deux dates	→ Calculer la **durée** entre les deux
`prix` + `quantité`	→ prix × quantité = revenu, prix / quantité = prix unitaire
`surface` + `nb_pièces`	→ surface / nb_pièces = taille moyenne par pièce
Variable très asymétrique	→ Transformation **log**
"améliorer la performance du modèle"	→ Créer des features ! (ratio, agrégation, temporel, binaire)

📝 7. Pipeline NLP (Données Textuelles)



⚡ Réflexes rapides

Si tu vois en exam...	Pense immédiatement à...
"pipeline textuel" / "données textuelles"	→ 6 étapes : Nettoyage → Tokenisation → Stop words → Stemming/Lemma → Vectorisation
"TF-IDF"	→ TF × IDF. Mot rare dans le corpus mais fréquent dans le doc = score élevé
"stemming vs lemmatisation"	→ Stemming = racine (rapide, imprécis) / Lemma = forme canonique (précis, lent)
"rôle de chaque étape"	→ Expliquer POURQUOI on fait chaque étape (pas juste la lister)

🎯 Arbre de Décision Global — Face à un problème d'exam