

Compte Rendu

Analyse du Dataset "MoMA Museum Collection - Artworks & Artists"

Date : 26 Novembre 2025

Auteur :	Younes El Baz
Numéro Étudiant :	22007219
Cours :	CAC 2

Table des matières

1	Introduction et Contexte	3
2	Chargement & Description des Données	3
2.1	Dimensions et Structure	3
2.2	Problèmes Initiaux	3
3	Nettoyage et Prétraitement	3
3.1	Gestion des Valeurs Manquantes	3
3.2	Ingénierie de Caractéristiques (Feature Engineering)	3
4	Analyse Exploratoire (EDA)	3
4.1	Distribution des Âges à l'Acquisition	3
4.2	Distribution Spatiale et Temporelle	4
5	Analyse Statistique & Corrélations	4
5.1	Corrélation entre l'Âge à l'Acquisition et les Caractéristiques	4
6	Analyse des Tendances Temporelles	4
6.1	Évolution des Acquisitions du MoMA	4
7	Exploration des Médiums & Catégories	4
7.1	La Prédominance du Print et de la Photographie	4
8	Modélisation & Approches Prédictives	4
8.1	Objectif	4
8.2	Approche et Résultats	4
9	Interprétation & Recommandations	5
9.1	Ce que révèle l'analyse :	5
9.2	Recommandations pour aller plus loin :	5
10	Conclusion	5

1. Introduction et Contexte

Le dataset **MoMA Museum Collection** regroupe plus de **150 000 œuvres** du *Museum of Modern Art* de New York. Cette analyse vise à explorer la richesse des données et à identifier les tendances et les relations clés au sein de la collection, notamment les dynamiques d'acquisition et les caractéristiques des œuvres.

2. Chargement & Description des Données

2.1. Dimensions et Structure

- Plus de 150 000 lignes (œuvres).
- Environ 50 colonnes initiales, incluant des informations sur l'artiste, l'œuvre, les dates, les dimensions, le médium, et les détails d'acquisition.

2.2. Problèmes Initiaux

- Nombre important de valeurs manquantes (e.g., dans les dimensions, les nationalités des artistes).
- Incohérences de formatage (dates, descriptions textuelles libres des médiums).

3. Nettoyage et Prétraitement

3.1. Gestion des Valeurs Manquantes

- Imputation pour les colonnes numériques (moyenne/médiane) et catégorielles (mode ou "Inconnu").
- Nettoyage intensif des dates pour créer des variables numériques exploitables (*YearCreated*, *AcquisitionYear*).

3.2. Ingénierie de Caractéristiques (Feature Engineering)

- Création d'une variable clé : *AgeAtAcquisition*, qui représente le temps écoulé entre la création de l'œuvre et son acquisition par le MoMA.

$$\text{AgeAtAcquisition} = \text{AcquisitionYear} - \text{YearCreated}$$

- Standardisation des descriptions de médiums pour regrouper les catégories.

4. Analyse Exploratoire (EDA)

4.1. Distribution des Âges à l'Acquisition

- Une majorité d'œuvres sont acquises peu de temps après leur création (distribution fortement asymétrique).
- Cependant, la queue de distribution révèle des acquisitions tardives d'œuvres historiques.

4.2. Distribution Spatiale et Temporelle

- **Nationalité** : Forte dominance des artistes américains, suivie par les artistes européens (français, allemands).
- **Dates de Création** : Pic d'œuvres créées au milieu du XX^e siècle (Art Moderne).

5. Analyse Statistique & Corrélations

5.1. Corrélation entre l'Âge à l'Acquisition et les Caractéristiques

- Corrélation faible entre les dimensions (hauteur, largeur) et *l'AgeAtAcquisition*.
- Le *Médium* (catégoriel) semble être le facteur le plus discriminant.

6. Analyse des Tendances Temporelles

6.1. Évolution des Acquisitions du MoMA

- On observe des périodes d'acquisition intenses, souvent en ligne avec des changements de politique muséale ou des périodes d'expansion.
- Les acquisitions récentes montrent un effort de diversification.

7. Exploration des Médiums & Catégories

7.1. La Prédominance du Print et de la Photographie

- Le Top 3 des médiums : **Print**, **Photography**, et **Drawing**.
- Ces catégories représentent une part majoritaire de la collection, confirmant une orientation spécifique du MoMA.

8. Modélisation & Approches Prédictives

8.1. Objectif

Tenter de prédire *l'AgeAtAcquisition* d'une œuvre en fonction de ses caractéristiques (médium, nationalité, année de création).

8.2. Approche et Résultats

- Utilisation de modèles de Régression (e.g., Régression Linéaire, Random Forest).
- Le modèle Random Forest offre les meilleures performances, confirmant :
 - forte importance du médium
 - influence modérée de la nationalité et de l'année de création

9. Interprétation & Recommandations

9.1. Ce que révèle l'analyse :

- La collection est très riche en **photographies et prints**.
- Les artistes modernes du XX^e siècle dominent numériquement.
- Les acquisitions suivent des politiques évolutives selon les décennies.
- Les caractéristiques des œuvres n'expliquent pas toujours bien l'âge à l'acquisition (*dataset* très artistique, non linéaire).

9.2. Recommandations pour aller plus loin :

1. Analyser les artistes sous-représentés (*female artists, nationalités émergentes*).
2. Étudier les réseaux artistiques (graphes artistes-médiums).
3. Développer un modèle de classification des médiums.
4. Mesurer l'évolution de la diversité artistique du MoMA.
5. Étudier comment la collection reflète l'évolution de l'art moderne et contemporain.

10. Conclusion

Cette étude du dataset du **MoMA Museum Collection** montre :

- ✓ Une collection immense et diversifiée
- ✓ Une forte orientation vers la photographie et les impressions
- ✓ Des dynamiques d'acquisition liées aux périodes historiques
- ✓ Une richesse analytique exploitant artistes, médiums, dimensions et dates
- ✓ Des possibilités de modélisations avancées malgré la nature artistique des données

L'analyse constitue une **base solide** pour aller plus loin dans la compréhension de l'évolution artistique du MoMA.