
Projet d'Analyse de Données

À vous de mener en **autonomie** un projet d'analyse de données !

Vous êtes libres de vos questionnements, une attention particulière sera portée lors de l'évaluation sur la motivation de ceux-ci.

Option n°1 : Typologie des jeux vidéo et succès sur Steam

Le but de ce projet est d'étudier la base de données disponible sur **kaggle** : Steam Games Dataset

kaggle.com/datasets/fronkongames/steam-games-dataset/data

Pour vous aider, quelques questions que vous pouvez *par exemple*¹ vous poser :

- Quelle est la qualité du jeu de données disponible ?
- Existe-t-il une typologie des jeux vidéo sur Steam ?
- Peut-on détecter des jeux atypiques (outliers) ?
- Quels facteurs sont associés au succès commercial des jeux ?
- Peut-on mettre en évidence des axes latents permettant de structurer l'espace des jeux vidéo (casual *vs.* hardcore, indie *vs.* AAA, narratif *vs.* action, *etc.*) ?
- Peut-on regrouper les jeux en classes homogènes ? Comment interpréter ces classes ?
- Les genres déclarés par les développeurs sont-ils cohérents avec la typologie obtenue par les méthodes statistiques ?
- Peut-on identifier des segments de marché distincts sur Steam ?
- Les jeux indépendants se distinguent-ils statistiquement des jeux AAA ?
- Existe-t-il une relation entre prix, popularité et genre ?
- Peut-on interpréter les clusters comme des niches de marché ?
- Comment la structure du marché évolue-t-elle avec l'année de sortie ?

On pourra, au besoin, appliquer des transformations aux données quantitatives, voire les transformer en variables catégorielles pour faciliter leur traitement. Ces choix devront être motivés.

Option n°2 : Style de musique et popularité

Le but de ce projet est d'étudier la base de données disponibles sur **kaggle** : 30000 Spotify Songs

kaggle.com/datasets/joebeachcapital/30000-spotify-songs

Pour vous aider, quelques questions que vous pouvez *par exemple*¹ vous poser :

- Quelle est la qualité du jeu de données disponible ?
- Existe-t-il une structure latente de la musique populaire sur Spotify ?
- Peut-on mettre en évidence des axes latents permettant de structurer l'espace des morceaux (dansabilité, énergie, complexité, *etc.*) ?
- Peut-on regrouper les morceaux en classes homogènes ? Comment interpréter ces classes ?

1. En particulier, il n'est pas requis de répondre à toutes ces questions.

- La popularité est-elle associée à certaines caractéristiques musicales ? Dépend-elle de leur style ?
- Les genres musicaux déclarés sont-ils cohérents avec la typologie obtenue par des méthodes statistiques ?
- Peut-on identifier des segments distincts de musique populaire (mainstream, niche, expérimental, etc.) ?
- Comment la structure de la musique populaire évolue-t-elle au cours du temps ?
- Comment les différentes mesures influent-elles sur l'appréciation du morceau considéré ?

On pourra, au besoin, appliquer des transformations aux données quantitatives, voire les transformer en variables catégorielles pour faciliter leur traitement. Ces choix devront être motivés.

Option n°3 : Le jeu de données de votre choix

Vous pouvez également choisir d'étudier un autre jeu de données pour ce projet.

- Votre jeu de données devra cependant être suffisamment complexe pour vous permettre de réinvestir la plupart des méthodes vues en cours ce semestre.
- Des exemples de sites pour trouver des données :
 - `kaggle` : [kaggle.com](https://www.kaggle.com)
 - `Hugging Face` : huggingface.co
 - `Our World in Data` : ourworldindata.org
 - La plateforme des données publiques françaises : [data.gouv](https://data.gouv.fr)
 - ...

*Merci dans ce cas de valider votre choix auprès de l'équipe enseignante **a priori**.*

Remarque importante : Il n'est pas interdit de réinvestir des méthodes vues dans d'autres cours (EMS et/ou Machine Learning typiquement). Ces méthodes ne doivent cependant pas se substituer à celles vues en cours d'Analyse de Données.

Évaluation du projet

Critères	Poids	Indicateurs attendus
Compréhension et Motivation du questionnement	5%	Clarté des questions posées, de la problématique étudiée, Justification du choix du jeu de données, Cohérence scientifique.
Prétraitement et Qualité des données	10%	Gestion des valeurs manquantes et aberrantes, Transformations et justification des choix.
Analyse factorielle	20%	Choix de la méthode (PCA, MCA, FAMD, <i>etc.</i>), Interprétation des axes, Qualité des visualisations.
Clustering	20%	Choix de la méthode, Justification du nombre de clusters, Validation (silhouette, inertie, <i>etc.</i>), Interprétation.
Interprétation et Prise de recul scientifique	20%	Capacité à expliquer les choix méthodologiques, Compréhension et interprétation des résultats en lien avec la problématique, Discussion critique des résultats, Prise de recul, Qualité des réponses aux questions.
Qualité de la soutenance	10%	Qualité pédagogique de la présentation, Clarté du propos, Structure de l'exposé, Figures pertinentes, lisibles et soignées.
Reproductibilité (Git)	15%	Code reproductible, Données en libre accès, Organisation du répertoire Git et des fichiers, Références, Travail sourcé.
Bonus	+5%	Analyse temporelle, Tests statistiques Méthodes avancées (UMAP, Bootstrap, <i>etc.</i>), Originalité.
Malus	-5%	Dépassement du temps lors de la soutenance, Mauvaise répartition du temps de parole entre les membres du groupe (soutenance et questions).

Remarques :

- Une attention particulière sera portée à la justification des choix méthodologiques.
- Les méthodes avancées ne doivent pas se substituer à celles vues en cours.
- L'originalité et la prise de recul critique seront valorisées.

Livrables :

1. Une soutenance orale par groupe de 3 à 4 étudiant-es.
 - Chaque membre du groupe doit être capable de répondre aux questions sur l'ensemble du projet.
2. Un dépôt **Git** contenant l'ensemble des codes nécessaires à la reproduction des analyses et des figures.
 - Un fichier **README** décrivant l'organisation du dépôt est attendu.
 - La reproductibilité des résultats sera évaluée.
 - Tout membre du groupe doit être en mesure d'expliquer l'ensemble des codes fournis.