|  |  |
| --- | --- |
|  | **SCI 1402** |
| Projet en science des données |
|  |

**Travail noté 1**

Plan de projet documenté (10%)

Feuille d’identité

Nom Prévost Prénom Jean-Marie

Numéro d’étudiant 22202685 Trimestre Hiver 2025

Adresse 4-196 rue Saint-Gérard, Saint-Jean-sur-Richelieu

      Code postal J2W 2L9

Téléphone Domicile s/o Travail s/o

Cellulaire 450-357-3043

Courriel prevost.jean-marie@univ.teluq.ca

Nom du professeur ou son délégué Richard Hotte

**Réservé à l’usage du professeur ou son délégué**

Date de réception       Date de retour

Note

Date d’envoi

# Table des matières

[Table des matières 2](#_Toc191722284)

[1. Introduction 3](#_Toc191722285)

[2. Contexte et motivations 3](#_Toc191722286)

[3. Portée et limites du projet 3](#_Toc191722287)

[3.1. Objectifs du projet 3](#_Toc191722288)

[3.2. Limites et contraintes 4](#_Toc191722289)

[4. Méthodologie 4](#_Toc191722290)

[4.1. Choix des jeux de données et justifications 4](#_Toc191722291)

[4.2. Traitements des données 4](#_Toc191722292)

[4.3. Analyse et modélisation 5](#_Toc191722293)

[4.3.1. Exploration et visualisation descriptive 5](#_Toc191722294)

[4.3.2. Analyse relationnelle entre variables 5](#_Toc191722295)

[4.3.3. Segmentation et modélisation exploratoire 5](#_Toc191722296)

[4.3.4. Projections et optimisations 5](#_Toc191722297)

[4.4. Outils et technologies envisagés 6](#_Toc191722298)

[5. Calendrier du Projet 7](#_Toc191722299)

[6. Défis anticipés et plan de contingence 8](#_Toc191722300)

[7. Conclusion 8](#_Toc191722301)

[8. Bibliographie 9](#_Toc191722302)

# Introduction

Ce document décrit l’exploration et la planification initiale du projet « Outil d’analyse de marché de la plateforme Steam » dans le cadre du cours SCI 1402, en clôture du programme « 4875, certificat en science de donnée », sous la supervision de Fatima Bensalma.

# Contexte et motivations

L’industrie du jeu vidéo est un milieu volatile dans lequel s’affrontent des tendances passagères, des courants soutenus, des événements ponctuels et des fluctuations saisonnières. Il s’agit également d’un secteur qui repose en grande partie sur ses succès majeurs pour compenser aux coûts de multiples échecs. (Thai, 2024)

La nature du médium rend difficiles les réorientations majeures, et il n’est pas rare de voir même des géants de l’industrie tenter d’exploiter une tendance pour finalement livrer un produit décalé du contexte réel à sa sortie quelques années plus tard — Battleborn par Gearbox Software ou Anthem par BioWare sont de forts exemples relativement récents.

Ayant moi-même œuvré dans le monde du développement de jeux indépendants, il m’a toujours intéressé de pouvoir brosser le portrait réel de l’industrie, d’en explorer les courants, voire d’en prédire la direction. Ce projet s’aligne donc tant aux exigences du cours qu’à cette curiosité personnelle.

# Portée et limites du projet

Ce projet vise à développer un outil d’analyse de marché pour la plateforme Steam sous forme de tableau de bord ou d’interface interactive. Il exploitera des ensembles de données accessibles sur Kaggle, rendus possibles par des sources telles que l’API Steam, SteamCharts, SteamSpy et SteamDB. (Riad, 2021)

Il permettra d’examiner les tendances du marché en combinant des données historiques et récentes (avec un décalage de quelques mois).

## Objectifs du projet

Étant difficile à estimer avec précision, le projet est conçu en niveaux progressifs, chacun s’appuyant sur le précédent. Selon ce que les ressources permettront, le projet vise les objectifs suivants :

1. **Analyse rétrospective des tendances du marché.**
   * Vue d’ensemble des tendances historiques de l’industrie.
   * Options de filtrage simples et, si possible, analyses avancées de type OLAP (slice, dice, drill down). (Varma, 2024)
   * Introduction d’intelligence d’affaires (BI) et d’analytique. (Bay Atlantic University, 2025)
2. **Analyse relationnelle et modélisation statistique**
   * Étude des relations entre différents facteurs (prix, popularité, évaluations).
   * Identification des éléments favorables ou défavorables à une métrique donnée.
3. **Développements supplémentaires (si le temps le permet)**
   * Projection des ventes via un modèle d’apprentissage automatique.
   * Recommandations optimisées pour un produit spécifique.
   * Évaluation des risques et du potentiel commercial d’un jeu (revenu, concurrence, volatilité).
   * Cartographie de marchés sous-exploités ou à fort potentiel.

## Limites et contraintes

Ce projet présente plusieurs contraintes importantes :

* **Données figées dans le temps** : L’outil repose sur des ensembles de données préexistants, sans recourir à un « scraping » direct de Steam.
* **Fiabilité des données** : Bien que les sources utilisées soient réputées, certaines métriques, comme les chiffres de ventes, sont des estimations. Toute projection sera donc sujette à des marges d’erreur additionnelles. (Galyonkin, 2021)
* **Facteurs externes non pris en compte** : Les tendances du marché étant influencées par des événements ponctuels (lancements majeurs, phénomènes viraux), ces fluctuations ne seront pas intégrées à l’analyse.
* **Contraintes de temps et d’expérience** : La portée du projet dépendra des ressources disponibles et de mon apprentissage progressif des outils et techniques utilisées.

Malgré ces défis, l’objectif est de produire un outil fonctionnel et pertinent, particulièrement utile aux développeurs indépendants cherchant à mieux comprendre le marché et à prendre des décisions éclairées.

# Méthodologie

## Choix des jeux de données et justifications

Les jeux de données seront sélectionnés en fonction de leur récence afin d'assurer une pertinence dans un contexte réel. L'accent sera mis sur la complémentarité des sources pour enrichir l'analyse et obtenir une vue d'ensemble fiable du marché.

Il est possible que les avis d’utilisateurs bruts soient étudiés, mais leur inclusion dépendra du succès de leur interprétation, la nature et l’important volume de ces données représentant des défis majeurs.

## Traitements des données

Le traitement des jeux de données sera somme toute un processus assez standard. Les jointures possibles étant le seul élément notable.

* Normalisation et nettoyage des données (gestion des valeurs manquantes, suppression des doublons, vérification de la cohérence des champs).
* Jointures basées sur l'identifiant unique de produit (« *appid* ») pour relier les différents ensembles de données.
* Extraction et transformation des variables pertinentes, selon les besoins des méthodes d’analyses.

## Analyse et modélisation

L’approche retenue est *exploratoire et flexible*, permettant d’adapter les analyses en fonction des résultats obtenus et des contraintes rencontrées. Plusieurs méthodes et niveaux d’analyse seront considérés :

### Exploration et visualisation descriptive (Niveau 1)

Production des statistiques descriptives principales : distribution des prix, popularité des genres et étiquettes, évaluations, etc.

Utilisation de diagrammes de dispersion, boîtes à moustaches et matrices de corrélation pour faciliter la consultation.

Tests de normalité des distributions et détection d’anomalies.

### Analyse relationnelle entre variables (Niveau 2)

Étude des corrélations (Pearson, Spearman) et régressions linéaires (simples et multiples) pour identifier les liens entre les variables (p. ex. l’impact du prix sur le nombre d’installations ou les évaluations).

Tests de signification statistique (tests T de Student, ANOVA) pour valider la puissance des relations.

Visualisation des relations via heatmaps de corrélation et régressions locales (LOESS).

### Segmentation et modélisation exploratoire (Niveau 3)

Utilisation de clustering (K-Means, DBSCAN) pour regrouper les jeux selon des critères comme la tarification et la réception critique.

ACP (Analyse en Composantes Principales) pour réduire la dimensionnalité et détecter des tendances sous-jacentes.

Expérimentation avec des arbres de décision pour comprendre l'influence des facteurs sur le succès commercial.

### Projections et optimisations (Niveau 3)

Expérimentation avec la modélisation prédictive via des régressions avancées (Ridge, Lasso) ou modèles d’ensemble (Random Forest, XGBoost) pour estimer la performance d’un jeu en fonction de ses caractéristiques.

Analyse de séries temporelles (ARIMA, Prophet) pour prédire l’évolution des tendances du marché et exploration des réseaux de neurones récurrents (RNN, GRU) pour capturer des dynamiques plus complexes dans les données historiques.

Simulation d’optimisation de prix via des techniques de recherche par grille et maximisation du revenu estimé.

## Outils et technologies envisagés

Également sujets à une phase d’exploration préliminaire, les outils et technologies prévus sont :

* **Langage et librairies** 
  + Python (avec Pandas, NumPy, Scikit-learn, TensorFlow si applicable)
* **Environnements**
  + Local, Kaggle Notebook
* **Visualisation**
  + Matplotlib, Seaborn, Plotly
* **Stockage et manipulation des données**
  + CSV, JSON
* **Interface et présentation des résultats**
  + Rapport statique (pdf, excel, notebook exporté, etc.)
  + Interface en ligne (par ex. Streamlit or Dash)
  + Terminal ou interface graphique locale

# Calendrier du Projet

|  |  |  |  |
| --- | --- | --- | --- |
| Phase | Tâche | Prévision (h) | Total (h) |
| Exploration initiale | Lecture du matériel de cours, exploration des outils et plateformes, exploration des jeux de données. | 5 | 5 |
| Planification et validation | Exploration d'idées de projet, définition des objectifs, sélection des outils et validation méthodologique (« proof of concept »). | 10 | 15 |
| Travail noté 1 | Rédaction du plan de projet, | 5 | 20 |
| Préparation des données | Sélection, chargement, nettoyage, jointures des jeux de données. | 5 | 25 |
| Analyse descriptive (« Niveau 1 ») | Étude exploratoire des tendances du marché (statistiques descriptives, heatmaps, visualisations). | 10 | 35 |
| Exploration des interfaces | Évaluation des options pour l'interface utilisateur (dashboard Excel, Streamlit/Dash, terminal/tkinter, etc.). | 15 | 50 |
| Analyse relationnelle (« Niveau 2 ») | Mise en place de modèles de corrélation, régression et clustering. | 10 | 60 |
| Modélisation avancée et projections (« Niveau 3 ») | Expérimentation avec les séries temporelles (ARIMA, Prophet, RNN/GRU) et optimisation des paramètres. | 15 | 75 |
| Travail noté 2 | Rédaction du rapport de mi-parcours (modélisation finale du projet). | 5 | 80 |
| Intégration | Finalisation des capacités de l'outil et préparation pour intégration avec interface. (API) | 15 | 95 |
| Développement de l'interface | Développement du tableau de bord ou de l'interface graphique de l'outil. | 20 | 115 |
| Finalisation du projet | Ajustements, mise en ligne (GitHub, dashboard). | 5 | 120 |
| Travail noté 3 | Soumission du produit final fonctionnel. | 0 | 120 |
| Travail noté 4 | Rédaction du rapport final. | 15 | 135 |

# Défis anticipés et plan de contingence

Le projet présente plusieurs défis potentiels, avec des stratégies d’atténuation prévues pour chacun :

* **Données incomplètes ou bruitées**
  + Traitement via imputation des valeurs manquantes et filtrage des données aberrantes pour améliorer la qualité des analyses.
* **Difficulté d’interprétation des avis utilisateurs**
  + Utilisation d’approches NLP (traitement automatique du langage) pour analyser ces données. Si les résultats d’expérimentations ne sont pas concluants, cette partie du projet sera abandonnée.
* **Volume de données élevé**
  + Mise en place de stratégies d’optimisation comme le prétraitement. Possibilité de limiter la portée des produits étudiés. (p. ex. période 2024-2025 uniquement)
* **Contraintes de temps**
  + Réduction de la portée analytique si nécessaire, en priorisant les fonctionnalités essentielles et l’interface.

# Conclusion

Ce projet vise à appliquer un large éventail de concepts clés de la science des données pour analyser les tendances du marché du jeu vidéo sur Steam. En exploitant les ensembles de données accessibles et diverses méthodes analytiques, on espère fournir un outil pertinent à la fois pour l’exploration rétrospective et l’aide à la prise de décision.

Plusieurs défis sont toutefois à anticiper, notamment la gestion du volume des données, les contraintes de temps et le manque d’expertise. Une approche modulaire a été retenue afin d’assurer des livrables itératifs et pertinents.

Si les résultats sont concluants, l’outil pourrait être affiné et enrichi par des modèles avancés et une interface plus flexible. À terme, il pourrait aider les développeurs indépendants à mieux comprendre le marché et à se prémunir contre de couteux échecs.

# Bibliographie

Bay Atlantic University. (2025, 02 28). *Business Intelligence vs Data Analytics.* Récupéré sur Bay Atlantic University: https://bau.edu/blog/business-intelligence-vs-data-analytics/

Galyonkin, S. (2021). *Learn about Steam Spy.* Récupéré sur steamspy: https://steamspy.com/about#Technology

Google Brain team. (2025). Récupéré sur TensorFlow: https://www.tensorflow.org/

Google LLC. (2025). Récupéré sur Kaggle: https://www.kaggle.com/

NumFOCUS, Inc. (2025). Récupéré sur Pandas: https://pandas.pydata.org/

NumPy team. (2025). Récupéré sur NumPy: https://numpy.org/

Plotly. (2025). *Dash Python User Guide.* Récupéré sur Dash: https://dash.plotly.com/

Plotly. (2025). *Data Apps for Production.* Récupéré sur Plotly: https://plotly.com/

Riad. (2021, 06 25). *Ultimate list of the best Analysis Tools for Steam Developers.* Récupéré sur codecks: https://www.codecks.io/blog/2021/best-steam-tools-for-game-developers/

scikit-learn. (2025). Récupéré sur scikit-learn: https://scikit-learn.org/

Snowflake Inc. (2025). *A faster way to build and share data apps.* Récupéré sur Streamlit: https://streamlit.io/

Thai, L. (2024, 04 22). *An Overview of the Current State of the Video Game Industry.* Récupéré sur Foley: https://www.foley.com/insights/publications/2024/04/video-game-industry-challenges-opportunities/

The Matplotlib development team. (2025). *Matplotlib: Visualization with Python.* Récupéré sur Matplotlib: https://matplotlib.org/

Varma, S. (2024, 04 12). *OLAP Operations.* Récupéré sur Scaler Topics: https://www.scaler.com/topics/olap-operations/

Waskom, M. (2025). *seaborn: statistical data visualization.* Récupéré sur seaborn: https://seaborn.pydata.org/