ĐẠI HỌC QUỐC GIA TP. HỒ CHÍ MINH

TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN

PHÂN TÍCH DỮ LIỆU DOANH SỐ BÁN GAME TOÀN CẦU

VIDEO GAME SALES ANALYSIS

Môn học: Lập Trình Cho Khoa Học Dữ Liệu
(Programming for Data Science)

Sinh viên:
Bàng Mỹ Linh – 23122009
Nguyễn Gia Bảo – 23122015
Lại Nguyễn Hồng Thanh – 23122018

GVHD:
ThS. Lê Nhựt Nam
ThS. Phạm Trọng Nghĩa

Thành phố Hồ Chí Minh, tháng 12 năm 2025

Mục lục

Chương 1

TỔNG QUAN

1.1 Thông tin chung

Thông tin	Chi tiết
Tên đồ án	Phân Tích Dữ Liệu Doanh Số Bán Game Toàn Cầu
Dataset	Video Game Sales (VGChartz)
Nguồn dữ liệu	Kaggle - gregorut/videogamesales
Số lượng bản ghi	Hơn 16,000 games
Khoảng thời gian	1980 - 2020
Ngày bắt đầu	15/11/2025
Ngày kết thúc	16/12/2025
Thời gian thực hiện	4 tuần

Bảng 1.1: Thông tin tổng quan về dự án

1.2 Mục tiêu

Đồ án nhằm mục đích thực hiện phân tích toàn diện dữ liệu doanh số bán game toàn cầu, bao gồm:

1. **Thu thập và tiền xử lý dữ liệu:** Làm sạch, chuẩn hóa và xử lý dữ liệu thiếu từ dataset gốc.
2. **Phân tích khám phá dữ liệu (EDA):** Khám phá các xu hướng, mẫu hình và mối quan hệ trong dữ liệu qua các chiều Platform, Genre, Publisher, Year và Region.
3. **Trả lời câu hỏi nghiên cứu:** Phân tích sâu 5 câu hỏi nghiên cứu về:
    - Sự kết hợp Platform-Genre tối ưu

- Vòng đời nền tảng và doanh số
- Xu hướng thể loại game theo thời gian
- Hiệu suất của các Publisher
- So sánh Publisher lớn và nhỏ

4. **Xây dựng mô hình dự đoán:** Phát triển và đánh giá các mô hình Machine Learning để dự đoán doanh số game.
5. **Trực quan hóa kết quả:** Tạo các biểu đồ, đồ thị trực quan để trình bày findings một cách hiệu quả.

1.3 Cấu trúc deliverables

Cấu trúc đồ án bao gồm:

- **3 Jupyter Notebooks:**
  1. `01_Data_Collection_Preprocessing_EDA.ipynb`
  2. `02_Analysis_Questions.ipynb`
  3. `03_Modeling.ipynb`
- **Dataset:** `vgsales.csv`
- **Tài liệu:**
  - `README.md` - Hướng dẫn sử dụng
  - `Team_Plan_Work_Distribution.pdf` - Tài liệu này

Chương 2

KẾ HOẠCH THỰC HIỆN

2.1 Timeline tổng quan

Tuần	Thời gian	Nội dung chính	Deliverable
1	15/11 - 21/11	Thu thập & Tiền xử lý dữ liệu, EDA cơ bản	Notebook 1 (50%)
2	22/11 - 28/11	EDA chi tiết, Phân tích câu hỏi 1-3	Notebook 1 (100%) Notebook 2 (60%)
3	29/11 - 05/12	Phân tích câu hỏi 4-5, Xây dựng models	Notebook 2 (100%) Notebook 3 (70%)
4	06/12 - 16/12	Đánh giá models, Hoàn thiện tài liệu	Notebook 3 (100%) Documentation

Bảng 2.1: Timeline thực hiện dự án theo tuần

2.2 Chi tiết kế hoạch theo tuần

2.2.1 Tuần 1: Thu thập và Khám phá Dữ liệu (15/11 - 21/11)

Mục tiêu: Hoàn thành thu thập dữ liệu, tiền xử lý và EDA cơ bản.

Ngày	Công việc	Kết quả mong đợi
15-16/11	• Download dataset từ Kaggle • Setup môi trường Python • Import thư viện cần thiết	Environment sẵn sàng, dataset đã tải
17-18/11	• Load và kiểm tra dữ liệu • Xử lý missing values • Kiểm tra duplicates	Dữ liệu sạch, không có lỗi
19-20/11	• Phân tích mô tả dataset • EDA theo Platform • EDA theo Genre	Hiểu tổng quan về dữ liệu
21/11	• Review và chỉnh sửa Notebook 1 • Tổng kết tuần 1	Notebook 1 hoàn thành 50%

Bảng 2.2: Kế hoạch chi tiết Tuần 1

2.2.2 Tuần 2: EDA Chi tiết và Phân tích Câu hỏi (22/11 - 28/11)

Mục tiêu: Hoàn thiện EDA và trả lời 3 câu hỏi đầu tiên.

Ngày	Công việc	Kết quả mong đợi
22-23/11	• EDA theo Year/Decade • EDA theo Publisher • Phân tích doanh số theo Region	Hiểu xu hướng theo thời gian và khu vực
24/11	• Phân tích Top Games • Phân tích correlation • Hoàn thiện Notebook 1	Notebook 1 hoàn thành 100%
25-26/11	• Phân tích Câu 1: Platform-Genre • Phân tích Câu 2: Vòng đời Platform • Visualization cho 2 câu hỏi	Insights về Platform-Genre tối ưu
27-28/11	• Phân tích Câu 3: Xu hướng Genre • Time series analysis • Review Notebook 2	Notebook 2 hoàn thành 60%

Bảng 2.3: Kế hoạch chi tiết Tuần 2

2.2.3 Tuần 3: Hoàn thành Phân tích và Modeling (29/11 - 05/12)

Mục tiêu: Trả lời câu hỏi 4-5 và xây dựng models ML.

Ngày	Công việc	Kết quả mong đợi
29-30/11	• Phân tích Câu 4: Hit game ratio • Định nghĩa và tính toán metrics • Visualization Publisher performance	Xác định Publisher hiệu quả nhất
01-02/12	• Phân tích Câu 5: So sánh Publisher • Statistical testing • Hoàn thiện Notebook 2	Notebook 2 hoàn thành 100%
03-04/12	• Feature engineering cho ML • Encoding categorical variables • Train-test split	Dữ liệu sẵn sàng cho modeling
05/12	• Training Linear Regression • Training Random Forest • Preliminary evaluation	Models cơ bản hoàn thành

Bảng 2.4: Kế hoạch chi tiết Tuần 3

2.2.4 Tuần 4: Đánh giá Models và Hoàn thiện (06/12 - 16/12)

Mục tiêu: Hoàn thiện modeling, đánh giá và viết tài liệu.

Ngày	Công việc	Kết quả mong đợi
06-07/12	• Training Gradient Boosting • Cross-validation • Hyperparameter tuning	Models được tối ưu hóa
08-09/12	• So sánh models (MAE, RMSE) • Feature importance analysis • Visualization results	Xác định model tốt nhất
10-12/12	• Viết phần kết luận • Tổng hợp tài liệu tham khảo • Hoàn thiện Notebook 3	Notebook 3 hoàn thành 100%
13-14/12	• Viết README.md • Viết Team Plan document • Review toàn bộ project	Documentation hoàn chỉnh
15-16/12	• Final review • Testing notebooks • Chuẩn bị nộp bài	Project sẵn sàng submit

Bảng 2.5: Kế hoạch chi tiết Tuần 4

Chương 3

PHÂN CÔNG CÔNG VIỆC

Thành viên	MSSV	Công việc	% Hoàn thành
Bàng Mỹ Linh	23122009	**Notebook 1:** Thu thập và load dataset, EDA theo Platform và Region, Phân tích Top Games, Visualization cho EDA cơ bản	100%
		Notebook 2: Trả lời Câu 1 (Platform-Genre combination), Trả lời Câu 4 (Hit game ratio analysis), Visualization cho 2 câu hỏi	
		Notebook 3: Thực hiện Gradient Boosting model, Feature importance analysis	
		Documentation: Viết README.md, Mô tả dataset và findings	
Nguyễn Gia Bảo	23122015	**Notebook 1:** Xử lý missing values và data cleaning, Kiểm tra duplicates, EDA theo Year/Decade/Publisher, Correlation heatmap và phân phối dữ liệu	100%
		Notebook 2: Trả lời Câu 2 (Platform life-cycle), Trả lời Câu 5 (Publisher size comparison), Statistical testing, Review toàn bộ Notebook 2	
		Notebook 3: Feature Engineering, Training Random Forest, So sánh performance models, Visualization cho model results	
		Documentation: Code comments và docstrings, Phần methodology	
Lại Nguyễn Hồng Thanh	23122018	**Notebook 1:** Review chất lượng dữ liệu, EDA theo Genre chi tiết, Phân tích correlation và distribution, Tổng kết EDA	100%
		Notebook 2: Trả lời Câu 3 (Genre trends over time), Time series analysis, Tổng hợp insights từ 5 câu hỏi	
		Notebook 3: Chuẩn bị dữ liệu cho ML (encoding, scaling), Training Linear Regression, Cross-validation và metrics, Feature importance visualization, Kết luận	
		Documentation: Team Plan & Work Distribution, Tài liệu tham khảo, Final review toàn bộ project	

Bảng 3.1: Phân công công việc chi tiết theo thành viên

Chương 4

KẾT LUẬN

4.1 Cam kết của nhóm

Nhóm chúng em cam kết:

1. Hoàn thành đúng deadline và đảm bảo chất lượng công việc
2. Phân công công việc công bằng, rõ ràng
3. Hỗ trợ lẫn nhau trong quá trình thực hiện
4. Tuân thủ quy trình làm việc và quality assurance
5. Thực hiện nghiêm túc theo kế hoạch đã đề ra
6. Liên tục cập nhật tiến độ và điều chỉnh kế hoạch nếu cần

4.2 Kỳ vọng kết quả

Sau khi hoàn thành dự án, nhóm kỳ vọng đạt được:

- 3 notebooks hoàn chỉnh với code chạy tốt và visualization đẹp
- Insights sâu sắc về thị trường game toàn cầu
- Models ML với performance chấp nhận được
- Documentation đầy đủ và chuyên nghiệp
- Kinh nghiệm làm việc nhóm và quản lý dự án data science
- Kỹ năng phân tích dữ liệu và machine learning được nâng cao