Steam Sales & Pricing Analysis

By Angel Ramirez (arami80)
Miguel Madrigal (mmadr5)
Alex Cruz Martinez (acruz200)
Christian Julias (cjulias82)
Mohsin Patel (mpate308)

https://github.com/mmadr5/SteamSalesAndPricingAnalysis.git

Problem

Big Idea:

Analyze how both game pricing and discount percentages relate to game attributes such as genres, reviews, and release dates, to identify patterns in how Steam does its pricing strategies.

Why does this matter?

- For developers: Understand common pricing trends and discounting practices.
- For consumers: Recognize how game features and reviews relate to discounts.
- For analysts: Provide insights into digital game marketplace behavior.

How we chose this problem: The famous Steam sales entices many people to buy their wanted game so want to know how discounts are applied across genres and game types.

Hypotheses:

- Games with higher discounts may have lower user reviews.
- AAA and Indie titles may show different discounting patterns.
- Older games are more likely to receive higher discounts.

Data

Source: We are using the Steam Store Games dataset from Kaggle (non-competition dataset) as our primary data source. To strengthen our analysis, we are also exploring additional datasets from sources like Zenodo or other public platforms to provide complementary insights.

Access: We have immediate access to the Kaggle dataset and are in the process of identifying other relevant datasets to integrate.

Effort:

- Minimal cleaning required on the Kaggle dataset (such as categorizing discounts and grouping genres).
- Additional effort may be needed to clean and merge data from secondary sources once identified.

Primary Dataset (Kaggle):

- **Size:** ~27,000 games
- **Type:** Tabular snapshot of current Steam store listings
- **Features:** Game titles, Original price, Discounted price, Discount percentage, Release dates, Genres and tags, and Review scores
- Potential Secondary Datasets:
  - Currently exploring datasets that could provide historical pricing, player activity, or gameplay metrics to enrich the analysis.

Limitations:

- The Kaggle dataset has no historical sale data or seasonal sale tracking.
- Additional datasets are still under consideration to help address these limitations and provide broader context to our findings.

Solution

- EDA:
  - Analyze price distributions.
  - Explore relationships between discounts and review scores.
  - Compare discount patterns across genres.
  - Look at age of games and likelihood of discounts.
- ML Models:
  - Classification: Predict whether a game has a "high" or "low" discount.
  - Regression: Predict discounted price based on original price, release year, and other features.
- Visualizations:
  - Heatmaps (correlation between features).
  - Scatterplots (discount % vs. review scores).
  - Boxplots (discounts across genres).

Expected Deliverables/Findings

End Result:

- Identify patterns in how discounts are applied across Steam games.
- Understand which types of games tend to be more heavily discounted.
- Provide insights for developers on typical discounting practices.

Next Steps:

- Finalize dataset cleaning and preprocessing.
- Perform EDA and build initial visualizations.
- Train first ML models for discount prediction.

Progress Report Goal:

- Completed EDA with at least five visualizations.
- Early results from classification and regression models.
- Preliminary insights into discount patterns and pricing strategies.