Individual Project
Data Science Fundamentals

Introduction

The main question that will be answered because of this project is "What makes a video game 'Popular' on Steam?". My desire to answer this stems from the fact that I enjoy playing video games as a hobby and prefer using the Steam platform to purchase my video games. Knowing the answer to this question will give me insight into why some games become popular, as I typically enjoy lesser-known games or games that don't 'get huge'.

Datasets

The dataset I will be using is from Nik Davis on Kaggle.com. While his data is a bit outdated, it is clean, content-heavy, and organized well to use. The dataset is hyperlinked here and in case that doesn't work, here is a direct link:
https://www.kaggle.com/datasets/nikdavis/steam-store-games/data?select=steam.csv

Methodology

Breaking down the main question into three smaller questions will help set up what to do and how to do it. Let's begin.

To reiterate the main question, "What makes a video game 'Popular' on Steam?". I will break this down into three questions as follows:

1. What game features are most strongly associated with popularity on Steam?
    a. Price, genres, tags, categories (Single-player, Multiplayer), etc.
2. Can we predict whether a game will be popular using only metadata available at release time?
    a. Classification model that predicts if a game is "popular" based on features like pricing, support platforms, genres, tags, release year, etc.
3. How do pricing and release timing relate to popularity?
    a. Do cheaper games tend to be more popular?

b. Does release year (or month?) matter for popularity?

Having these in mind will help answer the main question. However, how is the process going to look like? What will I do to help solve these problems?

Q1. What game features are most strongly associated with popularity on Steam?
I will first define a popularity label using the estimated number of owners and mark the top 25 percent as "popular." Then I will compare features like price, release year, genres, categories, and platforms between popular and non-popular games using summary statistics, plots, and a simple logistic regression model to see which features are most associated with popularity.

Q2. Can we predict whether a game will be popular using only metadata available at release time?
I will treat this as a binary classification problem using only prerelease information such as price, age rating, release year, platforms, genres, categories, and a small set of tags. After preprocessing the data and splitting it into train and test sets, I will train a logistic regression classifier (and possibly a random forest) and evaluate it with accuracy, precision, recall, F1 score, and a confusion matrix.

Q3. How do pricing and release timing relate to popularity?
I will group games into price bins and compute the fraction of popular games in each bin, then visualize this with bar charts. I will also examine popularity by release year and optionally fit a simple logistic regression with price and release year to see how these two variables relate to the chance that a game is popular.