CS 4412: Data Mining Proposal: Analyzing Video Game Sales Between Publishers, Genre, and Platform

Joshua Vach

***Abstract*—This report is a proposal for data mining a dataset containing 16598 games ranked from most sold to least sold. the dataset also contains sales data for various markets as well as the platform, publisher, genre, name, and release year for every instance. The goal will be to use this data set to find correlations between these data points and find how different games perform in different times and markets.**

I. Introduction

Video games have been around for decades and have waxed and waned in popularity and profitability over the years, with new genres out competing old ones and new companies taking over from old ones. Despite all of this, video games have become a staple of modern entertainment, and almost every household has at least some way to play them. Due to the massive size of the video game market, it is more important than ever to study the trends of video games in the past so that we can gain valuable info about how the future of gaming is going to be.

II. Dataset Description

The data set we will be using in specific is named ”Video games sales” that was last updated in Kaggle on December 08, 2018.

A. Source and Description

- **Source:** https://www.kaggle.com/datasets/gregorut/videogamesales/data
- **Description:** The data set has 16600 rows, which means that 16600 games are in the data set, and 11 columns, each game has a unique rank assigned to it that is ranked from highest selling to lowest selling. The data set is stored in CSV format and is 1.36 megabytes.

B. Data Attributes

- **Main Attributes:** The attributes that will be used aside from the name consist of Rank, Platform, Year, Genre, Publisher,NASales, EUSales, JPSales, OtherSales and GlobalSales.
- **Known Issues:** Since the data set is almost a decade out of date, there are several big recent games that are not present in the data set which could cause some skewing of results. This issue also shows up because some games still receive sales even decades after release which the sales of the past 8 years are not stored which could cause some sales to be incorrect. Another issue is that many other large video game markets, namely China and South Korea, are lumped into the ”OtherSales” attribute instead of having their own, so it is impossible to tell what games were popular there.

III. Discovery Questions

The research will focus on answering a few questions that can be found by researching the data set.

- Do publishers do better in their home markets or in other countries and does the platform or year affect this outcome? This question is interesting because it can show whether a company is more likely to find success in releasing their game abroad or not, and whether or not the platform matters in its success abroad.
- Which platforms sold the most video games each year and were these dominated by specific games or a more broad spread. This can help find which platform was the largest in a specific year and if it was just a fluke due to one big game or a larger dominance in the market.
- How have total sales changed over the years in the different major markets. This can tell us a lot about the economic situation in markets around the world because video games, being an entertainment good, are very likely to change in the face of economic growth or strife.

IV. Planned Techniques

The analysis will consist of preprocessing and transforming the data before using the data mining techniques, then interpreting the output of the data mining techniques and then visualizing them.

A. FP-Growth

Taking the metadata within the data set, we can use association rules to group data points that are similar to each other. These relationships can be further extrapolated in order to find patterns within the data. For example, Which publishers do well in each market and which platforms sell games in certain years.

B. K-Means

This will be used to divide the data into much more manageable clusters of various different types of information. Along with euclidean distance, there are multiple other data values that could be divided up in our dataset. For example, using rank to separate the higher grossing games from the lesser ones, dividing by year released to see how certain periods of time had certain trends, the platform the game was released on, or specific sales in a certain market. This can also be used to visually represent data better than if it is one large block of data.

C. Anomaly Detection

Detecting abnormalities and outliers in the data set is critical because in a data set based on sales, which are affected by so many things, it is important to be able to detect if something is weird and figure out why that data point is abnormal and draw conclusions based on that. For example, Wii Sports is so much higher than every other game because it came complementary with the Wii which inflated its numbers. Things like this happen a lot with video games so it's important to catch them and be able to explain them.

V. Preliminary Timeline

This timeline is not final it is up to change as things change. Below is the tentative timeline for project completion.

TABLE I
Timeline: Project Milestones

Milestone	**Period**	**Key Milestones**
M2	Week 8	Begin Work on Mining Techniques and Anomaly Detection
M3	Week 11	Continue Mining Techniques and Anomaly Detection and Begin Testing
M4	Week 14	Complete Testing and Write Final Report

A. Anticipated Challenges

- The biggest challenge will definitely be getting the basics of data mining and how to implement the techniques right. I have experience working on databases but never anything at this scale and will definitely require some time to get the basics under my belt.
- Another large hurdle will be the size of the dataset and the amount of variables. It will be difficult to figure out exactly what processes or data needs to be implemented or analyzed.
- The final issue will be implementing Anomaly Detection as I have never really done anything like that before it will definitely be an interesting time implementing it.