Video game sales

Group D21
Viljam Ilves, Ats Tragel

Business understanding

Background:

In the video game industry, there are many failed video game projects that at first, seemed to be on a road to great success, but after they are published, they don't kick off as well as it was thought they would. It could have just been a bad coincidence of being published at a bad time or maybe the game was over-hyped. Most of the time people have really interesting and well thought out ideas of what a game would look like and how it would work, but the game they have thought out might not target an audience big enough to be worth the effort put into it.

This project could be useful for companies who want to start making video games. Our project will bring out different aspects of what could potentially make a successful video game, giving insight to platform choice, genre choice and where it would be to best publish such a game. It would be especially useful for starting companies or indie game companies that do not have too many resources to make a game that would be risky for the survival of the company itself. We believe our project could potentially help companies get a good general direction of what games would be successful and some basic info about how that game would sell best.

Goals:

Put together useful visual and textual statistics which show how well a video game sold and the basic reason for it. Make predictions of video game sales by given input variables thus helping future developers think of a game idea.

Success:

Predict some upcoming title successes and get accurate predictions for other games already in our data. Predict if a game will be profitable to make. (Recommend games?)

Inventory of resources:

The main dataset we are working with is in .csv format. The dataset has info of a video games, their publisher and year published and its sales in North America, Europe, Japan and Worldwide. We are looking to add data of each game's average rating by critics and other useful statistics. We will be coding in Python and probably will use Jupyter Notebook.

Requirements, assumptions, and constraints:

This project will be completed before the project presentation deadline. By then, we will have completed all of the goals and requirements we have set for ourselves. We believe there will be no security or legal obligations because the data we are using is available for use for free on the internet.

Risks and contingencies:
We might run into some delays with finding appropriate data that could expand and make our predictions more precise. Also filtering and cleaning that data to add to our dataset could take some time. Another risk we might have is that the current dataset we have chosen has data from the earliest days of video games up to data from 5 years ago. We don’t consider this too big of a risk, because the data is from a timespan of more than 30 years and we think that gives us enough data that would hold up even today.

Terminology:
Gaming Platform - a computer system specially made for playing video games, ex: Nintendo 64

Action game - a game genre emphasizing physical challenges, hand–eye coordination and reflexes. It includes fighting games, shooters, and platformers.

Platform game/platformer - Any video game, or genre which involves heavy use of jumping, climbing, and other acrobatic maneuvers to guide the player-character between suspended platforms and over obstacles in the game environment.

Role-Playing/RPG - A game genre in which the human player takes on the role of a specific character "class" and advances the skills and abilities of that character within the game environment.

Costs and benefits:
We are planning to work around 30 hours each. We believe we will not be spending any money on this project.

Data-mining goals:
Our goal is to make a report and code with which it would be possible to predict how well a game will sell. The report will have our findings of what makes a videogame more sales than other games and what company/publisher is best at making games, which will be shown in graphs or other appropriate visual and textual form. The code will take simple variables, for example the publisher and game genre, to make predictions how well a game will sell in North America, Europe and elsewhere.

Data-mining success criteria:

The world of video game marketing is fairly chaotic, so we are quite content if we can make roughly correct predictions from our data. Hopefully our predictions and other results are around 90% accurate, we consider an accuracy above that to be a great success.

Data understanding

Gathering data

Outline data requirements

The data we will be working with has to be in .csv format. We are definitely going to need some data about sales in different areas and the genre of the games themselves

Verify data availability

The data that I want to be working with is accessible for free on the kaggle website. Should we want to add more data, the creator of the kaggle database has also added the Python script which he used to collect the data from a video game website. We believe we have the required data to complete this project

Define selection criteria

The main dataset we will be using is from kaggle: https://www.kaggle.com/gregorut/videogamesales. The most important data we are gonna look at and use are the game genre and the game sales in different parts of the world. We want those pieces of data to be there every time so any entries that are missing the data in those parts will most likely get removed. The platform on which the game is made is also important but we think of it more as an extra variable to add to make the prediction more precise so this data is going to be important as well.

Describing data:

The data from kaggle is made using a python script to extract data from the website http://www.vgchartz.com/gamedb/ . We are looking into using the same script to expand our data, as the data on the website has some extra columns that we could use for our project.
The Data from kaggle consists of video game titles, the platform it was made for, the publisher, number of sales in a region and year published with 16600 different entries. The initial data seems to be suitable for our data-mining goals.

Exploring data:

The data only has video games with sales greater than 100000 copies. The number of sales in the data is shown by millions of copies sold, counts both digital and physical as an individual copy. It's not clear how data of games that are in continuous development and are selling many copies to this day is shown in the dataset; are the sales showing only yearly number of copies sold or the number from its release date to the year in the dataset. It is most likely the number of sales at the time the dataset was made.

Verifying data quality:

The data in the main dataset is enough for us to make basic predictions. Adding 1 or 2 columns of useful data to the data, would give a pretty good base to make predictions from. The only uncertainty is the continuously updated game sales data but we think it is not going to be too big of a problem for us.

Planning:

1. Extract and clean data- This task should not take too long because it is fairly simple and the data is already extracted by another user. We will be removing any entries which do not meet our desired requirements. Ats - 2h
2. Add more data to the original dataset and remove unusable entries - This task might take a little bit longer but it also should not be too long. Ats - 5h
3. Make predictions - This task will probably take the longest so we will be putting the most effort into this one. Viljam 20h, Ats 15h
4. Visualization - This task might not take too long, depending on what kind of a visualization we decide to do. Viljam 5h
5. Make code to give recommendations - This will be the last task, hopefully taking long enough to fill the gaps remaining in our planned time frame of 30hours per person. Viljam 5h, Ats 8h

https://github.com/viljamilves/video-game-sales-D21