KAGGLE-Steam Store games

Group C12
Viljam Ilves, Ats Tragel

Business understanding
Background:

In the video game industry, there are many video game digital distribution services, Steam being one of the biggest, hosting many big video game titles titles and even more smaller ones.

This project could be useful for many different game developers. Our project will bring out different aspects of what could potentially make a video game more successful than before, giving insight to different things that could keep the players playing for a longer time. We believe our project could potentially help developers get a good general understanding of what to add to their games or what to modify in their games so they could increase how much time each player spends playing their game.

Goals:

Put together useful visual and textual statistics which show correlations between playtime and other game variables. Make predictions of video game playtime by given input variables thus helping future developers to potentially make their games more played.

Success:

Predict some title successes and get accurate predictions based on other games already in our data. Predict if a game will be more successful if it has more of some features (achievements, multiplayer etc)

Inventory of resources:

The main dataset we are working with is in .csv format. The dataset has info of the video game, their publisher, playtime of the game, the amount of achievements and different other variables the game has like genres. 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. Otherwise we don't believe there are other big risks.

Terminology:

Platform - a computer system, ex: Windows, Mac

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

Multi-player - A game where players can play on the same system or different ones cooperatively to achieve a common goal or compete against other players.

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 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 to make a game have a slightly longer playtime. The report will have our findings of what makes a videogame more playable than other games and what variables are best at extending playtime, which will be shown in graphs or other appropriate visual and textual form. The code will take simple variables, for example the ratings and achievements, to make predictions how long that game would be played.

Data-mining success criteria:

Video game playtime is a fairly simple statistic to follow but it is hard to predict it accurately based on only variables like achievements or ratings. 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 playtime in different games and other game variables that could have a correlation with playtime.

Verify data availability
The data that we 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 documentation on how to use the Steam Store API which he used to collect the data. 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/datasets/nikdavis/steam-store-games?select=steam_support_info.csv .
The most important data we are gonna look at and use are the game playtime and the game variables like achievements and categories. 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 categories column 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 Steam Store API to extract data from the Steam Store. https://steamspy.com/api.php . We are looking into using the API to expand our data, as the data on the website has might have newer data. The Data from kaggle consists of video game playtime, the categories of the game, the positive and negative ratings, owners of the game the price of the game. The initial data seems to be suitable for our data-mining goals.

Exploring data:
The data has most of the video games from the Steam Store. The main dataset has columns named: Average- and median-playtime (in minutes) , positive- and negative-ratings, price of the product, genres, etc.

Verifying data quality:
The data in the main dataset is enough for us to make basic predictions. Adding statistics of newer releases to the data would be useful and would give a pretty good base to make predictions from. Quality of the data looks very good.
Planning:

1. Explore the data and 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. Viljam 5h,

Ats 5h

2. Visualization - This task might not take too long, depending on what kind of a visualization we decide to do. Viljam 10h, Ats 5h
3. Make the model for predictions - This task will probably take the longest so we will be putting the most effort into this one. Viljam 10h, Ats 20h
4. Making our project presentable - this task is for polishing and finishing touches/fixing problems - Viljam 5h

https://github.com/anonsta/C12-Steam-Store-Games