SI618 Project Proposal 2

Sijun Tao (sijuntao@umich.edu)

Summarize and motivate your proposed project:

Steam is one of the most popular platforms on which people can search and find the games they are interested. As one of the Steam users and someone interested in being a game developer, I'm focusing my project on evaluating the games on Steam. What type of games are most popular? Which developer has the most games with high ratings? What game price setting is common among the games?... I mainly want to explore how players decide which games to play and how to evaluate them.

The dataset I plan to use was found on Kaggle (https://www.kaggle.com/datasets/nikdavis/steam-store-games). This dataset contains games released prior to May 2019 on Steam. It has 18 columns, including information about application ID, name, released date, language support, developer, publisher, platforms, genres, average playtime and so on. I will clean and do some data manipulation to this dataset for better exploring, such as explode the data to make each row has one genre.

Proposed Analyses

Analysis 1-3: Time analysis of games

1. What is the average price over the years? I will calculate the average price of the games released in the year and use ggplot to plot it to find if there is any tendency over the years.
2. What is the average price over the months? I will calculate the average price of the games released in the month. I will plot the median price per month for each individual year.
3. What is the average user play time over the years? I will use ggplot to plot the average user play time of the games released in each year. Through this, I can have an idea of how much time people would like to invest into games.

Analysis 4-8: Exploratory analysis of games

4. How much is the game price commonly set? I will find the price distribution of all the games. I will calculate the highest, lowest, median, quartiles price and other statistical data. Using histograms, I can find the most common game price.
5. What is the game price set for the games of the top 5 developers (have most developed games)? I will use box plot to visualize the data, finding the price setting feature for different developers.
6. Is there a relationship between positive ratings ratio and the game owner number (estimated number of owners)? I will create reg plot for the positive ratings and game owner number and encode genre as color.
7. Is there a relationship between positive ratings ratio and the number of achievements of the game? I will create scatter plot to see if higher positive ratings ratio leads to a larger number of achievements.
8. Whether the game price has an influence on the average user play time? I will use scatter plot which encodes different platforms on color to analyze.