STATS415 Project Proposal: Video Game Sales with Ratings

Group members: Runhe Lu, Yunguo Cai, Wenfei Yan
Lab Section: Lab 03
Data Resource: https://www.kaggle.com/rush4ratio/video-game-sales-with-ratings

Background

Our dataset is about the sales and ratings of video games with 16720 observations.
For each observation, there are 16 variables, including 4 categorical variables in the dataset:
Name, Platform, Year_of_Release, Genre, Publisher,
NA_Sales, EU_Sales, JP_Sales, Other_Sales, Global_Sales,
Critic_score, Critic_count, User_score, User_count, Developer and Rating.
Among these, Platform, genre, developer and publisher are categorical variables.

Objectives

We aim to predict how different factors would affect the sales of video games. Average ratings across the 13 genre groups will be calculated per platform, per region and per decade to be used as predictors. We will present you with the following results of our data exploration and visualization techniques.

1. Which one is the best console: Wii, PS or XBOX?
2. Do cross-platform games sell better than monopoly games?
3. Can we make some predictions on sale number of games based on their genre?
4. What's the overall trending in game market these years?
5. What's the relationship between the video games and user scores and critic scores?

Specification

We will apply the data mining methods discussed in class to the project as follows.

1. Obtain the sales of video games with respect to genre, platform, rating scores and timestamp using 2-by-2 histograph.
2. Compare the correlaition between platform, genere, rating scores and timestamp with heatmap.
3. Examine if the PCA components could be generalized to predict the platform of video games with an unknown genre.
4. Predict the sales of the overall trending in game market these years by regression.
5. Cluster users based on their rating behavior and find the highest-rating genre of video games specific to each clustering group.

Tools used: Language R