Data 410 Project Proposal

Group Members

- Aedan Wen 97764138
- Mitchell Joram 98653793
- Jong Gil Park 55979158

Dataset

Dimensions: 16719, 16
Columns: 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, Rating

We're going to be working with a video game sales dataset from 2016. The dataset contains a list of games released up to 2016 with sales over 100,000 copies. The dataset contains attributes for name, platform, publisher, genre, sales by region in millions of copies, critic score, and user score for each game.User_Count and Critic_Count are the number of users/critics who contibuted to that score. The dataset contains 16719 rows, but about 6900 rows are completed since many have missing data due to Metacritic not covering all the games for certain attributes. Our goal for the project is to analyze game sales by region and to find out which factors affect the total game sales in the region. Our primary interest lies in finding out the differences in variable importance between NA, EU and the rest of the world. Each game could be classified in various ways such as whether the game is a domestic game (I.E Ninentndo in Japan) or a foreign game, platform of the game. Additionally the rating of the game is presented in the dataset. Analyzing the regression between variables and total sales is the main goal for this project.

Relevant questions we can answer

- What are the differences between variable imporance in different regions?
- How have video sales increased over time? Has North America grown faster than EU or the rest of the world?
- Do games made domestically sell better than they do in the rest of the world? (Does ninendo/sony sell better in Japan?)

- Have the different platforms changed in popularity over time? (Is PC more popular than consoles now than it was in 2009?)
- Does critic score affect sales?
- What consoles sold the most games?
- Are certain genres or consoles selling better in different regions?

Possible Problems

- The scores for the games come from Metacritic which does not have a score for all the games. This means that we will have to decide what to do with the empty rows. Possible solutions to this are to simply delete the rows or use data manipulation techniques such as mean replacement to fill them out.
- Another possible issue is that our data may not be linearly correlated or our errors won't be normally distributed. In the event that this happens we will likely have to try data transformation to fix this.

Regression Techniques

We want our linear model to maintain our 5 linear assumptions. As previously mentioned, if our data is not linearly correlated or normally distributed we'll use transformations on our input to . Once we have a linear model we'll use studentized residual analysis to detect outliers and make sure we don't violate assumptions 1 through 3, leverage and influence point analysis, Q-Q plots to analyze the distribution of the random variables. We'll try to find interaction terms to see if any of our terms interact such as scores and console, and scores and publisher (domestic or foreign) when we analyze each region.

Data Source Link