IST718 – Big Data Analysis

Research Project Title: Video Game Sales

Yangyang Ye, Yang Zhou, Xuanbang Liu

Objective of our project: Our project is to analysis the sales of video games in US, Europe, Japan and all over the world from 1980s to 2010s. In this project we try to solve some problems: what’s the factors to influence video game sales? What’s the sales developed in the future? Which types of video games have the best sale? And so on. We use the data to do virtualization , ,linear regression , production and finally make some useful marketing strategies based on our analysis.

Dataset Description

We found the dataset in Kaggle. The resource of dataset:

<https://www.kaggle.com/datasets/gregorut/videogamesales?select=vgsales.csv>

This dataset has 11 columns and 16600 rows. We will do the predictor of global sales and sales in different area.

Data Exploration:

We first print the schema of the data and change some columns schema. We use describe function to do the statistical summary. Here is some columns result.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Summary | NA\_Sales | EU\_Sales | JP\_Sales | Other\_Sales | Global\_Sales |
| Count | 16598 | 16598 | 16598 | 16598 | 16598 |
| Mean | 0.264 | 0.146 | 0.078 | 0.048 | 0.538 |
| Stddev | 0.817 | 0.5 | 0.3 | 0.189 | 1.556 |
| Min | 0 | 0 | 0 | 0 | 0.01 |
| Max | 9.81 | 29.02 | 10.22 | 10.57 | 82.74 |

We do some basic show of the dataset, such as The frequency of Global sales. From the plot we could know that the sales of most video games are less than 10. Only a little video games have a high sales.

A picture containing text

Description automatically generated

Prediction: For example, we will predict the different Global sales in different genre in the next year. we will add more predication questions in the future.

Inference: We will try different model, like regression model correlation model and so on, to do the analysis. We will compare the accuracy of each model, and choose the best model to do the prediction and finally do the virtualization.

Packages: we plan to use spark packages. In the progress, we may add more in the future.