****

**CALIFORNIA STATE UNIVERSITY, FULLERTON**

**Data Collection and Preparation:**

Projecting the Popularity of Video Games

**ISDS 577**

**TEAM 2**

Annie Xu

Arundhathi Roy

Joshua Katana

Katherine Gonzalez

Sydney Horanic

**Data Details**

Our data is sourced from two publicly available datasets on Kaggle (“Video Game Sales” and “Video Games Sales 2019”). These data sets were originally scraped from a site called VGChartz.com that tracks global video game sales data across multiple platforms and regions. The datasets contain no personal identifiers or usage restrictions, making them suitable for public analysis. Both of the datasets are formatted as CSV files that could be uploaded and analyzed in many analytical tools like Tableau and R.

The first data set is [Video Game Sales](https://www.kaggle.com/datasets/gregorut/videogamesales), which contains a list of video games with sales greater than 100,000 copies. There are 16,598 records in this data set and 11 columns with variables for Rank, Name, Platform, Year, Genre, Publisher, NA\_Sales, EU\_Sales, JP\_Sales, Other\_Sales, and Global\_Sales. Our second data set is [Video Games Sales 2019](https://www.kaggle.com/datasets/ashaheedq/video-games-sales-2019). This data set is a continuation of the first data set and contains 55,792 records. It includes variables such as Rank, Name, Platform, Genre, ESRP Rating, Rublisher, Developer, Critic Score, User Score, Total Shipped, Global\_Sales, NA\_Sales, PAL\_Sales, JP\_Sales, Other\_Sales, and Year.

To prepare the data, we merged the two datasets together by performing a left join on the “Name” and “Platform” variables, preserving the broader dataset while appending additional details from the 2019 file, such as “Critic\_Score” and “User\_Score”. We then created a new variable called “Pop” that calculates a popularity metric by averaging the z-score standardized values of Critic and User scores. This variable is designed to capture the perceived audience success beyond sales data alone. This will help us see the difference between games that sold well commercially versus those that built strong fan bases or critical acclaim. Standardizing these scores allows us to give more analytical weight to scores that deviated from the mean.

To ensure data quality, we removed null values, dropped unnecessary columns (such as “ERBS”, which was 58% null), and inspected for any duplicates and outliers. Our final data set contains 16 columns and 16,598 rows.

Variables:

| Rank | Ranking of overall sales |
| --- | --- |
| Name | Name of the game |
| Platform | Platform of the game |
| Year | Year of release of the game |
| Genre | Genre of the game |
| Publisher | Publisher of the game |
| NA\_Sales | Sales in North America (in millions) |
| EU\_Sales | Sales in Europe (in millions) |
| JP\_Sales | Sales in Japan (in millions) |
| Other\_Sales | Sales in the rest of the world (in millions) |
| Global\_Sales | Total worldwide sales (in millions) |
| Critic\_Score | Critic score of the game on scale of 10 |
| User\_Score | User score of the game on scale of 10 |
| Norm\_Critic | Normalized Critic score |
| Norm\_User | Normalized User score |
| Pop | The average normalized values of User Score and Critic Score |

The file was saved under the name vgsales\_final. Throughout the report, R code is used to manipulate the data relative to the research question at hand. Questions 2.1, 2.2, and 3.1 add a “Console” column to ensure that each platform is correctly labeled as “Handheld” or “Stationary”. Question 3 also uses the 172 rows of data that contain both an critic and user score to create a linear model and project the missing values into 3,997 rows which only had one score available. This data set of 4,196 data points was used to decipher and organize insights into audience scores and perceptions.

In order to fairly judge sales across games in certain genres that were released in different years, the average sales per year was taken as Global\_Sales divided by 2018 minus the release year. This avg\_annual\_sales was used to accurately compare sales of games and was inserted by R code for relevant questions.