FDS Visualisation and Interpretation Exercise

B239883

18th November 2025

1 Dataset Overview

Link to dataset: `https://www.kaggle.com/datasets/gregorut/videogamesales?resource=download`

The dataset consists of 16,598 gaming titles ranked in ascending order of global sales in terms of a rank (a number from 1 to 16,5998). Each game is classified by name, platform, year of release, genre, publisher, NA sales, EU sales and Japanese sales

2 Data Exploration

I first used very basic methods like head to get a basic understanding of the data and to realise that nothing was majorly wrong with the data. I then used the method info to work out the data types of the features and realised that the year of release column was encoded as floats instead of integers. I then used the describe method to get a better understanding of the data, specifically if there was some outliers in the maximum global sales. I then checked for missing values using isnull and sum to notice 271 missing values in the "Year" column and 58 missing values in the "Publisher" column. Finally to check for duplicate rankings I used the nunique method to identify that no rankings were counted twice. After formulating my research question I cleaned the data accordingly.

3 Question

My question was, "can year of release predict global sales of games". I asked this question because I wanted to put in practice the concept of linear regression, however this turned out to not work in my favour.

4 Visualisation and Interpretation

Figure 1 is the first visualisation I created and shows the varying Global sales across time with almost all Sales across the years being relatively low, with some big outliers breaking over 20 million in Global sales.However since the dataset is so massive the visualisation is hard to read with many data point overlapping.

In figure 2 When looking at the Average global sales over time it becomes much clearer that big spikes at say 1985,1989 and 1992 correspond to widespread success in particular games as global sales spike dramatically and then quickly fall back down. However as a whole the plot shows a downward trend in global sales across time maybe implying that gaming across the globe may be fizzling out but not neceserally due to year of release. This is because the plot shows a clear non-linear realtionship meaning a low corelation between the variables.

This is evidenced further by the regression plot which shows that the year of release is a bad predictor for global sales. We can also use the summary function as shown below to obtain an R^2 value of 0.006, which is shockingly low and firmly answers the question that the year of release does not predict the

global sales, not even in the slightest. This implies that other features for exmaple genre, publisher or platforms has a much better chance at predicting the global success of a game.

.

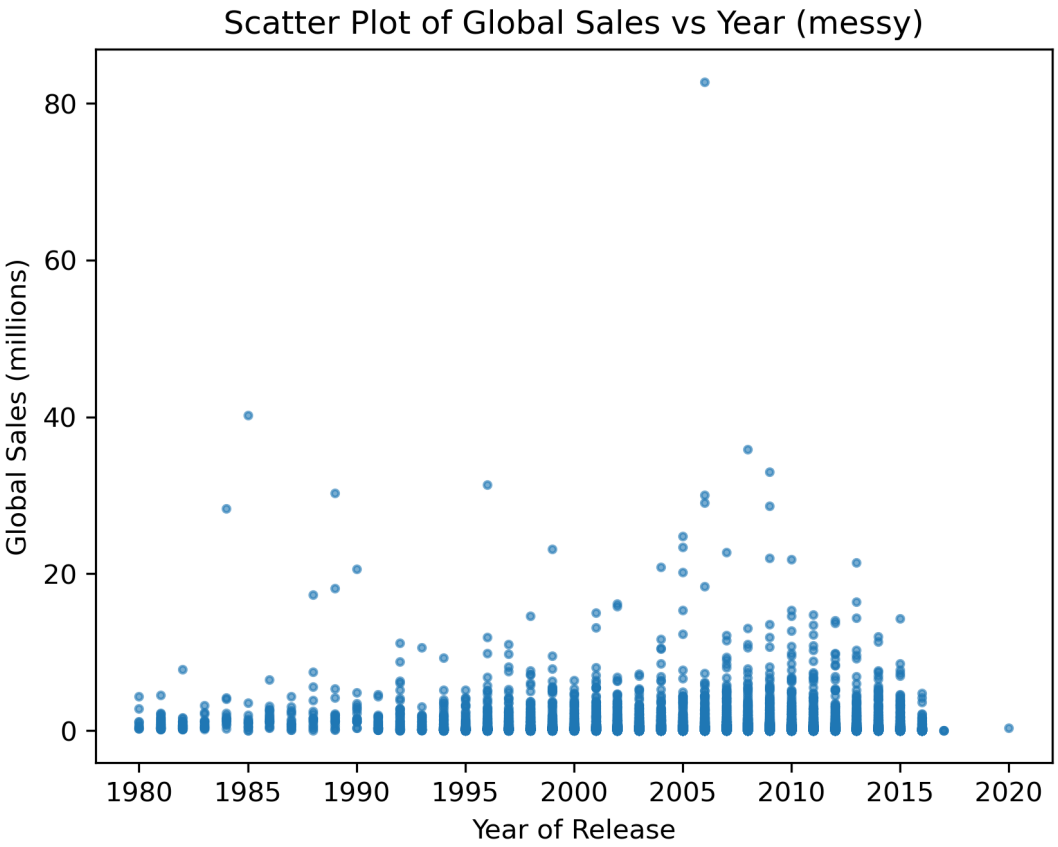

Figure 1: Messy scatter plot of the predictor variable and target variable

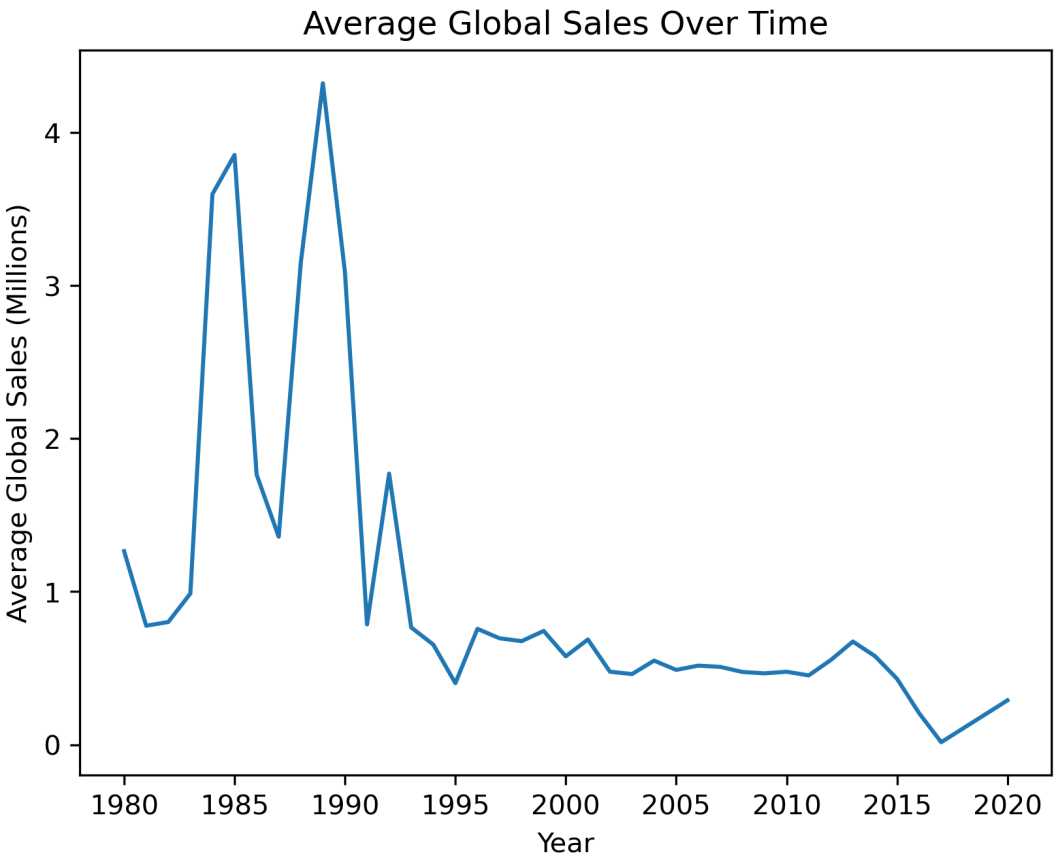

Figure 2: Clear line plot of average global sales per year

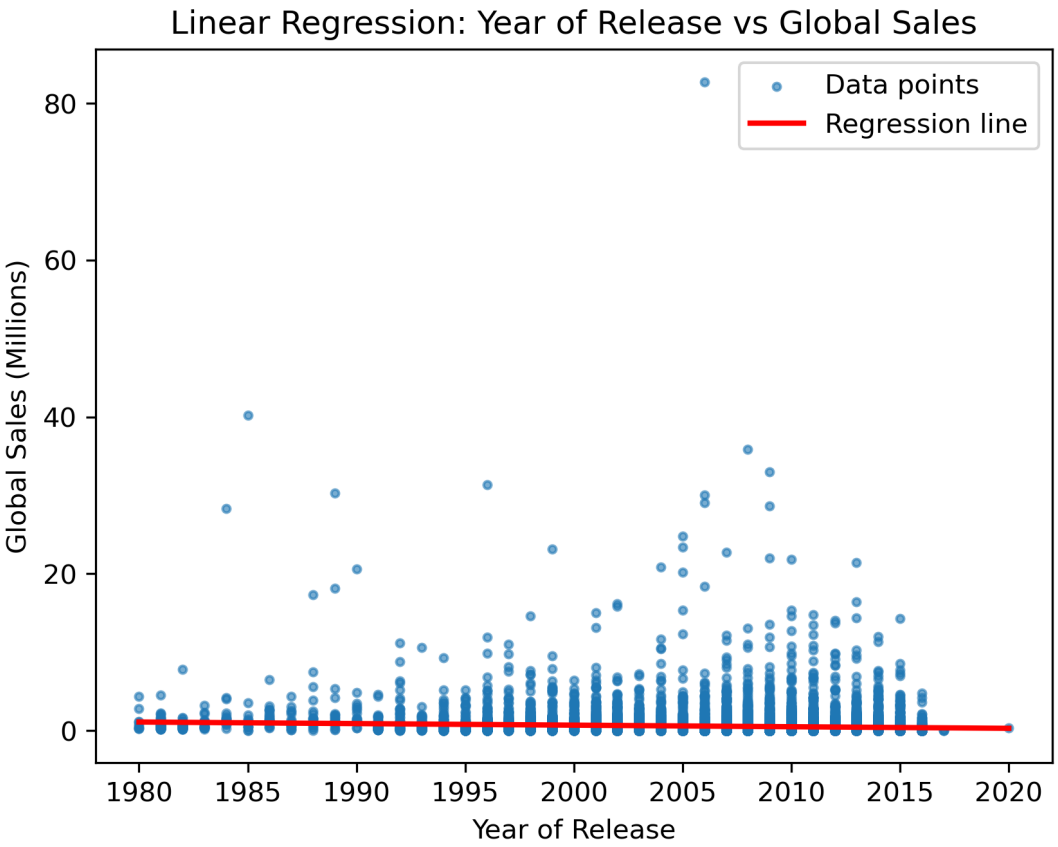

Figure 3: Regression line over the scatterplot shown in figure 1