Video Game Sales Analysis

Dataset: Video Game Sales

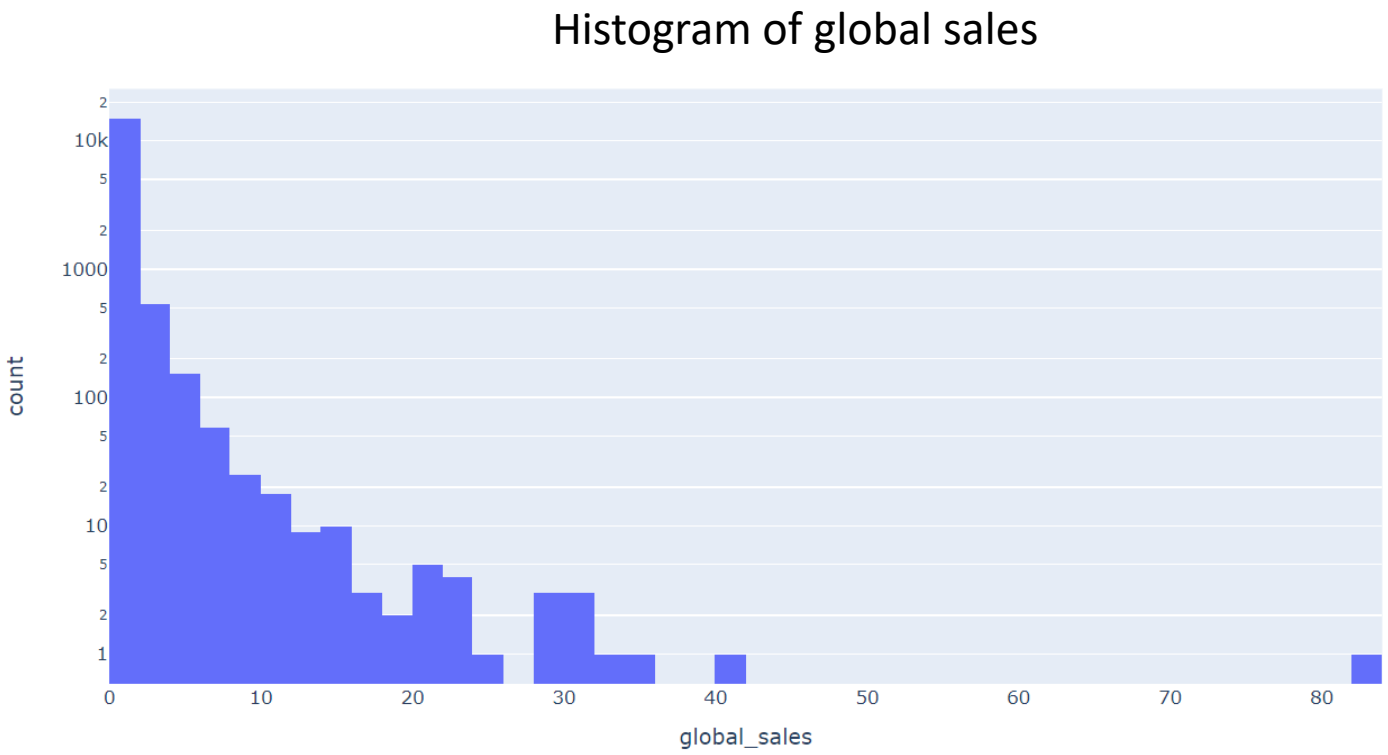
- Sourced from Kaggle, posted by Ulrik Thyge Pedersen
  - (https://www.kaggle.com/datasets/ulrikthygepedersen/video-games-sales)
- Contains 16600 entries
  - Records games from 1980 through 2020
- Each entry contains information about the game title, platform, release year, genre, publisher, and number of sales (in millions) in North America, Europe, Japan, other regions, and globally.
  - Some entries had missing information
  - Some years had negligible entries or were missing completely
    - 2017 and 2020 had single digit entries with less than a million global sales
    - 2018, 2019 had no entries

Motivations

- For this project, I wanted to determine the top publishers and top genres.
- I also tried to determine what genre was more profitable in general and what genre was more profitable for the top publisher.
- I also focused on the Action genre and tried to fit the data to a line to predict future sales.

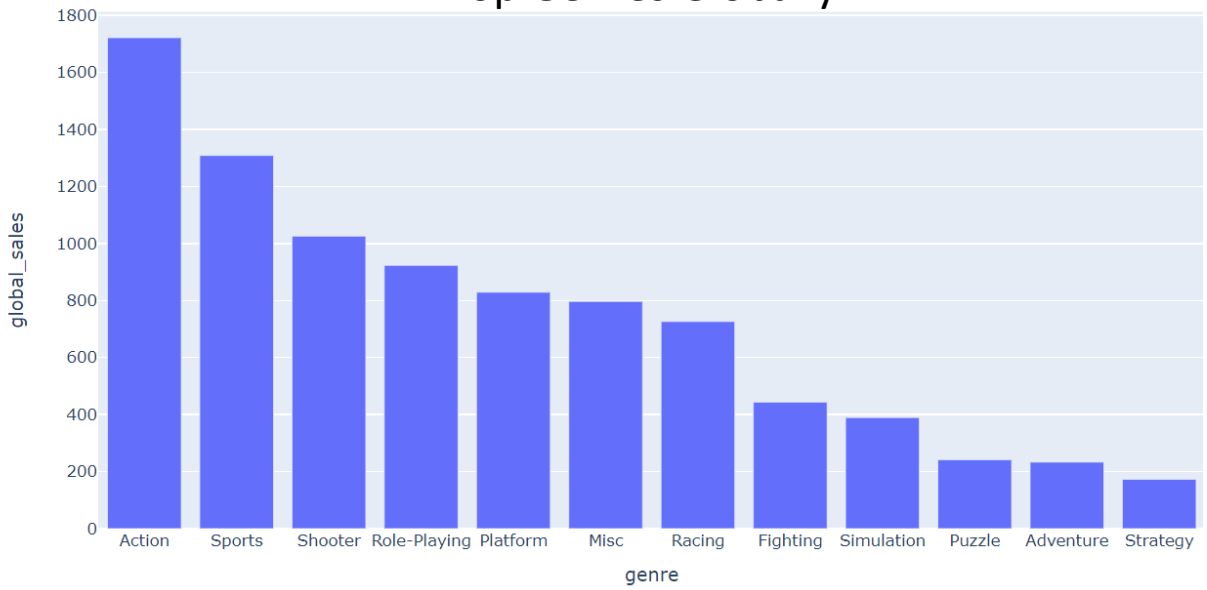
Exploring the Dataset

- Numeric summary of data does not tell us much
- Mean of global sales is about 0.5607, standard deviation is about 1.5921
- Majority of sales lies between 0 and 2 million sales.

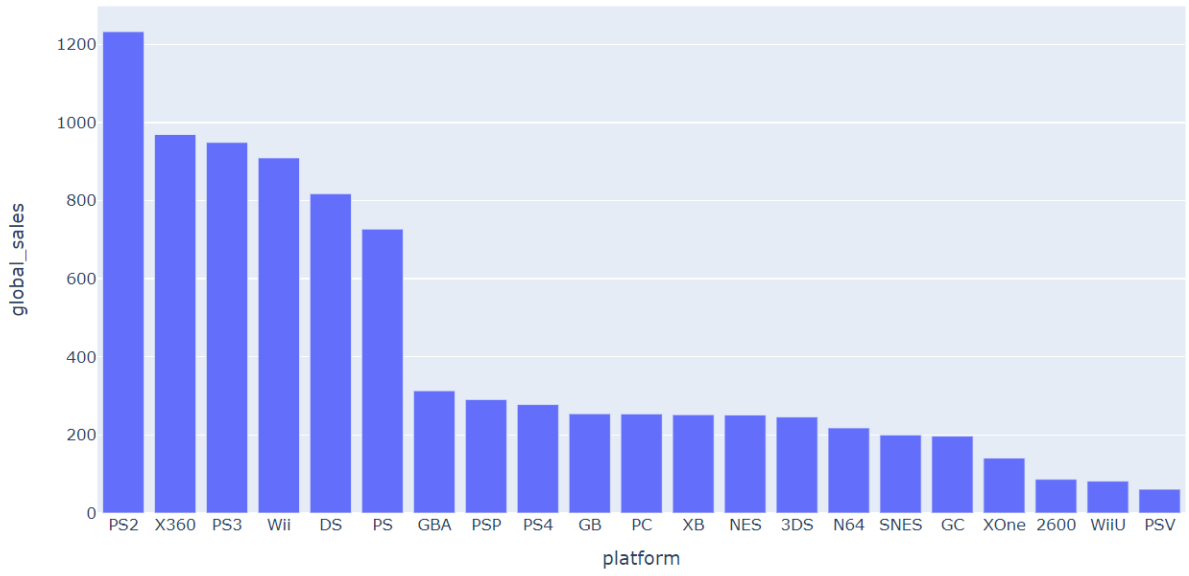


Looking at Sales

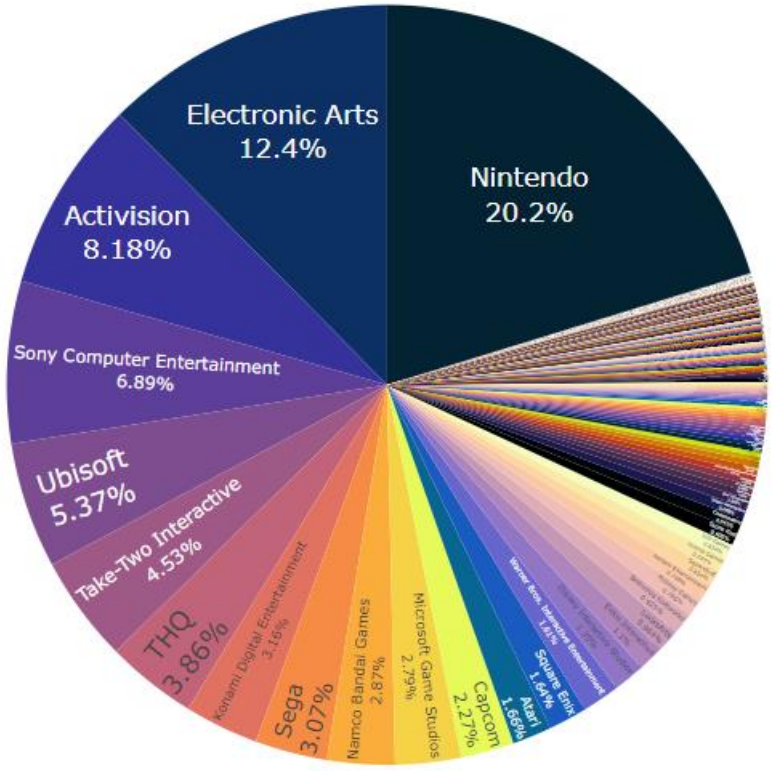
Top Genres Globally



Top Platforms Globally



Top Publishers: Share of Global Sales



Action vs Sports: Permutation Test

- We want to determine if Action games or Sports games are more profitable on average.
  - The observed difference in means was -0.033417 and the average of Sports games was greater.
- By resampling a subset of the data with only Action/Sports games, we can determine the proportion of samples where the resampled difference in means is greater than the observed.
- Result: $p = 0.2478$

Action vs Sports: Classic Hypothesis Test

- Reminder:
  - Null $H0$: The mean global sales of Action games is equal to the mean global sales of Sports games
    - $\mu_A = \mu_S$
  - Alternative HA: The mean global sales of Action games is less than the mean global sales of Sports games
    - $\mu_A < \mu_S$
- Using a t-Test: $T = -\frac{0.033417}{\sqrt{\frac{1.1802859^2}{3148}+\frac{2.1260558^2}{2255}}} = -0.67553746$ and we get $p = 0.2497$

Nintendo: Platformers or RPGs

- Nintendo is the top publisher by global sales and its top two genres are Platformers and RPGs.
- Like before, we want to determine which genre is more profitable on average, specifically for Nintendo.
  - In this case, the observed difference was 1.164083 and Platform games had a higher average.
- Result: $p = 0.0655$

Platformers or RPGs: Hypothesis Test

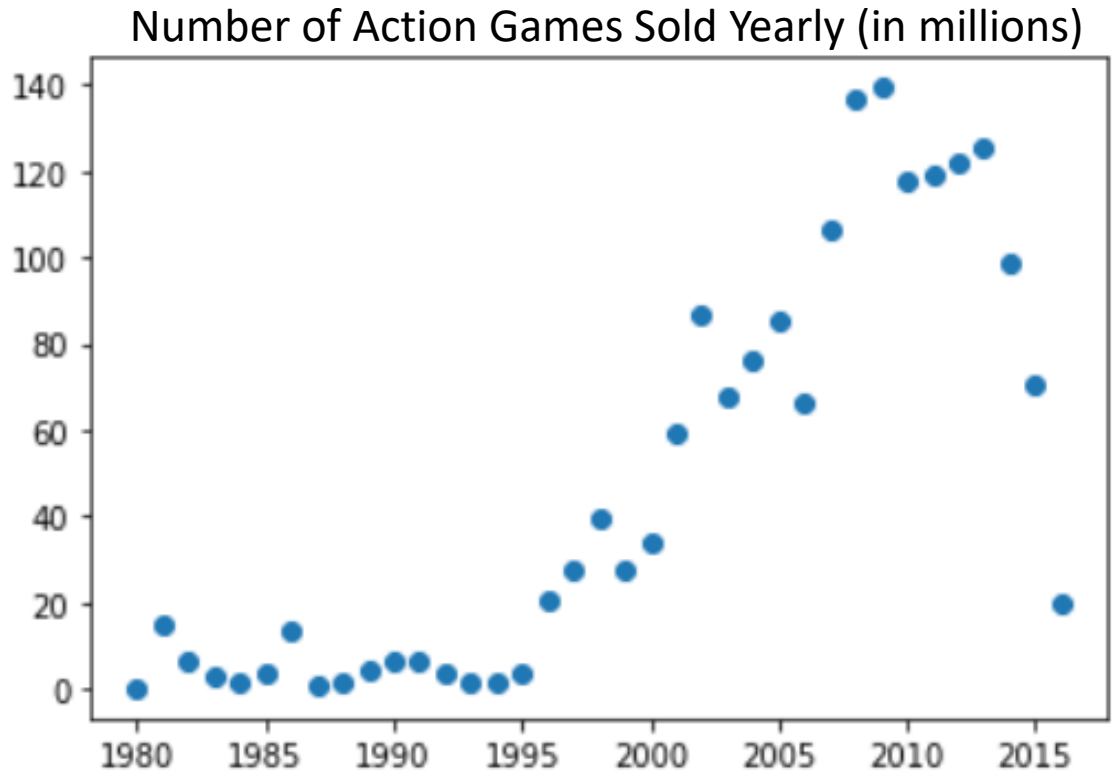
- Reminder:
  - Null $H0$: The mean global sales of Nintendo's Platformer games is equal to the mean global sales of their RPG games
    - $\mu_P = \mu_R$
  - Alternative HA: The mean global sales of Nintendo's Platformer games is greater than the mean global sales of their RPG games
    - $\mu_P > \mu_R$
- As before, we use a t-Test and get $T = 1.505$ and $p = 0.0669$

Results

- In our first test, we wanted to see if Sports games were more profitable compared to Action games, on average
  - Both tests provided very strong evidence that this was not the case
- In our second test, we wanted to see if Nintendo's Platformer games were more profitable compared to their RPG games, on average
  - Both tests had very marginal evidence that this was not the case

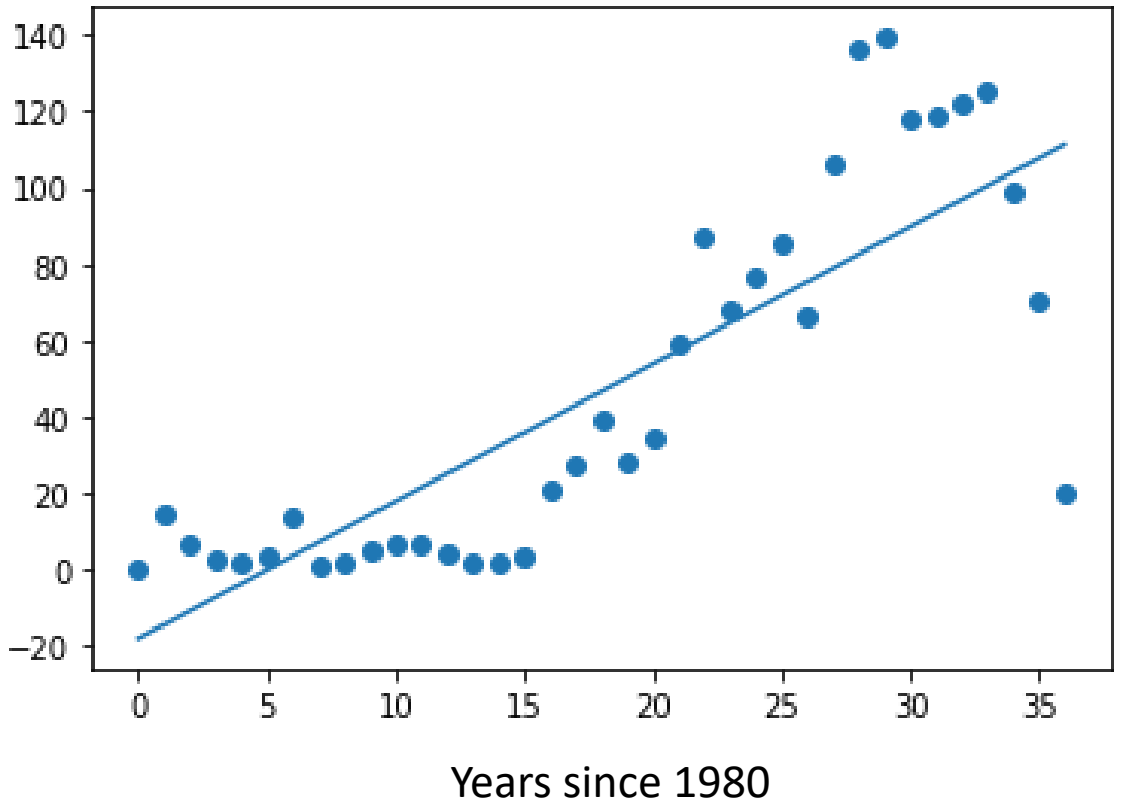
Fitting & Predicting: Linear Regression

- We would like to predict the number of sales in a year for action games.
- First, we'll try fitting our data to a line, but then also look at another fit.



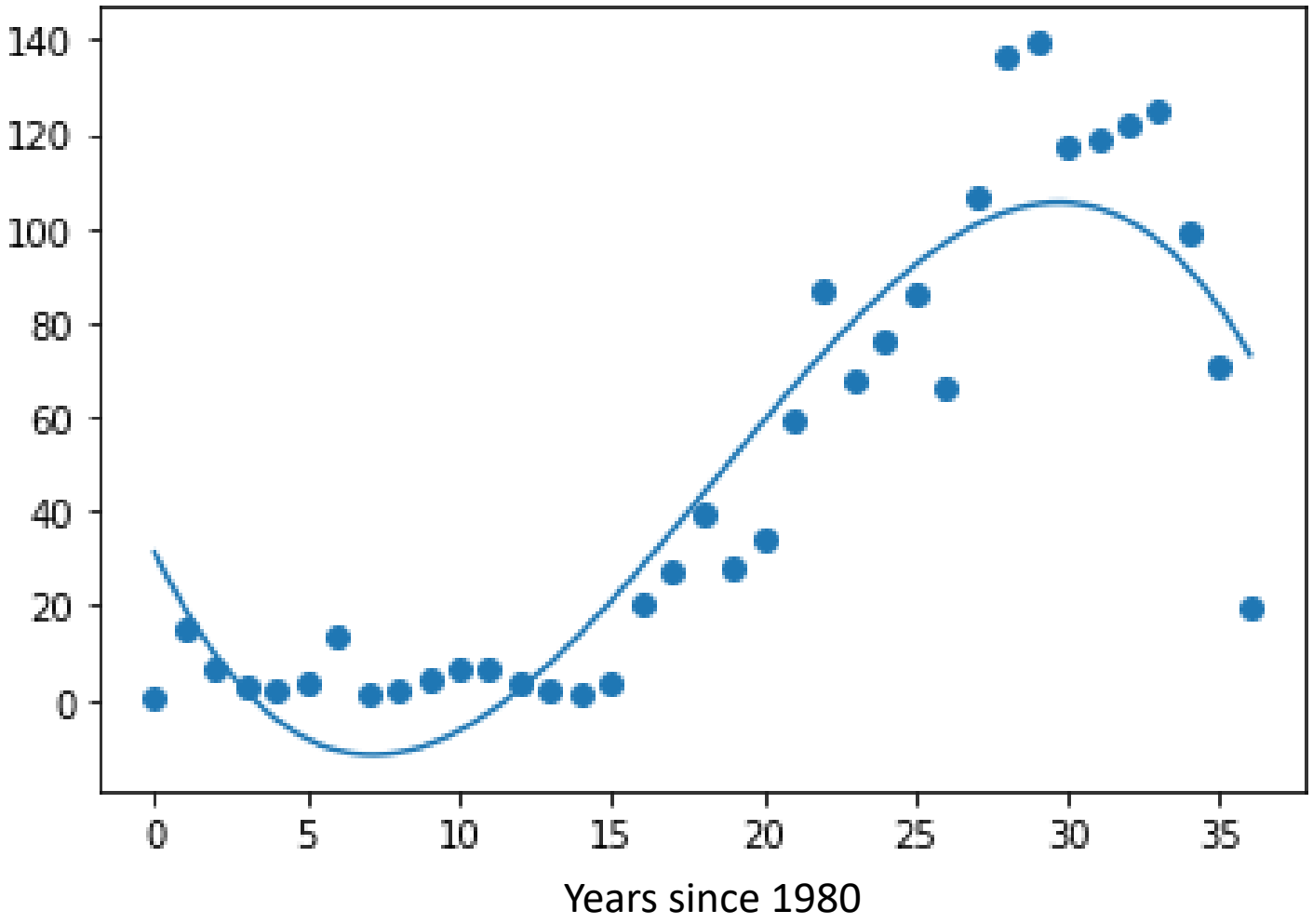
Linear Regression

- $y = -18.1257468 + 3.5923163x$
- Correlation coefficient
  - $r = 0.8178$
- Coefficient of determination
  - $r^2 = 0.6688$
- Decent correlation, however r^2 is low so the fit is not very good.
- Can also see that the sales drop towards the recent years, and a line won't capture that.



Polynomial (Cubic) Regression

- The linear fit was not very good, but a polynomial fit might do better
- To measure how good the fit is, we will use the squared error of the residuals and root mean squared error and compare it with the linear fit.



Comparison

- For polynomial regression, the squared error (SE) of the residuals is 12717.43 and the root mean squared error (RMSE) is 18.54.
- For linear regression, we get $SE = 26949.84$ and $RMSE = 26.99$
- We want the SE and RMSE to be as low as possible, so the cubic fit is better than the linear fit.

Conclusions

- I found that the top genre, globally, is Action followed by Sports
  - The results of the hypothesis testing suggested that Action games aren't more profitable on average than Sports games.
- Additionally, I determined the top publisher in number of global sales to be Nintendo
  - And there was marginal evidence that Nintendo's RPG games were less profitable on average compared to its Platformer games.
- When I tried to fit the yearly global sales for Action games, a linear fit was poor and a cubic fit was very good.
  - But the cubic fit ran into a problem of overfitting