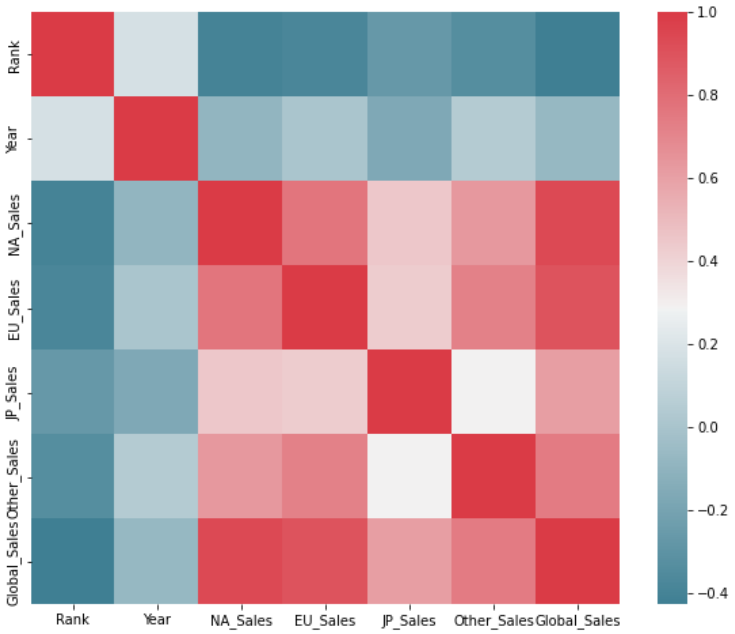
Multiple Regression Analysis on Video Game Data to Find Key Factors in Determining Sales Worldwide

Abstract: The history of video games dates back to the 1950s. Having great interest in a variety of games and consoles, I wanted to see how video games evolved to what we play these days. With video game sales data from Kaggle, I performed a series of data analysis to discover what kind of games tend to lead the market and which factors affect global sales and regional sales. Then, the results are explained with plots and graphs to visually demonstrate the relationship among variables in video game sales dataset.

Introduction: The purpose of the project is to determine how sales of the video games are compared and correlated to other factors such as publisher, genre, year, and other sales. The reason for choosing this topic is that the dataset used for the analysis satisfied all the criterias that determine the appropriate dataset. The criterias includes choosing a dataset that grabs my interest, having enough columns/rows, and preferably having little or no NaN values in the data. In order to determine the correlation of data, I used EDA to describe the data and visualization to illustrate it.

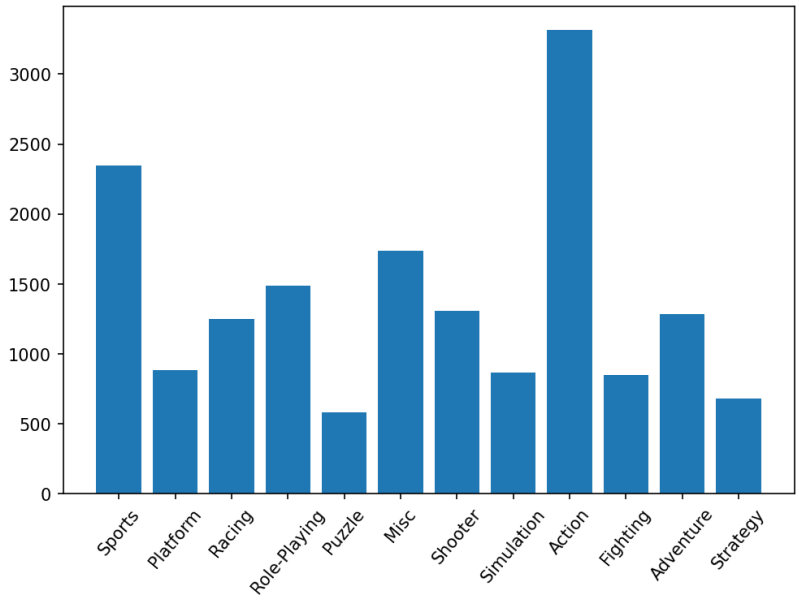
Data/EDA: The data selected for this project is “Video Game Sales” from Kaggle.com. The columns of the data are Rank, Name, Platform, Year, Genre, Publisher, NA_Sales, EU_Sales, JP_Sales, Other_Sales, and Global_Sales. It contains 11 columns and 11493 rows. The columns Name, Platform, Genre, and Publisher have string value, and the columns Rank, Year, NA_Sales, EU_Sales, JP_Sales, Other_Sales, and Global_Sales have double value. There are few outliers in sales columns that are much higher than other values, but they can’t be removed since they are important to consider. There are 271 null values in the Year column, and 58 of those in the Publisher column.

First data analyzed and visualized in the project is the correlation between each column. This is visualized using heatmap which displays the correlation of columns using red and blue. Red represents high correlation, and blue represents low correlation. The maximum value for correlation is 1, and the minimum value is -1. According to the heatmap, the pair of two distinct columns that has maximum correlation value is NA_Sales and Global_Sales. On the other hand, the pair with minimum correlation value is Global_Sales and Rank.

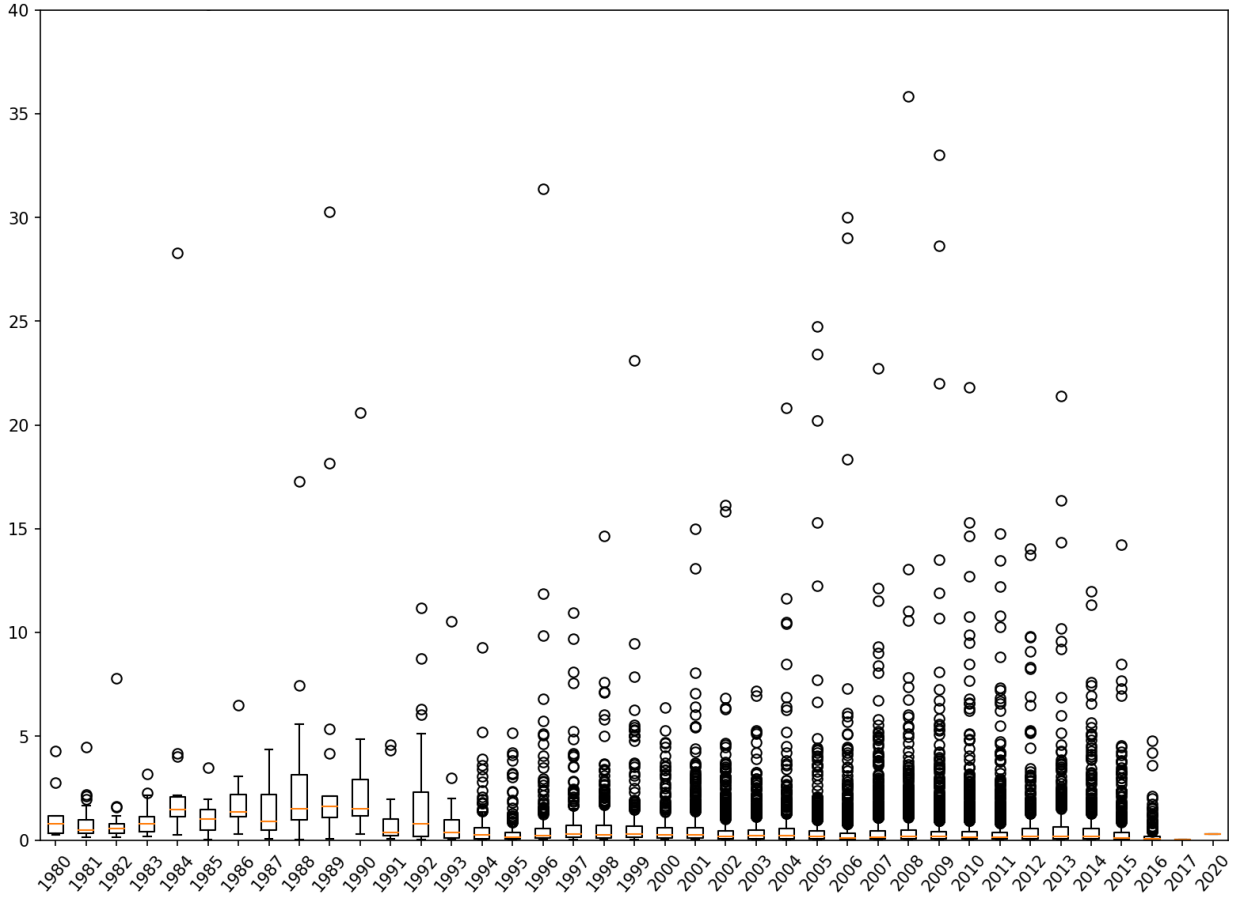


In addition to the heatmap, I used data.shape and a LinearRegression object from sklearn.linear_model library to find the detailed correlation values and predict the global sales. I found out that the correlations of Global_Sales to NA_Sales, EU_Sales, and Other_Sales are 0.941269, 0.903264, and 0.747964 respectively. Then, I constructed a linear regression model with columns Year, NA_Sales, EU_Sales, JP_Sales, and Other_Sales to predict global sales. The R^2 value of the linear regression model is 0.999988895, which is how much the model can describe the global sales. To test the model, I put the data of Wii Sports, the most popular game, into the model. The model displayed 82.734 for the predicted global sales, which is similar to the actual value, 82.74.

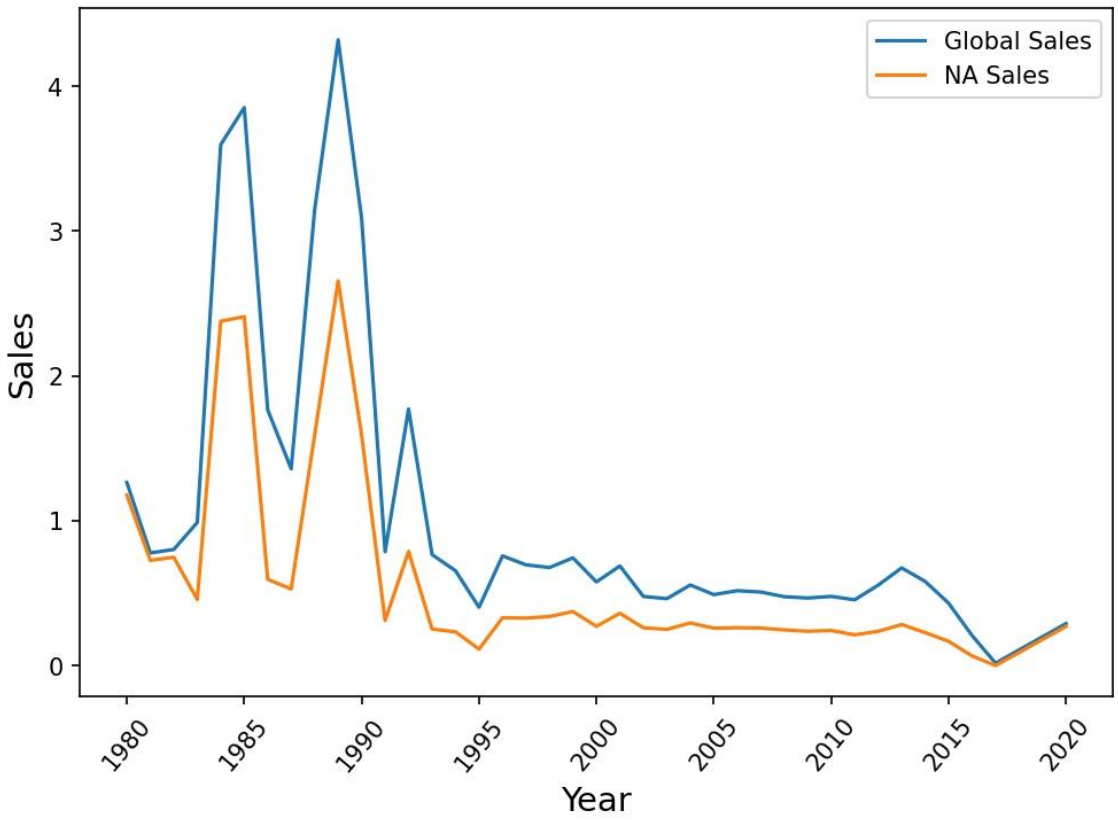
Then, I counted the number of each value from the Genre column and visualized it using a bar graph. The reason why I chose a bar graph is because its structure compares the frequency of genres clearly. I used the Counter object from the collection library to count the frequency of each genre. The bar graph displays that the most frequent genre is Action, followed by Sports, Mist, and Role-Playing. Puzzle is the least frequent genre according to the analysis.



Next, I visualized the values of the Global_Sales column corresponding to values in the Year column using a box plot. The year included in the box plot is between 1980 and 2020. The box plot demonstrates the median, interquartile range, and outlier values of global sales each year. The year with the highest median value is 1989, and the year with the largest interquartile range is 1992. The number of outliers is not many until 1993, and it increases drastically from 1994. The highest outlier value is located in the year 2008.



Finally, I analyzed the mean global sales and NA sales each year. To compare these two mean values distinctly, I visualized them in the same plot using line graphs. In order to calculate the mean values, I used the mean object from the statistics library. The graph demonstrates that the mean values of global sales is higher than the mean values of NA sales, and they follow similar trends throughout the domain. According to the graph, the maximum mean value of global sales and NA sales are about 4.3 and 2.6 respectively.



Conclusion: The project used several methods to find the specific factors that decide the global sales of video games. To visualize the correlation of each factor and global sales, I used heatmap which represents it with the difference of color. Then, the linear regression model calculated the estimated global sales according to the crucial factors, and I tested the model to find out that it is accurate to the actual global sales value. In addition, I visualized the data using different methods such as bar graph, box plot, and line graph to display the relationship between each factor. Finally, I added categorical variables on the model which uses variables such as publisher and genre to determine the global sales. By adding the categorical variables with one-hot encoded representation, the model achieved higher R^2 value, being able to describe the global sales more.