Data Analysis to Video Game Sales Data

Qifang Weng
Department of Data Analytic, Long Island University
Gurpreet Singh
20th December 2022

Abstract

Since the era of information and technology, video games are slowly becoming a norm to the young people and attracting everyone around them. After the year 2000, video games were developed in a more polished and under a more creative environment. Started from thousands, now millions of developers and programmers are in the Computer Science field. Not only Video games bring joy and happiness to the players and people, but the experiences are absolutely indescribable depending on the genres or styles. For the people born in the 2000s, video games gradually became an instinct or considerably joined daily routine. In recent years the video game market is in volatile condition, therefore, requires in depth research and marketing analysis. Furthermore, specific overview on the chosen historical video game company market and provide decision making advice as well as making prediction or forecast of the sales in the market between different continents.

Dataset Description:

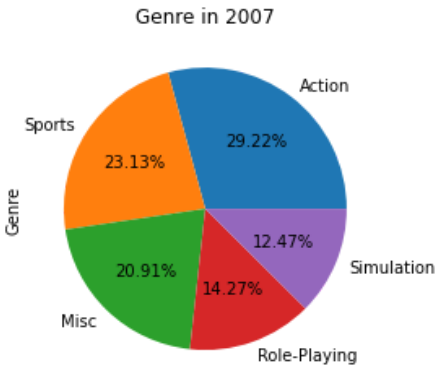
The dataset was originally collected through vgchartzfull. Vgchartzfull is a python script based on BeautifulSoup. It creates a dataset based on data based off the site(1). Beautiful Soup is a Python package for parsing HTML and XML documents. It creates a parse tree for parsed pages that can be used to extract data from HTML, which is useful for web scraping.

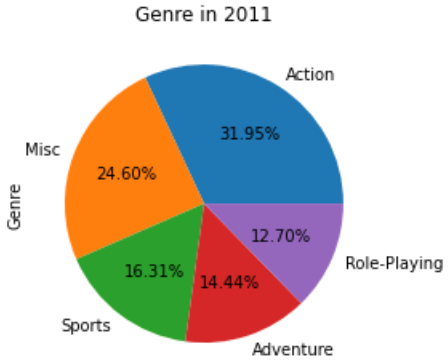
Data columns:

- **Rank** - Ranking of overall sales
- **Name** - The games name
- **Platform** - Platform of the games release (i.e. PC,PS4, etc.)
- **Year** - Year of the game's release
- **Genre** - Genre of the game
- **Publisher** - Publisher of the game
- **NA_Sales** - Sales in North America (in millions),unit of copies
- **EU_Sales** - Sales in Europe (in millions),unit of copies
- **JP_Sales** - Sales in Japan (in millions),unit of copies
- **Other_Sales** - Sales in the rest of the world (in millions),unit of copies
- **Global_Sales** - Total worldwide sales.

Dataset analysis, exploration, visualization, feature correlation, insight observed from data visualizations.

1. **The top 5 Genre: Throughout 2007-2011 Genre diversification**

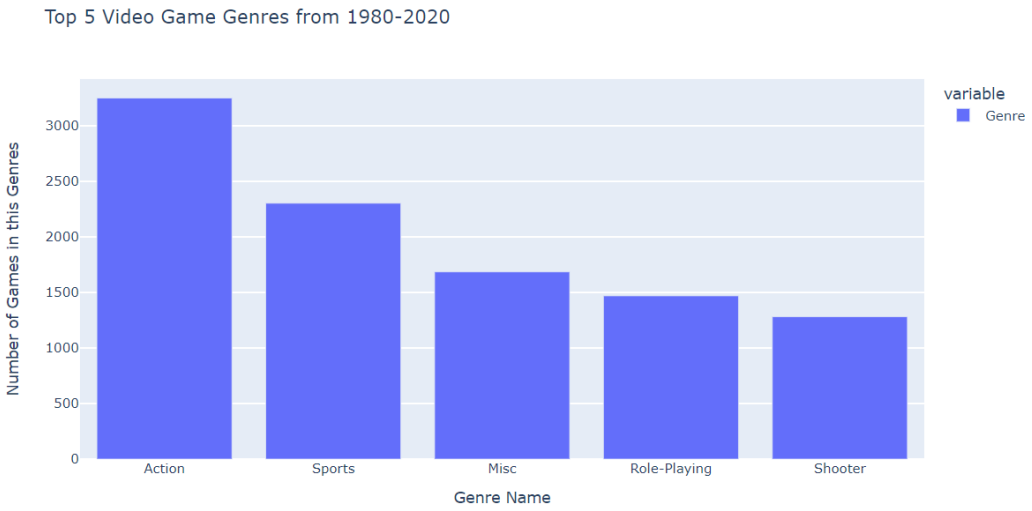


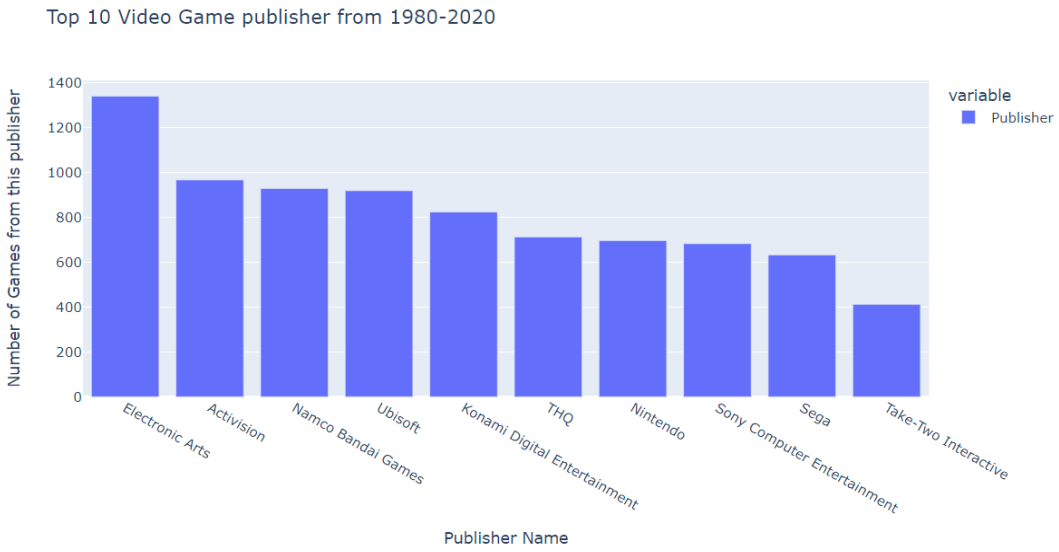


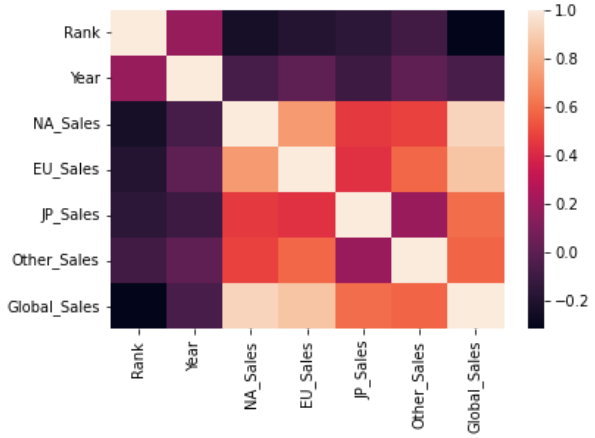
According to the pie chart visualization, the following genre count had positive increases within the four year period: Action, Misc. On the other hand, these genre counts had negative decreases within the four year period: Sports, Role-Playing. Last, the simulation genre had

dropped out of the top 5 genre count and surpassed by the genre Adventure. Referring to the dataset, the years from 2007-2011 had the highest number of games published. In 2007, there were a total of 1201 games, in 2008, there were 1428 games, in 2009, there were 1431 games, in 2010, there were 1257 games, and in 2011 there were 1136 games published.

2. Top 5 game genre and publisher and heatmap of correlation matrix

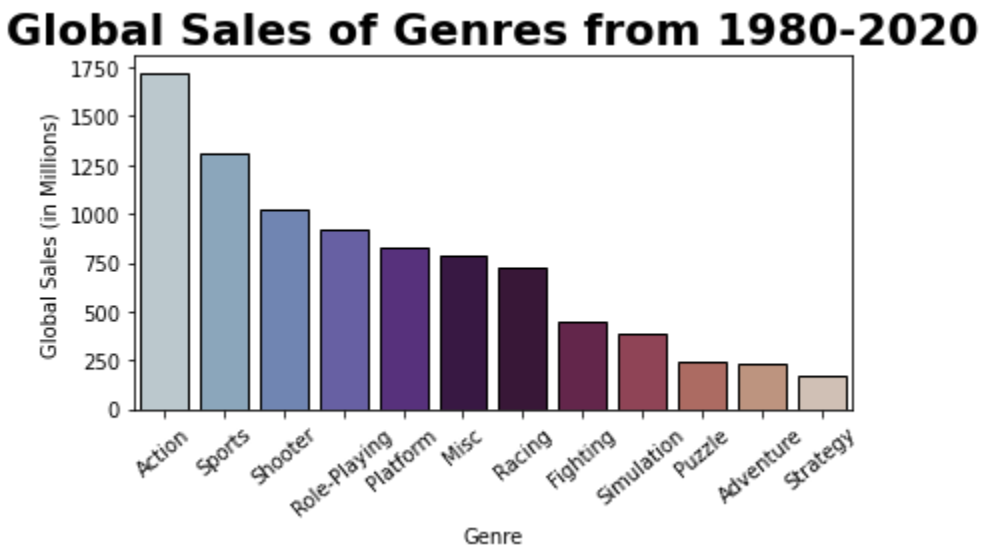






According to the visualizations above, Action and sport genre games are very popular throughout the period from 1980-2020. Derived that publisher Electronic Art and Activision published most games from 1980-2020, hence they are the targeted companies to be analyzed and create decision making recommendations with. As long as the correlation matrix heatmap goes, there are some relations between columns : NA_Sales, EU_Sales, JP_Sales and Other_Sales; however, not significant enough to be dropped according to the multicollinearity theory.

3. **More balanced visualization between genre and sales unit of copies.**



Referring to Global Sales based on genres: Action, sports, shooter, role-playing,platform genre had the most success in sales in units of copies. Therefore, companies and publishers tend to look at the sales and decide which genre of games to produce, hence capturing the most sales for their games. Furthermore, provided the dataset contains the most successful publishers such as Electronic Art and Activision, visualizations can be designed based on their historical sales and genres.

Next, there are many types of machine learning algorithms that can be used for forecasting and making predictions including the classification models. Classification models such as decision trees will use the acquired dataset and apply on the overall publishers and global sales to make predictions. For building the decision tree classification model, there are a set of rules to follow. First, separate the original dataset into a train dataset and a validation dataset. The machine learning technique will use the train dataset to view algorithms and apply it onto

the validation dataset or test set. At the end of the model, the confusion matrix is utilized to conclude the accuracy of the prediction and brings a percentage of accuracy to the user.

The first decision model will separate the dataset to 80 percent train and 20 percent test, and the second decision model will separate the dataset to 60 percent and 40 forty percent test. Additionally, the first model use the x columns: “JP_Sales” and “EU_Sales” to predict the y column “NA_Sales”, and the second model utilizes the x columns: “Other_Sales” and “EU_Sales” to predict the y column “NA_Sales”. The purpose of doing such prediction is to estimate the North America sales by other regions. Since there was some correlation between the region in terms of sales.

4. **Decision Tree model accuracies and total market sales by continent**

Model 1: train:Confusion Matrix (Accuracy 0.9506), test: Confusion Matrix (Accuracy 0.9401)

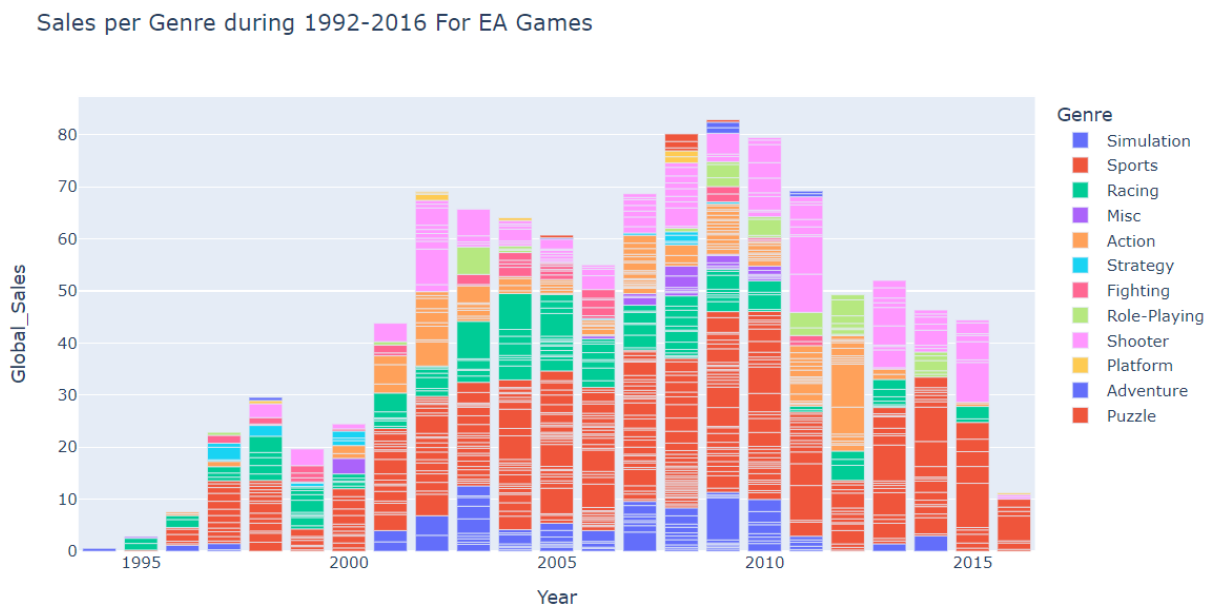
Model 2: train:Confusion Matrix (Accuracy 0.9486), test: Confusion Matrix (Accuracy 0.9450)

```
NA_Sales       4327.65
JP_Sales       1284.27
EU_Sales       2406.69
Other Sales     788.91
```

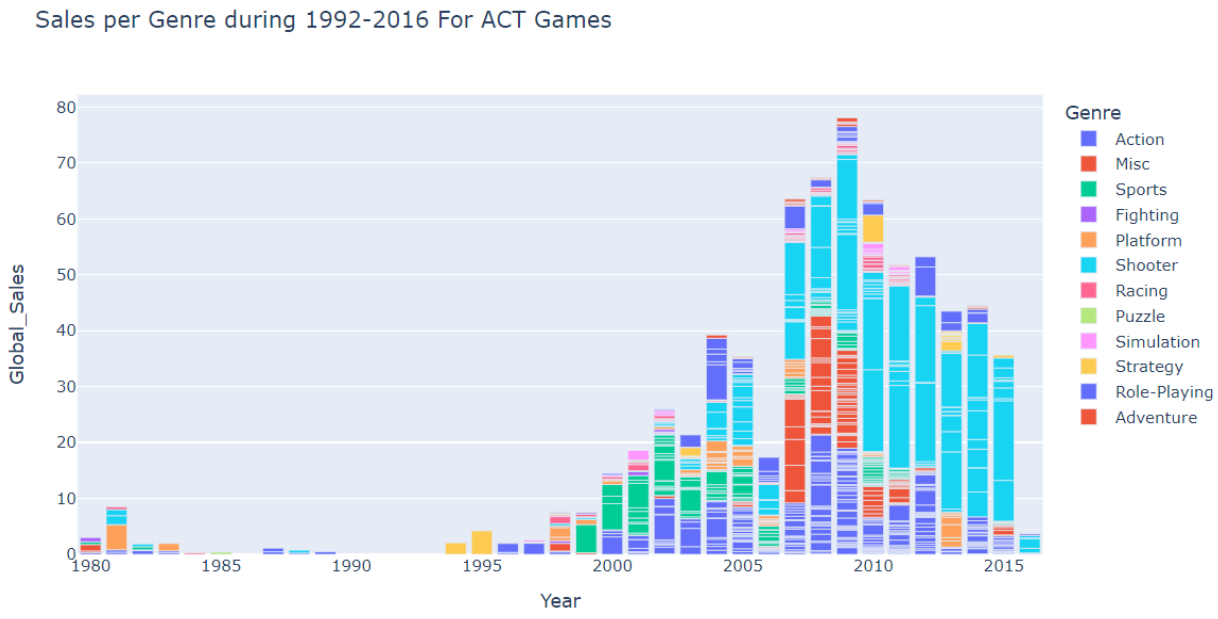
The accuracy demonstrates the performance of the model using corresponding columns. According to the Model 1, utilizing “JP_Sales” and “EU_Sales” the predictor columns to predict North America sales has above 94.01 percent on rates of success. Similarly, Model 2 utilizing “Other_Sales” and “EU_Sales” to predict North America sales has a 94.5 percent on rate of success. Thus, model 2 has greater accuracy at predicting the North America sales.

According to the calculation of each region's sales count, North America and Europe have the superior market compared to other regions. Therefore, the recommended region to release games along with exceptional advice in the targeted market is North America and Europe. Finally, The purpose of building the 2 models is to see the overall relationship on sales among the chosen 3 continents.

5. Overall global sales to Electronic Art and Activision



Genre	Global_Sales
Action	107.88
Adventure	4.09
Fighting	29.80
Misc	17.53
Platform	5.47
Puzzle	4.24
Racing	141.57
Role-Playing	34.36
Shooter	154.06
Simulation	86.66
Sports	451.14
Strategy	12.87



Genre	Global_Sales
Action	141.72
Adventure	5.41
Fighting	2.86
Misc	76.55
Platform	33.40
Puzzle	1.06
Racing	16.97
Role-Playing	46.78
Shooter	295.34
Simulation	8.26
Sports	75.16
Strategy	17.70

In recommendation and conclusion, the **Overall global sales to Electronic Art and Activision** figures showed that publisher Electronic Art should further produce more games in Genre : Sports and Shooter. Since the largest proportion of their sales are in these 2 genres. The next tiers of genres Electronic Art should produce is: Racing and Actions. Demonstrated by publisher Activision visualization, for publisher Activision, their sales mainly concentrated in genres: Shooter, Action. The tier 2 genre to produce for publisher Activision is Misc and sports.