CS 6140- MACHINE LEARNING

FINAL PROJECT

BY: SHIVANSH KASHYAP

TITLE: VIDEO GAMES SALES

PROBLEM STATEMENT

The primary goal of this project is to delve into the dataset to understand how the data is distributed in the gaming industry. This includes analyzing the sales figures, the frequency of game releases, and the diversity of genres and platforms. By analyzing the data, aim is to provide insights that could inform future decisions in the video game industry. This could include identifying trends in consumer preferences, successful genres or platforms, and potential growth markets.

INTEREST IN THE PROBLEM

Video game sales analysis is fascinating for several reasons, touching upon various aspects of business, technology, culture, and psychology. The video game industry is a multi-billion dollar sector that surpasses the revenue of industries like movies and music. Understanding sales patterns can reveal significant economic trends and inform investment and development strategies within the industry. Analyzing video game sales can shed light on consumer preferences and behavior. It can help in understanding what types of games are popular among different demographics, how gaming trends evolve over time, and what factors most influence purchasing decisions. The video game industry is often at the forefront of technological innovation. Sales analysis can highlight how technological advancements affect market trends and consumer adoption.

APPROACH USED

I have implemented methods to summarizing the basic features of the dataset with simple summaries and visualizations. This includes calculating mean, median, mode, range, and standard deviation for sales figures. I have also examined data over time to identify any consistent patterns or trends. This is useful in understanding how video game sales have evolved, how genres or platforms have gained or lost popularity, and seasonal trends.

Segmentation analysis has also been performed on the data based on various criteria like genre, platform, region, and publisher. This helps in understanding the performance of different segments and identifying which are the most lucrative or growing.

All of the above have been produced with data visualization using libraries such as Matplotlib and seaborn.

Regression analysis is also performed on the dataset to predict sales based on various factors and to understand the strength and type of relationships between variables. Both linear and multiple regression models have been used on the data.

RATIONALE BEHIND THE APPROACH

The used methods help in Identifying patterns over time, crucial in an industry where consumer preferences and technology evolve rapidly. Understanding trends is key to predicting future market behaviour. The video game market is diverse, with variations across genres, platforms, and regions. Segmenting the data allows for a more nuanced understanding of different market segments and their unique characteristics. By comparing different elements, such as games or time periods, you can identify what works well and what doesn't, helping in strategic decision-making and product development. Before establishing causation, it's important to identify if there's a relationship between variables. This can guide more detailed analyses and hypothesis testing. Regression Techniques were implemented to quantify the impact of various factors on sales and predicting future sales. This approach is central to understanding how different variables interact to influence outcomes. Complex data can be more easily understood through visual representation. It's an essential part of data analysis, facilitating the communication of insights to non-technical stakeholders.

LIMITATIONS FACED IN THE PROJECT

The video game industry is subject to rapid changes in technology and consumer references, making it challenging to predict long-term trends.There was Incomplete data which could lead to inaccurate conclusions.

As the dataset has been made using web scrapping on external data sources like sales tracking websites,it can limit the scope of analysis.

Complex models, if not properly regulated, can overfit to the training data and perform poorly on unseen data.

External factors like economic conditions, cultural trends, and marketing efforts are often difficult to quantify and incorporate into the analysis.

As we used consumer data, especially from online sources, privacy and ethical considerations had to be taken into account.

SETUP FOR THE PROJECT

Dataset:

The dataset is called vgsales which has been created by web scrapping data from vgchartz.com. it consists of columns such as

- 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)
- 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.

The script to scrape the data is available at https://github.com/GregorUT/vgchartzScrape.

It is based on BeautifulSoup using Python.

There are 16,598 records. 2 records were dropped due to incomplete information.

```
    Name             Platform             Year        Genre
Length:16323       Length:16323       2009   :1431   Length:16323
Class :character   Class :character   2008   :1428   Class :character
Mode  :character   Mode  :character   2010   :1259   Mode  :character
                                      2007   :1202
                                      2011   :1139
                                      2006   :1008
                                      (Other):8856
 Publisher            NA_Sales          EU_Sales          JP_Sales
Length:16323       Min.   : 0.0000   Min.   : 0.0000   Min.   : 0.00000
Class :character   1st Qu.: 0.0000   1st Qu.: 0.0000   1st Qu.: 0.00000
Mode  :character   Median : 0.0800   Median : 0.0200   Median : 0.00000
                   Mean   : 0.2655   Mean   : 0.1476   Mean   : 0.07868
                   3rd Qu.: 0.2400   3rd Qu.: 0.1100   3rd Qu.: 0.04000
                   Max.   :41.4900   Max.   :29.0200   Max.   :10.22000

 Other_Sales        Global_Sales
Min.   : 0.00000   Min.   : 0.0100
1st Qu.: 0.00000   1st Qu.: 0.0600
Median : 0.01000   Median : 0.1700
Mean   : 0.04834   Mean   : 0.5403
3rd Qu.: 0.04000   3rd Qu.: 0.4800
Max.   :10.57000   Max.   :82.7400
```

Implementation

Libraries used are:

1. **numpy**: Used for numerical computing and array manipulations.
2. **pandas**: Essential for data manipulation and analysis, particularly for working with structured data.
3. **scipy.stats**: Part of SciPy, this library provides functions for statistical operations.
4. **matplotlib**: A plotting library for creating static, interactive, and animated visualizations in Python.
5. **seaborn**: A data visualization library based on matplotlib, providing a higher-level interface for drawing attractive and informative statistical graphics.
6. **missingno (msno)**: A library specifically used for visualizing missing data in datasets.
7. **sklearn.preprocessing** Part of the scikit-learn library for model training.

Computing environment

Operating System: Windows 10

Language : Python 3

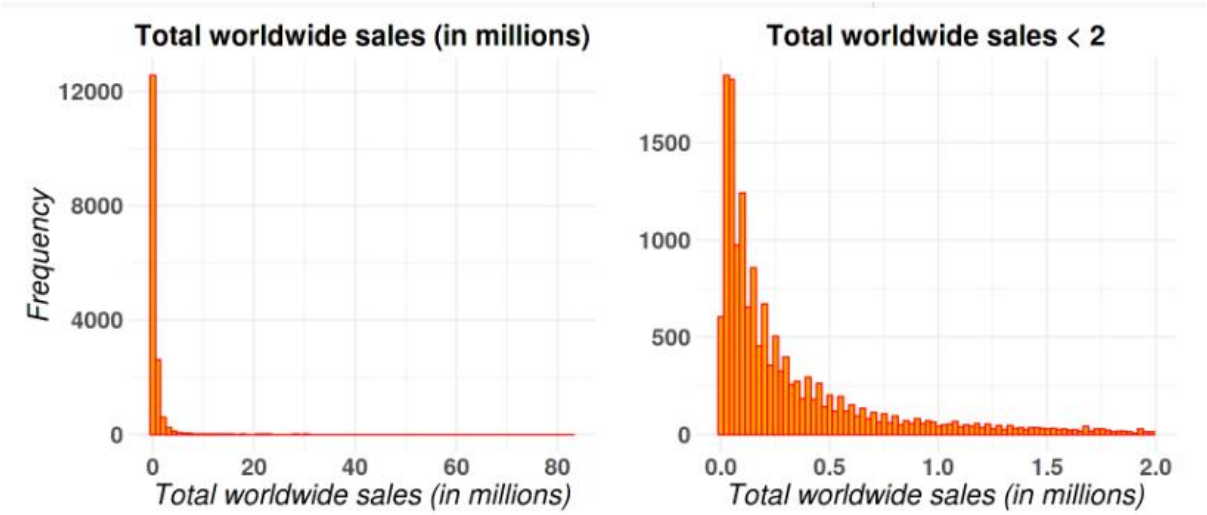
Compiler : Google Colab(basic)

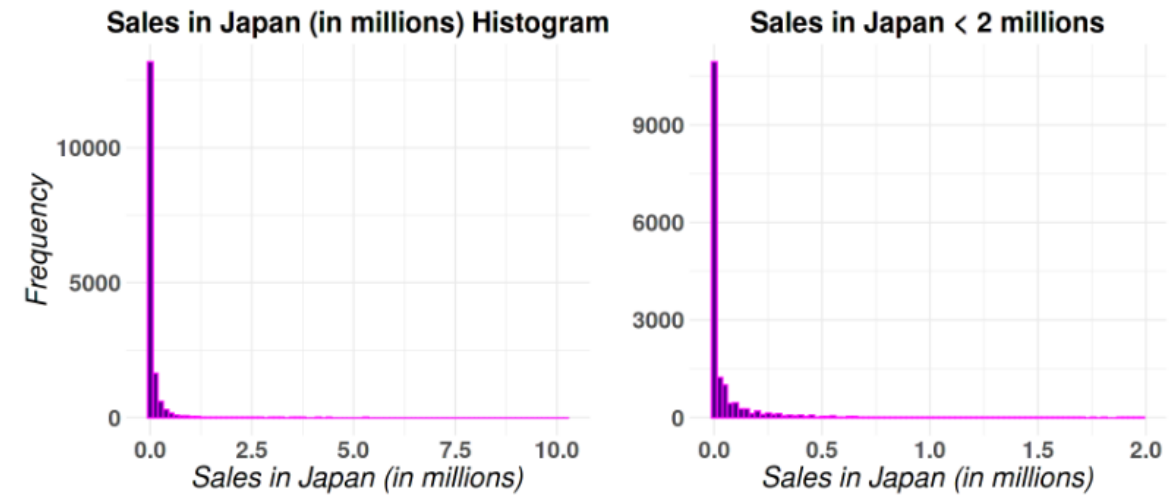
Limitations

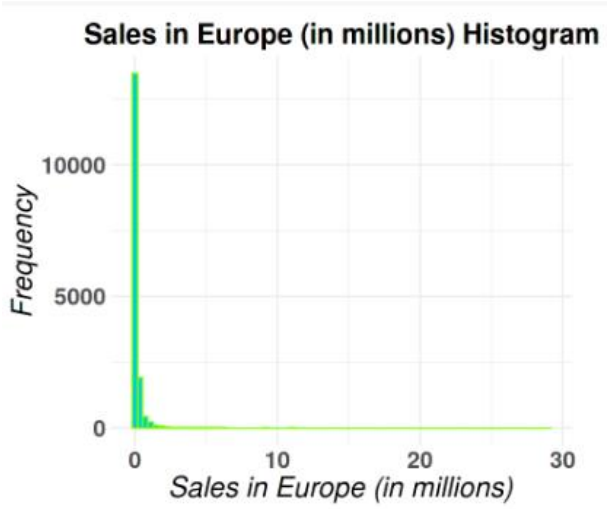
The project faced difficulties in some the analysis, such as data constraints or the inability to capture certain market dynamics. Recommendations for future research, such as incorporating additional variables like marketing spend, digital vs. physical sales, could be Utilized for better results.

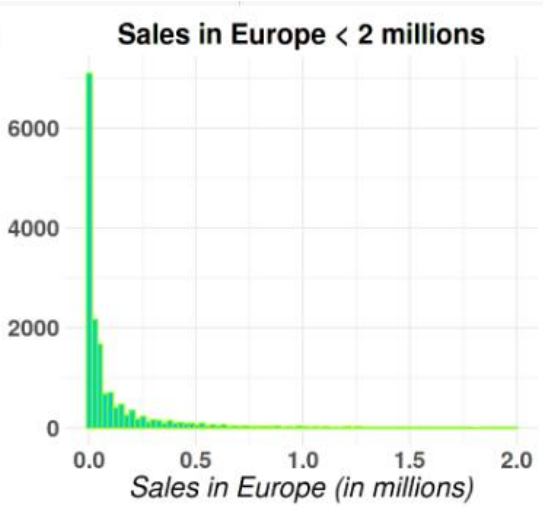
RESULTS

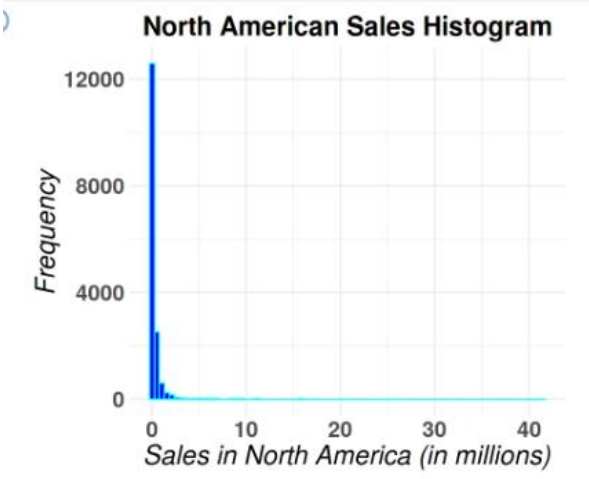
Sales as per region

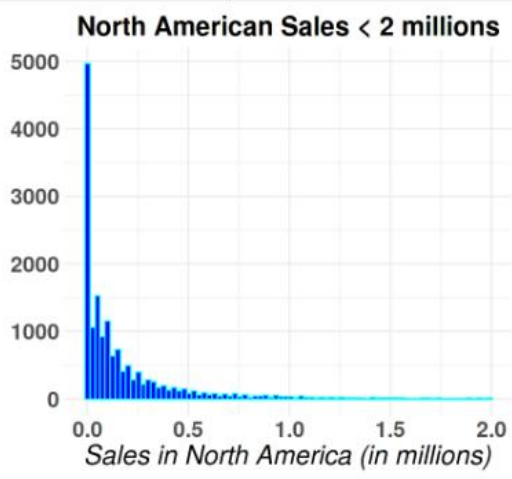




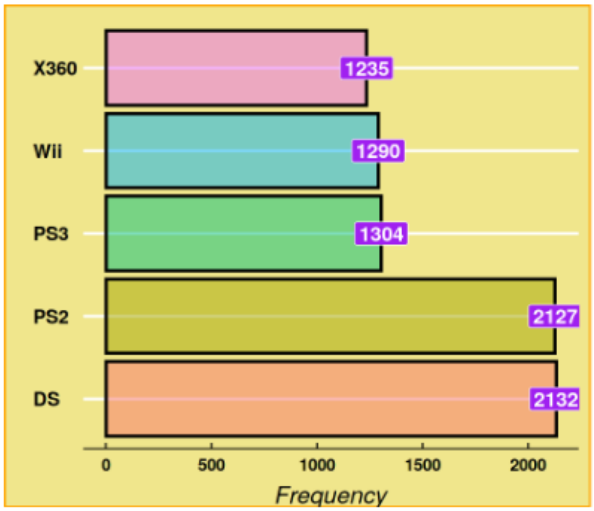


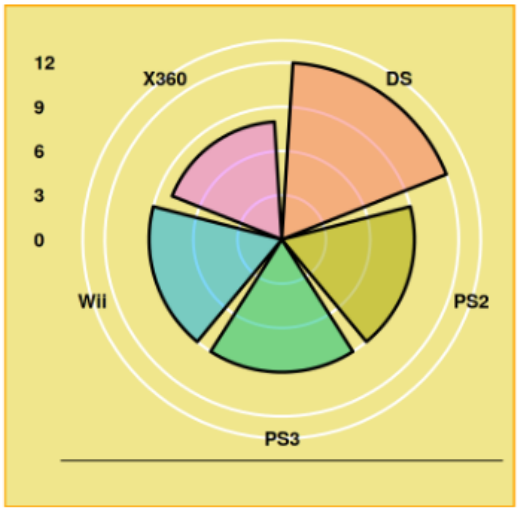


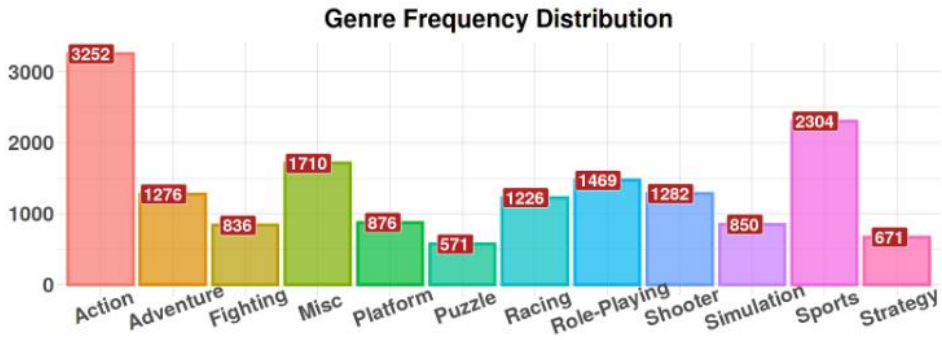




Frequecy of games released on platforms



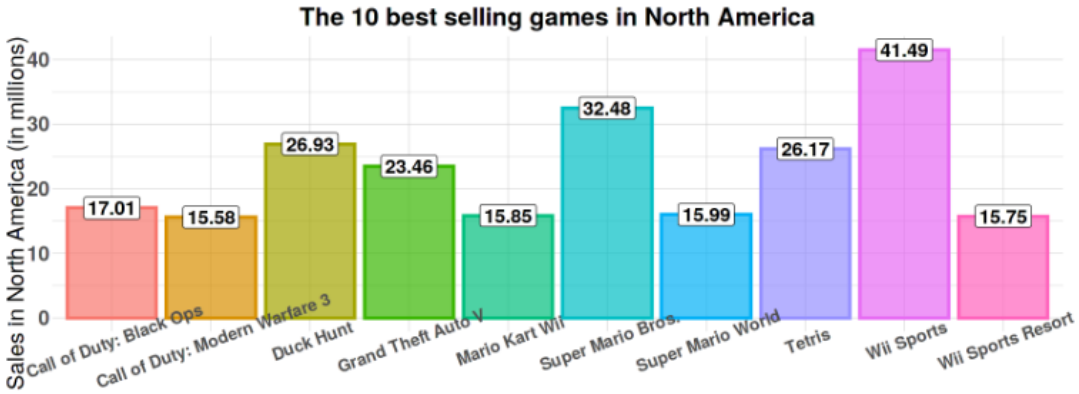


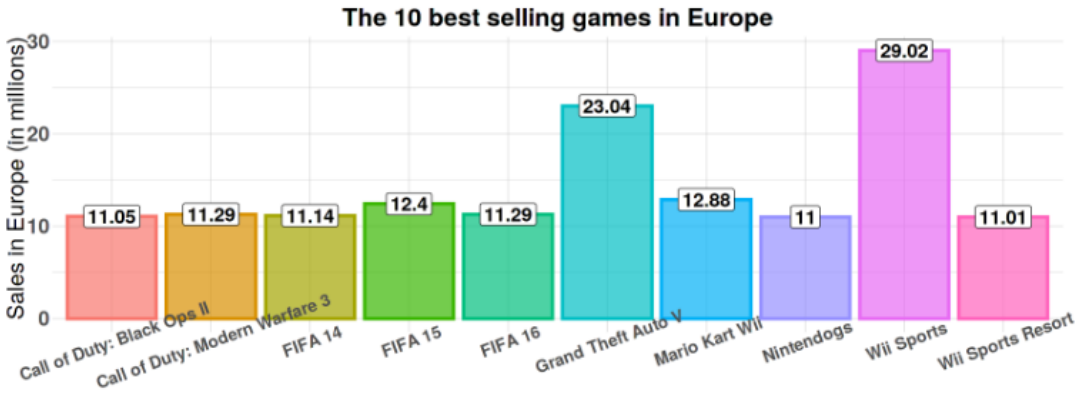


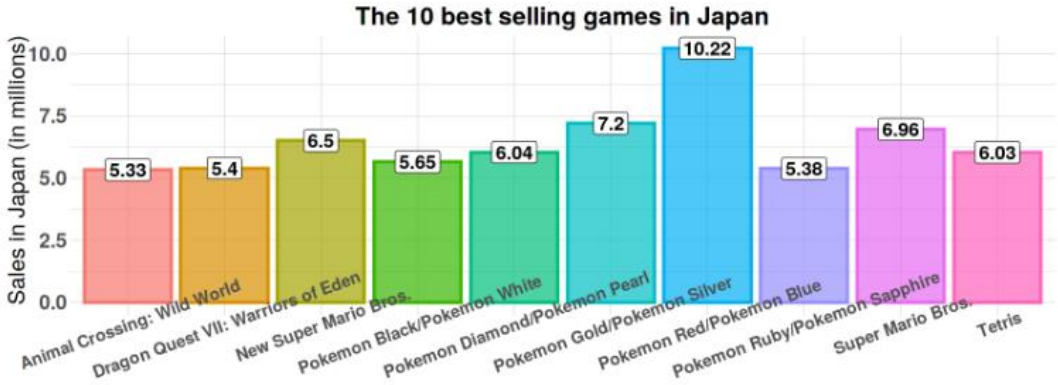
A data.frame: 5 × 3

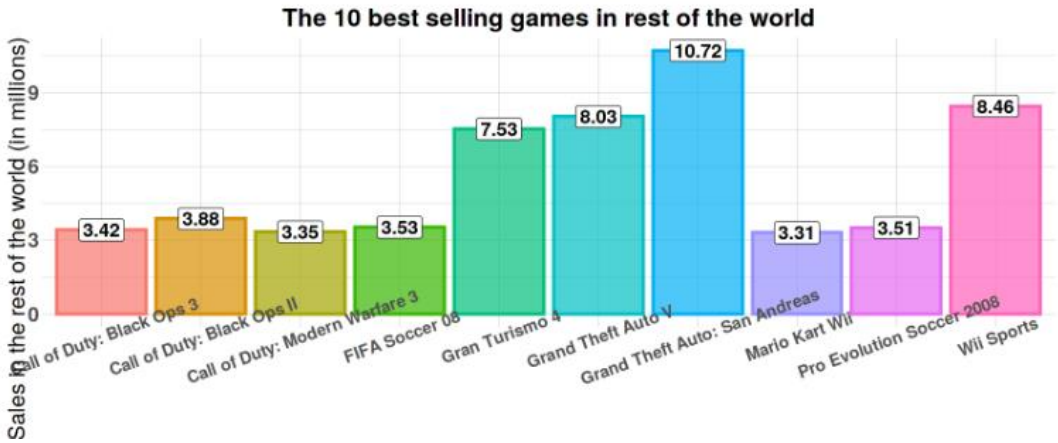
	DM	Variance	std
	<dbl>	<dbl>	<dbl>
NA_Sales	0.30947311	0.67516400	0.8216836
EU_Sales	0.19126483	0.25890063	0.5088228
JP_Sales	0.11670815	0.09709044	0.3115934
Other_Sales	0.06172115	0.03606483	0.1899074
Global_Sales	0.59452820	2.45206289	1.5659064

Best selling games of all time


The 10 best selling games in North America
Sales in North America (in millions)
0
10
20
30
40
17.01
15.58
26.93
23.46
15.85
32.48
15.99
26.17
41.49
15.75
Call of Duty: Black Ops
Call of Duty: Modern Warfare 3
Duck Hunt
Grand Theft Auto V
Mario Kart Wii
Super Mario Bros.
Super Mario World
Tetris
Wii Sports
Wii Sports Resort


The 10 best selling games in Europe
Sales in Europe (in millions)
0
10
20
30
11.05
11.29
11.14
12.4
11.29
23.04
12.88
11
29.02
11.01
Call of Duty: Black Ops II
Call of Duty: Modern Warfare 3
FIFA 14
FIFA 15
FIFA 16
Grand Theft Auto V
Mario Kart Wii
Nintendogs
Wii Sports
Wii Sports Resort

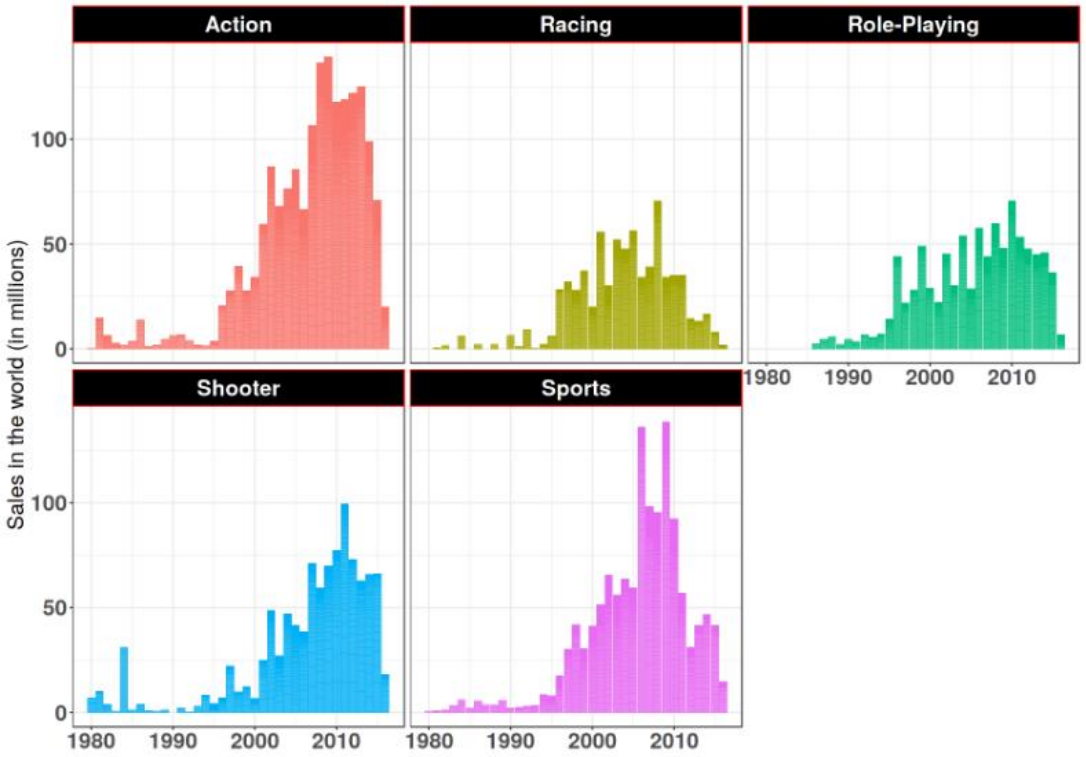




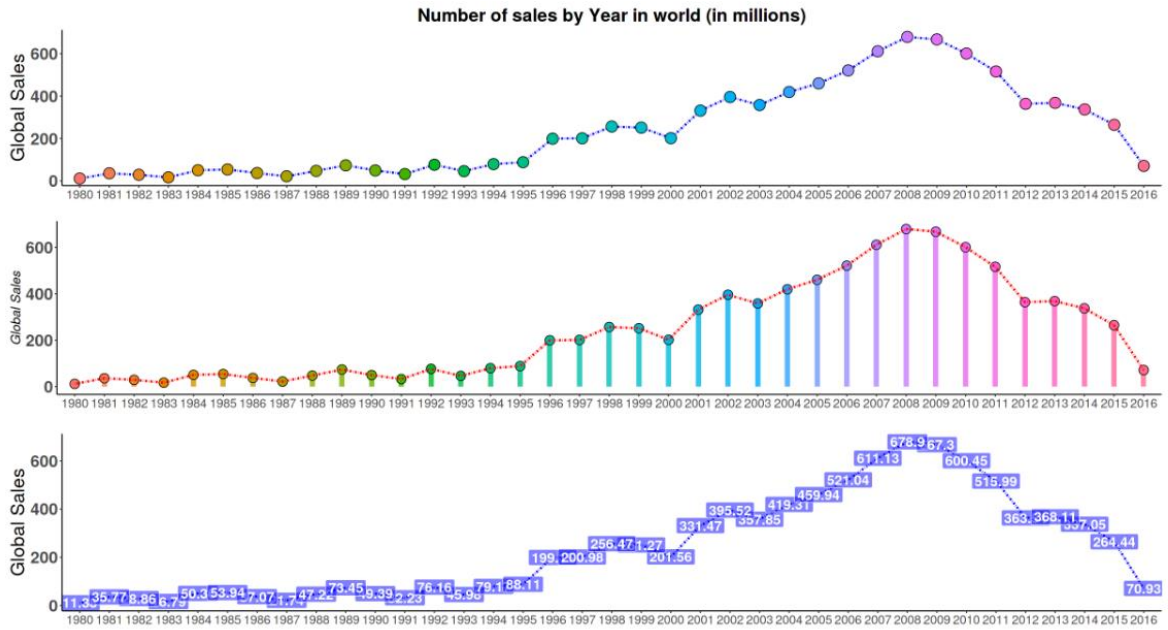
Best selling games from 1980 to 2016

	Name	Global_Sales
	<chr>	<dbl>
1	Wii Sports	82.74
2	Grand Theft Auto V	55.92
3	Super Mario Bros.	45.31
4	Tetris	35.84
5	Mario Kart Wii	35.82
6	Wii Sports Resort	33.00
7	Pokemon Red/Pokemon Blue	31.37
8	Call of Duty: Modern Warfare 3	30.83
9	New Super Mario Bros.	30.01
10	Call of Duty: Black Ops II	29.72

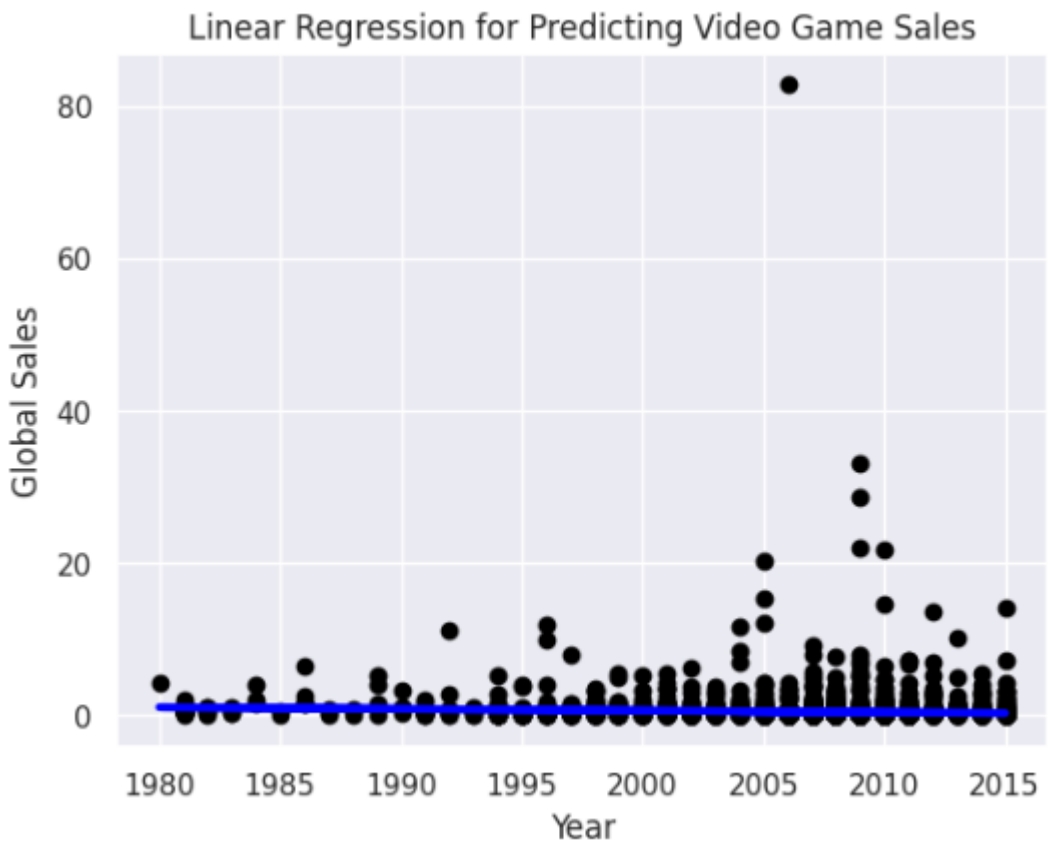
Sales by genre



Sales by year



Linear Regression


Linear Regression for Predicting Video Game Sales
Global Sales
0
20
40
60
80
1980
1985
1990
1995
2000
2005
2010
2015
Year

Quantitative Regression results

```
                         QuantReg Regression Results
==============================================================================
Dep. Variable:           Global_Sales   Pseudo R-squared:              0.01188
Model:                       QuantReg   Bandwidth:                     0.05749
Method:                 Least Squares   Sparsity:                      0.5550
Date:                Thu, 14 Dec 2023   No. Observations:              16250
Time:                        05:08:50   Df Residuals:                  16238
                                        Df Model:                         11
================================================================================================
                                   coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------------------
Intercept                       0.2000      0.005     40.744      0.000       0.190       0.210
Genre[T.Adventure]             -0.1400      0.009    -15.131      0.000      -0.158      -0.122
Genre[T.Fighting]               0.0100      0.011      0.927      0.354      -0.011       0.031
Genre[T.Misc]                  -0.0400      0.008     -4.821      0.000      -0.056      -0.024
Genre[T.Platform]               0.0800      0.011      7.559      0.000       0.059       0.101
Genre[T.Puzzle]                -0.0900      0.013     -7.196      0.000      -0.115      -0.065
Genre[T.Racing]            -1.167e-06      0.009     -0.000      1.000      -0.018       0.018
Genre[T.Role-Playing]          -0.0100      0.009     -1.137      0.256      -0.027       0.007
Genre[T.Shooter]                0.0300      0.009      3.267      0.001       0.012       0.048
Genre[T.Simulation]            -0.0400      0.011     -3.747      0.000      -0.061      -0.019
Genre[T.Sports]                 0.0300      0.008      3.957      0.000       0.015       0.045
Genre[T.Strategy]              -0.1000      0.012     -8.486      0.000      -0.123      -0.077
================================================================================================
Quantile: 0.75
                         QuantReg Regression Results
==============================================================================
Dep. Variable:           Global_Sales   Pseudo R-squared:              0.02020
Model:                       QuantReg   Bandwidth:                     0.05377
Method:                 Least Squares   Sparsity:                      1.789
Date:                Thu, 14 Dec 2023   No. Observations:              16250
Time:                        05:08:50   Df Residuals:                  16238
                                        Df Model:                         11
================================================================================================
                                   coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------------------
Intercept                       0.5100      0.014     37.135      0.000       0.483       0.537
Genre[T.Adventure]             -0.3500      0.026    -13.413      0.000      -0.401      -0.299
Genre[T.Fighting]               0.0400      0.030      1.320      0.187      -0.019       0.099
Genre[T.Misc]                  -0.1000      0.023     -4.302      0.000      -0.146      -0.054
Genre[T.Platform]               0.3098      0.030     10.485      0.000       0.252       0.368
Genre[T.Puzzle]                -0.2000      0.035     -5.722      0.000      -0.269      -0.131
Genre[T.Racing]                 0.0300      0.026      1.152      0.249      -0.021       0.081
Genre[T.Role-Playing]           0.0300      0.025      1.218      0.223      -0.018       0.078
Genre[T.Shooter]                0.2300      0.026      8.963      0.000       0.180       0.280
Genre[T.Simulation]            -0.0900      0.030     -3.020      0.003      -0.148      -0.032
Genre[T.Sports]                 0.0500      0.021      2.355      0.019       0.008       0.092
Genre[T.Strategy]              -0.2300      0.033     -6.988      0.000      -0.295      -0.165
================================================================================================
```

DISCUSSION

By running these various analysis and regression models we can learn that

1) There is a peak in popularity of a genre over a certain period of time irrespective of the region of the world.
2) The number of physical copies of video games has been steadily declining since 2014 due to improvement to internet capabilities.

3) The highest sales have been observed during 2008-2009. This was a time period of economic decline which in turn was highly favorable to the video game industry.
4) This shows that people were adapting video games as a source of entertainment as it is the cheapest form of media with the longest amount of entertainment provided.
5) Strategy games have been the least popular genre through history of gaming with the occasion high selling video game.
6) The highest sales of video game WII Sprots which is rated E for everyone implies sales can be boosted by targeting all age groups of people.

CONCLUSION

The project reveal how video game sales have evolved over the years, identifying peak periods of sales and possible reasons behind them such as technological advancements, popular game releases. The quantile regression analysis focusing on game genres and sales might uncover which genres are consistently top performers and how this varies across different sales quantiles. This could indicate which genres are niche but have dedicated fanbases versus those with widespread appeal.

The project might concludes which platforms have been most lucrative over time and how platform popularity impacts game sales. The project also highlights regional variations in video game sales, showing preferences for certain genres or platforms in different parts of the world.

The linear regression model offers predictive insights, such as forecasting future trends in game sales based on historical data. For stakeholders in the video game industry, such as game developers and marketers, these findings could inform strategic decisions about game development, marketing, and distribution.

The combined analysis provides a comprehensive overview of the video game market, offering valuable insights into consumer preferences, market trends, and potential future developments in the industry.