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

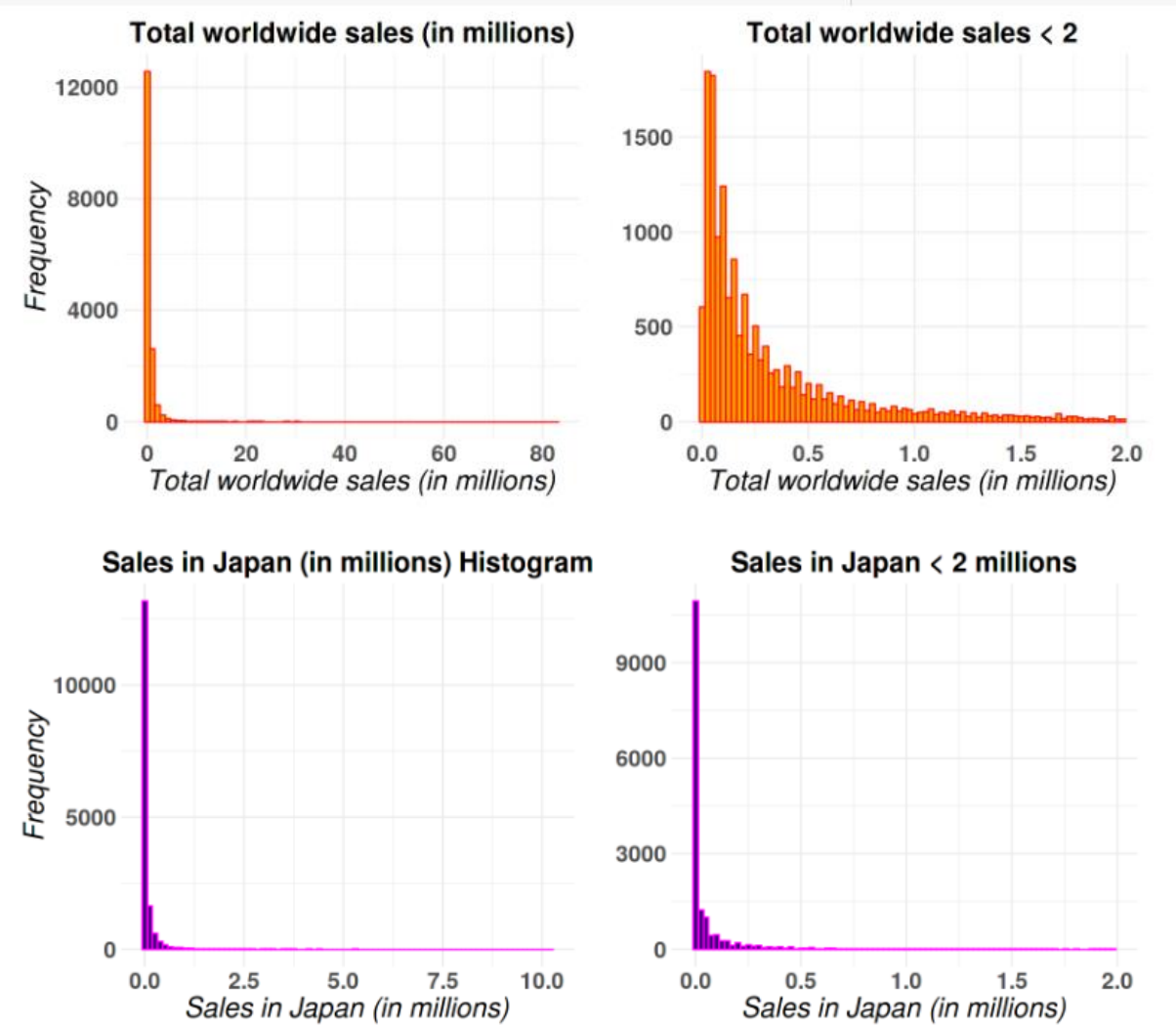
Operating System: Windows 10

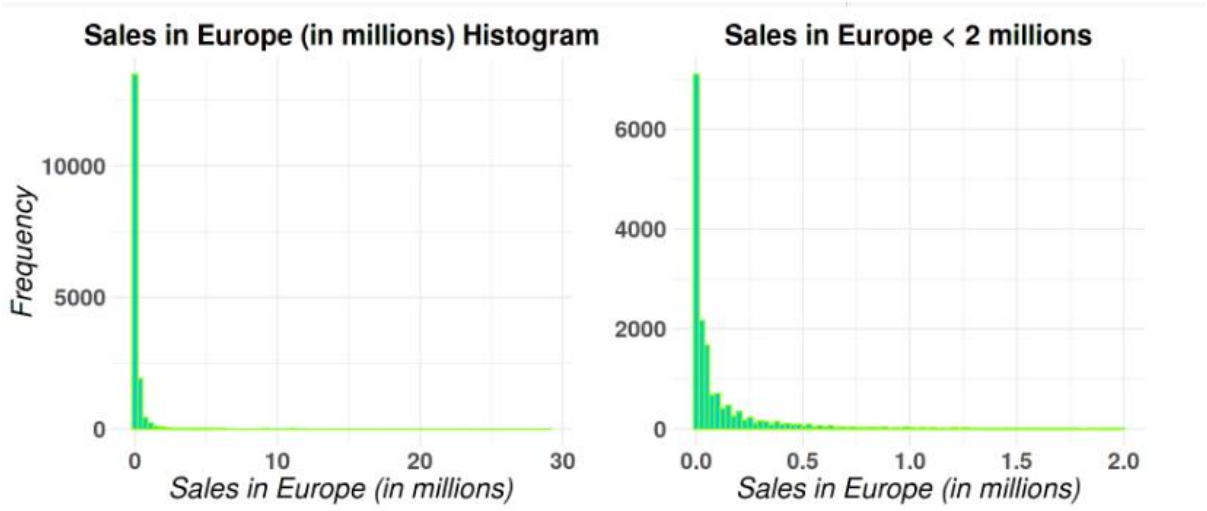
Language : Python 3

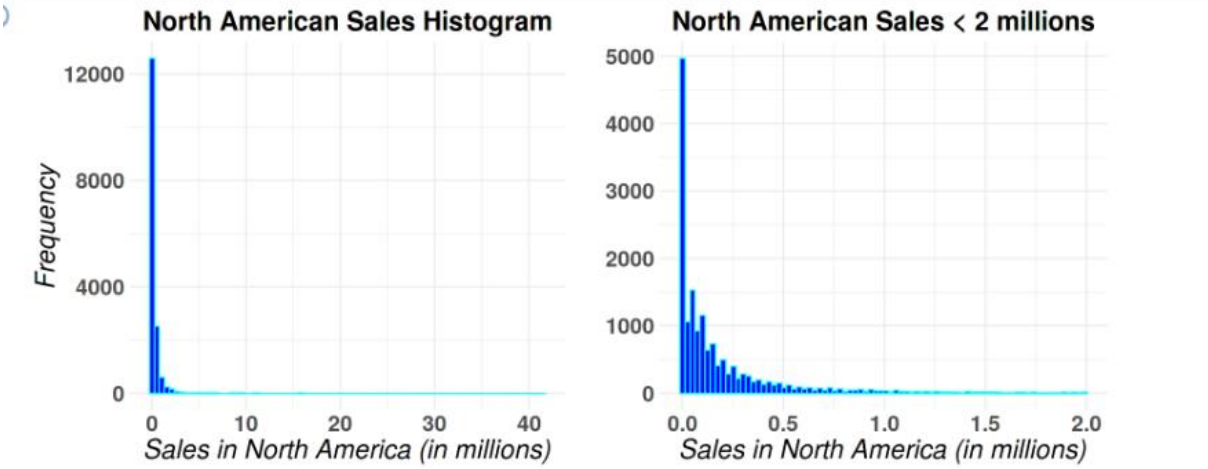
Compiler : Google Colab(basic)

RESULTS

Sales as per region




Sales in Europe (in millions) Histogram
Frequency
10000
5000
0
0
10
20
30
Sales in Europe (in millions)
Sales in Europe < 2 millions
6000
4000
2000
0
0.0
0.5
1.0
1.5
2.0
Sales in Europe (in millions)


North American Sales Histogram
Frequency
12000
8000
4000
0
0
10
20
30
40
Sales in North America (in millions)
North American Sales < 2 millions
5000
4000
3000
2000
1000
0
0.0
0.5
1.0
1.5
2.0
Sales in North America (in millions)