IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Video Games Sales Analysis: A Data Science Approach

[1]TM Geethanjali, [2]Ranjan D, [3]Swaraj HY, [4]Thejaskumar MV, [5]Chandana HP

[1]Assistant Professor, [2]Student, [3]Student, [4]Student, [5]Student
[1]Department of Information Science and Engineering,
[1] PES College of Engineering, Mandya, India

***Abstract**:* This paper aims to predict the top-selling video game sales in North America between 1983 and 2016. The dataset is collected from an internet platform known as Kaggle.com. The dataset was generated by vgchartz.com. Exploitation the dataset, the RStudio IDE tool and R-programming language are used for data cleaning, analysis, and representation. The machine learning algorithm used in this project is linear regression. Based on the Video Games Sales knowledge, it would be fascinating to know what area unit the required factors that make a game further successfully sold-out than others in North America. So, we'd would like to research what quite video games that area unit further successfully sold-out in North America. We have a tendency to tend to jointly would like to point the results of this Analysis in Associate in nursing intuitive methodology by visualizing outcome victimization ggplot2 in R. In this project, we have a tendency to tend to require NA_Sales (North America sales) as response variable and specialize in operative predictions by analyzing the rest of variables inside the k video games sales data. The results can facilitate film companies to know the key of generating an advertisement success game.

Index Terms **- North America, Kaggle, VGChartz, RStudio, R-programming, linear regression, gggplot2, NA_Sales.**

I. INTRODUCTION

In this modern world, video games are contended by all age teams of indivisual. These video games are contended on a video screen (on television, computer). There are many video games based on platform and genre like WII, NES, GB, etc., and puzzle, racing, shooting, etc., respectively. Video games are released by a publisher like Activision, Nintendo, etc., across different platforms. Video games have become a major platform of entertainment especially for kids and is used for business purpose as well. The video games such as NFS, Call of Duty, Battlefield and many more have produced exceptional income for about two to three decades and still earning.

Predictive modeling has helped a loads of people and organizations. This data science has several techniques, simulations, and machine learning. The main purpose of this study to seek out the issue that ends up in high video selling games Sales in North America. The dataset has 16,598 of data, and 11 attributes. This dataset is created based on the data from www.vgchartz.com/gamedb/. Then RStudio is used to predict which factor leads to top video game sales in North America.

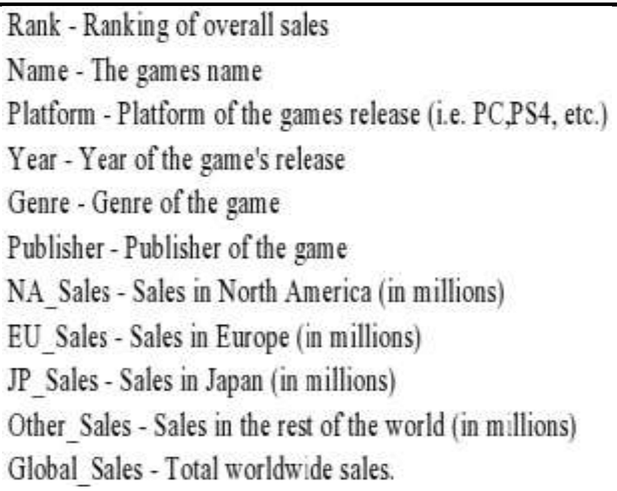
II. METHODOLOGY

A .Machine Learning Algorithm

The ML algorithm is a logic that grasp one step ahead when exposed to more information/data. When ML is exposed to training data it produces model. To build a Model, the Machine Learning Algorithm used here Linear Regression (Supervised Learning). It predicts the output values based on the input data fed. This algorithm builds a model based on the training data produced and predicts the new data.

B .Dataset

The RStudio is used to import the dataset. Dataset can be in excel or in CSV format. The dataset is reviewed and normalized. Normalization is changing the value of numeric columns of the dataset to common values and fit into a specific range. The attributes of the dataset are shown in the figure.



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)
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.

Figure: Attributes of Dataset

III. EXPERIMENT

This project/study uses a special video game sale dataset sold in different countries. This dataset is created by VGChartz.com.The RStudio is used to run an experiment.

The RStudio is free, open-source and an Integrated Development Environment (IDE) for R-language, used for statistical computing, programming, GUI, and Graphics. RStudio is most famous for Graphical capabilities, but in recent times it gained importance for analyzing data.

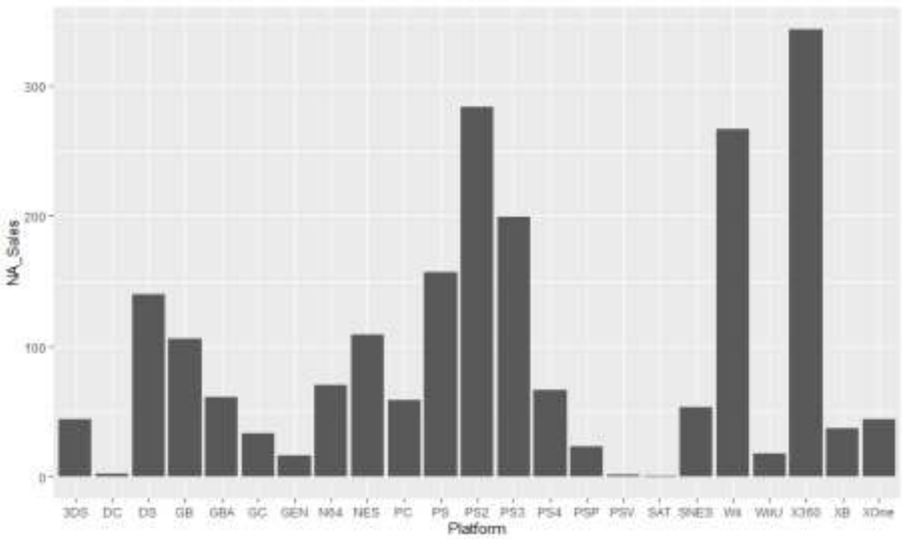
A .Data cleaning

At this stage, by using RStudio we import dataset and remove redundant, missing, duplicate, and unnecessary data for further processing. This stage is the most time-consuming stage in Data Science because to prevent wrongful prediction and get rid of the inconsistencies of data.

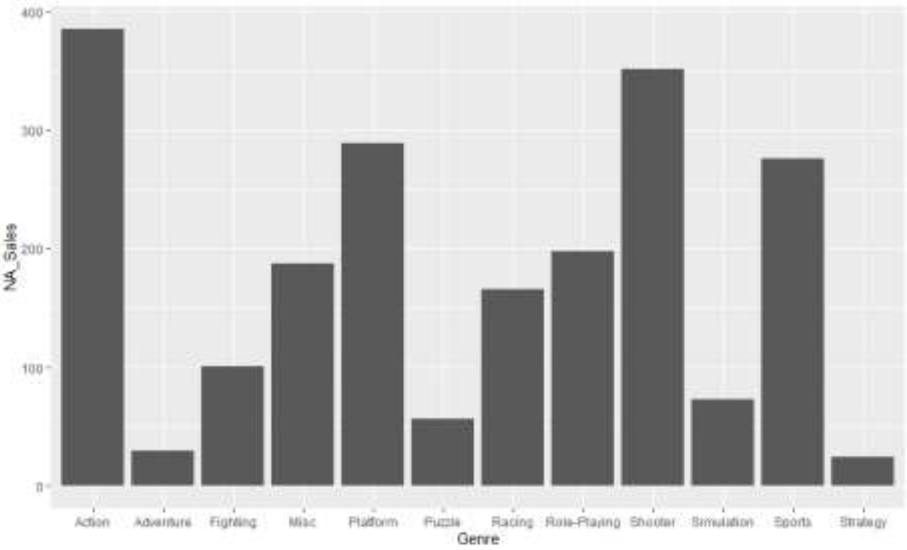
B .Data Exploration and Analysis

In this stage, we detect patterns, trends, and behavior in the data or dataset. This process makes further analysis easier because it excludes irrelevant data point and searches for no results data. It uses visualization which makes it easy to analyze. From our analysis, we concluded Platform attribute has mainly affected the Video Game Sales in North America.

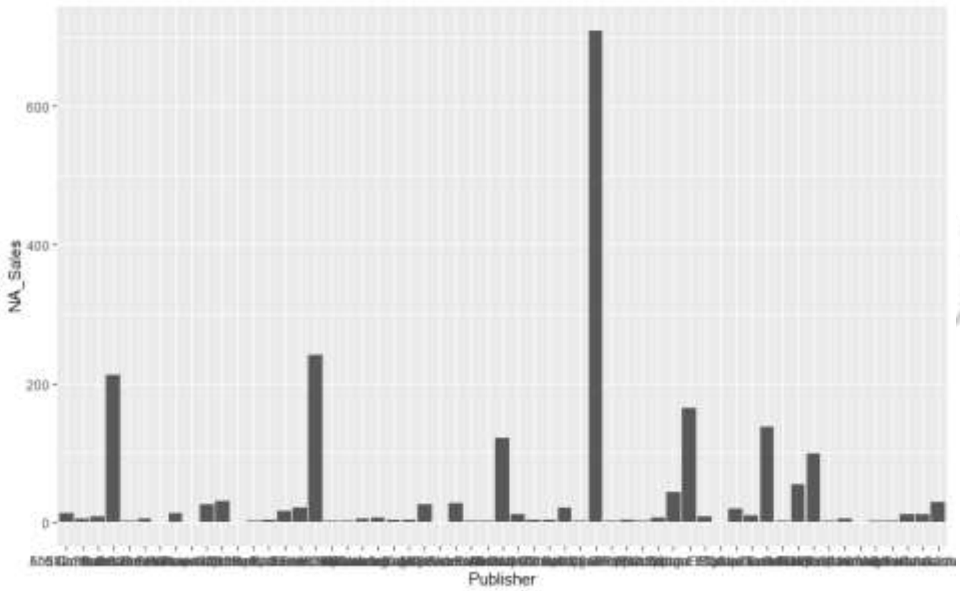
1. Platform vs NA_Sales (North America)



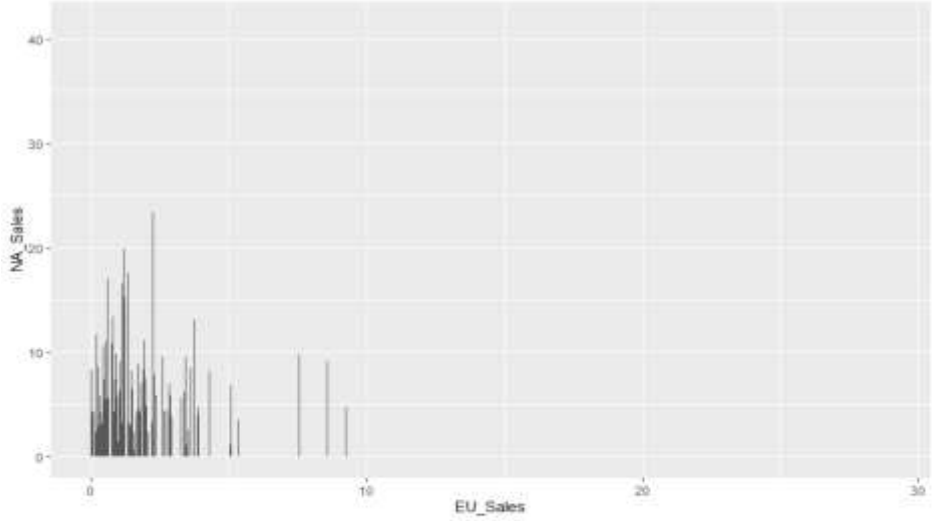
2. Genre vs NA_Sales (North America)



3. Publisher vs NA_Sales (North America)

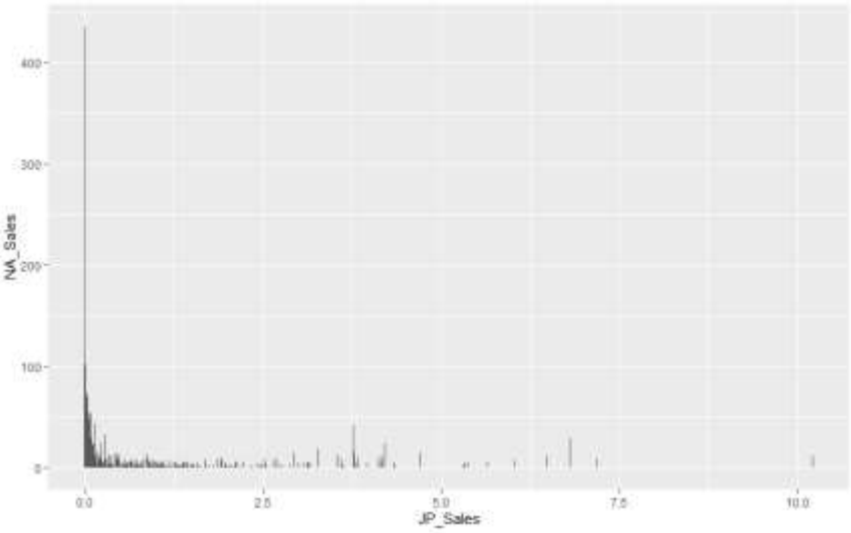


4. EU_Sales vs NA_Sales (North America)

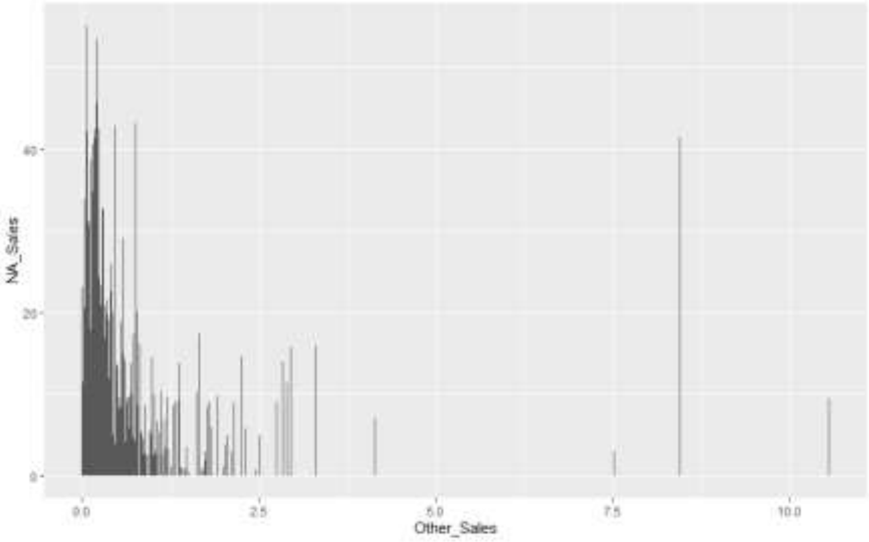


.

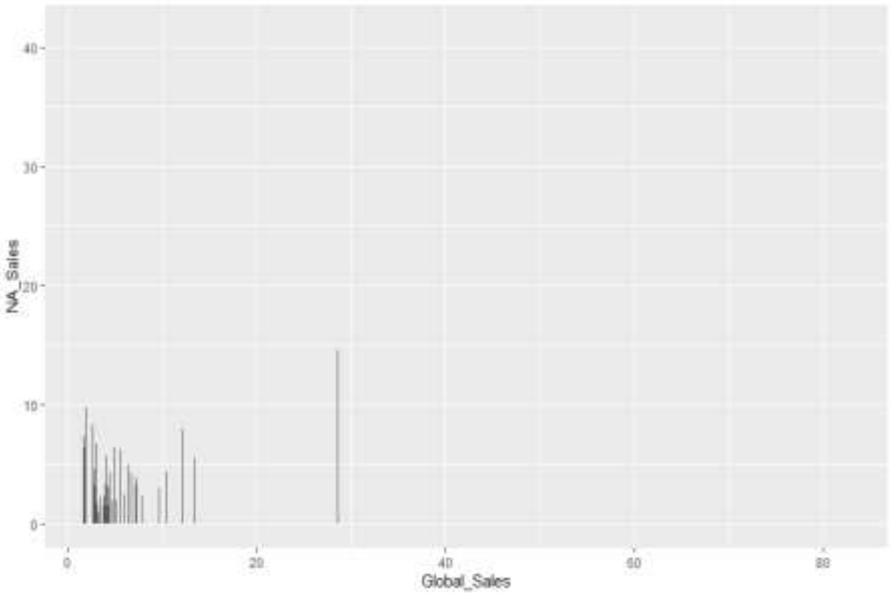
5. JP_Sales vs NA_Sales (North America)



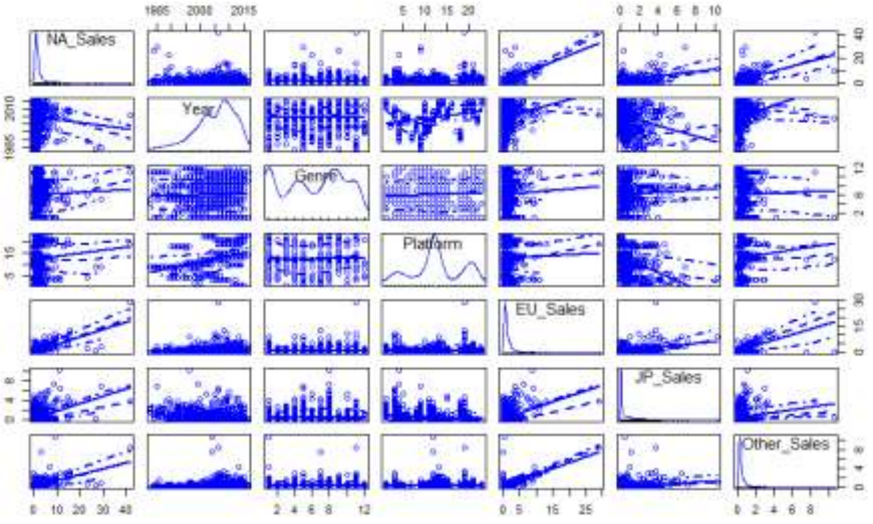
6. Other_Sales vs NA_Sales (North America)



7. Global_Sales vs NA_Sales (North America)



8. Scatterplot



So far we see that Platform of the videogame highly affects the sales of videogames in North America**.**

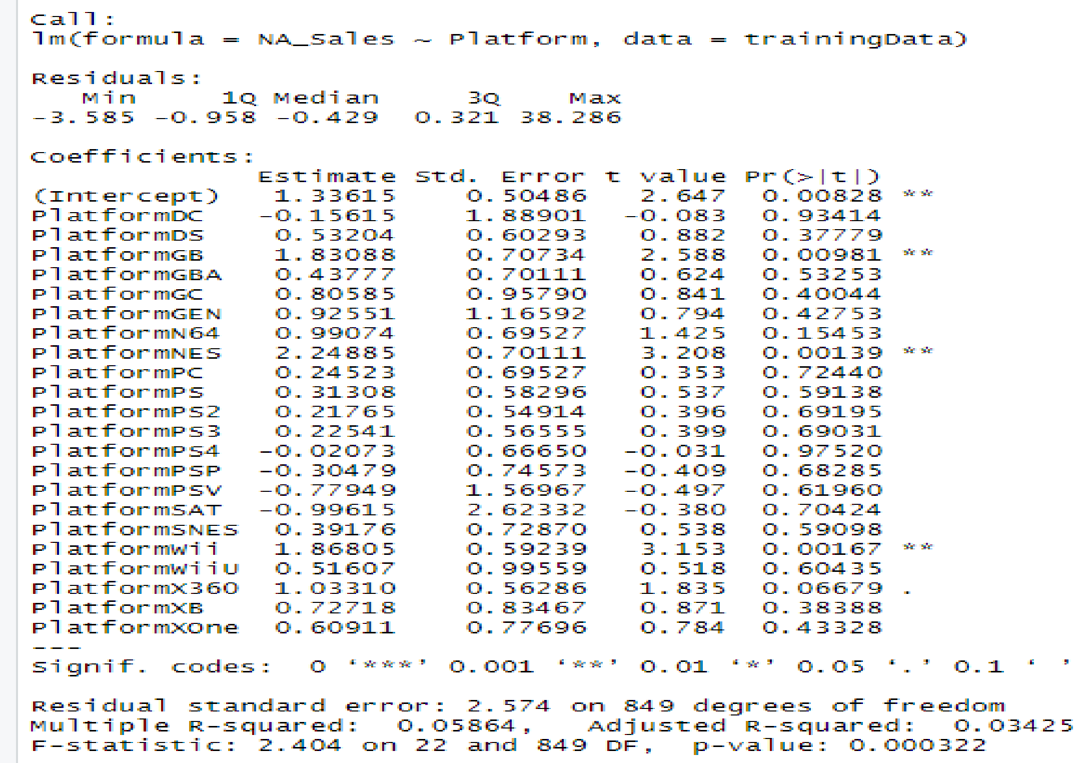
IV. Data Modeling

From exploratory Analysis, platform attributes highly affected the sales in North America. To build a model, we use a linear regression model. Following steps show the building of the model;

Step 1: Creating Training and Testing data.

Step 2: Building the model on training data.

Step 3: Review the model



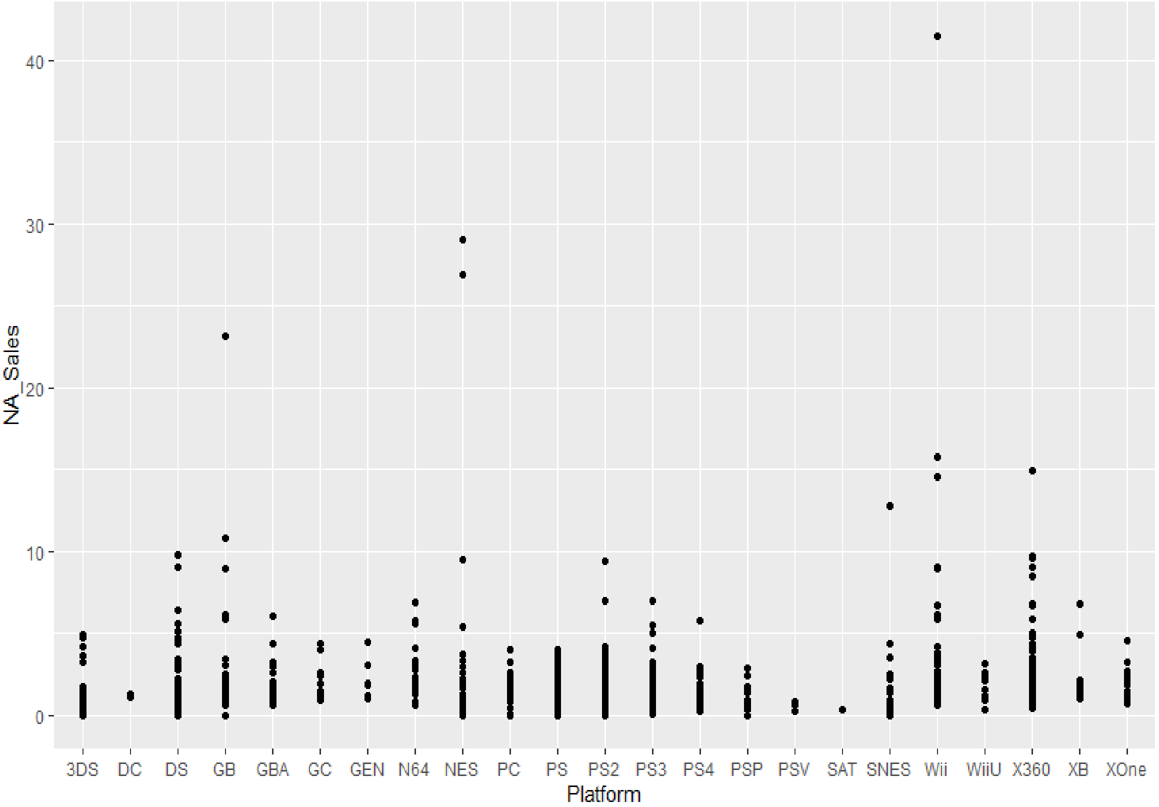
```
Call:
lm(formula = NA_Sales ~ Platform, data = trainingData)

Residuals:
   Min      1Q Median      3Q     Max
-3.585 -0.958 -0.429   0.321 38.286

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      1.33615    0.50486   2.647  0.00828 **
PlatformDC      -0.15615    1.88901  -0.083  0.93414
PlatformDS       0.53204    0.60293   0.882  0.37779
PlatformGB       1.83088    0.70734   2.588  0.00981 **
PlatformGBA      0.43777    0.70111   0.624  0.53253
PlatformGC       0.80585    0.95790   0.841  0.40044
PlatformGEN      0.92551    1.16592   0.794  0.42753
PlatformN64      0.99074    0.69527   1.425  0.15453
PlatformNES      2.24885    0.70111   3.208  0.00139 **
PlatformPC       0.24523    0.69527   0.353  0.72440
PlatformPS       0.31308    0.58296   0.537  0.59138
PlatformPS2      0.21765    0.54914   0.396  0.69195
PlatformPS3      0.22541    0.56555   0.399  0.69031
PlatformPS4     -0.02073    0.66650  -0.031  0.97520
PlatformPSP     -0.30479    0.74573  -0.409  0.68285
PlatformPSV     -0.77949    1.56967  -0.497  0.61960
PlatformSAT     -0.99615    2.62332  -0.380  0.70424
PlatformSNES     0.39176    0.72870   0.538  0.59098
PlatformWii      1.86805    0.59239   3.153  0.00167 **
PlatformWiiU     0.51607    0.99559   0.518  0.60435
PlatformX360     1.03310    0.56286   1.835  0.06679 .
PlatformXB       0.72718    0.83467   0.871  0.38388
PlatformXOne     0.60911    0.77696   0.784  0.43328
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.574 on 849 degrees of freedom
Multiple R-squared:  0.05864,   Adjusted R-squared:  0.03425
F-statistic: 2.404 on 22 and 849 DF,  p-value: 0.000322
```

Step 4: Plotting Training data model using Library ggplot



V. RESULTS AND DISCUSSION

From information modeling, the p-value (0.00032) is less than the significance value (typically ≤ 0.05). So we have a statistically significant model. Also, R-Sq. and Adj R-Sq. squares measures comparative to the initial model engineered on full information. Therefore we can conclude prime mercantilism **Video Games** in **North America** based on the platform are

- **GB**
- **NES**
- **Wii**

VI. ACKNOWLEDGMENT

I want to offer my true thanks to my counselor/tutor Assistant Prof. TMGeethanjali for huge help for this Data Science Project. Her direction helped me a great deal in information examination and composing this paper.

.

REFERENCES

[1] Amar Aziz , Shuhaida Ismail, Muhammad Fakri Othman, Aida Mustapha," Empirical Analysis on Sales of Video Game: Data mining Approach "Amar Aziz et al 2018 J. Phys.: Conf. Ser. 1049 012086

[2] "Fraud Detection of Credit Card Payment System by Genetic Algorithm" K.RamaKalyani, D.UmaDevi, International Journal of Scientific & Engineering Research Volume 3, Issue 7, July-2012 1 ISSN 2229-5518

[3] "Machine Learning Applied to Cervical Cancer Data" Dhwaani Parikh a, Vineet Menon. I.J. Mathematical Sciences and Computing, 2019, 1, 53-64 Published Online January 2019 in MECS (http://www.mecs-press.net) DOI: 10.5815/ijmsc.2019.01.05

[4] www.kaggle.com/ignacioch/predicting-vg-hits-1-million-sales-with-lr-rfc Predicting VG hits (>1 million sales) with LR, RFC¶ By Ignacio Chavarria (@ignacio_chr)

[5] *Video Game Sale Analysis,vh42720, November 29, 2017*.

[6] Chapter-4 "Knowledge from the data and Data Exploration Analysis", Ashish Patel, Jul 3, 2018