Mini-Project Final Report

DATA MINING AND PREDICTIVE ANALYSIS Lab (ICT 3171)

Project Title:

Video games sales analysis with ratings

Team Number:

9

Team Members:

Sl_No	Full Name	Reg_No	Roll_No
1	Adithya Rao Kalathur	200953015	04
2	Sushant Raj	200953062	21
3	Sathvik Chanda	200953078	25

Introduction:

Since the 1970s, when the video game industry first began to take off, it has experienced rapid expansion to the point where many individuals of all ages now play video games on a regular basis. With sales in the United States reaching $16.5 billion in 2015, it is a very lucrative market. In contrast, the U.S. film business generated $29.2 billion in sales in 2015. Since Valve established its Steam Store, the number of releases and digital sales of PC games have increased1. Thanks to a mechanism called Greenlight, which made it reasonably simple for developers to release their games on Steam without a publisher, which had been quite challenging up until then, this store experienced significant growth after 2012. The number of releases on the store rapidly increased as a result of greenlight over time, reaching over 10,000 by the end of 2016. As a result, it become more and harder for game designers to stand apart and even sell enough copies to pay for their game's creation. If the popularity of games could be predicted ahead, game developers could alter their creation to appeal to a wider audience or even abandon their endeavour to create a game if

it was unsuccessful. Therefore, evaluating a concept rather than a full game might be useful. Previous attempts to forecast the success of video games made projections based on the assumption that the game has already been launched. Additionally, there was no motivation to research the PC gaming market, hence they primarily dealt with console titles from before 2010. In order to evaluate a game's success based on descriptive information like genre, price, developer, or game features, this thesis will examine known characteristics that determine the success of video games, build a database of PC games because there isn't one already, and more.

Literature Survey:

1.Research Paper:

https://www.researchgate.net/publication/326510277_Empirical_Analysis_on_Sales_of_Video_Games_A_Data_Mining_Approach

Summary:

This essay investigates the elements that result in blockbuster video game sales. The Rapid Miner tool is used with the dataset to choose the features or components and create effective data estimation. In this study, the k-Nearest Neighbor (k-NN), Random Forest, and Decision Tree approaches were applied. There are various methods that have been used to compare the results and determine what makes a blockbuster video game. The methods are Random Forest, K-NN Result, and Decision Tree. The benefits of this study include the fact that it offers the most recent knowledge and aids in issue resolution. The disadvantage is that it doesn't provide data on sales for the current year.

2.Research Paper:

https://swer.wtamu.edu/sites/default/files/Data/303-1102-1-PB_0.pdf

Summary:

This study looks at North American video game sales by platform during the years of 2006 and 2011. The Kruskal-Wallis test is used in this study to compare eight distinct gaming platforms. According to the findings, Nintendo's Wii and DS are the best-selling game platforms, followed by the Xbox 360 in a distant second place, a number of Sony PlayStation platforms in a distant third place, computer games in a distant fourth place, and the sixth generation Nintendo GameCube in a distant last place in terms of sales. The benefit of this research is that it promotes hypothesis testing and issue solving. Its failure to provide information on sales relating to the current year is a negative.

3.Research Paper:

https://digitalcommons.iwu.edu/cgi/viewcontent.cgi?article=1107&context=econ_honproj

Summary:

The factors affecting the sales of video game software are examined in this essay. The literature that is currently available points to a variety of variables, including the hardware a game is released on and the game's genre. Several of these factors are included in our paper, but we also add a new one: quality. Although one would naturally anticipate that a game's quality would have a significant beneficial impact on its eventual sales, this topic has not been covered in the literature to date. The model takes into account a game's quality by taking into account the typical review score it obtains from expert critics. The findings suggest that quality does, in fact, play a significant effect in consumer choice. This research employed the emperical technique as its algorithm.The advantage of this research is it is helpful in challenging assumptions and critical thinking. The drawback is limited access of data and time constrain.

4.Research Paper:

https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.471.520&rep=rep1&type=pdf

Summary:

This study uses data from the video game industry to evaluate how product and consumer factors limit the impact of online consumer evaluations on product sales. The results show that less well-known games and games whose players have more Internet expertise are more influenced by online reviews. The article illustrates varying influence of consumer reviews among products in the same product category, and indicates that firms' online marketing techniques should be contingent on product and consumer attributes. In light of the rise in the share of niche items in recent years, they also analyse the ramifications of these findings. This research has the benefit of using both an exploratory study and a plot.The drawback of this research is some of the data was missing and the latest data was not available.

5.Research Paper:

https://www.questjournals.org/

Summary:

Exploratory data analysis will be done in this report to provide more information about video game sales. Data input, data inspection, data cleaning and coercion, data summary, data transformation & visualisation, and data explanation are all included in the process. This study has the benefit of offering the most recent knowledge and aiding in problem solutions. This study's flaw is that it doesn't provide data on sales for the current year.

6.Research Paper:

https://ijcrt.org/papers/IJCRT2005182.pdf

Summary:

The logic of the ML algorithm advances when it is exposed to new information or data. Models are created when ML is exposed to training data. The machine learning algorithm here used linear regression to create a model (Supervised Learning). Based on the input data that is provided, it forecasts the output values. Based on the training data generated, this algorithm creates a model and forecasts the fresh data. The benefit of this research is that it promotes critical thinking and assumptions-challenging. This study's flaw is that it doesn't provide data on sales for the current year.

7.Research Paper:

https://rpubs.com/atika_faradilla/858694

Summary:

The dataset used in this study, Video Game Sales with Ratings, was produced in 2016 and used information from the review website Metacritic. It was obtained via the data website Kaggle. Then, in order to pinpoint the crucial elements that go into the final model, they conducted a stepwise regression analysis using the statistical programme SPSS. High ratings are what cause games to appear higher on web sites or in better store locations. Over the years, a lot of games come and go, but the ones that are popular with users Critics are people who keep returning for new, "better" versions and ongoing sales. The benefit of this

study is that it fosters critical thinking and may be applied to a plot or exploratory investigation. The drawback of this is that some of the data were missing. If they removed the video games without a rating score as it would be hard to properly compare all the data with these Data missing. They still had over 7,000 data points. There were a few variables not included that we thought might Have been important to consider such as gender, as many males and females play different types of video games.

8. Research Paper

https://www.diva-portal.org/smash/get/diva2:1680040/FULLTEXT01.pdf

Summary:

Instead of repeating findings from earlier studies, this study's goal is to use its findings as a springboard for additional investigation. As it produced reliable projections in earlier research, Random Forest was utilised as the baseline model. Bootstrap Aggregating is a machine learning ensemble meta-algorithm, often known as bagging. It works by breaking up the data into smaller subsets to train each tree with. It is used to improve the performance of tree-based algorithms. The algorithm was successfully employed in a prior study by Schaer et al. (2022) to anticipate market potential using Google Trends, which gives this research an edge. The full value of using machine learning to anticipate market potential cannot be determined until this is done. The problem is that time has been the project's main constraint and has influenced decisions at every stage. Time constraints also prevented more thorough research into the preparation of video game entries utilised in the datasets. Self-admittedly, the evaluation of model performance is really basic.

9. Research Paper:

https://www.theseus.fi/bitstream/handle/10024/497979/e1700994ThesisRevised.pdf?sequenc=2&isAllowed=y

Summary :

This thesis' main objective was to investigate the market for the video game industry and identify the factors that contributed to its enormous success. The major objective was to examine the techniques, strategies, and trends used in the creation and design of video game projects that have fostered rapid expansion and achievement. The k-Nearest Neighbor algorithm and the Random Forest decision tree were both used in this project. The findings provide an overview

of the whole video game industry, including strategies and tactics used by businesses, how trends develop, and its general outlook for the future. This study has the benefit of offering the most recent knowledge and aiding in problem solutions.The drawback is some of the assertions made throughout this thesis may be proved false due to the limitations set upon this research and the video game industry being in constant growth.

10. Research Paper

https://www.slideshare.net/AksshitVijayvergia/video-game-sales-analysis-84181197

Summary:

A list of video games with sales of more than 100,000 copies can be found in this dataset. We must first establish a connection between R and SQL Server before performing EDA on R. By configuring an ODBC data source in Administrative Tools, we may accomplish this. To access the tables in the SQL Server database, R will use the RODBC package. Now that the SQL Server table has been cleansed, it can be imported into Renvironment. The sqldf package in R will then be used for analysis. After conducting the EDA, we came to the conclusion that some games were released on more platforms than others, which led to better sales of that particular game.The advantage and drawbacks of this research is that the null hypothesis can be rejected. Hence we can say console gaming is not dying. We might even say, it is doing better than PC gaming.

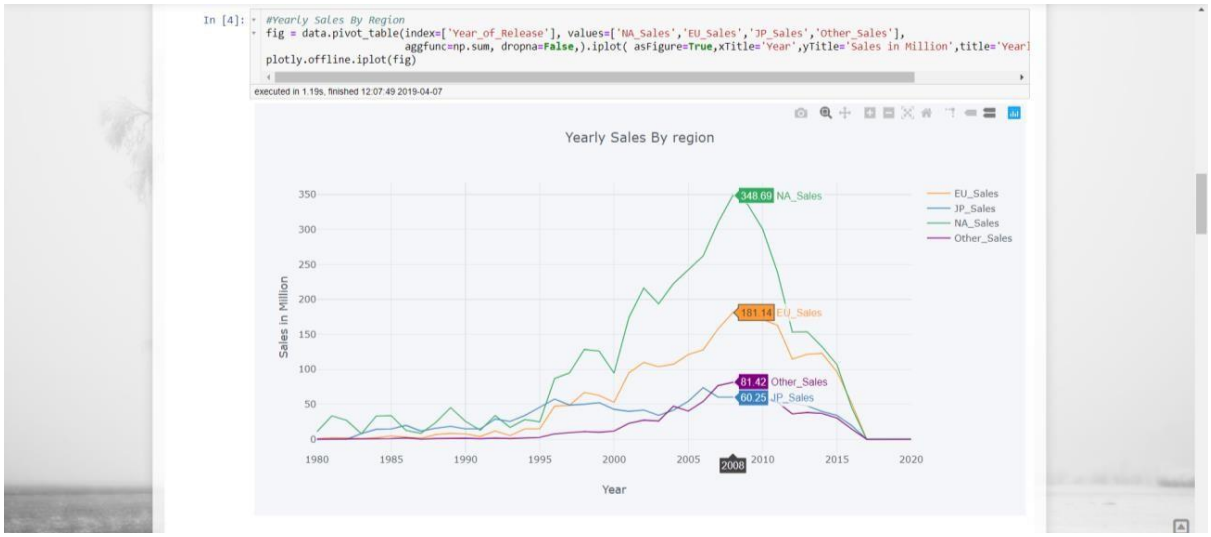
Methodology:

Data Preprocessing and Cleaning:

The data is first downloaded from Kaggle (https://www.kaggle.com/rush4ratio/video-gamesales-with-ratings) and then load into our IDE. We then explore the data and clean it respectively to benefit us for training and Exploration of data. From looking at the data we can point out data which are categorical or nominal or ratio respectively.

Data Exploration:

The data is then explored using data exploration libraries like Seaborn and Plotly. Various different kinds of plots are drawn to different inferences from the data to understand how they are linked with each other. An example is shown below, which compares sales of different regions across the globe.



Initial Overview of Data Set:

```
data.describe(include="all")

#The top value in User_Score column is "tbd".
#There are high outliers in sales columns (NA, EU, JP, Other, Global) and User_Count column.
```

	Name	Platform	Year	Genre	Publisher	NA	EU	JP	Other	Global	Critic_Score	Critic_Count
count	16448	16448	16448.000000	16448	16416	16448.000000	16448.000000	16448.000000	16448.000000	16448.00000	7983.000000	7983.000000
unique	11429	31	NaN	12	579	NaN	NaN	NaN	NaN	NaN	NaN	NaN
top	Need for Speed: Most Wanted	PS2	NaN	Action	Electronic Arts	NaN	NaN	NaN	NaN	NaN	NaN	NaN
freq	12	2127	NaN	3308	1344	NaN	NaN	NaN	NaN	NaN	NaN	NaN
mean	NaN	NaN	2006.488996	NaN	NaN	0.263965	0.145895	0.078472	0.047583	0.53617	68.994363	26.441313
std	NaN	NaN	5.877470	NaN	NaN	0.818286	0.506660	0.311064	0.187984	1.55846	13.920060	19.008136
min	NaN	NaN	1980.000000	NaN	NaN	0.000000	0.000000	0.000000	0.000000	0.01000	13.000000	3.000000
25%	NaN	NaN	2003.000000	NaN	NaN	0.000000	0.000000	0.000000	0.000000	0.06000	60.000000	12.000000
50%	NaN	NaN	2007.000000	NaN	NaN	0.080000	0.020000	0.000000	0.010000	0.17000	71.000000	22.000000
75%	NaN	NaN	2010.000000	NaN	NaN	0.240000	0.110000	0.040000	0.030000	0.47000	79.000000	36.000000
max	NaN	NaN	2020.000000	NaN	NaN	41.360000	28.960000	10.220000	10.570000	82.53000	98.000000	113.000000

Overview of the Data after Cleaning :

```
data = data.rename(columns={"Year_of_Release": "Year",
                            "NA_Sales": "NA",
                            "EU_Sales": "EU",
                            "JP_Sales": "JP",
                            "Other_Sales": "Other",
                            "Global_Sales": "Global"})
data = data[data["Year"].notnull()]
data = data[data["Genre"].notnull()]
data["Year"] = data["Year"].apply(int)
data["Age"] = 2018 - data["Year"]
data["User_Score"] = data["User_Score"].replace("tbd", np.nan).astype(float)
data.describe(include="all")
```

	Name	Platform	Year	Genre	Publisher	NA	EU	JP	Other	Global	Critic_Score	Critic_Count	U
count	16448	16448	16448.000000	16448	16416	16448.000000	16448.000000	16448.000000	16448.000000	16448.00000	7983.000000	7983.000000	7
unique	11429	31	NaN	12	579	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
top	Need for Speed: Most Wanted	PS2	NaN	Action	Electronic Arts	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
freq	12	2127	NaN	3308	1344	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
mean	NaN	NaN	2006.488996	NaN	NaN	0.263965	0.145895	0.078472	0.047583	0.53617	68.994363	26.441313	
std	NaN	NaN	5.877470	NaN	NaN	0.818286	0.506660	0.311064	0.187984	1.55846	13.920060	19.008136	
min	NaN	NaN	1980.000000	NaN	NaN	0.000000	0.000000	0.000000	0.000000	0.01000	13.000000	3.000000	
25%	NaN	NaN	2003.000000	NaN	NaN	0.000000	0.000000	0.000000	0.000000	0.06000	60.000000	12.000000	
50%	NaN	NaN	2007.000000	NaN	NaN	0.080000	0.020000	0.000000	0.010000	0.17000	71.000000	22.000000	
75%	NaN	NaN	2010.000000	NaN	NaN	0.240000	0.110000	0.040000	0.030000	0.47000	79.000000	36.000000	
max	NaN	NaN	2020.000000	NaN	NaN	41.360000	28.960000	10.220000	10.570000	82.53000	98.000000	113.000000	

Prediction Model:

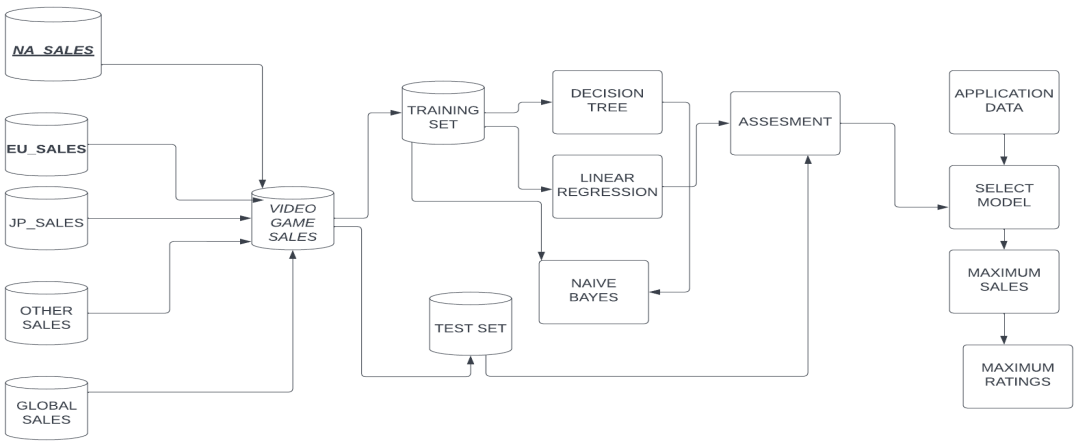
Before we start predicting model, we divide our dataset into training and testing data respectively. After splitting the data, we define two function which will be used to fit the models and calculate the respective errors obtained by different models. The machine learning models that are used on the dataset are: -

• Linear Regression

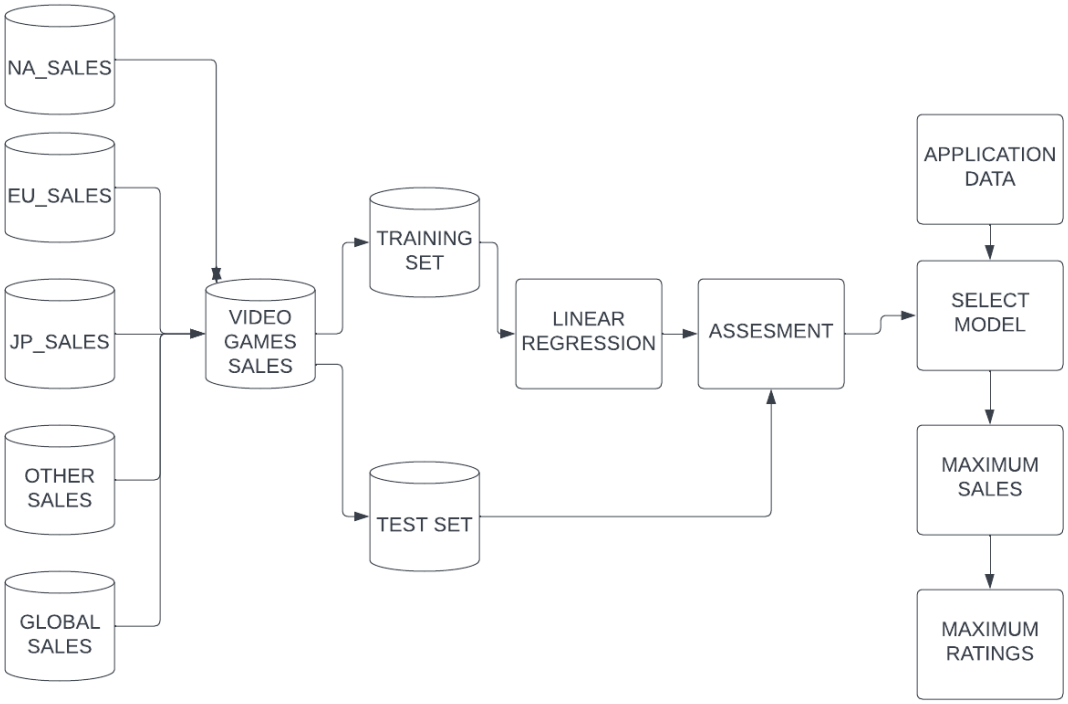
• Naïve Bayes

• K-Means

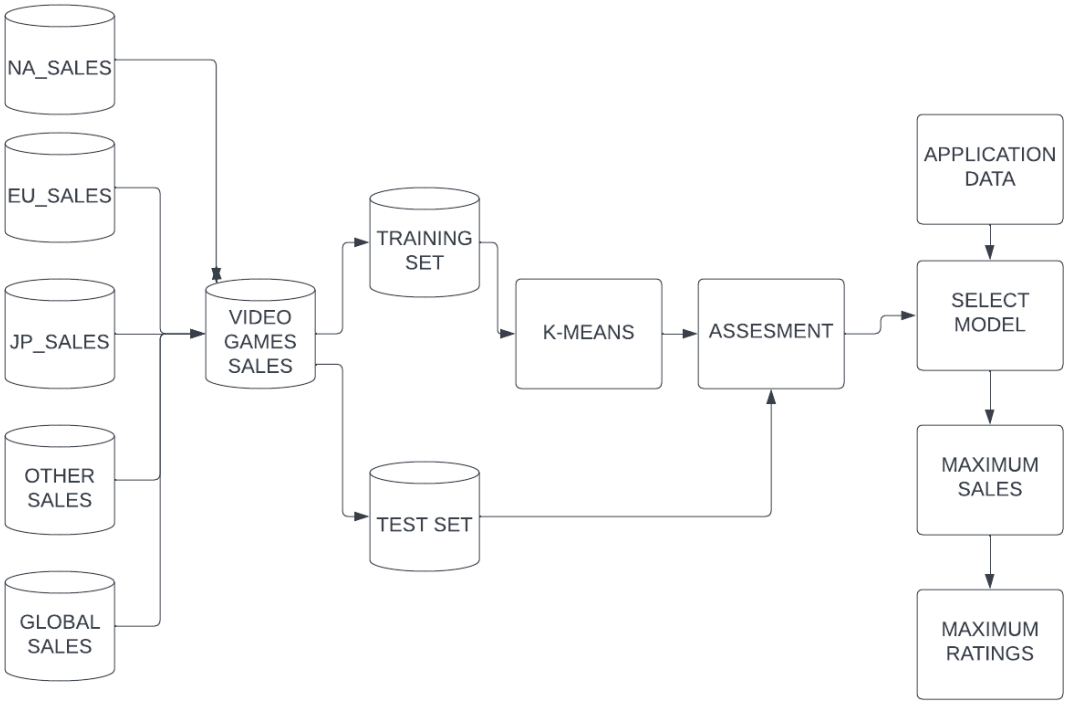
Block Diagram for all three algorithms:



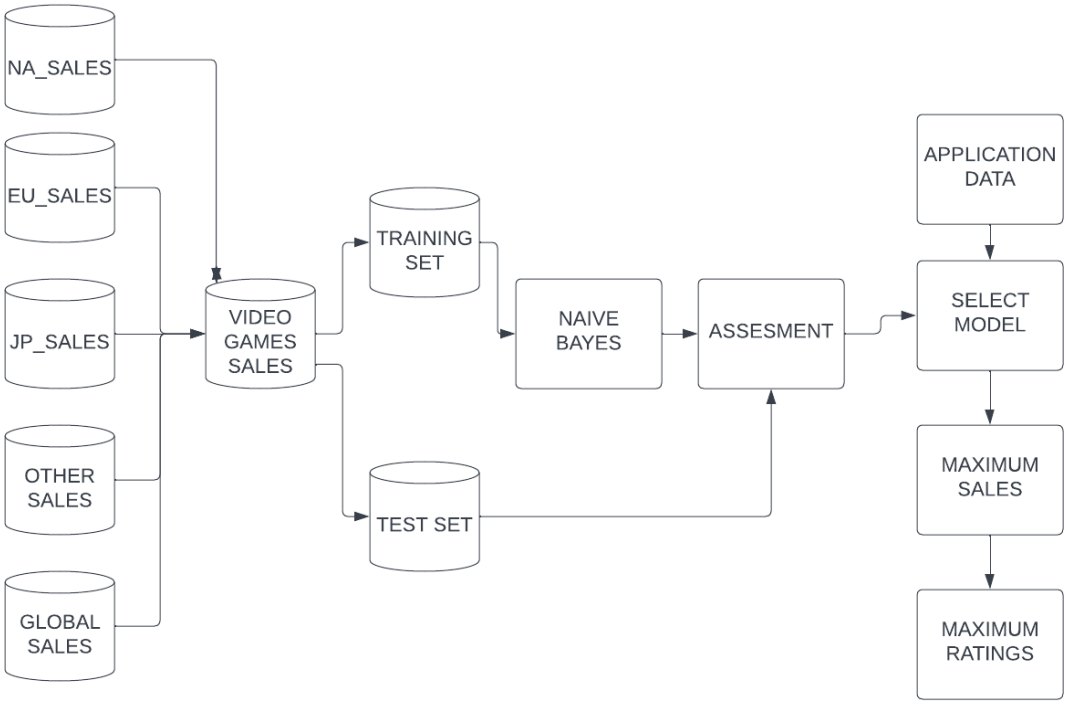
Block Diagram for Linear Regression Algorithm:



Block Diagram for K-Means Algorithm:



Block Diagram for Naïve Bayes Algorithm:



Models:

1.Linear Regression:

A machine learning algorithm based on supervised learning is linear regression. It executes a regression operation. Regression uses independent variables to model a goal prediction value. It is mostly used to determine how variables and forecasting relate to one another.It carries out the task of predicting the value of a dependent variable (y) based on a specified independent variable (x). Therefore, x (the input) and y (the output) are found to be linearly related by this regression technique (output).

2.K-Means Algorithm:

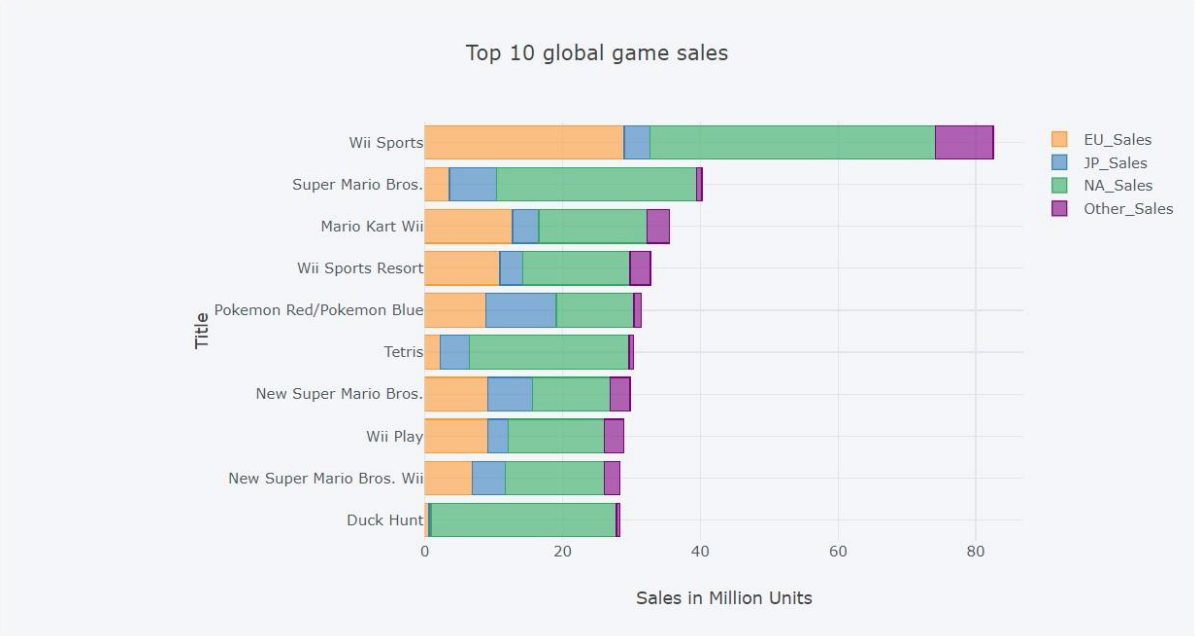
K-means is a centroid-based clustering algorithm where each data point is assigned to a cluster based on the distance between it and a centroid. Finding the K number of groups in the dataset is the objective. Each data point is assigned to a group iteratively, and over time, data points are clustered according to shared characteristics. In order to determine which group each data point should belong to, the goal is to reduce the sum of distances between the data points and the cluster centroid.

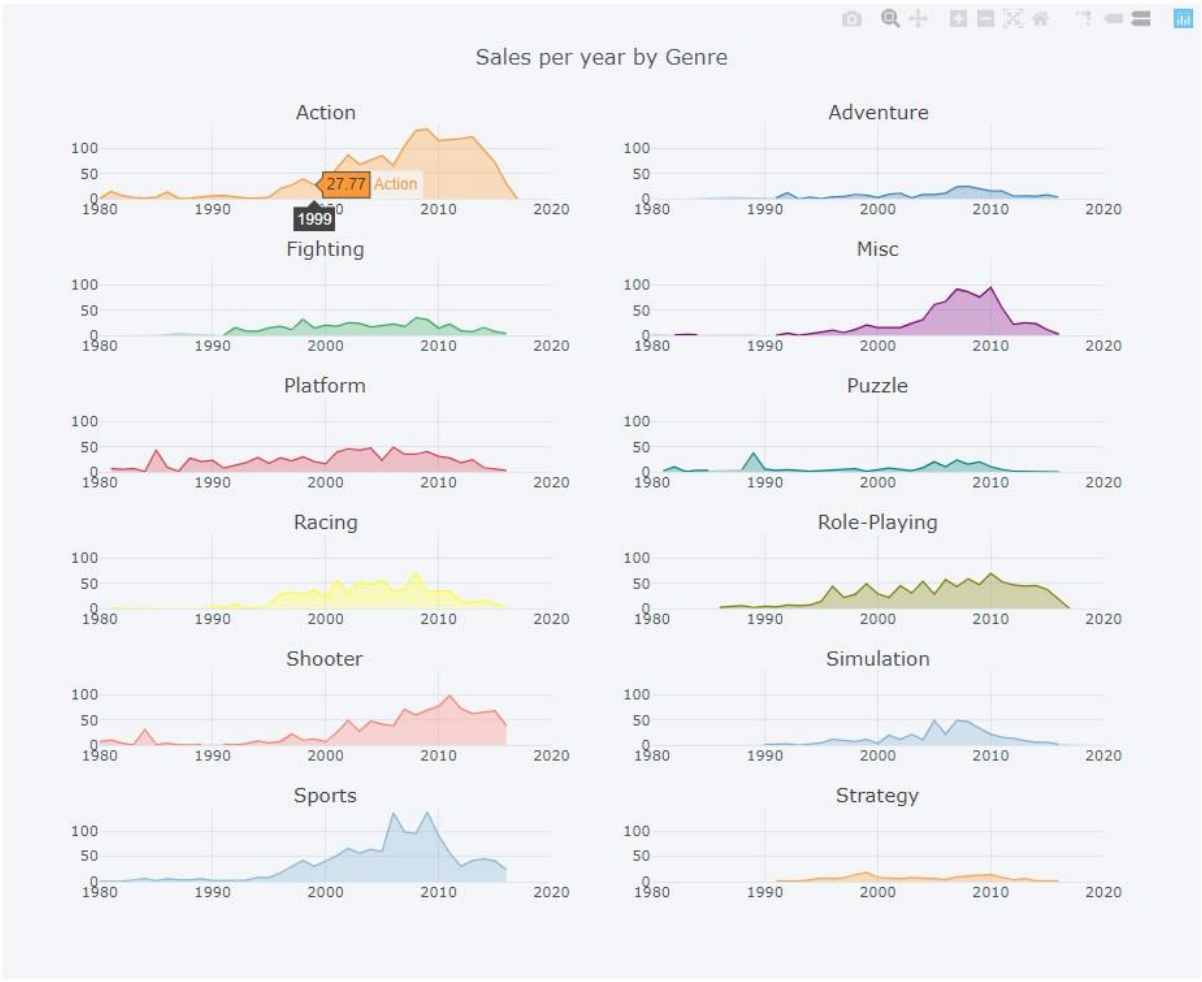
3.Naive Bayes:

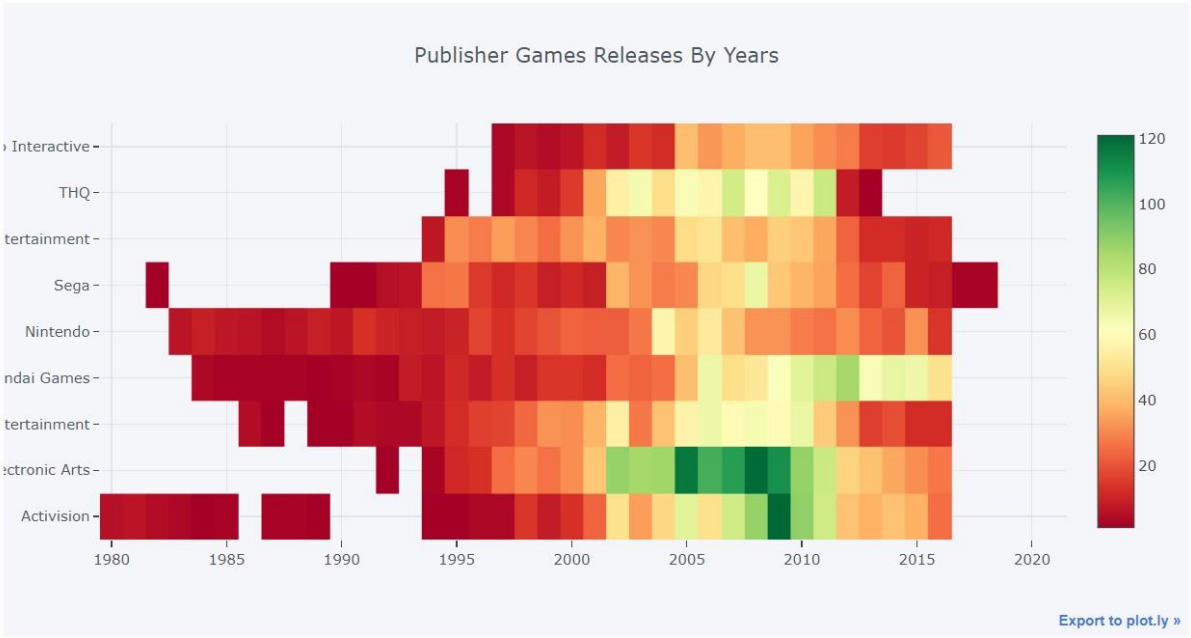
It is a classification method built on the Bayes Theorem and predicated on the idea of predictor independence. A Naive Bayes classifier, to put it simply, believes that the presence of one feature in a class has nothing to do with the presence of any other feature.

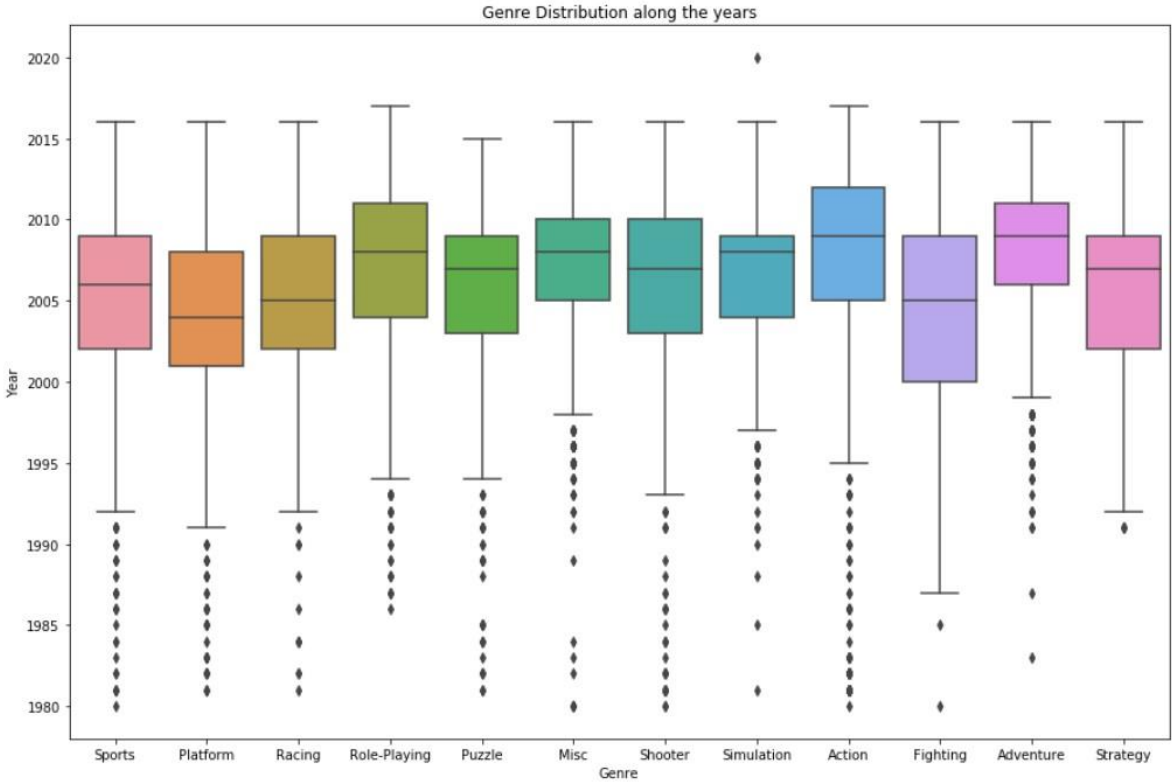
Data exploration and analysis :

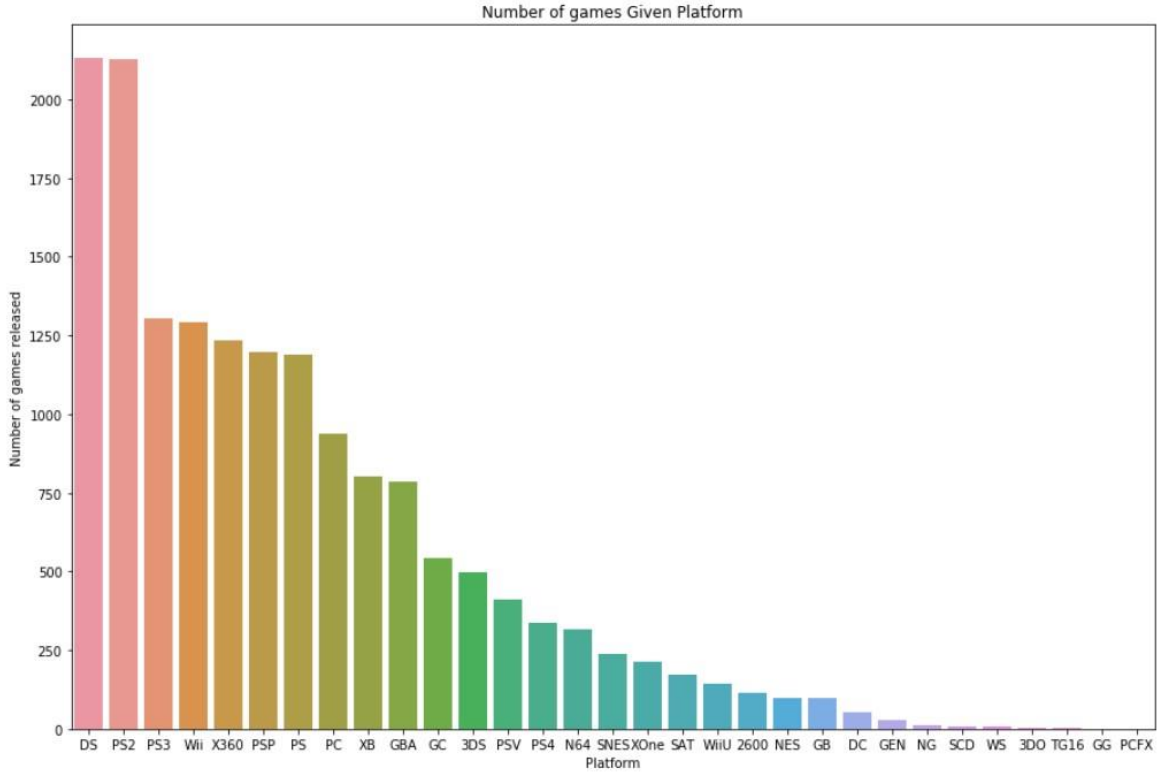
Below are few of the plots from EDA: -




Sales per year by Genre
Action
Adventure
Fighting
Misc
Platform
Puzzle
Racing
Role-Playing
Shooter
Simulation
Sports
Strategy
27.77 Action
1999
100
50
0
1980
1990
2000
2010
2020


Publisher Games Releases By Years
Interactive
THQ
tertainment
Sega
Nintendo
ndai Games
tertainment
ctronic Arts
Activision
1980
1985
1990
1995
2000
2005
2010
2015
2020
120
100
80
60
40
20
Export to plot.ly »


Genre Distribution along the years
Year
2020
2015
2010
2005
2000
1995
1990
1985
1980
Sports
Platform
Racing
Role-Playing
Puzzle
Misc
Shooter
Simulation
Action
Fighting
Adventure
Strategy
Genre


Number of games Given Platform
Number of games released
2000
1750
1500
1250
1000
750
500
250
0
DS PS2 PS3 Wii X360 PSP PS PC XB GBA GC 3DS PSV PS4 N64 SNES XOne SAT WiiU 2600 NES GB DC GEN NG SCD WS 3DO TG16 GG PCFX
Platform

Advantages and Disadvantages of each model:

The benefit of the linear regression technique is that it is straightforward to apply, offers a solution with low spatial complexity, allows for quick training, and assumes that a feature is significant based on its value. The downside is that it only applies if the answer is linear. Although it may not always be the case in real-world situations, the algorithm presupposes that the input features are mutually independent and that the input residuals (error) are normally distributed (no co-linearity).

The k-means method has the advantages of being quick, robust, simple to grasp, somewhat efficient, producing tighter clusters, versatile, simple to analyse, and having a lower computing cost. It improves precision, performs better with spherical clusters, but the number of cluster centres must be specified beforehand. If two sets of data are heavily overlapping, it is impossible to identify that there are two clusters. The outcomes are variable as a result of the various data representations, Euclidean distance might unfairly balance the variables, It provides the squared error function's local maxima. Occasionally, selecting the centroids at random is unsuccessful. Only if the meaning is clarified may it be used, can't deal with noisy data and outliers, Avoid working with the nonlinear data set. lacks coherence responsive to scale The computer could crash if it encounters particularly huge data sets. Issues with prediction.

Naive Bayes is suitable for solving multi-class prediction problems, if its assumption of the independence of features holds true, it can perform better than other models and requires much less training data, and it is better suited for categorical input variables than numerical variables. However, Naive Bayes has the drawback of assuming that all predictors (or features) are independent. This technique faces the "zero-frequency problem," where it gives zero probability to a categorical variable whose category in the test data set wasn't accessible in the training dataset. This limits the usability of this algorithm in real-world use situations. To solve this problem, it would be preferable if you employed a smoothing technique. However, you shouldn't take its probability outputs too seriously because sometimes its estimations are off.

Results:

1. With average sales of $264,667.430, North America is the top region.

2. The following are the platforms that are most popular:

• The most popular console in North America, earning $601.05 million overall, is the X360 (Microsoft).

• The PS3 (Sony) is the most popular system in Europe, grossing $343.71 million.

• In Japan, people like the Nintendo DS more than any other platform, spending a total of $175.57 million on it.

• The PS2 (Sony) is the most popular console overall among consumers in other regions, grossing $193.44 million.

3. Wii Sports, which has earned a total of $82.74 million worldwide, is the most popular game.

4. These are the top games:

• In North America, Europe, and other regions, Wii Sports has been the most popular game.

• In Japan, Pokemon Red/Blue is the most popular game.

5. Super Mario Bros., which was first released in 1985, is the oldest game currently selling $40.24 million worldwide. Pokemon Red/Blue (1996), Tetris (1989), and other games come after this.

6. Action games are the most popular worldwide, with $1751.18 million.

7. The top publishers are: • Nintendo, which earned a combined $816.87 million, $418.74 million, and $455.42 million in North America, Europe, and Japan, respectively.

• Electronic Arts is the top publisher in other areas, earning $129.77 million.

8. No region has outperformed the worldwide average in terms of sales. Global sales on average total $537,440.656 million.

This report lists the top games, platforms, publishers, genres, and other characteristics for various geographical areas. The outcomes will help to increase sales and customer satisfaction.

Conclusion and future work:

With the dataset, we were able to successfully conduct an exploratory data analysis and draw insightful conclusions. From there, we were able to further encode the dataset, train various machine learning algorithms, and develop

several models. The inaccuracy was subsequently reduced, and we obtained a final accuracy of 0.1757. Future work will focus on improving Model 1 and removing the overfitting condition since it was overfitting the data. To lower the error, Model 2 can be further improved. Other than the ones already mentioned, the introduction of new machine learning algorithms may be able to produce superior results. All of them might be viewed as future work.

References:

- Kaggle (https://www.kaggle.com/datasets)
- Wikipedia (https://en.wikipedia.org/wiki/Video_game)
- NOMIKOS, Petros M. (ed.). 2016 IEEE Conference on Computational Intelligence and Games. 2016 IEEE Conference on Computational Intelligence and Games. 2016.
- SIWEK, STEPHEN E. Video Games in the 21st Century: The 2017 Report [online]. 2017 [visited on 2017-04-12]. Available from: http : / / www . theesa . com / wp - content / uploads / 2017 / 02 / ESA _ EconomicImpactReport_Design_V3.pdf.
- MCCOURT, J. Year-end DEG Home Entertainment Spending [online]. 2016 [visited on 2017-04-12]. Available from: http://degonline. org/wp-content/uploads/2016/01/External_2015- Year- endDEG-Home-Entertainment-Spending1-5-2016.pdf.
- TAMASSIA, Marco; RAFFEY, William; SIFAZ, Rafet; DRACHENX, Anders; ZAMBETTA, Fabio; HITCHENS, Michael. Predicting Player Churn in Destiny: A

Hidden Markov Models Approach to Predicting Player Departure in a Major Online Game. In: NOMIKOS, Petros M. (ed.). 2016 IEEE Conference on Computational Intelligence and Games. 2016, p. 325.

- SHAKER, Noor; ABOU-ZLEIKHA, Mohamed. Transfer Learning for Cross-Game Prediction of Player Experience. In: NOMIKOS, Petros M. (ed.). 2016 IEEE Conference on Computational Intelligence and Games. 2016, p. 209.

DATA MINING AND PREDICTIVE ANALYSIS Lab (ICT 3171

ORIGINALITY REPORT

14%	14%	3%	%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	www.upgrad.com Internet Source	3%
2	www.questjournals.org Internet Source	3%
3	datascience.fm Internet Source	3%
4	ijariie.com Internet Source	2%
5	dergipark.org.tr Internet Source	1%
6	ijarcce.com Internet Source	1%
7	ejournal.unsrat.ac.id Internet Source	<1%
8	codesria.org Internet Source	<1%
9	ekmair.ukma.edu.ua Internet Source	<1%

10	www.ideals.illinois.edu Internet Source	<1 %
11	link.springer.com Internet Source	<1 %

Exclude quotes	On	Exclude matches	< 3 words
Exclude bibliography	On		