SALES ANALYSIS ON VIDEO GAMES

Madhumitha R [#], Vishakhaa S [#], Nithin Kodipyaka[#]

madhumitha.2020a@vitstudent.ac.in

vishakhaa.s2020@vitstudent.ac.in

kodipyaka.nithin2020@vitstudent.ac.in

***ABSTRACT*: Understanding the market for a given industry is important for product development and sales tactics. This can be accomplished by analysing previous sales and identifying helpful patterns. Over the last few years, the video game sector has seen substantial growth. The market has been evolving with the introduction of new gaming platforms and the creation of new genres. Identifying how industry developments have impacted the market might be critical in identifying prospects for a new product or to improve an existing one. This research examines how different factors such as Critic score, User score, and ratings effect global and country sales, and uses sales regression analysis and the XGBoost algorithm to forecast worldwide sales over time. We used sales data from games released up through 2016 for this project. What characteristics have you discovered a link between and the outcomes you've discovered? What were your predictions, and how correct were they? (Write this from the results of your ML code, and the algo you used is XGboost, not Random forest).**

I. KEY WORDS:

Video games, Ratings of expert critics, User ratings, machine learning, performance measures, random forest

II. INTRODUCTION

In today's society, video games are played by groups of people of various ages. On a video screen, these video games are played (on television, computer). There are numerous video games categorised by platform and genre, such as WII, NES, GB, and puzzle, racing, and shooting games. A publisher, such as Activision or Nintendo, releases video games on a variety of platforms. Video games have evolved into a major source of pleasure, particularly for children, and are also employed for corporate purposes. Video games like NFS, Call of Duty, Battlefield, and many others have made and continue to make a lot of money for the past two to three decades. Like any other kind of entertainment, video game creation and writing is typically a multi-disciplinary endeavour.

Due to regional tastes, sales of various sorts of games vary greatly between countries. Consumers in Japan buy far more handheld games than console and, in particular, PC games, with a strong predilection for games that appeal to local interests. Another significant distinction is that, despite their demise in the West, arcade games remain a significant part of the Japanese gaming business. In South Korea, computer games, particularly MMORPGs and real-time strategy games, are often favoured over console games. In China, computer games are also popular.

Aside from being entertaining, well-designed video games have been shown to be beneficial in education for people of all ages and comprehension levels. Video game learning principles have been identified as feasible strategies for reforming the US educational system. It has been observed that gamers have such a focused mindset while playing that they are unaware they are learning, and that if the same mentality could be applied at school, education would gain significantly. While playing video games, students are found to be "learning by doing" and developing creative thinking.

III. LITERATURE REVIEW:

Shidong Yu, Dansheng Yang, and colleagues (2019) used online transaction logs to assess commodity sales, which, unlike traditional visualisation approaches that study products from the perspective of the user, can assist merchants in making better sales decisions. This study proposes using dimensionality reduction to visually examine multi-dimensional time-oriented data sales trends. In addition, the volatility and dynamic performance of sales trends were employed to determine performance. This research also uses rule colour mapping rather than typical linear mapping to successfully show the contrasts between various objects in sales data with a large dynamic range and a skewed distribution of sales.

Chang Hoon is a Korean actor. Oh,Alan M. Rugman (2006) investigates the regional characteristics of top MNEs in the cosmetics and toiletry industry. Their study investigates whether MNEs can produce asymmetric Upstream Firm Specific Strategies and Downstream Firm Specific Strategies. The article analyses the geographic dispersion of countries based on their home region to see if the industry has a global or regional supply chain.

In their article, Daniel Baier and Wolfgang Polasek (2010) explicitly model a spatial component in the production function, and then apply a hierarchical technique to cluster regional sales. Stochastic Partial Derivatives (SPD) restrictions are used in the developed Cross Sectional Sales Response (CSSR) models. They're put through their paces with synthetic and pharma marketing data.

Navjot Kaur, Tabdil Sai Akhil, and others Visualize global video game sales, genre classification, and top games using critic counts and scores as dashboards. Line graphs, bar graphs, pie charts, and heatmaps are among the charts used in the article.

In their paper, Alice Yufa et al. (2019) look at a few independent variables and look at how they relate to global sales. The dependent variable was global sales, while the other variables were deemed independent variables. A platform-based analysis was also performed to determine which video game platform generated the most profit. A stepwise regression was used to run the modelled data. However, the paper does not go into detail about the relationship between genre and sales. The report claims that when the user score is greater, global sales are lower, but no explanation is given as to why.

The dataset was analysed by Jeffery Babb, Neil Terry, and others in 2013 for the years 2006 to 2011. They attempted to discover which part of the gaming industry market has a higher impact on sales in their research. They analysed eight gaming platforms and divided them into three levels based on their sales using the Kruskal-Wallis test. This report also discusses how the game's platform affects sales in home markets.

Yufan Zheng, Jianbin Li (2021), In this research, we present a new hybrid feature selection method that combines Pearson correlation coefficient and Random Forest Feature Selection (PCC-RFFS), and we apply machine learning methods in conjunction with PCC-RFFS to forecast video game sales. The Pearson correlation coefficient and feature ranking technique were utilised for feature selection, while Random Forest was employed as the machine learning algorithm to quantify the value of features and targets. The combination strategy is used in addition to both stages.

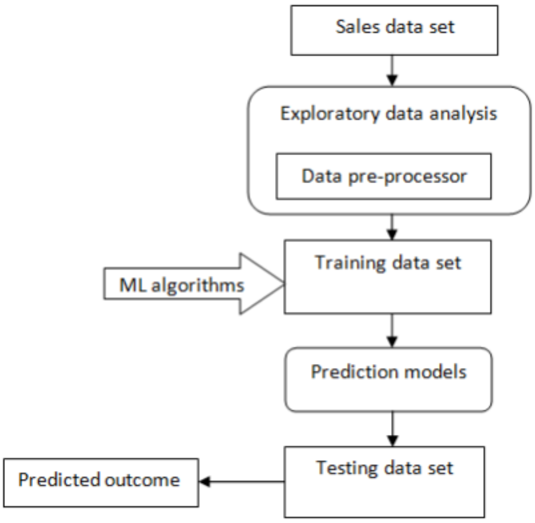
M.Kimura et al. (2015) examined the effects of various marketing tactics on console game sales in Japan in their article. In this paper, the author proposes a performance calculator model for estimating sales fluctuation. The performance of the second video game model can be predicted and forecasted based on the previous model's sales.

Traditional forecasting methods and machine learning algorithms for vegetable sales are compared by Yakul Turgupt et al (2021) in their article.

They used this to assess the Xgboost's performance using their data.

IV. METHODOLOGY:

To acquire the best forecasts, a few steps can be taken to gather data, analyse it, and model it.



V. DATASET:

We chose video game sales data for this research; our dataset has 11 variables and 500 samples, with a mix of categorical and numeric variables. Rank, Name of Video Game, Platform, Year, Genre, Publisher, North American Sales, Europe Sales, Japan Sales, Other Sales, and Global Sales are the categories.

COLUMN NAME	DESCRIPTION
Name	Name Of the Game
Platform	Console On Which the Game Is Running
Year_Of_Release	Year Of the Game Released
Genre	Game's Category
Publisher	Publisher
Na_Sales	Game Sales in North America (In Millions Of Units)
Eu_Sales	Game Sales in The European Union (In Millions Of Units)
Jp_Sales	Game Sales in Japan (In Millions Of Units)
Other_Sales	Game Sales in The Rest of The World, I.E. Africa, Asia Excluding Japan, Australia, Europe Excluding The E.U.
Global_Sales	Total Sales in The World (In Millions of Units)
Critic_Score	Aggregate Score Compiled By Metacritic Staff
Critic_Count	The Number of Critics Used In Coming Up With The Critic_Score
User_Score	Score By Metacritic's Subscribers
User_Count	Number Of Users Who Gave the User_Score
Developer	Party Responsible For Creating The Game
Rating	The ESRB Ratings (E.G. Everyone, Teen, Adults Only...Etc)

VI. ALGORITHM USED:

Random Forest: - Random Forest is a supervised machine learning technique that generates a forest of several trees at random. To generate a set of random trees, use the Random Forest operator. The random trees are generated in the same way that a tree is generated using the Random Tree operator. A given number of random tree models are included in the resultant forest model. The required number of trees is specified by the number of trees parameter. Why do we use random forest instead of decision tree? Decision trees are simple to design and work with training data rapidly, but they provide poorer accuracy owing to overfitting. Overfitting occurs when a model trains the data to the point where it degrades the model's performance on new data. As a result, the random forest appears. train (formula, dataset, method="rf,"trControl=trcontrol()) [where "rf" refers to the random forest approach].

VII. RESULTS & DISCUSSION:

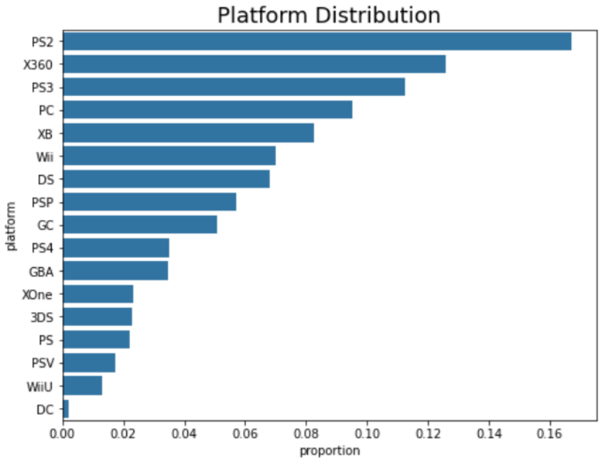


Fig 1

Play Station 2 accounts for 16 percent of the games in our database.

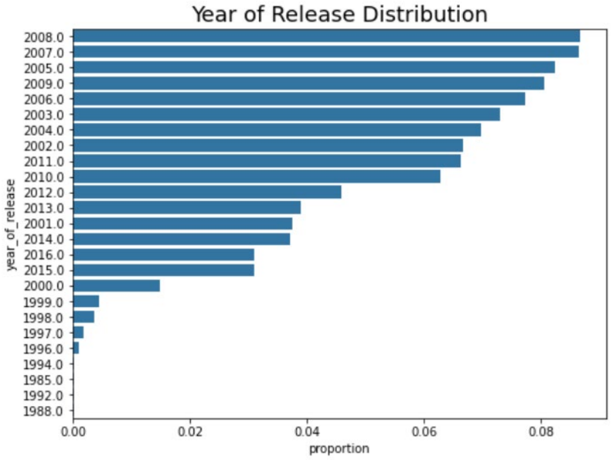


Fig 2

The majority of the games in the collection were released prior to 2011.

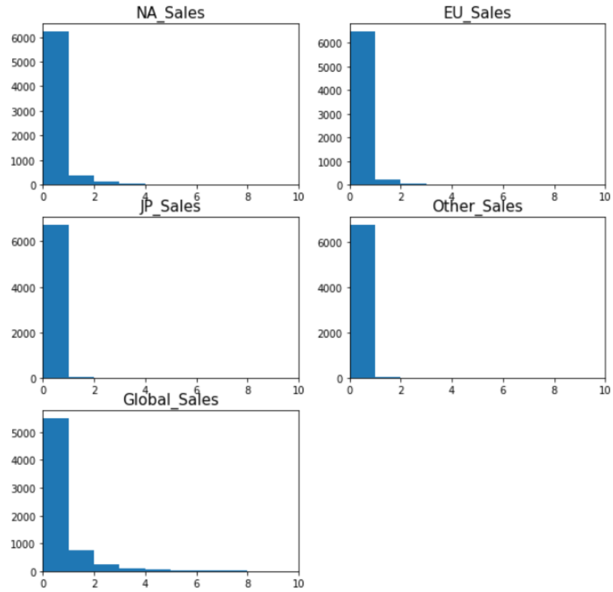


Fig 3

As can be seen from the graphs above, the vast majority of games fail to reach the 1 million sales mark.

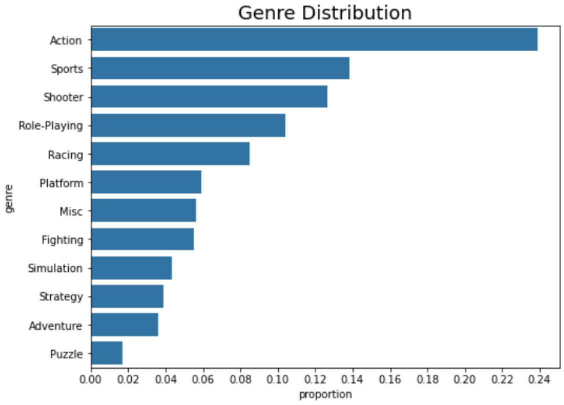


Fig 4

The majority of games contain action.

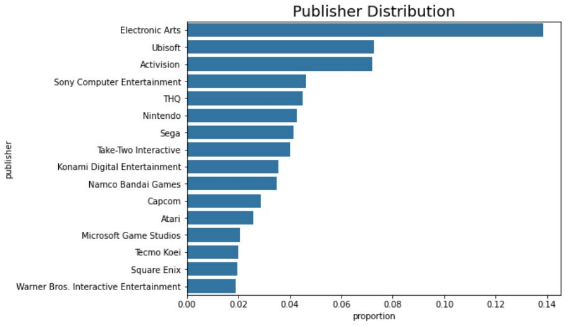


Fig 5

Electronic Arts has released games that have made a strong comeback in terms of sales.

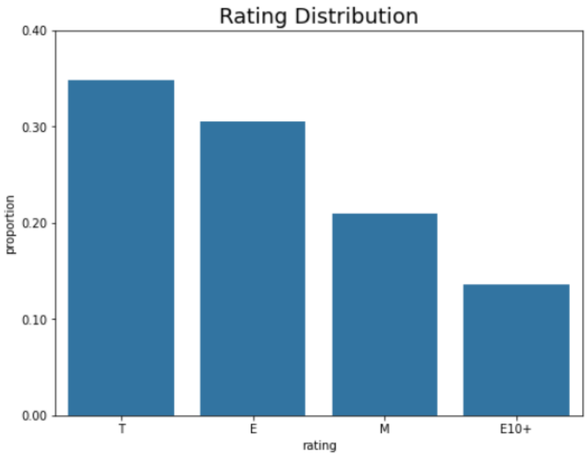


Fig 6

T(for teens) is assigned to over 35% of the games in the sample.

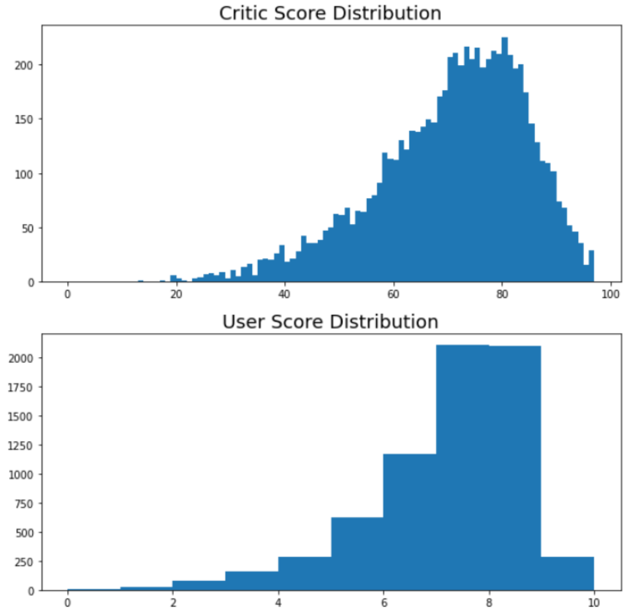


Fig 7

Critic score and user score are very similar.

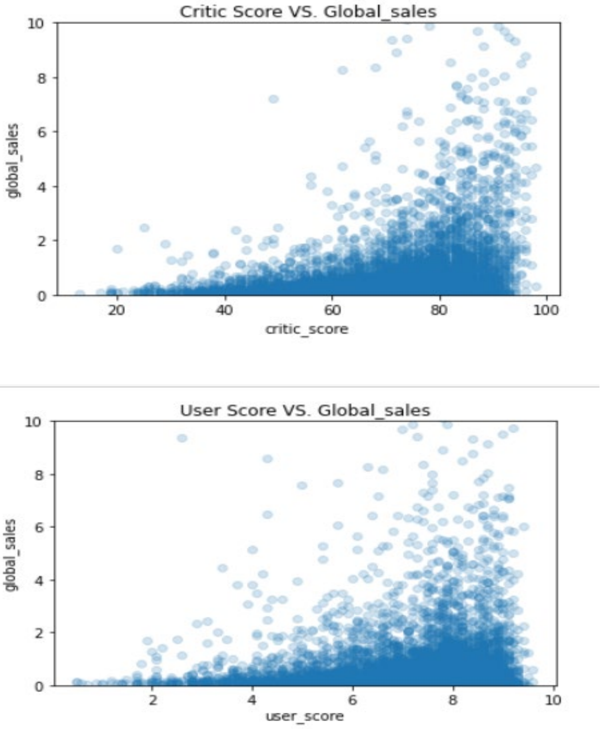


Fig 8

There is a positive relashionship between users score and global sales.Hence, users score does have an imapact on global sales

.

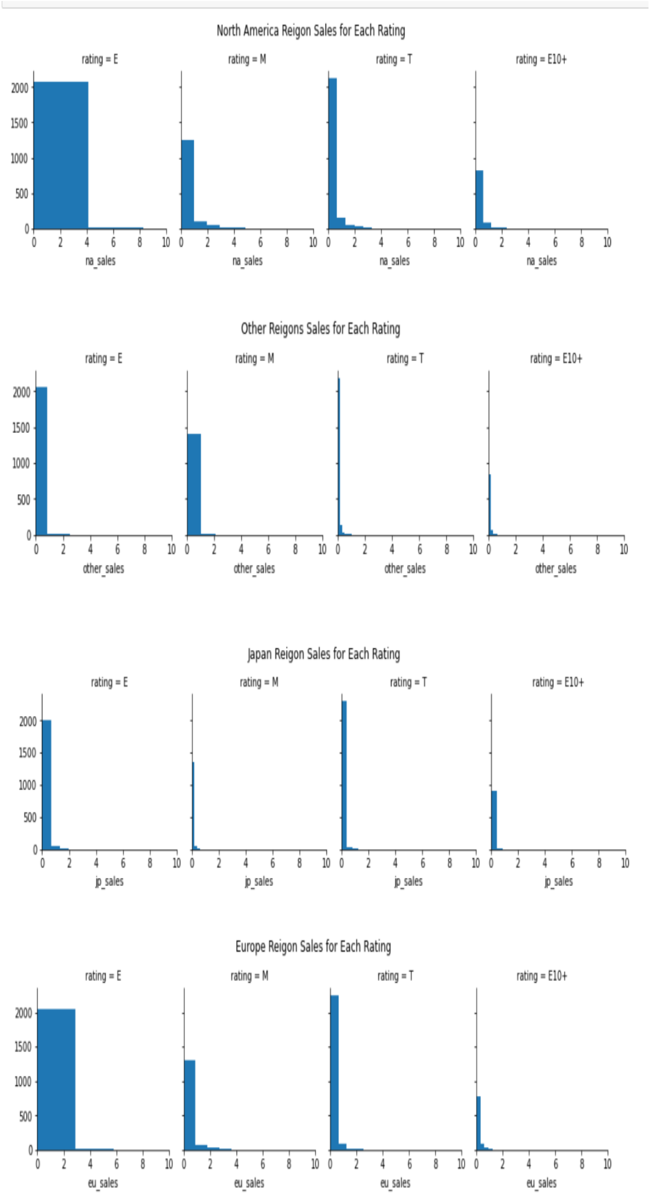


Fig 9

From the visuals above we can see that, Rating E(for everyone) sells the most in all regions.The games that have an M(for mature) rating sells the most in regions other than(NA, EU, JP)The games that have an T(for teens) rating sells the most in the NA region

.

.

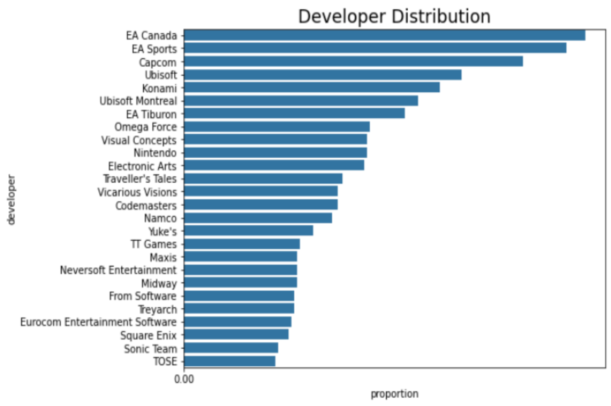


Fig 10

EA Canada owes a higher part

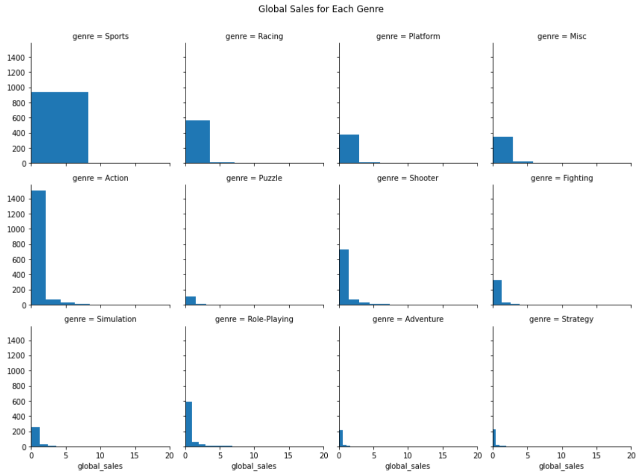


Fig 11

Sports, Action, and Racing games sell the most globally.Puzzle, and Strategy games sell the least globally.

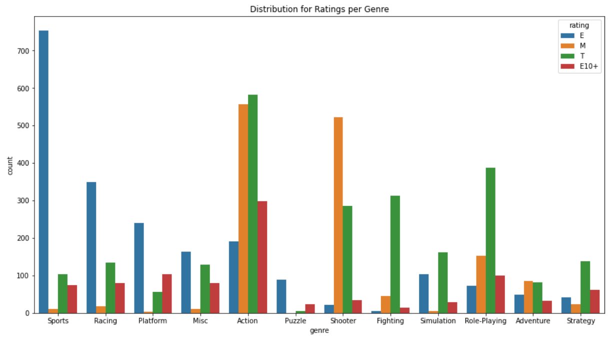


Fig 12

Fighiting, and Role-Playing games are mostly aimed towards teens.Sports games are aimed towards everyone

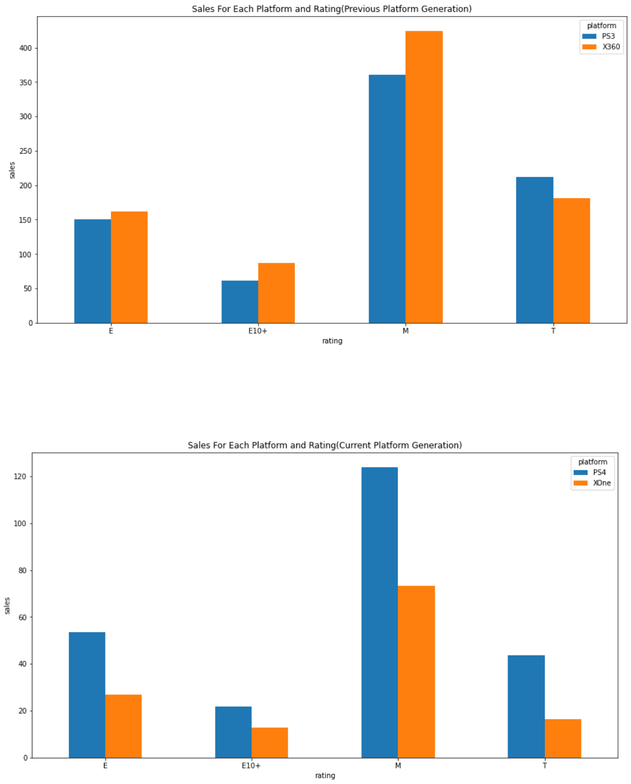


Fig 13

These visuals shows that the competition between Xbox 360, and PS3 was close. But in the current generation PS4 dominates the market.

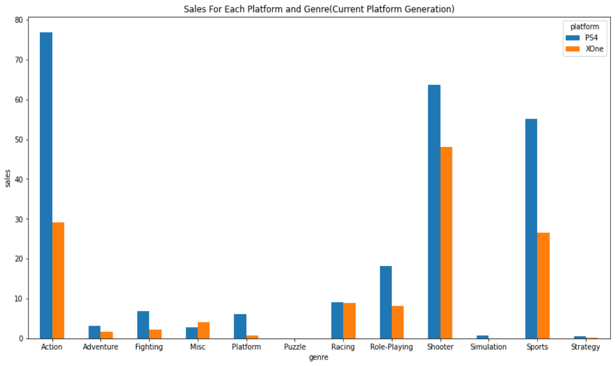


Fig 14

All the popular genres sell more on PS4.

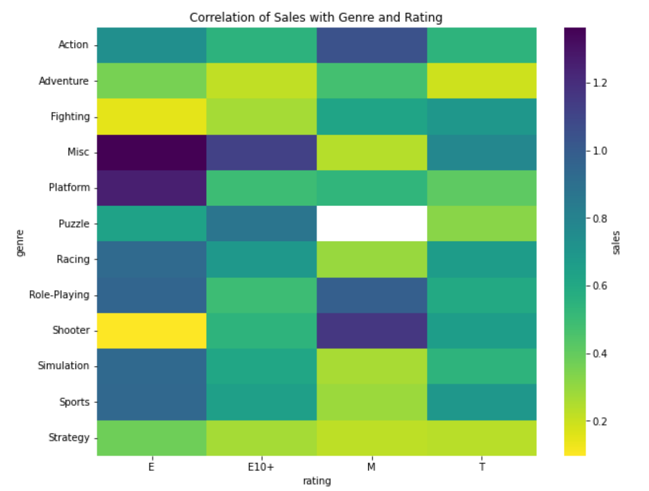
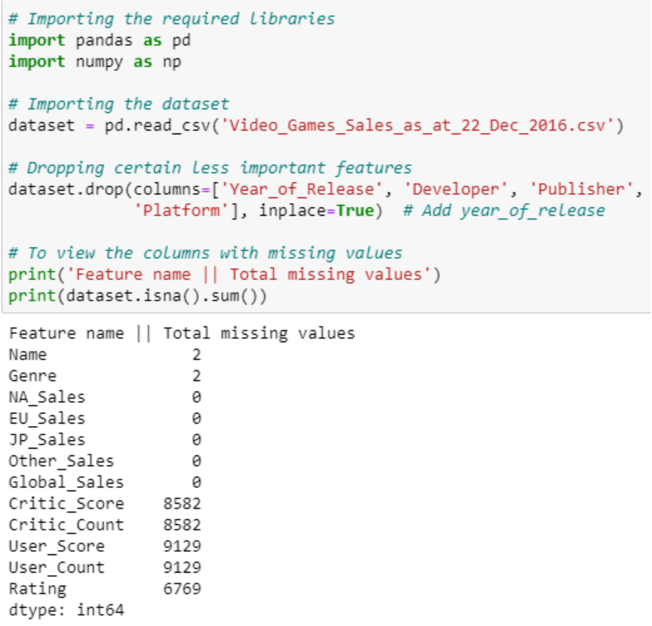


Fig 15

We can see that misc games that rated E(for everyone) sell the highest, while puzzle games that are rated M(for mature) sell the least.

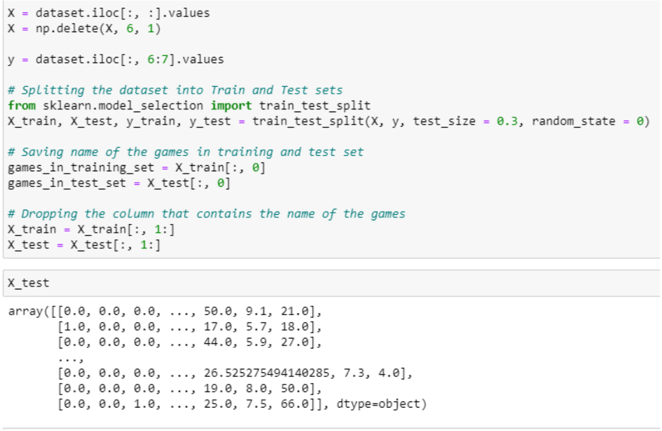
PREDICTING VIDEO GAME SALES USING ML:

We'll need to use a regression model because global video game sales are a continuous variable. Regression is a type of supervised machine learning technique that uses a set of independent features to predict a target variable (which should be a continuous value). Salary forecasting, real estate forecasting, and other applications are among them.



```
# Importing the required libraries
import pandas as pd
import numpy as np

# Importing the dataset
dataset = pd.read_csv('Video_Games_Sales_as_at_22_Dec_2016.csv')

# Dropping certain less important features
dataset.drop(columns=['Year_of_Release', 'Developer', 'Publisher',
             'Platform'], inplace=True)  # Add year_of_release

# To view the columns with missing values
print('Feature name || Total missing values')
print(dataset.isna().sum())
```

```
Feature name || Total missing values
Name               2
Genre              2
NA_Sales           0
EU_Sales           0
JP_Sales           0
Other_Sales        0
Global_Sales       0
Critic_Score    8582
Critic_Count    8582
User_Score      9129
User_Count      9129
Rating          6769
dtype: int64
```

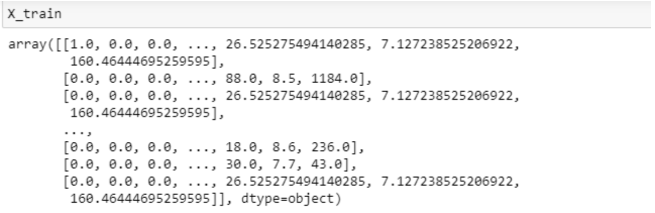


```
X = dataset.iloc[:, :].values
X = np.delete(X, 6, 1)

y = dataset.iloc[:, 6:7].values

# Splitting the dataset into Train and Test sets
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.3, random_state = 0)

# Saving name of the games in training and test set
games_in_training_set = X_train[:, 0]
games_in_test_set = X_test[:, 0]

# Dropping the column that contains the name of the games
X_train = X_train[:, 1:]
X_test = X_test[:, 1:]
```

```
X_test
```

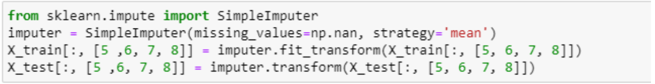
```
array([[0.0, 0.0, 0.0, ..., 50.0, 9.1, 21.0],
       [1.0, 0.0, 0.0, ..., 17.0, 5.7, 18.0],
       [0.0, 0.0, 0.0, ..., 44.0, 5.9, 27.0],
       ...,
       [0.0, 0.0, 0.0, ..., 26.525275494140285, 7.3, 4.0],
       [0.0, 0.0, 0.0, ..., 19.0, 8.0, 50.0],
       [0.0, 0.0, 1.0, ..., 25.0, 7.5, 66.0]], dtype=object)
```

Here, we set up 'X' and 'y,' with 'X' representing the set of independent variables and 'y' representing the target variable, Global Sales. Before the dataset is separated into training and test sets, the Global Sales column, which is present at index 6 in 'X,' is removed using the np.delete() method. The names of the games are saved in a separate array called 'games in training set' and 'games in test set,' as these names won't help us estimate global sales very well.

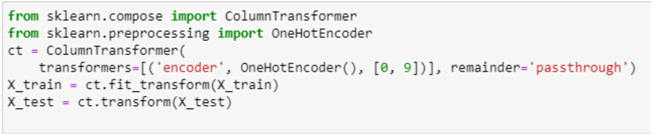


```
X_train
```

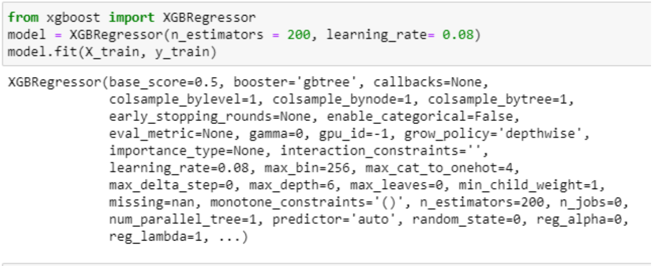
```
array([[1.0, 0.0, 0.0, ..., 26.525275494140285, 7.127238525206922,
        160.46444695259595],
       [0.0, 0.0, 0.0, ..., 88.0, 8.5, 1184.0],
       [0.0, 0.0, 0.0, ..., 26.525275494140285, 7.127238525206922,
        160.46444695259595],
       ...,
       [0.0, 0.0, 0.0, ..., 18.0, 8.6, 236.0],
       [0.0, 0.0, 0.0, ..., 30.0, 7.7, 43.0],
       [0.0, 0.0, 0.0, ..., 26.525275494140285, 7.127238525206922,
        160.46444695259595]], dtype=object)
```



```
from sklearn.impute import SimpleImputer
imputer = SimpleImputer(missing_values=np.nan, strategy='mean')
X_train[:, [5 ,6, 7, 8]] = imputer.fit_transform(X_train[:, [5, 6, 7, 8]])
X_test[:, [5 ,6, 7, 8]] = imputer.transform(X_test[:, [5, 6, 7, 8]])
```



```
from sklearn.compose import ColumnTransformer
from sklearn.preprocessing import OneHotEncoder
ct = ColumnTransformer(
    transformers=[('encoder', OneHotEncoder(), [0, 9])], remainder='passthrough')
X_train = ct.fit_transform(X_train)
X_test = ct.transform(X_test)
```



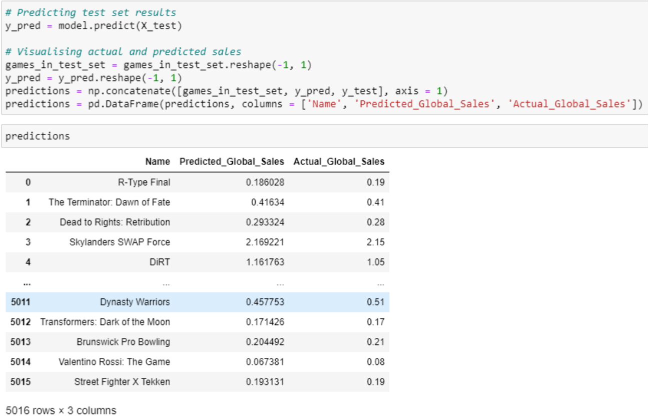
```
from xgboost import XGBRegressor
model = XGBRegressor(n_estimators = 200, learning_rate= 0.08)
model.fit(X_train, y_train)
```

```
XGBRegressor(base_score=0.5, booster='gbtree', callbacks=None,
             colsample_bylevel=1, colsample_bynode=1, colsample_bytree=1,
             early_stopping_rounds=None, enable_categorical=False,
             eval_metric=None, gamma=0, gpu_id=-1, grow_policy='depthwise',
             importance_type=None, interaction_constraints='',
             learning_rate=0.08, max_bin=256, max_cat_to_onehot=4,
             max_delta_step=0, max_depth=6, max_leaves=0, min_child_weight=1,
             missing=nan, monotone_constraints='()', n_estimators=200, n_jobs=0,
             num_parallel_tree=1, predictor='auto', random_state=0, reg_alpha=0,
             reg_lambda=1, ...)
```

In machine learning, imputation is a technique for replacing missing data with replaced values. To impute the columns with missing values, we'll use the Imputer class from the scikit-learn package, and to impute the columns with values of type string, we'll use CategoricalImputer from sklearn pandas, and replace the missing values with 'NA', which stands for Not Available.

ColumnTransformer and OneHotEncoder from the scikit-learn library are used to encode the categorical columns of 'X.' Each category contained in a categorical column of 'X' will be assigned to a separate column.

We'll use XGBRegressor to implement our model, specifically the regressor (where XGB stands for extreme gradient boosting). XGBoost is a decision tree-based ensemble machine learning technique comparable to the RandomForest algorithm. XGBoost, on the other hand, mixes trees that aren't too deep, unlike RandomForest, which uses fully developed trees. In addition, in compared to RandomForest, the number of trees joined in XGBoost is higher. Ensemble algorithms combine weak learners to create a powerful learner. When compared to gradient boosting, XGBoost includes additional performance and speed capabilities.



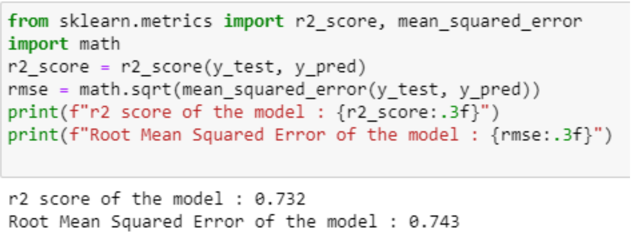
```
# Predicting test set results
y_pred = model.predict(X_test)

# Visualising actual and predicted sales
games_in_test_set = games_in_test_set.reshape(-1, 1)
y_pred = y_pred.reshape(-1, 1)
predictions = np.concatenate([games_in_test_set, y_pred, y_test], axis = 1)
predictions = pd.DataFrame(predictions, columns = ['Name', 'Predicted_Global_Sales', 'Actual_Global_Sales'])
```

```
predictions
```

	Name	Predicted_Global_Sales	Actual_Global_Sales
0	R-Type Final	0.186028	0.19
1	The Terminator: Dawn of Fate	0.41634	0.41
2	Dead to Rights: Retribution	0.293324	0.28
3	Skylanders SWAP Force	2.169221	2.15
4	DiRT	1.161763	1.05
...	...	...	...
5011	Dynasty Warriors	0.457753	0.51
5012	Transformers: Dark of the Moon	0.171426	0.17
5013	Brunswick Pro Bowling	0.204492	0.21
5014	Valentino Rossi: The Game	0.067381	0.08
5015	Street Fighter X Tekken	0.193131	0.19

5016 rows × 3 columns

The model.predict() method is used to predict Global Sales, which is the target variable 'y' for the games in the test set.



```
from sklearn.metrics import r2_score, mean_squared_error
import math
r2_score = r2_score(y_test, y_pred)
rmse = math.sqrt(mean_squared_error(y_test, y_pred))
print(f"r2 score of the model : {r2_score:.3f}")
print(f"Root Mean Squared Error of the model : {rmse:.3f}")
```

```
r2 score of the model : 0.732
Root Mean Squared Error of the model : 0.743
```

To evaluate the model's performance, we'll use r2 score and root mean squared error (RMSE), where the closer the r2 score is to 1 and the lower the magnitude of RMSE, the better.

As The fact that the r2 score is so near to 1 implies that the model is quite accurate. You can also try tweaking the XGBoost regressor's hyperparameters to improve model performance.

CONCLUSION:

Predicting sales is an important aspect of the strategic planning process. It enables a business to forecast how it will perform in the future. The three countries that make up North America account for a significant portion of overall video game sales revenue and are one of the world's most important markets for developers to target. We reasoned those good sales in North America would translate to strong global sales. We were unsure about the impact of genre. This is partly due to the lack of a clear definition of genre. A video game can have elements from a variety of genres; for example, an action game could include shooting, adventure, riddles, and role playing. Because video games contain so many moving pieces, it can be difficult to categorise them into a single genre. We proceeded into the study with the assumption that customers purchase games based on their overall qualities and features rather than their genre classification.

Predicting a company's sales is useful not just for identifying new opportunities, but also for identifying any bad tendencies that occur in the forecast. Finally, we conclude that video game sales predictions have been made, and we have determined which game has the highest global sales. We used XGB to apply Random Forest to anticipate video game sales.

VIII. REFERENCES :

1.Dongsheng Yang,Shidong Yu,Ying Hao, Visual Analysis of Sorting and Classification of Multidimensional Data, International Journal of Pattern Recognition and Artificial Intelligence, July 2020.

2.Jeffrey Babb, Neil Terry et al The Impact Of Platform On Global Video Game Sales September 2013 International Business & Economics Research Journal (IBER)

[3] David Buckley, Ke Chen and Joshua Knowles, “Predicting skill from game play input to a first

person shooter”, 2013, 978-1-4673-5311-3/13, IEEE.

3.Predicting Video Game Sales Based on Machine Learning and Hybrid Feature Selection Method November 2021 Conference: 2021 16th International Conference on Intelligent Systems and Knowledge Engineering (ISKE)

4.M.Kimura et al, Effects for console game sales in Japan market, Asian Pacific Journal of Marketing and Logistics, 12 January 2015

5.Chang Hoon Oh, Alan M. Rugman, Regional Sales of Multinationals in the World Cosmetics Industry,
European Management Journal,
Volume 24,
Issues 2–3, 2006,
Pages 163-173,
ISSN 0263-2373

6.Daniel Baier, D., Polasek, W. (2010). Marketing and Regional Sales: Evaluation of Expenditure Strategies by Spatial Sales Response Functions. In: Locarek-Junge, H., Weihs, C. (eds) Classification as a Tool for Research. Studies in Classification, Data Analysis, and Knowledge Organization. Springer, Berlin, Heidelberg.

https://doi.org/10.1007/978-3-642-10745-0_74

7. Akhil, Tabdil Sai, Navjot Kaur, B. Surendranath Reddy, B. Vinay, and K. Nanda Kishore. "Data Visualization on video games global sales analysis & Predictive analysis on Real Estate pricing in Boston."

8.Li, Jianbin, Yufan Zheng, Haoran Hu, Junhui Lu, and Choujun Zhan. "Predicting Video Game Sales Based on Machine Learning and Hybrid Feature Selection Method." In 2021 16th International Conference on Intelligent Systems and Knowledge Engineering (ISKE), pp. 497-502. IEEE, 2021.

9.Turgut, Yakup, and Mustafa Erdem. "Forecasting of Retail Produce Sales Based on XGBoost Algorithm."
In Global Joint Conference on Industrial Engineering and Its Application Areas,
pp. 27-43. Springer, Cham, 2020.

[11] P. Boinee, A. D. Angelis, and G. Foresti, "Meta random forests," International Journal of Computational Intelligence, vol. 2, no. 3, pp. 138-147, 2005.