<u>Video Games Sales Dataset</u>

PROJECT REPORT

Submitted by

Ankit Singh- 201020407
Priyank Singh- 201020442

Dr. Shyama Prasad Mukherjee International Institute of Information Technology, Naya Raipur

February-2022

1. INTRODUCTION

Video games sales Analysis dataset is a detailed analysis of different video games sales in different regions across the globe. The objective of the dataset is to help different video games companies to better understand their market region and customers and makes it easier for them to modify their products according to the specific needs, behaviors, and concerns of different types of customers. The dataset consists of several predictor variables like NA_Sales, EU_Sales, JP_Sales, Other_Sales, Global_Sales, etc. but for prediction, we have taken NA_Sales and EU_Sales as they are highly correlated.

1.1 Regression Analysis

Regression analysis is a solid technique for recognizing which factors affect a subject of revenue. The method involved with playing out a relapse permits you to unquestionably figure out which variables make the biggest difference, which elements can be overlooked, and the way that these elements impact one another.

Dependent Variable: This is the fundamental component so that we are having only one dependent variable in the dataset.

Independent Variables: These are the elements that we theorize affect our reliant variable. We are having 20 independent variables in the dataset.

1.2 Dataset

The objective of the Video Games Sales dataset is to help different video games companies to better understand their market region and customers and makes it easier for them to modify their products according to the specific needs, behaviors, and concerns of different types of customers.
For example, instead of spending money to market a new product to every customer in the company's database, a company can analyse which customer segment is most likely to buy the product and then market the product only on that particular segment.

```
import seaborn as sns
import matplotlib.pyplot as plt
sns.set_style("darkgrid")
df=pd.read_csv(r"C:\Users\ankit\Downloads\video_games_sales.csv")
print(df.head())
```

```
                      Name Platform  Year_of_Release         Genre Publisher  \
0               Wii Sports      Wii           2006.0        Sports  Nintendo
1        Super Mario Bros.      NES           1985.0      Platform  Nintendo
2           Mario Kart Wii      Wii           2008.0        Racing  Nintendo
3        Wii Sports Resort      Wii           2009.0        Sports  Nintendo
4  Pokemon Red/Pokemon Blue      GB           1996.0  Role-Playing  Nintendo

   NA_Sales  EU_Sales  JP_Sales  Other_Sales  Global_Sales  Critic_Score  \
0     41.36     28.96      3.77         8.45         82.53          76.0
1     29.08      3.58      6.81         0.77         40.24           NaN
2     15.68     12.76      3.79         3.29         35.52          82.0
3     15.61     10.93      3.28         2.95         32.77          80.0
4     11.27      8.89     10.22         1.00         31.37           NaN

   Critic_Count User_Score  User_Count Developer Rating
0          51.0          8       322.0  Nintendo      E
1           NaN        NaN         NaN       NaN    NaN
2          73.0        8.3       709.0  Nintendo      E
3          73.0          8       192.0  Nintendo      E
4           NaN        NaN         NaN       NaN    NaN
```

```
df.shape
```

```
(16719, 16)
```

2. LITERATURE

In this project, we used machine learning methods like Linear Regression, Multiple Regression, Polynomial Regression, Gradient Descent method, Regularization. In Regularization, there are two methods. They are Ridge and Lasso. These are the methods that we used in this project.

2.1 Linear Regression

In statistics, linear regression is a straight methodology for displaying the connection between a scalar reaction and at least one informative factor (otherwise called reliant and free factors). The instance of one informative variable is called simple linear regression. This term is unmistakable from multivariate direct relapse, where numerous related ward factors are anticipated, rather than a solitary scalar variable. The model assumes a linear relationship between the input variables (x) and the single output variable (y). This is all about linear regression.

2.2 Multiple Regression

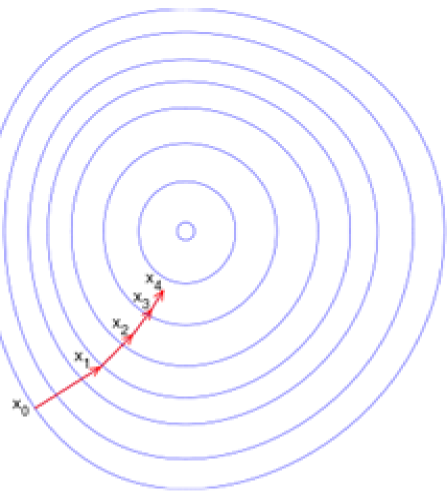
In statistics, linear regression is a straight methodology for displaying the connection between a scalar reaction and at least one informative factor for multiple, the interaction is called multiple linear regression. This term is unmistakable from multivariate direct relapse, where numerous related ward factors are anticipated, rather than a solitary scalar variable. The relationship between a single dependent variable and several independent variables. This is about multiple regression.

2.3 Polynomial Regression

In statistics, polynomial regression is a type of regression examination wherein the connection between the autonomous variable x and the reliant variable y is demonstrated as a most extreme limit polynomial in x. Polynomial regression fits a nonlinear connection between the worth of x and the comparing contingent mean of y, indicated E(y |x). Albeit polynomial regression fits a nonlinear model to the information, as a measurable assessment issue it is straight, as in the relapse work E(y | x) is direct in the obscure boundaries that are assessed from the information.

2.4 Gradient Descent method

Gradient descent is an iterative improvement calculation for tracking down the nearby least of a capacity. To observe the nearby least of a capacity utilizing slope plummet, we should make strides relative to the negative of the angle (create some distance from the inclination) of the capacity at the current point.



2.5 Regularization

In math, measurements, finance, software engineering, especially in machine learning and backward issues, regularization is the method involved with adding data to tackle a poorly presented issue or to forestall overfitting. Regularization can be applied to genuine capacities in badly presented improvement issues.

2.5.1 Ridge

Ridge regression is a method of estimating the coefficients of multiple-regression models in scenarios where linearly independent variables are highly correlated. The coefficients of numerous relapse models in situations where straight autonomous factors are profoundly related.

2.5.2 Lasso

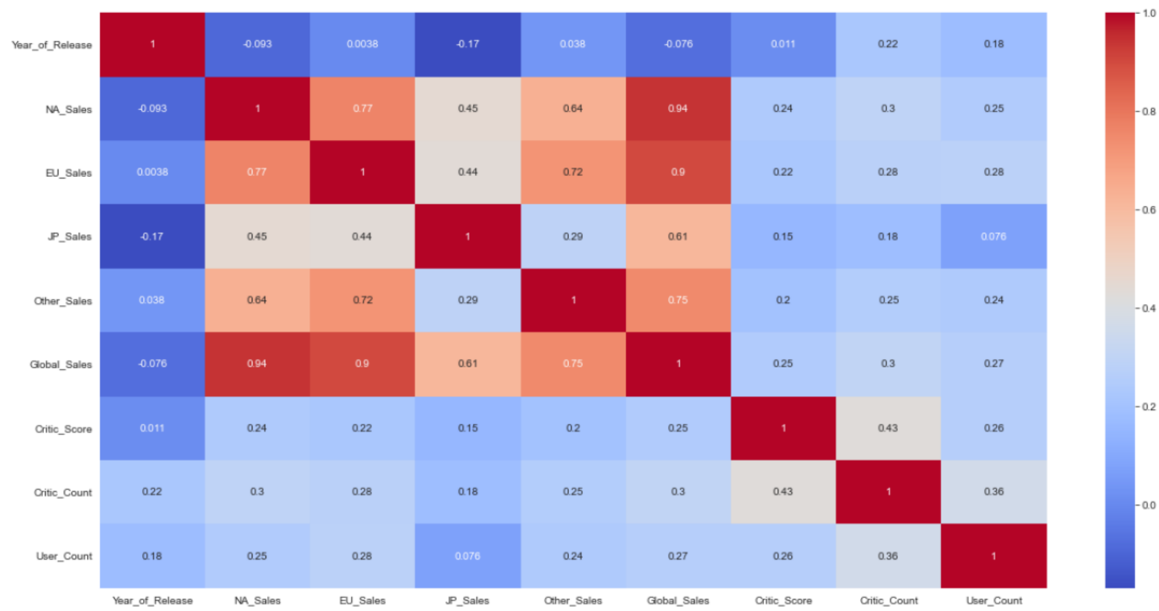
Lasso was initially figured out for linear regression models. This basic case uncovers a significant sum about the assessor. These incorporate its relationship to edge relapse and best subset determination and the associations between tether coefficient gauges thus called delicate thresholding. It likewise uncovers that (like linear regression) the coefficient gauges shouldn't be extraordinary assuming covariates are collinear.

3. Methodology

3.1 Data Preprocessing

Data preprocessing is the process of transforming raw data into an understandable format. It is also an important step in data mining as we cannot work with raw data. The quality of the data should be checked before applying machine learning or data mining algorithms.

Heatmap:



Different preprocessing techniques are:

- **Data Imputation**
  Imputation is the process of replacing missing data with substituted values
  We start with handling missing values in the dataset.

- **Encoding**

  Encoding is the process of converting data into a format required for a number of information processing needs, including: Program compiling and execution. Data transmission, storage and compression/decompression. Application data processing, such as file conversion.

- **Data Discretization**

  Data discretization is defined as a process of converting continuous data attribute values into a finite set of intervals and associating with each interval some specific data value.

- **Outlier Handling**

  Outlier trimming refers to simply removing the outliers beyond a certain threshold value. One of the main advantages of outlier trimming is it is extremely quick and doesn't distort the data. There are several ways to find the thresholds for outlier trimming.

- **Feature Selection**

  Feature selection is also known as Variable selection or Attribute selection. Feature selection is a process where you automatically select those features in your data that contribute most to the prediction variable or output in which you are interested.

NOTE: As our dataset was clean and does not have more null values. So, we don’t have to apply Data preprocessing techniques in our dataset.

3.2 Techniques

3.2.1 Linear Regression

➔ **ASSUMPTIONS**

- **Linearity:** The relationship between the features and target.
- **Homoscedasticity:** The error term has a constant variance.
- **Multicollinearity:** There is no multicollinearity between the features.
- **Independence:** Observations are independent of each other.
- **Normality:** The error(residuals) follows a normal distribution.

➔ **ALGORITHM**

Linear Regression is a supervised machine learning algorithm where the predicted output is continuous and has a constant slope. It's used to predict values within a continuous range, (e.g. sales, price) rather than trying to classify them into categories (e.g. cat, dog).

3.2.2 Multiple Regression

➔ **ASSUMPTIONS**

- **Linearity:** The relationship between the features and target.

- **Homoscedasticity:** The error term has a constant variance.
- **Multicollinearity:** There is no multicollinearity between the features.
- **Independence:** Observations are independent of each other.
- **Normality:** The error(residuals) follows a normal distribution.

➔ **ALGORITHM**

Multiple Linear Regression is one of the important regression algorithms which models the linear relationship between a single dependent continuous variable and more than one independent variable.

3.2.3 Polynomial Regression

➔ **ASSUMPTIONS**

The relationship between the dependent variable and any independent variable is linear or curvilinear (specifically polynomial). The independent variables are independent of each other. The errors are independent, normally distributed with mean zero and a constant variance (OLS).

➔ **ALGORITHM**

Polynomial Regression is a regression algorithm that models the relationship between a dependent(y) and independent variable(x) as nth degree polynomial.

3.2.4 Gradient Descent method

➔ **ASSUMPTIONS**

Stochastic gradient descent is based on the assumption that the errors at each point in the parameter space are additive. The error at point one can be added to the error at point two which can be added to the error at point three, and so on for all of the points.

➔ **ALGORITHM**

Gradient descent is an iterative optimization algorithm for finding the local minimum of a function. To find the local minimum of a function using gradient descent, we must take steps proportional to the negative of the gradient (move away from the gradient) of the function at the current point.

3.2.5 Regularization

3.2.5.1 Ridge

➔ **ASSUMPTIONS**

The assumptions are the same as those used in regular multiple regression: linearity, constant variance (no outliers), and independence. Since ridge regression does not provide confidence limits, normality need not be assumed.

➔ **ALGORITHM**

Ridge regression is a model tuning method that is used to analyse any data that suffers from multicollinearity. This method performs L2 regularization. When the issue of multicollinearity occurs, least-squares are unbiased, and variances are large, this results in predicted values being far away from the actual values.

3.2.5.2 Lasso

➔ **ASSUMPTIONS**

LASSO is not a type of model. It is a method for selecting variables and shrinking coefficients to adjust for complexity. The assumptions are those of the type of model it is applied to, which could be ordinary least squares regression, logistic regression, Cox proportional hazards, or other types of regression, each with their own assumptions.

➔ **ALGORITHM**

The goal of lasso regression is to obtain the subset of predictors that minimizes prediction error for a quantitative response variable. The lasso does this by imposing a constraint on the model parameters that causes regression coefficients for some variables to shrink toward zero.

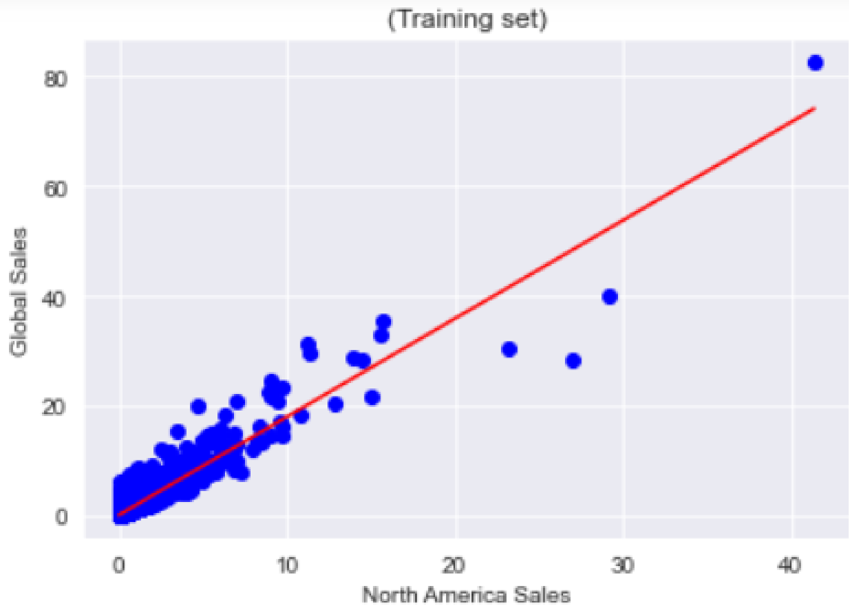
4.Experimentation and result

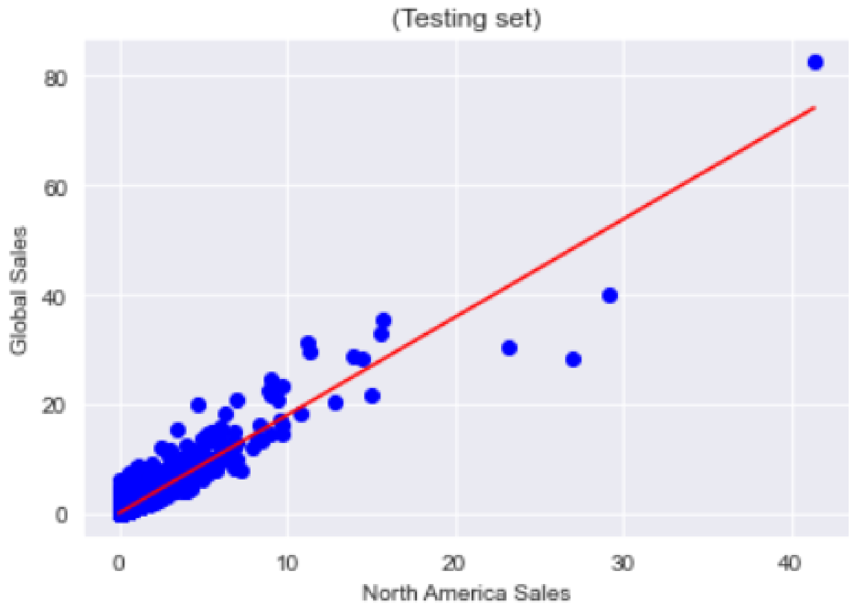
4.1 Linear Regression

4.1.1 Independent Variable: NA_Sales
Dependent Variable: Global_Sales

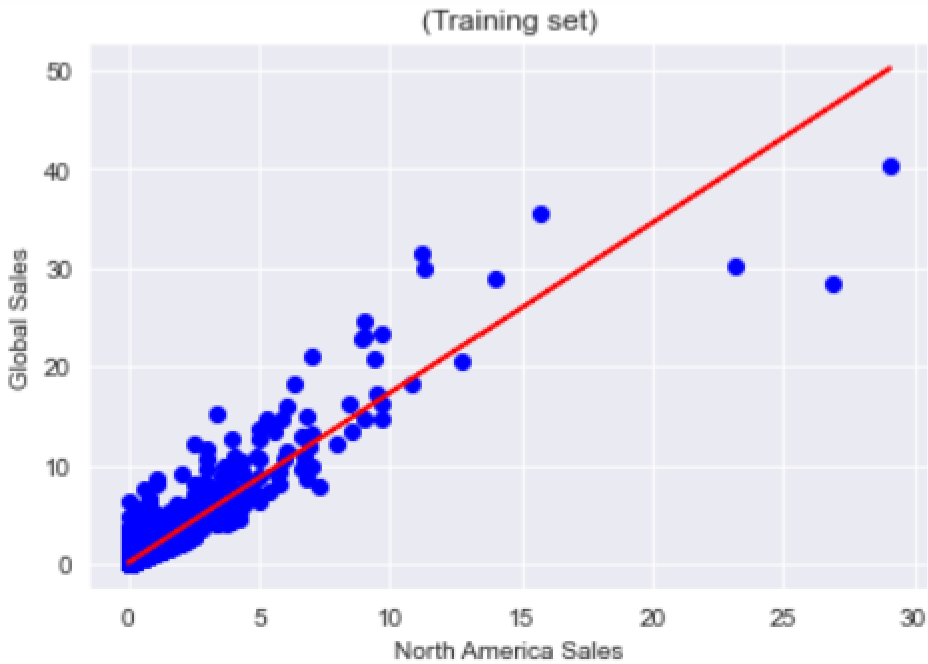
	Without Split	Train:Test 80:20	Train:Test 70:30	Train:Test 50:50
Coefficient	1.79053279	1.72171555	1.71170151	1.65720968
Intercept	0.06204106	0.07751573	0.07804729	0.09390077
R2 Score	0.8855007503060843	0.860	0.862	0.856
MSE	0.2743356033079391	0.2874924044611133	0.26142753890283615	0.26160755670132924
RMSE	0.5237705636134385	0.5361831818148657	0.5112998522421419	0.511475861308556
MAE	0.19745859404243615	0.2025194943758726	0.1993162015463381	0.20047404794672466
R2 Adjusted	0.885	0.860	0.862	0.856
AIC	2.583e+04	2.069e+04	1.856e+04	1.395e+04
BIC	2.584e+04	2.071e+04	1.857e+04	1.397e+04

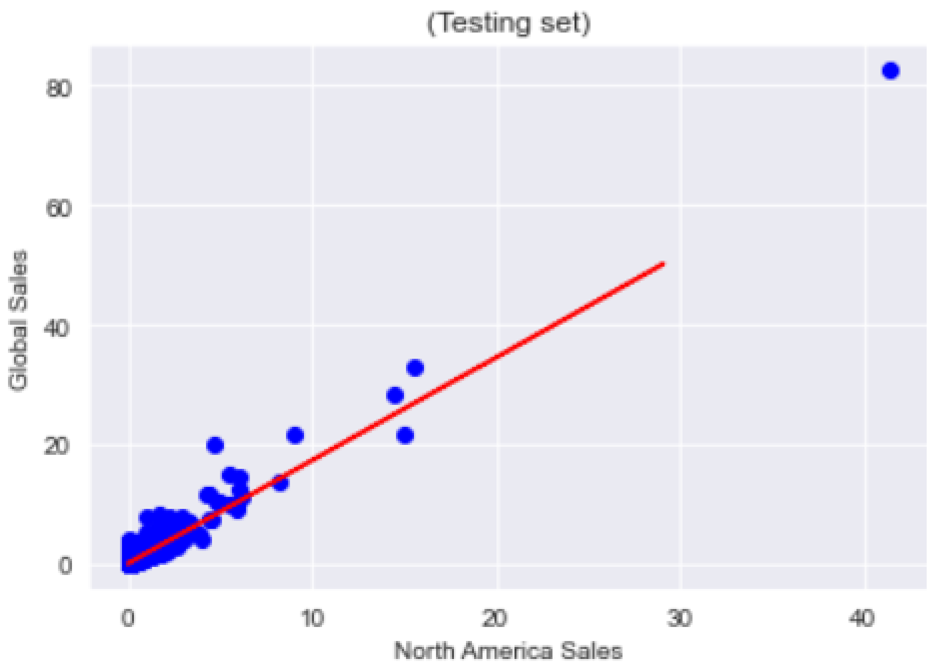
Graph for Without Split





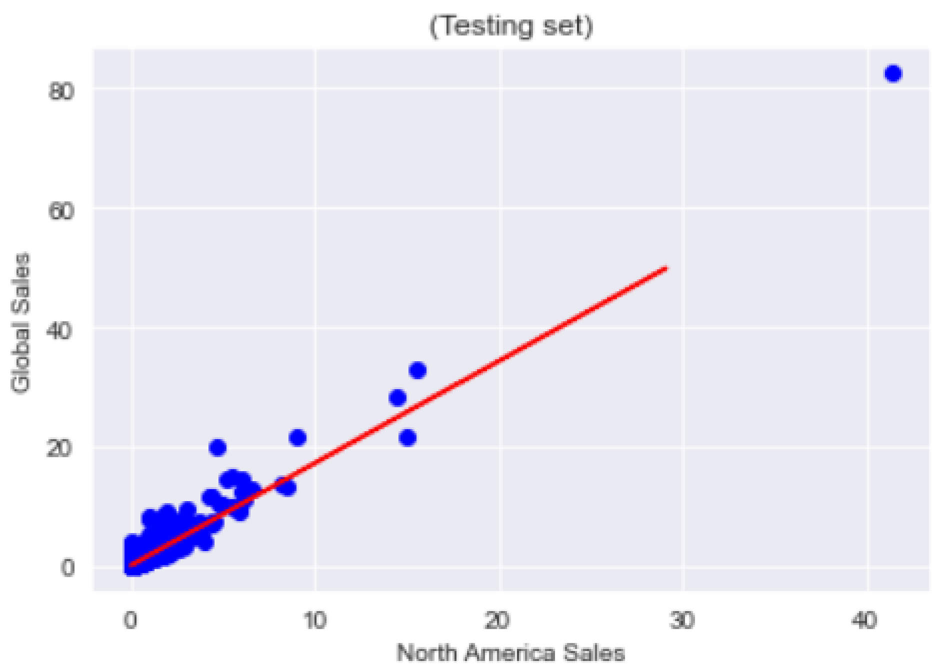
80:20


(Training set)
Global Sales
North America Sales
50
40
30
20
10
0
0
5
10
15
20
25
30

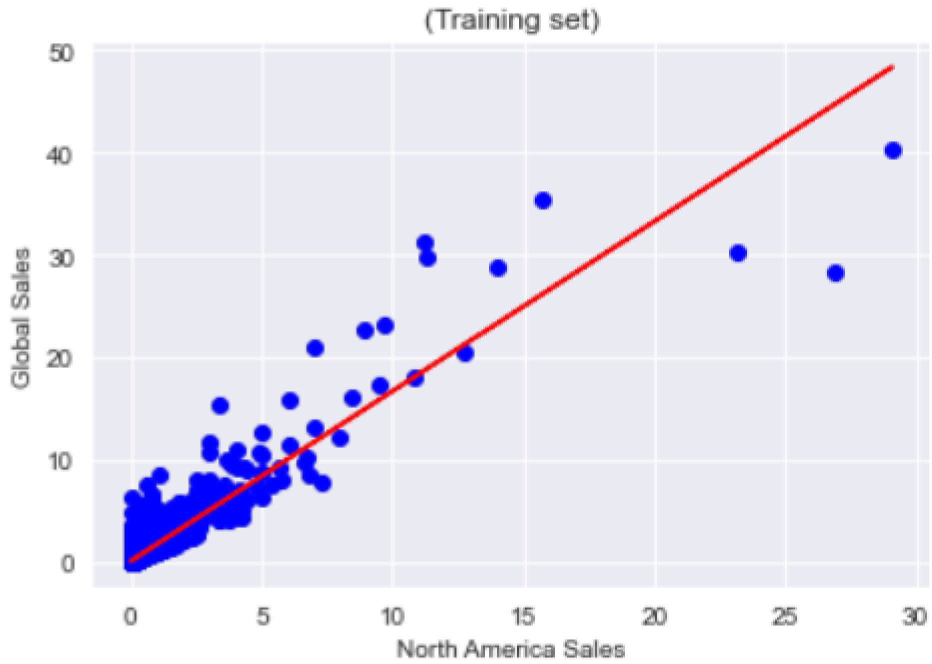

(Testing set)
Global Sales
North America Sales
80
60
40
20
0
0
10
20
30
40

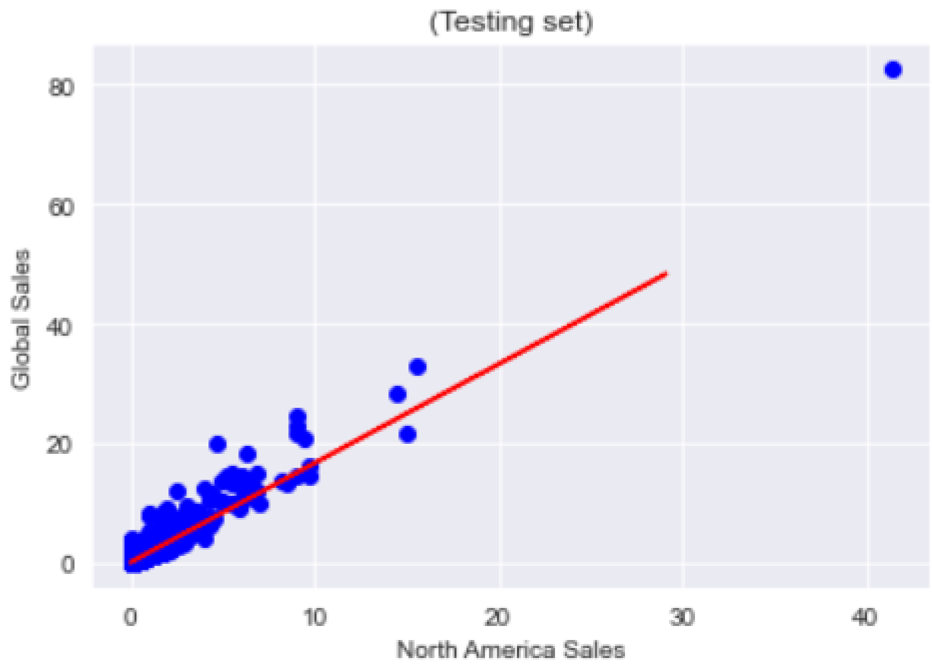
70:30


(Training set)
Global Sales
0
10
20
30
40
50
0
5
10
15
20
25
30
North America Sales


(Testing set)
Global Sales
0
20
40
60
80
0
10
20
30
40
North America Sales

50:50


(Training set)
Global Sales
North America Sales
0
10
20
30
40
50
0
5
10
15
20
25
30

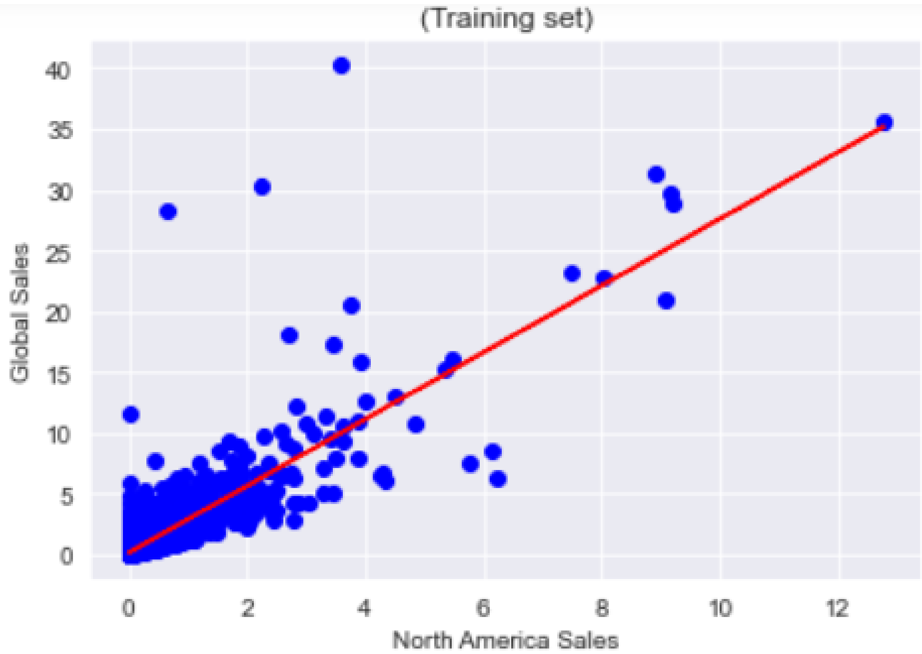

(Testing set)
Global Sales
North America Sales
0
20
40
60
80
0
10
20
30
40

4.1.2 Independent Variable: EU_Sales
Dependent Variable: Global_Sales

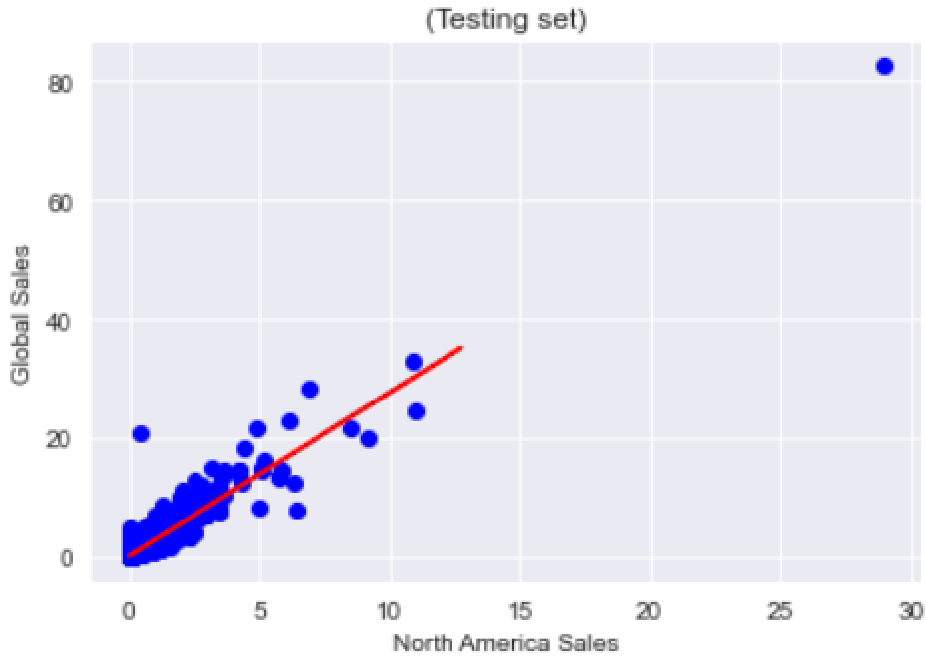
	Without Split	Train:Test 80:20	Train:Test 70:30	Train:Test 50:50
Coefficient	2.77191603	2.74859053	2.75628759	2.74528463
Intercept	0.13154605	0.13494423	0.13558062	0.13558035
R2 Score	0.812	0.752	0.746	0.727
MSE	0.44988652399731854	0.29945516691401924	0.27787095342730367	0.30947632867764374
RMSE	0.6707358078985485	0.5472249691982441	0.5271346634658962	0.5563059667823488
MAE	0.24532652449885314	0.2401978507261	0.23833325853770335	0.23946165826050744
R2 Adjusted	0.812	0.752	0.746	0.727
AIC	3.410e+04	2.836e+04	2.565e+04	1.932e+04
BIC	3.411e+04	2.837e+04	2.566e+04	1.934e+04

Graph for Without Split

50:50



(Training set)
Global Sales
North America Sales



(Testing set)
Global Sales
North America Sales

80:20

(Training set)
Global Sales
North America Sales
0
5
10
15
20
25
30
35
40
0
2
4
6
8
10
12

(Testing set)
Global Sales
North America Sales
0
20
40
60
80
0
5
10
15
20
25
30

70:30


(Training set)
Global Sales
North America Sales


(Testing set)
Global Sales
North America Sales

4.2 Multiple Regression

4.2.1 Independent Variable: NA_Sales, EU_Sales
Dependent Variable: Global_Sales

	Train:Test 80:20	Train:Test 70:30	Train:Test 50:50
Coefficient	1.14045928 1.37004663	1.17421485 1.3928549	1.16350508 1.44954647
Intercept	0.033562226774832116	0.022955536269016585	0.017677014780652867
R2 Score	0.968	0.962	0.958
MSE	0.0949326759565501	0.0970728832577 9118	0.09214129415614845
RMSE	0.30811146677225454	0.31156521509595897	0.30354784492094233
MAE	0.11569857408846232	0.11904647508893894	0.11896611438897403
R2 Adjusted	0.968	0.962	0.958
AIC	4560.	3889.	3058.
BIC	4582.	3911.	3079.

4.3 Polynomial Regression

4.3.1 Independent Variable: NA_Sales, Degree = 2
Dependent Variable: Global_Sales

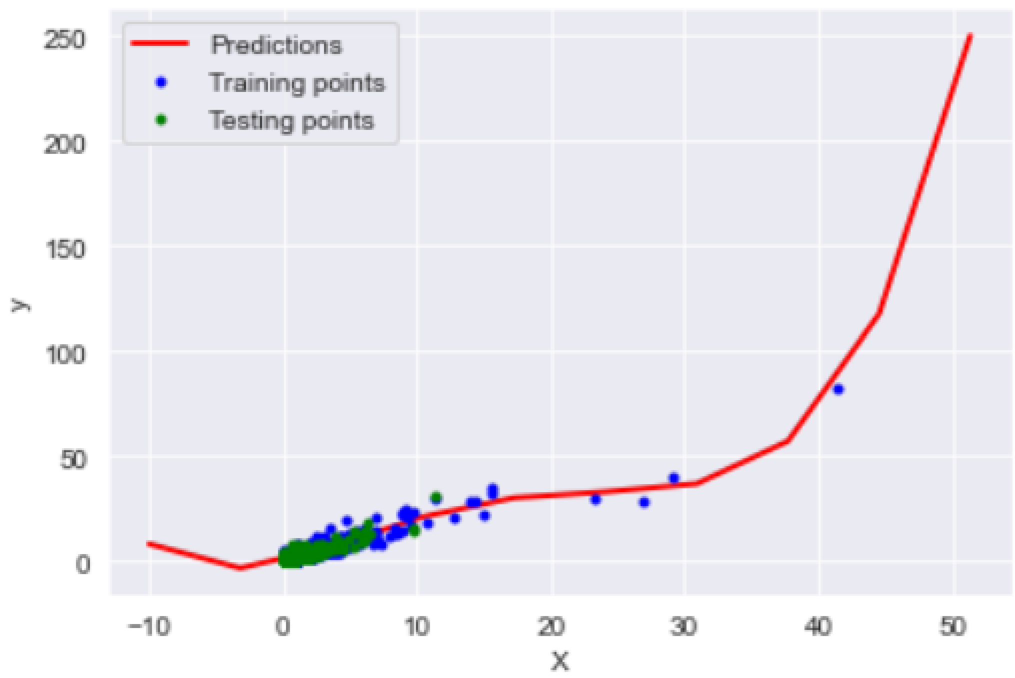
	Train:Test 80:20
Coefficient	1.8246
Intercept	0.055
R2 Score	0.891
MSE	0.246
RMSE	0.496
MAE	0.440
R2 Adjusted	0.891
AIC	2.092e+04
BIC	2.094e+04

4.3.2 Independent Variable: NA_Sales, Degree = 3
Dependent Variable: Global_Sales

	Train:Test 80:20
Coefficient	2.0165
Intercept	0.021
R2 Score	0.895
MSE	.258
RMSE	0.508
MAE	0.444
R2 Adjusted	0.895
AIC	2.032e+04
BIC	2.035e+04

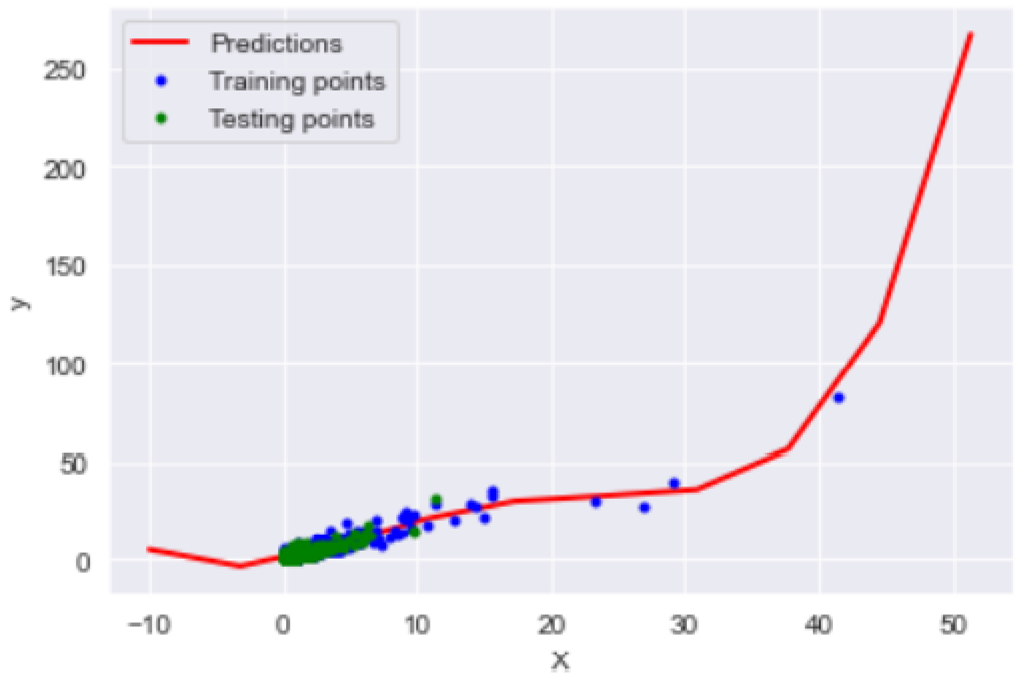
4.3.3 Independent Variable: NA_Sales, Degree = 4
Dependent Variable: Global_Sales

	Train:Test 80:20
Coefficient	1.5381
Intercept	0.089
R2 Score	.913
MSE	0.234
RMSE	0.483
MAE	0.443
R2 Adjusted	.913
AIC	1.781e+04
BIC	1.785e+04



4.3.4 Independent Variable: NA_Sales, Degree = 5
Dependent Variable: Global_Sales

	Train:Test 80:20
Coefficient	1.5564
Intercept	0.087
R2 Score	.913
MSE	0.233
RMSE	0.483
MAE	0.443
R2 Adjusted	.913
AIC	1.781e+04
BIC	1.786e+04



4.3.5 Independent Variable: EU_Sales, Degree = 2
Dependent Variable: Global_Sales

	Train:Test 80:20
Coefficient	2.7663
Intercept	0.136
R2 Score	0.839
MSE	0.284
RMSE	0.533
MAE	0.491
R2 Adjusted	949.083
AIC	2.846e+04
BIC	2.848e+04

4.3.6 Independent Variable: EU_Sales, Degree = 3
Dependent Variable: Global_Sales

	Train:Test 80:20
Coefficient	2.8309
Intercept	0.130
R2 Score	0.838
MSE	0.285
RMSE	0.534
MAE	0.491
R2 Adjusted	952.541
AIC	2.845e+04
BIC	2.848e+04

4.3.7 Independent Variable: EU_Sales, Degree = 4
Dependent Variable: Global_Sales

	Train:Test 80:20
Coefficient	2.7214
Intercept	0.137
R2 Score	0.836
MSE	0.288
RMSE	0.536
MAE	0.491
R2 Adjusted	961.815
AIC	2.843e+04
BIC	2.847e+04

4.3.8 Independent Variable: EU_Sales, Degree = 5
Dependent Variable: Global_Sales

	Train:Test 80:20
Coefficient	2.5113
Intercept	0.148
R2 Score	0.834
MSE	0.291
RMSE	0.539
MAE	0.492
R2 Adjusted	973.162
AIC	2.840e+04
BIC	2.844e+04

4.4 Ridge Regression

4.4.1 Independent Variable: NA_Sales
Dependent Variable: Global_Sales

	Train:Test 70:30	Train:Test 50:50
Coefficient	1.71303485	1.70422867
Intercept	0.084	0.086
R2 Score	0.887	0.895
MSE	0.2252666321959597	0.2398814557933692
RMSE	0.475	0.490
MAE	0.441	0.448
R2 Adjusted	0.887	0.895
AIC	1.896e+04	1.401e+04
BIC	1.897e+04	1.402e+04

4.4.2 Independent Variable: EU_Sales
Dependent Variable: Global_Sales

	Train:Test 70:30	Train:Test 50:50
Coefficient	2.61826398	2.64285984
Intercept	0.154	0.150
R2 Score	0.827	0.835
MSE	0.4494745811546051	0.4091031474324704
RMSE	0.670	0.640
MAE	0.496	0.499
R2 Adjusted	0.827	0.835
AIC	2.397e+04	1.782e+04
BIC	2.398e+04	1.784e+04

4.5 Lasso Regression

4.5.1 Independent Variable: NA_Sales
Dependent Variable: Global_Sales

	Train:Test 70:30	Train:Test 50:50
Coefficient	0.	0.
Intercept	0.538	0.540
R2 Score	0.887	0.895
MSE	1.8590657898994667	1.801
RMSE	1.363	1.342
MAE	0.762	0.765
R2 Adjusted	0.887	0.895
AIC	1.896e+04	1.401e+04
BIC	1.897e+04	1.402e+04

4.5.2 Independent Variable: EU_Sales
Dependent Variable: Global_Sales

	Train:Test 70:30	Train:Test 50:50
Coefficient	0.	0.
Intercept	0.538	0.540
R2 Score	0.827	0.835
MSE	1.8590657898994667	1.801
RMSE	1.363	1.342
MAE	0.762	0.765
R2 Adjusted	0.827	0.835
AIC	2.397e+04	1.782e+04
BIC	2.398e+04	1.784e+04

4.6 Gradient Descent(Train:Test = 70:30) (Iterations = 100)

4.6.1 Independent Variable: NA_Sales
Dependent Variable: Global_Sales

Learning Rate	0.1	0.001	0.5	0.05	1
Coefficient	1.58275299	1.87512098	1.01311339	1.73624256	0.77327016
Intercept	0.13617446	0.05189988	0.28272394]	0.06739965	0.36028936
R2 Score	0.8606589489724527	0.8349389769227513	0.7455325327708984	0.8546027268287029	0.6311049145975234
MSE	0.2872091655457477	0.3402230595545871	0.5245072315907924	0.29969222416658564	0.760364938193459
RMSE	5359189915889786	0.583286430113531	0.7242287149725509	0.5474415257966696	0.8719890699965562
MAE	0.2204977559868012	0.20027619336453004	0.31825289784301586	0.29969222416658564	0.387621536563743

4.6.2 Independent Variable: EU_Sales
Dependent Variable: Global_Sales

Learning Rate	0.1	0.001	0.5	0.05	1
Coefficient	2.07692312	2.76516476	0.7825022	2.41159935	0.39966067
Intercept	0.21177857	0.1272922	0.39341429	0.17025871	0.45644439
R2 Score	0.6649888547679437	0.7053909724902898	0.3462889441240369	0.6958540583743504	0.1913887350913719
MSE	0.690523508765569	0.6072468402479615	1.3474263720678195	0.6269042859534549	1.6667060061100394
RMSE	0.8309774417910326	0.7792604444266125	1.1607869623956928	0.7917728752321936	1.2910096847468029
MAE	0.2794946128895307	0.2503421657616041	0.4417994265961159	0.25997589249126574	0.5089629079448483

4.7 Gradient Descent(Train:Test = 70:30) (Iterations = 500)

4.7.1 Independent Variable: NA_Sales
Dependent Variable: Global_Sales

Learning Rate	0.1	0.001	0.5	0.05	1
Coefficient	1.7768936	1.91956011	0.89863659	1.62302559	0.68266089
Intercept	0.07670791	0.04222911	0.27306861	0.10146659	0.32778761
R2 Score	0.850086699407895	0.8259161732392442	0.6960962860134436	0.8608588218871382	0.5784661998456682
MSE	0.30900064015417233	0.35882082308308866	0.6264050074021377	0.28679718836731594	0.8688636270421035
RMSE	0.5558782601920786	0.5990165465853916	0.7914575209081898	0.5355344884947336	0.9321285464151946
MAE	0.20207686849257395	0.2012045816455104	0.32358642856881364	0.20352352553313574	0.8688636270421035

4.7.2 Independent Variable: EU_Sales
Dependent Variable: Global_Sales

Learning Rate	0.1	0.001	0.5	0.05	1
Coefficient	2.06497156	2.74037705	0.8706583	2.51735896	0.5504676
Intercept	0.22827036	0.13465723	0.39137537	0.20375051	0.4159554
R2 Score	0.663674023998731	0.7055164104453805	0.37822493116042166	0.7006736266740249	0.2553631296897173
MSE	0.6932336322020917	0.6069882880830132	1.281600055128481	0.6169702128326302	1.534842263491558
RMSE	0.8326065290412342	0.7790945309030305	1.132077760195156	0.7854745144386482	1.2388875104268176
MAE	0.28809738778024996	0.25265675254196773	0.4334779900395961	0.28004733996649583	0.4738713346308347

4.8 Gradient Descent(Train:Test = 70:30) (Iterations = 1000)

4.8.1 Independent Variable: NA_Sales
Dependent Variable: Global_Sales

Learning Rate	0.1	0.001	0.5	0.05	1
Coefficient	1.48020095	1.9500298	1.1732985	1.69832335	0.92632475
Intercept	0.11609683	0.01951202	0.21466791	0.07340948	0.291613
R2 Score	0.8562176338154301	0.819185245946646	.8003145864297885	0.8576471723564176	0.7089840598388555
MSE	0.2963635849416652	0.3726945810086984	0.4115907019520023	0.29341702634716077	0.5998407843046162
RMSE	0.5443928590105359	0.6104871669484121	0.6415533508228308	0.5416798190325728	0.7744938891331656
MAE	0.20794891301000595	0.19752738811308876	0.268719639610496	0.1943837153483185	0.33126859652796303

4.8.2 Independent Variable: EU_Sales
Dependent Variable: Global_Sales

Learning Rate	0.1	0.001	0.5	0.05	1
Coefficient	2.10822204	2.77871577	1.02573887	2.37532512	0.3603436
Intercept	0.2237972	0.12223024	0.38050151	0.19807471	0.43424344
R2 Score	0.66894813 35053206	0.70525785 04614286	0.43060258 081854097	0.69359710 18449272	0.17317662 436576997
MSE	0.68236266 0731631	0.60752123 07921913	1.17363947 25103552	0.63155631 48904302	1.70424472 9162249
RMSE	0.82605245 64043321	0.77943648 28465443	1.08334642 31308263	0.79470517 48229844	1.30546724 5534046
MAE	0.28539445 742634184	0.24870382 6019552	0.41584575 019144926	0.27335010 34108097	0.50072196 89037146

5.Conclusion

Hence, we can see that in this project we used various statistical learning techniques to make predictions from the given dataset which are better suited for machine learning models and for understanding the data in general and we have also presented variou model parameters along with their statistical scores to better choose which model to use.

6.Acknowledgement

Both of the group members thank everyone for their best wishes and helping us to make this project a success. We are very grateful to one's who dedicated their valuable time to guide us throughout this project. We would like to acknowledge IIIT-NR for providing the necessary resources and facilities to implement the project successfully.

7.References

https://www.kaggle.com/rush4ratio/video-game-sales-with-ratings