**The association between North American Sales vis-à-vis other areas**

Analysis of sales is vital for any organization. Sales determine the quantum of revenue generated and helps the organizations to implement strategies and tactics for obtaining optimum performance. Sales analysis also helps to identify the grey areas and subsequently provides platforms and avenues for correcting them. It is also very important to comprehend various variables that affect the sales. The identification of the variables that affect the sales is very important as we can make important decisions related to marketing, promotion, placing the right product at the right place and right time and so on.

With this backdrop, I am very much interested in sales analysis specially to understand what variables affect the sales. The dataset I will be using comes from Kaggle website that have taken into consideration various video games and their sales in different regions like North America, European Union, Japan, other areas and global sales. NA\_Sales represents the sales in North America, EU\_Sales denotes the sales in European Union, JP\_Sales represents the sales in Japan, Other\_Sales represents the sales in other areas and Global\_Sales represents the sales across the globe. Now, I am interested in analysing the relationship between the variables taken into consideration and also indulge in prediction.

I hypothesis that the best predictor of North America game sales will be Global sales.

The table below give the summary of all the independent variable and dependent variable.

**Table 1: Variable Summary Statistics**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **Frequency** | **Minimum** | **Maximum** | **Median** | **Mean** | **Std.Dev** |
| **North America Sales** | **16598** | **0** | **41.49** | **0.08** | **0.26** | **0.82** |
| **Japan Sales** | **16598** | **0** | **10.22** | **0** | **0.08** | **0.31** |
| **Europe Sales** | **16598** | **0** | **29.02** | **0.02** | **0.15** | **0.51** |
| **Other Sales** | **16598** | **0** | **10.57** | **0.01** | **0.05** | **0.19** |
| **Global Sales** | **16598** | **0.01** | **82.74** | **0.17** | **0.54** | **1.56** |

I chose to run a regression in order to understand the variation on the dependant variables by the independent variables. . NA\_Sales i.e. North America sales is our dependent variable and the independent variables will be JP\_Sales i.e. Japan Sales, EU\_sales i.e. European Union Sales, Other\_Sales and Global\_Sales. Thus, the multiple linear regression model is as follows:

Y = α + β1X1 + β2X2 + β3X3 + β4X4 + Є

Where,

Y = North America Sales i.e. NA\_sales

X1 = European Union Sales i.e. EU\_Sales

X2 = Japan Sales i.e. JP\_Sales

X3 = Other Sales i.e. Other\_Sales

X4 = Global sales i.e. Global\_Sales

β1, β2, β3 and β4 are the slope coefficients of EU\_Sales, JP\_Sales, Other\_Sales and Global\_Sales.

From the multiple regression model I retrieve the following equation:

Y = α + β1X1 + β2X2 + β3X3 + β4X4 + Є

North America Sales = -3.163e-04 - 9.998e-01\*Japan Sales -9.998e-01\*Europe Sales -9.995e-01\* Other Sales + 1.000e+00\* Global Sales + 0.005222

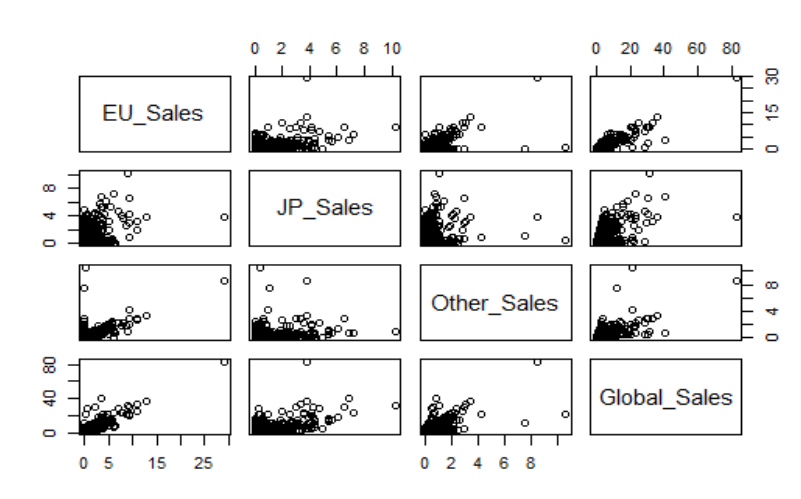
The equation shows the North America Sales is predicted to decrease 9.998e-01 when the Japan Sales variable goes up by one. The sale of North America decrease by the same rate as the Japan sale rate when Europe Sales goes up by one unit. North America sales also decreases by 9.995e-01 when other sales variable goes up by one. However, the North America sales increases by 1.000e+00 when Global sales goes up by one unit bearing all the other variable stays the same. The sale of North America decreases by -3.163e-04 when all the independent variable is 0, and is predicted to be 0.005222 when both variables are zero. The intercept in this case is not statistically significant as the sales of game in those regions zero will show that gaming is not popular in those regions and therefore would not make sense to include in this research.

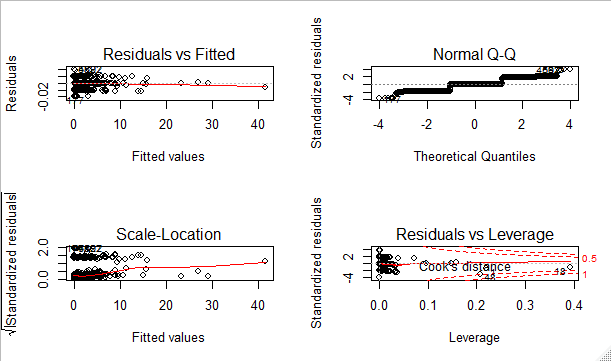
The model is a perfect fit with Adjusted . Looking at the *p* < .05 it can be stated that all the variables **i.e.** EU\_Sales, JP\_Sales, Other\_Sales and Global\_Sales all are less than 0.05. Thus at 5% level of significance, all the independent variables have statistically significant impact on North American Sales. Therefore, I can reject the null hypothesis and say that all the variables are a good predictor of the North America Sales. Now, looking at the t(16593) = 12424.04, it can be stated that the Global sales is the best predictor of North America Sales.

I have wanted to see if my model violates any assumptions or not. And by looking at the plots, I have noticed that my model clearly violates Residual Vs. Leverage as the cook distance travels further than leverage of 0.05. This means that there are some high leverage points present in the data.

Some potential weaknesses and gaps in my study was that I only took the sales of different region to predict the best variable and ignore other important variable such as platform and publisher which could prove to be better predictor than the independent variables I chose for this study. For future analysis, I could run the multiple regression test on all the variable present in the data to make sure I get the absolute best predictor of the North America Sales. Since, my model violates the residual vs leverage, this means that I could have recoded the variable to make sure that there are no high leverage points.

**Figure 1. Scatterplots between various independent variables**

****



**Table 1. Regression output**

======================================================

Dependent variable:

----------------------------------

North America Sales

------------------------------------------------------

Europe Sales -1.000\*\*\*

(0.0002)

JP\_Sales -1.000\*\*\*

(0.0002)

Other\_Sales -0.999\*\*\*

(0.0003)

Global\_Sales 1.000\*\*\*

(0.0001)

Constant -0.0003\*\*\*

(0.00004)

------------------------------------------------------

Observations 16,598

R2 1.000

Adjusted R2 1.000

Residual Std. Error 0.005 (df = 16593)

F Statistic 101,472,202.000\*\*\* (df = 4; 16593)

======================================================

Note: \*p<0.1; \*\*p<0.05; \*\*\*p<0.01