**Final Project Summary Report**

**Group Name: SAS R US**

**CSC 423**

**Abstract**

This project was designed to address the process of regression analysis as a tool for understanding and making predictions from diverse datasets.  The purpose of our analysis was to clean, explore, analyze, and predict from a large dataset trends and influences among the variables. The dataset that we chose as a team was a popular Video Game Sales dataset found on the data science projects website Kaggle.com.  Our analysis process was implemented beginning with a project proposal, which provided us with an understanding of our dataset through research, by defining independent and dependent variables, and by describing the methodology that we would follow throughout our analyses.  We then individually chose how we would preprocess, clean, explore, and analyze the data so that we could build our unique regression models which we utilized in order to describe which predictor variables have the most significant effects on the predicted variable of global video game sales, and also with the purpose of predicting and forecasting global video game sales given various possible scenarios.

Although our models differed and as such so did the variables that we found had the greatest effect, however we agreed that North American sales has the biggest impact on the global sales.

Some of us discovered more specific impacts such as genre or rating or platform, for instance one of us discovered, if the rating for video game is e(Everyone) that also has significant effect on global sales rather than the other variables. We also found that the global sales of video games on different platform may vary. Some of games launched on specific platforms like PC or PS2 may help increase the sales of video games while the games launched on other platforms such as XB or PS4 may lead to decrease of sales, while some of us examined the influence that user or critic scores have on overall sales.

We also could look at this from the point of view of the Developer or Publisher. In learning which games promote global sales, we can prioritize what type of game to develop based on our available developers and resources, and expected sales. We can use the data to determine which regions we want to target, and to decide to spend more money on marketing in a certain region to try to improve sales. We might even determine something like we should release two different types of game at the same time, because a player would not play both types, or we should release our Shooter at the same time as the competition because we dominate them. Or they could decide which platform to port to, based on previous successes of their games on other platforms.

**Introduction**

The data set that we as a team chose to analyze contains sales information of global and regional video games that were released from 1980 to 2016. There are total of 16719 samples and 16 different variables. In this dataset, we chose global\_sales as the dependent variable.

This data set includes 14 independent variables, with 9 of those variables numeric, 6 text variables, and one year variable.  The description of our variables is the following:

1. Name - Name of the game (Text).

2. Platform - Console on which the game runs, i.e. Wii, Ps3, etc... (Text).

3. Year\_of\_Release - Year the game was released (Number in format YYYY).

4. Genre - Category of the game (Text).

5. Publisher - Publisher i.e. Nintendo, etc.…(Text).

6. NA\_Sales - North America sales in millions (Numeric).

7. EU\_Sales - European sales in millions (Numeric).

8. JP\_Sales - Sales in Japan in millions (Numeric).

9. Other\_Sales - Game sales in the rest of the world - Africa, Asia excluding Japan, Australia, Europe excluding the E.U. and South American millions (Numeric).

10.   Global\_Sales - Total world sales in millions (Numeric).

11.   Critic\_Score - Aggregate score compiled from Metacritic website’s staff of reviewers (Numeric).

12.   Critic\_Count - Number of critics scoring the game (Numeric).

13.   User\_Score - Average score from subscribers of Metacritic website (Numeric).

14.   User\_Count - Number of users who scored the game (Numeric).

15.   Developer - Creators of the game (Text).

16.   Rating - ESRB ratings, i.e. Teen, Everyone, etc.…(Text).

Our dataset was found on the Kaggle website:

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

As we know, the gaming industry over the years has been becoming more prevalent. Video games have become one of the major sources of entertainment thanks to greater revolution in computer sciences and with such popularity, more data is becoming available to us to research this growing industry.

Our goal with this project is to investigate the sales of games by looking at platform, genre, regional sales, user and critic scores(as found via Metacritic website used in our dataset: <http://www.metacritic.com/>), as well as each of the additional factors described in the above independent variables.  Each of us individually chose our datasets from our original ~16,000 point data set and each of us, explored the data, and performed regression analysis on our own unique models that we developed through our data exploration and analysis processes. Our research intention is to find accurate models that can predict the factors which influence global sales of video games given multiple factors.

**Kristen Groom Section:**

Methodology

I began this project by exploring the dataset that we as a team chose:

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

and some of the references we found as a group including:

<https://towardsdatascience.com/predicting-hit-video-games-with-ml-1341bd9b86b0>.

<http://sdsu-dspace.calstate.edu/bitstream/handle/10211.10/1073/Ehrenfeld_Steven.pdf>.

[*http://fortune.com/2018/07/05/2018-best-selling-pc-games-so-far/*](http://fortune.com/2018/07/05/2018-best-selling-pc-games-so-far/)

I next preprocessed the data using python and python libraries.  I randomly chose 2000 data points to use in my analysis.

I recoded some of the variables and created dummy variables.  I then explored the data via scatterplots, boxplots, and correlation tables. I checked for multicollinearity and did not find any.

In my analysis phase, I performed linear regression on all of my variables using Proc Reg and looking for outliers, influential points, standardized estimates, and multicollinearity.  I looked at Adj R2, Goodness of Fit measures and I removed outliers and influential points which was incredibly time consuming giving the slowness of SAS. I looked at the significance of the variables and did notice multicollinearity between two of my Year\_of\_Release variables.  I removed one at a time and decided on which one to keep based on AdjR2 values. I reviewed the residual plots and noted any linear, constant variation, and independence violations.

Next, I split the data into training and testing sets and performed model selection on the training set using Adj R2 and selection methods.  I chose the best model and then once again checked Adj R2, Goodness of Fit, residuals, outliers and influential points and adjusted my model by fixing these issues.  Once I was satisfied with my final model, I continued with validation of the model by predicting test values, and then testing the goodness of the prediction results.

I then performed two predictions on my final model by finding an expected Global Sales given certain parameters, and I included Confidence Intervals and Prediction Intervals.

Analysis, model, results

***1) Data preprocessing: (Python)***

My data exploration first began with research into what kind of models and data analyses had been previously done on the data set that we had chosen, Video Game Sales found on the Kaggle website.  I reviewed kernals from this dataset, which are different analyses that are linked to the dataset and where people can post their work, and many of the kernals that I looked at were focused on data visualizations. In the report by *Chavarria, 2017,* I reviewed the author’s data visualizations, which will not be as relevant for my analysis but found them interesting.  I noted how he coded his variables and how he did not transform Year\_of\_Release which proved itself a significant predictor in his analysis.  I decided, however, to recode Year\_of\_Release into decades as stated below for a bit easier coding for dummy variables. The author also performed logistic regression to discover if a game would be a hit or not.  I will not be performing logistic regression, although this will be a process that I would like to practice and further explore. He used other machine learning models that we have not learned in this class, including classification such as Random Forest Classification.

I also reviewed the *Ehrenfeld, 2011,* aMaster’s Thesis*,* which provided a detailed explanation of the research process that the author performed to gather data into video game sales using internet language and ‘consumer interest’ (*pg. iv, Ehrenfeld, 2011*), ie message boards/reviews, in order to predict popularity.  He also used more complex analyses techniques than I will be using in this project.

Prior to beginning my own analysis, I also read an article about popular video games this year which gave me more information regarding popularity of PC games and years of release.  I learned about a term ‘A [long-tail game](http://fortune.com/2018/06/21/steam-how-much-youve-spent-on-computer-games/), an industry term for a title that continues to sell long after its initial release’(*pp.5, Morris, 2018).* I was curious to find out if this information would be useful to me as I began my analysis.

I began my initial data preprocessing by importing my dataset into a Jupyter notebook, in order to utilize python libraries such as pandas and numpy. The dataset that we started with had over 16000 data points, however, there were only approximately 6900 complete records, therefore, for my analysis I chose to remove all null values by removing the rows, which gave me my full dataset to work with.  I looked at the variables and their datatypes, I transformed User\_Count to numeric so that I could work with it in SAS. I looked at correlations among the variables and I also looked at the value counts of the categorical variables so I could decide if I wanted to recode any of them. I decided to recode Year\_of\_Release to simplify the categorization of dummy variables. I then decided to take a random sample of these rows by randomly ordering the data and then choosing the top 2000 rows.  This gave me my final dataset to import into SAS. (see example in appendix A1)

***2) Data Exploration (SAS)***

My first step in working with my video game sales dataset was to import the dataset into SAS.  I printed out the dataset and after looking at the frequencies (A2) of all of the categorical variables, I decided to recode the Year\_of\_Release to account for the decades that represent my data.  I then created dummy variables for all of the categorical data: Year\_of\_Release, platform, genre, and rating.(A3)

I next performed data exploration via scatterplots and gplots for Global\_Sales against the numeric x variables: *NA\_Sales EU\_Sales JP\_Sales Other\_Sales Critic\_Score Critic\_Count User\_Score User\_Count*. (A4)

My impression of the scatterplot matrix and gplots was that there seems to be a linear relationship between Global\_Sales and the other sales region x variables, and perhaps there will be multicollinearity issues.  I am also seeing a slight half-u shape with User\_Score and Critic\_Score which could mean that there may be a need to transform some of the x variables to see if the u-shapes will become more linear. I will further explore these possibilities throughout my analysis to decide whether or not I will need to perform a more complex polynomial regression analysis, such as quadratic regression to my data.

I also created gplots for Global\_Sales vs NA\_Sales against each of the qualitative variables.  These were difficult to read for the categorical variables which have many different values such as for the variable Platform, however, in doing these scatterplots, I can see that there are likely some outliers, some very popular games, years, and genres. *(A5)*

I created boxplots for all the categorical variables against Global\_Sales and due to the way that they are are situated near the bottom of the graph tells me that there is likely a need for a transformation of the y-variable.  I will move forward with the full model and look at the residuals before I decide to perform any transformations.

(A6)

I then checked the Pearson correlation table to check for multicollinearity. (A7)

I did not see obvious multicollinearity (r >= .9) among the x variables although I do see fairly high correlations among the various different region’s sales.  I’m guessing these variables would be excellent candidates for creating interaction variables. Given time constraints, I will not be creating interaction variable at this time, however, I will discuss this further in future work.  I will also check for multicollinearity further when fitting the models.

***3)   Data Analysis phase:***

I began my data analysis phase by fitting my full model, including vif, stb, r, and influence in order to look for multicollinearity, significance of the predictors, outliers, and influential points.  When I viewed the output, I saw very strange residuals, and model indicators. (A8) It is possible that I coded my dummy variables incorrectly especially my Year\_of\_Release variable that I recoded to represent 3 decades rather than individual years.  I have reviewed this coding and I do not see any obvious mistakes so I decided to move forward with a transformation of the y-variable and see what my residuals and values look like before I continue with my analysis. (A9)

After transforming Global\_Sales to lnsales (log(Global\_Sales), right away the output made more sense to me.  I do, however, see a clear pattern in the residual plots against the predicted value. They are clearly u-shaped.  I also see a slight s-shape to the normal probability plot. This could be an indicator of outliers and influential points so I will address these next. (A10)

The shapes that show clear patterns tell me that there are indicators of constant variation and independence violations.  I also recognized that the initial scatter plots appeared to violate linearity which means that I may be using the wrong model for the data.   I had considered further transformations of x variables, however, for the other regions of sales data, there are 0 values so log transformation would not work and for this project, I did not want to introduce too much complexity by performing different types of transformations on the y vs the x variables.

Looking at the Analysis of Variance table I can look for goodness of fit. (A9) I noticed that the F value is not very high but the p-value associated with it is very low, which indicates that there is at least one significant predictor of Global\_sales.  This means that we can reject the null hypothesis that states that there are 0 significant predictors, and we can accept the alternate hypothesis which tells us that there are 1 or more significant predictors. I can see that the Adj R2 is moderate at 0.5877, which tells me that the full model accounts for approximately 59% of the variability in Global Sales.  RMSE is very low 0.90293, which is good because we want to minimize error statistics.

The Parameter Estimates table tells me that there are significant variables and others that are not significant in that their p-values are > alpha = 0.05.  These insignificant variables are: EU\_Sales, pl2, pl8, pl9, pl13, r2, r3(although it is on the threshold of 0.0577), r4, r5, r6, g3, g5, g7, g9, g11, y2, and y3.  I also see very high indications of multicollinearity for y2 and y3, which could have contributed to the output issues that I noticed prior to the transformation but given that the output makes more sense after transformation of the y-variable, it seems that this was a positive choice in moving forward with a model.    I will remove y2 and then y3 and see which model has higher Adj R2 and if collinearity issues are still a problem.

As far as normal probability looks, the normal probability plots do look pretty linear and the line looks slightly pulled downward which may be an indicator of influential points and outliers. I do not see violations of normality. I will move forward with the analysis as is to see if removing influential points and outliers will improve the model.

After removing y3, I noticed that Adj R2 stayed almost the same, and vif for y2 decreased to approximately 2.02.  The standard estimates table shows that y2 is significant at p-value = 0.0002. (A11)

I checked the full model with y2 removed as well.  Although removing either y2 or y3 did not affect the Adj R2 significantly, removing y3 allowed it to be very slightly higher, so I removed y3 due to the multicollinearity issues.

***Checking for outliers, influential points, and viewing most significant variables:***

I started by removing only points with both red and blue arrows, indicating that they are both outliers and influential points and with each subsequent removal, I would check residual plots, normal probability plots, Goodness of fit, and Adj R2.  See example output of both outliers and influential points: (A12)

I continued to remove influential points and outliers until I noticed no further significant improvement in the model by reviewing Adj R2.

After outliers and many influential points were removed, and I did not see any further significant improvement in Adj R2, I reviewed the residuals and there is a very clear u-shaped pattern and possibly a need for transformation of some of the x-variables.  My final linear model had Adj R2 of 0.7939 which tells us that approximately 79% of the variability of Global Sales can be accounted for given this model. The RMSE 0.53928 which is very low. The F-Value is 171.22 which is fairly high and the p-value associated with this F-value is very low at <.0001.  This tells us that we can reject the null hypothesis that there are 0 significant predictors in this model and we can accept the alternate hypothesis that there is at least one significant predictor. I can see from the parameter estimates table that there are some insignificant predictors. Which I will address in model selection step.

The residual plots and normal probability plots are all showing the need for a quadratic model given the u\_shape of these plots.  I would like to move forward with a quadratic model, however, given time constraints in that this is an accelerated semester, I will move forward with the linear model knowing that this model is not good for my data.  I did initially try the proc glmselect command for quadratic equation although with so many variables there were hundreds of combinations for me to recode and I did not have time to do so. (A13)

I then split the full model into training and test sets creating a new\_y variable for lnsales.  This allows me to have the training and test set in the same dataset, only the train set (the rows that have a 1 for criteria ‘Selected’) will have the new\_y values and the test set will have null values represented by a period.  (A14)

***4)   Model Selection using train set:***

I performed the model selection method using adjrsq and stepwise

·      *Adj R2 method: (A15)*

·      *Stepwise method: (A16)*

I chose to use the stepwise selection model because it has much fewer variables and an almost equally high R2 compared with AdjR2.I noticed some insignificant variables and I did not want to overfit the model, I removed one at a time starting with the highest p-value variable, which was r6, and then I’ll recheck for significance.

I ended up removing 4 insignificant variables for this first draft of a final fitted model.

I then performed a stepwise selection with the remaining variables which gave me a final model with 14 variables, so three variables were removed.  I decided to remove the final two variables, one at a time, although when I removed one, then three more became insignificant. I was concerned that I would be underfitting the model, I did, however, decide to continue to remove insignificant variables so that the model has the least number of significant variables.  There are still some influential points, but I decided to not remove them because I did not want to remove too many points from the training set.

My final model before removing additional insignificant variables has an Adj R2 value of 0.8464, which tells me that this model accounts for approximately 85% of the variability of Global Sales of video games.  Looking for goodness of fit, I see that the F-Value is high at 571.38 with a very low p-value associated with it at alpha < .0001. This affirms to us that we can reject the null hypothesis that states that there are no significant predictors in our model.  We can accept the alternate hypothesis by saying that we have at least 1 significant predictor for Global Sales. I can see from the parameter estimates table that all of the remaining variables are significant. I also see that RMSE is very low at 0.41363, which tells me that there is low prediction error.

I can also see that there are no issues with multicollinearity from the vif values andlooking at Standardized Estimates I can see that North American sales is the highest predictor for Global Sales.

I continued to remove some influential points and insignificant variables one at a time, although there were some influential points remaining at this point. I decided to leave those influential points. The final model has all significant variables.

The residual plots definitely show a clear pattern and the normal probability plot has a somewhat u or s-shape.  Although I left some influential points, I believe this skew is based on using the incorrect model – linear vs quadratic.(A17)

***5)  Final model equation:***

Global\_Sales = -2.49907 + 2.48296 (NA\_Sales) + 2.09582(EU\_Sales) – 1.23200(pl7) + 2.50759(JP\_Sales) + 0.00424 (Critic\_Score) + 0.00073605 (User\_Count) – 0.19113 (pl10) + 2.06274(Other\_Sales) + 0.00159(Critic\_Count) – 0.02187 (User\_Score) – 0.58813(g6) + 0.11326(y3)

        where pl7=1 when Platform=’PC’, otherwise 0,

        and pl10=1 when Platform=’PS4’ otherwise 0,

        and g12=1 when Genre=’Strategy’, otherwise 0,

        and g6=1 when Genre=’Puzzle’, otherwise 0,

        and y3=1 when Year\_of\_Release=2010 (2010-2020), otherwise 0.

-Interpreting this model equation:

This model equation can be interpreted the following way:

It tells us that if we see the following increases in the variables, given all other factors are the same, we can expect to see the following increases or decreases in global video game sales revenue:

North American Sales increases by $1 million then we can expect Global Sales to increase by approximately 1098%.

European Sales increases by $1 million, then we can expect Global Sales to increase by approximately 713%.

When we go to PC sales from PS2 sales, we can expect decreases in Global Sales by approximately 233%.

Japanese sales increases by $1 million then we can expect Global Sales to increase by approximately 1128%.

Critic’s aggregate score increases by 1 point then we can expect to see Global Sales increase by approximately 0.4%.

Number of users who scored the game increases by 1 user, we can expect Global Sales to increase by approximately 0.07%

When we go to PS4 from PS2, we can expect decreases in Global Sales by approximately 21%.

Sales from other parts of the world, including Africa and Asia, increases by $1 million, we can expect to see approximately 687% increase in Global Sales.

Number of critics scoring the game increases by 1 critic, we can expect to see Global Sales increase by approximately 0.16%

Average score of users increases by 1point, we can interestingly expect to see approximately a 2.2% decrease in Global Sales.

When the genre of the game goes from Action to Puzzle, we can expect to see approximately an 80% decrease in Global Sales.

When we look at the decade from 2010 – present, vs 2000-2009, we can expect to see an approximate 12% increase in Global Sales.

***6)  Validation Process:***

After fitting the final model, I continued with validation steps to test the goodness of this model.  I did this by using the Proc Reg procedure on the final fitted model where the output was written to a new dataset in the column where new\_y = ‘.’ (A18). Next I computed the performance statistics – RMSE and MAE for the test set. (A19)

In order to check the goodness of the model I will look at MAE, RMSE, AdjR2, R2, and CV-R2 of the test set.

The model training set statistics:

RMSE: 0.41363

Adj R2: 0.8464

R2: 0.8479

F-Value: 571.38

p-value associated with F-Value: <0.0001

Test model:

RMSE: 0.62593

MAE: 0.46931

R = yhat ^2 = 0.87438\*\*2 = 0.7645403844

CV-R2 = |train R2 – test R2| ~ 0.0833596, which is < 0.3

The model is a good model, because I can see that both RMSE and MAE are very low, almost as low as the training model.  I can see that R2 is almost as high as the training set at approximately 0.76 and the Cross Validation statistic is low, much lower than 0.3.  It would be a good model if it were appropriate for our data, but given the violations in linearity, independence and constant variation, these results are not useful overall to us in a real world setting. Although it has been a useful exercise as far as practicing the process of linear regression, even with a model that is flawed.

***7)  Using the model for two predictions:***

1) Predicting Global Sales on a Teen rated role-playing game

for the XBox, released in 2015 that totaled 2million in North American Sales: (A20)

Predicted value: 2.5801, CI(2.2450, 2.9153), PI(1.7021, 3.4581)

What this means is that we can expect Global Sales to increase by approximately 1220% for a T-rated role-playing game for the Xbox that grossed $2million in North American Sales, with a Confidence Interval of (844%, 1745%), and Prediction Interval of (449%, 3076%).

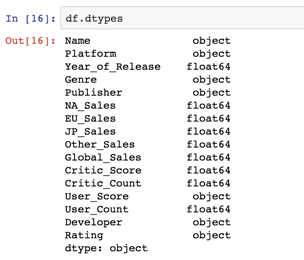
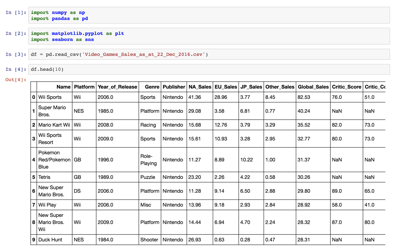
2) Predicting Global Sales for a E10 rated game that grossed .5 million in Japan, had 8.5 of user scores, 50 critics scored the game, and 100 users who scored the game.  (A21)

Predicted value: -1.2781, CI(-1.5030, -1.0533), PI(-2.1202, -0.4361)

What this means is that we can expect Global Sales to decrease by approximately 259% for a E10-rated game that grossed $0.5 million in Japanese Sales, with an average user score of 8.5, 50 critic scores, and 100 user scores, with a Confidence Interval of approximately(350%, 187%), and Prediction Interval of approximately(733%, 55%).

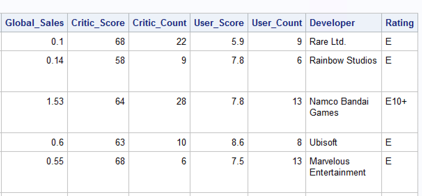
Appendix

(A1)

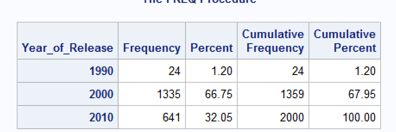


(A2)

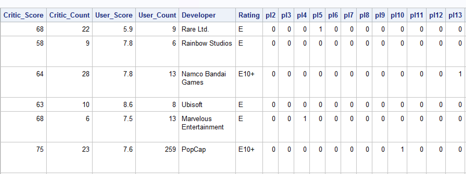


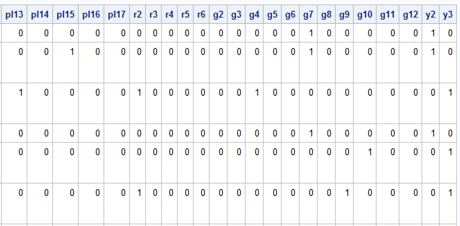


(After recoding dummy variables)

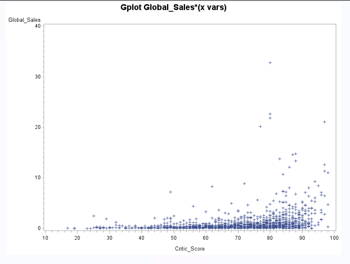
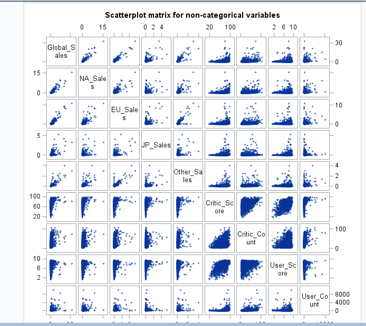


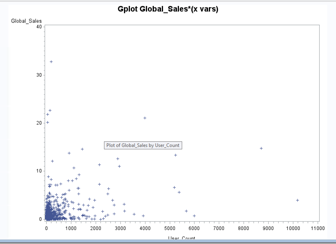
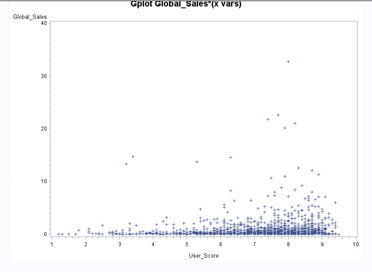
(A3)

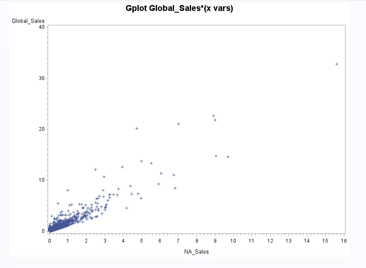
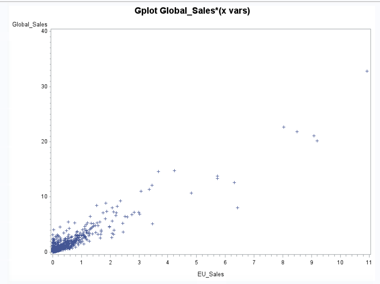


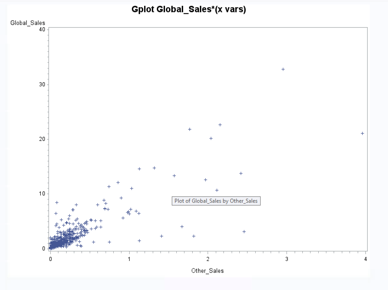
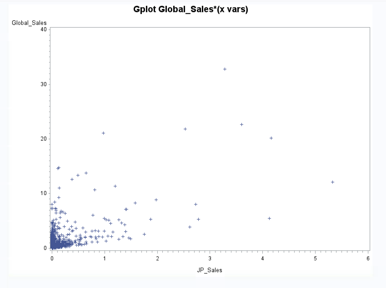
**

(A4)

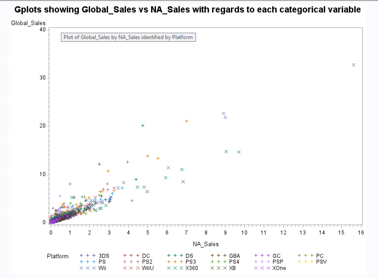
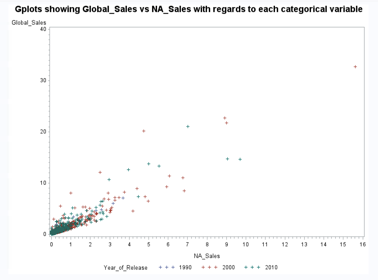
**

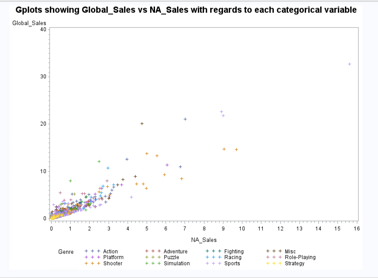
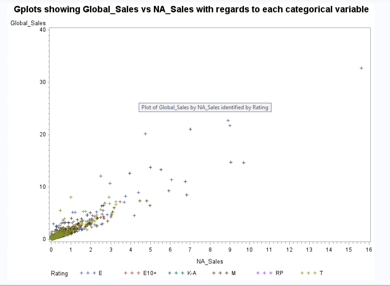
**

* *

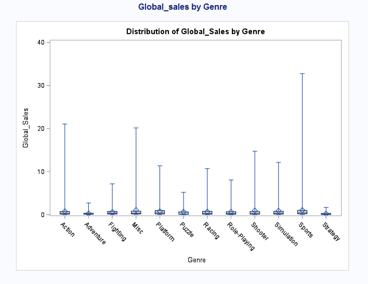
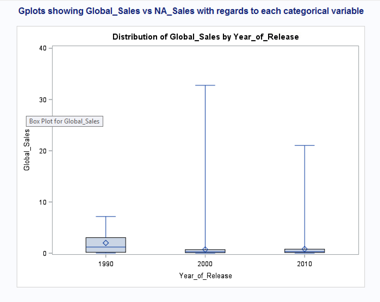
**

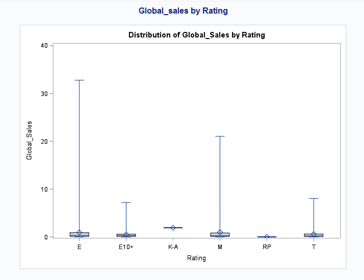
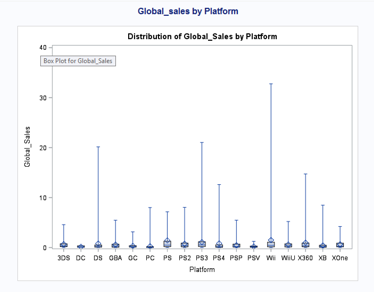
*(A5)*

**

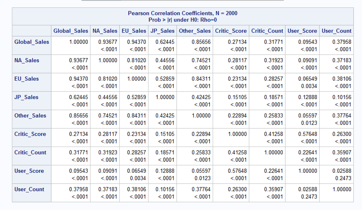
**

(A6)

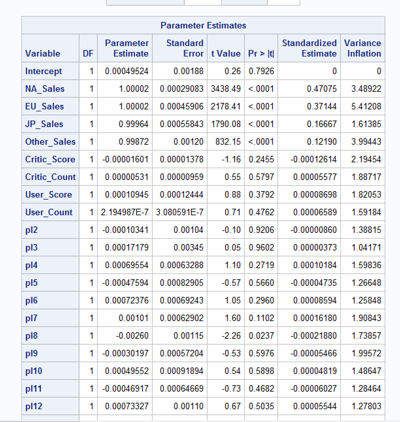
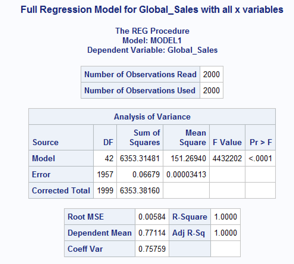
**

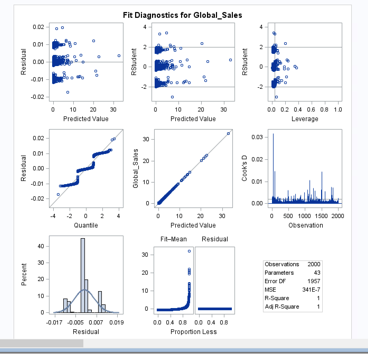
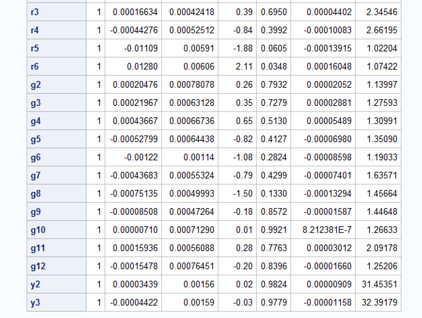
**

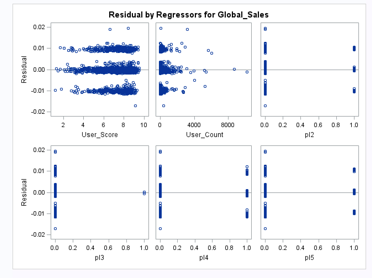
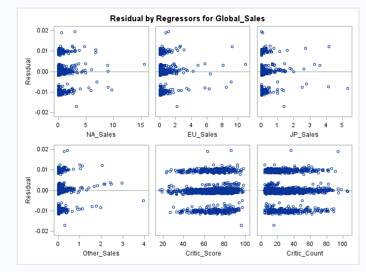
*(A7)*

**

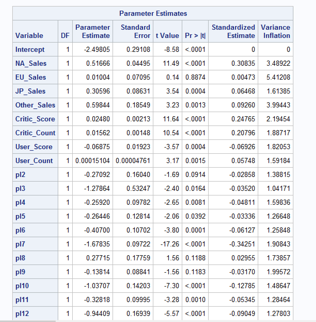
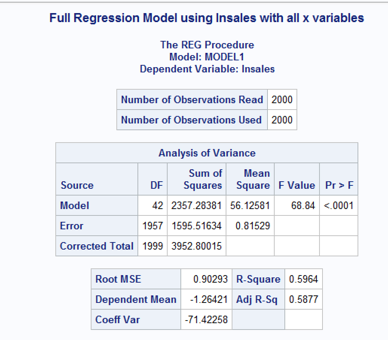
(A8)





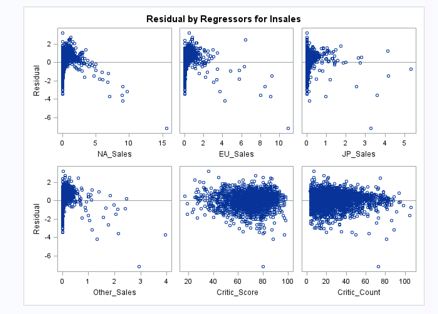
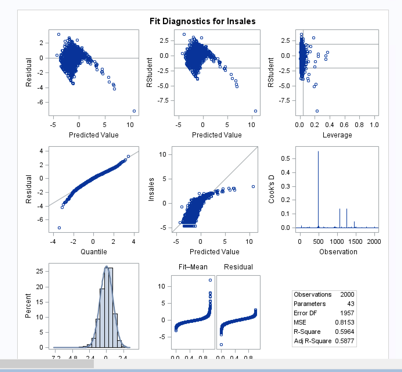
**

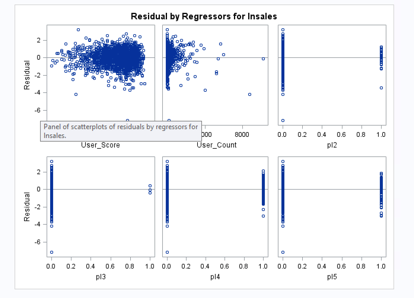
*(A9)*

**

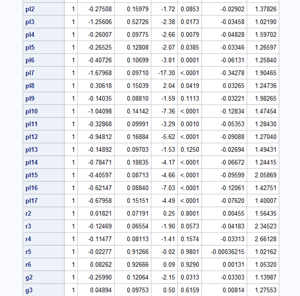
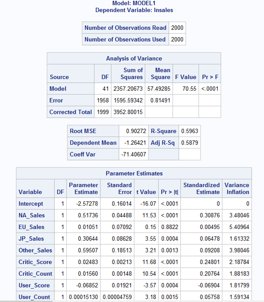
**

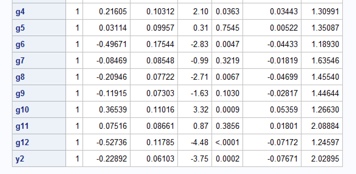
*(A10)*

**

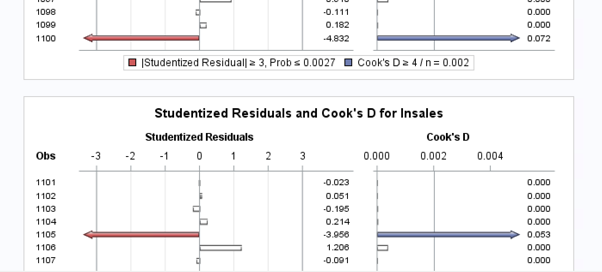


(A11)

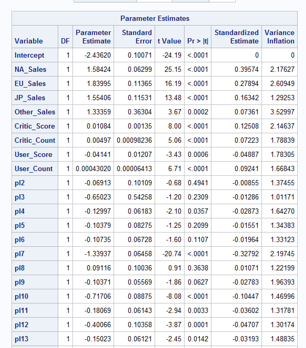
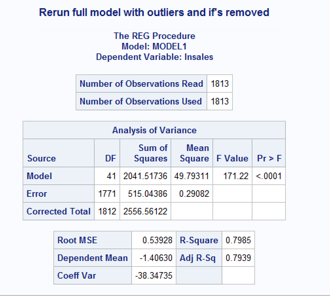


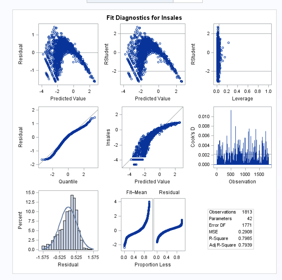
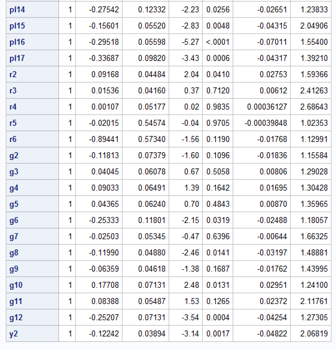


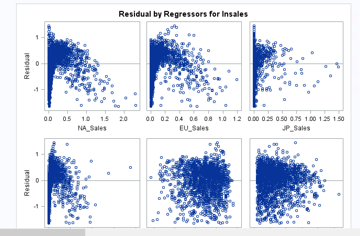
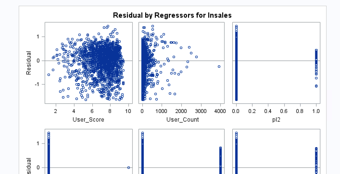
(A12)



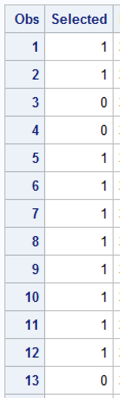
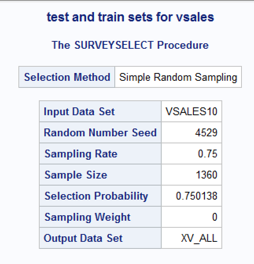
(A13)



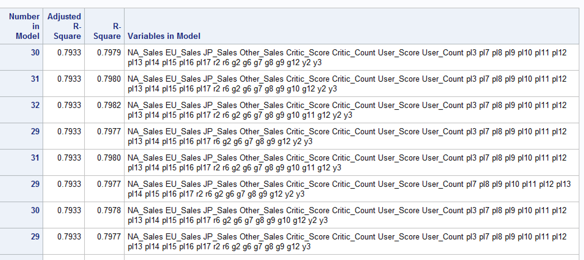
**

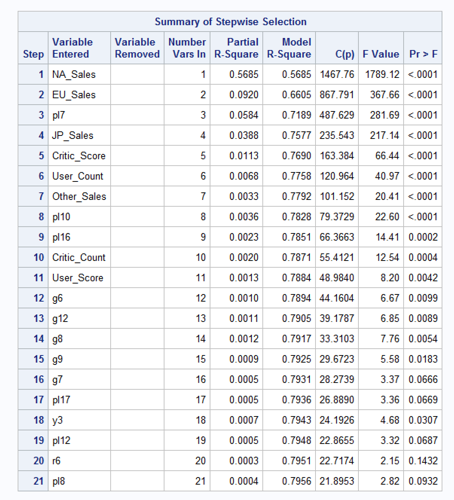
**

(A14)

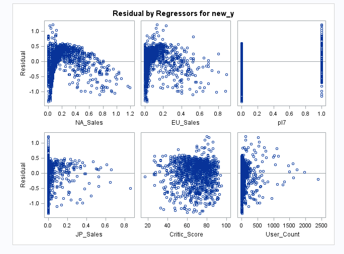
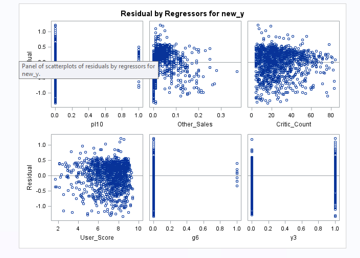
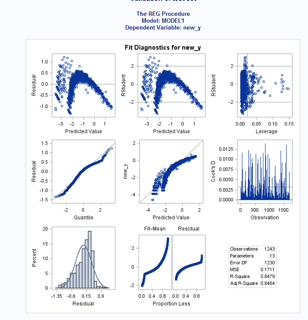
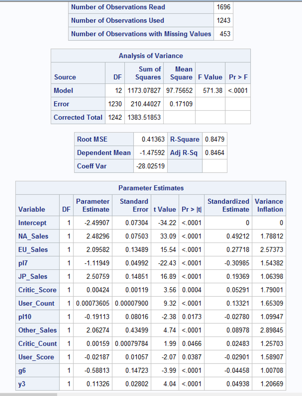


*(A15)*

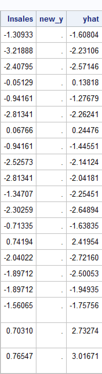
**

*(A16)*

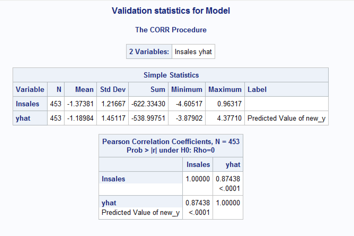
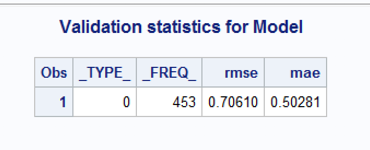
(A17)



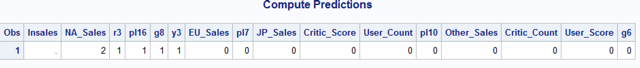
(A18)

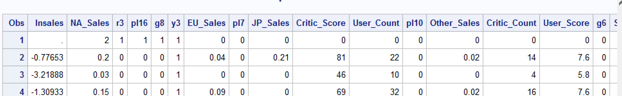


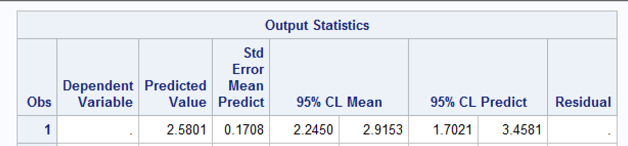
*(A19)*

**

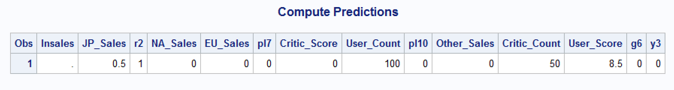
(A20)

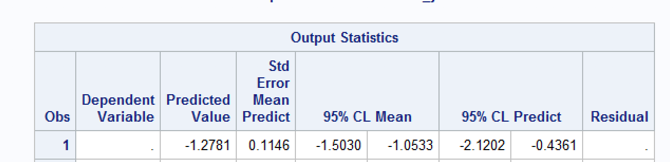






(A21)



**

Future work

I would have liked to perform Quadratic regression for this data because I could clearly see that linear regression was not the correct model for my data set.  And after we learned the techniques for polynomial regression, I began to try to implement a quadratic model. I was a half-day away from the due date and when I performed the Proc glmselect method, I realized that this would give me hundreds of combination of variables that I would need to recode, and knowing that SAS was extremely slow, I decided that it would not be possible for me to perform quadratic regression for this report.  I would like to continue to look into using other analysis platforms that are faster to continue to explore this data set, and I would also like to try other complex models that I read about in the articles and reports that I reviewed prior to beginning my own analysis.

It also would have been beneficial to create interaction terms, possibly for the variables for other regions of the world because they were highly correlated although not multicollinear.  I would look into the interaction term possibilities among the number of critics who scored the game, number of users who scored, and aggregate critic scores, although I did see a decrease in global sales with regard to user scores.  This could be an issue with the sample that I randomly chose, or it could also be that I used a less useful model so it’s difficult to accurately interpret the results.

I would also like to explore various visualization tools as I was inspired by the reports that I viewed and the amount of detail and interpretation that can come from well-executed visualizations.

References

Morris, Chris. 2018. 4 of 2018’s Best-Selling PC Games Are at Least 5 Years Old. Fortune.<http://fortune.com/2018/07/05/2018-best-selling-pc-games-so-far/> Accessed August 10, 2018.

Chavarria, Ignacio. 2017. Predicting Video Game Hits with Machine Learning. *Towards Data Science (Medium).* <https://towardsdatascience.com/predicting-hit-video-games-with-ml-1341bd9b86b0>. Accessed August 2, 2018.

Ehrenfeld, Steven E. 2011. ‘Predicting Video Game Sales Using an Analysis of Internet Message Board Discussions’, MA Thesis, San Diego State University, San Diego, California, USA.