Animesh Kansal

Jaahnavi Tiruthani

Medha Tiwary

Hanson Wu

Homework 4 Group 2

**1a. Final Model:**

**DATA** vgsales;

SET "C:\Users\bxl131330\Desktop\videogamesales\_main.sas7bdat";

**DATA** new;/\*Interaction variables creation\*/

SET vgsales;

CriticScSq = Critic\_Score \* Critic\_Score;

UserScSq = User\_Score \* User\_Score;

BothSc = Critic\_Score \* User\_Score;

**RUN**;

SET "C:\Users\bxl131330\Desktop\videogamesales\_main.sas7bdat";

**PROC** **GLM** data=new; /\*Dropped all interaction variables and Publisher\*/

CLASS Year\_of\_Release Rating Genre Platform Developer;

MODEL Global\_Sales = Critic\_Score Critic\_Count User\_Score User\_Count Year\_of\_Release Rating Genre Platform Developer;

**QUIT**;



**1b. Initial Model:**

/\*Creation of initial model\*/

**DATA** vgsales;

SET "C:\Users\bxl131330\Desktop\videogamesales\_main.sas7bdat";

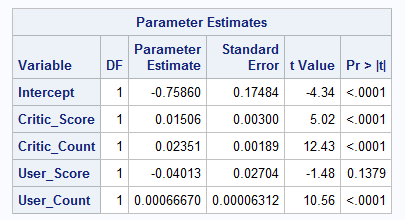
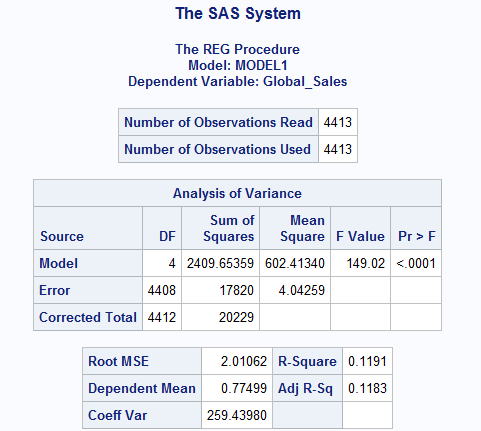
**PROC** **PRINT** data=vgsales;

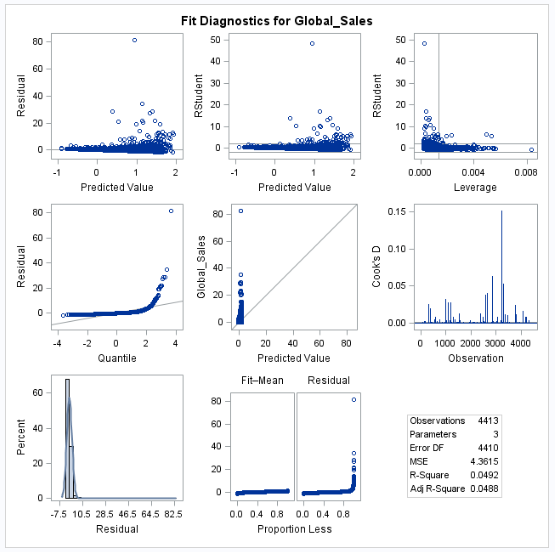
**RUN**;

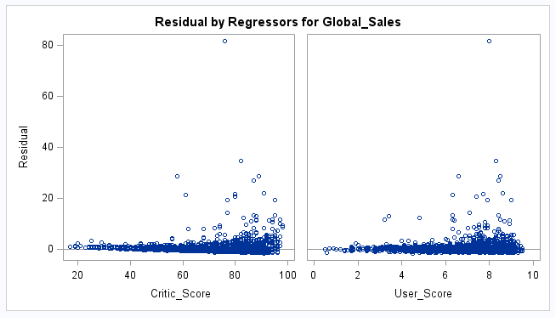
**PROC** **REG** data=vgsales;

MODEL Global\_Sales = Critic\_Score Critic\_Count User\_Score User\_Count;

**QUIT**;







Our initial model was a regression model based on the variables Critic Score, Critic Count, User Score, and User Count. It resulted in a model with a very low R2 value.

First, we needed to see the relationship between the individual variables and their effects on Global Sales. We did this by using the PROC SGPLOT function to plat the numeric variables against Global Sales. We discovered a positive correlation between both Critic Score and User Score to Global Sales. However, there was no obvious correlation in the variables Year of Release, Critic Count, and User Count. We would later discover this reasoning to be faulty because just because a relationship is not evident in a scatterplot does not mean that it does not exist.

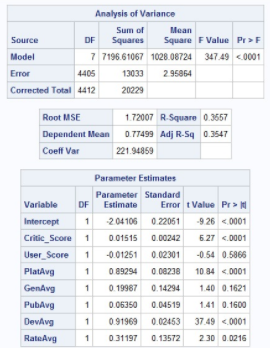
To see the relationship between the categorical variables and Global Sales, we utilized the PROC MEANS function to find the average Global Sales for each Platform, Genre, Publisher, Developer, and Rating. We found that the mean Global Sales for these variables differed significantly and concluded that they were necessary in our model. To quantify these variables, we merged the resulting tables created from the PROC MEANS functions all into one big master table. The new variables PlatAvg, GenAvg, PubAvg, DevAvg, and RateAvg represented the average global sales that videogames for each Platform, in each Genre, published by each Publisher, developed by each Developer, and rated each Rating could be expected to sell respectively. The code for this is posted at the end of this document.

We came to an altered model:

**PROC** **REG** data=merge5;

MODEL Global\_Sales = Critic\_Score User\_Score PlatAvg GenAvg PubAvg DevAvg RateAvg Year\_of\_Release Critic\_Count User\_Count;

**QUIT**;



This model had an R-Square value of 0.3657 which is a considerable improvement. However, the p-values for the variables User\_Score, GenAvg, and PubAvg were still not significant.

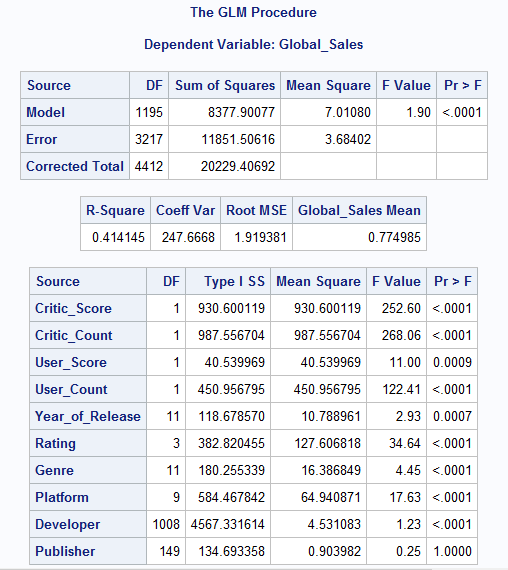
We decided to start using the PROC GLM function since many of our variables were not numeric and assigning numeric values using the PROC MEANS function was faulty logic since the variation among different Platforms, Genres, Publishers, Developers, and Ratings could vary greatly (e.g. one publisher may only have 1 data point, but another publisher may have over 50). We used CLASS to create dummy variables internally. Our first model using PROC GLM is as follows:

**PROC** **GLM** data=vgsales;

CLASS Year\_of\_Release Rating Genre Platform Developer Publisher;

MODEL Global\_Sales = Critic\_Score Critic\_Count User\_Score User\_Count Year\_of\_Release Rating Genre Platform Developer Publisher;

**QUIT**;



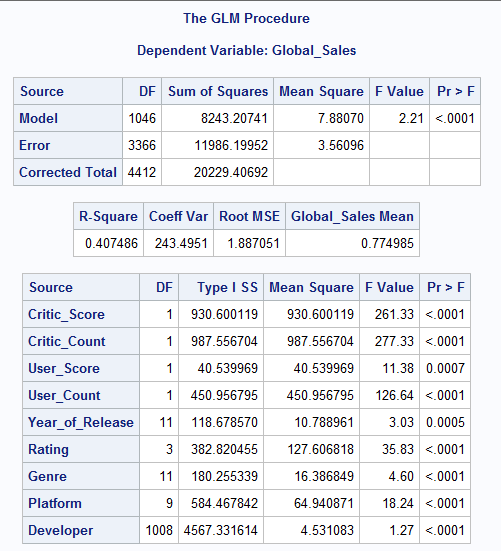
We got a higher R2 but the p-value for Publisher was 1.0 which was alarming. We repeated the procedure but did not use the Publisher variable and obtained the following results:

**PROC** **GLM** data=vgsales;

CLASS Year\_of\_Release Rating Genre Platform Developer;

MODEL Global\_Sales = Critic\_Score Critic\_Count User\_Score User\_Count Year\_of\_Release Rating Genre Platform Developer;

**QUIT**;



The R2 value is a little lower but all the variables had a significant p-value.

Our next steps were to test if any interaction variables were needed. We noticed that the relationships between Critic Score and Global sales as well as User Score and Global Sales was positive but not completely linear so the code below was implemented:

**DATA** new;

SET vgsales;

CriticScSq = Critic\_Score \* Critic\_Score;

UserScSq = User\_Score \* User\_Score;

BothSc = Critic\_Score \* User\_Score;

**RUN**;

**PROC** **GLM** data=new;

CLASS Year\_of\_Release Rating Genre Platform Developer;

MODEL Global\_Sales = BothSc Critic\_Score CriticScSq Critic\_Count User\_Score UserScSq User\_Count Year\_of\_Release Rating Genre Platform Developer;

**QUIT**;



This model has a higher R2 value. However, the p-value for the variable UserScSq (User­\_Score2) was not significant. We tried removing the interaction variable but that caused other variables to not have a significant p-value. We ended up having to remove all the interaction variables to let all p-values in the model become significant again.

**1c. Interpretation of results:**

About 40.75% of the variation in Global\_Sales can be explained by our model. At this point it is suspected that the variables interact in other ways than linearly which is why the model is not as robust as expected. In the next few steps, we have explored the correlation between the variables as well as introduced more variables in terms of mean to understand the sales against different parameters.

There is heteroskedasticity in the data since the residuals of sales are increasing exponentially against both user\_score and critic\_score. This problem is resolved in the next few steps.

**2. Verify Regression Assumptions**

Outliers:

**PROC** **GLM** data=new; /\*adds cook's d\*/

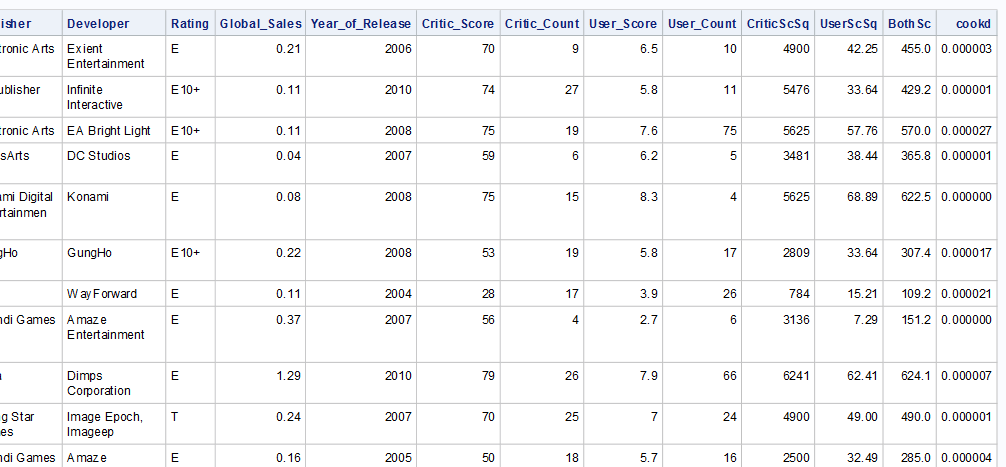
CLASS Year\_of\_Release Rating Genre Platform Developer;

MODEL Global\_Sales = Critic\_Score Critic\_Count User\_Score User\_Count Year\_of\_Release Rating Genre Platform Developer;

OUTPUT OUT = new1 cookd = cookd;

**RUN**;

**QUIT**;



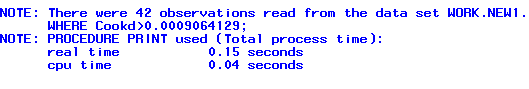
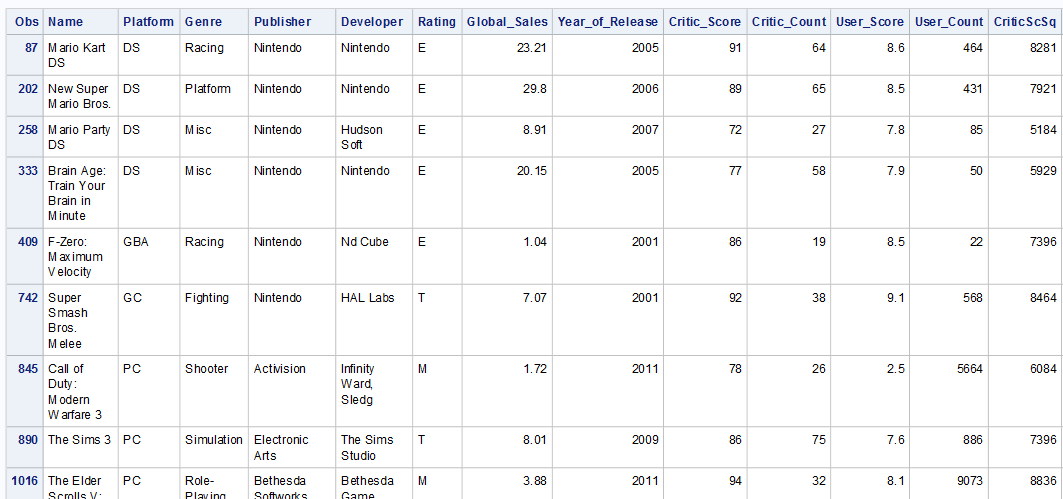
/\* prints the observations that are influential (cook's d > 4/ n) \*/

**PROC** **PRINT** DATA=new1;

VAR \_ALL\_;

WHERE Cookd > **4** / **4413**;

**run**; /\*42 observations were selected\*/

We had 42 outliers/influential points that we needed to drop.

**PROC** **GLM** data=new1;

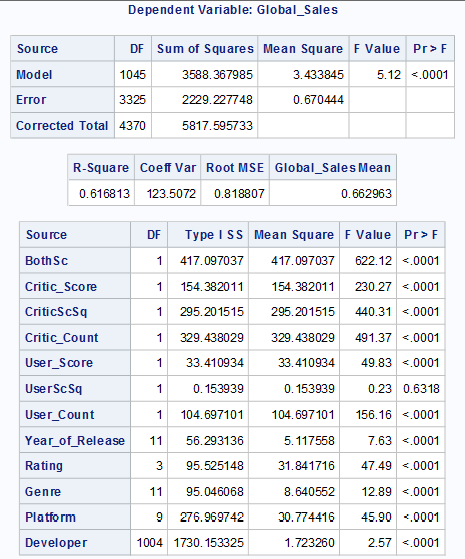
CLASS Year\_of\_Release Rating Genre Platform Developer;

MODEL Global\_Sales = Critic\_Score Critic\_Count User\_Score User\_Count Year\_of\_Release Rating Genre Platform Developer;

WHERE Cookd < **4** / **4413**; /\*Drops Outliers and influential points\*/

**RUN**;

**QUIT**;

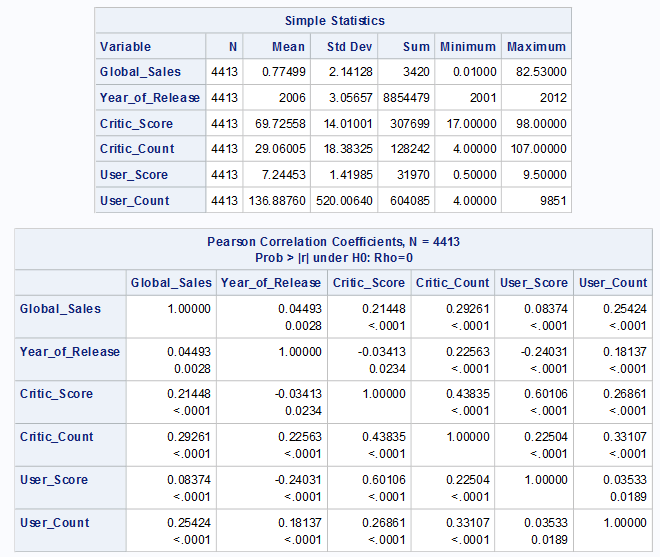


/\*Multicollinearity test\*/

**PROC** **CORR** data = vgsales;

VAR Global\_Sales Year\_of\_Release Critic\_Score Critic\_Count User\_Score User\_Count ;

**RUN**;



Correlation Matrix shows that, there is no multicollinearity in the primary variables to be used in the initial regression model. All mentioned variables are independent or in other words, all variables are not significantly correlated.

/\*Heteroscedasticity detection\*/

/\* How does Global Sales vary with Critic Score \*/

**PROC** **REG** data=vgsales;

MODEL Global\_Sales = Critic\_Score / hcc spec;

**RUN**;

**PROC** **REG** data=vgsales;

MODEL Global\_Sales = User\_Score / hcc spec;

**RUN**;

**PROC** **REG** data=vgsales;

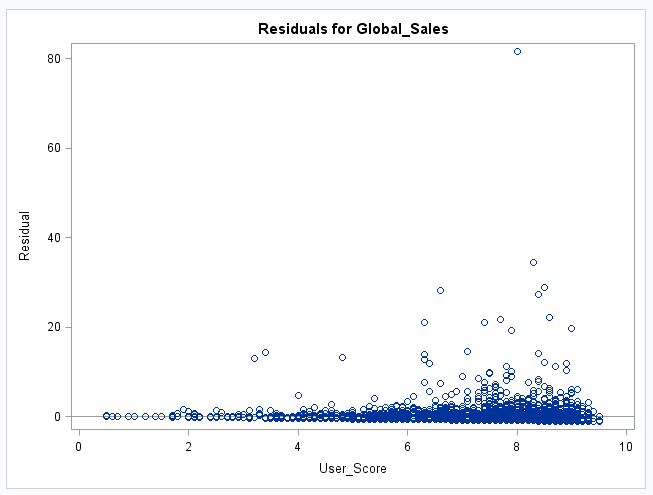
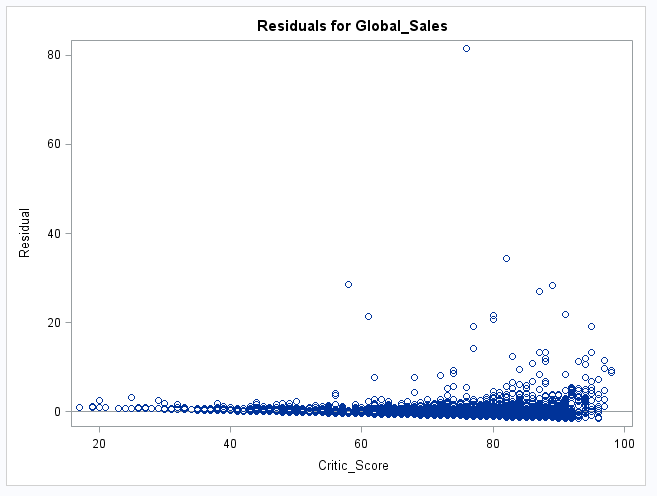
MODEL Global\_Sales = Critic\_Count / hcc spec;

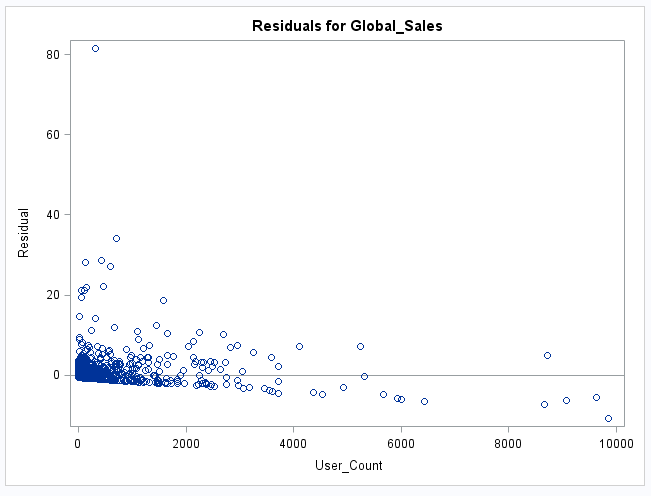
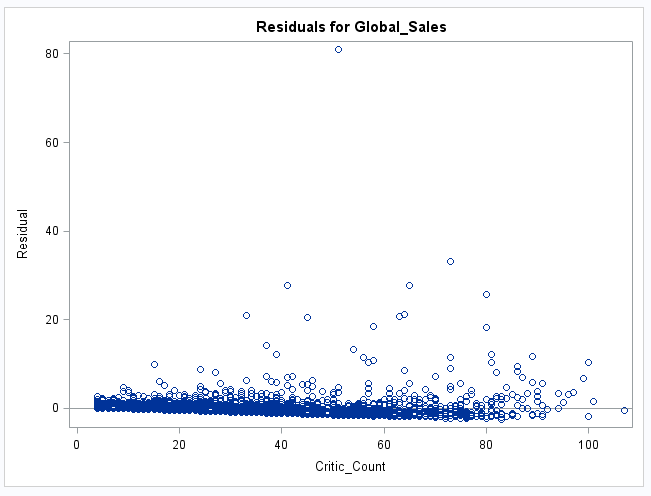
**RUN**;

**PROC** **REG** data=vgsales;

MODEL Global\_Sales = User\_Count / hcc spec;

**RUN**;





There is heteroscedasticity because the residuals increase as Critic\_Score and User\_Score increase and decrease as User\_Count decreases. To solve this we must transform Global\_Sales using log().

/\* using a log transform for the y variable to tackle heteroscedasticity \*/

**DATA** new2;

SET new1;

log\_Sales = log(Global\_Sales);

**RUN**;

**PROC** **GLM** data=new2;

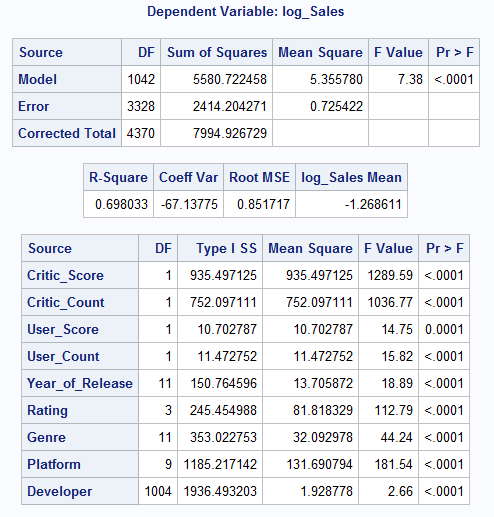
CLASS Year\_of\_Release Rating Genre Platform Developer;

MODEL log\_Sales = Critic\_Score Critic\_Count User\_Score User\_Count Year\_of\_Release Rating Genre Platform Developer;

WHERE Cookd < **4** / **4413**; /\*Drops Outliers and influential points\*/

**RUN**;

**QUIT**;



All these assumptions check, and the fixes made a significant impact on our model by greatly increasing the R2 value.

Code is below

**Code Dump**

**DATA** vgsales;

/\*SET "C:\Users\bxl131330\Desktop\videogamesales\_main.sas7bdat";\*/

SET "C:\Users\Hanson\Documents\Hanson\College\UTD Spring 2018\Predicive Analaysis using SAS\Datasets\videogamesales\_main.sas7bdat";

**RUN**;

/\*Creation of initial model\*/

**PROC** **REG** data=vgsales;

MODEL Global\_Sales = Critic\_Score Critic\_Count User\_Score User\_Count;

**QUIT**;

/\*

PROC REG data=vgsales;

MODEL Global\_Sales = Critic\_Score User\_Score;

QUIT;

\*/

/\*Checking variable relationships\*/

**proc** **sgplot** data = vgsales;

scatter X=User\_Score Y=Global\_Sales;

title 'User Score and Global Sales';

label User\_Score = 'User Score' Global\_Sales = 'Global Sales';

**run**; /\*Positive correlation with User Score\*/

**proc** **sgplot** data = vgsales;

scatter X=Critic\_Score Y=Global\_Sales;

title 'Critic Score and Global Sales';

label Critic\_Score = 'Critic Score' Global\_Sales = 'Global Sales';

**run**; /\*Positive correlation with Critic Score\*/

**proc** **sgplot** data = vgsales;

scatter X=Critic\_Count Y=Global\_Sales;

title 'Critic Count and Global Sales';

label Critic\_Count = 'Critic Count' Global\_Sales = 'Global Sales';

**run**; /\*No Correlation with Critic Count\*/

**proc** **sgplot** data = vgsales;

scatter X=User\_Count Y=Global\_Sales;

title 'User Count and Global Sales';

label User\_Count = 'User Count' Global\_Sales = 'Global Sales';

**run**; /\*No Correlation with User Count\*/

**proc** **sgplot** data = vgsales;

scatter X=Year\_of\_Release Y=Global\_Sales;

title 'Year of Release and Global Sales';

label Year\_of\_Release = 'Year of Release' Global\_Sales = 'Global Sales';

**run**; /\*No Correlation with Year of Release\*/

**PROC** **MEANS** DATA = vgsales NWAY;

CLASS Platform;

VAR Global\_Sales;

OUTPUT OUT = PlatAv (drop=\_TYPE\_ \_FREQ\_) MEAN = PlatAvg;

TITLE 'Platform and Average Global Sales';

**RUN**;

**PROC** **SORT** DATA=PlatAv;

BY PlatAvg;

**PROC** **PRINT** DATA=PlatAv;

**RUN**; /\*Platform matters\*/

**PROC** **SORT** DATA=PlatAv;

BY Platform;

**RUN**;

**PROC** **SORT** DATA=vgsales;

BY Platform;

**RUN**;

**DATA** merge1; /\*We are merging the two data sets to assign a numeric value to the categorical value 'Platform'\*/

MERGE vgsales PlatAv;

BY Platform;

**PROC** **PRINT** data=merge1;

**RUN**;

**PROC** **MEANS** DATA = vgsales NWAY;

CLASS Genre;

VAR Global\_Sales;

OUTPUT OUT = GenAv (drop=\_TYPE\_ \_FREQ\_) MEAN = GenAvg;

TITLE 'Genre and Average Global Sales';

**RUN**;

**PROC** **SORT** DATA=GenAv;

BY GenAvg;

**PROC** **PRINT** DATA=GenAv;

**RUN**; /\*Genre Matters\*/

**PROC** **SORT** DATA=GenAv;

BY Genre;

**RUN**;

**PROC** **SORT** DATA=merge1;

BY Genre;

**RUN**;

**DATA** merge2; /\*We are merging the two data sets to assign a numeric value to the categorical value 'Genre'\*/

MERGE merge1 GenAv;

BY Genre;

**PROC** **PRINT** data=merge2;

**RUN**;

**PROC** **MEANS** DATA = vgsales NWAY;

CLASS Publisher;

VAR Global\_Sales;

OUTPUT OUT = PubAv (drop=\_TYPE\_ \_FREQ\_) MEAN = PubAvg;

TITLE 'Publisher and Average Global Sales';

**RUN**;

**PROC** **SORT** DATA=PubAv;

BY PubAvg;

**PROC** **PRINT** DATA=PubAv;

**RUN**; /\*Publisher Matters\*/

**PROC** **SORT** DATA=PubAv;

BY Publisher;

**RUN**;

**PROC** **SORT** DATA=merge2;

BY Publisher;

**RUN**;

**DATA** merge3; /\*We are merging the two data sets to assign a numeric value to the categorical value 'Publisher'\*/

MERGE merge2 PubAv;

BY Publisher;

**PROC** **PRINT** data=merge3;

**RUN**;

**PROC** **MEANS** DATA = vgsales NWAY;

CLASS Developer;

VAR Global\_Sales;

OUTPUT OUT = DevAv (drop=\_TYPE\_ \_FREQ\_) MEAN = DevAvg;

TITLE 'Developer and Average Global Sales';

**RUN**;

**PROC** **SORT** DATA=DevAv;

BY DevAvg;

**PROC** **PRINT** DATA=DevAv;

**RUN**; /\*Developer Matters\*/

**PROC** **SORT** DATA=DevAv;

BY Developer;

**RUN**;

**PROC** **SORT** DATA=merge3;

BY Developer;

**RUN**;

**DATA** merge4; /\*We are merging the two data sets to assign a numeric value to the categorical value 'Developer'\*/

MERGE merge3 DevAv;

BY Developer;

**PROC** **PRINT** data=merge4;

**RUN**;

**PROC** **MEANS** DATA = vgsales NWAY;

CLASS Rating;

VAR Global\_Sales;

OUTPUT OUT = RateAv (drop=\_TYPE\_ \_FREQ\_) MEAN = RateAvg;

TITLE 'Publisher and Average Global Sales';

**RUN**;

**PROC** **SORT** DATA=RateAv;

BY RateAvg;

**PROC** **PRINT** DATA=RateAv;

**RUN**; /\*Games rated E and M have higher sales than games rated E10+ and T\*/

**PROC** **SORT** DATA=RateAv;

BY Rating;

**RUN**;

**PROC** **SORT** DATA=merge4;

BY Rating;

**RUN**;

**DATA** merge5; /\*We are merging the two data sets to assign a numeric value to the categorical value 'Rating'\*/

MERGE merge4 RateAv;

BY Rating;

**PROC** **PRINT** data=merge5;

**RUN**;

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*Adjusting and testing\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

**DATA** new;/\*Interaction variables creation\*/

SET vgsales;

CriticScSq = Critic\_Score \* Critic\_Score;

UserScSq = User\_Score \* User\_Score;

BothSc = Critic\_Score \* User\_Score;

**RUN**;

**PROC** **REG** data=merge5; /\*no\*/

MODEL Global\_Sales = Critic\_Score User\_Score PlatAvg GenAvg PubAvg DevAvg RateAvg Year\_of\_Release Critic\_Count User\_Count;

**QUIT**;

**PROC** **REG** data=merge5; /\*no\*/

MODEL Global\_Sales = Critic\_Score PlatAvg GenAvg PubAvg DevAvg RateAvg Year\_of\_Release Critic\_Count User\_Count;

**QUIT**;

/\*\*\*\*\*\*\*\*\* PROC GLM MODEL AND TESTING \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

**PROC** **GLM** data=vgsales; /\*All variables used, publisher gave p-value of 1.0, must drop\*/

CLASS Year\_of\_Release Rating Genre Platform Developer Publisher; /\*generates dummy variables internally\*/

MODEL Global\_Sales = Critic\_Score Critic\_Count User\_Score User\_Count Year\_of\_Release Rating Genre Platform Developer Publisher;

**QUIT**;

**PROC** **GLM** data=vgsales; /\*Dropped Publisher variable\*/

CLASS Year\_of\_Release Rating Genre Platform Developer;

MODEL Global\_Sales = Critic\_Score Critic\_Count User\_Score User\_Count Year\_of\_Release Rating Genre Platform Developer;

**QUIT**;

**PROC** **GLM** data=new; /\*Added some Interaction variables\*/

CLASS Year\_of\_Release Rating Genre Platform Developer;

MODEL Global\_Sales = BothSc Critic\_Score CriticScSq Critic\_Count User\_Score UserScSq User\_Count Year\_of\_Release Rating Genre Platform Developer;

**QUIT**;

**PROC** **GLM** data=new; /\*Dropped UserScSq (User\_Score\*User\_Score)\*/

CLASS Year\_of\_Release Rating Genre Platform Developer;

MODEL Global\_Sales = BothSc Critic\_Score CriticScSq Critic\_Count User\_Score User\_Count Year\_of\_Release Rating Genre Platform Developer;

**QUIT**;

**PROC** **GLM** data=new; /\*Dropped BothSc, added UserScSq back in\*/

CLASS Year\_of\_Release Rating Genre Platform Developer;

MODEL Global\_Sales = Critic\_Score CriticScSq Critic\_Count User\_Score UserScSq User\_Count Year\_of\_Release Rating Genre Platform Developer;

**QUIT**;

**PROC** **GLM** data=new; /\*Dropped UserScSq\*/

CLASS Year\_of\_Release Rating Genre Platform Developer;

MODEL Global\_Sales = Critic\_Score CriticScSq Critic\_Count User\_Score User\_Count Year\_of\_Release Rating Genre Platform Developer;

**QUIT**;

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*INITIAL FINAL MODEL\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

**PROC** **GLM** data=new; /\*Dropped all interaction variables and Publisher\*/

CLASS Year\_of\_Release Rating Genre Platform Developer;

MODEL Global\_Sales = Critic\_Score Critic\_Count User\_Score User\_Count Year\_of\_Release Rating Genre Platform Developer;

**QUIT**;

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*REGRESSION ASSUMPTIONS AREA\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

/\*OUTLIER AND INFLUENTIAL POINTS TESTING\*/

**PROC** **GLM** data=new; /\*adds cook's d\*/

CLASS Year\_of\_Release Rating Genre Platform Developer;

MODEL Global\_Sales = BothSc Critic\_Score CriticScSq Critic\_Count User\_Score UserScSq User\_Count Year\_of\_Release Rating Genre Platform Developer;

OUTPUT OUT = new1 cookd = cookd;

**RUN**;

**QUIT**;

/\* prints the observations that are influential (cook's d > 4/ n) \*/

**PROC** **PRINT** DATA=new1;

VAR \_ALL\_;

WHERE Cookd > **4** / **4413**;

**run**; /\*42 observations were selected\*/

/\* Estimate the model without the influential points \*/

**PROC** **GLM** data=new1;

CLASS Year\_of\_Release Rating Genre Platform Developer;

MODEL Global\_Sales = Critic\_Score Critic\_Count User\_Score User\_Count Year\_of\_Release Rating Genre Platform Developer;

WHERE Cookd < **4** / **4413**; /\*Drops Outliers and influential points\*/

**RUN**;

**QUIT**;

/\*Multicollinearity test\*/

**PROC** **CORR** data = vgsales;

VAR Global\_Sales Year\_of\_Release Critic\_Score Critic\_Count User\_Score User\_Count ;

**RUN**; /\*No variables are collinear\*/

/\*Heteroscedasticity detection\*/

/\* How does Global Sales vary with Critic Score \*/

**PROC** **REG** data=vgsales;

MODEL Global\_Sales = Critic\_Score / hcc spec;

**RUN**;

**PROC** **REG** data=vgsales;

MODEL Global\_Sales = User\_Score / hcc spec;

**RUN**;

**PROC** **REG** data=vgsales;

MODEL Global\_Sales = Critic\_Count / hcc spec;

**RUN**;

**PROC** **REG** data=vgsales;

MODEL Global\_Sales = User\_Count / hcc spec;

**RUN**;

/\* using a log transform for the y variable to tackle heteroscedasticity \*/

**DATA** new2;

SET new1;

log\_Sales = log(Global\_Sales);

**RUN**;

/\*\*\*\*\*\*\*\*\*\*\*FINAL FINAL MODEL\*\*\*\*\*\*\*\*/

**PROC** **GLM** data=new2;

CLASS Year\_of\_Release Rating Genre Platform Developer;

MODEL log\_Sales = Critic\_Score Critic\_Count User\_Score User\_Count Year\_of\_Release Rating Genre Platform Developer;

WHERE Cookd < **4** / **4413**; /\*Drops Outliers and influential points\*/

**RUN**;

**QUIT**;