REPORT
TOPIC – REGRESSION ANALYSIS

The final model selected by me uses log transformation of the dependent variable Global Sales and independent factors namely, Platform, Genre, Rating keeping PC, Strategy and E10+ as the base conditions respectively. It also uses the User Score and Critic Score both in the standardized forms, User count, critic count, an additional variable called the weighted score which is basically the average rating of the critic and user scores. I have transformed the year variable to represent 2 generations of the release and hence act as important market segments. This market segment has a joint effect with the platform on which the video game was released and hence accounts for an interaction effect.

I developed this model after a careful analysis and putting the variables through a series of transformations described later in the report. Post the transformations, linear regression for performed at every step to analyse the coefficients, their sign indicating whether or not these variables complemented their effect on the Y dependent, the p-values to see if these variables were significant and the R square. The R-square needed the utmost vigilance as I did not want to force any changes that would overfit the model and lead to poor predictions.

The GLM Procedure

Dependent Variable: log_global_sales

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	35	4129.286027	117.979601	177.77	<.0001
Error	3904	2590.913987	0.663656		
Corrected Total	3939	6720.200014			

R-Square	Coeff Var	Root MSE	log_global_sales Mean
0.614459	-68.49612	0.814651	-1.189339

Source	DF	Type I SS	Mean Square	F Value	Pr > F
Rating	3	75.681520	25.227173	38.01	<.0001
Platform	9	1495.607441	166.178605	250.40	<.0001
Genre	11	91.777425	8.343402	12.57	<.0001
Weighted_Score	1	219.190811	219.190811	330.28	<.0001
Critic_Score	1	1306.956196	1306.956196	1969.33	<.0001
User_Score	1	18.630524	18.630524	28.07	<.0001
User_Count	1	367.332349	367.332349	553.50	<.0001
Critic_Count	1	502.855779	502.855779	757.71	<.0001
Generation	1	30.585530	30.585530	46.09	<.0001
Generation*Platform	6	20.668453	3.444742	5.19	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Rating	3	71.8314090	23.9438030	36.08	<.0001
Platform	9	773.1260855	85.9028984	129.44	<.0001

Question 1 – Part B

In this part I will be explaining how I developed this model.

Initial Assumptions:

- The scale of the 2 scores – Critic and User were different. One in the range of 0-100 and the other in the range of 1-10. These rating are themselves the average rating respective of the number of ratings each video game was received. Therefore the count of ratings received for each game was different.
- We cannot put the categorical variables directly into regression and hence must be converted to dummy variables. Rating had just a few but platform and genre would be a time-consuming process. Hence it was better to use proc GLM where ever suited.
- Assessing the importance and influence of the dependent variables on the global sales. Name of the game, Publisher or the developer will not affect the global sales as much as the other variables would do.

Initial model:

Y = Global Sales
X = Platform Rating Genre User_Count User_Score Critic_Count Critic Score Year_of_Release

R-Square	Coeff Var	Root MSE	Global_Sales Mean
0.179467	251.0792	1.945827	0.774985

Source	DF	Type I SS	Mean Square	F Value	Pr > F
Rating	3	169.109515	56.369838	14.89	<.0001
Platform	9	505.124084	56.124898	14.82	<.0001
Genre	11	78.663662	7.151242	1.89	0.0361
Critic_Score	1	1187.767928	1187.767928	313.71	<.0001
User_Score	1	44.239411	44.239411	11.68	0.0006
User_Count	1	965.558067	965.558067	255.02	<.0001
Critic_Count	1	660.188247	660.188247	174.36	<.0001
Year_of_Release	1	19.865452	19.865452	5.25	0.0220

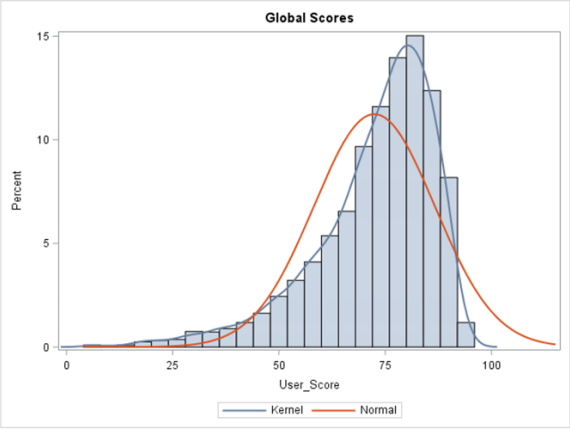
Source	DF	Type III SS	Mean Square	F Value	Pr > F
Rating	3	115.4985336	38.4995112	10.17	<.0001
Platform	9	622.7733146	69.1970350	18.28	<.0001
Genre	11	114.7940260	10.4358205	2.76	0.0014
Critic_Score	1	103.9904009	103.9904009	27.47	<.0001
User_Score	1	14.7779307	14.7779307	3.90	0.0483
User_Count	1	673.5194878	673.5194878	177.89	<.0001
Critic_Count	1	653.6114133	653.6114133	172.63	<.0001
Year_of_Release	1	19.8654520	19.8654520	5.25	0.0220

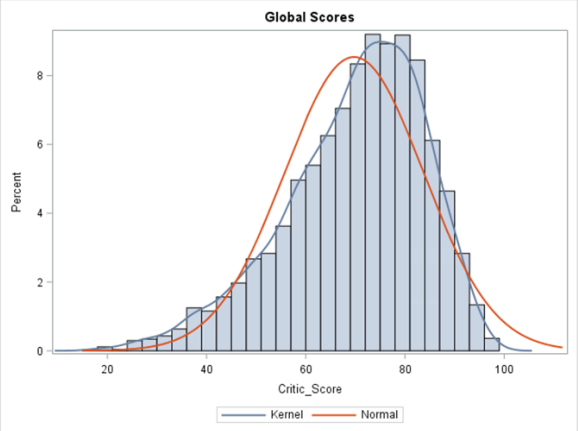
Parameter	Estimate		Standard Error	t Value	Pr > |t|
Intercept	73.13548902	B	32.84194538	2.23	0.0260
Rating E	0.39313205	B	0.10688511	3.68	0.0002
Rating M	-0.12551427	B	0.12329393	-1.02	0.3087
Rating T	-0.08665409	B	0.10392823	-0.83	0.4044
Rating E10+	0.00000000	B	.	.	.
Platform DS	1.09772964	B	0.15638081	7.02	<.0001
Platform GBA	0.73227316	B	0.19994259	3.66	0.0003
Platform GC	0.52045930	B	0.17652489	2.95	0.0032
Platform PS2	0.92637422	B	0.14389268	6.44	<.0001
Platform PS3	0.81936174	B	0.14106708	5.81	<.0001
Platform PSP	0.74542744	B	0.15736012	4.74	<.0001
Platform Wii	1.66885452	B	0.15319024	10.89	<.0001
Platform X360	0.57441792	B	0.14432307	3.98	<.0001
Platform XB	0.31626963	B	0.15879138	1.99	0.0465
Platform PC	0.00000000	B	.	.	.
Genre Action	0.24693290	B	0.16352353	1.51	0.1311
Genre Adventure	0.04442334	B	0.21769983	0.20	0.8383
Genre Fighting	0.31863944	B	0.20068933	1.59	0.1124
Genre Misc	0.62429122	B	0.19360417	3.22	0.0013
Genre Platform	0.27889691	B	0.19783407	1.41	0.1587
Genre Puzzle	-0.15406447	B	0.27964614	-0.55	0.5817
Genre Racing	0.30175998	B	0.18131754	1.66	0.0961

The screen shot shows only part of the parameter coefficients. The joint effect of all the levels of Rating, Platforms and Genre can be observed from the anova table generated by above and we can safely say that it is significant. But the R square is 0.179467 which is not good at all and needs improvement.

Feature Engineering:

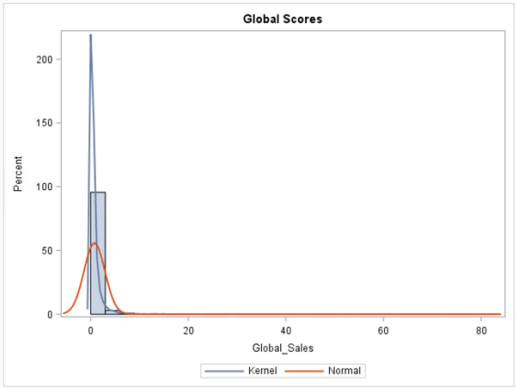
Before transforming the data, it is important we understand the skewness of the numerical variables, test for the presence of missing values and the correlation between the variables. I multiplied the user score by 10 and created another variable weighted score = [(user score * user count) + (critic score* critic count)] / critic count + user count.





Pearson Correlation Coefficients, N = 4413
Prob > |r| under H0: Rho=0

	log_global_sales	Critic_Score	User_Score	User_Count	Critic_Count	Weighted_Score
log_global_sales	1.00000	0.35058 <.0001	0.16202 <.0001	0.23055 <.0001	0.44016 <.0001	-0.08563 <.0001
Critic_Score	0.35058 <.0001	1.00000	0.60106 <.0001	0.26861 <.0001	0.43835 <.0001	0.15562 <.0001
User_Score	0.16202 <.0001	0.60106 <.0001	1.00000	0.03533 0.0189	0.22504 <.0001	0.23731 <.0001
User_Count	0.23055 <.0001	0.26861 <.0001	0.03533 0.0189	1.00000	0.33107 <.0001	-0.14291 <.0001
Critic_Count	0.44016 <.0001	0.43835 <.0001	0.22504 <.0001	0.33107 <.0001	1.00000	0.16885 <.0001
Weighted_Score	-0.08563 <.0001	0.15562 <.0001	0.23731 <.0001	-0.14291 <.0001	0.16885 <.0001	1.00000



From the other graphs we now know that the global sales are highly skewed to the right whereas the User scores and critic scores are also skewed. Thus, I transformed the y variable global scales into its log function and used it hence forth in regression. This spread the data pretty much normally. We can also see that there is a slight correlation among the User Scores and Critic Scores. So I analyzed the outlier points and discarded the once that were least important and way away from the 2 standard deviations. Once this was done i had better results by standardize the 2 variables around the mean and SD

Later I performed proc freq on platform*year_of_release to understand if the 2 variables had any significance. From the table below, I could identify that the platform on which this video games were release have a patter. They span for a duration and then either get replaced by their competition brand or have a new version of themselves that takes over. Hence, I divided the year into two generations; 2001-2006 as 1st gen and 2007-2012 as 2nd gen. It cannot be ruled out that the market conditions will vary with the 2 segments referring to how the need of the customers and their interests have evolved over time.

Frequency Percent Row Pct Col Pct	Table of Platform by Year_of_Release												
	Year_of_Release												
Platform	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	Total
DS	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	10 0.23 2.83 2.67	43 0.97 12.18 9.71	56 1.27 15.86 13.02	83 1.88 23.51 17.33	80 1.81 22.66 16.95	47 1.07 13.31 11.44	27 0.61 7.65 7.54	6 0.14 1.70 1.87	1 0.02 0.28 0.56	353 8.00
GBA	23 0.52 12.11 13.45	48 1.09 25.26 12.44	46 1.04 24.21 11.79	46 1.04 24.21 12.30	18 0.41 9.47 4.06	9 0.20 4.74 2.09	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	190 4.31
GC	18 0.41 6.36 10.53	81 1.84 28.62 20.98	77 1.74 27.21 19.74	39 0.88 13.78 10.43	49 1.11 17.31 11.06	17 0.39 6.01 3.95	2 0.05 0.71 0.42	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	283 6.41
PC	7 0.16 1.67 4.09	16 0.36 3.82 4.15	21 0.48 5.01 5.38	20 0.45 4.77 5.35	21 0.48 5.01 4.74	33 0.75 7.88 7.67	44 1.00 10.50 9.19	44 1.00 10.50 9.32	55 1.25 13.13 13.38	53 1.20 12.65 14.80	71 1.61 16.95 22.12	34 0.77 8.11 19.10	419 9.49
PS2	103 2.33 11.83 60.23	155 3.51 17.80 40.16	146 3.31 16.76 37.44	153 3.47 17.57 40.91	138 3.13 15.84 31.15	93 2.11 10.68 21.63	44 1.00 5.05 9.19	28 0.63 3.21 5.93	10 0.23 1.15 2.43	1 0.02 0.11 0.28	0 0.00 0.00 0.00	0 0.00 0.00 0.00	871 19.74
PS3	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	17 0.39 3.15 3.95	65 1.47 12.06 13.57	93 2.11 17.25 19.70	91 2.06 16.88 22.14	99 2.24 18.37 27.65	105 2.38 19.48 32.71	69 1.56 12.80 38.76	539 12.21
PSP	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	6 0.14 1.97 1.60	53 1.20 17.43 11.96	81 1.84 26.64 18.84	63 1.43 20.72 13.15	29 0.66 9.54 6.14	32 0.73 10.53 7.79	31 0.70 10.20 8.66	9 0.20 2.96 2.80	0 0.00 0.00 0.00	304 6.89

If the platforms were revised with the progressing years, we identify an interaction effect and thus build a segment response model.

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	37	4157.351738	112.360858	121.33	<.0001
Error	4117	3812.699650	0.926087		
Corrected Total	4154	7970.051388			

R-Square	Coeff Var	Root MSE	log_global_sales Mean
0.521622	-80.12770	0.962334	-1.201000

Source	DF	Type I SS	Mean Square	F Value	Pr > F
Rating	3	69.947478	23.315826	25.18	<.0001
Platform	9	1374.801069	152.755674	164.95	<.0001
Genre	11	117.290624	10.662784	11.51	<.0001
Weighted_Score	1	188.875684	188.875684	203.95	<.0001
Critic_Score	1	1383.047295	1383.047295	1493.43	<.0001
User_Score	1	14.340598	14.340598	15.49	<.0001
User_Count	1	317.236757	317.236757	342.56	<.0001
Critic_Count	1	647.673016	647.673016	699.37	<.0001
Generation	1	18.952715	18.952715	20.47	<.0001
Generation*Platform	8	25.186503	3.148313	3.40	0.0007

The r square has improved and all the variables seem to be significant even when in the combined form as denoted by the anova table. We see that generation is no longer significant but the interaction effect is. No matter how tempted we might be to delete this variable we shouldn't.

Ideally if the user scores are high then the global sales should also be high. But the coefficient has a negative sign which is strange. Thus, we haven't yet handled the influential points. This is when I started looking at whether I was meeting the assumptions of regression or not.
Hence, I performed the cook's d test to check how many of the points were greater then 4/n;
I used proc robust reg to add weights to these observations rather than just dropping them.

Obs	Name	Platform	Genre	Publisher	Developer	Rating	Global_Sales	Year_of_Release	Critic_Score	Critic_Count	User_Score	User_Count	log_^
22	Namco Museum	GBA	Misc	Namco Bandai Games	Mass Media	E	4.24	2001	0.6257664835	10	-0.1649949	6	
59	The Sims: House Party	PC	Simulation	Electronic Arts	Maxis	T	2.16	2001	0.2275462208	17	0.3935010763	30	
77	IL-2 Sturmovik	PC	Simulation	Blue Byte	1C, 1C Company	T	0.01	2001	1.5814951139	12	1.2312450415	130	
78	The Sims: Hot Date	PC	Simulation	Electronic Arts	Maxis	T	1.82	2001	1.1036307987	17	0.3935010763	70	
110	Super Smash Bros. Melee	GC	Fighting	Nintendo	HAL Labs	T	7.07	2001	1.6611391664	38	1.5104930299	568	
115	Rez	PS2	Shooter	Sony Computer Entertainment	UGA	E	0.05	2001	0.546122431	34	0.7658317275	28	
140	The Simpsons: Road Rage	XB	Racing	Electronic Arts	Fox Interactive	T	1.05	2001	-0.807826462	17	0.7658317275	6	
149	Gitaroo Man	PS2	Misc	THQ	Koei/Inis	E	0.05	2001	0.8646986411	21	1.1381623787	25	
150	Pac-Man Collection	GBA	Puzzle	Atari	Mass Media	E	2.94	2001	0.6257664835	13	0.3004184135	4	
198	RollerCoaster Tycoon 2	PC	Strategy	Atari	Chris Sawyer	E	1.25	2002	0.2275462208	22	0.9519970531	135	
207	Egg Mania: Eggstreme Madness	GC	Puzzle	Kemco	Kemco	E	0.01	2002	-0.568894304	7	0.4865837391	4	
274	Tetris Worlds	PS2	Puzzle	THQ	Blue Planet Software	E	2.08	2002	-2.161775355	7	-1.188904191	11	
281	Darkened Skye	GC	Adventure	TDK Mediactive	Boston Animation	T	0.01	2002	-0.807826462	12	-1.002738866	7	
336	Space Channel	GBA	Misc	Atari	Art	E	0.02	2002	-1.285690777	15	0.4865837391	5	

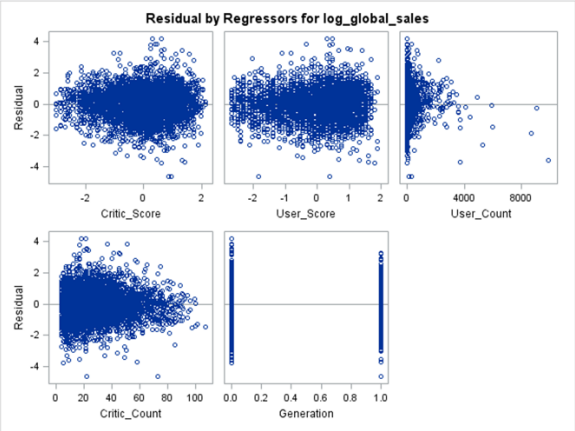
```
data vg_sales_reg;
set vg_sales_cookd;
/*Considering the base case as E10+ */
IF Rating = "T" THEN Rating_T=1;  ELSE Rating_T=0;
IF Rating = "M" THEN Rating_M=1;  ELSE Rating_M=0;
IF Rating = "E" THEN Rating_E=1;  ELSE Rating_E=0;
/*Considering the base case as Strategy */
IF Genre = "Action" THEN Genre_Action=1;  ELSE Genre_Action=0;
IF Genre = "Adventure" THEN Genre_Adventure=1;  ELSE Genre_Adventure=
IF Genre = "Fighting" THEN Genre_Fighting=1;  ELSE Genre_Fighting=0;
IF Genre = "Misc" THEN Genre_Misc=1;  ELSE Genre_Misc=0;
IF Genre = "Platform" THEN Genre_Platform=1;  ELSE Genre_Platform=0;
IF Genre = "Puzzle" THEN Genre_Puzzle=1;  ELSE Genre_Puzzle=0;
IF Genre = "Racing" THEN Genre_Racing=1;  ELSE Genre_Racing=0;
IF Genre = "Role-Playing" THEN Genre_RolePlaying=1;  ELSE Genre_RoleP
IF Genre = "Shooter" THEN Genre_Shooter=1;  ELSE Genre_Shooter=0;
IF Genre = "Simulation" THEN Genre_Simulation=1;  ELSE Genre_Simulati
IF Genre = "Strategy" THEN Genre_Strategy=1;  ELSE Genre_Strategy=0;
/*Considering the base case as PC */
IF Platform = "DS" THEN Platform_DS=1;  ELSE Platform_DS=0;
IF Platform = "GBA" THEN Platform_GBA=1;  ELSE Platform_GBA=0;
IF Platform = "GC" THEN Platform_GC=1;  ELSE Platform_GC=0;
IF Platform = "XB" THEN Platform_XB=1;  ELSE Platform_XB=0;
IF Platform = "PS2" THEN Platform_PS2=1;  ELSE Platform_PS2=0;
IF Platform = "PS3" THEN Platform_PS3=1;  ELSE Platform_PS3=0;
IF Platform = "PSP" THEN Platform_PSP=1;  ELSE Platform_PSP=0;
IF Platform = "Wii" THEN Platform_Wii=1;  ELSE Platform_Wii=0;
IF Platform = "X360" THEN Platform_X360=1;  ELSE Platform_X360=0;
run;
```

Since I cannot just enter the categorical variables directly into the proc robustreg I had to manually create all the dummy variables.
I then tested for the presence of multicollinearity as generation was not significant. we observe that all the values of vif are less than 10 indicating that there is not much multicollinearity in the dataset and the condition index is less than 10 for all eigen values meaning that they have a weak effect even if they do. Hence, there is no collinearity in the variables.

Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	-3.28034	0.09698	-33.82	<.0001	0
Rating_T	1	-0.12976	0.05308	-2.44	0.0145	2.89267
Rating_M	1	-0.29296	0.06320	-4.64	<.0001	2.82000
Rating_E	1	0.21234	0.05473	3.88	0.0001	2.90736
Platform_DS	1	2.04375	0.08109	25.20	<.0001	2.11705
Platform_GBA	1	2.23947	0.10071	22.24	<.0001	1.90723
Platform_GC	1	2.05218	0.08928	22.99	<.0001	2.17763
Platform_XB	1	2.01178	0.08405	23.94	<.0001	2.90866
Platform_PS2	1	2.52961	0.07263	34.83	<.0001	3.78250
Platform_PS3	1	2.08590	0.07291	28.61	<.0001	2.50446
Platform_PSP	1	2.02288	0.08092	25.00	<.0001	1.94078
Platform_Wii	1	2.35203	0.08012	29.36	<.0001	2.15947
Platform_X360	1	1.66543	0.07465	22.31	<.0001	2.97999
Genre_Action	1	-0.01260	0.06197	-0.20	0.8389	2.99508
Genre_Adventure	1	-0.47599	0.09888	-4.81	<.0001	1.39209
Genre_Fighting	1	-0.00788	0.08555	-0.09	0.9267	1.66022
Genre_Misc	1	0.26524	0.07561	3.51	0.0005	1.44377
Genre_Platform	1	-0.09775	0.07745	-1.26	0.2070	1.45780
Genre_Puzzle	1	-0.34132	0.12678	-2.69	0.0071	1.18886
Genre_Racing	1	-0.10029	0.06510	-1.54	0.1235	1.56866
Genre_RolePlaying	1	-0.40104	0.07290	-5.50	<.0001	2.11545

Number	Eigenvalue	Condition Index	
1	2.68075	1.00000	0.
2	2.20921	1.10156	0.
3	1.83409	1.20898	
4	1.56431	1.30908	
5	1.41046	1.37863	
6	1.28391	1.44498	0.
7	1.22942	1.47665	0.
8	1.19814	1.49580	0.
9	1.17284	1.51185	0.
10	1.16538	1.51668	0.
11	1.12642	1.54269	
12	1.08450	1.57222	0.
13	1.06229	1.58857	

Looking at the residual graphs I suspected the presence of heteroscedasticity and thus tested for it using the whites test. Since the p value is very very small and highly significant and heteroscedasticity consistent standard errors and t-values are higher than original values, we reject the null hypothesis that all regression assumptions are satisfied and prove the presence of heteroscedasticity. So I build a regression model that would fit the model based on the robust standard error.

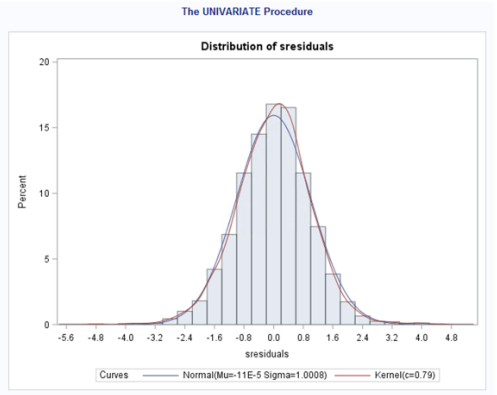


Root MSE	0.96457	R-Square	0.5185
Dependent Mean	-1.20100	Adj R-Sq	0.5151
Coeff Var	-80.31394		

Parameter Estimates								
						Heteroscedasticity Consistent		
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standard Error	t Value	Pr > |t|
Intercept	1	-3.28034	0.09698	-33.82	<.0001	0.10652	-30.80	<.0001
Rating_T	1	-0.12976	0.05308	-2.44	0.0145	0.05249	-2.47	0.0135
Rating_M	1	-0.29296	0.06320	-4.64	<.0001	0.06079	-4.82	<.0001
Rating_E	1	0.21234	0.05473	3.88	0.0001	0.05739	3.70	0.0002
Platform_DS	1	2.04375	0.08109	25.20	<.0001	0.09506	21.50	<.0001
Platform_GBA	1	2.23947	0.10071	22.24	<.0001	0.12096	18.51	<.0001

Last I also conducted a test for normality using proc univariate. Looking at the Goodness-of-Fit Tests for Normal Distribution we can clearly state that no matter which test we perform, the pvalues are all significant hence we reject the null that the distribution is normal.

Tests for Normality				
Test	Statistic		p Value	
Kolmogorov-Smirnov	D	0.017913	Pr > D	<0.0100
Cramer-von Mises	W-Sq	0.507298	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	3.379899	Pr > A-Sq	<0.0050



I then used only those points for the proc univariate whose cook d values < 4/n which gave me a nice almost normal distribution of errors shown above. Coming to the end I ran an autoreg that will estimate and forecast when the errors are autocorrelated or heteroscedastic.

The GLM Procedure

Dependent Variable: log_global_sales

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	35	4129.286027	117.979601	177.77	<.0001
Error	3904	2590.913987	0.663656		
Corrected Total	3939	6720.200014			

R-Square	Coeff Var	Root MSE	log_global_sales Mean
0.614459	-68.49612	0.814651	-1.189339

Source	DF	Type I SS	Mean Square	F Value	Pr > F
Rating	3	75.681520	25.227173	38.01	<.0001
Platform	9	1495.607441	166.178605	250.40	<.0001
Genre	11	91.777425	8.343402	12.57	<.0001
Weighted_Score	1	219.190811	219.190811	330.28	<.0001
Critic_Score	1	1306.956196	1306.956196	1969.33	<.0001
User_Score	1	18.630524	18.630524	28.07	<.0001
User_Count	1	367.332349	367.332349	553.50	<.0001
Critic_Count	1	502.855779	502.855779	757.71	<.0001
Generation	1	30.585530	30.585530	46.09	<.0001
Generation*Platform	6	20.668453	3.444742	5.19	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Rating	3	71.8314090	23.9438030	36.08	<.0001
Platform	9	773.1260855	85.9028984	129.44	<.0001
Genre	11	126.9363225	11.5396657	17.39	<.0001
Weighted_Score	1	458.4502219	458.4502219	690.79	<.0001
Critic_Score	1	220.0811859	220.0811859	331.62	<.0001
User_Score	1	10.5174601	10.5174601	15.85	<.0001
User_Count	1	164.3680081	164.3680081	247.67	<.0001
Critic_Count	1	524.9235415	524.9235415	790.96	<.0001
Generation	1	13.4989552	13.4989552	20.34	<.0001
Generation*Platform	6	20.6684532	3.4447422	5.19	<.0001

Parameter	Estimate		Standard Error	t Value	Pr > |t|
Intercept	-4.120954832	B	0.10696691	-38.53	<.0001
Rating E	0.200342168	B	0.04836176	4.14	<.0001
Rating M	-0.323182422	B	0.05513799	-5.86	<.0001
Rating T	-0.166211018	B	0.04639657	-3.58	0.0003
Rating E10+	0.000000000	B	.	.	.
Platform DS	2.476259570	B	0.13110996	18.89	<.0001
Platform GBA	2.843340624	B	0.11057964	25.71	<.0001
Platform GC	2.605395967	B	0.10212569	25.51	<.0001
Platform PS2	3.079556991	B	0.09345390	32.95	<.0001
Platform PS3	3.192792176	B	0.25232948	12.65	<.0001
Platform PSP	2.722092437	B	0.11614958	23.44	<.0001

- Analysis of variance: Model is significant since p<0.0001
- How well the model performed: R square = 0.614 which is not very good yet providing decent results than predicting the average value if at random.

- Anova tests: The second table gives us a clear understanding of what is the combined effect of the variables used even if categorical. We can state that since all these have p value less than 0.05, they are significant.
- Individual parameter estimates table provides us for a way to analyze if the variables have the correct effect or not. For example, we can say that since a video game which is suitable for everyone will be purchased the most and hence have the highest global sales among e, m, t, e10+ ratings. The sign for this coefficient is positive indicating that if there is one more purchase of the e rating video game then the sales will increase by 0.20 additionally than the other ratings.

Regression Assumptions and their violations

1. Outliers and influential observations

   Looking at the user scores and count scores, they had major outliers and influential points due to which we were getting an R-square of 0.02 – 0.03 in the beginning.
   Ideally if the user scores are high then the global sales should also be high. But the coefficient has a negative sign which is strange. Thus, we haven't yet handled the influential points.

   Detection:
   - By looking at the graphs
     Residual * Predicted values – Observations ideally should be near zero
     Rstudent * Pred values – observations should fall between 2 standard deviations
     Rstudent * leverage – observations should be low x and low y values
     QQ plot – points should be approximately around the normal distribution
     If any of the above is violated, then outliers and influential exist.
   - Test with the cook d value. If cook d > 4/n where n is the number of observations, then these points must be checked.

   Resolution:
   Either drop the points that had the cook d >4/n or use the robust standard errors that will minimize the effect of these points by giving different weights to different observations based on their influence.

   I tried both. Using robust regression dropped the r square and introduced funny coefficients. Hence, I dropped these points not affecting the R square much. The variance of these two columns was huge and drastically pulled the regression line away from the actual normal. Hence removing these was a very crucial step.

2. Multicollinearity among multiple variables

   When I performed a test with the interaction effect, the independent segments had a p-value not significant whereas its interaction effect was significant. I tested the violation of this assumption but everything seemed fine.

   Detection of this will tell us how much the estimates of other parameters will change when I drop this one variable.

   Detection:
   - If the collinearity is pairwise then, before using regression by developing a correlation matrix. If the correlation coefficient is above 0.7 then highly correlated whereas if it is less than 0.3 then they are comfortably independent.
   - If more variables are correlated, then after regression is done by comparing the variance inflation factors. If VIF > 10 then multicollinear.

   Resolution:
   We can use the PCA. Number of variables will be equal to the number of pca generated. For the eigen values generated we can check the condition index, if it is less than or near 10 then they have a weak effect on the dependent but if the values are close to 100 then they have a large effect.

   A certain number of PCA can be selected based on the amount of variance they explain.
   There was no significant impact on my model.

3. Heteroscedasticity in error term

   The assumption states that the error term for all the observations must have the same variance. This assumption was violated in my case and I could detect it using the graphs and whites test.

   Detection:
   - Graphs: if there is a pattern in the variance of the points then heteroscedasticity exists.
   - Whites test
   - Standard residual plots

   Resolution:
   - If simple, then can be solved simply by taking the log or sqrt type of transformations.
   - If complex, then we need to use robust errors to fit the model.

   In my case since the dependent was heavily skewed, I had to use the log transformations, but the other variables were too complex. Looking at the residual graphs I suspected the presence of heteroscedasticity and thus tested for it using the whites test. Since the p value is very small and highly significant and heteroscedasticity consistent standard errors and t-values are

higher than original values, we reject the null hypothesis that all regression assumptions are satisfied and prove the presence of heteroscedasticity. I used the proc model method to fit these robust errors into the regression model. There was a significant impact on my model.

4. Normality of error

This is tested after fitting the regression model. I did face this issue and had to use the proc univariate method to resolve. The errors were normalized to a great extent.

Detection:

- Check the residuals in the QQ plot
- Shapirowilks test, Kolmogorov smirnov

Resolution:

- Fixing the outliers generally solves this issue too

Looking at the Goodness-of-Fit Tests for Normal Distribution we can clearly state that no matter which test we perform, the p-values are all significant hence we reject the null that the distribution is normal. I then used only those points for the proc univariate whose cook d values < 4/n which gave me a nice almost normal distribution of errors shown above.

There was a significant impact on my model.