STA302 Fall 2023 Methods of Data Analysis 1
Final Project Proposal (Part 1)

Names of Group Members	Contribution to Proposal
Mojan Majid	A2, C4, C5
Luis Rodrigo Nieto Pizano	B. Data Description, Justifications and Summary
David Runesha	A3, A4, C3
Jacob Rebelo	A1,C1,C2

A. Research Question and Supporting Literature

1. *What is the research question you will be studying in this project? Be sure to explicitly refer to the variables under study and avoid using vague language to describe your study question.*

The research question we will be studying in this project is how different variables both numerical and categorical affect the total sales of video games. In our data set we have 2 numerical variables “critic score” and “user score” which we will be using in our model along with 3 other categorical variables “platform”, “Genre” and “Year of release”. Our model will consist of these 5 different variables and it is our goal to see if there exists a significant relationship between our predictor variables as just described and our response variable “Video game sales”. That is we wish to research whether there is a specific relationship which can help us predict if a video game will have higher sales. Furthermore using these data points we wish to develop a model that can predict for us the sales of a video game based off the specific attributes of that game.

2. *Provide an explanation for why a linear regression model would allow you to answer your research question. What aspect of your fitted model would give you the answer.*

A linear regression model would be appropriate for our research question because we expect to see a linear relationship between our response variable (Global_Sales) and the predictors (Platform, Genre, User_Score, etc). The linear regression coefficients are the aspect of the model that would allow us to estimate the game sales given information about a game. We can do this by plugging in the values of the predictors into the equation for our fitted line.

3. *Provide proper citations for 3 peer-reviewed academic research articles related to your specific research question or your topic of interest. For each, describe how the results of the article relate to your research question. Further, rank each article on a scale of 1 to 3 (1=not useful, 2=slightly useful, 3=very useful) based on how useful the article is in providing insight into the population relationship you wish to estimate. Justify this ranking.*

Citation	Description, ranking and justification
Cox, J. (2014). "What Makes a Blockbuster Video Game? An Empirical Analysis of US Sales Data," *Managerial and Decision Economics, John Wiley & Sons, Ltd., vol. 35*(3), pages 189-198, April. https://doi.org/10.1002/mde.2608	This study primarily focuses on a wide variety of factors that affect the likelihood of a video game becoming a blockbuster game. The conclusion of this study showed that three main factors were consistently and significantly associated with a higher likelihood of a given game achieving a very large volume of sales. These three factors were publisher, platform, and review scores. Though this study provides very solid and relevant context, It is the most broad of the three studies, therefore I ranked it 1 out of 3 usefulness.
Babb, J., Terry, N., & Dana, K. (2013). The Impact Of Platform On Global Video Game Sales. *International Business & Economics Research Journal (IBER)*, *12*(10), 1273–1288. https://doi.org/10.19030/iber.v12i10.8136	This research article primarily focuses on video game sales by platform in the global market, specifically between 2006 through 2011. This research, like ours, aims to identify what aspects of the video game market have the greatest impact on sales, with the difference being that this research focuses on the platform of video games, one of our variables. With that said, I would rank this research article as a 3, as the most useful article of these 3 on providing insight and context regarding our research topic.
Sherrick, B., Schmierbach, M. (2016) The Effects of Evaluative Reviews on Market Success in the Video Game Industry. *The Computer Games Journal 5*, 185–194. https://doi.org/10.1007/s40869-016-0027-y	This research article primarily focuses on the connection between professional/peer reviews of video games, and market and commercial success of said game. The paper concludes by suggesting that professional reviews predicted sales more than almost any other variable they considered. Though not using the exact critic score metric that we are using, this study does observe the effects of published professional feedback and opinions on video game sales, And for that reason I would give it a 2 rank of usefulness.

4. *Provide the database/library where you located the above academic papers. List the search terms used to find these papers, in addition to the number of results for each search term.*

Database/library searched	Search terms used	Number of results for each
Google scholar	"factors that impact video game sales"	418,000
Jstor	"platform video game sales"	8,225
Google Scholar	"critic score impact video game sales"	32,100

B. Data Description, Justifications and Summary

1. *Provide the website from which your chosen data was obtained/downloaded.*

Website:**	https://www.kaggle.com/datasets/gregorut/videogamesales

*** If your data was obtained from a data repository (e.g. Kaggle, UCI Repository, etc.), please state how your research question differs from the original purpose of these data.*

In the original dataset the provider didn't have any research question or specific purpose designated for this data. Also, the codes and dataset notebooks shared by other users differ from the objective that we are going to give it, that is, there is no evidence that other users have used this dataset for linear regression models.

2. *List the variables you have selected to be part of your preliminary model (minimum of 5 with at least one a categorical variable). Please give an understandable name to each variable rather than writing the name that appears in R.*

For each variable, justify why you have chosen to use this variable over others in the dataset, and what the role of each variable will be (e.g., predictor of interest, predictor informed by literature, confounder, etc.).

Variable Name	Justification for Use	Role in Model
Global_Sales	This numerical variable is the one that we are trying to predict and explain based on the predictor variables.	Response
Platform	We expect that this categorical variable will predict the value of the response, combined with all the other response variables	Predictor

Year_of_Release	We expect that this categorical variable will predict the value of the response, combined with all the other response variables	Predictor
Genre	We expect that this categorical variable will predict the value of the response, combined with all the other response variables	Predictor
Critic_Score	We expect that this numerical variable will predict the value of the response, combined with all the other response variables	Predictor
User_Score	We expect that this numerical variable will predict the value of the response, combined with all the other response variables	Predictor

3. *Produce a table of numerical summaries of the variables listed above. Summaries should be appropriate to the type of variable, and interesting/important characteristics about variables should be mentioned in an informative caption. Include your summary table below.*

Variable	Minimum value	1st Quantile	Median	Mean	3rd Quantile	Maximum value
Global_Sales	0.0100	0.1100	0.2900	0.7673	0.7500	82.5300
Critic_Score	13.00	62.00	72.00	70.25	80.00	98.00
User_Score	0.500	6.500	7.500	7.183	8.200	9.600

Platform	3DS	DC	DS	GBA	GC	PC	PS	PS2	PS3	PS4	PSP
Percentage	2.26	0.20	6.74	3.43	5.05	9.94	2.20	16.54	11.24	3.61	5.66
Platform	PSV	Wii	WiiU	X360	XB	XOne					
Percentage	1.71	6.96	1.29	12.49	8.21	2.39					

Year	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006
Percentage	0.11	0.20	0.37	0.43	1.48	3.71	6.60	7.24	6.92	8.15	7.66
Year	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	
Percentage	8.56	8.63	8.04	6.25	6.76	4.65	3.94	3.71	3.20	3.29	

Genre	Action	Adventure	Fighting	Misc	Platform	Puzzle	Racing	Role - playing
Percentage	23.86	3.83	5.48	5.60	5.84	1.71	8.50	10.37
Genre	Shooter	Simulation	Sports	Strategy				

Percentage	12.58	4.35	13.80	4.03				

C. Preliminary Model Results

1. *Fit your preliminary multiple linear model and present the estimated relationship. Present this information carefully so that it is easily readable and understandable.*

Coefficient	Estimate	Std.Error	T-Value	P-Value
(Intercept)	**-0.065372888**	0.711742881	-0.09184902	9.27E-01
Critic_Score	**0.045227998**	0.002109033	21.44490048	7.60E-99
User_Score	**-0.118617568**	0.020554051	-5.77100667	8.22E-09
PlatformDC	**-1.541934683**	0.548862987	-2.80932531	4.98E-03
PlatformDS	**0.037957268**	0.185492313	0.20462987	8.38E-01
PlatformGBA	**-0.547137197**	0.217923896	-2.51068014	1.21E-02
PlatformGC	**-0.660190898**	0.206431477	-3.19811159	1.39E-03
PlatformPC	**-0.896108201**	0.17076428	-5.24763259	1.59E-07
PlatformPS	**-0.222876401**	0.288199944	-0.7733395	4.39E-01
PlatformPS2	**-0.193641939**	0.185591381	-1.04337786	2.97E-01
PlatformPS3	**-0.044308859**	0.168120075	-0.26355484	7.92E-01
PlatformPS4	**0.223126477**	0.206320644	1.08145493	2.80E-01
PlatformPSP	**-0.426726184**	0.189955905	-2.24644863	2.47E-02
PlatformPSV	**-0.558455663**	0.225547732	-2.47599768	1.33E-02
PlatformWii	**0.640067885**	0.182693782	3.50350119	4.62E-04
PlatformWiiU	**-0.196331247**	0.247885948	-0.7920225	4.28E-01
PlatformX360	**-0.006204787**	0.16860496	-0.03680074	9.71E-01
PlatformXB	**-0.775696946**	0.197148415	-3.93458373	8.42E-05
PlatformXOne	**-0.117420312**	0.221248002	-0.53071807	5.96E-01
Year_of_Release1997	**0.070255445**	0.815994558	0.08609793	9.31E-01
Year_of_Release1998	**-0.438964703**	0.743853378	-0.59012262	5.55E-01
Year_of_Release1999	**-0.37966781**	0.734092919	-0.51719312	6.05E-01
Year_of_Release2000	**-1.099051811**	0.678837457	-1.61902057	1.05E-01
Year_of_Release2001	**-0.900426623**	0.676042165	-1.33190897	1.83E-01
Year_of_Release2002	**-1.042158689**	0.679591499	-1.53350754	1.25E-01
Year_of_Release2003	**-1.15991957**	0.67929963	-1.70752275	8.78E-02
Year_of_Release2004	**-1.019960599**	0.679175759	-1.50176237	1.33E-01
Year_of_Release2005	**-1.153708414**	0.678175342	-1.70119487	8.90E-02
Year_of_Release2006	**-1.054839108**	0.678960882	-1.55360807	1.20E-01
Year_of_Release2007	**-1.293877913**	0.679497962	-1.90416747	5.69E-02
Year_of_Release2008	**-1.271536931**	0.68001167	-1.86987516	6.15E-02

Year_of_Release2009	**-1.312599427**	0.680085457	-1.93005072	5.36E-02
Year_of_Release2010	**-1.142152105**	0.681672638	-1.67551408	9.39E-02
Year_of_Release2011	**-1.255331895**	0.68154365	-1.84189508	6.55E-02
Year_of_Release2012	**-1.264039616**	0.684264319	-1.84729728	6.47E-02
Year_of_Release2013	**-1.173670097**	0.686418689	-1.70984578	8.73E-02
Year_of_Release2014	**-1.376249228**	0.688817856	-1.99798716	4.58E-02
Year_of_Release2015	**-1.518372541**	0.693836027	-2.18837374	2.87E-02
Year_of_Release2016	**-1.837220337**	0.693999079	-2.64729506	8.13E-03
Year_of_ReleaseN/A	**-1.433886891**	0.692310753	-2.0711608	3.84E-02
GenreAdventure	**-0.309613748**	0.122323738	-2.53110109	1.14E-02
GenreFighting	**-0.190751023**	0.104753222	-1.82095614	6.87E-02
GenreMisc	**0.162063903**	0.103777383	1.56164954	1.18E-01
GenrePlatform	**0.096409492**	0.103419439	0.93221829	3.51E-01
GenrePuzzle	**-0.356451153**	0.177666108	-2.00629798	4.49E-02
GenreRacing	**0.016136813**	0.088673173	0.18198077	8.56E-01
GenreRole-Playing	**-0.176370571**	0.083674544	-2.10781634	3.51E-02
GenreShooter	**0.146300951**	0.077454331	1.88886726	5.90E-02
GenreSimulation	**-0.07282772**	0.116642361	-0.62436768	5.32E-01
GenreSports	**-0.160469782**	0.076563463	-2.0959055	3.61E-02
GenreStrategy	**-0.41328769**	0.122635347	-3.37005357	7.56E-04

2. *Justify your choice of how you included the categorical variable in your preliminary model. How does this choice contribute to answering your research question?*

Our dataset has many categorical variables and each with many different responses as seen in the table above. Our choice to include these predictor variables is justified because as we can see many have very small p values which indicate they are statistically significant to our model. Each one of these categorical predictors offers us insight to help answer our research question about what contributes to making a certain video game high selling. That is, our categorical variables of Genre, Year of release and platform offer statistically significant evidence towards our model of determining the sales of a video game.

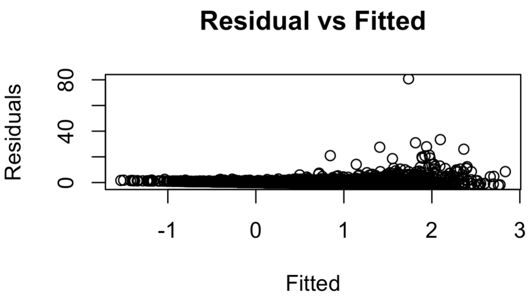
3. *Do your estimated coefficients align/agree with the results of your three peer-reviewed articles? Explain in what way they differ/agree and provide a reason why this might be the case.*

4. *Perform a complete assessment of the assumptions of your preliminary model. Do you observe violations of assumptions or conditions? Describe how you came to this conclusion, making explicit reference to any plots or other information that is relevant.*

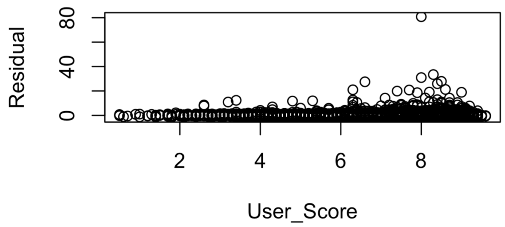
We observe a violation of the normality assumption based on the deviation and curving from the diagonal line that occurs in the QQ plot. We also have some evidence of a violation of the constant variance assumption due to the increase of the spread shown in the residual vs fitted, residual vs user_score, and residual vs critic_score plot. Our plots don’t indicate any signs of violation of the linearity or uncorrelated errors assumptions.

Based on the roughly random scatter around the diagonal line in the scatterplot of our response versus fitted values (refer to Rmd file) the first condition holds.
Based on the predictor scatterplots (refer to Rmd file) there are no non-linear patterns present so the 2nd condition holds as well.

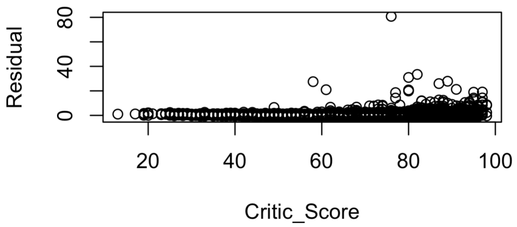
5. *Include all relevant plots created for assessing model assumptions below, with appropriate axis labels and captions.*



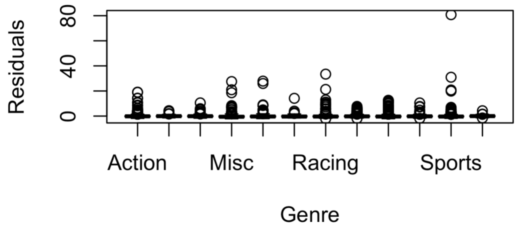
Residual vs User_Score



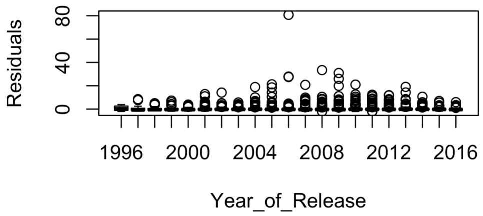
Residual vs Critic_Score



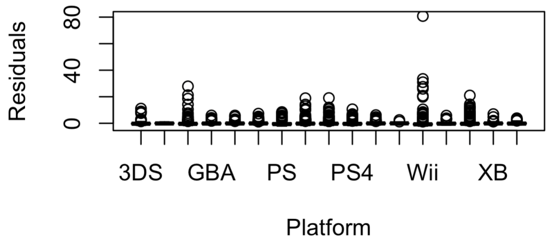
Residual vs Genre



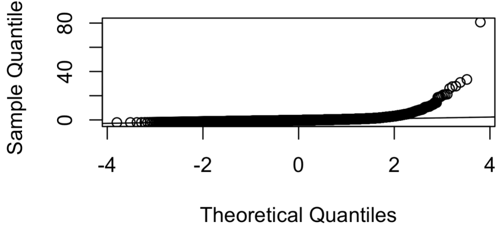
Residuals vs Year_of_Release



Residual vs Platform



Normal Q-Q Plot


Sample Quantiles
Theoretical Quantiles