# Video Game Regression Modelling Project

**Nat Berryman** 

### Summary

The purpose of this regression model is to predict the Global Sales of Video Games by analysing the Video Game Sales dataset from Kaggle.

#### **Outline**

- Business Problem
- Cleaning the Data
- Exploratory Data Analysis
- Models
- Conclusions

#### **Business Problem**

How can we predict the Global Sales of Video Games?

In order to solve this problem, I intended to answer the below questions:

- 1. Does console impact global sales?
- 2. Does genre impact global sales?
- 3. Does publisher impact global sales?

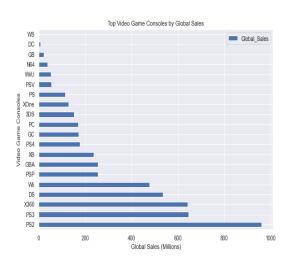
## Cleaning the Data

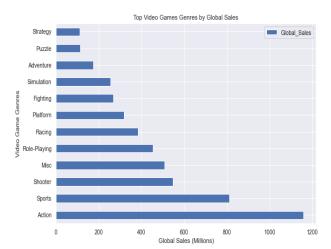
- Dropped unnecessary data
- Replaced or removed null values
- Narrowed data to only include sales between 2010-2016
- Using the empirical formula I removed outliers
- Addressed multicollinearity
- Split data set between continuous and categorical data
- Grouped Publisher data into Small, Medium and Large
- Created additional variables including total engagement and average score

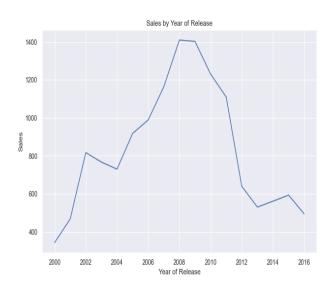
- No key features stood out that showed positive correlation with Global Sales.
- Further analysis required to identify any positive correlation with Global Sales.

ta /	<b>~</b> !	IU	ı y	J	<b>3</b>														
Genre_Adventure	1	-0.065	-0.11	-0.069	-0.056	-0.084	-0.095	-0.087	-0.072	-0.12	-0.061	-0.02	0.077	-0.07	0.029	-0.016	0.1	8.4e-05	-0.2
Genre_Fighting	-0.065	1	-0.076	-0.049	-0.04	-0.06	-0.068	-0.062	-0.051	-0.086	-0.043	-0.056	0.066	0.015	0.041	-0.051	-0.019	0.0081	0.032
Genre_Misc	-0.11	-0.076	1	-0.08	-0.065	-0.097	-0.11	-0.1	-0.083	-0.14	-0.07	0.12	-0.05	-0.037	-0.016	-0.074	0.078	-0.012	-0.0045
Genre_Platform	-0.069	-0.049	-0.08	1	-0.042	-0.063	-0.071	-0.065	-0.054	-0.091	-0.046	0.091	-0.047	-0.028	-0.0037	-0.045	-0.064	-0.0046	0.05
Genre_Puzzle	-0.056	-0.04	-0.065	-0.042	1	-0.052	-0.058	-0.054	-0.044	-0.074	-0.037	0.15	-0.092	-0.067	-0.0087	-0.0089	0.015	-0.00071	-0.067
Genre_Racing	-0.084	-0.06	-0.097	-0.063	-0.052	1	-0.087	-0.08	-0.066	-0.11	-0.056	-0.05	0.0096	0.06	0.0046	-0.01	-0.098	-0.028	0.03
Genre_Role_Playing	-0.095	-0.068	-0.11	-0.071	-0.058	-0.087	1	-0.09	-0.074	-0.12	-0.063	-0.03	0.067	-0.064	0.0066	0.018	0.079	0.074	0.0062
Genre_Shooter	-0.087	-0.062	-0.1	-0.065	-0.054	-0.08	-0.09	1	-0.068	-0.11	-0.058	-0.11	-0.019	0.12	-0.0021	0.08	0.013	-0.0046	0.057
Genre_Simulation	-0.072	-0.051	-0.083	-0.054	-0.044	-0.066	-0.074	-0.068	1	-0.095	-0.048	0.083	-0.086	-0.047	-0.011	0.08	0.0043	0.0041	-0.0062
Genre_Sports	-0.12	-0.086	-0.14	-0.091	-0.074	-0.11	-0.12	-0.11	-0.095	1	-0.08	-0.043	0.03	0.06	-0.0011	-0.063	-0.11	0.016	0.079
Genre_Strategy	-0.061	-0.043	-0.07	-0.046	-0.037	-0.056	-0.063	-0.058	-0.048	-0.08	1	-0.041	-0.034	-0.036	-0.0094		0.028	0.027	-0.071
Nintendo	-0.02	-0.056	0.12	0.091	0.15	-0.05	-0.03	-0.11	0.083	-0.043	-0.041	1	-0.63	-0.34	-0.037	-0.21	0.00018	-0.045	-0.028
Sony	0.077	0.066	-0.05	-0.047	-0.092	0.0096	0.067	-0.019	-0.086	0.03	-0.034	-0.63		-0.35	-0.038	-0.21	0.016	0.033	0.045
Microsoft	-0.07	0.015	-0.037	-0.028	-0.067	0.06	-0.064	0.12	-0.047	0.06	-0.036	-0.34	-0.35	1	-0.02	-0.11	-0.079	-0.028	0.09
Sega	0.029	0.041	-0.016	-0.0037	-0.0087	0.0046	0.0066	-0.0021	-0.011	-0.0011	-0.0094	-0.037	-0.038	-0.02	1	-0.012	0.013	0.026	-0.0084
PC	-0.016	-0.051	-0.074	-0.045	-0.0089	-0.01	0.018	0.08	0.08	-0.063		-0.21	-0.21	-0.11	-0.012	1	0.081	0.061	-0.16
Total_Engagement	0.1	-0.019	0.078	-0.064	0.015	-0.098	0.079	0.013	0.0043	-0.11	0.028	0.00018	0.016	-0.079	0.013	0.081	1	0.29	-0.073
Total_Score	8.4e-05	0.0081	-0.012	-0.0046	-0.00071	-0.028	0.074	-0.0046	0.0041	0.016	0.027	-0.045	0.033	-0.028	0.026	0.061	0.29	1	0.13
Global_Sales_log	-0.2	0.032	-0.0045	0.05	-0.067	0.03	0.0062	0.057	-0.0062	0.079	-0.071	-0.028	0.045	0.09	-0.0084	-0.16	-0.073	0.13	1
	2	9	9	ε	9	9	9	in .	9	20	à		₽	45	9	٧	t	2	9

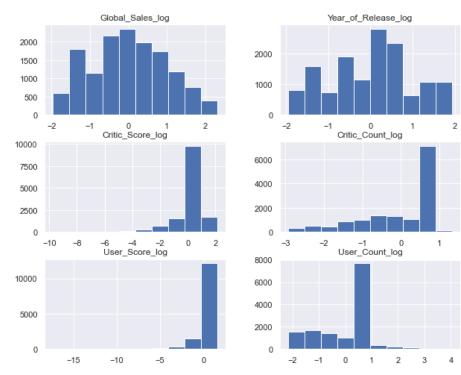
- Global Sales for Video Games are highest on the PS2 console while Action games are the highest selling following by Sports and Shooter
- Using the platform visualization I modelled two sets of data, individual platforms and then grouping platform together by maker e.g. PS, PS2 and PS3 under Sony and XOne, X360 under Microsoft
- Number of games peaked in 2008 and 2009 but had declined since







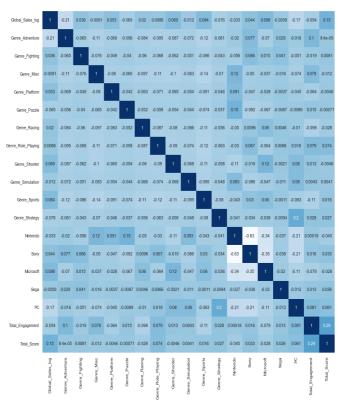
- Used mean normalization to standardise the data
- All features appeared ok however the critic and user variables are somewhat negatively skewed

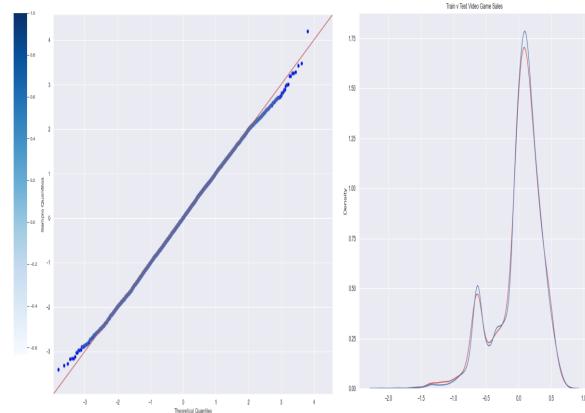


Unfortunately no models could provide concrete recommendations

Model / Features	R^2	Train MSE	Test MSE
Model 1: All Features	0.28	0.73	0.76
Model 2: All Features minus Publisher	0.19	0.80	0.82
Model 3: Genre, Grouped Platform, Total Engagement, Total Score	0.12	0.88	0.90
Polynomial 1st Degree: All features	0.12	-	-
Polynomial 2 <sup>nd</sup> Degree: All features	0.21	-	-
Polynomial 3 <sup>rd</sup> Degree: All features	-8.02	-	-

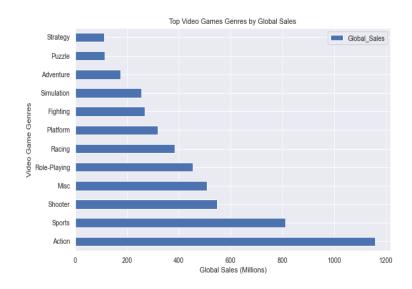
#### Model 1

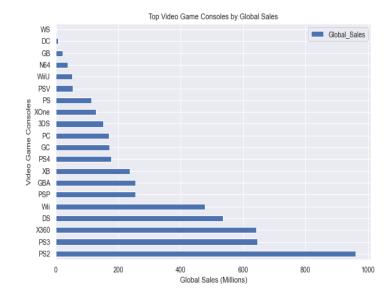




#### **Conclusions**

- While there is highly performing genres, platforms and publishers the Global Sales can't be accurately predicted using OLS and Polynomial Regression modelling
- It is suggested that advanced regression models are used to drive decisions
- It is however recommended that insights are used for focus areas such has highly performing genres, engagement and consoles





# Thank You!

Email: nathaniel.berryman@gmail.com

GitHub: @natberr

LinkedIn: linkedin.com/in/nathaniel-berryman/