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:

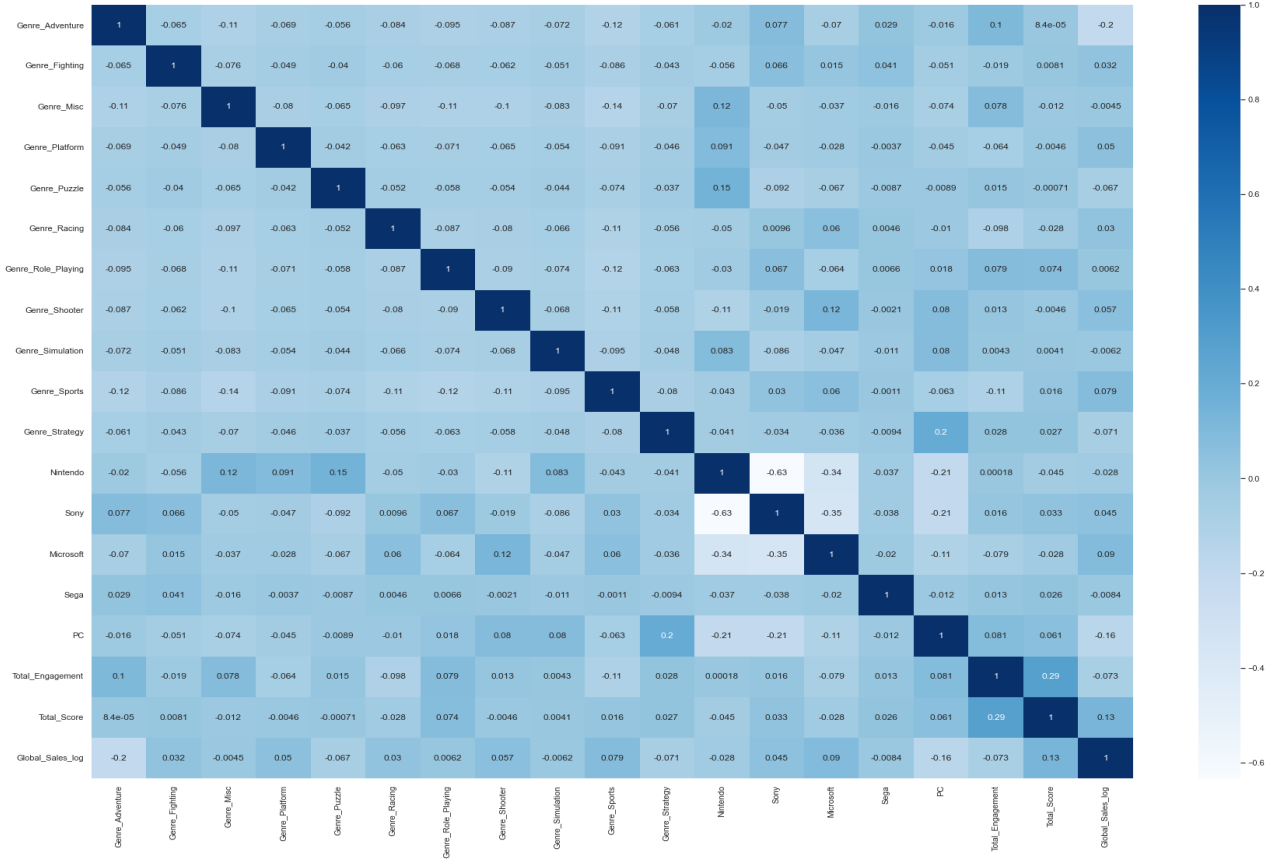
1. Does console impact global sales?
2. Does genre impact global sales?
3. Does publisher impact global sales?

Cleaning the Data

- Dropped unnecessary data
- Replaced or removed null values
- Narrowed data to only include sales between 2000-2016
- 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

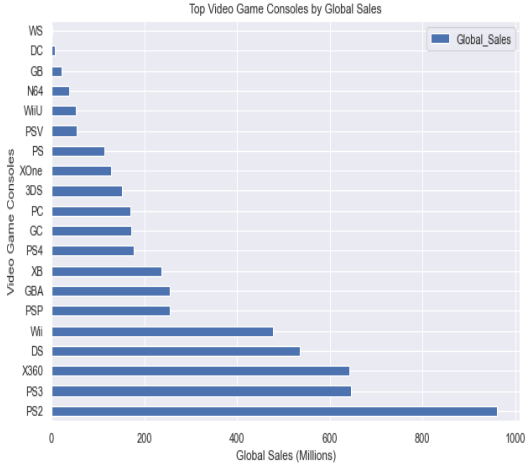
Exploratory Data Analysis

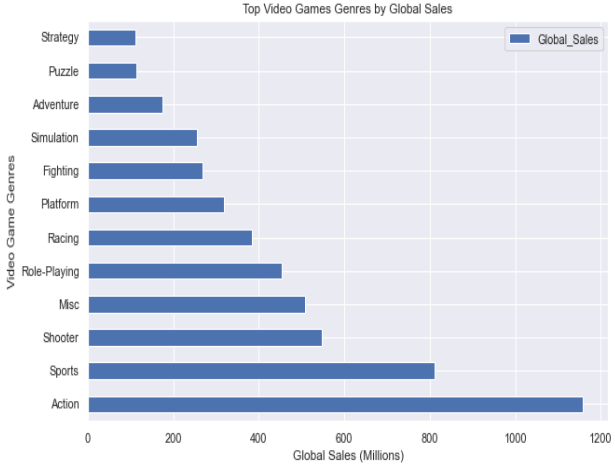
- No key features stood out that showed positive correlation with Global Sales.
- Further analysis required to identify any positive correlation with Global Sales.

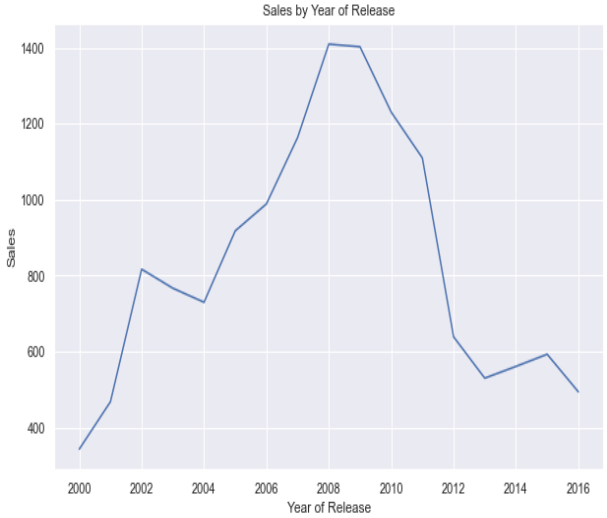


Exploratory Data Analysis

- 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

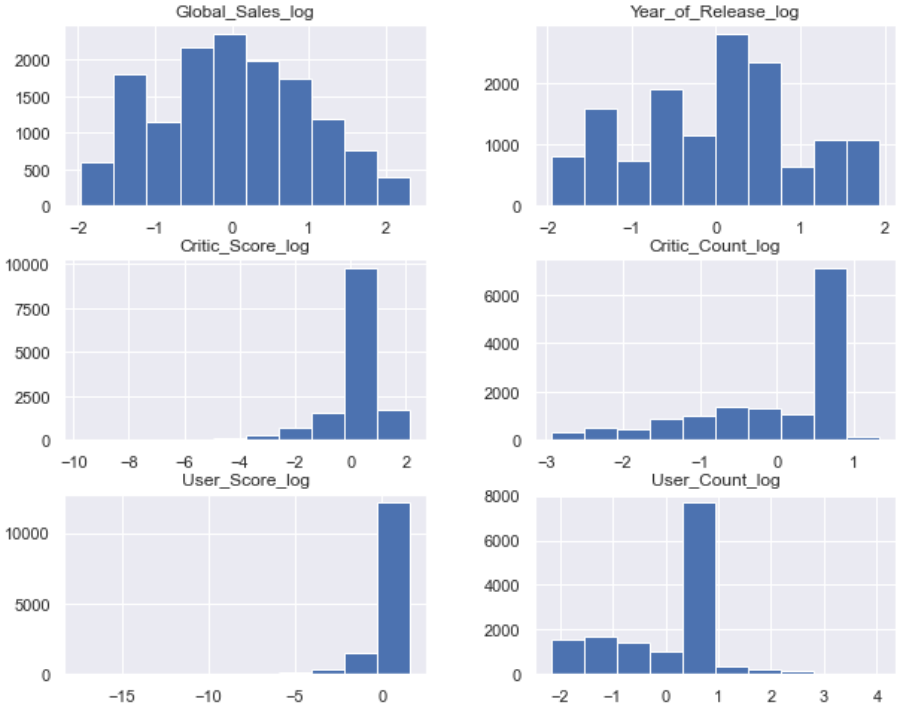






Exploratory Data Analysis

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

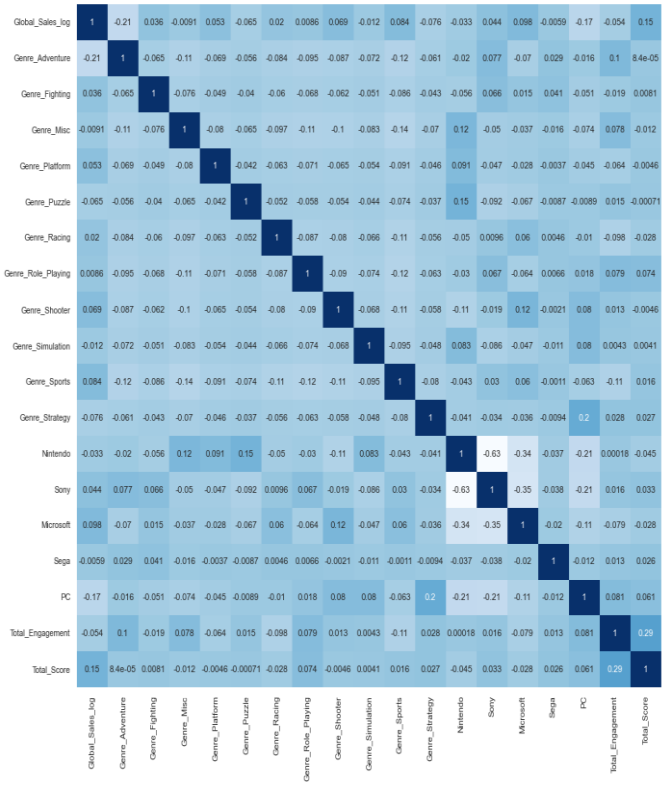


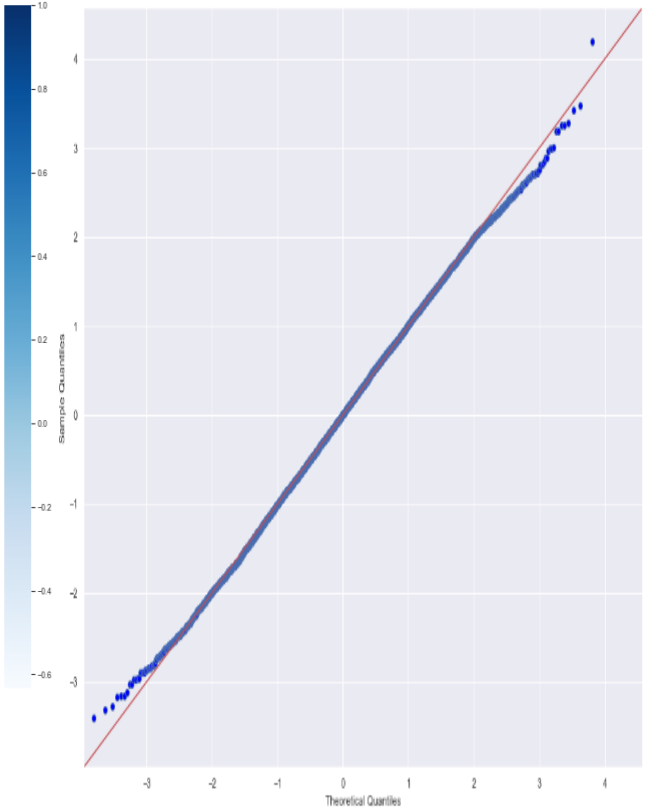
Exploratory Data Analysis

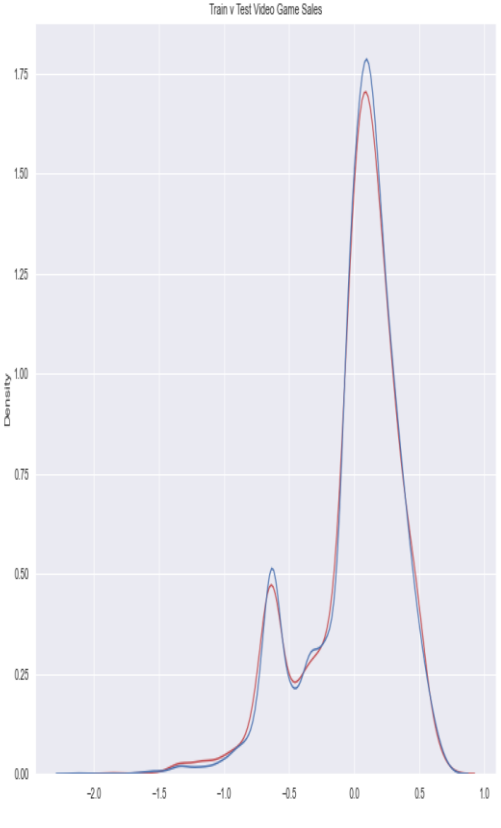
- Unfortunately no models could provide concrete recommendations

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 2nd Degree: All features	0.21	-	-
Polynomial 3rd Degree: All features	-8.02	-	-

Model 1

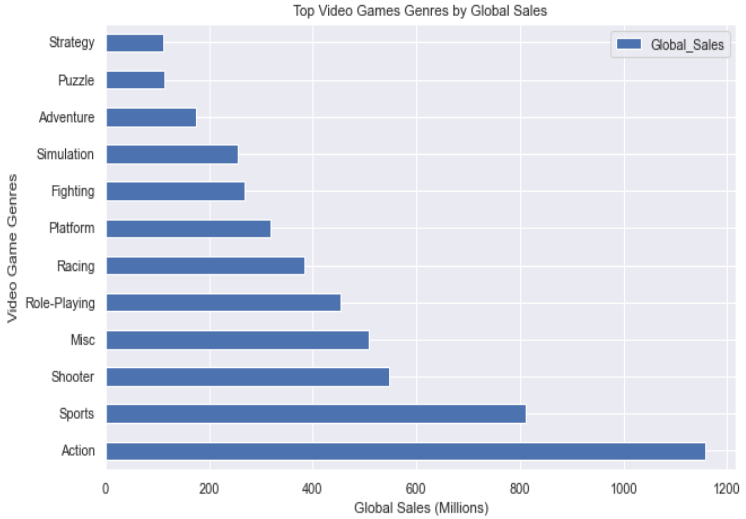

Global_Sales_log
Genre_Adventure
Genre_Fighting
Genre_Misc
Genre_Platform
Genre_Puzzle
Genre_Racing
Genre_Role_Playing
Genre_Shooter
Genre_Simulation
Genre_Sports
Genre_Strategy
Nintendo
Sony
Microsoft
Sega
PC
Total_Engagement
Total_Score

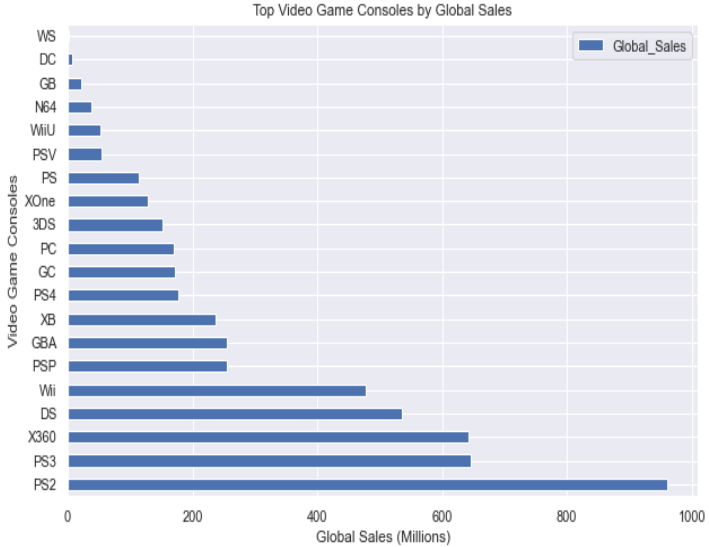

Sample Quantiles
Theoretical Quantiles


Train v Test Video Game Sales
Density

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/