# Exploring Video games sales data (2003-2004) using R programming language

**Explanation of dataset:** This dataset was provided by the instructor for one of the data analysis course and is based off of real world data. Data dictionary can be found in the appendix.

**Purpose:** To determine the relationship between Life to date sales and various other variables to find how to maximize revenue. The purpose is not to find a real-world root cause but to use my skills in R and data analysis on a real-world dataset.

**Skills and tools used:**

* R for data analysis
* Data cleaning and recoding
* Tidyverse, Pacman, rio, Janitor and QuantPsyc packages
* Data wrangling using filter and mutate command for recoding and computing variables
* Multiple and linear regression using scatter plots, correlation matrix

**Verifying the assumptions** for the validity of multiple linear regression:

We asked R to give us some diagnostic plots for our regression results that would help to verify if our regression model meets the assumptions for a valid linear regression.

Chart, scatter chart

Description automatically generated Chart, line chart

Description automatically generated

1. **Relation between X and Y is linear (linearity)** : If the relation was linear we would have seen an almost straight line in the residuals v/s fitted plot. Hence, this assumption is not satisfied as of now.
2. The **residuals are normally distributed:** As seen from the residual-fit plot and the Quantile-Quantile (Q-Q) plot, residuals are not normally distributed. Hence, this assumption is also violated.
3. **There is no multi-collinearity in data:** Using the scatterplot matrix (full matrix in appendix) it was clear that, First year sales is highly correlated with life to date sales, but it has no business meaning since we cannot say that the root cause of success in video games is making more revenue in the first year. Secondly, another scatterplot shows that first year units sold and life to date units sold are correlated which creates the same problem of inadequate business meaning. Hence, for our regression analysis, first year sales, first year units sold and last year units sold would be removed to make sure we get accurate regression results and its assumptions are not violated.

**Chart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generated**

1. **The value of residuals are independent**: This can be verified using the Durban-western test in R.

**Text

Description automatically generated with low confidence**

Since the p value is less than 0.05, we reject the null hypothesis that the error terms are independent, which means we do not have enough evidence to believe that errors are independently distributed. Hence, this assumption is violated.

**Roadblock:** The assumptions for a valid regression model are being violated leading to a bias or invalid model.

**Solution:** We will try to use natural log to make the distribution more normally distributed and avoid overfitting the model since we have too many variables that might be skewing with the distributions of the residuals.

**Re-Verifying** the assumptions of linear regression again with the log-transformed response variable (Logged LTD dollars):

We produced another set of diagnostic plots with the log-transformed response variable:

Chart, scatter chart

Description automatically generated Chart

Description automatically generated

1. **Linearity:** As the residual v/s fitted plot indicates the line is much straighter than it was before and the skew in the line might be attributed to the presence of very large outliers, once dealt with the outliers, we might see a fairly straight line throughout the model.
2. **Normal distribution of residuals:** We can see that Q-Q plots gives us a fairly straight line for the residuals, means that residuals are now normally distributed.
3. **There is no multi-collinearity:** These assumptions are still valid as we do not have any correlated variables in the model.
4. **Independence of error terms:** We re-ran the Dublin-Watson test and it returned the following results: Text

   Description automatically generated with medium confidence

Since the p value is less than 0.05, we reject the null hypothesis that the error terms are independent, which means we do not have enough evidence to believe that errors are independently distributed. Hence, this assumption is violated. This indicates there might be a chance of bias in the model.

**Limitations of the model and analysis:** The dataset is derived from a real-world data, and it is quite complex to get the data to match all our expectations, such as not meeting all the assumptions of the regression analysis. Since my objective is to showcase the application of my skills in R programming language, I still decided to carry on with the analysis. Given an ideal environment with more sophisticated tools (such as tools to solve non-linear regression) we would have derived some more useful results but given the limited availability of tools and nature of our data I will try my best to do the best analysis as per my knowledge.

**Analysis:** Running the regression analysis will show the coefficients and significance of various variables that influence Life to date sales for videogames.

Regression results with Ltd dollars (sales revenue) as response variable and other relevant variables as explanatory variables:

A screenshot of a computer

Description automatically generated with low confidence A screenshot of a computer

Description automatically generated with low confidence Text

Description automatically generated

These estimates gives a high level overview of how much each variable impacts the Ltd dollars, however to compare these explanatory variables with each other requires standardized coefficent estimates, hence we’ll get R to give us the standardized estimates.

**Text

Description automatically generated with low confidence**

As the standardized estimates indicates, review and other consoles (GCN<ps2<xbx) ranks as the highest driving factors (in terms of standard deviation) for ltd dollars. These variables being the driving factor for ltd sales is counter-intuitive since, review is given after the game is released. It cannot be said for a game to do better, we need to make a game that will get better review from critics. XBOX, GCN, PS2 are the indicator variables showing whether the game was released on these platforms, which means that games will earn higher if they are released on GCN than ps2 or Xbox but that cannot be the only be the deciding factor for making business decisions.

Text

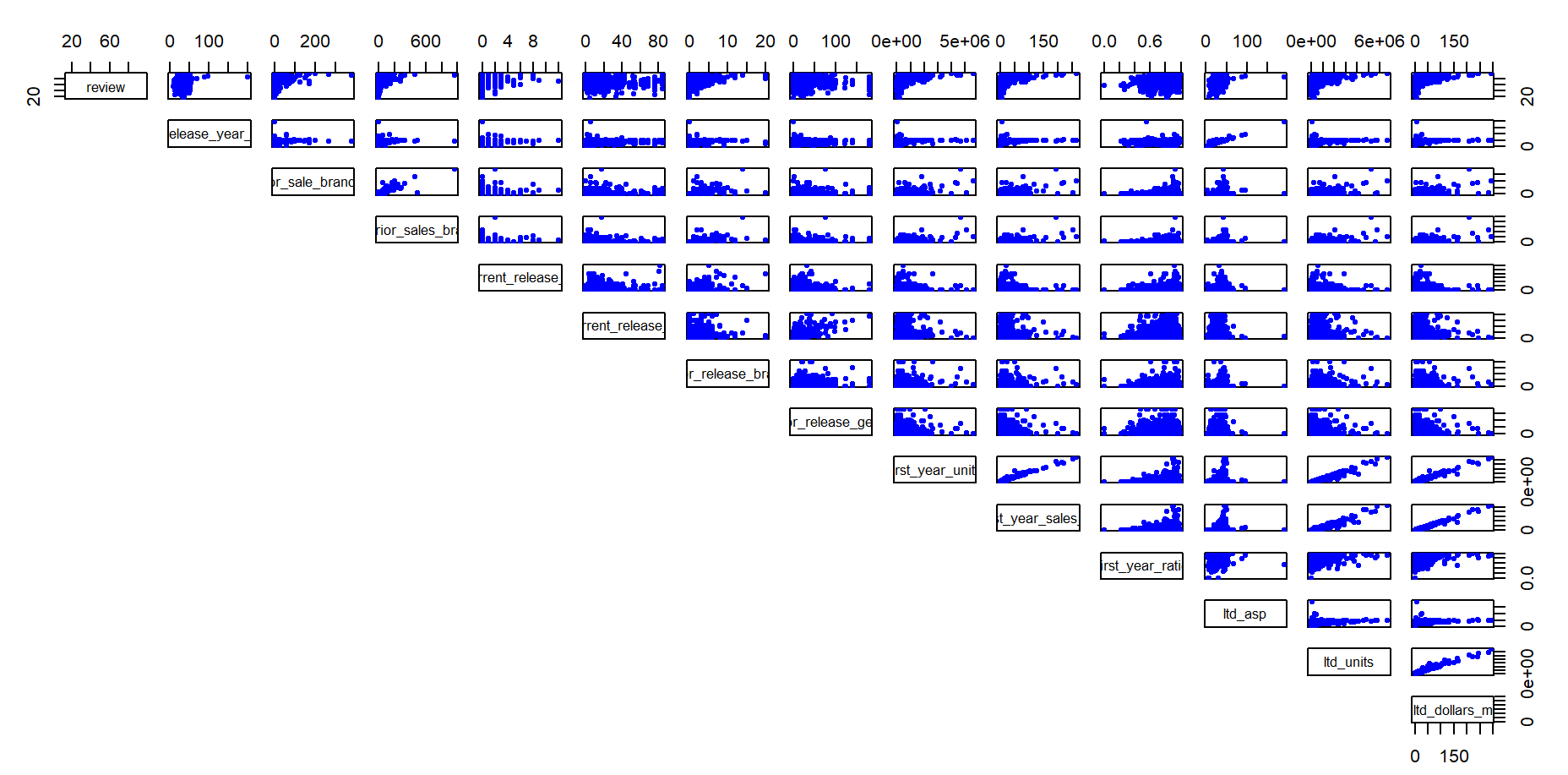
Description automatically generated with low confidence

The adjusted R-square is 0.5085 which means that the model explains almost 50% of variability in the response variable.

**Conclusion:** The results of the regression do not provide much insights into root cause of video game sales but gives an overview of how different variables affect the total life to date sales of video games. In business context, the video game companies look at these results and can eliminate any variables that the regression deems as insignificant and can further analyze the results to find useful patterns of data.

# Appendix

Correlation Matrix



Diagnostic plots for response variable before logging ltd dollars.

Graphical user interface, chart

Description automatically generated

Diagnostic Plots for log-transformed response variable:

A screenshot of a computer

Description automatically generated with low confidence