SC1015 Mini Project:

Video Game Sales Analysis

Bernard Chiang (U2120378C)
Chee Wen Zhan (U2122475L)
Tan Jia Ze (U2122410B)

» TABLE OF CONTENTS «

1 Motivation

2 Source

3 Cleaning

4 EDA

5 Model

6 Insight

» Problem Statement «

- Predict sale volume using game attributes
- Create a model to help new or aspiring game creators

» Dataset «

- Data was obtained from VGchartz
- Credible source with the latest data of video games sales
- Genre, Developer, Publisher of games etc.

» Cleaning «

	Rank	Name	Genre	Platform	Publisher	Developer	Vgchartz_Score	Critic_Score	User_Score	Total_Shipped	Total_Sales	NA_Sales	PAL_Sales
155	156	Marvel's Spider-Man	NaN	PS4	Sony Interactive Entertainment	Insomniac Games	8.0	9.1	NaN	20000000.0	NaN	NaN	NaN
167	168	God of War (2018)	NaN	PC	Sony Interactive Entertainment	SIE Santa Monica Studio	9.0	9.7	10.0	19500000.0	NaN	NaN	NaN
168	169	Grand Theft Auto V	NaN	PS4	Rockstar Games	Rockstar North	NaN	9.7	NaN	NaN	19390000.0	6060000.0	9710000.0
172	173	Brain Age: Train Your Brain in Minutes a Day	NaN	DS	Nintendo	Nintendo SDD	NaN	8.1	NaN	19010000.0	NaN	NaN	NaN

Cleaning

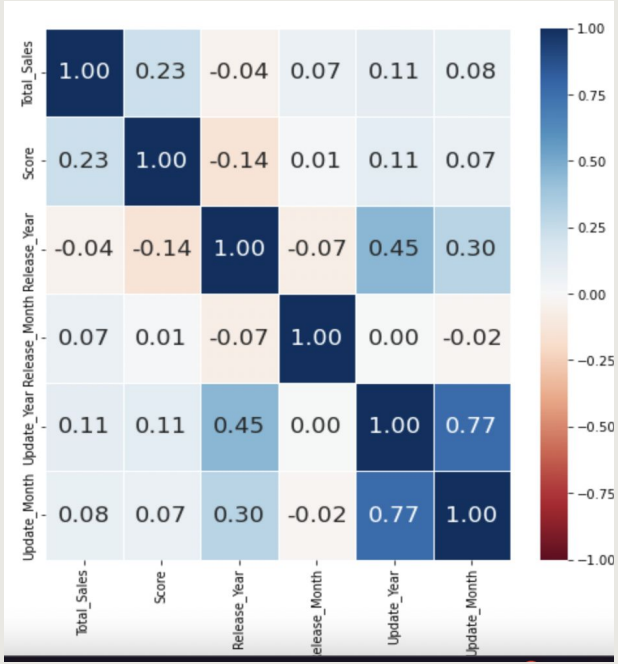
- Drop NULL values in mean Score & Total Sales columns
- Creating new columns Year, Month & Date
- Changing Release_Date & Last_Update into numerical values
- Adding Genre using github user *GregorUT*'s dataset

» Dataset «

	Name	Platform	Publisher	Developer	Total_Sales	Score	Release_Year	Release_Month	Update_Year	Update_Month	Genre
0	Wii Sports	Wii	Nintendo	Nintendo EAD	82900000.0	7.700000	2006	11	<NA>	<NA>	Sports
1	Super Mario Bros.	NES	Nintendo	Nintendo EAD	40240000.0	9.100000	1985	10	<NA>	<NA>	Platform
2	Mario Kart Wii	Wii	Nintendo	Nintendo EAD	37380000.0	8.666667	2008	4	2018	4	Racing
3	Wii Sports Resort	Wii	Nintendo	Nintendo EAD	33140000.0	8.533333	2009	7	<NA>	<NA>	Sports
4	New Super Mario Bros.	DS	Nintendo	Nintendo EAD	30800000.0	8.600000	2006	5	<NA>	<NA>	Platform

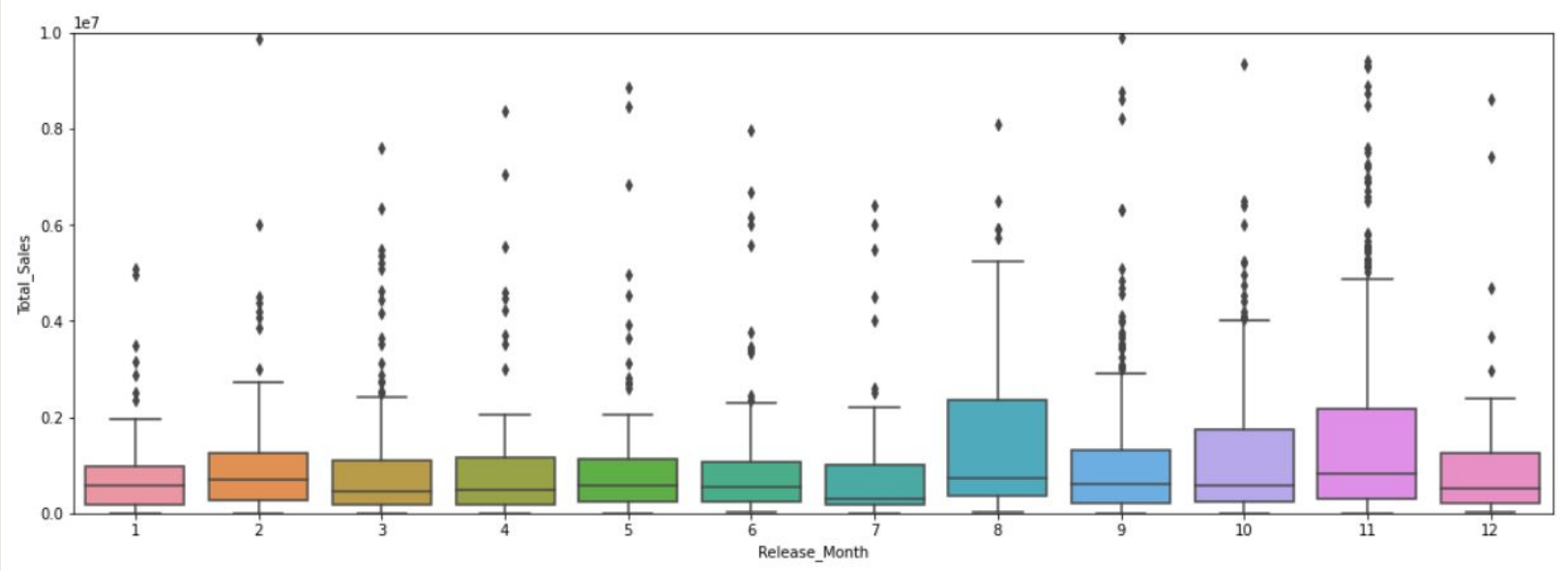
Exploratory Data Analysis

» EDA: Numeric Values «



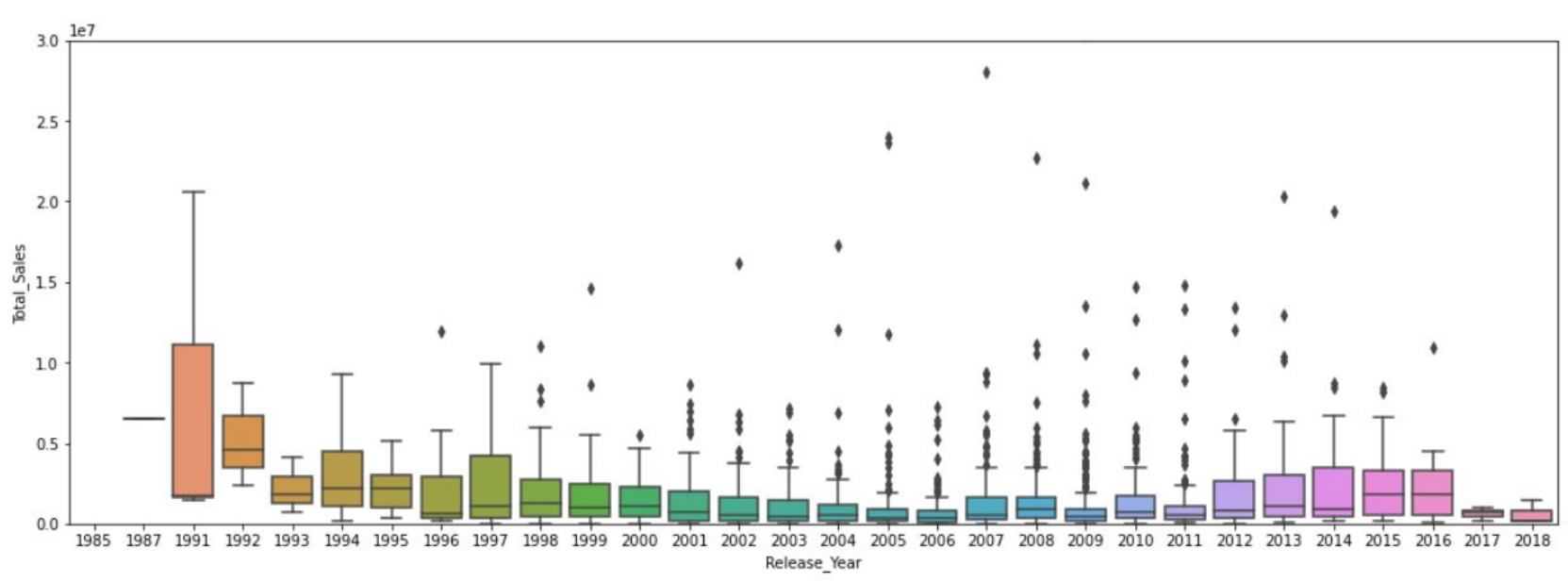
- High correlation between Update_Month and Update_ Year
- This coefficient is meaningless
- No other correlation observed between other datas
- Simple linear regression **may not be effective**

» EDA: Release Month «



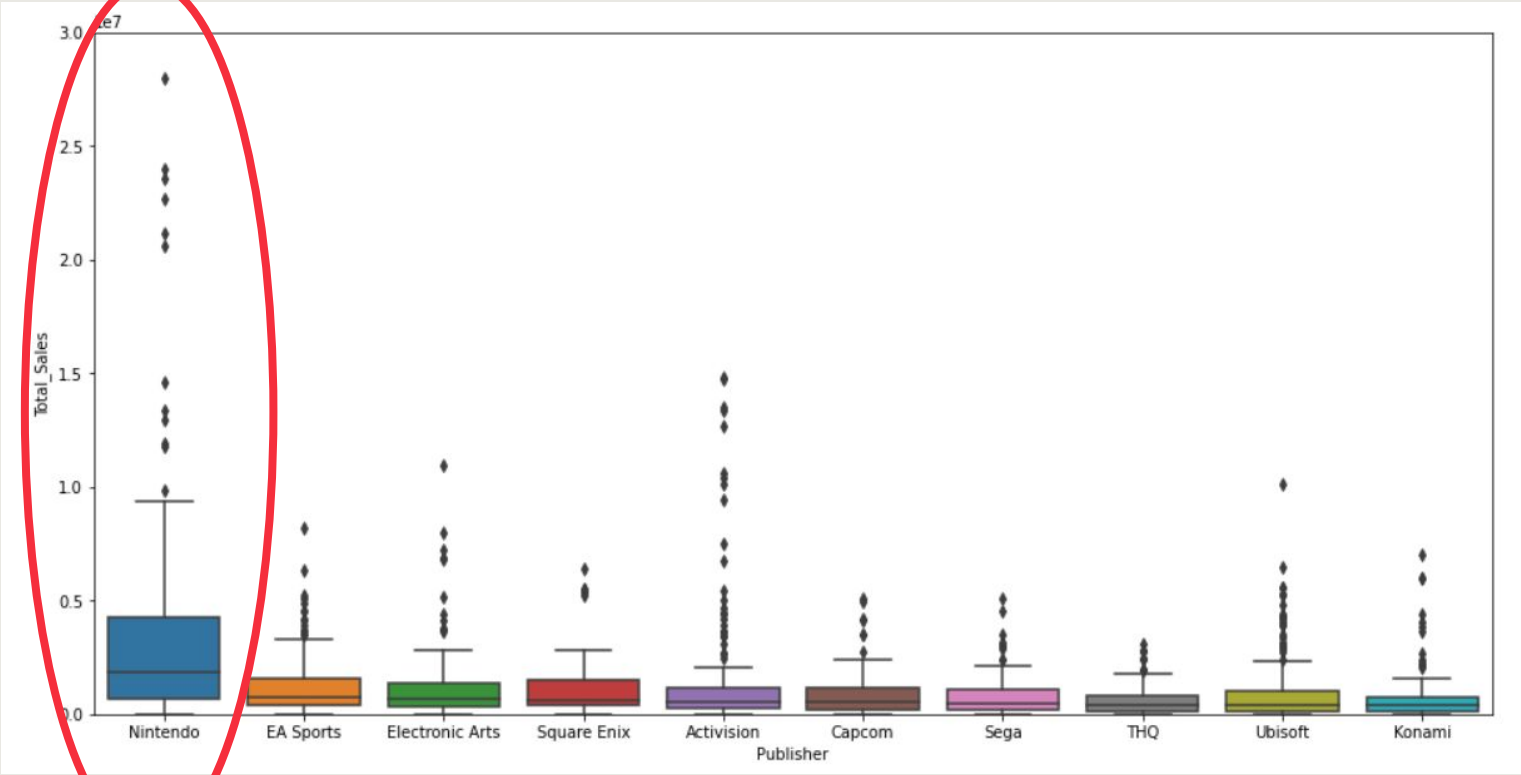
- Games enjoy higher sales during **summer** and **Christmas** seasons

» EDA: Release Year «

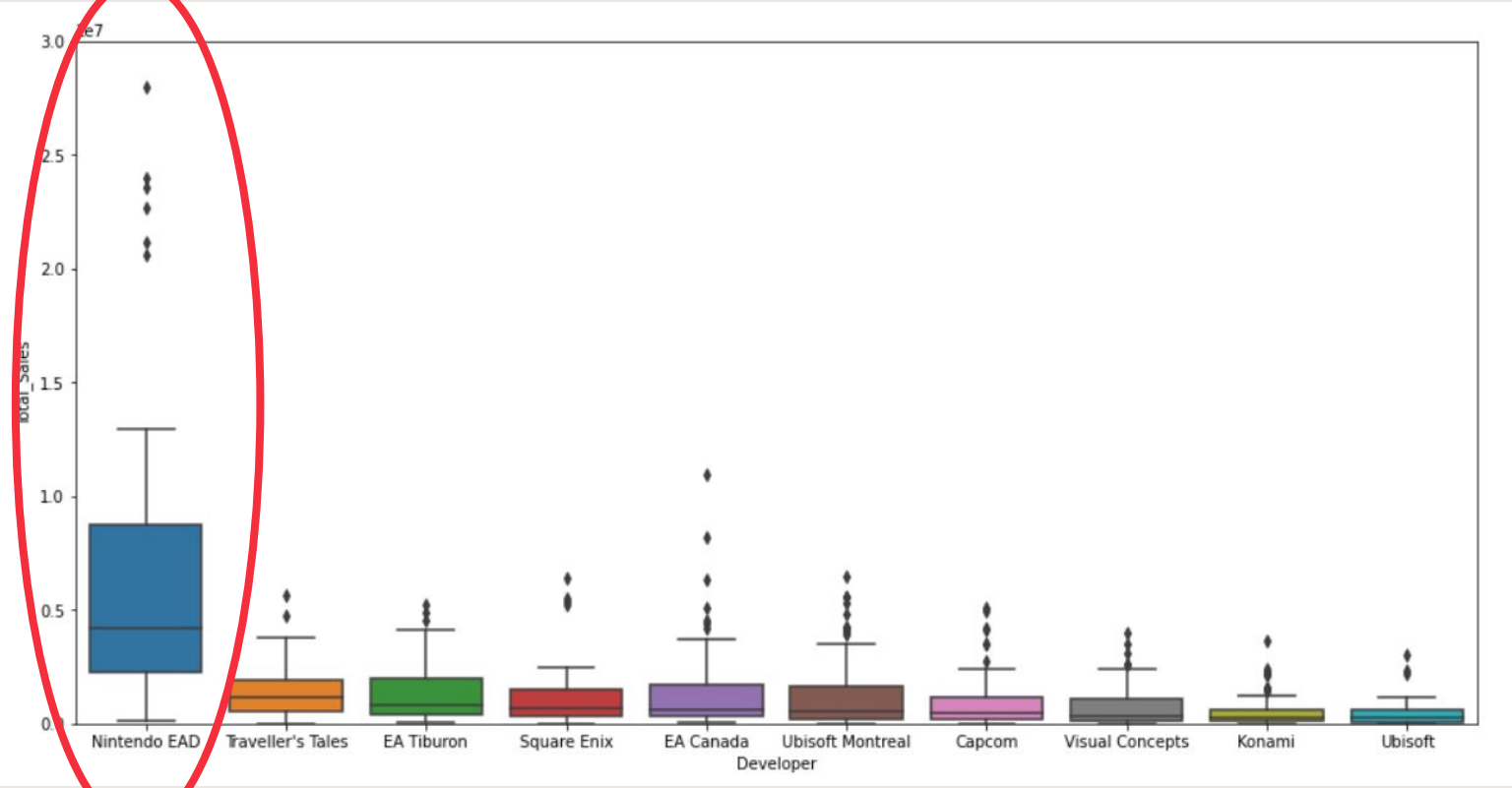


- Dip in sales around 2006-2011 region, due to **2008 financial crisis**
- Economic situation is an important factor

» EDA: Publishers «



» EDA: Developers «



» EDA: Pub/Dev «

- Nintendo has more experience in the industry & produced many ‘AAA’ titles
- Game creators can consider Nintendo as a safe and reliable developer/publisher to work with

» Model «

Regression

» Model Preparation «

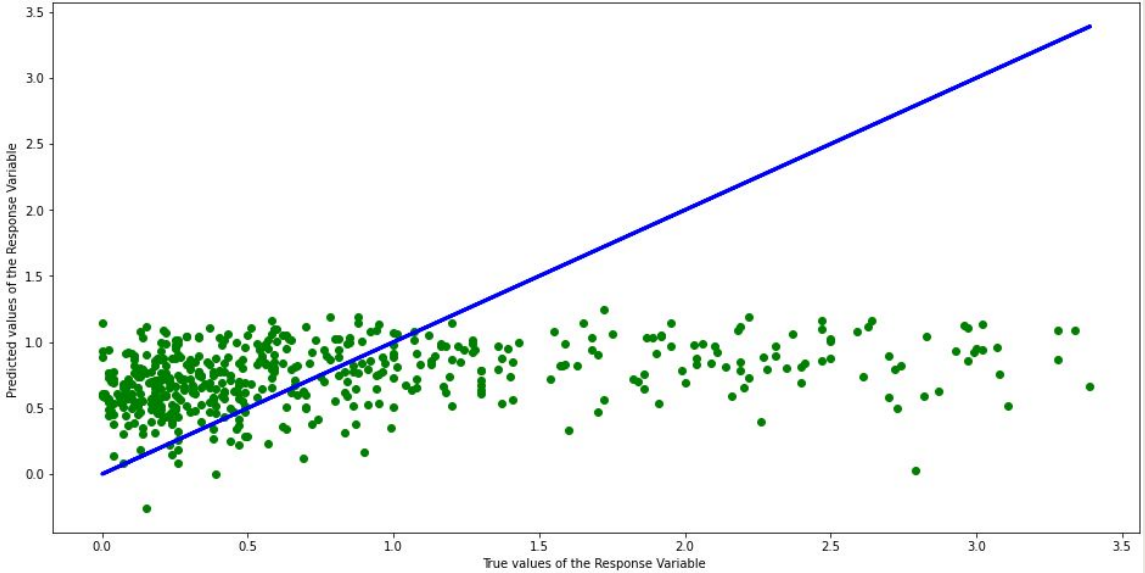
Predictors Used:

Genre, Platform, Publisher, Developer, Score, Release_Month

Response Variable:

Total_Sales - in millions

» Simple Linear Regression «

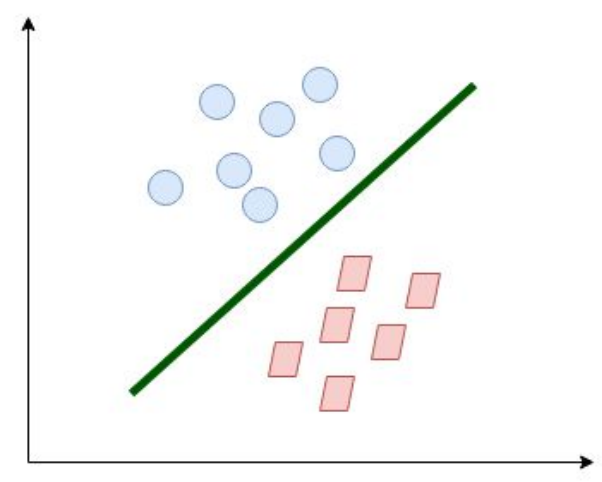


Evaluation Score:

MSE: 0.580947

RMSE: 0.762199

» Support Vector Regression «



Use of margin lines to decide on **decision boundaries** separating plotted points.

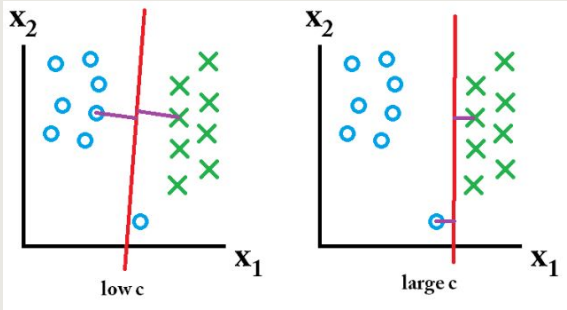
Evaluation Score:

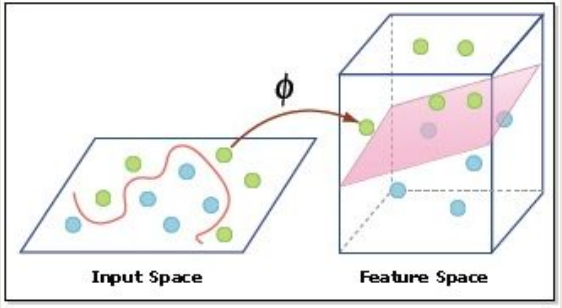
MSE: 0.659933

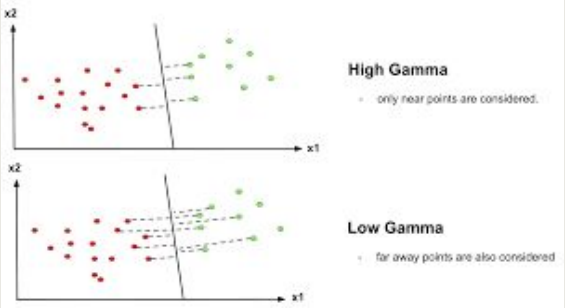
RMSE: 0.812362

» Model Tuning «

GridSearchCV







» Model Tuning « GridSearchCV

```
from sklearn.model_selection import GridSearchCV

hyper_grid = {'C': [0.1,1,10,100,1000], 'gamma':[1,0,1,0,0.1,0,0.01,0,0.001], 'kernel':['rbf']}

regressor = SVR(kernel='rbf')
regressor.fit(X_train, y_train)
grid = GridSearchCV(estimator = regressor,
                    param_grid = hyper_grid,
                    refit = True,
                    verbose = 3)
grid.fit(X_train, y_train)
```

Evaluation Score:

MSE: 0.512079

RMSE: 0.715597

» Insights using Model «

» Problem Statement «

- Predict sales price using game attributes
- Create a model to help new or aspiring game creators

Genre

Platform

Publisher

Developer

Score

Release-Month

» Combination « Generator

```
def generator(Genre = None, Platform = None, Publisher = None, Developer = None)
```

The idea is to create a black box function that allows users to input whichever genre, platform, etc. they have in mind. (Not necessary to specify all)

The generator will then use the model to predict which combination of predictors with the user specified ones fixed.

generator(Genre = ‘Action’, Platform = ‘X360’)

```
Your best combination is :-
Genre:  Action
Platform:  X360
Publisher:  THQ
Developer:  Krome Studios
Release Month:  October
Estimated Total Sales:  2.0  million copies!
```

Input

Output

» Conclusion «

Data-Driven Insights

- Games enjoy high sale volume during **summer** and **Christmas season**
- Games suffer low sale volume during **recession**
- Nintendo is a reliable publisher and developer for potential game developers

» Conclusion «
Blackbox Model

- Users fill in at least 1 attribute of their game
- Generator decide the best option for the rest of the attributes
- Useful for users who already **have a general idea** of their game, looking to **get a more complete image and guidance**

Future Improvements

- Scrape **Summary, Reviews** for game and perform **Natural Language Processing**
- Use other gaming websites to fill in blanks in the dataset

» THANK YOU «