Ukrainian Catholic University

Bachelor Thesis

Optimization of the product concept for the video game market by applying econometric modelling

Author:
Marko Zahartovskyi

Supervisor:
Dr. Kenny Ching

A thesis submitted in fulfillment of the requirements
for the degree of Bachelor of Science

in the

Department of Computer Sciences
Faculty of Applied Sciences

Lviv 2022

Declaration of Authorship

I, Marko ZAHARTOVSKYI, declare that this thesis titled, "Optimization of the product concept for the video game market by applying econometric modelling" and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed:

Date:

"Nobody in this industry knows what they're doing, we just have a gut assumption."

Cliff Bleszinski

UKRAINIAN CATHOLIC UNIVERSITY

Faculty of Applied Sciences

Bachelor of Science

Optimization of the product concept for the video game market by applying econometric modelling

by Marko ZAHARTOVSKYI

Abstract

The video game industry is a mix of technologies and art. This is a rapidly evolving market which captures the look of many dreamers. They are not marketing specialists or experts in statistics. They have only their vision to guide them. Because of this, it is tough to find the right product concept for this market and entering this industry brings many risks. This thesis is created to help independent game developers find an alternative to their "gut assumption" in a data-driven approach using econometric modelling tools.

Acknowledgements

I want to express my gratitude to my supervisor Dr.Kenny Ching who helped me to complete this thesis. Also I'm infinitely thankful to my parent, UCU, my classmates and friends, our APPS faculty and to everyone who supported me through last four years and continues to do this . . .

Contents

List of Figures

List of Abbreviations

OLS	**O**rdinary **L**east **S**quares
WLS	**W**eighted **L**east **S**quares
SSR	**S**um (of) **S**quares **R**egression
SST	**S**um (of) **S**quares **T**otal

List of Symbols

r^2	R-squared
e	margin of error(error in the model)
N	size of population
p	varience
Z	number from z-table according to precision level

For my mother and others who inspire me to live…

Chapter 1

Introduction

1.1 Motivation

The video game industry is a fast-growing and rapidly changing industry. It grew by 7.6% year on year in 2021, and analytics predict it to reach more than $200 billion via consumer spending in 2022, according to reports by NewZoo.(Wijman, 2021; Wijman, 2022)

One of the reasons for this growth is the growing number of indie developers. The accessibility of tools and reduced minimum level of skills needed to develop a game have led to the state of the market, which is commonly named "Indiepocalypse".(Ibrisagic et al., 2016; Wright, 2018) This term describes an overcrowded market due to the uncontrollable release of new products on the market with low entry requirements. As a result, any new product has a slight chance of being discovered by players among other similar products on the market despite the quality or innovation of this product. Around 37% of video games released on the Steam, an online store with a market share of around 75% of all game sales on PC, never was downloaded.(Orland, 2014)

This market state causes a high value of a good marketing campaign for the successful release of the product and to understand which product would be interesting for the audience. All of this depends on data and its analysis. International companies with multibillion budgets can allow themselves to fund a separate team of professionals to perform research and gather data or to hire independent marketing companies. At the same time, small independent developers, who sell their games on the same market, are on a tight budget and limited to skills inside their team. So, these small development teams depend on open data and research methods that are cheap and easy to use.

1.2 Problem Statement

This thesis concentrates on the early stage of the game development process - concept development. At this stage, developers can determine the price, release date and draft defining features of a future product like genre, gameplay features, art style, available languages in the game and type of controller used in game. Most available research papers try to predict sales based on data available only after the product already has been released, like critics' reviews or its place in different ratings.(Aziz et al., 2018; Yufa et al., 2019)

Also, unlike other works, in this thesis, we use econometrics for analysis because of its advantage in part of result interpretation. In other research papers, researchers use machine learning but, despite better precision of predictions, this approach is not practical for causal analysis. This advantage of the econometrical approach is vital in

game concept optimization. Because predicting sales can give only binary answers, is this number of copies sold enough for developers or not. However, using the econometrical model, developers should understand the influence of different features on their sales to more effectively fit product features for market needs.(Trněný, 2017)
So, in this thesis, we want to verify three hypotheses:

1. Is it possible to predict game sales?
2. Is it possible to analyse the effect of different game features on sales?
3. Does this approach satisfy the needs of independent developers?

All of this with econometric modelling and only data available in the concept development stage.

1.3 Structure

In next chapter we will go through all steps of the research from data gathering to results analysis. All steps will be discussed with the idea that any game developer should understand if not the logic than at least an implementation of methods, because developers should be able to replicate this research and make it useful for themselves.

Chapter 2

Main Part

2.1 Background

2.1.1 What is a game concept and its development process?

Creating the game concept is one of the earliest stages in the development process. The concept document, which is a result of this stage, should describe the essential features representing a future product. So, its development is a stage of deep market research and brainstorming about what is interesting, what sells good and what developers can add to make their product unique.(Schell, 2008; Dille, 2007)

Usually, independent developers rely on their game design understanding rather than a data-driven approach. As a result, they can miss critical interconnections between different features or some unobvious trends.(Wagner, 2021

2.1.2 Lack of data

Furthermore, even developers who want to adopt a data-driven approach can suffer from the lack of accessible data. Usually, big game development companies and online markets like Steam protect their sales-related data. So, the game developers community develops methods to calculate approximate game sales using different proxy data like the number of players' reviews on Steam.(Kontus, 2021

2.1.3 Game features

Game features in the context of this thesis include specific characteristics of the game, like some game mechanics or languages present in the game, and general descriptive terms, like the genre or mood of the game. Although, these features can describe only characteristics existing in other products released before. Any innovations or unique features which we cannot interpret in commonly used terms we cannot include in the model. So we should understand the limitations of the data-driven analysis in part of features we can describe in our research.

2.2 Dataset

2.2.1 Description

In this thesis, we use data gathered from the SteamSpy - web service collecting all available data about products on Steam. This service was developed by Sergey Galyonkin, director of publishing strategy at Epic Games. This service provides product pages with information accessible through Steam's Web API and with approximated number of owners calculated according to the method described by Kyle Orland from Ars Technica.

As a proxy of different game features, we use Steam tags. Steam tags are a user-defined set of tags defined by the game developers and players, which should describe the game for ease of discoverability on Steam. The complete Steam tags that we use consist of 340 game tags, 11 genres, 29 in-game localization languages, and 120 categories. So, in our final dataset, tags will be represented by 500 binary variables.

Other game-defining features that are not tagged are release date and price. Price is float number and equals the product's price on Steam in US dollars. For this thesis, we use twelve binary variables to define the game's release month. However, any other approach which describes release date as a categorical variable can be used.

As dependent variables, we use the number of copies owned by players on Steam at the time one month and three months from the start of sales. These variables are non-negative integers rounded to thousands.

2.2.2 Gathering

SteamSpy lacks the built-in functionality to download needed game information, so we must parse every game page separately. To do this, we use Python and Selenium library. However, this website has protection from parsing, so we must parse not the actual pages but their source code representation because if we go directly to this source view website cannot fixate that we made any calls to the server.

We can limit the games we analyze for this research by taking only games released in 2020-2021. We do this because the video game industry is rapidly changing. The data from the last two years is the minimal time range we can take with an assumption that its trends are homogenous and that size of the population is enough for analysis. With these limits size of the population equals 20153. Due to the bandwidth of the SteamSpy, downloading all of the data takes a few days, so we will take a sample.

To calculate the minimum size of the random sample, we use Cochran and Yamane formulas:

$$n = \frac{Z^2 p(1-p)}{e^2}$$

$$n = \frac{N}{1 - Ne^2}$$

With a margin of error, variance and level of precision equal to 0.01, 0.5 and 0.95 accordingly, we get the following calculations:

$$n = \frac{1.65^2 * 0.5(1-0.5)}{0.01^2} = 6806.25$$

$$n = \frac{20153}{1 - 20153 * 0.01^2} = 6683.58$$

So, the sample size of 6807 or more should be enough to get the correct understanding of the population with approximately 94% precision.

We will use data about 7414 random games released in 2020 and 2021 for this research.

2.2.3 Cleaning

Users and developers create tags on Steam to better describe games. As a result, some tags are so specific that they are almost unique and can be found only describing a few games on the market. So, to avoid mistakes in analysis connected to the low representation of these rare variables, we must clean data. We start with removing all of the variables encountered less than one hundred times in our data set. If we look at the list of deleted variables, we can see that they describe specific or abstract features of the games. Final quantity of independent variables in cleared dataset equals 206.

Finally, to check data on the bias, we compear the distribution of sales and genres between the population and our sample. As the figures 2.1, 2.2, 2.3 show, the distributions in the population and the sample of dependent and independent variables are nearly identical, so we can state that our sample is unbiased.

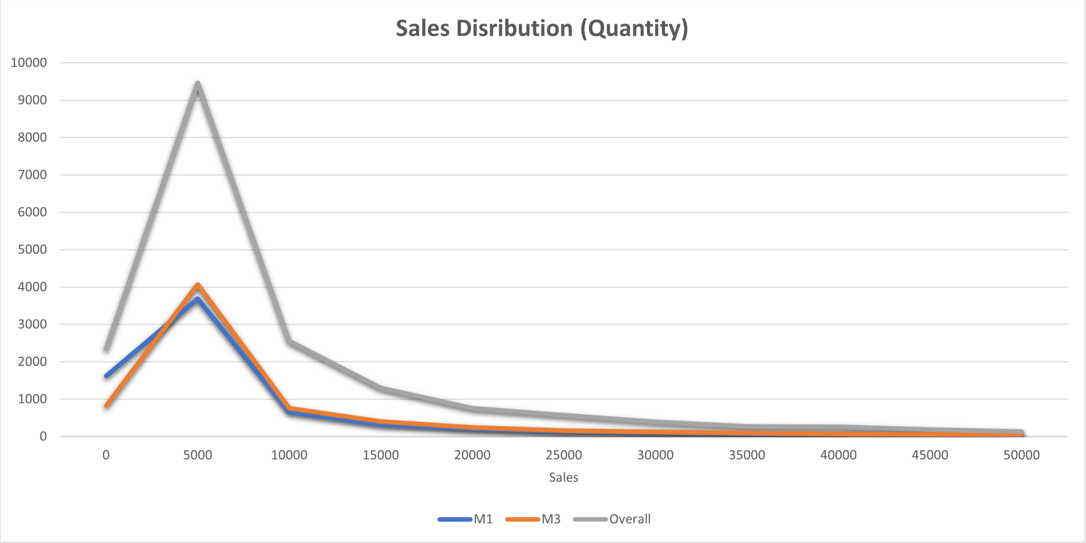


FIGURE 2.1: Distribution of sales in quantity of games in the sample. M1 - one month after release, M3- three months after release, Overall - from release till the day data was downloaded.

2.3 Methodology

We use Python with statsmodels and pandas libraries to conduct this study.

2.3.1 Data split

In the first stage, we must randomly split the dataset into training and testing groups. We use an 80/20 size ratio for these two sets of data. The training set is used to build an econometric model. The testing set is used to optimise the model and evaluate the results.

However, we have to eliminate outliers due to the significant gap between sales of the most successful game and the median. So, we sort data based on dependent variables and cut the percentage of top entries before we split the dataset.

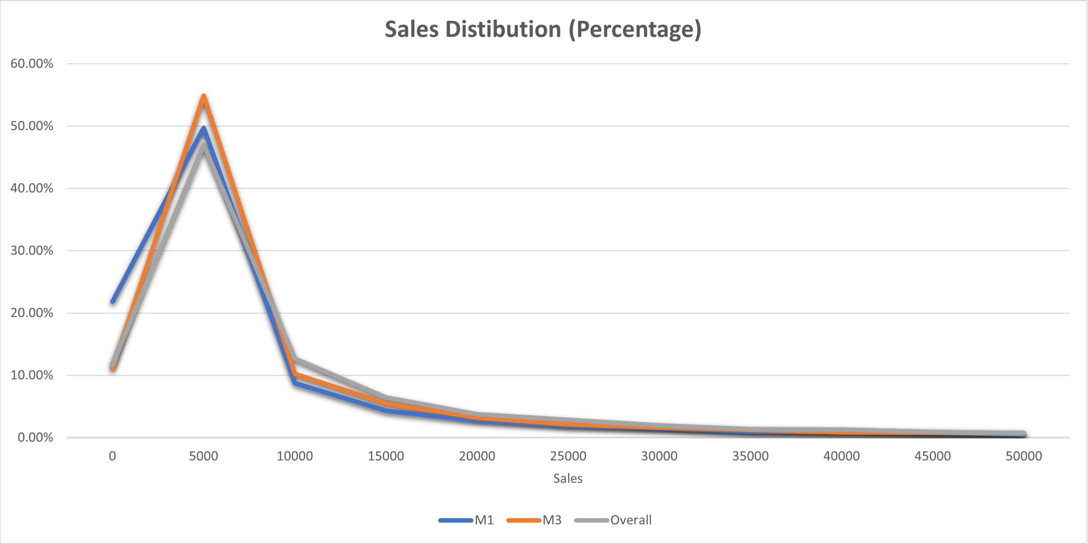


FIGURE 2.2: Distribution of sales in the proportion of the sample. M1 - one month after release, M3- three months after release, Overall - from release till the day data was downloaded.

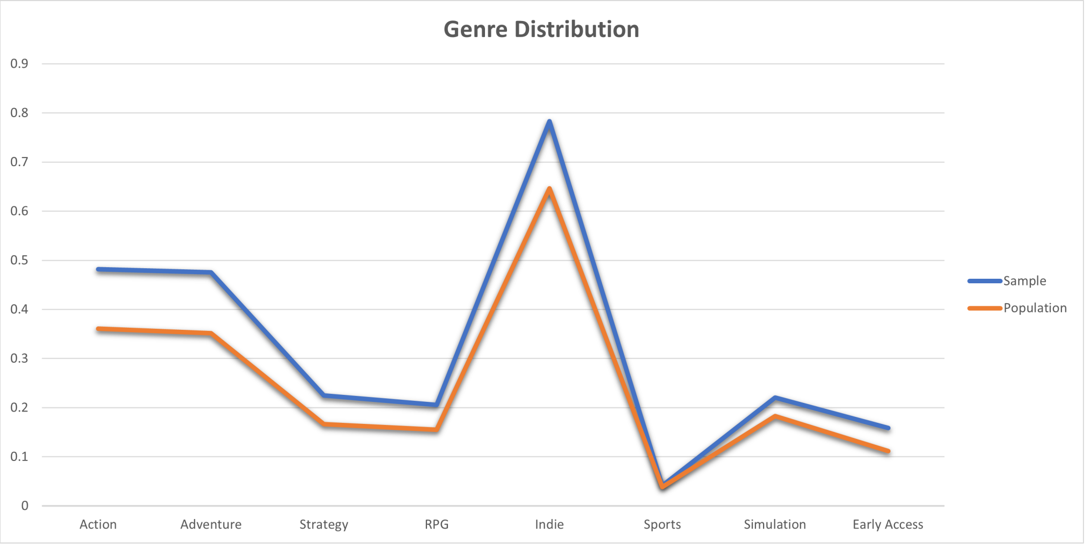


FIGURE 2.3: Distribution of game genres in the sample and population.

2.3.2 Econometric models

In the second stage, we build an econometric model. We use two different model variations: OLS and WLS.

OLS is the more accessible and the most popular model in econometrics. The goal of the OLS is to minimise SSR. The quality of the model can be evaluated using r^2. The higher the r^2 - the better model can explain the dependent variable.(Wooldridge, 2020)

$$Y_i = \beta_0 + \beta_1 X_i + e_i$$

$$SST = \sum_{i=1}^{n} (y_i - \bar{y})$$

$$SSR = \sum_{i=1}^{n} e_i^2$$

$$r^2 = 1 - \frac{SSR}{SST}$$

The limitation of the OLS is that it loses its effectiveness the less linear the correlation of the dependent variable is. To address this, we use the WLS model too. WLS is a generalisation of OLS. OLS works with the assumption of homoscedasticity, so when this assumption is not valid, OLS cannot provide a quality model. WLS addresses this problem by adding additional weight coefficients.

We can build both of these econometric models using statsmodels library.

2.3.3 Multicollinearity problem

Multicollinearity occurs when a few independent variables in a model are correlated. It can lead to losing the model's precision and wrong conclusions. So the first step of model optimisation is to check for multicollinearity and drop redundant variables.

To do this, we use VIF. It shows the presence of collinearity in a set of regression variables. The higher factor indicates higher collinearity of the independent variable with other variables in the model. There is not any established point when we can say that this variable has too high collinearity, and we should drop it. Usually, a threshold for VIF is set equal to 5 or 10 is used. For this thesis, we place a threshold equal to 5. When the VIF of an independent variable is higher than five, we consider that this variable is collinear with other variables and causes a multicollinearity problem in the model.

$$VIF_i = \frac{1}{1 - R^2}$$

2.3.4 Predictive model

After creating an econometric model, we can use it to calculate the dependent variable. So, we can use the model to predict game sales based on our input features.

As was stated in 2.3.2, we use r^2 to verify the precision of our predictive model. The model is considered precise if r^2 is 0.9 or higher and the minimum threshold to use the model for prediction purposes is when r^2 equals 0.8.

2.3.5 Model optimisation

At this stage, there are two possible directions for optimisation. The method how we optimise is standard for both approaches. We look at the p-value of the variables and drop variables with high p-values. P-value is a statistic which shows if the variable should be discarded. A higher p-value means a higher probability that a variable is insignificant.

The first approach starts with a basic model with all of the variables in, and then we drop variables with high p-values one by one rebuilding the model at each step of the loop. The second approach starts with a minimal number of variables and then adds variables while dropping variables with a high p-value.

However, it is crucial to prioritise logic and understanding of variables because otherwise, we can drop essential variables that we know are important because of the numbers. For example, all release dates are mutually unique variables, so their VIF equals infinite, but we understand what causes this, so we should not drop them all blindly. The same works for optimisation based on the p-value. When we see that genre has a low p-value, we should consider the cause and whether this variable may be insignificant.

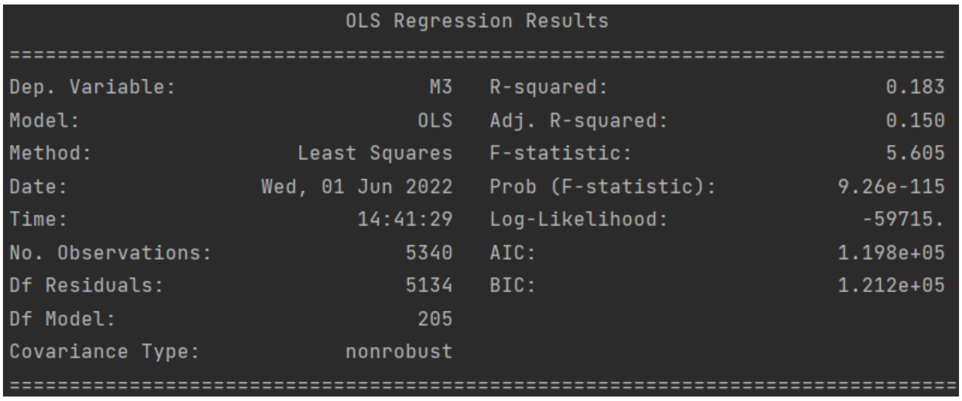
2.3.6 Feature analysis

The most crucial advantage of econometrics is the capability to perform causal analysis and understand how each variable affects the dependent variable. To do this, we look into coefficients accordingly to each variable. If a variable is statistically significant, we can interpret the coefficient as an impact this variable has on the dependent variable. As our variables(excluding price) are binary, the interpretation of coefficients is straightforward. The coefficient shows how much sales would this feature generate if the coefficient is positive and how much sales would drop if the coefficient is negative.

2.4 Results

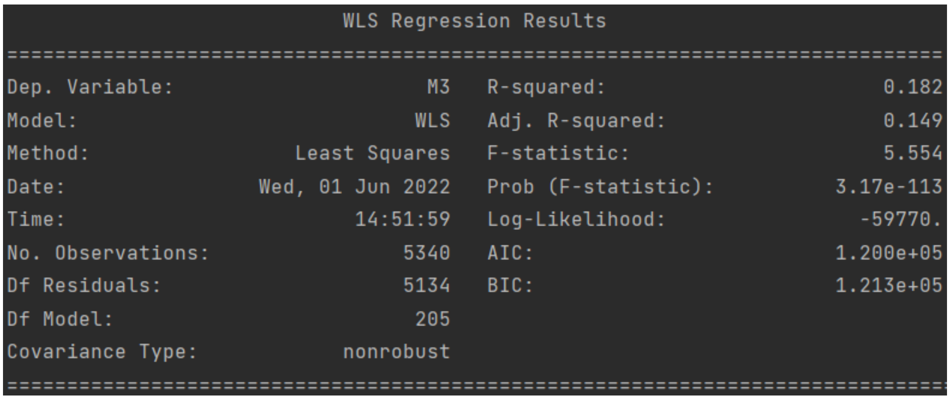
2.4.1 Basic model

Basic models include all variables from the cleaned dataset. Results of these models with full set of variables from cleaned dataset are shown on 2.4, 2.5. We see that r^2 for both methods are approximately equal and less than 0.2. So, we can confidently say that basic models are not precisive. We see that r^2 for both methods are approx-



```
                            OLS Regression Results
==============================================================================
Dep. Variable:                     M3   R-squared:                       0.183
Model:                            OLS   Adj. R-squared:                  0.150
Method:                 Least Squares   F-statistic:                     5.605
Date:                Wed, 01 Jun 2022   Prob (F-statistic):          9.26e-115
Time:                        14:41:29   Log-Likelihood:                -59715.
No. Observations:                5340   AIC:                         1.198e+05
Df Residuals:                    5134   BIC:                         1.212e+05
Df Model:                         205
Covariance Type:            nonrobust
==============================================================================
```

FIGURE 2.4: Results of basic OLS model.



```
                            WLS Regression Results
==============================================================================
Dep. Variable:                     M3   R-squared:                       0.182
Model:                            WLS   Adj. R-squared:                  0.149
Method:                 Least Squares   F-statistic:                     5.554
Date:                Wed, 01 Jun 2022   Prob (F-statistic):          3.17e-113
Time:                        14:51:59   Log-Likelihood:                -59770.
No. Observations:                5340   AIC:                         1.200e+05
Df Residuals:                    5134   BIC:                         1.213e+05
Df Model:                         205
Covariance Type:            nonrobust
==============================================================================
```

FIGURE 2.5: Results of basic WLS model.

imately equal and less than 0.2. So, we can confidently say that basic models are not precisive.

2.4.2 After multicollinearity fix

After calculating VIF, we see that variables in the figure have VIF higher than 0.5. We should drop these variables based on the numbers from the table, but instead, we should think about causality, as was mentioned in 2.3.5.

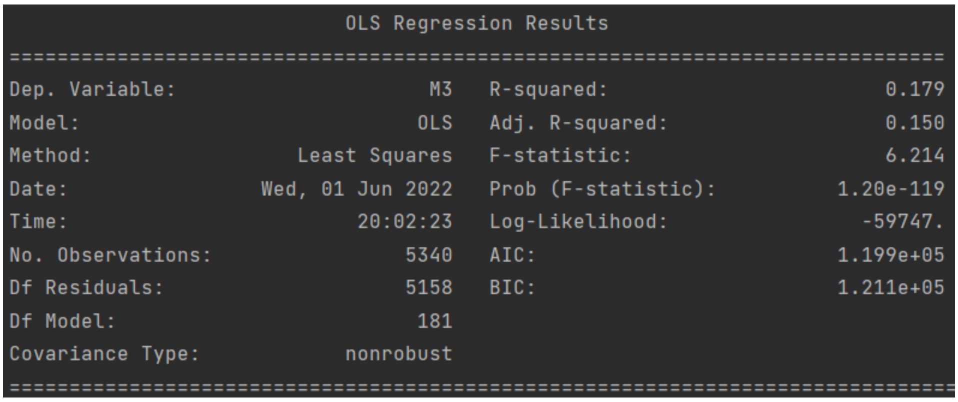
Localisation for the rear languages is uncommon and can be made only by big projects, and in such cases, developers can afford to add a lot of different languages. So, this is the cause why we see the correlation between languages. We should drop most of the languages except a few rare.

As already was mentioned, release dates are mutually exclusive. So, this causes infinite correlation. To eliminate this problem, we can leave only half of the months with the lowest p-values.

Var	VIF
January	inf
February	inf
March	inf
April	inf
May	inf
Jube	inf
July	inf
August	inf
September	inf
October	inf
November	inf
December	inf
Swedish	5.8
Greek	10.9
Romanian	9.8
Vietnamese	5.3
Norwegian	9.2
Hungarian	5.2
Finnish	9.7
Danish	9.7

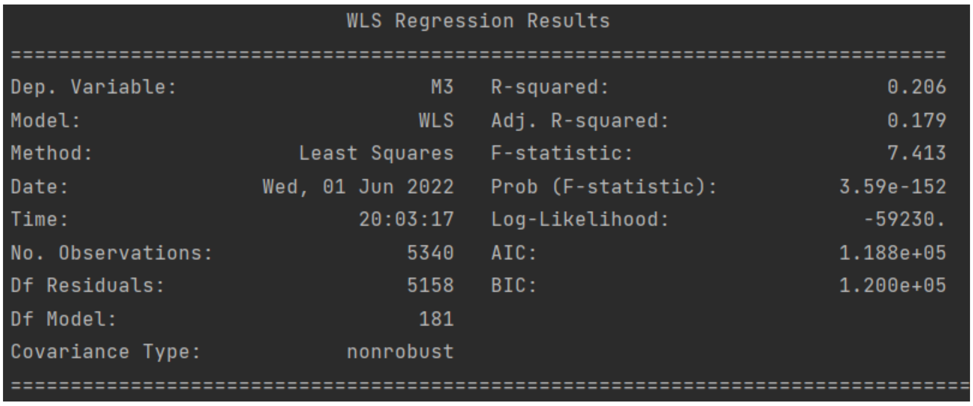
2.4.3 Predictions

After excluding variables causing multicollinearity issues, we can build new predictive models. The results of these models can be seen in the figures 2.6 and 2.7. r^2 is approximately the same as in the results of the basic models. Then we optimise with the first method described in the section 2.3.5. Results are shown in the figures 2.8 and 2.9—still the approximately same results as in previous models. After using the second method of optimisation from the section 2.3.5, we see results in the figures 2.10 and 2.11. r^2 is the lowest among models. These are the worst models that we tested. But they have an advantage in feature analysis due to the model's smaller number of independent variables.



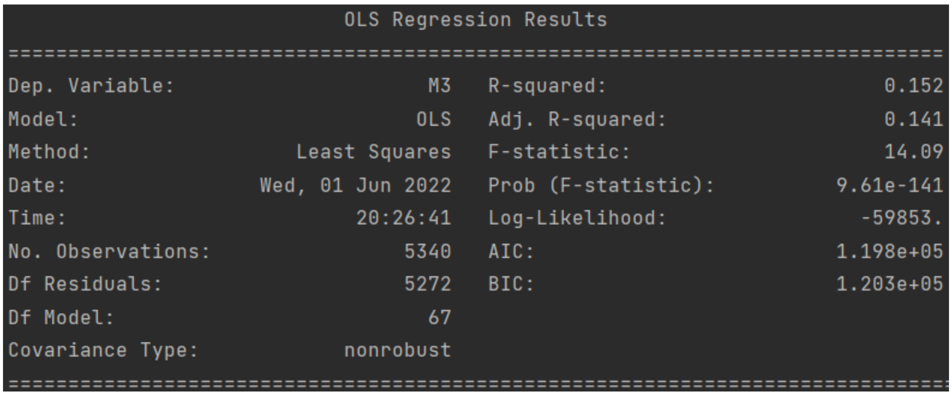
```
                            OLS Regression Results
==============================================================================
Dep. Variable:                     M3   R-squared:                       0.179
Model:                            OLS   Adj. R-squared:                  0.150
Method:                 Least Squares   F-statistic:                     6.214
Date:                Wed, 01 Jun 2022   Prob (F-statistic):         1.20e-119
Time:                        20:02:23   Log-Likelihood:                -59747.
No. Observations:                5340   AIC:                         1.199e+05
Df Residuals:                    5158   BIC:                         1.211e+05
Df Model:                         181
Covariance Type:            nonrobust
==============================================================================
```

FIGURE 2.6: Results of basic OLS model after multicollinearity fix.



```
                            WLS Regression Results
==============================================================================
Dep. Variable:                     M3   R-squared:                       0.206
Model:                            WLS   Adj. R-squared:                  0.179
Method:                 Least Squares   F-statistic:                     7.413
Date:                Wed, 01 Jun 2022   Prob (F-statistic):         3.59e-152
Time:                        20:03:17   Log-Likelihood:                -59230.
No. Observations:                5340   AIC:                         1.188e+05
Df Residuals:                    5158   BIC:                         1.200e+05
Df Model:                         181
Covariance Type:            nonrobust
==============================================================================
```

FIGURE 2.7: Results of basic WLS model after multicollinearity fix.

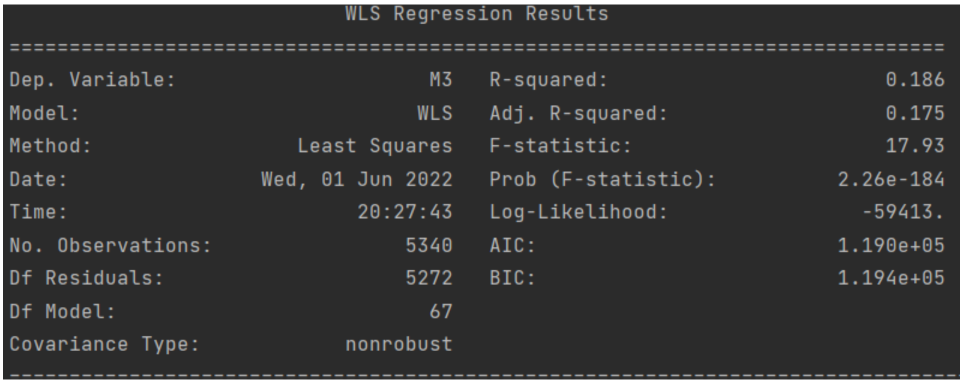


```
                            OLS Regression Results
==============================================================================
Dep. Variable:                     M3   R-squared:                       0.152
Model:                            OLS   Adj. R-squared:                  0.141
Method:                 Least Squares   F-statistic:                     14.09
Date:                Wed, 01 Jun 2022   Prob (F-statistic):         9.61e-141
Time:                        20:26:41   Log-Likelihood:                -59853.
No. Observations:                5340   AIC:                         1.198e+05
Df Residuals:                    5272   BIC:                         1.203e+05
Df Model:                          67
Covariance Type:            nonrobust
==============================================================================
```

FIGURE 2.8: Results of OLS model after optimisation with first method.

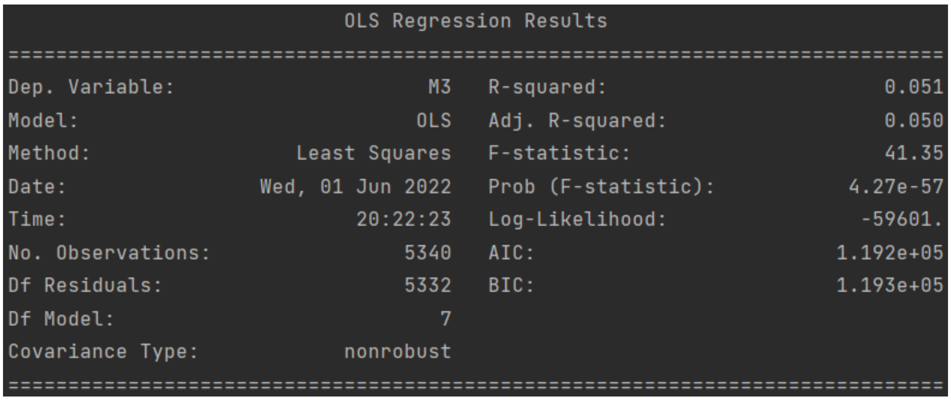
2.5 Discussion

We have seen that all predictive models were not precise according to low r^2, so econometric modelling should not be used for sales prediction for the video game market. However, we still can analyse distinct features, their combinations and their impact on sales. The possibility to predict the effect of different game elements on sales is more valuable than a blind prediction of these sales because it allows developers to make reasoned decisions when they understand what is essential in their game and how to distribute their limited resources.



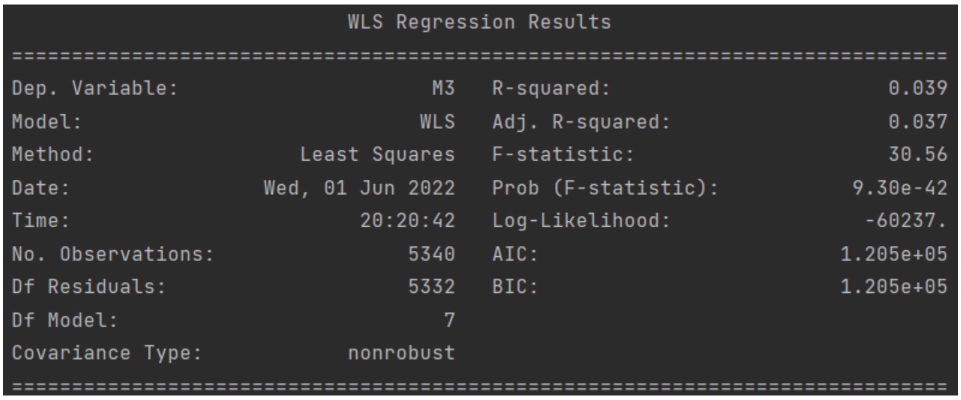
```
                            WLS Regression Results
==============================================================================
Dep. Variable:                   M3   R-squared:                       0.186
Model:                          WLS   Adj. R-squared:                  0.175
Method:               Least Squares   F-statistic:                     17.93
Date:              Wed, 01 Jun 2022   Prob (F-statistic):          2.26e-184
Time:                      20:27:43   Log-Likelihood:                -59413.
No. Observations:              5340   AIC:                         1.190e+05
Df Residuals:                  5272   BIC:                         1.194e+05
Df Model:                        67
Covariance Type:          nonrobust
------------------------------------------------------------------------------
```

FIGURE 2.9: Results of basic WLS model after optimisation with first method.



```
                            OLS Regression Results
==============================================================================
Dep. Variable:                   M3   R-squared:                       0.051
Model:                          OLS   Adj. R-squared:                  0.050
Method:               Least Squares   F-statistic:                     41.35
Date:              Wed, 01 Jun 2022   Prob (F-statistic):           4.27e-57
Time:                      20:22:23   Log-Likelihood:                -59601.
No. Observations:              5340   AIC:                         1.192e+05
Df Residuals:                  5332   BIC:                         1.193e+05
Df Model:                         7
Covariance Type:          nonrobust
==============================================================================
```

FIGURE 2.10: Results of basic OLS model after optimisation with second method.



```
                            WLS Regression Results
==============================================================================
Dep. Variable:                   M3   R-squared:                       0.039
Model:                          WLS   Adj. R-squared:                  0.037
Method:               Least Squares   F-statistic:                     30.56
Date:              Wed, 01 Jun 2022   Prob (F-statistic):           9.30e-42
Time:                      20:20:42   Log-Likelihood:                -60237.
No. Observations:              5340   AIC:                         1.205e+05
Df Residuals:                  5332   BIC:                         1.205e+05
Df Model:                         7
Covariance Type:          nonrobust
==============================================================================
```

FIGURE 2.11: Results of basic WLS model after optimisation with second method.

Furthermore, the methods described in this chapter are easy to replicate using Python or R. All needed data has open access or can be easily gathered by developers by creating their own set of tags. Python code is shown in Appendix B.

Chapter 3

Conclusion

3.1 Results

We have proposed three hypotheses that had to be verified. And as the result of the research, we can state a conclusion:

- It is impossible to predict sales of video games using econometric modelling and data available at the stage of concept development;
- It is impossible to predict sales of video games using econometric modelling and data available at the stage of concept development;
- With minimal programming skills and a basic understanding of mathematics, it is possible to use mentioned methods. As game development demands programming and often mathematical skill at a confident level, we can be sure that these methods are easy enough for game developers to use.

3.2 Recommendations

This thesis looked through this domain very superficially. In the case of actual implementation, the own set of game features should be developed. Also, it would be helpful to look deeper into each category of variables separately. For example, we can try to use not months but for different events like the Christmas holidays or summer vacation season as periods that can impact sales. Overall, this research was successful. It is not innovative in part of econometrics or data analysis, but it is vital for this industry and may change the game developers' approach. Unlike other research that has looked into machine learning techniques, the methods described in this research are "user friendly" and easy to replicate, which makes them more practical.

Appendix A

Variables

#	Title	Quantity	Average	Median
1	Price	7414	8.75	4.99
2	January	450	-	-
3	February	576	-	-
4	March	564	-	-
5	April	587	-	-
6	May	557	-	-
7	June	570	-	-
8	July	739	-	-
9	August	604	-	-
10	September	668	-	-
11	October	811	-	-
12	November	629	-	-
13	December	659	-	-
14	Hand-drawn	507	-	-
15	Sports	304	-	-
16	Realistic	594	-	-
17	Comedy	624	-	-
18	Driving	205	-	-
19	Turn-Based Tactics	332	-	-
20	Historical	264	-	-
21	Cute	1334	-	-
22	Swedish	265	-	-
23	Perma Death	230	-	-
24	Minimalist	780	-	-
25	Walking Simulator	280	-	-
26	Platformer	880	-	-
27	Shoot 'Em Up	380	-	-
28	Polish	761	-	-
29	Sandbox	457	-	-
30	Adventure	3525	-	-
31	Greek	209	-	-
32	Sexual Content	383	-	-
33	Japanese	1643	-	-
34	Tactical	370	-	-
35	2D	2806	-	-
36	Spanish - Spain	1583	-	-

37	Isometric	255	-	-
38	Combat	701	-	-
39	Base-Building	273	-	-
40	1980s	265	-	-
41	Controller	490	-	-
42	Single-player	7149	-	-
43	Violent	539	-	-
44	Female Protagonist	591	-	-
45	Fantasy	1104	-	-
46	Horror	822	-	-
47	Simplified Chinese	2194	-	-
48	Thriller	192	-	-
49	Pixel Graphics	1407	-	-
50	Robots	241	-	-
51	Emotional	252	-	-
52	Mature	179	-	-
53	Life Sim	159	-	-
54	Space	501	-	-
55	Nature	251	-	-
56	Action-Adventure	1039	-	-
57	Includes level editor	157	-	-
58	Romanian	218	-	-
59	Cyberpunk	225	-	-
60	Steam Achievements	3635	-	-
61	Beat ′em up	171	-	-
62	Czech	326	-	-
63	Logic	536	-	-
64	Multiple Endings	571	-	-
65	Wargame	154	-	-
66	German	1741	-	-
67	Action Roguelike	227	-	-
68	Ukrainian	282	-	-
69	Medieval	379	-	-
70	Full controller support	1580	-	-
71	Partial Controller Support	920	-	-
72	Linear	705	-	-
73	Investigation	275	-	-
74	Hack and Slash	286	-	-
75	FPS	592	-	-
76	Puzzle-Platformer	464	-	-
77	1990′s	305	-	-
78	Interactive Fiction	329	-	-
79	French	1708	-	-
80	Tactical RPG	146	-	-
81	Choices Matter	562	-	-
82	Anime	850	-	-
83	Steam Leaderboards	526	-	-
84	Difficult	602	-	-
85	Atmospheric	1604	-	-

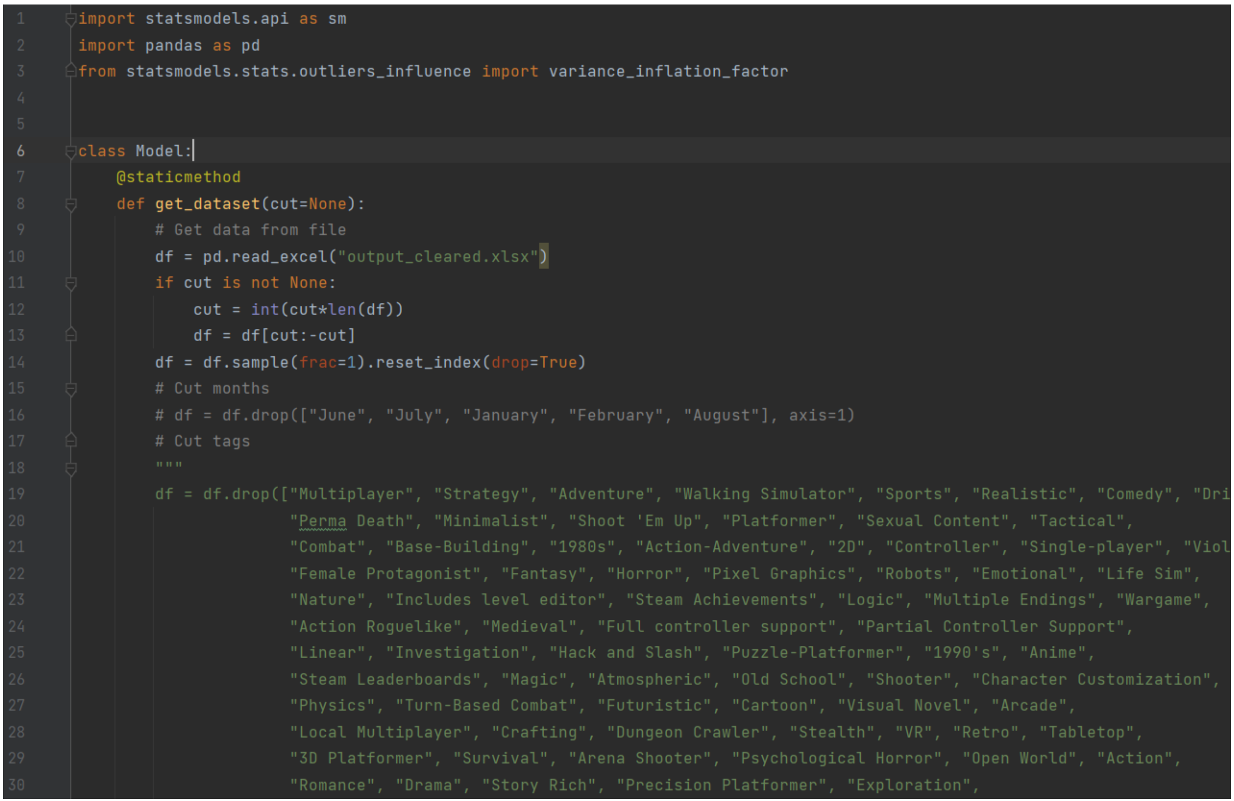
86	Magic	420	-	-
87	Military	231	-	-
88	Tower Defense	152	-	-
89	Old School	462	-	-
90	Match 3	131	-	-
91	Shooter	1032	-	-
92	Strategy	1666	-	-
93	Mystery	594	-	-
94	Choose Your Own Adventure	280	-	-
95	Steam Workshop	183	-	-
96	Multiplayer	836	-	-
97	Thai	279	-	-
98	2.5D	248	-	-
99	Cartoony	662	-	-
100	Side Scroller	522	-	-
101	Score Attack	383	-	-
102	Dating Sim	199	-	-
103	Turkish	617	-	-
104	Resource Management	291	-	-
105	Relaxing	941	-	-
106	Simulation	1634	-	-
107	Abstract	273	-	-
108	Stylized	792	-	-
109	Board Game	209	-	-
110	Top-Down Shooter	271	-	-
111	Cinematic	204	-	-
112	Funny	929	-	-
113	Early Access	1174	-	-
114	Nudity	329	-	-
115	Immersive Sim	249	-	-
116	Colorful	1598	-	-
117	Stats	263	-	-
118	Survival Horror	368	-	-
119	3D	1813	-	-
120	JRPG	322	-	-
121	Puzzle	1672	-	-
122	Dark	592	-	-
123	Runner	231	-	-
124	Italian	1166	-	-
125	2D Platformer	808	-	-
126	Experimental	258	-	-
127	Procedural Generation	416	-	-
128	Family Friendly	923	-	-
129	Shared/Split Screen	539	-	-
130	Top-Down	780	-	-
131	Post-apocalyptic	274	-	-
132	RPGMaker	162	-	-
133	Casual	3502	-	-

134	War	273	-	-
135	English	7287	-	-
136	Korean	1067	-	-
137	Traditional Chinese	1077	-	-
138	Singleplayer	5219	-	-
139	Vietnamese	230	-	-
140	PvE	513	-	-
141	Portuguese - Brazil	1098	-	-
142	Hidden Object	332	-	-
143	Racing	316	-	-
144	Surreal	264	-	-
145	Education	292	-	-
146	Zombies	276	-	-
147	Norwegian	233	-	-
148	Indie	5807	-	-
149	Arabic	296	-	-
150	RPG	1526	-	-
151	Bullet Hell	309	-	-
152	Clicker	231	-	-
153	Turn-Based Strategy	349	-	-
154	Gore	424	-	-
155	Building	454	-	-
156	Rogue-like	363	-	-
157	Narration	238	-	-
158	Steam Cloud	1837	-	-
159	Russian	1889	-	-
160	Point & Click	475	-	-
161	Fighting	161	-	-
162	Sci-fi	858	-	-
163	Hungarian	265	-	-
164	Great Soundtrack	289	-	-
165	Replay Value	218	-	-
166	Exploration	1290	-	-
167	Precision Platformer	270	-	-
168	Third-Person Shooter	252	-	-
169	Story Rich	1280	-	-
170	Drama	316	-	-
171	Rogue-lite	352	-	-
172	Romance	243	-	-
173	Aliens	220	-	-
174	Action	3572	-	-
175	Finnish	233	-	-
176	Open World	631	-	-
177	Psychological Horror	444	-	-
178	Arena Shooter	205	-	-
179	Survival	758	-	-
180	Steam Trading Cards	484	-	-
181	3D Platformer	411	-	-
182	Action RPG	409	-	-

183	Tabletop	198	-	-
184	Retro	863	-	-
185	VR	411	-	-
186	Stealth	239	-	-
187	Dungeon Crawler	315	-	-
188	Dutch	378	-	-
189	Crafting	311	-	-
190	Danish	230	-	-
191	Third Person	840	-	-
192	Local Multiplayer	335	-	-
193	Arcade	1332	-	-
194	Spanish - Latin America	574	-	-
195	Visual Novel	551	-	-
196	Cartoon	436	-	-
197	Futuristic	407	-	-
198	Portuguese	576	-	-
199	Turn-Based Combat	361	-	-
200	Physics	670	-	-
201	Character Customization	433	-	-
202	Management	383	-	-
203	First-Person	1178	-	-
204	Remote Play	688	-	-
205	PvP	1491	-	-
206	Co-Op	839	-	-
207	M1	5795	25266	2000
208	M3	6594	36752	3000

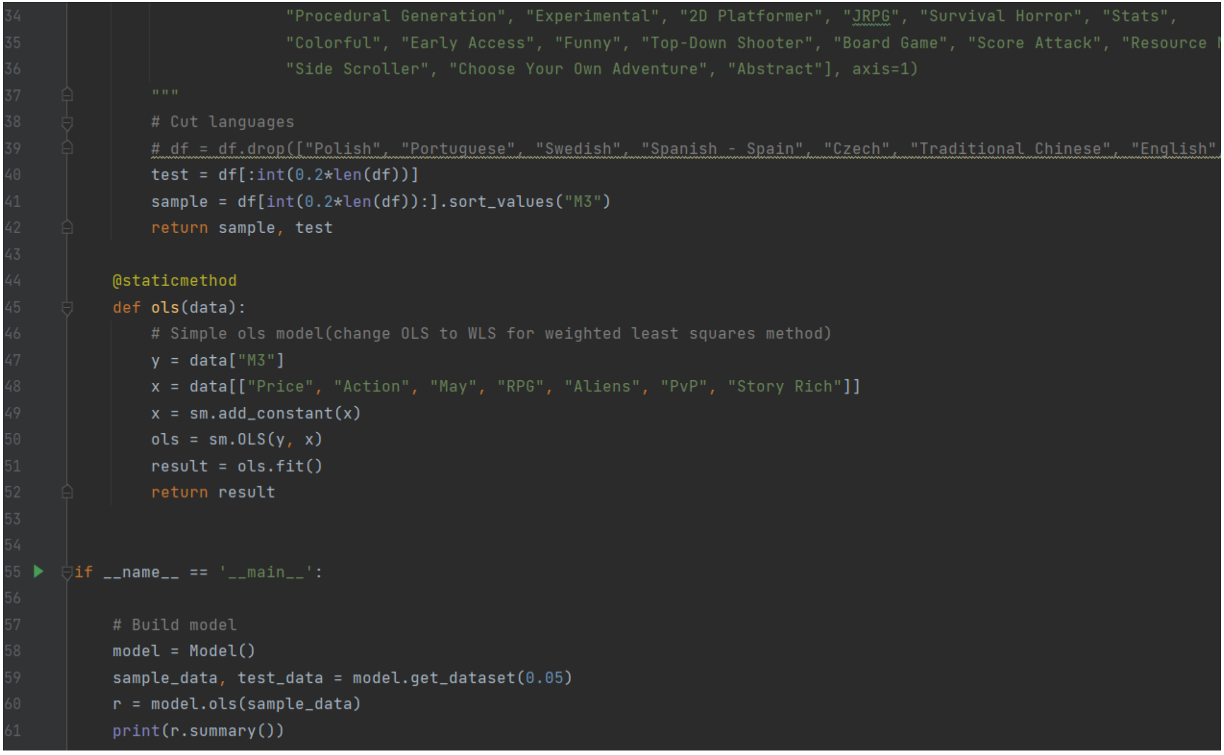
Appendix B

Code



```
import statsmodels.api as sm
import pandas as pd
from statsmodels.stats.outliers_influence import variance_inflation_factor


class Model:
    @staticmethod
    def get_dataset(cut=None):
        # Get data from file
        df = pd.read_excel("output_cleared.xlsx")
        if cut is not None:
            cut = int(cut*len(df))
            df = df[cut:-cut]
        df = df.sample(frac=1).reset_index(drop=True)
        # Cut months
        # df = df.drop(["June", "July", "January", "February", "August"], axis=1)
        # Cut tags
        """
        df = df.drop(["Multiplayer", "Strategy", "Adventure", "Walking Simulator", "Sports", "Realistic", "Comedy", "Dri
                      "Perma Death", "Minimalist", "Shoot 'Em Up", "Platformer", "Sexual Content", "Tactical",
                      "Combat", "Base-Building", "1980s", "Action-Adventure", "2D", "Controller", "Single-player", "Viol
                      "Female Protagonist", "Fantasy", "Horror", "Pixel Graphics", "Robots", "Emotional", "Life Sim",
                      "Nature", "Includes level editor", "Steam Achievements", "Logic", "Multiple Endings", "Wargame",
                      "Action Roguelike", "Medieval", "Full controller support", "Partial Controller Support",
                      "Linear", "Investigation", "Hack and Slash", "Puzzle-Platformer", "1990's", "Anime",
                      "Steam Leaderboards", "Magic", "Atmospheric", "Old School", "Shooter", "Character Customization",
                      "Physics", "Turn-Based Combat", "Futuristic", "Cartoon", "Visual Novel", "Arcade",
                      "Local Multiplayer", "Crafting", "Dungeon Crawler", "Stealth", "VR", "Retro", "Tabletop",
                      "3D Platformer", "Survival", "Arena Shooter", "Psychological Horror", "Open World", "Action",
                      "Romance", "Drama", "Story Rich", "Precision Platformer", "Exploration",
```

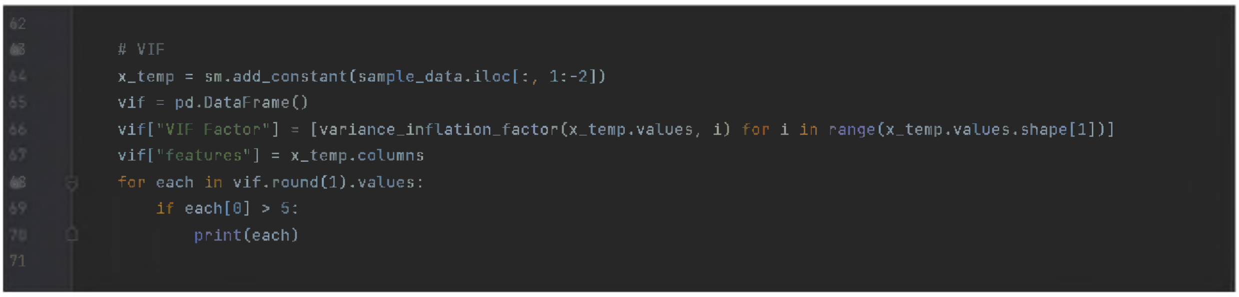


```
                      "Procedural Generation", "Experimental", "2D Platformer", "JRPG", "Survival Horror", "Stats",
                      "Colorful", "Early Access", "Funny", "Top-Down Shooter", "Board Game", "Score Attack", "Resource
                      "Side Scroller", "Choose Your Own Adventure", "Abstract"], axis=1)
        """
        # Cut languages
        # df = df.drop(["Polish", "Portuguese", "Swedish", "Spanish - Spain", "Czech", "Traditional Chinese", "English",
        test = df[:int(0.2*len(df))]
        sample = df[int(0.2*len(df)):].sort_values("M3")
        return sample, test

    @staticmethod
    def ols(data):
        # Simple ols model(change OLS to WLS for weighted least squares method)
        y = data["M3"]
        x = data[["Price", "Action", "May", "RPG", "Aliens", "PvP", "Story Rich"]]
        x = sm.add_constant(x)
        ols = sm.OLS(y, x)
        result = ols.fit()
        return result


if __name__ == '__main__':

    # Build model
    model = Model()
    sample_data, test_data = model.get_dataset(0.05)
    r = model.ols(sample_data)
    print(r.summary())
```



```
# VIF
x_temp = sm.add_constant(sample_data.iloc[:, 1:-2])
vif = pd.DataFrame()
vif["VIF Factor"] = [variance_inflation_factor(x_temp.values, i) for i in range(x_temp.values.shape[1])]
vif["features"] = x_temp.columns
for each in vif.round(1).values:
    if each[0] > 5:
        print(each)
```

Bibliography

Aziz, Amar et al. (July 2018). “Empirical Analysis on Sales of Video Games: A Data Mining Approach”. In: *Journal of Physics: Conference Series* 1049, p. 012086. ISSN: 1742-6588, 1742-6596. DOI: 10.1088/1742-6596/1049/1/012086. URL: https://iopscience.iop.org/article/10.1088/1742-6596/1049/1/012086.

Dille, Flint (2007). *The ultimate guide to video game writing and design*. New York: Watson-Guptill Publications. ISBN: 9781580650663.

Ibrisagic, Armin et al. (2016). *What Do We Mean When We Say "Indiepocalypse"?* San Francisco, CA. URL: https://www.gdcvault.com/play/1023441/What-Do-We-Mean-When.

Kontus, Karl (Aug. 2021). *How to Estimate Steam Video Game Sales in 2021?* en. URL: https://www.gamedeveloper.com/business/how-to-estimate-steam-video-game-sales-in-2021-.

Orland, Kyle (Apr. 2014). *Introducing Steam Gauge: Ars reveals Steam's most popular games*. en-us. URL: https://arstechnica.com/gaming/2014/04/introducing-steam-gauge-ars-reveals-steams-most-popular-games/ (visited on 05/26/2022).

Schell, Jesse (2008). *The art of game design: a book of lenses*. OCLC: ocn213839335. Amsterdam ; Boston: Elsevier/Morgan Kaufmann. ISBN: 9780123694966.

Trněný, Michal (2017). “Machine Learning for Predicting Success of Video Games”. en. MA thesis. Brno: Masaryk University. URL: https://is.muni.cz/th/k2c5b/diploma_thesis_trneny.pdf.

Wagner, Michael (Apr. 2021). *Marketing for Indie Devs: Market Research*. en. URL: https://www.gamedeveloper.com/business/marketing-for-indie-devs-market-research.

Wijman, Tom (Dec. 2021). *The Games Market and Beyond in 2021: The Year in Numbers*. en. URL: https://newzoo.com/insights/articles/the-games-market-in-2021-the-year-in-numbers-esports-cloud-gaming.

— (May 2022). *Games Market Revenues Will Pass $200 Billion for the First Time in 2022 as the U.S. Overtakes China*. en. URL: https://newzoo.com/insights/articles/games-market-revenues-will-pass-200-billion-for-the-first-time-in-2022-as-the-u-s-overtakes-china.

Wooldridge, Jeffrey M. (2020). *Introductory econometrics: a modern approach*. Seventh edition. Boston, MA: Cengage Learning. ISBN: 9781337558860.

Wright, Steven T. (Sept. 2018). *There are too many video games. What now?* en-US. URL: https://www.polygon.com/2018/9/28/17911372/there-are-too-many-video-games-what-now-indiepocalypse.

Yufa, Alice et al. (2019). “Predicting Global Video-Game Sales”. en. In: *Quest Journals Journal of Research in Business and Management*. Vol. 7, pp. 60–64. URL: https://www.questjournals.org/jrbm/papers/vol7-issue3/I07036064.pdf.