[2023] Pattern Recognition Projects (CS)

The objective of the projects is to prepare you to apply different machine learning algorithms to real-world tasks. This will help you to increase your knowledge about the workflow of the machine learning tasks. You will learn how to clean your data, applying pre-processing, feature engineering, regression, and classification methods. Each project will be delivered in milestones.

- The best three teams for each project will be honored.
- Registration starts: Friday 31/3/2023.
- Registration ends: Tuesday 4/4/2023.
- Delivering Milestone 1: 18/4/2023 11:59 PM Online.
- Delivering Milestone 2: Practical exam.
- Minimum number of members is 5 and the maximum is 6 or 7 with teams as 7 having an extra task mandatory
- You must deliver a detailed report for each milestone contains all your work (feature analysis, algorithms used in each module and the achieved accuracy for each one)

Note : Each report will be graded

In the first milestone, you will apply the followings :-

Preprocessing: Before building your models, you need to make sure that the dataset is clean and ready-to-use.

Regression: Apply different regression techniques (at least two) to find the model that fit your data with minimum error.

Milestone 1: 50%

- Preprocessing, Regression.

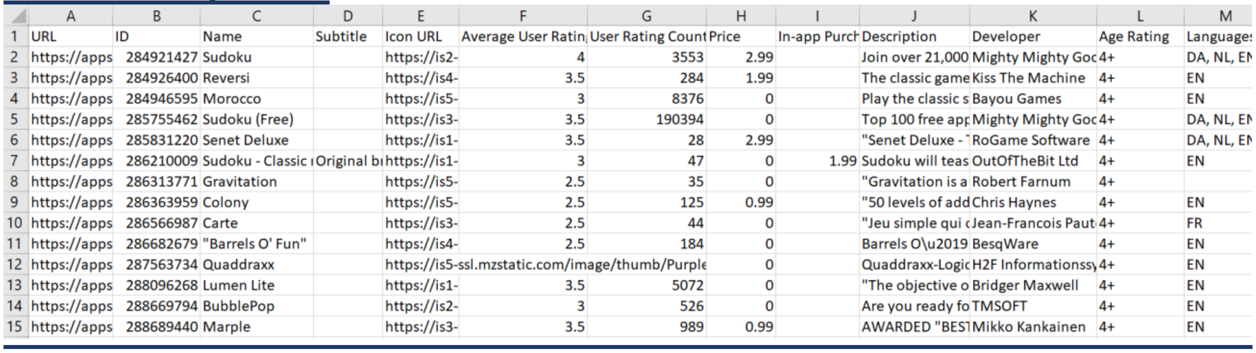
Milestone 1 Report Must Include:

- You must explain in details the **preprocessing techniques** you needed to apply on your dataset and how you implemented them.
- Perform **analysis** on the dataset as studied and explain how the features affect and relate to each other.
- You must explain what **regression techniques** you used (at least two).
- Mention the **differences** between each model and the acquired **results** (accuracy/error and so on).
- You must clearly mention **what features** you used or discarded to create your regression models.
- Explain what the **sizes** of your training, testing and validation sets are, if exist.
- Mention any further techniques that were used to **improve** the results (if exist).
- You should include **screenshots** of the resultant(s) regression line plots.
- Finally, write a **conclusion** about this phase of the project and what intuition you had about your problem and how it was proved/disproved.

Project(1): Game Application Success Prediction

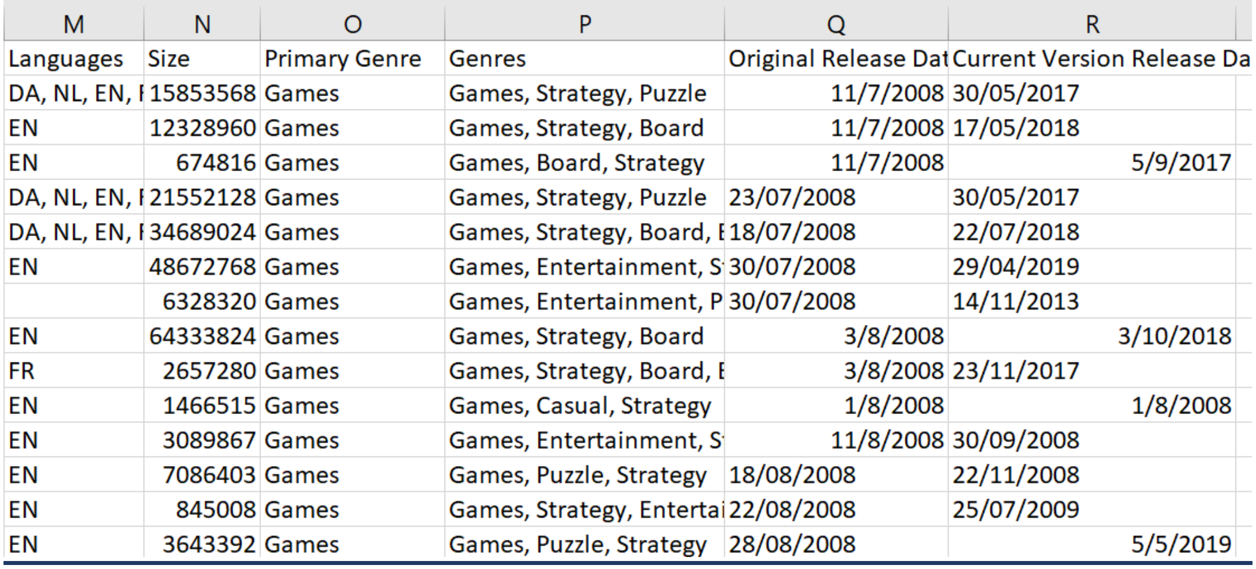
The mobile games industry is worth billions of dollars, with companies spending vast amounts of money on the development and marketing of these games to an equally large market. Using this data set, insights can be gained into this market.

Dataset Snapshots:



	A	B	C	D	E	F	G	H	I	J	K	L	M
1	URL	ID	Name	Subtitle	Icon URL	Average User Ratin	User Rating Count	Price	In-app Purch	Description	Developer	Age Rating	Language:
2	https://apps	284921427	Sudoku		https://is2-	4	3553	2.99		Join over 21,000	Mighty Mighty Goc	4+	DA, NL, EN
3	https://apps	284926400	Reversi		https://is4-	3.5	284	1.99		The classic game	Kiss The Machine	4+	EN
4	https://apps	284946595	Morocco		https://is5-	3	8376	0		Play the classic s	Bayou Games	4+	EN
5	https://apps	285755462	Sudoku (Free)		https://is3-	3.5	190394	0		Top 100 free app	Mighty Mighty Goc	4+	DA, NL, EN
6	https://apps	285831220	Senet Deluxe		https://is1-	3.5	28	2.99		"Senet Deluxe - T	RoGame Software	4+	DA, NL, EN
7	https://apps	286210009	Sudoku - Classic ı	Original bı	https://is1-	3	47	0	1.99	Sudoku will teas	OutOfTheBit Ltd	4+	EN
8	https://apps	286313771	Gravitation		https://is5-	2.5	35	0		"Gravitation is a	Robert Farnum	4+	
9	https://apps	286363959	Colony		https://is5-	2.5	125	0.99		"50 levels of add	Chris Haynes	4+	EN
10	https://apps	286566987	Carte		https://is3-	2.5	44	0		"Jeu simple qui c	Jean-Francois Paut	4+	FR
11	https://apps	286682679	"Barrels O' Fun"		https://is4-	2.5	184	0		Barrels O\u2019	BesqWare	4+	EN
12	https://apps	287563734	Quaddraxx		https://is5-ssl.mzstatic.com/image/thumb/Purple			0		Quaddraxx-Logic	H2F Informationssy	4+	EN
13	https://apps	288096268	Lumen Lite		https://is1-	3.5	5072	0		"The objective o	Bridger Maxwell	4+	EN
14	https://apps	288669794	BubblePop		https://is2-	3	526	0		Are you ready fo	TMSOFT	4+	EN
15	https://apps	288689440	Marple		https://is3-	3.5	989	0.99		AWARDED "BEST	Mikko Kankainen	4+	EN

~Dataset header Continued:



M	N	O	P	Q	R
Languages	Size	Primary Genre	Genres	Original Release Dat	Current Version Release Da
DA, NL, EN, I	15853568	Games	Games, Strategy, Puzzle	11/7/2008	30/05/2017
EN	12328960	Games	Games, Strategy, Board	11/7/2008	17/05/2018
EN	674816	Games	Games, Board, Strategy	11/7/2008	5/9/2017
DA, NL, EN, I	21552128	Games	Games, Strategy, Puzzle	23/07/2008	30/05/2017
DA, NL, EN, I	34689024	Games	Games, Strategy, Board, E	18/07/2008	22/07/2018
EN	48672768	Games	Games, Entertainment, S	30/07/2008	29/04/2019
	6328320	Games	Games, Entertainment, P	30/07/2008	14/11/2013
EN	64333824	Games	Games, Strategy, Board	3/8/2008	3/10/2018
FR	2657280	Games	Games, Strategy, Board, E	3/8/2008	23/11/2017
EN	1466515	Games	Games, Casual, Strategy	1/8/2008	1/8/2008
EN	3089867	Games	Games, Entertainment, S	11/8/2008	30/09/2008
EN	7086403	Games	Games, Puzzle, Strategy	18/08/2008	22/11/2008
EN	845008	Games	Games, Strategy, Entertai	22/08/2008	25/07/2009
EN	3643392	Games	Games, Puzzle, Strategy	28/08/2008	5/5/2019

Dataset Descriptions:

Feature	Description
ID	
Name	
Subtitle	The secondary text under the name
Icon URL	

Average User Rating	Rounded to nearest .5, requires at least 5 ratings
User Rating Count	Number of ratings internationally, null means it is below 5
Price	
In App Purchases	Prices of available in-app purchases
Description	
Developer	
Age Rating	Either 4+, 9+, 12+ or 17+
Languages	
Size	
Primary Genre	Main genre
Genres	Genres of the app
Original Release Date	
Current Version Release Date	

Milestone 1 tasks:

1. Apply pre-processing on the provided dataset. (You must preprocess all the features even if you won't use them later after feature selection)
2. Apply Feature Selection and Experiment with regression techniques to reduce the error on prediction of the "Average User Rating" (Deliver at least two regression models with significant difference).
3. Finish Milestone 1 Report.

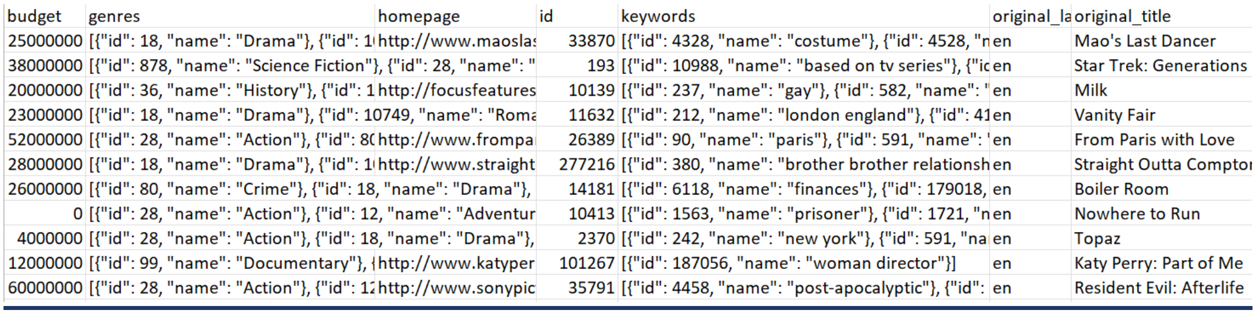
Bonus Task: Extract meaningful feature from description column

Note: You must preprocess all features, but model and feature selection can be done after that (i.e You can drop a feature only after preprocessing and with valid reason)

Project(2): Movie Popularity Prediction

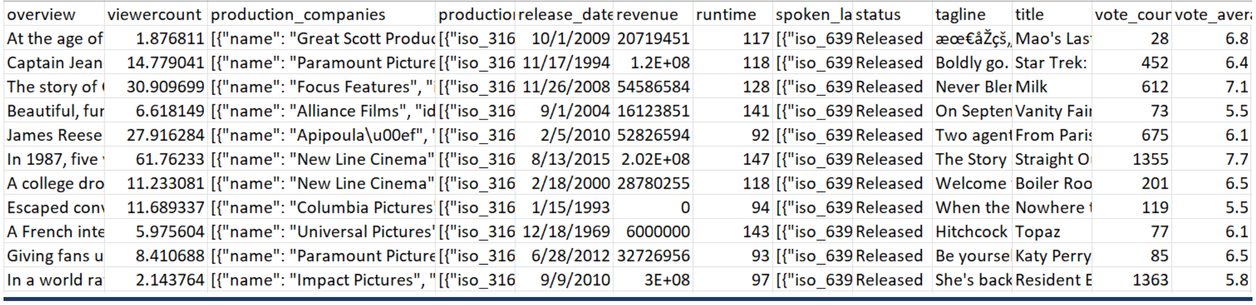
What can we say about the success of a movie before it is released? Are there certain companies (Pixar?) that have found a consistent formula? Given that major films costing over $100 million to produce can still flop, this question is more important than ever to the industry. Can we predict which films will be highly rated, whether or not they are a commercial success?

Dataset Snapshots:



budget	genres	homepage	id	keywords	original_la	original_title
25000000	[{"id": 18, "name": "Drama"}, {"id": 1	http://www.maoslas	33870	[{"id": 4328, "name": "costume"}, {"id": 4528, "n	en	Mao's Last Dancer
38000000	[{"id": 878, "name": "Science Fiction"}, {"id": 28, "name": "		193	[{"id": 10988, "name": "based on tv series"}, {"ic	en	Star Trek: Generations
20000000	[{"id": 36, "name": "History"}, {"id": 1	http://focusfeatures	10139	[{"id": 237, "name": "gay"}, {"id": 582, "name": "	en	Milk
23000000	[{"id": 18, "name": "Drama"}, {"id": 10749, "name": "Roma		11632	[{"id": 212, "name": "london england"}, {"id": 41	en	Vanity Fair
52000000	[{"id": 28, "name": "Action"}, {"id": 8(	http://www.frompa	26389	[{"id": 90, "name": "paris"}, {"id": 591, "name": '	en	From Paris with Love
28000000	[{"id": 18, "name": "Drama"}, {"id": 1	http://www.straight	277216	[{"id": 380, "name": "brother brother relationsh	en	Straight Outta Compto
26000000	[{"id": 80, "name": "Crime"}, {"id": 18, "name": "Drama"},		14181	[{"id": 6118, "name": "finances"}, {"id": 179018,	en	Boiler Room
0	[{"id": 28, "name": "Action"}, {"id": 12, "name": "Adventur		10413	[{"id": 1563, "name": "prisoner"}, {"id": 1721, "n	en	Nowhere to Run
4000000	[{"id": 28, "name": "Action"}, {"id": 18, "name": "Drama"},		2370	[{"id": 242, "name": "new york"}, {"id": 591, "na	en	Topaz
12000000	[{"id": 99, "name": "Documentary"}, {	http://www.katyper	101267	[{"id": 187056, "name": "woman director"}]	en	Katy Perry: Part of Me
60000000	[{"id": 28, "name": "Action"}, {"id": 12	http://www.sonypic	35791	[{"id": 4458, "name": "post-apocalyptic"}, {"id":	en	Resident Evil: Afterlife

~Dataset header Continued:



overview	viewercount	production_companies	productio	release_date	revenue	runtime	spoken_la	status	tagline	title	vote_cour	vote_avera
At the age of	1.876811	[{"name": "Great Scott Produ	[{"iso_316	10/1/2009	20719451	117	[{"iso_639	Released	æœ€åŽçš„	Mao's Las	28	6.8
Captain Jean	14.779041	[{"name": "Paramount Picture	[{"iso_316	11/17/1994	1.2E+08	118	[{"iso_639	Released	Boldly go.	Star Trek:	452	6.4
The story of	30.909699	[{"name": "Focus Features", "	[{"iso_316	11/26/2008	54586584	128	[{"iso_639	Released	Never Bler	Milk	612	7.1
Beautiful, fur	6.618149	[{"name": "Alliance Films", "id	[{"iso_316	9/1/2004	16123851	141	[{"iso_639	Released	On Septen	Vanity Fai	73	5.5
James Reese	27.916284	[{"name": "Apipoula\u00ef", '	[{"iso_316	2/5/2010	52826594	92	[{"iso_639	Released	Two agent	From Paris	675	6.1
In 1987, five	61.76233	[{"name": "New Line Cinema"	[{"iso_316	8/13/2015	2.02E+08	147	[{"iso_639	Released	The Story	Straight O	1355	7.7
A college dro	11.233081	[{"name": "New Line Cinema"	[{"iso_316	2/18/2000	28780255	118	[{"iso_639	Released	Welcome	Boiler Roo	201	6.5
Escaped con	11.689337	[{"name": "Columbia Pictures	[{"iso_316	1/15/1993	0	94	[{"iso_639	Released	When the	Nowhere	119	5.5
A French inte	5.975604	[{"name": "Universal Pictures'	[{"iso_316	12/18/1969	6000000	143	[{"iso_639	Released	Hitchcock	Topaz	77	6.1
Giving fans u	8.410688	[{"name": "Paramount Picture	[{"iso_316	6/28/2012	32726956	93	[{"iso_639	Released	Be yoursel	Katy Perry	85	6.5
In a world ra	2.143764	[{"name": "Impact Pictures", "	[{"iso_316	9/9/2010	3E+08	97	[{"iso_639	Released	She's back	Resident E	1363	5.8

Dataset Description:

Feature	Description
Budget	Cost of making the movie
Genres	A list of the genres that the movie belongs to. (i.e Avatar is a movie that has several genres some of which are action, adventure, science)
Homepage	
Id	
Keywords	
Original Language	
Original title	
Overview	General plot description
ViewerCount	Number of viewers
Production Companies	

Production Countries	
Release Date	
Revenue	Profit
Runtime	Movie duration in Minutes
Spoken Languages	
Status	
Tagline	
Title	
Vote Average	Average movie rating from 0 - 10
Vote Count	Number of voters for the average movie rating

Milestone 1 tasks:

1. Apply pre-processing on the provided dataset. (You must preprocess all the features even if you won't use them later after feature selection)
2. Apply Feature Selection and Experiment with regression techniques to reduce the error on prediction of the "Vote Average" (Deliver at least two regression models with significant difference).
3. Finish Milestone 1 Report.

Bonus Task: Using the second excel file in a meaningful way.

Note: You must preprocess all features, but model and feature selection can be done after that (i.e You can drop a feature only after preprocessing and with valid reason)