Machine Learning Projects (SC)

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 ends: 07/11/2019.
- ➢ Delivering Milestone 1: 30/11/2019.
- ➢ Delivering Milestone 2: Practical exam.
- ➢ Minimum number of members is 3 and the maximum is 5
- ➢ 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.

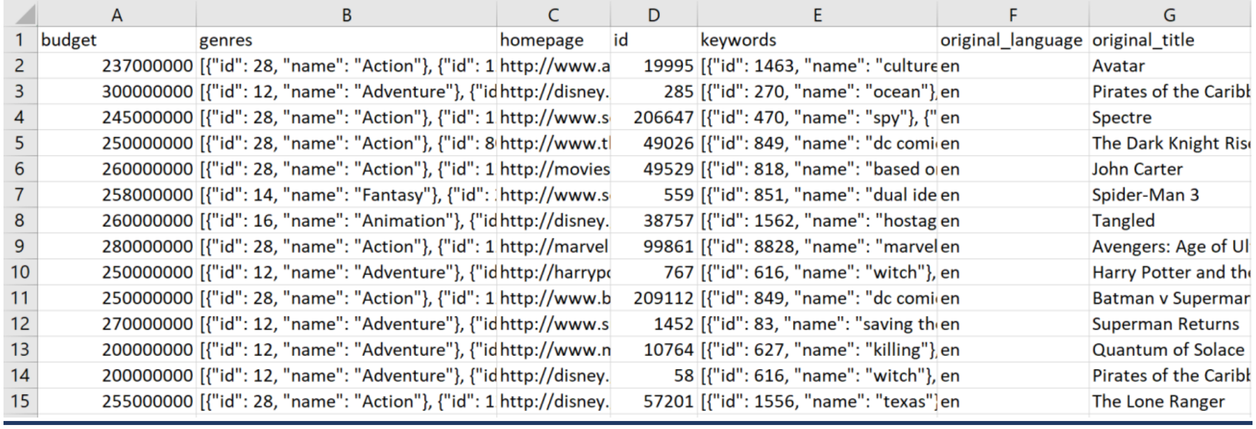
Milestone 2 Deliverables will be announced later.

Project(1): Predicting Movie Success

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?

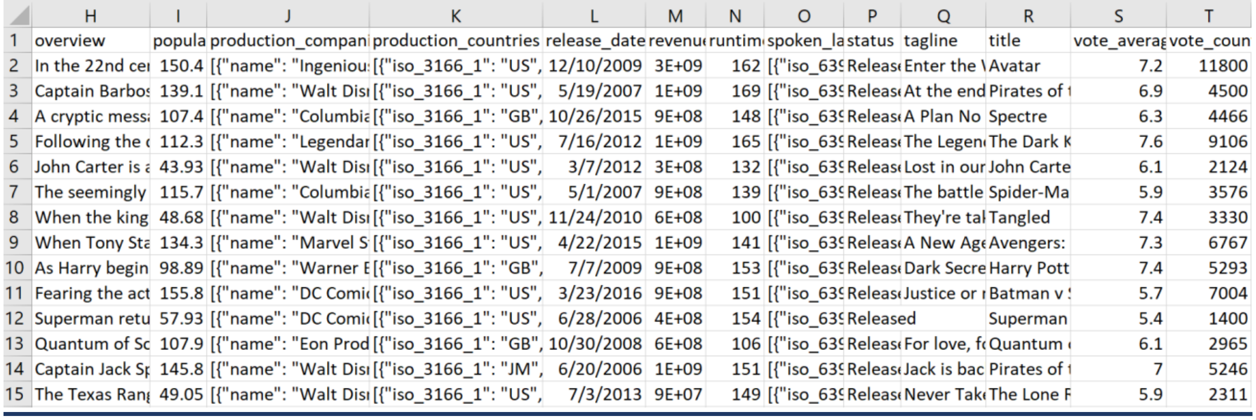
This is a great place to start digging in to those questions, with data on the plot, cast, crew, budget, and revenues of several thousand films.

Dataset Snapshots:



	A	B	C	D	E	F	G
1	budget	genres	homepage	id	keywords	original_language	original_title
2	237000000	[{"id": 28, "name": "Action"}, {"id": 1	http://www.a	19995	[{"id": 1463, "name": "culture	en	Avatar
3	300000000	[{"id": 12, "name": "Adventure"}, {"id	http://disney.	285	[{"id": 270, "name": "ocean"},	en	Pirates of the Caribb
4	245000000	[{"id": 28, "name": "Action"}, {"id": 1	http://www.s	206647	[{"id": 470, "name": "spy"}, {"	en	Spectre
5	250000000	[{"id": 28, "name": "Action"}, {"id": 8	http://www.t	49026	[{"id": 849, "name": "dc comi	en	The Dark Knight Ris
6	260000000	[{"id": 28, "name": "Action"}, {"id": 1	http://movies	49529	[{"id": 818, "name": "based o	en	John Carter
7	258000000	[{"id": 14, "name": "Fantasy"}, {"id":	http://www.s	559	[{"id": 851, "name": "dual ide	en	Spider-Man 3
8	260000000	[{"id": 16, "name": "Animation"}, {"id	http://disney.	38757	[{"id": 1562, "name": "hostag	en	Tangled
9	280000000	[{"id": 28, "name": "Action"}, {"id": 1	http://marvel	99861	[{"id": 8828, "name": "marvel	en	Avengers: Age of Ul
10	250000000	[{"id": 12, "name": "Adventure"}, {"id	http://harryp	767	[{"id": 616, "name": "witch"},	en	Harry Potter and th
11	250000000	[{"id": 28, "name": "Action"}, {"id": 1	http://www.b	209112	[{"id": 849, "name": "dc comi	en	Batman v Supermar
12	270000000	[{"id": 12, "name": "Adventure"}, {"id	http://www.s	1452	[{"id": 83, "name": "saving th	en	Superman Returns
13	200000000	[{"id": 12, "name": "Adventure"}, {"id	http://www.n	10764	[{"id": 627, "name": "killing"},	en	Quantum of Solace
14	200000000	[{"id": 12, "name": "Adventure"}, {"id	http://disney.	58	[{"id": 616, "name": "witch"},	en	Pirates of the Caribb
15	255000000	[{"id": 28, "name": "Action"}, {"id": 1	http://disney.	57201	[{"id": 1556, "name": "texas"}	en	The Lone Ranger

~Dataset header Continued:



	H	I	J	K	L	M	N	O	P	Q	R	S	T
1	overview	popula	production_compani	production_countries	release_date	revenu	runtim	spoken_la	status	tagline	title	vote_averag	vote_coun
2	In the 22nd ce	150.4	[{"name": "Ingeniou	[{"iso_3166_1": "US",	12/10/2009	3E+09	162	[{"iso_639	Releas	Enter the \	Avatar	7.2	11800
3	Captain Barbo	139.1	[{"name": "Walt Dis	[{"iso_3166_1": "US",	5/19/2007	1E+09	169	[{"iso_639	Releas	At the end	Pirates of	6.9	4500
4	A cryptic mess	107.4	[{"name": "Columbi	[{"iso_3166_1": "GB",	10/26/2015	9E+08	148	[{"iso_639	Releas	A Plan No	Spectre	6.3	4466
5	Following the	112.3	[{"name": "Legendar	[{"iso_3166_1": "US",	7/16/2012	1E+09	165	[{"iso_639	Releas	The Legen	The Dark K	7.6	9106
6	John Carter is	43.93	[{"name": "Walt Dis	[{"iso_3166_1": "US",	3/7/2012	3E+08	132	[{"iso_639	Releas	Lost in our	John Carte	6.1	2124
7	The seemingly	115.7	[{"name": "Columbi	[{"iso_3166_1": "US",	5/1/2007	9E+08	139	[{"iso_639	Releas	The battle	Spider-Ma	5.9	3576
8	When the king	48.68	[{"name": "Walt Dis	[{"iso_3166_1": "US",	11/24/2010	6E+08	100	[{"iso_639	Releas	They're tal	Tangled	7.4	3330
9	When Tony Sta	134.3	[{"name": "Marvel S	[{"iso_3166_1": "US",	4/22/2015	1E+09	141	[{"iso_639	Releas	A New Age	Avengers:	7.3	6767
10	As Harry begin	98.89	[{"name": "Warner E	[{"iso_3166_1": "GB",	7/7/2009	9E+08	153	[{"iso_639	Releas	Dark Secre	Harry Pott	7.4	5293
11	Fearing the act	155.8	[{"name": "DC Comi	[{"iso_3166_1": "US",	3/23/2016	9E+08	151	[{"iso_639	Releas	Justice or	Batman v	5.7	7004
12	Superman retu	57.93	[{"name": "DC Comi	[{"iso_3166_1": "US",	6/28/2006	4E+08	154	[{"iso_639	Released		Superman	5.4	1400
13	Quantum of Sc	107.9	[{"name": "Eon Prod	[{"iso_3166_1": "GB",	10/30/2008	6E+08	106	[{"iso_639	Releas	For love, f	Quantum	6.1	2965
14	Captain Jack Sp	145.8	[{"name": "Walt Dis	[{"iso_3166_1": "JM",	6/20/2006	1E+09	151	[{"iso_639	Releas	Jack is bac	Pirates of	7	5246
15	The Texas Ran	49.05	[{"name": "Walt Dis	[{"iso_3166_1": "US",	7/3/2013	9E+07	149	[{"iso_639	Releas	Never Tak	The Lone F	5.9	2311

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
Popularity	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

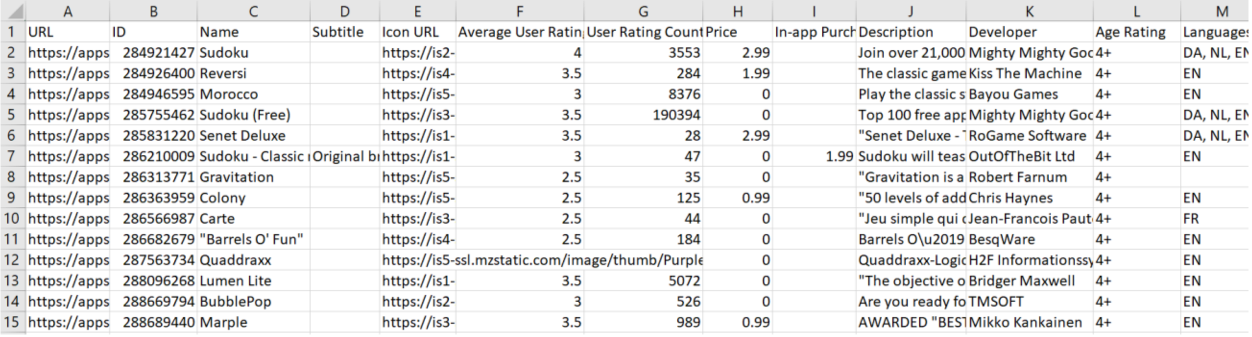
Milestone 1 tasks:

1. Apply pre-processing on the provided dataset.
2. Experiment with regression techniques to reduce the error on prediction of the average rating of a movie (Deliver at least two techniques).

Project(2): Predicting Mobile Game Success

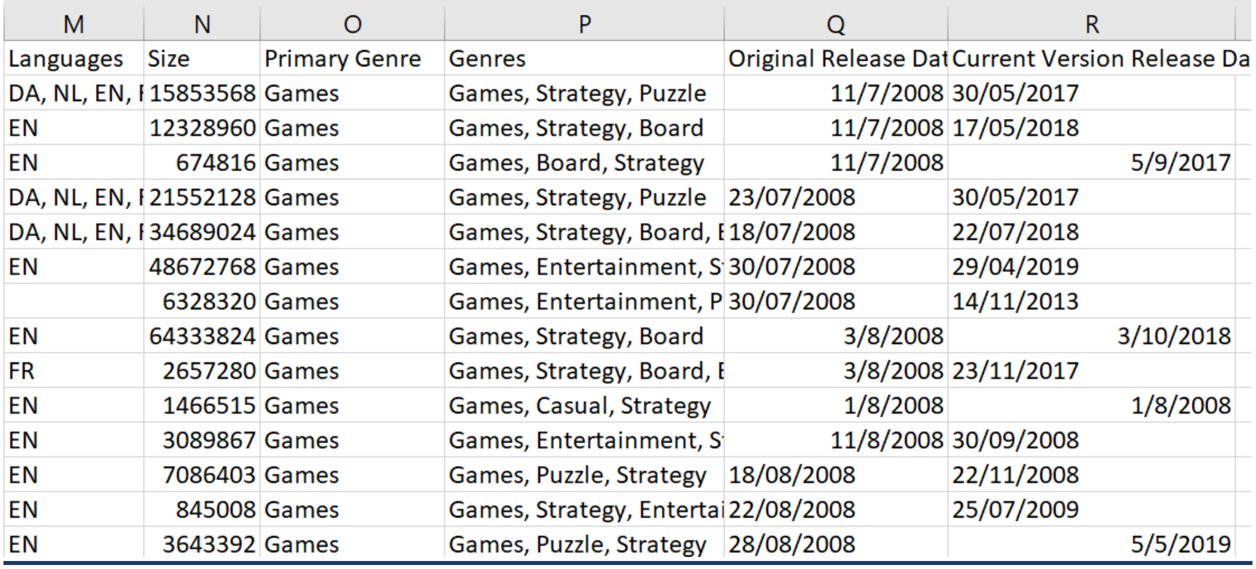
The mobile games industry is worth billions of dollars, with companies spending vast amounts of money on the development and marketing of these games to an equally large market. Using this data set, insights can be gained into this market.

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	Original b	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, F	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, F	21552128	Games	Games, Strategy, Puzzle	23/07/2008	30/05/2017
DA, NL, EN, F	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.
2. Experiment with regression techniques to reduce the error on prediction of the average user rating of a game app (Deliver at least two techniques).