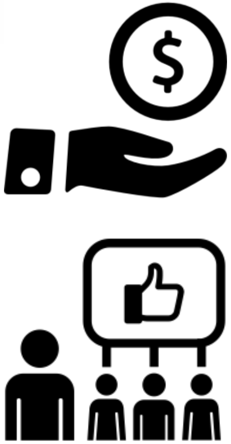
Actors Tours of Data Science

NTDS-2019

- Students: Andres Montero, Ariel Alba, Elias Poroma

EPFL, 22-01-2020

Introduction



Dataset

Using IMDb 5000 movies dataset

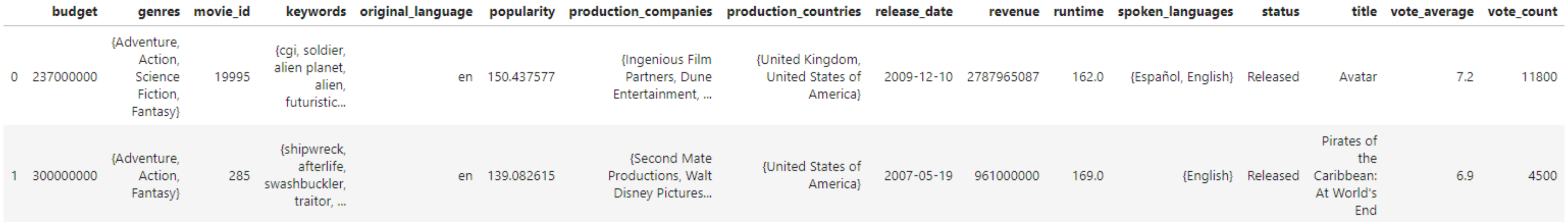
- Remove movies with missing information in important columns (budget, revenue, popularity, votes)
- Combine movie dataset with credits dataset (includes cast, crew and actors).
- Take the most important actors in each movie and create an actor dataset.

Tools used:

- Python
- Pandas
- Seaborn
- Gephi
- Pygsp
- Scikit-learn
- Networkx
- Matplotlib
- Python Louvain

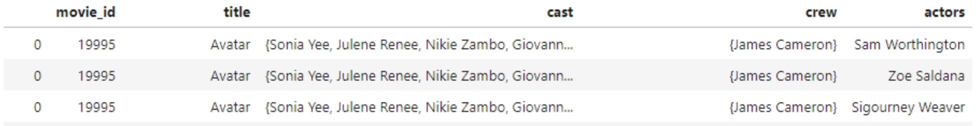
Dataset

Movies dataset



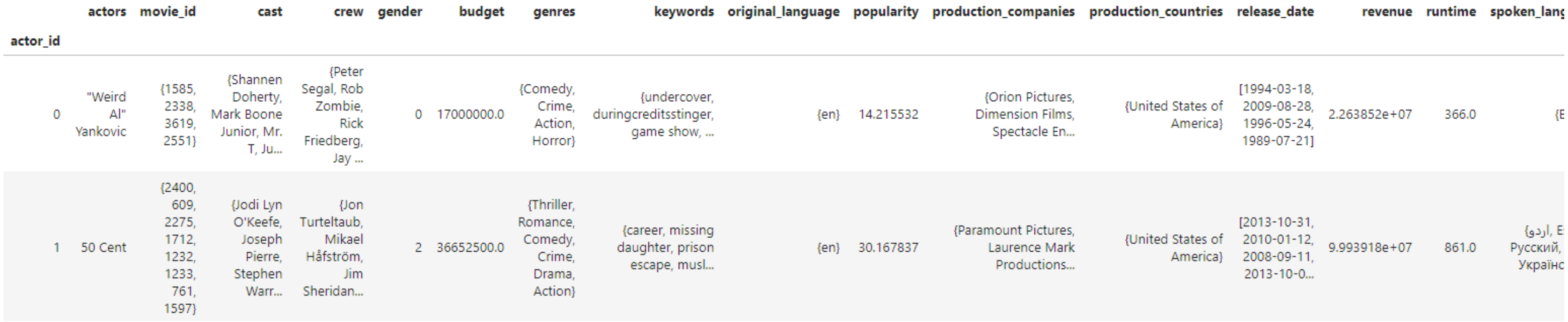
	budget	genres	movie_id	keywords	original_language	popularity	production_companies	production_countries	release_date	revenue	runtime	spoken_languages	status	title	vote_average	vote_count
0	237000000	{Adventure, Action, Science Fiction, Fantasy}	19995	{cgi, soldier, alien planet, alien, futuristic...	en	150.437577	{Ingenious Film Partners, Dune Entertainment, ...	{United Kingdom, United States of America}	2009-12-10	2787965087	162.0	{Español, English}	Released	Avatar	7.2	11800
1	300000000	{Adventure, Action, Fantasy}	285	{shipwreck, afterlife, swashbuckler, traitor, ...	en	139.082615	{Second Mate Productions, Walt Disney Pictures...	{United States of America}	2007-05-19	961000000	169.0	{English}	Released	Pirates of the Caribbean: At World's End	6.9	4500

Movie credits dataset



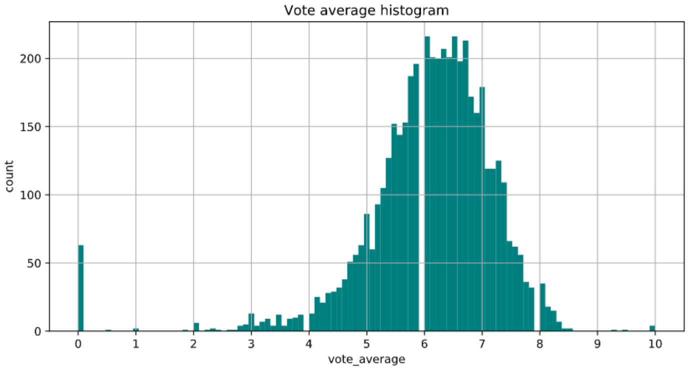
	movie_id	title	cast	crew	actors
0	19995	Avatar	{Sonia Yee, Julene Renee, Nikie Zambo, Giovann...	{James Cameron}	Sam Worthington
0	19995	Avatar	{Sonia Yee, Julene Renee, Nikie Zambo, Giovann...	{James Cameron}	Zoe Saldana
0	19995	Avatar	{Sonia Yee, Julene Renee, Nikie Zambo, Giovann...	{James Cameron}	Sigourney Weaver

NEW actors dataset

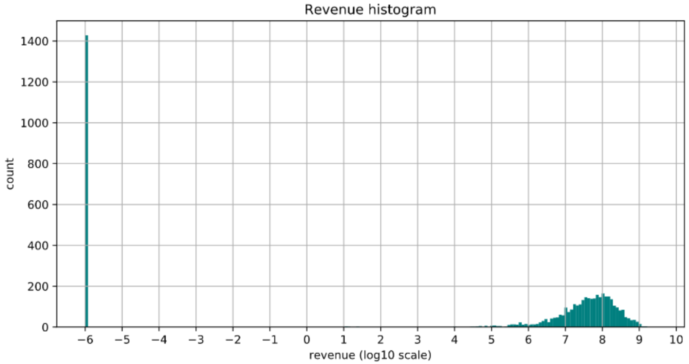


	actors	movie_id	cast	crew	gender	budget	genres	keywords	original_language	popularity	production_companies	production_countries	release_date	revenue	runtime	spoken_lang
actor_id																
0	"Weird Al" Yankovic	{1585, 2338, 3619, 2551}	{Shannen Doherty, Mark Boone Junior, Mr. T, Ju...	{Peter Segal, Rob Zombie, Rick Friedberg, Jay ...	0	17000000.0	{Comedy, Crime, Action, Horror}	{undercover, duringcreditsstinger, game show, ...	{en}	14.215532	{Orion Pictures, Dimension Films, Spectacle En...	{United States of America}	[1994-03-18, 2009-08-28, 1996-05-24, 1989-07-21]	2.263852e+07	366.0	{E
1	50 Cent	{2400, 609, 2275, 1712, 1232, 1233, 761, 1597}	{Jodi Lyn O'Keefe, Joseph Pierre, Stephen Warr...	{Jon Turteltaub, Mikael Håfström, Jim Sheridan...	2	36652500.0	{Thriller, Romance, Comedy, Crime, Drama, Action}	{career, missing daughter, prison escape, musl...	{en}	30.167837	{Paramount Pictures, Laurence Mark Productions...	{United States of America}	[2013-10-31, 2010-01-12, 2008-09-11, 2013-10-0...	9.993918e+07	861.0	{اردو, E Русский, Українс

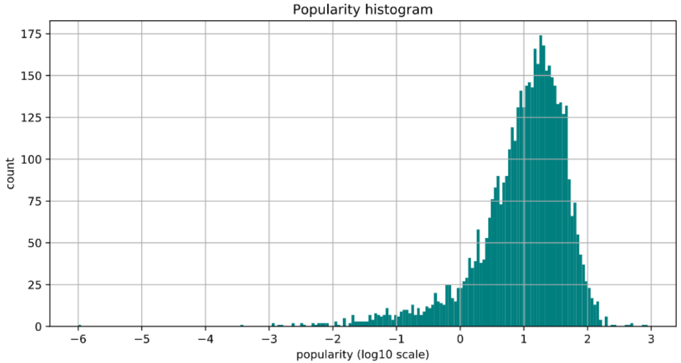
Dataset



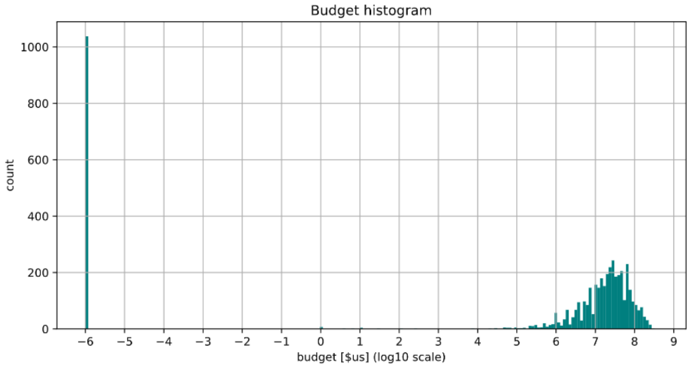
Vote average histogram



Revenue histogram



Popularity histogram



Budget histogram

Exploration

Graph created from the actor dataset, define weight between actors:

$$w_{ij} = \frac{0.3|movie_i \cap movie_j| + 0.3|cast_i \cap cast_j| + 0.2|crew_i \cap crew_j| + 0.1|genre_i \cap genre_j + 0.1|companies_i \cap companies_j}{0.3|movie_i \cup movie_j| + 0.3|cast_i \cup cast_j| + 0.2|crew_i \cup crew_j| + 0.1|genre_i \cup genre_j + 0.1|companies_i \cap companies_j}$$

Graph Properties:

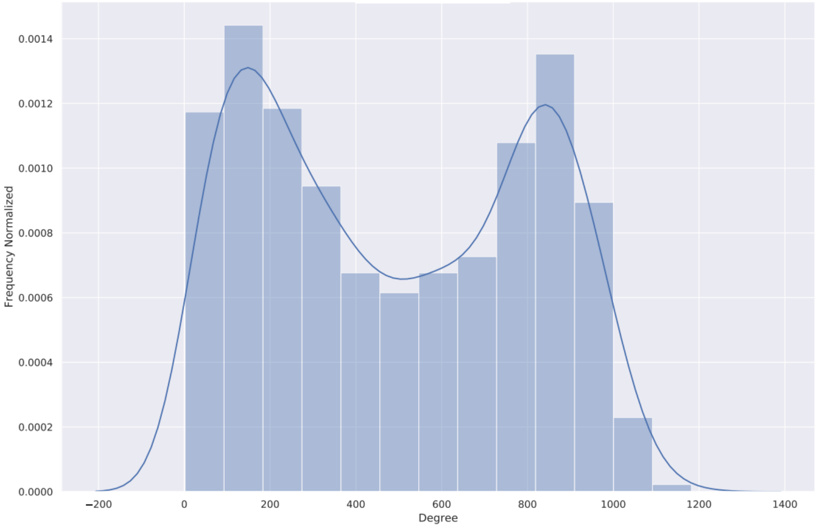
- 1 connected component.
- Diameter of the graph is 4.
- Type of graph is small world.

Nodes Properties:

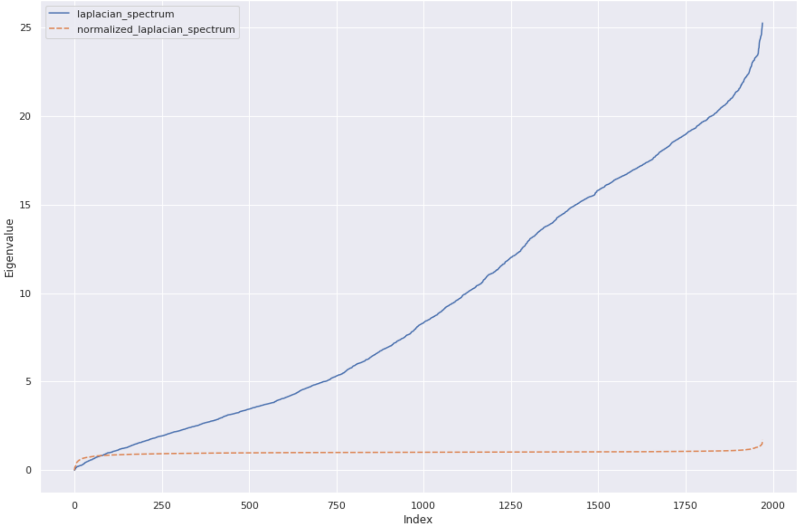
- Average Degree: 492.75
- Average clustering coefficient: 0.6016
- Hub nodes: 958 hubs.

Exploration

Degree Distribution

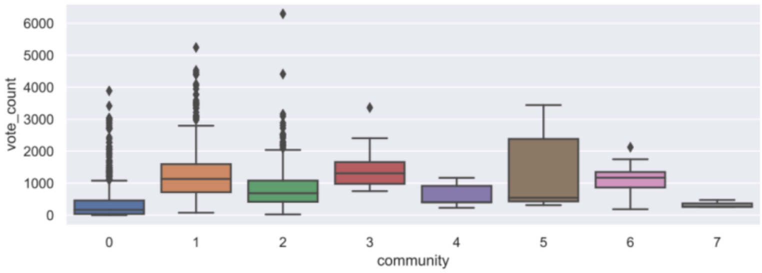


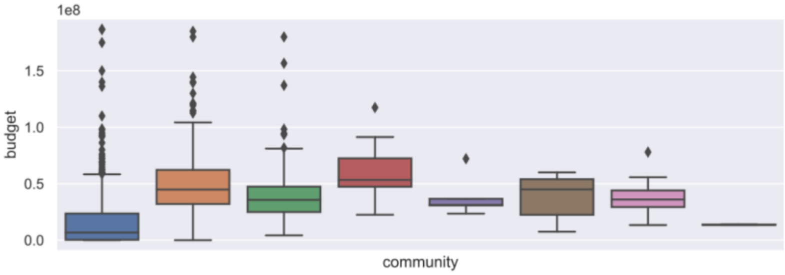
Laplacian and Normalized Laplacian

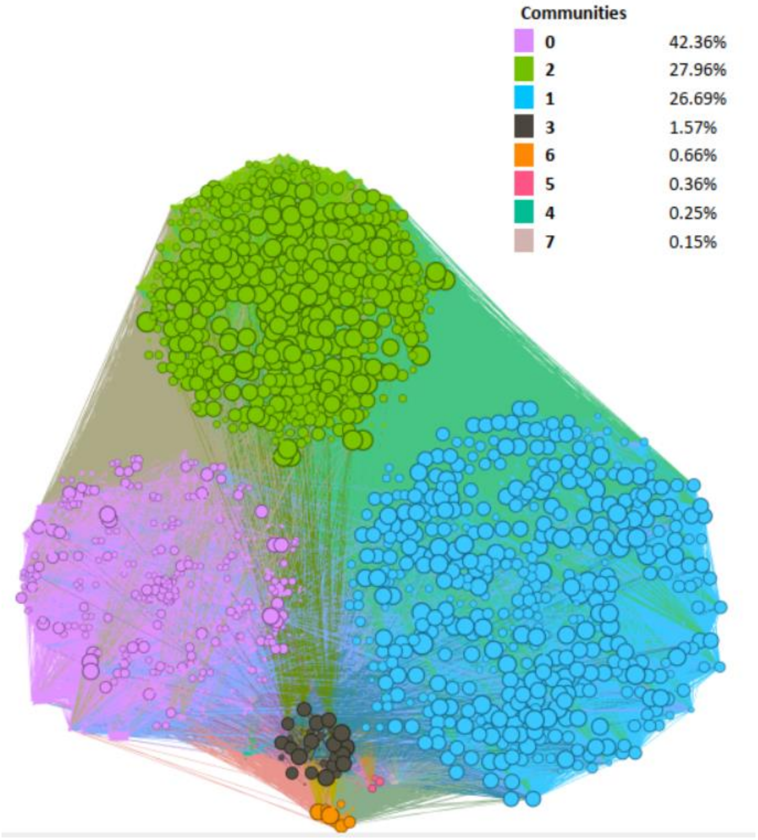


Exploitation

- Louvain: Community detection based on modularity.

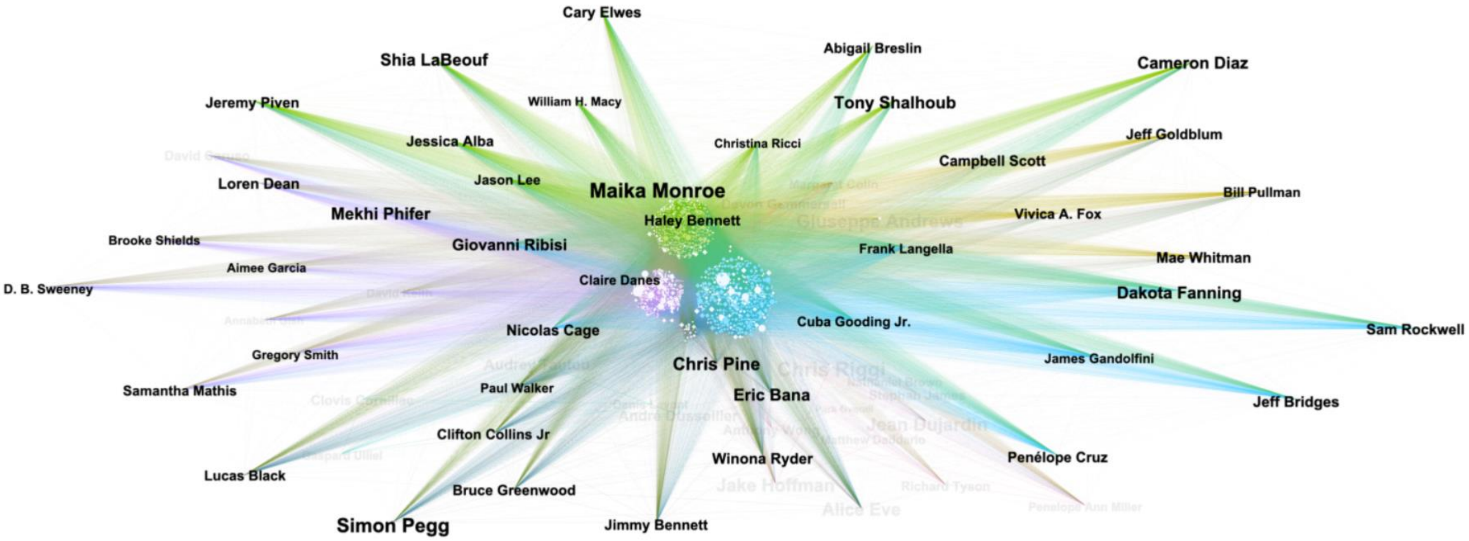






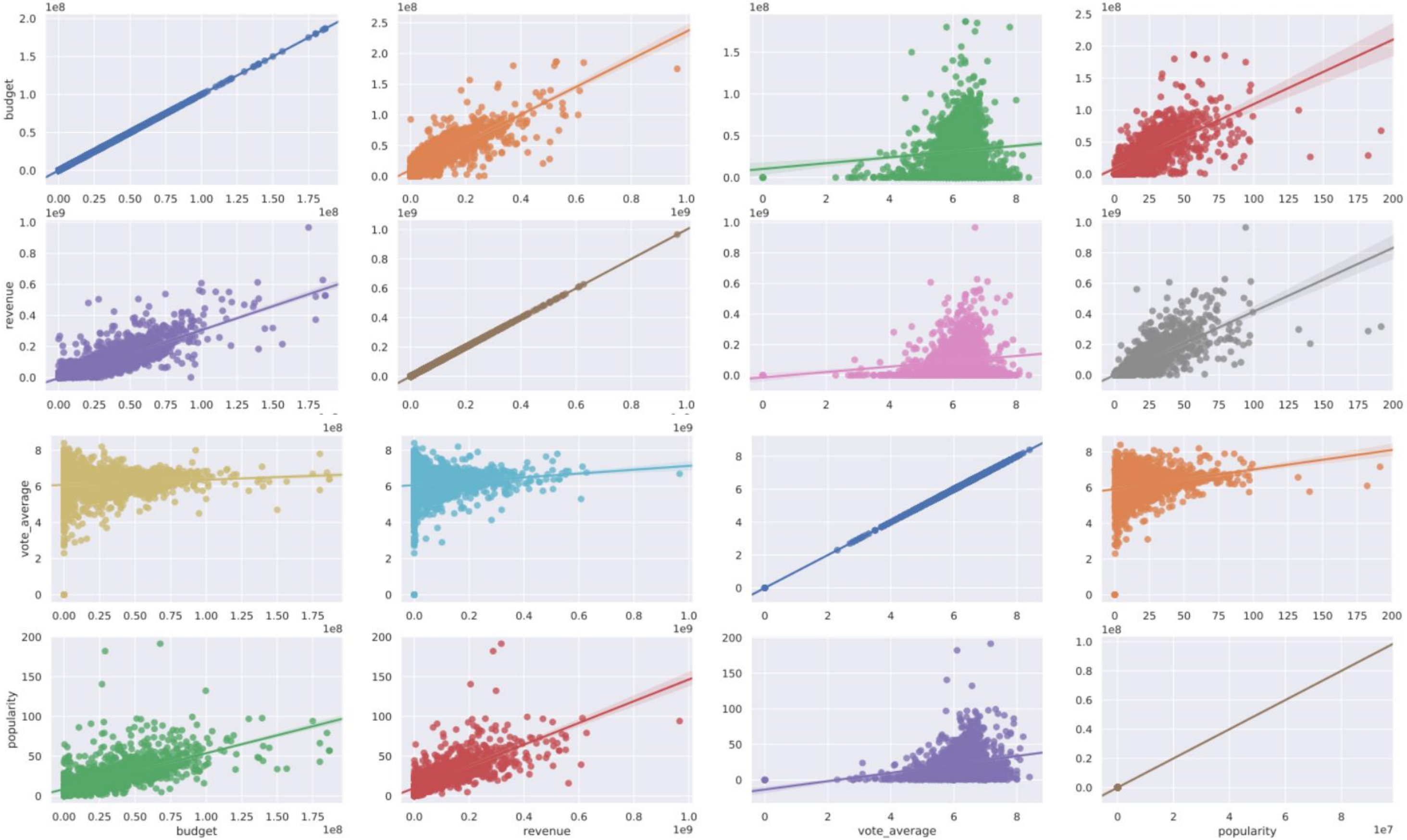
Exploitation

- Most representative actors based on revenue and popularity



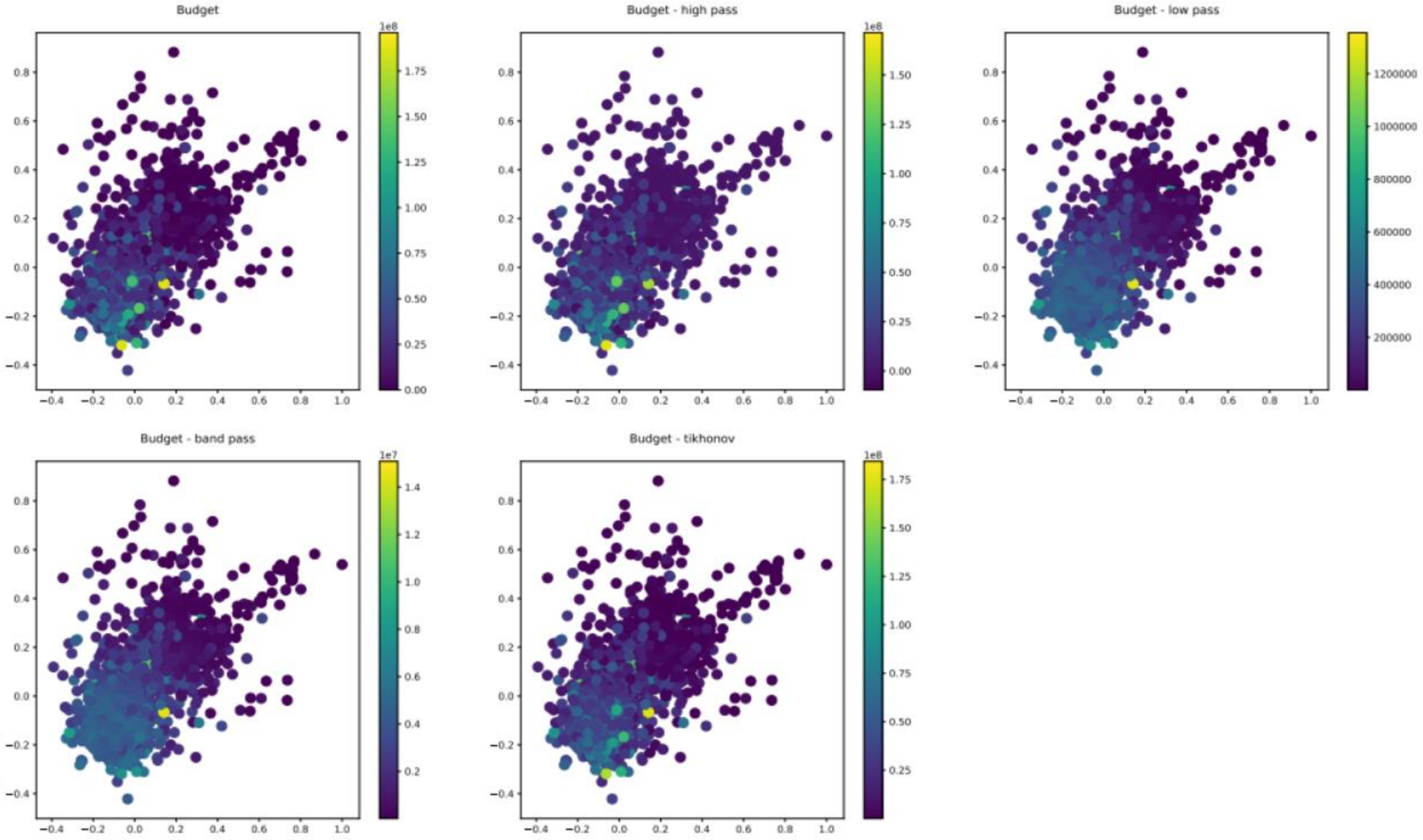
Exploitation

- Linear Regression



Exploitation

- Filters: Low pass, band pass, high pass and tikhonov.



Results and Conclusions

Signal / Method	V Lin Reg	V Lin Reg - Com	Lin Reg - HP	Lin Reg - HP - Com	Lin Reg - LP	Lin Reg - LP - Com	Lin Reg - BP	Lin Reg - BP - Com	Lin Reg - TK	Lin Reg - TK - Com
Budget	0.0537	0.0489	0.063	0.0583	0.0302	0.0295	0.0299	**0.0291**	0.0505	0.0465
Revenue	0.0447	0.0435	0.036	0.0354	0.0206	0.0206	0.0199	**0.0199**	0.0283	0.0281
Popularity	0.0267	0.0265	0.0647	0.0637	0.0437	0.043	0.0432	**0.0425**	0.0607	0.0596
Vote Average	0.0834	0.0827	0.0334	0.0699	0.0334	0.033	0.0333	**0.0334**	0.0519	0.0519

QUESTIONS?