Movie Dataset Visualization and IMDB Score Prediction

Grigor Keropyan[1]

[1]Department of Mathematcis and Mechanics
Yerevan State University

29 may 2020

About Dataset

https://www.kaggle.com/rounakbanik/the-movies-dataset

```
df
```

original_title	popularity	production_countries	release_date	revenue	runtime	vote_average	vote_count	vote_level
Toy Story	21.946943	United States of America	1995-10-30	373554033.0	81.0	7.7	5415.0	high_level
Jumanji	17.015539	United States of America	1995-12-15	262797249.0	104.0	6.9	2413.0	high_level
Grumpier Old Men	11.712900	United States of America	1995-12-22	104765121.0	101.0	6.5	92.0	high_level
Waiting to Exhale	3.859495	United States of America	1995-12-22	81452156.0	127.0	6.1	34.0	high_level
of the Bride Part II	8.387519	United States of America	1995-02-10	76578911.0	106.0	5.7	173.0	high_level

Figure: Dataset

Problem Definition

Explore Movies dataset. Using data visualization techniques detect some dependencies between dataset features. Particularly, how popular is the movie depend on the what is the original language of it. Finally business probelm is to predict a IMDB score of the movies based on the information available prior to the theatrical release.

Original language vs popularity

Spanish?
France?
English?
Armenian?

Original language vs popularity

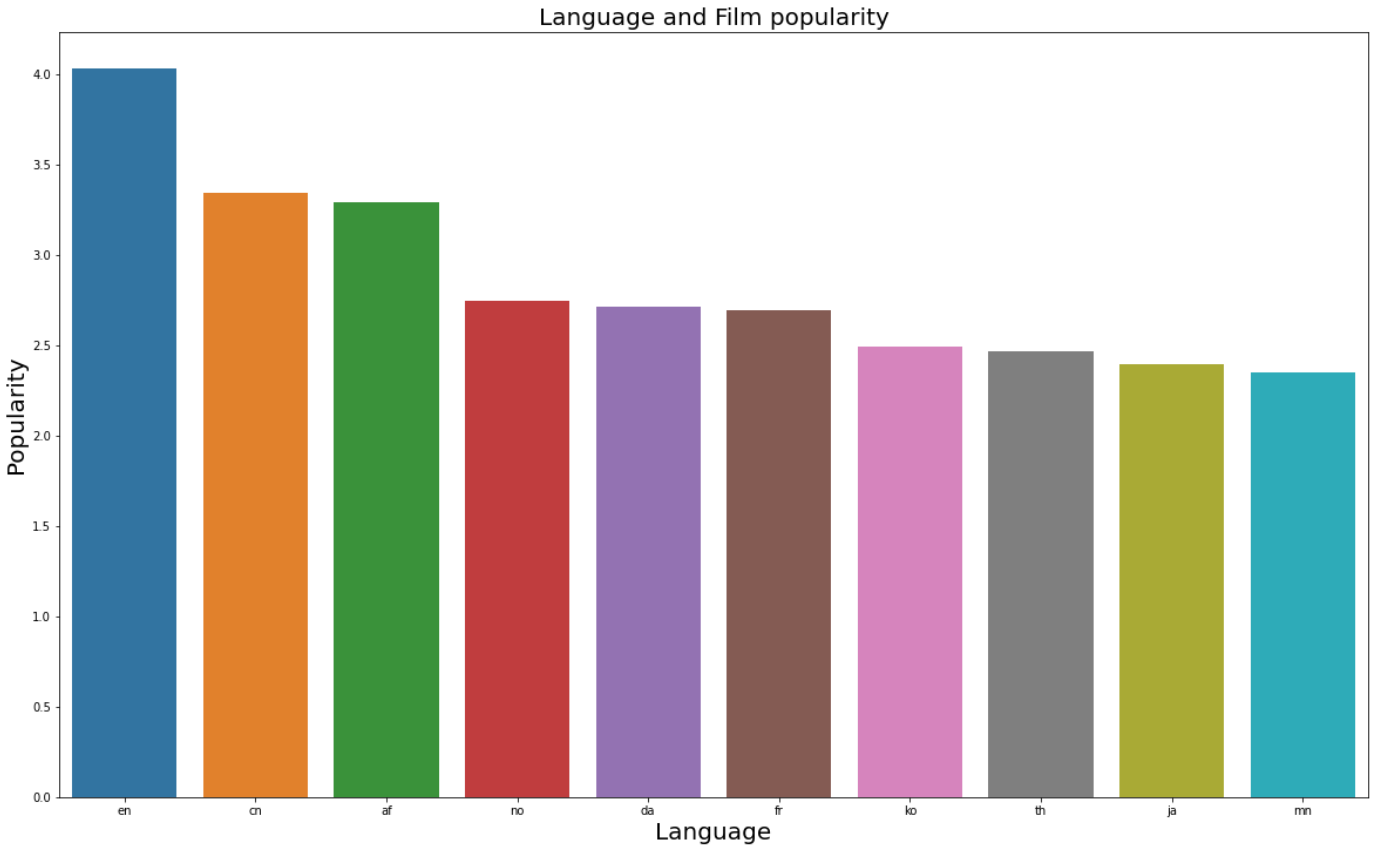


Figure: Movies popularity depends on their original language

Genre count in movies

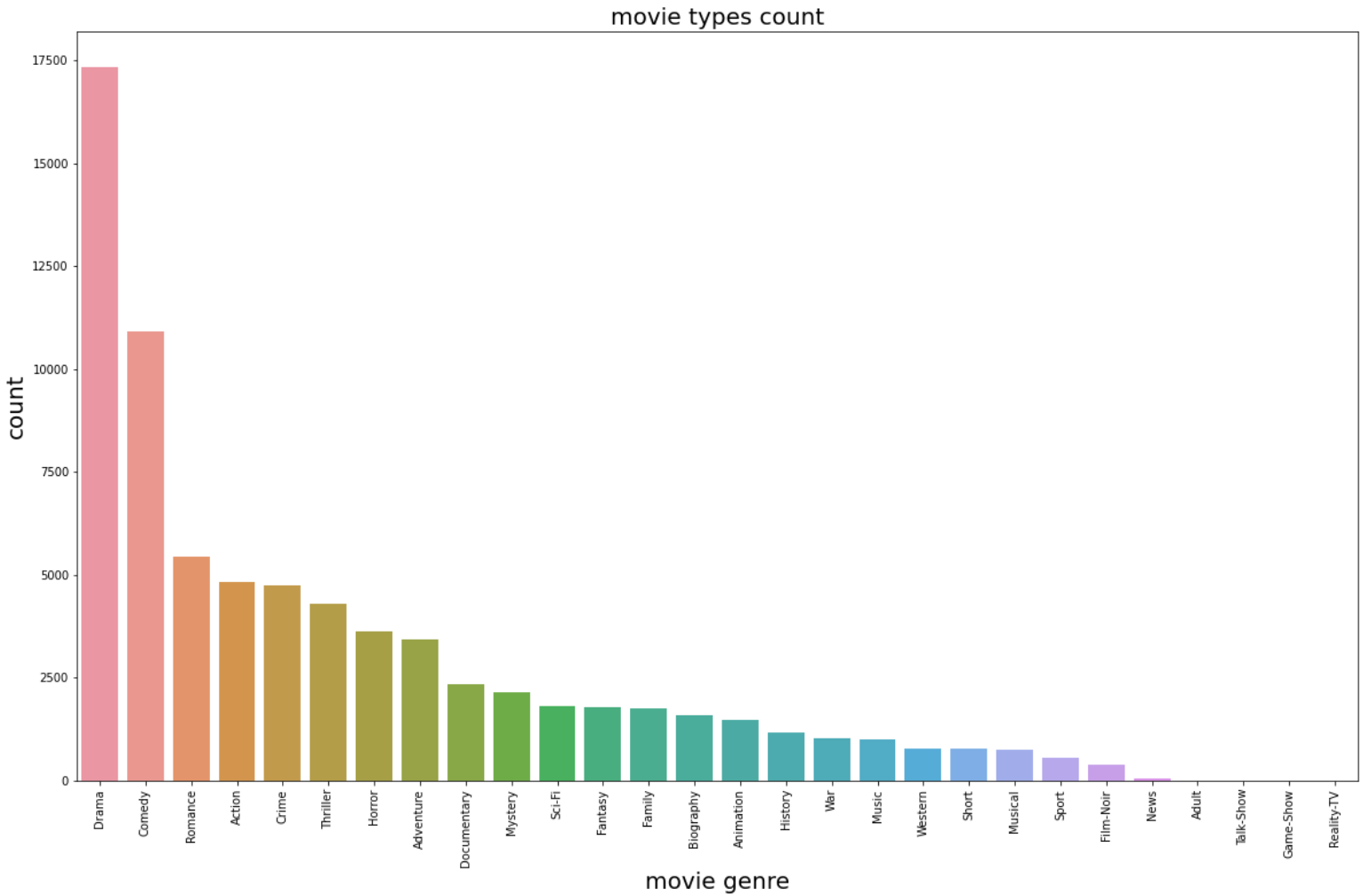


Figure: Genre count in movies

Runtime vs Popularity

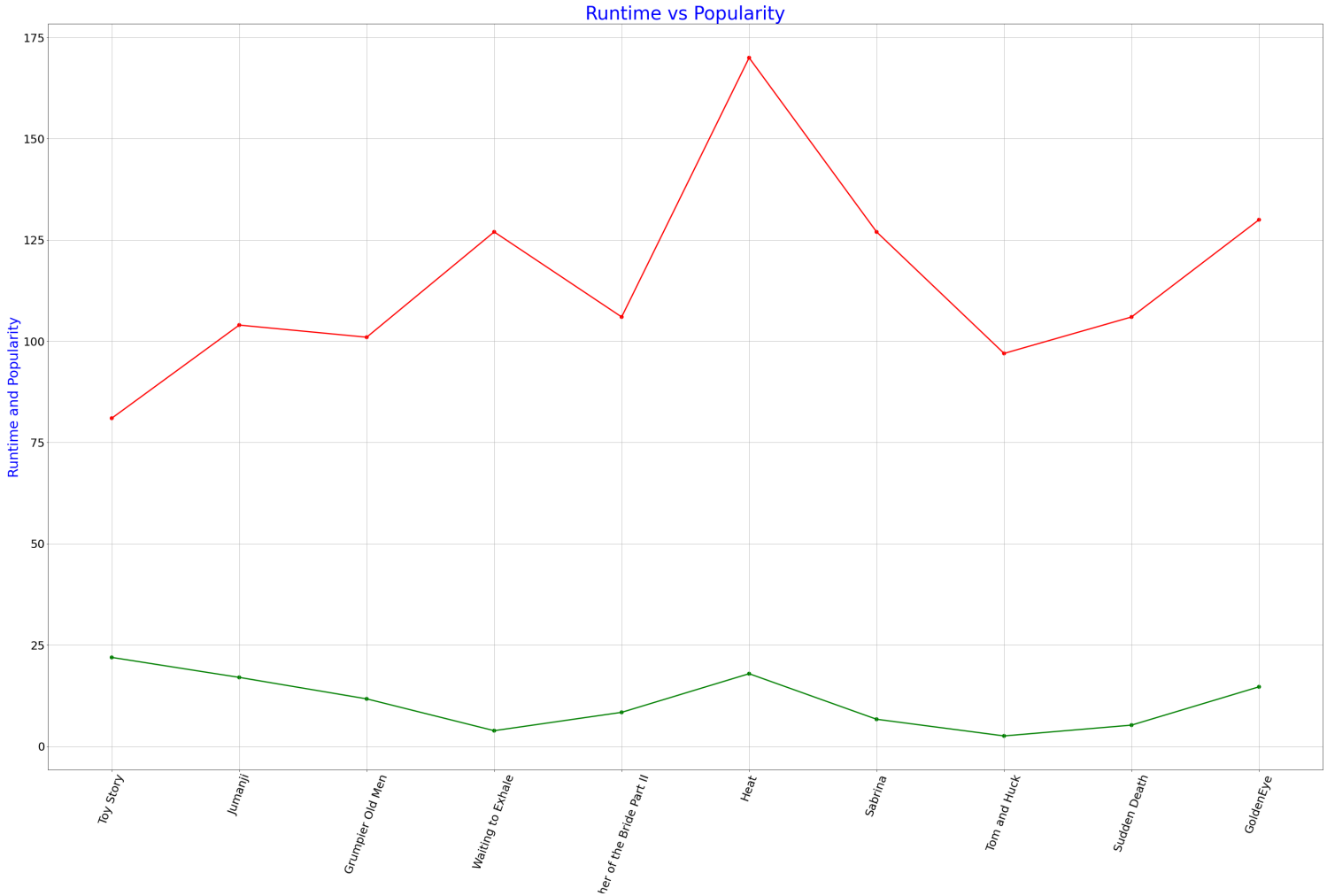
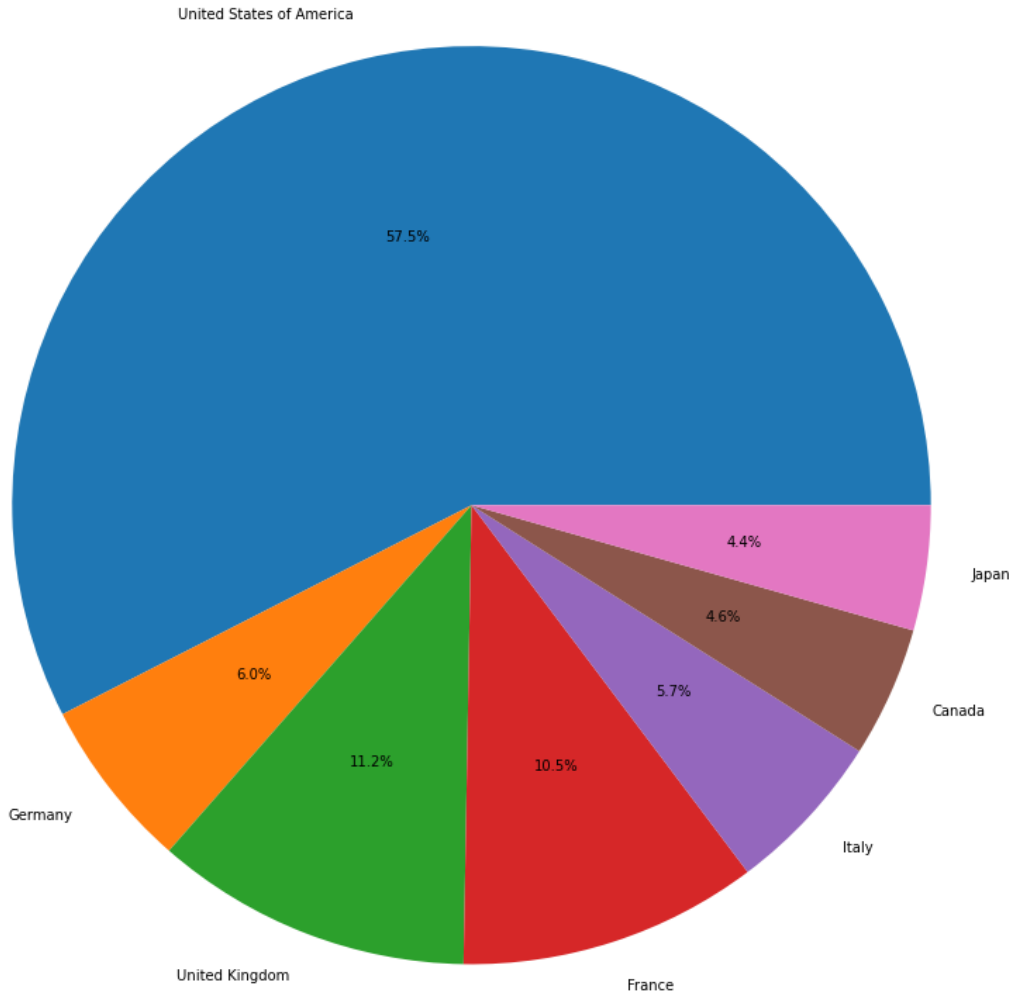


Figure: Runtime and Popularity connection

Production countries



Genres in WordCloud

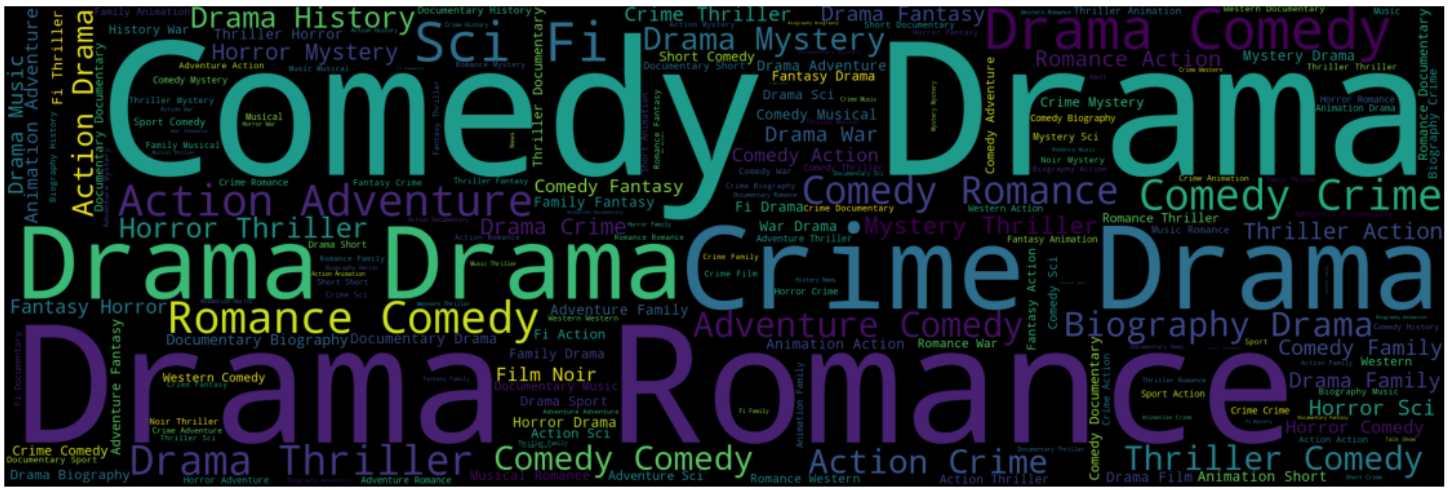


Figure: wordcloud visualization of movie genres

Ridge regression on IMDB score prediction

	alpha	train_r2	val_r2	test_r2	train_mse	val_mse	test_mse
4	100.0	0.492799	0.352405	0.390991	0.749172	0.923437	0.908345
3	10.0	0.634264	0.296577	0.346365	0.540217	1.003045	0.974906
2	5.0	0.662290	0.255346	0.306424	0.498821	1.061839	1.034478
1	1.0	0.693804	0.137866	0.191024	0.452273	1.229359	1.206599
0	0.1	0.699453	0.043802	0.097634	0.443929	1.363489	1.345892

Figure: Ridge regression result on different alphas

Conclusion from the plots

1. From the pie and bar plots we evidently understand that the most popular and movies come from the US. More than half of the movies are produced in United State Of America and most popular movies original language is in English.

Conclusion from the plots

2. From the wordcloud and bar plots we evidently understand that the vast majority of movies have a Drama genre.

Conclusion from the plots

3. Finally from the line graph we might be able to conclude that runtime of movies is connected with their popularity in some cases.

Conclusion from the regression

Ridge regression performed well on the dataset and this result could be used for new movies IMDB score prediction

Thank you !