MOVIE RATINGS PREDICTION

MATYÁŠ MATTANELLI

OUTLINE

- Introduction
- Data Collection
- Data Analysis
- Linear Regression
- Language Model
- Predictive analysis

INTRODUCTION

Movie information obtained from www.csfd.cz

The goal is to try to predict the ratings of a movie based on various information contained on the website

Statistical analysis was conducted to show which features are significant determinants of movie ratings

DATA COLLECTION

- Each movie has its own website
- Web scraping performed using the Python Selenium module
- Data for 47 776 movies were obtained
- Features: Genre, Country of origin, Year of release, Duration, Director, Preview, Number of ratings, Number of fans
- Ratings
  - Value between 0 and 100
  - Users evaluate the movie on a scale from 0 (worst) to 5 (best)

Vykoupení z věznice Shawshank

The Shawshank Redemption (více)

Drama / Krimi

USA, 1994, 142 min

Režie: Frank Darabont

Předloha: Stephen King (povídka)

Scénář: Frank Darabont

Kamera: Roger Deakins

Hudba: Thomas Newman

Hrají: Tim Robbins, Morgan Freeman, Bob Gunton, William Sadler, Clancy Brown, Gil Bellows, Mark Rolston, James Whitmore, Jeffrey DeMunn, Larry Brandenburg (více)

(další profese)

Obsahy (2)

zobrazit všechny obsahy

Mladý bankéř Andy Dufresne (Tim Robbins) je v roce 1947 odsouzen za vraždu své ženy a jejího milence. Přesto, že tento čin popírá, čeká na něj dvojnásobný doživotní trest v obávané věznici Shawshank. Andy se snaží přizpůsobit vězeňskému životu a po krušných začátcích se sblíží s Redem (Morgan Freeman), jenž si tu odpykává svůj doživotní trest už dvacet let. Sílu, jak přežít zdejší peklo, Andy nachází v tajném snu a skryté naději na svobodu... *(Magic Box)*

95%

1. nejlepší

2. nejoblíbenější

Hodnocení (112 127)	Fanklub (17 067)
golfista ★★★★★	
POMO ★★★★★	
kleopatra ★★★★★	
verbal ★★★	
Cival ★★★★★	
KevSpa ★★★★★	
Douglas ★★★★★	
kOCOUR ★★★★★	
Houdini ★★★★★	
don corleone ★★★★★	

1 - 10 >

OVLÁDACÍ PANEL

NETFLIX

APPLE TV+

O2TV

GOOGLE PLAY

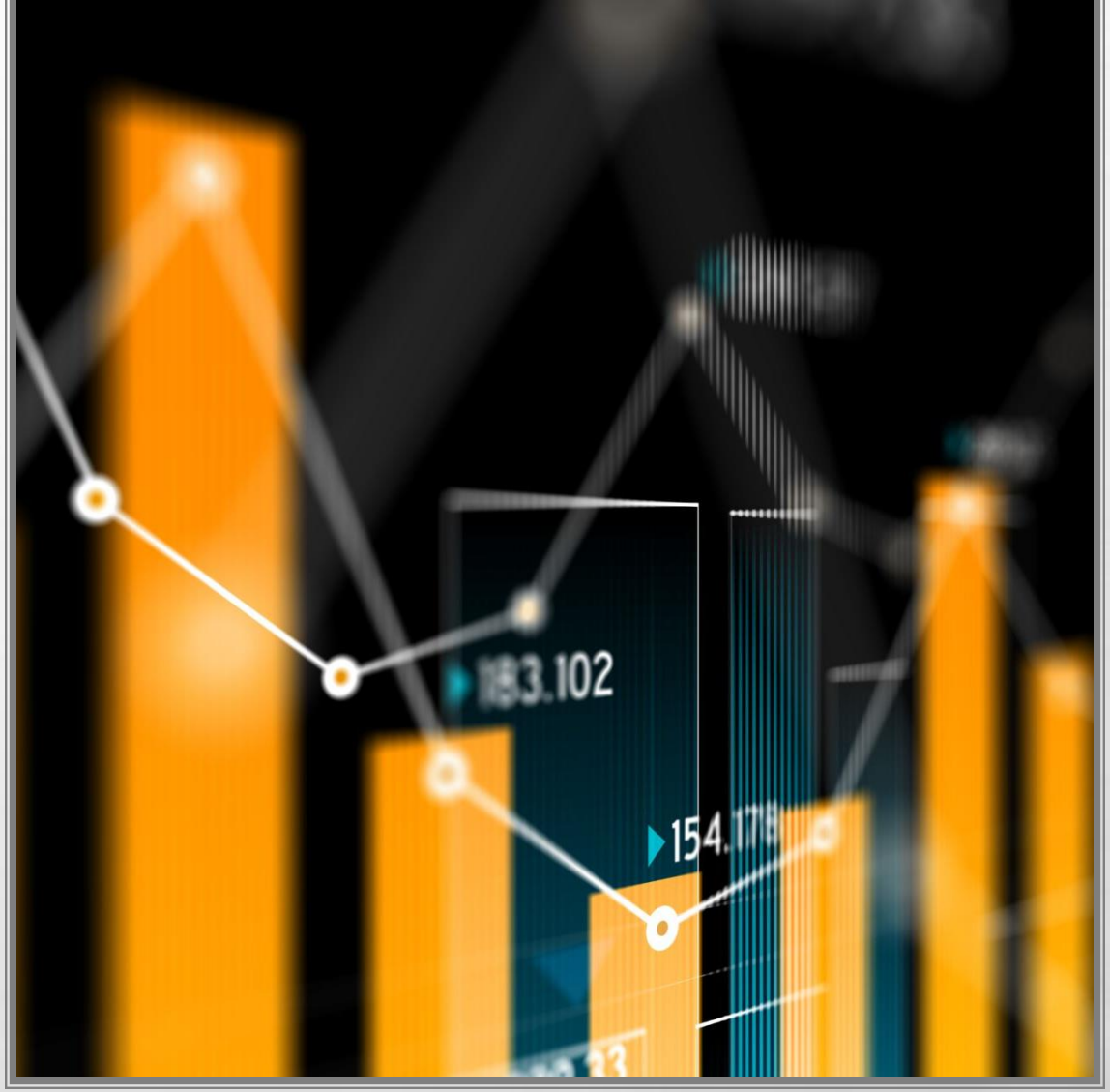
DVD

BLU-RAY

KNIHA

DATA COLLECTION

- Extensive data processing and feature engineering was required given the textual nature of the data
- A movie can be flagged with multiple genres and can be based in several countries
  - Dummy variable for each genre and each country
- Year was converted to the age of a movie
- Due to the high number of unique directors, the number of directed movies (prior to the given movie) was considered instead



DATA ANALYSIS

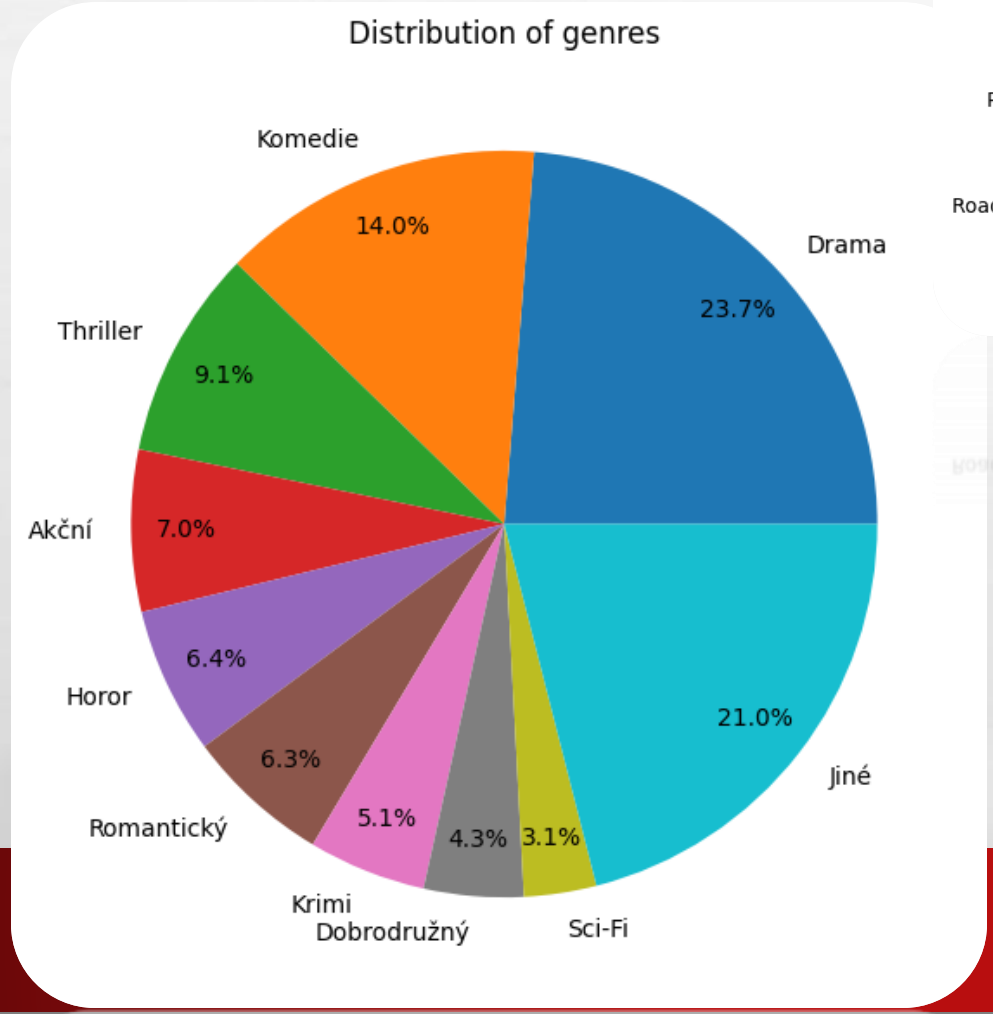
47 776 movies

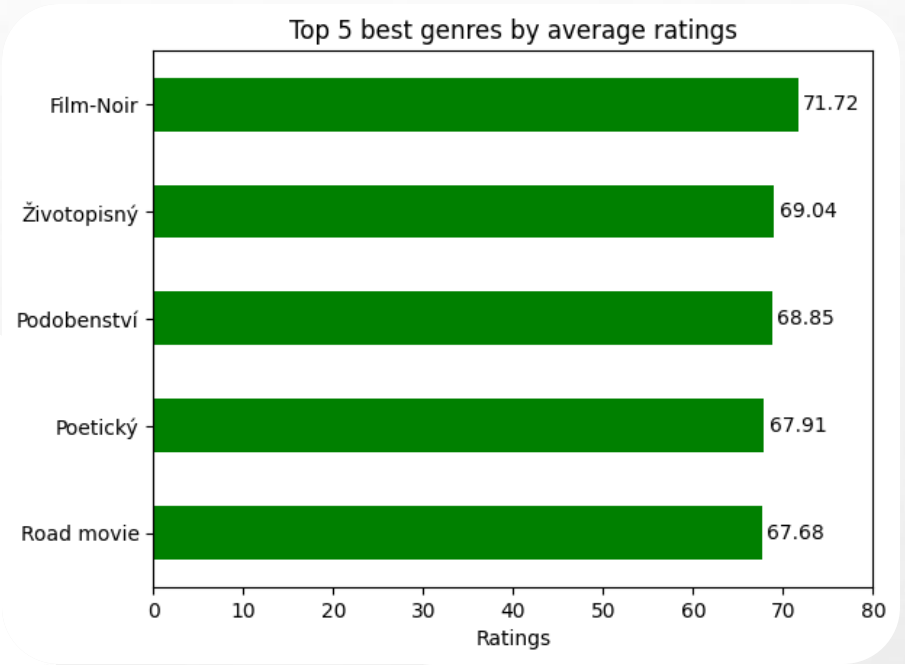
Period 1906 – 2023

200 features

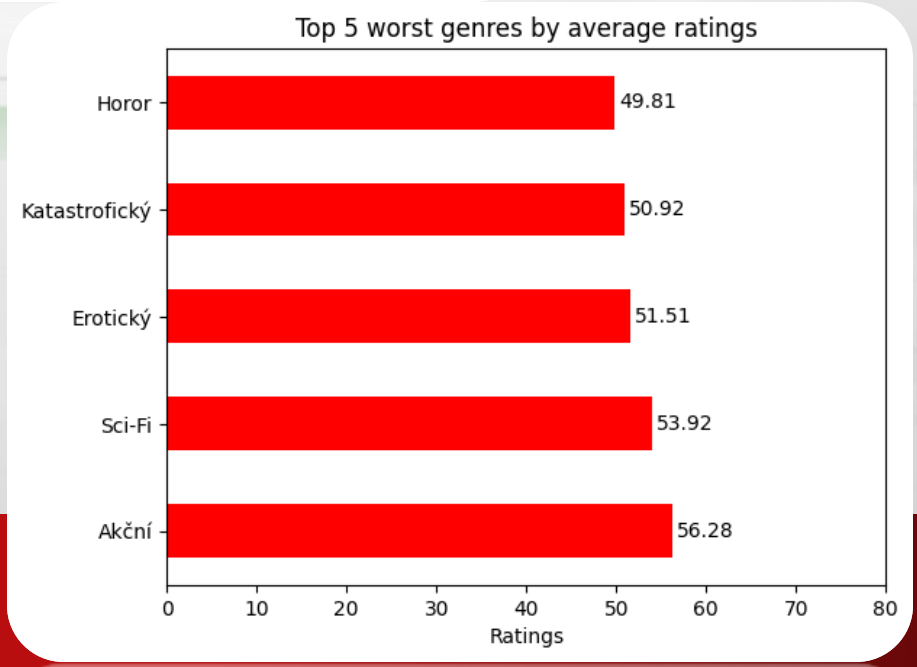
- 5 numerical features
- 195 categorical features (dummy variables)

GENRES

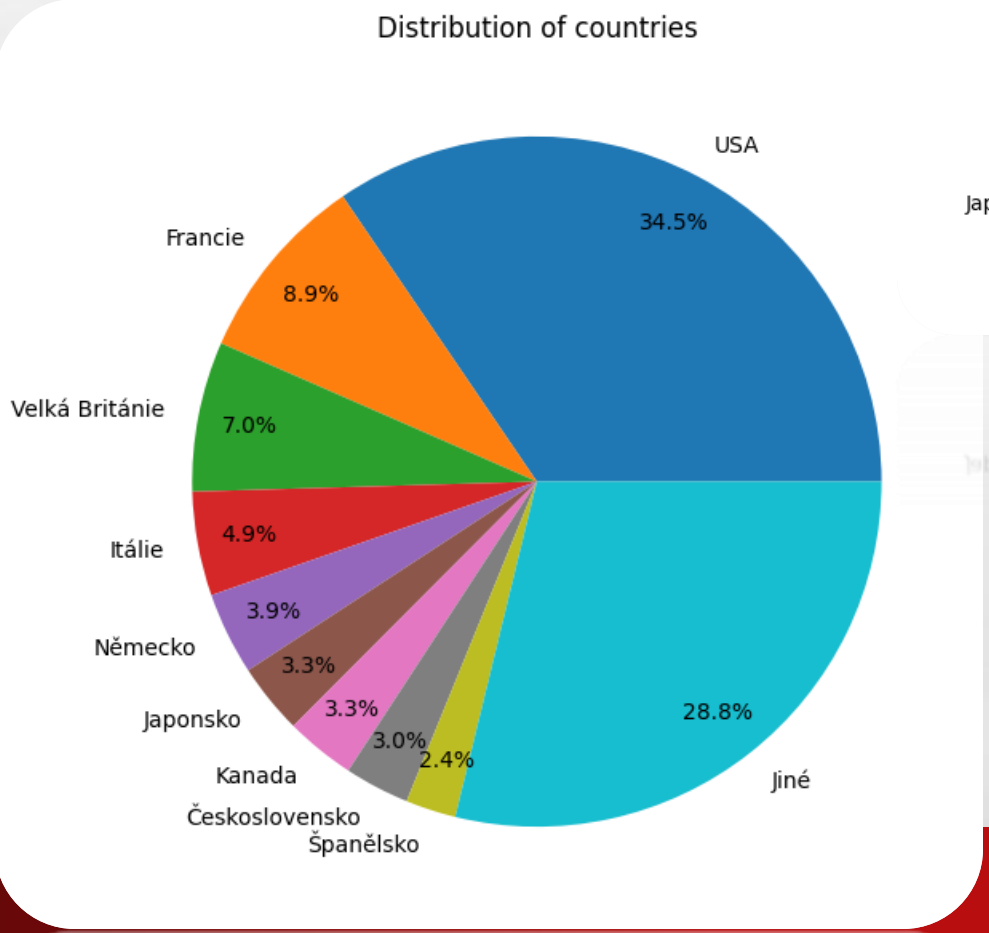


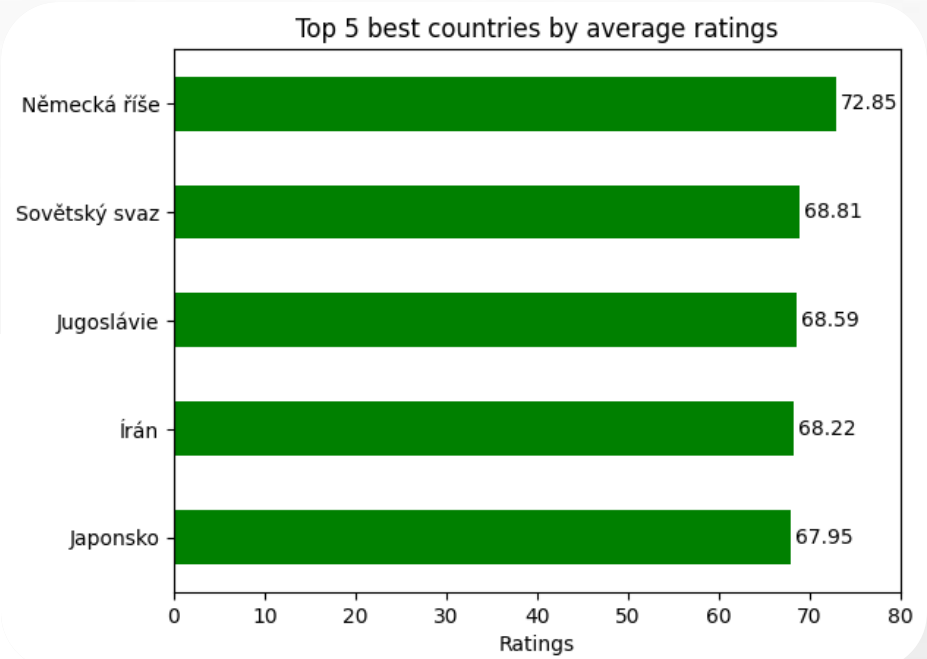


37 unique genres

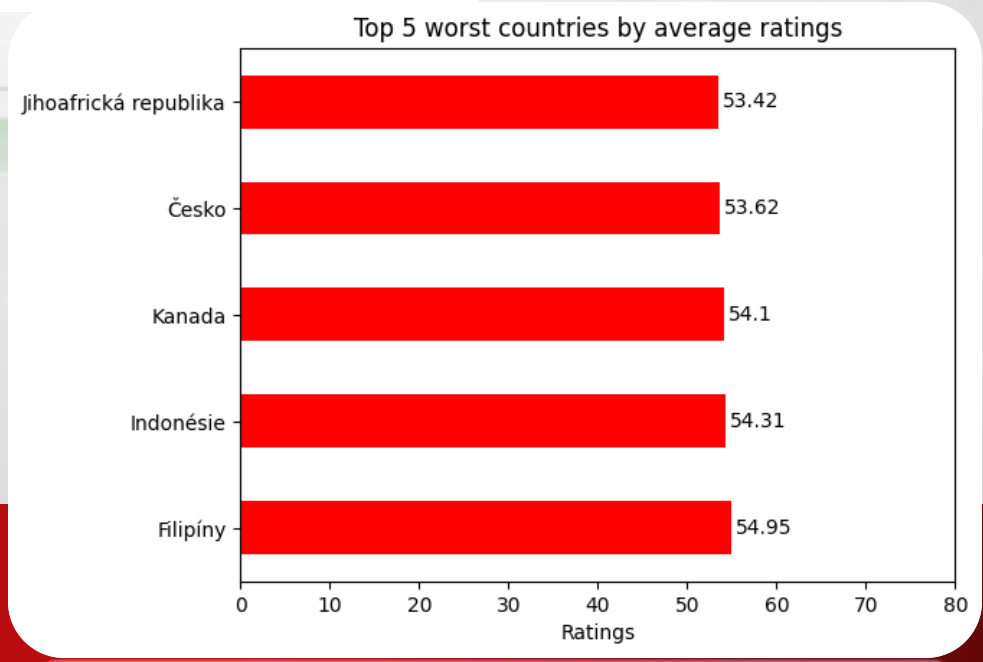


COUNTRIES

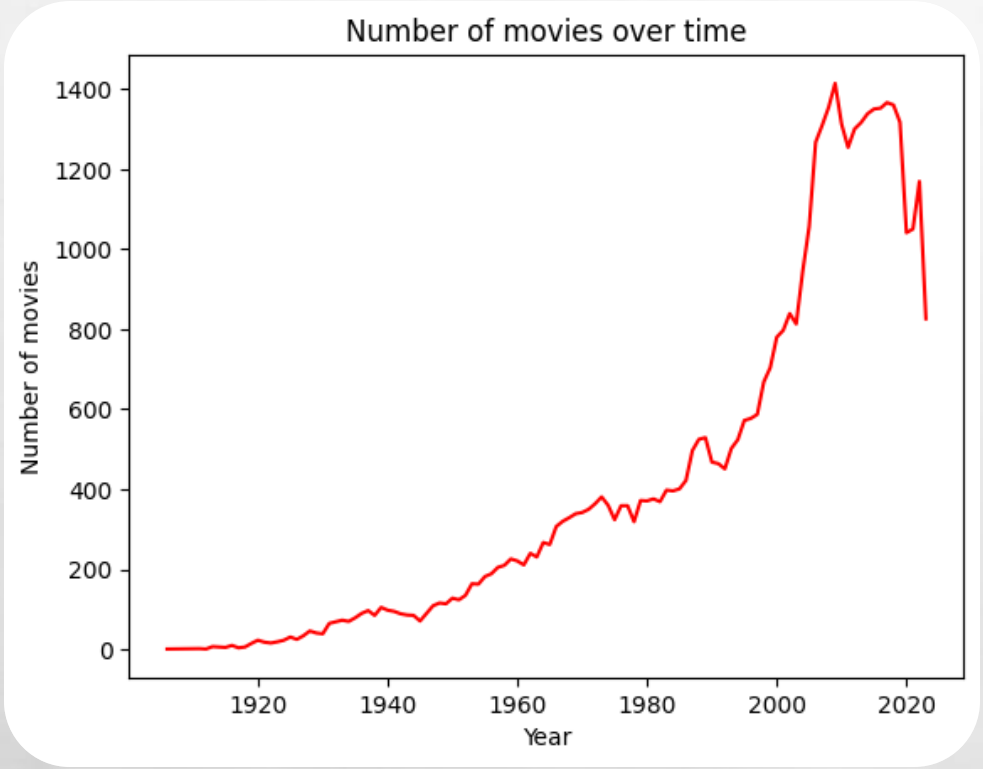


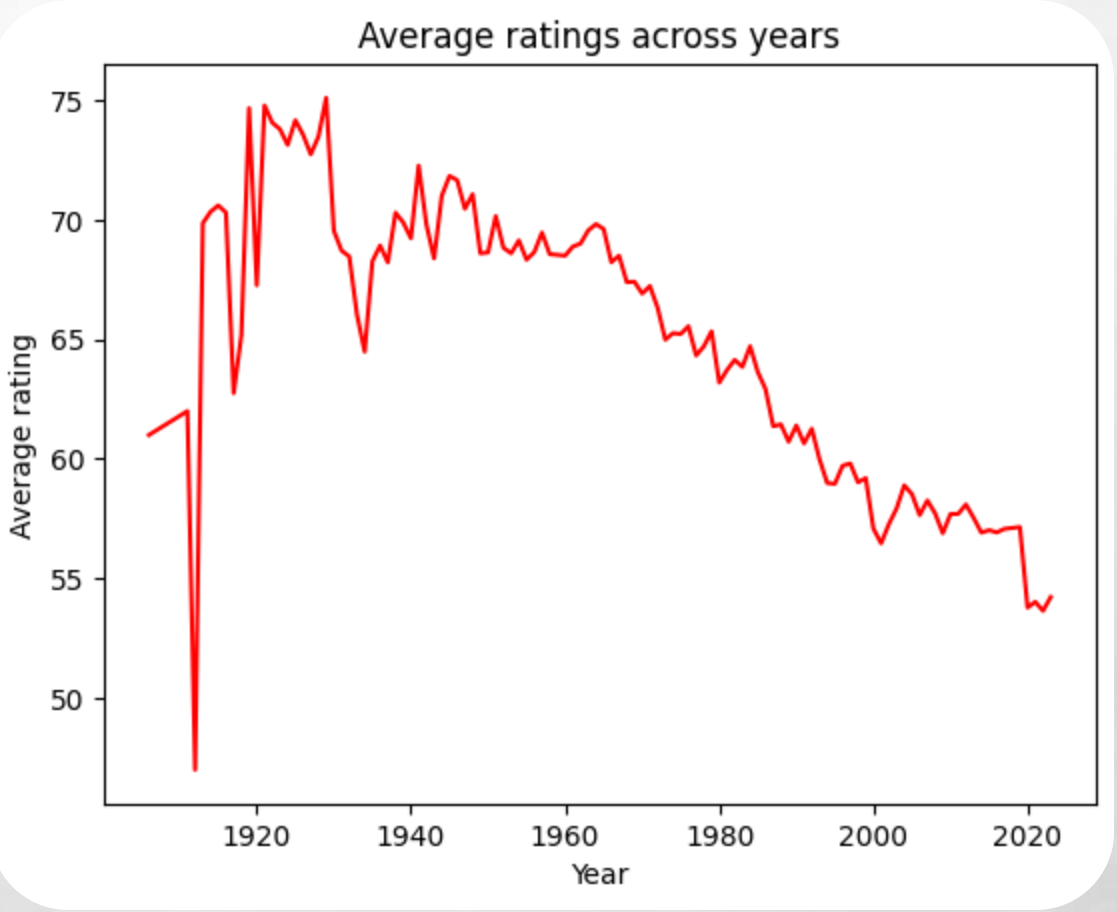


159 unique countries (including “unavailable”)

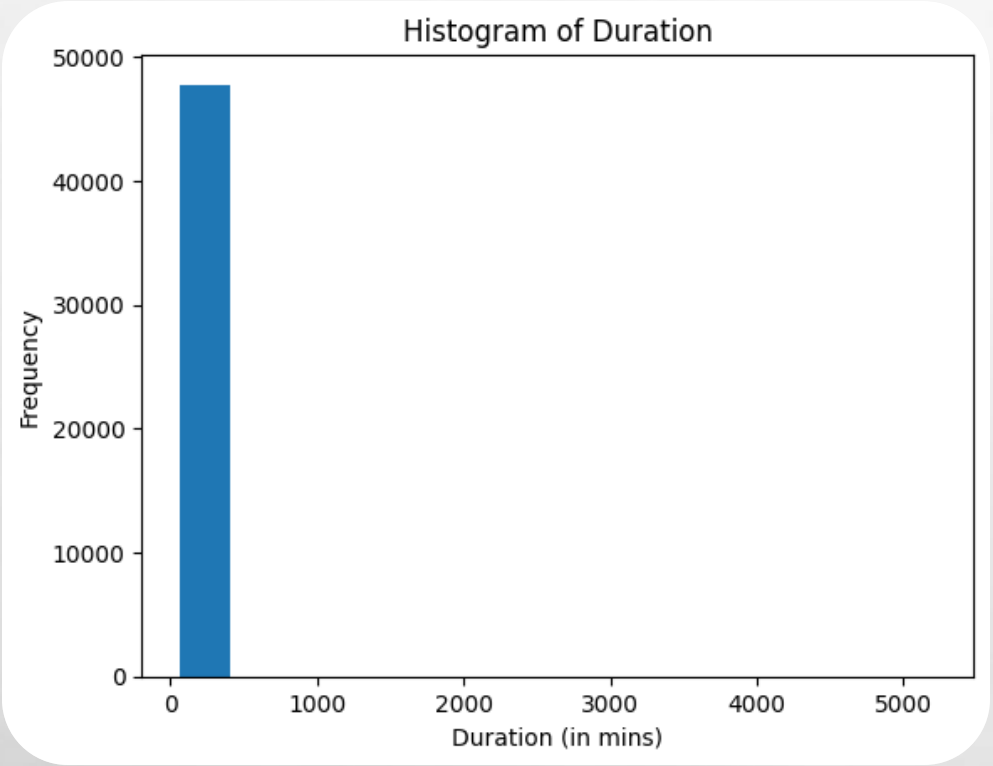


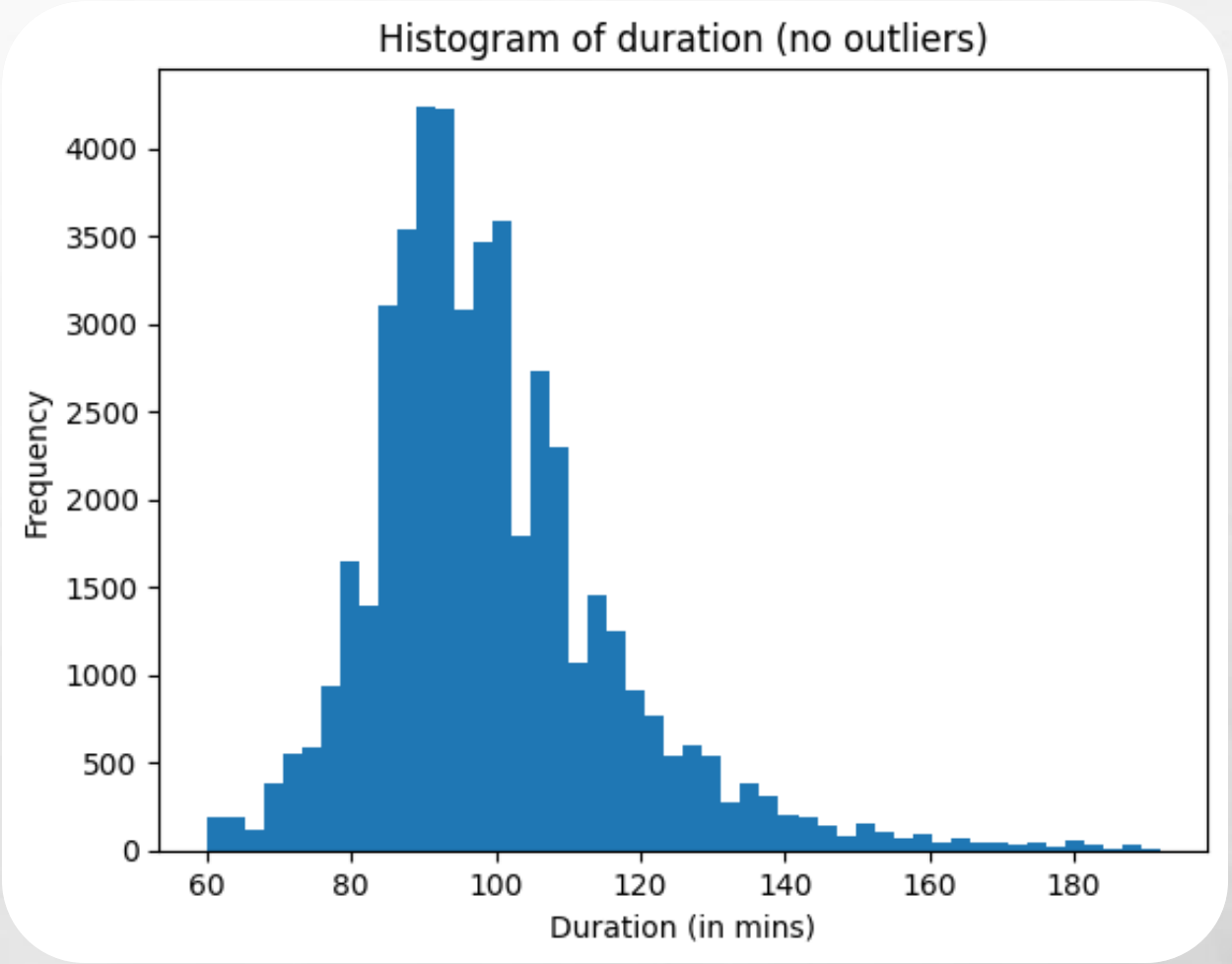
YEAR


Number of movies over time
Number of movies
0
200
400
600
800
1000
1200
1400
1920
1940
1960
1980
2000
2020
Year

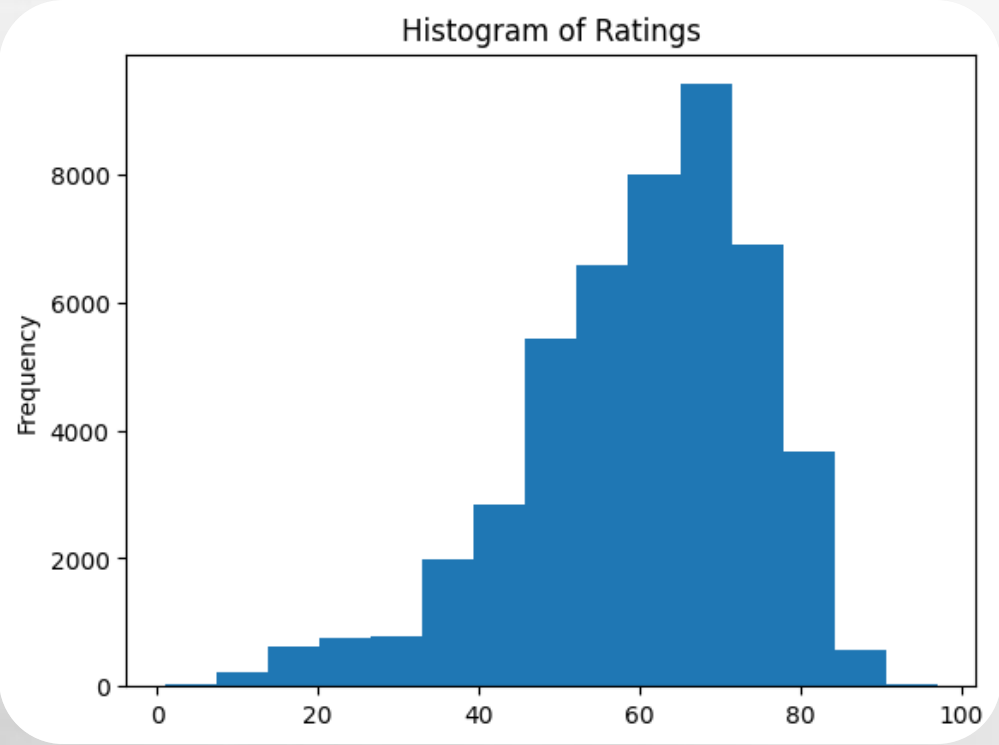

Average ratings across years
Average rating
50
55
60
65
70
75
1920
1940
1960
1980
2000
2020
Year

DURATION


Histogram of Duration
Frequency
Duration (in mins)
0
10000
20000
30000
40000
50000
0
1000
2000
3000
4000
5000


Histogram of duration (no outliers)
Frequency
Duration (in mins)
0
500
1000
1500
2000
2500
3000
3500
4000
60
80
100
120
140
160
180

RATINGS



Mean	60.10
Standard deviation	14.64
Minimum	1
25th percentile	52
Median	62
75th percentile	71
Maximum	97

CORRELATION ANALYSIS

Number of ratings	Number of fans	0.616
Ratings	Age	0.3
Ratings	Number of ratings	0.197
Ratings	Duration	0.189
Ratings	Number of fans	0.125
Duration	Number of ratings	0.1

CORRELATION ANALYSIS

Tajikistan	Uzbekistan	0.71
Czechia	Slovakia	0.36
Animated	Family movies	0.33
Somalia	Kenya	0.32
Ratings	Drama	0.30
Ratings	Horror	-0.28
Drama	Horror	-0.27
Thriller	Crime	0.26

LINEAR REGRESSION

- Regressions with several variable combinations were estimated to inspect the impact on the results
  - Only basic numerical variables (Duration, Number of directed movies, Age)
  - Regression with removed outliers
  - Additional "*ex post*" variables (Number of ratings, Number of fans)
  - Categorical variables
- The model performance is evaluated using Adjusted R-squared, MSE, MAE
- The significance of the estimated coefficients is tested using a t-test

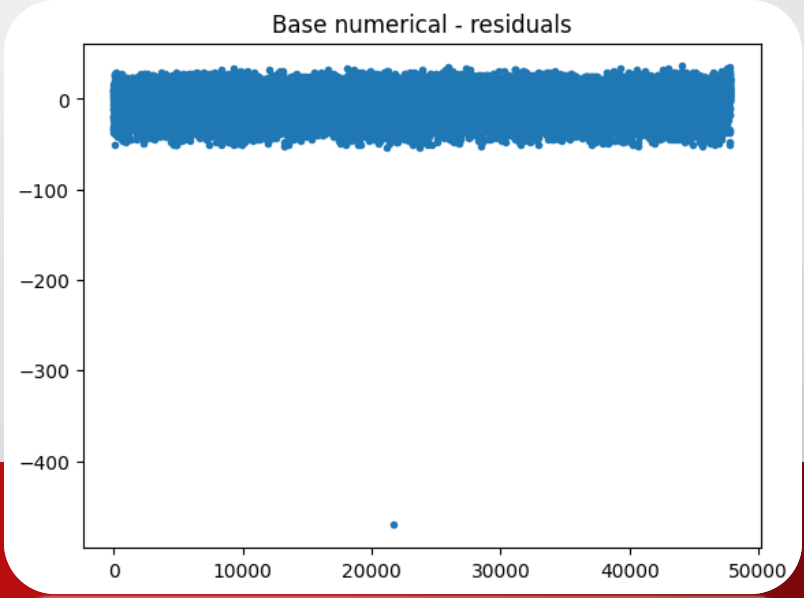
LINEAR REGRESSION

- **R-squared**
  - $R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \overline{y_i})^2}$
- **Adjusted R-squared**
  - $\bar{R}^2 = 1 - (1 - R^2)\frac{n-1}{n-p-1}$
- **t-test**
  - $t = \frac{\beta}{SE(\beta)}$
- **MSE**
  - $MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$
- **MAE**
  - $MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i|$

SIMPLE MODEL

Variable	Coef.	Std. Err.	t	p-value
Constant	45.08	0.227	198.19	0.0
Duration	0.095	0.002	46.95	0.0
No. of directed movies	0.035	0.011	3.12	0.002
Age	0.205	0.003	71.15	0.0

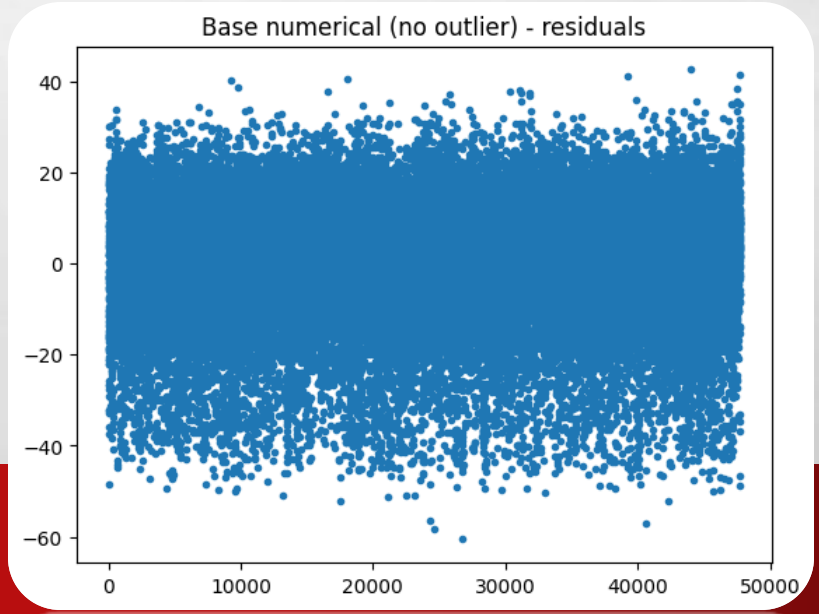
Adjusted R-squared	0.133
MSE	185.81
MAE	10.48



SIMPLE MODEL – NO OUTLIERS

Variable	Coef.	Std. Err.	t	p-value
Constant	26.62	0.365	73.0	0.0
Duration	0.28	0.004	79.63	0.0
No. of directed movies	-0.05	0.011	-4.73	0.0
Age	0.22	0.003	80.2	0.0

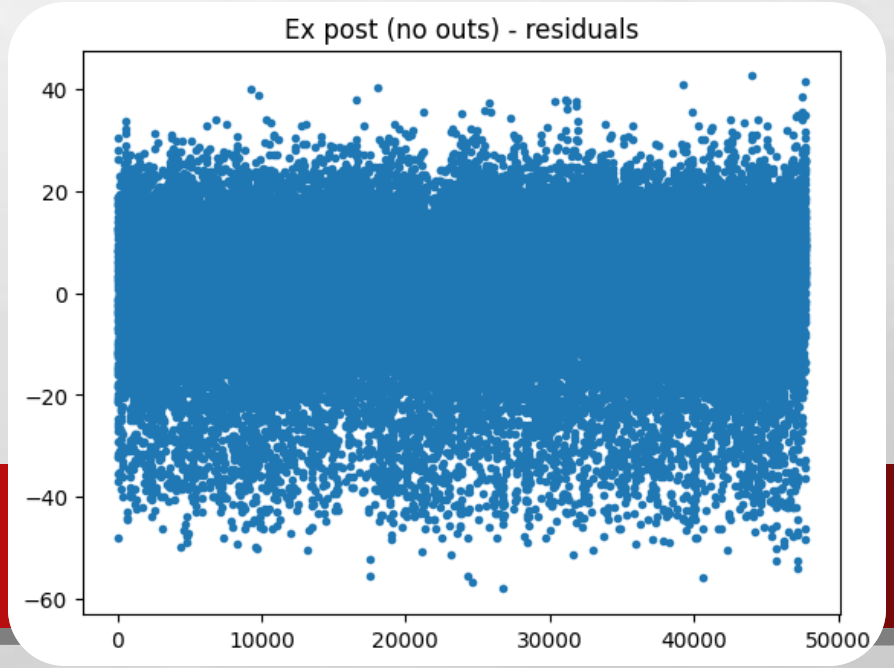
Adjusted R-squared	0.2
MSE	171.27
MAE	10.15



EX POST MODEL

Variable	Coef.	Std. Err.	t	p-value
Constant	28.09	0.361	77.84	0.0
Duration	0.26	0.003	73.36	0.0
No. of directed movies	-0.09	0.011	-8.13	0.0
Age	0.23	0.003	83.89	0.0
Number of ratings	0.0004	1.19e-05	33.93	0.0
Number of fans	-0.001	0.0	-2.94	0.003

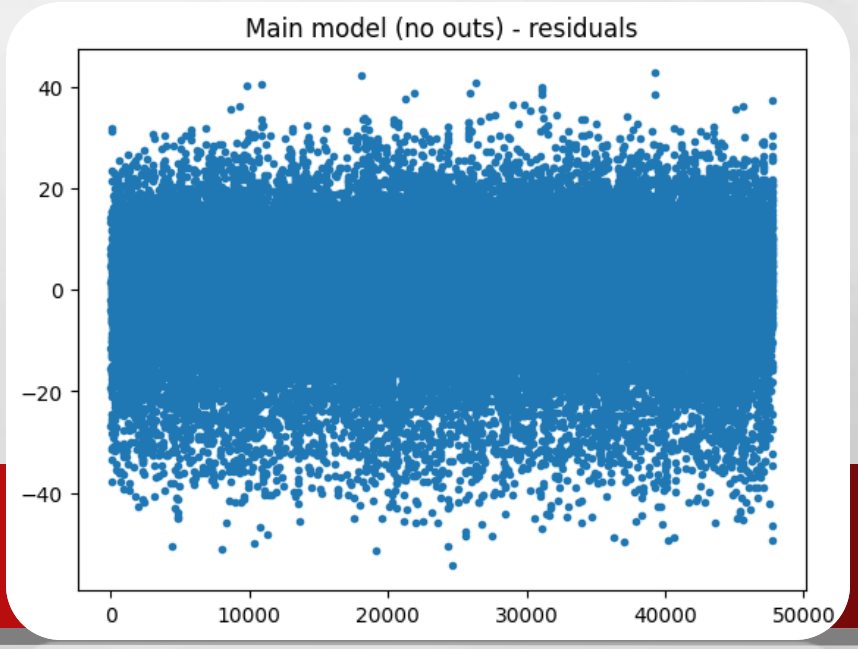
Adjusted R-squared	0.226
MSE	165.57
MAE	9.96



MAIN MODEL

Variable	Coef.	Std. Err.	t	p-value
Constant	29.57	0.39	75.62	0.0
Duration	0.21	0.004	59.9	0.0
No. of directed movies	0.02	0.01	1.5	0.135
Age	0.24	0.003	81.1	0.0

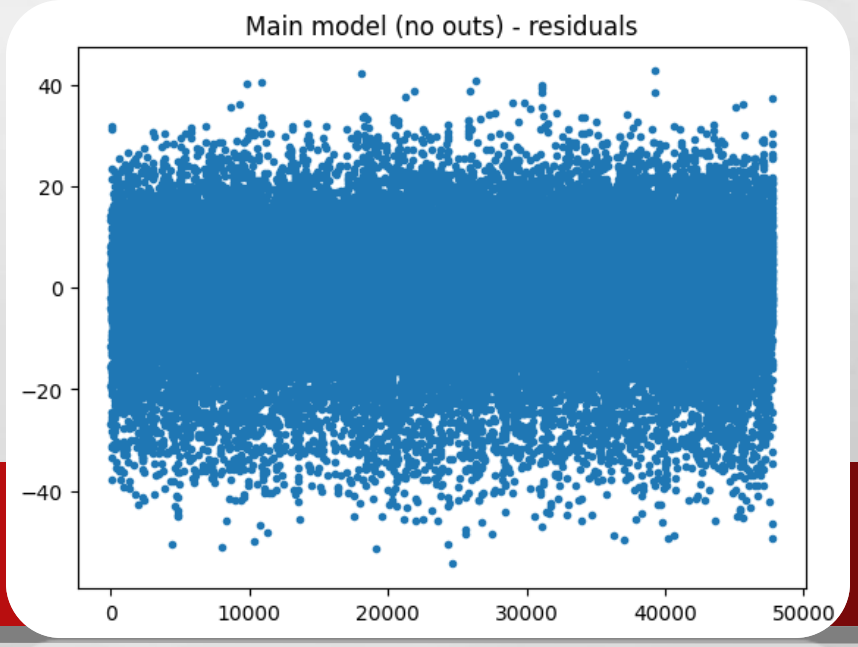
Adjusted R-squared	0.376
MSE	133.06
MAE	8.92



MAIN MODEL CONT.

Variable	Coef.	Std. Err.	t	p-value
Drama	5.42	0.13	41.15	0.0
Animated	12.55	0.3	41.62	0.0
Horror	-5.22	0.18	-28.82	0.0
USA	-3.25	0.15	-21.63	0.0
Czechoslovakia	-6.5	0.33	-19.93	0.0
Comedy	2.5	0.14	18.66	0.0
Action	-2.93	0.17	-17.05	0.0
Japan	4.69	0.29	16.11	0.0

Adjusted R-squared	0.376
MSE	133.06
MAE	8.92



COMPARISON

	Adjusted R-squared	MSE	MAE
Simple model	0.133	185.81	10.48
Simple model (no outs)	0.2	171.27	10.15
Ex post model	0.226	165.57	9.96
Main model	0.376	133.06	8.92
All variables (also ex post)	0.428	121.95	8.52

LANGUAGE MODEL

- CAN WE USE ADDITIONAL INFORMATION TO IMPROVE THE RESULTS?
- EACH MOVIE HAS A SHORT PREVIEW AVAILABLE
- DOES THE REVIEW CONTAIN INFORMATION THAT CAN HELP PREDICTING THE RATINGS?

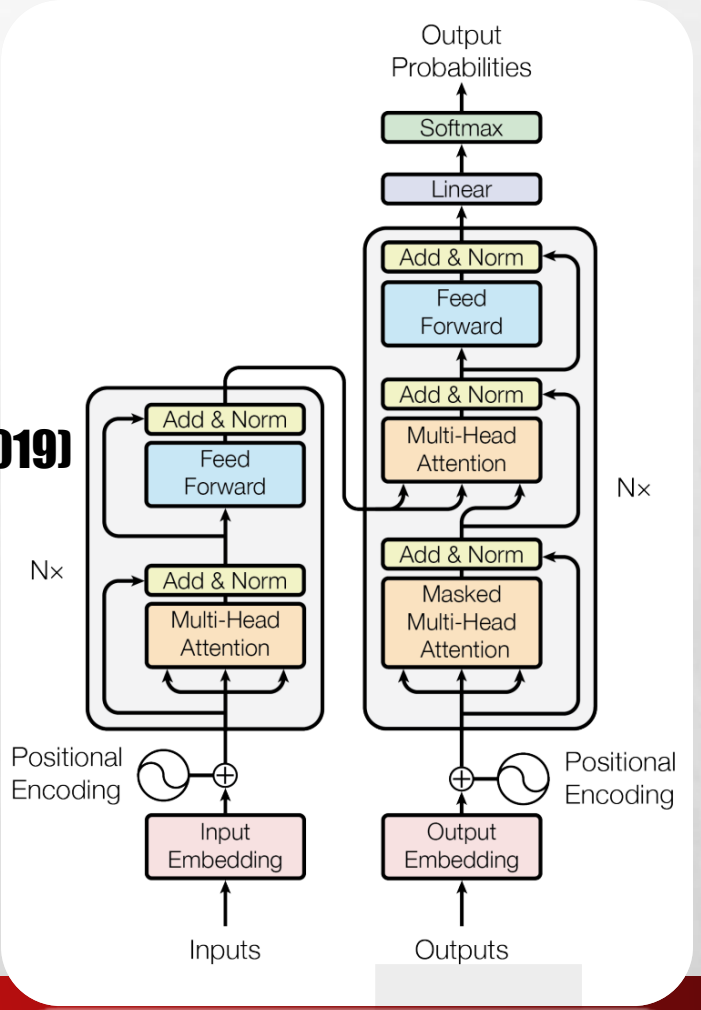
Mladý bankéř Andy Dufresne (Tim Robbins) je v roce 1947 odsouzen za vraždu své ženy a jejího milence. Přesto, že tento čin popírá, čeká na něj dvojnásobný doživotní trest v obávané věznici Shawshank. Andy se snaží přizpůsobit vězeňskému životu a po krušných začátcích se sb
odpykává svůj doživotní trest už dvacet let. Sílu, jak přežít zdejší p
naději na svobodu... *(Magic Box)*

Paul Edgecomb se vrací ve vzpomínkách do roku 1935, kdy byl zaměstnán v louisianské věznici jako hlavní dozorce. Tenkrát se tam setkal s výjimečným, byť duchem prostým mužem, který byl obdařen nejen velkým srdcem, ale také nadpozemskými schopnostmi. Byl to John Coffey, neprávem odsouzený na smrt za vraždu dvou malých holčiček. V té době trpěl Paul těžkým zánětem močového měchýře a také neměl šanci zbavit se
cyho. Jednoho dne chce s Paulem mluvit Coffey. Když se k
okovaného Paula pustí a on uvidí, jak černoch vypustil z úst
lestivý zánět zmizel. Coffey má zvláštní schopnost, díky které
noci převezou, dokonce oživí cvičenou myš jednoho vězně,
ho čeká smrt, přijímá svůj úděl odevzdaně a bez hořkosti. *(TV*

Dva policisté (Brad Pitt a Morgan Freeman) jsou na stopě geniálního vraha, zodpovědného za sérii děsivých vražd, jejichž oběti spojuje sedm smrtelných hříchů. V jedné z rolí tohoto kvalitního trhilleru, zasazeného do temného města nasáklého bolestí a zkázou, uvidíte také Gwyneth Paltrow. David Fincher (*Klub rváčů*, *Zodiac*, *Podivuhodný případ Benjamina Buttona*) s dokonalou znalostí našich nejhlubších obav pevně svírá otěže akce – fyzické, psychické i spirituální – neodvratně směřující k rozuzlení, které do hloubi otřese i tou nejzatvrzelejší duší. *(Magic Box)*

LANGUAGE MODEL

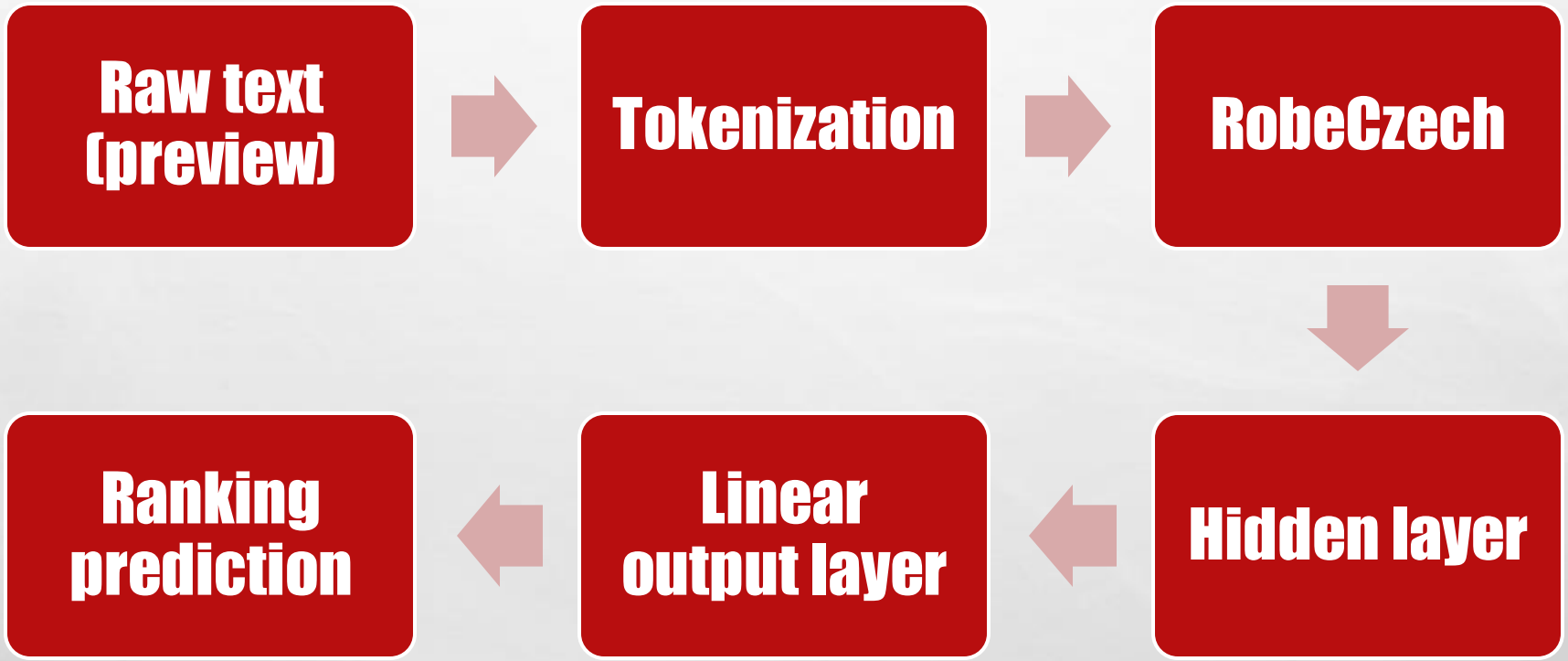
- Pretrained model: RobeCzech
  - Straka et al. (2021)
  - RoBERTa: A Robustly Optimized BERT Pretraining Approach (Liu et al., 2019)
  - Trained purely on Czech data
  - Based on Google's BERT model
    - Bidirectional Transformer architecture



LANGUAGE MODEL

- The model has to be adjusted for our specific purpose
- An additional fully connected layer is added on top of the pretrained model
- The output layer is linear with a single node
- During training, the pretrained model is initially frozen and only the new layers are trained
- After that, additional fine-tuning of the whole model with a small learning rate

LANGUAGE MODEL

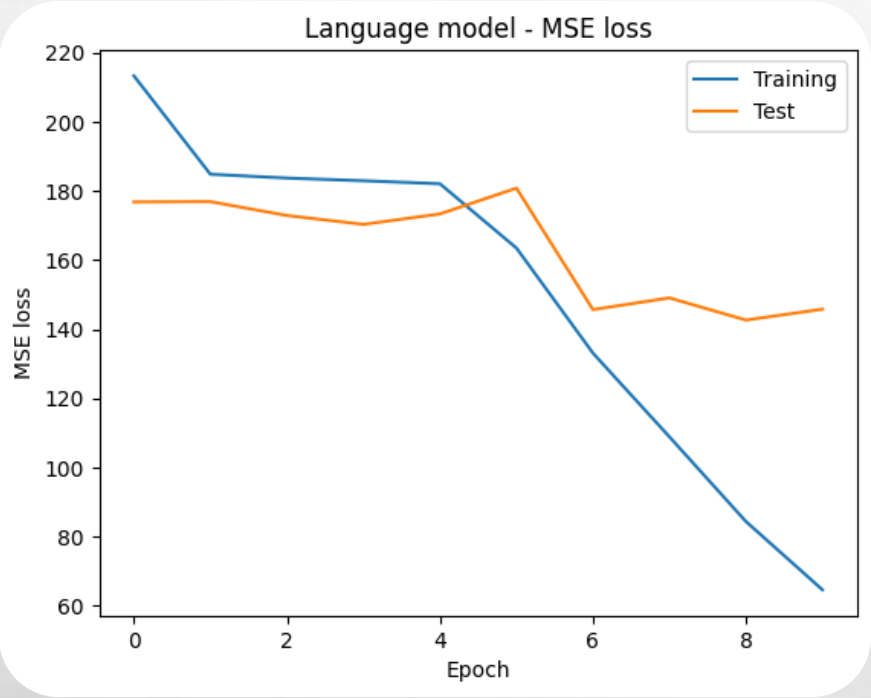


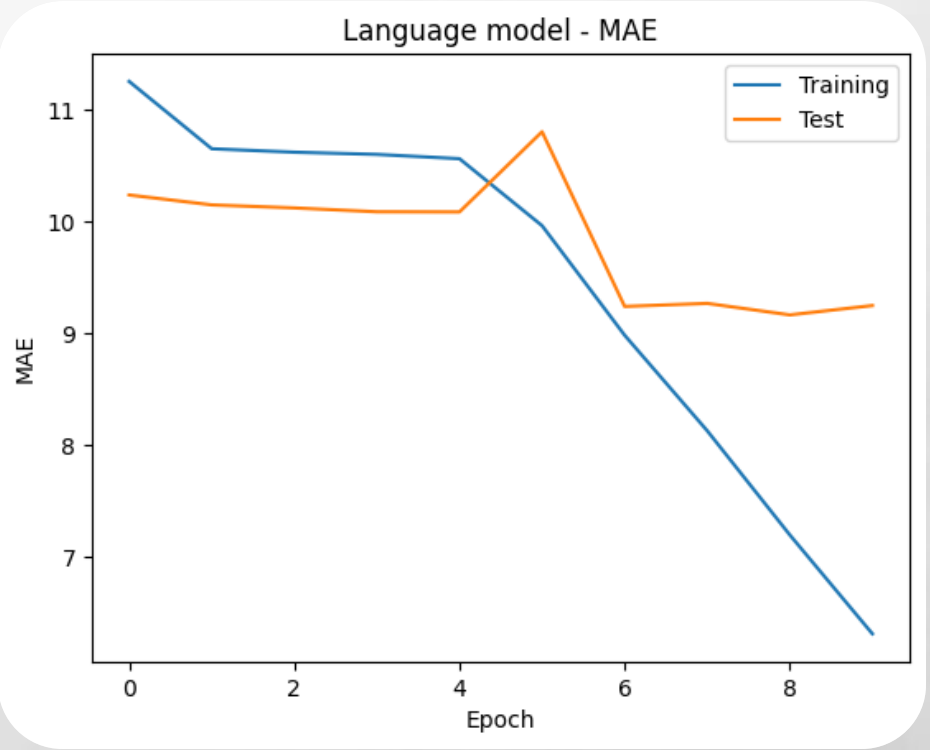
LANGUAGE MODEL

- ReLU activation function (hidden layer)
- MSE loss
- Adam optimizer
  - Learning rate
    - First 5 epochs: 0.001
    - Remaining 5 epochs (fine-tuning): 1e-5
- Batch size: 5
- Weights initialized using Xavier uniform distribution
- MAE evaluation metric
- 80% training, 20% test

$$W_{ij} \sim U\left[-\frac{\sqrt{6}}{\sqrt{fan_{in}+fan_{out}}}, \frac{\sqrt{6}}{\sqrt{fan_{in}+fan_{out}}}\right]$$

LANGUAGE MODEL





LANGUAGE MODEL – LINEAR REGRESSION

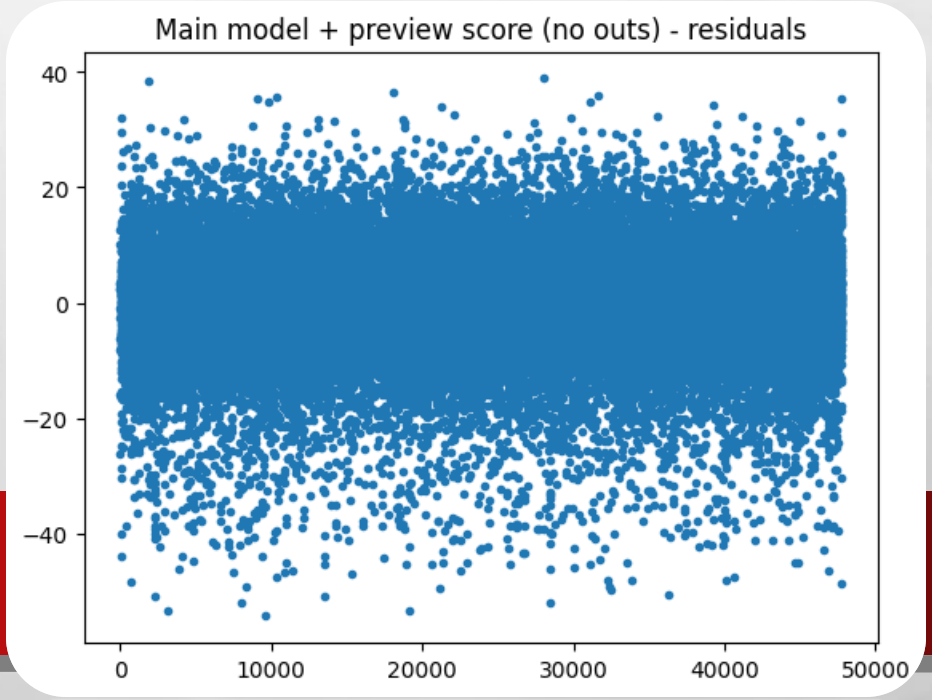
- Best MAE on test set: 9.16
  - Almost as good as linear regression which was evaluated on the train set!
  - Based purely on the preview

- Can we use the predictions of the language model in the linear regression to simulate “preview quality”?

LANGUAGE MODEL – LINEAR REGRESSION

Variable	Coef.	Std. Err.	t	p-value
Constant	-0.47	0.34	-1.37	0.17
Duration	0.12	0.003	42.76	0.0
No. of directed movies	-0.01	0.008	-1.27	0.2
Age	0.13	0.002	56.67	0.0
Preview score	0.72	0.004	181.85	0.0

Adjusted R-squared	0.632
MSE	78.38
MAE	6.52



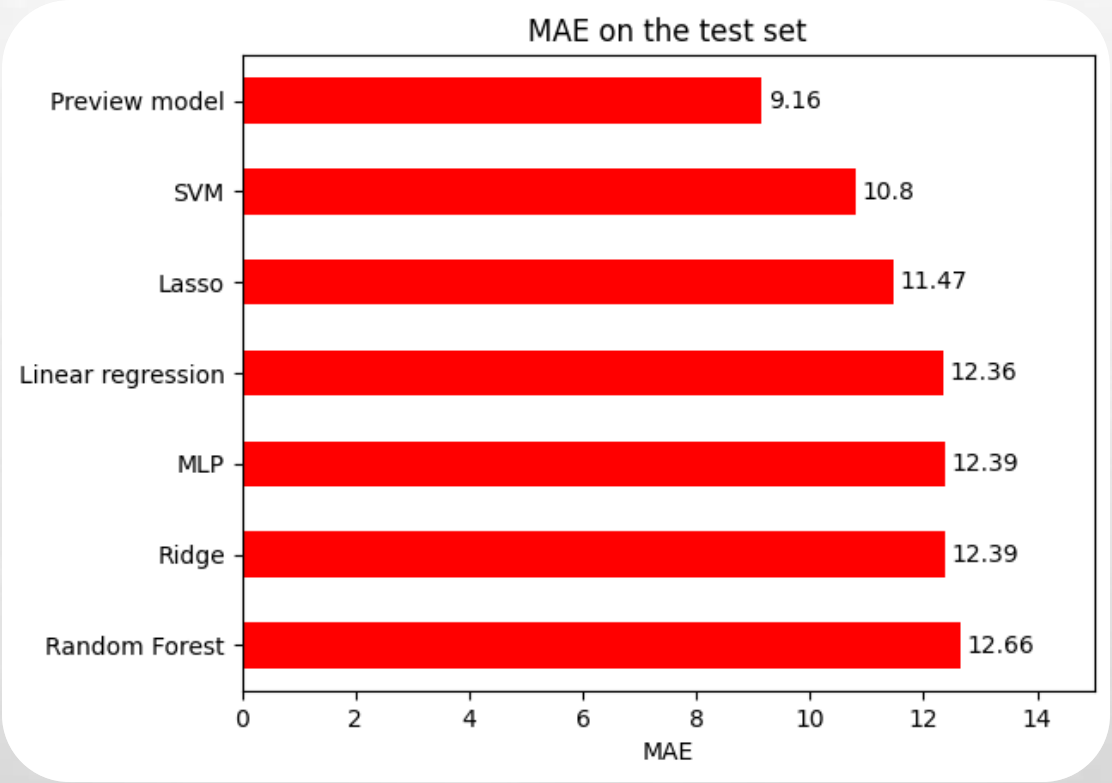
COMPARISON

	Adjusted R-squared	MSE	MAE
Simple model	0.133	185.81	10.48
Simple model (no outs)	0.2	171.27	10.15
Ex post model	0.226	165.57	9.96
Main model	0.376	133.06	8.92
All variables (also ex post)	0.428	121.95	8.52
Main model + preview score	0.632	78.38	6.52

PREDICTIVE ANALYSIS

- Comparison of performance of classifiers on the test set
- Evaluation metric: MAE
- No “ex post” variables considered
- Classifiers: Linear Regression, SVM, Random Forest, MLP

PREDICTIVE ANALYSIS



THANK YOU FOR YOUR ATTENTION

REFERENCES

- Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach. ArXiv, abs/1907.11692.
- Straka, M., Náplava, J., Straková, J., Samuel, D. (2021). RobeCzech: Czech RoBERTa, a Monolingual Contextualized Language Representation Model. In: Ekštein, K., Pártl, F., Konopík, M. (Eds) Text, Speech, and Dialogue. Tsd 2021. Lecture Notes in Computer Science(), vol 12848. Springer, cham. Https://doi.Org/10.1007/978-3-030-83527-9_17