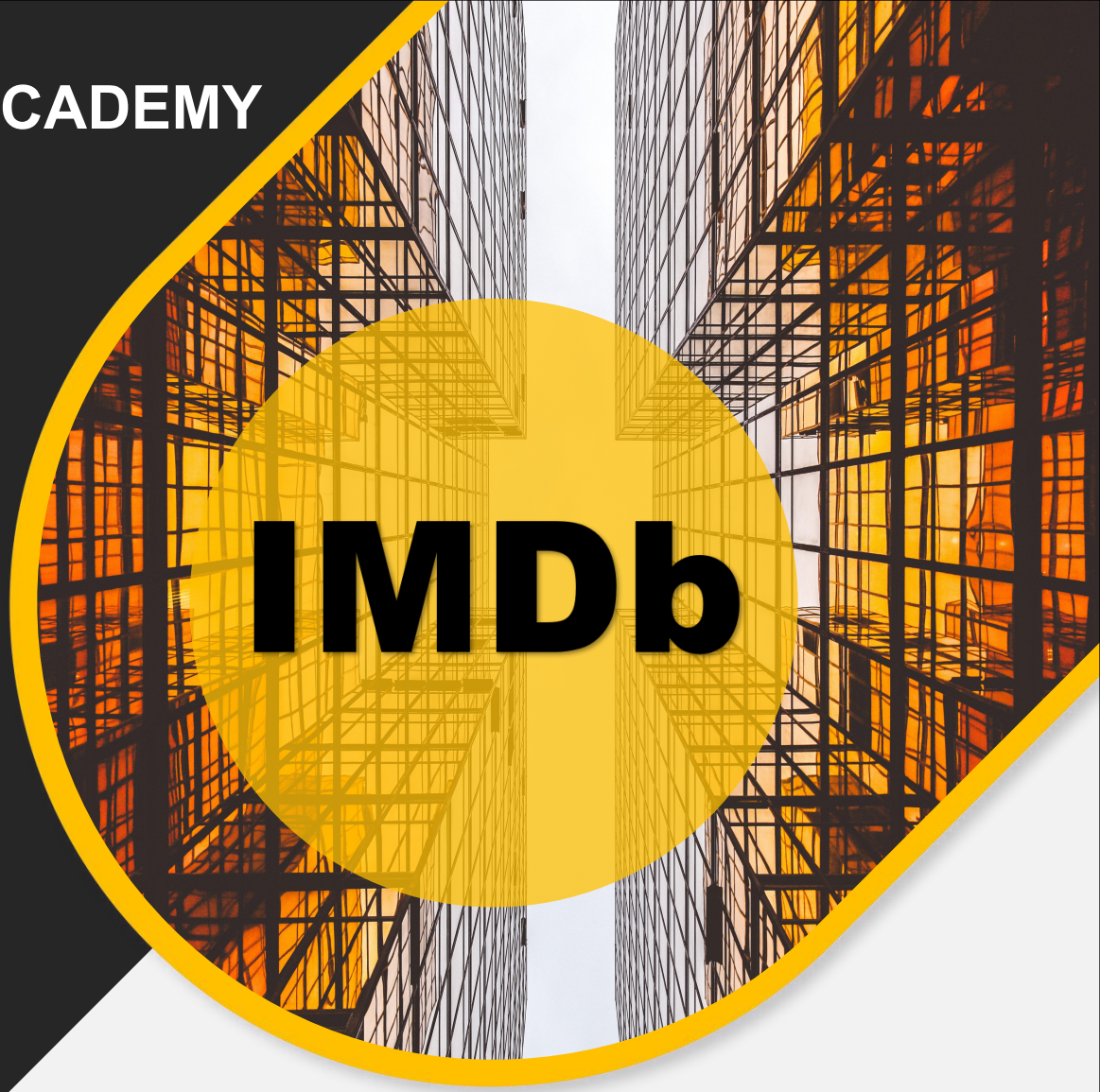
ISTANBUL DATA SCİENCE ACADEMY

Predicting Highest-Grossing Movies of Using Regression

Kübra Erensoy



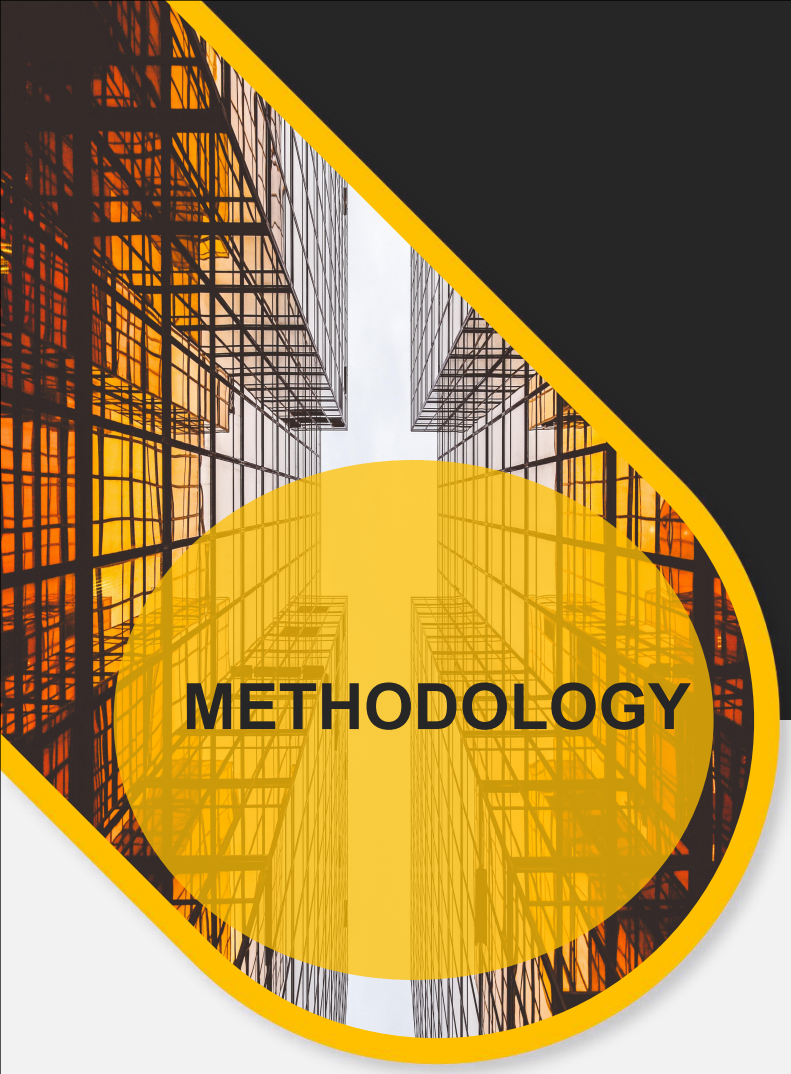
02 INTRODUCTION

Our Goal:

- predict the movie gross with machine learning models

Dataset:

- gathering with web scraping from IMDB website
- totally 1000 movies and 13 features



METHODOLOGY

DATA	TOOLS	EDA	REGRESSION
• Top 1000 Highest-Grossing Movies of All Time	• Python • Matplotlib • Pandas • Numpy • Seaborn • BeautifulSoup • Statsmodels • Scikit-learn	• Data cleaning • Data analysis and Visulizaions	• Data preparation and regresion

DATA PREPARATION

Data Cleaning

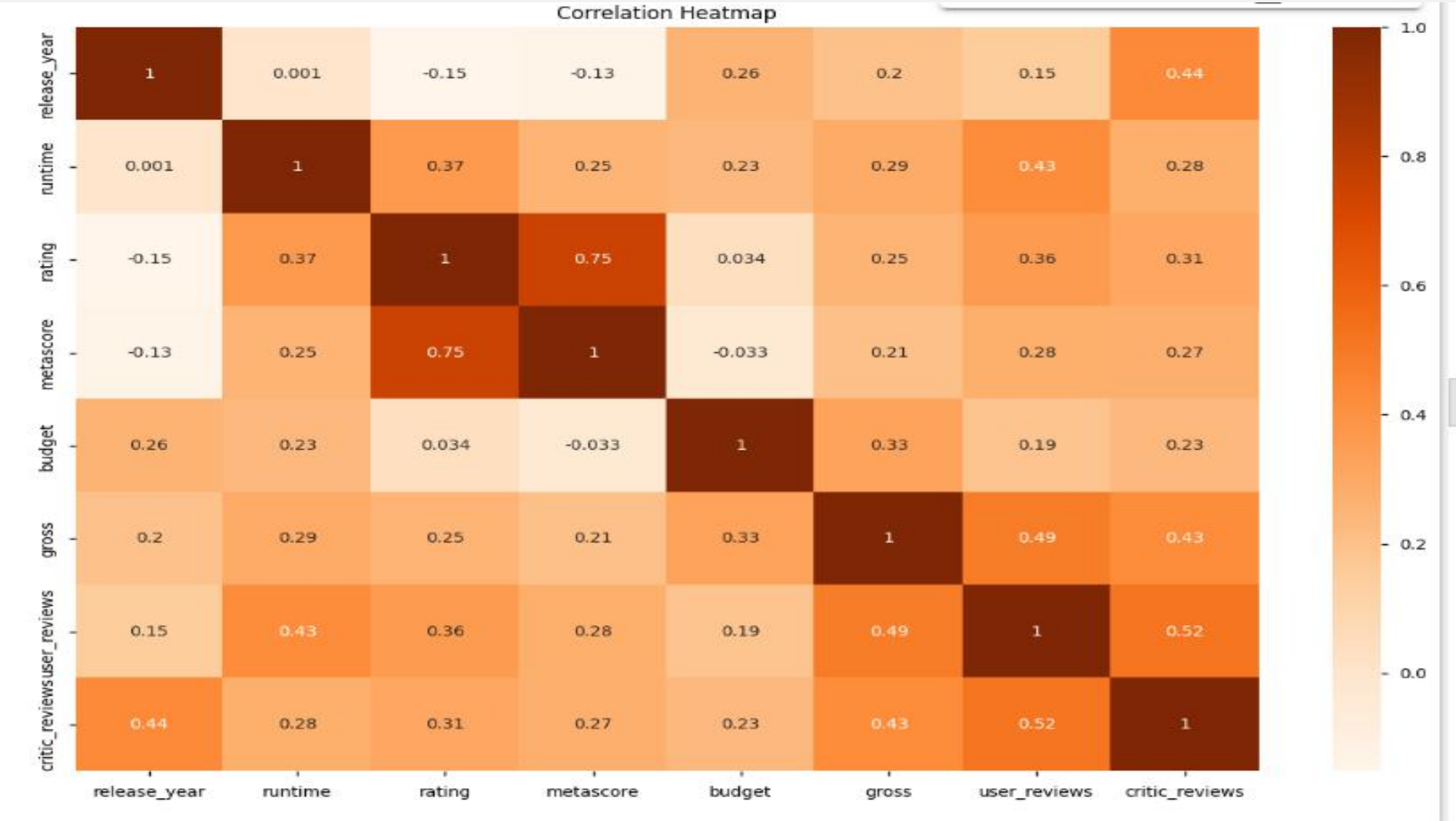
```
df.head(3)
```

	movie_name	release_year	runtime	director	genre1	genre2	genre3	rating	user_reviews	critic_reviews	metascore	budget	gross
0	Avatar	2009	2 hours 42 minutes	James Cameron	Action	Adventure	Fantasy	7.80	3.5K	479	83.00	$237,000,000 (estimated)	$2,847,379,794
1	Avengers: Endgame	2019	3 hours 1 minute	Anthony Russo	Action	Adventure	Drama	8.40	9.4K	589	78.00	$356,000,000 (estimated)	$2,797,501,328
2	Titanic	1997	3 hours 14 minutes	James Cameron	Drama	Romance	-	7.90	3.1K	331	75.00	$200,000,000 (estimated)	$2,201,647,264

	movie_name	release_year	runtime	director	genre1	genre2	genre3	rating	metascore	budget	gross	user_reviews	critic_reviews
0	Avatar	2009	162.00	James Cameron	Action	Adventure	Fantasy	7.80	83.00	237000000.00	2847379794.00	3500.00	479.00
1	Avengers: Endgame	2019	181.00	Anthony Russo	Action	Adventure	Drama	8.40	78.00	356000000.00	2797501328.00	9400.00	589.00
2	Titanic	1997	194.00	James Cameron	Drama	Romance	-	7.90	75.00	200000000.00	2201647264.00	3100.00	331.00

- Fill NAN values : Runtime,Metascore,Budget

DATA PREPARATION



FEATURE ENGINEERING

- Convert features to Dummy variables : “release_year”,”genre”
- Log Transformation to “gross”

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 995 entries, 0 to 994
Data columns (total 13 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   movie_name      995 non-null    object
 1   release_year    995 non-null    int64
 2   runtime         995 non-null    float64
 3   director        995 non-null    object
 4   genre1          995 non-null    object
 5   genre2          995 non-null    object
 6   genre3          995 non-null    object
 7   rating          995 non-null    float64
 8   metascore       995 non-null    float64
 9   budget          995 non-null    float64
 10  gross           995 non-null    float64
 11  user_reviews    995 non-null    float64
 12  critic_reviews  995 non-null    float64
dtypes: float64(7), int64(1), object(5)
memory usage: 101.2+ KB
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 995 entries, 0 to 994
Data columns (total 19 columns):
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
 0   release_year        995 non-null    int64
 1   user_reviews        995 non-null    float64
 2   critic_reviews      995 non-null    float64
 3   ratings             995 non-null    float64
 4   metascores          995 non-null    float64
 5   runtimess           995 non-null    float64
 6   budgets             995 non-null    float64
 7   genre_Action        995 non-null    uint8
 8   genre_Adventure     995 non-null    uint8
 9   genre_Animation     995 non-null    uint8
 10  genre_Biography     995 non-null    uint8
 11  genre_Comedy        995 non-null    uint8
 12  genre_Crime         995 non-null    uint8
 13  genre_Documentary   995 non-null    uint8
 14  genre_Drama         995 non-null    uint8
 15  genre_Fantasy       995 non-null    uint8
 16  genre_Horror        995 non-null    uint8
 17  genre_Mystery       995 non-null    uint8
 18  log_gross           995 non-null    float64
dtypes: float64(7), int64(1), uint8(11)
memory usage: 73.0 KB
```

07

Results

Machine Learning Models

SPLITTING MODELS
- 60% TRAIN
- 20% VALIDATION
- 20% TEST

Linear Regression

R^2 Score Value (Test): 0.4305748592173465

R^2 Score Value (Validation): 0.5326125444019985

MSE Value:: 0.1269308831998663

Statsmodel

R^2 Score: 0.45

MSE Value:0.2896224395715880

Ridge Regression

R^2 Score Value (Test): 0.421

R^2 Score Value (Validation): 0.4473

Degree 2 Polynomial Regression

R^2 Score Value (Test): 0.107

R^2 Score Value (Validation): 0.357

08 CONCLUSION

Linear Regression algorithm has the best results.

Polynomial Regression, Ridge Regression, Statsmodel has similar scores

FUTURE WORK

- Add more features: stars,company,director
- Add more movies

Thank you for listening to me :)