In a World…

Predicting the Gross Receipts for Movies During the Shutdown

Albert Lee

GREYHOUND

TENET

MULAN

Let’s Go to the Movies

- The Movie Database
- Box Office Mojo
- Internet Movie Database

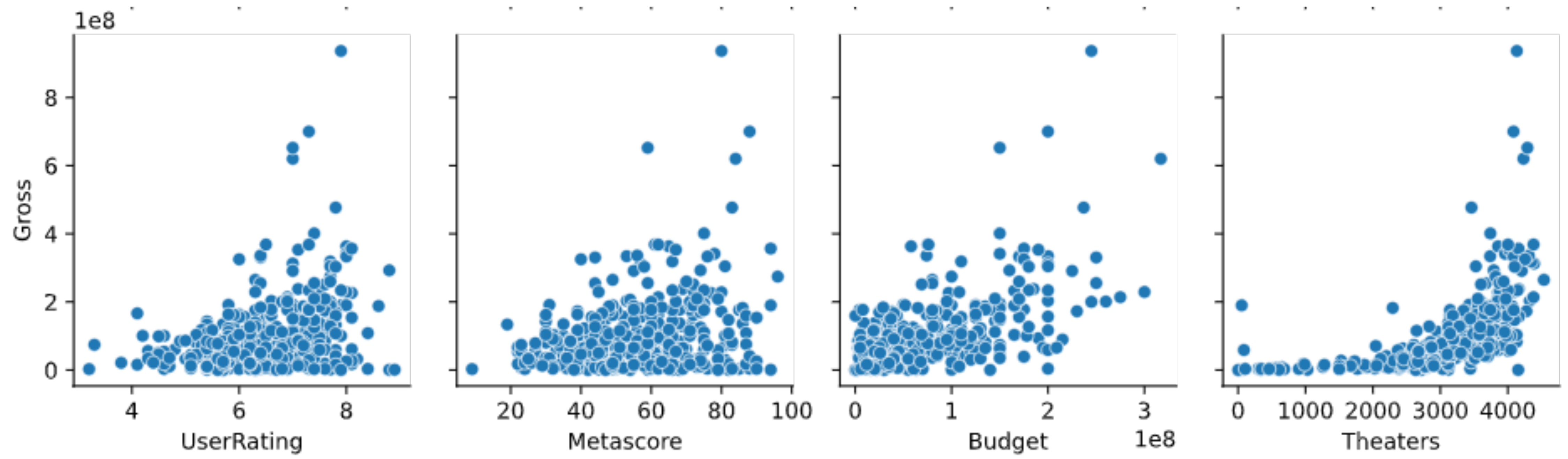
TMDB

Box Office Mojo
by IMDbPro

IMDb

Features

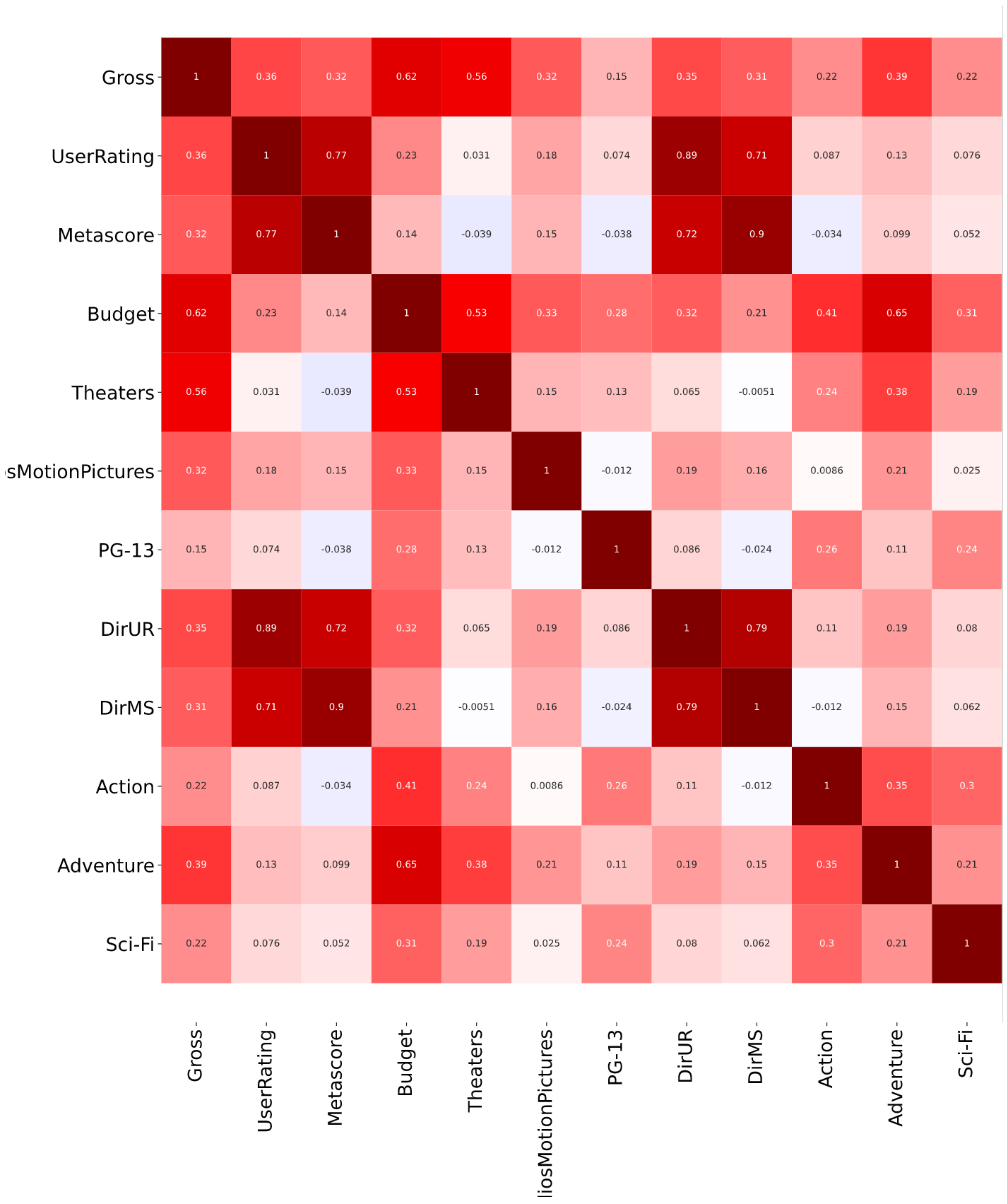
Shaping the model

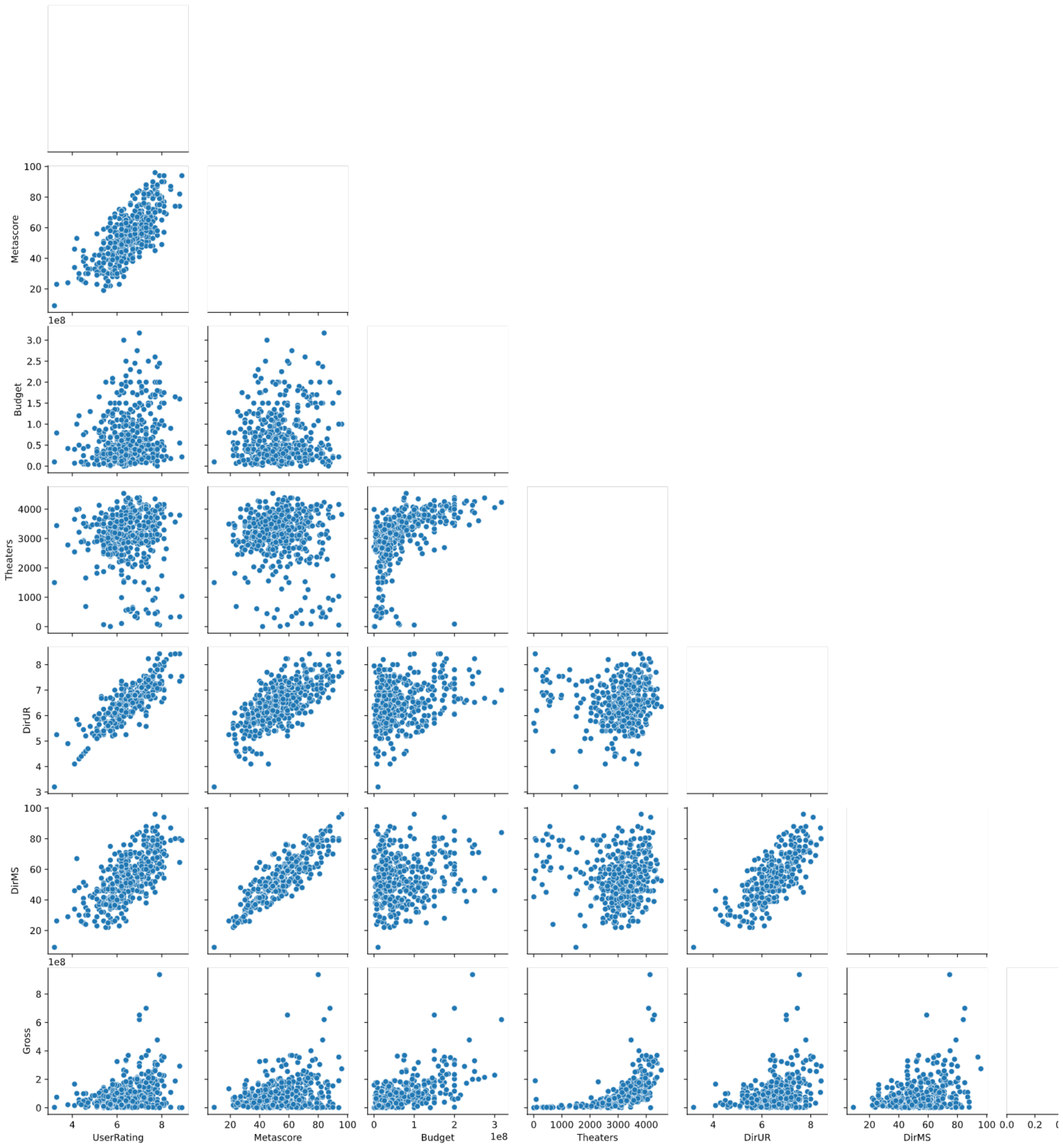


- Number of theaters vs gross shows high correlation.

- Separate month of release, MPAA rating, category.

Gross	Theaters	Budget	User Rating
Metascore	Disney	Universal	Fox
Sony	Paramount	Warner	PG
PG-13	R	Jan	Feb
Mar	Apr	May	Jun
Jul	Aug	Sep	Oct
Nov	Dec	Dir UR	Dir MS
Action	Adventure	Sci-fi	Animation
Comedy	Thriller	Drama	Music
Romance	Fantasy	Biography	Horror
Crime	Sport	Mystery	Theaters ^ 2

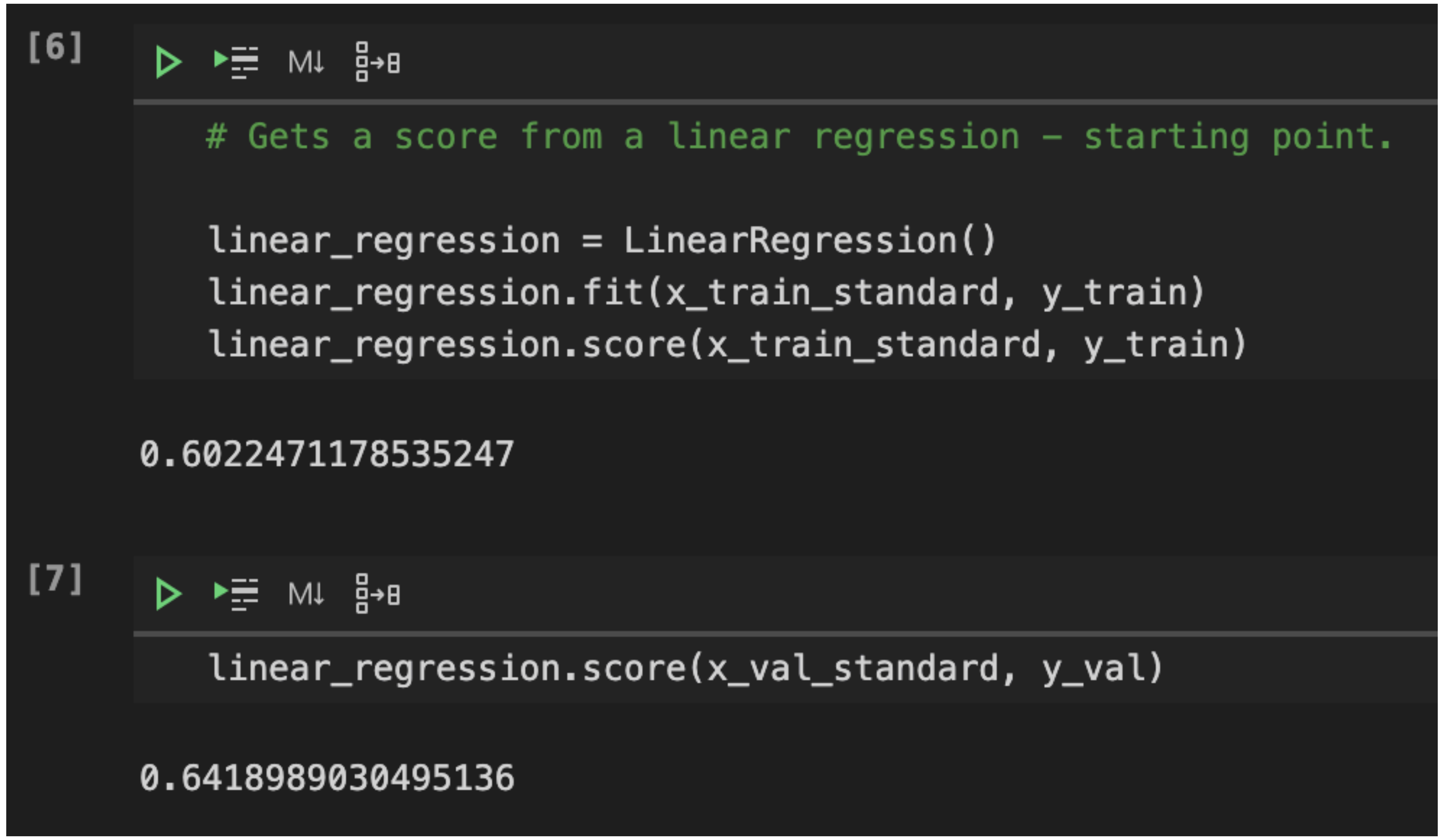

Gross
UserRating
Metascore
Budget
Theaters
sMotionPictures
PG-13
DirUR
DirMS
Action
Adventure
Sci-Fi
1 0.36 0.32 0.62 0.56 0.32 0.15 0.35 0.31 0.22 0.39 0.22
0.36 1 0.77 0.23 0.031 0.18 0.074 0.89 0.71 0.087 0.13 0.076
0.32 0.77 1 0.14 -0.039 0.15 -0.038 0.72 0.9 -0.034 0.099 0.052
0.62 0.23 0.14 1 0.53 0.33 0.28 0.32 0.21 0.41 0.65 0.31
0.56 0.031 -0.039 0.53 1 0.15 0.13 0.065 -0.0051 0.24 0.38 0.19
0.32 0.18 0.15 0.33 0.15 1 -0.012 0.19 0.16 0.0086 0.21 0.025
0.15 0.074 -0.038 0.28 0.13 -0.012 1 0.086 -0.024 0.26 0.11 0.24
0.35 0.89 0.72 0.32 0.065 0.19 0.086 1 0.79 0.11 0.19 0.08
0.31 0.71 0.9 0.21 -0.0051 0.16 -0.024 0.79 1 -0.012 0.15 0.062
0.22 0.087 -0.034 0.41 0.24 0.0086 0.26 0.11 -0.012 1 0.35 0.3
0.39 0.13 0.099 0.65 0.38 0.21 0.11 0.19 0.15 0.35 1 0.21
0.22 0.076 0.052 0.31 0.19 0.025 0.24 0.08 0.062 0.3 0.21 1
Gross
UserRating
Metascore
Budget
Theaters
liosMotionPictures
PG-13
DirUR
DirMS
Action
Adventure
Sci-Fi


Metascore
Budget
Theaters
DirUR
DirMS
Gross
1e8
1e8
1e8
UserRating
Metascore
Budget
Theaters
DirUR
DirMS
20 40 60 80 100
0.0 0.5 1.0 1.5 2.0 2.5 3.0
0 1000 2000 3000 4000
3 4 5 6 7 8
0 2 4 6 8
4 6 8
0 1 2 3
0.0 0.2

Getting Started

Establishing a baseline

- Start with a linear regression



```
# Gets a score from a linear regression - starting point.

linear_regression = LinearRegression()
linear_regression.fit(x_train_standard, y_train)
linear_regression.score(x_train_standard, y_train)
```

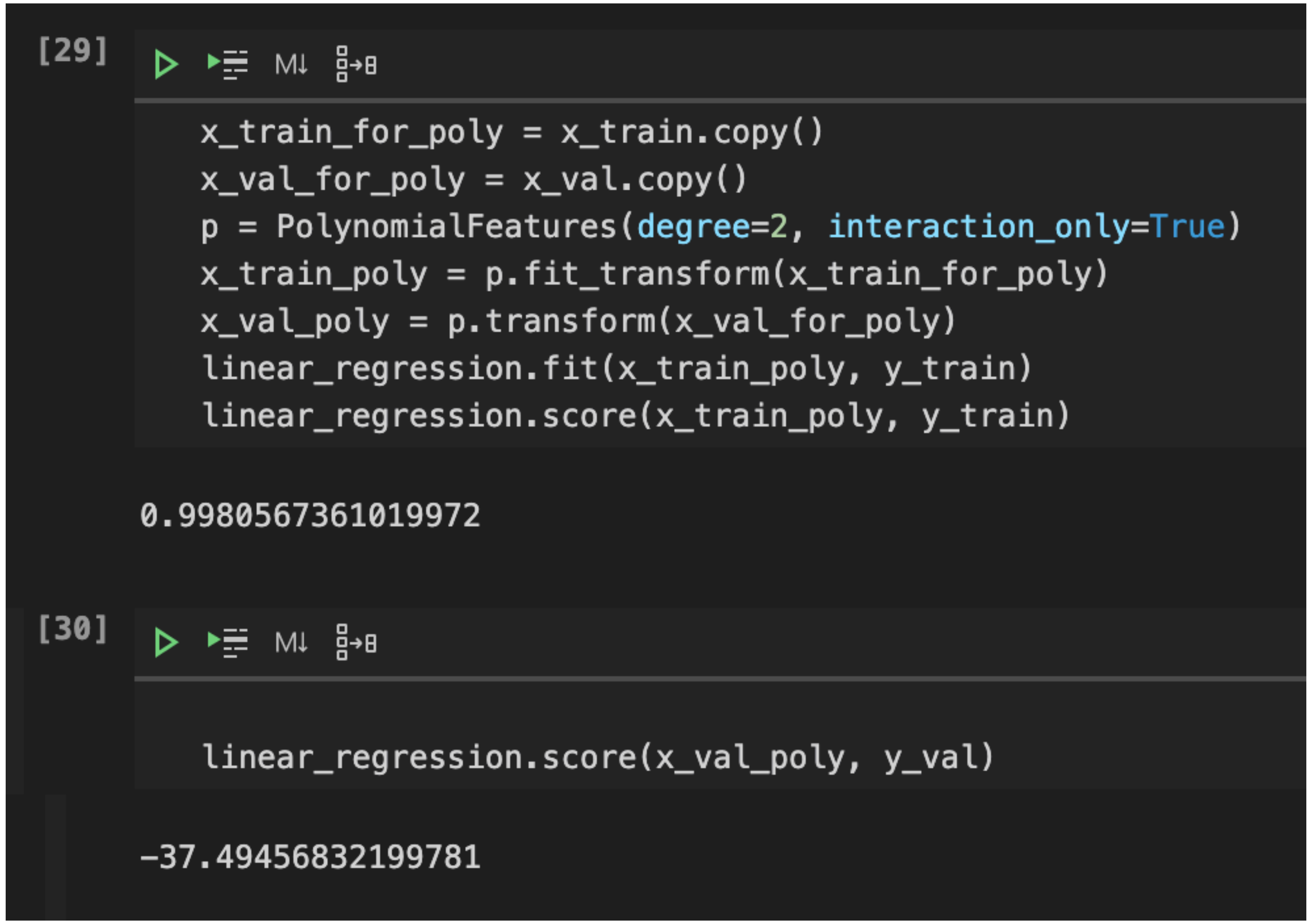
```
0.6022471178535247
```

```
linear_regression.score(x_val_standard, y_val)
```

```
0.6418989030495136
```

Overfit

Easy there, turbo



```
x_train_for_poly = x_train.copy()
x_val_for_poly = x_val.copy()
p = PolynomialFeatures(degree=2, interaction_only=True)
x_train_poly = p.fit_transform(x_train_for_poly)
x_val_poly = p.transform(x_val_for_poly)
linear_regression.fit(x_train_poly, y_train)
linear_regression.score(x_train_poly, y_train)
```

```
0.9980567361019972
```

```
linear_regression.score(x_val_poly, y_val)
```

```
-37.49456832199781
```

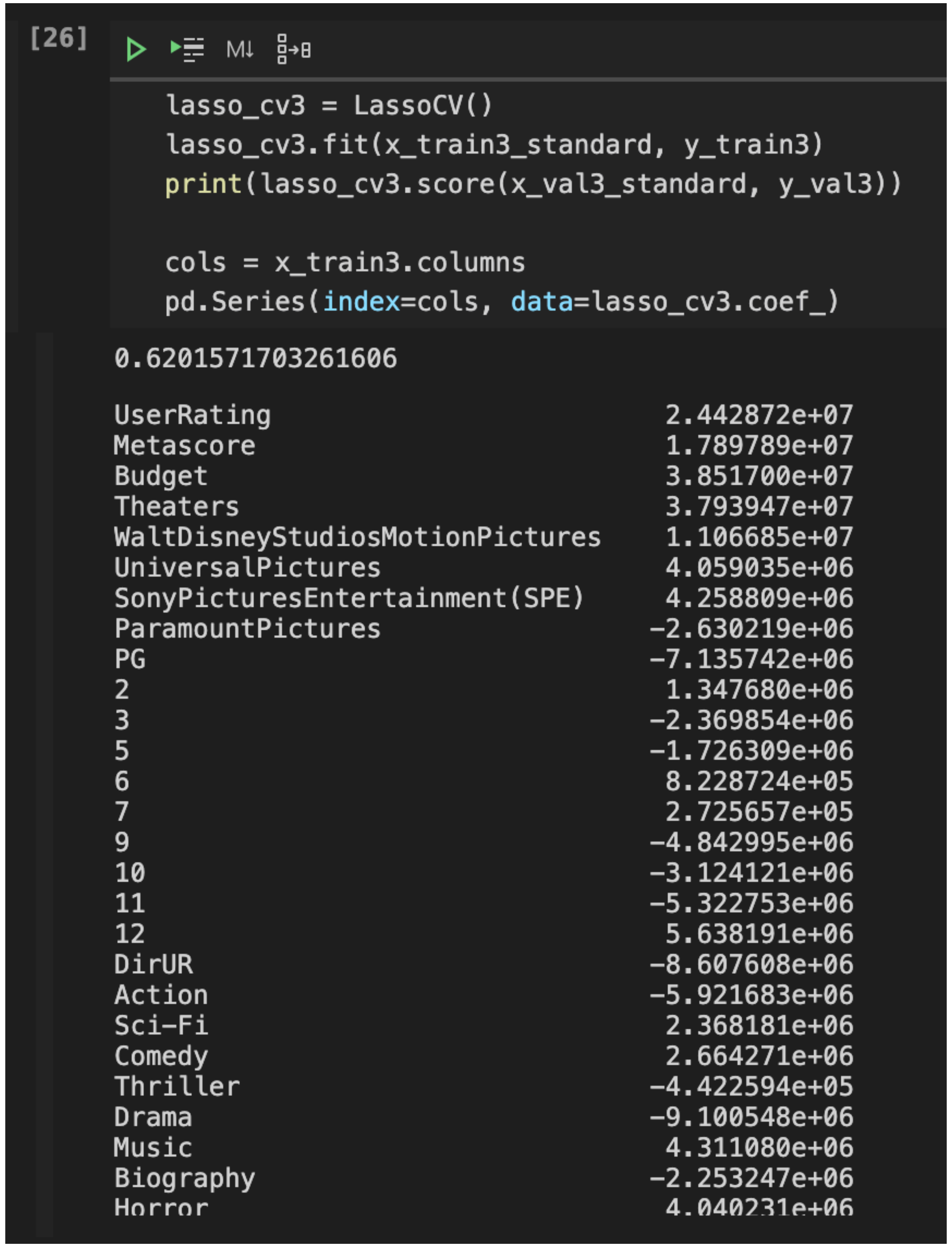
- Polyomial regression scores indicate overfitting
- Run LassoCV instead

Done?

No more zero coefficients

- Ran RidgeCV, but only notable change was to minimize a field that Lasso also removed.

- On the third pass, LassoCV has no more zero coefficients.



```
lasso_cv3 = LassoCV()
lasso_cv3.fit(x_train3_standard, y_train3)
print(lasso_cv3.score(x_val3_standard, y_val3))

cols = x_train3.columns
pd.Series(index=cols, data=lasso_cv3.coef_)
```

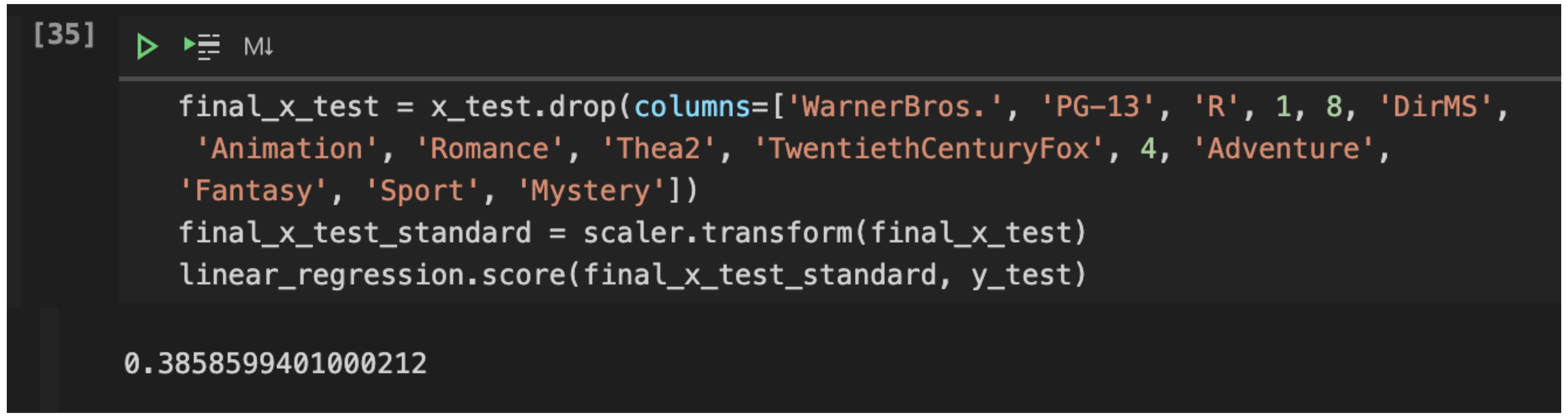
```
0.6201571703261606

UserRating                           2.442872e+07
Metascore                            1.789789e+07
Budget                               3.851700e+07
Theaters                             3.793947e+07
WaltDisneyStudiosMotionPictures      1.106685e+07
UniversalPictures                    4.059035e+06
SonyPicturesEntertainment(SPE)       4.258809e+06
ParamountPictures                   -2.630219e+06
PG                                  -7.135742e+06
2                                    1.347680e+06
3                                   -2.369854e+06
5                                   -1.726309e+06
6                                    8.228724e+05
7                                    2.725657e+05
9                                   -4.842995e+06
10                                  -3.124121e+06
11                                  -5.322753e+06
12                                   5.638191e+06
DirUR                               -8.607608e+06
Action                              -5.921683e+06
Sci-Fi                               2.368181e+06
Comedy                               2.664271e+06
Thriller                            -4.422594e+05
Drama                               -9.100548e+06
Music                                4.311080e+06
Biography                           -2.253247e+06
Horror                               4.040231e+06
```

And the Winner Is…

Not me

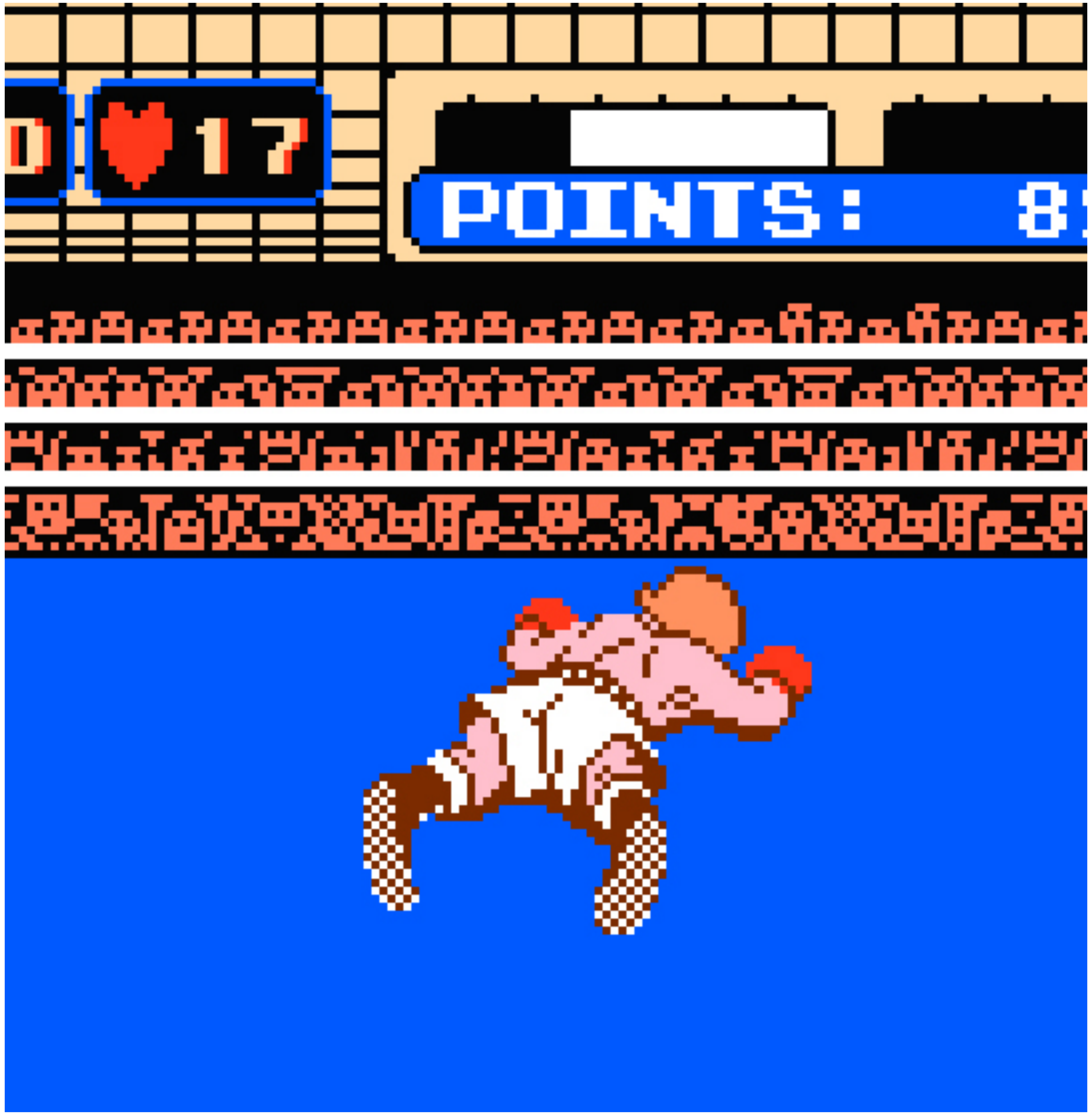
- Final score 0.38!



```
[35]
final_x_test = x_test.drop(columns=['WarnerBros.', 'PG-13', 'R', 1, 8, 'DirMS',
 'Animation', 'Romance', 'Thea2', 'TwentiethCenturyFox', 4, 'Adventure',
'Fantasy', 'Sport', 'Mystery'])
final_x_test_standard = scaler.transform(final_x_test)
linear_regression.score(final_x_test_standard, y_test)

0.3858599401000212
```

- Confirmed with Cross Validation.



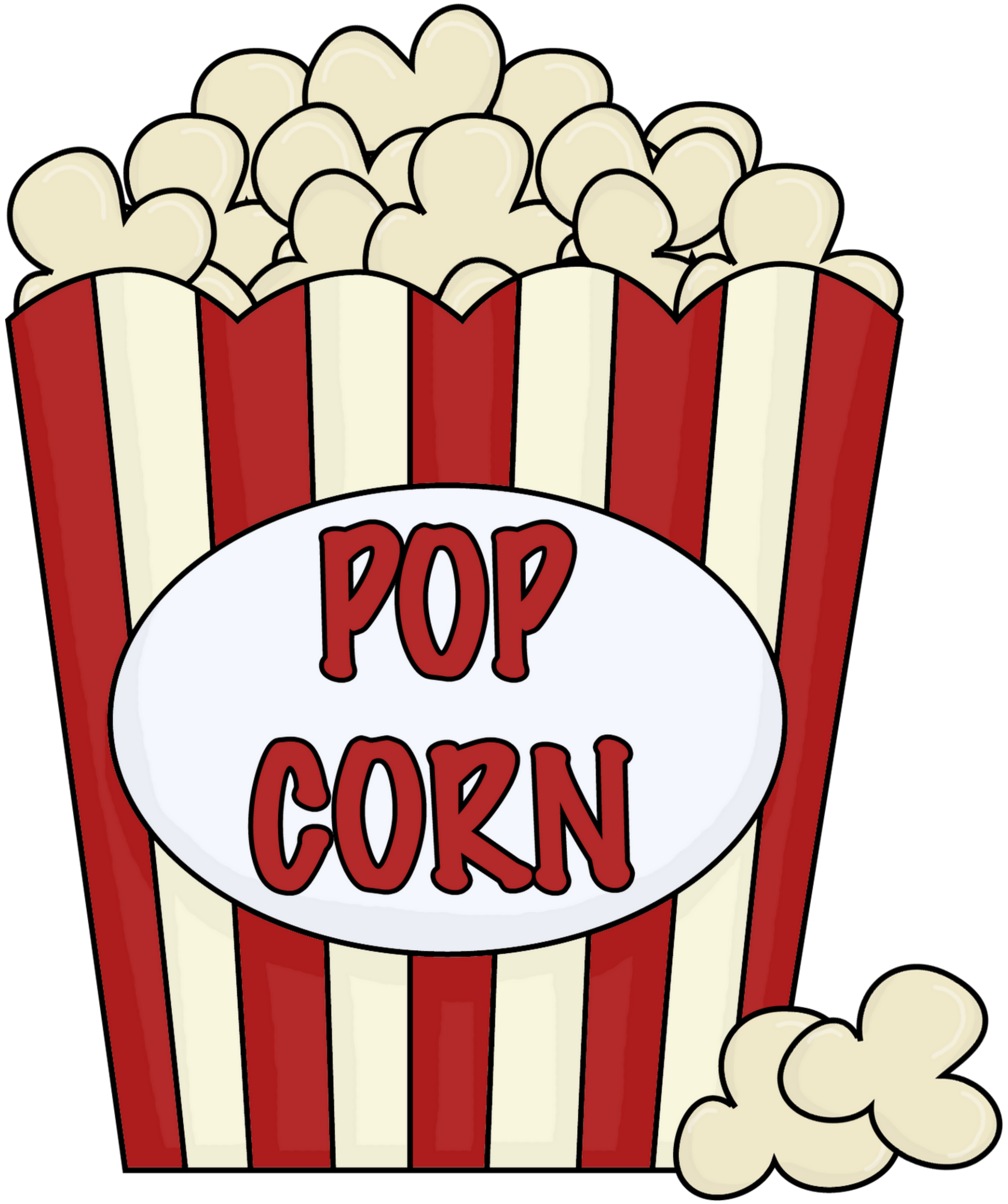
Predictions

What results will I get?

- Theaters and budgets have a notable correlation with movie grosses.

- The shutdown eliminated my best correlation.

- Ironically, I had to extrapolate the number of theaters to predict Gross



Missed It by That Much

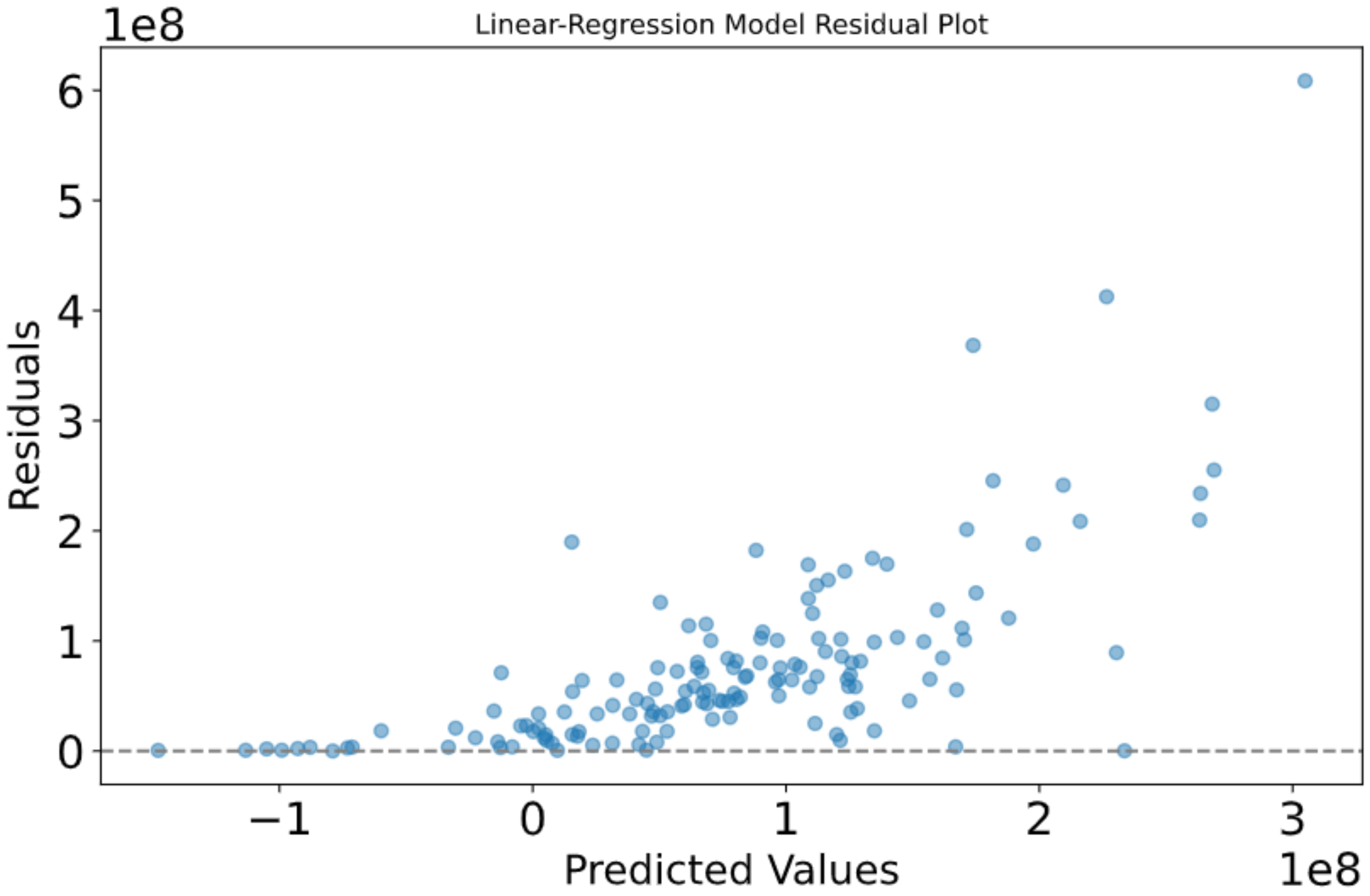
2.02625851e+15

8.25751765e+15

8.056192407e+15

Questions?

Residuals


Linear-Regression Model Residual Plot
1e8
6
5
4
3
2
1
0
Residuals
−1
0
1
2
3
Predicted Values
1e8