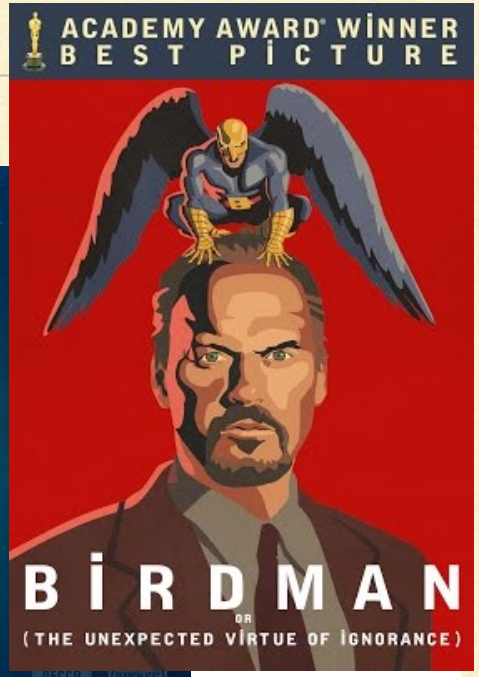
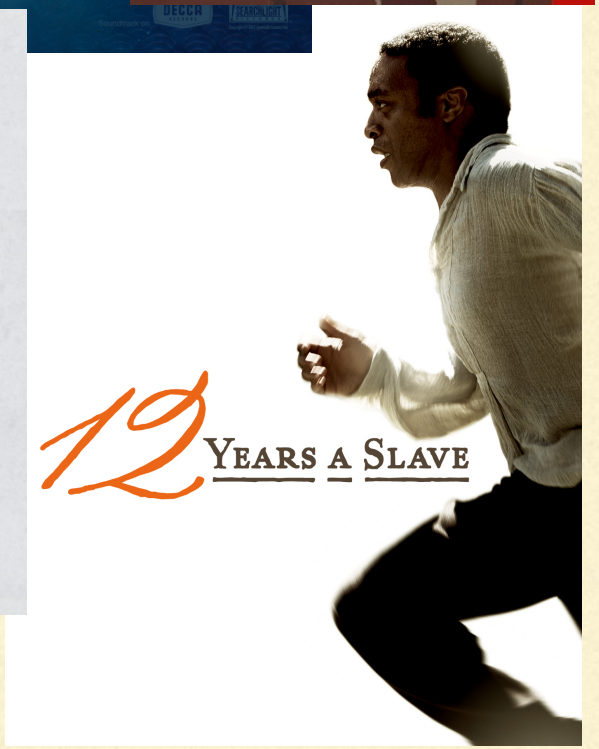
LINEAR REGRESSION ON INDEPENDENT FILM DATA


OFFICIAL SELECTION
TORONTO
INTERNATIONAL FILM FESTIVAL
2017
WINNER
GOLDEN LION BEST PICTURE
VENICE INTERNATIONAL
FILM FESTIVAL
2017
OFFICIAL SELECTION
BFI LONDON
FILM FESTIVAL
2017
A GUILLERMO DEL TORO FILM
THE SHAPE OF WATER
DIRECTED BY
GUILLERMO DEL TORO
SCREENPLAY BY
GUILLERMO DEL TORO & VANESSA TAYLOR
STORY BY
GUILLERMO DEL TORO
IN CINEMAS JANUARY 18


ACADEMY AWARD WINNER
BEST PICTURE
BIRDMAN
OR
(THE UNEXPECTED VIRTUE OF IGNORANCE)


MOONLIGHT


SPOTLIGHT
MARK RUFFALO
MICHAEL KEATON
RACHEL McADAMS
LIEV SCHREIBER
AND STANLEY TUCCI
BASED ON THE PULITZER PRIZE-WINNING INVESTIGATION


12 YEARS A SLAVE

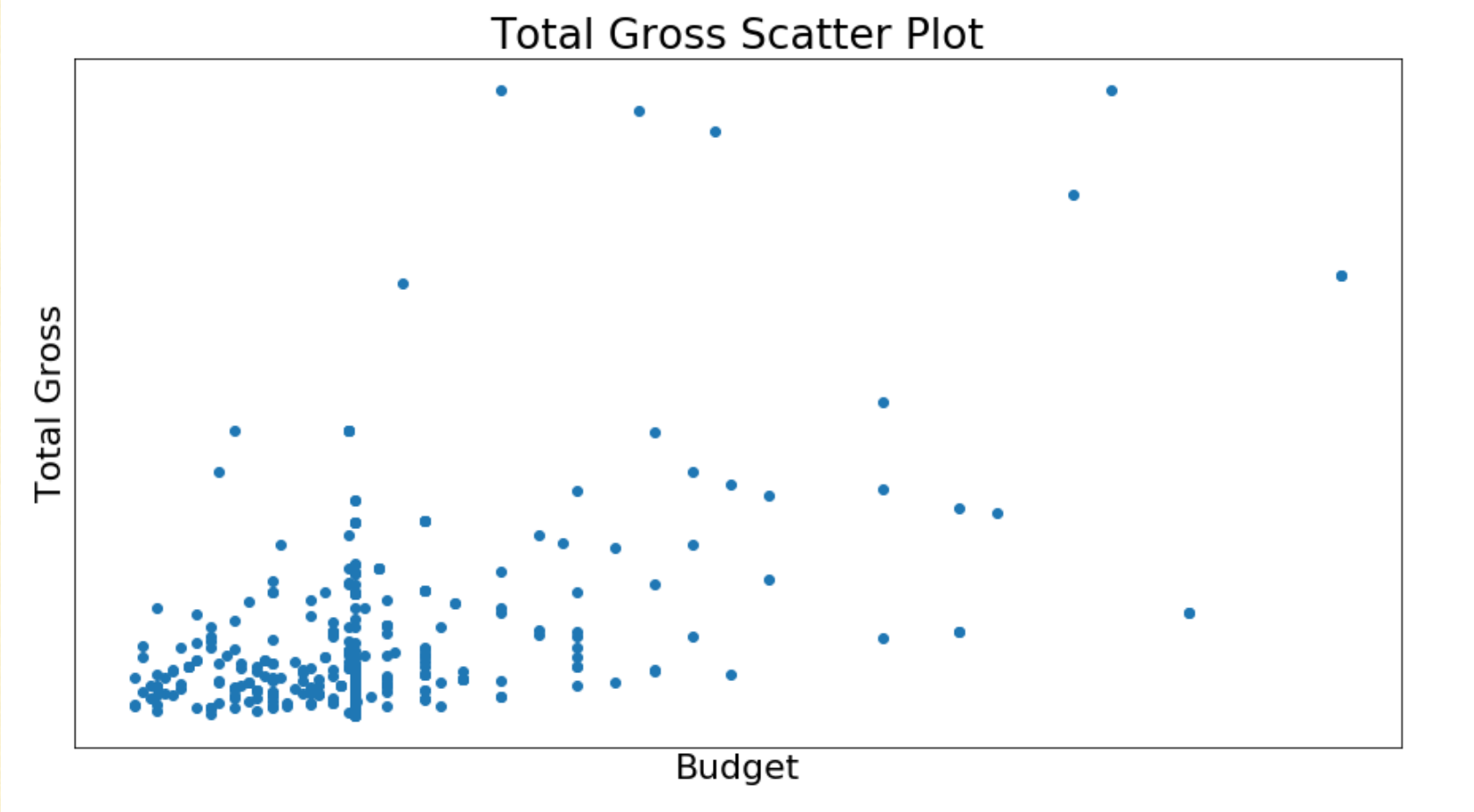
DATA SET

- Movies between 2008-2017 that are not "Big Six"
- Movie Revenue = theater ticket sales + disc sales
- Features:
  - Numerical - budget, runtime, days in release
  - Categorical - genre, MPAA rating, release date

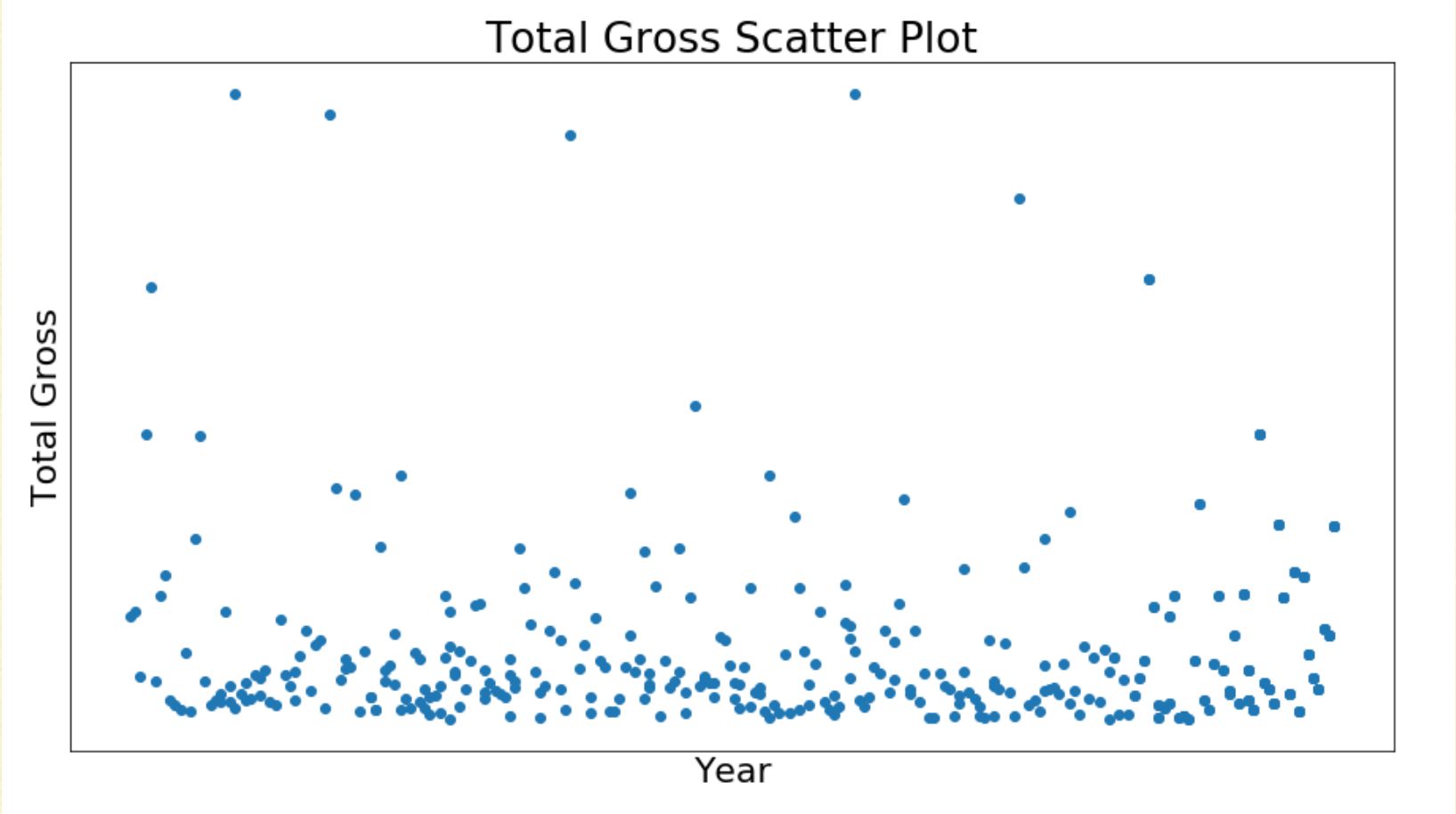
DATA CLEANING

- Adjust monetary values for inflation (2017)
- Impute mean values for NaNs in numerical features
- Merge DVD/Blu-ray sale data

EXPLORATORY DATA ANALYSIS



EXPLORATORY DATA ANALYSIS



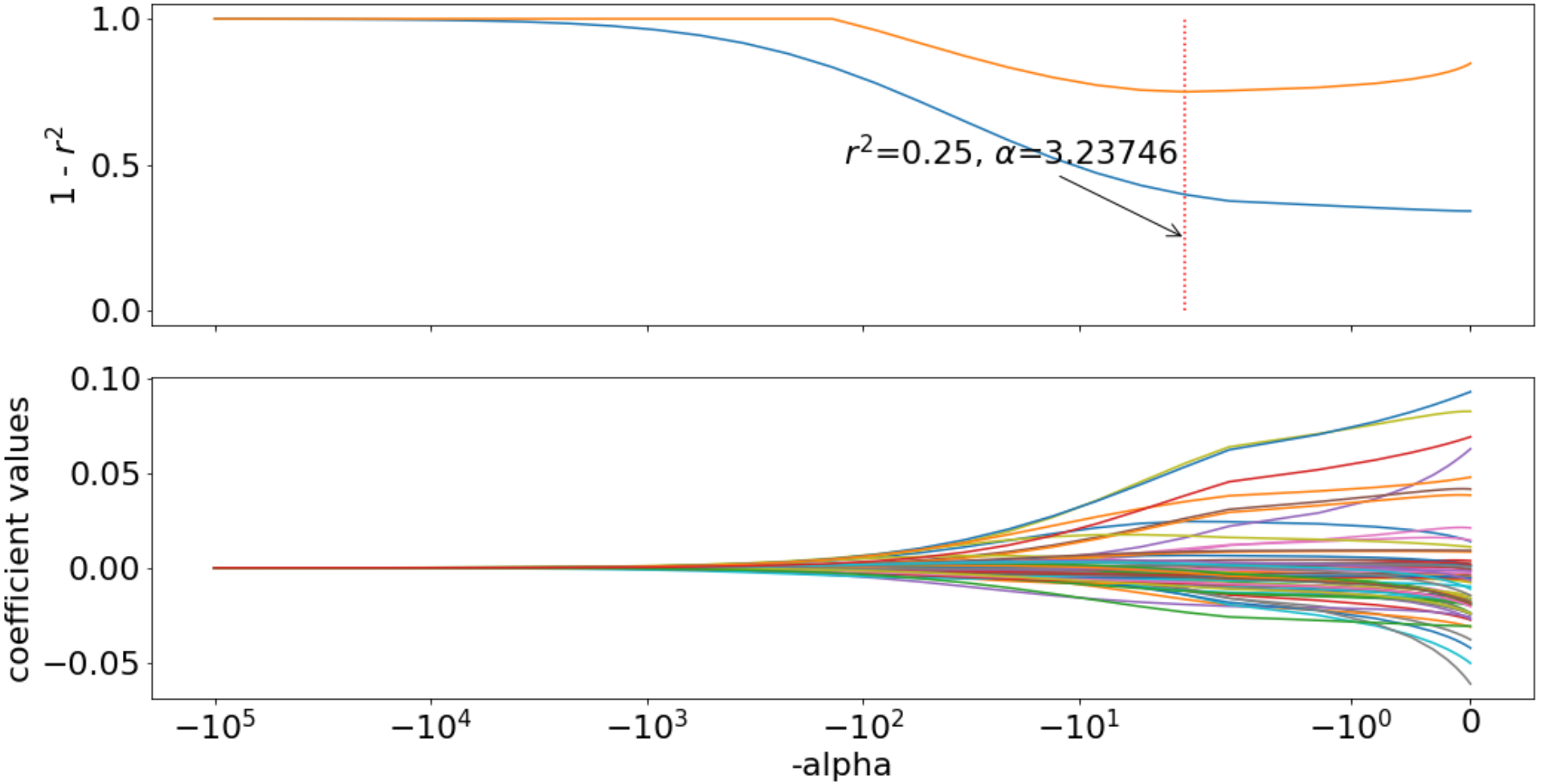
MODELING

- 67 total features
- Scikit-learn modeling - using “normalize” scaling

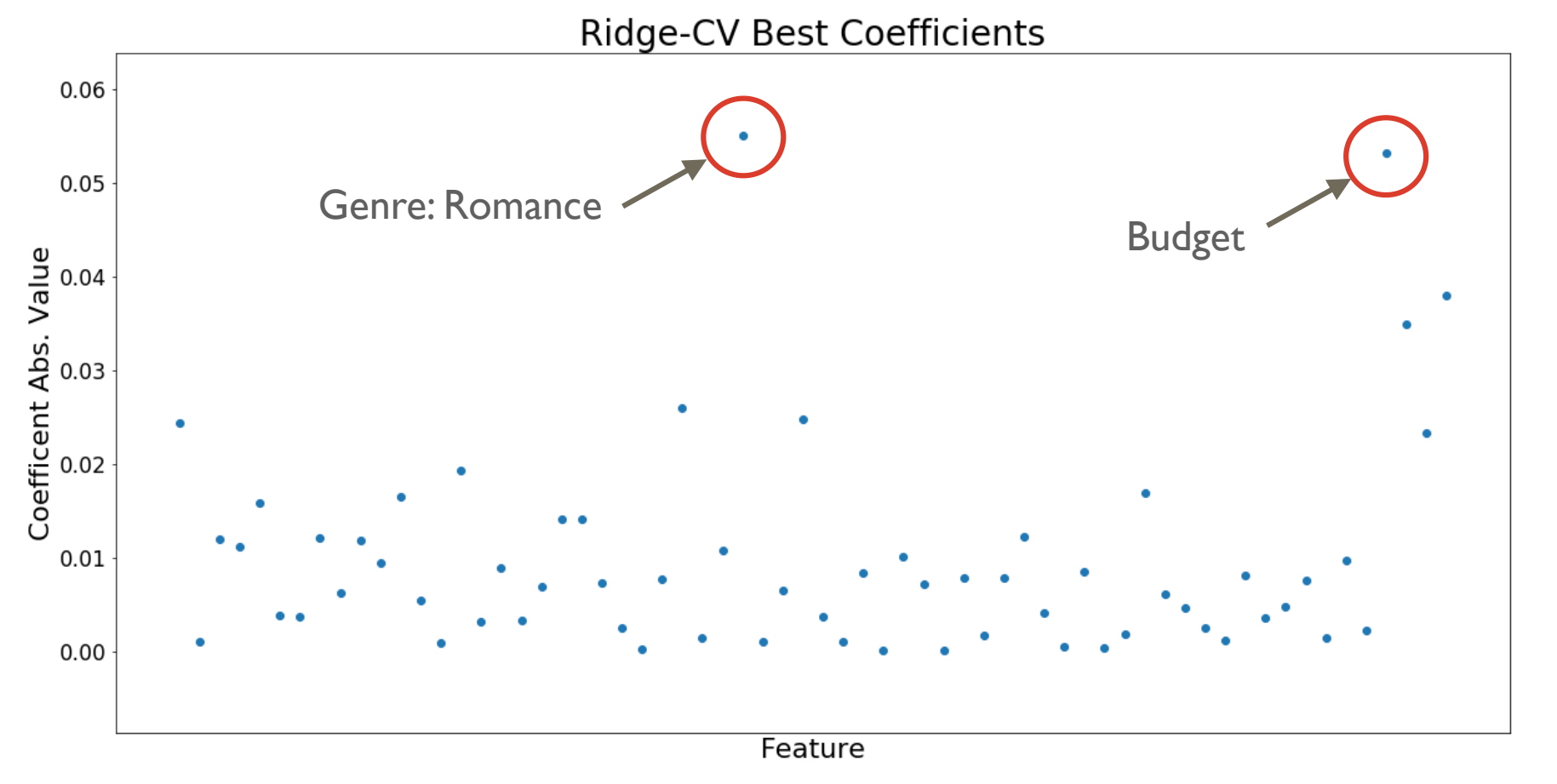
INITIAL MODEL

- Train R squared of 0.66 and Test R squared of -4.06e+26!

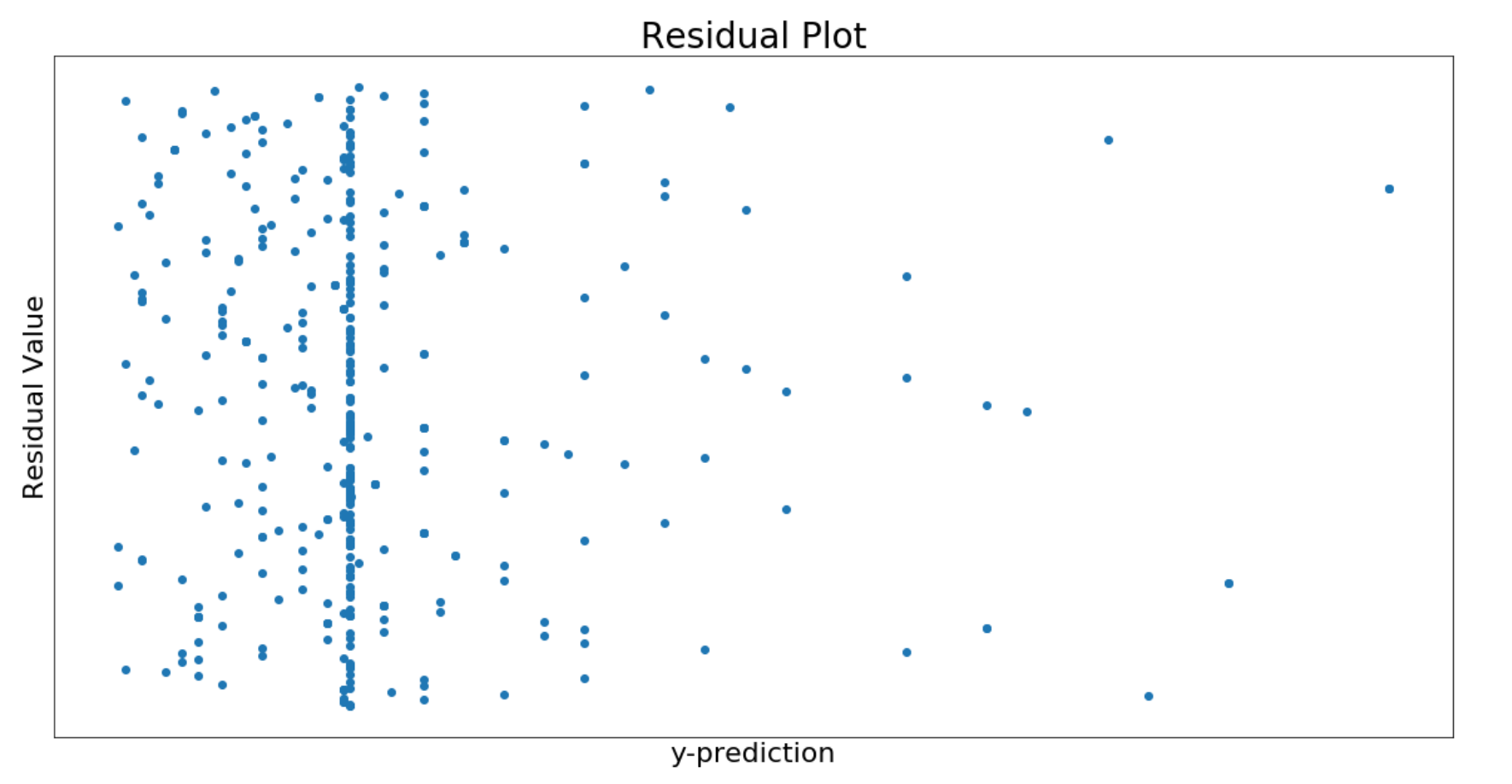
RIDGE-CV REGULARIZATION



RIDGE MODEL


Ridge-CV Best Coefficients
Genre: Romance
Budget
Coefficent Abs. Value
0.06
0.05
0.04
0.03
0.02
0.01
0.00
Feature

RESIDUALS



NEXT STEPS

- Transform to log-scale
- Increase data set
- Retrieve data for viewer/critic ratings
- Investigate retrieval of streaming data

LESSONS LEARNED

- Be thoughtful about feature selection
- Don’t over-filter during data-cleaning
- Models only as good as data put into them

THANK YOU