Movie Revenue Prediction

Mitchell Seiter

Design

- The purpose of this task was to try and build a model to determine if we could predict revenue of a movie based off certain features.
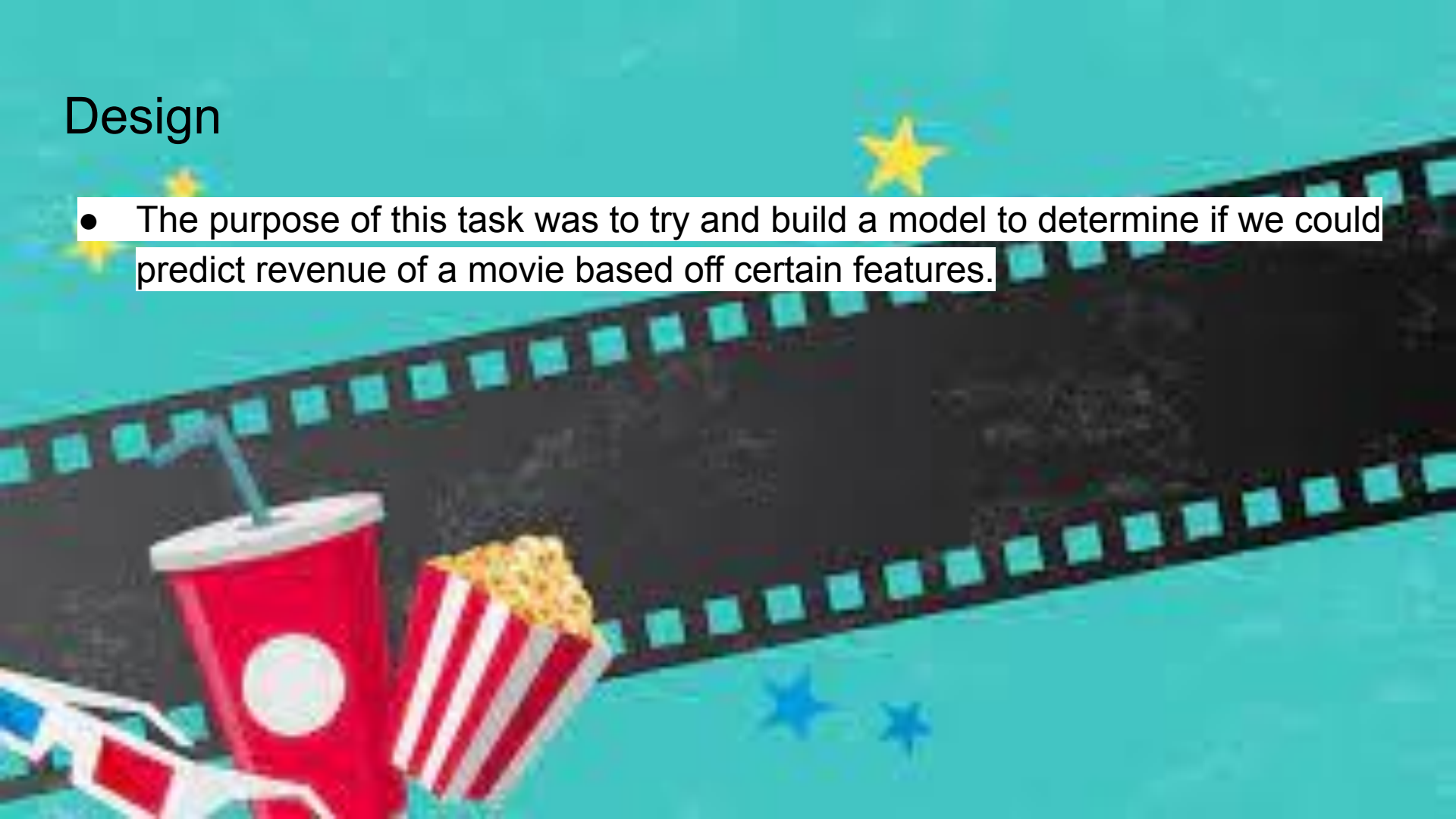
Data

- Using Selenium and BeautifulSoup I was able to scrape the data off of moviejojo.com
- Data consisted of 1000+ entries that were used in the model building
- Some key features included budget, genres, and production companies
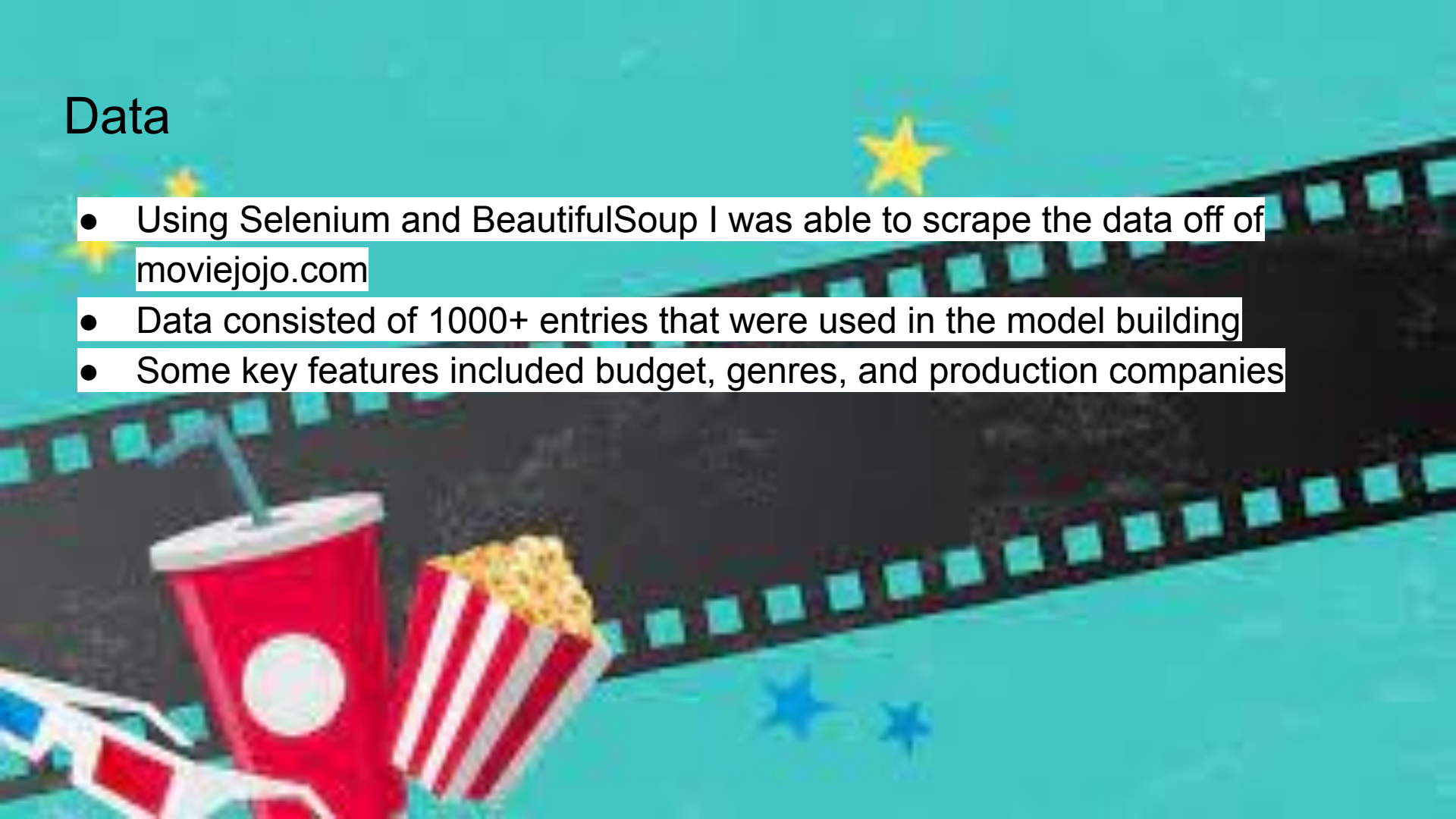
Algorithms

- Was able to build an algorithm that automated almost all of the scraping.
- Did some feature engineering on the DF to convert the categorical features to numerical
- Ran a CV LASSO regression, RIDGE regression, and simple linear regression
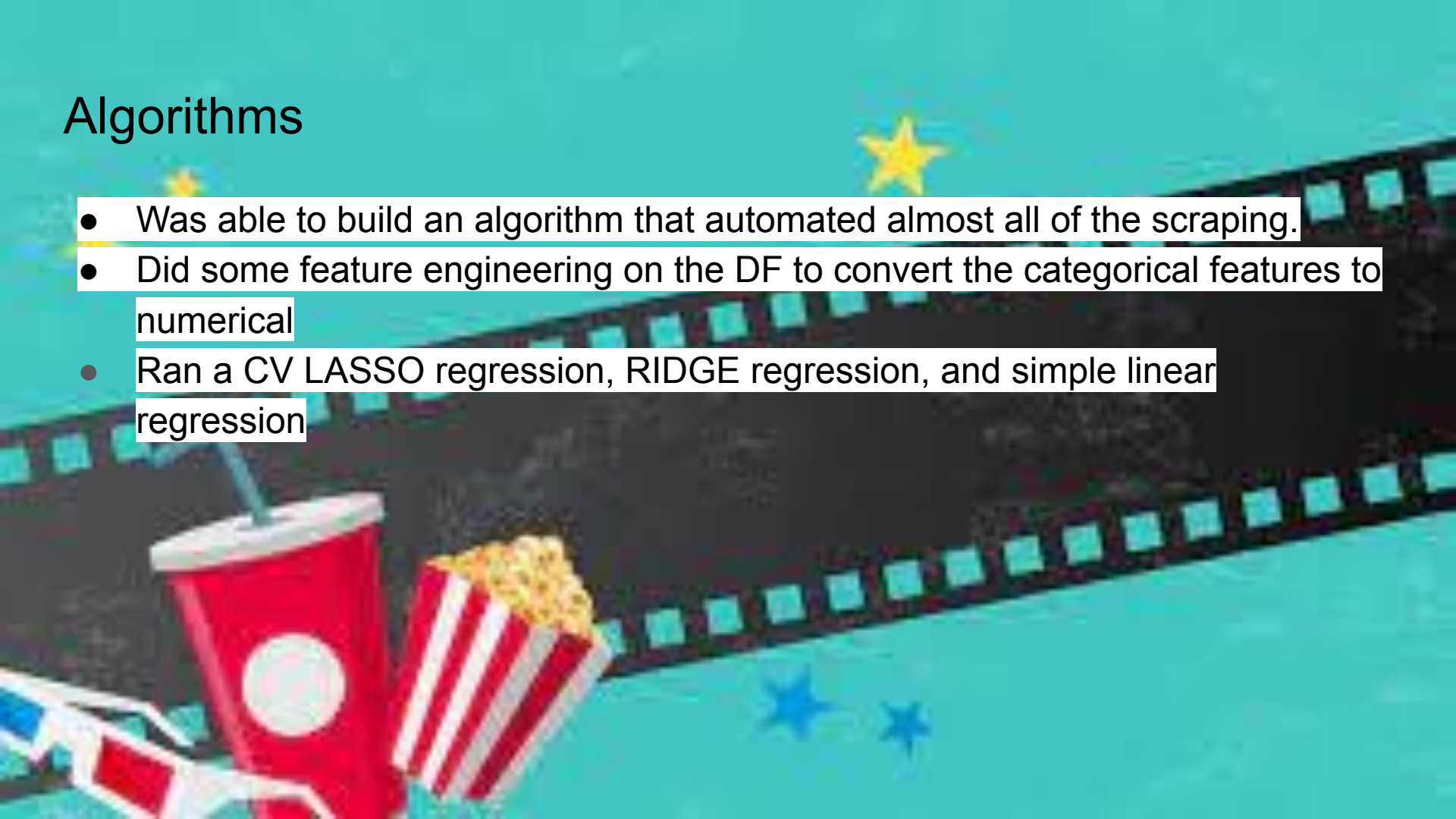
Results

('Budget', 54584281.17371638),
('Run Time', 29698626.504274845),
('Release Count', 61510756.94565977),
('WR**2', 73496923.9619472),
('NOM**2', 18200222.124388628),
("Genre_['Action,Adventure,Sci-Fi']", 15436799.068830376),
("Genre_['Action,Adventure,Thriller']", 13459836.623650305),
("Genre_['Adventure,Animation,Comedy,Family,Fantasy']", 3911101.1432901085),
("Genre_['Comedy']", 1869206.7963927921),
("Genre_['Comedy,Drama']", 1907145.3286956595),
("Genre_['Comedy,Drama,Romance']", 3058499.674553796),
("Genre_['Comedy,Romance']", 2559180.311129556),
("Genre_['Drama,Romance']", 1936628.2460110404),
("Genre_['Horror,Mystery,Thriller']", 3997116.53220809),
('Production_Co_Others', 17809303.333897013),
('Production_Co_Paramount Pictures', 2755555.5942156846),
('Production_Co_Sony Pictures Entertainment (SPE)', 9468982.256711828),
('Production_Co_Twentieth Century Fox', 26192548.53705725),
('Production_Co_Universal Pictures', 15075279.322821783),
('Production_Co_Walt Disney Studios Motion Pictures', 20960161.421733778),
('Production_Co_Warner Bros.', 1648265.6790653446),
('Release_Season_Summer', 23388064.664329324),
('Release_Season_Winter', 6019470.499782998),
('Movie Rating_PG', 6368544.281539175),
('Movie Rating_R', 11993862.42852944)]

Results
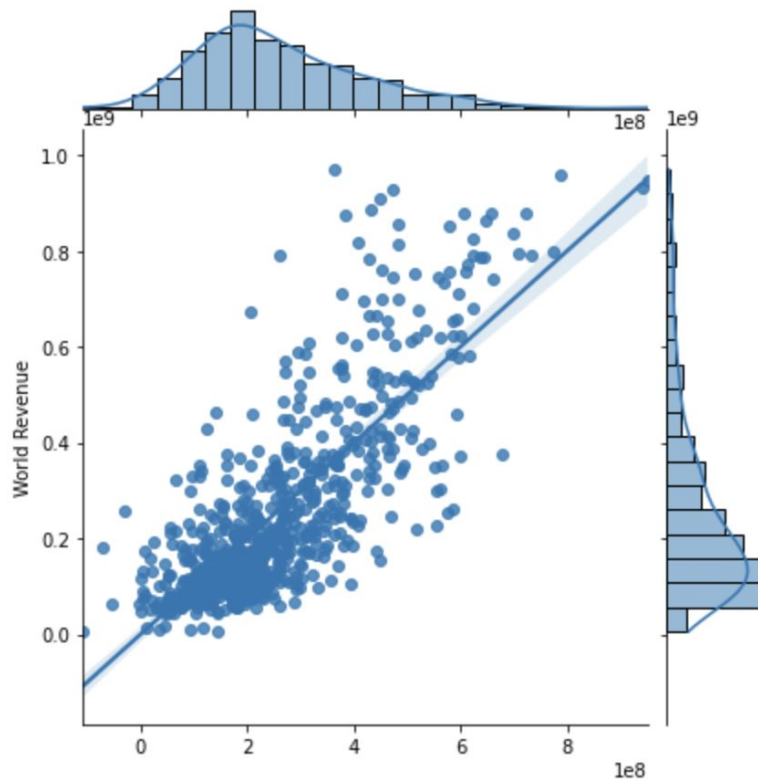
Results Cont.

MAE = 99911193.07667746 (Roughly $100 Million)

RMSE = 131170622.6754773

R^2 Values for LASSO:

0.6106474317594031 (Train)

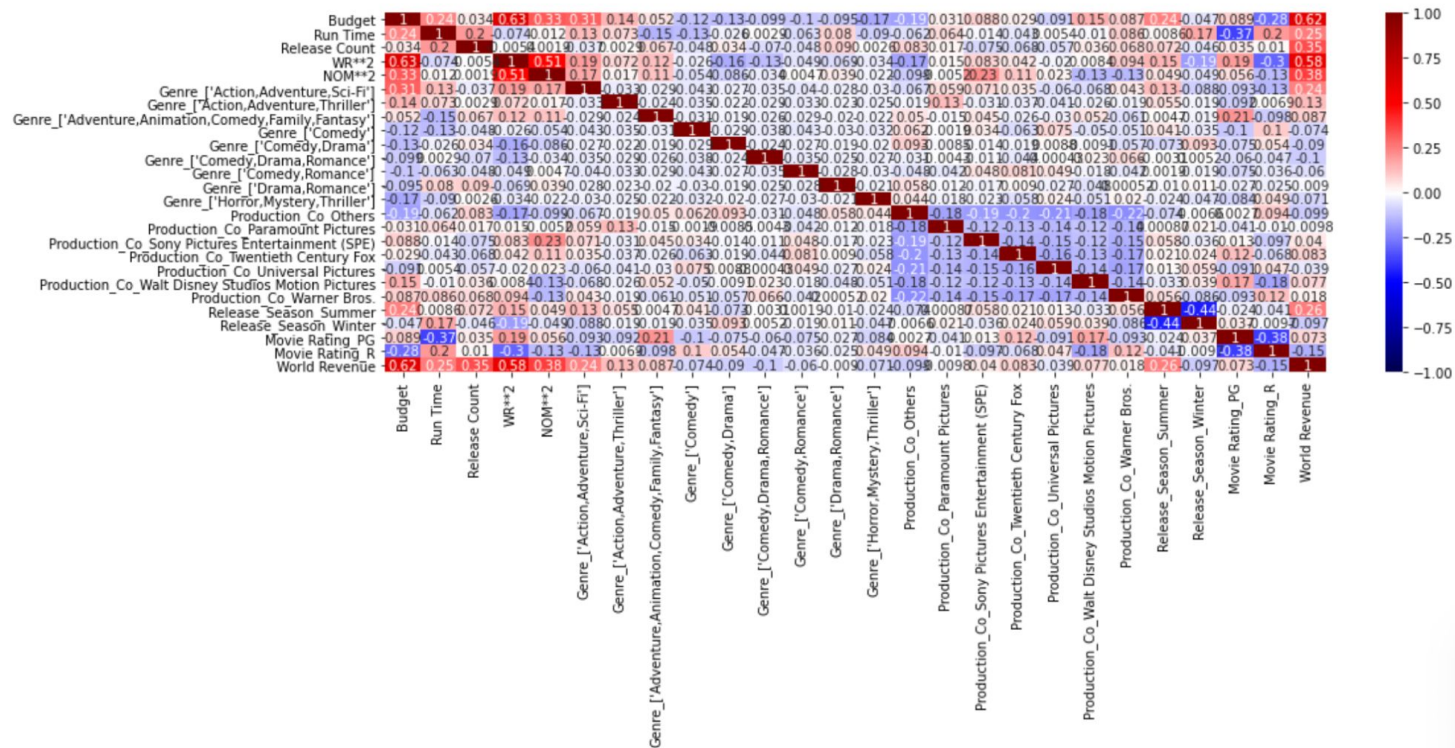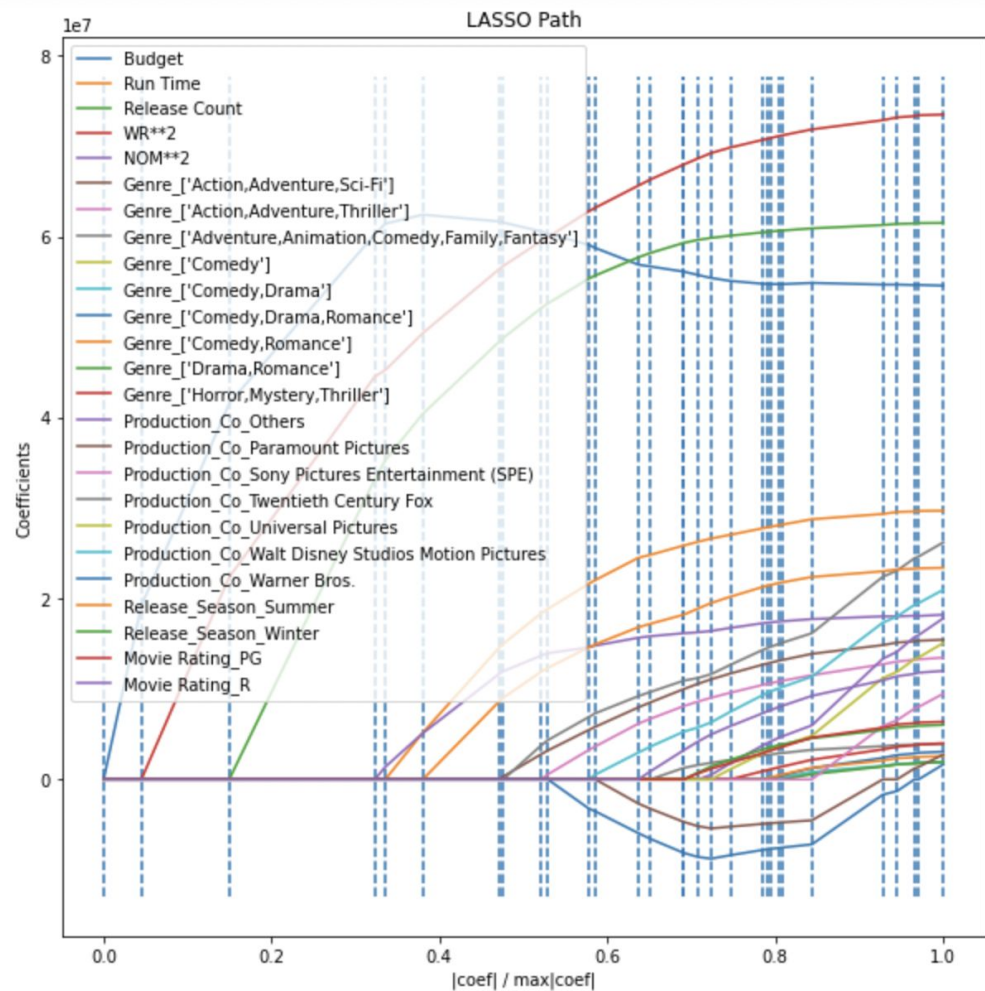0.6074711175196785 (Val)

0.535356485837577 (Test)

Ridge Regression test R^2: 0.545

Simple Linear R^2 0.5709442667716154

Appendix

Appendix

End