Movie revenue predictors analysis

METIS Data Science Bootcamp 2021 - Project 2

Gabriella Nemeth

Introduction

- The clients: newly launched film making company, international interest as well - target: total gross

- Project goal to define the **main predictors** of a movie which has effect on the total gross

- Additional the project adds the extra information how the score of the used model changes in case if the **director is female**

Methodology

- Web - scraping with **BeautifulSoup**
- **Python pandas** package to clean, amend data
- **Python scikit - learn, numpy** used for model creation
- **Genderdize package** to gender identification
- To visualisation of the model and the results with **Seaborn, Matplotlib** package

Project workflow

- Data collection:

  - Box Office Mojo (www.boxofficemojo.com) - **4071 movies' data extracted**

  - Wikipedia (www.wikipedia.com) -

  female directors list (over 1500 directors) - not up to date

  Genderdize package- 356 match on movie list ~ 8.7%

- Data cleaning, exploratory data analysis:

  - dropping outliers, missing values - **filled with median/0**

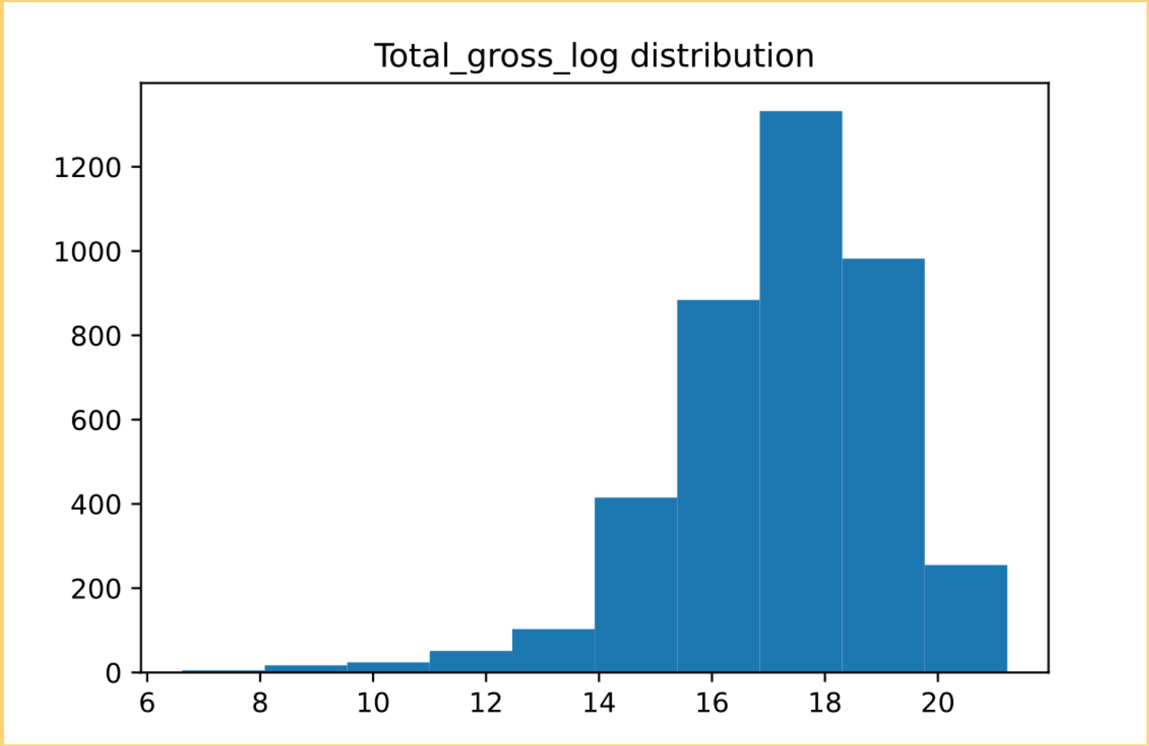
- Modelling

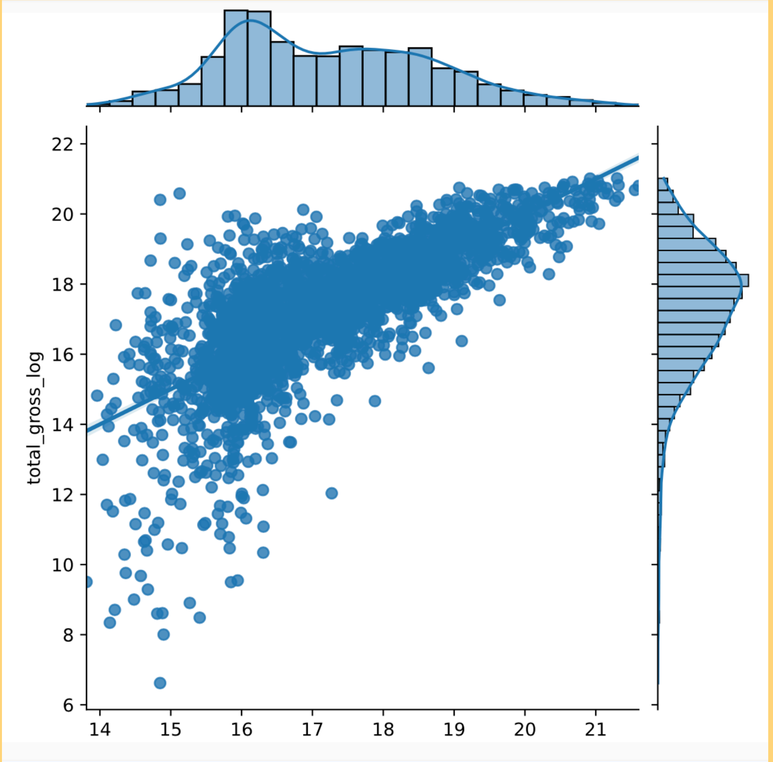
  - finding best fitting model via testing and feature engineering

- Present conclusions

Results

- Genres, rating do not have the expected correlation or effect on total gross
- Number of opening theatres shows an increasing effect
- **Director's gender DO NOT effect significantly the target in this model**
- R2 score: **0.56**, RMSE: **1.26**





Used log transformation on the target (total gross) to make the distribution normal and the results can be explained easier. The prediction (on the right) represents the target's distribution (on the left)

Conclusion

Most important predictors of total gross of a movie are:

- running time,
- budget
- number of opening theatre,
- genres - action

The gender of the director has no significant effect on the movies revenue.

Future work

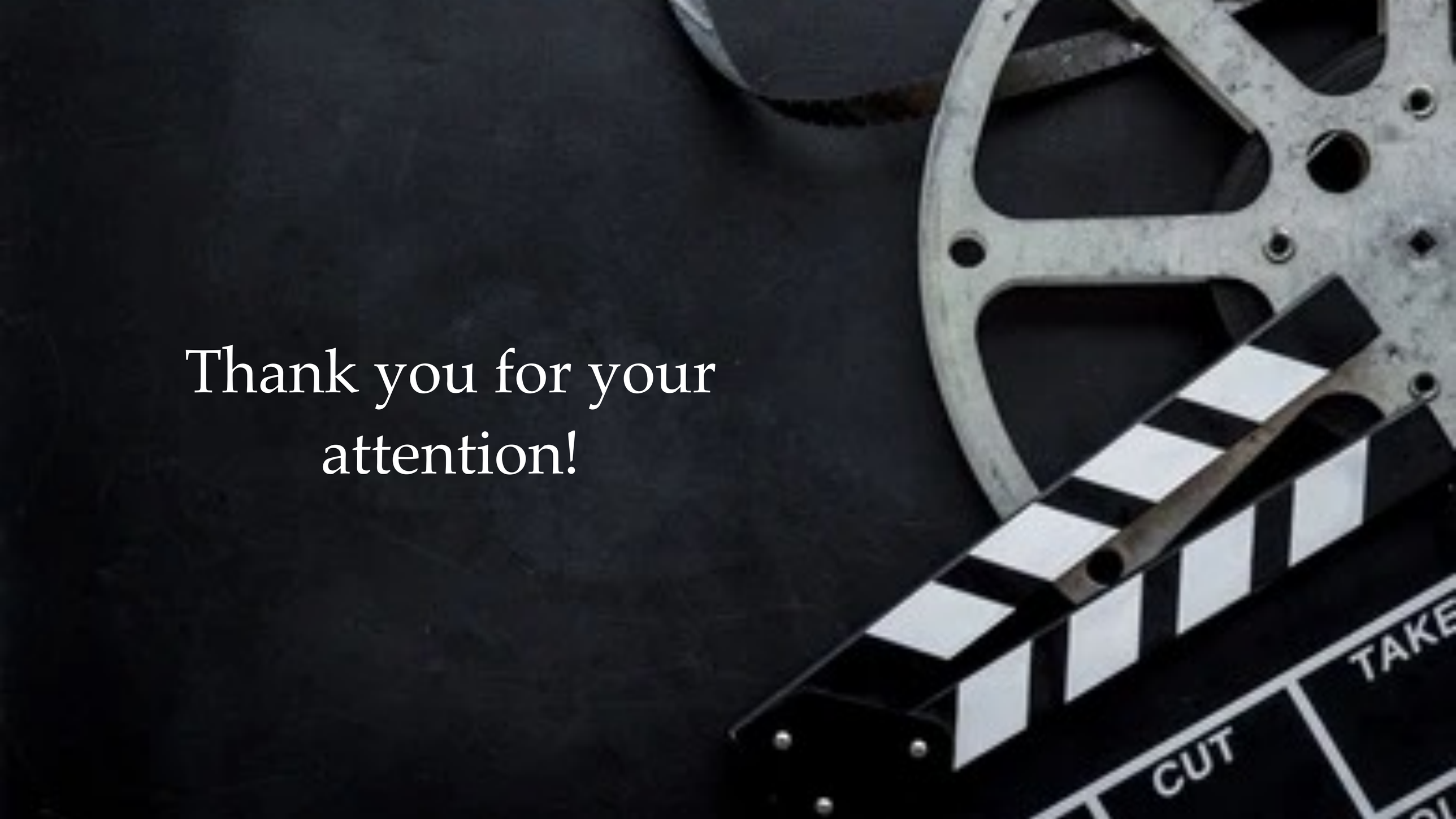
- To investigate further the below topics in movie industry:

  - Gender rates in movie's cast how effects the revenue

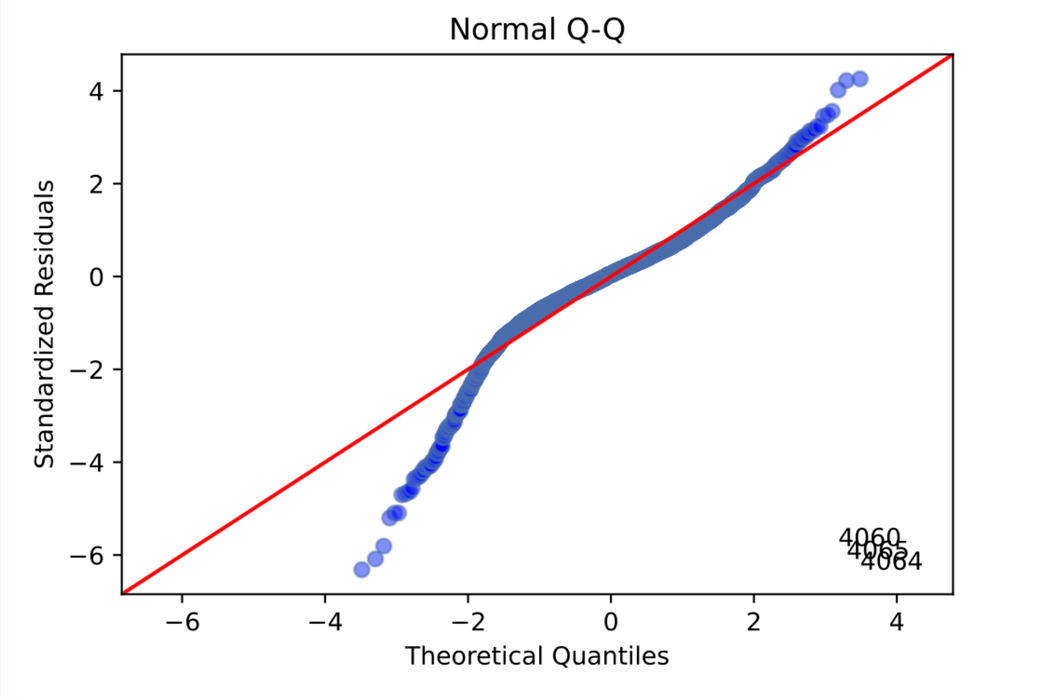
  - Set the target to domestic gross the see results more specifically for USA

  - Try to use more advanced modelling technics to increase the rate of the covered errors

Thank you for your attention!



Appendix



Normal Q - Q plot (left)
- Heavy-tailed

Pairplot of continuous numerical features (right)