Box Office Prediction Case Study Rubric

DS 4002 – Spring 2025 – Matthew Haid
Due: TBD
Submission format: Upload link to GitHub repo to Canvas

Individual Assignment

Why am I doing this?
This case study allows you to leverage your data science knowledge by using regression model analysis techniques to predict opening weekend box office revenue for movies based on pre-release critic reviews. As you work through this assignment, you will be exposed to how text data can be processed and used for real-world financial forecasting applications.

What am I going to do?
The GitHub repository for this case study can be found at:
🔗 https://github.com/SurvivrrHayde/CS3DS4002

You will be provided with a cleaned dataset containing pre-release critic review sentiment scores and movie metadata. You will use Python to perform exploratory data analysis (EDA), feature engineering, and regression modeling to predict a movie's opening weekend box office revenue.

You are expected to train at least one regression model (e.g., Linear Regression) and evaluate model performance using Root Mean Squared Error (RMSE) and R^2. You are encouraged to explore multiple models (e.g., Random Forest, XGBoost) to improve prediction accuracy.

Your final deliverables should include:
- A data dictionary
- Well-documented, commented source code
- Images of graphs plotted for the model evaluation and EDA
- A GitHub repository containing all materials used

Tips for success:
- Don’t overthink it. Focus on building a solid pipeline with EDA, feature engineering, modeling, and evaluation.
- Ensure your code is clean, runs without errors, and is well commented.
- Try multiple modeling approaches if possible.
- Focus on creating an understandable, reproducible workflow.
- It is recommended that you work in Python. Create a roadmap of objectives before starting your notebook.

How will I know I have Succeeded? You will meet expectations on CS3 Create Case Study when you follow the criteria in the rubric below.

Spec Category	Spec Details
Formatting	● One GitHub repository (submitted via link on Canvas) ○ Create a new GitHub repository for this assignment titled CS3_BoxOfficePrediction that contains: ■ README.md ■ LICENSE.md ■ Source Code File (Colab or Jupyter Notebook) ■ Your data (provided dataset) ■ Evaluation (Graphs, writeup, modeling, etc.) ■ REFERENCES.md
README.md	● Brief summary of what you've produced for the case study. ● This markdown file should provide enough information for people to understand the project contents and purpose.
Source Code File	● A well-documented Jupyter Notebook (or Colab notebook) containing your full workflow. ● The notebook must include: ○ Exploratory Data Analysis (EDA) ○ Feature engineering steps ○ At least one regression model trained and evaluated ○ Evaluation metrics (at minimum: RMSE and R^2) ○ Well-commented code and clean structure
REFERENCES.md	● Markdown file titled REFERENCES.md that includes any resources that helped you during the case study. ○ Use IEEE Documentation style citations.

	○ Include a brief 1–2 sentence statement under each citation explaining how each reference helped you.
Graphs and Plots	● Include plots that illustrate key findings in EDA. ○ Include graphs of model evaluation metrics when appropriate (e.g., residual plots, scatter plots of predictions vs. actuals)

Acknowledgements: Special thanks to Jess Taggart from UVA CTE for coaching on making this rubric. This structure is pulled from Streifer & Palmer (2020).