Day 1 - 02/06/2025

Topic: Assemble the next big blockbuster film.

Research question: Based on the predicted ratings and box office gross of a movie, can we accurately predict whether it will become a blockbuster?

- Based on the ratings and the gross prediction, do we predict a new movie to be a blockbuster or not?
- Decide the decision boundaries
- Possibly two parameters

Summary

- Meeting the supervisors
- Getting into groups
- Deciding the topic - Movie Dataset
- Looking at the dataset together
- Dividing homework tasks - and what we want to explore in the data
- Planning scraper
- Planning how we'll get the link to each movie (333 movies each) for the scraper

Logistic/progress meeting / rough timetable

1. Exploratory Analysis →
    - Intuition (which columns make sense to keep)
    - What models could work?
2. Preprocessing/Feature engineering →
    - Defining Parameters and Hyperparameters
3. Base models - try to beat these →
4. Hyperparameter tuning →
5. Experimentation →

** Meeting Thursday at 10:00 - 14:00 first week

Brainstorms

- Other datasets
- Using Summary (expanding dataset)
- Using themes (expanding dataset) + Plot keywords
- Using Text Mining for the Overview

What Does Our Data Look Like?

- Info
- Features
- Target Variable
- Missing Data

Homework:
Leon: Visualize distributions and empty entries.
Julia: Summary of each column, and do background research into each column. Check if the Star number in the dataset aligns with the actor role importance. Check if the scraping stuff is legal - ensure reproducibility.
Lettie: Identifying possible alternative RQs

Context
IMDB Dataset of top 1000 movies and tv shows.
You can find the EDA Process on -
https://www.kaggle.com/harshitshankhdhar/eda-on-imdb-movies-dataset
Please consider UPVOTE if you found it useful.

Content
Data:-
Poster_Link - Link of the poster that imdb using
Series_Title = Name of the movie
Released_Year - Year at which that movie released
Certificate - Certificate earned by that movie
Runtime - Total runtime of the movie
Genre - Genre of the movie
IMDB_Rating - Rating of the movie at IMDB site
Overview - mini story/ summary
Meta_score - Score earned by the movie
Director - Name of the Director
Star1,Star2,Star3,Star4 - Name of the Stars
No_of_votes - Total number of votes
Gross - Money earned by that movie

Poster_Link (string): The URL for the poster image of the film. This field contains long text inputs in URL format and is movie-specific. It is usually not analyzed but may be useful for visualization or UI purposes

Series_Title (string): The title of the movie. A string field, usually 2–5 words in length. It clearly defines the movie and is a label or index in analysis and visualizations.

Released_Year (integer): The release year of the movie. This integer field enables trends over time to be analyzed, for example, trends in ratings or genres by decade.

Certificate (string/categorical): This is the film's age/content rating (e.g., A, U, UA, PG-13, R). This is a short text field and usually contains few repeated categories. Useful for filtering by audience appropriateness.

Runtime (integer/numeric): The movie's length in minutes. This is a numeric field gleaned from lines such as "142 min". This can be used in duration-related analysis or to normalize ratings ?

Genre (string/categorical): A comma-separated list of genres relevant to the film (e.g., "Action, Drama"). A string field that may be split into multiple binary fields for genre-by-genre analysis.

IMDB_Rating (float): User rating on IMDB of the film, typically in the range 1-10. An integer field which may be used to monitor popularity or quality based on crowd reviews.

Overview (string): Brief synopsis or description of the film plot. This 20–50 word long text field can be used for natural language processing activities like sentiment analysis or summarization.

Meta_score (integer): Metacritic score (critic rating) usually on a scale of 0–100. This numerical field is an addition to IMDb ratings, and user and critic comparisons are feasible. Some are missing.

Director (string): The title of the director of the film. This is a text field and can be used for examining the effect of specific directors or comparing average ratings by director.

Star1 (string), Star2 (string), Star3 (string), Star4 (string): These columns have the lead actors of the film. They are all text fields. All together, they can be used for examining the network, star power, or casting behavior.

No_of_Votes (integer): The total number of user votes on IMDb. This numeric field is a measure of a movie's popularity or visibility. It can also vary significantly between movies.

Gross (integer, may have missing values): The worldwide box office gross, typically in USD. This field is numeric but may have missing values. It's processed for financial success and can be inflation-adjusted or region-adjusted.

Inspiration (from Kaggle)
Analysis of the gross of a movie vs the directors.
Analysis of the gross of a movie vs different - different stars.
Analysis of the No_of_votes of a movie vs the directors.
Analysis of the No_of_votes of a movie vs different - different stars.
Which actor prefers which Genre more?
Which combination of actors is getting the best IMDB_Rating maximum time?
Which combination of actors is getting good gross?

Day 2 - 03/06/2025

Plan:

- Git (morning) ✔
- Present our homework findings to each other ✔
- Cleaning Data ✔ (some done)
- Leons scraper plan
- Start collecting the links
- What models would we like to use? Talk to supervisors for guidance. ✔

Git Notes:
Fancy file saving system. Save files within the repo. Commit changes to refer back to in the future. If we want to make changes to a repo, we make a branch. Merge the branch into the main branch. Use git carefully. Basic actions: committing, branching, pulling/rebasing, pushing, merging. Git etiquette: use imperative, capitalize subject, use body to explain why and how + what (tags with issue tickets). Commit based on a theme, not quantity.

Git branch name - creating a new branch
Git branch - checking all branches
Git checkout -

What did we do today?

- LEARNING GIT!!!! - Super important, we noticed we don't know enough
- Presented our homework findings
- Establish an initial research question
- Looking for box office gross income
- Decided on using the simple Bert -> https://huggingface.co/lvwerra/bert-imdb
- Started looking at how we'd use NLTK -> https://medium.com/@khalidassalafy/sentiment-analysis-with-nltk-4afbb0bf6a49

Notes on Dataset:

- "Updated four years ago" - the dataset could be old (4-5 years old).

What is a blockbuster? -> Cambridge Dictionary
High money + high number of votes + high ratings = blockbuster
High money + high number of votes + low ratings = memes
Low money + low number of votes + low ratings = flop
Low money + low number of votes + high ratings = hidden gem

Alternative research questions:

1. How do cast and crew combinations influence a movie's success probabilities?
    - Stars
    - Director
    - Genre?
    - Number of votes
    - Rating
2. Are there specific genres, themes or runtimes that are most likely part of a well-acclaimed film?
    - Genre
    - Web scraper themes
    - Number of votes
    - Rating
3. How have the reactions to movies changed over time?
    - Released year
    - Certificate
    - Genre
    - Number of votes
    - Rating

Edited research question: Based on the other features, can we predict the gross income and ratings of a new film?

- Predict gross income and ratings (number of votes)

To do

1. Clean data
    - Web scraper
    - Fill gross
    - Remove unnecessary columns → link

Homework:
EVERYONE watches at least one video on GIT - especially branching and merging.
Julia and Lette: get an IMDb Pro account (free trial) for clean data collection (sacrifices must be made).

Day 3 - 04/06/2025

Plan:

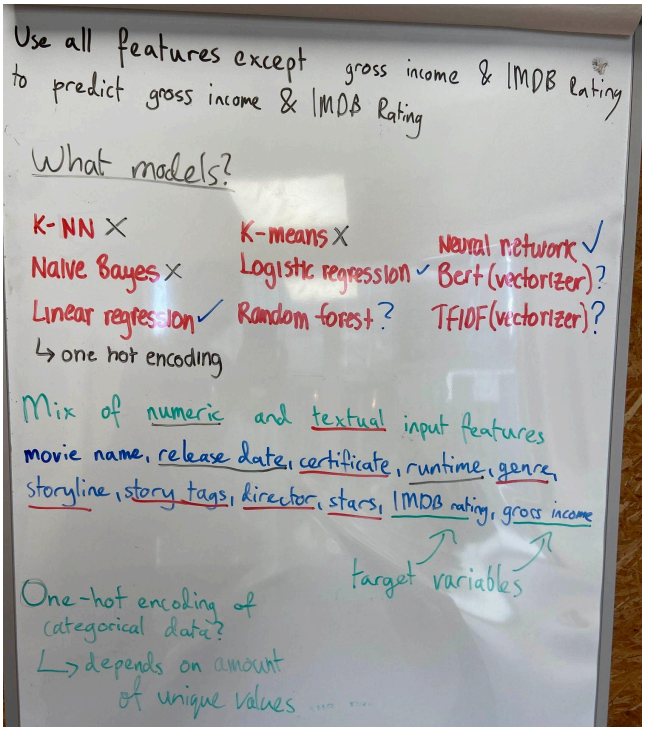
- Lettie:
    - Fill in gross income ✔

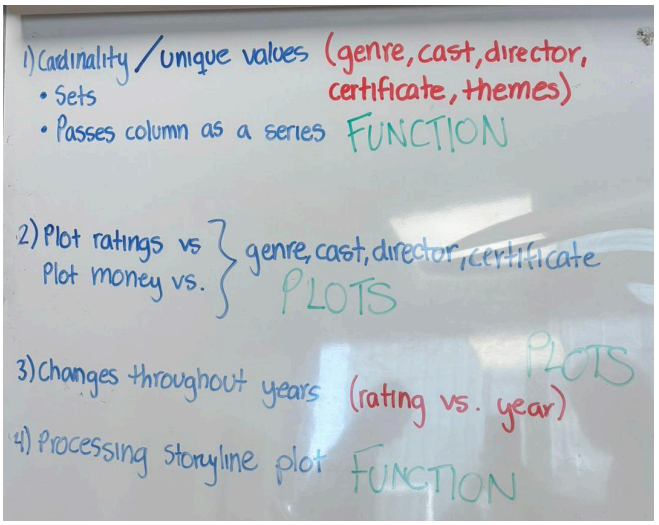
    - Remove useless column ✔
- Julia:
    - Start finding links (333+111) ✔
- Leon:
    - Fix histogram ✔
    - Do number of votes histogram
    - Start webscraper work ✔

Notes:
We've decided not to use the number of votes as it is not a clear indicator of popularity or how many people watched it (gross income would be a better predictor of this). This makes our data mostly textual.
We could possibly use one-hot encoding to create models that use linear regression, logistic regression, random forest, and neural networks.

BERT + TF IDF , Random forest, Neural Network - Top picks?

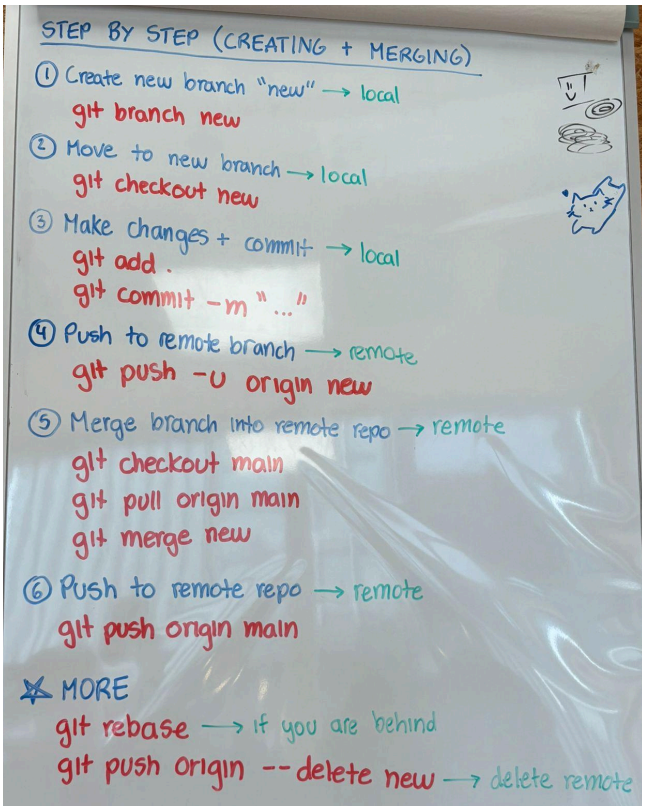
Our board:





Planning Functions for further Exploration:

- Function for the unique values: sets - cardinality
- Average of the year, maybe made into sets (70s,90s..., etc)
- Certificate (affecting how it reaches viewers)



- See score averages, and compare with higher scores (eg, above 7)
- Take a look at actor frequency and relationships: if any actors appear multiple times in the dataset (they are all top 1000). Same for the director.
- Function for lemmatization

What we did:
This day involved a lot of brainstorming. We changed our plan quite a bit, and will be using a new dataset provided by Maria. While waiting for the dataset, we brainstormed about what models would be fit for our project, taking our goals and data into account. We discussed preprocessing and decided on things we'd still like to take a look at within our data. We planned the architecture and started designing functions we could implement as soon as we got our new dataset from Maria. Besides the planning, we took time to speak to another group and get an insight into their plan of action. It was helpful to see how a different group is approaching their data and what steps they're prioritizing for the first week. During this exchange of ideas, it was brought to light that we could play around with our data a bit more than previously planned. Doing so, we gain a deeper understanding of the trends in our data and are better prepared for getting into the models, our work for week 2.

Day 4 - 05/06/2025

Plan:
- Google Meet with Christ at 10, introductions, and day plan. ✔
- 10:30 Meet at Lettie's: Discuss project outline and daily tasks. ✔
- Task Assignment:
    - Julia: work on the diary (clean text, add missing information, and push day 3), make functions with the actors ✔ and directors, look for patterns there, ✔ make the years into categories (and a histogram out of it).
    - Lettie: creating the Cardinality / unique value function, and constructing visualizations. ✔
    - Leon: processing the storyline/plot. ✔
    - Mehmet: exploring trends in the year and runtime. Is there an ideal runtime? + creating visualizations + correlations between different numeric values (runtime, average rating, total votes, gross worldwide box office) ✔
- Meeting at 15:30 to combine what we've done and be ready for the 16/16:30? Meeting. ✔

What we still need to do:
Julia finishes the actor duo analysis and starts the director analysis.

Leon finish creating dataset
Further exploratory data analysis, preparing for the new dataset.
Once we have the new dataset, apply what we've done.
Next week we should start with the models (by Tuesday).

Day 5 - 06/06/2025

Plan:

- Productively wait for the scraping
- Create functions to use on the next dataset
- Start exploring what we have from the dataset
- Julia finished the actor set and director set. Make changes to analysis to fit the new data.

Preprocessing:

- ☐ Remove tconst
- ☐ Remove titleType
- ☐ Remove Column7?
- ☐ Remove OriginalTitle
- ☐ Remove EndYear
- ☐ Remove numVotes

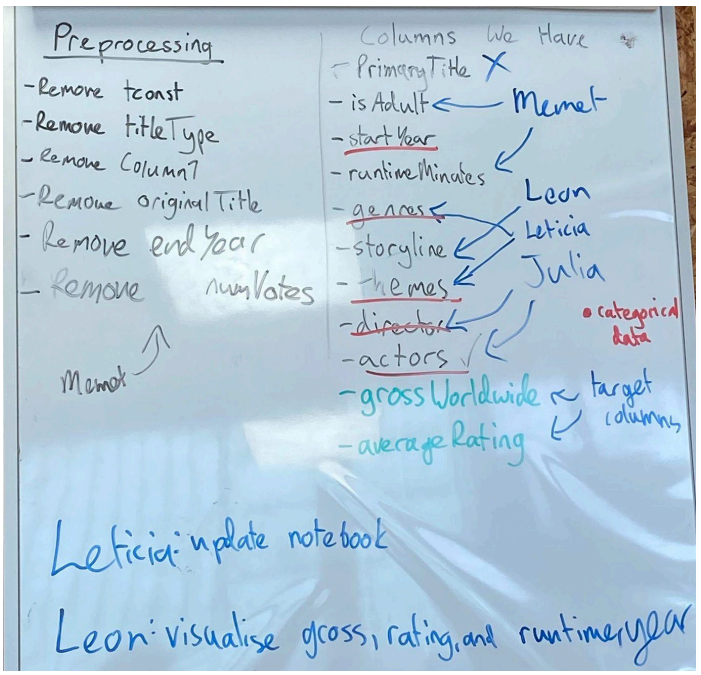
Columns we have:

- Primary Title
- isAdult
- StartYear
- RuntimeMinutes
- Genres
- Storyline
- Themes
- Directors
- Actors
- grossWorldwide
- averageRating

Conversation with Baran and Maria about blockbusters: we are rethinking about what we consider a blockbuster to be. This is about the rating and the gross income

Why is gross income missing??? - see this before it messes with out data (why only 20%).

Understand how we could quantify the qualitative data + names + categories t allow for the use of a neural network. - Maybe use the IMDb BERT.



Day 6- 10/06/2025

Plan:

- Meeting with Christ at 10:15
- Discuss task allocation within team
- Julia: actor and director EDA
- Mehmet: EDA
- Leon:
- Letti:

Day 7- 11/06/2025

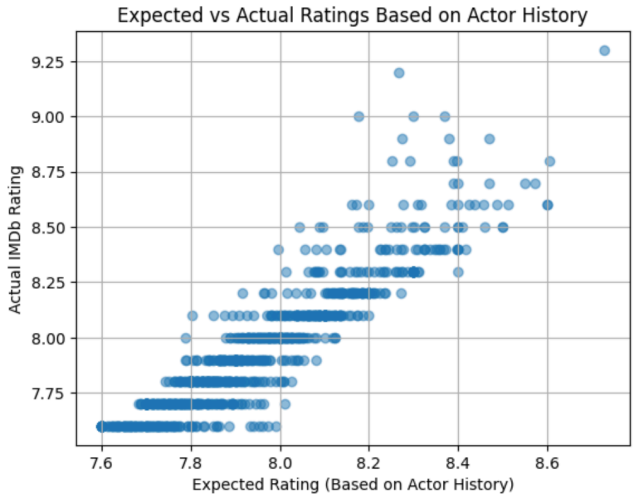
Plan:

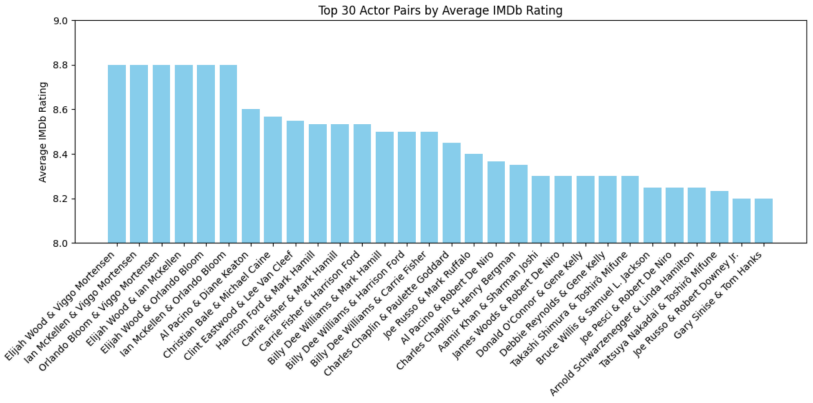
- Meeting at 10
- Check in with Maria 10:10-10:20
- Task Division
- Choosing between Regression and C;assification

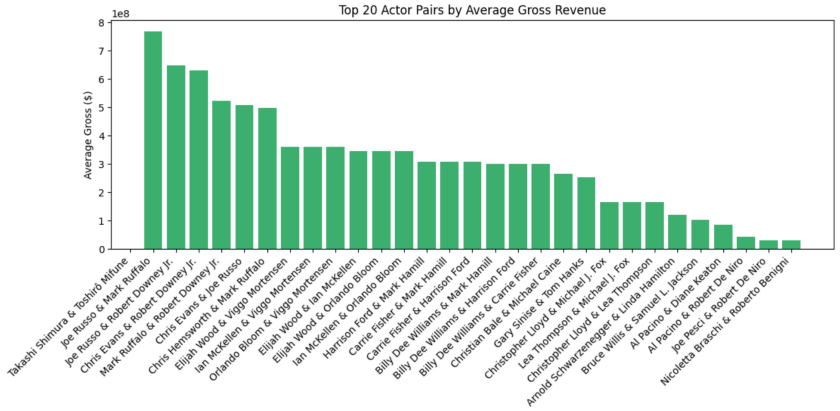
- We are using issues to divide tasks so we can also use them in commits
- We decided to focus on the presentation so that we have everything ready for friday and if we have extra time we can use that for our model and next steps

- Julia: finishing her EDA and committing it all. Decided to use number of votes to balance out bias in a few actor instances with high ratings.

$$\text{Efficiency} = \frac{\sum(\text{rating} \times \text{votes})}{\sum(\text{votes})}$$







Day 8- 12/06/2025

Plan:

- Meeting with Chris
- Julia add stuff to repo - chose which graphs to include in the presentation
- Lettie make read.me files for repo
- Everyone prepare their slide bit and time themselves
- Have a list of what you HAVE to say, and what you'd like to say
- How are we gonna structure the presentation flow-wise?
- Make a list of basic questions we have to answer - what people might ask us

Presentation Plan:

- Leon: introduction, previous dataset, web scraping
- Julia: EDA
- Mehmet: EDA, Model Choice
- Lettie: Future Work, Model Choice

Day 9- 13/06/2025

Interim Presentation Day!!

- Morning meet
- Rehearse with julia before she leaves
- Continue work on it and chose which graphs are staying (which are come impactful)
- Presentations
- Fix repo

-technical problems

Day 10- 16/06/2025

Plan:

- Meeting
- Presentation feedback
- Brainstorming next tasks
- Finishing up past tasks
- Research on SHAP values and lightgbm

Presentation Feedback: Points to work on for the next one
Minimal writing - only headlines
Give the most basics
Relax the boundary
Talk more about data cleaning - remove small movies
Data Visualization - color scheme - let the plot do the talking
Cut out names or colors
When are they necessary
Simplest possible headline information
HEADLINES (Go back to basics more!!)
Why , is it gonna add something
Manual evaluation for credibility
Question every time you do something

Good
- Flowchart
- Working through the RQ (imagined scenario)
- Target variable categories is a good starting point (table)
- A goal in the EDA overview is very good
-GOAL!! - why are we doing this

Bad
- Too much text per slide
- The dataset synopsis and functions could be simpler
- Less in the target variable
- Huge class imbalance, maybe relax the boundaries
- Should have spoken more about data cleaning
- Data visualisations were not visible, or it was mentioned that the patterns are what matter, not the names
- Explain why we are using the methods we are using (PCA)
- Specifically for the sentiment analysis, make sure to check what we think the feeling is vs what the model says
- Went over time, we could trim it

Day 11- 17/06/2025

Plan:
- Meeting Maria in the morning
- New tasks division
- Memet - continue PCA + feature engineering
- Letti - LightGBM

- Julia: Research on best hyperparameters for LightGBM + Help Lettie
- Leon: Start working on the model

Day 12- 18/06/2025

Plan:
- Continue/finish assigned tasks
- Leon: Preprocess the directors. Finished. Created two new columns, one for average gross worldwide and average average rating. Removed movies that have made less than 1000 USD.
- Lettie: SHAP values added to an exploratory LightGBM model, preprocessing -> giving each actor a numerical value.
- Julia: arrived midway, finished off the hyperparameter exploration, reviewed eda so leon can review it.

Train tests split chronologically, we are predicting the future from the past. Test set is 2024, 2025.

Justify choices and resources
Reproducibility
Where they can loosen things up

Leons Notes on Repo Feedback:
- A bit more info about what the project is about in Readme
- Be more specific in issues, frame as mini experiment
- Certain (leon's) commits too long, if using commas in commit = writing too much
- Don't commit errors
- Folder for cleaned dataset
- Gradient colour to graphs
- Too many branches, clean them up
- No need to customise merge branch commits
- Use more section subheadings in eda notebook
- DON'T COMMIT AT 2AM

Day 13- 19/06/2025

Plan:
- Meeting with Maria 10:15
- Group Meeting to assign tasks and plan for next week
- What has to be done by tomorrow and by next week.

- Allocate time to tasks. What day do we want to have everything done by?
    - 2 days to work on the presentation
    - Today we will start working on the model, tomorrow we all work on the model.
    - Have the model ready latest by monday.
    - Start Evaluation
    - Thursday and Friday presentations only! -> everything done on wednesday!
- Create Issues for today -> Julia and Lettie
- Catch up Memet on tasks for today
- Work and reconvene a bit before checkout.

Julia: Start presentation, Describe Model, start model part of the notebook. Put things there as If we have the preprocessed columns.
Lettie: Create function for the Model
Leon: review Julia's eda, change the df to movie_dataset, make actors work, organize, create headings and subheadings for eda and model.

Notes on Leon's feedback:
- too much code, too little explanation
- Whenever you print something, do it as an f-string, so there is context given
- Comment explaining the results
- Plot and then discuss
- Subheadings
- Gradient in graphs
- Describe my function
- Discuss the outliers
- Explain correlations
- Fix comments, be more specific
- Talk about the graph
- Short description of the function
- Say it
- Look at the distribution of actors after using the function, don't forget to explain the function.
- Remove the weight bit from the score function
- Why efficiency
- Be specific, give numbers and names.
- Directors explain that there is a cutoff and it is not normal.
- Talk about distribution
- Write explanation

Frequency encoding and tinting encoding for the genres

Day 14- 20/06/2025

Plan:

- Train and test our model
- We all work together to make the model
- We all test it together
- Solving issues
- Julia fix eda

Advice from Chris

- Test it out as two regression models - point of comparison
- Maybe there's an optimal point - confidence in the selection
- Different thresholds for the gross

Day 15- 23/06/2025

Plan:

- Morning Check in
- Presentation Vera - Bias and Stereotypes in AI
- Check in with Maria
- Julia and Leon deciding on tasks
- Leon: split the preprocessing and model notebooks, work on regression model
- Julia: work on model notebook comments, research on cross validation and hyperparameters, initiate discussion section - understand document
- Memet: improving model_ splitting and deleting certain columns, debug MSE issue, reduce mse.
- Lettie: Update LightGBM regression model and start classification model

To do:

- Clean repo
- Improve model
- Explain + Discuss results
- Bad results can also be useful

Day 16- 24/06/2025

Plan:

- Meeting online
- Task Division

Day 17- 25/06/2025

Plan:

- Meeting
- Leticia: Presentation
- Leon: Create new notebook and add some preliminary info, add descriptions and make graphs better
- Julia: Gridsearch for classification, presentation, results and discussion notebook, fix eda.
- Memet: Confusion matrix, try to run model on 100M

True positive worth more

Day 18- 26/06/2025

-Finish results_discussion notebook
-Clean up notebooks
-Update READMEs
-Presentation
-Ideal blockbuster?
-Actor + director is proxy for budget of movie -> add to discussion

LEON: look over results_discussion notebook, finish preprocessing notebook
LETICIA: READMEs, model notebook
MEMET: results_discussion notebook
JULIA: fix randomsearch bug, comments for eda part