A GUIDING DOCUMENT FOR ANALYZING THE MOVIES DATASET

# INTRODUCTION

This is a case study that grants us an opportunity to choose a question that is of

interest to us and figure out answers using data skills. We’ll go through the following

steps of data analysis: **ask**, **prepare**, **process**, **analyze**, **share**, and **act**.

# SCENARIO

We are junior data analysts who have tasked ourselves with leading a project involving

everything from defining the business task all the way through presenting our

data-driven recommendations. We are going to choose a topic, ask the right questions,

identify a fresh dataset and ensure its integrity, conduct analysis, create compelling

data visualizations, and prepare a presentation.

# ASK

Five questions will guide our case study:

1. ***What type of company does our client represent, and what are they asking us to accomplish?***  
   Our clients are movie enthusiasts around the world as well as, perhaps, movie makers who might be interested in the factors associated with the success of movies at the box office.
2. ***What are the key factors involved in the business task we are investigating?***  
   Key factors for predicting the success of a movie might include all the data we can find about the movie like the cast, directors, producers, movie studio, budget, genre, and other related data points.
3. ***What type of data will be appropriate for our analysis?***  
   Data about different attributes of movies from a source that is closely related to the movie industry and is reliable for the purposes of inference about future movies.
4. ***Where will we obtain that data?***  
   We have obtained the data from the user Rounak Banik on Kaggle(https://www.kaggle.com/datasets/rounakbanik/the-movies-dataset) which is credited as coming from The Movie Database(https://developer.themoviedb.org/docs). Perhaps, in a later iteration of this work, we may derive the data directly from the source.
5. ***Who is our audience, and what materials will help us present to them effectively?***  
   Our audience consists of general movie enthusiasts at large and perhaps, in particular, stakeholders(like those who make the movies themselves) in the movie industry to whom data-backed insights into box office success would be useful. Presenting insights to such people effectively would involve publishing the insights where the audience is. Movie enthusiasts may be found in movie fan sites and groups on social media. Major stakeholders like movie directors, actors and studio heads may be harder to get a hold of. However, our insights could get to them via personal communication or through virality on social media.

We will produce a report with the following deliverables:

1. A clear statement of the business task we have selected to investigate
2. A description of all data sources used
3. Documentation of any cleaning or manipulation of data
4. A summary of our analysis
5. Supporting visualizations and key findings
6. Based on what we discover, a list of additional deliverables we think would be helpful to include for further exploration
7. Our top high-level insights based on our analysis

We’ll use the following Case Study Roadmap as a guide:

## CASE STUDY ROADMAP — ASK

### GUIDING QUESTIONS

1. ***What topic are we exploring?***  
   We intend to explore the factors surrounding movie box office success and failure
2. ***What is the problem we are trying to solve?***  
   At the moment, we are interested in answering only one question: How well can we predict box office success or failure based on data about the movie?
3. ***What metrics will we use to measure our data to achieve our objective?***  
   We could use some kind of "distance measure" to measure the difference between **predicted** box office revenues and **actual** box office revenues. Possibly suitable measures might be **mean squared error** or **absolute error**. A suitable metric would help us measure how successful our results are.
4. ***Who are the stakeholders?***  
   All movie enthusiasts and leaders in the movie industry.
5. ***Who is your audience?***  
   The same as our stakeholders.
6. ***How can our insights help our client make decisions?***  
   Major players in the movie making industry can make better decisions about the kinds of movies that make more money. And for ordinary movie enthusiasts, this is one more piece of information for them to add to their arsenal of knowledge.

### DELIVERABLES

***A clear statement of the business task we have selected to investigate***: We are

going to investigate the extent to which the box office revenue of movies can be

predicted based on other data surrounding it.

# PREPARE

For the prepare phase, these questions will guide us

1. ***Where is your data located?***The final version of the cleaned dataset will be hosted on our github repository. The original is located on Kaggle at <https://www.kaggle.com/datasets/rounakbanik/the-movies-dataset>.
2. ***How is the data organized?***  
   The data is organized into multiple files:
   1. **movies\_metadata.csv:** The main Movies Metadata file. Contains information on 45,000 movies featured in the Full MovieLens dataset. Features include posters, backdrops, budget, revenue, release dates, languages, production countries and companies.
   2. **keywords.csv:** Contains the movie plot keywords for our MovieLens movies. Available in the form of a stringified JSON Object.
   3. **credits.csv:** Consists of Cast and Crew Information for all our movies. Available in the form of a stringified JSON Object.
   4. **links.csv:** The file that contains the TMDB and IMDB IDs of all the movies featured in the Full MovieLens dataset.
   5. **links\_small.csv:** Contains the TMDB and IMDB IDs of a small subset of 9,000 movies of the Full Dataset.
   6. **ratings\_small.csv:** The subset of 100,000 ratings from 700 users on 9,000 movies.
3. ***Are there issues with bias or credibility in this data?*** [***Does your data ROCCC***](https://www.coursera.org/learn/data-preparation/lecture/lHirM/what-is-bad-data)?  
   ***Reliable***: Data was retrieved by Rounak from grouplens.org from a paper published by F. Maxwell Harper and Joseph A. Konstan(2015)  
   ***Original***: Although we aren’t getting this from the original source, we shall assume this to be reliable until we find issues with this. We shall use the original in a later project  
   ***Comprehensive***: Yes  
   ***Current***: Not very recent, but it’s current enough to give us results  
   ***Cited***: Yes, from GroupLens.org
4. ***How are you addressing licensing, privacy, security, and accessibility?***

The data is under the CCO: Public Domain license

(<https://creativecommons.org/publicdomain/zero/1.0/>) which gives us the right to

copy, modify, distribute our work, even for commercial purposes, all without asking

Permission.

1. ***How did you verify the data’s integrity?***   
   To be answered later
2. ***How does it help you answer your question?***  
   The data contains important variables like genre, cast, director, crew, and others that will help answer the questions we have
3. ***Are there any problems with the data?***  
   To be answered later.

# PROCESS

1. ***What tools are you choosing and why?***  
   Excel to view the data and Python to process the data because it has a plethora of libraries and packages for that purpose
2. ***Have you ensured your data’s integrity?***

●

What steps have you taken to ensure that your data is clean?

●

How can you verify that your data is clean and ready to analyze?

●

Have you documented your cleaning process so you can review and share those results?