# 42184 Data Science for Mobility
# 42577 Introduction to Business Analytics course
# Challenge statement

Welcome to this year's challenge!

The topic this year is *Movie industry*. After two years of lockdown, we all became movie experts, didn't we? Wouldn't be nice to learn something new about this business and understand which factors contribute most to a movie's success?

In this project, we invite you to appropriate this question and use your best Data Sciences skills to explore what makes a movie successful. We do not expect you to crack the show business with a single Data Sciences project. Instead, we want you to address the mandatory questions (below) but also seek yourself for new questions, new data, new insights.

You have access to the "Comprehensive IMDb Data" dataset[1], where you can find 22 feature-engineered indicators extracted from IMDb, for more than 7000 movies. The dataset includes movies released in the USA between 1980-2020. The dataset was designed to understand which features contribute most to a movie's box office success. It is in itself an interesting Data Sciences exploration and a very good starting point for your project.

## Project structure

The project has three components:

- *Prediction challenge* - where all groups need to address the same problem (30%)
- *Exploratory component* - where each group is invited to choose their own research question and explore the data accordingly (40%)
- *Report (30%) -* Each group should deliver one or more jupyter-notebooks, that should be self-explanatory in each step (or block). This will function as a report, so it should have introduction and conclusions, besides the individual comments and reflections. However, there are some rules about the structure of the report, which should follow the 4-part outline showed below:

  Section 1: Introduction + Data analysis and visualization (Max 1 page)

  Section 2: Prediction Challenge (Max 1 page)

  Section 3: Exploratory Component (Max 1 page)

  Section 4: Conclusions (Max ½ page)

[1] https://www.kaggle.com/austinwolff/comprehensive-imdb-data

The page limit refers to **code and text**. Figures and results (e.g. data visualization, numerical outputs) should not be considered. To check the length of your report, you can export it as a PDF in A4 format. Also, you can add as many appendices as you want (we do not want you to limit yourself). But the report should be self-explanatory, appendices should only contain what you cannot include in the report.

## Introduction to the data

Figure 1 shows the variables you will have in this dataset. The data is provided as a CSV file. Notice that the variables require a lot of treatment in order to be usable (e.g. NaNs, categorical, different scales, IDs).

| | 0 | 1 | 2 |
|---|---|---|---|
| **titleId** | tt0081505 | tt0118460 | tt0118460 |
| **title** | The Shining | The Shining | The Shining |
| **rating** | R | R | R |
| **region** | US | US | US |
| **genre** | Drama | Drama | Drama |
| **released** | 1980-06-13 | 1980-06-13 | 1980-06-13 |
| **year** | 1980 | 1980 | 1980 |
| **month** | June | June | June |
| **day** | 13 | 13 | 13 |
| **score** | 8.4 | 8.4 | 8.4 |
| **director** | Stanley Kubrick | Stanley Kubrick | Stanley Kubrick |
| **writer** | Stephen King | Stephen King | Stephen King |
| **star** | Jack Nicholson | Jack Nicholson | Jack Nicholson |
| **country** | United Kingdom | United Kingdom | United Kingdom |
| **budget** | 1.9e+07 | 1.9e+07 | 1.9e+07 |
| **gross** | 4.69988e+07 | 4.69988e+07 | 4.69988e+07 |
| **company** | Warner Bros. | Warner Bros. | Warner Bros. |
| **runtime** | 146 | 146 | 146 |
| **category** | actor | actor | actress |
| **nconst** | nm0000197 | nm0001836 | nm0308371 |
| **primaryName** | Jack Nicholson | Steven Weber | Cynthia Garris |
| **knownForTitles** | tt0407887,tt0119822,tt0073486,tt0071315 | tt0118460,tt0112896,tt0098948,tt0105414 | tt0108941,tt0105428,tt0118460,tt0094919 t |

Figure 1. *Dataframe view*

The *prediction challenge* consists of two parts. You are expected to predict both **the expected revenue (gross)** and the **IMDb score (score)** of a movie, conditioned on any other variable you choose.

- Part 1 – The objective is to predict the expected revenue of a movie. The training set will correspond to 75% of the dataset and test set will be 25%. You can shuffle the

data. You can use any sklearn regression or classification model you want, including those not taught in the class. As a benchmark, we expect you to be able to predict the test set with an $R^2$ at least 0.6 (or an f1 score at 0.7).

- Part 2 – Same as the previous section but you need to predict the IMDb score. The idea is for you to experiment (and discuss in the group) with the concept of generalizability/transferability Would the model be the same? What about the features? You can use any sklearn regression or classification model you want, including those not taught in the class. Explain the differences with respect to the previous case (in terms of $R^2$ / f1).

In both part 1 and part 2, if you want to use a development set, you need to extract it from the train set, i.e. you should not change the test set as above proposed. Alternatively, you can also try to use old movies to predict new movies (e.g. predict the revenue of movies from 2010). In this case, what do you expect to happen (again, in terms of metrics)?

In the *exploratory component*, each group needs to address at least one new research question. Here, we expect you to formulate your own question, follow the data sciences cycle. The project will be positively valued with one or more of the following extensions:

- Extension of the dataset (preferably using Python APIs) with other relevant data (more movies, more features, ..)
- Generation and analysis of insightful visualizations;
- Usage of different techniques with respect to those in Part 1 (e.g. dimensionality reduction, clustering, classification, time series)

Some example research questions:

- How the success of a movie is related to the age/career of the leading actor?
- How do older movies influence the new ones? (e.g. did *The Shining* start a new trend in the *Drama* genre?)
- Consider other aspects, such as social media, Netflix, and changes in the movie industry (Streaming? Piracy?)
- Relationships with societal trends (e.g. happiness indexes, global warming)

**Note:** The ordering of tasks we mention is **not** mandatory. In other words, if you prefer to start with the exploratory component, and then go to the prediction challenge, this is very acceptable. You can mention that in the report (or invert Section 2 and 3).

**Evaluation**

The evaluation of the report will be based on the following criteria:

- Clarity - self-explanatory nature of the notebooks
- Thoroughness - Each research question deserves to be explored to the right amount of depth
- Insightfulness - It's important to go beyond the surface of the conclusions

- Honesty - While it's fine to use others' code (as a starting point), these shouldn't generally be the actual deliverable **and** the appropriate ethical practice is to **always** reference the source of that code you used.

**Rules**

- Each group should consist of 4 students. Exceptions are allowed for other forms, but only with strong justification.
- The submission of the project shall be a zip file with all the notebooks. This zip file should contain the surnames of the group members (for example, for Pablo, Anders, Suarez, and Mila, it should be Pablo_Anders_Mila.zip).
- At the end of the report, there must be a section where **individual contributions are clearly clarified**. In case of doubts on individual contributions or authenticity of the report, the teachers will call the group for an oral defence. This section should **not be part of the page counts.**
- Meeting the deadlines for the milestones is important, including for non-evaluated milestones. A penalty of 10% is given for each extra day of delay

**Report**

The report must be in the form of a jupyter notebook. The structure should be the one describes in page 1.

**Important dates**

- October 11 – Announcement of this challenge statement
- October 25 – Communication of group members (to camara@dtu.dk and guica@dtu.dk)
- November 15 – Descriptive statistics notebook – this notebook should present preliminary analysis on the dataset, and other datasets obtained by the group, including data preparation, data cleaning, initial analysis of patterns and insights from the data. Submit through CampusNet
- November 29 – Final submission – all materials, including report notebook. Submit through CampusNet