**Basic Python: Project**

In this project, you will work with data from the entertainment industry. You will study a dataset with records on movies and shows. The research will focus on the “Golden Age” of television, which began in 1999 with the release of *The Sopranos* and is still ongoing.

The aim of this project is to investigate how the number of votes a title receives from IMDb users impacts its ratings. The assumption is that highly-rated shows (we will focus on TV shows, ignoring movies) released during the “Golden Age” of television also have the most votes.

This project is similar to the tasks you will be getting in your job as a data professional. Many business decisions are initially born as assumptions, your contribution as an expert in the data domain is to answer the question “Did the assumption formulated before the study appear to be true?”

**Description of the data**

The data is stored in the file /datasets/movies\_and\_shows.csv.

Description of columns:

* 'name' — first and last name of actor (director)
* 'character' — character played (for actors)
* 'role ' — the person’s contribution to the title (it can be in the capacity either of actor or director)
* 'title ' — title of movie (show)
* 'type ' — show or movie
* 'genres' — list of genres under which the movie (show) falls
* 'release\_year' — year when the movie (show) was released
* 'imdb\_score' — score on IMDb
* 'imdb\_votes' — votes on IMDb

**Instructions for completing the project**

A template of a Jupyter Notebook is provided for you where you can write your code, along with descriptions of your analysis. To complete the project, you must fill in every code cell in the template and edit markdown cells where the template asks you to explain your results.

In later sprints, the work will be more independent. This time, to help you get started, we’ll provide a notebook containing:

* The project structure
* Guidelines for analysis
* Conclusions

Your project will consist of three stages:

Stage 1: Data overview. The notebook has premade cells with instructions for what kind of code to write, as well as text blocks where you can write your observations.

Stage 2: Data preprocessing. In this stage, you’ll clean up column names and address missing and duplicate values. Follow the outline provided in the notebook and be sure to write about your observations at the end of this section.

Stage 3: Data analysis. This is the meat of your project. Walk through the coding steps needed to evaluate the assumption and comment on your results in the appropriate blocks. Finally, summarize the overall project in the “Conclusion” section which is equal to “Presenting Results” stage we mentioned in the lessons.