**Project-Based Course Overview**

Welcome to **Project: Analyze Box Office Data with Seaborn and Python**!This is a project-based course which should take approximately 1.5 hours to complete. Before diving into the project, please take a look at the course objectives and structure:

**Course Objectives**

In this course, we are going to focus on two learning objectives:

1. Produce data visualizations with Seaborn.
2. Apply graphical techniques used in exploratory data analysis (EDA).

By the end of this course, you will be able to generate publication-quality graphs using Seaborn and Python to analyze data.

**Course Structure**

This course is divided into 3 parts:

1. Course Overview: This introductory reading material.
2. Analyze Box Office Data with Seaborn and Python Project: This is the hands on project that we will work on in Rhyme.
3. Graded Quiz: This is the final assignment that you need to pass in order to successfully complete the course and earn a Course Certificate.

**Project Structure**

The hands on project on Analyzing Box Office Data with Seaborn and Python is divided into the following tasks:

**Task 1: Data Loading and Exploration**

* In this task, you are introduced to the project and learning outcomes.
* Once you are familiarized with the Rhyme interface, we begin working in [Jupyter Notebooks](https://jupyter.org/), a web-based interactive computational environment for creating notebook documents.
* Next, we will import essential libraries such as [NumPy](https://numpy.org/), [pandas](https://pandas.pydata.org/), [Seaborn](https://seaborn.pydata.org/), and [matplotlib](https://matplotlib.org/).
* Lastly, load the data sets into memory using pandas and display the first few entries. The TMDB data set contains 7400 movies and a variety of metadata.

**Task 2: Visualizing the Target Distribution**

* The TMDB Box Office Prediction [database](https://www.kaggle.com/c/tmdb-box-office-prediction/data) was initially released so that teams could treat it like a regression problem and predict the worldwide box office revenue of 4400 movies provided in the test file.
* We will use this feature rich database for data analysis and data visualization in this project.
* In the next project, we will use the same database for feature engineering and feature visualization.
* To proceed to feature analysis, we first need to visualize the target. Using Seaborn's [distplot](https://seaborn.pydata.org/generated/seaborn.distplot.html) function, we will plot the distribution of movie revenues. To illustrate the skew, we will also plot the distribution of revenue on a logarithmic scale using [np.log1p](https://docs.scipy.org/doc/numpy/reference/generated/numpy.log1p.html).

**Task 3: Comparing Film Revenue to Budget**

* Just as we did in Task 2, we will create two subplots: one for the distribution of the budget; the other to plot the log budget against log revenue.
* Perform the log transformation to make the distributions more manageable.

**Task 4: Do Official Homepages Impact Revenue?**

* Use the data to demonstrate the effect of an official homepage on movie revenue.
* Create appropriate features to indicate the presence of official movie homepages. We can use Seaborn's [catplot](https://seaborn.pydata.org/generated/seaborn.catplot.html) function to plot the revenues for movies with and without homepages using our newly created features.

**Task 5: Distribution of Languages across Films**

* As Hollywood is known globally, we would expect the data to support our intuition that English movies generate the highest revenue.
* In this task, we will use box plots to test this hypothesis. We might be surprised that our intuitions about the world may not always align.

**Task 6: Common Words in Film Titles and Descriptions**

* Identify trends across movie titles and descriptions and their effect on revenue.
* Generate word clouds for movie titles and descriptions. Word Cloud is a data visualization technique used for representing text data in which the size of each word indicates its frequency or importance.

**Task 7: How do Film Descriptions Impact Revenue?**

* To identify which words have the highest impact on revenue, we tokenize and vectorize movie titles and descriptions, fit a linear regression model to it, and use [ELI5](https://eli5.readthedocs.io/en/latest/overview.html) to display the high impact words.

**Meet the Instructor**

I'm Snehan Kekre, a machine learning and data science instructor at Rhyme. I will graduate in 2021 with a BSc in Computer Science and Artificial Intelligence from Minerva Schools at KGI, based in San Francisco. My passion to grow the ML and AI community drives me to create accessible, skills-based, machine learning projects at Rhyme.