Welcome to **Exploratory Data Analysis With Python and Pandas**. This is a project-based course which should take approximately 2 hours to finish. 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 **three** learning objectives:

1. Apply practical Exploratory Data Analysis (EDA) techniques on any tabular dataset using Python.
2. Produce data visualizations using Seaborn and Matplotlib
3. Identify and handle duplicate and missing data

By the end of the course, you will be able to dive in and explore a sales dataset from a supermarket retail chain, and apply those skills to any tabular dataset using Python. You will learn to use external Python packages such as Pandas, Numpy, Seaborn, Matplotlib and more to retrieve valuable information from your dataset.

**Course Structure**

This course is divided into 3 parts:

1. Course Overview: This introductory reading material.
2. **Exploratory Data Analysis With Python and Pandas:** 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 finish the course successfully.

**Project Structure**

The hands on project on **Exploratory Data Analysis With Python and Pandas** is divided into following tasks:

**Task 1: Initial Data Exploration**

* In this task, we are introduced to the project and learning outcomes.
* Once we are familiarized with the Rhyme interface, we begin working in [Jupyter Notebooks](https://jupyter.org/" \t "_blank), a web-based interactive computational environment for creating notebook documents.
* Next, we will import essential libraries such as NumPy, Pandas, Seaborn, Matplotlib and so on.
* We use Pandas to read in the data, get a brief glimpse of the first few rows, and calculate some quick summary statistics of the numeric columns.

**Task 2: Univariate Analysis**

* In this task, we conduct univariate analysis on both continuous and categorical variables.
* We first plot the distribution of customer ratings with seaborn and also overlay the mean, 25th and 75th percentile quantiles calculated using Numpy.
* We then use Pandas' .hist() method to plot the distribution for all numeric variables.
* Using Seaborn's .countplot() method, we see the frequency distribution of 'Branch' and 'Payment' which are categorical variables.

**Task 3: Bivariate Analysis**

* In this task, we conduct bivariate analysis on both continuous and categorical variables.
* We use Seaborn to plot scatterplots and regression plots to identify the relationship between customer rating and gross income.
* Additionally, we use Seaborn to plot a boxplot to check the difference in aggregate sales figures between the three branches of supermarkets, and to compare sales patterns between men and women.
* We plot a time series graph to check for trends in gross income over a period of three months.

**Task 4: Dealing With Duplicate Rows and Missing Values**

* In this task, we identify and deal with duplicate rows and missing values in our dataset.
* We calculate the number of duplicate rows and delete them using Pandas.
* We then do the same with missing values, but instead of deleting those rows, we replace missing values by the means of their respective columns.
* We explore our dataset using Pandas Profiler to see how we can automate a lot of exploratory data analysis given certain conditions are met.

**Task 5: Correlation Analysis**

* In this task, we conduct correlation analysis on the numeric variables in our dataset.
* We use Numpy to calculate the correlation between two numeric variables.
* We then use pandas to calculate a correlation matrix to show all pairwise correlations of numeric variables.
* Finally, we use seaborn to plot the calculated correlation matrix as a heatmap that is easily interpretable.