**Course Objectives:**

In this course, you will build and evaluate a simple linear regression model using Python. You will employ the scikit-learn module for calculating the linear regression, while using pandas for data management, and seaborn for plotting. You will be working with the very popular Advertising data set to predict sales revenue based on advertising spending through mediums such as TV, radio, and newspaper.

By the end of this project, you will be able to

* Explain the core ideas of linear regression to technical and non-technical audience
* Build a simple linear regression model in Python with scikit-learn
* Employ Exploratory Data Analysis (EDA) to small data sets with seaborn and pandas

**Rhyme Platform:**

This course runs on Coursera's hands-on platform called Rhyme. On Rhyme, you do projects in a hands-on manner in your browser. You will get instant access to pre-configured cloud desktops that have all the software and data you will need. So, you can just focus on the learning. For this project, this means instant access to a cloud desktop with Jupyter and Python 3.7 with all the necessary libraries pre-installed.

**Course Structure:**

**Task 1: Introduction and Overview**

You will be introduced to the Rhyme interface and the learning environment. You will be provided with a cloud desktop with Jupyter Notebooks and all the software you will need to complete the project. Jupyter Notebooks are very popular with Data Science and Machine Learning Engineers as one can write code in cells and use other cells for documentation. We will also introduce the model we will be building as well the [Advertising dataset](http://www-bcf.usc.edu/~gareth/ISL/Advertising.csv) for this Project.

**Task 2: Loading the Data and Importing Libraries**

In this task, we will load the very popular [Advertising dataset](http://www-bcf.usc.edu/~gareth/ISL/Advertising.csv) about various costs incurred on advertising by different mediums such as through TV, radio, newspaper, and the sales for a particular product. Next, we will briefly explore the data to get some basic information about what we are going to be working with.

**Task 3: Removing the Index Column**

Cleaning and pre-processing data is a vital process in data analysis and machine learning. In this task, we will take a look at our imported data and ascertain what to remove and what to keep for further analysis.

**Task 4: Exploratory Data Analysis (EDA)**

It's good practice to first visualize the data before doing any analysis and model building. If the data is high dimensional, at least examine few slices using simple techniques like box plots. In this task, we will look at the distributions of our response variable , ***Sales***, and also looked at the distributions of our predictors, ***TV***, ***Radio***, and ***Newspaper***.

**Task 5: Relationship between Predictors and Response**

Here we visualize pairwise correlations between our predictors and response. We also create a heatmap of the corresponding correlation matrix. This is done so as to ascertain whether there exists any relationship between the sales revenue and the expenditure on the various advertisement channels.

**Task 6: Creating the Simple Linear Regression Model**

Linear regression is an approach for modeling the relationship between a scalar dependent response variable y and one or more predictors (or independent variables) denoted ***X***. The case where we have just one predictor is known as simple linear regression. In this task, we are going to use the *LinearRegression* estimator from *sklearn.model\_selection*to create our ordinary least squares (OLS) linear regression model. We will use *train\_test\_split*to split our data into training and test sets.

**Task 7: Evaluation and Model Parameters**

In this task we will use the *intercept\_* and *summary()* methods to produce human-readable output of our model coefficients and the OLS regression results.

**Task 8: Making Predictions with the Model**

Now that we have trained our OLS regression model on the training data, we will use *predict()*method to make predictions on the test data. Note that the predictions are made on data the model has never seen during train time.

**Task 9: Model Evaluation Metrics**

Evaluation metrics for classification problems, such as accuracy, are not useful regression problems. We need a metric designed to evaluate continuous values. In this task, we calculate and implement the three most common evaluation metrics for regression: Mean Absolute Error (**MAE**), Mean Squared Error (**MSE**), and Root Mean Squared Error (**RMSE**).