Introduction

We'll cover the following

- A Dataset and a Machine Learning Problem, What Should You Do?
  - Overview of the Main Steps:
    - 1. Exploratory Data Analysis
    - 2. Prepare the data for machine learning algorithms
    - 3. Transformation Pipelines in Scikit-Learn
    - 4. Assess Machine Learning Algorithms
    - 5. Fine-Tune Your Model
    - 6. Present the Solution
    - 7. Launch, Monitor, and Maintain the System
  - Important Preliminary Steps

A Dataset and a Machine Learning Problem, What Should You Do?

Say you have been recently hired as a Data Scientist to work on a project and you have been given some real estate data. How can you approach the problem in a **systematic and structured way** rather than ending up with a spaghetti code? What are the steps to follow?

In this section, we are going to **deconstruct** the main step needed to work on a ML project via a real end-to-end example with code. We are going to work with a challenge based on a **Kaggle Competition**.

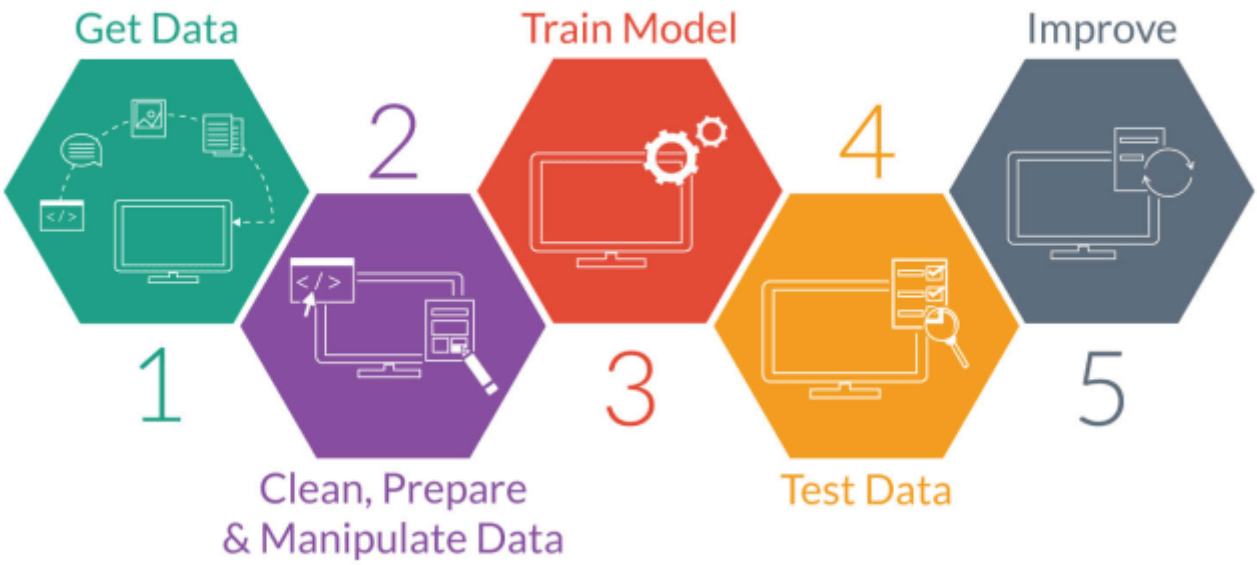
Overview of the Main Steps:

There isn't a golden approach that every data scientist must adopt that would work for every single project. However, there are some good practices and steps recommended by the sages of the field that one should keep in mind. For

this project, we are going to adopt an approach that is an adaptation of advice

collected from various books and articles on the subject (details in *"Study Material"*), and from personal experience.

Of course, this is to serve as a reference skeleton to guide you through your projects; you should add or delete steps based on the specific needs of your project.



Here are the main steps we are going to go through:

1. Exploratory Data Analysis

- Understand the data structure
- Discover and visualize the data to gain insights
    - Explore numerical attributes
    - Look for correlations among numerical attributes
    - Explore categorical attributes

2. Prepare the data for machine learning algorithms

- Deal with missing values
- Handle outliers
- Deal with correlated attributes
- Handle text and categorical attributes
- Feature scaling

3. Transformation Pipelines in Scikit-Learn

4. Assess Machine Learning Algorithms

4. Assess Machine Learning Algorithms #
    - Train and evaluate multiple models on the training set
    - Comparative analysis of the models and their errors
    - Evaluation Using Cross-Validation
5. Fine-Tune Your Model #
6. Present the Solution #
7. Launch, Monitor, and Maintain the System #

Important Preliminary Steps

Before jumping into coding and playing with the data, if you have been given a **real problem** to work on, your first question should be to ask your manager/the stakeholders/the owner of the project what exactly the **business objective** is. Just building a machine learning model is not the end goal, they are likely interested in the benefits from the solution, so get an understanding of how the company expects to use your work and benefit from it.

Why is this an important step for a real project?
Because having a clear understanding of the business objective will determine:

- how you frame the problem
- what algorithms you will select
- what performance measure you will use to evaluate your model
- the amount of effort you should spend tweaking your final model
- how you present your solution

What does **framing the problem** mean?
It means that first you need to understand if you are dealing with supervised, unsupervised, or Reinforcement Learning? Have you been given a classification task, a regression task, or something else?

Now without further ado, let's move on to the next lesson and ***learn by practicing***!