***Question :***

Explain how will you build the dataset. If we tell you the requirements.

***Answer :***

Building dataset completely depends on the problem which we trying to resolve.

There are multiple ways by which we can collect or build the dataset required for us.

As per the problem statement, First of all I will try to check if I can get the required data from below options.

1. **Kaggle** is one of the widely used platform which provides almost every kind of and every problem specific free datasets. We can easily download it and start working.
2. **Google dataset search** is a brilliant tool which google has launched for data scientists who are in search of data. We can search a dataset by keywords and it provides bunch of websites which provides that dataset or similar dataset.
3. There are other websites also which provides the raw data for us to work on.

e.g. Data.gov, [UCI Machine Learning Repository](http://archive.ics.uci.edu/ml/index.php), [Google Public Datasets](https://cloud.google.com/bigquery/public-data/) and many more.

1. Many people who already worked on the problem make the dataset available publicly (GitHub Repository). Even if not, we can try to get it by contacting and requesting them.

If I could not find the required data from above sources then I will go with below methods.

1. Web Scraping is one of best way to collect the data now a days. There many advantages of it. The biggest and most important advantage is we can get the latest data. Another thing is we can fetch data in the way we want. Any number of fields and as many number of records as we want we can get by scraping the website.
2. Another way to build the dataset is to create a synthetic data. If I have found a dataset from the above sources but it is not in expected format then in that case we can generate synthetic data as per our requirement using the already existing dataset.
3. Data augmentation is also one of the famous method to generate a artificial dataset. If I have a dataset which is very small in size or we need a huge amount of data to train our deep learning models, then we can make use of Data augmentation technique.