Great Expectations - Getting Started

To set up **Great Expectations (GX)** locally for validating a CSV dataset downloaded from Kaggle, follow these steps:

### **1. Install Great Expectations**

Ensure you have Python 3.6 or above installed. It's recommended to use a virtual environment to manage dependencies.

# Create and activate a virtual environment

python -m venv gx\_env

# On Windows

gx\_env\Scripts\activate.bat

# On Unix or MacOS

source gx\_env/bin/activate

# Install Great Expectations

pip install great\_expectations

### **2. Initialize a Great Expectations Project**

Initialize GX in your project directory:

great\_expectations init

This command creates a great\_expectations directory with the necessary configuration.

### **3. Add Your CSV Data**

Place your Kaggle CSV dataset into a directory within your project, for example, data/.

### 

### 

### **4. Configure a Datasource**

A Datasource tells GX where your data resides. Configure it to point to your local CSV files:

great\_expectations datasource new

When prompted:

* **Select the data source type:** 1. Files on a filesystem (for processing with Pandas or Spark)
* **Select the data connector:** 1. Pandas
* **Enter the path of the root directory where the data files are stored:** data/

This sets up a Datasource pointing to your CSV files.

### **5. Create an Expectation Suite**

An Expectation Suite is a collection of assertions about your data. To create one:

great\_expectations suite new

Follow the prompts to name your suite and select the Datasource you configured. GX will open a Jupyter notebook for you to define expectations.

### **6. Define Expectations**

In the Jupyter notebook:

* **Load a Batch of Data:** GX will provide code to load a sample of your data.
* **Add Expectations:** Use methods like expect\_column\_values\_to\_not\_be\_null or expect\_column\_values\_to\_be\_unique to define your data quality checks.

After adding your expectations, save the notebook.

### 

### 

### 

### **7. Validate Your Data**

To validate your data against the expectations:

great\_expectations checkpoint new my\_checkpoint

Follow the prompts to configure the checkpoint, specifying your Datasource, data asset, and Expectation Suite. Once configured, run:

great\_expectations checkpoint run my\_checkpoint

GX will validate your data and generate a Data Docs report, which you can view in your browser.

### **Additional Resources**

For more detailed guidance, consider the following resources:

* **DigitalOcean Tutorial:** [How To Test Your Data With Great Expectations](https://www.digitalocean.com/community/tutorials/how-to-test-your-data-with-great-expectations)
* **Qxf2 Blog:** [Data validation using Great Expectations with a real-world scenario](https://qxf2.com/blog/data-validation-great-expectations-real-example/)

These tutorials provide step-by-step instructions and examples to help you effectively set up and use Great Expectations for your data validation needs.