**Preprocessing Report: Environmental Site Potential (ESP) Vegetation Dataset**

**1. Data Source**

* **Dataset Name**: us\_140esp.csv
* **File Size**: 3.1 MB
* **Source**: [LANDFIRE Data Portal](https://landfire.gov)
* **Theme**: *Environmental Site Potential (ESP)* — used to represent the vegetation that could exist on a site based on environmental and physical conditions.

**2. Objective**

The goal was to preprocess the ESP vegetation CSV file to prepare it for wildfire risk classification modeling. This included cleaning, filtering, and standardizing vegetation classification labels and corresponding metadata (e.g., color codes, region names, etc.).

**3. Data Columns Overview**

The ESP file contains the following key fields:

* VALUE: Category code used in raster data.
* MZ: Map zone number.
* ZONE\_NAME: Geographical zone name.
* ESP\_2 and ESP\_Name: Vegetation type identifiers.
* ESP\_LF and ESPLF\_Name: LANDFIRE’s refined classification name.
* R, G, B: RGB integer color codes for raster visualization.
* Red, Green, Blue: Normalized color codes (0–1 range).

**4. Preprocessing Steps**

**a. Loading the Data**

* Read the CSV using pandas for efficient tabular processing.

**b. Handling Missing or NoData Values**

* Rows with VALUE = -9999 or ESPLF\_Name = 'NoData' were dropped.
* Any NaN values in key vegetation descriptor columns (ESPLF\_Name, ESP\_Name) were also excluded.

**c. Simplifying Vegetation Labels**

* Combined ESP\_Name and ESPLF\_Name to create a unified label for vegetation classification.
* Created a mapping from VALUE codes to vegetation types for raster translation.

**d. Color Normalization (if needed for visualization)**

* Ensured color data was preserved for raster styling if using GIS software.
* Optionally converted RGB (0–255) to normalized (0–1) formats where necessary.

**e. Filtered Output**

* Retained only columns required for the wildfire model (e.g., VALUE, ESPLF\_Name, ZONE\_NAME).
* Saved a clean version as processed\_vegetation.csv.