# K-Means Clustering on Pokémon Dataset

Vaibhav Chourasia

October 24, 2025

## 1. Introduction

Clustering is an unsupervised machine learning technique used to group data points based on similarity. The objective is to understand the underlying mathematics and computational workflow behind clustering algorithms while applying it to a real-world dataset — the Pokémon Legendary dataset.

The dataset provides a variety of attributes such as HP, attack, defense, and speed that allow us to cluster Pokémon based on their combat performance characteristics.

## 2. Why K-Means Instead of DBSCAN

Although both K-Means and DBSCAN are clustering algorithms, they differ in behaviour:

- K-Means is simple, fast, and effective when clusters are roughly spherical and of similar size.
- DBSCAN handles non-spherical and arbitrarily shaped clusters better, but it is sensitive to its parameters ($\varepsilon$ and MinPts) and struggles with varying densities.

Given that the Pokémon dataset contains structured numerical features (combat stats) that naturally form compact clusters (e.g., weak vs strong Pokémon), K-Means was chosen for its interpretability and stable performance.

## 3. Mathematical Derivation of K-Means

The K-Means algorithm partitions $n$ observations into $k$ clusters, minimizing the within-cluster sum of squared distances.

### Step 1: Initialization

Randomly select $k$ points as initial centroids:

$$C = \{c_1, c_2, ..., c_k\}$$

### Step 2: Cluster Assignment

Assign each point $x_i$ to the nearest centroid:

$$\text{Cluster}(x_i) = \arg\min_j \|x_i - c_j\|$$

where $\|x_i - c_j\|$ denotes Euclidean distance.

### Step 3: Centroid Update

Recalculate each centroid as the mean of all points assigned to it:

$$c_j = \frac{1}{n_j} \sum_{x_i \in C_j} x_i$$

where $n_j$ is the number of data points in cluster $C_j$.

### Step 4: Convergence

Repeat steps 2 and 3 until centroids stabilize:

$$\|C^{(t)} - C^{(t-1)}\| < \text{tolerance}$$

The algorithm minimizes the objective function:

$$J = \sum_{j=1}^{k} \sum_{x_i \in C_j} \|x_i - c_j\|^2$$

## 4. Dataset Description

The dataset used is the Pokémon Legendary Dataset. It contains both numeric and categorical attributes describing each Pokémon's physical and combat characteristics.

- Shape: (801, 14)
- Columns: pokedex_number, name, attack, defense, height_m, hp, percentage_male, sp_attack, sp_defense, speed, type, weight_kg, generation, is_legendary
- Legendary Distribution:
  - 0 (Non-Legendary): 731 Pokémon
  - 1 (Legendary): 70 Pokémon

### Features Used for Clustering

Only combat-relevant numerical features were selected:

hp, attack, defense, sp_attack, sp_defense, speed

# 5. Data Preprocessing

Before clustering, the features were normalized using **Min-Max Scaling** to ensure all variables contribute equally:

$$X' = \frac{X - X_{\min}}{X_{\max} - X_{\min}}$$

The resulting scaled values were within [0, 1].

**Columns used for clustering:**

[hp, attack, defense, sp_attack, sp_defense, speed]

**Minimum values:**

All columns: 0.0

**Maximum values:**

All columns: 1.0

**Training samples:** 640
**Validation samples:** 161

# 6. Results and Visualizations

K-Means was trained with $k = 3$ clusters. After convergence, the first 10 validation cluster assignments were:

$$[1, 0, 0, 0, 2, 1, 0, 1, 1, 0]$$

## Legendary vs Non-Legendary Distribution

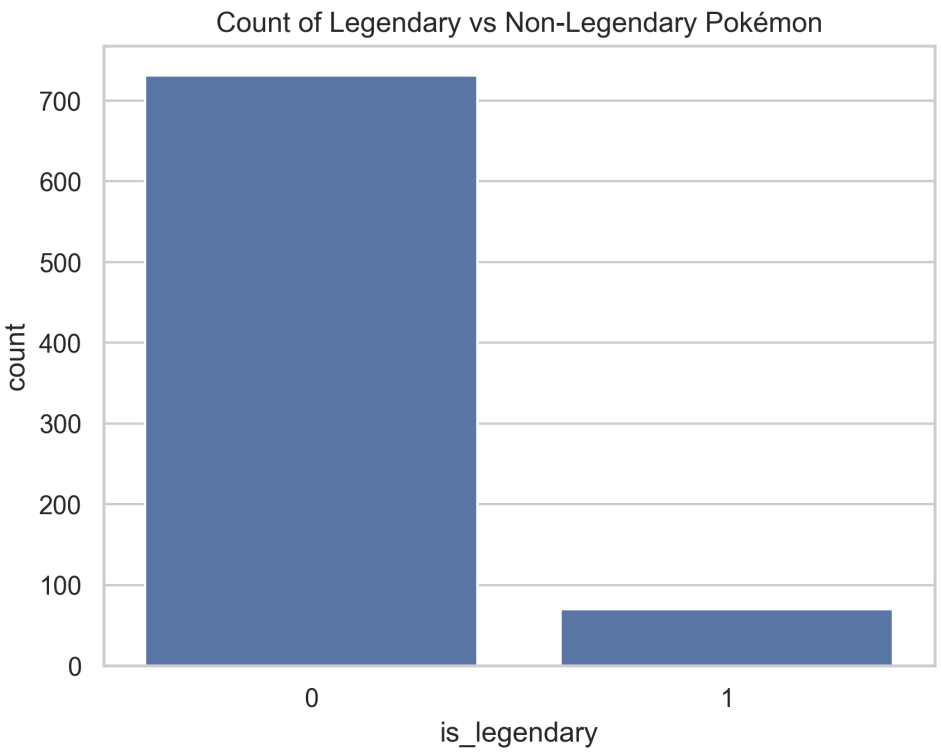


Figure 1: Count of Legendary vs Non-Legendary Pokémon

## K-Means Clusters (Attack vs HP)

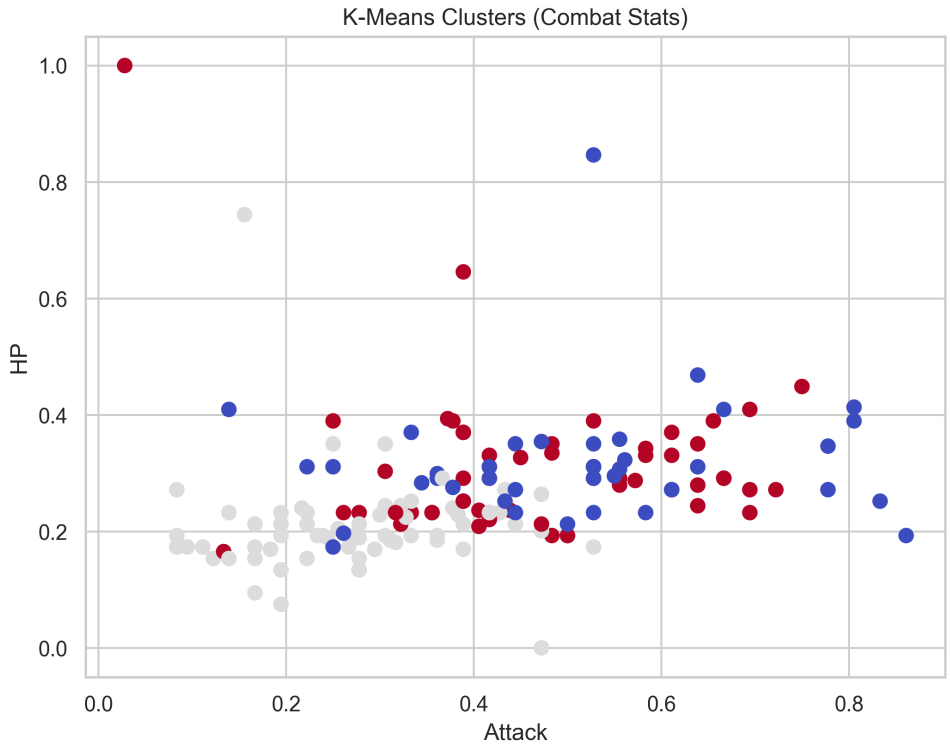


Figure 2: K-Means Clusters based on Attack and HP

### Clusters with Centroids

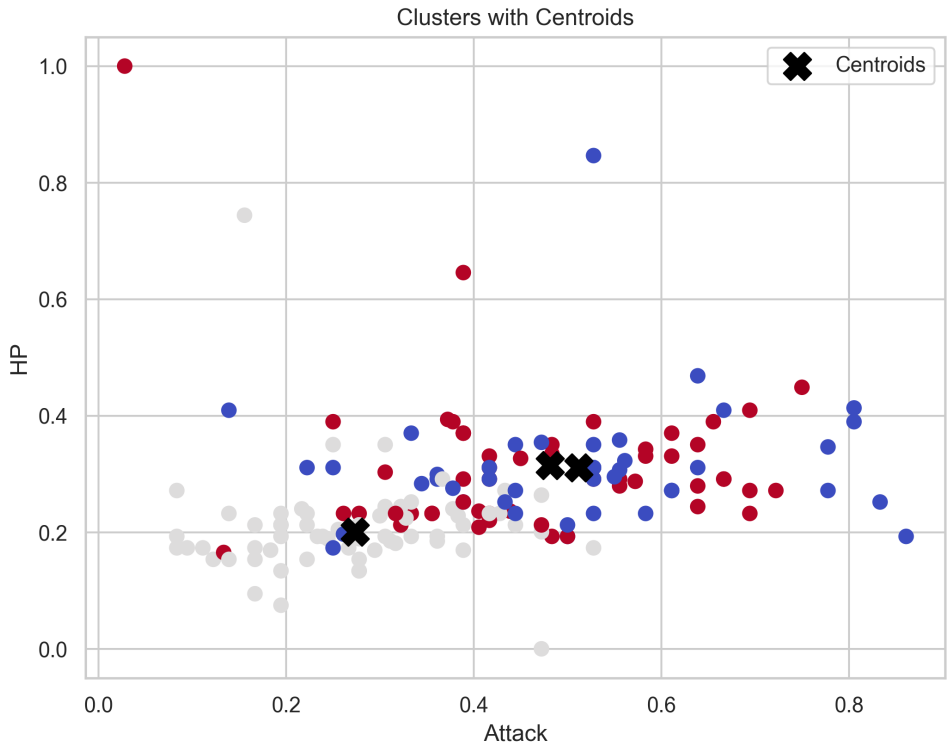


Figure 3: Cluster Visualization with Centroids (Attack vs HP)

### Evaluation Metric — Purity Score

The purity score measures how well cluster labels correspond to the actual "is_legendary" labels:

$$\text{Purity} = \frac{1}{N} \sum_{k} \max_{j} |C_k \cap T_j|$$

**Obtained Purity Score:** 0.9068

## 7. Discussion and Inference

The K-Means clustering from scratch successfully identified meaningful groupings within the Pokémon dataset. The three clusters correspond broadly to different strength levels of Pokémon, with Legendary ones forming smaller, distinct groups.

A high purity score (0.9068) confirms that the clusters align well with true labels, validating both the normalization and implementation.

## 8. Conclusion

This project demonstrates that K-Means can achieve high-quality clustering results. By scaling features, initializing centroids carefully, and updating iteratively, the algorithm accurately grouped Pokémon based on their combat characteristics.