Cluster Analysis and K-Medians

MATH/CMPT 370

Amanda Landi

March 9, 2017

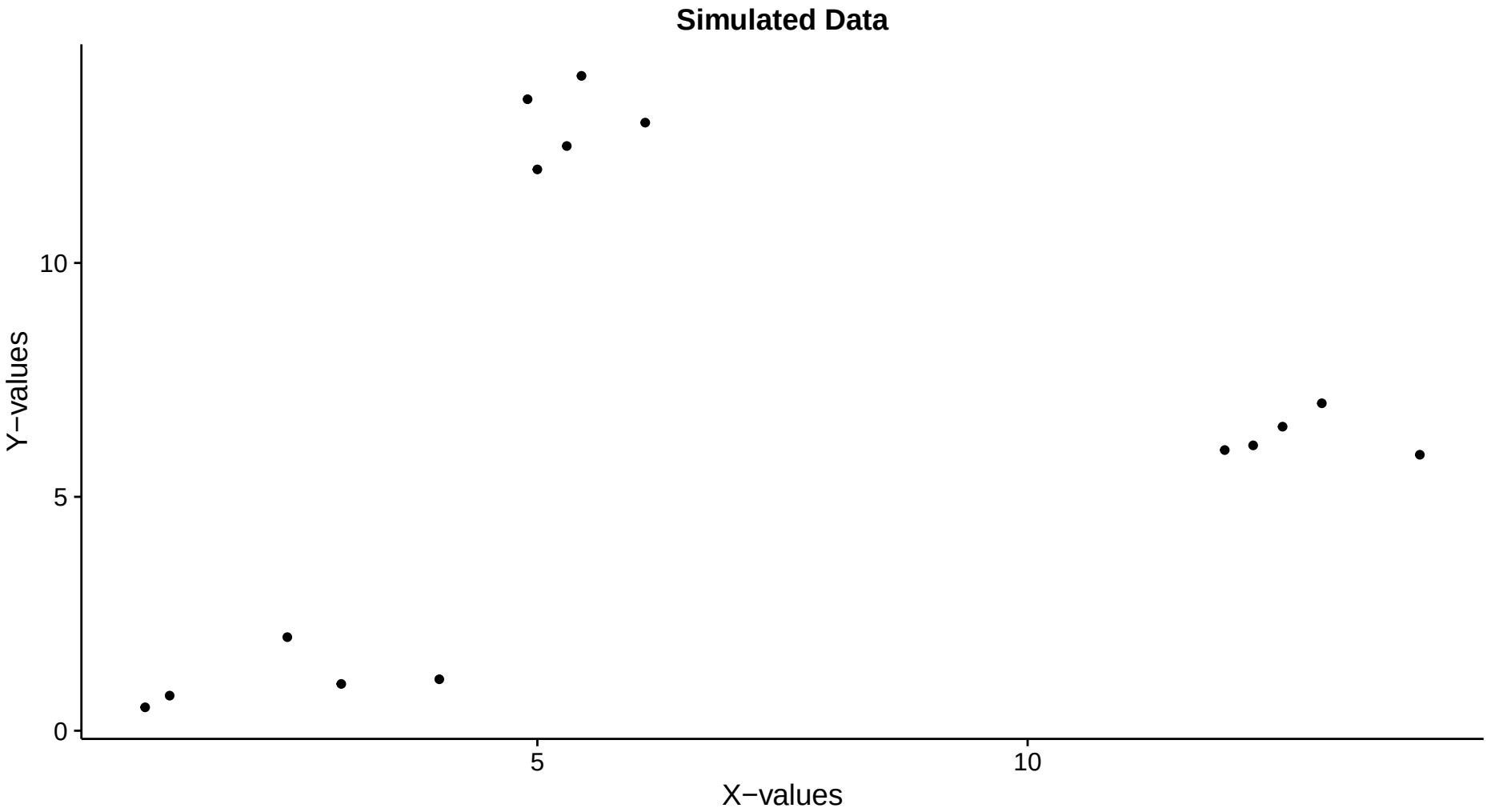
Outline

In this lesson, we discuss cluster analysis and explore the unsupervised machine learning method known as K-medians.

1. K-Medians Method and The Difference
2. K-Medians Compared to K-Means: Iris Dataset
3. In-class Activity
4. Python and R Libraries

Clustering Example

We will consider the simulated data from yesterday where clusters are clearly defined.



K-Medians Method and The Difference

Given a set of observations $\{x_1, x_2, \ldots, x_n\}$, where each observation is an m-dimensional real vector, k-medians clustering aims to partition the n observations into k sets $P = \{P_1, P_2, \ldots, P_k\}$ so as to minimize

$$\sum_{i=1}^{k} \sum_{x \in P_j} ||x - m_i||_1$$

where m_i is the geometric median of the points in P_i and $||\mathbf{v}||_1 = |v_1| + |v_2| + \cdots + |v_n|$.

- The geometric median is defined as

$$m = \arg\min_{y \in \mathbb{R}^m} \sum_{x \in P_j} ||x - y||_2$$

- Why? Median is resistant to outliers.
- There isn't a closed form that can determine the geometric median.
  - There is a nice closed form for the $||\cdot||_2^2$ cost. See the math homework!

K-Medians, The Method

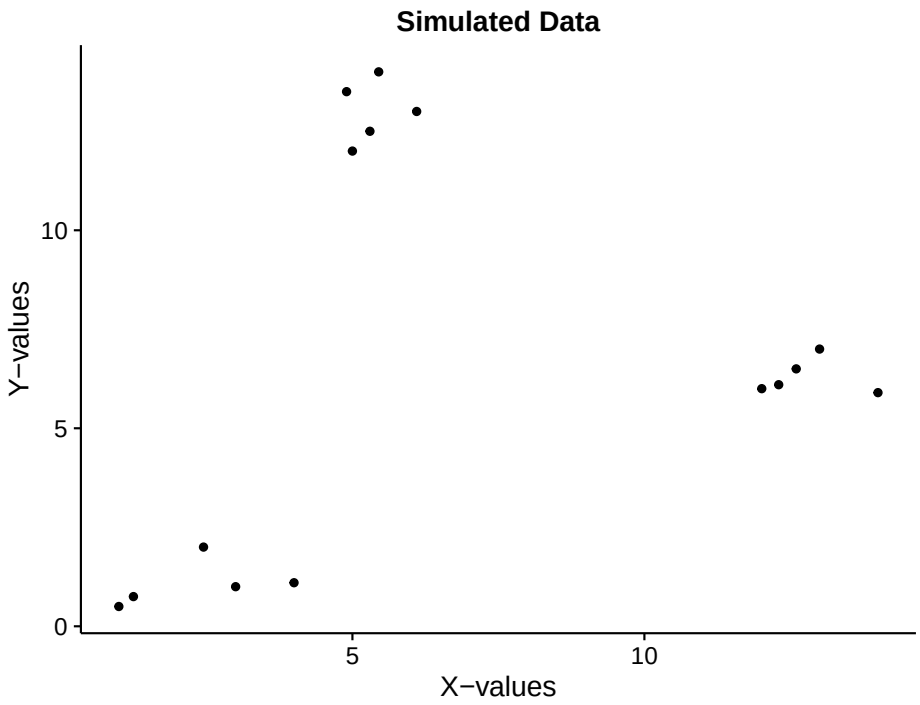
Outline for Code

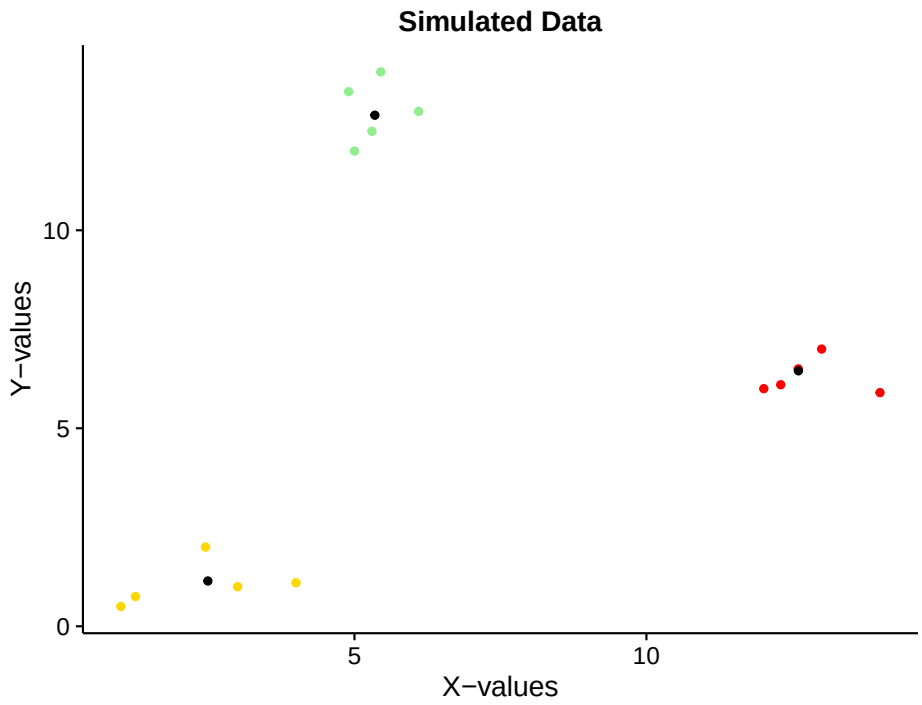
1. Parameters: data, k, maxiter
2. Initialize the k medians (using k means)
3. Create a vector for cluster assignments
4. Then, within our maxiters
    - we loop over each data point
        - calculate its distance to the medians
        - the median closest to point gives the cluster assignment
    - update the median

Again

- Algorithm has "converged" when assignments no longer change
- There is no guarantee a global optimum is found. So, when to stop?
- Many different implementations - also, many methods to find geometric median

Simulated Data Results



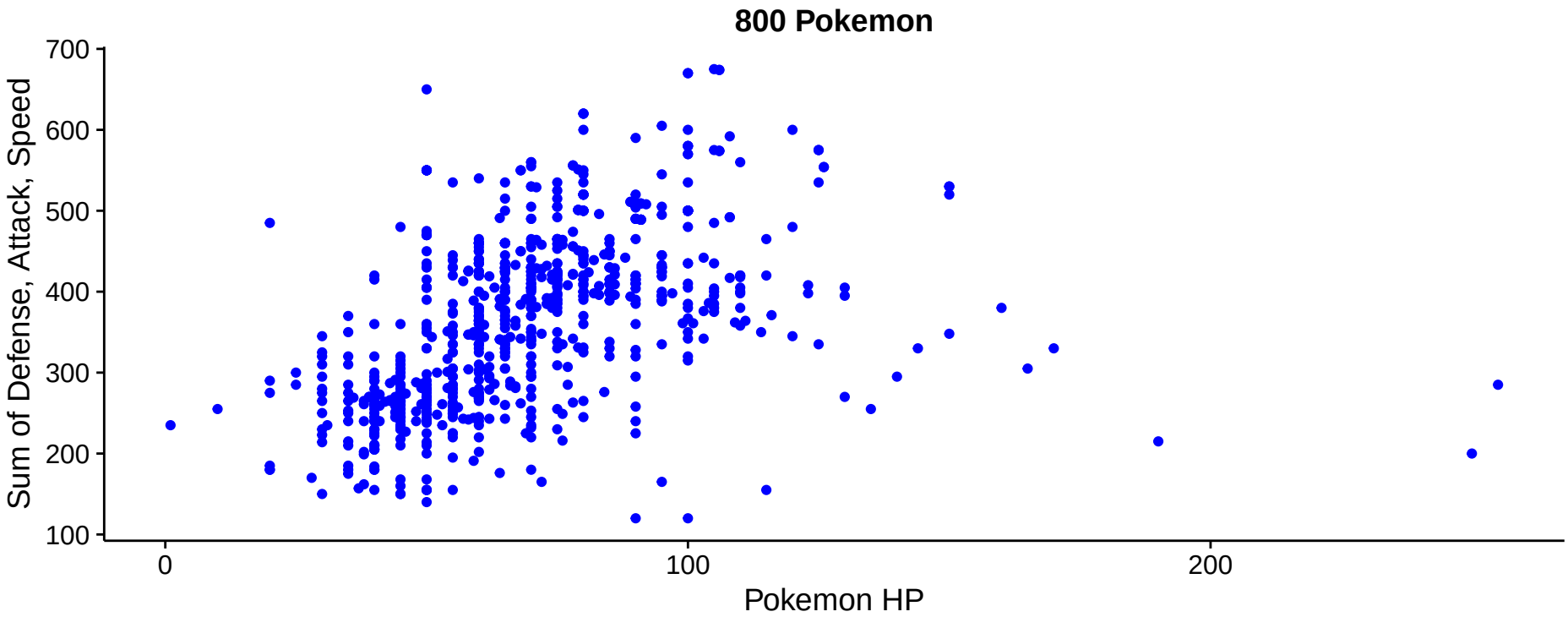


Note, K-medians run 10 iterations.

Pokemon

"... Gotta Catch Em All"

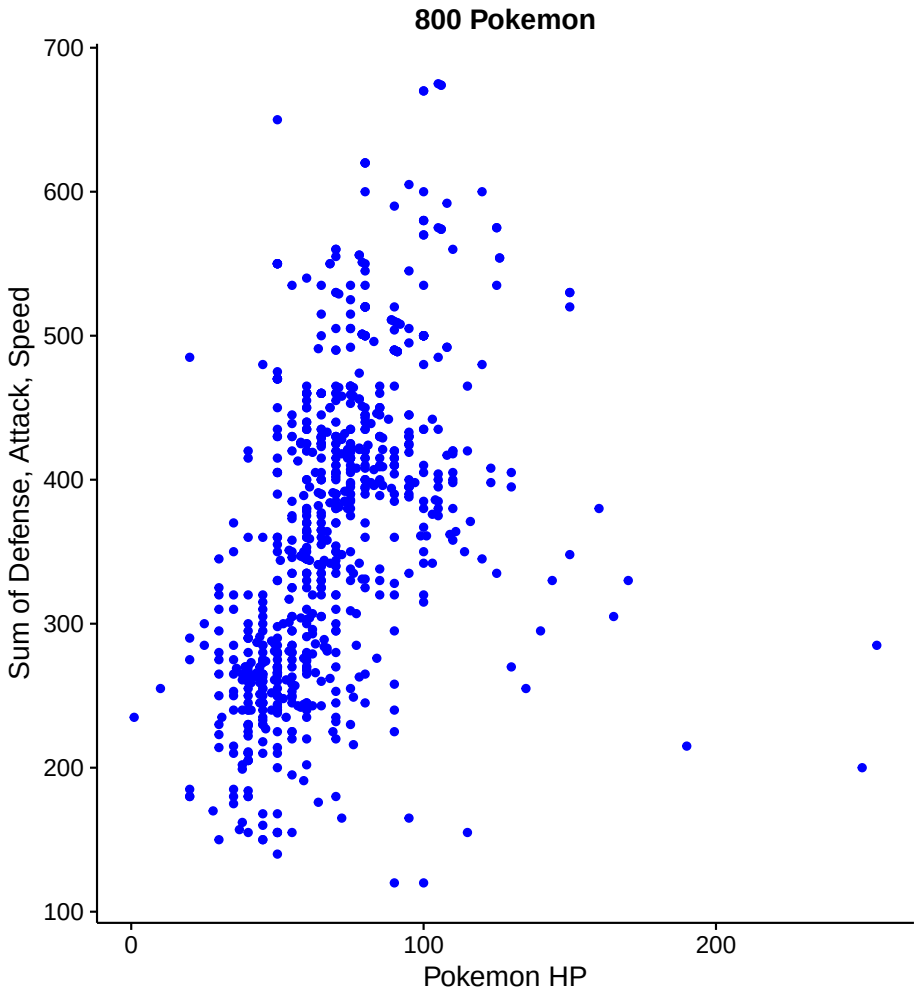
There are 800 Pokemon, classified by stats. Some are stronger than others!

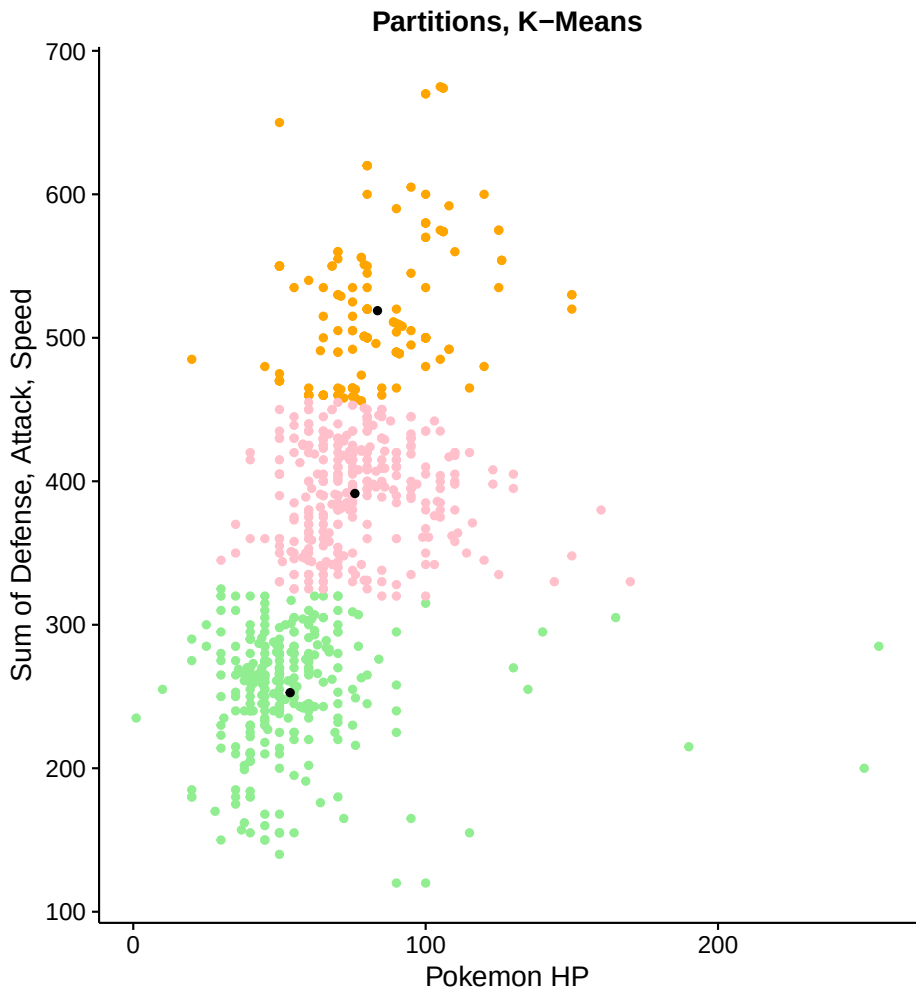


Example - PTCG

K-Means Result

"... it's You and Me"

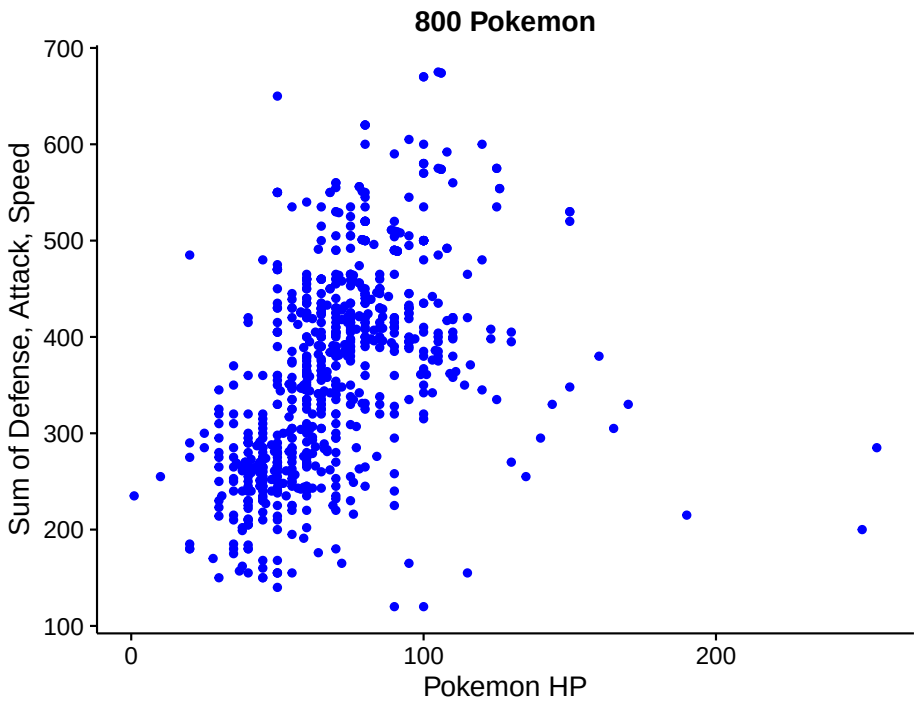


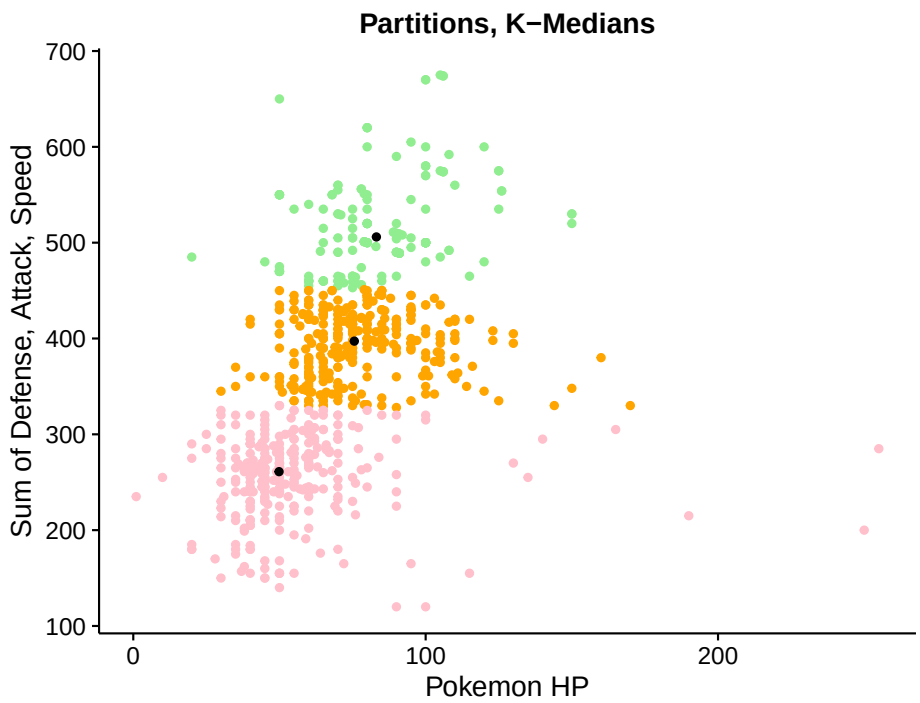


K-Medians

"... I Know It's My Destiny"

We'll explore the K-Medians method, now, using the same data. Why median?

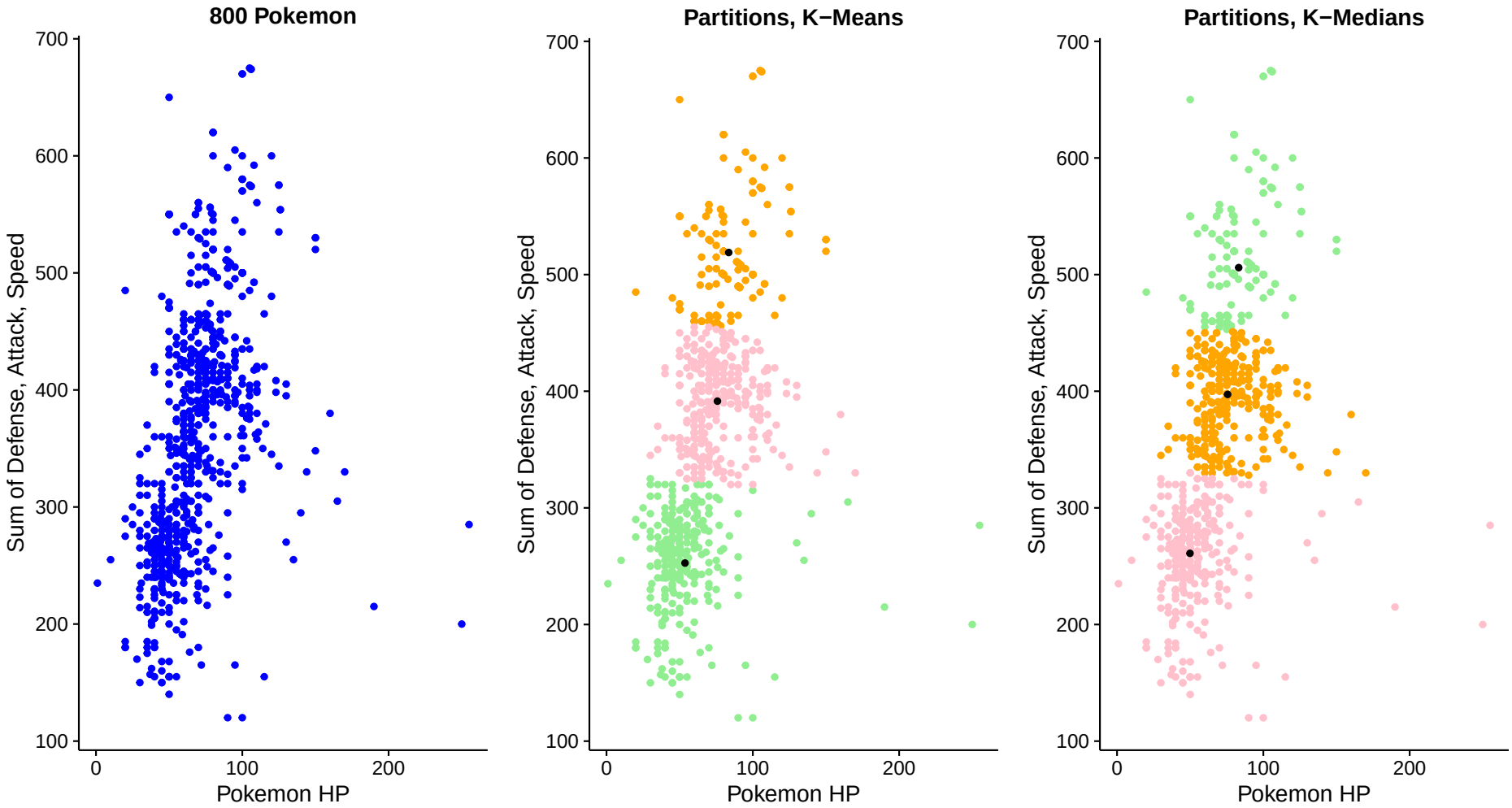




How do we measure accuracy?

K-Medians v. K-Means

"... You're My Best Friend, In A World We Must Defend"



In-class Activity (20 - 25 minutes)

- Using the description of K-Medians, attempt to implement your own K-medians.
- For a guide on how to begin, please look at the outlined code on Git

`https://src-code.simons-rock.edu/git/MATH_CMPT_370_S17/K_medians`

- Practice your data visualization – Others need to be able to understand your results.

What you don't finish in class, please do for homework.

Using Pre-Created Functions

- R has a function called *kmeans* available in the **stats** library. To use the function, you must provide a matrix of the data to cluster and an initial set of centers.
- R has a function called *kGmedian* available in the **Gmedian** library.
- Python has a *kmeans* and *kmedians* function in the **pyclustering.cluster** module

References

- Vardi, Y. and Zhang, C. **A modified Weiszfeld algorithm for the Fermat-Weber location problem**. 2001.
- Whelan, C., Harrell, G. and Wang, J. **Understanding the K-Medians Problem**. 2015.

The Median as 1-Norm Minimizer

We look at the optimization problem

$$\min_y J(y) = \min_y \sum_{i=1}^{n} |x_i - y|$$

The median is the point at which 50% of data points are to the left and 50% of data points are to the right. If we consider the convexity of the absolute value function $f(t) = |t|$, its minimum point IS the median!