K-Means Clustering using C++

By Vivian Tran

What is K-Means Clustering

- Data Clustering Technique
- What does it do
  - Creates/Finds groups within a random dataset. For this instance, I will be using a pokemon statistic dataset.
  - The number of groups is represented through variable K.

Background on Dataset

- The dataset that this project will operate on contains data on 721 Pokemon. This data includes the corresponding number, name, first and second type, and basic stats: HP, Attack, Defense, Special Attack, Special Defense, and Speed.

A	B	C	D	E	F	G	H	I	J	K
#	Name	Type 1	Type 2	Total	HP	Attack	Defense	Sp. Atk	Sp. Def	Speed
1	Bulbasaur	Grass	Poison	318	45	49	49	65	65	45
2	Ivysaur	Grass	Poison	405	60	62	63	80	80	60
3	Venusaur	Grass	Poison	525	80	82	83	100	100	80
3	VenusaurM	Grass	Poison	625	80	100	123	122	120	80
4	Charmande	Fire	-	309	39	52	43	60	50	65
5	Charmeleo	Fire	-	405	58	64	58	80	65	80
6	Charizard	Fire	Flying	534	78	84	78	109	85	100
6	CharizardM	Fire	Dragon	634	78	130	111	130	85	100
6	CharizardM	Fire	Flying	634	78	104	78	159	115	100
7	Squirtle	Water	-	314	44	48	65	50	64	43
8	Wartortle	Water	-	405	59	63	80	65	80	58
9	Blastoise	Water	-	530	79	83	100	85	105	78
9	BlastoiseM	Water	-	630	79	103	120	135	115	78
10	Caterpie	Bug	-	195	45	30	35	20	20	45
11	Metapod	Bug	-	205	50	20	55	25	25	30
12	Butterfree	Bug	Flying	395	60	45	50	90	80	70
13	Weedle	Bug	Poison	195	40	35	30	20	20	50
14	Kakuna	Bug	Poison	205	45	25	50	25	25	35
15	Beedrill	Bug	Poison	395	65	90	40	45	80	75
15	BeedrillMe	Bug	Poison	495	65	150	40	15	80	145
16	Pidgey	Normal	Flying	251	40	45	40	35	35	56
17	Pidgeotto	Normal	Flying	349	63	60	55	50	50	71

Figure 1: Preview of Pokemon Dataset

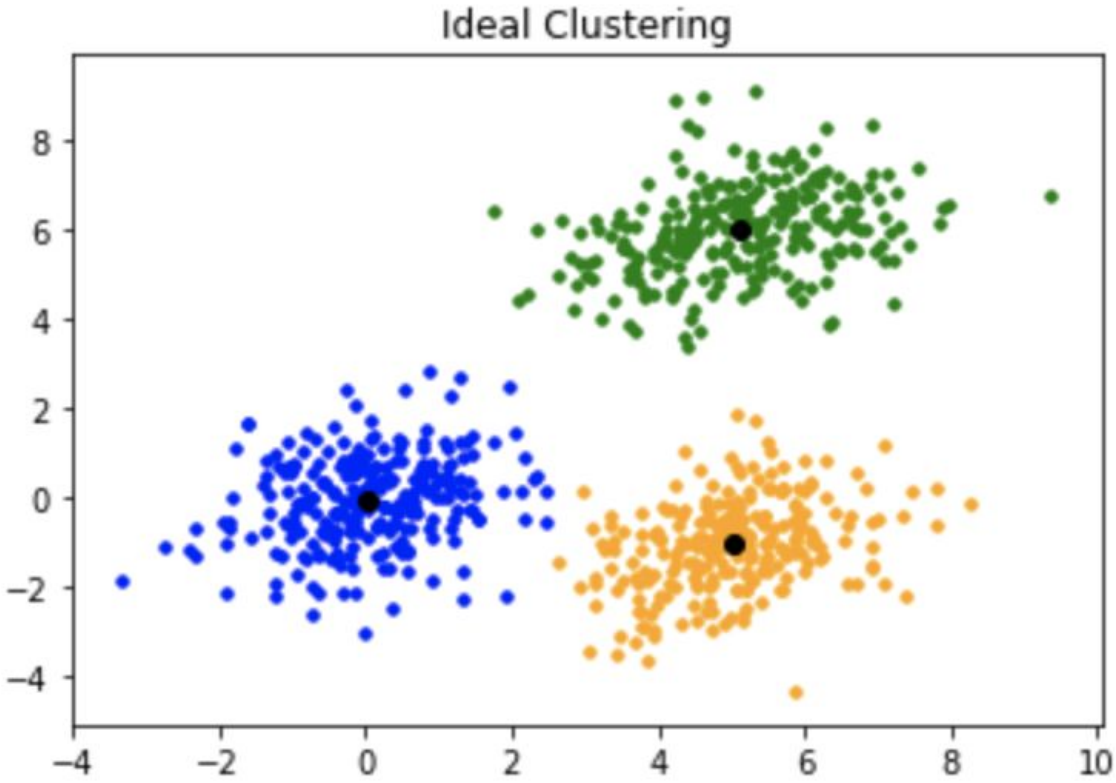


Figure 1: Ideal Clustering Imagery
https://www.geeksforgeeks.org/ml-k-means-algorithm/

Problem and Motivation

Problem: To take the Pokemon dataset and make several and yet separate observations into an appropriate amount of clusters

Motivation: To cluster the Pokemon based off of their skills with a number of clusters;

To analyze what Pokemon would be the strongest

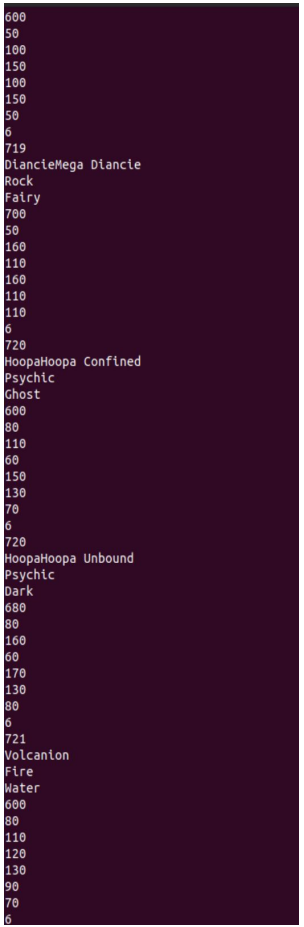
MPI Commands Used

- MPI_Init:
  - Initialization
- MPI_Comm_rank:
  - Determines rank
- MPI_Comm_size:
  - Determines size
- MPI_Finalize:
  - Finalization

- MPI_Bcast:
  - Broadcasts the message from the process
- MPI_Allreduce:
  - Combines values and sends results to all the other processes
- MPI_Barrier:
  - Blocks processes until all other processes are complete

Technical Difficulties and how they were addressed

- Trouble reading and parsing through PokedexCluster.csv file
    - Solution: Managed to find a method to read the file and use getline method to parse through data into elements using the comma as a separator



Difficulties unable to solve at this time

- Ability to isolate the data into the appropriate categories
- Randomizing and selecting the initial centroids from the dataset

Shortcomings:
None